PRML MINOR PROJECT

PREPARED BY

Abhaymani Singh(B21EE001)

Garvit Gangwal(B21EE019)

Gaurav Naval(B21EE020)

1. Project Overview

We are given the **Retail Dataset** and in this Minor Project we analyse the dataset using different concepts and implement an end-to-end machine learning pipeline for the task given in the project.

2. Preprocessing

We start with the preprocessing of the Retail Dataset which is the first and foremost step while implementing any machine learning pipeline.

1) All the rows containing **null values** are removed from the dataset making the dataset smaller (406829 x 8)
2) From **“Invoice Date”** date,year and month are separated and made into new columns to perform monthly and yearly revenue analysis.
3) **Duplicate Values** along with **negative values** were removed from the the Dataset making the dataset smaller(392692 x 11)
4) **Revenue** for each row is calculated and a new dataframe is made with a year-month revenue column for ease of revenue analysis.

3. Visualisation(Dataset)

The goal of visualisation is to make a complex dataset more understandable.

1) A line+scatter plot of **monthly revenue** is obtained with year-month being on the x axis and revenue on the y axis.
2) Similarly a **monthly growth rate** (line + scatter) plot is obtained.
3) Barplot of “**top 10 customers by sales**” and “**top 10 products by sale**” is plotted.
4) A **countplot** with variable countries is plotted showing the frequency of occurrence of different countries.

5) Count plot of **sales by year** and **sales by month** is also plotted.here months are mentioned on the x axis with encoding.
6) Line Plot of **sales by date** is drawn with revenue on the y axis and dates being mentioned on the x axis

4. Machine Learning Algorithms

Through this Minor project we have used various ML algorithm for analysis which include the following:

1) **K Means Clustering** : K-means is an unsupervised machine learning algorithm used for clustering similar data points together in a dataset.
2) **Hierarchical Clustering** : hierarchical clustering is a bottom-up approach that builds a tree-like hierarchy of clusters, known as a dendrogram, without the need to specify the number of clusters in advance.
3) **PCA** : PCA (Principal Component Analysis) is a popular **dimensionality reduction** technique used in machine learning to identify patterns and relationships in high-dimensional data.
4) **LDA** : LDA (Linear Discriminant Analysis) is a dimensionality reduction technique that finds a **linear combination** of features which characterises or separates two or more classes of objects or events.

