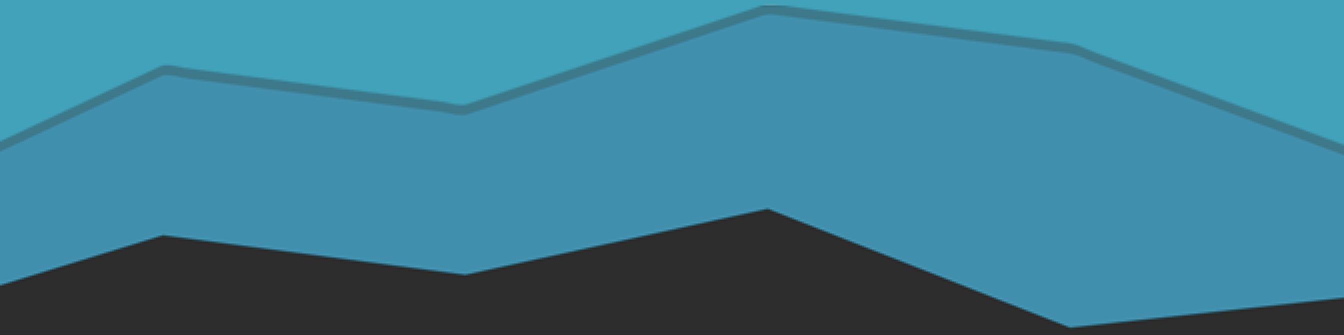
DATA + DESIGN

A simple introduction to preparing and visualizing information



By Trina Chiasson, Dyanna Gregory, and all of these people.

With support from Infoactive (https://infoactive.co/) and the Donald W. Reynolds Journalism Institute (http://www.rjionline.org/).

DATA + DESIGN

A simple introduction to preparing and visualizing information

By Trina Chiasson, Dyanna Gregory, and all of these people, with support from Infoactive (https://infoactive.co/) and the Donald W. Reynolds Journalism Institute (http://www.rjionline.org/)

Over 50 people worked hard over the course of many months to create this book, which we're delighted to offer as a free resource. If you do use or build upon this work, make sure to give credit to the contributors who made it happen by including a link to infoactive.co/data-design (https://infoactive.co/data-design).

The source code can be found and forked on Github (https://github.com/infoactive/data-design/).

TABLE OF CONTENTS

Introduction

Data Fundamentals

Collecting Data

Preparing Data

Visualizing Data

What Not To Do

Conclusion

PREFACE

I hadn't written a book before, let alone an international, open-source book with more than 50 contributors from 14 different countries.

It started with a message on Kickstarter:

> *Hi Trina! Stats dork from Chicago here….Do you have any plans to include tutorials for basic data cleaning and data selection techniques for users who may not have any statistics background?*

At the time, I didn't know that this one message would turn into a book, a community, and a global endeavor to make information design more accessible.

The message author, Dyanna Gregory, was a statistical programmer who knew the challenges of teaching university-level stats to those who don't identify as math nerds. I was an entrepreneur building Infoactive, a web application to help people create interactive infographics and data visualizations. I was also a Reynolds Fellow at the Donald W. Reynolds Journalism Institute where my goal was to find ways to simplify the process of making data visualizations in newsrooms. I had launched a Kickstarter campaign to support Infoactive, and we had nearly 1,500 backers who were excited about breaking down barriers in data visualization.

We all believed in the vision of making data simple.

But working with data can be far from simple. Data come in all different shapes, sizes, and flavors. There's no one-size-fits-all solution to collecting, understanding, and visualizing information. Some people spend years studying the topic through statistics, mathematics, design, and computer science. And many people want a bit of extra help getting started.

Dyanna and I began talking about what a plain-language data resource would look like. Usability was a big priority for us. It's hard to write a technical book that's easy for less-technical readers to digest, but we believed that it was an important challenge to tackle. We wanted to create a free resource that was well-designed, joyful to read, and easy to understand.

Of course, the information would need to be accurate and we wanted to cover a range of different data concepts. We needed technical writers with an in-depth understanding of data, math, and statistics. We also needed editors who could comb over the content to make adjustments for simplicity and understandability and ensure that the chapters had a friendly, conversational tone. Our mission was to translate geek into non-geek — to make technical concepts more accessible to less-technical people.

Dyanna and I made a call for contributors. Neither of us expected to see such a strong, positive response. Frankly, it blew our minds. Messages began pouring in from people from all over the globe who told us about their experiences working with data and design, or their lack thereof. What struck me the most was the number of self-identified "non-math people" who were hungry for a resource that could introduce them to data concepts in a manner that was simple, approachable, and even fun. They were motivated to pitch in and volunteer their time to make it happen.

Dyanna and I kicked off the project in February with a Write-A-Thon in Chicago. We invited writers, data analysts, programmers, designers, and others to come together in person and talk about the project as a whole. We thought through the process, talked about data, opened up our laptops, and started writing.

After that, we contacted everyone who applied to contribute, and began figuring out who was going to do what. For an all-volunteer project with more than 50 contributors spread across many time zones, it was important to stay organized. We found project managers for different sections of the book, chose writers and editors for each chapter, and put together research and distribution teams. Chapter by chapter, the vision began to turn into a reality.

For me, the best part of this project was having the honor of working with some of the smartest, funniest, most creative people I've ever met. Again and again, I've been blown away by their creativity, passion, and support. I'm beyond grateful that I had the experience of creating this book with such an incredible group of people. *Data + Design* is truly a community effort.

This book isn't a final product. It's the beginning of a community process to improve our collective understanding of data and design. We're releasing the first edition now, but we're already working on more chapters for future releases and thinking about ways that we can improve. Together we can build upon it, translate it, transform it, and make it better with every iteration.

Data + Design is open source and available on Github. It's free for anyone to read, download, remix, and re-distribute for noncommercial purposes. We invite you to join us. Email ebook@infoactive.co to get involved.

I also want to thank the Donald W. Reynolds Journalism Institute (RJI) (http://www.rjionline.org/) for supporting *Data + Design*. RJI's support for journalism and data storytelling played an instrumental role in bringing this project to life.

TRINA CHIASSON
CO-FOUNDER & CEO, INFOACTIVE
2013-2014 REYNOLDS FELLOW

FOREWORD

Data are all around us and always have been. Everything throughout history has always had the potential to be quantified: theoretically, one could count every human who has ever lived, every heartbeat that has ever beaten, every step that was ever taken, every star that has ever shone, every word that has ever been uttered or written. Each of these collective things can be represented by a number. But only recently have we had the technology to efficiently surface these hidden numbers, leading to greater insight into our human condition.

But what does this mean, exactly? What are the cultural effects of having easy access to data? It means, for one thing, that we all need to be more data literate. It also means we have to be more design literate. As the old adage goes, statistics lie. Well, data visualizations lie, too. How can we learn how to first, effectively read data visualizations; and second, author them in such a way that is ethical and clearly communicates the data's inherent story?

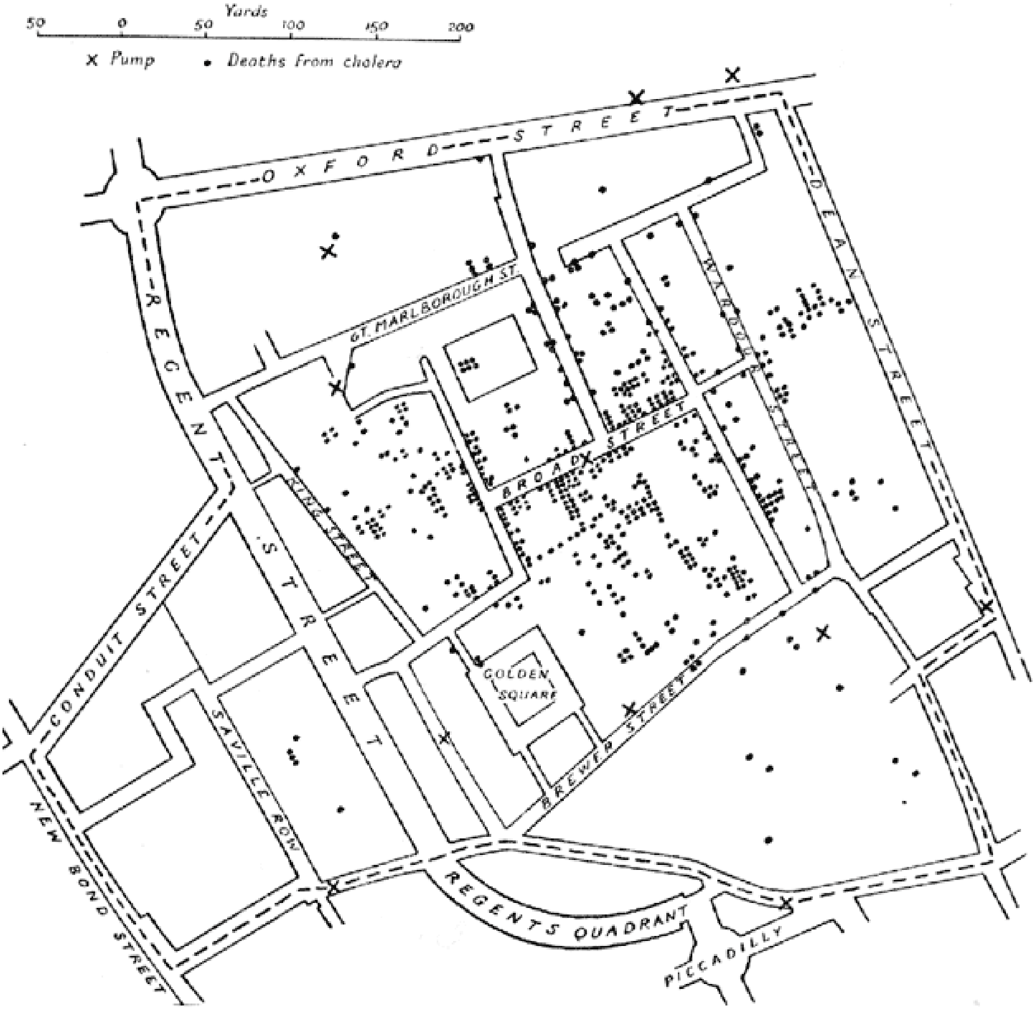
> *"At the intersection of art and algorithm, data visualization schematically abstracts information to bring about a deeper understanding of the data, wrapping it in an element of awe.*
>
> *Maria Popova, Stories for the Information Age, Businessweek (http://www.businessweek.com/innovate/content/aug2009/id20090811_137179.htm)*

My favorite description of data visualization comes from the prolific blogger, Maria Popova, who said that data visualization is "at the intersection of art and algorithm." To learn about the history of data visualization is to become an armchair cartographer, explorer, and statistician.

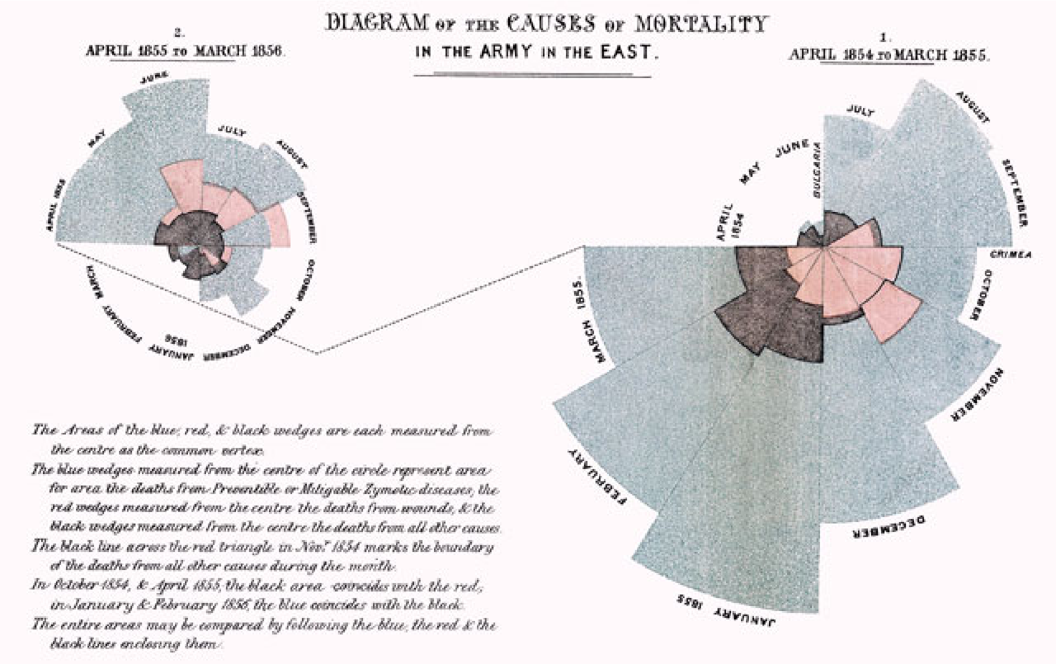
Early visual explorations of data focused mostly on small snippets of data gleaned to expand humanity's understanding of the geographical world, mainly through maps. Starting with the first recognized world maps of the 13th century, scientists, mathematicians, philosophers, and sailors used math to visualize the invisible. Stars and suns were plotted, coastlines and shipping routes charted. Data visualization, in its native essence, drew the lines, points, and coordinates that gave form to the physical world and our place in it. It answered questions like “Where am I?”, “How do I get there?”, and “How far is it?”

Early data visualizations were also used to answer questions pertaining to issues of public health. Epidemiologist John Snow’s 1854 London cholera map was created to record instances of cholera in a London neighborhood, pinpointing the cause of the outbreak to a single well. This knowledge gained from patterns in lists of names, numbers, and locations was then used to persuade London’s populace to install sewer systems to alleviate the proliferation and spread of disease. The human brain is particularly adept at recognizing patterns, and a good data visualization, like Snow’s, optimizes displays of these patterns through effective use of Gestalt theory, design principles, and color. (Or lack of it, as this case may be.)

Snow’s visualization, with its absence of color, optimizes Gestalt’s theories of visual perception, most notably “Proximity” and “Figure and Ground.” The small black dots, each one representing a single case of cholera are small black figures standing out in contrast against the ground: in this graphic, the lines and white space representing streets. The proximity of these dots around the affected well are what enabled Snow to determine the exact source of the outbreak. Today, even with our advanced computing systems and sophisticated tools for creating data visualizations, there is little you could do to improve the effectiveness of this chart. It is simple, beautiful, and true: a data visualization that saved lives.



Florence Nightingale, famous more for her nursing skills than her analytic prowess, was nonetheless also a master data scientist and storyteller. Through data presented via her signature Coxcomb diagram (also known as polar or rose charts), she convinced the British army to invest in sanitation measures after illustrating that the majority of deaths in the Crimean War were the result of preventable diseases caused by the horrible sanitary conditions in hospitals. “Why are we sick?” she asked, then answering the question herself by giving visual form to data.



Looking at this graph, it is readily apparent that preventable diseases outnumbered all other causes of death. The area in blue represents deaths by preventable diseases, measured from the center, with red representing deaths caused by injuries and black indicating all other causes. Design principles at play here include the addition of color theory to take advantage of more Gestalt principles: “Similarity” and “Continuity”. Color makes it easy for us to tell which segments belong to which category. It also helps to draw the eye in a continuous path around the graphic, making it easier to read.

There is debate over the quality of this chart. Some claim it one of the best, most memorable visualizations ever created, not solely because of its visual communication strength, but in spite of it. It is remembered because of the change it inspired. Others deride it, claiming it’s just a glorified pie chart, suffering from the same misrepresentation of the information by distorting the data: in terms of visual perception, humans have a hard time accurately judging measures represented by differences in area. Despite their ubiquity, pie charts, for this very reason, are an incredibly poor way to visualize data. A simple stacked bar chart with reference lines, while not as beautiful or visually intriguing, would have communicated more effectively and on a quicker read.

We still ask the same kinds of questions that Snow and Nightingale posed, and as curious humans, probably always will. But the major differences today are that we have the means with which to gather much larger, richer datasets. And we also have the tools with which to automate visualizing our data.

This isn't to say that small datasets, like the ones Nightingale and Snow observed, are any less valuable or interesting. Thanks to data gathering technology and advances in wearable computing and the Internet of Things, to the contrary. My friend Jen Lowe, a data scientist and artist working in New York, recently published her heartbeat on the internet (http://datatelling.com/projects/onehuman-heartbeat/). As a flat, static visualization, it would be beautiful but not especially informative. But by adding interactivity and live data feeds of her pulse via wearable sensors and API calls, her heartbeat is a living, beating, digital thing, viewable by anyone, anywhere, at any time. What you have is insight into another human being like you've never seen before.

Unique insight is the essence of data, both big and small, and the result of the tools that allow us to access, probe, poke, prod, dissect, visualize, and hopefully, make sense of it. Tools which, through the democratization of data visualization, allow us to change our lens on the world, creating pictures of humanity from different perspectives, bringing into focus stories about humanity and the world that were previously invisible, allowing us insight into ourselves like we've never seen before.

CALLIE NEYLAN
SENIOR DESIGNER, MICROSOFT

HOW TO USE THIS BOOK

Structure of the Book

The book's chapters follow pretty closely the actual set of steps that need to be accomplished before data can be visualized, from the design of the survey to the collection of the data to ultimately its visualization. Really pretty simple. So if you're fairly new to the game, it is probably going to make sense for you to start at the beginning and read the whole way through the book. Likewise, if you're familiar with some of the content, but not everything, there's no harm in just skipping to the chapters that are relevant to your case: you aren't going to be scrambling to catch up if you skip around!

If you don't have much (or any) experience with survey design and data collection and visualization, it would probably be best to read this book all the way through before you start collecting your data. Some of the visualizations you want to make when everything is said and done might require some well-planned decisions up front. If you're reading this book part-by-part as you go, there's a better chance that you might not consider something until it's too late!

Common Terms and Definitions

Throughout the book, there may be times when we introduce specialized terms and phrases that are rather specific to the topic being discussed. Whenever this is the case, we will provide a link to the term in our glossary. You're welcome to reference this at any time, but we do our best to point you in that direction when a new term is used.

Advanced Chapters

This book does contain some chapters that go a bit above and beyond just the basics that we think you should know. We do denote these chapters with a warning at the beginning of the chapter so that you know to expect a more in-depth look at the topic. Reading these chapters (although we recommend it!) is not necessary for an adequate understanding of the broader topic being discussed; however, there are times when we feel that it's useful to provide you with easy access to the information and that you could benefit from a more thorough understanding of the topic.

Examples: The Good, the Bad, and the Ugly

One of our favorite parts about this book (although maybe we're a little biased) is the examples that we use throughout: many of these concepts can get kind of abstract and having a concrete example to refer to really helps to keep things straight. That said, there are a few different styles that we use to indicate good and bad examples, dos and don'ts, etc.

QUESTION	ANSWER
What are we hoping to find?	The answer lives here.

GOOD EXAMPLES

Good examples are always accompanied by a green or blue bar, a checkmark, or a thumbs-up symbol, depending on what's being presented (a graphic, a data table, a procedure, something else).

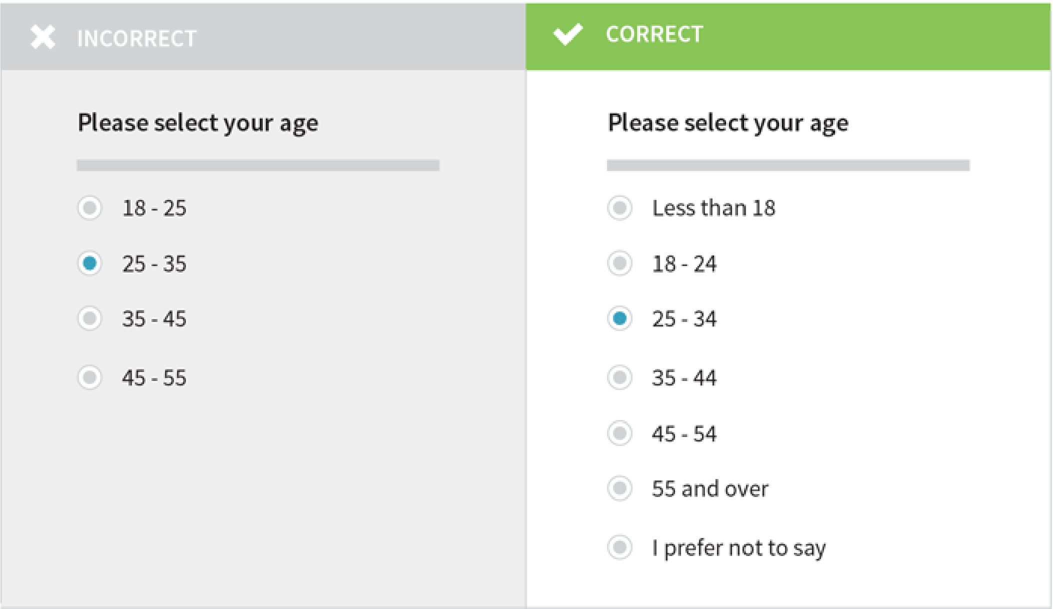
BAD EXAMPLES

Conversely, whenever we illustrate what **not** to do in a given situation, we will include a gray bar or cross out the example with a large X or include a thumbs-down symbol to accompany it. These different styles will look like:

PROS	CONS
The benefits of an action go here.	And the negatives go here.

✓ DO	✗ DON’T
This column lists good tips and advice.	This column shows what you should avoid.

Sometimes we also mark good and bad examples using images



✗ INCORRECT

Please select your age

○ 18 - 25
● 25 - 35
○ 35 - 45
○ 45 - 55

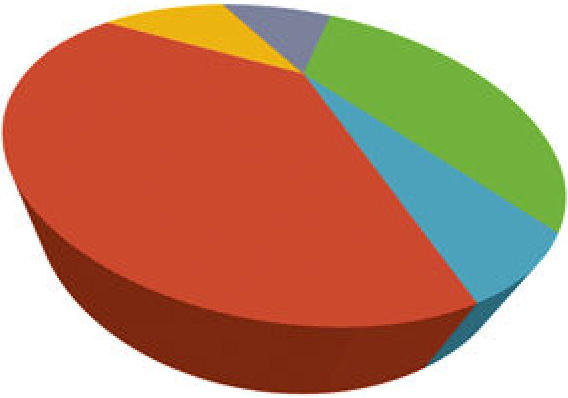
✓ CORRECT

Please select your age

○ Less than 18
○ 18 - 24
● 25 - 34
○ 35 - 44
○ 45 - 54
○ 55 and over
○ I prefer not to say

An image preceded by a gray box with an “x” means that it’s a graphical example of a practice that you should avoid.

✖ THIS IS AN EXAMPLE OF WHAT NOT TO DO



WARNINGS

Other times, we will just want to make sure that you're careful and don't forget to consider an important point. When this is the case we will include a warning box, identified by a red exclamation mark. (Actually, we already used one of these up above!)

Making Contributions and Corrections

If you're reading through the book and notice a typo, or think that something is incorrectly or ambiguously explained, or get a great idea for a new section or chapter, let us know! There are two ways to do this: you can either fill out this form (https://docs.google.com/a/infoactive.us/forms/d/1LsafHUV-BPQHmQsXHR40UsXS4f0c_jySgMrF9vMloF4/viewform?usp=send_form) or, if you're a little bit more technical, follow the directions that we have on GitHub (https://github.com/infoactive/data-design#can-i-make-edits).

Likewise, if you just aren't sure about something after reading about it, ask us for a better explanation! It's likely that we didn't realize that we could have explained it better (but that doesn't mean we don't want to!). Tweet us at @DataDesignBook (http://twitter.com/DataDesignBook), email ebook@infoactive.co, or check out HelpMeViz (http://helpmeviz.com/) for even more resources. We'll do our best to get your question squared away and will make sure that future editions do a better job addressing the issue.

DATA FUNDAMENTALS

When you cook, you go through certain steps in a consistent order. You figure out which ingredients you need, you gather them, you prep them, and finally you cook them and present your finished dish.

Creating a data visualization is a lot like cooking. You decide what data you need, you collect it, you prepare and clean it for use, and then you make the visualization and present your finished result.

When you're cooking, it helps to understand what each ingredient does so you know what you can do with it. For example, salt and sugar look similar but achieve very different results!

In this section, we'll talk about what the basic ingredient groups of data are and what can and can't be done with them so you have a sense of how to properly work with them later on.

CHAPTER 1

BASIC DATA TYPES

BY MICHAEL CASTELLO

There are several different basic data types and it’s important to know what you can do with each of them so you can collect your data in the most appropriate form for your needs. People describe data types in many ways, but we’ll primarily be using the levels of measurement known as nominal, ordinal, interval, and ratio.

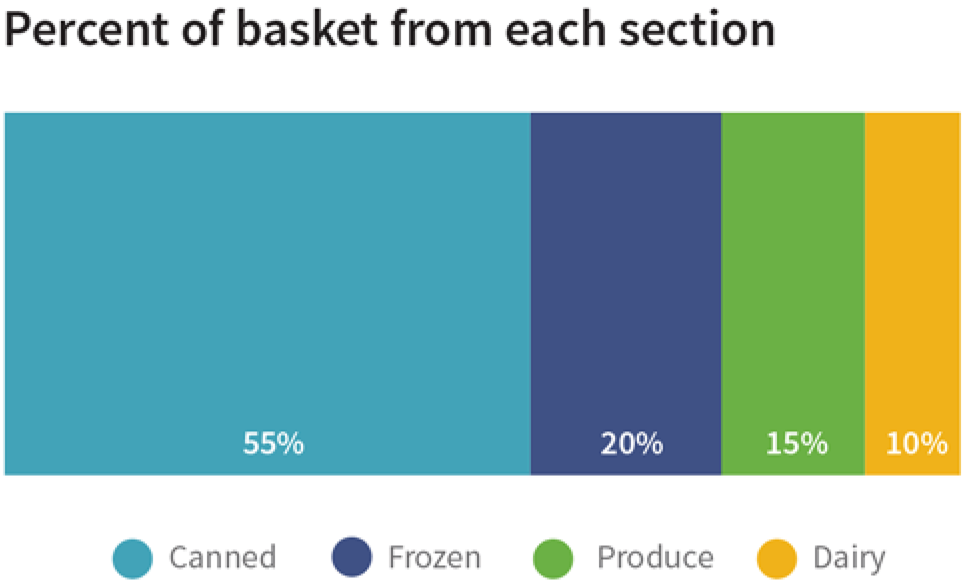
Levels of Measurement

Let’s say you’re on a trip to the grocery store. You move between sections of the store, placing items into your basket as you go. You grab some fresh produce, dairy, frozen foods, and canned goods. If you were to make a list that included what section of the store each item came from, this data would fall into the nominal type. The term nominal is related to the Latin word “nomen,” which means “pertaining to names;” we call this data nominal data because it consists of named categories into which the data fall. Nominal data is inherently unordered; produce as a general category isn’t mathematically greater or less than dairy.

NOMINAL

Nominal data can be counted and used to calculate percents, but you can’t take the average of nominal data. It makes sense to talk about how many items in your basket are from the dairy section or what percent is produce, but you can’t calculate the average grocery section of your basket.

When there are only two categories available, the data is referred to as dichotomous. The answers to yes/no questions are dichotomous data. If, while shopping, you collected data about whether an item was on sale or not, it would be dichotomous.

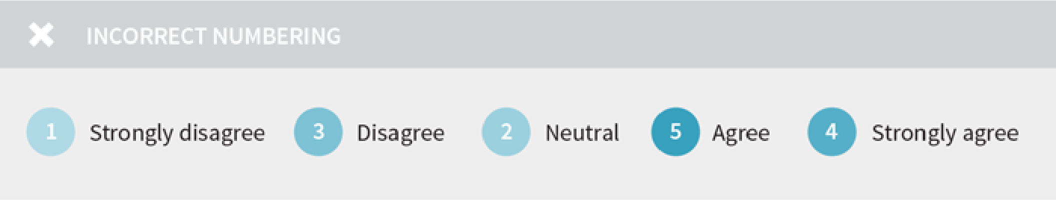


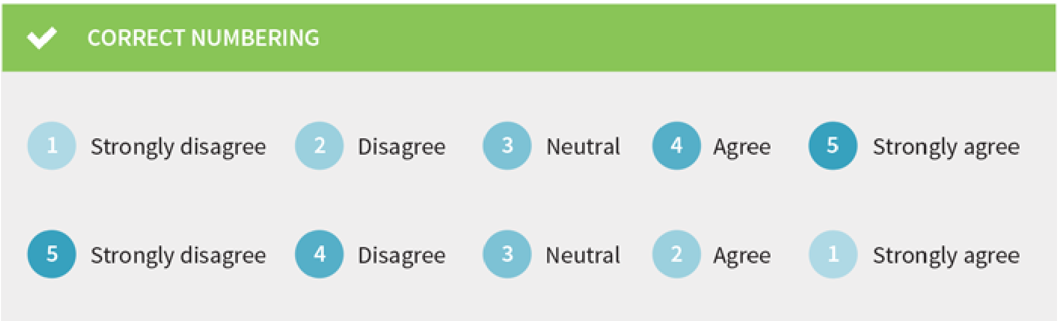
ORDINAL

At last, you get to the checkout and try to decide which line will get you out of the store the quickest. Without actually counting how many people are in each queue, you roughly break them down in your mind into short lines, medium lines, and long lines. Because data like these have a natural ordering to the categories, it's called ordinal data. Survey questions that have answer scales like "strongly disagree," "disagree," "neutral," "agree," "strongly agree" are collecting ordinal data. No category on an ordinal scale has a true mathematical value. Numbers are often assigned to the categories to make data entry or analysis easier (e.g. 1 = strongly disagree, 5 = strongly agree), but these assignments are arbitrary and you could choose any set of ordered numbers to represent the groups. For instance, you could just as easily decide to have 5 represent "strongly disagree" and 1 represent "strongly agree."

The numbers you select to represent ordinal categories do change the way you interpret your end analysis, but you can choose any set you wish as long as you keep the numbers in order.

It is most common to use either 0 or 1 as the starting point.





Like nominal data, you can count ordinal data and use them to calculate percents, but there is some disagreement about whether you can average ordinal data. On the one hand, you can't average named categories like "strongly agree" and even if you assign numeric values, they don't have a true mathematical meaning. Each numeric value represents a particular category, rather than a count of something.

On the other hand, if the difference in degree between consecutive categories on the scale is assumed to be approximately equal (e.g. the difference between strongly disagree and disagree is the same as between disagree and neutral, and so on) and consecutive numbers are used to represent the categories, then the average of the responses can also be interpreted with regard to that same scale.

> Some fields strongly discourage the use of ordinal data to do calculations like this, while others consider it common practice. You should look at other work in your field to see what usual procedures are.

INTERVAL

Enough ordinal data for the moment… back to the store! You've been waiting in line for what seems like a while now, and you check your watch for the time. You got in line at 11:15am and it's now 11:30. Time of day falls into the class of data called interval data, so named because the interval between each consecutive point of measurement is equal to every other. Because every minute is sixty seconds, the difference between 11:15 and 11:30 has the exact same value as the difference between 12:00 and 12:15.

Interval data is numeric and you can do mathematical operations on it, but it doesn't have a "meaningful" zero point – that is, the value of zero doesn't indicate the absence of the thing you're measuring. 0:00 am isn't the absence of time, it just means it's the start of a new day. Other interval data that you encounter in everyday life are calendar years and temperature. A value of zero for years doesn't mean that time didn't exist before that, and a temperature of zero (when measured in C or F) doesn't mean there's no heat.

RATIO

Seeing that the time is 11:30, you think to yourself, "I've been in line for fifteen minutes already…???" When you start thinking about the time this way, it's considered ratio data. Ratio data is numeric and a lot like interval data, except it *does* have a meaningful zero point. In ratio data, a value of zero indicates an absence of whatever you're measuring—zero minutes, zero people in line, zero dairy products in your basket. In all these cases, zero actually means you don't have any of that thing, which differs from the data we discussed in the interval section. Some other frequently encountered variables that are often recorded as ratio data are height, weight, age, and money.

Interval and ratio data can be either discrete or continuous. Discrete means that you can only have specific amounts of the thing you are measuring (typically integers) and no values in between those amounts. There have to be a whole number of people in line; there can't be a third of a person. You can have an *average* of, say, 4.25 people per line, but the actual count of people has to be a whole number. Continuous means that the data can be any value along the scale. You can buy 1.25 lbs of cheese or be in line for 7.75 minutes. This doesn't mean that the data have to be able to take all possible numerical values – only all the values within the bounds of the scale. You can't be in line for a negative amount of time and you can't buy negative lbs of cheese, but these are still continuous.

> For simplicity in presentation, we often round continuous data to a certain number of digits. These data are still continuous, not discrete.

To review, let's take a look at a receipt from the store. Can you identify which pieces of information are measured at each level (nominal, ordinal, interval, and ratio)?

Date: 06/01/2014 Time: 11:32am				
Item	Section	Aisle	Quantity	Cost (US$)
Oranges—Lbs	Produce	4	2	2.58
Apples—Lbs	Produce	4	1	1.29
Mozzarella—Lbs	Dairy	7	1	3.49
Milk—Skim—Gallon	Dairy	8	1	4.29
Peas—Bag	Frozen	15	1	0.99
Green Beans—Bag	Frozen	15	3	1.77
Tomatoes	Canned	2	4	3.92
Potatoes	Canned	3	2	2.38
Mushrooms	Canned	2	5	2.95

Variable Type Vs. Data Type

If you look around the internet or in textbooks for info about data, you'll often find variables described as being one of the data types listed above. Be aware that many variables aren't exclusively one data type or another. What often determines the data type is how the data are collected.

Consider the variable age. Age is frequently collected as ratio data, but can also be collected as ordinal data. This happens on surveys when they ask, "What age group do you fall in?" There, you wouldn't have data on your respondent's individual ages – you'd only know how many were between 18-24, 25-34, etc. You might collect actual cholesterol measurements from participants for a health study, or you may simply ask if their cholesterol is high. Again, this is a single variable with two different data collection methods and two different data types.

The general rule is that you can go down in level of measurement but not up. If it's possible to collect the variable as interval or ratio data, you can also collect it as nominal or ordinal data, but if the variable is inherently only nominal in nature, like grocery store section, you can't capture it as ordinal, interval or ratio data. Variables that are naturally ordinal can't be captured as interval or ratio data, but can be captured as nominal. However, many variables that get captured as ordinal have a similar variable that can be captured as interval or ratio data, if you so choose.

Ordinal Level Type	Corresponding Interval/Ratio Level Measure	Example
Ranking	Measurement that ranking is based on	Record runners' marathon times instead of what place they finish
Grouped scale	Measurement itself	Record exact age instead of age category
Substitute scale	Original measurement the scale was created from	Record exact test score instead of letter grade

It's important to remember that the general rule of "you can go down, but not up" also applies during analysis and visualization of your data. If you collect a variable as ratio data, you can always decide later to group the data for display if that makes sense for your work. If you collect it as a lower level of measurement, you can't go back up later on without collecting more data. For example, if you do decide to collect age as ordinal data, you can't calculate the average age later on and your visualization will be limited to displaying age by groups; you won't have the option to display it as continuous data.

When it doesn't increase the burden of data collection, you should collect the data at the highest level of measurement that you think you might want available later on. There's little as disappointing in data work as going to do a graph or calculation only to realize you didn't collect the data in a way that allows you to generate what you need!

Other Important Terms

There are some other terms that are frequently used to talk about types of data. We are choosing not to use them here because there is some disagreement about their meanings, but you should be aware of them and what their possible definitions are in case you encounter them in other resources.

CATEGORICAL DATA

We talked about both nominal and ordinal data above as splitting data into categories. Some texts consider both to be types of categorical data, with nominal being unordered categorical data and ordinal being ordered categorical data. Others only call nominal data categorical, and use the terms "nominal data" and "categorical data" interchangeably. These texts just call ordinal data "ordinal data" and consider it to be a separate group altogether.

QUALITATIVE AND QUANTITATIVE DATA

Qualitative data, roughly speaking, refers to non-numeric data, while quantitative data is typically data that is numeric and hence quantifiable. There is some consensus with regard to these terms. Certain data are always considered qualitative, as they require pre-processing or different methods than quantitative data to analyze. Examples are recordings of direct observation or transcripts of interviews. In a similar way, interval and ratio data are always considered to be quantitative, as they are only ever numeric. The disagreement comes in with the nominal and ordinal data types. Some consider them to be qualitative, since their categories are descriptive and not truly numeric. However, since these data can be counted and used to calculate percentages, some consider them to be quantitative, since they are in that way quantifiable.

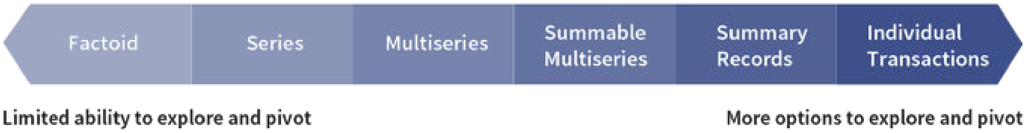
To avoid confusion, we'll be sticking with the level of measurement terms above throughout the rest of this book, except in our discussion of long-form qualitative data in the survey design chapter. If you come across terms "categorical," "qualitative data," or "quantitative data" in other resources or in your work, make sure you know which definition is being used and don't just assume!

CHAPTER 2

ABOUT DATA AGGREGATION

BY ALISTAIR CROLL

When trying to turn data into information, the data you start with matter a lot. Data can be simple factoids—of which someone else has done all of the analysis—or raw transactions, where the exploration is left entirely to the user.



Level of Aggregation	Number of metrics	Description
Factoid	Maximum context	Single data point; No drill-down
Series	One metric, across an axis	Can compare rate of change
Multiseries	Several metrics, common axis	Can compare rate of change, correlation between metrics
Summable multiseries	Several metrics, common axis	Can compare rate of change, correlation between metrics; Can compare percentages to whole
Summary records	One record for each item in a series; Metrics in other series have been aggregated somehow	Items can be compared
Individual transactions	One record per instance	No aggregation or combination; Maximum drill-down

Most datasets fall somewhere in the middle of these levels of aggregation. If we know what kind of data we have to begin with, we can greatly simplify the task of correctly visualizing them the first time around.

Let's look at these types of aggregation one by one, using the example of coffee consumption. Let's assume a café tracks each cup of coffee sold and records two pieces of information about the sale: the gender of the buyer and the kind of coffee (regular, decaf, or mocha).

The basic table of these data, by year, looks like this These are completely made up coffee data, BTW.:

Year	2000	2001	2002	2003	2004	2005	2006	2007	2008
Total sales	19,795	23,005	31,711	40,728	50,440	60,953	74,143	93,321	120,312
Male	12,534	16,452	19,362	24,726	28,567	31,110	39,001	48,710	61,291
Female	7,261	6,553	12,349	16,002	21,873	29,843	35,142	44,611	59,021
Regular	9,929	14,021	17,364	20,035	27,854	34,201	36,472	52,012	60,362
Decaf	6,744	6,833	10,201	13,462	17,033	19,921	21,094	23,716	38,657
Mocha	3,122	2,151	4,146	7,231	5,553	6,831	16,577	17,593	21,293

Factoid

A factoid is a piece of trivia. It is calculated from source data, but chosen to emphasize a particular point.

> **Example:** 36.7% of coffee in 2000 was consumed by women.

Series

This is one type of information (the dependent variable) compared to another (the independent variable). Often, the independent variable is time.

Year	2000	2001	2002	2003
Total sales	19,795	23,005	31,711	40,728

In this example, the total sales of coffee *depends* on the year. That is, the year is independent ("pick a year, any year") and the sales is dependent ("based on that year, the consumption is 23,005 cups").

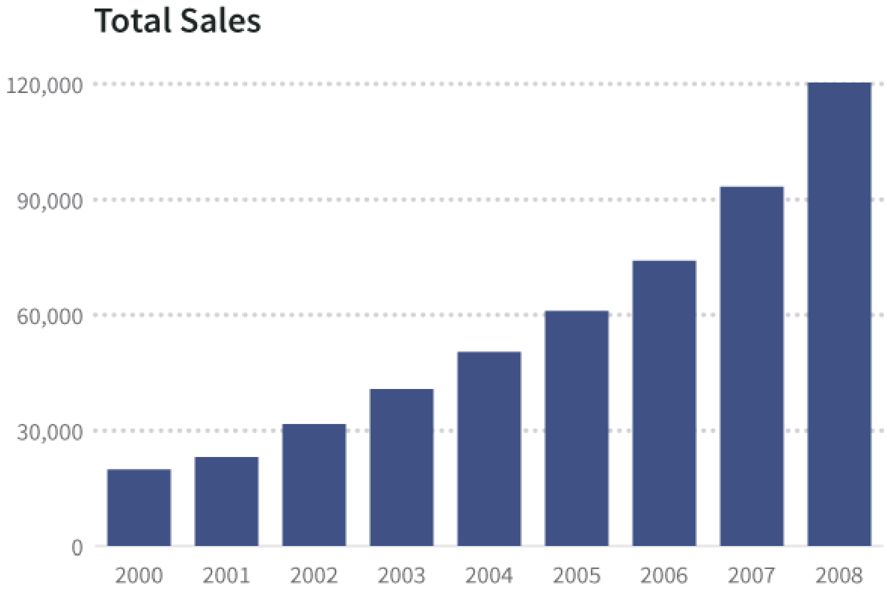
A series can also be some other set of continuous data, such as temperature. Consider this table that shows how long it takes for an adult to sustain a first-degree burn from hot water. Here, water temperature is the independent variable US Government Memorandum, Consumer Product Safety Commission, Peter L. Armstrong, Sept. 15, 1978.:

Water Temp °C (°F)	Time for 1st Degree Burn
46.7 (116)	35 minutes
50 (122)	1 minute
55 (131)	5 seconds
60 (140)	2 seconds
65 (149)	1 second
67.8 (154)	Instantaneous

And it can be a series of non-contiguous, but related, information in a category, such as major car brands, types of dog, vegetables, or the mass of planets in the solar system National Space Science Data Center, NASA: http://nssdc.gsfc.nasa.gov/planetary/factsheet/planet_table_ratio.html:

Planet	Mass relative to earth
Mercury	0.0553
Venus	0.815
Earth	1
Mars	0.107
Jupiter	317.8
Saturn	95.2
Uranus	14.5
Neptune	17.1

In many cases, series data have one and only one dependent variable for each independent variable. In other words, there is only one number for coffee consumption for each year on record. This is usually displayed as a bar, time series, or column graph.



In cases where there are several dependent variables for each independent one, we often show the information as a scatterplot or heat map, or do some kind of processing (such as an average) to simplify what's shown. We'll come back to this in the section below, *Using visualization to reveal underlying variance*.

Multiseries

A multiseries dataset has several pieces of dependent information and one piece of independent information. Here are the data about exposure to hot water from before, with additional data US Government Memorandum, Consumer Product Safety Commission, Peter L. Armstrong, Sept. 15, 1978.:

Water Temp °C (°F)	Time for 1st Degree Burn	Time for 2nd & 3rd Degree Burns
46.7 (116)	35 minutes	45 minutes
50 (122)	1 minute	5 minutes
55 (131)	5 seconds	25 seconds
60 (140)	2 seconds	5 seconds
65 (149)	1 second	2 seconds
67.8 (154)	Instantaneous	1 second

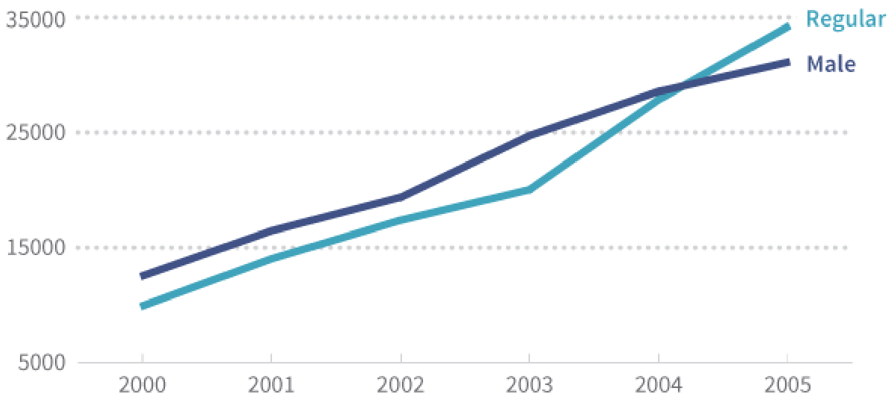
Or, returning to our coffee example, we might have several series:

Year	2000	2001	2002	2003	2004	2005
Male	12,534	16,452	19,362	24,726	28,567	31,110
Regular	9,929	14,021	17,364	20,035	27,854	34,201

With this dataset, we know several things about 2001. We know that 16,452 cups were served to men and that 14,021 cups served were regular coffee (with caffeine, cream or milk, and sugar).

We don't, however, know how to combine these in useful ways: they aren't related. We can't tell what percentage of regular coffee was sold to men or how many cups were served to women.

In other words, multiseries data are simply several series on one chart or table. We can show them together, but we can't meaningfully stack or combine them.

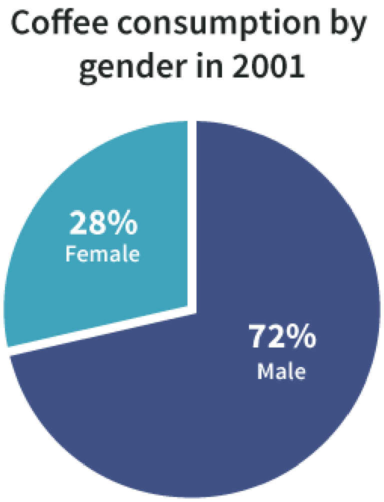


Summable Multiseries

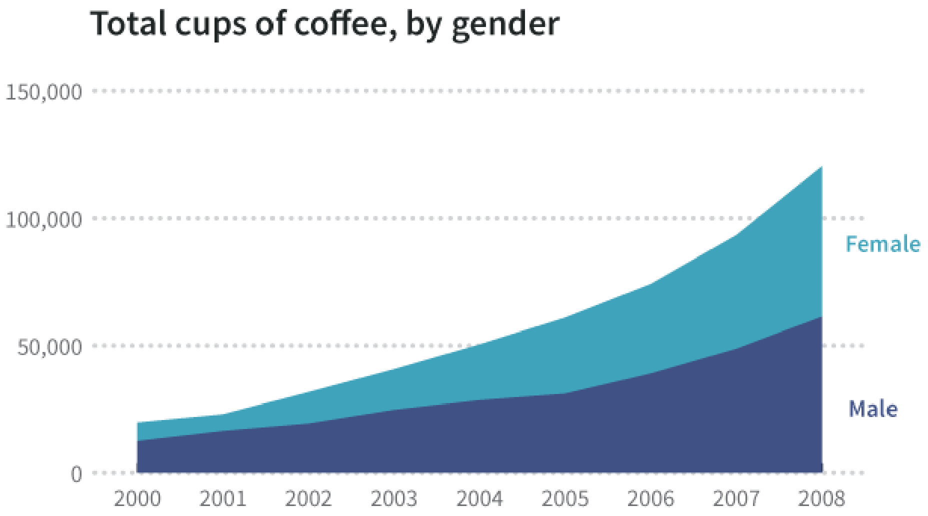
As the name suggests, a summable multiseries is a particular statistic (gender, type of coffee) segmented into subgroups.

Year	2000	2001	2002	2003	2004	2005	2006	2007	2008
Male	12534	16452	19362	24726	28567	31110	39001	48710	61291
Female	7261	6553	12349	16002	21873	29843	35142	44611	59021

Because we know a coffee drinker is either male or female, we can add these together to make broader observations about total consumption. For one thing, we can display percentages.



Additionally, we can stack segments to reveal a whole:



One challenge with summable multiseries data is knowing which series go together. Consider the following:

Year	2000	2001	2002	2003	2004
Male	12534	16452	19362	24726	28567
Female	7261	6553	12349	16002	21873
Regular	9929	14021	17364	20035	27854
Decaf	6744	6833	10201	13462	17033
Mocha	3122	2151	4146	7231	5553

There is nothing inherent in these data that tells us how we can combine information. It takes human understanding of data categories to know that *Male + Female = a complete set* and *Regular + Decaf + Mocha = a complete set*. Without this knowledge, we can't combine the data, or worse, we might combine it incorrectly.

IT'S HARD TO EXPLORE SUMMARIZED DATA

Even if we know the meaning of these data and realize they are two separate multiseries tables (one on gender and one on coffee type) we can't explore them deeply. For example, we can't find out how many women drank regular coffee in 2000.

This is a common (and important) mistake. Many people are tempted to say:

- 36.7% of cups sold in 2000 were sold to women.
- And there were 9,929 cups of regular sold in 2000.
- Therefore, 3,642.5 cups of regular were sold to women.

But this is wrong. This type of inference can only be made when you know that one category (coffee type) is evenly distributed across another (gender). The fact that the result isn't even a whole number reminds us not to do this, as nobody was served a half cup.

The only way to truly explore the data and ask new questions (such as "How many cups of regular were sold to women in 2000?") is to have the raw data. And then it's a matter of knowing how to aggregate them appropriately.

Summary Records

The following table of summary records looks like the kind of data a point-of-sale system at a café might generate. It includes a column of categorical information (gender, where there are two possible types) and subtotals for each type of coffee. It also includes the totals by the cup for those types.

Name	Gender	Regular	Decaf	Mocha	Total
Bob Smith	M	2	3	1	6
Jane Doe	F	4	0	0	4
Dale Cooper	M	1	2	4	7
Mary Brewer	F	3	1	0	4
Betty Kona	F	1	0	0	1
John Java	M	2	1	3	6
Bill Bean	M	3	1	0	4
Jake Beatnik	M	0	0	1	1
Totals	5M, 3F	16	8	9	33

This kind of table is familiar to anyone who's done basic exploration in a tool like Excel. We can do subcalculations:

- There are 5 male drinkers and 3 female drinkers
- There were 16 regulars, 8 decafs, and 9 mochas
- We sold a total of 33 cups

But more importantly, we can combine categories of data to ask more exploratory questions. For example: *Do women prefer a certain kind of coffee?* This is the kind of thing Excel, well, excels at, and it's often done using a tool called a Pivot Table.

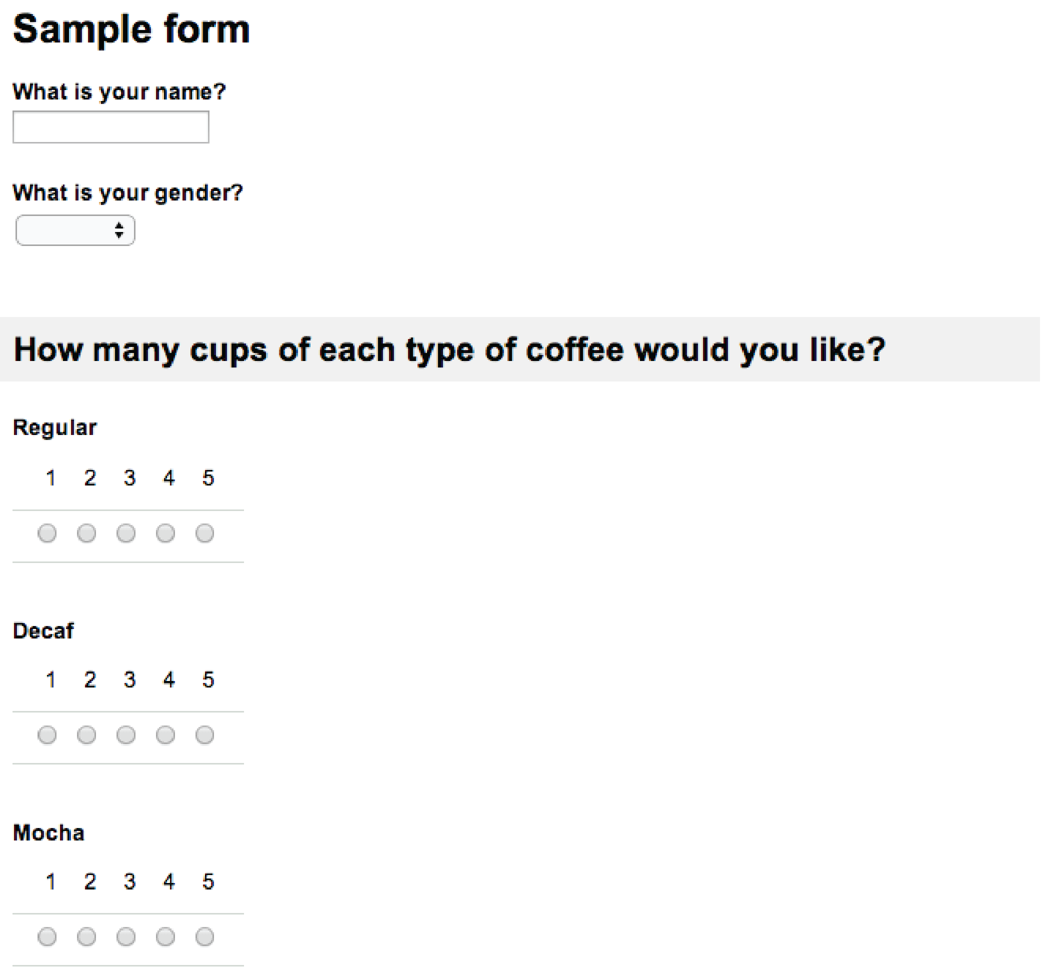
Here's a table looking at the average number of regular, decaf, and mocha cups consumed by male and female patrons:

Row Labels	Average of Regular	Average of Decaf	Average of Mocha
F	2.67	0.33	0.00
M	2.00	1.75	2.00
Grand Total	2.29	1.14	1.14

Looking at this table, we can see a pretty clear trend: Women like regular; men seem evenly split across all three types of coffee There aren't enough data to make a statistically reliable statement like this. But this is all made-up data anyway, so stop thinking so much about coffee consumption..

The thing about these data, however, is they have still been aggregated somehow. We summarized the data along several dimensions—gender and coffee type—by aggregating them by the name of the patron. While this isn't the raw data, it's close.

One good thing about this summarization is that it keeps the dataset fairly small. It also suggests ways in which the data might be explored. It is pretty common to find survey data that looks like this: for example, a Google Form might output this kind of data from a survey that says:



Sample form

What is your name?

What is your gender?

How many cups of each type of coffee would you like?

Regular

1	2	3	4	5
○	○	○	○	○

Decaf

1	2	3	4	5
○	○	○	○	○

Mocha

1	2	3	4	5
○	○	○	○	○

Producing the following data in the Google spreadsheet:

Timestamp	What is your name?	Gender?	Regular	Decaf	Mocha
1/17/2014 11:12:47	Bob Smith	Male		4	3

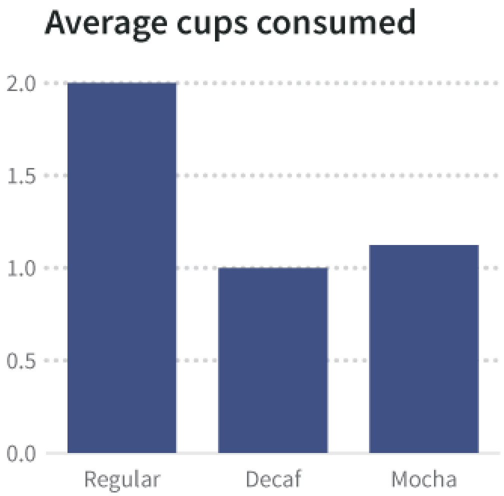
USING VISUALIZATION TO REVEAL UNDERLYING VARIANCE

When you have summary records or raw data, it's common to aggregate in order to display them easily. By showing the total coffee consumed (summing up the raw information) or the average number of cups per patron (the mean of the raw information) we make the data easier to understand.

Consider the following transactions:

Name	Regular	Decaf	Mocha
Bob Smith	2	3	1
Jane Doe	4	0	0
Dale Cooper	1	2	4
Mary Brewer	3	1	0
Betty Kona	1	0	0
John Java	2	1	3
Bill Bean	3	1	0
Jake Beatnik	0	0	1
Totals	16	8	9
Averages	2	1	1.125

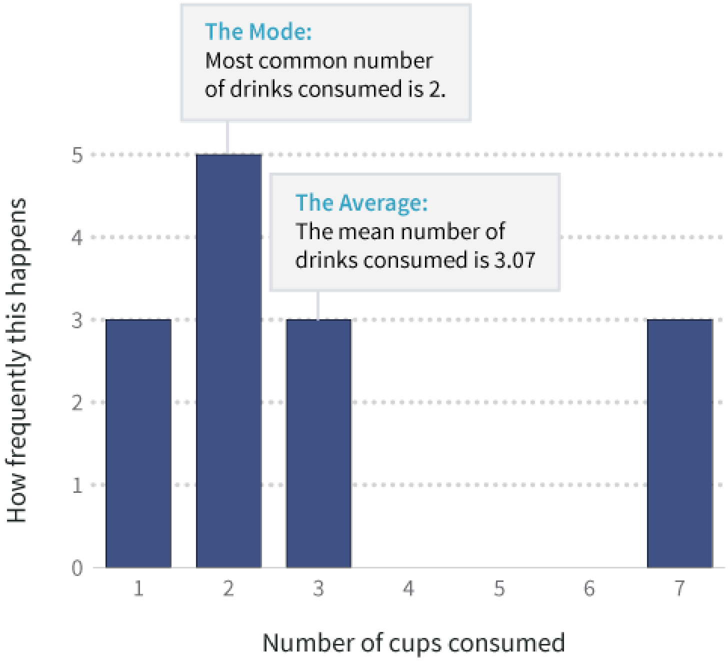
We can show the average of each coffee type consumed by cup as a summary graphic:



But averages hide things. Perhaps some people have a single cup of a particular type, and others have many. There are ways to visualize the spread, or variance, of data that indicate the underlying shape of the information, including heat charts, histograms, and scatterplots. When keeping the underlying data, you can wind up with more than one *dependent* variable for each *independent* variable.

A better visualization (such as a histogram, which counts how many people fit into each bucket or range of values that made up an average) might reveal that a few people are drinking a lot of coffee, and a large number of people are drinking a small amount.

