3

Visualization

CONTENTS

3.1 Introduction

Statistical graphs can offer very effective means for formally presenting the findings from a data analysis or simulation study. They are also an essential part of exploratory data analysis as we saw in Section 2.7. In this chapter, we consider the question of how to visually present data in informative and effective ways. We begin with an overview of the basic types of statistical graphs and how to select the appropriate graph given the type of variable(s) we are analyzing (Section 3.2). We also provide a framework and a set of guidelines for making effective plots (Section 3.3); these include considerations for making the data stand out in the visual presentation, ways to facilitate important comparisons

between groups or against a benchmark, and approaches for augmenting a visualization to create a context for interpreting the visual display. We demonstrate these concepts by providing several example visualizations, including addressing the issue of how to create visualizations for large data. These examples are presented without the code we used to create them. Later in Section 3.5, we introduce the two primary models in *R* for creating graphs – the painter's model in base *R* and the grammar of graphics model in `ggplot2`.

Implicit in the guidelines for making statistical graphs is that we are attempting to convey a message in as clear and concise a manner as possible. This means that an important aspect of data visualization is discerning the message that we have discovered in our data and selecting an effective graph for conveying that message. Before we begin with an overview of plot types, we demonstrate the process of creating a visualization that best represents the message in the data.

3.1.1 Composing a Graph of Voter Registration Trends

There is no shortage of poorly designed statistical graphs. We use one as an example here because we can clearly separate the process of improving a graph by following the basic principles of plotting from the more nuanced considerations of how to best convey the underlying message in our data. These data are from online voter registration summaries published by the California Secretary of State at `http://www.sos.ca.gov/elections/voter-registration/voter-registration-statistics`. Figure 3.1 shows a bar plot created from data available at this site. A version of this plot first appeared online at `swivel.com`, which is no longer an active site.

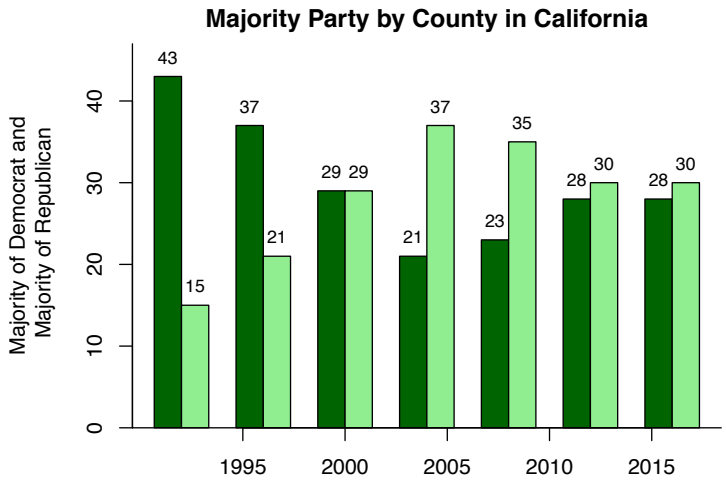

Figure 3.1: Distribution of California Voters by County. *The bar chart shown here imitates one that appeared on `swivel.com`. It has many flaws. Figure 3.2 fixes these problems, but does not address the question of how well the plot conveys the story in the data. Figure 3.3 replaces the bar chart with a more informative line plot that better conveys the message about changes in voter registration from 1992 to 2016.*

There are many obvious stylistic problems with this plot, e.g.,

- Tick marks: X-axis tick marks are at 5-year intervals and do not line up with the

locations of the bars so the viewer has to work too hard to figure out that the bars correspond to measurements made at 4-year intervals.

- Color: Atypical use of light and dark green for the Democratic and Republican parties (traditionally represented with blue and red, respectively).
- Legend: The lack of a legend means that we cannot discern which color represents which party.
- Axis Label: Y-axis label does not indicate what are the units of measurement.
- Title: Confusing title does not illuminate the content of the plot, i.e., what is meant by 'Majority Party by County in California'?

The numbers on top of each bar help elucidate what are the data. We see that the heights of each pair of light and dark green bars sum to 58, which we determine (with an Internet search) to be the number of counties in California. Furthermore, we can visit the voter registration page where the data come from to determine which colors represent the Democratic and Republican parties. With this additional information in hand, we can address all of the concerns listed above. The revised bar chart appears in Figure 3.2. In this plot, we have located the year labels (1992, 1996, etc.) below each pair of bars, used the traditional red and blue for the 2 parties, added a legend to associate color with party, replaced the y-axis label with one that specifies the units of measurement, added an x-axis label to indicate that the years are presidential election years, and modified the title slightly (with all of the other changes the title now seems adequately informative).

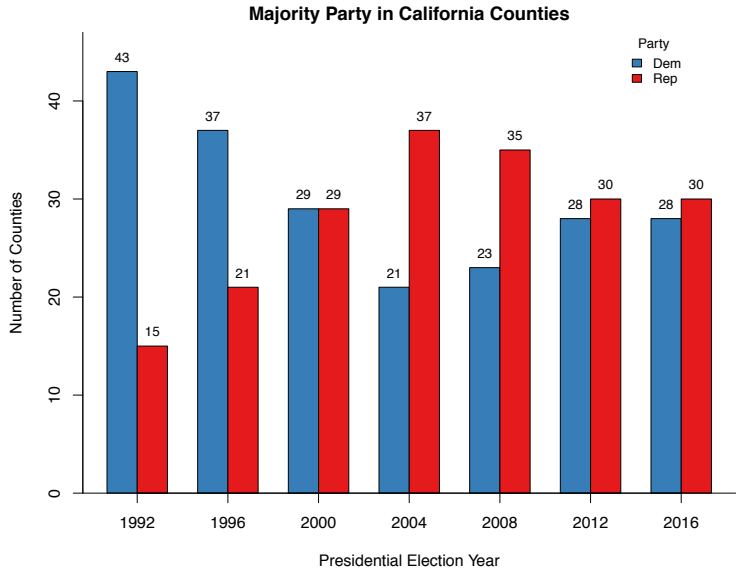

Figure 3.2: Bar Chart of Majority Party in California Counties – Revisited. *This bar chart addresses many of the problems found with the bar chart in Figure 3.1, including the inaccurate y-axis label, ill-positioned tick marks on the x-axis, poor choice of colors, and lack of legend.*

Before we declare Figure 3.2 a success, let's ask ourselves what message was the creator of this plot trying to convey and is this the appropriate plot for doing so? It seems that

the graph is trying to show the change in voter registration over the past 7 presidential elections. However, it's people who register to vote, not counties. County size is a lurking variable–small counties tend to be rural and conservative–so counting counties overstates the Republican presence. Rather than count counties, let's make a plot that tallies voter registration. To do this, we revisit the registration Web site to obtain these figures. There we find that voters can decline to affiliate with a party and there are several other parties with which a voter can register. To effectively observe the trends in registration, we need to include these other possibilities in our plot because they are not insignificant. What kind of plot should we make? We have registration figures over time so a line plot seems appropriate. Also, given that the California population has grown dramatically in the past 25 years, rather than compare raw registration numbers, we scale them by each year's total registration and compare percentages.

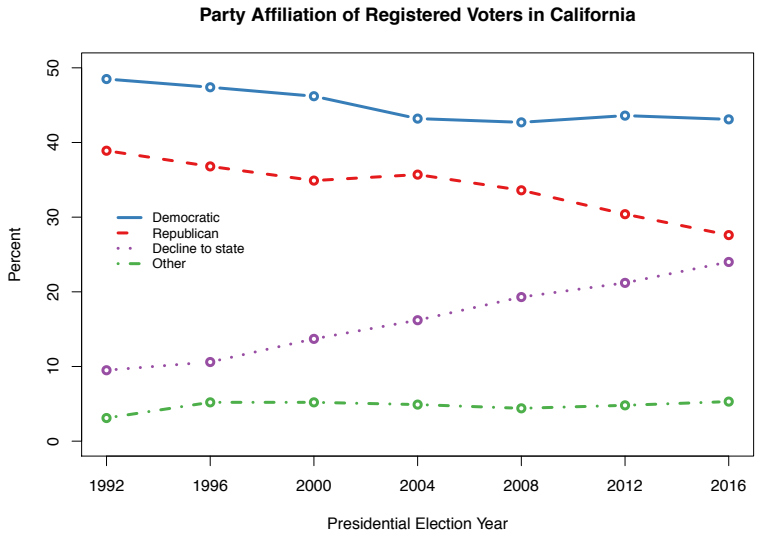

Figure 3.3: Distribution of California Voters by Party. *The line chart addresses the essential problem with the bar chart in Figure 3.2, i.e., we are interested in the change in voter registration over the years, not in the number of counties that are majority Republican or Democratic. Here we see that the percentage of registered Democrats and Republicans have declined over this 25-year period, the percentage of unaffiliated voters has dramatically increased, and that the gap between Democrats and Republicans has grown from about 10% to 15%.*

We have entirely overhauled the plot (see Figure 3.3). From this new graph, we get a more interesting and accurate depiction of the voter registration trends in California. We see that: the percentage of registered Democrats and Republicans have declined over this 25-year period; the percentage of Democrats was about 10% higher than the Republicans in the earlier years but the spread has grown recently to about 15%; and the percentage of unaffiliated voters has dramatically increased from about 10% to about 25% over this period. This aspect of making meaningful statistical graphs that accurately convey the story in the data follows from experience and experimentation, in addition to abiding by graphics guidelines (Section 3.3).

3.2 Data Types and Plot Choice

Chapter 2 introduced many of the basic plots within the context of cleaning and formatting data and carrying out exploratory data analysis. There we connected the choice of plot to the data type, and discussed how to read and interpret the various kinds of plots. In this section, we provide a brief overview of these basic plots. We do this in the context of one set of data that includes a range of variables and data types. These data are described in Section 3.2.2. We conclude this section by examining a second set of data with 100s of thousands of records, in order to address some of the considerations that arise when creating visualizations of relatively large amounts of data (Section 3.2.6). We begin with an introduction of the terminology we use to reference various components of a plot.

3.2.1 Terminology

The common terms we use to describe various pieces in a plot are shown in Figure 3.4. They include: the plot title, x and y axes and their respective labels, tick marks, and tick mark labels; within the plotting region, plotting symbols, reference lines, and labels on these lines and symbols; and a legend with its title, keys, and labels.

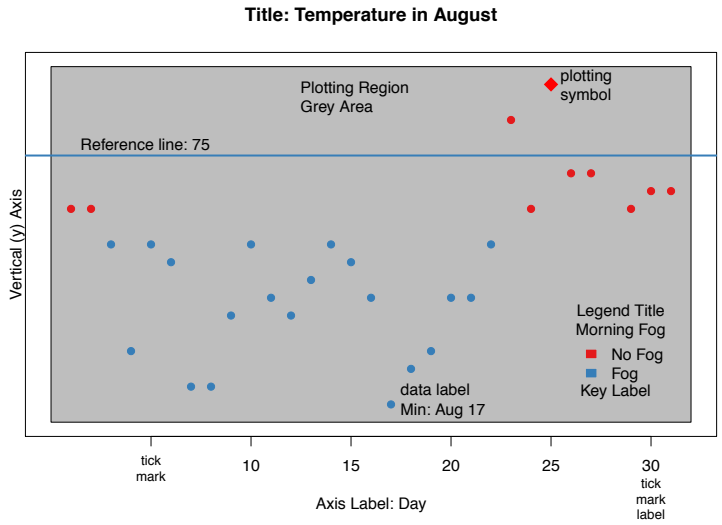

Figure 3.4: Graph Terminology. *This annotated plot provides a reference for the terminology we use in describing and critiquing plots.*

3.2.2 The Kaiser Family Data

The Child Health and Development Studies (CHDS) is a comprehensive investigation of all pregnancies that occurred between 1960 and 1967 among women who received prenatal care in the Kaiser Foundation Health Plan in the San Francisco–East Bay area and delivered at any one of the Kaiser hospitals in northern California. Over 15,000 families participated in the CHDS. The babies and their parents were followed through adolescence. The study had

TABLE 3.1: Infant Health Data Dictionary

Variable	**Definition**
id	identification number
date	birth date where 1096 = January1,1961
gestation	length of gestation in days
wt	birth weight in ounces (999 unknown)
parity	total number of previous pregnancies including fetal deaths and still births, 99 = unknown
race	mother's race 0-5 = white, 6 = mexican, 7 = black, 8 = asian, 9 = mixed, 99 = unknown
age	mother's age in years at termination of pregnancy, 99 = unknown
ed	mother's education 0 = less than 8th grade, 1 = 8th-12th grade – did not graduate; 2 = HS graduate-no other schooling; 3 = HS + trade; 4 = HS + some college; 5 = college graduate; 6 & 7 = trade school HS unclear; 9 = unknown
ht	mother's height in inches to the last completed inch, 99 = unknown
wt	mother prepregnancy wt in pounds, 999 = unknown
drace	father's race, coding same as mother's race.
dage	father's age, coding same as mother's age.
ded	father's education, coding same as mother's education.
dht	father's height, coding same as for mother's height
dwt	father's weight coding same as for mother's weight
marital	1 = married, 2 = legally separated, 3 = divorced, 4 = widowed, 5 = never married
inc	family yearly income in $2500 increments 0 = under 2500, 1 = 2500-4999, ..., 8 = 12500-14999, 9 = 15000+, 98 = unknown, 99 = not asked
smoke	mother's smoking status: 0 = never, 1 = smokes now, 2 = until current pregnancy, 3 = once did, not now, 9 = unknown
time	If mother quit, how long ago? 0 = never smoked, 1 = still smokes, 2 = during current pregnancy, 3 = within 1 year, 4 = 1 to 2, 5 = 2 to 3, 6 = 3 to 4, 7 = 5 to 9, 8 = 10+ years ago, 9 = quit and don't know, 98 = unknown, 99 = not asked
number	number of cigarettes smoked per day for past and current smokers 0 = never, 1 = 1-4, 2 = 5-9, 3 = 10-14, 4 = 15-19, 5 = 20-29, 6 = 30-39, 7 = 40-60, 8 = 60+ cigarettes, 9 = smoke but don't know, 98 = unknown, 99 = not asked

several goals, one of which was to examine the effect of the mother smoking during pregnancy on the baby. The `babies` data frame provides a subset of this information collected for 1236 babies–baby boys born during one year of the study who lived at least 28 days and were single births (i.e., not one of a twin or triplet). The information available for each baby, including how the variables are coded, is provided in Table 3.1.

For background, the gestation period is reported in days and the typical gestation is 40 weeks, or 280 days. In the CHDS, the start date of the pregnancy is self-reported by the mother. Additionally, the father's height, weight, smoking status, and education are reported by the mother. At the time of the study, little was known about the adverse effects

of smoking on health. For example, the link between smoking and lung cancer was first reported by the Surgeon General in 1964, the first warnings appeared on cigarette packages in 1965, and the effects of smoking on the unborn had not been widely studied.

