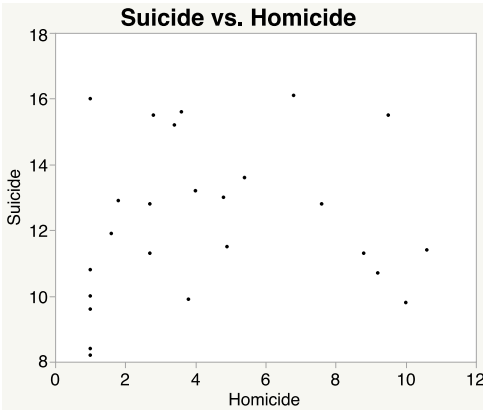
Chapter 4 – Scatterplots and Correlation

4.1 (a) Explanatory: number of lectures attended; response: grade on final exam. **(b)** Explanatory: time exercising; response: calories burned. **(c)** Explanatory: time spent online using Facebook is explanatory; response: GPA (assuming that more time on Facebook means less time studying). **(d)** Explore the relationship.

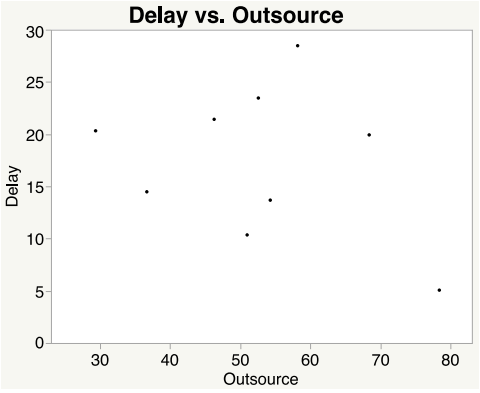
4.2 Sea surface temperature is the explanatory variable; coral growth is the response variable. Both variables are quantitative.

4.3 Answers will vary. Examples: weight, sex, blood pressure, country of origin, etc.

4.4 A scatterplot is provided. For convenience, homicide rate (explanatory variable) is on the horizontal axis and suicide rate (response variable) is on the vertical axis.



4.5 A scatterplot is provided. Outsource percent is the explanatory variable and should be on the horizontal axis. Delay percent is the response and should be on the vertical axis. These data do not support concerns of the critics.

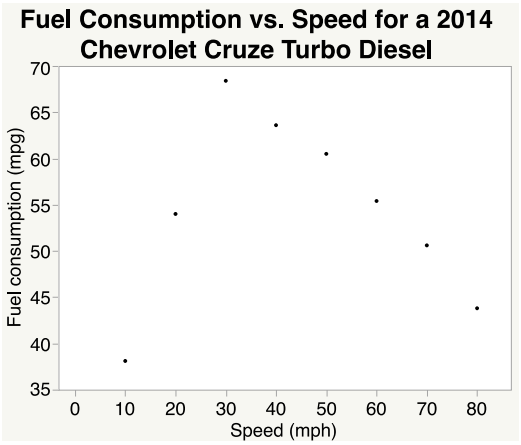


4.6 Answers may vary. Some may see the scatterplot to show a curved

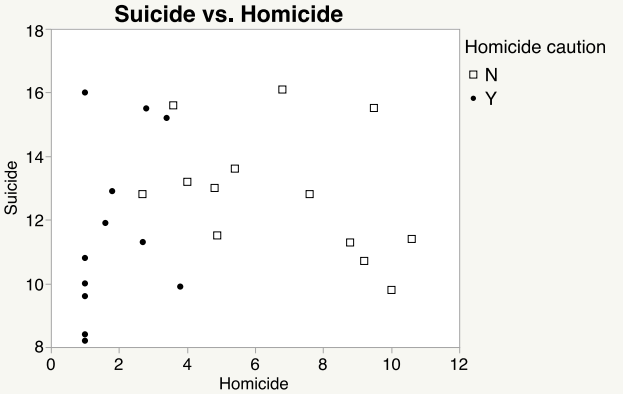
relationship—high in the middle, lower at the extremes. This relationship has moderate strength, and there are possible deviations to this relationship for counties with high suicide rates and homicide rates near the extremes (for example, Clermont and Montgomery counties). Others may see no real relation and describe the plot as looking like a random scatter of points. Thus, there is no direction and no form in the relationship between homicide and suicide rates, a weak relationship between the variables, and no deviations from the overall pattern.

4.7 One could consider Hawaiian to be an outlier, with a very high outsourcing percent and a very low delay percent. Without Hawaiian, there is a slightly positive relationship; with Hawaiian the relationship is slightly negative.

