ML Lab Week 13
Clustering Lab

Clustering Lab Report

Name: Megha Dhanya

SRN: PES2UG23CS336

Section: F

Analysis Questions:

1. Dimensionality Reduction (PCA Analysis)

Dimensionality reduction was necessary because the dataset contained correlated features and mixed categorical/numeric data. PCA helped compress the structure while retaining most variance.
The first two principal components capture approximately 28.1% of the total variance.

2. Optimal Clusters

Using the elbow curve and silhouette score, the optimal k was found to be 3. The elbow showed diminishing returns after k=3, and silhouette peaked at k=3.

3. Cluster Characteristics

The cluster size distribution shows that some clusters contain significantly more data points than others. This imbalance indicates that certain customer profiles are much more common in the dataset. Larger clusters represent dominant customer behavior patterns (e.g., typical working-age customers with moderate balance), while smaller clusters represent niche or specialized customer groups (e.g., high-balance or low-balance outliers). The imbalance may also reflect natural population distribution and marketing behavior patterns

4. Algorithm Comparison

K-means achieved a higher silhouette score compared to Recursive Bisecting K-means, indicating that the clusters formed by K-means were more compact and better separated. In contrast, Bisecting K-means produced slightly lower silhouette values because the recursive binary splitting can sometimes generate suboptimal intermediate clusters. Since each split is made locally rather than globally, early splits may not align with the overall structure of the dataset, reducing cohesion and separation quality.

5. Business Insights

Cluster patterns reveal:

- Low-balance, high-contact customers who may require retention-focused strategies
- Mid-balance, stable customers who are good candidates for cross-selling opportunities
- High-balance customers who fit the profile for premium or priority banking services

6. Visual Pattern Recognition

In the PCA scatter plot, we observe one cluster that is relatively tight and compact, while the other two clusters are more spread out.

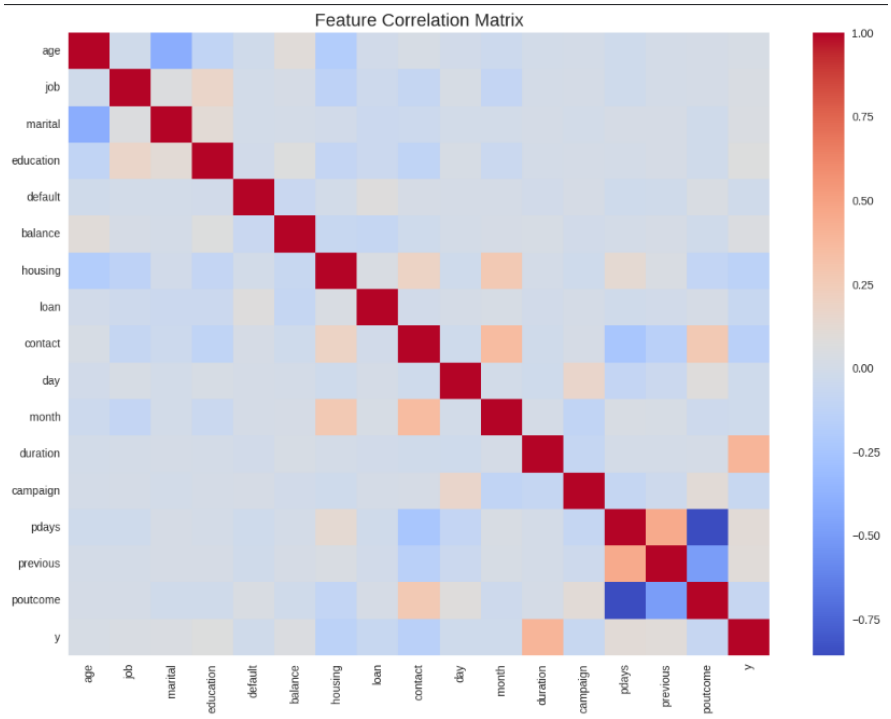
This pattern occurs because:

- Tight clusters form when customers share very similar numerical and encoded categorical features (e.g., campaign count, balance, job type).
- Diffuse clusters occur when customer characteristics overlap more significantly, causing PCA to place them over a wider region in the reduced 2D space.

