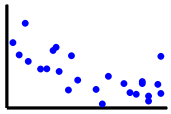


TOPIC D6: RELATIONSHIPS BETWEEN QUANTITATIVE VARIABLES

SPOTLIGHT: MEASURING CLIMATE

One reason why people relocate to a different city is climate. The book *Cities Ranked & Rated* presents data on the key components of climate – temperature, precipitation, cloud cover, humidity, and hazards – that people may use in deciding on a place to live. We will be using some of the data from this book to explore the relationships between different weather measurements for a group of cities.

Climate actually refers to the physical characteristics of a location that will affect the weather that we observe. What factors determine climate of a metropolitan area? One obvious factor is a place's latitude or north-south location. Generally places farther south tend to be warmer, and those further north are colder and have greater seasonal changes. The altitude of an area can have a large effect on climate. A location's higher altitude means less dense air and generally less humidity – this means less oxygen which places greater strain on the human circulatory system. The availability of nearby water (a lake or an ocean) can have a significant impact on climate. Water helps to moderate a place's temperature, but the water's moisture can affect local precipitation such as "lake effect" snows from Lake Erie or Lake Michigan. Wind direction can also affect climate – much of the climate in the United States is governed by the west to east movement of air across the middle latitudes of North America. Landforms, such as mountains and valleys, can have a large impact on climate. For example, mountain ranges can block winds, creating a drier, less humid climate. Also storm tracks affect the climate of cities along their path. Cities located near common storm tracks will experience greater swings in weather and strong storms.

The climate or physical characteristics of cities actually are fixed. But the weather we observe depends on the interaction of climate factors such as latitude, winds, and storms, and can exhibit considerable variation. *Cities Ranked & Rated* measures

weather for a location in different ways. The average minimum temperature in January and the average maximum temperature in July are recorded – this gives a person an idea about the temperature range for a city. In addition, the book presents the average number of days where the high temperature exceeds 90 or the low temperature is below freezing. Precipitation is measured both by the annual inches of rain and snow combined and the number of inches of snowfall. The average number of days in the year with at least some measurable rain or snow is recorded. The July relative humidity is measured for a city – this is the moisture content of air relative to temperature and greater humidity usually refers to less comfort. Other measures of comfort are made, including the average number of mostly sunny days, and a score measuring the risk of tornados and hurricanes.

PREVIEW

Often we collect more than one variable from each individual in a dataset. We do so because we are interested in studying the relationship between the variables. When we do this, we typically have

- *a response variable* – a variable that we are mainly interested in
- *an explanatory variable* – a variable that we think might be helpful in explaining some of the variation in the response variable

In this topic we describe some general ways of looking at the relationship between two variables.

In this topic your learning objectives are to:

- Understand how a pattern in a scatterplot tells us about the direction and the strength of a relationship.
- Understand what a QCR and a correlation coefficient tell us about the relationship between two quantitative variables.

NCTM Standards

✓In Grades 6-8, all students should select, create, and use appropriate graphical representations of data, including scatterplots.

✓In Grades 6-8, all students should make conjectures about possible relationships between two characteristics of a sample on the basis of scatterplots of the data.

✓In Grades 9-12, all students should, for bivariate measurement data, be able to determine correlation coefficients using technological tools.

RELATIONSHIPS - SCATTERPLOTS

In many situations, we will collect two or more measurements from individuals. For example, we might collect the pulse rate and weight from a number of students, or we might collect different weather measurements from different cities. In this situation, we are often interested in the relationships between the measurements. We will study relationships by

- graphing the data to get a picture of the association between two measurements
- computing a summary value, called a correlation coefficient, to describe the strength of the relationship between the variables
- drawing a "best line" of fit to describe how one variable changes as a function of the other variable

Weather data

From the usatoday.com web site, there is a weather section that describes the climate of different cities in the United States. For ten cities, the below table displays the average high daily temperatures (in degrees Fahrenheit) in January and July, the average amount of precipitation (inches of rain or melted snow) in January and July, the average dew point (a humidity measure in Fahrenheit degrees) in January and July, and the latitude in degrees. Here the data units are the cities, and we are recording seven

measurements for each data unit. Note that there are several blank values in the table – the dew point was not given for two cities.

City	January temp	July temp	Jan precip	July precip	Jan dew	July Dew	Latitude
Aberdeen	20	85	0.5	3	3	59	45
Akron	33	83	2.7	4	19	61	40
Albuquerque	47	92	1.4	1.3	18	49	35
Amarillo	49	91	0.5	2.8	19	58	35
Aspen	34	80	2	1.5			39
Atlanta	52	89	4.7	5.3	32	68	33
Bakersfield	57	98	1	0	39	51	35
Bar Harbor	33	77	4.7	3.3			44
Chicago	29	84	1.7	3.6	14	62	41
Miami	76	89	2	6	58	73	25

Let's focus on two variables

- the average high daily temperature in January (called Jan temp)
- the average high daily temperature in July (called July temp)

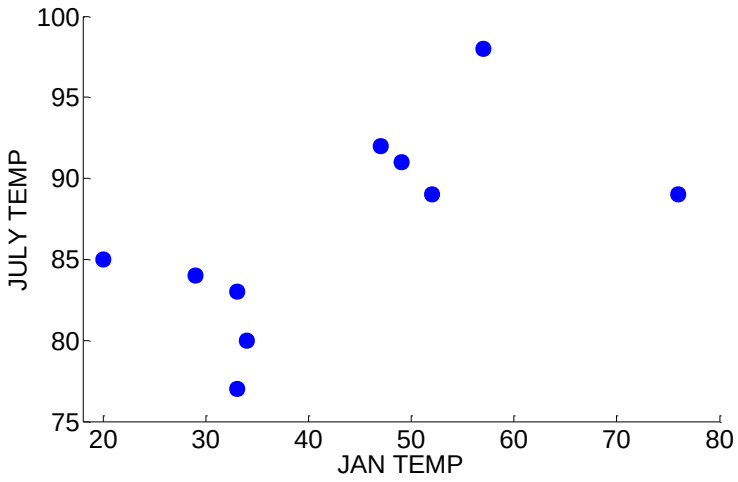
Do you think these two variables are associated? If I told you that the average daily high temperature in January of an unknown city was 50 degrees, would this give any information about the average daily high temperature of the city in July? I think the answer should be "yes." If the average high temp in January is 50 degrees, I think that it is a city in a warm climate (at least, warmer than Ohio), and so I would also expect the average high temperature in July also to be high.

Scatterplot

How can we study the relationship between Jan temp and July temp?

We begin by constructing a graph called a scatterplot. This is a generalization of the dotplot where you are plotting points along two axes.

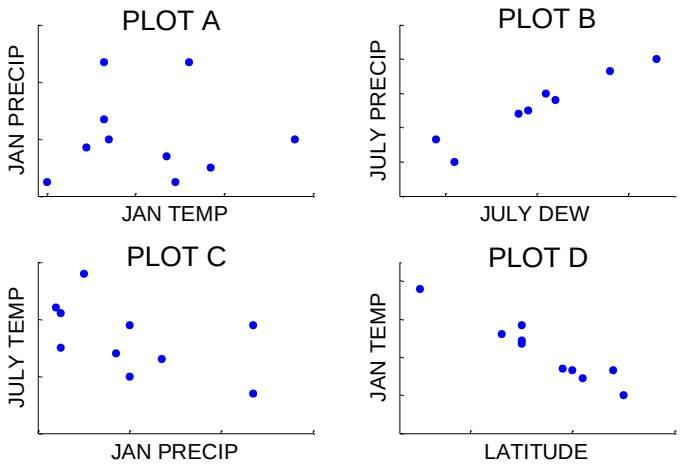
First we construct a Cartesian grid, where the values of one variable (here Jan temp) are along one axis and the values of the second variable (July temp) are along the second axis. Then for each city we place a dot corresponding to the values of the two variables. If we do this plotting for all 10 cities, we get the following scatterplot.



Scatterplot Patterns

We detect association between a pair of variables by finding a pattern in this scatterplot. What type of patterns are we looking for?

In the figure below, we show four scatterplots labeled PLOT A, PLOT B, PLOT C, PLOT D.



1. First, look at PLOT A that graphs the January temperature against the January precipitation. As we look at the points, we don't detect any general drift in the point (either upward or downward) as you scan the points from left to right. The conclusion is

that there is at most a weak relationship between a city's January temperature and its January precipitation. If we are told a city's temperature in January, this gives us little information about that city's precipitation that month.

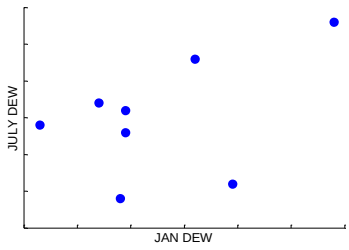
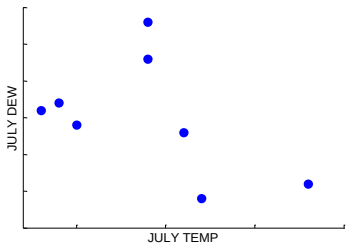
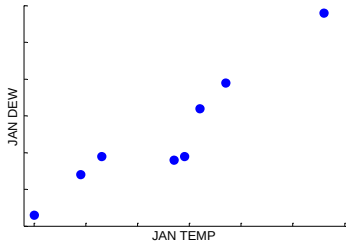
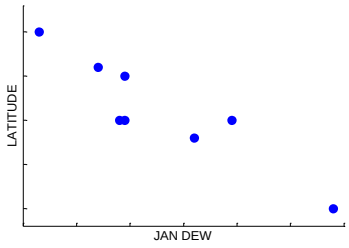
2. In PLOT B we see the points drift from the lower-left to the upper-right sections of the plot. This indicates that the two variables July dew and July precipitation are *positively associated*. Cities with small dew points in July tend to be associated with small July precipitation values, and large values of July dew and large values of the July precipitation go together. If you know anything about weather, this makes sense – cities with high humidity tend to have more rain.

