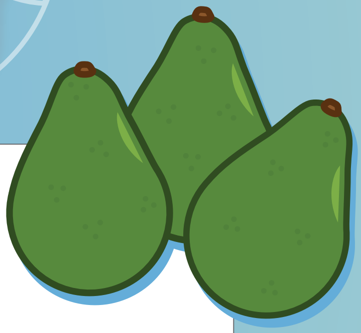


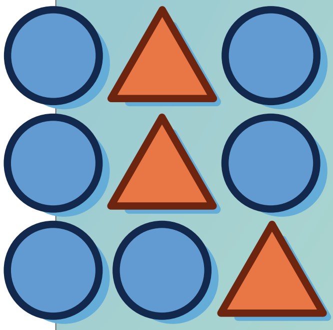
Unit 3

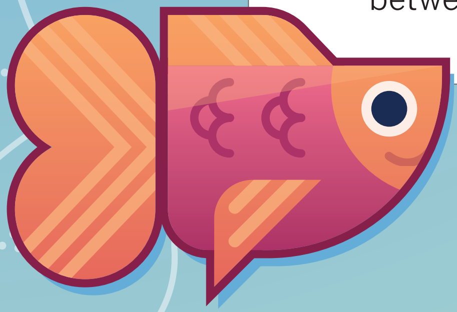
Describing Data

In this unit, you will use dot plots, histograms, box plots, and measures of center and spread to analyze and describe one-variable data sets. You will also use scatter plots, correlation coefficients, and lines of best fit to describe two-variable data sets.

Essential Questions

- What tools can you use to help determine if there is an association in a set of data?
- How can you compare data using measures of center and spread?
- How can you use the correlation coefficient and line of best fit to describe the relationship between two variables?





Summary | Lesson 1

Different types of questions lead to different types of data:

- **Quantitative data** has values that are numbers, measurements, or quantities instead of words. It's sometimes called *numerical data*.

  How many pets do you have? is a question that produces quantitative data.

- *Categorical data* has values that are categories, such as colors, words, or zip codes.

  What's your favorite animal? is a question that produces categorical data.

It's important to be specific when writing survey questions, so you can gather the exact type of data you need.

Try This

Determine whether each survey question will produce categorical or quantitative data.

a How many languages do you speak?

b Are you left-handed or right-handed?

c What is your height?

d What is your phone number?

It can be helpful to represent categorical data in a *two-way table*. Two-way frequency tables show counts and totals for each group. For example, 105 students take dance and are 13 or older.

A *relative frequency table* shows the same data as percentages out of the total number of responses in each group. This is also sometimes called a *conditional frequency* table.

The relative frequency table in this example shows the percentages of students who take or do not take dance class based on the age groups.

You can use this representation to see if the data presents evidence that there is an *association* between the two variables.

Arts Studio Students

	Take Dance Class	Do Not Take Dance Class	Total
13 or Older	105	121	226
Under 13	89	62	151
Total	194	183	377

Arts Studio Students

	Take Dance Class	Do Not Take Dance Class	Total
13 or Older	≈46.5%	≈53.5%	100.0%
Under 13	≈58.9%	≈41.1%	100.0%

In these tables, there is evidence of an association in the data between age and whether a student takes dance class. This association can be seen when comparing the students under 13 who take dance class (≈58.9%) to the students 13 or older who take dance class (≈46.5%).

Try This

Here are two tables that show some athletes' states of mind during a track meet and whether or not they meditated beforehand.

a Complete each table.

Two-Way Table

	Meditated	Did Not Meditate	Total
Calm	45		53
Anxious		21	
Total	68	29	97

Relative Frequency Table

	Meditated	Did Not Meditate	Total
Calm	≈85%		100%
Anxious			100%

b Choose *one* number from each table and explain what it means.

Summary | Lesson 3

Two-way tables and total relative frequency tables can be helpful for making decisions.

For example, compare this total relative frequency table and its corresponding two-way table. The tables show the data gathered by Tyani's principal about which field trip the students would prefer to take: going whale watching or going to an art museum.

The relative frequency table shows that Tyani's homeroom has a stronger preference for whale watching than the other homerooms, but the two-way table shows that there are more students overall who would prefer to go to the art museum.

Total Relative Frequency Table

	Whale Watching	Art Museum	Total
Tyani's Homeroom	8.4%	6%	14.4%
Other Homerooms	38.4%	47.2%	85.6%
Total	46.8%	53.2%	100%

Two-Way Table

	Whale Watching	Art Museum	Total
Tyani's Homeroom	21	15	36
Other Homerooms	96	118	214
Total	117	133	250

Try This

This table shows some athletes' states of mind during a track meet and whether or not they meditated beforehand.

	Meditated	Did Not Meditate
Calm	≈66%	
Anxious		≈72%
Total	100%	100%

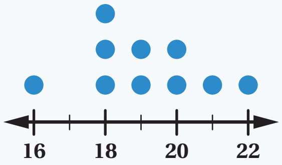
a. Complete the relative frequency table.

b. Based on this data, is there an association between meditating and state of mind during a track meet? Explain your thinking.

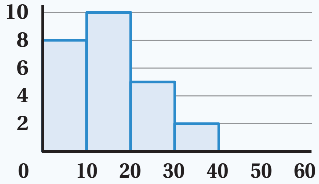
Summary | Lesson 4

It can be helpful to visualize quantitative data using dot plots or histograms.

Dot plots present each data point as a dot at its value on a number line. Dots with the same value are stacked on top of one another.



Histograms group data into rectangular bins. The height of each rectangle shows how many values are in that bin.



A dot plot is useful for observing the frequency of individual data, and the number of points in a data set. Histograms are useful for representing large data sets and observing the shape of a distribution.

Try This

Here is a histogram of how students rated the season fall/autumn on a scale from 0–10.