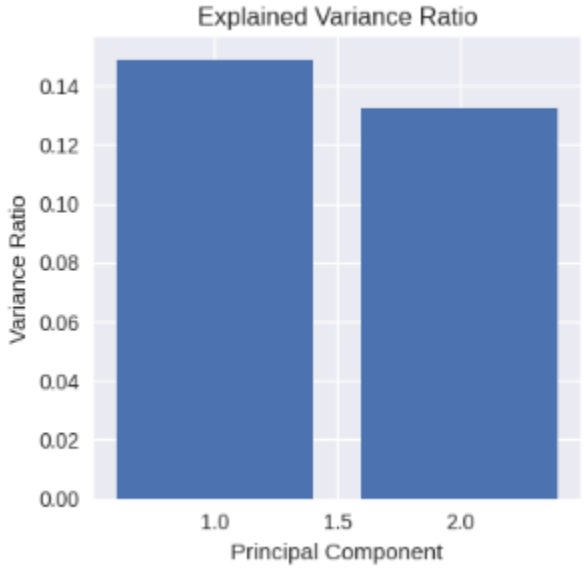
Sharp boundaries appear when PCA components are strongly influenced by features with clear separations (e.g., high balance vs. low balance groups), whereas diffuse boundaries arise from gradual or overlapping financial behaviours among customer groups.