Consider this histogram of the number of cups per patron. All we did was tally up how many people had one cup, how many had two, how many had three, and so on. Then we plotted how *frequently* each number occurred, which is why this is called a frequency histogram.



The *average* number of cups in this dataset is roughly 3. And the mode, or most common number, is 2 cups. But as the histogram shows, there are three heavy coffee drinkers who've each consumed 7 cups, pushing up the average.

In other words, when you have raw data, you can see the exceptions and outliers, and tell a more accurate story.

Even these data, verbose and informative as they are, aren't the raw information: they're still aggregated.

Aggregation happens in many ways. For example, a restaurant receipt usually aggregates orders by table. There's no way to find out what an individual person at the table had for dinner, just the food that was served and what it cost. To get to really specific exploration, however, we need data at the transaction level.

Individual Transactions

Transactional records capture things about a specific event. There's no aggregation of the data along any dimension like someone's name (though their name may be captured). It's not rolled up over time; it's instantaneous.

Timestamp	Name	Gender	Coffee
17:00	Bob Smith	M	Regular
17:01	Jane Doe	F	Regular
17:02	Dale Cooper	M	Mocha
17:03	Mary Brewer	F	Decaf
17:04	Betty Kona	F	Regular
17:05	John Java	M	Regular
17:06	Bill Bean	M	Regular
17:07	Jake Beatnik	M	Mocha
17:08	Bob Smith	M	Regular
17:09	Jane Doe	F	Regular
17:10	Dale Cooper	M	Mocha
17:11	Mary Brewer	F	Regular
17:12	John Java	M	Decaf
17:13	Bill Bean	M	Regular

These transactions can be aggregated by any column. They can be cross-referenced by those columns. The timestamps can also be aggregated into buckets (hourly, daily, or annually). Ultimately, the initial dataset we saw of coffee consumption per year results from these raw data, although summarized significantly.

Deciding How to Aggregate

When we roll up data into buckets, or transform it somehow, we take away the raw history. For example, when we turned raw transactions into annual totals:

- We anonymized the data by removing the names of patrons when we aggregated it.
- We bucketed timestamps, summarizing by year.

Either of these pieces of data could have shown us that someone was a heavy coffee drinker (based on total coffee consumed by one person, or based on the rate of consumption from timestamps). While we might not think about the implications of our data on coffee consumption, what if the data pertained instead to alcohol consumption? Would we have a moral obligation to warn someone if we saw that a particular person habitually drank a lot of *alcohol*? What if this person killed someone while driving drunk? Are data about alcohol consumption subject to legal discovery in a way that data about coffee consumption needn't be? Are we *allowed* to aggregate some kinds of data but not others?

> *Can we address the inherent biases that result from choosing how we aggregate data before presenting it?*
>
> The big data movement is going to address some of this. Once, it was too computationally intensive to store all the raw transactions. We had to decide how to aggregate things at the moment of collection, and throw out the raw information. But advances in storage efficiency, parallel processing, and cloud computing are making on-the-fly aggregation of massive datasets a reality, which should overcome some amount of aggregation bias.

COLLECTING DATA

Gathering data is similar to gathering ingredients for a recipe. If you want to create an excellent dish, you'll want to start with the right ingredients, which means you'll have to make a bunch of decisions beforehand.

For example, if you need honey, do you want a generic honey or a specific variety like orange blossom? Does the brand of honey matter? Does it need to be raw or organic? Do you prefer to get the honey from your local farmer or a supermarket? And who's going to get all these ingredients? If you don't have time, are you willing to pay someone to get them for you, even if it means you might not get exactly what you want?

You have to make similar decisions before gathering data. You could get all of your data by asking open-ended questions on paper surveys that you pass out to people in a random place, but that probably won't make the most sense. Different types of data are best collected in different sorts of ways from different places, and sometimes you might not have enough time or money to collect the data you need by yourself.

In this section, we'll talk about data collection methods to help you figure out where and how to best obtain the information you're looking for, and how to sort out what information you really need. Let's go shopping!

CHAPTER 3

INTRO TO SURVEY DESIGN

BY GINETTE LAW

Most people think conducting a survey is as simple as writing a bunch of questions and asking people to answer them. Easy, right? Well, sort of: it's a little more complicated than that if you want to collect the best data possible. Don't worry, though: we're about to go over some key elements of survey design, starting with the purpose of your survey.

Purpose of a Survey

The first step in designing a good survey is to identify its purpose *before* you create it. A good survey collects accurate and verifiable data that allow you to make concrete claims. This is easier to do if you have a clear purpose to guide you when deciding what information you want to collect your respondents.

A survey generally helps you do one or more of the following:

- describe something;
- describe how things are related;
- explain a relationship; or
- influence something

So how do you come up with a nice, well-defined purpose for your survey? Try asking yourself the following questions:

- What are you trying to achieve with your survey?
- What, precisely, do you want to know?
- Why is this important to know?
- Is there other information that could be useful? Why?
- Is conducting a survey the right method for the type of data you're collecting?

To see this process in action, let's say you're a media tycoon trying to increase your profits from television subscribers, so you're interested in the potential returns from expanding to an internet market. You decide the main question and purpose of your survey are:

RESEARCH QUESTION	PURPOSE
What percentage of television viewers watch their favorite television shows online?	**Describe a variable** The percentage of people who watch television programs online.

After thinking about it, you realize that it would be helpful to get more details, so you ask yourself some of those questions we just suggested:

QUESTION	ANSWER
What are you trying to achieve with your survey?	To evaluate the profit potential of the internet market.
What do you want to know precisely?	How many people currently or would like to be able to stream shows online.
Why is this important to know?	It will help determine where we can improve our online service and whether it's worth the investment.
Is there other information that could be useful? Why?	Which age group is more likely to watch TV online.
Is conducting a survey the right method for the type of data you're collecting?	Yes.

Based on your answers to these questions, you expand the scope of your survey slightly:

RESEARCH QUESTION	PURPOSE
How can we maximize profits from online viewership?	**Influence something:** Find out what needs to be done to improve our online service.
Do younger television users tend to stream more television programs online than older viewers?	**Describe how things are related:** Describe how a viewer's age is related to how much TV he or she watches online.
If one group watches more programs online, why?	**Explain a relationship:** Explain why one age group prefers watching TV online more than another age group does.

Now you have a nice set of clearly-defined questions that your survey should address. The next step is to choose the best type of survey to help you expand your business.

Data and numbers can be very powerful, but sometimes people conduct surveys for the wrong reasons. You will miss the opportunity to gain meaningful insight with your survey if:

- You don't really care about the results; you just want to show people you've got numbers.
- You misuse the data collected.
- You are more concerned by how people will receive your findings than you are with having reliable results.
- You have already determined what the results "should" be.

Remember, we do research to gain insight and to test hypotheses. That means it's important to try and collect the most accurate and representative data possible and not just the data that support your own preconceived ideas and biases.

Types of Surveys

Now that you've identified the purpose of your survey, you can use this to help you choose a type of survey. When it comes to choosing a survey type, the first big decision you'll have to make is *how* you want to distribute your survey to respondents. Surveys can be self-administered or administered.

Self-Administered and Administered Surveys

Self-administered simply means that respondents fill out a questionnaire by themselves, whereas administered means that an interviewer asks the questions. Each of these methods has advantages and disadvantages, summarized below.

Self-Administered Surveys

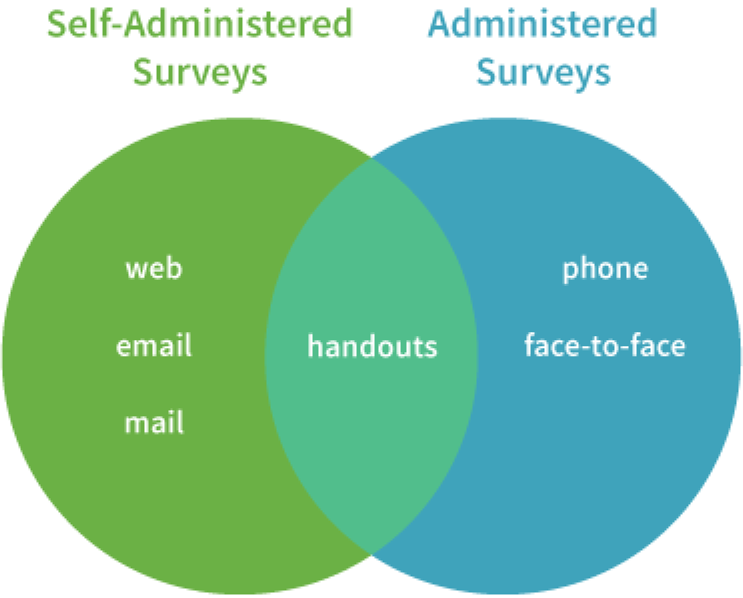
PROS	CONS
• Good for limited budgets • Large geographic reach • Respondents can answer when convenient and at their own pace	• More difficult to distribute to populations with limited literacy • Risk of respondents misinterpreting questions • Risk of lower completion rate • Requires extra effort to verify that the respondent was actually the person to whom the survey was addressed

Administered Surveys

PROS	CONS
• Tighter quality control • Interviewers can clarify questions for respondents • Can collect richer and more in-depth information • Easier to reach specific or marginalized populations (e.g., the elderly or the homeless)	• Can be expensive • Can be time consuming • Risk of interviewer effect (e.g., an interviewer's interpretation of a question could bias some of the results)

The table above mentions the interviewer effect. This is a phenomenon whereby the interviewers themselves influence the responses. This could happen because of the way they ask questions or phrase explanations, or because of another social factor altogether. For example, maybe the interviewer is incredibly good-looking and the respondent unintentionally gives answers that are biased towards impressing the interviewer!

Self-administered and administered surveys are each associated with certain types of surveys, as shown below:



Types of Self-Administered Surveys

WEB AND EMAIL SURVEYS

One of the easiest and quickest ways you can conduct a survey is online. You can either have people fill out questionnaires directly on a website or send them a questionnaire via e-mail. There are even survey sites that let you design questionnaire forms, collect data, and analyze responses in real-time at an affordable rate. You'll find links to a few of these sites in the resources section.

Don't forget to read the Privacy Policies and Terms of Use before signing up though—especially if you're dealing with sensitive data—because these services are often subject to the laws of the countries in which they're based.

The downside to all of this convenience is that only regular internet or computer users are likely to fill out your survey.

WEB AND EMAIL

PROS	CONS
• Fairly fast results • Integrated support tools for images, sounds, video • Relatively cheap	• Limited to people who have internet access • Respondents can be a self-selected group of regular computer-users

WEB

PROS	CONS
• Can invite participants through social media • Can use survey sites	• Survey sites are subject to the laws of the countries in which they're based

EMAIL

PROS	CONS
• Easier to survey a specific group	• Need email addresses in order to distribute survey

MAIL SURVEYS

Mail surveys are much like email surveys except that you're sending a paper version of the questionnaire by mail. They aren't limited to the population of internet users, but they can be more expensive because you have to pay for printing, paper, envelopes, and postage. You may also end up with a lower response rate because some people think it's a pain to mail anything, not just your survey. Including a pre-addressed, stamped envelope with the survey will often improve your overall response rate, but will, again, increase the overall cost to you.

MAIL

PROS	CONS
• Open to a wider population	• More expensive • Lower response rate

Types of Administered Surveys

PHONE SURVEYS

Compared to online surveys, phone surveys may take less time for you to collect all the data you need because respondents answer questions immediately. However, if you need to complete a certain number of surveys and most people hang up on you, it might longer to collect data by phone!

It helps if questions are short and clear so that respondents can easily understand them over the phone. Try also to not take more than 15 to 20 minutes of your respondents' time (both to be respectful of their time and to avoid losing their attention).

Another thing to keep in mind is that the interviewer and the respondent can't see each other. This can be both good and bad: good because it lessens the interviewer effect but bad because both parties may miss non-verbal cues.

PROS	CONS
• Relatively fast results • Higher quality control because trained interviewers administer the survey	• Can be more expensive • Can't use images or video when asking questions • Need a list of reliable phone numbers

FACE-TO-FACE SURVEYS

Unlike phone surveys, face-to-face surveys allow the interviewer and the respondent to see each other's facial expressions and body language. This can be helpful because the additional physical cues can help the interviewer and the respondent understand better understand each other; however, it can also lead to the respondent being further influenced by the interviewer's behavior and appearance.

Face-to-face surveys are limited geographically if interviewers and respondents have to meet in person. If they don't have to meet in person, the surveys can be conducted online using video conferencing software such as Skype (http://www.skype.com/) or Google Hangouts (https://www.google.com/hangouts/).

👍 PROS	👎 CONS
• Can collect more in-depth information • Good for specific or hard-to-reach populations • Better for long surveys	• Can be more expensive • May have a stronger interviewer effect • Can take longer to train interviewers and to complete enough surveys

Handouts

So far, we've talked about self-administered and administered surveys, but one of the most frequently-encountered type of survey is actually a combination of these. Say you want to hand out paper surveys and have people complete and return them immediately. The survey itself is self-administered, but since you have a trained person there who is available to answer questions, it also has features of an administered survey.

For this reason, handouts can be a good option if you have a limited budget but want to make sure you have someone available when the survey is completed to clarify any questions your respondents might have.

One of the disadvantages of handouts is that people may be rushed to complete the survey if you are catching them in passing, which can affect the quality of your data. You will also be limited to the population that is physically present in the location where you are giving the survey. This may not be an issue if you are targeting a specific group, such as college students, shoppers at a particular store, or residents of a certain region. If you're looking for a more general audience, however, you may consider handing the survey out in several different locations to reach a more diverse audience.

Choosing a Survey Type

Whew! That is a lot of information so far. As you can see, there are many types of surveys, each with its own pros and cons. With so many factors to consider, how do you decide which one to choose? Let's walk through a few scenarios of choosing a survey type. Remember, you're a media tycoon trying to expand your media empire.

Let's say you want to evaluate the potential of the internet market. If you're only interested in the online viewing habits of internet users, then a web or email survey cheaply and conveniently targets the group you're interested in. On the other hand, if you want to know how much of the general television viewing audience watches shows online, but you don't want to spend the time or the money to train interviewers, then a mail survey is a better option.

But wait a minute: you're a media tycoon! You have lots of money *and* you're a total data geek. You want to collect more in-depth information like:

- How do families watch television programs in their homes?
- Which family members tend to watch TV online?
- What type of television program does each family member like to watch?
- Do they watch programs together?
- If so, what types of programs do they watch together? Do they ever fight about what to watch?

In this case, a face-to-face survey is a great option because you can interview all the members of a household at the same time and collect higher quality data about the dynamics of television viewing in families.

Now let's say you've analyzed the data and the numbers look promising, so you decide to create a new online viewing experience for your users. You want to emphasize how great and cutting-edge your new service is by creating a catchy name and logo for it. If you want to see what people think about different logos, then phone surveys won't be helpful. On the other hand, if you want to see what people think about different names and slogans, then phone surveys are fine. If your new service is specifically targeting a younger audience of avid internet users, then you can avoid the extra cost of phone surveys and use an online survey instead.

If you have the time and resources to do so, you can also consider using more than one type of survey. For example, you may want to do an online and phone version of the same survey to increase response rates and improve your sample in both size and diversity. Or you may start by doing an online survey and then conduct a face-to-face survey afterwards to gain more insight from your initial results.

We hope these scenarios show you that choosing a survey type isn't a one-size-fits-all process. There isn't a nice and tidy formula that you can use to pick the perfect survey. Instead, it's up to your good judgment to balance the resources you have with the goals you're trying to achieve.

CHAPTER 4

TYPES OF SURVEY QUESTIONS

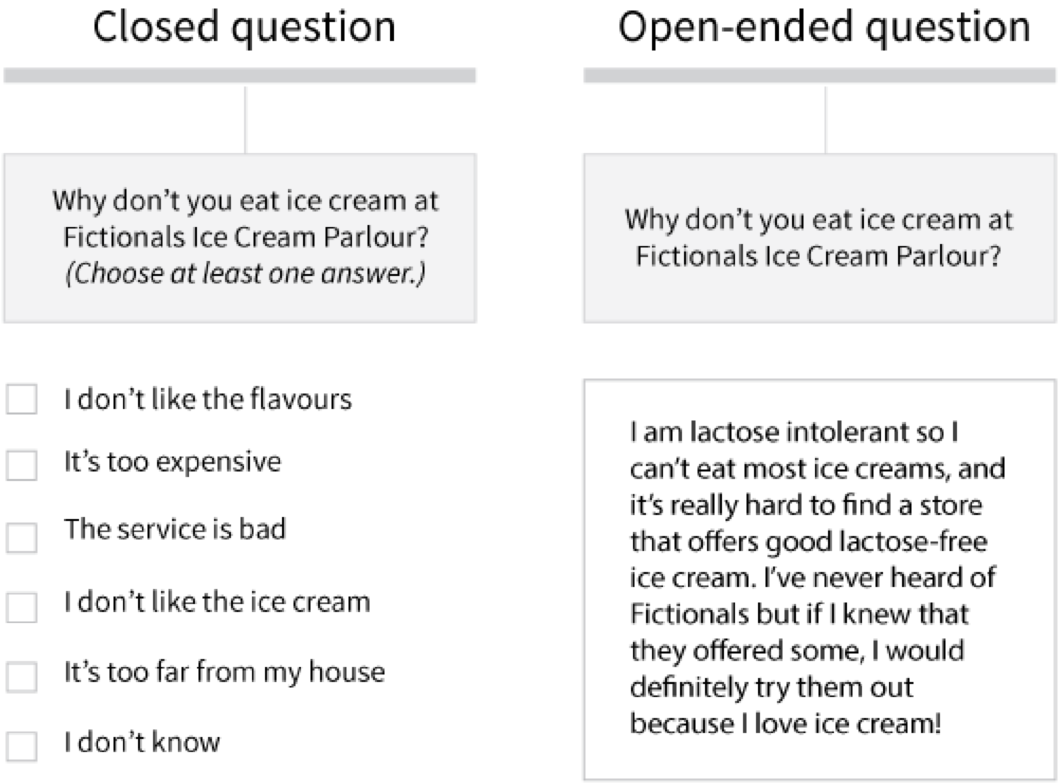
BY GINETTE LAW

After you've decided what type of survey you're going to use, you need to figure out what kinds of questions to include. The type of question determines what level of data can be collected, which in turn affects what you can do with the data later on. For example, if you ask about age and you record it as a number (e.g., 21, 34, 42), you'll have numeric data that you can perform mathematical operations on. If you instead ask for a person's age group (e.g., 18-24, 25-34, 35-44), you'll have ordered categorical data: you'll be able to count how many people were in each group, but you won't be able to find the mean age of your respondents. Before you create your survey, it's important to consider how you want to analyze your final results so you can pick question types that will give you the right kind of data for those analyses.

Open vs. Closed Questions

Open (or open-ended) questions are used when you want to collect long-form qualitative data. You ask the person a question without offering any specific answers from which to choose. This is often a good option to discover new ideas you may have overlooked or did not know existed. In a closed question, however, a limited number of specific responses are offered to answer the question. Closed questions are used when you already have an idea what categories your answers will fall into or you're only interested in the frequency of particular answers.

Let's pretend you work for an ice cream parlour called Fictionals. You want to know why some people don't come to your store. Here are open and closed questions you could ask to find out their reasons.



Closed question

Why don't you eat ice cream at Fictionals Ice Cream Parlour? *(Choose at least one answer.)*

□ I don't like the flavours

□ It's too expensive

□ The service is bad

□ I don't like the ice cream

□ It's too far from my house

□ I don't know

Open-ended question

Why don't you eat ice cream at Fictionals Ice Cream Parlour?

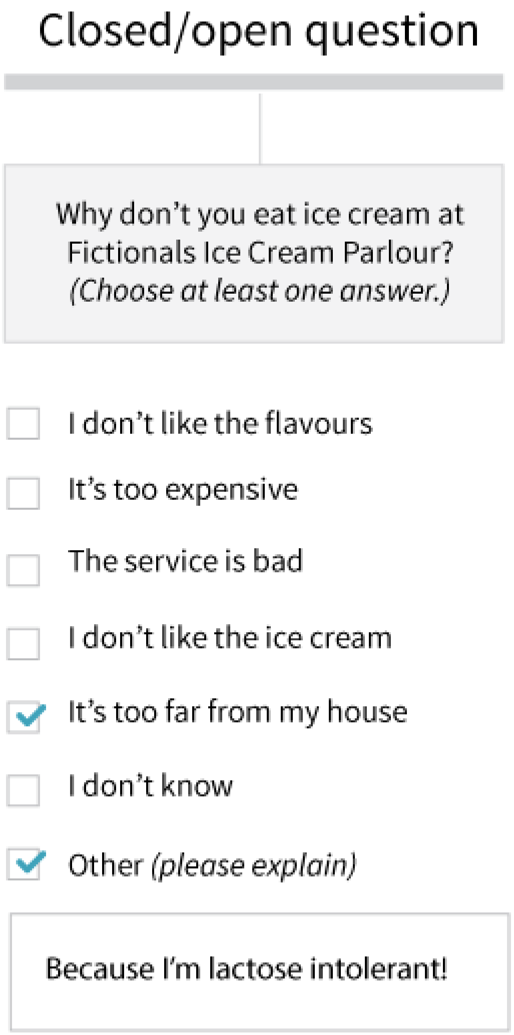
I am lactose intolerant so I can't eat most ice creams, and it's really hard to find a store that offers good lactose-free ice cream. I've never heard of Fictionals but if I knew that they offered some, I would definitely try them out because I love ice cream!

In the closed question, you offer the most probable reasons why people would not go eat ice cream at your store, whereas in the open question, you just ask why they don't eat at Fictionals and allow respondents to answer however they want.

When you ask an open-ended question you gain new insight to why some people might not try your ice cream: it's not because they don't like ice cream or they don't like Fictionals, it's because they can't eat it! With this new insight, you could choose to offer a new type of ice cream and attract a clientele you didn't appeal to before.

This is a distinct advantage of open-ended questions, and it's particularly useful if you're doing exploratory research and will primarily be reading through the responses to get a sense of how people responded. However, it takes considerably more work to analyze and summarize these data.

When you think that most people will give more or less the same answer, you can combine an open and closed question to get the best results. This is usually done by offering a number of predefined choices, but leaving room for participants to explain their response if it falls under the "Other" category. Here is how the same question would look if you transformed it into a combined open and closed question:



Closed/open question

Why don't you eat ice cream at Fictionals Ice Cream Parlour? *(Choose at least one answer.)*

☐ I don't like the flavours

☐ It's too expensive

☐ The service is bad

☐ I don't like the ice cream

☑ It's too far from my house

☐ I don't know

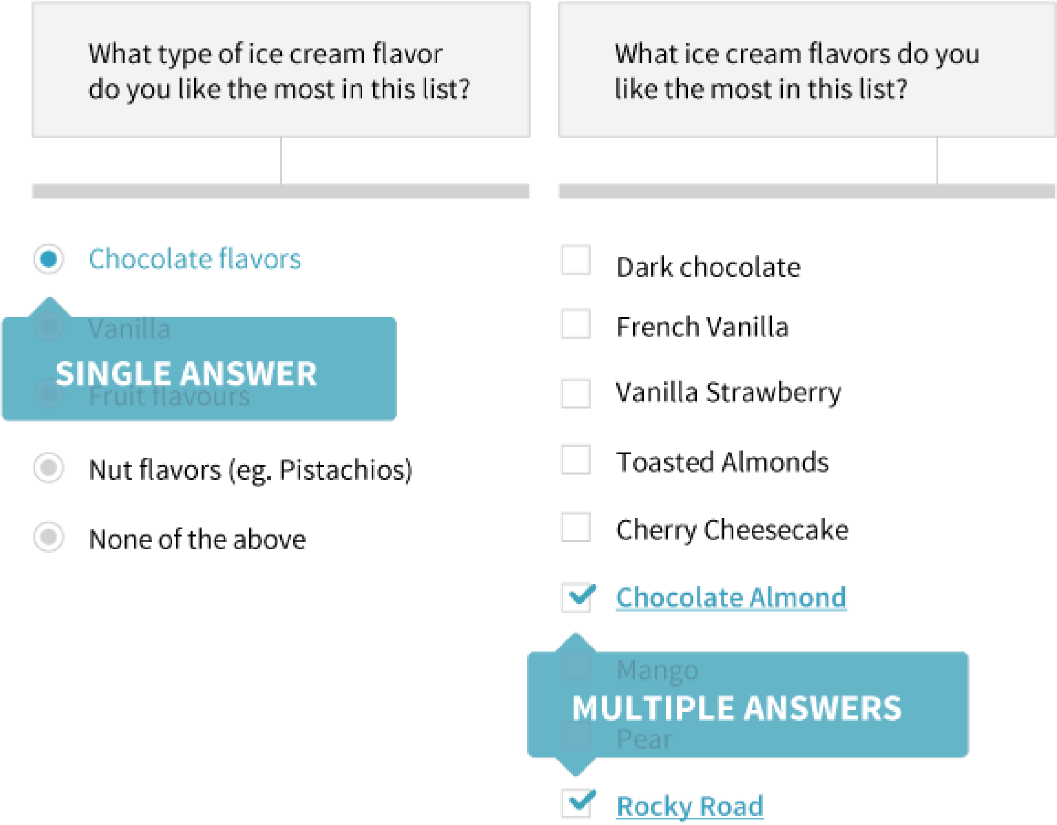
☑ Other *(please explain)*

Because I'm lactose intolerant!

Closed Question Types

MULTIPLE CHOICE QUESTIONS

Sometimes you'll want the respondent to select only one of several answers. If your survey is online, you use radio buttons (or circles), which indicate to choose just one item. If you want to allow the respondent to choose more than one option, you should use a checkbox instead. Here is what it looks like in practice:



What type of ice cream flavor do you like the most in this list?

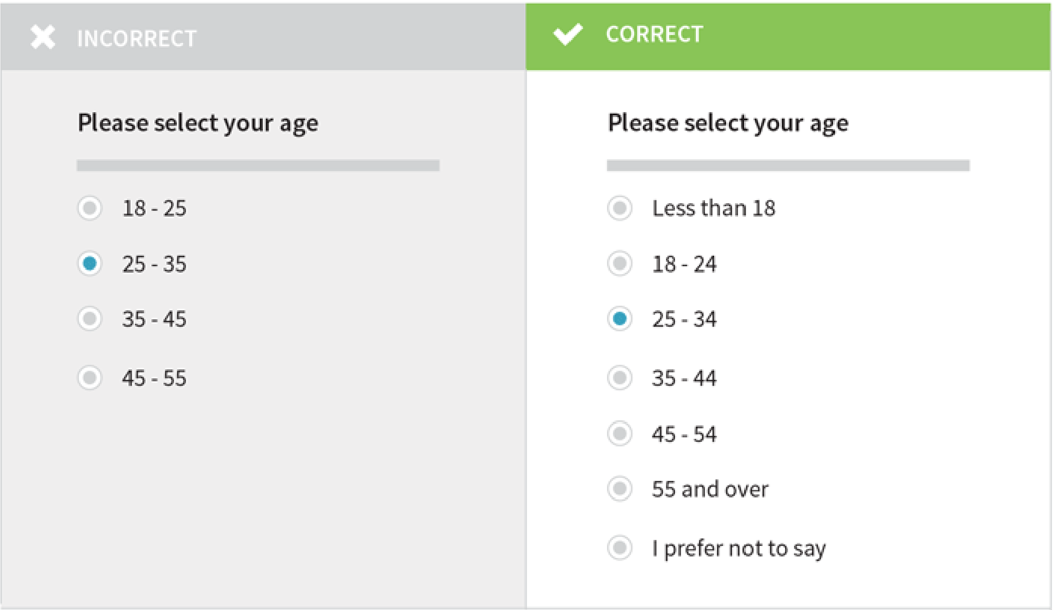
- ● Chocolate flavors
- ○ Vanilla
- ○ Fruit flavours
- ○ Nut flavors (eg. Pistachios)
- ○ None of the above

What ice cream flavors do you like the most in this list?

- □ Dark chocolate
- □ French Vanilla
- □ Vanilla Strawberry
- □ Toasted Almonds
- □ Cherry Cheesecake
- ☑ Chocolate Almond
- □ Mango
- □ Pear
- ☑ Rocky Road

If your survey is a paper or administered survey, you would use printed or verbal directions to indicate how many options the respondent should pick (e.g., "Please choose one"; "Please select up to 3"; "Please select all that apply").

Answer choices should not overlap and should cover all possible options. This is called having “mutually exclusive and exhaustive” categories and is crucial for sound survey design. Consider the situation below:



✕ INCORRECT

Please select your age

○ 18 - 25
● 25 - 35
○ 35 - 45
○ 45 - 55

✓ CORRECT

Please select your age

○ Less than 18
○ 18 - 24
● 25 - 34
○ 35 - 44
○ 45 - 54
○ 55 and over
○ I prefer not to say

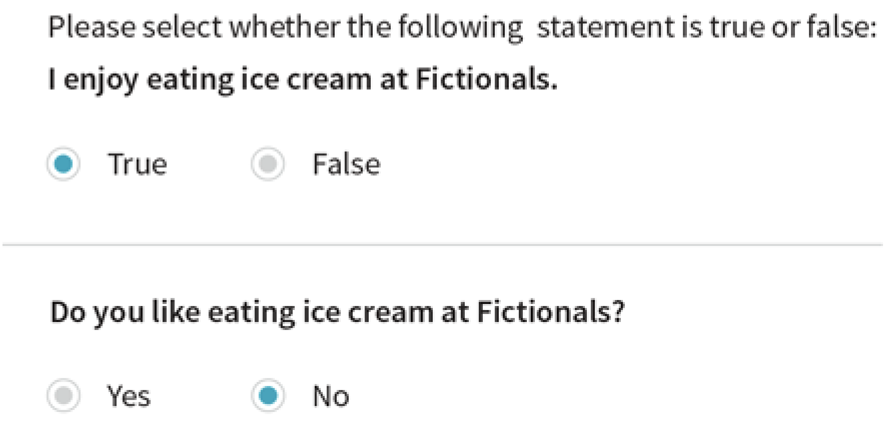
In the left set of choices, someone who is 25 wouldn’t know whether to select the 18-25 button or the 25-35 button, and someone who is 60 wouldn’t have any option to select at all! The set of age categories on the right is better because it includes all possible age groups, the age choices don’t overlap, and it allows people the choice to withhold that information if they choose to.

It is advisable to include a “Prefer not to answer” option for questions that may be of a personal nature, such as race, ethnicity, income, political affiliation, etc. Without this option, people may skip the question altogether, but you won’t be able to tell later on who skipped the question on purpose and who missed it accidentally, which can be an important difference when you analyze your data.

It won’t always be practical to list every possible option. In that case, including “None of the above” or “Other” choices may be helpful. This makes it less likely that people will just skip the question and leave you with missing data. You may also consider including a “Not Applicable” (i.e., “N/A”) choice for questions that may not apply to all of your respondents, or a “Don’t know” choice for questions where a respondent may not not have the information available.

DICHOTOMOUS QUESTIONS

Dichotomous questions are a specific type of multiple choice question used when there are only two possible answers to a question. Here are some classic examples of dichotomous type questions:



Please select whether the following statement is true or false:

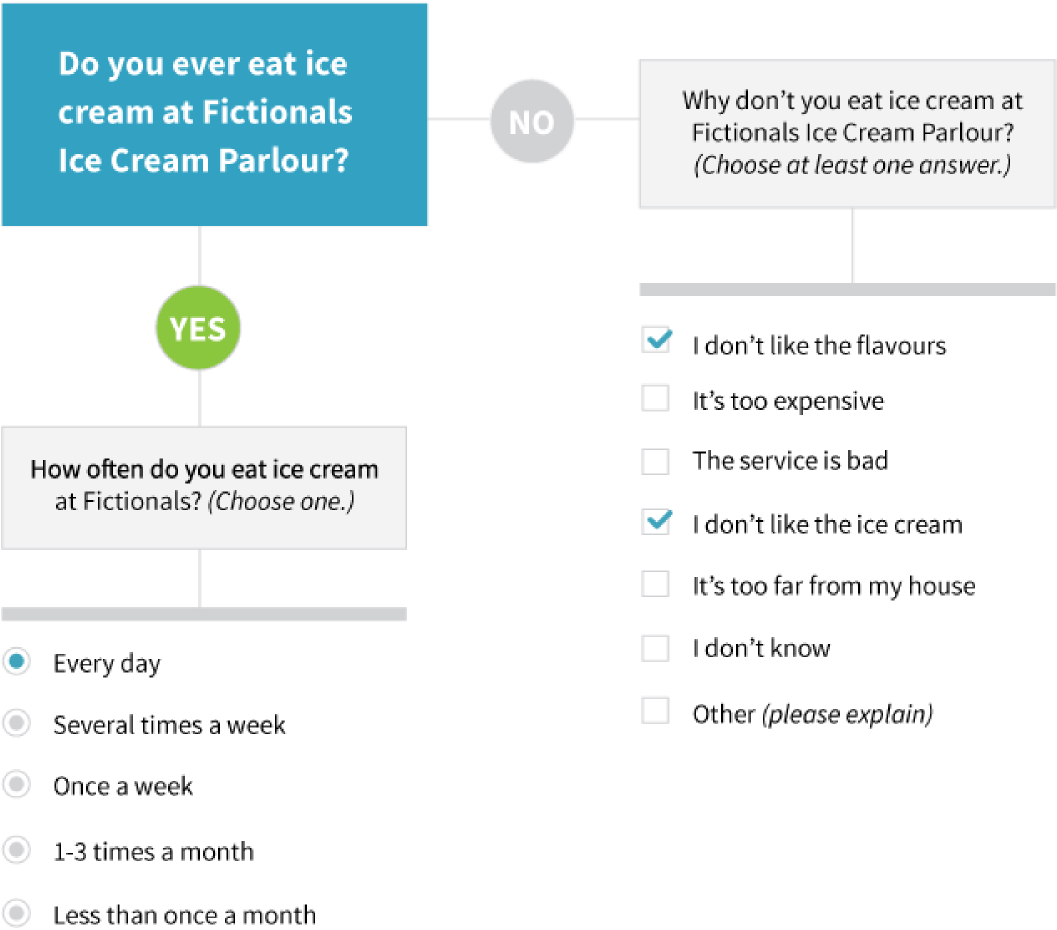
I enjoy eating ice cream at Fictionals.

● True ○ False

Do you like eating ice cream at Fictionals?

○ Yes ● No

You can use dichotomous questions to determine if a respondent is suited to respond to the next question. This is called a filter or contingency question. Filters can quickly become complex, so try not to use more than two or three filter questions in a row. Here's an example of the logic behind a filter question:



Do you ever eat ice cream at Fictionals Ice Cream Parlour?

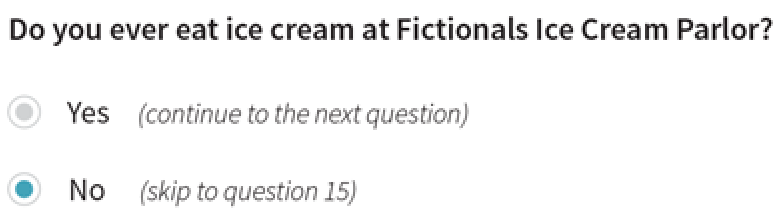
Why don't you eat ice cream at Fictionals Ice Cream Parlour? *(Choose at least one answer.)*

☑ I don't like the flavours
□ It's too expensive
□ The service is bad
☑ I don't like the ice cream
□ It's too far from my house
□ I don't know
□ Other *(please explain)*

How often do you eat ice cream at Fictionals? *(Choose one.)*

● Every day
○ Several times a week
○ Once a week
○ 1-3 times a month
○ Less than once a month

In an online survey or administered survey, answering "Yes" to the question "Do you ever eat ice cream at Fictionals Ice Cream Parlour?" would take a respondent to the question about how often they do, while selecting "No" would open the checkbox response question asking why they don't eat there. In a paper survey, all questions would need to be listed on the page and couldn't be presented conditionally based on a specific response. You could use a flowchart form like in the diagram above, or you could employ skip directions, as illustrated below.



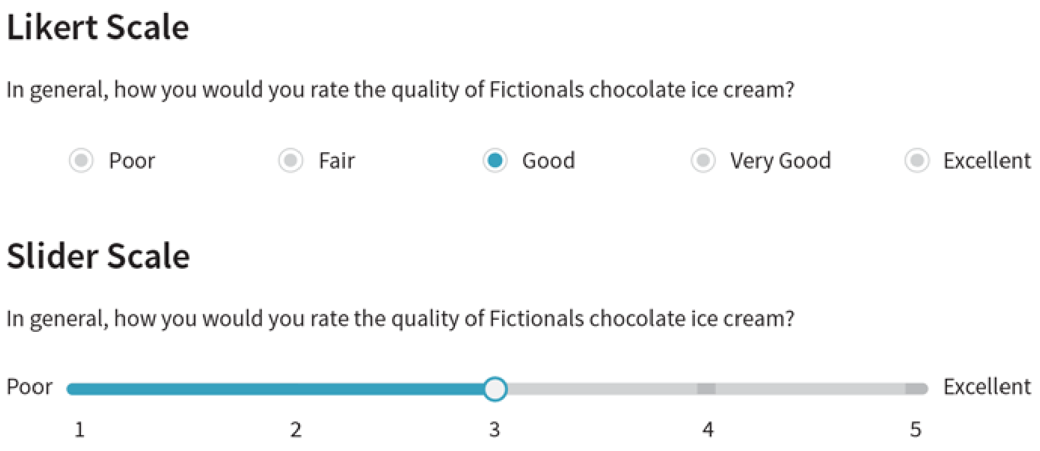
Do you ever eat ice cream at Fictionals Ice Cream Parlor?

○ Yes *(continue to the next question)*

● No *(skip to question 15)*

SCALED QUESTIONS

Scaled questions are used to measure people's attitudes, opinions, lifestyles and environments. There are many different types but the most commonly used are Likert and slider scales. The main difference between the two is that a Likert scale consists of ordered categories whereas a slider scale allows respondents to make a mark indicating their response anywhere along a scale. As a result, a Likert scale-type question will give you data that are measured at an ordinal level while a slider scale will give you data that are measured at an interval level.



Likert Scale

In general, how you would you rate the quality of Fictionals chocolate ice cream?

○ Poor ○ Fair ● Good ○ Very Good ○ Excellent

Slider Scale

In general, how you would you rate the quality of Fictionals chocolate ice cream?

In the example above, the same question is asked twice, once with a Likert scale and once with a sliding scale. Notice that in the Likert scale there are five categories. Scales with an odd number of categories allow participants to agree, disagree, or indicate their neutrality; scales with an even number of categories only allow participants to agree or disagree. If you want to allow for neutrality, scales with an odd number of categories are usually preferred; however, you should use an even number if you want to require the respondent to choose one direction or the other. This is called a “forced question.”

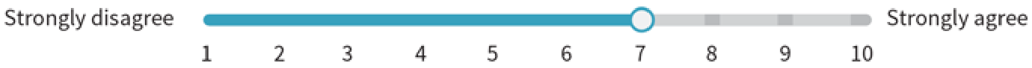
LIKERT SCALE

Likert scales are usually limited to a five-category scale, as it is often difficult to clearly display a larger number of categories. The most common Likert scales you will find are:

- *Opinions and attitudes:* “How much do you agree with…?”
  Possible answers: strongly agree, agree, neither agree or disagree, disagree, strongly disagree
- *Frequency:* “How often do you…?”
  Possible answers: always, often, sometimes, rarely, never
- *Quality:* “In general, how do you rate the quality of…?”
  Possible answers: excellent, very good, good, fair, poor
- *Importance:* “How important would you say {topic} is…”
  Possible answers: very important, important, somewhat important, not at all important

SLIDER SCALE

Slider scales are useful when you want to have a more precise reading of a respondent’s views. These scales are most easily implemented in an online survey, since the computer will calculate the position of the marker; when used on a paper survey, you will need to have someone measure the marker’s location on the scale manually. Slider scales may be more practical than Likert scales when conducting a survey that is being translated into multiple languages since text categories are not always easy to translate.



> When creating a scale, whether Likert or slider, never forget to label the first and last points in your scale. Otherwise, your respondents may misinterpret your scale and give you inaccurate or misleading data.

Question Wording

It's extremely important to think about how you word your questions when developing a survey. Here are a few key things to keep in mind:

FOCUS

Each question should be specific and have a defined focus.

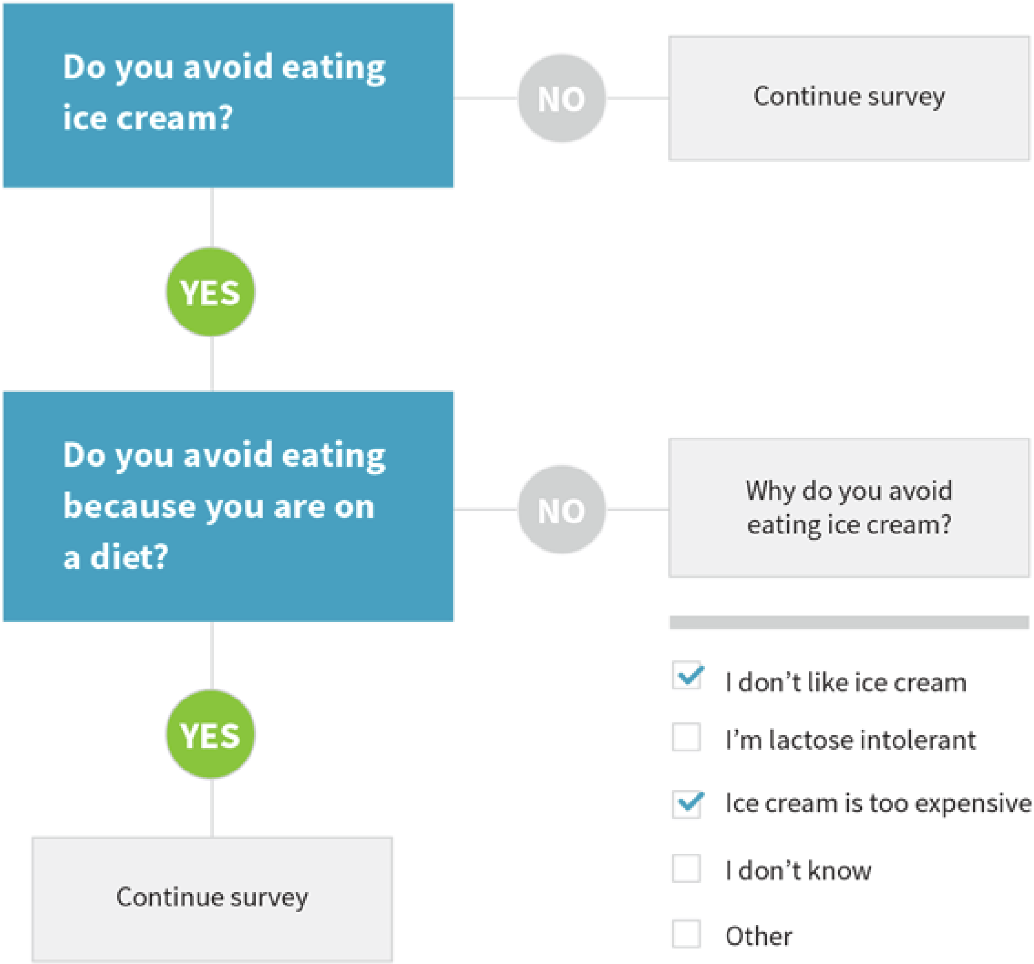
> How many times have you eaten ice cream in the last month?

This question focuses on one issue (frequency of ice cream eating) and is specific (a certain time period).

> Do you avoid eating ice cream because you are on a diet? [Answer: Yes, I avoid ice cream, but not because I am on a diet]

This question is poorly worded for a few reasons: it is both leading and *too* specific. It assumes that the respondent is in fact on a diet and that this is why he or she doesn't eat ice cream. A better wording would be "Why do you avoid eating ice cream?" with "I'm dieting" as one of a number of options that the participant can choose from.

You should also avoid using the word "and" if it is connecting two different ideas within a single question. Remember, each question should focus on only one issue at a time: otherwise, you won't be collecting the best data that you can. By compounding multiple thoughts into a single question, you reduce the accuracy of participants' responses and thereby limit the claims that you can make from those data. Instead, consider using filter questions to obtain the desired information. For example:


Do you avoid eating ice cream?
NO
Continue survey
YES
Do you avoid eating because you are on a diet?
NO
Why do you avoid eating ice cream?
☑ I don't like ice cream
□ I'm lactose intolerant
☑ Ice cream is too expensive
□ I don't know
□ Other
YES
Continue survey

PRECISION

Not everyone will interpret all words and phrasings in the same way, even if the meaning seems obvious to you. To reduce the chance of participants misinterpreting your question, you can parenthetically define ambiguous terms or give additional context as appropriate. Try also to avoid confusing wordings, such as those that use double negatives or many subordinate clauses. Be careful to also not use words that are loaded or have many highly-emotional connotations.

✓ DO	✗ DON'T
• Have you eaten chocolate ice cream in the past month? • Ice cream with peanuts should not be permitted. • Ice cream vendors should not disclaim having a peanut-free shop.	• Rocky Road, Mint Chocolate Chip, Dutch Chocolate and Hot Fudge ice creams are Fictionals' best sellers. Have you eaten any chocolate ice cream in the last month? • Ice cream with peanuts should be banned. • Ice cream vendors should accept to have a peanut-free shop.

BREVITY

Your questions should be relatively short, except where additional wording is absolutely necessary to provide context or to clarify terminology. Long, complex questions can quickly become confusing to participants and increase the chances that they will respond without fully understanding what is being asked, or worse, skip the question entirely! If your question seems to be getting too long, consider whether there is any unnecessary information that you can take out or if you can split it up into several smaller questions. Alternately, you may want to have a short paragraph of explanatory text that is separate from the question itself: that way participants will have all necessary background information, but can also easily pick out and digest what is actually being asked.

✓ DO	✗ DON'T
Please list your three favorite ice creams in order of preference.	Can you please tell us what ice cream flavors you like and what are your first, second, and third favorites?

BIASED AND LEADING QUESTIONS

Biased or leading questions can easily skew your answers if you do not pay close attention to your wordings. Avoid over-emphasizing your statements and keep an eye out for questions that create "social desirability effects" where respondents may be more likely to answer according to what society views as proper or socially or morally acceptable.

✓ DO	✗ DON'T
Do you agree that ice cream vendors in our city should offer peanut-free products?	Do you agree that ice cream vendors that serve peanuts are a severe hazard for the well-being of our children?

Notice how the second question tries to bias the respondent by using strong or affective phrases such as "severe hazard." You should try to keep your questions as value-free as you can: if the question itself suggests how the participant is expected to answer, it should be reworded.

A FEW FINAL THINGS…

There are few final things to consider when developing your survey questions:

- If possible, consider varying the type of questions you use to keep your respondents engaged throughout the survey. You have a variety of question types to choose from, so mix it up where you can! (But do keep in mind that each type of question has a specific purpose and gives you more or less specific data.)
- Think about how certain words can have different interpretations and remember that meanings are often embedded in culture and language. Be sensitive to different cultures if you are examining several groups within a population.

Designing Your Questionnaire

Once you have identified the purpose of your survey and chosen which type you will use, the next step is to design the questionnaire itself. In this section, we'll look at the structure, layout and ordering of questions. We'll also talk about some key things to keep in mind once you're done.

QUESTIONNAIRE STRUCTURE

Do you remember when you were taught in grade school that every essay should have an introduction, body and conclusion? Questionnaires should also have a certain structure. The major parts of a questionnaire include the introduction, main topic, transitions between topics, demographic questions, and conclusion.

INTRODUCTION

Always start your survey with a brief introduction which explains:

- the purpose of the survey;
- who is conducting it;
- the voluntary nature of the participant's involvement;
- the respect for confidentiality; and
- the time required to complete the survey.

MAIN TOPICS

Generally, it is best to start with general questions and then move on to more specific questions.

- Ask about objective facts before asking more subjective questions.
- Order questions from the most familiar to least.
- Make sure that the answer to one question does not impact how the participant interprets the following question.

TRANSITIONS BETWEEN TOPICS

Use transitions between groups of questions to explain a new topic or format. For example: "The next few questions are related to the frequency of your TV viewing habits. Please choose the answer that best describes your situation."

DEMOGRAPHIC QUESTIONS

Unless you are using demographic questions as filtering criteria for survey eligibility, it is usually better to put them near the end of a survey. Typical demographic questions include:

- gender;
- age;
- income;
- nationality or geographic location;
- education; and
- race or ethnicity.

CONCLUSION

Thank the participants for their contribution and explain how it has been valuable to the project. Reiterate that their identities will be kept confidential and that results will be anonymized. If you wish, you can also include your contact information in case they have any additional questions related to the survey and also ask for their contact information if you are offering an incentive for completing the survey.

If you contact respondents in advance to invite them to participate in the survey, you should also explain why they have been selected to participate, when the survey will take place, and how they can access it.

GENERAL LAYOUT

There are a few good rules to follow when designing the layout of your survey:

- Put your introduction and conclusion on separate pages from your questions.
- The format should be easy to follow and understand.
- Check that filter questions work as intended.

- Stick to questions that meet your goals rather than just asking every question you can think of. In general, survey completion rates tend to diminish as surveys become longer.
- Leave enough space for respondents to answer open-ended questions.
- Make sure to include all likely answers for closed-ended questions, including an “Other” option in case respondents feel that none of the provided answers suits them. Where appropriate, also include “Not Applicable” and “Choose not to respond” options.
- Ensure that questions flow well and follow a logical progression. Begin the survey with more general questions and then follow with more specific or harder issues. Finish with general demographic topics (e.g., age, gender, etc.), unless you are using these to screen for eligibility at the beginning of the survey. Group questions by theme or topic.

Some Further Considerations

CONSIDER YOUR AUDIENCE

Think about the people who will be participating in your survey. Let’s say, for example, that you want to conduct a survey in a school where many students have recently immigrated to the country as refugees with their families and don’t yet speak the local language well.

In what language(s) should you conduct the survey? Should you only conduct the survey in the local language or instead prepare it in several languages? If you’re preparing different versions of the survey, who will ensure consistency among the translations?

What kind of vocabulary should you use? These are students who may not be taking the questionnaire in their native language. You will probably want to use fairly basic language, especially if many are not native speakers. We have not specified their age, but it would be useful to take this information into consideration when thinking about vocabulary.

WORD INSTRUCTIONS AND QUESTIONS CLEARLY AND SPECIFICALLY

It is important that your instructions and questions are simple and coherent. You want to make sure that everyone understands and interprets the survey in the same way. For example, consider the difference between the following two questions.

- When did you first notice symptoms? ______________________
- When did you first notice symptoms (e.g., MM/YYYY)? __ __ / __ __ __ __

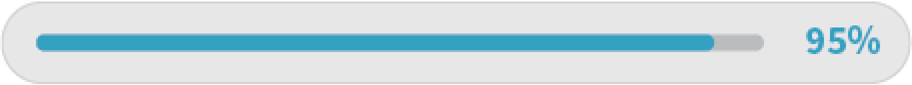
If you want the question to be open to interpretation so that people can give answers like, “after I came back from my last trip,” then the first option is okay. However, if you’re really wondering how long each person has been experiencing symptoms, the second version of the question lets the respondent know that you’re interested in the specific time at which the symptoms first occurred.

PAY ATTENTION TO LENGTH

We often want to collect as much information as possible when conducting a survey. However, extremely long surveys quickly become tedious to answer. Participants get tired or bored which in turn decreases the completion rate. After a while, you risk getting inaccurate responses from people who are trying to finish the survey as rapidly as possible. When reviewing your questionnaire, ask yourself, “Do I really need this question? Will it bring any valuable data that can contribute to my reason for doing this survey?”

TRACK PROGRESS

It is good to indicate the respondents’ progress as they advance through the survey. It gives them an idea of their progression and encourages them to continue on. In a self-administered survey, a progress bar such as the one below could be placed at the beginning of each new section or transition.



In an administered survey, the interviewer can simply make statements such as:

- “We are halfway done.”
- “There are only two more sections left.”
- “Only a few more questions to answer.”

TRAIN INTERVIEWERS

If you are conducting an administered survey, you will have to prepare your interviewers ahead of time, explain their tasks, and make sure that they understand all of the questions properly. You may also want them to be supervised when they are first conducting the survey to monitor for quality to control for interviewer effects.

PRETEST

Pretesting means testing your survey with a few initial respondents before officially going out in the field. This allows you to get feedback to improve issues you might have with length, technical problems, or question ordering and wording. Sometimes this step gets skipped if there’s a tight deadline to meet, but you shouldn’t underestimate its value. It can save a lot of time and money in the end.

ANONYMIZE RESPONSES

We briefly mentioned response anonymization earlier when we talked about questionnaire structure. Sometimes, people are afraid to give an honest response for fear that it might be traced back to them. By anonymizing responses you will get more accurate data while protecting the identity of your participants. There are times when you may need to keep respondents’ contact information associated with their responses, like if you’ll be conducting a follow-up survey and will need to know what their initial answers were. Be clear in your survey about whether you’re anonymizing responses or not so respondents know if you’re storing their contact information.

You should always anonymize any results you present, whether the data you collect are anonymous or not.

INCENTIVES

The use of incentives can be contentious. You want to encourage as many people as possible to participate in your survey, but you also don't want people completing the survey just to get a reward. Larger reward amounts can also raise ethical concerns of participant coercion (i.e., having individuals participate when they would otherwise refuse due to personal discomfort, risk of repercussions, etc.).

If the survey is long and time consuming, consider giving a small reward or stipend if a respondent completes the questionnaire. If the survey is shorter, consider doing a prize drawing from respondents who complete the survey and choose to provide their contact information.

Wrapping Things Up

As you have noticed, there are a number of factors you will need to consider when you're designing your survey and deciding what types of question to use. Survey creation can seem daunting at first, but remember this is the foundation of your data analysis. Getting this step right will make the rest of your work much easier later on.

For further reading, please see our resources section.

CHAPTER 5

ADDITIONAL DATA COLLECTION METHODS

BY DYANNA GREGORY

Not all types of information are easily gathered using surveys. Surveys are self-report tools that take place at a single point in time, so exact measurements, external impressions of reactions, and data about things that happen over time can be difficult to capture. Nevertheless, these are important pieces of information that you may need, so let's discuss a few of the other data collection tools you can use to better collect these types of data.

Method	Good if:
Direct measurement	Values need to be exact; Information likely to be biased if self-reported
Focus Groups	Don't know exactly what you want to ask yet; Interested in full spectrum of reactions or multiple topics; Interested in group dynamic; Interested in decision process
Observation	What you're measuring is easily and publicly observable; You want to make notes about participant behavior
Examination of existing documents	The data you are interested in are already recorded elsewhere (e.g. on receipts, logs of web traffic, etc.)
Diaries	Need to track variables over time; Information likely to be biased if recalled later

Direct Measurement

There are some variables that should be measured rather than surveyed if you're trying to obtain an exact, correct statistic. Many medical variables, for example, are difficult if not impossible to gather accurately using a survey. Let's say you need to collect data on participants' weight at the beginning of a study. There are a few common reasons someone might report an inaccurate number.

- Lack of information: they didn't actually know how much they weighed when asked

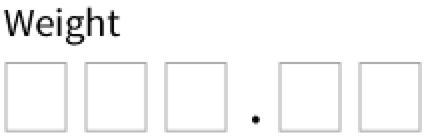
- Social expectation: they felt there was a “correct” answer they were supposed to give
- Ease of response: they knew about how much they weighed but didn’t think the exact answer was needed; “I’m about 170 lbs, so I’ll say that.”

Whether your data need to be exact depends on how you’re using the the information. If you’re not concerned about having a precise measurement and an estimate will work, then a survey might be fine as long as it asks something people will be able to reasonably estimate. If you have a variable is likely to be incorrectly self-reported and it is important that these data are current and accurate, direct measurement should be used instead of a survey.

In direct measurement, you use an appropriate device to measure the variable and then record the value in the dataset. This is often done for health-related variables that a person wouldn’t be able to “just know.” The measurements are usually captured on forms and the data is transferred from the forms into the dataset.

It is important for your forms to be clear about how you want the measurement to be recorded. You should indicate the preferred units and the precision you want the measurement captured with. One of the easiest ways to communicate this is by allotting a specific number of boxes so the person taking the measurement knows how many digits you want recorded.

The image below shows an example for how we might set up a form for capturing adult weight. Here, we want the measurement in pounds, and we want the number recorded to two digits after the decimal place.



Weight

□□□.□□

When relevant, include a place to record what device was used to take the measurement.

Focus Groups

Sometimes it's helpful to watch people while they're responding to your questions, see their thought processes, and observe them interacting with others. You may also have a wider variety of topics you would like to cover than would make sense for a survey or you might not be sure exactly what questions need to be asked yet. Focus groups can help in all these situations.

A basic focus group works a lot like a facilitated book club meeting. A small group of people (usually 6 to 12 individuals) is gathered to discuss their thoughts on a specific topic, and this discussion is led by a trained moderator. The moderator asks a variety of questions in order to get more in-depth opinions than a survey would answer alone. Focus groups often have more flexibility than surveys in that the questions are not entirely pre-determined. The moderator has more freedom to explore the answers the respondents provide and formulate new questions based on those responses. Additionally, since participants are in a group, the answers one person gives may cause another to think of answers they might not have otherwise.

However, both the group setting and the presence of the moderator can create bias in participant responses. It is important to keep this in mind when reviewing and analyzing focus group data.

Observation

Sometimes the data you need to collect are a matter of observation. Let's go back to Fictionals Ice Cream Parlour for a moment. You recently purchased new furniture for the store, and you're considering a couple of different layouts for it. You want to see which layout seems to work best for customer flow, so you set up the furniture one way for a few days and record your personal observations about customer movement within the shop. Then you switch the furniture to the other layout and again record what you notice. These data can help you figure out what other questions you might want to ask or what other data you need before making your decision.

