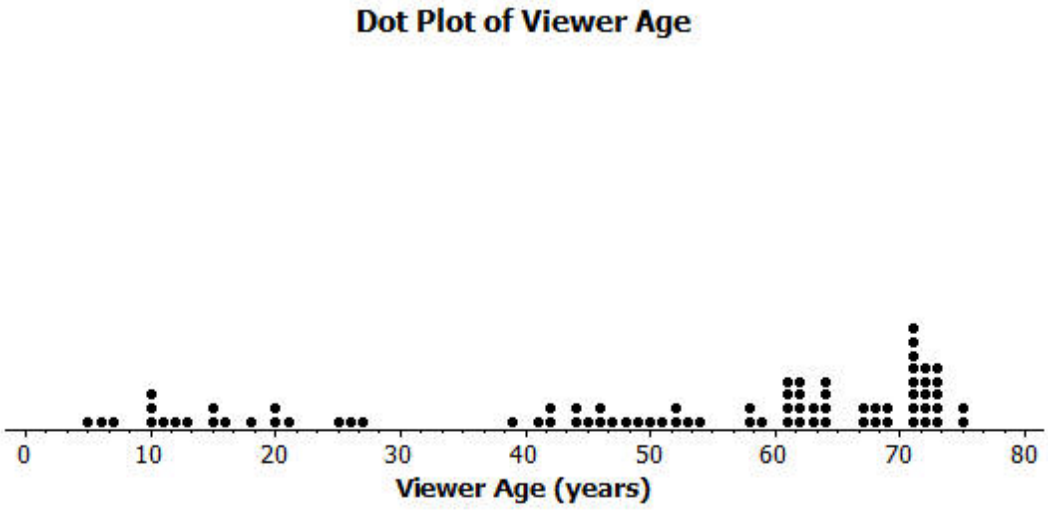
Lesson 5: Box Plots

Opening Exercise

Consider the following scenario.

A television game show, *Fact or Fiction*, was cancelled after nine shows. Many people watched the nine shows and were rather upset when it was taken off the air. A random sample of eighty viewers of the show was selected. Viewers in the sample responded to several questions. The dot plot below shows the distribution of ages of these eighty viewers.



A data distribution that is not symmetrical is described as *skewed*. In a skewed distribution, data "stretch" either to the left or to the right. The stretched side of the distribution is called a *tail*.

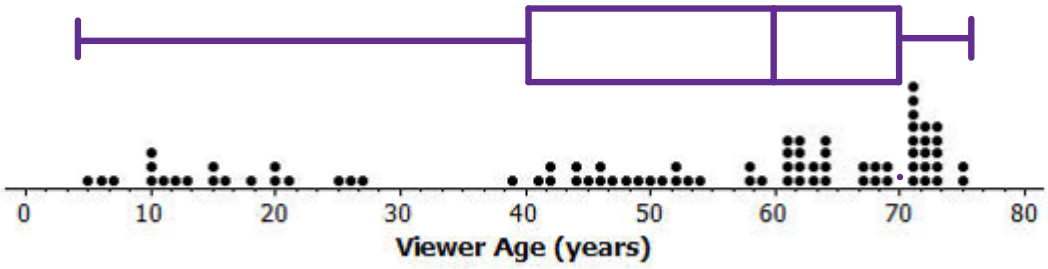
1. Would you consider this data set to be skewed? Explain your thinking.

This work is derived from Eureka Math ™ and licensed by Great Minds. ©2015 Great Minds. eureka-math.org
This file derived from ALG I-M2-TE-1.3.0-08.2015

BY-NC-SA This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

Exploratory Challenge 1 – Constructing and Interpreting the Box Plot

2. Using the dot plot in the Opening Exercise, construct a box plot **over** the dot plot by completing the following steps. Recall that there are 80 data points in the dot plot.



i. Locate the middle 40 observations, and draw a box around these values.

ii. Calculate the median, and then draw a vertical line in the box at the location of the median.

$\frac{59+61}{2} = \frac{120}{2} = 60$

iii. Draw a line that extends from the upper end of the box to the largest observation in the data set.

iv. Draw a line that extends from the lower edge of the box to the minimum value in the data set.

