Summer 2022 Data Science Intern Challenge

Answers(highlighted in yellow)

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.

Issue is not taking the total_items into account, we need to first calculate the cost_per_item which is order_amount divided by total_items and then calculate the AOV

b. What metric would you report for this dataset?

Average of (order_amount divided by total_items)

c. What is its value?

387.7428

Python code -

```
df = pd.read_csv('Shopify Winter Data Science Intern Challenge Data Set.csv') #read csv
df.dropna(subset=['order_id'], inplace=True) #drop any nan values
print(df.tail())
print(df.nunique()) #all unique value counts in each of the columns
df["cost_per_item"] = df["order_amount"]/df["total_items"] #calculate the cost per item which is order_amount divided by total_items
print(df['cost_per_item'].mean()) #average order 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?

```
SELECT COUNT(*) FROM [Orders] as A, [Shippers] as B WHERE A.ShipperID =
B.ShipperID AND B.ShipperName = 'Speedy Express'
```

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

```
SELECT A.LastName from [Employees] AS A,
(SELECT A.EmployeeID, COUNT(*) FROM [Orders] AS A, [Employees] AS B where
A.EmployeeID = B.EmployeeID
GROUP BY A.EmployeeID
ORDER BY COUNT(*) DESC
LIMIT 1) AS B WHERE A.EmployeeID = B.EmployeeID
//I can achieve the same using RANK window function as well
```

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

```
SELECT ProductName FROM Products AS A,(
SELECT C.ProductID, count(C.OrderID) FROM [Customers] AS A, [Orders] AS B,
[OrderDetails] AS C where A.CustomerID = B.CustomerID AND A.Country = 'Germany'
AND B.OrderID = C.OrderID
GROUP BY (C.ProductID)
ORDER BY count(C.OrderID) DESC
LIMIT 1) AS B WHERE A.ProductID = B.ProductID
```