Statistical Thinking for the 21st Century

Chapter 4

Data Visualization

4.2 Principles of good visualization

Many books have been written on effective visualization of data. There are some principles that most of these authors agree on, while others are more contentious. Here we summarize some of the major principles.

4.2.1 Show the data and make them stand out

Let's say that I performed a study that examined the relationship between dental health and time spent flossing, and I would like to visualize my data. Figure 4.4 shows four possible presentations of these data.

1. In panel A, we don't actually show the data, just a line expressing the relationship between the data. This is clearly not optimal, because we can't actually see what the underlying data look like.

Panels B-D show three possible outcomes from plotting the actual data, where each plot shows a different way that the data might have looked.

2. If we saw the plot in Panel B, we would probably be suspicious – rarely would real data follow such a precise pattern.

3. The data in Panel C, on the other hand, look like real data – they show a general trend, but they are messy, as data in the world usually are.

4. The data in Panel D show us that the apparent relationship between the two variables is solely caused by one individual, who we would refer to as an *outlier* because they fall so far outside of the pattern of the rest of the group. It should be clear that we probably don't want to conclude very much from an effect that is driven by one data point. This figure highlights why it is *always* important to look at the raw data before putting too much faith in any summary of the data.

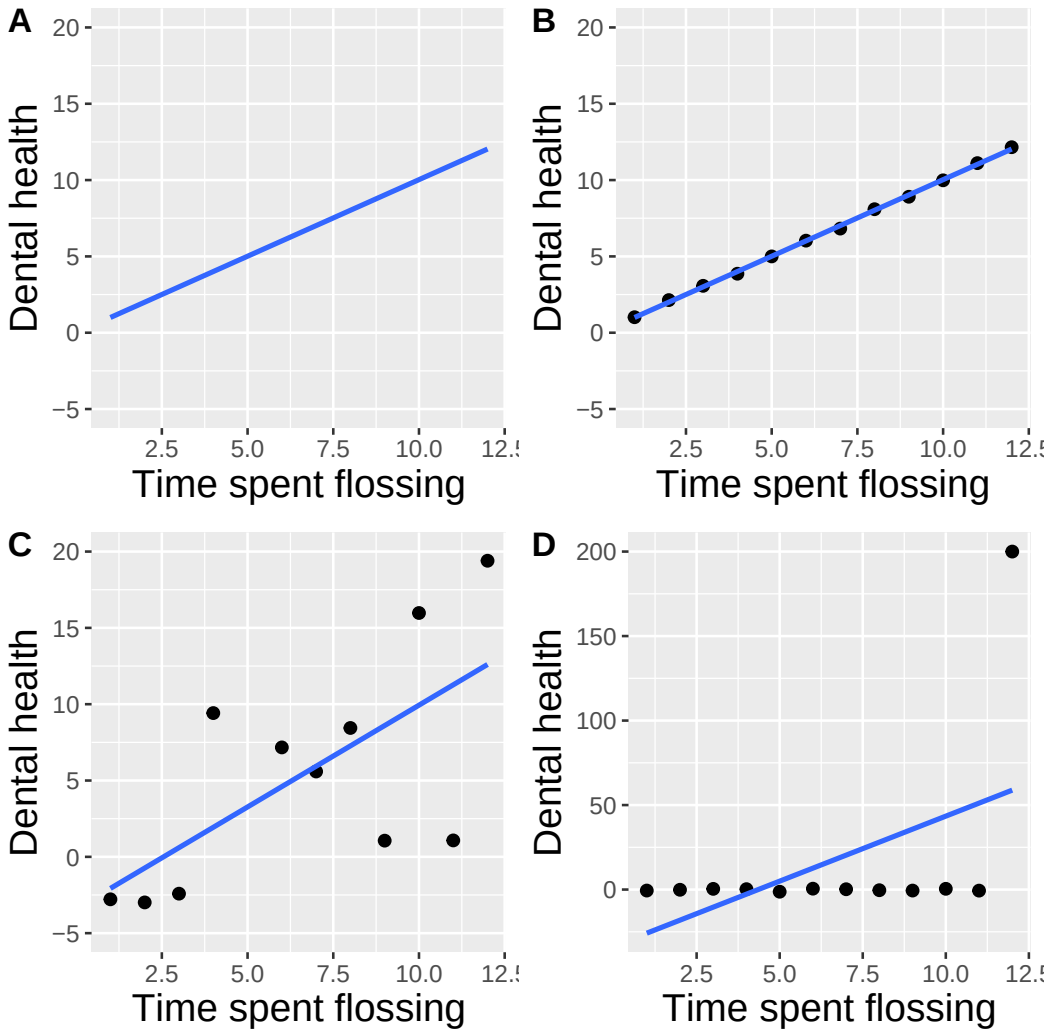

Figure 4.4: Four different possible presentations of data for the dental health example. Each point in the scatter plot represents one data point in the dataset, and the line in each plot represents the linear trend in the data.

