Shopify Summer 2022 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.

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

   Answer -

   The definition of the metric Average Order Value (AOV) is the average total of orders over a defined period. We are given the set of transactions over a 30-day period and while calculating we get an Average Order Value (AOV) of $3145.13. While this is correct when considering the actual definition and formula of average order value i.e., Total Revenue/Total Number of Orders getting a value of $3154.13 seems unreasonably high for an affordable item like sneakers.

   The one property of the dataset given to us which is not being considered while calculating the AOV is the quantity of sneakers that are being bought per order and bulk ordering. If we do a further analysis of the transactions on the data, we can observe that number of sneakers per order almost as high as 2000 sneakers for one order. To get a more effective evaluation of this dataset we should focus on the amount spent per sneaker as each sneaker store sells only one type of shoe.

   In dataanalysis.py I have loaded the csv file into a data frame using the pandas library.

   In this data frame I add a new row called per_sneaker_cost which calculates the price for one sneaker (order amount/total items) in each transaction.

   Now if we calculate the average order value (AOV) on the per_sneaker_cost we obtain a much more realistic value of $387.74 which represents the average order value for a single sneaker at the 100 shops.

2. What metric would you report for this dataset?

   Answer -

On further analysis of the dataset, I can use other measures of central tendency to get a better idea of the statistics behind the data. I calculate the mean, median and mode of the order amount field and obtain values of $3145.13, $284, and $153 respectively.

The mean of the dataset (same as Average Order Value (AOV)) is heavily skewed again because of the various outlier data points, however the median is a relatively reasonable value as it is the midpoint of the sorted order amount values. Therefore, we can consider median as an alternative measure to calculating Average Order Value (AOV).

An interesting metric that could have been considered is Revenue Per Visitor (RPV) which is how much revenue is generated each time a customer visits the sneaker store. (Total Revenue/Total Number of Unique Visitors). However, for this metric to be considered we would need more data on the number of visitors who visited the shop but did not purchase anything as only transaction data is not enough.

3. What is its value?

   Answer - The value of the median of the dataset is $284.0.

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.

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

   ```
   SELECT count(O.OrderID) as speedy_express_total
   FROM Orders O,Shippers S
   WHERE S.ShipperID=O.ShipperID
   AND S.ShipperName='Speedy Express';
   ```

   Answer – 54

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

   ```
   SELECT E.LastName,COUNT(O.OrderID) AS num_orders
   FROM Employees E,Orders O
   WHERE E.EmployeeID=O.EmployeeID
   GROUP BY E.LastName
   ORDER BY COUNT(O.OrderID) DESC;
   ```

   Answer – Peacock with 40 Orders.

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

```
SELECT SUM(OD.Quantity) as Quantity,P.ProductName
FROM OrderDetails OD,Orders O,Products P,Customers C
WHERE C.CustomerID=O.CustomerID AND OD.ProductID=P.ProductID AND
OD.OrderID=O.OrderID and C.country='Germany'
GROUP BY P.ProductName
ORDER BY SUM(OD.Quantity) DESC;
```

Answer – Boston Crab Meat with 160 Orders.