EXERCISES

1. Association. Suppose you were to collect data for each pair of variables. You want to make a scatterplot. Which variable would you use as the explanatory variable and which as the response variable? Why? What would you expect to see in the scatterplot? Discuss the likely direction, form, and strength.

a) Apples: weight in grams, weight in ounces
b) Apples: circumference (inches), weight (ounces)
c) College freshmen: shoe size, grade point average
d) Gasoline: number of miles you drove since filling up, gallons remaining in your tank

2. Association. Suppose you were to collect data for each pair of variables. You want to make a scatterplot. Which variable would you use as the explanatory variable and which as the response variable? Why? What would you expect to see in the scatterplot? Discuss the likely direction, form, and strength.

a) T-shirts at a store: price each, number sold
b) Scuba diving: depth, water pressure
c) Scuba diving: depth, visibility
d) All elementary school students: weight, score on a reading test

3. Association. Suppose you were to collect data for each pair of variables. You want to make a scatterplot. Which variable would you use as the explanatory variable and which as the response variable? Why? What would you expect to see in the scatterplot? Discuss the likely direction, form, and strength.

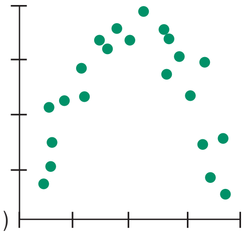
a) When climbing mountains: altitude, temperature
b) For each week: ice cream cone sales, air-conditioner sales
c) People: age, grip strength
d) Drivers: blood alcohol level, reaction time

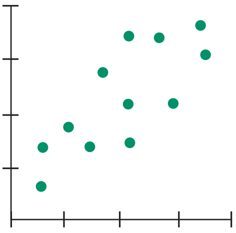
4. Association. Suppose you were to collect data for each pair of variables. You want to make a scatterplot. Which variable would you use as the explanatory variable and which as the response variable? Why? What would you expect to see in the scatterplot? Discuss the likely direction, form, and strength.

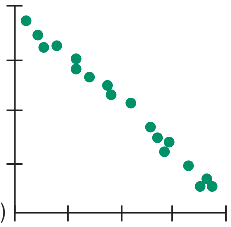
a) Long-distance calls: time (minutes), cost
b) Lightning strikes: distance from lightning, time delay of the thunder
c) A streetlight: its apparent brightness, your distance from it
d) Cars: weight of car, age of owner

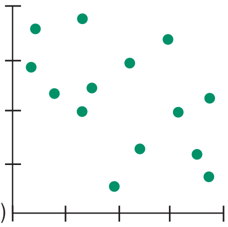
5. Scatterplots. Which of the scatterplots at the top of the next column show

a) little or no association?
b) a negative association?
c) a linear association?
d) a moderately strong association?
e) a very strong association?


(1)

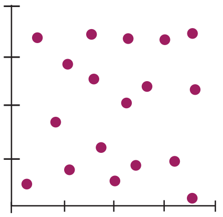

(2)

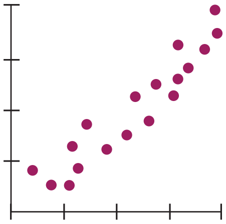

(3)

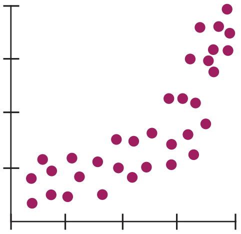

(4)

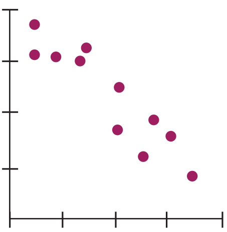
6. Scatterplots. Which of the scatterplots below show

a) little or no association?
b) a negative association?
c) a linear association?
d) a moderately strong association?
e) a very strong association?

