# 1
# *Data exploration*

Supply and demand, chocolate and peanut butter, education and income . . . some things just go hand in hand. In each case, a particular idea about how things work turns upon the interpretation of an observed relationship between things we can measure. To do this correctly requires care, judgment—and the right toolkit. The goal of this chapter is to equip you with some basic visual and numerical tools for exploring multivariate data sets, with an eye towards finding interesting relationships among variables.

*Cases and variables.* In statistics, we typically refer to the *cases* and *variables* of a data set. The cases are the basic observational units that we're interested in: people, houses, cars, guinea pigs, etc. The variables are the different kinds of information we have about each case—for example, the horsepower, fuel economy, and vehicle class for a car. We typically organize a data set into a *data frame*. A data frame is like a simple spreadsheet where each case is a row and each variable is a column, like in Table 1.1.

Variables come in two basic kinds. Numerical variables are represented by a number, like horsepower. Categorical variables are described by the answer to a multiple-choice question, like vehicle class. This chapter will describe some strategies for summarizing relationships among both kinds of variables, as well as some further refinements to this basic "numerical versus categorical" distinction.

Table 1.1: A simple example of a data frame. Each case is a car, and there are five variables: horsepower, city gas mileage, highway gas mileage, weight (in pounds), and vehicle class.

| | Horsepower | CityMPG | HighwayMPG | Weight | Class |
|---|---|---|---|---|---|
| BMW 325xi | 184 | 19 | 27 | 3461 | Sedan |
| Chevrolet Corvette | 350 | 18 | 25 | 3248 | Sports |
| Mercedes-Benz CL500 | 302 | 16 | 24 | 4085 | Sedan |
| Dodge Neon | 132 | 29 | 36 | 2626 | Sedan |
| Acura MDX | 265 | 17 | 23 | 4451 | SUV |

## Variation across categories

MANY OF the data sets you'll meet will involve categories: chocolate or vanilla; rap or country; Toyota, Honda, or Hyundai; butcher or baker or candlestick maker. A simple, effective way to summarize these categorical variables[1] is to use a *contingency table*. On the Titanic, for example, a simple two-way table reveals that women and children survived in far greater numbers than adult men:

[1] Categorical variables are sometimes referred to as *factors*, and the categories themselves as the *levels* of the factor. The R statistical software package uses this terminology.

| | Girl | Woman | Boy | Man |
|---|---|---|---|---|
| Survived | 50 | 242 | 31 | 104 |
| Died | 22 | 74 | 51 | 472 |

Table 1.2: A two-way table, because there are two categorical variables by which cases are classified. The data are available in the R package `effects`. Originally compiled by Thomas Cason from the *Encyclopedia Titanica*.

We call this a two-way or bivariate table because there are two variables are being compared: survival status versus type of person. The categories go along the rows and columns of the table; the cell counts show how many cases fall into each class. The process of sorting cases into the cells of such a table is often called *cross-tabulation*.

We can also make multi-way tables that show more than two variables at once. Given the constraints of a two-dimensional page, multiway tables are usually displayed as a series of two-way tables. As the following three-way table reveals, richer passengers, of either sex, fared better than others.

| | Cabin Class | 1st | 2nd | 3rd |
|---|---|---|---|---|
| Female | Survived | 139 | 94 | 106 |
| | Died | 5 | 12 | 110 |
| Male | Survived | 61 | 25 | 75 |
| | Died | 118 | 146 | 418 |

Table 1.3: An example of a *multi-way table*, where counts are classified by cabin class, sex, and survival. NB: passengers of unknown age are included in this table, but not the previous one.

Tables are almost always the best way to display categorical data sets with few classifying variables, for the simple reason that they convey a lot of information in a small space.[2]

[2] This animation provides some good guidelines for formatting tables.

*Ordinal and binary variables.* If a categorical variable has only two options (heads or tails, survived or died), we often call it an indicator, binary, or dummy variable. (These names can be used interchangeably.)

Some categories have a natural ordering, like measures of severity for a hurricane, or responses to a survey about consumer satisfaction. (Has your experience with our call center been Atrocious, Merely Bad, Acceptable, Good, or Excellent?) These are called *ordinal variables*. Ordinal variables differ from numerical variables in that, although they can be placed in a definite order, they cannot be compared using the laws of arithmetic. For example, we can't subtract "Good" from "Excellent" and get a meaningful answer, in the way we can subtract $1000 from $5000 and get a number.

*Relative risk*

The relative risk, sometimes also called the risk ratio, is a widely used measure of association between two categorical variables. To introduce this concept, let's examine a tidbit of data from the PREDIMED trial, a famous study on heart health conducted by Spanish researchers that followed the lifestyle and diet habits of thousands of people over many years, beginning in 2003.[3]

[3] Estruch R, Ros E, Salas-Salvado J, et al. Primary prevention of cardiovascular disease with a Mediterranean diet. N Engl J Med 2013;368:1279-1290. The full text of the article is available at http://www.nejm.org/doi/full/10.1056/NEJMoa1200303

The main purpose of the PREDIMED trial was to assess the effect of a Mediterranean-style diet on the likelihood of someone experiencing a major cardiovascular event (defined by the researchers as a heart attack, stroke, or death from cardiovascular causes). But as part of the study, the researchers also collected data on whether the trial participants were, or had ever been, regular smokers. The table below shows the relationship between smoking and whether someone experienced a cardiovascular event during the study period.

| | Current or former smoker? | |
|---|---|---|
| | No ($n = 3892$) | Yes ($n = 2432$) |
| No event | 3778 | 2294 |
| Event | 114 | 138 |

Let's compare the absolute risk of cardiovascular events for smokers, versus that of non-smokers.[4] Among the smokers, 138 of 2432 people (5.7%) experienced an event; while among the non-smokers, 114 of 3892 people (2.9%) experienced an event. To compute the relative risk of cardiovascular events among smokers, we take the ratio of these two absolute risks:

[4] By "absolute risk," we simply mean the chance of an event happening.

$$\text{Relative risk} = \frac{138/2432}{114/3892} = 1.94\,.$$

This ratio says that smokers were 1.94 times more likely than non-smokers to experience a cardiovascular event during the study.[5]

More generally, for any event (a disease, a car accident, a mortgage default) and any notion of "exposure" to some factor (smoking, driving while texting, poor credit rating), the relative risk is

$$\text{Relative risk} = \frac{\text{Risk of event in exposed group}}{\text{Risk of event in non-exposed group}}.$$

The relative risk tells us how much more (or less) likely the event is in one group versus another. It's important to remember that the relative risk (in our example, 1.94 for smokers) is quite different from the *absolute risk* (in our example, 0.057 for smokers). This distinction is often missed or elided in media coverage of health issues. See, for example, this blog post from the UK's cancer-research funding body about news reports of cancer studies.

[5] Of course, this doesn't prove that the smoking caused the cardiovascular events. One could argue that the smokers may have had other systematically unhealthier habits that did them in instead, and the smoking was merely a marker of these other habits. We'll soon talk about this issue of confounding much more.

## Variation of numerical variables

FIGURE 1.1 depicts a histogram of daily average temperatures in two American cities—San Diego, CA, and Rapid City, SD—for every day from January 1995 to November 2011. Temperature is an example of a *numerical variable*, or something for which numerical comparisons are meaningful (twice as far, six times as fast, $17 cheaper, and so forth). Numerical variables can be *discrete* or *continuous*. Temperature is continuous; we measure it in arbitrarily small increments. Marbles, on the other hand, are discrete; we count them on our fingers and toes.

