Module 10.1

Scatter Plots And Trend Lines

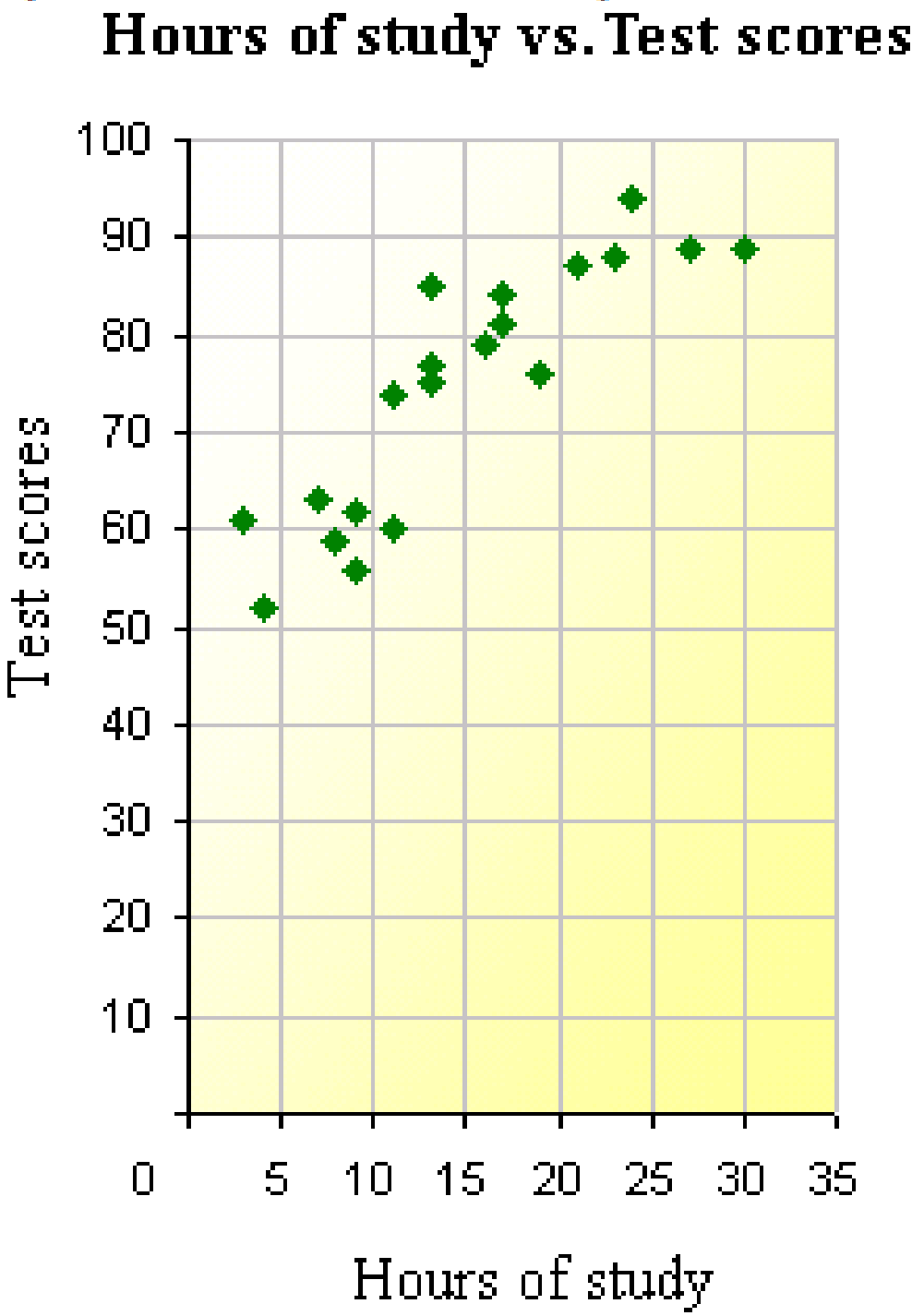
How can you describe the relationship
between two variables
and use it to make predictions?

P. 435

Scatter Plot

A Scatter Plot is a graph made by plotting ordered pairs to show the relationship between two variables.

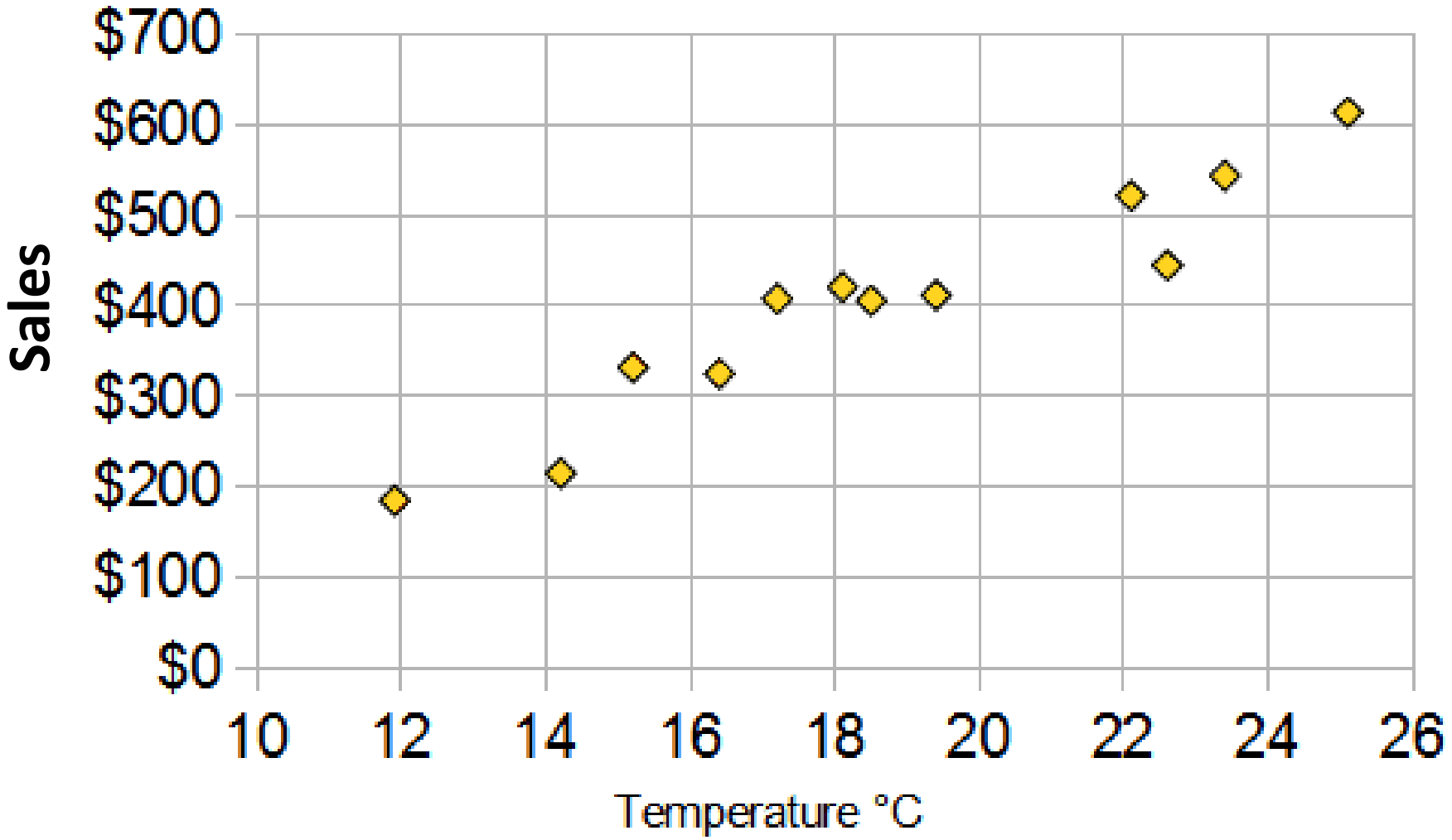
In this example, the scatter plot shows the hours of study and test scores of 20 students.



Here's another example.

The local ice cream shop keeps track of how much ice cream they sell versus the noon temperature on that day. Here are their figures for the last 12 days.

Ice Cream Sales vs Temperature	
Temperature °C	**Ice Cream Sales**
14.2°	$215
16.4°	$325
11.9°	$185
15.2°	$332
18.5°	$406
22.1°	$522
19.4°	$412
25.1°	$614
23.4°	$544
18.1°	$421
22.6°	$445
17.2°	$408

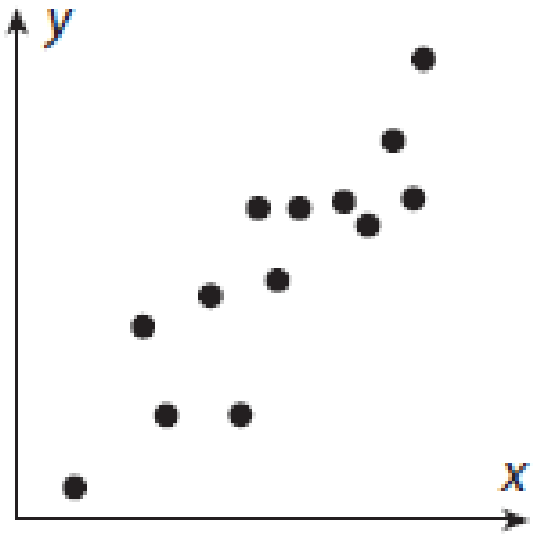


Correlation

The word Correlation is made of **Co-** (meaning "together"), and **Relation**

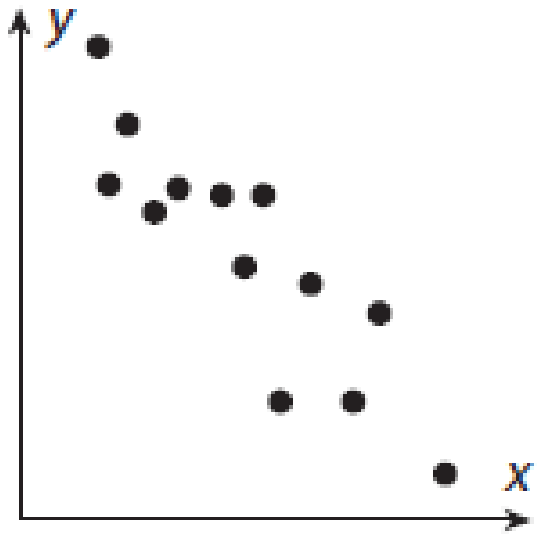
A correlation is a measure of the strength and direction of the relationship between 2 variables.

