intern.R

dzwu

2022-05-21

```
rm(list = ls());
Sys.setlocale(category = 'LC_ALL', locale = 'English_United States')
```

```
## [1] "LC_COLLATE=English_United States.1252;LC_CTYPE=English_United States.1252;LC_MONETARY=English_U
```

```
setwd('C:/Users/dzwu/Documents/study/document/job/shopify')
df = read.csv('2019 Winter Data Science Intern Challenge Data Set - Sheet1.csv')


#Question a: a better way to evaluate the data.

#First check whether the data are collected in the same period.
df$created_at = as.Date(df$created_at)
df$month = months(df$created_at,)

#The time period is correct for all of the data.
unique(df$month)
```

```
## [1] "March"
```

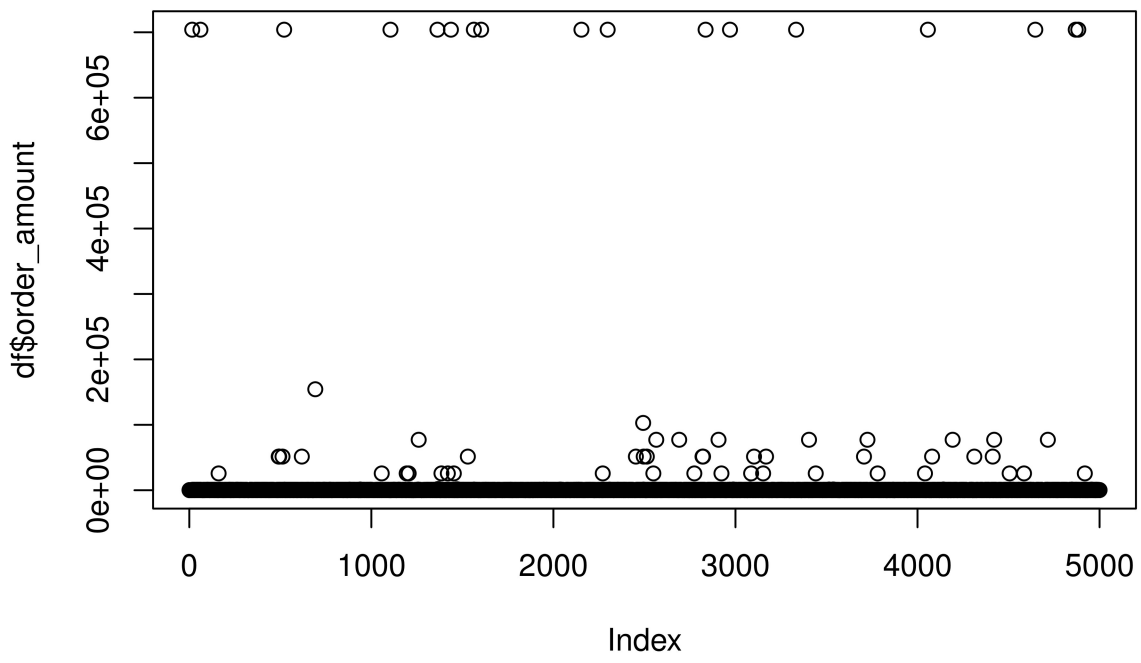
```
df = df[,-8]


#Then summarize order_amount and total_items to see whether any missing value or abnormal value exists.
summary(df$order_amount)
```

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

```
summary(df$total_items)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    1.000    1.000    2.000    8.787    3.000 2000.000
```

```
#By plotting the order_amount, we can see a few values that are far higher than the majority.
#This may indicate something that we need to worry about.
plot(df$order_amount)
```



