Pattern Recognition and Machine Learning
Minor-Project

Team Members:--

Sourabh Malviya [B21EE070]
Tushar Sharma [B21CS095]

Problem Statement:--

A company that sells some of the product, and you want to know how well the selling performance of the product. You have the data that we can analyse, but what kind of analysis can we do? Well, we can segment customers based on their buying behaviour on the market. Your task is to classify the data into the possible types of customers which the retailer can encounter.

Possible Solution:-

We need to analyse the data precisely. We made different visualisations after preprocessing the data. This gave us a good visual analysis. We need to perform customer segmentation. This is done by using various clustering methods.

Preprocessing:-

First, the file is read using the function "pd.read_excel()". Now, we have checked the missing values present in different columns, using the function, "isnull.any()".
Following are the obtained columns and their status of missing values.

```
InvoiceNo      False
StockCode      False
Description     True
Quantity       False
InvoiceDate    False
UnitPrice      False
CustomerID      True
Country        False
dtype: bool
```

Two columns, "Description" and "CustomerID" are having missing values.
Now, preprocessing is done. Some columns which are not of use are dropped like "InvoiceNo".
Then, missing values in "Description " and "CustomerID" columns are replaced by its mode value.

```
The mode for the column Description is  WHITE HANGING HEART T-LIGHT HOLDER
Again checking missing values in Description column
False
The mode value for CustomerID is  17841.0
Again checking missing values in CustomerID column
False
```

By this, a preprocessed data frame is obtained.

Data Visualisation:-

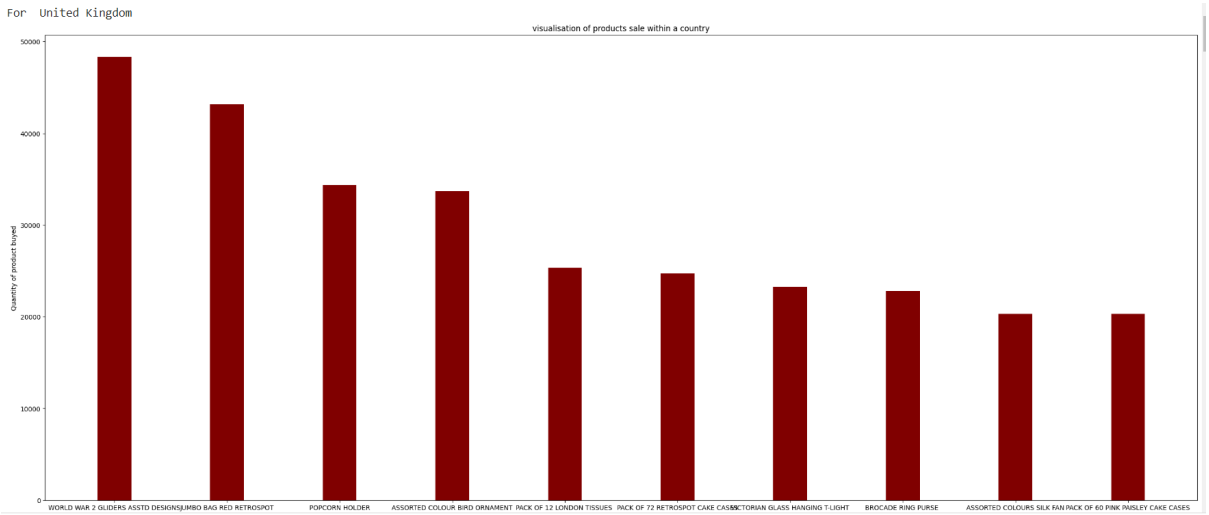
Now, to visualise and analyse the dataset variation with features we have calculated the count of different products for each and every country present. These counts are stored in the dictionary.

```
{'United Kingdom': {'WHITE HANGING HEART T-LIGHT HOLDER': 19584, 'WHITE METAL LANTERN': 1779, 'CREAM CUPID HEARTS COAT HANGER': 1411, 'KNITTED UNION FLAG
```

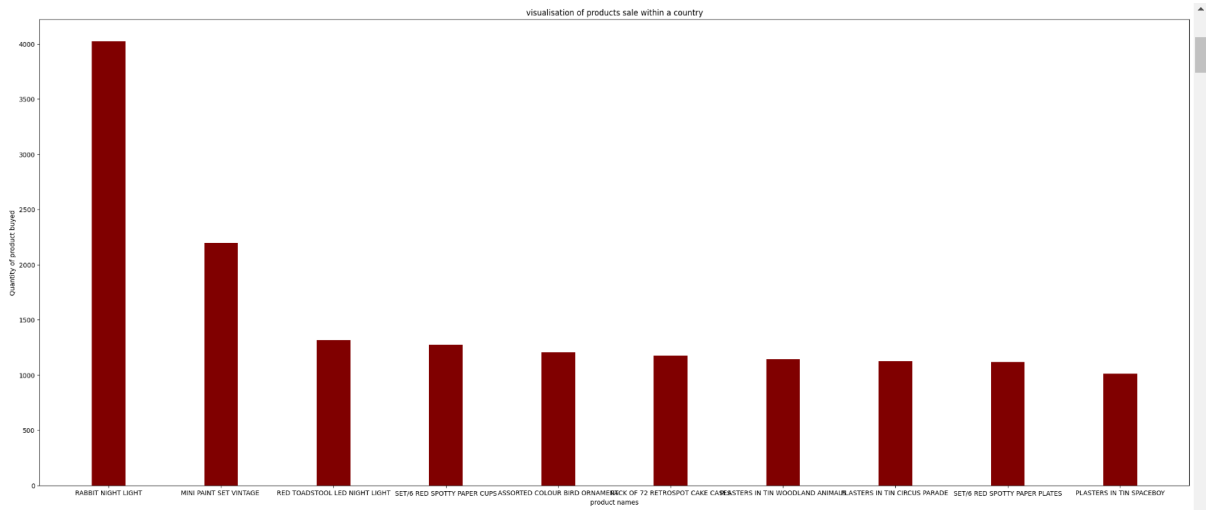
Now, the count of products are sorted and top selling products are found and plotted.

