Chapter 6 – Scatterplots, Association, and Correlation

Section 6.1

1. Association.

a) Either weight in grams or weight in ounces could be the explanatory or response variable. Greater weights in grams correspond with greater weights in ounces. The association between weight of apples in grams and weight of apples in ounces would be positive, straight, and perfect. Each apple's weight would simply be measured in two different scales. The points would line up perfectly.

b) Circumference is the explanatory variable, and weight is the response variable, since one-dimensional circumference explains three-dimensional volume (and therefore weight). For apples of roughly the same size, the association would be positive, straight, and moderately strong. If the sample of apples contained very small and very large apples, the association's true curved form would become apparent.

c) Shoe size is the explanatory variable, and GPA is the response variable, since it is more likely that one would want to predict GPA given a student's shoe size. There would be no association between shoe size and GPA of college freshmen.

d) Number of miles driven is the explanatory variable, and gallons remaining in the tank is the response variable. The greater the number of miles driven, the less gasoline there is in the tank. If a sample of different cars is used, the association is negative, straight, and moderate. If the data is gathered on different trips with the same car, the association would be strong.

2. Association II.

a) Price for each T-Shirt is the explanatory variable, and number of T-Shirts sold is the response variable. The association would be negative, straight (until the price became too high to sell *any* shirts), and moderate. A very low price would likely lead to very high sales, and a very high price would lead to low sales.

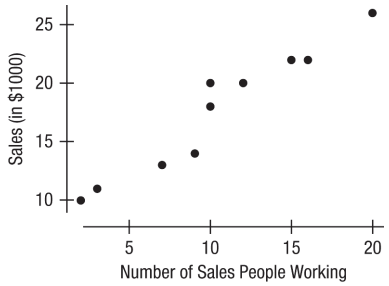
b) Depth of the water is the explanatory variable, and water pressure is the response variable. The deeper you dive, the greater the water pressure. The association is positive, straight, and strong. For every 33 feet of depth, the pressure increases by one atmosphere (14.7 psi).

c) Depth of the water is the explanatory variable, and visibility is the response variable. The deeper you dive, the lower the visibility. The association is negative, possibly straight, and moderate if a sample of different bodies of water is used. If the same body of water has visibility measured at different depths, the association would be strong.

d) At first, it appears that there should be no association between weight of elementary school students and score on a reading test. However, with weight as the explanatory variable and score as the response variable, the association is positive, straight, and moderate. Students who weigh more are likely to do better on reading tests because of the lurking variable of age. Certainly, older students generally weigh more and generally are better readers. Therefore, students who weigh more are likely to be better readers. This does not mean that weight causes higher reading scores.

3. Bookstore sales.

a) The scatterplot is shown below.



b) There is a positive association between bookstore sales and the number of sales people working.

3. (continued)

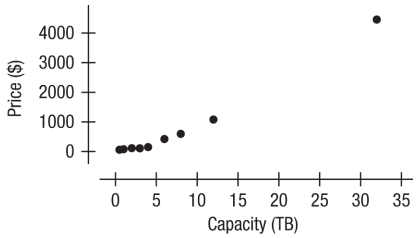
c) There is a linear association between bookstore sales and the number of sales people working.

d) There is a strong association between bookstore sales and the number of sales people working.

e) The relationship between bookstore sales and the number of sales people working has no outliers.

4. Disk drives 2016.

a) The scatterplot is shown below.



b) There is a positive association between price and capacity of disk drives.

c) There is a curved association between price and capacity of disk drives.

d) There is a strong, but not linear, association between price and capacity of disk drives.

e) There are no outliers. The 32 TB drive is far from the other sizes, but its price seems reasonable given the rest of the pattern.

Section 6.2

5. Correlation facts.

a) True.

b) False. The correlation will remain the same.

c) False. Correlation has no units.

6. Correlation facts II.

a) False. This is a very weak association.

b) False. Standardizing does not change the correlation.

c) True.

Section 6.3

7. Bookstore sales again.

This conclusion is not justified. Correlation does not demonstrate causation. The analyst argues that the number of sales staff working causes sales to be higher. It is possible (perhaps more plausible) that the store hired more people as sales increased. The causation may run in the opposite direction of the analyst's argument.

