Lesson 57 – Scatter Plots, Correlation, & Trend Lines

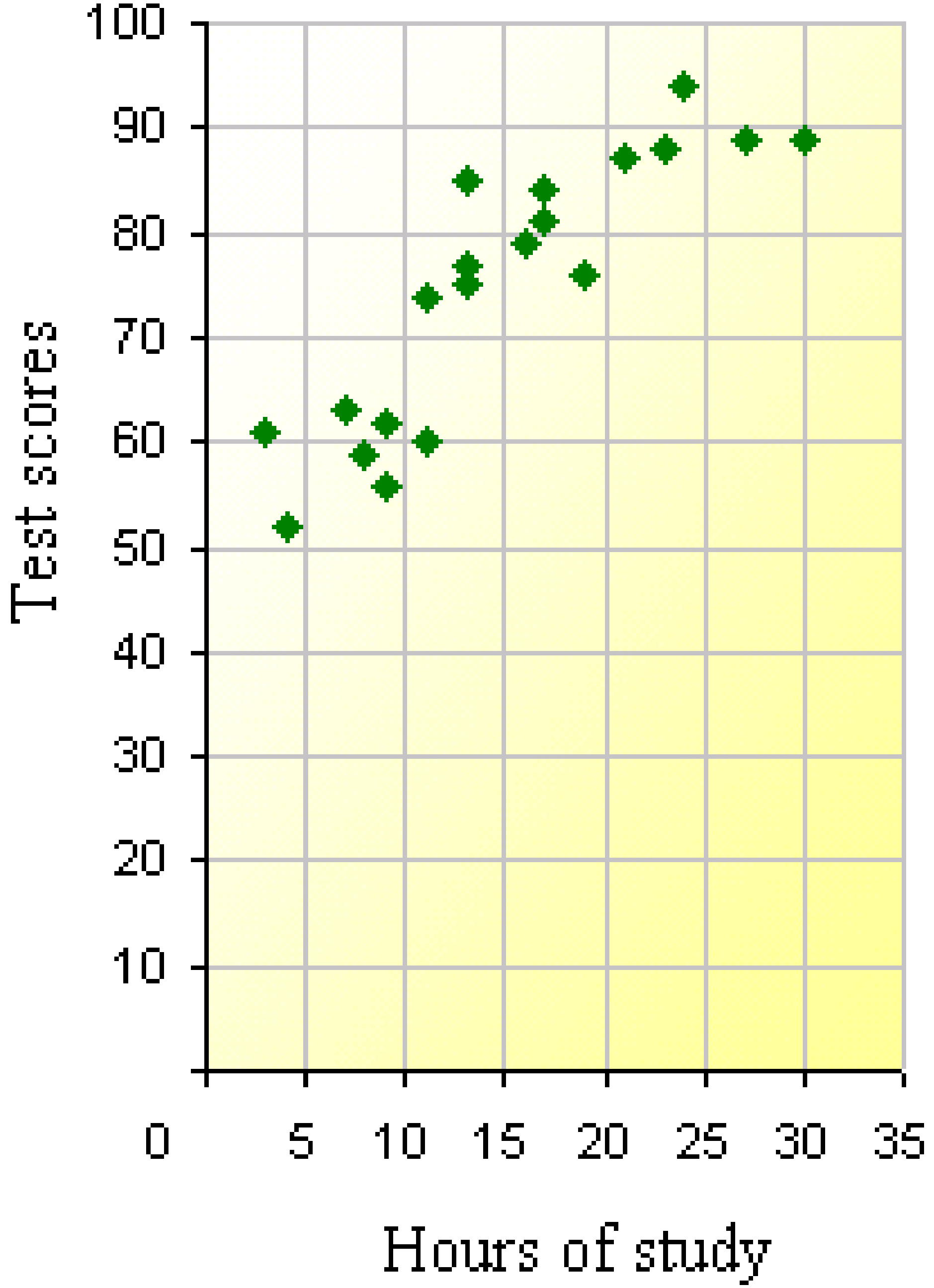
A **Scatter Plot** is a graph made by plotting ordered pairs to show the relationship between two quantitative variables.

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

This is bivariate data, meaning it involves the relationship between an independent variable (hours of study) and a dependent variable (test scores).

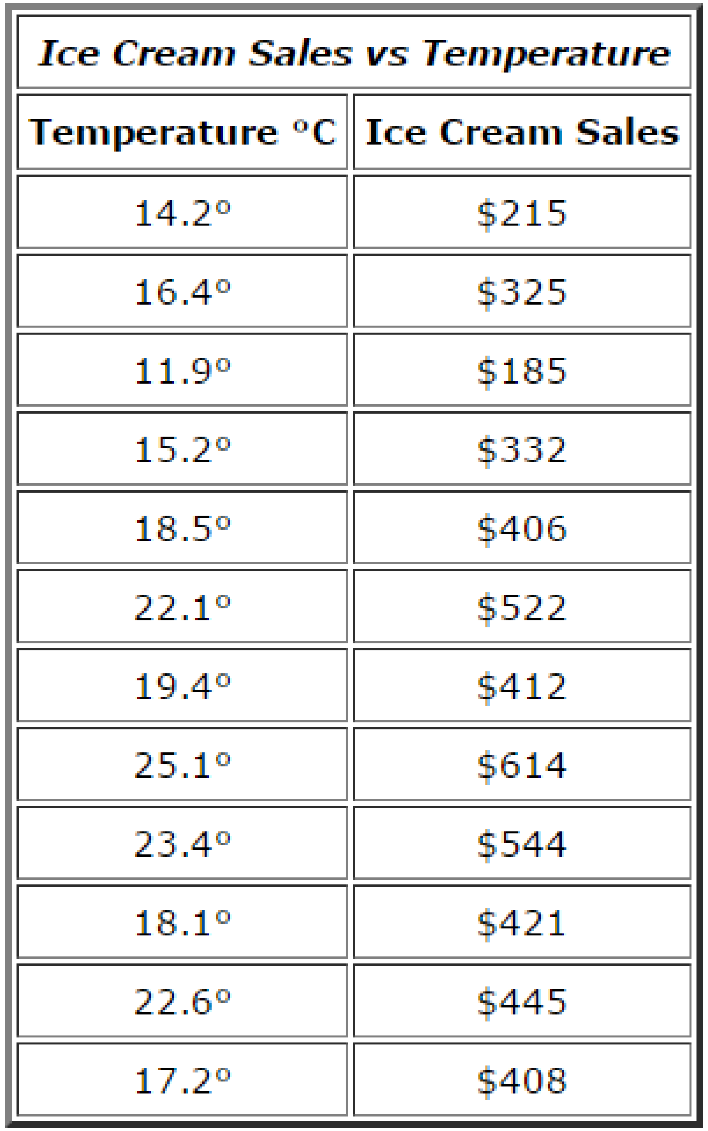
Other names for the I.V. are **explanatory** and **predictor**. Another name for the D.V. is **response**.

Hours of study vs. Test scores



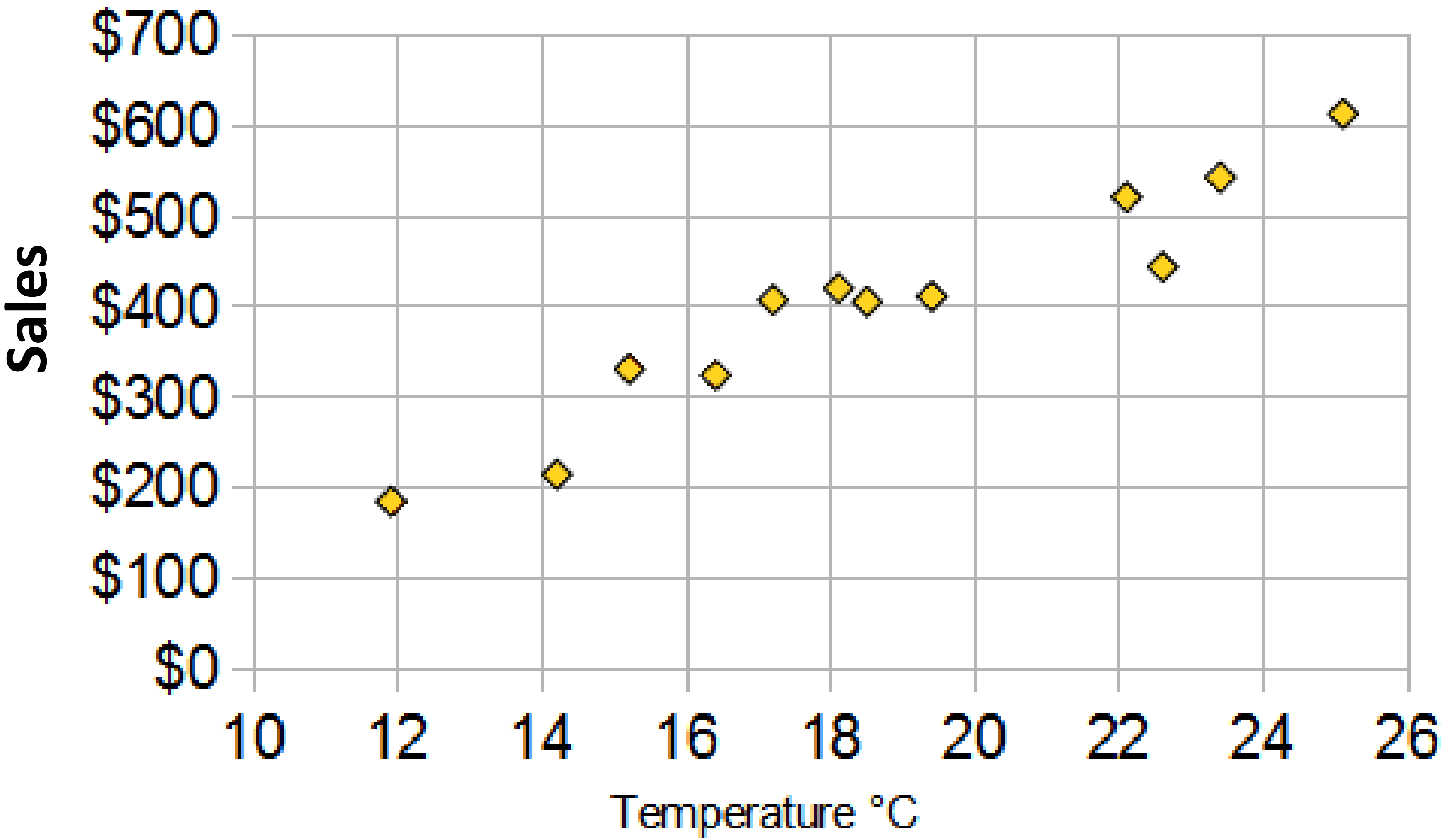
Here’s another example.

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



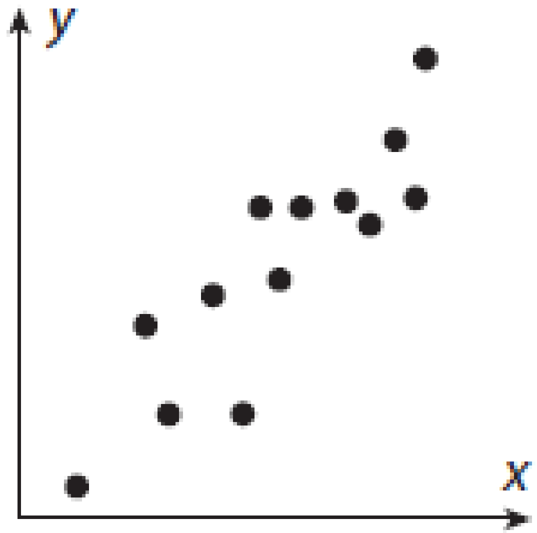
Correlation

P. 435

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

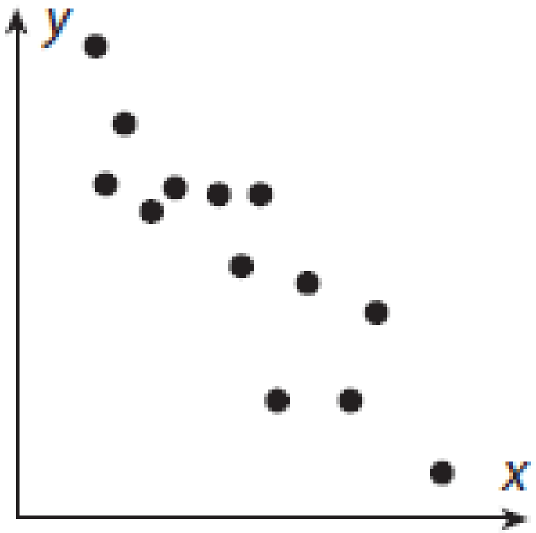
A correlation is a measure of the strength and direction of the relationship between 2 variables.

