Winter 2021 Data Science Intern Challenge

Question 1

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.

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

The number of items per order usually varies between 1 and 8 items, however, as shown in fig. 1, 17 out of 5000 orders had a total number of items equal to 2000. These orders are obviously unusual events that create skewness in the data and significantly influence the average order value. Consequently, the average is not representative of the data set. Alternatively, the outliers should be discarded or a different metric that gives more representative information about the data should be used.

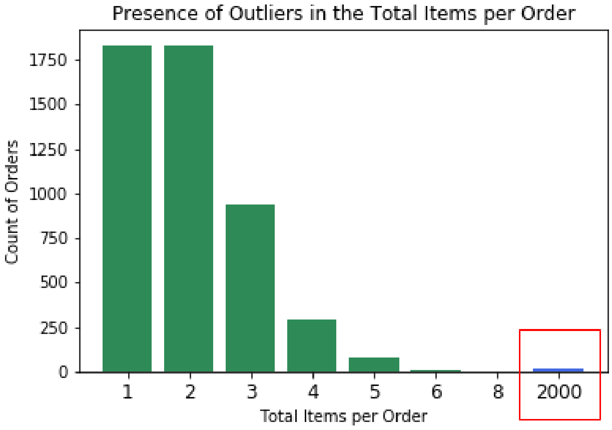

Figure 1: The presence of outliers in the number of ordered items causes the average metric to become irrelevant in terms of reflecting the usual events present in the data.

1.b What metric would you report for this dataset?

For this dataset, I would use the Median. The Median is determined by ranking the data from largest to smallest, and then identifying the value which is located in the middle. In that sense, unlike the mean, the median is robust against outliers and won't show sensitivity to unusual events.

Metric	Total Items	Order Amount
Mean	8.7872	$3145.128
Median	2	$284.0

Table 1: Mean Vs Median for the data analysis of the sneakers shops.

1.c What is its value?

As shown in table 1, the median shows a value of $284 for the order amount and the value of 2 for the total items per order. Knowing that these shops are selling sneakers, these values makes more sense than the ones displayed using the average.

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.

2.a How many orders were shipped by Speedy Express in total?

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

The result is: 54

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

```
SELECT LastName FROM Employees
WHERE EmployeeID=(SELECT EmployeeID FROM
                     (SELECT EmployeeID, COUNT(EmployeeID) AS TotalOrders FROM Orders
                     GROUP BY EmployeeID
                     ORDER BY TotalOrders DESC
                     LIMIT 1))
```

The last name is: Peacock

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

```
SELECT ProductID FROM (SELECT ProductID, SUM(Quantity) AS TotalQuantity FROM OrderDetails
                     JOIN (SELECT CustomerID AS CI1, OrderID AS OI1 FROM Orders
                          JOIN (SELECT CustomerID AS CI2 FROM Customers
                              WHERE Country='Germany')
                          ON CI1 = CI2)
                     ON OrderID = OI1
                     GROUP BY ProductID
                     ORDER BY TotalQuantity DESC
                     LIMIT 1)
```

The product ID is: 40