8. Blizzards.

The director's conclusion is not justified. The lurking variable is the severity of the blizzard. Particularly severe blizzards require more snowplows, and they also prevent people from leaving home, where they are more likely to make online purchases, especially since they have to leave home to go to a store.

Section 6.4

9. Salaries and logs.

Since $\log_{10} 10,000 = 4$, $\log_{10} 100,000 = 5$, and $\log_{10} 1,000,000 = 6$, the plotted points will be (1, 4), (15, 5), and (30, 6). The plot of these three points will lie very close to a straight line.

10. Dexterity scores.

The reciprocal re-expression is straighter. The points to plot for the re-expression are (4, 0.4), (9, 0.5), and (12, 0.67).

Chapter Exercises.

11. Association III.

a) Altitude is the explanatory variable, and temperature is the response variable. As you climb higher, the temperature drops. The association is negative, possibly straight, and weak to moderate.

b) At first, it appears that there should be no association between ice cream sales and air conditioner sales. When the lurking variable of temperature is considered, the association becomes more apparent. When the temperature is high, ice cream sales tend to increase. Also, when the temperature is high, air conditioner sales tend to increase. Therefore, there is likely to be an increase in the sales of air conditioners whenever there is an increase in the sales of ice cream. The association is positive, straight, and moderate. Either one of the variables could be used as the explanatory variable.

c) Age is the explanatory variable, and grip strength is the response variable. The association is neither negative nor positive, but is curved, and moderate in strength, due to the variability in grip strength among people in general. The very young would have low grip strength, and grip strength would increase as age increased. After reaching a maximum (at whatever age physical prowess peaks), grip strength would decline again, with the elderly having low grip strengths.

d) Reaction time is the explanatory variable, and blood alcohol content is the response variable. As blood alcohol level increase, so does the time it takes to react to a stimulus. The association is positive, probably curved, and strong. The scatterplot would probably be almost linear for low concentrations of alcohol in the blood, and then begin to rise dramatically, with longer and longer reaction times for each incremental increase in blood alcohol content.

12. Association IV.

a) Time spent on the legal consultation is the explanatory variable, and cost of the consultation is the response variable. The longer you spend talking, the more the consultation costs. The association is positive, straight, and strong, since it is likely you will be charged an hourly rate.

b) Time delay of the thunder is the explanatory variable, and distance from lightning is the response variable. The longer it takes the thunder to reach your ears, the farther away you are from the strike. The association is positive, straight, and fairly strong, since the speed of sound is not a constant. Sound travels at a rate of around 770 miles per hour, depending on the temperature.

c) Brightness is the explanatory variable, and distance from the streetlight is the response variable. The less bright the streetlight appears, the further away from the light you are. The association is negative, curved, and strong. Distance and light intensity follow an inverse square relationship. Doubling the distance to the light source reduces the intensity by a factor of four.

d) Weight of the car is the explanatory variable, and age of the owner is the response variable. There is likely very little association between the weight of the car and the age of the owner. However, some might say that older drivers tend to drive larger cars. If that is the case, there may be a positive, straight, and very weak association between weight of a car and the age of its owner.

13. Scatterplots.

a) None of the scatterplots show little or no association, although # 4 is very weak.

b) #3 and #4 show negative association. Increases in one variable are generally related to decreases in the other variable.

c) #2, #3, and #4 all show a linear association.

d) #2 shows a moderately strong association.

e) #1 and #3 each show a very strong association. #1 shows a curved association and #3 shows a straight association.

14. Scatterplots II.

a) #1 shows little or no association.

b) #4 shows a negative association.

c) #2 and #4 each show a linear association.

d) #3 shows a moderately strong, curved association.

e) #2 and #4 each show a very strong association, although some might classify the association as merely "strong".

15. Performance IQ scores vs. brain size.

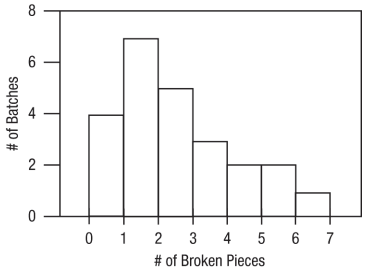
The scatterplot of IQ scores vs. Brain Sizes is scattered, with no apparent pattern. There appears to be little or no association between the IQ scores and brain sizes displayed in this scatterplot.