We have prepared these data for analysis by, e.g., converting values of 99 and 999 into `NA`s, formatting variables such as smoking status and education as factors, and collapsing levels with only a few observations into other levels. The cleaning and formatting process we carried out is outlined in the exercises of Chapter 2.

3.2.3 Univariate Plots

We described in Chapter 2 that when selecting a plot to create a visualization, we need to consider the data type. In particular, we determine whether or not the variable represents a quantitative or qualitative measurement. Although there are exceptions, the data type tends to dictate the kinds of plots most appropriate for the data values. We use plots to visualize the distribution of observations across the variable's values.

Quantitative Variables
With quantitative data, we want to know about the basic locations and number of high-density regions (i.e., those regions where a large fraction of the observations crowd together), whether or not there are a few observations with unusually large or small values, gaps where no data are observed, the size of tails, and the symmetry or skewness of the distribution. On the other hand, with qualitative variables, we tend to simply summarize the proportion (or counts) of observations that take on each possible category.

Take for example the baby's birth weight (measured in ounces) in the Kaiser study. Figure 3.5 shows 4 different statistical graphs of this variable, including (clockwise from top left) a rug plot, histogram, density curve, and normal-quantile plot. We describe each of these in turn.

Rug Plot
With only a few observations, the rug plot provides a simple representation of the distribution of a quantitative variable. In a rug plot, each observation is marked by a tick along the x-axis, i.e., 'a yarn in the rug'. With more than a handful of observations, the rug plot is typically not adequate for conveying the distribution because there's too much over plotting (ticks plotted on top of one another) or just too many ticks marks to see the shape of the distribution, i.e., we have trouble distinguishing between high and low density regions. There are more than 1200 babies in our data frame, and the rug plot in Figure 3.5 reveals little about the distribution of birth weight. Instead, we want to create a more informative representation of the distribution with a histogram or density curve.

Histogram
To make a histogram of the birth weights, we divide the x-axis into intervals that span the range of the data values. For each interval (or bin), we find the percentage of observations with values that fall into this bin and create a bar over the interval with an area that equals this percentage. Essentially, we are smoothing the observations evenly across each bin, i.e., we don't know the exact location of observations in a bin. The histogram of birth weight shows a unimodal distribution (one high-density region) that is centered at about 120 ounces. There do not appear to be any gaps or unusually large or small data values, the distribution looks roughly symmetric, and the tails appear neither long nor short.

The intervals in the birth weight histogram in Figure 3.5 are each 5 ounces wide. However, the intervals in a histogram need not be all the same width. As mentioned already, the defining property of the histogram is that the area of a bar equals the proportion of observations in the corresponding bin. Note that we can recover this percentage/proportion

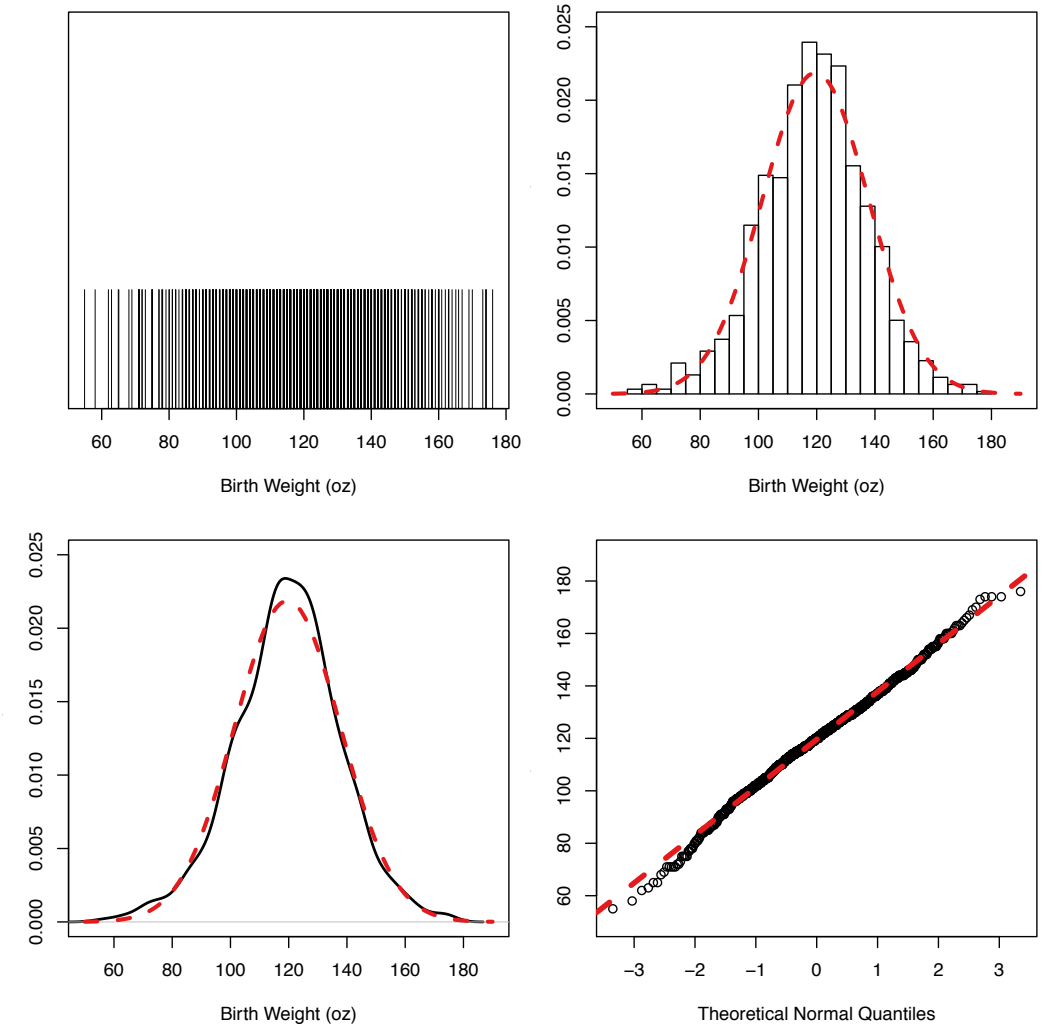

Figure 3.5: Distribution of Birth Weight. *These 4 plots show the distribution of birth weight for a subset of the babies in the CHDS. These are (top left to bottom right) a rug plot, histogram, density curve, and normal-quantile plot. The rug plot shows values of individual observations but with so many records it's difficult to see the general shape of the distribution. The histogram and density curve reveal a similar shape–a unimodal, symmetric distribution. A normal curve is overlaid on the histogram and density curve to show that the distribution is roughly normal. However, the normal quantile plot offers a better visual comparison to the normal. Generally, the data quantiles and normal quantiles follow a straight line which indicates the data closely follow the normal curve. The downward curve on the left indicates the empirical distribution has a slightly longer left tail than the normal.*

by multiplying the height of the bar by the width of the bar. In other words, the units of the bar's height is a density, such as percent per ounce.

Figure 3.6 provides another example of a histogram for the parity of the pregnancy, i.e., the number of previous pregnancies. Here 0 corresponds to the woman's first pregnancy, 1 to her second, etc. We describe the distribution of parity as unimodal with a peak at 0, skewed right (there is more area to the right of the mode than the left), a long right tail with some mothers having parity of 6 or more, and short left tail since it's not feasible to have values below 0. Notice that in this histogram, most bins are 1 unit wide, but the 3 rightmost bins are wider. They are 2, 3, and 3 units wide, respectively. We often use wider bins in the tails of a distribution to further smooth the data. The mother's `parity` is a discrete quantitative variable because only integer values are possible, i.e.,

```
table(babies$parity)
```

```
  0   1   2   3   4   5   6   7   8   9  10  11  13
315 310 238 168  83  52  32  16   8   7   4   2   1
```

We have used a bin that combines the counts for 11, 12, and 13 in the histogram because the proportions for these values are low and we want to smooth them out over the distribution's tail.

Additionally, to make it clear that `parity` can take only nonnegative integer values, we center the bins on the integers, e.g., the bin from 1.5 to 2.5 contains those mothers with a parity of 2. For baby's birth weight, since the measurements are to the nearest ounce, we need to know the interval convention, i.e., whether the bin from 120 to 125 is open on the left and closed on the right or closed on the left and open on the right, inorder to determine whether the bin includes the babies with a birthweight of 120 or 125. This distinction is less important because these measurements ideally can be measured to a finer precision.

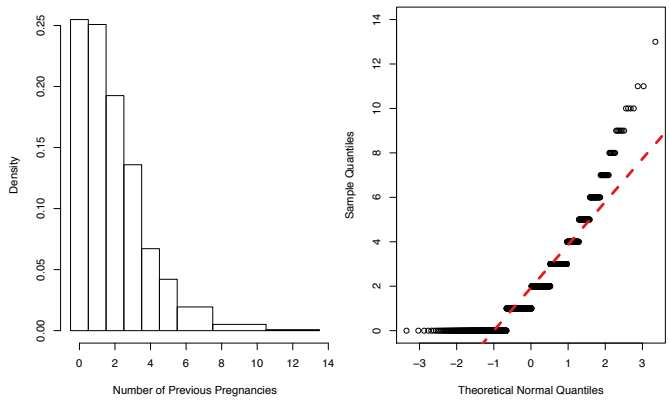

Figure 3.6: Histogram of Parity of the Pregnancy. *Parity has a discrete distribution (only integer values are possible). The histogram (left) reveals the distribution is unimodal, skew right, with a long right tail where a few women had 6 to 13 previous pregnancies. The steps in the normal-quantile plot (right) are due to the discreteness in the distribution. The very long initial step indicates the empirical distribution has no left tail and the curvature on the right indicates a long right tail.*

Density Curve

The density curve provides a smooth representation of a distribution. With the density

curve, the area under the curve for a particular interval approximates the proportion of values in that interval. (The total area under the curve is 1.) The density curve for birth weight (bottom left in Figure 3.5) has a similar shape as the histogram in that figure. We again see a unimodal distribution that is centered at about 120 ounces, symmetric, with neither long or short tails. Note that we do not make a density curve for parity because of the discrete nature of the values for this variable.

Normal Curve

We overlaid a normal curve on both the histogram and density curve for birth weight in Figure 3.5. The normal curve is often used as an idealization of an empirical distribution. It can align well with the distribution of a quantitative characteristic of a population, such as the height of fathers, width of Dungeness crab shells, weight of locally grown tomatoes, and circumference of trees. When we match a normal curve to data, we choose from a family of similarly shaped curves, where each curve is characterized by its center (point of symmetry) and spread. The normal curve that is super-posed on the histogram and density curve in Figure 3.5 has a center that matches the average birth weight and a spread that matches the SD (standard deviation) of birth weight. (Recall that the SD is a measure of spread, which is defined as the square root of the average of the squared deviations from the observations to the average, i.e., in *R* it is

```
sqrt(mean( (bwt - mean(bwt))^2 ))
```

These 2 quantities, center and spread, completely determine a normal curve. Also, when we describe the length of a distribution's tails, we often compare them to the tails of a normal curve.

Normal-Quantile Plot

The normal curve that is layered over the histogram and density curves in Figure 3.5 appears to follow the histogram and density curve quite closely. However, a normal-quantile plot offers a better visual comparison (lower right, Figure 3.5). In this plot, we see that the points roughly follow a line, which indicates that the data roughly follow the normal curve. Generally, a normal-quantile plot compares the quantiles of the data to those of the normal curve. That is, the scatter plot consists of quantile pairs: (qth quantile of the idealized normal curve, qth quantile of the data), for many qs between 0 and 1. If the data's distribution is close to the normal curve, then the points roughly fall on a line. Deviations from a line indicate differences between the distributions. We have added a line to the plot with intercept and slope that match the mean and SD of our data because the normal quantiles used in the plot are for the standard normal (mean 0 and SD 1).

As another example, we make a normal-quantile plot for parity (on the right in Figure 3.6). Most noticeable are the steps in the plot. These are due to parity taking only discrete values. The birth weight measurements are discrete in the sense that weight is measured to the nearest ounce, but the steps from the duplicate values are not noticeable given the range of data values. We saw in the histogram that this distribution is clearly not normal because it is skewed right and has a long right and short left tail. These features appear in a normal quantile plot via curvature; specifically, the short left tail is indicated by the long step at the lower left, and the long right tail is evident from the upward turn on the right end of the curve.

More generally, we can make a quantile-quantile plot that compares the quantiles of any two distributions, including comparing observed quantiles from two sets of data (see Figure 3.10) and comparing the empirical quantiles from data to the quantiles of a theoretical distribution other than the normal. See the exercises for more practice with reading quantile plots.

Qualitative Variables
A qualitative variable takes on a fixed and finite set of possible values, where each value corresponds to a category. For example, in the Kaiser study, mother's smoking status can belong to one of 5 possible categories : she never smoked, smoked during pregnancy, smoked until she was pregnant, smoked once but not now, or has an unknown smoking status. We are interested in the distribution of smoking status, i.e., the proportions in each category. (We exclude those mothers with an unknown smoking status). We can compute these proportions with

```
table(babies$smoke) / sum(table(babies$smoke))
```

```
 Never Current  Until   Once
 0.444   0.395  0.077  0.084
```

These proportions can be arranged in several different visual formats, such as a bar plot, dot chart, and pie chart, as shown in Figure 3.7.

Bar Plot
Unlike a histogram, the area of the bar in a bar plot has no meaning–only the height of the bar conveys the distributional information. For this reason, the 2 bar plots in the top row of Figure 3.7 are equivalent. In both we see that there are nearly equal numbers of mothers who never smoked and those who smoked during pregnancy. We also see that less than 10% of the mothers quit smoking when they became pregnant and about the same proportion quit smoking before becoming pregnant.

Dot Chart
The dot chart (bottom right of Figure 3.7) takes the lack of meaning in the width of a bar in a bar chart to its extreme conclusion and eliminates the bars entirely. In a dot chart, the proportion of mothers in each category are located along the category's respective line. Again, we see clearly that the smokers make up slightly less than 40% of the mothers and the never smokers about 44%. The bars in a bar chart and dots in a dot chart can be arranged vertically or horizontally. The dot chart here is arranged horizontally while the bars are arranged vertically in the 2 bar plots.

Pie Chart
The pie chart (bottom left of Figure 3.7) conveys these proportions through angles in the pieces of pie. For example, the slice of pie corresponding to the `Never` group makes up 44% of total pie. Comparisons between slices in a pie can be difficult to make, e.g., it's difficult to discern that the `Current` slice of pie is smaller than the `Never` piece. In general, we can more accurately compare lengths of bars and locations of dots on a line than angles in slices of pies so bar charts and dot charts are typically preferred over pie charts.

3.2.4 Bivariate Plots

When we have more than one variable, we typically want to examine the relationship between variables. For example, we may want to observe the relationship between the heights of mothers and fathers to see if, for example, mothers who are above average in height tend to have partners who are are above average in height.

We may also want to examine the relationship between 2 categorical variables to see, for example, if the distribution of never and current smokers is the same across different levels of education. In general, more highly educated mothers tend to live healthier lives which can impact the health of the new born and we want to see if this relationship is born out in the data.

