Shopify Data Science Challenge

Hanhee Yang

5/11/2022

Question 1

Given some sample data, write a program to answer the following

On Shopify, we have exactly 100 sneaker shops, and each of these shops sells only one model of shoe. We want to do some analysis of the average order value (AOV). When we look at orders data over a 30 day window, we naively calculate an AOV of $3145.13. Given that we know these shops are selling sneakers, a relatively affordable item, something seems wrong with our analysis.

a) Think about what could be going wrong with our calculation.
b) Think about a better way to evaluate this data.
c) What metric would you report for this dataset? What is its value?

a) Think about what could be going wrong with our calculation

To start off, the required data was loaded.

```
library(magrittr)
library(dplyr)
library(ggplot2)
library(readxl)
dataset <- read_excel("dataset.xlsx")
```

I then loaded the summary of the dataset to observe how the overall data looked.

```
summary(dataset)
```

```
##     order_id       shop_id          user_id       order_amount
##  Min.   :   1   Min.   :  1.00   Min.   :607.0   Min.   :    90
##  1st Qu.:1251   1st Qu.: 24.00   1st Qu.:775.0   1st Qu.:   163
##  Median :2500   Median : 50.00   Median :849.0   Median :   284
##  Mean   :2500   Mean   : 50.08   Mean   :849.1   Mean   :  3145
##  3rd Qu.:3750   3rd Qu.: 75.00   3rd Qu.:925.0   3rd Qu.:   390
##  Max.   :5000   Max.   :100.00   Max.   :999.0   Max.   :704000
##   total_items       payment_method       created_at
##  Min.   :   1.000   Length:5000        Min.   :2017-03-01 00:08:09
##  1st Qu.:   1.000   Class :character   1st Qu.:2017-03-08 07:08:04
##  Median :   2.000   Mode  :character   Median :2017-03-16 00:21:20
##  Mean   :   8.787                      Mean   :2017-03-15 22:20:37
##  3rd Qu.:   3.000                      3rd Qu.:2017-03-23 10:39:57
##  Max.   :2000.000                      Max.   :2017-03-30 23:55:35
```

$3145 was the estimated mean of the order amount, which refers to the $3145.13 that was described as the average order value (AOV) in the dataset.

Initially, I assumed that the reason the average order value was high was because of order value outliers in the data that could result for the increase in the average order value of the data. These outliers could be a resultant of bulk purchases by a customer, an error in the system, or an unknown reason that needs to be analyzed.

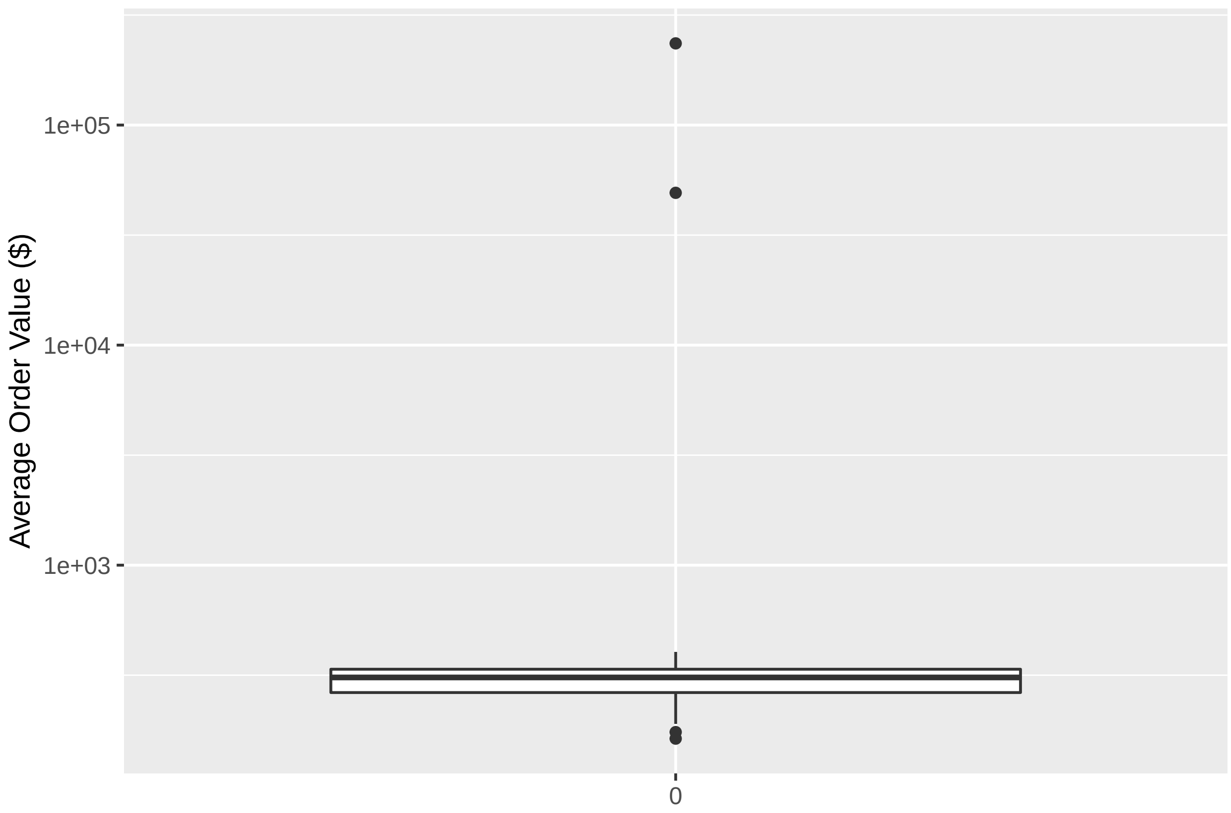
I first decided to figure out what the outliers of the dataset could be or if outliers could exist.

```
top_aov <- dataset %>%
  group_by(shop_id) %>%
  summarize(aov = mean(order_amount)) %>%
  arrange(desc(aov))

head(top_aov)
```

```
# A tibble: 6 x 2
  shop_id     aov
    <dbl>   <dbl>
1      42 235101.
2      78  49213.
3      50    404.
4      90    403.
5      38    391.
6      81    384
```

