Shopify Data science Challenge

AOV

Taiwo Owoseni

May 19th 2022

Question

Given some sample data, write a program to answer the following:click here to access the required data set

In 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 $3,145.13. Given that we know these shops are selling sneakers, a relatively affordable item, something seems wrong with our analysis.

1. Think about what could be going wrong with our calculation. Think about a better way to evaluate this data.
2. What metric would you report for this data set?
3. What is its value?

Answer to Question :

1. Presence of uutliers in the data set
2. AOV without outliers
3. AOV : $299 dollars

Approach

Here are the list of steps I intend to use to evaluate if the AOV is a good measurement or not:

- **Investigating the data set** : Visually understanding the data and look for (filter) outliers.
- **Observe how AOV vary with and without outliers**

Load Libraries and Data

```
head(data)
```

```
## # A tibble: 6 x 7
##   order_id shop_id user_id order_amount total_items payment_method created_at
##      <dbl>   <dbl>   <dbl>        <dbl>       <dbl> <chr>          <chr>
## 1        1      53     746          224           2 cash           2017-03-13 1~
## 2        2      92     925           90           1 cash           2017-03-03 1~
## 3        3      44     861          144           1 cash           2017-03-14 4~
## 4        4      18     935          156           1 credit_card    2017-03-26 1~
## 5        5      18     883          156           1 credit_card    2017-03-01 4~
## 6        6      58     882          138           1 credit_card    2017-03-14 1~
```

Data Cleaning

Change the data type of some of the columns for easier manipulation

```
data <- data |>
  mutate(order_id = as.factor(order_id))|>
  mutate(shop_id = as.factor(shop_id))|>
  mutate(user_id = as.factor(user_id))|>
  mutate(payment_method = as.factor(payment_method))|>
  mutate(created_at = as.Date(created_at))
data
```

```
## # A tibble: 5,000 x 7
##    order_id shop_id user_id order_amount total_items payment_method created_at
##    <fct>    <fct>   <fct>          <dbl>       <dbl> <fct>          <date>
##  1 1        53      746              224           2 cash           2017-03-13
##  2 2        92      925               90           1 cash           2017-03-03
##  3 3        44      861              144           1 cash           2017-03-14
##  4 4        18      935              156           1 credit_card    2017-03-26
##  5 5        18      883              156           1 credit_card    2017-03-01
##  6 6        58      882              138           1 credit_card    2017-03-14
##  7 7        87      915              149           1 cash           2017-03-01
##  8 8        22      761              292           2 cash           2017-03-08
##  9 9        64      914              266           2 debit          2017-03-17
## 10 10       52      788              146           1 credit_card    2017-03-30
## # ... with 4,990 more rows
```

Investigating the data set

- $$\text{Calculate unit cost} = \frac{\text{Order Amount}}{\text{Total Items}}$$
- $$\text{AOV} = \frac{\text{Summed Order Amount}}{\text{Number of Orders}}$$

```
AOV <- sum(data$order_amount)/ nrow(data)
AOV
```

```
## [1] 3145.128
```

Expectation: I assume that the distribution of the AOV is not skewed . My assumption is from the question **"the stores sell the same kind of sneakers"** .

Plotting AOV per Shop

I am going to create the AOV per shop. The table below shows the aov in each shop.

The AOV at **shops 42 and 78** is very high approximately **235, 101 dollars and 49,213 dollars respectively.**

```
shop_aov <- data |>
  group_by(shop_id = as.factor(shop_id))|>
  summarise(aov =  sum(order_amount)/ n())|>
  arrange(desc(aov))
shop_aov
```

```
## # A tibble: 100 x 2
##    shop_id     aov
##    <fct>     <dbl>
##  1 42      235101.
##  2 78       49213.
```

