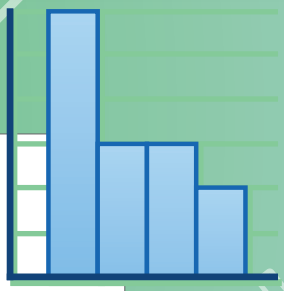
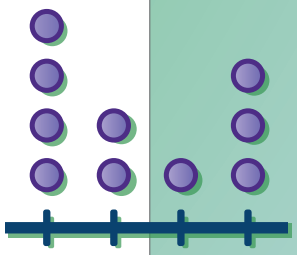


Unit 8

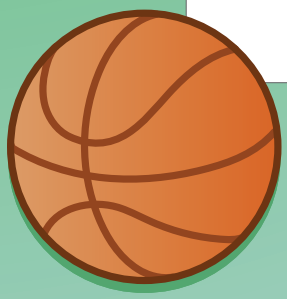
Describing Data

Statistics is the science of collecting and analyzing data. It is one of the most relevant aspects of mathematics in everyday life, and it is also used by researchers in many fields, such as sociologists identifying and studying problems in society. In all cases, knowing what is typical is critical to understanding what is not.



Essential Questions

- What are statistical questions and how are they used?
- What are different ways to represent numerical data?
- How do we measure the center of a data set?
- How do we measure the spread of a data set?



Summary | Lesson 1

You can use surveys and measurements to collect data to answer questions about a topic. This data can be either **categorical data** or **numerical data**. Categorical data can be sorted into categories. Numerical data are numbers, quantities, or measurements that can be meaningfully compared. Some data that contain numbers, like addresses or dates, are categorical because the numbers are not quantities or measurements.

Here are some examples.

Categorical Data

- Favorite color
- Food people eat for lunch
- People's phone numbers

Numerical Data

- Number of people in each class
- Weights of dogs
- Ages of people in your school

Try This

Antwon asked his friends the following question: *How many hours do you spend on your phone each week?*

Here is the data he collected.

12	24	25	0	40	0	28

a Antwon claims that most of his friends use their phone for 3 or more hours each day.

Do you agree? Explain your thinking.

b Antwon also asked his friends: *Do you have more than 2 pets?*

Will this question produce numerical data or categorical data? Explain your thinking.

Summary | Lesson 2

A **statistical question** is a question that needs more than one piece of data to answer it.

Here is an example:

- "Which classroom in your school has the most books?" is a statistical question because you need to know the number of books in *each* classroom to answer it.
- "How many books are in your classroom?" is not a statistical question because you only need to know the number of books in *one* classroom to answer it.

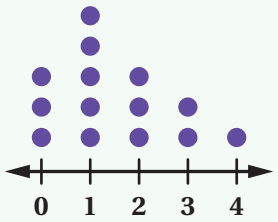
You can organize data that answers a statistical question into a list or a **dot plot**.

List

0, 0, 0, 1, 1, 1, 1, 1, 2, 2, 2, 3, 3, 4

Lists allow you to see all of the data. Lists can be used for both numerical and categorical data.

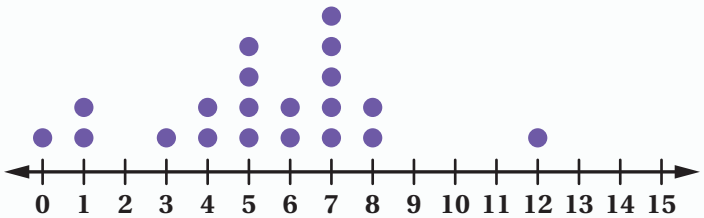
Dot Plot



Dot plots are a visual representation of numerical data and allow you to compare multiple data sets.

Try This

This dot plot shows the number of books that 20 students in a 6th grade class read in one month.



a What do the 2 dots plotted at 4 tell us about this situation?

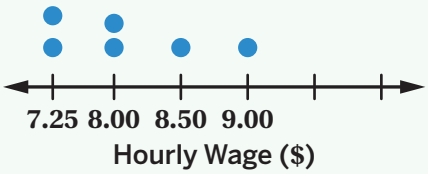
b Julian claims that most students in the class read 6 books or less this month. Do you agree with him? Explain your thinking.

