Project Title :

Lab Week 13 - Clustering

Name:

Nidhi Nitin Nag

SRN:

PES2UG23CS385

Course Name:

ML Lab

Submission Date:

November 11, 2025

Section :

F

1. 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?

Dimensionality reduction was helpful here because a lot of the original features were correlated with each other. When features are strongly related, they don't really add new information and instead make the model more noisy and harder to work with. PCA helps by combining those related features into a smaller set of components that still carry most of the important information.

From the explained variance plot, the first two principal components together capture a little over 28 percent of the total variance. That might not sound very high, but since this dataset has many features that overlap in what they represent, reducing it down to two components still gives a cleaner and more compact view of the main structure in the data.

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.

Based on both the elbow curve and the silhouette score, the best number of clusters for this dataset seems to be **3**.

From the elbow curve, the biggest drops in inertia happen between k=1 to k=3. After k=3, the curve starts to flatten, which usually means that adding more clusters doesn't improve things much anymore. That's the classic "elbow" shape.

The silhouette score for k=3 was around **0.39**, which is not super high but still decent for real-world data. Silhouette scores normally drop if we force too many clusters, so 3 clusters strikes a balance between compactness and separation.

So taking both together, 3 clusters feels like the most reasonable choice here.

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?

When the cluster sizes are compared, it is seen that both methods ended up creating one larger cluster and two smaller ones. This basically tells us that the data itself is unevenly distributed. In other words, there's one big group of customers who share very similar characteristics, and then there are smaller groups that behave differently.

The bigger cluster probably represents the most common customer profile in the dataset. These are the people whose features fall into the typical range, so the algorithm naturally groups a lot of them together. The smaller clusters are the customers who stand out in some way. They might have unusual spending behavior, different financial histories, or other patterns that set them apart.

Customers aren't always split evenly into categories in real life, so the clustering just reflects that natural imbalance.

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?

K-means performed better because it had a higher silhouette score (0.39) compared to bisecting K-means (0.29). This means its clusters were more compact and better separated. Bisecting K-means likely did worse because the data doesn't have a strong hierarchical structure, so the forced splits created less natural 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?

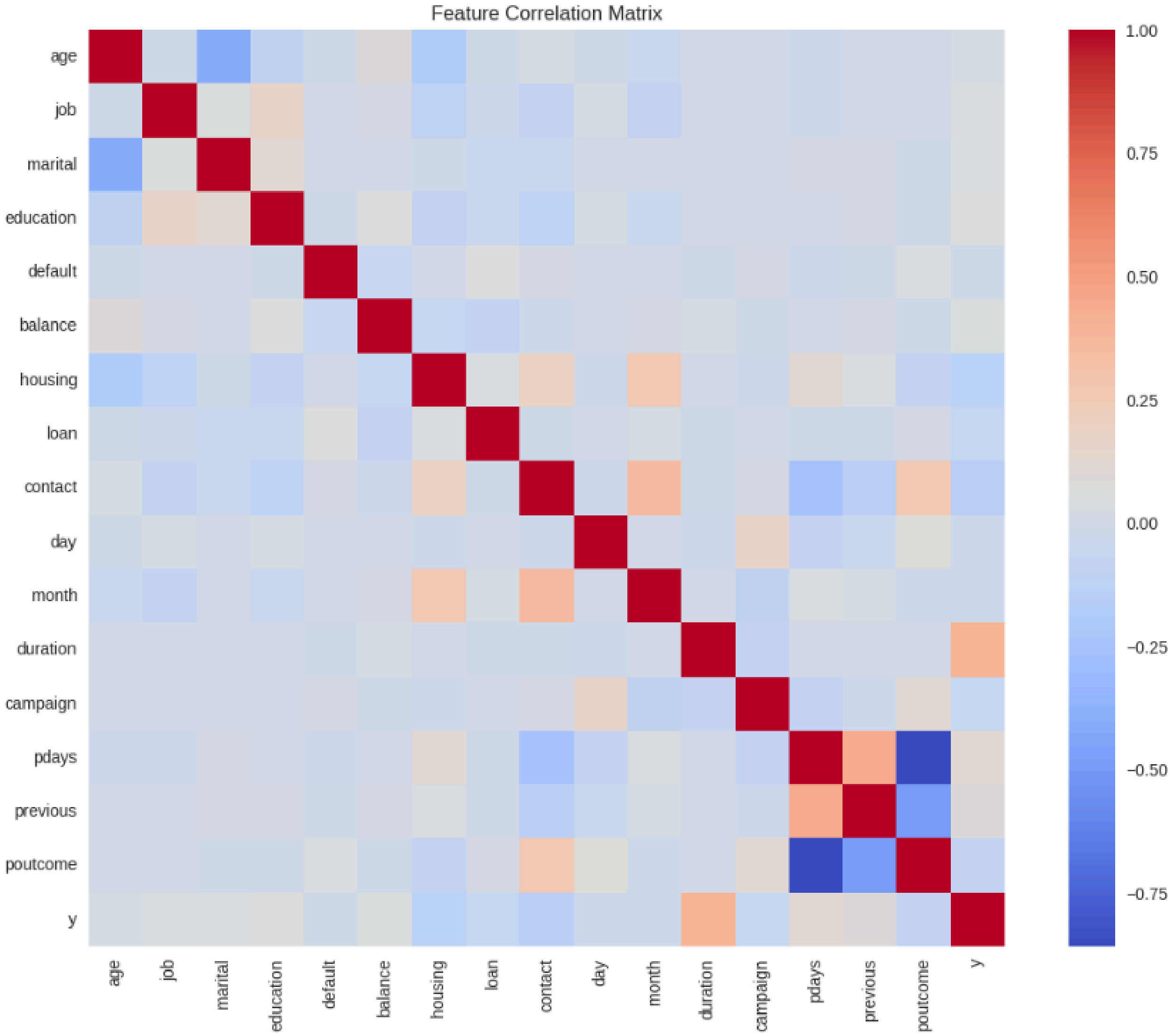
The clusters in the PCA space suggest that the bank's customers naturally fall into a few distinct groups. Each group shows different patterns in behavior and characteristics, meaning they likely respond differently to marketing efforts. This helps the bank target each segment more effectively instead of treating everyone the same. For example, one cluster might represent more stable customers, while another could include people with higher variability in finances or engagement. Understanding these groups allows the bank to tailor campaigns, improve outreach, and focus resources where they will have the most impact.