A histogram is a great way to depict the distribution of a numerical variable. To construct one, we first partition the range of possible outcomes (here, temperatures) into a set of disjoint intervals ("bins"). Next, we count the number of cases that fall into each bin. Finally, we draw a rectangle over each bin whose height is equal to the count within each bin.[6]

[6] Technically this is called a frequency histogram; one could also make a *density histogram* in which the heights of the bars are scaled appropriately so that the total area of all the bars sums to 1.

The histogram in Figure 1.1 suggest two obvious, meaningful questions we can ask about a numerical variable like temperature: where is the middle of the sample, and how much does a typical case vary from the middle?

You're probably already aware of more than more way to answer the question, "Where is the middle?"

- There's the sample mean, written as $\bar{y}$. If we have $n$ data

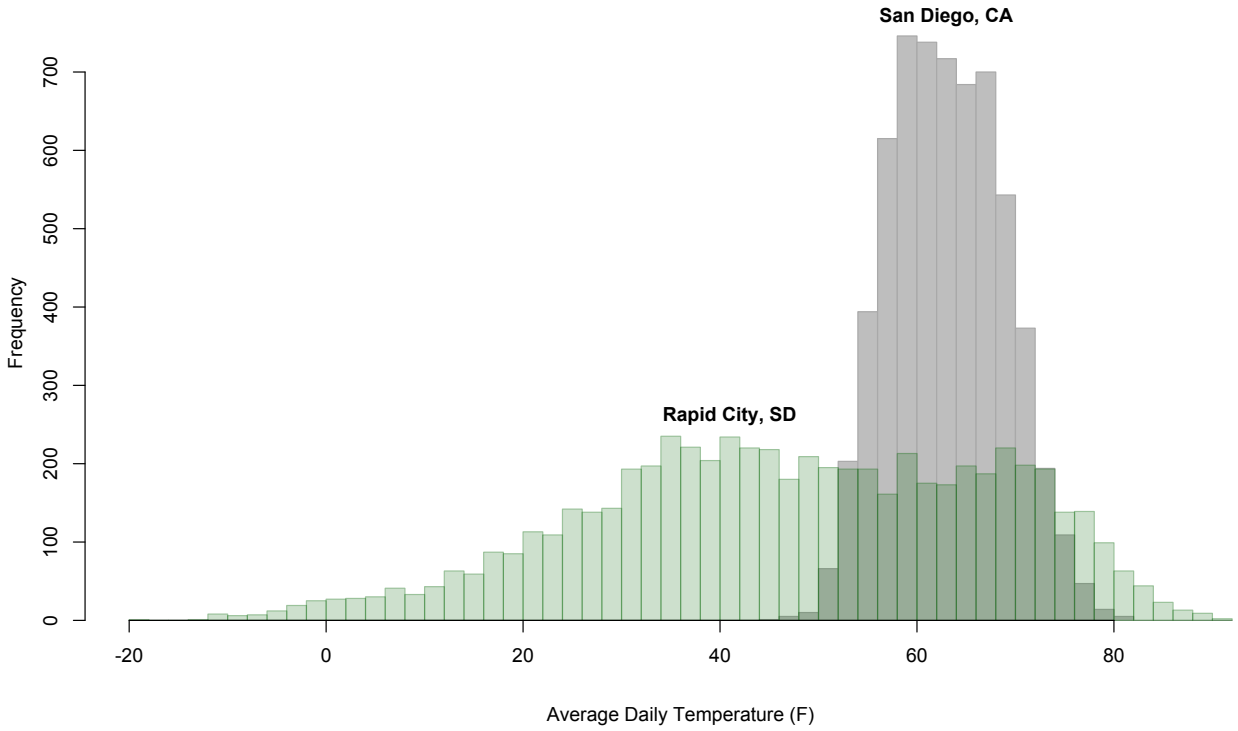



Figure 1.1: Daily average temperatures for San Diego and Rapid City, 1995–2011. These data are visualized in a histogram, which is a simple and effective way to depict the variation of a single numerical variable across many cases.

points $\{y_1, \ldots, y_n\}$, then

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \,.$$

The subscript $i$'s run from case 1 to case $n$, where $n$ is the number of data points in the sample. In many data sets the actual ordering of cases won't matter, and will just reflect the arbitrary ordering of the rows in your data frame.[7]

- There's the median, or the halfway point in a sample.
- There's also the mode, or the most common value.

These different ways of quantifying the middle value all have different properties. For example, the median is less sensitive than the mean to extreme values in your sample; there can be more than one mode in a sample, but only one mean or median.[8]

[7] An obvious exception is in the analysis of time-series data, where the ordering of observations in time may be highly meaningful.

[8] For example, consider the data set $\{1, 2, 3, 3, 4, 4, 5\}$.

*Sample standard deviation and sample variance*

Another important question is, "How spread out are the data points from the middle?" Figure 1.1 drives home the importance

of dispersion in making useful comparisons. Not only are average temperatures lower overall in Rapid City than in San Diego, but they are also a lot more variable: the coldest days are much colder in Rapid City, but the hottest days are hotter, too.

As with the notion of "middle" itself, there is more than one way of quantifying variability, and each way is appropriate for different purposes. Let's follow the line of thinking that leads us to the *standard deviation*, which is probably the most common way of measuring dispersion. Suppose we choose to measure the middle of a sample $y_1, \ldots, y_n$ using the mean, $\bar{y}$. Each case varies from this middle value by its *deviation*, $y_i - \bar{y}$. Why not, therefore, just compute the average deviation from the mean? Well, because

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}) &= \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{n}{n}\bar{y} \\ &= \bar{y} - \bar{y} \\ &= 0\,. \end{aligned}$$

The positives and negatives cancel each other out. We could certainly fix this by taking the absolute value of each deviation, and then averaging those:

$$M = \frac{1}{n}\sum_{i=1}^{n}|y_i - \bar{y}|\,.$$

This quantity is a perfectly sensible measure of the "typical deviation" from the middle. Fittingly enough, it is called the *mean absolute deviation* of the sample.

But it turns out that, for the purposes of statistical modeling, a quantity called the *sample variance* makes more sense:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2\,.$$

That is, we *square* each deviation from $\bar{y}$, rather than take the absolute value. Remember that when we square a negative number, it becomes positive, so that we don't have the problem of the positives and negatives cancelling each other out.

The definition of sample variance raises two questions:

(1) Why do we divide by $n - 1$, when dividing by $n$ would seem to make more sense for computing an average?

(2) Why do we square the deviations, instead of taking absolute values as above?

To answer the first question: we divide by $n-1$ rather than $n$ for obscure technical reasons that, despite what you may read in other statistics textbooks, just aren't that important. (It has to do with "unbiased estimators," which, despite the appealing name, are overrated.) Mainly we use $n-1$ to follow convention.

As for the second question: because sums of squares are special! In all seriousness, there are deep mathematical reasons why we choose to measure dispersion using sums of squared deviations, rather than the seemingly more natural sums of absolute deviations. You'll learn why in a future chapter, but it you want a preview, think about Pythagoras and right triangles. . . .

Of course, computing the sample variance leaves us in the awkward position of measuring variation in the *squared* units of whatever our variable is measured in. This is not intuitive; imagining telling someone that the mean temperature in Rapid City over the last 17 years was 47.3 degrees Fahrenheit, with a sample variance of 402 degrees squared. This is a true statement, but nearly uninterpretable.

Luckily, this is easily fixed by taking the square root of the sample variance, giving us the sample standard deviation:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}\,. \tag{1.1}$$

Now we're back to the original units, and an interpretable measure of "typical deviation from the middle"—for Rapid City, 20.1 degrees. This looks about right from the histogram below; the blue dot is the sample mean, and the blue line stretches 1 sample standard deviation to either side of the mean.