Summary | Lesson 3

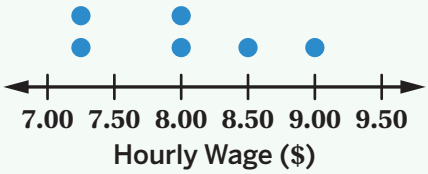
You can use dot plots to visualize data and compare different data sets. When creating a dot plot it is important to use a consistent scale on the number line and a consistent amount of vertical space between the dots.

Here is an example of two dot plots showing the same set of data, one with mistakes and one done accurately.

Dot Plot With Mistakes



Accurate Dot Plot



Notice:

- The scale of the dot plot with mistakes is inconsistent, but the scale of the accurate dot plot is consistent with $0.50 between each tick mark.
- The height of the dots above $7.25 and $8.00 should be the same because there are the same number of dots above each.

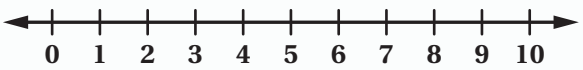
Try This

Zola asked 8 of their friends: *How many hours did you work in the school store this week?*

They collected the responses in this table.

8	0	2	5	5	2	1	5

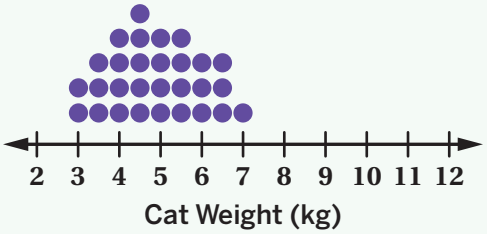
Make a dot plot using this data.



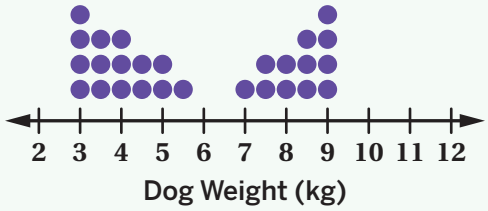
Summary | Lesson 4

You can use **center**, **spread**, and **shape** to describe the distribution of data on a dot plot. The center is a number that represents a typical value. The spread describes how alike or different the values in a distribution are, often in relation to the center.

Here are some examples.



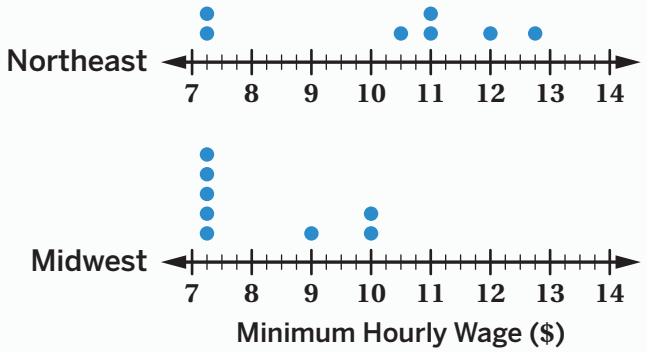
- The data is symmetric.
- There are no gaps in the data.
- The center of the data is between 4 and 5 kilograms.
- Most of the data is clumped together.
- The data is shaped like a mountain with the most dots on 4.5 kilograms.



- There is a gap in the data between 5.5 and 7 kilograms.
- The data is spread out in two clumps, one with a peak at 3 kilograms and another peak at 9 kilograms.
- The center of the data is approximately 6 kilograms.
- The data is shaped like two triangles with the most dots on 3 kilograms and 9 kilograms.

Try This

Here are dot plots that show the minimum wages of states in two different parts of the U.S.



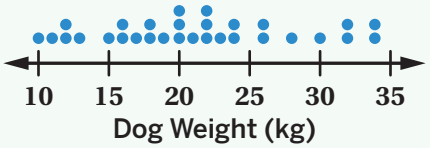
a Which dot plot has a larger spread?

b What do these dot plots tell us about minimum wages in different parts of the U.S.?