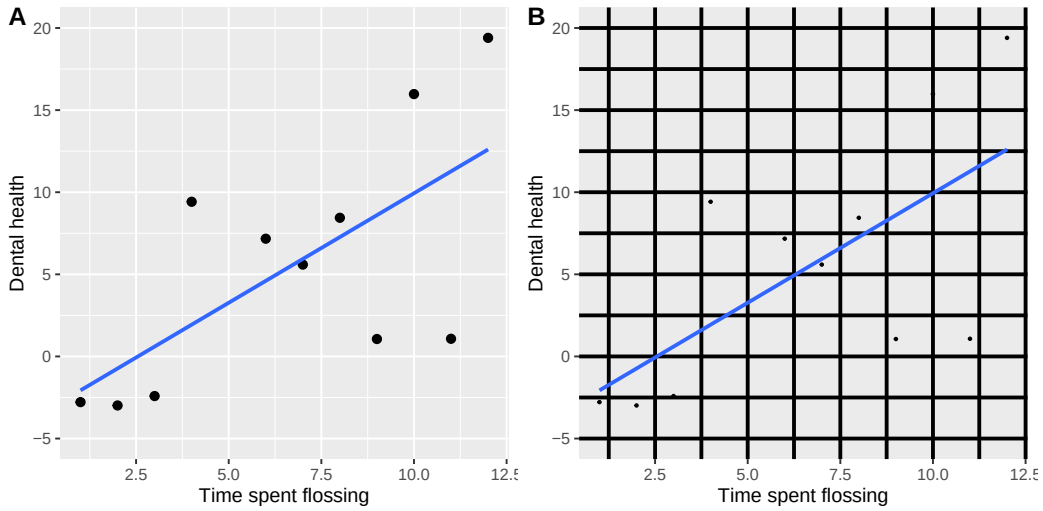

Figure 4.5: An example of the same data plotted with two different data/ink ratios.

4.2.2 Maximize the data/ink ratio

Edward Tufte has proposed an idea called the data/ink ratio:

$$data/ink\ ratio = \frac{amount\ of\ ink\ used\ on\ data}{total\ amount\ of\ ink}$$

The point of this is to minimize visual clutter and let the data show through. For example, take the two presentations of the dental health data in Figure 4.5. Both panels show the same data, but panel A is much easier to apprehend, because of its relatively higher data/ink ratio.

4.2.3 Avoid chartjunk

It's especially common to see presentations of data in the popular media that are adorned with lots of visual elements that are thematically related to the content but unrelated to the actual data. This is known as *chartjunk*, and should be avoided at all costs.

Figure 4.6: An example of chart junk.

One good way to avoid chartjunk is to avoid using popular spreadsheet programs to plot one's data. For example, the chart in Figure 4.6 (created using Microsoft Excel) plots the relative popularity of different religions in the United States. There are at least three things wrong with this figure:

- it has graphics overlaid on each of the bars that have nothing to do with the actual data
- it has a distracting background texture
- it uses three-dimensional bars, which distort the data

4.2.4 Avoid distorting the data

It's often possible to use visualization to distort the message of a dataset. A very common one is use of different axis scaling to either exaggerate or hide a pattern of data. For example, let's say that we are interested in seeing whether rates of violent crime have changed in the US. In Figure 4.7, we can see these data plotted in ways that either make it look like crime has remained constant, or that it has plummeted. The same data can tell two very different stories!

