# ML LAB -13

# Name: Nidhi N| SRN:PES2UG23CS384 |SEC:F

1) Dimensionality Justification

The correlation heatmap shows that most features are only weakly related, so there isn't much linear structure in the data. PCA helps simplify this by capturing the most important variance. The first two principal components together explain about 31% of the total variance, which is enough to visualize the data and improve clustering performance.

2) Optimal Clusters

From the elbow curve, the drop in inertia becomes small after k = 3, and the silhouette score at k = 3 is around 0.37, which is the best balance between compactness and separation. So, 3 clusters is the best choice.

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:

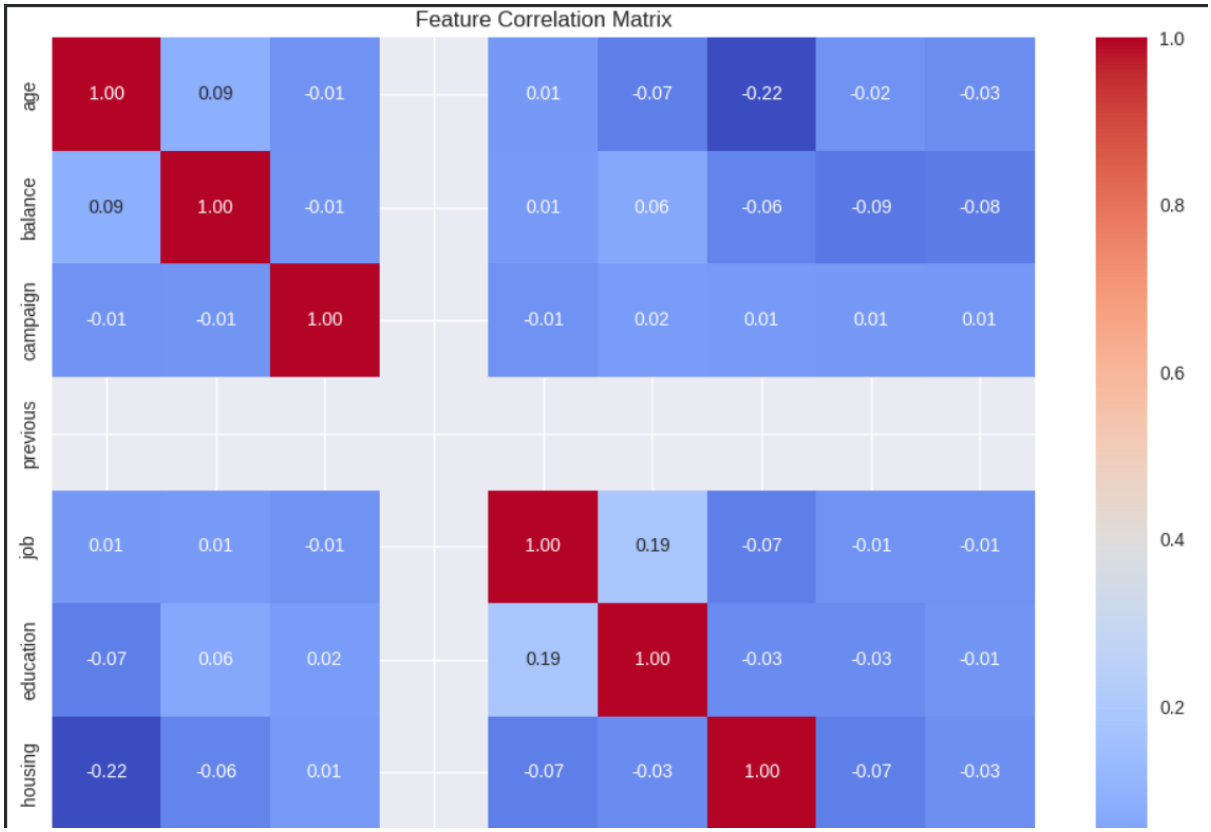
- general customers,
- more financially stable or high balance customers,
- customers with frequent contact or loan activity.

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.