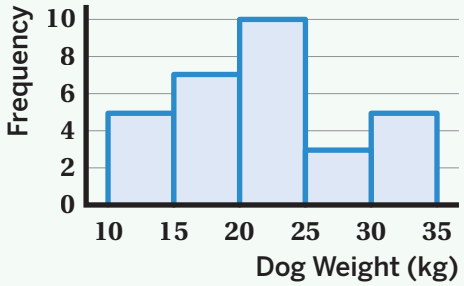
Summary | Lesson 5

You can use dot plots and **histograms** to visualize numerical data. Here is an example of a data set of the weights of 30 dogs presented in a dot plot and in a histogram.

Dot Plot



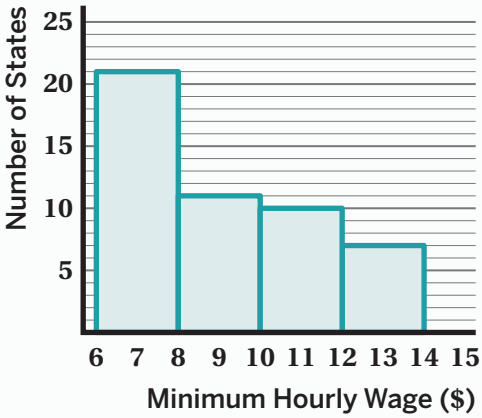
Histogram



In a histogram, data values are grouped into bins that cover a range of values, and each **bin** has the same width. The height of each bar represents the total number of values in that range, including the left boundary (least value) but excluding the right boundary (greatest value). For example, the height of the tallest bar, from 20 to 25, represents weights of 20 kilograms up to (but not including) 25 kilograms.

Try This

Here is a histogram that shows minimum wages in the U.S. in 2020.



a How many states have a minimum wage of less than $10.00?

b Michigan has a minimum wage of $10.33. Which bin should it go into?

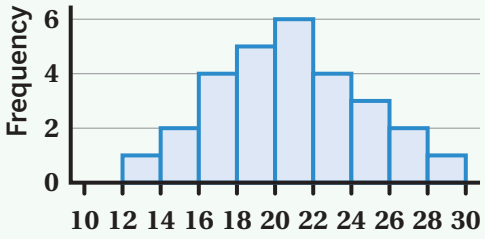
c Adriana claims that this histogram represents the minimum wages of 4 different states.

Is her claim correct? Explain your thinking.

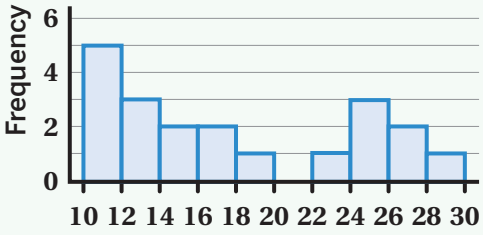
Summary | Lesson 6

You can use histograms to compare numerical data sets using their shape, center, and spread. Here are two histograms with very different shapes and features.

Histogram A



Histogram B



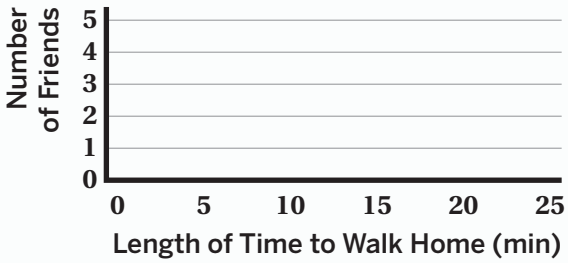
Shape	Histogram A is shaped like a mountain. It is symmetric with a peak around 21.	Histogram B has two clusters, with peaks around 11 and 25. There is a gap in the data between 20 and 22.
Center	The center is approximately 21.	The center is approximately 17.
Spread	Most values are in the middle and there are roughly the same amount on each side of 21.	Most values are on the left then get lower as it goes to the right.

Try This

Marco recorded how many minutes it takes each of his friends to walk home from school.

22	5.5	13	20	12
3	18	21.5	5	13

Create a histogram of this data.



