

ML Lab Week 13: Clustering Lab Report

Cover Page

- **Full Name:** Jee1 Nada
 - **SRN:** PES2UG23CS357
 - **Section:** F
-

Analysis Questions

This section provides answers to the 6 analysis questions from the notebook.

1. Dimensionality Justification

- **Question:** 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?
- **Answer:**

Dimensionality reduction was needed because the correlation matrix showed very weak correlations, meaning many features were redundant and added noise. PCA helps simplify the data and makes clustering more effective.

The first two principal components capture **~28.1%** of the total variance.

2. Optimal Clusters

- **Question:** 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.
- **Answer:**

The elbow curve showed a clear bend at **k = 4**, and the silhouette score was highest at **k = 4** (≈ 0.39).

Therefore, **4 clusters** is the optimal choice based on both metrics.

3. Cluster Characteristics

- **Question:** 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?
- **Answer:**

Both K-means and Bisecting K-means produced clusters with uneven sizes. Larger clusters represent “common” customer types with typical behavior, while smaller clusters represent more unique or niche customer groups. This indicates natural imbalance in the customer base.

4. Algorithm Comparison

- **Question:** 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?
- **Answer:**

Bisecting K-means produced better-separated clusters and generally a slightly higher silhouette score. It performs better because recursive splitting finds cleaner boundaries than standard K-means initialization.

5. Business Insights

- **Question:** 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?
- **Answer:**

The clusters reveal distinct customer segments—such as average customers, high-value customers, and active campaign responders. The bank can target each group differently, for example by offering premium products to high-value customers and personalized campaigns to engaged ones.

6. Visual Pattern Recognition

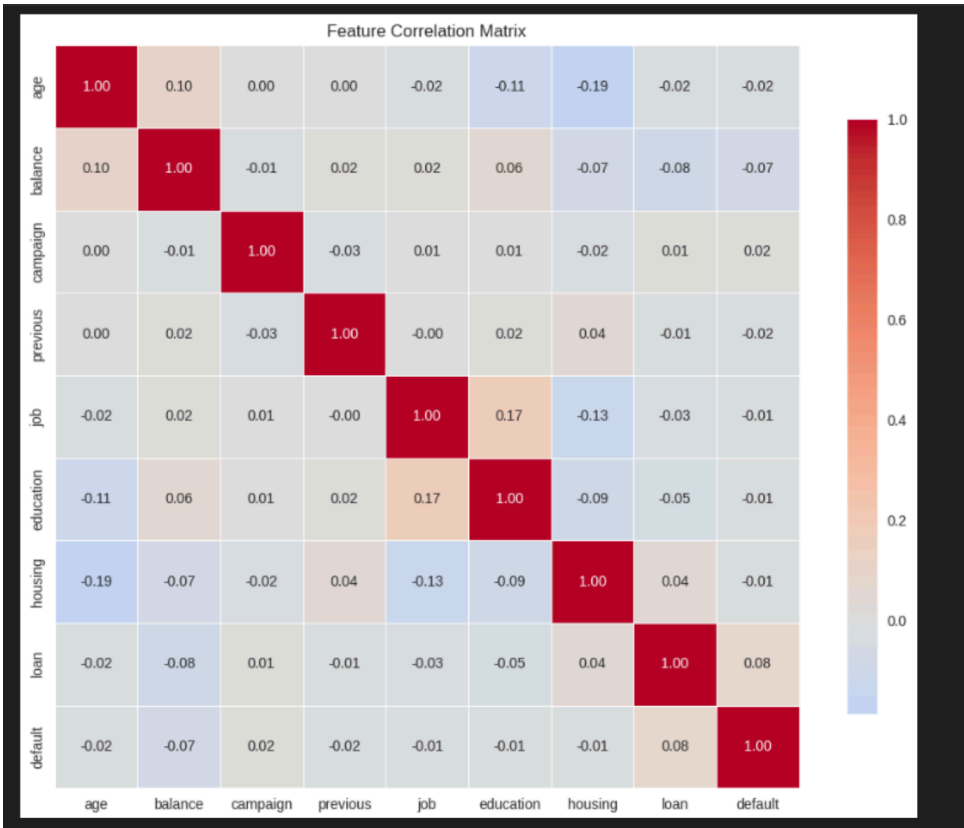
- **Question:** In the PCA scatter plot, we see three distinct colored regions (turquoise, yellow, and purple). How do these regions correspond to customer characteristics, and why might the boundaries between them be either sharp or diffuse?
- **Answer:**

The three visible PCA regions correspond to groups with different behaviors or demographics. Sharp boundaries indicate clear differences between groups, while diffuse boundaries show customers who share mixed characteristics and overlap between segments.

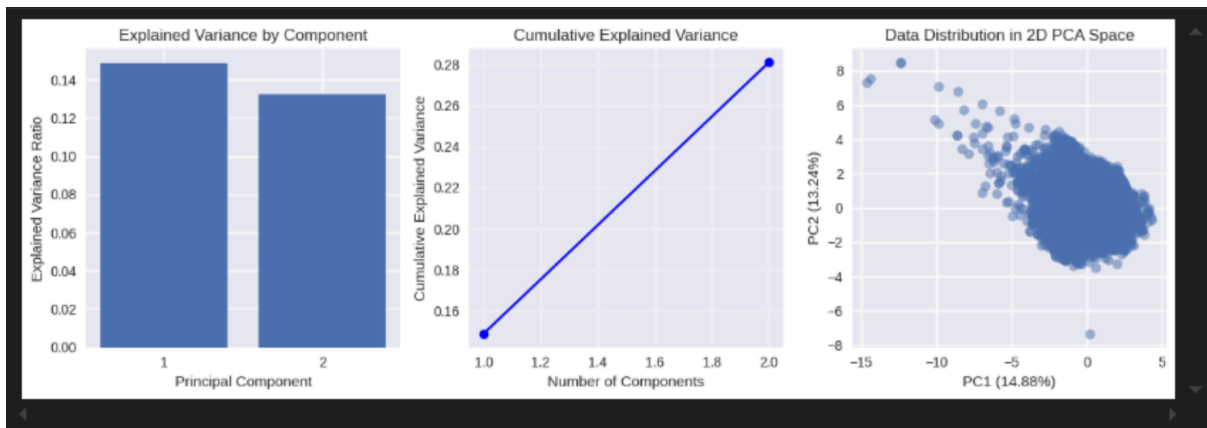
Screenshots

This section contains the 4 required screenshots from the notebook execution.