4.8 (a) A scatterplot is provided. Speed is explanatory. **(b)** The relationship is curved—high in the middle, lower at the extremes. This makes sense because it takes less fuel to travel the same distance at moderate speeds than at slow or fast speeds. **(c)** Below-average (that is, bad) values of "fuel consumption" are found with both low and high values of "speed." **(d)** The relationship is very strong—there is little scatter around the curve, so the curve is very useful for prediction.

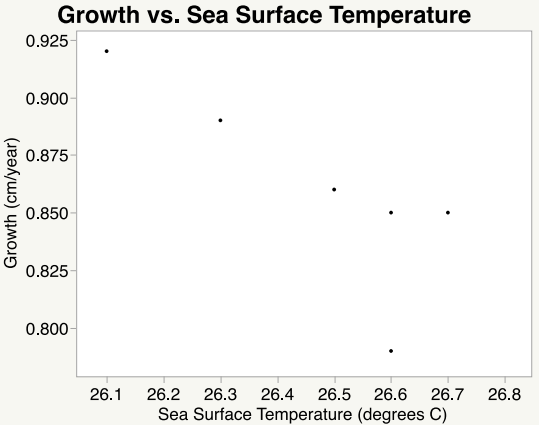


4.9 (a) A scatterplot is provided. Caution counties are marked with closed circles, the others with open squares.



(b) For both types of counties, there appears to be no relationship between homicide and suicide rates. The caution counties form a band on the left portion of the graph (low homicide rates), while the non-caution counties form a band on the right portion (high homicide rates).

4.10 (a) A scatterplot is provided. Temperature is the explanatory variable.

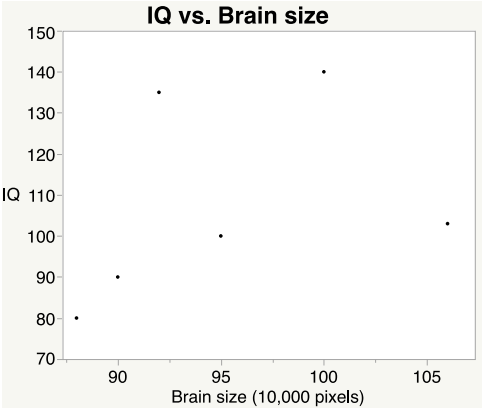


(b) $\bar{x}$ = 26.47 degrees, s_x= 0.23 degrees, $\bar{y}$ = 0.86 cm, and s_y = 0.04 cm. See the table provided for the standardized scores. The correlation is r = –4.37/(6 – 1) = – 0.874. This is consistent with the strong, negative association depicted in the scatterplot.

z_x	z_y	$z_x z_y$
1.00	–0.25	–0.25
0.57	–0.25	–0.14
0.57	–1.75	–1.00
0.13	0	0
–0.74	0.75	–0.56
–1.61	1.50	–2.42
	sum	–4.37

(c) Software will give a value of –0.811. The more precision you used in each step, the closer you'll get to that value. The answer in part (c) is erroneous at the hundredths place due to rounding.

4.11 (a) A scatterplot is provided. Brain size is the explanatory variable.



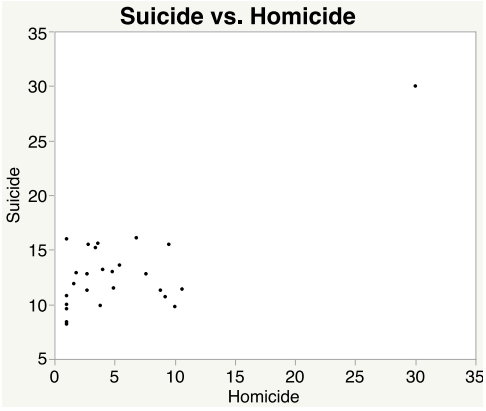
(b) $\bar{x}$ = 95.17 (10,000 pixels), s_x = 6.77 (10,000 pixels), $\bar{y}$ = 108 points, and s_y = 24.29 points. See the table provided for the standardized scores. The correlation is r = 1.87/(6 – 1) = 0.374. This is consistent with the weak, positive association depicted in the scatterplot.

z_x	z_y	$z_x z_y$
0.71	1.32	0.94
–0.76	–0.74	0.56
–0.03	–0.33	0.01
–0.47	1.11	–0.52
–1.06	–1.15	1.22
1.60	–0.21	–0.34
	sum	1.87

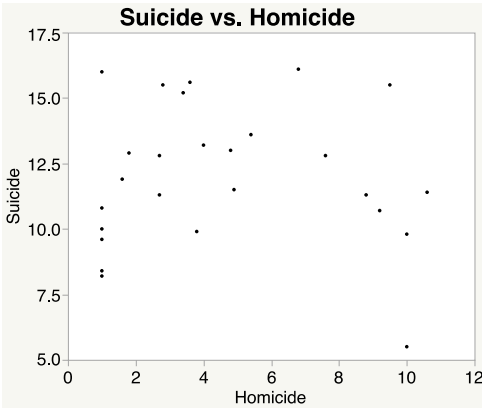
(c) Software will give a value of 0.377. The more precision you used in each step, the closer you'll get to that value. The answer in part (c) is erroneous at the thousandths place due to rounding.

