Chapter 2: Telling a story with quantitative data

Our schooling tends to separate language from numbers. In English and language arts classes we read stories and learn how to tell our own. In math classes, we learn to work with numbers. This split between language and numbers hides how we also tell stories with numbers.

Many well-known scientists, engineers, analysts, and other professionals talk about the research they read as telling "a beautiful story,"[i] or their work as finding "a story now one else has found"[ii] or using "data to tell a meaningful story that resonates both intellectually and emotionally with an audience."[iii]

What do we mean, though, by telling a story with data? Let's start with one of my favorite data visualizations, created by the dating website OKCupid. OKCupid asks its users hundreds of questions about their beliefs, habits, desires, and preferences and uses this information to match up potential romantic partners. A side benefit of all these questions is that the website operators have access to lots of social data which they can use to tell stories such as the following:

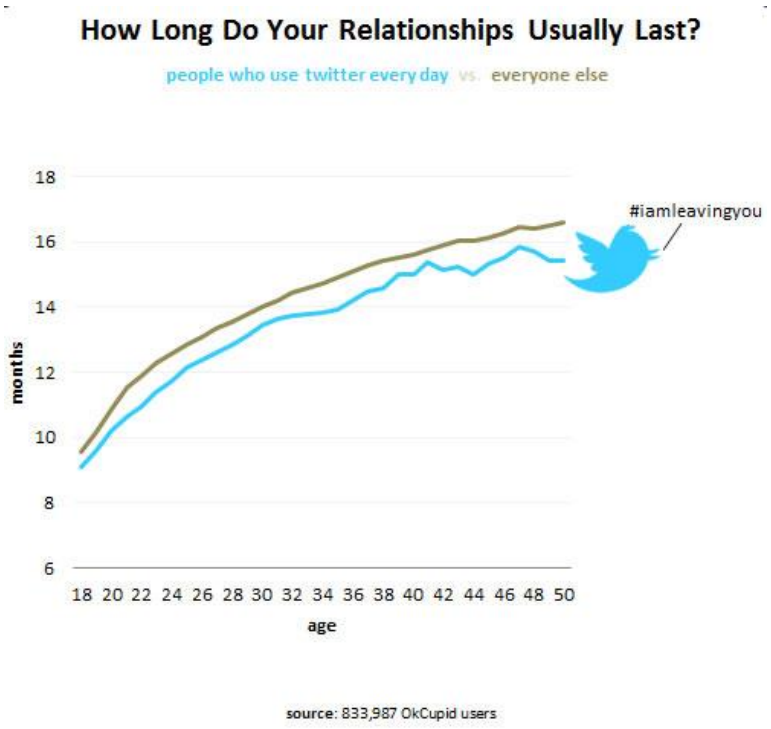


Figure 2.1: Length of relationship (in months) of OKCupid users by age.

This graph actually tells us multiple stories. The first—and more interesting story—is that daily Twitter users average shorter relationships than others. The second story is that relationship length increases as people age. This second story is less surprising than the first.

In order to create this story, someone had to dig through gigabytes of data to determine what was interesting. They had to make decisions about how to measure variables such as relationship

length. And then the writer had to choose to present the data in a certain way—choosing a line graph rather than a bar chart or table—to make us come to a certain conclusion. All of these choices amount to the creation of a story.

Even though we are accustomed to thinking about data as "factual," the small choices writers make in how to present that data affect the story it tells. To illustrate, take a look at tables 2.1a and 2.1b below. Both tables present data on the number of medals earned by country in the 2008 Summer Olympics. However, despite drawing on the same numbers, these two tables tell different stories.

Can you guess which version was favored by US media outlets?

Table 2.1a 2008 Olympic Medals

Country	*Gold*	*Silver*	*Bronze*	*Total*
USA	36	38	36	**110**
China	51	21	28	**100**
Russia	23	21	29	**73**
Britain	19	13	15	**47**
Australia	14	15	17	**46**
Germany	16	10	15	**41**
France	7	16	18	**41**

Table 2.1b 2008 Olympic Medals

Country	*Gold*	*Silver*	*Bronze*	*Total*
China	**51**	21	28	100
USA	**36**	38	36	110
Russia	**23**	21	29	73
Britain	**19**	13	15	47
Germany	**16**	10	15	41
Australia	**14**	15	17	46
S. Korea	**13**	10	8	31

Table 2.1a tells the story that the USA "won" the 2008 Olympics by 10 medals whereas Table 2.1b suggests that China is the winner by 15 gold medals. You are probably not surprised to learn that variations on Table 2.1a appeared in most US newspapers whereas Table 2.1b was favored by the rest of the world.

Both versions, of course, are completely accurate. They draw on the same factual data, but the differences in how these tables prioritize certain data turn them into competing stories.

We, of course, tell stories about data not just with visualizations but also with words. Whether in a popular magazine or academic journal, writers use captions and paragraphs to 'tell a story' about their data. However, rather than a story, many beginning writers tend to describe what their data is *about*—for instance, they might write "Figure 2.1 shows Twitter use and relationship length" or "Table 2.1 presents the medals winnings from the 2008 Olympics." Such statements are factual, but they do not tell stories. Readers want to know *how* Twitter use seems to affect relationships. They want to know *which country* dominated the Olympics.

Table 2.2 below illustrates the difference between describing data and telling a story about data.

Table 2.2: Descriptions versus Stories about data

Description (no story)	***Story***
Table 2.1 shows Olympic medals by country	China took home the most gold medals in 2008.
We show the most popular dog breeds 1970-present	Labrador retrievers have been the most popular dog breed for over three decades.
Figure X shows profits from 2010-2015	Figure X shows that profits nearly doubled between 2010 and 2015
Our data compares user perceptions of the old versus new interface.	Our data shows that the new interface is less popular with users than the old one.
The chart below show the percentage of students from each major.	The chart below shows that the majority of students were either Business or Computer Science majors.

