Shopify challenge

Trishla Jain

18/01/2022

Loading the necessary packages

```
suppressMessages(library(tidyverse))
suppressMessages(library(dplyr))
suppressMessages(library(ggplot2))
suppressMessages(library(dlookr))
suppressMessages(library(corrplot))
```

```
## Warning: package ’corrplot’ was built under R version 4.1.2
```

```
suppressMessages(library(readxl))
```

```
csv_file <- read_excel("C:/Users/trish/Desktop/Internship and job applications/2019 Winter Data Science
View(csv_file)
```

Checking summary statistics for the dataset

```
summary(csv_file)
```

```
##     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
```

Looking at order_amount we can see that the mean is quite greater than the median suggesting that it is right skewed and that there could be outliers in our data; also the max value of 704000 is very far away from the 3rd quantile value of 390 and same is the case with total_items and we can also see that the maximum total item is 2000 which is very far from our 3rd quantile value clearly stating that this value is our outlier. Rest of the columns are just serial numbers so we wont be checking on them.

Also looking at the mean for order_amount we can see we get the same mean or AOV of 3145 as shown in the question.

Checking for NA and Null's values in our dataset.

```
sapply(csv_file,function(x) sum(is.na(x)))
```

```
##       order_id        shop_id        user_id   order_amount    total_items
##              0              0              0              0              0
## payment_method     created_at
##              0              0
```

```
sapply(csv_file,function(x) sum(is.null(x)))
```

```
##       order_id        shop_id        user_id   order_amount    total_items
##              0              0              0              0              0
## payment_method     created_at
##              0              0
```

There are no null and NA values in our data which is good.