4.12 r would not change; units do not affect correlation.

4.13 (a) $r = 0.170$. **(b)** A scatterplot is provided. With Point A included, the correlation increases to 0.746.

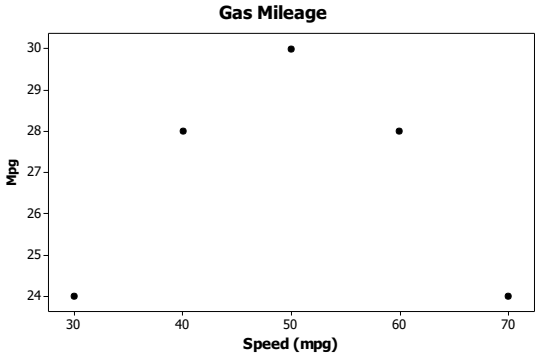


(c) A scatterplot is provided. With Point B, the correlation drops to −0.017.



(d) Point A strengthens the positive linear association because, when A is included, the points of the scatterplot seem to actually have a linear relationship (your eye is drawn to that point at the upper right). Meanwhile, Point B (at the lower right of the graph) deviates from the pattern, weakening the association.

4.14 A scatterplot is provided. In computing the correlation, note that $\bar{x}$ = 50 mph, s_x = 15.8114 mph, $\bar{y}$ = 26.8 mpg, and s_y = 2.6833 mpg. Refer to the table of standardized scores provided, then note that $r = 0/4 = 0$. The correlation is zero because these variables do not have a straight-line relationship; the association is neither positive nor negative. Remember that correlation only measures the strength and direction of a *linear* relationship between two variables.



Z_x	Z_y	Z_xZ_y
−1.2649	−1.0435	1.3199
−0.6325	0.4472	−0.2829
0	1.1926	0
0.6325	0.4472	0.2829
1.2649	−1.0435	−1.3199
	sum	0

4.15 (b) is the engine size. The engine size will be used to predict gas mileage.

4.16 (c) a negative association. The association should be negative since cars with bigger engines tend to have lower gas mileages.

4.17 (a) price of beer (per ounce) = \$0.44, price of a hot dog = \$1.25.

4.18 (c) 0.1.

4.19 (c) $-1 \le r \le 1$. Correlations range from –1 to 1 inclusive.

4.20 (c) no straight-line pattern, but there might be a strong pattern of another form. A correlation close to 0 might arise from a scatterplot with no visible pattern, but there could be a nonlinear pattern. See Exercise 4.14, for example.

4.21 (c) either –1 or 1, we can't say which. Because we are not told how the *x*-values and *y*-values vary together, we cannot tell whether the correlation will be −1 or +1.

4.22 (a) 1. There would be a perfect, positive linear association. The line would be Exam2 = Exam1 − 10.

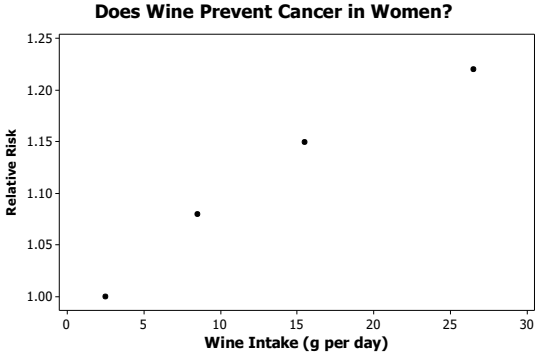
4.23 (b) 0.3. Correlation is unaffected by units.

4.24 (b) $r = 0.3$. Computation with a calculator or software gives $r = 0.298$.

4.25 (a) The lowest first-round score was 66, scored by one golfer. This golfer scored 74 in the second round. **(b)** Clarke scored 84 in the second round and 76 in the first round. **(c)** The correlation is small but positive, so closest to $r = 0.25$. Knowing a golfer's first-round score would not be very useful in predicting a second-round score.

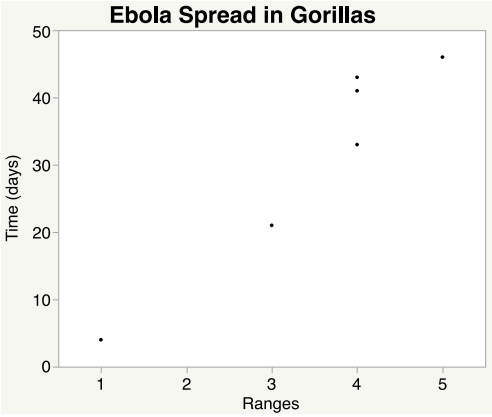
4.26 (a) Overall, there is a slightly negative association between these variables. **(b)** There is general disagreement—low BRFSS scores correspond to greater happiness, and these are associated with higher-ranked states (the least happy states, according to the objective measure). **(c)** It is hard to declare any of the data values as "outliers." It does not appear that any of the values are obviously outside of the general pattern. Perhaps one value (Rank = 8, BRFSS = 0.30) is an outlier, but this is hard to say.

