Unit 10: Scatterplots

SUMMARY OF VIDEO

The video opens with views of manatees, large gentle sea creatures that live along the coast of Florida. Each year, a number of manatees are killed by powerboat propellers. From 1977 to 2011 the number of yearly manatee deaths appeared to be rising, but then so were the number of powerboat registrations. One avenue of investigation into manatee deaths is to look at the data for clues to the relationship between the number of powerboats registered in Florida in a given year and the number of manatees killed by powerboats. These are both quantitative variables; that is, they are measured in meaningful numerical units for which arithmetic operations make sense.

Powerboat registrations help explain manatee deaths, so powerboat registrations is the explanatory variable and manatees killed by powerboats is the response variable. The scatterplot in Figure 10.1, with the explanatory variable on the horizontal axis and the response variable on the vertical axis, shows the relationship between the two variables.

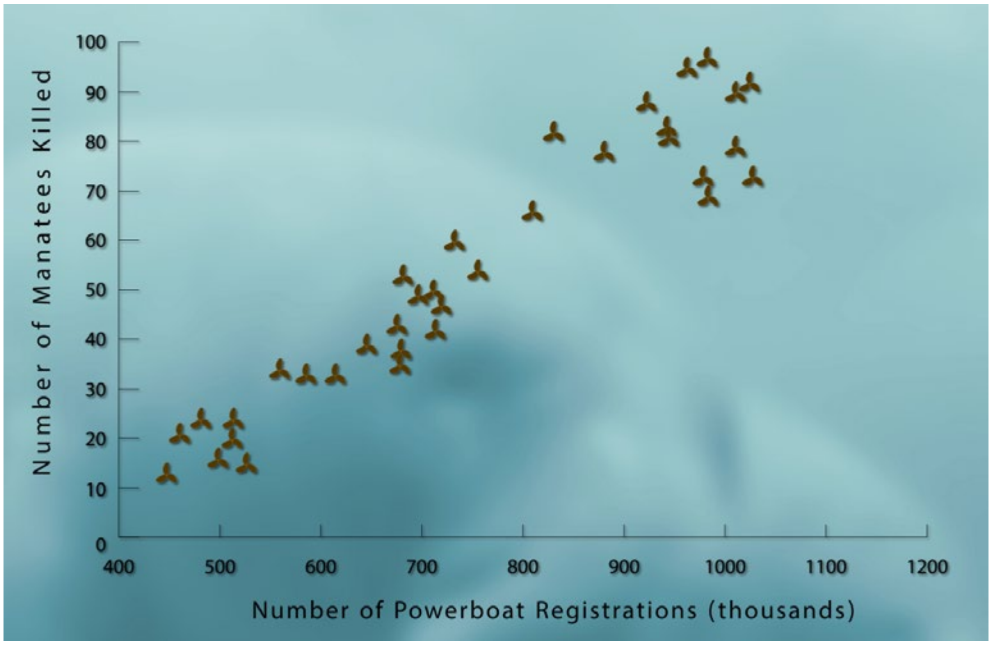

Figure 10.1. Scatterplot of manatee deaths and powerboat registrations.

The relationship shown in the scatterplot confirms our suspicions: as powerboat registrations rose, so did manatee fatalities. Our scatterplot shows positive association because as one variable increased, the other also tended to increase. An example of negative association is shown in Figure 10.2, a graph of hypothetical data on the time to make a pie (response variable) versus numbers of tries in making this type of pie (explanatory variable). As the number of pies produced increased, the time required decreased because we get more efficient in making this type of pie.

