lights. The data in **TrafficFlow** show results of one experiment[56] that simulated buses moving along a street and recorded the delay time (in seconds) for both a fixed time and a flexible system of lights. The simulation was repeated under both conditions for a total of 24 trials.

(a) What is the explanatory variable? What is the response variable? Is each categorical or quantitative?

(b) Use technology to find the mean and the standard deviation for the delay times under each of the two conditions (*Timed* and *Flexible*). Does the flexible system seem to reduce delay time?

(c) The data in **TrafficFlow** are paired since we have two values, timed and flexible, for each simulation run. For paired data we generally compute the *difference* for each pair. In this example, the dataset includes a variable called *Difference* that stores the difference *Timed* – *Flexible* for each simulation run. Use technology to find the mean and standard deviation of these differences.

(d) Use technology to draw a boxplot of the differences. Are there any outliers?

DRAW THESE SIDE-BY-SIDE BOXPLOTS

Exercises 2.156 and 2.157 examine issues of location and spread for boxplots. In each case, draw side-by-side boxplots of the datasets on the same scale. There are many possible answers.

[56]Lammer, S. and Helbing, D., "Self-Stabilizing Decentralized Signal Control of Realistic, Saturated Network Traffic," Santa Fe Institute, Santa Fe, NM, working paper No. 10-09-019, September 2010.

2.156 One dataset has median 25, interquartile range 20, and range 30. The other dataset has median 75, interquartile range 20, and range 30.

2.157 One dataset has median 50, interquartile range 20, and range 40. A second dataset has median 50, interquartile range 50, and range 100. A third dataset has median 50, interquartile range 50, and range 60.

2.158 Examine a Relationship in StudentSurvey From the **StudentSurvey** dataset, select any categorical variable and select any quantitative variable. Use technology to create side-by-side boxplots to examine the relationship between the variables. State which two variables you are using and describe what you see in the boxplots. In addition, use technology to compute comparative summary statistics and compare means and standard deviations for the different groups.

2.159 Examine a Relationship in USStates Exercise 2.149 examines the relationship between region of the country and level of physical activity of the population of US states. From the **USStates** dataset, examine a different relationship between a categorical variable and a quantitative variable. Select one of each type of variable and use technology to create side-by-side boxplots to examine the relationship between the variables. State which two variables you are using and describe what you see in the boxplots. In addition, use technology to compute comparative summary statistics and compare means and standard deviations for the different groups.

2.5 TWO QUANTITATIVE VARIABLES: SCATTERPLOT AND CORRELATION

In Section 2.1, we look at relationships between two categorical variables, and in Section 2.4, we investigate relationships between a categorical and a quantitative variable. In this section, we look at relationships between two quantitative variables.

DATA 2.9 **Presidential Approval Ratings and Re-election**

When a US president runs for re-election, how strong is the relationship between the president's approval rating and the outcome of the election? Table 2.26 includes all the presidential elections since 1940 in which an incumbent was running and shows the presidential approval rating at the time of the election and the margin of victory or defeat for the president in the election.[57] The data are available in **ElectionMargin**. ■

[57]Data obtained from *http://www.fivethirtyeight.com* and *http://www.realclearpolitics.com.*

Table 2.26 *Presidential approval rating and margin of victory or defeat*

Year	Candidate	Approval	Margin	Result
1940	Roosevelt	62	10.0	Won
1948	Truman	50	4.5	Won
1956	Eisenhower	70	15.4	Won
1964	Johnson	67	22.6	Won
1972	Nixon	57	23.2	Won
1976	Ford	48	−2.1	Lost
1980	Carter	31	−9.7	Lost
1984	Reagan	57	18.2	Won
1992	G. H. W. Bush	39	−5.5	Lost
1996	Clinton	55	8.5	Won
2004	G. W. Bush	49	2.4	Won
2012	Obama	50	3.9	Won

Example 2.32

(a) What was the highest approval rating for any of the losing presidents? What was the lowest approval rating for any of the winning presidents? Make a conjecture about the approval rating needed by a sitting president in order to win re-election.

(b) Approval rating and margin of victory are both quantitative variables. Does there seem to be an association between the two variables?

Solution

(a) Three presidents lost, and the highest approval rating among them is 48%. Nine presidents won, and the lowest approval rating among them is 49%. It appears that a president needs an approval rating of 49% or higher to win re-election.

(b) In general, it appears that a higher approval rating corresponds to a larger margin of victory, although the association is not perfect.

Visualizing a Relationship between Two Quantitative Variables: Scatterplots

The standard way to display the relationship between two quantitative variables is to extend the notion of a dotplot for a single quantitative variable to a two-dimensional graph known as a *scatterplot*. To examine a relationship between two quantitative variables, we have *paired* data, where each data case has values for both of the quantitative variables.

Scatterplot

A **scatterplot** is a graph of the relationship between two quantitative variables.

A scatterplot includes a pair of axes with appropriate numerical scales, one for each variable. The paired data for each case are plotted as a point on the scatterplot. If there are explanatory and response variables, we put the explanatory variable on the horizontal axis and the response variable on the vertical axis.

Example 2.33

Draw a scatterplot for the data on approval rating and margin of victory in Table 2.26.

Solution

We believe approval ratings may help us predict the margin of victory, so the explanatory variable is approval rating and the response variable is margin of victory. We put approval rating on the horizontal axis and margin of victory on the vertical axis. The 12 data pairs are plotted as 12 points in the scatterplot of Figure 2.48. The point corresponding to Roosevelt in 1940, with an approval rating of 62% and a margin of victory of 10 points, is indicated. We notice from the upward trend of the points that the margin of victory does tend to increase as the approval rating increases.

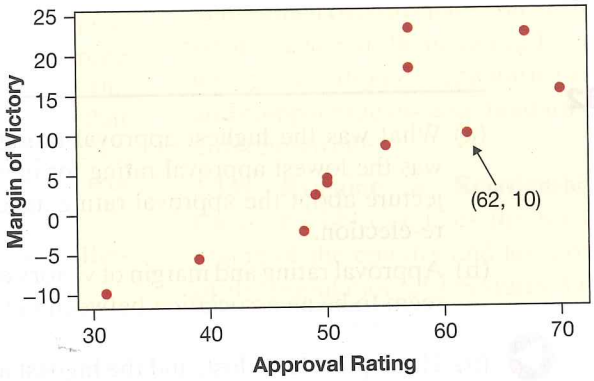


Figure 2.48 *Approval rating and margin of victory*

Interpreting a Scatterplot

When looking at a scatterplot, we often address the following questions:

- Do the points form a clear trend with a particular direction, are they more scattered about a general trend, or is there no obvious pattern?
- If there is a trend, is it generally upward or generally downward as we look from left to right? A general upward trend is called a *positive* association while a general downward trend is called a *negative* association.
- If there is a trend, does it seem to follow a straight line, which we call a *linear association*, or some other curve or pattern?
- Are there any outlier points that are clearly distinct from a general pattern in the data?

For the presidential re-election data in Figure 2.48, we see a positive association since there is an upward trend in margin of victory as approval increases. While the points certainly do not all fall exactly on a straight line, we can imagine drawing a line to match the general trend of the data. There is a general linear trend, and it is a relatively strong association.

Example 2.34

Scatterplots Using Data from Florida Lakes

Four scatterplots are shown in Figure 2.49 using data from the **FloridaLakes** dataset, introduced in Data 2.4 on page 71. For each pair of variables, discuss the information contained in the scatterplot. If there appears to be a positive or negative association, discuss what that means in the specific context.

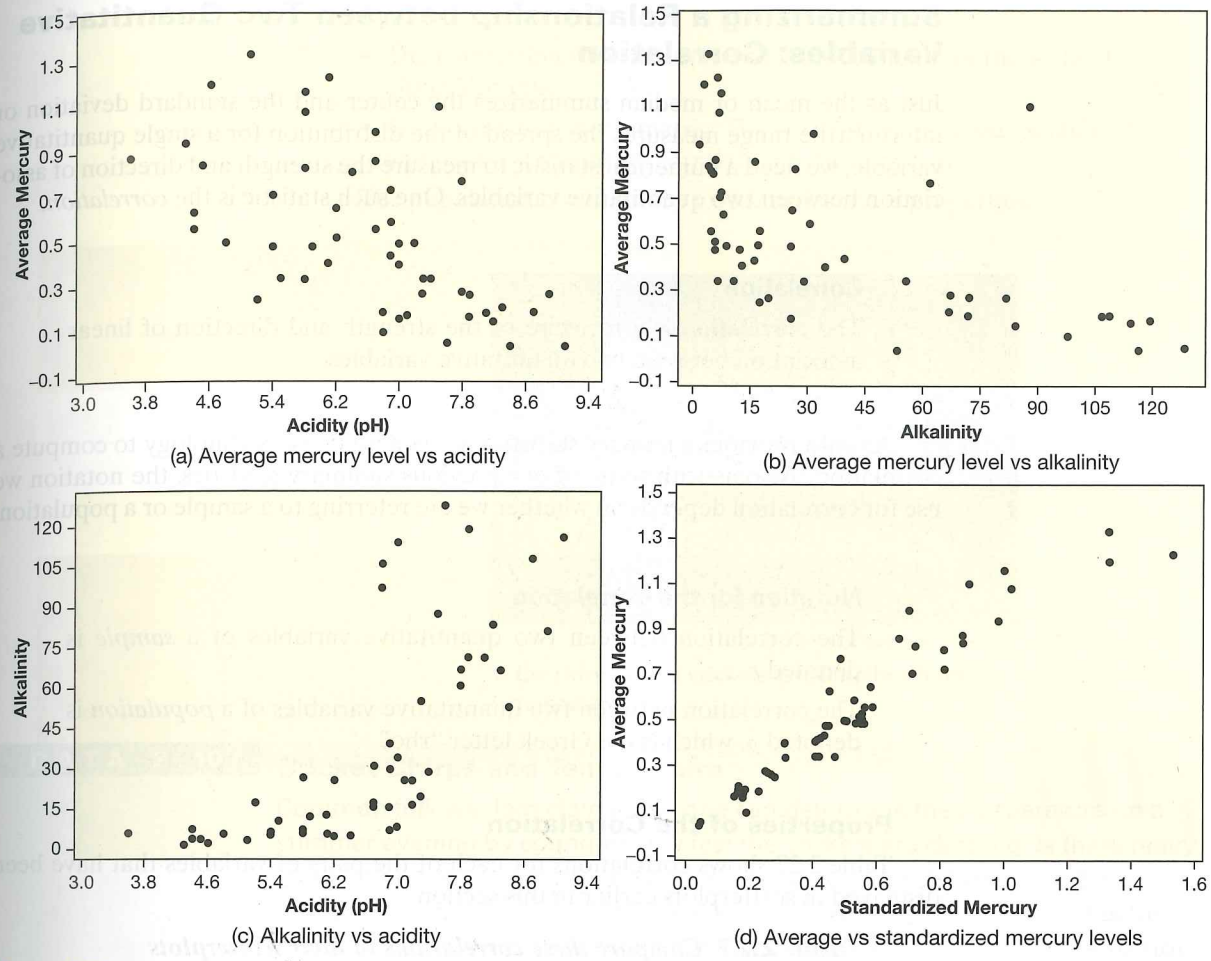


Figure 2.49 *Scatterplots of data from Florida lakes*

Solution

(a) Acidity appears to have a negative linear association with average mercury level, but not a strong one as the points are scattered widely around any straight line. Since the association is negative, larger values of acidity tend to be associated with smaller levels of mercury.

(b) Alkalinity also is negatively associated with average mercury level, with a slightly stronger association along a more curved trend. One lake with a high average mercury level around 1.1 ppm also has a high alkalinity at almost 90 mg/L and is clearly away from the general trend of the data. Note that neither of the values for this lake would be considered outliers for the individual variables, but the data pair stands out in the scatterplot so it is considered an outlier. Since the association is negative, larger values of alkalinity tend to be associated with smaller levels of mercury.

(c) There is a positive association since the acidity increases with alkalinity along a curved pattern. Since the association is positive, larger values of acidity tend to be associated with larger values of alkalinity.

(d) The average mercury levels show a strong positive association with the standardized mercury levels that fit fairly closely to a straight line. Since the association is positive, larger levels of standardized mercury tend to be associated with larger levels of mercury.

Summarizing a Relationship between Two Quantitative Variables: Correlation

Just as the mean or median summarizes the center and the standard deviation or interquartile range measures the spread of the distribution for a single quantitative variable, we need a numerical statistic to measure the strength and direction of association between two quantitative variables. One such statistic is the *correlation*.

Correlation

The **correlation** is a measure of the strength and direction of linear association between two quantitative variables.

As with previous summary statistics, we generally use technology to compute a correlation. Also as with some of our previous summary statistics, the notation we use for correlation depends on whether we are referring to a sample or a population.

Notation for the Correlation

The correlation between two quantitative variables of a *sample* is denoted r.

The correlation between two quantitative variables of a *population* is denoted ρ, which is the Greek letter "rho."

Properties of the Correlation

Table 2.27 shows correlations for each of the pairs of variables that have been displayed in scatterplots earlier in this section.

Table 2.27 *Compare these correlations to their scatterplots*

Variable 1	Variable 2	Correlation
Margin of victory	Approval rating	0.86
Average mercury	Acidity	−0.58
Average mercury	Alkalinity	−0.59
Alkalinity	Acidity	0.72
Average mercury	Standardized mercury	0.96

Notice that all the correlations in the table are between −1 and +1. We see that a positive correlation corresponds to a positive association and a negative correlation corresponds to a negative association. Notice also that correlation values closer to 1 or −1 correspond to stronger linear associations. We make these observations more precise in the following list of properties.

Properties of the Correlation

The sample correlation r has the following properties:

- Correlation is always between −1 and 1, inclusive: $-1 \le r \le 1$.
- The sign of r (positive or negative) indicates the direction of association.
- Values of r close to +1 or −1 show a strong linear relationship, while values of r close to 0 show no linear relationship.

- The correlation r has no units and is independent of the scale of either variable.
- The correlation is symmetric: The correlation between variables x and y is the same as the correlation between y and x.

The population correlation ρ also satisfies these properties.

Is the chirp rate of crickets associated with the temperature?

DATA 2.10

Cricket Chirps and Temperature

Common folk wisdom claims that one can determine the temperature on a summer evening by counting how fast the crickets are chirping. Is there really an association between chirp rate and temperature? The data in Table 2.28 were collected by E. A. Bessey and C. A. Bessey,[58] who measured chirp rates for crickets and temperatures during the summer of 1898. The data are also stored in **CricketChirps**. ■

Table 2.28 *Cricket chirps and temperature*

Temperature (° F)	54.5	59.5	63.5	67.5	72.0	78.5	83.0
Chirps (per minute)	81	97	103	123	150	182	195

Example 2.35

A scatterplot of the data in Table 2.28 is given in Figure 2.50.

(a) Use the scatterplot to estimate the correlation between chirp rate and temperature. Explain your reasoning.

(b) Use technology to find the correlation and use correct notation.

(c) Are chirp rate and temperature associated?

Solution

(a) Figure 2.50 shows a very strong positive linear trend in the data, so we expect the correlation to be close to +1. Since the points do not all lie exactly on a line, the correlation will be slightly less than 1.

[58] Bessey, E. A. and Bessey, C. A., "Further Notes on Thermometer Crickets," *American Naturalist*, 1898; 32: 263–264.

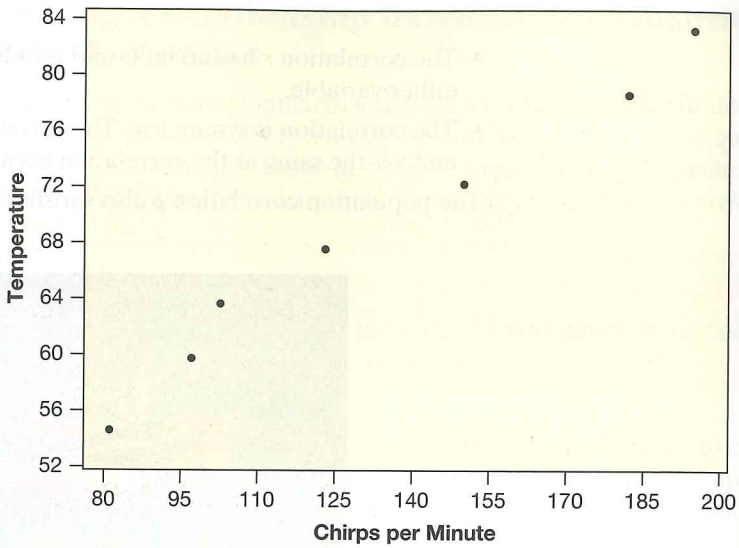


Figure 2.50 *Scatterplot of chirp rate and temperature*

(b) We use the notation r for this sample correlation. Using technology, we see that $r = 0.99$, matching the very strong positive linear relationship we see in the scatterplot.

(c) Yes, cricket chirp rates and air temperature are strongly associated!

Correlation Cautions

Example 2.36

Figure 2.51 shows the estimated average life expectancy[59] (in years) for a sample of 40 countries against the average amount of fat[60] (measured in grams per capita per day) in the food supply for each country. The scatterplot shows a clear positive

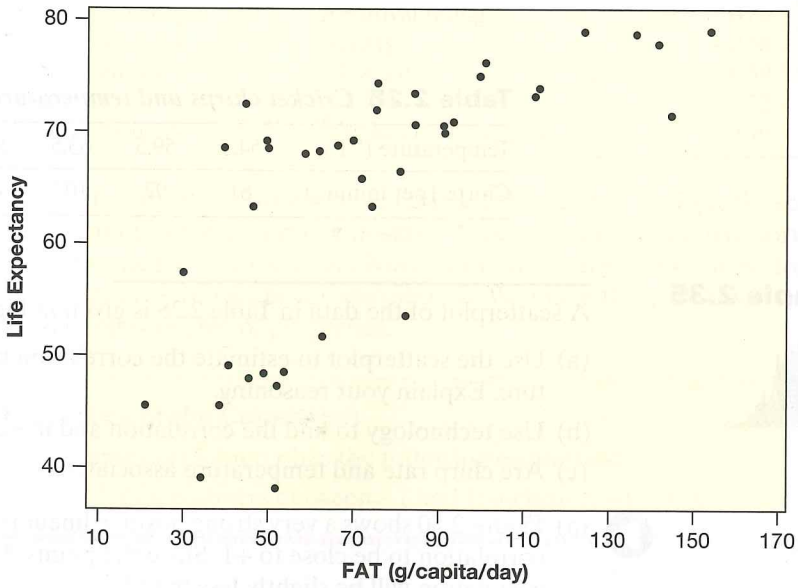


Figure 2.51 *Life expectancy vs grams of fat in daily diet for 40 countries*

[59]United Nations Development Program, *Human Development Report 2003.*
[60]Food and Agriculture Organization of the United Nations.

association ($r = 0.70$) between these two variables. The countries with short life expectancies all have below-average fat consumption, while the countries consuming more than 100 grams of fat on average all have life expectancies over 70 years. Does this mean that we should eat more fat in order to live longer?

Solution

No! Just because there is a strong association between these two variables, it would be inappropriate to conclude that changing one of them (for example, increasing fat in the diet) would *cause* a corresponding change in the other variable (lifetime). An observational study was used to collect these data, so we cannot conclude that there is a causal relationship. One likely confounding variable is the wealth of the country, which is associated with both life expectancy and fat consumption.

A strong correlation does not necessarily imply a causal association! As we saw in Chapter 1, we need to be aware of confounding variables and we need to pay attention to whether the data come from an experiment or an observational study.

Correlation Caution #1

A strong positive or negative correlation does not (necessarily) imply a cause and effect relationship between the two variables.

Example 2.37

Core body temperature for an individual person tends to fluctuate during the day according to a regular circadian rhythm. Suppose that body temperatures for an adult woman are recorded every hour of the day, starting at 6 am. The results are shown in Figure 2.52. Does there appear to be an association between the time of day and body temperature? Estimate the correlation between the hour of the day and the woman's body temperature.

Solution

There is a regular pattern with temperatures rising in the morning, staying fairly constant throughout the day, and then falling at night, so these variables are associated. Despite this association, the correlation between these two variables will be

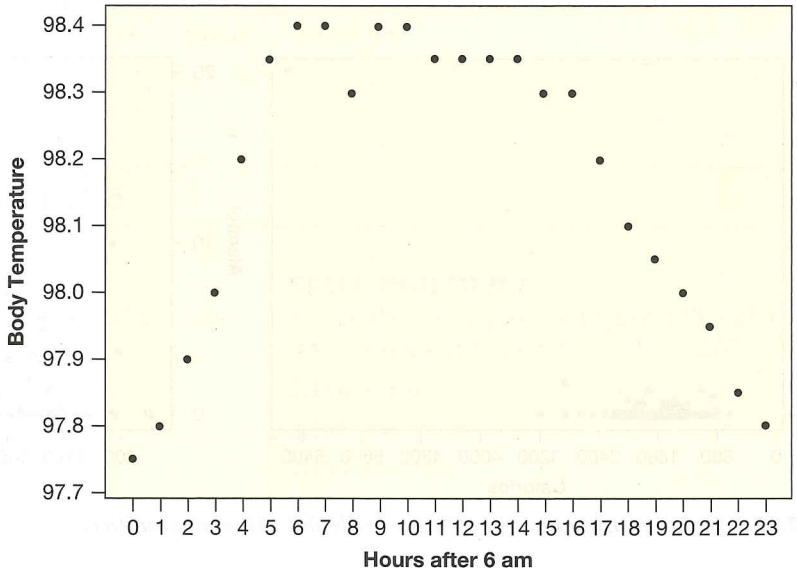


Figure 2.52 *Hourly body temperatures*

near zero. (For Figure 2.52 the actual correlation is $r = -0.08$.) The beginning hours appear to have a positive association but the trend is negative for the later hours. Remember that correlation measures the strength of a *linear* relationship between two variables.

Correlation Caution #2

A correlation near zero does not (necessarily) mean that the two variables are not associated, since the correlation measures only the strength of a *linear* relationship.

DATA 2.11 Effects of Diet on Levels of Retinol and Beta-carotene

In a study on the association between diet and levels of retinol and beta-carotene in the blood stream, researchers recorded a variety of dietary and demographic variables for the subjects. Variables include alcohol consumption, average daily calories, age, gender, multivitamin use, fat grams per day, fiber grams per day, smoking habits, etc. The data are available in **NutritionStudy**. ■

Example 2.38

Figure 2.53 shows the alcohol consumption (drinks per week) and average daily caloric intake for 91 subjects who are at least 60 years old, from the data in **NutritionStudy**. Notice the distinct outlier who claims to down 203 drinks per week as part of a 6662 calorie diet! This is almost certainly an incorrect observation. The second plot in Figure 2.53 shows these same data with the outlier omitted. How do you think the correlation between calories and alcohol consumption changes when the outlier is deleted?