4.27 (a) A scatterplot is provided. It reveals a very strong, positive linear relationship between wine intake and relative risk for cancer. We expect correlation to be close to +1.



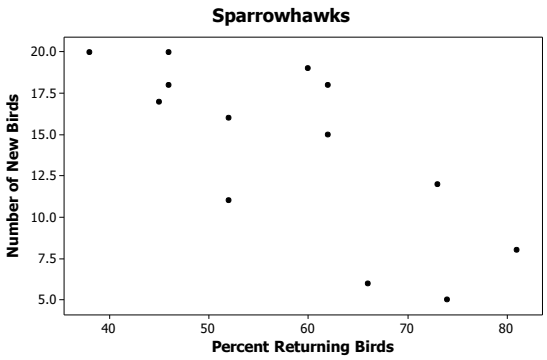
(b) Using software, $r = 0.9851$. The data suggest that women who consume more wine tend to have higher risk of breast cancer. However, this is an observational study, and no causal relationship can be determined. The women who drink more wine may differ in many respects from the women who drink less wine.

4.28 (a) A scatterplot is provided. It suggests a strong positive linear association between distance and time with respect to the spread of Ebola.



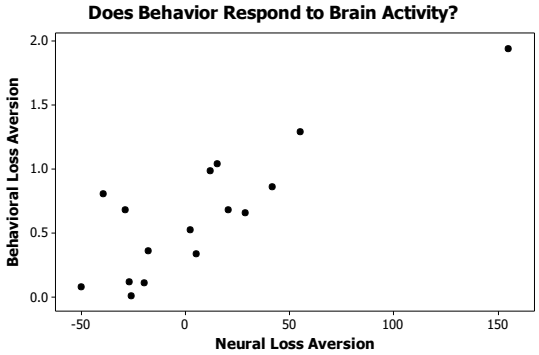
(b) $r = 0.9623$. This is consistent with the pattern described in part (a). **(c)** The correlation would not change, since it does not depend on units.

4.29 (a) A scatterplot is provided. It shows a negative, somewhat linear relationship. Because the relationship is linear, correlation is an appropriate measure of strength: $r = -0.7485$.



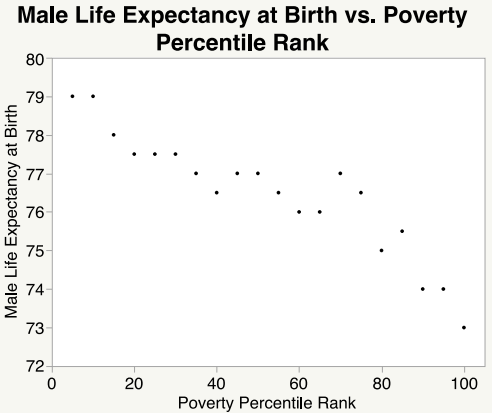
(b) Because this association is negative, we conclude that the sparrowhawk is a long-lived territorial species.

4.30 (a) The scatterplot is shown; note that neural activity is explanatory (and so should be on the horizontal axis).



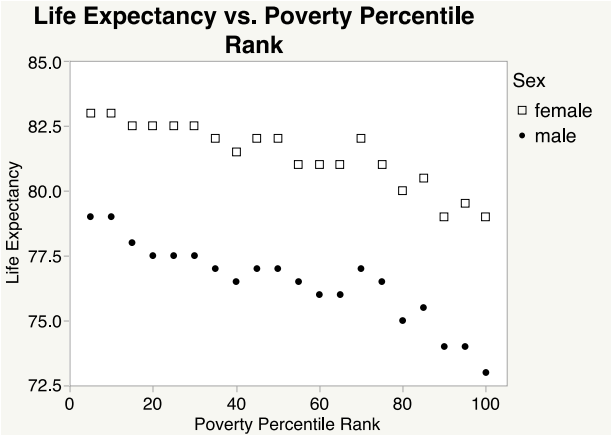
(b) The association is moderately strong, positive, and linear. The outlier is in the upper right corner (behavioral score is 1.94). **(c)** For all points, $r = 0.8486$. Without the outlier, $r = 0.7015$. The correlation is greater with the outlier because it fits the pattern of the other points; if one drew the line suggested by the other points, the outlier would extend the length of the line and would therefore decrease the relative scatter of the points about that line.

