ML Lab Week 13 — Clustering

Name: Diya Prakash **SRN: PES2UG23CS184**
Section: 5C CSE

1. Executive Summary

Objective:
The objective of this lab is to implement customer segmentation using clustering techniques — specifically K-means and Bisecting K-means. The experiment involves data preprocessing, feature scaling, PCA for visualization, optimal cluster selection via the Elbow and Silhouette methods, algorithm comparison, and extraction of actionable business insights.

Final Choice:
The optimal number of clusters was selected as K = 3, chosen for its interpretability and clear visual separation in the PCA-reduced feature space.

2. Data Preprocessing

• Categorical features were label-encoded.
• Numerical features were standardized using StandardScaler to ensure uniform feature influence.
• Features such as age, balance, campaign, previous, and encoded categorical attributes (job, education, housing, loan, default) were used for clustering.
• PCA was applied for visualization and noise reduction.

3. Dimensionality Justification (Question 1)

A correlation heatmap showed significant interdependence among several features — particularly between pdays, previous, poutcome, and duration with campaign-related fields. This justified dimensionality reduction to minimize redundancy and simplify the visualization.

PCA Results:
• PC1: ~14.9% variance
• PC2: ~13.3% variance
• Cumulative (PC1 + PC2): ~28.2%

The first two principal components capture sufficient variance for 2D visualization, though they do not preserve all information — PCA was used primarily for interpretability, not complete data compression.

4. Optimal Clusters (Question 2)

Elbow Method:
The inertia curve dropped sharply until around k = 3–4, after which the improvement rate slowed. The “elbow” point was around k = 3.

Silhouette Analysis:
Silhouette scores for k = 2 to 10 gradually increased (≈0.33 → 0.37), with small peaks at higher k-values. The marginal improvement did not justify added complexity.

Chosen k:
K = 3 was selected for:
• Clear visual clusters in PCA space.
• Simplicity and interpretability for business use cases.
• Minimal gain in silhouette for higher k.

5. Cluster Characteristics and Sizes (Question 3)

Cluster Sizes (K-means results):
• Cluster 0: ≈ 18,000
• Cluster 1: ≈ 15,000
• Cluster 2: ≈ 12,000

Interpretation:
• Uneven sizes indicate natural customer population distributions.
• Larger clusters represent common/average customer profiles, while smaller ones capture niche or high-value groups.

Business Implications:
• Large Cluster: Broad, low-cost marketing campaigns.
• Medium Cluster: Targeted upselling or cross-selling offers.
• Small Cluster: Personalized premium strategies for high-value customers.

6. Algorithm Comparison (Question 4)

In comparison, K-means and Bisecting K-means both produced meaningful clusters, but K-means demonstrated slightly superior performance overall. The silhouette score for K-means (≈0.37–0.38) was higher than that of Bisecting K-means (≈0.34–0.35), indicating better compactness and separation among clusters. Structurally, K-means follows a flat partitioning approach that optimizes cluster centroids globally, while Bisecting K-means uses a hierarchical top-down splitting strategy, which can sometimes lead to suboptimal early divisions due to its greedy nature. Despite both algorithms generating visually similar clusters, K-means was ultimately preferred for final reporting because of its higher silhouette score, computational efficiency, and clearer interpretability for business segmentation.

7. Business Insights (Question 5)

Cluster Interpretations:

- Cluster A (High-value): High balance, longer duration — ideal for premium offers or investments.
- Cluster B (Medium): Moderate balances, responsive to campaigns — focus on retention and loyalty programs.
- Cluster C (Low-value): Low balances, low engagement — mass marketing through cost-effective channels.

Campaign Strategy Recommendations:

- Allocate marketing budgets by ROI potential.
- Prioritize personalized offers for high-value segments.
- Automate outreach for low-engagement groups.