- Correlation is **Positive** when the values **increase** together, and
- Correlation is **Negative** when one value **decreases** as the other increases



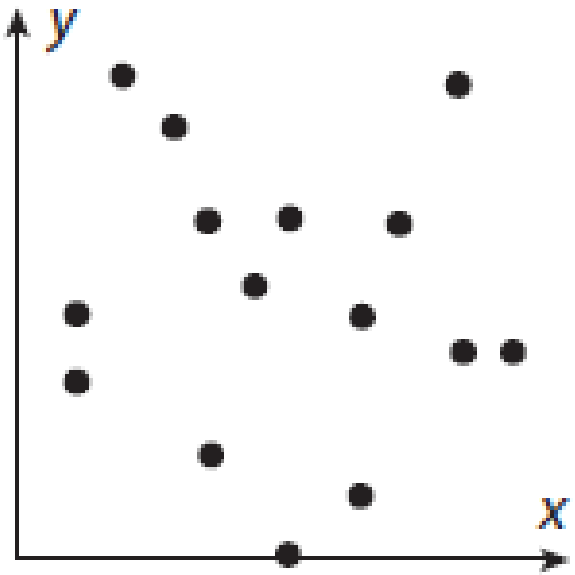
Positive correlation

Positive slope



Negative correlation

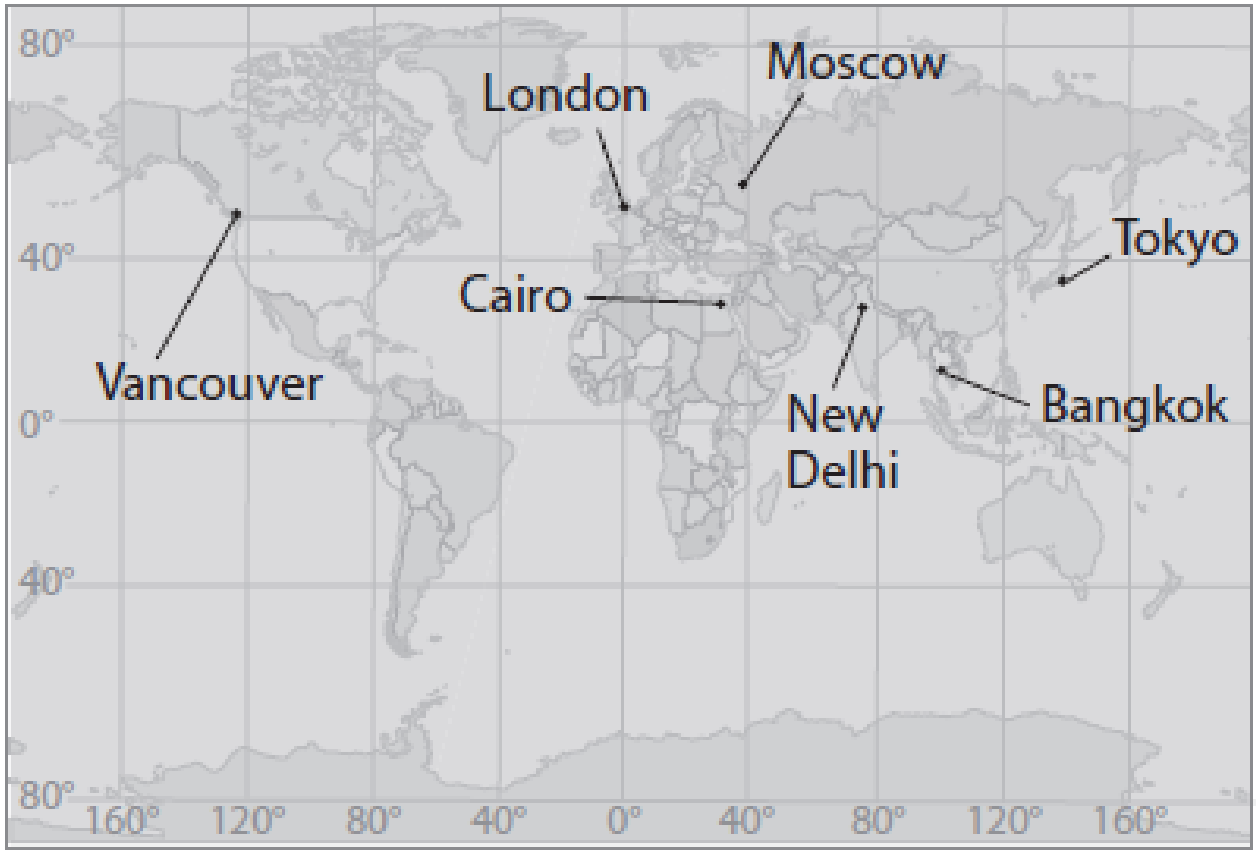
Negative slope



No correlation

(A) The table below presents two-variable data for seven different cities in the Northern hemisphere.

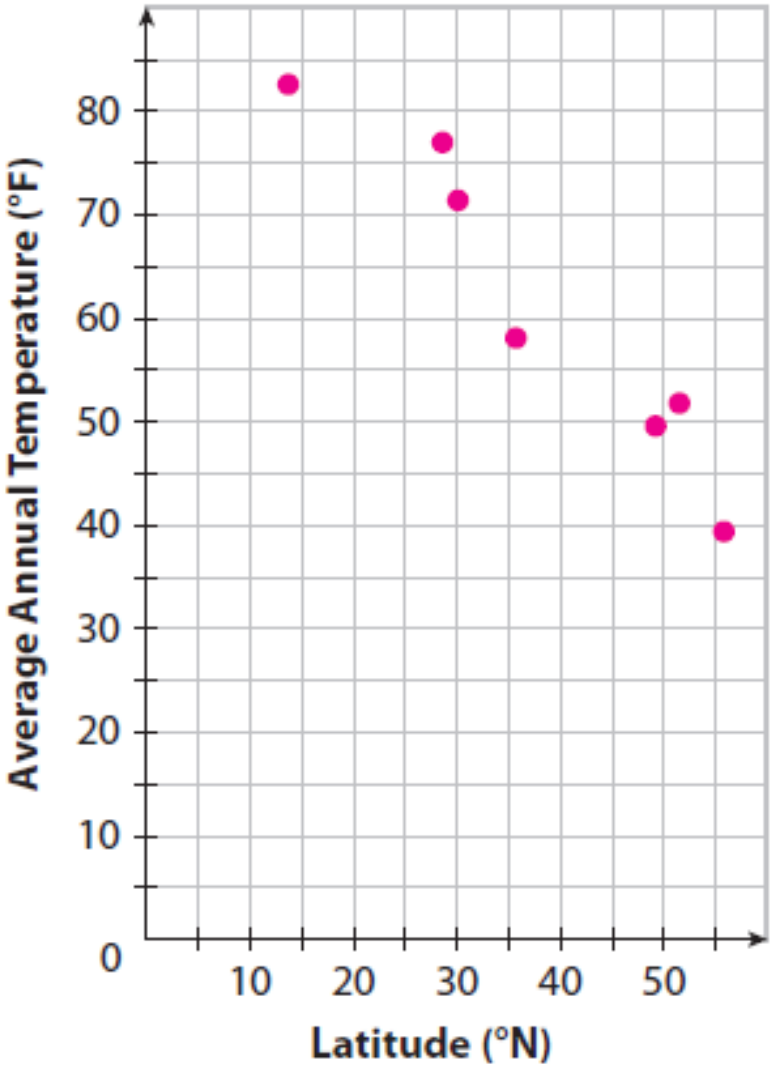
City	Latitude (°N)	Average Annual Temperature (°F)
Bangkok	13.7	82.6
Cairo	30.1	71.4
London	51.5	51.8
Moscow	55.8	39.4
New Delhi	28.6	77.0
Tokyo	35.7	58.1
Vancouver	49.2	49.6



The two variables are Latitude and Temperature.

P. 436

B Plot the data on the grid provided.



The ordered pairs are:
(13.7,82.6) for Bangkok
(30.1,71.4) for Cairo
(51.5,51.8) for London
(55.8,39.4) for Moscow
(28.6,77.0) for New Delhi
(35.7,58.1) for Tokyo
(49.2,49.6) for Vancouver

C The variables are **negatively** correlated.

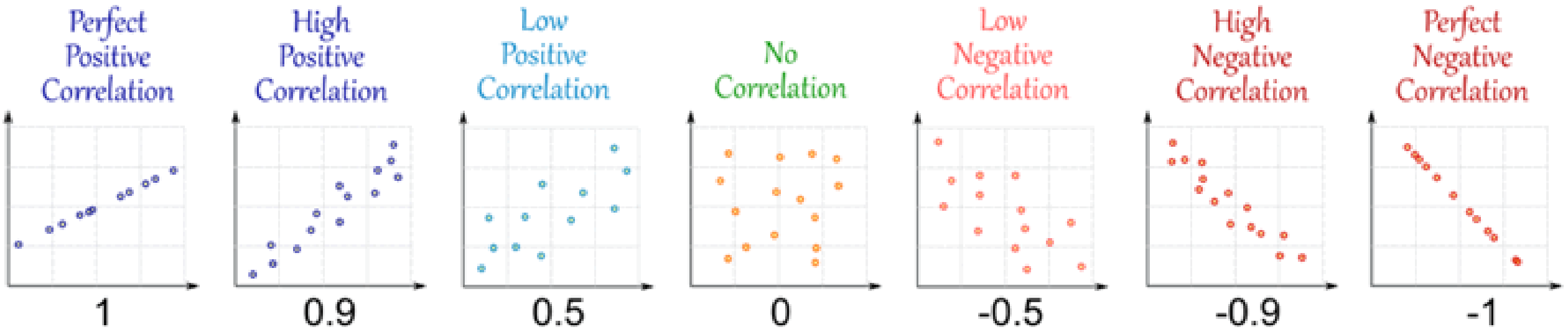
Reflect

1. **Discussion** Why are the points in a scatter plot not connected in the same way plots of linear equations are?

A straight line (or any connected trace) on a graph indicates a continuous set of points. Solutions to linear equations, for example, can be found anywhere along the line that represents the solution. Data in a scatter plot are represented by discrete points. Line segments between points would incorrectly imply either data or function along segments between the scattered points.

The measurement of the correlation is called "The Correlation Coefficient" and is denoted by the letter ***r***, which can range from 1 to –1.

Here are some linear correlations and their values:



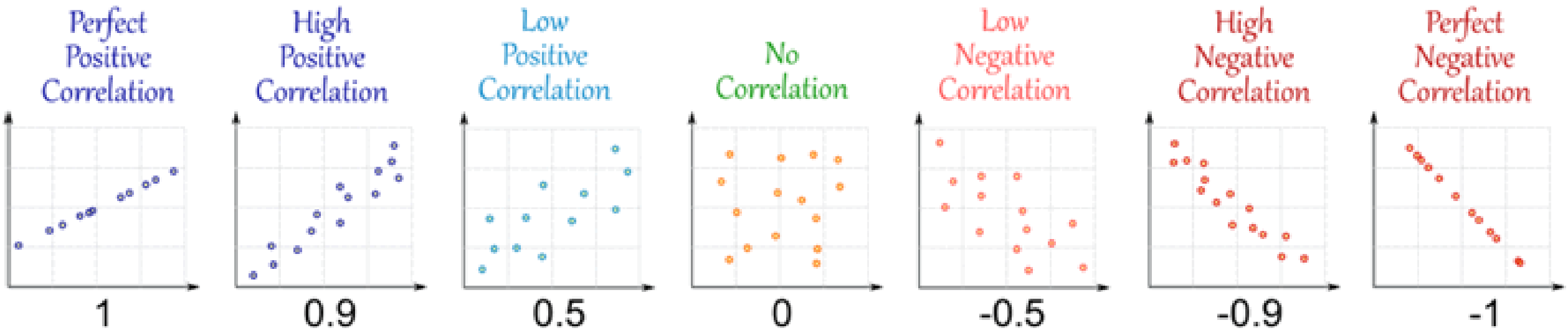
The values are very linked

The values don't seem linked at all

The values are very linked

The value shows **how good the correlation is** (not how steep the line is), and if it is positive or negative.

Note: A correlation of 0.9 is of equal strength to –0.9.



Strongly correlated data points look more like points that lie in a straight line,
and have values of ***r*** that are closer to 1 or -1.
Weakly correlated data points are spread out and will have values of ***r*** closer to 0.

FYI: There's a formula/calculation to determine the precise number for ***r***, but we won't be learning it.

Which of the following usually have a positive correlation? Select all that apply.

a. **The number of cars on an expressway and the cars' average speed**
b. **The number of dogs in a house and the amount of dog food needed**
c. **The outside temperature and the amount of heating oil used**
d. **The weight of a car and the number of miles per gallon**
e. **The amount of time studying and the grade on a science exam**

A car manufacturer collects data on the number of gallons of gasoline left in the gas tank after driving for different numbers of miles. The manufacturer creates a scatter plot of the data and determines that the correlation coefficient is –0.92.

Select *three* true statements based on this correlation coefficient.

A. There is no correlation between the number of miles driven and the gallons of gasoline left in the tank.
B. There is a weak correlation between the number of miles driven and the gallons of gasoline left in the tank.
C. There is a linear correlation between the number of miles driven and the gallons of gasoline left in the tank.
D. There is a strong correlation between the number of miles driven and the gallons of gasoline left in the tank.
E. There is a negative correlation between the number of miles driven and the gallons of gasoline left in the tank.

A student is trying to determine whether there is an association between the number of years of education and the amount of money a person makes. Which of the following would be a reasonable correlation coefficient and interpretation for this situation?

A. The correlation coefficient is −5.1, which indicates no association between the number of years of education and the amount of money a person makes.
B. The correlation coefficient is 8.2, which indicates a strong positive linear association between the number of years of education and the amount of money a person makes.
C. The correlation coefficient is 0.79, which indicates a strong positive linear association between the number of years of education and the amount of money a person makes.
D. The correlation coefficient is −0.94, which indicates a weak negative linear association between the number of years of education and the amount of money a person makes.