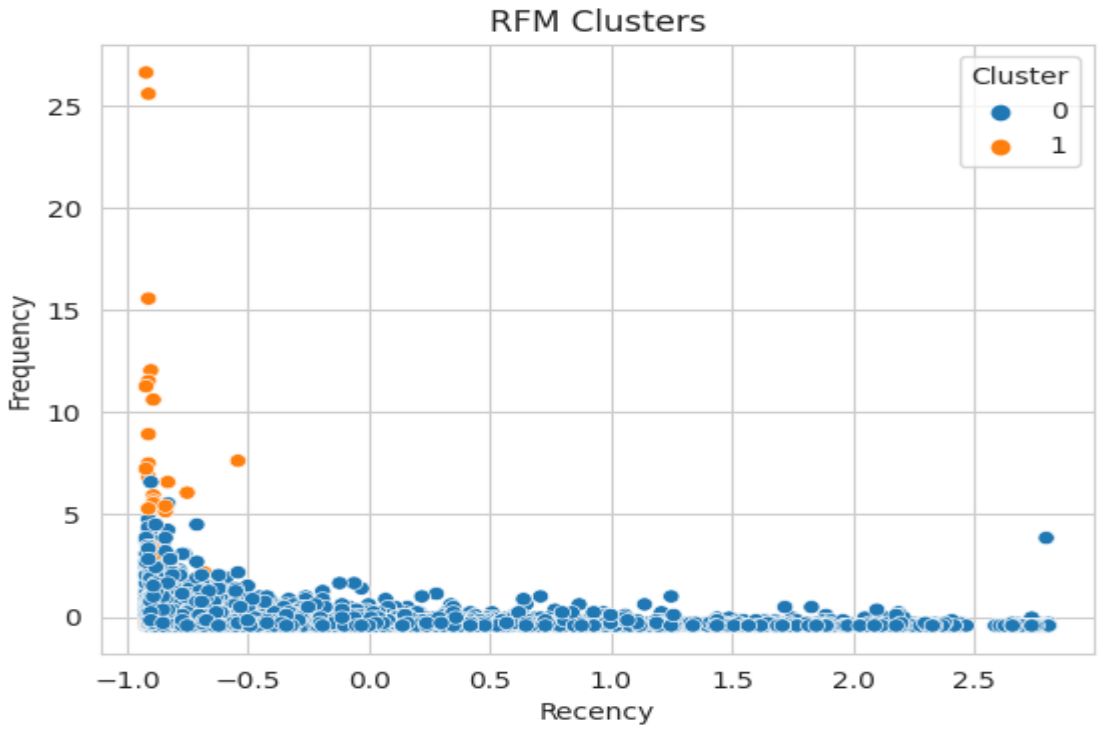
5. Application of ML Algorithms

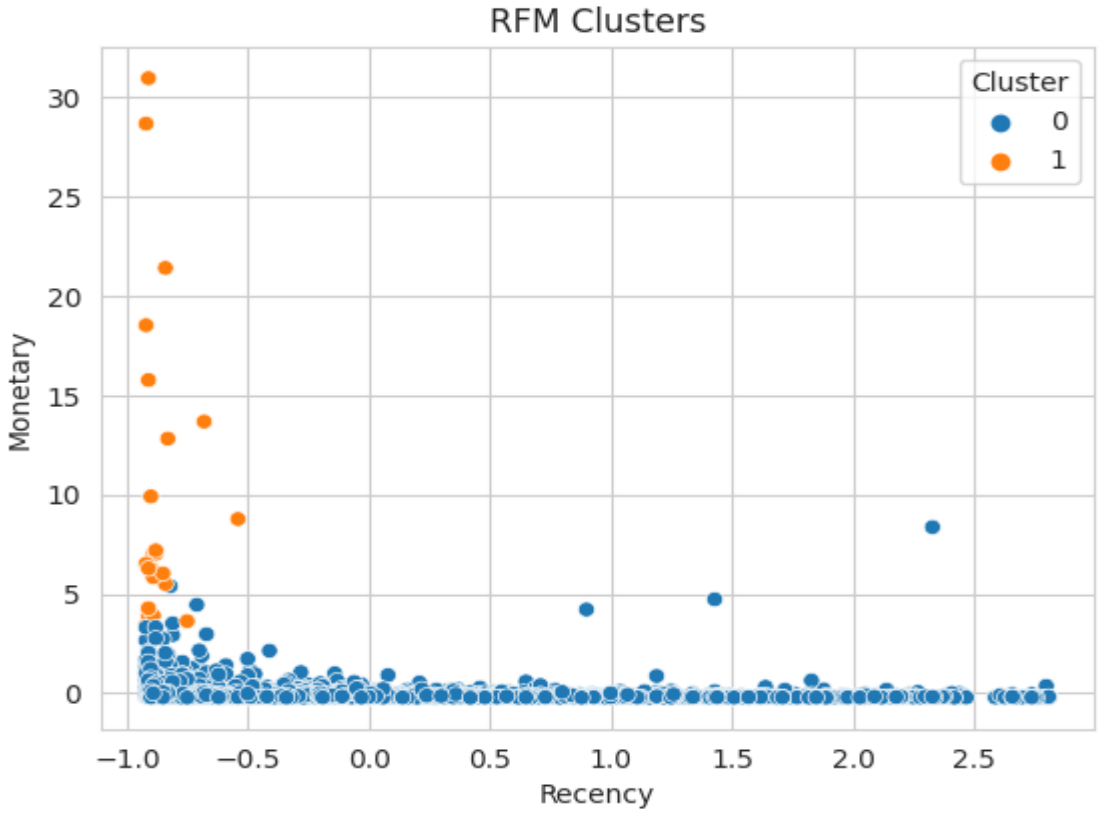
1) K Means

On The Basis Of RFM

- We performed a customer segmentation analysis using the **RFM (Recency, Frequency, Monetary)** model to segment customers based on their purchase behaviour

- RFM data frame is formed and it is standardised to obtained normalised dataframe
- The optimal number of clusters is determined using the elbow method and silhouette score.**(optimal clusters=2,silhouette_score=0.895)**
- K Means algorithm is applied with the optimal number of clusters on the scaled dataset.
- The clusters are added to dataframe rfm_df and then visualised using scatter plots.


