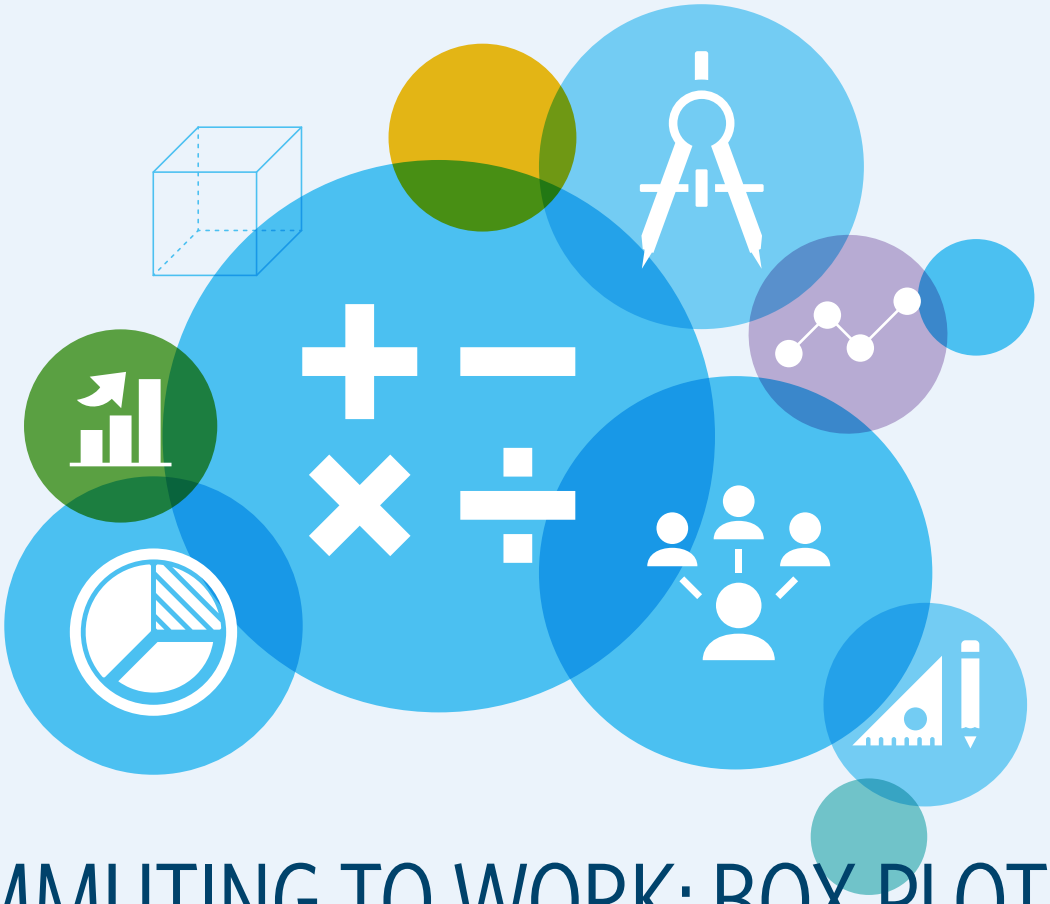


COMMUTING TO WORK: BOX PLOTS, CENTRAL TENDENCY, AND OUTLIERS

TEACHER VERSION

Subject Level:
High School Math

Grade Level:
9-12

Approx. Time Required:
45 minutes

Learning Objectives:

- Students will be able to analyze and draw conclusions about the effect of an outlier on the mean, median, and shape of a data distribution.
- Students will be able to assess the accuracy of measures of central tendency and variability in the presence of outliers.
- Students will be able to assess the utility of outliers, specifically that they may illuminate influential factors relevant to the data set that may otherwise not have been evident.
- Students will be able to make precise claims about the data set and nature of outliers.

Activity Description

Students will calculate various measures of central tendency using data on the number of people who bike to work in select states. Students will then create a box plot to represent the data set and answer conceptual questions about the impact of the data set's outlier.

