Chapter 1

Data Visualization

1.1 Introduction

Visualizing your data is one of the most important thing you can learn to do. There are certain features and patterns in the data that cannot be uncovered with summary statistics alone. And as the old adage goes: "a picture is worth a thousands words". Good data visualizations can convey much more information than a block of text and a string of numbers. This chapter introduces a series of plot types for both categorical and continuous data types. 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 (multivariate) variables.

placeholder for discussion of graphics for confirmatory vs exploratory analysis.

This chapter uses several data sets introduced in Chapter **??** and described in **??**. 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 **??** containing a discussion of graphical capabilities in other computer software programs. Additionally links and references to external learning resources have been provided at the end of the chapter.

There are three levels of visualizations that can be created, examples are shown from left to right in Figure 1.1.

- **For your eyes only:** Made by the analyst, for the analyst. These plots are quick and easy to make, using the default options without any annotation or context. These are meant to be looked at once or twice for exploratory analysis to understand your data.

- **For an internal report:** Some chosen plots are then cleaned up to be shared with others for example in an weekly team meeting. These need to

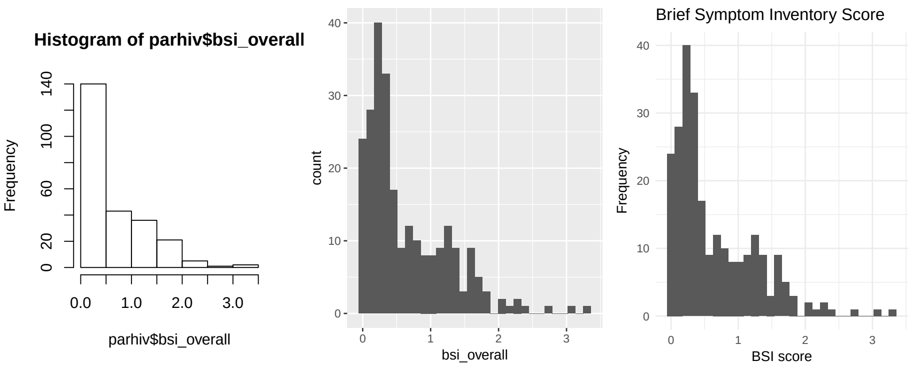


Figure 1.1: Three levels of graphic quality and completeness

completely stand on their own, but can be slightly less than perfect. Axes labels, titles, colors as needed, annotations and other captions.

- **For publication or external report:** These are meant to be shared with other stakeholders such as the public, your client, or administration. Very few plots make it this far. These have all the bells and whistles as a plot for a report, but additionally are often saved to an external file of specific size or file type, with high resolution. For print publication in most journals or books figures also need to be in black and white (possibly grayscale).

1.2 Univariate

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

1.2.1 Categorical data

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

Tables

Tables are the most common way to organize and display summary statistics of a categorical variable using just numbers. It is typically the case that you will see both the frequency (N), and the percent in the table for each category.

Education Level	N	%
Less than HS	79	43.2%
HS/GED	57	31.1%
Post Secondary	47	25.7%

Table 1.1: Education level among mothers with HIV

Table 1.1 shows that a quarter (47, 25.7%) of mothers in this data set have post-secondary school education level.

Program	Method
R	table()
SAS	PROC FREQ
STATA	table, tabulate
SPSS	

Table 1.2: Summary of methods to create tables

Bar Charts

A barchart or barplot takes these frequencies, and draws bars for each category along the X-axis where the height of the bars is determined by the frequencies seen in the table (Figure 1.2). A reasonable modification is to put the percentages on the y axis, and drop the bar for missing employment data (Figure 1.3).

The ordering of categories is important for readability. Nearly all statistical software packages will set the automatic factor ordering to alphabetical. If your data is ordinal in nature, tables and plots should read left to right along with that ordering, such as the educational level example above.

Cleveland Dot Plot

Bars use a lot of ink, and the width of the bar is meaningless. Cleveland Dot Plots **cite Cleveland** provides an alternative method to display the same information in a cleaner manner (Figure 1.4). This is especially helpful when there are a large number of categories to visualize. Since marital status is a nominal variable and not ordinal, this summary data is best displayed in descending order of frequency.

The standard dotplot plots the value of each record as a separate dot. Cleveland dot plots differ from the standard dot plot because they plot summary data,

