ML Lab 13: Clustering

Name:	SRN:	Section:
Dinakar Emmanuel	PES2UG23CS178	5C

Analysis questions

Dimensionality justification

The Feature correlation matrix (screenshot 1) shows moderate correlation between the features. PCA transforms these correlated features into a new set of uncorrelated principal components, which improves the performance of clustering algorithms like K-means that are sensitive to redundant data.
The first 2 principal components capture 28% of the total variance

Optimal clusters

1. Elbow Method for Optimal K: plots number of clusters (K) against the inertia (within-cluster sum of squares). Inertia generally decreases as the number of clusters increases because the points are closer to their assigned centroids. The elbow point in this plot is the point where the rate of decrease in inertia sharply changes, which appears to be K = 3

2. Silhouette Score for Different K: The silhouette score measures how similar an object is to its own cluster compared to other clusters. A higher silhouette score indicates that the object is well-matched to its own cluster and poorly matched to neighboring clusters. Plotting the silhouette score for different K values can help identify the number of clusters that results in the highest average silhouette score, suggesting a better-defined clustering structure. There is a large spike at K = 3 which suggests that the optimal number of clusters for this problem is 3.

Cluster characteristics

Both algorithms result in imbalanced cluster sizes

1. K-means (screenshot 4):

Number of clusters	Number of data points (customers)
0	≈ 15000
1	≈ 10000
2	≈ 20000

2. Bisecting K-means (screenshot 5):

Number of clusters	Number of data points (customers)
0	≈ 20000
1	≈ 16000
2	≈ 9500

Both follow the pattern of having one large cluster with 2 smaller clusters. This could be representative of the common/mainstream customers (large cluster) and high profile/niche clients (small clusters) of the bank.

Algorithm comparison

Silhouette scores

Standard method	Recursive bisecting
0.3900	0.3369

Considering only the above scores, the K-means algorithm performed better. The scatter plot (screenshot 2) shows that the data has a globular shape. Both algorithms work well with globular shapes. Bisecting k-means performs greedy splits by splitting the larger cluster. The split may seem optimal at that moment but may or may not lead to the globally optimal solution.

Business insights

1. Large cluster: This is the bank's primary customer base. Marketing here should focus on retention and loyalty.
2. Niche segment 1 and 2:
   a. One of the groups could represent the wealthy clientele, should be targeted with investment opportunities, and premium banking perks.
   b. The other group could be targeted with loan products, such as mortgage refinancing, car loan offers, or balance transfer promotions.

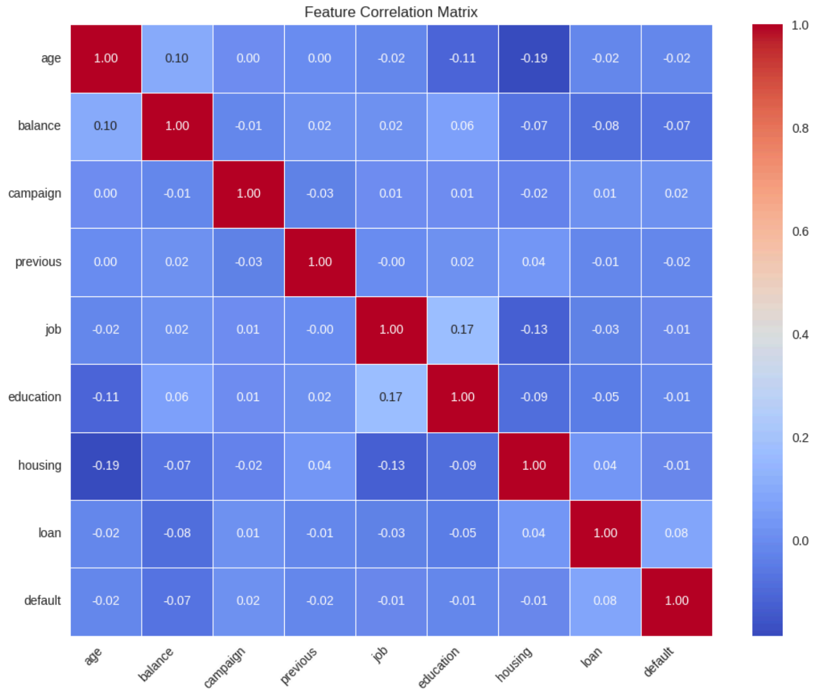
Visual pattern recognition

The large turquoise region represents the "average" customer, while the yellow and purple regions represent the two distinct niches that differ most from the average (as defined by the principal components).

