Practical Multivariate Analysis
Sixth Edition

Abdelmonem Afifi, Susanne May, Robin Donatello, Virginia A. Clark

Contents

Part I

Preparation for Analysis

Chapter 4

Data Visualization

4.1 Introduction

Visualizing data is one of the most important things we can do to become familiar with the data. There are often features and patterns in the data that cannot be uncovered with summary statistics alone. As the old adage goes: "a picture is worth a thousand words." Good data visualizations often convey much more information than a block of text or a table full of numbers. This chapter introduces a series of plot types for both categorical and continuous data. We start with visualizations for a single variable only (univariate), then combinations of two variables (bivariate), and lastly a few examples and discussion of methods for visualizing relationships between more than two variables (multivariate). Additional graphs designed for a specific analysis setting are introduced as needed in other chapters of this book.

This chapter uses several data sets described in Appendix A. Specifically, we use the parental HIV and the depression data sets to demonstrate different visualization techniques.

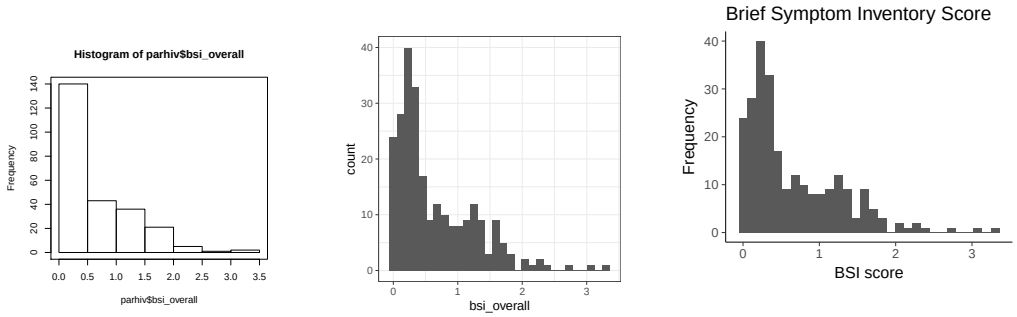
All graphics in this chapter are made using R, with section 4.5 containing a discussion of graphical capabilities to create these graphs in other statistical software programs.

There are three levels of visualizations that can be created, with examples shown in Figure 4.1a, b and c.

- **For your eyes only (4.1a):** Made by the analyst, for the analyst, these plots are quick and easy to create, using the default options without any annotation or context. These graphs are meant to be looked at once or twice for exploratory analysis in order to better understand the data.
- **For an internal report (4.1b):** Some chosen plots are then cleaned up to be shared with others, for example in a weekly team meeting or to be sent to co-investigators participating in the study. These plots need to be capable of standing on their own, but can be slightly less than perfect. Axis labels, titles, colors, annotations and other captions are provided as needed to put the graph in context.
- **For publication or external report (4.1c):** These are meant to be shared with other stakeholders such as the public, your collaborator(s) or administration. Very few plots make it this far. These plots should have all the "bells and whistles" as they appear in formal reports, and are often saved to an external file of a specific size or file type, with high resolution. For publication in most printed journals and books, figures typically need to be in black and white (possibly grayscale).

4.2 Univariate Data

This section covers how to visualize a single variable or characteristic. We start with plots for categorical data, then cover plots for continuous data. Visualization is one of the best methods to identify univariate outliers, skewness, low frequencies in certain categories and/or other oddities in the distribution of the data.



(a) *For your eyes only* **(b)** *For an internal report* **(c)** *For publication*

Figure 4.1: *Three levels of graphic quality and completeness*

4.2.1 Categorical Data

Categorical (nominal or ordinal) data are summarized by reporting the count, or frequency, of records in the data set that take on the value for each category of the variable of interest (See Section **??** for a review of data type classifications.) Common visualizations for the counts of categorical data include tables, dot plots, and pie charts. The subsections below discuss and demonstrate each of these types.

Tables

A **table** is the most common way to organize and display summary statistics of a categorical variable using just numbers. Typically the table shows both the frequency (N) and the percent for each category.

Table 4.1: *Education level among mothers with HIV*

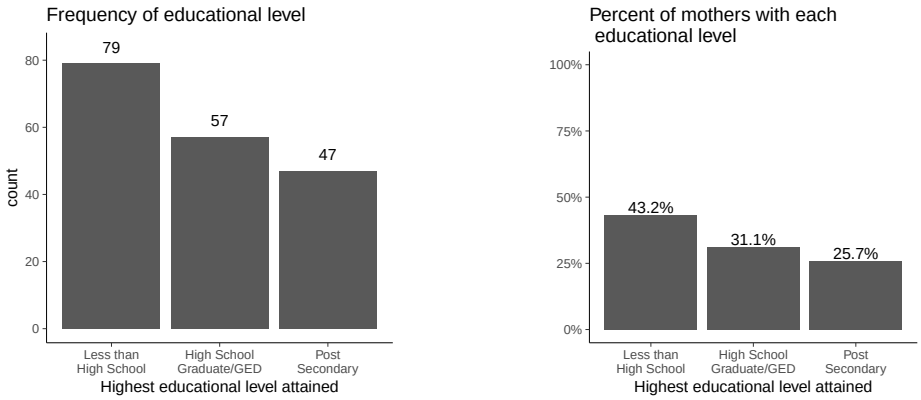
Education Level	N	Percent
Less than High School	79	43.2%
High School Graduate/GED	57	31.1%
Post Secondary	47	25.7%

Table 4.1 shows that about a quarter (47, 25.7%) of mothers in the parental HIV data set have post-secondary school education level.

Bar Charts

A **bar chart** (Figure 4.2) takes these frequencies and draws bars for each category (shown along the horizontal axis) where the heights of the bars are determined by the frequencies seen in the table (Figure 4.2a). A reasonable modification is to put the percentages on the vertical axis (Figure 4.2b). This is a place to be cautious however. Some programs by default will exclude the missing data before calculating percentages, so the percentages shown are for available data. Other programs will display a bar for the missing category and display the percentages out of the full data set. For either choice it is advised that the analyst understand what the denominator is.

The ordering of categories is important for readability. Nearly all statistical software packages will set the automatic factor ordering to alphabetical, or according to the numerical value that is assigned to each category. If the data are ordinal, tables and plots should read left to right along with that ordering, such as the educational level example in Figure 4.2. Sometimes there is a partial ordering such as years of high school education and then different degrees that are not necessarily



(a) *A bar chart comparing frequencies* **(b)** *A bar chart comparing percentages*

Figure 4.2: *Two bar charts showing the distribution of highest level of education attained*

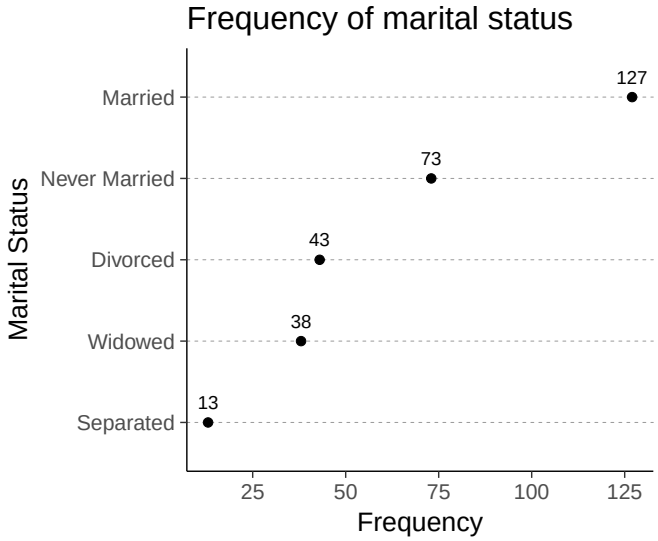


Figure 4.3: *A Cleveland dot plot of marital status*

easy to summarize in years (can one year of vocational school be considered equivalent to one year of college or one year of community college?). In these situations it is left to the researcher to decide on how to define and justify the order of the categories using subject matter expertise.