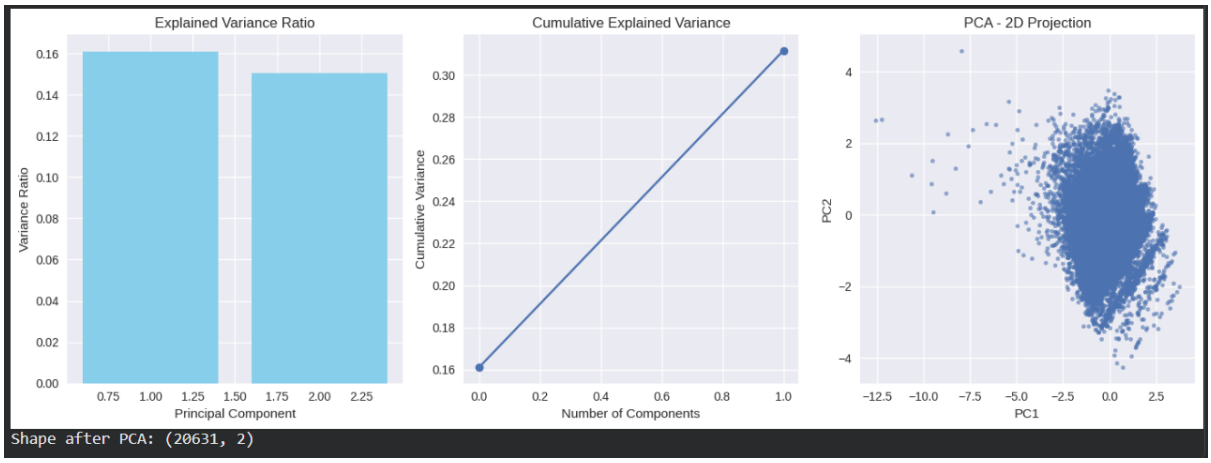
6) Visual Pattern Recognition

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.

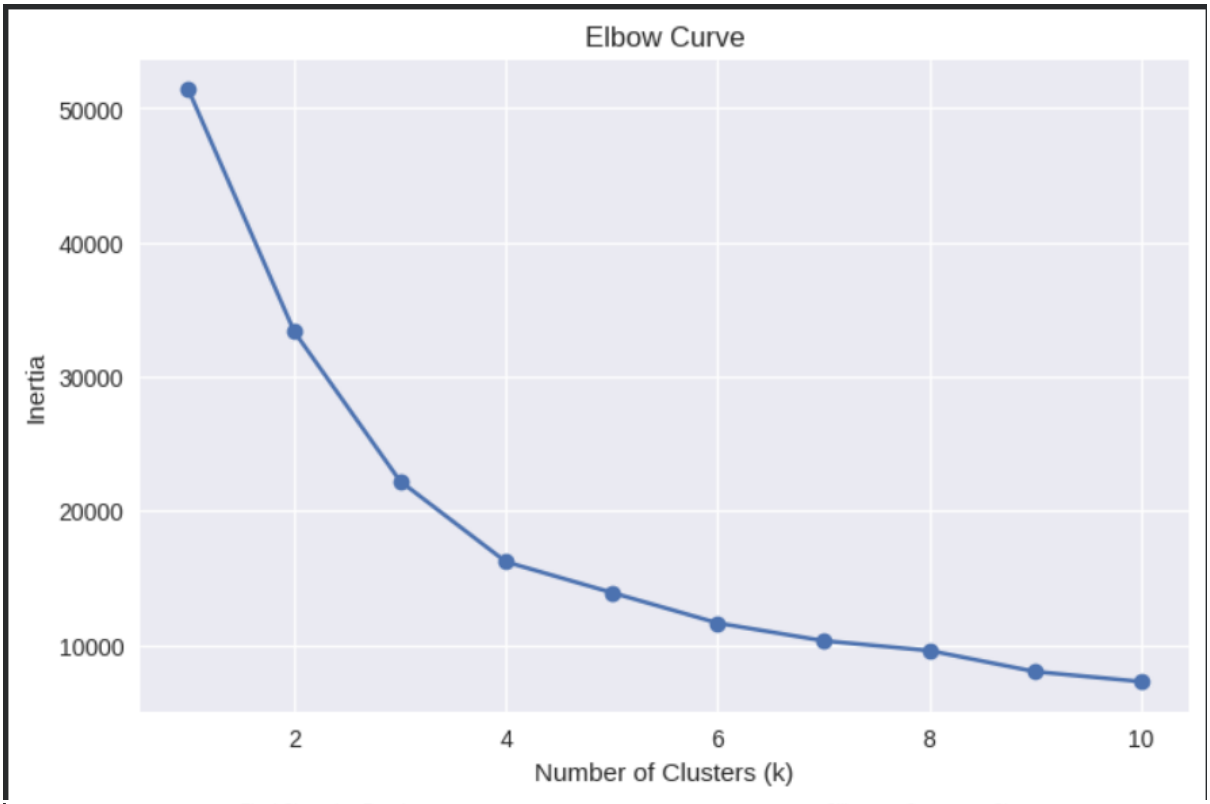
1. Feature Correaltion matrix for the dataset