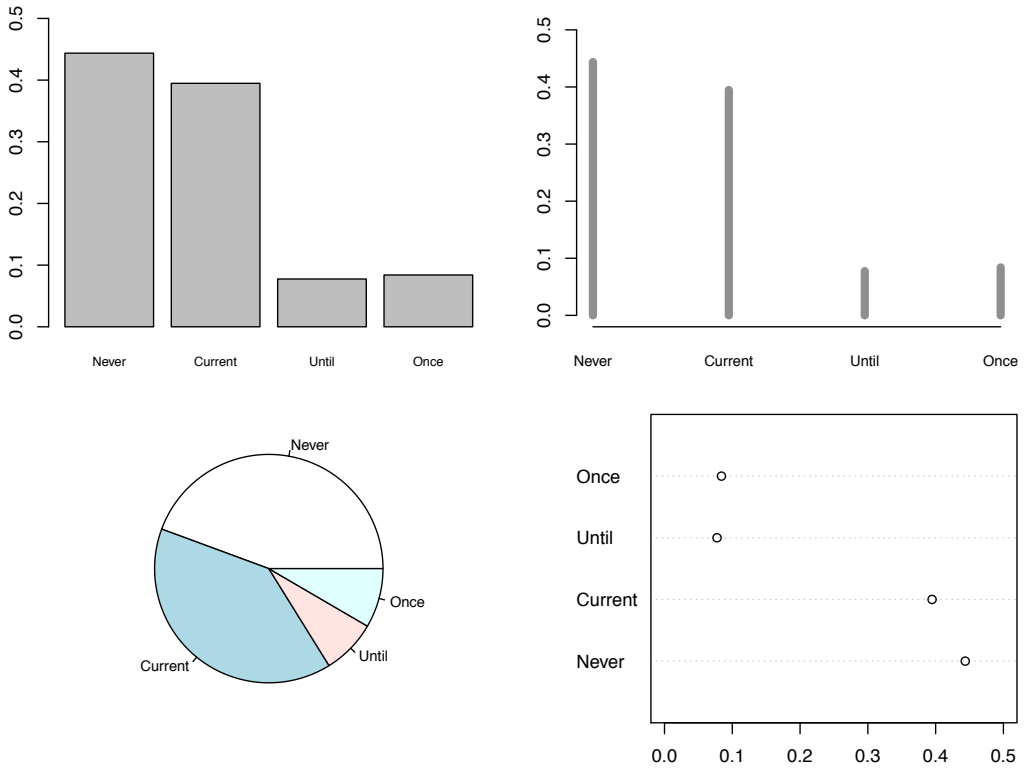

Figure 3.7: Distribution of Mother's Smoking Status. *These 4 plots display the proportion of mothers in each smoking status via (top left to bottom right) bar charts with wide and narrow bars, pie chart, and dot chart. The labels denote whether the mother* Never *smoked,* Current*ly smokes (in pregnancy), smoked* Until *she became pregnant, or smoked* Once *and quit before pregnancy. The bar and dot charts use length to represent the proportion of each type of smoker. The pie chart uses angles which are typically harder to compare accurately.*

A core interest of ours is in the relationship between the mother's smoking status and her baby's birth weight; that is, we want to examine the relationship between a qualitative and quantitative variable. To do this, we can compare the distribution of birth weight for the group of smokers and never smokers to see if these distributions are the same. If not, we want to know how they differ. We provide examples of bivariate plots that can begin to address these various situations.

Scatter Plot

Mother's and father's height are both quantitative variables. With quantitative variables, we typically use a scatter plot to examine their relationship. For the scatter plot in Figure 3.8, each point corresponds to a (mother, father) pair. The scatter of points display a weak positive association; in other words, mother's above average in height tend to be associate with father's who are also above average in height, but there are many exceptions to this 'rule'. We can compute the linear correlation between these variables with

```
cor(babies$ht, babies$dht, use = "complete.obs")
```

```
[1] 0.34
```

Correlations range between -1 and +1 with values of ± 1 indicating a perfect linear relationship and a value of 0 indicating the lack of a linear relationship. The correlation of 0.34 is weak, but it does indicate that there is a small positive association. However, since the correlation is small, there is a great deal of variability in the father's height among the mothers with the same height. We also note that a correlation coefficients can be strong even when the relationship beween 2 variables is clearly not linear so it is good practice to plot the data in order to view the shape of the relationship.

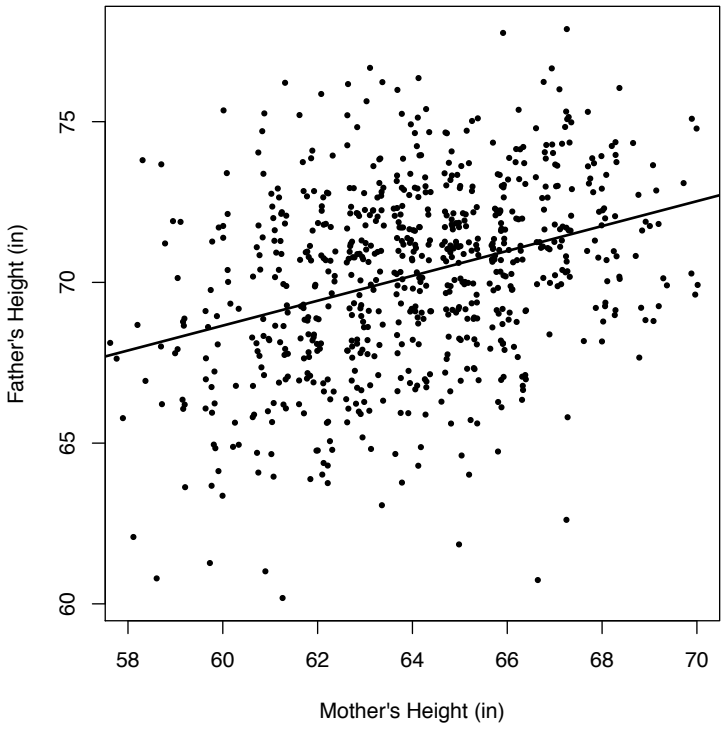

Figure 3.8: Heights of Mothers and Fathers. *This scatter plot shows a weak linear association between the heights of mothers and fathers in the Kaiser study. The line fitted to these points has a slope of about 0.39 and an intercept of about 45. According to this line, a mother who is 64 inches tall tends to be with a father 70.2 inches tall on average, and a mother 2 inches taller tends to be with a father 71 inches tall on average. The correlation is 0.34, indicating the variability of father's height for mothers who are, e.g., 64 or 66 inches tall is quite large.*

Two Qualitative Variables

When we examine the relationship between 2 qualitative variables, we examine proportions of one variable within the subgroups defined by the other variable. For example, with education and smoking status, we can compare the proportion of never, current, until pregnant, and once smokers within each education level with

```
        < 12th    HS Trade Some Col College
Never   0.345 0.459 0.371    0.466   0.498
Current 0.530 0.400 0.486    0.345   0.297
```

```
Until   0.065 0.091 0.043    0.071   0.082
Once    0.060 0.050 0.100    0.118   0.123
```

Alternatively, we can compare the proportion of some high school, high school, trade school, some college, and college educated mothers across each smoking status with

```
        < 12th     HS Trade Some Col College
Never    0.126  0.371 0.048    0.254   0.200
Current  0.219  0.364 0.070    0.211   0.135
Until    0.137  0.421 0.032    0.221   0.189
Once     0.117  0.214 0.068    0.340   0.262
```

Which of these tables of proportions is most helpful? The answer depends on what we consider to be the important comparison. If we condition on education level and examine the distribution of smoking status for each level of education, then we can compare education levels across smoking status. We associate education level with healthy life choices, e.g., alcohol consumption, so we want to see how education level varies across smoking status. This comparison corresponds to the proportions in the first table. In Figure 3.9, we have created (clockwise from top left) a mosaic plot, side-by-side bar chart, line plot, and stacked bar chart from this table.

Mosaic Plot and Stacked Bar Chart
The stacked bar chart is the most difficult to read because, for example, the top and bottom of each sub-rectangle that corresponds to a particular education level moves up and down across smoking status. The mosaic plot has a similar issue but it has several advantages over the stacked bar plot. The width of the vertical slices correspond to the proportion of smoking levels among the entire population under study; the gaps between the vertical and horizontal slices assist in the comparison; and the shading of a rectangle indicates whether the proportion is more (blue) or less (red) than expected under the assumption of there being no relationship between the two variables.

Side-by-Side Bar Chart and Line Plot
The line and side-by-side bar chart are very similar in how they convey information. In the side-by-side bar chart, the bars for education status are grouped together for each level of smoking. Also, since education is an ordered qualitative variable, the bars are arranged according to this order. The line plot connects dots for each education level across smoking status. The line plot has two main advantages over the bar chart: the dots for education level within each smoking status appear directly above/below one another, and the dots for the same education level are connected across smoking status. Both the these features make it easier for us to make comparisons. We can more easily see whether the ordering of education level remains the same across the smoking status and if the differences between education grow or shrink with smoking status.

One Qualitative and One Quantitative Variable
When we examine the relationship between one quantitative and one qualitative variable, we use the qualitative variable to divide the data into groups and compare the distribution of the quantitative variable across the groups. For example, we can compare the distribution of birth weight for babies born to mothers with different smoking statuses.

Quantile-Quantile Plot
We can compare 2 quantitative distributions with a quantile-quantile plot. That is, we can plot pairs of quantiles from 2 empirical distributions (this is in contrast to the normal-quantile plot where we plot the quantiles of an empirical distribution against those of the

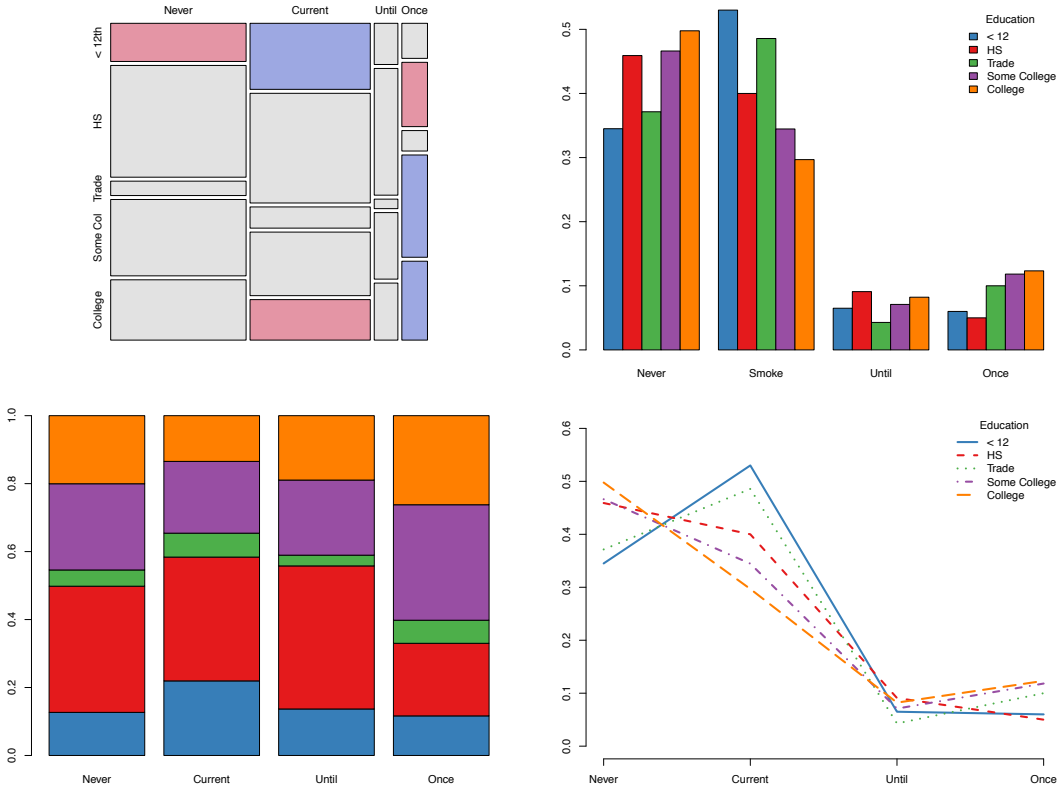

Figure 3.9: Mother's Smoking Status and Education. *The 4 statistical graphs in this figure offer alternative approaches to comparing education level for different types of smokers. These are (from top left to bottom right) the mosaic plot, side-by-side bar chart, stacked bar chart, and line plot. Each plot displays the proportion of mothers with a particular education level within smoking status, i.e., the proportions for high school education add to 1 across the 4 smoking statuses. The line plot has advantages over the bar chart due to the close proximity of education level values within a smoking status and to the connecting line segments that help compare education level across smoking status. The stacked bar plot is problematic because the rectangles for a particular education level are not aligned, i.e., both the lower and upper sides of the rectangles move up and down across smoking status.*

theoretical normal). For example, in Figure 3.10 (top left), we compare the distribution of birth weight for never smokers and current smokers. Here we see that the points roughly fall on a line, which indicates the distributions have roughly the same shape. However, this line is shifted down from the reference line that has intercept 0 and slope 1, which points to a shift left in the distribution of weight for the smokers' babies in comparison to the never smokers.

Super-posed Density Curves
Alternatively, we can super-pose (place on the same plot) the density curves for each subgroup (Figure 3.10, top right). There we confirm that the density curve for the smokers is shifted to the left of the curve for the never-smokers. It also appears that the spread in birth weight is larger for the smokers and that the never-smoker distribution has longer tails than the smoker group. How do these features appear in the quantile-quantile plot? Note the slope of the points and the curvature at the extremes. We see that the smaller spread for never-smokers is reflected in the slightly steeper line of points, and the unusually large and small values for the never-smokers appears in the curvature of the points at the two ends of the birth weight values. However, the shift seems the most pronounced difference between the never and current smokers.

Side-by-Side Box Plots and Violin Plots
Yet another approach compares summary statistics for the subgroups with side-by-side box plots. Recall that the box marks the lower and upper quartiles and the line in the interior of the box denotes the median. Whiskers are drawn to the nearest observation that is within 1.5 IQRs of the quartiles and points beyond that are marked individually. Figure 3.10 contains box plots of birth weight for all 4 smoking statuses. We see that the IQR for the never smokers is about 5 ounces smaller than the IQR for the current, until pregnant, and once smokers. Also evident are the large number of outliers in the never smoked group; there are many unusually small and large birth weights in this group. Alternatively, we can juxtapose 4 density curves in a violin plot. In the bottom right of Figure 3.10, the density curves for 4 subgroups of mothers are plotted vertically and reflected about their respective axes to create violin-shaped plots that are similar in purpose to box plots.