Summary | Lesson 7

A **statistic** is a single number that measures something about a data set. One way to measure the center of a data set is by determining the **mean**, or average, of all the data values. You can think of the mean as "an equal share."

For example, suppose this data set represents how many liters of water are in 5 bottles: $1, 4, 2, 3, 0$. To calculate the mean, you first add up all of the values to determine the total (10 liters), then divide that sum by the number of values (5 bottles). This example can be represented by the expression $(1 + 4 + 2 + 3 + 0) \div 5$, or $10 \div 5$. So, the mean amount of water in the 5 bottles is 2 liters (per bottle). The mean is a whole number in this example, but it is possible for the mean to be a decimal number.

Try This

Six friends played games together at the Fly-Score Arcade.

Here are the number of tickets that each friend won.

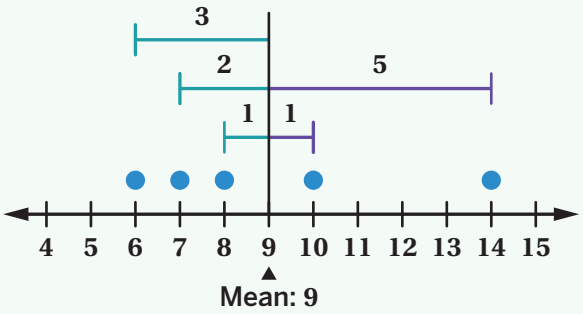
7	3	4	6	8	2

a Calculate the mean number of tickets for this data. Show your thinking.

b What does the mean tell us about this situation?

Summary | Lesson 8

The distance between a data point and the mean is called an **absolute deviation**. For example, the dot plot below shows a data set with a mean of 9. The absolute deviation of the point at 14 is 5 because 14 is 5 units away from 9.

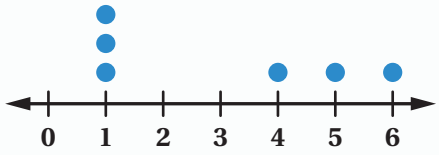


The sum of the absolute deviations to the right of the mean is equal to the sum of the absolute deviations to the left of the mean. You can use this to check whether a number is the mean of a data set. For example, if 9 is the correct mean of the data set shown, the sum of the absolute deviations to the left of 9 should be equal to the sum on the right side.

- The sum of the deviations on the left is $3 + 2 + 1 = 6$.
- The sum of the deviations on the right is $5 + 1 = 6$.
- The sums are equal so 9 is the correct mean.

Try This

Ava made a dot plot to show how many bubbles she and her friends popped in two seconds.

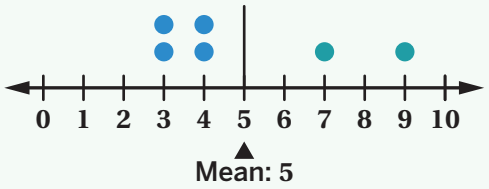


Ava says the mean of this data is 2. Is her statement correct?

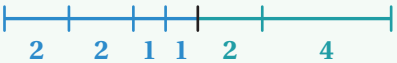
Use absolute deviations to show or explain your thinking.

Summary | Lesson 9

You can describe how spread out the values in a data set are with a single number, the **mean absolute deviation (MAD)**. The MAD is calculated by determining the mean of the absolute deviations (i.e., the average of the distances between each data value and the mean).



Mean: 5



$2 + 2 + 1 + 1 + 2 + 4 = 12$

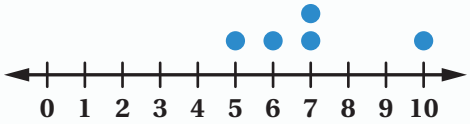
$12 \div 6 = 2$

The MAD is 2.

The mean absolute deviation is an example of a measure of spread. A measure of spread is a way to measure the consistency of the values in a data set. The smaller the value of the MAD, the less spread out the data points are around the mean, and the more consistent the data is. The larger the MAD, the more spread out the data points are around the mean, and the less consistent the data is.

Try This

Titus made a dot plot to show the number of baskets he made during each round of basketball practice last week.



a Calculate the mean of this data.