3. Recall that the five values used to construct the box plot make up the 5-number summary. What is the 5-number summary for this data set of ages?

Minimum age: 6

Lower quartile or Q1: 40

Median age: 60

Upper quartile or Q3: 70

Maximum age: 76

This work is derived from Eureka Math ™ and licensed by Great Minds. ©2015 Great Minds. eureka-math.org
This file derived from ALG I-M2-TE-1.3.0-08.2015

BY-NC-SA This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

4. A. What percent of the data does the box part of the box plot capture?

50%

B. What percent of the data fall between the minimum value and Q1?

25%

C. What percent of the data fall between Q3 and the maximum value?

25%

5. Why do we use the median for a box plot?

It's not impacted by the skew

6. What are the advantages and challenges to using a box plot?

Fill in each blank with the appropriate word from the word bank.

7. Each section is called a quartile, since the data is split into 4 sections (quarters).

8. The box is also called the Interquartile range or IQR

9. Each section holds $\frac{1}{4}$ of the data.

10. The IQR can be determined by subtracting the first quartile from the third quartile.

Word Bank

first	four	Interquartile Range
IQR	one-fourth or 25%	quarters
quartile	section	third

This work is derived from Eureka Math ™ and licensed by Great Minds. ©2015 Great Minds. eureka-math.org
This file derived from ALG I-M2-TE-1.3.0-08.2015

BY-NC-SA This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

Exploratory Challenge 2 – Comparing Data

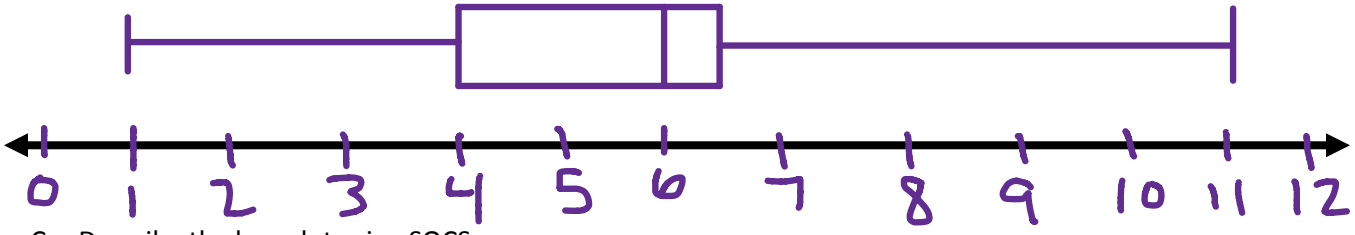
11. Ron is taking a survey to find out how many pencils each of his friends have. The data is below.

    Number of pencils in their pencil pouch: 1, 2, 4, 4, 4, 4, 5, 5, 6, 6, 6, 6, 6, 7, 8, 10, 11

    A. What is the 5- Number Summary for this data?

    Minimum = 1; Q1 = 4; Median = 6; Q3 = 6.5; Maximum = 11

    B. Draw the box plot below.



    C. Describe the box plot using SOCS.

12. Neville joins the group and has 3 pencils in his pencil pouch. The updated data is below.

    Number of pencils in their pencil pouch: 1, 2, 3, 4, 4, 4, 4, 5, 5, 6, 6, 6, 6, 6, 7, 8, 10, 11

    A. What is the 5- Number Summary for this data?

    Minimum = 1; Q1 = 4; Median = 5.5; Q3 = 6; Maximum = 11

    B. Draw the box plot below.

    C. Describe the box plot using SOCS.

This work is derived from Eureka Math ™ and licensed by Great Minds. ©2015 Great Minds. eureka-math.org
This file derived from ALG I-M2-TE-1.3.0-08.2015

BY-NC-SA This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

13. Did Neville's data change the box plot significantly?

Exploratory Challenge 2 – Comparing Data

14. Hermione joins the group and has 20 pencils in her pencil pouch. Do you think 20 an outlier for this data set? Explain your thinking.

A data distribution may contain extreme data (unusually large or unusually small relative to the median and the IQR). A box plot can be used to display extreme data values that are identified as ***outliers***. We often use a dot (•) or an asterisk (*) to identify outliers on a box plot.