The first column simply describes what data the writer collected or what the intended goal of the analysis was. Imagine this was all you had read about the data. You would learn almost nothing. By contrast, the second column tells us what we *learn* from the data. You could read the sentences in this column and "get" what the data says without reading anything else.

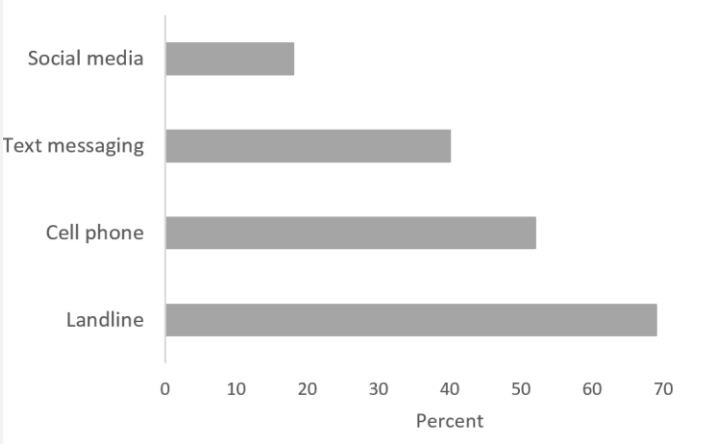
You might be feeling that telling stories with both words and visuals is redundant. However, both are critical. Without words guiding them to what is most important about your data, readers can misread your data or fail to grasp its significance. And if you provide words without visuals, readers will question your credibility and ask for more detail. The story you tell about your data needs to be reinforced in multiple formats.

We also need to consider the story of our data when working with qualitative data, such as quotations, observations or descriptions. We may present a particular quotation or description as evidence, but we need to then interpret it for our own purposes to show the story it is emphasizing. Chapter 6will provide a detailed look at how to tell a story about data primarily in the form of *words* and not numbers.

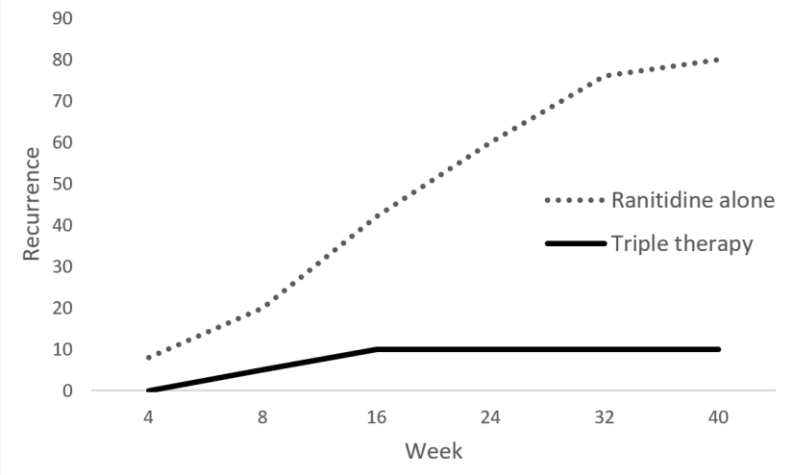
Exercise 2.1:
The goal of this exercise is to give you practice telling stories about data. Below you will find four data visualizations accompanied by captions that describe but do not tell a story about the visualization's content.

For each visualization, write one complete sentence that communicates its main story. Then, if the visualization has multiple stories, write an additional one or two sentences communicating these secondary stories.

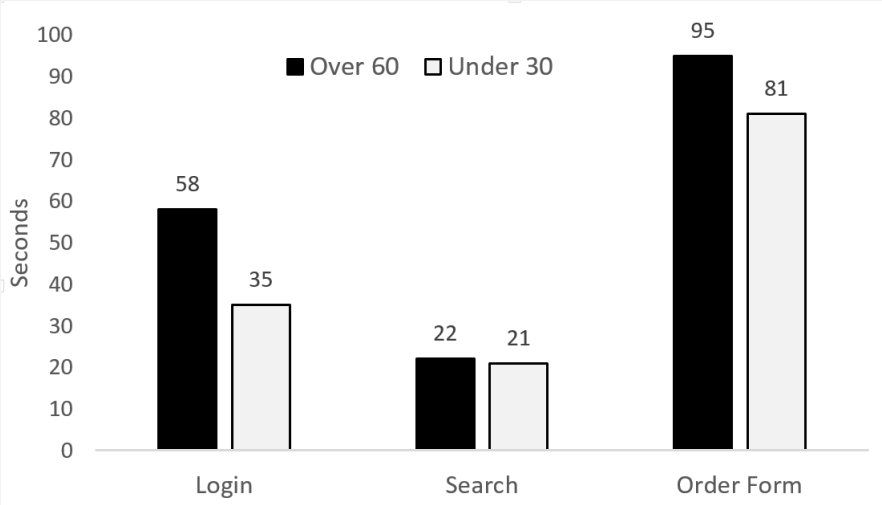
A. Percentage of adults who feel secure sharing private data via different communication channels



B: Recurrence of gastric ulcers for the year after successful healing with ranitidine alone or triple therapy plus ranitidine.



C. Times to complete three different virtual store functions by adults over 60 years and under 30 years.



D. Average hours studied per week, GPA, and degree completion of students in three different majors.

Major	*Avg Hrs. studied/wk*	*Avg GPA*	*5-year Degree completion*
Business	22.4	2.9	58%
Humanities	25.1	3.2	61%
Sciences	29.8	2.8	52%

Not all stories are equal

As we learned in the last chapter, not all stories about data are equal. We shape our stories to suit our audience and purpose—for instance, the nationality of our audience might affect how we report data about Olympics medals. However, we must be careful to remain credible and ethical in our story-telling.

Writers have to continually balance their purpose—their desire to persuade readers that their data has a meaningful story to tell—with the needs of their audience and the need to maintain credibility. When writers exaggerate either their visuals or their words to make their story appear more compelling than it might otherwise, they sacrifice their credibility.

