PES2UG23CS371
Navneet Nayak
Section 'F'

Analysis Questions

Q. 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?

A: Reducing dimensionality with PCA removes multicollinearity, concentrates the signal in a few orthogonal components, and makes clustering and visualization reliable and interpretable in 2D. 28.12% of the variance is captured by the first 2 principal components.

Q. 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.

A: Both the elbow curve and silhouette score graph suggest that the optimal number of clusters for this dataset is 3.
Elbow Curve: Curve begins to flatten after k = 3
Silhouette Score: Highest score is seen for k = 3

Q. 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?

A: In both K-means and Bisecting K-means, it is noticed that one cluster is slightly larger than the others, and has a trail. Some clusters are larger than the others as they probably represent the base/most common cluster. In context of customer segments, the largest cluster can represent the average customer, while the others represent more niche segments.

Q. 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?

A: K-Means got a silhouette score of 0.39, while Bisecting K-Means got a silhouette score of 0.3602. K-Means is performing better here. It has a higher score, which means there is less overlapping between the clusters.

Q. 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?

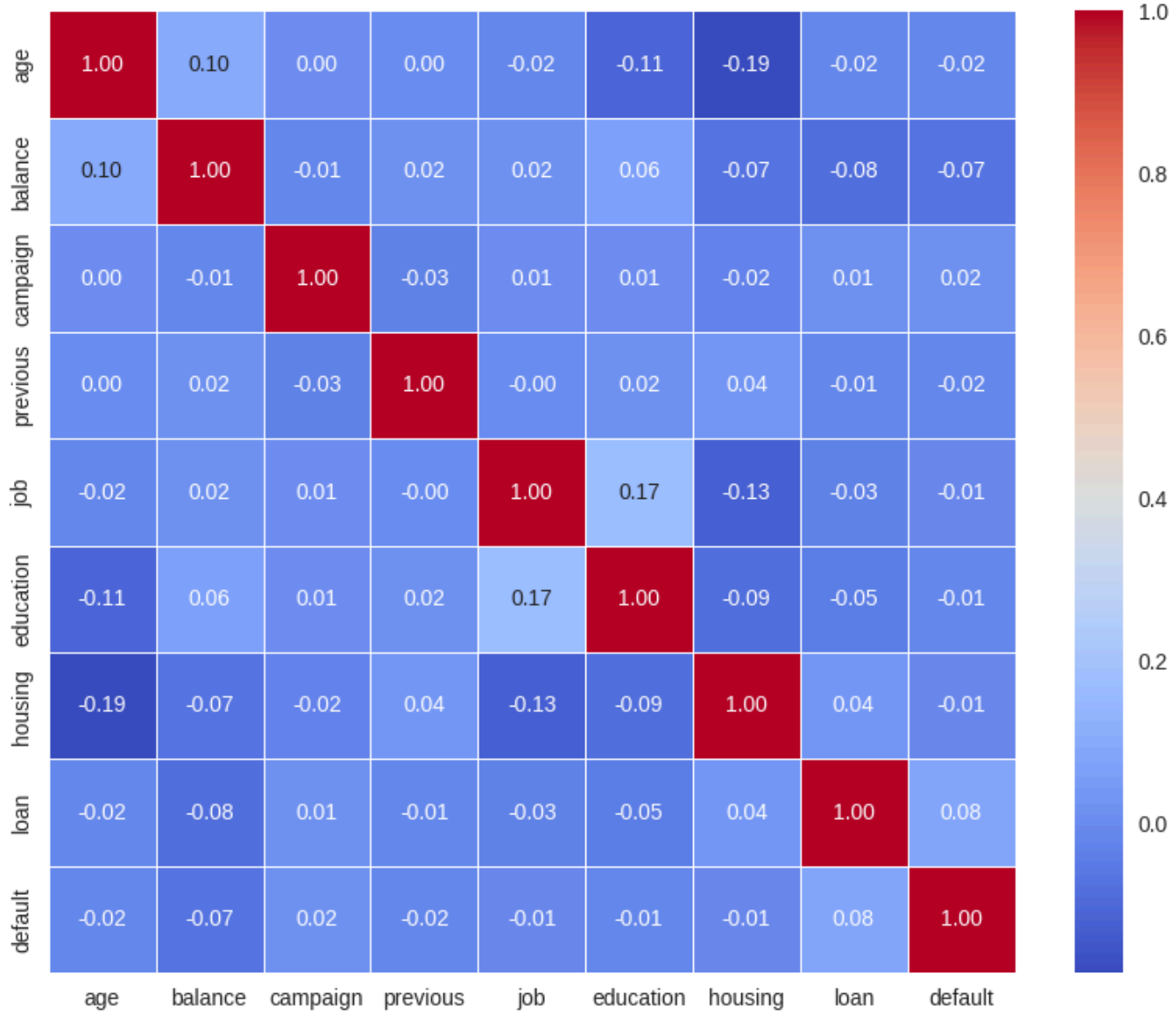
A: Based on the clustering results, the bank can take three different marketing strategies for the 3 different clusters as they represent 3 different customer segments. Customers of the same category receive the same type of marketing.