You can use observation in this way to gain insight into naturalistic behavior. This can be especially useful if your subjects of interest are not human and can't answer survey questions: scientists rely on observation as a data collection technique all the time!

One of the major shortcomings of this method is that the presence of an observer changes an observation. Every person sees an event from their own perspective, and their report of that event is influenced by that perspective. You can decrease this bias by having several observers so that the data gathered represents multiple viewpoints.

Examination of Existing Documents

In some cases, the data that you need already exist as parts of other documents and your data collection is really a matter of getting all of that information into one place.

As the manager of Fictionals Ice Cream Parlour, you want to take a look back at your sales for the last six months to see if your recently-added menu items have been profitable. This information is already available to you through your receipts or POS software data. You just have to get it set up in a way that allows you to easily work with and analyze it.

Other existing documents that are frequently used to compile information include books, newspapers, web traffic logs, and webpages. There are also entire datasets that are available for use. These are covered in more detail in the chapter on Finding External Data.

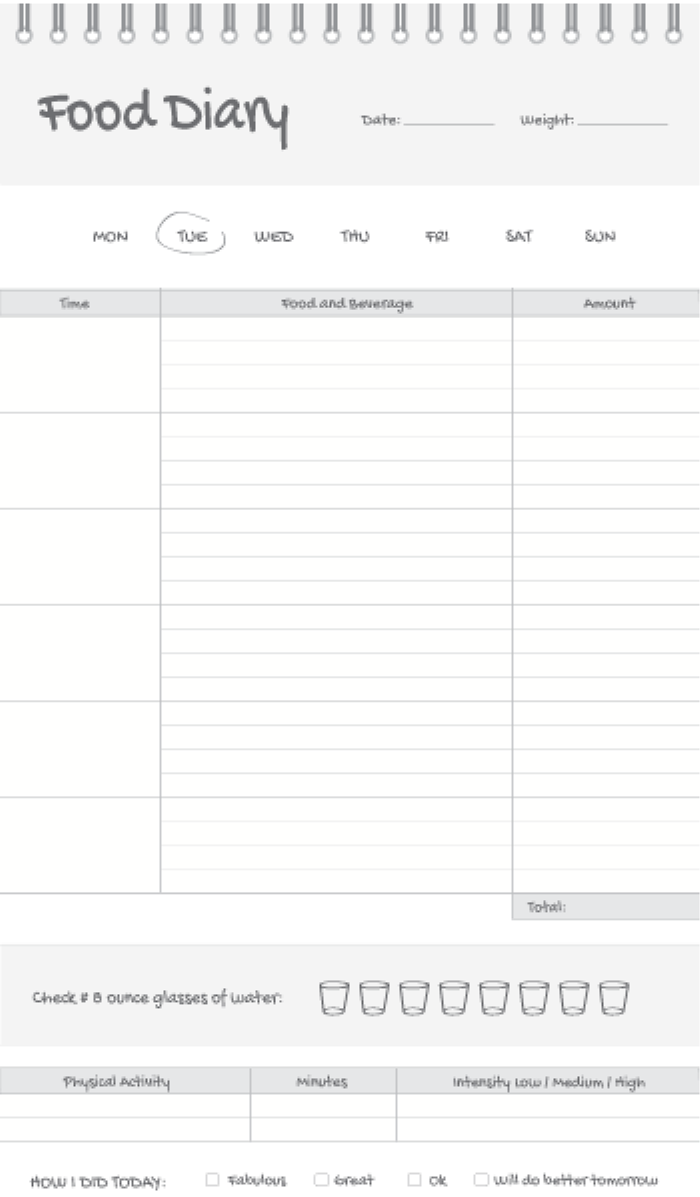
When you're using other documents as the main source of your data, you should first set up a data collection plan, much the way that you design a survey. The plan should detail what pieces of data you're looking for, the level of measurement you want to capture them at, the time frame you need (e.g. do you only want data from the last 6 months? the last 12?), and how much data you need (e.g. do you want to look at all the receipts or just a sample of them?).

If any of the sources are ones that you don't own, make sure to properly cite them. It's important to credit others' work, and it's also important to be able to support your research if anyone challenges your information later on.

Diaries

Diary forms can be useful if you’re collecting data from people about things that are happening over extended periods of time. They are particularly helpful if you them to record a lot of details that could easily be misremembered or forgotten.

You might have seen something similar to this before:



Food Diary

Date: ____________ Weight: ____________

MON [✓ TUE] WED THU FRI SAT SUN

Time	Food and Beverage	Amount
		Total:

Check # 8 ounce glasses of water: □ □ □ □ □ □ □ □

Physical Activity	Minutes	Intensity Low / Medium / High

HOW I DID TODAY: □ Fabulous □ Great □ Ok □ Will do better tomorrow

Diary forms are often used for tracking things like meals, medications, or exercise but can be used for anything you want to record over time. They can be paper forms or computerized, but should follow the same design principles as surveys.

Fields should be well-labelled and the instructions should be very clear about how (and how often) the diary should be completed.

Using Multiple Collection Methods

When you're considering a project as a whole, it is possible that not all the research questions you're trying to address can be answered using data collected from just one of the methods discussed so far. You may find that your survey will need to be supplemented with some direct measurements, or you may need to have your focus group participants complete diary forms.

Just as you can benefit from a combination of survey types, you can also benefit from incorporating multiple types of data collection if the data you need would most appropriately be gathered in different ways. Consider what approaches make the most sense for the data you need, and then select the best choices that are practical within your budgetary and time constraints.

CHAPTER 6

FINDING EXTERNAL DATA

BY JANE FOO

Running your own study to collect data is not the only or best way to start your data analysis. Using someone else's dataset and sharing your data is on the rise and has helped advance much of the recent research. Using external data offers several benefits:

Time / Cost	Can decrease the work required to collect and prepare data for analysis
Access	May allow you to work with data that requires more resources to collect than you have, or data that you wouldn't otherwise have access to at all
Community	Promotes new ideas and interesting collaborations by connecting you to people who are interested in the same topic

Where to Find External Data

All those benefits sound great! So where do you find external data? To help narrow your search, ask yourself the following questions:

Scope	What is the scope of the data you're looking for? What are the: • geographic boundaries? • specific data attributes (such as age range)? • time periods?
Type	What type of data are you looking for? Do you need: • statistics? • research data? • raw data? • data that have been collected using a specific method?
Contribution	How will the data contribute to your existing data analysis? Do you need several external datasets to complete your analysis?

PUBLIC DATA

Once you have a better idea of what you're looking for in an external dataset, you can start your search at one of the many public data sources available to you, thanks to the open content and access movement (http://book.openingscience.org/) that has been gaining traction on the Internet. Many institutions, governments, and organizations have established policies that support the release of data to the public in order to provide more transparency and accountability and to encourage the development of new services and products. Here's a breakdown of public data sources:

Source	Examples
Search Engines	Google (http://www.google.com)
Data Repositories	re3data.org DataBib (http://databib.org/) DataCite (http://www.datacite.org/) Dryad (http://datadryad.org/) DataCatalogs.org (http://datacatalogs.org/) Open Access Directory (http://oad.simmons.edu/oadwiki/Data_repositories) Gapminder (http://www.gapminder.org/data) Google Public Data Explorer (https://www.google.com/publicdata/directory) IBM Many Eyes (http://www.manyeyes.com/software/analytics/manyeyes/datasets) Knoema (http://knoema.com/atlas//)
Government Datasets	World Bank (http://data.worldbank.org/) United Nations (http://data.un.org/) Open Data Index (https://index.okfn.org/) Open Data Barometer (http://www.opendataresearch.org/project/2013/odb) U.S. Government Data (https://www.data.gov/) Kenya's Open Data Initiative (https://opendata.go.ke/)
Research Institutions	Academic Torrents (http://academictorrents.com/) American Psychological Association Other professional associations Academic institutions

If you decide to use a search engine (like Google) to look for datasets, keep in mind that you'll only find things that are indexed by the search engine. Sometimes a website (and the resource associated with it) will be visible only to registered users or be set to block the search engine, so these kinds of sites won't turn up in your search result. Even still, the Internet is a big playground, so save yourself the headache of scrolling through lots of irrelevant search results by being clear and specific about what you're looking for.

If you're not sure what to do with a particular type of data, try browsing through the Information is Beautiful awards (http://www.informationisbeautifulawards.com) for inspiration. You can also attend events such as the annual Open Data Day (http://opendataday.org/) to see what others have done with open data.

Open data repositories benefit both the contributors and the users by providing an online forum to share and brainstorm new ways to study and discuss data. In some cases, data crowdsourcing has led to new findings that otherwise would have developed at a much slower rate or would have not been possible in the first place. One of the more publicized crowdsourcing projects is Foldit (http://fold.it/portal/info/about) from the University of Washington, a Web-based puzzle game that allows anyone to submit protein folding variations which are used by scientists to build new innovative solutions in bioinformatics and medicine. And recently, Cancer Research UK released a mobile game called Genes in Space (http://scienceblog.cancerresearchuk.org/2014/02/04/download-our-revolutionary-mobile-game-to-help-speed-up-cancer-research/) that tasks users with identifying cancer cells in biopsy slides which in turn helps researchers cut down data analysis time.

NON-PUBLIC DATA

Of course, not all data is public. There may come a time when you have access to a special collection of data because of your status within a particular network or through an existing relationship. Or maybe you come across a dataset that you can buy. In either case, you typically have to agree to and sign a license in order to get the data, so always make sure that you review the Terms of Use *before* you buy. If no terms are provided, insist on getting written permission to use the dataset.

Assessing External Data

Let’s say you’ve found a dataset that fits your criteria. But is the quality good enough?

Assessing data quality means looking at all the details provided about the data (including metadata, or “data about the data,” such as time and date of creation) and the context in which the data is presented. Good datasets will provide details about the dataset’s purpose, ownership, methods, scope, dates, and other notes. For online datasets, you can often find this information by navigating to the “About” or “More Information” web pages or by following a “Documentation” link.

Feel free to use general information evaluation techniques when reviewing data. For instance, one popular method used by academic libraries is the CRAAP Test, which is a set of questions that help you determine the quality of a text. The acronym stands for:

Currency	Is the information up-to-date? When was it collected / published / updated?
Relevancy	Is the information suitable for your intended use? Does it address your research question? Is there other (better) information?
Authority	Is the information creator reputable and has the necessary credentials? Can you trust the information?
Accuracy	Do you spot any errors? What is the source of the information? Can other data or research support this information?
Purpose	What was the intended purpose of the information collected? Are other potential uses identified?

Finally, when you review the dataset and its details, watch out for the following red flags:

- Details of data collection method not stated
- No contact information
- Unclear ownership of data
- No authoritative entities or credentials associated with data collector or data
- Terms of use or license includes details that raises questions (e.g. data can't be used for scientific study)
- Inconsistent or missing metadata
- Large chunks of missing data without explanation or reference points
- Raw data looks "too perfect"
- Published articles challenge or question the dataset

Using External Data

So now you have a dataset that meets your criteria and quality requirements, and you have permission to use it. What other things should you consider before you start your work?

Checklist	
Did you get all the necessary details about the data?	Don’t forget to obtain variable specifications, external data dictionaries, and referenced works.
Is the data part of a bigger dataset or body of research?	If yes, look for relevant specifications or notes from the bigger dataset.
Has the dataset been used before?	If it has and you’re using the data for an analysis, make sure your analysis is adding new insights to what you know has been done with the data previously.
How are you documenting your process and use of the data?	Make sure to keep records of licensing rights, communication with data owners, data storage and retention, if applicable.
Are you planning to share your results or findings in the future?	If yes, you’ll need to include your data dictionary and a list of your additional data sources.

Your answers to these questions can change the scope of your analysis or prompt you to look for additional data. They may even lead you to think of an entirely new research question.

The checklist encourages you to document (a lot). Careful documentation is important for two big reasons. First, in case you need to redo your analysis, your documentation will help you retrace what you did. Second, your documentation will provide evidence to other researchers that your analysis was conducted properly and allow them to build on your data findings.

Giving Credit to External Data Sources

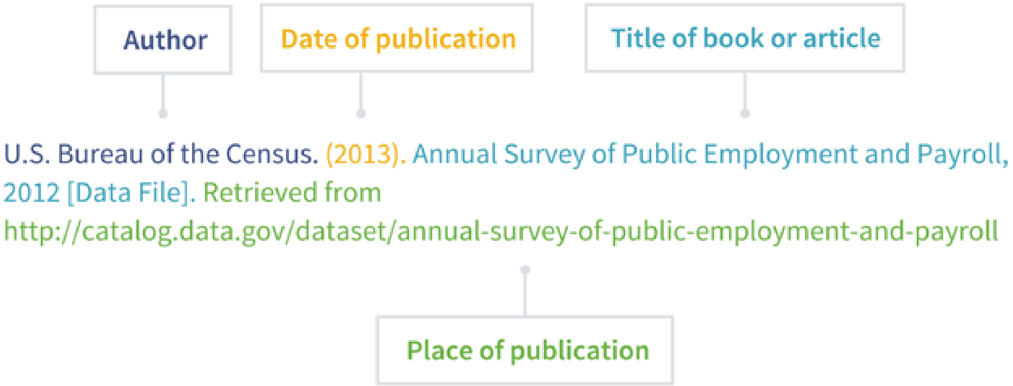
Simply put, crediting the source of your external dataset is the right thing to do. It’s also mandatory. Ethical research guidelines state that crediting sources is required for any type of research. So always make sure that you properly credit any external data you use by providing citations.

Good citations give the reader enough information to find the data that you have accessed and used. Wondering what a good citation looks like? Try using an existing citation style manual from APA (https://owl.english.purdue.edu/owl/section/2/10/), MLA (https://owl.english.purdue.edu/owl/section/2/11/), Chicago (http://www.chicagomanualofstyle.org/tools_citationguide.html), Turabian (http://www.press.uchicago.edu/books/turabian/turabian_citationguide.html), or Harvard (http://guides.is.uwa.edu.au/harvard). Unlike citations for published items (like books), citations for a dataset vary a great deal from style to style.

As a general rule, all styles require the author and the title. In addition, editor, producer or distributor information (location, publication date), access date (when you first viewed the data), details about the dataset (unique identifier, edition, material type), and the URL may be needed. For government datasets, use the name of the department, committee or agency as the group / corporate author.

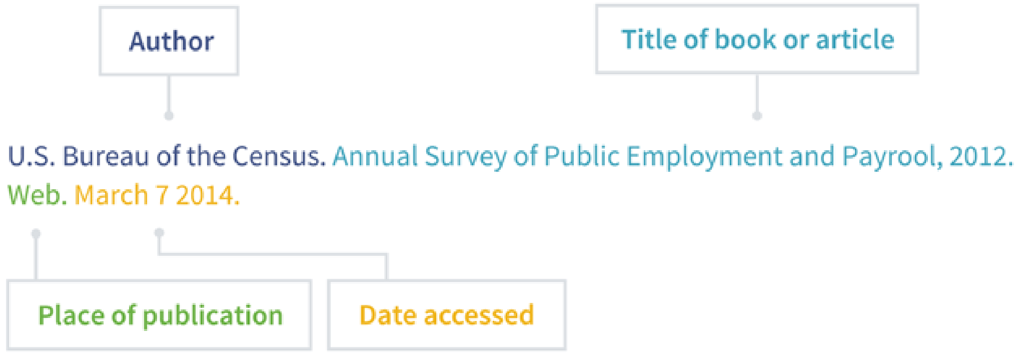
For example, let's say you're using the U.S. Census Annual Survey of Public Employment and Payroll.

The APA Style Manual (Publication Manual of the American Psychological Association, 6th edition) would cite this the following way:



U.S. Bureau of the Census. (2013). Annual Survey of Public Employment and Payroll, 2012 [Data File]. Retrieved from http://catalog.data.gov/dataset/annual-survey-of-public-employment-and-payroll

while the MLA Style Manual (MLA Handbook for Writers of Research Paper, 7th edition) cites the same census data as:



Data repositories and organizations often have their own citation guidelines and provide ready citations that you can use “as is”. The Interuniversity Consortium for Political and Social Research (ICPSR) (http://www.icpsr.umich.edu/icpsrweb/landing.jsp), The National Center for Health Statistics (http://www.cdc.gov/nchs/), Dryad (http://datadryad.org/), PANGAEA (http://www.pangaea.de/), and Roper Center Data (http://www.ropercenter.uconn.edu/) all provide guidelines for citing their datasets.

This chapter gives you a brief look into external data: the important takeaway is that we are only at the start of a significant growth in data thanks to the technologies that now make massive data storage and processing an affordable reality. Open datasets in particular have the potential to become a de facto standard for anyone looking for data to analyze.

PREPARING DATA

After you've collected the ingredients you want for your recipe, you have to prepare them before you start cooking. This is the prep step, in which you wash, chop, and otherwise get your ingredients ready. Many of us dislike this step because it isn't much fun, but if you've ever had a dish with dirty vegetables, you know firsthand how much the end result can suffer if this step is done poorly or skipped altogether!

As in cooking, many also dislike the prep step when working with data, but it is necessary. You can't just grab the data, throw it into a visualization, and expect it to come out right. If you want to create the best graphics you can, you need to make sure your data is prepared correctly. Some of your data may need to be separated into different groups, some may need converting, and some may need to be cleaned, like the vegetables.

In this section, we'll talk about some of the common preparation and cleaning tasks encountered when working with data, and about ways that you can potentially decrease the amount of time you'll need to spend doing them. No matter how you slice it, if you get your data set up the right way, it will make everything a go a lot smoother when you get to the visualization stage.

CHAPTER 7

GETTING DATA READY FOR CLEANING

BY OZ DU SOLEIL

One aspect of data that is difficult both to learn and to teach is how to get your data in shape so that it's useful. There are a few common preparation tasks that you might encounter when working with data, especially if you're handling data someone else collected. While we can use computers to perform many of the actions required for data preparation, there is a certain amount of manual involvement in figuring out exactly which tasks need to be done. For this reason, many people admittedly hate this part of working with data, but someone has to clean this stuff up!

Whether you're a highly-paid data analyst with billions of records or a 1-person business with a 90-person contact list, you'll be faced with messy data at some point. Unfortunately, data preparation isn't a straightforward task and there's no one right way to go about doing it. Each dataset is unique and some techniques may be used just once in a lifetime, as you'll see in the examples below.

Separating Data

The first part of data preparation is separating data into the fields that will be most useful to you.

Have you ever been given a dataset where you couldn't directly access the information you needed? Maybe complete addresses were in a single field, preventing you from getting statistics about specific cities and states. Perhaps a vendor sent an inventory spreadsheet with 6000 numbered parts but their numbering system includes the warehouse code followed by the part number and you need the part numbers by themselves.

You Want	They Provide
C77000S	GA3C77000S
W30000P	GA1W30000P
D21250G	DE1D21250G

Consider the challenges presented in the dataset below:

Mall	Address	City	State
Warm Willows Mall Peters Road Marrison, MI			
Jaspers Martinson & Timberlake Rds Reed, FL			
Lara Lafayette Shoppes 17 Industrial Drive Elm, CT			

You want the mall name, address, city, and state to be stored in separate fields. The dataset has hundreds of malls in it so splitting the entries apart by hand would take a lot of time. The mall names are bold, which makes it easy to visually distinguish where the mall name stops and the address starts, but not all the addresses begin with numbers, so no standard tool exists for separating the data into different fields. Code can be written that will recognize the bold and plain font weights and peel them apart, but since this is a rare situation, it is likely that you won't have it right at hand. This is a case where you might write code (or have someone write it for you) that you'll never use for this purpose again.

We can't teach you everything about how to split data in this chapter since every situation is unique. However, there are some strategies that are useful in many cases. We'll go through a few of these, as well as some common challenges and their sources, to help get you started.

Let's Begin

A straightforward example that many people encounter is separating first names from last names. You may receive a dataset where full names are contained in a single field and you need the first and last names to be separate, or there may already be separate fields for first and last names, but some of the entries have the full name in one of the fields.

All in one field:

Name
Keith Pallard
Fumi Takano
Rhonda Johnson
Warren Andersen
Juan Tyler
Cicely Pope

Two fields, but not all info is in the correct place:

First Name	Last Name
	Keith Pallard
Fumi Takano	
Rhonda	Johnson
Warren Andersen	
Juan	Tyler
Cicely	Pope

When datasets come to us this way, the challenge is seemingly easy to resolve. There are common ways of pulling apart the first names and last names for Keith, Fumi, and Warren. The easiest way is to look for the space, break the name apart at the space, and voila!

This is simple enough, but when we add in reality, things get complicated quickly. What if, for instance, your dataset has thousands of records? You'll be spending a lot of time manually separating names, and there are far more potential name combinations than just first and last name.

Middle Initials	Martina C. Daniels
Professional Designations	Lloyd Carson DVM
2-part Last Names	Lora de Carlo
Prefixes	Rev Herman Phillips
Suffixes	Jimmy Walford III
Hyphenated Last Names	Tori Baker-Andersen
Last Name First	Kincaid Jr, Paul
2-part First Names	Ray Anne Lipscomb
Prefixes/Suffixes	Rev Rhonda-Lee St. Andrews-Fernandez, DD, MSW
Other fields incorrectly included	Murray Wilkins 993 E Plymouth Blvd
No/Missing First Name	O’Connor
No/Missing Last Name	Tanya
No name at all	
I have no earthly idea!	JJ
Not a person’s name at all	North City Garden Supply

Now what? Let’s Make Some Decisions

Let’s say we’re faced with separating the names so that you can sort by last name and that our list has 500 names (too large to consider reformatting by hand).

Before starting, we need to know certain things:

- Why is it important to parse this specific field? Is anything being hurt by having names all in one field?
- What do we want the result to look like?
    - Is it important to keep ‘Rev.’ and create a separate field to hold other titles like Dr., Mrs., Capt., etc.?
    - Should ‘Jr’ stay with the last name or should there be a separate field for it?

    - Should middle initials be kept? In their own field? With the first name? With the last name?
    - Do we want to keep professional designations?
- Is it worth the cost of getting it done? In simple cases, you may be able to hire someone else to complete the work, but this might not be ideal for complicated ones. What if a professional said they could fix the list for $1000?
- What do we do with incomplete or incorrect entries?

These questions should be answered before you start working with your data. If you just dive in, you can create a larger mess, and these decisions can actually make the process much easier. For example, if you decide that the professional designations don't matter, there might be an easy way to get rid of them and simplify other parts of the project.

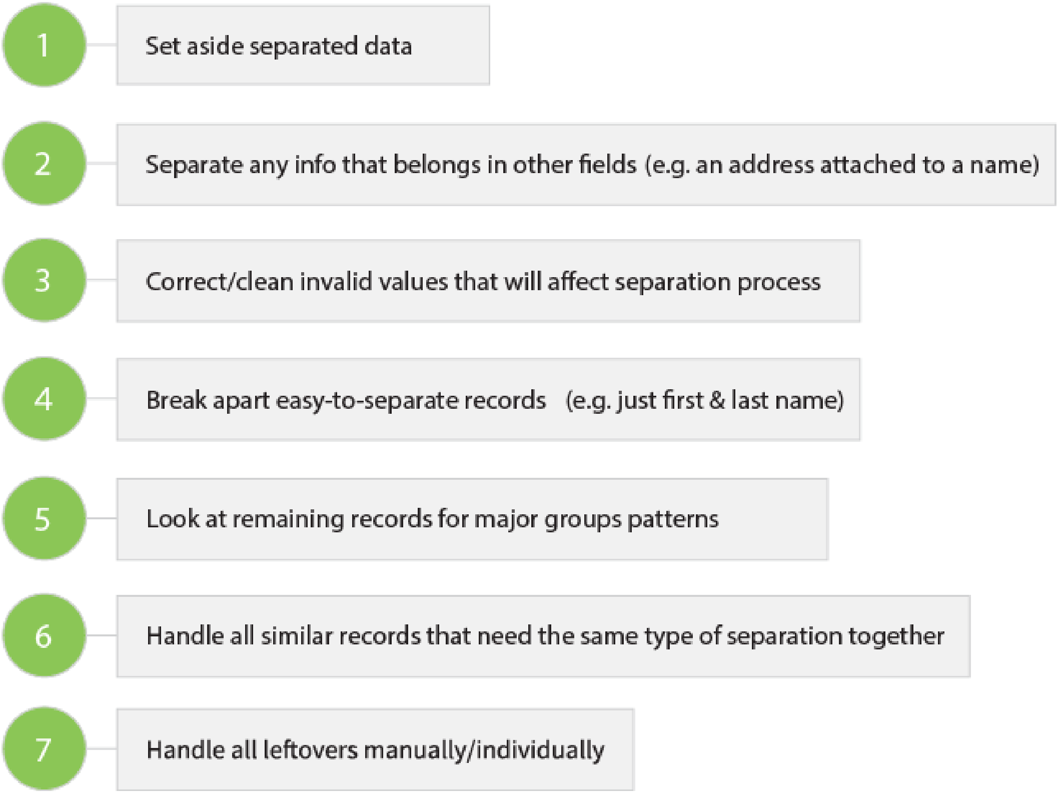
Say you're working with a list of 20,000 names, 19,400 from the US and 600 from Turkey. The honorific titles in the US come *before* the name (e.g. **Mr.** John Smith), whereas Turkish honorific titles come *after* the name (e.g. Jon Smith **Bey**). You're trying to figure out if you need separate fields for the Turkish data or another dataset altogether, so you ask the client what their preference is.

Their answer is simple. They aren't conducting business in Turkey and they don't care if you delete those records. GREAT! Just 19,400 records left.

Now, how do we break up these data?

There are so many techniques, we could write an entire book on this subject alone. You can break apart the data in Excel, or if you have programming skills, you can use Python, SQL, or any number of other languages. There's too much for us to cover in this chapter, but for some good references, check out our Resources Appendix. For right now, we'll cover some basic starting strategies so you'll be better equipped to take on those references when you're ready.

You should *always* make a copy of your dataset before doing any data preparation or cleaning in case a change you make is incorrect and you need to refer to the original.



1. Set aside separated data
2. Separate any info that belongs in other fields (e.g. an address attached to a name)
3. Correct/clean invalid values that will affect separation process
4. Break apart easy-to-separate records (e.g. just first & last name)
5. Look at remaining records for major groups patterns
6. Handle all similar records that need the same type of separation together
7. Handle all leftovers manually/individually

LOOK FOR LOW-HANGING FRUIT

In many instances, most of the data that need to be separated are fairly simple. Out of 500 names, you may discover that 200 have just a first and last name. Set these aside. Leave them alone. Then look at the remaining 300 names.

IDENTIFY THE ODDITIES

Look through your dataset for the no-names, complicated names, incomplete names, non-person names and any entries you don't know what to do with. Set these aside. We'll say these comprise 40 more names.

LOOK FOR SIMILARITIES

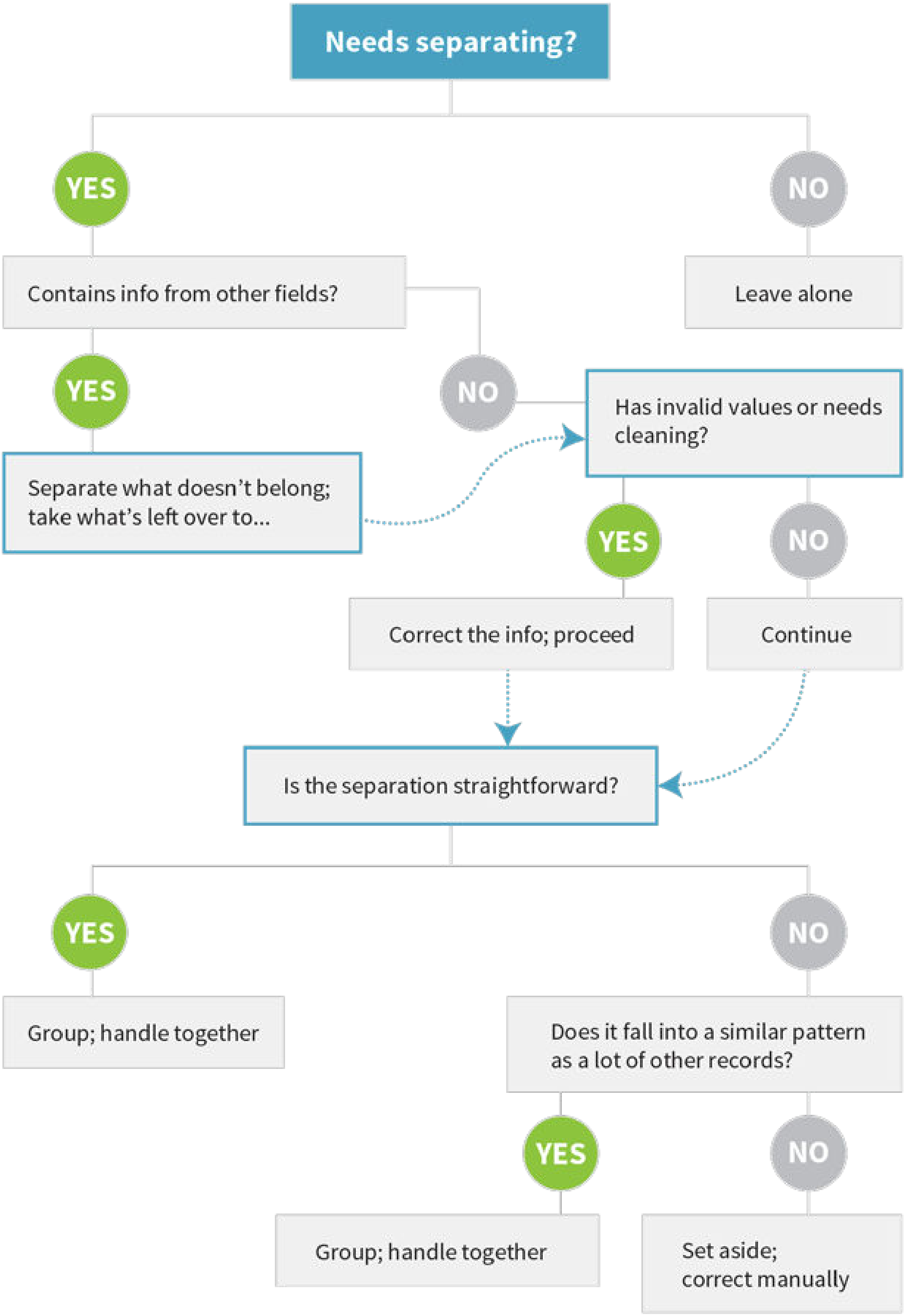
Of the 260 remaining names, maybe 60 are complicated by professional alphabet soup after their names. Whether you're deleting these professional designations or putting the designations in their own field separate from the name, work with them all at once. Now the designations are separate (or deleted) and you have a field with just names. For those that have only a first and last name, add those to the 300 that were set aside at the beginning.

We put the 2-part last names and other remaining types into their own groups.

MANUAL EFFORT

The 40 oddities you identified at the start may come down to just re-typing manually, depending what is odd about them.

> Sometimes, when we're handling data in real life, records take care of themselves. For example, maybe a non-name record turns out to be a duplicate of a record that's complete. In that case, we can just delete the incorrect record and move on.


Needs separating?
YES
NO
Contains info from other fields?
Leave alone
YES
NO
Has invalid values or needs cleaning?
Separate what doesn’t belong; take what’s left over to...
YES
NO
Correct the info; proceed
Continue
Is the separation straightforward?
YES
NO
Group; handle together
Does it fall into a similar pattern as a lot of other records?
YES
NO
Group; handle together
Set aside; correct manually

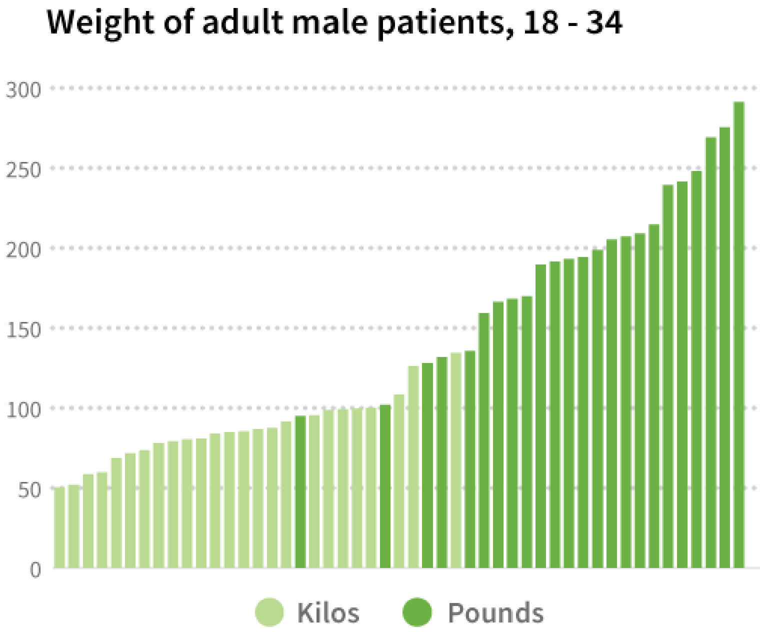
Commonly Problematic Fields

Depending on how data are collected, there are other fields you may be interested in splitting up besides names. Some of the most common are:

- Address information
- Phone numbers, particularly if you need the area code to be separate
- Emails, if you're curious about domain information
- Date, if you only want year or month

UNITS AND UNIT CONVERSION

Another important data preparation task is making sure that all the data in a single field are given in the same units. Ideally, you would have specified the unit type on the input form, but you may be working with someone else's dataset or the data may have been collected directly from other sources where the information was compiled by machines that measure in particular units. For example, you might have details from medical records from multiple countries and the patients' weights could be in pounds for some records but in kilograms for others. It is important to convert all the numbers to either pounds or kilograms so that they are all on the same scale, otherwise the records cannot be directly compared to each other and any visualization you do of the original data would look quite strange indeed!

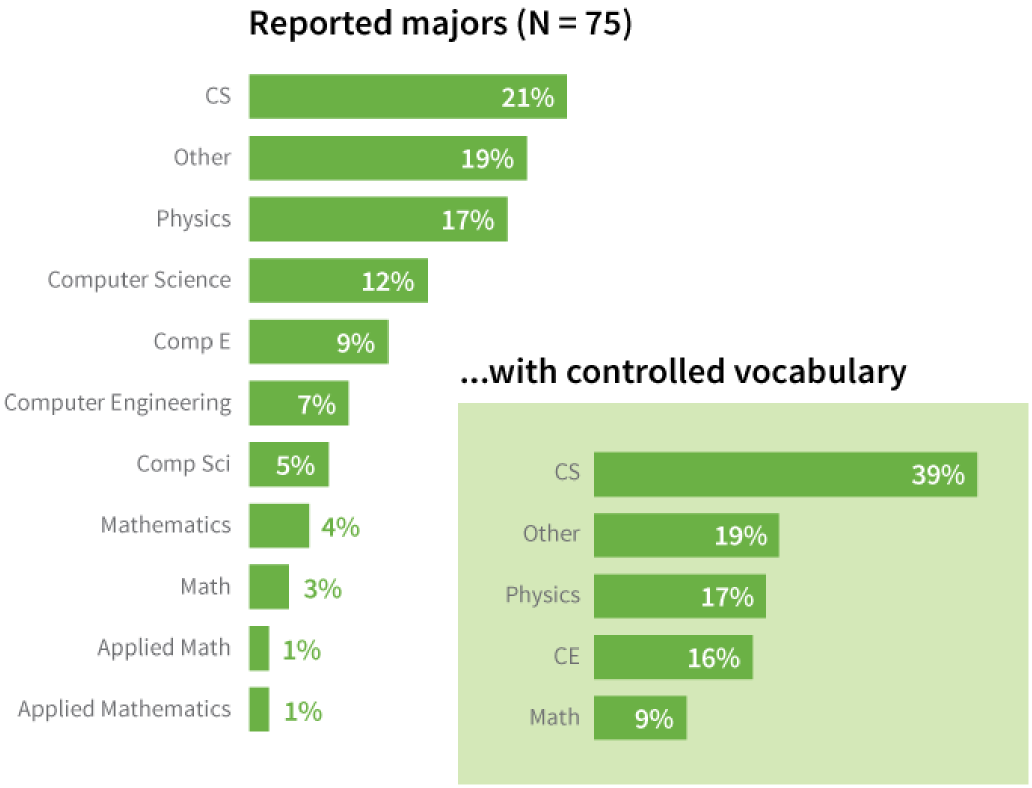


You should do a quick scan of the variables in your dataset to identify any fields that could potentially report multiple units so that you can figure out if these conversions are necessary. If they are, you might need to reference an additional field to identify what the original units are that you should be converting from so you know what records need to be changed. If there isn't a specific field that lists the units themselves, other fields like geographic location can be helpful. If you cannot find a field that lets you know what the units are but you suspect that a conversion is necessary, you should contact the original provider of the data to obtain that information. It may turn out that you actually don't need to convert anything and your data are just strange, but it's better to check and make sure than to ignore a potential error in your data.

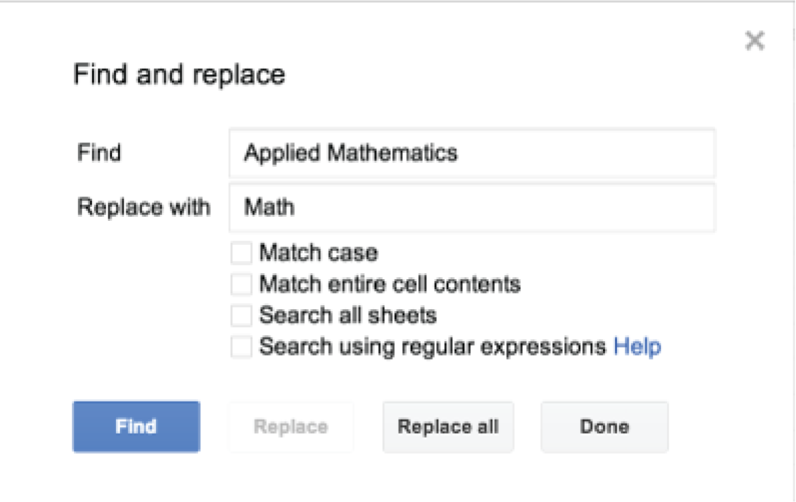
Another type of conversion that is sometimes less obvious is data type conversion. It is important to make sure that all of the data in a single field is being stored as the same type or your visualization might not appear correctly, depending on how your software interprets the information. For example, "80" might look like a number to you, but the computer might actually be storing it as a string of text rather than in a numeric form. Some visualization software will go ahead and treat any text that looks like a number as a number, but other software will not. It is a good idea to check that each of your fields (or variables) is stored as a single, specific data type, such as text or numeric. This will ensure that anything that is captured in that field is being treated as the data type that you mean for it to be.

CONTROLLING FOR INCONSISTENCIES

One of the most work-intensive tasks in cleaning data is handling inconsistent information. For example, we see multiple languages popping up in a single dataset more often now that the internet allows us to easily collect data from users all around the world. This can sometimes create problems when you're trying to group things up for visualization purposes, depending on how your data entry form is designed and what data from the set you're using. Perhaps you do a survey of college students and one of the text fields asks for their major. One student might enter “Math” while another enters “Mathematics” and another types “Applied Mathematics.” You know that all these are considered the same major on your campus, but a computer or visualization software would not group these records together. You would need to create a single controlled vocabulary term (e.g. change them all to say “Math” or “Mathematics”) or create a separate coded field if you wanted to have them treated as part of the same category.



Although the computer can aid in changing values, problems with data inconsistency often have to be handled semi-manually. If you are looking through text fields, much like when you do data coding, find and replace is your best friend. As long as you know what the main variants are of the data that you want to control, you can quickly change them all to be the same value.



Find and replace

Find: Applied Mathematics

Replace with: Math

□ Match case
□ Match entire cell contents
□ Search all sheets
□ Search using regular expressions

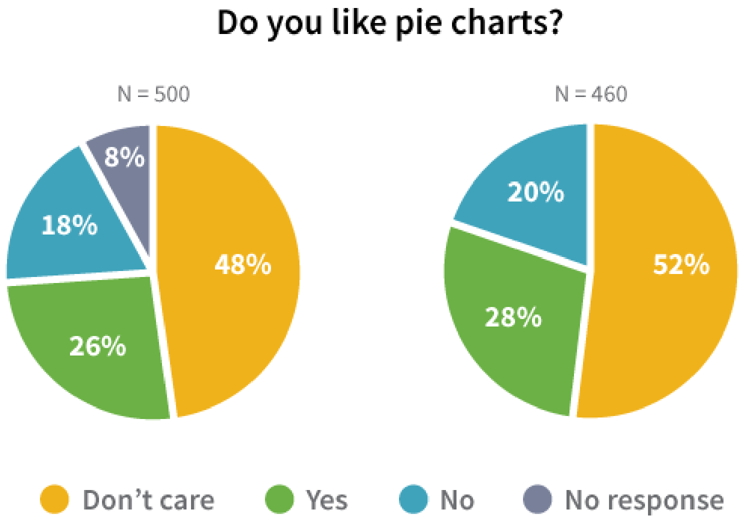
MISSING VALUES

One of the most frustrating problems is when data fields are simply blank or are incomplete. If data have not been collected, you may be able to return to the source to fill in what is missing, but you may also no longer have access to the source. It may also be possible that you do not know who the source was, for example, in the case of an anonymous survey. If you are not able to obtain the data, it is important to handle the missing values correctly. You should set pre-defined values that you can enter in where the missing values would be so when someone looks at your dataset, they know that the data for that field are actually missing and you didn't just forget to type them in.

These values should fall well outside the range of your reasonable data so that they clearly stand for missing data and nothing else. For datasets that do not deal with negative numbers, "-9" is often used as this default value, at least in the numeric fields. "999" is another popular choice for datasets that do not use large numbers, and a single period (.) is used by some statistical software programs that may be used with datasets with both negative and large numbers. For text fields, a single dash (-) is commonly used as the indicator for a missing value.

Keep in mind that a missing value is not inherently the same thing as an intentional non-response! You don't have the particular information that the question was asking about in either case, but when someone actively chooses not to answer, that in itself is a piece of data you wouldn't have if the question were unintentionally skipped. Those data aren't missing: you know exactly where they are, the respondent just doesn't want you to have them. As discussed in the Survey Design chapter, it is good to include a "Prefer not to answer" option for questions that may be of a personal nature, such as race or ethnicity, income, political affiliation, etc. That way, you can designate a code for this type of response so when you are going through your dataset later on, you know the difference between the respondents that purposely chose not to provide you a given piece of information and the data that are just missing altogether.

It is important to note that for purposes of basic descriptive visualization, missing values are described by including a non-responder category or by noting a change in the sample size. However, in inferential statistics, missing values may be dealt with in a variety of other ways, from exclusion to imputation to full analysis methods like use of EM algorithms.



MINIMIZING THE BURDEN OF DATA PREPARATION

The best solutions are preventive. If you are the one creating the form for user input, do whatever you can to prevent receiving data that will require intensive handling during the data preparation stages. In the Types of Data Checks chapter, we'll talk about different strategies for minimizing the number of data preparation tasks that need to be performed.

If you're not the one collecting the data but can speak with the people who are, try to work with them to identify and resolve any problematic data collection points using the strategies in the Types of Data Checking chapter as a guide.

POST-DATA PREPARATION

Once your data are separated into the fields you want, converted into the units and types you intend to work with, and terminology is controlled, you're ready to proceed to the data cleaning phase where you'll check for actual errors in the data. In the next three chapters, we'll talk about the basic process of data cleaning, different processes for performing checks on your data to find problems, and what you can and can't do with data cleaning.

CHAPTER 8

DATA CLEANING

BY MARIT BRADEMANN AND DYANNA GREGORY

Now that you have a prepped dataset, you're ready to get it clean. What does that mean though? What exactly are clean data and what do we have to do get them that way?

Well, when we clean data, we're going through and identifying incorrect information — wrong numbers, misspellings, etc. — and deciding whether to correct them (if they are correctable) or to remove them altogether. Like in data preparation, many data cleaning tasks are a combination of computerized tasks and manual work, since it is important for you to review the potential errors the computer identifies to see if they are, in fact, errors. Some of the items your computer flags as problems may turn out to just be extreme observations so it's critical that you remain involved in the process. Don't just automate it all or you risk the possibility of deleting valid data!

So What Do We Do?

Let's start with some of the most basic data cleaning procedures. We're going to use Excel for many of these these examples, but you can use any spreadsheet or data manipulation software to perform these procedures. A list of programs is available in the Appendix.

Range Checks

Range checks are a very straightforward procedure that we use on numeric fields to see if any values in the dataset are above or below the most extreme acceptable values for that variable. Let's use an example of homework scores. Pretend that you're a college professor and your teaching assistants entered the first set of homework scores for the semester. You want to make sure they entered everything correctly, so you go into the dataset and sort by the column that contains the scores for the first homework, graded on a scale of 0-100. You see the first few rows:

Student ID	HW 1 Score
679372531	980
673540288	99
674082892	97
673923590	96

There is a score of 980, so one of the TAs probably accidentally typed a zero after a score that should have been 98. You would want to flag the record and ask the TAs what the actual score should have been.

Visual scans of data during range checks can also reveal other potential problems even within the official bounds of the data:

Student ID	HW 1 Score
674472019	78
679029425	75
671822390	74
671278927	9

Here, one score is much lower than the others. In this situation, it is possible that the score should have been entered as a 90 instead. This is another record that would make sense to flag and check with the source record. This is a good example of why it is important to be personally involved in the data checking process for variables that carry a lot of weight, like student grades or primary outcome variables for research projects. However, hand-checking all the variables in a dataset can be very time-consuming, especially if the dataset is large. Additionally, not all variables are created equal: it is up to you to decide which variables require involved personal attention and which can be checked automatically.

Range checks will work no matter what your range is. Maybe you're expecting a decimal between 0 and 1, or your variable is normal body temperature in Fahrenheit so you're expecting mostly values between 97.0 and 99.0, allowing for people that run cold and hot. You can always look and see if there are values stored in that variable group that are too low or high.

You can also use "Filter" commands to check for all values that are outside the acceptable range of a variable. However, this doesn't catch values that are inside your range but that look "off" based on the rest of your data. Here, a basic range filter would detect the "980" score but not the "9" score. If you use filtering to do your range checks, it is a good idea to also use another method to look at the overall distribution of your data to catch any values that might seem "strange" in comparison to your other values.

Spell Check

Spell Check is another basic check that you can use to find problems in your dataset. We suggest doing this field-by-field rather than trying to do it across the whole dataset at once. The reason for this is that a word that might be considered a misspelling in one variable could be a valid word for another variable. A good example of this is the first name field. If you have a dataset with a first name field, many of those entries could trigger a spell check alert even though they are legitimate names. If instead you focus on a single field at a time, you can more quickly work through the dataset. In the example from the data preparation chapter where students were listing their major on a survey, say one of the students had just pulled an all-nighter and accidentally typed "Mtahmeitcs" instead of "Mathematics." A spell check on the "Major" field in your dataset would quickly identify the misspelling and you could change it to "Math" or "Mathematics," depending on which controlled vocabulary term you chose.

Pattern Matching/Regular Expressions

Another slightly more advanced type of data check involves pattern matching. This is the sort of check that you can use, for example, to make sure all the entries in a field are an email address. This involves something called regular expressions (often shortened to regex), which give you a way of telling the computer, “I only want things that look like {this} to be stored in that variable. Tell me if something in there doesn’t look like {this}.” The way that you indicate what {this} should be varies from program to program and can look a little complicated if you’ve never worked with it before. If you have ever used an asterisk (*) as a wildcard for searching, that’s actually part of a regex expression, so you already know a piece of it!

There are also pattern matching options in Excel and some advanced filter options that sometimes work even better. Check the resources section for links to more on regex and Excel filters and pattern matching.

Combination of Fields

You can also use combinations of fields for data checking. This is sometimes actually necessary because you have to look at all the fields together to tell if one or more of the fields are incorrect. If you do any of your banking online, you do this all the time without even realizing it. Your online bank record shows you several different fields that all have to make sense together, and if they don’t, red flags immediately go up in your mind. You have the source of the transaction, the amount of the transaction, the unit of currency that it’s in, if it’s a credit or a debit, the date the transaction occurred and the total balance of your account afterwards. All of these items are part of a dataset, and you’re doing a check on that dataset every time you pull up your account online to make sure everything looks okay. If the amount of the transaction was different from what you were expecting or the total of your account after the transaction was off, you would mentally flag it and call the bank to see what was up.

It’s the same with any other dataset. There may be fields that have to make sense together. Imagine that you’re working on a medical study of hospital patients and you’re tracking the medications they take daily by using three separate fields for medication type, the amount of the quantity of medication being administered, and the unit of the medication. So, for example, if the dataset read, “Aspirin, 500, mg” that would mean the patient took 500 mg of aspirin each day. Now imagine that you received a record that said, “Morphine, 200, lbs.” What would your reaction be? It’s logical that a hospital patient would be taking morphine and 200mg is a reasonable dosage, so the number alone wouldn’t raise flags, but even 1lb of morphine would kill someone so there’s definitely a problem there. You would probably want to go back to the patient’s record or to whoever entered the data to get the correct units.

If any of these fields are free response, there are an infinite number of combinations that you can receive. As such, you should go through your dataset early and identify groups of variables like this that need to work together so you can check records periodically as they come in for problems. Again, since this can be a time-consuming process, you need to decide how critical the cleanliness of these particular fields is to your end result, be it a specific visualization, a statistical analysis, or a general report.

What Happens if We Don’t Clean Our Data?

As many of these cleaning procedures can be time-intensive, you will often have to decide which variables are worth cleaning, and which procedures you can justify using. But what if we just skip the data cleaning process altogether and leave the data “dirty”? The answer to that isn’t easy because it all depends how dirty your data are in the first place. At best, you’ll get lucky and your data will be minimally dirty and you won’t have any real impact on your end report. At worst, your results will be incorrect due to errors in your dataset that you could have potentially corrected if you had gone through data cleaning procedures.

Of course, your data may be in relatively good shape to begin with. If that’s the case, you might be able to ignore the cleaning process with little impact on your end product. However, until you at least go through the basic checks involved in data cleaning, there’s no real way for you to know how clean or dirty your data are. That said...

Accept That Most Datasets are Never 100% Clean

Data cleaning is just like house cleaning—you won't ever catch everything. As hard as you may try, you can't force people to input 100% correct data, and we make errors ourselves as we work with data. You want your data to be as accurate as possible, but there will always be a little dust in the corners so it's important to accept that datasets are never perfect and to develop a sense for what is "good enough" for your purposes.

For example, you have a list of 1,000 contacts from your database of 100,000 contacts and you notice that 2 of the 1,000 have the first and last names together in one field. Do you take on a project of combing through and correcting all 100,000 records?

It depends.

You may make a formal decision that the data are either clean enough for your purposes or too dirty for what you want to accomplish. This really depends on the variable in question and what your intended end output is. If you're responsible for checking that students correctly reported their majors for an internal report and you think that there was a 0.01% error rate, that's probably of a much lower concern than if you're checking a variable that is of critical importance to a safety report and you think there's a possibility of a 5% error rate. Over time, you'll get a better sense of how clean or dirty a dataset is, relatively speaking, and how labor-intensive the cleaning process will be for a particular group of variables. At that point, it's good to consult the key stakeholders for the end product to see how much cleaning they agree is sensible to pursue. You should always aim to have your data as clean as possible but always remember that it won't be 100% perfect.

Data preparation and cleaning have costs. If you hire someone to do this work for you, the cost is financial. If you're going to do it yourself, it costs you or someone on your team time (and maybe a little sanity). So if you'll never use a phone or fax number or need to refer to someone as Reverend, you may make the decision to delete those variables, stop collecting them, or just not worry about cleaning them in the future.

After Data Cleaning: Please Be Kind and Document!

Once we've cleaned our data, we're left with a brand new problem: how can we (and others!) verify that what we've done is correct and that we haven't corrupted the data by making these changes? After all, the processed data may look vastly different from the raw data we started out with.

The simple answer is to document everything, particularly if you think you might want to share your data later on with a statistician or other researchers. When you're cleaning your data, it's always a good idea to save any changes as an entirely separate file: this way you're always able to go back and look at what changed between the raw and processed data, what rows and columns were dropped, etc. It also ensures that you can go back to the unprocessed data if you ever want to slice things up a different way that might involve different cleaning procedures.

You should be careful to write a set of instructions as you go, documenting exactly what was done in each step to identify bad data and which data points were removed. It's crucial to write this while you're actually cleaning your data: it's always easier to document as you go than it is to try and remember every step that you took after all is said and done. If you're using point-and-click software to manage your data (like Excel), you should take special care to record exactly what steps were taken in cleaning the data since everything is done by hand, rather than by computer code that can be easily re-run later on. A good rule of thumb is that if you aren't able to easily follow the instructions you wrote and end up with the same results a second time, you shouldn't expect anyone else to be able to.

Regardless of how you choose to do it, good documentation of your cleaning procedures ensures that you can always justify why certain data points were removed and others weren't and that others are able to verify that the data were cleaned competently and correctly. Additionally, if you think you might want to share the data with a statistician later on for further analysis, being able to see both the raw data and what operations were done on those data will make the statistician's job much easier and quicker.

CHAPTER 9

TYPES OF DATA CHECKS

BY IAN TEBBUTT

In the last chapter, we looked at data cleaning and the checking processes that are necessary to make that happen. Here, we'll take a more in-depth look at data checking and talk about other validation processes, both before and after cleaning occurs.

Data checking is crucial if you and your audience are going to have confidence in its insights. The basic approach is quite straightforward: you have fields of data and each of those fields will have expected values. For instance, an age should be between 0 and 120 years (and in many cases will be less than 80 years). Transaction dates should be in the recent past, often within the last year or two, especially if you're dealing with an internet-only data source, such as a Twitter stream.

However, data checking, although easy to understand and important to do, is a complex problem to solve because there are many ways data can be wrong or in a different format than we expect.

When to Check

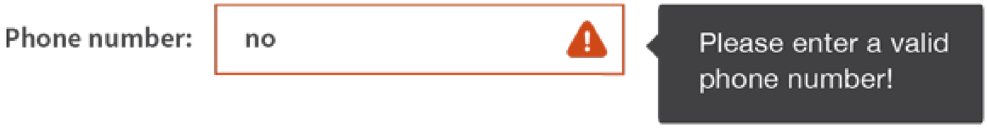
Consider this dataset from a telecom company with information about customers who are changing phone numbers. Here, they provided the database but didn't check the data which were aggregated from hundreds of smaller phone service providers. The database is still in daily use and is a great example of why checking is important. Imagine you're tracking the types of phone charges by age. The example below shows a few of the issues.

- PhoneNumberType mixes codes and text
- Age field has 0 meaning unknown—not good for averages, and is 112 a genuine age?

RecordId	PhoneNumberType	Age	New Customer	Price
1	MOBILE	0	NO	$12.45
2	Mobile	47	Y	12 45
3	Land Line	34	YES	37
4	LandLine	23	YES	1.00
5	LL	112	Y	$1K

The basic rule for data checking is check early, check often. By checking early, when data are first entered, there is a chance to immediately correct those data. For example, if your "New Customer" field expects the values YES or NO, but the user enters something different, such as an A or a space, then the user can be prompted to correct the data. If the validation isn't done until later then incorrect values will reach the database; you'll know they're wrong but will be unable to fix the issue without returning to the user to ask for the information. Occasionally, it's possible to compare incorrect fields with other linked datasets and then use that information to fix the original data. That can be complex and lead to further issues, since you have to decide which data source is correct.

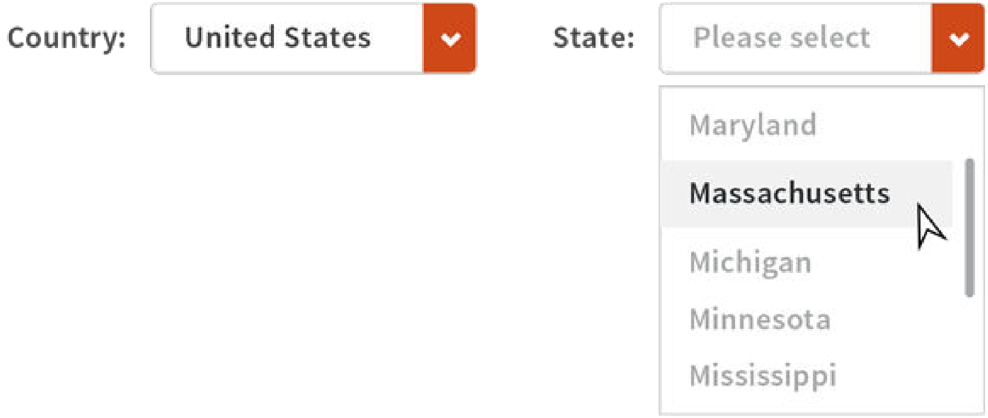
If you're in the happy position of controlling how the data are gathered, you have a great advantage, as one of the easiest forms of checking is done as soon as the data are entered. This type of data check is called a front-end check or a client-side check because it happens at the moment that the user enters the data, before the data are submitted to the database. This is most commonly done by making sure that your data collection application or web page is designed to only accept valid types of input. You have probably encountered this type of data validation yourself when filling out forms on the web before.



Phone number: no

Please enter a valid phone number!

For example, states and countries should be selected from a list and if you're dealing with international data, the choice of country should limit which state choices are available.



Country: United States

State: Please select

In this way your system will be constrained to only allow good data as they are entered. The approach isn't perfect though. A classic internet speed bump is data entry that waits until submission before letting on there was an issue in a field at the beginning of the form. A better approach is to check every field as it is entered, but that has other disadvantages as it can be harder to code and can result in a continual stream of checking requests being sent to the server and potential error messages being returned to the user. As a compromise, some simple checking and validation can be carried out entirely in the browser while leaving the more complicated work for server-side processing. This will be necessary as some checking will require data or processes that are only available on the server itself. This often occurs when secure values—credit card verification, for instance—are being checked.

Other good survey design policies to consider to minimize data preparation time include:

- Decide how you want names to be input in advance. Is it okay for people to add things like Jr, DVM, PhD, CPA after their names or do you want these to be stored separately from the name itself? If so, do you want professional designations to be in a different field than suffixes like Jr, II, III? Do you want first and last names to be separate fields?