We demonstrate credibility by ethically and accurately displaying and discussing our data. Here are four common ways writers lose credibility by telling inaccurate or unethical stories about their data:

1. Going against common conventions for reporting or displaying data in the hopes readers will misinterpret the data
2. Leading readers to believe that small differences are actually large (or vice-versa)
3. Leaving out important context or information that would change readers' understanding of the data
4. Exaggerating the conclusions we can draw from the data

Avoiding these four mistakes is much easier said than done. It is not always clear whether a given claim is credible. For instance, take #2 above. There are situations where very small differences can have big effects. For instance, if we are talking about the size of a medical implant, very miniscule differences in size or weight could have a major impact on whether or not our bodies would accept or reject the implant. Likewise, it is unclear when writers are leaving out information (#3) or exaggerating conclusions (#4). In the case of a medical implant, a 0.02% difference in size could lead us to claim that an implant should or should not be used, whereas the same difference would be negligible if we were talking about changes in the growth of the entertainment industry.

In such instances, we need to look to the audience and purpose of the data story to decide whether or not the reporting is accurate and ethical. In the case of medical implants, the audience likely understands (or can easily be made to understand) that small variations can literally be a matter of life or death. Thus, the audience will not conclude that writers are exaggerating if they magnify small differences in medical implants, whereas if we are talking about profits of a major industry, our criteria for evaluating claims will likely differ.

Let's go through these four ways writers can lose credibility one-by-one. For each, we will look at clear-cut cases where a data story is inaccurate or unethical and then we will discuss more borderline cases that depend on audience, purpose and context.

1. **Writers should avoid going against common conventions for reporting data in the hopes that readers will misinterpret the data**
   One of the most infamous examples of flouting graph conventions can be seen in Figure 2.2 below. See if you can figure out why this graph caused such a controversy when published by Reuters news.

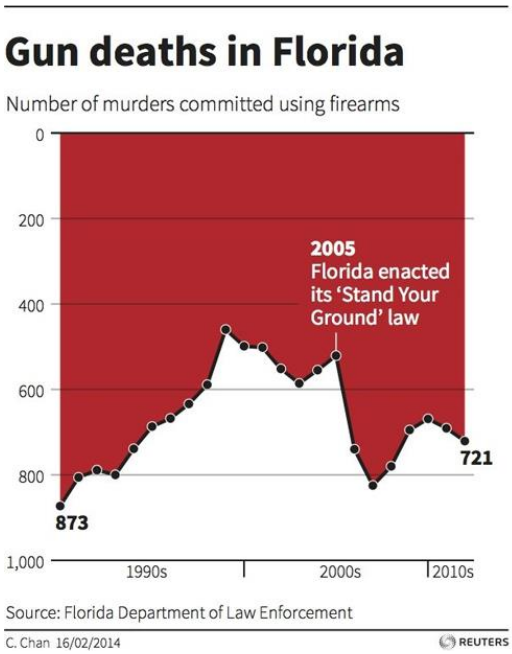


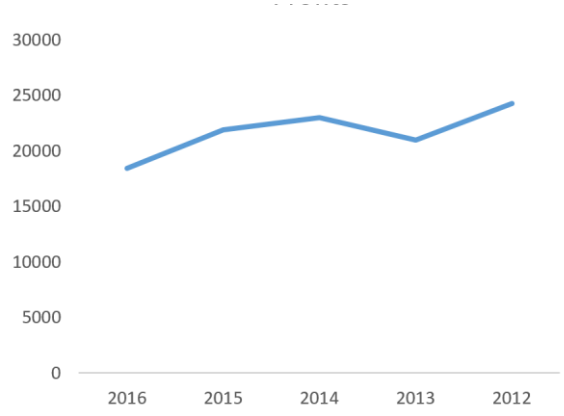
Figure 2.2: Reuters' infamous 'stand your ground' graph

As you hopefully noted, this graph reverses the normal way we are used to looking at the y-axis of a graph by placing 0 at the top of the graph rather than at the bottom where it would intersect with the x-axis. As a consequence, any reader looking quickly at this graph would conclude that gun homicides decreased after the "stand your ground law" was enacted whereas the data actually shows gun homicides increasing. Most readers perceived this graph as outright deceitful—a crass attempt to make readers draw false conclusions to serve a particular political stance.

It is worth noting that the author of this graph, Christine Chan, claims her intent was not to deceive; rather, she states she was trying to increase the emotional impact of the graph by giving the impression that blood is dripping down the page. Few readers were persuaded by this defense. Her deliberate flouting of graph conventions causes most readers to consider it not an example of artistic license but a "Liar, liar, pants on fire" deception.

Below, we see another example of a graph that tries to deceive by flouting common conventions. Imagine a store manager presenting version a of the graph below to a potential investor rather than the more accurate version b.

Version a



Version b

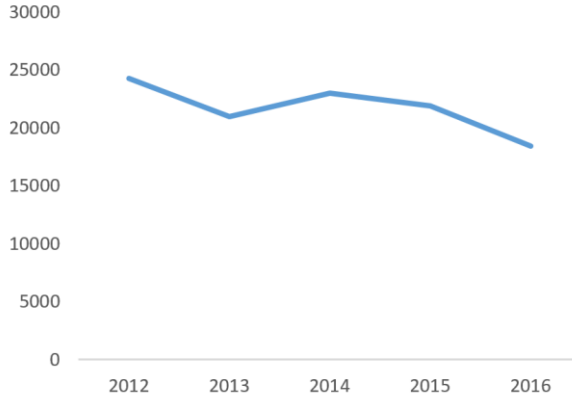


Figure 2.3: Store profits by year. Version (a) is an unethical attempt to disguise falling profits by reversing the x-axis whereas version (b) is a credible visualization of the same data

Such attempts to trick readers into believing the opposite of what a trend actually is are clearly unethical and cause a writer to lose credibility. In fact, it is easy to imagine a store manager losing his or her job or being subject to a law suit after using version a in a presentation.

