*FULL NAME: MRUNAL MANJUNATH KUDTARKAR*

*SRN: PES2UG23CS354*

*SECTION: F*

**ANALYSIS QUESTIONS**

**1.**

Dimensionality reduction was necessary because the dataset contained several correlated features, which can distort distance-based clustering. PCA helps by transforming the data into uncorrelated components, improving clustering performance and enabling meaningful 2D visualization.

The first two principal components capture approximately 28**%** of the total variance (around 14% each). While not high, this is sufficient for visualizing cluster separation in 2D space.

**2.**

The elbow curve shows a sharp drop in inertia up to k=3, after which the decrease becomes gradual. This suggests that 3 clusters are optimal, balancing compactness and simplicity.

Silhouette scores peak at k=3, indicating that this configuration yields the best separation between clusters. Scores decline for higher k values, suggesting increased overlap.

The optimal number of clusters is 3, supported by both the elbow method and silhouette analysis. This choice offers a good trade-off between intra-cluster cohesion and inter-cluster separation.

**3.**

In K-means, clusters are moderately balanced but not equal. In Bisecting K-means, the recursive splitting often leads to more uneven cluster sizes, especially if dominant clusters are split late.

Larger clusters suggest that many customers share similar financial profiles—such as age, balance, or loan status—indicating common behavioral patterns. Smaller clusters may represent niche or outlier segments, useful for targeted marketing.

**4.**

K-means achieved a higher silhouette score (0.39) compared to Bisecting K-means, indicating better-defined clusters. This is likely because K-means optimizes all clusters simultaneously, while Bisecting K-means may inherit suboptimal splits from earlier stages.

**5.**

The clusters reveal distinct customer segments:

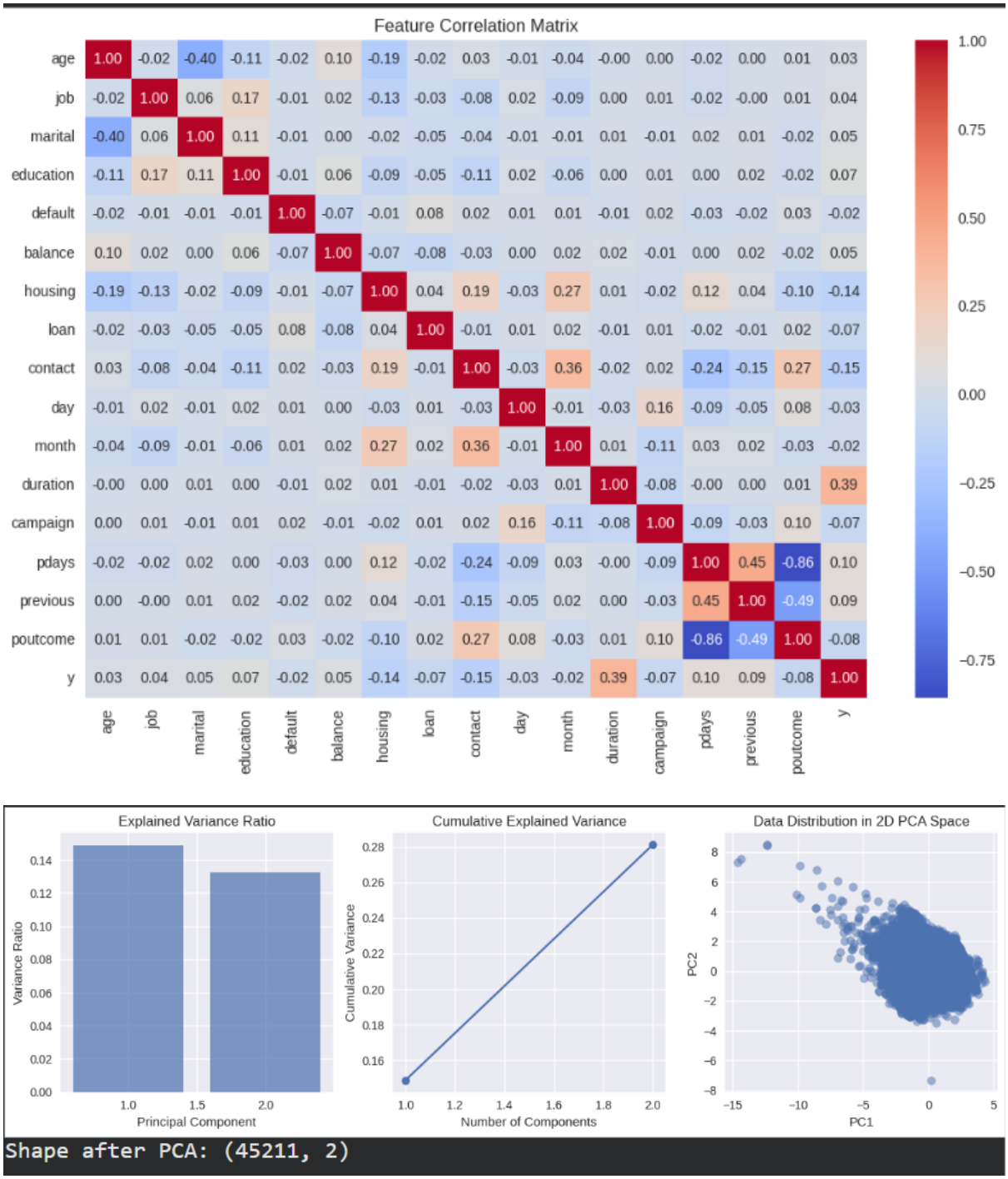
- One group may represent stable savers with high balances.
- Another may include frequent campaign responders.
- A third might consist of customers with loans or defaults.