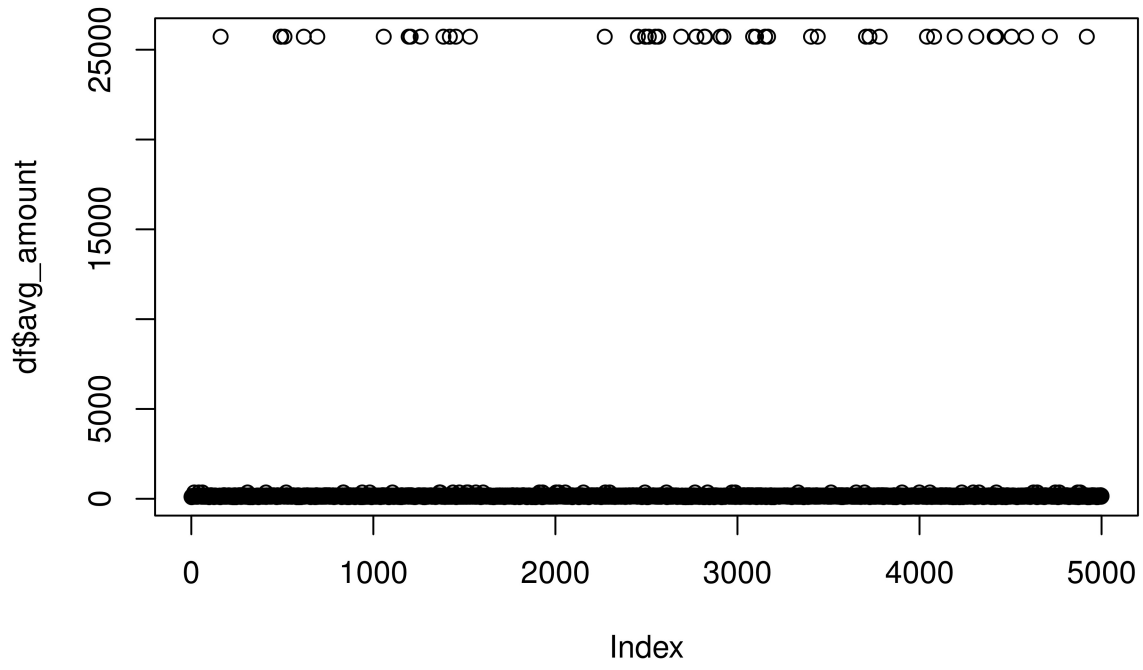

```
#Using the current data to calculate the AOV(average order value) will end up with $3145,
#It is unusual for orders of sneakers.
mean(df$order_amount)
```

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

```
#Calculate the order_amount per item for each order.
df$avg_amount = df$order_amount / df$total_items
summary(df$avg_amount)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    90.0   133.0   153.0   387.7   169.0 25725.0
```

```
#We can see that there are 46 observations with a abnormal order_amount per item(higher than $25000).
plot(df$avg_amount)
```



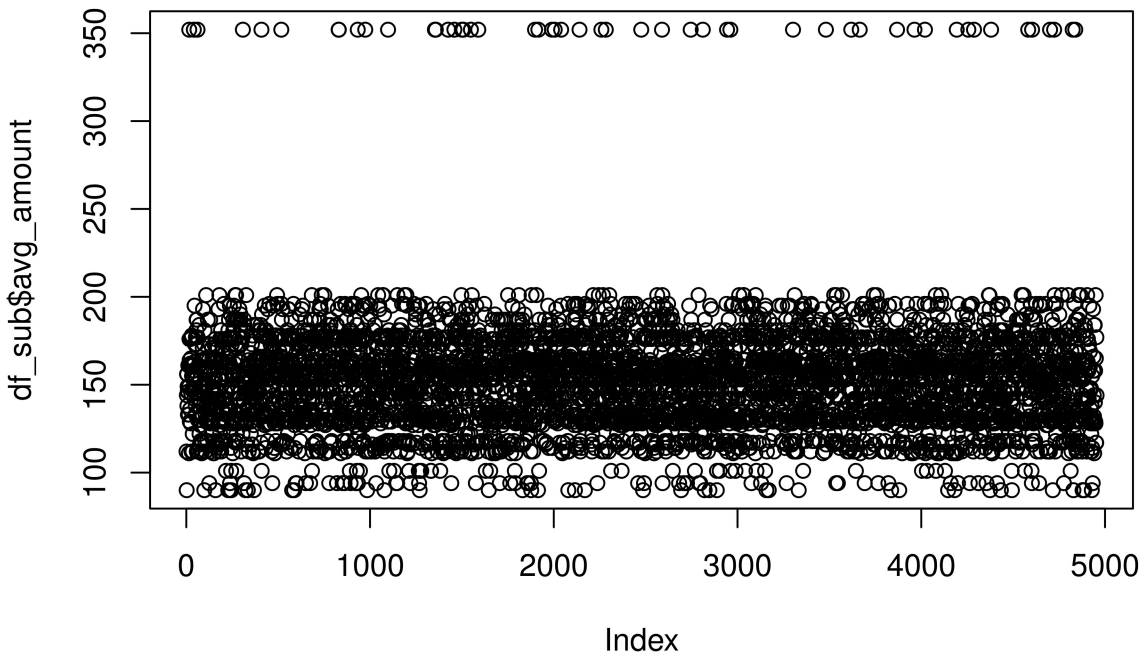

```
length(df$avg_amount[df$avg_amount > 500])
```

```
## [1] 46
```

```
#Since I cannot check the source of these data, I simply drop these observations.
#And now the order_amount per item seems to be more reasonable.
#But the AOV is still at $2717,indicating some other problems.
df_sub = subset(df, avg_amount <= 500)
summary(df_sub$order_amount)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      90     163     284    2717     390  704000
```

```
plot(df_sub$avg_amount)
```




```
#I notice that several observations recorded hundreds or even thousands of items in a single order.
#Since only 100 sneaker shops is selling and each of these shops sells only one model of shoe,
#the order with more than 100 items may indicates the existence of resellers.
#These outliers may disturb the evaluation of AOV.
summary(df_sub$total_items)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##     1.000     1.000     2.000     8.851     3.000 2000.000
```

```
#So I contain observations that with fewer than 100 items in the new data set.
df_sub_2 = subset(df_sub, total_items <= 100)

#Now the AOV is $302.58, which is plausible for orders of sneakers.
summary(df_sub_2$order_amount)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    90.0   163.0   284.0   302.6   387.0  1760.0
```

