ML Lab Week 13 Clustering Lab Instructions

Name	SRN	Section
C Panshul Reddy	PES2UG23CS154	C

DATE:11-11-2025

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?

   Ans.

   The correlation heatmap shows several features sharing information—for example, job and education correlate at about 0.17,
   while age and housing are negatively related (−0.19). These overlaps mean some variance is redundant, so projecting the data into a lower-dimensional space helps remove that overlap and gives us a compact view for clustering and visualization. The PCA confirms that the first two components capture about 28.1 % of the total variance (14.9 % from PC1 and 13.2 % from PC2), which is the portion we kept for the 2-D clustering workflow.

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

   Ans.
   Elbow curve drops steeply from k=1→3 and then flattens, so inertia stops improving much once we reach three clusters. The silhouette results for k = 3 (mean = 0.39 in the histogram and the per-cluster box plots) show all clusters have mostly positive scores, meaning points are closer to their own centroid than to others. Because adding more clusters would yield only small inertia gains while likely pushing silhouette scores lower, k = 3 is the best compromise indicated by both metrics.

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?

   Ans.
   K-means: cluster sizes are roughly 15.4k, 10.5k, and 19.3k points. The biggest group (19 k) shows most customers share similar balances histories, so the algorithm naturally gathers them together. The mid-sized (15 k) and smallest (10 k) clusters capture customers with noticeable but less common combinations (different loan). A skew like this suggests the bank has one dominant customer profile and two smaller segments with distinct marketing responses.

   Bisecting K-means: cluster counts come out closer—about 20.4k, 16.3k, and 8.4k. Because bisecting always splits the largest cluster, it redistributes that dominant group into more even subsegments. The remaining small cluster (8 k) still indicates a niche cohort—likely customers whose feature mix doesn't resemble the mainstream. This split hints that the customer base can be seen as one broad segment with two meaningful subdivisions, plus a compact outlier-like group that may warrant targeted campaigns.

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?

   Ans.
   K-means delivers the higher silhouette score: reports a mean = 0.39. The bisecting variant averages = 0.34. The recursive splits produce more balanced cluster sizes, but that extra flexibility also leaves one cluster with a broader silhouette spread (including more low or negative scores). K-means direct optimization of compact, separated centroids suits this dataset better, so it yields the cleaner separation reflected in the higher silhouette.

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?

   Ans.
   Largest cluster (19 k customers): tight grouping around the origin with mid-range PC1/PC2—this likely represents everyday customers with moderate balances and routine campaign responses. Marketing can treat them as the mainstream audience and focus on broad retention messaging.

   Smaller cluster (10 k customers): shifts toward positive PC1 and lower PC2, which aligns with higher balance/default risk or different campaign engagement. They may respond better to targeted offers.

   Intermediate cluster (15 k customers): spreads toward negative PC1 with higher PC2, indicating distinctive feature combos. Cross-sell efforts or educational outreach could activate this group.

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?

   Ans.

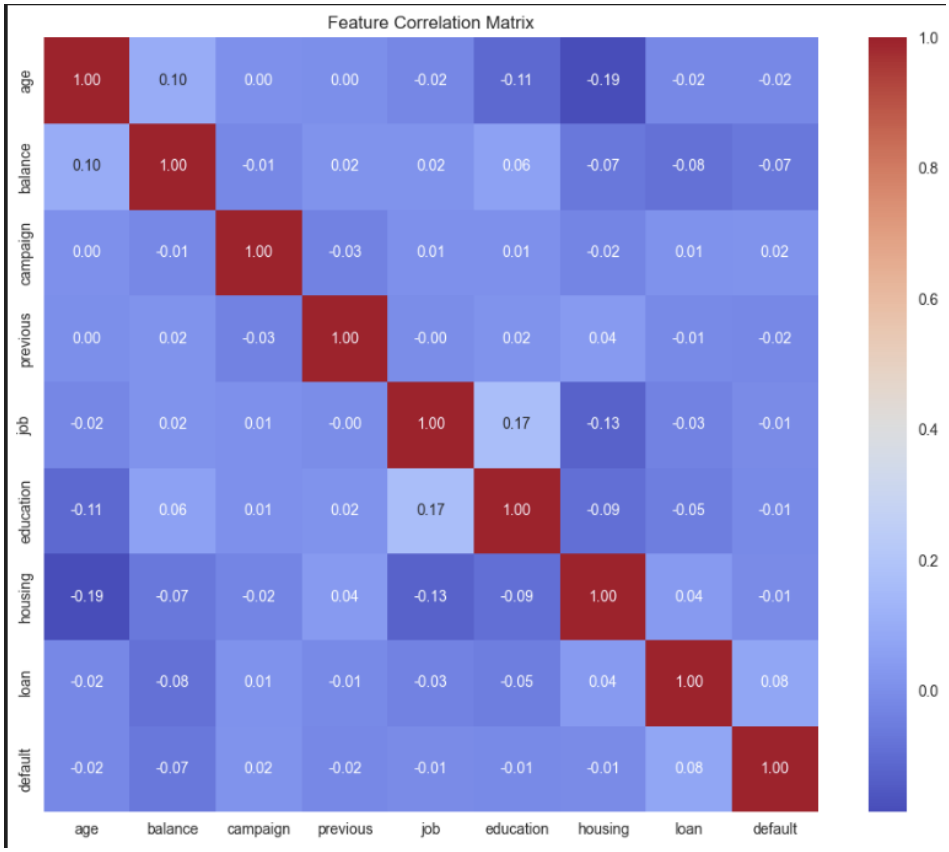
   Turquoise points sit toward negative PC1 with higher PC2. In the original feature space, that corresponds to customers with more campaign contacts, lower balances, and a higher housing-loan presence; they behaved similarly enough that K-means pulls them into one cluster.