Following plots are obtained for visualising the datasets:---

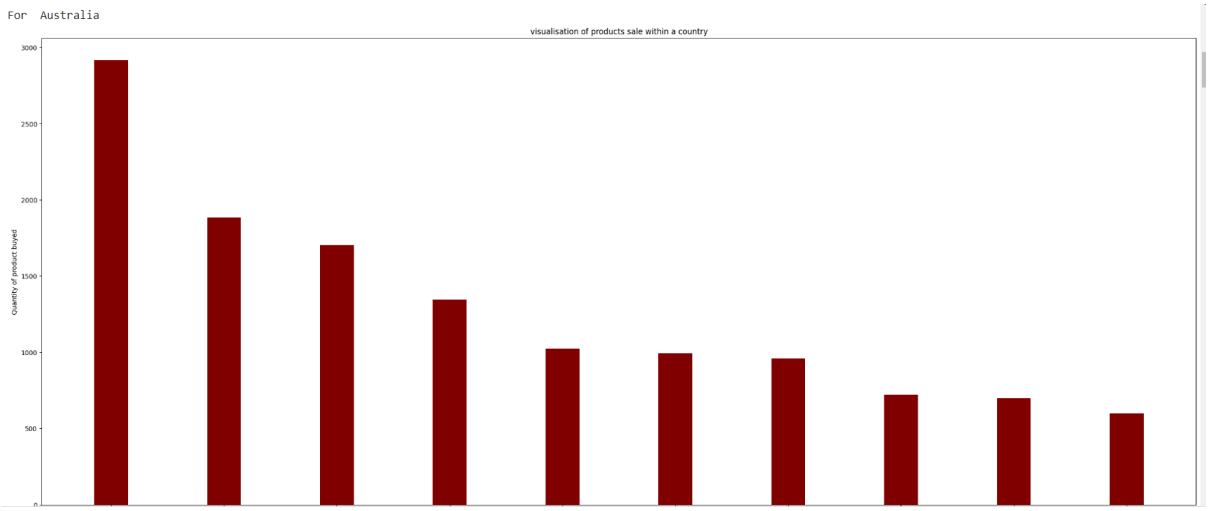
For united states-



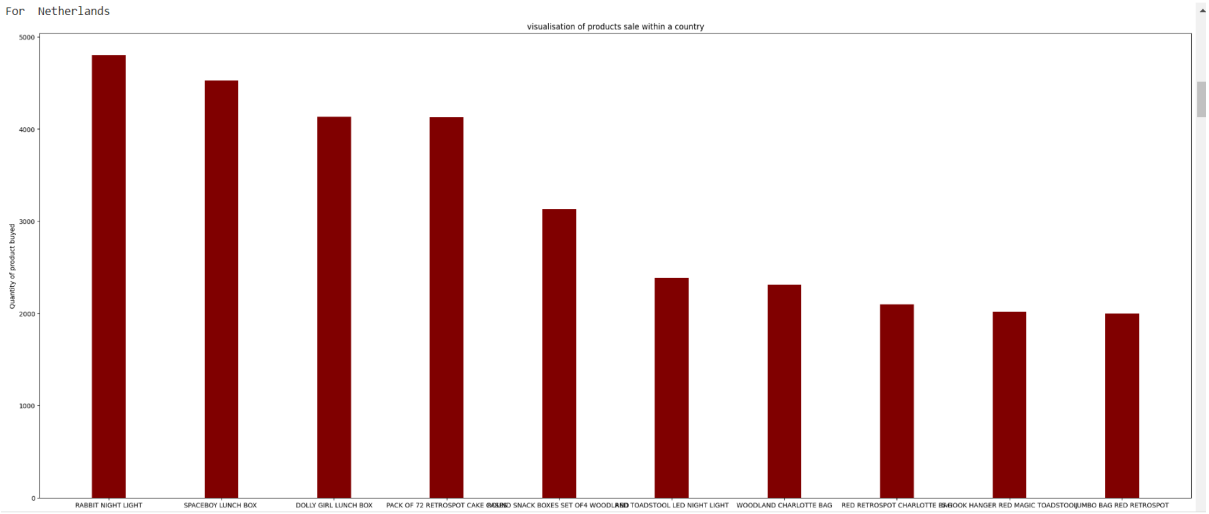
For France



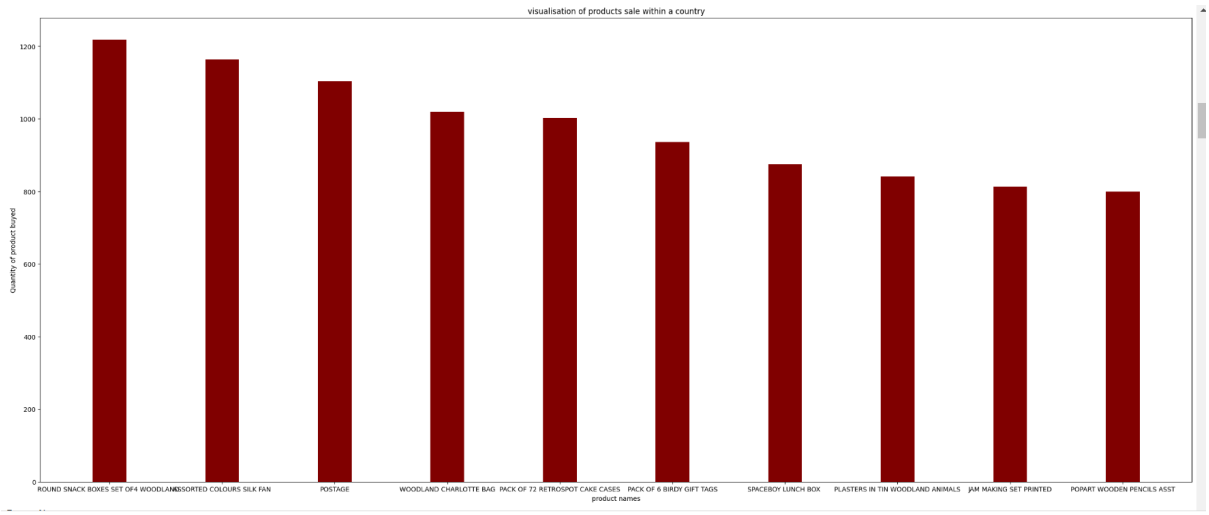
For Australia



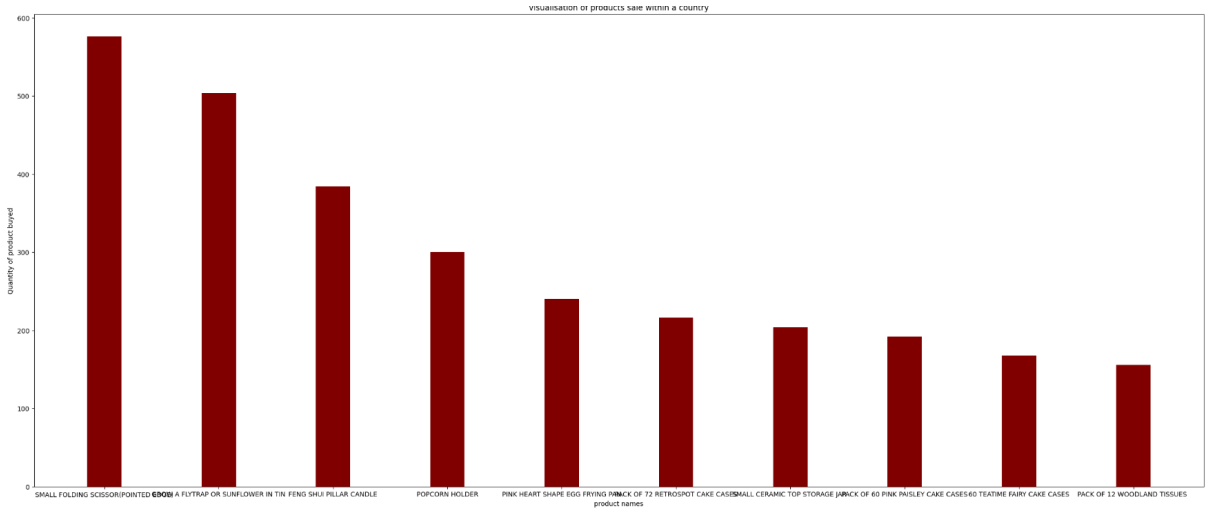
For Netherlands



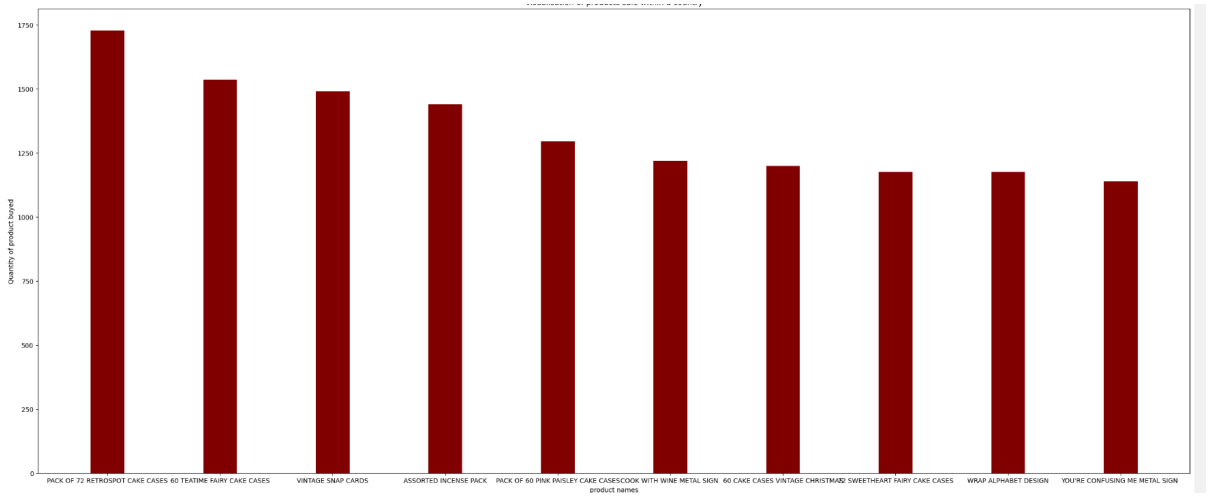
For Germany