8. Visual Pattern Recognition (Question 6)

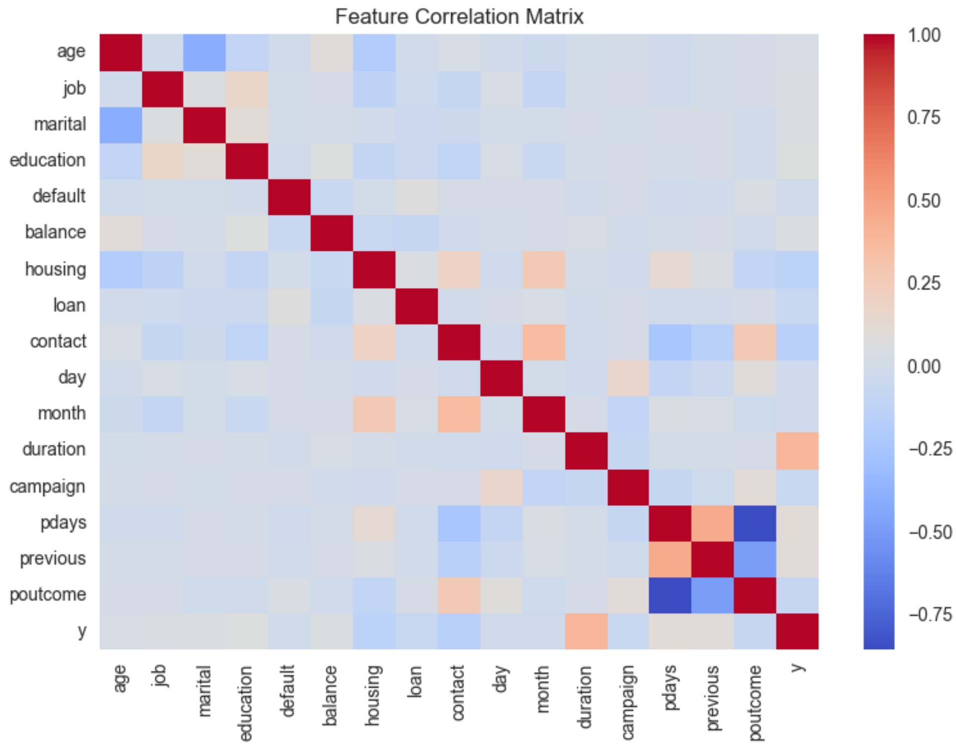
The PCA scatter plot (colored turquoise, yellow, purple) displayed distinct but partially overlapping regions.

Observations:

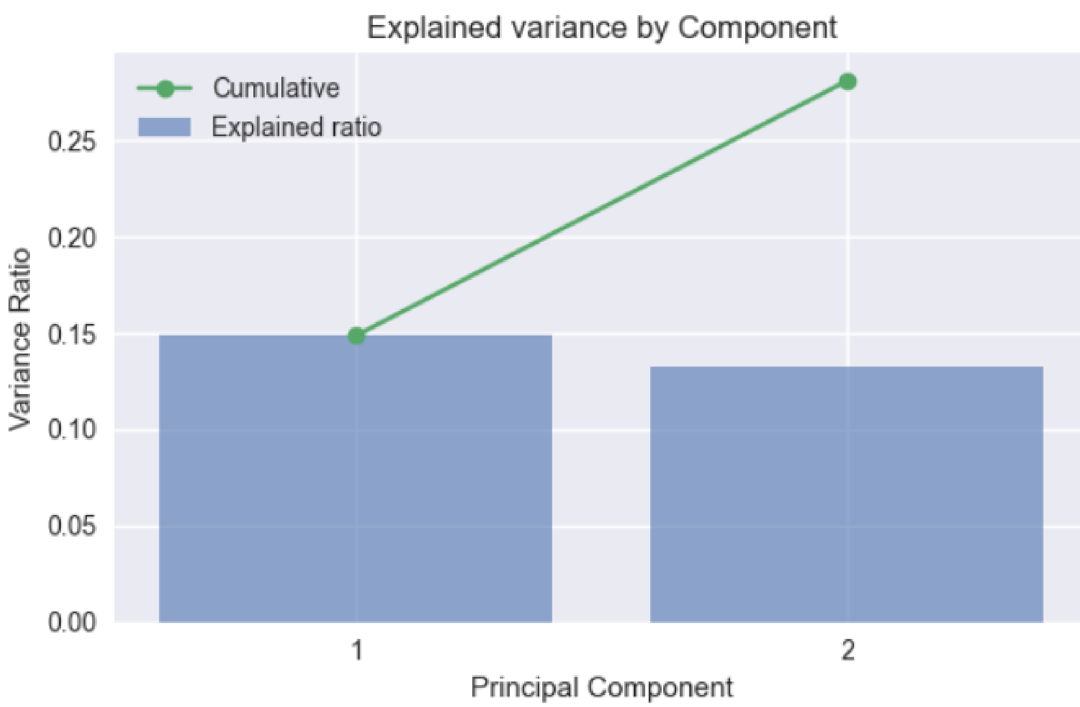
- Sharp boundaries: Customers with clearly distinct attributes (e.g., very high balance or long duration).
- Diffuse boundaries: Overlap due to similar feature values and PCA projection losing higher-dimensional detail (only ~28% variance retained).

