Winter 2021 Data Science Intern Challenge

Please complete the following questions, and provide your thought process/work. You can attach your work in a text file, link, etc. on the application page. Please ensure answers are easily visible for reviewers!

**Question 1:** Given some sample data, write a program to answer the following: [click here to access the required data set](https://docs.google.com/spreadsheets/d/16i38oonuX1y1g7C_UAmiK9GkY7cS-64DfiDMNiR41LM/edit#gid=0)

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.

It seems the order\_amount is imbalanced in this dataset. Compare with most orders, there are few orders with huge order\_amount. If it just simply uses total revenue / number of orders to calculate AVO ($3145.13), the AOV will be dominated and miss leaded by these huge order\_amount. Thus, I will consider remove outlier (huge order\_amount) record from the dataset and then recalculate the AOV with following R Code.

######################### Average Order Value ######################

# Read the csv file

dataset <- read.csv("Challenge.csv")

#Print the structure of the dataframe

str(dataset)

#outlier

outl <- boxplot(dataset$order\_amount,plot=FALSE)

datasetout <- dataset[!dataset$order\_amount %in% outl$out,]

#Calculate AVO

AVO <- sum(datasetout[,4])/nrow(datasetout)

The recalculate the AOV is $293.72 which is relatively affordable.

1. What metric would you report for this dataset?

For this dataset, I would like use the payment\_method data to analysis and report the structure of customer payment. The following R Code show the process of payment method analysis:

######################### Payment Method Analysis ######################

# Read the csv file

dataset <- read.csv("Challenge.csv")

#Calculate the amount payment

sum(dataset[,4])

#Calculate the payment of credit card, debit and cash

credit <- dataset[which(dataset$payment\_method=='credit\_card'),]

debit <- dataset[which(dataset$payment\_method=='debit'),]

cash <- dataset[which(dataset$payment\_method=='cash'),]

#Calculate the % of each payment method

CreditPayment <- sum(credit[,4])/sum(dataset[,4])\*100

DebitPayment <- sum(debit[,4])/sum(dataset[,4])\*100

CashPayment <- sum(cash[,4])/sum(dataset[,4])\*100

The analysis shows the credit card payment is 82.32% of the amount payment; the debit payment is 10.27% of the amount payment, the cash payment is only 7.40% of the amount payment.

1. What is its value?

Knowing the payment structure of each store may help the merchants understand their customer more. The payment method builds up the customer shopping hobbit and consumption ability. It also helps to plan promotional activities and develop the cooperation with bank and commercial institution.

**Question 2:** For this question you’ll need to use SQL. [Follow this link](https://www.w3schools.com/SQL/TRYSQL.ASP?FILENAME=TRYSQL_SELECT_ALL) 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.

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

Answer:

For Speedy Express, The ShipperID is “1” which can find in Shippers table. Thus, the SQL query is:

SELECT COUNT(ShipperID) AS SpeedyExpress FROM Orders

WHERE ShipperID='1'

There are 54 orders were shipped by Speedy Express in total.

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

Answer:

To find answer, it needs to count the orders of each employee with the following SQL query:

SELECT EmployeeID, COUNT(EmployeeID) FROM Orders

GROUP BY EmployeeID

In the Result table, it is clear EmployeeID 4 with the most orders which exactly 40 orders. With checking the Employees table, the last name of EmployeeID 4 is Peacock.

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

Answer:

To find answer, it needs to join OrderDetails, Orders and Customers tables with keys for the query as bellow:

SELECT ProductID, COUNT(ProductID), SUM(Quantity) FROM OrderDetails

INNER JOIN Orders

ON OrderDetails.OrderID = Orders.OrderID

INNER JOIN Customers

ON Customers.CustomerID = Orders.CustomerID

WHERE Customers.Country="Germany"

GROUP BY ProductID

In the Result table, it shows the ProductID 40 is the most ordered product which ordered 4 times in total 160 quantity by the customers in Germany. With checking the Products table, the ProductID 40 is Boston Crab Meat.

\*\*\*\*

SELECT ProductID, COUNT(ProductID) FROM OrderDetails

INNER JOIN Orders

ON OrderDetails.OrderID = Orders.OrderID

INNER JOIN Customers

ON Customers.CustomerID = Orders.CustomerID

WHERE Customers.Country="Germany"

GROUP BY ProductID

\*\*\*\*

In the Result table, it shows the ProductID 31 is the most ordered product which ordered 5 times by the customers in Germany. With checking the Products table, the ProductID 31 is Gorgonzola Telino.