Summer 2022 Data Science Intern Challenge

Question 1:

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?

The program in the Git repo for your reference.

a)

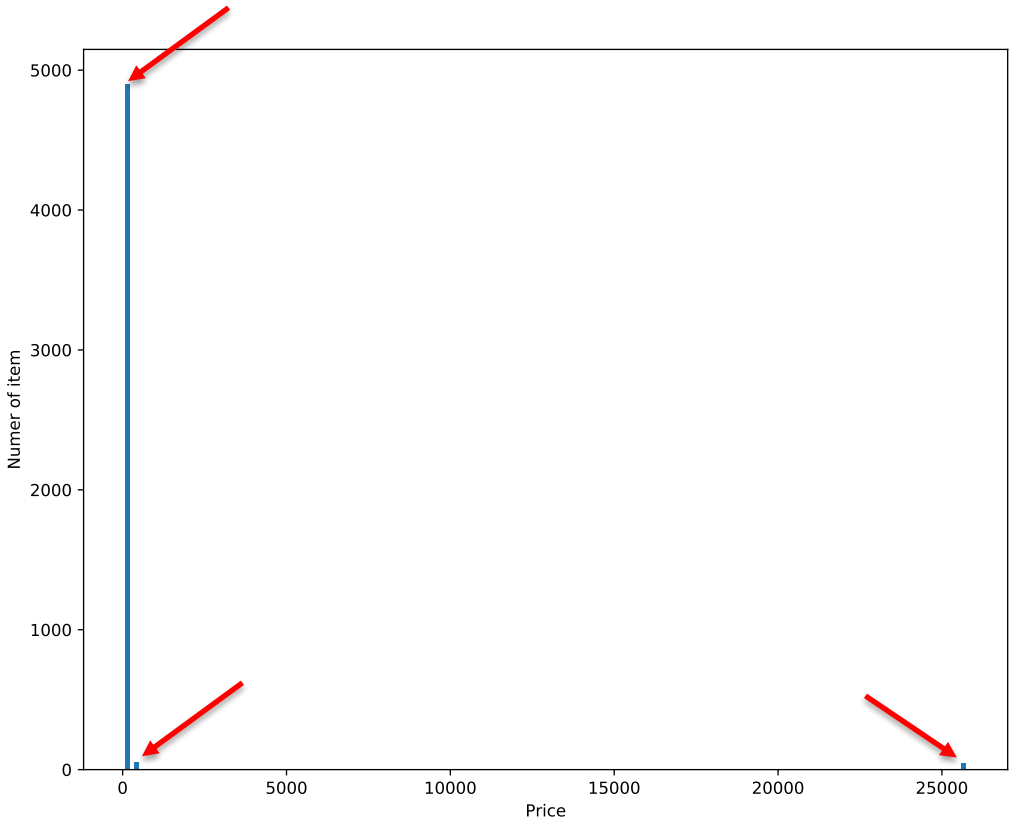
The naïve analysis would be dividing the total amount of orders in a month by the number of orders, which is:

Total amount of order / number of orders = **$3145.128**

Since some of the orders are in **bulk** (such as 2000 items per order) and some items are more **expensive** (such $25k shoes), we will get an AOV of $3k per order.

b)

To understand the data better, I calculated the **actual price of each item** for each store. Dividing the amount of order ($) by the number of orders will be the price of each purchased item. I added this information as another column into the dataset. Now let's have look at the prices of the purchased items:



As shown in the picture above (with the red arrows) there are **three** main price ranges. Two are under $5,000 and the other one is above $25,000.

Now it makes more sense why AOV was close to $3,000. **Maybe one way to have a more meaningful AOV is to categorize our customers based on their orders into 3 categories**, items with **low** prices, **medium** prices, or **high** prices, and then calculate the AOV separately for each group.

Also, we have the **mean, min, max, median** and **mode** along with the AOV to understand the statistics of the dataset better. For example, before splitting the data into different categories we have:

```
print('min:',min(df['product']),'max:',max(df['product']), 'mean:',np.mean(df['product']),
      'median:',np.median(df['product']), 'mode:',stats.mode(df['product']))
```

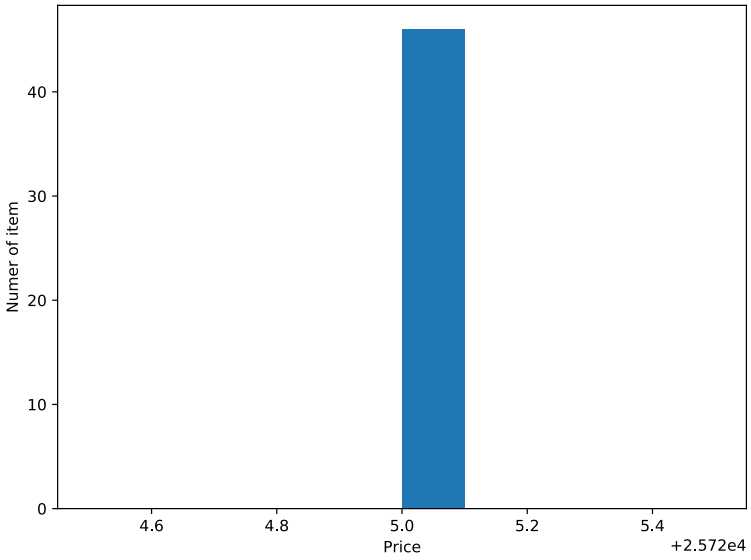
```
min: 90.0 max: 25725.0 mean: 387.7428 median: 153.0 mode: ModeResult(mode=array([153.]), count=array([256]))
```

Mean: is the shoe with the price of $387.74.
Median: the middle value of all orders is the shoe with the price of $153.
Mode: the most frequently occurring order is the shoe with the price of $153.