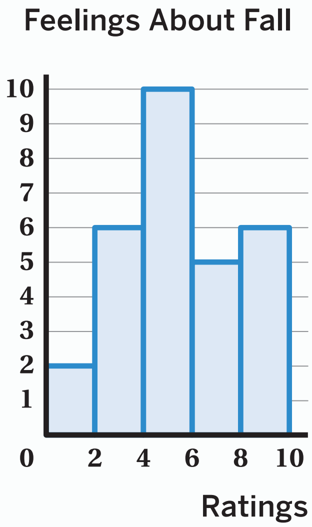
Determine whether each statement is true, false, or cannot be determined using the histogram.

a There are 29 total ratings.

b The highest rating was a 9.

c The lowest rating was less than 2.

d There are 10 ratings of 6 or higher.



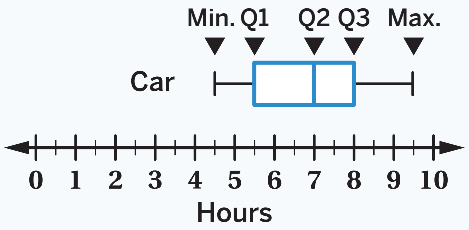
Summary | Lesson 5

A *box plot* is one way to visualize quantitative data. The data is divided into four sections using five values: the *minimum*, the *maximum*, and three quartiles. *Quartiles* divide a data set into four sections, or quarters. Each quarter represents 25% of the data.

Quartile 1 (Q1) is the median of the lower half of the data. Quartile 2 (Q2) is also the *median* of the entire data set, which divides the data into two halves. Quartile 3 (Q3) is the median of the upper half of the data.

The minimum, maximum, median, and quartiles are all examples of statistics. A **statistic** is a single number that measures something about a data set.

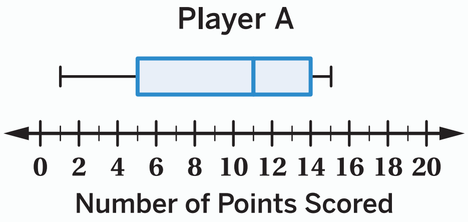
Box plots are useful for representing large data sets, especially those with extreme values. They are less helpful for seeing individual data points.

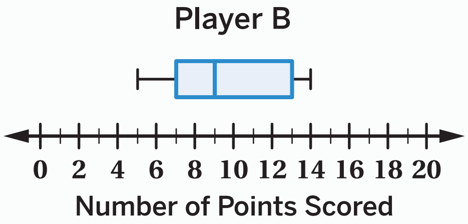


Box plots can be used to make statements about percentages of data. This box plot represents data on road trip travel times. According to the box plot, 50% of travelers had a travel time of under 7 hours. However, 75% of the recorded travel times were less than 8 hours. The longest road trip recorded was 9.5 hours.

Try This

Two basketball players recorded their points for each game in the season.





Determine whether each statement is true, false, or cannot be determined using the box plots.

a The median of Player B's data is 9 points.

b Player A played 15 games this season.

c Player A scored 0 points in at least one game.

d Player B scored between 7 and 9 points in about the same number of games as they scored between 9 and 13 points.

Summary | Lesson 6

Here are some terms that can help us describe the *shape* of a data set.

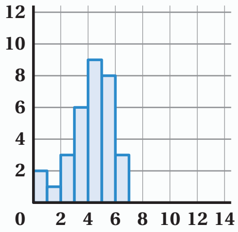
- A data set is **bell-shaped** when most of the data is at the center and there are fewer points farther from the center. When presented in a dot plot or histogram, the data looks like a bell.
- A data set is **bimodal** when it has two very common data values. These appear in a dot plot or histogram as two peaks.
- A data set is **skewed** when more values are concentrated on one end of the data than the other.
- A data set is **symmetric** when you can draw a vertical line of symmetry through it.
- A data set is **uniform** when the data values are evenly distributed.

Try This

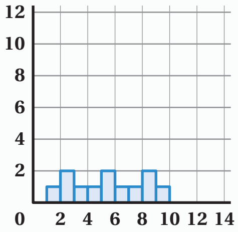
Match each histogram with the best description of its shape.

Bimodal Bell-shaped Skewed Symmetric

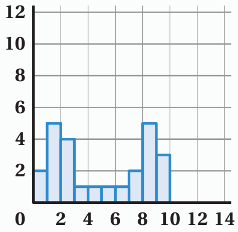
a.



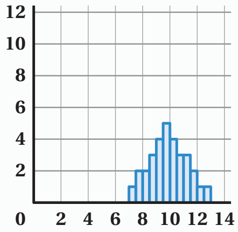
b.



c.



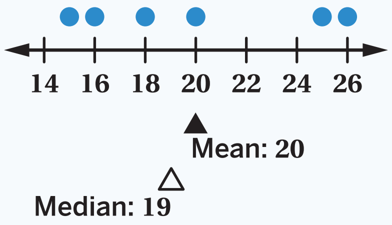
d.



Summary | Lesson 7

A **measure of center** is a single number that summarizes all of the values in a data set. *Mean* and *median* are measures of center that are used to describe a typical value of a data set.

- The mean is also called the average of a data set. To calculate the mean, you can add up all the data values, and divide by the number of data points.
- The median is the middle value of a data set when the values are in numerical order. If there are two values in the middle of the data set, then the median is the middle of those two values.



Here is one way to determine the mean of the data in this dot plot.

$15 + 16 + 18 + 20 + 25 + 26 = 120$

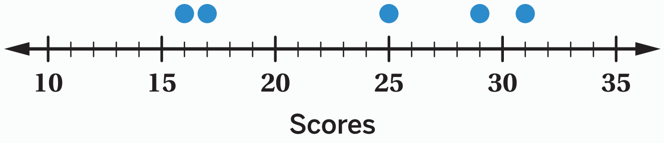
$120 \div 6 = 20$

The median is the value halfway between 18 and 20, so the median is 19.