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



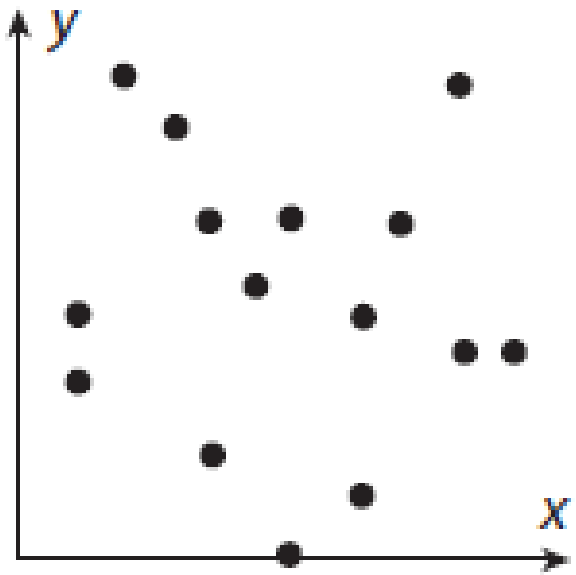
Positive correlation

Positive slope



Negative correlation

Negative slope

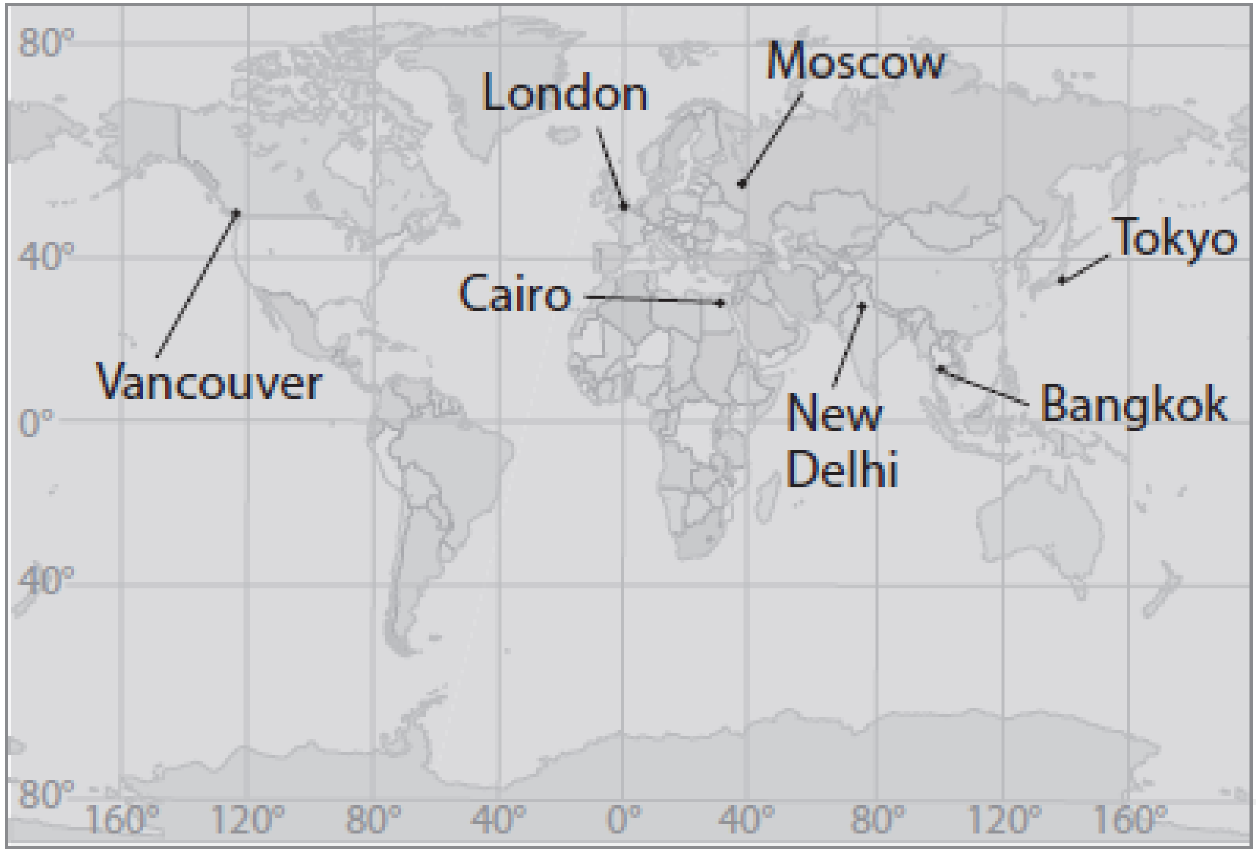


No correlation

P. 435

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

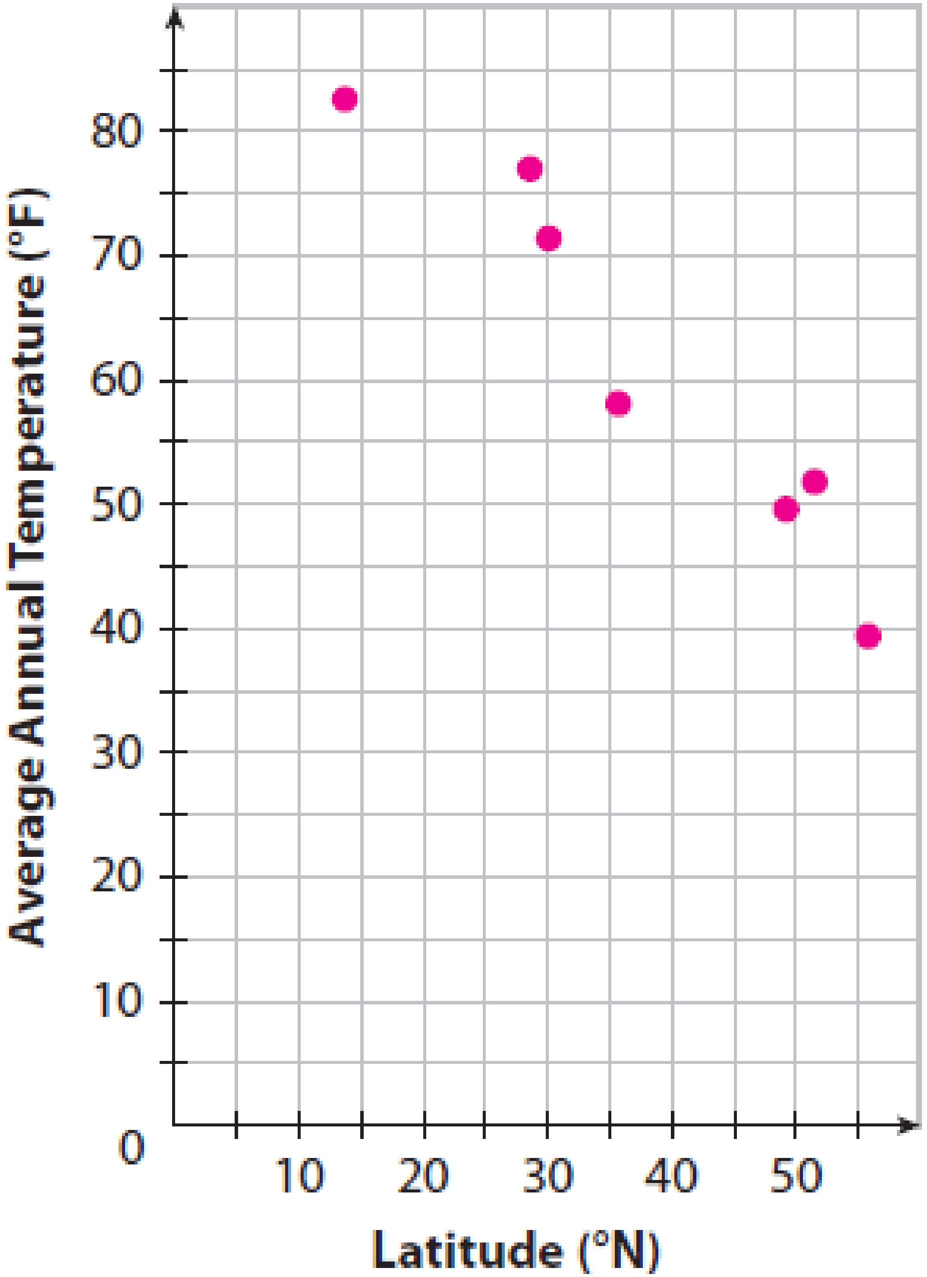


The two variables are Latitude and Temperature.

Here the ordered pairs just look like regular numbers. But when you graph them…

B Plot the data on the grid provided.

P. 436



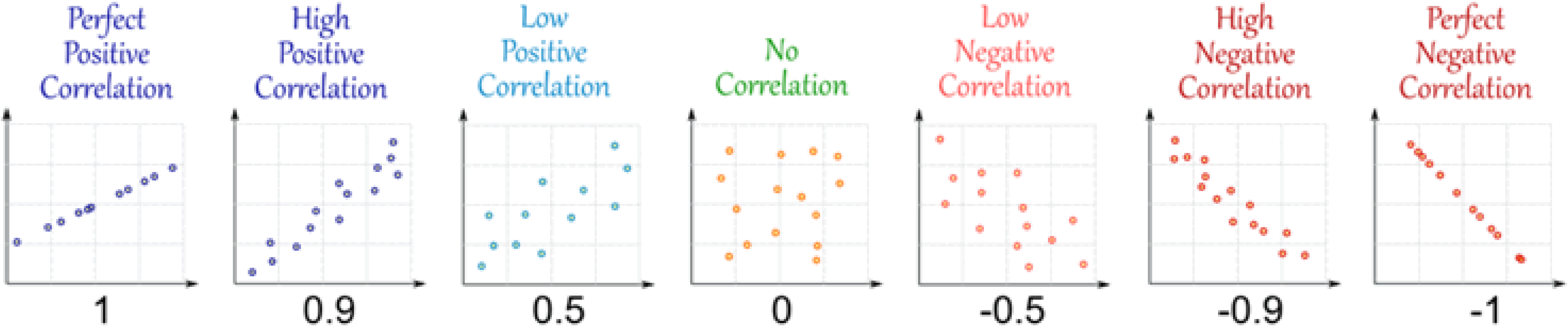
The ordered pairs are:
(13.7,82.6) for Bangkok
(30.1,71.4) for Cairo
(51.5,51.8) for London
(55.8,39.4) for Moscow
(28.6,77.0) for New Delhi
(35.7,58.1) for Tokyo
(49.2,49.6) for Vancouver

C The variables are negatively correlated.

P. 436

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:



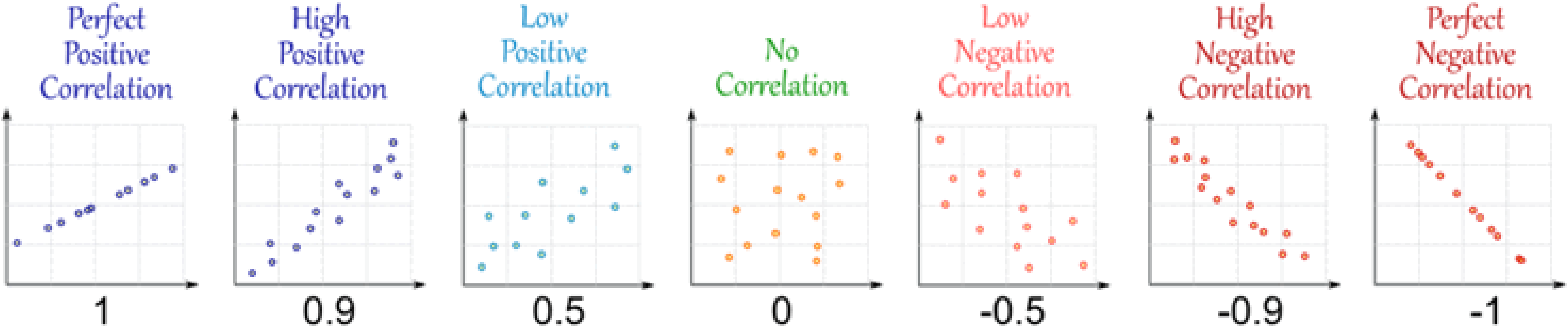
The values are **very** linked

The values don't seem linked at all

The values are **very** linked

The value shows **<u>how good the correlation is</u>** - not how steep the line is – and if it’s positive or negative.

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

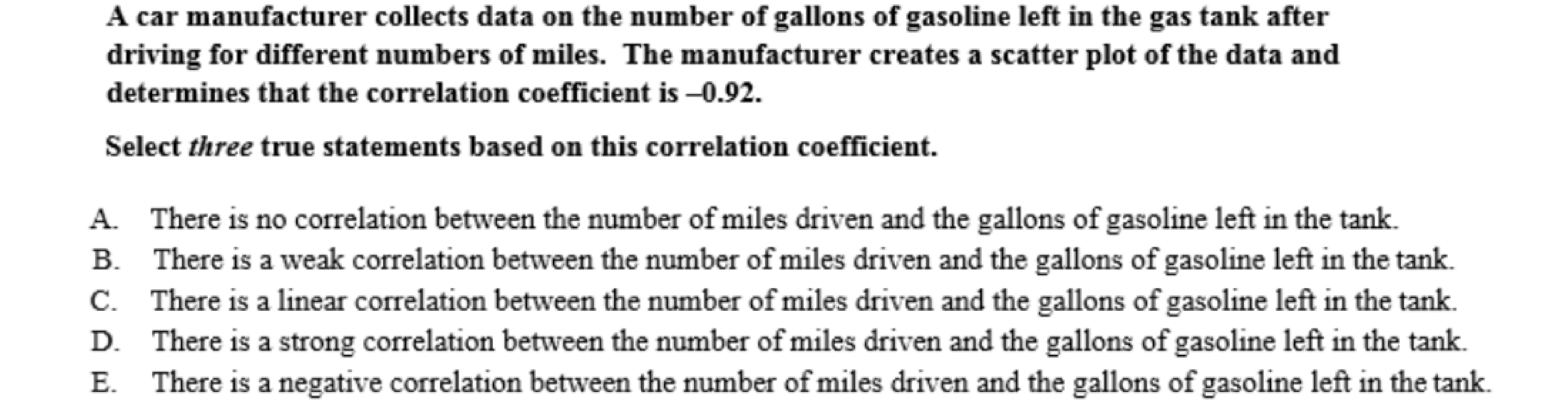


Strongly correlated data points look more like points that lie in a straight line,
and have values of *r* that are closer to 1 or -1.
Weakly correlated data points are spread out and will have values of *r* closer to 0.

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 math 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.

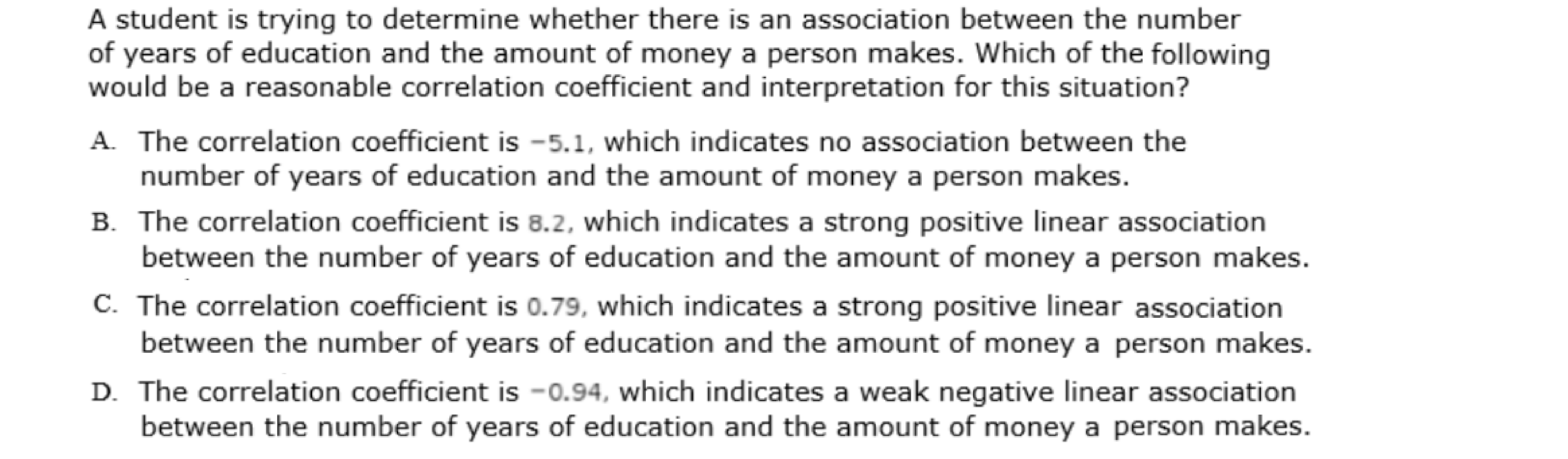
A. There is no correlation between the number of miles driven and the gallons of gasoline left in the tank.

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.