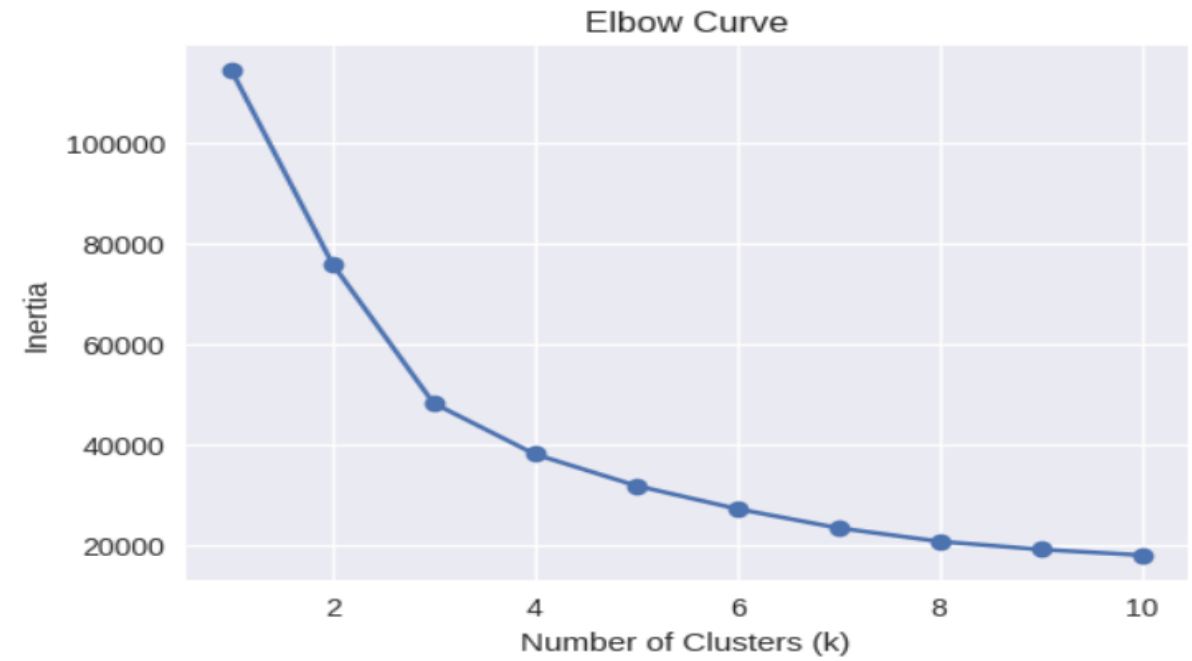
These insights can guide personalized marketing, such as offering savings plans to one group and loan restructuring to another.