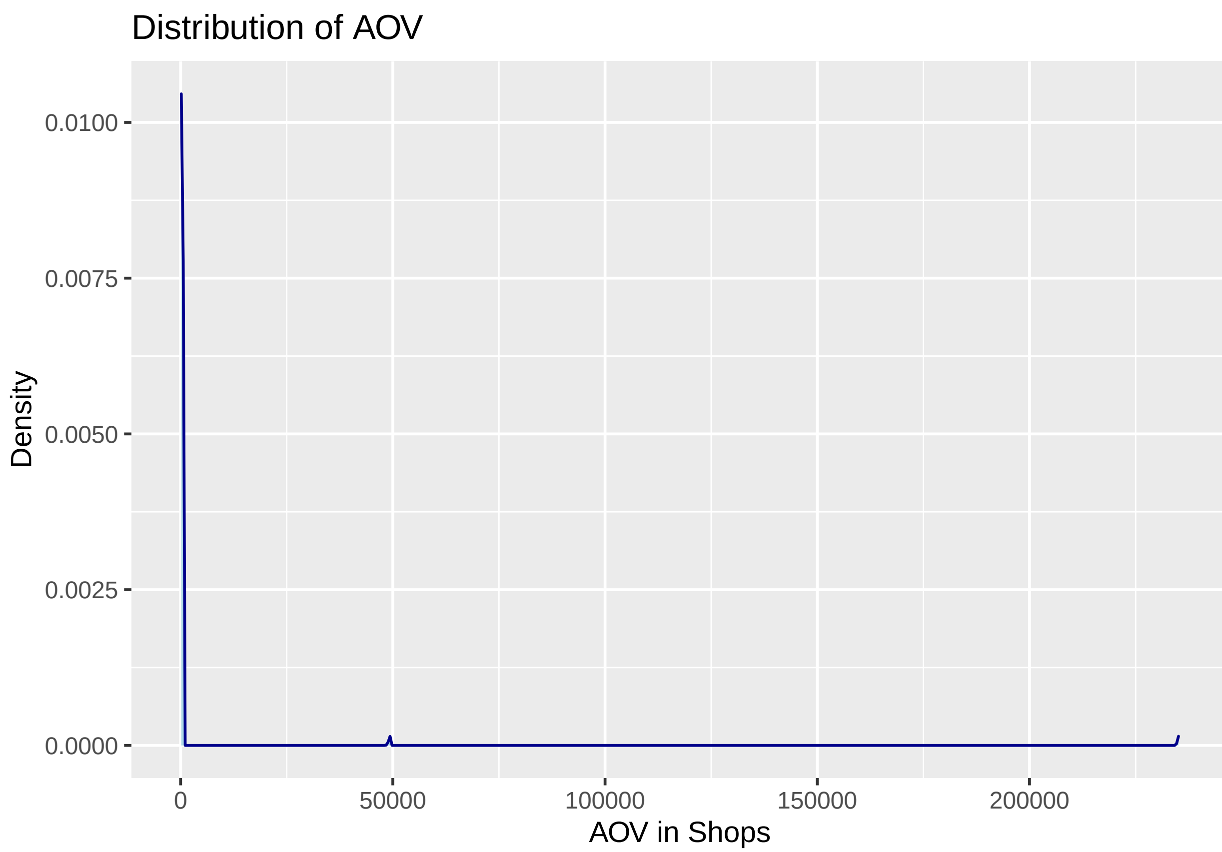
```
##   3 50          404.
##   4 90          403.
##   5 38          391.
##   6 81          384
##   7 6           384.
##   8 89          379.
##   9 33          376.
##  10 51          362.
## # ... with 90 more rows
```

Distribution of AOV

The plot below shows that most of the shops have a low AOV and few have high AOV. These are the little peaks at around 50,000 and above 200,000

Result: The plot below proves that the shops are not selling the sneakers at a normally distributed price. The plot is right skewed. This indicates presence of outliers in the data.

```
shop_aov |>
   ggplot(aes(x= aov)) +
  geom_density(color="darkblue", fill="lightblue") +
  ggtitle('Distribution of AOV') +
  ylab('Density') +
 xlab('AOV in Shops')
```



Next, I will remove the outliers in the data

```
outliers <- boxplot(shop_aov$aov, plot = FALSE)$out
paste0("NUmbers of Outliers: ",  length(outliers))
```