This does not mean, however, that all attempts to go against convention are unethical. If readers are immediately aware that they need to change their reading practices—and if the shift in conventions helps the writer communicate an important story—going against convention may be acceptable. For instance, Figure 2.4 below uses some unusual conventions to tell a story about the enactment of stand your ground laws. Take a moment to read this graph and see if you can determine its story.

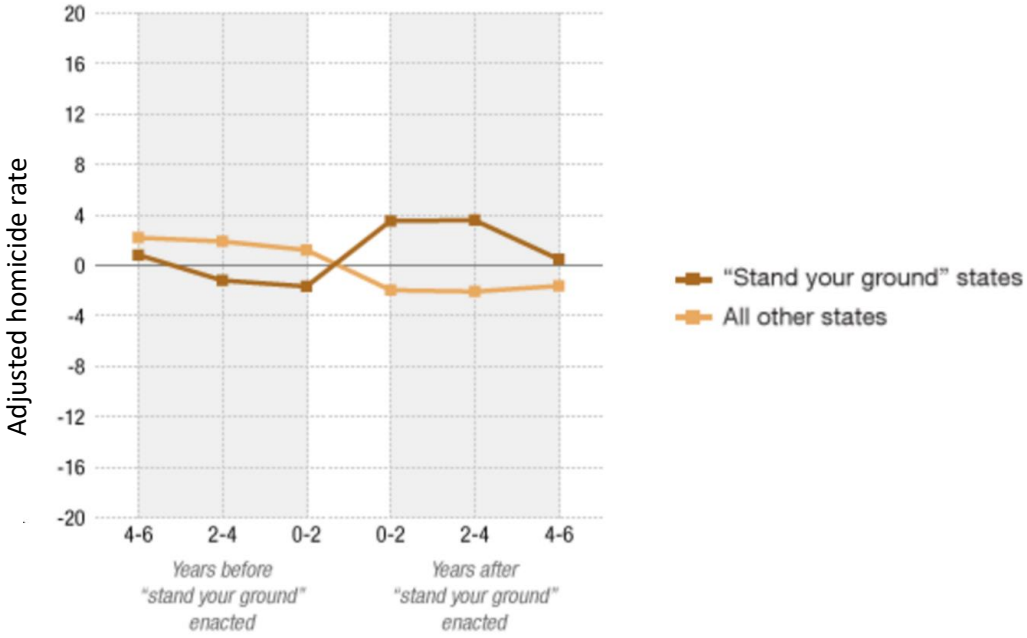


Figure 2.4: Homicide rates in states that have stand your ground laws versus rates in states that don't have the laws. The vertical y-axis represents an adjusted homicide rate that takes into account a state's population, pre-existing crime trends and other factors.[iv] "Stand your ground" laws remove the duty to retreat in the case of self-defense. Florida was the first state to enact such a law in 2005, and the law played a prominent role in the trial of George Zimmerman's shooting of Trayvon Martin.

You probably identified the story of this graph as something like "adjusted homicide rates initially increased in states enacting stand your ground laws while they decreased in other states." Although Figure 2.4 uses some unconventional techniques that require effort to understand, it also clearly calls our attention to these techniques with its shading and dark horizontal line demarcating the zero on the y-axis. The unconventional depictions of the x- and y- axes in this graph do not necessarily hurt the writers' credibility, although many audiences may not want to invest the effort necessary to understand the graph.

What *might* make readers challenge the graph's credibility is its reliance on an "adjusted homicide rate." Although the caption tells us some of the factors included in this adjusted rate, it is hard to tell whether it is credible without more information. If we track down the original data, we will find that it draws on some sophisticated statistical analyses by two economists. Our assessment of the graph's credibility will ultimately depend on how much we trust these economists and their techniques. Becoming a critical reader of data gets easier the more data-driven arguments you encounter.

We will discuss the pros and cons of using non-intuitive variables such as 'adjusted' rates in Chapter ___. The point you should take from this discussion is that breaking from accepted

conventions is unethical if it leads to blatant misreading of the data, but may be acceptable if readers are aware that their reading practices need to change to digest a complicated story.

2. Writers should not deceive readers into thinking that small differences are actually large

Writers sometimes use visual effects to distort our impressions of the data. Note for instance, how the 3D pie chart on the left exaggerates the number of Muslims compared to the chart on the right:

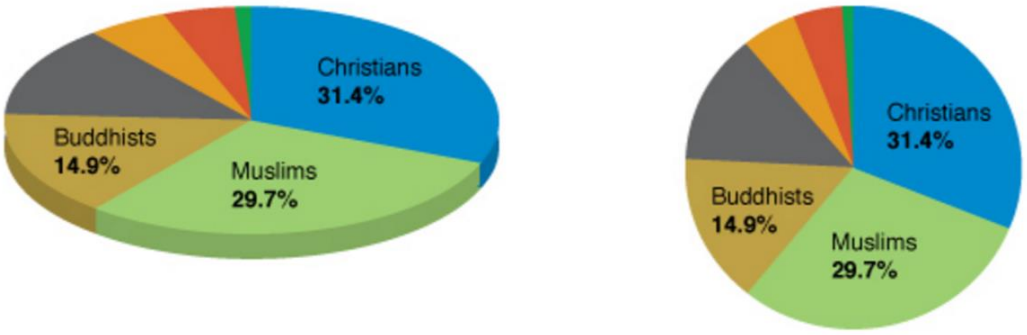


Figure 2.5: Religions in the world: The pie chart on the left uses 3D effects to make the proportions of Muslims appear larger than it otherwise would seem.

Another common way to misrepresent data is to truncate an axis in order to distort the scale. Consider Figure 2.5 below.

