Summer 2022 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!

My answers to question 1 and question 2 are located below question 2. The code I wrote to determine my answer for question 1 is given in the file “Question 1 notebook” in this repository. Thanks!

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.
> b. What metric would you report for this dataset?
> c. What is its value?

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?
b. What is the last name of the employee with the most orders?
c. What product was ordered the most by customers in Germany?

Question 1 Answer:

a)

The reason the AOV is so high is because the mean price of the sneakers includes some significant outliers that skew the mean higher than it would be without them. Removing the amounts ordered from the table that have a Z score above 3 standard deviations will remove the outliers and cause the data to make more sense. After removing the outliers, the mean amount ordered in the dataset was $723.

To confirm that the outlier transactions were not due to reporting or human error, I calculated the price of each item in these transactions. The mean price per item in these large transactions was $352, indicating that it may have just been somebody buying in bulk to resell.

We can also just use the median amount ordered instead of the mean, as this is not as sensitive to outliers and makes more sense at a median amount ordered value of $284.

b)

I would report the median order amount, since the median order amount is not as sensitive to outliers as the mean order amount is.

c)

The median order value is $284 for these shops.

Question 2 Answer:

a)

```
SELECT count(OrderID) FROM Orders a
inner join shippers b
on a.ShipperID = b.ShipperID
Where ShipperName in ('Speedy Express')
```

Answer: 54

b)

```
SELECT LastName, MAX(NetOrders) FROM
(Select *, COUNT(DISTINCT OrderID) as NetOrders FROM
(SELECT a.OrderID, b.EmployeeID, b.LastName, b.FirstName FROM Orders a
Inner Join Employees b
ON a.EmployeeID = b.EmployeeID)
GROUP BY EmployeeID
ORDER BY COUNT(DISTINCT OrderID) DESC)
```

Answer: Peacock, 40 orders

c)

```
SELECT ProductName, SUM(Quantity) as TotalOrders
From Products a, Orders b, OrderDetails c, customers d
WHERE d.CustomerID=b.CustomerID AND d.Country = 'Germany' and b.OrderID=c.OrderID
AND c.ProductID=a.ProductID
GROUP BY ProductName
ORDER BY TotalOrders DESC
LIMIT 1
```

Answer: Boston Crab Meat, 160 orders