Prithvi Kiran (B21CS001)
Akshat Jain(B21CS005)
Ankana Chowdhury(B21CS010)

PRML Minor Project

Project:

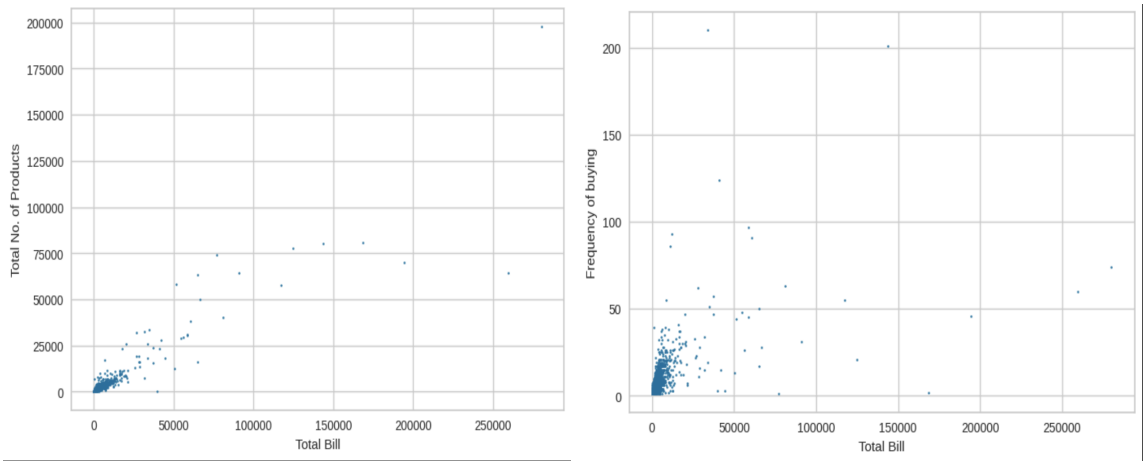
A company that sells some of the product, and you want to know how well the selling performance of the product. You have the data that we can analyze, but what kind of analysis can we do? Well, we can segment customers based on their buying behavior on the market. Your task is to classify the data into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

Objective : The objective of our project is to analyze an online retail dataset and segment customers based on purchasing behavior.We are doing this to help the company understand its customers better , identify high-value customers and tailor marketing strategies to target different segments effectively.

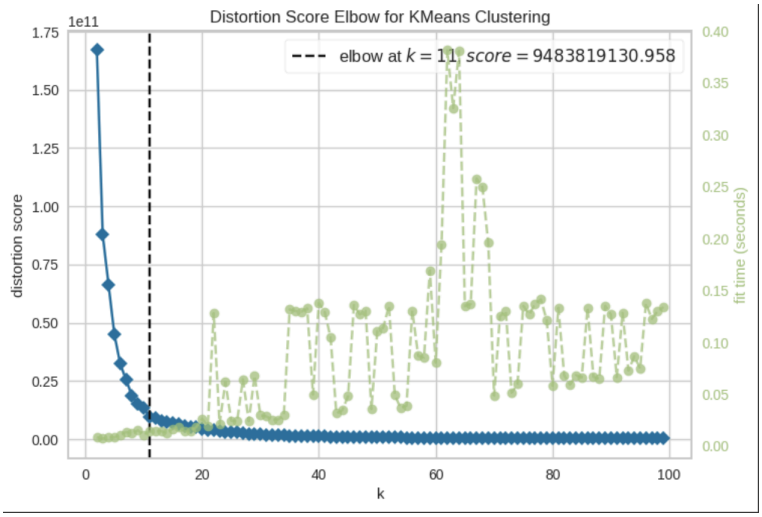
To achieve this objective we have performed the following tasks(/processes) :

1. Data Preprocessing : We have preprocessed the dataset by handling missing values ,we checked the unique items present in the data and also the count of each unique item present in the data, removing canceled orders , and calculating relevant features such as total price per transaction.
2. Feature Aggregation : We have aggregated the data per customer to calculate:
   Recency(X4) : Time since the last highest purchase w.r.t the reference date
   Frequency(X3) : Number of purchases made by a customer
   Monetary Value(X1) : Total amount spent by a customer to make the purchase
3. Created a new dataframe df2 using these aggregated features.
   Rows with negative values in any of the columns were removed from df2.