4.31 (a) The scatterplot is shown; note that poverty percentile rank is explanatory (and so should be on the horizontal axis).



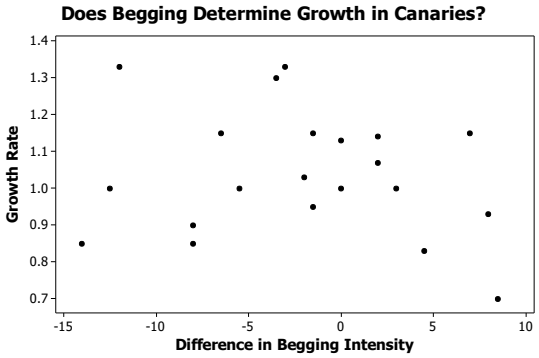
(b) The association is strong, negative, and linear; $r = -0.924$.

4.32 (a) The scatterplot is shown; note that poverty percentile rank is explanatory (and so should be on the horizontal axis). Males are marked with solid circles; females are marked with open squares.



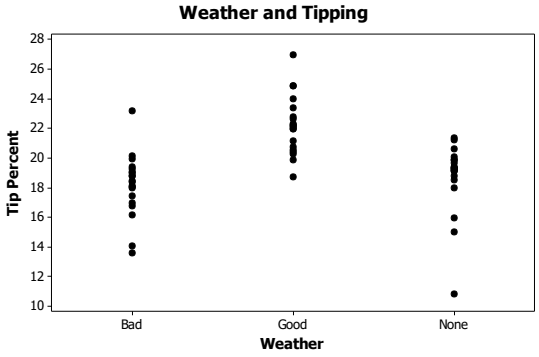
(b) Both males and females have a strong, negative, and linear pattern of life expectancy and poverty percentile rank. For each level of poverty percentile rank, females have a higher life expectancy than males. Additionally, the association between poverty percentile rank and life expectancy is less negative for females than for males.

4.33 (a) The scatterplot is provided.



(b) The scatterplot suggests that there is not a linear relationship between relative growth rate and difference in begging intensity (we can see a bit of a curved relationship). Here, $r = -0.1749$. r is not helpful here because the relationship is not linear. **(c)** Neither theory is strongly supported, but the latter is more strongly supported. That is, growth rate increases initially as begging intensity increases but then levels off or decreases as parents begin to ignore increases in begging by the foster babies.

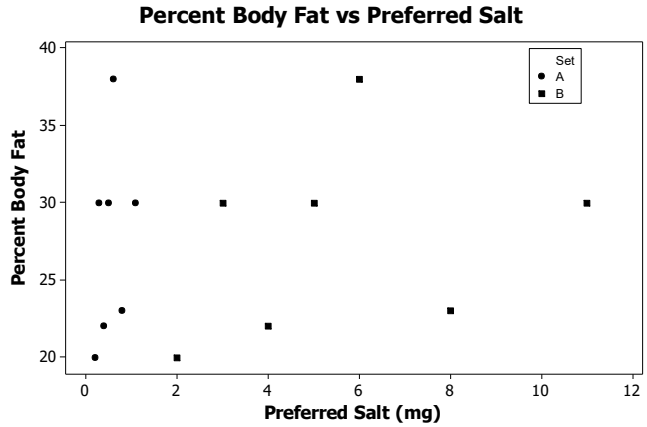
4.34 (a) The plot provided suggests that "Good" weather reports tend to yield higher tips.



(b) The explanatory variable is categorical, not quantitative, so *r* cannot be used. Notice that we can arrange the categories any way, and these different arrangements would suggest different associations. Hence, it doesn't make sense to discuss a relationship direction here.

4.35 (a) The correlation would not change, as correlation does not depend on units. **(b)** The correlation would not change. By subtracting 0.25 from all risks, each point in the scatterplot moves "down" by 0.25, but the strength and direction of the linear relationship between risk and wine intake does not change. **(c)** There would be a perfect, positive linear relationship, with $r = +1$.

4.36 (a) The scatterplot is provided. Set B (the mad scientist's set) has stretched out the *x*-values, but the pattern is still the same.



(b) Units do not impact correlation. For both data sets, $r = 0.298$.