3.2.5 Conveying Relationships between 3 or More Variables

We can use the plots described already to create 2-dimensional visualizations for 3 (or more) variables. For example, if we have 3 or more qualitative variables, we can subdivide the data according to the combinations of levels of these variables and compare proportions with line plots, dot charts, side-by-side bar charts, and mosaic plots as in Figure 3.9. We typically organize the lines, dots, and bars into groups according to a combination of levels from two (or more) variables. We can examine the relationship between one quantitative variable and 2 or more qualitative variables with side-by-side box/violin plots or overlaid density curves. Again, the box/violin plots are organized according to the combination of categories of the qualitative variables. The specific organization depends on which comparison we want to focus on, similar to the decision made when we compared education levels across smoking status (Figure 3.9). With 2 quantitative variables and 1 or more qualitative variables, we often make scatter plots where the points are color coded according to the categories of the qualitative variable or where the shape of the plotting symbol denotes the category. Additionally, we juxtapose scatter plots in a grid (keeping the range of the axes the same across the various plots).

Figure 3.11 is an example where we examine three variables: birth weight, mother's height, and smoking status. The scatter plot displays the relationship between the baby's birth weight and mother's height. As with the scatter plot of mother's and father's height,

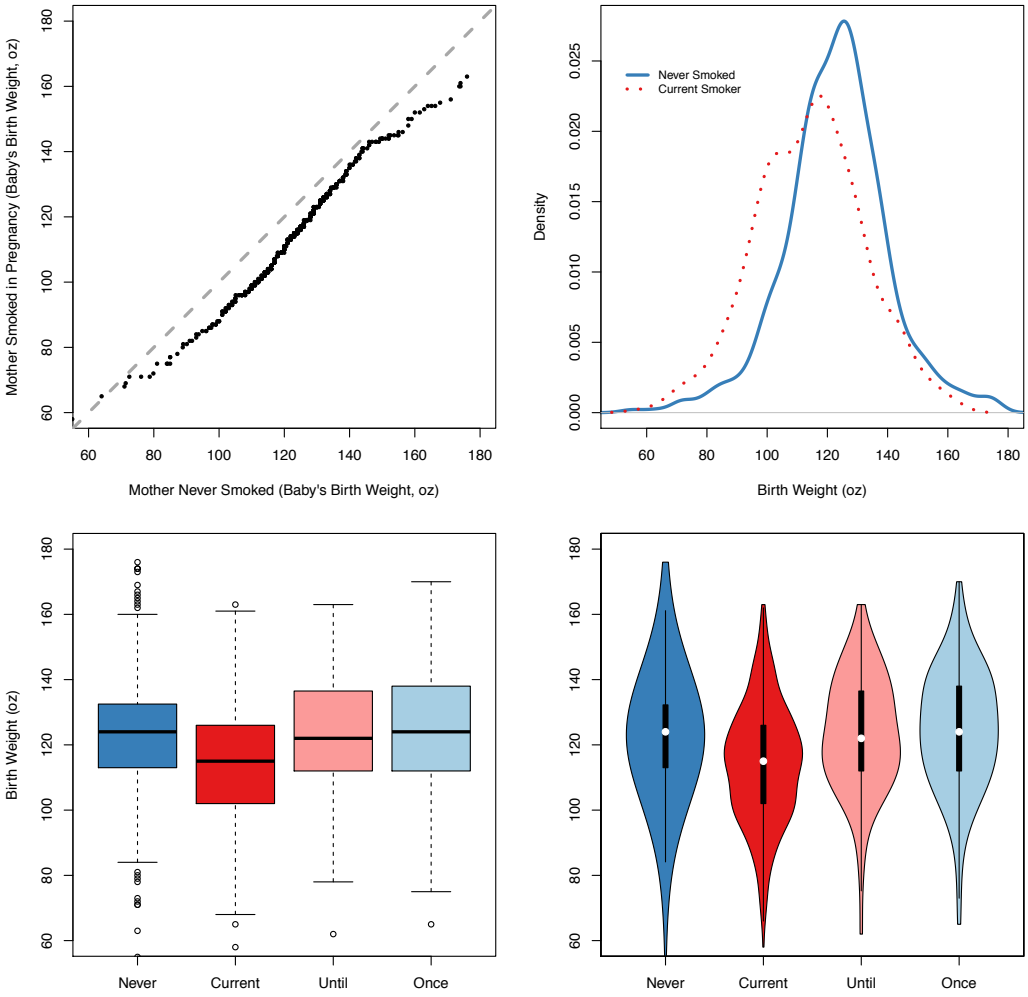

Figure 3.10: Mother's Smoking Status and Baby's Birth Weight. *These plots compare the distribution of birth weight for mothers with different smoking statuses. A comparison of the birth weight of babies born to mothers who never smoked versus those who smoked during their pregnancy is made with a quantile-quantile plot (top right) The quantiles roughly fall on a line which indicates similar distributional shape. The straight line added to this plot has slope 1 and intercept 0 and can be used to discern differences between the distributions of these two groups. Particularly, the downward shift of the points in comparison to this line indicates a shift to lower birth weights for smokers. The plot on the top right overlays (or super-poses) the density curves for these 2 groups and confirms the noted shift. The birth weight distribution for all 4 smoking statuses are provided in side-by-side box and violin plots (bottom left and right, respectively). It appears that those who quit smoking have a distribution with a median that matches the non-smokers but a slightly larger variability.*

we find a weak linear association. The points in the scatter plot have been color coded according to whether the mother never smoked (blue) or currently smokes (red). We have added two curves to this plot, the red dashed line for the smokers and the blue solid line for the never smokers. Fore each group of mothers (never or current smokers) the curve displays the average birth weight of the babies born to mothers of the same height. We see that the both curves are roughly linear, indicating the taller mothers tend to have heavier babies. We also see that the curves have roughly the same slope but curve for the smokers is about 10 ounces lower than for the never smokers.

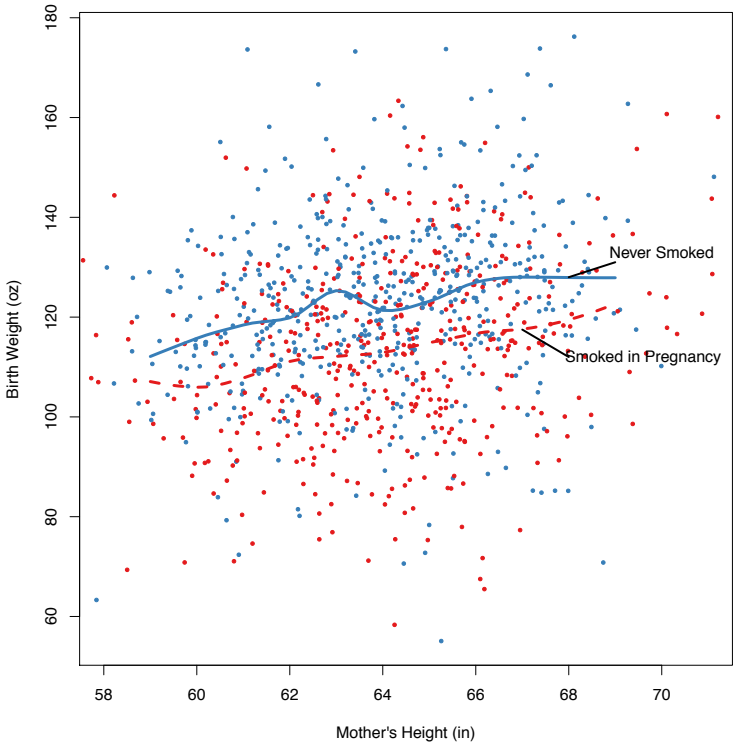

Figure 3.11: Baby's Birth Weight and Mother's Height. *Mother's height and baby's birth weight have a weak linear relationship with a correlation coefficient of about 0.20. The points in this scatter plot are color-coded according to the mother's smoking status (blue for never smoked and red for smoked during pregnancy). Two curves added to this plot show the average birth weight for babies born to mothers with the same height (red dashed is the average for mothers who smoked during pregnancy and solid blue for mothers who never smoked). We see that the relationship between birth weight and mother's height is roughly linear for both groups of mothers and babies born to never-smokers consistently weigh about 10 oz more than those born to current-smokers for all heights.*

Visualizations of 3 or more quantitative variables are more difficult to make. We often create scatter plots of all pairs of variables. However, plots of pairs of variables may not be adequate for revealing higher dimensional relationships. We can also create a scatter plot with 2 of the variables and color the points according to the level of the 3rd variable. Another technique, multi-dimensional scaling, creates a scatter plot that tries to maintain

TABLE 3.2: San Francisco Housing

Variable	**Definition**
county	County name
city	City name
zip	Zip code
street	Street address
price	Sale price in dollars
br	Number of bedrooms
lsqft	Size of lot (ft^2)
bsqft	Size of building (ft^2)
year	Year house was built
long	Longitude of house location
lat	Latitude of house location
wk	Week sale reported by SF Chronicle, e.g., "2003-04-21"

the distances between all pairs of observations, where the distance between two observations is computed with all of the variables. (See Section 13.9 for an example).

3.2.6 Scalability – Real Estate Data with 500,000 records

When we have a large number of observations, some of the plots we have described can be problematic for displaying the relationship between variables. As an example, we examine a catalog of houses sold between April 27 2003 and November 16 2008 in the San Francisco Bay Area. For each of the 500,000 sales, we have the sale date and price along with the house location, lot size, building size, number of bedrooms, and year it was built. See Table 3.2 for a description of the variables and their units of measurement. These data were scraped from the San Francisco Chronicle web site [2]. (See also [13] for an analysis of these data, [1] for an example of how to scrape these data from the Web, and the exercises in Chapter 2 for how to clean and format the `housing` data frame for analysis.)

Over plotting is the main problem with large amounts of quantitative data. For this reason, rug plots are not feasible. Instead, we create a histogram or density curve to view a variable's distribution. We can alleviate the problem with over plotting in a scatter plot in a variety of ways. We can: divide the data into sub groups and juxtapose scatter plots of these groups; examine a subset of the data where particular variables are held constant; or make plots of summaries or aggregates of the data, rather than individual points.

For example, we can select one year of data, say 2004, and a subset of cities, say those within 10 miles of Berkeley. With this reduced collection of sales, the data are likely to be more similar because housing prices don't change as much in 1 year in comparison to 8 years and the houses are in a smaller geographic area. Figure 3.12 shows side-by-side box plots of sale price in the 12 neighboring cities. Rather than arranging the box plots alphabetically according to city name, they have been arranged according to median sale price. Notice that we have also plotted price on a log scale in order to accomodate the tremendous range in prices. We might consider removing a few outliers so that we can zoom in on the bulk of the data. Nonetheless, it is clear that houses in Piedmont are clearly the most expensive (Piedmont is a small city surrounded that has many large, stately homes and schools with high test scores.)

With this same subset, we examine the relationship between the sale price and size of

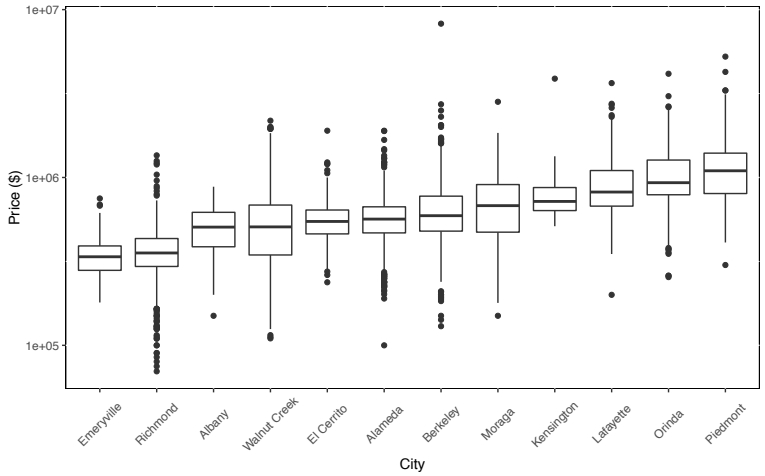

Figure 3.12: Housing Sale Prices by City. *These side-by-side box plots of sale price are arranged in order according to the median sale price in the city. Price is plotted on a log-scale. Included here are sales for 2004 from 12 cities within 10 miles of Berkeley, California.*

the house with a set of scatter plots. Rather than examine sale price directly, we compute the price per square foot of the building. We have juxtaposed 12 scatter plots, one for each city, in a 3 by 4 array in Figure 3.13. Notice that the range of the x- and y-axes remain the same across all 12 scatter plots to help us compare cities.

We also added a smooth curve that shows the relationship between price per square foot and building size. The curve averages price per square foot given building size for houses in all 12 cities (so the curve is identical on all 12 plots). This curve shows a general trend and helps compare cities across plots. The curvature indicates that smaller homes cost more per square foot than large ones, i.e., there is an entry cost to buying a home and the cost of additional space is cheaper.

Furthermore, we made the color of the points in the scatter plot partially transparent so that overplotting is ameliorated to some extent. The color corresponds to the number of bedrooms. It's evident that the 1 bedroom houses are more prevalent in the cities with lower sale prices.

For another approach, we can include all of the 2004 data for all 12 cities in one plot, as in Figure 3.14 (left plot). Now there are too many observations to color code the points by the number of bedrooms. Instead, we color code the points according to density. This way, we visualise the shape of the bivariate distribution of price per square foot and building size. That is, the deep red/purple area has the greatest density of sales; the peak is located at about 1250 square feet and $350 per square foot (which corresponds to prices from about $350,000 to $400,000).

To make yet another kind of summary, we can plot smooth curves rather than individual points. The plot on the right of Figure 3.14 shows no individual points, and instead displays curves of the average price per square foot as a function of building size. We have superposed 12 curves, one for each of our cities. From these curves, we see that each city follows this pattern of a high entry point and decreasing cost per square foot for larger homes. We also see that the curves for all of the cities are roughly the same shape and those for the more expensive cities are shifted higher than the cheaper cities. The lumpiness in a couple of these curves, e.g., Moraga, may be due to the number of houses sold in that size range.

