Shopify

Yongjia Huang

5/12/2022

Fall 2022 Data Science Intern Challenge

Question 1:

```
# import data
data <- read.csv("2019 Winter Data Science Intern Challenge Data Set - Sheet1.csv")
head(data)
```

```
##   order_id shop_id user_id order_amount total_items payment_method
## 1        1      53     746          224           2           cash
## 2        2      92     925           90           1           cash
## 3        3      44     861          144           1           cash
## 4        4      18     935          156           1    credit_card
## 5        5      18     883          156           1    credit_card
## 6        6      58     882          138           1    credit_card
##            created_at
## 1 2017-03-13 12:36:56
## 2 2017-03-03 17:38:52
## 3  2017-03-14 4:23:56
## 4 2017-03-26 12:43:37
## 5  2017-03-01 4:35:11
## 6 2017-03-14 15:25:01
```

```
any(is.na(data))
```

```
## [1] FALSE
```

```
# since the data don't had any missing value, we can use it
```

Now let look at the summary table for the spreadsheet

```
summary(data)
```

```
##     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        Length:5000
##  1st Qu.:   1.000   Class :character   Class :character
##  Median :   2.000   Mode  :character   Mode  :character
##  Mean   :   8.787
##  3rd Qu.:   3.000
##  Max.   :2000.000
```

From the summary table, we observed that the average order value is $3145, which is extremely higher than expected since the sneaker that is selling in 100 stores are consider affordable. We also observed that the maximum order value is $704000 and maximum items that purchase is 2000. Now the best way to analysis the data is to find the shoe price for each shop since one shop only selling one model of shoe.

```
shoe_price <- numeric(0)
for (i in seq(100)){
  data_copy <- data[data$shop_id == i,]
  shoe_price <- c(shoe_price, sum(data_copy$order_amount) / sum(data_copy$total_items))
}
shoe_price
```

```
##   [1]   158    94   148   128   142   187   112   132   118   148   184   201
##  [13]   160   116   153   156   176   156   163   127   142   146   156   140
##  [25]   130   176   169   164   163   153   129   101   173   122   164   130
##  [37]   142   190   134   161   118   352   181   144   142   166   145   117
##  [49]   129   193   187   146   112   133   171   117   147   138   178   177
##  [61]   158   160   136   133   154   161   131   136   131   173   164   160
##  [73]   165   153   128   155   156 25725   181   145   177   177   129   153
##  [85]   172   130   149   176   196   178   160    90   114   134   168   153
##  [97]   162   133   195   111
```

Now observed that most of the store having the affordable shoe price, however, there is one store it shoe price is $25725, which is extremely high.

```
which(shoe_price == 25725)
```

```
## [1] 78
```

This given that the store_id 78 had a shoe price that is very high. Now look at all the purchase at store_id 78

```
data_copy <- data[data$shop_id == 78,]
data_copy$order_amount / data_copy$total_items
```

```
##  [1] 25725 25725 25725 25725 25725 25725 25725 25725 25725 25725 25725 25725
## [13] 25725 25725 25725 25725 25725 25725 25725 25725 25725 25725 25725 25725
## [25] 25725 25725 25725 25725 25725 25725 25725 25725 25725 25725 25725 25725
## [37] 25725 25725 25725 25725 25725 25725 25725 25725 25725 25725
```

Now it can see that the shoe price for this store is $25725, which is unreasonable, this mean that the it maybe the mistake on the system that input a wrong item price. If we look at the shoe price for other store, most of the stores selling shoe price that is below $200. Since we don't know the actual shoe price for store 78, we can assume that the store is mistakenly input the cents value for shoe instead of dollar. Then let assume that the original value for the shoe is $257.25, and plug it into the original data set.

```
price_assumption <- 25725/100
data_new <- data
index <- which(data$shop_id == 78)
for (i in index){
  data_new[i,]$order_amount = data_new[i,]$total_items * price_assumption
}
```

```
AOV <- sum(data_new$order_amount) / sum(data_new$total_items)
```

Since Average Order Value (AOV) = Revenue / number of orders, then we had the new value: 306.9118263

Hence, the better way to analysis the data is to split the data into different stores to see if they are in the price that we expected, if yes, then we can calculate the average order value by total dolloar of order amount divided by total number of orders. Therefore, the metrix for this dataset is order_amount and order_items.

Problem 2

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

```
SELECT COUNT(ShipperID)

FROM [Orders]

WHERE ShipperID == 1
```

Number of Records: 1

COUNT(ShipperID)
54

Answer: The total orders that were shipped by Speedy Express were 54.

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

```
Select
    Employees.LastName,
    count(Orders.OrderID)
AS Numbers From Orders
Inner JOIN Employees ON Orders.EmployeeID == Employees.EmployeeID
Group by
    Employees.LastName
Order By
    Numbers DESC
```

Number of Records: 9

LastName	Numbers
Peacock	40
Leverling	31
Davolio	29
Callahan	27
Fuller	20
Suyama	18
King	14
Buchanan	11
Dodsworth	6

Answer: The last name of employee with the most orders is Peacockered the most by customers in Germany was Gorgonzola Telino.

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

```
Select
        Products.ProductName,
        Customers.Country,
        Count(Orders.OrderID)
As Number From Customers
Inner Join Orders on Customers.CustomerID = Orders.CustomerID
Inner Join OrderDetails on Orders.OrderID = OrderDetails.OrderID
Inner Join Products On OrderDetails.ProductID = Products.ProductID
Where Country = 'Germany'
Group By
        ProductName
Order By
        Number DESC
```

Result:

Number of Records: 45

ProductName	Country	Number
Gorgonzola Telino	Germany	5
Lakkalikööri	Germany	4
Boston Crab Meat	Germany	4
Tunnbröd	Germany	3
Mozzarella di Giovanni	Germany	3
Inlagd Sill	Germany	3
Teatime Chocolate Biscuits	Germany	2
Tarte au sucre	Germany	2
Rössle Sauerkraut	Germany	2
Rhönbräu Klosterbier	Germany	2
Pâté chinois	Germany	2
Perth Pasties	Germany	2
Konbu	Germany	2
Gudbrandsdalsost	Germany	2

Answer: The product that was ordered the most by customers in Germany was Gorgonzola Telino.