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!

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.

b. What metric would you report for this dataset?

c. What is its value?
This is analyzed with both excel (2019 Winter Data Science Intern Challenge Data Set.xlsx) and python(Question_1.ipynb).

a. There are some outliers on Average item price ($ 25725) from Store 78 and total_items (2000) from Store 42. After removing these outliers, the average order value is $ 302.58 (*shown in Delete outlier on order amount*). The normal item prices are lower than $ 500, while the normal total_items for other stores are between 1 to 8. Median of all the orders will be a better way to evaluate the data. It will be $284.

b. For this dataset, I would report the median of order_amount, total_items, and Average price. If we want to digger deep, the following metrics will be reported:
    i. Check if there are any difference between different payment methods.
    ii. Check which shop is most ordered and earns the most revenue
    iii. Check which user places the most order and spends the most money

c. The value of the median value of all orders is $ 284.

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?
b. What is the last name of the employee with the most orders?
c. What product was ordered the most by customers in Germany?
This can be answer in Question 2.sql