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.

OUTLIERS: Checking all the data description and scatterplot, will find out that there should be outliers from a few shops, having much higher order amount than others, which would be the reason driving up the AOV.
So, we should do the AOV separately, the reasonable AOV should exclude the shops that have a much higher order amount.
For the shop(s) having special situations, we should see how to deal with it once we have more detailed information.

b. What metric would you report for this dataset?

I'd report the average order amount and average total items from each shop, with the numbers and figures to help explain. I'd also check the price for one pair of sneakers, the average and median values from these three.
Therefore, the reasonable AOV should be: AOV(G) = order amount / order count (without outlier)

c. What is its value?

General AOV(G) should be: $300.16
For those outliers, it's also good to know the AOV:
Higher item amount for each order AOV(H_Item): $235101.49 (Shop_id 42)
Higher price for each shop AOV(H_P): $49213.04 (Shop_id 78)

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(O.OrderID) FROM Orders O
JOIN Shippers S
ON O.ShipperID = S.ShipperID
WHERE S.ShipperName = 'Speedy Express'
```

A: 54 orders

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

```
SELECT E.LastName, COUNT(O.OrderID) AS Order_Num
FROM Orders O
LEFT JOIN Employees E
ON O.EmployeeID = E.EmployeeID
GROUP BY E.LastName
ORDER BY Order_Num DESC
LIMIT 1
```

A: Peacock (with 40 orders)

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

```
SELECT SUM(OD.Quantity) AS Product_Num,  P.ProductName, C.Country
FROM OrderDetails OD
JOIN Orders O, Products P
ON O.OrderID = OD.OrderID and P.ProductID = OD.ProductID
JOIN Customers C
ON C.CustomerID = O.CustomerID
WHERE C.Country = 'Germany'
GROUP BY P.ProductName
ORDER BY Product_Num DESC
LIMIT 1
```

A: Boston Crab Meat (with 160 orders)