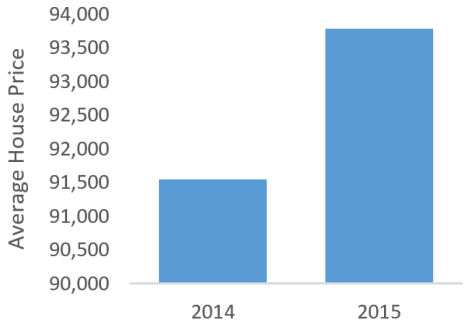


Figure 2.6: Average housing prices in 2014 and 2015

The y-axis begins at 90,000, making an increase of 2% look like housing prices have more than doubled. Here is how this same data would appear if the y-axis began at 0.

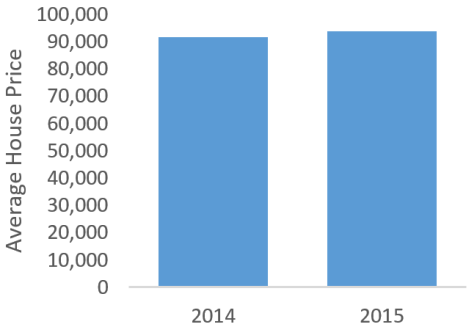


Figure 2.6 revised: Average housing prices in 2014 and 2015

With the y-axis at zero, the differences between 2014 and 2015 are barely noticeable.

Since most homeowners would not find a 2% increase in housing prices surprising (particularly if that rate was consistent with inflation for the year), the original version of Figure 2.5 seems unethical. The writer appears to be trying to alarm readers where no cause for alarm exists.

However, it is more debatable whether Figure 2.6 below is acceptable.

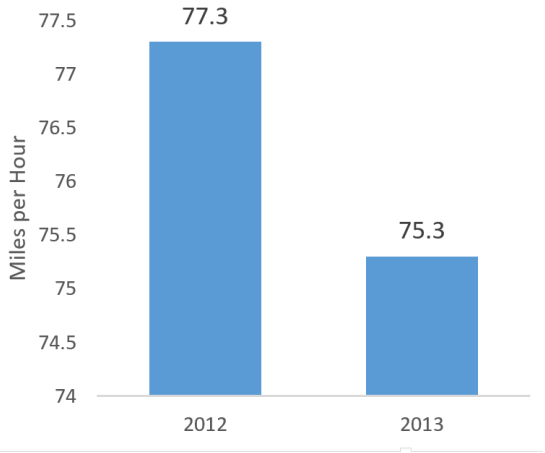


Figure 2.7: Average knuckleball velocity of Pitcher R.A. Dickey

Like the original version of Figure 2.5, Figure 2.6 begins the y-axis at an arbitrary point rather than at zero. This choice makes the decrease of two miles per hour appear severe. However, there are two major differences between this graph and the original version of Figure 2.5

First, the author of Figure 2.6 has included data labels over the bars, drawing our attention to the exact values. Readers can easily calculate in their heads the difference between 2012 and 2013.

Second, a sports fan could easily argue that a two mph decrease in pitch speed *is* a dramatic drop that could affect a game. This context, especially when combined with the data labels, lends some justification to the writer's decision to begin the y-axis at a place other than zero.

Whether or not Figure 2.6 is credible depends on our **audience**. Some audiences will state that the y-axis of a bar graph should always begin at zero while others will say it depends. This brings us back to one of the basic themes of this book: while our culture tends to treat numbers as inherently factual, in reality, numbers are often open to interpretation. What determines if data presentation is ethical or not largely depends on if a writer is attempting to deceive a reader into seeing a story that the data does not support.

For Discussion:
Ask an instructor or teaching assistant in your intended major to look at Figure 2.6 and say whether or not it is ethical. Then ask that instructor if starting a y-axis at a point other than zero would ever be acceptable in their field. What factors would contribute to its acceptability?

3. Writers should not leave out important context or information that would change our understanding of the data

Imagine a writer has a table of data with dozens of variables. It is not necessary to report every single variable that she collected if that data does not tell an interesting story. There is nothing wrong with being selective if her goal is to avoid overwhelming her audience.

However, if she leaves out data simply because it contradicts a story that she wants to tell, she is undermining her credibility. Imagine for instance that a sales manager has the following chart reporting sales by month for an entire year:

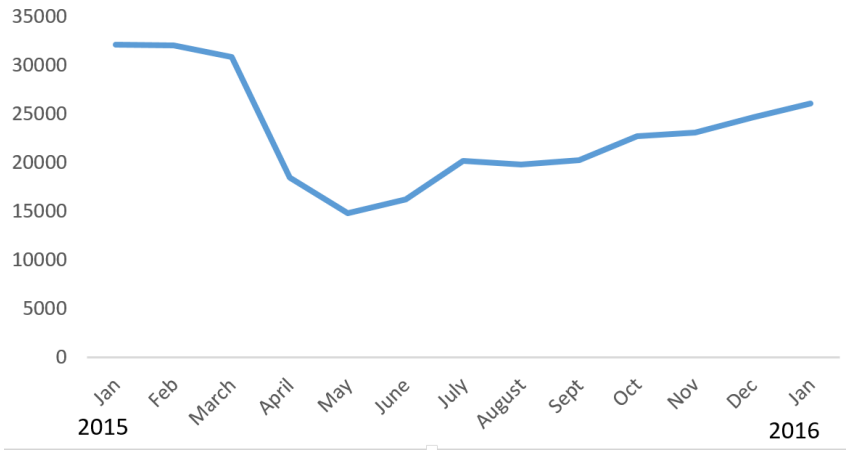


Figure 2.8: Monthly sales Jan 2015-Jan 2016

If the manager was concerned about the fact that sales overall have dropped, she might be tempted to have her graph start in May so that her story would focus on the steady increase from May through January. This technique would hide the fact that her profits in January

2016 were actually lower than they were a year before. Such manipulation seems deliberately deceptive and would hurt her credibility.

