PART I

A STRONG FOUNDATION

Chapter 1

Data Visualization: A Primer

This book is about real-world dashboards and why they succeed. In many of the scenarios, we explain how the designers use visualization techniques to contribute to that success. For those new to the field, this chapter is a primer on data visualization. It provides enough information for you to understand why we picked many of the dashboards. If you are more experienced, this chapter recaps data visualization fundamentals.

Why Do We Visualize Data?

Let's see why it's vital to visualize numbers by beginning with Table 1.1. There are four groups of numbers, each with 11 pairs. In a moment, we will create a chart from them, but before we do, take a look at the numbers. What can you see? Are there any discernible differences in the patterns or trends among them?

Let me guess: You don't really see anything clearly. It's too hard.

Before we put the numbers in a chart, we might consider their statistical properties. Were we to do that, we'd find that the statistical properties of each group of numbers are very similar. If the table doesn't show anything and statistics don't reveal much, what happens when we *plot* the numbers? Take a look at Figure 1.1.

Now do you see the differences? *Seeing* the numbers in a chart shows you something that tables and some statistical measures cannot. We visualize data to harness the incredible power of our visual system to spot relationships and trends.

This brilliant example is the creation of Frank Anscombe, a British statistician. He created this set

TABLE 1.1 Table with four groups of numbers: What do they tell you?

Group A		Group B		Group C		Group D	
x	y	x	y	x	y	x	y
10.00	8.04	10.00	9.14	10.00	7.46	8.00	6.58
8.00	6.95	8.00	8.14	8.00	6.77	8.00	5.76
13.00	7.58	13.00	8.74	13.00	12.74	8.00	7.71
9.00	8.81	9.00	8.77	9.00	7.11	8.00	8.84
11.00	8.33	11.00	9.26	11.00	7.81	8.00	8.47
14.00	9.96	14.00	8.10	14.00	8.84	8.00	7.04
6.00	7.24	6.00	6.13	6.00	6.08	8.00	5.25
4.00	4.26	4.00	3.10	4.00	5.39	19.00	12.50
12.00	10.84	12.00	9.13	12.00	8.15	8.00	5.56
7.00	4.82	7.00	7.26	7.00	6.42	8.00	7.91
5.00	5.68	5.00	4.74	5.00	5.73	8.00	6.89

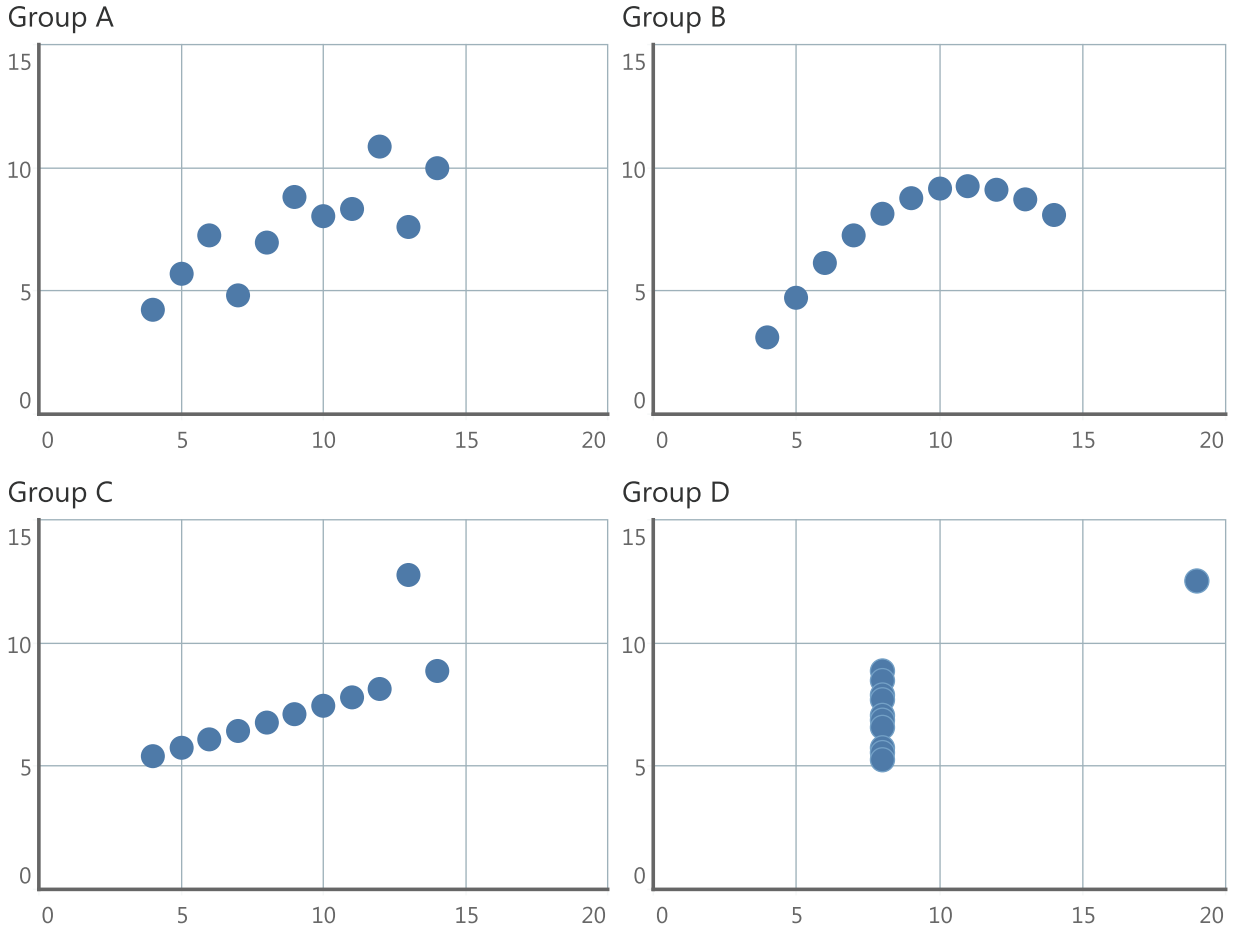


FIGURE 1.1 Now can you see a difference in the four groups?

of numbers—called "Anscombe's Quartet"—in his paper "Graphs in Statistical Analysis" in 1973. In the paper, he fought against the notion that "numerical calculations are exact, but graphs are rough."

Another reason to visualize numbers is to help our memory. Consider Table 1.2, which shows sales numbers for three categories, by quarter, over a four-year period. What trends can you see?

Identifying trends is as hard as it was with Anscombe's Quartet. To read the table, we need to look up every value, one at a time. Unfortunately, our short-term memories aren't designed to store many pieces of information. By the time we've reached the fourth or fifth number, we will have forgotten the first one we looked at.

Let's try a trend line, as shown in Figure 1.2.

TABLE 1.2 What are the trends in sales?

Category	2013 Q1	2013 Q2	2013 Q3	2013 Q4	2014 Q1	2014 Q2	2014 Q3	2014 Q4
Furniture	$463,988	$352,779	$338,169	$317,735	$320,875	$287,934	$319,537	$324,319
Office Supplies	$232,558	$290,055	$265,083	$246,946	$219,514	$202,412	$198,268	$279,679
Technology	$563,866	$244,045	$432,299	$461,616	$285,527	$353,237	$338,360	$420,018
Category	**2015 Q1**	**2015 Q2**	**2015 Q3**	**2015 Q4**	**2016 Q1**	**2016 Q2**	**2016 Q3**	**2016 Q4**
Furniture	$307,028	$273,836	$290,886	$397,912	$337,299	$245,445	$286,972	$313,878
Office Supplies	$207,363	$183,631	$191,405	$217,950	$241,281	$286,548	$217,198	$272,870
Technology	$333,002	$291,116	$356,243	$386,445	$386,387	$397,201	$359,656	$375,229

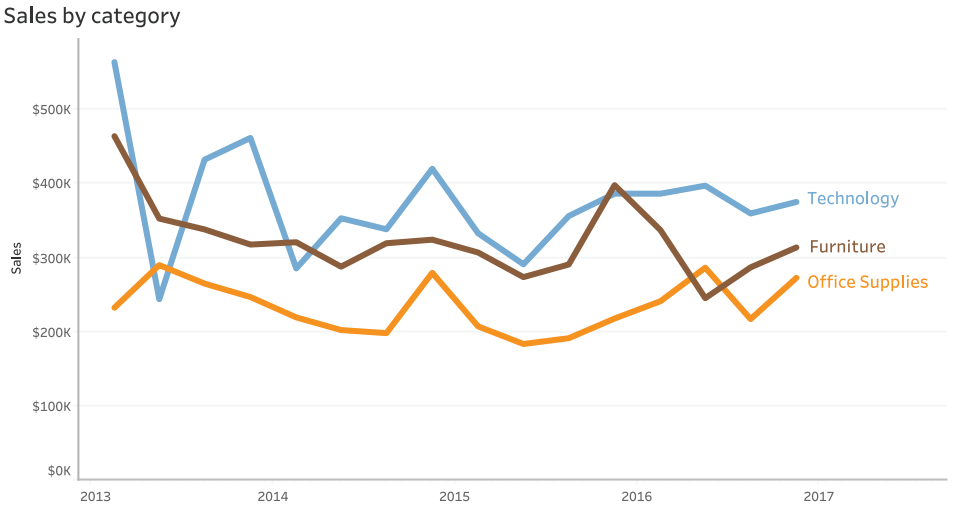


FIGURE 1.2 Now can you see the trends?

Now we have much better insight into the trends. Office supplies has been the lowest-selling product category in all but two quarters. Furniture trends have been dropping slowly over the time period, except for a bump in sales in 2015 Q4 and a rise in the last two quarters. Technology sales have mostly been the highest but were particularly volatile at the start of the time period.

The table and the line chart each visualized the same 48 data points, but only the line chart lets us see the trends. The line chart turned 48 data points into three *chunks* of data, each containing 16 data points. Visualizing the data hacks our short-term memory; it allows us to interpret large volumes of data instantly.

How Do We Visualize Data?

We've just looked at some examples of the power of visualizing data. Now we need to move on to *how* we build the visualizations. To do that, we first need to look at two things: preattentive attributes and types of data.

Preattentive Attributes

Visualizing data requires us to turn data into marks on a canvas. What kind of marks make the most sense? One answer lies in what are called "preattentive attributes." These are things our brains process in milliseconds, before we pay attention to everything else. There are many different types. Let's look at an example.

Look at the numbers in Figure 1.3. How many 9s are there?

How did you do? It's easy to answer the question—you just look at all the values and count the 9s—but it takes a long time. We can make one change to the grid and make it very easy for you. Have a look at Figure 1.4.

2 2 5 6 7 1 1 6 9 1
9 1 7 5 5 5 6 2 5 9
4 5 2 9 6 9 7 6 4 6
8 1 5 7 8 5 6 6 6 7
7 2 3 6 8 9 1 7 9 1
3 8 6 8 4 5 6 9 4 5
4 9 9 2 3 7 1 9 1 2
3 7 8 1 6 1 5 6 1 6
5 6 6 8 6 6 9 1 2 6
3 2 4 2 6 9 4 2 7 1

FIGURE 1.3 How many 9s are there?

Now the task is easy. Why? Because we changed the color: 9s are red, and all the other numbers are light gray.

Color differences pop out. It's as easy to find one red 9 on a table of hundreds of digits as it is on a 10-by-10 grid. Think about that for a moment: Your brain registers the red 9s before you consciously addressed the grid to count them. Check out the grid of 2,500 numbers in Figure 1.5. Can you see the 9?

It's easy to spot the 9. Our eyes are amazing at spotting things like this.

2 2 5 6 7 1 1 6 9 1
9 1 7 5 5 5 6 2 5 9
4 5 2 9 6 9 7 6 4 6
8 1 5 7 8 5 6 6 6 7
7 2 3 6 8 9 1 7 9 1
3 8 6 8 4 5 6 9 4 5
4 9 9 2 3 7 1 9 1 2
3 7 8 1 6 1 5 6 1 6
5 6 6 8 6 6 9 1 2 6
3 2 4 2 6 9 4 2 7 1

FIGURE 1.4 Now it's easy to count the 9s.

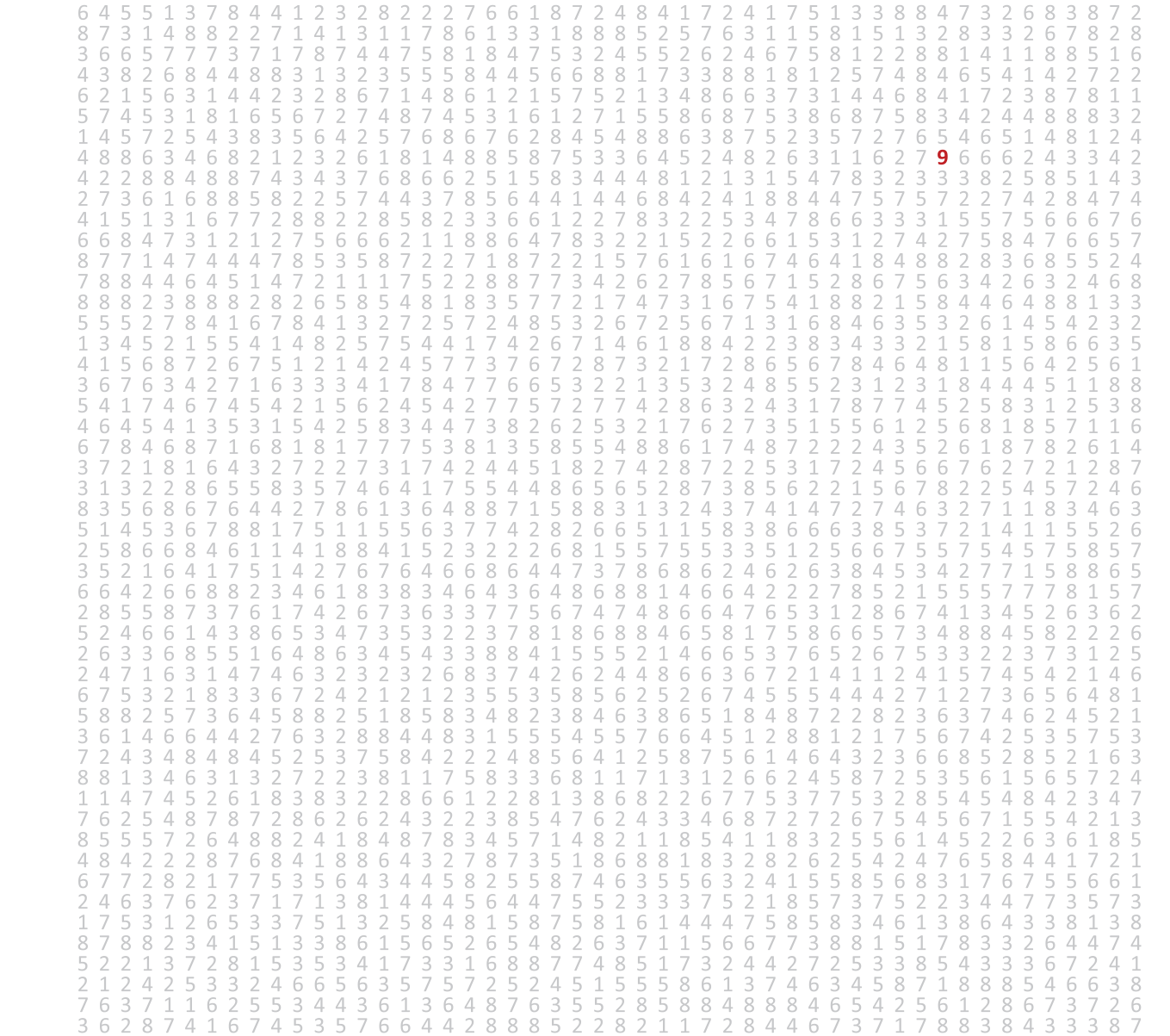


FIGURE 1.5 There is a single 9 in this grid of 2,500 numbers. We wager you saw it before you started reading any other numbers on this page.