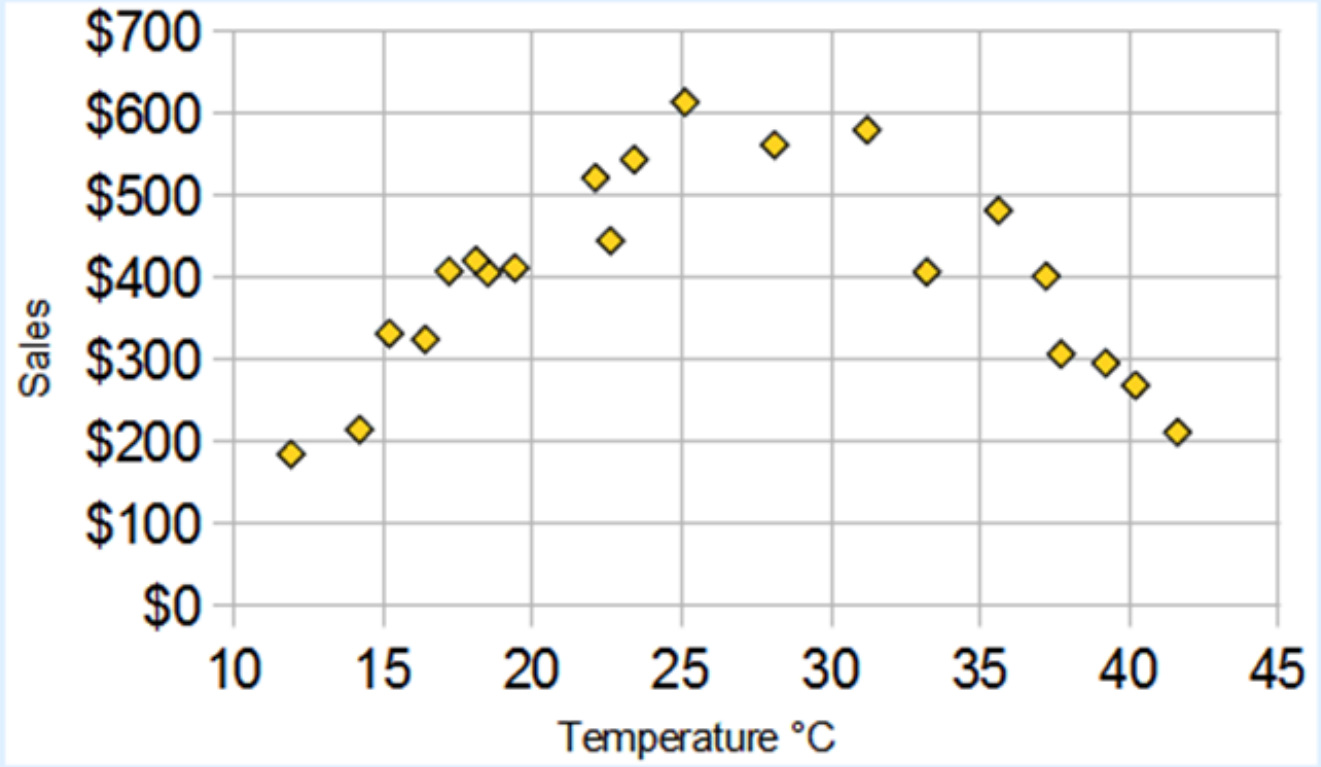
Correlation Is Not Good at Curves

The correlation calculation only works well for relationships that follow a straight line.

Our Ice Cream Example: **there has been a heat wave!**

It gets so hot that people aren't going near the shop, and **sales start dropping**.

Here is the latest graph:



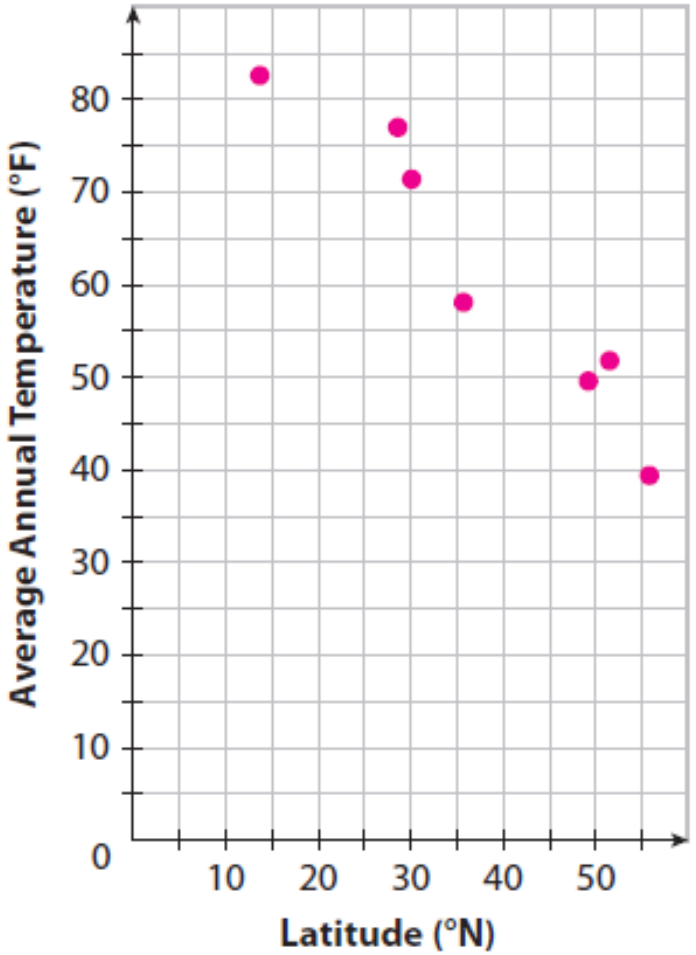
We can visually see the data does have a correlation:
It follows a nice curve that reaches a peak around 25° C.

But the linear correlation calculation isn’t "smart" enough to see this; it’s value is 0, which means "no correlation".

Example 1 Use a scatter plot to estimate the value of r. Indicate whether r is closer to $-1, -0.5, 0, 0.5$, or 1.

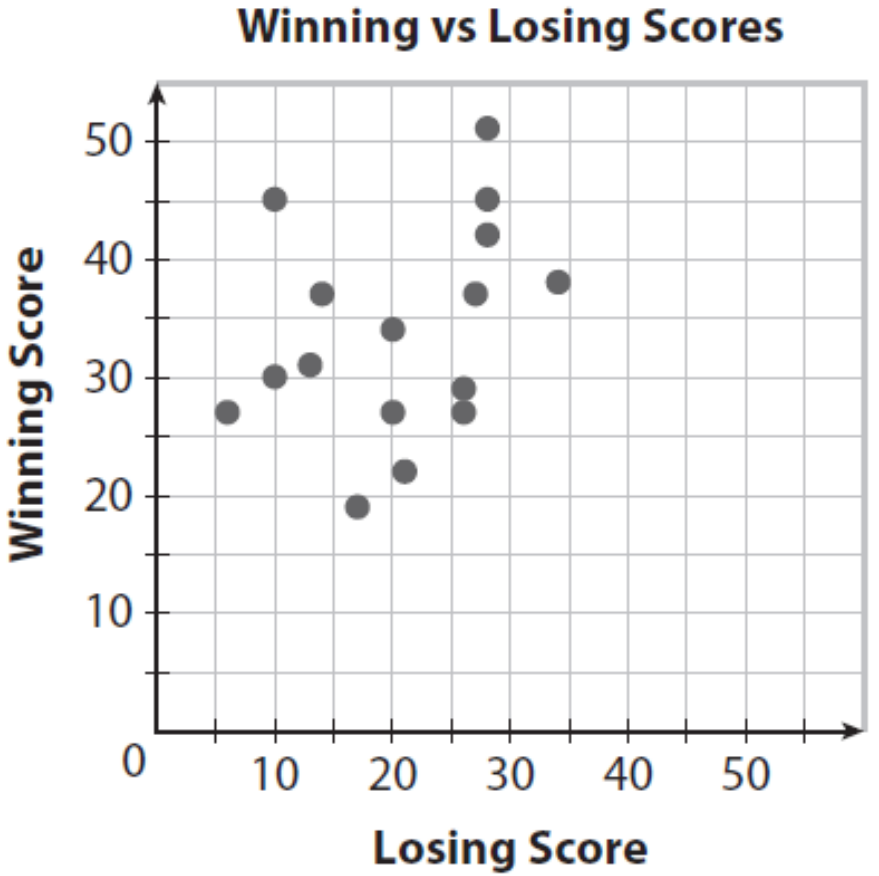
P. 437

(A) Estimate the r-value for the relationship between city latitude and average temperature using the scatter plot you made previously.



This is strongly correlated and has a negative slope, so r is close to -1.

(B)

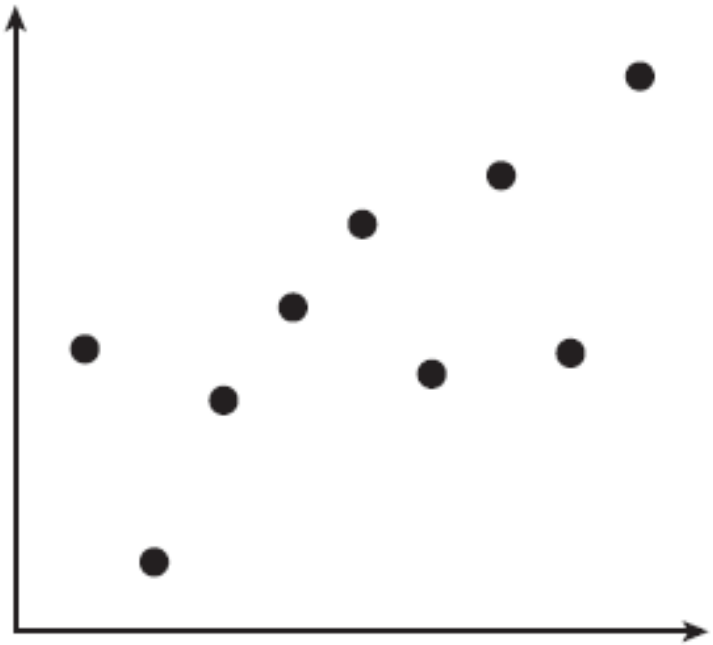


This data represents the football scores from one week with winning score plotted versus losing score.

r is close to **0**.

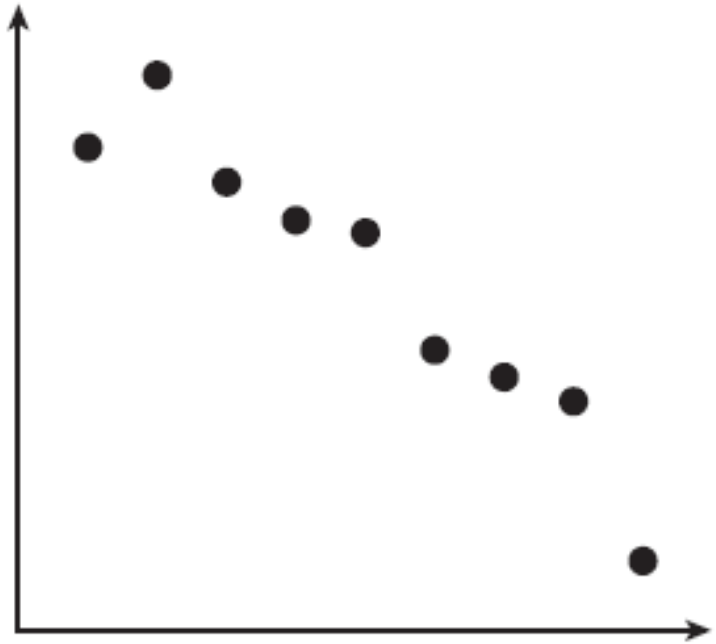
Your Turn

2.



r is close to

3.