Another case of leaving out important context occurs when we lack information that would help us put data in perspective. See if you can figure out what additional information might change our interpretation of Figure 2.8 below:

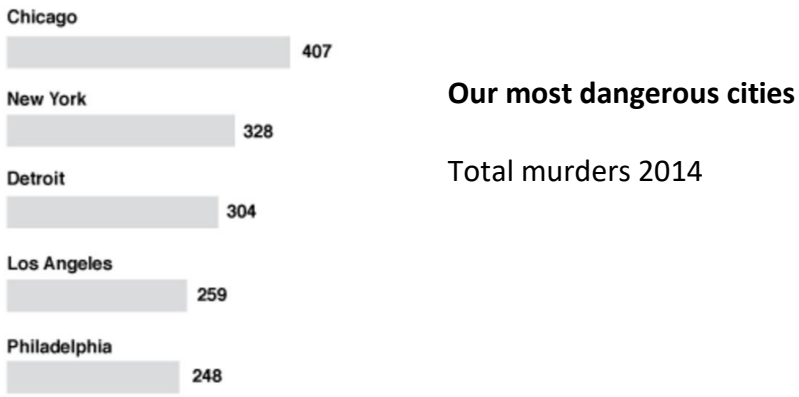


Figure 2.9 Total murders by city 2014.

A good question to ask yourself when assessing a data visualization's credibility is always **"Out of how many?"**

The problem with Figure 2.8 is that Chicago and New York are the two largest cities in the U.S. Thus, it is no surprise that their murder rate is higher than other cities. While this graph is ethical, it is not quite an accurate representation of which city is most dangerous. To better illustrate which city has the highest murder rate (and therefore could be said to be most dangerous), we should adjust the numbers *per capita* (or for each person) as in Figure 2.9:

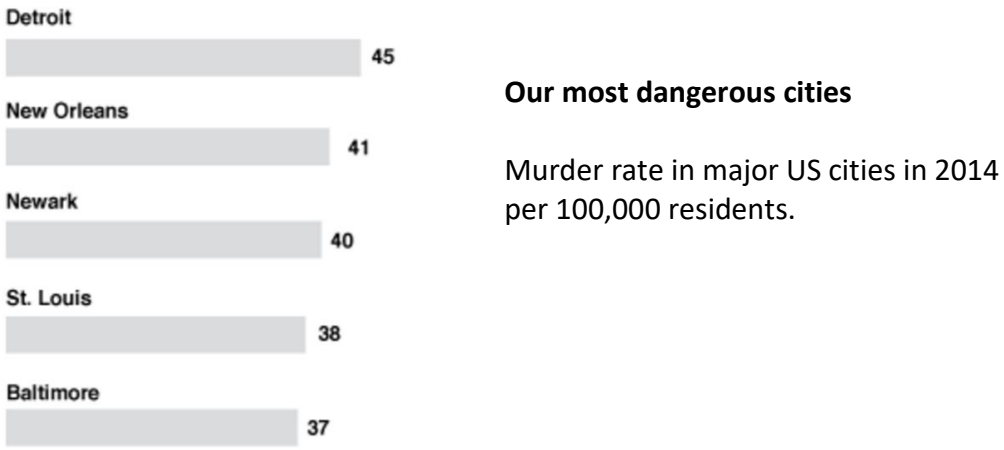


Figure 2.9: 2014 Murder rate per 100,000 residents

Figure 2.9 provides appropriate context for the data by clarifying "Out of how many?" In this representation, we see that Detroit has the highest murders per capita. In fact, when we look at murder rates as a proportion of city population, Chicago and New York do not make the top of the list.

As you become more experienced in writing with data, you will become more adept at recognizing when numbers need additional context to be understood.

4. **Writers should carefully word claims to avoid exaggerating what the data can say**
This last principle is very closely tied to the previous one and can be a particularly hard line to define. On the one hand, we want to capture our readers' interest and make them care about our story. On the other hand, we don't want to overstate that story in ways that will hurt our credibility.

One of the most common exaggerations writers make when discussing their data is to mistake correlation for cause. Just because two variables are correlated does not mean that one necessarily causes the other. For instance, flossing one's teeth is strongly correlated with a longer life span, but that does not mean that flossing *causes* us to live longer. This claim *might* be true. Or it may be that people with a healthy life style (e.g., regular exercise, good diet, not smoking) are more likely to floss than those with an unhealthy life style. The most we can claim is that flossing is *associated* with longer life spans.

Return your attention briefly to Figure 2.1 at the beginning of this chapter. If we were to describe this graph's story as "Figure 2.1 shows that daily Twitter use shortens relationships," we would be exaggerating what the data can tell us in order to create a sensational story. The data in Figure 2.1 can tell us that there is a *correlation* between Twitter use and relationship length, but not that one *causes* the other. It is just as likely that failed relationships drive people to post on Twitter as it is that Twitter use drives people out of relationships. Or there could be a third factor connected to both Twitter use and relationships that is the real cause here. For instance, both daily Twitter use and short relationships could be caused by some underlying personality factor that is the real explanation for the data in Figure 2.1

Mistaking cause and correlation is not the only way that writers exaggerate their data stories. Consider Figure 2.10 below and think about what stories might be told about this data.

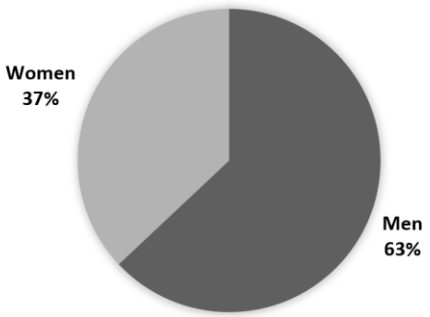


Figure 2.10: Fatal automobile accidents by gender of driver

If we were to write "This data shows that women are safer drivers than men," we would clearly be exaggerating what data can say. First, this data only looks at fatal accidents rather than accidents overall so it is not fair to conclude that women are safer overall. Secondly, it may be that there are more male drivers than female drivers, and this difference accounts for more accidents by men.

