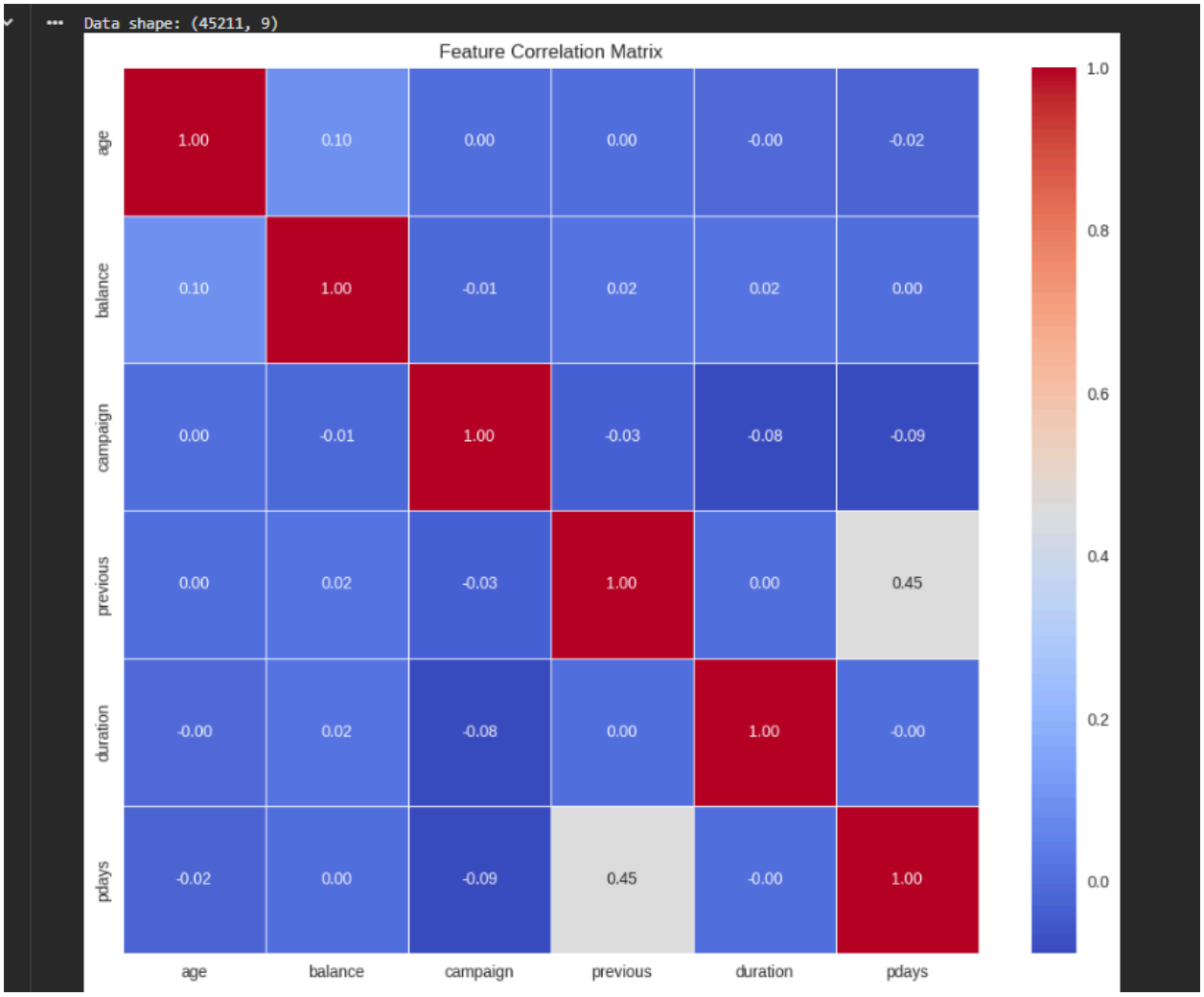
