Statistics 10 - Homework 1

Name: Preyasi Gaur (UID 705704939)

Exercise 1

1. There are 15 variables contained in the dataset. Age is the numerical variable.
2. Given the dataset, we cannot answer (ii). For (ii), we require data on the temperature of the water which is not present in the dataset.
3. To answer these questions, I would use a histogram to plot the data between shark attacks and age. This would help show the distribution of data, and the most frequently occurring age data. More so, it would help to measure the range of the data.

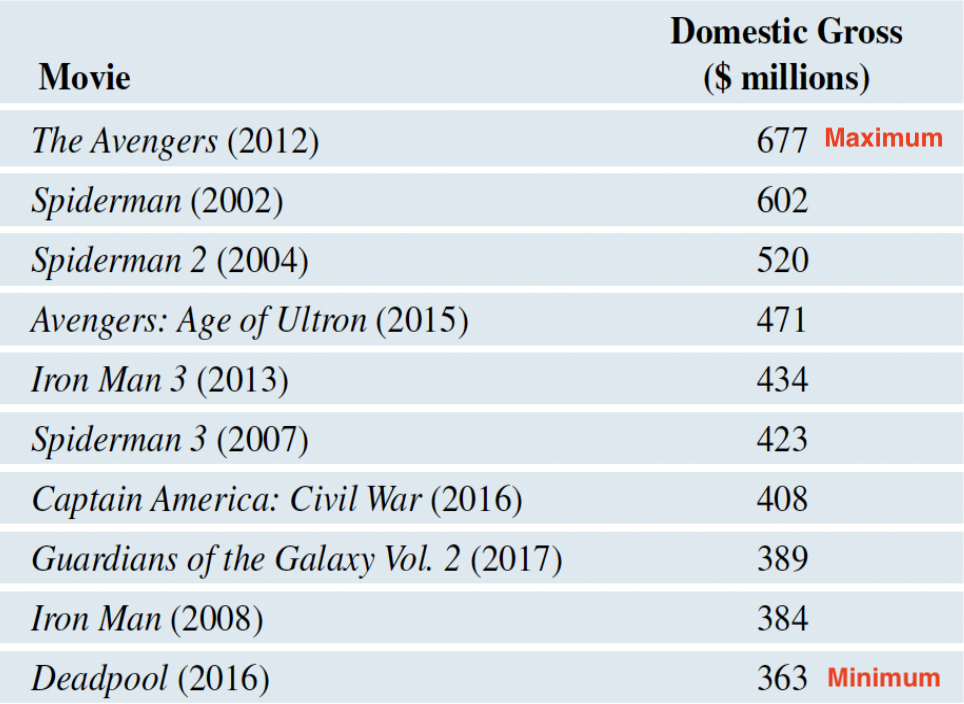
Depending on the spread of the data, we would also use median/mean to find the center of the data. The interquartile range (IQR) can also tell us about the spread of the distribution.

Exercise 2

1. The distribution is left-skewed (negatively skewed)
2. The mean score here would be lesser than the median as the distribution is left-skewed. Here, as the mean is calculated by taking into account the few, but small observations, it would lead to it being lesser than median, which does not depend on the observations.
3. As the graph is negatively skewed, the five number summary will be useful to gain information about the spread and center. The median would help determine the center. The IQR, minimum and maximum will help show the spread of the distribution.

Exercise 5

1. The data is arranged in the descending order



Clearly here, **maximum (in $ millions) = 677** and **minimum (in $ millions) = 363**.   
As there are 10 observations, the median is the midpoint of 5th and 6th observations i.e the midpoint of 434 and 423.

**median(in $ millions) = = 428.5**

Q1 is the midpoint of the first half of the data. Thus, **Q1 (in $ millions)= 389**.

Q3 is the midpoint of the second half of the data. Thus, **Q3 (in $ millions) = 520**

**The Interquartile Range (IQR) (in $ millions)** = Q3 - Q1

= 520 - 389

= **131**

Thus, the five number summary (all values in $ million) is:

Minimum Q1 Median Q3 Maximum

363 389 428.5 520 677

1. Putting the five number summary (all values in $ million) in context:

Minimum Q1 Median Q3 Maximum

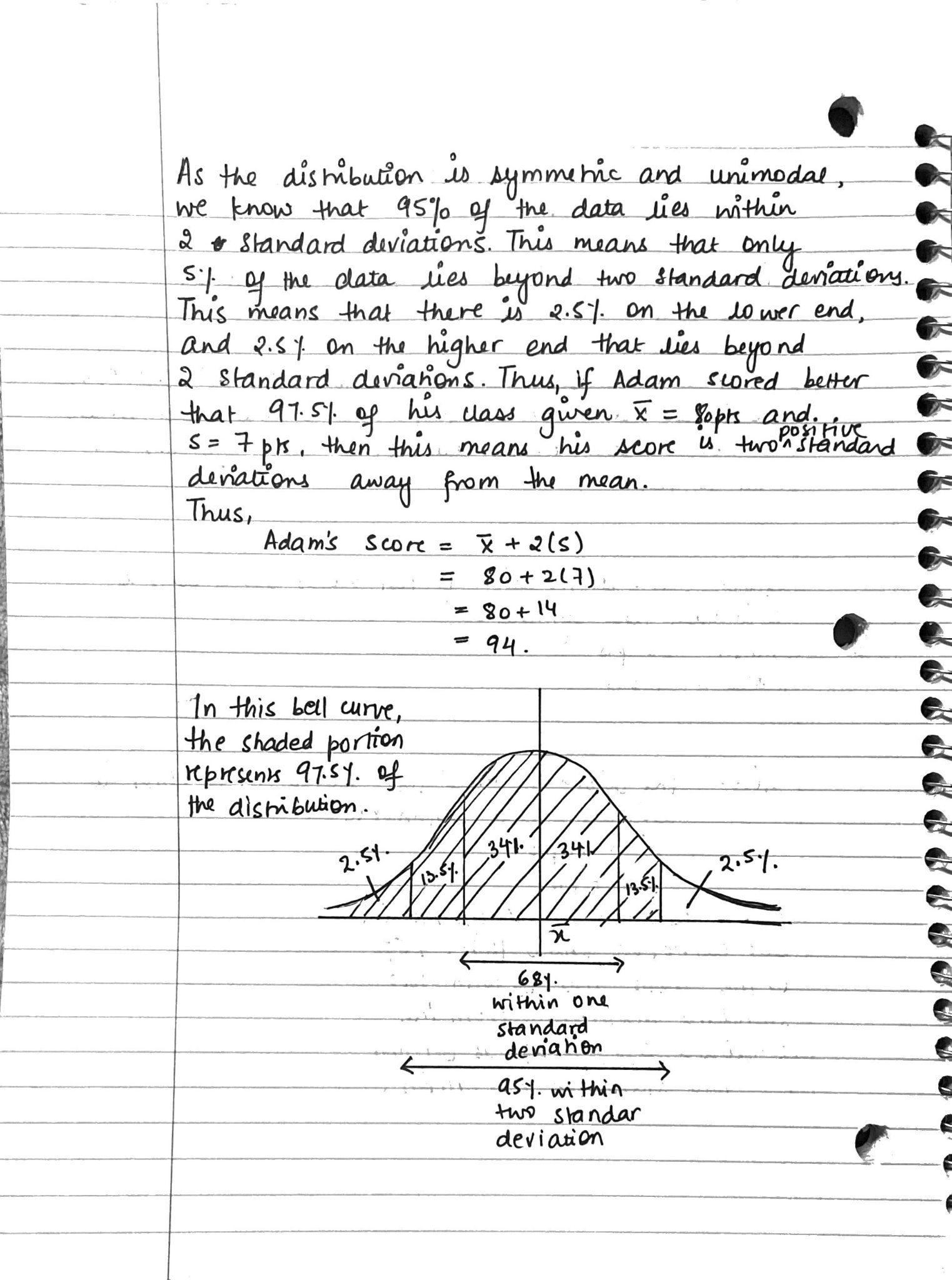
363 389 428.5 520 677

Thus, from the five number summary we see that the highest grossing movie on the U.S box office as of fall 2017 (i.e.the highest value) was Avenger, which collected $677 million. We also see that the tenth highest grossing movie on the U.S box office as of fall 2017 (i.e the lowest value) was Deadpool which collected $363 million. This tells us that the domestic box office collections of the top ten highest grossing movies as of fall 2017 lie between $363 million and $677 million.   
The median divides the dataset into half and is representative of central tendency. Here, the median is $428.5 million and this means that of the top ten highest domestic grossing movies of fall 2017, half collected more than $428.5 million and half collected below that.   
Q1 and Q3 help tell us that 50% of the top ten highest grossing domestic movies collections as of fall 2017 lie in the range of $389 million and $520 million. And more so, it also tells us that 50% of the observations lie beyond this range. We can also use the IQR (in $million) = 131 to determine the potential outliers in the dataset i.e movies that performed really poorly or exceedingly well.   
Upper limit : Q3 + 1.5IQR = 716.5