Cleveland Dot Plot

Bars use a lot of ink, and the width of the bar is typically meaningless. **Cleveland dot plots** (Cleveland, 1993) provide an alternative method to display the frequencies using less ink and space than bar charts (Figure 4.3). A dot plot is especially helpful when there are a large number of categories to visualize. We use marital status as an example. Because it is a nominal variable (in contrast to the previous ordinal variable examples), these summary data are best displayed in descending order of frequency.

The standard dot plot depicts the value of each record as a separate dot. Cleveland dot plots differ from the standard dot plot because they plot summary data, not raw data. After summarizing

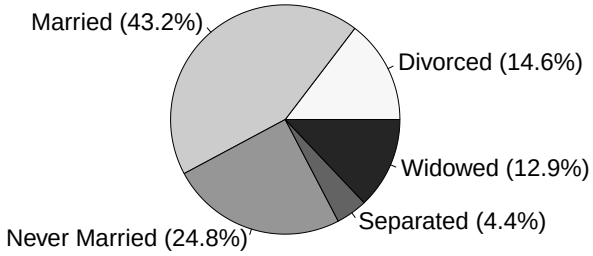


Figure 4.4: *A fully labeled pie chart of marital status. Note that percentages may not always add up to 100% due to rounding*

the data, such as calculating the frequency of records per category, we now only have one data point per category to plot.

Pie Charts

Each wedge of a **pie chart** (Figure 4.4) contains an internal angle indicating the relative proportion of records in that category. However, human eyes cannot distinguish between angles that are close in size as well as they can distinguish between heights of bars (or lines or dots). A necessary component to make any pie chart interpretable is having labels with names and percentages for each wedge.

4.2.2 Continuous Data

Continuous data by definition can take on infinite possible values, so the above plots that display frequencies of records within a finite number of categories do not apply here unless the continuous data are categorized into distinct groups (e.g. income brackets). To visualize continuous data, we need to display the actual value or the distribution of the data points directly. Common plot types include: stem-leaf plots, histograms, density graphs, boxplots, violin plots, and qqplots. So, what do these plots depict and how are they generated?

Stem-leaf plots

The **stem-leaf plot** (Tukey, 1972) demonstrates how numbers placed on a line can describe the shape of the distribution of observed values and provide a listing of all individual observations in the same plot. Since this type of graphic includes each individual data point, the usefulness and readability diminishes as the number of data points increases. Figure 4.5 displays the values of age for individuals in the depression data set.

Because stem-leaf plots display the value of every observation in the data set, the data values can be read directly. The first row displays data from 15 to 19 years of age, or, the second half of the 10s place. Note that this study enrolled only adults, so the youngest possible age is 18. There are five 18 year olds and five 19 year olds in the data set. From this plot one can get an idea of how the data are distributed and know the actual values (of ages in this example). The second row displays data on ages between 20 and 24, or, the first half of the 20s. The third row displays data on ages between 25 and 29, or, the second half of the 20s, and so forth.

Often we are not interested in the individual values of each data point, rather we want to examine the distribution of the data or summary measures of the distribution. Example questions might be:

```
1 | 8888899999
2 | 00000011111122222222233333333333444444444
2 | 55555566666666677778888889999
3 | 000000111112222222222333333444444444
3 | 555566666677777889
4 | 000001222222222333333344
4 | 555566677777788889999
5 | 00000111111222233444
5 | 55556667777778888888999999999
6 | 000000011111222233444
6 | 555556667788889
7 | 000001112233444
7 | 5778899
8 | 011233333
8 | 9
```

Figure 4.5: *A stem-leaf plot of the individual age in the depression data set*

where is the majority of the data? Does the distribution look symmetric around some central point? Around what values do the bulk of the data lie? For example, the distribution of ages of individuals in the depression data set ranges from 18 to 89, is slightly skewed with a right tail, unimodal, with a mean around 45.

Histograms

Rather than showing the value of each observation, we often prefer to think of the value as belonging to a *bin*, or an interval. The heights of the bars in a **histogram** display the frequencies of values that fall into those bins. For example if we grouped the ages of individuals in the depression data set into 10 year age bins, the frequency table looks like this:

(15,25]	(25,35]	(35,45]	(45,55]	(55,65]	(65,75]	(75,85]	(85,95]
57	61	42	41	51	26	15	1

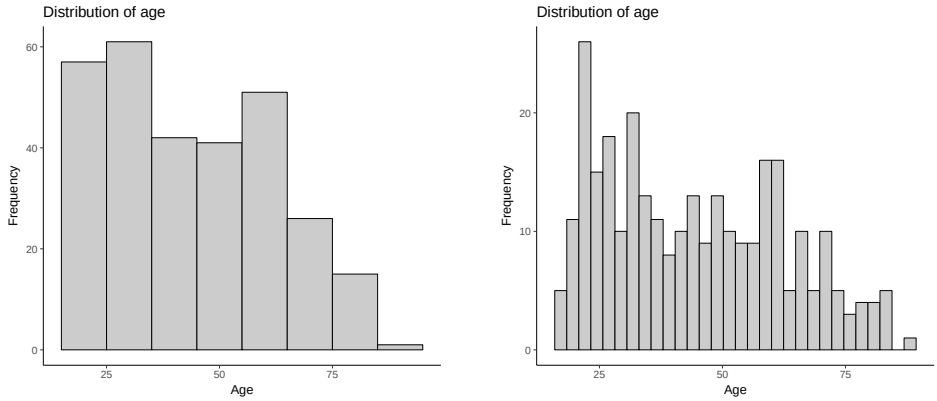
In this table, the notation (15,25] indicates that this bin includes values that are between 15 and 25, including 25 but excluding 15, and so forth.

To create a histogram, the values of a continuous variable are plotted on the horizontal axis, with the height of the bar for each bin equal to the frequency of data within that bin (Figure 4.6a). The main difference between a histogram and a bar chart is that bar charts plot a categorical variable on the horizontal axis, so the vertical bars are separated. The horizontal axis of a histogram is continuous, so the bars touch each other, and thus there is no gap between bins because the bins represent directly adjacent categories.