One of the major controversies in statistical data visualization is how to choose the Y axis, and in particular whether it should always include zero.

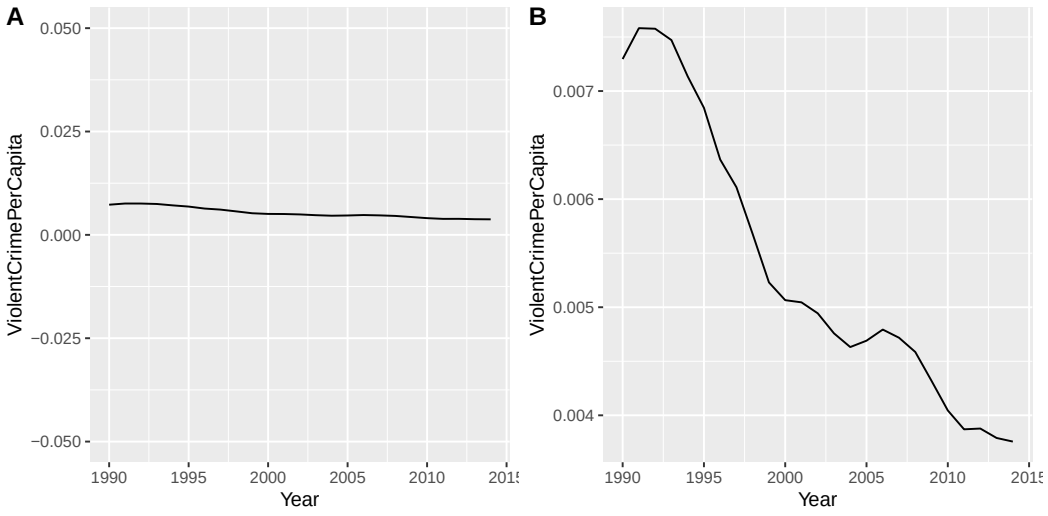

Figure 4.7: Crime data from 1990 to 2014 plotted over time. Panels A and B show the same data, but with different axis ranges. Data obtained from https://www.ucrdatatool.gov/Search/Crime/State/RunCrimeStatebyState.cfm

In his famous book "How to lie with statistics", Darrell Huff argued strongly that one should always include the zero point in the Y axis. On the other hand, Edward Tufte has argued against this:

> "In general, in a time-series, use a baseline that shows the data not the zero point; don't spend a lot of empty vertical space trying to reach down to the zero point at the cost of hiding what is going on in the data line itself." (from https://qz.com/418083/its-ok-not-to-start-your-y-axis-at-zero/)

There are certainly cases where using the zero point makes no sense at all. Let's say that we are interested in plotting body temperature for an individual over time. In Figure 4.8 we plot the same (simulated) data with or without zero in the Y axis. It should be obvious that by plotting these data with zero in the Y axis (Panel A) we are wasting a lot of space in the figure, given that body temperature of a living person could never go to zero! By including zero, we are also making the apparent jump in temperature during days 21-30 much less evident. In general, my inclination for line plots and scatterplots is to use all of the space in the graph, unless the zero point is truly important

