# Shopify Data science Challenge

AOV

Taiwo Owoseni

May 9th 2022

### Question

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

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.

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?

### 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 in 30-days**
- **Try other metric:** Customer retention?

### 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~
```

### 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
```

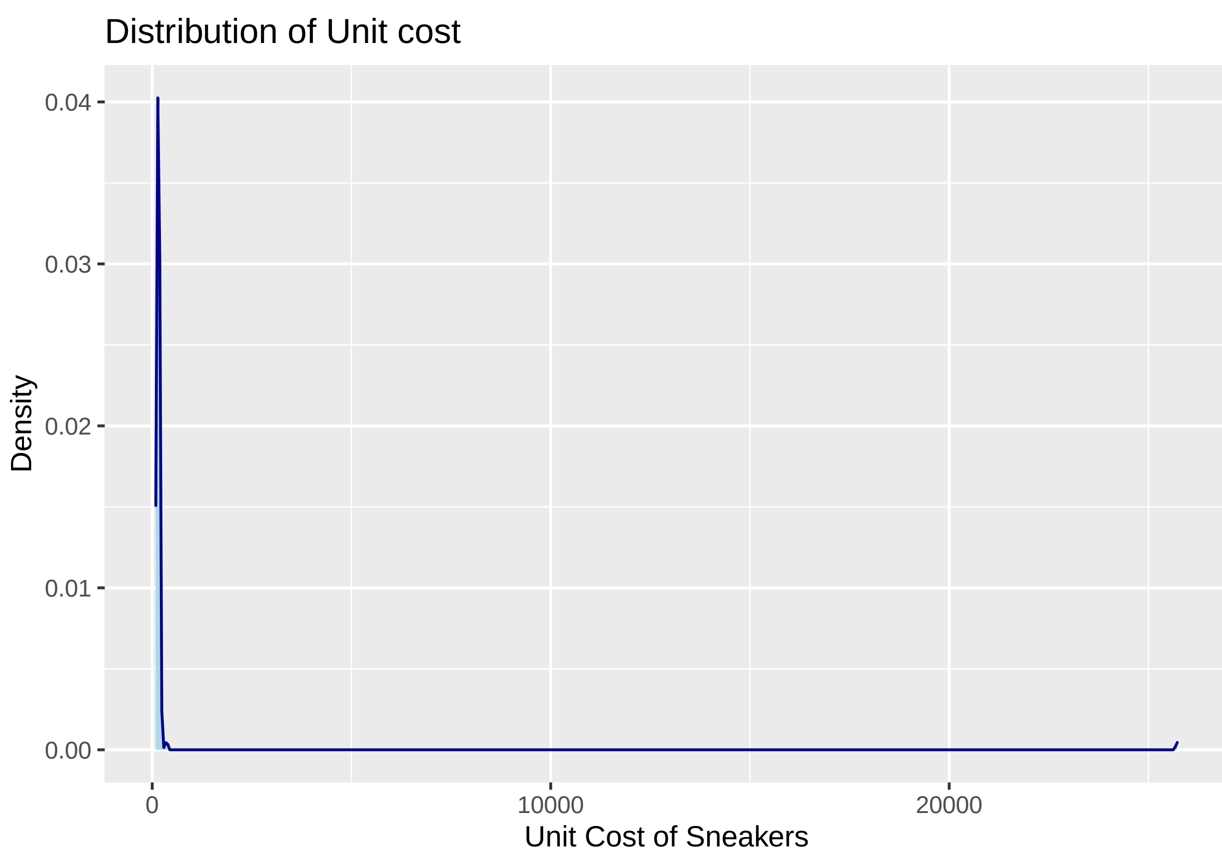
```
## [1] 3145.128
```

```
data$unit_cost <- data$order_amount / data$total_items
```