RFM Clusters
Cluster
0
1
Frequency
25
20
15
10
5
0
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
Recency


RFM Clusters
Cluster
0
1
Monetary
30
25
20
15
10
5
0
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
Recency

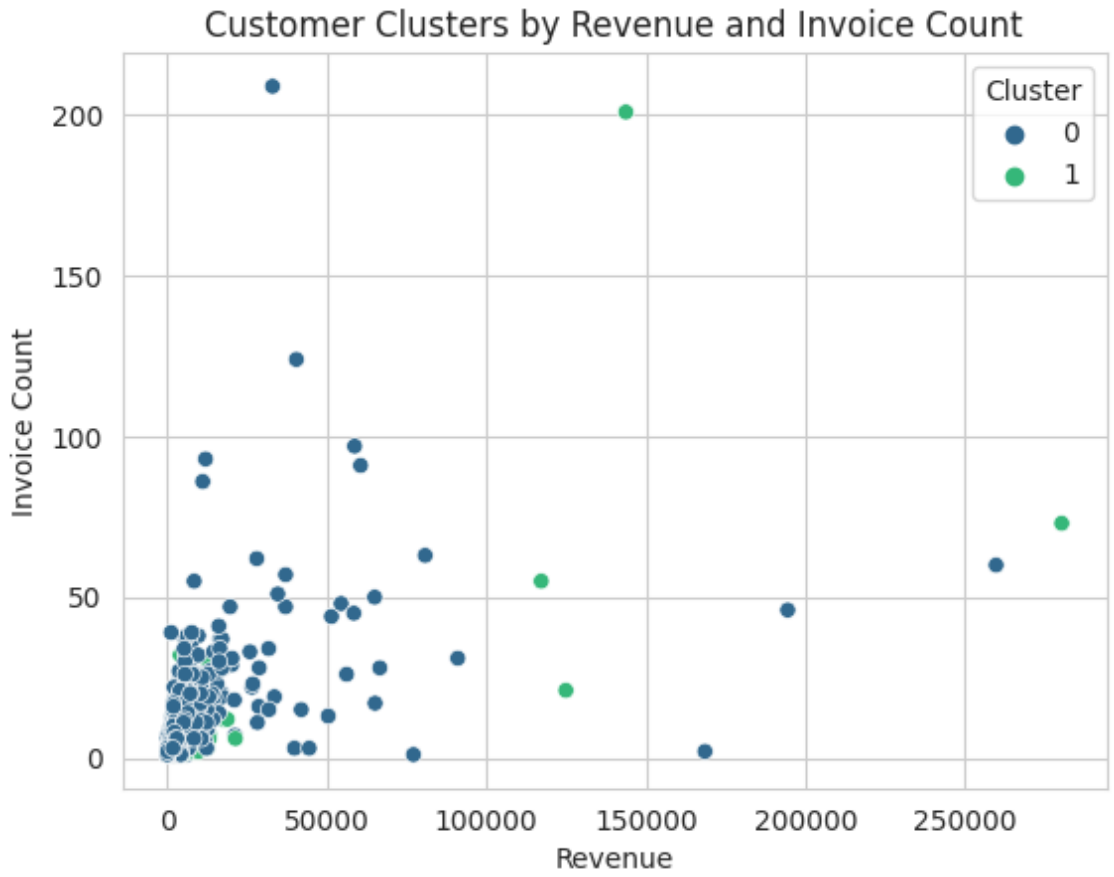
On The Basis Of Quantity

- Clustering on the basis of quantity is performed.Relevant columns like customer id and quantity are extracted and stored in a new dataframe.
- The optimal number of clusters is determined using the elbow method and silhouette score.**(optimal clusters=2,silhouette_score=0.9798)**
- K Means algorithm is applied with the optimal number of clusters on the scaled dataset.
- The clusters are then visualised using scatter plots.

