Zofia Zhengzheng Wang zofia.wang@mail.utoronto.ca

Fall 2021 Data Science Intern Challenge

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.

Answer: For Question 1, I used Jupyter Notebook for coding. The original average order value (AOV) provided above doesn't take into account the quantities of sneakers in different orders. The order value can be very high if the quantity is large, thus it is not a good metric to evaluate the dataset. Additionally, outliers in the prices of sneakers are not considered.

b. What metric would you report for this dataset?

Answer: The first alternative metric is average item value (AIV), which calculates the average price of sneakers in different shops, given each of these shops sells only one model of shoe. This metric makes more sense to evaluate the values of sneakers. Another alternative method is to remove records with abnormal sneaker prices before calculating the AOV. And here I used IQR method to detect outliers. According to the IQR, the prices of sneakers that are outside of the range [79, 223] would be considered outliers.

c. What is its value?

Answer:

The value of the first alternative metric AIV is 387.7428.

The value of the second alternative metric AOV after removing outliers is 300.1558.

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 Shippers.ShipperName, COUNT(*) AS order_number FROM Orders
JOIN Shippers
ON Orders.ShipperID=Shippers.ShipperID
WHERE ShipperName='Speedy Express'
```

ShipperName	order_number
Speedy Express	54

Answer: 54 orders were shipped by Speedy Express in total.

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

```
SELECT LastName, COUNT(Orders.OrderID) FROM Employees
JOIN Orders
ON Employees.EmployeeID=Orders.EmployeeID
GROUP BY Employees.EmployeeID
HAVING COUNT(*)= (SELECT max(order_number) FROM (SELECT COUNT(OrderID) AS
order_number FROM Orders GROUP BY EmployeeID) )
```

LastName	COUNT(Orders.OrderID)
Peacock	40

Answer: The last name of the employee with the most orders is Peacock.

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

```
SELECT ProductName, SUM(OrderDetails.Quantity) AS Total_Quantity,
COUNT(Orders.OrderID) AS Number_of_Orders FROM Products
JOIN OrderDetails ON OrderDetails.ProductID=Products.ProductID
JOIN Orders ON Orders.OrderID=OrderDetails.OrderID
JOIN Customers ON Customers.CustomerID=Orders.CustomerID
WHERE Customers.Country='Germany'
GROUP BY Products.ProductID
ORDER BY Total_Quantity DESC LIMIT 1
```

ProductName	Total_Quantity	Number_of_Orders
Boston Crab Meat	160	4

Answer: Since the maximum total quantity of the product ordered by customers in Germany is 160, the product which was ordered the most by customers in Germany is Boston Crab Meat.