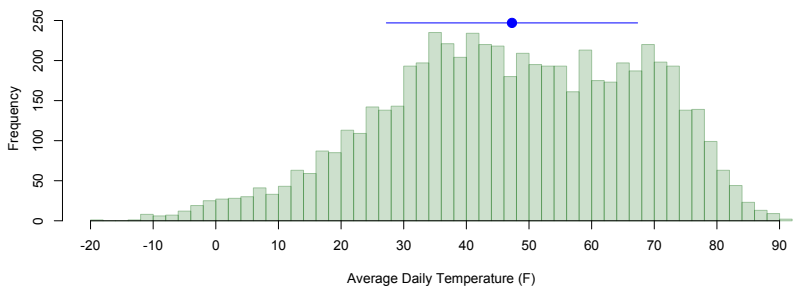



Figure 1.2: The histogram shows average daily temperatures in Rapid City. The blue dot is the sample mean, and the blue line shows an interval encompassing one sample standard deviation to either side of the sample mean.

Two other simple measures of spread are worth mentioning briefly. First, there's the *range*, or the difference between the largest

and smallest values in the sample. There's also the *interquartile range*, or the difference between the $75^{\text{th}}$ and $25^{\text{th}}$ percentiles. This is robust to extreme values, since it involves only the middle 50% of the sample.

*Percentiles, quantiles, and coverage intervals*

Another useful way to summarize the variation of a numerical variable across cases is to compute a set of percentiles, also called quantiles. A familiar example is the median: it happens that exactly 50% of the daily average temperatures in Rapid City fall below fall below 47.6 degrees, and we call this point the median (or the 50th percentile). Similarly, 10% of days in Rapid City are colder than 20.7 degrees, and 90% of days are colder than 73.2 degrees; these are the 10th and 90th percentiles, respectively. A quantile is just a percentile expressed in terms of a decimal fraction; the 80th percentile and 0.8 quantile are the same number.

A common way to summarize a distribution of a numerical variable is to quote a *coverage interval* defined by two percentiles, like the 10th and 90th percentiles (which covers 80% of the cases) or the 2.5th and 97.5th percentiles (which covers 95% of the cases). So, for example, we might quote an 80% coverage interval for daily average temperatures in Rapid City as (20.7, 73.2), whose endpoints are formed from the 10th and 90th percentiles.

*Standardization by z-scoring*

Which temperature is more extreme: 50 degrees in San Diego, or 10 degrees in Rapid City? In an absolute sense, of course 10 degrees is a more extreme temperature. But what about in a relative sense? In other words, is a 10-degree day more extreme *for Rapid City* than a 50-degree day is *for San Diego*? This question could certainly be answered using quantiles, which you've already learned how to handle. But let's discuss a second way: by calculating a z-score for each temperature.

The z-score of some quantity $x$ is the number of standard deviations by which $x$ is above its mean. If a z-score is negative, then the corresponding observation is below the mean.

To calculate a z-score for a number $x$, we subtract the corresponding mean $\mu$ and divide by the standard deviation $\sigma$:

$$z = \frac{x - \mu}{\sigma}.$$

For a 50-degree day in San Diego, this is:

$$z = \frac{50 - 63.1}{5.7} \approx -2.3\,.$$

Or about 2.3 standard deviations below the mean. On the other hand, for a 10-degree day in Rapid City, the z-score is

$$z = \frac{10 - 47.3}{20.1} \approx -1.9\,.$$

Or about 1.9 standard deviations below the mean. Thus a 50-degree day in San Diego is actually more extreme than a 10-degree day in Rapid City! The reason is that temperatures in Rapid City are both colder on average (lower mean) and more variable (higher standard deviation) than temperatures in San Diego.

As this example suggests, z-scores are useful for comparing numbers that come from different distributions, with different statistical properties. It tells you how extreme a number is, relative to other numbers from that some distribution. We often think of the normal distribution as a useful reference here for interpreting $z$-scores. The normal distribution has the property that about 68% of observations fall within $z = 1$ standard deviation of the mean, and about 95% fall within $z = 2$ standard deviations.

### Variation between, and within, groups

A COMMON situation is that we have both categorical and numerical data about each case in a data set. For example, Table 1.4 below shows the average SAT math and verbal scores, stratified by college, for undergraduates in the incoming fall of 2000 freshmen class at the University of Texas at Austin. All 5,191 students who went on to receive a bachelor's degree within 6 years are included; those who dropped out, for whatever reason, are not.

The table tells you something about how the numerical variables (test scores) change depending upon the categorical variable (college), and they are superficially similar to the contingency tables we just encountered. They highlight interesting and useful facts about variation between the groups. Math skills, for example, are probably more important for engineering majors than English majors, and this is reflected in the differences between the group-level means.

| College | Average SAT Math | Average SAT Verbal |
|---|---|---|
| Architecture | 685 | 662 |
| Business | 633 | 597 |
| Communications | 592 | 609 |
| Education | 555 | 546 |
| Engineering | 675 | 606 |
| Fine Arts | 597 | 594 |
| Liberal Arts | 598 | 590 |
| Natural Sciences | 633 | 597 |
| Nursing | 561 | 555 |
| Social Work | 602 | 589 |

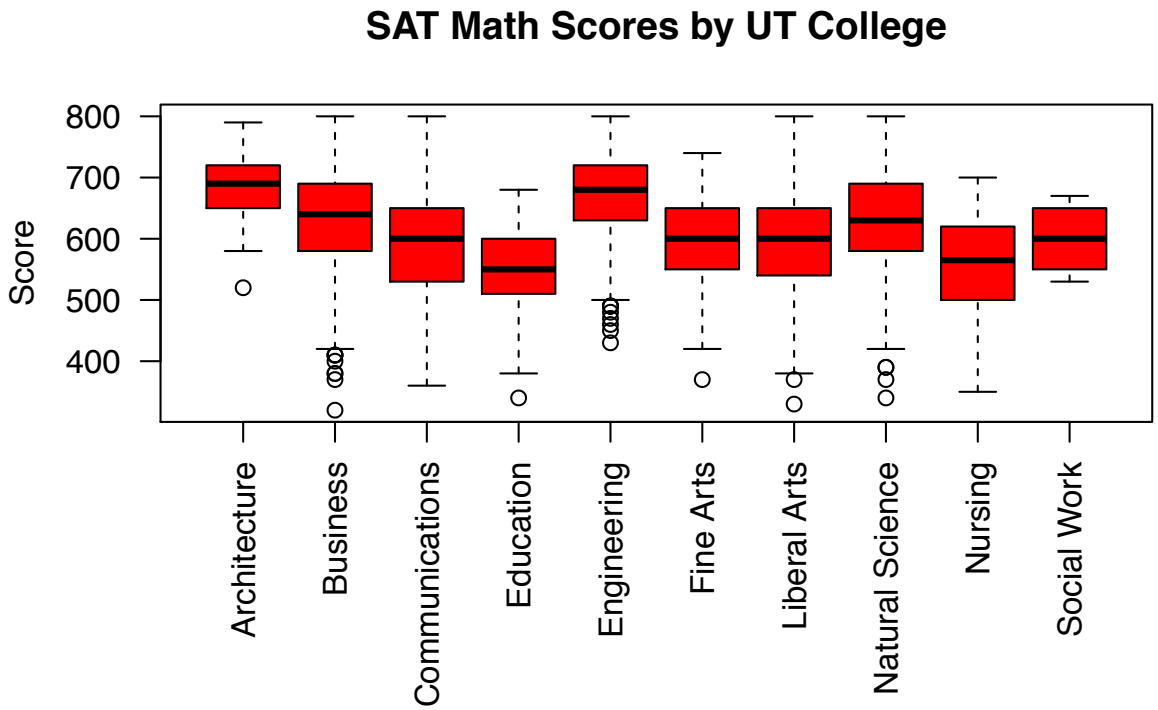
Table 1.4: Average SAT math and verbal scores, stratified by college, for entering freshmen at UT–Austin in the fall of 2000. Collected under the Freedom of Information Act from the state of Texas.