When a data set includes extreme values that are much larger or smaller than most of the data, the value of the mean and median can be very different. Extreme values impact the mean more than they impact the median.

Try This

Here is a dot plot of Oscar's scores from his newest video game.



a Calculate the mean and median of Oscar's data.

b Oscar plotted his last score incorrectly. He actually scored a 6 instead of a 16.

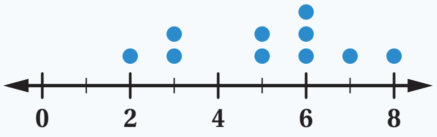
Which measure(s) of center would be affected by this mistake? Explain your thinking.

Summary | Lesson 8

The **standard deviation** of a data set is a **measure of spread**, or a way to measure how spread out the values in a data set are from its mean. Measures of spread can help you describe how consistent a data set is. If a data set is very consistent and much of the data clustered together, the standard deviation will be small. If the data is spread out, the standard deviation will be larger.

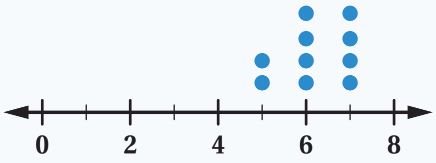
For example, compare these two dot plots.

Dot Plot A



Standard deviation: 1.81

Dot Plot B



Standard deviation: 0.75

The data in Dot Plot A has a greater standard deviation than the data in Dot Plot B because the data in A is more spread out, or variable. Since the data values in Dot Plot B are more consistent, B has a lower standard deviation.

One way to calculate the standard deviation of a data set is to use the Desmos Graphing Calculator. In the calculator, use the functions menu or type $\text{stdevp}(\ )$ and then insert a list with the data values. For example, to calculate the standard deviation of Data Set A, you can type $\text{stdevp}([2, 3, 3, 5, 5, 6, 6, 6, 7, 8])$.

Try This

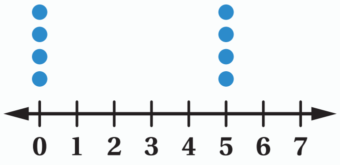
Here are three data sets.

a Which data set do you think has the lowest standard deviation?

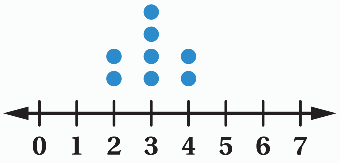
Explain your thinking.

b Use a graphing calculator and the Unit 3 Calculator Guide to calculate the mean and standard deviation for each data set.

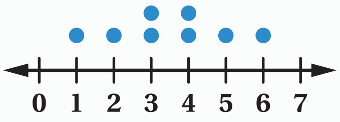
Data Set A



Data Set B



Data Set C



Summary | Lesson 9

You can use statistics, like *mean*, *median*, or standard deviation, to help you compare data sets and make conclusions about the real world. Different statistics can reveal different information about the data sets. Comparing the mean of two data sets can help determine which one generally has higher or lower values. Comparing the standard deviation of two data sets can help determine which one is more consistent.

For example, here are statistics about the high temperatures in Metropolis and Springtown over one week.

	Mean (°F)	Standard Deviation (°F)
Metropolis	74	6
Springtown	74.43	2.06

When you compare the means, you can see that both Metropolis and Springtown had similar high temperatures in that week. But when you compare the standard deviations, you can see that the temperature in Springtown was more consistent than the temperature in Metropolis.

Try This

Three people travel to work each day. Here are the means and standard deviations for each of their travel times.

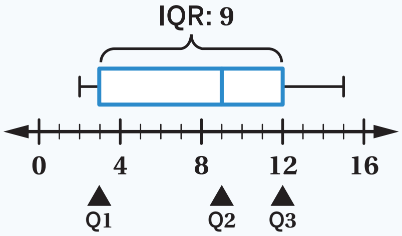
Person A	Person B	Person C
Mean: 29.71 minutes Standard Deviation: 2.60 minutes	Mean: 38.71 minutes Standard Deviation: 3.10 minutes	Mean: 28.14 minutes Standard Deviation: 8.68 minutes

a Which commuter's travel times were most consistent?
Use statistics to justify your claim.

b Which commuter had the longest commute time?
Use statistics to justify your claim.

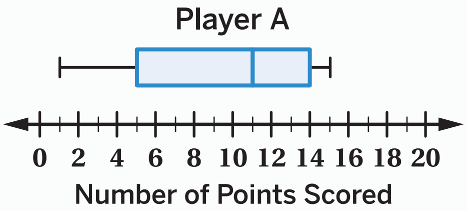
The *interquartile range* (IQR) is a statistic that measures the spread of a data set. In a box blot, the IQR is the width of the box. Measures of spread, like IQR, help us determine how consistent the data within a set is. The IQR represents the middle half of the data set, and we calculate it by determining the distance from Q1 to Q3. This makes IQR a more resistant measure of spread for skewed data than standard deviation.

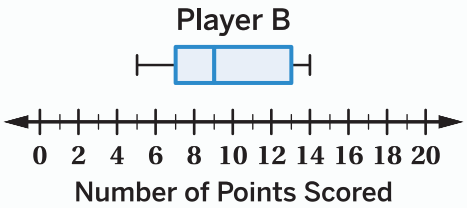
Here is an example of a data set where Q1 $= 3$ and Q3 $= 12$. The IQR of this data set is 9, because $12 - 3 = 9$.



Try This

Two basketball players recorded their points for each game in the season.