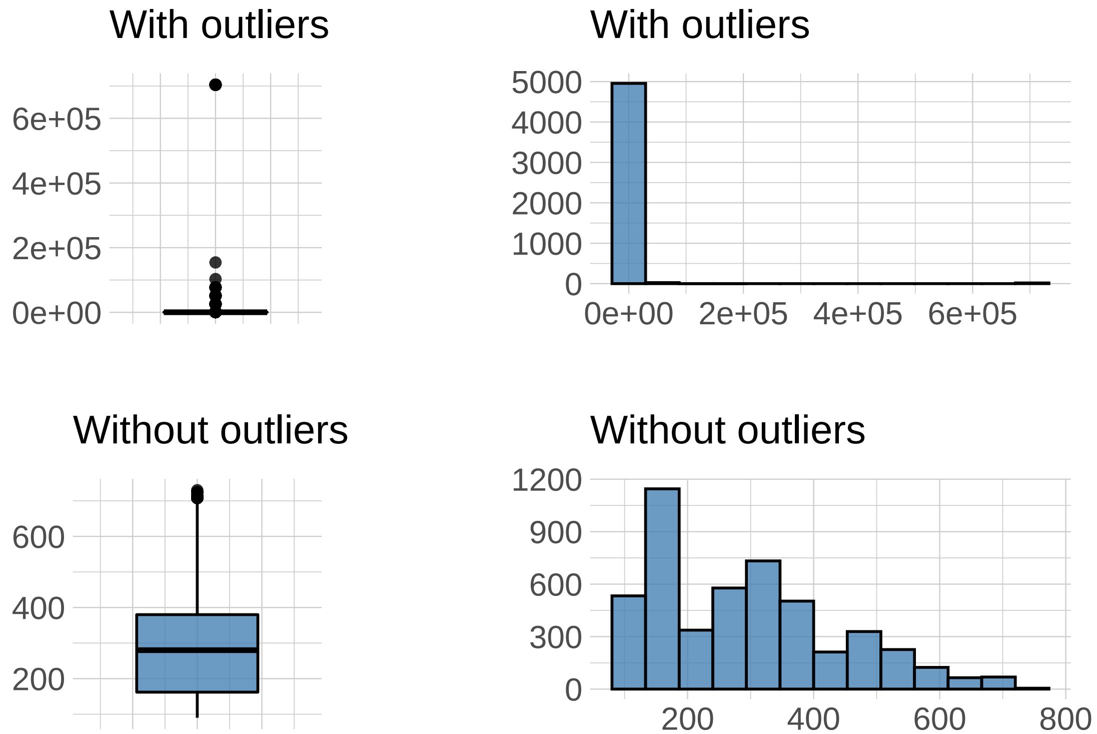
```
diagnose_outlier(csv_file)
```

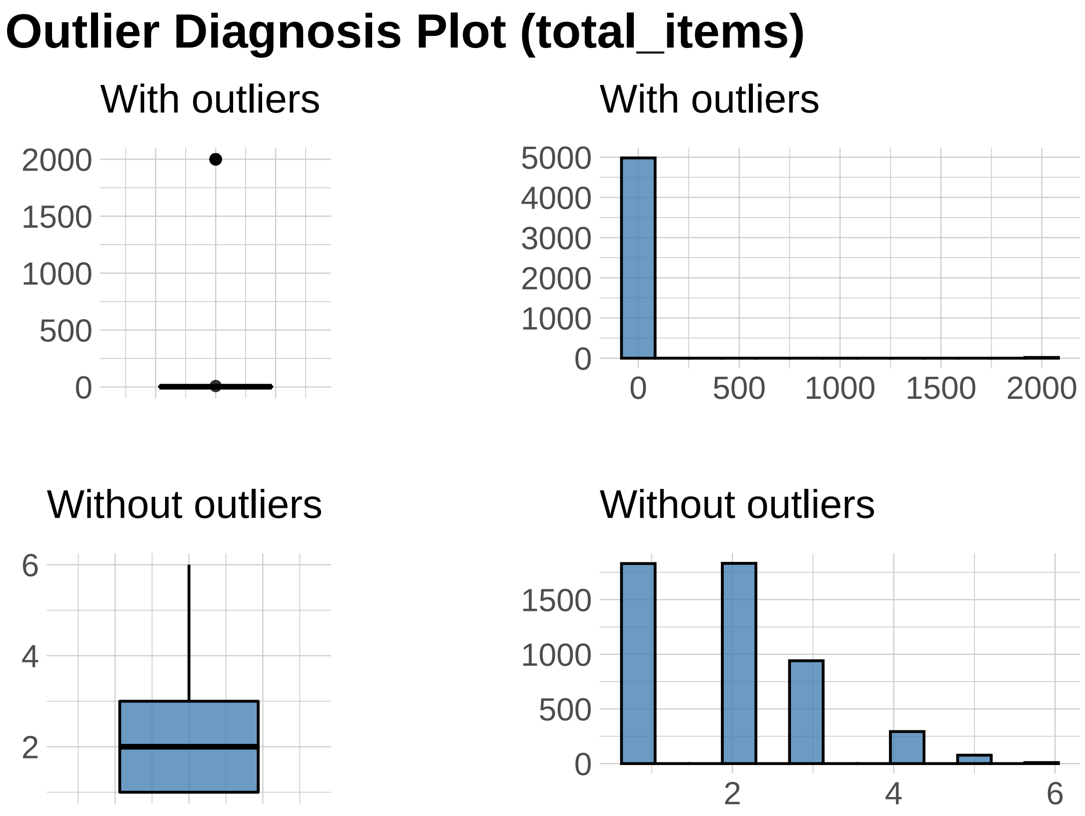
```
## # A tibble: 5 x 6
##   variables    outliers_cnt outliers_ratio outliers_mean with_mean without_mean
##   <chr>               <int>          <dbl>         <dbl>     <dbl>        <dbl>
## 1 order_id                0           0              NaN    2500.        2500.
## 2 shop_id                 0           0              NaN      50.1         50.1
## 3 user_id                 0           0              NaN     849.         849.
## 4 order_amount          141           2.82       101408.   3145.         294.
## 5 total_items            18           0.36         1889.      8.79         1.99
```

We can see that there are very few outliers in our dataset.

```
plot_outlier(csv_file %>%
      select(order_amount,total_items))
```

Outlier Diagnosis Plot (order_amount)