The choice of the size of each bin can highlight or hide some features in the data. If there is no scientifically motivated or otherwise prespecified bin size, we might start with the default value for the chosen statistical software package, as we did in Figure 4.6b, and then adjust as necessary. Figure 4.6b displays the same data on ages in the depression data set using the default value of range/30 chosen by the ggplot2 package in R. For this example the range of ages is 18 to 89, thus the binwidth is $\frac{(89-18)}{30} = 2.4$. This choice of bin width shows the most frequent age at around 23, unlike Figure 4.6a where it appears to be over 25.

Kernel density plots

Instead of plotting bars for each bin, we can sometimes get a better (or different) idea of the true shape of the distribution by creating a **kernel density plot**. The kernel density is a function, $f(x)$, that is generated from the data set, similar to a histogram. Density plots differ from histograms in that this function is a smooth continuous function, not a stepwise discrete function that creates bars with flat tops.



(a) *A histogram with 8 bins* **(b)** *Using bins of width range/30*

Figure 4.6: *Histograms displaying the distribution of age in the depression data set*

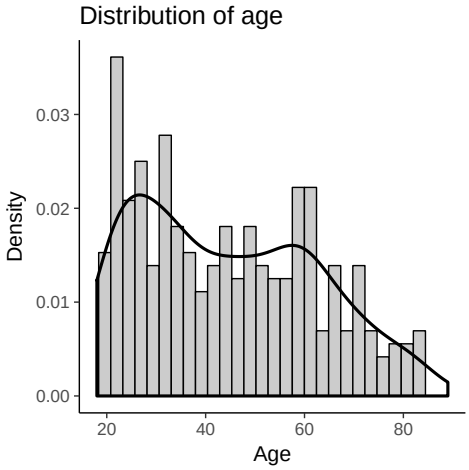


Figure 4.7: *Density plot for the distribution of age in the depression data set*

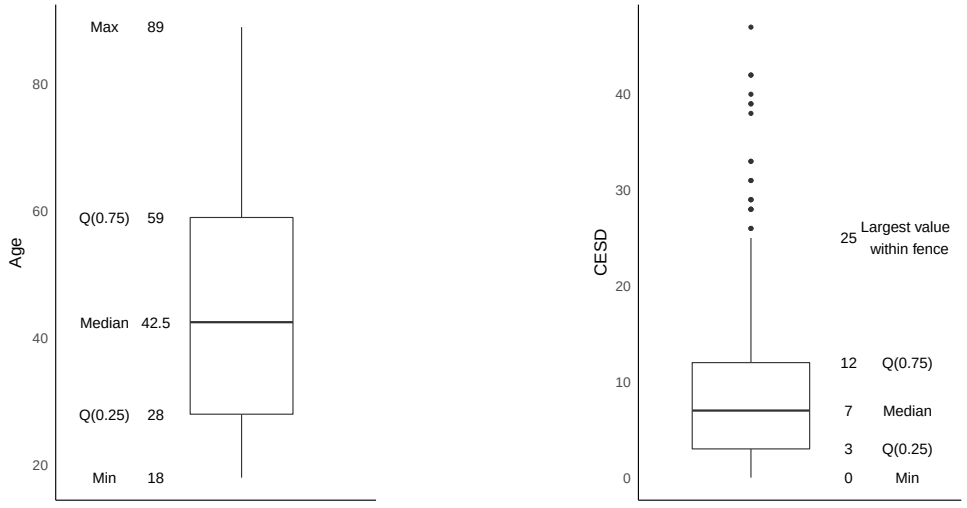
Figure 4.7 shows that the density line smooths out the multitude of peaks and valleys in the histogram, providing a better idea of the general shape or trend of the data.

Notice that the vertical axis on a **density plot** is no longer the frequency or count, but the value of the kernel density, which can be thought of as the relative frequency. While the density curve in Figure 4.7 is overlaid on top of the histogram with the smaller bin width, the first peak of the density plot is around 25, which is more representative of Figure 4.6a which uses the wider bin size. This highlights the importance of looking at multiple types of graphics to fully understand the distribution of the data.

Boxplots and Violin plots

A **boxplot** (also called **box-whisker** plots) display the five number summary (Min, $Q(0.25)$, Median, $Q(0.75)$, Max) in graphical format, where $Q(0.25)$ indicates that 25% of the data are equal to or below this value. The data are split into equal sized quarters, i.e., same number of data points are in each of the four sections of a boxplot (Figure 4.8a).

The box outlines the middle 50%, or the interquartile-range ($IQR = Q(0.75) - Q(0.25)$) of the data, and the horizontal lines (whiskers) extend from the 1st quartile ($Q(0.25)$) down to the



(a) *An unadorned boxplot of age* **(b)** *A modified boxplot of CESD*

Figure 4.8: *An unadorned, and a modified boxplot*

minimum value, and upwards from the third quartile $Q(0.75)$ to the maximum value. This means that in Figure 4.8a, the same number of individuals in the depression data set are between the 10 year span of ages between 18 and 28, as there are between the 30 year span of ages between 59 and 89.

Some statistical packages plot the **modified boxplot** by default. We first define the **fences**, i.e., $Q(0.25) - 1.5 * IQR$ and $Q(0.75) + 1.5 * IQR$. Then, in the modified boxplot, the whiskers do not extend all the way out to the maximum and minimum, but out to the data points that are just inside the fences as calculated by the $1.5 * IQR$ rule. In the modified boxplots, outliers are typically denoted as points or dots outside the fences. For example, consider the continuous measure of depression, variable CESD (Center for Epidemiological Studies Depression) in Figure 4.8b. Here the upper whisker extends to 25, the maximum value inside $1.5 * IQR$. The points above 25 are considered potential outliers.

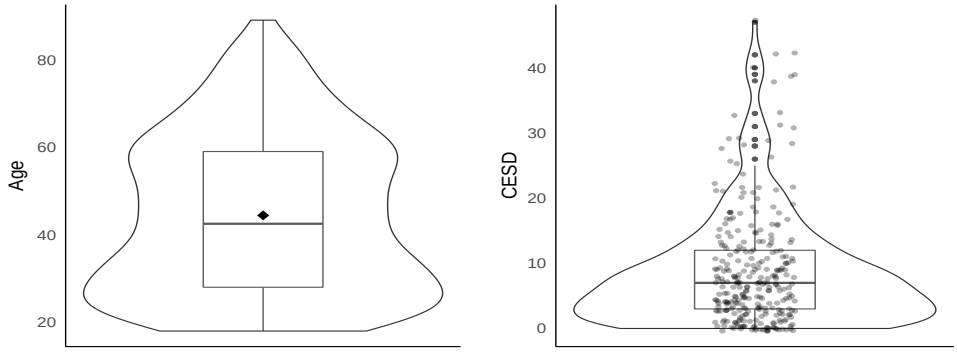
Additions that make boxplots much more informative are displayed in Figure 4.9: 1) adding the mean as a point, 2) adding a **violin** plot to show the density (reflected around the mid-line of the boxplot), and 3) adding the points directly, but jittered (where equal values are moved slightly apart from each other) to avoid plotting symbols on top of each other and thus making them difficult or impossible to identify. Violin plots are not commonly used, but they can be very informative in that they can display the shape of the kernel density in the same graph as the boxplot.