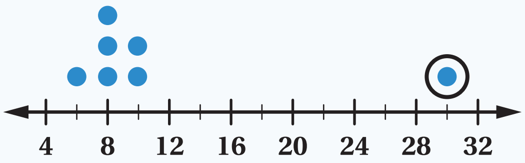


a Determine the median and interquartile range (IQR) for each player.

b Which player was more consistent in their points scored? Use statistics to justify your claim.

c Which player generally scored more points? Use statistics to justify your claim.

Outliers are data values that are far from the other values in the data set. In this data set, the circled data point is an outlier.

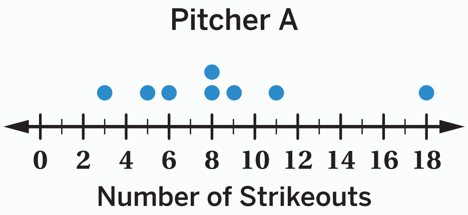


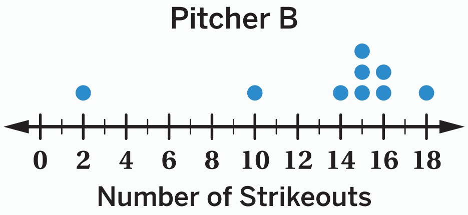
You can identify outliers using dot plots, box plots, and technology tools such as graphing calculators. You can also identify outliers using the *IQR*. Outliers are values further than 1.5 times the IQR below Q1 or above Q3.

When deciding which measure of center is appropriate to represent a data set, it's important to identify any outliers. Outliers have a big impact on the *mean*, but they don't impact the *median* very much.

Try This

Let's compare the pitchers on two different baseball teams. These dot plots show the number of strikeouts each pitcher threw in the recent games.





a Use a graphing calculator to make a box plot and identify any outliers in each data set.

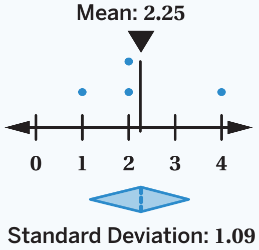
b If you removed any outliers from each data set, what would be more affected, the mean or the median? Explain your thinking.

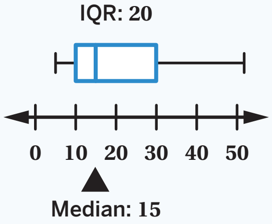
Summary | Lesson 12

It can be helpful to use measures of center and measures of spread to compare data sets. You can choose which measure of center or spread to use based on the shape of the data:

- When data distributions are symmetric or bell-shaped, you can use the mean and standard deviation to compare.
- When data distributions are skewed or contain outliers, you can use the median and IQR because the mean and standard deviation are both affected by extreme values.

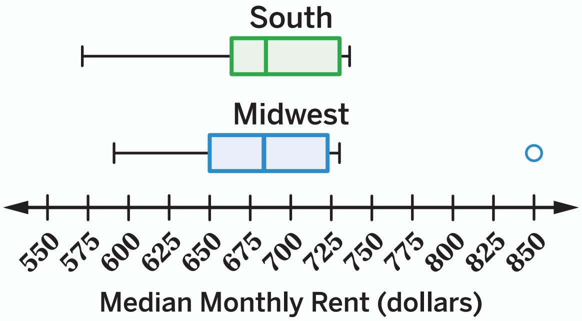
Comparing different data sets can reveal important information about change over time in different situations, like minimum wage or median rent. Comparing measures of center, like mean or median, can help you determine if the data has increased, decreased, or stayed the same. Comparing measures of spread, like standard deviation or IQR, can help you determine if the data has become more or less consistent over time.





Try This

Here is a box plot showing the median monthly rent for eight states in the South and Midwest in 2019.



a What measures of center and spread would you use to compare the rents in each group of states?

Explain your thinking.

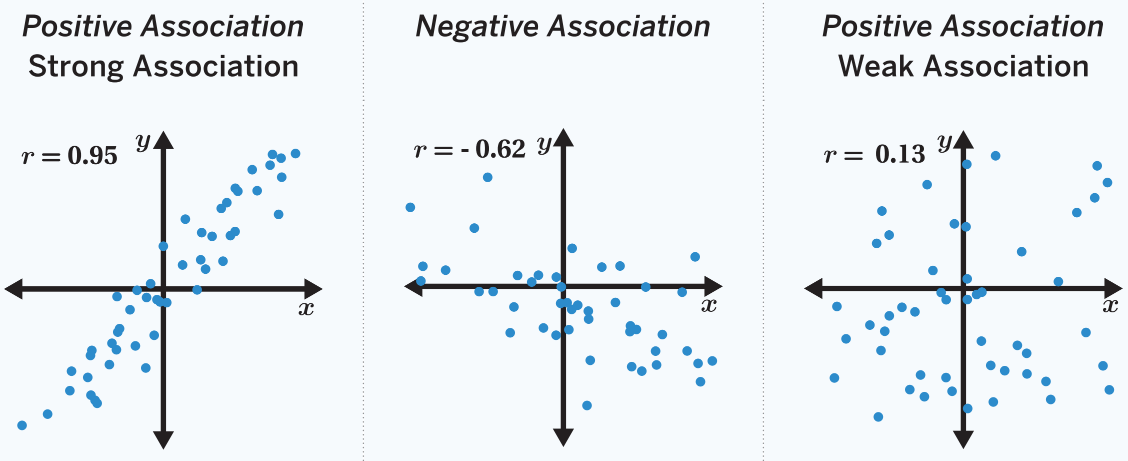
b How did rents in the Midwest compare to rents in the South in 2019?

Use statistics to justify your claims.

A *scatter plot* is a graph of plotted points that shows the relationship, or association, between two variables. When there is a *linear association* between the variables, you can describe the *strength* and *direction* of the association using the **correlation coefficient** (also called the **r-value**). The correlation coefficient is a number between -1 and 1 that describes the strength and direction of a linear association between two variables in a scatter plot.

