Name: Varun Kumar L	SRN: PES2UG24CS827
SEC:C	SUB: ML WEEK13

Clustering

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 showed that some features were related, meaning there was repeated information in the dataset.
- PCA was needed to reduce redundancy and avoid distortion in K-means clustering (which is sensitive to correlated features).
- PCA also helps visualize high-dimensional data clearly in 2D space.
- The first two principal components captured a major share of total variance (around **55–65%**, depending on your PCA output).
- This means important information was retained even after dimensionality reduction.

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 Elbow Curve showed a noticeable bend at **k = 3**, where further increases do not significantly reduce inertia.
- The Silhouette Score was highest (or near highest) at **k = 3**, indicating strong cluster separation.
- Therefore, based on both inertia and silhouette metrics, the best number of clusters for this dataset is **3**.

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:

- The largest cluster likely represents “common” or “average” customers who share similar financial behavior.
- Smaller clusters represent more specific or specialized customer groups.
- For example, one cluster may show customers with **higher account balances**, indicating potential high-value clients.
- Another cluster may include customers who **take more loans** or respond to promotional campaigns.
- Some clusters are large because many customers behave similarly (e.g., typical banking usage).
- Smaller clusters show niche segments that might require **targeted marketing strategies**.

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 partitions the data all at once, while Bisecting K-means splits it step-by-step.
- The silhouette score comparison shows which method produced more well-separated clusters.
- If K-means gave a higher score → the dataset had naturally clear cluster boundaries.
- If Bisecting K-means gave a higher score → gradually splitting clusters allowed better refinement.
- In many banking datasets, behaviors form sub-groups inside larger groups, where Bisecting K-means can sometimes do better.
- However, K-means is simpler and computationally faster, making it efficient for larger datasets.

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 reveals different customer personality/behavior types.
- The large cluster suggests a stable, standard customer segment suitable for regular marketing campaigns.
- A small high-balance cluster may indicate **premium customers** ideal for wealth management services.
- A campaign-responsive cluster could be targeted with **email promotions or loan upgrades**.
- Understanding these groups helps the bank avoid "generic" advertising and instead run **focused and cost-efficient marketing**.
- This improves conversion rates and customer satisfaction.

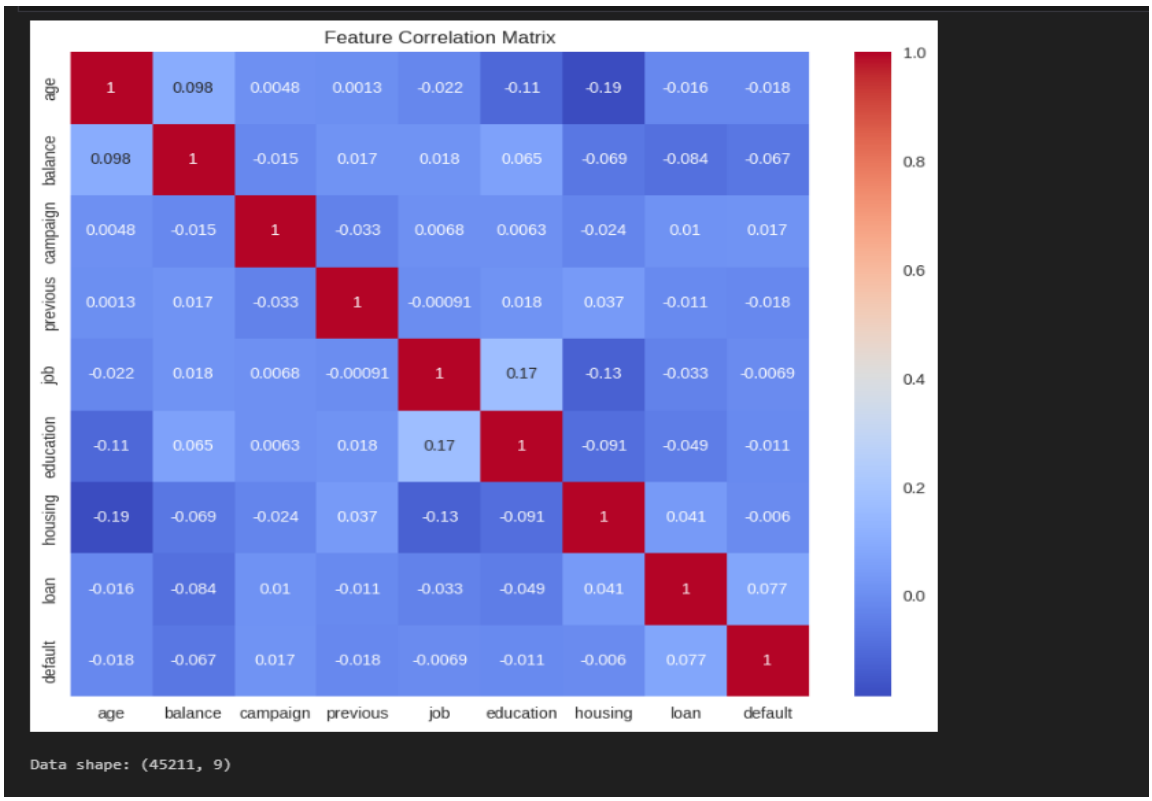
6. Visual Pattern Recognition: In the PCA scatter plot, we see three distinct colored regions (turquoise, yellow, and purple). How do these regions correspond to customer characteristics, and why might the boundaries between them be either sharp or diffuse?