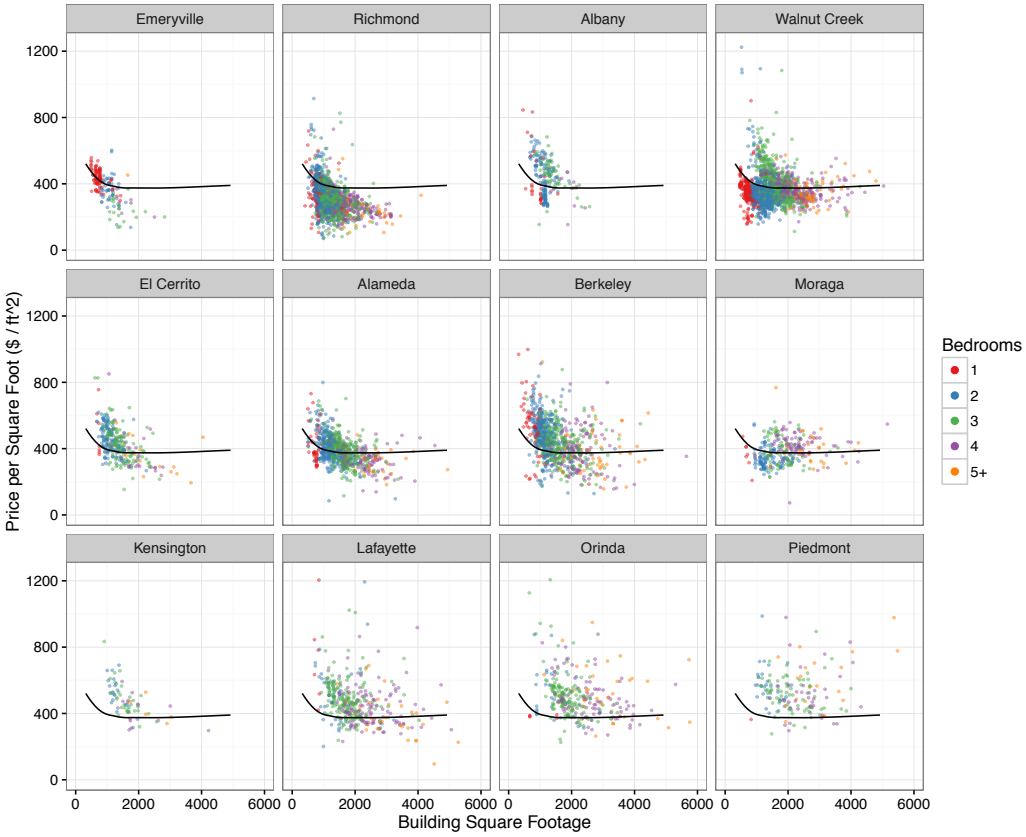

Figure 3.13: Sale Price and Building Size for 12 Cities. *Plotted here are the price per square foot against the size of building for all houses sold in 2004 in 12 cities within 10 miles of Berkeley, California. The data for each city appear in separate (juxtaposed) scatter plots, which are arranged from least to most expensive. The plots have common x-axis and y-axis scales across the 12 plots, and the curve of average price as a function of building size for all cities is super-posed on each plot. The points are color-coded according to the number of bedrooms in the house.*

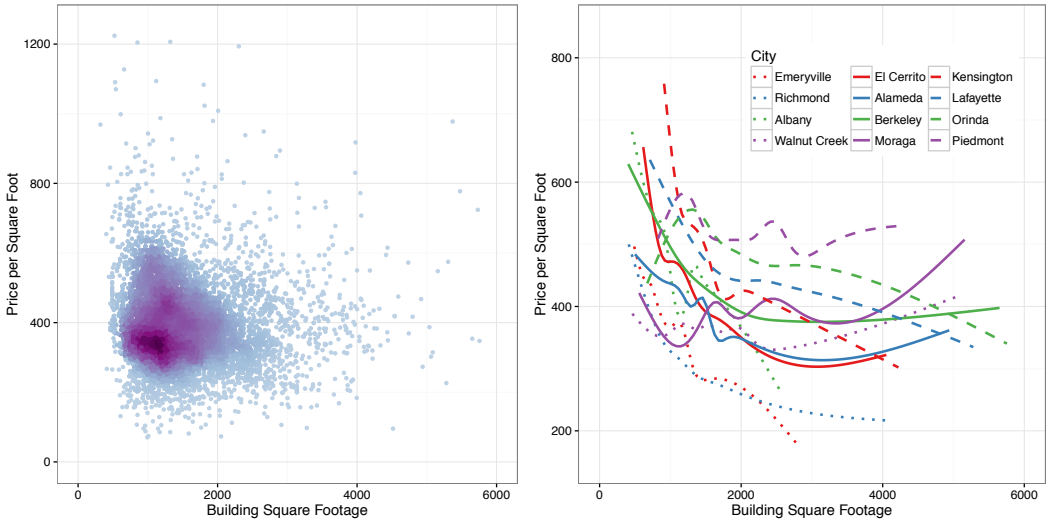

Figure 3.14: Sale Price and Building Size. *The scatter plot (left) shows the relationship between the price per square foot and size of building for all houses sold in 2004 in 12 cities in the East San Francisco Bay Area. The points are color-coded according to the density of the number of houses in each price-size combination. The smooth curves (right) show the average price per square foot as a function of the building size for these 12 cities. The curvature in the relationship between price per square foot and building size is similar across all 12 cities with the cities with higher median sale prices remaining consistently above the other cities.*

3.2.7 Measurements in Time

The original data set contains weekly sale prices for 6 years so we may want to observe trends in prices over time. When we include time as a variable in a plot, we typically place it along the x-axis and make a line plot that connects the y-values across time. In this case, we have multiple measurements for each time, i.e., all of the sales for the week, so we make a time plot with an aggregate of sale price for each week. Given the skewness of the distribution of sale price, we typically summarize sale price by a median rather than a mean. However, we can also examine other quantiles of the data. In Figure 3.15 (modeled after a figure in [13]), we make a line plot for each of the 9 weekly price deciles. That is, the 10th percentile, 20th percentile, ..., and 90th decile of sale price for each week's sales. Housing prices were seen to fluctuate tremendously over this short time period as the housing market boomed and crashed. One way to explore the impact on the different priced homes is to normalize the weekly deciles by their value at the start of our window of observation. In other words, we divide the weekly 10th percentile of sale prices by the 10th percentile on Apr 27, 2003, and similarly scale the other percentiles. These normalized percentile line plots appear in Figure 3.15. They are color-coded with a grey-scale that ranges from dark grey for the 10th percentile to light grey for the 90th percentile. It's evident from this plot that the lower priced homes suffered the greatest volatility in this period. They proportionally surpassed the higher-priced homes in 2006 and lost the greatest value in 2008.

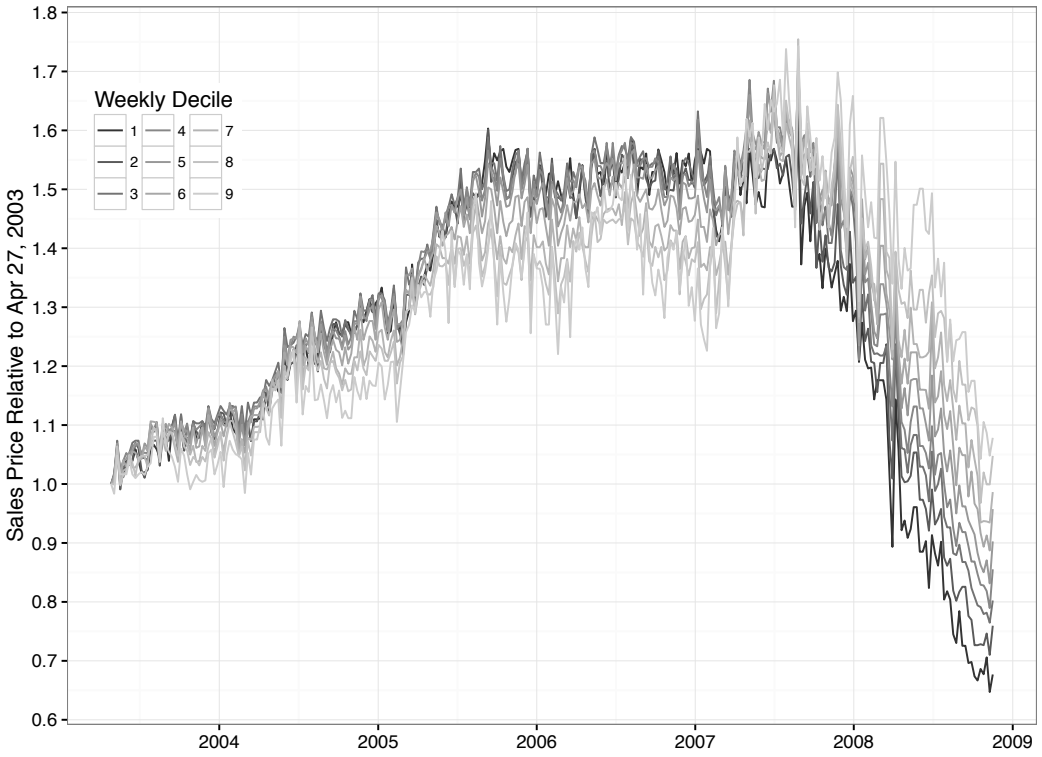

Figure 3.15: Deciles of Weekly Sale Prices. *These 9 line plots show the change in weekly housing prices over a 6-year period. Each line represents a decile in weekly house prices. These are normalized to the respective decile value for sales in the week of Apr 27, 2003 (the beginning of the data collection period). The least expensive houses show the greatest variability, rising the most and dropping the most, relative to their starting position.*

3.2.8 Geographic Data

The housing data include geographic information. In addition to street address, we have the latitude and longitude of each house sold. With this information, we can examine spatial relationships. For example, in Figure 3.16, we have plotted the upper quartile of sale price across a latitude-longitude grid. The upper quartile is coded as a color ranging from blue ($100,000) to red ($10,000,000). These colored tiles use alpha transparency so that we can discern the city names beneath the them. Information is revealed in this map that is not apparent from the scatter plots, smooth curves, box plots, and time series plots we have examined already. Furthermore, the background map shows land masses, bodies of water, cities, and roadways; these details provide additional context for interpreting the patterns in housing prices. In Figure 3.16, we see that sale prices on the San Francisco peninsula are higher than those in the east bay; houses along the Golden Gate are the most expensive in the city; and housing in Marin County north of San Francisco also command high-prices. The regions with very few colored tiles include national, state, and regional park areas. Additionally, the bands of color that run along both sides of the bay indicate that houses in the hills command relatively higher prices than those in the flats.

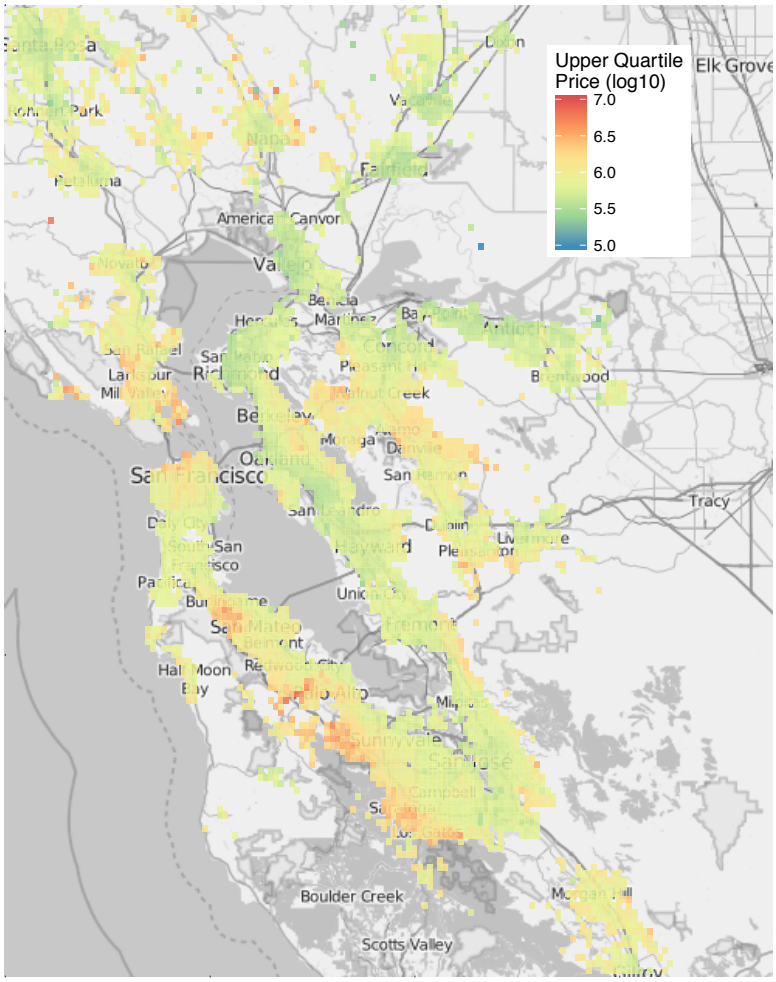

Figure 3.16: Housing Price Map. *This map of the San Francisco Bay Area shows the locations of the houses sold in 2004. The points are color-coded according to the upper quartile in sale price at that geographic location. It is evident that prices on the peninsula (San Francisco and south) are higher than those in the east bay; houses in the Golden Gate area of San Francisco are the most expensive in the city; north of San Francisco is also a high-priced area; and houses in the east bay and on the peninsula with higher elevation command relatively higher prices. The city of Piedmont, which has the highest prices in the east bay appears as a small red area surrounded by Oakland.*

3.3 Guidelines

In this section, we introduce guidelines for making effective statistical graphs through a collection of before-and-after plots. The 'before' plots are adapted from or inspired by plots we have encountered in the news and on the Web. For each, we describe several problems with the 'before' plot and how they are addressed in the 'after' version.

Word clouds such as the one shown in Figure 3.17 have gained tremendous popularity. The word cloud randomly arranges words found in text documents and scales them according to their frequency of occurrence. These visualizations are terrific for t-shirt designs, but they are not very effective at communicating the essential information in the data and in many cases can be misleading. The data underlying the pair of graphs in Figure 3.17 come from job listings on Kaggle. All postings for 'data science' or 'data scientist' were scraped from Kaggle's job postings in January, 2015. The listed skills were extracted from these ads and tallied. The word cloud displays the top 20-25 terms in a random pattern, and the height of each word is proportional to the frequency of occurrence in the listings. The random arrangement makes it difficult to compare the frequency of skills, e.g., is statistics or python listed more often? Furthermore, short terms are not correctly represented. For example, based on the area that the letter R covers, it appears that *R* occurs in about 1 listing to 4 or 5 for *Python*. However, we see in the dot chart in Figure 3.17 that this is clearly not the case. In fact, *R* appears more often than *Python* in these listings. The problem is that in the word cloud the height of the letters in a term is proportional to the frequency so a term with the same frequency as another but with more letters covers a greater area and appears to have a greater frequency than the shorter term. Although the dot chart is not as eye-catching as the word cloud, it orders terms according to their frequency making them easy to compare accurately.

The line plot on the left of Figure 3.18 is a remake of a plot that was presented by Congressman Chaffetz (R-UT), chairman of the US House Oversight Committee in the 2015 hearings investigating federal funding of Planned Parenthood (`https://oversight.house.gov/interactivepage/plannedparenthood/`). This plot originally appeared in a report by Americans United for Life (`http://www.aul.org/`). It demonstrates a clear violation of good graphics principles. The lines are meant to show the change in the types of procedures carried out by Planned Parenthood from 2006 to 2013. At first impression it appears the number of cancer screenings plummeted while the number of abortions sky-rocketed in this 6-year period. However, close inspection reveals that the plot has no y-axis, and the lines for cancer screenings and abortions are drawn on different scales. That is, Planned Parenthood performed 327,000 abortions and 935,573 cancer screenings in 2013, yet cancer screenings appear below abortions in this plot. One way to fix this problem, is to use the same vertical scale for both lines. However, since the number of cancer screenings in 2006 is nearly 10 times the number of abortions (2,007,371 compared to 289,750), the increase in abortions from 289,750 to 327,000 appears roughly flat in this plot. Given that there are only 4 numbers in this plot (we do not have annual figures from 2006 to 2013), it may be more informative to report the proportion of the total number of procedures via a dot chart (on the right in Figure 3.18). We have grouped the dots according to the type of procedure to make it easier to compare the change from 2006 to 2013. There we see that cancer screenings dropped from 87% of procedures in 2006 to 74% in 2013, and this drop was mirrored by an increase in abortion procedures (13% to 26%). The actual changes are not nearly as dramatic as those depicted in the original plot.