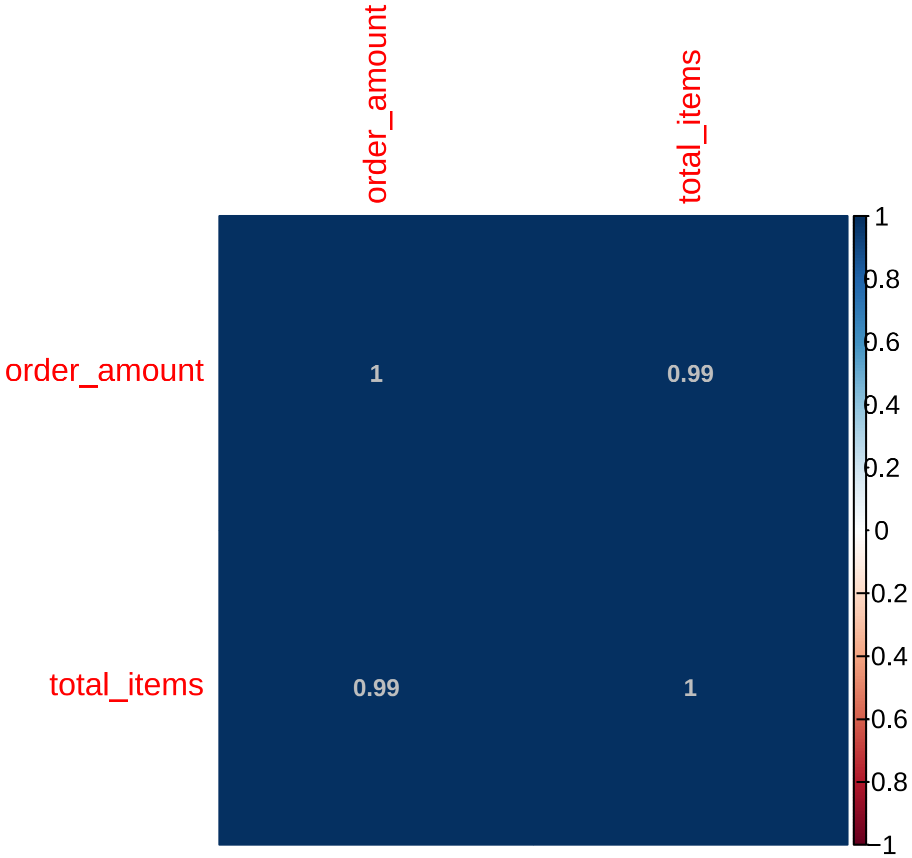




For order amount we can see we can reduce the right skewness after removing the outlier and looking at the boxplot we can see that it looks almost normal distribution. For Total_items we we can see we get a better box plot after removing the outlier although the graph has barely improved.

Checking correlation

```
corrplot(cor(csv_file %>% dplyr::select(order_amount,total_items)),
         method = "color",
         addCoef.col="grey",
         order = "AOE", number.cex=0.75)
```



We can see that the items are highly correlated which makes sense as when the total items increase the order amount would also increase.

Removing outliers:

```
# checking how many records we would be removing
count(subset(csv_file, csv_file$total_items >= 2000))
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1    17
```

```
# removing the outliers
csv_file_noOut <- csv_file[!(csv_file$total_items >= 2000),]
```

Validating our changes

```
diagnose_outlier(csv_file_noOut)
```

```
## # A tibble: 5 x 6
##   variables    outliers_cnt outliers_ratio outliers_mean with_mean without_mean
##   <chr>               <int>          <dbl>         <dbl>     <dbl>        <dbl>
## 1 order_id                0          0               NaN    2501.        2501.
## 2 shop_id                 0          0               NaN      50.1         50.1
## 3 user_id                 0         0               NaN      850.         850.
## 4 order_amount          124         2.49          18794.     754.         294.
## 5 total_items             1         0.0201            8        1.99         1.99
```

```
count(csv_file_noOut)
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1  4983
```

```
# checking how many records we would be removing
count(subset(csv_file_noOut, csv_file_noOut$order_amount >= 715))
```

