Chapter 3

Visualizing Data

One of the first tasks of data analysis should be to look at our data. Certainly, you should open up the dataset and eyeball it to make sure everything looks okay, that the file isn’t mysteriously corrupted, etc., but that is not what I mean when I say “look”. Nor am I talking about data visualization in the sense of the masterpieces of stalwarts such as Alberto Cairo, Mona Chalabi, Andy Kirk, Robert Kossara, Giorgia Lupi, David McCandless, Cole Nussbaumer Knaflic, Randy Olson, Lisa Charlotte Rost, John Schwabish, Nathan Yau, Stephanie Evergreen, Martin Wattenberg, Fernanda Viégas, and many others. We are not going to create visuals that rival these artists’ stunning works. Rather, our goal will be simpler, to use appropriate tables and graphics when analyzing and describing our data, both so that we understand what the data are telling us and can communicate the resulting story to our audience. To accomplish these tasks we will obey simple rules that help us make wise decisions.

3.1 Visualizing Nominal/Ordinal Data

3.1.1 Bar-charts and Frequency Tables

If a variable is measured in nominal or ordinal terms either (i) a bar-chart or (ii) a frequency table are very effective displays of how the variable is distributed: What value seems to be most common? Are high values as likely as low values? You can see both these visuals below, drawn from data about the relative abundance of different species. First, the frequency table.

<table>
<thead>
<tr>
<th>Activity</th>
<th>Frequency</th>
<th>Proportion</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disturbing tiger kill</td>
<td>5</td>
<td>0.06</td>
<td>6</td>
</tr>
<tr>
<td>Fishing</td>
<td>8</td>
<td>0.09</td>
<td>9</td>
</tr>
<tr>
<td>Forest products</td>
<td>11</td>
<td>0.12</td>
<td>12</td>
</tr>
<tr>
<td>Fuelwood/timber</td>
<td>5</td>
<td>0.06</td>
<td>6</td>
</tr>
<tr>
<td>Grass/fodder</td>
<td>44</td>
<td>0.50</td>
<td>50</td>
</tr>
<tr>
<td>Herding</td>
<td>7</td>
<td>0.08</td>
<td>8</td>
</tr>
<tr>
<td>Sleeping in house</td>
<td>3</td>
<td>0.03</td>
<td>3</td>
</tr>
<tr>
<td>Toilet</td>
<td>2</td>
<td>0.02</td>
<td>2</td>
</tr>
<tr>
<td>Walking</td>
<td>3</td>
<td>0.03</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>88</td>
<td>1.00</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 3.1: Factors associated with human-killing tigers in Chitwan National Park, Nepal
Notice how the frequency table is constructed. You have each human activity listed, followed by the frequency (the number of times someone was killed by a tiger while engaging in each activity), then this column with this frequency converted into a proportion, and finally the proportion converted into a percentage. None of this should be a mystery to you:

\[
\text{proportion} = \frac{\text{frequency}}{\text{Total}}
\]

\[
\text{percentage} = \left( \frac{\text{frequency}}{\text{Total}} \right) \times 100 = \text{proportion} \times 100
\]

The last row in the table shows you the total frequency (we have a total of 88 humans killed), the total proportion (which must sum to 1), and the total percentage (which should sum to 100). I love frequency tables such as these because they show all the data and the story; most kills occurred while the individuals were cutting grass for cattle fodder, followed by while they were out gathering other forest products, and least of all while they were going to the toilet\(^1\). We could improve upon this table and bar-chart by organizing it in such a way that the most dangerous activity is listed/plotted first. This would have the added benefit of quickly drawing the reader’s/viewer’s eyes to the most dangerous activity.

