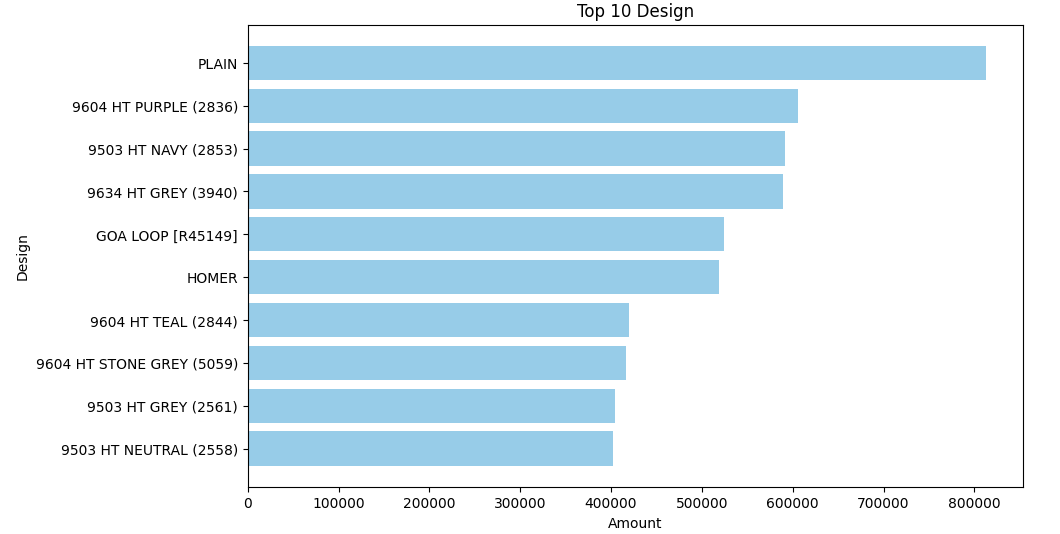
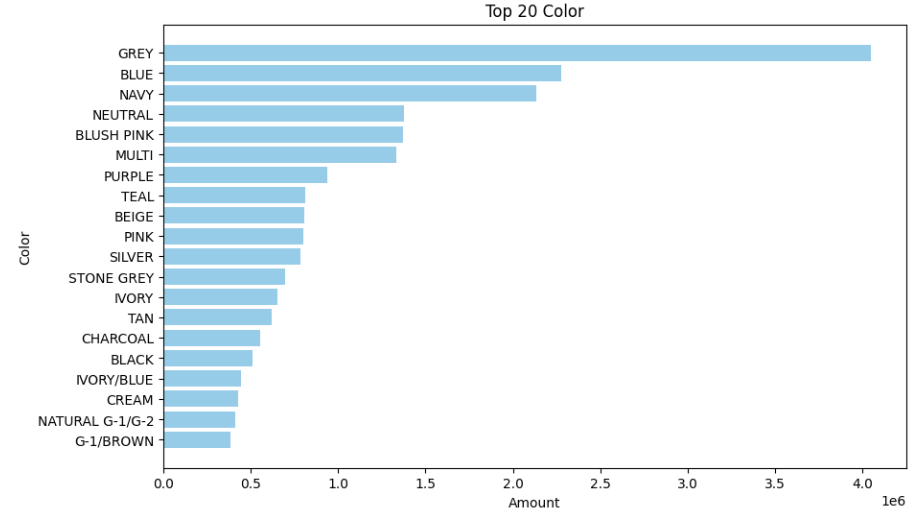
MLUL1 Group Assignment - Group 11