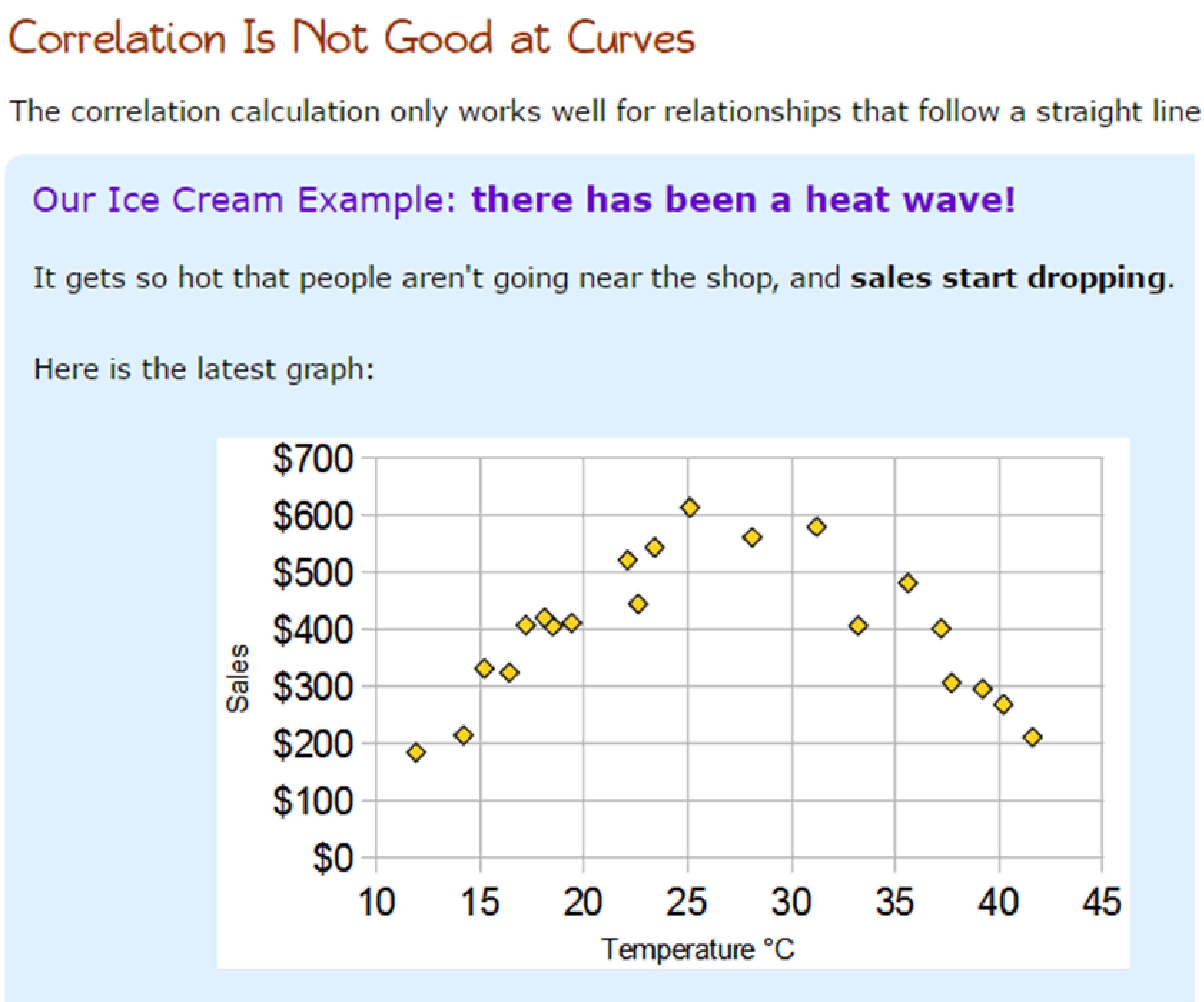
A student is trying to determine whether there is an association between the number of years of education and the amount of money a person makes. Which of the following would be a reasonable correlation coefficient and interpretation for this situation?

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**.

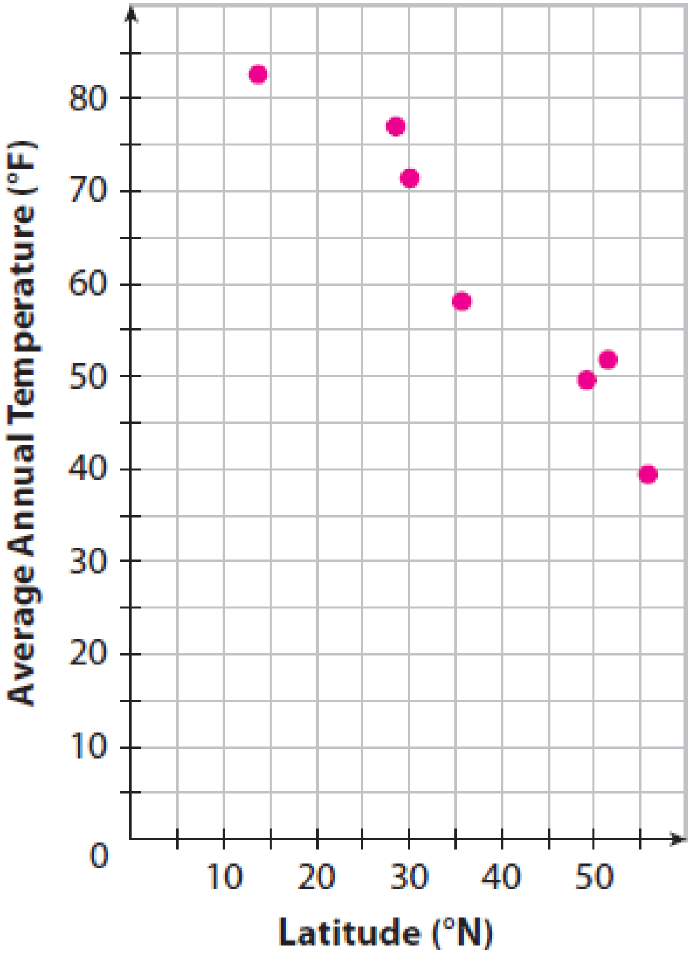
Here is the latest graph:

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

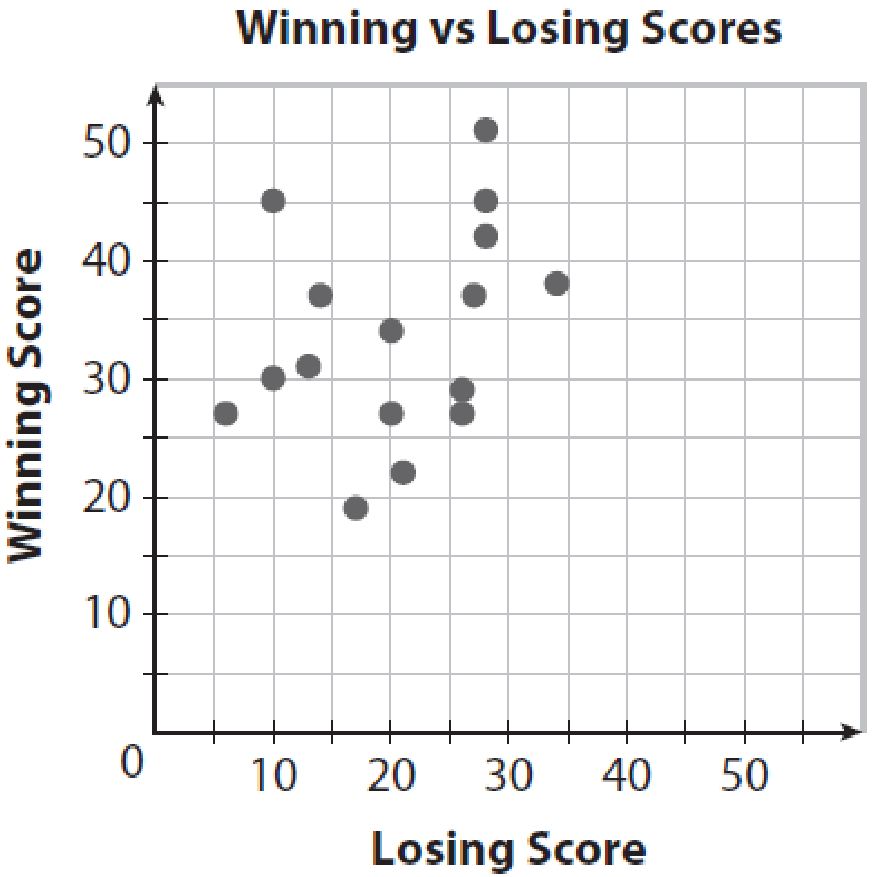
But the linear correlation calculation isn't "smart" enough to see this; it's value is 0, which means "no correlation".

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.

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.

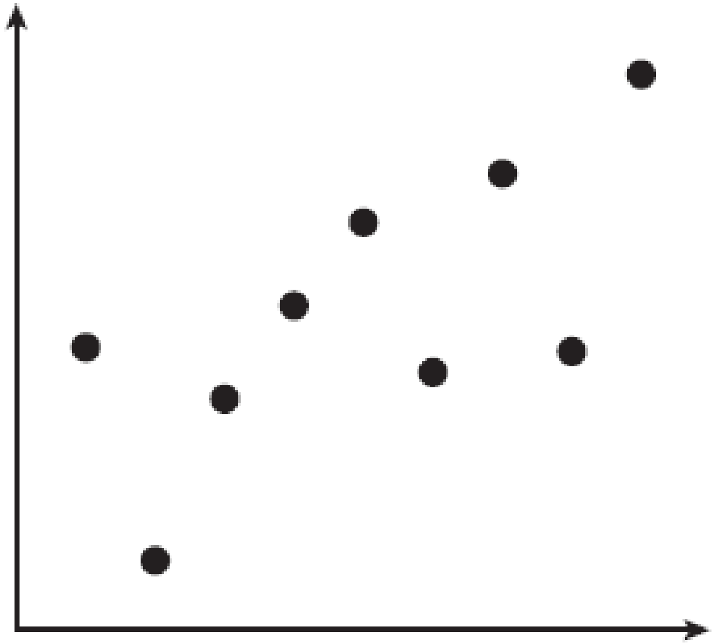


This data represents the football scores from one week with winning score plotted versus losing score.

r is close to 0.

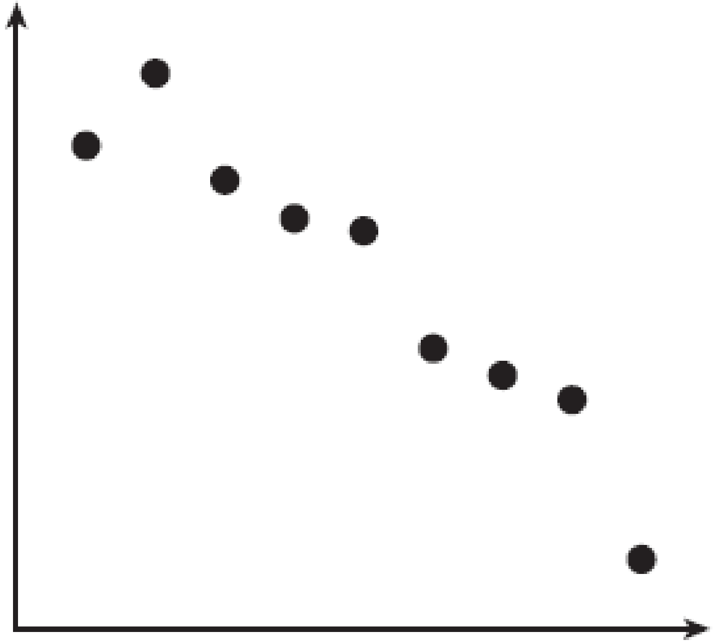
Your Turn

2.



r is close to

3.



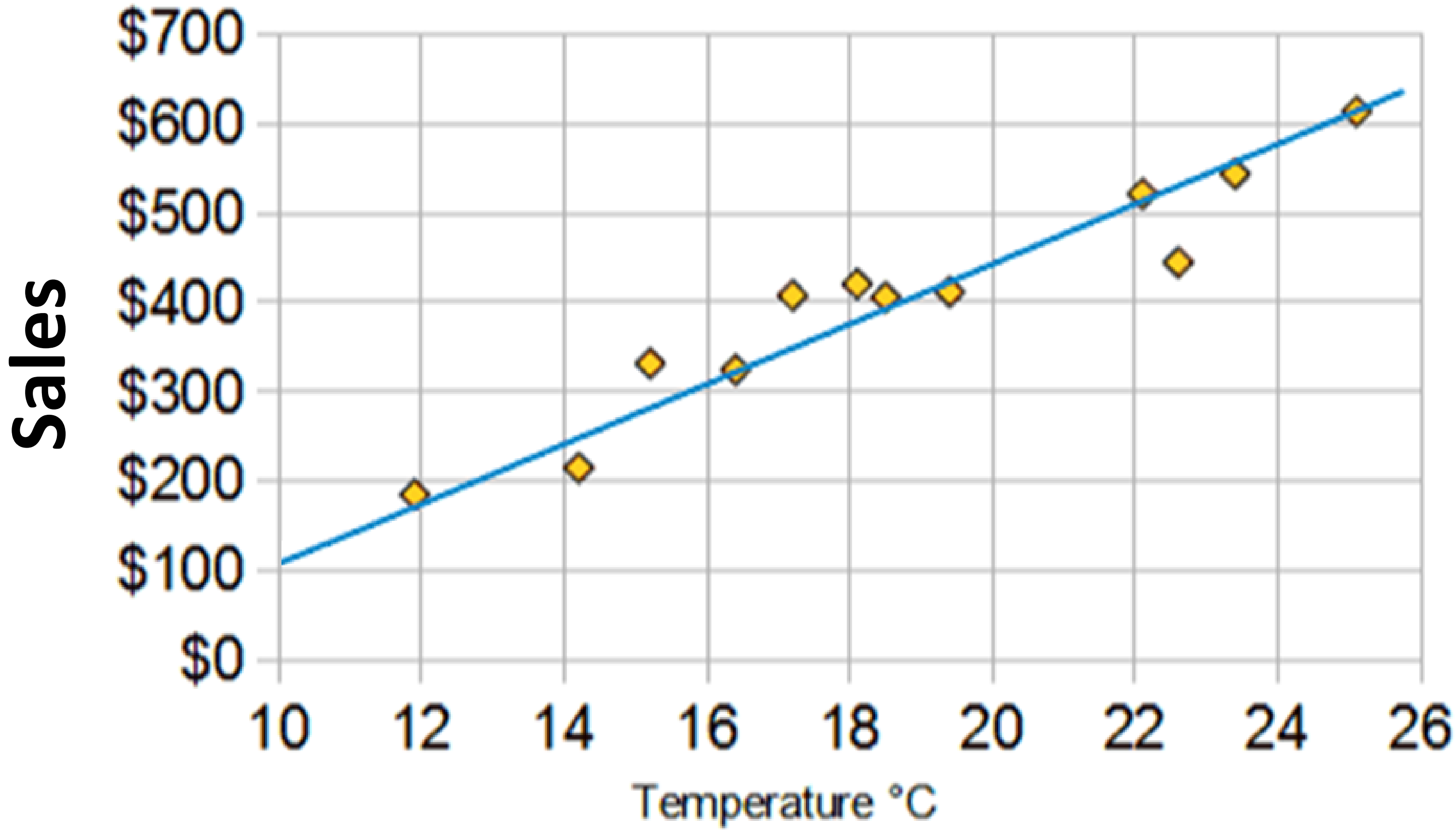
r is close to

Line Of Fit a.k.a. Line Of Best Fit a.k.a. Trend Line a.k.a. Regression Line a.k.a. Least Squares Line