\(^1\)Indoor plumbing is lacking in many still developing countries, particularly in the rural areas. In some of these countries, particularly India, tigers aren’t the most dangerous animals.
3.1. Visualizing Nominal/Ordinal Data

Table 3.2: Frequency of Prayer

<table>
<thead>
<tr>
<th>Response Category</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Never</td>
<td>196</td>
</tr>
<tr>
<td>2. Once a week or less</td>
<td>226</td>
</tr>
<tr>
<td>3. A few times a week</td>
<td>264</td>
</tr>
<tr>
<td>4. Once a day</td>
<td>346</td>
</tr>
<tr>
<td>5. Several times a day</td>
<td>456</td>
</tr>
<tr>
<td>Total</td>
<td>1488</td>
</tr>
</tbody>
</table>

Figure 3.2: Frequency of Prayer

What if the variable was an ordinal level variable, say something like an individual’s frequency of praying? We would be facing the same options, a bar-chart or a frequency table. However, we would have to be cautious here in making sure both the table and the bar-chart categories are logically assigned. That is, notice the Response Category; we start with the option “Never” and each category that follows corresponds to a category that reflects praying more often than the preceding category. This order would have to be maintained so that we would be unable to arrange the table or bar-chart in ascending/descending order of the “Frequency” column. If we made that mistake we would be destroying the natural order that exists in the ordinal variable we have before us.

You might be wondering, what about pie charts? Why not use those with nominal/ordinal data? There are two camps, those who hate them and those who think they may be useful. If you are interested, read What do you mean I’m not supposed to use Pie Charts?! but I for one do not use them since I find them less useful than bar-charts.

3.1.2 Contingency Tables and Bar-charts

Frequency tables and bar-charts are also useful when you have two nominal/ordinal variables to work with and are interested in exploring difference between the two or more groups reflected in one variable versus whatever is being measured by the second variable. Let us use a specific example, one where we ask if religiosity differs between Liberals, Moderates, and Conservatives. First the table.

If you look at Table 3.3, you see the frequencies reported in what we call a contingency table or a crosstabulation – where the distributions of two nominal/ordinal variables are jointly displayed. Reading such a table should be simple. In brief, You see how many Liberals said they prayed “Never”, “Once a week...
Table 3.3: Political Ideology and Religiosity

<table>
<thead>
<tr>
<th></th>
<th>Liberal</th>
<th>Moderate</th>
<th>Conservative</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Never</td>
<td>62</td>
<td>46</td>
<td>36</td>
<td>144</td>
</tr>
<tr>
<td>2. Once a week or less</td>
<td>53</td>
<td>63</td>
<td>52</td>
<td>168</td>
</tr>
<tr>
<td>3. A few times a week</td>
<td>53</td>
<td>69</td>
<td>76</td>
<td>198</td>
</tr>
<tr>
<td>4. Once a day</td>
<td>76</td>
<td>89</td>
<td>107</td>
<td>272</td>
</tr>
<tr>
<td>5. Several times a day</td>
<td>69</td>
<td>99</td>
<td>176</td>
<td>344</td>
</tr>
<tr>
<td>Total</td>
<td>313</td>
<td>366</td>
<td>447</td>
<td>1126</td>
</tr>
</tbody>
</table>

Table 3.4: Political Ideology and Religiosity: Row Percentages

<table>
<thead>
<tr>
<th></th>
<th>Liberal</th>
<th>Moderate</th>
<th>Conservative</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Never</td>
<td>43.06</td>
<td>31.94</td>
<td>25.00</td>
</tr>
<tr>
<td>2. Once a week or less</td>
<td>31.55</td>
<td>37.50</td>
<td>30.95</td>
</tr>
<tr>
<td>3. A few times a week</td>
<td>26.77</td>
<td>34.85</td>
<td>38.38</td>
</tr>
<tr>
<td>4. Once a day</td>
<td>27.94</td>
<td>32.72</td>
<td>39.34</td>
</tr>
<tr>
<td>5. Several times a day</td>
<td>20.06</td>
<td>28.78</td>
<td>51.16</td>
</tr>
</tbody>
</table>

or less”, “A few times a week”, “Once a day”, or “Several times a day”. What pattern is evident from this table? Of the 447 conservatives we see most of them saying they pray several times a day, followed by once a day, and the fewest saying they never pray. The pattern is similar for liberals and moderates, although the differences between the numbers responding “Never” and “Several times a day” is small for liberals than it is for conservatives.