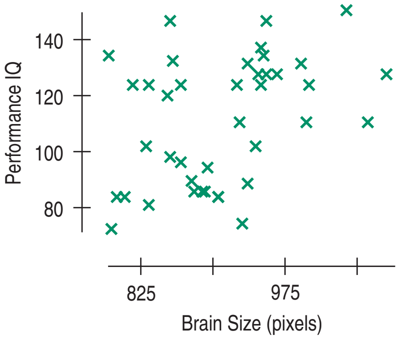

(1)


(2)

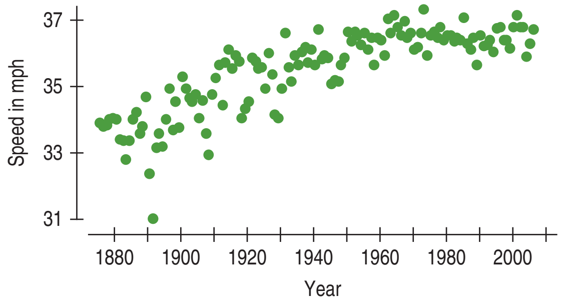

(3)


(4)

T **7. Performance IQ scores vs. brain size.** A study examined brain size (measured as pixels counted in a digitized magnetic resonance image [MRI] of a cross section of the brain) and IQ (4 Performance scales of the Weschler IQ test) for college students. The scatterplot shows the Performance IQ scores vs. the brain size. Comment on the association between brain size and IQ as seen in the scatterplot on the next page.

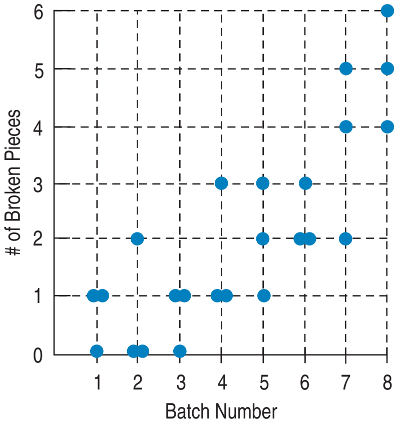


T **8. Kentucky Derby 2006.** The fastest horse in Kentucky Derby history was Secretariat in 1973. The scatterplot shows speed (in miles per hour) of the winning horses each year.



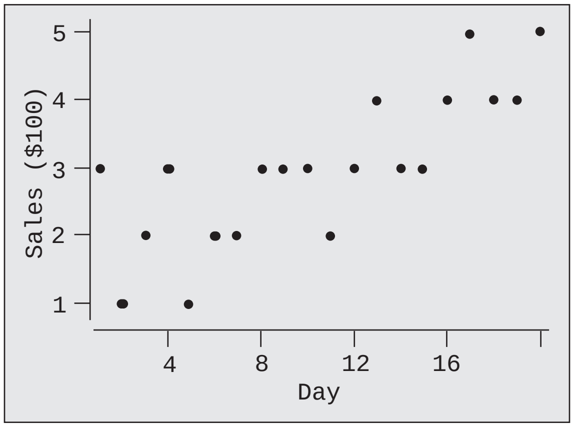
What do you see? In most sporting events, performances have improved and continue to improve, so surely we anticipate a positive direction. But what of the form? Has the performance increased at the same rate throughout the last 130 years?

9. Firing pottery. A ceramics factory can fire eight large batches of pottery a day. Sometimes a few of the pieces break in the process. In order to understand the problem better, the factory records the number of broken pieces in each batch for 3 days and then creates the scatterplot shown.



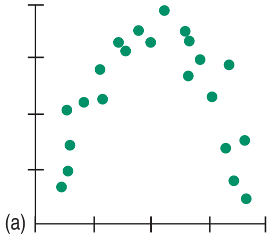
a) Make a histogram showing the distribution of the number of broken pieces in the 24 batches of pottery examined.
b) Describe the distribution as shown in the histogram. What feature of the problem is more apparent in the histogram than in the scatterplot?
c) What aspect of the company's problem is more apparent in the scatterplot?

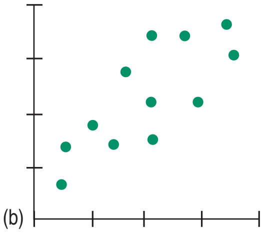
10. Coffee sales. Owners of a new coffee shop tracked sales for the first 20 days and displayed the data in a scatterplot (by day).