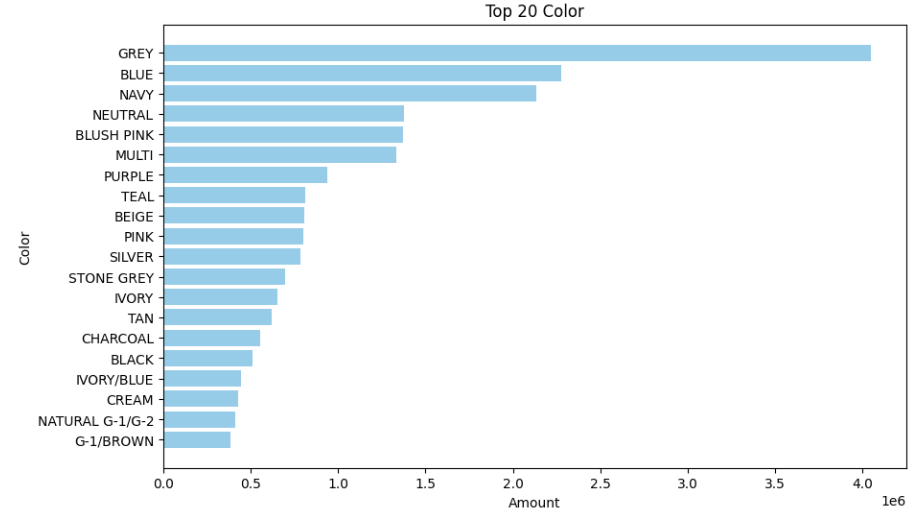
## 

## Exploratory Data Analysis:

### **Insight 1: Product Attribute & Customer Analysis**

Top 10 Carpet Designs, Top 10 Carpet Colors , Top 10 Customers

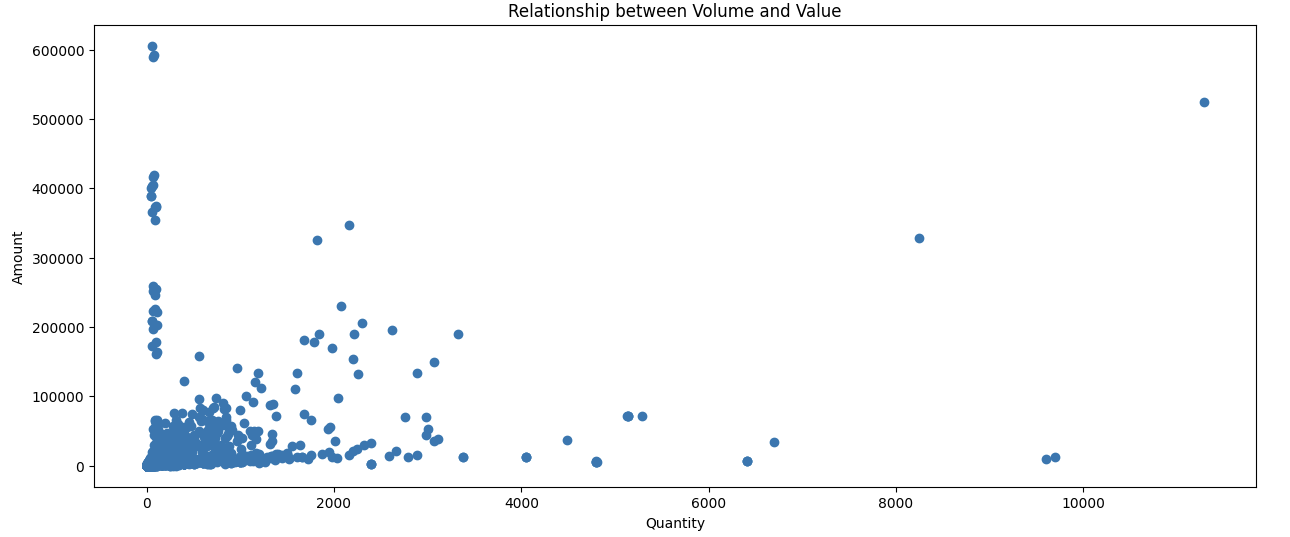




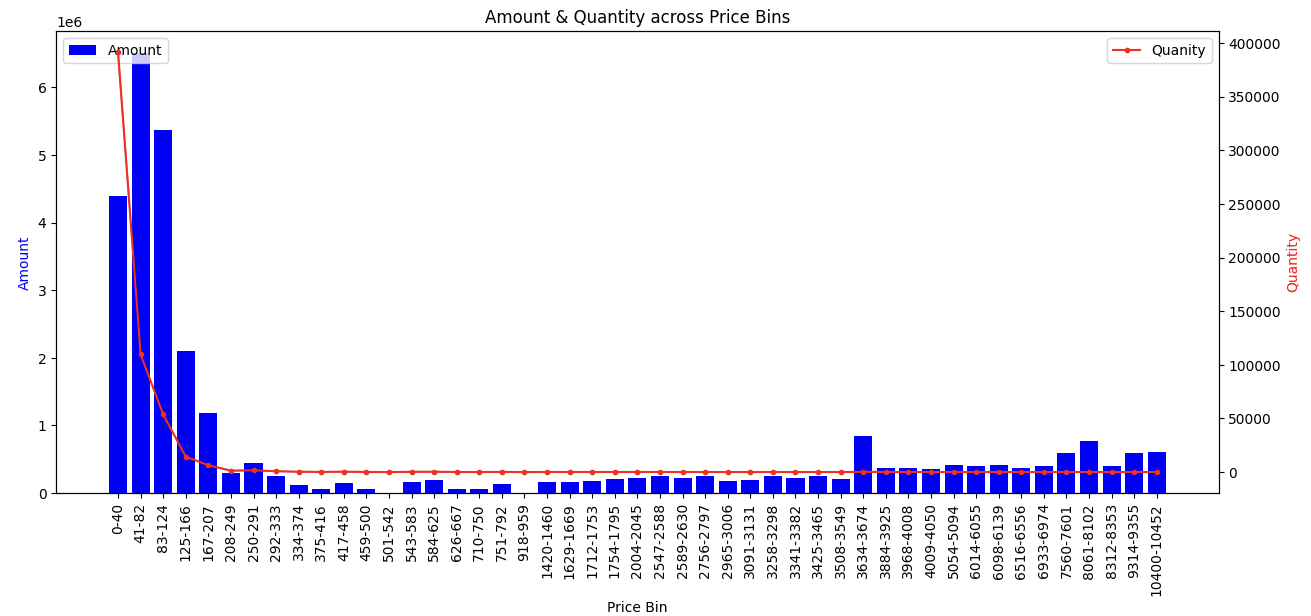
Analysis provides the information about which carpets are high in demand and generates more revenue when it comes to Carpet attributes such as Carpet Design or Color. Also Top 10 customers gives the understanding of focused customers for Campo carpet manufacturers as they bring the maximum revenue.

### **Insight 2: Amount , Quantity ordered and Price Relationships**

**Quantity vs Amount for Products:** Based on Product category data associated prices , Quantity and Revenue generated we see clusters of amount getting generated at low quantity which means the majority of products are value tier products. Products with low Quantity and high amount indicate the Premium products.