An outlier is defined to be any data value that is more than $1.5 \times (IQR)$ away from the nearest quartile.

Lower Boundary = Q1 – 1.5 x IQR

Upper Boundary = Q3 + 1.5 x IQR

This work is derived from Eureka Math ™ and licensed by Great Minds. ©2015 Great Minds. eureka-math.org
This file derived from ALG I-M2-TE-1.3.0-08.2015

BY-NC-SA This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

15. Hermione joins the group and has 20 pencils in her pencil pouch. The updated data is below.

Number of pencils in their pencil pouch: 1, 2, 3, 4, 4, 4, 4, 5, 5, 6, 6, 6, 6, 7, 8, 10, 11, 20

A. What is the 5- Number Summary for this data?

Minimum = 1; Q1 = 4; Median = 6; Q3 = 7; Maximum = 20

B. Calculate the IQR (interquartile range).

$7 - 4 = 3$

C. Do you think 20 is an outlier? How can we know for sure?

$1.5 \times 3 = 4.5$ $7 + 4.5 = 11.5$

D. Determine if 20 is an outlier for this data set.

Yes

E. Draw the box plot below.

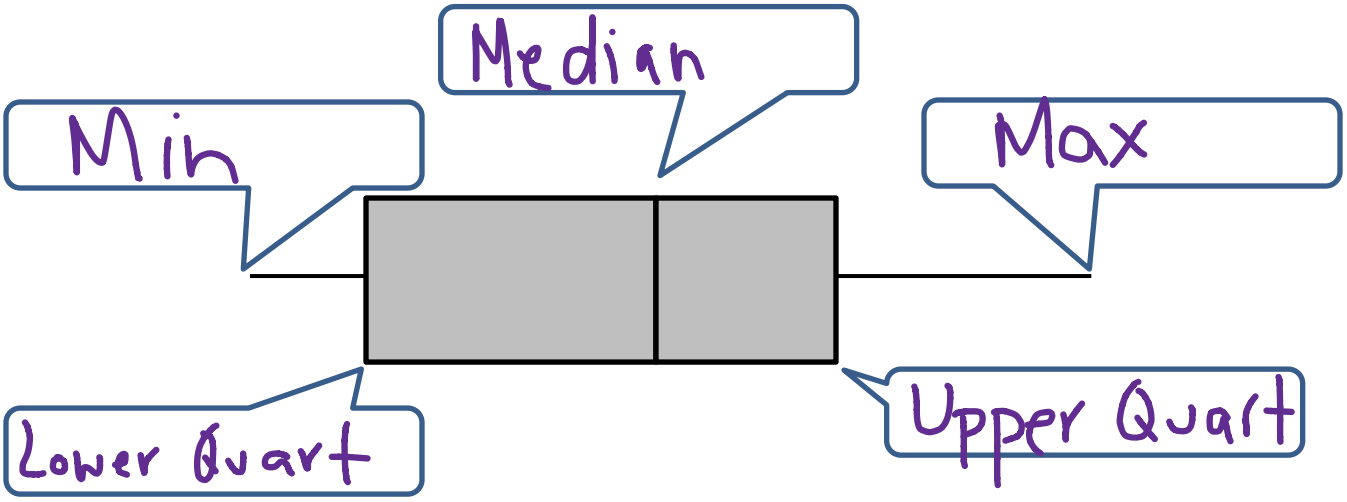
F. How did the box plot change by adding Hermione's 20 pencils? What parts changed very little? What parts changed significantly?

This work is derived from Eureka Math ™ and licensed by Great Minds. ©2015 Great Minds. eureka-math.org
This file derived from ALG I-M2-TE-1.3.0-08.2015

BY-NC-SA This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

Lesson Summary

16. Use the diagram and the word list to identify the five-number summary that makes up a box plot. Then complete the sentences.