On The Basis Of Country

- We perform clustering analysis on customer data based on two features: '**Revenue**' and '**InvoiceNo**' grouped by country.

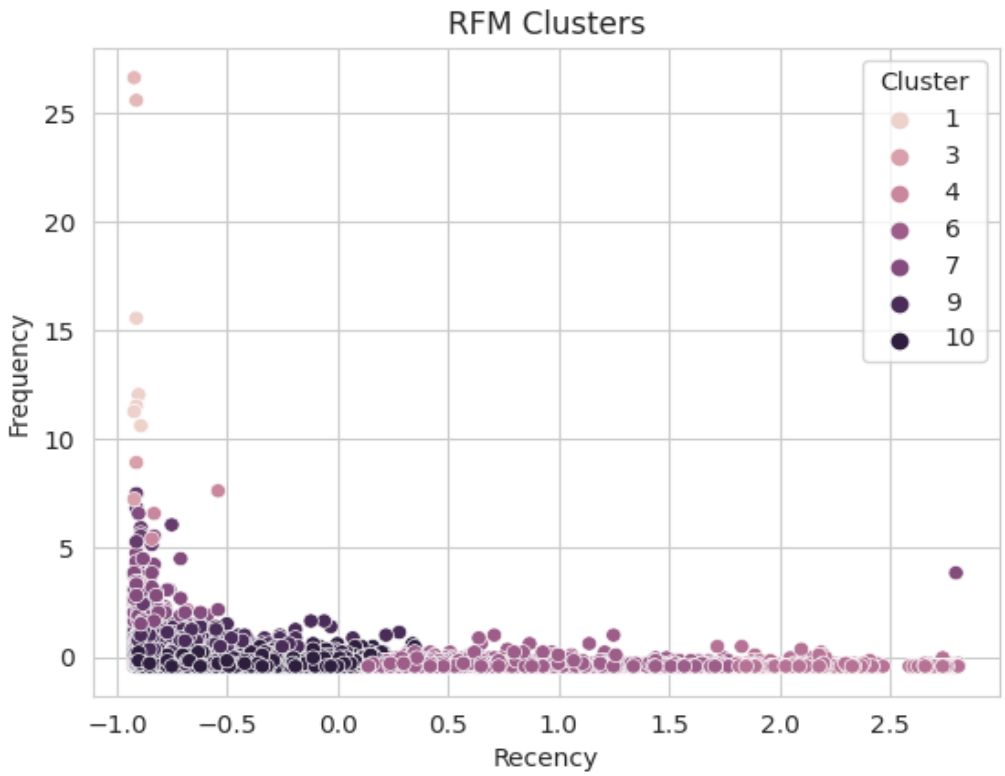
- The necessary features are extracted from the original dataset by grouping the data by 'CustomerID' and 'Country' and calculating the number of unique invoices and total revenue for each customer
- The optimal number of clusters is determined using the elbow method and silhouette score.(**optimal clusters=2,silhouette_score=0.9798**)
- K Means algorithm is applied with the optimal number of clusters on the scaled dataset.
- The clusters are then visualised using scatter plots.