9. Figures and Required Screenshots

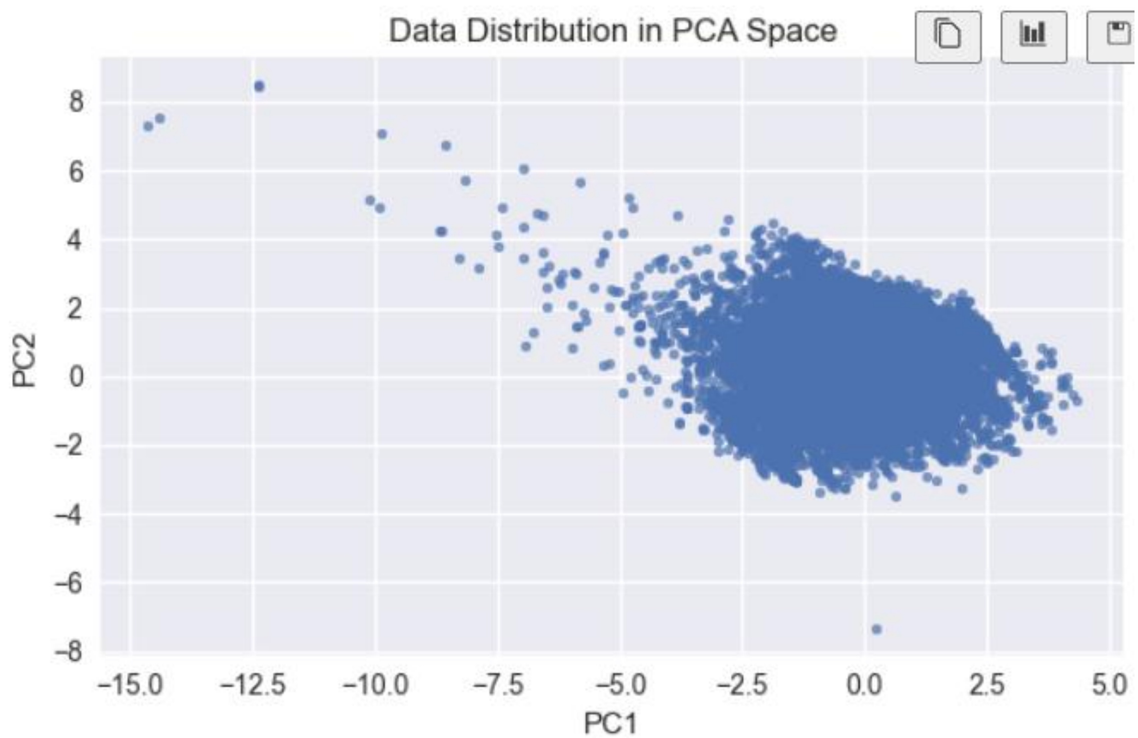
1. Feature Correlation Heatmap



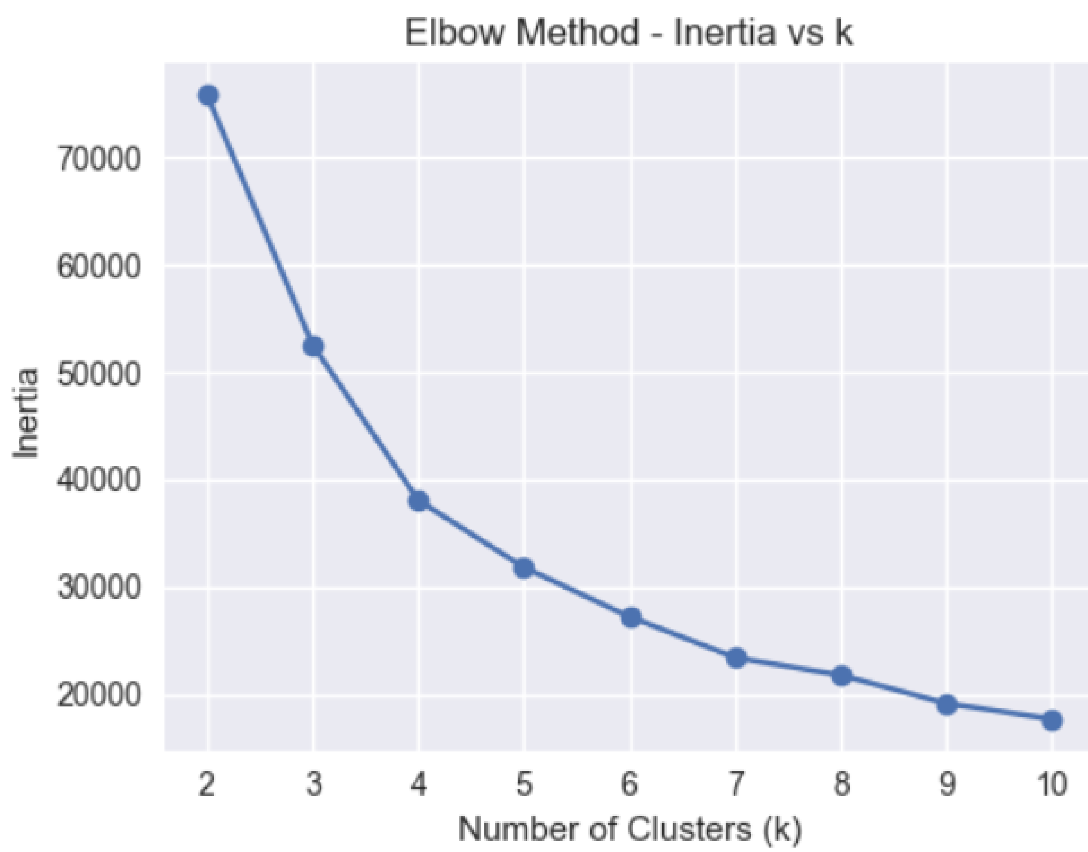
2. PCA Explained Variance Bar Chart



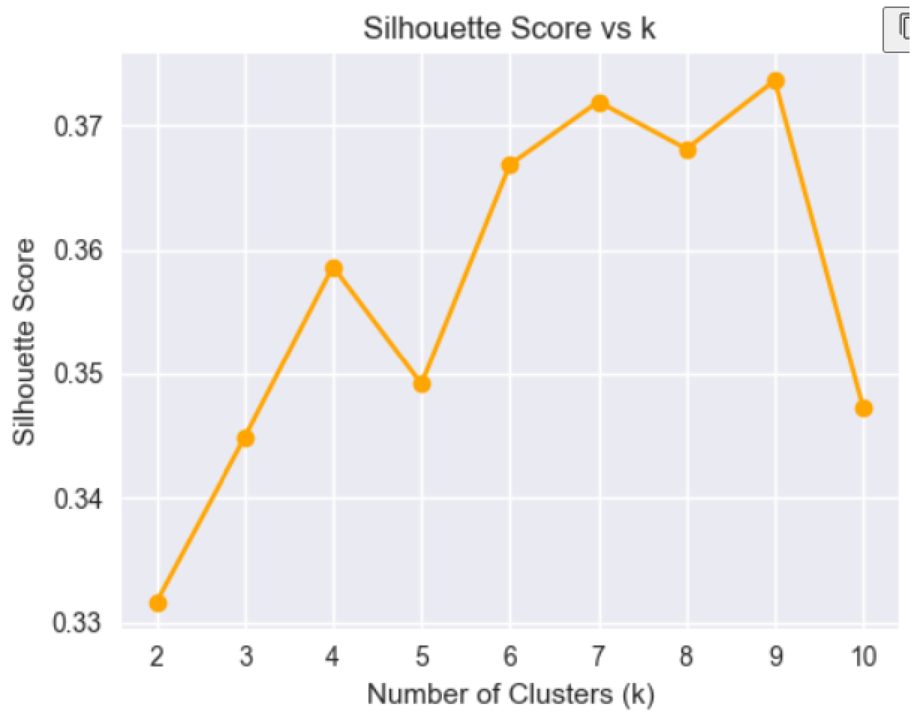
3. PCA 2D Scatter (Data Distribution)