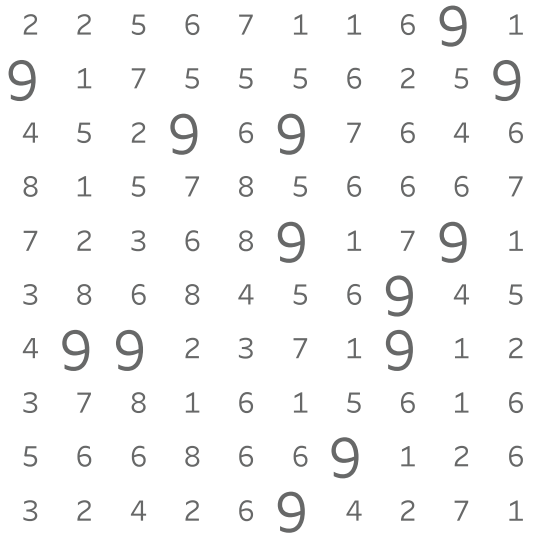


FIGURE 1.6 Differences in size are easy to see too.

FIGURE 1.7 Coloring every digit is nearly as bad as having no color.

Color (in this case, hue) is one of several *preattentive attributes*. When we look at a scene in front of us, or a chart, we process these attributes in under 250 milliseconds. Let's try out a couple more preattentive features with our table of 9s. In Figure 1.6, we've made the 9s a different size from the rest of the figures.

Size and hue: Aren't they amazing? That's all very well when counting the 9s. What if our task is to count the frequency of *each* digit? That's a slightly more realistic task, but we can't just use a different color or size for each digit. That would defeat the preattentive nature of the single color. Look at the mess that is Figure 1.7.

It's not a complete disaster: If you're looking for the 6s, you just need to work out that they are red and then scan quickly for those. Using one color on a visualization is highly effective to make one category stand out. Using a few colors, as we did in Figure 1.2 to distinguish a small number of categories, is fine too. Once you're up to around eight to ten categories, however, there are too many colors to easily distinguish one from another.

To count each digit, we need to aggregate. Visualization is, at its core, about encoding aggregations, such as frequency, in order to gain insight. We need to move away from the table entirely and encode the frequency of each digit. The most effective way is to use length, which we can do in a bar chart. Figure 1.8 shows the frequency of each digit. We've also colored the bar showing the number 9.

Since the task is to count the 9s in the data source, the bar chart is one of the best ways to see the results. This is because length and position are best for quantitative comparisons. If we extend the example one final time and consider which numbers are most common, we could sort the bars, as shown in Figure 1.9.

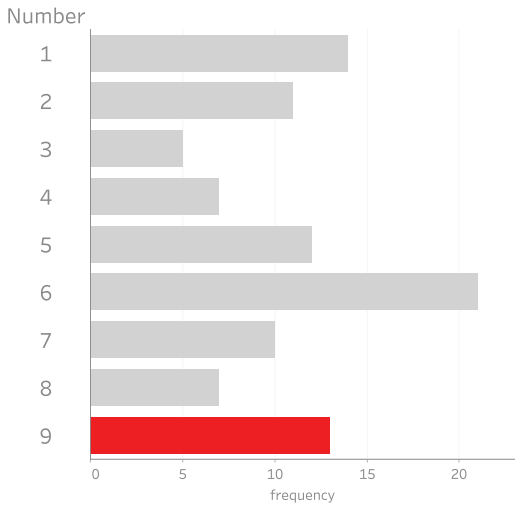


FIGURE 1.8 There are 13 9s.

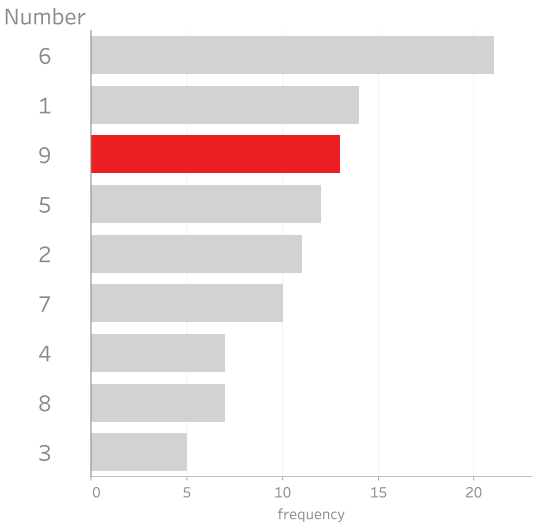


FIGURE 1.9 Sorted bar chart using color and length to show how many 9s are in our table.

This series of examples with the 9s reemphasizes the importance of visualizing data. As with Anscombe's Quartet, we went from a difficult-to-read table of numbers to an easy-to-read bar chart. In the sorted bar chart, not only can we count the 9s (the original task), but we also know that 9 was the third most common digit in the table. We can also see the frequency of every other digit.

The series of examples we just presented used color, size, and length to highlight the 9s. These are three of many preattentive attributes. Figure 1.10 shows 12 that are commonly used in data visualization.

Some of them will be familiar to you from charts you have already seen. Anscombe's Quartet (see Figure 1.1) used position and spatial grouping. The x- and y-coordinates are for position, while spatial grouping allows us to see the outliers and the patterns.

Preattentive attributes provide us with ways to encode our data in charts. We'll look into that in more detail in a moment, but not before we've talked about data.

To recap, we've seen how powerful the visual system is and looked at some visual features we can use to display data effectively. Now we need to look at the different types of data, in order to choose the best visual encoding for each type.

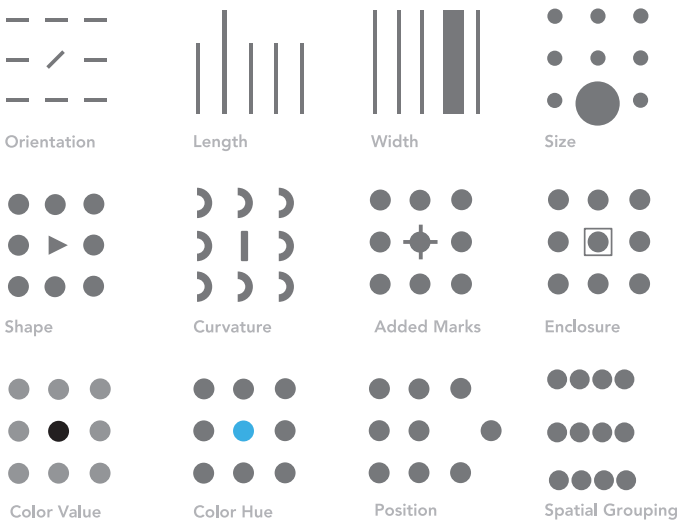


FIGURE 1.10 Preattentive features.

Types of Data

There are three types of data: categorical, ordinal, and quantitative. Let's use a photo to help us define each type.

Categorical Data

Categorical (or *nominal*) *data* represents *things*. These things are mutually exclusive labels without any numerical value. What nominal data can we use to describe the gentleman with me in the Figure 1.11?

- His *name* is Brent Spiner.
- By *profession* he is an actor.
- He played the *character* Data in the TV show *Star Trek: The Next Generation*.

FIGURE 1.11 One of your authors (Andy, on the right) with a celebrity.
Source: Author's photograph

Name, profession, character, and TV show are all categorical data types. Other examples include gender, product category, city, and customer segment.

Ordinal Data

Ordinal data is similar to categorical data, except it has a clear order. Referring to Brent Spiner:

- Brent Spiner's *date of birth* is Wednesday, February 2, 1949.
- He appeared in all seven *seasons* of *Star Trek: The Next Generation*.
- Data's *rank* was lieutenant commander.
- Data was the fifth of six *androids made* by Dr. Noonien Soong.

Other types of ordinal data include education experience, satisfaction level, and salary bands in an organization. Although ordinal values often have numbers associated with them, the interval between those values is arbitrary. For example, the difference in an organization between pay scales 1 and 2 might be very different from that between pay scales 4 and 6.

Quantitative Data

Quantitative data is the numbers. *Quantitative* (or *numerical*) data is data that can be measured and aggregated.

- Brent Spiner's *date of birth* is Wednesday, February 2, 1949.
- His *height* is 5 ft 9 in (180 cm) tall.
- He made 177 *appearances* in episodes of *Star Trek*.
- Data's positronic brain is capable of 60 trillion *operations per second*.

You'll have noticed that date of birth appears in both ordinal and quantitative data types. Time is unusual in that it can be both. In Chapter 31, we look in detail about how you treat time influences your choice of visualization types.

Other types of quantitative measures include sales, profit, exam scores, pageviews, and number of patients in a hospital.

Quantitative data can be expressed in two ways: as discrete or continuous data. Discrete data is presented at predefined, exact points—there's no "in between." For example, Brent Spiner appeared in 177 episodes of *Star Trek*; he couldn't have appeared in 177.5 episodes. Continuous data allows for the "in between," as there is an infinite number of possible intermediate values. For example, Brent Spiner grew to a height of 5 ft 9 in but at one point in his life he was 4 ft 7.5 in tall.

Encoding Data in Charts

We've now looked at preattentive attributes and the three types of data. It's time to see how to combine that knowledge into building charts. Let's look at some charts and see how they encode the different types of data. Sticking with *Star Trek*, Figure 1.12 shows the IMDB.com ratings of every episode of *Star Trek: The Next Generation*.

Table 1.3 shows the different types of data, what type it is, and how it's been encoded.

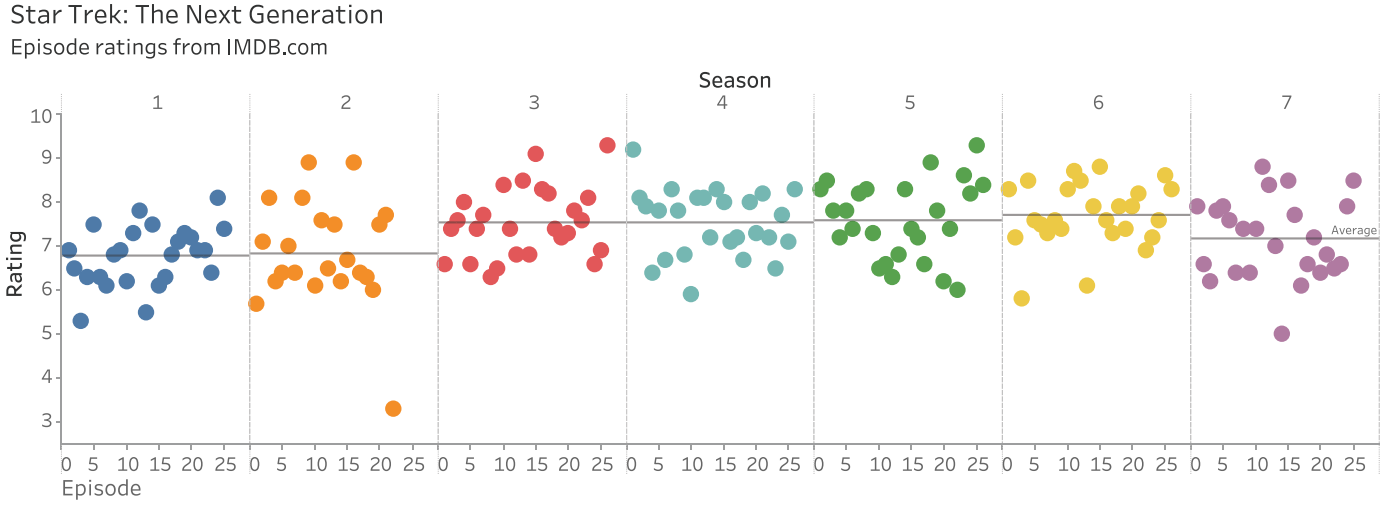


FIGURE 1.12 Every episode of *Star Trek: The Next Generation* rated.
Source: IMDB.com

TABLE 1.3 Data used in Figure 1.12.

Data	Data Type	Encoding	Note
Episode	Categorical	Position	Each episode is represented by a dot. Each dot has its own position on the canvas.
Episode Number	Ordinal	Position	The x-axis shows the number of each episode in each season.
Season	Ordinal	Color Position	Each season is represented by a different color (hue). Each season also has its own section on the chart.
IMDB rating	Ordinal	Position	The better the episode, the higher it is on the y-axis.
Average season rating	Quantitative	Position	The horizontal bar in each pane shows the average rating of the episodes in each season. There is some controversy over whether you should average ordinal ratings. We believe that the practice is so common with ratings it is acceptable.

Let's look at a few more charts to see how preattentive features have been used. Figure 1.13 is from *The Economist*. Look at each chart and see if you can work out which types of data are being graphed and how they are being encoded.

Table 1.4 shows how each data type is encoded.