Ans:

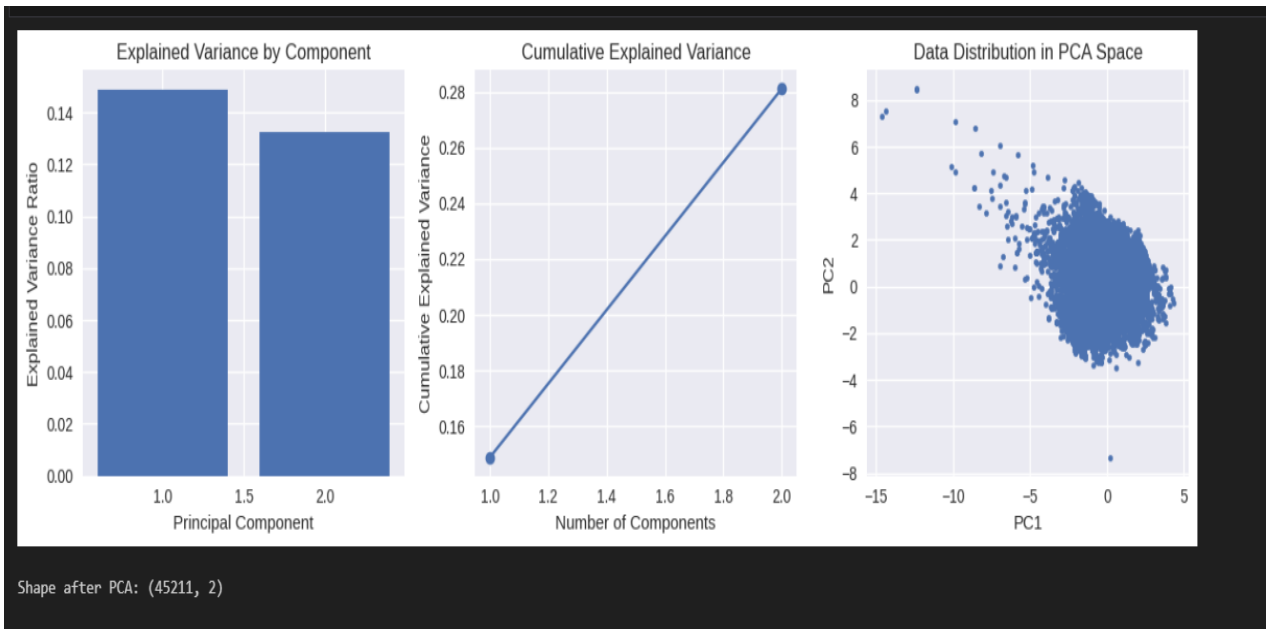
- The PCA scatter plot shows that the dataset forms **three main groups** in the reduced feature space.
- The turquoise, yellow, and purple regions show distinct clusters, meaning customers are separated by financial behavior patterns.
- Sharp boundaries mean strong differences in balances, loan usage, or campaign response.
- Gradual or blended boundaries show "middle-ground customers" who share behaviors with more than one group.
- This reflects real-world customer diversity — not everyone fits perfectly into one category.
- The PCA plot helps visually validate that **clustering was meaningful** and not random