4.3 Bivariate Data

Next we introduce graphical methods to explore relationships between two variables. Many of the same plotting types such as boxplots and histograms introduced for univariate exploration will be used again here.

4.3.1 Categorical versus Categorical Data

To compare the distribution of one categorical variable across levels of another categorical variable, primarily tables are created. Tables come under several names including **cross-tabulation**, **contingency** tables and **two-way** tables. Table 4.2 displays the frequency of gender by education level for



(a) *Add a violin plot & the mean* **(b)** *Add a violin plot & jittered points*

Figure 4.9: Boxplot enhancements

individuals in the depression data set. The value in each cell is the number of records in the data set with that combination of factor levels. For example there are 4 males with less than a high-school (HS) degree, 26 females who have completed a Bachelor's (BS) degree, and 8 males who have completed a Masters (MS) degree.

Table 4.2: *Two-way frequency table of gender by educational level*

	<HS	Some HS	HS Grad	Some college	BS	MS	PhD
Male	4	19	39	18	17	8	6
Female	1	42	75	30	26	6	3

When group sizes are not comparable, it is more informative to compare percents instead of frequencies. There are three types of percentages that can be calculated, with each one having its own purpose. Table 4.3 displays a table of **cell percents**, where the denominator is the entire sample. There are 13.3% of all respondents in this data set who are male and have graduated high school. Table 4.4 displays the **row percents**, where the denominator is the row total. Relatively more males than females completed a four year degree: 15.3% of males completed a BS degree, compared to 14.2% of females. Table 4.5 displays the **column percents**, where the denominator is the column total. The majority of PhD graduates were male; 66.7% of the PhD graduates were male and 33.3% female.

Table 4.3: *Cell percents: Percent out of the entire data set*

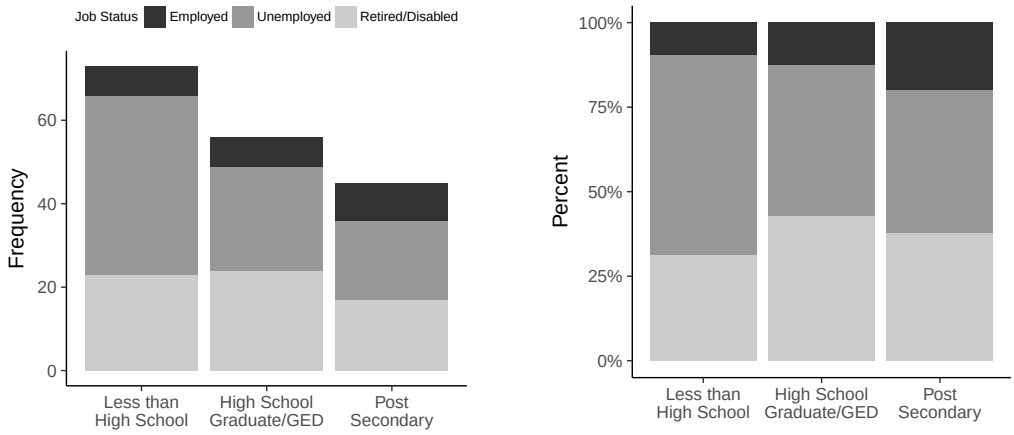
	<HS	Some HS	HS Grad	Some college	BS	MS	PhD	Total
Male	1.4	6.5	13.3	6.1	5.8	2.7	2.0	37.8
Female	0.3	14.3	25.5	10.2	8.8	2.0	1.0	62.2
Total	1.7	20.7	38.8	16.3	14.6	4.8	3.1	100.0

Table 4.4: *Row percents: Percent of educational level within each gender*

	<HS	Some HS	HS Grad	Some college	BS	MS	PhD	Total
Male	3.6	17.1	35.1	16.2	15.3	7.2	5.4	100.0
Female	0.5	23.0	41.0	16.4	14.2	3.3	1.6	100.0

Table 4.5: *Column percents: Percent of gender within each educational level*

	<HS	Some HS	HS Grad	Some college	BS	MS	PhD
Male	80.0	31.1	34.2	37.5	39.5	57.1	66.7
Female	20.0	68.9	65.8	62.5	60.5	42.9	33.3
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0



(a) *Frequency on the vertical axis*

(b) *Percents on the vertical axis*

Figure 4.10: *Distribution of current job status within highest education attained in the parental HIV data set*

Bar Charts

To visually compare the distribution of one categorical variable within levels of another categorical variable, we return to bar charts. Figure 4.10 compares the distribution of job status within the highest educational level attained using the parental HIV data set.

Stacked bar charts can be informative when plotting percentages instead of counts. Figure 4.10b shows how the proportion of observations in each job status category compare across each level of highest educational level attained. This plot is created by plotting column percentages, so that all percents within a column add up to 100%. The group of respondents whose highest education level is post secondary has the highest proportion of employed respondents compared to the other two educational level groups.

The default for many software programs is a stacked bar chart as shown in Figure 4.10. For few categories this option could be acceptable. However, consider the proportion of those with HS/GED who are unemployed, is it bigger or smaller than the percent of those with post secondary degrees who are currently unemployed? It is difficult to tell in a stacked bar chart, but much easier to see the difference with the bars placed side by side (Figure 4.11).

The Cleveland Dot Plot can also be done across groups. Figure 4.12 demonstrates a slight variation where the dot is placed at the end of the solid line, instead of on a reference line as in Figure

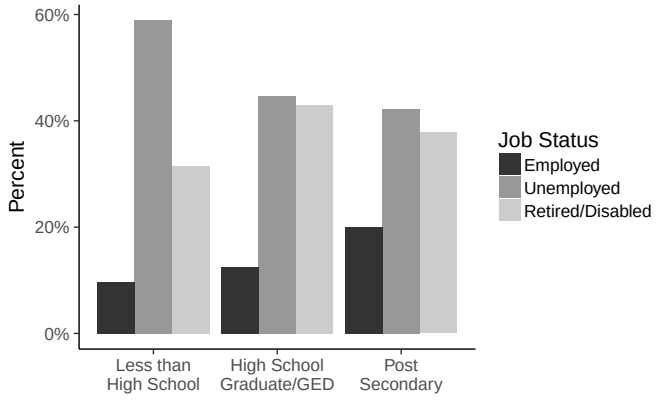


Figure 4.11: *Side by side bar chart depicting the percent of current job status within highest education level attained in the parental HIV data set*

4.3. This is also the first demonstration of **paneling**, where the data for each level of the grouping variable are set apart from the other levels using a rectangular border or frame. This method helps to visually separate the groups.

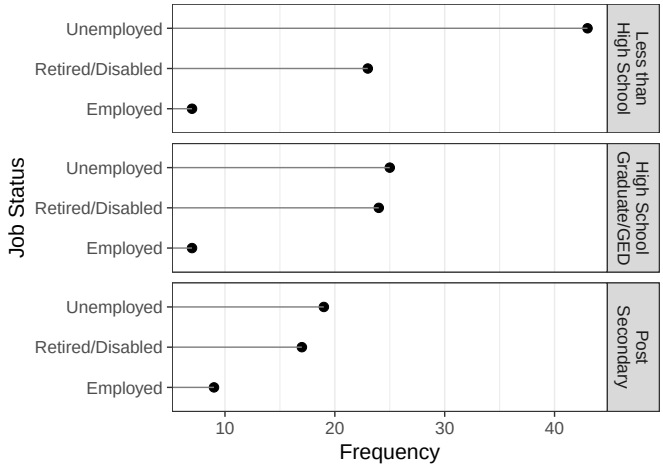


Figure 4.12: *Dot plot demonstrating the frequency of job status within highest education attained*

Mosaic Plot

Bar plots and dot plots show either the row or column percents of a bivariate comparison. They compare the distribution of one categorical variable within levels of a second categorical variable. **Mosaic plots** provide a graphical method to compare the association between two categorical variables.

