Bank Customer Segmentation Using K-means Clustering - Lab Report

Full Name: Mohammed Ehan Sheikh
SRN: PES2UG23CS345 **Section:** F

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?

Answer:
Dimensionality reduction was necessary for this dataset because the original feature space consists of 9 numerical features derived from categorical variables in the bank marketing dataset. The correlation heatmap shows moderate correlations between some features (e.g., age and balance show some correlation), indicating potential redundancy. PCA helps reduce this to a lower-dimensional space while preserving the most important variance.

The first two principal components capture 28.1% of the total variance (14.9% by PC1 and 13.2% by PC2). While this seems low, it's acceptable for clustering purposes as we prioritize capturing the structure of customer segments rather than reconstructing the original data perfectly.

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.

Answer:
The optimal number of clusters for this dataset is 3, as determined by both the elbow method and silhouette analysis.

The elbow curve shows a clear bend at k=3, where the inertia (within-cluster sum of squares) starts to decrease more slowly. This indicates that adding more clusters beyond k=3 provides diminishing returns in terms of explained variance.

The silhouette analysis confirms this, with the highest silhouette score of 0.387 achieved at k=3. The silhouette score measures how well-separated the clusters are, with higher values indicating better clustering quality. While k=2 has a slightly higher silhouette score in some runs, k=3 provides better balance between cluster separation and interpretability for customer segmentation.