The World Resources Institute (`http://www.wri.org/`) provides data on historical carbon dioxide (CO2) emissions from fuel combustion (`http://cait.wri.org`). The

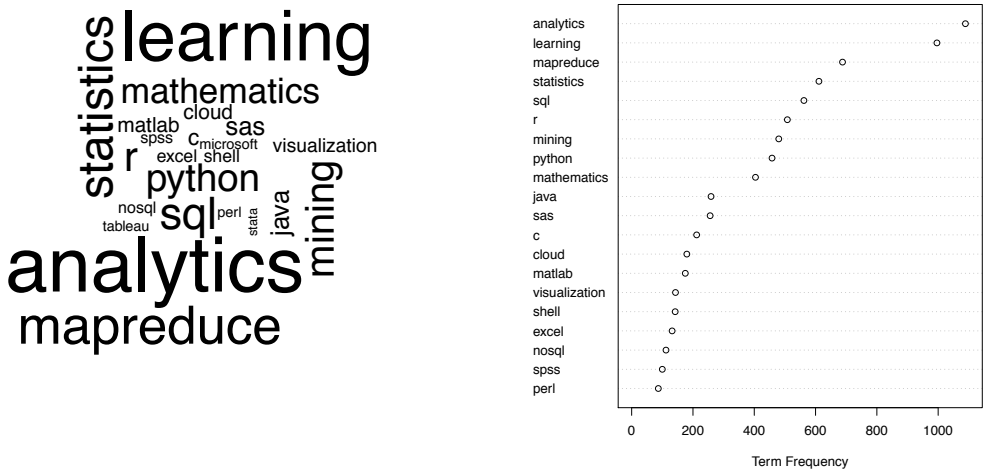

Figure 3.17: Kaggle Job Postings. *The word cloud (left) is an attractive graphic but it is difficult to compare the frequencies of terms due to their random arrangement. In contrast, comparisons are easy and accurate with the dot chart (right). Moreover, the word cloud is visually misleading because the height, not the area, of each word is proportional to its frequency. These terms are from job postings for a data scientist on Kaggle in January 2015.*

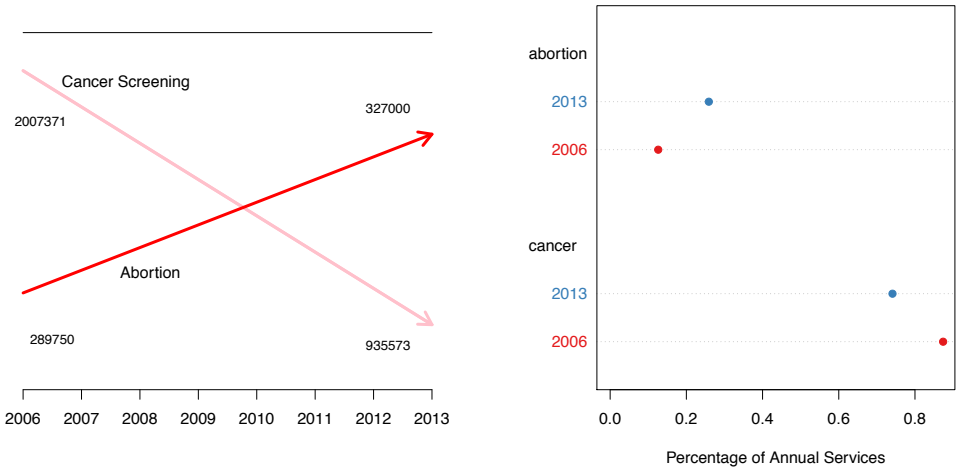

Figure 3.18: Planned Parenthood Services. *The line plot (left) shows Planned Parenthood's decrease in cancer screenings (pink) and increase in abortion services (red) from 2006 to 2013. This plot gives a misleading comparison of the change because the two line segments are on different scales (note that the 327,000 abortions appears above 935,573 cancer screenings for 2013. On the other hand, the dot chart (right) shows the change in the proportions of services for these two years. There we see that the percentage of cancer screenings dropped from 87% to 74% and the abortions rose accordingly from 13% to 26%.*

data that we have downloaded provide country annual CO2 emissions dating back to 1850. We have plotted trends since 1950 for the 14 countries that emitted the greatest amount of CO2 in 2012. The plot on the left of Figure 3.19 is an example of the colorful plot that we often see made with data like these. Since we have emissions over time, it is natural to make a line plot to see trends in emissions. However, stacking these line plots makes it difficult to compare country trends. Only the first country's data and the total for all 14 countries have a horizontal base that enables us to accurately perceive changes in emissions. For all of the other countries, the change from one year to the next is the length of a vertical segment where both the top and bottom of the segment move up and down from one year to the next, which makes it difficult to assess the trend and size of change. In contrast, the super-posed line plots (Figure 3.19, right) are not stacked, i.e., all values are rendered relative to the horizontal axis. We have also used a log scale. With this scale the data fill the plotting region, we can more easily compare countries with emissions that are different orders of magnitude, and we can see the kind of the growth some countries have undergone in this period. We have also used only 7 colors in this plot, rather than 14, because most people have trouble distinguishing between more than about 7 colors. We use 2 different line types for each color to distinguish between countries, e.g. Brazil and Japan have the same color but Japan's line is dashed.

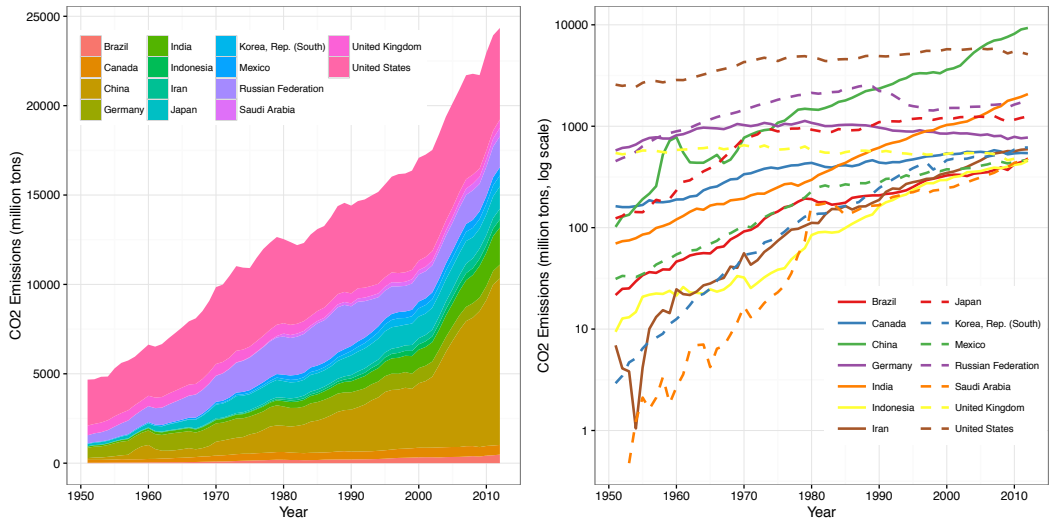

Figure 3.19: CO2 Emission Trends. *The stacked line plot (left) displays the rise in CO2 emissions from 1950 to 2012 for the 14 countries with the highest emissions in 2012. It is difficult to compare the countries because the base line for each country jiggles up and down with the emissions of the country below it. Only the first country at the bottom of the plot has a straight base line, but the tremendous change in China's emissions over this 60-year period masks any changes for, e.g., Brazil and Canada. In contrast, the line plot (right) is not stacked so the countries can be directly compared. Also the emissions are displayed on a log scale making it easier to see how both the small and large countries are changing.*

The graphs in Figure 3.20 examine the improvements in the manufacturing of the Intel chip for desktop computers since 1974, when Intel released its first microprocessor for a home computer (the 8080 chip). The plot on the left of this figure tracks the number of transistors on each new desktop model of microprocessor introduced from 1974 to 2004. (These data are from `http://computer.howstuffworks.com/microprocessor1.htm`.) Intel now

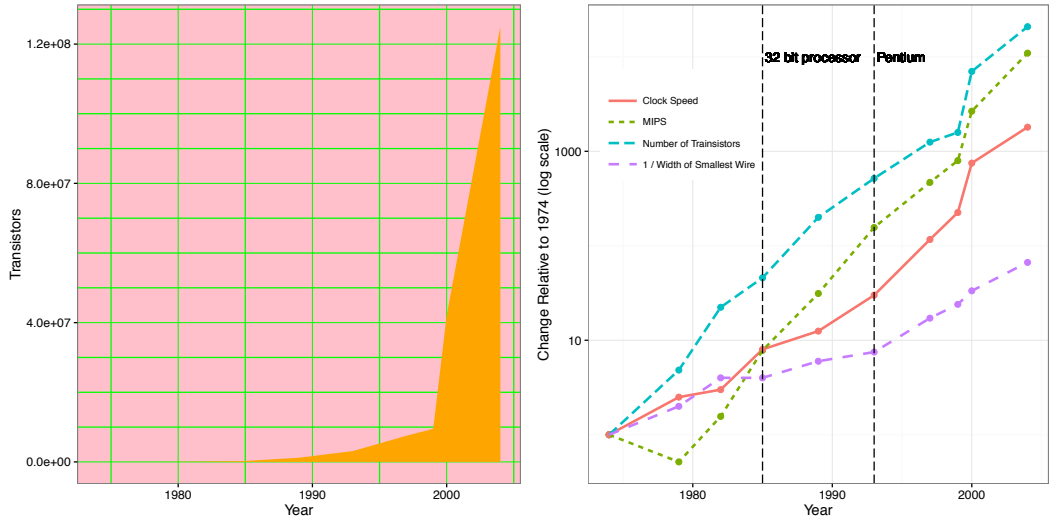

Figure 3.20: Intel Chip Benchmark Scores. *The line plot of the number of transistors on an Intel chip from 1972 to 2004 (left) demonstrates many flaws. The value for 2004 is orders of magnitude greater than the values from the 1970s through 1990s making it difficult to see how the relationship has changed over time. Furthermore, the use of color for the background, grid lines, and the region between the line and axis are too bright, distracting, and offer no insights. Alternatively, the plot on the left uses color to differentiate between the various measures of improvement, the y-axis is on a log-scale so that we can see how quickly these scores have grown, and two important design changes for the chip are denoted with reference lines.*

manufactures microprocessors with multiple cores and millions of transistors, but the new technology is not as easily compared to the earlier technology. For example, clock speeds in 2014 remain close to those of the Pentium 4, but the newer CPU can get more work done in one cycle so comparing clock speed is not informative. We made the visualization on the left in Figure 3.20 as an example of glaringly poor graphics design. It breaks many of the guidelines for good graphics:

- The data do not fill the plotting region because the number of transistors has grown so quickly that the most recent years take up most of the vertical range. This makes it very difficult to assess the kind of growth, e.g., we cannot tell whether the growth is exponential or slower/faster.

- The colors are very bright which makes it difficult to examine the plot closely. We typically want to be able to study a plot for minutes at a time so want to choose colors that facilitate careful inspection. Bright colors such as these have an after image effect that interferes with our inspection.

- Color should be used sparingly and should convey information. In this plot, orange fills the region between the data curve and the x-axis to no particular purpose. Also, the grid lines and background should use colors such as grey, white and black because they recede and do not interfere with the view of the data.

- The plot has obtrusive grid lines; they are too numerous, too thick, and too bright. We

want grid lines to assist us in reading off values from the data curve and not to dominate or interfere with our perception of the data.

- In addition to the number of transistors, our data include other aspects of the chip, such as the width of the smallest wire (in microns), the clock speed or the number of clock cycles a CPU can perform per second (in Mega Hertz or Giga Hertz), and the number of instructions the chip can execute per second (MIPS – millions of instructions per second). Ideally, we want our plot to be rich with information about these additional aspects of the chip.

The plot on the right of Figure 3.20 remakes the original to address these issues. The y-axis is on a log scale so that we can see the rate of improvement. All color has been eliminated, except for the color of the lines. These multiple curves show the changes in clock speed, size of the wire, and MIPS, in addition to the number of transistors. Since they are measured in different units, the values for each variable are scaled by its 1974 measurement. In addition, we have added 2 reference lines that mark major developments in the chip–the 32-bit processor and the Pentium processor.

These pairs of before-and-after plots have introduced many of the considerations that we take into account when designing a statistical graph that effectively conveys a story. We summarize the ideas presented in these examples into a set of topics and guidelines for good graphics.

3.3.1 Scale

When we choose the scale of an axis, we try to fill the plotting region with data. See for example, Figure 3.19, where the log transformation and limits on the y-axis create line plots that make it easy to see that some country emissions have remained flat and others have grown exponentially in this 60-year period. The log-transformation makes some lines appear at a 45-degree angle, which we call banking to 45 degrees. Banking makes it easier to ascertain a trend in the data. If we want to highlight some structure in the data that is not readily visible from the scale that we have chosen, then we can make a second plot for a subset of the data that uses a scale that brings out this additional feature. We can place this second plot in an inlay to the main plot or next to the original plot. Additionally, we may want to drop a few unusually large observations in order to get a better view of the main portion of the data. See for example Figure 3.14 where unusually expensive and large houses are not included in the scatter plot nor in the calculation of smooth curve. If we do not include all of the data in the graph, then we need to mention in the caption or on the plot itself that we have dropped some observations.

Depending on the measurements, it may not be necessary to include 0 on the axis, especially if including it makes it difficult to fill the data region. For example in the y-axis for the box plots in Figure 3.12, there is no need to include 0 as the lowest prices are near $100,000. On the other hand, the bars in Figure 3.7 include 0 so the heights of the bars for 10% have the correct proportion when compared to the bars for 40%, i.e., they are 1/4 as tall.

3.3.2 Position

When we juxtapose plots that contain different subsets of the data, we want to use the same limits on the axes of the plots in order to facilitate comparisons across plots (see Figure 3.13). We want to arrange bars in a bar chart and dots in a dot chart in increasing order and side-by-side box plots in increasing order of the median (Figure 3.12). An exception to this

convention is when the groups are naturally ordered, such as with education level. If groups are defined by two qualitative variables, then we want to arrange the groups in a way that emphasizes the important difference; see for example the grouping of the bars for level of education within smoking status in Figure 3.9.

Comparisons are often easiest if we superpose plots. See for example dot plots in Figure 3.9 (bottom right), density curves in Figure 3.10 (top right), line plots in Figure 3.19 (right), and smooth curves in Figure 3.14 (right). If superposition is not feasible then side-by-side comparisons are preferable (Figure 3.9, top right) to stacking bars or curves because stacking makes it difficult to compare subgroups.