Table 1.4 does differ from a contingency table, however, in one crucial respect: the entries in the table are not counts, but group-level averages. Notice that, to depict between-group variation, the table has reduced each college to a typical case, represented by some hypothetical student who earned the college-wide average SAT scores on both the math and verbal sections. In doing so, it has obscured the underlying variability of students *within* the colleges. But as our example of city temperatures demonstrated, sometimes this variability is an important part of the story as well.

*Boxplots*

This is where boxplots are useful: they allow you to assess variability both between and within the groups. In a boxplot, like the ones shown in Figure 1.3, there is one box per category. (The top panel shows a boxplot for SAT Math scores; the bottom, for SAT Verbal scoers.) Each box shows the *within-group variability*, as measured by the interquartile range of the numerical variable (SAT score) for all cases in that category. The middle line within each box is the median of that category, and the differences between these medians give you a sense of the *between-group variability.* In this boxplot, the whiskers extend outside the box no further than 1.5 times the interquartile range. Points outside this interval are shown as individual dots.

A table like 1.4 focuses exclusively on the between-group variability; it reduces each category to a single number, and shows how those numbers vary from one category to the next. But in



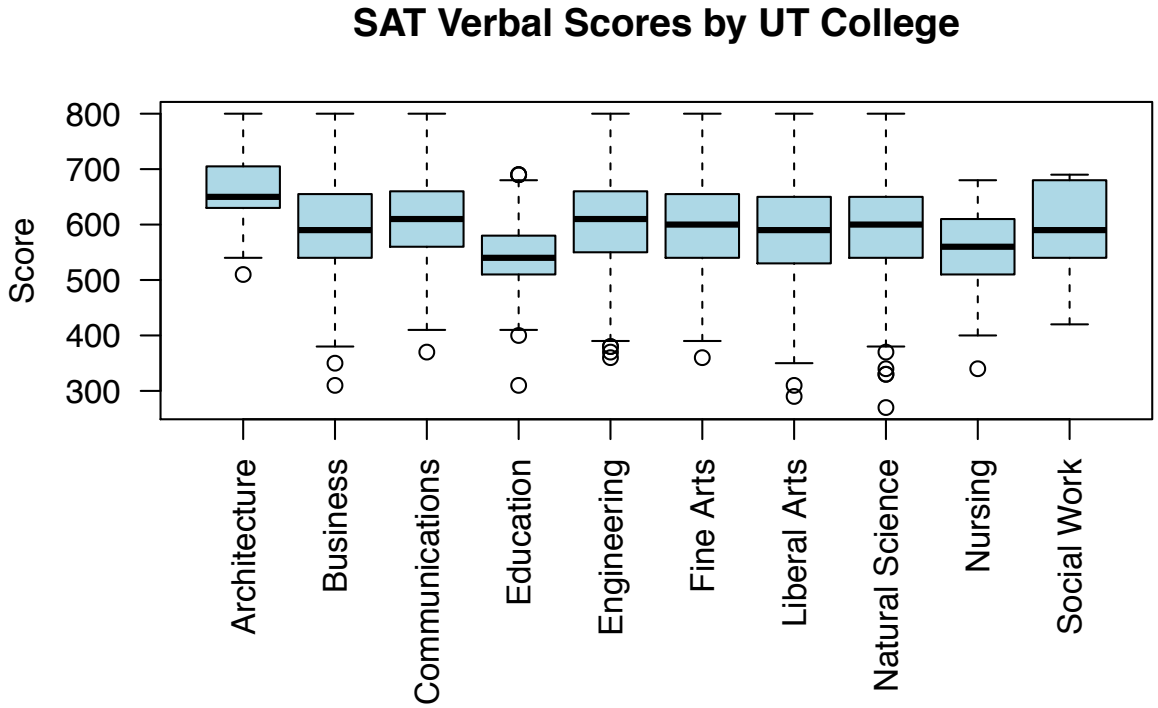




Figure 1.3: Boxplots of the full data set used to form the means in Table 1.4.

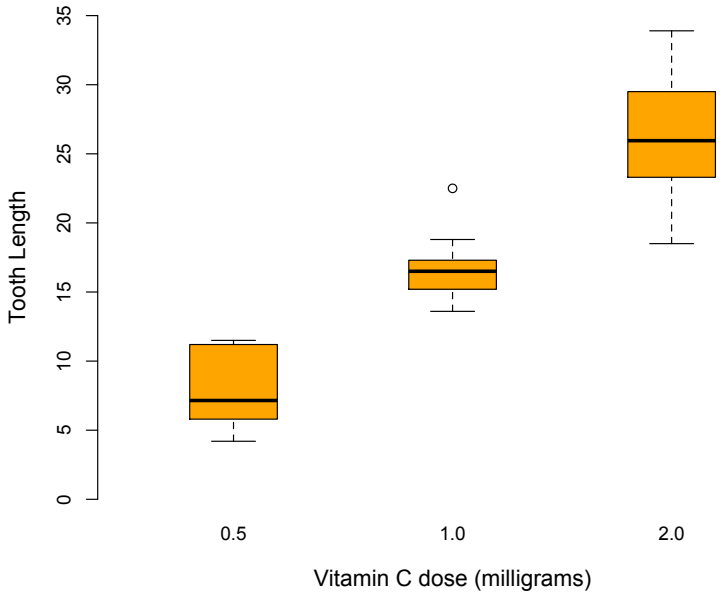



Figure 1.4: For comparison, the table of within-group means is below. Notice how the within-group variability evident in the boxplots at left simply disappears when presented in the form of a summary table, below:

| Dose (mg) | Tooth len. |
|---|---|
| 0.5 | 7.98 |
| 1.0 | 16.77 |
| 2.0 | 26.14 |

many data sets, it is actually the within-group variability that matters most. For example, as Figure 1.3 shows, SAT scores vary much more within a college as they do between colleges. For example, there is 52-point difference in average SAT math scores between Architecture students and Natural Science students. But within Natural Sciences, the interquartile range is nearly twice as large: 100 points.

The situation is quite different Figure 1.4. These boxplots show the growth of guinea pigs' teeth versus their daily dosage of Vitamin C. Like humans, but unlike most other mammals, guinea pigs need Vitamin C to keep rollin', yet they cannot synthesize their own. Their vitamin C intake is strongly predictive of their overall health, measured in this case by the length of their teeth. In this boxplot, we see comparatively more variability between the groups, whose boxplots almost don't overlap.


The same comparison will come up again and again: between-group variability (the differences between typical or average group members) versus within-group variability (the variation of cases within a single group). We'll soon make this comparison mathematically rigorous, but these examples convey the essence of the idea:

- A UT student's college tells you something, though not ev-

erything, about his or her likely SAT scores.

- A guinea pig's Vitamin C regimen tells you something, though not everything, about its tooth growth. But in a relative sense, it tells you more than a UT student's college tells you about his or her SAT scores.

Always remember that a table of group-wise means does not depict "data" as such, but an abstraction of some typical group member. This abstraction may be useful for some purposes. But within-group variability is also important, and may even be the dominant feature of interest. In this case, presenting the group-wise means alone, without the corresponding plots or measures of variability, may obscure more than it reveals.