16. Kentucky derby 2017.

Winning speeds in the Kentucky Derby have generally increased over time. The association between year and speed is positive, moderately strong, and nonlinear, with a generally increasing winning speeds until 1973. Speeds level off after 1973.

17. Firing pottery.

a) A histogram of the number of broken pieces is shown below.

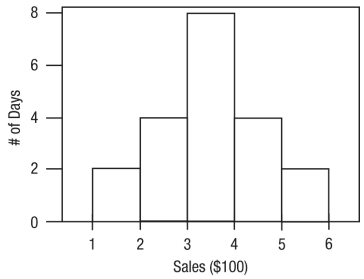


b) The distribution of the number broken pieces per batch of pottery is skewed right, centered around 1 broken piece per batch. Batches had from 0 and 6 broken pieces. The scatterplot does not show the center or skewness of the distribution.

c) The scatterplot shows that the number of broken pieces has a positive and somewhat linear relationship. The number of broken pieces increases as the batch number increases. If the 8 daily batches are numbered sequentially, this indicates that batches fired later in the day generally have more broken pieces. This information is not visible in the histogram.

18. Coffee sales.

a) A histogram of daily sales is shown below.



b) The scatterplot shows a positive linear relationship. In general, the sales have been increasing over time. The histogram does not show this.

c) The histogram shows that the mean of the daily sales for the coffee shop was around $350 (between $300 and $400), and that this happened on 8 days. The scatterplot does not show this.

19. Matching.

a) 0.006 **b)** 0.777 **c)** –0.923 **d)** –0.487

20. Matching II.

a) –0.977 **b)** 0.736 **c)** 0.951 **d)** –0.021

21. Politics.

The candidate might mean that there is an *association* between television watching and crime. The term correlation is reserved for describing linear associations between quantitative variables. We don't know what type of variables "television watching" and "crime" are, but they seem categorical. Even if the variables are quantitative (hours of TV watched per week, and number of crimes committed, for example), we aren't sure that the relationship is linear. The politician also seems to be implying a cause-and-effect relationship between television watching and crime. Association of any kind does not imply causation.

22. Car thefts.

It might be reasonable to say that there is an *association* between the type of car you own and the risk that it will be stolen. The term correlation is reserved for describing linear associations between quantitative variables. Type of car is a categorical variable.

23. Coasters 2015.

a) It is appropriate to calculate correlation. Both height of the drop and speed are quantitative variables, the scatterplot shows an association that is linear, and there are no outliers.

b) There is a strong, positive, linear association between drop and speed; the greater the height of the initial drop, the higher the top speed.

24. Antidepressants.

a) It is appropriate to calculate correlation. Both placebo improvement and treated improvement are quantitative variables, the scatterplot shows an association that is linear, and there are no outliers.

b) There is a strong, positive, straight association between placebo and treated improvement. Experiments that showed a greater placebo effect also showed a greater mean improvement among patients who took an antidepressant.

25. Streams and hard water.

It is not appropriate to summarize the strength of the association between water hardness and pH with a correlation, since the association is not linear.

26. Traffic headaches.

It is not appropriate to summarize the strength of the association between highway speed and total delay with a correlation. The scatterplot is not linear, it has two high outliers, and there is a cluster of 8 cities that appear to be atypical. Correlation is not appropriate for such a relationship.

27. Cold nights.

The correlation is between the number of days since January 1 and temperature is likely to be near zero. We expect the temperature to be low in January, increase through the spring and summer, then decrease again. The relationship is not Straight Enough, so correlation is not an appropriate measure of strength.

28. Association V.

The researcher should have plotted the data first. A strong, curved relationship may have a very low correlation. In fact, correlation is only a useful measure of the strength of a linear relationship.

29. Prediction units.

The correlation between prediction error and year would not change, since the correlation is based on z-scores. The z-scores are the same whether the prediction errors are measured in nautical miles or miles.

30. More predictions.

The correlation between prediction error and year would not change, since the correlation is based on *z*-scores. The *z*-scores of the prediction errors are not changed by adding or subtracting a constant.

31. Correlation errors.

a) If the association between GDP and infant mortality is linear, a correlation of –0.772 shows a moderate, negative association. Generally, as GDP increases, infant mortality rate decreases.

b) Continent is a categorical variable. Correlation measures the strength of linear associations between quantitative variables.