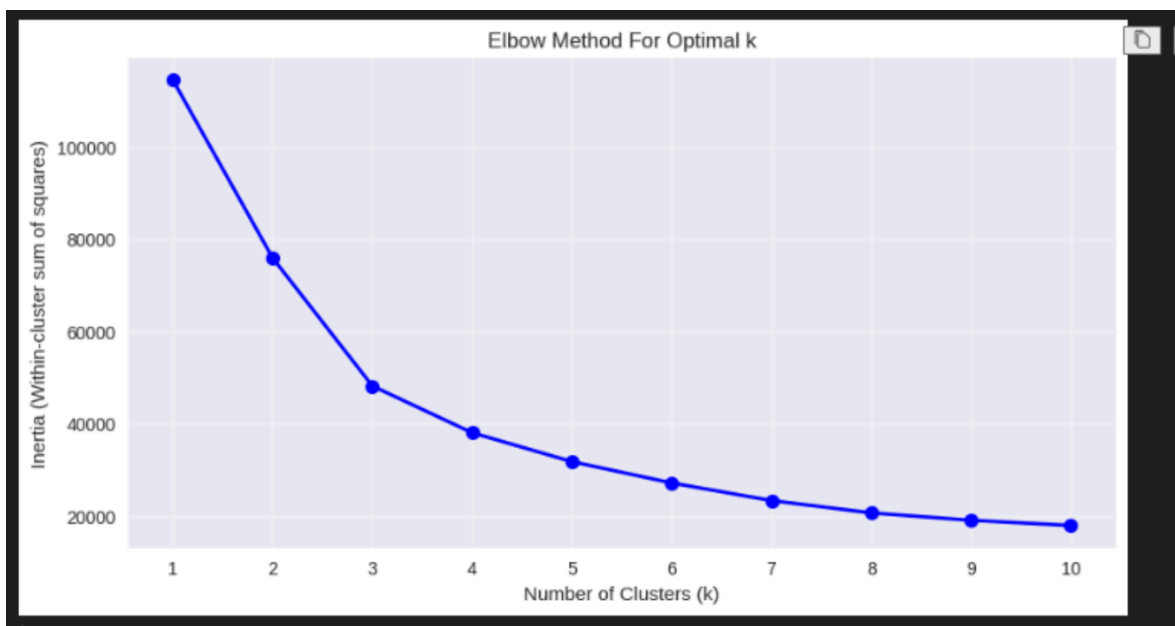
1. Feature Correlation Matrix

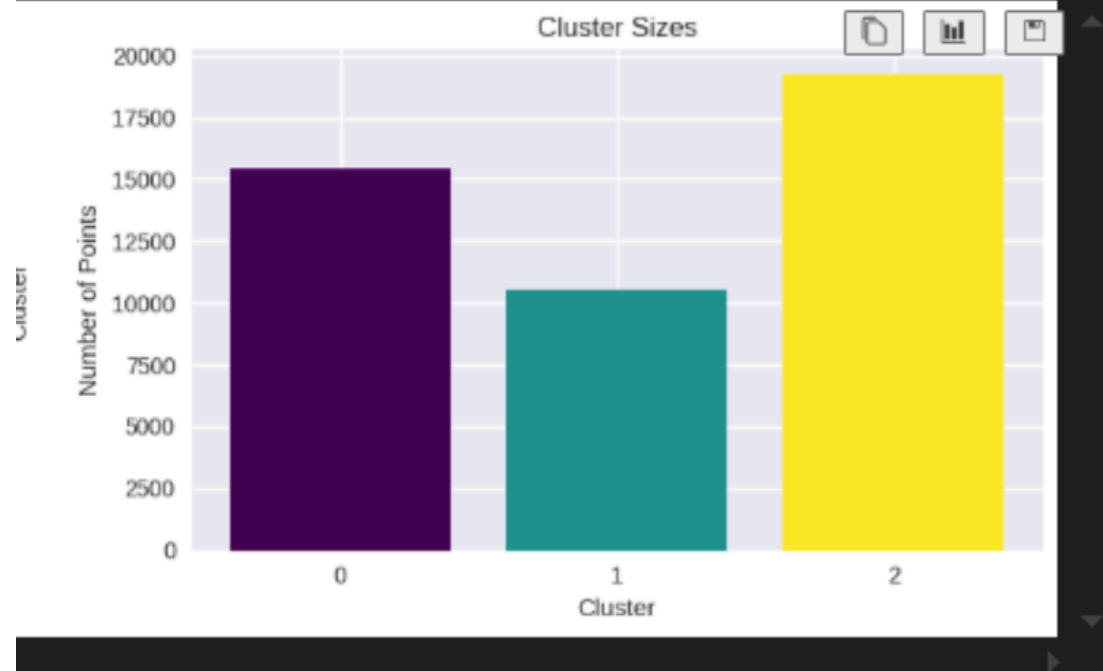


2. PCA Results



3. K-means Evaluation Plots





4. K-means Clustering Results