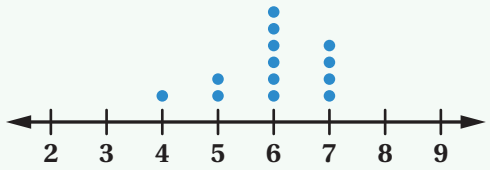
b Calculate the MAD of this data.

Summary | Lesson 10

When comparing two data sets, the mean allows you to compare the average value of each set and the mean absolute deviation (MAD) allows you to compare their spreads.

Here is an example showing the ages of two groups of kids who started learning to read and who started learning to ride a bike.

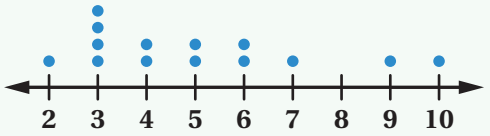
Learning to Read



Mean: 6 years **MAD:** ≈ 0.62 years

- On average, these kids started learning to read around 6 years old, which is older than the mean age for learning to ride a bike.
- The value of the MAD is smaller, meaning the data is less spread out.

Learning to Ride a Bike



Mean: 5 years **MAD:** ≈ 1.86 years

- On average, these kids started learning to ride a bike around 5 years old, which is younger than the mean age for learning to read.
- The value of the MAD is greater, meaning the data points are more spread out.

Try This

This table shows the salaries, in millions of dollars, for 8 top-earning women in Hollywood in 2019. The mean of these women's salaries is $33.7 million.

Calculate the MAD of the data.

Use the table to organize your thinking.

Salaries (in millions)	Absolute Deviations
$56	
$44	
$35	
$34	
$28	
$25	
$24	
$23.5	

You can describe a data set using another measure of center called the **median**. The median is the "middle" value in a data set when the values are listed in order from least to greatest (or greatest to least). Half of the data values are less than or equal to the median, and half of the data values are greater than or equal to the median.

To determine the median from an ordered representation of the data, you can repeat a process of eliminating the pairs of least and greatest values.

Here are some examples.

Odd Number of Values

~~0~~ ~~1~~ ~~1~~ (2) ~~2~~ ~~4~~ ~~5~~

Once all pairs have been eliminated, only one value remains in the middle, making it the median.

Median: 2

Even Number of Values

~~0~~ ~~1~~ ~~1~~ (1 2) ~~2~~ ~~4~~ ~~5~~

Once all pairs have been eliminated, two values remain.

Their average is the median.

$(1 + 2) \div 2 = 1.5$

Median: 1.5

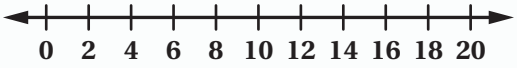
Try This

Jayla launched her toy car 7 times and recorded each distance in inches.

Distances Jayla's car traveled:
[5, 6, 17, 19, 9, 20, 18]

What is the median distance the car traveled?

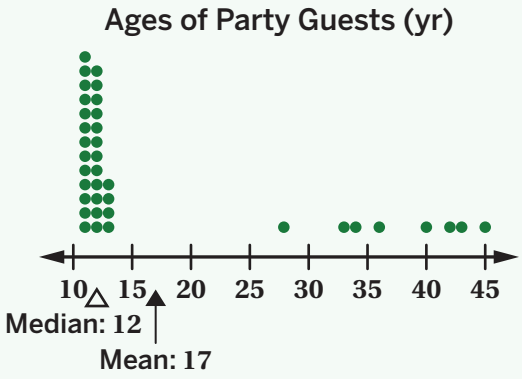
Create a dot plot if it helps you with your thinking.



Summary | Lesson 12

You can use mean and median to tell you different things about a data set. One measure might be more appropriate depending on the shape of the data and the situation.

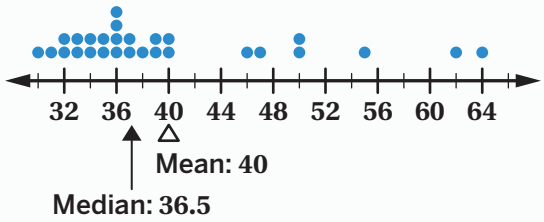
Here is data from a 12-year-old's birthday party. The median of this data set is 12. The mean of this data set is 17. In this situation, the median is a better measure to represent the ages of the party guests because most guests are 11 to 13 years old. The mean has shifted away from where most of the data is because the older ages are added to the younger ages.