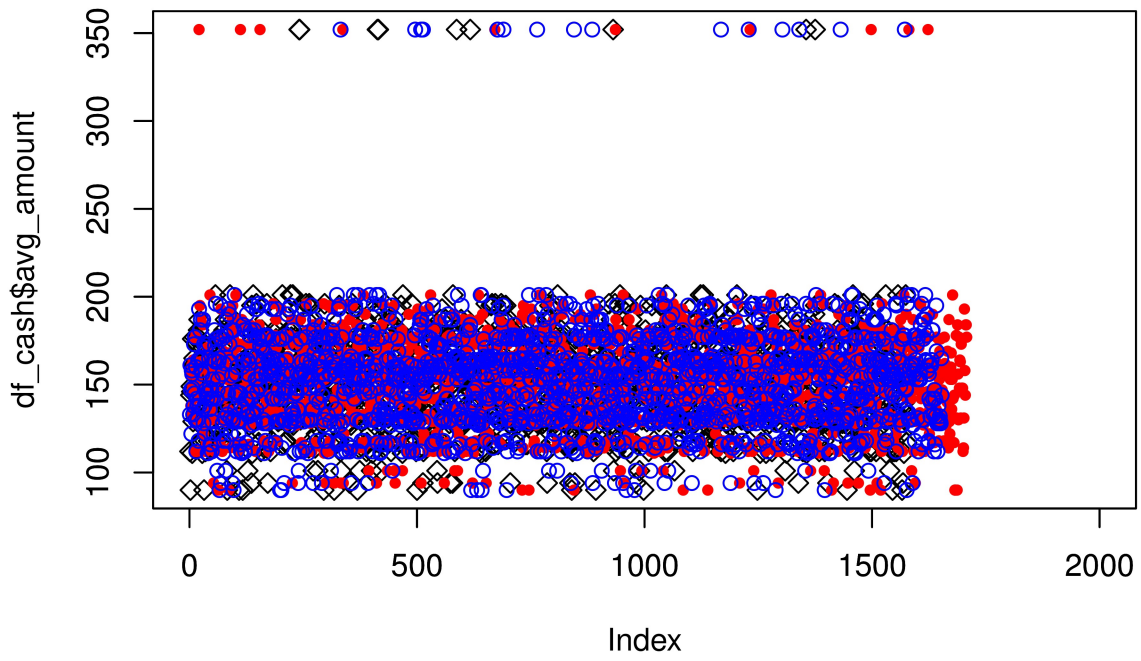
```
AOV = mean(df_sub_2$order_amount)
AOV
```

```
## [1] 302.5805
```

