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.

Since this data is not normally distributed , a simple mean will not represent the data fairly as we see two stores that skew the results. a better approach would be calculating the median of shoe prices since each store is selling a single model of shoe.

b. What metric would you report for this dataset?

The best metric in my opinion is user percentiles applied to the single shoe prices for each model. the table below shows the median and 75th.

	Order Amount	Shoe Price
mean	3145.12	387.74
25%	163.00	133.00
50%	**284.00**	**153.00**
75%	**390.00**	**169.00**

c. What is its value?

153.00 $ Which is the median of the shoe costs.

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.

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

Since we can see the SpeedyExpress is shipperID 1 we can write the following Query to get the total number of orders shipped by Speedy Express.

```
SELECT
  Count(OrderID) as Total_Orders_SpeedyExpress
FROM
  Orders
WHERE
  ShipperID = 1;
```

or if the shipperID may change at a later time we can use

```
SELECT
  Count(OrderID) as Total_Orders_SpeedyExpress
FROM
  Orders as o
  JOIN Shippers as s ON o.ShipperID = s.ShipperID
WHERE
  s.ShipperName = 'Speedy Express';
```

Answer : Peacock , with Total Orders of 40

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

The orders table contains the employee ID which can be linked to the employees table to get the last name of the employee. and a group and order by function to filter and sort to our desired result. we limit to 1 combined with a Descending order to get the top result.

```
SELECT
  e.LastName as LastName,
  Count(OrderID) as Total_Orders
FROM
  Orders as o
  JOIN Employees as e ON o.EmployeeID = e.EmployeeID
GROUP BY
  e.LastName
ORDER BY
  Count(OrderID) DESC
LIMIT
  1;
```

Answer : Peacock , with Total Orders of 40

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

Here we can user the same orders table and the customers filtered by country to get the result. But to get the product info we need two more tables to join on , the products table and the order details table.

```
SELECT
  p.ProductName as ProductName,
  Count(*) as Total_Orders
FROM
  Orders as o
  JOIN Customers as c ON o.CustomerID = c.CustomerID
  JOIN OrderDetails as od ON o.OrderID = od.OrderID
  JOIN Products as p ON od.ProductID = p.ProductID
WHERE
  c.Country = 'Germany'
GROUP BY
  p.ProductName
ORDER BY
  Count(*) DESC
LIMIT
  1;
```

Answer : Gorgonzola Telino , with Total Orders of 5