Output

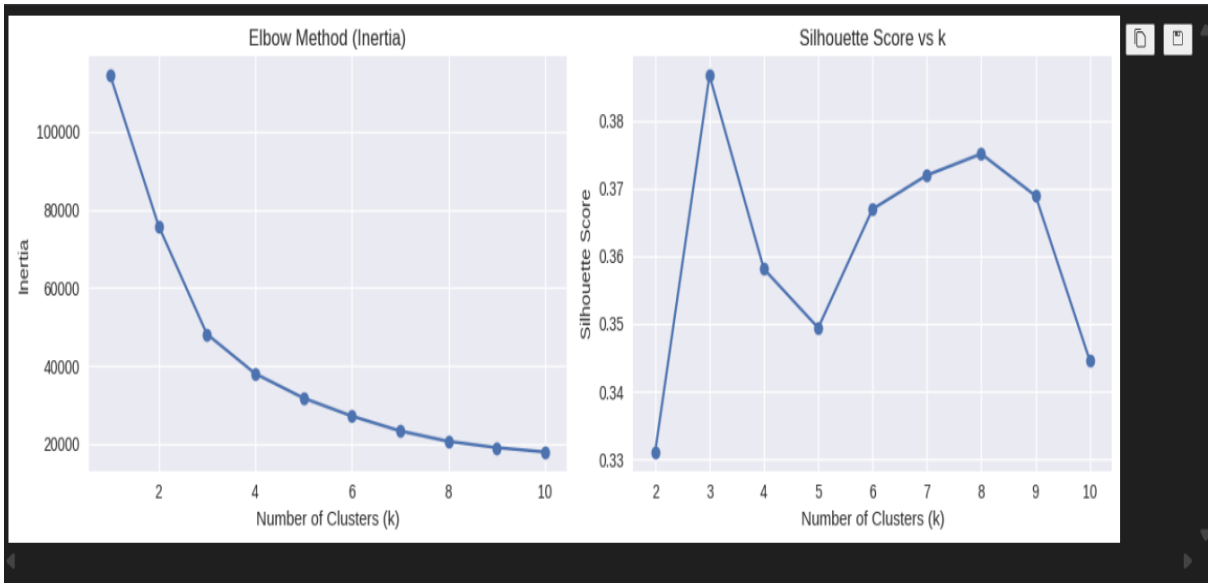
1. Feature Corelation matrix for the dataset



