Fall 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!

Candidate Name: Jahnvi Sikligar

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.
As we know that Average order value(AOV) is a critical metric that allows businesses to keep track of whether they want to scale their profits and revenue growth. The mathematical formulation to compute AOV is as follows:

$$\frac{Total\ Revenue}{\text{Number of Orders}}$$

In this case, Total revenue = 'order_amount' and Number of orders = 'total_items' But when checked with the data using df.describe() method it can be seen that $3145.13 is wrongly rounded off AOV. The wrongly identified value is computation of average of values of column – 'total_items' . The correct way to compute it in context to this dataset is as follows:

$$\frac{Sum\ of\ 'order\_amount'}{Sum\ of\ 'total\_items'}$$

Accordingly the calculated new AOV is as follows: ***$357.92***. This new computed value is not completely correct as the data still contains outliers.

Working with Outliers:

As it can be seen from the graph below that the data is skewed distribution. Therefore, we will be using IQR(Inter quartile range) to remove outliers from the data to get a normal distribution graph. The new Average order value is: ***$150.61.***

b. What metric would you report for this dataset?
The correct metric for this dataset would be the median of the 'order_amount' column.

c. What is its value?
The value of median is ***$284.00*** which has been computed with outliers in the dataset. But after the removal of outliers from dataset, the median value changes to ***$280.00***.

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 *
FROM Orders
LEFT JOIN Shippers
ON Orders.ShipperID = Shippers.ShipperID
WHERE ShipperName = "Speedy Express"
```

Output:

Number of Records: 54

OrderID	CustomerID	EmployeeID	OrderDate	ShipperID	ShipperName	Phone
10249	81	6	1996-07-05	1	Speedy Express	(503) 555-9831
10251	84	3	1996-07-08	1	Speedy Express	(503) 555-9831
10258	20	1	1996-07-17	1	Speedy Express	(503) 555-9831
10260	55	4	1996-07-19	1	Speedy Express	(503) 555-9831
10265	7	2	1996-07-25	1	Speedy Express	(503) 555-9831
10267	25	4	1996-07-29	1	Speedy Express	(503) 555-9831
10269	89	5	1996-07-31	1	Speedy Express	(503) 555-9831
10270	87	1	1996-08-01	1	Speedy Express	(503) 555-9831
10274	85	6	1996-08-06	1	Speedy Express	(503) 555-9831
10275	49	1	1996-08-07	1	Speedy Express	(503) 555-9831

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

```
SELECT Employees.LastName, Count(Orders.OrderID) As NumberOfOrders From
Orders
INNER Join Employees ON Orders.EmployeeID = Employees.EmployeeID
Group by LastName Order by NumberOfOrders ASC
Limit 10;
```

Output:

Result:

Number of Records: 9

LastName	NumberOfOrders
Dodsworth	6
Buchanan	11
King	14
Suyama	18
Fuller	20
Callahan	27
Davolio	29
Leverling	31
Peacock	40

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

```
SELECT p.ProductName, SUM(Quantity) AS TotalQuantity
FROM Orders AS o, OrderDetails AS od, Customers AS c, Products AS p
WHERE c.Country = "Germany" AND od.OrderID = o.OrderID AND od.ProductID =
p.ProductID AND c.CustomerID = o.CustomerID
GROUP BY p.ProductID
ORDER BY TotalQuantity DESC
LIMIT 10;
```

Output:

Result:

Number of Records: 10

ProductName	TotalQuantity
Boston Crab Meat	160
Gorgonzola Telino	125
Tunnbröd	105
Steeleye Stout	100
Teatime Chocolate Biscuits	95
Mozzarella di Giovanni	86
Chang	84
Lakkalikööri	75
Fløtemysost	75
Camembert Pierrot	75