Solution

The correlation between alcohol consumption and daily calories is $r = 0.72$ with the outlier present, but only $r = 0.15$ when that data point is omitted. What initially might look like a strong association between alcohol consumption and daily calories turns out to be much weaker when the extreme outlier is removed.

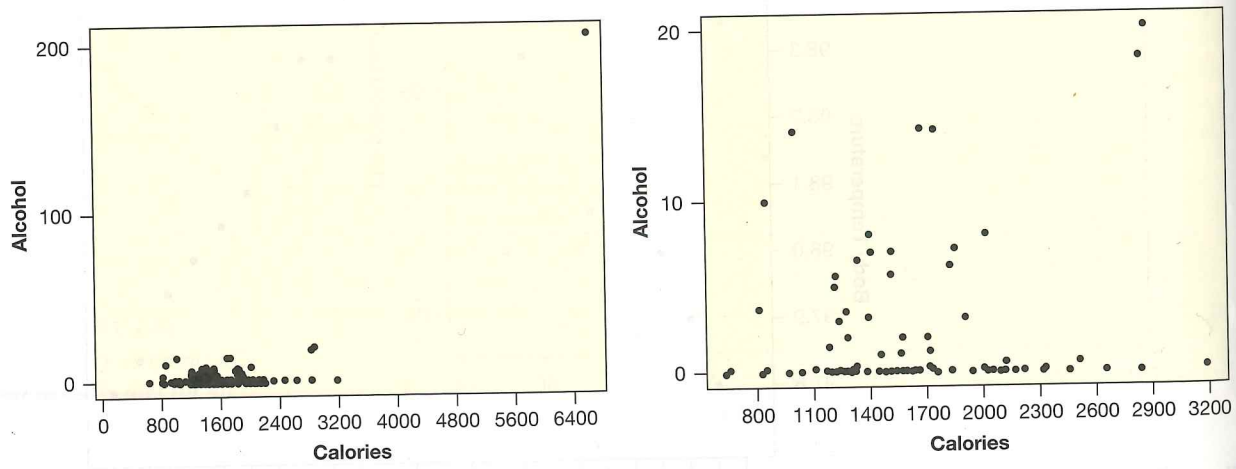


Figure 2.53 *Alcohol consumption vs calories (with and without an outlier)*

Correlation Caution #3

Correlation can be heavily influenced by outliers. Always plot your data!

A Formula for Correlation

We routinely rely on technology to compute correlations, but you may be wondering how such computations are done. While computing a correlation "by hand" is tedious and often not very informative, a formula, such as the one shown below, can be helpful in understanding how the correlation works:

$$r = \frac{1}{n-1}\sum\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$$

Essentially this involves converting all values for both variables to z-scores, which puts the correlation on a fixed -1 to $+1$ scale and makes it independent of the scale of measurement. For a positive association, large values for x tend to occur with large values of y (both z-scores are positive) and small values (with negative z-scores) tend to occur together. In either case the products are positive, which leads to a positive sum. For a negative association, the z-scores tend to have opposite signs (small x with large y and vice versa) so the products tend to be negative.

SECTION LEARNING GOALS

You should now have the understanding and skills to:

- Describe an association displayed in a scatterplot
- Explain what a positive or negative association means between two variables
- Interpret a correlation
- Use technology to find a correlation
- Recognize that correlation does not imply cause and effect
- Recognize that you should always plot your data in addition to interpreting numerical summaries

Exercises for Section 2.5

SKILL BUILDER 1

Match the scatterplots in Figure 2.54 with the correlation values in Exercises 2.160 to 2.163.

2.160 $r = -1$

2.161 $r = 0$

2.162 $r = 0.8$

2.163 $r = 1$

SKILL BUILDER 2

Match the scatterplots in Figure 2.55 with the correlation values in Exercises 2.164 to 2.167.

2.164 $r = 0.09$

2.165 $r = -0.38$

2.166 $r = 0.89$

2.167 $r = -0.81$

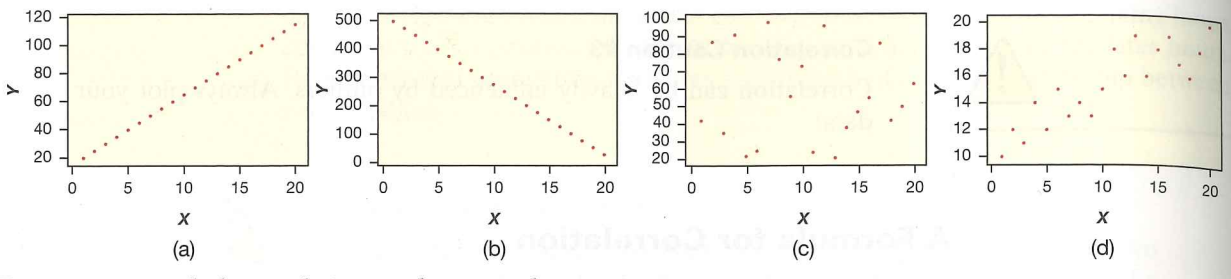


Figure 2.54 *Match the correlations to the scatterplots*

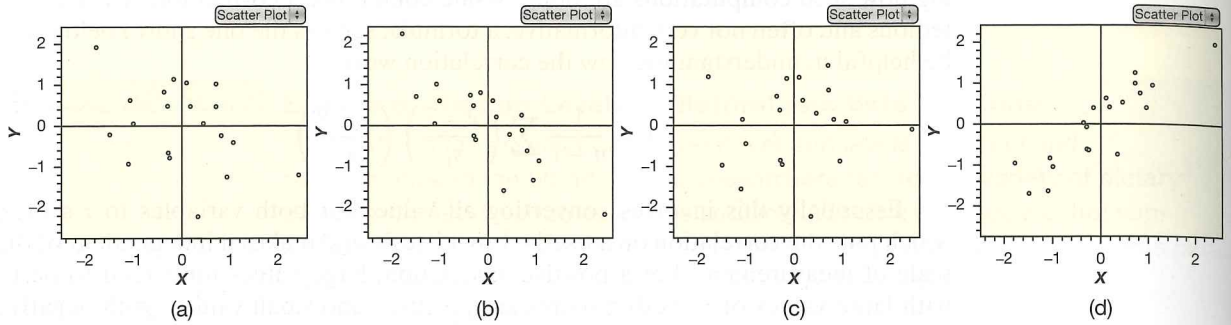


Figure 2.55 *Match the correlations to the scatterplots*

SKILL BUILDER 3

In Exercises 2.168 to 2.173, two quantitative variables are described. Do you expect a positive or negative association between the two variables? Explain your choice.

2.168 Size of a house *and* Cost to heat the house.

2.169 Distance driven since the last fill-up of the gas tank *and* Amount of gas left in the tank.

2.170 Outside temperature *and* Amount of clothes worn.

2.171 Number of text messages sent on a cell phone *and* Number of text messages received on the phone.

2.172 Number of people in a square mile *and* Number of trees in the square mile.

2.173 Amount of time spent studying *and* Grade on the exam.

SKILL BUILDER 4

In Exercises 2.174 and 2.175, make a scatterplot of the data. Put the X variable on the horizontal axis and the Y variable on the vertical axis.

2.174

X	3	5	2	7	6
Y	1	2	1.5	3	2.5

2.175

X	15	20	25	30	35	40	45	50
Y	532	466	478	320	303	349	275	221

SKILL BUILDER 5

In Exercises 2.176 and 2.177, use technology to find the correlation for the data indicated.

2.176 The data in Exercise 2.174.

2.177 The data in Exercise 2.175.

2.178 Light Roast or Dark Roast for Your Coffee? A somewhat surprising fact about coffee is that the longer it is roasted, the less caffeine it has. Thus an "extra bold" dark roast coffee actually has less caffeine than a light roast coffee. What is the explanatory variable and what is the response variable? Do the two variables have a negative association or a positive association?

2.179 Mother's Love, Hippocampus, and Resiliency Multiple studies[61] in both animals and humans show the importance of a mother's love (or the unconditional love of any close person to a child) in a child's brain development. A recent study shows that children with nurturing mothers had a substantially larger area of the brain called the hippocampus than children with less nurturing mothers. This is important because other studies have shown that the size of the hippocampus matters: People with large hippocampus area are more resilient and are more likely to be able to weather the stresses and strains of daily life. These observations come from experiments in animals and observational studies in humans.

[61] Raison, C., "Love key to brain development in children," cnn.com, *The Chart*, March 12, 2012.

(a) Is the amount of maternal nurturing one receives as a child positively or negatively associated with hippocampus size?

(b) Is hippocampus size positively or negatively associated with resiliency and the ability to weather the stresses of life?

(c) How might a randomized experiment be designed to test the effect described in part (a) in humans? Would such an experiment be ethical?

(d) Can we conclude that maternal nurturing in humans causes the hippocampus to grow larger? Can we conclude that maternal nurturing in animals (such as mice, who were used in many of the experiments) causes the hippocampus to grow larger? Explain.

2.180 Commitment Genes and Cheating Genes In earlier studies, scientists reported finding a "commitment gene" in men, in which men with a certain gene variant were much less likely to commit to a monogamous relationship.[62] That study involved only men (and we return to it later in this text), but a new study, involving birds this time rather than humans, shows that female infidelity may be inherited.[63] Scientists recorded who mated with or rebuffed whom for five generations of captive zebra finches, for a total of 800 males and 754 females. Zebra finches are believed to be a monogamous species, but the study found that mothers who cheat with multiple partners often had daughters who also cheat with multiple partners. To identify whether the effect was genetic or environmental, the scientists switched many of the chicks from their original nests. More cheating by the biological mother was strongly associated with more cheating by the daughter. Is this a positive or negative association?

[62]Timmer, J., "Men with genetic variant struggle with commitment," *http://www.arstechnica.com*, reporting on a study in *Proceedings of the National Academy of Science*, 2009.

[63]Millus, S., "Female infidelity may be inherited," *Science News*, July 16, 2011, p. 10.

2.181 Social Jetlag Social jetlag refers to the difference between circadian and social clocks, and is measured as the difference in sleep and wake times between work days and free days. For example, if you sleep between 11 pm and 7 am on weekdays but from 2 am to 10 am on weekends, then your social jetlag is three hours, or equivalent to flying from the West Coast of the US to the East every Friday and back every Sunday. Numerous studies have shown that social jetlag is detrimental to health. One recent study[64] measured the self-reported social jetlag of 145 healthy participants, and found that increased social jetlag was associated with a higher BMI (body mass index), higher cortisol (stress hormone) levels, higher scores on a depression scale, fewer hours of sleep during the week, less physical activity, and a higher resting heart rate.

(a) Indicate whether social jetlag has a positive or negative correlation with each variable listed: BMI, cortisol level, depression score, weekday hours of sleep, physical activity, heart rate.

(b) Can we conclude that social jetlag causes the adverse effects described in the study?

2.182 Football, Brain Size, and Cognitive Scores Exercise 2.143 on page 102 introduces a study that examines the association between playing football, brain size as measured by left hippocampal volume (in μL), and percentile on a cognitive reaction test. Figure 2.56 gives two scatterplots. Both have number of years playing football as the explanatory variable while Graph (a) has cognitive percentile as the response variable and Graph (b) has hippocampal volume as the response variable.

(a) The two corresponding correlations are -0.465 and -0.366. Which correlation goes with which scatterplot?

[64]Rutters, F., et al., "Is social jetlag associated with an adverse endocrine, behavioral, and cardiovascular risk profile?" *J Biol Rhythms*, October 2014; 29(5):377–383.

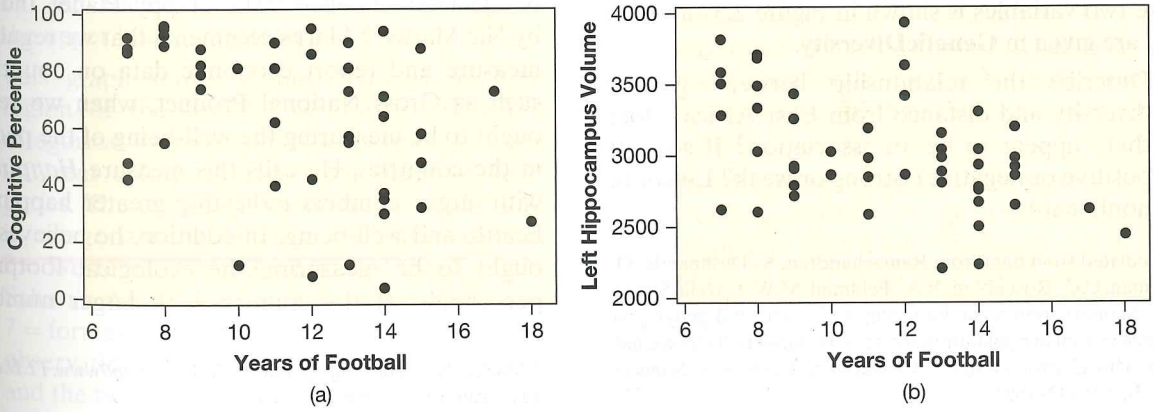


Figure 2.56 *Football, cognitive percentile, and brain size*

(b) Both correlations are negative. Interpret what this means in terms of football, brain size, and cognitive percentile.

2.183 NFL Pre-Season Does pre-season success indicate regular season success in the US National Football League? We looked at the number of pre-season wins and regular season wins for all 32 NFL teams over a 10-year span.

(a) What would a positive association imply about the relationship between pre-season and regular season success in the NFL? What would a negative association imply?

(b) The correlation between these two variables is $r = 0.067$. What does this correlation tell you about the strength of a linear relationship between these two variables?

2.184 What's Wrong with the Statement? A researcher claims to have evidence of a strong positive correlation ($r = 0.88$) between a person's blood alcohol content (BAC) and the type of alcoholic drink consumed (beer, wine, or hard liquor). Explain, statistically, why this claim makes no sense.

2.185 Help for Insomniacs In Exercise 1.23, we learned of a study in which participants were randomly assigned to receive or not receive cognitive behavioral therapy (CBT), and then reported whether or not they experienced any sleep improvement. One news magazine reporting this study said "Sleep improvements were strongly correlated with CBT." Why is this an incorrect use of the statistics word *correlation*?

2.186 Genetic Diversity and Distance from Africa It is hypothesized that humans originated in East Africa, and migrated from there. We compute a measure of genetic diversity for different populations,[65] and the geographic distance of each population from East Africa (Addis Ababa, Ethiopia), as one would travel over the surface of the earth by land (migration long ago is thought to have happened by land). The relationship between these two variables is shown in Figure 2.57 and the data are given in **GeneticDiversity**.

(a) Describe the relationship between genetic diversity and distance from East Africa. Does there appear to be an association? If so, it is positive or negative? Strong or weak? Linear or nonlinear?

(b) Which of the following values gives the correlation between these two variables: $r = -1.22$, $r = -0.83$, $r = -0.14$, $r = 0.14$, $r = 0.83$, or $r = 1.22$?

(c) On which continent is the population with the lowest genetic diversity? On which continent is the population that is farthest from East Africa (by land)?

(d) Populations with greater genetic diversity are thought to be better able to adapt to changing environments, because more genetic diversity provides more opportunities for natural selection. Based only on this information and Figure 2.57, do populations closer or farther from East Africa appear to be better suited to adapt to change?

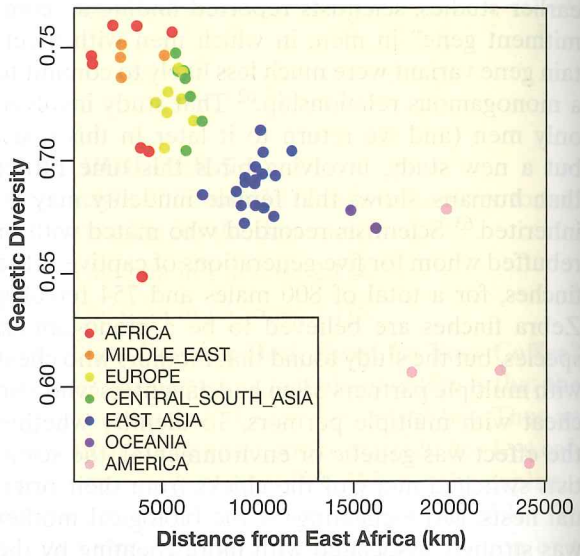


Figure 2.57 ***Genetic diversity of populations by distance from East Africa***

2.187 The Happy Planet Index The website *TED.com* offers free short presentations, called TED Talks, on a variety of interesting subjects. One of the talks is called "The Happy Planet Index," by Nic Marks.[66] Marks comments that we regularly measure and report economic data on countries, such as Gross National Product, when we really ought to be measuring the well-being of the people in the countries. He calls this measure *Happiness*, with larger numbers indicating greater happiness, health, and well-being. In addition, he believes we ought to be measuring the ecological footprint, per capita, of the country, with larger numbers

[65] Calculated from data from Ramachandran, S., Deshpande, O., Roseman, C.C., Rosenberg, N.A., Feldman, M.W., Cavalli-Sforza, L.L. "Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa," *Proceedings of the National Academy of Sciences*, 2005, 102: 15942–15947.

[66] Marks, N., "The Happy Planet Index," *http//www.TED.com/talks*, August 29, 2010.

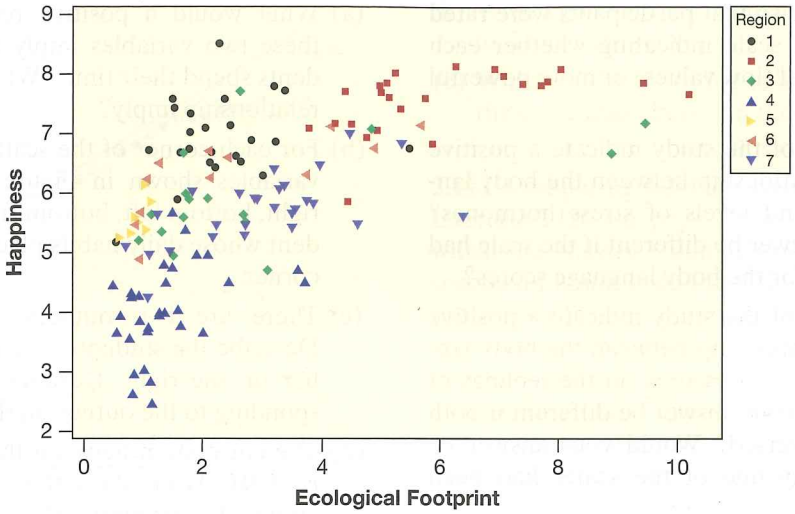


Figure 2.58 *Happiness and ecological footprint*

indicating greater use of resources (such as gas and electricity) and more damage to the planet. Figure 2.58 shows a scatterplot of these two quantitative variables. The data are given in **HappyPlanetIndex**.

(a) Does there appear to be a mostly positive or mostly negative association between these two variables? What does that mean for these two variables?

(b) Describe the happiness and ecological footprint of a country in the bottom left of the graph.

(c) Costa Rica has the highest *Happiness* index. Find it on the graph and estimate its ecological footprint score.

(d) For ecological footprints between 0 and 6, does a larger ecological footprint tend to be associated with more happiness? What about for ecological footprints between 6 and 10? Discuss this result in context.

(e) Marks believes we should be working to move all countries to the top left of the graph, closer to Costa Rica. What attributes does a country in the upper left of the graph possess?

(f) This graph shows a third variable as well: region of the world. One way to depict a categorical variable on a scatterplot is using different colors or shapes for different categories. The code is given in the top right, and is categorized as follows: 1 = Latin America, 2 = Western nations, 3 = Middle East, 4 = Sub-Saharan Africa, 5 = South Asia, 6 = East Asia, 7 = former Communist countries. Discuss one observation of an association between region and the two quantitative variables.

(g) If the goal is to move all countries to the top left, how should efforts be directed for those in the bottom left? How should efforts be directed for those in the top right?

2.188 Ages of Husbands and Wives Suppose we record the husband's age and the wife's age for many randomly selected couples.

(a) What would it mean about ages of couples if these two variables had a negative relationship?

(b) What would it mean about ages of couples if these two variables had a positive relationship?

(c) Which do you think is more likely, a negative or a positive relationship?

(d) Do you expect a strong or a weak relationship in the data? Why?

(e) Would a strong correlation imply there is an association between husband age and wife age?

2.189 Is Your Body Language Closed or Open? A closed body posture includes sitting hunched over or standing with arms crossed rather than sitting or standing up straight and having the arms more open. According to a recent study, people who were rated as having a more closed body posture "had higher levels of stress hormones and said they felt less powerful than those who had a more open pose."[67]

(a) What are the variables in this study? Is each variable categorical or quantitative? Assume participants had body language rated on a numerical scale from low values representing more closed to larger values representing more

[67] "Don't Slouch!" *Consumer Reports OnHealth*, February 2011; 23(2):3.

open. Assume also that participants were rated on a numerical scale indicating whether each felt less powerful (low values) or more powerful (higher values).

(b) Do the results of the study indicate a positive or negative relationship between the body language scores and levels of stress hormones? Would your answer be different if the scale had been reversed for the body language scores?

(c) Do the results of the study indicate a positive or negative relationship between the body language scores and the scores on the feelings of power? Would your answer be different if both scales were reversed? Would your answer be different if only one of the scales had been reversed?

2.190 SAT Scores: Math vs Verbal The **StudentSurvey** dataset includes scores on the Math and Verbal portions of the SAT exam.

(a) What would a positive relationship between these two variables imply about SAT scores? What would a negative relationship imply?

(b) Figure 2.59 shows a scatterplot of these two variables. For each corner of the scatterplot (top left, top right, bottom left, bottom right), describe a student whose SAT scores place him or her in that corner.

(c) Does there appear to be a strong linear relationship between these two variables? What does that tell you about SAT scores?

(d) Which of the following is most likely to be the correlation between these two variables?

$-0.941, \quad -0.605, \quad -0.235, \quad 0.445, \quad 0.751, \quad 0.955$

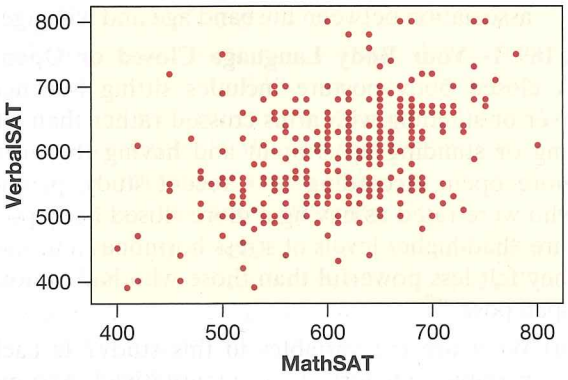


Figure 2.59 *MathSAT score and VerbalSAT score*

2.191 Exercising or Watching TV? The **StudentSurvey** dataset includes information on the number of hours a week students say they exercise and the number of hours a week students say they watch television.

(a) What would a positive relationship between these two variables imply about the way students spend their time? What would a negative relationship imply?