So what if we were to write "This data shows that male drivers are more likely to get into fatal accidents than female drivers?" This claim seems a little more credible. But, again, we need to consider how much time men and women spend behind the wheel. If men drive twice as much as women then it should not be a surprise that they get into the majority of fatal accidents.

Thus, a safe story to tell about Figure 2.10 would be that "Figure 2.10 shows that a majority of fatal vehicle accidents involve a male driver." To make additional claims about the relative safety of male and female drivers, we need additional information. We can hypothesize about what additional inferences we might draw from the data, but we need to clearly distinguish these inferences from our main story by using language such as "this data could suggest…." Or "these findings might imply…." We will further discuss the language choices for distinguishing confident claims we can make about our data from more tentative hypotheses and inferences in Chapter ___.

Does this mean that the caption of "our most dangerous cities" for Figure 2.8 and 2.9 is exaggerated? After all, danger involves more than just the murder rate. The answer to this question again goes back to the expectations of our audience. If we were writing for an audience of scientists, this would be overstating our data, but for a general newspaper audience, such a claim might be acceptable.

Be careful, however, about over-applying the advice in this section. Many writers are so afraid of overstating their data story that they retreat to not making a story at all. Telling a non-story—such as "Figure 2.10 shows the proportion of fatal accidents by male and female drivers"—is just as problematic as exaggerating your story. Just as reading more data will build your skills in assessing that data, the more you practice writing about data the more you will gain experience in striking the balance between telling a story and taking care not to overstate that story.

SUMMARY

Effective writing about data involves focusing readers' attention on the stories our data tells. Small choices in the design of a visual display of data can make a major difference in the story it communicates. This is one reason why it is important to tell the data's story in both visuals and words.

When discussing data, writers need to balance credibility with purpose and audience. We need to tell readers what data means—i.e., what lessons we learn from the data—while avoiding the temptation to exaggerate what the data tells us. We need to tell a compelling story without sacrificing our credibility.

Common ways that writers sacrifice their credibility include:

1. Going against common conventions for reporting or displaying data in the hopes readers will misinterpret the data.
2. Leading readers to believe that small differences are actually large (or vice-versa).
3. Leaving out important context or information that would change readers' understanding of the data.
4. Exaggerating the conclusions we can draw from the data.

In addition to avoiding these mistakes when writing about your own data, you should also be on the lookout for these exaggerations when you read others' data. Newspapers, blogs, and other popular media often prioritize entertaining their audience over accurately presenting their data. Critical readers need to watch for misrepresentations and know what questions to ask before accepting what they read as truth. Remember, a good starting place to ask about data is: out of how many?

Exercise 2.2:

Rate each data story below on a scale of 1-4 for how credible it is.

4 = Completely credible
3 = Probably credible
2 = Probably not credible
1 = Not at all credible

For each example that you rate less than "completely credible," indicate which of the four principles above it seems to violate (note: some may violate multiple principles).

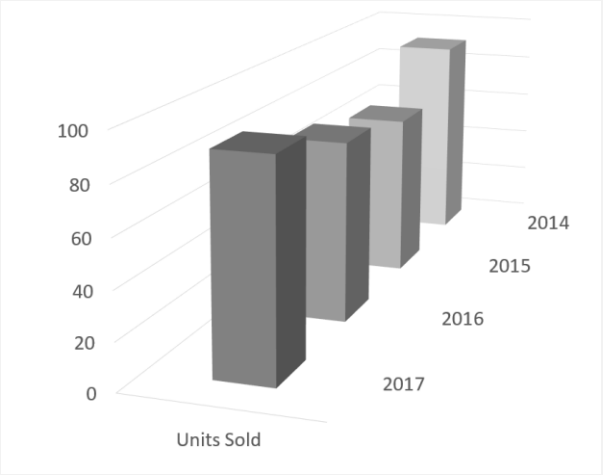
The goal of this exercise is to generate good questions about what is and is not credible in discussing data. Focus on your understanding and on raising good questions rather than obsessing over whether you are obtaining the "correct" answer.

A.

Gender	***Number of questions answered correctly***
Female	16
Male	22

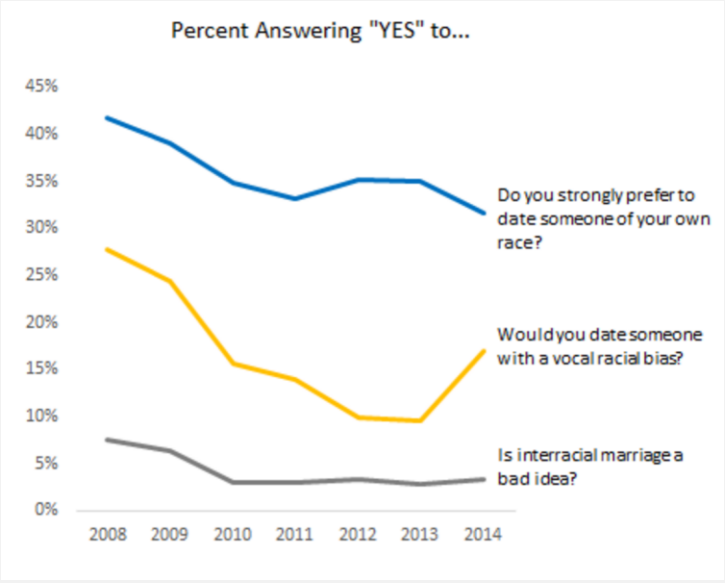
Male students answered more questions correctly than their female peers

B.