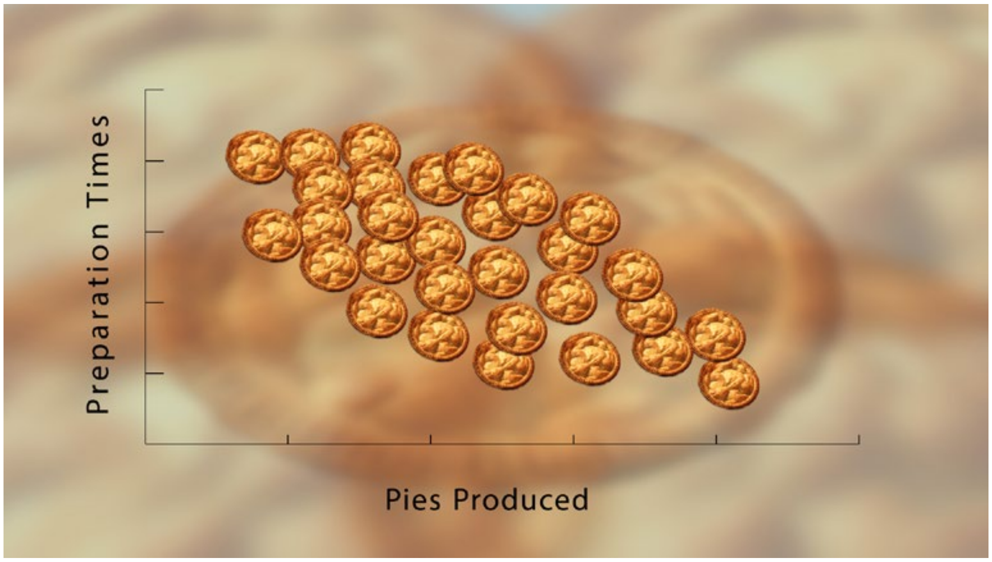

Figure 10.2. Scatterplot of preparation times and pies produced.

When looking at a scatterplot, think about the overall pattern, how strong it is, and its direction. The manatee scatterplot has an overall pattern that is linear, as is shown in Figure 10.3; the points lie roughly in a straight line. The points stay pretty close to that line – hence, the relationship is strong. We've already noted that the relationship is positive and so is the slope of the line.

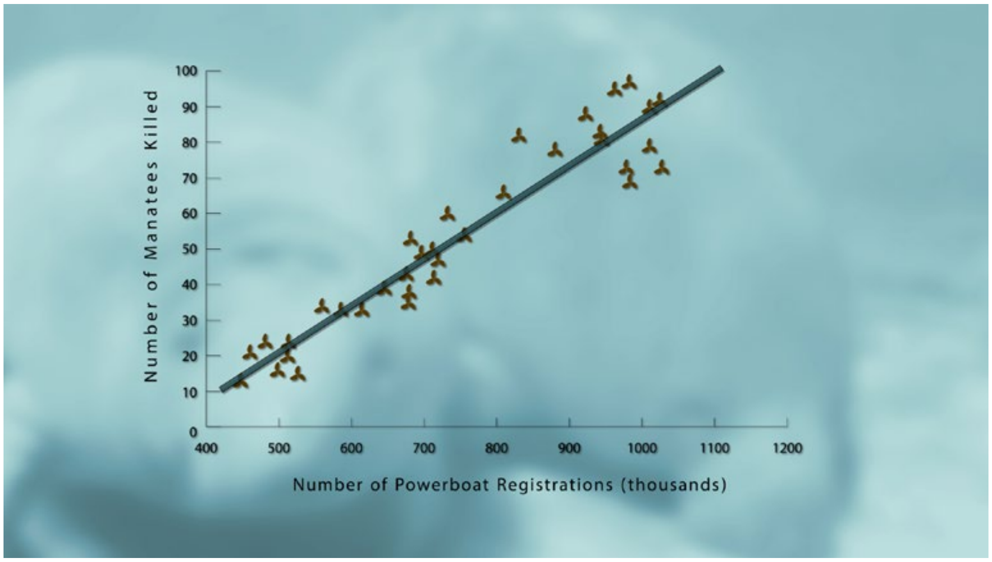

Figure 10.3. Linear relationship between the variables.

Keep in mind that not all relationships are linear. For example, the relationship shown in Figure 10.4 shows a curved pattern.