Word Bank:

Lower Quartile	Upper Quartile	Maximum	Median	Minimum

- Nonsymmetrical data distributions are referred to as skewed.
- Left-skewed or skewed to the left means the data spread out longer (like a tail) on the left side.
- Right-skewed or skewed to the right means the data spread out longer (like a tail) on the right side.
- The center of a skewed data distribution is described by the median.
- Variability of a skewed data distribution is described by the interquartile range (IQR).
- The IQR describes variability by specifying the length of the interval that contains the middle 50 % of the data values.
- Outliers in a data set are defined as those values more than $1.5 \times (IQR)$ from the nearest quartile. Outliers are usually identified by an "*" or a "•" in a box plot.

This work is derived from Eureka Math ™ and licensed by Great Minds. ©2015 Great Minds. eureka-math.org
This file derived from ALG I-M2-TE-1.3.0-08.2015

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

Homework Practice Set

An advertising agency researched the ages of viewers most interested in various types of television ads. Consider the following summaries:

Ages	Target Products or Services
30–45	Electronics, home goods, cars
46–55	Financial services, appliances, furniture
56–72	Retirement planning, cruises, health-care services

1. The mean age of the people surveyed is approximately 50 years old. As a result, the producers of the show decided to obtain advertisers for a typical viewer of 50 years old.

   A. According to the table, what products or services do you think the producers will target?

   B. Based on the sample, what percent of the people surveyed about the *Fact or Fiction* show would have been interested in these commercials if the advertising table is accurate?

2. The show failed to generate the interest the advertisers hoped. As a result, they stopped advertising on the show, and the show was cancelled. Kristin made the argument that a better age to describe the typical viewer is the median age.

   A. What is the median age of the sample?

   B. What products or services does the advertising table suggest for viewers if the median age is considered as a description of the typical viewer?

   C. What percent of the people surveyed would be interested in the products or services suggested by the advertising table if the median age were used to describe a typical viewer?

This work is derived from Eureka Math ™ and licensed by Great Minds. ©2015 Great Minds. eureka-math.org
This file derived from ALG I-M2-TE-1.3.0-08.2015

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

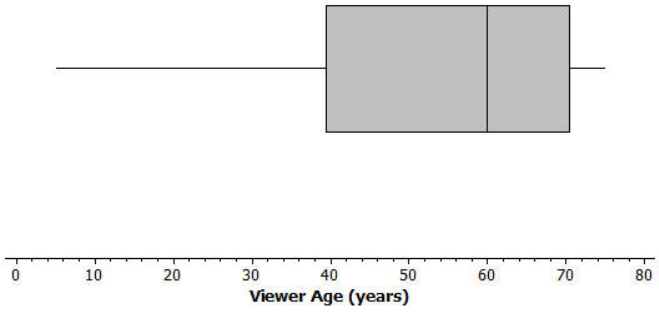
3. A. What percent of the viewers have ages between Q1 and Q3?

   B. The difference between Q3 and Q1, or Q3 – Q1, is called the **interquartile range**, or **IQR**. What is the IQR for this data distribution?

4. Do you think producers of the show would prefer a show that has a small or large interquartile range? Explain your answer.

5. Do you agree with Kristin's argument that the median age provides a better description of a typical viewer? Explain your answer.

