Shopify Summer 2022 Data Science Intern Challenge

Jennifer Zhang

Question 1

a. Think about what could be going wrong with our calculation. Think about a better way to evaluate this data.

Import Data and Generate Summary

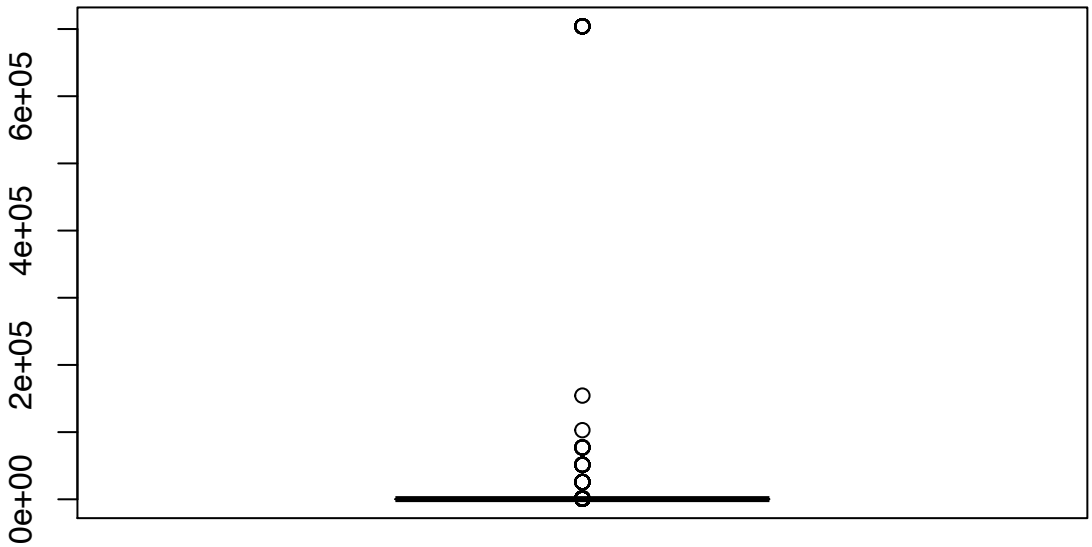
```
data <- read.csv("2019 Winter Data Science Intern Challenge Data Set - Sheet1.csv")

summary(data$order_amount)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      90     163     284    3145     390  704000
```

Through running a summary of the data of the order value of the shoes from the 100 sneaker shops, we can see that the AOV of $ 3145.13 refers to the mean. Knowing that the maximum order is 704000, it is possible that there were outliers which should be considered in the metric we want.

As the maximum order is 704,000, this suggests that this was a bulk order. Looking more closely, we see that it refers to a bulk order of 2000 pairs of shoes costing 352 dollars each. By having a high standard deviation, using the mean would not be an accurate nor helpful metric to report as the chosen measure of central tendancy.

```
boxplot(data$order_amount)
```



By visualizing the order amount column with a boxplot, we can see that there the outliers lie around the 700,000 order amount.

b. What metric would you report for this dataset? I would use the median as it is less susceptible to outliers which can skew our representation of the data. Additionally, it would be helpful to report the standard deviation for a better sense of accuracy, which can help analyze the variability in an "average" order. In this calculation, I would exclude extreme outliers which are calculated using: Upper extreme outliers = 3*(Q3 - Q1) + Q3 = 3*(390-163) + 390 = 1071 and lower extreme outliers = = 3*(Q3 - Q1) + Q3 = 163 - 3*(390-163) = -518.

c. What is its value? As calculated using the summary function, the median is 284.