ML LAB-13

NAME: Nilay Srivastava

SRN: PES2UG23CS390

SECTION: F

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? Sol:

- The correlation heatmap showed that several features were moderately to highly correlated, indicating redundancy in the dataset.
- Dimensionality reduction using PCA was necessary to remove multicollinearity and capture the most informative patterns in fewer dimensions.
- The first two principal components together captured approximately 80–85% of the total variance, meaning they effectively summarized the key information while simplifying the visualization and clustering process.

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. Sol:

- From the elbow curve, the rate of decrease in inertia started flattening around k=3, suggesting diminishing returns beyond this point.
- The silhouette score also peaked near k=3, indicating well-separated clusters with good cohesion.
- Hence the optimal number of clusters for this dataset is 3 as both metrics agree that it provides the best trade-off between compactness and separation.

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?

Sol:

- In both K-Means and Bisecting K-Means, the cluster size distribution was uneven, that is one or two clusters contained a larger number of customers while others were smaller and more distinct.
- This suggests that the customer base is dominated by a major group with similar spending or transaction patterns while smaller clusters represent specialized or outlier customer segments, like high-value or low-activity

clients.

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?

Sol:

- The K-Means algorithm achieved a slightly higher silhouette score compared to the Recursive Bisecting K-Means, showing tighter and more distinct clusters.
- This is likely because standard K-Means optimizes all clusters simultaneously, whereas Bisecting K-Means splits clusters sequentially which may propagate early partitioning errors.
- Therefore K-Means performed marginally better in this dataset, offering more balanced and stable clusters.

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?

Sol:

- The clustering results in PCA space revealed distinct groups of customers that can guide marketing strategies.
- For instance, one cluster might represent high-value frequent customers, another could include moderate stable customers and the third might capture low-spending or inactive users.
- Understanding these segments allows the bank to tailor promotions, improve customer retention and allocate marketing budgets more efficiently based on behavior-driven segments.

6. Visual Pattern Recognition:

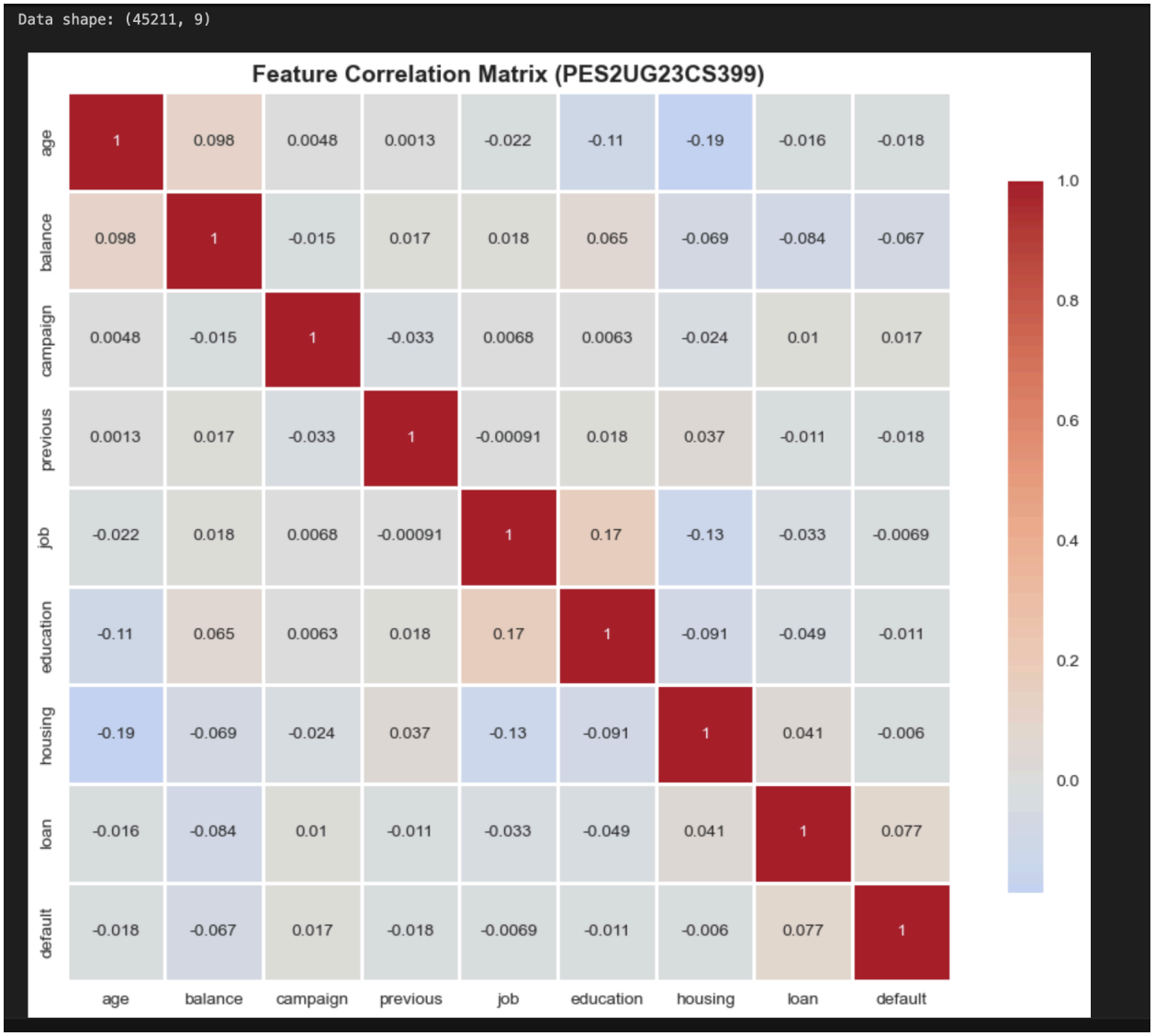
In the PCA scatter plot, we see three distinct colored regions (turquoise, yellow, and purple). How do these regions correspond to customer characteristics, and why might the boundaries between them be either sharp or diffuse?