3. PLOT C illustrates *negative association* between the two variables. The points tend to drift from the upper-left to the lower-right sections of the plot. This means that large values of January precipitation tend to be associated with small values of July temperature. Likewise small values of January precipitation go together with large values of the July temperature.

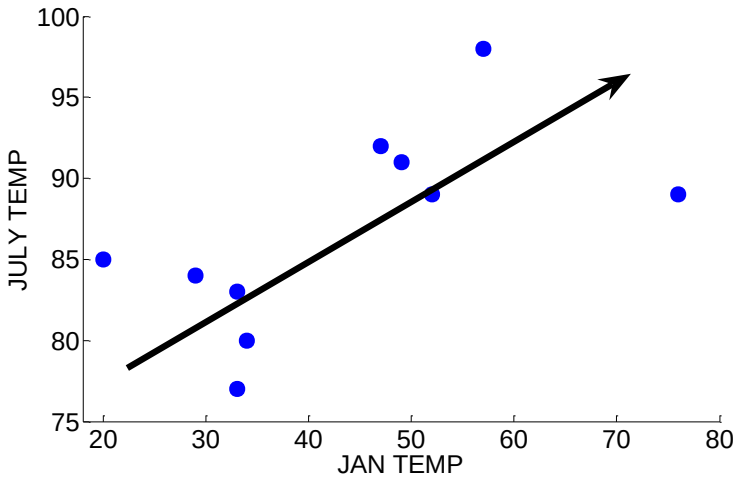
4. In PLOT D, we see that the variables latitude and January temperature are also negatively associated since the graph has the same downward drift pattern as PLOT C. But note that the points in PLOT D are more clustered about a line than the points in PLOT C. This tells us that latitude and January temperature of cities have a stronger association than January precipitation and July temperature.

Generally, when we look at a scatterplot, we identify both the *direction* and the *strength* of the association. Using our weather data, we construct four scatterplots in the figure below that illustrate different types of direction and strength of relationships. All four graphs illustrate some association between the weather variables. The left two graphs illustrate positive relationships where small (large) values of one variable are associated with small (large) values of the second variable. The right two graphs demonstrate negative relationships where small (large) values of the first variable are associated with large (small) values of the second variable. The bottom two graphs demonstrate stronger relationships where the points tend to follow a line.

	DIRECTION	
STRENGTH	Positive Association	Negative Association

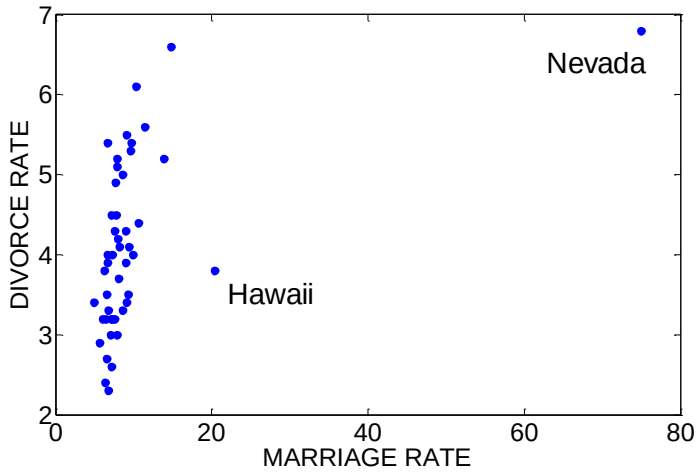
Weak		
Strong		

Let's return to our example where we constructed a scatterplot of the average January high daily temperature and the average July high daily temperature for a group of cities. What type of association do we see in this plot?



I hope you agree that the points have a positive drift, indicating that the January high temperature and the July high temperature are positively associated. This makes sense. Cities that are unusually cold in January (like Aberdeen and Chicago) tend also to be cooler in July; likewise, warm January temperatures (like those in Bakersfield and Miami) tend to be associated with warm July temperatures.

When we were describing a distribution of a single batch in topic D2, we were interested in the general shape of the data and also in observations that were far away from the general pattern. Similarly, when we look at scatterplots, we may observe outlying points that don't agree with the general pattern. For example, suppose we are interested in the relationship between a state's marriage rate (the number of marriage per 1000 people) and its divorce rate (the number of divorces for each 1000 people). We collect the marriage and divorce rates for all 50 states from 2001 data and construct a scatterplot shown below.



Generally there is a strong positive association in this graph – states with high marriage rates also tend to have high divorce rates. But there are two points corresponding to the states Hawaii and Nevada that don't follow the general pattern. Since these points seem special, it is helpful to label them in the scatterplot. Nevada is a very popular place for couples to visit to acquire a quick wedding certificate and Hawaii is a popular place to get married due to its nice climate. For both states, we observe a much higher marriage rate than one would expect based on the data from the remaining states. In practice, it is important to identify both the general relationship between two quantitative variables and the particular observations like Hawaii and Nevada that don't follow the general relationship pattern.

PRACTICE: INTERPRETING SCATTERPLOTS

A number of measurements were made on 38 1978-79 model automobiles. For each car, the variables measured were

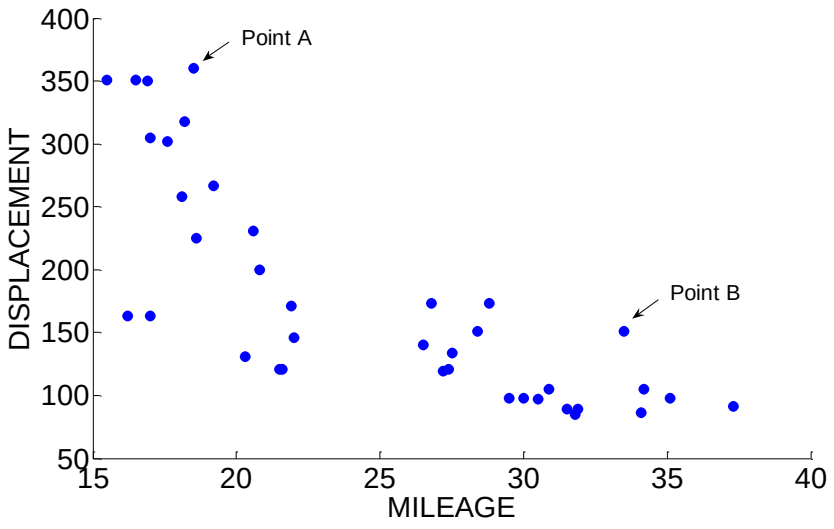
- the mileage in miles per gallon (MPG)
- the weight in thousands of pounds
- the drive ratio
- the horsepower
- the displacement of the car in cubic inches

A table of this data is shown below.

Car	MPG	Weight	Drive_Ratio	Horsepower	Displacement
Buick_Estate_Wagon	16.9	4.36	2.73	155	350
Ford_Country_Squire_Wagon	15.5	4.054	2.26	142	351
Chevy_Malibu_Wagon	19.2	3.605	2.56	125	267
Chrysler_LeBaron_Wagon	18.5	3.94	2.45	150	360
Chevette	30	2.155	3.7	68	98
Toyota_Corolla	27.5	2.56	3.05	95	134
Datsun_510	27.2	2.3	3.54	97	119
Dodge_Omni	30.9	2.23	3.37	75	105
Audi_5000	20.3	2.83	3.9	103	131
Volvo_240_GL	17	3.14	3.5	125	163
Saab_99_GLE	21.6	2.795	3.77	115	121
Peugeot_694_SL	16.2	3.41	3.58	133	163
Buick_Century_Special	20.6	3.38	2.73	105	231
Mercury_Zephyr	20.8	3.07	3.08	85	200
Dodge_Aspen	18.6	3.62	2.71	110	225
AMC_Concord_D/L	18.1	3.41	2.73	120	258
Chevy_Caprice_Classic	17	3.84	2.41	130	305
Ford_LTD	17.6	3.725	2.26	129	302
Mercury_Grand_Marquis	16.5	3.955	2.26	138	351
Dodge_St_Regis	18.2	3.83	2.45	135	318
Ford_Mustang_4	26.5	2.585	3.08	88	140

Ford_Mustang_Ghia	21.9	2.91	3.08	109	171
Mazda_GLC	34.1	1.975	3.73	65	86
Dodge_Colt	35.1	1.915	2.97	80	98
AMC_Spirit	27.4	2.67	3.08	80	121
VW_Scirocco	31.5	1.99	3.78	71	89
Honda_Accord_LX	29.5	2.135	3.05	68	98
Buick_Skylark	28.4	2.67	2.53	90	151
Chevy_Citation	28.8	2.595	2.69	115	173
Olds_Omega	26.8	2.7	2.84	115	173
Pontiac_Phoenix	33.5	2.556	2.69	90	151
Plymouth_Horizon	34.2	2.2	3.37	70	105
Datsun_210	31.8	2.02	3.7	65	85
Fiat_Strada	37.3	2.13	3.1	69	91
VW_Dasher	30.5	2.19	3.7	78	97
Datsun_810	22	2.815	3.7	97	146
BMW_320i	21.5	2.6	3.64	110	121
VW_Rabbit	31.9	1.925	3.78	71	89

A scatterplot of DISPLACEMENT and MILEAGE is shown below.