If many observations have the same values, we can add a small amount of random noise to these values in order to reduce the amount of over plotting and better see the underlying relationships (Figure 3.8).

3.3.3 Shape

Length is easier to compare and assess than angle, and for this reason barplots and dot charts are preferable to pie charts (Figure 3.7). However, stacked bars and line plots hinder comparisons as they lack a constant base line. Each dimension of a multidimensional plot (e.g., the depth in a 3-dimensional bar) should represent a variable, otherwise it is unnecessary. Similarly, maps that fill geographic regions can be misleading if the density of observations is not constant over these regions.

3.3.4 Aggregates

With large amounts of data, we often want to visualize aggregates of the data. Smooth curves or lines can be plotted in addition to or instead of scatter plots to make average relationships more apparent (Figure 3.14, right). Alpha transparency shows darker colors when symbols are over plotted, which can help reveal high density regions (Figure 3.13). Color can represent density in a smooth scatter plot (Figure 3.14, left).

When the data have been collected according to some sampling design and the observations represent different numbers of individuals, we need to incorporate the sampling design in any plots that we make (Figure 2.24).

3.3.5 Color

If we use color in a plot, the color should represent information and not be gratuitous. Depending on the type of information being represented, different color palettes are preferable. For categorical data, we want to use a collection of colors that are easily distinguishable and where one does not stand out more than another. For continuous data, we want to use a sequential gradation that emphasize one end of the spectrum of values over the other, or we want a diverging palette that emphasizes the two extremes of the spectrum over the middle. The choice between a sequential and diverging palette depends on the message being conveyed. For example, with cancer rates, we want to use a sequential palette that increases the brightness and saturation for high rates. On the other hand, with two-party election results we want a diverging palette with two distinct hues for low and high values. Both ends of the palette use bright saturated colors to distinguish between one party's dominance over the other.

Colors can be specified in several ways in *R*, including by name, such as `cornflowerblue` and `lightgreen`. We can also use a triple of numbers for the amount of red, green and blue light to add together. These numbers are typically specified in hexadec-

imal and they range from 0 and 256, i.e. between 00 and FF. For example, the color called `cornflowerblue` is a combination of 100 red, 149 green and 237 blue, or in hexadecimal we express it as `"#6495ED"`. We discuss the various ways to specify colors in greater detail in Chapter 10. Creating a sequential, diverging, or qualitative palette of colors that appropriately conveys the underlying values in a variable is a difficult task, and we recommend using those that have been developed by researchers, such as Cindy Brewer's palettes (see the functions in the `RColorBrewer` package [8] for examples).

Plots are meant to be examined for long periods of time so we want to use colors that we can stare at without impeding our ability to perceive the important information in the plot. For example, we want to avoid colors that create an after-image when we look from one part of the graph to another. Similarly, we should avoid using combinations of colors that color-blind people have trouble distinguishing between. Furthermore, people have trouble distinguishing between more than about 7 to 9 colors so we should limit the number of colors we use in a plot. Finally, colors can appear different when, e.g., printed on paper or projected on a screen, so they should be chosen with the medium in which they will be presented in mind.

3.3.6 Context

Depending on whether our plots are part of an informal exploratory analysis or a formal presentation, we include different amounts of contextual information. However, even with EDA we want to include some context so that when we return to an analysis we can easily determine what we have plotted. It's good practice to consistently use informative labels on axes (including the units of measurement), labels on tick marks, and titles. Plot captions should describe what has been plotted and point out the important features of the plot.

Ideally, we include additional context to help tell the data's story. For example, reference markers and lines provide benchmarks to compare against and other external information that's helpful in interpreting the results. Examples include the background map of the plot of sale locations in Figure 3.16 and the reference lines in the plot of Intel chip historical development in Figure 3.20. We also use color and plotting symbols to include additional variables in a plot, e.g., the points in the scatterplots in Figure 3.13 are colored according to the number of bedrooms in the house.

3.3.7 Over Arching Considerations

The previous guidelines are organized according to various features of a plot. When we examine a plot that we have made, in addition to considering this checklist, we can ask ourselves more holistic questions about the visualization. The following three questions are abstracted from Cleveland's *Elements of Graphing Data* and they help serve as a framework for developing a statistical graph. We connect these questions with the above topics and guidelines.

Do the Data Stand Out?
Essentially, this is a question of scale, transformations, and banking. That is, we want the data to fill the plotting region in order to best reveal it's structure. By structure we mean the shape of a distribution or the relationship between variables. Other considerations in answering this question address whether the data are obscured because, e.g., plotting symbols are too small, there's too much over plotting, graphing elements cover the data, colors are too bright, and extraneous glyphs detract from the visual perception. We want to avoid hiding the data in our visualization.

Does the Plot Facilitate Comparisons?

When we create a plot, we need to keep in mind what is the important comparison and we want to be sure our plot has emphasized this comparison well. A comparison can be a simple benchmark or reference or a partition of the data into subgroups that we want to compare. We consider whether or not we can further reveal a relationship by providing more information or context by, e.g., using color to denote another variable. Additionally, we consider which kind of plot makes it easy for the reader to make accurate comparisons and whether we want to superpose graphing elements on the same plotting region, e.g., density curves, or juxtapose plots, e.g., a grid of scatter plots for subgroups. We have seen that aligning dots along the same axis is one of the easiest and most accurate approaches to compare proportions, much better than stacking bars or dividing pies into slices. The proper choice of a color palette can also facilitate comparisons. Lastly, identifying individual points can assist in making comparisons of one or a few cases against the rest.

Can We Add Information?
We want our visualizations to be information rich without distracting from or cluttering the main message. We nearly always want our axes labeled (an exception is with a map). Our plots should contain titles, and when needed, legends. We can add information with color and plotting symbols, reference markers, and point labels. This additional information provides a context for interpreting the findings. Importantly, we want to clearly describe these findings and what we have plotted in a comprehensive caption.

3.4 Iterative process

Making a good statistical graph is an iterative process of discovery. After we make a plot, we examine it for ideas on how to improve that plot and we consider whether we should make an entirely different plot. We consider the 3 questions in Section 3.3.7 as we continue to try to uncover the story in the data and find a way to present it clearly and effectively. For example, in making the plot about Intel chips, we went through several stages of development. Our first visualization (top left in Figure 3.21) was similar to the plot on the left in Figure 3.20, without all of the garish bells and whistles. When we examined this plot, we realized that we needed to transform the number of transistors in order to fill the data region and get a sense of the rate of growth. This led us to the second plot (top right) in Figure 3.21. We use base 10 logarithm rather than the natural logarithm to make it easier for the viewer to read the original units, e.g., 1,000 vs 10,000 transistors. This plot provides a reasonable visualization, but we asked ourselves whether we can include more information. The problem with including other measurements of the chip is that these variables have incompatible units. Since it's the pace of the change that interests us, we can convert the data into the same units by scaling each to their respective value in 1974. To create the next plot, which appears in the bottom left of Figure 3.21, we needed to perform several data manipulations (see the exercises). At this point, we think that the plot is basically complete, except for the addition of context to help in understanding the message. The final version (bottom right) has informative axis labels and legends, and we have added two reference lines to indicate major advances in chip technology. Note that we would normally add a title to our plot, but we use the figure title for the plot title. The caption describes what we have plotted and points out important features in the plot.

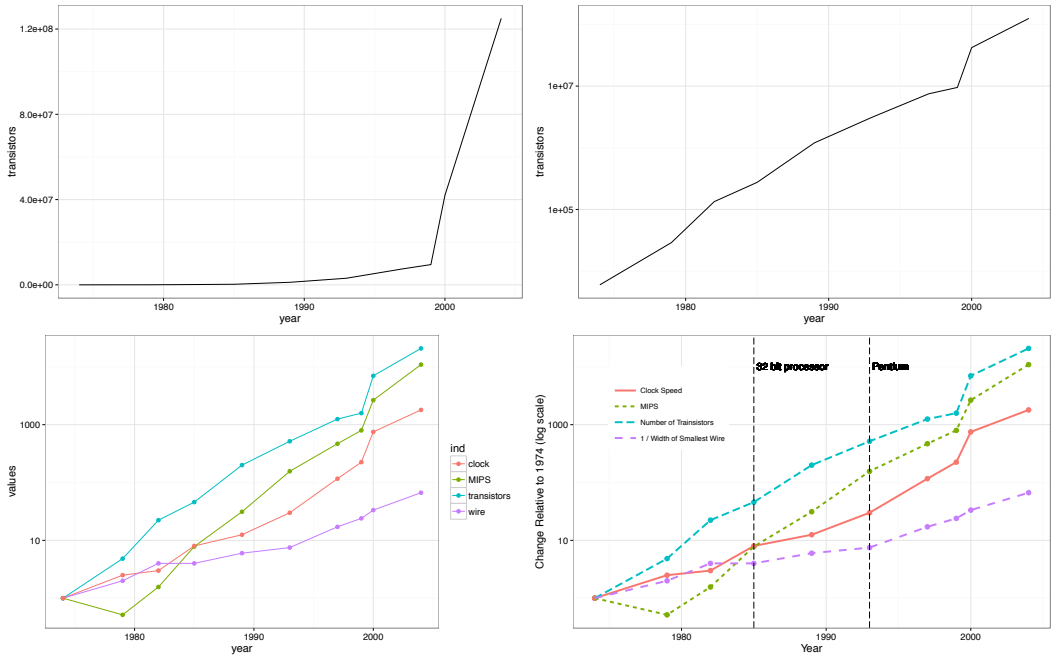

Figure 3.21: Intel Chip Benchmark Scores – An Iterative Visualization Process. *These 4 graphs represent the 4 steps in the process of creating the visualization of the development of the Intel chip. They are (from top left to bottom right) the initial plot of the number of transistors, transforming the number of transistors to log-scale, transforming the measurements of all the variables so they are relative to the 1972 values, and adding context with more informative labels, legends, and references markers.*

3.5 *R*s Graphics Models

There are two models for creating statistical graphs in *R*: the painter's model, which the functions in base *R* provide, and the object-based model, which follows Grammar of Graphics [14] and has been implemented in the `ggplot2` package [**?**] using grid graphics approach of the `grid` package [7]. We have used both techniques to create the plots in this chapter and other chapters. However, since these two models take quite different approaches, we encourage the beginner to focus on one method at first. We briefly describe them both in this section.

3.5.1 Painter's Model in Base *R*

The graphics functionality in base *R* creates a statistical graph from a high-level function, such as plot(). There are a host of functions for making different kinds of plots, including hist(), plot(), boxplot(), dotchart(), barchart(), and mosaicplot() to make a histogram, scatter plot, box plot, dot chart, bar chart, and mosaic plot, respectively. A call to one of these functions initiates a new plot. We can think of this as a new plot on a new 'page', like a painter starting a new painting on a new canvas. This function call creates a complete plot.

In addition to these high-level plotting functions, there are several low-level functions

that can add more to the current plot. These functions include those to add a line, additional connected line segments, additional points, shapes, text, and legends, the corresponding functions are abline(), lines(), points(), polygon(), text(), and legend(), respectively.

The high-level plot functions have many common arguments to adjust the appearance of the plot. In most plots, we can modify the default labels for the axes (*xlab* and *ylab*), range of the axes (*xlim* and *ylim*), title of the plot (*main*), color of points and lines (*col*), plotting symbol and size (*pch* and *cex*), and type and thickness of line, (*lty* and *lwd*). There are many other parameters, some of which make sense for a particular type of plot, e.g., *vertical* to indicate whether the bars in a bar chart are to be vertical or horizontal; *freq* to indicate whether the area of the bars in a histogram should be counts or proportions; *breaks* to specify the number or location of the intervals in a histogram; *groups* to indicate whether the dots in a dot chart are to be grouped by a categorical variable; and *labels* to change the default labels for the dots.

Finally, the par() function can be used to globally control many plotting parameters. Some of the arguments to par() are exclusive to this function. Two that are quite useful are *mar* to control the size of the plot margins and *mfrow* to divide the canvas into subpanels for multiple plots. We highly recommend reading the documentation for par() to get a sense of the tremendous flexibility available for making plots with base *R* functionality.

We close this section with 2 examples of code that we used to create 2 of the plots in this chapter. Our first example creates the scatter plot in Figure 3.11. We begin by making a vector of muted green and purple to use for the color of the points and lines in the plot. We specify these with their RGB (red-green-blue hexadecimal values) as

```
smokeColors = c(Never = "#1b9e77", Current = "#7570b3")
```

Notice that we assigned names to these elements to match the levels of the `smoke` variable in `babies`.

We call plot() with

```
with(babies[babies$smoke == "Never" | babies$smoke == "Current", ],
     plot(x = jitter(ht, amount = 0.5),
          y = jitter(bwt, amount = 0.5),
          xlim = c(58,71),
          pch = 19, cex = 0.4,
          col = smokeColors[as.character(smoke)],
          xlab = "Mother's Height (in)",
          ylab = "Birth Weight (oz)"))
```

We use with() to make the scatter plot because it makes it easier to specify the subset of current and never smokers and avoid using `$`-notation when we specify the variables in plot(). Notice that we used jitter() to add a small amount of random noise to both height and birth weight to avoid overplotting. We also shrank the plotting symbol to limit the amount of over plotting. We have specified the limits of the x-axis to zoom in on the main portion of the data; a few unusually small/tall mothers are not included in the plot. The argument *col* is set to the expression `smokeColors[as.character(smoke)]`. This expression uses indexing by name to create a vector of reds and blues for the current and never smokers.

Now that we have created the basic scatter plot with the high-level function plot(), we use loess() to fit two smooth curves to the data, for the current and never smokers. Later we add these 2 curves to this scatter plot using the lower-level function lines(). We fit the two curves with

```
bwtN.lo = loess(bwt ~ ht,
```

```
            data = babies[babies$smoke == "Never", ], span = 0.5)
bwtS.lo = loess(bwt ~ ht,
            data = babies[babies$smoke == "Current",], span = 0.5)
```

The first argument to loess() is a formula that indicates we want to model birth weight as a function of mother's height. The fit is a smooth curve that essentially averages birth weight for mothers with similar heights.

Next, we use the information from the fits in `bwtN.lo` and `bwtS.lo` to make predictions of birth weight for a dense grid of heights from 59 to 69 inches. This way we can draw a 'curve' with line segments that connect the sequence of predictions. We create the set of heights with

```
gridHt = data.frame(ht = seq(59, 69, 0.2))
```

Then, we call predict() and pass it the fitted loess object (`bwtN.lo` or `bwtS.lo`) and the vector of heights in `gridHt` to find the predicted birth weights with

```
pred.bwtN = predict(bwtN.lo, gridHt, se = FALSE)
pred.bwtS = predict(bwtS.lo, gridHt, se = FALSE)
```

Lastly, we add these pairs (height, predicted birth weight) to the scatter plot and connect the dots with lines(). We do this with

```
lines(x = gridHt$ht, y = pred.bwtN,
      col = smokeColors["Never"], lwd = 3)
lines(x = gridHt$ht, y = pred.bwtS,
      col = smokeColors["Current"], lwd = 3, lty = 3)
```

Notice that we matched the color of the curves to the associated group, used different line types to further help distinguish between the curves, and made the line thicker so that it stands out from the point cloud.