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?

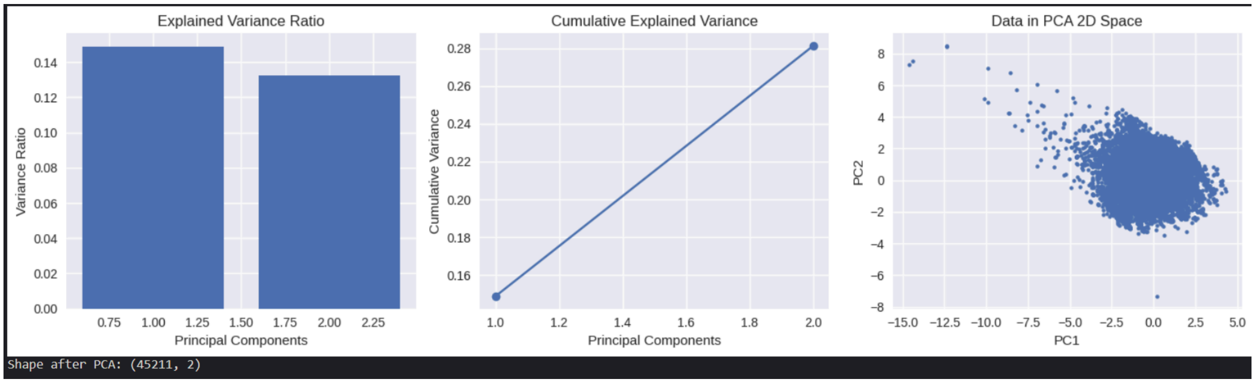
The three colored regions in the PCA plot represent groups of customers who share similar overall patterns in their data. Each region reflects a different combination of financial behavior and demographic traits once everything is compressed into two PCA components. The boundaries look sharp in some areas because certain customers are clearly different from others, making the split more obvious. In other areas the boundaries are more diffuse because these customers share overlapping characteristics, so the separation isn't as clean. Overall, the clusters show that some customer groups are very distinct, while others blend gradually into each other.

