ML Lab Week 13

Name:

Maya Nithyanand Bhagath

SRN:

PES2UG23CS331

Section:

F

Date:

13/11/2025

Analysis Questions

1. Dimensionality Justification: Based on the correlation heatmap and explained variance ratio from PCA, why was dimensionality reduction necessary for this dataset? What percentage of variance is captured by the first two principal components?

The heatmap shows that the dataset contains many features. A high-dimensional space can make clustering algorithms struggle due to the curse of dimensionality, and makes visualization very difficult. Therefore, dimensionality reduction was necessary for this dataset.

The first principal component captures about 15% of the total variance, and the second principal component captures about 13%. In total, they capture approximately 28% of the total variance.

2. Optimal Clusters: Looking at both the elbow curve and silhouette scores, what is the optimal number of clusters for this dataset? Justify your answer using both metrics.

The optimal number of clusters for this dataset is 3. In the elbow curve, we can see that this is where the curve bends sharply, and the rate of decrease slows down after this point. The average Silhouette score is 0.39 for this, which indicates that the clusters are reasonably well-separated.

3. Cluster Characteristics: Analyse the size distribution of clusters in both K-means and Bisecting K-means. Why do you think some clusters are larger than others? What might this tell us about the customer segments?

The cluster sizes for the 3 clusters are 11350, 20156, and 13705. Cluster 1 is larger than Clusters 0 and 2. This could indicate that Cluster 1 has a large core segment that represents the most common user profiles. Clusters 0 and 2 could represent smaller and more niche segments.

4. Algorithm Comparison: Compare the silhouette scores between K-means and Recursive Bisecting K-means. Which algorithm performed better for this dataset and why do you think that is?

The Silhouette scores for K-means and Recursive K-Means were 0.39 and 0.29 respectively. This means that the K-means performed better, and its clusters were more compact and well-separated. Recursive Bisecting K-Means probably did worse since the data does not have a strong hierarchical structure.

5. Business Insights: Based on the clustering results in the PCA space, what insights can you draw about customer segmentation that might be valuable for the bank's marketing strategy?

Based on the clustering results in the PCA space, we can see that there are three main customer segments based on the nine original features, with the centroids of each cluster clearly visible. The clusters are mostly spread across the horizontal axis, , confirming that PC1 is the most important dimension for separating the segments.

The yellow cluster is largest and most densely packed, indicating that this corresponds to a large core segment. The yellow and purple segments are more concentrated, suggesting that they represent more niche segments.

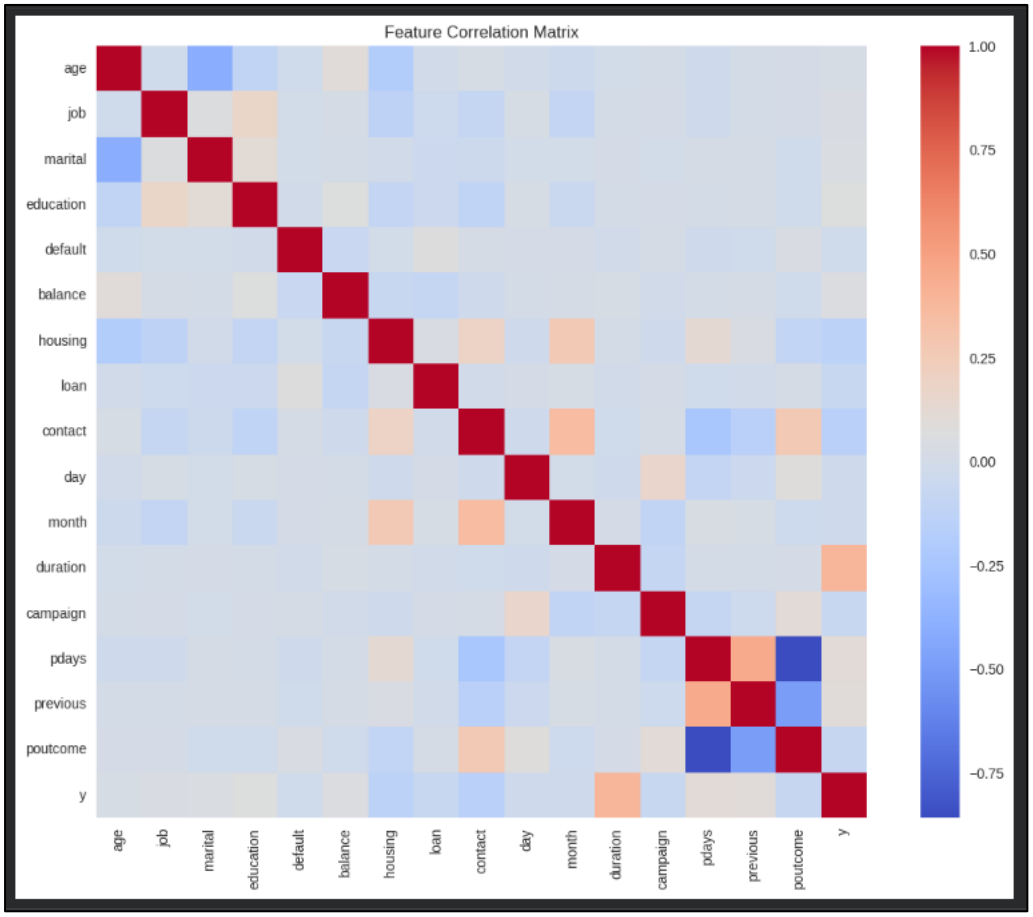
6. Visual Pattern Recognition: In the PCA scatter plot, we see three distinct coloured regions (turquoise, yellow, and purple). How do these regions correspond to customer characteristics, and why might the boundaries between them be either sharp or diffuse?