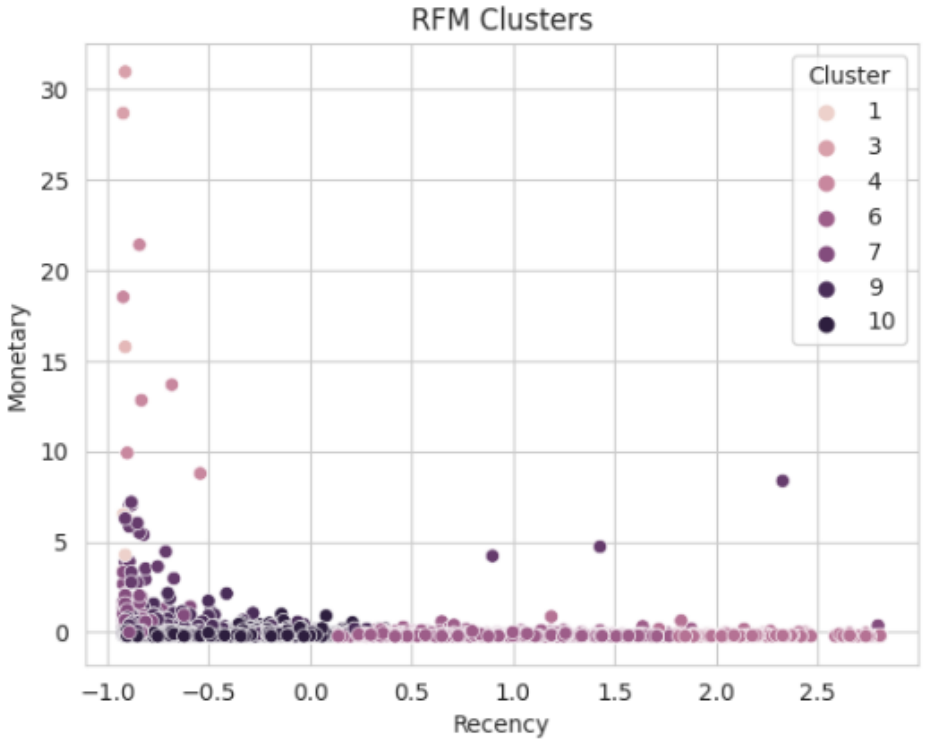


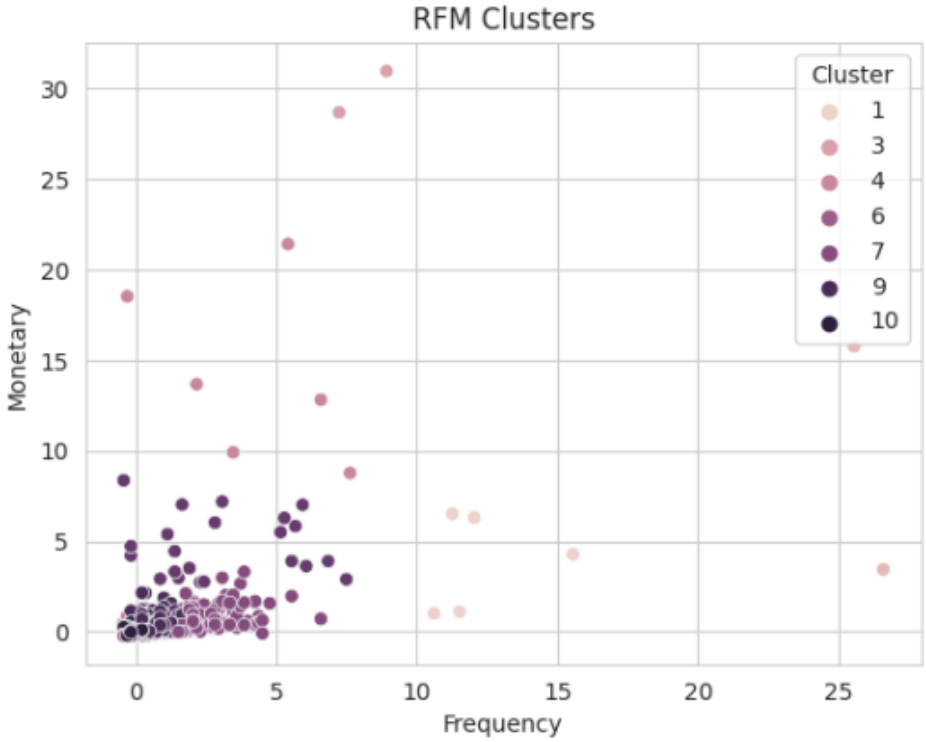
2) Hierarchical Clustering

On The Basis Of RFM

- First, the **linkage matrix** is computed using the linkage() function from the scipy.cluster.hierarchy module.
- Next, the **dendrogram** is plotted using the dendrogram() function. This allows us to visualise the hierarchical structure of the clusters.
- Then, the fcluster() function is used to obtain cluster labels based on a maximum distance threshold.
- Finally, the **cluster labels** are added to the original RFM dataset as a new column called 'clusters'. This allows us to analyse the RFM data by cluster and identify patterns or insights within each cluster.
- Clusters are then visualised using scatter plots.