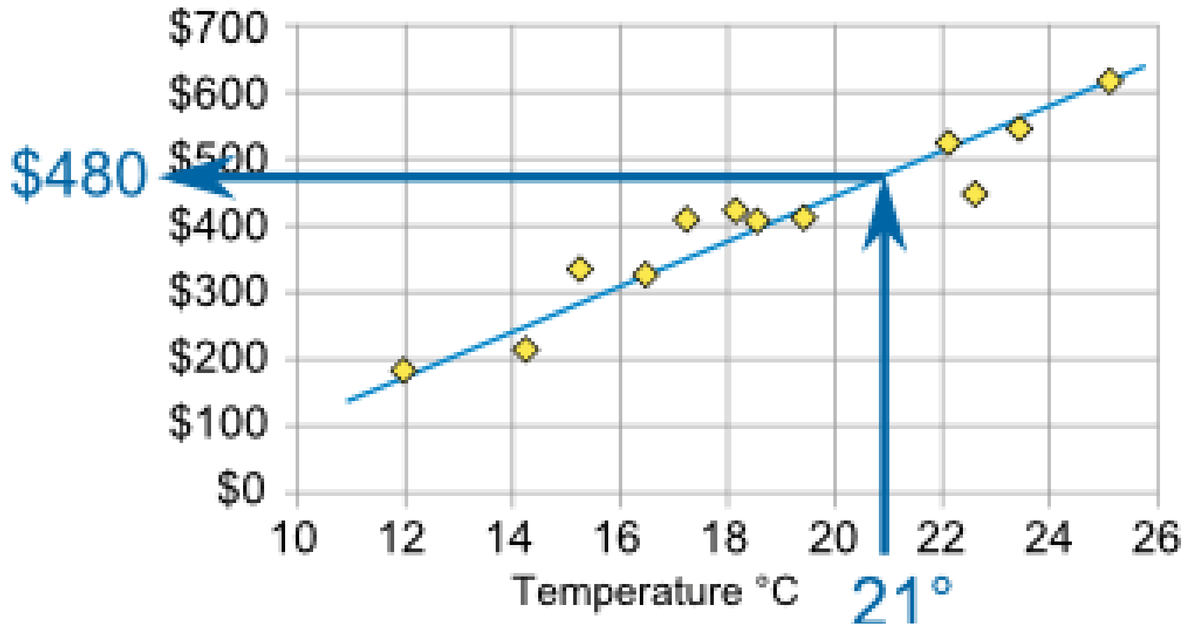
P. 438

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

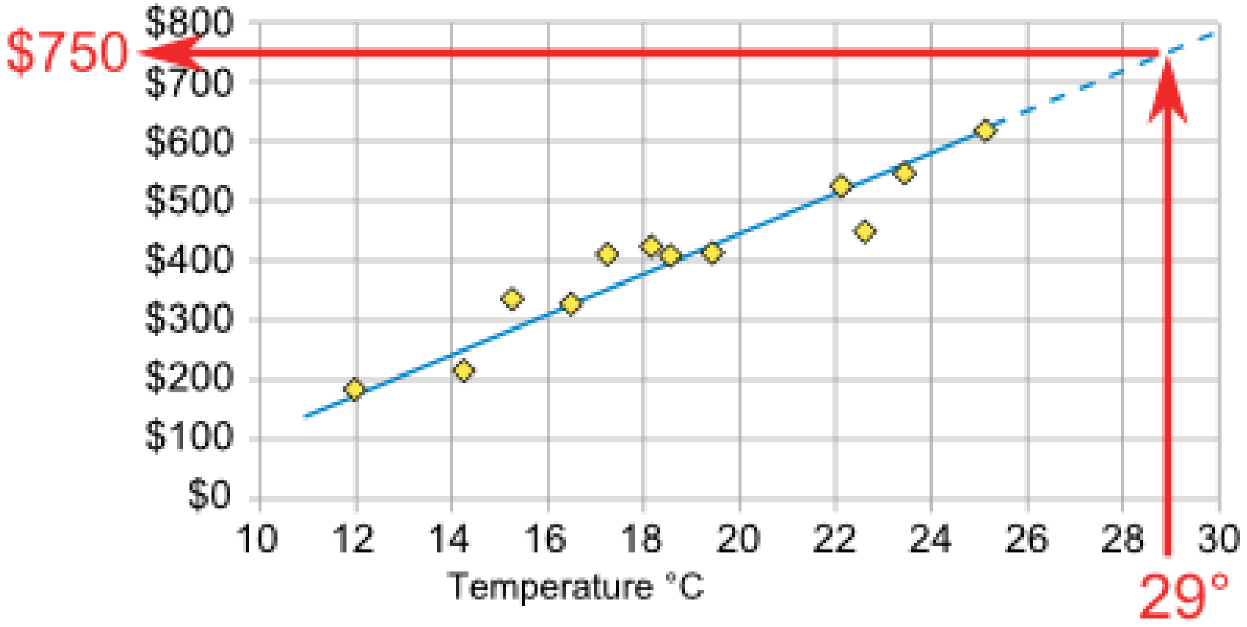
Remember the graph of the ice cream shop's sales?



Interpolation and Extrapolation



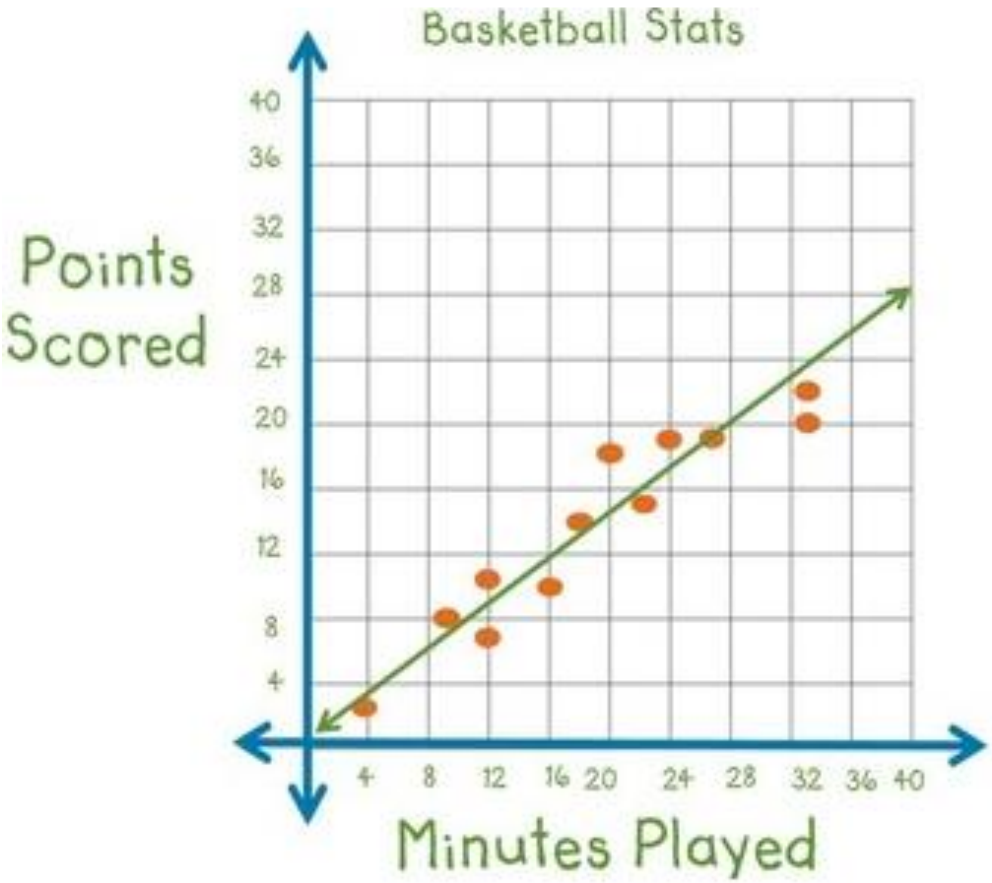
Interpolation is where we find a value **inside** our set of data points. Here we use it to estimate the sales at 21 °C.

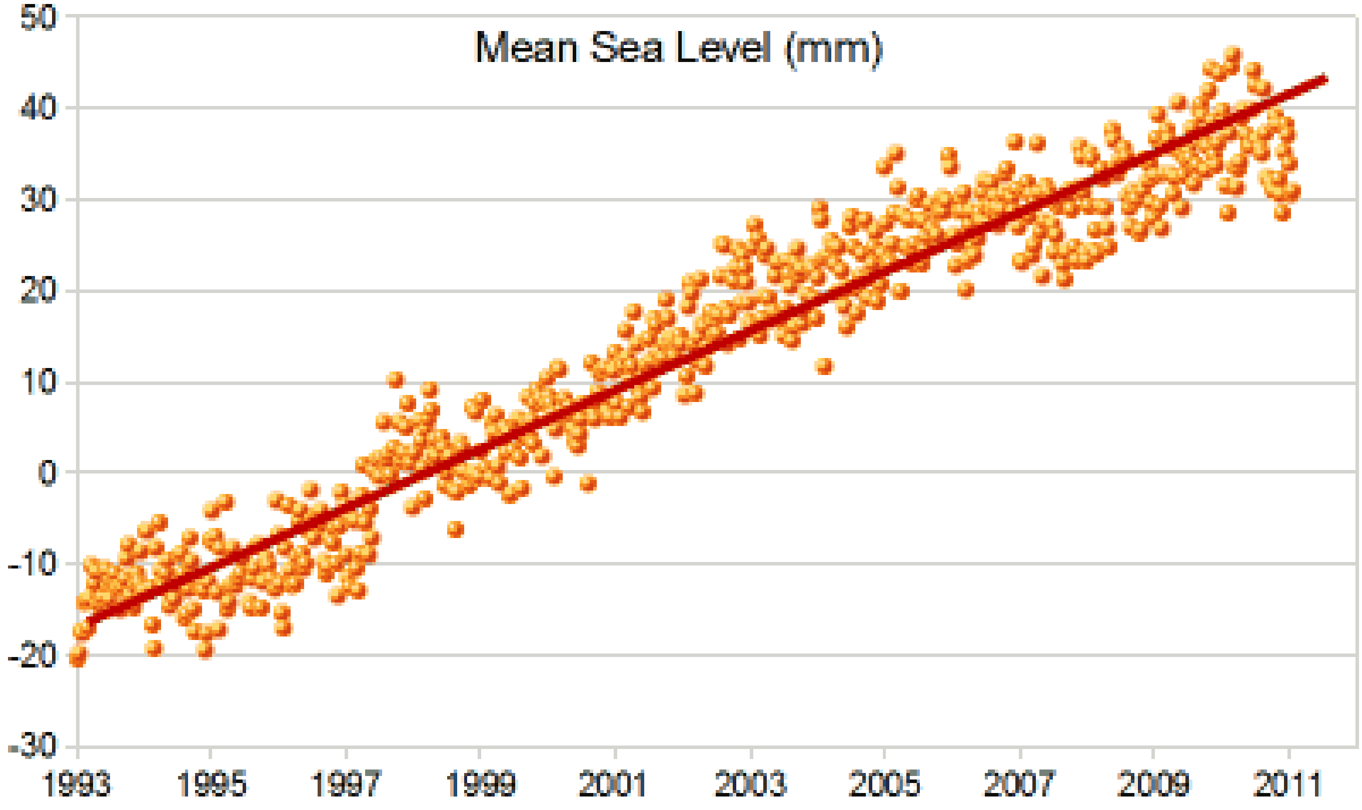


Extrapolation is where we find a value **outside** our set of data points.
Here we use it to estimate (predict) the sales at 29 °C (which is higher than any value we have).

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

Using a ruler, draw the line that the data points appear to be clustered around. It's 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, so as many as there are below the line are also above the line. There's no perfect line to draw. The more the points are spread out, the more lines of fit that can be drawn.





Your Turn

P. 440

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?

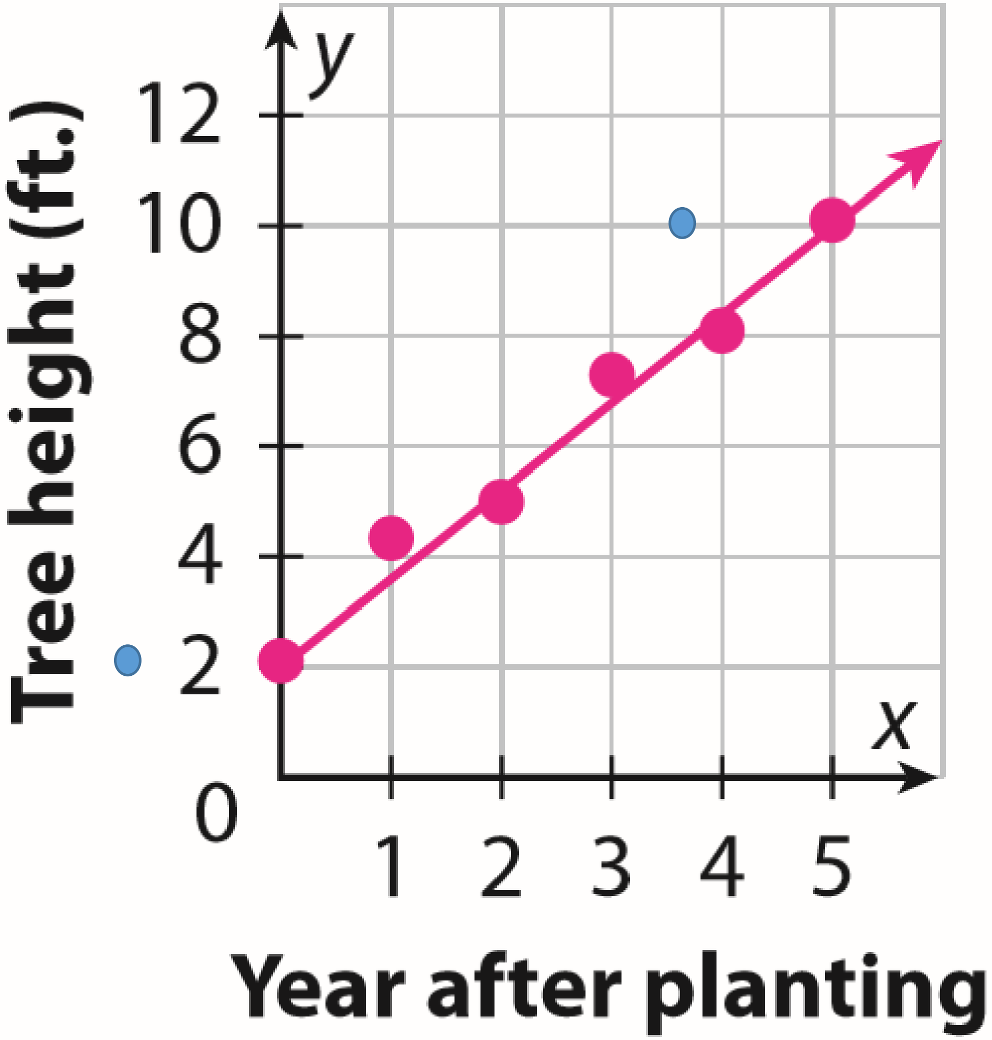
$$y = mx + b$$

Rise =

Run =

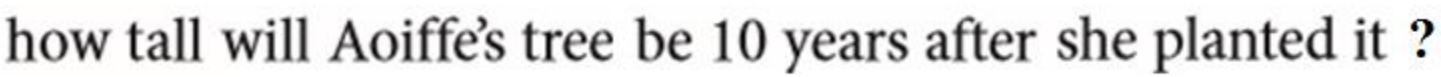
Y-intercept =

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



Equation for Line Of Fit is $y = \frac{8}{5}x + 2$

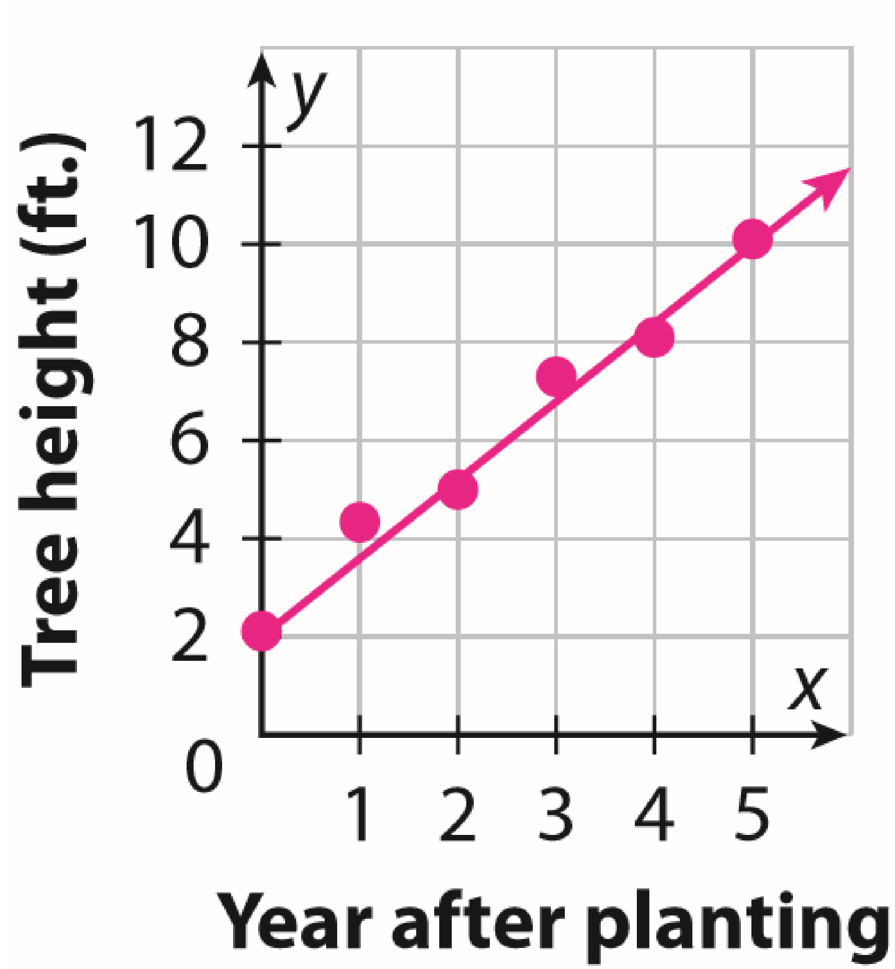
P. 441



how tall will Aoiffe's tree be 10 years after she planted it ?