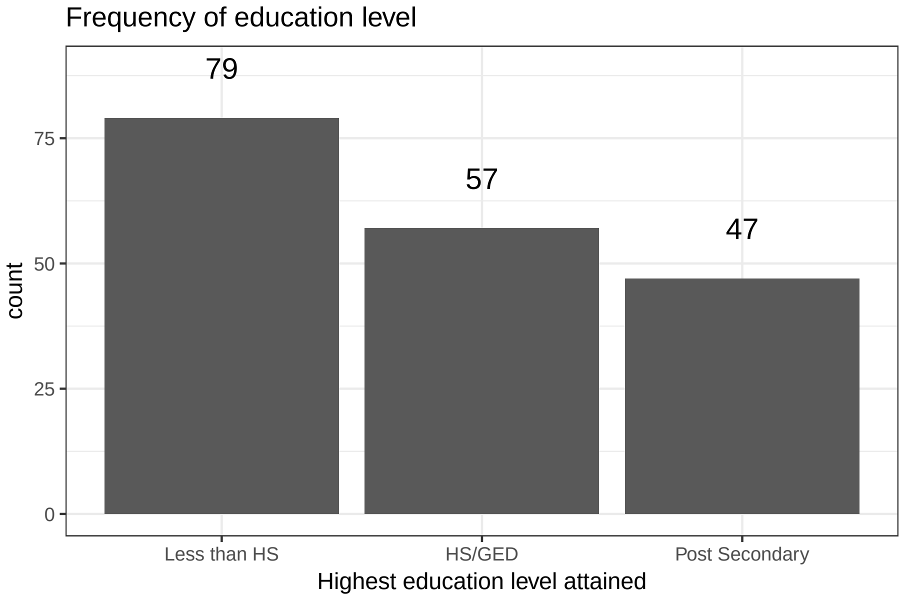


Figure 1.2: A bar chart with frequencies displayed

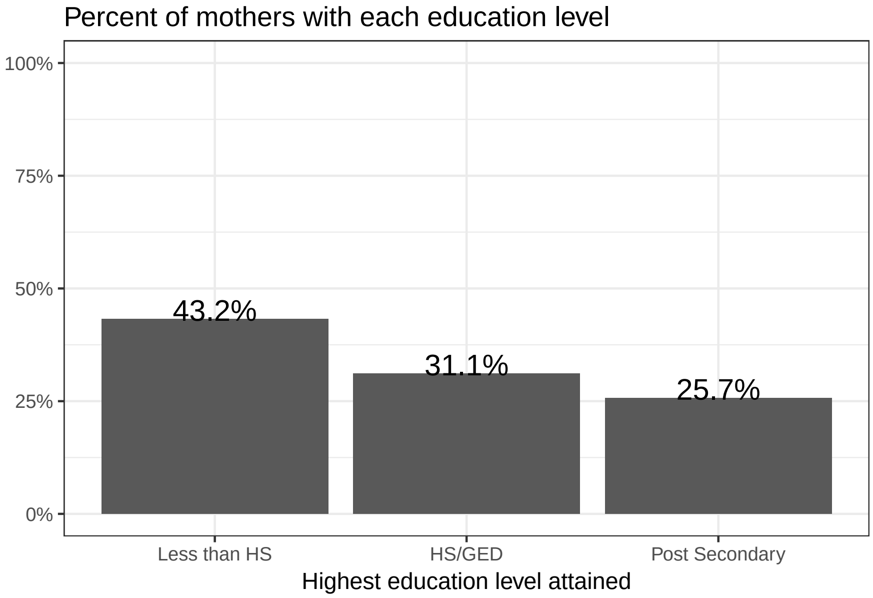


Figure 1.3: A bar chart with percentages on the y-axis

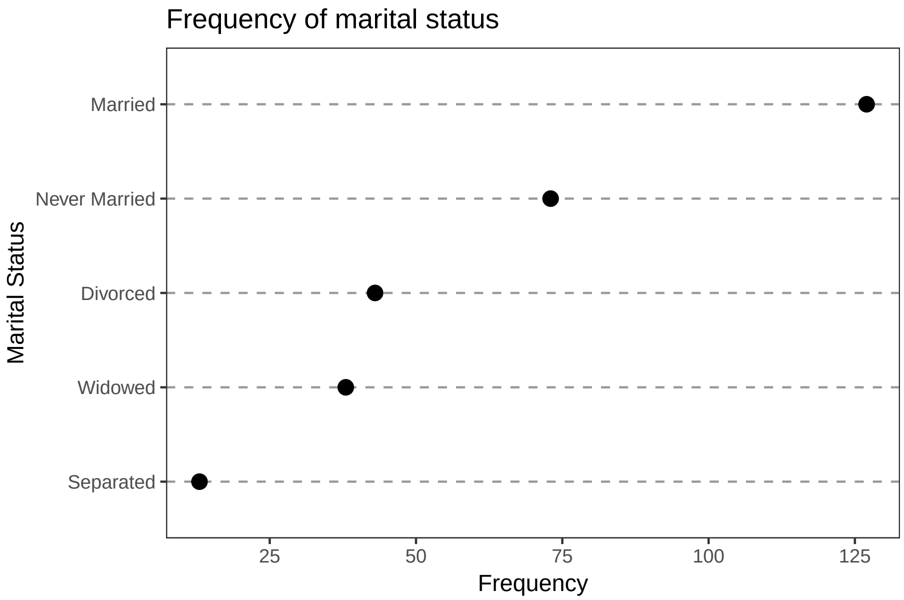


Figure 1.4: A Cleveland Dot Plot of marital status

not raw data. However, after you summarize the data you are effectively putting a dot for each data point, it's just that you only have one data point (the frequency) per category now.

Program	Method
R	dotchart(), geom_point()
SAS	PROC SGPLOT data=; dot type / response=
STATA	graph dot
SPSS	select counts from frequency table, right click and select "create graph", then select "dot"

Table 1.3: Summary of methods to create Cleveland Dot Plots

Pie Charts

Each wedge of a pie chart (Figure 1.5) contains an internal angle equal to the relative proportion of records in that category. However, human eyeballs can't distinguish between angles that are close in size as well as we can with heights. A necessary component to make any pie chart interpretable is labels with names and percentages for each wedge.

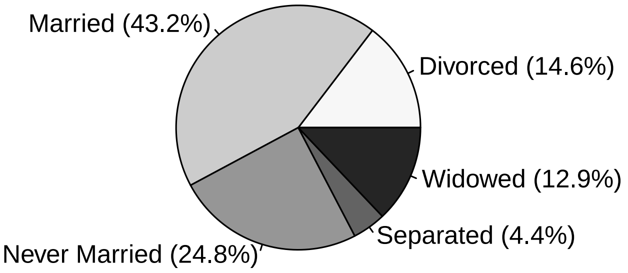


Figure 1.5: A fully labeled pie chart of marital status

Waffle Charts

Waffle charts convey the same information about relative frequencies for each level of a categorical variable. Figure 1.6 displays a grid of appropriately shaded squares, where each square represents a certain number of records in that category. The dessert plots (pie, waffle) tend to be found more in infographics than professional publications.

Numerical data types

Continuous data by definition has infinite possible values the variable can take on, so the above plots that display frequencies of records within a finite number categories do not apply here. To visualize numeric data we tend to want to display the actual value or the distribution the data points directly. Common plot types include: histogram, density, boxplots, violin plots, and qqplots. So how are these plots created?

First we introduce a stem-leaf plot **cite tukey** that demonstrates how values placed on a line can describe the shape of the distribution of values. Figure 1.7 displays the distance in feet it takes for a car to stop. The data used is from the `cars` **- cite** training data set built into `R`, since a stem-leaf plot is only useful for a relatively small data set.

Since stem-leaf plots display the value of every observation in the data set, the data values can be read directly. The first row reads that the first car took

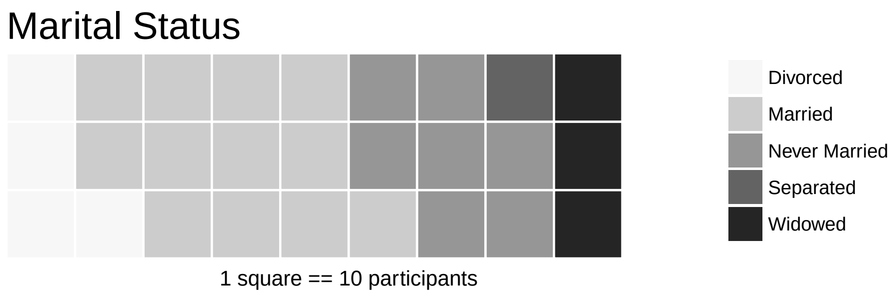


Figure 1.6: A waffle chart of marital status

```
 0 | 24
 1 | 004678
 2 | 0024666688
 3 | 22244466
 4 | 002668
 5 | 024466
 6 | 0468
 7 | 06
 8 | 045
 9 | 23
10 |
11 |
12 | 0
```

Figure 1.7: A stem-leaf plot of stopping distance in feet of cars from the 1920s

2' to stop, the second car took 4'. The second row is in the 10's place; two cars took 10' to stop, one at 14', one at 16' and so on. From this plot one can get an idea of of how the data is distributed.