It seems as though Shop ID of 42 & 78 seem to be major outliers in the dataset. This can be proven using a boxplot.

```
boxplot <- qplot(factor(0),aov,data=top_aov,geom='boxplot')+
  scale_y_continuous(trans='log10')

print(boxplot + labs(
  y = "Average Order Value ($)", x = " "
))
```



From the plot above, shop 42 and 78 are the two points on the top that are way outside of the whiskers of the box plot, meaning they are definite outliers in the model.

Why is this happening?
Because the average order value from these shops were significantly higher than the average, I was curious and wanted to investigate more into why this was the case.

Shop 42 First of I will be analyzing shop 42.

```
# Shop 42
shop_42 <- dataset[dataset$shop_id==42,]

# Factoring out dataset to shop 42 and user 607
shop42_607 <- dataset[dataset$shop_id==42 & dataset$user_id==607,]

# Dataset of shop 42 without user 607
not_607 <-dataset[dataset$shop_id==42 & dataset$user_id!=607,]

# Printing average order value with and without user 607. Printing total order value from user 607.
avg_not_607 <- mean(not_607$order_amount)
cat("The average order value without user 607 is $", avg_not_607)
```

```
The average order value without user 607 is $ 652.2353
```

```
avg_42 <- mean(shop_42$order_amount)
cat("\nThe average order value without user 607 is $", avg_42)
```

```
The average order value without user 607 is $ 235101.5
```

```
sum_607 <- sum(shop42_607$order_amount)
cat("\nThe total order value from user 607 is $", sum_607)
```

```
The total order value from user 607 is $ 11968000
```

The dataset for shop 42 seemed almost normal. To my surprise, I found that user 607 specifically had a very peculiar behavior. Without user 607, the average order value is $652, which seems very normal for an average shoe spending size. However, user 607 specifically order sizes of $704,000 of 2000 items at exactly 4am at the time of every purchase. This shot the average up dramatically to $235101 for shop 42.This was done 17 times throughout the month for a total spending of $11,968,000.

This kind of spending from User 607 from a single Shopify store did not seem very realistic to me. Take for example, ColourPop, one of the most successful Shopify stores, creates an annual revenue of $15.25 million. The fact that this store would receive 80% of ColourPop's annual revenue in one month just does not seem correct.

I believe that there is some kind of suspicious activity or error in the system that could be causing this, where I would need more information about the store and user to investigate. Another possible scenario is the user of an automated system that user 607 set up to constantly buy $704,000 worth of material every morning at 4am. These mass purchases could be done by a large corporation to stock up on material from a Shopify store to keep up with cycle inventory.

```
shop_78 <- dataset[dataset$shop_id==78,]
shop_78
```

Shop 78

```
# A tibble: 46 x 7
   order_id shop_id user_id order_amount total_items payment_method
      <dbl>   <dbl>   <dbl>        <dbl>       <dbl> <chr>
 1      161      78     990        25725           1 credit_card
 2      491      78     936        51450           2 debit
 3      494      78     983        51450           2 cash
 4      512      78     967        51450           2 cash
 5      618      78     760        51450           2 cash
 6      692      78     878       154350           6 debit
 7     1057      78     800        25725           1 debit
 8     1194      78     944        25725           1 debit
 9     1205      78     970        25725           1 credit_card
10     1260      78     775        77175           3 credit_card
# ... with 36 more rows, and 1 more variable: created_at <dttm>
```

Shop 78 seems more reasonable. The sneaker(s) is very pricey ($25,725). Though not common, there are sneakers such as the Nike Mag 2016 (Auto-Lacing) [priced at $26,000 currently] that are purchased at this price range. There are different user_ids that purchase the sneakers at very humanly reasonable times. This shop may have been an outlier solely because of the large price value that the store asks its customers compared to the other shops. More data about the store would be needed to determine if this assumption is true.

b) Think about a better way to evaluate this data.

