Summer 2022 Data Science Intern Challenge

Name: Chaitanya Shekhar Deshpande
Email ID: deshpandec998@gmail.com

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

> 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.

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

Answer: **There is actually nothing wrong with the Average Order Value calculation. By definition, Average Order Value is Total Revenue / Total Order = 15725640/5000 = $3145.128 (Please refer the attached code for these values). The value is high because there are a few very costly transactions in the middle, for example: for order_id 16, the order_amount is 704000, because of which the AOV is so high.However, the thing which is getting ignored is the total number of items per order as well. We need to consider the total_items per order while evaluating our data. If we need to evaluate the average price of a pair of sneakers, a better approach would be to evaluate the amount per item, and then total the same, and evaluate the average value over the number of orders. This would help us provide a more relatively accurate value per sneaker per order. For instance, for order_id 16, the cost per sneakers' pair is 704000/2000 = $ 352, which makes sense. Another way to solve this could be to remove such data, but it won't make sense, as it will lead to inaccuracies.**

b. What metric would you report for this dataset?

Answer: **As suggested in the above answer, a good metric would be to evaluate the amount per item, total for each order, and then take an average over the number of orders, i.e. 5000. For example, considering the 1st order in the dataset, the order_amount value is 224, and the total_items are 2, so the amount per item is 224/2=112. We evaluate the same for all orders, take its sum and then divide by 5000 to get this metric value.**

c. What is its value?

Answer: **After coding out the above metric, we get the value as 387.7428.**
<u>**Code**</u>: **The executing code is also present in the current GitHub link.**

```
import pandas as pd
import numpy as np

data=pd.read_csv("/content/sample_data/2019 Winter Data Science Intern Challenge Data Set - Sheet1.csv”)

total_revenue=data["order_amount"].sum()
total_revenue

# Output: 15725640

AOV=total_revenue/len(data)
AOV

# Output: 3145.128

# New Metric Evaluation begins here:

total=0
for i in range(len(data)):
  total+=data["order_amount"][i]/data["total_items"][i]
new_metric=total/len(data)

new_metric
# Output: 387.7428
```

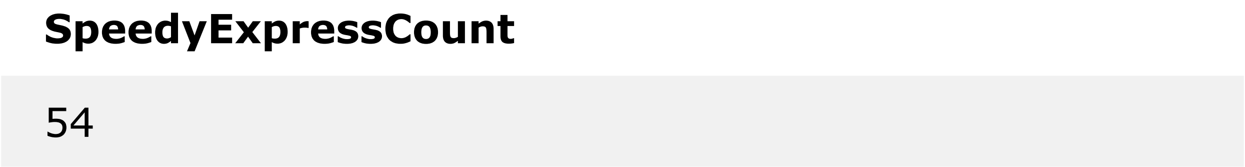
Question 2: For this question you’ll need to use SQL. [Follow this link]() to access the data set required for the challenge. Please use queries to answer the following questions. Paste your queries along with your final numerical answers below.

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

Answer:

```
SELECT COUNT(Orders.OrderID) AS SpeedyExpressCount
FROM Orders
INNER JOIN Shippers ON Orders.ShipperID = Shippers.ShipperID WHERE
Shippers.ShipperName="Speedy Express";
```

<u>**OUTPUT: 54**</u>



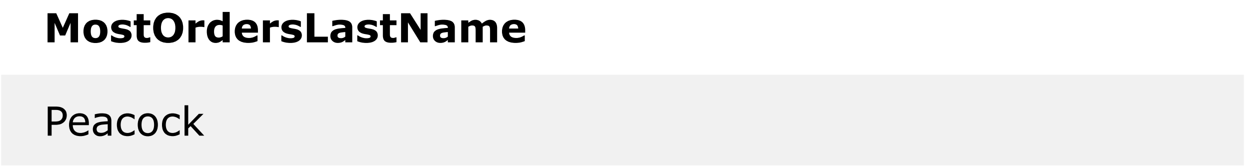
SpeedyExpressCount
54

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

Answer:

```
SELECT TOP 1 LastName AS MostOrdersLastName FROM (SELECT
Employees.LastName, COUNT(Orders.OrderID) AS TotalOrders FROM Orders
INNER JOIN Employees ON Orders.EmployeeID = Employees.EmployeeID
GROUP BY LastName) ORDER BY TotalOrders DESC;
```

OUTPUT: Peacock



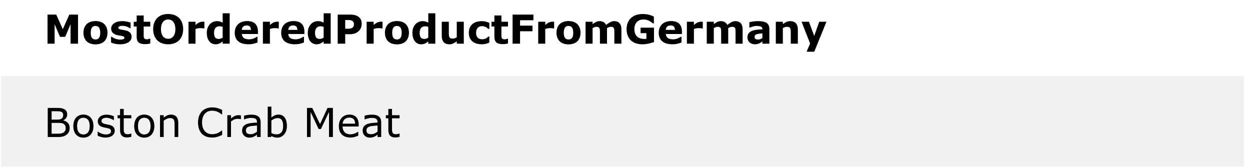
MostOrdersLastName
Peacock

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

Answer:

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

OUTPUT: Boston Crab Meat



MostOrderedProductFromGermany
Boston Crab Meat