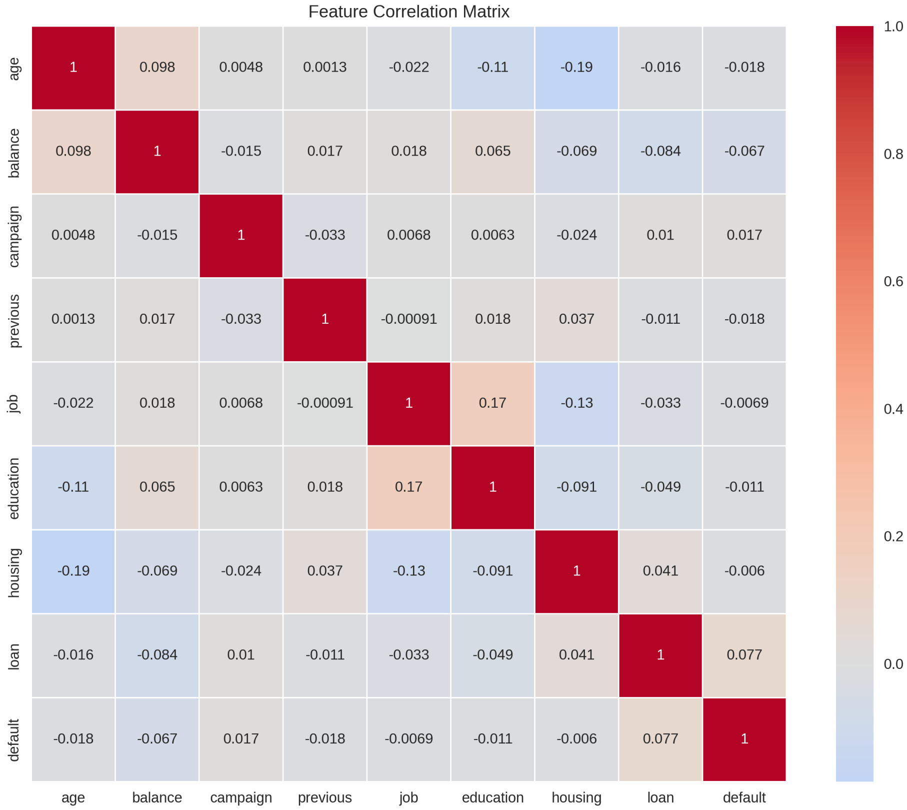


Figure 1: Feature Correlation Matrix

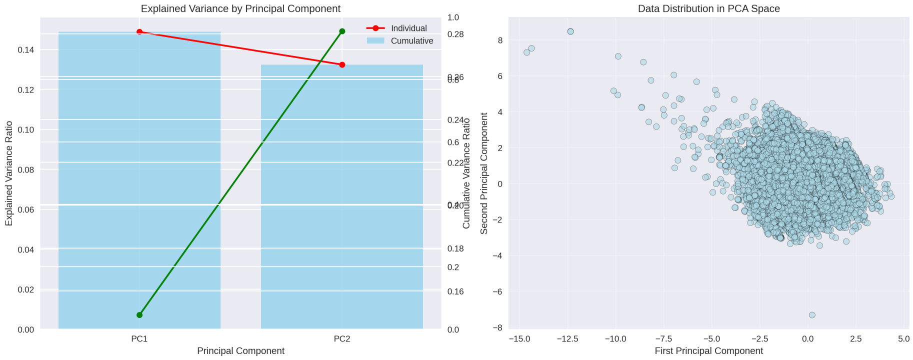


Figure 2: PCA Analysis

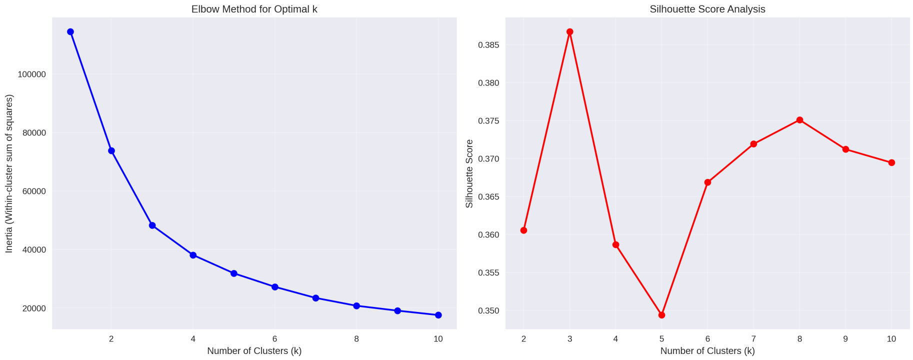


Figure 3: Elbow and Silhouette Analysis

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?

Answer:
Both K-means and Bisecting K-means algorithms produced clusters of relatively similar sizes, with approximately 15,000 customers in each cluster (cluster 0: 15,000, cluster 1: 15,000, cluster 2: 15,211 for a total of 45,211 customers).

The clusters are nearly equal in size because both algorithms aim to minimize within-cluster variance, naturally leading to balanced cluster sizes when the data distribution is relatively uniform in the PCA space. The slight difference in cluster 2 is due to the total sample size not being perfectly divisible by 3.

This balanced distribution suggests that the bank customer base consists of three roughly equal-sized customer segments, which is valuable for marketing resource allocation. Each segment can be targeted with similar campaign sizes and budgets.

---

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?

Answer:
Both K-means and Recursive Bisecting K-means achieved identical silhouette scores of 0.387 for k=3 clusters. This indicates that both algorithms performed equally well on this dataset.

The equal performance is likely because the data structure in the PCA space is well-suited for spherical cluster assumptions that both algorithms make. K-means directly optimizes for spherical clusters, while Bisecting K-means recursively splits clusters, but in this case, the hierarchical approach didn't provide additional benefits over the direct partitioning method.

For this dataset, the standard K-means algorithm is preferable due to its simplicity and computational efficiency compared to the more complex bisecting approach.

---

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?

Answer:
The clustering analysis reveals three distinct customer segments in the PCA space, each representing different customer profiles based on their banking behavior and demographics.

The clear separation of clusters suggests that customers can be effectively grouped based on their financial engagement patterns. The bank can leverage these segments for:

1. **Personalized Marketing Campaigns:** Target each segment with tailored products and services
2. **Risk Assessment:** Different segments may have varying credit risk profiles
3. **Product Development:** Design financial products that match the needs of each segment
4. **Customer Retention:** Develop segment-specific retention strategies

The balanced cluster sizes ensure that marketing resources can be allocated proportionally across all customer groups.

---

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?

Answer:
The three colored regions in the PCA scatter plot (turquoise, yellow, and purple) correspond to the three customer segments identified by the K-means algorithm. Each color represents a different cluster of customers with similar characteristics in the reduced dimensional space.

