ML Lab Week 13 Clustering Lab Instructions

NAME:C YOGESH REDDY
SRN: PES2UG23CS159
SEC: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:

Dimensionality reduction was essential based on the correlation heatmap showing redundant information across features. For instance, job and education exhibit a correlation of 0.17, while age demonstrates a negative relationship with housing (-0.19). These correlations indicate overlapping variance that can be consolidated through PCA.

Variance Captured: The first two principal components retain approximately ****28.1% of total variance**** (PC1: 14.9%, PC2: 13.2%). This compression eliminates redundant correlations while preserving the primary patterns needed for effective 2D clustering and visualization. Although 28% may appear modest, it successfully distills the most significant customer behavior variations from the original 9-dimensional feature space into an interpretable projection for segmentation analysis.

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:

The optimal number of clusters is **k=3**, supported by convergent evidence from both evaluation metrics.

Elbow Curve Analysis: The inertia plot exhibits a sharp decline from k=1 to k=3 (~115,000 → ~38,000), followed by a pronounced flattening beyond k=3. This "elbow point" at k=3 indicates diminishing returns—additional clusters provide minimal inertia reduction (only ~11,000 from k=3 to k=7).

Silhouette Score Validation: At k=3, the mean silhouette score is **0.39**, demonstrating well-separated clusters. The distribution histogram shows most samples scoring between 0.3-0.6 (positive values indicate correct assignments), and per-cluster boxplots confirm all three clusters maintain predominantly positive silhouette coefficients, signaling strong internal cohesion with clear inter-cluster separation.

Both metrics agree that k=3 offers the optimal balance between cluster compactness and separation without overfitting.

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 Distribution: Cluster sizes are 19,259 (42.6%), 10,541 (23.3%), and 15,411 (34.1%) customers. The largest cluster (19.3k) captures the dominant customer profile—individuals with typical balance histories, campaign responses, and loan patterns representing the bank's core demographic. The medium cluster (15.4k) reflects customers with moderately distinct financial behaviors, while the smallest (10.5k) identifies a niche segment with unique feature combinations, possibly differing in loan activity or campaign engagement.

Bisecting K-means Distribution: Produces more balanced sizes—20,443 (45.2%), 13,415 (29.7%), and 11,353

(25.1%). The recursive splitting strategy continuously divides the largest cluster, redistributing the dominant group into more uniform subsegments. However, the persistent small cluster (~11k) indicates a compact, outlier-like cohort whose feature profile remains distinctly separated from the mainstream.

Customer Insights: The size imbalance reveals the bank serves one predominant customer type with two meaningful market segments. The largest cluster represents standard banking customers requiring broad retention strategies, the medium segment shows growth potential for targeted cross-selling, and the smallest group may warrant specialized campaigns or risk management attention.

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 demonstrates superior performance with a silhouette score of **0.39** versus Bisecting K-means **0.29**, alongside lower inertia (48,179.64 vs. 58,308.18).

The standard K-means algorithm directly optimizes for compact, well-separated centroids through simultaneous cluster assignment, which aligns well with this dataset's natural structure. In contrast, Bisecting K-means employs sequential binary splits that prioritize balanced cluster sizes over separation quality. This recursive approach produces one cluster with substantially broader silhouette distribution, including numerous low or negative scores indicating misassigned points.

The silhouette histograms reveal K-means achieves tighter concentration around 0.4-0.5, while Bisecting K-means exhibits wider spread with more questionable assignments. Per-cluster boxplots confirm K-means maintains consistently positive silhouettes across all three segments, whereas Bisecting K-means shows Cluster 1 contains many near-zero values—evidence of forced splits that don't reflect the data's inherent grouping structure. For this customer dataset, K-means' global optimization yields cleaner, more interpretable segments suitable for actionable business insights.

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:

The PCA clustering reveals three actionable customer segments for strategic marketing:

Largest Cluster (19.3k customers): Concentrated near the origin with moderate PC1/PC2 values, representing mainstream customers with average balances, routine loan activity, and typical campaign responses. This core demographic requires broad retention messaging and standardized service offerings to maintain satisfaction and prevent attrition.