**6.**

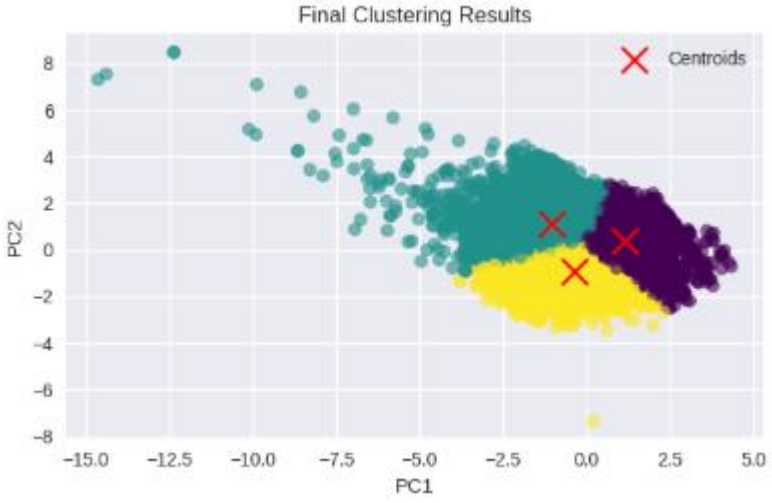
The turquoise, yellow, and purple regions likely correspond to distinct customer traits (like employment type, financial status). Sharp boundaries suggest strong feature separation, while diffuse edges indicate overlapping or transitional profiles common in real-world data.