Plug it in!

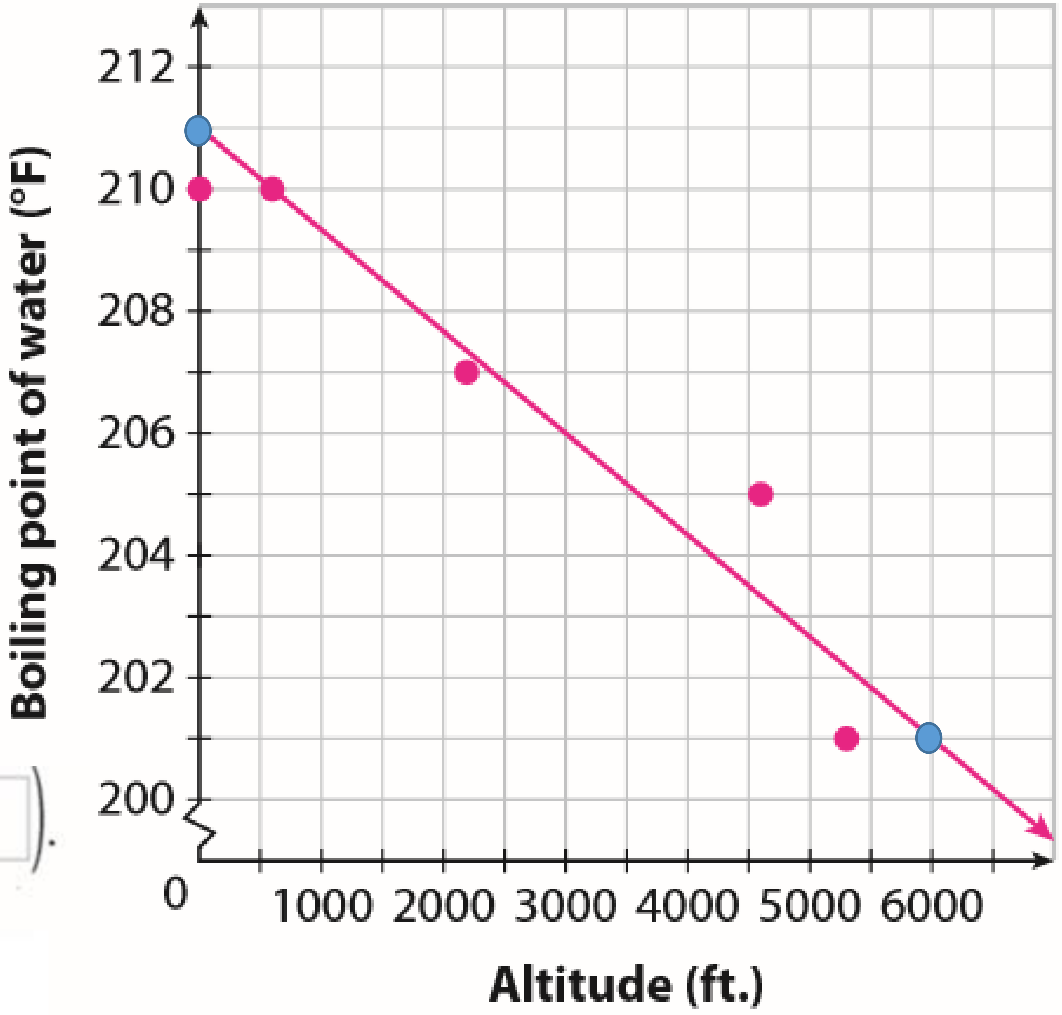
Is this Interpolation or Extrapolation?



P. 439

(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 $(\boxed{0}, \boxed{211})$ and $(\boxed{6000}, \boxed{201})$.

$$m = \frac{rise}{run} = \frac{y_2 - y_1}{x_2 - x_1} = \frac{\boxed{-10}}{\boxed{6000}} = -0.00167$$

$b = \boxed{211}$ so The equation is of this line of fit is $\boxed{y = -0.00167x + 211}$

Ⓑ 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)

P. 441

$$y = -0.00167x + 211$$

Mexico City: $y =$ □ □ + □ $= 197.74$

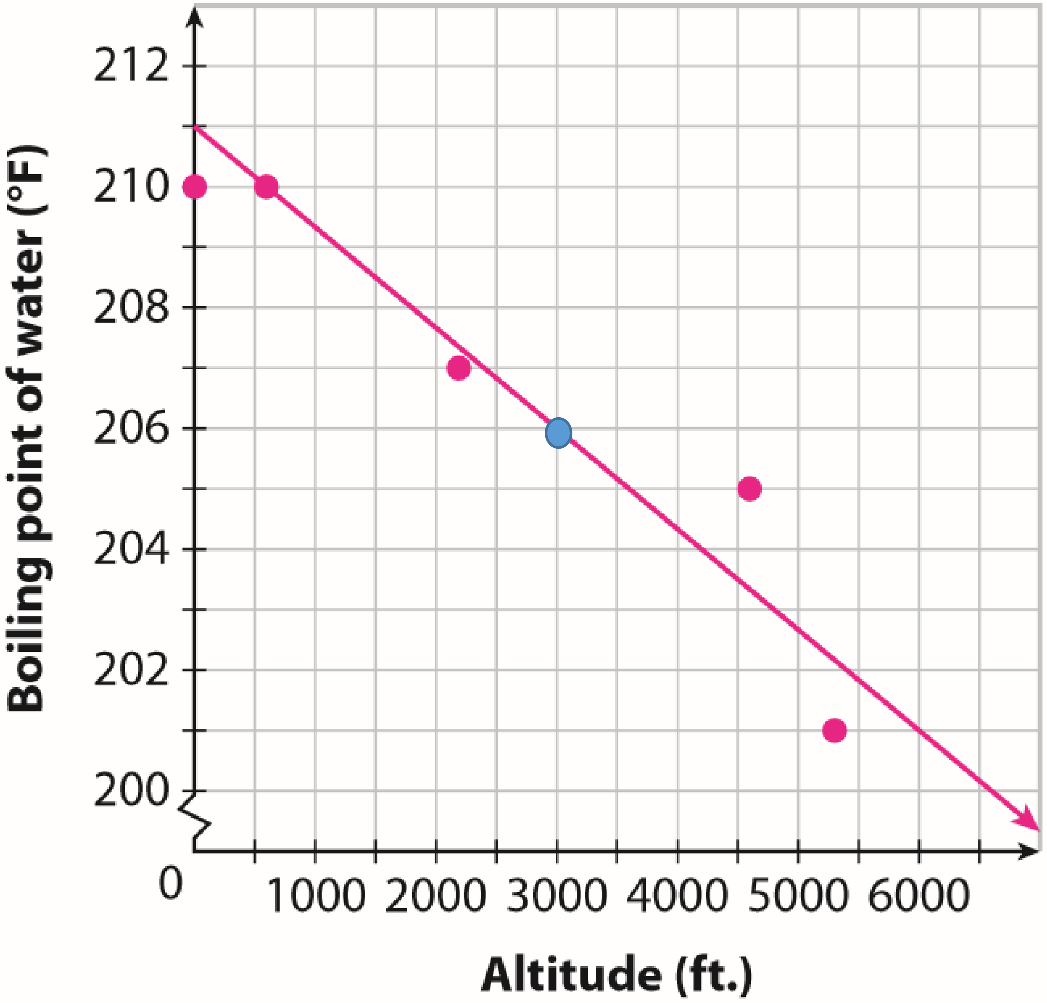
Fargo: $y =$ □ □ + □ $= 205.99$

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

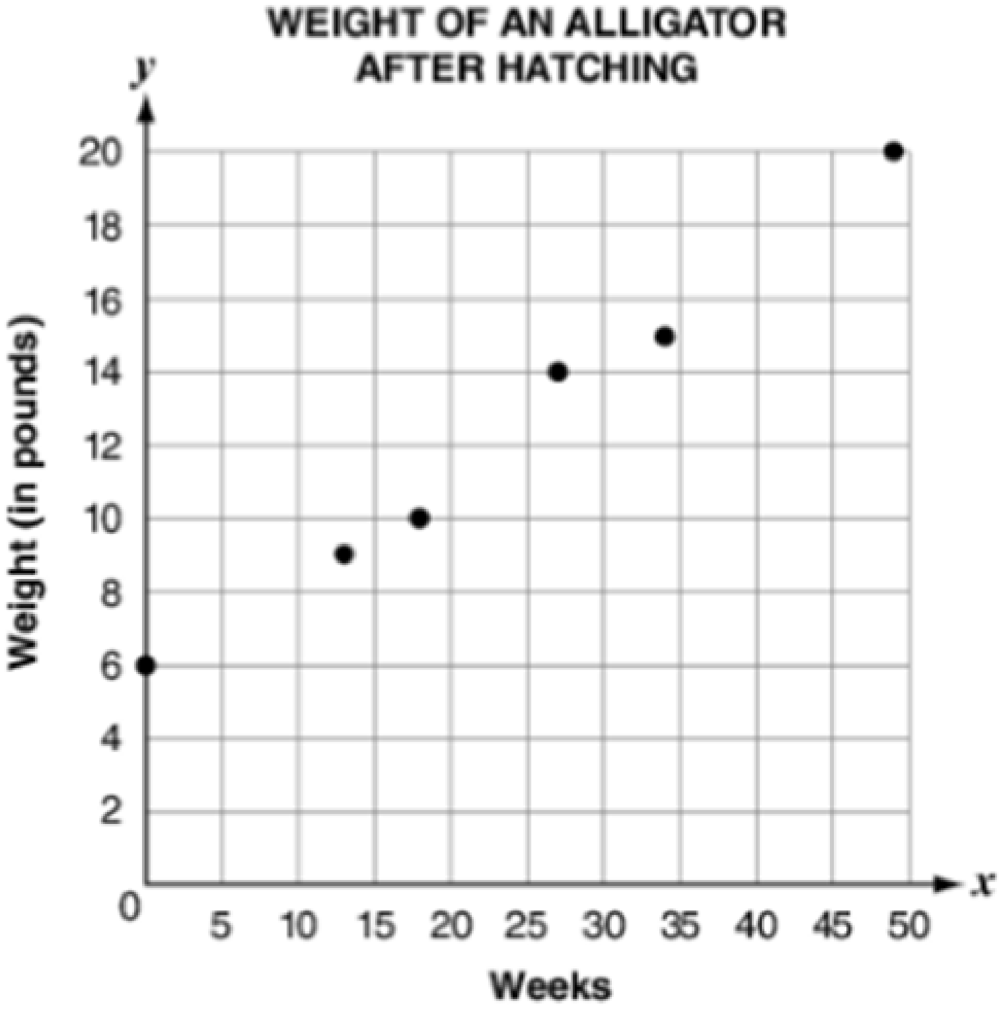
The boiling point in Fargo, North Dakota is a more reliable prediction because it is an interpolation, while Mexico City is an extrapolation.

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.



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



A. $w = 0.25n + 6$

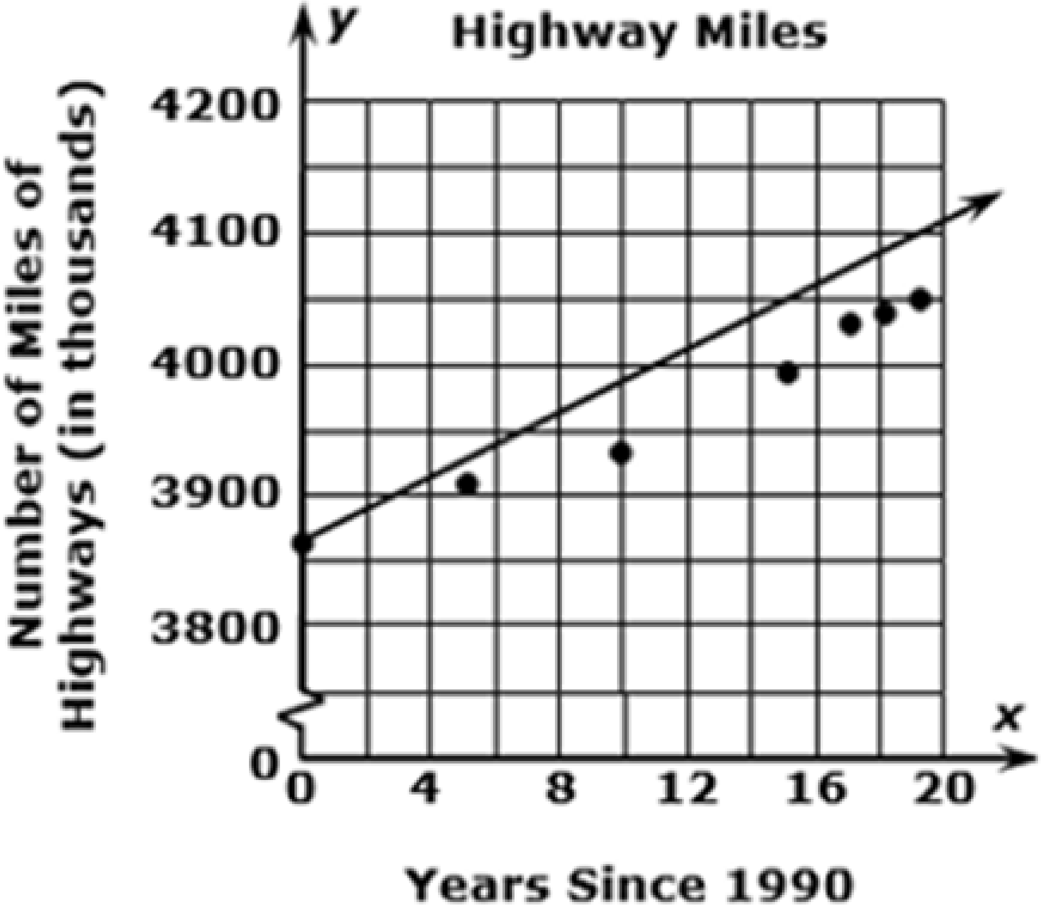
B. $w = 0.65n + 6$

C. $w = 6n + 0.25$

D. $w = 6n + 0.65$

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

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?