Figure 4.13 compares job status to educational level by visualizing the cell proportions. The heights of the boxes correspond to the marginal distribution of educational level, and the widths of the boxes correspond to the marginal distribution of job status. The area of each smaller rectangle is proportional to the percent of data with that combination of levels. Using Table 4.6 as a numerical reference, 4% of responses in the parental HIV data set have a GED and are employed, whereas

24.7% have less than a HS education and are currently unemployed. This may seem like a high proportion of unemployment, but recall that these data were collected in the early nineties, where there was very limited treatment options for HIV positive individuals.

Table 4.6: *Cell percentages for the combination of educational level and job status*

	Employed	Unemployed	Retired/Disabled
Less than High School	4.0	24.7	13.2
High School Graduate/GED	4.0	14.4	13.8
Post Secondary	5.2	10.9	9.8

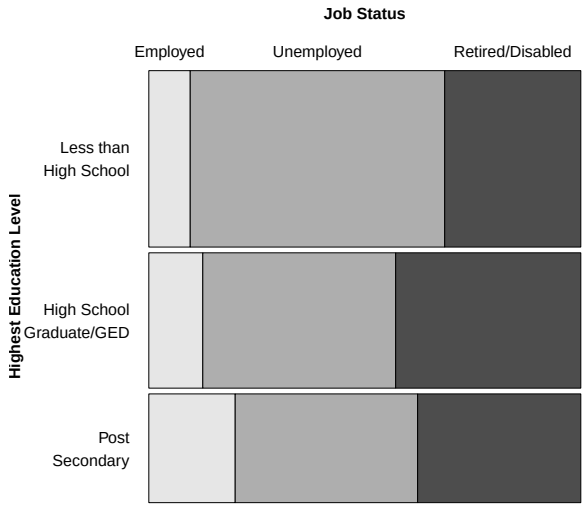


Figure 4.13: *Mosaic plot comparing job status and educational level*

4.3.2 Continuous versus Continuous Data

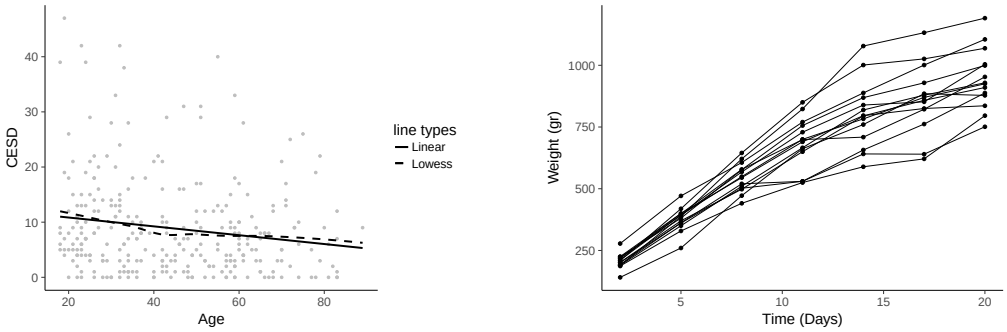
The most common method of visualizing the relationship between two continuous variables is the **scatterplot** (Figure 4.14a). Lines are often added to help see the trend in the data points. The two most common best fit methods are the straight line (shown as a solid line) and the **lowess** smoother line (shown as a dashed line). In this plot the points have been colored grey to place the emphasis on the lines. These two methods are both regression techniques discussed in Chapter **??**.

Line Plots

Line plots connect the points with a line. This is typical in time series, or **profile plots** where the goal is to track the data on an individual or a population over time. One line is plotted per individual. For data sets with larger number of individuals this process can create an unreadable plot. We suggest plotting data on a random subset of individuals to explore the data or create multiple such plots for subsets if feasible.

Figure 4.14b uses the mice data set described in Appendix A where the weights of mice were measured periodically for about a month. The mice grew almost at the same rate until about 8 days, and then started to separate due to individual and treatment characteristics. This particular type of plot is also known as a **spaghetti plot** or a **growth curve** because it typically presents a measure of

growth over time. We use this plot again in Chapter **??** when discussing how to analyze longitudinal data.



(a) *Scatterplot of age against CESD, with a dashed lowess line and solid best fit linear line*

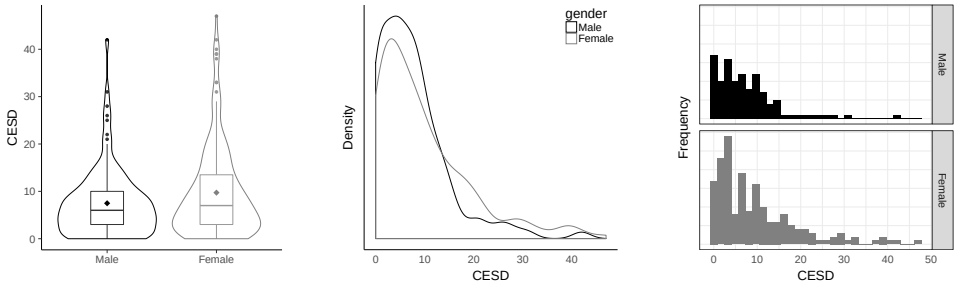
(b) *Weight over time for 14 mice*

Figure 4.14: *Two types of scatterplots for examining the relationship between two continuous variables. Points in the scatterplot (a) are not connected to other points whereas in the line plot (b) points from the same mouse are connected with a line*

4.3.3 Continuous versus Categorical Data

When comparing the distribution of a continuous variable across levels of a categorical variable, the same types of plots seen for a single continuous variable can be used including histograms, density plots, boxplots and violin plots.

Figure 4.15 demonstrates how to plot the distribution within each group side by side (a), overlaying plots onto the same plotting grid (b), or to create a grid of **panels** (c) with one group per panel. It is very important to use a shared or common axis when comparing conditional distributions across groups.