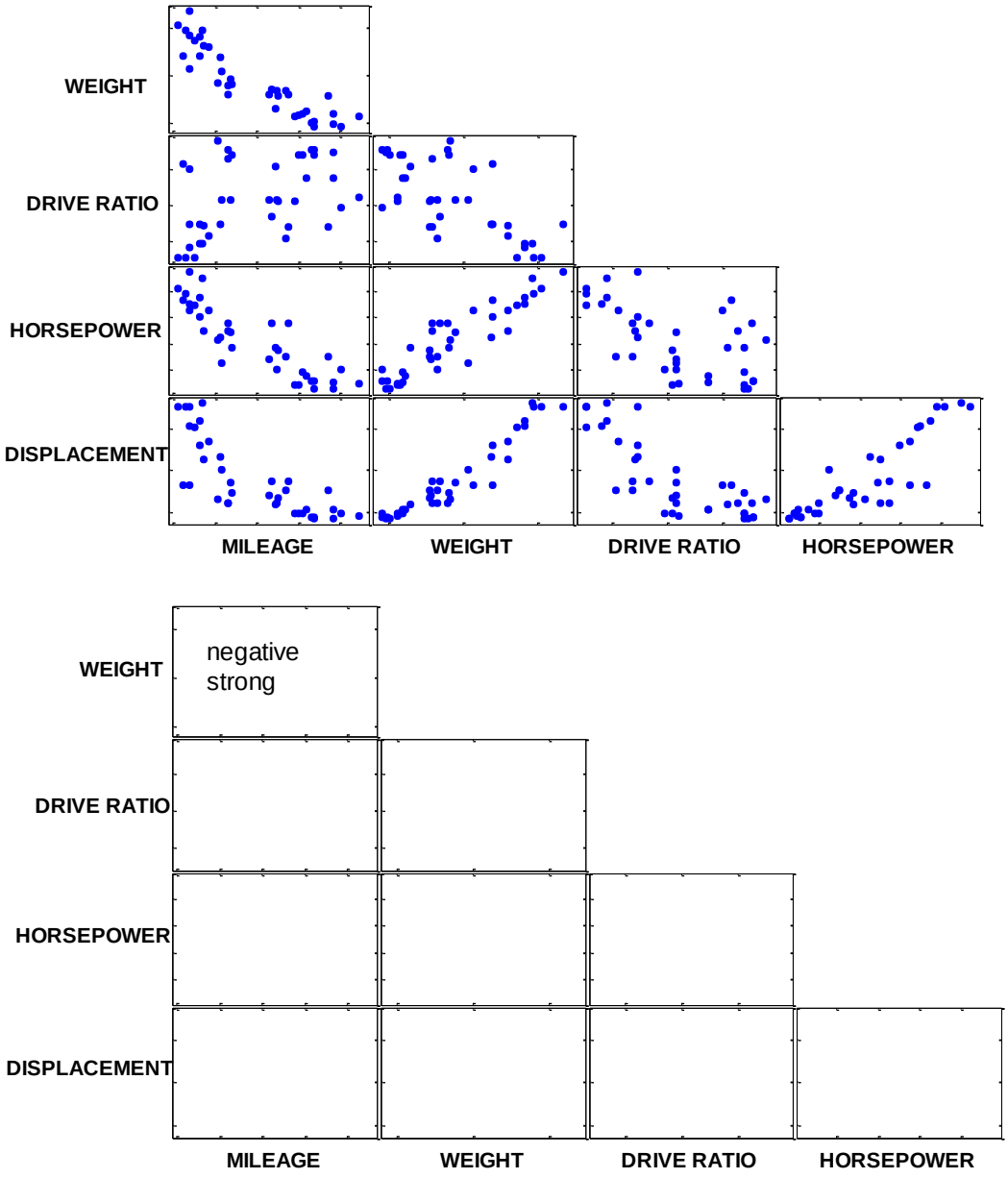


1. Identify the cars corresponding to Point A and Point B in the scatterplot.

2. What is the general pattern of the scatterplot? Is there a relationship between the displacement of the car and its mileage? Describe to a layman what this means?

3. Circle one point in the scatterplot corresponding to a car that has a small displacement and a small mileage. Is this car unusual with respect to the general pattern in the scatterplot that you described in question 2?

4. The following figure displays scatterplots of all 10 possible pairs of variables among MILEAGE, WEIGHT, DRIVE RATIO, HORSEPOWER and DISPLACEMENT. For each scatterplot, describe the general pattern (negative, positive, or little) and if the pattern indicates a positive or negative association, state if the association is strong or weak. Place your answers in the empty boxes below. (The first box has been completed for you.)

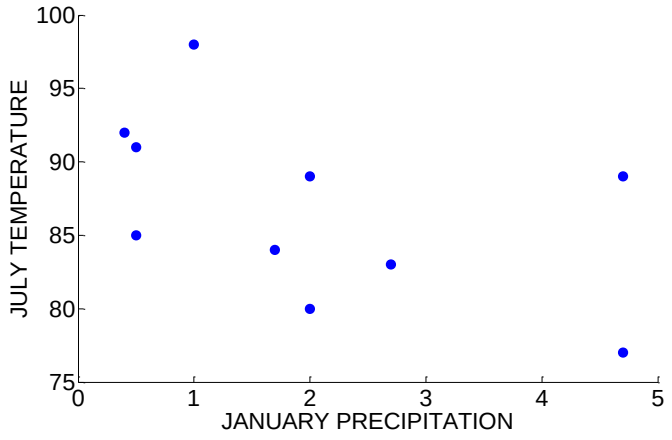


INTRODUCTION – LOOKING AT WEATHER DATA

Let's return to our weather dataset that describes the climate of different cities in the United States. We focus on the average precipitation in January (labeled "Jan precip") and the average high temperature in July (labeled "July temp") for 10 cities.

City	Jan precip	July temp
Aberdeen	0.5	85
Akron	2.7	83
Albuquerque	0.4	92
Amarillo	0.5	91
Aspen	2	80
Atlanta	4.7	89
Bakersfield	1	98
Bar Harbor	4.7	77
Chicago	1.7	84
Miami	2	89

We already learned in this topic that a good first step in exploring the relationship between a city's January precipitation and its July temperature is to construct a scatterplot.



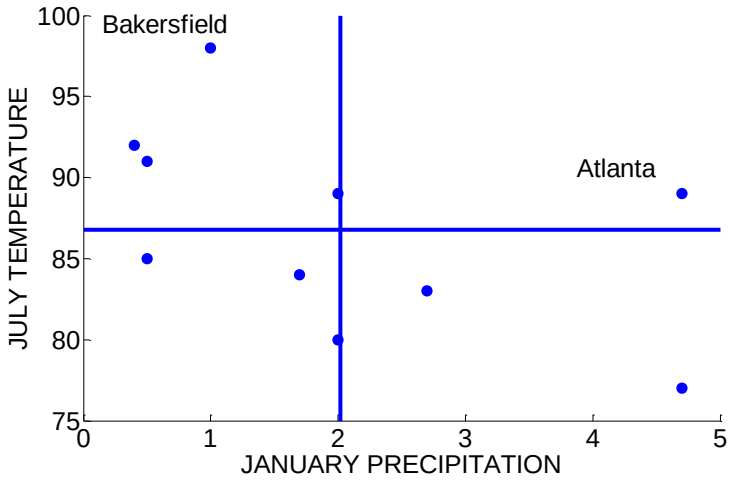
We see a negative trend in this display. This means that cities that have low precipitation in January tend to have high temperatures in July. Also, high precipitation in January and low July temperatures tend to go together.

We can summarize the relationship that we see in this scatterplot in a couple of ways.

1. First, we want to describe the *strength* of the association that we see in the plot by using a number called a *correlation coefficient*. We will describe how to compute and interpret this measure of association.
2. Second, a useful way of describing the positive association is by fitting a line to the plot. We will describe two ways of fitting a "best line" to the points called a *least-squares line* and a *median-median line*, and discuss how we can use these lines to predict the value of one variable knowing the value of the second variable. We'll talk more about best fitting lines in the second half of this topic.

A SIMPLE CORRELATION FORMULA – THE QCR

To motivate a simple formula for measuring a relationship, suppose that we divide the scatterplot into four regions by drawing horizontal and vertical lines at the respective means of the two variables. The mean of the January precipitation values is 2.02 inches and we draw a vertical line at this value. Also, we draw a horizontal line at the value 86.8 degrees (the mean July temperature value).



We have labeled two points that correspond to the January precipitation and July temperature for the cities Atlanta and Bakersfield. Atlanta has an above-average precipitation in January and an above-average temperature in July and therefore this point

is to the right of the vertical line and above the horizontal line. Bakersfield, in contrast, is left of the vertical line and above the horizontal line, which means, respectively, that this city has a below-average January precipitation and an above-average July temperature.

Note that most of the points fall in the upper left and lower right sections of the plot. In these regions, the cities either have above-average precipitations and below-average temperatures, or below-average precipitations and above-average temperatures. A simple of measure of association finds the number of points in the upper right and lower left sections (quadrants I and III), subtracts the number of points in the upper left and lower right sections (quadrants II and IV), and divides the result by the number of points (n). This simple measure, called the Quadrant Count Ratio or QCR for short, is defined as

$$QCR = \frac{(\#\ of\ \ points\ in\ Quadrants\ I\ and\ III) - (\#\ of\ \ points\ in\ Quadrants\ II\ and\ IV)}{n}.$$

In our example, we count 1 point in Quadrant I, 4 points in Quadrant II, 3 points in Quadrant III, and 2 points in Quadrant IV. Also there are n = 10 points in our graph. So the Quadrant Count Ratio is given by

$$QCR = \frac{(1+3)-(4+2)}{10} = -0.2.$$

The QCR has some attractive properties as a measure of association. If all of the points are in quadrants I and III, then QCR = +1, and if all of the points are in quadrants II and IV, then QCR = -1. If most of the points fall in quadrants I and III, the measure of association will be positive. Similarly, if the points generally fall in quadrants II and IV, then the association will be negative.

In this example, the QCR is negative, reflecting a negative association between a city's January precipitation and its July temperature.

THE CORRELATION COEFFICIENT

A second, more traditional way of measuring the relationship between two quantitative variables is by means of a correlation coefficient. As in the QCR, we motivate this measure of relationship by dividing the scatterplot into four regions by drawing horizontal and vertical lines at the means of the two variables. For a correlation, we use more information that simply the number of points in the four regions.

We compute a correlation coefficient in two steps:

STEP 1: The correlation can be shown to only depend on the variables though their standardized scores introduced in topic D4. So we first find two standardized scores or z-scores corresponding to the two variables for each city. For each January precipitation value, we standardize it by subtracting its mean and dividing by its standard deviation. Likewise, for each July temperature value, we standardize by subtracting its mean and dividing by its standard deviation.