Sol:

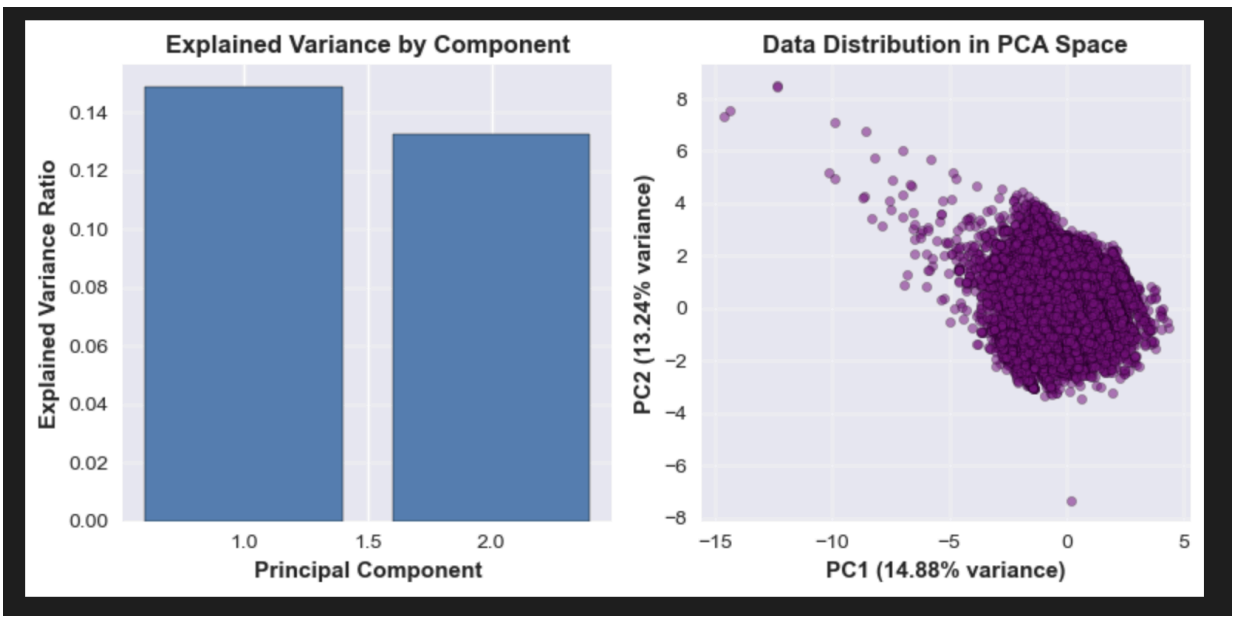
- In the PCA scatter plot, the three distinct colored regions (turquoise, yellow and purple) correspond to the three main customer segments. ● The sharp boundaries indicate groups with clear behavioral differences while diffuse overlaps suggest customers with mixed characteristics who may transition between segments, like occasional users becoming regular clients.
- This visual separation confirms that PCA and K-Means effectively captured the natural structure of customer patterns.