- Set up forms so that phone numbers and dates can only be input the way you want them to be stored (more about dates below). Determine if you want to collect office extensions for office phone numbers. If so, set up a separate extension field.

Trust Nobody

No matter how well you have designed your form and how much validation you have put into your front-end checks, an important rule of thumb is never trust user data. If it has been entered by a person, then somewhere along the line there will be mistakes. Even if there are client side checks, there should always be server side or back-end checks, too—these are the checks that happen after the data are submitted. There are many good reasons for this. You might not have designed the data gathering tools and if not, you could have different front end applications providing data. While some may have excellent client side checks, others might not. Unclean or unchecked data may arrive in your system through integration with other data services or applications. The example telecom database had too many owners with little oversight between them, resulting in a messy dataset. A small amount of extra effort up front saves us time and frustration down the road by giving us a cleaner dataset.

A second golden rule is to only use text fields where necessary. For instance, in some countries it's normal to gather address data as multiple fields, such as Line1, Line2, City, State, Country, postcode, but in the UK it's very common to just ask for the postcode and the street number as those two pieces of information can be then be used to find the exact address. In this way the postcode is validated automatically and the address data are clean since they aren't not entered by the user. In other countries, we have to use text fields, and in that case lots of checking should occur.

Commas in data can cause all kinds of problems as many data files are in comma separated (CSV) format. An extra comma creates an extra unexpected field and any subsequent fields will be moved to the right. For this reason alone it's good to not cut/paste data from an application; instead save to a file and then read into your next application.

Title	Name	FamilyName	Address1	Address2	Town	State	Country	
Bill	Short		13	A The Street	Hastings	VIC	AUS	
	Mr	William	Tall	27 The Close		Guildford	VIC	AUS

An additional comma in the second record has pushed the data to the right. This is an easy problem for a human to spot, but will upset most automated systems. A good check is to look for extra data beyond where the last field (“country”) would be. This example emphasizes the importance of combining computerized and manual approaches to data checking. Sometimes a quick visual scan of your data can make all the difference!

Data Formats and Checking

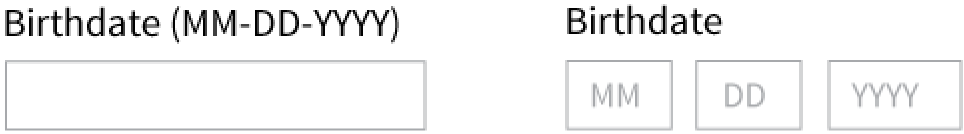
When dealing with numbers there are many issues you should check for. Do the numbers make sense? If you’re handling monetary figures, is the price really $1,000,000 or has someone entered an incorrect value? Similarly, if the price is zero or negative, does that mean the product was given away or was someone paid to remove it? For accounting reasons many data sources are required to record negative prices in order to properly balance the books.

In addition, checking numeric values for spaces and letters is useful, but currency and negative values can make that hard as your data may look as follows. All of these are different and valid ways of showing a currency, and contain non-numeric character values.

```
$-1,123.45
(1123.45)
-US$1123.45
-112345E-02
```

Letters in numbers aren't necessarily wrong, and negative values can be formatted in a variety of ways.

Dates also exhibit many issues that we have to check for. The first is the problem of differences in international formatting. If you see the date 1/12/2013, that's January 12, 2013 in America, but in the UK it's December 1. If you're lucky, you'll receive dates in an international format such as 2014-01-12. As a bonus, dates in this standardized format (http://whatis.techtarget.com/definition/ISO-date-format) can be sorted even if they're stored as text. However, you might not be lucky, so it's important to check and make sure you know what dates your dates are really supposed to be, particularly if you're receiving data from respondents in multiple countries with different date formats. A good way to handle this if you are designing the data entry form is to make the date field a calendar button field, where the user selects the date off a calendar instead of entering it manually. Alternatively, you can specify the format next to the entry box as a sort of instruction for the user.



Another checking task that you may encounter is the analysis and validation of others' work to make sure the visualizations and numbers actually make sense. This can happen in a work situation where you need to proof other people's work of others or online where public visualizations will sometimes provide the underlying data so you can try your own analysis. In both instances the first check is to just recalculate any totals. After that, look at the visualization with a critical eye: do the figures make sense, do they support the story or contradict it? Checking doesn't have to be just for errors. It can be for understanding, too. This will give you good experience when moving on to your own data checking and is the first thing to try when someone passes you a report.

Data Versions

Another big source of data checking problems is the version of the data you're dealing with.

As applications and systems change over the years, fields will be added, removed, and—most problematic—their purpose will be changed. For instance the Australian postcode is 4 digit and is stored in a 4 character field. Other systems have been changed to use a more accurate 5 digit identifier called the SLA. When data from those systems are combined, we often see the 5 digit values chopped down to fit into a postcode field. Checking fields for these kinds of changes can be hard: for postcodes and SLAs, the lookup tables are in the public domain, but it takes additional investigation to realize why a location field with 4 digits matches values from neither table.

You should consider collecting additional fields which won't be part of the actual visualization or final report but will give you important info about your records, like when they were created. If new fields are added after the dataset is created, any existing records won't have that field filled and if someone is improperly handling the dataset, the older records may have new numeric fields filled with zeroes. This will throw off averages and the effect on your visualizations would be huge. If you have the record creation date, you can go through and change the incorrectly added zeroes to a missing value to save your data. For those fields that have been removed, a similar issue might be seen. It's relatively rare for unused fields to be removed from data but they can sometimes be repurposed, so figuring out the meaning of a specific piece of data can be challenging if the functional use of a field has changed over time.

Amount	PaymentType	ServerId	CreatedOn
$100	CC		
$143	Cash		
$27	Amex	237	3/1/2013
$45	Cash	467	3/1/2013

Here you can see the effect of adding two new fields, ServerId and CreatedOn, to an existing data source. It’s likely that change was put into place 03/01/2013 (March 1, 2013), so if your analysis is only looking at data since that date then you can track sales/server. However, there’s no data before that in this source, so if you want to look at what was happening on January 1, 2013, you need to find additional data elsewhere.

One of the most important checking issues is that the meaning of fields and the values in them may change over time. In a perfect world, every change would be well-documented so you would know exactly what the data means. The reality is that these changes are rarely fully documented. The next best way of knowing what the data in a field really represents is to talk the administrators and users of the system.

These are just some of the steps that you can take to make sure you understand your data and that you’re aware of potential errors. In the next chapter, we’ll talk about some of the other sneaky errors that may be lurking in your data, and how to make sense of those potential errors.

CHAPTER 10

WHAT DATA CLEANING CAN AND CAN'T CATCH

BY DYANNA GREGORY

Now that we understand what data cleaning is for and what methods and approaches there are to shape up our dataset, there is still the question of what cleaning can and can't catch.

A general rule for cleaning a dataset where each column is a variable and the rows represent the records is:

- if the number of incorrect or missing values in a row is greater than the number of correct values, it is recommended to exclude that row.
- if the number of incorrect or missing values in a column is greater than the number of correct values in that column, it is recommended to exclude that column.

It should be made clear that exclusion is not the same as deletion! If you decide that you don't want to include a row or column in your analysis or visualization, you should set them aside in a separate dataset rather than deleting them altogether. Once data are deleted, you can't retrieve them any longer, even if you realize later on that there was a way to fill in the missing values. Unless you are absolutely certain that you will not use a record or variable again, do not just delete it.

In the last few chapters, we have talked about several different processes for data cleaning and have seen the types of problems they can help identify and fix. When we're searching for errors and mistakes, we are able to detect potential problems such as:

- inconsistent labels, misspellings, and errors in punctuation;
- outliers, invalid values, and extreme values;
- data that aren't internally consistent within the dataset (e.g. 200 lbs. of morphine);
- lack or excess of data;
- odd patterns in distributions; and

- missing values.

What we haven’t talked a lot about yet is what data cleaning can’t catch. There may be incorrect values that are nevertheless both within the acceptable range for the data and that make complete sense. For example, if someone enters the number 45 instead of 54 into your dataset and your valid range of numbers is 0-100, it will be unlikely that you’ll catch that error unless that field is one that you’re cross-checking with another field or you’re verifying the information with an outside source record.

Similar to that, you may be receiving information from an online survey form and the person filling it out may have selected the button for “Strongly Agree” when they actually meant to select “Strongly Disagree.” Again, unless this answer is somehow cross-checked with another variable or source, you will have no easy way to detect this error. Sometimes this type of error is more critical than others. If a person selects “Strongly Agree” instead of “Agree” on an opinion survey, that is unlikely to have the same impact on the results as if someone accidentally marks the wrong gender on a form for a research study where you are using gender as a grouping category for treatment assignments.

Data cleaning also can’t tell if a missing value is truly missing (i.e. the question was accidentally skipped or the data were not collected for some reason) or the question was purposely skipped (i.e. the participant declined to answer) unless “Prefer not to answer” was an answer choice. This may be relevant in some cases (particularly in demographics), though in others, you may decide to just treat both as missing data. This is why, as mentioned before, you need to include a “Prefer not to answer” choice for any question of a personal nature where you want to know if the data are truly missing, since some people may actively choose to not answer questions about race/ethnicity, income, political affiliation, sexual orientation, etc.

CHAPTER 11

DATA TRANSFORMATIONS

BY KIRAN PV

This chapter covers more advanced statistical concepts than some of the others but we wanted to include a brief introduction to data transformations in case you encounter them. If you need to do your own transformation, check out the resources in our Appendix for additional tips.

When you take a digital photo, sometimes the picture comes out in a way that makes certain features hard to see. Maybe the colors are too dark or too light, the photo is blurry, or the objects in the image are too small and you want to zoom in on them. You can pop the photo into something like Instagram or Photoshop, tweak it, apply a filter or two, and transform the picture so it's much clearer and easier to understand.

Sometimes we need to transform data, too. You might get a set of data where, if you visualize it as is, there will be parts that are difficult to see. Also, if you're going to do statistical testing of your data, many common tests make specific assumptions about the distribution of the data (e.g. that the data are normally distributed). In the real world, we often come across data that don't meet these assumptions. Data transformations are one way to handle both of these problems. Here, we'll talk about some of the more common transformations so that when you encounter these terms in the future you'll have an idea what was done to the data.

Data transformations are one of the common manipulation procedures which can reveal features hidden in the data that are not observable in their original form. We can transform the distribution of data to make it easier to see and so that any required assumptions of statistical tests are met. We usually do this by replacing one variable with a mathematical function operating on that variable. For example, you could replace a variable *x* by the logarithm of *x* or by square root of *x*.

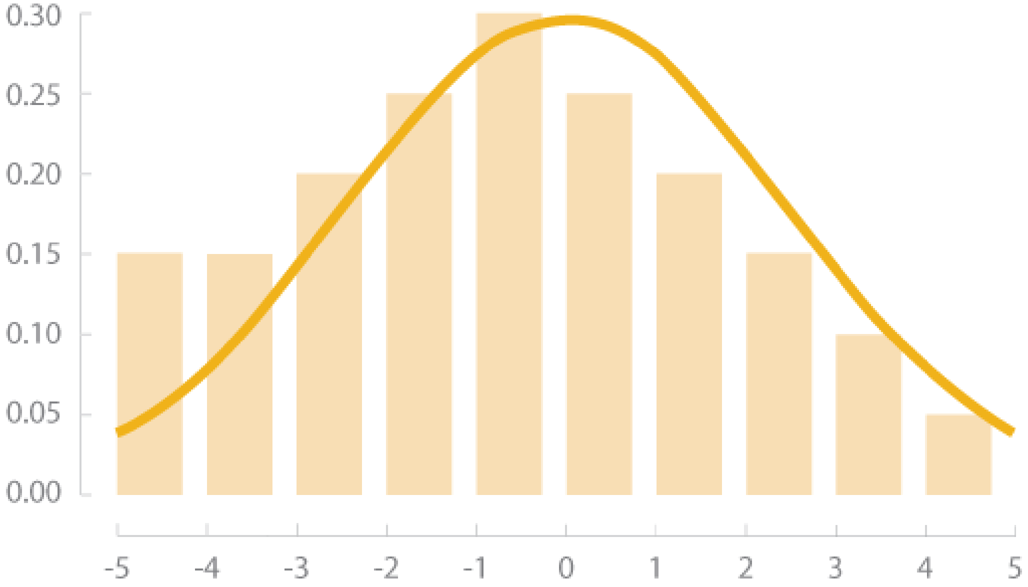
Never perform the transform directly on your original data! Either create an additional column to hold the new values for each variable you're transforming or make a copy of your entire dataset.

Normal Distribution and Skewness in Data

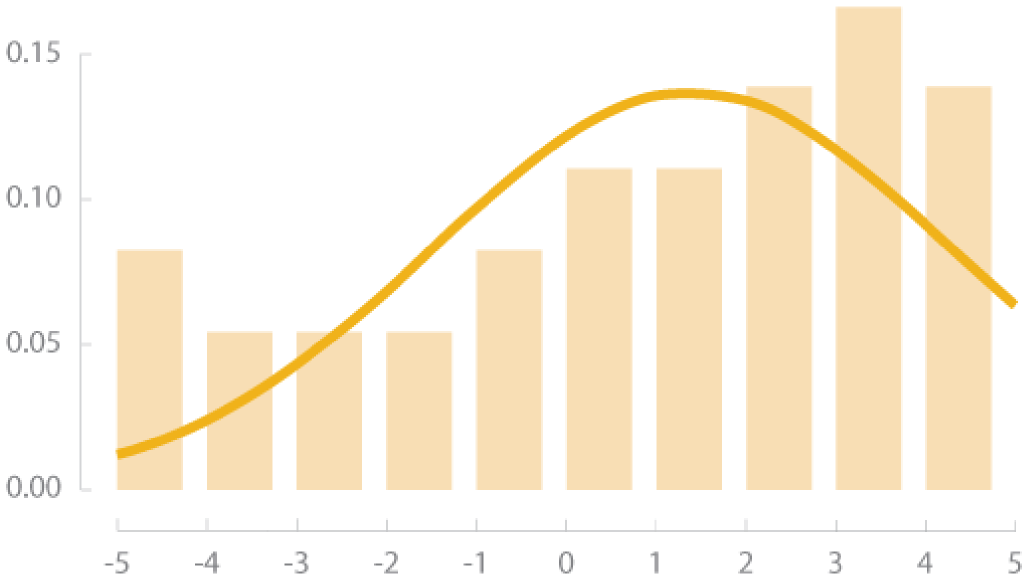
One of the most frequently-encountered assumptions of statistical tests is that data should be normally distributed. You may have heard of the normal distribution referred to as a "bell curve" before; this is because a normal distribution takes the shape of a bell, with the data spread around a central value. Some of the data examples that commonly follow a normal distribution are related to human measurements such as height, weight, life span, and scores on IQ tests.

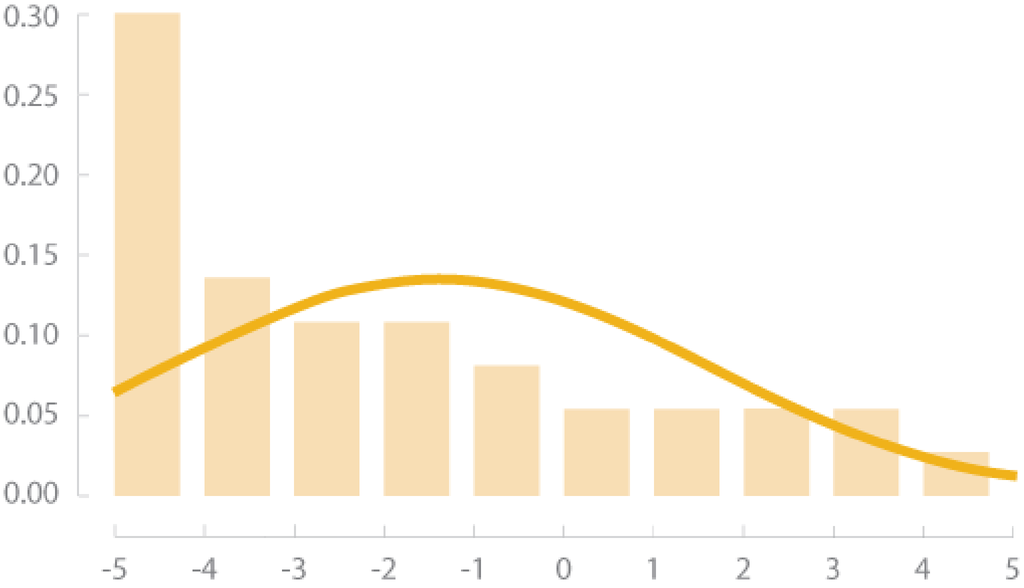
Unlike a normal distribution, which is symmetric around the mean value, skewed data tend to have more observations either to left side or to right side. Right skewed data have a long tail that extends to right whereas left skewed data will have a long tail extending to the left of the mean value. When data are *very* skewed, it can be hard to see the extreme values in a visualization. If you notice that your data distribution is skewed, you might consider transforming it if you're doing statistical testing or if the data are difficult to visualize in their original state.

NORMAL DISTRIBUTION



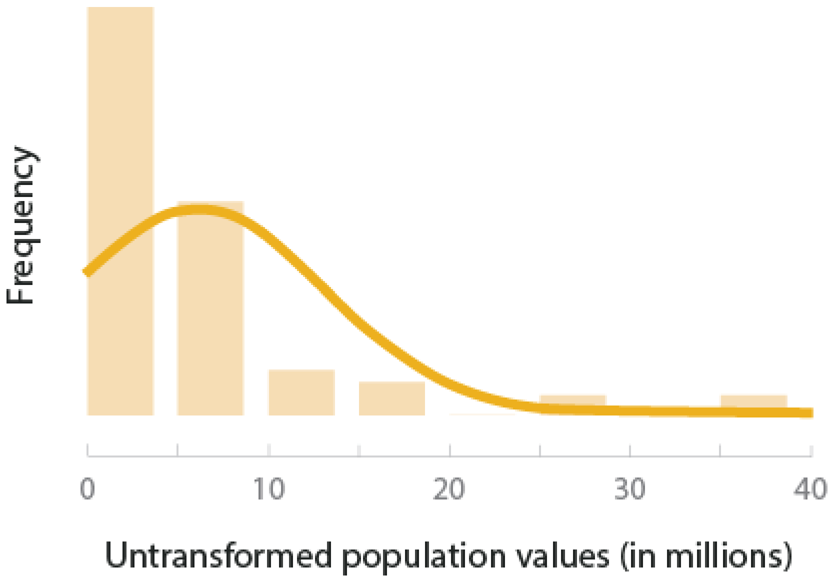
LEFT SKEW



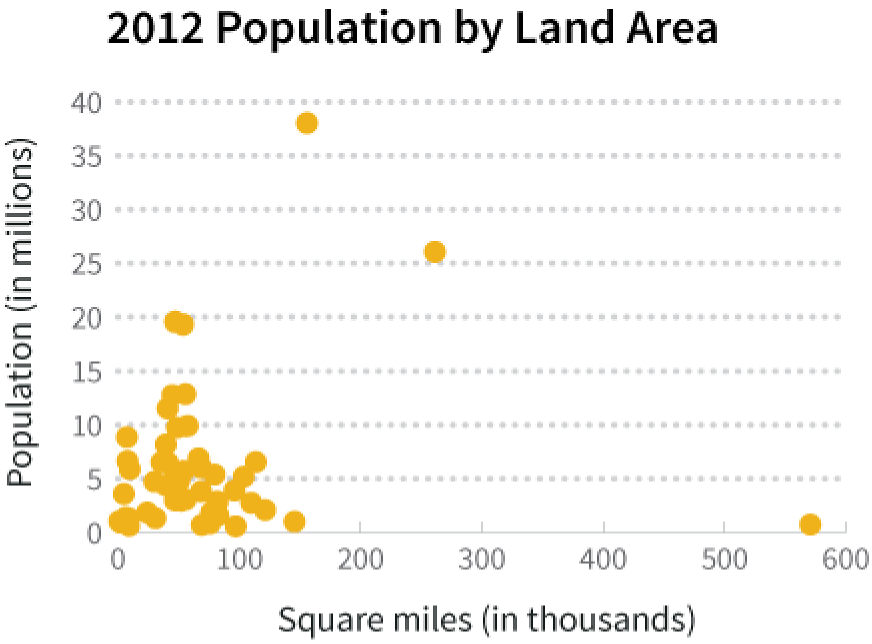

RIGHT SKEW

Understanding Transformations Using Sample Data

Let's use the population and land area of the 50 US states from 2012 to see how transformations work on actual data. The first step in transformation is to evaluate the distribution of the data. Then you can decide what transformation is appropriate (if one is needed). We can start by constructing a histogram of the population data and a scatterplot of the population-area data to get a better sense of how they're distributed.



The histogram above shows that the distribution of population values is right skewed. This is reasonable to expect because the majority of states' populations lie in the range of 1-10 million. If we want to do statistical testing that relies on a normal distribution assumption, these data will need to be transformed.

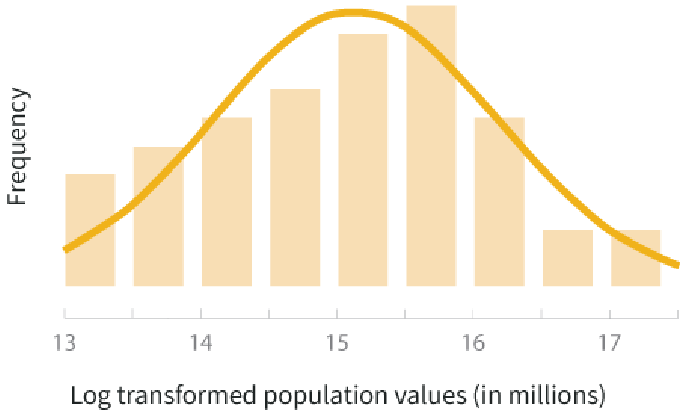


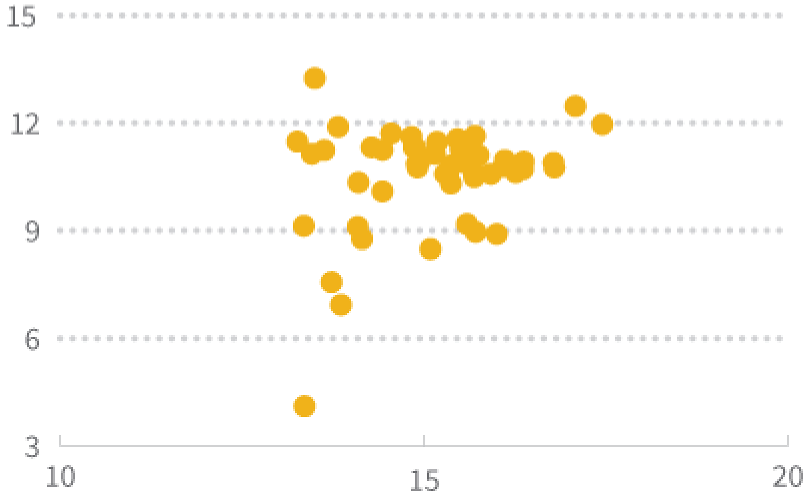
In the scatter plot above, you can see that most of the data points are clustered in the bottom left corner of the graph, making it hard to see how population and land area are related. We can't just scale the graph differently to "zoom in" on that corner because we'd knock California and Alaska off the chart. We can, however, use transformations to help make the data easier to view.

There are many transformation methods that can be applied in either of these situations, but let's look at a couple of the common ones to see how they can affect both a visualization and the shape of a distribution.

LOG TRANSFORM

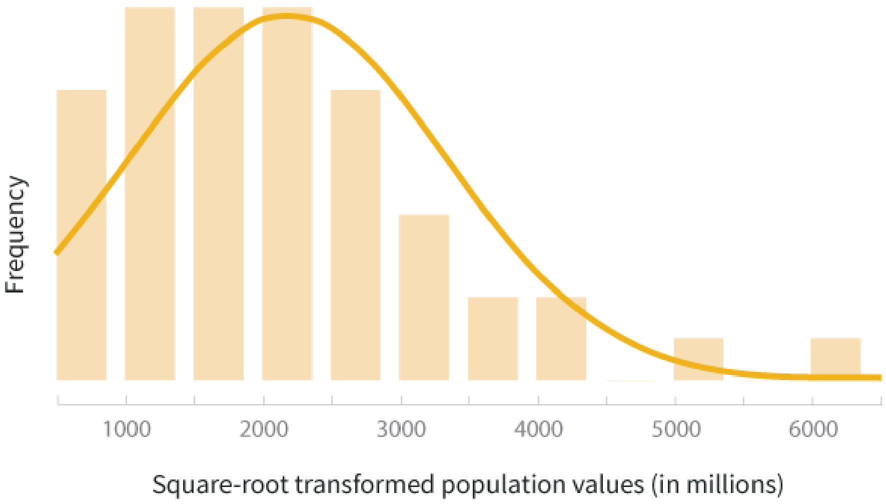
To do a logarithmic transformation, you calculate the log of each value in the dataset and use those transformed values rather than your raw data. Log transforms tend to have a major effect on distribution shape, and in visualizations can bring extreme outliers closer to the rest of the data so graphs aren't stretched out as much. You can either use natural logs (*ln*) or logs with base 10. The graphs below show the histogram of population data after a natural log transformation is applied and what the scatterplot looks like if you use a natural log transformation on both the population and land area variables.

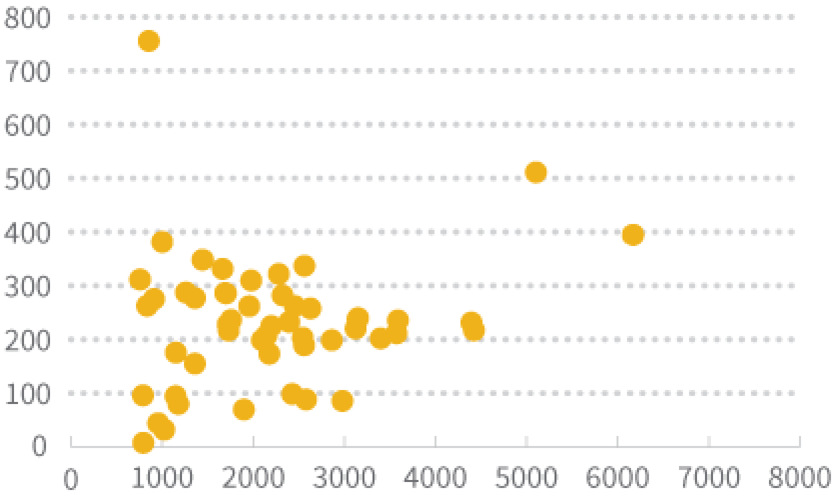




SQUARE ROOT TRANSFORM

The square root transformation uses the square root of each value instead of the log, and has a more moderate effect on the distribution shape. The two graphs below show the histogram of population data and the scatterplot of population by land area, both after square root transformation is applied .





Choosing the Right Transform

As you develop a better understanding of different transformation methods, you might wonder how to pick between them. The answer to this question is not straightforward and although there are formal statistical methods for selecting a transformation, we often need to use trial-and-error combined with knowledge of different transformations. A general strategy is to apply some of the most frequently used transforms such as log, square root, square, reciprocal, and cube root, and then choose the best one after observing the results.

Looking at the transformed histograms above, the log transformed data seems to be a better fit to the normal distribution while the square root transformed data still carries the right skew. In this example, if you're doing a statistical test that has assumes the data are normally distributed, the log transformation would be a better method to use than the square root transformation.

On the other hand, if your primary purpose in the example above is to visualize the relationship between state population and land area, the square root transformation does a better job of spreading out the data and making it easier to view than the log transformation.

Common Transformations

Method	Math Operation	Good for:	Bad for:
Log	ln(x) $\log_{10}(x)$	Right skewed data $\log_{10}(x)$ is especially good at handling higher order powers of 10 (e.g. 1000, 100000)	Zero values Negative values
Square root	$\sqrt{x}$	Right skewed data	Negative values
Square	x^2	Left skewed data	Negative values
Cube root	$x^{1/3}$	Right skewed data Negative values	Not as effective at normalizing as log transform
Reciprocal	1/x	Making small values bigger and big values smaller	Zero values Negative values

Caveats about Transformation

Since data transformation methods involve the application of a mathematical function to your data, you need to be careful when reporting or interpreting any insights derived from the transformed data because a transformation changes the unit of the data. For example, when we apply a logarithmic function to a population variable, the unit of measurement becomes the log of the population. When you're sharing results, your audience may assume that the calculated numbers or visualizations they're seeing are based on raw data, so if the values have been transformed, you should clearly communicate what transformation was used, as well as what units the data are being displayed in.

If you use transformed data to calculate statistical values like means, you should back-transform the final results and report them in their original units. To back-transform, you just do the *opposite* of the mathematical function you used in the first place. For example, if you did a square root transformation, you would back-transform by squaring your end result.

You may not see transforms every day, but when you do, it's helpful to know why they were used and how they affect your data. It's important to be able to see different parts of the picture when working with data, and transformations give you another tool to help you do just that!

VISUALIZING DATA

You finally have everything collected, prepped, and cleaned. Now it's time to cook! When you're combining ingredients, some will stand out in your dish, some will play a supporting role, and some will seemingly fade away. Despite these differences, you still need to make sure each ingredient is incorporated properly in order for the dish to succeed as a whole.

You also need to think about how to present the dish in terms of emphasis, functionality, and appropriateness. For example, if you want to highlight the veggies in a vegetable soup, you may want to chop them instead of puree them. And you have to find bowls because you're not going to serve this delicious soup on a flat plate, right?

Similarly, deciding which graph types and presentations to use depends on the data behind your visualization and the narrative it supports. Some variables will stand out in your visualization, while others disappear into the calculations for the end result. And then there are other elements to consider - like fonts, colors, and icons.

In this section, we'll talk about crafting visualizations that help you best tell the story of your data. We'll cover tips on choosing what to visualize, deciding which graph types make sense, and giving the right finishing touches to a beautiful and accurate presentation that your audience will be excited to consume.

CHAPTER 12

DECIDING WHICH AND HOW MUCH DATA TO ILLUSTRATE

BY MARGIE HENRY

Let’s lay some groundwork for successful data presentation. If done thoughtfully, it will go a long way in helping you determine which aspects of your data to visualize and how. We’ll begin with a little brainstorming. You can do this in just a few thoughtful moments alone or working as a team. Your work here is two-fold: define your message and define your intended audience. You can flip this sequence around, but we’ll begin with defining your message.

Determine Your Message

Before tackling which data to present, take a few minutes to decide what you want to say. Close down your latest social media craze, step back from your computer, and consider the exact message you want to communicate. Ask yourself, “What do I know, what does it mean, and why do I believe it’s important?”

Consider a dataset containing observations on different types of caffeinated beverages and the effects of their consumption. Don’t stop at “caffeine affects the body.” You never want to present information that solicits a “well, duh” response. Dig deeper. Be more specific. What do your data say about how caffeine affects the body? Are the effects all good, all bad, or maybe an interesting combination of both? Do the effects change with a person’s age and/or sex? Are some caffeinated beverages better or worse for overall health? Your answer should be concise: short, sweet, and to the point. A statement such as “Coffee has an ability to reduce the risk of certain diseases and ailments when consumed in moderation because it contains key antioxidants.” goes a lot further than our original example. Even better, it establishes a pretty clear focus for our visuals and some common language to use with our audience.

Right about now you should be having a flashback to English 101. That’s because determining your key message is just like writing a good thesis statement. If you can’t summarize your key message in a few concise sentences then you probably need a better grasp of the topic. Sound harsh? Maybe, but not as harsh as presenting information to a crowd of your yawning disinterested peers. Fight the urge to skip this step! If you're the paper-and-pencil type, go ahead and write your message down! You can use it as a reference throughout your data visualization process.

Simply put, your chances of creating a compelling, well-organized visual argument are immeasurably greater if you begin with a clear and focused message.

Understand Your Audience

You've determined your message. Let's now consider the importance of understanding your audience. This knowledge will go just as far in helping you determine which and how much of your data to illustrate.

Take another couple of minutes and ask yourself "what information is most valuable to my audience," "what role will my visuals play in this dialogue," and "what action(s) do I want to incite?" Would you spend time explaining algebra to a group of engineers? (The correct answer is no.) What would be the point? The better you know your audience, the better your chances of creating a successful visual presentation.

Let's imagine presenting data on "Environmental Conservation in the New Millennium" in the following scenarios: (1) on a small-scale blog visited mostly by lay environmentalists; (2) in a classroom of high school students; and (3) at a fundraising event for an environmental conservation organization. Would you create and explain your data the same way to each audience? Hopefully not. You should be able to make a few assumptions about what's most relevant to present even if you've never met a single audience member.

In our first scenario, we can assume visitors are already interested in conservation. They may have spent time doing actual research. A portion are return visitors who may rely on your specific perspective; they might see you as a content area expert. Your site is, most likely, not the only blog on which they rely, but one day it could be their favorite! At minimum, we can assume they've stumbled upon your blog intentionally, and not because of issues with autocomplete. In this instance, breadth and depth are key. You can take more time to explore, deconstruct and restructure the data. If the intention of your site is to incite further exploration, you can presents visuals that pose questions or make viewers question their own beliefs.

Our high school student scenario is a bit different. You can assume that your audience possesses very little familiarity with the topic. (Though, as always, some members will know more than others.) Attendance may be mandatory, not voluntary: keeping their interest will be key. You'll want to present fascinating, high-level, attention-grabbing visuals, that address immediate and pressing issues. Approach your vocabulary carefully: explain less-common terminology, and include more visual indicators of good/bad, positive/negative. Your visual display is intended to clearly present the importance of conservation, leaving little room for doubt.

At last, we have our fundraiser attendees. This audience needs to feel that environmental conservation is a cause worthy of their monetary support. It will likely be a mixed crowd: interested donors, their disinterested partners (who just came for free food and drinks), field experts, employees, and interns. You can assume they'll expect a balance of sentiment, the need for urgency, and solid fact. We've assumed the crowd is mixed, so you'll want to use language that is both familiar and easily understood while not appearing condescending. This audience expects to have their interest in the importance of conservation confirmed and your visuals should accommodate this. As with your student group, leave no obvious question unanswered.

> Presenting emotion-driven content doesn't mean leaving out key facts if they don't fit into your ideal storyline. Be extra careful when sharing cause-driven content, and do your best to ensure that your values don't interfere with an accurate presentation of the data!

Now that we've discussed the importance of determining a key message and understanding its audience, let's delve into deciding which data to illustrate.

Deciding Which Data to Illustrate

You can begin the process by expanding your key message into a narrative or story. Our goal is to present a sequence or set of facts which gradually leads your audience to the key message. The data you choose to illustrate should set the context, establish the main points of interest, and explain how these are interconnected. Be intentional in what you present, but do not censor data to further your argument. Your visual story should be based on what the data—and not only what you want to—say.

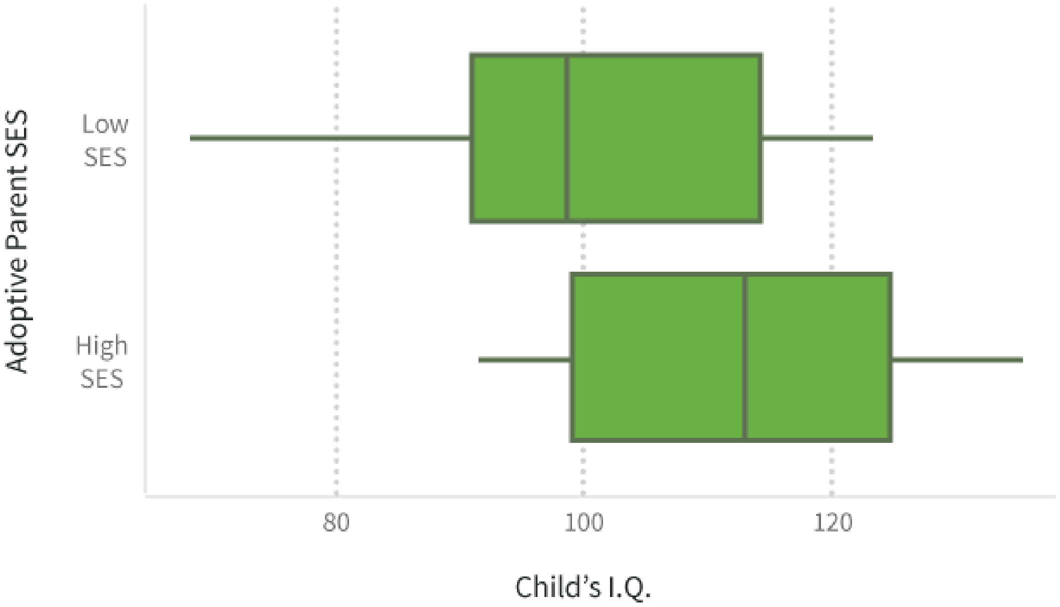
Take, for example, the following table presenting the I.Q. scores of children who were adopted at a young age and the socioeconomic status (based on income and occupation) of both their adoptive and birth parents. These data are taken from C. Capron and M. Duyme's 1989 study, "Children's IQs and SES of Biological and Adoptive Parents in a Balanced Cross-Fostering Study," published in the *European Bulletin of Cognitive Psychology*.

I.Q.	Adoptive Parent SES	Birth Parent SES
136	High	High
99	High	High
121	High	High
133	High	High
125	High	High
131	High	High
103	High	High
115	High	High
116	High	High
117	High	High
94	High	Low
103	High	Low
99	High	Low
125	High	Low
111	High	Low
93	High	Low
101	High	Low
94	High	Low

Let's discuss two possible narratives that you could create from this dataset: "Children's Intelligence Limited by Adoptive Parents' SES," and "Adopted Children's Intelligence Influenced by Both Biological And Adoptive Parents' SES".

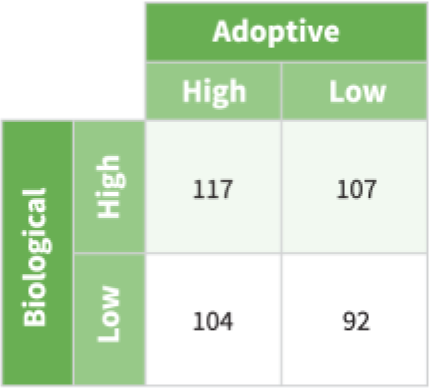
CHILDREN'S INTELLIGENCE LIMITED BY ADOPTIVE PARENTS' SES

We can create a story supporting the first message by solely looking at the adoptive parents' socioeconomic status: children of those adoptive families with a high SES had a mean I.Q. of nearly 112 whereas those adopted by a low SES family had a mean I.Q. of 99. But, this narrative would only include half of the relevant information: it leaves out entirely the SES of the child's biological parents. Understandably, this could play just as big a role as the family's socioeconomic status would likely impact the level and quality of prenatal care, and, in turn, the in utero development of the child.



ADOPTED CHILDREN'S INTELLIGENCE INFLUENCED BY BOTH BIOLOGICAL AND ADOPTIVE PARENTS' SES

A little more boring of a title, but far more accurate. When we include both the adoptive and biological parents' SES we get a much better picture of the impact that each has on the child's I.Q. Specifically, we see:



		Adoptive	
		High	Low
Biological	High	117	107
	Low	104	92

So, more correctly, a child's I.Q. is a function of both his or her biological and adoptive parents' socioeconomic status. If both have a high SES, the child's I.Q. will tend to be the highest. If one has a high SES and the other a low SES (it doesn't matter which set of parents has which), the child will typically have an average I.Q. And finally, if both have a low SES, the child will tend to have a below-average I.Q.

Our first example is a clear illustration of what happens when you create a story based on what you want to say, and not what the data say. Unfortunately, applications of data such as this are neither uncommon nor farfetched. We see this done on the news and during casual conversation. The omission of key facts and related variables creates a visual that is full of misinformation. It lacks credibility and presents obvious biases. The second instance presents far less outright bias, is a plausible story based on the data available, presents context, introduces all variables, and explains how the variables are connected. Although it will usually result in a less-sensationalized title, a full presentation of all relevant data is the only way to maintain a credible and airtight argument.

Deciding How Much Data to Illustrate

In previous sections we've gone over how to determine a key message, the importance of identifying the audience, and a process for isolating facts to illustrate. We can work on determining how much of our data we need to visualize.

If illustrating data is supposed to make information more digestible, then care should be taken not to present more than the audience expects, or more than they need to be able to understand your message. As you decide how much data to illustrate, keep in mind the idea that more is not always synonymous with better unless it's meaningful and presented in support of your key message. In most instances, your visuals will be taken as part of a narrative, contents in a storehouse, or maybe a combination of both.

As previously discussed, a narrative is a simply a story presenting a sequence of facts which gradually lead your audience to the key message. When you think of the narrative, think of written reports, PowerPoint presentations, and individual articles in newspapers and magazines or online. You want to illustrate just enough data for your audience to easily identify and understand your perspective without becoming exhausted. Each illustration should have a specific purpose. Avoid including visuals simply because they impress. As a test, try removing one or more illustrations or rearranging the presentation order. Does your narrative still make sense? Each illustration should stand alone, without too much verbal or written explanation, but if it doesn't add to the audience's understanding, it's probably not needed.

For the audience members wanting more, you can always provide links or references to additional takes on your data along with detailed commentary to contextualize and further explain the information. If you'd like to impress a technically savvy audience, a graphical appendix could be even be shared as a GitHub (https://github.com/) repository or a gallery of code gists hosted on bl.ocks.org.

A storehouse, on the other hand, can be thought of as an information repository. Usually consisting of multiple narratives and stand-alone content, this is an example of when more can be better. Unlike those of a narrative, storehouse visitors are less susceptible to data fatigue. They respond well to large quantities of data because they expect to spend time building or enhancing their understanding of a topic. The storehouse doesn't need to focus on presenting a single message. Its audience seeks new treatments of data, a diversity of perspectives, and various dissections of a topic or content area. In the storehouse scenario, the main criterion for deciding how much data to illustrate should be whether something will create redundancy. If you illustration fails to add something new to the mix or to expand on a topic, it can likely be omitted.

To exemplify, let's imagine a cinephile and store manager. Both are browsing a blog filled with upcoming movie release dates, reviews, and critiques of directors. The cinephile spends hours on the site, soaking up each and every visual and reading through its content. The manager simply wants to know what popular movies he should order for the next holiday season. The manager probably wouldn't want to spend hours trying to find his answer. For our cinephile, more is better; for the manager, less is more.

Editing and Revising

Here's a frequent and frustrating occurrence: you did your brainstorming, made a bunch of visualizations, and edited down to the best subset to include in your project. You were careful not to overwhelm your audience and you made sure that your illustrations covered the most important key points without being redundant.

How maddening, then, to field questions in a presentation, or see comments in a story or blog post, calling for the very visualizations that you left on the cutting room floor! You second-guess your calls, resist the urge to argue with the person asking the question, grit your teeth and grumble.

It's okay. If you do a good job and engage your audience, they will naturally be curious and want more information. They might want to see the same data presented in a different way, to dig down, or to zoom out. If these questions mirror the decisions you were making in your selection process, that's good news! It means you are on the same wavelength as your audience, and that they are involved and interested in the story your data tell.

There are several ways to keep (but de-emphasize) the visualizations that did not make the cut in your main collection. For slideshows, it is common practice to have a collection of extra slides after the "thank you" or conclusion slide that contain information that might be interesting but that won't fit within the time limit. "Yes, I do have that broken down by [industry sector/year/country/gender]," you say confidently as you flip to the prepared slide. Voila!

Another way to do this would be to publish interactive versions of your visualizations that allow the viewers to dive in and explore the information themselves. If you're able to share the raw datasets, that's even better! That way, those who wish to dig deeper and understand the data in new ways will have the option to do so. We'll talk more about static and interactive graphics later in the Print vs. Web chapter.

If you're looking for early feedback and you're not exactly sure where to turn, you can check out HelpMeViz (http://helpmeviz.com/), a community site where you can post your works-in-progress and receive friendly suggestions on how to improve. Getting feedback from your audience and revising your visuals to better fit their needs is all a part of the process!

CHAPTER 13

GRAPHING THE RESULTS OF CHECKBOX RESPONSES

BY ELLEN COOPER

This chapter focuses on checkbox responses or multiple response questions, where a question can be answered with more than one answer, if applicable.

Checkboxes Vs. Radio Buttons

Let’s say you’re doing a survey and you’re interested in what multimedia devices your respondents have used over the last six months. You would use a checkbox response question if you wanted to find out all of the multiple devices that people used over the six-month period. A radio button only allows respondents to select a single answer, so you could only use it to find out, for example, which one device a person used most often during that same six-month period. Each type of question has merit; which you should use just depends on the purpose of your question and how you are going to use the results.

What a Checkbox Question Really Is

So here’s the most important thing to know about checkbox questions, and it’s why you have to consider how you graph the results of checkbox questions differently than you do the results of other types of questions. Checkbox questions aren’t really their own question type! They’re actually just a shorthand way to write a *series* of yes/no questions. A respondent checks a box if an answer choice applies and leaves it blank if it doesn’t.

We have the checkbox format because it makes surveys more streamlined and easier to understand. In the example below, we asked, “Which of the following electronic devices have you used in the past 6 months? Please select all that apply.” The premise behind the question is that it’s likely that a respondent could use more than one electronic device over a 6-month period, such as a cell phone and a tablet.

If we were to pose this as a series of yes/no questions, it would read something like this:

In the last 6 months, have you used a/an:	
Desktop PC?	Y / N
Desktop Mac?	Y / N
iPad?	Y / N
Tablet (other than an iPad)?	Y / N
Laptop (Mac or PC)?	Y / N
Cell phone?	Y / N

With the checkbox question, survey respondents only need to check the answers that apply to them, while in a series of yes/no questions, they would need to respond to every question, even if all their answers were "No". With a checkbox question, you can simply provide a "None" option at the bottom of your choice list to handle this situation. When several yes/no questions are related, checkbox questions also prevent repetition of instructions, since all the questions are grouped into one.

These changes can help improve survey readability, flow, length, and overall response rates. However, if you want to handle the resulting data correctly, it is very important for you to remember that the underlying structure of a checkbox is actually a series of dichotmous questions.

How Checkbox Answers are Received

How your results or raw data are compiled will, of course, depend on the program you are using to design and distribute your survey. One of the more common formats is shown in the table below; this particular data structure reflects how a checkbox question serves as a quick way to represent a series of yes/no questions. A "1" is shown when a device was selected and a "0" if a device was not selected.

Date	Q1_PC	Q1_Mac	Q1_Tablet	Q1_iPad	Q1_Laptop	Q1_Cellphone	Q1_None
10/02/2013	1	0	1	0	0	1	0
10/01/2013	0	1	0	1	1	1	0
09/30/2013	1	0	0	1	0	1	0
09/30/2013	1	0	1	0	0	1	0
09/30/2013	0	1	0	0	1	1	0
09/30/2013	0	1	1	0	0	1	0
09/30/2013	1	0	0	0	1	1	0
09/27/2013	1	0	0	0	1	1	0
09/26/2013	1	0	0	0	1	1	0
09/26/2013	0	0	0	0	0	0	1
	6	3	3	2	5	9	1

You might also receive results like this:

Date	Response
10/02/2013	PC, Tablet, Cellphone
10/01/2013	Mac, iPad, Tablet, Cellphone
09/30/2013	PC, iPad, Cellphone
09/30/2013	PC, Tablet, Cellphone
09/30/2013	Mac, Laptop, Cellphone
09/30/2013	Mac, Tablet, Cellphone
09/30/2013	PC, Laptop, Cellphone
09/27/2013	PC, Laptop, Cellphone
09/26/2013	PC, Laptop, Cellphone
09/26/2013	None

Or like this:

Date	Q1_PC	Q1_Mac	Q1_Tablet	Q1_iPad	Q1_Laptop	Q1_Cellphone	Q1_None
10/02/2013	Q1_PC		Q1_Tablet			Q1_Cellphone	
10/01/2013		Q1_Mac		Q1_iPad	Q1_Laptop	Q1_Cellphone	
09/30/2013	Q1_PC			Q1_iPad		Q1_Cellphone	
09/30/2013	Q1_PC		Q1_Tablet			Q1_Cellphone	
09/30/2013		Q1_Mac			Q1_Laptop	Q1_Cellphone	
09/30/2013		Q1_Mac	Q1_Tablet			Q1_Cellphone	
09/30/2013	Q1_PC				Q1_Laptop	Q1_Cellphone	
09/27/2013	Q1_PC				Q1_Laptop	Q1_Cellphone	
09/26/2013	Q1_PC				Q1_Laptop	Q1_Cellphone	
09/26/2013							Q1_None

All three of the above examples represent the same answers, but they're formatted in different ways. Since different survey collection tools format checkbox responses in different ways, you may need to reformat your data to match the specific format required by the visualization software you are using.

Let's take a look at a summary of possible responses to the checkbox question posed above.

Table 1 Electronic Devices Used	Total
PC	421 (84%)
Mac (desktop)	300 (60%)
Tablet (any kind)	285 (57%)
iPad	185 (37%)
Laptop	200 (40%)
Cell phone (any kind)	450 (90%)

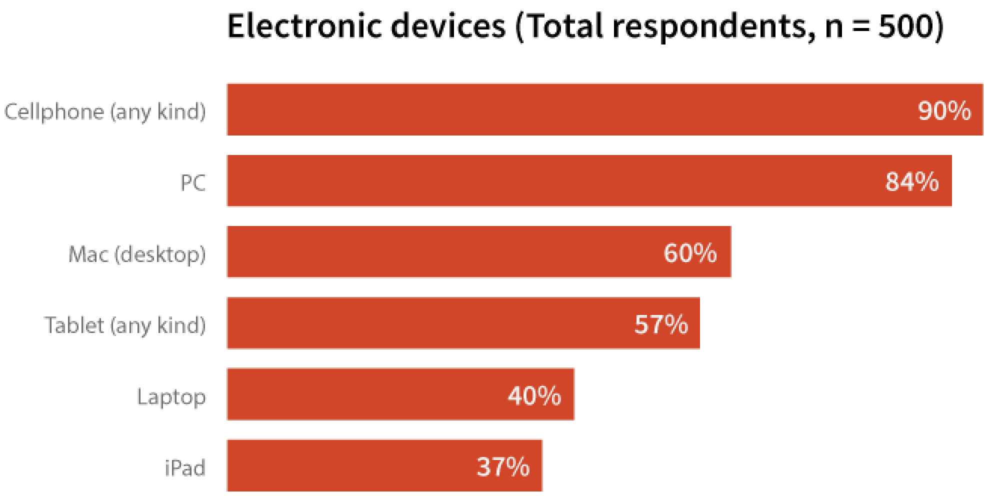
You may notice that the total number of responses (1,841) is greater than the number of people that did the survey (N=500)! Why? It's because of the whole "a checkbox is really a bunch of yes/no questions rolled into one" thing. The total possible number of checked boxes in a checkbox question? It's the (# of "real" answer options) X (# of respondents). (Here, a "real" answer option means one that isn't "None," "N/A" or "Prefer not to Answer," since selecting one of those options would prevent a person from choosing any other answers in additional to that.) For this survey, there were 6 device options (aside from "None") that a person could select and there were 500 people that answered the survey. So the total number of boxes that had the potential to get checked during the survey was 3000, not just 500.

Displaying Your Results

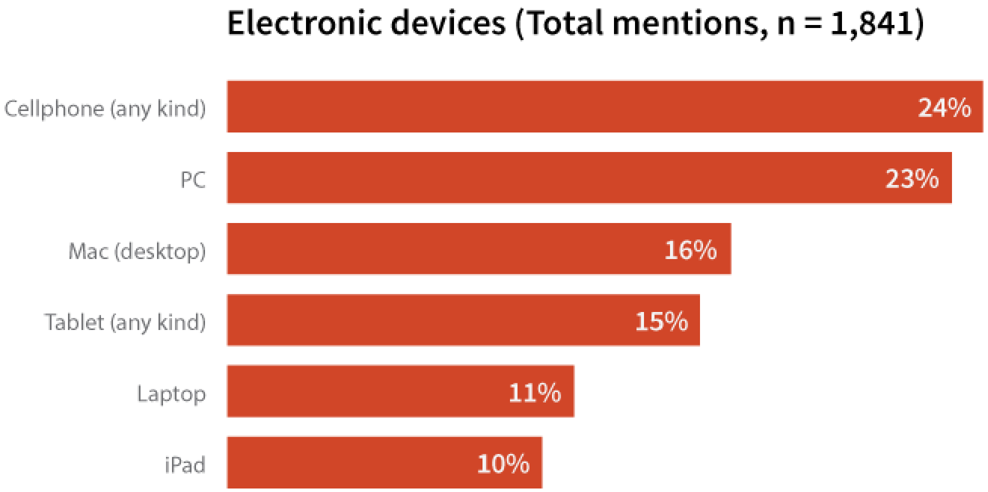
Since the total number of responses is greater than the number of respondents, you need to use some caution when creating graphs based on these data. There are a few different ways to illustrate your results, depending on what your overall question of interest is.

BAR CHARTS

One way is to construct a bar graph and base the percentages on the number of respondents that selected each answer choice, like in the graph below. Clearly, cellphones (90%) and PCs (84%) are the most commonly-cited electronic devices used in the past six months.



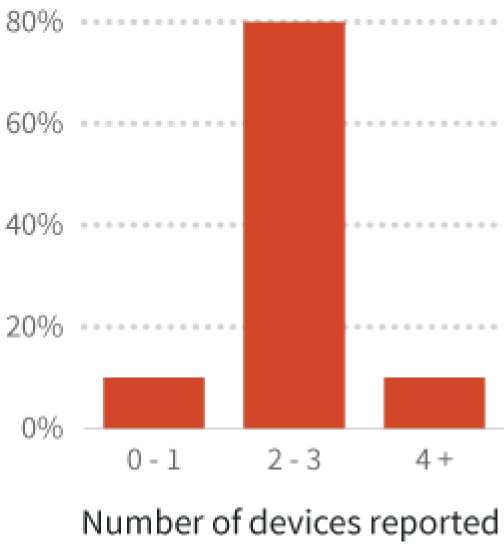
However, the fact that the total adds to more than 100% can be unsettling for some. An alternative way to illustrate the results is to base the percentages on the total mentions (1,841). Analyzing results based on mentions is useful when you want the percentages to total 100%.



Keep in mind that this way of displaying data is based on the number of mentions of a device, not the number of consumers who use that device. While you may be tempted to say, “24% of consumers used a cellphone in the past six months,” the bar chart above isn’t actually displaying that information.

Rather than the number of respondents (500), the chart shows the number of responses (1,841) to our survey on electronic devices. So, it would be better to say, “Based on all devices mentioned, cellphones were mentioned approximately one-quarter (24%) of the time, followed closely by PCs (23%).” This percentage represents the number of cellphone mentions (450) out of the total mentions (1,841) and accurately reflects the information on display.

Depending on your question of interest, you can also group your data. Maybe you’re more interested in reporting how many devices people used rather than exactly what devices were. You could make a column chart like the one below.

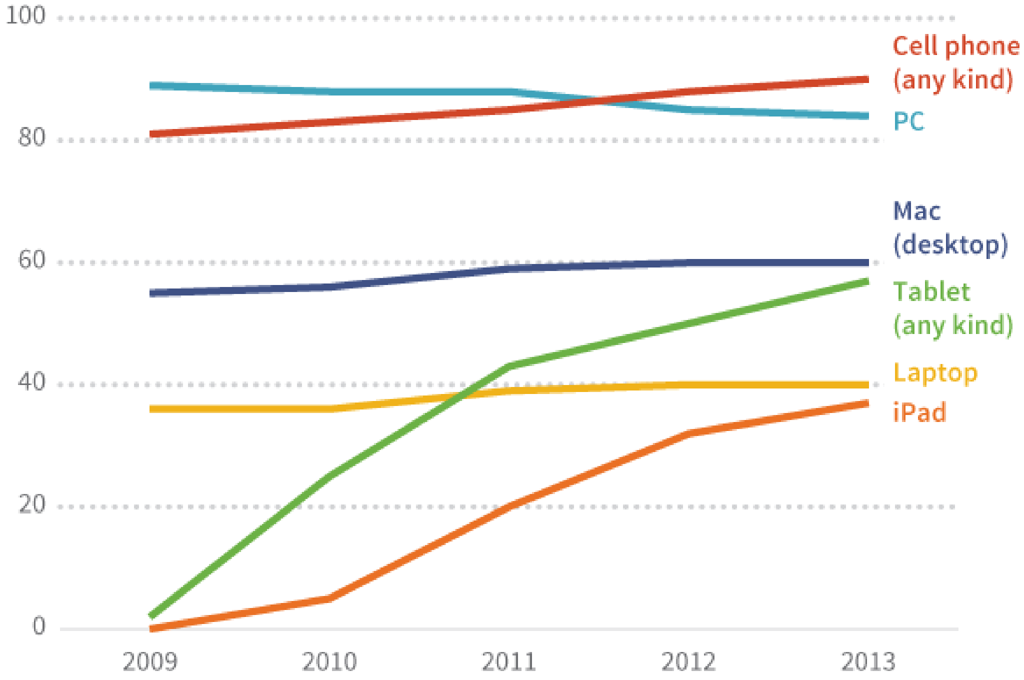


WARNING ABOUT PIE CHARTS AND CHECKBOX QUESTIONS

Don’t use pie charts if you’re basing your percentages on the number of respondents that selected each answer choice! Pie charts are used to represent part-to-whole relationships and the total percentage of all the groups has to equal 100%. Since the possible sum of the percentages is greater than 100% when you base these calculations on the number of respondents, pie charts are an incorrect choice for displaying these results.