The formulas for the two standardized scores, z_x and z_y are respectively

$$z_x = \frac{January\ precip - mean(January\ precip)}{std\ dev(January\ precip)}$$

$$z_y = \frac{July\ temp - mean(July\ temp)}{std\ dev(July\ temp)}$$

We found the standardized scores for all cities and placed them in the below table. Let's check the calculations for Atlanta. The mean January precipitation (across cities) was 2.02 inches and the standard deviation was 1.61 inches; for the July temperatures the mean and standard deviation are 86.8 and 6.21 degrees, respectively. Atlanta's January precipitation was 4.7 inches, so its standardized score was

$$z_x = \frac{4.7 - 2.02}{1.61} = 1.66.$$

-- this means that Atlanta's precipitation in January was approximately 1.66 standard deviations above the mean. Atlanta's July temperature was 89 degrees and the corresponding standardized score is

$$z_y = \frac{89 - 86.8}{6.21} = 0.35.$$

Likewise, Atlanta's July temperature was about a third of a standard deviation higher than the mean.

STEP 2. After we find the two standardized scores for each city, we take the products of the standardized scores – these products are placed in the "Product" column of the table.

City	Jan precip	July temp	z_x	z_y	Product
Aberdeen	0.5	85	-0.94	-0.29	0.27
Akron	2.7	83	0.42	-0.61	-0.26
Albuquerque	0.4	92	-1.01	0.84	-0.85
Amarillo	0.5	91	-0.94	0.68	-0.64
Aspen	2	80	-0.01	-1.1	0.01
Atlanta	4.7	89	1.66	0.35	0.58
Bakersfield	1	98	-0.63	1.8	-1.13
Bar Harbor	4.7	77	1.66	-1.58	-2.62
Chicago	1.7	84	-0.2	-0.45	0.09
Miami	2	89	-0.01	0.35	0.00
SUM					-4.55

The correlation coefficient, denoted by r, is computed by dividing the sum of the products by the number of cases minus one. Or if you like formulas, r is given by

$$r = \frac{sum(z_x \times z_y)}{(number\ of\ \ cases) - 1}.$$

Here the correlation coefficient is given by

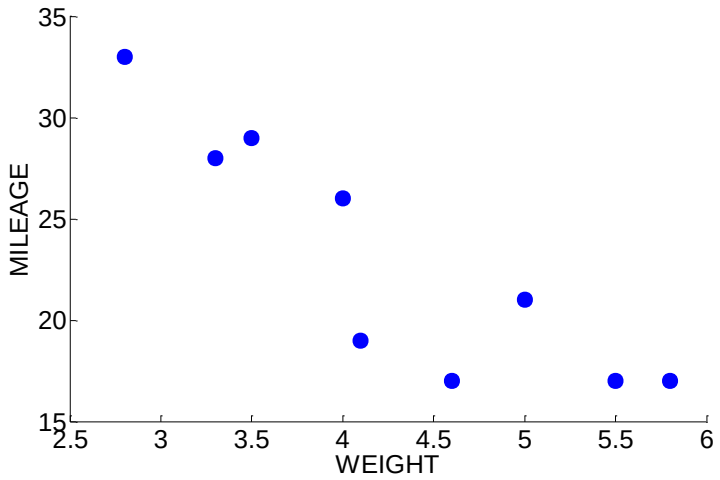
$$r = \frac{-4.55}{10-1} = -0.51$$

We'll discuss how to interpret this value shortly – we'll see that the negative value of r indicates that there is a negative association between the January precipitation and the July temperature.

PRACTICE: COMPUTING THE CORRELATION COEFFICIENT

The author collected the weight (in thousands of pounds) and the highway mileage for nine 2004 cars manufactured by Chevrolet. A table of the data is shown below together with a scatterplot of the two variables.

Car	Weight	Highway MPG	z_x	z_y	Product
Avalanche	5.8	17	1.48	-0.98	-1.45
Cavalier	2.8	33			
Corvette	3.3	28	-0.97	0.82	-0.8
Impala	3.5	29	-0.77	0.98	-0.75
S-10	4.1	19	-0.19	-0.66	0.13
Suburban	4.6	17			
Tahoe	5.5	17	1.19	-0.98	-1.17
TrailBlazer	5	21	0.7	-0.33	-0.23
Venture	4	26			



1. The mean and standard deviation of the car weights are given by 4.29 and 1.02 thousand pounds, respectively. The mean and standard deviation of the mileages are 23 and 6.1 mpg. Using these values, fill in the missing standardized scores in the table.

2. Draw vertical and horizontal lines on the scatterplot corresponding to the mean weight and the mean mileage of the cars. From looking at the scatterplot, compute the value of the QCR. What does this say about the relationship between car weight and car mileage?

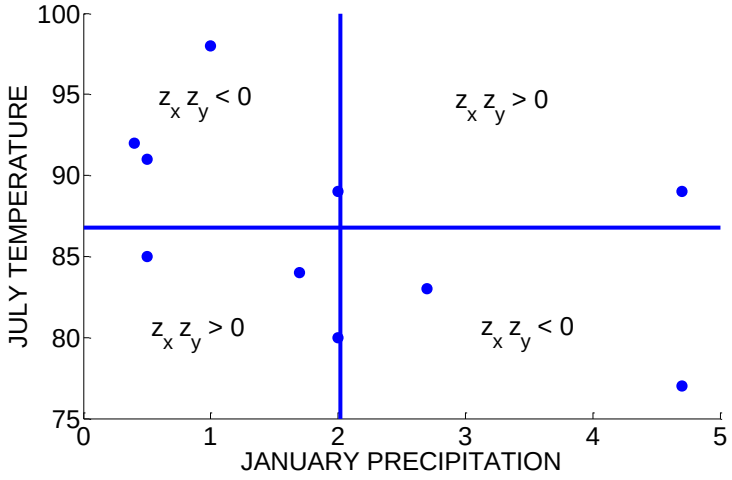
3. Compute the value of the correlation.

4. In this example, the pattern in the scatterplot was _____ and the sign of the correlation was _____ .

INTERPRETING THE CORRELATION COEFFICIENT

Okay, we compute the correlation (or more precisely, we have a computer compute a correlation for you). What does it mean? Here are some basic facts about the correlation coefficient r.

THE SIGN OF R: Recall that the numerator of r is the sum of the products of the standardized scores $z_x \times z_y$. This product will be positive in the upper right and lower left quadrants of the scatterplot, and the product will be negative in the two other quadrants. (See the below figure.) When the data primarily fall in the upper left and lower right sections, we will be adding up a lot of negative products, and r will be negative. This happened in our example. Similarly, when the data fall in the lower left and upper right sections, there will be many positive products of standardized scores, and r will be positive. So the sign of the correlation is informative about the general positive or negative pattern in the scatterplot.

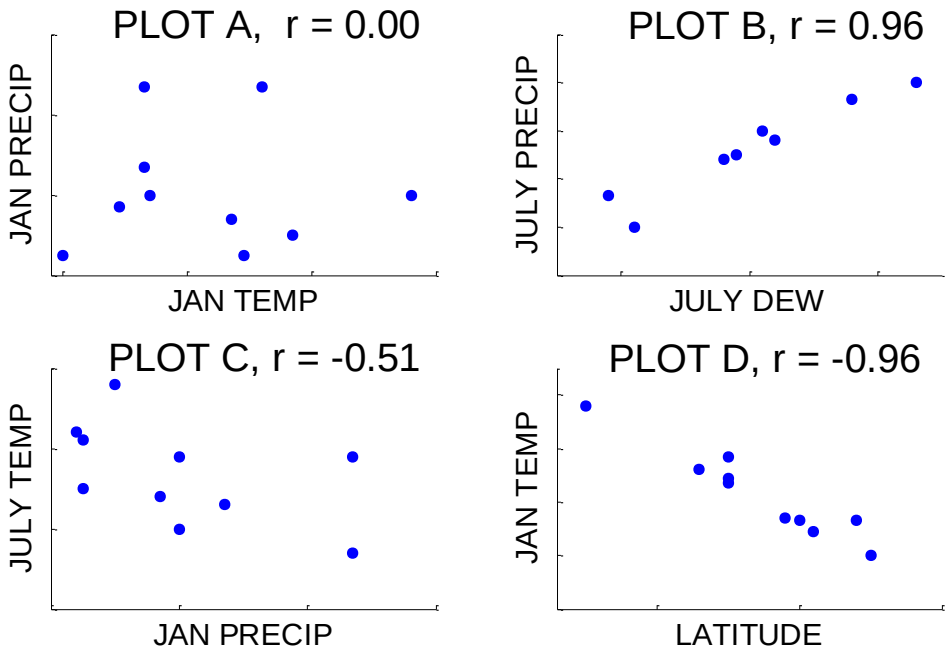


THE SIZE OF R: The value of the correlation coefficient will range from -1 to +1. If r is close to +1, then the points will fall closely to a line with positive slope. Likewise, a correlation value r close to -1 corresponds to a scatterplot where the points fall close to a line with negative slope. A correlation value close to zero means that there is little straight-line association in the graph.

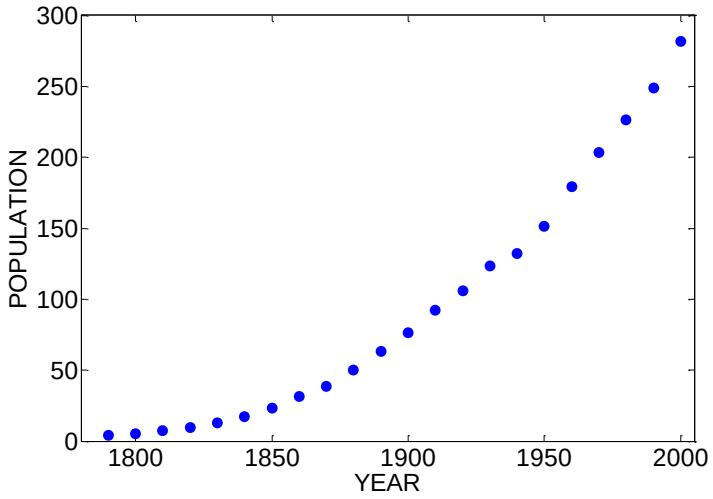
To illustrate the interpretation of r, let us revisit the weather example where we considered scatterplots of four pairs of the variables.