32. More correlation errors.

a) Correlation must be between –1 and 1, inclusive. Correlation can never be 1.22.

b) Assuming the relation is linear, the strong correlation shows a relation, but it does not show causality.

33. Height and reading.

a) Actually, this *does* mean that taller children in elementary school are better readers. However, this does *not* mean that height causes good reading ability.

b) Older children are generally both taller and are better readers. Age is the lurking variable.

34. Smartphones and life expectancy.

a) No. It simply means that in countries where smartphone use is high, the life expectancy tends to be high as well.

b) General economic conditions of the country could affect both smart phone use and life expectancy. Richer countries generally have more smartphone use and better health care. The economy is a lurking variable.

35. Correlation conclusions I.

a) No. We don't know this from correlation alone. The relationship between age and income may be non-linear, or the relationship may contain outliers.

b) No. We can't tell the form of the relationship between age and income. We need to look at the scatterplot.

c) No. The correlation between age and income doesn't tell us anything about outliers.

d) Yes. Correlation is based on *z*-scores, and is unaffected by changes in units.

36. Correlation conclusions II.

a) No. We don't know this from correlation alone. The relationship between fuel efficiency and price may be non-linear, or the relationship may contain outliers.

b) No. We can't tell the form of the relationship between fuel efficiency and price. We need to look at the scatterplot.

c) No. The correlation between fuel efficiency and price doesn't tell us anything about outliers.

d) No. Correlation is based on *z*-scores, and is unaffected by changes in units.

37. Baldness and heart disease.

Even though the variables baldness and heart disease were assigned numerical values, they are categorical. Correlation is only an appropriate measure of the strength of linear association between quantitative variables. Their conclusion is meaningless.

38. Sample survey.

Even though ZIP codes are numbers, they are categorical variables representing different geographic areas. Likewise, even though the variable *Datasource* has numerical values, it is also categorical, representing the source from which the data were acquired. Correlation is only an appropriate measure of the strength of linear association between quantitative variables.

39. Income and housing.

a) There is a positive, moderately strong, linear relationship between *Housing Cost Index* and *Median Family Income*, with several states whose *Housing Cost Index* seems high for their *Median Family Income*, and one state whose *Housing Cost Index* seems low for their *Median Family Income*.

b) Correlation is based on *z*-scores. The correlation would still be 0.65.

c) Correlation is based on *z*-scores, and is unaffected by changes in units. The correlation would still be 0.65.

d) Washington D.C. would be a moderately high outlier, with *Housing Cost Index* high for its *Median Family Income*. Since it doesn't fit the pattern, the correlation would decrease slightly if Washington D.C. were included.

e) No. We can only say that higher *Housing Cost Index* scores are associated with higher *Median Family Income*, but we don't know why. There may be other variables at work.

f) Tau says that there is an association between median income and housing costs, but it makes no claims about the form of this association. It also has no requirement that the relationship be linear. Here it appears that the plot "thickens" from left to right. That could affect the correlation, but not tau.

40. Interest rates and mortgages 2015.

a) There is a negative and strong relationship between mortgage loan amount and interest rate, although the pattern for interest rates between 4% and 6% does not fit the pattern. There are no outliers in the relationship.

b) Correlation is based on *z*-scores. The correlation would still be –0.85.

c) Correlation is based on *z*-scores, and is unaffected by changes in units. The correlation would still be –0.85.

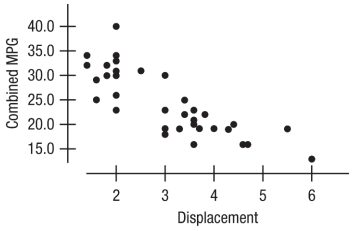
d) The given year has a very high mortgage amount for an interest rate that is that high. It doesn't fit the overall pattern, so the correlation would weaken (get closer to zero).

e) No. We can only say that lower interest rates are associated with larger mortgage amounts, but we don't know why. There may be other economic variables at work.

f) Tau says that there is a tendency for higher interest rates to be associated with less money loaned. Unlike the correlation, Tau does not require that we assume linearity—which isn't true here.

41. Fuel economy 2016.

a) A scatterplot of expected fuel economy vs. displacement is shown below.



