Summer 2022 Data Science Intern Challenge

Name: Vaibhav Chugh

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!

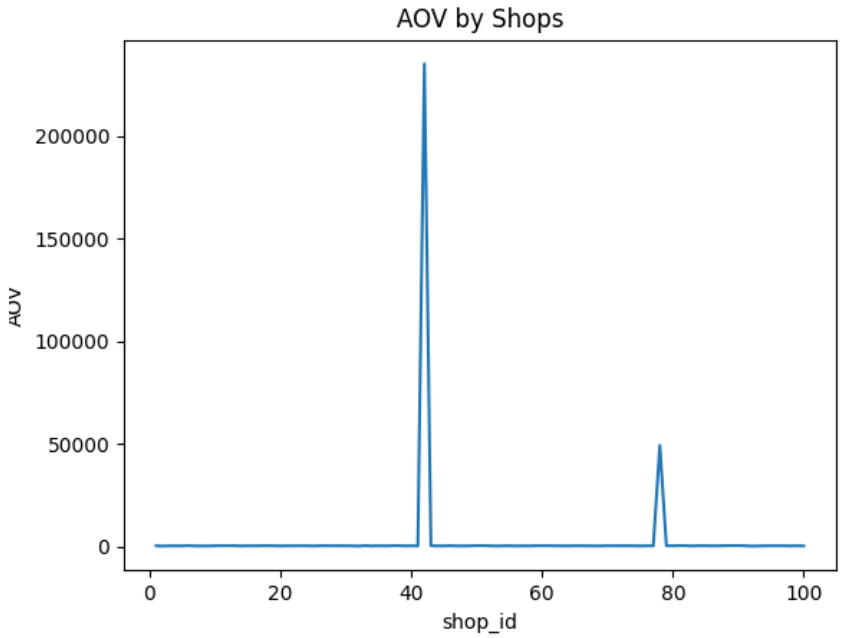
Question 1: Given some sample data, write a program to answer the following: click here to access the required data set

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.

- **Think about what could be going wrong with our calculation. Think about a better way to evaluate this data.**

  AOV is calculated as revenue / total_no_of_orders. The value given in the question is naively calculated by summing up all the order_amounts, dividing by the total number of rows. After doing some data processing and exploratory data analysis (code for which is attached as exp_data_anal.py), I calculated the AOV for each shop separately.

  Graph:



As we can see, there is a huge bump for one shop and a small bump for another shop and the rest of the AOV values are closely related.

Diving deeper into the dataset, these bumps belong to shops 42 and 78. After seeing the data, the order_amounts are really high due to huge orders being placed by a particular user for shop 42 and certain users for shop 78, but the number of orders are comparatively less, hence the huge AOV values.
So if we calculate the AOV for all the shops collectively, these values mentioned above act as outliers, and as AOV is a mean value, we already know means are affected by outliers. This makes sense on a logical level as well, if 98 shops have similar AOVs but the rest 2 have huge AOVs, the total AOV is not going to be representative of all the 100 shops.

Since the 2 shops have exponentially higher AOV values, we can calculate the new AOV by removing these outliers.

➢ **What metric would you report for this dataset?**

I would report the AOV value by removing an exponentially high outlier i.e. by not taking into consideration the sale for shop IDs 42 and 78.

Another way is by reporting the median instead of calculating the mean, since median is not sensitive to outliers but mean is.

- **What is its value?**

  New AOV value after removing shops 42 and 78 = 300.16

  Median value = 284.0

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.

- **How many orders were shipped by Speedy Express in total?**

  Query:

```
SELECT count(*)
FROM Orders
WHERE ShipperID=1;
```

  Answer:
  54

- **What is the last name of the employee with the most orders?**

  Query:

```
SELECT A.EID, max(A.TotalEmpOrders) as TotalOrders, E.LastName
FROM
(       SELECT EmployeeID as EID, count(OrderID) as TotalEmpOrders
        FROM Orders GROUP BY EmployeeID
) as A, Employees E
WHERE E.EmployeeID=A.EID;
```

  Output:

EID	TotalOrders	LastName
4	40	Peacock

  Answer: Peacock

➢ **What product was ordered the most by customers in Germany?**

Query:

```
SELECT ProductName, max(TotalOrders) as Orders
FROM
(SELECT Products.ProductName, count(*) as TotalOrders FROM Orders INNER JOIN
Customers ON Orders.CustomerID = Customers.CustomerID INNER JOIN OrderDetails
ON Orders.OrderID = OrderDetails.OrderID INNER JOIN Products ON
OrderDetails.ProductID = Products.ProductID WHERE Customers.Country = 'Germany'
GROUP BY Products.ProductName
)
```

Output:

ProductName	Orders
Gorgonzola Telino	5

Answer:
Gorgonzola Telino