OVER TIME

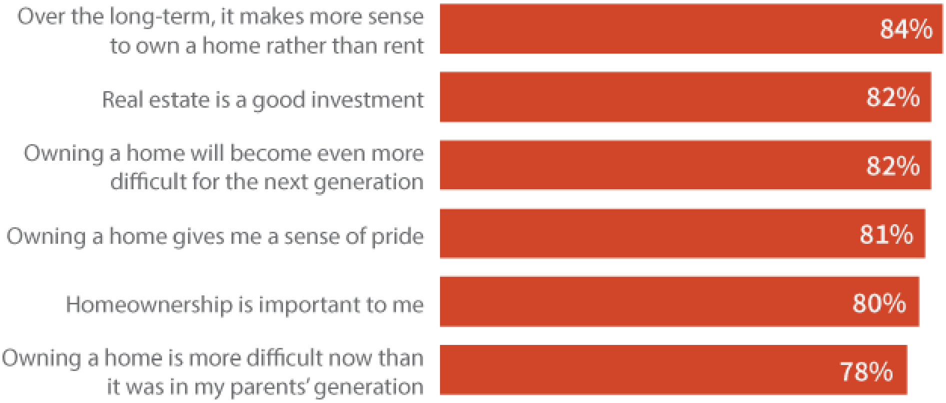
If your checkbox responses have been collected over a period of time, say 2009–2013, you could display the responses in a line chart as shown below.



Attitudinal Measurements

So far, we've been looking at the use of checkbox questions to gather data on basic counts (e.g. electronic devices used). Checkbox responses can also be used to assess agreement with attitudinal statements. The graph below shows attitudes of homeowners towards owning their home. In this case, since the statistic of interest is what percentage of homeowners agree with each statement, it is probably best to keep the graph as is, with the total exceeding 100%.

Please select the statements that most closely match how you feel about owning a home. Select up to three responses.



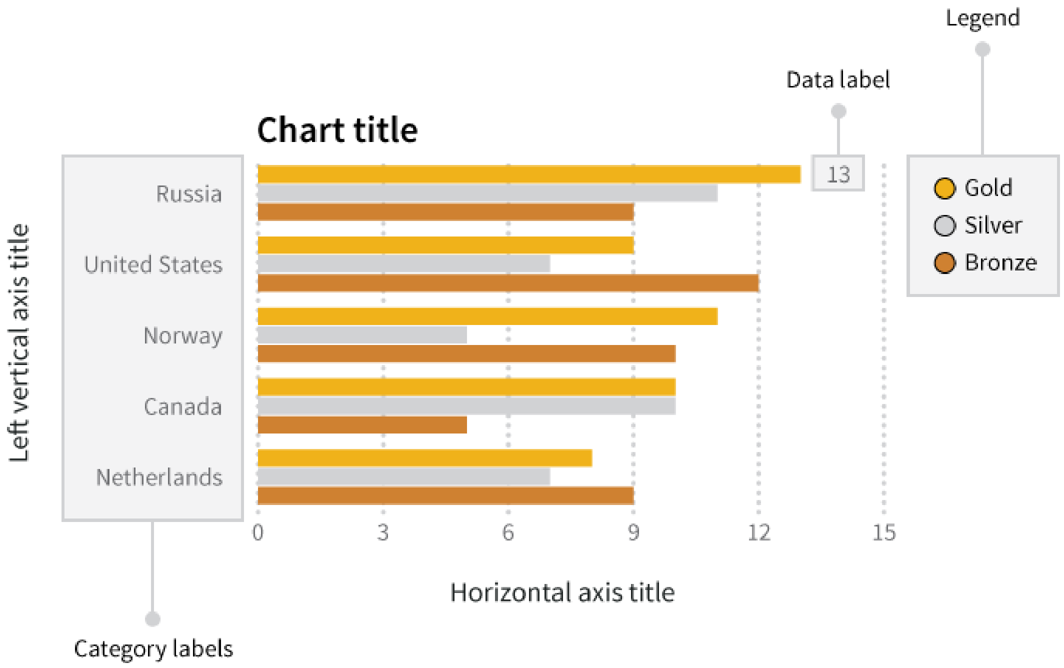
CHAPTER 14

ANATOMY OF A GRAPHIC

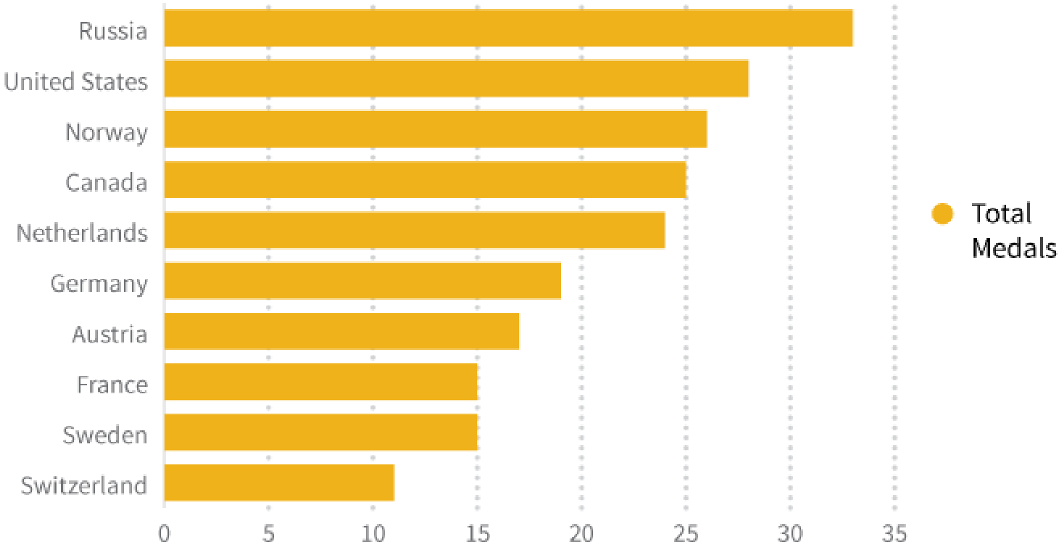
BY AMANDA MAKULEC

Once you've built a draft of your chart, the next step in creating an impactful visualization is making sure all of its elements are labeled appropriately. The text components of a graph give your reader visual clues that help your data tell a story and should allow your graph to stand alone, outside of any supporting narrative.

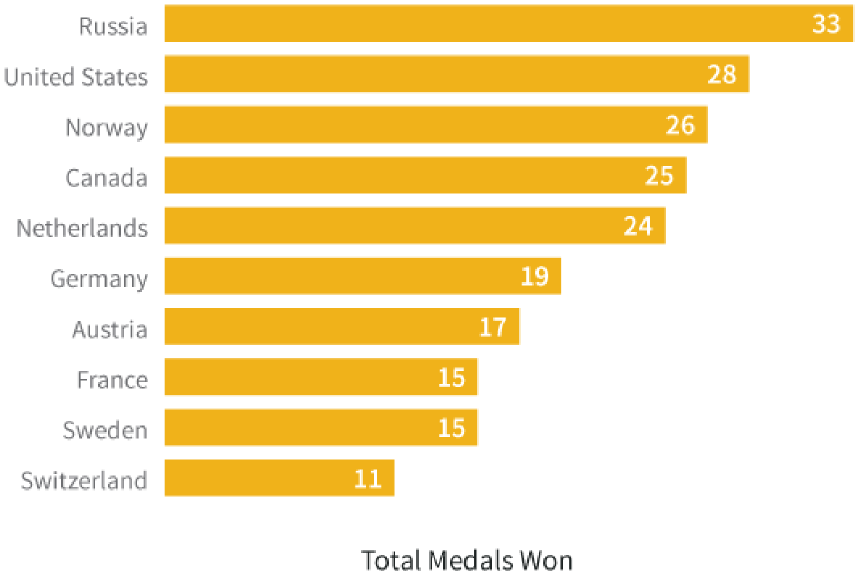
Let's make sure a few terms are totally clear up front. When we're talking about data labels, category labels, axis labels, and other pieces of text on your graph or chart, we're generally referring to what is mapped out in Figure 1.



One of the most impactful ways to show the value added by text labels is by showing the same visualization with and without these labels. Let's take these two graphs showing the number of Olympic medals won by country at the Sochi Winter Olympics in 2014.



Total Medals



Top ten countries by total medal count

Sochi Winter Olympics, 2014

Both graphs display the same data (and show how home team Russia ran away with the medal count!). The first has the default text generated by Excel, giving limited information to the reader, while the texted added in the second shows how smart use of simple text can help your data tell a story.

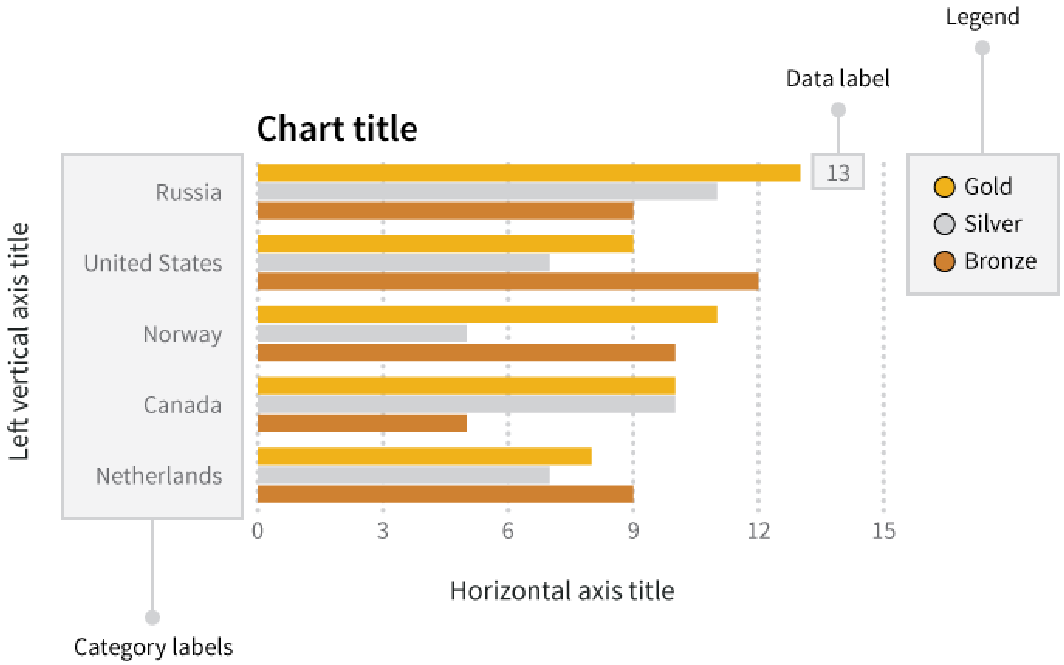
In Figure 2, the simple addition of different text clues helps the reader understand what he or she is reading. Check out the the chart title, axis labels, legend, and data labels. Each of these pieces of information is an opportunity to help your reader understand the message you want to share through your graph.

As important as text can be to help you tell your story, it uses prime real estate and should be as simple and direct as possible. Over-labeling axes or data points, or writing excessively long chart titles, is a quick way to clutter your visualization and reduce its impact. Graphs that pack a punch are often elegant in their simplicity. With the electronic tools we have for developing graphs and charts, changes can be made with the click of a button and a few keystrokes, so don't be afraid to experiment with your labels to help you strike this balance! Try adding different labels, editing them, and then taking them away as needed until you find the right balance to most effectively communicate your message.

In most cases, you want a chart or graph to be able to stand alone, outside of any narrative text. Sometimes that means tacking on explanatory text to clearly articulate your message. This could be written as a caption above or beneath the graph to clearly articulate your message to a specific audience. For the sake of simplicity, we'll focus here on the essential labels and legends within your graph, rather than crafting additional explanations.

Clear and Relevant Axes

Let's be clear about a couple of terms up front so there's no confusion about how we're using them: category labels are the ticks and categories along the axis that identify the data. Axis titles are the short text titles included along an axis to identify what variable is being displayed. Refer back to Figure 1 for a quick reference if you need to!



You should use axis titles to give your reader a summary description of the variable you're displaying. Depending on the data you're working with, you may have extremely long and descriptive names for each of your variables (if someone built your data table and added lots qualifying information), or very short, abbreviated variable names (as is often the case with machine-readable data). What you write along your horizontal and vertical axes should start with the axis title and be followed by the units you're using to display information (percent, dollars, etc.). The axis titles remind your reader what information you're sharing with them and help limit confusion about the units.

Labeling Categories and Axes

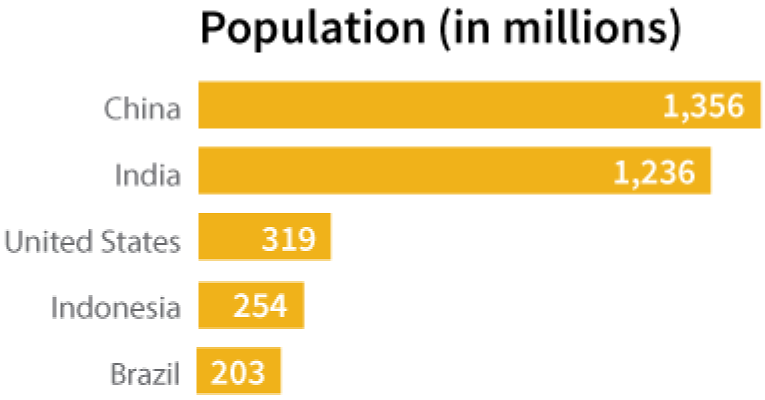
It's important to find a balance between giving enough information to your audience and keeping your text simple. Make sure you include enough information to explain the data you're presenting while keeping the text simple enough so it doesn't clutter the graph.

In our Olympic medal example, the horizontal axis shows the number of medals won at the Sochi 2014 Olympic Winter games. When creating a graph, it's common to go through a couple of iterations of the axis labels and chart titles. Once you've picked an axis title, think about how your reader could interpret that information. An axis title of "Medals" seems straightforward enough, but what exactly do we mean? Medals won? Medals lost? Medals won in 2014 alone? Medals won in any winter Olympic games? Your chart title can help clarify these questions, and your axis title also gives you the opportunity to tell your story.

Instead of Medals, let's title the horizontal axis with "Total Olympic Medals Won (2014)." Using the word "total" ensures the reader knows you're talking about all medals, not only one kind (gold, silver, or bronze) or from one event. "Olympic" ensures your reader knows that you've shared only the Olympic medals, not a total of Olympic and Paralympic medals for the same year. And finally, "2014" reminds your reader in what year the medals were won. Alternately, as we've shown here, you can include some of those details in the chart title, which we'll talk about later in this chapter.

Units

When we refer to units, we're talking about things like "percentages", "dollars", "millions", or other descriptors that make sure your reader knows what kind of information you're displaying. These are important if you're sharing data that are in percents instead of counts or if a multiplier has been used (for example, millions of dollars instead of just dollars).



In this graph of population by country, the words "in millions" make a big difference. This is a surprisingly easy thing to forget, so don't let it happen to you! Population data from http://www.census.gov/popclock/

It's important to be clear in labeling your units. Some quick checks to walk through when you're assessing your axis titles:

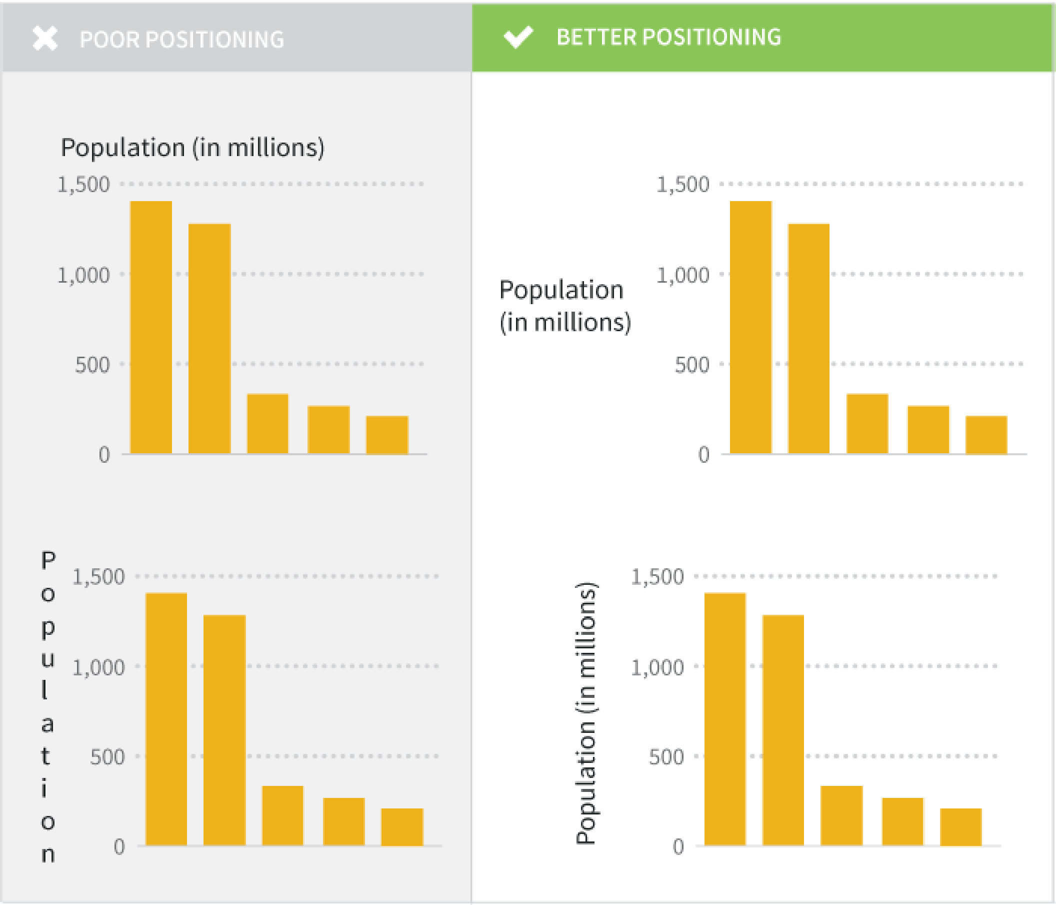
- Make sure you're consistent if you're displaying multiple types of data using a dual axis graph: you don't want to display two variables with one in months and a second in years on the same axis!
- Instead of using abbreviations or acronyms, write out the full unit (for example, millimeters instead of mm).
- Symbols can sometimes be a helpful alternative to words in text, but make sure to clearly state what a symbol represents. For example, you could use "$" instead of the word "dollar." However, this symbol is used with the currency of more than 30 countries, so you should specify which country's currency you mean (e.g. US$, A$, Mex$).

There are occasions where, for the sake of visual simplicity, you can get away with omitting the axis title if the categories you're sharing are self explanatory. In our Olympic medal graph, we've listed the country names along the vertical axis. If you're confident your audience will be able to recognize that the fields listed are countries, there's no need to over complicate the graph by adding an axis title that says, "Country." If you have any doubt about your audience's knowledge of your categories, err on the side of caution and label your axes.

Position and Format

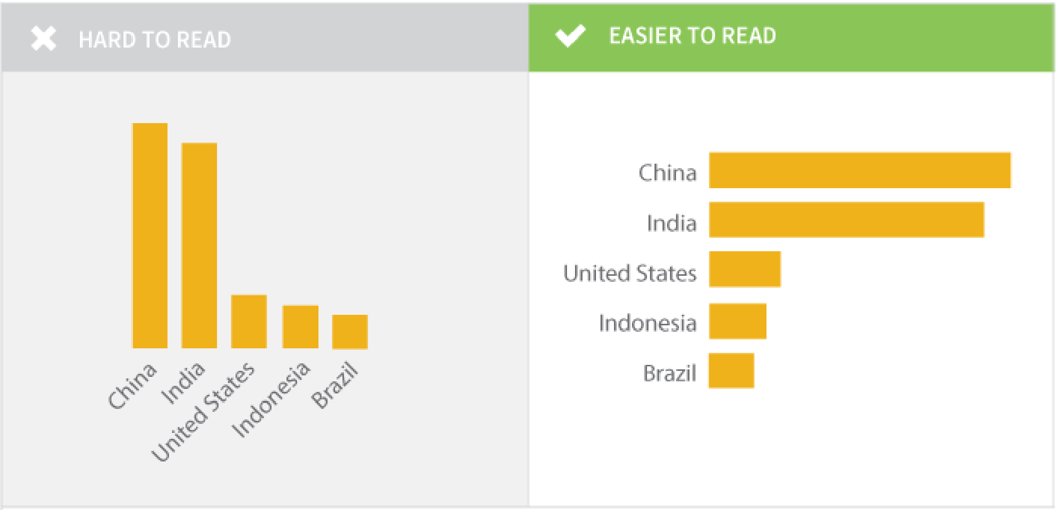
Horizontal axis titles should be aligned parallel to the axis labels. Format them in a font size large enough to read, but not so large that it dominates the graph.

There's some debate over whether or not vertical axis labels should be aligned parallel to the axis or not. On the one hand, aligning the text vertically makes it very clear that it's directly associated with the vertical axis. There's often more room to write the axis title text if it's rotated at a 90-degree angle. On the other hand, humans are not very good at reading vertical text, and readers may find themselves squinting and turning their heads to understand the chart. If you have enough room and the axis title text is short, consider keeping the text level instead of rotating it. We've also included two examples below that show what to avoid when placing your vertical axis titles.

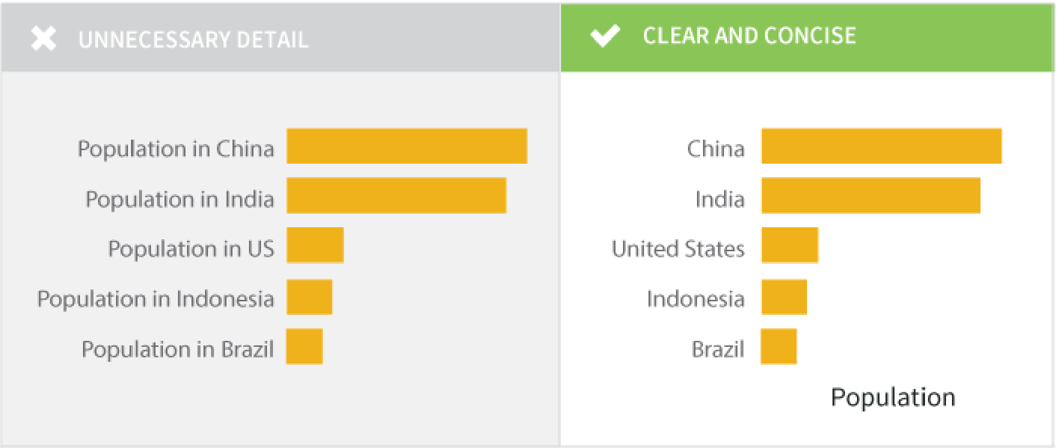


Individual category labels should be clearly marked too. Make sure any words and numbers are easy to read: your reader should not have to play contortionist games with his or her neck to get a good look at what you're trying to share in your graph. A few tricks for formatting your axis labels:

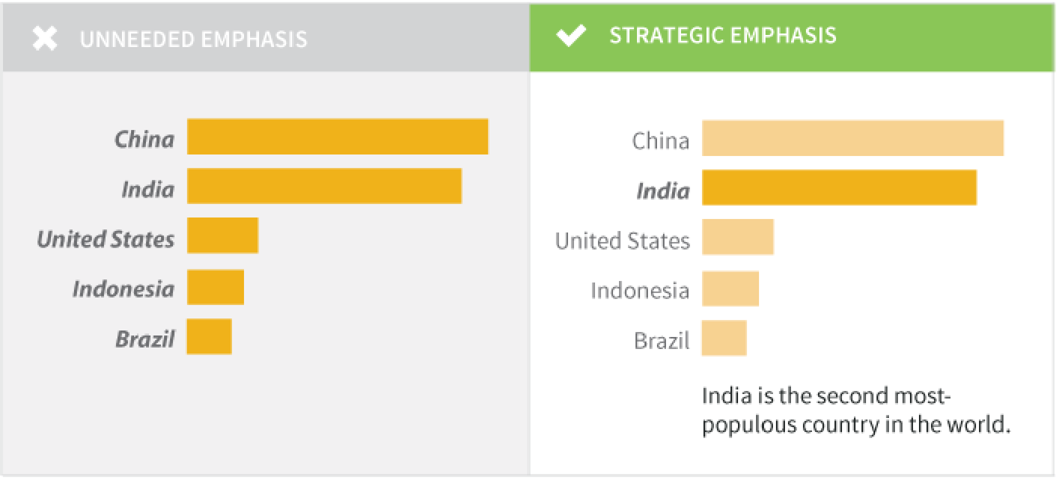
- Where possible, avoid rotating text on an angle, which is often used in vertical bar charts to force long text labels to fit but is difficult to read. If you find yourself with a horizontal axis using rotated, dense labels, try changing your chart type to a horizontal bar which will make the labels easier to read.



- Simplify your labels. Even if your data were originally formatted with long strings of text for each category you're trying to display, think about what you can simplify without losing meaning.



- Use boldface text and italics sparingly and only where they have meaning. For example, you might use boldface for a category label for a data point you're highlighting.



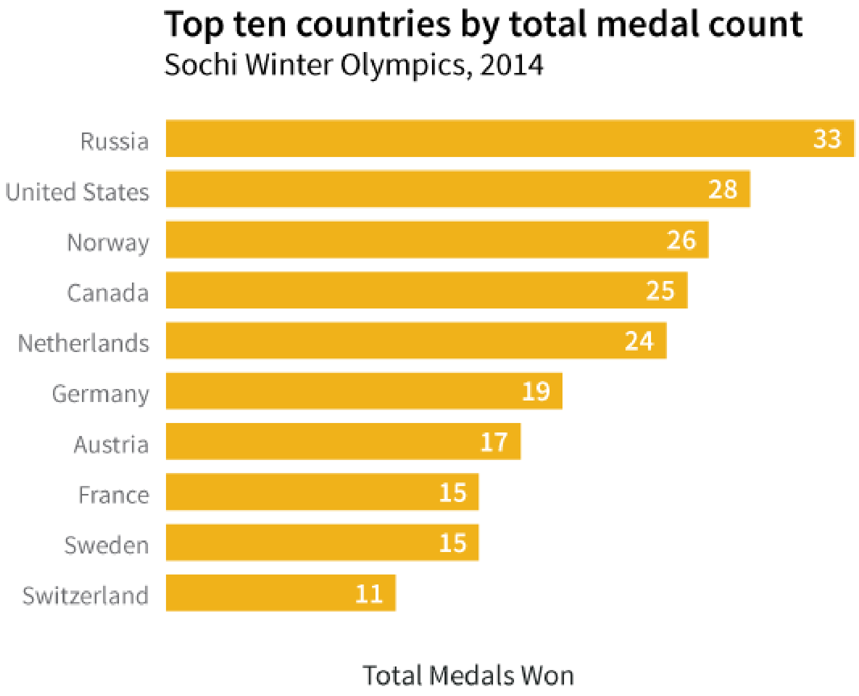
LABELING YOUR DATA

To label or not to label, that is the question. A data label is most commonly added as the data value associated with a bar or point on a graph. When you want to make sure your audience has access to the precise data points, adding data labels works well. What labels to add, if any, depend on your audience and the story you're trying to tell with your chart or graph. Labels can add detail and value to your visualization, but can just as easily clutter it and confuse your reader.

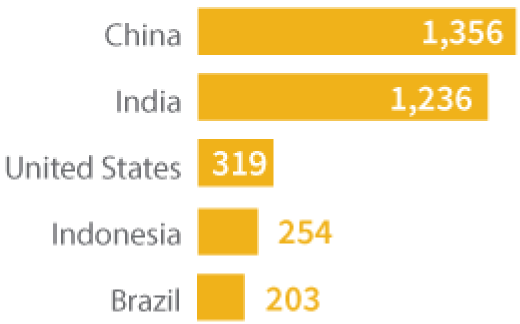
You don't have to limit your data labels to the data values themselves, though the numeric values are the most common ones added. Instead, you can label the bars/dots/etc. to highlight specific pieces of information with a text description where appropriate. For example, if you're creating a complex chart that shows the number of Olympic medals won by a country over the past 50 years you may want to label the year with the highest total as "most medals" to help it jump out for the reader for its value (and not just the raw number).

When deciding how you can use data labels to create impact, think about your audience. Do they need to know the precise numbers you're displaying on your chart? Are you trying to show a trend? Are you using a graph to help your audience easily see highs and lows? Depending on what information your audience needs, you could choose one of three approaches towards labels.

OPTION 1: LABEL EVERY POINT

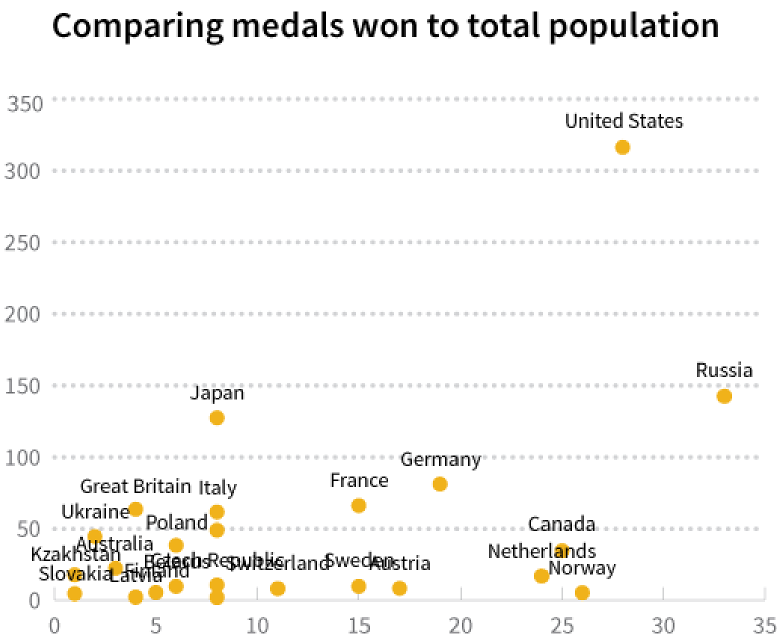


Data labels for every point ensure your reader can easily read the precise value of each data point. On bar and column charts, this also eliminates the need for grid lines. Color and position are important to consider when adding all of these values to the graph. For bar charts or pie charts, placing the labels inside the bars can help the reader quickly identify the important information. On a bar chart, you can place the data label on the inside or the outside of the bar. Placing the label inside the bar often makes it more clear where the bar ends, and reduces the chance that the reader will misinterpret the data.



In this bar chart, two of the data labels are too big to fit neatly inside the bars. A reader glancing quickly at the chart might accidentally view those bars as being longer than they actually are. Be aware of how the positioning of labels can affect a viewer's visual perception of the data values.

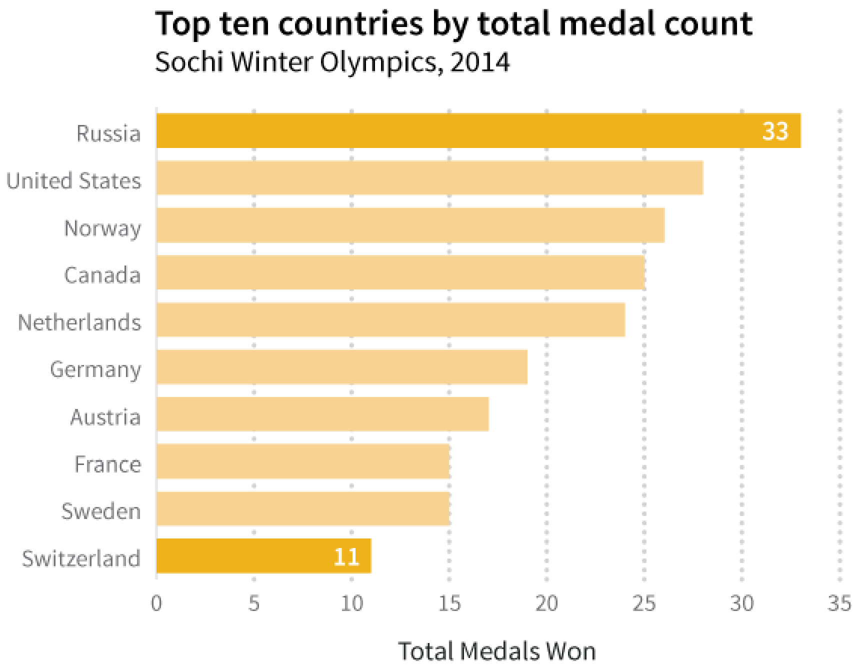
Labeling each point can provide detailed information, but also make your graph look messy when you have a large number of fields. Scatterplots are a great example of a graph type that typically looks messy with all data points labeled, particularly if the labels are text (country names) like in Figure 4. So many labels! So much overlap! So…hard to read.



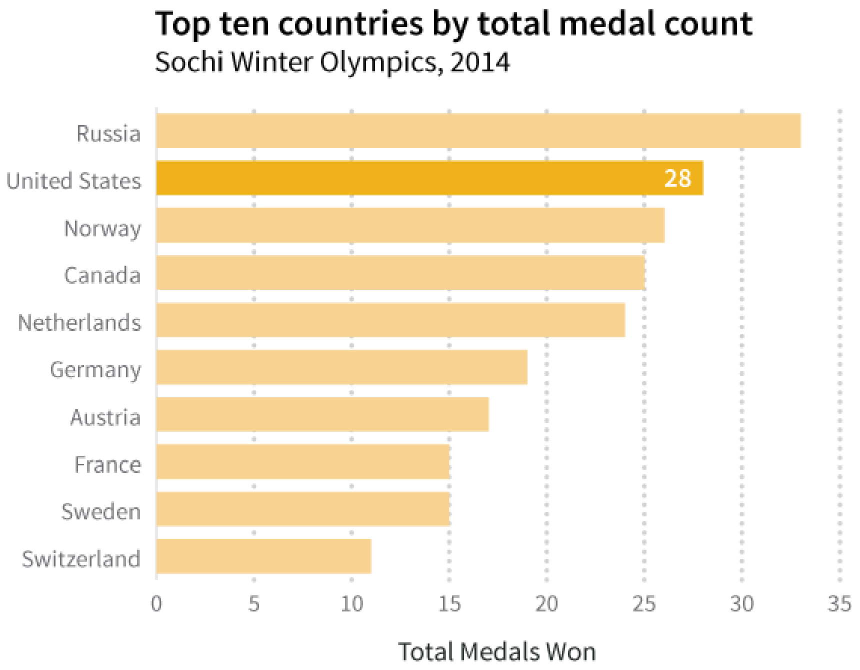
Determine if your reader really needs to know each precise value or if you're trying to show trends, highs, or lows, in which case move onto Option 2.

OPTION 2: LABEL THE DATA YOU WANT TO HIGHLIGHT

When you're trying to emphasize how specific data points compare to others (for example, highlighting the data from one country across many in the same region) or emphasizing the high and low values, having selective data labels can guide your reader. Here are two examples for different audiences.

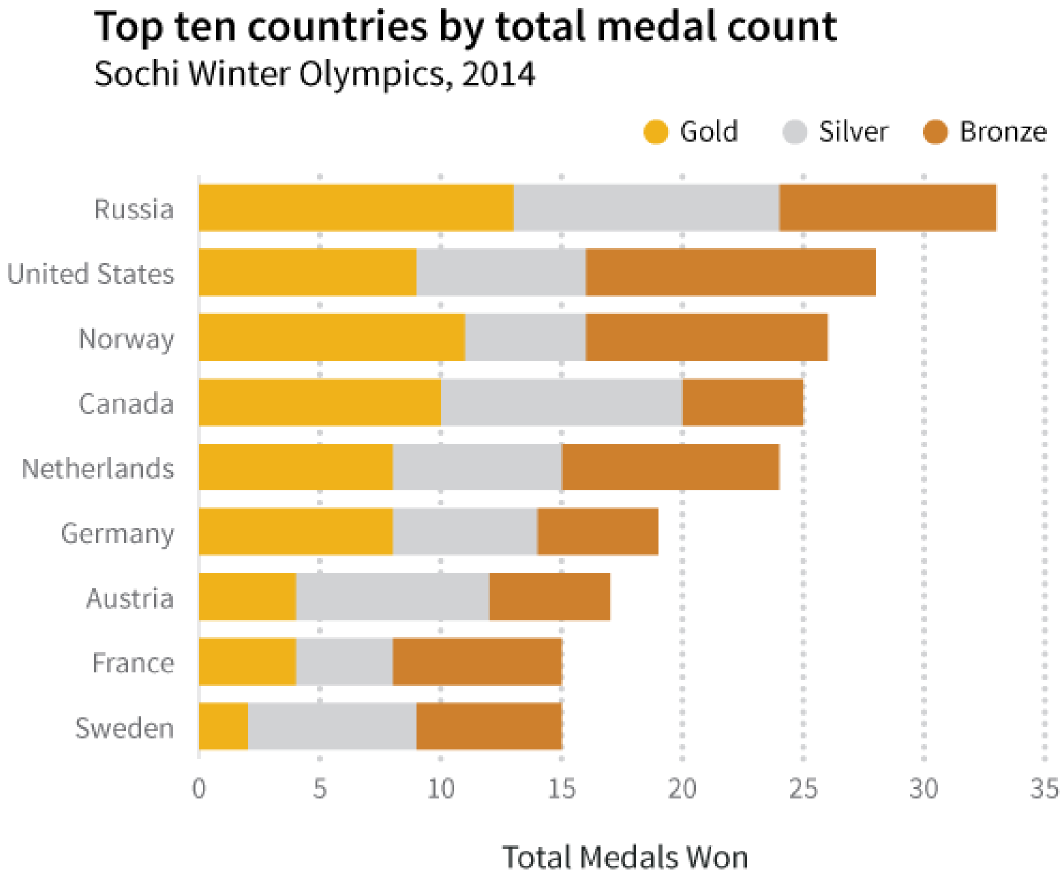


If your audience is interested in the range of values for number of medals won by the top ten countries, label the high and low values. By highlighting the highest and lowest values, you establish the range of the values. Additionally formatting the choices, like making the high and low bars different colors, also calls out these values.



If your audience is from one specific country and wants to see how their medal wins compare to the other top ten countries' medal totals, highlight and provide the data point for the single county (for example, the United States).

OPTION 3: NO DATA LABELS



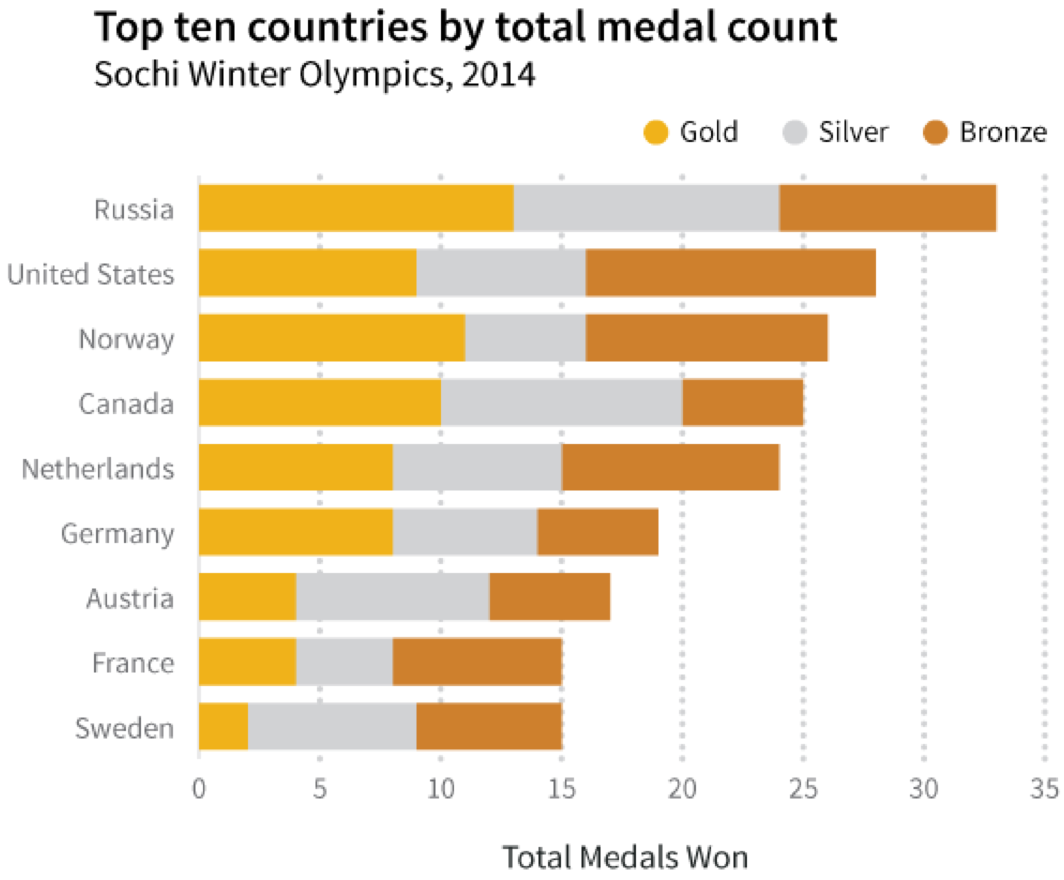
We often build graphs to simplify the visual presentation of information. If you can effectively convey your message without complicating your visual with extra labels and text, just say no to adding data labels. Reserve the use of labels for the occasions when you need that level of detail: make your default to be no labels and then add as needed.

Because of the way that the human brain interprets angles and area, it can be tough to visually assess the values associated with slices of pie charts. For that reason, it’s a good idea to add data labels to pie charts for clarity.

Simple and Powerful Legends

A legend ensures that your reader knows what your graph’s colors or patterns mean. When you’re only showing one kind of data where color or pattern doesn’t carry any special meaning, you don’t need a legend (as in the earlier bar charts, only showing total medal count). If you have multiple kinds of data represented on your graph or chart, a legend is essential.

Say that instead of showing total medal count, you want to show the total count broken down by medal types (gold, silver, and bronze). You could include them all one bar chart with a different color to represent each medal type.



Here, each color represents a single medal type and is clearly defined in the legend on the top of the graph. Consider how your reader will look at your graph: obscuring the legend in a corner can make it difficult to find quickly, while laying it over white space on the chart (for example, where there are short bars) can be distracting.

This graph is a great example of the importance of position for the legend. By aligning the legend along the top of the bar chart, it's easy to visually place which color aligns with which data category. While the placement of the legend is partially an aesthetic choice, its placement can visually assist your reader in interpreting your graph.

Concise but Impactful Titles

At last! You've labeled your axes. You've added (or decided not to add) data labels. You've ensured that the categories or values along your axes are clear. Now comes the opportunity to tell your story in 1-2 lines of text: your chart title. The title should succinctly tell your reader what message this graph or chart has been designed to convey. It doesn't have to be a complete sentence, but it should be free of complicated language and syntax.

It's important to think about who your audience is and where your graph is going to be published when writing your chart title: are you a journalist adding the chart to an article, a blogger creating a very visual post, or perhaps a scientist going for a peer-reviewed publication? Scientists may need to clearly articulate the relationship displayed between their variables whereas a journalist may want to give some spin to the title to emphasize the message they want to share.

The most basic graph titles clearly state the relationship between the independent variable on the horizontal axis and the dependent variable on the vertical axis. For example, we called our medal count graph, "Top Ten Countries by Medal Count, Sochi Winter Olympics 2014." This title tells that reader that we're looking at the total medals won (medal count) by country, limited to the top ten winners at the Sochi Winter Olympic Games in 2014. Each element of the chart title we've created adds value and tells something to the reader. There's nothing wrong with titles like these: while somewhat dry, they're straightforward and clearly tell your reader what is on the graph.

You can also approach a title as an opportunity to state the conclusion your reader should draw from your graph. For example, take where we highlighted the US medal count. If you want to highlight the US ranking second in overall medal count, you could title the graph, "US ranks second for total medals won, Sochi Winter Olympics 2014." Often you see these leading titles in newspapers and other publications telling a specific story. The chart title provides an extra opportunity to emphasize and amplify your message—use it!

Looking for Feedback

One of the best ways to see if your chart or graph clearly communicates your message is to ask an impartial party to take a look. This is the time to call in that favor and ask a friend or colleague for five minutes and some honest feedback. Ask your reviewer to identify if there's any information they feel is missing. Remember not to take it personally if there are edits that need to happen; the best charts and graphs happen through multiple drafts and with the help of collaborators.

The contributors for this book have also set up a group to give you feedback on your visualizations. If you're looking for honest feedback from an impartial party, HelpMeViz (http://helpmeviz.com/) is a community feedback site that can help.

This final check is a great way to make sure you've provided enough detail in your axis labels, axis titles, chart title, and legend. If some thoughtful changes are suggested, go back to the file and make a few edits: it will be time well spent, and at the end of the day your visualization will be better for it!

CHAPTER 15

IMPORTANCE OF COLOR, FONT, AND ICONS

BY HOWARD COALE

Form (Frequently) Tyrannizes Function

Listening to someone with a good voice sing a few bars of “I Want to Hold Your Hand” pretty much beats out the experience of hearing someone mutter “$E=MC^2$” any day, but that preference isn’t a reflection on the value of the information in those two phrases. Rather, it describes how the experience of information can change the perception of its value. In other words, as information designers, we aspire to be singers, not mumblers. No, that doesn’t mean “change everything to make it pretty”: it means you must first and foremost acknowledge that you are not presenting information, you are designing an experience of information.

In this chapter, we’ll cover a few basic design principles and tools that can be used to sway, focus, clarify, enhance, and sharpen your audience’s experience and understanding of data (and specifically fonts, hierarchy, color and icons).

THE MYTH OF OBJECTIVITY

Before we dive into design fundamentals, we need to understand just how much design can impact the perception of information. At its heart, data visualization is a form of storytelling, and no story is ever told from an objective point of view. The graphic representation of data has been used just as effectively by dictators and terrorists to justify acts of violence as it has by scientists to report groundbreaking research.

To that point, a critical rule in data visualization is to realize you’re creating, at some level, a biased point of view. If you don’t, you’ll have little chance of successfully judging the honesty, accuracy, elegance or utility of your solution. Even the most typical chart should seek to captivate and inspire that part of the human mind that yearns for meaning and narrative while in the act of seeking information.

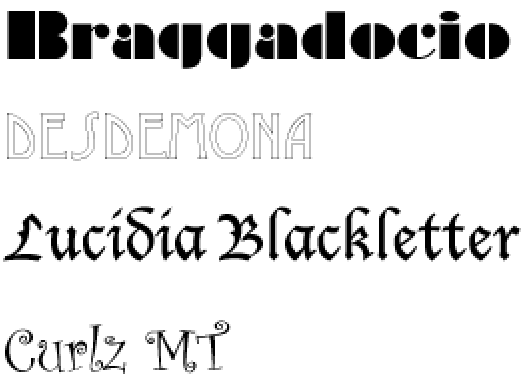
You might be a researcher reading this who wants to ask: “I’m making a graph to show how many times a rat turned left or right in a maze. Why does a graph that displays that information necessarily reflect a bias or a narrative perspective?” The answer is that though the data show the rat turned left 35 times and right 22 times, the way the presentation of that data is designed will — usually within one or two seconds — teach the viewer what is important and what is not. Being aware of that fact during the design process can make your graphic representations of data much more effective and puts you in a better position to think through what you want to express before, and while, you design.

Fonts & Text

LEGIBILITY AND READABILITY

When designing the viewer experience for a given set of information, the most important rule of all is to strive for ruthless simplicity in every choice you make. In the “narrative” of the information you are presenting, the style of presentation should never demand attention (that is, the design should complement the data, not the other way around!). That applies to everything from the design of an illustrative representation like a chart to something as seemingly simple as font selection. In the world of typography, the balance between legibility and readability is a decision that a typographer makes when designing a typeface (more commonly known as a font). Not all fonts that are legible (i.e., one character can be easily distinguished from another) are necessarily readable (i.e., designed in such a way that supports and enhances reading so that the reader easily absorbs the text’s meaning without being distracted by the font).

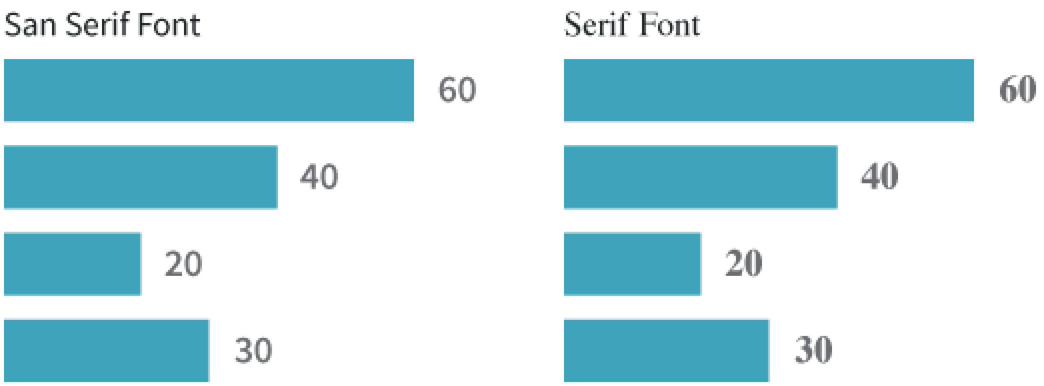
For instance, there are many fonts that are designed specifically to stand out. For example:



These “display typefaces” are good examples of fonts that are legible, but not readable. They were created to emphasize small pieces of text, like a title or heading, or to convey a specific style rather than communicate information quickly and transparently. In general, you should not use display fonts in data visualization because they create additional work for the reader. You can’t afford your viewer struggling for a second, or even half a second, trying to understand what something says. Nor you do you want a specific font skewing the meaning of your data because your audience automatically associates that font with a specific type of content. We’re looking at you, Charlie Brown, when we say no Comic Sans!

This is one way that data visualization and storytelling are similar. If your audience becomes too aware of the style in which a story is told, they may be distracted from the meaning, or the valuable details in the content. The bottom line is that you should use fonts that are legible and strive to present them in a readable fashion.

Fonts for data visualization that do a good job of communicating the message without calling attention to themselves include Helvetica, Arial, and Gill Sans. In general, it’s best to use sans serif fonts for numbers in charts of any sort because serif fonts tend to have more “visual noise” to be interpreted. For example:



Believe it or not, your mind takes a fraction of a second longer to understand the numbers on the right than it does to understand the numbers on the left. Why? Serif fonts have stylistic variations in their endpoints, widths, curves and shapes that adds more for the eye to take in at any given moment.

Since data visualization seeks to speed the process of insight, you want to stay away from anything that slows down the viewer's understanding, even if by an infinitesimal amount.

Hierarchy

On the subject of speeding up comprehension, one of the designer's greatest tools is use of hierarchy. Hierarchy refers to the ways in which we weight and organize information so that the audience can more quickly and easily scan and, more importantly, understand what they're seeing or reading.

Perhaps the most familiar application of hierarchy is the good old-fashioned outline that we all learned in grade school. By breaking the information into titles, headings, supporting information, and details, you create a clear map by which your audience can quickly scan the information, absorb the content, and look for the details they're interested in.

Data visualizations with a clear hierarchy tell the reader how to process the information and in what order, whether from the most important information to the least, or from the big idea to the supporting details. Hierarchy can also give your viewer the freedom to choose where to enter into a given set of information, which is especially useful when presenting large amounts of data.

Fonts can help establish hierarchy through variations in size (11pt vs. 14pt), weight (bold, medium, regular), and emphasis (italics, capitals). Color, when used selectively and consistently, can also help delineate hierarchy by creating emphasis and guiding the viewer's eye in a way that facilitates understanding. Other common tools include line breaks, bullet points, numbered lists, indentations, etc. The means and methods to establish hierarchy are endless, but remember that the guiding rule is ruthless simplicity in all things. You don't want the tools you use to establish hierarchy to distract your audience from the data experience you're trying to create.

Color

CONCEIVE IN BLACK AND WHITE, THEN ADD COLOR

To paraphrase the designer Paul Rand, if a design doesn't make sense in black and white, it will make less sense when color is added. That's not meant to imply that graphic representations of data should always be black and white. It only means that balance of line and solid, dark and light, layout and structure are the foundation on which any chart or visualization is built and, as such, the design of these elements needs to be able to stand on its own.

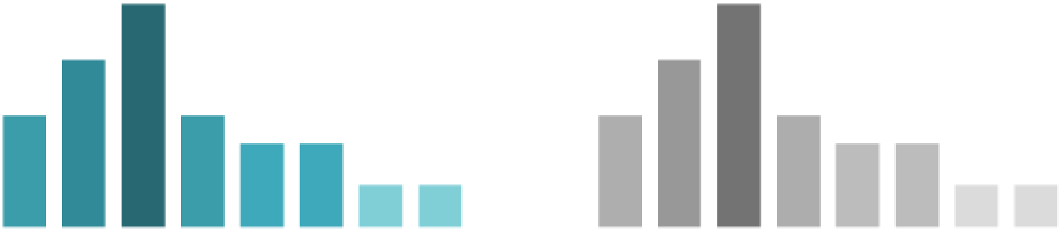
The assumption is that you are making visualizations because you want to provide clarity, insight, and context. Color is a tool, but it should never be the starting point. Instead, conceive and plan your charts and visualizations in black and white whenever possible. Then, apply color conservatively with the primary objective of highlighting and enhancing comprehension. Here are three basic tips to keep in mind when assessing how to use color in graphic representations of data.

DON’T FORGET GRAY

Gray is one of the most important colors in your palette because it offsets the colors around it. In the world of information design, it is critical for modifying, clarifying and balancing almost any color palette, and taming the overall graphic comprehension of charts.

COLORS CAN BE USED TO EXPRESS VALUE

Don’t forget that color can be used as a more conceptual axis to designate value. That said, when using colors to represent a scale, or gradations of value of any sort, look at what you’ve created with all the color/saturation removed. This helps guarantee that the various tints you are using appear as truly different values to the eye.



THE ROLE OF WHITE SPACE IN READING COLOR

If you were painting, you could merge colors in one seamless quilt. But in information design, every color you use must have a reason for being there and should be distinct. If you are creating a chart in which a very slow gradation or change from one color to another has meaning — as it might if representing climate data, election turnouts, or other continuous variables — then the changing and merging of colors is central to interpreting the data. However, in many cases, colors need to be distinct since they will represent a unique value or attribute, and to help the viewer's eye read them as distinct and clear, the colors need room to breathe. That's what white space can do for you.

Icons

A SHORT HISTORY

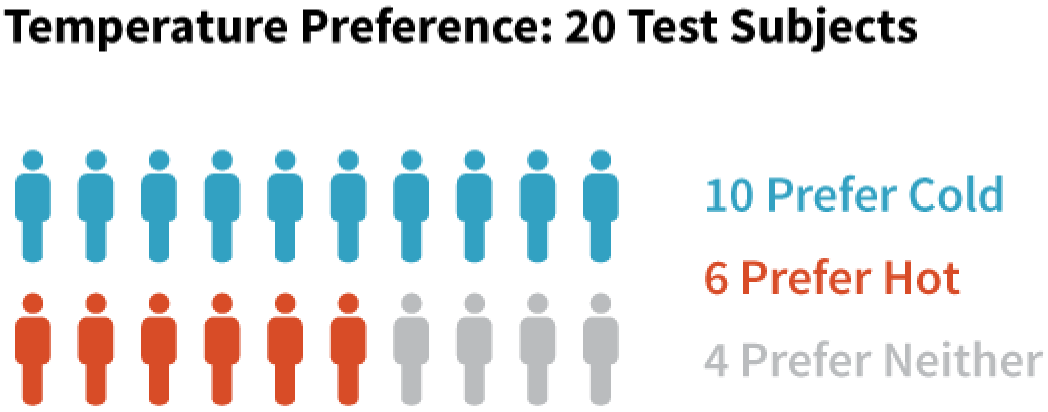
The first graphical user interface (GUI) was introduced by Xerox in 1981 in the Star workstation, also called the 8010 Information System, but the ideas that inspired it were first explored by one of its chief designers, David Canfield Smith, in his 1975 PhD dissertation titled, *Pygmalion: A Creative Programming Environment*. A couple years later, Steve Jobs famously "borrowed" the idea from Smith and Xerox for the Macintosh. Only a few decades later, two thirds of the seven billion people on Earth use mobile phones that are largely driven by graphic, icon-based interfaces.

Yet, simply drawn, easy-to-remember pictograms have been used since Neolithic times to represent information that was too complex to explain in a small amount of space. Within the practice of data visualization itself, iconography has been used as far back as the early 15th century, borrowing from other disciplines that sought precision and accuracy in drawings like architecture, military planning, cartography, alchemy, and astronomy.

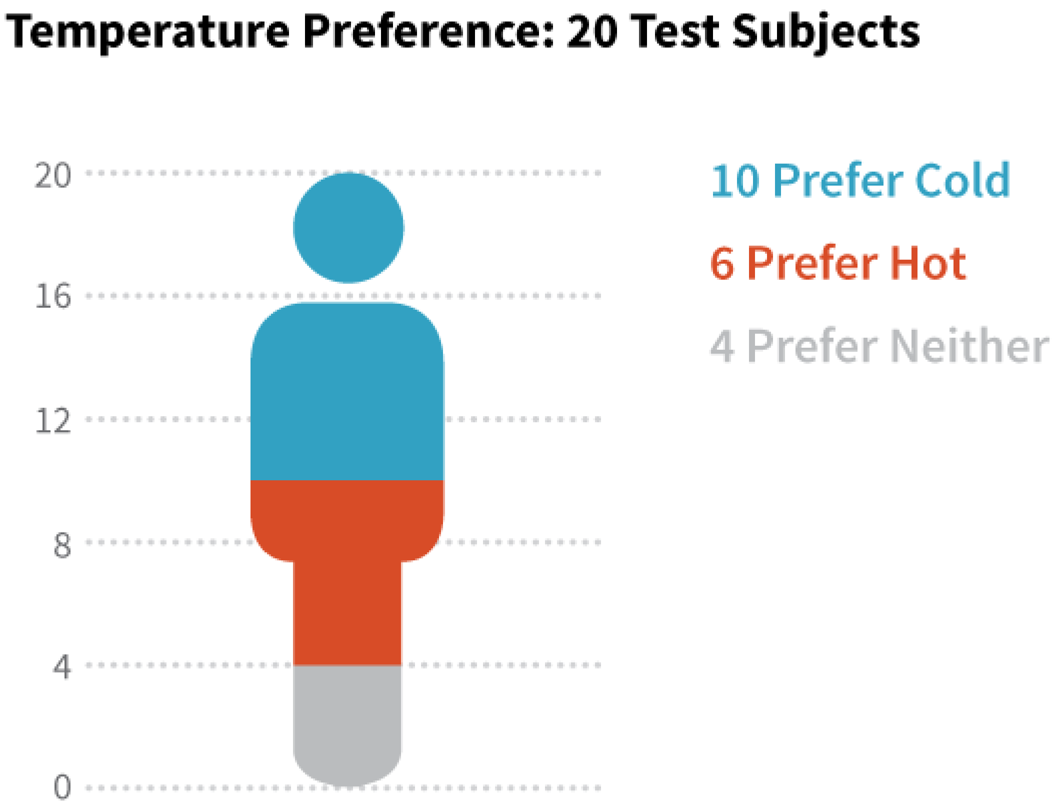
USE SKEPTICISM & LEAN TOWARD THE LITERAL

The use of icons is now so widespread and commonplace that they are often added without much thought. People feel they should "just be there" in order for an app to feel like an app or a website to feel user-friendly. As a result, we often see them without ever wondering why they were put there in the first place.