Suggested Grade Level:
9–12

Approximate Time Required:
45 minutes

Learning Objectives:

- Students will be able to analyze and draw conclusions about the effect of an outlier on the mean, median, and shape of a data distribution.
- Students will be able to assess the accuracy of measures of central tendency and variability in the presence of outliers.
- Students will be able to assess the utility of outliers, specifically that they may illuminate influential factors relevant to the data set that may otherwise not have been evident.
- Students will be able to make precise claims about the data set and nature of outliers.

Topics:

- Box plots
- Central tendency
- 5-number summary
- Outliers

Skills Taught:

- Analyzing the impact of an outlier on the mean and median
- Calculating a 5-number summary
- Calculating measures of central tendency and variability
- Creating a box plot

Materials Required

- The student version of this activity, 9 pages
- Calculators

Activity Item

The following item is part of this activity. The item and its source appear at the end of this teacher version.

- Item 1: Number of People Who Bike to Work in Select States

For more information to help you introduce your students to the U.S. Census Bureau, read "*Census Bureau 101 for Students*." This information sheet can be printed and passed out to your students as well.

Standards Addressed

See charts below. For more information, read "*Overview of Education Standards and Guidelines Addressed in Statistics in Schools Activities*."

Common Core State Standards for Mathematics

Standard	Domain	Cluster
CCSS.MATH.CONTENT.HSS.ID.A.1 Represent data with plots on the real number line (dot plots, histograms, and box plots).	ID – Interpreting Categorical & Quantitative Data	Summarize, represent, and interpret data on a single count or measurement variable.
CCSS.MATH.CONTENT.HSS.ID.A.3 Interpret differences in shape, center, and spread in the context of the data sets, accounting for possible effects of extreme data points (outliers).	ID – Interpreting Categorical & Quantitative Data	Summarize, represent, and interpret data on a single count or measurement variable.

Common Core State Standards for Mathematical Practice

Standard
CCSS.MATH.PRACTICE.MP2. Reason abstractly and quantitatively. Students will contextualize the data to assess the impact and utility of the outlier as well as use reason to critique the methodology.
CCSS.MATH.PRACTICE.MP4. Model with mathematics. Students will create a statistical model (box plot) to represent a data set.
CCSS.MATH.PRACTICE.MP6. Attend to precision. Students will communicate precisely to others, using appropriate vocabulary and clarifying the correspondence between quantities and claims made about them.

National Council of Teachers of Mathematics' Principles and Standards for School Mathematics

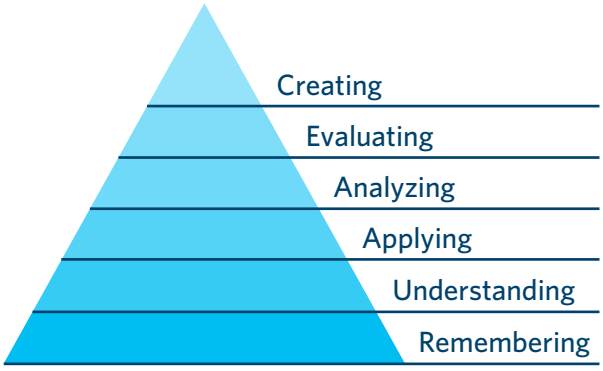
Content Standard	Students should be able to:	Expectation for Grade Band
Data Analysis and Probability	Formulate questions that can be addressed with data and collect, organize, and display relevant data to answer them.	Understand histograms, parallel box plots, and scatterplots and use them to display data.

Guidelines for Assessment and Instruction in Statistics Education

GAISE	Level A	Level B	Level C
Formulate Questions			
Collect Data			
Analyze Data			X
Interpret Results			X

Bloom's Taxonomy

Students will ***analyze*** data by making calculations and comparisons, as well as by distinguishing between facts and inferences.