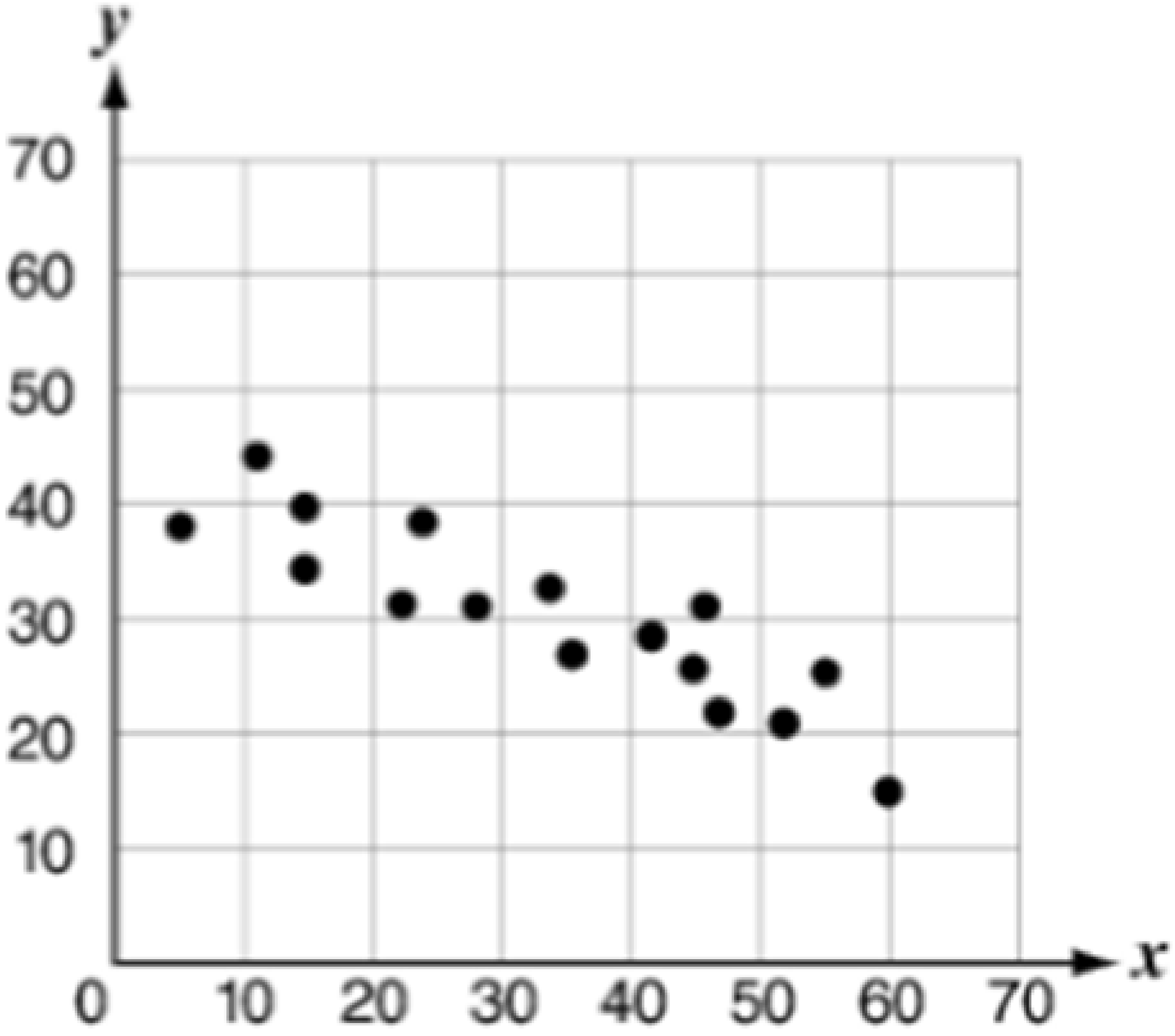
A. The line drawn by the student represents an adequate fit as the line passes through the first plotted data point.

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.

Which function best fits the data in this scatter plot?



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

B. $y = -2x + 45$

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

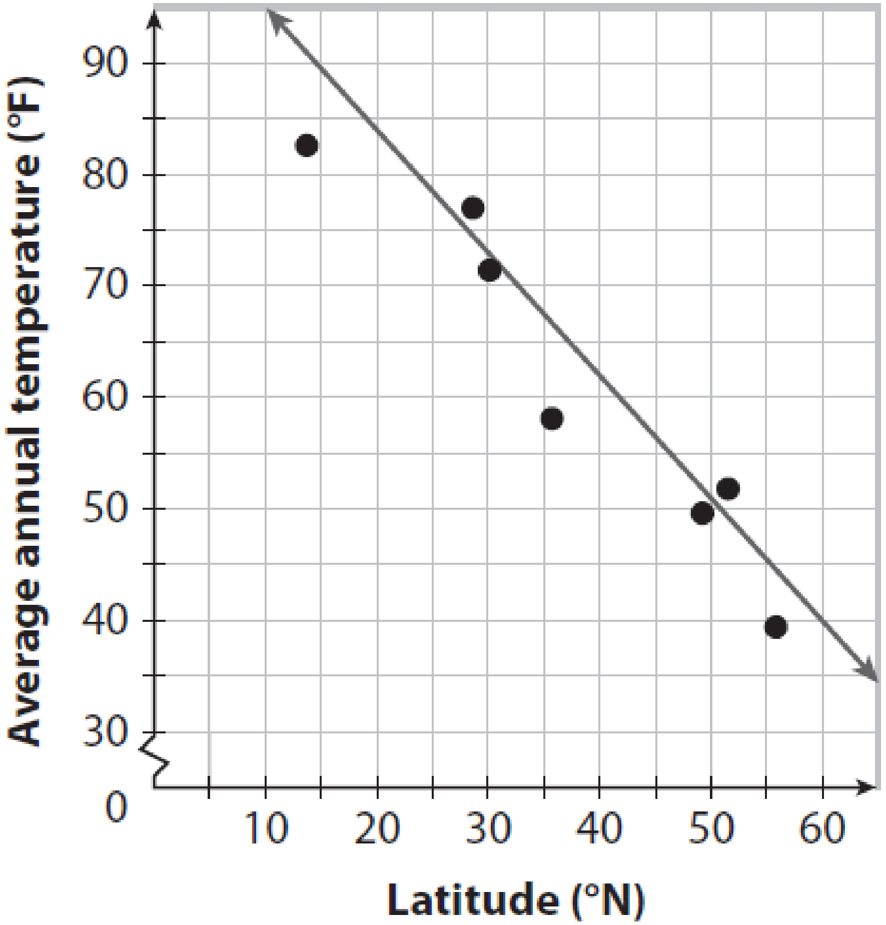
D. $y = -4x + 45$

To determine the Line of Fit, you need the slope and the y-intercept.
But what if the Line of Fit doesn't reach the *y* axis? How can you find the y-intercept?

P. 438

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

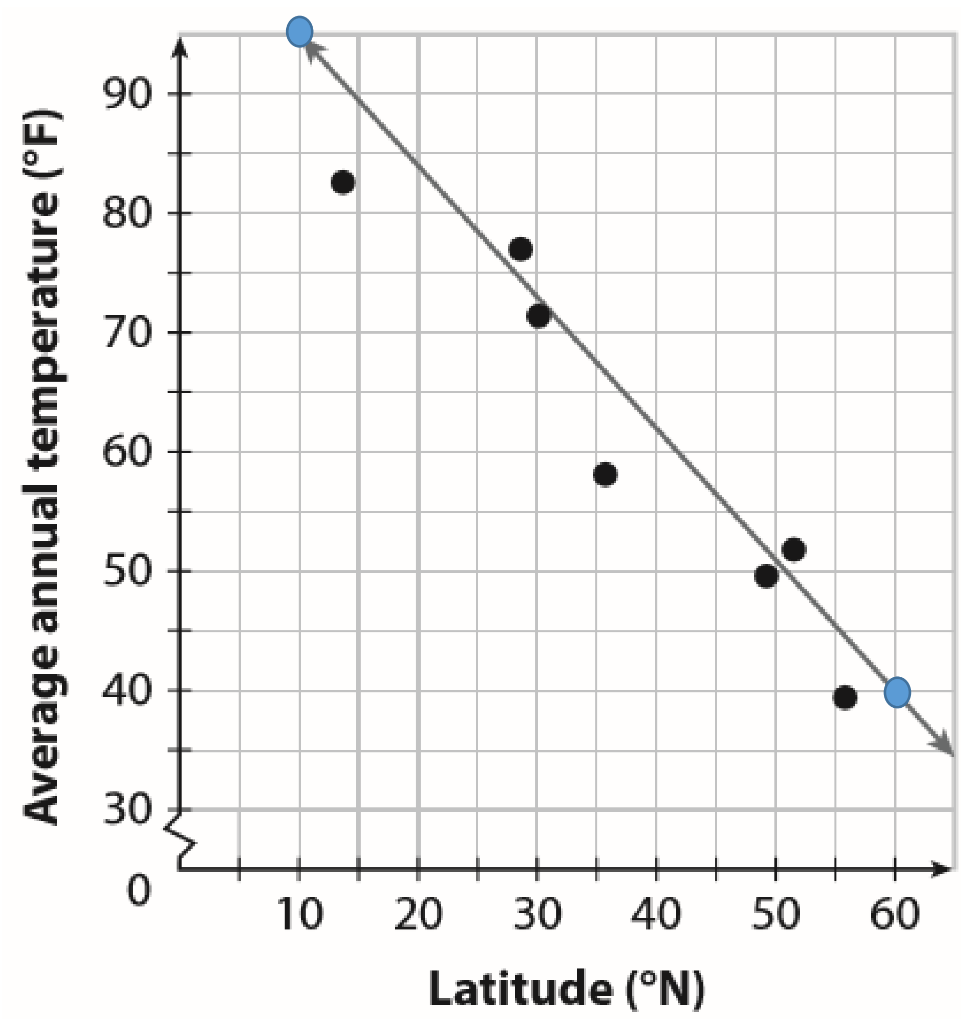
Ⓐ Go back to the scatter plot of city temperatures and latitudes and add a line of fit.



Once a Line of Fit has been drawn onto the scatter plot:

P. 438

- **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.**
- **Plug in any point's *x* and *y* to determine the *b* (y-intercept).**
- **Write the equation for the line.**



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.

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

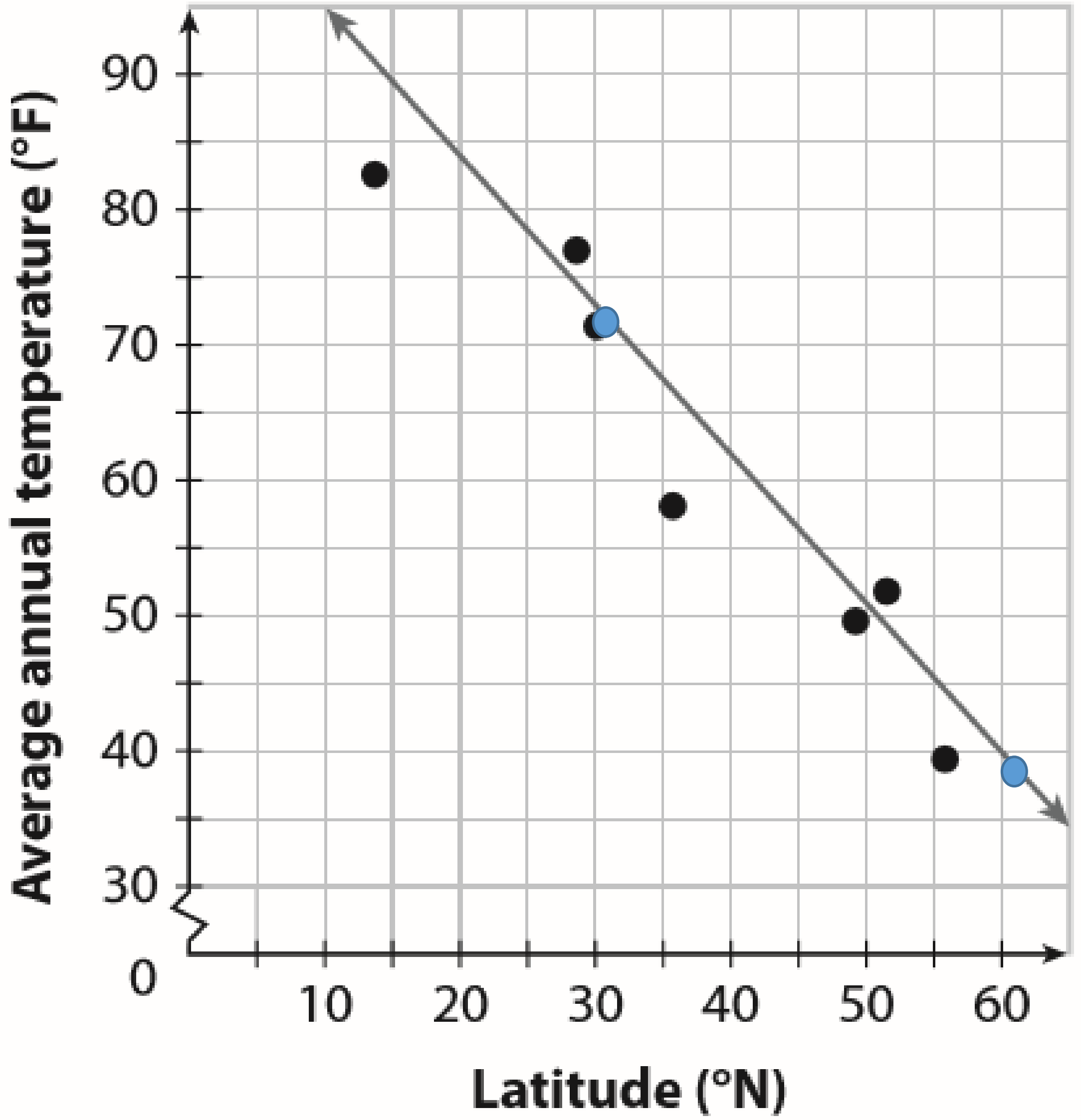
P. 440

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

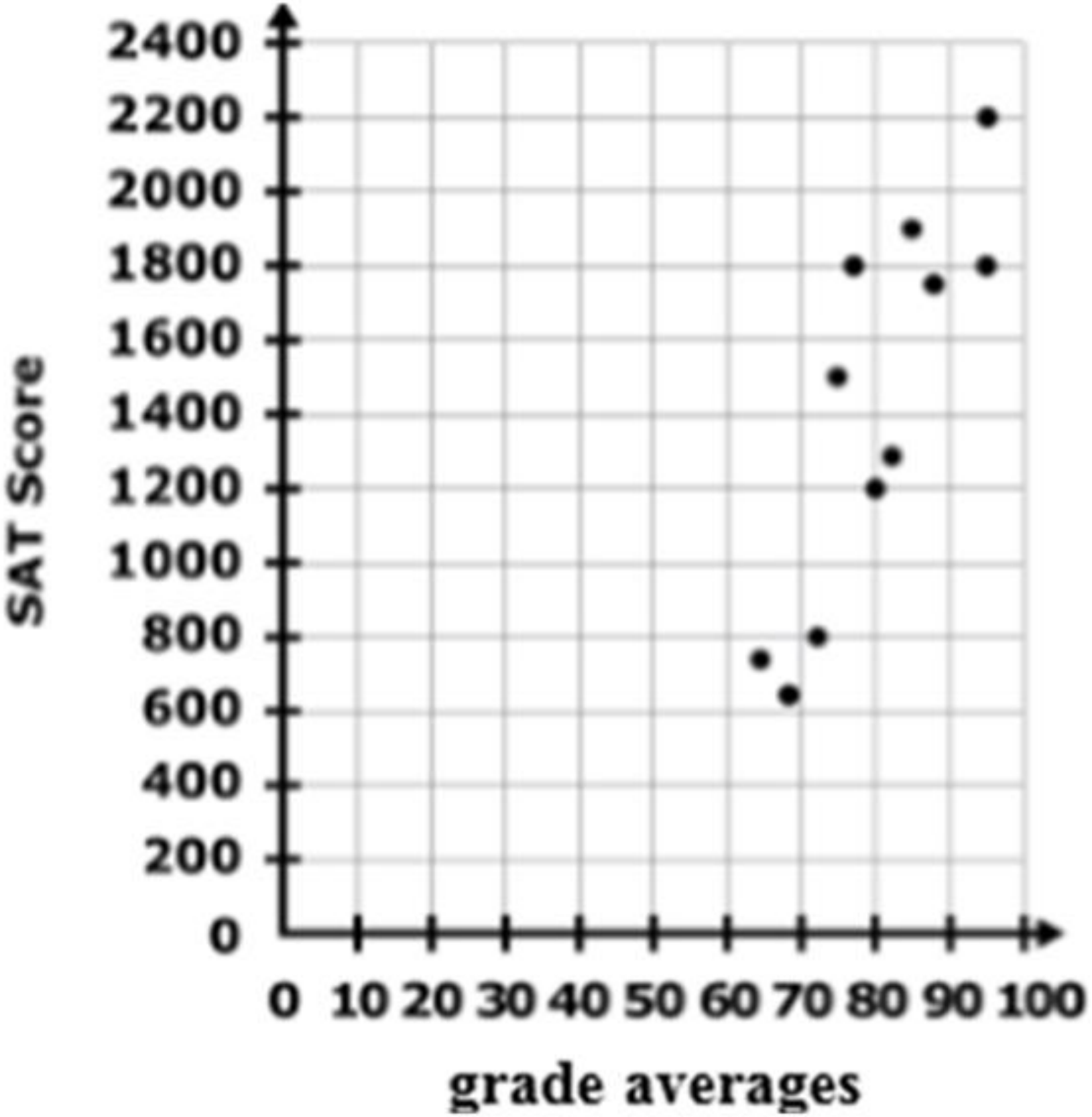
$$y = -1.1x + 106$$

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

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



The correlation of (old-style) SAT scores and grade averages for a high school is represented by the scatterplot below.