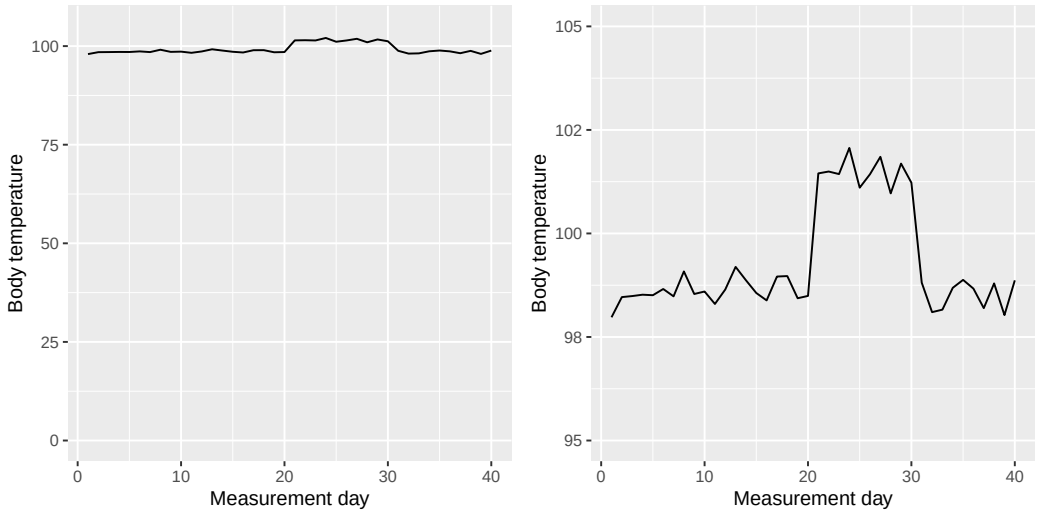

Figure 4.8: Body temperature over time, plotted with or without the zero point in the Y axis.

to highlight.

Edward Tufte introduced the concept of the *lie factor* to describe the degree to which physical differences in a visualization correspond to the magnitude of the differences in the data. If a graphic has a lie factor near 1, then it is appropriately representing the data, whereas lie factors far from one reflect a distortion of the underlying data.

The lie factor supports the argument that one should always include the zero point in a bar chart in many cases. In Figure 4.9 we plot the same data with and without zero in the Y axis. In panel A, the proportional difference in area between the two bars is exactly the same as the proportional difference between the values (i.e. lie factor = 1), whereas in Panel B (where zero is not included) the proportional difference in area between the two bars is roughly 2.8 times bigger than the proportional difference in the values, and thus it visually exaggerates the size of the difference.

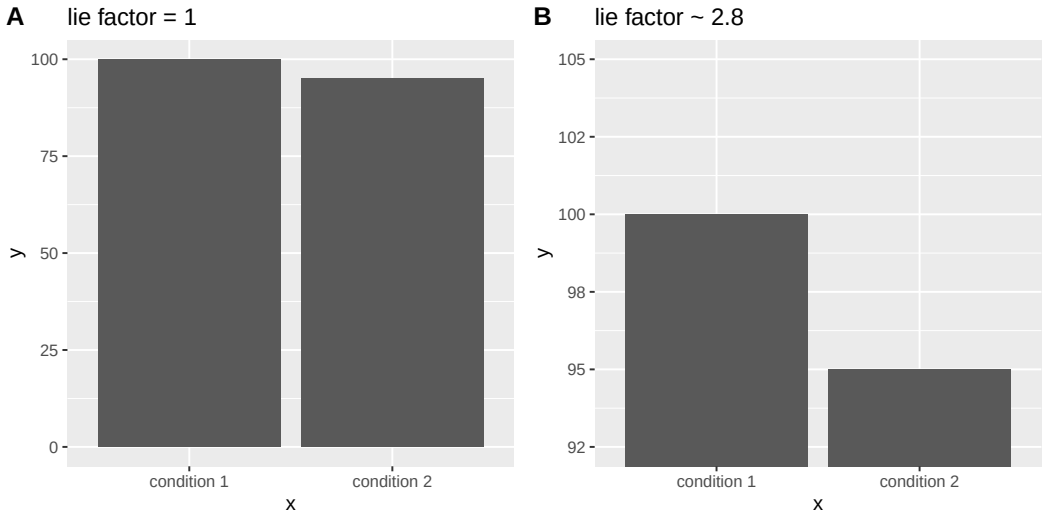

Figure 4.9: Two bar charts with associated lie factors.