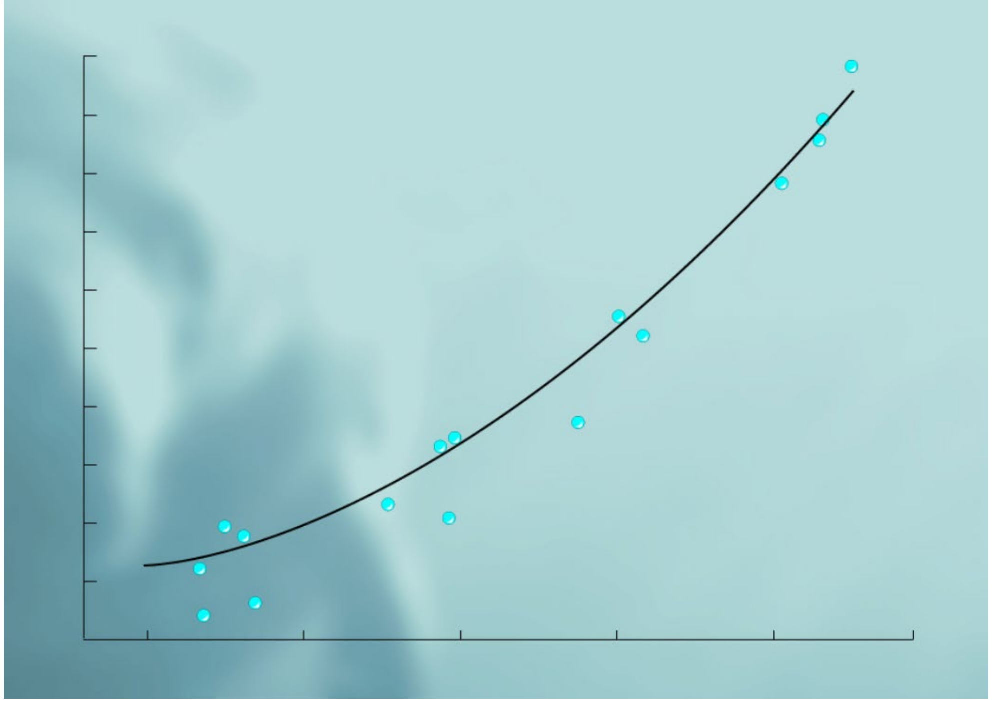

Figure 10.4. Relationship with a curved pattern.

Although a scatterplot shows the nature of a relationship between two variables, it doesn’t prove that one variable causes the changes in the other. Even so, identifying a relationship between powerboat registrations and manatee deaths provided sufficient evidence for the state of Florida to consider additional ways to protect the manatees from those deadly propellers.

STUDENT LEARNING OBJECTIVES

A. Know how to make a scatterplot of quantitative bivariate data.

B. Recognize when there is an explanatory/response distinction and put the explanatory variable on the horizontal axis in making a scatterplot.

C. Recognize patterns in a scatterplot, especially positive or negative association and a linear (straight line) pattern.

D. Recognize outliers in a scatterplot and understand that these points should be investigated.

CONTENT OVERVIEW

A **bivariate data** set consists of measurements or observations on two variables from the same individual or subject. When the two variables are quantitative, such as the height and weight of a group of children, we can display the data in a **scatterplot** by plotting the ordered pairs of data values.

In plotting bivariate quantitative data, we need to decide which variable to put on the horizontal axis and which to put on the vertical axis. In many cases, it is correct to put either variable on the horizontal axis. But if one variable, the **explanatory variable**, explains or causes changes in the other variable, the **response variable**, the explanatory variable is put on the horizontal axis. In plotting a variable against time, time is always on the horizontal axis.

A fundamental strategy of data analysis is: Make a graph of your data, then look for an overall pattern and for striking deviations from that pattern. In applying this strategy to relationships between two quantitative variables, the basic graph is the scatterplot. Look for the direction and shape of the pattern and identify any deviations from the overall pattern. **Positive association** and **negative association** describe the direction. Linear or curved (quadratic, exponential, and so on) describe the shape. Not all plots have either a clear direction or a clear shape; hence, a simple description may not be possible in all cases. Outliers are the most noticeable kind of deviation from the overall pattern.

KEY TERMS

For **bivariate data** measurements or observations are recorded on two attributes for each individual or subject under study. For example, in the video segment on manatees, the two attributes are number of manatee deaths and powerboat registrations and the subject is the year. For **multivariate data** measurements or observations are recorded on two or more attributes for each individual or subject under study.

A **response variable** measures an outcome of a study. The response variable is always plotted on the vertical axis of a scatterplot. An **explanatory variable** attempts to explain the observed outcomes. The explanatory variable is always plotted on the horizontal axis of a scatterplot.

Two variables are **positively associated** when above-average values of one tend to accompany above-average values of the other and below-average values of one tend to accompany below-average values of the other. In a scatterplot a positive association would appear as a pattern of dots in the lower left to the upper right. Two variables are **negatively associated** when above-average values of one accompany below-average values of the other, and vice versa. In a scatterplot a negative association would appear as a pattern of dots in the upper left to the lower right. (See Figure 10.5.)