Often you are not interested in the individual values of each data point, but the distribution of the data. In other words, where is the majority of the data? Does it look symmetric around some central point? Around what values do the bulk of the data lie? For example, the distribution of distance in feet it takes a car to stop is is unimodal, centered around 25', with an outlier at 120' stopping distance.

Histograms

Rather than showing the value of each observation, we prefer to think of the value as belonging to a *bin*. The height of the bars in a histogram display the frequency of values that fall into those of those bins. For example if we grouped the ages of individuals in the `depression` data set into 10 year bins, the frequency table would look like this:

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

To create a histogram, the values of the continuous variable are plotted on the x-axis, with the heights of the bars for each bin equal to the frequency of records within that bin (Figure 1.8). The main difference between a histogram and a barchart is that barcharts plot a categorical variable on the x-axis, so the vertical bars are separated. The x-axis of a histogram is continuous, so the bars touch each other, there is no gap between bins.

The size of the bins can highlight, or hide features in the data. It is recommended to start with the default value for your chosen statistical software package, and adjust as necessary. Figure 1.9 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`. This choice of bin width more clearly shows the most frequent age is around 23, unlike Figure 1.8 where it appears to be over 25.

Kernel density plots

Instead of plotting bars for each bin, you can sometimes get a better (or different) idea of the true shape of the distribution by plotting the kernel density plot. Figure 1.10 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 y-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.

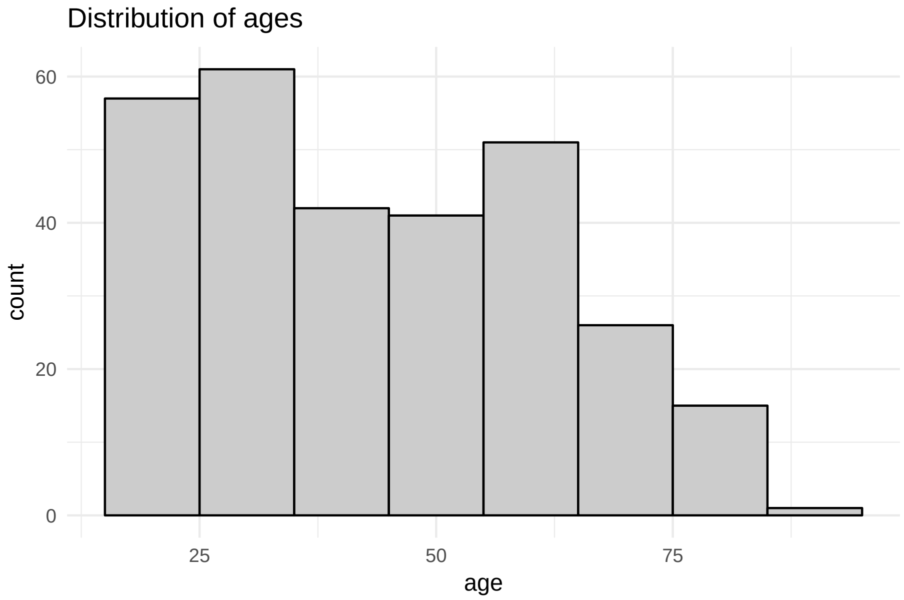


Figure 1.8: A histogram displaying the distribution of ages in the depression data set.

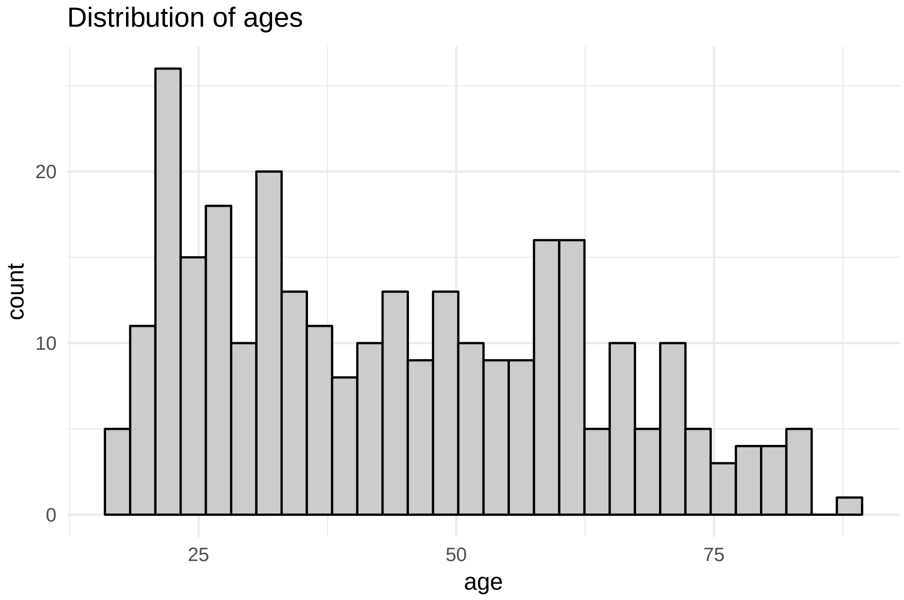


Figure 1.9: A histogram of ages using a bin width of range/30.

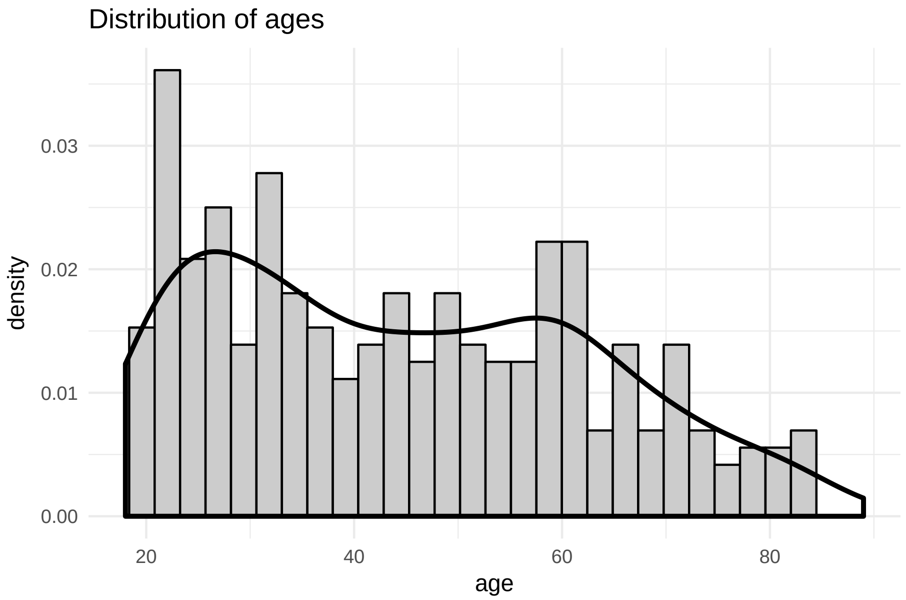


Figure 1.10: Density plot for the distribution of ages in the Depression data set.