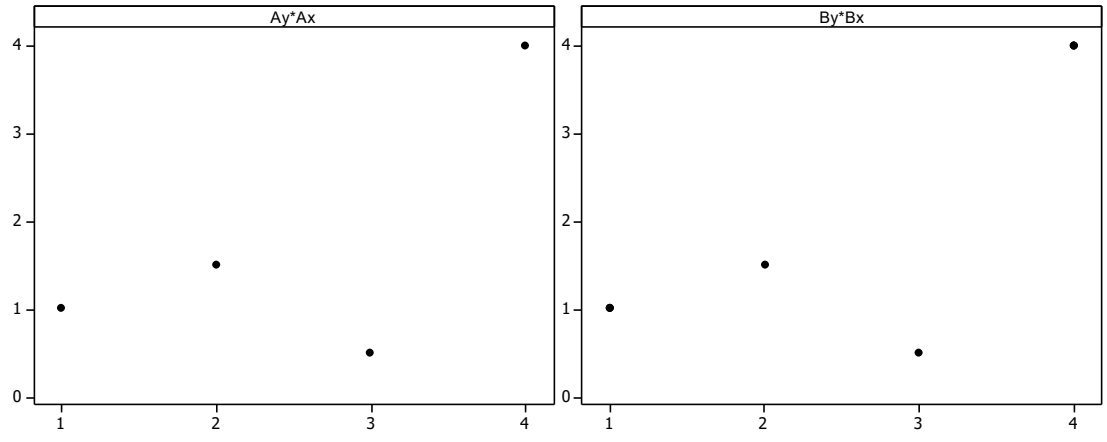
4.37 Explanations and sketches will vary, but should note that correlation measures the strength of the association, not the slope of the line. The hypothetical Fund A and Fund B mentioned in the report, for example, might have a linear relationship having line of slope 2 or ½.

4.38 (a) Small-cap stocks have a lower correlation with municipal bonds, so the relationship is weaker. **(b)** She should look for a negative correlation (although this would also mean that this investment tends to *decrease* when bond prices rise).

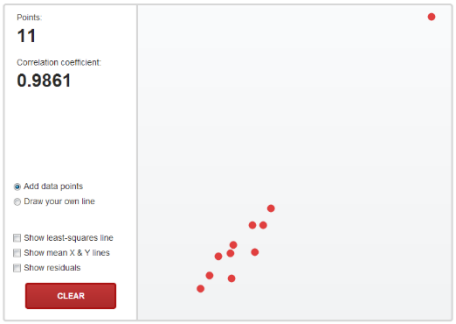
4.39 The person who wrote the article interpreted a correlation close to 0 as if it were a correlation close to –1 (implying a negative association between teaching ability and research productivity). Professor McDaniel's findings mean there is little linear association between research and teaching ability. For example, knowing that a professor is a productive researcher gives little information about whether the professor is a good or bad teacher. Also, remember that correlation is only meaningful if both variables are quantitative—and here, there is no guarantee that this is the case.

4.40 (a) Because sex has a nominal scale, we cannot compute the correlation between sex and any other variable. There is a strong *association* between sex and income. Some writers and speakers use "correlation" as a synonym for "association," but this is not correct. **(b)** A correlation of $r = 1.09$ is impossible, because r is restricted to be between –1 and 1. **(c)** Correlation has no units, so "$r = 0.63$ centimeter per kilogram" is incorrect.

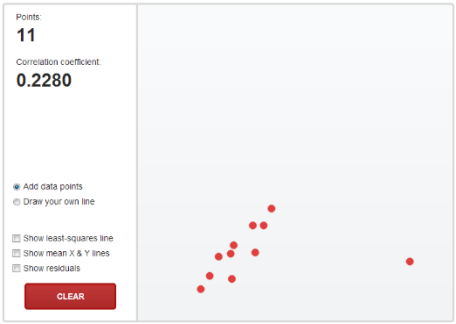
4.41 (a) Scatterplots are given. To the naked eye, the two plots look identical. **(b)** For data set A, $r = 0.664$. For data set B, $r = 0.834$. The increase in r is due to more points at (1, 1) and (4, 4). Because these are not visible (several points are represented by one in the scatterplot), you would not expect the difference in r if you simply looked at the plots.



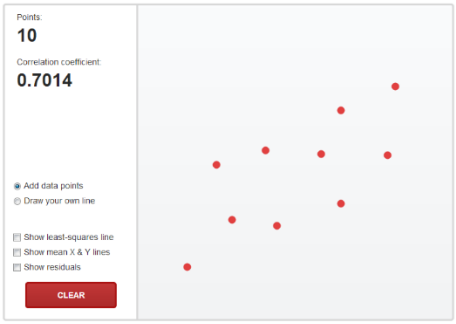
4.42 (a) The correlation will be closer to 1. One possible answer is shown (see first scatterplot given).



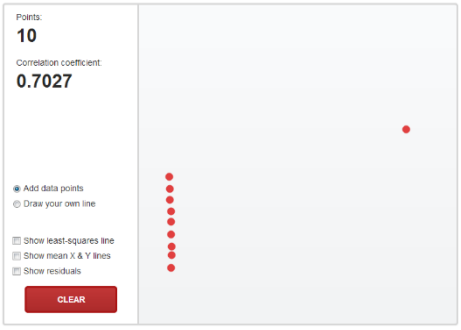
(b) Answers will vary, but the correlation will decrease, and can be made negative by dragging the point down far enough (see second scatterplot given).