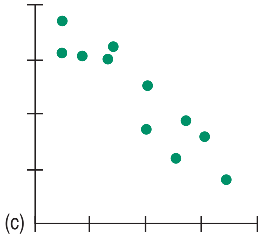


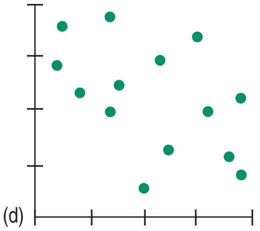
a) Make a histogram of the daily sales since the shop has been in business.
b) State one fact that is obvious from the scatterplot, but not from the histogram.
c) State one fact that is obvious from the histogram, but not from the scatterplot.

11. Matching. Here are several scatterplots. The calculated correlations are −0.923, −0.487, 0.006, and 0.777. Which is which?

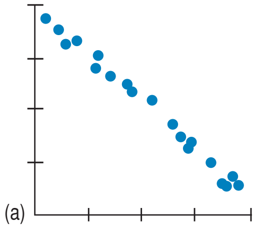

(a)

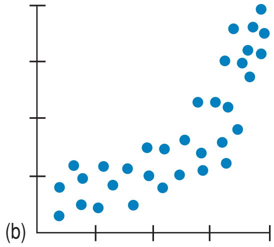

(b)

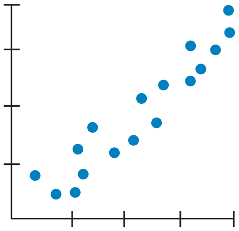

(c)


(d)

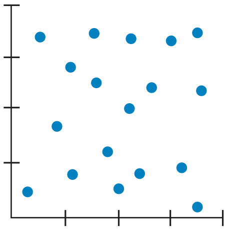
12. Matching. Here and on the next page are several scatterplots. The calculated correlations are −0.977, −0.021, 0.736, and 0.951. Which is which?


(a)


(b)



(c)

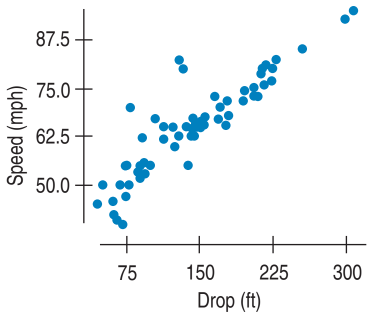


(d)

13. Politics. A candidate for office claims that "there is a correlation between television watching and crime." Criticize this statement on statistical grounds.

14. Car thefts. The National Insurance Crime Bureau reports that Honda Accords, Honda Civics, and Toyota Camrys are the cars most frequently reported stolen, while Ford Tauruses, Pontiac Vibes, and Buick LeSabres are stolen least often. Is it reasonable to say that there's a correlation between the type of car you own and the risk that it will be stolen?

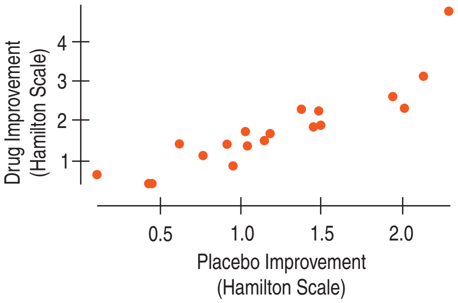
T **15. Roller coasters.** Roller coasters get all their speed by dropping down a steep initial incline, so it makes sense that the height of that drop might be related to the speed of the coaster. Here's a scatterplot of top *Speed* and largest *Drop* for 75 roller coasters around the world.



a) Does the scatterplot indicate that it is appropriate to calculate the correlation? Explain.

b) In fact, the correlation of *Speed* and *Drop* is 0.91. Describe the association.