- In Plot A where there was little association between the January precipitation and the January temperature, the value of r is close to zero.

- In Plot B, there is a strong positive association, and r is close to one.
- Plots C and D both show negative associations between the corresponding variables and the values of r for both datasets are negative. But note that Plot D shows a stronger relationship than Plot C which is reflected in the corresponding values of the correlation. The variables latitude and January temperature have an association close to a straight line; the correlation value is r = -0.96.



To emphasize the second point, the correlation coefficient r measures the strength of a *straight-line relationship* between the two variables. It is possible that there is a strong relationship between two variables, but the relationship is not linear and so r is not the best measure of this association. For example, suppose we record the population of the United States for the years 1790, 1800, ..., 2000. Here is a scatterplot of the population against year:

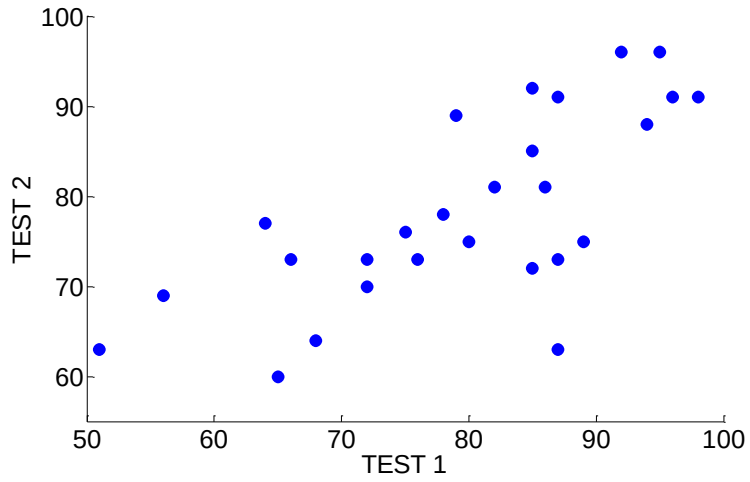


Is there a strong relationship between year and population? Yes, there is, but it isn't a straight-line relationship. In the early years, the population of the U.S. increased exponentially and this accounts for the significant curvature in the graph. The correlation coefficient for these early years is 0.959. This example illustrates several points:

- Although r measures the strength of a linear relationship, a high value of r need not mean that the variables are linearly associated.
- It is not sufficient to compute a correlation coefficient – one should graph the data *and* compute a correlation coefficient to understand the relationship between two variables.

PRACTICE – INTERPRETING CORRELATION

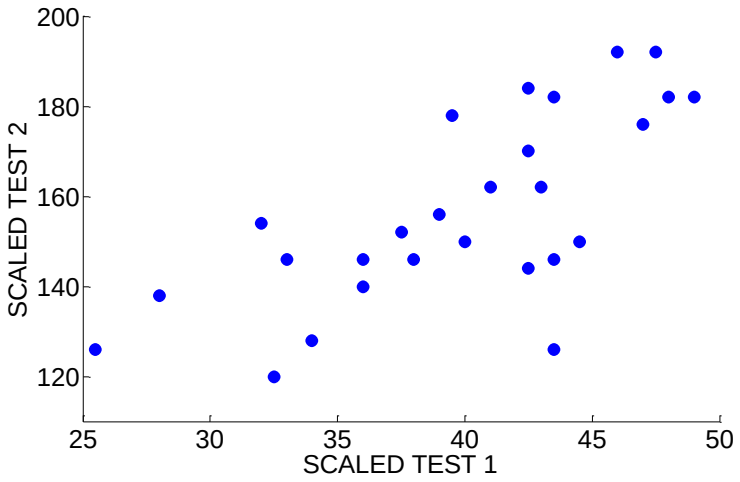
For a recent class, an instructor collected the student grades for two tests, each test scored on a scale from 0 to 100. A scatterplot of the TEST 1 score and the TEST 2 score is displayed below.



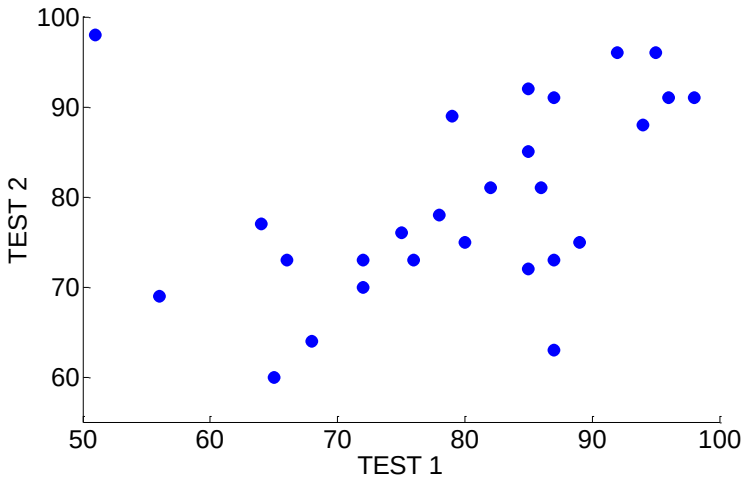
1. From looking at the pattern in the scatterplot, select the value of the correlation coefficient from the list of values below.

-.91 -.71 -.21 0 +.21 +.71 +.91

2. Suppose that the instructor decides on changing the scale of the two tests to give the tests different weights in the determination of a final grade. She decides on dividing the TEST 1 score by 2, so that the new TEST 1 range is 0 to 50, and multiplying the TEST 2 score by 2, so that the new TEST 2 range is 0 to 200. A scatterplot of the scaled TEST 1 and scaled TEST 2 scores is shown below. How would this rescaling affect the value of the correlation coefficient? Explain why there would be a change or no change in the value in r. (Hint: How does this rescaling affect the standardized scores of TEST1 and TEST2?)



3. Return to the original test scores TEST 1 and TEST 2. Suppose that there was a mistake in scoring TEST 2 for one student – his grade of 63 should have been 95. A scatterplot of the adjusted TEST 1 and TEST 2 scores is shown below. How will this change affect the value of the correlation r?

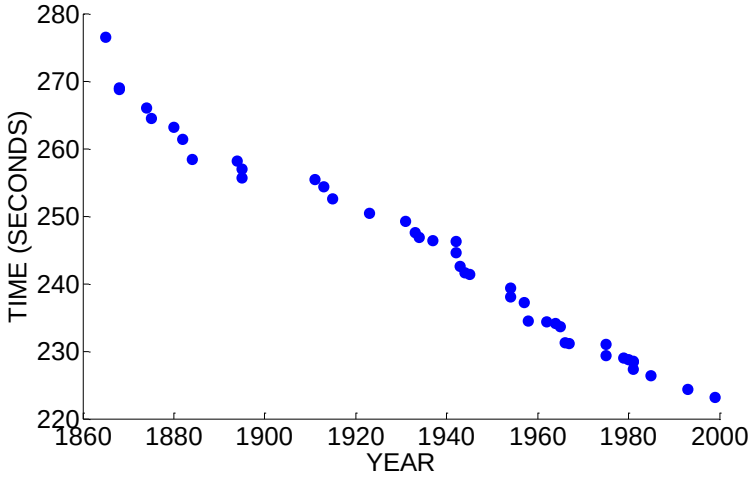


4. It turns out that the correlation value for the adjusted test scores is r = .41. (The correlation for the original test scores was r = .71.) What does this say about the sensitivity of the correlation coefficient to unusual values?

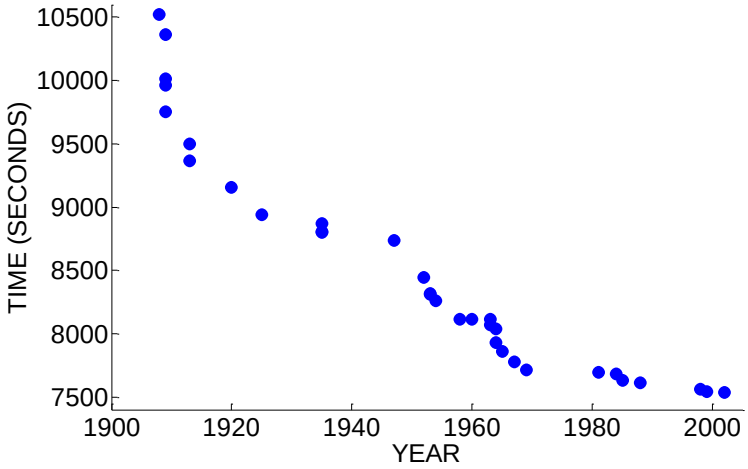
5. It is interesting to look at the pattern of record times for different athletic events. The first scatterplot below graphs the world record time in the men's mile race against the

year in which the record was set. The second scatterplot graphs the world record time in the men's marathon race and the year.

WORLD RECORD TIME FOR THE MILE



WORLD RECORD TIME FOR THE MARATHON



Which graph displays a stronger straight-line relationship? The correlation values for the two scatterplots are –0.95 and –0.99. Which value is associated with which scatterplot?

TECHNOLOGY ACTIVITY – GUESSING CORRELATIONS

DESCRIPTION: Here we will get some experience looking at scatterplots and guessing at the value of the correlation. Even if you know little initially about scatterplot patterns and the corresponding correlation values, you'll start understanding this connection when you see enough scatterplots.

PART 1: Guess the Correlation Game

For this part, you only need to know that a correlation is simply a measure of the association pattern in the scatterplot.

1. Open up the Fathom document "guess_correlation.ftm."
2. Select the scatterplot. Hit the Apple-Y key to simulate a new scatterplot.
3. Guess at the value of the correlation – put your guess in the "Guess at Correlation" column of the table.
4. Now scroll right to see the actual value of the correlation. Put this in the "Actual value of Correlation" column. Also compute your error – the distance between your guess and the actual value.
5. Scroll back (left) so you can see the scatterplot. Repeat steps 2, 3, 4 – keep going until you have done 20 guesses at the correlations for 20 scatterplots. When you are done, compute the TOTAL ERROR in your 20 guesses.

