Exploratory Challenge: Outliers

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.



A data distribution may contain extreme data (specific data values that are unusually large or unusually small relative to the median and the interquartile range). A box plot can be used to display extreme data values that are identified as **outliers**.

Each “\*” in the box plot represents the ages of four people from this sample. Based on the sample, these four ages were considered outliers.

1. Estimate the values of the four ages represented by an \*.

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

1. What is the median age of the sample of ages from Kenya? What are the approximate values of Q1 and Q3? What is the approximate IQR of this sample?
2. Multiply the IQR by $1.5$. What value do you get?
3. Add $1.5×(IQR)$ to the 3rd quartile age (Q3). What do you notice about the four ages identified by an \*?
4. Are there any age values that are less than $Q1-1.5×(IQR)$? If so, these ages would also be considered outliers.
5. Explain why there is no \* on the low side of the box plot for ages of the people in the sample from Kenya.

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:



1. How many flights left more than $60$ minutes late?
2. Why is this data distribution considered skewed?
3. 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?
4. Draw a box plot over the dot plot of the flights for December.
5. What is the interquartile range, or IQR, of this data set? Be sure to show your work.
6. 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?
7. Based on the December data, write a brief description of the BigAir flight distribution for December.
8. Calculate the percentage of flights with delays of more than $1$ hour. Were there many flight delays of more than $1$ hour?
9. 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.
10. 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.