Mean and mode of $153 indicate that we might have **outliers** in our dataset. After splitting the data into different group (with regards to different modes), we will have different distribution of purchases and accordingly we can calculate different AOVs:

c)

1- Group one: customers who buy more expensive items with price over than $25000:



And:

```
print('min:',min(product_1),'max:',max(product_1), 'mean:',np.mean(product_1),
      'median:',np.median(product_1), 'mode:',stats.mode(product_1))
```

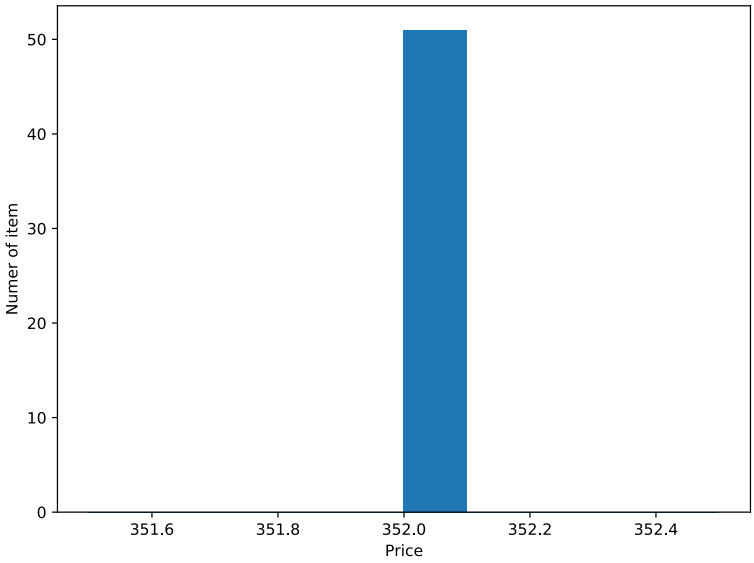
```
min: 25725.0 max: 25725.0 mean: 25725.0 median: 25725.0 mode: ModeResult(mode=array([25725.]), count=array([46]))
```

All statistics are the same and **AOV** will be:

Total amount of order / number of orders = **$49,213.04**

For shoes with price over 25,000, this AOV makes sense.

2- Group two: customers who spend on items with price 250<$<25000:



And:

```
print('min:',min(product_2),'max:',max(product_2), 'mean:',np.mean(product_2),
      'median:',np.median(product_2), 'mode:',stats.mode(product_2))
```

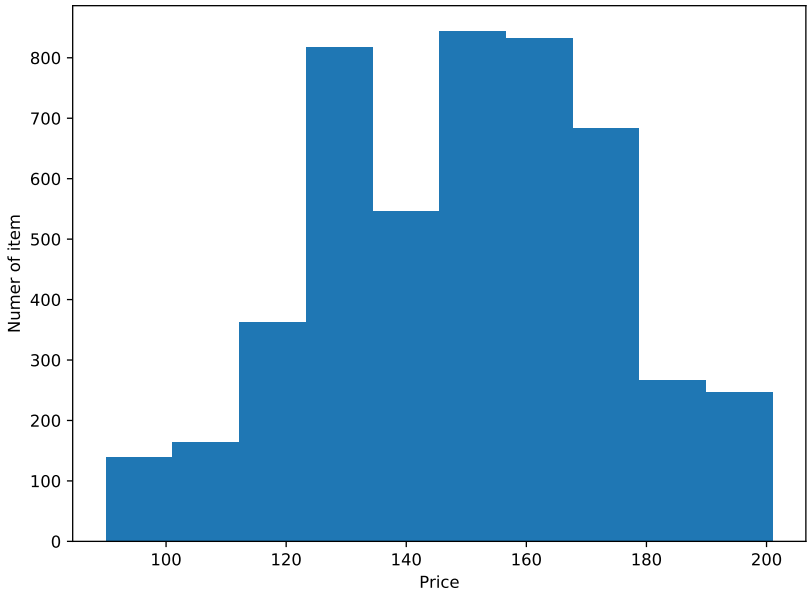
```
min: 352.0 max: 352.0 mean: 352.0 median: 352.0 mode: ModeResult(mode=array([352.]), count=array([51]))
```

All statistics are the same and **AOV** will be:

Total amount of order / number of orders = **$235,101.5**

This is the price range where customers purchased in bulk! I think that can be an amazing AOV.