Trial	Guess at Correlation	Actual value of Correlation	Error
1			
2			
3			
4			
5			
6			
7			

8			
9			
10			
11			
12			
13			
14			
15			
16			
17			
18			
19			
20			
TOTAL			

PART 2: Using Fathom to analyze your correlation data.

To see how you did in this guessing activity,

- Open up a new Fathom document.
- Open up a new Collection and new Case Table.
- Create two variables, called GUESS and ACTUAL.
- Put the data from the table above into the Case Table.
- Construct a scatterplot of the GUESS, ACTUAL data.
- Compute the correlation between GUESS AND ACTUAL.
- Interpret what the value of this correlation is telling you.

WRAP-UP

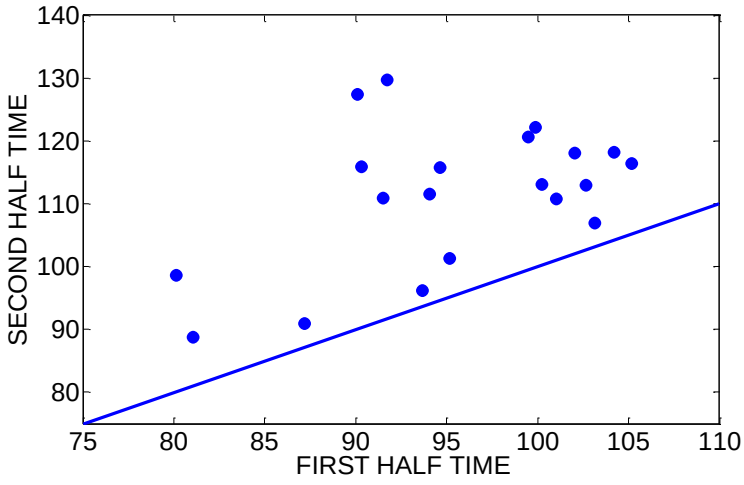
In this topic, we were introduced to different strategies for understanding the relationships between two quantitative variables. In studying this relationship, typically there is a *response variable* that is of main interest and an *explanatory variable* that we

believe is helpful in understanding the variation in the response variable. In this topic, we were introduced to the *direction* and *strength* of association in a scatterplot. The QCR and the correlation coefficient are statistics that measures the strength of the relationship between the variables. The correlation only measures the straight-line relationship, so one should also construct a scatterplot to check this straight-line assumption.

EXERCISES

1. **Boston Marathon Running Times**

A random selection of runners from the 2003 Boston Marathon was selected. For each runner, the time (in minutes) to run the first half of the race and the time to run the second half were recorded. The below figure displays a scatterplot of the first-half time and the second-half time.



a. Describe the pattern of association in the scatterplot. Explain why you see a relationship between these two variables.

b. The line FIRST HALF TIME = SECOND HALF TIME is drawn on the figure. Note that all of the points fall above this line. Can you explain why?

c. Circle one point (label it "S") where the runner got tired and took much longer to complete the second half than the first half.

d. Circle a second point (label it "F") where the runner took about the same time to complete both the first and second halves of the race.

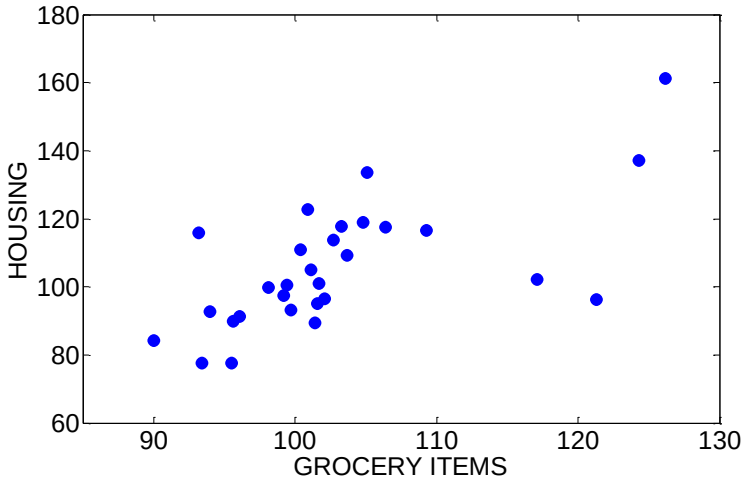
e. Circle the point corresponding to the runner that had the fastest total time in this race and circle the point corresponding to the runner with the slowest total time.

2. **Cost of Living Indices**

The ACCRA Cost of Living Index (www.costofliving.org) is a quarterly report that compares basic living expenses in various U.S. cities. The index measures the price level for consumer goods and services relative to the average, and a value of 100 represents an average price level. In a 2000 report, the index values for a number of American cities are listed. The table below lists the index values for 29 cities for the categories grocery items and housing.

City	Grocery items	Housing	City	Grocery items	Housing
Anchorage, Alaska	124.3	137.1	Manchester, N.H.	104.8	119.0
Phoenix, Ariz.	101.7	100.9	Albuquerque, N.M.	102.7	113.8
Sacramento, Calif.	121.3	96.2	Charlotte, N.C.	98.1	99.8
San Diego, Calif.	126.2	161.3	Cincinnati, Ohio	99.2	97.5
Colorado Springs, Colo.	103.3	117.7	Cleveland, Ohio	109.3	116.5
Jacksonville, Fla.	101.4	89.5	Oklahoma City, Okla.	95.5	77.7
Atlanta, Ga.	103.7	109.2	Salem, Ore.	100.4	111.0
Springfield, Ill.	99.7	93.1	Philadelphia, Pa.	105.1	133.6
New Orleans, La.	102.1	96.5	Memphis, Tenn.	95.6	89.9
Baltimore, Md.	94.0	92.6	Austin, Tex.	93.2	115.9
Lansing, Mich.	100.9	122.8	El Paso, Tex.	93.4	77.5
Minneapolis, Minn.	101.1	105.0	San Antonio, Tex.	90.0	84.2
Billings, Mont.	99.4	100.4	Salt Lake City, Utah	106.4	117.6
Omaha, Neb.	96.1	91.2	Cheyenne, Wyo.	101.6	95.1
Las Vegas, Nev.	117.1	102.2			

The figure below displays a scatterplot of these two index values.



a. Describe the general pattern of this scatterplot.

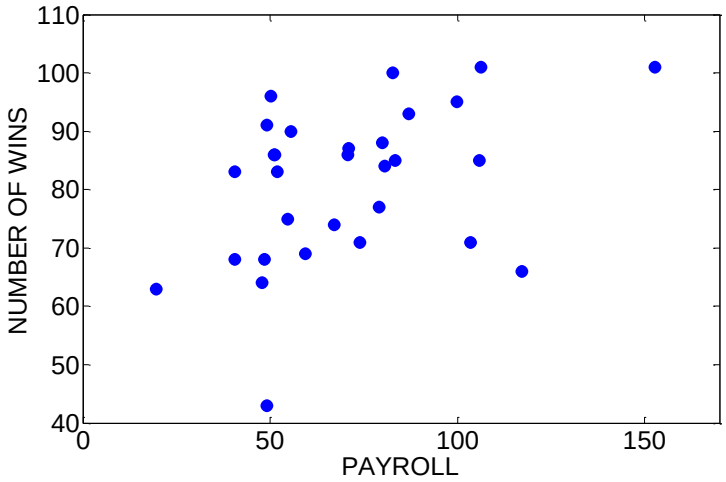
b. There are four points that seem to stand out from the main group. Identify the cities that correspond to these four points and explain how they are different from the remaining cities.

c. Circle the point corresponding to a city that has the lowest index value for grocery items and circle a second point corresponding to a city that has the lowest index value for housing.

d. Choose a city that you believe is similar in costs to the costs of your hometown. Is this city above or below average with respect to costs in this group of cities?

3. **Baseball Team's Payroll and Winning**

In professional baseball, there is currently a great variation in the amount of money that teams pay for ballplayers. That raises the question: are teams with high payrolls generally more successful in winning games than teams with low payrolls? To answer this question, the total team payroll (in millions of dollars) and the number of season wins was recorded for the 2003 baseball season. A scatterplot of payroll against number of wins is shown below.



a. Describe the general pattern in this scatterplot. Are teams with high payrolls generally more successful in winning games?

b. Circle a point (label it "A") corresponding to a team that had a high payroll but had a relatively small number of wins.

c. Circle a second point (label it "B") corresponding to a team who had a small payroll but was very successful in winning games.

d. Circle any points that seem to deviate from the general pattern in the plot. Provide how these points are outliers.

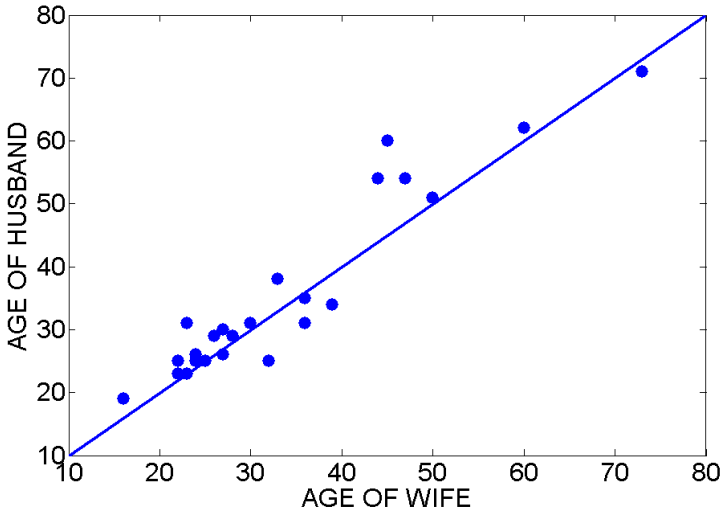
4. **Marriage Ages**

Listed below are the ages of a sample of 24 couples taken from marriage licenses filed in Cumberland Country, Pennsylvania, in June and July of 1993.