4.3 Accommodating human limitations

Humans have both perceptual and cognitive limitations that can make some visualizations very difficult to understand. It's always important to keep these in mind when building a visualization.

4.3.1 Perceptual limitations

One important perceptual limitation that many people (including myself) suffer from is color blindness. This can make it very difficult to perceive the information in a figure (like the one in Figure 4.10) where there is only color contrast between the elements but no brightness contrast. It is always helpful to use graph elements that differ substantially in brightness and/or texture, in addition to color.

Even for people with perfect color vision, there are perceptual limitations that can make some plots ineffective. This is one reason why statisticians *never* use pie charts: It can be very difficult for humans to accurately perceive

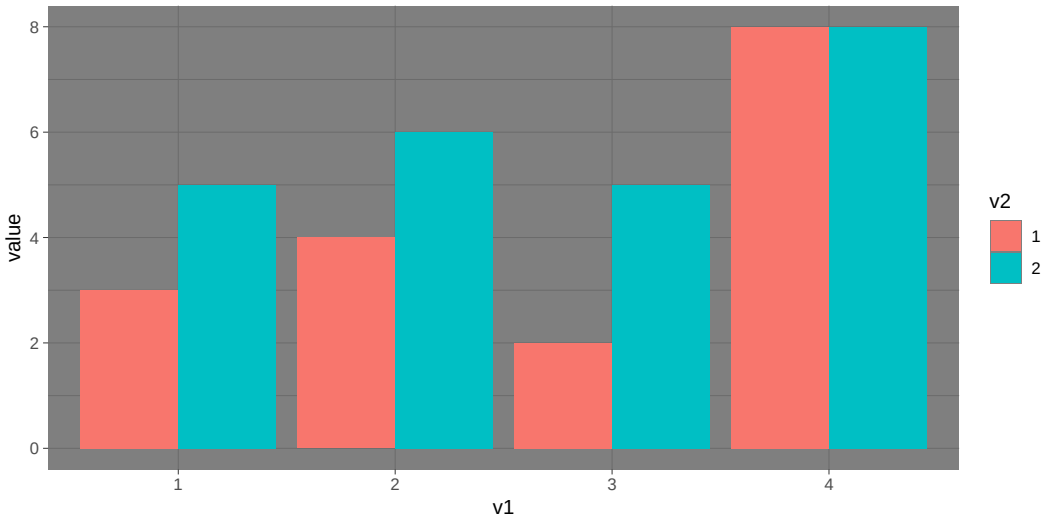

Figure 4.10: Example of a bad figure that relies solely on color contrast.

differences in the volume of shapes. The pie chart in Figure 4.11 (presenting the same data on religious affiliation that we showed above) shows how tricky this can be.

This plot is terrible for several reasons. First, it requires distinguishing a large number of colors from very small patches at the bottom of the figure. Second, the visual perspective distorts the relative numbers, such that the pie wedge for Catholic appears much larger than the pie wedge for None, when in fact the number for None is slightly larger (22.8 vs 20.8 percent), as was evident in Figure 4.6. Third, by separating the legend from the graphic, it requires the viewer to hold information in their working memory in order to map between the graphic and legend and to conduct many "table look-ups" in order to continuously match the legend labels to the visualization. And finally, it uses text that is far too small, making it impossible to read without zooming in.

Plotting the data using a more reasonable approach (Figure 4.12), we can see the pattern much more clearly. This plot may not look as flashy as the pie chart generated using Excel, but it's a much more effective and accurate representation of the data.