The story could be helped a great deal if we calculated percentages for these frequencies. We have two choices when calculating these percentages, we could calculate these as row percentages where we ask “what percent of those who said Never were Liberal, Moderate, and Conservative, respectively?” We could then repeat this for the other categories of religiosity. The result is shown in Table 3.4. This table shows quite clearly that 51.16% of those who say they pray several times a day tend to be Conservative. Likewise, most (43.06%) of those who say they never pray tend to be Liberal.

If we calculated column percentages we would be able to answer such questions as: “What percent of Liberals said they never pray, pray once a week or less, a few times a week, once a day, several times a day?” The same could then be asked of Moderates and Conservatives, respectively. If we used the column percentages shown in Table 3.5 it would be obvious that Moderates and Conservatives tend to be more religious than Liberals. The essential takeaway here is that how you calculate the percentages (row versus column) depends upon what story you want to highlight, the question you want to ask and answer.

Graphing these tables is easily done as well, and very effective; see Figure 3.3.

Table 3.5: Political Ideology and Religiosity: Column Percentages

<table>
<thead>
<tr>
<th></th>
<th>Liberal</th>
<th>Moderate</th>
<th>Conservative</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Never</td>
<td>19.81</td>
<td>12.57</td>
<td>8.05</td>
</tr>
<tr>
<td>2. Once a week or less</td>
<td>16.93</td>
<td>17.21</td>
<td>11.63</td>
</tr>
<tr>
<td>3. A few times a week</td>
<td>16.93</td>
<td>18.85</td>
<td>17.00</td>
</tr>
<tr>
<td>4. Once a day</td>
<td>24.28</td>
<td>24.32</td>
<td>23.94</td>
</tr>
<tr>
<td>5. Several times a day</td>
<td>22.04</td>
<td>27.05</td>
<td>39.37</td>
</tr>
</tbody>
</table>
3.2 Visualizing Interval/Ratio Data

We have several visuals we could draw with numeric data that are either interval or ratio levels of measurement. Let us see these first before we look at the one frequency table that could be used with interval/ratio data.

3.2.1 The Histogram

A histogram is used with a single numeric variable and looks like a bar-chart except there are no gaps between consecutive bars unless there are missing data. The example that follows use a popular dataset known as hsb2, which contains information about 200 randomly selected students from a national survey of high school seniors called the High School and Beyond survey. The variables in this dataset include:

(1) id = student id
(2) female = (0/1)
(3) race = ethnicity (1=hispanic 2=asian 3=african-amer 4=white)
(4) ses = (1=low 2=middle 3=high)
(5) schtyp = type of school (1=public 2=private)
(6) prog = type of program (1=general 2=academic 3=vocational)
(7) read = standardized reading score
(8) write = standardized writing score
(9) math = standardized math score
(10) science = standardized science score
(11) socst = standardized social studies score

I’ll use read (Reading scores on a standardized test) to plot a histogram. Notice a few things about the histogram in Figure 3.4. The height of the bars, representing how often a particular score occurs, varies a
great deal. A few students have done poorly while a few have done very well, but the rest are distributed over the middle range of test scores.

We could farther break this histogram apart by asking if male and female students (female), private versus public students (schtyp), or students of different races/ethnicities (race) perform differently on the reading test.

These plots are hard to read for a number of reasons. First, with just \( n = 200 \) students in all we have more students in some groups than in others and as a result the histograms look very thin for the groups with fewer data points, making it difficult to tease out any patterns. Second, within any group there are too many different test scores so we don’t see a clear pattern at all. To fix these problems we construct groups of scores, turning what is a numeric variable into an ordinal variable. Let us build this by first creating a grouped frequency table and then plotting this table as a histogram.

### 3.2.2 Grouped Frequency Tables

Because numeric variables take on too many values it is often easier to see their distribution by grouping the numeric values. For example, we could start by seeing what are the lowest and highest reading scores. These turn out to be 28 and 76, respectively. We can build the groups as follows:

1. Calculate difference between maximum and minimum values, which turns out to be \( 76 - 28 = 48 \)
2. Decide how many groups we want. Good practice suggests no fewer than 4/5 and no more than 6/7. Say we go with 5 groups.
3. Divide the gap between the maximum and minimum values by the desired number of groups: \( \frac{48}{5} = 9.6 \) and round up to the nearest whole number = 10. This tells us how wide each group should be.
4. The groups could thus be \( 28 - 38, 38 - 48, 48 - 48, 58 - 68, 68 - 78 \).