4. Visualizing the aggregated data using matplotlib.pyplot library.

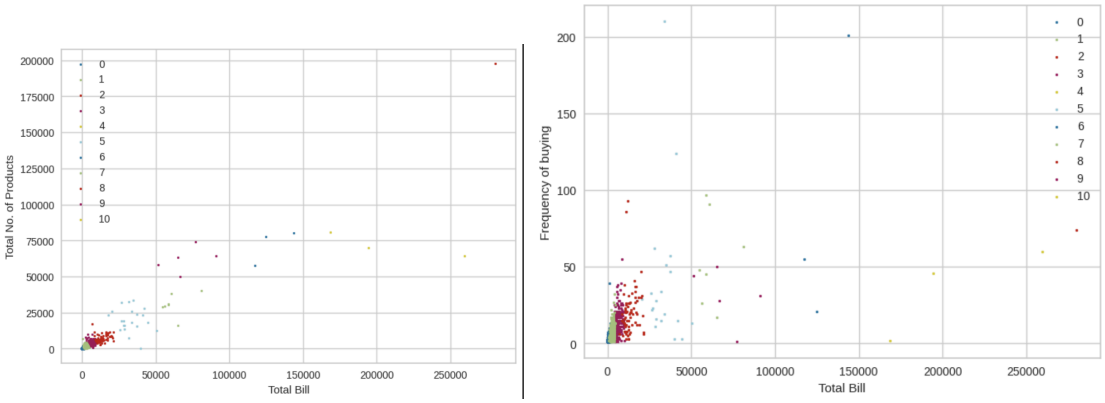


5. KMeans Clustering : KMeans clustering with k(1,30) clusters was applied on the aggregated data (df2).Then we found out the optimal number of clusters required using the “Elbow” Method , which came out to be 11.

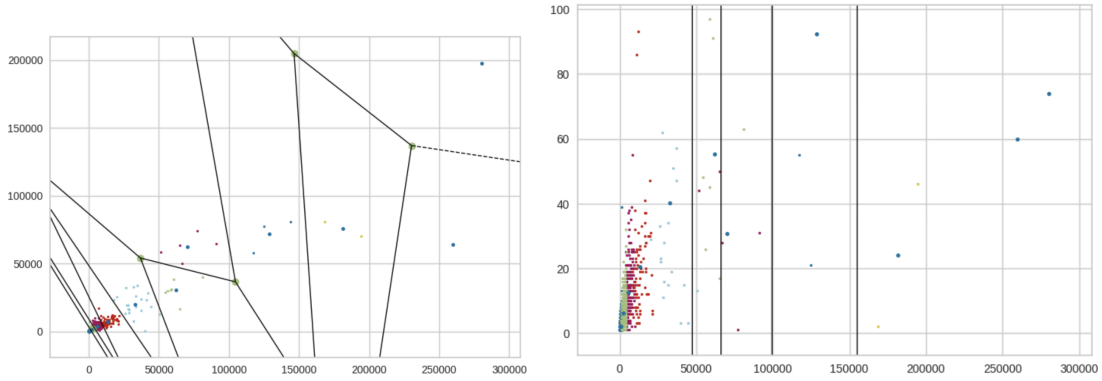


6. After applying KMeans clustering on the aggregated data we visualized the resulting clusters using scatter plots , we plotted the scatterplots for Total Bill vs Total No. of Products and Total Bill vs Frequency of buying with different colors