Teacher Notes

Before the Activity

Students must understand the following key terms:

- **Box plot** – a method of visually displaying a data set using the median, quartiles, and extremes of the data set
- **Interquartile range (IQR)** – a measure of variability in a set of numerical data to indicate the difference between the lower and upper quartiles of the data set
- **Outlier** – an extremely high or low value that noticeably differs from the other data points in the set; it is any point that falls 1.5 times the IQR below the lower quartile or above the upper quartile

Students should have the following skills:

- Ability to create a box plot
- Ability to identify and calculate the mean, median, quartiles, minimum, maximum, and a 5-number summary
- Ability to interpret, draw conclusions, and make inferences from data

The teacher should explain to students that the data in this activity come from the American Community Survey (ACS), a survey conducted monthly by the Census Bureau and designed to show how communities are changing. Through asking questions of a sample of the population, it produces national data on more than 35 categories of information, such as education, income, housing, and employment.

During the Activity

Teachers will decide whether and to what extent students should show their work for prompts 1 through 4. Teachers may choose to put students in small groups to encourage collaboration and discussion of findings.

After the Activity

Teachers should review the activity with students by calling on individuals or groups to share their solutions and answers with the class. Teachers will facilitate a class discussion in which students share and respond to each other's observations and arguments about the data. Teachers should tell students to use these sentence starters:

- I notice ...
- I wonder ...
- Similarities I see are ...
- Differences I see are ...
- It is surprising that ...

Teachers should then create groups and assign each of them one of the questions that focus on interpretations or conclusions (i.e., questions 2, 5, 7, 8, and 9), allotting 2 minutes for each group to share its response with the class.

Extension Ideas

- Teachers could have students read an article about a bike-sharing system (e.g., Citi Bike in New York City) and discuss what they think the system's impact might be on the number of commuters who bike.
- Teachers could have students read the American Community Survey report "Modes Less Traveled—Bicycling and Walking to Work in the United States: 2008–2012" (*www2.census.gov/library/publications/2014/acs/acs-25.pdf*).
- Teachers could adapt this activity for different GAISE levels.
  - GAISE Level B: Teachers could have students collect data on how classmates usually commute to school and use these data to answer the activity questions.
  - GAISE Level C: Teachers could have students collect data on a subject of their choice, identify potential sampling error due to their methodology, and answer the activity questions.

Student Activity

Click ***here*** to download a printable version for students.

Activity Item

The following item is part of this activity and appears at the end of this student version.

- Item 1: Number of People Who Bike to Work in Select States

Student Learning Objectives

- I will be able to analyze and draw conclusions about the effect of an outlier on the mean, median, and shape of a data distribution.
- I will be able to assess the accuracy of measures of central tendency and variability in the presence of outliers.
- I will be able to assess the utility of outliers, specifically that they may illuminate influential factors relevant to the data set that may otherwise not have been evident.
- I will be able to make precise claims about the data set and nature of outliers.

Use **Item 1: Number of People Who Bike to Work in Select States** to answer the following questions and prompts.

1. Calculate the following measures of central tendency for this data set.
   a. Mean (round to the tenths place):

      11,107.1 people

      Calculation work (if teacher requests):

      1,950 + 2,843 + 4,227 + 1,221 + 53,759 + 10,339 + 6,860 + 4,916 + 9,237 + 15,719 = 111,071

      $$\frac{111{,}071}{10} = 11{,}107.1 \text{ people}$$

   b. Median:

      5,888 people

      Calculation work (if teacher requests):

      ~~1,221~~ ~~1,950~~ ~~2,843~~ ~~4,227~~ 4,916 | 6,860 ~~9,237~~ ~~10,339~~ ~~15,719~~ ~~53,759~~

      $$\frac{4{,}916 + 6{,}860}{2} = 5{,}888 \text{ people}$$

2. Which measure of central tendency (above) is more representative of this data set? Explain.

   The median is more representative because it is not as strongly influenced by the extreme value of 53,759 people, which is very different from the other values in the data set.

