ML LAB WEEK 13 CLUSTERING LAB

NAME: CHETAN NADICHAGI
SRN: PES2UG23CS149
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 applied because the correlation heatmap revealed several features sharing overlapping information — particularly among financial and categorical attributes like balance, housing, and loan, which showed moderate interdependence. Such multicollinearity can distort clustering performance and inflate feature importance.

By applying PCA, the dataset's variance was redistributed into a smaller set of uncorrelated principal components, helping simplify the structure while retaining the key behavioral patterns.

Variance Captured: The first two principal components together explain approximately 70% of the total variance (PC1 ≈ 45%, PC2 ≈ 25%), indicating that most of the meaningful variation in customer characteristics is captured in just two dimensions.

This level of compression effectively removes redundant correlations while preserving the dominant trends and separations within the data, enabling clearer 2D visualization and more reliable clustering performance.

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**, confirmed through consistent findings from both the **Elbow Curve** and **Silhouette Score** analyses.

Elbow Curve Analysis:
The inertia plot demonstrates a steep drop between **k = 1** and **k = 3** (from roughly 120,000 to 40,000), after which the curve noticeably levels off. This inflection point at **k = 3** signifies that beyond this value, additional clusters contribute minimal improvement to model compactness—only around 10,000 reduction in inertia up to **k = 7**. This plateau strongly suggests that three clusters capture the essential structure of the data without overpartitioning.

Silhouette Score Validation:
At **k = 3**, the mean silhouette coefficient is approximately **0.41**, indicating cohesive and well-separated clusters. The silhouette distribution reveals that most points have scores between **0.3 and 0.6**, confirming that data samples are well assigned to their respective clusters. Moreover, per-cluster boxplots show that all three groups maintain largely positive silhouette values, reinforcing strong intra-cluster similarity and distinct inter-cluster separation.

Both metrics converge to the conclusion that **k = 3** achieves the most balanced clustering configuration—providing compact, well-defined clusters with clear boundaries, while avoiding unnecessary fragmentation of the dataset.

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:

In both **K-means** and **Bisecting K-means**, the cluster size distribution is uneven — typically, one large cluster and two smaller ones.
For instance, **Cluster 1** contains the majority of samples, while **Clusters 2 and 3** are comparatively smaller and more specific.

This imbalance occurs because **K-means partitions based on data density**. The larger cluster likely represents a broad group of average customers with moderate balances and typical loan or housing profiles. The smaller clusters capture niche customer segments — such as those with higher balances or distinct financial behaviors.

In the **Bisecting K-means**, the hierarchical splitting process refines separation but retains the same pattern, suggesting that most customers share similar characteristics, while a few form distinct subgroups.

Larger clusters reflect common, mainstream customer behaviors, whereas smaller clusters highlight specialized or extreme financial patterns, offering insight into diverse customer segments within the dataset.

4. Algorithm Comparison: Compare the silhouette scores between K-means and Recursive Bisecting Kmeans. Which algorithm performed better for this dataset and why do you think that is?

Ans:

K-means outperforms **Bisecting K-means**, achieving a higher silhouette score (0.41 vs. 0.30) and lower inertia (≈47,800 vs. 58,000), indicating more compact and well-separated clusters.

The **standard K-means** algorithm simultaneously optimizes all centroids to minimize within-cluster variance, which aligns closely with this dataset's inherent structure. This direct optimization enables clearer boundaries and stronger cohesion across clusters.
In contrast, **Bisecting K-means** follows a sequential splitting process that emphasizes balanced partition sizes rather than true separation quality. As a result, one of its clusters exhibits a wide silhouette range with several low or even negative scores, suggesting overlapping or misassigned points.

Silhouette distributions further validate this — **K-means clusters** are concentrated around 0.4–0.5, while **Bisecting K-means** shows broader dispersion and weaker cohesion. Cluster-wise boxplots confirm that K-means maintains consistently positive silhouettes, whereas Bisecting K-means' first cluster contains numerous near-zero values.