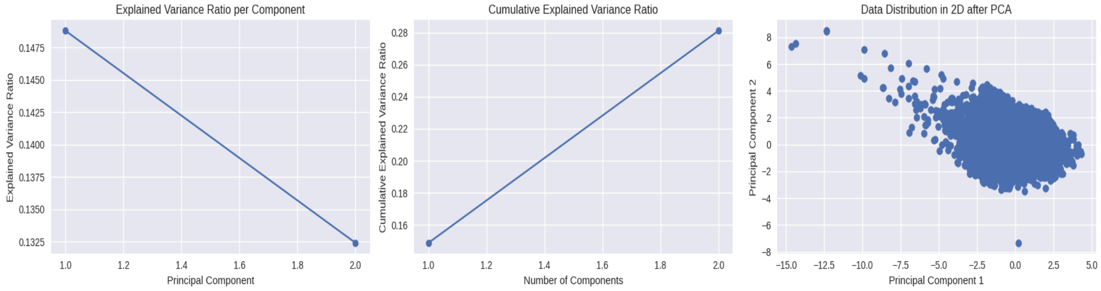
The boundaries between the clusters are visibly diffuse rather than sharp. This is because human behavior, especially in finance, exists on a continuum. Customers don't fit into perfect, discrete boxes. A diffuse boundary represents a customer who could be part of different segments.