Notice that these groups span, start to finish, all the values of reading scores. But we’ll have to decide which group to include 38, 48, 58, and 68 in. Should 38 go in 28-28 or in 38-48? The choice doesn’t matter so long as we are consistent. Let us choose a rule that says include 38 in 38-48, 48 in 48-58, 58 in 58-68, and 68 in 68-78. Using this rule we now find each reading score and drop it into its group. Then we calculate how many scores fall in each group, creating our Frequency column.

Now it is easy to see that the largest number of students (68) appear to fall in the 38-48 group of reading
Figure 3.5: Histogram of Reading Scores by Various Groups
Table 3.6: Grouped Frequency Table of Reading Scores

<table>
<thead>
<tr>
<th>Grouped Scores</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>28-38</td>
<td>14</td>
</tr>
<tr>
<td>38-48</td>
<td>68</td>
</tr>
<tr>
<td>48-58</td>
<td>62</td>
</tr>
<tr>
<td>58-68</td>
<td>36</td>
</tr>
<tr>
<td>68-78</td>
<td>20</td>
</tr>
<tr>
<td>Total</td>
<td>200</td>
</tr>
</tbody>
</table>

Table 3.7: Grouped Frequency Table of Reading Scores with Percentages

<table>
<thead>
<tr>
<th>Grouped Scores</th>
<th>Frequency</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>28-38</td>
<td>14</td>
<td>7</td>
</tr>
<tr>
<td>38-48</td>
<td>68</td>
<td>34</td>
</tr>
<tr>
<td>48-58</td>
<td>62</td>
<td>31</td>
</tr>
<tr>
<td>58-68</td>
<td>36</td>
<td>18</td>
</tr>
<tr>
<td>68-78</td>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td>Total</td>
<td>200</td>
<td>100</td>
</tr>
</tbody>
</table>

scores while the smallest frequency (14) occurs in the 68-78 group. Let us add to this table a percentage column.

The percentages make it even easier to see how the distribution breaks down; 34 percent of the students have scores in the 38-48 range while only 10 percent scored in the highest bracket (68-78). We can also break this down by the three groups we used earlier.

Histograms were once quite popular but there is a better visual for looking at our numeric variables, one called the box-plot that we’ll see in the next chapter. I am not a huge fan of histograms because grouping decisions influence the story being told. My advice would be to use grouped frequency tables instead of histograms to present a summary overview of your numeric data.
Figure 3.6: Histogram of Grouped Reading Scores by Specific Groups
3.2.3 Scatterplots

With two numeric variables, a scatterplot comes in very handy if we want to explore how one variable might be related to another. For example, we may want to ask whether students who score high on the reading test also tend to score high on the mathematics test. This could be visually explored via a scatterplot, as shown below. The goal should be to look for a pattern: Does one variable increase as the other increases or does one variable decrease as the other increases? Or does there seem to be no relationship at all?

Quite clearly, students who do well in Reading also tend to do well in Mathematics. You can see this by virtue of the upward, left-to-right tilt of the cloud of points.

What about Science scores and Mathematics scores? Is there any relationship between doing well in Science and doing well in Mathematics?

Of course, just like everything else we could break this down by any nominal/ordinal variable. The visual below shows you the breakouts by the student’s sex.
Figure 3.9: Scatterplot of Science Scores and Mathematics Scores, by Sex
Chapter 3. Visualizing Data

3.2.4 Line Graphs

If we have time-series data, such as the Presidential Approval data we saw earlier, then a line graph works well because it shows you how the outcome/phenomenon varies over time. Let us take another example, stock prices. Say I have invested in both IBM and LinkedIn stocks. How are these stocks doing?

These plots don’t work well just for financial or election data, they are ideally suited for any phenomenon that is measured over time. For example, the size of the immigrant population over the years, and even the number of lynx pelts reported in Canada per year from 1752 to 1819.