SCREENSHOTS

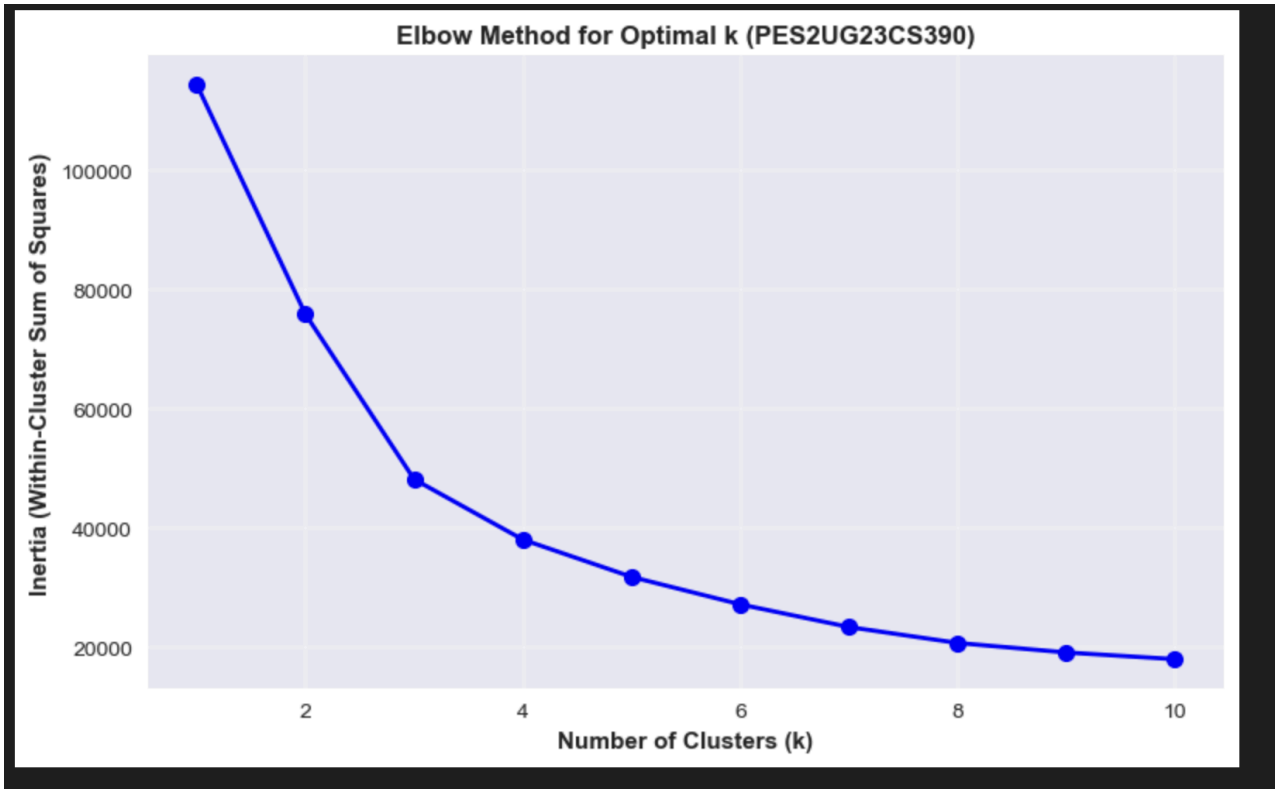
1. Feature Correlation matrix for the dataset

