**Summer 2022 Data Science Intern Challenge**

**Thi Giang**

January 7th, 2022

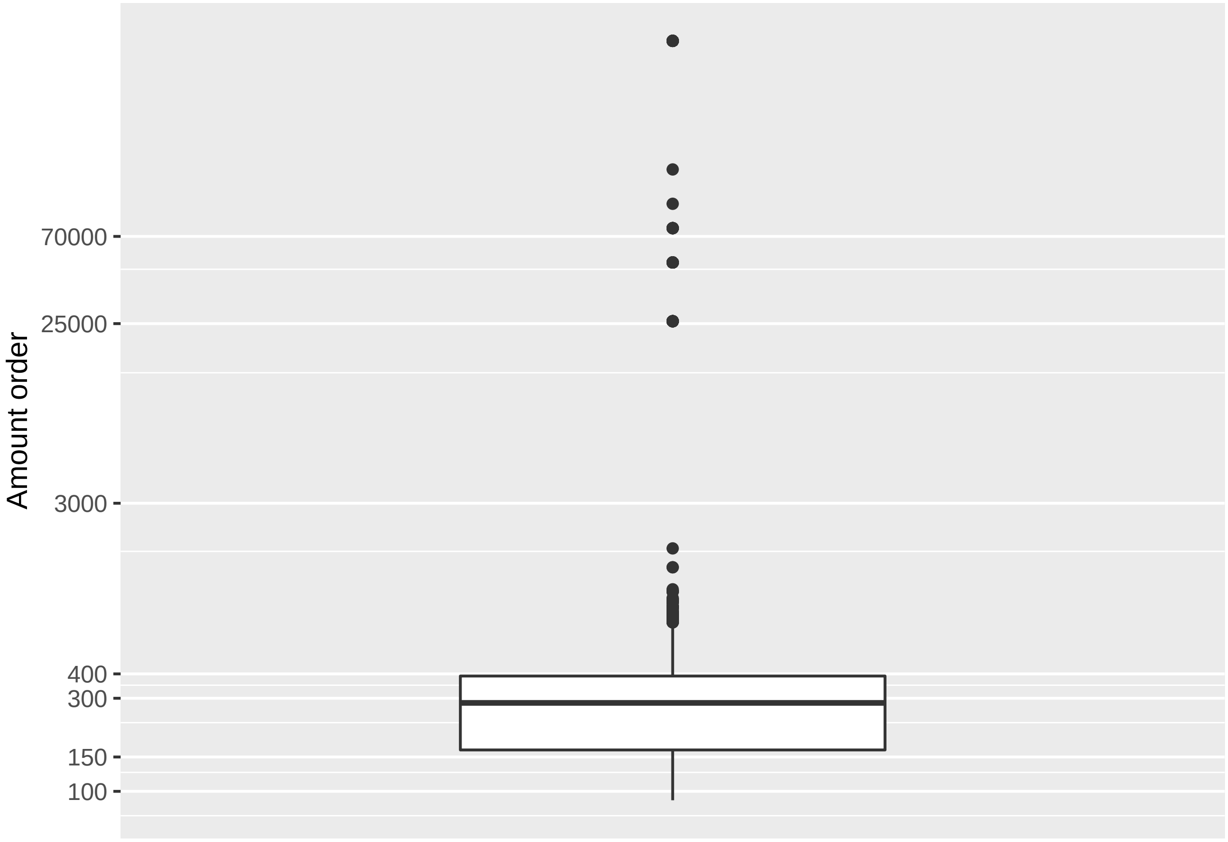
```
setwd("/Users/gabati/Documents/Internship/Shopify/")
library(readxl)
library(ggplot2)
challenge<-read_excel("Data.xlsx")
```

### Question 1a: Think about what could be going wrong with our calculation. Think about a better way to evaluate this data.