3.2.5 Polar Charts

These charts are helpful for visualizing data that might have otherwise been explored via a bar-chart. For example, say you want to look at the miles per gallon given by a number of different automobiles. We could use a bar-chart, as shown below:
3.2. Visualizing Interval/Ratio Data

Figure 3.12: Number of lynx pelts reported in Canada per year from 1752 to 1819.

Figure 3.13: Miles per Gallon
Quite obviously, the Toyota Corolla has the best fuel economy, followed by the Fiat 128, while the Cadillac Fleetwood is tied with the Lincoln Continental for the worst fuel economy. A polar chart would present the same information as seen in the bar-chart, albeit in a more aesthetically pleasing manner.

3.3 Some Essential Rules for Good Visualizations

There are several rules, some favored by this expert or that, but regardless of the source of the rule, its merit is not in doubt. So here are the ones I try to follow (although I do break these at times, some times by design and other times unintentionally). It all starts of course with Edward Tufte’s maxims: show the data, tell the truth, help the viewer think about the information rather than the design, encourage the eye to compare the data, make large data sets coherent.

A more specific listing of the rules I try to keep in mind follows:

1. Do not include anything in your visualization that is not very informative. Give titles and subtitles where needed and make these stand on their own. At the same time, do not clutter your charts and tables with too much information otherwise the reader gets lost.
Use colors wisely. Is this visual to be printed on a color printer? Is it part of a presentation in a poorly lit room? Choose bright colors that stand out from each other and yet are visible on a printed page or on the projection screen in the room. Remember, a fair share of men seeing the visual could be color-blind so use color palettes designed for color-blind individuals. If this is new to you, read *How a dog sees a rainbow,* and 12 other images that explain how we see color. Pay attention to what is being plotted: Use **sequential** colors for ordered data values (that go from low to high or vice-versa); **qualitative** colors nominal data, and; **diverging** colors if the goal is to put equal emphasis on mid-range critical values and extremes at both ends of the data range but emphasize the middle with light colors and the low/high extremes with dark colors that have contrasting hues.

The visualization is not your personal Mona Lisa; it must be effective for the target audience and help you tell the story. Do not get carried away in creating it.

Use a table if a table would be more effective than a graph but remember, while tables are useful for audiences that will want to see more (not less of the actual data you have) graphs are favored by those who want a quick take-away.

Always have your y-axis (or x-axis if relevant) starting at zero. If you don’t do this you can misrepresent the data. If you are forced to truncate the axis, point this out to the reader so that they can interpret dips and swells in line charts or differences between heights of frequency bars with care.

Combine multiple graphs/tables into a single figure if you can so long as it does not lead to information overload. Go back and look at the scatterplots and line charts and try to visualize what these might look like if the y-axis had been forced to start at 0; they would certainly look different. For illustration purposes I have let the plotting software pick the starting coordinates.

For non-technical audiences you should round up all percentages/proportions to no more than one decimal place and ideally to no decimal place unless doing so distorts the picture. For technical audiences, stay with two or four decimal places.

Above all, start with pencil and paper and consider all alternate visualizations possible by drawing a rough sketch of what the finished product should look like.

Consider using a choropleth map if geographic variation is one of the key narratives.

Be prepared to alter your visual if it does not work; we are all hesitant to delete a page or a graphic and start from scratch but this locks you into something that isn’t working to begin with and you will be stuck in a rut.

Avoid jargon like the plague. We often think we sound intelligent when we use jargon but this distracts from the oral/written presentation. Think of your audience and write and present in a manner that will resonate with them.

Show as much data as you can; it vastly improves the effectiveness of the visual narration.

### 3.4 Practice Problems

**Problem 1**

Download the monthly Lake Erie levels dataset from here and construct an appropriate chart to display the data. Be sure to label the x- and y-axis, and to title the chart.

**Problem 2**

Download the number of births per 10,000 of 23 year old women, U.S., 1917-1975 from here and construct an appropriate chart to display the trend in the data. Be sure to label the x- and y-axis, and to title the chart.