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

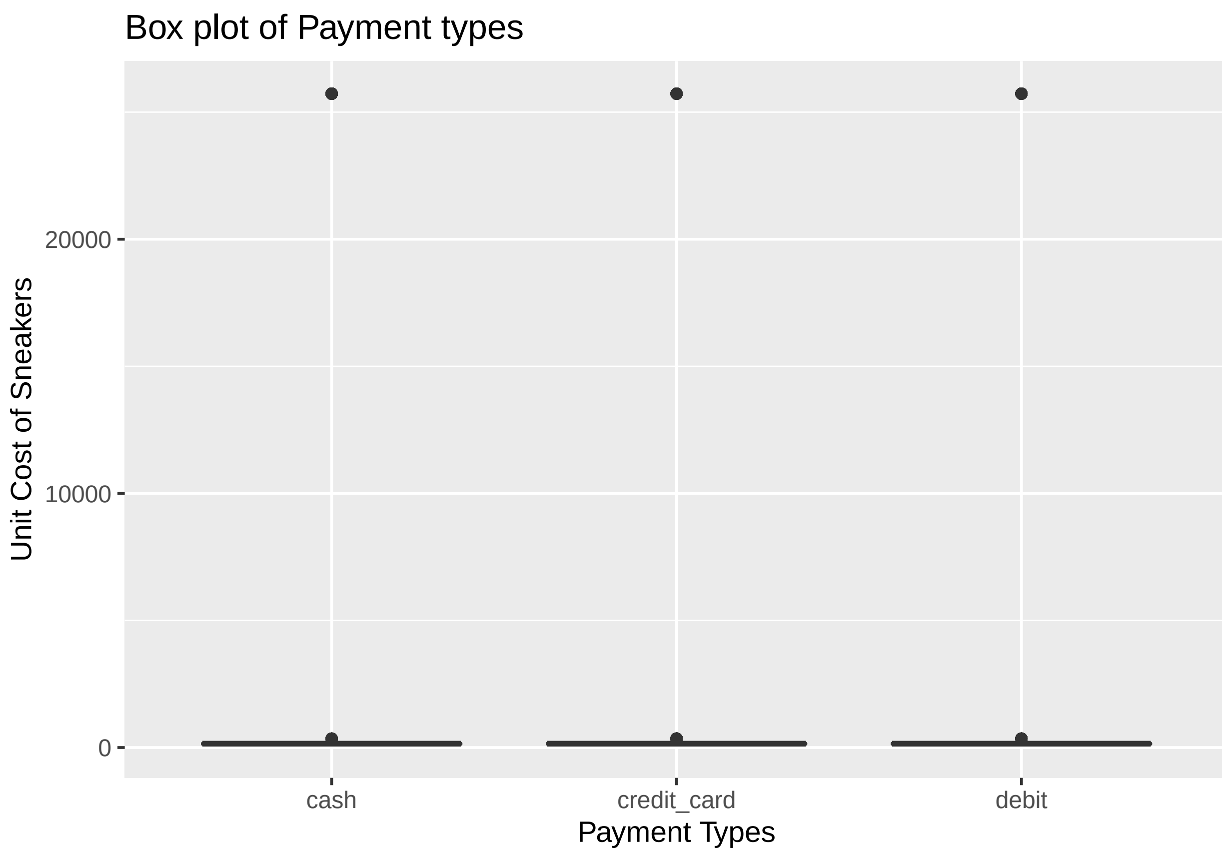
**Result:** The plot below proves that the shops are not selling the sneakers at the an normally distributed price. The plot is right skewed. This indicates presence of outliers in the data. I will plot a box plot to see these outliers in the different payment method.

```
data |>
   ggplot(aes(x= unit_cost)) +
  geom_density(color="darkblue", fill="lightblue") +
  ggtitle('Distribution of Unit cost ') +
  ylab('Density') +
 xlab('Unit Cost of Sneakers')
```