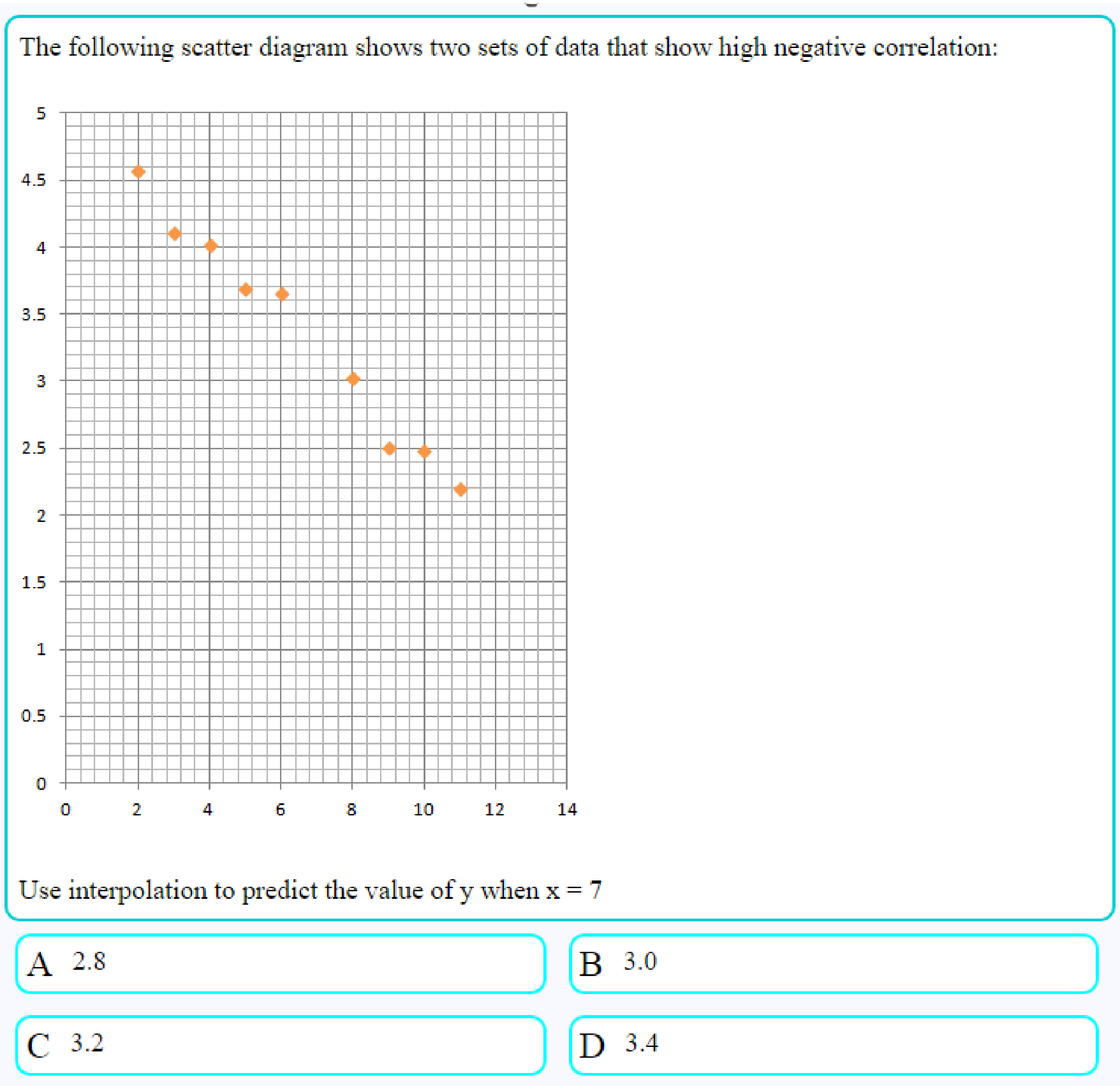


1) Calculate a Line Of Fit.

2) Based on your line (trend) – what's the likely lowest grade average?

3) Is this an **Interpolation** or **Extrapolation** ?

4) Someone calculated the Line of Fit to be $\boldsymbol{y = 40x - 1800}$. Based on that trend, what would be the likely SAT score for someone with a grade average of 95? Of 75?



The following scatter diagram shows two sets of data that show high negative correlation:

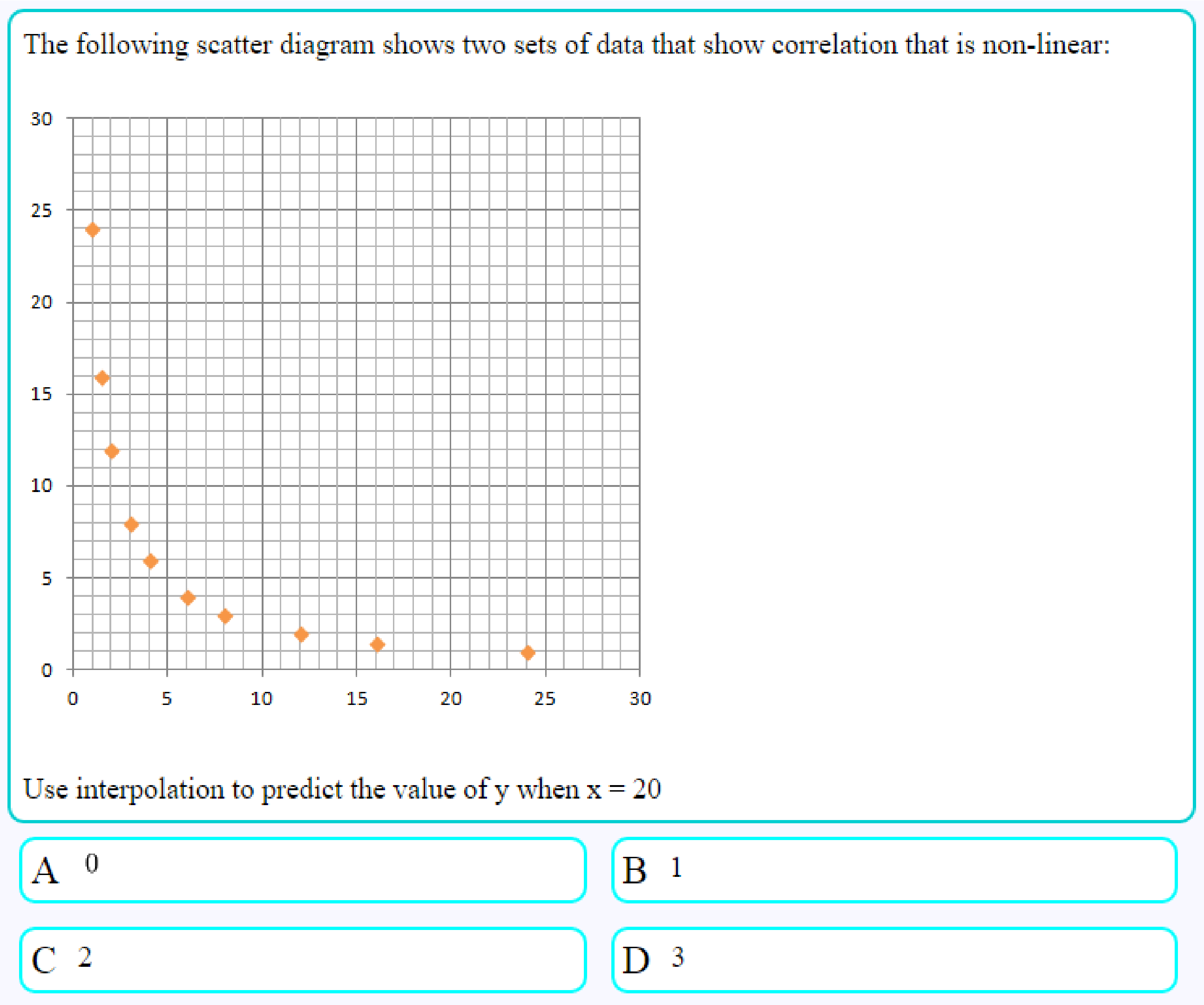
Use interpolation to predict the value of y when x = 7

A 2.8

B 3.0

C 3.2

D 3.4



The following scatter diagram shows two sets of data that show correlation that is non-linear:

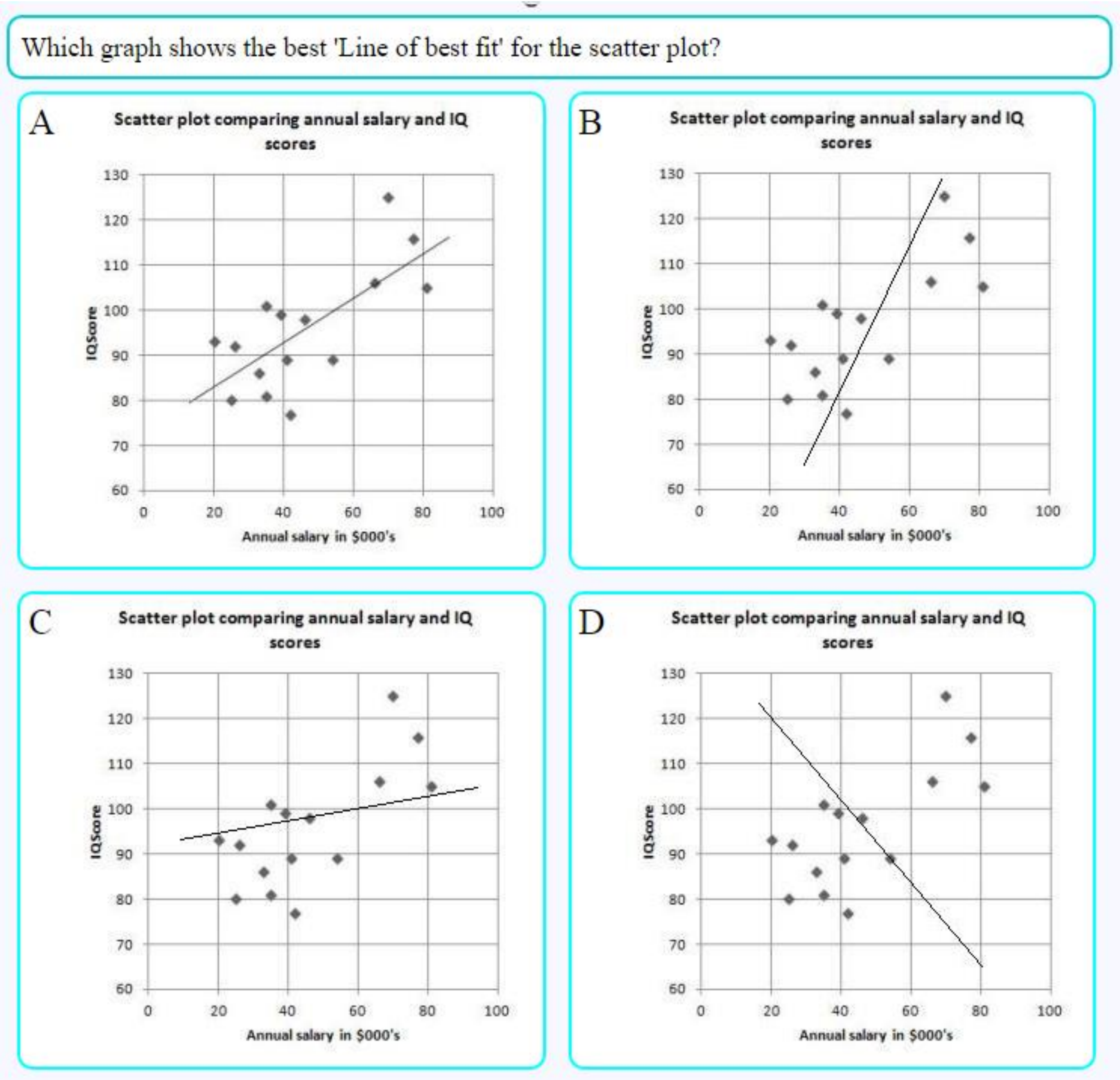
Use interpolation to predict the value of y when x = 20

A 0

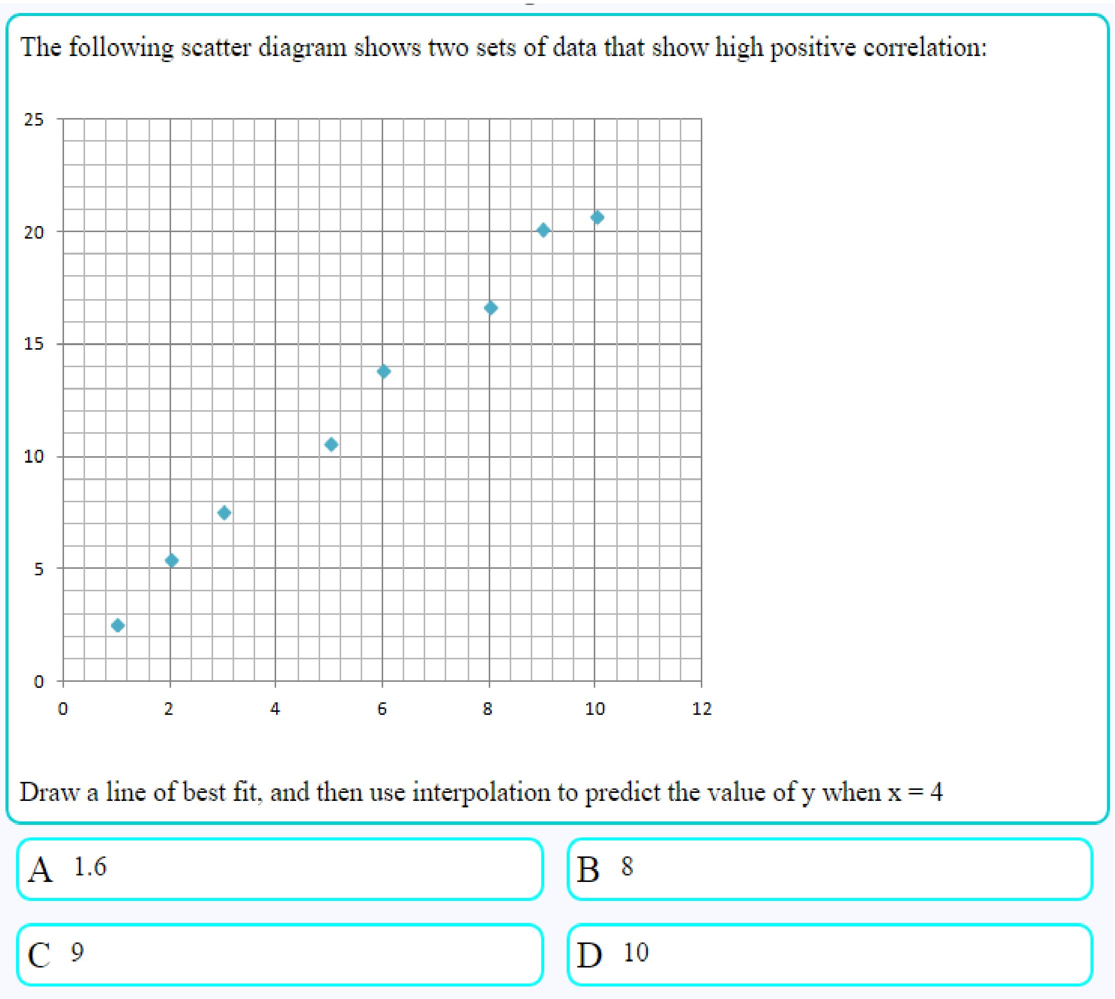
B 1

C 2

D 3



Which graph shows the best 'Line of best fit' for the scatter plot?