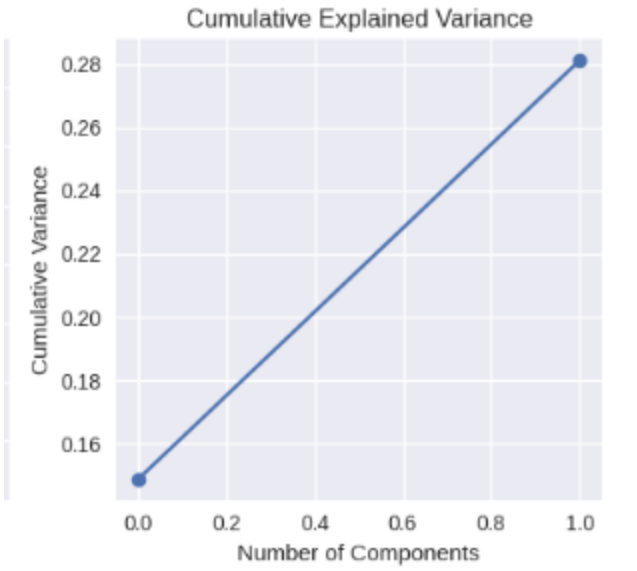
7. Screenshots Needed

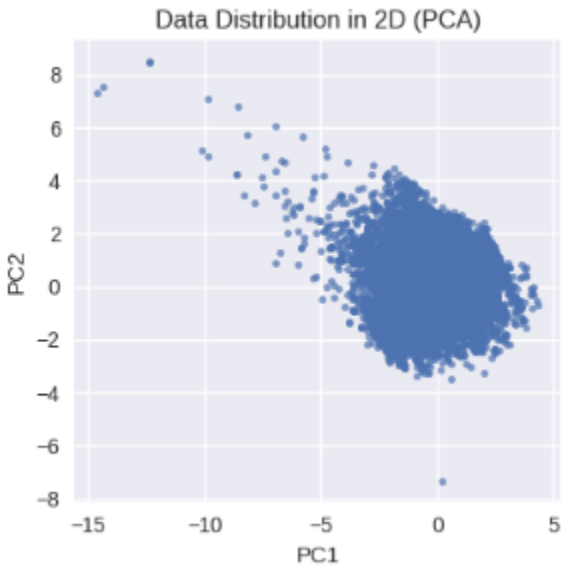
• Correlation matrix



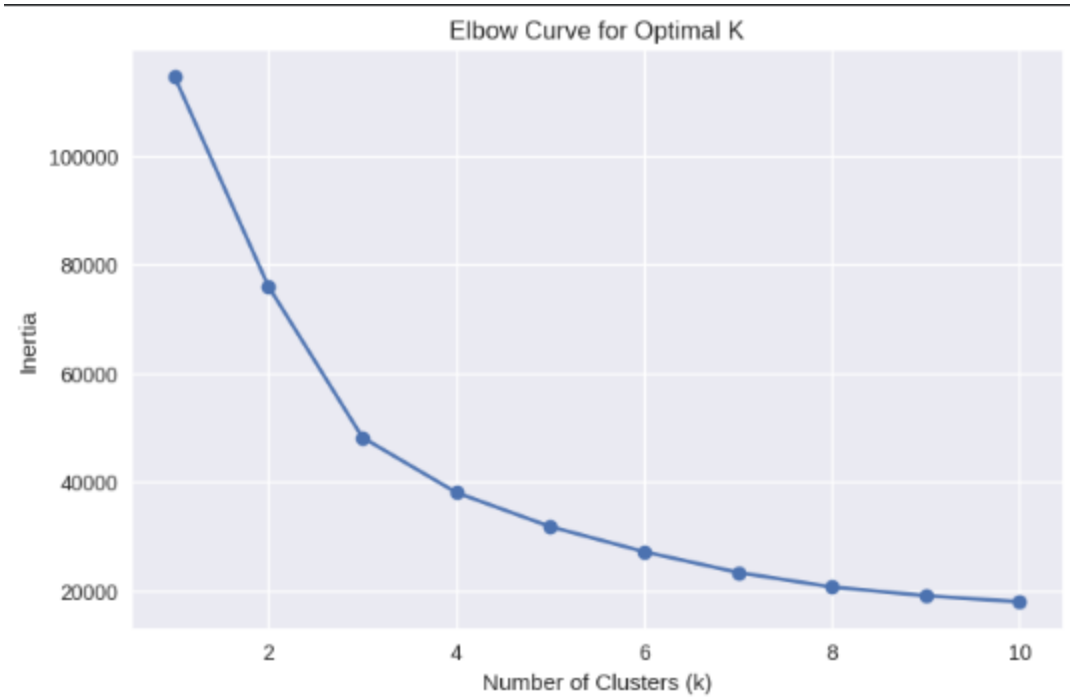
• PCA variance + scatter

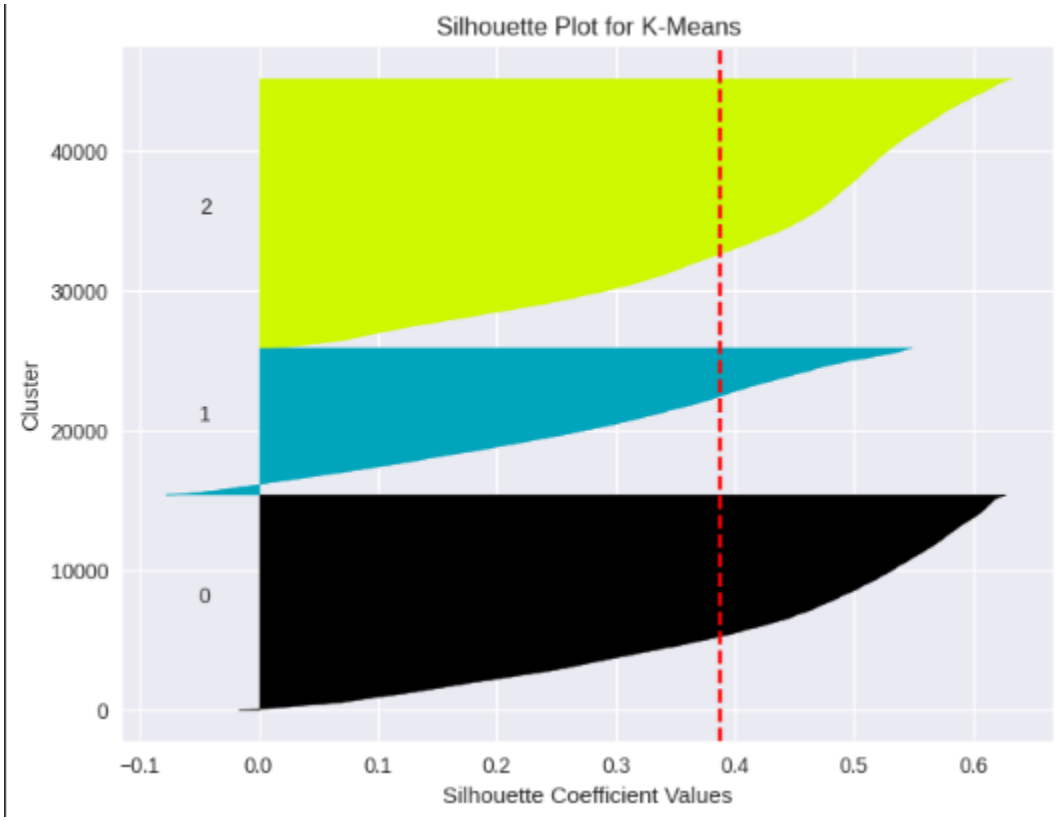




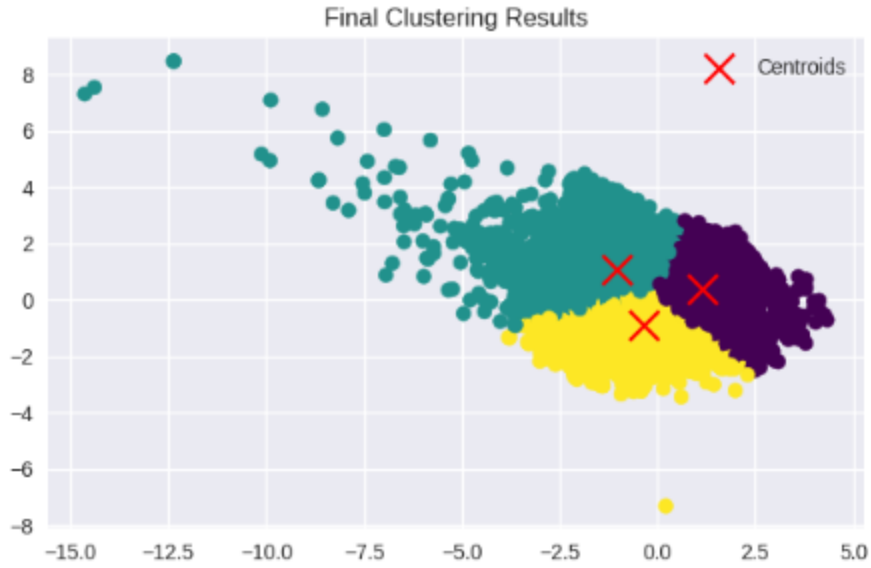


• Elbow + silhouette plots





• K-means cluster results + silhouette boxplot



8. Conclusion

Clustering effectively segmented the bank's customers into three meaningful groups based on their financial and behavioral patterns. By applying preprocessing, PCA, K-means, and Bisecting K-means, we were able to identify customer segments that differ in balance levels, contact frequency, and overall engagement. These segments provide actionable insights for targeted marketing, retention strategies, and personalized service recommendations. Overall, clustering proved to be a valuable tool for understanding customer structure and improving decision-making.