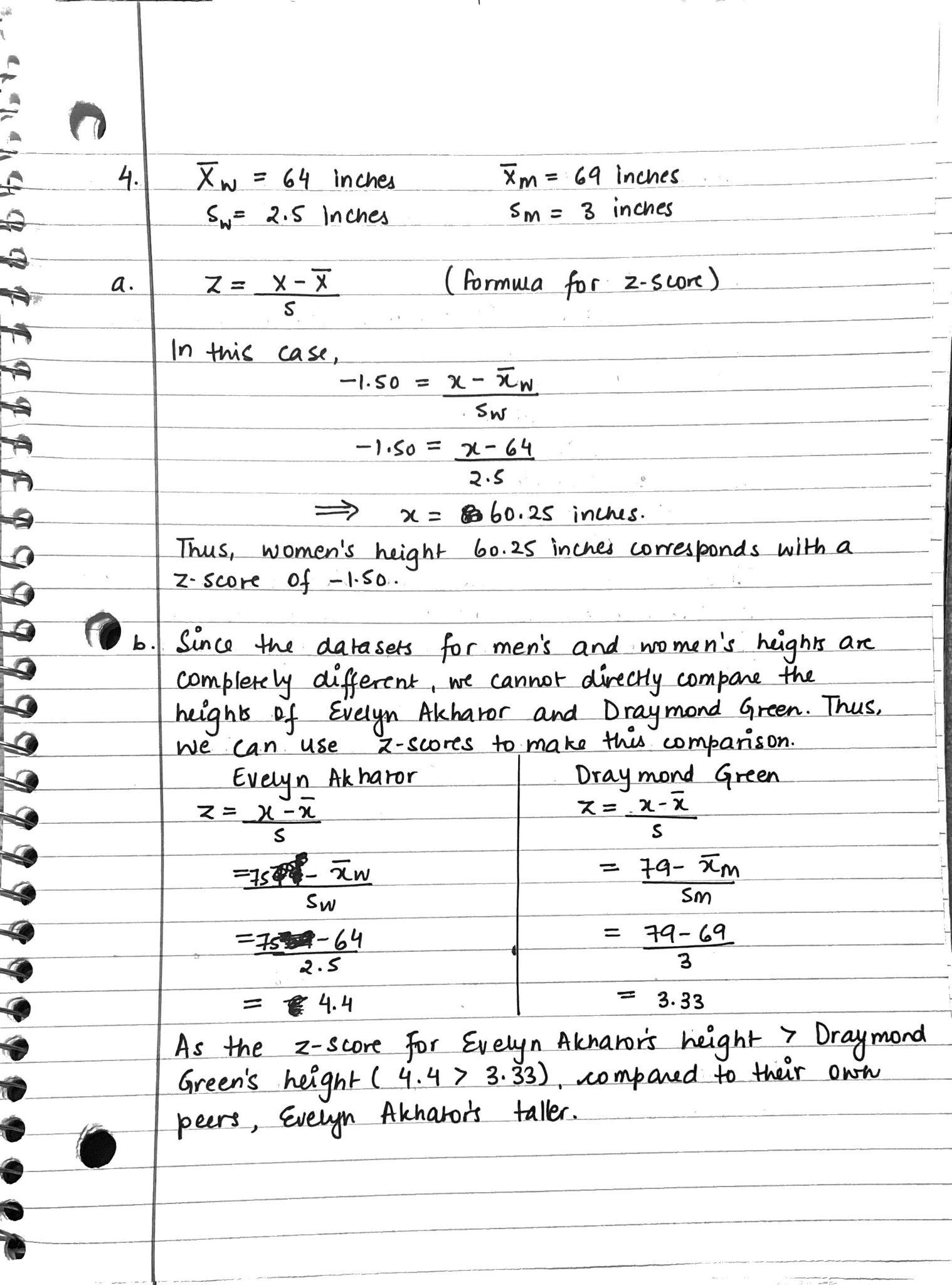
Lower limit: Q1 - 1.5IQR = 166.5

As all the values lie between the range of $166.5 million and $716.5 million, we see that there are no outliers.

Exercise 3



Exercise 4



Exercise 6