Answer:

As given in the begining that the shops sell a affordable item—snearker, but the AOV of $3145.13 per oder is too high. The AOV here was calculated by taking the average revenue of all 100 shops over 5,000 orders (total orders of 100 shops in 30 days). Let have a quick look into the data by observing the boxplot:

```
ggplot(data=data.frame(challenge$order_amount), aes(y=challenge$order_amount))+
  geom_boxplot()+
  scale_x_discrete()+
  coord_trans(y= "log10")+
  ylab("Amount order")+
  scale_y_continuous(breaks=c(100,150,300,400,3000,25000, 70000))
```



**There are two things could be going wrong here.**

1. First, the mean is extremely sensitive with outliers but this dataset has many potential outliers. There is a shop that sell sneakers with extremely high price (shop 78 with price of $25,725 per pair). Shop 42 that sales with large quantity per order (2000 items) compared to the majority of the data. It makes these orders become outliers of the dataset. Therefore, if we just naively take the average per order, we will get high AOV but it does not reflect correctly customer buying behavior or their amount purchase each time.

```
table(challenge$total_items)
```

```
##
##    1    2    3    4    5    6    8 2000
## 1830 1832  941  293   77    9    1   17
```

```
cat("Shop that had the most quantity sold in one order is shop", challenge$shop_id[which.max(challenge$t
```

```
## Shop that had the most quantity sold in one order is shop 42
```

```
unit_price<-challenge$order_amount/challenge$total_items
cat("Shop that sold the most expensive item is shop ", challenge$shop_id[which.max(unit_price)], " at $
```

```
## Shop that sold the most expensive item is shop 78 at $25725/ item
```

2. Second, as we all aware that the error might present during collecting the data, I suspect that some observations in this dataset are potentially wrong data. If we look closer on the orders of shop 42, we see that there are multiple orders have the EXACT SAME details about user id, order amount, date and time (see the result below). This may happen when the payment was cancled due to error but the program still recorded the order in database. Plus, these transactions have the highest order amounts of $704,000 in the whole dataset. So if these transactions are wrong, it obviously indicates that the AOV is misleading.

```
 challenge[c(521,4647,61,16,2298,1437,2154,1363,1603,1563,4869,1105,333,4883,2836,2970,4057),]
```

```
## # A tibble: 17 x 7
##    order_id shop_id user_id order_amount total_items payment_method
##       <dbl>   <dbl>   <dbl>        <dbl>       <dbl> <chr>
##  1      521      42     607       704000        2000 credit_card
##  2     4647      42     607       704000        2000 credit_card
##  3       61      42     607       704000        2000 credit_card
##  4       16      42     607       704000        2000 credit_card
##  5     2298      42     607       704000        2000 credit_card
##  6     1437      42     607       704000        2000 credit_card
##  7     2154      42     607       704000        2000 credit_card
##  8     1363      42     607       704000        2000 credit_card
##  9     1603      42     607       704000        2000 credit_card
## 10     1563      42     607       704000        2000 credit_card
## 11     4869      42     607       704000        2000 credit_card
## 12     1105      42     607       704000        2000 credit_card
## 13      333      57     794          441           3 credit_card
## 14     4883      42     607       704000        2000 credit_card
## 15     2836      42     607       704000        2000 credit_card
## 16     2970      42     607       704000        2000 credit_card
## 17     4057      42     607       704000        2000 credit_card
## # ... with 1 more variable: created_at <dttm>
```

### Think about a better way to evaluate this data.

As I suspected that there might be wrong observations in the data, I would have double checked and clea

I think a better way to evaluate this dataset is to separate the outliers and the majority of the data a

### Question 1b What metric would you report for this dataset?

One quick way to deal with outlier is to use Median which is not affected by outliers. However, if we us

Instead, we can separate the data into two groups. The first group include the orders that have regular

### Question 1c. What is its value?

If we simply want to choose a better metric, median of the data which is $284 is a better choice compar

```r
cat("Median is $", median(challenge$order_amount), sep="")
```

```
## Median is $284
```