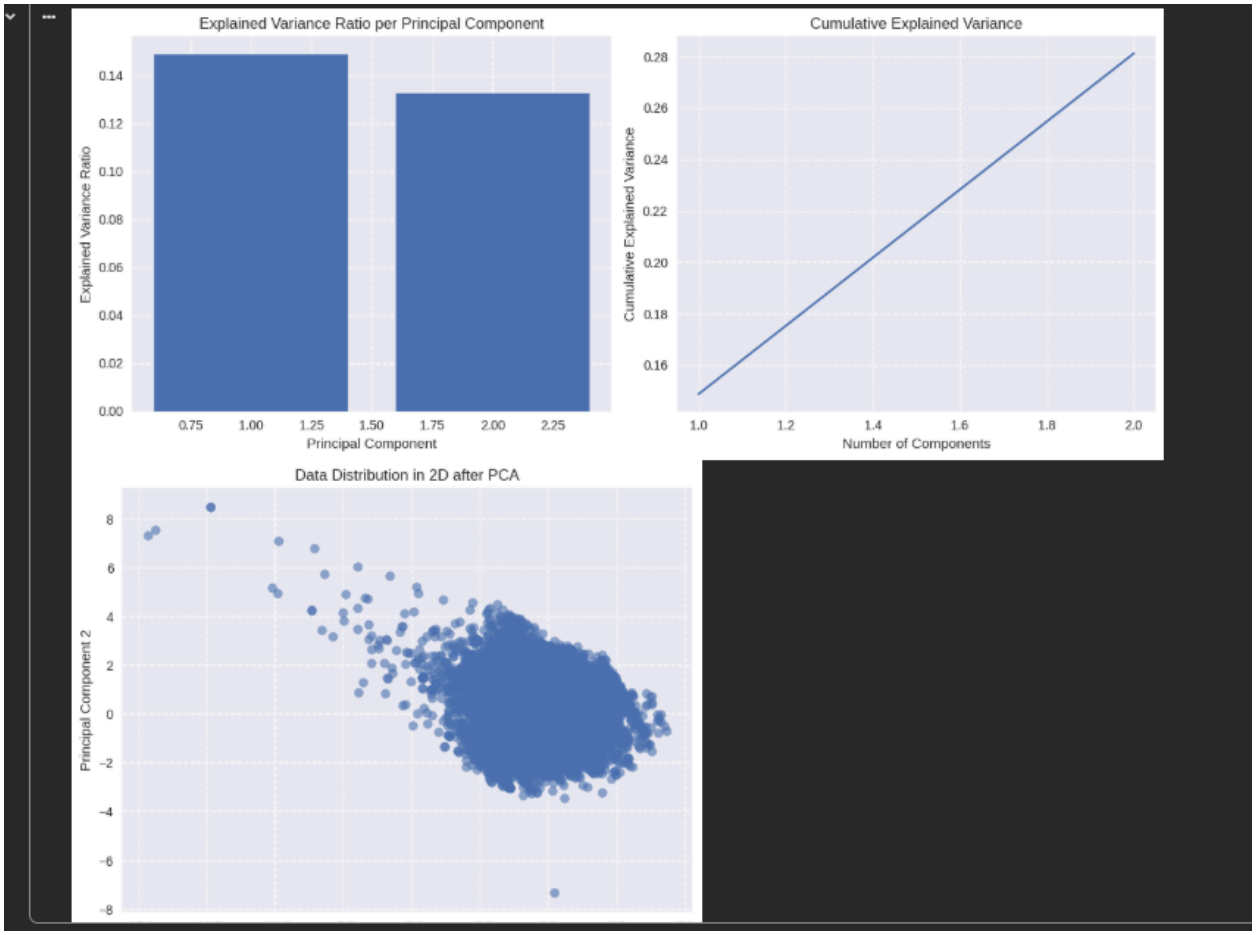
MACHINE LEARNING WEEK 13 LAB

