# Q1 Shopify DS Intern Markdown

Zetong Jin

1/13/2022

```
library(readxl)
data = read_excel("/Users/thomas/Downloads/2019 Winter Data Science Intern Challenge Data Set.xlsx")
attach(data)
mean(order_amount) # wrong answer
```

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

```
summary(order_amount)
```

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

The AOV of $3145.13 comes out directly by calculate the mean value of the order amount which is not correct. By calculating the mean value of the order amount, we did not consider the total items that a customer bought. Thus, this number would be wrong. A better way to evaluate this data is to group by the shop and use the total revenue/total order to get the correct AOV.

## AOV = total Revenue/Total order

## AOV by each shop

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
aov_1 <- data %>%
  group_by(shop_id) %>%
  summarise(total_revenue = sum(order_amount),
            total_orders = sum(total_items)) %>%
  transmute(aov_1 = total_revenue/total_orders) %>%
  ungroup() %>%
  summarise(average_ov = mean(aov_1)) #the mean value for each AOV
aov_1
```

```
## # A tibble: 1 x 1
##   average_ov
##        <dbl>
## 1       408.
```

#sneaker_price = order_amount/total_items #Price for each sneaker

```
aov_2 <- data %>%
  group_by(shop_id) %>%
  mutate(sneaker_price = order_amount/total_items) %>%
  summarise(sneaker_price = mean(sneaker_price)) %>%
  select(shop_id, sneaker_price) %>%
  arrange(desc(sneaker_price))
aov_2
```

```
## # A tibble: 100 x 2
##    shop_id sneaker_price
##      <dbl>         <dbl>
##  1      78         25725
##  2      42           352
##  3      12           201
##  4      89           196
##  5      99           195
##  6      50           193
##  7      38           190
##  8       6           187
##  9      51           187
## 10      11           184
## # ... with 90 more rows
```

shop no.78 has the most expensive sneaker price (25725) which is not normal

#Exclude shop No.78 (Delete outlier)

```
aov_3 <-  data %>%
  group_by(shop_id) %>%
  filter(shop_id != 78) %>%
  summarise(total_revenue = sum(order_amount),
            total_orders = sum(total_items)) %>%
  transmute(aov_3 = total_revenue/total_orders) %>%
  ungroup() %>%
  summarise(aov_3 = mean(aov_3))
aov_3
```

```
## # A tibble: 1 x 1
##   aov_3
##   <dbl>
## 1  152.
```

We get the final AOV for each shop - 152 With the outlier, the AOV would be 408 and now it appears to decrease a lot.