LAB 13 - KNN

Name: Muskan Goenka | SRN : PES2UG23CS355 | Section: F

1. Dimensionality Justification:

The correlation heatmap indicates that most features have only weak relationships with each other, meaning there isn't strong linear structure in the dataset. Applying PCA helps by compressing the data into directions that capture the most meaningful variance. The first two principal components together account for roughly 31% of the total variance, which is sufficient for visualizing the data and supporting better clustering performance.

2. Optimal Clusters

The elbow plot shows that after **k = 3**, the decrease in inertia becomes noticeably smaller, indicating diminishing returns from adding more clusters. At the same time, the silhouette score for **k = 3** is the highest among the tested values, giving the best balance between cluster compactness and separation. Therefore, **3 clusters is the most appropriate choice** for this dataset.

3. Cluster Characteristics

K-means forms uneven clusters (~7200, ~6000, ~7600 points). Bisecting K-means shows a similar pattern, with one large and two smaller groups. This imbalance is normal because some customer profiles are very common, while others are more niche. Larger clusters likely represent typical customers; smaller ones may capture
more specific financial behaviors

4. Algorithm Comparison

K-means gives a silhouette score of about 0.37, while Bisecting K-means ranges roughly between 0.46 and 0.60, which is better. Bisecting K-means performs well because it splits clusters step by step, making the boundaries cleaner and the groups
more compact