Even so, when creating data visualizations, you should regard the use of icons with caution. If you can make something clear in a small space without using an icon, don't use one. Why? Because icons can make things more obscure, not less. As soon as anyone asks, "What does this icon mean?", you've lost the battle for simplicity. This is true for icons that are used as metaphorical identifiers in infographics or data visualizations, as well as for functional icons that perform a task or reveal more information within the context of an interactive graph or infographic.

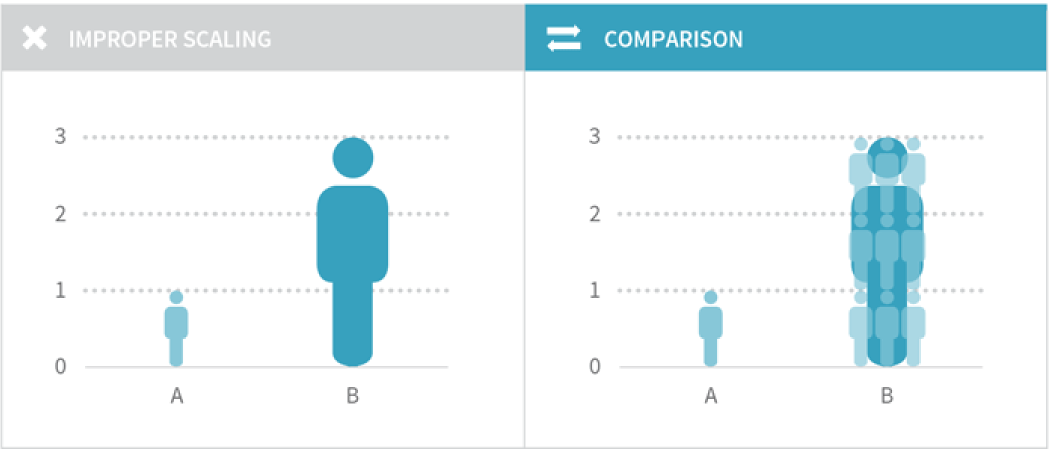


It would be very hard to misunderstand what the icons are meant to represent here: one person-shaped icon equals one person. This is a very literal representation, and that is why it is so easily understood. In fact, the more literal and simple an icon is, the better it is. It is a very similar issue to the one faced when selecting fonts: if the viewers spend too much time acclimating to an elaborate presentation, they will often miss the full value and substance of what is actually being presented. Here is another version of the same information:



The exact same data is being conveyed here, but it takes the viewer slightly longer to understand it. Why? Before the viewer can correctly interpret the data, they have to create a complex metaphor in their head. First, they're being asked to imagine that this icon represents all twenty test subjects, but they're then also being asked to divide that metaphorical person by another set of criteria. As simple as they may seem, icons are still asking the mind to pretend they represent something else. The more you ask symbols to represent, the less literal and less clear they become. When it comes to using symbols and icons of any sort, stick to one simple meaning each.

You should also make sure that if you use icons in graphs, you properly handle their sizing. A common mistake is to scale both dimensions of an icon, which leads to an incorrect visual comparison. In the image below, the height of the person figure has been tripled to indicate that Group B is three times the value of Group A. However, in order to keep the ratio of the image, the width has also been tripled. This results in the larger figure actually being nine times the size of the smaller one rather than just three.

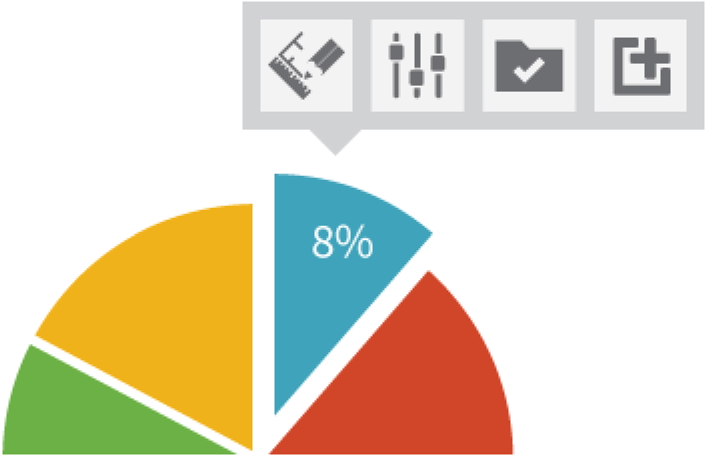


Note how in the improperly scaled pictogram bar graph, the image for **B** is actually 9 times as large as **A**.

STAY COGNIZANT OF FUTURE NEEDS

It is not uncommon for new traders to join an investment bank and have someone walk them through the company's software on the trading floor, with the assurance: "Don't worry, you'll understand what all the icons mean after you've been here a few weeks." A mass proliferation of icons across an interface or data visualization often indicates that options have been added after the overall planning, design and (in the circumstances of interactive data visualization) development has taken place.

Here is a simplified example of the sort of icon-based fold-out palette one frequently finds in financial analysis dashboards that have gone through several versions in which capabilities have been added over time:



One of the greatest challenges in the graphical representation of data is how to ruthlessly simplify even while the complexity of choice for the viewer or user increases. As in the example above, sometimes icons are used to solve this problem; but as you can see, this can create another problem—confusion.

The best way to avoid this is take time during the initial conceptual work on any given data visualization to assess what the extreme requirements might be in future iterations. Ask yourself what options or views might be added in the future? Or, in the circumstance of complex interactive data views, what possible functionality might be introduced that would change the nature of the overall visualization?

These are just a few of the core guidelines for using icons in infographics or complex data visualizations.

CHAPTER 16

PRINT VS. WEB, STATIC VS. INTERACTIVE

BY SARAH RIGDON

Whether you're designing graphics for static environments or for interactive ones, it's important to address a few key things up front. First of all, what do we mean by static and interactive?

Defining Print Vs. Web; Static Vs. Interactive

You can design static graphics for print, the web, or both. With static graphics, you'll provide everything you want your viewers to see without requiring any input from them. The most a static image demands from a viewer is to turn a page or scroll or perhaps enlarge an image. Typically, static graphics display information that isn't subject to change.

You can also design interactive graphics for print or the web or both. Interactive graphics typically provide viewers with the option to dive deeper into the information, often requiring some degree of input. Now, you may be thinking, "Interactive print?" Actually, yes, though of course it's not as common as interactive graphics on the web. Some examples include:

- Interactive projects
- Art installations
- Flip books/analog animations
- Popup books

So, static isn't a stand-in for print, nor is interactive a stand-in for web.

One more thing: the antonym to static isn't interactive. It's changing, moving, or dynamic. Motion graphics don't require viewer interaction, but they're not static, either. Any given cable news network will give you examples of these. Do they display necessary information? Not usually. Adding motion to infographics can keep viewers' attention, at least at first. But over the course of a segment or a show, viewer fatigue sets in. Motion graphics often look impressive, but you should always ask whether motion is necessary for conveying the information and be wary of overusing it.

Now let’s shift gears and talk about some of the core design concepts you should consider when creating static and interactive print-based and web-based graphics.

Color

As for color, there are a lot of questions to take into consideration before you break out your palette, especially if you’re working in print. Are you working with full color or black and white? If you’re working with color, consider what you’re gaining that makes it worth the cost difference from black and white. After all, grayscale and two-tone printing can still communicate a lot of information. Is a full-color print job really necessary?

How many colors are available for the print job? What kind of paper are you working with? Colors work differently with news print, magazine, and printer paper. Or are you getting fancy and putting the graphic on merchandise? A good merchandising company will work with you to make sure your t-shirts, totes, buttons, or stickers look great, but you’ll want to anticipate these concerns in advance.

On the web, a lot of these restrictions go away, but not all. Many designers have to work within a website’s color palette. They have to ask questions such as: does the website’s green color scheme completely clash with this bright pink graphic? Does it subtly complement the branding and logo design? The answers to each of these can have an impact on how the viewer receives the final design.

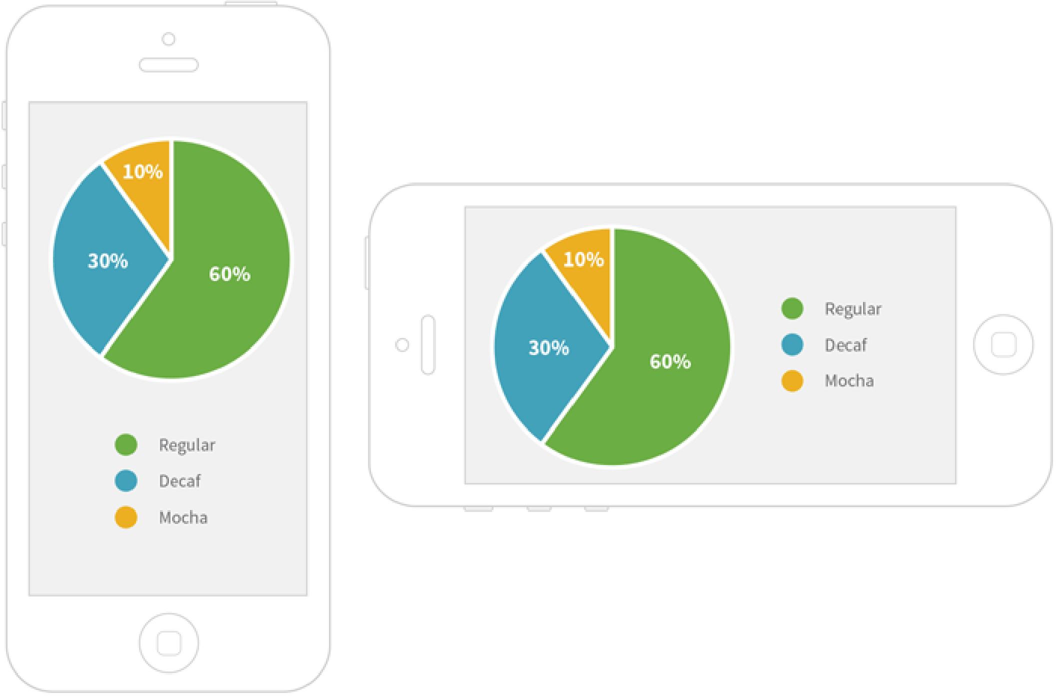
Size

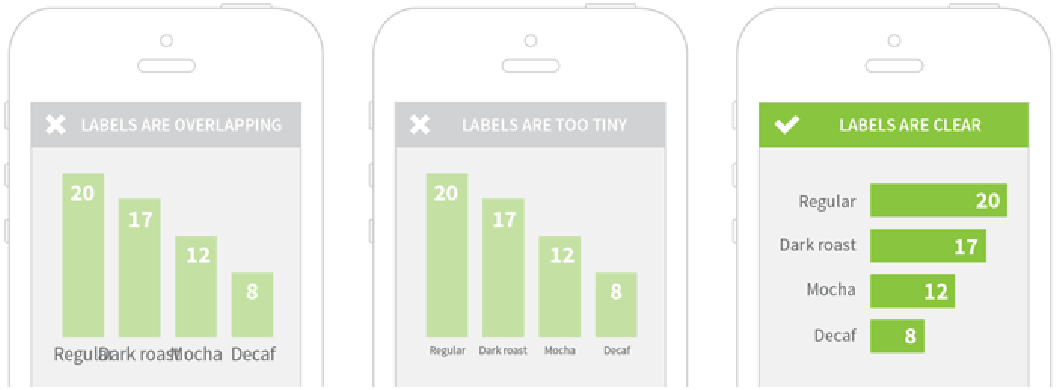
Step back and take a look at the space you’re working with. You have to consider size for both print and the web.

With print, you have the page size constraining your graphics, but that can also be a blessing because you only need to design for a single width and height.

On the web, you can often include image-zooming functionality or allow viewers to open a larger version of a graphic in a separate browser tab, but you should consider the effects that zooming and scrolling can have on a user's experience. A well-designed graphic should provide key information in a clear and elegant manner with minimum effort required by the user.

Mobile devices are also a concern, as more people are using cell phones and tablets to browse the web. Keep in mind that nearly two-thirds of cell-phone owners now use their phones to browse the web, and twenty percent use their phones to do *most* of their web browsing. Cell Internet Use 2013. Pew Research Internet Project. "Main Findings." Duggan and Smith. 16 September 2013: http://www.pewinternet.org/2013/09/16/main-findings-2/ It's hard to see a lot of detail on a tiny touchscreen, but you don't want users to have to squint to understand your graphic! You could build a responsive graphic — one that automatically changes its layout according to the dimensions of the viewer's screen size — or you could design a completely different graphic specifically for display on small mobile devices. Given the state of mobile web browsing trends, it's something you should consider when publishing a visualization on the web.

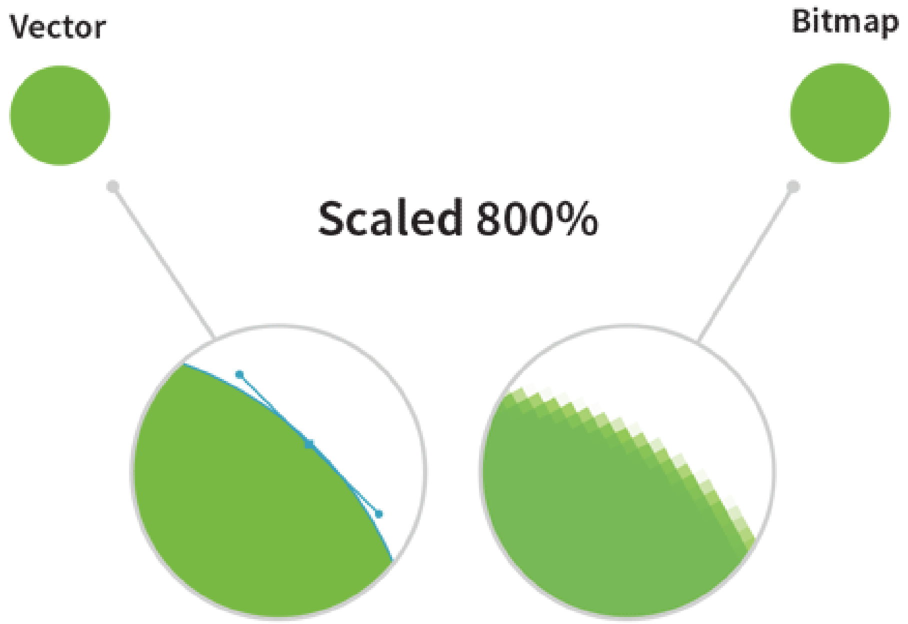




Many mobile devices (and increasingly, laptop and desktop monitors) have high-resolution displays, often referred to as *retina* displays. A static image designed at the standard web resolution of 72 dpi (dots per inch) will usually look blurry when viewed on a retina display. You’ll need to make double-resolution retina versions of your image-based graphics if you want them to look sharp on, say, a modern iPad. You can also create them using web fonts and SVGs, and they will always look crisp, no matter what the screen resolution. One downside to this is that some older browsers (notably Internet Explorer 8 and below) do not natively support SVG format.

WHAT'S AN SVG?

SVG stands for Scalable Vector Graphic: it's a web-friendly format for displaying vector images—images that can scale to any dimension without pixellating by using mathematical expressions to draw paths. Vector images will look crisp even when blown up to very large sizes, unlike bitmap images (also known as raster graphics), which use bits of specific colors which are mapped to certain regions in the image. Bitmap images will look blurry when viewed at a high resolution, but they are often necessary when the color combinations are complex. A photograph is a great example of a bitmap image, and a simple circle is a great example of an image that can easily be drawn as a vector image.



Many of the tools for building web-based interactive charts and graphs (such as D3.js (http://d3js.org/)) use SVGs to render visualizations, especially interactive ones. Most of the static infographics, however, are published in bitmap image formats, such as JPEG or PNG.

Size doesn’t just refer to how big an image appears. Pay attention to the size of your file as well. People all around the world use mobile devices to browse the web. Not only can it take a long time to download a large image across a mobile network, but in some parts of the world, each megabyte of data is an added cost. Be nice to your viewers and work to reduce your final image size.

Layout

With static images, the point is to provide everything you want to convey in one or more frames. There are no mouseovers, you can’t click around and select certain details and un-select certain others (e.g., show only state rodeos from 1988 to 1996). Make sure that your key points are visible. If you’re creating a static version of an interactive graphic, keep in mind that you might not be able to show everything that was accessible in your original interactive graphic. If possible, provide URLs so that your readers can also access the raw data and the interactive version of your graphic. That way, they’ll have the option to dig deeper and find more information about the parts of the data that didn’t make it into the print version.

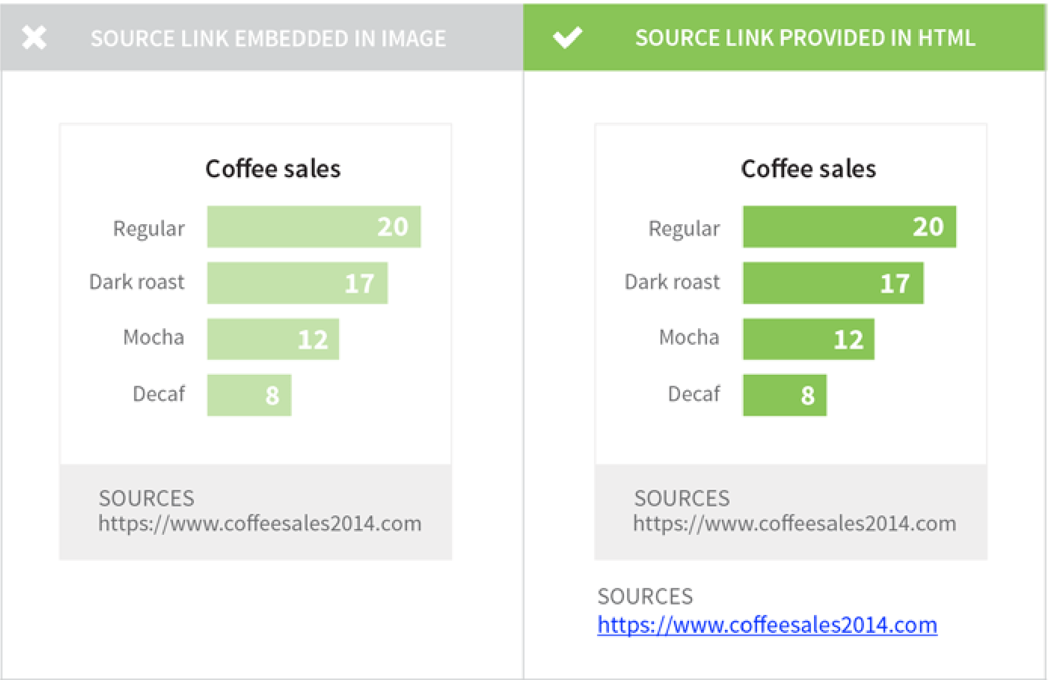
With interactive graphics, make sure to test any clickable controls that manage the interaction. Are there buttons or slides off to the side, or are they a part of the graphic? Does the viewer have to scroll around to see everything with several options engaged? Test your graphic in different browsers to make sure it all fits—and functions—the way you want.

A NOTE ON STATIC GRAPHICS ON THE WEB

These aren’t necessarily interactive, save in that viewers can enlarge them. Because the graphic itself doesn’t change, the main considerations of layout still apply. But just by being on the web, static graphics give us more interactivity than print. For example, they can be linked to their data source, another related graph, or a different web page. Mobile screen sizes still remain a challenge for static graphics specifically because their layout doesn’t change to accommodate smaller screen real estates or different width-to-height ratios.

> *Text labels and visual details that are clear on a larger screen can sometimes be difficult or impossible to read on a mobile device.*

Although URLs displayed as part of a static image can't be directly clicked on, it's common for thoughtful publishers to paste reference links just below the graphic so that readers can easily access the source data.

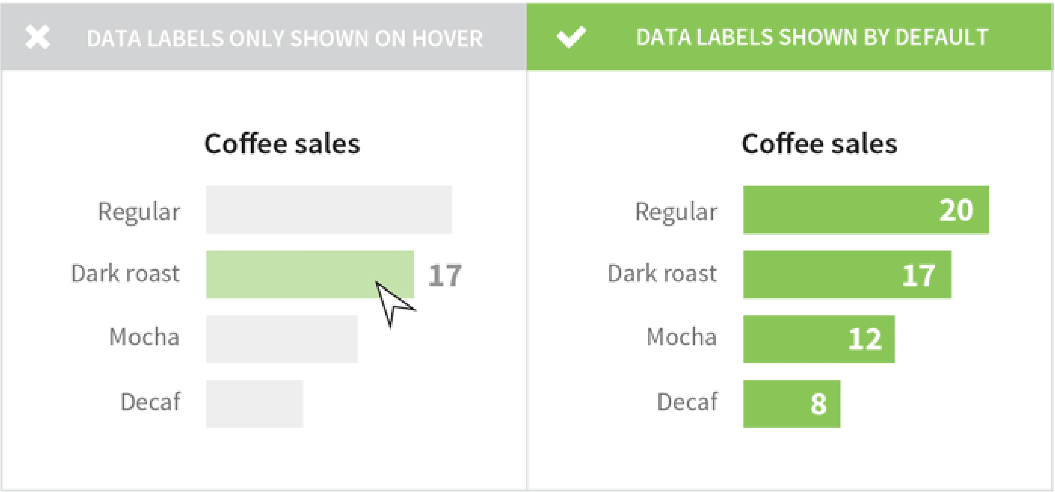


WEIGHING THE NEED FOR INTERACTIVITY

Making interactive graphics is getting easier all the time with new software and applications. Interactivity can be not only extremely impressive but also useful, highlighting different types of information more effectively than a static image ever could. But sometimes, you'll come across an interactive image that actually makes the information more confusing. Other times, you'll come across an interactive image that seems neat at first, but as you play with it, you realize it's just bells and whistles. You're not learning anything new, and the information could just as clearly have been presented in a static image. What a let-down!

Why spend time making a graphic interactive when it doesn't need it? Interactivity is a tool in the designer's toolbox. Interactivity should be used when it improves upon a static graphic or adds value to the information and does something that the static image alone could not.

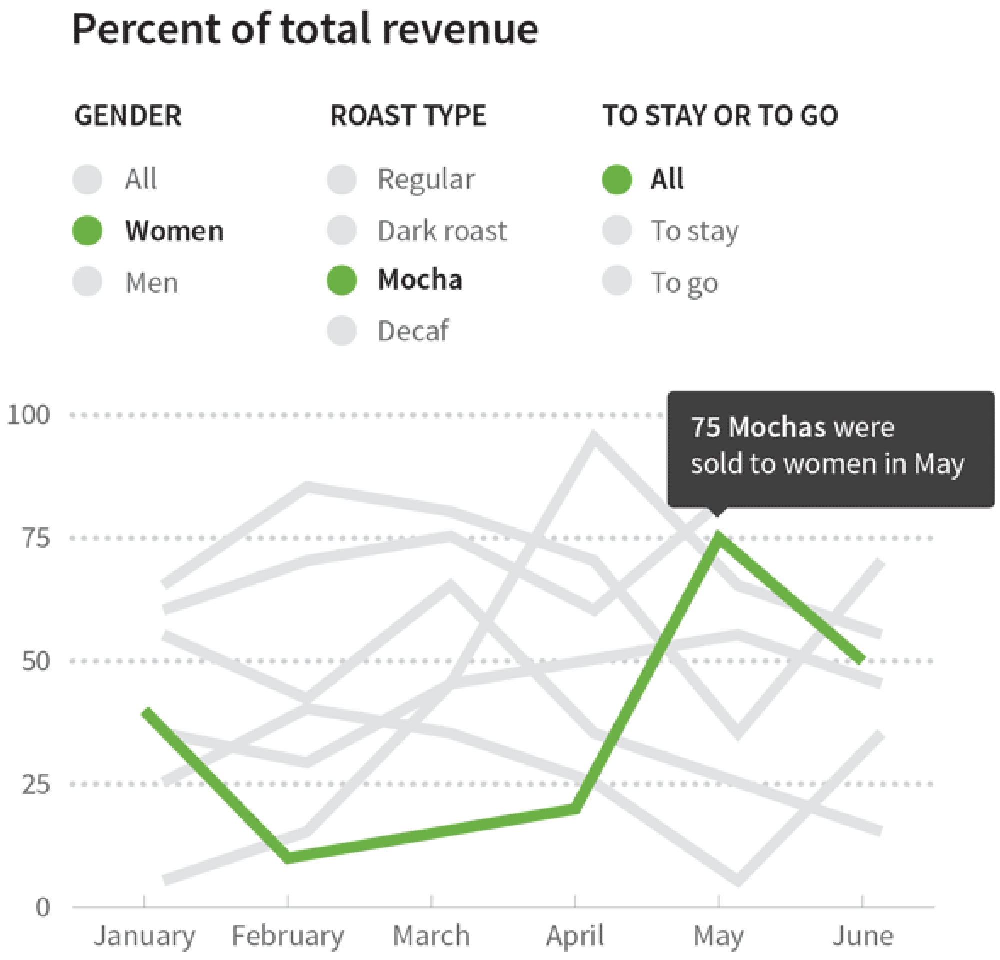
For example, consider the following bar chart:



In the figure above, the interactive labels (e.g., the hover effects) aren't necessary for the viewer's comprehension. The labels can be added to the chart *before* hovering, which makes the chart easier to understand and requires less effort from the viewer. When you're presenting a small, simple chart with a single variable, interactivity can actually detract from your message.

> There are plenty of amazing use cases where interactivity adds value to a visualization, but you shouldn't make yours interactive just because it looks cool. Make sure you're actually improving the user's experience. For example, interactivity can help greatly enhance any graphic that illustrates a rich data set you want the user to explore. Let's say you want to show how different groups of consumers drink coffee. An interactive visualization might allow your user to filter by demographic groups and explore coffee consumption patterns.

Interactive graphics can add a lot of value when you have a rich dataset that you want the user to explore. For example, let's say that you want to know how different groups of consumers drink coffee. An interactive visualization might allow you to filter by demographic groups and explore coffee consumption patterns.



In static graphics, you can't rely on visual effects like movement or color changes to make your point. On the other hand, while interactive graphics can incorporate visual effects to show trends across time, they do introduce room for technical difficulties. Be mindful of compatibility across computer systems and browsers. Some of the more complex interactive visualizations require a lot of computational power and can be slow on less-powerful machines such as smart phones. Visualizations made using Flash won't render on iPhones or iPads, and those made using SVGs won't work in Internet Explorer 8 or below. Interactive visualizations built for a computer with a mouse may be difficult or impossible to use on touchscreen devices if the creator forgets to test and optimize the graphics for touch-based interactions.

Any medium for sharing a visualization — whether static or interactive — will have its strengths and its limitations. Think about your story, your audience, and how readers are likely to view your visualizations. Ask for feedback and continue to iterate on your process. And if you're looking for some friendly feedback and suggestions, you can post your works-in-progress on HelpMeViz (http://helpmeviz.com/).

WHAT NOT TO DO

Some cooking don’ts:

- Don’t mix baking soda and vinegar.
- Don’t serve pink chicken.
- Don’t throw water on a grease fire.

There are also data visualization don’ts. This section is about those.

CHAPTER 17

PERCEPTION DECEPTION

BY MIZEL DJUKIC

> “ *With great power comes great responsibility.*

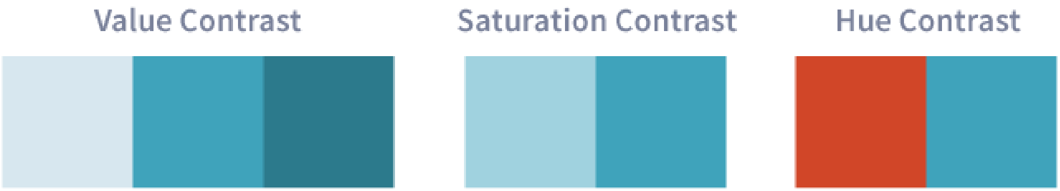
So far you’ve picked up a bunch of tools you can use to build compelling stories through visualizations. If the story you tell is powerful enough, it can change people’s minds and move them to action. But before you rush out there to change the world with your newfound powers, let’s go over some of the accidents your visuals might cause because of the quirky way our brains perceive things.

How We Perceive Differences

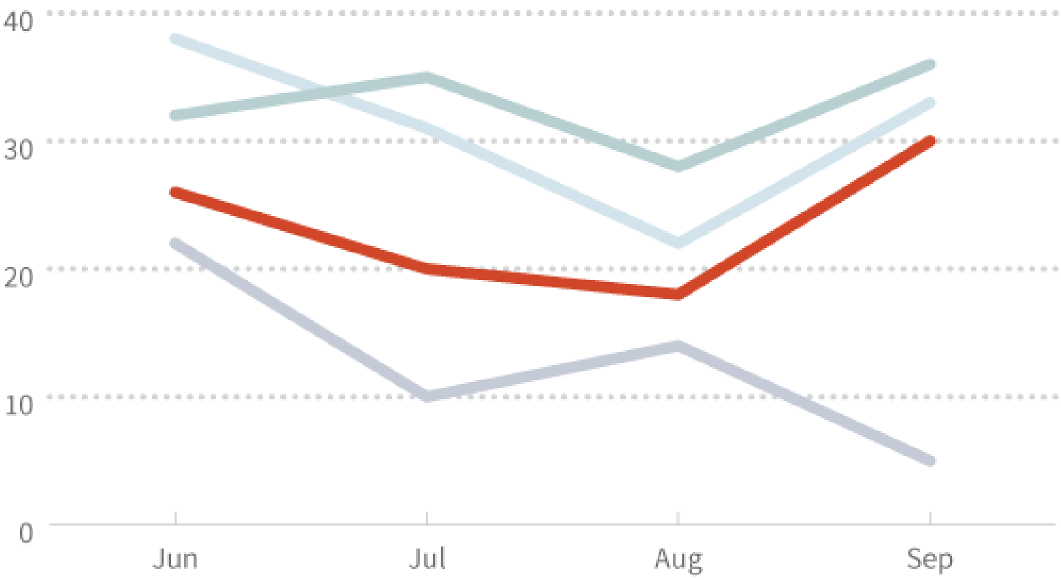
Our brains love a difference. Even when we’re just relaxing in front of our computers watching cat videos and drinking iced tea, our brains are constantly looking for boundaries in our environment, differentiating between our feet and the floor, our hands and the keyboard, our cup and the table. How we perceive these boundaries depends on a variety of factors, including color, contrast, shape, and perspective.

DIFFERENCES IN COLOR AND CONTRAST

When it comes to color and contrast, the example that probably comes to mind is black and white. It’s very easy for us to distinguish between the two because one is so dark and the other is so light. Varying the lightness and darkness of a color (known as the value) is one way to establish contrast, but it’s not the only way. You can also vary the saturation (or chroma) so that a color looks pale and faded or bright and vivid. Or you can establish contrast by simply using different colors (or hues) that are easy to tell apart from each other.



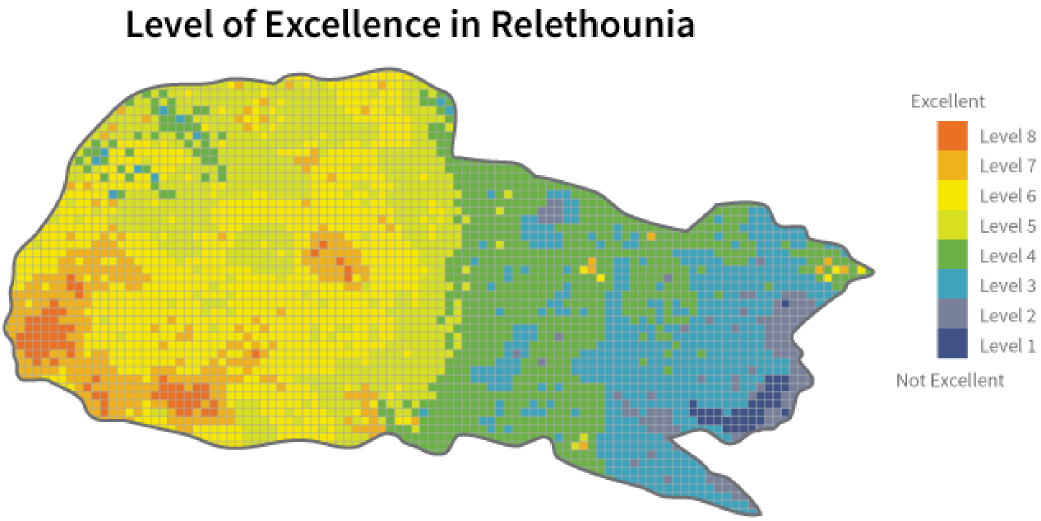
How our brains perceive contrast depends on the context. The higher the contrast between objects, the more different we think they are. Cairo, Alberto. The Functional Art. Berkeley: New Riders, 2013. Print. For example, take a look at the following graph:



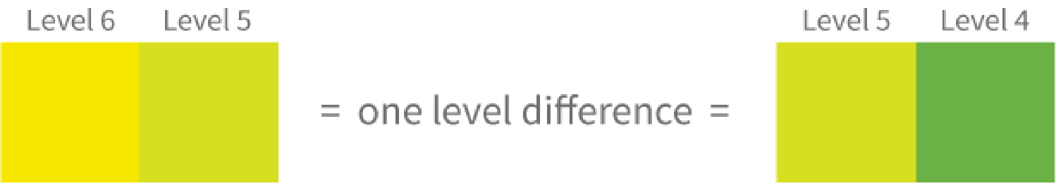
Anything stand out to you? Probably the red line, for a few reasons. First, the red line is more saturated than the other lines. Second, red and blue are good contrasting colors. Third, to our eyes, reds and yellows come to the foreground while blues recede to the background. Rockwell, Ken. The Secret: What Makes a Great Photo. Web. 1 March 2014. http://www.kenrockwell.com/tech/basics.htm

Because of the high contrast between the red line and everything else, our brains will naturally focus on it, and we'll probably think that there's something special about the red line. Having the red line stand out like this is fine if there's a good reason for doing so. For example, if each line represents the level of happiness over time in different neighborhoods, then the red line can highlight *your* neighborhood. However, if it turns out that there's nothing special about the red line, then it winds up being distracting and confusing.

Now take a look at the following map which shows the level of excellence in different parts of a land called Relethounia:



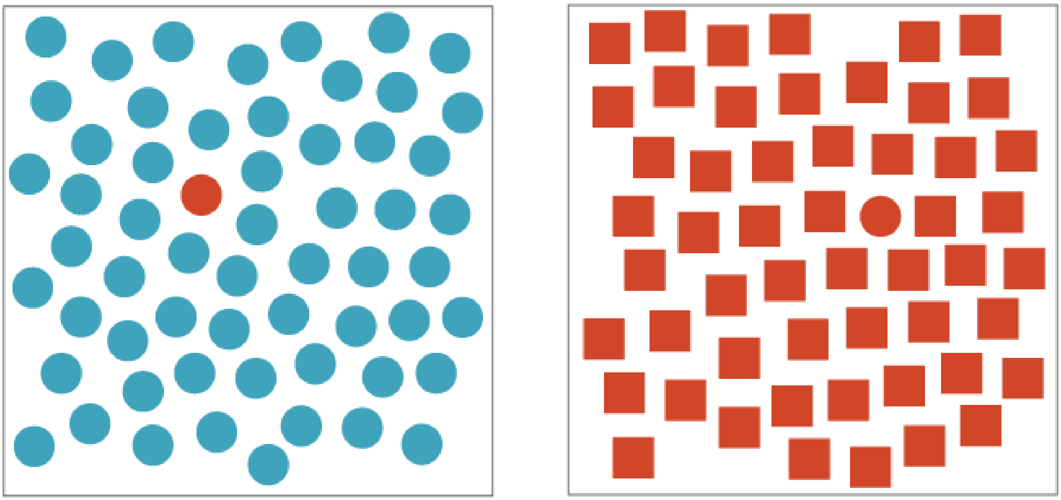
What stands out to you this time? Maybe that there seems to be some sort of division between the western and the eastern parts of the country? This apparent division is the result of the relatively high contrast between the light green and the dark green areas. So it seems like there's a bigger difference between these two parts than between, for example, the yellow and the light green areas. But when we look at the legend, we see that the light green parts differ from the yellow and the dark green parts by the same amount: one level of excellence.



So be careful with color and contrast. When people see high contrast areas in your graphs, they'll probably attach some importance to it or think there's a larger difference, and you want to make sure your graphs meet that expectation.

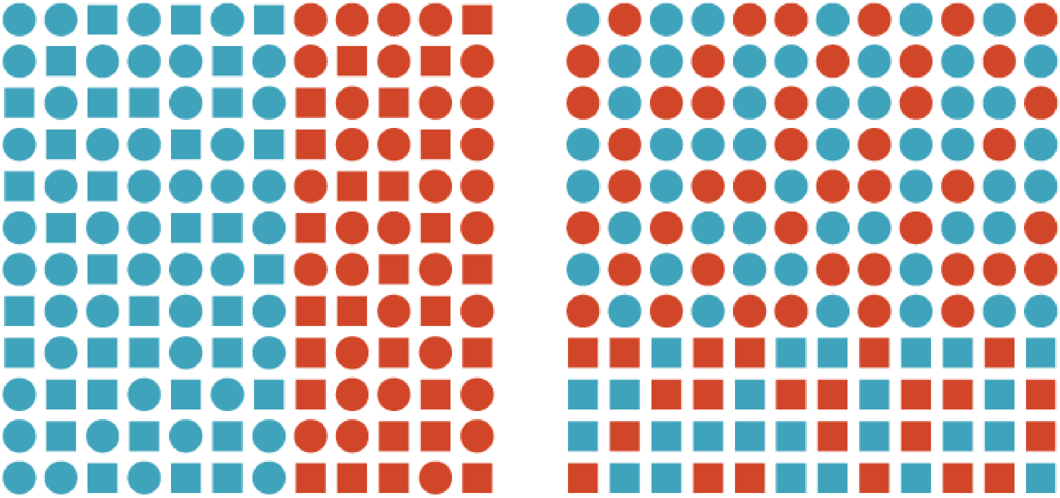
DIFFERENCES IN SHAPE

Our brains have a much harder time telling the difference between shapes. To illustrate, see how long it takes you to find the red circle in each of the figures below:



Adapted from: Healey, Christopher G., Kellogg S. Booth, and James T. Ennis. "High-Speed Visual Estimation Using Preattentive Processing." ACM Transactions on Computer-Human Interaction 3.2 (1996): 4.

Did you find the circle faster in the first figure? Most of us do because our brains are better at differentiating between colors than between shapes. This is true even if some other factor is randomly changing and interferes with what we see. Healey, Christopher G., Kellogg S. Booth, and James T. Enns. “High-Speed Visual Estimation Using Preattentive Processing.” ACM Transactions on Computer-Human Interaction 3.2 (1996): 107-135. Web. 16 June 2014. For example, see how long it takes you to find the boundary in each of the figures below:



Adapted from: Healey, Christopher G., Kellogg S. Booth, and James T. Ennis. “High-Speed Visual Estimation Using Preattentive Processing.” ACM Transactions on Computer-Human Interaction 3.2 (1996): 5.

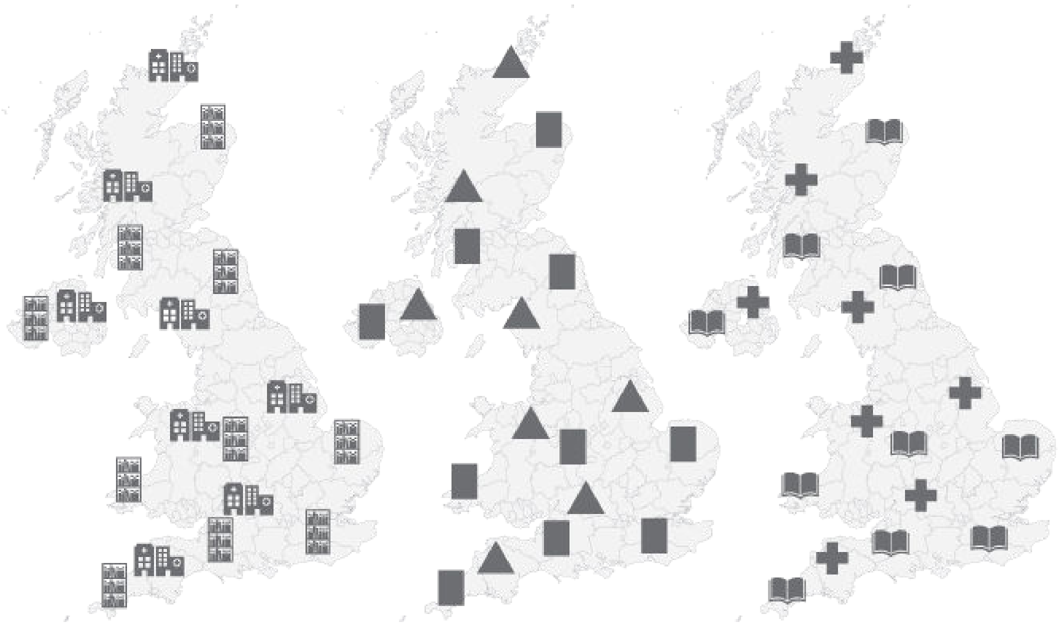
Most of us find the color boundary (between red and blue) in the left figure more quickly than the shape boundary (between circles and squares) in the right figure, even though both figures have another factor that’s randomly changing and interfering with our ability to detect the boundary. In the left image, the objects are organized by color while the shape is randomly changing. The reverse is true in the right image: the objects are organized by shape while the color is randomly changing.

All of this brings us to the question of whether or not it’s a good idea to use icons or pictograms in our visualizations because the simplest icons are defined by their form, not color. Luckily for us, the chapter on The Importance of Color, Font, and Icons offers some wisdom:

> “ *The rule should be: if you can make something clear in a small space without using an icon, don’t use an icon. […] As soon as anyone asks, “What does this icon mean?” you’ve lost the battle to make things ruthlessly simple.*

Icons can be a clever and useful way to illustrate an idea while making your infographic more attractive, but in order to be effective, they have to strike the right balance between detail and clarity. Too much detail leads to icons that are hard to distinguish from each other; too little detail leads to icons that are too abstract. Both overcomplicated and oversimplified shapes can lead to a confused audience.

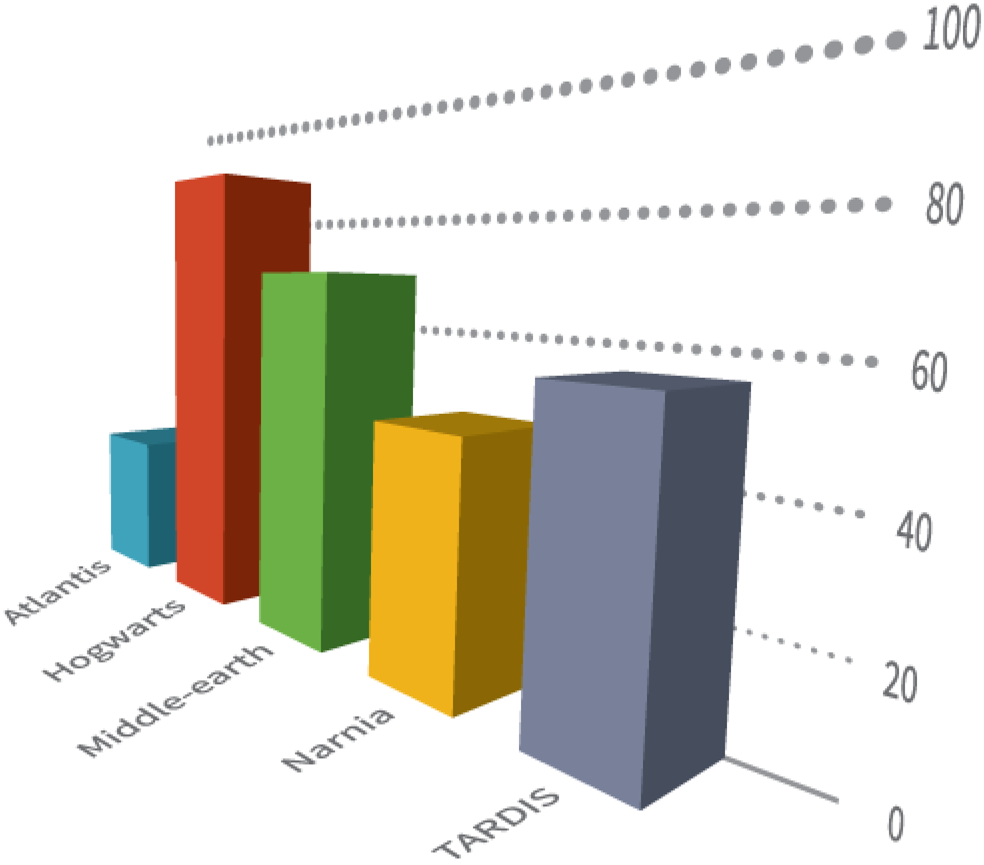
Now you be the judge. Good icons can communicate an idea without the benefit of color. Each of the three maps below shows the locations of hospitals and libraries in an area. Based solely on the icons, which map is the quickest and easiest for you to understand?

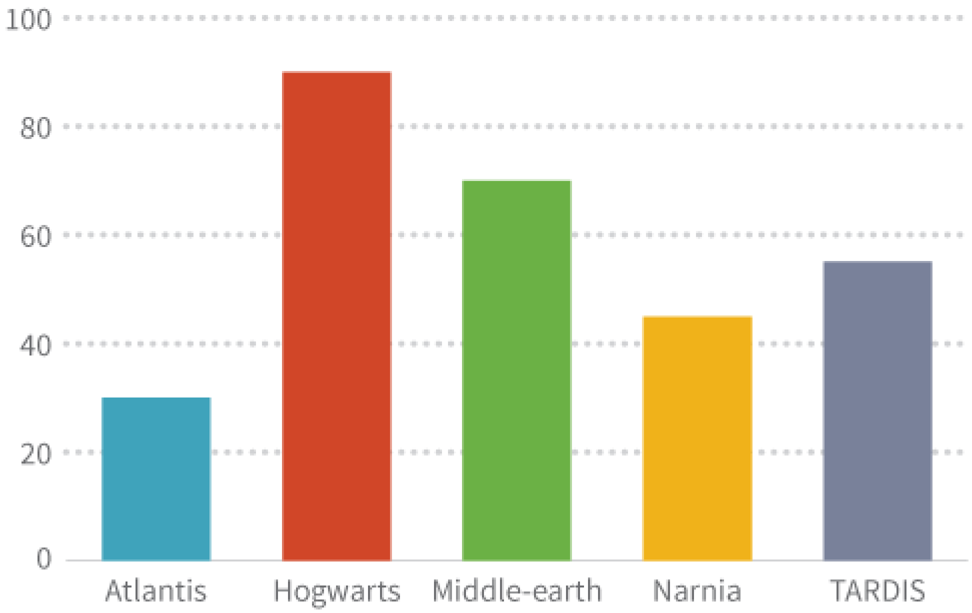


DIFFERENCES IN PERSPECTIVE

We live in a 3-dimensional world with 3D movies, 3D printing, and 3D games. It can be very tempting to add some of this 3-dimensional razzle dazzle to your 2-dimensional visualizations. Most of the time, though, people end up more dazed than dazzled by 3D graphs.

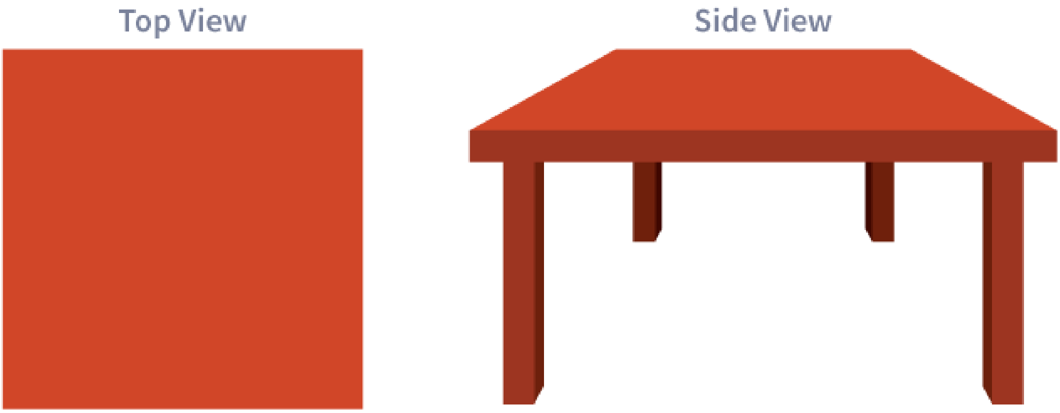
As mentioned earlier, our brains already have a hard enough time distinguishing 2-dimensional shapes from each other. Adding a third dimension to these forms often adds to confusion as well. But don't take our word for it. Which bar graph do *you* think is easier to read?


100
80
60
40
20
0
Atlantis
Hogwarts
Middle-earth
Narnia
TARDIS



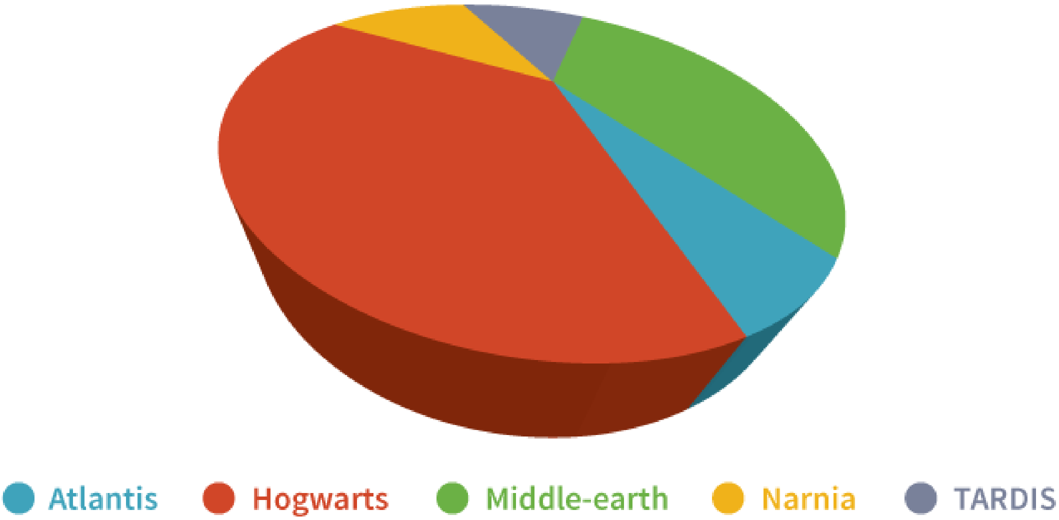
Harder-to-read graphs aren't the only reason to avoid 3D in your visualizations. The other reason is that our brains associate differences in perspective with differences in size. You've probably experienced this phenomenon many times, whether you're looking down a street or up at a skyscraper: objects in the foreground appear bigger while those in the background appear smaller.

For example, take a simple square table. You know that because it's a square, all four sides are the same size. But when you look at it from the side:

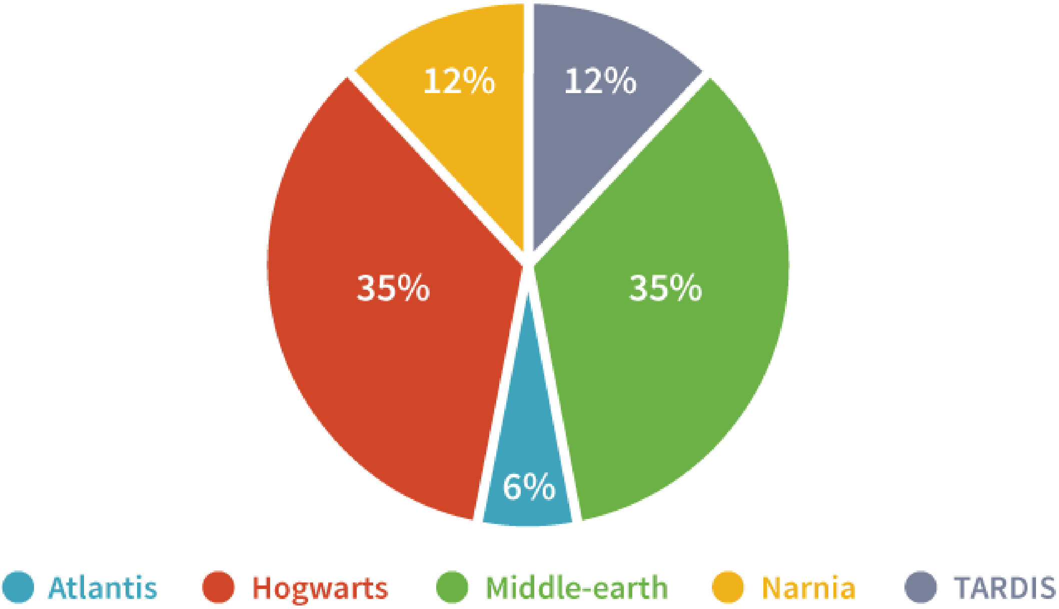


It seems like the side closer to you in the front is bigger than the side farther from you in the back.

For this reason, many types of 3D visualizations can be misleading. In the pie chart below, which slice do you think is bigger: the red slice or the green slice? the blue slice or the yellow slice?



In terms of which piece represents more data, the red slice and the green slice are actually the same size. As for the blue and the yellow slices, the yellow slice is bigger: in fact, it's twice as big as the blue slice. This is not clear from the 3D pie chart because of the way perspective distorts the size of the slices, but it is clear when we look at the same pie chart in 2D form:



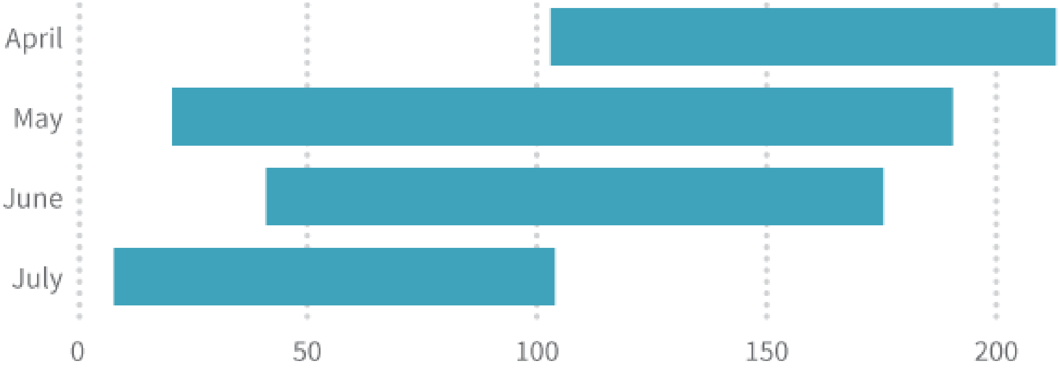
So unless you have a really, really, really good reason to use 3D graphs, it's better to avoid them. But just because 3D *graphs* aren't great doesn't mean 3D is bad in general. For example, 3D visualizations can be very useful in some types of scientific modeling, such as a 3-dimensional model of the brain that highlights active regions, or a 3-dimensional map that models the topography of an area. There are definitely instances when 3D enhances a visualization. Just make sure the 3D you add actually helps your audience see what you're trying to show instead of getting in the way.

How We Compare Differences

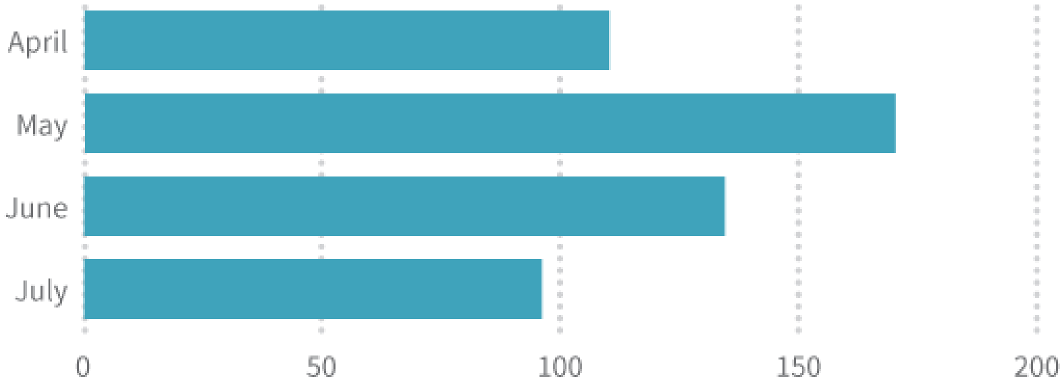
So what have we learned so far? That our brains are pretty good at seeing differences in color and contrast, but less good at interpreting differences in shape and perspective. Generally speaking, it takes us longer to understand a data visualization when there are more differences to process. So help a brain out and minimize the differences! One way to do this is to be *consistent* in your graphs.