If separate the dataset into groups that has similar characteristic, the AOV for group 1 is $302.58.

```r
shop42_index<-which(challenge$order_amount==704000)
shop78_index<-which(challenge$shop_id==78)
outlier<-c(shop42_index,shop78_index)
group1<-challenge[-outlier,]
cat("The average order value of group 1 is $", round(sum(group1$order_amount)/length(group1$order_amoun
```

```
## The average order value of group 1 is $302.58
```

```r
AOV_TABLE<-data.frame()
for (i in 1:100)  {
  Total_order_amount<-sum(challenge$order_amount[which(challenge$shop_id==i)])
  Shop_id<-i
  Total_order<-length(which(challenge$shop_id==i))
  Total_items<-sum(challenge$total_items[which(challenge$shop_id==i)])
  Price_unit<-Total_order_amount/Total_items
  AOV_by_shop<-round(Total_order_amount/Total_order,2)
  by_shop<-cbind(Shop_id,Total_order,Total_items,Total_order_amount, Price_unit,AOV_by_shop)
  AOV_TABLE<-rbind(AOV_TABLE,by_shop)
}
AOV_TABLE
```

```
##     Shop_id Total_order Total_items Total_order_amount Price_unit AOV_by_shop
## 1         1          44          86              13588        158      308.82
## 2         2          55         102               9588         94      174.33
## 3         3          48          99              14652        148      305.25
## 4         4          51         103              13184        128      258.51
## 5         5          45          92              13064        142      290.31
## 6         6          59         121              22627        187      383.51
## 7         7          56         109              12208        112      218.00
## 8         8          46          84              11088        132      241.04
## 9         9          59         117              13806        118      234.00
## 10       10          53         119              17612        148      332.30
## 11       11          49          95              17480        184      356.73
## 12       12          53          93              18693        201      352.70
## 13       13          63         136              21760        160      345.40
## 14       14          58         121              14036        116      242.00
## 15       15          52         105              16065        153      308.94
## 16       16          41          71              11076        156      270.15
## 17       17          53         100              17600        176      332.08
## 18       18          51         112              17472        156      342.59
## 19       19          64         126              20538        163      320.91
## 20       20          52         103              13081        127      251.56
## 21       21          46         100              14200        142      308.70
## 22       22          48          90              13140        146      273.75
## 23       23          55         112              17472        156      317.67
## 24       24          55         126              17640        140      320.73
## 25       25          48          86              11180        130      232.92
## 26       26          49          95              16720        176      341.22
## 27       27          54         107              18083        169      334.87
## 28       28          43          84              13776        164      320.37
## 29       29          58         118              19234        163      331.62
## 30       30          56         108              16524        153      295.07
## 31       31          47          98              12642        129      268.98
## 32       32          42          79               7979        101      189.98
## 33       33          40          87              15051        173      376.28
## 34       34          50          96              11712        122      234.24
## 35       35          52         104              17056        164      328.00
## 36       36          50          98              12740        130      254.80
## 37       37          48         115              16330        142      340.21
## 38       38          35          72              13680        190      390.86
## 39       39          41          82              10988        134      268.00
## 40       40          48          88              14168        161      295.17
## 41       41          59         127              14986        118      254.00
## 42       42          51       34063           11990176        352   235101.49
## 43       43          58         107              19367        181      333.91
## 44       44          39          71              10224        144      262.15
## 45       45          58         110              15620        142      269.31
## 46       46          43          90              14940        166      347.44
## 47       47          47          84              12180        145      259.15
## 48       48          40          83               9711        117      242.78
## 49       49          53         115              14835        129      279.91
## 50       50          44          92              17756        193      403.55
## 51       51          46          89              16643        187      361.80
## 52       52          41          89              12994        146      316.93
## 53       53          68         130              14560        112      214.12
## 54       54          50         104              13832        133      276.64
## 55       55          48          92              15732        171      327.75
## 56       56          37          69               8073        117      218.19
## 57       57          53         107              15729        147      296.77
## 58       58          59         109              15042        138      254.95
## 59       59          60         121              21538        178      358.97
## 60       60          47          93              16461        177      350.23
## 61       61          50         109              17222        158      344.44
## 62       62          43          83              13280        160      308.84
## 63       63          58         113              15368        136      264.97
## 64       64          43          88              11704        133      272.19
## 65       65          54         116              17864        154      330.81
## 66       66          53         103              16583        161      312.89
## 67       67          37          77              10087        131      272.62
## 68       68          47          88              11968        136      254.64
## 69       69          60         121              15851        131      264.18
## 70       70          59         117              20241        173      343.07
## 71       71          66         130              21320        164      323.03
## 72       72          46          89              14240        160      309.57
## 73       73          58         118              19470        165      335.69
## 74       74          38          76              11628        153      306.00
## 75       75          42          79              10112        128      240.76
## 76       76          42          87              13485        155      321.07
## 77       77          50          90              14040        156      280.80
## 78       78          46          88            2263800      25725    49213.04
## 79       79          54          98              17738        181      328.48
## 80       80          45          93              13485        145      299.67
## 81       81          59         128              22656        177      384.00
## 82       82          42          83              14691        177      349.79
## 83       83          42          81              10449        129      248.79
## 84       84          59         132              20196        153      342.31
## 85       85          35          67              11524        172      329.26
## 86       86          52         111              14430        130      277.50
## 87       87          52         102              15198        149      292.27
## 88       88          50         101              17776        176      355.52
## 89       89          61         118              23128        196      379.15
## 90       90          49         111              19758        178      403.22
## 91       91          54         110              17600        160      325.93
## 92       92          42          76               6840         90      162.86
## 93       93          59         111              12654        114      214.47
## 94       94          45         100              13400        134      297.78
## 95       95          39          74              12432        168      318.77
## 96       96          51         110              16830        153      330.00
## 97       97          48          96              15552        162      324.00
## 98       98          58         107              14231        133      245.36
## 99       99          54          94              18330        195      339.44
## 100     100          40          77               8547        111      213.68
```