2. SCREENSHOTS

1. Feature Correaltion matrix for the dataset

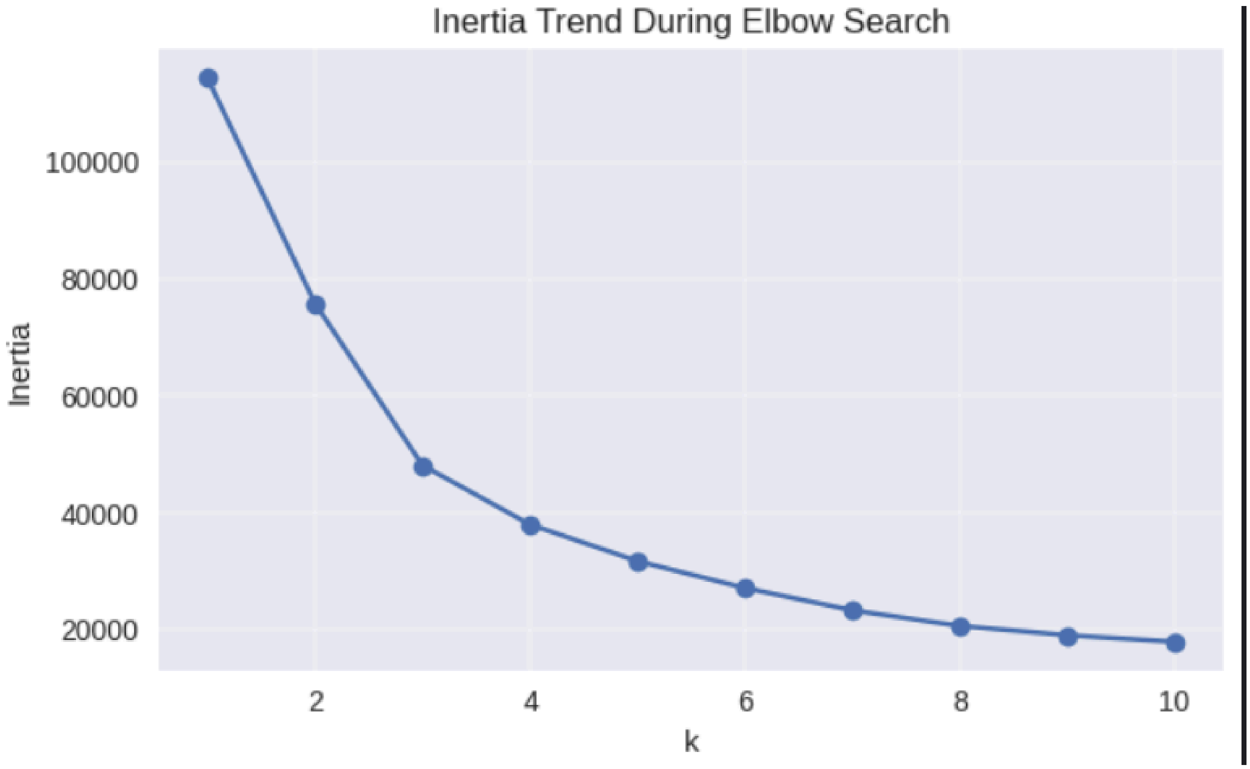


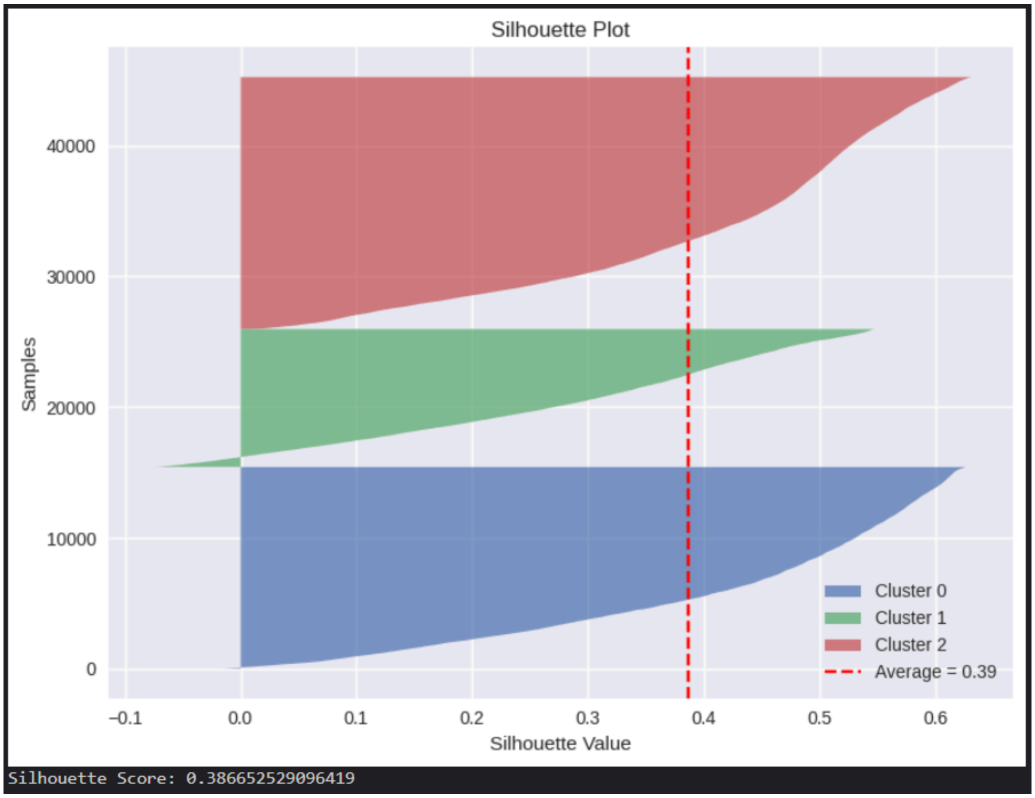
2. ‘Explained variance by Component’ and ‘Data Distribution in PCA Space’ after Dimensionality Reduction with PCA



```
Shape after PCA: (45211, 2)
```