Q. 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?

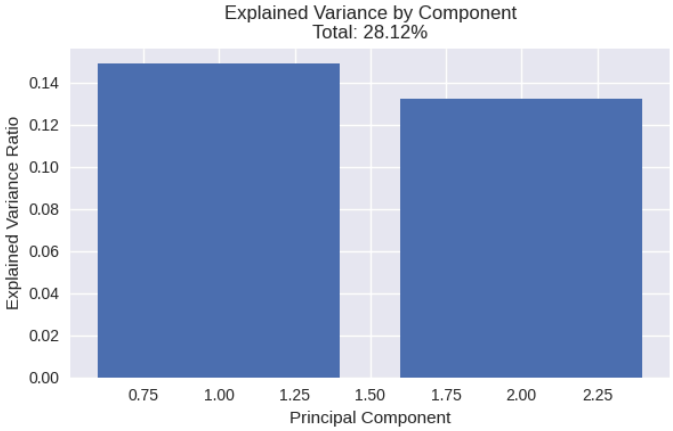
A: The 3 regions represent the 3 major customer segments, and each segment represents a certain type of consumer behavior. Sharp boundaries appear where clusters differ clearly in the principal components. Diffusing boundaries can suggest that some customers share traits of multiple segments.

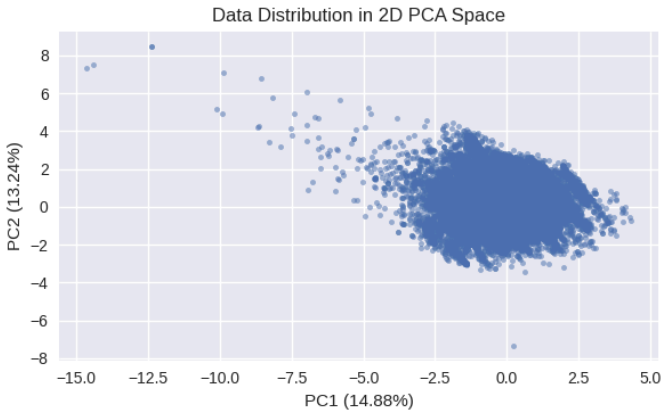
Screenshots

I. Feature Co-relation matrix:

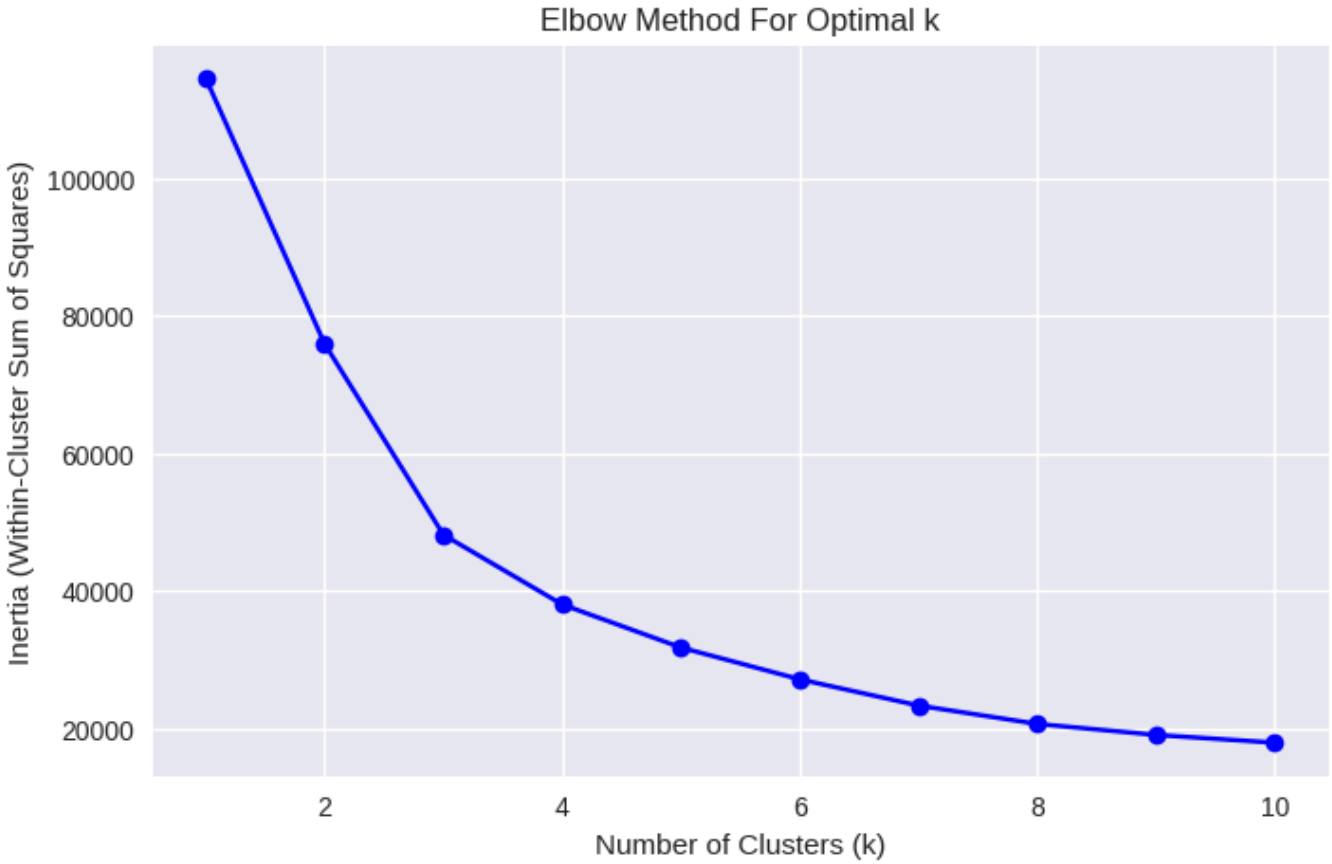


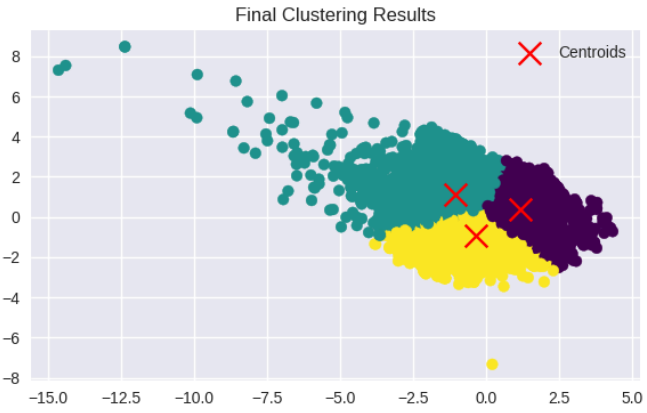
I. 'Explained variance by Component' and 'Data Distribution in PCA Space' after Dimensionality Reduction with PCA:

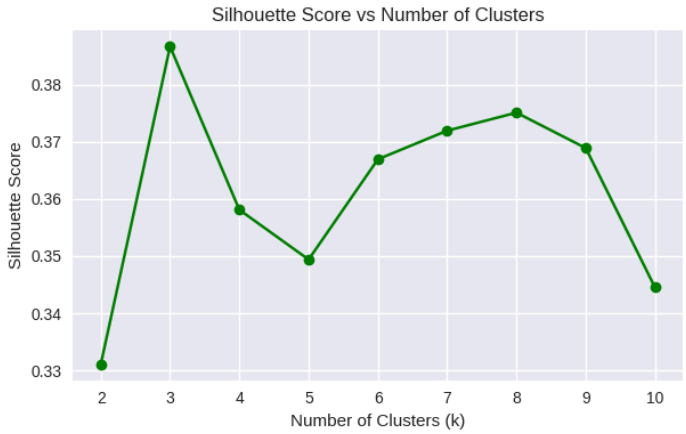




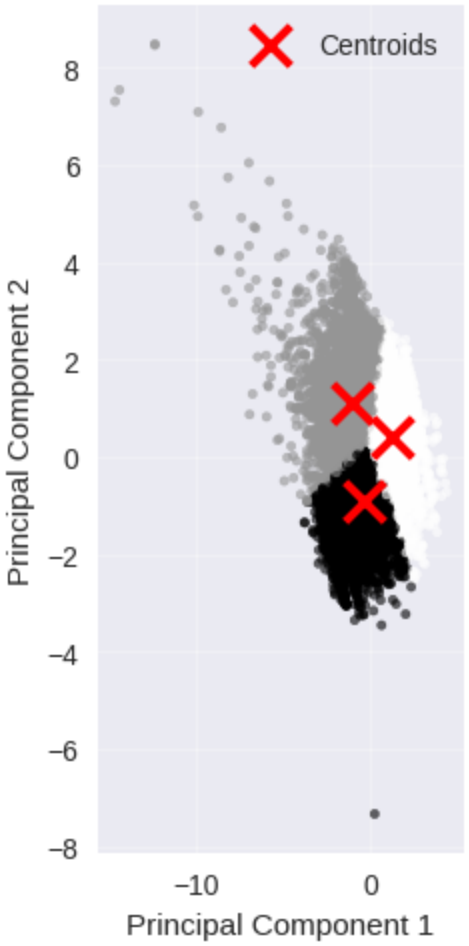
I. ‘Inertia Plot’ and ‘Silhoutte Score Plot’ for K-means:



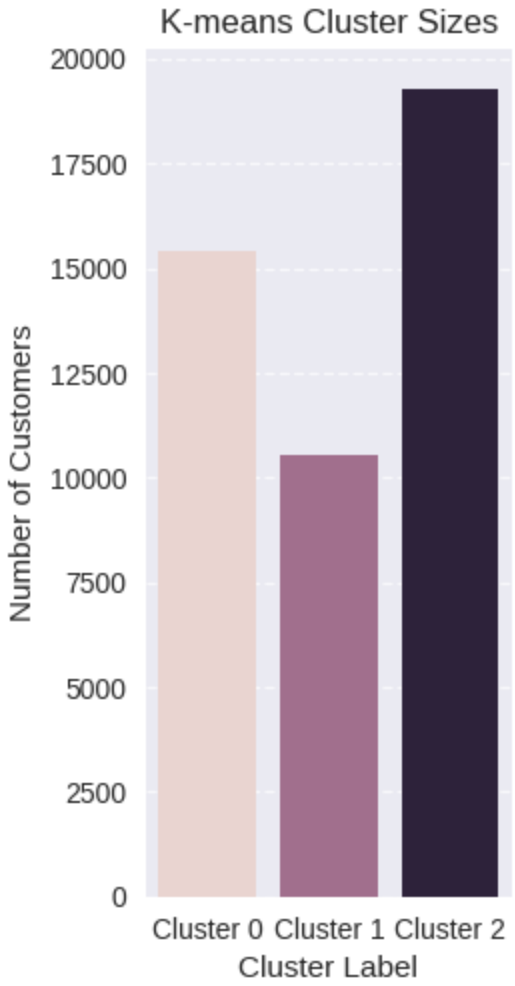




I. K-means Clustering Results with Centroids Visible (ScatterPlot):



I. K-means Cluster Sizes (Bar Plot):



I. Silhouette distribution per cluster for K-means (Box Plot):