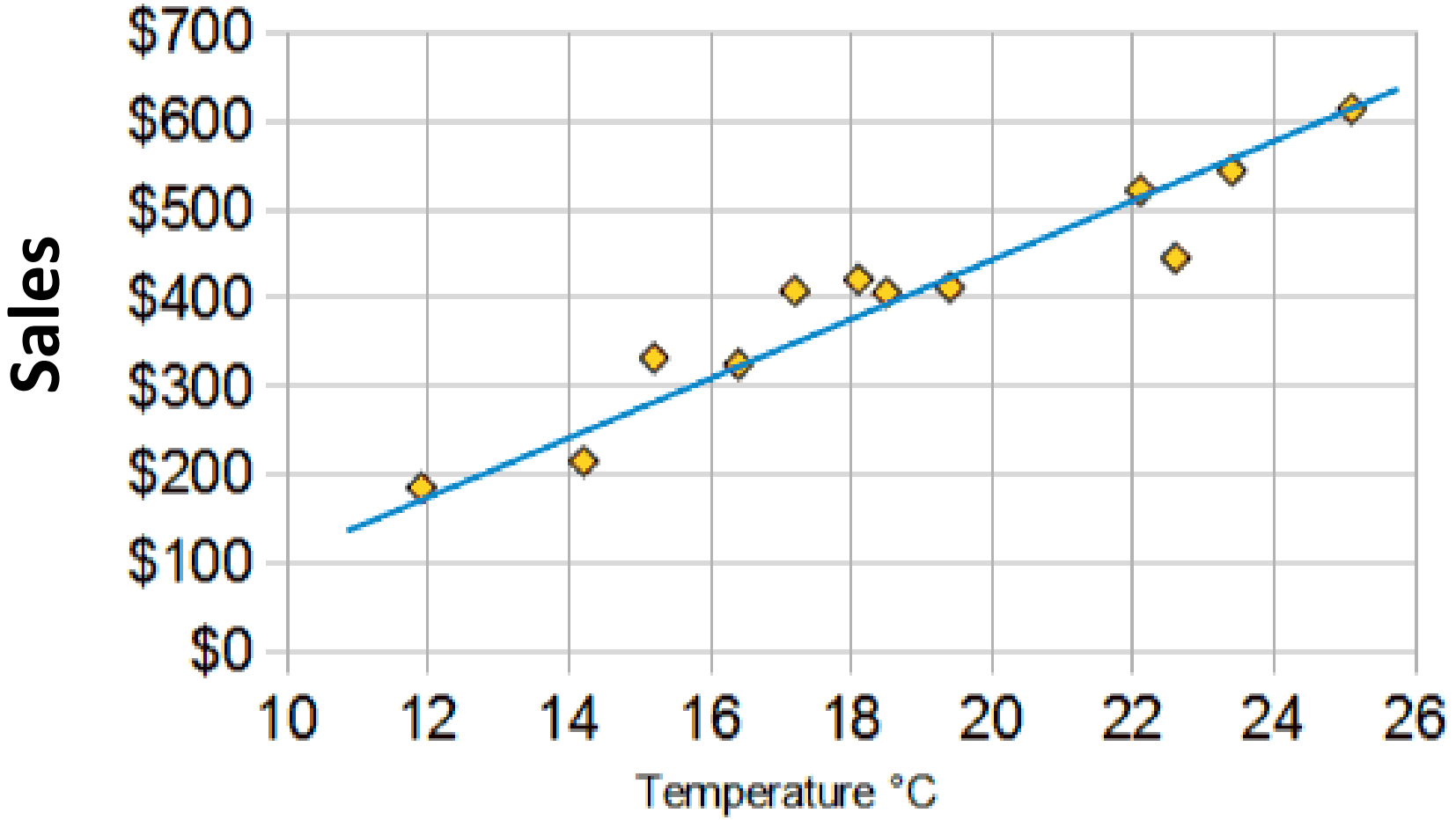
r is close to

Line Of Fit a.k.a. Trend Line

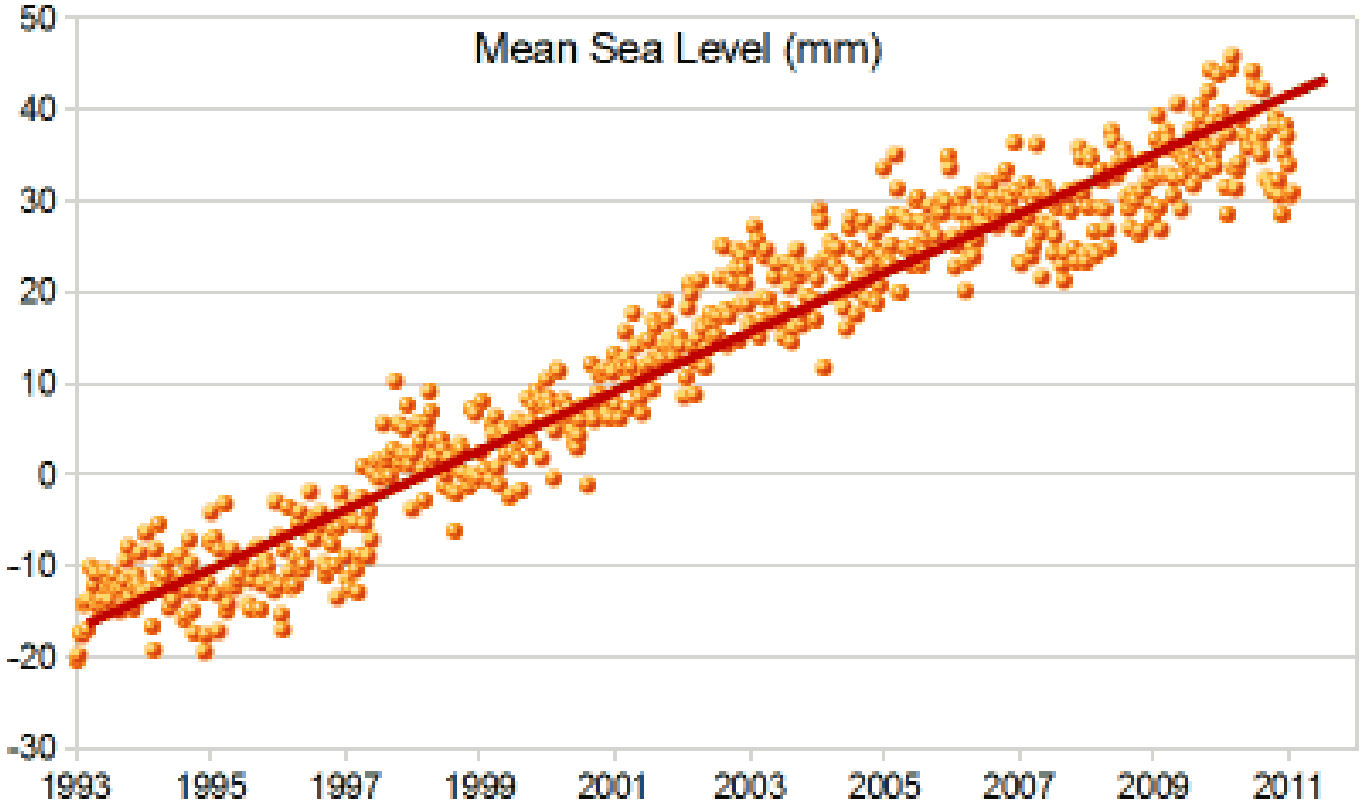
P. 438

A **line of fit** or a **trend line** is a line through a set of two-variable data that illustrates the correlation. It can be used to make predictions.

Remember the graph of the ice cream shop's sales?



Using a straight edge, draw the line that the data points appear to be clustered around. It is not important that any of the data points actually touch the line; instead the line should be drawn as straight as possible and should go through the middle of the scattered points. There is no perfect line to draw. The more the points are spread out, the more lines of fit that can be drawn.

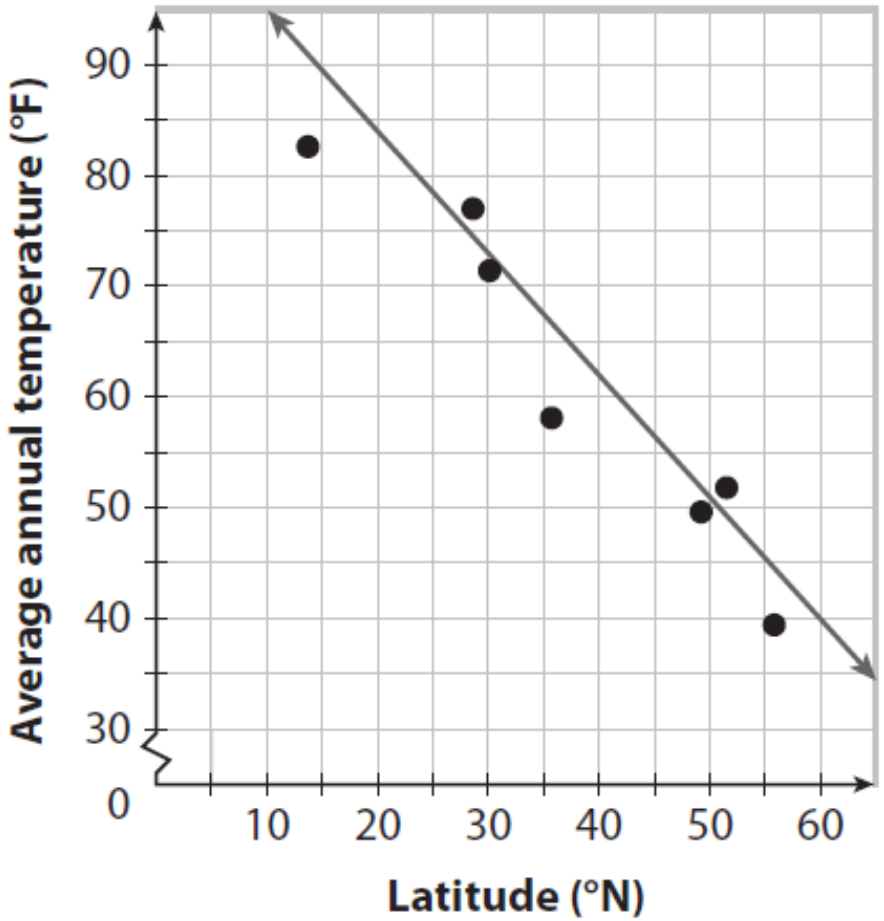


When there is a strong correlation between the two variables in a set of two-variable data, you can use a line of fit as the basis to construct a linear model for the data.

P. 438

Example 2 **Determine a line of fit for the data, and write the equation of the line.**

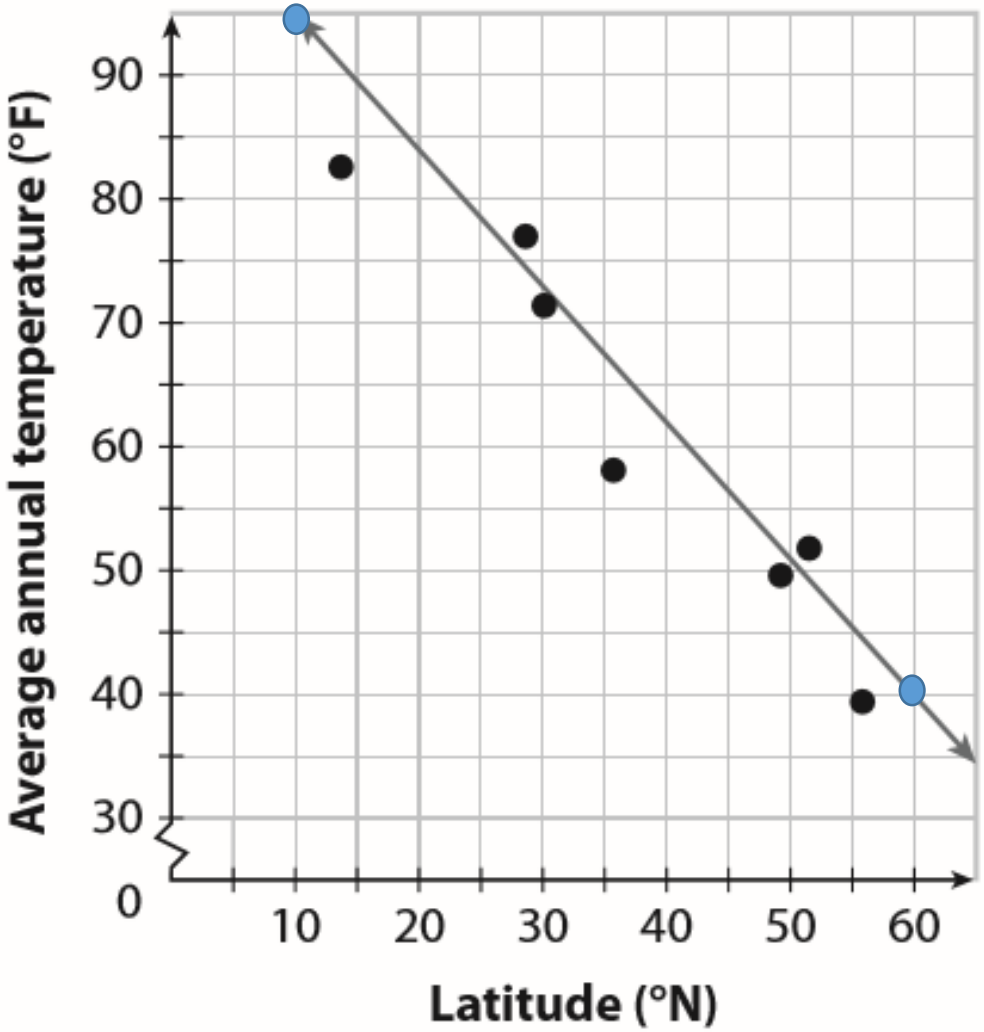
(A) Go back to the scatter plot of city temperatures and add a line of fit.