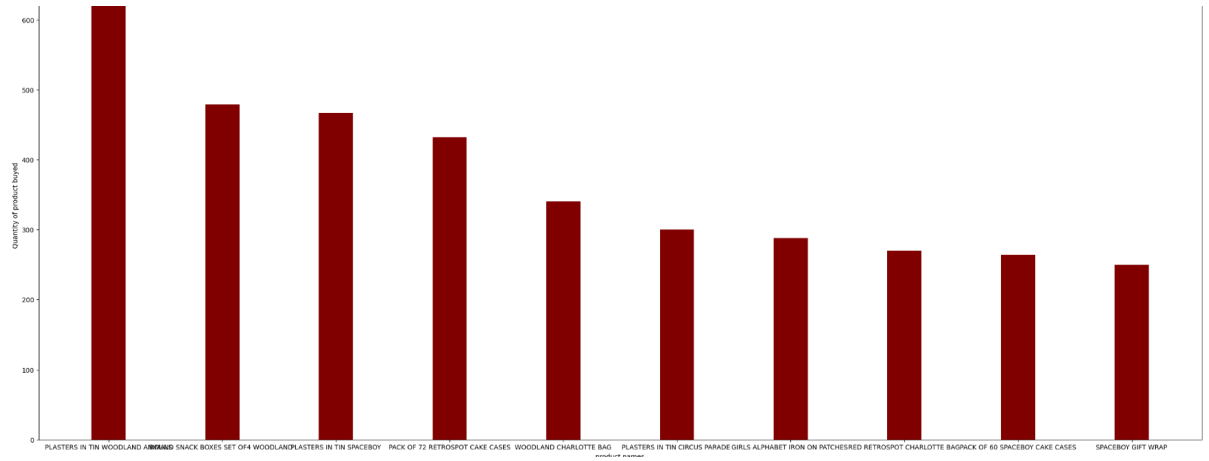
For Norway



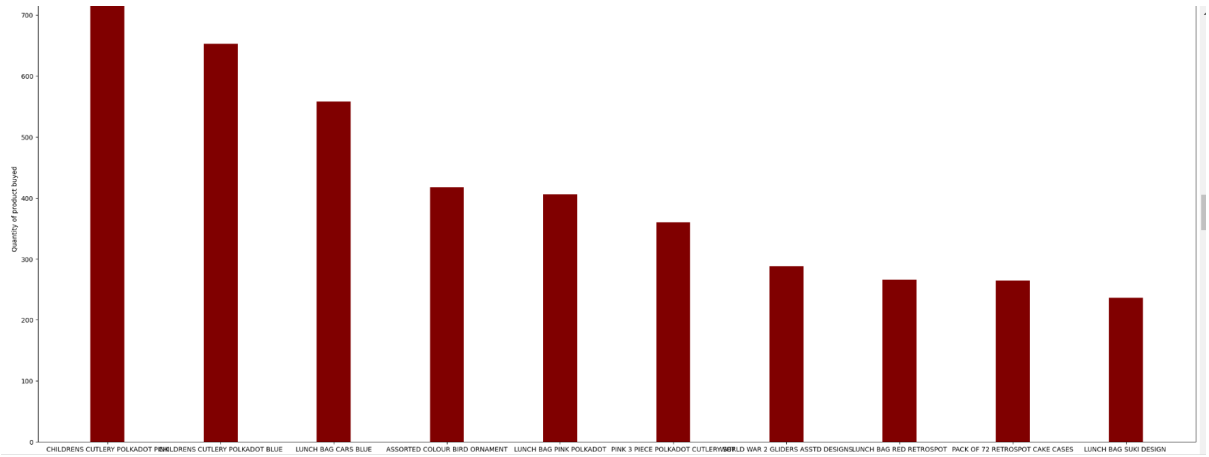
For Eire



For Switzerland

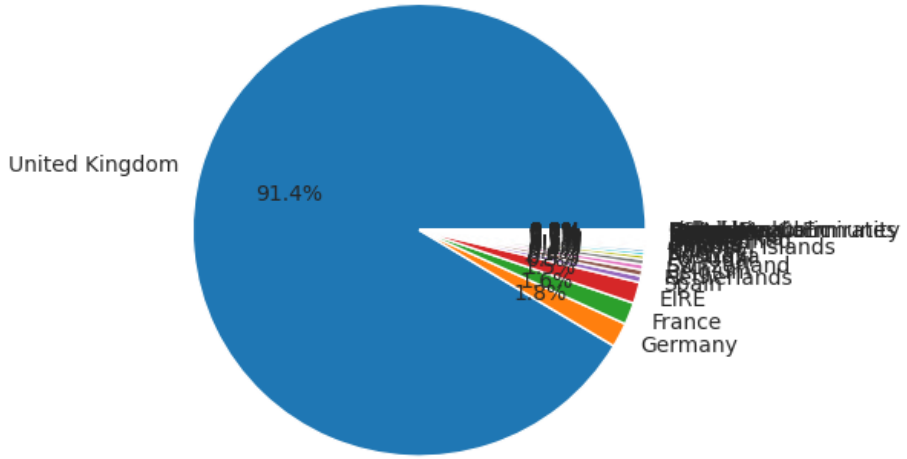


For spain

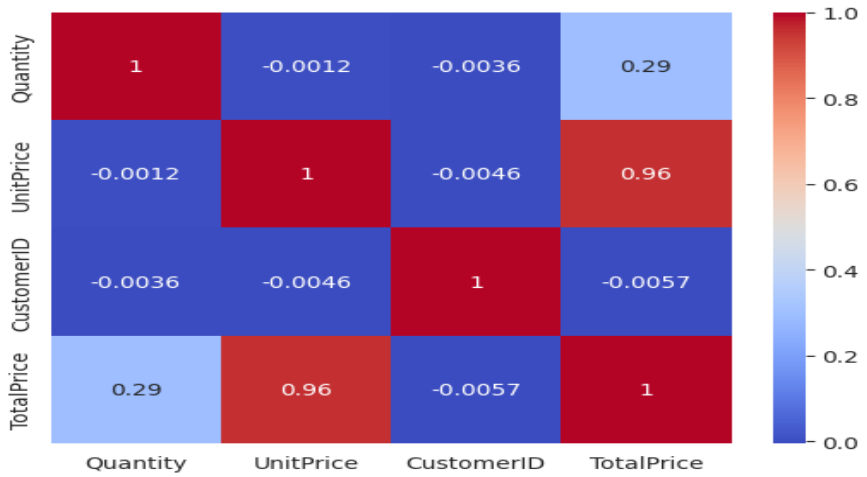


Other plots are in a collab file.

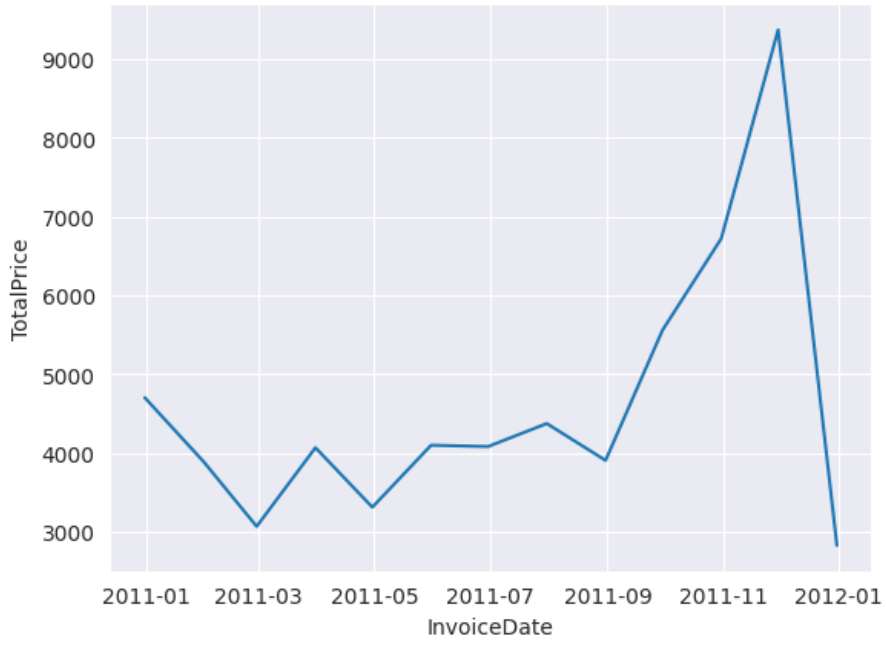
There were a large number of countries but the United States is one of the largest customers. Following we have Germany and France.



We also tried to analyse the data based on its correlation matrix. Here is what we got.



Analysing and visualising the monthly sales.



We also tried to analyse the dataset by splitting and separating it by country. We printed top 10 revenue and quality StockCodes for each country.

Another analysis done was based on the words used in Description. With this analysis we can infer what kind of words are most used in the products with maximum sales. Tus, we can conclude the products which may give maximum profit. Here, the plot.

We can say that the items with words such as Bag and Heart in description are the most sold.