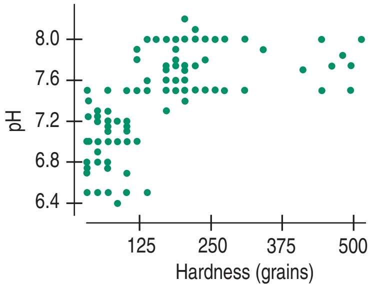
T **16. Antidepressants.** A study compared the effectiveness of several antidepressants by examining the experiments in which they had passed the FDA requirements. Each of those experiments compared the active drug with a placebo, an inert pill given to some of the subjects. In each experiment some patients treated with the placebo had improved, a phenomenon called the *placebo effect*. Patients' depression levels were evaluated on the Hamilton Depression Rating Scale, where larger numbers indicate greater improvement. (The Hamilton scale is a widely accepted standard that was used in each of the independently run studies.) The scatterplot at the top of the next column compares mean improvement levels for the antidepressants and placebos for several experiments.



a) Is it appropriate to calculate the correlation? Explain.

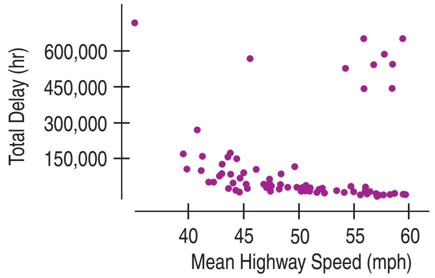
b) The correlation is 0.898. Explain what we have learned about the results of these experiments.

T **17. Hard water.** In a study of streams in the Adirondack Mountains, the following relationship was found between the water's pH and its hardness (measured in grains):



Is it appropriate to summarize the strength of association with a correlation? Explain.

18. Traffic headaches. A study of traffic delays in 68 U.S. cities found the following relationship between total delays (in total hours lost) and mean highway speed:



Is it appropriate to summarize the strength of association with a correlation? Explain.

19. Cold nights. Is there an association between time of year and the nighttime temperature in North Dakota? A researcher assigned the numbers 1–365 to the days January 1–December 31 and recorded the temperature at 2:00 a.m. for each. What might you expect the correlation between *DayNumber* and *Temperature* to be? Explain.

20. Association. A researcher investigating the association between two variables collected some data and was surprised when he calculated the correlation. He had expected to find a fairly strong association, yet the correlation was near 0. Discouraged, he didn't bother making a scatterplot. Explain to him how the scatterplot could still reveal the strong association he anticipated.

21. Prediction units. The errors in predicting hurricane tracks (examined in this chapter) were given in nautical miles. An ordinary mile is 0.86898 nautical miles. Most people living on the Gulf Coast of the United States would prefer to know the prediction errors in miles rather than nautical miles. Explain why converting the errors to miles would not change the correlation between *Prediction Error* and *Year.*

22. More predictions. Hurricane Katrina's hurricane force winds extended 120 miles from its center. Katrina was a big storm, and that affects how we think about the prediction errors. Suppose we add 120 miles to each error to get an idea of how far from the predicted track we might still find damaging winds. Explain what would happen to the correlation between *Prediction Error* and *Year,* and why.

23. Correlation errors. Your Economics instructor assigns your class to investigate factors associated with the gross domestic product (*GDP*) of nations. Each student examines a different factor (such as *Life Expectancy, Literacy Rate,* etc.) for a few countries and reports to the class. Apparently, some of your classmates do not understand Statistics very well because you know several of their conclusions are incorrect. Explain the mistakes in their statements below.

a) "My very low correlation of −0.772 shows that there is almost no association between *GDP* and *Infant Mortality Rate.*"

b) "There was a correlation of 0.44 between *GDP* and *Continent.*"

24. More correlation errors. Students in the Economics class discussed in Exercise 23 also wrote these conclusions. Explain the mistakes they made.

a) "There was a very strong correlation of 1.22 between *Life Expectancy* and *GDP.*"

b) "The correlation between *Literacy Rate* and *GDP* was 0.83. This shows that countries wanting to increase their standard of living should invest heavily in education."

25. Height and reading. A researcher studies children in elementary school and finds a strong positive linear association between height and reading scores.

a) Does this mean that taller children are generally better readers?

b) What might explain the strong correlation?

26. Cellular telephones and life expectancy. A survey of the world's nations in 2004 shows a strong positive correlation between percentage of the country using cell phones and life expectancy in years at birth.

a) Does this mean that cell phones are good for your health?

b) What might explain the strong correlation?