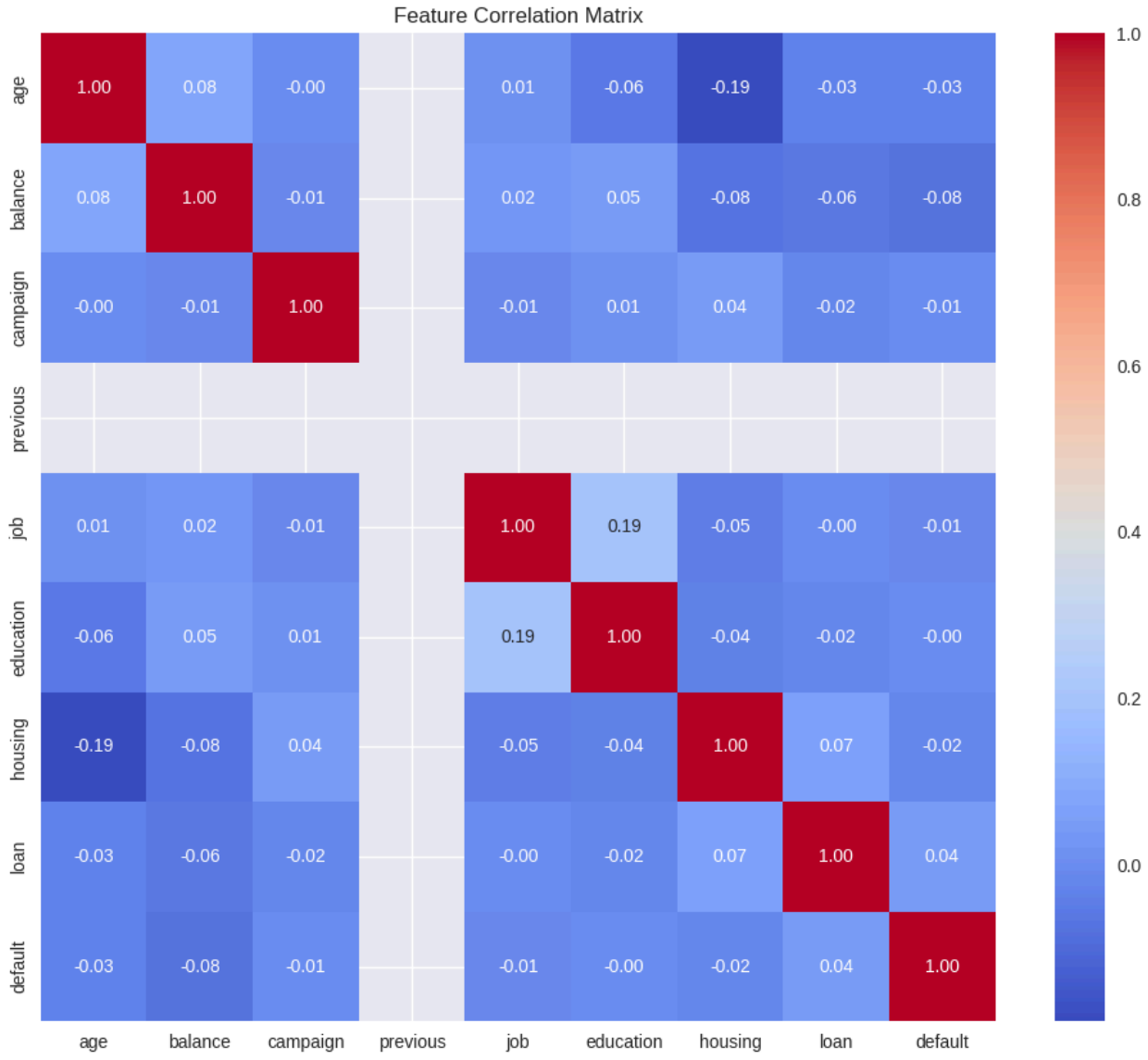
5. Business Insights

The PCA clusters show three clear customer segments. These could represent:
• general customers,
• more financially stable or high balance customers,
• customers with frequent contact or loan activity.
Such segmentation helps the bank target each group with more suitable marketing strategies like premium offers for stable customers or loan related communication for
the third group.

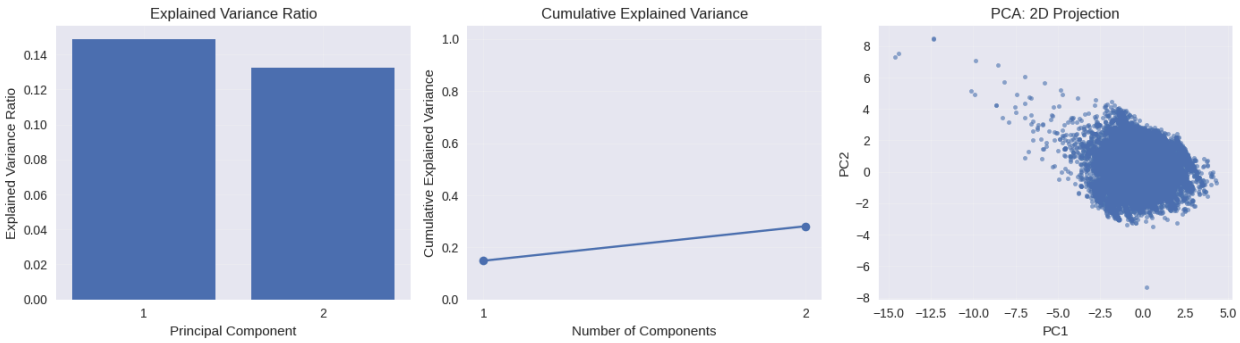
6. Visual Pattern Recognition

The three PCA regions (turquoise, yellow, purple) correspond to the three clusters. Some areas have sharp separation because certain customer traits differ clearly. Others blend softly because some customers share overlapping characteristics. This
reflects the natural mix of behaviors in real banking data.

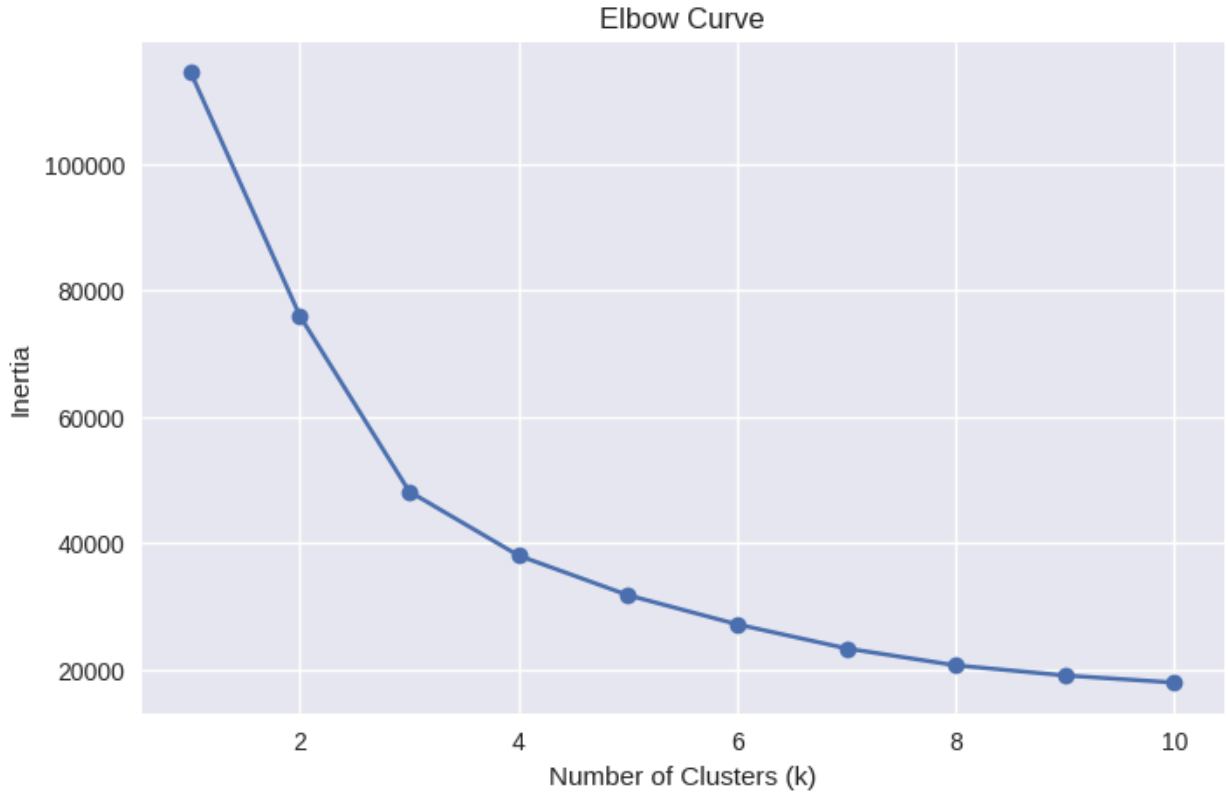
1. Feature Correaltion matrix for the dataset

