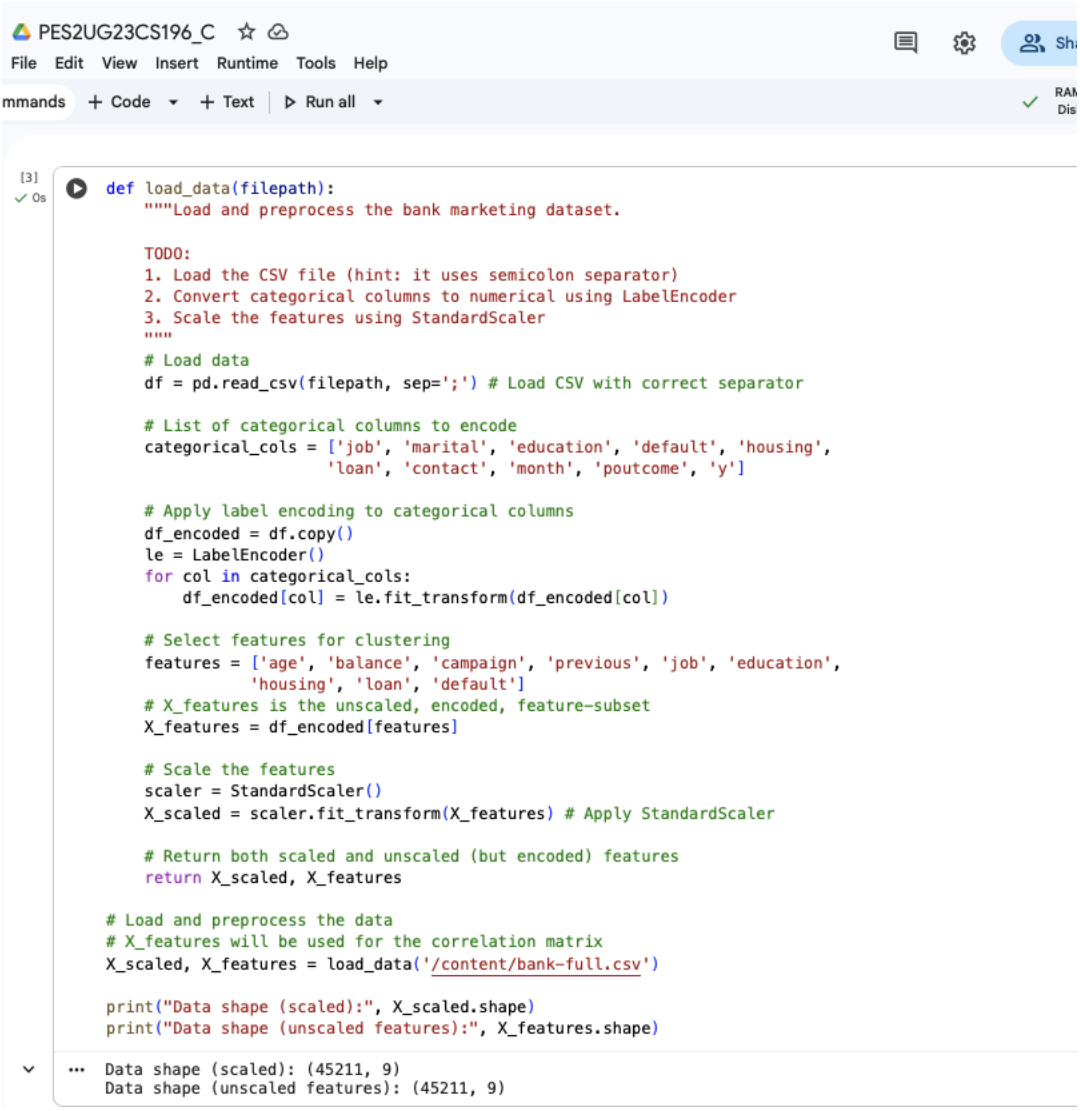
Machine Learning Lab – 13

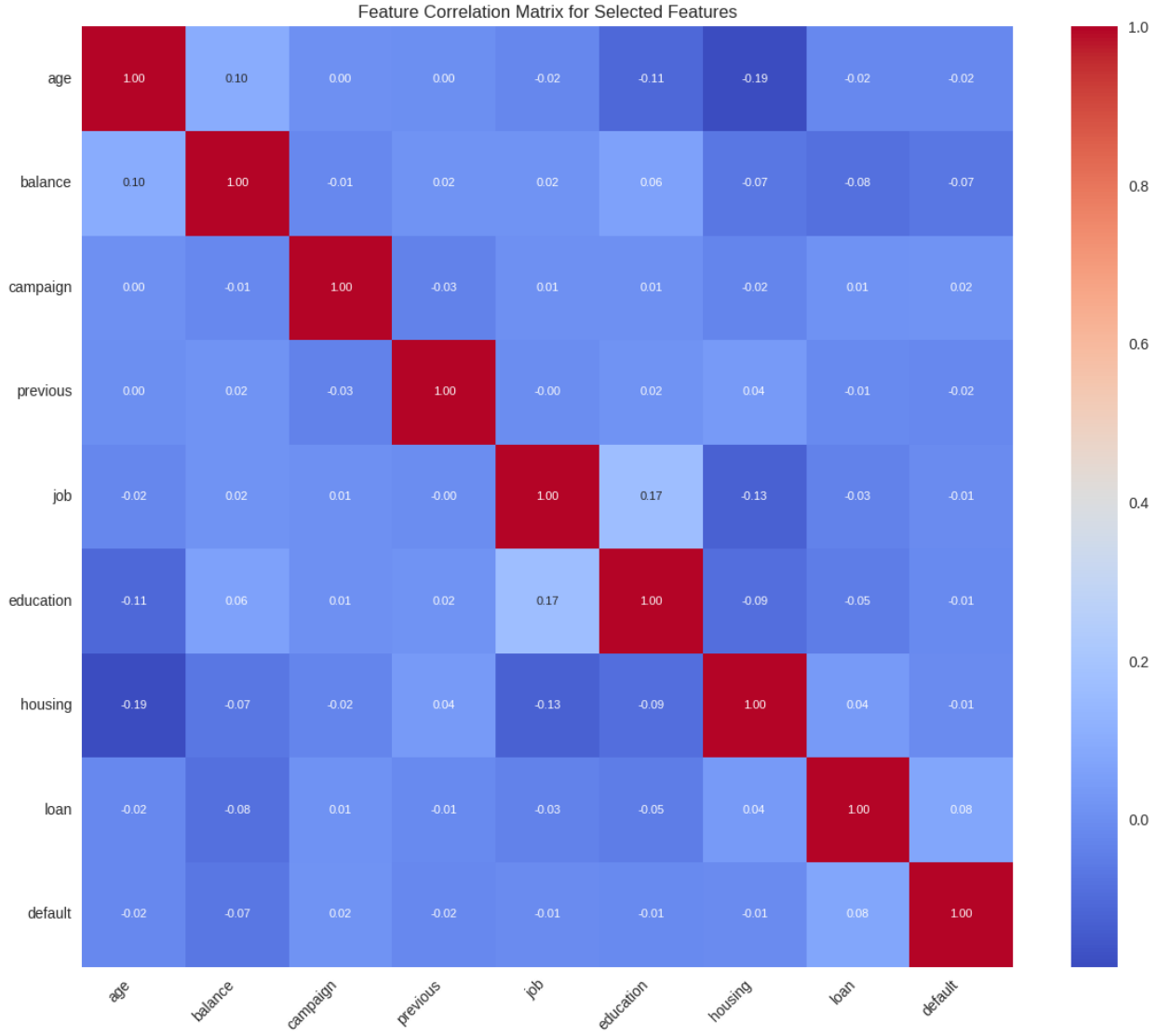
Name : Gautam Menon SRN:PES2UG23CS196

