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.

   There are 5000 unique orders `(=COUNT(UNIQUE(A2:A5001)))` and the total of order amount is $15,725,640.00 (`=SUM(D2:D5001)`). To calculate average order cost is dividing the total order amount divided by total number of orders. And in fact, it gives the value above of $3145.13.

   A better evaluation would be checking the summary of the columns where it shows the min and the max value of each column. By using the MIN and the MAX functions of the spreadsheet, one could see that the max of the order amount is $704,000.00 which looks abnormal.

   By sorting the order_amount column, we would see some figures repeated such as 704000, 77175, 51450 and 25725. There are a total of 63 orders where the total amount looks abnormally higher than the rest of the orders. We may narrow down the orders to shops 42 and 78. These orders need further investigation perhaps there may be designer pieces sold for rather higher prices or there may be wholesale customers.

b. What metric would you report for this dataset?
I would remove the total 63 orders where the figures need more investigation and would sum up the rest of the orders and divide them into the remainder number of orders.

c. What is its value?
$302.58

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] O
      LEFT JOIN Shippers S ON O.ShipperID = S.ShipperID
WHERE ShipperName = "Speedy Express"

OR -
SELECT COUNT(*)
FROM [Orders]
WHERE ShipperID = 1
```

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

```
SELECT MAX(C.mostorders), C.LastName, C.FirstName
      FROM (
            SELECT COUNT(*) as mostOrders, O.EmployeeID,
            E.LastName, E.FirstName
            FROM [Orders] O
                LEFT JOIN Employees E ON O.EmployeeID =
            E.EmployeeID
            GROUP BY O.EmployeeID
            ) C
```

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

```
SELECT ProductName, Country, MAX(totalOrdered) as totalOrdered
FROM (
      SELECT OD.OrderID, C.Country, P.ProductName,
      SUM(OD.Quantity) AS totalOrdered
      FROM [Orders] O
            LEFT JOIN Customers C ON O.CustomerID = C.CustomerID
            LEFT JOIN OrderDetails OD ON O.OrderID = OD.OrderID
            LEFT JOIN Products P ON OD.ProductID = P.ProductID
      WHERE C.Country = 'Germany'
      GROUP BY P.ProductName
      )
```