Boxplots and Violin plots

Boxplots (a.k.a box-whisker plots) display the five number summary (Min, Q_1, Median, Q_3, Max) in graphical format. The data is split into equal sized quarters, the same number of data points are in each of the four sections of a boxplot (Figure 1.11).

The box outlines the middle 50%, or the interquartile-range ($\text{IQR} = Q_3 - Q_1$) of the data, and the horizontal lines (whiskers) extend from the 1st quartile (Q_1) down to the minimum value, and upwards from the third quartile Q_3 to the maximum value. By plotting the number of data points instead of the value of the data, this in the depression data set are between ages 18 and 28 as there are 59 and 89.

Some statistical packages will plot the *modified boxplot* by default. This is where the whiskers do not extend all the way out to the max and min, but out to the data points that are just inside the fences as calculated by the 1.5*IQR rule. **cite or explain?** In the modified boxplots, outliers are typically denoted as points or dots outside the whiskers. For example consider the continuous measure of depression, `cesd`; notice here the whisker extends to 25, the maximum value inside 1.5*IQR.

Additions that make boxplots much more informative are displayed in Figure **??**: 1) adding the mean as a point, 2) adding a violin plot to show the density (reflected around the mid-line of the boxplot), 3) adding the points directly, but jittered to avoid over-plotting. Violin plots are not commonly used, but they

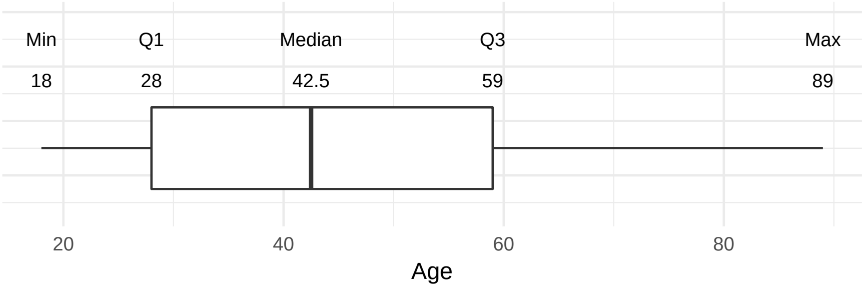


Figure 1.11: Boxplot for the distribution of ages in the depression data set

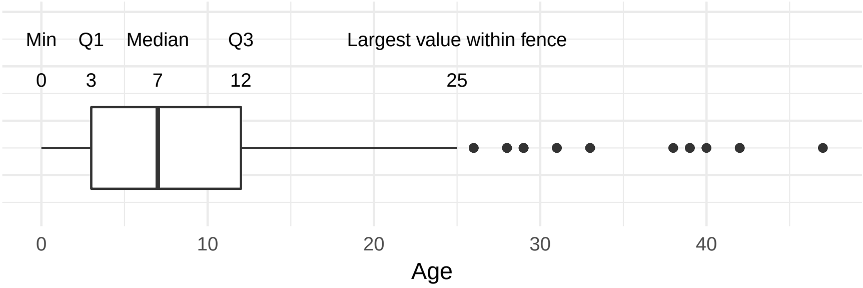
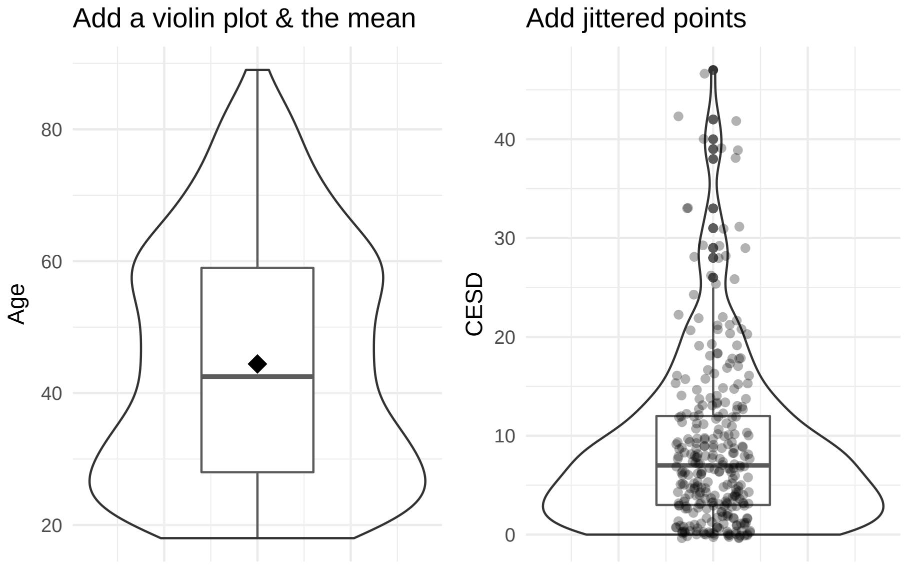


Figure 1.12: Modified boxplot for the distribution of CESD from the depression data set. Dots represent data points outside the upper fence.

can be very informative in that it can you the shape of the kernel density.


Add a violin plot & the mean

Add jittered points

1.3 Bivariate

Next we move on to introducing graphical methods to explore relationships between two variables. Many of the same plotting types such as boxplots and histograms that were introduced for univariate exploration will be used again here. This section is organized by data types **need better descriptor here**.

Categorical v Categorical

To compare the distribution of one categorical variable across levels of another categorical variable, primarily tables are created. These come by several names including cross-tabulations, contingency tables and two-way tables. Table 1.4 displays the frequency table of gender by education status, the values in the cells are the number of records on the data set with that combination of factor levels. There are 4 males with less than a HS degree, and 30 females with some college.

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

Table 1.4: Two-way frequency table of gender by educational status.

When group sizes are not comparable, it is more informative to compare percents instead of frequencies. There are three types of percents you can calculate, each one has it's own purpose. Table 1.5 displays a table of *cell percents*, where the denominator is the entire sample. There are 13.3% of all respondents in this data set are males who have graduated high school. Table 1.6 displays the *row percents*, where the denominator is the row total. More males than females completed a four year degree: 15.3 % of males got a BS degree, compared to 14.2% of females. Table 1.7 displays the *column percents*, where the denominator is the column total. PhD graduates were majority male; 66.7% of the PhD graduates were male and 33.3% female.