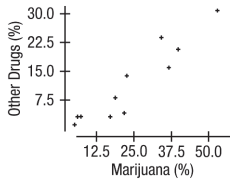
b) There is a strong, negative, linear association between displacement and mileage of the selected vehicles. There don't appear to be any outliers. All of the cars seem to fit the same pattern. Cars with larger engines tend to have lower mileage.

c) Since the relationship is linear, with no outliers, correlation is an appropriate measure of strength. The correlation between displacement and mileage of the selected vehicles is $r = -0.797$.

d) There is a strong linear relationship in the negative direction between displacement and highway gas mileage. Lower fuel efficiency is associated with larger engines.

42. Drug abuse.

a) A scatterplot of percentage of teens who have used other drugs vs. percentage who have used marijuana in the U.S. and 10 Western European countries is shown below.



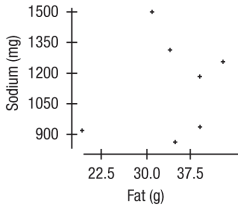
b) The correlation between the percent of teens who have used marijuana and the percent of teens who have used other drugs is.

c) The association between the percent of teens who have used marijuana and the percent of teens who have used other drugs is positive, strong, and linear. Countries with higher percentages of teens who have used marijuana tend to have higher percentages of teens that have used other drugs.

d) These results do not confirm that marijuana is a "gateway drug". An association exists between the percent of teens that have used marijuana and the percent of teens that have used other drugs. This does not mean that one caused the other.

43. Burgers.

a) There is no apparent association between the number of grams of fat and the number of milligrams of sodium in several brands of fast food burgers. The correlation is only $r = 0.199$, which is close to zero, an indication of no association. One burger had a much lower fat content than the other burgers, at 19 grams of fat, with 920 milligrams of sodium. Without this (comparatively) low fat burger, the correlation would have been $r = -0.325$. The plot could have explanatory and predictor variables swapped.

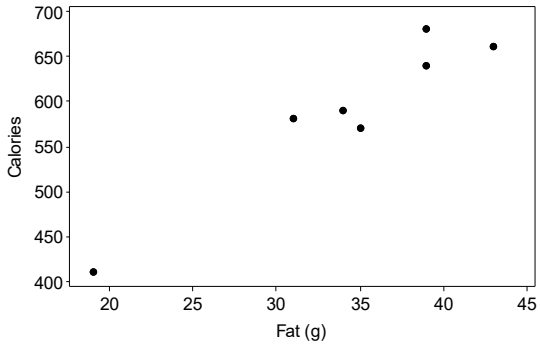


b) Spearman's rho is slightly negative. Using ranks doesn't allow the outlier to have as strong an influence and the remaining points have little or no association.

44. Burgers again.

a) The correlation between the number of calories and the number of grams of fat in several fast food burgers is $r = 0.961$. The association between the number of calories and the number of grams of fat in several fast food burgers is positive, straight, and strong. Typically, burgers with higher fat content have more calories. Even if the outlier at 410 calories and 19 grams of fat is set aside, the correlation is still quite strong at 0.837.

44. (continued)



b) Spearman's rho is only 0.83 because it doesn't allow the outlying point to have as strong an influence on the calculation.

45. Attendance 2016.

a) Number of runs scored and attendance are quantitative variables, the relationship between them appears to be straight, and there are no outliers, so calculating a correlation is appropriate.

b) The association between attendance and runs scored is positive, straight, and moderate in strength. Generally, as the number of runs scored increases, so does attendance.

c) There is evidence of an association between attendance and runs scored, but a cause-and-effect relationship between the two is not implied. There may be lurking variables that can account for the increases in each. For example, perhaps winning teams score more runs and also have higher attendance. We don't have any basis to make a claim of causation.

46. Second inning 2016.

a) Winning teams generally enjoy greater attendance at their home games. The association between home attendance and number of wins is positive, somewhat straight, and weak. The correlation is 0.457.

b) Wins. The correlation, $r = 0.457$ for wins and attendance is slightly higher than the correlation, $r = 0.431$ for runs and attendance.

c) The correlation between number of runs scored and number of wins is $r = 0.646$, indicating a possible moderately strong association. However, since there is no scatterplot of wins vs. runs provided, we can't be sure the relationship is straight. Correlation may not be an appropriate measure of the strength of the association.

