Fall 2021 Data Science Intern Challenge

Please complete the following questions, and provide your thought process/work. You can attach your work in a text file, link, etc. on the application page. Please ensure answers are easily visible for reviewers!

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.

   ANS: AOV of $3145.13 is high mean value for an affordable item like sneakers, it can be assumed that there are outliers in this dataset.
   Dataset analysis shows standard deviation of 41282.52 for column order_amount, min value of $90.00 and max of $704000 which clearly indicates presence of outliers.
   So, in such case mean will not represent the data accurately and won't be a good measure. Mean is a good estimate when data is normally distributed.

   In this case where the data has outliers and is skewed, median would be more appropriate for AOV.
   To evaluate, the outliers should be checked whether they're wrong or not.

   1. **Wrong values –** It should be rectified or the values can be dropped.
   2. **Right values –**
      - AOV values can be calculated for different cases like –
        i. Shops selling avg-priced shoes.
        ii. Shops selling high-priced shoes.
        iii. Shops selling high number of shoes.
      - Or the outliers can be capped towards the end of right tail of distribution.

b. What metric would you report for this dataset?

ANS: Most prominent points of analysis:

1. No. of sneakers purchased at a time is less than or equal to 8 except for 17 times when user_id 607 bought 2000 pairs of sneakers from shop_id 42 each having order_amount of $704000. This clearly is an outlier and should be investigated.

2. In shop_id 78, one pair of sneakers are sold for $25725.0 since each shop only sells one model of shoe. In this case order_amount is high due to expensive shoes.

   Median amount of price per shoe pair is $153.0.
   So, for this dataset, I would check if the outliers mentioned above are accurate or not.
   Even after removing extreme values, the graph for order_amount looks skewed. In such cases, **median** provides a better estimate than mean since it is less affected by outliers and skewed data.

c. What is its value?

ANS: The median value is $284.0

Kindly review the analysis done in Shopify_analysis.ipynb

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?

   ANS: SELECT COUNT(OrderID) FROM Orders
   WHERE ShipperID =
   (SELECT ShipperID FROM Shippers
   WHERE ShipperName = 'Speedy Express');

   Returns – **54**

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

ANS:

```
SELECT e.LastName, COUNT(*) AS order_count
FROM Employees e
JOIN Orders o ON e.EmployeeID = o.EmployeeID
GROUP BY e.EmployeeID
ORDER BY order_count DESC
LIMIT 1;
```

RETURNS – **Peacock**

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

ANS:

```
SELECT p.ProductName, SUM(d.Quantity) AS Quantity
FROM Products p JOIN OrderDetails d
   ON p.productID = d.productID
JOIN Orders o
   ON d.orderID = o.orderID
JOIN Customers c
   ON o.customerID = c.customerID
WHERE c.Country == 'Germany'
GROUP BY p.productName
ORDER BY Quantity DESC
LIMIT 1;
```

RETURNS – **ProductName Boston Crab Meat**