27. Correlation conclusions I. The correlation between *Age* and *Income* as measured on 100 people is $r = 0.75$. Explain whether or not each of these possible conclusions is justified:

a) When *Age* increases, *Income* increases as well.

b) The form of the relationship between *Age* and *Income* is straight.

c) There are no outliers in the scatterplot of *Income* vs. *Age.*

d) Whether we measure *Age* in years or months, the correlation will still be 0.75.

28. Correlation conclusions II. The correlation between *Fuel Efficiency* (as measured by miles per gallon) and *Price* of 150 cars at a large dealership is $r = -0.34$. Explain whether or not each of these possible conclusions is justified:

a) The more you pay, the lower the fuel efficiency of your car will be.

b) The form of the relationship between *Fuel Efficiency* and *Price* is moderately straight.

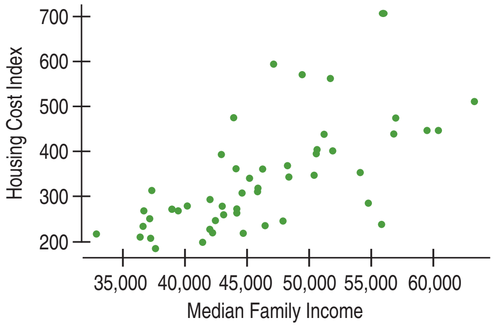
c) There are several outliers that explain the low correlation.

d) If we measure *Fuel Efficiency* in kilometers per liter instead of miles per gallon, the correlation will increase.

29. Baldness and heart disease. Medical researchers followed 1435 middle-aged men for a period of 5 years, measuring the amount of *Baldness* present (none = 1, little = 2, some = 3, much = 4, extreme = 5) and presence of *Heart Disease* (No = 0, Yes = 1). They found a correlation of 0.089 between the two variables. Comment on their conclusion that this shows that baldness is not a possible cause of heart disease.

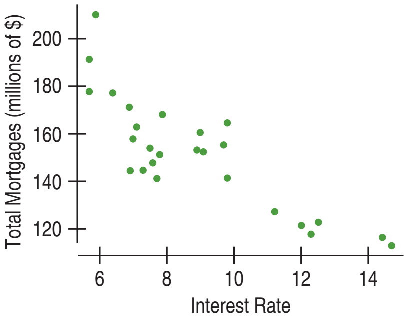
30. Sample survey. A polling organization is checking its database to see if the two data sources it used sampled the same zip codes. The variable *Datasource* = 1 if the data source is MetroMedia, 2 if the data source is DataQwest, and 3 if it's RollingPoll. The organization finds that the correlation between five-digit zip code and *Datasource* is −0.0229. It concludes that the correlation is low enough to state that there is no dependency between *Zip Code* and *Source of Data.* Comment.

T **31. Income and housing.** The Office of Federal Housing Enterprise Oversight (www.ofheo.gov) collects data on various aspects of housing costs around the United States. Here is a scatterplot of the *Housing Cost Index* versus the *Median Family Income* for each of the 50 states. The correlation is 0.65.



a) Describe the relationship between the *Housing Cost Index* and the *Median Family Income* by state.
b) If we standardized both variables, what would the correlation coefficient between the standardized variables be?
c) If we had measured *Median Family Income* in thousands of dollars instead of dollars, how would the correlation change?
d) Washington, DC, has a Housing Cost Index of 548 and a median income of about $45,000. If we were to include DC in the data set, how would that affect the correlation coefficient?
e) Do these data provide proof that by raising the median income in a state, the Housing Cost Index will rise as a result? Explain.

32. Interest rates and mortgages. Since 1980, average mortgage interest rates have fluctuated from a low of under 6% to a high of over 14%. Is there a relationship between the amount of money people borrow and the interest rate that's offered? Here is a scatterplot of *Total Mortgages* in the United States (in millions of 2005 dollars) versus *Interest Rat*e at various times over the past 26 years. The correlation is −0.84.



a) Describe the relationship between *Total Mortgages* and *Interest Rate.*
b) If we standardized both variables, what would the correlation coefficient between the standardized variables be?
c) If we were to measure *Total Mortgages* in thousands of dollars instead of millions of dollars, how would the correlation coefficient change?
d) Suppose in another year, interest rates were 11% and mortgages totaled $250 million. How would including that year with these data affect the correlation coefficient?
e) Do these data provide proof that if mortgage rates are lowered, people will take out more mortgages? Explain.