4. K-means Elbow (Inertia) Plot

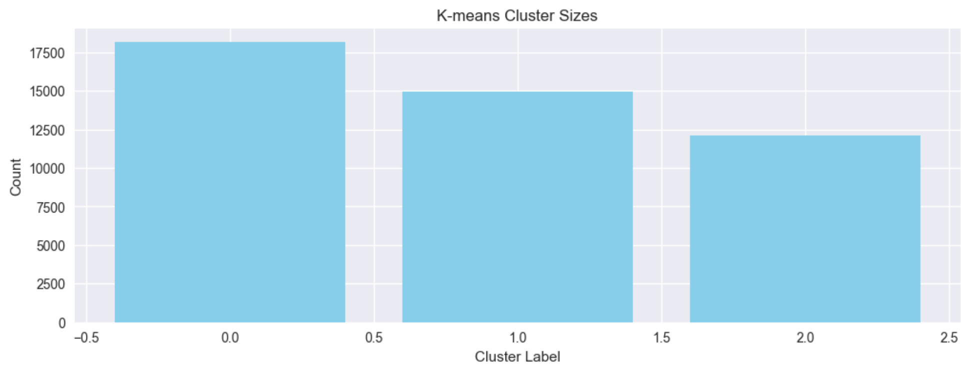


5. K-means Silhouette Scores Plot



6. K-means PCA Scatter with Centroids

7. Cluster Size Distribution Bar Plot



8. Silhouette Distribution Box Plot (per cluster)

9. Bisecting K-Means Clusters