Fall 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.

Assumptions:

- Each shop's performance measured as AOV is equally important.

The issue with this calculation is that the mean is very sensitive to outliers. In the dataset, there are multiple transactions made by the same user_id 607 consisting of 2000 items which has a big impact on the current metric. A better way could be to use a metric that is more robust to outliers, such as the median. This metric is used in the Box-Plot visualisation because of its robustness to outliers and wouldn't require any additional pre-processing such as removing outliers from the dataset. Another alternative when working with such data, is to use the logarithm on those values. Then we can use our baseline model which is calculating the mean and reverse the log transform by exponentiating the output, in this case the mean. Using the logarithm with the base 10, would yield a mean of $285.02 compared to $3145.12.

b. What metric would you report for this dataset?

Given this problem, I would opt to use the median for it's property of being robust to outliers and for it's interpretability. Explaining the concepts of transforming the values isn't as straightforward that explaining what consists of the median. It is also a good metric for reporting on the performance of all the stores, as the value returned will be the 50th percentile.

c. What is its value?

The median of the order value is 284$.

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(*) FROM Orders
JOIN Shippers USING(ShipperID)
WHERE ShipperName = 'Speedy Express';
```

Answer: 54

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

```
SELECT count(*) as TotalOrders, FirstName, LastName FROM Orders
JOIN Employees USING(EmployeeID)
GROUP BY EmployeeID
ORDER BY TotalOrders DESC
LIMIT 5;
```

Answer: Peacock

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

```
SELECT COUNT(*) as ProductSales, ProductName, ProductID from Orders
JOIN OrderDetails USING(OrderID)
JOIN Products USING(ProductID)
JOIN Customers USING(CustomerID)
WHERE Country = "Germany"
GROUP BY ProductID
ORDER BY ProductSales DESC LIMIT 5;
```

Answer: Gorgonzola Telino