Once a line of fit has been drawn onto the scatter plot:

- Choose two points on the line to write an equation for the line.
  These DO NOT have to be original data points.
- Calculate the slope.
- Write the equation for the line.

P. 438



The points (10, 95) and (60, 40) appear to be on the line.

$$m = \frac{40 - 95}{60 - 10} = -1.1$$

$$y = mx + b$$

$$95 = -1.1(10) + b$$

$$106 = b$$

The model is given by the equation

$$y = -1.1x + 106$$

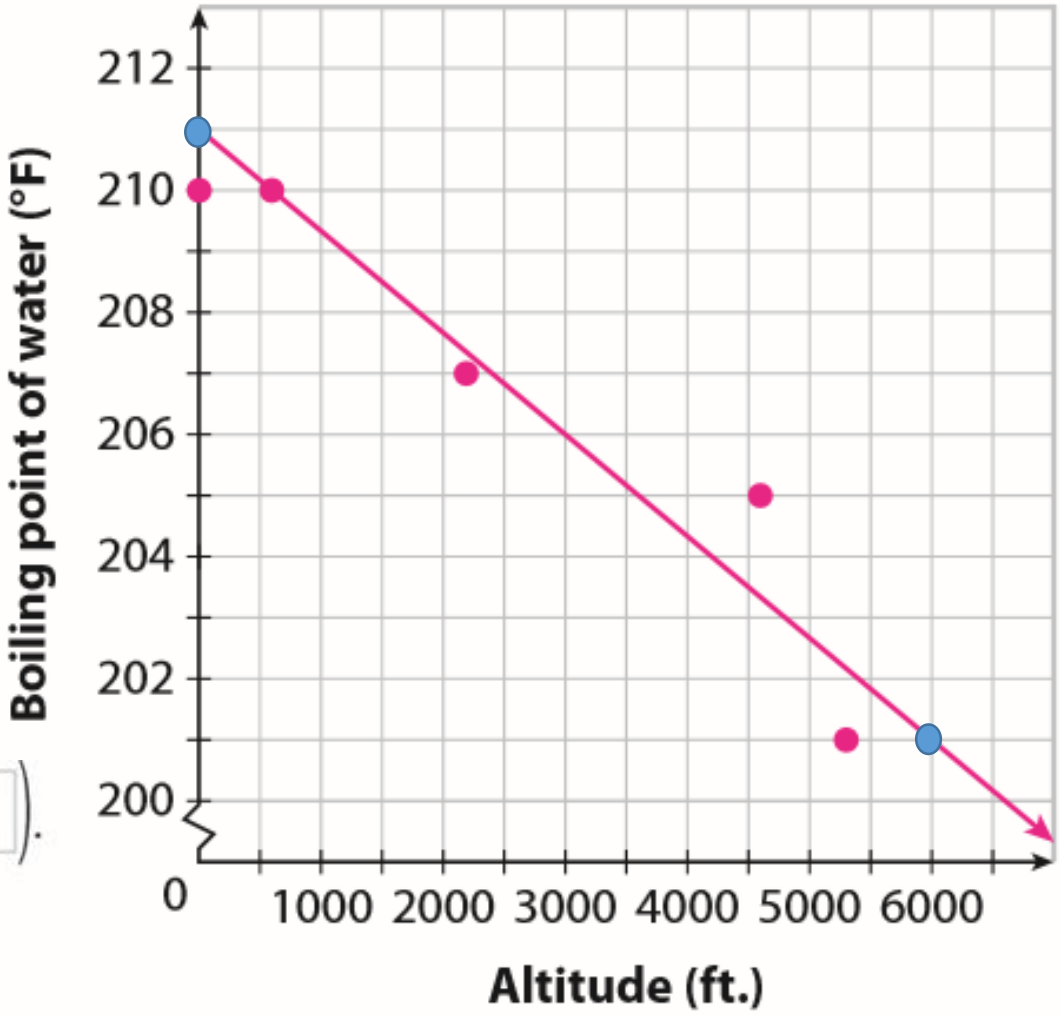
Reflect

4. In the model from Example 2A, what do the slope and y-intercept of the model represent?

The slope is negative and shows a drop in average annual temperature of ~11°F for every 10° increase in latitude. The y-intercept at 0° latitude is the average annual temperature at the equator.

(B) The boiling point of water is lower at higher elevations because of the lower atmospheric pressure. The boiling point of water in some different cities is given in the table.

City	Altitude (feet)	Boiling Point (°F)
Chicago	597	210
Denver	5300	201
Kathmandu	4600	205
Madrid	2188	207
Miami	6	210



A line of fit may go through points $(0, 211)$ and $(6000, 201)$.

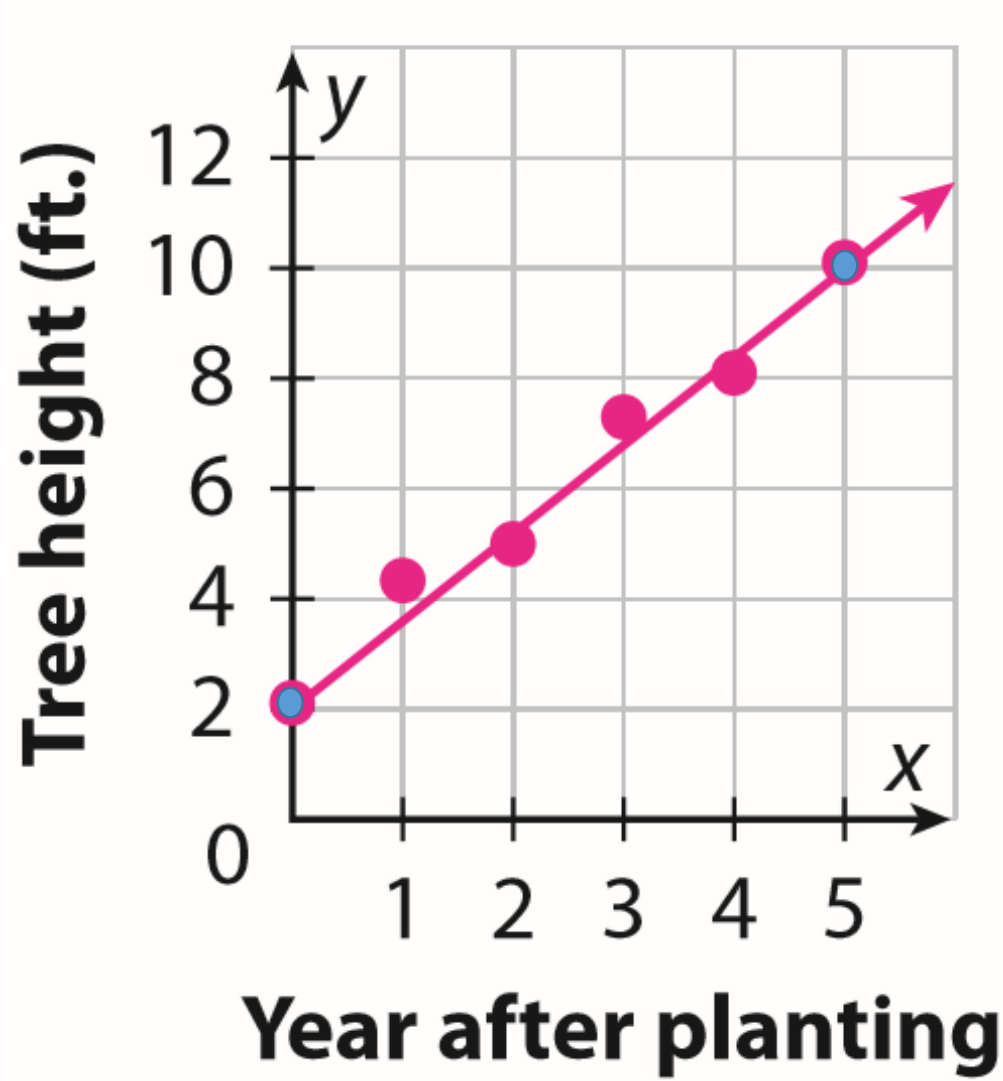
$$m = \frac{-10}{6000} = -0.00167 \qquad b = 211$$

The equation is of this line of fit is $y = -0.00167x + 211$

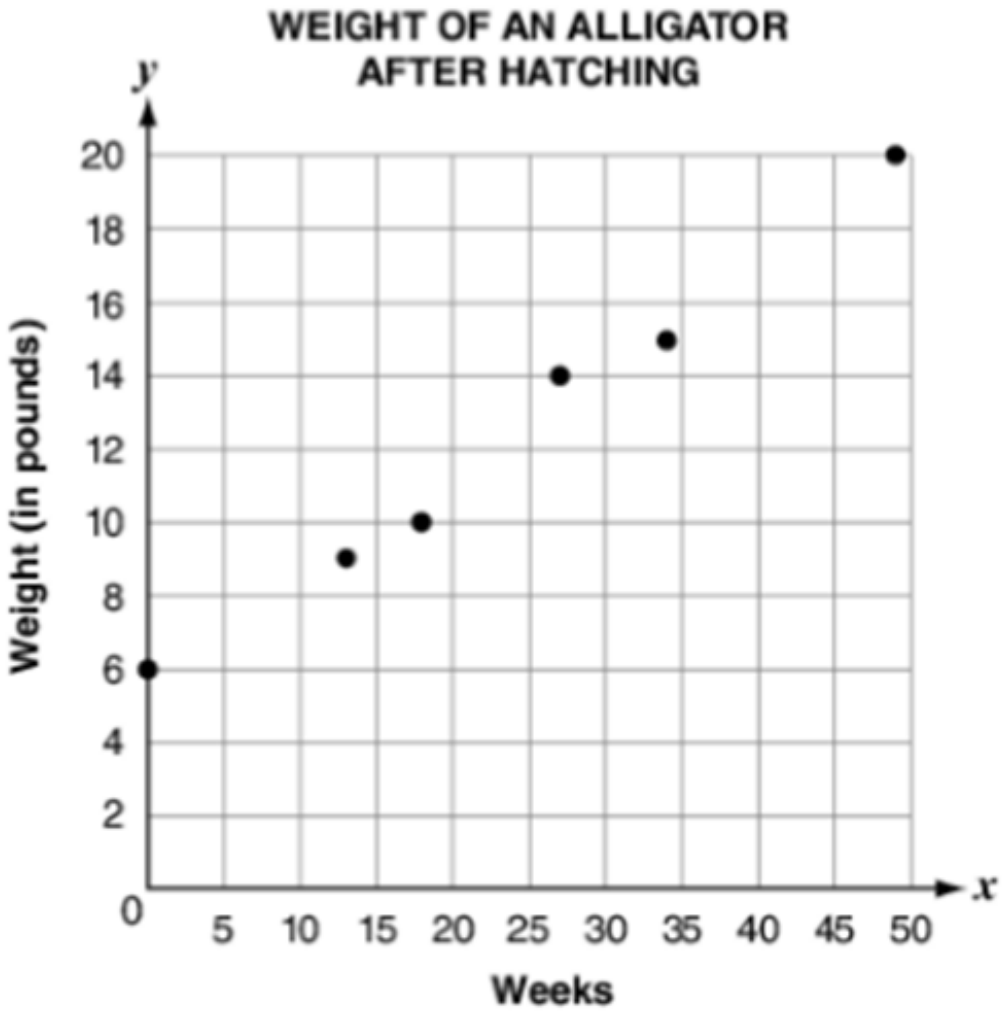
Your Turn

5. Aoiffe plants a tree sapling in her yard and measures its height every year. Her measurements so far are shown. Make a scatter plot and find a line of fit if the variables have a correlation. What is the equation of your line of fit?

