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.

Program to run: Question1.py
Github repo to clone: https://github.com/ruiding-code/Shopify-2022-Data-Science-Challenge.git

a. Think about what could be going wrong with our calculation. Think about a better way to evaluate this data.
    i. **We seem to have some outlier/invalid values that need to be dealt with. The data could also be skewed by huge values, since the average is abnormally high. Question1.py contains my analysis on this dataset.**
b. What metric would you report for this dataset?
    i. **I chose to report the median since the dataset is skewed (after removing invalid values).**
c. What is its value?
    i. **284$ per order!**

Question 2 on the next page.

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?
    a. **Answer: 54**
    b. **Query: SELECT COUNT(*) FROM Orders JOIN Shippers ON (Orders.ShipperID = Shippers.ShipperID) WHERE Shippers.ShipperName = "Speedy Express";**

b. What is the last name of the employee with the most orders?
    a. **Answer: 40**
    b. **Query: SELECT Employees.LastName, COUNT(Employees.LastName) as "Occurence" FROM Orders JOIN Employees ON (Orders.EmployeeID = Employees.EmployeeID) GROUP BY Employees.LastName ORDER BY "Occurence" DESC LIMIT 1;**

c. What product was ordered the most by customers in Germany?
    a. For this question, if we want the product that was ordered the most in separate orders:
        i. **Answer: Gorgonzola Telino (ProductID: 31)**
        ii. **Query:**

```
SELECT Products.ProductName, Products.ProductID,
COUNT(ProductName) as "Occurence"
FROM OrderDetails
JOIN Products ON (Products.ProductID = OrderDetails.ProductID)
JOIN Orders ON (Orders.OrderID = OrderDetails.OrderID)
JOIN Customers on (Customers.CustomerID = Orders.CustomerID)
WHERE Customers.Country = "Germany"
GROUP BY Products.ProductName
ORDER BY "Occurence" DESC
LIMIT 1;
```

    b. If it is the product that was ordered in the highest quantity:
        i. **Answer: Boston Crab Meat (ProductID: 40)**
        ii. **Query:**

```
SELECT Products.ProductName, Products.ProductID,
SUM(OrderDetails.Quantity) as "TotalQuantityOrdered"
FROM OrderDetails
JOIN Products ON (Products.ProductID = OrderDetails.ProductID)
JOIN Orders ON (Orders.OrderID = OrderDetails.OrderID)
JOIN Customers on (Customers.CustomerID = Orders.CustomerID)
WHERE Customers.Country = "Germany"
GROUP BY OrderDetails.Quantity
ORDER BY "TotalQuantityOrdered" DESC
LIMIT 1;
```