(b) For each corner of the scatterplot of these two variables shown in Figure 2.60 (top left, top right, bottom left, bottom right), describe a student whose daily habits place him or her in that corner.

(c) There are two outliers in this scatterplot. Describe the student corresponding to the outlier on the right. Describe the student corresponding to the outlier on the top.

(d) The correlation between these two variables is $r = 0.01$. What does this correlation tell you about the strength of a linear relationship between these two variables?

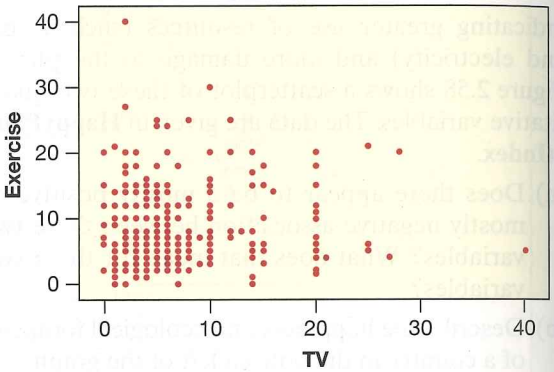


Figure 2.60 *Number of hours a week of exercise and of television watching*

2.192 Females Rating Males on OKCupid The OKCupid dating site provides lots of very interesting data.[68] Figure 2.61 shows a scatterplot of the age of males that females find most attractive, based on the age of the females doing the rating. The X-variable is the age of heterosexual females using the OKCupid site. For each age, the Y-variable gives the age of males that are rated most attractive by women at that age. So, for example, the dot that is farthest left shows that 20-year-old women find men who are age 23 the most attractive. The $Y = X$ line is also shown on the graph, for reference. (The comparable graph for males is given in Exercise 2.193.)

(a) At what age(s) do women find men the same age as themselves the most attractive?

(b) What age men do women in their early 20s find the most attractive: younger men, older men, or men the same age as themselves?

[68]Matlin, C., "Matchmaker, Matchmaker, Make Me a Spreadsheet," *http://fivethirtyeight.com*, September 9, 2014. Based on data from *http://blog.okcupid.com*.

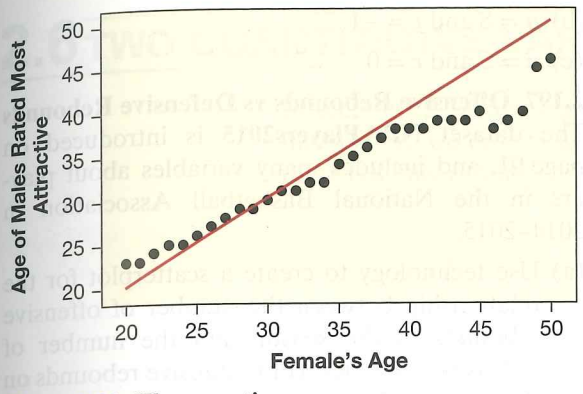


Figure 2.61 *Women rating men*

(c) What age men do women in their 40s find the most attractive: younger men, older men, or men the same age as themselves?

(d) Which of the following is likely to be closest to the correlation between these two variables?

0.9, 0, −0.9

2.193 Males Rating Females on OKCupid Exercise 2.192 introduced data showing the age of males that females find most attractive, based on the age of the females doing the rating. Here we examine the ratings males give for females. Figure 2.62 shows a scatterplot of the age of females that males find most attractive, based on the age of the males doing the rating. The X-variable is the age of heterosexual males using the OKCupid site. For each age, the Y-variable gives the age of females that are rated most attractive by men at that age. So, for example, the dot that is farthest left shows that 20-year-old men find women who are also age 20 the most attractive. The $Y = X$ line is shown on the graph, for reference.

(a) At what age(s) do men find women the same age as themselves the most attractive?

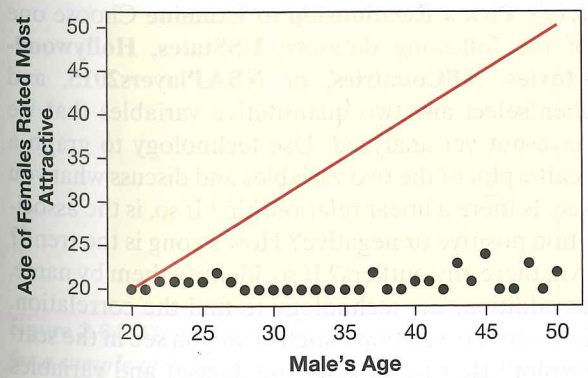


Figure 2.62 *Men rating women*

(b) What age range for women do *all* ages of men find most attractive?

(c) Which of the following is likely to be closest to the correlation between these two variables?

0.9, 0, −0.9

2.194 Comparing Global Internet Connections In Exercise 2.120 on page 92, we discuss a study in which the Nielsen Company measured connection speeds on home computers in nine different countries in order to determine whether connection speed affects the amount of time consumers spend online.[69] Table 2.29 shows the percent of Internet users with a "fast" connection (defined as 2Mb or faster) and the average amount of time spent online, defined as total hours connected to the Web from a home computer during the month of February 2011. The data are also available in the dataset **GlobalInternet**.

(a) What would a positive association mean between these two variables? Explain why a positive relationship might make sense in this context.

(b) What would a negative association mean between these two variables? Explain why a negative relationship might make sense in this context.

Table 2.29 *Internet connection speed and hours online*

Country	Percent Fast Connection	Hours Online
Switzerland	88	20.18
United States	70	26.26
Germany	72	28.04
Australia	64	23.02
United Kingdom	75	28.48
France	70	27.49
Spain	69	26.97
Italy	64	23.59
Brazil	21	31.58

(c) Make a scatterplot of the data, using connection speed as the explanatory variable and time online as the response variable. Is there a positive or negative relationship? Are there any outliers? If so, indicate the country associated with each outlier and describe the characteristics that make it an outlier for the scatterplot.

[69] "Swiss Lead in Speed: Comparing Global Internet Connections," *http://www.nielsen.com/us/en/insights/news/2011/swiss-lead-in-speed-comparing-global-internet-connections.html*, April 1, 2011.

(d) If we eliminate any outliers from the scatterplot, does it appear that the remaining countries have a positive or negative relationship between these two variables?

(e) Use technology to compute the correlation. Is the correlation affected by the outliers?

(f) Can we conclude that a faster connection speed causes people to spend more time online?

2.195 Iris Petals Allometry is the area of biology that studies how different parts of a body grow in relation to other parts. Figure 2.63 shows a scatterplot[70] comparing the length and width of petals of irises.

(a) Does there appear to be a positive or negative association between petal width and petal length? Explain what this tells us about petals.

(b) Discuss the strength of a linear relationship between these two variables.

(c) Estimate the correlation.

(d) Are there any clear outliers in the data?

(e) Estimate the width of the petal that has a length of 30 mm.

(f) There are at least two different types of irises included in the study. Explain how the scatterplot helps illustrate this, and name one difference between the types that the scatterplot makes obvious.

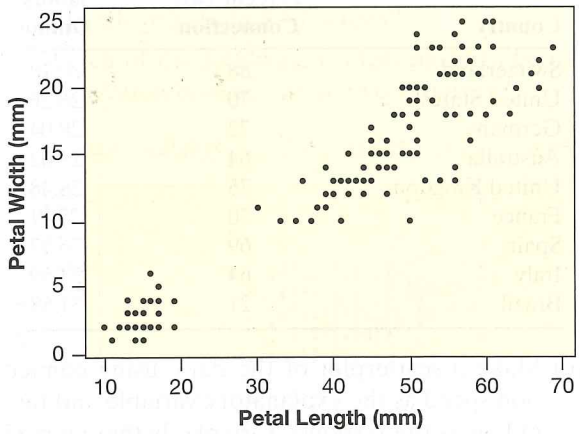


Figure 2.63 *Iris petals*

2.196 Create a Scatterplot Draw any scatterplot satisfying the following conditions:

(a) $n = 10$ and $r = 1$

(b) $n = 8$ and $r = -1$

(c) $n = 5$ and $r = 0$

2.197 Offensive Rebounds vs Defensive Rebounds The dataset **NBAPlayers2015** is introduced on page 91, and includes many variables about players in the National Basketball Association in 2014–2015.

(a) Use technology to create a scatterplot for the relationship between the number of offensive rebounds in the season and the number of defensive rebounds. (Put offensive rebounds on the horizontal axis.)

(b) Does the relationship appear to be positive or negative? What does that mean for these two variables? How strong is the relationship?

(c) There appear to be two outliers in the top right. Who are they?

(d) Use technology to find the correlation between these two variables.

2.198 Do Movies with Larger Budgets Get Higher Audience Ratings? The dataset **HollywoodMovies** is introduced on page 95, and includes many variables for movies, including *Budget* and *AudienceScore*.

(a) Use technology to create a scatterplot to show the relationship between the budget of a movie, in millions of dollars, and the audience score. We want to see if the budget has an effect on the audience score.

(b) There is an outlier with a very large budget. What is the audience rating for this movie and what movie is it? For the data value with the highest audience rating, what is the budget and what movie is it?

(c) Use technology to find the correlation between these two variables.

2.199 Pick a Relationship to Examine Choose one of the following datasets: **USStates**, **HollywoodMovies**, **AllCountries**, or **NBAPlayers2015**, and then select any two quantitative variables that we have not yet analyzed. Use technology to graph a scatterplot of the two variables and discuss what you see. Is there a linear relationship? If so, is the association positive or negative? How strong is the trend? Are there any outliers? If so, identify them by name. In addition, use technology to find the correlation. Does the correlation match what you see in the scatterplot? Be sure to state the dataset and variables you use.

[70] R.A. Fishers's iris data downloaded from *http://lib.stat.cmu.edu/DASL/Datafiles/Fisher'sIrises.html*.

2.6 TWO QUANTITATIVE VARIABLES: LINEAR REGRESSION

In Section 2.5, we investigate the relationship between two quantitative variables. In this section, we discuss how to use one of the variables to predict the other when there is a linear trend.

Image Source/Getty Images, Inc.

Can we predict the size of a tip?

DATA 2.12

Restaurant Tips

The owner[71] of a bistro called *First Crush* in Potsdam, New York, is interested in studying the tipping patterns of its patrons. He collected restaurant bills over a two-week period that he believes provide a good sample of his customers. The data from 157 bills are stored in **RestaurantTips** and include the amount of the bill, size of the tip, percentage tip, number of customers in the group, whether or not a credit card was used, day of the week, and a coded identity of the server. ■

For the restaurant tips data, we want to use the bill amount to predict the tip amount, so the explanatory variable is the amount of the bill and the response variable is the amount of the tip. A scatterplot of this relationship is shown in Figure 2.64.

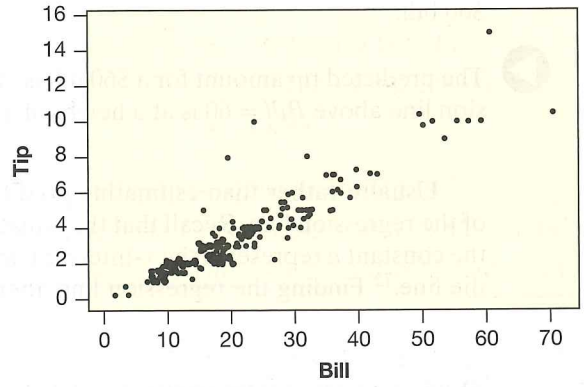


Figure 2.64 ***Tip vs Bill for a sample of First Crush customers***

[71] Thanks to Tom DeRosa for providing the tipping data.

Example 2.39

(a) Use Figure 2.64 to describe the relationship between the bill amount and the tip amount at this restaurant.

(b) Use technology to find the correlation between these two variables.

(c) Draw a line on the scatterplot that seems to fit the data well.

Solution

(a) Figure 2.64 shows a strong positive linear relationship in the data, with a few outliers (big tippers!) above the main pattern.

(b) Using technology, we see that the correlation is $r = 0.915$, reinforcing the fact that the data have a strong positive linear relationship.

(c) There are many lines we could draw that fit the data reasonably well. Try drawing some! Which of the lines you drew do you think fits the data the best? One line that fits the data particularly well is shown in Figure 2.65.

The Regression Line

The process of fitting a line to a set of data is called *linear regression* and the line of best fit is called the *regression line.* The regression line for the restaurant tips data is shown in Figure 2.65 and we see that it seems to fit the data very well. The regression line provides a model of a linear association between two variables, and we can use the regression line on a scatterplot to give a predicted value of the response variable, based on a given value of the explanatory variable.

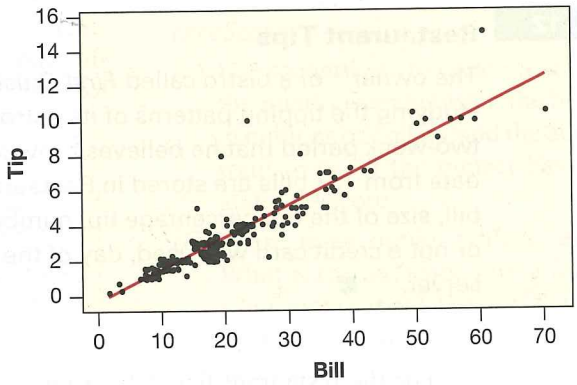


Figure 2.65 ***How well does this line fit the data?***

Example 2.40

Use the regression line in Figure 2.65 to estimate the predicted tip amount on a $60 bill.

Solution

The predicted tip amount for a $60 bill is about $10, because the point on the regression line above *Bill* = 60 is at a height of about $10 on the vertical *Tip* scale.

Usually, rather than estimating predictions using a graph, we use the equation of the regression line. Recall that the equation for a line is given by $y = a + bx$ where the constant a represents the y-intercept and the coefficient b represents the slope of the line.[72] Finding the regression line, then, means finding values for the slope and

[72] You may have learned the equation for a line as $y = mx + b$. Statisticians prefer to use $y = a + bx$. In either case, the coefficient of x is the slope and the constant term is the vertical intercept.

intercept of the line that best describes the linear trend of the data. This can be done on many calculators and computer programs.

To help distinguish between the predicted and observed values of the response variable, we often add a "hat" to the response variable name to denote the predicted value. Thus if our data pairs are (x, y) with x as the explanatory variable and y as the response variable, the regression line is given by

$$\hat{y} = a + bx$$

Explanatory and Response Variables

The regression line to predict y from x is NOT the same as the regression line to predict x from y. Be sure to always pay attention to which is the explanatory variable and which is the response variable! A regression line is always in the form

$$\widehat{Response} = a + b \cdot Explanatory$$

For the restaurant tips data, the equation of the regression line shown in Figure 2.65 is

$$\widehat{Tip} = -0.292 + 0.182 \cdot Bill$$

The y-intercept of this line is -0.292 and the slope is 0.182.

Using the Equation of the Regression Line to Make Predictions

The equation of the regression line is often also called a *prediction equation* because we can use it to make predictions. We substitute the value of the explanatory variable into the prediction equation to calculate the predicted response.

Example 2.41

Three different bill amounts from the **RestaurantTips** dataset are given. In each case, use the regression line $\widehat{Tip} = -0.292 + 0.182 \cdot Bill$ to predict the tip.

(a) A bill of \$59.33
(b) A bill of \$9.52
(c) A bill of \$23.70

Solution

(a) If the bill is \$59.33, we have

$$\begin{aligned} \widehat{Tip} &= -0.292 + 0.182 \cdot Bill \\ &= -0.292 + 0.182(59.33) \\ &= 10.506 \end{aligned}$$

The predicted size of the tip is 10.506 or about \$10.51.

(b) For a bill of \$9.52, we have $\widehat{Tip} = -0.292 + 0.182(9.52) = 1.441 \approx \$1.44$.

(c) For a bill of \$23.70, we have $\widehat{Tip} = -0.292 + 0.182(23.70) = 4.021 \approx \$4.02$.

The predicted value is an estimate of the average response value for that particular value of the explanatory variable. We expect actual values to be above or below this amount.

Residuals

In Example 2.41, we found the predicted tip for three of the bills in the restaurant tips dataset. We can look in the dataset to see how close these predictions are to the actual tip amount for those bills. The *residual* is the difference between the observed value and the predicted value. On a scatterplot, the predicted value is the height of the regression line for a given *Bill* amount and the observed value is the height of the particular data point with that *Bill* amount, so the residual is the vertical distance from the point to the line. The residual for one data value is shown in Figure 2.66.

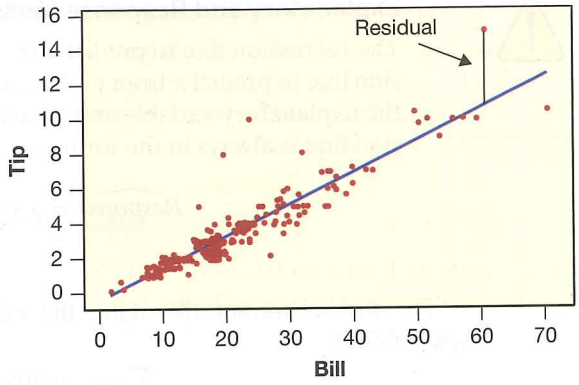


Figure 2.66 ***A residual is the vertical deviation from a point to the line***

Residuals

The **residual** at a data value is the difference between the observed and predicted values of the response variable:

$$Residual = Observed - Predicted = y - \hat{y}$$

On a scatterplot, the residual represents the vertical deviation from the line to a data point. Points above the line will have positive residuals and points below the line will have negative residuals. If the predicted values closely match the observed data values, the residuals will be small.

Example 2.42

In Example 2.41, we find the predicted tip amount for three different bills in the **RestaurantTips** dataset. The actual tips left by each of these customers are shown below. Use this information to calculate the residuals for each of these sample points.

(a) The tip left on a bill of \$59.33 was \$10.00.

(b) The tip left on a bill of \$9.52 was \$1.00.

(c) The tip left on a bill of \$23.70 was \$10.00.

Solution

(a) The observed tip left on the bill of \$59.33 is \$10.00 and we see in Example 2.41(a) that the predicted tip is \$10.51. The observed tip is a bit less than the predicted tip. We have

$$Residual = Observed - Predicted = 10.00 - 10.51 = -0.51$$

(b) The observed tip left on the bill of \$9.52 is just \$1.00, and we see in Example 2.41(b) that the predicted tip for a bill this size is \$1.44, so

$$Residual = Observed - Predicted = 1.00 - 1.44 = -0.44$$

(c) The observed tip left on a bill of \$23.70 (the first case in the dataset) is \$10.00 and we see in Example 2.41(c) that the predicted tip is only \$4.02. The observed tip is quite a bit larger than the predicted tip and we have

$$Residual = Observed - Predicted = 10.00 - 4.02 = 5.98$$

This is one of the largest residuals. The server would be quite happy to receive this extra large tip!

Example 2.43

Data 2.9 on page 106 introduced data that show the approval rating of a president running for re-election and the resulting margin of victory or defeat for the president in the election. The data are in **ElectionMargin**.

(a) The regression line for these 12 data points is

$$\widehat{Margin} = -36.8 + 0.839(Approval)$$

Calculate the predicted values and the residuals for all the data points.

(b) Show the residuals as distances on a scatterplot with the regression line.

(c) Which residual is the largest? For this largest residual, is the observed margin higher or lower than the margin predicted by the regression line? To which president and year does this residual correspond?

Solution

(a) We use the regression line to find the predicted value for each data point, and then subtract to find the residuals. The results are given in Table 2.30. Some of the residuals are positive and some are negative, reflecting the fact that some of the data points lie above the regression line and some lie below.

(b) See Figure 2.67. At a given approval rating, such as 62, the observed margin (10) corresponds to the height of the data point, while the predicted value (15.23) corresponds to the height of the line at an approval rating of 62. Notice that in this case the line lies above the data point, and the difference between the observed value and the predicted value is the length of the vertical line joining the point to the line.

(c) The largest residual is 12.16. The observed margin of victory is 23.2, high above the predicted value of 11.04. We see in Figure 2.67 that this is the point with the greatest vertical deviation from the line. Looking back at Table 2.26 on page 107, we see that this residual corresponds to President Nixon in 1972.

Table 2.30 ***Predicted margin and residuals for presidential incumbents***

Approval	Actual Margin	Predicted Margin	Residual
62	10.0	15.23	−5.23
50	4.5	5.17	−0.67
70	15.4	21.94	−6.54
67	22.6	19.43	3.17
57	23.2	11.04	12.16
48	−2.1	3.49	−5.59
31	−9.7	−10.76	1.06
57	18.2	11.04	7.16
39	−5.5	−4.05	−1.45
55	8.5	9.36	−0.86
49	2.4	4.33	−1.93
50	3.9	5.17	−1.27

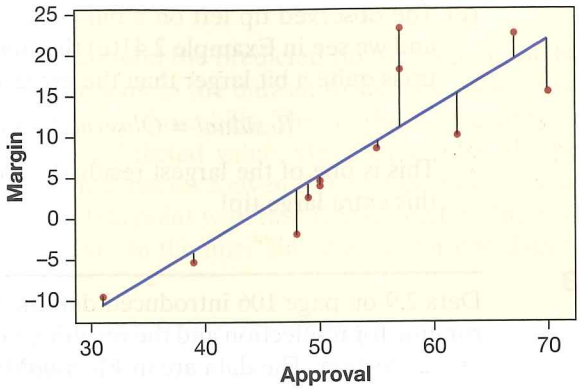


Figure 2.67 *Residuals measure vertical deviations from the line to the points*

What Does "Line of Best Fit" Mean?

How can we determine which line is the best fit for a set of data? And what do we even mean by "best fit"? Our goal is to find the line that provides the best predictions for the observed values of the response variable. The line that fits the data best should then be one where the residuals are close to zero. In particular, we usually try to make the squares of the residuals, $(y - \hat{y})^2$, small. The *least squares line* is the line with the slope and intercept that makes the sum of the squared residuals as small as it can possibly be.

Least Squares Line

The **least squares line**, also called the **line of best fit**, is the line that minimizes the sum of the squared residuals.

Throughout this text, we use the terms *regression line* and *least squares line* interchangeably.

We should expect observed values to fall both above and below the line of best fit, so residuals are both positive and negative. This is one reason why we square them. In fact, if we add up all of the residuals from the regression line, the sum will always be zero.

Interpreting the Slope and Intercept of the Regression Line

Recall that the regression line for the **RestaurantTips** data is

$$\widehat{Tip} = -0.292 + 0.182 \cdot Bill$$

How can we interpret the slope 0.182 and intercept −0.292?

Recall that for a general line $y = a + bx$, the slope represents the change in y over the change in x. If the change in x is 1, then the slope represents the change in y. The intercept represents the value of y when x is zero.

Interpreting the Slope and Intercept of the Regression Line

For the regression line $\hat{y} = a + bx$,

- The slope b represents the predicted change in the response variable y given a one unit increase in the explanatory variable x.
- The intercept a represents the predicted value of the response variable y if the explanatory variable x is zero. The interpretation may be nonsensical since it is often not reasonable for the explanatory variable to be zero.