When most of the data is in one place but there are a few data points far away from the group (like in the dot plot shown), the mean and median are likely to be far apart.

Try This

Here is a dot plot that shows the ages of teachers at a school.



a Is the mean or median a better measure of center to represent the ages of the teachers?

Explain your thinking.

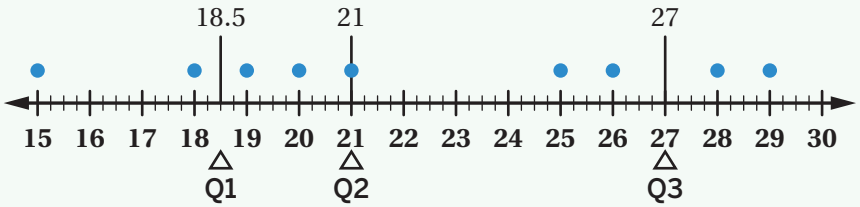
b Why do you think the mean is higher than the median?

c When are the mean and median likely to be far apart?

Summary | Lesson 13

You can describe the middle half of a data set by dividing it into four equal sections called **quartiles**.

You can determine the value of the quartiles by splitting the entire data set in half and then splitting the halves again. The middle half is all the data points that are between Q1 and Q3. Representations such as dot plots are helpful for identifying quartiles to describe data sets.



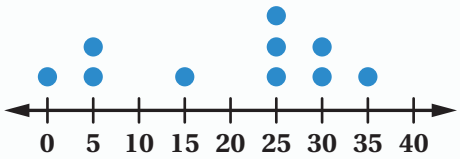
The *first quartile (Q1)* is the median of the lower half of the data set.

The *second quartile (Q2)* is the median of the entire data set.

The *third quartile (Q3)* is the median of the upper half of the data set.

Try This

Aki made a dot plot to track how many squirrels their dog chased during 10 separate walks.



Determine the values of Q1, Q2 (median), and Q3.

Q1	
Q2 (Median)	
Q3	

Summary | Lesson 14

You can create a **box plot** to visualize a data set. While a box plot shows the same data as a dot plot, it gives us new information about the data. Rather than showing every data point, a box plot separates the data into quartiles.

We can use box plots to describe the spread of the data in two ways.

- The **range** represents the difference between the *maximum* and *minimum* values of a data set. It describes the overall spread of the data.

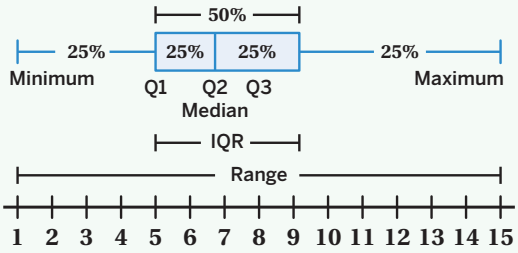
  Range: $15 - 1 = 14$

- The **interquartile range (IQR)** represents the range of the middle 50% of the data (between Q3 and Q1). It describes how spread out the middle of the data is.

  IQR: $9.25 - 6.75 = 2.5$

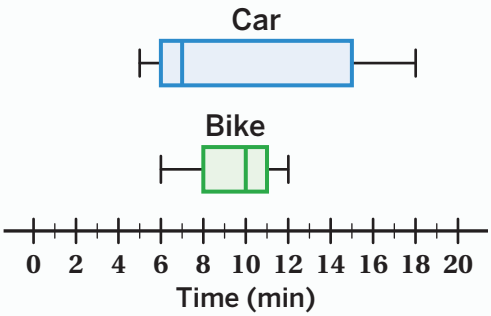
Box plots do not show how many data points are in each set, or the values of any individual data points, except the minimum and maximum.

