

Winter 2022 Data Science Intern Challenge

Harmanpreet Kaur

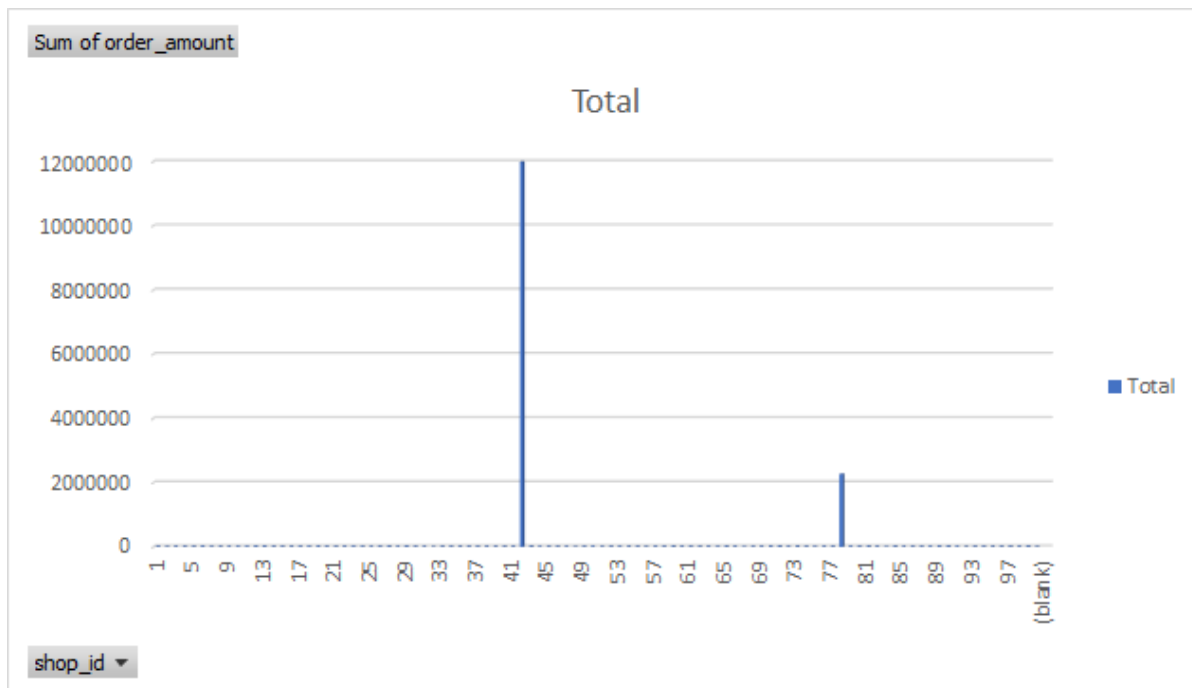
Question 1: Given some sample data, write a program to answer the following: [click here to access the required data set](#)

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.

Answer: AOV value of \$3,145.13 has been calculated by taking the Mean of the Order amount. Since the data provided is skewed, therefore Mean is not a true representation of Average Order Value. The Standard deviation from Mean is \$41,282.5 which explains that the values vary \$41,282.5 from the mean and hence concluding that mean is not a true representation of Average Order Value in this dataset.

As we can see from the below graph, the dataset is skewed heavily mainly due to do two stores, store #43 and store #78. If we remove data for these two stores, then the average of the remaining data set is \$300.15, which is comparable to the average order value.



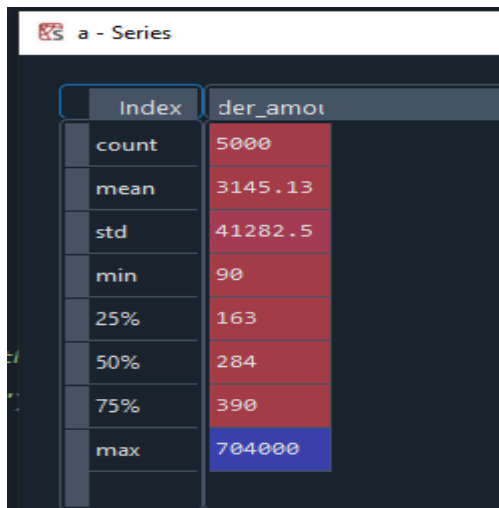
Winter 2022 Data Science Intern Challenge

Harmanpreet Kaur

Alternatively, median would be a more appropriate central tendency to use as it is more robust and also, since median calculates the middle value of the data set, it diminishes the effect of outliers. The median value for this dataset is \$284.

