Shopify
Winter 2021 Data Science Intern Challenge

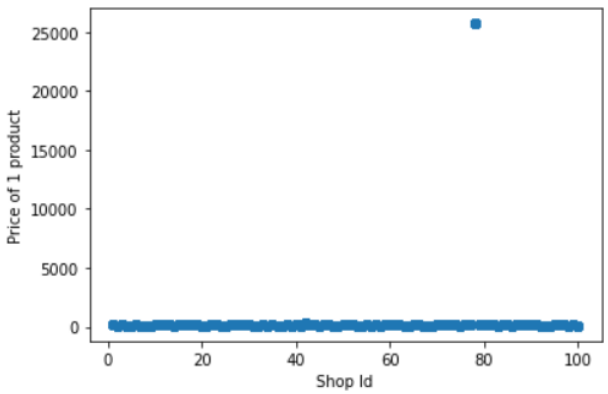
Question 1: Given some sample data, write a program to answer the following: click here to access the required data set

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.
Solution:
Initially,only the order amount column is considered for calculation of AOV which is wrong as each order amount is for different quantity of items which are provided in the total_items column. So to get the correct value of AOV the correct approach will be to consider both the order_amount and total_items columns and perform the calculations as shown in the avg_per_product function in the code.
Initial AOV calculated:
AOV_naive = sum(order_amount)/Total number of data points
Corrected AOV:
AOV = sum(order_amount)/sum(total_items)

```
Naive AOV = 3145.128
Correct AOV = 357.92152221412965
```

b. What metric would you report for this dataset?
Solution:



The scatter plot represents the cost of each product by the shop under a single order id and it can be clearly seen that there is presence of outliers which hinder the calculation of the Average Order Value and make it biased towards value of outliers. So, we can conclude that even if we use the new formula for calculation the AOV the AOV calculated will be wrong and biased towards the outliers. The best metric to use in such a situation will be **median** because median is less effected by the outliers than the mean(earlier AOV).

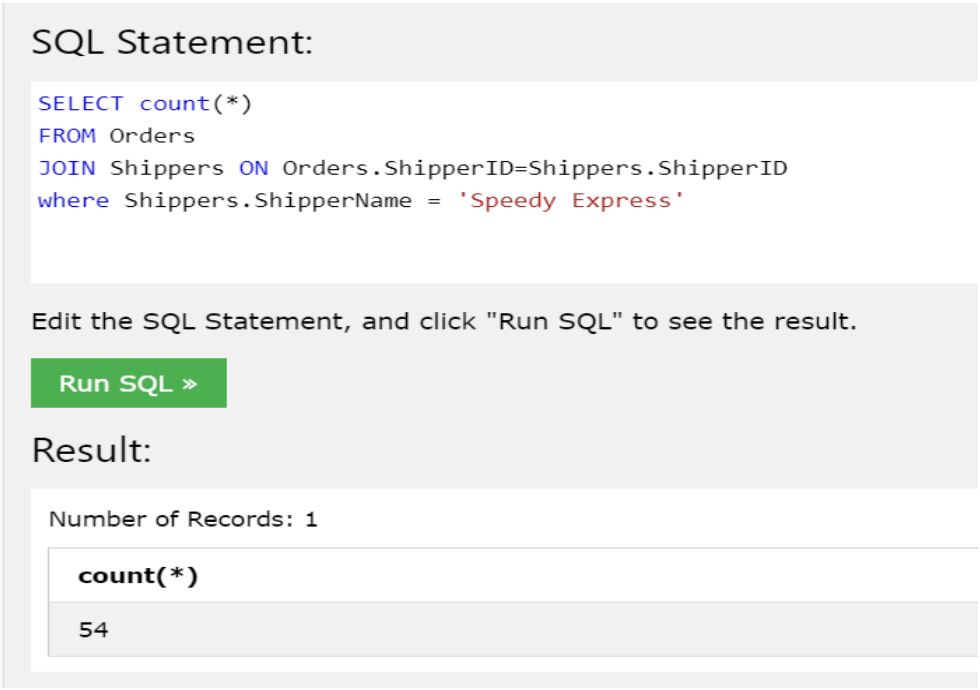
c. What is its value?
Solution:
The median value or new AOV value is `352.0`

Note: *The code can be found in the notebook named shopify.ipynb*

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?
Solution:
Query:

```
SELECT count(*)
FROM Orders
JOIN Shippers ON Orders.ShipperID=Shippers.ShipperID
where Shippers.ShipperName = 'Speedy Express'
```



b) What is the last name of the employee with the most orders?
Solution:
Query:

```
SELECT Employees.LastName,SUM(OrderDetails.Quantity) TotalOrders
FROM Orders
JOIN OrderDetails ON OrderDetails.OrderID = Orders.OrderID
JOIN Employees ON Employees.EmployeeID = Orders.EmployeeID
Group by
Orders.EmployeeID
Order BY
TotalOrders DESC LIMIT 1
```

SQL Statement:

```
SELECT Employees.LastName,SUM(OrderDetails.Quantity) TotalOrders
FROM Orders
JOIN OrderDetails ON OrderDetails.OrderID = Orders.OrderID
JOIN Employees ON Employees.EmployeeID = Orders.EmployeeID
Group by
Orders.EmployeeID
```

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

Run SQL »

Result:

Number of Records: 1

LastName	TotalOrders
Peacock	3232

c) What product was ordered the most by customers in Germany?
Solution:
Query:
SELECT Products.ProductName,MAX(Products.Unit) MostOrderedProduct
FROM Products
JOIN Suppliers on Suppliers.SupplierID = Products.SupplierID
where Suppliers.Country = 'Germany'

SQL Statement:

```
SELECT Products.ProductName,MAX(Products.Unit) MostOrderedProduct
FROM Products
JOIN Suppliers on Suppliers.SupplierID = Products.SupplierID
where Suppliers.Country = 'Germany'
```

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

Run SQL »

Result:

Number of Records: 1

ProductName	MostOrderedProduct
Thüringer Rostbratwurst	50 bags x 30 sausgs.

Here since the products that are ordered by the customers are only supplied by the suppliers in the Germany so I have used the suppliers table for the location rather than the customers table.