Name: DHRUV HEGDE
Student ID: PES2UG23CS172
Course Name: MACHINE LEARNING
SECTION: C

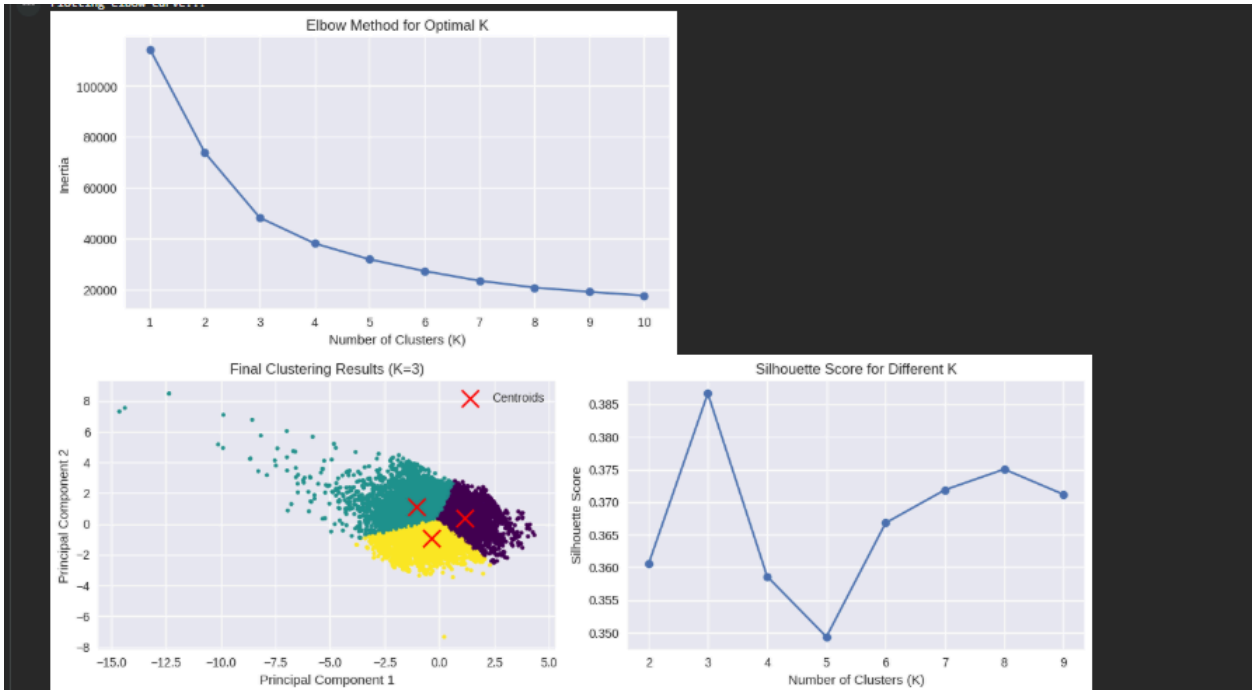
1. Feature Correlation matrix for the dataset



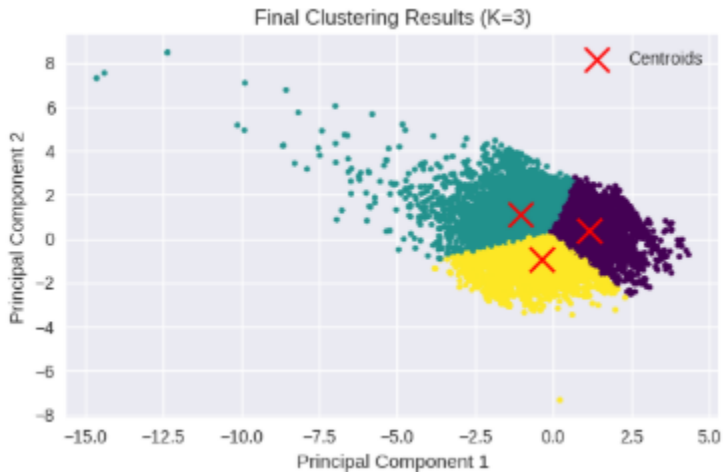
2. ‘Explained variance by Component’ and ‘Data Distribution in PCA Space’ after Dimensionality Reduction with PCA