(a) *Side by side*

(b) *Overlapping*

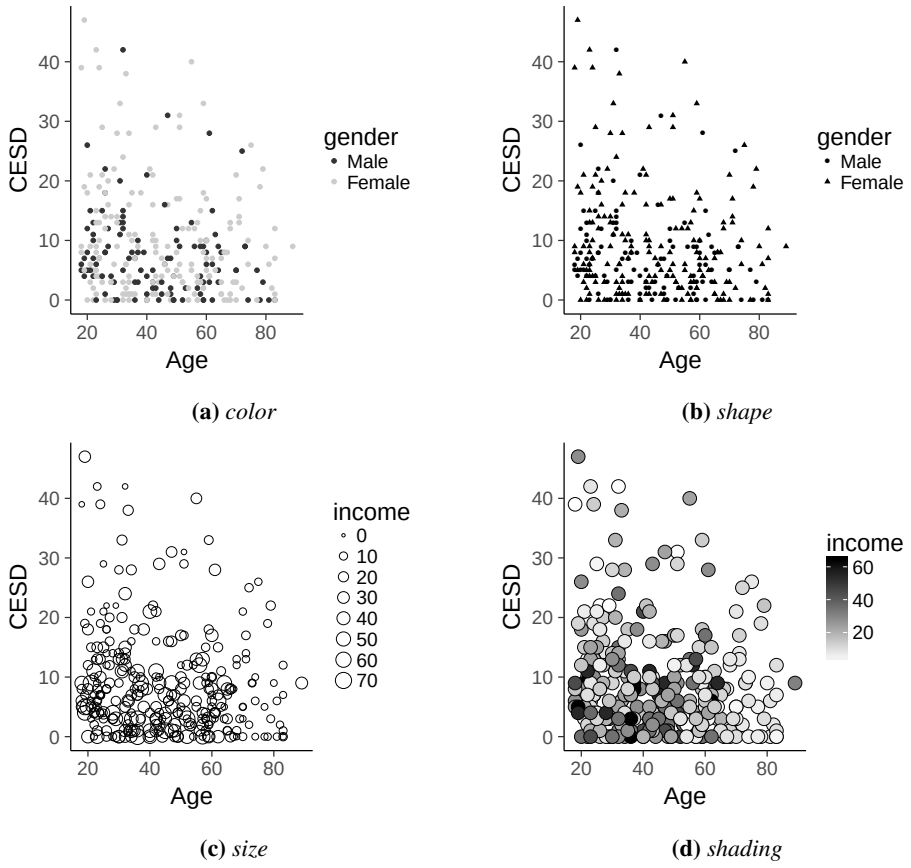
(c) *Paneled*

Figure 4.15: *Three methods to compare the distribution of the continuous variable CESD across levels of the categorical variable gender*

4.4 Multivariate Data

The techniques of applying colors, shadings, positioning and paneling of data from multiple groups to visualize bivariate relationships can be extended to visualize relationships among more than two variables simultaneously.

Figure 4.16 demonstrates how a third dimension can be added onto a bivariate scatterplot by changing the (a) color or (b) shape of the points according to the level of a third categorical variable, or by changing the (c) size or (d) fill shade of the points according to a continuous variable. For example, plot (a) allows us to see that the points in the low range of age with high CESD score are primarily female (grey dots), and plot (d) that those in the higher income levels (darker shades) tend to have a CESD score below 10.



(a) *color* **(b)** *shape*

(c) *size* **(d)** *shading*

Figure 4.16: *Scatterplot of CESD as a function of age, with a third characteristic included using different methods*

There are many other ways to examine a multivariate relationship. Even on each of these plots just discussed a fourth layer could be added, such as changing the size of the point by income in plots (a) and (b) and shape of the point by gender in plots (c) and (d). Another method to examine a multivariate relationship is to use paneling in two dimensions. Figure 4.17 demonstrates how we can examine the histogram of overall BSI (brief symptom inventory) score for each combination of employment status and highest educational level attained.

A **scatterplot matrix** is a common tool to quickly examine the bivariate relationships between multiple continuous variables simultaneously. Figure 4.18 demonstrates a publication-ready version of a scatterplot matrix that has many features added, including the pairwise correlation (see Chapter **??** for details), univariate histograms and lowess lines on the scatterplots. Each diagonal rectangle displays the histogram of a particular variable. This single plot lets us identify characteristics of the data such as (1) the distribution of BSI is skewed with a long tail to the right as demonstrated by the tall bar representing frequency of responses for low values of BSI, and a few very short bars for high values of BSI, (2) the parent bonding care sub-scale is also considered skewed since the bars

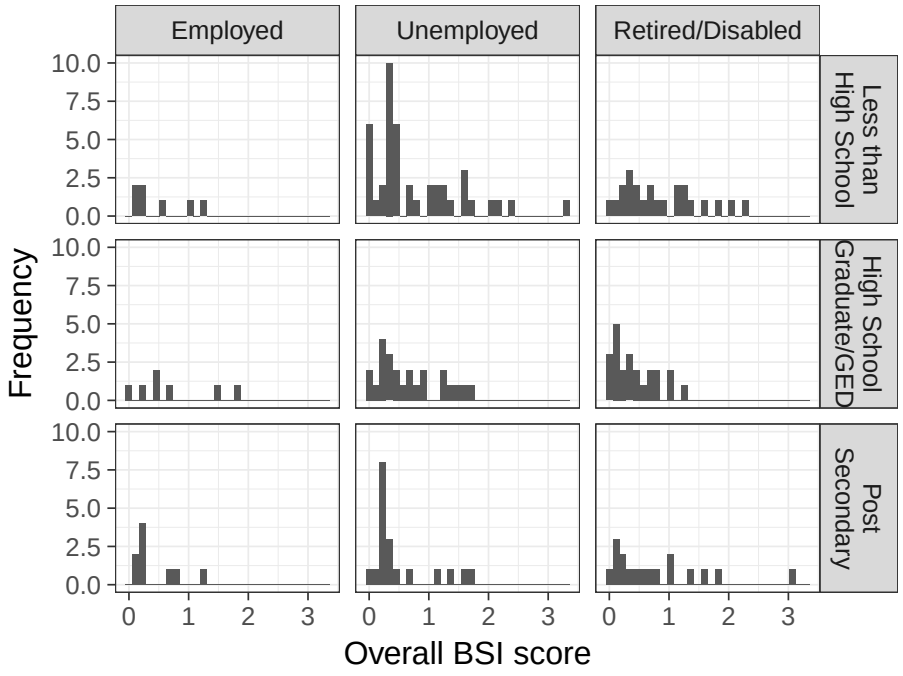


Figure 4.17: *A histogram of overall BSI score paneled on the combination of two other variables: employment status and highest educational level attained*

are short for low values of the sub-scale and increase in height as the scale increases, (3) there is a moderate positive correlation ($r = 0.37$) between the age a youth starts smoking and drinking and (4) none of the other three variables seem to be correlated with BSI since the lowess lines through the scatterplots are all approximately horizontal.