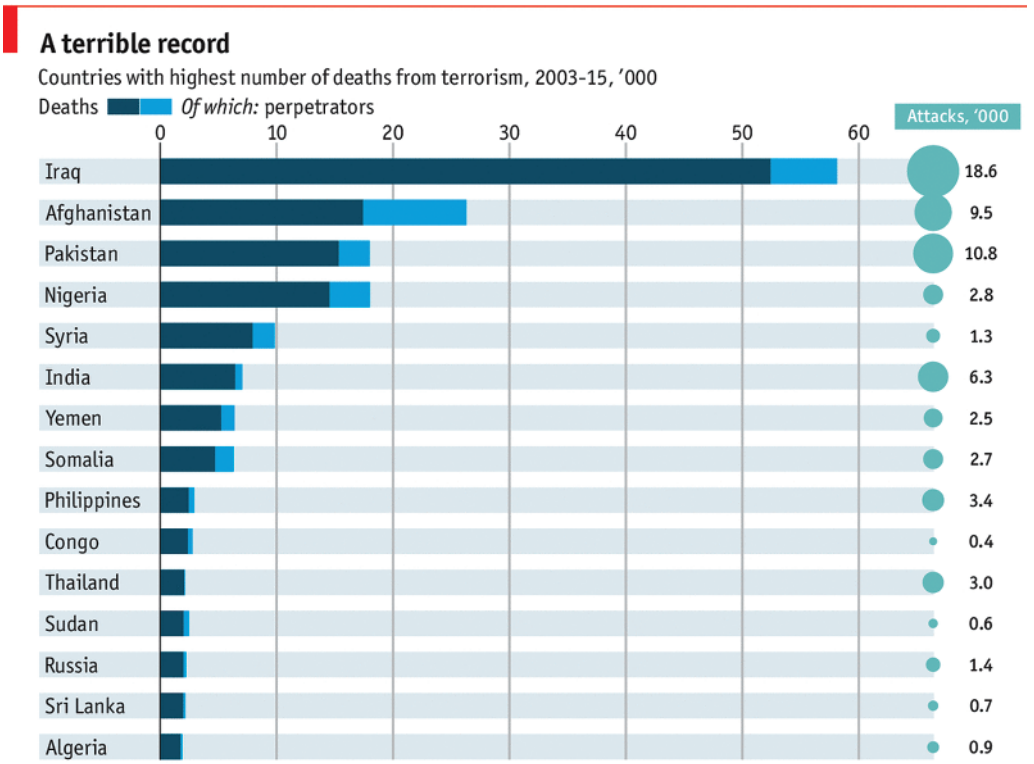


FIGURE 1.13 "A terrible record" from *The Economist*, July 2016.
Source: START, University of Maryland. *The Economist*, http://tabsoft.co/2agK3if

TABLE 1.4 Data used in the bar chart in Figure 1.13.

Data	Data Type	Encoding	Note
Country	Categorical	Position	Each country is on its own row (sorted by total deaths).
Deaths	Quantitative	Length	The length of the bar shows the number of deaths.
Death type	Categorical	Color	Dark blue shows deaths of victims, light blue shows deaths of the perpetrators.
Attacks	Quantitative	Size	Circles on the right are sized according to the number of attacks.

Let's look at another example. Figure 1.14 was part of the Makeover Monday project run by Andy Cotgreave and Andy Kriebel throughout 2016. This entry was by Dan Harrison. It takes data on malaria deaths from the World Health Organization. Table 1.5 describes the data used in the chart.

How did you do? As you progress through the book, stop and analyze some of the views in the scenarios: Think about which data types are being used and how they have been encoded.

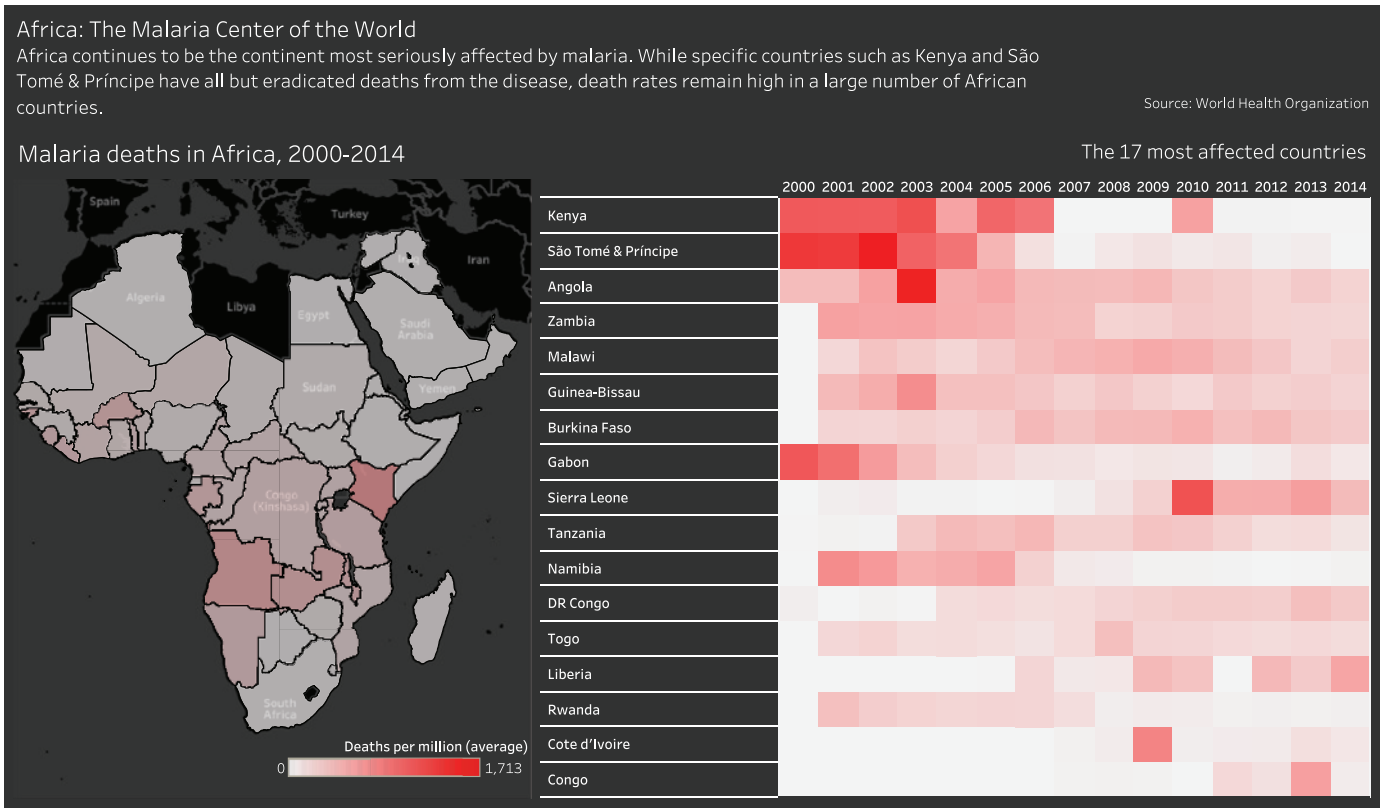


FIGURE 1.14 Deaths from malaria, 2000–2014.
Source: World Health Organization. Chart part of the Makeover Monday project

TABLE 1.5 Data used in the bar chart in Figure 1.14.

Data	Data Type	Encoding	Note
Country	Categorical	Position	The map shows the position of each country. In the highlight table, each country has its own row.
Deaths per million	Quantitative	Color	The map and table use the same color legend to show deaths per million people.
Year	Ordinal	Position	Each year is a discrete column in the table.

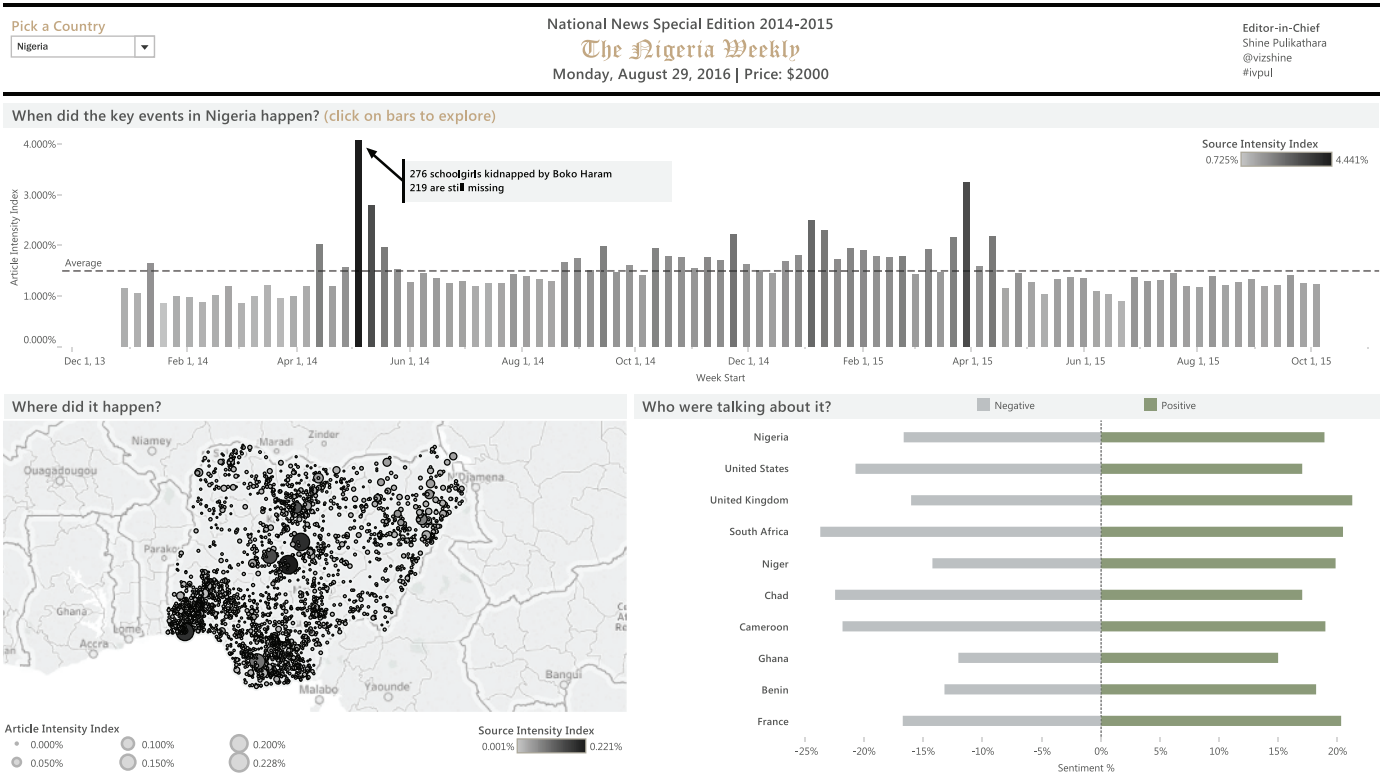


FIGURE 1.15 Winning visualization by Shine Pulikathara during the 2015 Tableau Iron Viz competition.
Source: Used with permission from Shine Pulikathara.

Color

Color is one of the most important things to understand in data visualization and frequently is misused. You should not use color just to spice up a boring visualization. In fact, many great data visualizations don't use color at all and are informative and beautiful.

In Figure 1.15, we see Shine Pulikathara's visualization that won the 2015 Tableau Iron Viz competition. Notice his simple use of color.

Color should be used purposefully. For example, color can be used to draw the attention of the reader, highlight a portion of data, or distinguish between different categories.

Use of Color

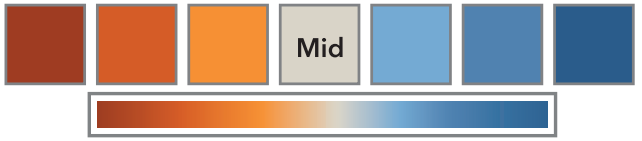
Color should be used in data visualization in three primary ways: *sequential*, *diverging*, and *categorical*.

In addition, there is often the need to *highlight* data or *alert* the reader of something important. Figure 1.16 offers an example of each of these color schemes.

Use of Color in Data Visualization

Sequential
color is ordered from low to high

Diverging
two sequential colors with a neutral midpoint



Categorical
contrasting colors for individual comparison

Highlight
color used to highlight something

Alert
color used to alert or warn reader

FIGURE 1.16 Use of color in data visualization.

Sequential color is the use of a single color from light to dark. An example is encoding the total amount of sales by state in blue, where the darker blue shows higher sales and a lighter blue shows lower sales. Figure 1.17 shows the unemployment rate by state using a sequential color scheme.

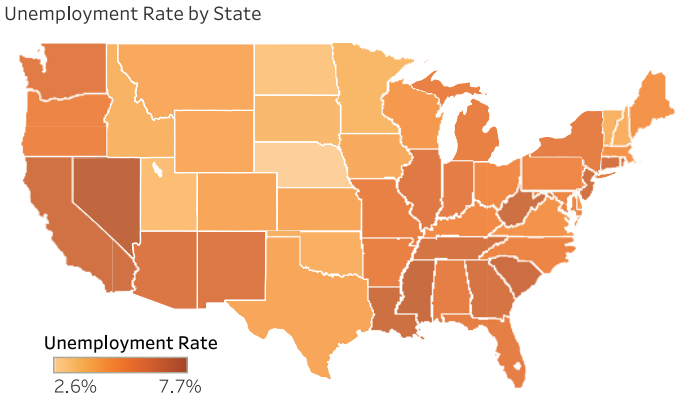


FIGURE 1.17 Unemployment rate by state using a sequential color scheme.

Diverging color is used to show a range diverging from a midpoint. This color can be used in the same manner as the sequential color scheme but can encode two different ranges of a measure (positive and negative) or a range of a measure between two categories. An example is the degree to which electorates may vote Democratic or Republican in each state, as shown in Figure 1.18.