Couple	Husband	Wife	Couple	Husband	Wife
1	25	22	13	25	24
2	25	32	14	23	22
3	51	50	15	19	16
4	25	25	16	71	73
5	38	33	17	26	27
6	30	27	18	31	36
7	60	45	19	26	24
8	54	47	20	62	60
9	31	30	21	29	26
10	54	44	22	31	23
11	23	23	23	29	28

12 34 39 24 35 36

The following scatterplot displays the relationship between husband's age and wife's age. The line drawn on the scatterplot is a 45 degree-line where the husband's age would equal the wife's age.



a. Does there seem to be an association between husband's age and wife's age? If so, is it positive or negative? Would you characterize it as strong, moderate, or weak? Explain.

b. Look back at the original listing of the data to determine how many of the 24 couples' ages fall exactly on the line. In other words, how many couples listed the same age for both the man and the woman on their marriage license?

c. Again looking back at the data, for how many couples is the husband younger than the wife? Do these couples fall above or below the line drawn in the scatterplot?

d. For how many couples is the husband older than the wife? Do these couples fall above or below the line drawn in the scatterplot?

e. Summarize what one can learn about the ages of marrying couples by noting that the majority of couples produce points which fall above the 45 degree line.

5. **Direction and Strength of Association**

Describe the relationship between the following pairs of variables as NEGATIVE, POSITIVE, or LITTLE. If there is a positive or negative relationship, describe the strength of the relationship (STRONG, MODERATE, or WEAK).

Pair of variables	Direction of association	Strength of association
Height and armspan		
Height and shoe size		
Height and GPA		
SAT score and college GPA		
Latitude and average January temperature of American cities		
Lifetime and weekly cigarette consumption		
Serving size and calories of fast food sandwiches		
Airfare and distance to destination		
Cost and quality rating of peanut butter brands		
Course enrollment and average student evaluation		
Number of absences and grade in a statistics class		

6. **Direction and Strength of Association**

Suppose you are a math teacher who is interested in understanding which variables are related to the student's final grade that is measured as a percentage from 0 to 100. Below is a list of variables collected on the students that may help to explain the student's final grade. For each variable, give the direction of the relationship of the variable with the student's final grade. Also rank the variables in order of most associated with final grade to least associated with final grade.

a. Student's IQ.

b. Number of absences from class.

c. Student's score on the first test

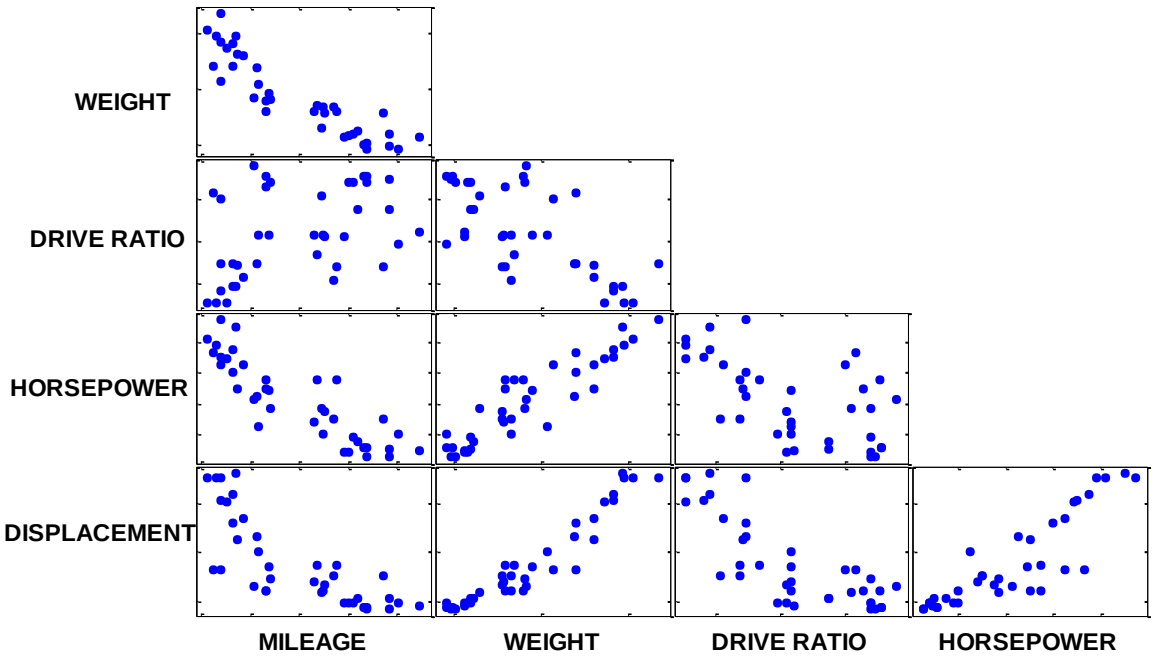
d. Student's height.

e. Number of hours that the student studies on average each week.

f. Student's grade point average.

7. **Car Measurements**

Below we have displayed a scatterplot matrix of the measurements WEIGHT, DRIVE RATIO, HORSEPOWER, DISPLACEMENT, and MILEAGE that were made on 38 1978-79 model automobiles.



Using the list of correlation values

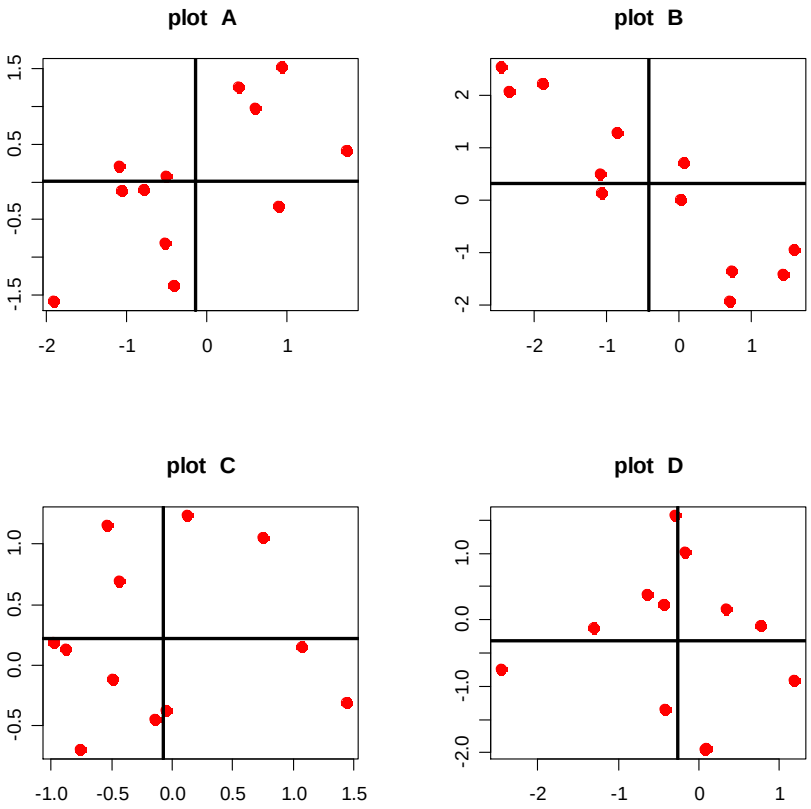
0.42, -0.69, -0.79, -0.90, 0.95 -0.80, 0.87, -0.87, 0.92, -0.59

write down the correlation above each scatterplot in the matrix.

8. **Computing Values of the QCR**

For each of the following scatterplots, (1) describe the direction and strength of the relationship between the two variables and (2) compute the value of the Quadrant

Count Ratio (QCR). The horizontal and vertical lines are drawn at the means of the two variables.



Scatterplot A:

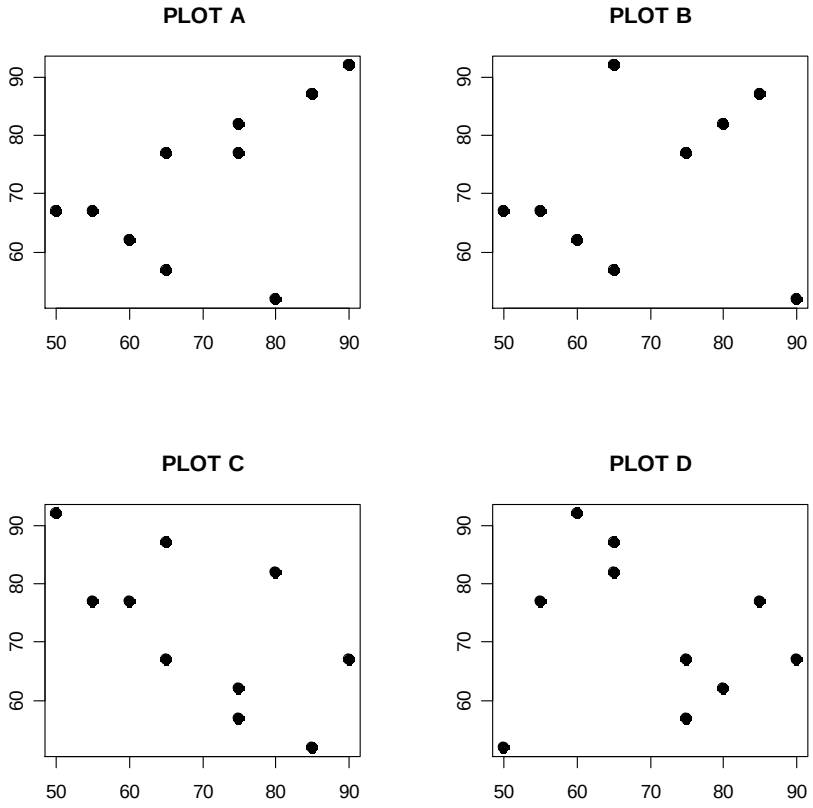
Scatterplot B:

Scatterplot C:

Scatterplot D:

9. **Two Measures of Association**

The following display shows four scatterplots of hypothetical test scores.



(a) List the four scatterplots in the order of most negatively associated to most positively associated. (Which plot is most negatively associated? Next, which plot is next most negatively associated, and so on.)