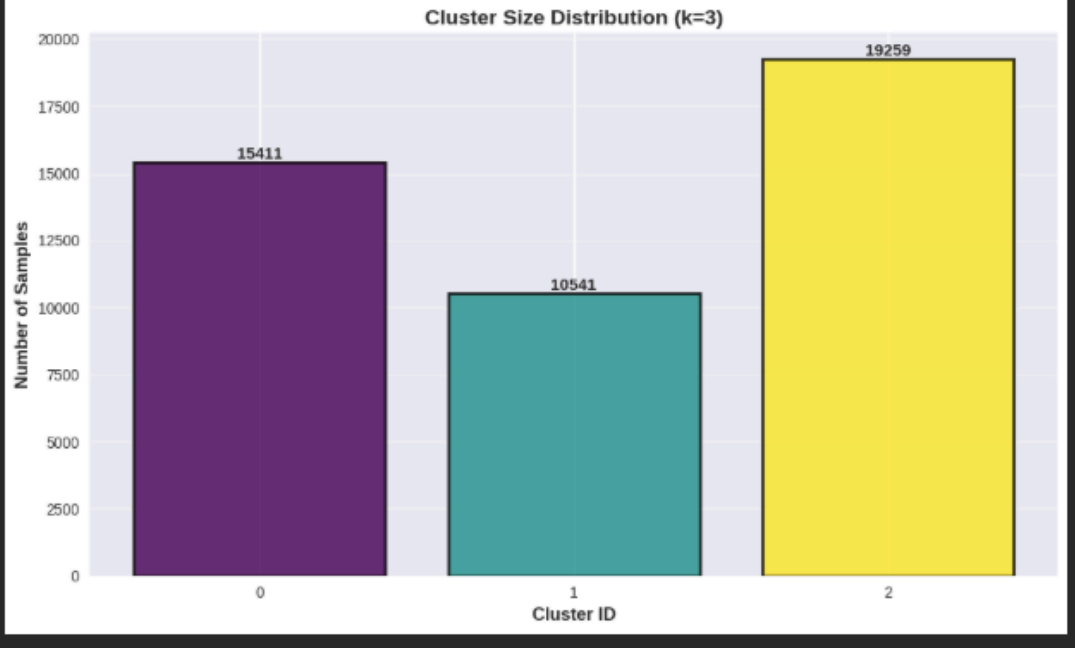
3. ‘Inertia Plot’ and ‘Silhouette Score Plot’ for K-means