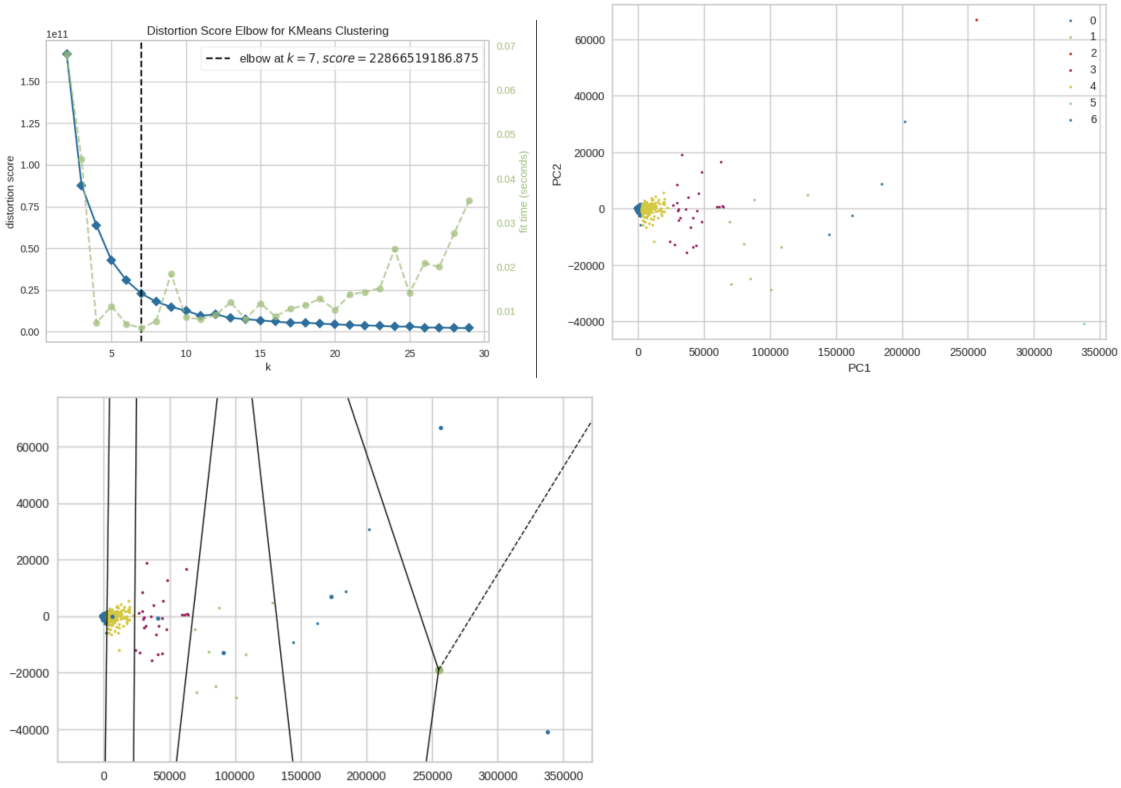
representing each cluster.



7. We plotted Voronoi diagrams for the KMeans clustering results:
   a.Extracted the cluster centers and created Voronoi diagrams using Voronoi function from the scipy.spatial module.
   b.Visualized them using voronoi_plot_2d function



8. Applied PCA with 4 components to the data and plotted the ‘cumulative explained variance’ to determine the optimal number of components, which came out to be 2.
9. Applied KMeans with 7 clusters on PCA transformed data with 2 components and visualized the resulting clusters using scatter plots.



10. RFM :

    a.Created an RFM DataFrame using Recency, Frequency, and Monetary values calculated earlier.

    b.Assigned scores to each customer based on quantiles for Recency , Monetary and Frequency values : [Used ‘pd.qcut()’ function to assign Recency, Frequency and Monetary scores based on quantiles]

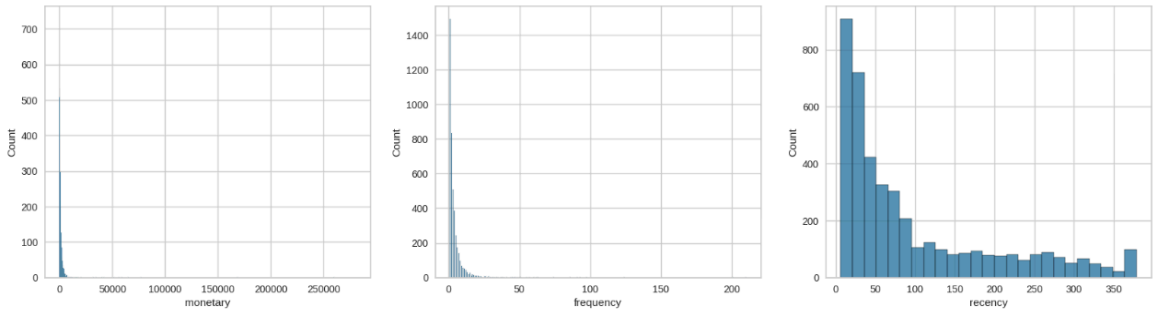
    (1).Recency was divided into three bins (3: most recent , 1: least recent)