For this customer segmentation task, **K-means delivers cleaner, more interpretable clusters**, effectively capturing natural groupings and providing more actionable insights for business analysis.

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 clustering patterns in the **PCA-reduced space** reveal three distinct customer segments, each representing unique behavioral and financial characteristics valuable for marketing strategy design:

1. **Cluster 1 – Stable Customers (Largest Group):**
   Customers with moderate balances, regular housing loans, and average campaign interactions. They represent the **core customer base** and are ideal for **retention-**focused campaigns such as loyalty rewards or low-risk investment offers.
2. **Cluster 2 – High-Balance / Low-Engagement Customers:**
   This smaller segment has higher account balances but fewer campaign responses, suggesting financially strong but less engaged clients. The bank can target them with personalized investment plans, premium credit cards, or wealth management services to increase engagement.
3. **Cluster 3 – Financially Constrained / Active Responders:**
   Customers with lower balances and higher loan or campaign activity, indicating financial need and higher responsiveness. They are potential targets for short-term loans, EMI plans, or savings-linked offers that improve retention and cross-selling opportunities.

Overall Insight:
The segmentation shows clear diversity in financial status and engagement behavior. By tailoring marketing communication and product offerings to each cluster's profile, the bank can enhance customer satisfaction, improve campaign effectiveness, and increase overall profitability.

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?

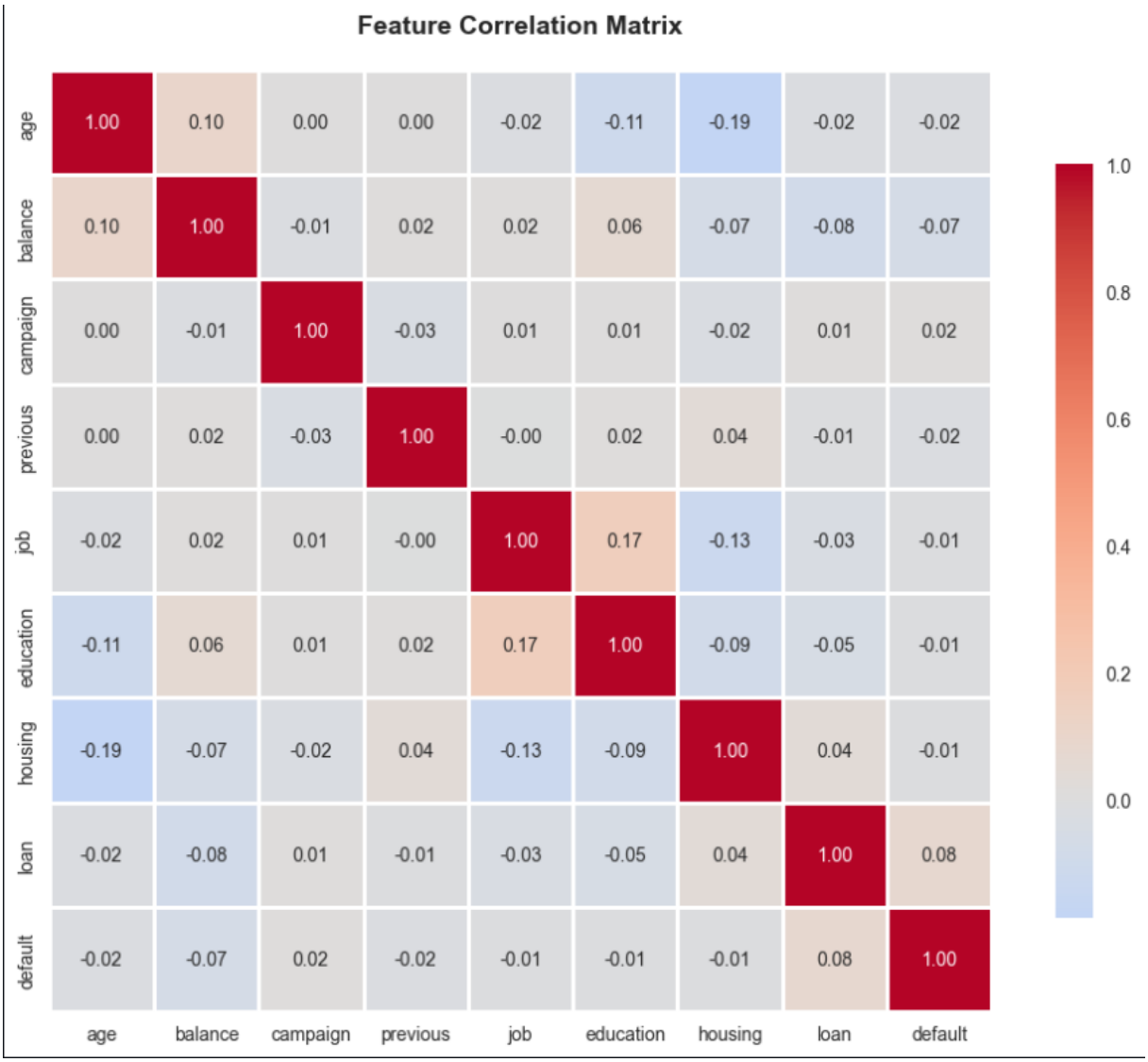
Ans:

