Winter 2019 Data Science Intern Challenge

Question 1: Given some sample data, write a program to answer the following:

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.
b. What metric would you report for this dataset?
c. What is its value?

Answer:

a. In this dataset, there are some anomalous data. Some orders show an unexplained total purchase amount of $25725 for 1 item. Others show bulk orders of 2000 units for $704000. These will prevent us from calculating a realistic AOV.

b. By removing the anomalous data points, we can calculate a real AOV based on the remaining data. In this case, since every order is either small or very large, we can easily clean the dataset. We remove the first anomalous data type by only considering orders with an average item value < $1000, and we can remove both by only considering orders with a total order value < $5,000.

c. Our AOV after cleaning the first anomalous data type is $2717. This is our AOV_bulk, which contains bulk orders. Our second AOV after cleaning both anomalous data types is $302.58. This is our AOV_real, which represents the typical order value of 1-3 items.

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?
b. What is the last name of the employee with the most orders?
c. What product was ordered the most by customers in Germany?

Answer:

d. SELECT * FROM Orders WHERE ShipperID = 1
=> Total: 54

e. SELECT EmployeeID, COUNT(EmployeeID) FROM Orders GROUP BY EmployeeID ORDER BY COUNT(EmployeeID)
=> EmployeeID 4 = 40
=> "Peacock"

f. SELECT ProductID, SUM(Quantity) FROM [OrderDetails] WHERE OrderID IN (SELECT OrderID FROM [Orders] WHERE CustomerID IN (SELECT CustomerID FROM [Customers] WHERE Country = 'Germany')) GROUP BY ProductID ORDER BY SUM(Quantity)
=> ProductID 40 = 160
=> "Boston Crab Meat"

Please include your answers as an attachment on your application