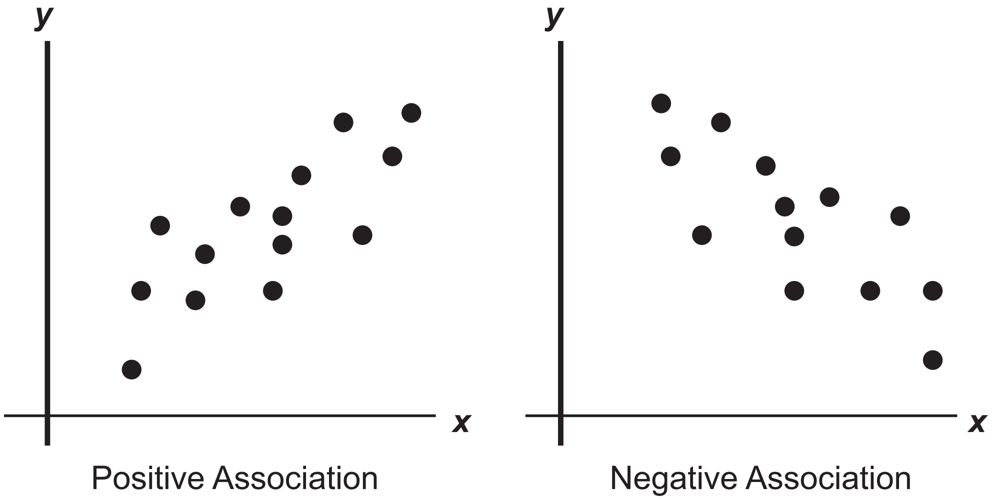

Figure 10.5. Scatterplots illustrating positive and negative association.

A scatterplot has **linear form** when the dots appear to be randomly scattered on either side of a straight line**.** However, sometimes the data form a curved pattern. In that case, we say the scatterplot has **nonlinear form**. Figure 10.6 shows two scatterplots, one with linear form and one with nonlinear form.

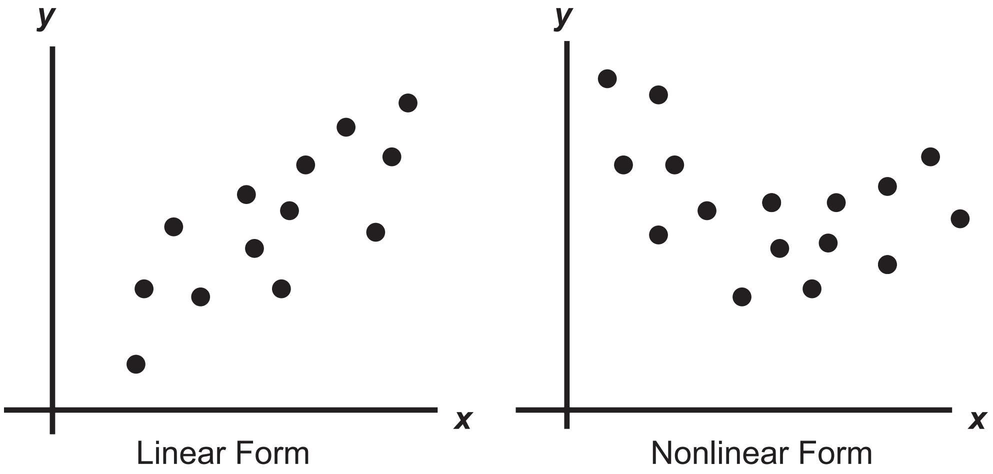

Figure 10.6. Scatterplots illustrating linear and nonlinear form.

THE VIDEO

Take out a piece of paper and be ready to write down answers to these questions as you watch the video.

1. What is a manatee?

2. What does a *scatterplot* show about the relationship between the number of powerboats registered in Florida and the number of manatees killed by powerboats?

3. Why is the number of boats plotted on the horizontal axis of this scatterplot?

4. What trend would you expect to see in a scatterplot of two variables that have a negative association?

UNIT ACTIVITY:

RELATIONSHIPS BETWEEN PARENT AND STUDENT HEIGHTS

This activity is based on class data. You will need to measure your height. If you are male, measure your father's height; if you are female, measure your mother's height.

Record the height data from the class in Tables 10.1 and 10.2, which follow this activity.

1. Make a scatterplot of female student's height versus mother's height. (Mother's height goes on the horizontal axis.)

2. Make a scatterplot of male student's height versus father's height. (Father's height goes on the horizontal axis.)

3. Compare the patterns of the two scatterplots from questions 1 and 2.

4. Next make a scatterplot in which you can visualize three variables: student height, parent height, and student gender. Use parent's height for the horizontal axis and student's height for the vertical axis. Use different symbols (or different colors) for data from males and females.