*Dot plots*

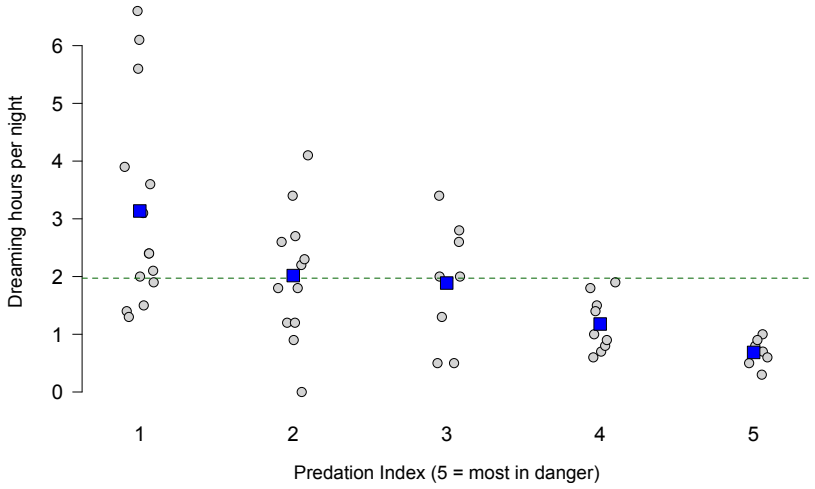
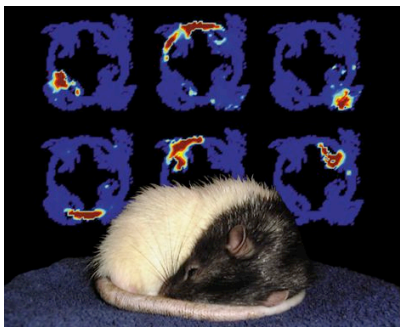



Figure 1.5: Dreaming hours per night versus danger of predation for 50 mammalian species. Data from: "Sleep in Mammals: Ecological and Constitutional Correlates," Allison and Cicchetti (1976). *Science*, November 12, vol. 194, pp. 732-734. Photo of the dreaming critter from the MIT News office (web.mit.edu/newsoffice/2001/dreaming.html).

The *dot plot* is a close cousin of the boxplot. For example, the plot in Figure 1.5 depicts a relationship between the length of a mammal's dreams (as measured in a lab by an MRI machine) and the severity of the danger it faces from predators. Each dot is a single species of mammal—like, for example, the dreaming critter at right. The predation index is an ordinal variable running from 1 (least danger) to 5 (most danger). It accounts both for how likely an animal is to be preyed upon, and how exposed it is when sleeping. Notice the direction of the trend—you'd sleep poorly too if you were worried about being eaten.



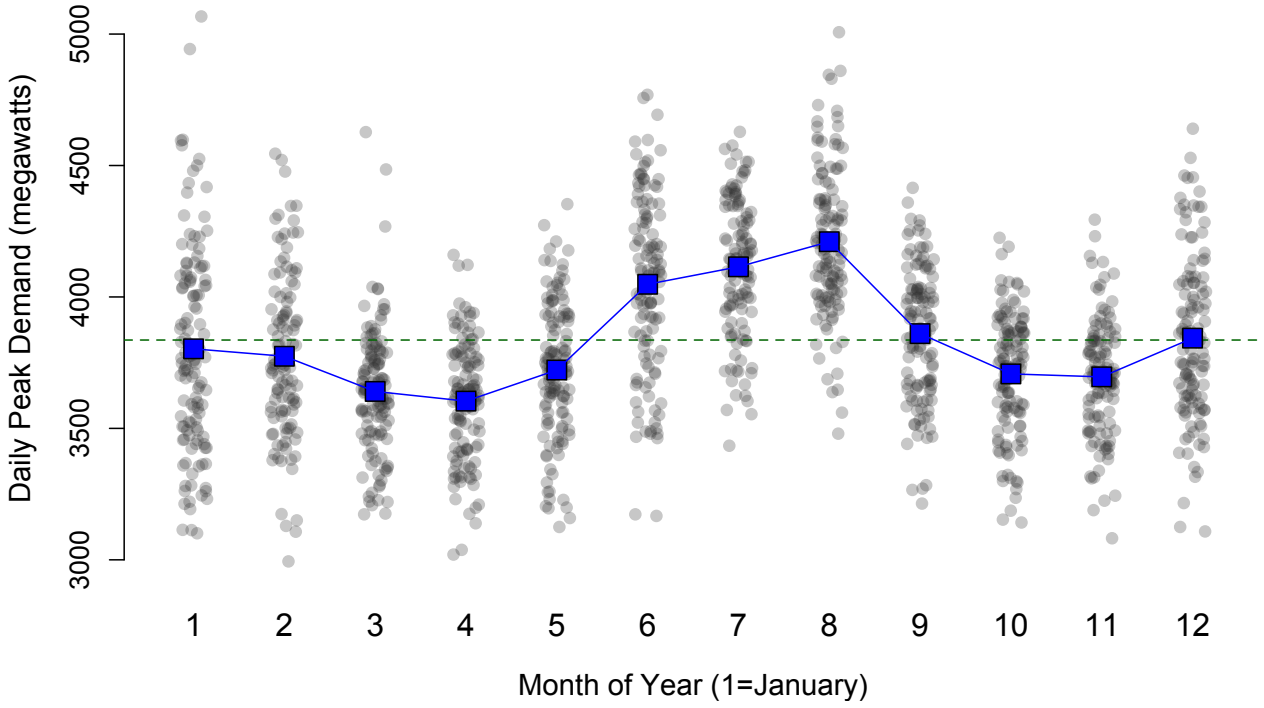



Figure 1.6: Daily peak electricity demand (stratified by month) in Raleigh, NC from 2006–09. The dashed line is the average peak demand for the whole data set, and the blue dots are the month-by-month means.

As you can see, the dot plot is useful for small data sets, when a boxplot is no simpler than just plotting the cases group by group. Strictly speaking, the points should all line up vertically with their corresponding values of predation index, on the $x$-axis. But a small amount of artificial horizontal jitter has been added to the dots, which allows the eye to distinguish the individual cases more easily.

Dot plots can also be effective for larger data sets. In Figure 1.6 we see four years of data on daily peak electricity demand for the city of Raleigh, NC, stratified by month of the year. Both the between-group and within-group variation show up clearly.

*Group means and grand means*

If you looked carefully, you may have noticed two extra features of the dot plots in Figures 1.5 and 1.6. The square blue dots show the *group means* for each category. The dotted green line shows the *grand mean* for the entire data set, irrespective of group identity.

Notice that, in plotting these means along with the data, we have implicitly partitioned the variability:

$$\begin{aligned}\text{Individual case} &= \text{Group mean} + \text{Deviation of that case}\\ \text{Individual case} &= \text{Grand mean} + \text{Deviation of group} + \text{Deviation of that case}\end{aligned}$$

This is just about the simplest statistical model we can fit, but it's still very powerful. We'll revisit it soon.