For a second example, we demonstrate how to make the bar chart in Figure 3.2. This bar chart is made from the following matrix of counts of counties,

```
                        [,1] [,2] [,3] [,4] [,5] [,6] [,7]
Majority.of.Democrats     43   37   29   21   23   28   28
Majority.of.Republicans   15   21   29   37   35   30   30
```

Again, we begin by creating a vector of colors with

```
require(RColorBrewer)
partyColor = brewer.pal(3, "Set1")[1:2]
```

A call to barplot() make the basic bar chart for us with

```
barplot(countyCounts,
        beside = TRUE,
        col = partyColor,
        axes = FALSE,
        xlab = "Presidential Election Year",
        ylab = "Number of Counties",
        ylim = c(0, 47),
        main = "Majority Party in California Counties")
```

Note that we have asked that the bars corresponding to the 2 rows in the matrix (the number of Democratic and Republican majority counties) to be plotted side by side and that their colors match the traditional party colors, which we provided in *col*. We also provide labels for both axes and a title. This plot is unusual in that we have turned off the automatic drawing of the axes with `axes = FALSE`. We do this because we want to control the location and labels for the tick marks on the x-axis.

We complete the plot with several lower-level function calls to add axes, a box around the plot, text above each bar, and a legend. The axes are numbered 1 through 4 beginning at the bottom and moving clockwise around the perimeter of the plot. We begin by adding the y-axis on the left side of the plot with

```
axis(2)
```

This call, `axis(2)`, uses the default value for the location and labels of the tick marks. Then, we add the x-axis along the bottom of the plot with

```
axis(1, at = seq(2, 22, by = 3), labels = seq(1992,2016,4))
```

Here we place ticks between each pair of bars and label them with the election year. Note that the actual scale on the x-axis runs from 1 to 23, not from 1992 to 2016; the years are simply labels on the tick marks.

We use text() to add the county counts above their respective bars. To do this, we provide the (x, y) coordinates for the top of each bar and specify the *pos* argument as `3` so the text is placed above this point and doesn't interfere with the bars. We call text() with

```
text(x = rep(c(1.5, 2.5), 7) + rep(seq(0, 18, by =3), each = 2),
     y = countyCounts,
     label = countyCounts, pos = 3, cex = 0.8)
```

Notice that we also shrink the text to 80% of its normal size so that it doesn't detract from the bars, and again the x-axis runs from 1 to 23. To make the axes more visually pleasing, we add an L-shaped box with

```
box(bty = "L")
```

Lastly, we add a legend to the plot with

```
legend(x = 18, y = 47, legend = c("Dem", "Rep"),
       fill = partyColor, cex = 0.8, bty = "n", title = "Party")
```

Notice that when we specified *ylim* in the original call to barplot(), we created enough vertical space in the plot for the numbers above the tallest bin and the legend.

3.5.2 Grammar of Graphics Model in `ggplot2`

The implementation of the grammar of graphics in `ggplot2` takes a different approach to constructing statistical graphs. We begin by defining an empty plot object with ggplot(). Then we add layers to the plot by specifying the graphics shapes, called geoms, with which to view the data, e.g., plotting symbols and lines. These are added to the plot object with the `+` operator. Each layer can have its own data (which must be in a data frame) and aesthetic mapping. This mapping connects variables in the data frame to a feature, such as x and y locations, color, and size. As an example the aesthetic `aes(x = Yr, y = CO2, color = Ctry)` maps year to the x-axis, CO2 emissions to the y-axis, and uses color to

denote country. This aesthetic mapping can be used to add points or lines to a plot. The geometric shape (e.g., point or line) is also specified in the layer.

We can modify the scales for these aesthetics to, e.g. use a log-scale for the y-axis, provide a special label on the x-axis, and specify a palette for the colors. We do this by adding scale() functions to the plot. That is, scale_y_continuous() has parameters *name* to specify the axis label, *breaks* to provide the location of the tick marks, *trans* to use a transformation such as log, *limits* to denote the range of the scale, etc. Scale functions are also available for discrete-valued axes and for color, shape, line type, size, etc.

Additionally, details related to, e.g., the appearance of axes, size of text, and background color for the plotting region are specified through themes.

When we create a plot, we call ggplot() and then add layers of data and, if desired, scale and theme specifications. We provide 2 examples by reviewing the code used to create the line plots in Figure 3.19 and the grid of scatter plots in Figure 3.13.

We begin by creating the plot on the right side of Figure 3.19. The first step is to create a plot object with ggplot(). We can specify the data and the aesthetic mapping in this function call, or we can do this in the layers, or both. When these are specified in ggplot(), they are available to all layers and they can be overridden in a specific layer. Since we have only 1 layer in this plot (the lines), it makes no difference where we provide the data and mapping so we specify them in the call to ggplot(). We provide this information in ggplot() with

```
co2Plot = ggplot(data = co2Top14,
                 mapping = aes(x = Yr, y = CO2, color = Ctry,
                               linetype = Ctry))
```

We have created a plot object, not a plot. After we specify the geometric shapes, we can print/see the plot. Notice that we map `Yr` to the x-axis, `CO2` to the y-axis, and `Ctry` is mapped to both color and line type. When we add a layer for lines to this plot, we see how this mapping is rendered. We add this layer with

```
co2Plot + layer(geom = "line", stat = "identity",
                position = "identity",
                params = list(size = 1))
```

Each layer has a geometric shape and a statistic. Depending on these, we can pass additional parameters to control the geom. In this example, we simply want to make the lines thicker so we set *params* to a list with a named `size` element. The *position* argument specifies how to handle over plotting. The value `'identity'` indicates that we are not jittering or performing any sort of repositioning to address over plotting.

Several short cut functions are available for layers. That is, each type of geom has a default statistic and position and the layer can be called with geom_XXX(), where `XXX` is the name of the *geom* and the default values for *stat* and *position* for this geom are provided. The addition of the above layer to `co2Plot` is equivalent to

```
co2Plot + geom_line(size = 1)
```

This one layer produces a plot very similar to the plot in Figure 3.19, except for the scale of the y-axis, the axis labels, locations of the tick marks for both axes, choice of color, line type for the lines, and the appearance and location of the legend. We modify the scales for the x and y axes and for the color and line type with scale components that we add to our plot. We change the x and y axes by adding the following calls to `co2Plot` with

```
co2Plot + geom_line(size = 1) +
  scale_x_continuous(name = "Year",
                     breaks = seq(1950, 2010, 10)) +
  scale_y_continuous(
    name = "CO2 Emissions (million tons, log scale)",
    breaks = c(1, 10, 100, 1000, 10000), trans = "log10") +
```

Note the trailing + indicates that we plan to add more component specifications to the visualization.

We want to use solid and dashed lines for each of 7 colors from a Brewer palette so that we can differentiate between the 14 countries. To do this, we add calls to scale_linetype_manual() and scale_color_manual() to the above plot object with

```
  scale_linetype_manual(
    name = "",
    values = rep(c("solid", "dashed"), each = 7),
    guide = guide_legend(nrow = 7)) +
  scale_color_manual(
    name = "",
    values = rep(brewer.pal(7, "Set1"), 2)) +
```

By setting the *name* of the line type and color scales to the same value, the legends for these two scales are combined into 1.

Lastly, we use the black and white theme for the background colors and grid lines and we modify the location and appearance of the legend with

```
  theme_bw() +
  theme(legend.position = c(0.75, 0.25),
        legend.text = element_text(size = 8),
        legend.key.width = unit(2, "line"),
        legend.key = element_blank())
```

In addition to placing the legend inside the plotting region, we shrank the labels in the legend, widened the key so that the line types are clear, and eliminated the box around the key.

For our second example, we create the plot in Figure 3.13. We need to create the variable price per square foot and to collapse the number of bedrooms to 1 through 5, where 5 represents 5 or more. We add these new variables to our original data frame because `ggplot2` expects the variables to be collected in one data frame. We do this with

```
housing04S$ppsf = housing04S$price/housing04S$bsqft
housing04S$br5 = housing04S$br
housing04S$br5[housing04S$br5 > 5] = 5
```

Note that `housing04S` contains the sales for 2004 for 12 cities in the East Bay, as described in Section 3.2.6.

We also want to add a smooth curve to each of the panels in the figure. This curve is created by averaging over all cities. We can fit the curve to our data and then create a data frame with equi-spaced values for building square footage and the respective prediction of price per square foot. We do this with

```
housingSmooth = loess(ppsf ~ bsqft, data = housing04)
housingPreds = data.frame(bsqft = seq(320, 5000, by = 100),
housingPreds$pred = predict(housingSmooth, newdata = housingPreds)
```

The data frame `housingPreds` contains two variables `bsqft` and `pred`.

We have completed our data preparation and can begin to make our plot. We 'add' together the following function calls to build the plot, beginning with creating the plot object with

```
ggplot(data = housing04,
       mapping = aes(x = bsqft, y = ppsf)) +
```

Notice that we have specified the aesthetic for mapping the x and y axes but not the color mapping because the color aesthetic is needed only in the point layer. We add the point layer with

```
  geom_point(aes(col = factor(br5)), size = 0.5, alpha = 0.5) +
```

In addition to specifying the mapping of the number of bedrooms to the color aesthetic, we also set the size of the points to be 1/2 their regular size and we set the transparency level to 1/2. These settings help with over plotting because smaller points do not overlap as much and, when they do, the transparency allows us to see the point density and colors.

To create the panels of scatter plots, one for each city, we use facet() as follows:

```
  facet_wrap(~ cityO) +
```

The argument to facet() is a formula that says that we want to see the dependency on cities. The variable `cityO` is an ordered factor that we have created so the cities are ordered according to median price.

The curve in each panel is a second layer. This layer is created from line segments that connect price per square foot predictions (`pred`) for the grid of building square footage values (`bbsf`) in our `housingPreds` data frame. We add this line layer with

```
  geom_line(data = housingPreds, mapping = aes(y = pred)) +
```

Notice that we need to specify the data frame in this layer because it is different from the default data frame provided in ggplot(). We also map the y-axis to `pred` because it too differs from the y-axis aesthetic supplied in ggplot().

Now that we have added our 2 layers, we control the scales for the 2 axes and the color with scale_xxx_xxx() functions. We provide the title and the limits for the x and y scales with.

```
  scale_y_continuous("Price per Square Foot",
                     limits = c(0, 1500)) +
  scale_x_continuous("Building Square Foot",
                     limits = c(0, 6000)) +
```

In addition, we specify the colors to use and map them to labels in the color scale with

```
  scale_color_manual("Bedrooms",
                     labels = c(1:4, "5+"),
                     values = brColors,
                     guide = guide_legend(
                       override.aes = list(size= 2, alpha = 1))) +
  theme_bw()
```

Notie that we also override the point size and alpha transparency that was set in geom_point() for legend readability. Our last addition specifies that we want to use the black and white theme for the background, axes, etc.

3.6 Creating Unique Plots

After some practice making standard plots in base *R* or `ggplot2` and some experience in following the graphics guidelines, you will be ready to design your own unique plots. We saw an example of this in Figure 2.15. A more typical visual representation of these data would be a graph that arranges circles (nodes) for the individuals and represents the emails by lines or arrows from the sender to the recipients. However, with so many individuals and emails involved, it is difficult to arrange this network on a page so that it is readable. The BioFabric plot was created to address this problem.

A second example appears in Figure 4.7. This plot is unique in that it aims to create a visualization for missing data. We can think of it as 56 rug plots, one for each weather station. The yarns are placed at the days when we have a precipitation measurement at the station, including the days when 0 inches of rainfall was recorded. The vertical stripes in the plot reveal that there are time periods when we have no data recorded for all 56 weather stations (these are the winter months). Additionally, the horizontal bands of white indicate that some stations have not been in operation as long as others and that a few stations were out of operation for lengthy periods of time.

According to Wainer [11], we don't want to re-invent a visualization if it has been done well. Moreover, if we can use a standard plot, such as those that appear in this chapter, then it's typically a good idea to use that rather than invent our own because viewers are familiar with how to read and understand the existing format. However, there are times when we want to adapt one of these standard visualizations (as in Figure 4.7) or create an entirely new kind of visualization (as in Figure 2.15) to overcome a limitation or present information from a new perspective. And we may even develop a function to create a specialized visualization. For example, in Section 7.7, we write a function to visually examine the results of a simulation study (see Figure 7.16).

3.7 Summary

There are many resources on how to design informative graphics. Some of our favorites are [3, 4, 5, 10, 11]. There are also many resources on how to create visualizations in *R*. For `lattice` plots see [9]. The grammar of graphics by Wilkinson [14] has been implemented in *R* by Wickham in the `ggplot2` package; see [12] for more details. For an in-depth treatment of all graphics models in *R* see [6].

3.8 Exercises

Bibliography

[1] Code to download and process SF housing sales data. `https://github.com/hadley/sfhousing`, 2009.

[2] Bay Area Home Sales - Weekly Updates. `http://www.sfgate.com/webdb/homesales/`, 2016.

[3] William S. Cleveland. *The Elements of Graphing Data*. Wadsworth Advanced Books and Software, Monterey, CA, 1985.

[4] Andrew Gelman. Why tables are really much better than graphs. *Journal of Computational and Graphical Statistics*, 20:3–7, 2011.

[5] Andrew Gelman and Antony Unwin. Infovis and statistical graphics: different goals, different looks. *Journal of Computational and Graphical Statistics*, 22:2–28, 2013.

[6] Paul Murrell. *R Graphics*. Chapman & Hall/CRC, Boca Raton, FL, 2005.

[7] Paul Murrell. `grid`: The grid graphics package. `http://cran.r-project.org/package=grid`, 2011. *R* package version 2.16.0.

[8] Erich Neuwirth. `RColorBrewer`: ColorBrewer palettes. `http://cran.r-project.org/web/packages/RColorBrewer`, 2011. *R* package version 1.0-5.

[9] Deepayan Sarkar. *Lattice: Multivariate Data Visualization with R*. Springer-Verlag, New York, 2008. `http://lmdvr.r-forge.r-project.org/figures/figures.html`.

[10] Edward Tufte. *The Visual Display of Quantitative Information*. Graphics Press, Connecticut, 1983.

[11] Howard Wainer. How to display data badly. *The American Statistician*, 38:137–147, 1984.

[12] Hadley Wickham. *`ggplot2`: Elegant graphics for data analysis*. Springer, New York, 2009.

[13] Hadley Wickham, Deborah Swayne, and David Poole. Bay Area blues: The Effect of the housing crisis. In Toby Segaran and Jeff Hammerbacher, editors, *Beautiful Data: The Stories Behind Elegant Data Solutions*, pages 303–322. O'Reilly Media, Inc., Sebastopol, CA, 2009.

[14] Leland Wilkinson. *The Grammar of Graphics*. Springer Science & Business Media, 2006.