Winter 2021 Data Science Intern Challenge

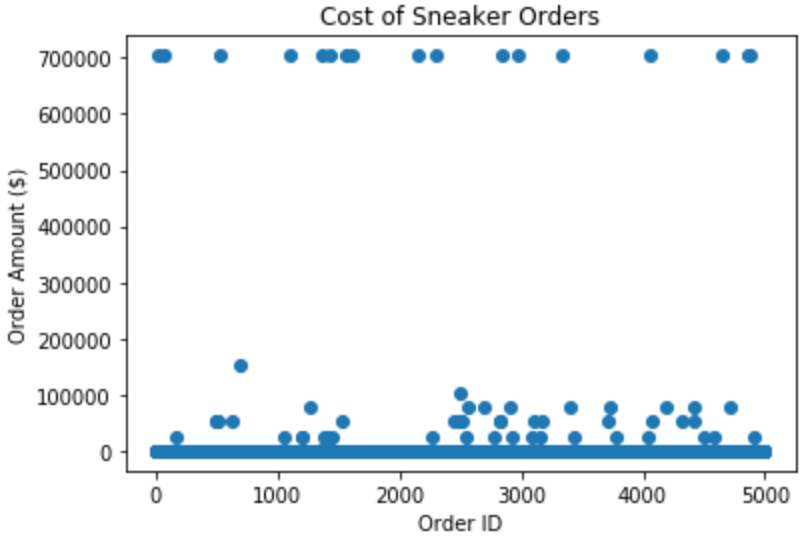
Please complete the following questions, and provide your thought process/work. You can attach your work in a text file, link, etc. on the application page. Please ensure answers are easily visible for reviewers!

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.

From an empirical examination of the dataset, it appears that the average is getting skewed by several large outliers- in this scenario, these are likely bulk orders and expensive designer shoes.



b. What metric would you report for this dataset?

Looks like there are a bunch of orders at the $700K level that are skewing the average. But what would be a better indicator of the AOV than the raw average? Three options that I will calculate are the Root Mean Square (RMS), the median, and the average with all orders > $700K removed.

The RMS provides a way to use all of the data, but outliers are weighted differently because the RMS takes the root of the sum of squares of the data points. The downside is that the outliers are so far away from the rest of the data that they could still skew the RMS. On the other hand, the median should end up very close to the majority of the orders. However, it entirely ignores the overall shape of the data. Lastly, the average with all orders > $700K removed will be more accurate to the AOV, but still includes the outliers around ~$50K and purposely throws away data, which is bad practice. We will use all three of these metrics to get a holistic view of the AOV.

c. What is its value?

RMS: 3,535.89

Median: 284.00

Average with All Orders > $700K Removed: 754.09

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(OrderID) FROM [Orders]
```

Answer: 196

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

```
SELECT Employees.LastName
FROM Orders
JOIN Employees on Employees.EmployeeID=Orders.EmployeeID
GROUP BY Orders.EmployeeID
ORDER BY COUNT(Orders.EmployeeID) DESC
LIMIT 1
```

Answer: Peacock

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

```
SELECT Products.ProductName
FROM Customers
LEFT JOIN Orders on Orders.CustomerID = Customers.CustomerID
LEFT JOIN OrderDetails on OrderDetails.OrderID = Orders.OrderID
LEFT JOIN Products on OrderDetails.ProductID = Products.ProductID
WHERE Country = 'Germany'
GROUP BY Products.ProductID
ORDER BY COUNT(OrderDetails.ProductID) DESC
LIMIT 1
```

Answer: Gorgonzola Telino