CONSISTENT ALIGNMENT

Our brains are fairly good at comparing differences in length, but only when things share a common reference point. Cleveland, William S. and Robert McGill. "Graphical Perception and Graphical Methods for Analyzing Scientific Data." Science 229.4716 (1985): 828-833. Web. 16 June 2014. For example, try to compare how long the bars are in the following graph:



Now try to do the same thing for this graph:

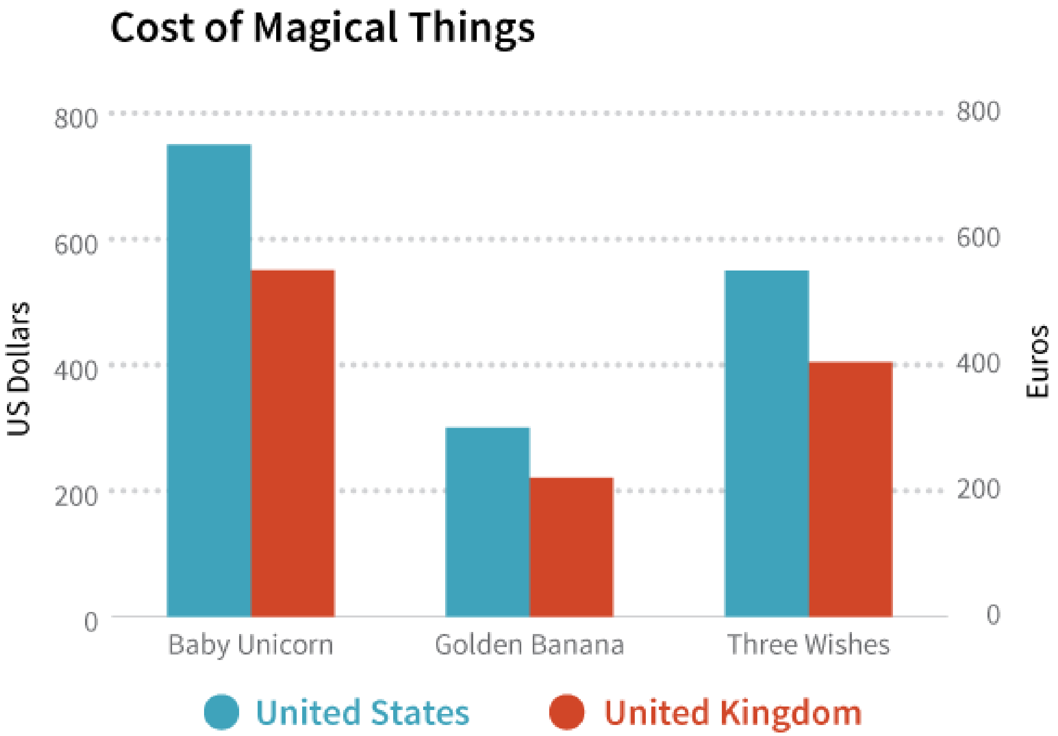


Wasn't it easier to compare the bars in the second graph? Even though both bar graphs have a horizontal axis with labels and tick marks, the second graph minimizes the differences our brains have to process by making all of the bars start from the same reference point. Consistent alignment allows our brains to focus on the lengths without having to worry about different starting points for each bar.

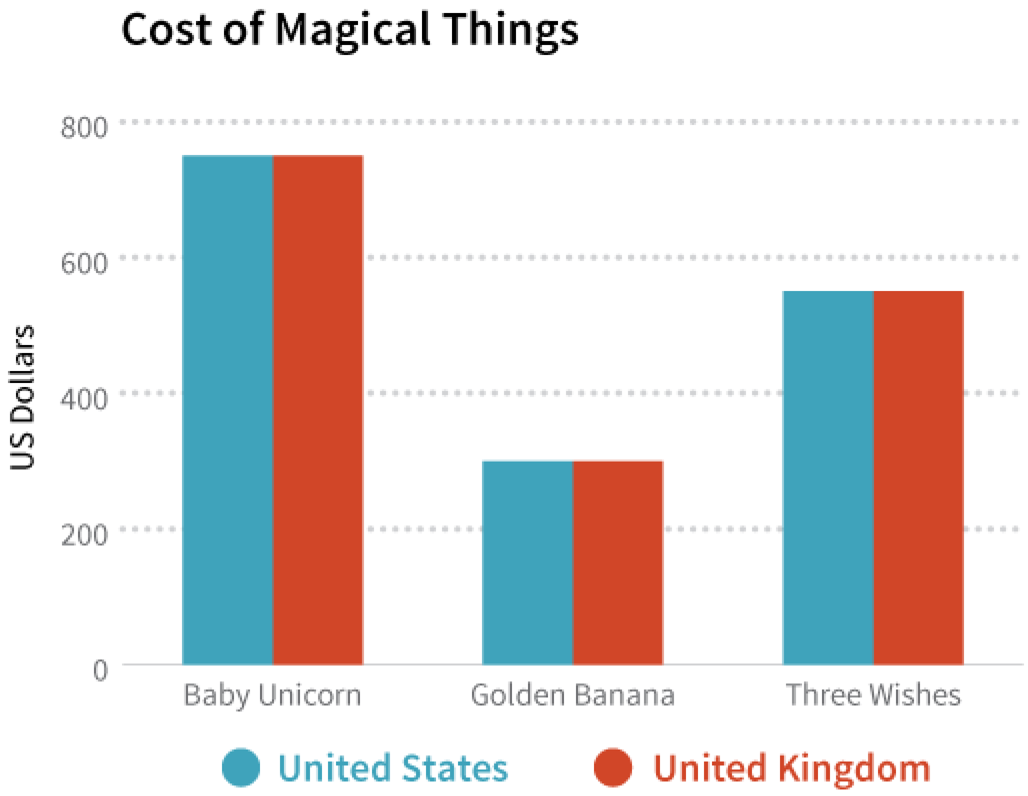
CONSISTENT UNITS

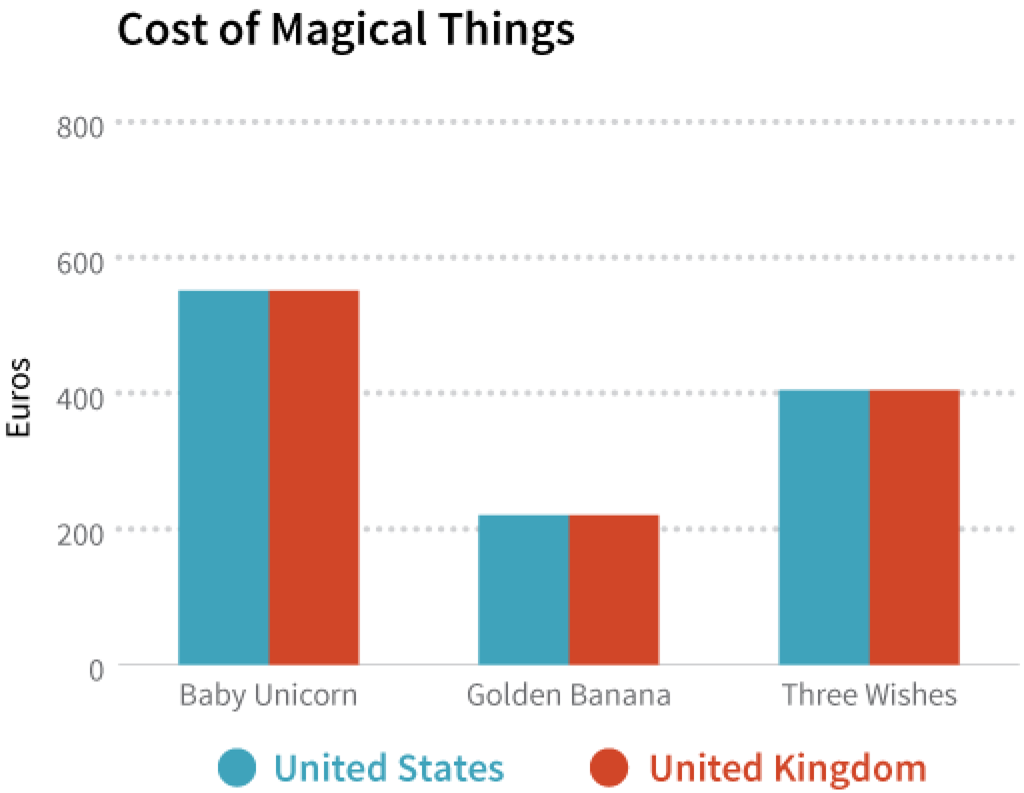
There may come a time when you want to show two sets of data in the same graph: something like a double line graph or a double bar graph, for instance. This is fine as long as you stick to the principle of consistency.

For example, the following bar graph compares how much it would cost to buy a few magical things in the United States versus the United Kingdom:



It seems like things are a bargain in the UK! But hold on because in this graph, it’s not enough to simply compare the bar lengths. There’s another difference we have to account for: the different units on the vertical axes. On the left side, the units are in US dollars. On the right side, the units are in euros. So in order to accurately compare prices, we need to either convert dollars to euros or euros to dollars. Some people can do this kind of currency conversion on the spot, and they are to be commended for their admirable math skills. The rest of us are going to look at that task and wonder why the maker of this bar graph is being so cruel to us. So don’t be cruel. Convert different units to consistent units so that your audience doesn’t have to. Like this:

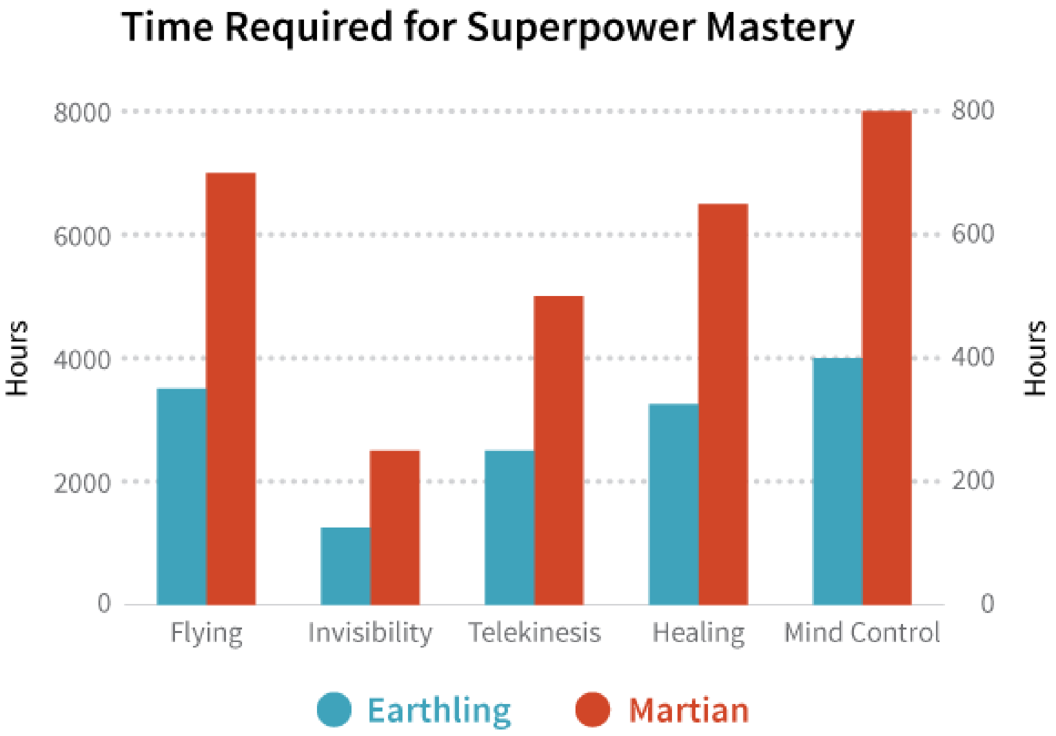

Cost of Magical Things
US Dollars
800
600
400
200
0
Baby Unicorn
Golden Banana
Three Wishes
United States
United Kingdom



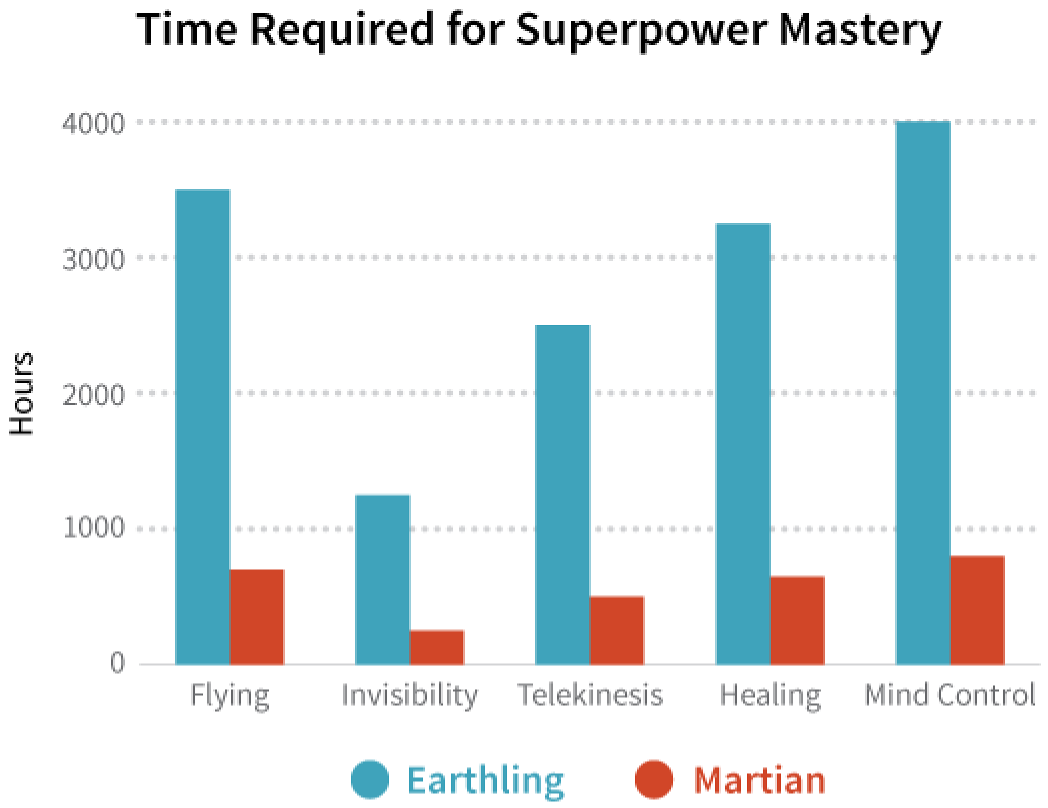
As it turns out, magical things cost the same in the US and the UK.

CONSISTENT SCALES

In addition to having consistent units, your graphs should have consistent scales. For example, the bar graph below compares how long it takes an Earthling and a Martian to master different superpowers:



The heights of the bars make it seem like Martians take longer to master superpowers, but again, the vertical axes are different. Even though the units are consistent this time (both axes use hours), the scales are different. On the left side, the scale is in thousands of hours for Earthlings. On the right side, the scale is in hundreds of hours for Martians. The same data with consistent scales would look like this:



A graph with consistent scales makes it visually obvious that Martians master superpowers much more quickly than Earthlings do. Maybe it's because they have less gravity on their planet.

Transparency

When it comes to data visualization and transparency, keep in mind that the data we show are just as important as the data we don't. We want to give readers enough context and any other additional information they may need to accurately interpret our graphs.

CUMULATIVE GRAPHS

A cumulative value tells you how much *so far*: it's a total number that adds up as you go along. For example, let's say you're keeping track of how many superhero sightings there are in your neighborhood:

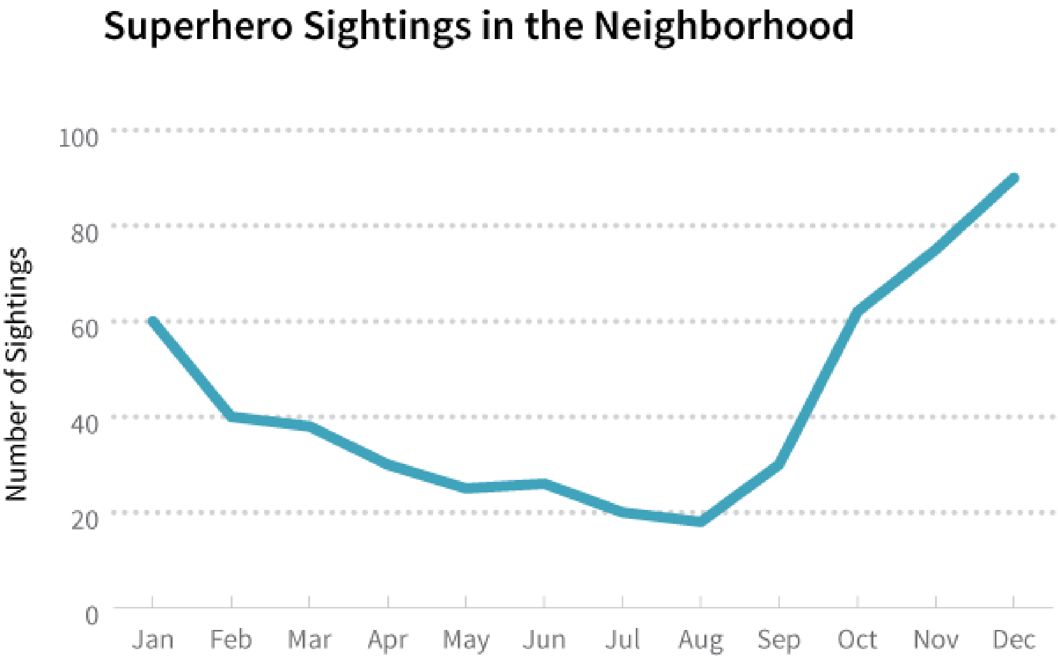
Month	Number of Superhero Sightings	Cumulative Number of Superhero Sightings
January	60	60
February	40	100 = 60 + 40
March	38	138 = 60 + 40 + 38
April	30	168 = 60 + 40 + 38 + 30
May	25	193 = 60 + 40 + 38 + 30 + 25
June	26	219 = 60 + 40 + 38 + 30 + 25 + 26
July	20	239 = 60 + 40 + 38 + 30 + 25 + 26 + 20
August	18	257 = 60 + 40 + 38 + 30 + 25 + 26 + 20 + 18
September	30	287 = 60 + 40 + 38 + 30 + 25 + 26 + 20 + 18 + 30
October	62	349 = 60 + 40 + 38 + 30 + 25 + 26 + 20 + 18 + 30 + 62
November	75	424 = 60 + 40 + 38 + 30 + 25 + 26 + 20 + 18 + 30 + 62 + 75
December	90	514 = 60 + 40 + 38 + 30 + 25 + 26 + 20 + 18 + 30 + 62 + 75 + 90

The first column of values gives you the number of sightings in each month, independent of the other months. The second column gives you the number of sightings *so far* by adding the numbers as you go down the column. For example, in January there were 60 superhero sightings, so the total so far is 60. In February there were 40 sightings, which means the total so far is 100 because the 40 sightings in February get added to the 60 sightings in January. And so on and so forth for each month that follows.

Why are we doing all this math? Because there's something called a cumulative graph, which shows cumulative values.

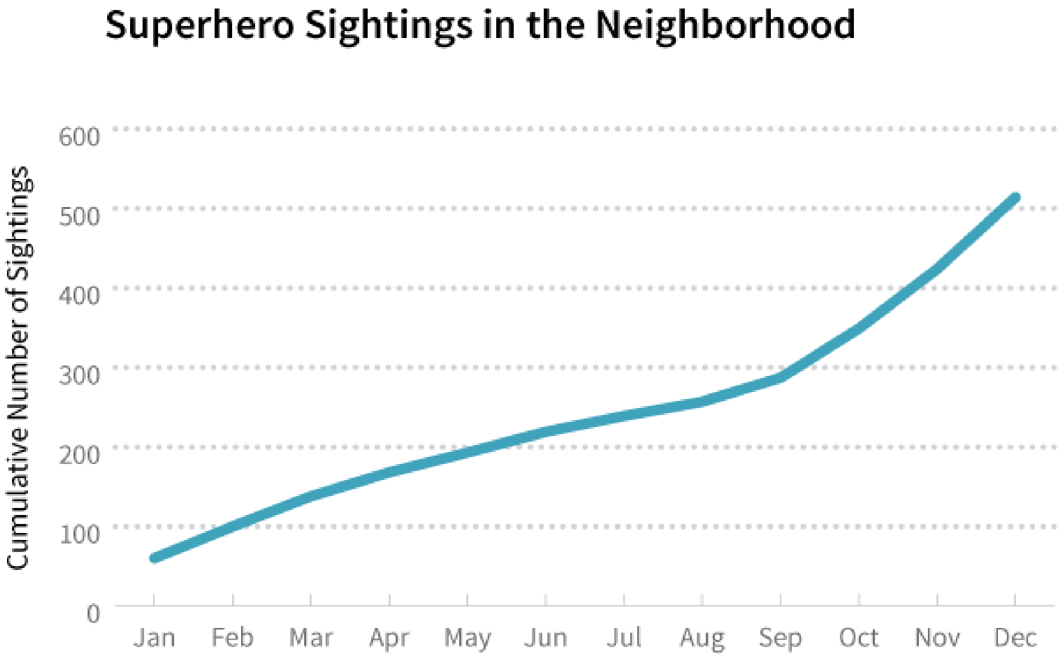
If you're going to show people a cumulative graph, it's important that you tell them it's a cumulative graph.

Why? Because this is what a regular line graph of the monthly, non-cumulative values looks like:



Notice how the line dips and rises: the number of monthly sightings generally decreases through August, then increases until the end of the year. Superheroes are apparently busier during the holiday months.

Compare the regular line graph above with the cumulative graph below:

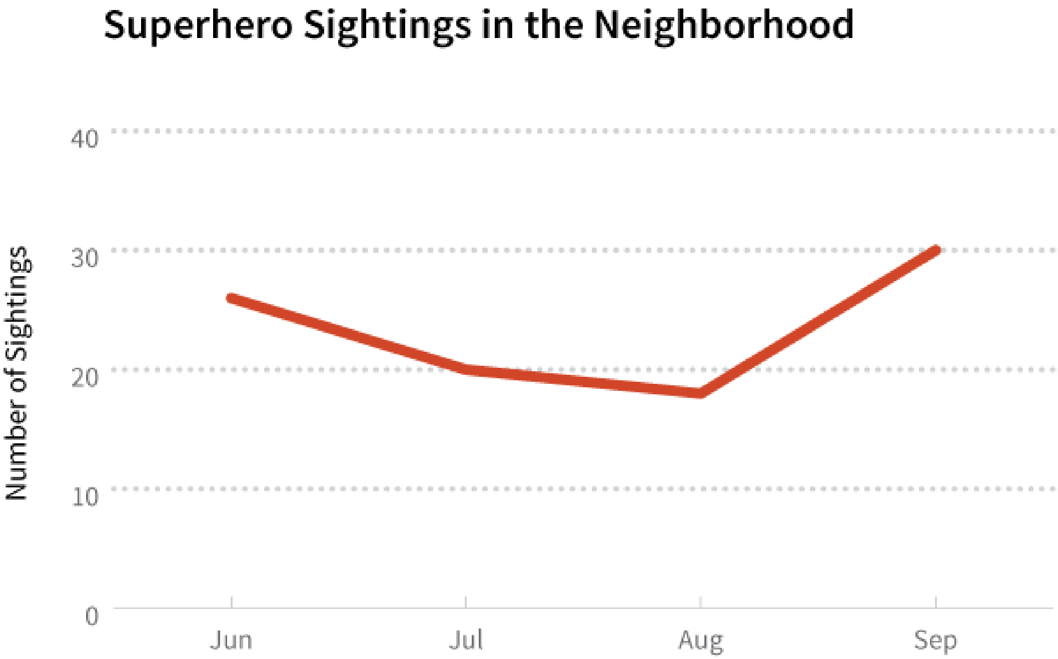


Do you see how the general decrease in superhero activity from January through August is not obvious in the cumulative graph? The line in the cumulative graph keeps going up because the numbers keep adding up and getting bigger. This upward trend masks other patterns in the data. It's easy to look at a cumulative graph and mistakenly think it's showing a trend that consistently increases when that's not really the case.

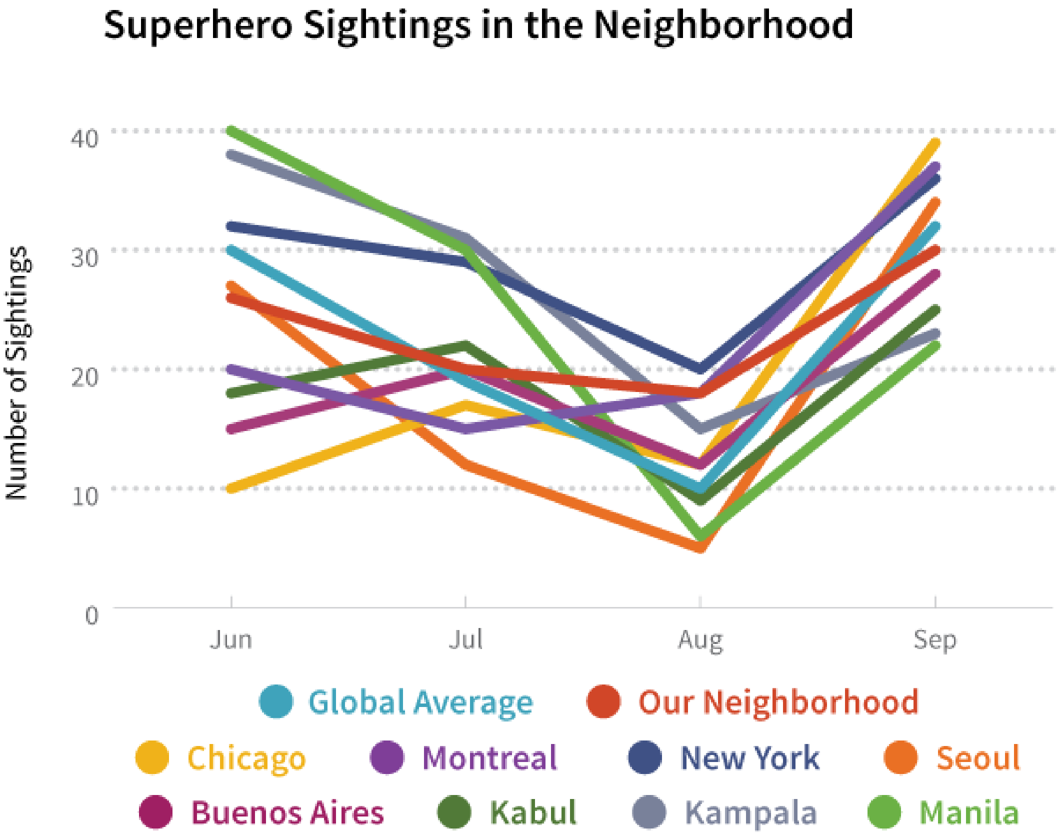
So when it comes to cumulative graphs, if you show it, let them know it, or else people might think the numbers are going up, up, and away.

CONTEXT

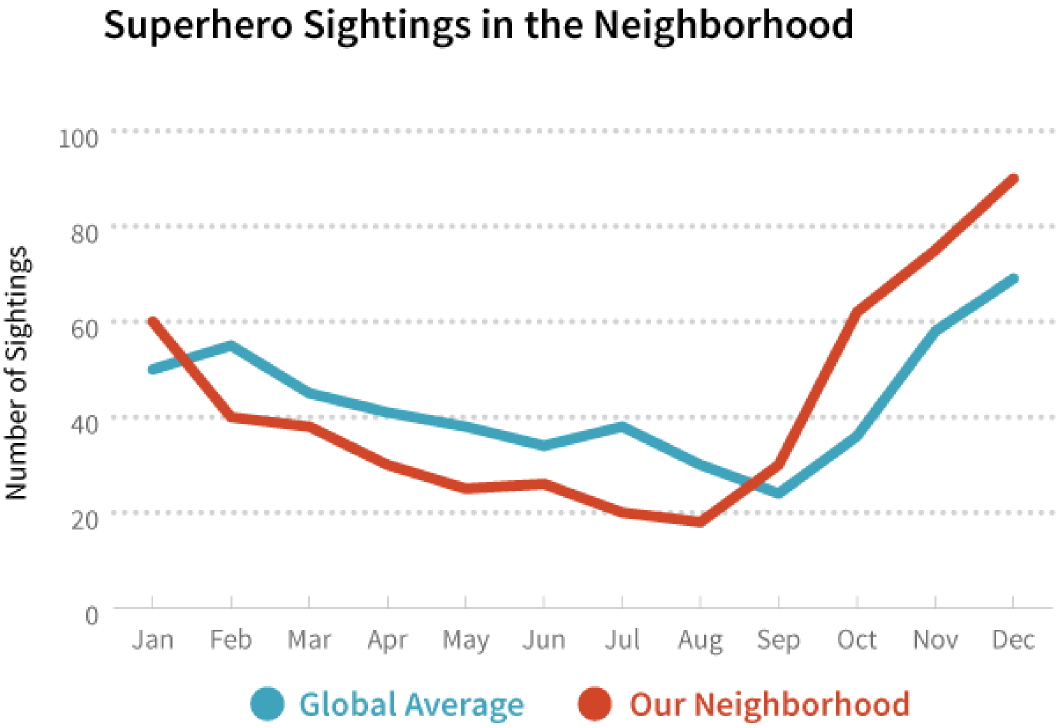
Without context, the stories our graphs tell have no point of reference that we can use to make a comparison. For example, is the following trendline average or unusual?



Alternatively, too much data can overwhelm an audience. In this case, the context gets lost among all the information and instead becomes an overcomplicated truth. For example, isn't it harder to pick out the red trendline in the graph below?



This is why data visualization is challenging! You have to figure out how to present your data in a way that allows readers to clearly see trends and accurately compare differences. For example, let's take a look at the following graph:



If you want to show your neighbors how the number of superhero sightings has changed over time, then presenting data that runs the course of a year instead of a few months provides more context. And instead of showing trendlines for dozens of cities around the world, your graph will be less cluttered and more relevant to your audience if you just show the trendline for your neighborhood. Finally, it's a good idea to include some point of reference—like a trendline for the global average—so that your neighbors get a sense of how typical or atypical their neighborhood is.

The next and last chapter of this book goes into further detail about the importance of context and points out other ways graphics can mislead an audience. Just remember: being consistent and transparent goes a long way in your visualizations.

CHAPTER 18

COMMON VISUALIZATION MISTAKES

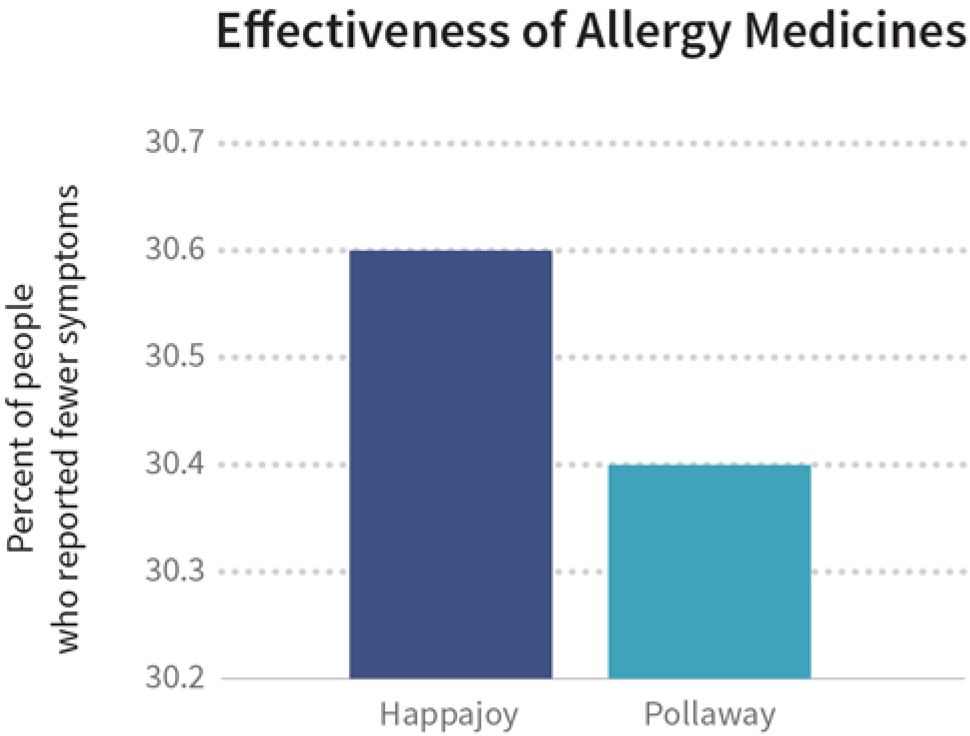
BY KATHY CHANG, KATE EYLER-WERVE, AND ALBERTO CAIRO

Welcome to the last (but certainly not least) chapter of the book! We hope you've learned enough to appreciate how much good data and good design can help you communicate your message. Professional data visualizers get excited by the stories they want to tell as well, but sometimes they forget to follow some best practices while doing so. It happens to the best of us. So in this chapter, we're going to cover what those best practices are.

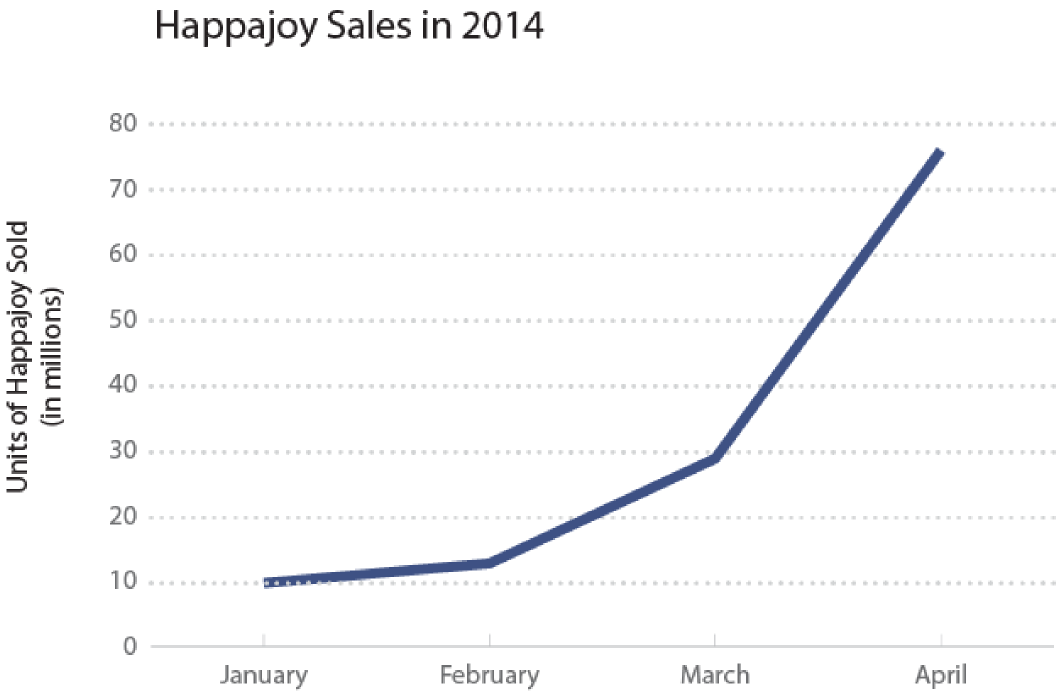
Don't Truncate Axes

One of the ways a graph can be distorted is by truncating an axis. This happens when an axis is shortened because one or both of its ends gets cut off.

Sometimes a distortion like this is really obvious. For example, let's say there are two allergy medicines called Happajoy and Pollaway. The bar graph below compares how effective these two medicines are at reducing the tearful, congested misery known as allergy symptoms. If you quickly glance at the bars, you may think that Happajoy is twice as effective as Pollaway is because its bar is twice as tall. But if you examine the graph more closely, you'll see that the y-axis is truncated, starting from 30.2 and going up to only 30.7 percent. The truncated y-axis makes the difference between the two bars look artificially high. In reality, Happajoy's effectiveness is only 0.2% higher than Pollaway's, which is not as impressive as the results implied by the bar graph.

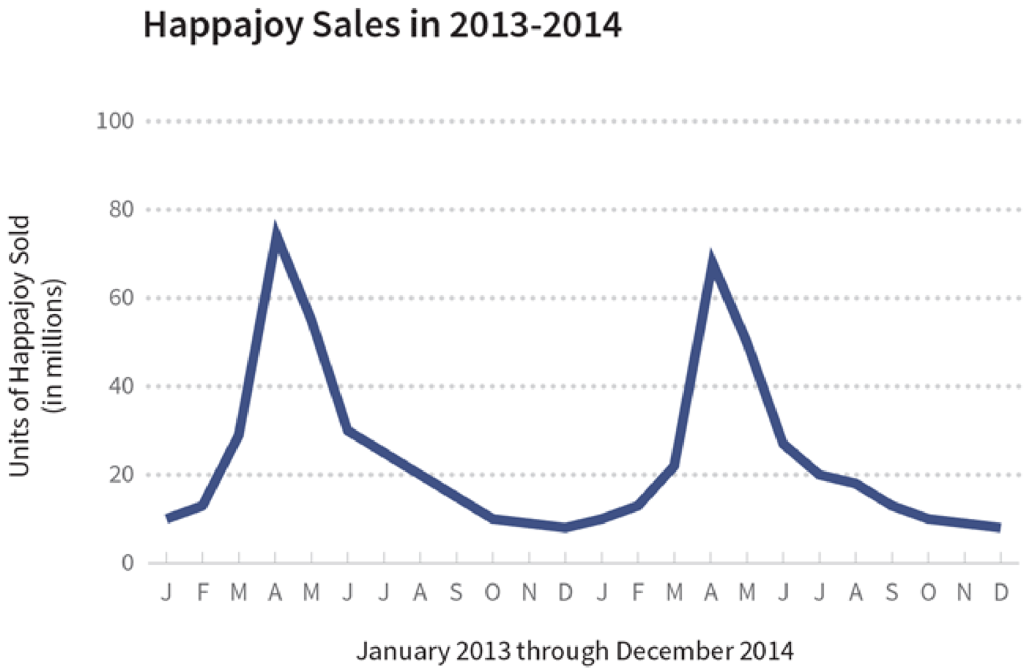


Sometimes a truncated axis and the resulting distortion can be more subtle. For example, the next graph shows the quantity of Happajoy sold from January through April 2014.



At first glance, there doesn't appear to be a truncation issue here. The y-axis starts at zero, so that's not a problem. The critical thing to understand is that it's the x-axis that's been truncated this time: we're seeing sales from less than half the year. Truncating a time period like this can give the wrong impression, especially for things that go through cycles. And—you guessed it—the sale of allergy medicine goes through a seasonal cycle since allergy symptoms are typically higher in the spring and lower in the winter.

What would be a better way to show sales of Happajoy? Something like the graph below:



This graph shows the same dataset of Happajoy sales, except this time the y-axis is proportional and the x-axis covers two full years instead of just a few months. We can clearly see that sales of Happajoy went down in the winter and up in the spring, but that the rate of sales didn't change much from year to year. In fact, sales were a little lower in 2014 than in 2013.

When you compare the last two graphs of Happajoy sales, do you see how different their stories are? If you were an investor in Happajoy's company and you saw the graph with truncated axes, you might dance happily through the streets because it seems like the company is doing really well. On the other hand, if you saw the graph with proportional axes, you might reach for some aspirin instead of Happajoy because of the headache you'd get worrying about the overall decrease in sales.

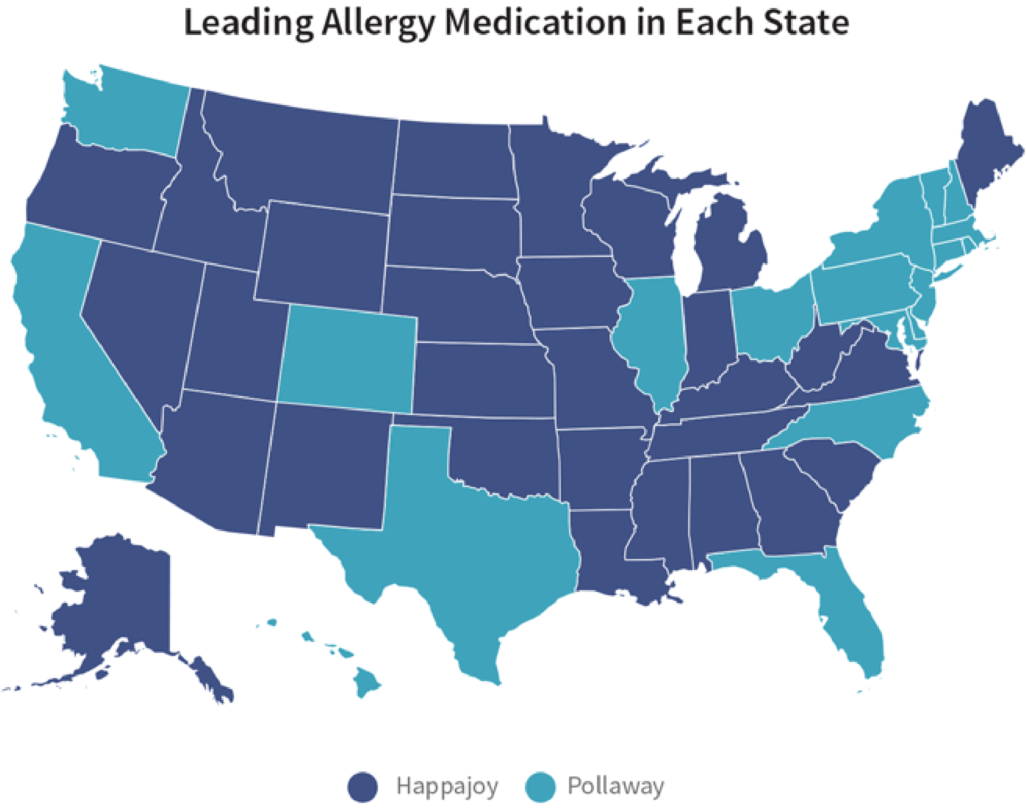
So watch out for truncated axes. Sometimes these distortions are done on purpose to mislead readers, but other times they're the unintentional consequence of not knowing how truncated axes can skew data.

Don't Omit Key Variables

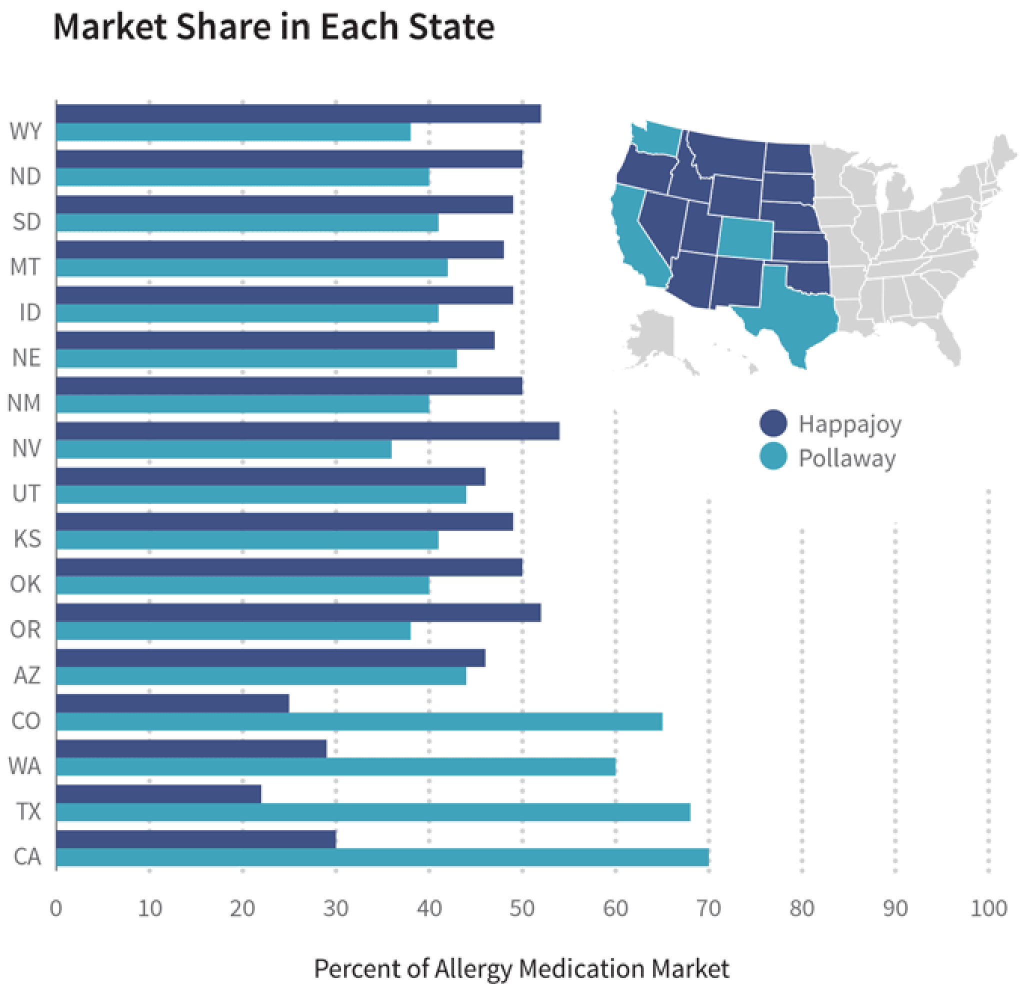
> “ *The first principle is that you must not fool yourself — and you are the easiest person to fool.*
>
> *- Richard Feynman, 1974 Caltech Graduation Address*

You know how famous people are sometimes criticized for something they said, and they often reply that they were quoted out of context? Context is important, especially when it comes to data. It's very easy to fool yourself by leaving out variables that could affect how you interpret the data. So whenever you're examining a variable and its relationships, carefully consider the ecosystem in which that variable exists and deliberately seek out other variables that could affect the one you're studying.

This is easier said than done. For example, the map below shows each state's market leader in allergy medicine: Happajoy is the leader in dark blue states, while Pollaway is the leader in light blue states. On the surface, it might seem like Happajoy is the market leader nationally, ahead of Pollaway. But to get the complete picture you have to pay attention to other variables.



For example, the bar graph below shows a breakdown of market share in each state. (We're only going to look at the western continental states for now.) The margins by which Happajoy leads are significantly less than the margins by which Pollaway leads.



Combine the information from the bar graph with the table below. The total sales in states where Happajoy is the leader is also significantly less than the total sales in states where Pollaway is the leader. When you add up the numbers, Pollaway's total sales are more than twice that of Happajoy's. Assuming that a similar pattern holds for the rest of the country, would it still be accurate to say that Happajoy is the national market leader in allergy medicine?

States	Happajoy	Pollaway
Wyoming (WY)	299,734	219,037
North Dakota (ND)	349,814	279,851
South Dakota (SD)	408,343	341,675
Montana (MT)	482,400	422,100
Idaho (ID)	782,040	654,360
Nebraska (NE)	872,320	798,080
New Mexico (NM)	1,043,000	834,400
Nevada (NV)	1,489,860	993,240
Utah (UT)	1,313,300	1,256,200
Kansas (KS)	1,414,140	1,183,260
Oklahoma (OK)	1,907,500	1,526,000
Oregon (OR)	2,027,480	1,481,620
Arizona (AZ)	3,014,380	2,883,320
Colorado (CO)	1,297,000	3,372,200
Washington (WA)	2,069,100	4,138,200
Texas (TX)	5,733,200	17,720,800
California (CA)	7,608,000	26,628,000
Sales Totals	32,111,611	64,732,343

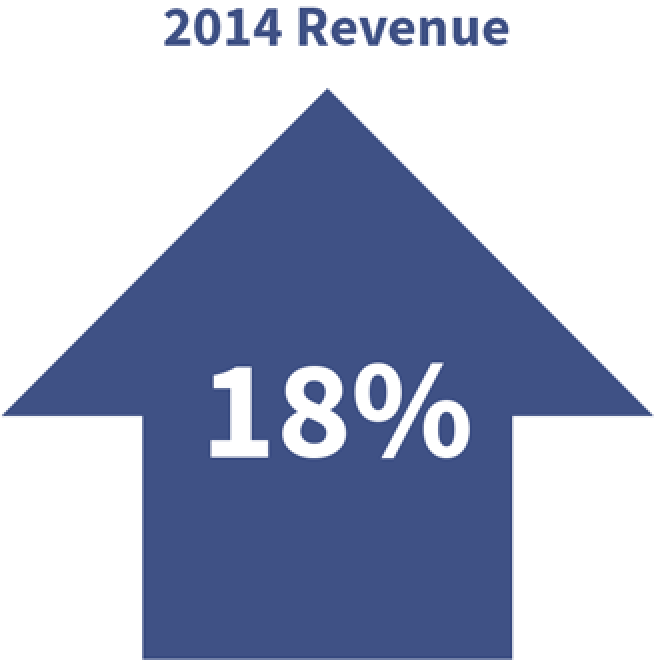
The lesson here is that if you want to provide a fair picture of what's going on, you have to expand your scope to show the variables that put things in their proper context. That way, you can provide your readers with a more complete and nuanced picture of the story you're trying to tell.

Don't Oversimplify

Life is complicated, right? Data can be complicated, too. Complicated is hard to communicate, so it's only natural to want to simplify what your data are saying.

But there is such a thing as simplifying too much. Oversimplifying is related to the previous point about not expanding the scope enough to provide a clear picture of what's going on.

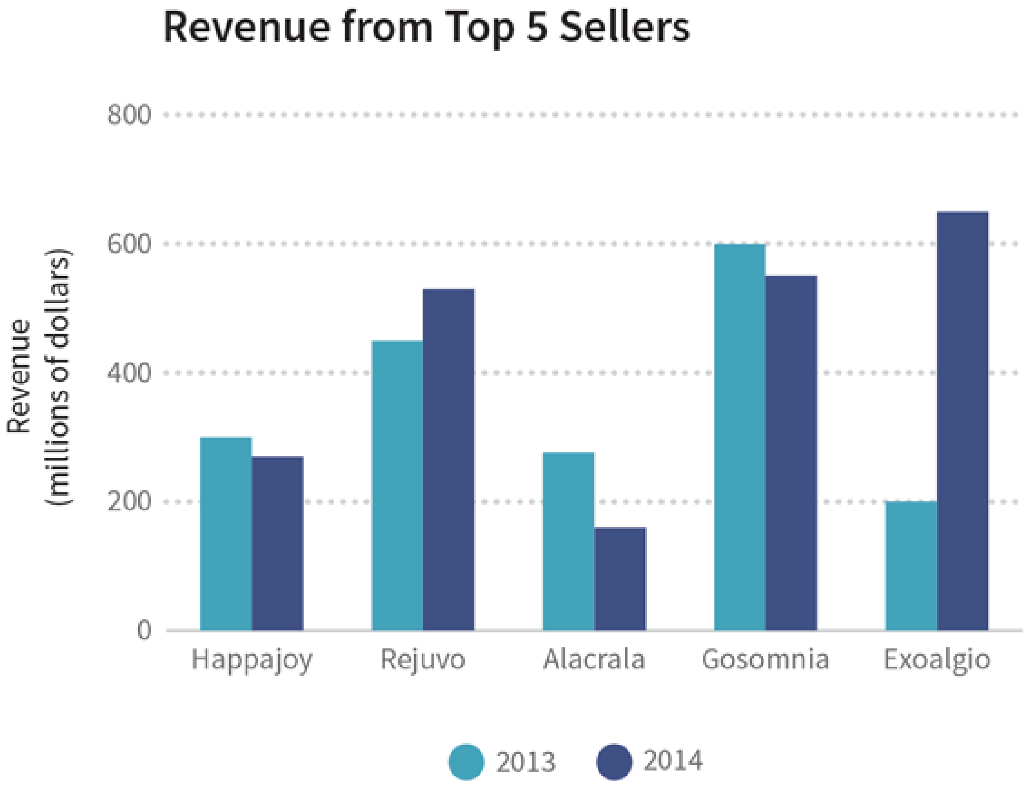
For example, let's say you're an investor of RediMedico, the maker of Happajoy, and you attend the annual sales meeting. The CEO of RediMedico starts off the presentation with the following graphic:



Now the investor in you might look at that graphic and start daydreaming about all the wonderful things you'll do with such a great return on your investment. But then the data pro in you kicks in, and you start wondering about what that 18% increase means. You ask yourself:

- Compared to what?
- Compared to when?
- Compared to whom?

These are all worthwhile questions to answer with a visualization! Thankfully, the CEO of RediMedico agrees and presents the next graphic, which compares the revenues from the five top-selling medicines RediMedico makes:



If we do some number-crunching, we see that the average increase in revenue between 2013 and 2014 is indeed 18%. However, we also see that this increase is primarily due to a whopping 225% increase in revenue from a single medicine, Exoalgio. Revenue from 3 out of 5 medicines actually dropped. So RediMedico's first graphic tells part of the truth, while the second graphic tells the whole truth by presenting the details behind the single number.

Using a graphic with a single number and no breakdowns is like writing a news headline without the news story. Keep the headline—that revenue improved by 18%—and then provide the context and the background to flesh out the full story.

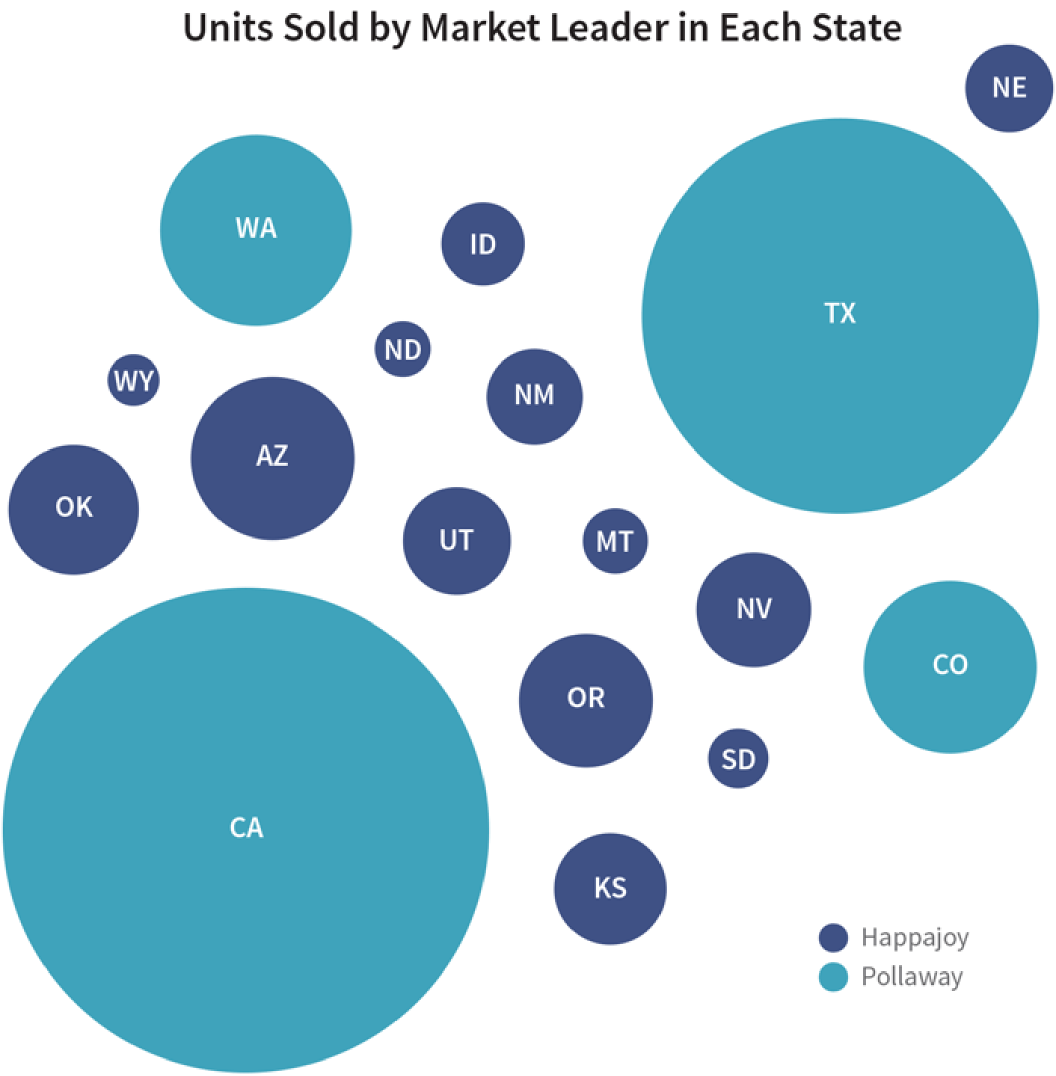
Try to be true to the underlying complexity by digging deeper into the data and providing readers with a better understanding of the numbers you're presenting.

Don't Choose the Wrong Form

Creating a data visualization is a balancing act between form and function. When choosing a graphic format for your data, you'll have to figure out how to effectively communicate to your audience in an aesthetically pleasing way. This may seem like a daunting task, but fear not! There's actually a lot of research that can help us with this. In fact, you've already been introduced to some of this research in the previous chapter: Cleveland and McGill's "Graphical Perception" paper, in which they rank how well different graphic forms help people make accurate estimates.