5. Write a brief description about what can be learned from the scatterplot drawn for question 4.

HEIGHT DATA FOR FEMALE STUDENTS

Student Name	Class (sophomore, junior, etc.)	Student’s Height (inches)	Mother's Height (inches)

Table 10.1. Female students and mothers’ heights

HEIGHT DATA FOR MALE STUDENTS

Student Name	Class (sophomore, junior, etc.)	Student’s Height (inches)	Father's Height (inches)

Table 10.2. Male students and fathers’ heights.

EXERCISES

1. In each of the following situations, tell whether you would be interested simply in exploring the relationship between the two variables or whether you would want to view one of the variables as an explanatory variable and the other as a response variable. In the latter case, state which variable is the explanatory variable and which is the response variable.

a. The amount of time spent studying for a statistics exam and the grade on the exam.

b. The weight and height of a person.

c. The amount of yearly rainfall and the yield of a crop.

d. Hand length and foot length of a person.

2. A study was conducted on mercury (Hg) concentrations in fish taken from Lake Natoma in California. The researchers were concerned that mercury concentration levels in sample fish tissue might differ depending on the lab testing the fish. Fish tissue samples from 10 largemouth bass were sent to two labs, Columbia Environmental Research Center (CERC) and University of California, Davis (UC Davis), for inter-laboratory comparison. The data appear in Table 10.3. Mercury concentration is measured in micrograms of mercury per gram of fish tissue (dry weight).

CERC Hg (μg/g dry wt.) *x*	UC Davis Hg (μg/g dry wt.) *y*
2.76	2.67
2.31	2.35
1.75	1.63
1.27	1.19
3.66	3.54
0.88	0.91
1.10	1.04
0.66	0.68
2.00	2.05
3.24	3.13

Table 10.3. Mercury concentration in fish as determined by two labs.

a. Make a scatterplot of the data in Table 10.3.

b. Is the relationship between mercury concentrations determined by the two labs an example of positive association or negative association or neither. Explain.

c. Does the relationship have linear or nonlinear form? Explain.

3. Satellites are one of the many tools used for predicting flash floods, heavy rainfall, and large amounts of snow. Geostationary Operational Environmental Satellites (GOES) collect data on cloud top brightness temperatures (measured in degrees Kelvin (°K)). It turns out that colder cloud temperatures are associated with higher and thicker clouds, which could be associated with heavier precipitation. Data consisting of cloud top temperature measured by a GOES satellite and rainfall rate measured by ground radar appear in Table 10.4. Because ground radar can be limited by location and obstructions, having an alternative for predicting the rainfall rates can be useful.

Temperature (°K)	Radar Rain Rate (mm/h)	Temperature (°K)	Radar Rain Rate (mm/h)
195	150	203	44
196	150	204	39
197	150	205	39
198	118	206	35
199	109	207	38
200	95	208	31
201	63	209	20
202	66	210	24

Table 10.4. Sixteen data pairs of (temperature, rain rate) data.

a. In this situation, which variable is the explanatory variable and which is the response variable? Explain your choice.

b. Make a scatterplot of the data in Table 10.4.

c. Is the association between the variables positive or negative? Does the pattern of the dots in your scatterplot appear roughly linear? Explain.

4. The video discussed the linkage between powerboat registrations in Florida and manatee deaths. Table 10.5 contains data from 1977 to 2011 on the number of manatee deaths and the number of powerboat registrations in Florida. A time-series graph is a scatterplot where time is on the horizontal axis and the variable being measured is on the vertical axis.

a. Make a time-series graph of yearly manatee deaths. Would you describe the pattern as linear or nonlinear? Explain.

b. Make a time-series graph of the yearly number of powerboat registrations in Florida. Would you describe the pattern as linear or nonlinear? Explain.

Year	Powerboat Registrations (thousands)	Manatee Deaths
1977	447	13
1978	460	21
1979	481	24
1980	498	16
1981	513	24
1982	512	20
1983	526	15
1984	559	34
1985	585	33
1986	614	33
1987	645	39
1988	675	43
1989	711	50
1990	719	47
1991	681	53
1992	679	38
1993	678	35
1994	696	49
1995	713	42
1996	732	60
1997	755	54
1998	809	66
1999	830	82
2000	880	78
2001	944	81
2002	962	95
2003	978	73
2004	983	69
2005	1010	79
2006	1024	92
2007	1027	73
2008	1010	90
2009	982	97
2010	942	83
2011	922	88

Table 10.5. Manatee data for years 1977 – 2011.