In group 2, the observations are divided into two catagories. One is order with high quantity and one is order with high price item. Therefore, I used AOV_TABLE to extract the AOV of shop 78 which is $49213.04 per order. This number is more suitable to be the estimation of the average order value for this shop. Shop 42 sells sneaker with price of $352 per items. This price is on the high side compared to the usual price of sneaker, but it is still affordable. This shop had 51 orders in 30 days window. Among those, 17 orders are big orders with the total sell of 2000 items each. The AOV for this shop is $235101.49 because of the big transactions. This shop can actually divide their customer into regular and high spender. They can use the AOV of group 1 to target on customers that spend normally and have some offer, reward program to high spenders to keep them back to the shop.

**Question 2: For this question you'll need to use SQL. Follow this link to access the data set required for the challenge. Please use queries to answer the following questions. Paste your queries along with your final numerical answers below.**

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

Answer: 54 orders were shipped by Speedy Express in total. Solution 1:

```
SELECT COUNT(Orders.OrderID) AS ShipBySpeedy FROM Orders
WHERE Orders.ShipperID=(SELECT Shippers.ShipperID FROM Shippers WHERE ShipperName='Speedy Express');
```

Solution 2 (using join tables)

```
SELECT Count(Orders.ShipperID) AS ShipBySpeedy FROM Orders
  JOIN Shippers On Orders.ShipperID = Shippers.ShipperID
    WHERE ShipperName='Speedy Express';
```

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

Answer: The last name of the employee with most orders is Peacock with total of 40 orders.

```
SELECT Employees.LastName
FROM Employees
WHERE Employees.EmployeeID=
  (SELECT EmployeeID
  FROM (SELECT COUNT(Orders.OrderID) AS TotalOrder, Orders.EmployeeID FROM Orders
        GROUP BY Orders.EmployeeID
        ORDER BY TotalOrder DESC
        LIMIT 1));
```

As the question asked for total orders not total quantity so I use Orders table to get the EmployeeID with highest number of order, and get the LastName of that employee from the Employee table.

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

Answer: Boston Crab Meat is the product which was ordered the most by custoners in Germany with 160 orders.

```
SELECT Products.ProductName FROM Products
WHERE Products.ProductID=
  (SELECT ProductID FROM
    (SELECT ProductID, SUM(Quantity) AS TotalQuantity FROM
      ((SELECT Orders.OrderID FROM Orders
          JOIN Customers ON Orders.CustomerID=Customers.CustomerID
          WHERE Customers.Country="Germany") AS GermanyOrders
          JOIN OrderDetails ON GermanyOrders.OrderID=OrderDetails.OrderID)
   GROUP BY ProductID
   ORDER BY TotalQuantity DESC LIMIT 1));
```