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!

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.
    - Upon analysis of the dataset, it appears that are a few outliers that bring up the average. Some betters ways would be to use median instead of the mean, remove outliers from the dataset and calculate the mean of of the outliers and non outliers or normalize each order value of each order by the total items then taking the mean of this.

b. What metric would you report for this dataset?
    - I think the best metric to report is the median as it is the simplest and quickest to report. However, the mean average of outliers might also be important to investigate if there were errors in collecting this data.

c. What is its value?
    - The median of all orders was 284.00
    - The outliers were defined as any order with an amount over 1000. Their mean was 200587.24 and the non outliers mean was 301.06
    - The mean of the value per total items ordered was 387.74

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?

- **Input**:
  ```
  SELECT COUNT(*) FROM Orders
  WHERE Orders.ShipperID = (SELECT ShipperID FROM Shippers WHERE
  ShipperName = 'Speedy Express');
  ```
- **Output**: 54

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

- **Input**:
  ```
  Select Orders.EmployeeID, LastName, COUNT(Orders.EmployeeID)
  From Orders
  INNER JOIN Employees ON Orders.EmployeeID = Employees.EmployeeID
  GROUP BY Orders.EmployeeID
  ORDER BY COUNT(Orders.EmployeeID) DESC;
  ```
- **Output**: Peacock

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

- **Input:**
  ```
  SELECT Orders.OrderID, OrderDetails.ProductID, ProductName,
  SUM(Quantity) FROM Customers
  INNER JOIN Orders ON Customers.CustomerID = Orders.CustomerID
  INNER JOIN OrderDetails ON Orders.OrderID = OrderDetails.OrderID
  INNER JOIN Products ON OrderDetails.ProductID = Products.ProductID
  WHERE Customers.Country = 'Germany'
  GROUP BY ProductName
  ORDER BY SUM(Quantity) DESC;
  ```
- **Output**: Boston Crab Meat