Removing Outliers

```
## [1] "NUmbers of Outliers: 2"
```

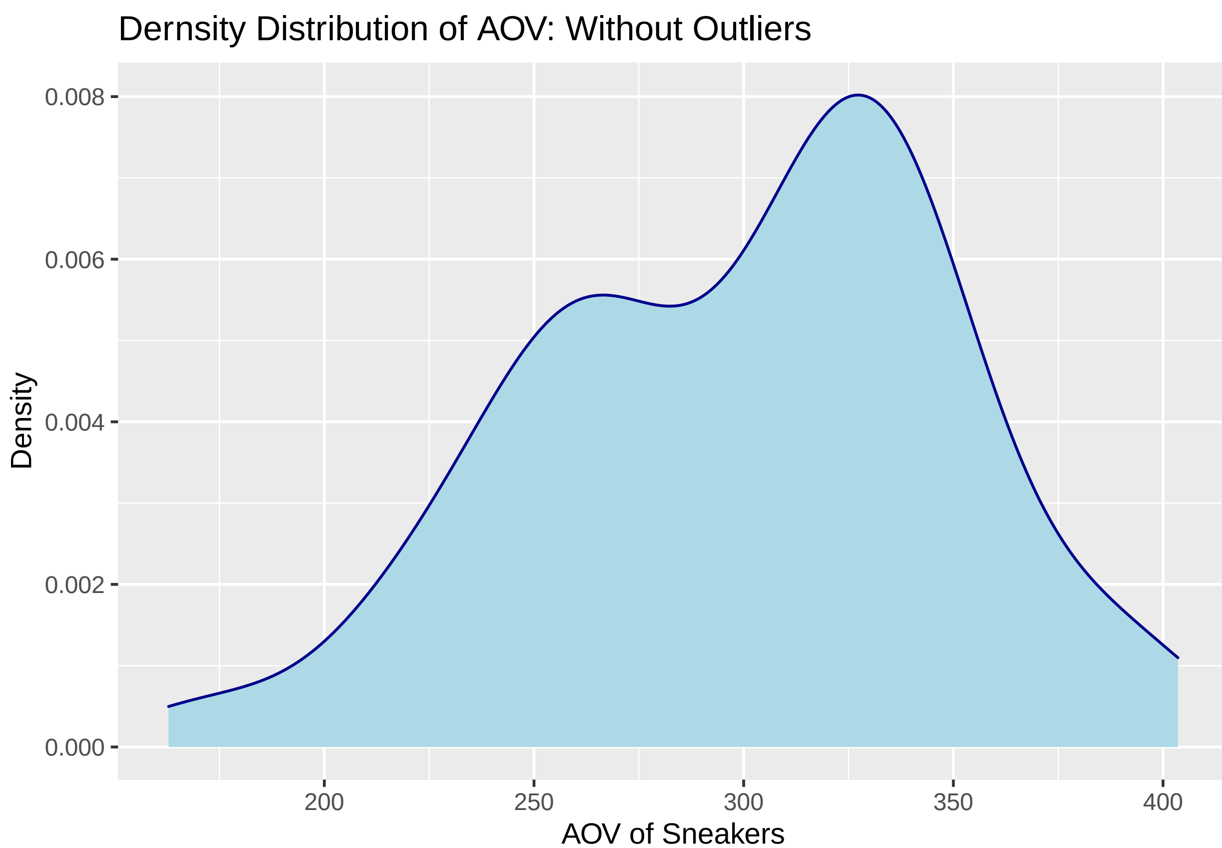
```
shop_aov_clean <- shop_aov[!(shop_aov$aov %in% outliers), ]
#shop_aov_clean
```

I will reproduce the two plot above to see if removing outliers has an effect on **unit cost**, hence have an effect on **AOV.**

Observation

There are two peaks in the distributions. This could indicate two groups in the shops. Overall, this is a better distribution than the first

```
shop_aov_clean |>
   ggplot(aes(x= aov)) +
  geom_density(color="darkblue", fill="lightblue") +
  ggtitle('Dernsity Distribution of AOV: Without Outliers ') +
  ylab('Density') +
 xlab('AOV of Sneakers')
```



Calculate average AOV after Outlier removal The AOV drastically reduced from **$3,145 to $299.6 dollars.** AOV is easily affected by outliers. In this dataset, if outliers are removed, AOV would be a good metric

```
AOV_no_outlier <- mean(shop_aov_clean$aov)
AOV_no_outlier
```

```
## [1] 299.6824
```

Answer to the Questions;

1.

Comments

- AOV is a good metric but sensitive to outliers. It's however possible that the outliers in the data are not as a result of data entry error. Some stores might be owned by a celebrity and people would love to buy from them. Hence the high order amount.
- What should matter to the store is if they are able to retain customers. Given the time period of the data set, it would be difficult to calculate customer retention form scratch.