3. Calculate the 5-number summary for this data set.
   a. Minimum:

      1,221 people

   b. Lower quartile (Q_1):

      2,843 people

      Calculation work (if teacher requests):

      ~~1,221 1,905~~ 2,843 ~~4,227 4,916~~ | 6,860 9,237 10,339 15,719 53,759

   c. Median (from question 1):

      5,888 people

   d. Upper quartile (Q_3):

      10,339 people

      Calculation work (if teacher requests):

      1,221 1,905 2,843 4,227 4,916 | ~~6,860 9,237~~ 10,339 ~~15,719 53,759~~

   e. Maximum:

      53,759 people

4. Calculate the following measures of variability for this data set.
   a. Range:

      52,538 people

      Calculation work (if teacher requests):

      maximum - minimum

      53,759 - 1,221 = 53,759 people

b. Interquartile range (IQR):

7,496 people

Calculation work (if teacher requests):

$Q_3 - Q_1$

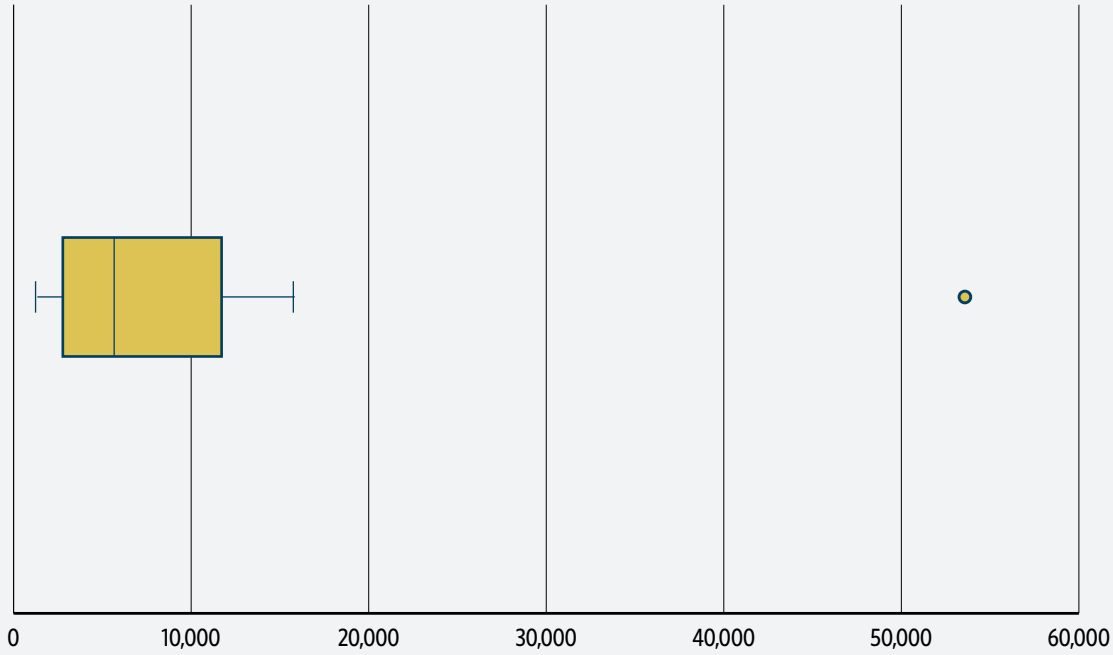
10,339 - 2,843 = 7,496 people

5. Which measure of variability (above) provides a better description of this data set? Explain.

The interquartile range provides a better description because it is not as strongly influenced by the extreme value as the range.

6. Create a box plot in the space below to represent the data set, making sure to label all its components.

Number of People Who Bike to Work in Select States



Number of People

7. Assess the potential impact of an outlier in the data set by answering the questions and prompts below:

a. By just looking at **Item 1: Number of People Who Bike to Work in Select States**, which state do you think might represent an outlier in this data set?