The closer the r-value is to 0, the weaker the linear association. The closer the r value is to -1 or 1, the stronger the linear association.

These graphs show examples of different associations and their correlation coefficients.



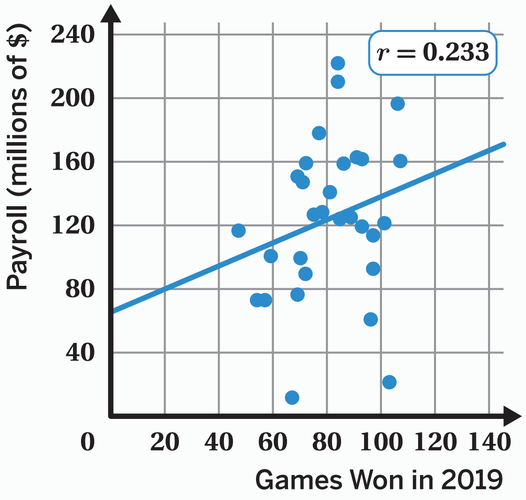
Try This

Lucy was curious about the relationship between how often baseball teams win games and how much money they spend on their players (called payroll).

She found data about games won in 2019 and the team's payroll (in millions of dollars).

a What is the correlation coefficient for Lucy's data?

b Use the correlation coefficient to describe the strength and direction of the association between wins in 2019 and payroll.



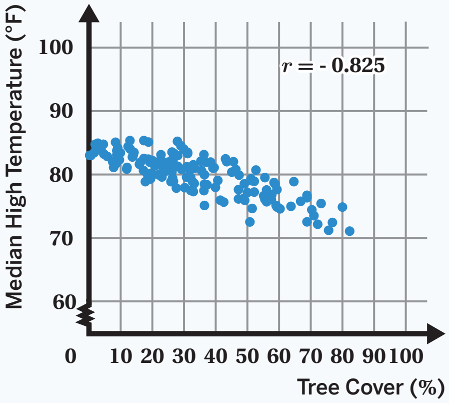
You can use a correlation coefficient to analyze the relationship between two variables and determine whether there is an association between them. The correlation coefficient, or r-value, describes the strength and direction of the relationship that may exist between two variables.

- A positive r-value means that as one variable increases, the other variable also increases.
- A negative r-value means that as one variable increases, the other variable decreases.
- The closer the r-value is to 1 or -1, the stronger the correlation.

People may use correlations in data to understand and address issues in their community.

For example, this scatter plot shows data on tree cover and temperature for 150 blocks in Detroit, Michigan.

The r-value is -0.825. This means there is a negative and strong relationship between the amount of tree cover and median high temperature in Detroit neighborhoods. Community members may use this correlation to advocate for more trees to be planted in different neighborhoods across the city.

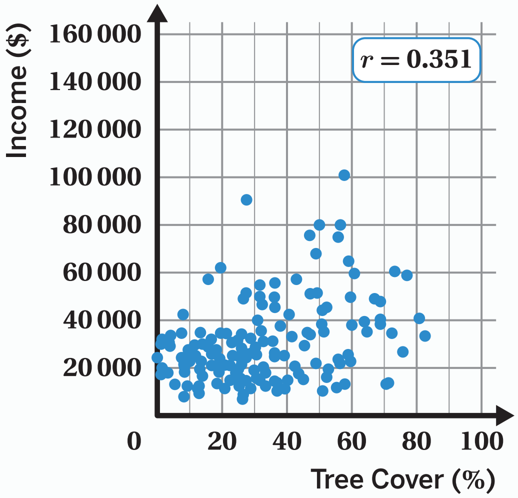


Try This

Here is a scatter plot with data about the percentage of tree cover recorded in an area of Detroit, Michigan and the income of residents who live there.

a Use the correlation coefficient to describe the association between tree cover and income.

b How might someone use this data and r-value to advocate for their community?



Summary | Lesson 15

While the correlation coefficient can help you understand the general relationship between two variables, a *line of fit* can help you make predictions about specific values in a data set. Points along the line of fit represent the likely value of unknown data in the data set.

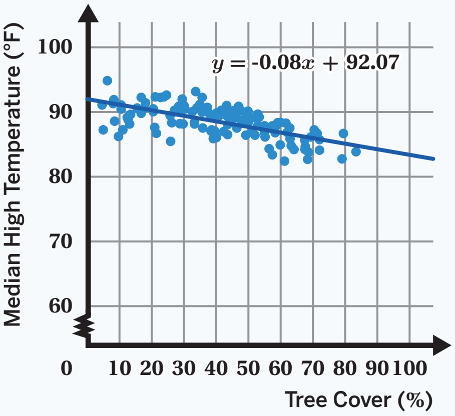
You can use the equation of the line of fit to better understand the data. The y-intercept represents a potential initial value and the slope of the line describes the rate that the variables change in relationship to one another.

For example, this scatter plot shows data on tree cover and temperature for 150 blocks in Austin, Texas. The equation of the line of fit is $y = -0.08x + 92.07$.

The slope is -0.08. This means that when the tree cover increases by 1% in Austin, the predicted temperature decreases by 0.08°F.

The y-intercept is 92.07. This means that if the tree cover in Austin is 0%, the predicted temperature is 92.07°F.

You can use the line to predict that if a block in Austin has 80% tree cover, the temperature will be about 85°F.