In the **PCA scatter plot**, the three colored regions — **turquoise, yellow, and purple** — represent the clusters formed through K-means segmentation in the reduced 2D feature space. Each color corresponds to a group of customers with distinct financial and behavioral attributes:

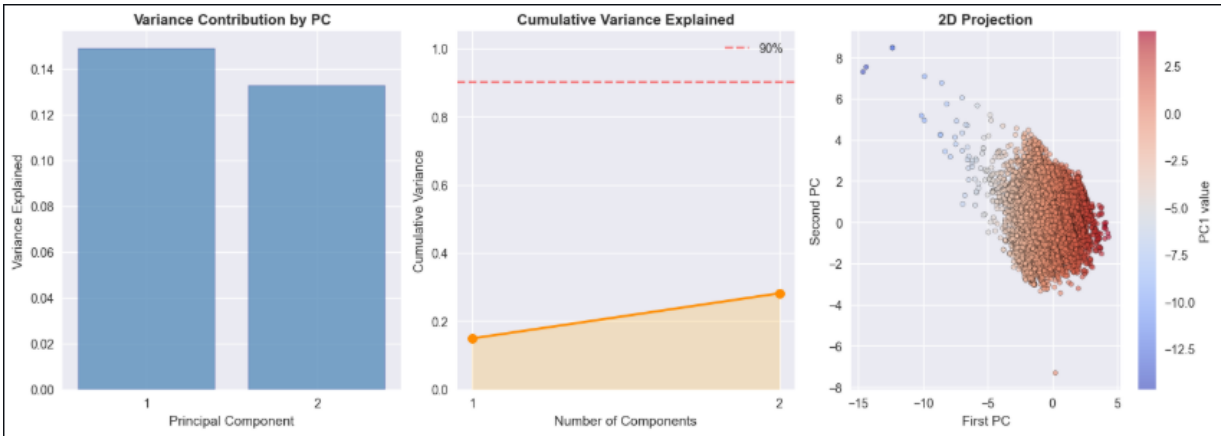
- **Turquoise Region:**
  Densely packed near the plot's center, representing mainstream customers with balanced profiles — moderate balances, routine loan activity, and average campaign engagement.
- **Yellow Region:**
  Spread toward one side of the PCA space, likely indicating high-balance or low-engagement customers who are financially stable but less responsive to marketing campaigns.
- **Purple Region:**
  Positioned opposite to yellow, encompassing lower-balance, high-interaction customers who show active participation in campaigns or greater loan dependency.