Years after Planting	Height (ft)
0	2.1
1	4.3
2	5
3	7.3
4	8.1
5	10.2



The scatter plot below shows how the weight of a baby alligator changed after hatching.



A. $w = 0.25n + 6$

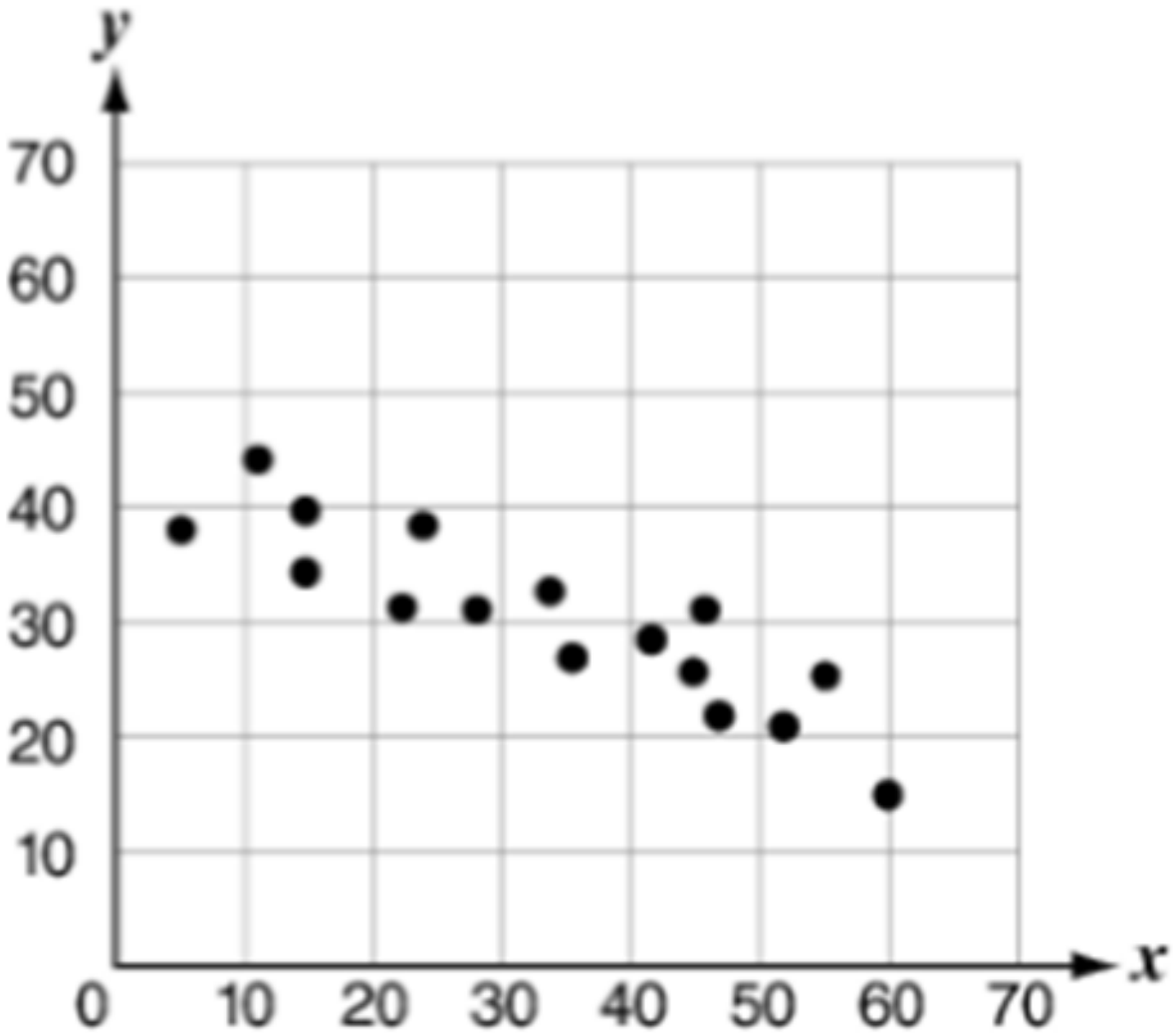
B. $w = 0.65n + 6$

C. $w = 6n + 0.25$

D. $w = 6n + 0.65$

Which equation best represents the weight, *w,* of this alligator *n* weeks after hatching?

Which function best fits the data in this scatter plot?



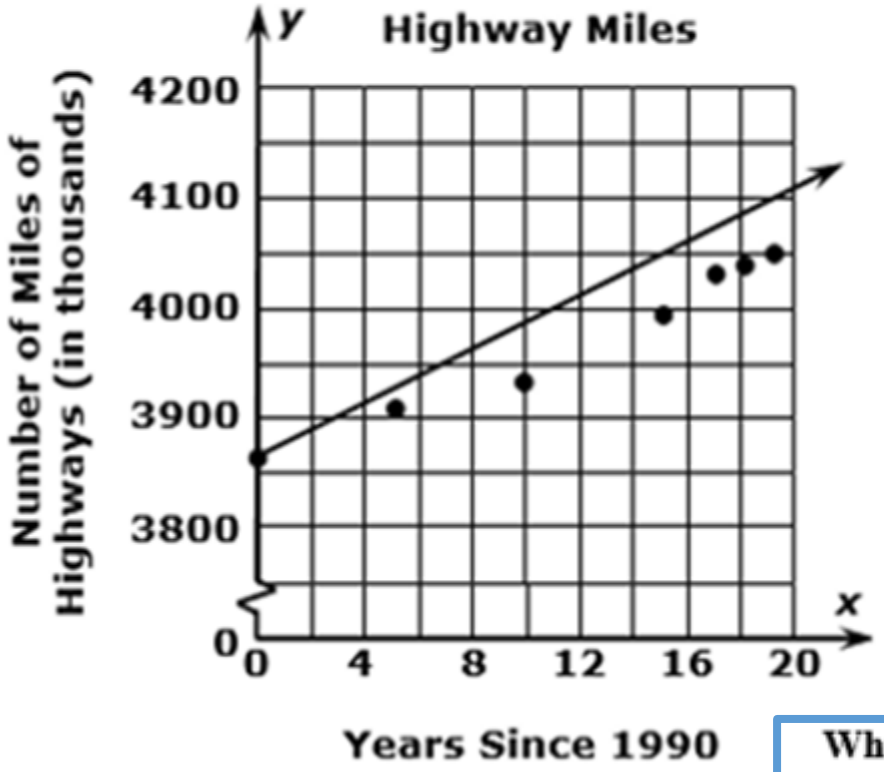
A. $y = -\frac{1}{2}x + 45$

B. $y = -2x + 45$

C. $y = -\frac{1}{4}x + 45$

D. $y = -4x + 45$

A student researches the number of miles in the U.S. highway system since 1990 and creates the scatter plot below. The student draws a line to fit a linear function for the data points on the scatter plot.



Which of the following statements describes the adequacy of the line drawn by the student?

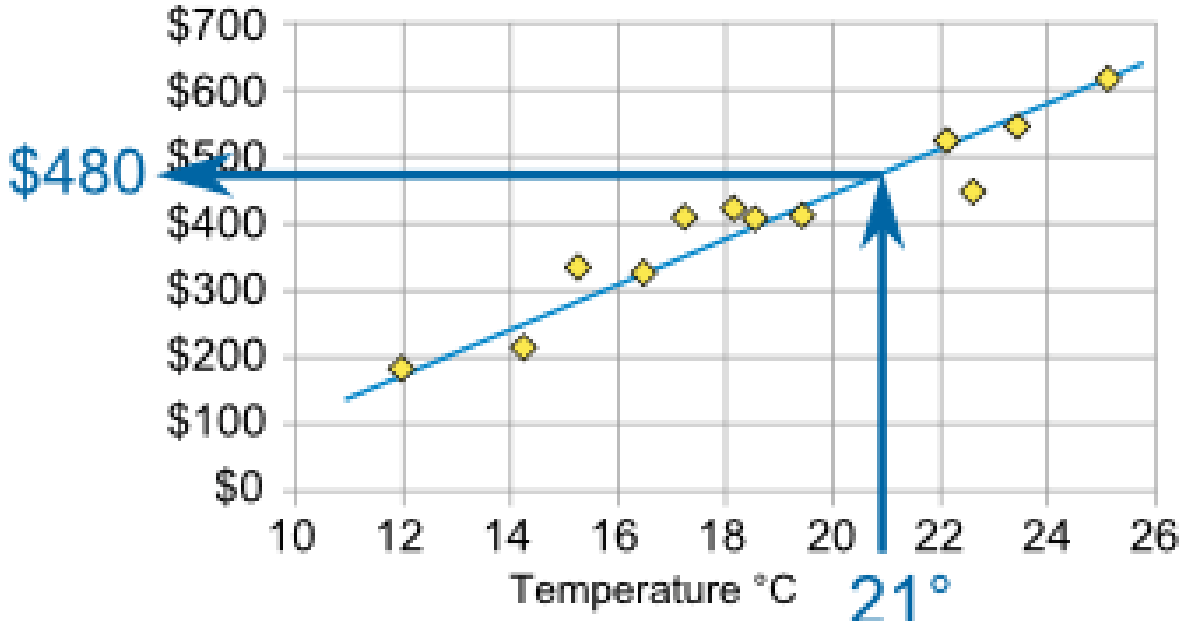
A. The line drawn by the student represents an adequate fit as the line passes through the first plotted data point.

B. The line drawn by the student represents an adequate fit as the line shows the same linear trend as the data points.

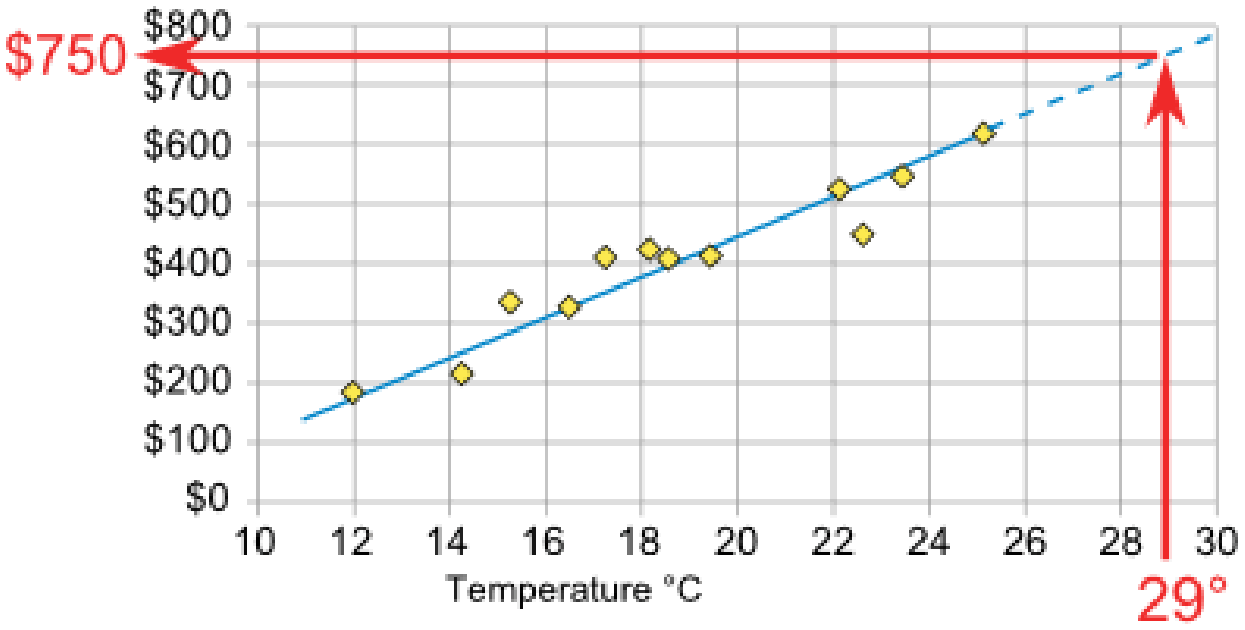
C. The line drawn by the student does not represent an adequate fit as the line should be closer to more of the data points.