Clustering of the dataset:--

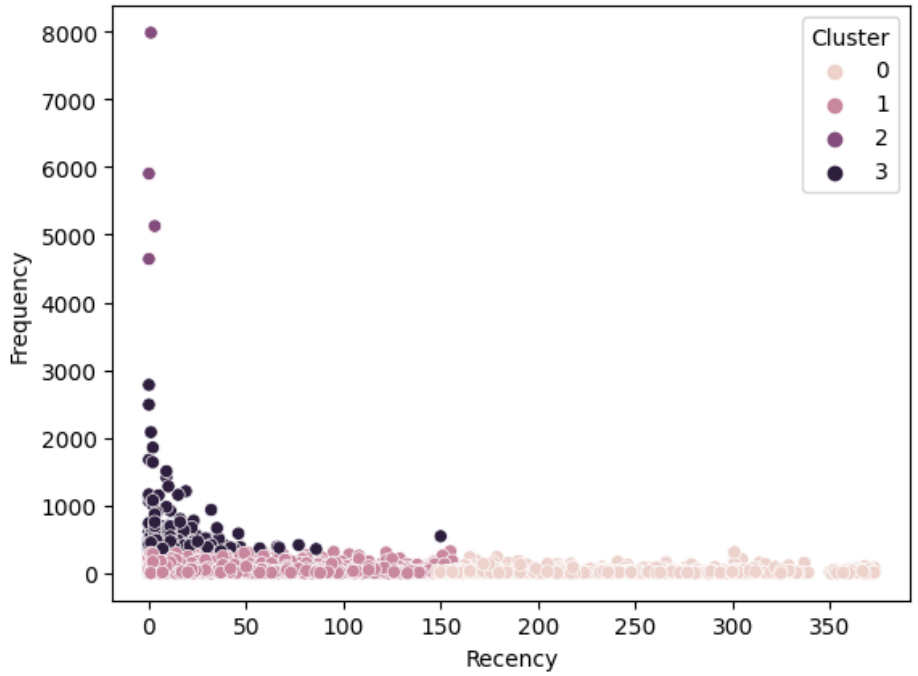
We used clustering algorithms such as K-Means or Hierarchical clustering to group customers with similar RFM scores into segments. This enabled us to understand the different types of customers that the retailer is dealing with.

Applying K-Means Clustering using the RFM scores

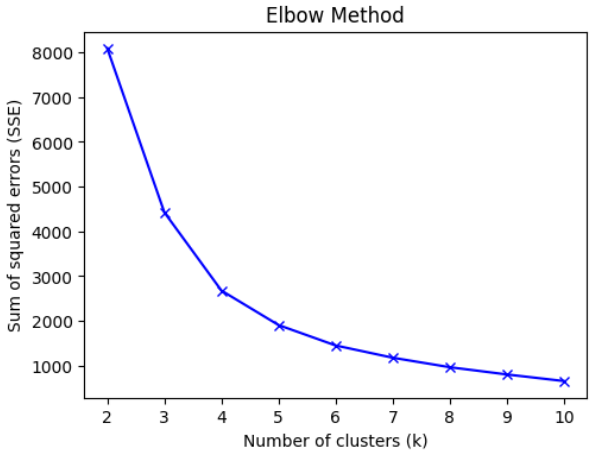
The code in the file calculates the Recency, Frequency, and Monetary scores for each customer, and assigns them to quartile labels based on their values. The labels are then combined into a single RFM segment string and the RFM score is calculated as the sum of the quartile labels. This allows us to segment the customers based on their RFM characteristics, which can help to identify different groups of customers with different needs and behaviours. Here are a few rows.

CustomerID	Recency	Frequency	Monetary	R	F	M	RFM_Segment	RFM_Score
12346.0	325	2	0.221120	1	1	1	111	3
12347.0	1	182	20.128213	4	4	4	444	12
12348.0	74	31	3.432023	2	2	2	222	6
12349.0	18	73	8.077038	3	3	3	333	9
12350.0	309	17	1.880267	1	1	1	111	3

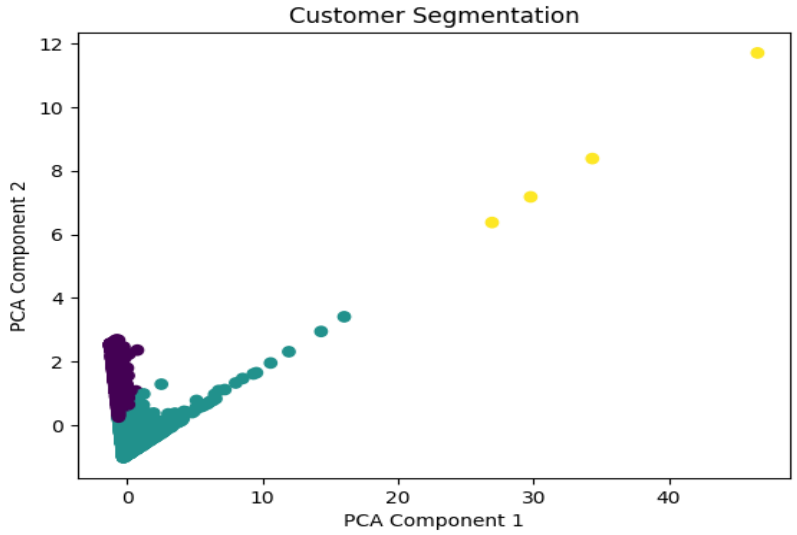
Then, we applied K Means on the ‘rfm’ dataset. Which is basically the above dataset.



We used the Elbow Method and silhouette scores to do optimal clustering.



Here is the final plot.



Applying the K-Means clustering on the preprocessed dataset on different set of features- Quantity and Unit Price

First, needed columns are selected for initial clustering.Data standardisation is also done.Then, with the scaled data k means is applied and cluster summary is stored on the basis of features, “Quantity” and “UnitPrice”.