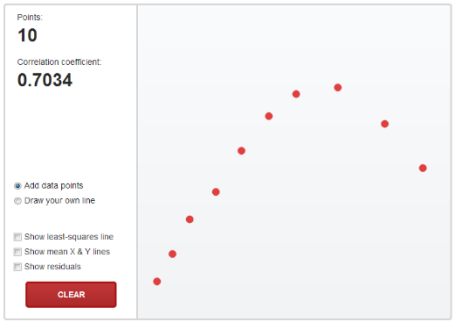
4.43 (a) Because two points determine a line, the correlation is always 1.
(b) Sketches will vary; an example is shown. Note that the scatterplot must be positively sloped, but r is affected only by the scatter about the line, not by the steepness of the slope of that line.



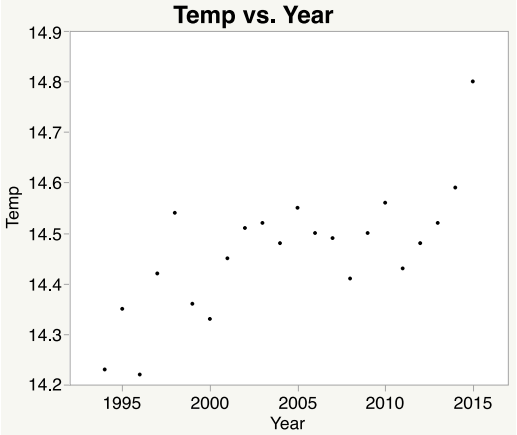
(c) The first nine points cannot be spread from the top to the bottom of the graph because, in such a case, the correlation cannot exceed about 0.66 (this is based on experience—lots of playing around with the applet). One possibility is shown.



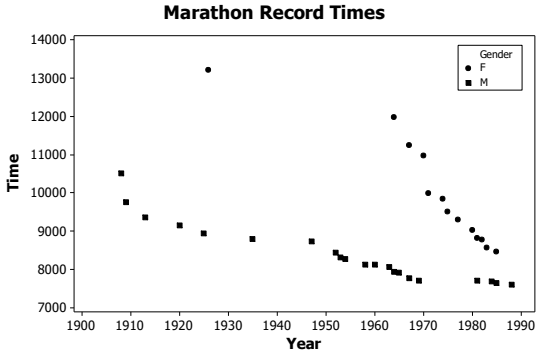
(d) To have $r = 0.7$, the curve must be higher at the right than at the left. One possibility is shown.



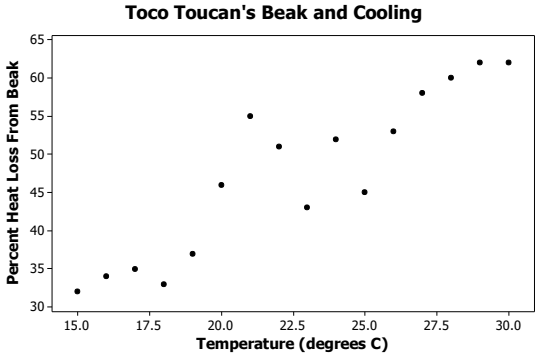
4.44 PLAN: To investigate global warming, we'll create a scatterplot and look for an increasing (positive) pattern. SOLVE: The plot suggests that temperatures have been increasing overall, but there seems to have been a slowing between 2000 and 2010; the relationship appears to be curved (cubic) in nature. Because of the curved nature of the relationship, the correlation between year and temperature (r = 0.7203) may not be useful here. CONCLUDE: Over time, average global temperatures have increased, but the increase may not be linear.



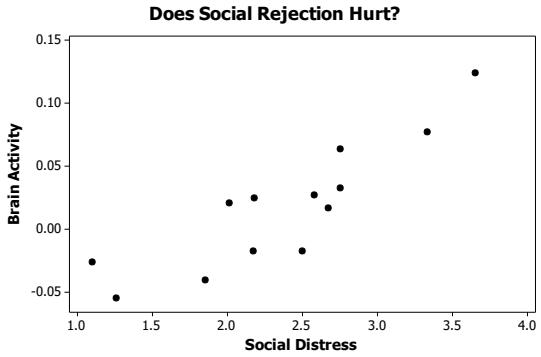
4.45 (a and b) PLAN: To study the improvements in running times between men and women, we'll plot the data on the same scatterplot. We will not use correlation, but we will examine the plot to see if women are beginning to outrun men. SOLVE: The plot is provided. By inspection, one might guess that the "lines" that fit these data sets will meet around 1998. This is how the researchers made this leap. CONCLUDE: Men's and women's times have, indeed, grown closer over time. Both sexes have improved their record marathon times over the years, but women's times have improved at a faster rate. In fact, as of February 2017, the world record time for men had continued to be faster than the world record time for women. The difference is currently about 748 seconds (just over 12 minutes), where in the data plotted, the difference was about 856 seconds.