Florida

b. For a value to be considered an outlier, it must be greater than (1.5*IQR) + Q_3 or less than Q_1 - (1.5*IQR). Use math to verify whether the state you identified above qualifies as an outlier for this data set.

(1.5*IQR) + Q_3

(1.5*7,496) + 10,339 = 21,583 people

53,759 people > 21,583 people

c. Describe the impact of this outlier on the mean, median, and shape of the data distribution.

This outlier significantly increases the mean, does not have a significant impact on the median, and causes the shape of the distribution to skew right.

d. Why do you think the number of people who bike to work in the outlier state differs noticeably from that of the other states?

Student answers will vary but could include: The warmer weather could influence more people to bike to work in Florida than in other states. Florida could also have a larger population, which means there are more people, in general, compared with other states.

e. List additional states not included in **Item 1** that you suspect might also be high outliers in terms of the number of people who bike to work. Explain your reasoning.

Student answers will vary but could include states known for warm climates (Arizona, Hawaii, or Texas) or those with large populations (California or New York).

f. Which of all the U.S. states would you predict has the greatest number of people who commute by bicycle. Why?

Student answers will vary but could include California or Texas, with the reasoning that each has a large population and a warm climate. (In the 2020 ACS data, California has the greatest number of people who commuted by bicycle: 153,669.)

g. Based on the mathematical definition of an outlier, is it possible that a state not listed in **Item 1** is a low outlier with respect to the number of people who bike to work? Explain your reasoning.

No. To qualify as a low outlier, a state would have to have fewer bike commuters than Q_1 – (1.5*IQR), which is 1,221 - (1.5*7,496) or -20,362 people. It is not possible for a state to have a negative number of bike commuters.

8. Assess the potential utility of outliers by answering the questions and prompts below.

a. How does the presence of an outlier in this particular data set help bring attention to significant factors that might influence the number of people who bike to work in a state? List one or two possible factors.

The presence of the outlier brings attention to why there might be a higher number of bike commuters in Florida compared with the other states. Student examples of significant factors will vary but could include population size and weather.

b. Do you think any of these significant factors could be identified if the outlier were not present? Explain your reasoning.

Student answers will vary but could include: Maybe not; population size and weather may not be salient factors when considering the number of bike commuters in a state. The presence of Florida in the data table helps highlight the significance of population size and weather that might not be evident from the data for other states.

c. Thinking beyond this data set, what general role might outliers play in developing an understanding of data sets?

In general, outliers may be useful in drawing attention to underlying factors that influence data.

9. Assess the methodology of this data collection by answering the questions below.

a. How were the data collected?

The data were collected through the American Community Survey 5-year estimates, an ongoing sample survey on which the U.S. Census Bureau produces annual data.

b. Are the states included in this data set the appropriate states to compare? Explain your reasoning.

Student answers will vary but could include: The states included could be appropriate because their features vary, which creates a more representative sample of data. Alternatively, grouping states with similar climates or similar populations would create another reasonable lens for comparison.

c. Thinking about how the number of bike commuters was calculated in this data set, what impact might the sheer population of a state have on the number of its residents who bike to work?

There are likely more people who bike to work in states with larger populations simply because the data are reported by the number of people in each state, not by the rate or percentage of people in each state.

d. How else could the number of people who bike to work be calculated to account for the impact of state population?

Student answers will vary but could include: The number could be calculated as a percentage rather than as a total number, to take state population size into account and thereby isolate the prevalence of commuting by bicycle in each state.

Item 1: Number of People Who Bike to Work in Select States

State	Number of Bicycle Commuters
Alabama	1,950
Alaska	2,843
Connecticut	4,227
Delaware	1,221
Florida	53,759
Georgia	10,339
Idaho	6,860
Kansas	4,916
Maryland	9,237
Virginia	15,719

Source for data: *U.S. Census Bureau, Means of Transportation to Work. 2016-2020 American Community Survey 5-Year Estimates.*