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

> 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.

The first thing that comes to mind with a higher AOV is that there may be some outliers in the dataset. Upon doing some quick analysis, we can see that there are definitely some higher values in the order_amount column. Another thing I noticed upon inspecting the data was that the values in the total_items column vary. Corresponding with the higher AOV value of $704,000 is 2000 total_items. I was able to recreate the calculation that obtained an AOV of $3145.13, which was just taking the average of the order_amount column values. This cannot be an accurate measurement as the number of items vary for each order.

b. What metric would you report for this dataset?

Instead, we should attempt to normalize the values of the order_amount in order to get an accurate AOV measure. In order to do this, we can divide the order_amount column by the total_items column to get a normalized order amount, a column titled norm_order_amount. Then, we can take the average of the values in this column to get an accurate AOV value.

c. What is its value?

$387.74

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? **196**

```
SELECT COUNT(OrderID)
FROM ORDERS JOIN SHIPPERS
WHERE ShipperName == 'SpeedyExpress';
```

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

```
SELECT max(order_count), LastName
FROM Employees JOIN (SELECT COUNT(OrderID) as order_count, EmployeeID
FROM Orders
GROUP BY EmployeeID);
```

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

```
SELECT ProductName FROM(
SELECT MAX(prod_count), ProductID as prod_id
FROM
(SELECT COUNT(*) as prod_count, ProductID
FROM Orders JOIN OrderDetails on Orders.OrderID
WHERE CustomerID IN (SELECT CustomerID FROM Customers WHERE Country ==
'Germany')
GROUP BY ProductID) ), Products
WHERE prod_id == ProductID;
```