Example 2.44

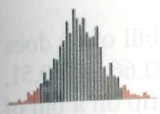
For the **RestaurantTips** regression line $\widehat{Tip} = -0.292 + 0.182 \cdot Bill$, interpret the slope and the intercept in context.

Solution

The slope 0.182 indicates that the tip is predicted to go up by about \$0.182 for a one dollar increase in the bill. A rough interpretation is that people in this sample tend to tip about 18.2%.

The intercept −0.292 indicates that the tip will be −\$0.292 if the bill is \$0. Since a bill is rarely zero dollars and a tip cannot be negative, this makes little sense.

Example 2.45



In Example 2.34 on page 108, we consider some scatterplots from the dataset **FloridaLakes** showing relationships between acidity, alkalinity, and fish mercury levels in $n = 53$ Florida lakes. We wish to predict a quantity that is difficult to measure (mercury level of fish) using a value that is more easily obtained from a water sample (acidity). We see in Example 2.34 that there appears to be a negative linear association between these two variables, so a regression line is appropriate.

(a) Use technology to find the regression line to predict *Mercury* from pH, and plot it on a scatterplot of the data.

(b) Interpret the slope of the regression line in the context of Florida lakes.

(c) Put an arrow on the scatterplot pointing to the data for Puzzle Lake, which has an acidity of 7.5 and an average mercury level of 1.10 ppm. Calculate the predicted mercury level for Puzzle Lake and compare it to the observed mercury level. Calculate the residual.

Solution

(a) We use technology to find the regression line:

$$\widehat{Mercury} = 1.53 - 0.1523 \cdot pH$$

For the scatterplot, since we are predicting mercury level from pH, the pH variable goes on the horizontal axis and the mercury variable goes on the vertical axis. The line is plotted with the data in Figure 2.68.

(b) The slope in the prediction equation represents the expected change in the response variable for a one unit increase in the explanatory variable. Since the slope in this case is −0.1523, we expect the average mercury level in fish to decrease by about 0.1523 ppm for each increase of 1 in the pH of the lake water.

(c) See the arrow in Figure 2.68. The predicted value for Puzzle Lake is $\widehat{Mercury} = 1.53 - 0.1523 \cdot (7.5) = 0.388$ ppm. The observed value of 1.10 is quite a bit higher than the predicted value for this lake. The residual is $1.10 - 0.388 = 0.712$, the largest residual of all 53 lakes.

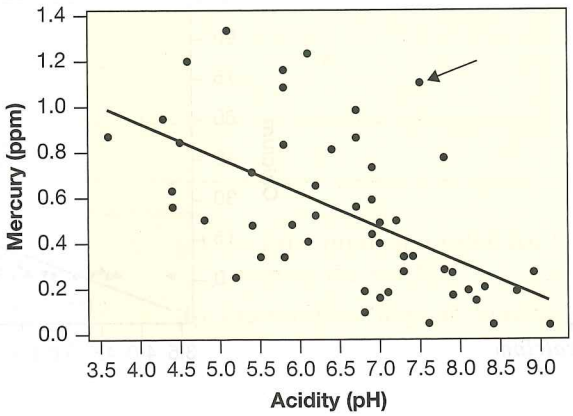


Figure 2.68 *Using acidity to predict average mercury level in fish*

Notation for the Slope

We have seen that we use the notation b for the slope of a regression line that comes from a sample. What about the regression line for a population? The dataset on presidential elections used to create the regression line $\widehat{Margin} = -36.8 + 0.839 \cdot Approval$ in Example 2.43 represents the population of all relevant US presidential elections since 1940. As we have seen with other quantities, the notation we use for the slope of the regression line of a population is different than the notation we use for the slope of the regression line of a sample. For the slope of a regression line for a population, we use the Greek letter β (beta).

Regression Cautions

In the solution to Example 2.44, we see that predicting the tip for a bill of \$0 does not make any sense. Since the bill amounts in that dataset range from \$1.66 to \$70.51, it also would not make sense to use the regression line to predict the tip on a bill of \$1000. In general, it is not appropriate to use regression lines to make predictions using values of the explanatory variable that are far from the range of values used to create the line. This is called *extrapolating* too far from the original data.

Regression Caution #1

Avoid trying to apply a regression line to predict values far from those that were used to create it.

Example 2.46

In Example 2.45 on page 129, we used the acidity (pH) of Florida lakes to predict mercury levels in fish. Suppose that, instead of mercury, we use acidity to predict the calcium concentration (mg/l) in Florida lakes. Figure 2.69 shows a scatterplot of these data with the regression line $\widehat{Calcium} = -51.4 + 11.17 \cdot pH$ for the 53 lakes in our sample. Give an interpretation for the slope in this situation. Does the intercept make sense? Comment on how well the linear prediction equation describes the relationship between these two variables.

Solution

The slope of 11.17 in the prediction equation indicates that the calcium concentration in lake water increases by about 11.17 mg/l when the pH goes up by one. The intercept does not have a physical interpretation since there are no lakes with a pH of zero and a negative calcium concentration makes no sense. Although there is clearly a positive association between acidity and calcium concentration, the relationship is not a linear one. The pattern in the scatterplot indicates a curved pattern that

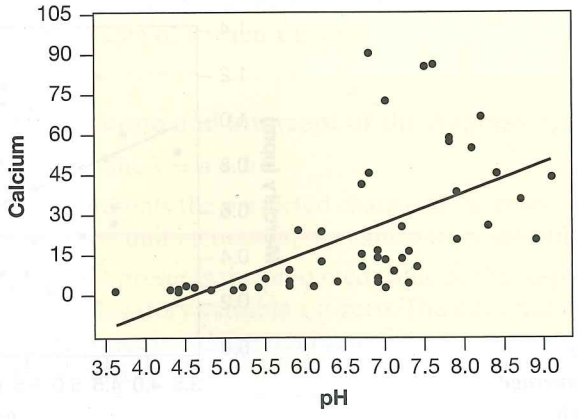


Figure 2.69 *Using acidity to predict calcium concentration*

increases more steeply as pH increases. The least squares line predicts negative calcium concentrations (which are impossible) for pH levels as large as 4.5, which are within the domain of lakes in this sample.

The correlation between acidity and average mercury levels in Figure 2.68 is −0.575 while acidity and calcium concentration in Figure 2.69 have a correlation of 0.577. Although these correlations are close in magnitude, linear regression is a more appropriate model for the first situation than it is for the second. It is always important to plot the data and look for patterns that may or may not follow a linear trend.

Regression Caution #2

Plot the data! Although the regression line can be calculated for *any* set of paired quantitative variables, it is only appropriate to use a regression line when there is a linear trend in the data.

Finally, when we plot the data, we also look for outliers that may exert a strong influence on the regression line, similar to what we see for correlation in Figure 2.53 on page 114.

Regression Caution #3

Outliers can have a strong influence on the regression line, just as we saw for correlation. In particular, data points for which the explanatory value is an outlier are often called *influential points* because they exert an overly strong effect on the regression line.

SECTION LEARNING GOALS

You should now have the understanding and skills to:

- Use technology to find the regression line for a dataset with two quantitative variables
- Calculate predicted values from a regression line
- Interpret the slope (and intercept, when appropriate) of a regression line in context
- Calculate residuals and visualize residuals on a scatterplot
- Beware of extrapolating too far out when making predictions
- Recognize the importance of plotting your data

Exercises for Section 2.6

SKILL BUILDER 1

In Exercises 2.200 to 2.203, two variables are defined, a regression equation is given, and one data point is given.

(a) Find the predicted value for the data point and compute the residual.

(b) Interpret the slope in context.

(c) Interpret the intercept in context, and if the intercept makes no sense in this context, explain why.

2.200 *Hgt* = height in inches, *Age* = age in years of a child.
$\widehat{Hgt} = 24.3 + 2.74(Age)$; data point is a child 12 years old who is 60 inches tall.

2.201 *BAC* = blood alcohol content (% of alcohol in the blood), *Drinks* = number of alcoholic drinks.
$\widehat{BAC} = -0.0127 + 0.018(Drinks)$; data point is an individual who consumed 3 drinks and had a *BAC* of 0.08.

2.202 *Weight* = maximum weight capable of bench pressing (pounds), *Training* = number of hours spent lifting weights a week.
$\widehat{Weight} = 95 + 11.7(Training)$; data point is an individual who trains 5 hours a week and can bench 150 pounds.

2.203 *Study* = number of hours spent studying for an exam, *Grade* = grade on the exam.
$\widehat{Grade} = 41.0 + 3.8(Study)$; data point is a student who studied 10 hours and received an 81 on the exam.

SKILL BUILDER 2

Use technology to find the regression line to predict Y from X in Exercises 2.204 to 2.207.

2.204

X	3	5	2	7	6
Y	1	2	1.5	3	2.5

2.205

X	2	4	6	8	10	12
Y	50	58	55	61	69	68

2.206

X	10	20	30	40	50	60
Y	112	85	92	71	64	70

2.207

X	15	20	25	30	35	40	45	50
Y	532	466	478	320	303	349	275	221

2.208 Concentration of CO_2 in the Atmosphere Levels of carbon dioxide (CO_2) in the atmosphere are rising rapidly, far above any levels ever before recorded. Levels were around 278 parts per million in 1800, before the Industrial Age, and had never, in the hundreds of thousands of years before that, gone above 300 ppm. Levels are now over 400 ppm. Table 2.31 shows the rapid rise of CO_2 concentrations over the 50 years from 1960–2010, also available in **CarbonDioxide**.[73] We can use this information to predict CO_2 levels in different years.

(a) What is the explanatory variable? What is the response variable?

(b) Draw a scatterplot of the data. Does there appear to be a linear relationship in the data?

(c) Use technology to find the correlation between year and CO_2 levels. Does the value of the correlation support your answer to part (b)?

(d) Use technology to calculate the regression line to predict CO_2 from year.

(e) Interpret the slope of the regression line, in terms of carbon dioxide concentrations.

(f) What is the intercept of the line? Does it make sense in context? Why or why not?

(g) Use the regression line to predict the CO_2 level in 2003. In 2020.

(h) Find the residual for 2010.

Table 2.31 ***Concentration of carbon dioxide in the atmosphere***

Year	CO_2
1960	316.91
1965	320.04
1970	325.68
1975	331.08
1980	338.68
1985	345.87
1990	354.16
1995	360.62
2000	369.40
2005	379.76
2010	389.78

2.209 The Honeybee Waggle Dance When honeybee scouts find a food source or a nice site for a new home, they communicate the location to the rest of the swarm by doing a "waggle dance."[74] They point in the direction of the site and dance longer for sites farther away. The rest of the bees use the duration of the dance to predict distance to the site.

[73] Dr. Pieter Tans, NOAA/ESRL, *http://www.esrl.noaa.gov/gmd/ccgg/trends/*. Values recorded at the Mauna Loa Observatory in Hawaii.

[74] Check out a honeybee waggle dance on YouTube!

Table 2.32 *Duration of a honeybee waggle dance to indicate distance to the source*

Distance	Duration
200	0.40
250	0.45
500	0.95
950	1.30
1950	2.00
3500	3.10
4300	4.10

Table 2.32 shows the distance, in meters, and the duration of the dance, in seconds, for seven honeybee scouts.[75] This information is also given in **HoneybeeWaggle**.

(a) Which is the explanatory variable? Which is the response variable?

(b) Figure 2.70 shows a scatterplot of the data. Does there appear to be a linear trend in the data? If so, is it positive or negative?

(c) Use technology to find the correlation between the two variables.

(d) Use technology to find the regression line to predict distance from duration.

(e) Interpret the slope of the line in context.

(f) Predict the distance to the site if a honeybee does a waggle dance lasting 1 second. Lasting 3 seconds.

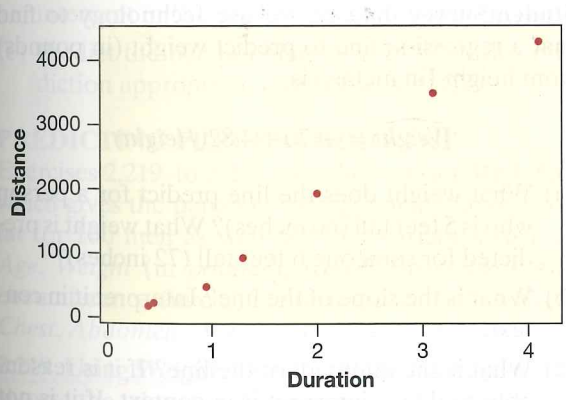


Figure 2.70 *Using dance duration to predict distance to source*

2.210 Is It Getting Harder to Win a Hot Dog Eating Contest? Every Fourth of July, Nathan's Famous in New York City holds a hot dog eating contest, which we discuss in Exercise 2.110. Table 2.21 on page 89 shows the winning number of hot dogs eaten every year from 2002 to 2015, and the data are also available in **HotDogs**. Figure 2.71 shows the scatterplot with the regression line.

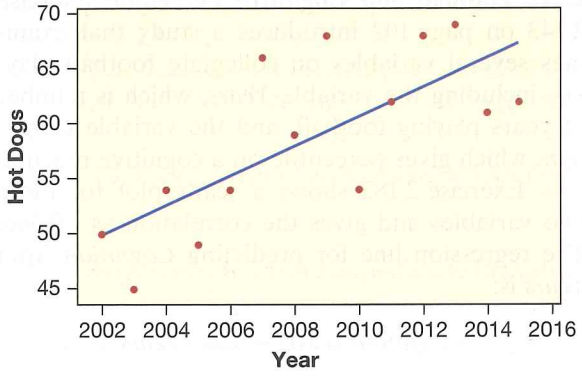


Figure 2.71 *Winning number of hot dogs*

(a) Is the trend in the data mostly positive or negative?

(b) Using Figure 2.71, is the residual larger in 2007 or 2008? Is the residual positive or negative in 2010?

(c) Use technology to find the correlation.

(d) Use technology to find the regression line to predict the winning number of hot dogs from the year.

(e) Interpret the slope of the regression line.

(f) Predict the winning number of hot dogs in 2016. (*Bonus*: Find the actual winning number in 2016 and compute the residual.)

(g) Why would it not be appropriate to use this regression line to predict the winning number of hot dogs in 2025?

2.211 Oxygen and Lung Cancer A recent study[76] has found an association between elevation and lung cancer incidence. Incidence of cancer appears to be lower at higher elevations, possibly because of lower oxygen levels in the air. We are told that "for every one km rise in elevation, lung cancer incidence decreased by 7.23" where cancer incidence is given in cases per 100,000 individuals.

(a) Is this a positive or negative association?

[75] Seeley, T., *Honeybee Democracy*, Princeton University Press, Princeton, NJ, 2010, p. 128.

[76] Simeonov, K.P., and Himmelstein, D.S., "Lung cancer incidence decreases with elevation: Evidence for oxygen as an inhaled carcinogen," *Peer Journal*, 2015; 3:e705.

(b) Which of the following quantities is given in the sentence in quotes: correlation, slope of regression line, intercept of regression line, or none of these?

(c) What is the explanatory variable? What is the response variable?

2.212 Football and Cognitive Percentile Exercise 2.143 on page 102 introduces a study that examines several variables on collegiate football players, including the variable *Years*, which is number of years playing football, and the variable *Cognition*, which gives percentile on a cognitive reaction test. Exercise 2.182 shows a scatterplot for these two variables and gives the correlation as −0.366. The regression line for predicting *Cognition* from *Years* is:

$$\widehat{Cognition} = 102 - 3.34 \cdot Years$$

(a) Predict the cognitive percentile for someone who has played football for 8 years and for someone who has played football for 14 years.

(b) Interpret the slope in terms of football and cognitive percentile.

(c) All the participants had played between 7 and 18 years of football. Is it reasonable to interpret the intercept in context? Why or why not?

2.213 Football and Brain Size Exercise 2.143 on page 102 introduces a study that examines several variables on collegiate football players, including the variable *Years*, which is number of years playing football, and the variable *BrainSize*, which is volume of the left hippocampus in the brain measured in μL. Figure 2.72 shows a scatterplot of these two variables along with the regression line. For each of the following cases, estimate from the graph the number of years of football, the predicted brain size, the actual brain size, and the residual.

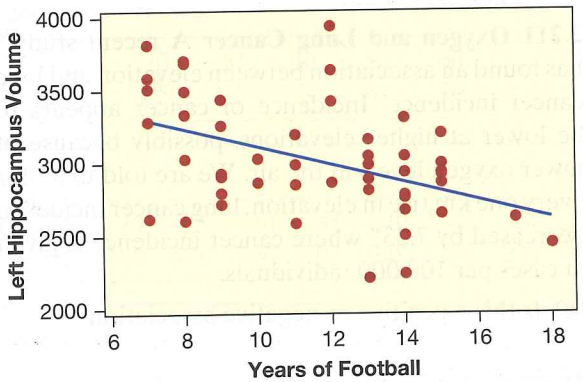


Figure 2.72 *Relationship of football experience and brain hippocampus size*

(a) The case with 18 years of football experience.

(b) The case with the largest positive residual.

(c) The case with the largest negative residual.

2.214 Runs and Wins in Baseball In Exercise 2.150 on page 104, we looked at the relationship between total hits by team in the 2014 season and division (NL or AL) in baseball. Two other variables in the **BaseballHits** dataset are the number of wins and the number of runs scored during the season. The dataset consists of values for each variable from all 30 MLB teams. From these data we calculate the regression line:

$$\widehat{Wins} = 34.85 + 0.070(Runs)$$

(a) Which is the explanatory and which is the response variable in this regression line?

(b) Interpret the intercept and slope in context.

(c) The San Francisco Giants won 88 games while scoring 665 runs in 2014. Predict the number of games won by San Francisco using the regression line. Calculate the residual. Were the Giants efficient at winning games with 665 runs?

2.215 Presidential Elections In Example 2.43 on page 127, we used the approval rating of a president running for re-election to predict the margin of victory or defeat in the election. We saw that the least squares line is $\widehat{Margin} = -36.76 + 0.839\,(Approval)$. Interpret the slope and the intercept of the line in context.

2.216 Height and Weight Using the data in the **StudentSurvey** dataset, we use technology to find that a regression line to predict weight (in pounds) from height (in inches) is

$$\widehat{Weight} = -170 + 4.82(Height)$$

(a) What weight does the line predict for a person who is 5 feet tall (60 inches)? What weight is predicted for someone 6 feet tall (72 inches)?

(b) What is the slope of the line? Interpret it in context.

(c) What is the intercept of the line? If it is reasonable to do so, interpret it in context. If it is not reasonable, explain why not.

(d) What weight does the regression line predict for a baby who is 20 inches long? Why is it not appropriate to use the regression line in this case?

2.217 NFL Pre-Season Using 10 years of National Football League (NFL) data, we calculate the following regression line to predict regular season wins

(*Wins*) by number of wins in the 4 pre-season games (*PreSeason*):

$$\widehat{Wins} = 7.5 + 0.2(PreSeason)$$

(a) Which is the explanatory variable, and which is the response variable in this regression line?

(b) How many wins does the regression line predict for a team that won 2 games in pre-season?

(c) What is the slope of the line? Interpret it in context.

(d) What is the intercept of the line? If it is reasonable to do so, interpret it in context. If it is not reasonable, explain why not.

(e) How many regular season wins does the regression line predict for a team that wins 100 pre-season games? Why is it not appropriate to use the regression line in this case?

2.218 Is the Honeybee Population Shrinking? The **Honeybee** dataset contains data collected from the USDA on the estimated number of honeybee colonies (in thousands) for the years 1995 through 2012.[77] We use technology to find that a regression line to predict number of (thousand) colonies from year (in calendar year) is

$$\widehat{Colonies} = 19,291,511 - 8.358(Year)$$

(a) Interpret the slope of the line in context.

(b) Often researchers will adjust a year explanatory variable such that it represents years since the first year data were colleected. Why might they do this? (*Hint:* Consider interpreting the y-intercept in this regression line.)

(c) Predict the bee population in 2100. Is this prediction appropriate (why or why not)?

PREDICTING PERCENT BODY FAT

Exercises 2.219 to 2.221 use the dataset **BodyFat**, which gives the percent of weight made up of body fat for 100 men as well as other variables such as *Age*, *Weight* (in pounds), *Height* (in inches), and circumference (in cm) measurements for the *Neck*, *Chest*, *Abdomen*, *Ankle*, *Biceps*, and *Wrist*.[78]

2.219 Using Weight to Predict Body Fat Figure 2.73 shows the data and regression line for using weight to predict body fat percentage. For the case with the largest positive residual, estimate the values of both variables. In addition, estimate the predicted body fat percent and the residual for that point.

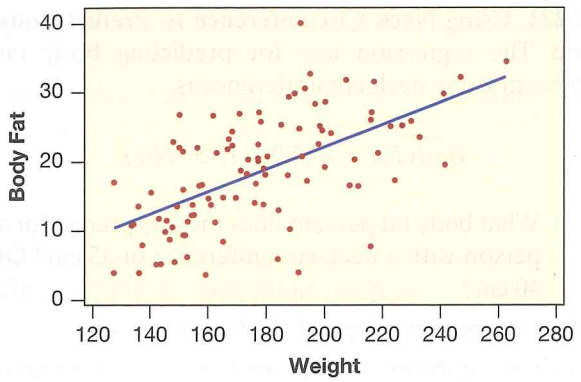


Figure 2.73 *Using weight to predict percent body fat*

2.220 Using Abdomen Circumference to Predict Body Fat Figure 2.74 shows the data and regression line for using abdomen circumference to predict body fat percentage.

(a) Which scatterplot, the one using *Weight* in Figure 2.73 or the one using *Abdomen* in Figure 2.74, appears to contain data with a larger correlation?

(b) In Figure 2.74, one person has a very large abdomen circumference of about 127 cm. Estimate the actual body fat percent for this person as well as the predicted body fat percent.

(c) Use Figure 2.74 to estimate the abdomen circumference for the person with about 40% body fat. In addition, estimate the residual for this person.

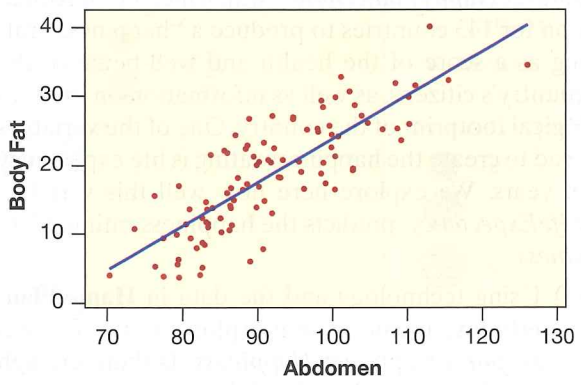


Figure 2.74 *Using abdomen circumference to predict percent body fat*

[77] USDA National Agriculture and Statistical Services, *http://usda.mannlib.cornell.edu/MannUsda/viewDocumentInfo.do?documentID=1191*. Accessed September 2015.