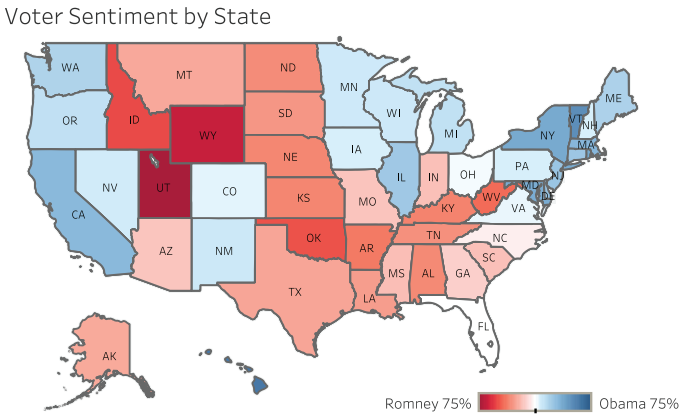


FIGURE 1.18 Degree of Democratic (blue) versus Republican (red) voter sentiment in each state.

Diverging color can also be used to show the weather, with blue showing the cooler temperatures and red showing the hotter temperatures. The midpoint can be the average, the target, or zero in cases where there are positive and negative numbers. Figure 1.19 shows an example with profit by state, where profit (positive number) is shown in blue and loss (negative number) is shown in orange.

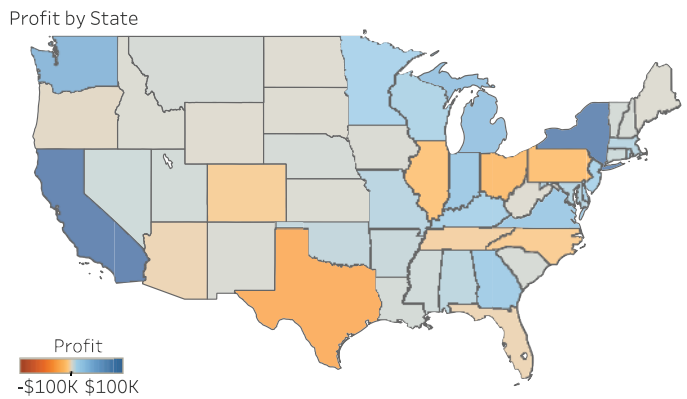


FIGURE 1.19 Profit by state using a diverging color scheme.

Categorical color uses different color hues to distinguish between different categories. For example, we can establish categories involving apparel (e.g., shoes, socks, shirts, hats, and coats) or vehicle types (e.g., cars, minivans, sport utility vehicles, and motorcycles). Figure 1.20 shows quantity of office supplies in three categories.

Highlight color is used when there is something that needs to stand out to the reader, but not alert or alarm them. Highlights can be used in a number of ways, as in highlighting a certain data point, text in a table, a certain line on a line chart, or a specific bar in a bar chart. Figure 1.21 shows a slopegraph with a single state highlighted in blue.

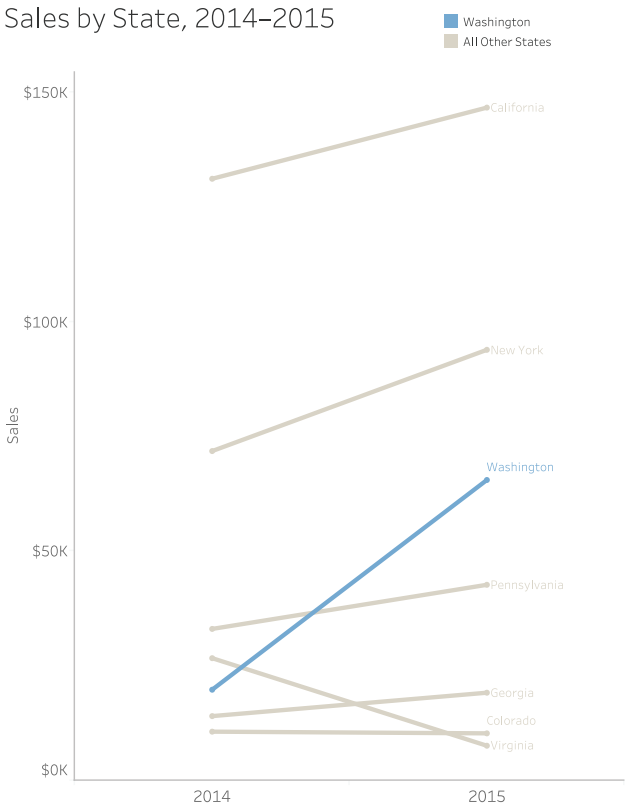


FIGURE 1.21 Slopegraph showing sales by state, 2014–2015, using a single color to highlight the state of Washington.

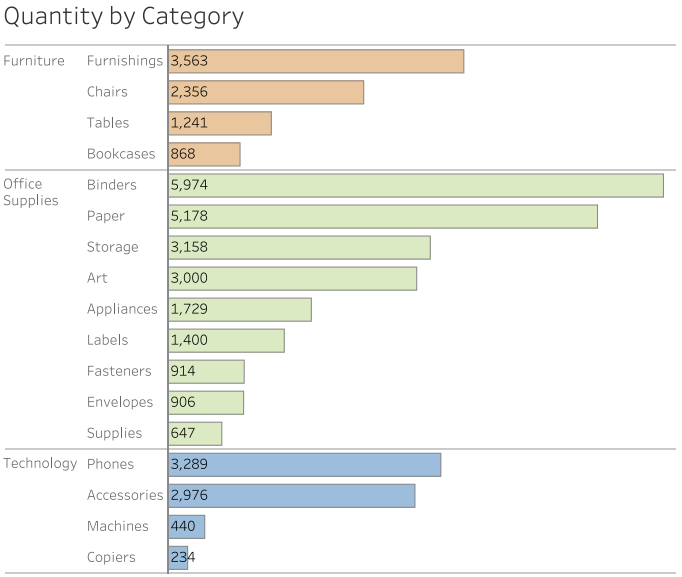


FIGURE 1.20 Quantity of office supplies in three categories using a categorical color scheme.

Alerting color is used when there is a need to draw attention to something for the reader. In this case, it's often best to use bright, alarming colors, which will quickly draw the reader's attention, as in Figure 1.22.

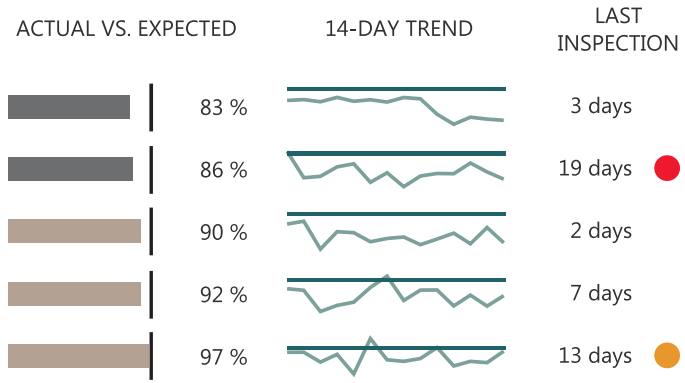


FIGURE 1.22 Red and orange indicators to alert the reader that something on the dashboard needs attention.

It is also possible to have a *categorical-sequential* color scheme. In this case, each category has a distinct hue that is darker or lighter depending on the measurement it is representing. Figure 1.23 shows an example of a four-region map using categorical colors (i.e., gray, blue, yellow, and brown) but at the same time encoding a measure in those regions using sequential color; let's assume that sales are higher in states with darker shading.

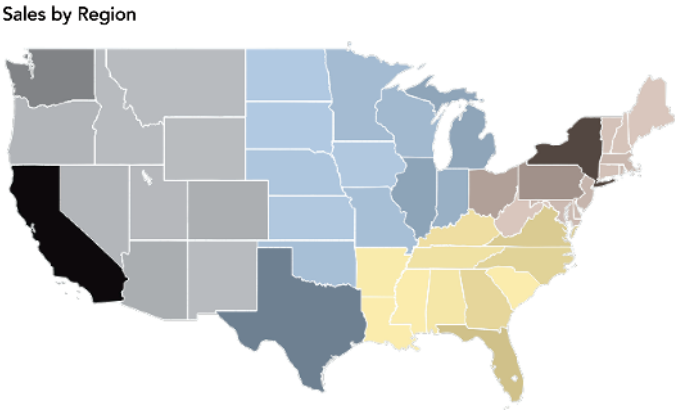


FIGURE 1.23 Sales by region using four categorical colors and the total sales shown with sequential color.

Color Vision Deficiency (Color Blindness)

Based on research (Birch 1993), approximately 8 percent of males have color vision deficiency (CVD) compared to only 0.4 percent of females. This deficiency is caused by a lack of one of three types of cones within the eye needed to see all color. The deficiency commonly is referred to as "color blindness", but that term isn't entirely accurate. People suffering from CVD can in fact see color, but they cannot distinguish colors in the same way as the rest of the population. The more accurate term is "color vision deficiency." Depending on which cone is lacking, it can be very difficult for people with CVD to distinguish between certain colors because of the way they see the color spectrum.

There are three types of CVD:

1. *Protanopia* is the lack of long-wave cones (red weak).
2. *Deuteranopia* is the lack of medium-wave cones (green weak).
3. *Tritanopia* is the lack of short-wave cones (blue). (This is very rare, affecting less than 0.5 percent of the population.)

CVD is mostly hereditary, and, as you can see from the numbers, it primarily afflicts men. Eight percent of men may seem like a small number, but consider that in a group of nine men, there is more than a 50 percent chance that one of them has CVD. In a group of 25 men, there is an 88 percent chance that one of them has CVD. The rates also increase among Caucasian men, reaching as high as 11 percent. In larger companies or when a data visualization is presented to the general public, designers must understand CVD and design with it in mind.

The primary problem among people with CVD is with the colors red and green. This is why it is best to avoid using red and green together and, in general, to avoid the commonly used traffic light colors. We discuss this issue further in Chapter 33 and offer some solutions for using red and green together.

Seeing the Problem for Yourself

Let's look at some examples of how poor choice of color can create confusion for people with CVD.

In Figure 1.24, the chart on the left uses the traditional traffic light colors red, yellow, and green. The example on the right is a protanopia simulation for CVD.

One common solution among data visualization practitioners is to use blue and orange. Using blue instead of green for *good* and orange instead of red for *bad* works well because almost everyone (with very rare exceptions) can distinguish blue and orange from each other. This blue-orange palette is often referred to as being "color-blind friendly."

Using Figure 1.25, compare the blue/orange color scheme and a protanopia simulation of CVD again.

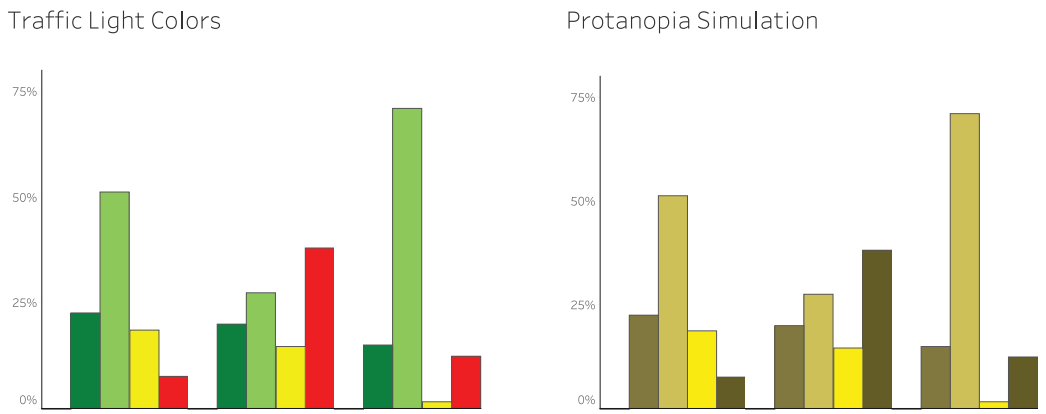


FIGURE 1.24 Bar chart using the traffic light colors and a protanopia simulation. Notice the red and green bars in the panel on the right are very difficult to differentiate from one another for a person with protanopia.

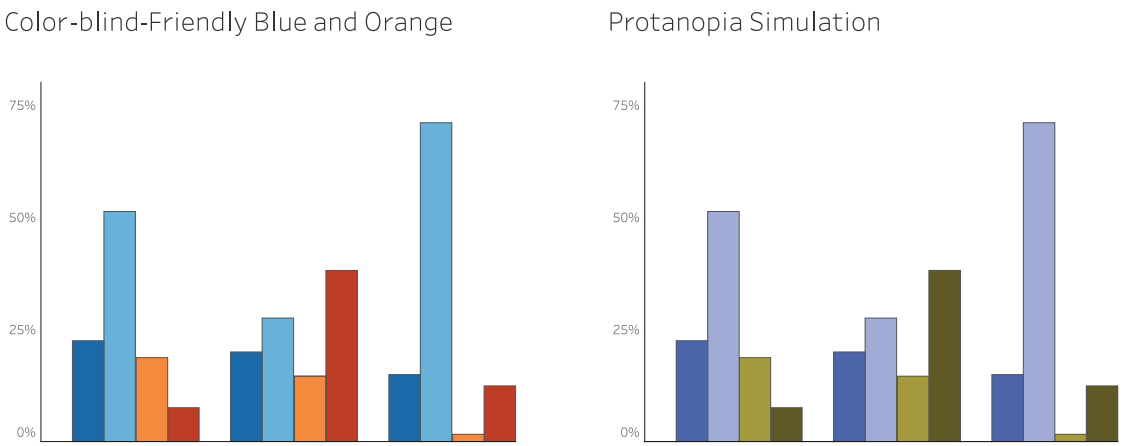


FIGURE 1.25 Bar chart using a color-blind-friendly blue and orange palette and a protanopia simulation.

The Problem Is Broader Than Just Red and Green

The use of red and green is discussed frequently in the field of data visualization, probably because the traffic light color palette is prevalent in many software programs and is commonly used in business today. It is common in Western culture to associate red with bad and green with good. However, it is important to understand that the problem in differentiating color for someone with CVD is much more complex than just red and green. Since red, green, and orange all appear to be brown for someone with strong CVD, it would be more accurate to say "Don't use red, green, brown, and orange together."

Figure 1.26 shows a scatterplot using brown, orange, and green together for three categories. When applying protanopia simulation, the dots in the scatterplot appear to be a very similar color.

One color combination that is frequently overlooked is blue and purple together. In a RGB (red-green-blue) color model, purple is achieved by using blue and red together. If someone with CVD has issues with red, then he or she may also have issues with purple, which would appear to look like blue. Other color combinations can be problematic as well. For example, people may have difficulty with pink or red used with gray or gray used together with brown.