RFM Clusters
Cluster
1
3
4
6
7
9
10
Monetary
Recency
30
25
20
15
10
5
0
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5

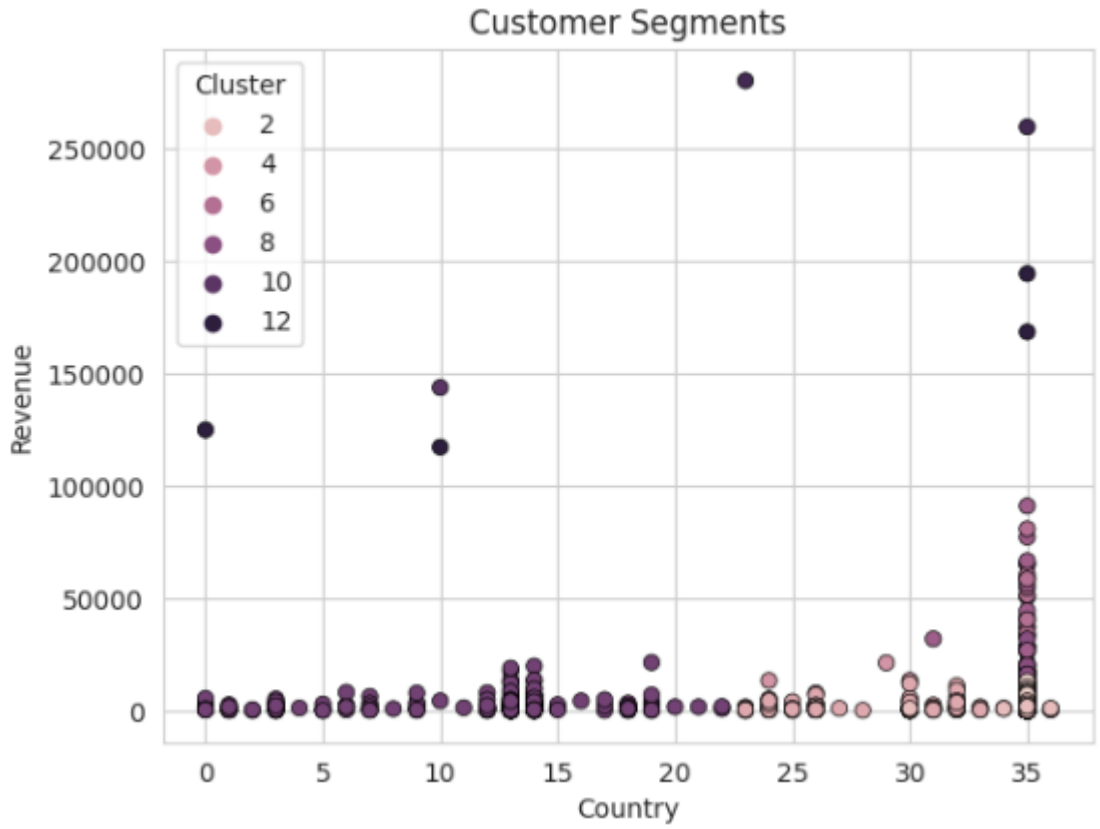

RFM Clusters
Cluster
1
3
4
6
7
9
10
Monetary
Frequency
30
25
20
15
10
5
0
0
5
10
15
20
25

On The Basis Of Quantity

- The **linkage matrix** is computed using the 'ward' method, which minimises the variance of the clusters being merged at each step.
- The **dendrogram** is plotted using the linkage matrix, which shows the hierarchical relationship between clusters and helps to determine the optimal number of clusters.
- The maximum distance **threshold** is set to 20, and the fcluster function is used to obtain **cluster labels** based on the distance between clusters.
- **Cluster labels** are added to the customer data.