Try This

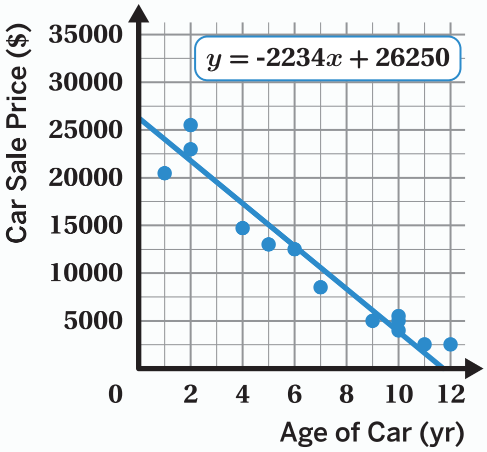
Kwasi was curious about the relationship between the ages of cars and their values. He found data on the ages of several cars (in years) and their sale prices (in dollars) and created a line of fit.

a What does each value represent?

-2,234:

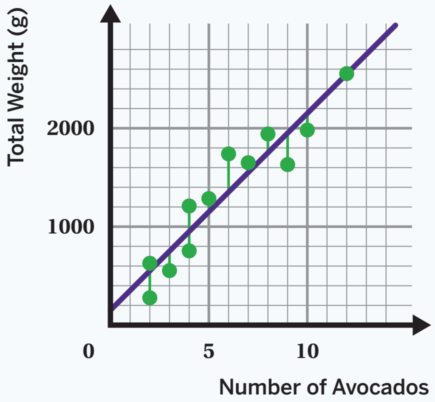
26,250:

b If a car is 3 years old, what does the line of fit predict its price will be?



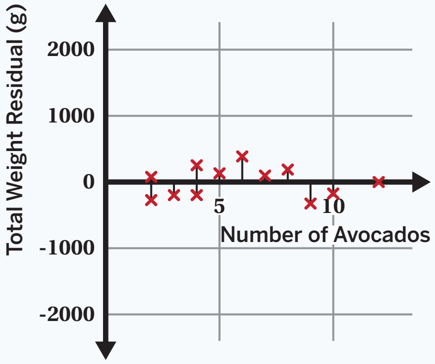
You can use residuals to determine how well a line fits a data set. A **residual** is the difference between the actual y-value of a data point and the value predicted by the line of best fit.

Here is a scatter plot with data on the number of avocados and their weights. The residuals are represented with lines connecting each point to the line of best fit.



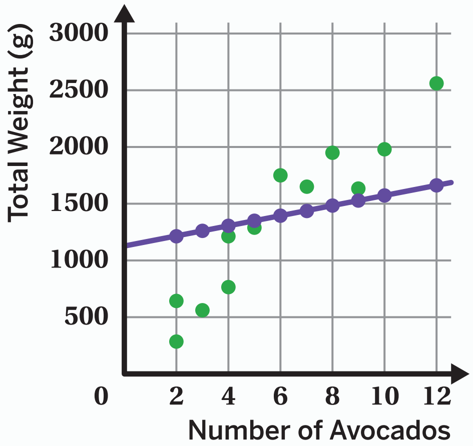
You can also create a residual plot to analyze how well a line fits a data set. A **residual plot** is a scatter plot of residual values for a data set.

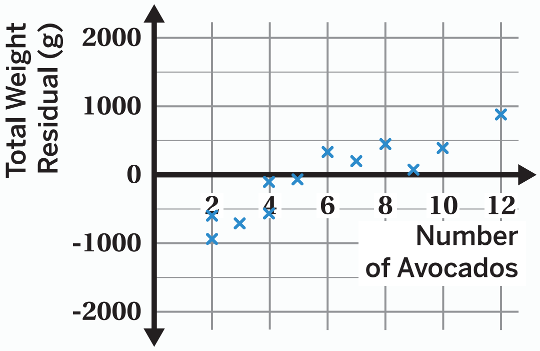
Here is the residual plot of the graph of avocado weights. The closer a point is to the x-axis, the closer that point is to the line of best fit. A line is a good fit for the data if the points on the residual plot are close to the x-axis and are randomly dispersed above and below the axis.



Try This

Here is a line of fit for some data about avocados and the residual plot for the line of fit.





a Explain why the point on the residual plot that represents 8 avocados is positive and the point for 3 avocados is negative.

b Use the residual plot to explain why this line is *not* a good fit for the data.

Summary | Lesson 17

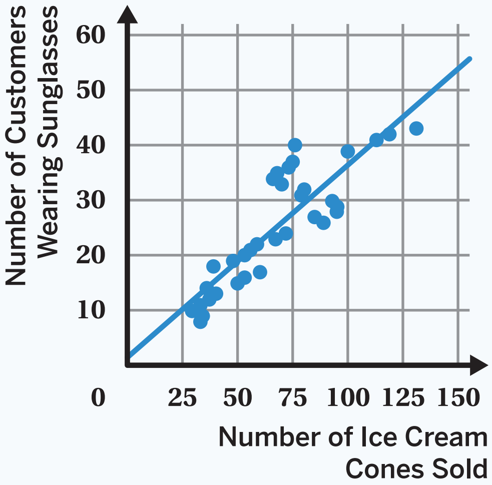
The **line of best** fit is the line on a scatter plot that best represents the trend created by the points in a data set.

Instead of sketching a line of fit, you can use a graphing calculator to precisely generate the equation of the line of best fit from a scatter plot. The equation of the line of best fit can help you interpret information about a situation, or allow you to substitute values into the equation to make predictions. You can also use a graphing calculator to calculate the correlation coefficient for a given data set.

Here is an example of a line of best fit generated by a graphing calculator, and the information about the line that a graphing calculator will show you.