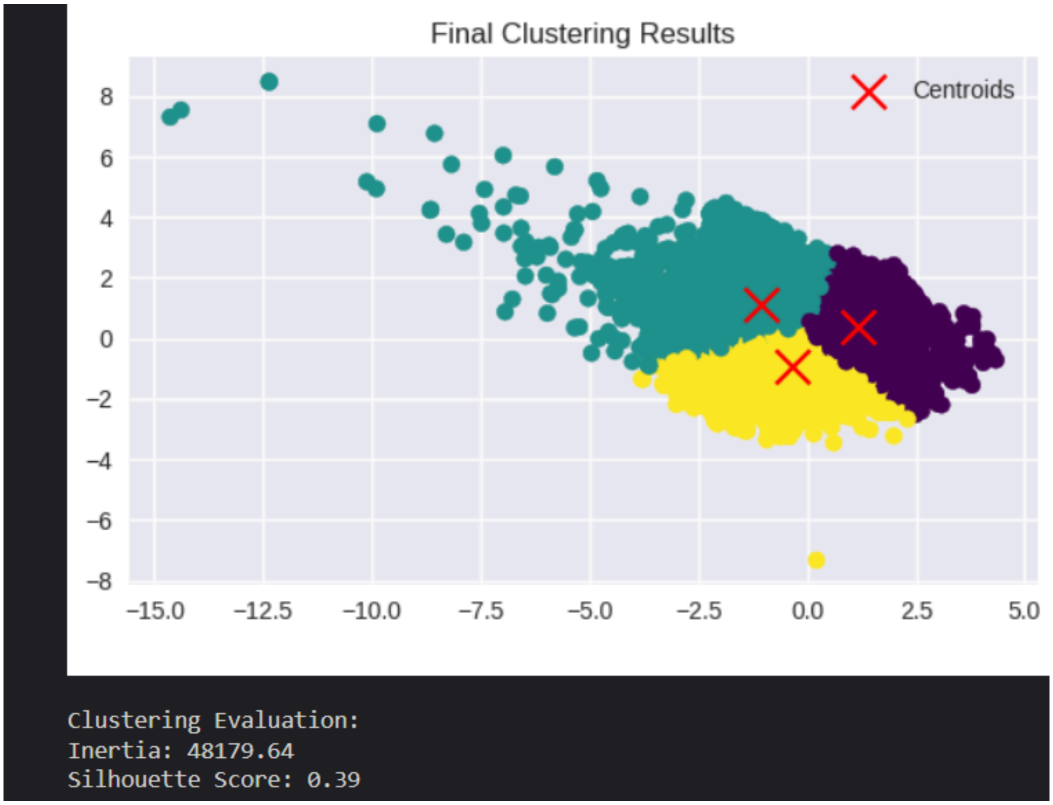
3. ‘Inertia Plot’ and ‘Silhoutte Score Plot’ for K-means





```
Silhouette Score: 0.386652529096419
```

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



```
Clustering Evaluation:
Inertia: 48179.64
Silhouette Score: 0.39
```