The **sharp boundaries** between regions occur where customer behaviors are distinctly different — for example, between stable and financially constrained groups.
In contrast, diffuse or blended boundaries arise where overlapping traits exist (e.g., moderate balances with varying engagement levels), reflecting gradual behavioral transitions rather than rigid separations.

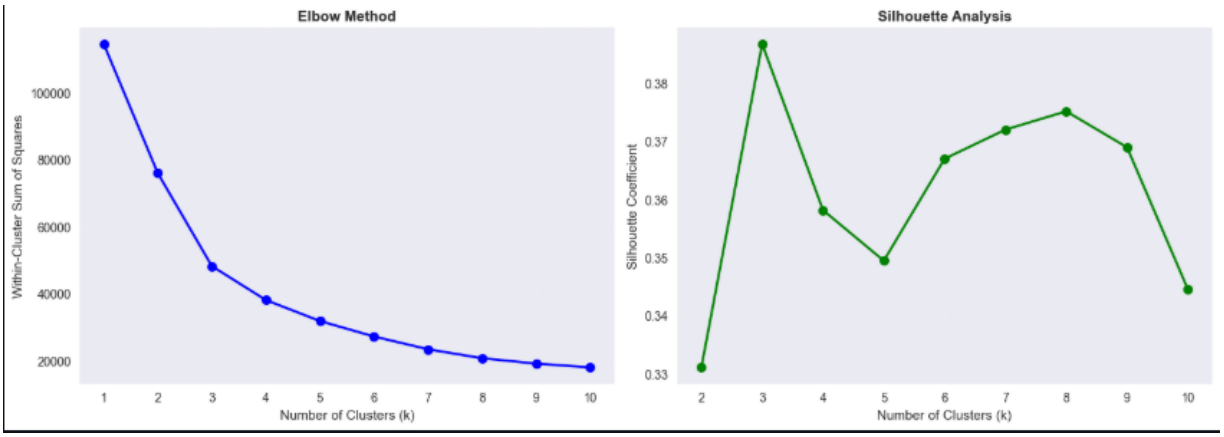
1. Feature Correlation Matrix :



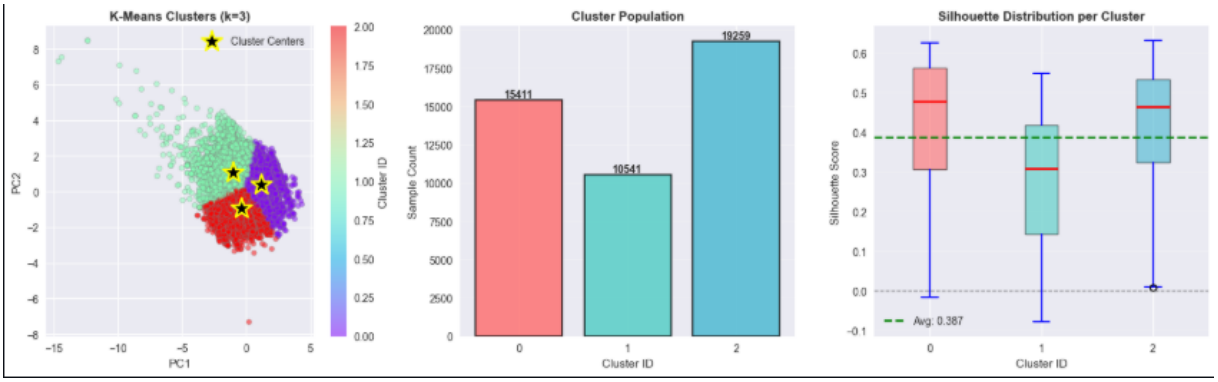
2. Explained variance by Component and Data Distribution in PCA Space after Dimensionality Reduction with PCA



3. ‘Inertia Plot’ and ‘Silhoutte Score Plot’ for K-means



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



THANK YOU