33. Fuel economy 2007. Here are advertised horsepower ratings and expected gas mileage for several 2007 vehicles. (http://www.kbb.com/KBB/ReviewsAndRatings)

Vehicle	Horsepower	Highway Gas Mileage (mpg)
Audi A4	200	32
BMW 328	230	30
Buick LaCrosse	200	30
Chevy Cobalt	148	32
Chevy TrailBlazer	291	22
Ford Expedition	300	20
GMC Yukon	295	21
Honda Civic	140	40
Honda Accord	166	34
Hyundai Elantra	138	36
Lexus IS 350	306	28
Lincoln Navigator	300	18
Mazda Tribute	212	25
Toyota Camry	158	34
Volkswagen Beetle	150	30

a) Make a scatterplot for these data.
b) Describe the direction, form, and strength of the plot.
c) Find the correlation between horsepower and miles per gallon.
d) Write a few sentences telling what the plot says about fuel economy.

34. Drug abuse. A survey was conducted in the United States and 10 countries of Western Europe to determine the percentage of teenagers who had used marijuana and other drugs. The results are summarized in the table.

	Percent Who Have Used	
Country	**Marijuana**	**Other Drugs**
Czech Rep.	22	4
Denmark	17	3
England	40	21
Finland	5	1
Ireland	37	16
Italy	19	8
No. Ireland	23	14
Norway	6	3
Portugal	7	3
Scotland	53	31
USA	34	24

a) Create a scatterplot.
b) What is the correlation between the percent of teens who have used marijuana and the percent who have used other drugs?
c) Write a brief description of the association.
d) Do these results confirm that marijuana is a "gateway drug," that is, that marijuana use leads to the use of other drugs? Explain.

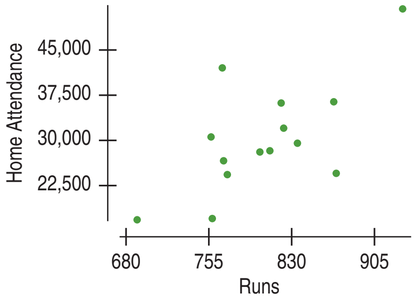
35. **Burgers.** Fast food is often considered unhealthy because much of it is high in both fat and sodium. But are the two related? Here are the fat and sodium contents of several brands of burgers. Analyze the association between fat content and sodium.

Fat (g)	19	31	34	35	39	39	43
Sodium (mg)	920	1500	1310	860	1180	940	1260

36. **Burgers II.** In the previous exercise you analyzed the association between the amounts of fat and sodium in fast food hamburgers. What about fat and calories? Here are data for the same burgers:

Fat (g)	19	31	34	35	39	39	43
Calories	410	580	590	570	640	680	660

37. **Attendance 2006.** American League baseball games are played under the designated hitter rule, meaning that pitchers, often weak hitters, do not come to bat. Baseball owners believe that the designated hitter rule means more runs scored, which in turn means higher attendance. Is there evidence that more fans attend games if the teams score more runs? Data collected from American League games during the 2006 season indicate a correlation of 0.667 between runs scored and the number of people at the game. (http://mlb.mlb.com)

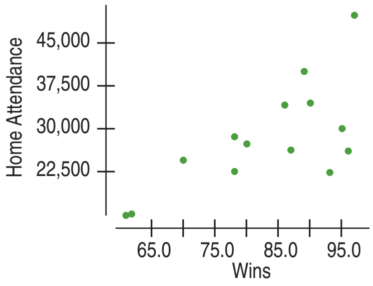


a) Does the scatterplot indicate that it's appropriate to calculate a correlation? Explain.
b) Describe the association between attendance and runs scored.
c) Does this association prove that the owners are right that more fans will come to games if the teams score more runs?

38. **Second inning 2006.** Perhaps fans are just more interested in teams that win. The displays below are based on American League teams for the 2006 season. (http://espn.go.com) Are the teams that win necessarily those which score the most runs?