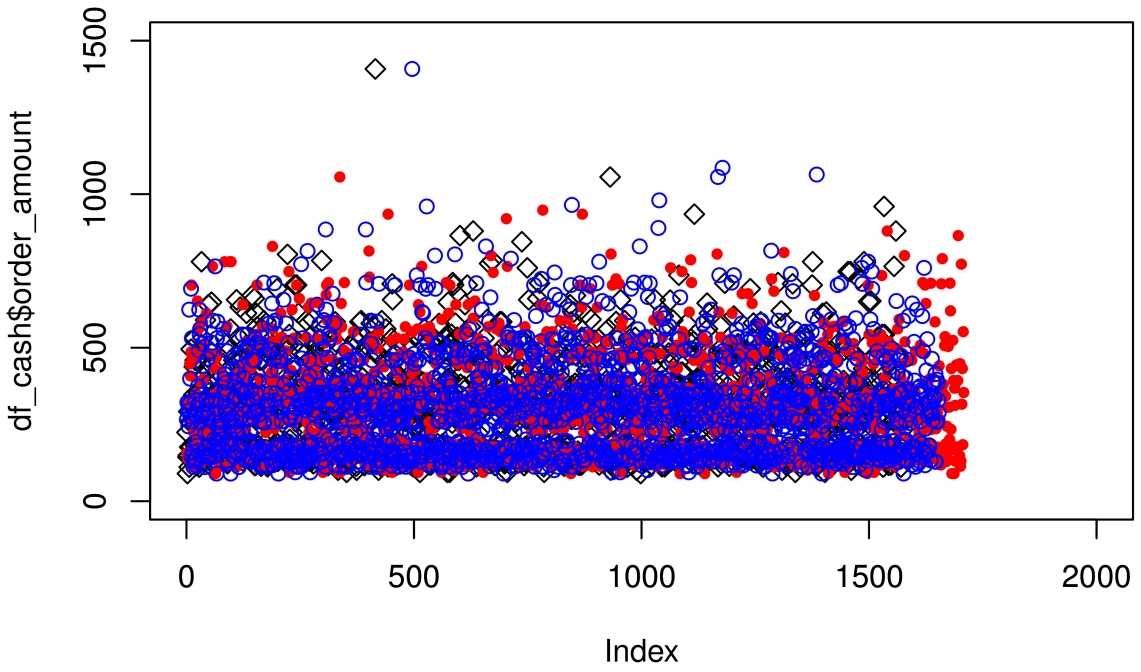
```r
#Then I plot the order_amount per item of each payment method to see whether there's a difference.
unique(df_sub_2$payment_method)
```

```
## [1] "cash"        "credit_card" "debit"
```

```r
df_credit = subset(df_sub_2, payment_method == 'credit_card')
df_debit = subset(df_sub_2, payment_method == 'debit')
df_cash = subset(df_sub_2, payment_method == 'cash')

plot(df_cash$avg_amount, type = 'p', pch = 5, xlim = c(0,2000))
par(new = T)
plot(df_credit$avg_amount, type = 'p', pch = 20, col = rgb(1, 0, 0), xlab = '', ylab = '', xlim = c(0,2
par(new = T)
plot(df_debit$avg_amount, col = rgb(0, 0, 1), xlab = '', ylab = '', xlim = c(0,2000))
```




```r
#I also do the same thing for order_amount, and no significant difference has been found.
#So AOV can be simply concluded as $302.58.
plot(df_cash$order_amount, type = 'p', pch = 5, ylim=c(0,1500), xlim = c(0,2000))
par(new = T)
plot(df_credit$order_amount, type = 'p', pch = 20, col = rgb(1, 0, 0), ylim=c(0,1500), xlab = '', ylab
par(new = T)
plot(df_debit$order_amount, col = rgb(0, 0, 1), ylim=c(0,1500), xlab = '', ylab = '', xlim = c(0,2000))
```




```
##############################################################################################
##############################################################################################

#Question b and c: other metrics I would report for this dataset.

#We can calculate AOF(average order frequency)
#The AOF during 2017-03 is 16,
#indicating that the average amount of orders placed by each customer during the time is 16.
AOF = nrow(df_sub_2) / length(unique(df_sub_2$user_id))
AOF
```

```
## [1] 16.45667
```

```
#We can also calculate the ACV( Average customer Value).
#The ACV during 2017-03 is $4979.47,
#indicating the average revenue value that each customer brings to the business during the time is 4979
ACV = sum(df_sub_2$order_amount) / length(unique(df_sub_2$user_id))
ACV
```

```
## [1] 4979.467
```

```
#ACV can calculated by another way, which is multiplying AOV by AOF.
#The result is the same as the previous method
ACV_2 = AOF * AOV
ACV_2
```

```
## [1] 4979.467
```

```
#If we can have the access to the data of a longer time span,
#we can calculate Average Customer Lifespan (ACL) and the Customer Lifetime Value (CLV).
```