## More than one numerical variable

Our basic tool for visualizing a bivariate relationship between two numerical variables is the *scatter plot*. Figure 1.7 shows a plot of the daily returns for Microsoft stock versus Apple stock for every trading day in 2015. Every dot corresponds to a day. The location of the dot along the horizontal axis shows the Apple return, and the location on the vertical axis shows the Microsoft return, for that day. In this case, we can see that Microsoft and Apple stocks tend to move up and down together. (Most stocks do.) We can also see the speckling of outliers: those points that are visibly separate

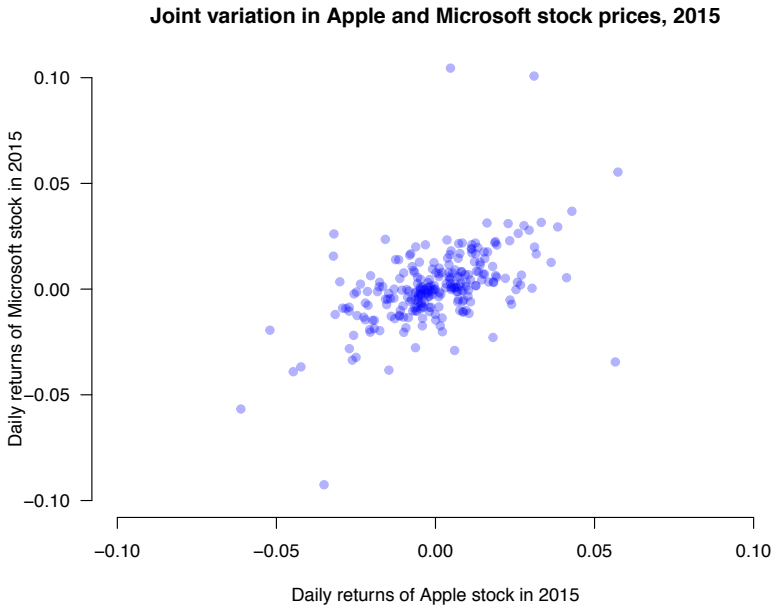



Figure 1.7: A scatter plot of the daily returns for Microsoft stock, versus those of Apple stock, for every trading day in 2015. The daily return is the implied interest rate from holding a stock from the end of one trading day to the end of the next. For example, Apple stock closed at $105.95 per share on January 7th and at $110.02 on January 8th. Thus the return for January 8th was

$$\frac{110.02 - 105.95}{105.95} \approx 0.038\,,$$

or about a 3.8% daily return. On the same day, holders of Microsoft stock enjoyed a 2.9% return.

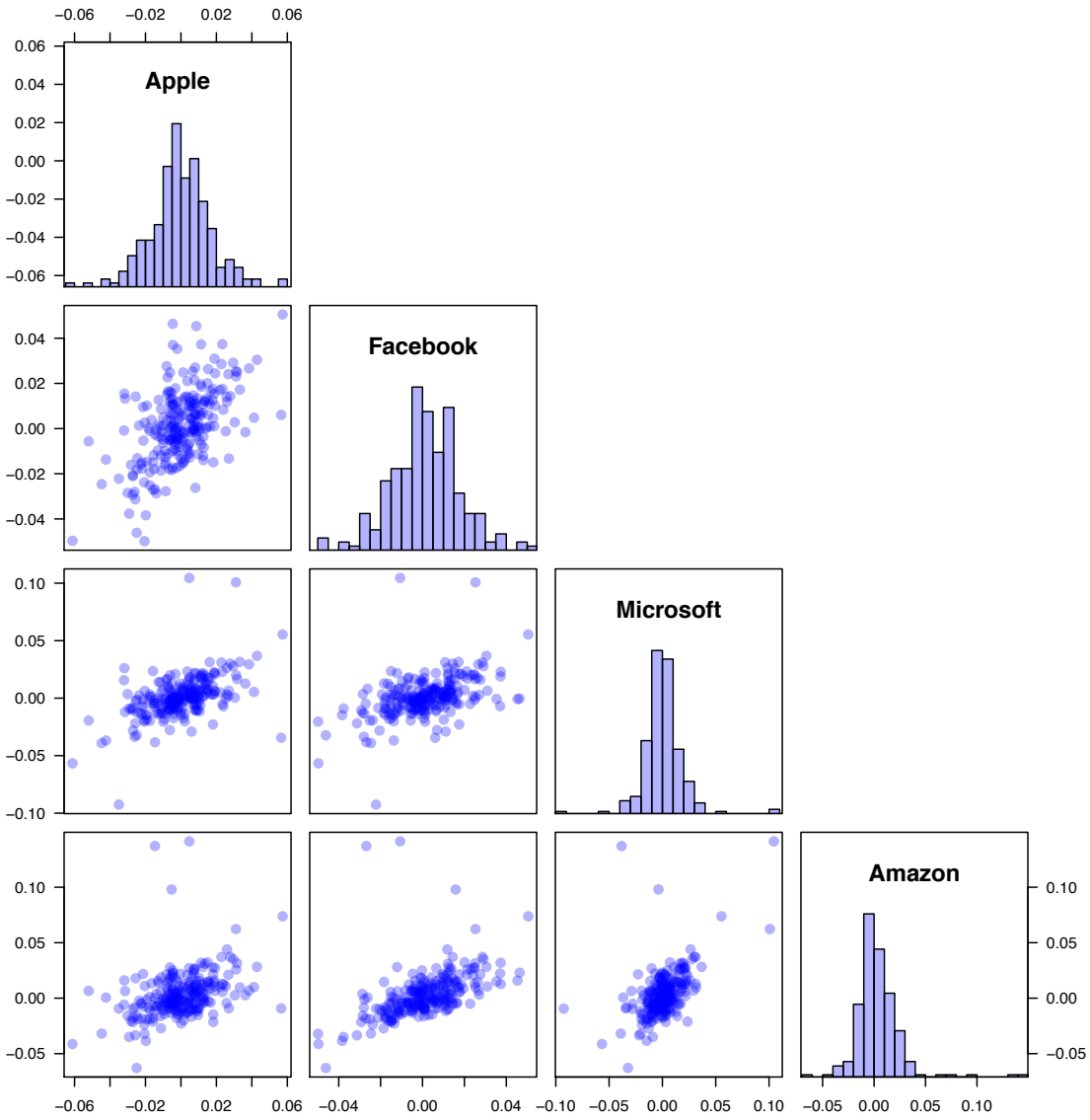



Figure 1.8: A pairs plot: a matrix of four pairwise scatter plots for the daily returns of Apple, Facebook, Microsoft, and Amazon stocks in 2015. The histograms along the diagonal also label the rows and columns of the matrix: e.g. the plot in the second row has Facebook returns along the vertical axis, while the plots in the second column both have Facebook returns along the horizontal axis.

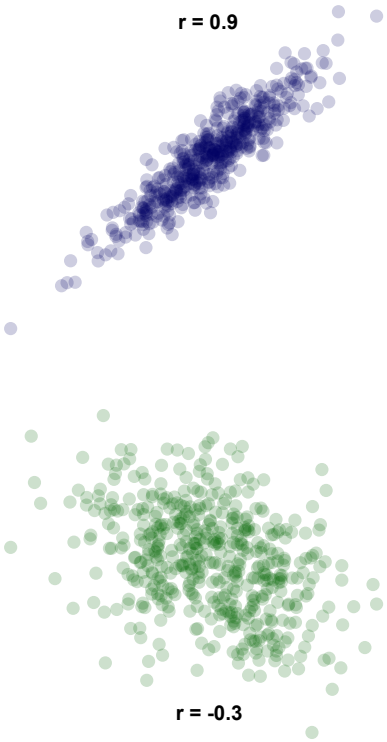
from the main cloud and that represent very good (or bad) days for holders of these two stocks.

