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.

A more thorough approach would require special attention to these extreme values. For instance, the problem for shop 42 seems to be the addition of '000' in the processing of total items for specific orders. We could first contact the store to ensure that this is indeed an error. If it is the case, correcting these values in our analysis would go a long way.

b. What metric would you report for this dataset?

Assuming that we do not have enough time to contact the merchant and assess the extreme values issue, we could use the median metric as it is much less sensible to extreme values. This will provide a much more reliable number.

c. What is its value?

```
The order amount median for a 30 day period is: 284.0
```

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 COUNT(DISTINCT(OrderID)) as 'Orders shipped by Speedy Express'
FROM [Orders] o
LEFT JOIN [Shippers] s
ON o.ShipperID = s.ShipperID
WHERE s.ShipperName = 'Speedy Express'
```

answer: 54

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

```
SELECT E.LastName,
      Count(DISTINCT O.OrderID) AS TotalOrders
FROM [Orders] O
INNER JOIN [Employees] E
  ON O.EmployeeID = E.EmployeeID
GROUP BY O.EmployeeID
Order BY Count(DISTINCT O.OrderID) DESC
LIMIT 1
```

answer: Peacock

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

```
WITH cte as (
SELECT * FROM [Orders] o
LEFT JOIN [OrderDetails] od
ON o.OrderID = od.OrderID
LEFT JOIN [Customers] c
ON o.CustomerID = c.CustomerID
WHERE c.Country = 'Germany'
)
SELECT ProductID, SUM(Quantity)
from cte
GROUP BY ProductID
ORDER BY SUM(Quantity) DESC
LIMIT 1
```

answer: 160 units of product 40 were ordered by customers based in Germany, making it the most ordered product by Germans.