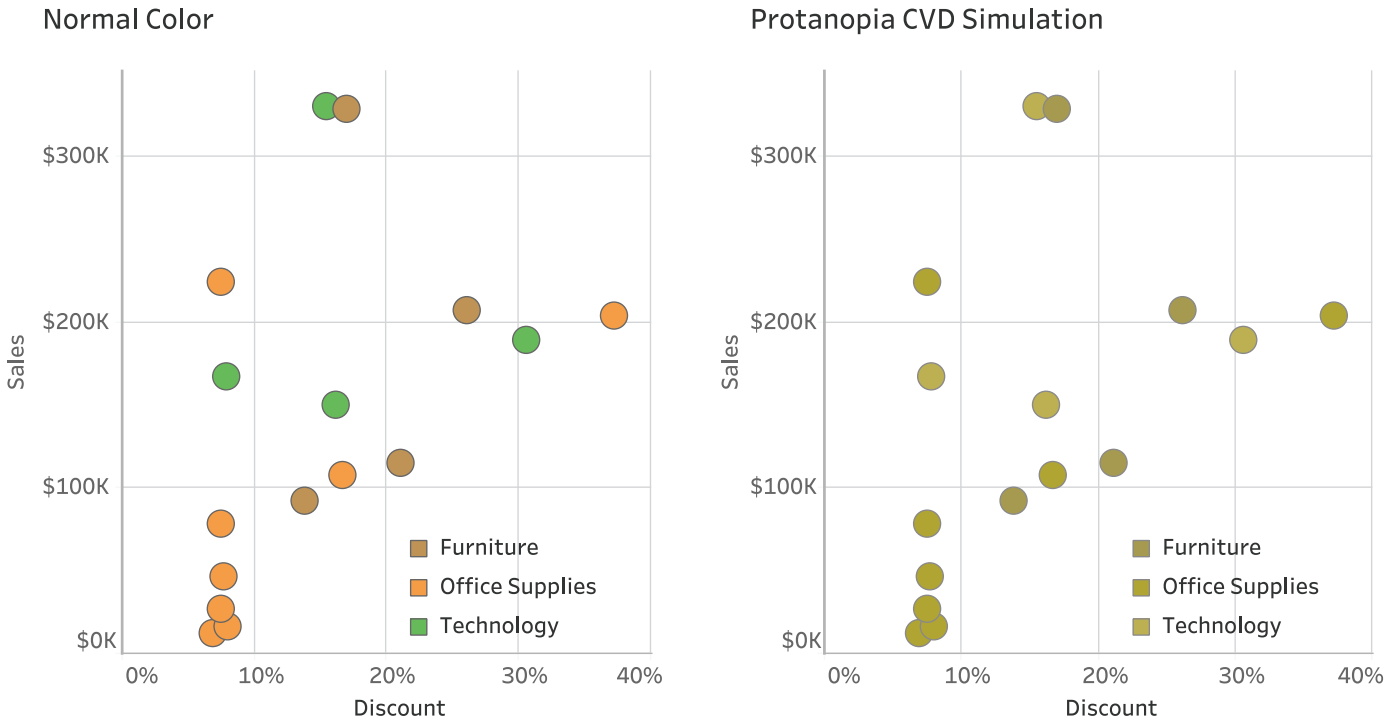


FIGURE 1.26 Scatterplot simulating color vision deficiency for someone with protanopia.

Figure 1.27 shows another scatterplot, this time using blue, purple, magenta, and gray. When applying deuteranopia simulation, the dots in the scatterplot appear to be a very similar color of gray.

It's important to understand these issues when designing visualizations. If color is used to encode data and it's necessary for readers to distinguish among colors to understand the visualization, then consider using color-blind-friendly palettes. Here are a few resources that you can use to simulate the various types of CVD for your own visualizations.

Adobe Illustrator CC. This program offers a built-in CVD simulation in the View menu under Proof Setup.

Chromatic Vision Simulator (free). Kazunori Asada's superb website allows users to upload images and simulate how they would appear to people with different form of CVD. See http://asada.tukusi.ne.jp/webCVS/

NoCoffee vision simulator (free). This free simulator for the Chrome browser allows users to simulate websites and images directly from the browser.

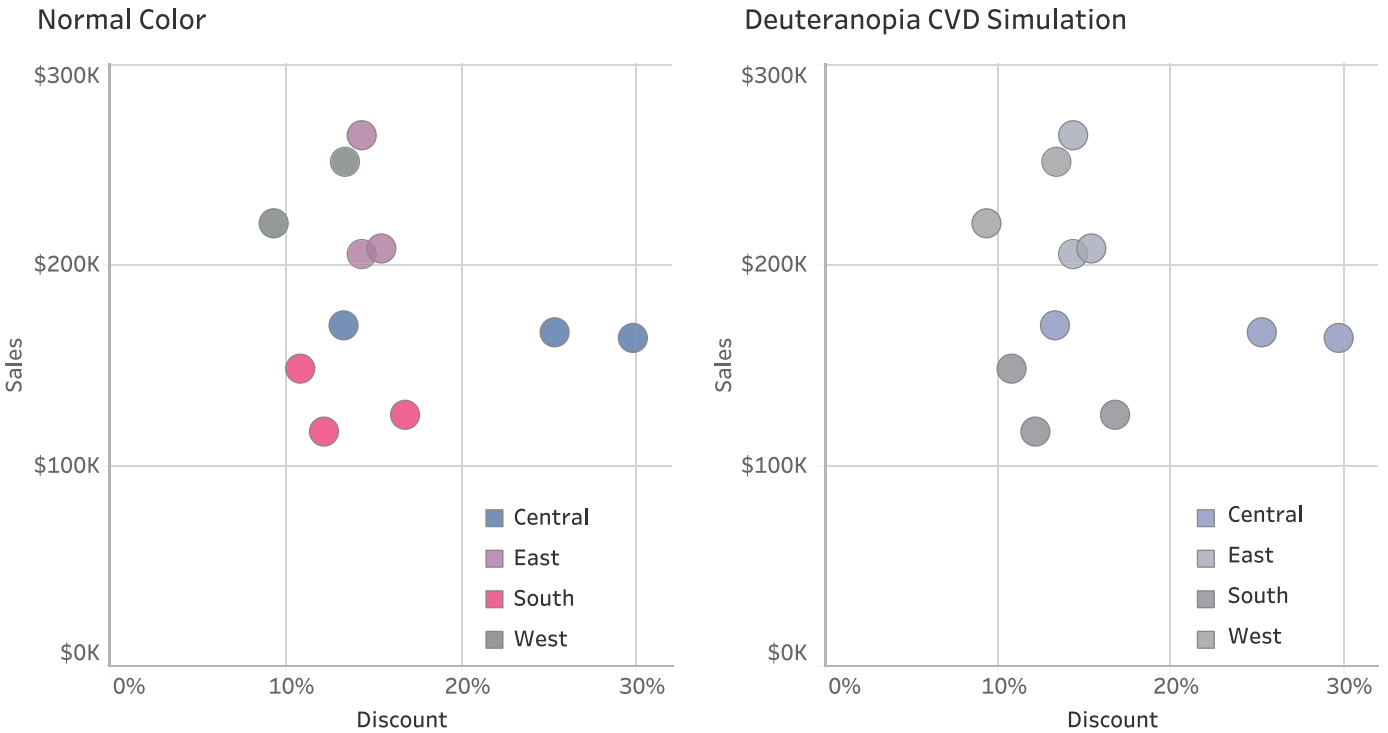


FIGURE 1.27 Scatterplot simulating color vision deficiency for someone with deuteranopia.

Common Chart Types

In this book, you will see many different types of charts. We explain in the scenarios why many of the charts were chosen to fulfill a particular task. In this section, we briefly outline the most common chart types. This list is intentionally short. Even if you use only the charts listed here, you would be able to cover the majority of needs when visualizing your data. More advanced chart types seen throughout the book are built from the same building blocks as these. For example, sparklines, which are shown in Chapters 6, 8, and 9, are a kind of line chart. Bullet charts, used in Chapter 17, are bar charts with reference lines and shading built in. Finally, waterfall charts, shown in Chapter 24, are bar charts where the bars don't have a common baseline.

Bar Chart

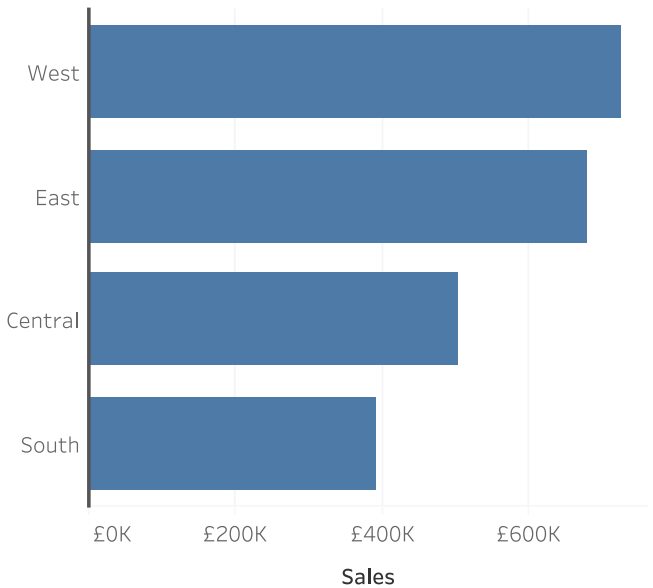


FIGURE 1.28 Bar chart.

A bar chart (see Figure 1.28) uses length to represent a measure. Human beings are extremely good at seeing even small differences in length from a common baseline. Bars are widely used in data visualization because they are often the most effective way to compare categories. Bars can be oriented horizontally or vertically. Sorting them can be very helpful because the most common task when bar charts are used is to spot the biggest/smallest items.

Time-Series Line Chart

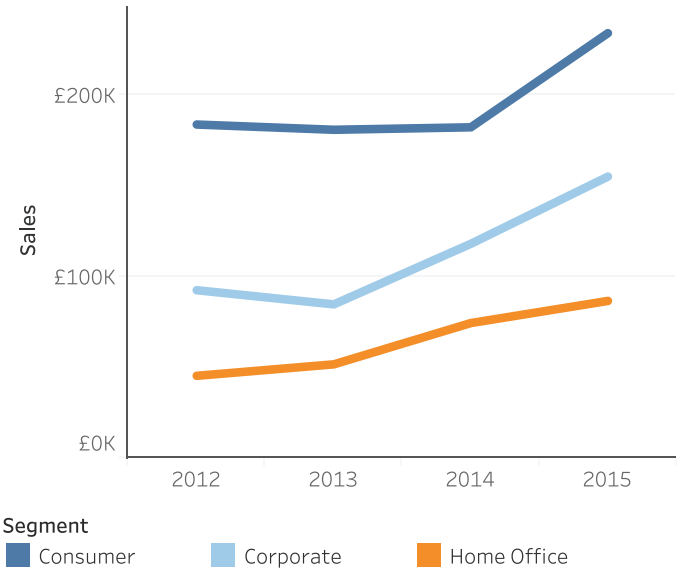


FIGURE 1.29 Time-series line chart.

Line charts (see Figure 1.29) usually show change over time. Time is represented by position on the horizontal x-axis. The measures are shown on the vertical y-axis. The height and slopes of the line let us see trends.

Scatterplot

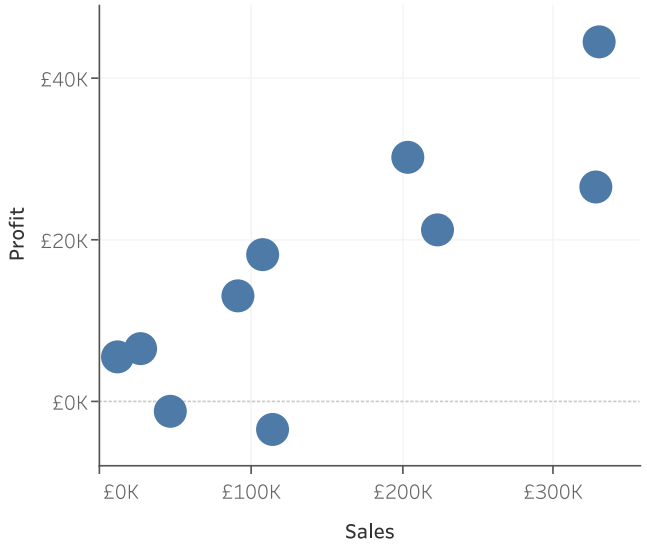


FIGURE 1.30 Scatterplot.

A scatterplot (see Figure 1.30) lets you compare two different measures. Each measure is encoded using position on the horizontal and vertical axes. Scatterplots are useful when looking for relationships between two variables.

Dot Plot

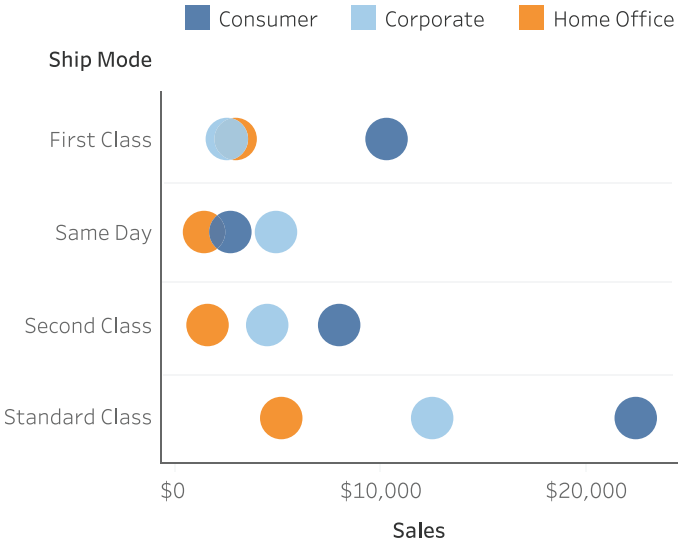


FIGURE 1.31 Dot plot.

A dot plot (see Figure 1.31) allows you to compare values across two dimensions. In our example, each row shows sales by ship mode. The dots show sales for each ship mode, broken down by each segment. In the example, you can see that corporate sales are highest with standard class ship mode.