Correlation

	Wins	Runs	Attend
Wins	1.000		
Runs	0.605	1.000	
Attend	0.697	0.667	1.000



a) Do winning teams generally enjoy greater attendance at their home games? Describe the association.
b) Is attendance more strongly associated with winning or scoring runs? Explain.
c) How strongly is scoring more runs associated with winning more games?

39. **Thrills.** People who responded to a July 2004 Discovery Channel poll named the 10 best roller coasters in the United States. The table below shows the length of the initial drop (in feet) and the duration of the ride (in seconds). What do these data indicate about the height of a roller coaster and the length of the ride you can expect?

Roller Coaster	State	Drop (ft)	Duration (sec)
Incredible Hulk	FL	105	135
Millennium Force	OH	300	105
Goliath	CA	255	180
Nitro	NJ	215	240
Magnum XL-2000	OH	195	120
The Beast	OH	141	65
Son of Beast	OH	214	140
Thunderbolt	PA	95	90
Ghost Rider	CA	108	160
Raven	IN	86	90

40. **Vehicle weights.** The Minnesota Department of Transportation hoped that they could measure the weights of big trucks without actually stopping the vehicles by using a newly developed "weight-in-motion" scale. To see if the new device was accurate, they conducted a calibration test. They weighed several stopped trucks (static weight) and assumed that this weight was correct. Then they weighed the trucks again while they were moving to see how well the new scale could estimate the actual weight. Their data are given in the table on the next page.

WEIGHTS (1000S OF LBS)

Weight-in-Motion	Static Weight
26.0	27.9
29.9	29.1
39.5	38.0
25.1	27.0
31.6	30.3
36.2	34.5
25.1	27.8
31.0	29.6
35.6	33.1
40.2	35.5

a) Make a scatterplot for these data.
b) Describe the direction, form, and strength of the plot.
c) Write a few sentences telling what the plot says about the data. (*Note*: The sentences should be about weighing trucks, not about scatterplots.)
d) Find the correlation.
e) If the trucks were weighed in kilograms, how would this change the correlation? (1 kilogram = 2.2 pounds)
f) Do any points deviate from the overall pattern? What does the plot say about a possible recalibration of the weight-in-motion scale?

41. Planets (more or less). On August 24, 2006, the International Astronomical Union voted that Pluto is not a planet. Some members of the public have been reluctant to accept that decision. Let's look at some of the data. (We'll see more in the next chapter.) Is there any pattern to the locations of the planets? The table shows the average distance of each of the traditional nine planets from the sun.

Planet	Position Number	Distance from Sun (million miles)
Mercury	1	36
Venus	2	67
Earth	3	93
Mars	4	142
Jupiter	5	484
Saturn	6	887
Uranus	7	1784
Neptune	8	2796
Pluto	9	3666

a) Make a scatterplot and describe the association. (Remember: direction, form, and strength!)
b) Why would you not want to talk about the correlation between a planet's *Position* and *Distance* from the sun?
c) Make a scatterplot showing the logarithm of *Distance* vs. *Position*. What is better about this scatterplot?

42. Flights. The number of flights by U.S. Airlines has grown rapidly. Here are the number of flights flown in each year from 1995 to 2005.
a) Find the correlation of *Flights* with *Year.*
b) Make a scatterplot and describe the trend.
c) Note two reasons that the correlation you found in (a) is not a suitable summary of the strength of the association. Can you account for these violations of the conditions?

Year	Flights
1995	5,327,435
1996	5,351,983
1997	5,411,843
1998	5,384,721
1999	5,527,884
2000	5,683,047
2001	5,967,780
2002	5,271,359
2003	6,488,539
2004	7,129,270
2005	7,140,596

JUST CHECKING Answers

1. We know the scores are quantitative. We should check to see if the Straight Enough Condition and the Outlier Condition are satisfied by looking at a scatterplot of the two scores.
2. It won't change.
3. It won't change.
4. They are likely to have done poorly. The positive correlation means that low scores on Exam 1 are associated with low scores on Exam 2 (and similarly for high scores).
5. No. The general association is positive, but individual performances may vary.