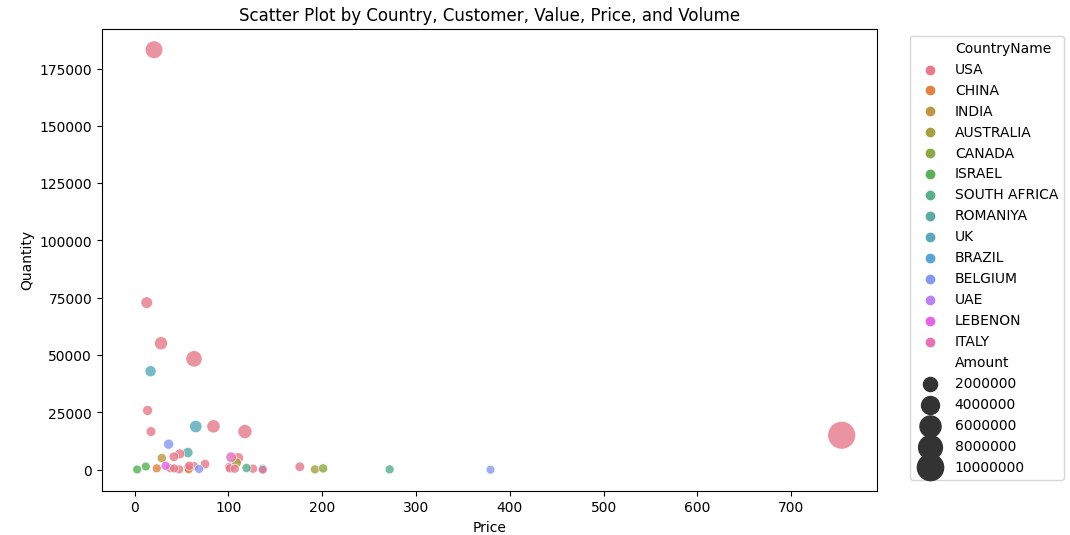


### **Insight 3 :Price Bin Analysis**



Products are categorized into price bins and associated Amount and Quantity is plotted. We can see low price driving the volume and Amount whereas high price products are driving significant Amount with low volumes

### Insight 4 : Customer performance across countries



Bubble represents the Customer with color = Country , Size = Amount , X axis Price and Y Quantity. USA customers are generating maximum sales with both low and high price products.

Most other country customers are selling value tier products that get low to medium ranges sales compared to the market.

### Insight 5 : Customer share of business and Index

### 

This represents the market share across all countries to understand customers behavior with Index. It denotes positive index is more for customers purchasing premium products whereas less than zero indicates that customers are focused to drive volume across consumers by selling value tier carpets or running promotions with reduced prices.

## Customer Segmentation

### The features

1. The number of Orders made by the customer
2. The monetary value of the customer

We have a scatter plot of the 2 features:

The data is widespread, so we have to apply suitable scaling on it.

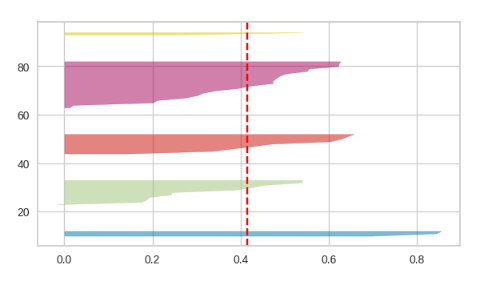
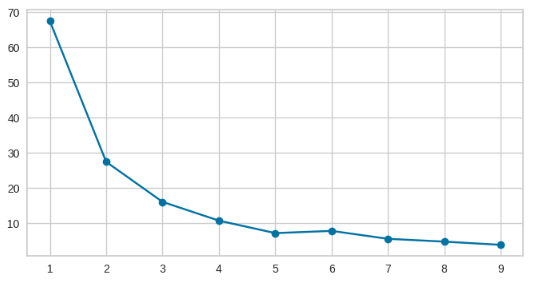
### Scaling the data

We have tried different Scaling methods, such as Standard Scaling, Min-Max Scaling and finally Logarithmic Scaling.

The output of the Logarithmic Scaling was satisfactory.

### Determining the number of clusters

We went ahead with the logarithmic scaled data and apply the elbow method and silhouette visualization to identify the number of clusters best suited for the data.



The Silhouette visualization for **5 clusters** shows 5 distinct clusters with no customer wrongly allocated.