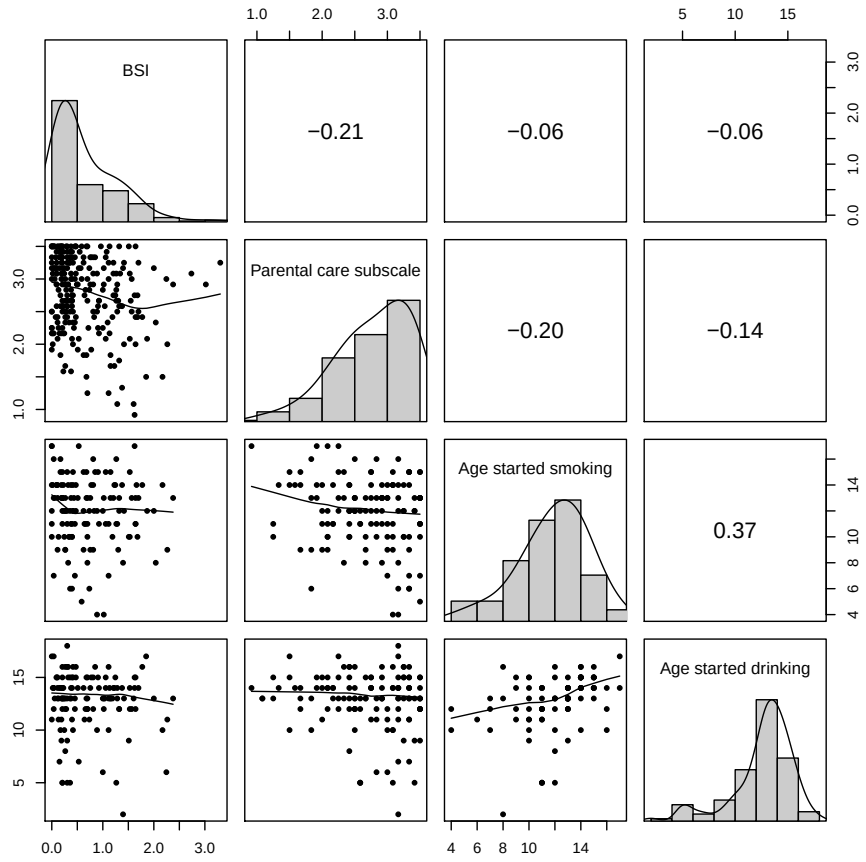


Figure 4.18: *Scatterplot matrix with histograms and density plots along the diagonal, and pairwise correlation values above the diagonal*

4.5 Discussion of computer programs

Each general statistical software package has commands or procedures to produce many, if not all, of the plots or visualizations we describe in this chapter. Table 4.7 shows which command can be used to produce a particular plot using the three major packages discussed in this book.

All graphics in this chapter were produced in R Version 3.3.1 (R Core Team, 2017). The full R code and colored versions can be found on the CRC Press web supplemental site.

Additional notes for Table 4.7:

- **R:** Entries that are in `monospace font` are functions within Base R. Entries in normal font are packages that contain functions (not specifically listed here) that are used to create the selected plot.
- **SAS:** All entries are individual procedures, called PROCs. Not all are part of BASE SAS. PROC GPLOT, GCHART, and GTL are part of SAS/GRAPH. PROC TEMPLATE is listed here as part of the Graph Template Language, which provides full customization of SAS Graphics.
- **SPSS:** With the exception of creating tables, all available graphics are best built using the Chart

Builder. Table entries provide guidance for the reader to find the appropriate selection. The Chart Builder also has tools to easily change the color and shape of the point (or marker).

- **Stata:** Entries that are in `monospace font` are functions, normal font indicates an option within that function. Entries marked with an asterisk * are user written functions.

Table 4.7: *Software commands for plotting*

	Visualizations	R	SAS	SPSS	Stata
Univariate					
Categorical	Table	`table`	FREQ	FREQUENCIES	`table`, `tabulate`
	Bar Chart	`plot`, ggplot2	GCHART	Bar	`graph bar`, `catplot*`,`tabplot*`
	Dot Plot	`dotchart`, ggplot2	SGPLOT	Scatter	`graph dot`
			SGPLOT	-Summary Point Plot	`graph dot`
	Pie Chart	`pie`	GCHART	Pie	`graph pie`
Continuous	Stem-Leaf	`stem`	SGPLOT, UNIVARIATE	EXAMINE	`stem`
	Histograms	`hist`, ggplot2	SGPLOT, UNIVARIATE	Histogram	`histogram`
	Kernel Density	`plot`, ggplot2	SGPLOT, UNIVARIATE	Histogram	`kdensity`
	Boxplot	`boxplot`, ggplot2	BOXPLOT, SGPLOT	Boxplot	`graph box`
	Violin Plots	ggplot2	TEMPLATE	-	`vioplot`
Bivariate					
Cat v Cat	Two-way table	`table`	FREQ	CROSS TABS	`table`, `tabulate`
	Bar Chart	`plot`, ggplot2	GCHART	Bar	`graph bar`, `catplot*`, `tabplot*`
	Mosaic Plot	`mosaicplot`, vcd	TEMPLATE	-	`spineplot`*[a]
Cont v Cont	Scatterplot	`plot`, ggplot2	GPLOT, SGPLOT	Scatter	`scatter`
				-Scatterplot Matrix	
	Line Plot	`plot`, ggplot2	GPLOT, SGPLOT	Line	`line`
Cont v. cat	grouping	`plot`, ggplot2	BOXPLOT, SGPLOT	Histogram, Boxplot	`histogram`, `graph box`
Multivariate					
	paneling	ggplot2, lattice	SGPANEL	Groups tab	`by`
	colors, size	`plot`, ggplot2	GPLOT, GCHART	-	`color`, msize, weight
	scatterplot matrix	psych, lattice, `pairs`, car	SGSCATTER	Scatter/Dot	`graph matrix`

[a]Technically not a mosaic plot (Hummel, 1996)

4.6 What to watch out for

- **Avoid complexity.** We advise against using too many enhancements on a single plot since doing so can confuse the reader instead of providing a better understanding. The purpose of most graphics is to understand distributional patterns, and to identify odd data points. Not all layers provide illumination. For example, in Figure 4.16c there is much over-plotting in the lower left so that it is difficult to see if there is a pattern emerging. In this case coloring the points for different income levels may be more helpful than changing the size, although barely.
- **Choose colors mindfully.** All plots in this textbook are either in black and white or shaded using a grayscale. This is a necessary adjustment for black and white printing, but also is a consideration for colorblind readers. We recommend using a colorblind friendly color pallet for publications involving color.
- **Do not add extra dimensions.** We do not demonstrate plots such as 3D pie charts, or 3D bar charts in this text. In these cases the third dimension does not provide true information, and is considered "chart junk" that can be very misleading. This is not a global recommendation, there are circumstances in which a third dimension does contain additional information that may be useful to include in the visualization.
- **Be truthful with the scaling.** Be mindful of the scaling of the vertical axis. For example, Figure 4.2b plots a percentage on the vertical axis with a high value of near 50%. We scale the vertical axis to 100% here to put the difference in percentages in the context of the overall range. A 2% point difference between categories can appear huge if the vertical axis only has a total range of say 5%. Displaying a vertical axis that is too large or too small relative to the data is one of the most common ways in which graphics can be misleading.
- **Check publishing guidelines.** If the goal is to publish using graphics, then be sure to check the rules carefully. Some publications have rules regarding features such as whether or not there is a box around the plot versus only showing the horizontal and vertical axes.
- **Be consistent with selected themes.** For example, if the first plot has a clear background and a box outlining the edge, all subsequent plots should have that same theme. If a second categorical variable controls the color or shape of the points for one plot, then all subsequent plots that also use that same categorical variable should have the same color and/or shape scheme applied.
- **Do not over-interpret.** One should not judge statistical significance based on graphs unless they are specifically designed for such. Even if they are (seemingly) designed for it (e.g., graphing confidence intervals for group comparisons or 95% pointwise confidence intervals or confidence bands for graphs) they can produce different and potentially misleading results and interpretations as compared to appropriate hypothesis tests.
- **Plotting with missing data.** Each software program has slightly different default values for how missing data are handled in different plots. For example, the `table` function in R does not automatically include a column for missing data, but a bar chart created using ggplot2 will show a bar for the missing category. In all software package, values that are missing will be omitted from continuous data plots such as histograms and scatterplots, with typically no mention of the sample size being used to create the plot shown.