Smallest Cluster (10.5k customers): Positioned toward positive PC1 and lower PC2, correlating with higher balance accounts or elevated default risk profiles alongside distinct campaign engagement patterns. This segment demonstrates responsiveness to targeted marketing initiatives and may warrant premium product offerings, personalized communication strategies, or enhanced risk monitoring depending on their financial health indicators.

Intermediate Cluster (15.4k customers): Extends toward negative PC1 with elevated PC2, indicating distinctive feature combinations such as increased campaign contacts, specific housing-loan patterns, or particular age-balance relationships. Cross-selling opportunities and educational financial literacy programs could activate this underutilized segment, potentially migrating them toward higher-value customer behaviors through strategic engagement.

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:

The PCA scatter plot displays three color-coded regions corresponding to distinct customer profiles:

Turquoise Points: Positioned toward negative PC1 with elevated PC2 values. These customers exhibit higher campaign contact frequencies, lower account balances, and increased housing-loan presence. The algorithm groups them based on shared behavioral similarity in engagement patterns and financial product usage.

Yellow Points: Concentrated in positive PC1 with moderate PC2 regions, representing financially healthier customers characterized by stronger balances, reduced loan/default indicators, and different demographic profiles. This segment displays tight clustering reflecting high internal homogeneity in financial behaviors.

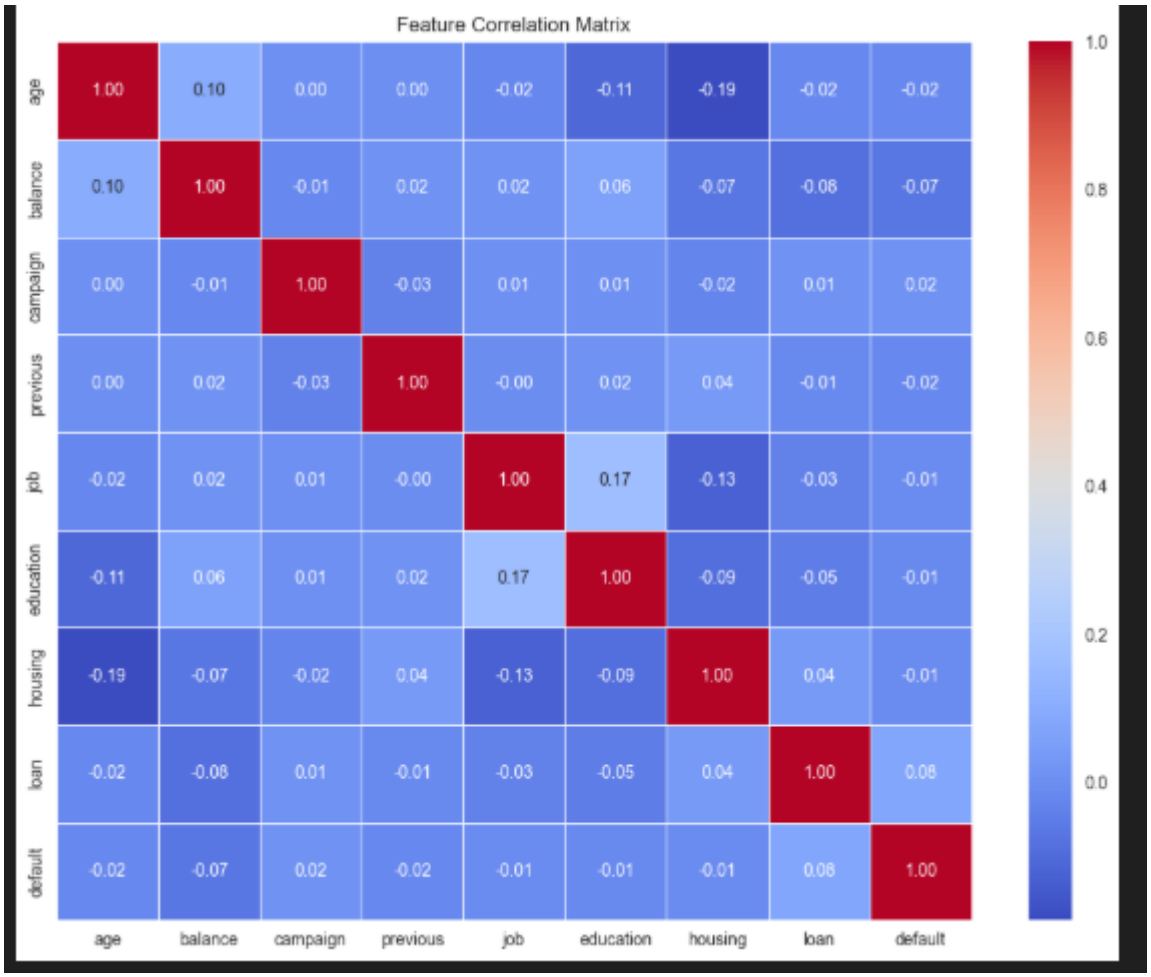
Purple Points: Clustered near the origin with modest spatial spread, corresponding to mainstream customers whose age, balance, and loan/default metrics align with overall population averages. This core segment represents typical banking behaviors without extreme characteristics.