	<HS	BS	HS Grad	MS	PhD	Some college	Some HS
Male	1.4	5.8	13.3	2.7	2.0	6.1	6.5
Female	0.3	8.8	25.5	2.0	1.0	10.2	14.3

Table 1.5: Cell percents: Percent out of the entire data set.

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

Table 1.6: Row percents: Percent of educational category within each gender.

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

Table 1.7: Column percents: Percent of gender within each educational category.

Bar Charts

To visually compare the distribution of one categorical variable within levels of another categorical variable, we come back to bar charts. Figures 1.13 and 1.14 compare the distribution of job status within the highest educational level attained using the `PARHIV` data set (because it has fewer levels than the educational status variable on the `depress` data set).

The default for many software programs is a stacked barchart as shown in Figure 1.13 For few categories, or for few comparisons this 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 1.15).

The Cleveland Dot Plot can also be done across groups. Figure 1.16 demonstrates a slight variation where the dot is placed at the end of the solid line,

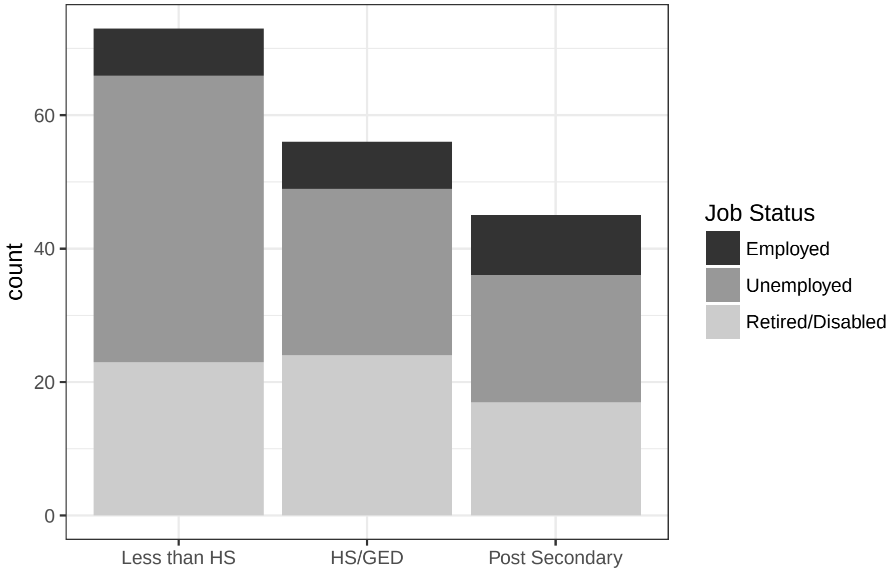


Figure 1.13: Frequency of current job status within highest education attained

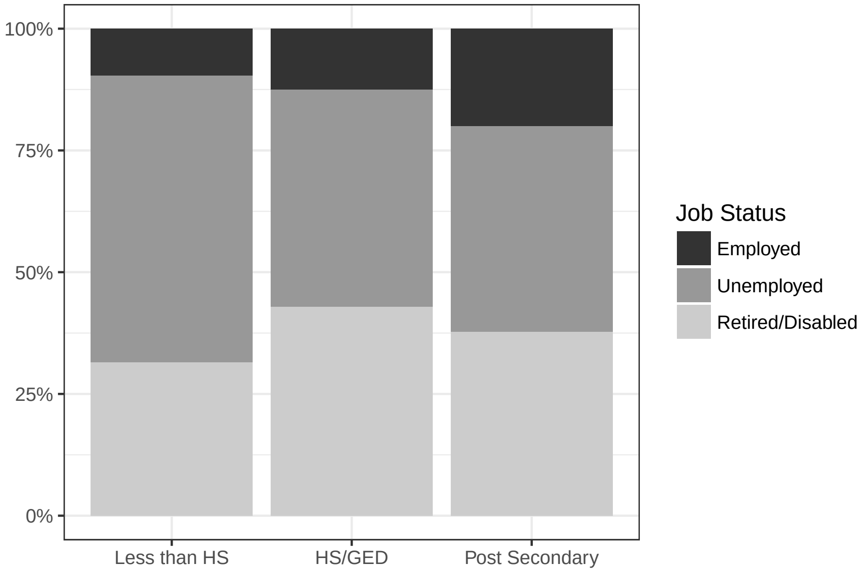


Figure 1.14: Percent of current job status within highest education attained

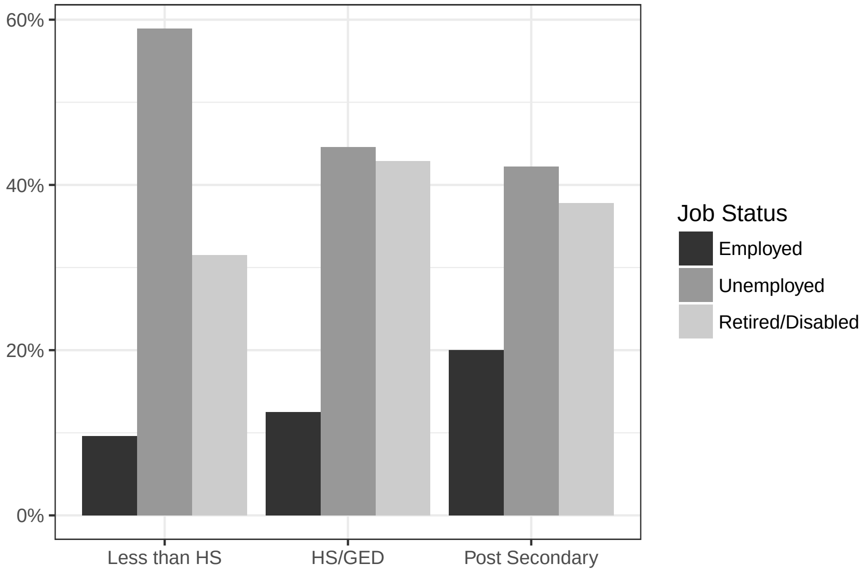


Figure 1.15: Side by side bar chart depicting the percent of current job status within highest education attained

instead of on a reference line as in Figure 1.4. This is also the first demonstration of *paneling*, where the data for each level of the grouping variable is set apart from the other levels using a rectangular border or frame. This helps to visually separate the groups.