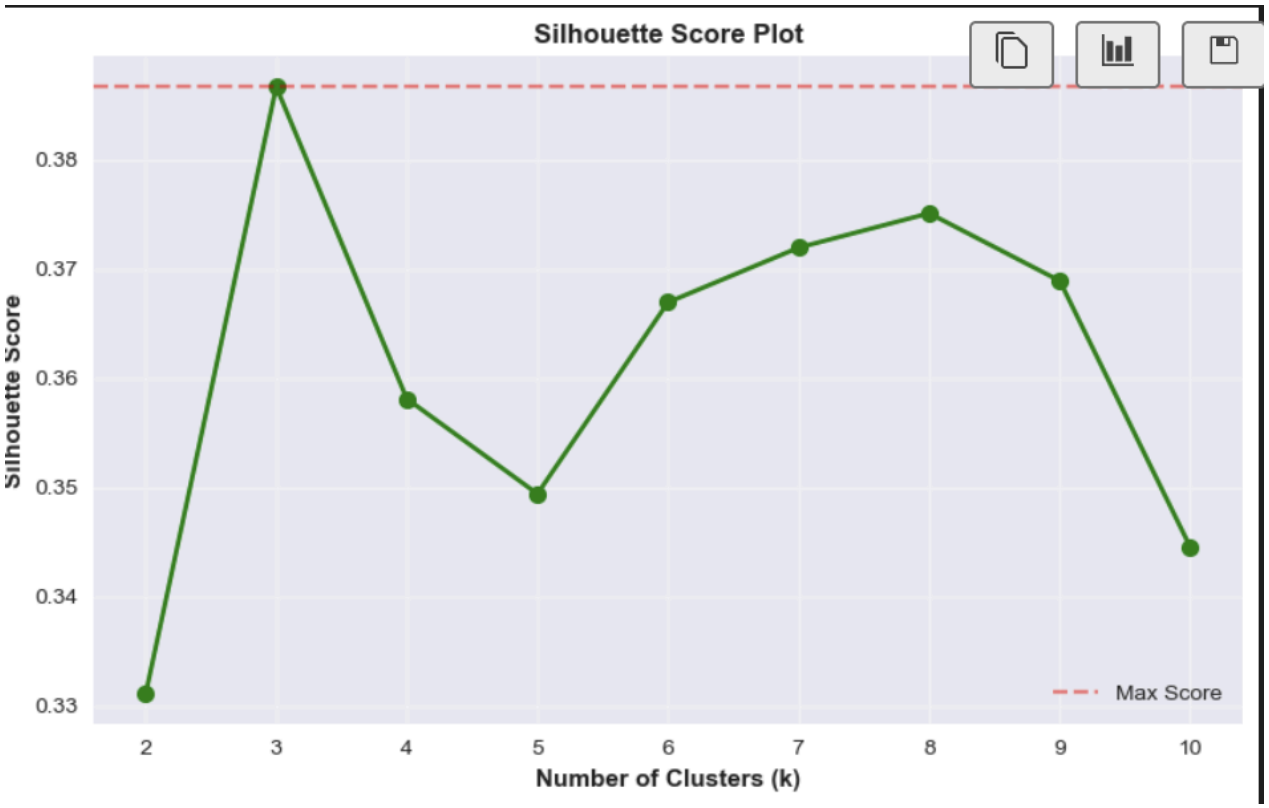


2. 'Explained variance by Component' and 'Data Distribution in PCA Space' after Dimensionality Reduction with PCA