The number of units sold increased from 2014-2017

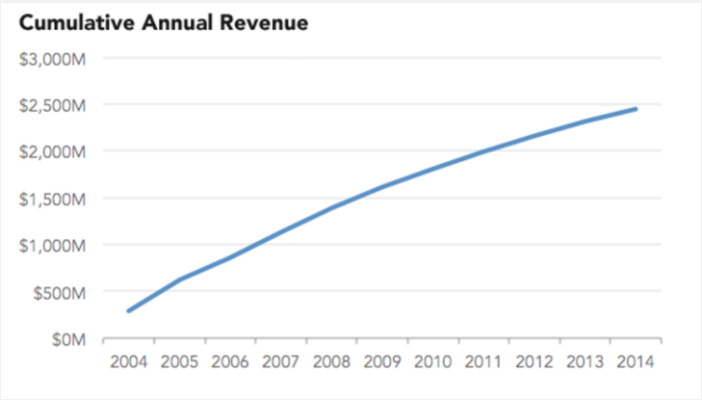
C.



Responses of OKCupid users to questions about race.

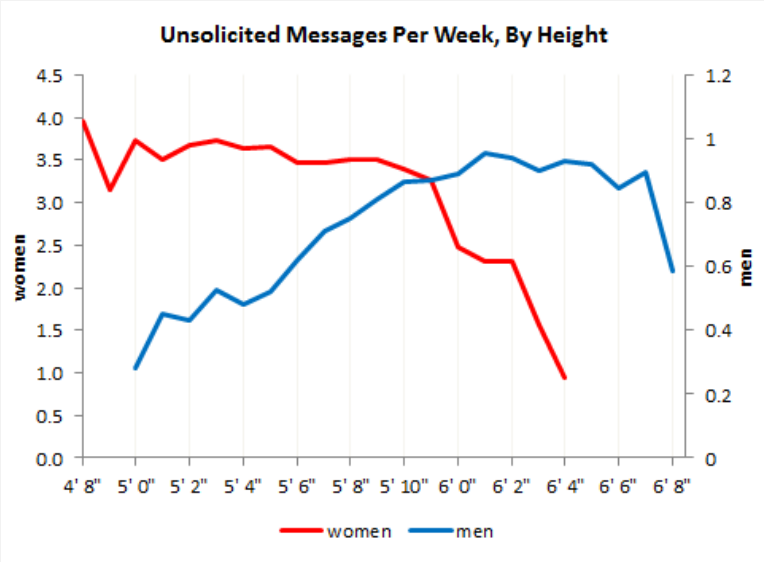
Racist attitudes toward dating appear to have improved slightly since 2008.

D.



Revenue has steadily increased over the last decade.

E.



Shorter women receive more unsolicited emails than taller women (especially women above 5'10") while the opposite trend is true for men.

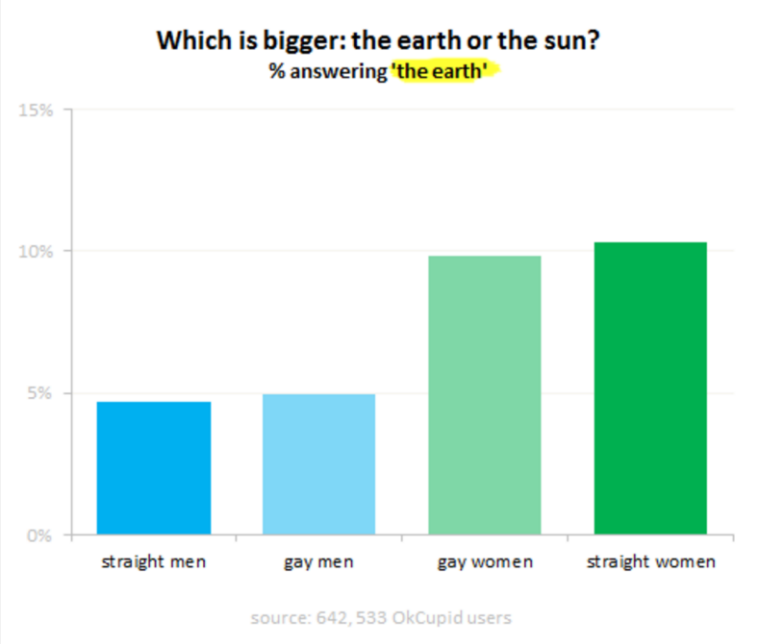
Average weekly unsolicited message totals by height; you can think of these as the number of times a person is "hit on" out of the blue each week on **OkCupid**. The genders are plotted on different scales because of the eternal fact that men almost always make the first move, so women get many more unsolicited messages.

F.

Gasoline prices dropped by over one-third between 2012-2015

Average price of a gallon of gasoline in 2015 dollars

G.



According to the OKCupid database, women are nearly twice as likely as men to believe that the earth is bigger than the sun.

H.

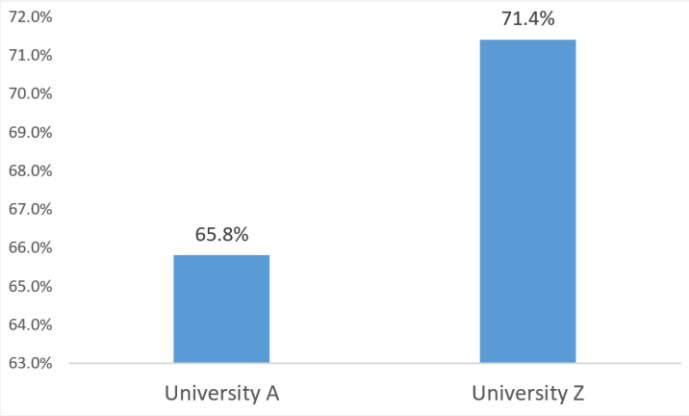


Figure H: The five-year graduation rates of students at two area universities

Students at University Z are much more likely to graduate within five years than students at University A

[i] Rymer 1988?
[ii] Steven Levitt Intro to Freakanomics
[iii] https://www.thinkwithgoogle.com/articles/tell-meaningful-stories-with-data.html

[iv] *Alyson Hurt, Adam Cole & David Schultz/NPR. "*__'Stand Your Ground' Linked To Increase In Homicides"__ Originally published 13 Jan 2013. http://www.npr.org/2013/01/02/167984117/-stand-your-ground-linked-to-increase-in-homicide. Accessed 20 August 2016.