[78] A sample taken from data provided by R. Johnson in "Fitting Percentage of Body Fat to Simple Body Measurements," *Journal of Statistics Education*, 1996, *http://www.amstat.org/publications/jse/v4n1/datasets.johnson.html*.

2.221 Using Neck Circumference to Predict Body Fat The regression line for predicting body fat percent using neck circumference is

$$\widehat{BodyFat} = -47.9 + 1.75 \cdot Neck$$

(a) What body fat percent does the line predict for a person with a neck circumference of 35 cm? Of 40 cm?

(b) Interpret the slope of the line in context.

(c) One of the men in the study had a neck circumference of 38.7 cm and a body fat percent of 11.3. Find the residual for this man.

2.222 Predicting World Gross Revenue for a Movie from Its Opening Weekend Use the data in **HollywoodMovies** to use revenue from a movie's opening weekend (*OpeningWeekend*) to predict total world gross revenues by the end of the year (*WorldGross*). Both variables are in millions of dollars.

(a) Use technology to create a scatterplot for this relationship. Describe the scatterplot: Is there a linear trend? How strong is it? Is it positive or negative? Does it look as if revenue from a movie's opening weekend is a good predictor of its future total earnings?

(b) The scatterplot contains an outlier with an unusually high world gross. Use the dataset to identify this movie.

(c) Use technology to find the correlation between these variables.

(d) Use technology to find the regression line.

(e) Use the regression line to predict world gross revenues for a movie that makes 50 million dollars in its opening weekend.

2.223 Using Life Expectancy to Predict Happiness In Exercise 2.187 on page 118, we introduce the dataset **HappyPlanetIndex**, which includes information for 143 countries to produce a "happiness" rating as a score of the health and well-being of the country's citizens, as well as information on the ecological footprint of the country. One of the variables used to create the happiness rating is life expectancy in years. We explore here how well this variable, *LifeExpectancy*, predicts the happiness rating, *Happiness*.

(a) Using technology and the data in **HappyPlanetIndex**, create a scatterplot to use *LifeExpectancy* to predict *Happiness*. Is there enough of a linear trend so that it is reasonable to construct a regression line?

(b) Find a formula for the regression line and display the line on the scatterplot.

(c) Interpret the slope of the regression line in context.

2.224 Pick a Relationship to Examine Choose one of the following datasets: **USStates**, **StudentSurvey**, **AllCountries**, or **NBAPlayers2015**, and then select any two quantitative variables that we have not yet analyzed. Use technology to create a scatterplot of the two variables with the regression line on it and discuss what you see. If there is a reasonable linear relationship, find a formula for the regression line. If not, find two other quantitative variables that do have a reasonable linear relationship and find the regression line for them. Indicate whether there are any outliers in the dataset that might be influential points or have large residuals. Be sure to state the dataset and variables you use.

2.225 The Impact of Strong Economic Growth In 2011, the Congressional Budget Office predicted that the US economy would grow by 2.8% per year on average over the decade from 2011 to 2021. At this rate, in 2021, the ratio of national debt to GDP (gross domestic product) is predicted to be 76% and the federal deficit is predicted to be $861 billion. Both predictions depend heavily on the growth rate. If the growth rate is 3.3% over the same decade, for example, the predicted 2021 debt-to-GDP ratio is 66% and the predicted 2021 deficit is $521 billion. If the growth rate is even stronger, at 3.9%, the predicted 2021 debt-to-GDP ratio is 55% and the predicted 2021 deficit is $113 billion.[79]

(a) There are only three individual cases given (for three different economic scenarios), and for each we are given values of three variables. What are the variables?

(b) Use technology and the three cases given to find the regression line for predicting 2021 debt-to-GDP ratio from the average growth rate over the decade 2011 to 2021.

(c) Interpret the slope and intercept of the line from part (b) in context.

(d) What 2021 debt-to-GDP ratio does the model in part (b) predict if growth is 2%? If it is 4%?

(e) Studies indicate that a country's economic growth slows if the debt-to-GDP ratio hits 90%. Using the model from part (b), at what growth rate would we expect the ratio in the US to hit 90% in 2021?

[79] Gandel, S., "Higher growth could mean our debt worries are all for nothing," *Time Magazine*, March 7, 2011, p. 20.

(f) Use technology and the three cases given to find the regression line for predicting the deficit (in billions of dollars) in 2021 from the average growth rate over the decade 2011 to 2021.

(g) Interpret the slope and intercept of the line from part (f) in context.

(h) What 2021 deficit does the model in part (f) predict if growth is 2%? If it is 4%?

(i) The deficit in 2011 was $1.4 trillion. What growth rate would leave the deficit at this level in 2021?

2.7 DATA VISUALIZATION AND MULTIPLE VARIABLES

In Sections 2.1 through 2.6 we consider basic graphs that can be used to visualize the distribution of a single variable (such as a histogram or a barchart), or a relationship between two variables (such as a scatterplot or side-by-side boxplot). Often we may wish to extend these basic methods or create an entirely new type of graph to convey more information. For example, we may wish to display more than two variables, to incorporate geographic information, to track data over time, or to connect our graph to the specific applied context of our dataset in other ways. Our only guiding principle is to facilitate quick and accurate interpretation of data, and this allows plenty of room for creativity. Good data visualization ("data viz") can involve elements of statistics, computer graphics, and artistic design.

Here we explore some additional, more advanced, techniques to visualize data, including ways to visually explore relationships between multiple variables.

Augmented Scatterplots for More than Two Variables

The scatterplot, introduced in Section 2.5, displays a relationship between two quantitative variables by letting the horizontal axis correspond to one variable, the vertical axis correspond to the other variable, and plotting a point for each pair of values. For a basic scatterplot the type of "point" (for example, a black circle) does not change. However, we can incorporate other variables by letting the size, shape, or color of the points vary. A scatterplot in which the size of the point depends on another variable is sometimes called a *bubble chart* or *bubble plot*.

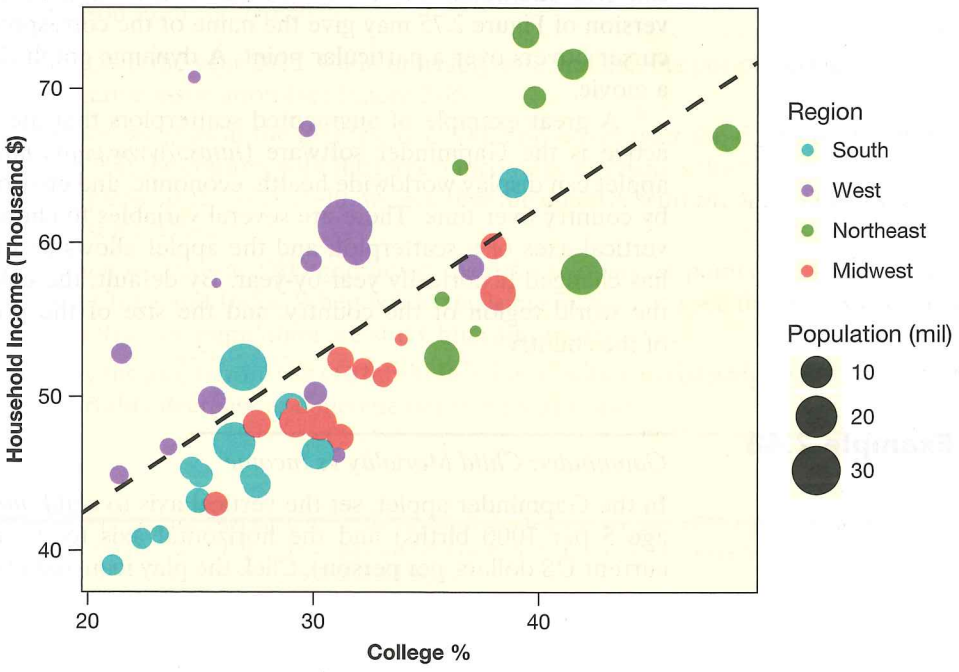


Figure 2.75 *Augmented scatterplot showing various information for all US states*

For example, Figure 2.75 gives a scatterplot of household income vs percent with a college degree for all US States, using the **USStates** dataset. The size of each point is proportional to the population of the state, and the color of each point indicates the region of the state (South, West, Northeast, or Midwest). In this single figure we are showing four variables: three quantitative variables (*HouseholdIncome*, *College*, *Population*) and one categorical variable (*Region*). The regression line for predicting *HouseholdIncome* from *College* is also shown. This augmented scatterplot can be used to answer questions about any combination of these four variables.

Example 2.47

Figure 2.75 shows income, education, population, and region of US states.

(a) What is the approximate population of the state with the highest household income, and in what region is it located?

(b) On average, which region of the US has the highest rate of college education?

(c) Is there a region of the US that appears to have high household income relative to college education rate?

Solution

(a) Comparing the size of the point with highest income with the legend on the right, we see that the state's population is slightly smaller than 10 million (somewhere between 5 and 10 million), and by its color we see that it is in the northeast region. The state is Maryland.

(b) The northeast states (green points) tend to be toward the right of the plot, indicating that these states generally have higher college education rates.

(c) Western states (purple points) generally appear to have high household income relative to their college education rate. There are several purple points that are far above the regression line.

If we are not restricted to the printed page, scatterplots and other graphs can be augmented further by making them *interactive* or *dynamic*. An interactive graph can give additional information based on user input. For example, an interactive version of Figure 2.75 may give the name of the corresponding state if your mouse curser hovers over a particular point. A dynamic graph can change over time, like a movie.

A great example of augmented scatterplots that are both dynamic and interactive is the Gapminder software (*https://www.gapminder.org/tools*). This online applet can display worldwide health, economic, and environmental data aggregated by country over time. There are several variables to choose for the horizontal and vertical axes of a scatterplot, and the applet allows us to see how the scatterplot has changed historically year-by-year. By default, the color of a given point gives the world region of the country, and the size of the points give the population of the country.

Example 2.48

Gapminder: Child Mortality vs Income

In the Gapminder applet, set the vertical axis to *child mortality rate* (deaths under age 5 per 1000 births) and the horizontal axis to *income per person* (GDP in current US dollars, per person). Click the play icon and observe how the scatterplot

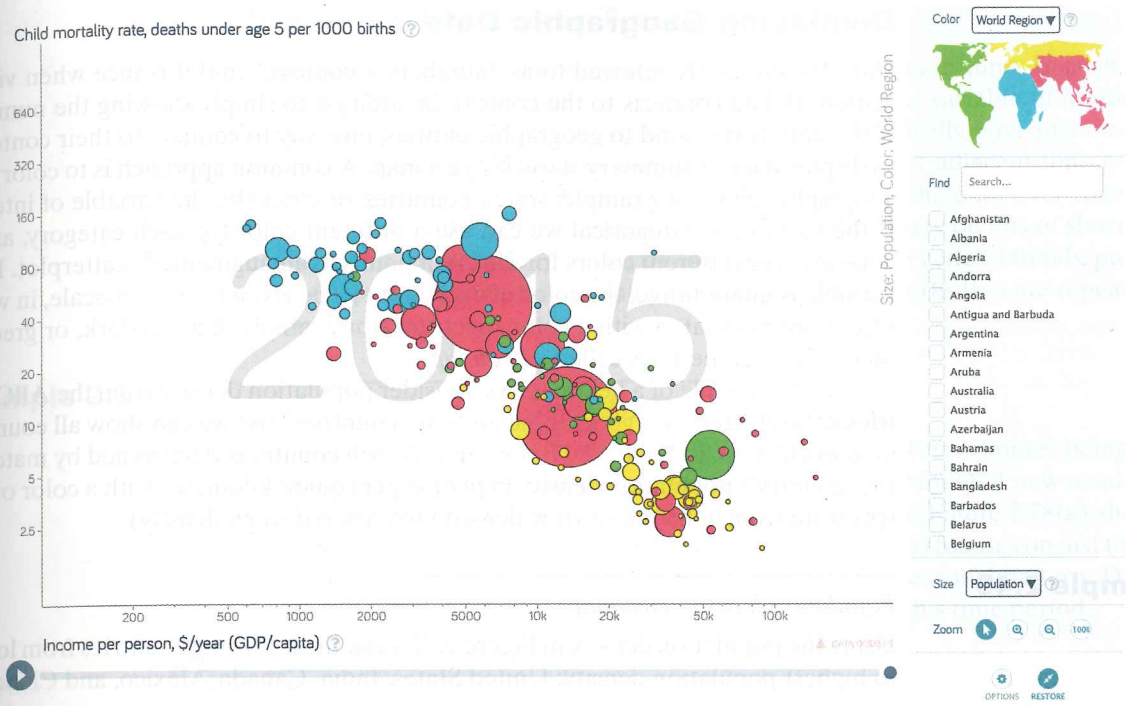


Figure 2.76 *A screenshot of the Gapminder software**

changes over time.[80] This dynamic scatterplot allows us to answer the following questions:

(a) In 2015, do child mortality rate and income per person have a positive or negative association?

(b) In 1800, which country has the largest income per person? In 2000?

(c) In 1800, what are the two largest countries by population? In 2015?

(d) In general, what is the worldwide trend in child mortality rate and income per person from 1800 to 2015?

Solution

(a) Pause at the year 2015. Child mortality rate and income per person have a clear negative association (see Figure 2.76).

(b) When we start at the first frame (1800) and hover over the rightmost point, we see that the country with the largest income per person is the Netherlands. When we pause at the year 2000, we see that the country with the highest income per person is Qatar.

(c) When we start at 1800 and hover over the two largest points, we see that they are China and India. When we do the same in 2015, we see that the two largest countries by population are still China and India.

(d) Play the graph from start to finish. In general, with each passing year we see child mortality decrease and income per person increase.

[80]For a very exciting commentary on this dynamic scatterplot, see the TED talk by Hans Rosling, a founder of Gapminder: *https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen.*

*FREE TO USE! CC- BY GAPMINDER.ORG.

Displaying Geographic Data

Data are popularly referred to as "numbers + context," and it is nice when visualization of data connects to the context, in addition to simply showing the numbers. If the data correspond to geographic entities, one way to connect to their context is to display data or summary statistics on a map. A common approach is to color each geographic unit (for example: states, countries, or cities) by the variable of interest. If the variable is categorical we can use a different color for each category, analogous to using different colors for different points in an augmented scatterplot. If the variable is quantitative, the color of each unit can be given by a color-scale, in which values are associated with a color spectrum (for example: light to dark, or green to blue); this is sometimes called a *heatmap*.

As an example of a heatmap, we consider population density from the **AllCountries** dataset. Here our geographic units are countries, and we can show all countries on a world map. In Figure 2.77, the color of each country is determined by matching that country's population density in people per square kilometer with a color on the spectrum from light yellow (low density) to dark red (high density).

Example 2.49

Population Density Heatmap

Using the population density in Figure 2.77, rank the following countries from lowest to highest population density: United States, India, Canada, Mexico, and China.

Solution

Looking at their colors, we see that Canada (light yellow) is the least dense, then the United States (yellow), then Mexico (orange), then China (light red), then India (red).

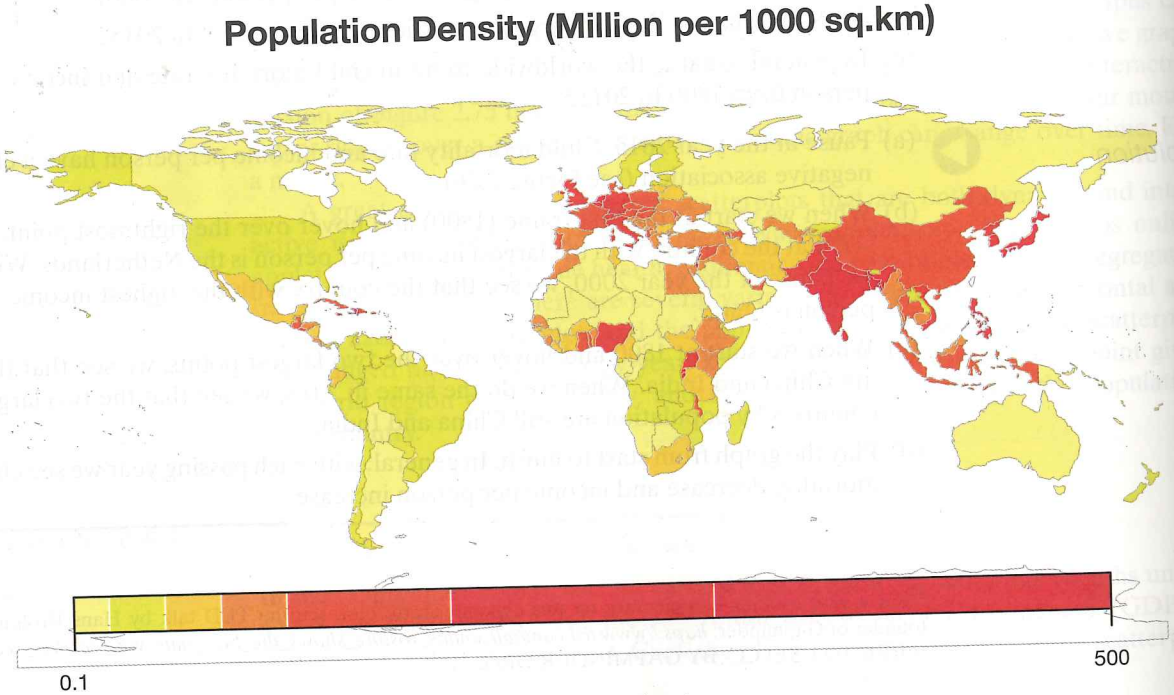


Figure 2.77 *A map of world countries, colored by population density.*

Displaying Data Over Time

In many situations, data are measured over multiple points in time. In such cases we often plot *time* on the horizontal axis and a quantitative variable of interest on the vertical axis to produce a *time series plot*. Since we typically have just one data value at each time period, we usually connect the points for adjacent time periods with line segments to help visualize trends and pattern in the data over time (and may omit plotting the points themselves.) Although there may be lots of short term fluctuations in a time series plot, we try to look for long term general trends, possible striking departures from trends, and potential seasonal patterns that are repeated at regular intervals.

Example 2.50

Quarterly Housing Starts (2000–2015)

A common measure of economic activity is the number of new homes being constructed. Data in **HouseStarts** include the number (in thousands) of new residential houses started in the US[81] each quarter from 2000 to 2015. Figure 2.78(a) shows a time series plot of the data over all 16 years and Figure 2.78(b) focuses on just the five year period from 2011 to 2015 to better see the quarterly seasonal pattern. Discuss what these plots tell us about trends in housing starts over this time period.

Solution

In Figure 2.78(a) we see that housing starts in the US generally increased until about 2006 when a substantial decline started that lasted until around 2009 (corresponding to the worldwide economic slowdown often called the Great Recession). Housing starts began to recover after 2009 and show modest increases through 2015, but still lag well below the levels of the early 2000's.

In Figure 2.78(b) we examine the recovery more closely and notice a seasonal pattern that occurs throughout the time series. Housing starts tend to be low in the winter (Q1=January-March), rise rapidly to a high level in the spring (Q2=April-June), drop slightly, but stay pretty high in the summer (Q3=July-September), and then fall quite a bit in the fall (Q4=October-December). This pattern, a clear reflection of the influence of favorable building weather, repeats over the five-year period, even as the overall trend shows a steady increase.

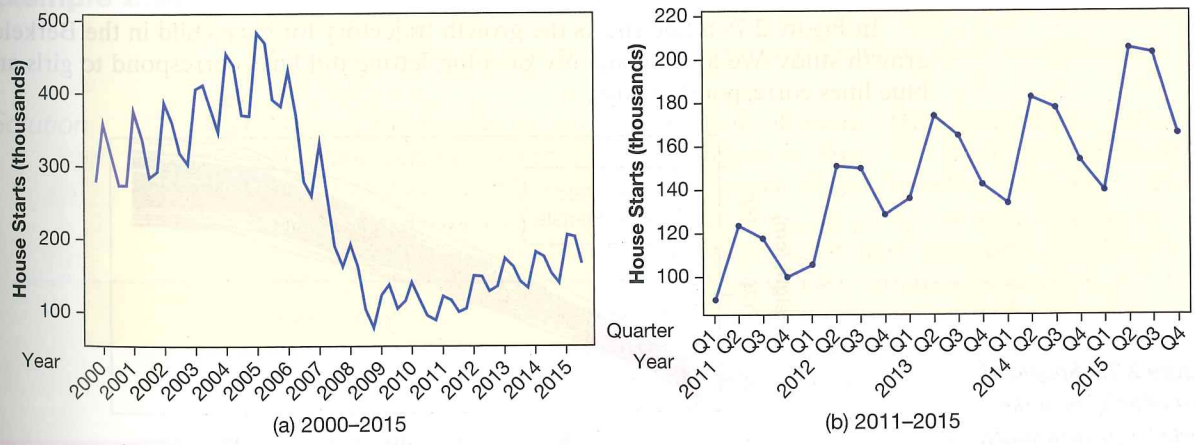


Figure 2.78 ***Quarterly housing starts in the United States***

[81]Data on housing starts downloaded from the US Census Bureau website at *https://www.census.gov/econ/currentdata/*

DATA 2.13 **Berkeley Growth Study**

In the 1940s and 1950s, the heights of 39 boys and 54 girls, in centimeters, were measured at 30 different time points between the ages of 1 and 18 years as part of the University of California Berkeley growth study.[82] The data are available in **HeightData**. ■

The Berkeley growth data involves many repeated measures over time (*age*), so it makes sense to plot the data in a chart where *age* is given on the horizontal axis, and the *height* of the children is given on the vertical axis. To display the data for a single child, we could simply make a time series plot of *height* that shows the measurements at each of the 30 different *age* times for that child. To show the data for all children at once we can use a *spaghetti plot*. In a spaghetti plot, we plot all available measurements and connect the points within each subject (in this case, each child) with a line, like many strands of spaghetti.

This is a spaghetti pot, NOT a spaghetti plot

In Figure 2.79 a line shows the growth trajectory for each child in the Berkeley growth study. We again make use of color, letting red lines correspond to girls and blue lines correspond to boys.

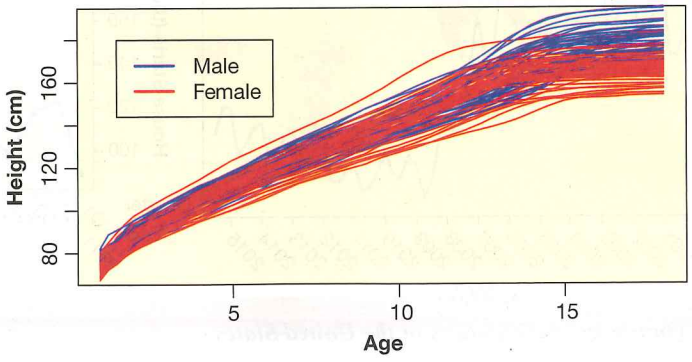


Figure 2.79 ***Spaghetti plot of heights in the Berkeley growth study, colored by gender***

[82] Tuddenham, R.D., and Snyder, M.M., (1954). "Physical growth of California boys and girls from birth to age 18", *University of California Publications in Child Development*, 1, 183–364.