D. The line drawn by the student does not represent an adequate fit as the line should be below all of the data points, not above.

Interpolation and Extrapolation



Interpolation is where we find a value **inside** our set of data points. Here we use it to estimate the sales at 21 °C.



Extrapolation is where we find a value **outside** our set of data points.
Here we use it to estimate (predict) the sales at 29 °C (which is higher than any value we have).

Careful: Extrapolation can give misleading results because we are in "uncharted territory".

Example 3 **Use the linear fit of the data set to make the required predictions.**

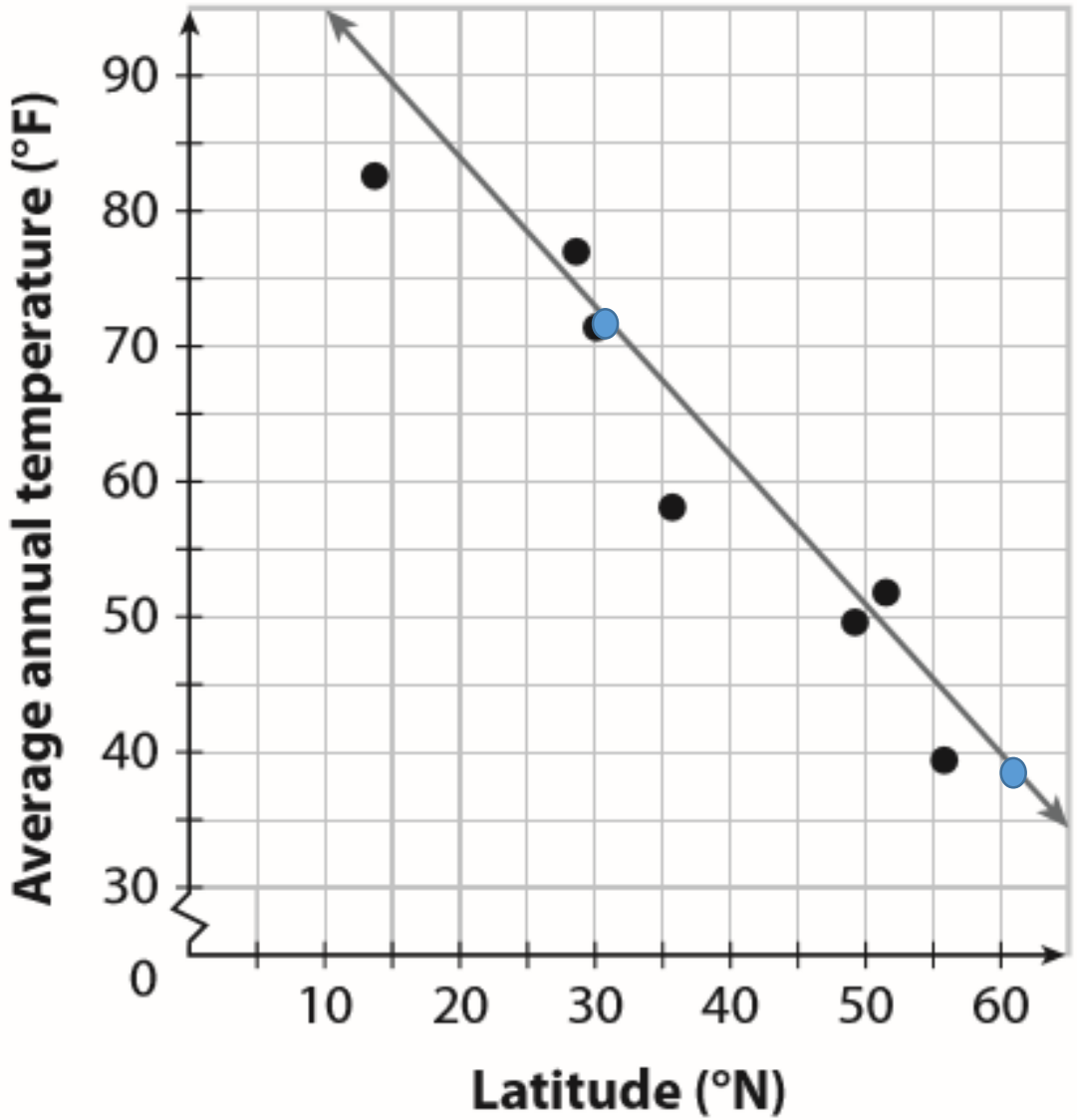
P. 440

(A) Use the model constructed in Example 2A to predict the average annual temperatures for Austin (30.3°N) and Helsinki (60.2°N).

$$y = -1.1x + 106$$

Austin: $y = -1.1 \cdot 30.3 + 106 = 72.67$ °F

Helsinki: $y = -1.1 \cdot 60.2 + 106 = 39.78$ °F

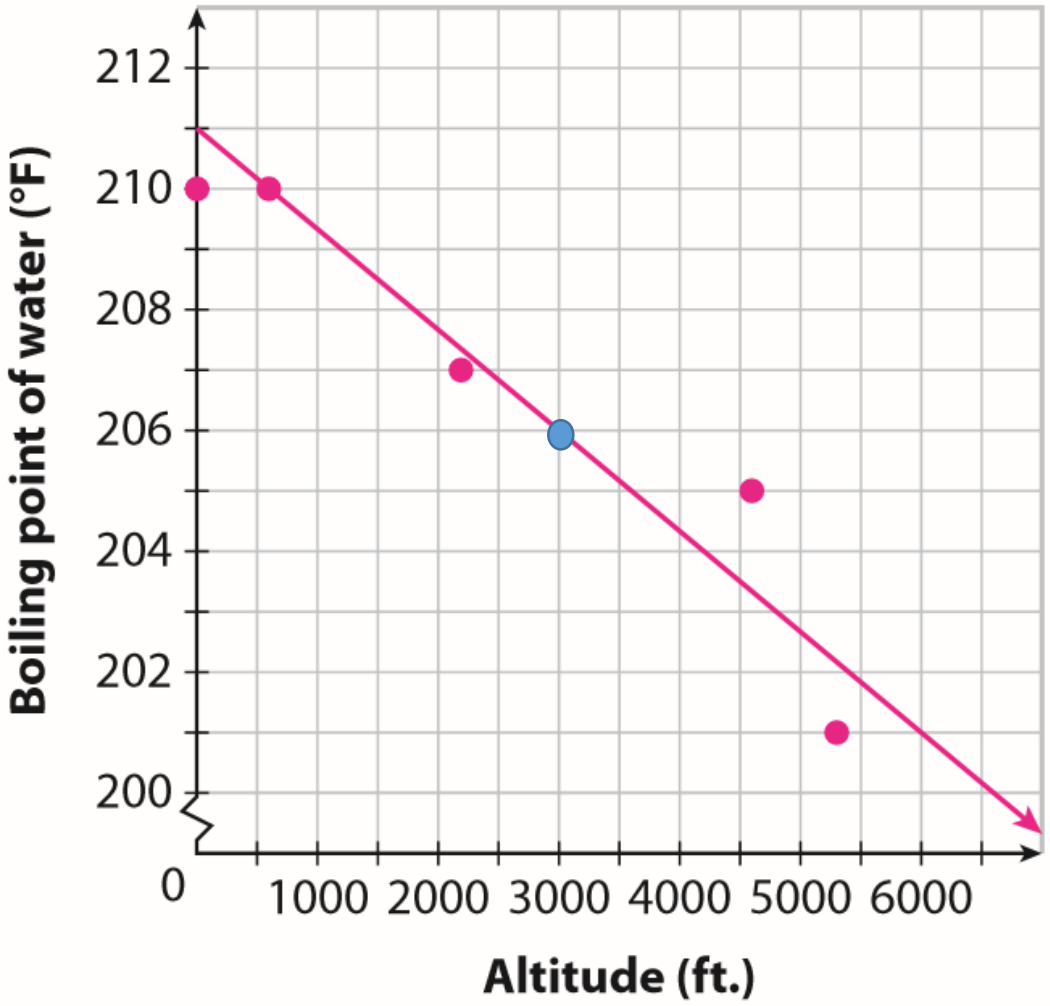


B Use the model of city altitudes and water boiling points to predict the boiling point of water in Mexico City (altitude = 7943 feet) and in Fargo, North Dakota (altitude = 3000 feet)

$$y = -0.00167x + 211$$

Mexico City: $y =$ ____ $\cdot 7943 +$ ____ $= 197.74$

Fargo: $y =$ ____ $\cdot 3000 +$ ____ $= 205.99$



Which prediction would you expect to be more reliable? Why?

The boiling point in Fargo, North Dakota is a more reliable prediction because it is an interpolation, while Mexico City is an extrapolation.

Is it possible to make a prediction based on a scatter plot with no correlation?

No; no correlation means that there is no relationship between the variables and the points on the graph show no pattern.

