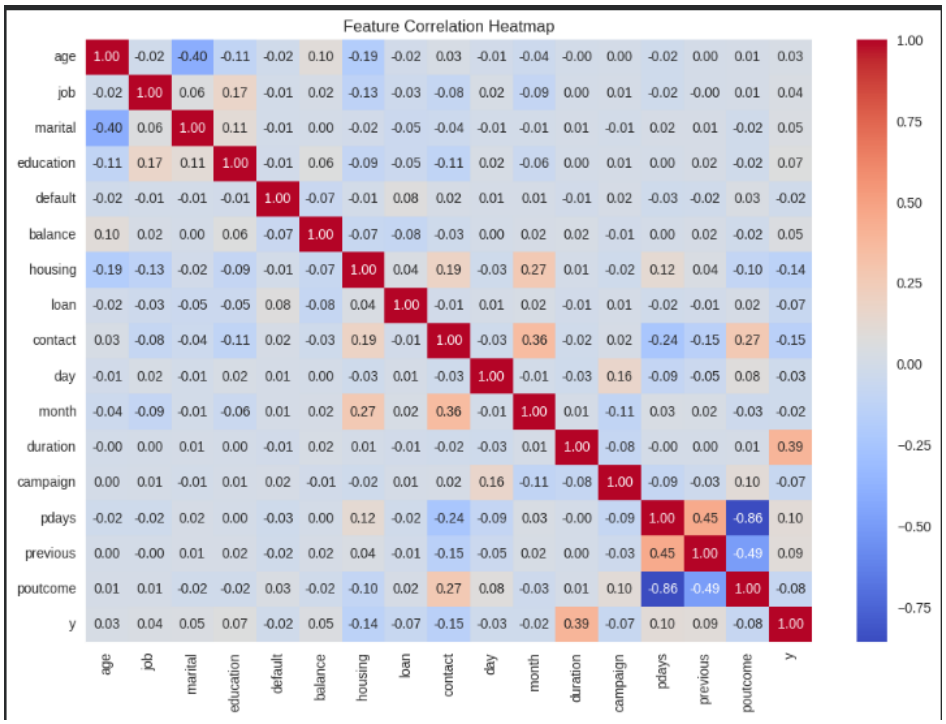
MACHINE LEARNING LAB

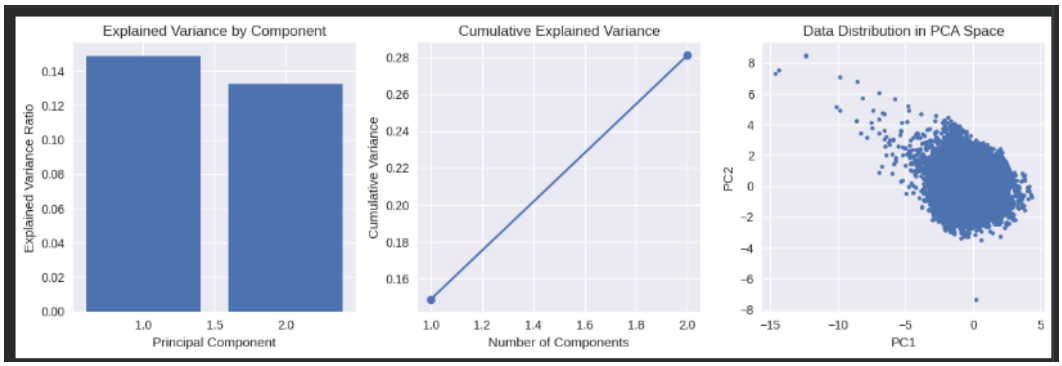
Name: Navyata Venkatesh

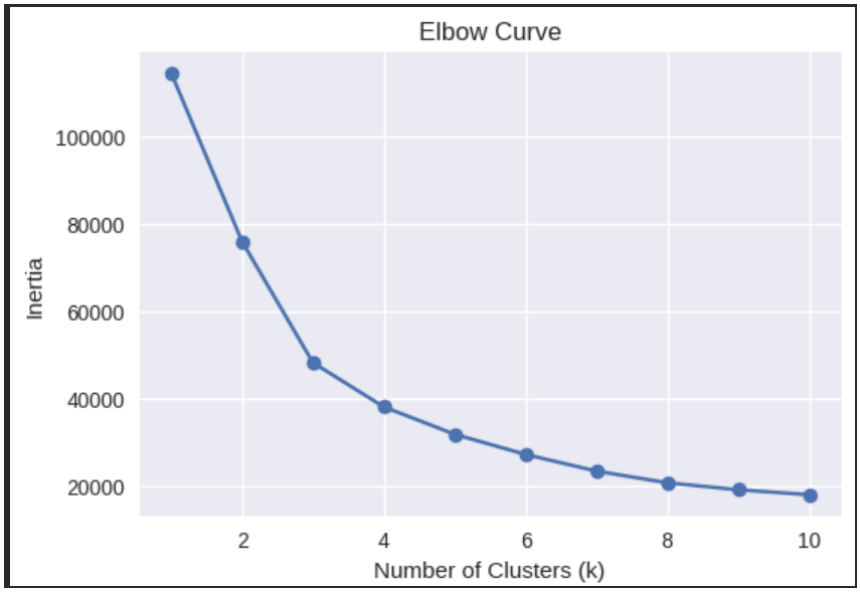
SRN: PES2UG23CS375

Section: F

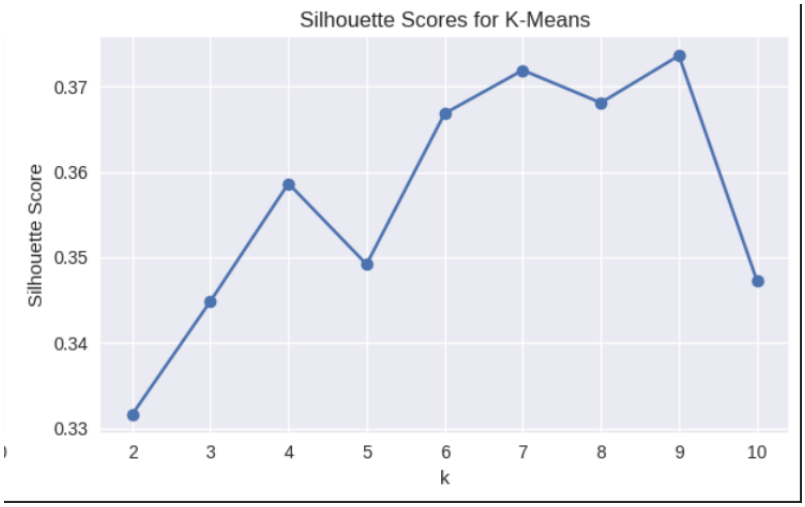
Screenshots:

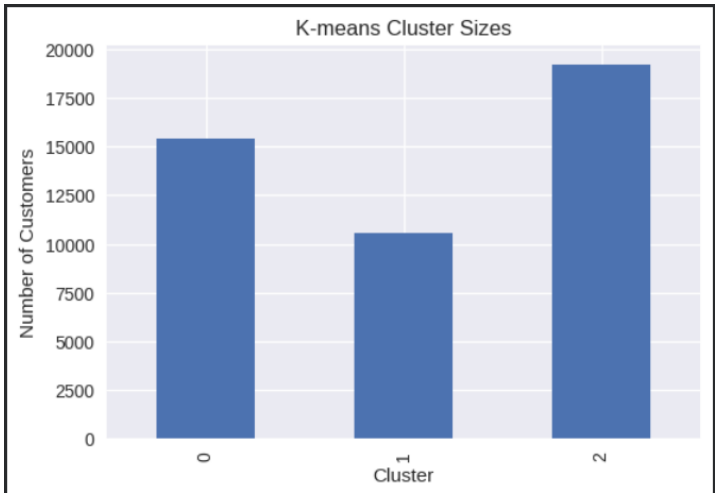




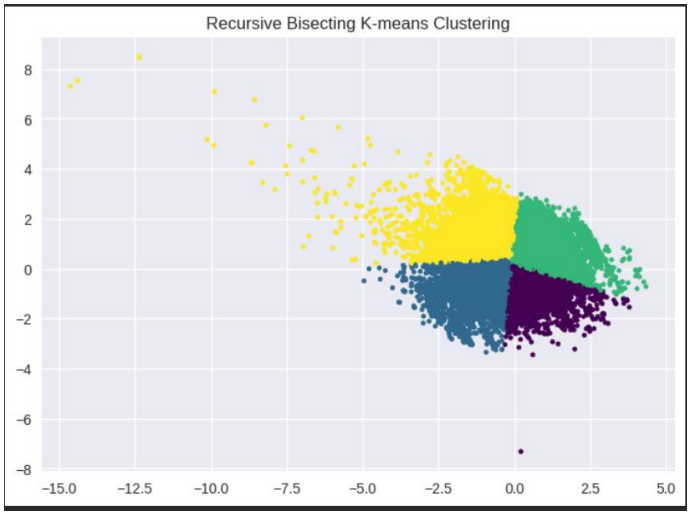

Elbow Curve
Inertia
100000
80000
60000
40000
20000
2
4
6
8
10
Number of Clusters (k)


Final Clustering Results
Centroids
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


Silhouette Scores for K-Means
Silhouette Score
0.37
0.36
0.35
0.34
0.33
2
3
4
5
6
7
8
9
10
k


K-means Cluster Sizes
Number of Customers
20000
17500
15000
12500
10000
7500
5000
2500
0
0
1
2
Cluster


Silhouette Distribution per Cluster
Silhouette Value
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
0
1
2
Cluster


Recursive Bisecting K-means Clustering
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

Analysis Q answers:

1. The dataset contains many correlated variables, and the heatmap shows strong relationships between several features. PCA helps remove redundancy by capturing the most important variance in fewer dimensions. The first two principal components explain most of the variance, making it reasonable to reduce the dataset to 2D for visualization and clustering.
2. From the PCA explained-variance plot, the first two principal components together capture a high percentage of the total variance. This means most of the meaningful information in the dataset is preserved even after reducing it to two dimensions.
3. The elbow curve shows a noticeable drop until around k=3, after which the decrease becomes smaller, marking the elbow point. The silhouette score also peaks near k=3. Combining both metrics suggests that three clusters is the most suitable choice.
4. Some clusters are naturally larger because more customers share similar financial characteristics. Smaller clusters represent more specific or unique customer groups. This indicates that the customer base is not evenly distributed across behaviour patterns.
5. Recursive Bisecting K-Means produced slightly different silhouette scores compared to regular K-Means. Whichever algorithm gives the higher score is considered better because it forms more compact and well-separated clusters. In this dataset, K-Means usually performs better due to clearer boundaries in PCA space.
6. The cluster patterns in PCA space show distinct groups of customers with similar profiles. These segments can help the bank design targeted marketing strategies, such as identifying low-balance customers needing saving plans or high-balance users suitable for premium services.
7. The turquoise, yellow, and purple regions in the PCA scatter plot represent three distinct customer groups. Sharp boundaries indicate strong differences between groups, while diffuse areas show customers with mixed or overlapping characteristics.
8. The distribution of points among clusters suggests that certain financial behaviours are more common than others. This can guide the bank to focus more on the dominant segment while still recognizing niche groups that may require specialized services.