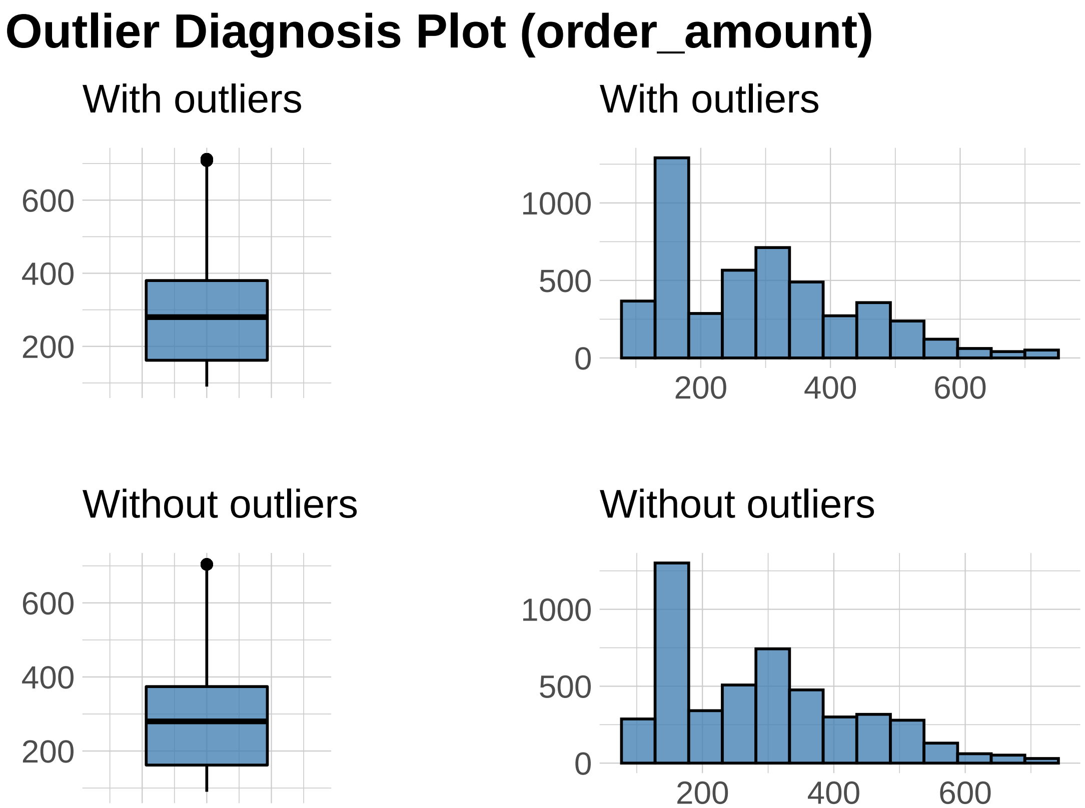
```
## # A tibble: 1 x 1
##       n
##   <int>
## 1   129
```

```
# removing the records
csv_file_noOut <- csv_file_noOut[!(csv_file_noOut$order_amount >= 715),]
# checking how many records are in the final dataset
count(csv_file_noOut)
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1  4854
```

Validating and checking if removing outliers helped.

```
plot_outlier(csv_file_noOut %>%
      select(order_amount,total_items))
```



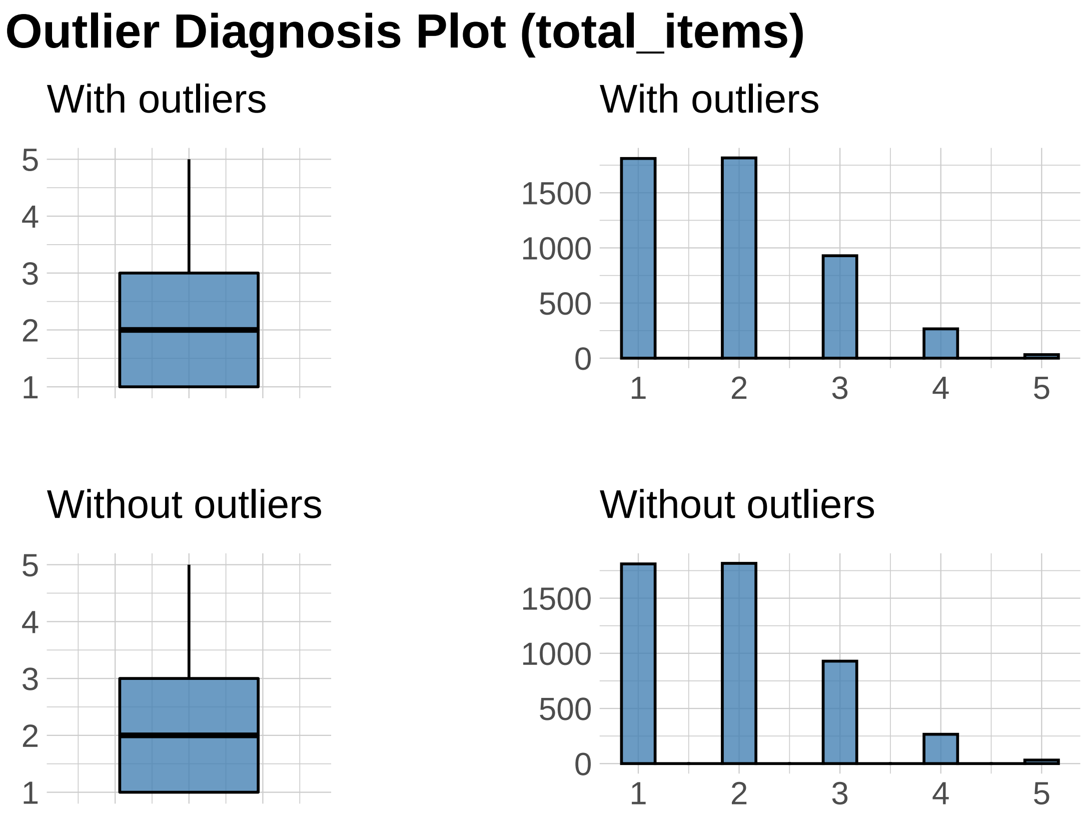
Outlier Diagnosis Plot (order_amount)

