Summer 2022 Data Science Intern Challenge

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.

To diagnose the problem, I conducted the following analysis.

- **Exploratory Data Analysis**:

  As the max value (704,000) of order amount is much larger than its median (284), and the standard deviation (41,282.5) of order amount is super large, I suspect that there may exist extreme (high) values in the order values because average order value can be easily affected by those extreme (high) values. To figure out the reason behind, I conducted deep-dive analysis from 4 perspectives.

- **Deep-dive Analysis**:

  - **Check if any anomalies exist in the order amount by plotting the box plot.**
    In the 5,000 data, there are 141 data associated with an order amount larger than Q3+1.5IQR, which are outliers according to the boxplot.

  - **Breakdown the target metric into component metrics.**
    As the target metric, AOV, can be decomposed into total_items and price_per_item, we bring in the component metric into analysis to check whether there are anomalies in the them that leads to anomalies in AOV. And I found ‘total_items = 2000’ and ‘price_per_item= 25725’ are anomalous data points. By removing data points related to the two anomalous data points, the value of AOV decreased from 3145.13 to 302.58, which is much normal.

  - **Segment AOV by shop, user and payment method.**
    By, segmenting AOV by shop, I found the order records NO.42 and NO.78 shop are anomalous. Particularly, the price of sneaker models sold at NO.42 shop is normal, but NO.607 user purchased 17 times at NO.42 shop and each time the NO.607 user

purchased 2000 pairs of sneakers which in total is 704,000 order amount. In addition, the price of sneakers sold at NO.78 shop is 25,725 per pair, which is very expensive, so that every order amount at NO.78 shop is over 25,725. And there is no obvious anomaly with regards to particular payment method.

The above anomalous data of NO.607 user at NO.42 shop and all orders at NO.78 shop lead to a high value of AOV. If we removing the above anomalous data, the AOV will drop from 3145.13 to 302.58.

- **Check temporal factor by plotting the time series trend of AOV.**
  There is no evidence of progressive increase of AOV, but we can say there are sudden increase in AOV because once the NO.607 user purchased at NO.42 shop or any order is made at NO.78 shop the AOV will increase at that day.

  Worth to mention, there are two different order but all the information of the order are the same including the timestamp of purchase. Therefore, I suspect that there may be some manual mistake or technical issue cased mistake in this dataset.

```
[51] data[(data['order_id']==2970)|(data['order_id']==2836)]
```

	order_id	shop_id	user_id	order_amount	total_items	payment_method	created_at	price_per_item
2835	2836	42	607	704000	2000	credit_card	2017-03-28 04:00:00	352.0
2969	2970	42	607	704000	2000	credit_card	2017-03-28 04:00:00	352.0

b. What metric would you report for this dataset?

a) **Median Order Value:**
Compared with the averaged order value, the median order value is a more robust metric and will not easily affected by outliers.

b) **Revised AOV (after removing outliers):**
If we remove the extreme values of order amount made by NO.607 user at NO.42 shop and all orders made at NO.78 shop, we can get a revised AOV.

Personally, I prefer the first one, median order value to report for this dataset.

c. What is its value?

The value of median order value for this dataset is 284.

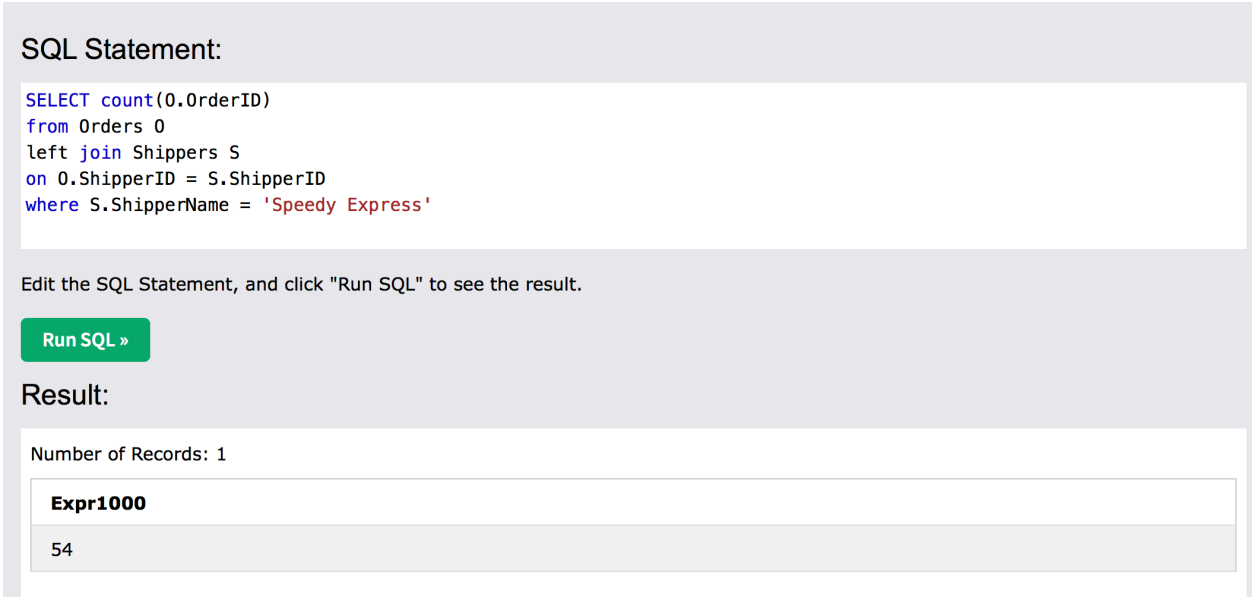
Question 2: For this question you'll need to use SQL. Follow this link to access the data set required for the challenge. Please use queries to answer the following questions. Paste your queries along with your final numerical answers below.

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

SQL code:

```
select count(O.OrderID)
from Orders O
left join Shippers S
on O.ShipperID = S.ShipperID
where S.ShipperName = 'Speedy Express'
```

Snapshot:



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

SQL code:

```
select E.LastName
from Employees E
right join (select top 1 EmployeeID from (select EmployeeID, count(OrderID) as cnt from
Orders group by EmployeeID) order by cnt desc) O
on E. EmployeeID = O. EmployeeID
```

Snapshot:

SQL Statement:

```
select E.LastName
from Employees E
right join (select top 1 EmployeeID from (select EmployeeID, count(OrderID) as cnt from Orders group by
EmployeeID)order by cnt desc) O
on E.EmployeeID = O.EmployeeID
```

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

Run SQL »

Result:

Number of Records: 1

LastName
Peacock

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

SQL Code:

```
select P.ProductName,TMP.ProductID, TMP.cnt
from Products P
join
(SELECT OD.ProductID, sum(OD.Quantity) as cnt
FROM Orders O
left join OrderDetails OD
on O.OrderID = OD.OrderID
where O.CustomerID in (SELECT CustomerID FROM [Customers] where Country = 'Germany')
group by OD.ProductID
order by sum(OD.Quantity) desc limit 1) TMP
on P.ProductID = TMP.ProductID
```

SQL Statement:

```
select P.ProductName,TMP.ProductID, TMP.cnt
from Products P
join
(SELECT OD.ProductID, sum(OD.Quantity) as cnt
FROM Orders O
left join OrderDetails OD
```

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

Run SQL »

Result:

Number of Records: 1

ProductName	ProductID	cnt
Boston Crab Meat	40	160