A simple way to visualize three or more numerical variables is via a *pairs plot*, as in Figure 1.8. A pairs plot is a matrix of simpler plots, each depicting a bivariate relationship. In Figure 1.8, we see scatterplots for each pair of the daily returns for Microsoft, Facebook, Apple, and Amazon stocks. The histograms on the diagonal serve a dual purpose: (1) they show the variability of each stock in isolation; and (2) they label the rows and columns, so that you know which plots compare which variables.

*Sample correlation.* The *sample correlation coefficient* is a standard measure of the strength of linear dependence between two variables in a sample. If we label the first variable as $x_1, \ldots, x_n$ and the second as $y_1, \ldots, y_n$, then the correlation coefficient is defined as

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}, \tag{1.2}$$

where $s_x$ and $s_y$ are the sample standard deviations of the $X$ and $Y$ variables. At right you see scatter plots that depict examples of strong positive (top) and weak negative (bottom) correlation. Sample correlation is between 1 and $-1$, which are the extremes of perfect positive and perfect negative correlation.




To summarize the correlation among a set of more than two variables, we typically calculate a *correlation matrix* whose entry in row $i$, column $j$ is the correlation between variable $i$ and variable $j$. For the four stocks depicted in Figure 1.8, the correlation matrix is below. Notice that the matrix is symmetric and has ones along the diagonal (because a variable is perfectly correlated with itself):

| | Apple | Microsoft | Facebook | Amazon |
|---|---|---|---|---|
| Apple | 1.00 | 0.52 | 0.55 | 0.36 |
| Microsoft | 0.52 | 1.00 | 0.47 | 0.52 |
| Facebook | 0.55 | 0.47 | 1.00 | 0.50 |
| Amazon | 0.36 | 0.52 | 0.50 | 1.00 |

*Caveats.* A key fact to remember is that correlation measures the strength of *linear* dependence. If two variables don't fall roughly along a straight line in a scatter plot, then correlation can be misleading. For example, consider Figure 1.9: four different data sets, four different stories about what's going on. Yet all have the same correlation: $r = 0.816$.

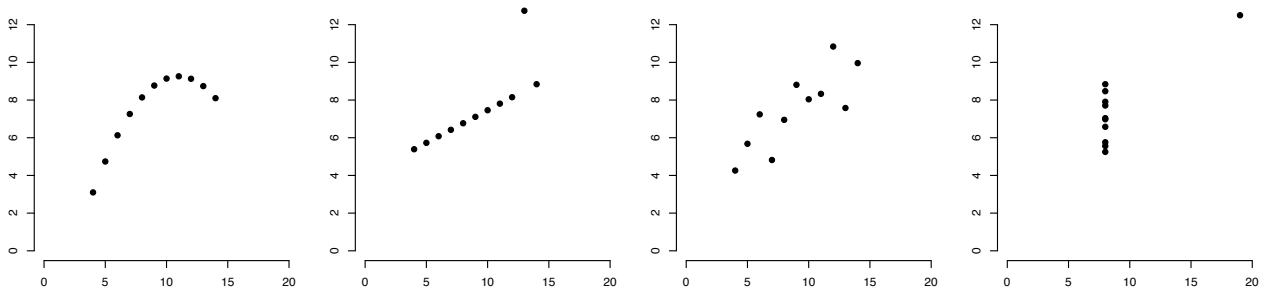


Figure 1.9: above. Data taken from F.J. Anscombe, "Graphs in Statistical Analysis." *American Statistician*, 27 (1973), pp. 17–21

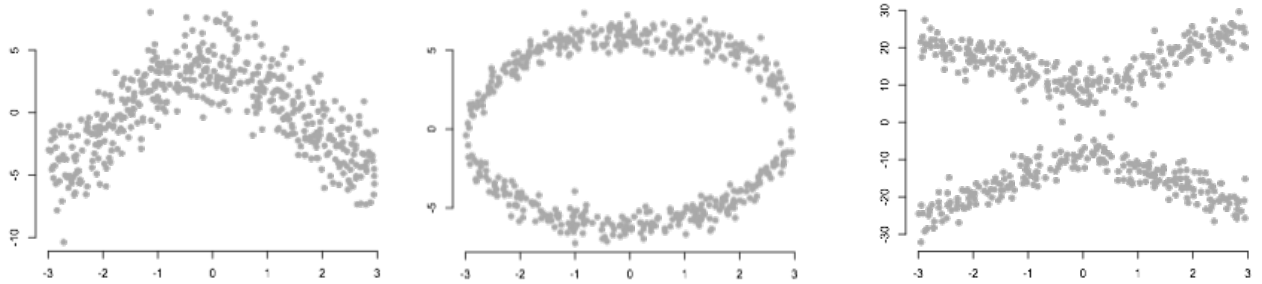


Figure 1.10: left. Each panel shows obvious dependence, but has a sample correlation of $r = 0$.

Another important fact is that a sample correlation of 0 ("uncorrelated") does not necessarily mean that two variables are unrelated. In fact, the correlation coefficient is so intimately tied up with the assumption of a linear relationship that it breaks down entirely when used to quantify the strength of nonlinear relationships. In each of the three plots in Figure 1.10, for example, there is an obvious (nonlinear) relationship between the two variables. Yet the sample correlation coefficient for each of them turns out to be exactly zero.

The lesson of these two plots is that you should always plot your data. After all, a sample correlation coefficient is just one number. It can only tell you so much about the relationship between two variables, and a scatterplot (or boxplot, or dot plot) is a much, much richer summary of that relationship.

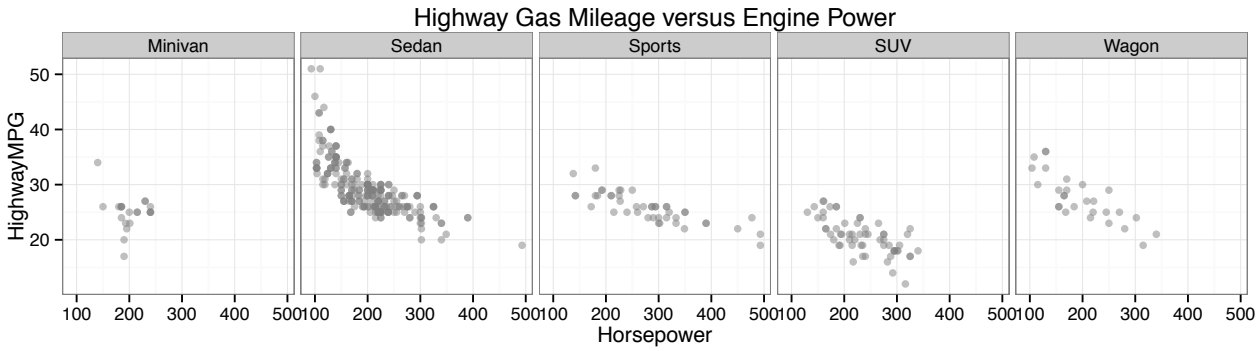



Figure 1.11: Highway gas mileage versus engine power for 387 vehicles in five different classes.

## Further multivariate plots

In visualizing data, we are usually constrained by the limitations of the two-dimensional page or screen. Nevertheless, there are many cool techniques for showing more than two variables at once, despite these limitations.

### *Lattice plots*

Figure 1.11 shows three variables from a data set on 387 vehicles: the highway gas mileage, the engine power (in horsepower), and the class of the vehicle (minivan, sedan, sports car, SUV, or wagon). This is done via a *lattice plot*, which displays the relationship between two variables, stratified by the value of some third variable. In this case the main relationship of interest is between mileage and engine power, and the stratifying variable is vehicle class. Notice how figure 1.11 repeats a scatterplot of MPG versus horsepower five times: one plot for the vehicles in each class. To facilitate comparisons across the strata, both the horizontal and vertical axes are identical in each plot.