3- Group two: customers who spend on item with price less than $250:




```
# cal stats
print('min:',min(product_3),'max:',max(product_3), 'mean:',np.mean(product_3),
      'median:',np.median(product_3), 'mode:',stats.mode(product_3)[0][0])
```

```
min: 90.0 max: 201.0 mean: 150.40016316540894 median: 153.0 mode: 153.0
```

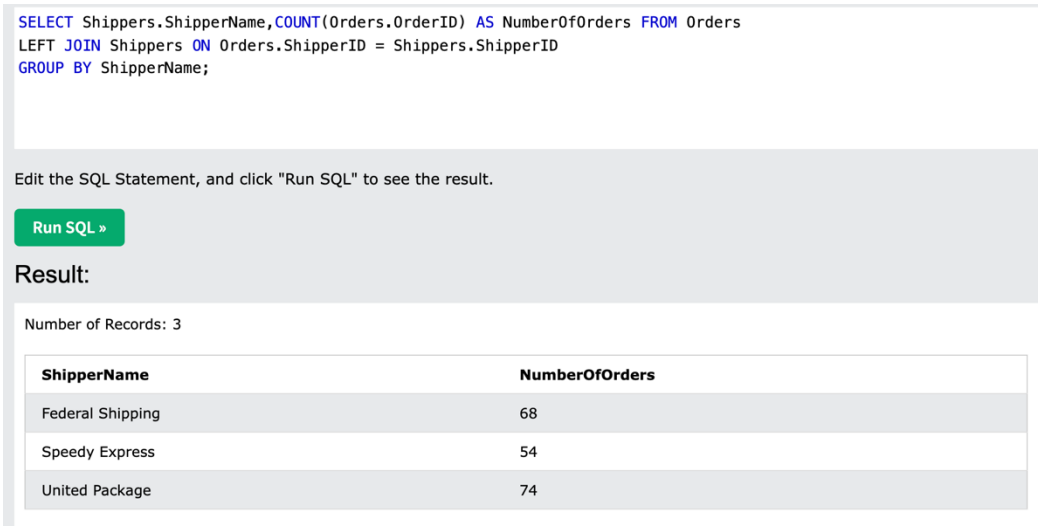
And **AOV** will be:

Total amount of order / number of orders = **$300.15**

Question 2:

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

SELECT Shippers.ShipperName,COUNT(Orders.OrderID) AS NumberOfOrders FROM Orders
LEFT JOIN Shippers ON Orders.ShipperID = Shippers.ShipperID
GROUP BY ShipperName;



```
SELECT Shippers.ShipperName,COUNT(Orders.OrderID) AS NumberOfOrders FROM Orders
LEFT JOIN Shippers ON Orders.ShipperID = Shippers.ShipperID
GROUP BY ShipperName;
```

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

Run SQL »

Result:

Number of Records: 3

ShipperName	NumberOfOrders
Federal Shipping	68
Speedy Express	54
United Package	74

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

SELECT Employees.LastName, COUNT(Orders.OrderID) AS NumberOfOrders FROM Orders
LEFT JOIN Employees ON Employees.EmployeeID= Orders.EmployeeID
GROUP BY LastName;

Run SQL »

Result:

Number of Records: 9

LastName	NumberOfOrders
Buchanan	11
Callahan	27
Davolio	29
Dodsworth	6
Fuller	20
King	14
Leverling	31
Peacock	40
Suyama	18

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

SELECT SUM(OrderDetails.Quantity) AS TotalNumOrder, Products.ProductName, Customers.Country
FROM Products
LEFT JOIN OrderDetails ON OrderDetails.ProductID= Products.ProductID
LEFT JOIN Orders ON Orders.OrderID= OrderDetails.OrderID
LEFT JOIN Customers ON Customers.CustomerID= Orders.CustomerID
Where Customers.Country = 'Germany'
Group by ProductName
ORDER BY TotalNumOrder DESC
LIMIT 1

```
LEFT JOIN Customers ON Customers.CustomerID= Orders.CustomerID
Where Customers.Country = 'Germany'
Group by ProductName
ORDER BY TotalNumOrder DESC
LIMIT 1
```

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

Run SQL »

Result:

Number of Records: 1

TotalNumOrder	ProductName	Country
160	Boston Crab Meat	Germany

If you were to open a Shopify store, what would you sell and why (or tell us about your store if you have one!) (200-word limit)

If I was supposed to open a Shopify store, I would either sell my own product or I would sell the best-selling products in the current market.

I would sell my own product, if I created some art crafts such as paintings, hand-made jewellery, or custom phone case designs. I would then take advantage of low-cost advertisements and utilize the power of social media (Twitter, YouTube, Facebook, Instagram, Pinterest and etc.) as a marketing platform as well as using machine learning tools to predict people's behaviour to leverage my revenue. Also, Shopify has apps and integrations that I would use to evaluate my store statistics.

To determine the best-selling products, I would need access to some statistics such as the best-selling products in the existing ecommerce platforms such as amazon, Facebook market, Shopify, etc. I think this information can also be extracted via mining the websites, texts, and images on different media platforms, to find the most occurring or discussed products, their reviews, and their ratings. I did a quick search, and the best-selling product on amazon appears to be toys and games, baby products, books, and electronics.