Choropleth Map

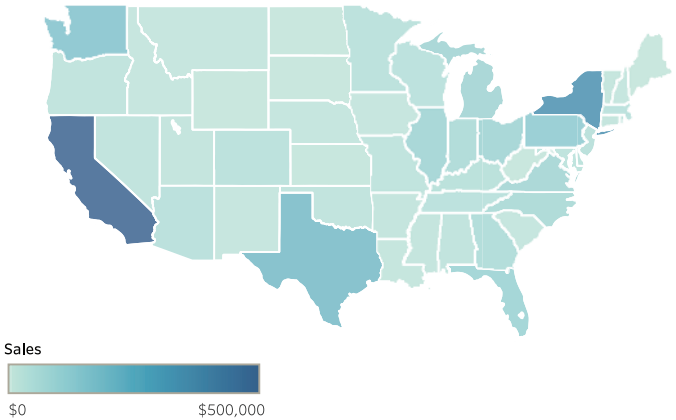


FIGURE 1.32 Choropleth map.

A choropleth (also known as a filled) map (see Figure 1.32) uses differences in shading or coloring within predefined areas to indicate the values or categories in those areas.

Symbol Map

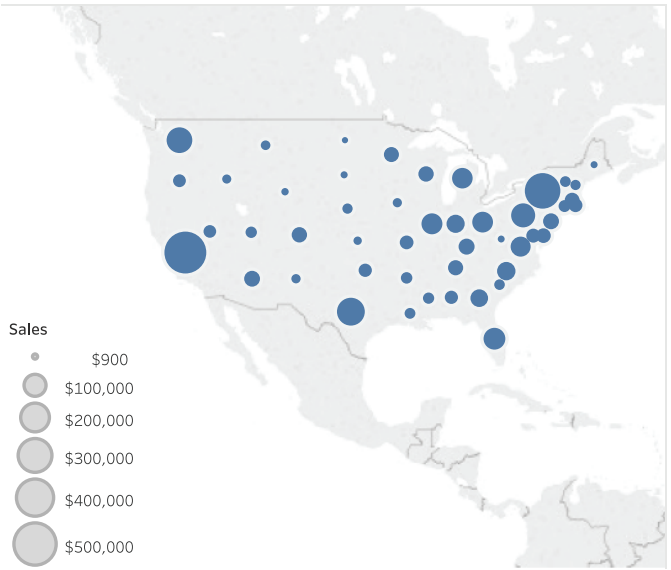


FIGURE 1.33 Symbol map.

A symbol map (see Figure 1.33) shows values in specific places. These could be the center points of large regions (e.g., the center of each U.S. state) or specific locations determined by an exact latitude/longitude measurement.

Avoid pie charts

Why isn't there a pie chart? Pie charts are common charts, but they are flawed. We don't recommend you use them. Check out the section titled "When Our Visual Processing System Betrays Us" for details.

Table

Sometimes you do need to be able to look up exact values. A table (see Figure 1.34) is an acceptable way to show data in that situation. On most dashboards, a table shows details alongside summary charts.

$111K	$131K	$138K	$154K
$132K	$117K	$157K	$215K
$77K	$68K	$79K	$106K

FIGURE 1.34 Table.

Highlight Table

Adding a color encoding to your tables can transform them into highly visual views that also enable exact lookup of any value. (see Figure 1.35.)

$111K	$131K	$138K	$154K
$132K	$117K	$157K	$215K
$77K	$68K	$79K	$106K

FIGURE 1.35 Highlight table.

Bullet Graph

A bullet graph (see Figure 1.36) is one of the best ways to show actual versus target comparisons. The blue bar represents the actual value, the black line shows the target value, and the areas of gray shading are performance bands.

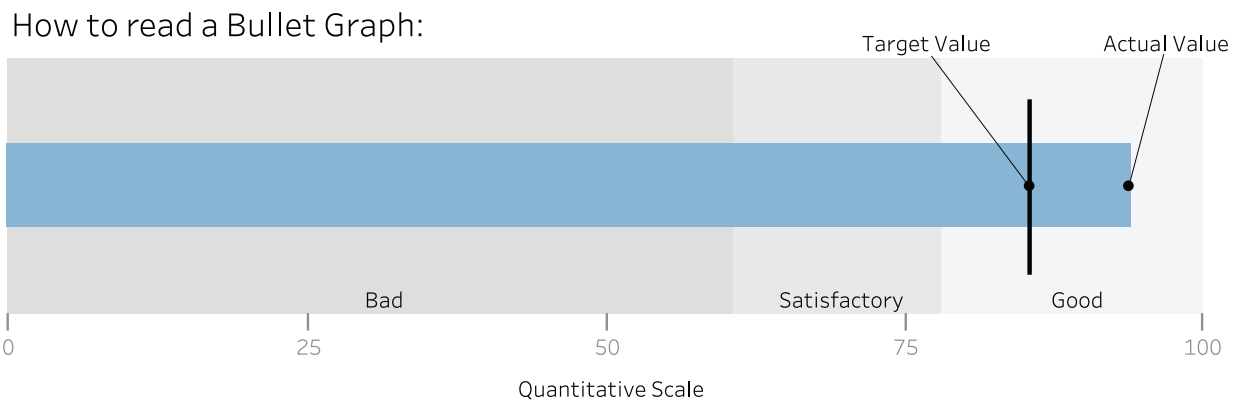


FIGURE 1.36 Bullet Graph.

When Our Visual Processing System Betrays Us

We have talked about how to use preattentive attributes to craft good data visualizations. By using those attributes, we can use the power of our visual system to our advantage. Unfortunately, our visual system also can be confused easily. In this section, we look at some common pitfalls.

Our eyes can be fooled in countless different ways. Figures 1.37 and 1.38 show two optical illusions.

In Figure 1.38, the top appears to be a well-lit gray surface and the bottom appears to be a poorly lit white surface that is in shadow. However, there is no shadow. Dr. Lotto added the gradient and shadows to the image. Our minds can't help but to see the shadow, making the top appear to be much darker than the bottom, but if you cover up the middle of the image, it becomes clear that the top and the bottom are exactly the same color.

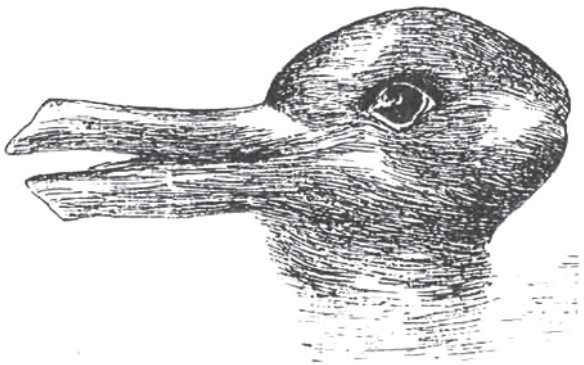


FIGURE 1.37 Is it a duck or a rabbit?

Source: Public domain. https://commons.wikimedia.org/w/index.php?curid=667017

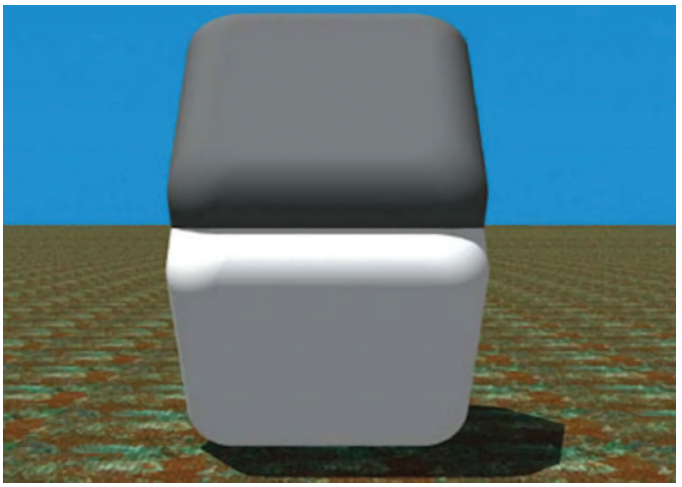


FIGURE 1.38 Does the top appear darker than the bottom? Put your thumb or finger over the center line and then try again.

Source: Image by R. Beau Lotto.

Ambiguity in images makes for playful illusions, but this can be disastrous if your data visualizations confuse instead of clarify. In the previous section, we looked at the power of preattentive attributes. Now it's time to look into the problems with some preattentive attributes. Throughout the book, we discuss which preattentive attributes are being used in the scenarios and why they work in each case.

When we visualize data, we are, for the most part, trying to convey the value of the measure in a way that can be interpreted most accurately in the shortest time possible. Some preattentive attributes are better than others for this purpose.

Figure 1.39 shows the number of deaths each day from various diseases in Africa. Each circle is sized according to the number of deaths. We have removed all the labels except the one for malaria (552 deaths per day). How many deaths per day are there from diarrhea? How much bigger is the HIV/AIDS circle than the diarrhea circle?

How did you do? The actual answers are shown in Figure 1.40.

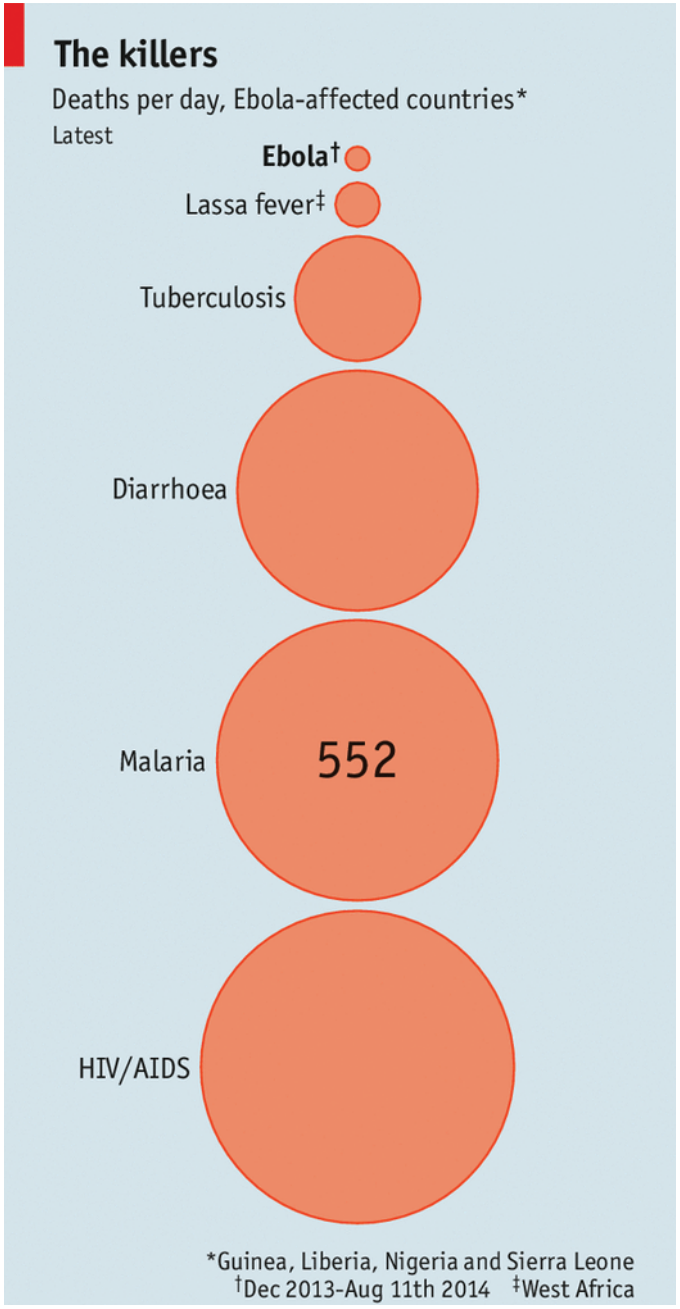


FIGURE 1.39 Deaths per day in Ebola-affected countries in 2014. We removed all labels except the one for malaria. Can you estimate the number of deaths from the other diseases?

Source: World Health Organization; U.S. Centers for Disease Control and Prevention; *The Economist*, http://tabsoft.co/1w1vwAc

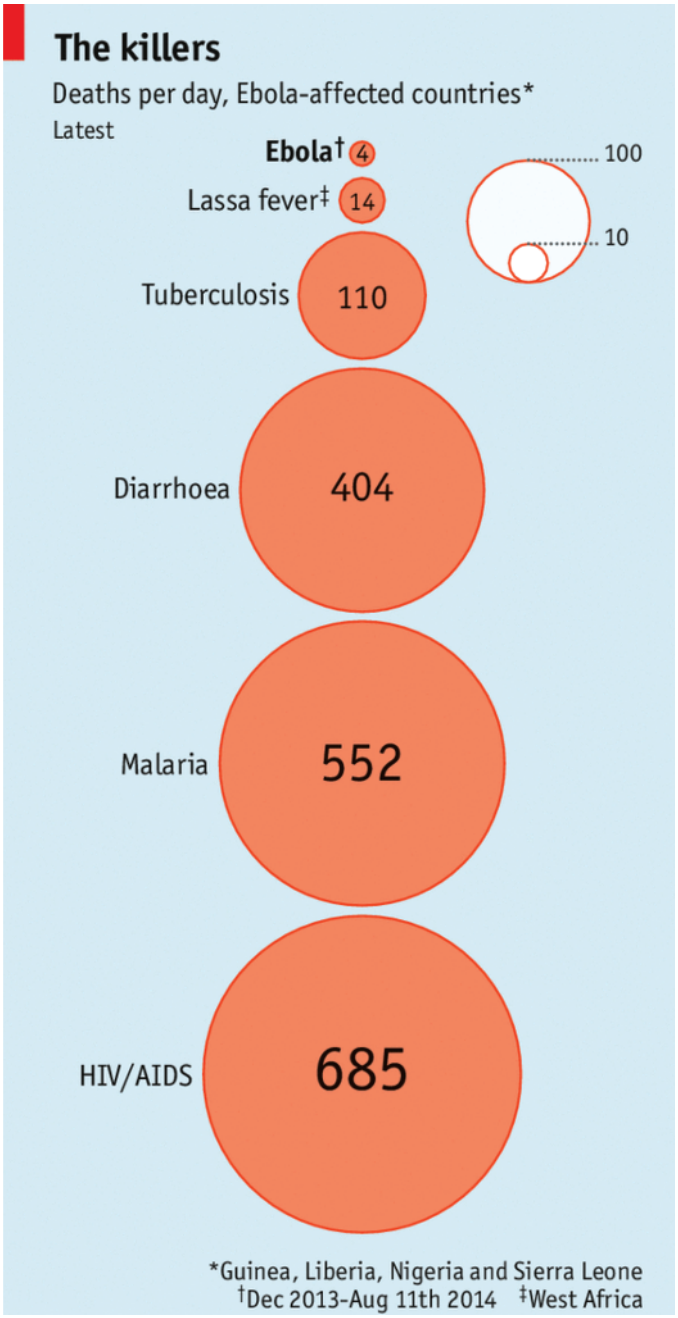


FIGURE 1.40 Deaths per day of various diseases in Ebola-affected countries, with labels added.