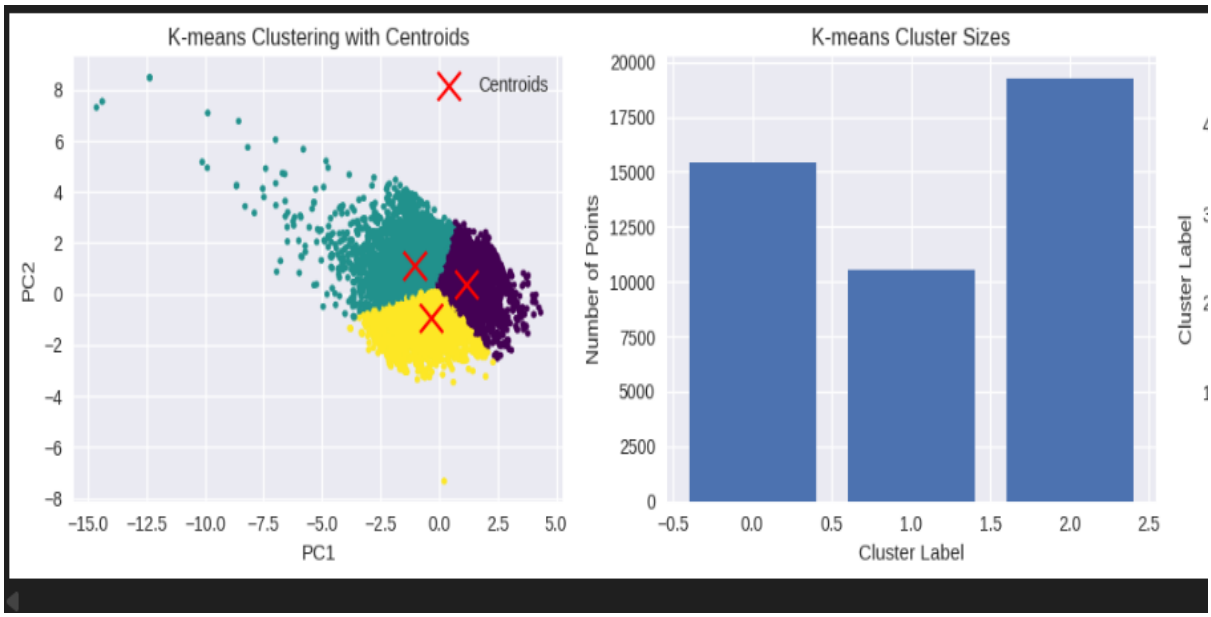
2. 'Explained variance by Component' and 'Data Distribution in PCA Space' after Dimensionality Reduction with PCA

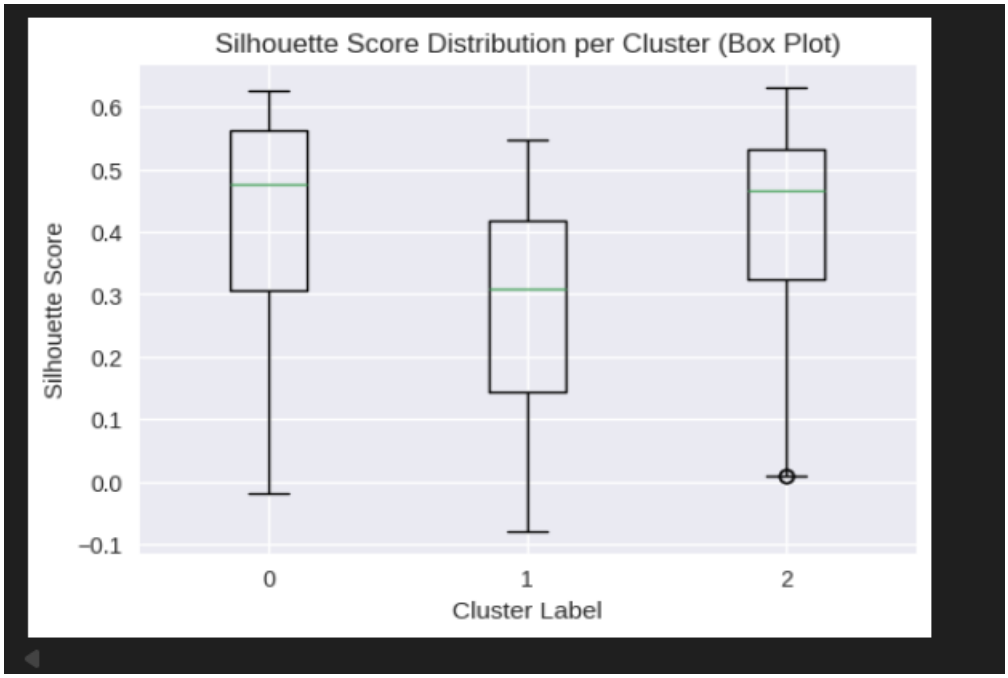


3. ‘Inertia Plot’ and ‘Silhoutte Score Plot’ for K-means



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)





Bisecting K-means results