SHOPIFY SUMMER 2022 DATA SCIENCE INTERN CHALLENGE

Name: Venkata Satyanarayana Chivatam
Email: Venkata.chivatam@mail.mcgill.ca

Question 1:

Part A:

Think about what could be going wrong with our calculation. Think about a better way to evaluate this data.

According to the dataset given, there are two major outliers affecting the average order value.

i) Shop no 42 with order_amount of 704000 with same user_id(607).
ii) Shop no 78 with order_amount of 25725 per each pair of shoes

But, only shop no 42 looks suspicious as it has the same user_id purchasing every time. Shop no 78 could be a genuine shop with a really expensive shoe model. Besides that there are only 301 users buying products multiple times from different shops which is not an issue as they all are buying in a similar manner.

A better way to evaluate this data would be to remove all the data points of shop no 42 outlier and check for Median. Considering the median would also mitigate the shop 78 outlier affecting the metrics but will let the data points stay for further investigation if required.

Part B:

What metric would you report for this dataset?

As stated in the part A, I would report Median as a metric for this dataset instead of mean or mode because using mean would be vulnerable to shop no 42 outlier and if any other exists, and mode is usually used for categorical variables.

Part C:

What is its value?

Mean with outliers : 3145.13
Mean without outliers : 754.7
****Median with and without outliers : 284.5**

Please checkout the ipynb file for the corresponding code.

SHOPIFY SUMMER 2022 DATA SCIENCE INTERN CHALLENGE

Question 2:

Part A:

```
SELECT COUNT(*) FROM [Orders]
WHERE ShipperID = (SELECT ShipperID FROM [Shippers]
                                    WHERE ShipperName = "Speedy Express")
```

Count = 54

Part B:

```
SELECT LastName FROM [Employees]
WHERE EmployeeID = (SELECT EmployeeID FROM [orders]
                                    GROUP BY EmployeeID
                                    ORDER by Count(EmployeeID) DESC)
```

LastName = Peacock

Part C:

```
SELECT ProductName FROM Orders
JOIN Customers ON Orders.CustomerID = Customers.CustomerID
JOIN OrderDetails ON Orders.OrderID = OrderDetails.OrderID
JOIN Products ON OrderDetails.ProductID = Products.ProductID
WHERE Country = "Germany"
GROUP BY ProductName ORDER BY SUM(Quantity) DESC LIMIT 1;
```

ProductName = Boston Crab Meat