### Understanding the Clusters

**Cluster 0** are customers with low order and low Amount => Chunks

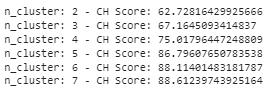
**Cluster 1** are customers who have a moderately high number of orders and moderately high amount billed. => Stars

**Cluster 2** are customers with low orders but high amount billed => Exclusive

**Cluster 3** are customers with moderate orders and moderate amount billed => Consistent

**Cluster 4** are customers with high Order and high Amount => High Rollers

### Evaluating the clusters

We applied two metrics as below:

#### Calinski-Harabasz Index:

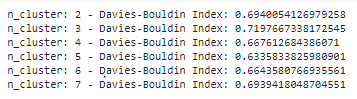
Relative comparison among clusters of

different sizes show that, when cluster size is

5 the index is high. This makes the clustering

a good one.

#### Davies-Bouldin Index:



A lower value for the index means good clusters. The Index for 5 clusters show a low value.

## Product Segmentation

We started with the following features,

Total Quantity of Item ordered

Total revenue by the Item

Item’s share in revenue

The number of customers buying the item

Average quantity of Item per Order

Average price of Item

A regression analysis of the above show that we can use one as the proxy of others and finalize the features.

### The features

Number of Customers buying the Item

Item’s share in revenue

Average Quantity per Order

Average Price of Item

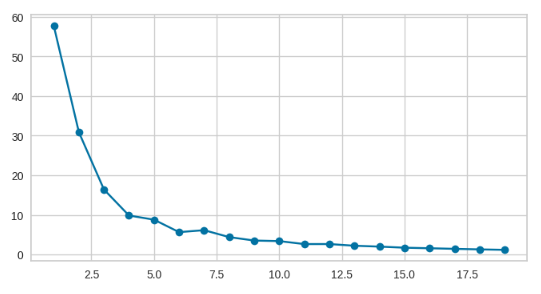
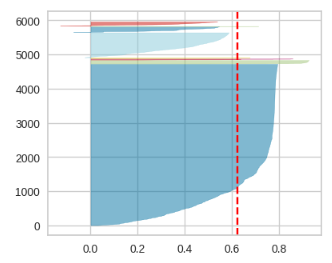
### Scaling the Data

We started scaling the data using the Standard Scaling and then Min-Max Scaling.

The Output from the Min-Max Scaling looks satisfactory.

### Determining the number of clusters

We went ahead with the Min-Max scaled data and apply the elbow method and silhouette visualization to identify the number of clusters best suited for the data.

The Silhouette visualization for **7 clusters** shows 7 distinct clusters with few customers wrongly allocated.

### Understanding the Clusters

**Cluster 0** are the products that have a very low price with very low orders, hence a low share in revenue. This constitutes bulk of the Items(approx 80%).

**Cluster 1** are products which are very high priced but since they have low orders so controls the mid-section when it comes to revenue. It is a very low populated cluster.

**Cluster 2** are the products that have a low price with low orders, hence a low share in revenue.

**Cluster 3** are the products that have a very low price with high orders and low-moderate share in revenue.

**Cluster 4** are the products that have a very low price with moderate orders and moderate share in revenue.

**Cluster 5** are the products that have a moderate price with very low orders and low share in revenue.

**Cluster 6** are the products that have a moderate price with very low orders and good share in revenue.

### Evaluating the clusters

We applied two metrics as below on the K-means clusters:

#### Calinski-Harabasz Index:

Relative comparison among clusters of different sizes, a high value signify good clustering..

A CH score of 11011 wrt other clusters show that 7-clusters give a good outcome

#### Davies-Bouldin Index:

A Davies-Bouldin score of 0.55 is a good score to have given the scores for other K values.

By HDBSCAN, number of Clusters is 1533.

It is too hard to understand and interpret that many clusters.

## 3 keys Lessons learnt

* Univariate and bivariate Data Analysis to identify trends in the data
* Understanding and implementing the different scaling methods and different clustering algorithms to cluster the customers and products
* Drawing inference from the clusters created and giving them labels to treat each cluster differently.