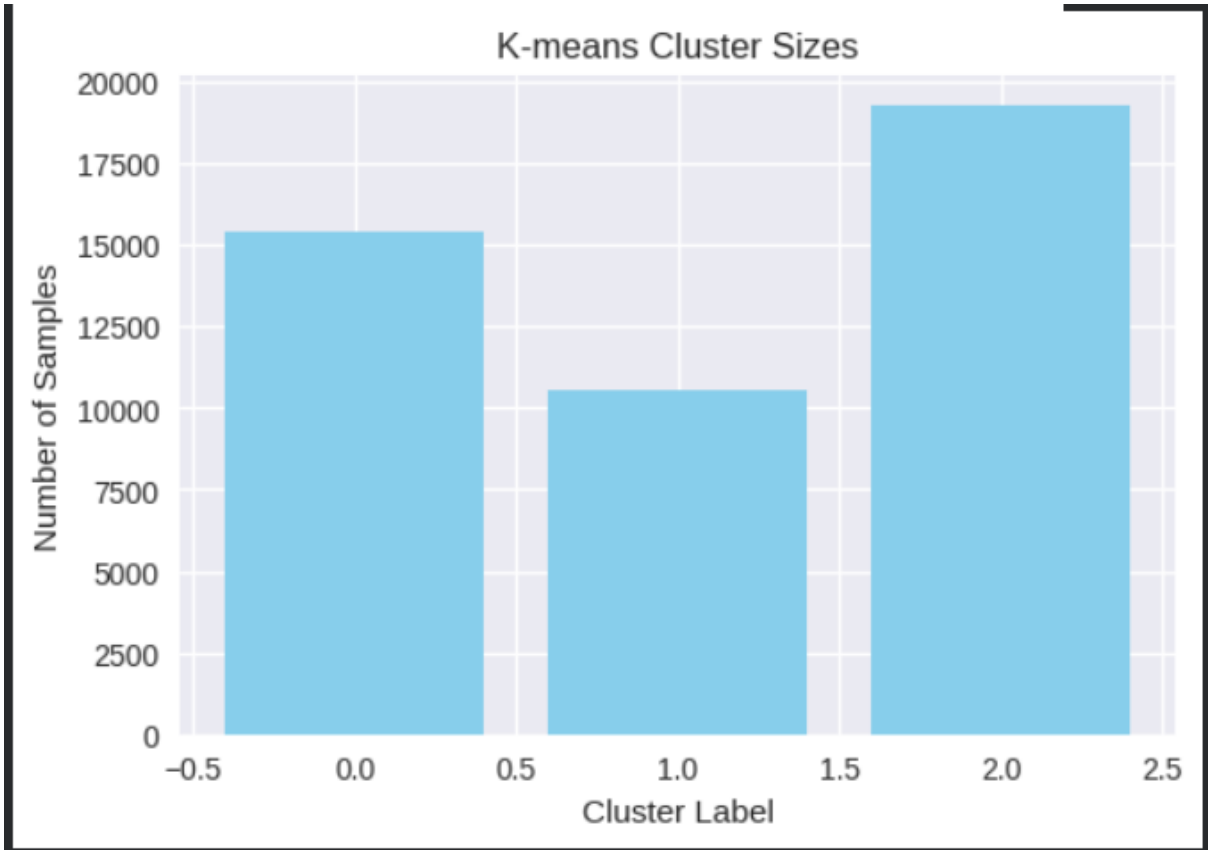
**SCREENSHOTS:**

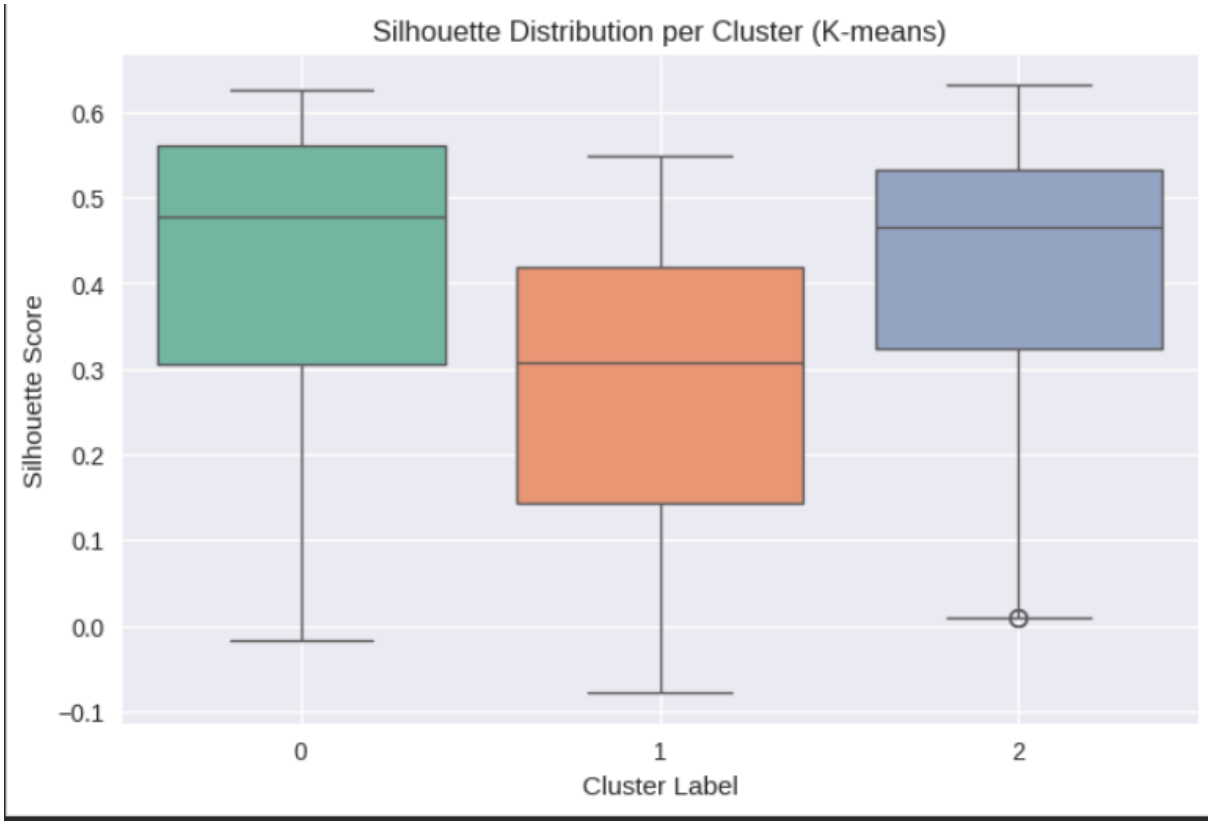



Elbow Curve
Inertia
100000
80000
60000
40000
20000
2
4
6
8
10
Number of Clusters (k)


Silhouette Scores
Score
0.38
0.37
0.36
0.35
0.34
0.33
2
3
4
5
6
7
8
9
10
Number of Clusters (k)


Final Clustering Results
Centroids
PC2
8
6
4
2
0
−2
−4
−6
−8
−15.0
−12.5
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
PC1


K-means Cluster Sizes
Number of Samples
20000
17500
15000
12500
10000
7500
5000
2500
0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
Cluster Label


Silhouette Distribution per Cluster (K-means)
Silhouette Score
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
0
1
2
Cluster Label