4.7 Summary

Data visualization is a powerful tool that can be used to explore and understand your data. Data values that need to be recoded can easily be identified and trends in the data can be uncovered. Graphics can be used to confirm that the data meet certain criteria, or assumptions that are needed for further statistical analysis such as the specialized analysis methods presented in the remainder of the book. Data visualizations can also enhance understanding of statistical analysis results, but do not serve as a substitute. Even though the saying "a picture is worth a thousand words" may be

true, and a graph can provide more information than a block of text or a table of numbers, when it comes to more than a few variables, the options for graphically representing the data are limited.

We have demonstrated a wide variety of visualizations in this chapter. Some plots require detailed written explanations and are more suitable for reports or publications that do not have length restrictions. Authors should attempt to strike a balance between complexity and interpretability of graphics, yet always aim to elucidate characteristics of data relevant to the question being asked or answered.

There are many other types of graphics that we do not discuss such as heatmaps and geographic maps. These are typically considered specialized graphics for specific analyses. We present some specialized plots in the appropriate chapters of this book but do not attempt to cover all possible ways to display information visually. We recommend looking at Edward Tufte's pioneering work for more ideas and guidelines on how to create informative graphics (Tufte, 1986, 2006).

For the programming language details on how to make the graphics shown in this chapter and others, we refer readers to this book's supplemental webpage and reference books such as the R Graphics Cookbook by Chang (2013), R Graphics by Murrell (2011), ggplot2: Elegant Graphics for Data Analysis by Wickham (2016), Statistical Graphics in SAS Kuhfeld (2010), A Visual Guide to Creating Graphs Interactively by Matange and Bottitta (2016), Handbook of Statistical Graphics using SAS by Der and Everitt (2014), the IBM SPSS Statistics 24 Brief Guide (IBM, 2016), Building SPSS Graphs to Understand Data by Aldrich and Rodriguez (2012), Speaking Stata Graphics by Cox (2014), and A Visual Guide to Stata Graphics by Mitchell (2012).

4.8 Problems

4.1 Using your statistical package, compute a scatterplot of income versus employment status from the depression data set. From the data in this table, decide if there are any adults whose income is unusual considering their employment status. Are there any adults in the data set whom you think are unusual?

4.2 Using a statistical software program, compute histograms for mothers' and fathers' heights and weights from the lung function data set described in Section **??** and in **Section A.5** of Appendix A. The data and codebook can be obtained from the web site listed in Section **A.7** of Appendix A. Describe cases that you consider to be outliers.

4.3 From the lung function data set in the previous problem, determine how many families have one child, two children, and three children between the ages of 7 and 18.

4.4 For the lung function data set, produce a two-way table of gender of child 1 versus gender of child 2 (for families with at least two children). Comment.

4.5 For the lung cancer data set (see the codebook in **Table 13.1 of Section 13.3 and Section A.7** of Appendix A for how to obtain the data) use a statistical package of your choice to

a) compute a histogram of the variable Days

b) for every other variable produce a frequency table of all possible values.

4.6 For the lung cancer data set in the previous problem,

a) produce a separate histogram of the variable Days for small and large tumor sizes (0 and 1 values of the variable Staget)

b) compute a two-way frequency table of the variable Staget versus the variable Death

c) comment on the results of (a) and (b).

4.7 For the depression data set, determine if any of the variables have observations that do not fall within the ranges given in Table **??**, codebook for depression data.

4.8 For the lung function data set, create a new variable called AGEDIFF = (age of child 1) – (age of child 2) for families with at least two children. Produce a frequency count of this variable. Are there any negative values? Comment.

4.9 Using the parental HIV data set create a few visualizations to explore the relationship between the variables of interest listed below. In a few sentences describe the information about the given relationship you learn from each graph, and what specific features of the graph led you to that conclusion.

a) The relationship between the age when a child starts smoking and when they start drinking.

b) Ethnicity, a choice of one neighborhood characteristic, and financial situation of the household.

c) Attendance of religious services and level of religiousness/spiritualism.

4.10 Using the lung function data explore and describe the following relationships:

a) How does the residential area affect the lung function of the parents?

b) For the oldest child, plot the relationship between FEV1 on (i) age; (ii) height; (iii) weight using a scatterplot and lowess line.

4.11 Using a scatterplot matrix, repeat the previous problem for fathers' measurements instead of those of the oldest child. Did you find the same pattern of relationships between body measurements and FEV1 in fathers as you did for the oldest child?

4.12 Using the mice data, create a profile plot for the average weight of mice per group over time.

Bibliography

James O. Aldrich and Hilda M. Rodriguez. *Building SPSS Graphs to Understand Data*. SAGE Publications, Inc, 2012. ISBN 1452216843.

Winston Chang. *R Graphics Cookbook*. O'Reilly Media, Inc., 2013. ISBN 1449316956, 9781449316952.

William S. Cleveland. *Visualizing Data*. Hobart Press, 1993. ISBN 0963488406.

Nicholas J. Cox. *Speaking Stata Graphics*. Stata Press, 2014. ISBN 978-1-59718-144-0.

Geoff Der and Brian S. Everitt. *Handbook of Statistical Graphics Using SAS*. CRC Press Inc, 2014. ISBN 1466599030.

J. Hummel. Linked bar charts: Analysing categorical data graphically. *Computational Statistics*, pages 23–33, 1996.

IBM. *IBM SPSS Statistics 24 Brief Guide*. IBM Corp, 2016.

Warren F. Kuhfeld. *Statistical Graphics in SAS: An Introduction to the Graph Template Language and the Statistical Graphics Procedures*. SAS Publishing, 2010. ISBN 1607644851.

Sanjay Matange and Jeanette Bottitta. *SAS ODS Graphics Designer by Example: A Visual Guide to Creating Graphs Interactively*. SAS Institute, 2016. ISBN 1612901913.

Michael N. Mitchell. *A Visual Guide to Stata Graphics, Third Edition*. Stata Press, 2012. ISBN 1597181064.

Paul Murrell. *R Graphics, Second Edition (Chapman & Hall/CRC The R Series)*. CRC Press, 2011. ISBN 1439831769.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL `https://www.R-project.org`.

Edward R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, USA, 1986. ISBN 0-9613921-0-X.

Edward R. Tufte. *Beautiful Evidence*. Graphics Press, 2006. ISBN 1930824165.

John W. Tukey. Some graphic and semi-graphic displays. In TA Bancroft, editor, *Statistical papers: in honor of Georges W. Snedecor*. Iowa, Iowa State university press, 1972. Editing assisted by Susan Alice Brown.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis (Use R!)*. Springer, 2016. ISBN 331924275X.

Index