CODE

```
df = pd.read_csv("WinterChallenge.csv")
a = df.order_amount.describe()
```



Index	der_amoi
count	5000
mean	3145.13
std	41282.5
min	90
25%	163
50%	284
75%	390
max	704000

Minimum = 90 and Maximum = 704000

b. What metric would you report for this dataset?

Answer: Median would be more appropriate to use for this dataset.

I also decided to remove the outliers as there is a very large range between 3rd quartile, which is \$390 and max value, which is \$704,000. To remove the outliers, we can use the Interquartile Range - IQR Method. Using IQR method, we can find the upper range ($Q3 + (1.5 * IQR)$) and lower range ($Q1 - (1.5 * IQR)$) of the data.

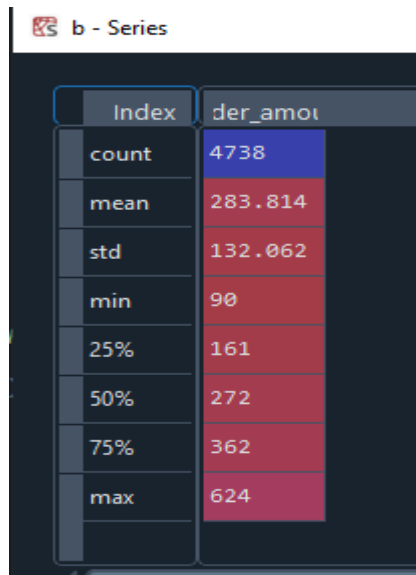
CODE

```
q1 = df.order_amount.quantile(q=0.25)
q2 = df.order_amount.quantile(q=0.5)
q3 = df.order_amount.quantile(q=0.75)
iqr = q3 - q1
```

```
df_truncated = df[(df.order_amount < q2 + iqr * 1.5) & (df.order_amount > q2 - iqr * 1.5)]
b = df_truncated.order_amount.describe()
```

Winter 2022 Data Science Intern Challenge

Harmanpreet Kaur



Index	der_amot
count	4738
mean	283.814
std	132.062
min	90
25%	161
50%	272
75%	362
max	624

After removing the outliers, as shown above, the value of mean becomes \$283.814 and median goes slightly down to \$272. So, mean can also be used to calculate AOV, by removing the outliers.

c. What is its value?

Answer : The value is \$284, if we use median as the measure to calculate AOV of the whole dataset. Also, we can use mean by removing the outliers and the value in that case is \$283.814 .

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?

Answer. There are 54 orders . Orders were counted by joining the tables 'Orders' and 'Shippers' based on shipper ID and then getting the count from table Orders.

CODE:

```
Select Count(*) AS SpeedyExpressOrders
From [Orders]
Join [Shippers]
ON [Shippers].ShipperID = [Orders].ShipperID
Where [Shippers].ShipperName = 'Speedy Express'
```

Winter 2022 Data Science Intern Challenge

Harmanpreet Kaur

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

Answer: Last name of the Employee with most Orders is 'Peacock' with 40 Orders. To find the last name of the employee with the most orders, we first join table 'Employees' with 'Orders' by Employee ID field. Then we can count the number of orders for each Employee ID and display their corresponding Last Names in descending order. Limiting this to 1, will provide the last name of the employee with the highest number of orders.

CODE:

```
Select [Employees].LastName, Count(*) AS OrderNums
From [Orders]
Join [Employees]
ON [Orders].EmployeeID = [Employees].EmployeeID
Group by [Employees].LastName
Order by OrderNums DESC
Limit 1
```

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

Answer: Boston Crab Meat was ordered most i.e 160 times by customers in Germany. To find the product most ordered in Germany, we need the product name, country and number of orders. Hence, we need to join 3 tables - Orders, Order Details and Products, using Order ID, Customer ID and product ID.

CODE:

```
Select [Customers].Country,
       [OrderDetails].ProductID,
       [Products].ProductName,
       SUM([OrderDetails].Quantity) AS "Total"
From [Orders]
Join [Customers]
ON [Customers].CustomerID = [Orders].CustomerID
Join [OrderDetails]
ON [OrderDetails].OrderID = [Orders].OrderID
Join [Products]
ON [Products].ProductID = [OrderDetails].ProductID
Where [Customers].Country = 'Germany'
Group by [OrderDetails].ProductID
Order by Total DESC
limit 1
```