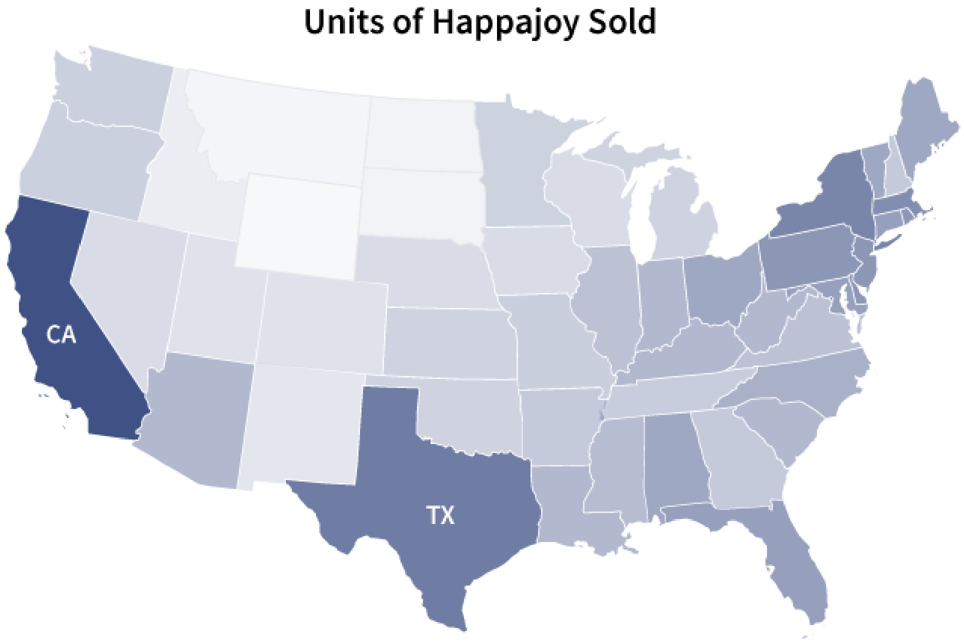
You can use this scale to help you choose the best graphic representation of your data. Area and shading are good at giving readers a general overview that helps them understand the big picture. Graphs with a common baseline (such as bar graphs) are good at helping readers make accurate comparisons.

Since we've already looked at examples of bar graphs and line graphs in this chapter, let's take a look at a couple graphics that use area and shading.

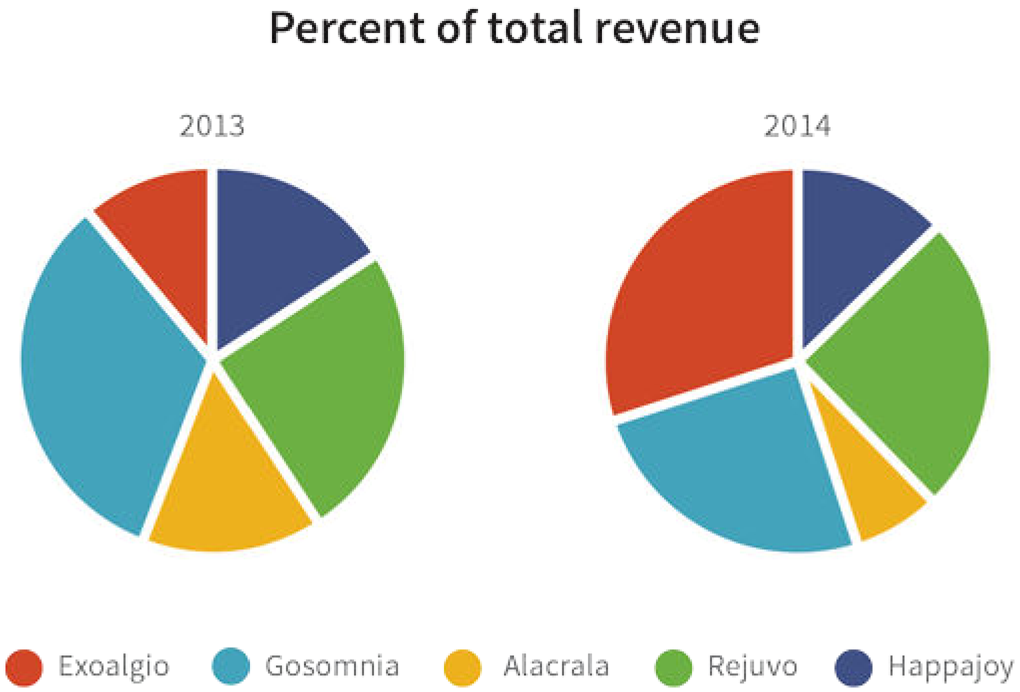


The bubble graphic uses area to display the units sold of the top selling allergy medicine in some states. Based on bubble size, you can generally tell that more Happajoy was sold in Arizona than in New Mexico. But can you tell by how much? Is the Arizona bubble three times bigger than the New Mexico bubble? Four times? It's hard to tell. It's even harder to tell when the bubble sizes are closer together: Did Utah or Kansas sell more Happajoy?

We run into the same problem with the next graphic, which uses shading to represent Happajoy sales: California is definitely darker than Texas, but how much darker? Two times? Three times? Who knows? This is why area and shading are better for giving an overall picture instead of making precise comparisons.



In addition to area and shading, angles also give us a tough time when it comes to making accurate estimates. This is why it's so hard to compare two pie charts, as in the example below.



It’s already hard to precisely compare the slices within the *same* pie chart. It’s even harder to compare slices across different pie charts. If the goal of this graphic is to help readers compare revenues from one year to the next, then something like a bar chart would have been a better choice.

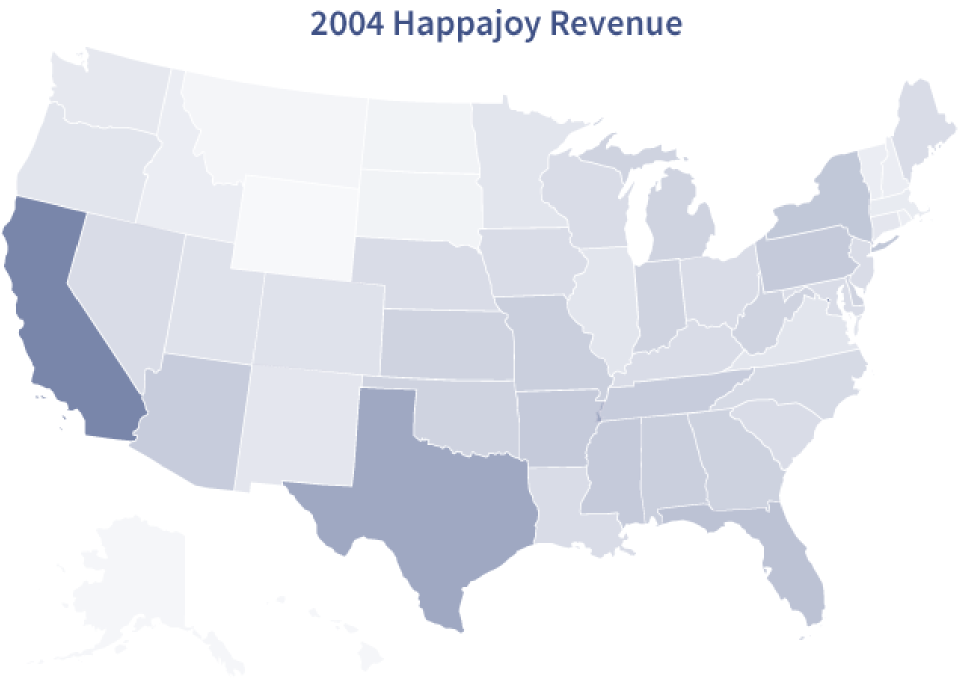
That’s the key thing: think about which graphic forms will best facilitate the tasks you want your readers to do.

Do Present Data in Multiple Ways

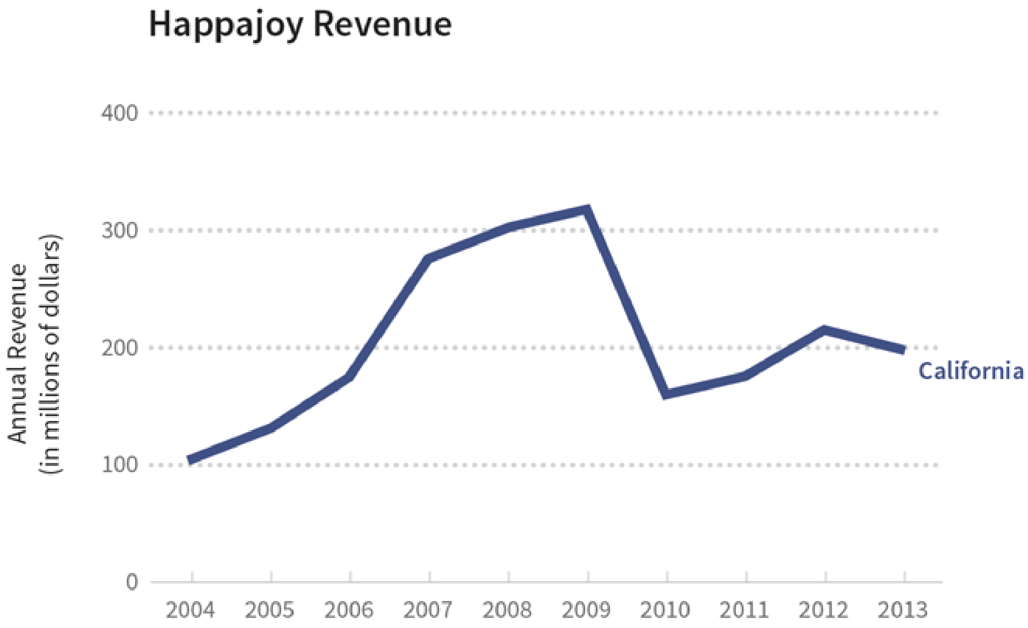
We just covered how different graphic forms are good at doing different things. So what do you do when you have a lot of data and you want to show different aspects of those data, but you also don’t want to overwhelm your audience with an overly complicated graphic? One way to deal with this challenge is to present your data in multiple ways. You can individually show multiple graphics, each one showing a different aspect of the data, so that taken together your audience gets a more accurate picture of the data as a whole.

For example, let's say the CEO of RediMedico wants to show investors how Happajoy has been selling in the United States since it was first introduced to the market ten years ago. The available data consists of Happajoy sales in every state for every year since 2004. You're the lucky data pro who gets to figure out how to present both the big picture and the small picture inside this data.

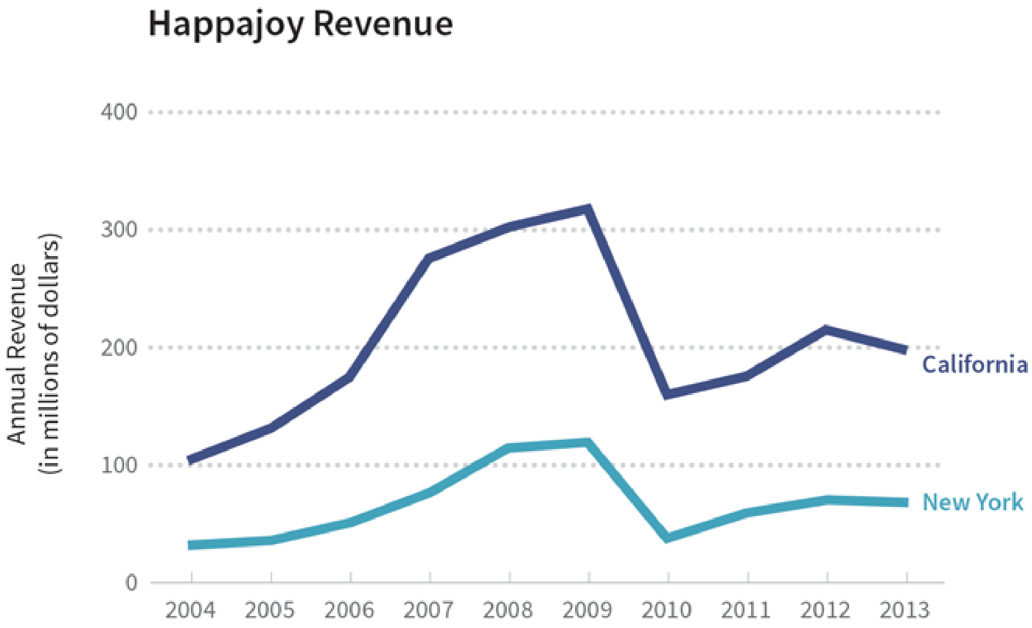
Let's start with the big picture. Remember how graphic forms that use area or shading are best at giving a general overview? For every year that Happajoy has been on the market, a map that uses shading to represent sales can give people a general sense of how sales have changed across time and location:



Now let's move on the small picture. Let's say RediMedico started to advertise heavily in California and New York a few years ago, and the investors are wondering how sales in those states are doing. Using the same dataset, you can give a more detailed picture of the sales in one state:



Or you can compare the sales between different states:

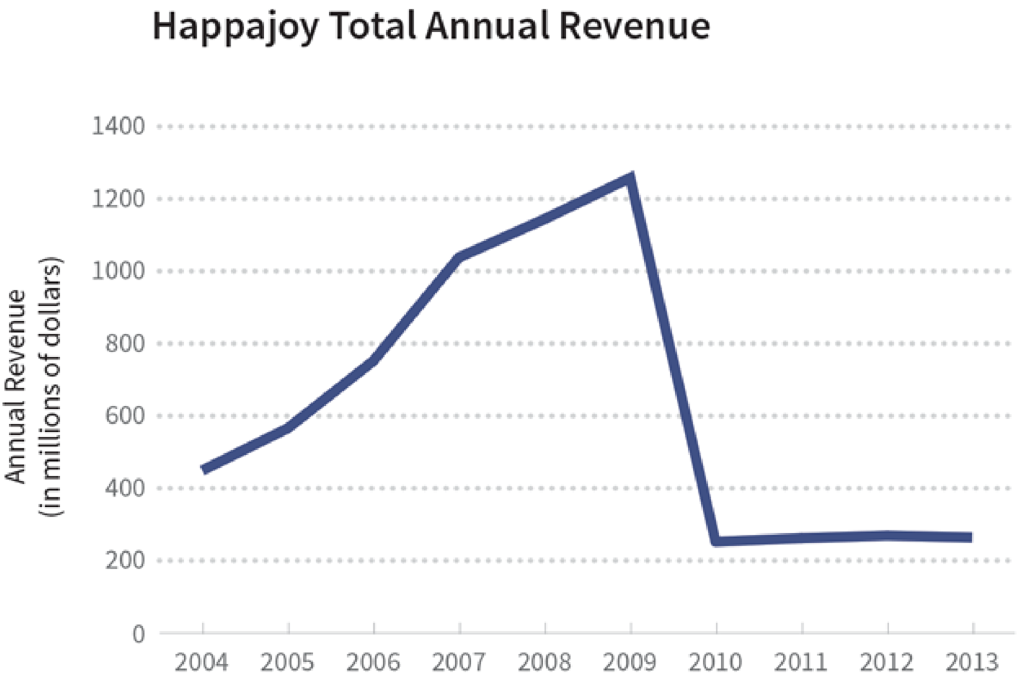


See? Same data, different presentations. So if you want to show your readers different sides of the same story, give them multiple graphic forms.

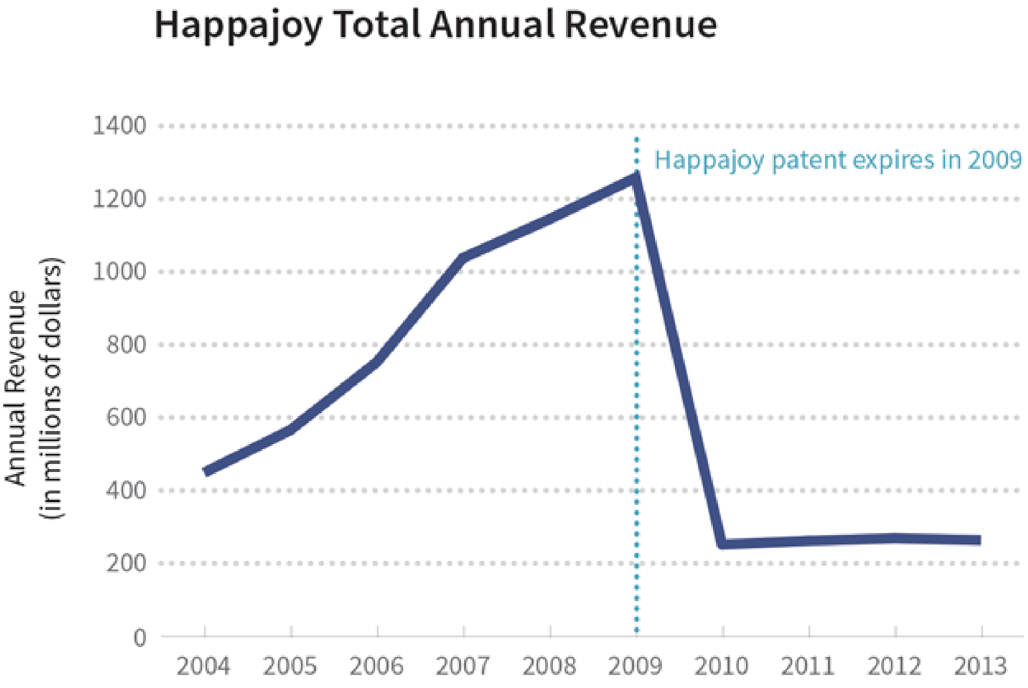
Do Annotate with Text

They say that a picture is worth a thousand words, but that doesn't mean you should forget about words entirely! Even your most beautiful and elegant visualizations can use text to highlight or explain things to your audience. This is especially useful when you're presenting multiple layers of information because your annotations can help readers connect the various pieces into an understandable whole. And obviously, you're not reading this book in order to make boring visualizations, right? Good visualizations engage an audience, so adding text is a great way to address questions that may come up as curious readers examine your graphic.

For example, let's go back Happajoy sales. If you see a graphic like the following:



Then you might be wondering what happened between 2009 and 2010. Why was there such a sharp drop in revenue? In this case, it would be helpful to add some text:



So whenever you create a visualization, think about the “So what?” of your graphic: Why should people care about the information you’re presenting? Add annotations to help them understand why they should care. Write a good headline that catches their attention, a good introduction that highlights interesting data points, and a good narrative that structures your visualization logically. Good writing is an important part of good visualizations.

Case Study of an Awesome Infographic

To close out this chapter, let’s take a look at all of these pro tips in action by going through a visualization made by real data pros: an infographic by The New York Times about breast cancer (http://www.nytimes.com/2013/10/16/health/uganda-fights-stigma-and-poverty-to-take-on-breast-cancer.html?_r=0#g-graphic). The designers organized the information as a narrative with a step-by-step structure. This is an interactive graphic, so it’s best if you click through the link to get the full experience.

On the first screen, you see a bubble graphic that gives you a general sense of which countries have the most new cases of breast cancer . After clicking “Begin”, you see a scatterplot with proportional axes. The scatterplot shows that there is an inverse correlation between breast cancer detection and mortality: as more women are detected with breast cancer, fewer women die from it. A scatterplot is a good way to show correlations between variables, and a bubble graphic is a good way to show a general overview, so the designers chose graphic forms that matched well with what they wanted to show.

Notice how the designers use text to write a good headline that grabs the reader’s attention (“Where Does Breast Cancer Kill?”) and to highlight another aspect of this scatterplot—that highly developed countries have higher diagnosis rates (and lower mortality rates) while the opposite is true for the developing world. As you keep clicking “Next”, the designers guide you deeper into the scatterplot by highlighting a cluster of countries and providing an annotation that gives you further insight into that cluster. Without these notes, we would be left with a relatively straightforward scatterplot that leaves little room for us to explore the data further. By providing useful and well-placed annotations, the designers help us see relationships that we otherwise may have missed.

The designers also present the data in multiple ways. They use color to add another layer of detail: the development status of various countries. In addition, if you’re curious about the statistics for a specific country, you can mouse over that country’s dot to get those numbers.

Finally, by adding useful annotations and showing the data in multiple ways, the designers present the data within a context that doesn’t leave out important variables or oversimplify. By following through on some good data visualization practices, the designers created a clear, balanced, and engaging infographic.

RESOURCES A

Chapter 3: Intro to Survey Design

Online survey sites

Kwik Surveys, http://www.kwiksurveys.com

Free Online Surveys, http://www.freeonlinesurveys.com

Google Forms, http://www.google.com/google-d-s/createforms.html

Survey Gizmo, http://www.surveygizmo.com

Survey Monkey, https://www.surveymonkey.com

Survey Planet, https://www.surveyplanet.com

Chapter 4: Types of Survey Questions

Conducting Surveys

Dierckx, Didier. “The Case of ‘Likert Scales v. Slider Scales,' *Market Research* (blog) Checkmarket.com, https://www.checkmarket.com/2012/08/likert_v_sliderscales/ (http://http://stattrek.com/regression/linear-transformation.aspx#)

Fink, Arlene. *How to Conduct Surveys: A Step-by-Step Guide,* 3rd ed., Thousand Oaks, California: SAGE Publications, Inc., 2006.

Fowler, Floyd J. *Survey Research Methods*, 5th ed., Thousand Oaks, California: SAGE Publications, Inc., 2014.

Gray, George and Neil Guppy. *Successful Surveys: Research Methods and Practice,* 3rd ed., Scarborough, Ontario: Thomson Nelson, 2003.

"Scaled Questions," Instructional Assessment Resources, The University of Texas at Austin, http://www.utexas.edu/academic/ctl/assessment/iar/teaching/plan/method/survey/responseScale.pdf (http://http://www.utexas.edu/academic/ctl/assessment/iar/teaching/plan/method/survey/responseScale.pdf)

Chapter 7: Getting Data Ready for Cleaning

Splitting delimited text into columns in Excel

"Split text into different cells," Microsoft Office Support, http://office.microsoft.com/en-us/excel-help/split-text-into-different-cells-HA102809804.aspx

Splitting strings

JAVASCRIPT

"JavaScript String split() Method," W3Schools, http://www.w3schools.com/jsref/jsref_split.asp

VISUAL BASIC

"Split Function (Visual Basic)," Microsoft Developer Network, http://msdn.microsoft.com/en-us/library/6x627e5f(v=vs.90).aspx

String Operations in Python

“Common string operations,” The Python Standard Library, Python Software Foundation, https://docs.python.org/2/library/string.html

Missing Data

Osborne, Jason W., “Dealing with Missing or Incomplete Data: Debunking the Myth of Emptiness,” Ch 6, Best Practices in Data Cleaning, Los Angeles, CA: SAGE Publications, Inc, 2013:

Chapter 8: Data Cleaning

Spreadsheet software

“List of spreadsheet software,” Wikipedia, http://en.wikipedia.org/wiki/List_of_spreadsheet_software

Regular Expressions

GUIDES AND TESTERS

“Regular Expressions - User Guide,” Zytrax, http://www.zytrax.com/tech/web/regex.htm

RegExr v2.0, http://www.regexr.com

JAVASCRIPT

“JavaScript RegExp Reference,” W3Schools, http://www.w3schools.com/jsref/jsref_obj_regexp.asp

PYTHON

“Regular expression operations,” The Python Standard Library, Python Software Foundation, https://docs.python.org/2/library/re.html

Chapter 10: What Data Cleaning Can and Can't Catch

Data Transformations

"Transformations to Achieve Linearity," Stat Trek, http://stattrek.com/regression/linear-transformation.aspx#

Glossary

Aggregation bias

Aggregation bias occurs when incorrect inferences are made from data that have been aggregated, or summarized. This includes making inferences about the characteristics of the parts of a whole based on the characteristics of the whole itself.

Aggregation

Aggregation refers to the process by which data have been collected and summarized in some way or sorted into categories.

Applied statistics

Applied statistics is a type of statistics that involves the application of statistical methods to disciplines and areas of study.

Arithmetic mean

An arithmetic mean, often simply called a mean, is a type of average, or measure of central tendency, in which the middle of a dataset is determined by adding its numeric values and dividing by the number of values.

Attitudinal statement

An attitudinal statement is a type of response given to a scaled question on a survey that asks an individual to rate his or her feelings toward a particular topic.

Axis label

An axis label is used on a graph to denote the kind of unit or rate of measurement used as the dependent or independent variable (or variables), and can be found along an axis of a graph.

Back transformation

Back transformation is the process by which mathematical operations are applied to data in a dataset that have already been transformed, in order to back transform, or revert, the data to their original form.

Back-end check

A back-end check, also called a server-side check is a type of data validation for datasets gathered electronically, and is performed at the back end, or after data are stored in an electronic database.

Bar graph

A bar graph or chart uses horizontal or vertical bars whose lengths proportionally represent values in a dataset. A chart with vertical bars is also called a column graph or chart.

Cartogram

A cartogram is a map that overlays categorical data onto a projection and uses a different color to represent each category. Unlike a heat map, a cartogram does not necessarily use color saturation to depict the frequency of values in a category.

Categorical data

Categorical data are data, quantitative or qualitative, that can be sorted into distinct categories.

Category label

A category label is used on a graph to denote the name of a category, or a group, of data and may be descriptive or a range of numeric values.

Chart title

A chart title is the description assigned to the graph and includes a summary of the message aimed at the target audience and may include information about the dataset.

Checkbox response

A checkbox response refers to an answer given to a question in a survey administered in electronic form, for which one or more responses can be selected at a time, as may be indicated by a checkbox an individual clicks on.

Chroma

Chroma is the saturation, or vividness, of a hue.

Closed question

A closed, or closed-ended, question is a type of question featured in a poll or survey that requires a limited, or specific kind of, response and is used to collect quantitative data or data that can be analyzed quantitatively later on.

Codebook

A codebook documents the descriptions, terms, variables, and values that are represented by abbreviated or coded words or symbols used in a dataset, and serves as a means for coding and decoding the information.

Color theory

Color theory refers to principles of design focuses on colors and the relationships between them.

Continuous variable

A continuous variable, or continuous scale, has an unlimited number of possible values between the highest and lowest values in a dataset.

Correlation

Correlation measures the degree of association, or the strength of the relationship, between two variables using mathematical operations.

CRAAP test

The CRAAP test denotes a set of questions a researcher may use to assess the quality of source information across five criteria: currency, relevance, authority, accuracy, and purpose.

Data cleaning

Data cleaning, also called data checking or data validation, is the process by which missing, erroneous, or invalid data are determined and cleaned, or removed, from a dataset and follows the data preparation process.

Data label

A data label is used on a graph to denote the value of a plotted point.

Data preparation

Data preparation is the process by which data are readied for analysis and includes the formatting, or normalizing, of values in a dataset.

Data transformation

Data transformation is the process by which data in a dataset are transformed, or changed, during data cleaning and involves the use of mathematical operations in order to reveal features of the data that are not observable in their original form.

Data visualization

Data visualization, or data presentation, is the process by which data are visualized, or presented, after the data cleaning process, and involves making choices about which data will be visualized, how data will be visualized, and what message will be shared with the target audience of the visualization. The end result may be referred to as a data visualization.

Data

Data are observations, facts, or numeric values that can be described or measured, interpreted or analyzed.

Dependent variable

A dependent variable is a type of variable whose value is determined by, or depends on, another variable.

Descriptive statistics

Descriptive statistics is a type of applied statistics that numerically summarizes or describes data that have already been collected and is limited to the dataset.

Diary

A diary is a data collection method in which data, qualitative or quantitative, are tracked over an extended period of time.

Dichotomous question

A dichotomous question is a type of closed question featured in a poll or survey that requires an individual to choose only one of two possible responses.

Direct measurement

Direct measurement is a type of measurement method that involves taking an exact measurement of a variable and recording that numeric value in a dataset.

Discrete variable

A discrete variable, or a discrete scale, has a limited number of possible values between the highest and lowest values in a dataset.

External data

External data refer to data that a researcher or organization use, but which have been collected by an outside researcher or organization.

Factoid

A factoid, or trivial fact, is a single piece of information that emphasizes a particular point of view, idea, or detail. A factoid does not allow for any further statistical analysis.

Filter

A filter is a programmed list of conditions that filters, or checks, items that meet those conditions and may specify further instructions either for the filtered items.

Focus group

A focus group is a data collection method used for qualitative research in which a group of selected individuals participate in a guided discussion.

Forced question

A forced question is a type of scaled question featured in a survey that requires an individual to choose from a give range of possible responses, none of which is neutral.

Front-end check

A front-end check, also called a client-side check, is a type of data validation for datasets gathered electronically, and is performed at the front end, or before data are stored in an electronic database.

Graphical user interface (GUI)

A graphical user interface, or GUI, is a type of interface that allows a user to interact with a computer through graphics, such as icons and menus, in place of lines of text.

Heat map

A heat map is a graph that uses colors to represent categorical data in which the saturation of the color reflects the category’s frequency in the dataset.

Histogram

A histogram is a graph that uses bars to represent proportionally a continuous variable according to how frequently the values occur within a dataset.

Hue

A hue, as defined in color theory, is a color without any black or white pigments added to it.

Independent variable

An independent variable is a type of variable that can be changed, or manipulated, and determines the value of at least one other variable.

Inferential statistics

Inferential statistics is a type of applied statistics that makes inferences, or predictions, beyond the dataset.

Infographic

An infographic is a graphical representation of data that may combine several different types of graphs and icons in order to convey a specific message to a target audience.

Interactive graphic

An interactive graphic is a type of visualization designed for digital or print media that presents information that allows, and may require, input from the viewer.

Interviewer effect

Interviewer effect refers to any effect an interviewer can have on subjects such that he or she influences the responses to the questions.

Invalid data

An invalid data are values in a dataset that fall outside the range of valid, or acceptable, values during data cleaning.

Leading question

A leading question is a type of question featured in a poll or survey that prompts, or leads, an individual to choose a particular response and produces a skewed, or biased, dataset.

Legend

A legend is used on a graph in order to denote the meaning of colors, abbreviations, or symbols used to represent data in dataset.

Legibility

Legibility is a term used in typography and refers to the ease with which individual characters in a text can be distinguished from one another when read.

Line graph

A line graph uses plotted points that are connected by a line to represent values of a dataset with one or more dependent variables and one independent variable.

Median

A median is a type of average, or measure of central tendency, in which the middle of a dataset is determined by arranging its numeric values in order.

Metadata

Metadata are data about other data, and may be used to clarify or give more information about some part or parts of another dataset.

Missing data

Missing data are values in a dataset that have not been stored sufficiently, whether blank or partial, and may be marked by the individual working with the dataset.

Mode

A mode is a numeric value that appears most often in a dataset.

Motion graphic

A motion graphic is a type of visualization designed for digital media that presents moving information without need for input from the viewer.

Multiseries

A multiseries is a dataset that compares multiple series,or two or more dependent variables and one independent variable.

Normal distribution

A normal distribution, often called a bell curve, is a type of data distribution in which the values in a dataset are distributed symmetrically around the mean value. Normally distributed data take the shape of a bell when represented on a graph,the height of which is determined by the mean of the sample, and the width of which is determined by the standard deviation of the sample.

Open content

Open content, open access, open source, and open data are closely-related terms that refer to digital works that are free of most copyright restrictions. Generally, the original creator has licensed a work for use by others at no cost so long as some conditions, such as author attribution, are met (See: Suber, Peter. *Open Access*, Cambridge, Massachusetts: MIT Press, 2012). Conditions vary from license to license and determine how open the content is.

Open question

An open, or open-ended question, is a type of question featured in a survey that does not require a specific kind of response and is used to collect qualitative data.

Order bias

Order bias occurs when the sequencing of questions featured in a survey has an effect on the responses an individual chooses, and produces a biased, or skewed, dataset.

Outlier

An outlier is an extremely high or extremely low numeric value that lies outside the distribution of most of the values in a dataset.

Pattern matching

Pattern matching is the process by which a sequence of characters is checked against a pattern in order to determine whether the characters are a match.

Peer-to-peer (P2P) network

A peer-to-peer network, often abbreviated P2P, is a network of computers that allows for peer-to-peer sharing, or shared access to files stored on the computers in the network rather than on a central server.

Pie chart

A pie chart is a circular graph divided into sectors, each with an area relative to whole circle, and is used to represent the frequency of values in a dataset.

Population

A population is the complete set from which a sample is drawn.

Probability

Probability is the measure of how likely, or probable, it is that an event will occur.

Qualitative data

Qualitative data are a type of data that describe the qualities or attributes of something using words or other non-numeric symbols.

Quantitative data

Quantitative data are a type of data that quantify or measure something using numeric values.

Radio response

A radio response refers to an answer given to a question in a poll or survey administered in electronic form, for which only one response can be selected at a time, as may be indicated by a round radio button an individual clicks on.

Range check

A range check is a type of check used in data cleaning that determines whether any values in a dataset fall outside a particular range.

Range

A range is determined by taking the difference between the highest and lowest numeric values in a dataset.

Raw data

Raw data refer to data that have only been collected, not manipulated or analyzed, from a source.

Readability

Readability is a term used in typography and refers to the ease with which a sequence of characters in a text can be read. Factors affecting readability include the placement of text on a page and the spacing between characters, words, and lines of text.

Sample

A sample is a set of collected data.

Sampling bias

Sampling bias occurs when some members of a population are more or less likely than other members to be represented in a sample of that population.

Scaled question

A scaled question is a type of question featured in a survey that requires an individual to choose from a given range of possible responses.

Scatterplot

A scatterplot uses plotted points (that are not connected by a line) to represent values of a dataset with one or more dependent variables and one independent variable.

Series graph

A series graph proportionally represents values of a dataset with two or more dependent variables and one independent variable.

Series

A series is a dataset that compares one or more dependent variables with one independent variable.

Shade

Shade refers to adding black to a hue in order to darken it.

Skewed data

Skewed data are data with a non-normal distribution and tend to have more values to the left, as in left-skewed, or right, as in right-skewed, of the mean value when represented on a graph.

Stacked bar graph

A stacked bar graph is a type of bar graph whose bars are divided into subsections, each of which proportionally represent categories of data in a dataset that can be stacked together to form a larger category.

Standard deviation

A standard deviation is a measure of how much the values in a dataset vary, or deviate, from the arithmetic mean by taking the square root of the variance.

Static graphic

A static graphic is a type of visualization designed for digital or print media that presents information without need for input from the viewer.

Statistics

Statistics is the study of collecting, measuring, and analyzing quantitative data using mathematical operations.

Summable multiseries

A summable multiseries is a type of multiseries with two or more dependent variables that can be added together and compared with an independent variable.

Summary record

A summary record is a record in a database that has been sorted,or aggregated, in some way after having been collected.

Tint

Tint refers to adding white to a hue in order to lighten it.

Transactional record

A transactional record is a record in a database that has not yet been sorted, or aggregated, after collection.

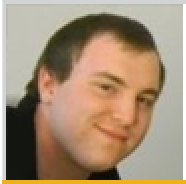
Value (color)

Value, or brightness, refers to the tint, shade, or tone of a hue that results black or white pigments to a base color.

Variance

Variance, or statistical variance, is a measure of how spread out the numeric values in a dataset are, or how much the values vary, from the arithmetic mean.

Contributors



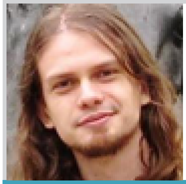
MIHAI BADESCU

Web Developer

Research, Code

Mihai is a junior web developer living in Chicagoland. If he's not dreaming of working with data, he's off somewhere actually working with it. He also has a thing for 90s Japanese cooking shows.

@mihaibad (https://twitter.com/mihaibad)



VITOR BAPTISTA (HTTP://VITORBAPTISTA.COM/)

Data Visualization Developer

Open Knowledge Foundation

Editor

Vitor Baptista is a data visualization developer at the Open Knowledge Foundation working on the CKAN data portal platform. He sees the access, consumption, and interpretation of data of all kinds as an exciting way to help people understand the world and make it better.

@vitorbaptista (https://twitter.com/vitorbaptista) website (http://vitorbaptista.com/)

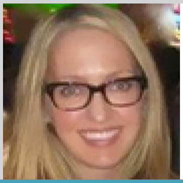
PETR BELES (HTTP://WWW.2150.CH)

Senior ETL Consultant

2150 GmbH

Editor

Petr is an expert in Data Modeling, ETL Architecture and Development. He's currently focused on agile business intelligence, including automated testing and data vault modeling. Always eager to share and expand his knowledge, Petr is a regular speaker at TDWI conferences in Switzerland and Germany.

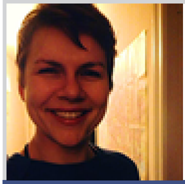


JESSICA BERGER
Research Study Coordinator
Northwestern University Feinberg School of Medicine

Technical Editor

Jessica works in psychiatric clinical research and has a passion for building databases and queries. She also enjoys opportunities to write and edit presentations and publications on the subject. She recently began work toward a degree in epidemiology and biostatistics at Northwestern University Feinberg School of Medicine.

@nerdfighter_14 (https://twitter.com/nerdfighter_14)



MARIT BRADEMANN
Data Enthusiast

Writer

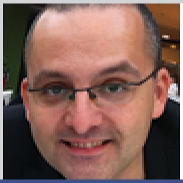
Marit's interest in data is fueled by her support for transparency and open access to information. She holds a Master's degree in quantitative sociology and regularly contributes to projects in the cultural, educational, and scientific spheres. She loves traveling, connecting, and writing about self-improvement, society, and, of course, (open) data.

AMANDA BUENZ MAKULEC (HTTP://WWW.DATAVIZHUB.CO)
Monitoring & Evaluation Associate
John Snow Inc.

Writer

Amanda works at the intersection of global health monitoring, evaluation, and communications. She loves how well-designed visualizations can make data accessible and even help improve health programs. Co-founder of the datavizhub.co community, she enjoys connecting with other viz enthusiasts to share ideas.

@abmakulec (https://twitter.com/abmakulec) website (http://www.datavizhub.co)



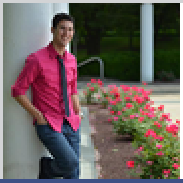
ALBERTO CAIRO (HTTP://WWW.THEFUNCTIONALART.COM)

Professor & Director of Visualization Program
University of Miami

Writer

Alberto teaches infographics and visualization at the University of Miami School of Communication. He also serves as director of the visualization program in the University's Center for Computational Science. Author of *The Functional Art*, Albert has led workshops and consulted media and educational institutions in more than 20 countries.

@albertocairo (https://twitter.com/albertocairo) website (http://www.thefunctionalart.com)



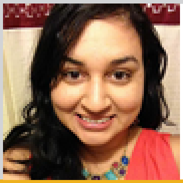
MICHAEL A. CASTELLO (HTTP://CASTELLO.ME)

PhD/MD Candidate
Loma Linda University

Writer

Currently working toward a PhD in human physiology, Michael Castello researches Alzheimer's disease and creates downright fascinating scientific presentations on the subject. Outside the lab, Michael writes, builds things, plays computer games, and hypothesizes about the future.

@castello (https://twitter.com/castello) website (http://castello.me)



MICHELLE CAZARES

Distribution

Michelle is currently dabbling in the business world, so learning how to work with data has been an interesting journey. She is set to start her MBA in Fall 2014 and hopes to be able to use the e-book to help her throughout her courses. In her spare time, she likes to read, play videogames (N7 at heart!), and learn about the world.

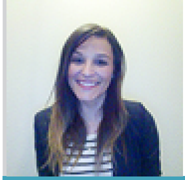
@andromeda_N7 (https://twitter.com/andromeda_N7)

KATHY CHANG (HTTP://NYCHANG.COM/)

Editor, Writer, Distribution

The daughter of immigrants, Kathy remembers when NYC subway trains were covered in graffiti and kind of misses that. Nowadays she rides her trusty Brompton wherever she can. She loves bacon, hates cilantro, and enjoys learning about learning.

website (http://nychang.com/)



JEANA CHESNIK

E-Learning Developer, Massage Envy Franchising, LLC

Editor

Jeana designs, develops, and implements online training programs to support employee development at the corporate and national level. She is currently working with a team at the Franchise Support Center to establish learning and development metrics that evaluate the effectiveness of learning outcomes and make changes as needed.

@Jeanacee (https://twitter.com/Jeanacee)

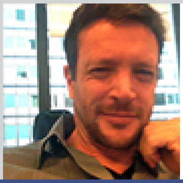
TRINA CHIASSON (HTTP://TRINA.CH/)

Co-founder & CEO
Infoactive

Organizer

Trina is the co-founder of Infoactive and a fellow at the Reynolds Journalism Institute where she works to make data visualization more accessible for journalists. She enjoys turning numbers into stories and making data more human.

@trinachi (https://twitter.com/trinachi) website (https://infoactive.co/)



HOWARD COALE (HTTP://WWW.SAPIENT.COM/)

NY Practice Lead & Creative Director, Sapient Global Markets

Writer

Howard leads the experienced, talented, ambitious, and (mostly) good-humored New York team at Sapient Global Markets, which focuses on data visualization and user experience design.

website (http://www.sapient.com/)

ELLEN COOPER (HTTP://QUADRARESEARCH.COM)

Senior Researcher
Quadra Research

Writer

Ellen has extensive experience designing, implementing, and managing market research studies. Twenty years ago, she launched Quadra Research to provide qualitative and quantitative research studies for industries including utilities, non-profits, packaged goods, real estate, and finance. She is a Certified Advertising Agency Practitioner, corporate member of the Marketing Research and Intelligence Association, and Certified Marketing Research Professional.

website

BILL COPPINGER (HTTP://WWW.WHALESONGSERVICES.COM.AU)

Owner
WhaleSongServices

Research

Bill Coppinger has over 30 years' experience in Education and Training. He co-founded the I*EARN Network in Australia and chaired the International I*EARN Management Team for 3 years. He has a passion for supporting the uptake of evidence-based planning and data visualization. He owns and operates WhaleSongServices.

@billcoppinger (https://twitter.com/billcoppinger) website (http://www.whalesongservices.com.au)

COLLEEN CRESSMAN

Editor

A graduate student in Library & Information Science at Simmons College in Boston, MA, Colleen is interested in the effects of open access on resources for education, research and scholarship. She also studies the relationship between digital literacy and learning.

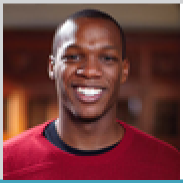
@namsserc (https://twitter.com/namsserc)

ALISTAIR CROLL (HTTP://WWW.SOLVEFORINTERESTING.COM)
Founder
Solve For Interesting

Writer

Alistair works in web performance, big data, cloud computing, and startup acceleration. Co-founder of Networkshop and Coradiant, he's helped create several startups, accelerators, and conferences. He chairs various tech events, including O'Reilly's Strata conference and the International Startup Festival, and has written four books on analytics, technology, and entrepreneurship, including *Lean Analytics*.

@acroll (https://twitter.com/acroll) website (http://www.solveforinteresting.com)



JARED CROOKS (HTTP://WWW.NOURIBAR.COM)
President
NOURI

Technical Editor

Jared is a scientist and innovator working in emerging technology, energy, and big data. He is the co-founder of NOURI—an "out of this world" food company with the mission of making the world a better place by ending child hunger—and the author of several children's books.

@jacrooks (https://twitter.com/jacrooks) website (http://www.nouribar.com)

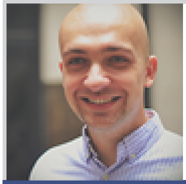
ERSIN DEMIROK

Sabancı University

Writer

Ersin lives a double life, well sort of. He's a researcher in the management and organization department of a national university and also a business developer for investors in the mining and energy industry. He digs being able to present any sort of investment opportunity clearly and quickly with inspiring and simple data visualizations.

@demiroker (https://twitter.com/demiroker)



MIZEL DJUKIC

Senior Product Manager
Millennial Media

Writer

Mizel is a Senior Product Manager in the mobile tech industry. He builds tools that use big data to tell stories effectively and allow users to make better, data-driven business decisions.

@mizthediz (https://twitter.com/mizthediz)

CANDICE DOUGLAS (HTTP://WWW.THESOCIALPOST.CO.ZA)

Chief of Social at The Social Post

Editor

Candice's love for putting words together beautifully began early in her life. Her love for photography came next. Today, she is paid to chat on social media, capture light through her camera lens, and help people translate their message into words for the rest of the world to read.

@TheSocialPostSA (https://twitter.com/TheSocialPostSA) website (http://www.thesocialpost.co.za)

STEVE DREW

Co-Founder, Veraison Vineyards

Technical Editor

Quite the globetrotter, Steve lived in New York and London before moving to South Africa in 2007 where he bought a Burgundian-sized vineyard and farm house five minutes from the Franschhoek village center.

OZ DU SOLEIL (HTTP://DATASCOPIC.NET)

DataScopic

Writer

Oz is a Chicago-based freelancer who has been working with Excel and data for 15 years. He's a passionate speaker and trainer devoted to data literacy. When Oz isn't elbow-deep in data, he's known for sriracha, bowties, and good bourbon.

@Oz_Dragoncookie (https://twitter.com/Oz_Dragoncookie) website (http://datascopic.net)

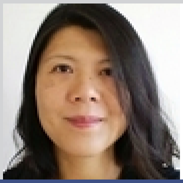
KATE EYLER-WERVE

Project Manager, Mightybytes

Editor, Writer

Kate helps organizations build websites that both deliver against business objectives AND look good. Kate also trains people on methods and models for integrating content strategy, user experience, and web design. She's co-authored two books: *Return on Engagement: Content Strategy and Web Design Techniques for Digital Marketing*, and O'Reilly Media's *The Civic Apps Competition Handbook*.

@innokate (https://twitter.com/innokate)



JANE FOO

Digital Systems Librarian, Seneca College

Writer

With a background in user-centered design, Jane has experience in systems testing (including integration and usability), systems implementation, UI Design, and training. She credits an undergraduate course on environmental psychology with prompting a lifelong interest in human-computer interaction and the impact of technology and visual design on everyday activities.

@grumpel (https://twitter.com/grumpel)

PHILIPP GASSNER (HTTP://WWW.GREENCHALLENGEACCEPTED.ORG/)

Interace Expert & Founder of Green Challenge Accepted

Distribution

Philipp is an environmental scientist and communication expert with a passion for knowledge, people, and the environment. Dedicated to communicating the importance of sustainability, he turns data into drama, numbers into narrative, and stats into stories, translating knowledge into practice to change institutions.

@GrnChllngAccptd (https://twitter.com/GrnChllngAccptd) website (http://www.greenchallengeaccepted.org/)

COLLIN GRAVES (HTTP://WWW.COLLINGRAVES.COM/)

Co-Founder, Vitalane

Technical Editor

Collin stumbled upon the art and psychology of design while working as a mechanic in the Air Force. He went on to found a few award-winning startups and now works as a startup consultant, UX designer, and freelance web developer.

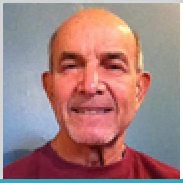
@GravesCollin (https://twitter.com/GravesCollin) website (http://www.collingraves.com/)

DYANNA GREGORY

Co-Organizer, Data + Design

Organizer, Writer

Dyanna is a librarian and biostatistician who has been obsessive about data cleaning since the age of 13 and loves to teach statistics to people of all ages, especially those who think they hate math.



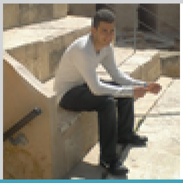
TOM GRUBISICH (HTTP://LOCALAMERICA.COM/#)

Editorial Director, Local America

Editor

Creator of and inspiration behind Local America, Tom is a veteran journalist with more than 30 years' experience. His previous gigs include: editor at the Washington *Post*, managing editor of AOL local news, founder of a chain of community newspapers in metro Washington, and consulting Web editor at the World Bank.

@TomGrubisich (https://twitter.com/TomGrubisich) website (http://localamerica.com/#)



NIDHAL HADDAD

Student, ESSAI

Editor

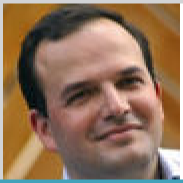
Nidhal is pursuing a national degree at The Higher School of Statistics and Data Analysis in Tunisia (ESSAI). His current and past projects include PCA analysis, tree-based models, and network analysis. His thesis project focuses on market share prediction in the airline industry.

MARGIE HENRY

Senior Analyst, Rakuten Marketing

Writer

Margie's relationship with data began in college when she nearly failed her first statistics class and decided the subject was a beast to be conquered. When she's not cranking out reports and crunching numbers, Margie addresses the growing need for evidence-based decision making and performance measurement among organizations in the arts, culture, and commercial creative sectors.



CAMILO HERRERA

President, RADDAR

Editor

Camilo serves as research consultant and strategic advisor to more than 100 brands. He has authored academic articles on microeconomics, cultural capital, and marketing, including several on the Columbian consumer.

website

MELISSA HILL DEES

Programs & Communication, Volunteer Center of Morgan County

Editor

Melissa is an integrator, collaborator, innovator, and relationship builder. She has 20 years of successful marketing experience working with industry leaders and 8 years in entrepreneurial electronic marketing and CRM for small businesses.

SAURABH JAIN (HTTP://ABOUT.ME/SAURABHJAIN1099)

Analytics Lead, Fractal Analytics

Writer

Saurabh is an analytics lead for one of the largest consumer goods companies in the world. On a typical day, he tackles tough big data in order to improve decision management, predict and influence consumer behavior, and optimize business results.

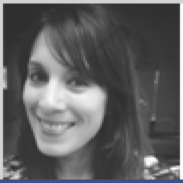
@saurabhjain1099 (https://twitter.com/saurabhjain1099) website (http://about.me/saurabhjain1099)

KATE KRETSCHMANN

Project Coordinator, Arizona Geological Survey

Project Manager

Kate coordinates community building and project development for the NSF EarthCube Initiative, a data and knowledge management system that will allow for cross-discipline data sharing across the geosciences. She comes to her work with a background in journalism, science reference publishing, and online media.



GINETTE LAW

Writer

Ginette collects, explores, and reports data for companies and organizations, including the World Economic Forum, ictQatar, Huawei, and the World Wide Web Foundation. When not geeking out over data, Ginette enjoys the finer things in life… like ice cream.

@ggeeksd (https://twitter.com/ggeeksd)

AN LE

Engagement Leader, Officience

Editor

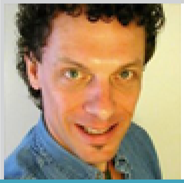
Born and living in South Central Vietnam in the time of Facebook and Twitter, Thanh An is an engagement leader at the global consulting firm Officience. He's driven by a strong entrepreneurial spirit and a fascination with mass data digitization.

BONNIE MEISELS

Designer, Social Entrepreneur, Real Estate Broker

Other

As a graduate of Parson's School of Design, Bonnie designed handbags & accessories in New York. After moving back to Montreal, she is applying her passion for designing and making a difference to her practice as a socially conscious real estate broker.



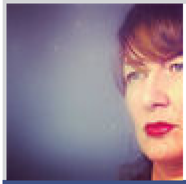
TOM MORRIS

Product Manager & Software Engineer

Technical Editor

Tom is a product manager, software geek, and athlete who lives and works in Boston, Massachusetts.

@tfmorris (https://twitter.com/tfmorris)



CALLIE NEYLAN (HTTP://NINETEENTHIRTYFOUR.ORG/)

Senior Software Designer, Microsoft

Writer

Callie was born in Aspen, Colorado, before movie stars discovered it. She's worked at various startups, software companies, and design firms, and as a senior interaction designer at NPR. She now works with some of the best engineers in the world creating data visualization and visual analytics tools.

@neylano (https://twitter.com/neylano) website (http://nineteenthirtyfour.org/)

SANJIT OBERAI (HTTP://WWW.INDIASPEND.COM)

Deputy Editor, IndiaSpend

Technical Editor

Sanjit is Deputy Editor of IndiaSpend, a non-profit organisation and India's first Data Journalism initiative which utilises open data to empower citizens decision making process. He likes creating stories around data with the help of free to use tools and believes that story telling is the way forward. He has worked with PaGaLGuY and Firstpost.com. He lives in Mumbai.

@sanjit_oberai (https://twitter.com/sanjit_oberai) website (http://www.indiaspend.com)

TARA PHELPS (HTTP://WWW.TECHTOSUCCESS.COM/)

Founder, Success Technologies

Editor

Tara the Technology Matchmaker is a business automation and strategy expert whose passion is helping coaches, entrepreneurs, and small business owners use technology to save time and money, create more meaningful relationships with clients and prospects, and achieve greater success in business.

@taraphelps (https://twitter.com/taraphelps) website (http://www.techtosuccess.com/)

KIRAN PV

Writer

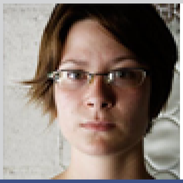
Kiran provides data science training at Jigsaw Academy, an Online School of Analytics based in Bangalore, India. He holds a dual degree in engineering from IIT Madras, India, and his technical expertise spans SAS, Python, Hadoop, and Tableau. Outside of work, Kiran writes, travels, and cheers for his favorite club, Manchester United in EPL.

CHRISTINE RIDGWAY

National Clinical Educator at Teleflex Medical Australia

Editor

Christine's passion is communicating complex topics in simple and approachable ways, empowering people to perform at their highest level. She is a National Clinical Educator at Teleflex Medical Australia, specializing in designing learning resources to support the sale and use of complex medical devices and, ultimately, improve patient outcomes.



SARAH RIGDON

Writer & Consultant

Writer

Sarah likes translating complex ideas into good stories. A fan of civic tech and mass collaboration, she's an alumna of NASA's Open Innovation Program and a volunteer at Code for DC and Code for Houston. She led communications for the 2013 International Space Apps Challenge and helped organize the TechLady Hackathon + Training Day.

@sarah_rigdon (https://twitter.com/sarah_rigdon)

ADAM ROBBINS

Research

Adam Robbins works with communities in the Midwest and East Asia to achieve LGBT equality. He helped Minnesotans United for All Families achieve the first victory against an anti-marriage constitutional amendment, trained with the New Organizing Institute, and is currently working with a startup in Beijing to measure and develop the LGBT market in China.

@robbinsadam (https://twitter.com/robbinsadam)

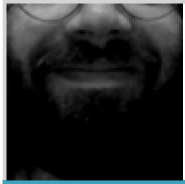
ANASTASIA SAVVINA (HTTP://ABOUT.ME/ANASTASIASAVVINA/BIO)

Sr. Copywriter, Kaiser Permanente

Editor

Anastasia is a senior copywriter by day and freelance writer in her off hours. She received a bachelor's degree in Literature & Writing from UC San Diego and is grateful/relieved/ecstatic that she gets to use that education at work. After writing, her favorite things are traveling, backpacking, and trying new foods.

website (http://about.me/anastasiasavvina/bio)



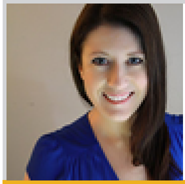
ALFREDO SERAFINI (HTTP://SERALF.IT)

Freelance Consultant

Editor

Alfredo has a diverse professional background that includes working as a consultant for information retrieval and linked data, teaching programming languages, and designing soundscapes. You can follow his various pursuits on his "designed to be unfinished" website.

@seralf (https://twitter.com/seralf) website (http://seralf.it)



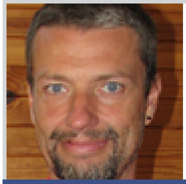
CHRISTINA SHOMAKER (HTTP://WWW.DUMBLEBEE.COM/)

Founder, Dumblebee

Distribution

Christina is a longtime tech nerd and founder of Dumblebee, the advice-based social network. Her never-ending curiosity has led her to work in many fields, including anthropology, photography, graphic design, museum research, and real estate.

@DumblebeeIt (https://twitter.com/DumblebeeIt) website (http://www.dumblebee.com/)



IAN TEBBUTT (HTTPS://WWW.HEALTHROUNDTABLE.ORG/DEFAULT.ASPX)

Senior Data Analyst

Writer

Ian knows big data. He's worked with it for more than two decades, doing everything from UK TV rating calculations to fraud protection for Australian banks to health datasets for hospitals. His team of data experts is small but powerful, known for producing good data and presenting it well.

website (https://www.healthroundtable.org/default.aspx)

KEVIN TAYLOR

NGS Program Manager, NetApp

Editor

Kevin has spent the past 15 years in Business Intelligence, focusing primarily on the effective delivery of information and insight. Downright passionate about visualization, he finds beauty in simplicity and loves reading and following the pioneers of the industry. Outside the 9-to-5, he enjoys spending time with his beautiful wife Katherine and their two beautiful twin daughters, Kaseyleigh and Kennedy.

@KevinTaylorNC (https://twitter.com/KevinTaylorNC)

JESSICA THOMAS (HTTP://WWW.SPEAKYOURDESIGN.COM)

Marketing Communications Manager

Distribution, Editor

Jessica develops communications plans and thought leadership platforms for clients in architecture, engineering, design, higher education, and the arts. She enjoys translating technical jargon into plain language and believes in the power of good design to communicate even the most complex information.

@speakyourdesign (https://twitter.com/speakyourdesign) website (http://www.speakyourdesign.com)

ALEX WALL (HTTP://ALEXWALL.CO)

Digital Marketing Strategist, Roar Media

Editor

Alex Wall is a digital marketing strategist in Coral Gables, Florida, and freshly addicted to data visualization.

@alexlwall (https://twitter.com/alexlwall) website (http://alexwall.co)

BEATRIZ WALLACE (HTTP://BEATRIZWALLACE.COM)

Visiting Professor, Duquesne University

Project Manager

A New Orleans native, Beatriz currently lives in Pittsburgh, working as a Visiting Professor at Duquesne University. She helped implement digital storytelling for children exposed to family violence in rural Missouri and is currently working with University of Missouri Professor Berkley Hudson on a project that explores photographs of the South.

@BigMuddyHeart (https://twitter.com/BigMuddyHeart) website (http://beatrizwallace.com)

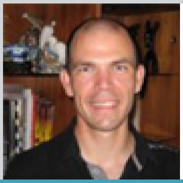
CHRISTOPHER WETHERILL (HTTP://WWW.CHRISWETHERILL.ME)

PhD Student, Virginia Tech

Technical Editor

Christopher is a PhD student in Virginia Tech's Translational Biology, Medicine, and Health program. He also enjoys cat yodeling on the side.

website (http://www.chriswetherill.me)



GREG ZOBEL (HTTP://WWW.DRGBZ.COM/)

Assistant Professor, Western Oregon University

Editor

Gregory Zobel lives in Oregon's Willamette Valley. He collects books and loves budgies (better known as parakeets to the uninitiated).

@drgbz (https://twitter.com/drgbz) website (http://www.drgbz.com/)