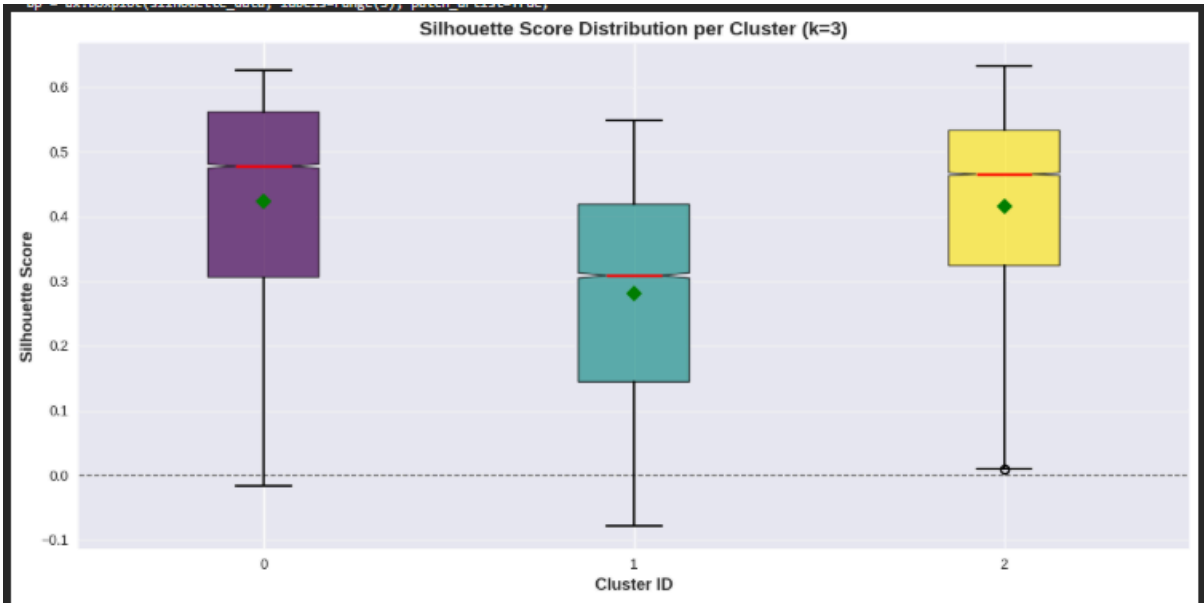


4. K-means Clustering Results with Centroids Visible (Scatter Plot) K-means Cluster Sizes (Bar Plot) Silhouette distribution per cluster for K-means (Box Plot)



```
============================================================
CLUSTER SIZE DISTRIBUTION
============================================================
```





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: Based on the correlation heatmap, some features show moderate correlation, indicating partial redundancy in the dataset. Hence, dimensionality reduction through PCA was necessary to remove multicollinearity and simplify analysis.
The first two principal components together explain 28.12% of the total variance, capturing the main data patterns while reducing noise and complexity.

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 inertia (within-cluster sum of squares) drops sharply from K = 1 to 3,and then the rate of decrease flattens noticeably after K = 3. Seeing the graph can see that the peak occurs at K = 3, where the silhouette score ≈ 0.385, the highest among all tested values.
This indicates the clusters at K = 3 are well-separated therefore 3 is the optimal number of clusters.

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 (a) Natural density differences in data- If a region of your data space has higher density , it will form a larger cluster.

K-Means produced one dominant cluster (Cluster 2) and one smaller cluster (Cluster 1), indicating one broad customer group and a few outlier or niche segments.

Bisecting K-Means split the data more hierarchically, redistributing cluster membership more evenly but still showing a dominant group.

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 (0.39) achieved a higher silhouette score than Recursive Bisecting K-Means (0.34).
This indicates that K-Means produced clusters that are slightly more cohesive and better separated for this dataset.
The dataset likely has well-separated, spherical clusters, which suits K-Means' assumptions (equal variance and compactness).
Recursive Bisecting K-Means, on the other hand, splits data hierarchically into two clusters at a time.This can sometimes over-partition the data or produce imbalanced clusters, leading to a lower silhouette score.

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: From a business perspective, this tells us that marketing or personalization strategies should prioritize the large segment for broad

campaigns, while using the smaller clusters to target niche customer needs or premium opportunities.

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 reveals three distinct customer clusters, each representing a unique behavioral profile. The turquoise cluster likely corresponds to low-engagement or cost-sensitive customers, the yellow cluster represents typical or average customers, and the purple cluster includes high-value or loyal customers. The clear separation between the yellow and purple regions suggests well-defined behavioral differences between mainstream and premium segments, while the more diffuse transition between turquoise and yellow implies gradual shifts in behavior among mid-range customers. These visual patterns highlight both clear market divisions and potential areas for cross-segment targeting or customer migration strategies.