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

Answers:

a. The AOV is influenced by outliers and skewed data. Here you are getting an AOV of $3145.13 because you are simply calculating the mean of all order amounts. Shop IDs 42 and 78 are the outliers as their order amounts lie in the range [352, 70400] and [25725, 154350] respectively. The range for other shops is much lesser.

b. A better way to evaluate this data would be to use the mode instead. Mode is not influenced by outliers. Thus mode is a better measure for tracking the revenue earned per order for businesses. As this is the most frequent order executed by customers. If the modal values are low, implementing strategies to increase the mode will have a positive impact on the revenue.

c. The mode for all 100 shops is $153

Additionally I am a bit confused about the need for finding mode or AOV across all 100 shops. This makes sense only if all 100 shops fall under the same parent organization/owner. If these

are separate entities then we should calculate the above metrics specific to each shop data. I have done the same in the notebook (my_notebook.ipynb). A lot more insight can be derived from this approach.

Also I see that one order can constitute more than one pair of sneakers. Hence dividing the order amount by the total_items will get us the price of sneakers at each shop. We can further use this new feature in analysing the data and how tis influences the AOV per shop. But I don't think this lies within the scope of this challenge.

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?

54

```
SELECT COUNT(*) AS num_orders
FROM Orders o
LEFT JOIN Shippers s
        ON o.ShipperID = s.ShipperId
WHERE s.ShipperName = 'Speedy Express'
```

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

Peacock

```
SELECT LastName
FROM Employees
LEFT JOIN Orders
        ON Employees.EmployeeID = Orders.EmployeeID
GROUP BY Orders.EmployeeID
ORDER BY COUNT(*) DESC
```

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

Boston Crab Meat

```
SELECT ProductName
FROM Products
LEFT JOIN OrderDetails
ON Products.ProductID = OrderDetails.ProductID
LEFT JOIN Orders
ON OrderDetails.OrderID = Orders.OrderID
LEFT JOIN Customers
ON Orders.CustomerID = Customers.CustomerID
WHERE Country = 'Germany'
GROUP BY OrderDetails.ProductID
```

```
ORDER BY COUNT(*)*Quantity DESC
LIMIT 1
```