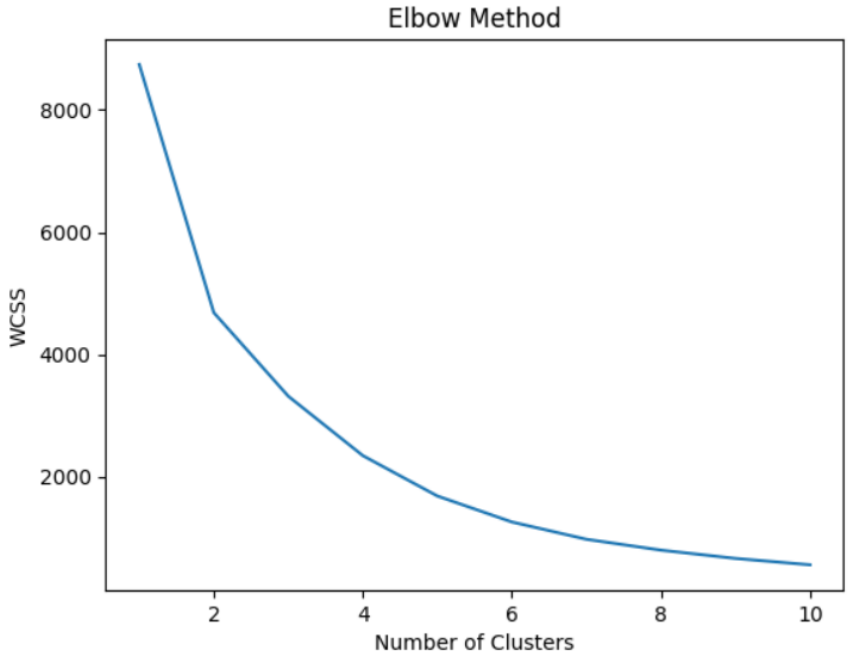
Obtained cluster summary:--

```
The cluster summary is as follows--
            Quantity    UnitPrice
cluster
0           9.839308     4.054723
1          -0.806452  9730.436452
2      -77605.000000     1.560000
```

Later, the elbow method is applied to find the optimal number of clusters for the k means algorithm. Here, we have varied the number of clusters from 1 to 11 and for each iteration wcss (Within-Cluster Sum of Square) i.e. the sum of the

square distance between points in a cluster and the cluster centroid is found and plotted to obtain the elbow graph.

Elbow point:--

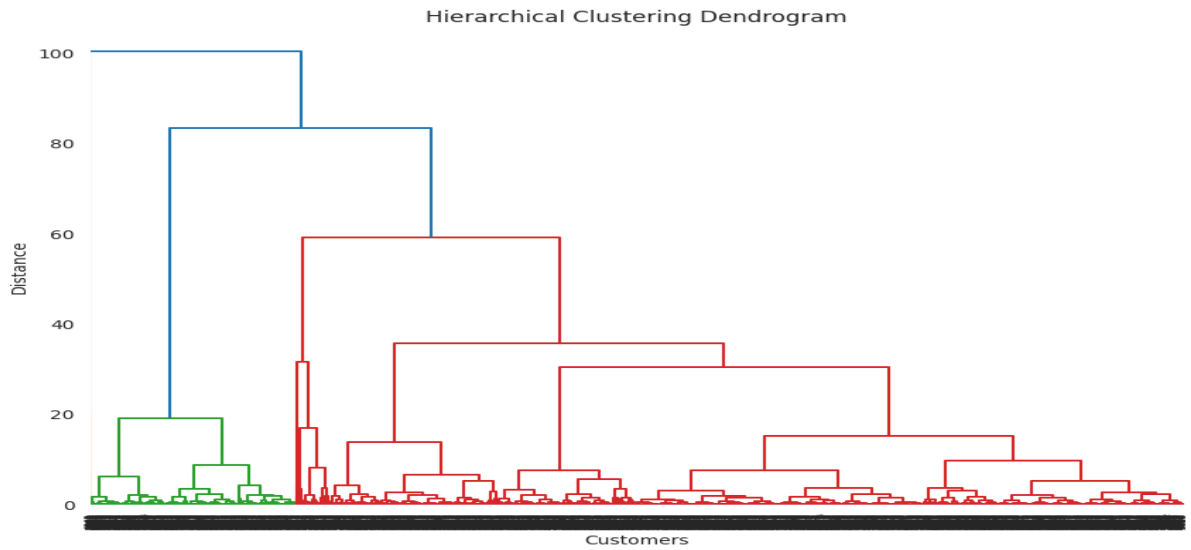


Elbow point is coming at 3. We use this value for k means clustering.

