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.

Given such a high mean value, we can assume that there are outliers present in our data set.
Therefore, a better metric would be Median Order Volume.
However, in the work below I demonstrate that this is not the best approach. The better way to evaluate the data is to recalculate the Average Order Volume after removing the large orders, and removing transactions involving store 78 (given that store 78 is charging $25,725 per pair of shoes).

b. What metric would you report for this dataset?

The metric I use is Average Small Order Volume, after removing the over priced orders from store 78.
(In case of ATS: Median Order Volume)

c. What is its value?

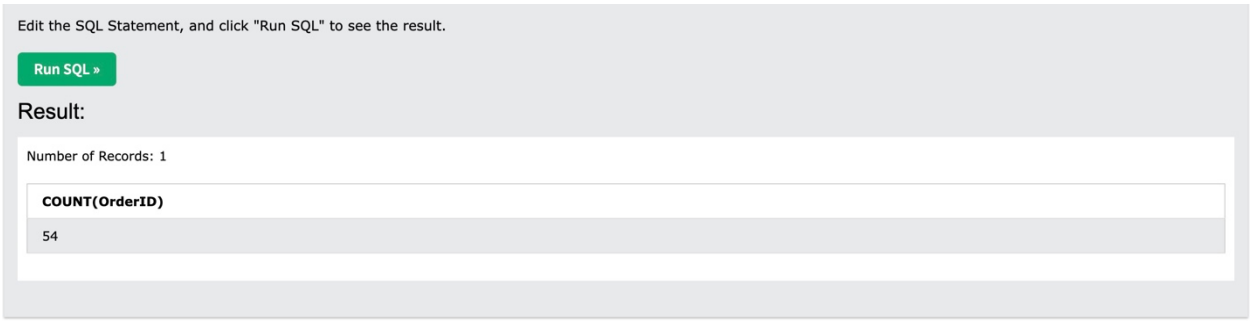
302.58
(In case of ATS: 284)

Here is the code - https://github.com/Rijul25/Shopify

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.

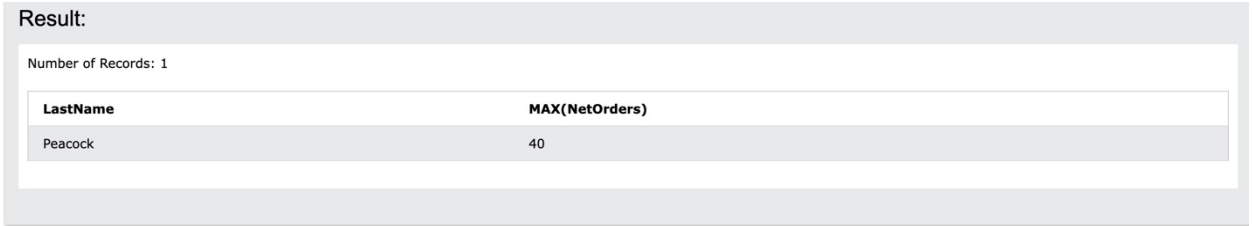
a. How many orders were shipped by Speedy Express in total?

```
SELECT COUNT(OrderID) FROM Orders
WHERE ShipperID =
(SELECT ShipperID FROM Shippers
WHERE ShipperName = 'Speedy Express')
```



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

```
SELECT LastName, MAX(NetOrders) FROM
(Select *, COUNT(DISTINCT OrderID) as NetOrders FROM
(SELECT o.OrderID, e.EmployeeID, e.LastName, e.FirstName
FROM Orders o Inner Join Employees e
ON o.EmployeeID = e.EmployeeID)
GROUP BY EmployeeID
ORDER BY COUNT(DISTINCT OrderID) DESC)
```



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

```
SELECT ProductName FROM Products
WHERE ProductID IN (SELECT TOP 1 ProductID FROM OrderDetails
WHERE OrderID IN (
SELECT OrderID FROM Orders
```

```
WHERE CustomerID IN (SELECT CustomerID FROM Customers
WHERE Country = 'Germany'))
GROUP BY ProductID
ORDER BY COUNT(OrderDetailID) DESC);
```

Returns "Gorgonzola Telino"