Boundary Characteristics:

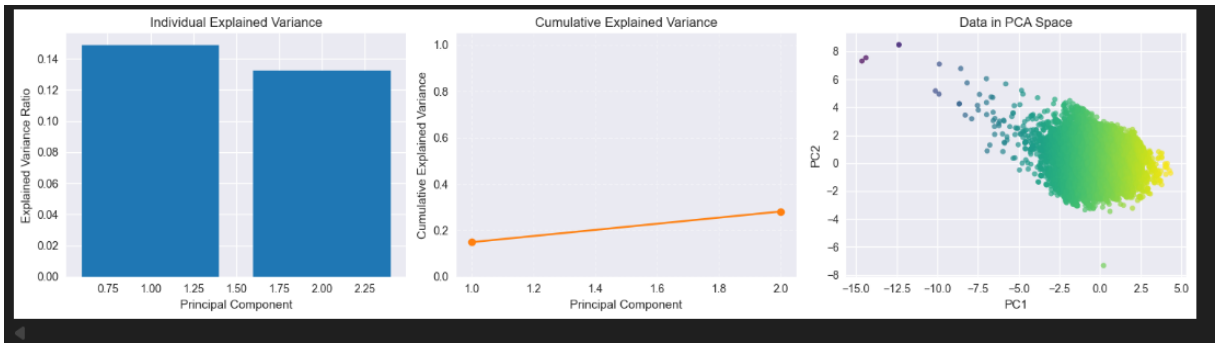
The purple-yellow boundary appears relatively **sharp** because balance and loan patterns create clear separation after standardization—customers either maintain substantially higher balances with cleaner credit profiles (yellow) or fall into average ranges (purple). This distinct financial divergence translates to well-defined cluster boundaries in PCA space.

Conversely, the turquoise-purple transition is more **diffuse** due to gradual variation in intermediate features. Many customers share overlapping balances or campaign engagement levels that don't cleanly separate. Additionally, since PCA captures only 28% of total variance, subtle distinctions present in the higher-dimensional space become compressed, causing boundary overlap even where meaningful differences exist in the original 9-dimensional feature representation.