**Box Plot : Payment Method** Visually, there are outliers in the data. The box of the boxplot are closer to $0 unit cost.

```r
data |>
 ggplot(aes(x= payment_method , y= unit_cost)) +
  geom_boxplot() +
  ggtitle('Box plot of Payment types') +
  xlab('Payment Types') +
  ylab('Unit Cost of Sneakers')
```




**Most expensive Shop**

The cost of a sneaker at **shop 78** is : **25,725 dollars.**

```r
data |> filter(unit_cost == max(unit_cost))
```

```
## # A tibble: 46 x 8
##    order_id shop_id user_id order_amount total_items payment_method created_at
##       <dbl>   <dbl>   <dbl>        <dbl>       <dbl> <chr>          <chr>
##  1      161      78     990        25725           1 credit_card    2017-03-12 ~
##  2      491      78     936        51450           2 debit          2017-03-26 ~
##  3      494      78     983        51450           2 cash           2017-03-16 ~
##  4      512      78     967        51450           2 cash           2017-03-09 ~
##  5      618      78     760        51450           2 cash           2017-03-18 ~
##  6      692      78     878       154350           6 debit          2017-03-27 ~
##  7     1057      78     800        25725           1 debit          2017-03-15 ~
##  8     1194      78     944        25725           1 debit          2017-03-16 ~
##  9     1205      78     970        25725           1 credit_card    2017-03-17 ~
## 10     1260      78     775        77175           3 credit_card    2017-03-27 ~
## # ... with 36 more rows, and 1 more variable: unit_cost <dbl>
```

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

**Removing Outliers**

```
## [1] "NUmbers of Outliers: 97"
```

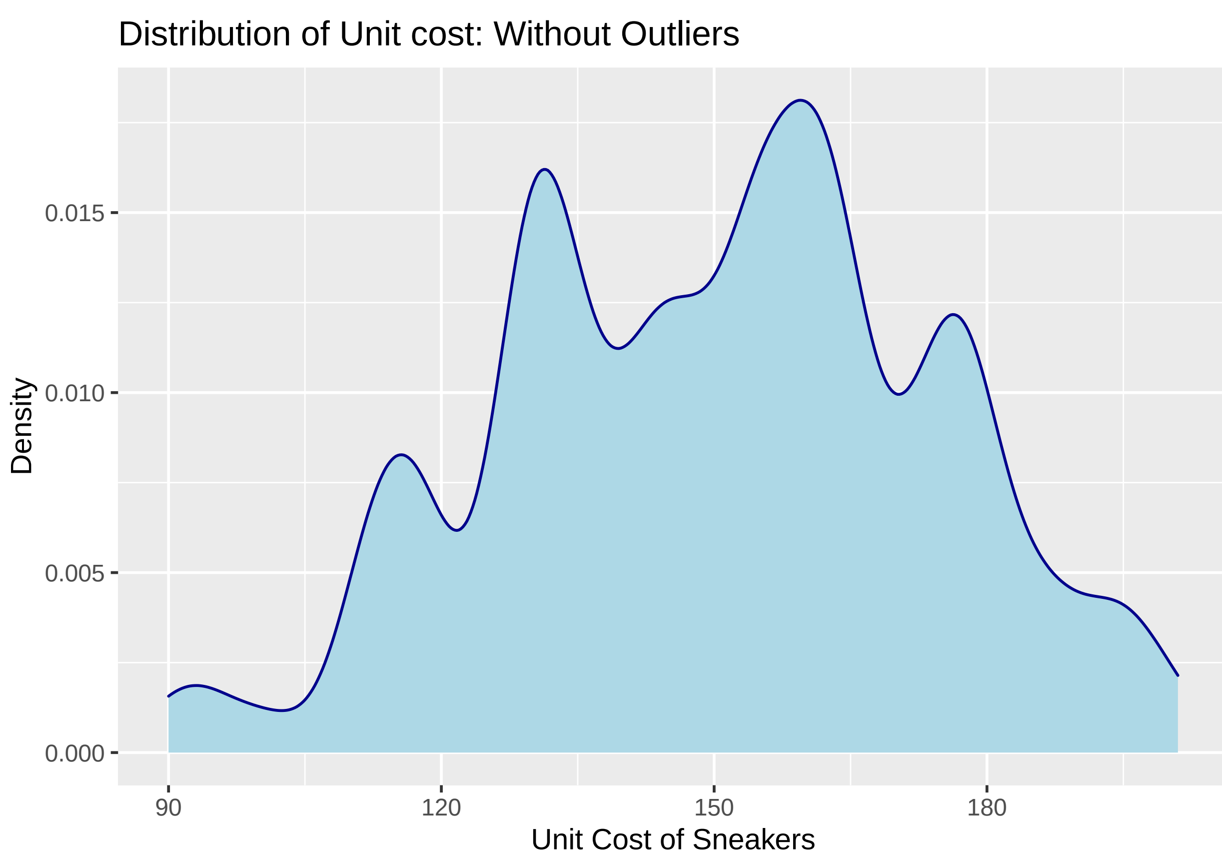
```
data_out <- data[!(data$unit_cost %in% outliers), ]
```

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**

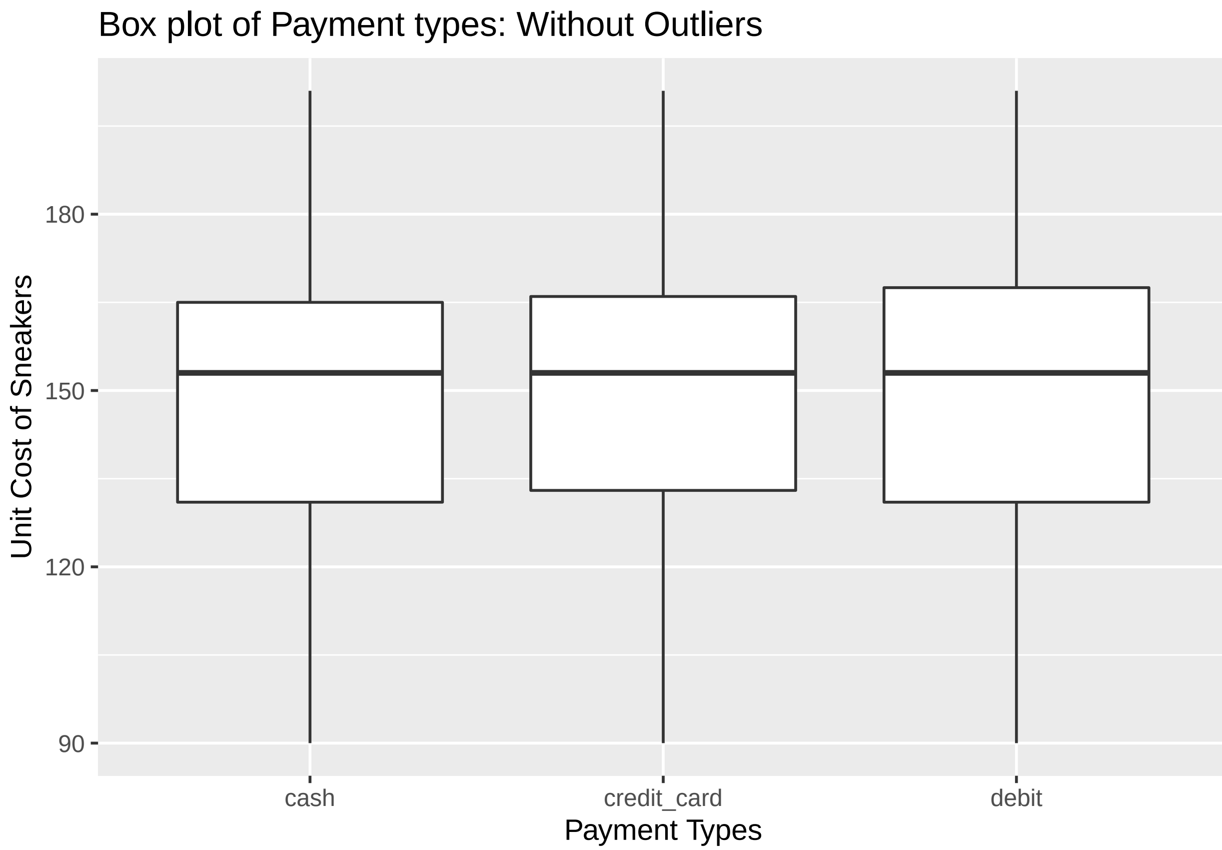
There are peaks in the distributions. This could indicate different groups in the shops. The boxplot looks great.

```
data_out |>
   ggplot(aes(x= unit_cost)) +
  geom_density(color="darkblue", fill="lightblue") +
  ggtitle('Distribution of Unit cost: Without Outliers ') +
  ylab('Density') +
 xlab('Unit Cost of Sneakers')
```




```
data_out |>
 ggplot(aes(x= payment_method , y= unit_cost)) +
  geom_boxplot() +
  ggtitle('Box plot of Payment types: Without Outliers') +
```

```
  xlab('Payment Types') +
  ylab('Unit Cost of Sneakers')
```




**Calculate AOV after Outlier removal** The AOV drastically reduced from **$3,145 to $300 dollars.**

```
AOV_no_outlier <- sum(data_out$order_amount)/ nrow(data_out)
AOV_no_outlier
```

```
## [1] 300.1558
```

**How Does the AOV vary in 30-days?** In other to understand if AOV is a great metric; I will look at how it varied in the month in two cases:

- **Data with outlier : data**
- **Data without outlier :data_out**

**Data with outlier : data** The time series for the AOV shows a spike in sales for the Beginning and **end** of the month. For the rest of the month, the spike is somewhat consistent.

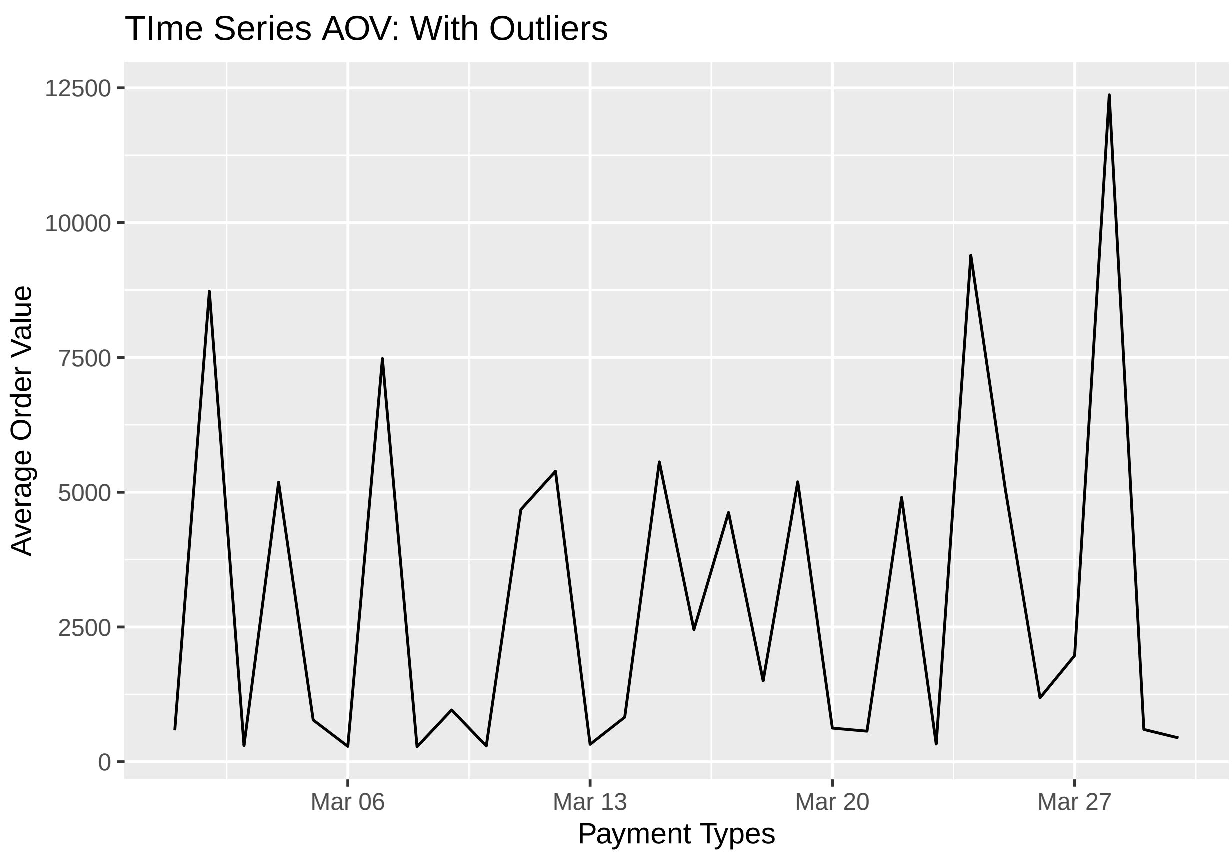
```
data$created_at <-as.Date(data$created_at)

#data$date <- as.Date(mdy_hms(data$created_at))
data$date <- as.Date(data$created_at, "%Y%m%d")

aov_data_mth <- data |>
  group_by(created_at)|>
```

```
  summarize( aov = sum(order_amount)/n())

 ggplot(aov_data_mth, aes(x= created_at, y= aov)) +
  geom_line() +
  xlab('Time Created') +
  ggtitle('TIme Series AOV: With Outliers') +
  xlab('Payment Types') +
  ylab('Average Order Value')
```

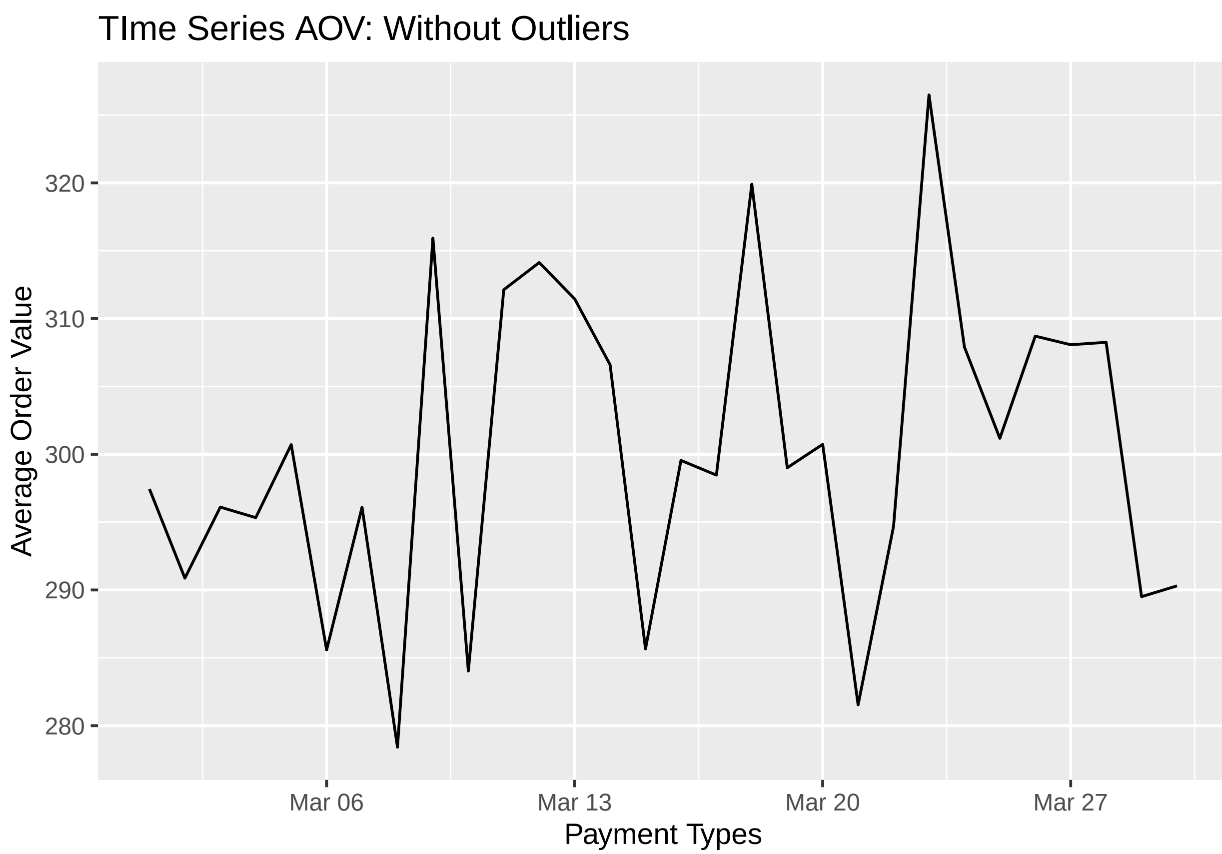



**Data with outlier : data_out** The time series for the AOV for the data without outlier shows a spike in AOV for the middle of the month. This is opposite of the first time series (the one with outliers). In the **end** of the month, their is a decline in AOV and in the beginning of the month, AOV is small.

```
data_out$created_at <-as.Date(data_out$created_at)
data_out$date <- as.Date(data_out$created_at, "%Y%m%d")

aov_data_mth_out <- data_out |>
  group_by(created_at)|>
  summarize( aov = sum(order_amount)/n())

 ggplot(aov_data_mth_out, aes(x= created_at, y = aov)) +
  geom_line() +
  xlab('Time Created') +
  ggtitle('TIme Series AOV: Without Outliers') +
  xlab('Payment Types') +
  ylab('Average Order Value')
```




**Comments**

- AOV only shows growth in a specific time period. This isn't a great metric. It' very 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. I will be using a library called cohort.

**General Retention Rate for all the Shops**

```
retention_data <- function(data) {
  cohort_days <- data %>%
  cohort_table_day(user_id, date)%>%
   shift_left_pct() %>%
   pivot_longer(-cohort) %>%
   mutate(time = as.numeric(str_remove(name,'t')),
          name = str_replace(name,'t', 'Day'))

   cohort_days[cohort_days == 0] <- NA
   cohort_days
}
```

**Retention with Outliers**

```r
data_cohort <- retention_data(data)
data_cohort_out <- retention_data(data_out)
data_cohort
```

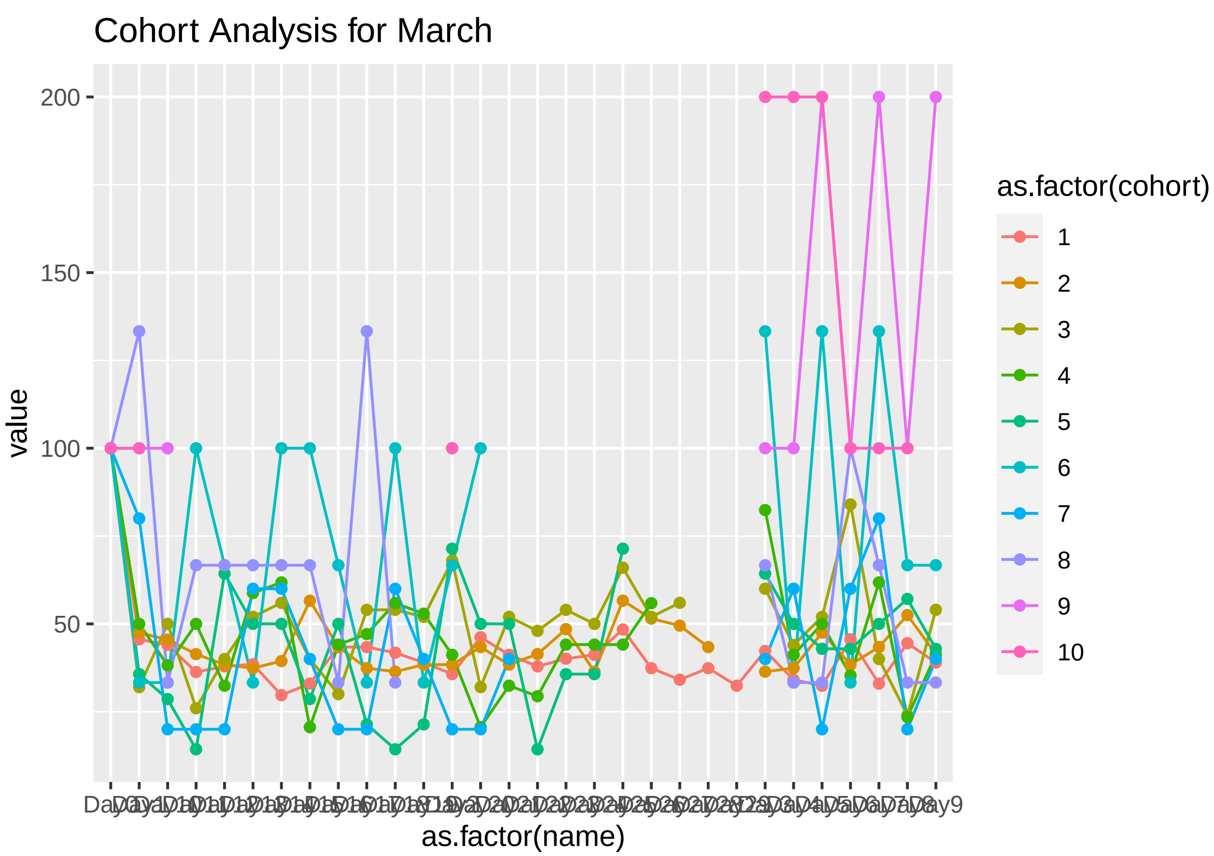
```
## # A tibble: 300 x 4
##    cohort name  value  time
##     <int> <chr> <dbl> <dbl>
##  1      1 Day0  100      NA
##  2      1 Day1   45.6     1
##  3      1 Day2   35.7     2
##  4      1 Day3   42.3     3
##  5      1 Day4   34.1     4
##  6      1 Day5   32.4     5
##  7      1 Day6   45.6     6
##  8      1 Day7   33       7
##  9      1 Day8   44.5     8
## 10      1 Day9   39       9
## # ... with 290 more rows
```

Interpretation of visual isn't clear for the Cohort analysis. I believe if the data spans more months, it will be easier to interpret the cohort analysis visual.

```r
ggplot(data_cohort, aes(x = as.factor(name), y = value, color = as.factor(cohort), group = cohort)) + ge
```

```
## Warning: Removed 1 row(s) containing missing values (geom_path).
```

```
## Warning: Removed 78 rows containing missing values (geom_point).
```



**Retention with Outliers**

- Conclusively, AOV is a good metric but it is easily affected by outliers.
- Customer retention is also a good metric to use when data has outliers.