    (2).Monetary Value was divided into three bins (3: highest spenders, 1: lowest spenders)

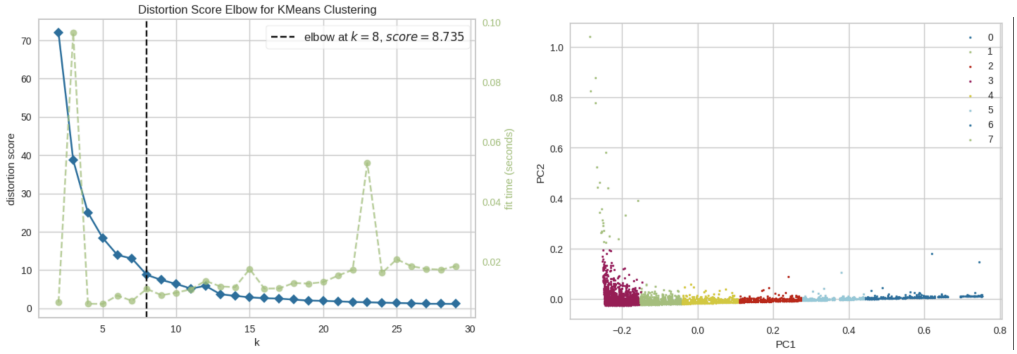
    (3).Frequency was divided into three bins (3: most frequent buyers , 1: least frequent buyers)

11. Combined all these scores to create an RFM score for each customer - created a new column ‘rfm_score’ in ‘rfm’ DataFrame. (the scores were concatenated as strings so that they are in a more readable format this helps in identifying the customer’s behavior easily)

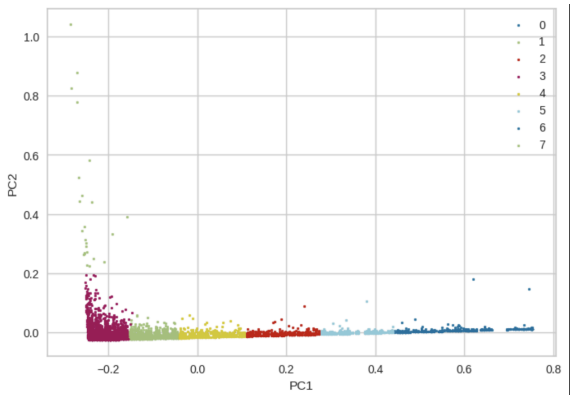
12. Histograms of RFM Values : Plotted histograms for Recency, Frequency and Monetary values to visualize their distribution across the customers.



13. KMeans Clustering on Scaled RFM Data :
    a.Applied MinMaxScaler to scale the RFM data (‘rfm_scaled’)
    b.Preformed KMeans clustering again to find an optimal number of clusters using the “Elbow” Method.

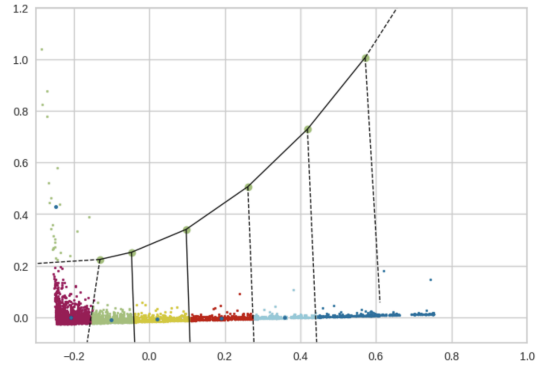


14. PCA on Scaled RFM Data :
    a.Applied PCA with 2 components to ‘rfm_scaled’ the scaled RFM data to reduce its dimensionality.
    b.Used the “Elbow” method to find the optimal number of clusters for KMeans clustering on the PCA transformed RFM data (‘d2’).
15. KMeans Clustering with 8 clusters on PCA transformed RFM Data (‘d2’).