$y = 0.351312x + 1.31984$

STATISTICS	RESIDUALS
$r^2 = 0.7642$	plot residuals
$r = 0.8742$	plot residuals

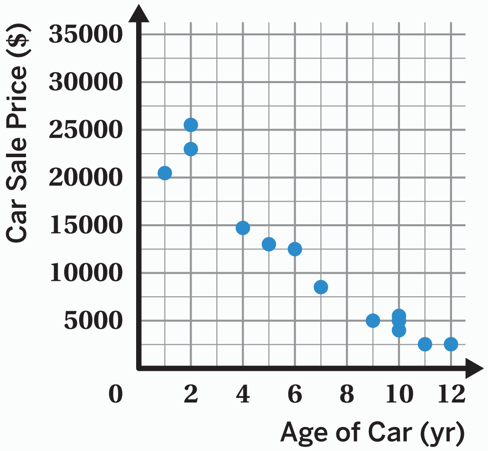


Try This

Kwasi was curious about the relationship between the ages of cars and their values. He found data on the ages of several cars (in years) and their sale prices (in dollars).

a Use the information from the graphing calculator to determine the equation of the line of best fit and the correlation coefficient.

b What does the correlation coefficient say about the association between car age and sale price?



$y = -2270.38x + 26886.7$

STATISTICS	RESIDUALS
$r^2 = 0.9215$	plot residuals
$r = -0.96$	plot residuals

Summary | Lesson 18

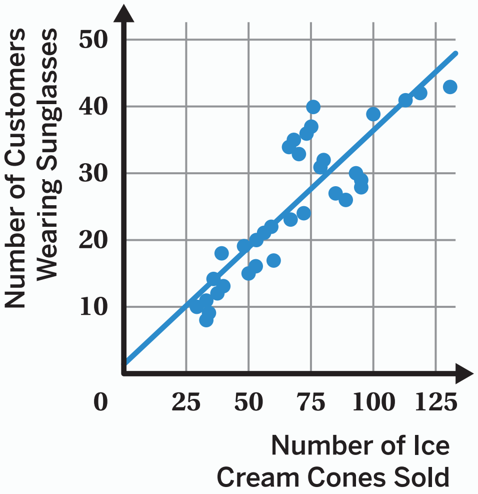
A **correlation**, sometimes called an association, is a relationship between two or more variables. One type of correlation is **causation**, which describes a relationship where a change in one variable causes a change in the other variable.

But not every correlation is a *causal relationship*. For example, a third variable can cause the relationship between two variables to change. Correlation can even be caused by coincidence: if lots of variables are considered, then it's very likely that at least two of them will be somewhat correlated.

Media, such as article headlines, might present relationships that are correlated as causal. To be a critical consumer of information, check the sources behind claims and ask questions about what conclusions can be made from a data set.

Try This

Nyanna noticed a trend at an ice cream shop. She recorded the number of ice cream cones sold and the number of customers wearing sunglasses one day. Then she calculated the r-value, which was 0.87.



a What does the 0.87 say about the association between these two variables?

b Do you think one of the variables causes the other? If not, what else could be affecting the relationship?

Explain your thinking.

We can use statistical tools to help us better understand our communities and the world. Scatter plots help us visualize large amounts of data, lines of best fit help us see trends and relationships or associations in that data, and correlation coefficients (r-values) help us describe the strength and direction of any associations that we find.

While all of these tools help us describe associations and test predictions, we need to explore the variables and relationships further to determine things like the cause and the effect on real people. Statistical tools are powerful, but they are just one step in understanding issues that affect real communities.

Try This

a What are two variables you could explore the association between using the statistics tools in this unit?

b Write a question about your variables that you could answer using the statistics tools from this unit.

c Write a question about your variables that you could *not* answer using the statistics tools from this unit.

Try This | Answer Key

Lesson 1

a Quantitative data

b Categorical data

c Quantitative data

d Categorical data.

Caregiver Note: Even though these are numbers, phone numbers are not measurements or quantities (e.g., 123-456-7890 is not less than 234-567-8901 in any meaningful way), so this is categorical data.

Lesson 2

a

Two-Way Table

	Meditated	Did Not Meditate	Total
Calm	45	**8**	53
Anxious	**23**	21	**44**
Total	68	29	97

Relative Frequency Table

	Meditated	Did Not Meditate	Total
Calm	≈ 85%	**≈ 15%**	100%
Anxious	**≈ 52%**	**≈ 48%**	100%

Caregiver Note: One strategy for calculating the missing values in the relative frequency table is to divide each value in the two-way table by the total for the row. For example, 23 people meditated out of the 44 total anxious people, so the percentage of anxious people who meditated is $\frac{23}{44}$ or ≈ 52%.

b *Responses vary.*

- 23 is the number of athletes who meditated before the meet and were anxious during the meet.
- ≈ 52% of all the athletes who were anxious meditated before the meet.

Lesson 3

a Responses shown in the table.

b Yes. *Explanations vary.* The group that meditated had a much lower percentage of athletes who were anxious. More than half of the athletes who meditated were calm, but only about a quarter of the athletes who did not meditate were calm.

	Meditated	Did Not Meditate
Calm	≈ 66%	**≈ 28%**
Anxious	**≈ 34%**	≈ 72%
Total	100%	100%

Lesson 4

a True.

Caregiver Note: One strategy for determining the total number of ratings is to add the ratings in each bin. $2 + 6 + 10 + 5 + 6 = 29$ ratings.

b Cannot be determined.

Caregiver Note: The histogram shows that the highest rating was between 8 and 10 but does not say precisely what the rating is.

c True.

Caregiver Note: The bin with the lowest ratings has ratings between 0 and 2.

d False.

Caregiver Note: There are 11 ratings higher than 6.

Lesson 5

a True.

Caregiver Note: The median is the value of quartile 2. The box plot shows that the median for Player B is 9 points.

b Cannot be determined.

Caregiver Note: Box plots do not show the number of data points, only groups of 25% of the data.

c False.