Obtained clusters:--

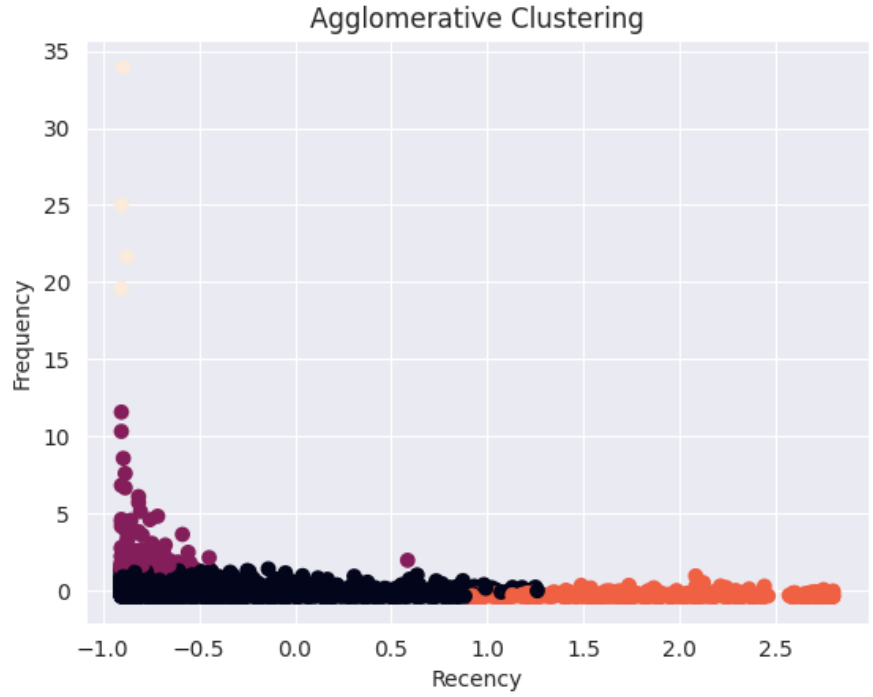
Performing Hierarchical clustering

Selecting the features to be used in clustering as Recency, Frequency, and Monetary, and standardising the data using the Z-score transformation, we performed the Hierarchical clustering. We used the Ward method and plotted a dendrogram to visualise the clusters.

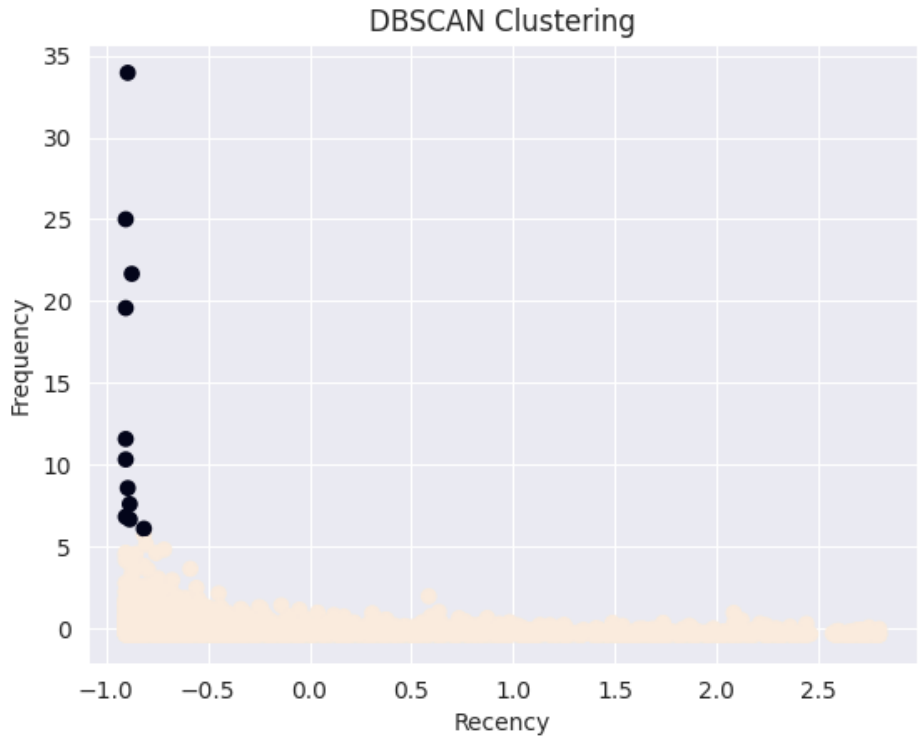


Agglomerative Clustering

Taking the same features as in the Hierarchical clustering, we performed Agglomerative clustering.



DBSCAN Clustering



Conclusions

1. The United States has our main market. We have a majority of customers from here.
2. We also find that the sales are at peak towards the end of a year, i.e. New Year.
3. The items with words such as Bag and Heart in description are the most sold.
4. We also tried to analyse the dataset by splitting and separating it by country. We printed top 10 revenue and quality StockCodes for each country. We can improve our revenue by increasing these stocks in the respective countries where sales are less.
5. From the clusterings we found that we can classify our customers into 4 categories.
   Cluster 0: High Recency and Low Frequency: Customers in this cluster are those who made a purchase a long time ago and have not made a purchase since then. These are the least valuable customers and may be considered for re-engagement strategies.

Cluster 1: Low Recency and Low Frequency: Customers in this cluster are those who make infrequent purchases and have not made a purchase recently. They may be considered for re-engagement strategies to encourage them to make a purchase.

Cluster 2: High Recency and High Frequency: Customers in this cluster are those who make frequent purchases but have not made a purchase recently. They may be considered for targeted marketing campaigns to encourage repeat purchases.

Cluster 3: Low Recency and High Frequency: Customers in this cluster are those who make frequent purchases and have made a purchase recently. They are the most valuable customers and should be targeted with loyalty programs and personalised offers to encourage them to continue purchasing.

Therefore, Cluster 3 is most important to us.

6. We got 3 clusters while doing K-Means clustering on the features of Quality and UnitPrice. And, 3 clusters in agglomerative using rfm characteristics. We got 1 cluster and some noise with DBSCAN clustering. Hence, K means using rfm is best for customer segmentation.

References

1. https://archive.ics.uci.edu/ml/datasets/online+retail
2. Stack Overflow
3. https://www.kaggle.com/datasets/lakshmi25npathi/online-retail-dataset
4. https://towardsdatascience.com/customer-segmentation-using-k-means-clustering-d33964f238c3
5. https://neptune.ai/blog/customer-segmentation-using-machine-learning