47. Coasters 2015 sampled.

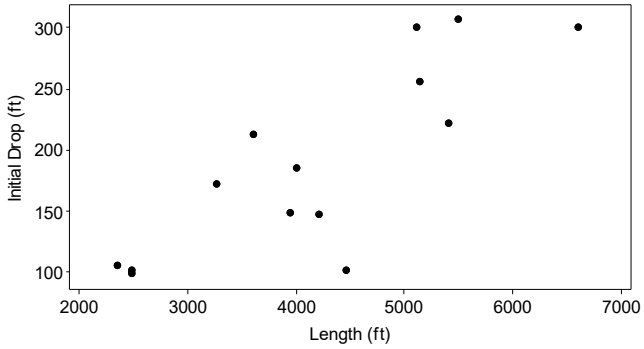
The distribution of the mean ride length of samples of 60 coasters is unimodal and symmetric, so we could use the 68–95–99.7 Rule for it. The distribution of median ride length of samples of 60 coasters is too skewed for the rule.

48. Housing prices sampled.

The distribution of the IQR of home age for samples of 105 homes is unimodal and symmetric, so we could use the 68–95–99.7 Rule for it. The distribution of the range of home age for samples of 105 homes has more than one mode, gaps, and is not symmetric. The 68–95–99.7 Rule is not appropriate here.

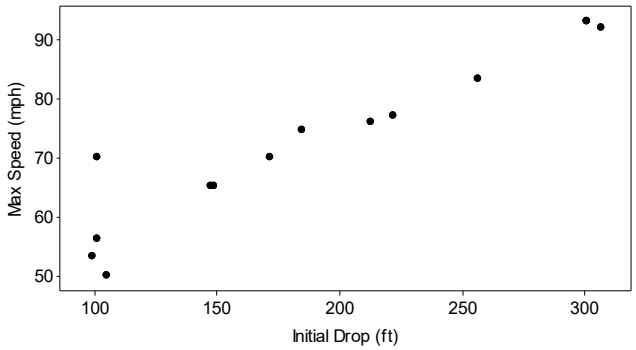
49. Thrills 2013.

The scatterplot below shows that the association between duration and length is linear, positive, and moderate, with no outliers. Generally, rides on coasters with a greater length tend to last longer. The correlation between length and duration is 0.736, indicating a moderate association.

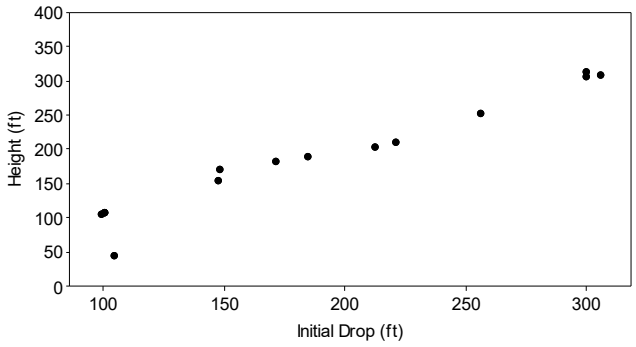


50. Thrills II.

a) With a correlation of 0.950, there is a very strong, positive, and linear relationship between the initial drop of a roller coaster and its maximum speed. It appears that the maximum speed of a roller coaster is directly related to the height of the first drop.



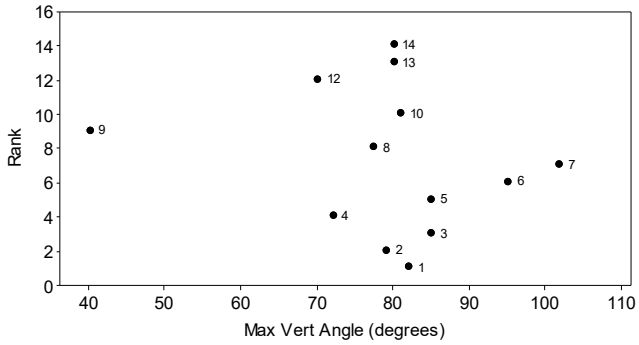
b) Similar to part (a), the height and initial drop are directly related. The relationship is strong, positive, and reasonably linear and has a correlation of 0.974.



c) The initial drop of a coaster clearly affects the height and speed of the coaster. This is apparent, as most coasters start with a tall ascent that is the largest and the fastest. The initial drop is also moderately correlated with the duration of the ride ($r = 0.523$).