This plot allows the viewer to make comparisons based on the the length

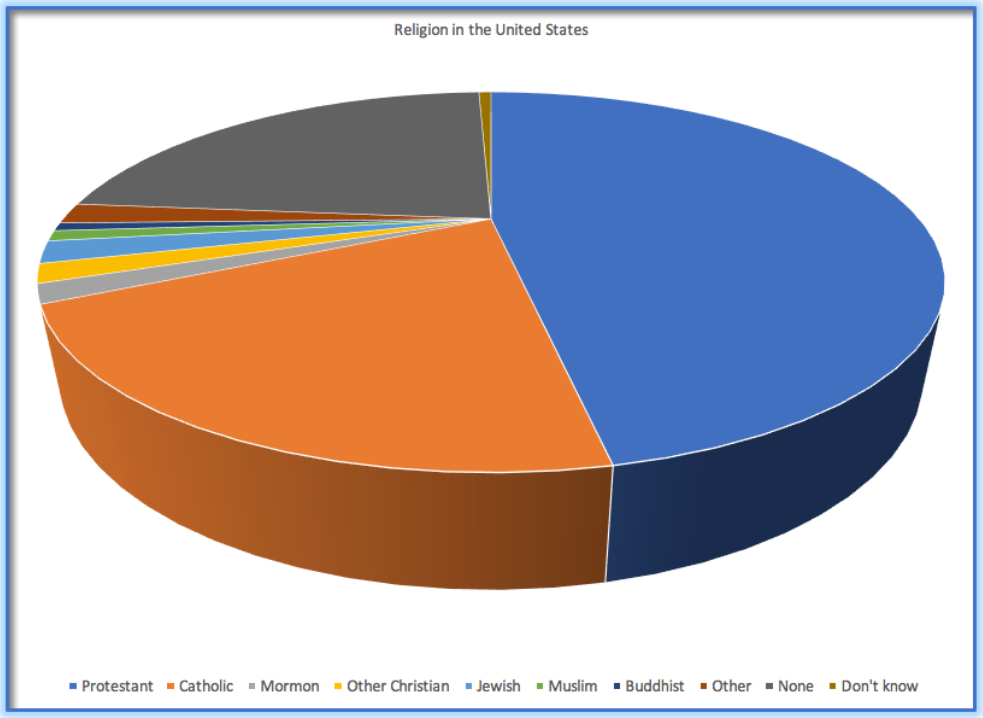

Figure 4.11: An example of a pie chart, highlighting the difficulty in apprehending the relative volume of the different pie slices.

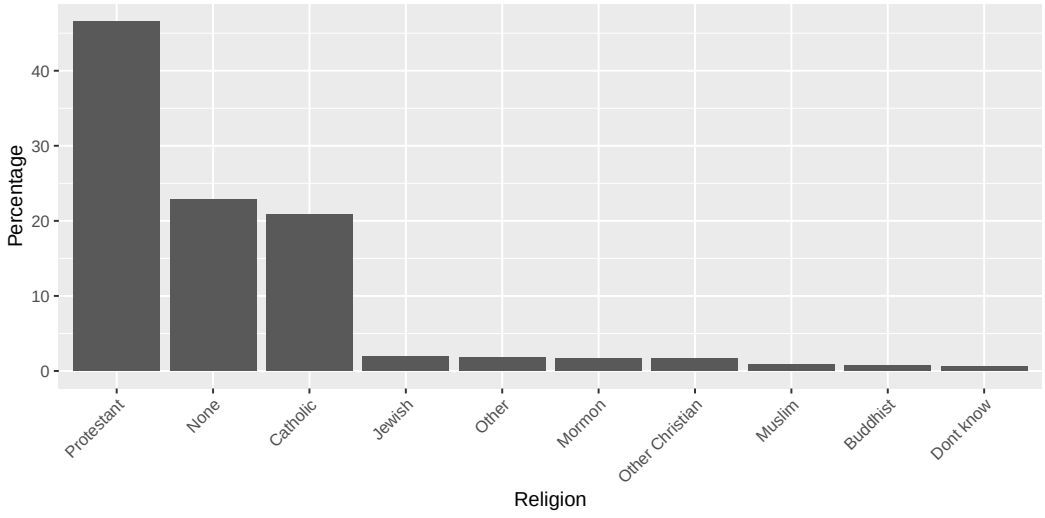

Figure 4.12: A clearer presentation of the religious affiliation data (obtained from http://www.pewforum.org/religious-landscape-study/).

of the bars along a common scale (the y-axis). Humans tend to be more accurate when decoding differences based on these perceptual elements than based on area or color.