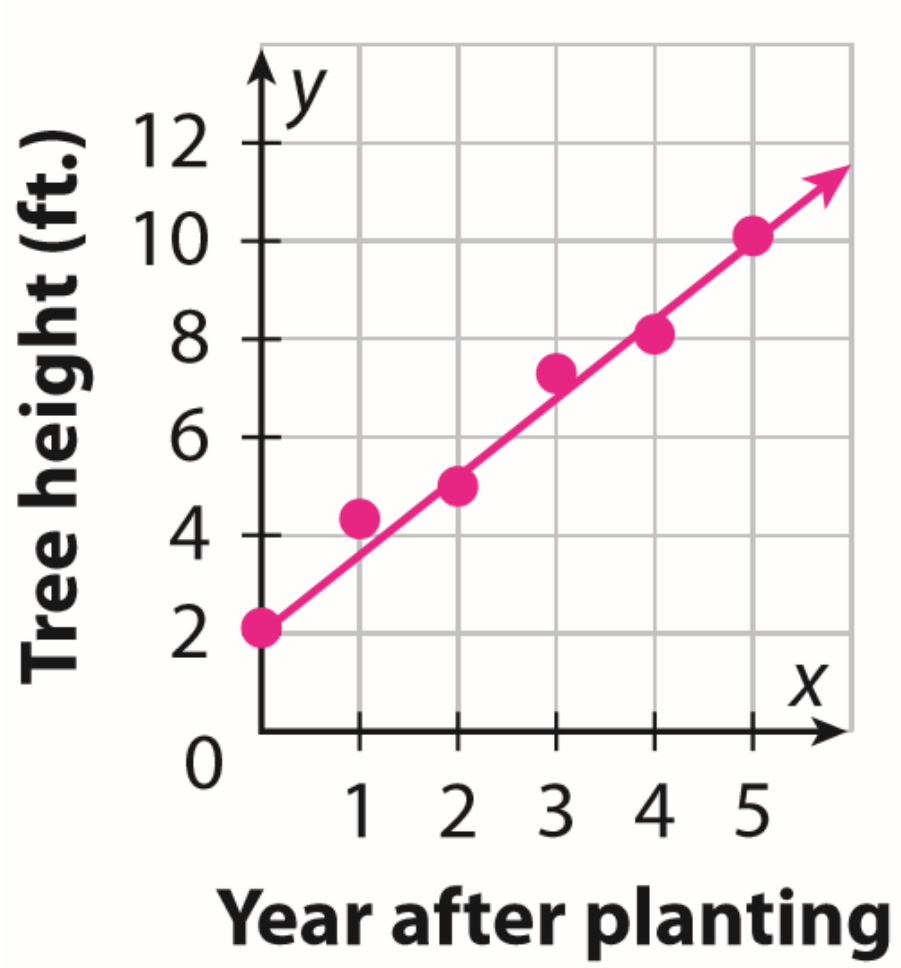
Your Turn

P. 441

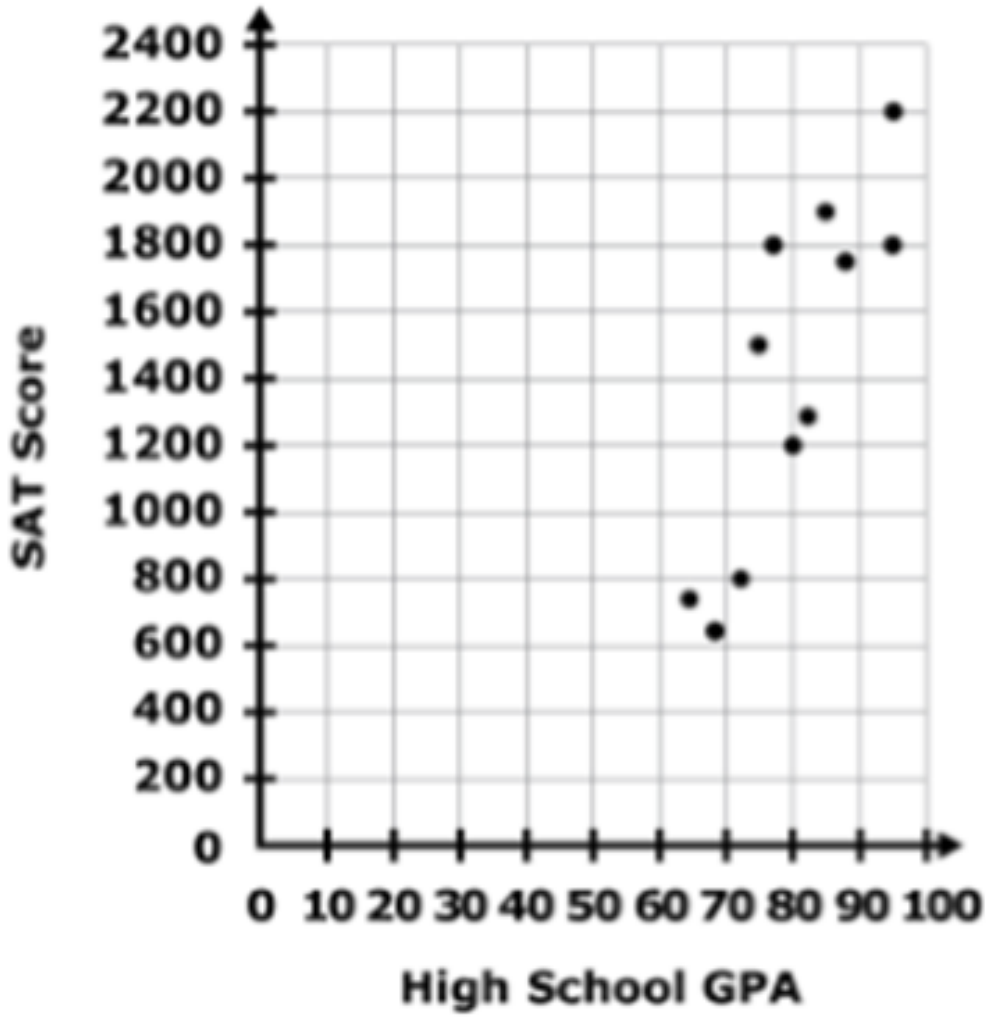
7. Use the model constructed in YourTurn 5 to predict how tall Aoiffe's tree will be 10 years after she planted it.

Equation for Line Of Fit is

$$y = \frac{8}{5}x + 2$$



The correlation of SAT scores and grade point averages (GPAs) for a random sample of high school students is represented by the scatterplot below.



The approximate line of best fit is given by the equation $y = 40x - 1800$. Based on this trend, which *best* predicts the SAT score for a high school student with a GPA of 95?

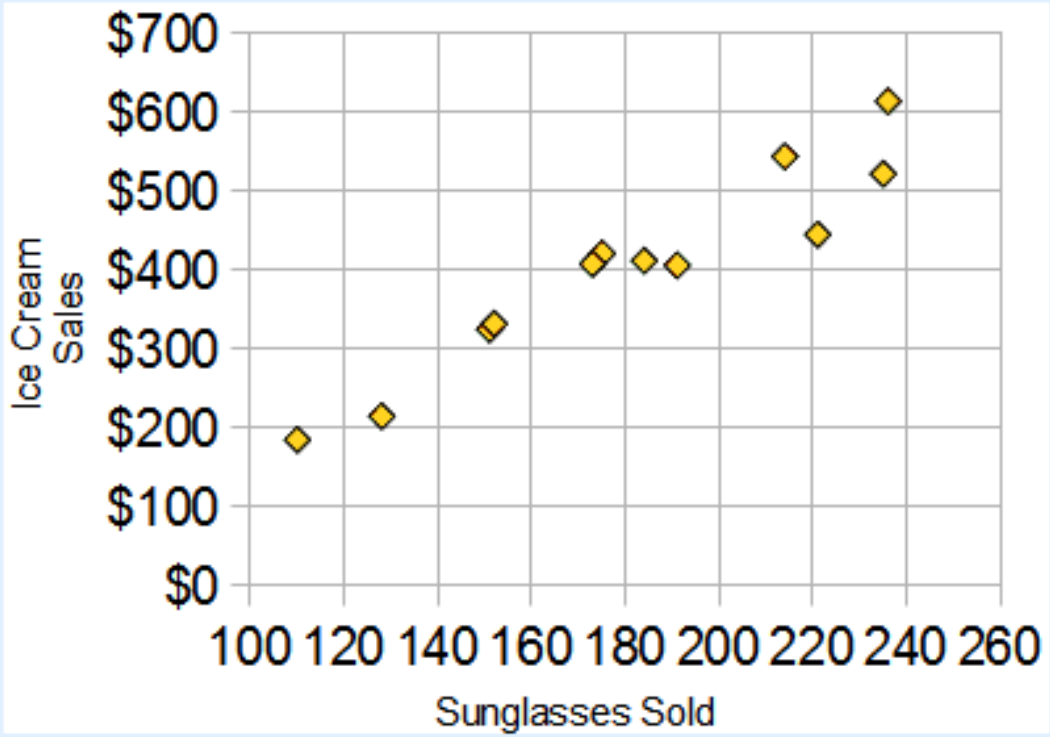
P. 441

Correlation Is Not Causation

A correlation does **NOT** mean that one thing causes the other. There could be other reasons the data has a good correlation!

Example: Sunglasses vs Ice Cream

Our Ice Cream shop finds how many sunglasses were sold by a big store for each day and compares them to their ice cream sales:



The correlation between Sunglasses and Ice Cream sales is high

Does this mean that sunglasses make people want ice cream?

The manager of an ice cream shop studies its monthly sales figures and notices a positive correlation between the average air temperature and how much ice cream they sell on any given day.

P. 441

The two variables are ice cream sales and average air temperatures.
It is likely that warmer air temperatures cause an increase in ice cream sales.
It is doubtful that increased ice cream sales cause an increase in air temperatures.

Describe whether changing either variable is likely, doubtful, or unclear to cause a change in the other variable.

Shoe size increases…So does their reading ability…

Traffic on Biscayne Blvd increases…So do ATM tardies…

A person's height increases…So does their weight…

A student's test scores increase…So does their grade…

A traffic official in a major metropolitan area notices that the more profitable toll bridges into the city are those with the slowest average crossing speeds.

The variables are _____ and _____.

It is [likely | doubtful | unclear] that increased profit causes slower crossing speed.

It is [likely | doubtful | unclear] that slower crossing speeds cause an increase in profits.

Each car pays the same toll regardless of the speed it crosses.

A swim team collects data on the number of laps each member swims in the pool and the time it takes to swim those laps. The team plots their data on a scatter plot. Which statement *most likely* interprets their results?

A. There is likely to be correlation between the number of laps and the time it takes to swim those laps but not causation.

B. There is likely to be causation between the number of laps and the time it takes to swim those laps but not correlation.

C. There is likely to be both correlation and causation between the number of laps and the time it takes to swim those laps.

D. There is likely to be neither correlation nor causation between the number of laps and the time it takes to swim those laps.

A classroom teacher keeps track of the amount of time it takes each of her students to complete a puzzle. The table below shows the teacher's results.

		Time to Solve Puzzle		
		Less Than 2 Minutes	More Than 2 Minutes	Total
Gender	Girls	8	4	12
	Boys	10	5	15
	Total	18	9	27

Based on the data in the table, which statement best describes whether there is evidence to support that the puzzle is easier to solve for one gender than the other?

A.

There is evidence that the puzzle is easier for the boys because 10 boys took less than 2 minutes as compared with 8 girls.

B.

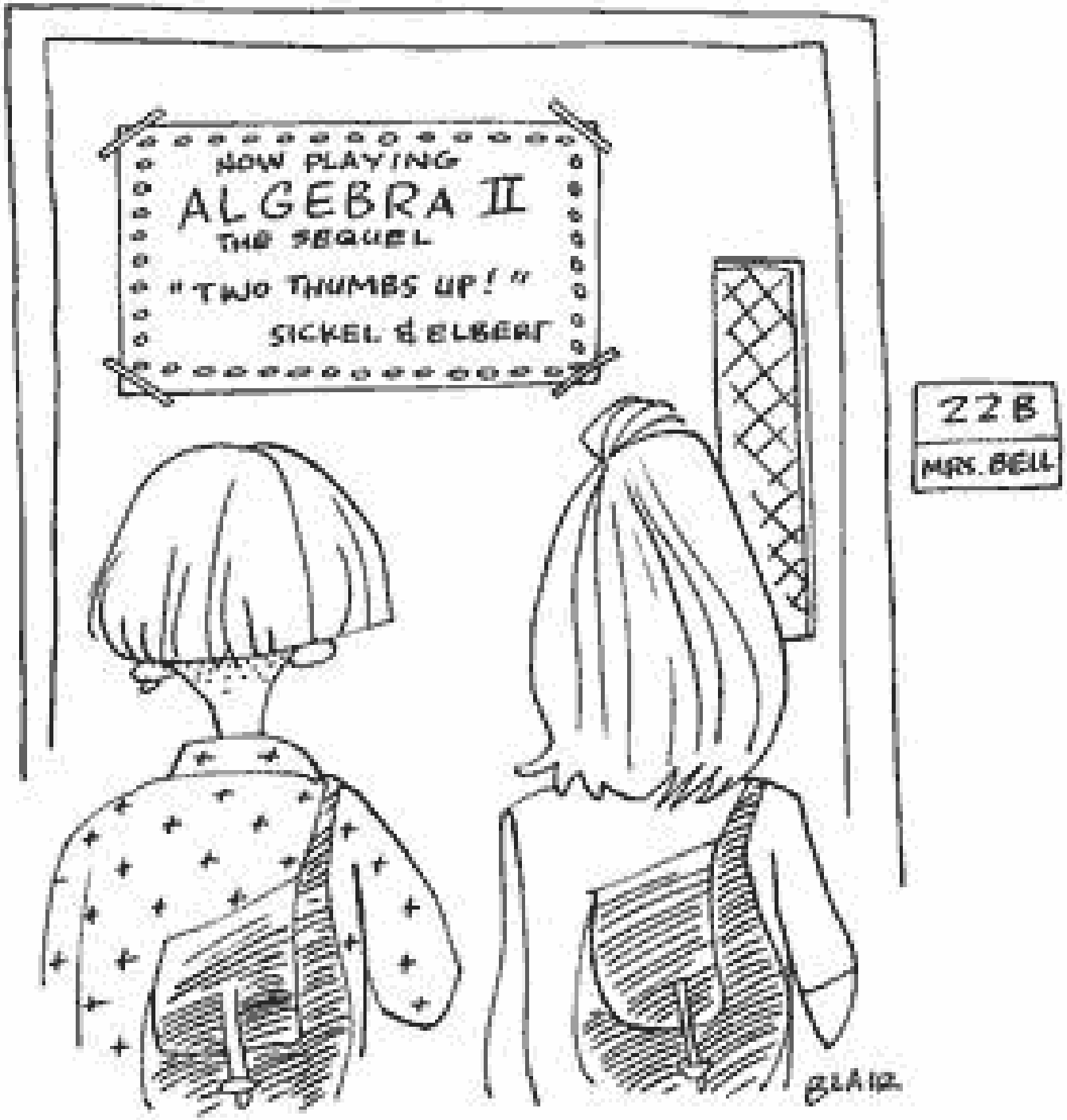
There is evidence that the puzzle is easier for the girls because only 4 girls took more than 2 minutes as compared with 5 boys.

C.

There is no evidence to prove that the puzzle was easier for either gender because equal percentages of boys and girls took less than and more than 2 minutes.

D.

There is no evidence to prove that the puzzle was easier for either gender because the number of boys is not the same as the number of girls.


NOW PLAYING
ALGEBRA II
THE SEQUEL
"TWO THUMBS UP!"
SICKEL & ELBERT
22B
MRS. BELL
BLAIR