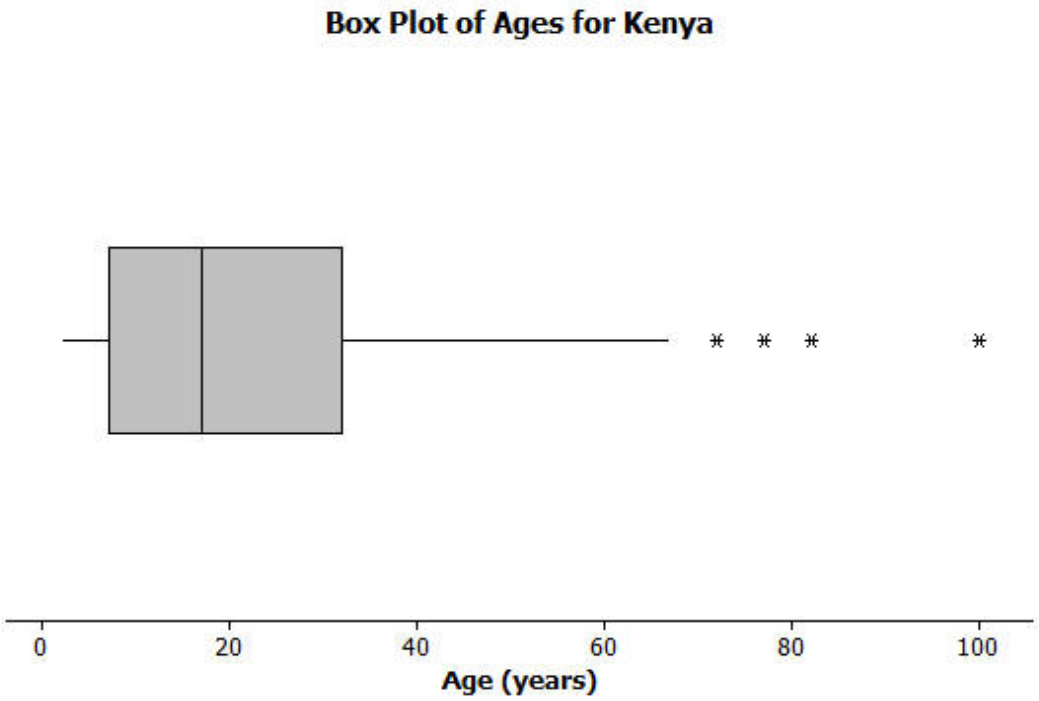
6. Which ages, if any, do you think are outliers for the viewer ages in the box plot below?



This work is derived from Eureka Math ™ and licensed by Great Minds. ©2015 Great Minds. eureka-math.org
This file derived from ALG I-M2-TE-1.3.0-08.2015

This work is licensed under a
Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

Students at Waldo High School are involved in a special project that involves communicating with people in Kenya. Consider a box plot of the ages of 200 randomly selected people from Kenya.



The four "*"s in the box plot represents the ages of four people from this sample. Based on the sample, these four ages were considered outliers.

7. Estimate the values of the four ages represented by an *.

Remember: An outlier is defined to be any data value that is more than $1.5 \times (IQR)$ away from the nearest quartile.

This work is derived from Eureka Math ™ and licensed by Great Minds. ©2015 Great Minds. eureka-math.org
This file derived from ALG I-M2-TE-1.3.0-08.2015

This work is licensed under a
Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

8. A. What is the median age of the sample of ages from Kenya?

B. What are the approximate values of Q1 and Q3?

C. What is the approximate IQR of this sample?

D. Multiply the IQR by 1.5. What value do you get?

E. Add $1.5 \times (IQR)$ to the third quartile age (Q3). What do you notice about the four ages identified by an *?

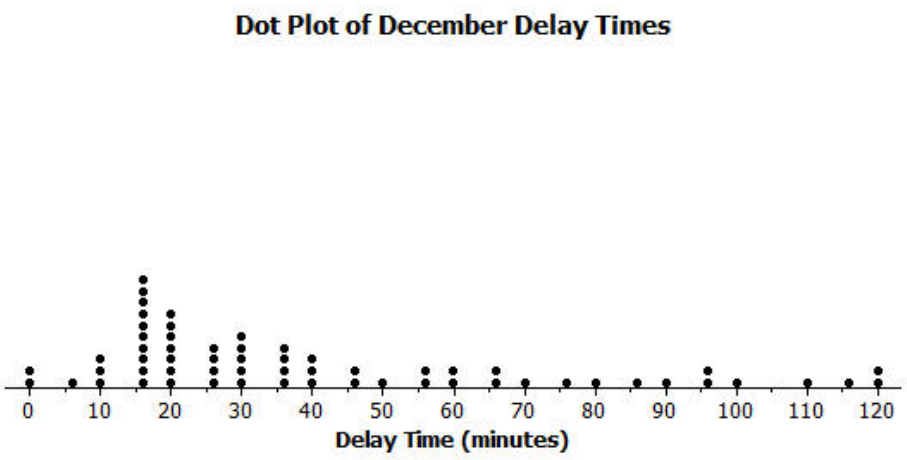
F. Are there any age values that are less than $Q1 - 1.5 \times (IQR)$? If so, these ages would also be considered outliers.