51. Thrills III.

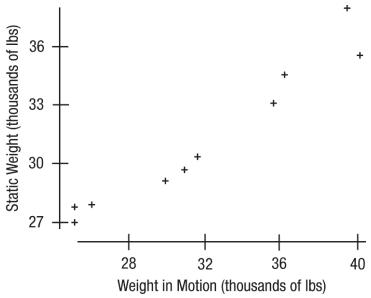
a) Lower rank is better, so variables like speed and duration should have negative associations with rank. We would expect that as one variable (say length of ride) increases, the rank will improve, which means it will decrease. In the scatterplot below, the points are labeled with roller coaster rank.



b) Max vertical angle is the only variable with a negative correlation with rank, and that is due almost entirely to a single coaster, Nemesis.

52. Vehicle weights.

a) A scatterplot of the Static Weight vs. Weight-in-Motion of the test truck is shown below.



b) The association between static weight and weight-in-motion is positive, very strong, and linear. There may be a hint of a curve in the scatterplot.

c) The new scale is able to predict the static weight fairly well, except possibly at the high end. It may be possible to predict the static weight from the new scale accurately enough to be useful. But the weight-in-motion measurements seem a bit too high.

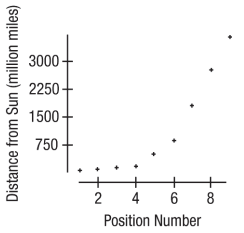
d) The correlation between static weight and weight-in-motion is $r = 0.965$.

e) Weighing the trucks in kilograms instead of pounds would not change the correlation. Correlation, like *z*-score, has no units. It is a numerical measure of the degree of linear association between two variables.

f) At the higher end of the weight-in-motion scale, there is one point where the weight in motion is much higher than the static weight. The points all lie above the line $y = x$, so the predicted values are not the same as the actual values, so the new scale may have to be recalibrated.

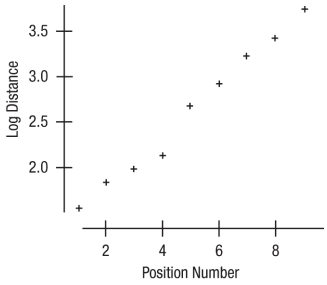
53. Planets (more or less).

a) The association between position number of each planet and its distance from the sun (in millions of miles) is very strong, positive and nonlinear. The scatterplot is shown below.



b) The relationship between Position Number and distance from the sun is not linear. Correlation is a measure of the degree of *linear* association between two variables.

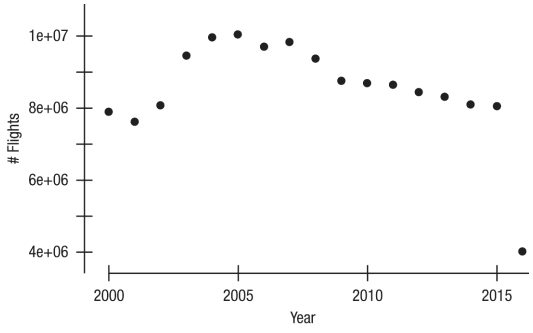
c) The scatterplot of the logarithm of distance versus Position Number (shown below) still shows a strong, positive relationship, but it is straighter than the previous scatterplot. It still shows a curve in the scatterplot, but it is straight enough that correlation may now be used as an appropriate measure of the strength of the relationship between logarithm of distance and Position Number, which will in turn give an indication of the strength of the association.



54. Flights 2016.

a) The correlation between the year and the number of flights is $r = -0.415$.

b) From 2000 until 2005, the number of flights generally increased. But since then, flights have generally decreased, and the value for 2016 is an extreme low outlier.



c) Correlation is not appropriate since the relationship is not linear. There are two trends in the data and the outlier for 2016.

d) The point for 2016 is based on incomplete data and shouldn't be included in any overall summary or plot.

e) Kendall's tau can be used in spite of the nonlinearity and outlier, but without a consistent pattern of growth or shrinking, it won't find much.