Source: World Health Organization; U.S. Centers for Disease Control and Prevention; *The Economist*, http://tabsoft.co/1w1vwAc

Most people underestimate the size of the bigger circles. The point is that while size is preattentive, we're not able to tell the differences with any accuracy. Consider the same data shown as a bar chart in Figure 1.41.

In the bar chart, we are encoding the quantitative variable, deaths per day, using length. Notice how accurately you can see the differences. This is why the bar chart is such a reliable chart to use: Length is one of the most efficient preattentive attributes for us to process.

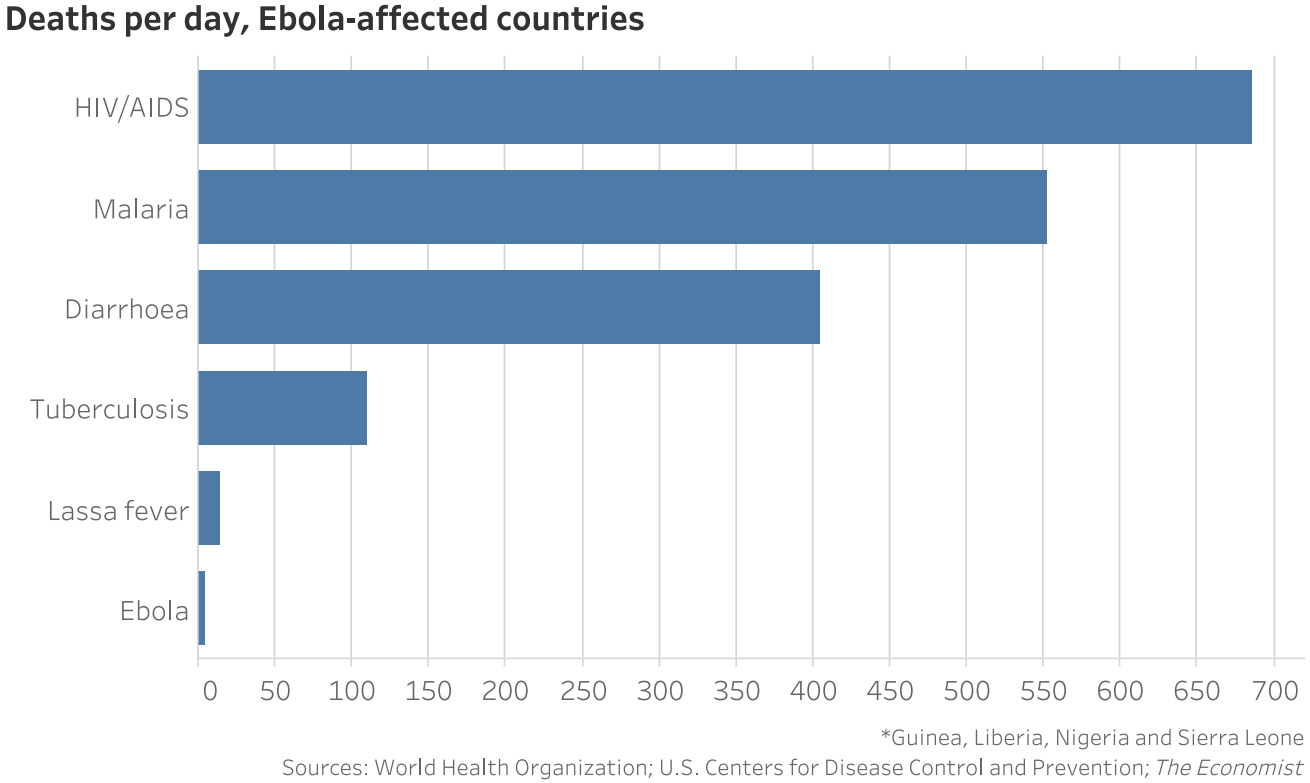


FIGURE 1.41 Bar chart version of the circle charts.

Yet using multiple preattentive attributes in one chart can lead to problems. Figure 1.42 shows a scatterplot of sales and profit for a fictional sales company. Position is used for sales (x-axis) and profit (y-axis). Color shows different segments, and shape shows the categories of products. Which category has, on average, the highest profits?

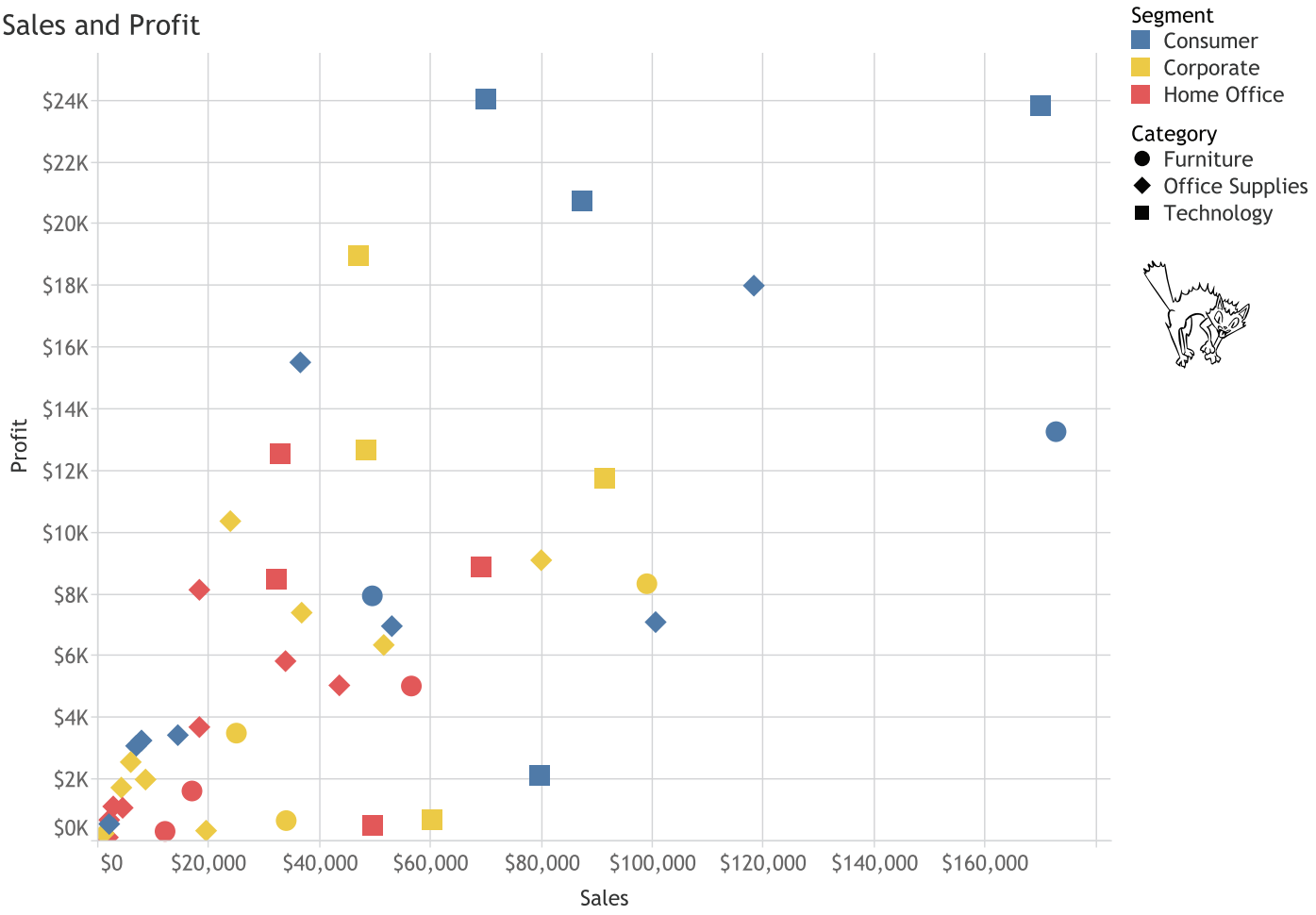


FIGURE 1.42 Scatterplot using shape and color. Which category has the highest profits?

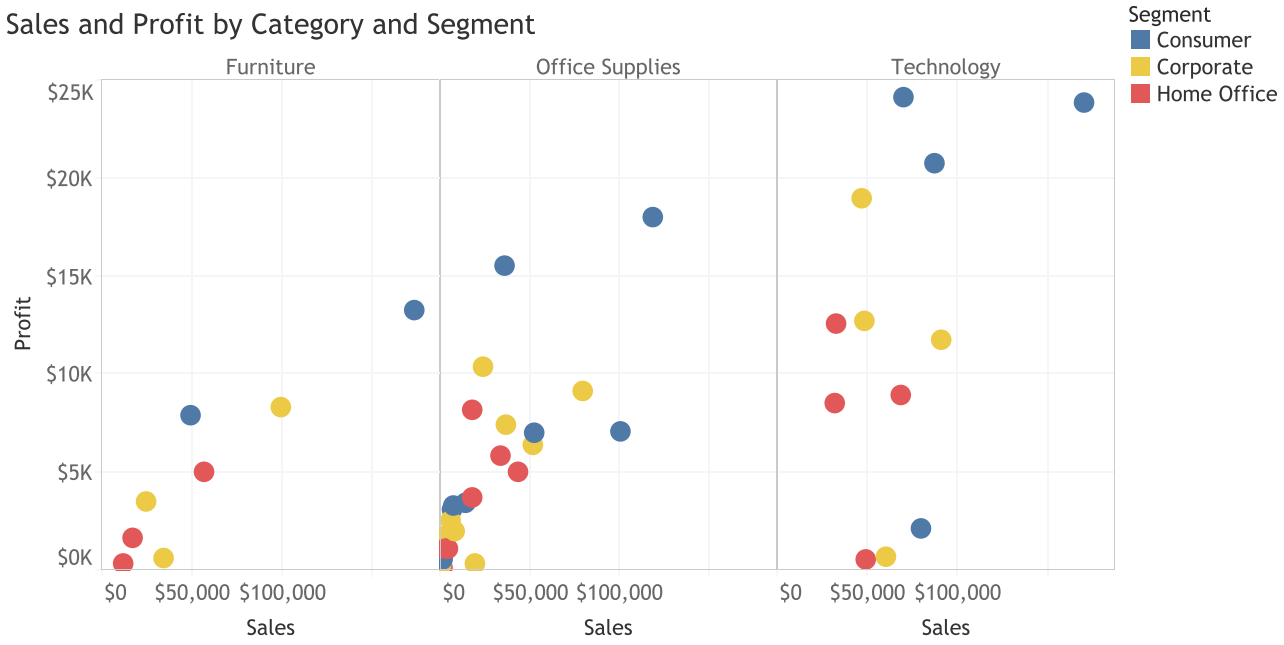


FIGURE 1.43 Sales and profit with one column for each category.

It's almost impossible to see anything, isn't it? Mixing position, color, and shape does not make for easy reading. What's a better solution? How about using position to represent category, breaking the single scatterplot into three panels? This is shown in Figure 1.43.

The result is much clearer. Now you can even see that technology sales, on average, have a higher range of profits than furniture and office supplies. That insight was certainly not apparent in the first scatterplot.

To close this section, let's look at some chart types you might be surprised *not* to see in our common chart types. The first is the pie chart. Sure, it is a common chart, but we do not recommend you use it. Let's see why pie charts don't play well with our visual system.

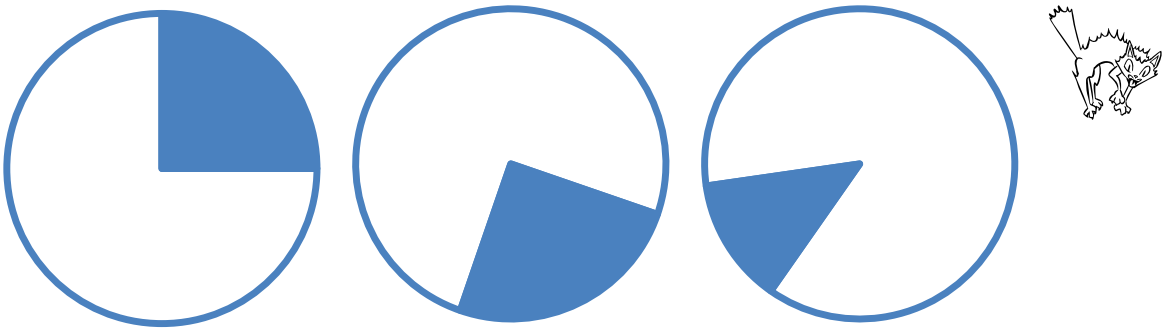


FIGURE 1.44 What percentage of each pie does the blue segment represent?

Look at Figure 1.44. What percentage of each circle is covered by the blue segment?

The one on the left is pretty easy: 25 percent. The middle? It's a little harder. It's also 25 percent, but because it's not aligned to a horizontal or vertical axis, it's harder to determine. And on the right? It's 13 percent. How did you do? We are simply not able to make accurate estimates of angle sizes, and if accurate estimates are the goal, it's a problem.

Let's look at another pie. The biggest slice in Figure 1.45 is easy to spot. But what about the second, third, and fourth biggest slices?

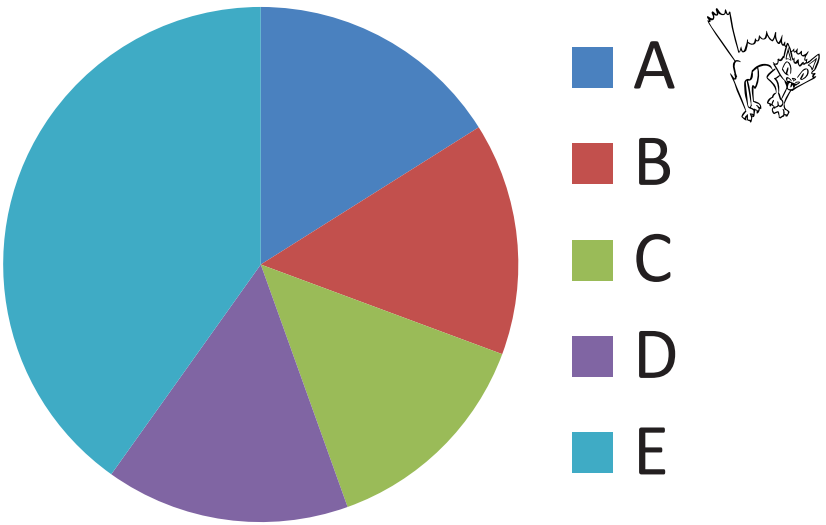


FIGURE 1.45 Can you order the slices from biggest to smallest?