(b) For each plot, compute the value of the QCR. (The mean value of the horizontal variable is 70 and the mean value of the vertical variable is 72.)

(c) Do the values of the QCR follow the same ordering as your ordering in part a?

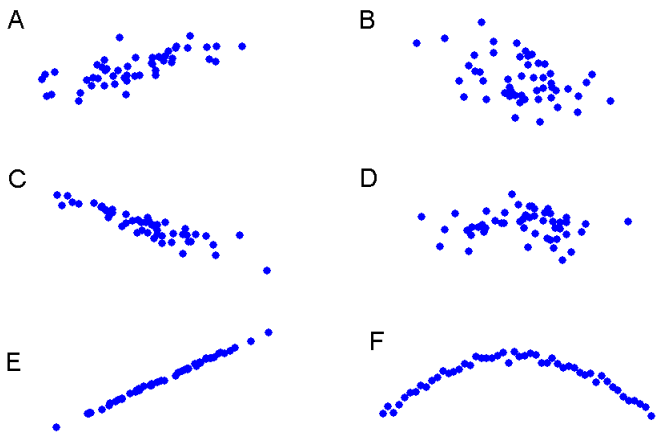
(d) The correlation values of Plots A, B, C, and D are respectively 0.53, 0.14, -0.64, and -0.14. Do these correlation values following the same ordering as your ordering in part b?

(e) Explain why the QCR and the correlation lead to different orderings of the association in the four scatterplots.

10. **Estimating Correlations**

Six scatterplots are shown in the figure below. Using the list of possible correlation values, find the correlation for each scatterplot.

Scatterplot	Correlation	Scatterplot	Correlation
A		D	
B		E	
C		F	



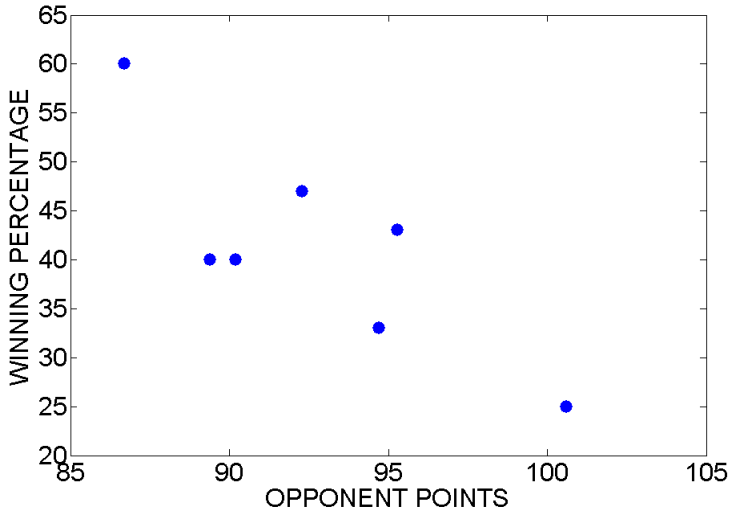
Possible correlation values: -0.99, -0.91, -0.73, -0.43, -0.14, 0.00, 0.43, 0.73, 0.91, 0.99

11. **Points Scored and Winning for Basketball Teams**

The table below gives the average number of points scored by an opponent and the number of wins and losses for seven professional basketball teams. Suppose you are interested in computing the correlation between a team's winning percentage (Win_Pct) and the points scored by the opponent (Opp). In basketball, for a team to be successful in winning, it is important to play good defense to keep the opponent from scoring many points. A scatterplot of these two variables is shown below.

Team	Opp	Wins	Losses	Win_Pct	ZX	ZY	product
Miami	89.4	21	32	40	-0.72	-0.10	0.07

Washington	94.7	16	33	33	0.43	-0.74	-0.32
Orlando	100.6	13	40	25			
New Jersey	86.7	30	20	60			
Philadelphia	90.2	21	31	40	-0.54	-0.10	0.05
New York	92.3	25	28	47	-0.09	0.54	-0.05
Boston	95.3	23	30	43	0.57	0.17	0.10



a. From the scatterplot, estimate the value of the correlation.

b. In the table, some of the standardized scores (for both opponent points and winning percentage) have been computed. Complete this table and use the results to compute the correlation. (The mean and standard deviation of opponent points are 92.7 and 4.6; the mean and standard deviation of winning percentage are 41.1 and 11.0.)

12. **Computing Correlations**

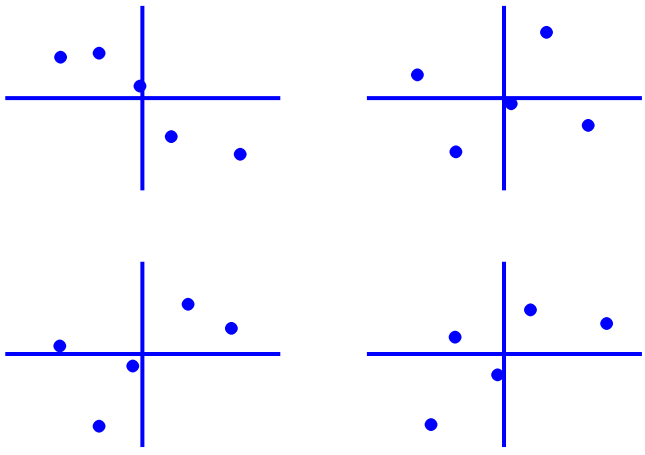
The table shows values of the standardized scores for the x and y variables for four small hypothetical datasets. Compute the correlation for each dataset. When you are completed, write the values of the correlations on the top of the corresponding scatterplots in the figure.

Dataset A	
ZX	ZY
-.63	-1.56
-1.20	.18
-.14	-.26
.67	1.08
1.30	.56

Dataset B	
ZX	ZY
-.63	.96
-1.19	.88
-.03	.25
.42	-.85
1.42	-1.23

Dataset C	
ZX	ZY
-.70	-1.18
-1.26	.49
.11	-.13
.62	1.41
1.23	-.60

Dataset D	
ZX	ZY
-.71	.37
-1.06	-1.53
-.09	-.45
.38	.95
1.49	.66

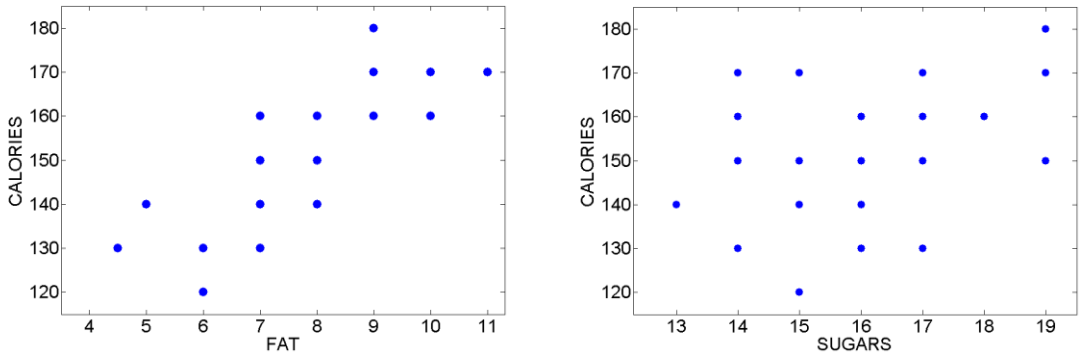


13. **Calories in Ice Cream**

The following table gives the calories, grams of fat, and grams of sugars in a ½ cup serving of each of the flavors of Breyer's ice cream.

Flavor	calories	fat	sugars	Flavor	calories	fat	sugars
carmel fudge	160	7	18	banana fudge chunk	170	9	19
Vanilla	140	8	15	vanilla fudge brownie	160	9	16
french vanilla	150	8	15	cherry chocolate chip	150	8	17
van/choc/straw	140	8	15	peanut butter & fudge	170	10	15
butter pecan	170	11	14	dulce de Leche	150	7	19
Chocolate	150	8	16	lactose free vanilla	160	9	17
mint chocolate chip	160	9	17	mocha almond fudge	170	9	15
strawberry	120	6	15	butter almond	160	10	14
rocky road	150	8	19	calcium rich vanilla	130	7	14
cookies & cream	160	8	16	carmel praline crunch	180	9	19
vanilla fudge twirl	140	7	15	fresh banana	140	5	16
Peach	130	6	16	homemade vanilla	140	7	13
Coffee	140	8	15	extra creamy vanilla	150	8	14
cherry vanilla	140	8	16	extra creamy chocolate	140	7	15
Chocolate chip	160	9	17	take two	150	8	15
Chocolate chip cookie dough	170	9	17	take two (sherbet)	130	4.5	17
vanilla & choc fudge checks	170	9	17				

The left scatterplot below graphs the fat content against the calories and the right scatterplot graphs sugar content and calories.



a. Describe the pattern of association in each graph.

b. Which graph has a stronger association?

c. Estimate the correlation in each graph.

d. Suppose someone tells you that it would be easy to manufacture a "low calorie" ice cream if they could eliminate the fat content. Based on the graphs above, can you agree with this statement? Why or why not?

14. **Body Measurements**

There was a recent study (Mohanty, Babu, and Nair, *Journal of Orthopedic Surgery*, 2001, 19-23) that investigated the relationship between different body parameters. The objective was to find particular body parameter measurements that correlate best with height. Sitting height, standing height, arm span, and leg lengths for 505 healthy women from South India between the ages of 20 and 29 were measured. The table presents correlations that were computed from these data.

Physical measurements	Correlation
Sitting height and arm span	0.561
Standing height and arm span	0.816
Sitting height and leg length	0.294
Standing height and leg length	0.842

a. Based on the table, describe the relationship between a person's sitting height and one's arm span.

b. Based on the table, what is the best predictor of a woman's standing height? Explain.

c. Note that there is a relatively weak association between a woman's sitting height and her leg length. Can you explain why there is a relatively weak association?

d. Think of another physical measurement that you would believe would have a weak relationship with standing height. Explain why the relationship would be weak.