Why AOV is not a good metric for store

I assume that the reason the average of AOV of all the Shopify sneaker stores is to determine the performance of sneakers stores in general.

If the reason for evaluating this data is to determine the performance of Shopify sneaker stores, simply stating the average order value is not reasonable. As seen in the data, there are shops that receive large order values, but a small number of orders, while there are shops that receive smaller order values, but larger number of orders. This current metric would favorably view the store with larger order values compared to the smaller order values because the AOV would be bigger if the order values are larger.

Also the method of simply just naively applying AOV is very susceptible to being impacted by outliers.

Creating a categorized metric of store performance

One way retail industries broadly measure store performance is by using Gross Margin Return on Investment (GMROI), which is calculated by dividing the gross margin by the average inventory cost. By calculating the GMROI for each shop ID and creating an average GMROI for each store would be a better way to evaluate this data. This way, stores with a lower gross revenue could potentially be categorized in a similar GMROI with a store with a higher gross revenue.

If the scale of a store is also important in categorizing store performance, the stores could be put into brackets of gross revenue (ex: Low, Medium, High) before the GMROI would be calculated, so that the scale of the business could be considered. The only problem with this method is that the average inventory cost or any costs associated with the store is not accessible in this data.

Though there are other metrics to calculate store performance, since the problem initially started off with calculating AOV, I will use a modified AOV to evaluate this data. A big problem in the dataset was the outliers. Outliers are generally calculated as being outside of the 1.5 interquartile range (IQR) below the first quartile (Q1) or above the third quartile (Q3). Thus, a better way to calculate the AOV would be first to remove the outliers for order amount and then calculate the AOV.

I would compare this modified AOV to the median AOV of the original data as also a verification to see if the modified AOV is really an okay representation of the average store AOV.

c) What metric would you report for this dataset? What is its value?

I will use a modified AOV analysis system by removing outliers initially to calculate the AOV.

```r
# Calculating the interquartile range, high, and low
Q1 <- unname(quantile(dataset$order_amount, probs=c(.25)))
Q3 <- unname(quantile(dataset$order_amount, probs=c(.75)))

IQR <- Q3-Q1
High <- max(unname(Q3 + 1.5*IQR),0)
Low <- max(unname(Q1 - 1.5*IQR),0)

# Getting rid of outliers
count = 1
outliers <- c()
for (i in 1:nrow(dataset)){
  if (dataset$order_amount[i]>High){
    outliers <- append(outliers,i)
  }
}
```

```
# modified dataset
dataset2 <- dataset[-outliers,]

# modified AOV
aov2 <- mean(dataset2$order_amount)

# Calculating the median
med <- quantile(dataset$order_amount, probs=c(.5))

cat("The modified AOV is",aov2)
```

```
The modified AOV is 293.7154
```

```
cat("\nThe median order value is",med)
```

```
The median order value is 284
```

The modified AOV and median seem to hover around $290 for the ordered value. This is the value that Shopify shops should use to target marketing strategies around to maximize profit.

Question 2

a) How many orders were shipped by Speedy Express in total?

```
SELECT ShipperName,COUNT(OrderID) as TotalOrders
FROM Orders
INNER JOIN Shippers
ON Shippers.ShipperID = Orders.ShipperID
WHERE ShipperName = ‘Speedy Express’;
```

There were 54 orders shipped by Speedy Express in total.

b) What is the last name of the employee with the most orders?

```
SELECT LastName,Count(OrderID) as MostLastName
FROM Orders
INNER JOIN Employees
ON Orders.EmployeeID=Employees.EmployeeID
GROUP BY LastName
ORDER BY MostLastName DESC
LIMIT 1;
```

The last name of the employee with the most orders was Peacock with 40 orders.

c) What product was ordered the most by customers in Germany?

```
SELECT ProductName, SUM(Quantity) AS MostProdGermany
FROM Orders, Products, Customers, OrderDetails
ON Customers.CustomerID = Orders.CustomerID
AND Customers.Country = ‘Germany’
AND Orders.OrderID = OrderDetails.OrderID
AND OrderDetails.ProductID = Products.ProductID
GROUP BY ProductName
ORDER BY MostProdGermany DESC
LIMIT 1;
```

The Product that was ordered most by customers in Germany was Boston Crab Meat (with a total order size of 160).