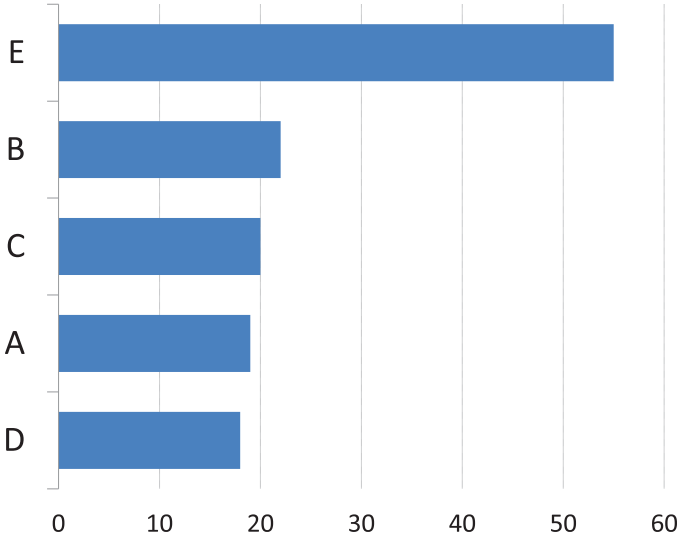


FIGURE 1.46 Bars make it very easy to see small differences in size.

That was really hard. Now look at the same data, shown in a bar chart, in Figure 1.46.

The sorted bar chart made it very easy to distinguish size differences: Length is such an effective visual attribute, we can see very small differences with ease.

To make effective dashboards, you must resist the temptation to use purely decorative chart types.

Let's look at one more example in order to keep you away from the lure of the circles. Sometimes people acknowledge the power of bars but then get tempted to put them in a circle, fashioning what is known as a donut chart. Figure 1.47 shows an example.

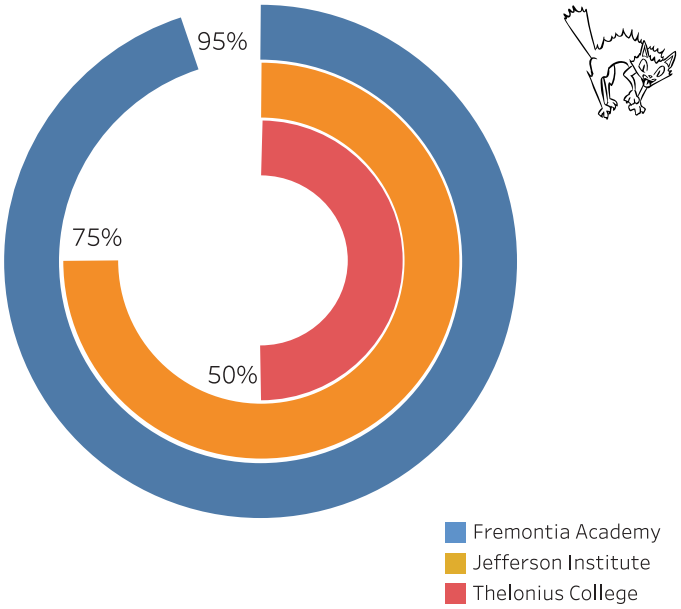


FIGURE 1.47 A concentric donut chart (also called a radial bar chart).

"What's the problem?" you may ask. "The comparison seems easy."

Although you may be able to make the comparisons, you are in fact working considerably harder than you need to be. Really. Let us prove it to you.

Let's suppose you wanted to compare the heights of three famous buildings: One World Trade Center, the Empire State Building, and the Chrysler Building. (See Figure 1.48.)

Now, that's an easy comparison. With virtually no effort, we can see that One World Trade Center (blue) is almost twice as tall as the Chrysler Building (red).

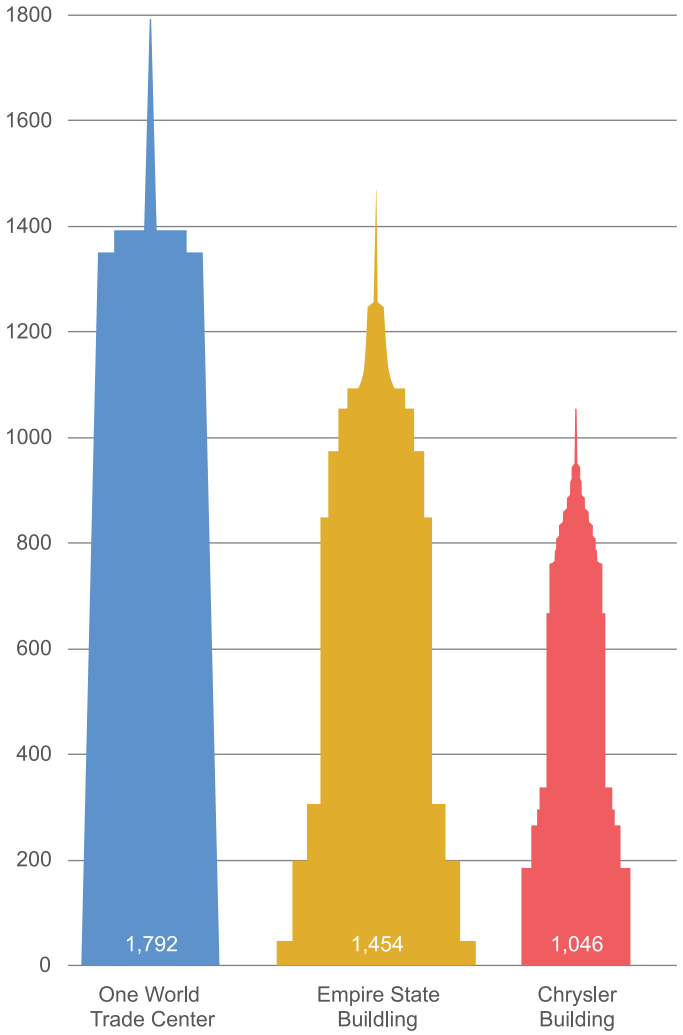


FIGURE 1.48 Comparing the size (in feet) of three large buildings.

Now let's see how easy the comparison is with donuts. (See Figure 1.49.)

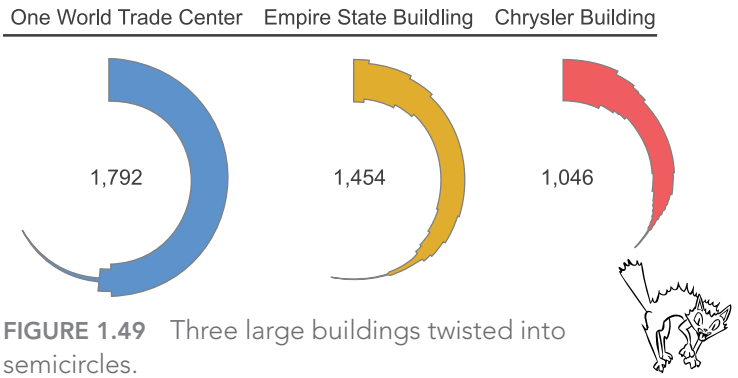


FIGURE 1.49 Three large buildings twisted into semicircles.

Figure 1.50 presents the same buildings rendered using a concentric donut chart. Can you tell the difference in heights of the buildings in this chart?

FIGURE 1.50 Three skyscrapers spooning.

Yikes!

So, with this somewhat contrived but hopefully memorable example, we took something that was simple to compare (the silhouettes of buildings) and contorted them into difficult-to-compare semicircles.

Every Decision Is a Compromise

However you choose to show your data, you will emphasize one feature over another. Let's have a look at an example. Table 1.6 shows a table of numbers. Let's imagine they are sales for two products, A and B, over 10 years.

Figure 1.51 shows eight different ways to visualize this data. Each uses a different mix of preattentive attributes.

Notice the compromises in the charts labeled 1 and 2. A standard line chart (1) showing each product lets us compare each product's sales very accurately. The area chart (2) lets us see total sales over time with ease, but now it is harder to compare the two products. You can't, in a single chart, answer every possible question or comparison. What you do need to do is assess whether the chart you do choose answers the question being asked.

TABLE 1.6 How would you visualize this data?

	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	Total
A	100	110	140	170	120	190	220	250	240	300	1,840
B	80	70	50	100	130	180	220	160	260	370	1,620

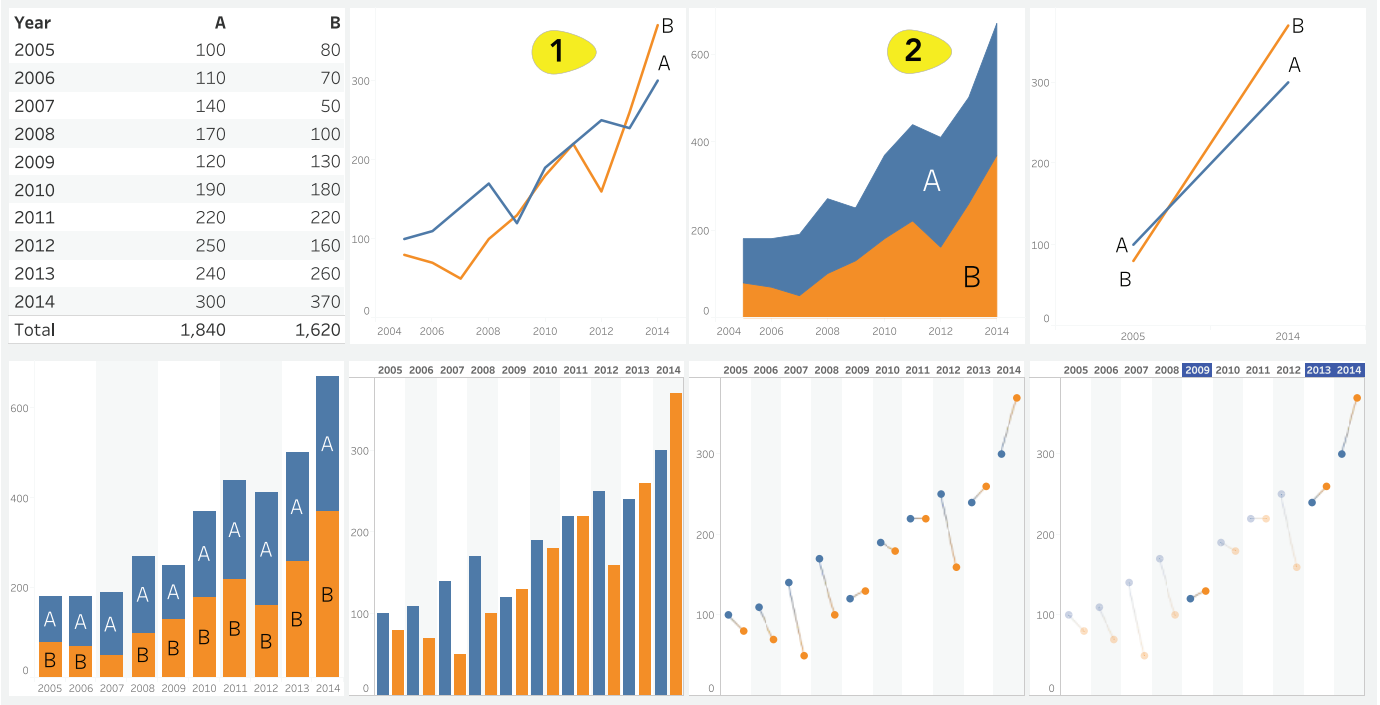


FIGURE 1.51 Eight different ways of visualizing the data.

Designing Dashboards That Are Functional and Beautiful

You now have a suitable vocabulary to interpret the charts in the scenarios in this book. Note that we have not offered a primer on graphic design. Instead, in each scenario, we point out where and how graphic design elements, such as white space, fonts, grid layout, and so on, contribute to the clarity of the dashboards.

We maintain that a dashboard must first be truthful and functional, but there are reasons you should go the extra mile to make dashboards that are elegant as well. We recommend considering the lessons from classic design books, such as *The Design of Everyday Things* by Donald A. Norman (Basic Books, 2013). In it, Norman says:

> Products [should] actually fulfill human needs while being understandable and usable. In the best of cases, the products should also be delightful and enjoyable, which means that not only must the requirements of engineering, manufacturing, and ergonomics be satisfied, but attention must be paid to the entire experience, which means the aesthetics of form and the quality of interaction. (p. 4)

Summary

This chapter has gone through the basics of data visualization. If you are new to visualization, you now have enough knowledge to interpret the charts in this book. You will be able to decode most of the charts you encounter. There is also a glossary of charts at the back of the book for further reference.

You might be inspired to find out more. There are many superb books on the theory and application of this science. Some of the examples in this chapter are based on examples first used in some of these books. Here are our recommendations:

Alberto Cairo's *The Functional Art* (New Riders, 2013). Alberto Cairo is an author who understands the need to balance functionality with beauty in charts. This book is an inspiring introduction to information graphics and visualization.

Stephen Few's *Now You See It* (Analytics Press, 2009). This is a practical and commonsense guide to table and graph design. It goes into great detail about each of the main chart types, explaining clearly when to use them and how to construct them well.

Cole Nussbaumer Knaflic's *Storytelling with Data* (Wiley, 2015). This is the data visualization guide for business professionals. It's an accessible look at not only the anatomy of charts but also at how to design charts to communicate messages effectively.

Colin Ware's *Information Visualization: Perception for Design* (Morgan Kauffman, 2013). This book has been called the bible of data visualization. In over 500 pages, it covers every aspect of the science of perception and its role in data visualization. It's an invaluable resource for anyone practicing data visualization.

Colin Ware's *Visual Thinking for Design* (Elsevier, 2008). Colin Ware presents a detailed analysis of the mechanics of visual cognition. The book teaches us how to see as designers by anticipating how others will see our designs. It's a fun book to read and makes detailed information about cognitive science a breeze to digest.