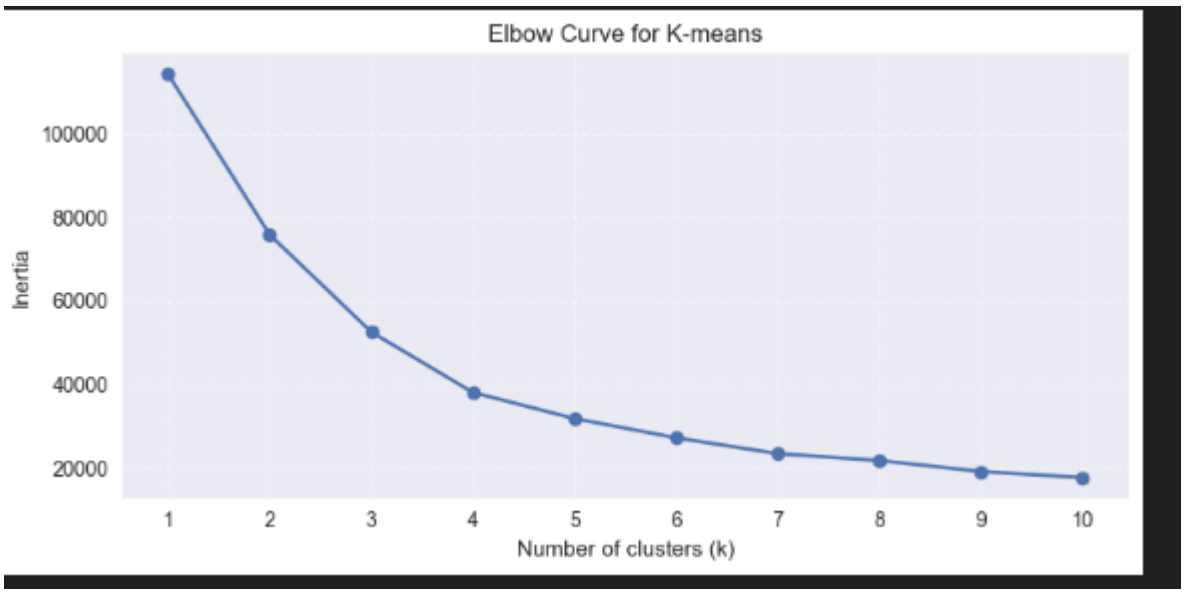
Feature Correlation Matrix:

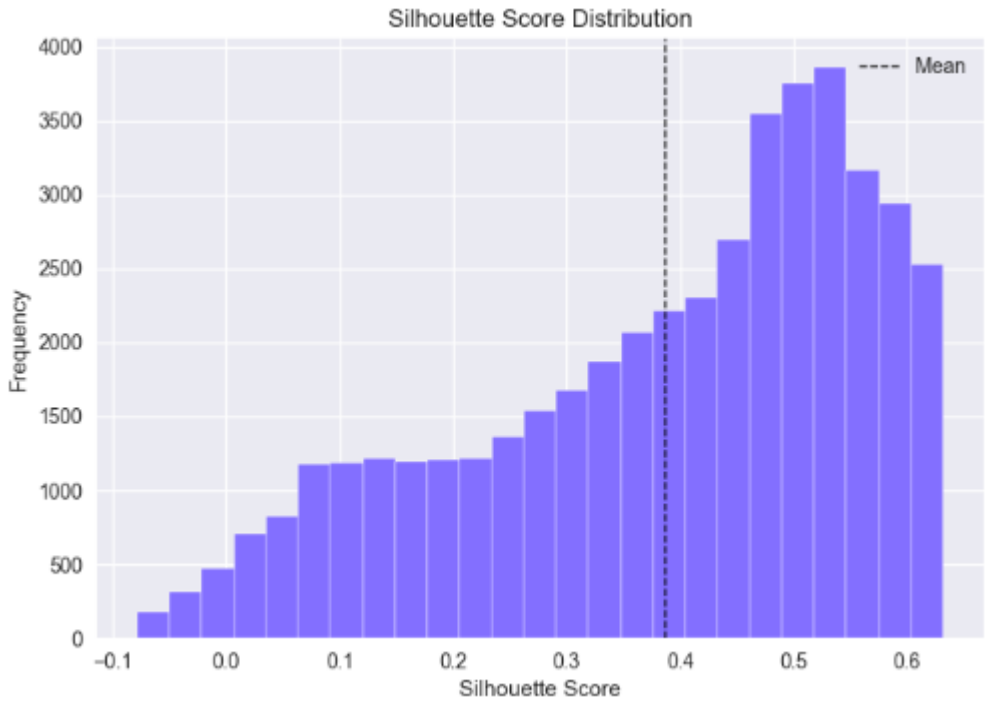


Explained variance by Component and Data Distribution in PCA Space after Dimensionality Reduction with PCA