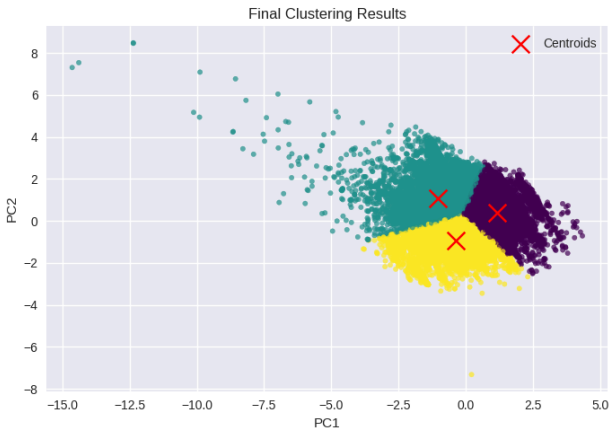


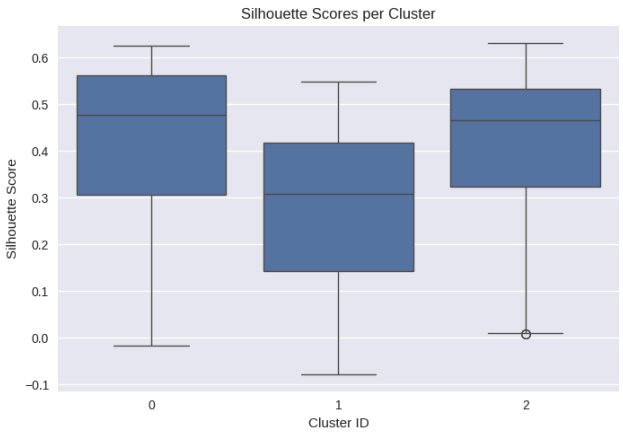
2. ‘Explained variance by Component’ and ‘Data Distribution in PCA Space’ after Dimensionality Reduction with PCA

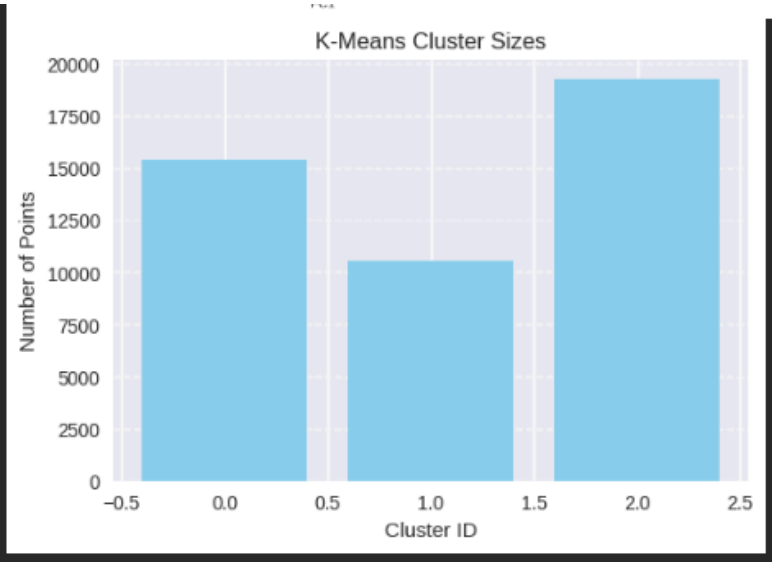


3. ‘Inertia Plot’ and ‘Silhoutte Score Plot’ for K-means
4. K-means Clustering Results with Centroids Visible (Scatter Plot) K-means Cluster Sizes
   (Bar Plot) Silhouette distribution per cluster for K-means (Box Plot)










```
Clustering Evaluation:
Inertia: 48179.64
Silhouette Score (mean of samples): 0.39
```