Another term for a lattice plot is a trellis plot.

The figure suggests several facts:

- Nobody makes a powerful minivan.
- The overall MPG–horsepower trend is negative for all classes.
- The SUVs have the worst gas mileage overall, and in particular have worse mileage than the sports cars and wagons despite having similar or lower power. (Compare the average vertical location in the SUV panel versus the others).
- The MPG–horsepower relationship becomes nonlinear for

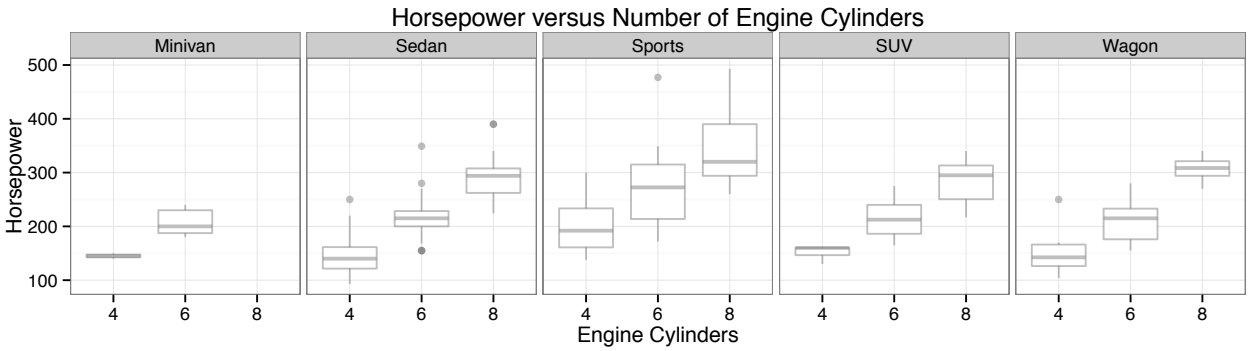



Figure 1.12: Highway gas mileage versus engine power for 387 vehicles in five different classes.

sedans at low horsepower, but perhaps not for wagons.

- As engine power increases, the dropoff in gas mileage looks steeper for SUVs than for sports cars.
- For a fixed level of engine power, there is considerable variability in fuel economy. (Pick a fixed point on the horizontal axis and focus on the cars near there. Now look at the corresponding variability along the vertical axis for those cars.)

We can make a lattice of boxplots as well. For example, Figure 1.12 shows boxplots of engine power versus number of engine cylinders, stratified by vehicle class. This suggests an explanation for the fact that engine power is not a perfect predictor of fuel economy: some cars get more power out of a smaller engine, and are presumably more efficient as a result.

*With a numerical variable.* In Figure 1.11, the stratifying variable is categorical. But we can also stratify a data set according to a numerical variable, by *discretizing* that variable into bins—much in the same way we do when we make a histogram. Figure 1.13 shows the latitude, longitude, and depth (in kilometers) beneath the earth's surface for the epicenter of every earthquake recorded since 1963 near Fiji, an island in the South Pacific Ocean. The "depth" variable has been discretized into nine equal-length bins. The nine panels show the latitude and longitude of the quakes whose depths fell in each interval, labeled at the top of each panel.

As depth increases (going left to right, top to bottom), a spatial pattern emerges. The shallower earthquakes are at the intersection of two major tectonic plates. The deeper quakes emanate from the Tonga Trench—35,702 feet below the sea at its deepest point.[9]

[9] And the final resting place of 3.9 kilograms of radioactive plutonium-238 from the ill-fated Apollo 13 mission.

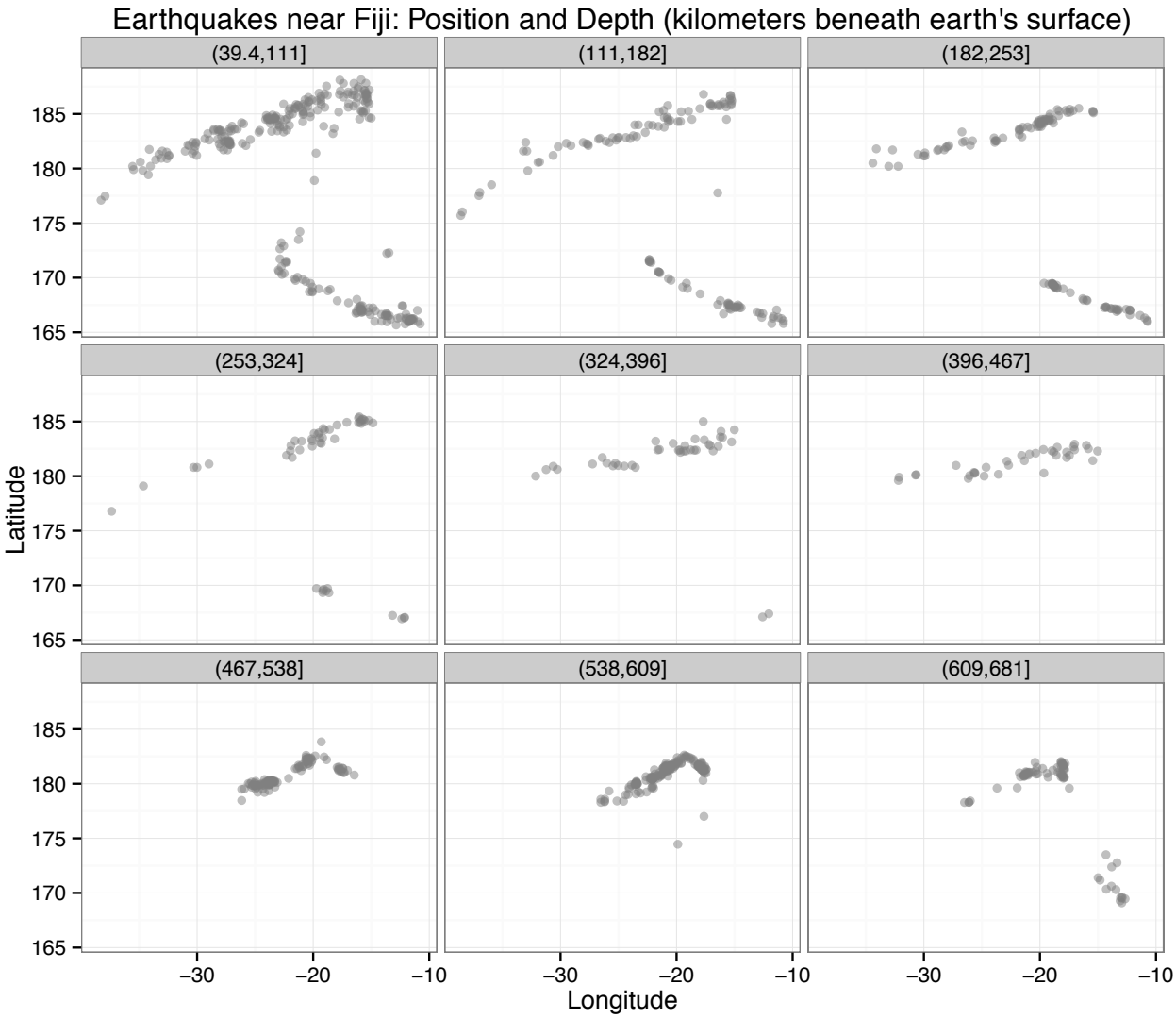



Figure 1.13: Earthquakes in Fiji: latitude versus longitude for quakes within each of nine different depth intervals. Here the range of depths beneath the earth's surface (in kilometers) is labeled at the top of each panel.