The three distinct coloured regions are defined by the principal components PC1 and PC2. The points are derived from the nine normalized features, and the location of each point corresponds to the weighted value of those original features.

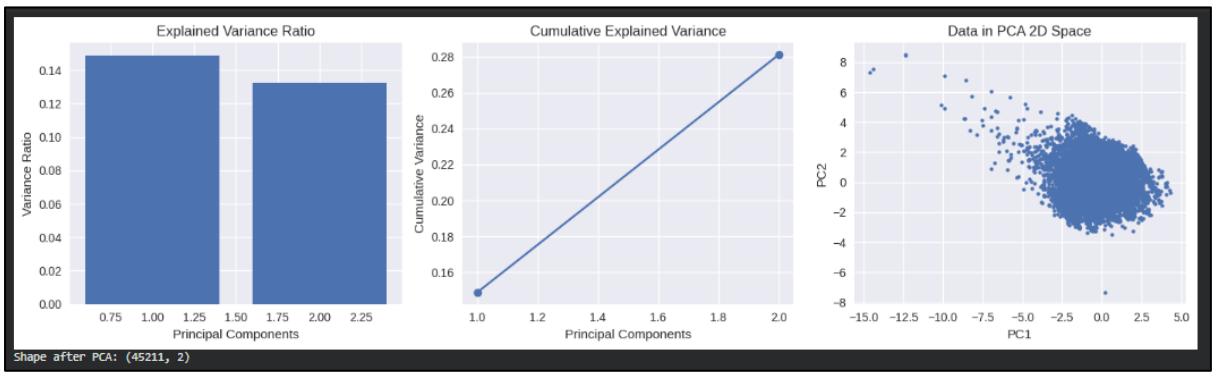
The boundaries between the regions might be sharp since each point is assigned to the closest centroid. However, some data distribution is diffuse. The points belonging to different clusters overlap significantly in the centre of the graph. This indicates that the customer segments are not perfectly well-separated.

Screenshots

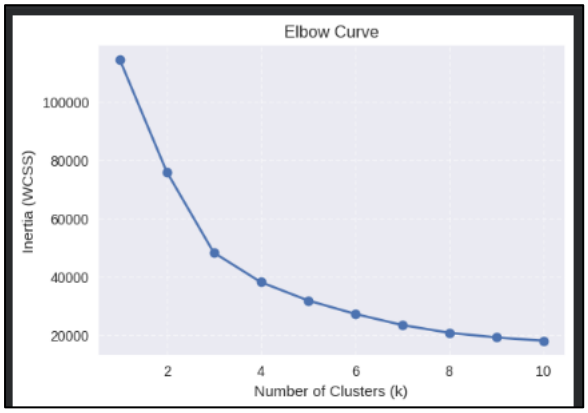
1. Feature Correlation matrix for dataset

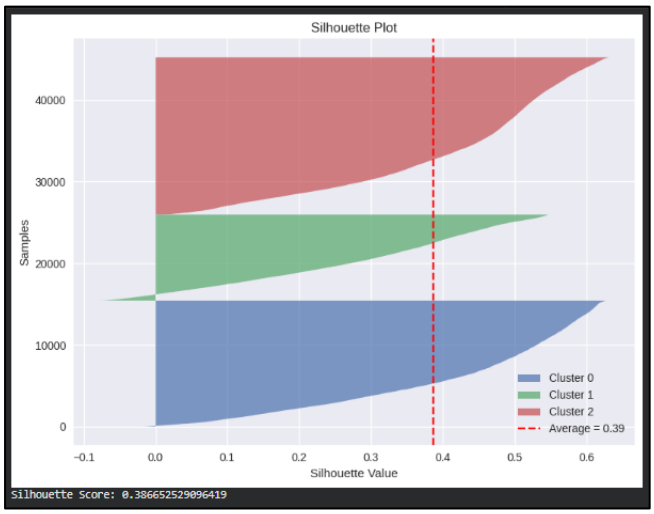


2. 'Explained variance by Component' and 'Data Distribution in PCA Space' after Dimensionality Reduction with PCA



3. ‘Inertia Plot’ and ‘Silhouette Score Plot’ for K-means





4. K-means Clustering Results with Centroids Visible (Scatter Plot)