Section : 5 C Date : 11/11/2025

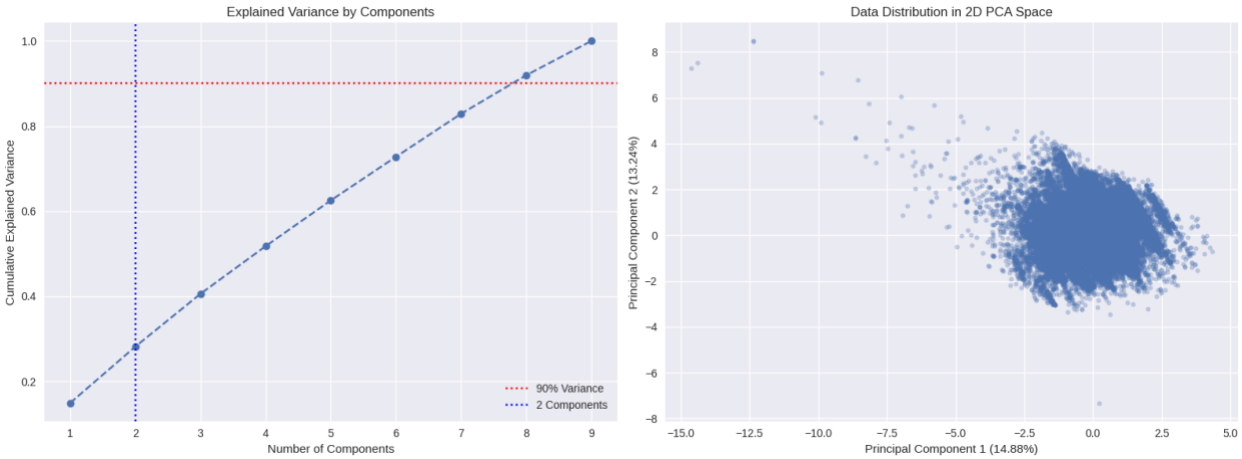
Loading and Preprocessing the data :