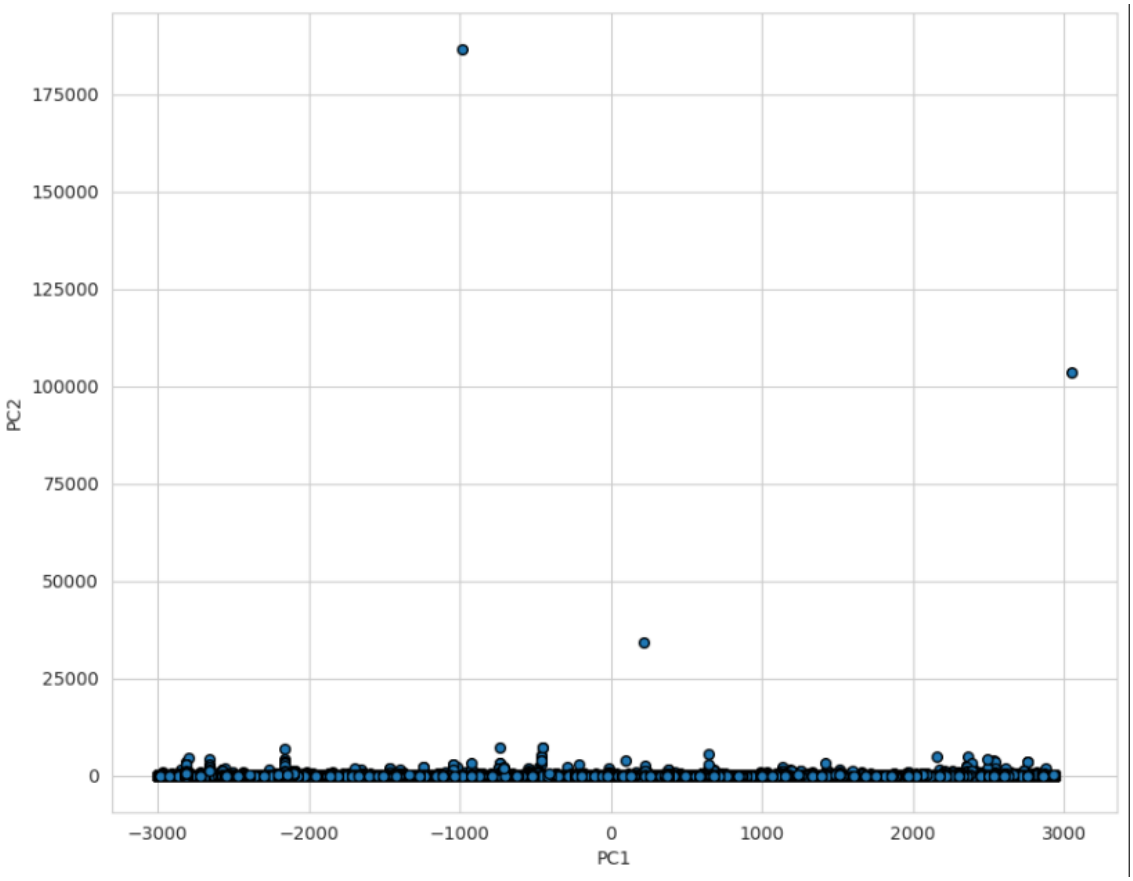
On The Basis Of Country

- It is implemented in the similar manner as above.
- First, the **linkage matrix** is computed using the 'ward' method
- Then the dendrogram is plotted.
- The '**max_d**' variable is set to 20, which is the maximum distance threshold for forming clusters.
- The '**fcluster**' function is used to obtain the cluster labels based on this distance threshold.
- These cluster labels are then added to the customer_data1 dataset as a new feature called 'Cluster'.
- Finally, a scatter plot is created to visualise the clusters.



3) PCA

- **Principal Component Analysis (PCA)** is applied on the online retail dataset and then KMeans clustering is on the reduced dimensional data.
- The optimal number of clusters are determined using the elbow method and plot the resulting graph.**(optimal = 3**)
- The clusters are visualised in a scatter plot with the cluster centres in black.



4)LDA

- We transform the data into two components using the **LinearDiscriminantAnalysis class.**
- KMeans clustering is applied with **three clusters** on the transformed data.
- The clusters are visualised using a scatter plot of the transformed data points.
- The resulting plot shows the clusters after LDA in two dimensions, with the x-axis and y-axis representing the **two LDA components.**