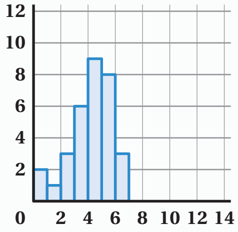
Caregiver Note: The minimum for Player A is 1 point.

d True, or cannot be determined.

Caregiver Note: On a box plot, each section represents about 25% of the data. Even though the section between 9 and 13 points looks larger than the section between 7 and 9 points, they each include about one quarter of the total games. Depending on the exact number of games, they may be slightly different, but that information isn't shown in the box plot.

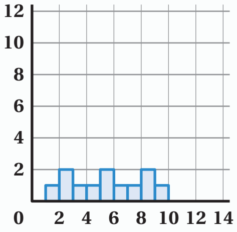
Lesson 6

a.



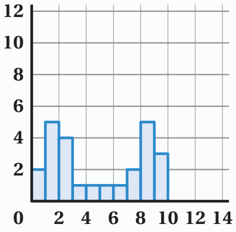
Skewed

b.



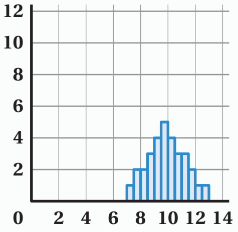
Symmetric

c.



Bimodal

d.



Bell-shaped

Lesson 7

a Mean: 23.6, Median: 25

b The mean would be affected. *Explanations vary*. Every value is used when calculating the mean, so changing any value affects the mean. The median is not affected by the change because 25 is still the middle value.

Lesson 8

a Data Set B. *Explanations vary*. The data points in Set B are closest together, so the variation between them is lowest and the standard deviation is lowest.

b Data Set A: Mean: 2.5, Standard Deviation: 2.5

Data Set B: Mean: 3, Standard Deviation: 0.7

Data Set C: Mean: 3.5, Standard Deviation: 1.5

Caregiver Note: The Calculator Guide includes instructions on how to use the Desmos Graphing Calculator to calculate the mean and standard deviation of a data set. Use the stdevp( ) *function in the calculator because each of these data sets is a population rather than a sample.*

Lesson 9

a Person A. *Explanations vary*. Person A's travel times are the most consistent because they have the smallest standard deviation.

b Person B. *Explanations vary*. I think Person B has the longest commute time because their mean travel time is the longest.

Lesson 10

a Player A: Median: 11, IQR: 9

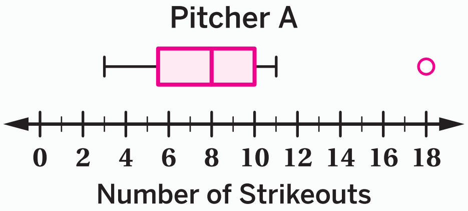
Player B: Median: 9, IQR: 6

b Player B. *Explanations vary*. Player B is more consistent because their data has a smaller IQR, which is a measure of spread.

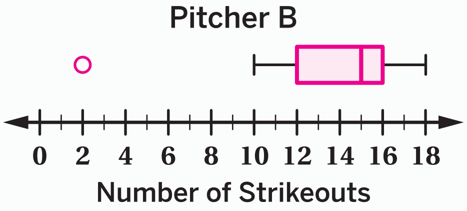
c Player A. *Explanations vary*. Player A generally scored more points because the median of their data was higher.

Lesson 11

a



Outlier: 18



Outlier: 2

b The mean would be more affected. *Explanations vary*. Every value is used when calculating the mean, so changing any value affects the mean. The median is not affected by the change because 15 is still the middle value after removing the outlier.

Lesson 12

a *Responses and explanations vary*. I would use the median and IQR to compare the rents because the data is not symmetrical. There is an outlier at about $850 which affects the mean and standard deviation.

b *Responses vary*. The median rents for both regions are alike. The rents in the South are more consistent since the rents in southern states have a smaller IQR compared to the IQR of the midwestern states.

Lesson 13

a 0.233

b There is a weak positive association between games won in 2019 and payroll in millions of dollars.

Lesson 14

a There is a weak positive relationship between percentage of tree cover and income in Detroit, Michigan.

b *Responses vary*. Since tree cover and income are not inherently related, someone might show the data to a city council to justify adding more trees into neighborhoods with lower incomes, which would bring the correlation coefficient closer to 0.

Lesson 15

a -2,234: The sale price of a car drops $2,234 every time the age of the car increases by 1 year. 26,250: If a car is new (0 years old), it would sell for $26,250.

b $19,548, or about $20,000

One strategy for predicting using a model is to substitute the known value into the line of fit.

$y = -2234(3) + 26250$

$y = -6702 + 26250$

$y = 19548$

Lesson 16

a *Responses vary*. The point on the residual plot that represents 8 avocados is positive because the value of the actual data point is greater than the line of fit predicts. The point that represents 3 avocados is negative because the value of the actual data point is less than the line predicts.

b *Responses vary*. The line is not a good fit for the data because many of the points are far from the x-axis. The residuals start off all negative and then turn positive, which shows that the line does not follow the pattern of the data. If the line was a good fit, the residuals would be close to the x-axis and not follow any pattern.

Lesson 17

a Line of Best Fit: $y = -2270.38x + 26886.7$; Correlation Coefficient: $r = -0.96$

b There is a strong negative relationship between car age and sale price.

Lesson 18

a The 0.87 says that there is a strong positive relationship between the number of ice cream cones sold and the number of customers wearing sunglasses.

b *Responses vary*. I don't think one variable causes the other. It doesn't make sense that the number of sunglasses worn or the number of ice cream cones sold cause each other. Both of these likely increase when the weather is warm.

Lesson 19

a *Responses vary*. The number of miles of highway in a city and the amount of time stuck in traffic.

b *Responses vary*. Does the amount of time stuck in traffic increase or decrease as the number of miles of highway increases?

c *Responses vary*. What should we do to help reduce traffic in our community?