Mosaic Plot

Bar plots and dot plots plot 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 1.17 compares job status to educational level by visualizing the cell proportions. The width of the boxes correspond to the marginal distribution of educational level, and the height of the boxes correspond to the marginal distribution of job status. The area of each smaller square is proportional to the percent of data with that combination of levels. Using Table 1.8 as a numerical reference, 4% of responses in the `PARHIV` data set have a GED and are employed, whereas 24.7% have less than a HS education and are currently unemployed.

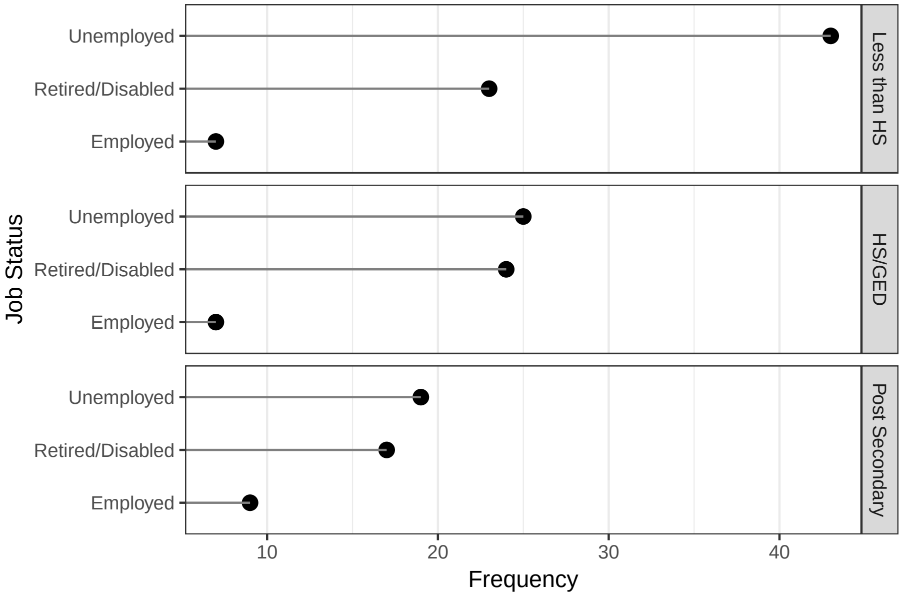


Figure 1.16: Dot plot demonstrating the frequency of Job status within highest education attained

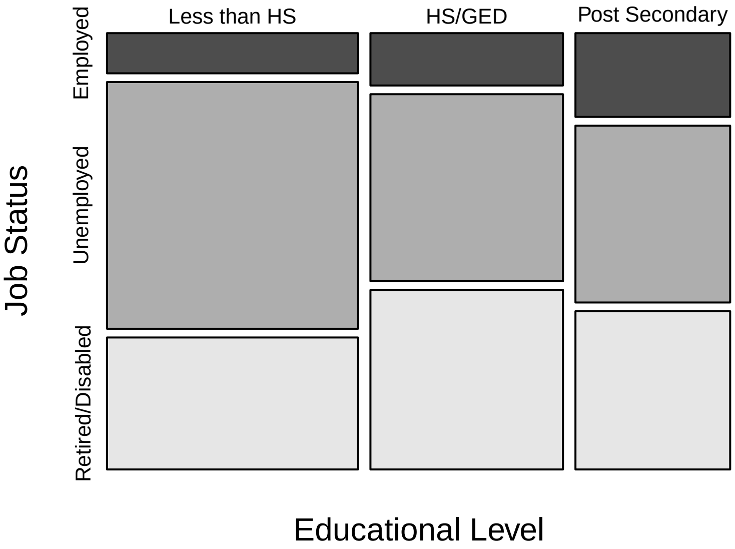


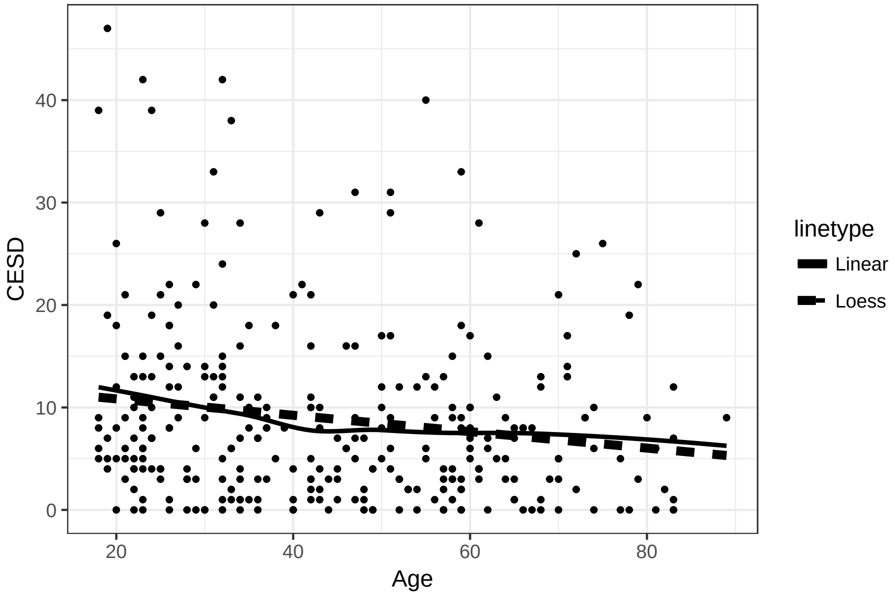
Figure 1.17: Mosiac plot comparing Job status and Educational level

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

Table 1.8: Cell percentages for the combination of Educational level and Job Status

Continuous v Continuous

The most common method of visualizing the relationship between two continuous variables is by using a scatterplot (Figure ??). Lines are often added to help see the trend in the data points. The two most common "best fit" straight line (thin solid) and the "lowess" smoother line (thick dashed).



Line Plots

Line plots are simply connecting points with a line. This is typical in time series, or profile plots where you are wanting to track the behavior of an individual or population over time. One line is plotted per individual. For data sets with larger number of individuals this can create an unreadable plot, we suggest plotting data on a random subset of individuals instead.