Example 2.51

Using Figure 2.79, showing childhood growth:

(a) Generally when do girls tend to stop growing in height? When do boys stop growing in height?

(b) Was the tallest participant in this study from ages 2 to 12 a boy or a girl?

Solution

(a) For girls, growth appears to slow substantially between ages 13 and 14 (after this, the slope of the red lines are very small). For boys, growth appears to slow between the ages of 15 and 16.

(b) A red line is above all others between age 2 and age 12, so the tallest individual between these ages is a girl.

Breaking it Down

Many of the basic plots that we've learned in previous sections can easily be extended just by breaking it down by a categorical variable. In other words, instead of looking at one plot containing all the cases, we break it down into several plots, where each plot includes only the cases within a certain subgroup. Often, breaking the data down into subgroups can reveal new and important insights.

DATA 2.14 **Discrimination among the Developmentally Disabled?**

The California Department of Developmental Services (DDS) allocates funds to support developmentally disabled California residents (such as those with autism, cerebral palsy, or intellectual disabilities) and their families. We'll refer to those supported by DDS as DDS consumers. An allegation of discrimination was made when it was found that average annual expenditures were substantially lower for Hispanic consumers than for white non-Hispanic consumers (who, for simplicity below, we refer to simply as white consumers.) The dataset **DDS** includes data on annual expenditure (in $), ethnicity, age, and gender for 1000 DDS consumers.[83] Do these data provide evidence of discrimination? ■

Example 2.52

Expenditure by Ethnicity

Compare annual expenditure for Hispanic consumers versus white consumers.

Solution

Figure 2.80 provides a visual comparison of annual DDS expenditure values for Hispanic consumers versus white consumers. From the graph, it is immediately apparent that expenditures tend to be much higher for white consumers than for Hispanic consumers. Computing means for each group, we find that the average annual expenditure is $11,066 for Hispanic consumers, as opposed to $24,698 for white consumers. Based on this analysis of two variables, it appears that these data provide very strong evidence of discrimination (we could formalize this with techniques we'll learn in Chapter 4 or 6; this difference is *extremely* significant), with expenditures much higher, on average, for white non-Hispanics than for Hispanics.

[83] Taylor, S.A. and Mickel, A.E. (2014). "Simpson's Paradox: A Data Set and Discrimination Case Study Exercise," *Journal of Statistics Education*, **22**(1). The dataset has been altered slightly for privacy reasons, but is based on actual DDS consumers.

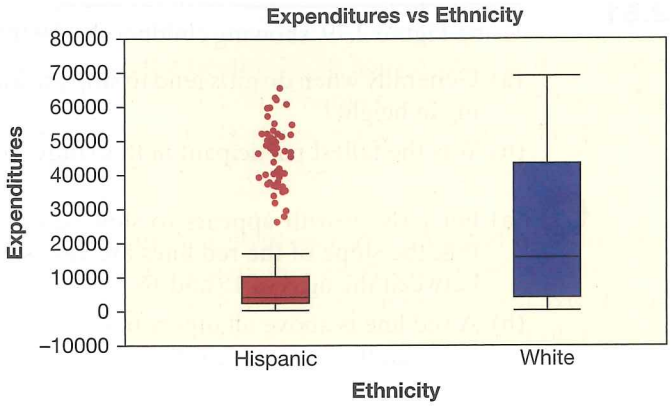


Figure 2.80 *Annual expenditure for Hispanic consumers vs white consumers*

An initial bivariate exploration of expenditure by ethnicity finds very strong evidence of discrimination. What happens if we "break it down" by age? To make the data easier to visualize in separate groups, we'll work with age group categories rather than age itself.

Example 2.53

Expenditure by Ethnicity, Broken Down by Age Group

Compare annual expenditure for Hispanic consumers versus white consumers, broken down by age group.

Solution

Figure 2.81 provides a visual comparison of annual expenditure values for Hispanic residents versus white residents, broken down by age group. Each set of side-by-side boxplots can be interpreted in the usual way, except that now we have a separate set of side-by-side boxplots for each age group (with age groups shown along the top). Take a close look at this plot, and in particular, look at the comparison of Hispanic to white expenditures *within each age group*. We now see that, within *each* age group, expenditures are actually *higher* for Hispanics than for whites!

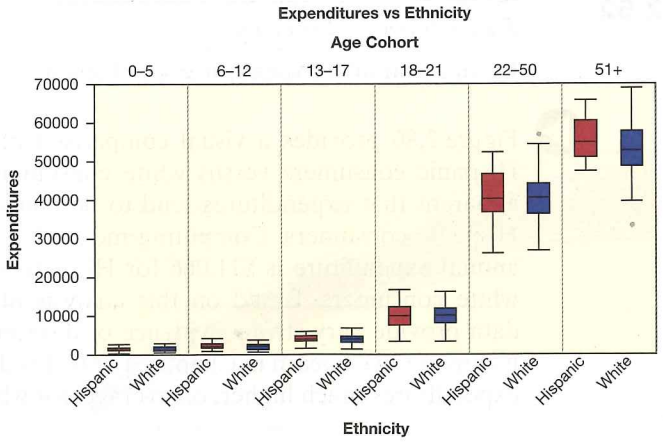


Figure 2.81 *Expenditure for Hispanic vs white, by age group*

In Example 2.53 we find that within age groups, expenditures are consistently higher for Hispanics than for whites, yet in Example 2.52 we find that expenditures overall are higher for whites than for Hispanics. How is this possible?? The explanation lies in one additional data visualization.

Example 2.54

Ethnicity by Age Group

Compare ethnicity counts by age group.

Solution

Figure 2.82 shows a bar chart displaying the number of people within each ethnic category, broken down by age group. For example, the first red bar goes up to 44 on the y-axis, showing that there are 44 Hispanic consumers within the 0-5 years old age group. This plot shows that the sample contains many more Hispanic children than white children, and many more white adults than Hispanic adults.

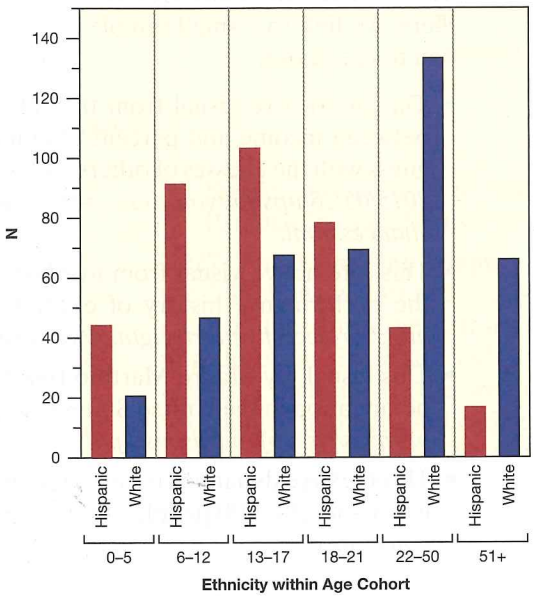


Figure 2.82 *Ethnicity counts by age group*

Combining the information from Figure 2.81 and Figure 2.82, we see that white consumers tend to be older than Hispanic consumers (Figure 2.82), and older consumers tend to receive higher expenditures than younger consumers (Figure 2.81). This explains why the white consumers receive higher expenditures overall: not because of discrimination, but just because they tend to be older. These visualizations are shown together in Figure 2.83, illustrating this point. Here age is a confounding variable, as introduced in Section 1.3: age is associated with both ethnicity and expenditure, confounding the relationship. Failing to account for the

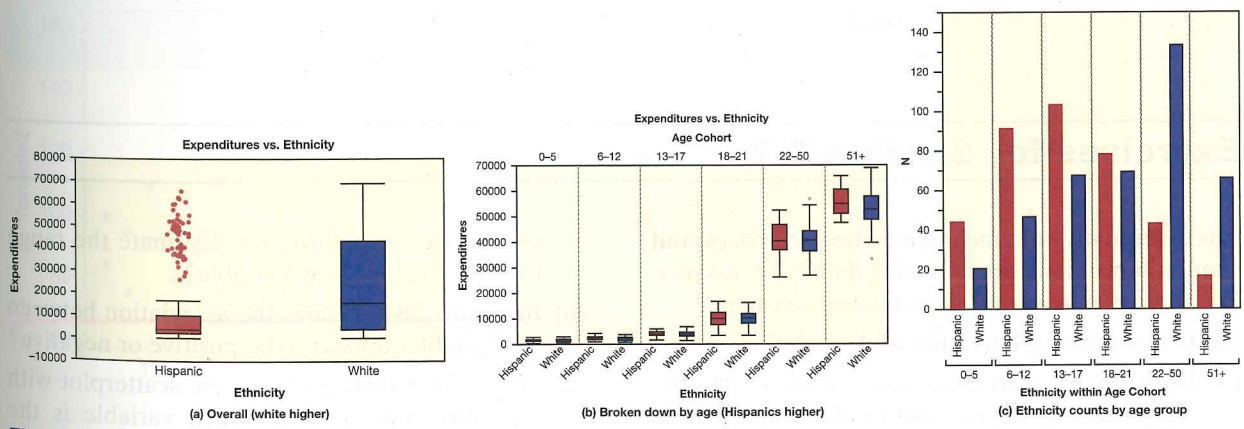


Figure 2.83 *Comparing the visualizations of DDS expenditure by ethnicity, with age as a confounding variable*

confounding variable can paint a very misleading picture. Luckily, DDS called in a statistician, and was not falsely charged with discrimination.

This is an example of *Simpson's Paradox*, which occurs when the relationship between two variables switches when the data are broken down by a third variable. Although a true reversal is rare, it is often true that the relationship between two variables differs when a third variable is taken into account. Incorporating more than just two variables can be important for revealing the true story the data have to tell.

Other Visualization Examples

The internet has many creative, interactive, and fascinating visualizations of data. Here, we link to a small sample of great visualization examples that we encourage you to check out:

- This interactive visual from the *NY Times* allows you to guess the relationship between income and percent of children attending college, then compares your guess with the guesses of others and the reality: *http://www.nytimes.com/interactive/2015/05/28/upshot/you-draw-it-how-family-income-affects-childrens-college-chances.html.*
- This interactive visual from *fivethirtyeight.com* is an interactive spaghetti plot of the performance history of every team in the national basketball association: *http://projects.fivethirtyeight.com/complete-history-of-the-nba/.*
- This visual, by Mauro Martino (*mamartino.com*), shows the increase in political polarization in the United States Congress since 1948: *http://www.mamartino.com/projects/rise_of_partisanship/.*
- This creative dynamic visualization by Nathan Yau (*flowingdata.com*) shows a day in the life of 1,000 people : *http://flowingdata.com/2015/12/15/a-day-in-the-life-of-americans/.*

SECTION LEARNING GOALS

You should now have the understanding and skills to:

- Interpret information from a variety of data visualizations
- Recognize ways to include multiple variables, and the value of including additional variables, in a display
- Recognize ways to include geographic data in a display
- Recognize ways to display time-dependent data
- Recognize that there are many effective and creative ways to display data

Exercises for Section 2.7

There are many links included in these exercises, and some will probably break during the life of this text. We apologize in advance for any inconvenience.

2.226 Considering Direction of Association

(a) How many variables are included in the scatterplot in Figure 2.84(a)? Identify each as categorical or quantitative. Estimate the range for Variable1 and for Variable2.

(b) In Figure 2.84(a), does the association between the variables appear to be positive or negative?

(c) Figure 2.84(b) shows the same scatterplot with regression line added. Which variable is the

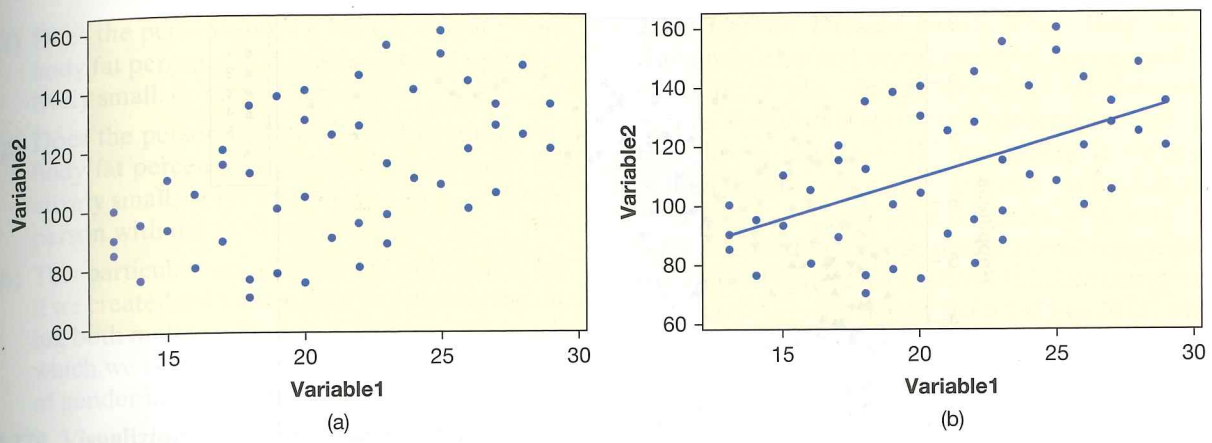


Figure 2.84 *Describe the association between these variables*

response variable? Does the line show a positive or negative association?

(d) Figure 2.85(a) shows the same scatterplot with a third variable included. Is the new variable categorical or quantitative? If categorical, how many categories? If quantitative, estimate the range.

(e) In Figure 2.85(a), if we only consider cases in Group A, does the association between Variable1 and Variable2 appear to be positive or negative? How about in Group B? Group C? Group D?

(f) Figure 2.85(b) shows the same scatterplot as Figure 2.85(a) with regression lines added within each of the four groups. Does the regression line for Group A show a positive or negative association? How about Group B? Group C? Group D?

(g) What happens to the direction of association shown in Figure 2.84 when we add the additional information contained in Variable3 as in Figure 2.85? (This is an example of Simpson's Paradox for quantitative variables.)

2.227 Visualizing the Happy Planet Index Figure 2.86 shows a scatterplot illustrating three different variables from the dataset **HappyPlanetIndex**, introduced in Exercise 2.187. The variable *Happiness* is a measure of the well-being of a country, with larger numbers indicating greater happiness, health, and well-being. The variable *Footprint* is a per capita measure of the ecological impact of a country on the environment, with larger numbers indicating greater use of resources (such as gas and electricity) and more damage to the planet. A third variable, *Region*, is given by the code shown in the top right, and is categorized as follows: 1 = Latin America, 2 = Western nations, 3 = Middle East, 4 = Sub-Saharan Africa, 5 = South Asia, 6 = East Asia, 7 = former Communist countries.

(a) Classify each of the three variables as categorical or quantitative.

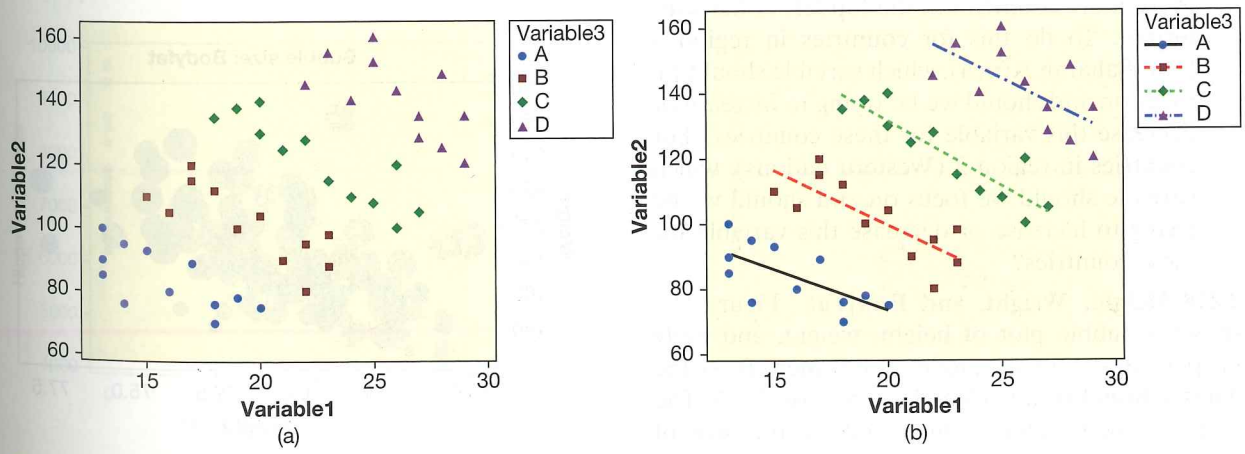


Figure 2.85 *Describe the association in each group*

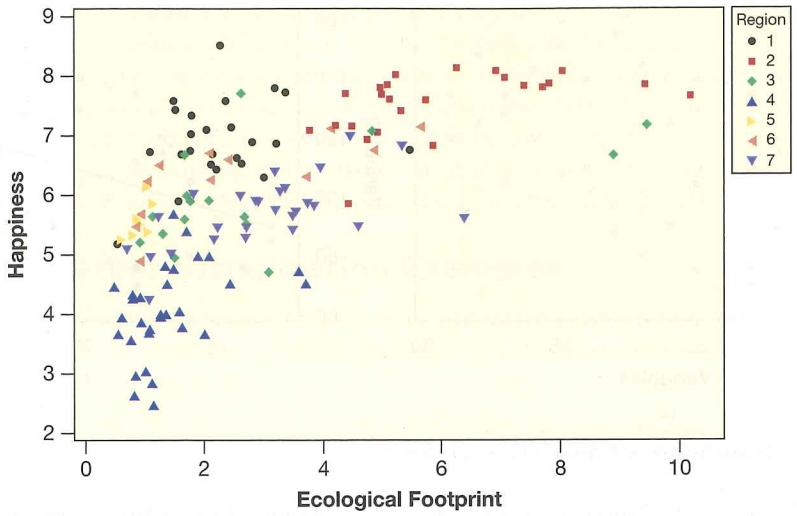


Figure 2.86 *Happiness, ecological footprint, and region*

(b) Which two regions seem to have the greatest happiness score? Which region seems to have the greatest ecological footprint?

(c) Which region seems to have the lowest happiness score? Does the ecological footprint tend to be high or low in that area?

(d) Looking at the scatterplot overall and ignoring region, does there appear to be a positive relationship between happiness score and ecological footprint?

(e) Considering only region 2 (Western nations), does there appear to be a positive relationship between happiness score and ecological footprint?

(f) The country with the highest happiness score is Costa Rica. Is it in the top left, top right, bottom left, or bottom right of the scatterplot?

(g) According to Nic Marks, the developer of the Happy Planet Index, we should be trying to move more countries to the top left of the scatterplot. To do this for countries in region 4 (Sub-Saharan Africa), which variable should we focus on and should we be trying to increase or decrease this variable for these countries? For countries in region 2 (Western nations), which variable should we focus on and should we be trying to increase or decrease this variable for these countries?

2.228 Height, Weight, and BodyFat Figure 2.87 shows a bubble plot of height, weight, and body fat percentage for a sample of 100 men, from the dataset **BodyFat**, introduced in Exercise 2.219. The body fat percentage is indicated by the size of the bubble for each case.

(a) How many variables are shown in the scatterplot? Identify each as categorical or quantitative.

(b) Ignoring bubble size, does there appear to be a positive or negative relationship between height and weight?

(c) Do the bubbles tend to be larger on the top half of the scatterplot or the bottom half? Interpret this in context and in terms of the relevant variables.

(d) Body fat percentage depends on more than just height and weight. There are two cases who are about 66 inches tall, one weighing about 125 pounds and the other about 140 pounds. Which has the larger body fat percentage?

(e) There are two cases weighing about 125 pounds, one about 66 inches tall and the other about 67 inches tall. Which has a larger body fat percentage?

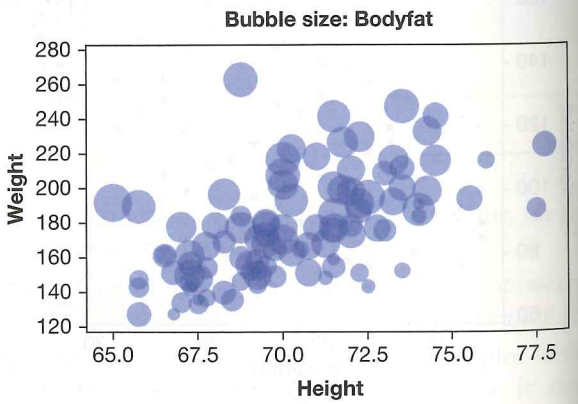


Figure 2.87 *Height, weight, and body fat*

(f) Does the person with the largest weight have a body fat percentage that is relatively large, relatively small, or pretty average?

(g) Does the person with the largest height have a body fat percentage that is relatively large, relatively small, or pretty average? How about the person with the third largest height?

(h) This particular sample included only males, but if we created a similar graph for a dataset including both males and females, indicate one way in which we could incorporate the fourth variable of gender in the graph.

2.229 Visualizing Football and Brain Size Exercise 2.143 introduces a study in which the number of years playing football and the size of the hippocampus in the brain were recorded for each person in the study. There were three different groups in the study: football players who had been diagnosed with at least one concussion, football players who had never been diagnosed with a concussion, and a control group of people who had never played football. Figure 2.88(a) shows a graph that incorporates all three of these variables.

(a) Identify each variable as quantitative or categorical.

(b) Why are all the blue dots stacked up on the left?

(c) Overall, does there appear to be a positive or negative association (or no association) between years playing football and hippocampus size?

(d) Figure 2.88(b) shows the same graph with regression lines for the two groups of football players. Which of the groups has the line that is lower on the graph? What does this tell us in the context of the three variables?

(e) Which of the groups has the line with the steeper slope?

2.230 Carbon Dioxide Levels Over Time Scientists are concerned about global warming and the effect of carbon dioxide emissions on the atmosphere. Figure 2.89 shows the concentration of carbon dioxide (CO2) in the atmosphere, in parts per million (ppm), over two different time intervals. Often, plots of data over time can look very different depending on the time interval selected for the graph. Figure 2.89(a) shows the concentration[84] of CO2 during the period from 1959 to 2015, while Figure 2.89(b) shows the concentration[85] over a very different window: the last 400,000 years!!

(a) Using Figure 2.89(a), estimate the CO2 concentration in 1960. Estimate the CO2 concentration in 2015.

(b) Is CO2 concentration primarily increasing, decreasing, or oscillating up and down during the period from 1959 to 2015?

(c) In the period of time shown in Figure 2.89(b), is CO2 concentration primarily increasing, decreasing, or oscillating up and down?

(d) In Figure 2.89(b), locate the portion of data that is shown in Figure 2.89(a). What does the curve look like on that piece?