G. Explain why there is no * on the low side of the box plot for ages of the people in the sample from Kenya.

This work is derived from Eureka Math ™ and licensed by Great Minds. ©2015 Great Minds. eureka-math.org
This file derived from ALG I-M2-TE-1.3.0-08.2015

BY-NC-SA This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

Consider the following scenario. Transportation officials collect data on flight delays (the number of minutes a flight takes off after its scheduled time). Consider the dot plot of the delay times in minutes for 60 BigAir flights during December 2012.



9. How many flights left more than 60 minutes late?

10. Why is this data distribution considered skewed?

11. Is the tail of this data distribution to the right or to the left? How would you describe several of the delay times in the tail?

12. Draw a box plot over the dot plot of the flights for December.

13. What is the interquartile range, or IQR, of this data set?

14. The mean of the 60 flight delays is approximately 42 minutes. Do you think that 42 minutes is typical of the number of minutes a BigAir flight was delayed? Why or why not?

15. Based on the December data, write a brief description of the BigAir flight distribution for December.

This work is derived from Eureka Math ™ and licensed by Great Minds. ©2015 Great Minds. eureka-math.org
This file derived from ALG I-M2-TE-1.3.0-08.2015

BY-NC-SA This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

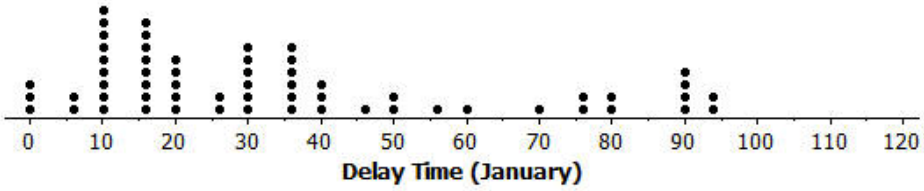
16. Calculate the percentage of flights with delays of more than 1 hour. Were there many flight delays of more than 1 hour?

17. BigAir later indicated that there was a flight delay that was not included in the data. The flight not reported was delayed for 48 hours. If you had included that flight delay in the box plot, how would you have represented it? Explain your answer.

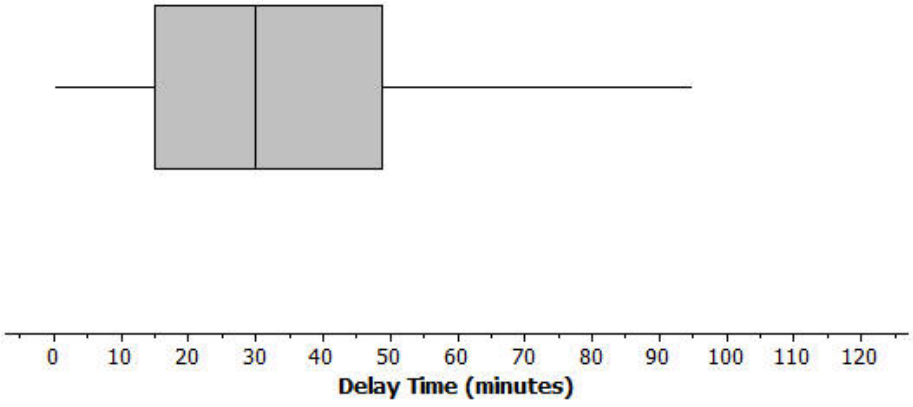
This work is derived from Eureka Math ™ and licensed by Great Minds. ©2015 Great Minds. eureka-math.org
This file derived from ALG I-M2-TE-1.3.0-08.2015

This work is licensed under a
Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.

18. A. Consider a dot plot and the box plot of the delay times in minutes for 60 BigAir flights during January 2013. How is the January flight delay distribution different from the one summarizing the December flight delays? In terms of flight delays in January, did BigAir improve, stay the same, or do worse compared to December? Explain your answer.



Box Plot of January Delay Times



B. Do you think this data set contains any outliers? Explain your thinking.

This work is derived from Eureka Math ™ and licensed by Great Minds. ©2015 Great Minds. eureka-math.org
This file derived from ALG I-M2-TE-1.3.0-08.2015

This work is licensed under a
Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License.