Figure 1.18 uses the `mice` data set described in **Appendix A** where the weight of mice were measured periodically for about a month. The mice grow almost at the same rate until about 8 days, then they start separating due to individual and treatment characteristics. This particular type of plot is also known as a growth curve, specifically because a measure of growth is plotted

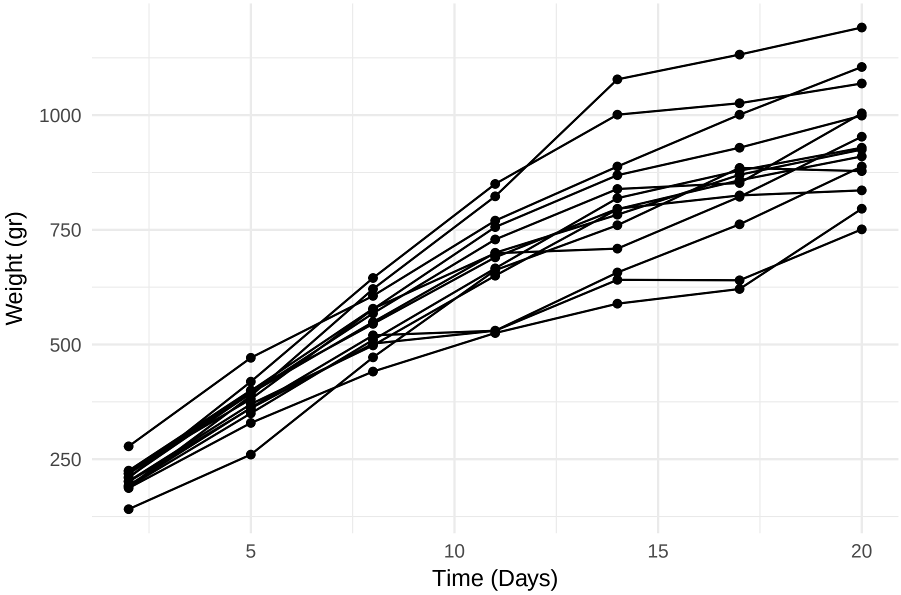


Figure 1.18: Weight over Time for 14 Mice

over time. We will see this plot again in Section **??** when discussing how to analyze longitudinal data.

Continuous v. Categorical

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

Figure 1.19 demonstrates how you can 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.

1.4 Multivariate

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

Figure 1.20 demonstrates how you can take bivariate scatterplot and add a third dimension by changing the a) color or b) shape the points according to the level of a third categorical variable, or change the c) sized or d) fill shade of

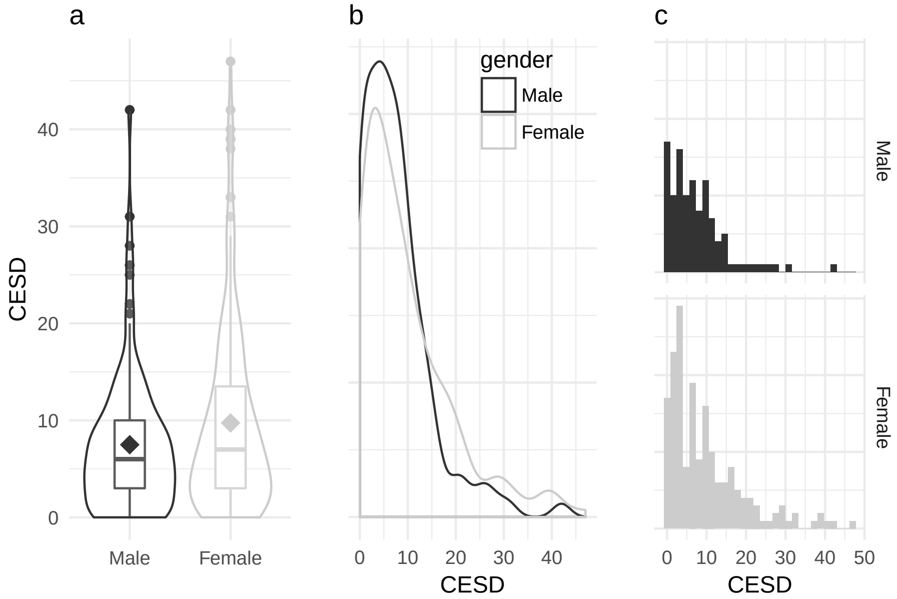


Figure 1.19: Three methods to compare the distribution of a continuous variable (CESD) across levels of a categorical variable (Gender).

the points according to a continuous variable.

There are many other ways, and in fact on each of these plots you could add yet a fourth layer, such as changing the size by income and shape by gender.Another method to examine a multivariate relationship is to use paneling in two dimensions. Figure **??** demonstrates how we can examine the distribution of Overall BSI score for each combination of employment status and highest educational level attained.

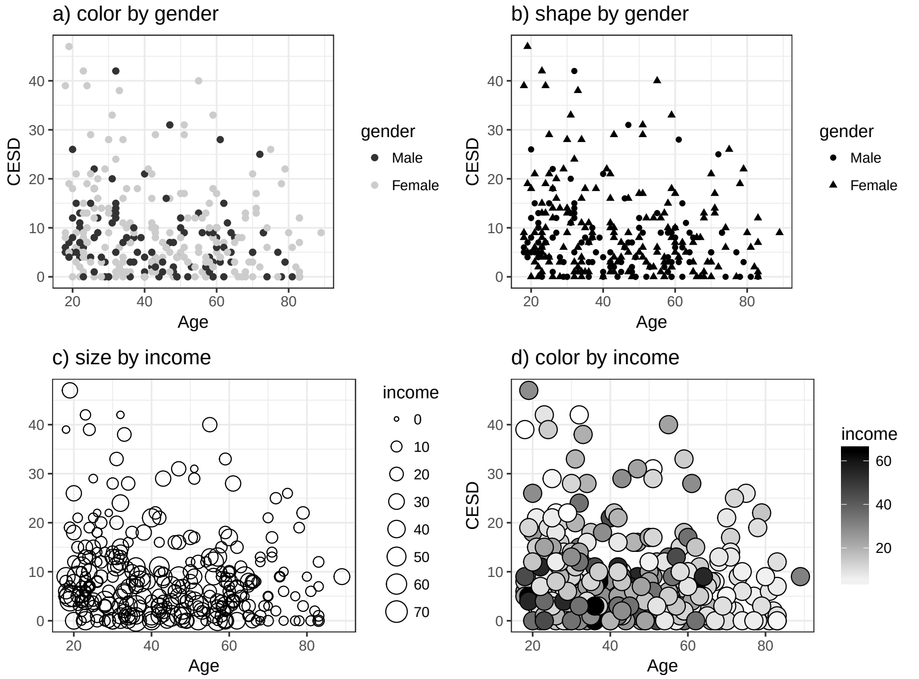
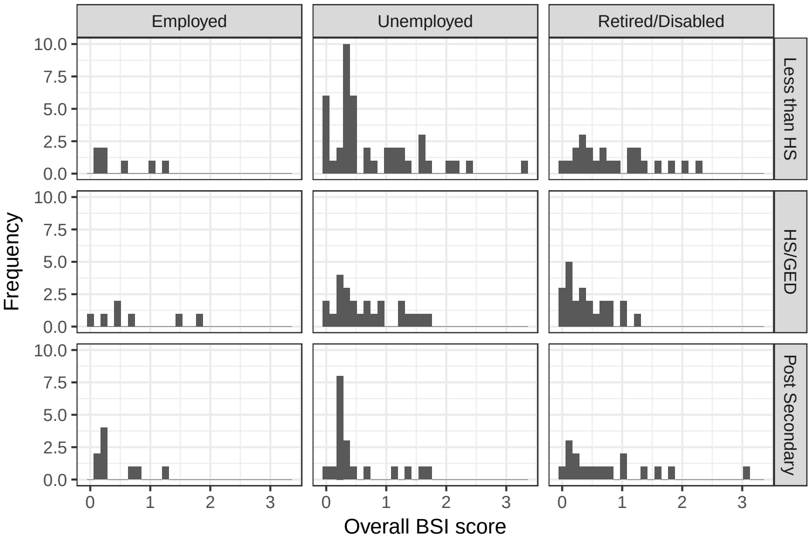