**Problem 3**

Download the winning times for the men’s 400 m final in each Olympic Games from 1896 to 1996 from here and construct an appropriate chart to display these data. Be sure to label the x- and y-axis, and to title the chart.
Problem 4

Use the data-set from Practice Problem 4 in the preceding Chapter. Construct a frequency table for belief in life after death, showing both the frequencies and the relative frequencies (as percentages). Based on the table, what do most people seem to believe? Report the percentage for your answer. Construct an appropriate chart for these data, making sure to label all axis and providing a title.

Problem 5

Construct an appropriate chart that shows the relationship between high school GPA and college GPA. Label both axis and title the chart. What does this plot show? Are the two positively/negatively related? How strong would you guess is the relationship?

Problem 6

Construct a contingency table of vegetarianism against belief in life after death. What percent of vegetarians believe in life after death? What percent of those who believe in life after death are vegetarians? Use an appropriate chart to show the relationship between vegetarianism and belief in life after death. Label everything.

Problem 7

Construct a grouped frequency table with five groups of the variable age. Report both the frequencies and the relative frequencies (as percentages). Plot the relative frequencies using an appropriate chart. Label everything as usual. What is the modal age group?

The next set of questions revolve around the 2016 Boston Marathon race results available here. The dataset contains the following variables:

- Bib = the runner’s bib number
- Name = the runner’s name
- Age = the runner’s age (in years)
- M/F = the runner’s gender (M = Male; F = Female)
- City = the runner’s home city
- State = the runner’s home state
- Country = the runner’s home country
- finishtime = the runner’s time (in seconds) to the finish line

Problem 8

What countries had the largest and second-largest number of runners in the race?

Problem 9

What was the distribution of the runners’ gender? Use a suitable chart to reflect the distribution.

Problem 10

Construct a grouped frequency table of the runners’ age, using the following groupings – 18-25; 25-32; 32-39; 39-46; 46-53; 53-60; 60-67; 67-74; 74-81; 81-88. Also construct a grouped histogram. What is the modal age group?

Problem 11

Draw a scatter-plot of runners’ age and finish times. Does this show any relationship? If it does, what sort of a relationship?

Problem 12

Using the grouping structure listed in Problem 11, construct country-specific histograms of finish times for runners from each the following countries – AUS, BRA, CAN, CHN, FRA, GBR, GER, ITA, JPN, MEX,
3.4. Practice Problems

and USA. Are finish times skewed for each country? Is the direction of the skew similar? What country seems to have the least skew? What country seems to have the most skew?

Use the data-set from Practice Problem 5 in the preceding Chapter to answer the following questions.

Problem 13
Were employees who had a workplace accident more likely to leave than employees who did not have an accident? Briefly explain your conclusion with appropriate charts/tables.

Problem 14
Are low salary employees more likely to leave than medium/high salary employees? Why do you conclude as you do? Briefly explain your reasoning with appropriate charts/tables.

Problem 15
Construct an appropriate chart that shows the distribution of the number of years employees spent in the company. What patterns do you see in this chart?

Download the 2017 County Health Rankings data and the accompanying codebook. Filter the downloaded data to cases where chrreleaseyear == 2017 and countycode == 0. The result will be state-level data in the 2017 database. Note: This will include the national estimates (evident from rows where statecode == 0) so be sure to also discard the national estimates. You will be using the resulting data for the problems that follow.

Problem 16
You should see the following measures (measurename)

(i) Adult obesity
(ii) Children in poverty
(iii) Diabetes monitoring
(iv) High school graduation
(v) Premature death
(vi) Preventable hospital stays
(vii) Unemployment rate

Looking at pairs of variables in turn, briefly (20 words or less) state whether you would expect a positive/negative/no relationship between the variables in each pair and why?

Problem 17
Construct scatterplots of (i) High school graduation and Adult obesity (ii) Adult obesity and Diabetes monitoring (iii) Adult obesity and Premature death (iv) Adult obesity and Preventable hospital stays (v) premature deaths and Preventable hospital stays (vi) Adult obesity and Unemployment rate (vii) Children in poverty and High school graduation (viii) Children in poverty and Adult obesity

Problem 18
How do the scatterplots you constructed in the preceding problem stack up against your expectations you listed in Problem 16? Which plots surprised you and why? Be brief.