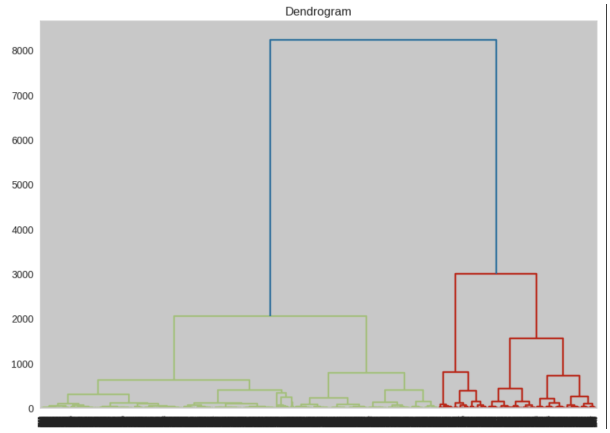


16. Voronoi DIagrams for KMeans Clustering Results on PCA transformed RFM data (‘d2’):

a.Extracted cluster centers and created Voronoi diagrams using the ‘Voronoi’ function from the ‘scipy.spatial’ module.
b.Visualized them using the ‘voronoi_plot_2d’ function.



17. Hierarchical CLustering with Dendrogram :
    a.Selected Recency and Frequency scores from RFM DataFrame and stored them in ‘selected_data’.
    b.Performed agglomerative hierarchical clustering using the ‘linkage’ function from the ‘scipy.cluster.hierarchy’ module with “Ward’s” Method for linkage criterion and Euclidean distance metric.
    c.Plotted a dendrogram using the ‘dendrogram’ function from the module.
    [The dendrogram shows how individual customers are grouped into clusters (represented by branches) that are progressively merged into clusters as we move up the tree.]



18. Performance evaluation was done using elbow method and silhouette score.

Conclusion :

In this project, we analyzed an online retail dataset and segmented customers on their purchasing behavior using various techniques, including KMeans clustering, PCA dimensionality reduction , RFM analysis and Hierarchical Clustering.The company

benefits a lot from this as segmentation of customers helps the company tailor its marketing strategies, improve product offerings and enhance overall business performance.

We applied KMeans clustering on both the original aggregated data and PCA-transformed data to identify customer segments.We also visualized these clusters using scatter plots , histograms and Voronoi Diagrams to gain insights into the characteristics of each segment.

We even performed RFM analysis by assigning Recency, Frequency, and Monetary scores to each customer based on quantiles (these are values that divide the dataset into equal intervals with each interval containing the same proportion of data). Then we combined these scores to create a single RFM score for each customer that helped us in easier segmentation and analysis of customers based on their overall purchasing behavior.

We also applied PCA on scaled RFM data and performed KMeans clustering with an optimal number of clusters determined using the “ Elbow” Method.This method allowed us to visualize and cluster customers in lower dimensional space (which usually helps to create more distinct and interpretable segments).This method helps us view patterns and relationships between customers that might not be as evident when working with higher dimensional or non-transformed data.

Finally we explored Hierarchical Clustering as an alternative method (or another method) to segment customers.After visualizing the dendrogram we were able to visualize how customers were clustered together based on their similarity in Recency and Frequency Scores.Visualizing the dendrogram provided an additional perspective on customer segmentation beyond KMeans clustering and PCA dimensionality reduction.It allowed us to explore hierarchical relationships between customer groups based on their RFM scores and gain further insights on how these groups could be further segmented. (which can be useful for additional targeted marketing strategies)

Through these techniques we successfully segmented customers based on their purchasing behavior.These insights can help tailor marketing strategies , improve product offerings (develop new products or increase production of specific products that appeal to specific segments of customers) and enhance overall business performance.

What can be done in the future to further improve the segmentation (on similar datasets) ? :

We could explore additional clustering algorithms or feature extraction techniques to further improve customer segmentation and incorporating domain knowledge or external factors like seasonal trends could provide a more comprehensive understanding of customer behavior.