Screenshots

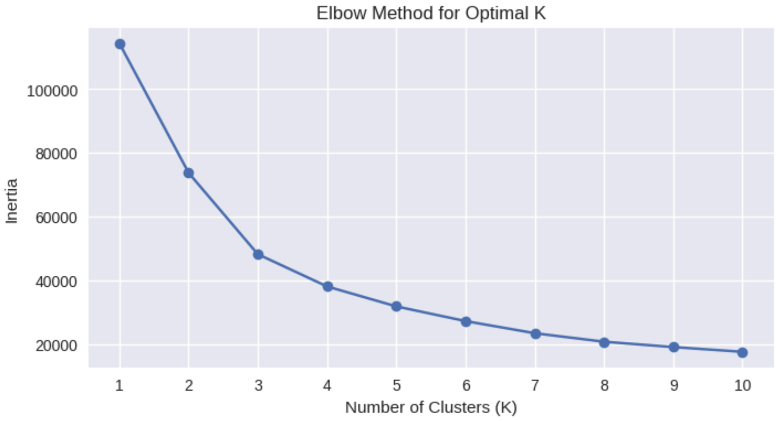
1. Feature correlation matrix

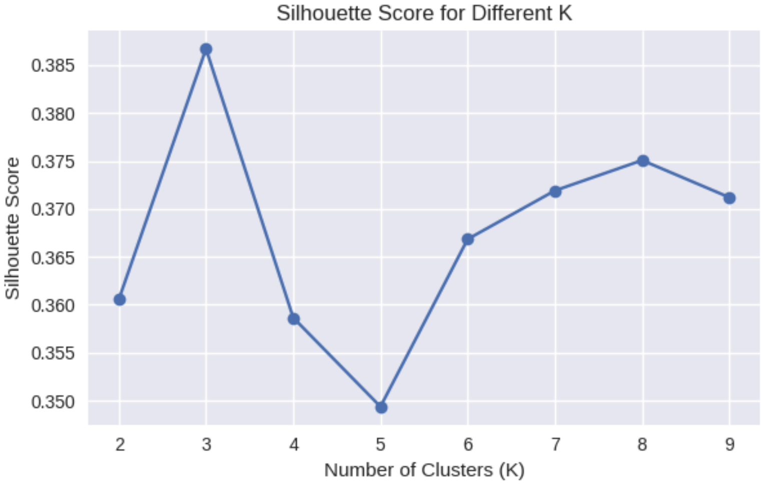


2. Explained variance (+ Cumulative) & data distribution in PCA space

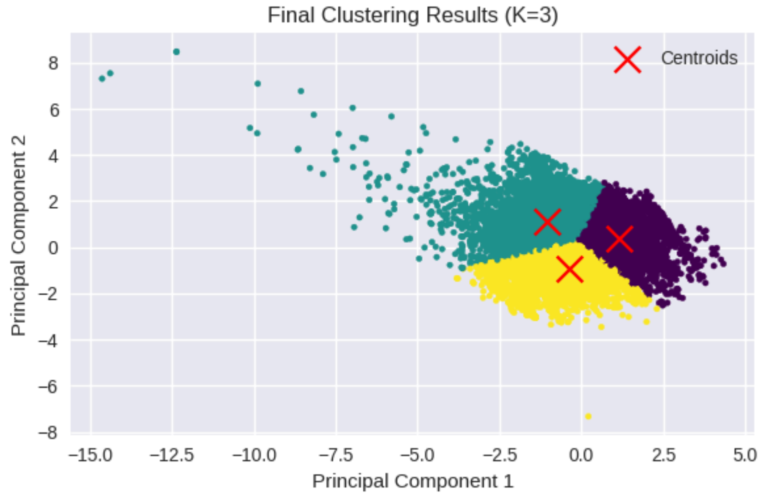


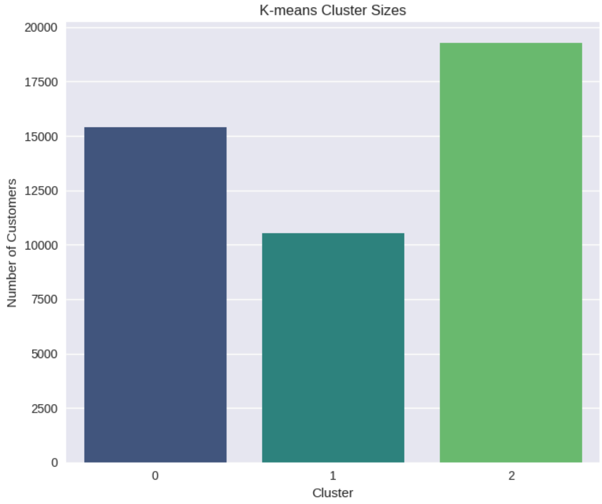
3. K-means: Inertia and silhouette scores plots

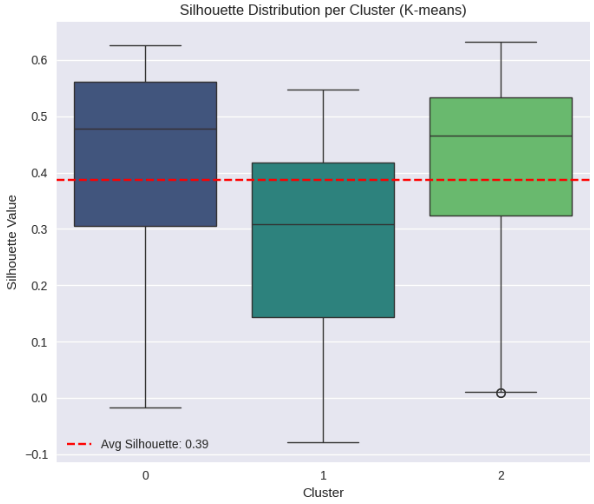




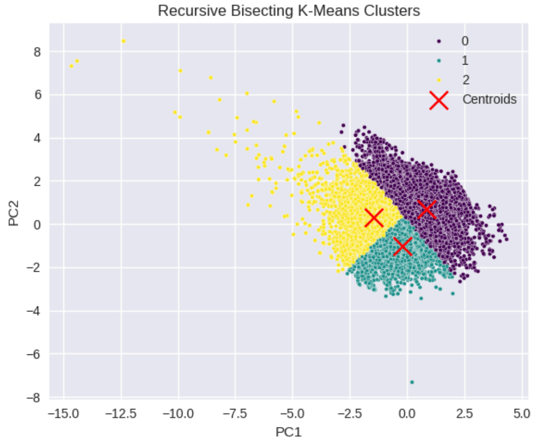
4. K-means: scatter + cluster sizes + silhouette distribution

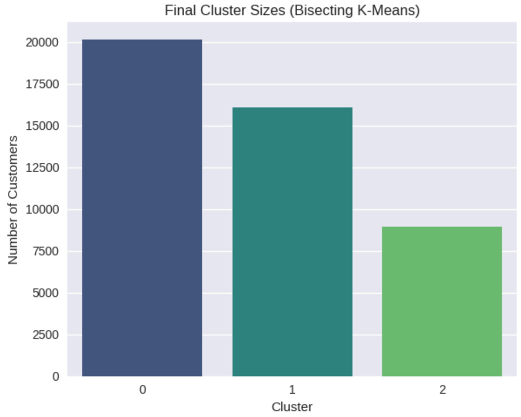

Final Clustering Results (K=3)

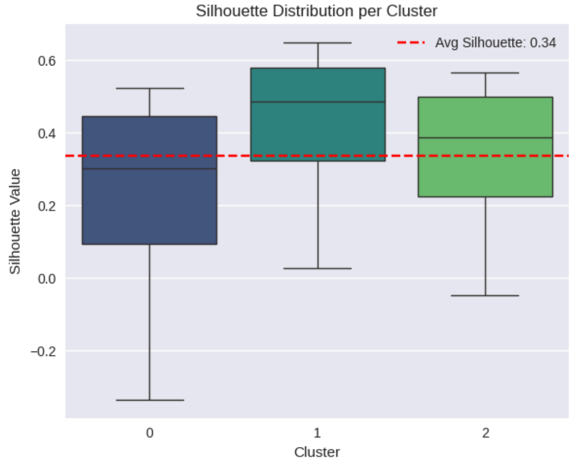

K-means Cluster Sizes


Silhouette Distribution per Cluster (K-means)

5. Bisecting K-means: scatter + cluster sizes + silhouette distribution


Recursive Bisecting K-Means Clusters
Centroids
PC2
PC1


Final Cluster Sizes (Bisecting K-Means)
Number of Customers
Cluster


Silhouette Distribution per Cluster
Avg Silhouette: 0.34
Silhouette Value
Cluster