(e) What was the highest concentration of CO2 ever in the 400,000 year history of the Earth, before 1950?

(f) For more data visualization on this subject, watch the 3-minute animated video "CO2 Movie" at *http://www.esrl.noaa.gov/gmd/ccgg/trends/history.html*.

[84]Dr. Pieter Tans, NOAA/ESRL (*www.esrl.noaa.gov/gmd/ccgg/trends/*) and Dr. Ralph Keeling, Scripps Institution of Oceanography (*scrippsco2.ucsd.edu/*).

[85]Data from the Vostok Ice Core project, Barnola, J.M., Raynaud, D., Lorius, C., and Barkov, N.I., *http://cdiac.ornl.gov/ftp/trends/co2/vostok.icecore.co2*

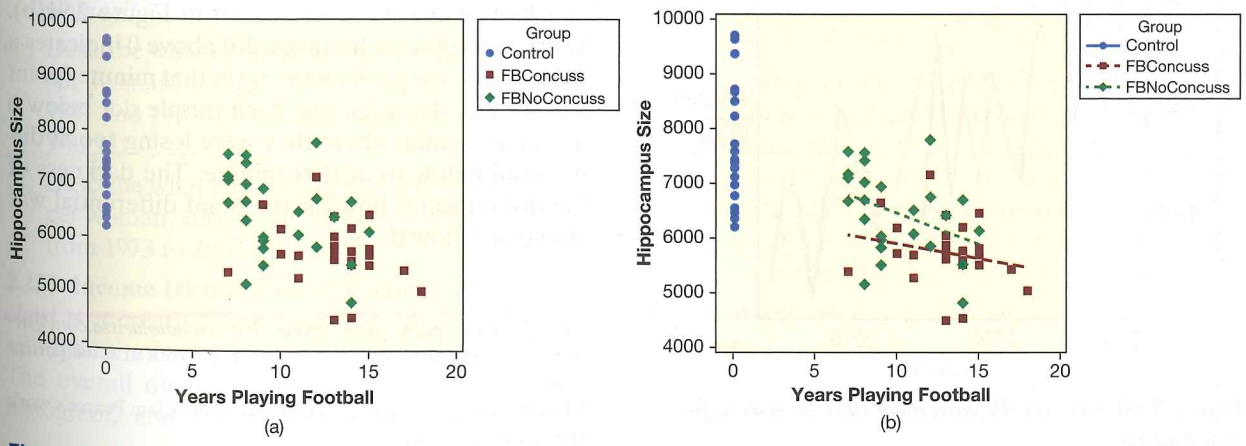


Figure 2.88 *Brain size, football experience, and concussions*

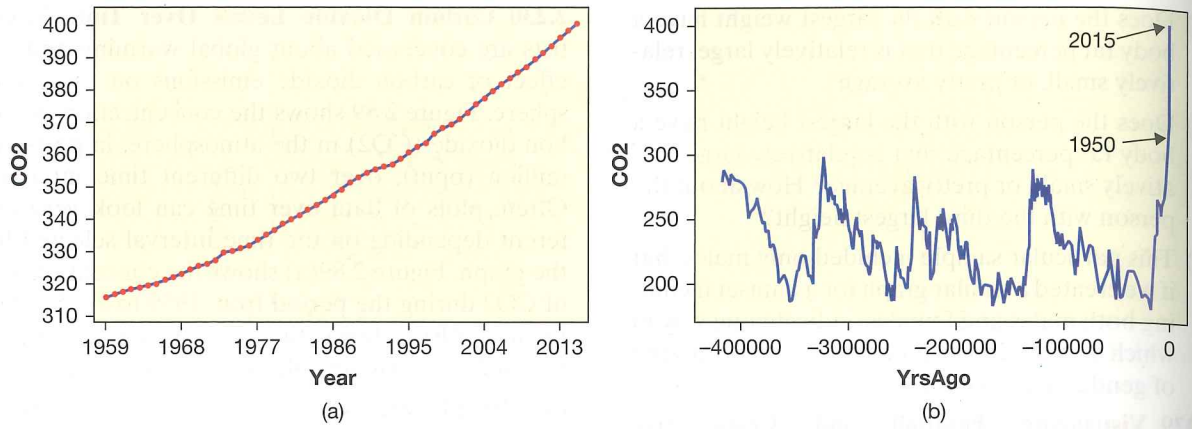


Figure 2.89 *Carbon dioxide levels over time, in two very different windows*

2.231 Forty-Yard Dash at the NFL Combine Every year the National Football League invites 335 draft eligible college football players to a scouting combine where they participate in a variety of drills and exercises. One of the more popular drills, called the 40-yard dash, is the time it takes each player to run 40-yards. We computed the average 40-yard dash time every season from 1990 to 2016 for all players at two positions: wide receiver (WR), who receive passes from the quarterback, and defensive cornerback (DC), who try to stop the wide receivers from catching the ball. These data are presented as a spaghetti plot in Figure 2.90.

(a) Describe the general trend of both positions. Are the players getting faster or slower?

(b) In 2016 which position had a faster average 40 time?

(c) Does one position appear to be consistently faster than the other?

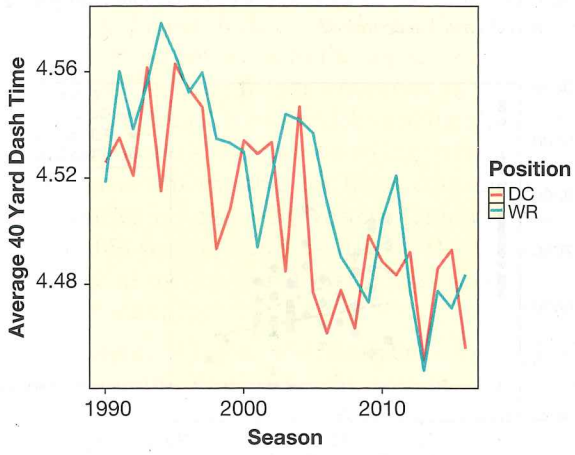


Figure 2.90 *Average 40 yard dash time by season for WR and DC*

2.232 Golden State Warriors: One Game During a record breaking season the Golden State Warriors of the National Basketball Association (NBA) won 24 straight games to start the 2015–2016 season. Adam Pearce plotted the point differential, Golden State points – Opponent points, each minute of the first 16 games of the streak.[86] One of those games, a 119 to 104 victory on November 6th, is plotted in Figure 2.91(a).

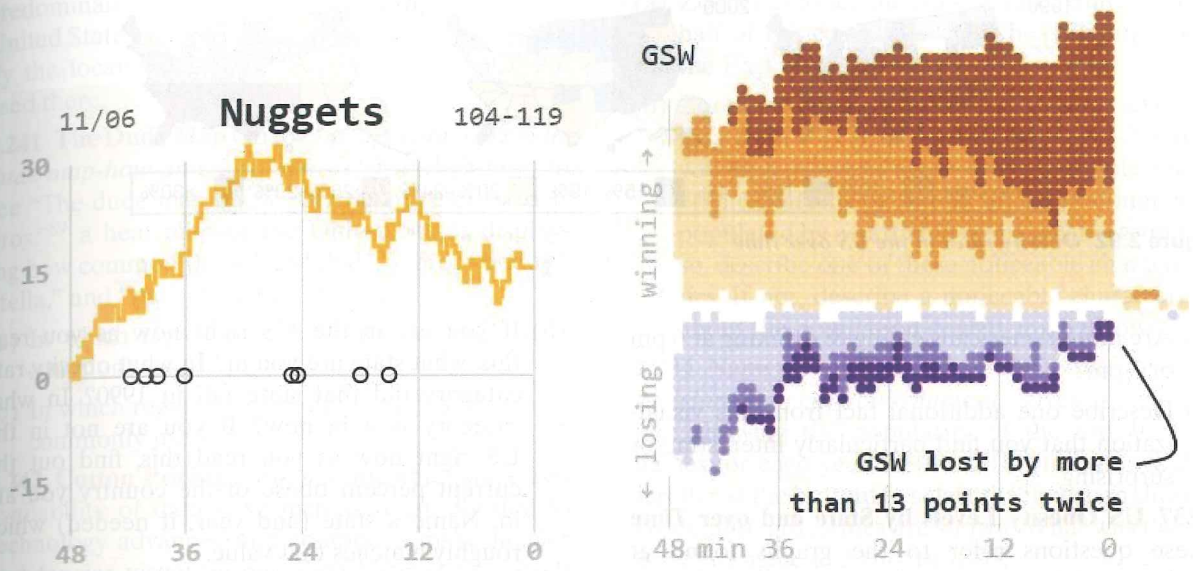
(a) Were the Warriors ever losing in this game (point differential below 0)?

(b) The game is split into quarters, demonstrated by the minutes remaining where 48 to 36 is 1st quarter, 36 to 24 is 2nd quarter, 24 to 12 is 3rd quarter, 12 to 0 is 4th quarter. In which quarter did the Warriors have their largest lead?

2.233 Golden State Warriors: First Half of Season Exercise 2.232 plotted the Golden State Warriors point differential, Golden State points – Opponent points, each minute of one game during their record breaking 2015–2016 season. Adam Pearce[87] of *roadtolarissa.com* also plotted the game state (winning or losing or tied) every minute of every game over the first half of the Warriors season in Figure 2.91(b). Within this plot each orange dot above 0 indicates a game where they were winning at that minute (point differential above 0), and each purple dot below 0 indicates a game where they were losing (point differential below 0) at that minute. The darkness of the dot indicates how far the point differential was above or below 0.

[86] Used with permission from *http://roadtolarissa.com/gsw-streak/*. Check out more of Adam Pearce's work at *roadtolarissa.com*.

[87] Used with permission. Check out more of Adam Pearce's work at *roadtolarissa.com*.



(a) Over one game

(b) Over 41 games

Figure 2.91 *Score difference by minutes*

(a) How many games were the Warriors losing at the 24-minute mark (halftime)?

(b) How many games were the Warriors losing at the 0-minute mark (the end of the game)?

2.234 Marriage Age vs Number of Children Using the Gapminder software (*https://www.gapminder.org/tools*), set the vertical axis to *Age at 1st marriage (women)* and the horizontal axis to *Babies per woman*. This scatterplot shows the mean age at which woman marry, and the mean number of children they have, for various countries. Click the play icon and observe how the scatterplot changes over time, then answer the following questions:

(a) Overall, is there a positive or negative association between *Age at 1st marriage* and *Babies per woman*?

(b) Describe what happens to the number of babies per woman, and age at 1st marriage, between 1941 and 1943 in Russia (at the height of World War II).

(c) Describe what happens to the number of babies per woman, and the age at 1st marriage, in Libya from 1973 to 2005.

2.235 Income Distribution by Country A *mountain chart* is a creative way to display the distribution of a quantitative variable over different categories. The overall distribution is shown as a smoothed histogram, and the area underneath is colored according to different categories. The distribution of the variable for each category corresponds to the size of its colored area. The Gapminder software includes a mountain chart that shows the distribution of income broken down by world region and country (*http://www.gapminder.org/tools/mountain*). Click the play icon to see how the mountain chart changes from 1800 to present day, then answer the following questions:

(a) Extreme poverty is defined as living on less than $2 a day. In 1970, the majority of people living in extreme poverty came from which world region?

(b) In 2015, the majority of people living in extreme poverty came from which world region?

(c) Describe the shape of the worldwide distribution of income in 1970. In 2015.

2.236 A Day in the Life This creative dynamic visualization by Nathan Yau (*flowingdata.com*) shows a day in the life of 1000 different representative Americans based on survey data: *http://flowingdata.com/2015/12/15/a-day-in-the-life-of-americans/*. Watch the dynamic visualization over the entire day, then answer the following questions:

(a) At 6 am, what are the majority of Americans doing?

(b) At 10 am, what is the most common activity for Americans?

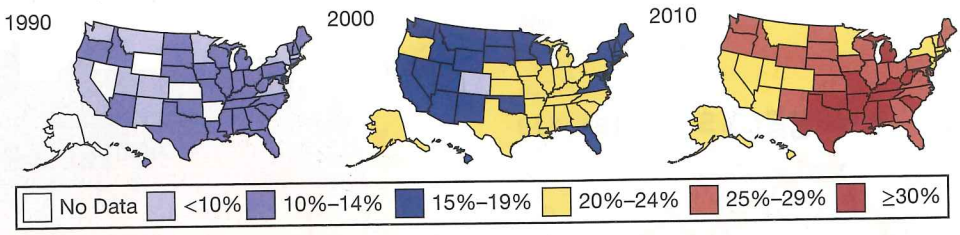


Figure 2.92 *Obesity rates in the US over time*

(c) Are more Americans eating & drinking at 6 pm or 7 pm?

(d) Describe one additional fact from this visualization that you find particularly interesting or surprising.

2.237 US Obesity Levels by State and over Time These questions refer to the graphs found at *http://stateofobesity.org/adult-obesity/* which show a sequence of maps of US states, colored by the proportion of the adult population classified as obese, for many different years between 1990 and 2014. Three of the maps, from 1990, 2000, and 2010, as well as a key for the color coding of percent obese, are shown in Figure 2.92.

(a) The cases are US states and one variable is year. Another variable is percent obese in that state, which could be a quantitative variable but here is classified into categories, making it a categorical variable. Using the legend given in Figure 2.92 (and ignoring the category of "No Data"), how many different categories are shown? (There is one more category already in the online graphs as this book goes to print and maybe another one by the time you are reading this!)

(b) Using Figure 2.92, what appears to be the highest category needed for any state in 1990? In 2000? In 2010?

2.238 US Obesity Levels by State over Many Years Exercise 2.237 deals with some graphs showing information about the distribution of obesity rates in states over three different years. The website *http://stateofobesity.org/adult-obesity/* shows similar graphs for a wider selection of years. Use the graphs at the website to answer the questions below.

(a) What is the first year recorded in which the 15-19% category was needed, and how many states are in that category in that year? What is the first year the 20-24% category was needed? The 25-29% category? The 30-34% category? The 35%+ category?

(b) If you are in the US right now as you read this, what state are you in? In what obesity rate category did that state fall in 1990? In what category is it in now? If you are not in the US right now as you read this, find out the current percent obese of the country you are in. Name a state (and year, if needed) which roughly matches that value.

2.239 Spaghetti Plots of US Obesity Levels by State over Many Years Exercises 2.237 and 2.238 look at geographic plots of obesity rates in different years. The website *http://stateofobesity.org/adult-obesity/* also shows a spaghetti plot (on the right) which tracks the obesity rate of each state for the years from 1990. Hovering over any strand highlights that state (click to select it or click on the state in the map) and then you can point along the strand to see the obesity rate for that year. Use this graph to answer the questions below.

(a) Between 1990 and 2014, did the percent obese more than double: for every state or no states or just some states?

(b) Is there more variability in obesity rates between states in 1990 or in 2014?

(c) Identify the state with the largest percent obese in 1990, and give the state name and the percent obese at that time. In addition, identify the state with the smallest percent obese in 2014, and give the state name and the percent obese at that time.

2.240 What Do You Call a Sweetened Carbonated Beverage? If you reach for a sweetened carbonated beverage, do you refer to it as soda, pop, coke, or a soft drink? Different regions of the United States use different terms, as shown in this heat map: *discovermagazine.com/galleries/2013/june/regional-us-language-dialect.*[88] If you live in the United States, specify where you live and which term is

[88] "Soda or Pop? Maps Show Americans' Colorful Dialect Differences," *Discover Magazine*, 6/7/13. *discovermagazine.com/galleries/2013/june/regional-us-language-dialect.* Visualization by Joshua Katz (NC State University), Data from Bert Vaux (Cambridge University).

predominantly used there. If you do not live in the United States, choose a location in the US and specify the location and which term is predominantly used there.

2.241 The Dude Map Go to *http://qz.com/316906/the-dude-map-how-american-men-refer-to-their-bros/* to see "The dude map: How Americans refer to their bros,"[89] a heat map of the United States displaying how common the words "dude," "bro", "buddy," "fella," and "pal" are across the US.

(a) In which region of the country is "bro" most commonly used?

(b) In which region of the country is "buddy" most commonly used?

2.242 Cotton Pricing Although the abundance and availability of data have increased rapidly due to technology advances, and computers make fancier and fancier visualizations, data visualization itself is not new. This link *http://www.handsomeatlas.com/us-census-statistical-atlas-1880/manufactures-specific-cotton-goods* provides a data visualization of the consumption and price of cotton from 1880. Zoom in on the graph to better see certain features.

(a) In 1880, which state spent the most money (per capita) on cotton? Which spent the least?

(b) In which year (1825 to 1880) was cotton most expensive?

2.243 A Map of All Americans! Visualization can often be an effective way to make sense of very large datasets. The Census Dot Map at *http://demographics.coopercenter.org/DotMap/* displays the race and location of *every* American recorded by the 2010 Census; that's over 300 million data points displayed simultaneously on one map! If you were counted in the US Census in 2010, you can find your dot on the map!

(a) Zoom out to see the entire United States. Which half of the country is more heavily populated, the East or the West?

(b) Zoom in to see the town/city your school is located in.[90] (Click on "Add map labels" to help you find it.) Are there any noticeable racial/ethnic patterns, with any areas predominantly populated by a particular racial/ethnic group? If so, describe one of these noticeable characteristics. If not, describe a noticeable characteristic about the population density in your town.

2.244 Names over Time The website *http://www.visualcinnamon.com/babynamesus* gives a spaghetti plot showing the popularity of the top 10 baby names for each year 1880 to 2014 (use the window scroller at the bottom to select the timespan shown). By default, girl names are shown. What was the top baby girl name in 2014? In 1880?

FISHER'S IRIS DATA

Exercises 2.245 to 2.247 refer to the data in **FisherIris**, from a paper published in 1936 by Sir R.A. Fisher, widely considered the father of modern statistics.[91] The cases are 150 irises and there are five variables: Type of iris is categorical, while petal width, petal length, sepal width, and sepal length (all in millimeters) are quantitative. Sepals are the green leaves underneath the petals, providing support for the petals.

2.245 Petal Length and Petal Width Figure 2.93(a) shows a scatterplot of the two quantitative variables petal length and petal width.

(a) Explain how the scatterplot appears to show at least two different types of irises.

[89] Sonnad, N., "The dude map: How Americans refer to their bros," *Quartz*, 12/23/14.

[90] If outside the US, pick a city to look at and specify which city you pick.

[91] Fisher, R.A., "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, 7(2), 1936, 179–188.

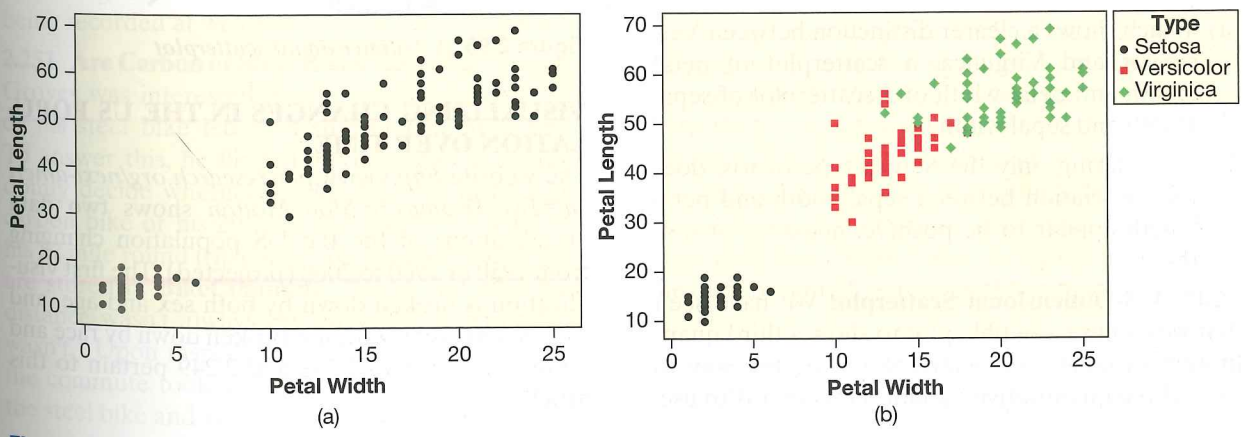


Figure 2.93 *Petal length and petal width*

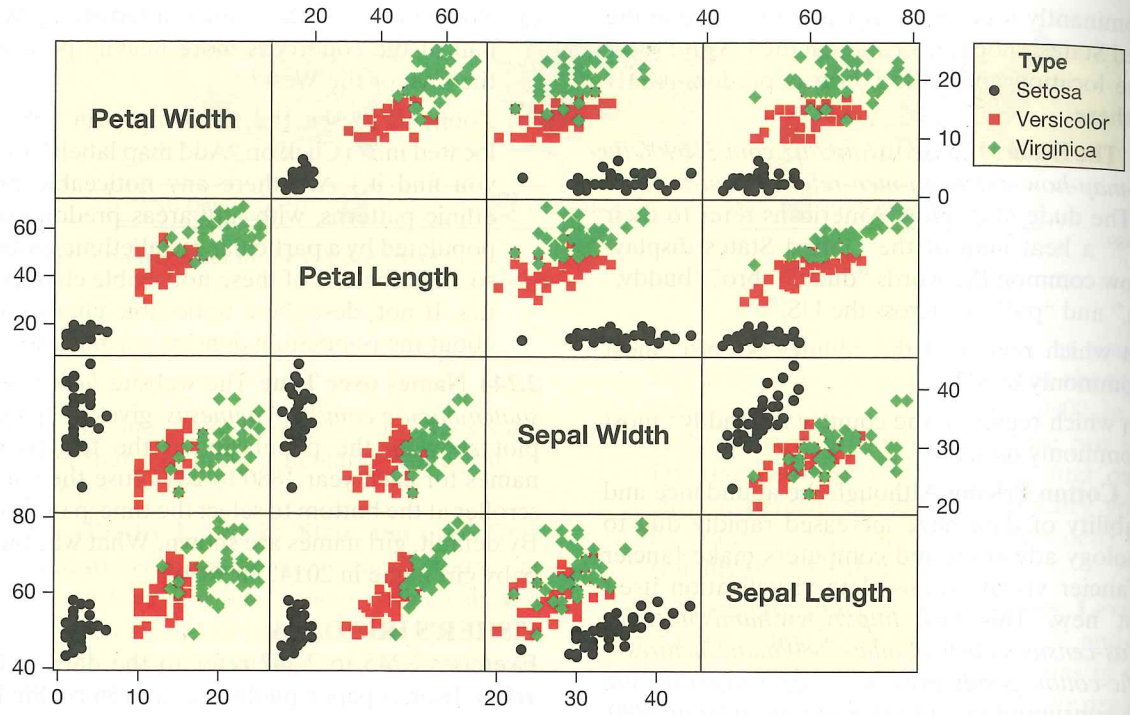


Figure 2.94 *A scatterplot matrix*

(b) Figure 2.93(b) shows the same scatterplot with the *Type* categorical variable included. We can now see the that there are three different types included: Setosa, Versacolor, and Virginica. Which type has the smallest petals? Which type has the largest petals?