The following scatter diagram shows two sets of data that show high positive correlation:

Draw a line of best fit, and then use interpolation to predict the value of y when x = 4

A 1.6

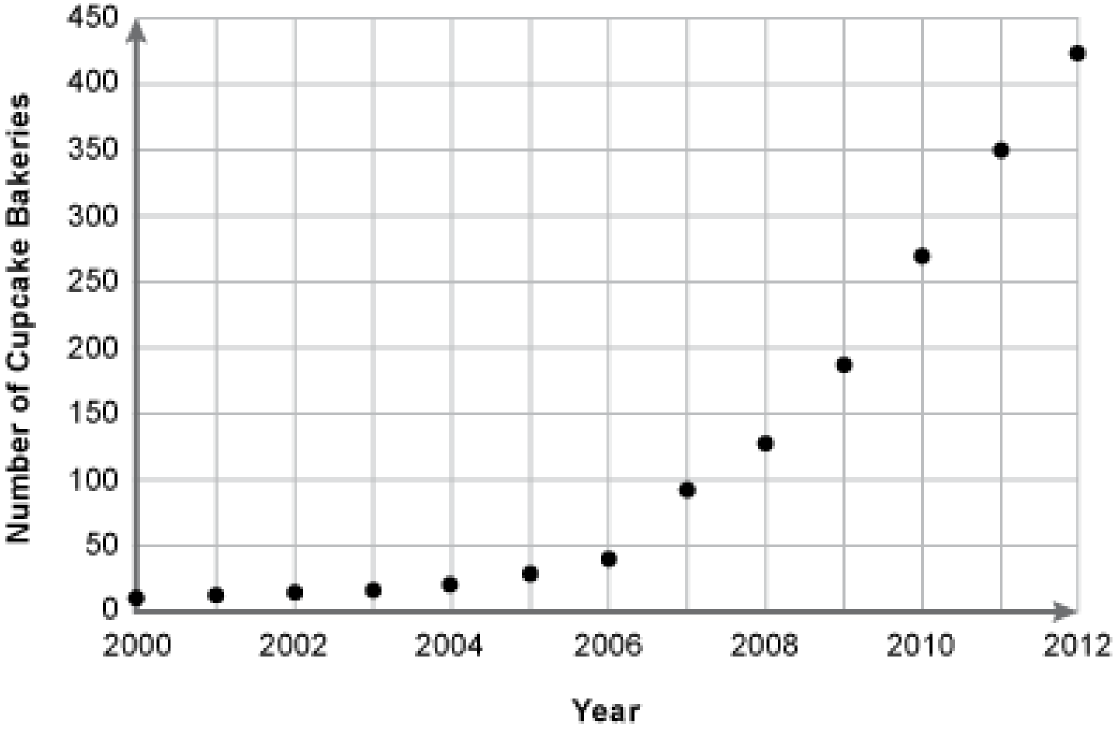
B 8

C 9

D 10

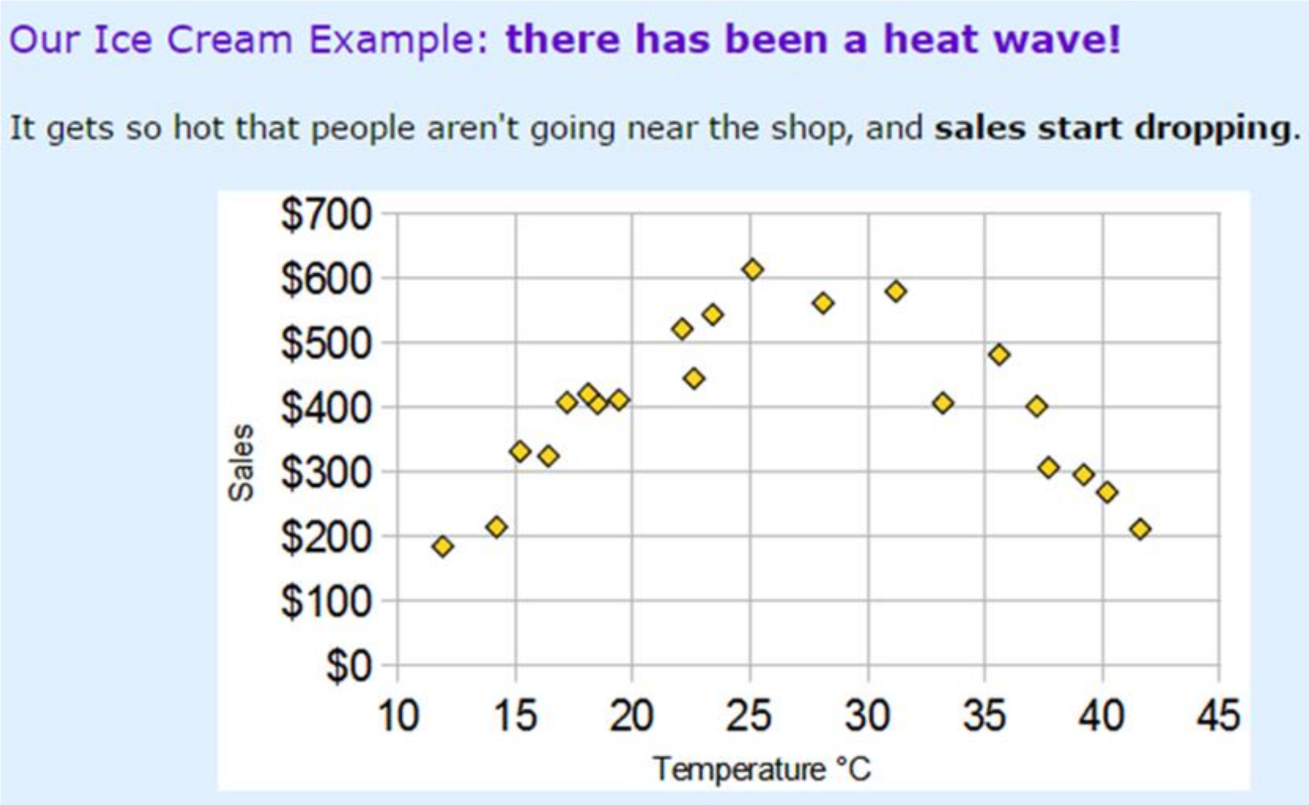
Exponential Model

Diagram 1 – The scatterplot shows the number of cupcake bakeries from 2000 through 2012.



Requires a Line-Of-Fit that's Exponential

Quadratic Model



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

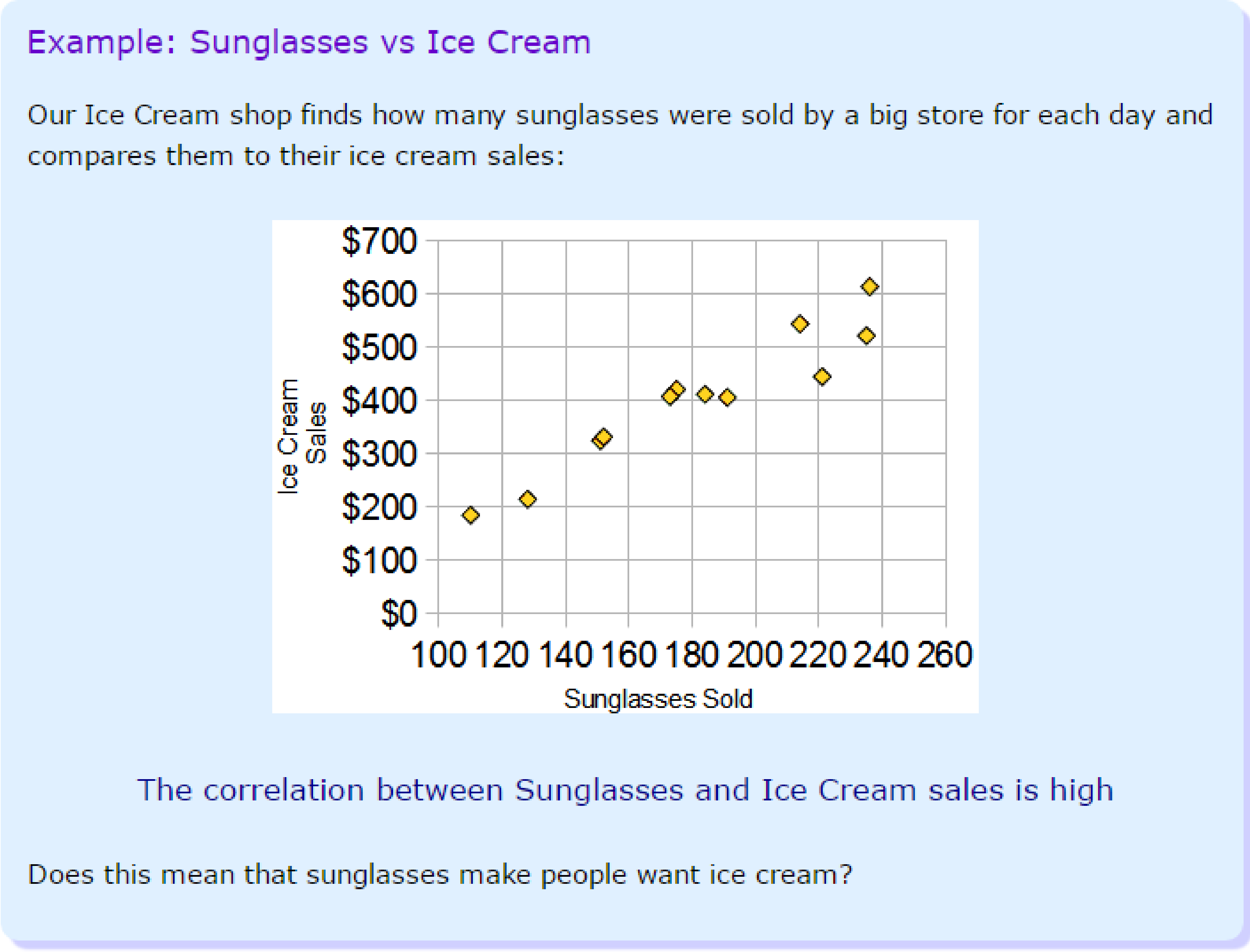
It gets so hot that people aren't going near the shop, and **sales start dropping.**

Requires a Line-Of-Fit that's Quadratic

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?

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

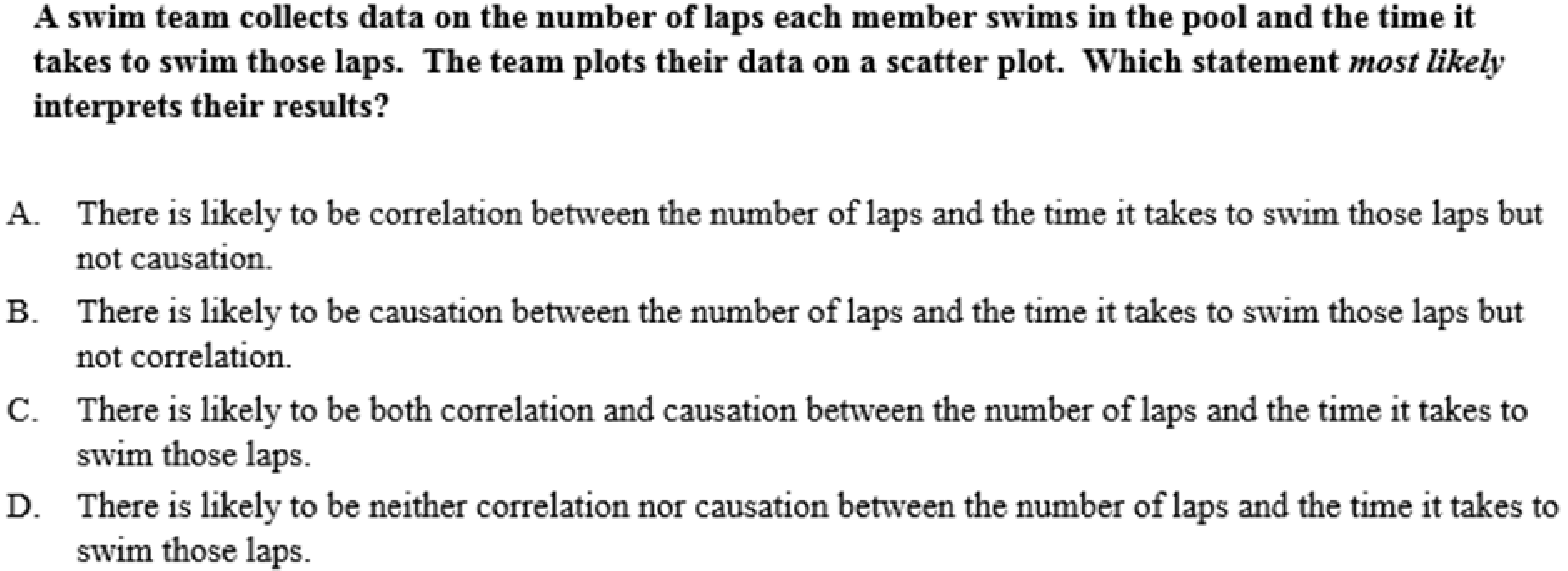
P. 441+

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 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.

P. 442

Can a data set be likely to show causation without showing a strong correlation?

No; a data set must show a strong correlation in order for causation to be likely.

Can a data set be likely to show correlation without showing causation?

Yes; for example, the number of sunglasses sold on a hot day plotted against the sales of ice cream.