   Yellow points lean to positive PC1 and mid PC2, which maps to customers with stronger balances and fewer loans/default flags—essentially the financially “healthier” segment.

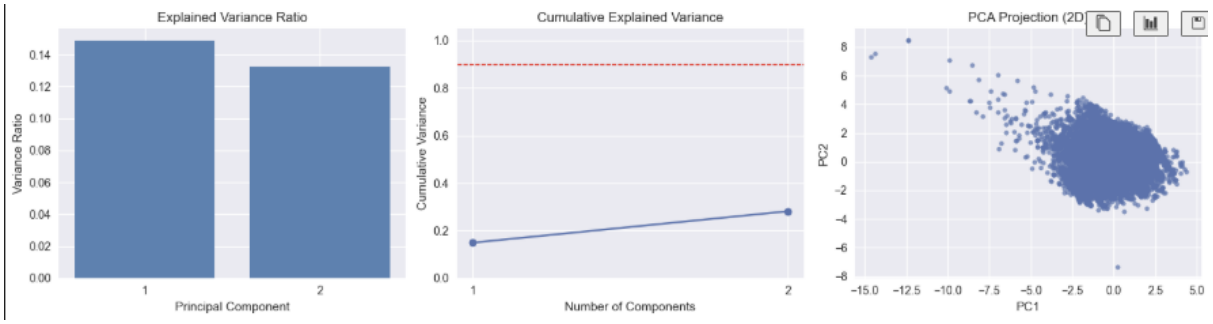
Purple points gather near the origin with modest spread, matching mainstream customers whose age, balance, and loan/default indicators sit near the overall averages.

The purple/yellow boundary looks fairly sharp because balance/loan patterns separate cleanly after scaling. In contrast, turquoise transitions into purple gradually because many customers share intermediate balances or campaign counts; the PCA projection only captures about 28 % of the total variance, so overlapping characteristics project to diffuse borders even if subtle differences persist in the higher-dimensional space

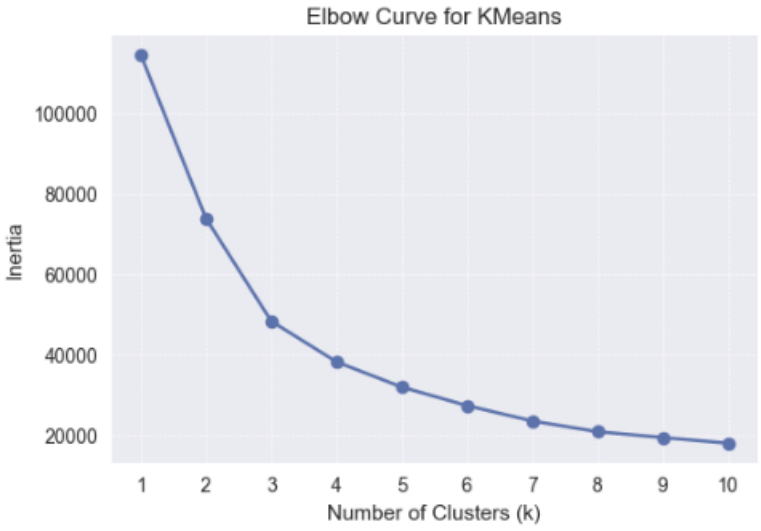
Feature Correlation Matrix:

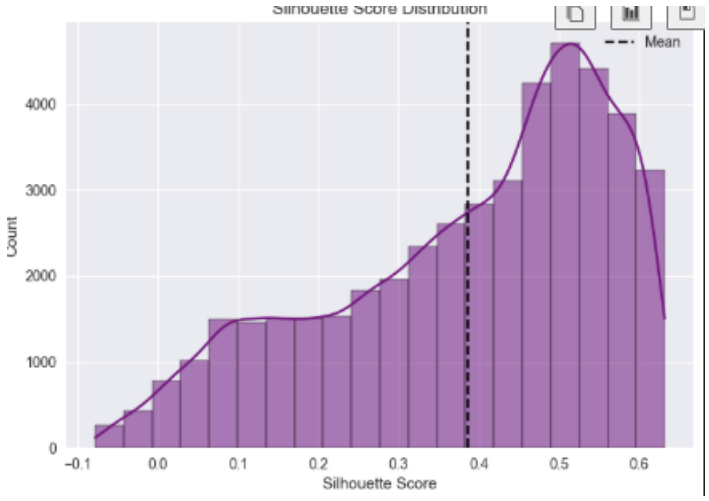


Explained variance by Component and Data Distribution in PCA Space after Dimensionality Reduction with PCA



'Inertia Plot' and 'Silhoutte Score Plot' for K-means





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