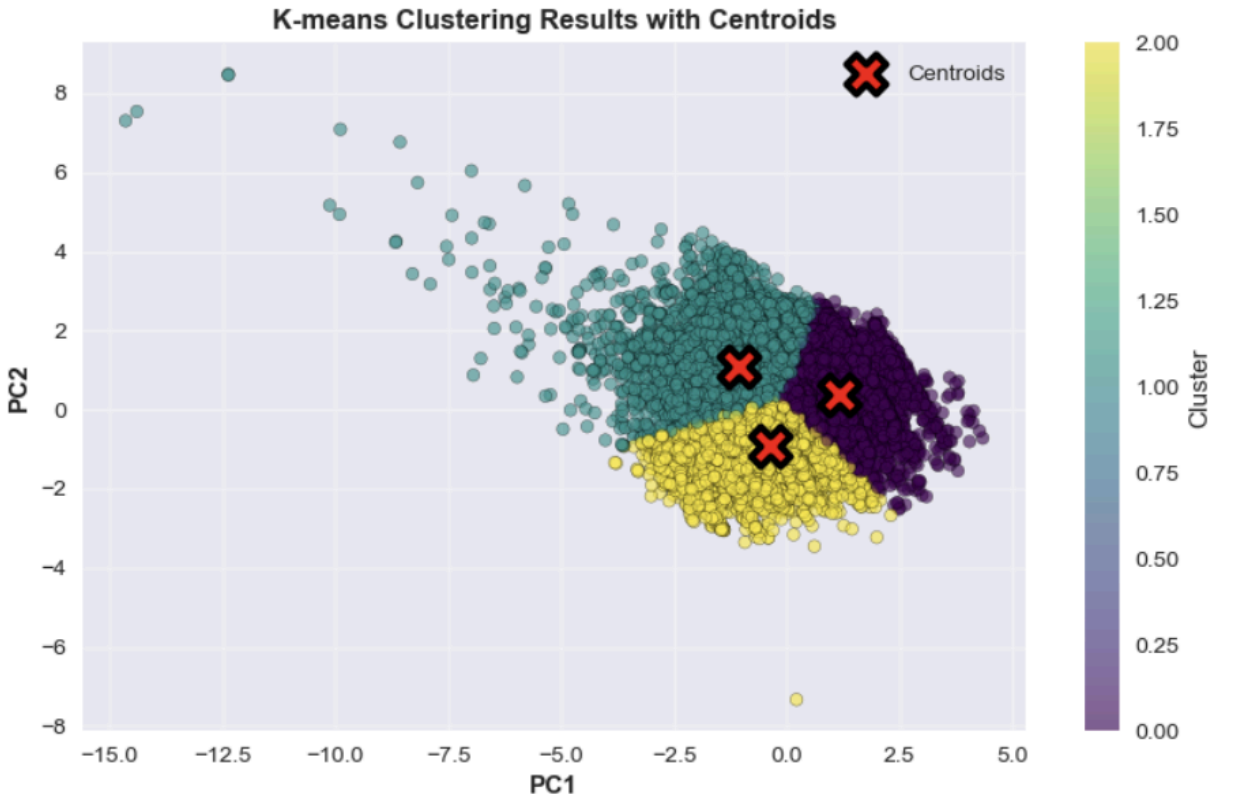


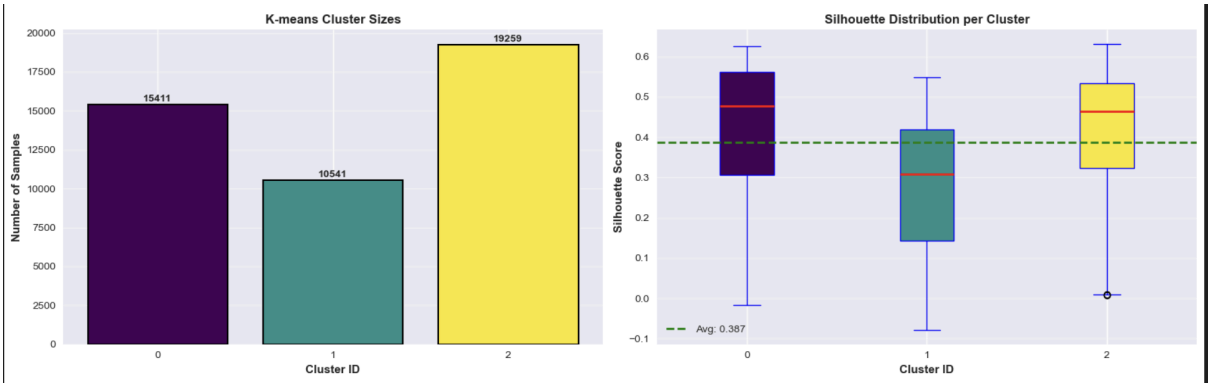
3. ‘Inertia Plot’ and ‘Silhouette Score Plot’ for K-means





4. K-means Clustering Results with Centroids Visible (Scatter Plot) K-means Cluster Sizes (Bar Plot)
Silhouette distribution per cluster for K-means (Box Plot)


K-means Clustering Results with Centroids
Centroids
PC1
PC2
Cluster
8
6
4
2
0
−2
−4
−6
−8
−15.0
−12.5
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00


K-means Cluster Sizes
Number of Samples
Cluster ID
15411
10541
19259
20000
17500
15000
12500
10000
7500
5000
2500
0
0
1
2
Silhouette Distribution per Cluster
Silhouette Score
Avg: 0.387
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
0
1
2
Cluster ID