Summer 2022 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.

   It is possible that there are outliers in the dataset or repeated values.

   First, checking that dataset we can see there are no missing values and 100 unique shop IDs. Of the 5000 orders, the mean order value is $3145.13, the standard deviation is 41282.539 (high relative to the mean), the minimum order value is $90.00, and the maximum order value is $704000.00. We can see that the maximum order value is coming mainly from shop_id=42, and given the number of shoe orders being 2000, the average price per shoe would amount to $352. This would not be an outlier.

   A quick scan from the first 200 values in the dataset, we see that shop_id=78 has sold a single pair of shoes for $25725. Given the information that all the shoes are relatively affordable, this may be an outlier or a case of misprice.

   The best way to evaluate this dataset is using order value.

b. What metric would you report for this dataset?

   The median order value is the most appropriate metric to report for this dataset.

c. What is its value?

The median order value is $248.

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?

```
SELECT ShipperName, COUNT(Orders.ShipperID)
FROM Orders
INNER JOIN Shippers ON Orders.ShipperID = Shippers.ShipperID
GROUP BY Orders.ShipperID
```

There are 54 orders shipped by Speedy Express.

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

```
SELECT LastName, COUNT(Orders.EmployeeID)
FROM Orders
INNER JOIN Employees ON Orders.EmployeeID = Employees.EmployeeID
GROUP BY Orders.EmployeeID
Order by COUNT(Orders.EmployeeID) Desc;
```

The last name of the employee with the most orders is "Peacock."

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

```
SELECT ProductName, MAX(ProductCount)
FROM(
    SELECT Products.ProductName as ProductName, COUNT(Products.ProductsName) as ProductCount
    FROM Orders
    INNER JOIN Customers ON Orders.CustomerID = Customers.CustomerID
    INNER JOIN OrderDetails ON Orders.OrderID = OrderDetails.OrderID
    INNER JOIN Products ON Products.ProductID = OrderDetails.ProductID
    WHERE Country = "Germany"
    GROUP BY Products.ProductName
    Order by COUNT(Products.ProductName) Desc
    );
```

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