REVIEW QUESTIONS

Team	Attendance	Price	Team	Attendance	Price
New York Knicks	18,234	117.50	Cleveland Cavaliers	16,487	48.62
Los Angeles Lakers	19,012	99.25	Orlando Magic	17,306	43.65
Miami Heat	19,399	67.00	New Jersey Nets	17,445	37.06
Chicago Bulls	20,007	68.37	Houston Rockets	17,148	41.00
Boston Celtics	18,252	68.55	Utah Jazz	17,302	42.10
Phoenix Suns	15,809	60.63	Philadelphia 76ers	16,797	39.25
Los Angeles Clippers	18,475	51.47	Detroit Pistons	14,695	41.26
San Antonio Spurs	17,999	58.45	Golden State Warriors	17,752	34.13
Toronto Raptors	16,926	46.98	Atlanta Hawks	15,815	36.13
Dallas Mavericks	18,686	49.45	Minnesota Timberwolves	17,407	34.50
Portland Trail Blazers	17,615	48.40	New Orleans Hornets	15,577	30.49
Denver Nuggets	17,191	47.30	Washington Wizards	15,636	23.64
Milwaukee Bucks	15,033	46.00	Indiana Pacers	15,258	30.59
Oklahoma City Thunder	17,962	47.15	Charlotte Bobcats	15,757	29.27
Sacramento Kings	15,299	48.17	Memphis Grizzlies	16,493	22.95

Table 10.6. NBA attendance and ticket prices 2012/13 Season

1. Table 10.6 gives the average attendance and average ticket price for each of the 30 teams in the National Basketball Association in the 2012-2013 season. We want to investigate the relationship between ticket prices and attendance.

a. Give some reasons why you might expect a positive association between ticket prices and attendance. Then give some reasons why you might expect a negative association.

b. Make a scatterplot of the data to display the relationship between price and attendance. Explain your choice about what graph to make. Describe the relationship in words. Are there any outliers?

2. Table 10.7 contains data on the number of doctors in the United States (in thousands) for the 30-year period from 1970 to 1999.

Year	*x,* Years Since 1970	Total Number of Physicians (thousands)	Number of Women Physicians (thousands)
1970		330	25
1975		390	36
1980		470	55
1985		550	80
1990		620	100
1995		720	150
1999		800	190

Table 10.7. Number of physicians by year.

a. Notice that the values for *x* in the second column, the number of years since 1970, are missing. Complete this column.

b. Make a scatterplot of the total number of physicians versus *x*. (We are using *x* as the explanatory variable.) Leave room to extend the horizontal axis out to 40 (for the year 2010).

c. Does your scatterplot from (b) have linear form? If not, describe the nonlinear nature of the scatterplot.

d. Make a scatterplot of the number of women physicians versus *x*. Leave room to extend the horizontal axis out to 40 (for the year 2010).

e. Look at your scatterplot from (d). Does your scatterplot from (d) have linear form? If not, describe the nonlinear nature of the scatterplot.

f. In 2010 ($x = 40$), the total number of physicians was 850,000 and the total number of female physicians was 250,000. Add this information to your scatterplots in (b) and (d). Does the added data point fit the overall pattern of the original data or deviate from the overall pattern? Explain.

3. Some students are good in mathematics and others are better at writing. So, the question is whether there is any relationship between ability in math and ability in writing. The SAT, a standardized test for college admissions that is widely used in the United States, has both a Math and a Writing section. Table 10.8 contains Math and Writing SAT scores for 20 randomly chosen students accepted by a university.

Math SAT, *y*	Writing SAT, *x*	Math SAT, *y*	Writing SAT, *x*
440	410	680	580
550	570	440	430
520	520	440	450
420	470	390	430
550	620	460	600
650	560	460	520
610	620	520	570
610	520	540	530
340	470	420	430
600	540	550	480

Table 10.8. Math and Writing SAT scores.

a. Make a scatterplot of the Writing and Math SAT scores. (Use Writing SAT for the horizontal axis.)

b. Based on your scatterplot, does there appear to be a relationship between Writing and Math SAT scores? If so, describe the relationship.