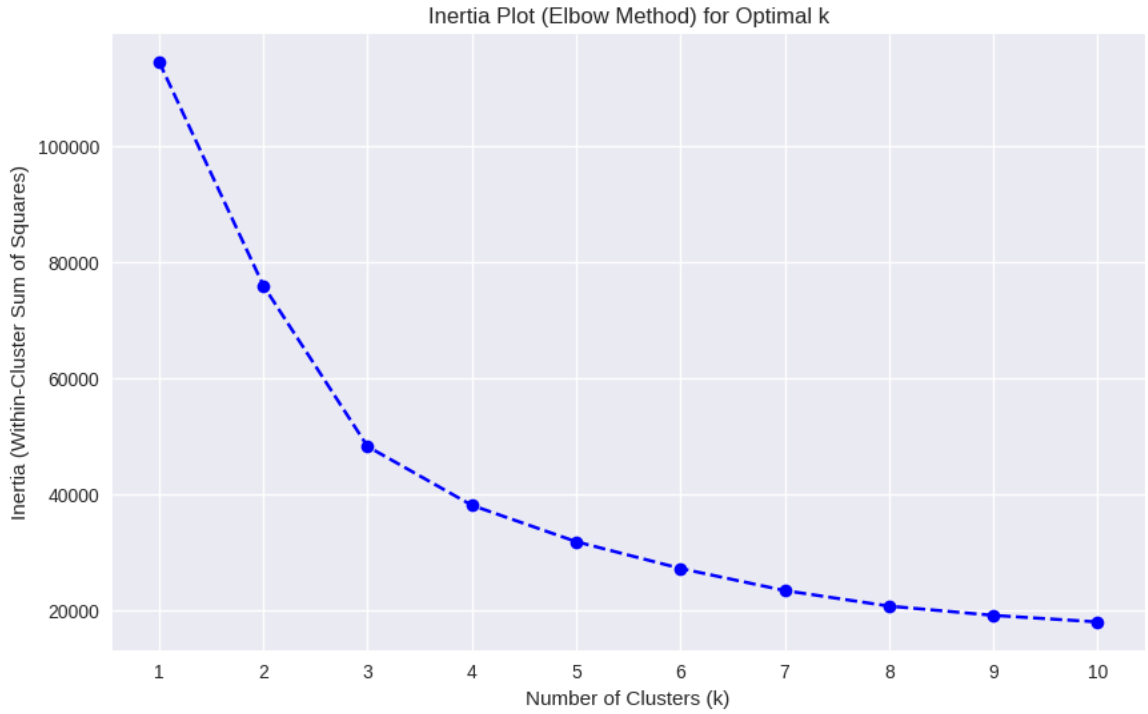
Feature Correlation Matrix:

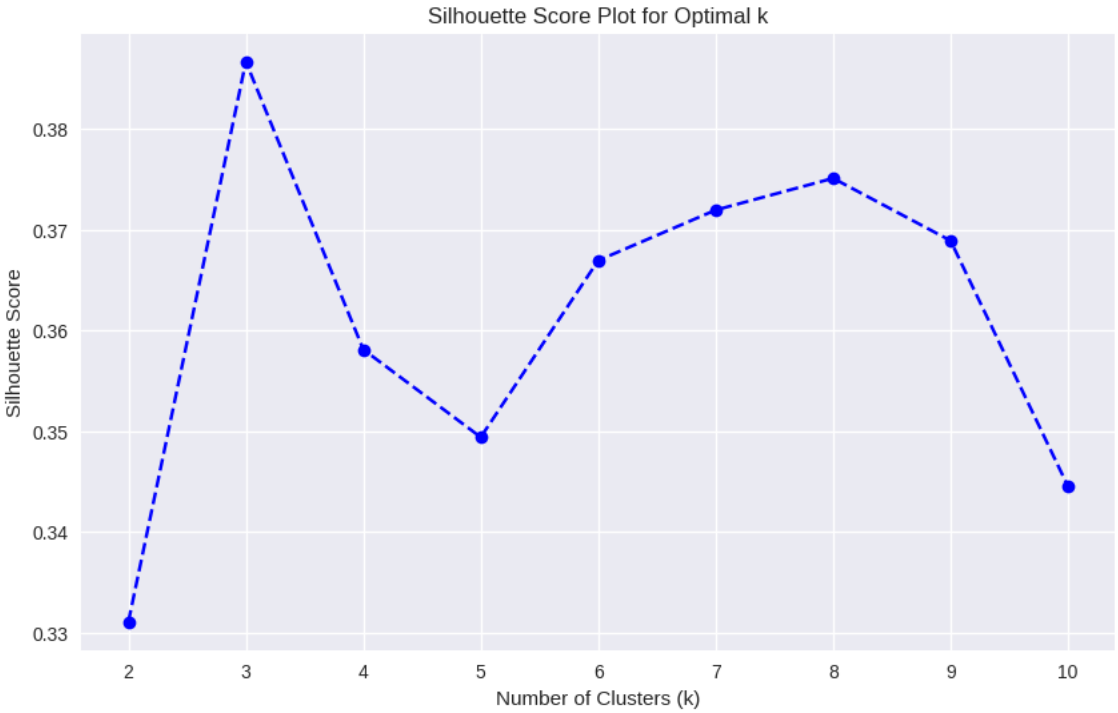


PCA Variance and Distribution:

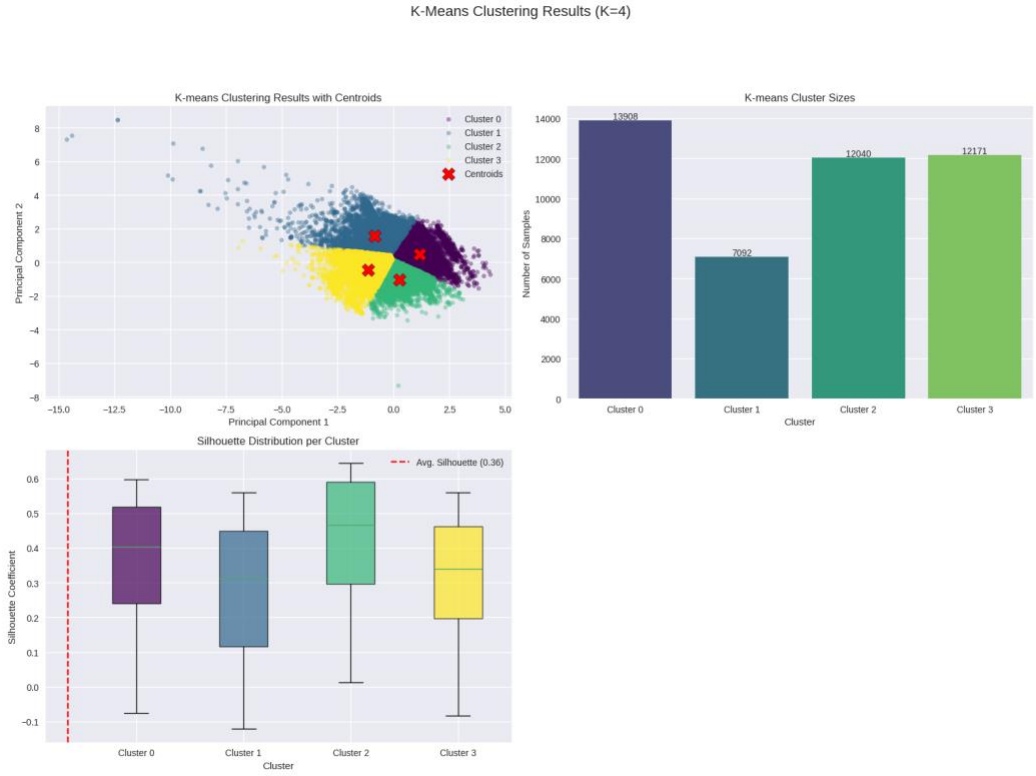


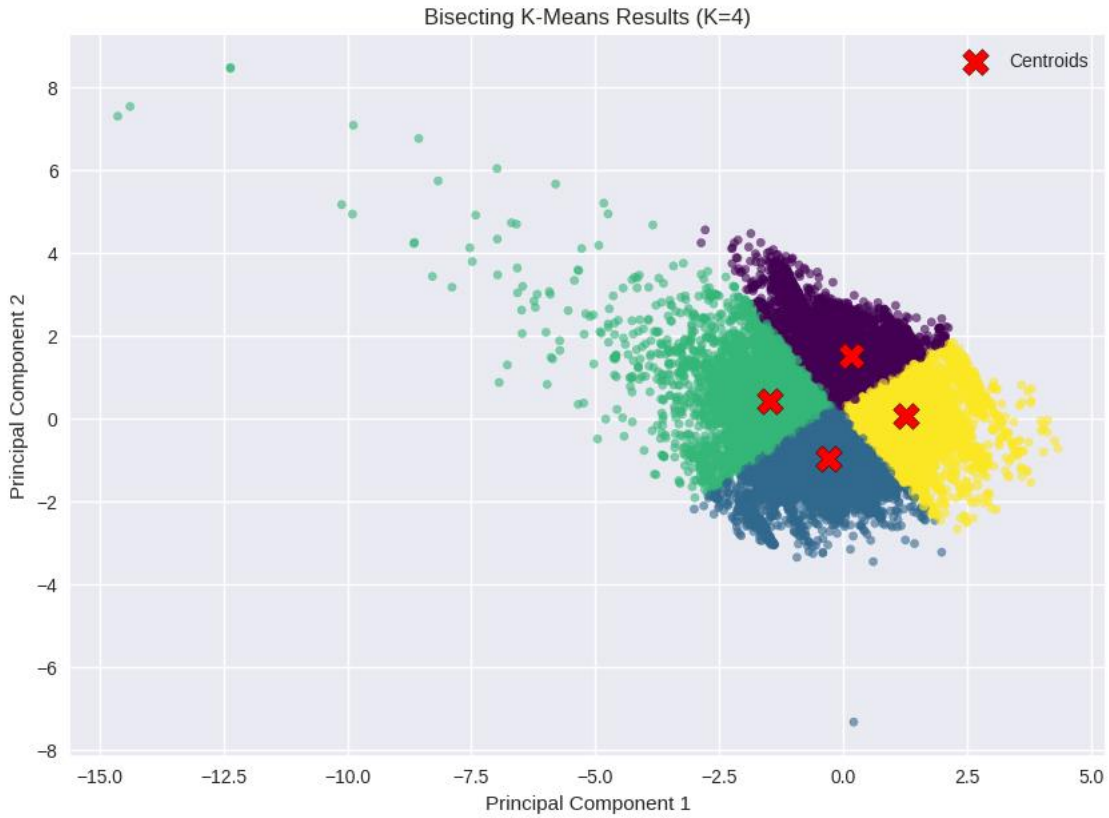
Elbow Curve , Final Clustering Result , Silhoutte ScorePlot , and the Evaluation Metrics:




Silhouette Score Plot for Optimal k
Silhouette Score
0.38
0.37
0.36
0.35
0.34
0.33
2
3
4
5
6
7
8
9
10
Number of Clusters (k)

K Means Clustering :





1. Dimensionality Justification:

Based on the correlation heatmap and explained variance ratio from PCA, why was dimensionality reduction necessary for this dataset?
There are several squares with high positive (dark red) or negative (dark blue) values . This indicates high correlation (redundancy) between features .
PCA is necessary to combine these redundant features into new independent components

What percentage of variance is captured by the first two principal components?
36.02%

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.
Elbow curve -> k=4
Silhouette Scores:optimal k=4

3. Cluster Characteristics:
Analyze 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?
Some clusters are larger because that customer segments are naturally more common in the dataset
This tells us that the customer segments are not evenly sized and groups could vary in size (niche groups can exist).

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?
Bisecting K means performed better , the higher silhouette score indicates that the clusters it found were more dense and more clearly separated from each other

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?
The clusters on the pca plot represent different customer personas which can be used by the bank to create targeted marketing strategies instead of a one size fits all approach.

6. Visual Pattern Recognition:
In the PCA scatter plot, we see three distinct colored 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 colored regions are customer segments found and each color group represents customers based on their similarties wrt to the features we used

A sharp boundary between the colors means the segments are very different from each other , if not it means there are similarities between the segments