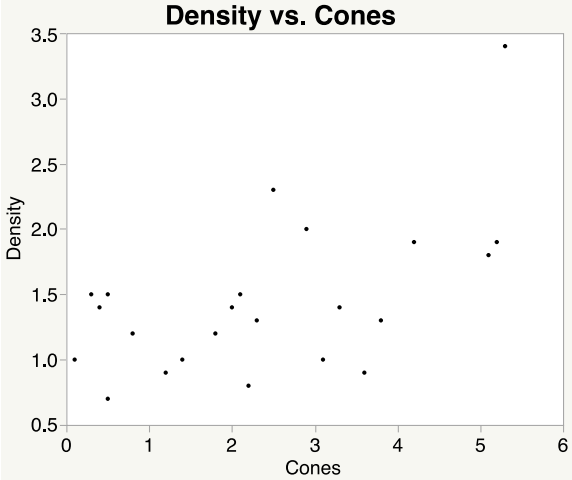
4.46 PLAN: To investigate the relationship between the outside temperature and the percent of total heat loss due to the beak, we plot the percentage heat loss from the beak against the outside temperature. We'll compute the correlation, if the relationship looks to be reasonably linear. SOLVE: The plot is given. Notice that there is a reasonably strong linear relationship. It seems reasonable to use correlation to describe this relationship's strength and direction. In fact, $r = 0.9143$. CONCLUDE: When the outside temperature increases, a greater percentage of total heat loss is due to beak heat loss. That is, the beak plays a more important role in cooling down the toco toucan as the weather outside becomes hotter.



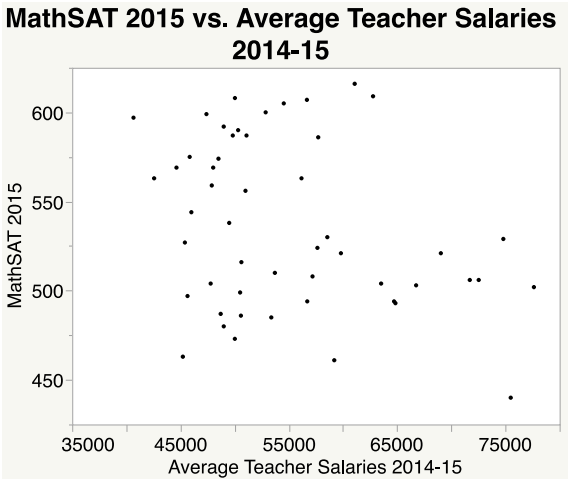
4.47 PLAN: We wish to explore the relationship between social distress and brain activity. We begin with a scatterplot and compute the correlation, if appropriate. SOLVE: A scatterplot shows a fairly strong, positive, linear association. There are no particular outliers; each variable has low and high values, but those points do not deviate from the pattern of the rest. The relationship seems to be reasonably linear, so we compute $r = 0.8782$. CONCLUDE: Social exclusion does appear to trigger a pain response: higher social distress measurements are associated with increased activity in the pain-sensing area of the brain. However, no cause-and-effect conclusion is possible since this was not a designed experiment.



4.48 PLAN: We wish to explore the relationship between cone supplies and squirrel density the following spring. We begin with a scatterplot and compute the correlation, if appropriate. SOLVE: A scatterplot shows a moderately strong, positive, linear association. The point at the upper right (5.3, 3.4) may be an outlier. This point seems to make the linear relationship appear more positive. Including the possible outlier, $r = 0.564$. If the outlier is omitted, $r = 0.4406$. CONCLUDE: The positive association supports the idea that squirrel populations increase when the cone supply is higher in the previous autumn. However, the relationship is somewhat weak; squirrels in the Yukon may have other good food sources.



4.49 PLAN: Because we want to know if higher teacher salaries lead to higher Mathematics SAT scores, salary is the explanatory variable and SAT score is the response. We'll create a scatterplot and compute the correlation between the two variables, if appropriate. SOLVE: The scatterplot is given. If there is a linear relationship, it is very weak. The plot almost looks like a "C." Most notable is that states with the highest average teacher salaries seem to have low Mathematics SAT scores. $r = -0.308$. CONCLUDE: The relationship between teacher salaries and Mathematics SAT scores (based on averages by state) is weakly linear and decreasing; these data do *not* support the idea that higher teacher salaries lead to greater student accomplishment (as measured by Mathematics SAT scores).



4.50 and **4.51** are Web-based exercises.