Shopify Data Science

Graydon Hope
linkedin.com/in/graydon-hope
905 975-4721
graydonho@gmail.com

Question 1: Sneaker Shop

a) What could be going wrong with the calculation? What is a better way to evaluate the data?

a. Problem(s)

- The order amounts are unevenly distributed. The largest 62 orders skew the average data severely. This is an imbalanced dataset.
- Concretely, most of the orders are $1760 or less.
- It is not a good idea to include smaller personal order statistics with larger business-related orders.

b. Possible Solution – Removing / Splitting up Outlier Orders

- We can split our calculation into two parts so that outliers are removed from the general case.
- The first part will exclude the outlier orders with values of $25,725.
- Additionally, we can include the higher value orders by dividing by the quantity associated with them.

b) What Metric to Report?

a) AOV = Sum(Order Amount / Total Items) / Number of Orders. Although this approach works much better, I would still like to remove outliers as well.

b) Value evaluated is **$387.74** which is a much more reasonable value. In my Python script I also calculated the value removing the outlier values and it is**:** **$151.07** which is an even more reasonable value for the common case.

Question 2: SQL Queries (All code runs but if you copy/paste, please format it like shown!).

a) **54 entries found**. Since I could not assume that OrderID was unique, I used the distinct keyword.

```
SELECT DISTINCT OrderID
FROM Orders
WHERE ShipperID = (
        SELECT ShipperID
        FROM Shippers
        WHERE Shippers.ShipperName = "Speedy Express"
    );
```

b) **Handel** is the last name with most orders

```
SELECT right(
        CustomerName,
        len(CustomerName) - CHARINDEX(' ', CustomerName)
    ) as LastName
FROM Customers
WHERE CustomerID IN(
        SELECT TOP 1 CustomerID
        FROM Orders
        GROUP BY CustomerID
        ORDER BY count(CustomerID) DESC
    );
```

c) **Product 40 (Boston Crab Meat)** was ordered the most by the German customers.

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

Analysis Used for Q1:

- Most expensive common single item: $201
- Most expensive outlier single item: $25,725
- Least expensive single item: $90
- Out of 5000 orders, 4938 of them are $1760 or under. Then there is a jump up to very large orders which include a range of values from $25,725 to $704,000
- We see that the **total** of all orders is: $15,725,640 and dividing by 5000, we find the **average** of: $3,145.128 as mentioned in the question.
- As previously stated, the range of common items is: [$90, $201]. However, there is an item available which is worth $25,725. We see this value come up when we look at some of the higher order values with lower quantities (ex: $71,175 / 3 = $25,725)