2.246 A Scatterplot Matrix There are five variables in the dataset, and all are included in the *scatterplot matrix* shown in Figure 2.94. The graph shows scatterplots for each pair of quantitative variables, with the *Type* categorical variable included on each. For example, the second one down on the left is the same scatterplot as in Figure 2.93(b), with petal width on the horizontal axis and petal length on the vertical axis.

(a) Which shows a clearer distinction between Versicolor and Virginica: a scatterplot of petal length and petal width or a scatterplot of sepal length and sepal width?

(b) Considering only the Setosa type of iris, does the association between sepal width and petal length appear to be positive, negative, or neither?

2.247 A 3-Dimensional Scatterplot We have seen that we can use a bubble plot to show a third quantitative variable on a scatterplot. Another way to show three quantitative variables together is to use a *3-dimensional scatterplot*, such as the one showing petal length, petal width, and sepal length in Figure 2.95. In this case, which variable is on the vertical axis? Which color dots are highest up (meaning they have the largest values for that variable)?

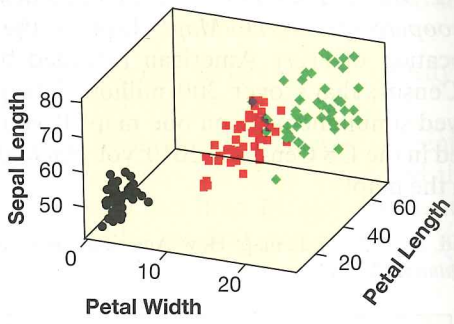


Figure 2.95 *A 3-dimensional scatterplot*

VISUALIZING CHANGES IN THE US POPULATION OVER TIME

The website *http://www.pewresearch.org/next-america/#Two-Dramas-in-Slow-Motion* shows two data visualizations of the the US population changing from 1950 or 1960 to 2060 (projected). The first visualization is broken down by both sex and age, and the second visualization is broken down by race and ethnicity. Exercises 2.248 and 2.249 pertain to this article.

2.248 Look at the dynamic visualization titled "U.S. Age Pyramid Becomes a Rectangle" (the first visualization on the page). The data visualization is essentially histograms of age, but the histograms are turned sideways and broken down by sex, comparing males and females.

(a) Go to View Data for 1950 (the baby boomer band will be at the bottom). Describe the distribution of ages (for either sex) in 1950.

(b) In 1950, is the age distribution left-skewed, right-skewed, or symmetric? (Think of what the histogram would look like on a regular number line with 0 on the left.)

(c) Click the Animation Control button to watch the distribution change over time. How does the (projected) age distribution in 2060 differ from the age distribution in 1950? (Note the age distribution in 2060 is close to what is referred to as a "uniform" distribution, a histogram that is essentially flat.)

(d) In 2060, are there projected to be more males or females in the 85+ range?

2.249 Look at the visualization titled "Changing Face of America" (the second visualization on the page). This is a new kind of visualization in which the total for each year is scaled to 100%, and the colors are shaded according to the percentage of the population comprised of each racial/ethnic group.

(a) Which racial/ethnic group is decreasing the most in terms of percentage of the US population?

(b) Which racial/ethnic group is increasing the most in terms of percentage of the US population?

(c) Which racial/ethnic group is staying the most constant in terms of percentage of the US population?

2.250 The Wind Map The website *hint.fm/wind/* shows the current wind patterns across the US. In order to generate this map, what two variables are being recorded at weather stations across the US?

2.251 Are Carbon or Steel Bikes Faster? Dr. Jeremy Groves was interested in whether his carbon bike or his steel bike led to a shorter commute time. To answer this, he flipped a coin each day to randomly decide whether to ride his 20.9 lb (9.5 kg) carbon bike or his 29.75 lb (13.5 kg) steel bike for his 27 mile round trip commute. His data for 56 days are stored in **BikeCommute**. (We're not sure why distance wasn't always the same, but apparently it wasn't.) Upon inspection of the data, he finds that the commute took an average of 107.8 minutes on the steel bike and 108.3 minutes on the carbon bike, suggesting the steel bike is faster, but he also finds that the average speed on the steel bike was 15.05 mph and the average speed on the carbon bike was 15.19 mph, suggesting that the carbon bike is faster. What's going on?!?

(a) Using Figure 2.96, what is the most obvious difference between commutes on the steel and the carbon bike?

(b) Use your answer to part (a) to explain why the carbon bike is slightly faster in terms of average speed, but yields a longer commute time, on average.

(c) In this study investigating whether the steel or the carbon bike yields a shorter commute time, a confounding variable is present. What is the confounding variable?

(d) What advice would you give to Dr. Groves to minimize his commute time?

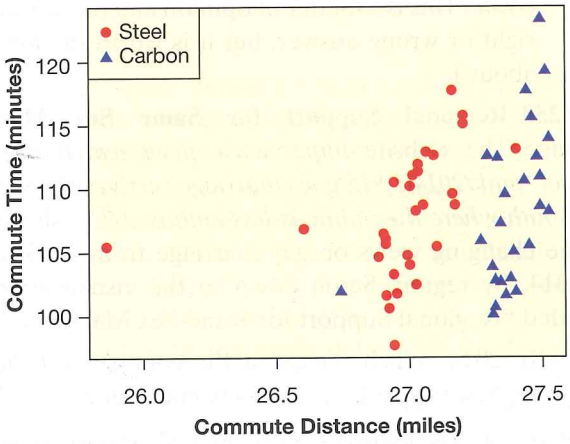


Figure 2.96 ***Commute distance and time, by type of bike***

2.252 Political Polarization Pew Research Center collected data on the same 10 political value questions from 1994 to 2014 and combines these responses to place each person on a scale ranging from consistently liberal to consistently conservative.[92] Visit *http://www.people-press.org/2014/06/12/section-1-growing-ideological-consistency/#interactive* to see a visualization of responses (in the form of a smoothed histogram), broken down by political party, changing over time. Click "Animate data from 1994–2014" to dynamically watch the distribution of responses changing over time.

[92] "Political Polarization in the American Public," *Pew Research Center*, June 12, 2014, *http://www.people-press.org/2014/06/12/section-1-growing-ideological-consistency/#interactive*

(a) Describe what happened to the median Democrat value in the years 1994–2014.

(b) Describe what happened to the median Republican value in the years 1994–2014.

(c) Is there more political polarization (less overlap between parties) in 1994 or 2014?

(d) In what year did the two medians start moving rapidly away from each other?

(e) By default, the general population results are shown. Click "POLITICALLY ACTIVE" just above the visualization to see results only for the third of the public who are most politically active. In 2014, are the politically active people more or less politically polarized than the general population?

(f) Read the first two paragraphs of the article. Do you think you learn more from reading the text or from looking at the data visualization? (*Note*: This is a matter of opinion and there is no right or wrong answer, but it is worth thinking about.)

2.253 Regional Support for Same Sex Marriage The website *http://www.pewresearch.org/fact-tank/2014/10/15/gay-marriage-arrives-in-the-south-where-the-public-is-less-enthused/*[93] shows the changing views on gay marriage from 2003 to 2014, by region. Scroll down to the visualization titled "Regional Support for Same-Sex Marriage."

(a) In 2014, which region of the country had the highest support for same-sex marriage?

(b) In 2014, which region of the country had the lowest support for same-sex marriage?

(c) Although regions have different starting levels of support, the *increase* in support for same-sex marriage is remarkably consistent across many of the different regions. Which region of the country displayed the smallest increase? The largest?

(d) The author (or the statistician / data scientist) decided to display these data with a separate time series plot for each region. Name two other ways these data could have been visualized.

2.254 Kidney Stone Treatment A study[94] collected data comparing treatments for kidney stones. Two of the treatments studied were open surgery and percutaneous nephrolithotomy. Treatment was deemed "successful" if, after three months, the kidney stones were either eliminated or less than 2 mm. The latter treatment (nephrolithotomy) is cheaper and less invasive, but is it as successful? Results are shown in Figure 2.97, first overall and then broken down by stone size. (If the answers are not obvious visually, in each case you can calculate the proportion of successes using the numbers shown on the graph.)

(a) When all stone sizes are considered, which treatment is more successful?

(b) When only small kidney stones are considered, which treatment is more successful?

(c) When only large kidney stones are considered, which treatment is more successful?

[93] Lipka, M., "Gay marriage arrives in the South, where the public is less enthused" *Pew Research Center*, October 15, 2014.

[94] Charig, R., Webb, D.R., Payne, S. (1986). "Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy." *British Medical Journal* (Clinical Residents Edition), 292(6524): 879–882.

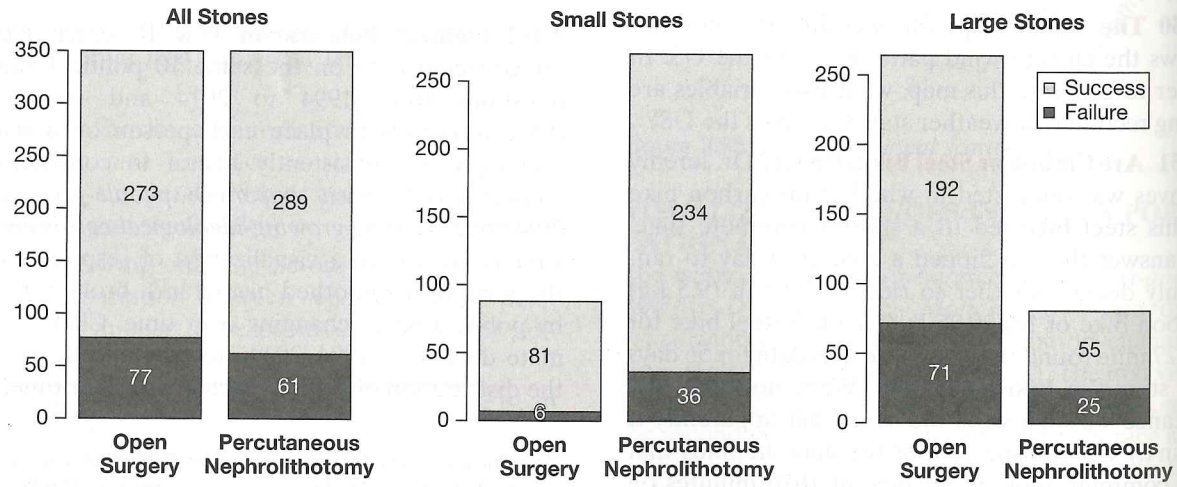


Figure 2.97 *Success rates, first for all stones, and then broken down by stone size*

(d) Which stone size results in higher success rates, regardless of treatment type?

(e) Which treatment is more commonly used for small stones?

(f) Which treatment is more commonly used for large stones?

(g) This is an example of Simpson's Paradox. Use your answers to parts (d) to (f) to explain how one treatment can be better for both small AND large stones, yet the other treatment appears to be better overall.

(h) Do you think this was a randomized experiment, with treatment randomly assigned? Why or why not?

2.255 Draw a Graph for Income vs College! The *New York Times* created an interesting interactive graph on income versus percent of children who attend college[95]. They ask you to first draw what you think the graph might look like, based on your intuition, and then compare your guess to the actual graph. The site also then shows a visualization of everyone's guesses. Go to *http://www.nytimes.com/interactive/2015/05/28/upshot/you-draw-it-how-family-income-affects-childrens-college-chances.html*, make your guess by drawing a line on the plot, and then click "I'm Done". How did you do? (The title above the plot showing your guess and reality tells you how you did—just reproduce the title phrase for your answer.)

2.256 What's Really Warming the World? Scientists at NASA collected data to study which forces, including both natural and human factors, are responsible for the increase in observed temperature in the last two centuries.[96] Go to *http://www.bloomberg.com/graphics/2015-whats-warming-the-world/* to see their resulting data visualization, an animated spaghetti plot (click the down arrow or scroll down with your mouse to see each new line appear). According to this data visualization, what is warming the world?

2.257 The Racial Divide The website *http://vallandingham.me/racial_divide/#pt* uses data from the US Census to visualize where whites and blacks live in different cities. Figure 2.98 gives a heat map of all the census tracks in St. Louis, with each track colored according to the racial composition (white to black).

[95] Aisch, G., Cox, A., and Quealy, K., "You Draw It: How Family Income Predicts Children's college Chances," *New York Times*, May 28, 2015.

[96] Roston, E. and Migliozzi, B., "What's Really Warming the World?," *Bloomberg*, June 24, 2015.

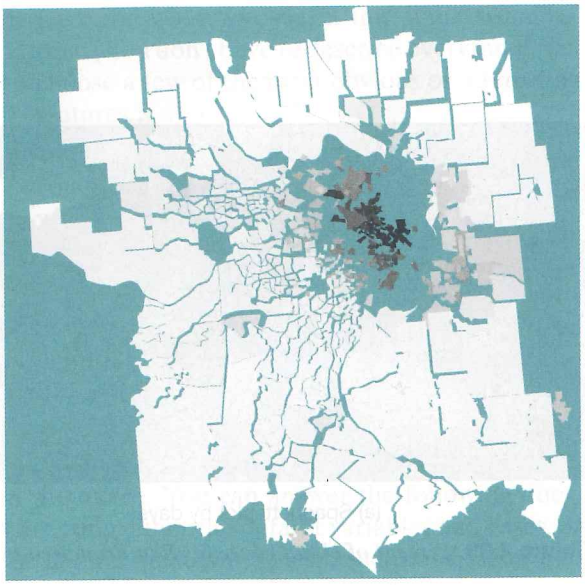


Figure 2.98 ***The census tracks of St. Louis, colored by racial composition***

Also, the space between tracks is shown proportional to the change in racial composition between neighboring tracks. Comment on what you see.

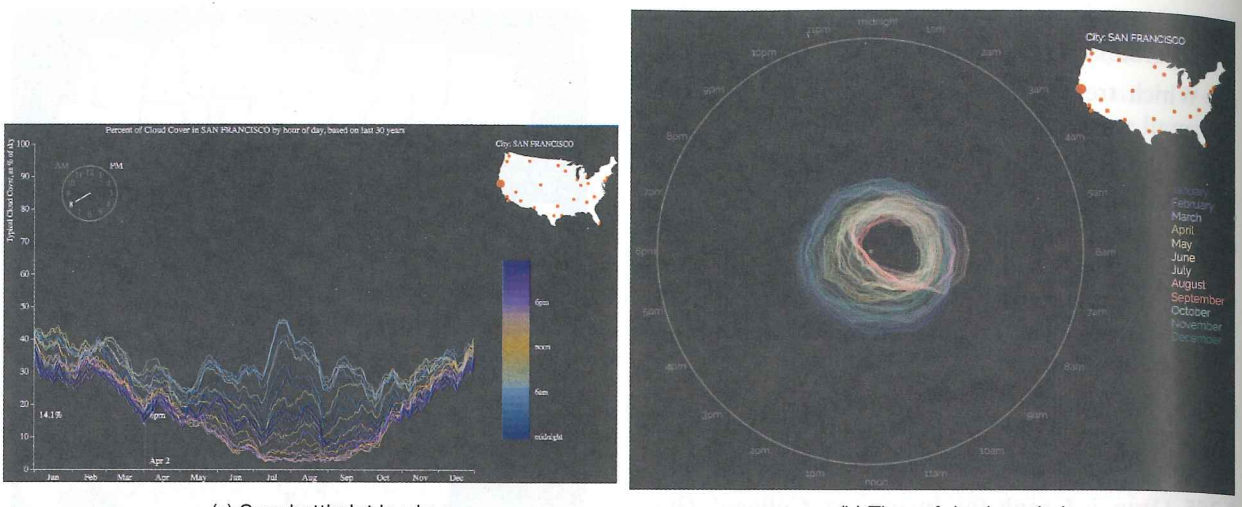
2.258 Cloud Cover in San Francisco Often, the same dataset can be visualized in many different ways. Figure 2.99 shows two different visualizations of San Francisco's typical cloud cover (as a percent of the sky) for each day of the year and time of day, based on the same data from the last 30 years.[97] Figure 2.99(a) shows a spaghetti plot with each hour of the day depicted with a separate strand, and Figure 2.99(b) displays the data with a different curve for every day of the year (the center of the circle is 0% cloud cover, the outer circle is 100%). Both visualizations are created using the same data, and both convey the same information; you may use whichever you find more intuitive to answer the following questions.

(a) Do mornings or evenings typically have more cloud cover, in general?

(b) Which season (winter, spring, summer, or fall) typically shows the most variability in cloud cover throughout the day?

(c) Which visualization do you find easier to interpret? (The answer may depend on the question of interest and there is no right or wrong answer.)

[97] Visualizations created by Zan Armstrong and used with permission. Check out more of her work at *zanarmstrong.com*.



(a) Spaghetti plot by days

(b) Time of day in a circle

Figure 2.99 ***Percent of cloud cover for San Francisco by time and day***

2.259 Cloud Cover in San Francisco–Online The plots in Exercise 2.258 on cloud cover in San Francisco can be found online at

weatherlines.zanarmstrong.com

if you prefer Figure 2.99(a) or

weather.zanarmstrong.com

if you prefer Figure 2.99(b). In the interactive display you can hover over points to get more information. You can also click on the map to change the city or the drop down menu to change the weather statistic that is plotted. Use the interactive plots at this website to answer the questions below.

(a) In San Francisco, approximately what time of day has the highest percent cloud cover in August?

(b) Which season tends to be the least windy for Chicago (the "Windy City")?

SCIENCE DATA STORIES VIDEOS

Data visualizations can also include video. In 2016, Science magazine hosted a Data Stories competition, in which participants upload a short video visualizing and telling a story with data. The winners can be viewed at *www.sciencemag.org/projects/data-stories/winners*. Exercises 2.260 to 2.263 pertain to the winners and finalists in this competition.

2.260 The Corporate Winner and People's Choice Winner was Daniel Gallagher from NASA's Scientific Visualization Studio, for his video "Martian Atmosphere Loss Explained by NASA." Watch this video and briefly describe the main message of the video.

2.261 The Professional Winner was RJ Andrews from Info We Trust, for the video "Are Gazelles Endangered?"

(a) Watch this video. What data are this video conveying?

(b) You can interact with the data and learn about other animals at this site: *http://www.infowetrust.com/endangeredsafari/*. Go to this site and hover over an animal on the interactive visualization. Indicate what animal you chose, whether its population is increasing or decreasing, and the endangered status of the animal. (These details appear at the bottom when you hover over an animal.)

2.262 The Student Winner was Ulf Aslak Jensen, for the video "How People Gather: An Interactive Visualization Approach." Watch this video, and answer the following questions:

(a) What data are this video displaying?

(b) You can explore the data shown in the video on your own at *https://ulfaslak.com/portfolio/Visualisation/*. Interact with the data, and report one of your findings.

2.263 The finalists can be viewed at *http://www.sciencemag.org/projects/data-stories/finalists*. Pick a video that interests you, watch it, and answer the following questions:

(a) Give a link to the chosen video.

(b) What data are being displayed in the video?

(c) What did you learn from the video?

GOOGLE PUBLIC DATA

The link *http://www.google.com/publicdata/directory* brings up some data visualizations created from public online data. Hovering your mouse over the little circles below the visualization brings you to a different data visualization (or you can just wait for the image to change). Many of the visualizations are dynamic, and by clicking them and then pressing the play button, you can watch them change over time. Also, when you click on the image, you can then edit it, including changing the variables and cases (often countries) displayed, or even the way the data are visualized. You can also hover over the points to see case labels and get more accurate information. Exercises 2.264 to 2.266 pertain to this website.

2.264 Find an example of an augmented scatterplot and click on the image. You can answer the following questions using either the default variables and cases, or else use the menu on the left to select variables and cases you are more interested in.

(a) Take a screenshot of the visualization (use the most recent year if multiple years are available) and include it.

(b) Which variable is displayed on the x-axis? The y-axis? The color of the points? The size of the points?

(c) Describe what you see in this static visualization. (You don't have to describe everything, just choose a few of the most obvious or interesting features.)

(d) Choose one point and hover over it to see which country (or case) it corresponds to. Give the values of each variable for this country (or case).

(e) If this is a dynamic graph, press play to watch the trend over time. Comment on what is happening over time.

(f) If this is a dynamic graph, choose one case, identify its name, and explain how this case changes over time.

2.265 Find an example of plot displaying geographic data and click on the image. You can answer the following questions using either the default variables and cases, or else use the menu on the left to select variables and cases you are more interested in.

(a) Take a screenshot of the visualization (use the most recent year if multiple years are available) and include it.

(b) Which variable is displayed by color? By point size (if the plot has points)?

(c) Describe what you see in this static visualization. (You don't have to describe everything, just choose a few of the most obvious or interesting features.)

(d) Choose one point and hover over it to see which country (or case) it corresponds to. Give the variable value(s) for this country (or case).

(e) If this is a dynamic graph, press play to watch the trend over time. Comment on what is happening over time.

(f) If this is a dynamic graph, choose one case, identify it's name, and explain how this case changes over time.

2.266 Find an example of a spaghetti plot and click on the image. You can answer the following questions using either the default variables and cases, or else use the menu on the left to select variables and cases you are more interested in.

(a) Take a screenshot of the visualization and include it.

(b) What is this plot displaying?

(c) Describe the overall trend in this visualization.

(d) Choose one case, identify it, and describe the trend for this particular case.

2.267 Find your own! Find your own data visualization online.[98]

(a) Include a screenshot of the visualization.

(b) What data are being displayed?

(c) Describe the story told by the visualization.

2.268 Monthly City Temperatures The data file **CityTemps** contains the average monthly temperature (in °C) for the cities of Moscow (Russia), Melbourne (Australia), and San Francisco (United States) in each of the years 2014 and 2015.

(a) Use time series plots and/or spaghetti plots to compare monthly temperatures between Moscow and San Francisco.

(b) Use time series plots and/or spaghetti plots to compare monthly temperatures between Melbourne and San Francisco.

2.269 Create Your Own: Augmented Scatterplot Using any of the datasets that come with this text that include at least two quantitative variables and at least one categorical variable (or any other

[98] If you need inspiration, check out *flowingdata.com, driven-by-data.net/*, or search for "New York Times Data Visualization" in Google images.

dataset that you find interesting and that meets these conditions), use statistical software to create an augmented scatterplot that identifies the dots by the category that they are in. Indicate the dataset, the cases, and the variables that you use. Comment (in context) about any interesting features revealed in your plot.

2.270 Create Your Own: Bubble Plot Using any of the datasets that come with this text that include at least three quantitative variables (or any other dataset that you find interesting and that meets this condition), use statistical software to create a bubble plot of the data. Indicate the dataset, the cases, and the variables that you use. Specify which variable represents the size of the bubble. Comment (in context) about any interesting features revealed in your plot.

2.271 Create Your Own: Be Creative!! Create your own data visualization, and describe it. Be creative!!