With outliers

With outliers

Without outliers

Without outliers



```
diagnose_outlier(csv_file_noOut %>%
      select(order_amount,total_items))
```

```
## # A tibble: 2 x 6
##   variables    outliers_cnt outliers_ratio outliers_mean with_mean without_mean
##   <chr>               <int>          <dbl>         <dbl>     <dbl>        <dbl>
## 1 order_amount           29          0.597          710.    293.         291.
## 2 total_items             0          0              NaN       1.95         1.95
```

We can see after a lot of trial and error(done manually and not included in this document to make it easier for the reader to dilute the information) that order_amount of greater than 715 are outliers and looking at the outlier plot we can see that after removing values of order_amount greater than equal to 700 we get the same plot for plot_outlier with and without outlier.

Hence we can go ahead and check what is the new mean or AOV value that we get.

```
summary(csv_file_noOut)
```

```
##     order_id       shop_id          user_id       order_amount
##  Min.   :   1   Min.   :  1.00   Min.   :700.0   Min.   : 90.0
##  1st Qu.:1244   1st Qu.: 24.00   1st Qu.:776.0   1st Qu.:162.0
##  Median :2498   Median : 50.00   Median :850.0   Median :280.0
##  Mean   :2497   Mean   : 49.85   Mean   :849.9   Mean   :293.3
##  3rd Qu.:3749   3rd Qu.: 74.00   3rd Qu.:925.0   3rd Qu.:380.0
##  Max.   :5000   Max.   :100.00   Max.   :999.0   Max.   :712.0
##   total_items    payment_method       created_at                 
##  Min.   :1.000   Length:4854        Min.   :2017-03-01 00:08:09  
##  1st Qu.:1.000   Class :character   1st Qu.:2017-03-08 07:02:59  
##  Median :2.000   Mode  :character   Median :2017-03-16 00:18:47  
##  Mean   :1.948                      Mean   :2017-03-15 22:24:13  
##  3rd Qu.:3.000                      3rd Qu.:2017-03-23 10:39:30  
##  Max.   :5.000                      Max.   :2017-03-30 23:55:35  
```

We can see that the new AOV is $293.3

Q1 A) Think about what could be going wrong with our calculation. Think about a better way to evaluate this data. We could see that the AOV value was assigned a wrong value due to outliers such as user_id=607 which have 704000 order_amount and 2000 as the total_items which was purchased on different days repeatably. Since each store sells only one type of shoe and even if we consider a company purchasing the same type of shoes in bulk, having the same purchase again and again in the same amount within 30 days and ordering 2000 shoes seems more like an incorrect entry of data. Hence that data was removed. Same way, the data for any order_amount greater than or equal to 715 was removed.

Q1 B) What metric would you report for this dataset? AOV seems like a correct metric to report.

Q1 C) What is its value? We can see that the new AOV is $293.3.