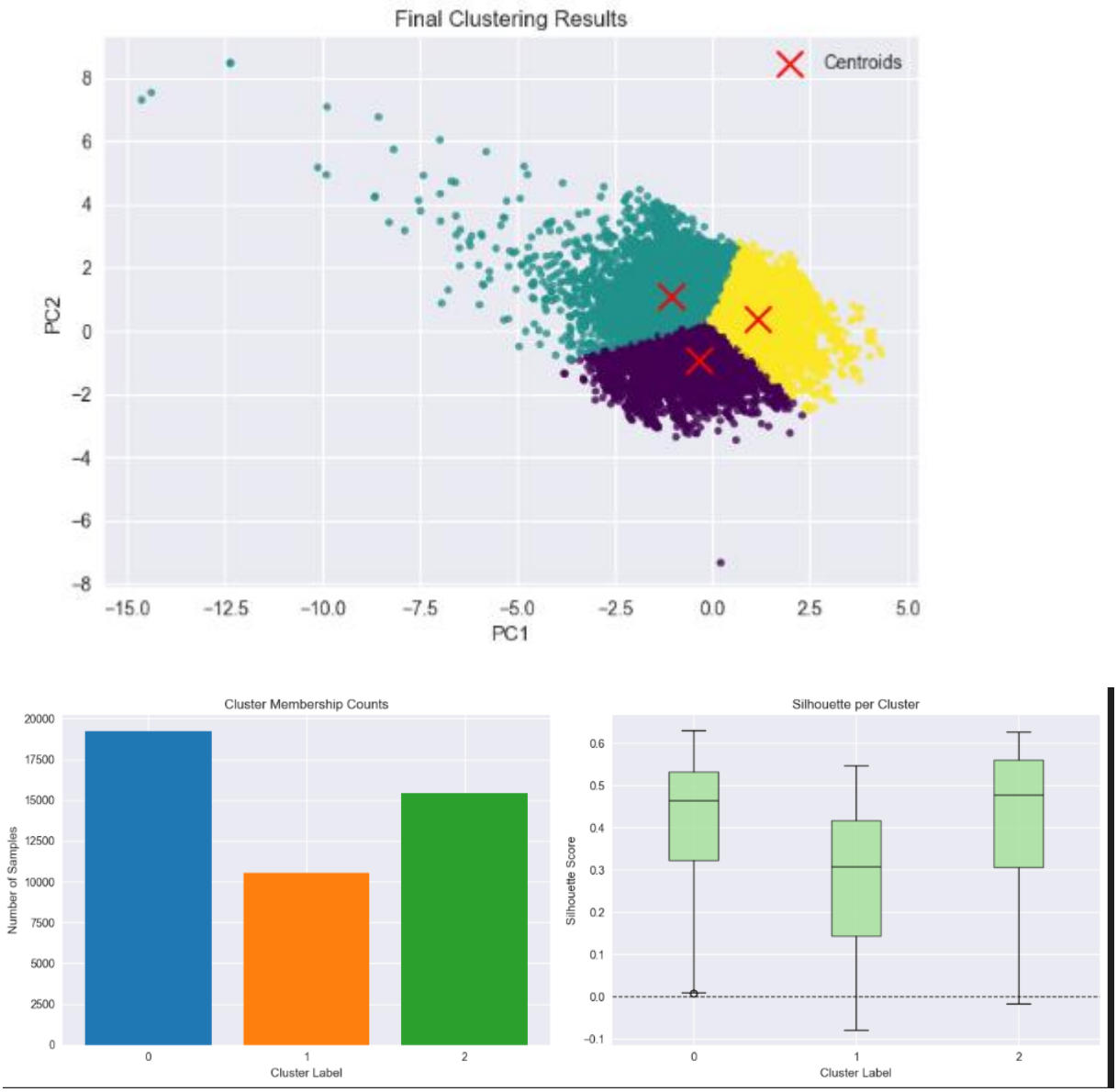


'Inertia Plot' and 'Silhoutte Score Plot' for K-means




Silhouette Score Distribution
Mean
Frequency
Silhouette Score
4000
3500
3000
2500
2000
1500
1000
500
0
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6

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



Analysis Questions: