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?

(a) There is a big problem that there is no cleaning in this problem. The problem states that this is an affordable item, a sneaker, then AOV cannot be around 3000 unless every customer ordered around 10 sneakers which are not possible in general.
After observation, there are many orders which sold one shoe for 25725 dollars which are unreasonable in this case. Remove those rows to get a better estimation.
There are some orders which ordered 2000 items at once, they are likely to be sale store or salesman but not individual, they are not targets of analysis so remove those orders too.

(b) The metric used is still AOV. AOV is calculated by (Revenue of all orders / number of orders). Besides, as there is only one type of item, Average Selling Price(ASP) is also a good metric in this case. The formula for ASP is (Revenue / number of sold items), ASP would give direct information about how this model of sneaker sold

(c) After data cleaning, the AOV is 302.6, and the ASP is 151.7

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?

Q2.a query and result

```
SELECT "Speedy Express" as name, COUNT(distinct OrderID) total_order
FROM Orders o join Shippers s on o.ShipperID = s.ShipperID
WHERE ShipperName = "Speedy Express"
```

Number of Records: 1

name	total_order
Speedy Express	54

Q2.b query and result

```
SELECT LastName, count(distinct OrderID) as total_order
FROM Employees e join Orders o on e.EmployeeID = o.EmployeeID
GROUP BY e.EmployeeID
ORDER BY count(distinct OrderID) DESC
LIMIT 1
```

Number of Records: 1

LastName	total_order
Peacock	40

Q2.c query and result

SQL Statement:

```
SELECT ProductName, count(distinct o.OrderID) as total_order
FROM Orders o join OrderDetails od on o.OrderID = od.OrderID join Customers c on c.CustomerID =
o.CustomerID join Products p on p.ProductID = od.ProductID
WHERE c.Country = "Germany"
GROUP BY p.ProductID
ORDER BY count(distinct o.OrderID) DESC
```

Edit the SQL Statement, and click "Run SQL" to see the result.

Run SQL »

Result:

Number of Records: 1

ProductName	total_order
Gorgonzola Telino	5