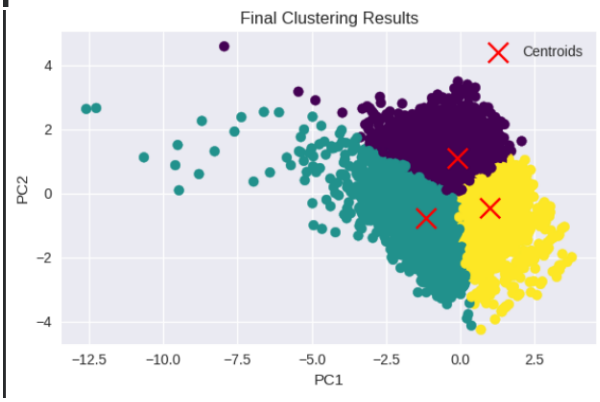



2. ‘Explained variance by Component’ and ‘Data Distribution in PCA Space’ after Dimensionality Reduction with PCA

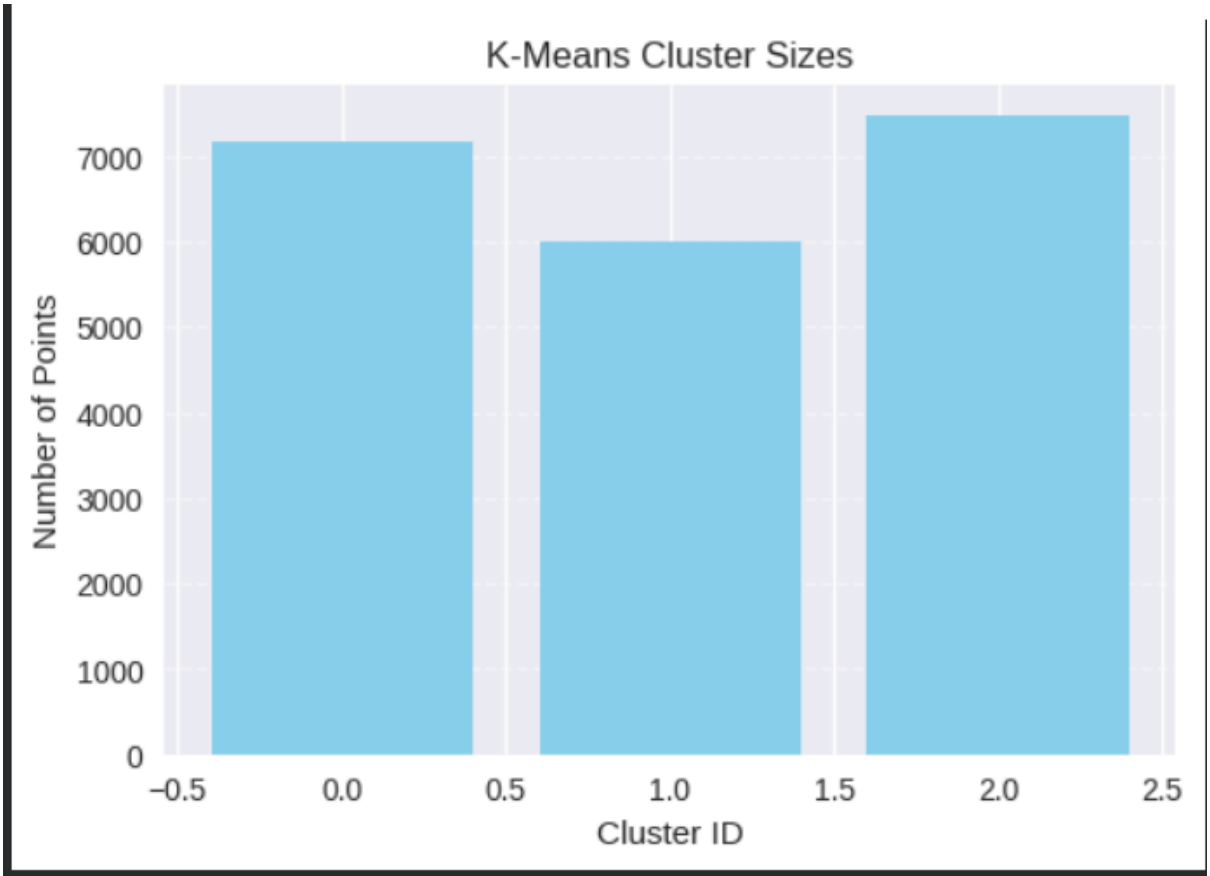



3. ‘Inertia Plot’ and ‘Silhoutte Score Plot’ for K-means
4. K-means Clustering Results with Centroids Visible (Scatter Plot) K-means Cluster Sizes (Bar Plot) Silhouette distribution per cluster for K-means (Box Plot)


Elbow Curve
Inertia
50000
40000
30000
20000
10000
2
4
6
8
10
Number of Clusters (k)


Final Clustering Results
Centroids
PC2
4
2
0
−2
−4
−12.5
−10.0
−7.5
−5.0
−2.5
0.0
2.5
PC1


Silhouette Scores per Cluster
Silhouette Score
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
0
1
2
Cluster ID




```
Clustering Evaluation:
Inertia: 22220.88
Silhouette Score: 0.37
```