Min.	Q1	Median	Q3	Max.
1	5	6.75	9.25	15



Try This

Keya made box plots to show how long it takes for her to get to school by car and by bike.



a For each box plot, determine the median, IQR, and range.

b If Keya wants a more predictable way of traveling, should she go to school by car or by bike?

Explain your reasoning.

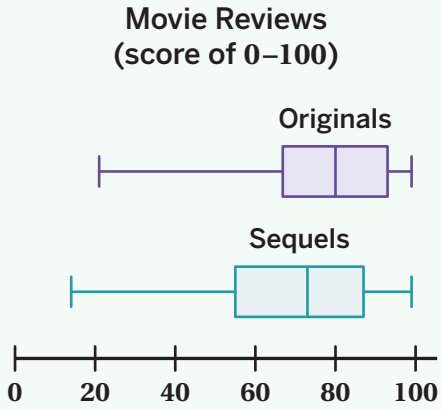
	Car	Bike
Median		
IQR		
Range		

You can make many different claims about data using box plots.

For example, these box plots show data about movie review scores for originals and sequels.

You can use the IQR or the median to make claims about specific information, as well as general trends, including:

- The IQR for originals is 20 and, for sequels, the IQR is 30. This tells us that the middle 50% of the data is more spread out for sequels.
- The median score for originals is higher than the median score for sequels. This tells us that original movies typically had higher review scores than sequels.



Originals	Sequels
Median: 80 IQR: 20	Median: 75 IQR: 30

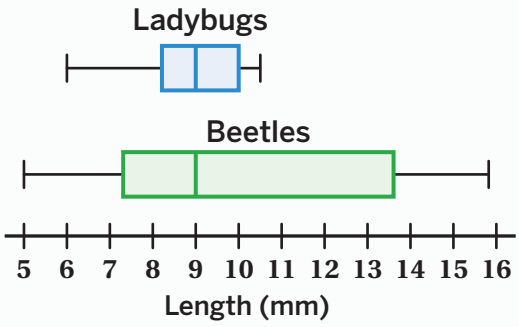
Try This

These box plots represent the length data for a collection of ladybugs and a collection of beetles.

a Complete the table to estimate the median and IQR of each box plot.

b Compare the IQR of each bug collection.

What do these numbers tell you about the lengths of ladybugs and beetles?



	Median	IQR
Ladybugs		
Beetles		

Summary | Lesson 16

You can answer statistical questions using statistics and data displays. Here is data about how many hours of movies some students and teachers watched last month.

Students				
0	0	5	10	10
20	25	60	75	100

Teachers				
0	15	15	20	20
30	30	35	35	40

To make a claim about the movie habits of these students and teachers, you could create data displays like box plots or calculate measures of center (mean or median) and measures of spread (MAD or IQR).

Students	Teachers
Mean: 30.5 MAD: 28.7 Median: 15 IQR: 55	Mean: 24 MAD: 10 Median: 25 IQR: 20

From this data, you might claim that the teachers watched more movies than the students because their median is higher. You might also claim that the students watch more movies because their mean is higher.

Try This

Over a two-week period, Jada and Mai recorded the number of math problems they got for each school day.

Mai thinks that she typically has more math problems than Jada.

Is Mai's thinking correct? Explain your response.

Mai

2	15	20	0	5
25	1	0	5	17

Jada

4	0	23	7	5
0	4	10	8	5

Lesson 1

a *Responses vary.* I don't agree with Antwon. He has data on how many hours his friends use their phones for the whole week, but he can't tell how many hours they use their phones per day. He should ask a different question.

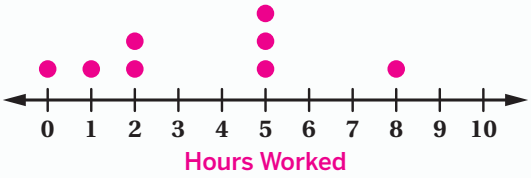
b Categorical data. *Explanations vary.* Even though Antwon is asking about a specific number of pets, 2, the answers will be "Yes" or "No," which are categorical.

Lesson 2

a The 2 dots tell us that there are 2 students in the class who read 4 books this month.

b Yes. Responses vary. I agree with Julian because 12 out of the 20 students read 6 books or less, and that is more than half of the students.