Figure 1.20: Scatterplot of CESD as a function of age, with a third variable added using a) color, b) shape, c) size, and d) shading



A *scatterplot matrix* is a common tool to quickly examine the bivariate relationship between multiple continuous variables simultaneously. Figure 1.21 demonstrates a publication ready version of a scatterplot matrix that has many features added including the pairwise correlation, univariate histograms and lowess lines on the scatterplots. This single plot lets us identify characteristics of the data such as `parent_bonding` is skewed right, `AGESMOKE` and `AGEALC` has a high frequency of 0's, there is a moderate positive correlation between `AGESMOKE` and `AGEALC`, and that none of the three covariates look to be significantly associated with `log_bsi_overall`.

1.5 Discussion of computer programs

All of these graphics were produced in R.

- R: ggplot2 can handle most anything you throw at it. Base graphics is fully customizable down to the last pixel.
- SAS/GRAPH: GCHART, GPLOT
- STATA:
- SPSS:

1.6 What to watch out for

We advise against changing too many features of a graphic at once. It can confuse the reader instead of provide understanding. Remember, the purpose of graphics are to understand distributional patterns, and to identify odd data

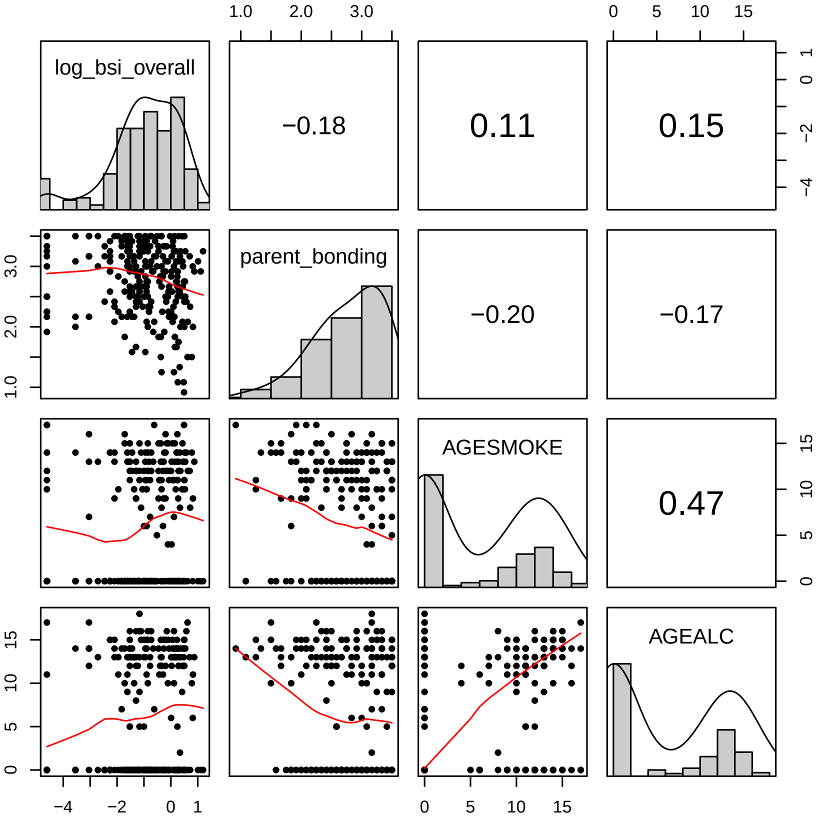


Figure 1.21: Scatterplot matrix with histograms and density plots along the diagonal, and pairwise correlation values above the diagonal

points. Not all layers provide illumination - take for example Figure 1.20 c); there is so much over-plotting in the lower left that it is impossible to see if there is a pattern emerging. In this case coloring by income is more helpful than changing the size, albeit barely.

All plots in this textbook are shaded using a grayscale, this is a necessary adjustment for black and white printing, but also is a consideration for colorblind readers. If you are able to publish color graphics, we recommend using a colorblind friendly color pallet.

We do not demonstrate plots such as 3D pie charts, or 3D bar charts in this text. In these cases the third dimension is false, it is considered “chart junk” and can be very misleading.

Be mindful of the scaling of your vertical axis on all plots. For example, Figure 1.3 plots a percentage on the y-axis with a high value of near 50%. We scale the y-axis to 100% here to put the difference in percentages in context of the overall range. A 2% point difference between categories can appear huge if the y-axis only has a total range of say 5%. Altering the y-axis to be too large or too small relative to the data is one of the most common ways graphics can be misleading.

If you are looking to publish your graphics be sure to check the rules carefully. Some publications have rules regarding features such as whether or not there is a box completely around the plot versus only showing the x and y axes.

Be consistent with your plotting themes. If you use a clear background and a box around the plot for one plot, you should apply that same theme to all plots. If you change the color of the points by a categorical variable for one plot, then all subsequent plots that also use that same categorical variable should have the same color scheme applied.

1.7 Summary

We have demonstrated a wide variety of plots in this chapter. Some plots are more appropriate for a report where you have more page space to write an explanation of the plot, others are better suited for a manuscript. Just as you have to consider your audience to determine the level of complexity of an analysis, so do you have to assess what visualizations are common in your field. For example, Violin plots are not commonly used, but are close enough to a histogram that it would be simple to explain how to read the plot in a manuscript. Pie charts have such a low ink-to-information ratio that it would be better to show a table or write the percentages into the text of a manuscript rather than taking up valuable figure space.

There are many other types of graphics that we do not discuss directly such as heatmaps, dendograms, geographic maps. These are typically considered specialized graphics for specific analyses. We present some of these 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 work as he is a current leader in the data visualization field for more ideas and guidelines to create informative graphics.

1.8 Problems

1.1 draw a plot?

1.2 draw a different plot?

1.3 This list is auto-numbered with the chapter number included.