Lesson 3



Lesson 4

a Northeast

b Responses vary. These dot plots tell us that minimum wages in states in the Northeast are more different or further apart from each other. In the Midwest, minimum wages of states are more similar.

Lesson 5

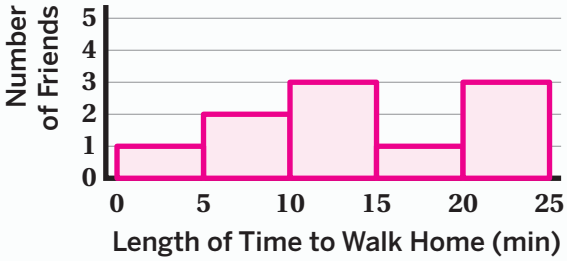
a 32 states

Caregiver Note: One strategy is to add the number of states in the 6–8 bin, which is 21, and the number of states in the 8–10 bin, which is 11.

b Bin 10–12

c No. *Explanations vary.* The 4 rectangles do not each represent a state; they represent bins that organize the data. The height of each rectangle tells you how many states are in each bin.

Lesson 6



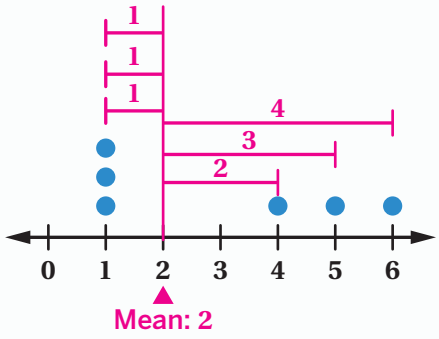
Lesson 7

a 5 tickets. *Work varies.* $\frac{7+3+4+6+8+2}{6} = 5$ (or equivalent)

b *Responses vary.* In this situation, the mean tells us how many tickets each friend would get if they shared all of the tickets equally.

Lesson 8

No. *Explanations vary.* Ava is not correct because the sum of the deviations on the left of 2 is not equal to the sum of the deviations on the right of 2.
Left: $1 + 1 + 1 = 3$
Right: $2 + 3 + 4 = 9$



Lesson 9

a 7 baskets

b 1.2 baskets

Caregiver Note: Here is a strategy to calculate the MAD. Find each data point's distance from the mean and add the sum of all the distances. Then divide the sum by the number of data points. $\frac{2+1+0+0+3}{5} = 1.2$

Lesson 10

The MAD is approximately $8.6 million.

Salaries (in millions)	Absolute Deviations
$56	22.3
$44	10.3
$35	1.3
$34	0.3
$28	5.7
$25	8.7
$24	9.7
$23.5	10.2

Lesson 11

17 inches

Lesson 12

a. Median. *Explanations vary.* The median, 36.5, is a better measure of center to represent the ages of the teachers than the mean, 40, because most of the teachers are in their thirties.

b. The mean is higher than the median because all those older ages, like 64, mixed with the younger ages and made the mean greater. The median is just the middle number, so it doesn't matter that a few of the ages are really high.

c. The mean and median are likely to be far apart when most of the data is in one place and there are a few data points that are far away.

Lesson 13

Q1	5
Q2 (Median)	25
Q3	30

Lesson 14

a

	Car	Bike
Median	7 min	10 min
IQR	9 min	3 min
Range	13 min	6 min

b By bike. *Explanations vary*. Keya should choose the bike because the range of times by car is greater than the range of times by bike.

Lesson 15

a *Estimates vary*.

	Median	IQR
Ladybugs	9	2
Beetles	9	6.5

b *Responses vary*. The medians of the two collections are the same, but the IQR of the ladybugs is much smaller. This tells us that a typical ladybug length is similar to a typical beetle length, but the ladybugs are more alike in their lengths than the beetles are in their lengths.

Lesson 16

Responses vary.

- No, because the medians of both data sets are the same: 5 math problems.
- Yes, because Jada has many lower values, so their mean, 6.6 problems, is lower than Mai's mean, 9 problems.