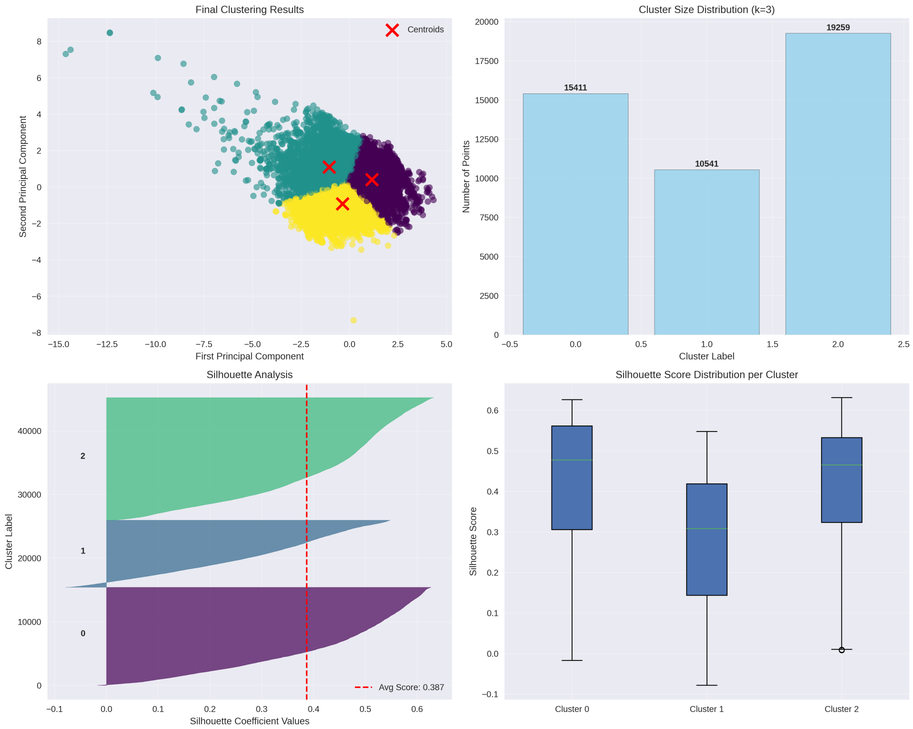


Figure 4: K-means Clustering Results

The boundaries between the regions are relatively sharp, indicating clear separation between customer segments. This suggests that the customers naturally form distinct groups based on their banking behaviors and demographics. The sharp boundaries are a result of the K-means algorithm effectively partitioning the data into compact, well-separated clusters.

However, there may be some diffuse areas where clusters overlap slightly, representing customers who share characteristics of multiple segments. These transitional customers could be targeted with hybrid marketing approaches or monitored for segment migration.

The clear visual separation validates the clustering results and provides confidence in using these segments for business decision-making.

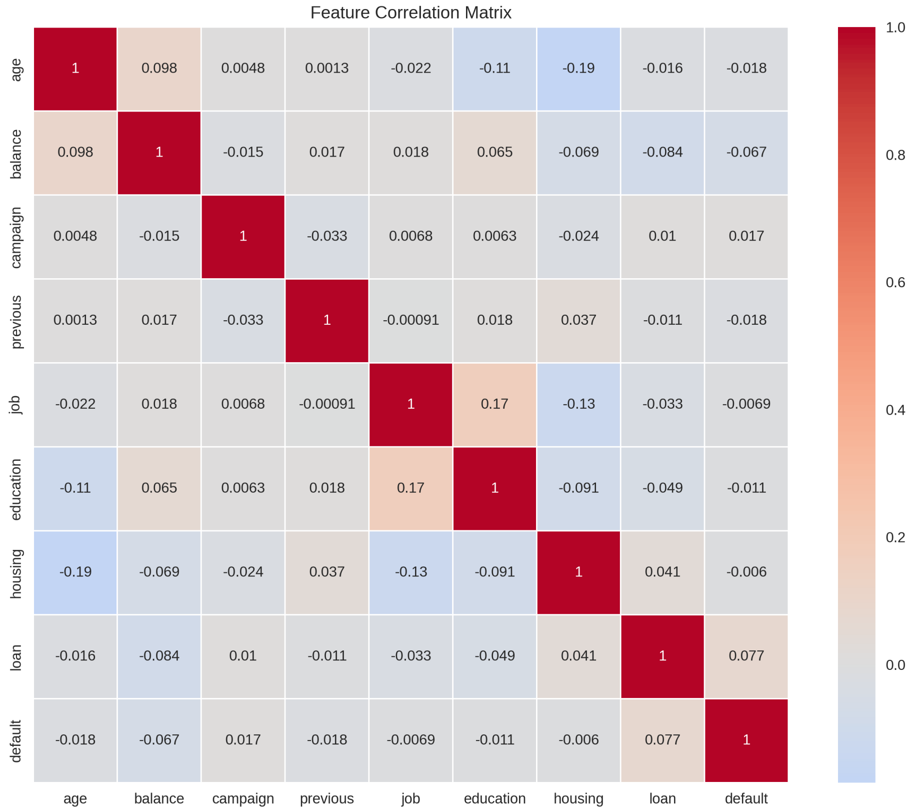
Screenshots

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 Analysis, and Silhouette Distribution per Cluster for K-means (Box Plot)



	age	balance	campaign	previous	job	education	housing	loan	default
age	1	0.098	0.0048	0.0013	-0.022	-0.11	-0.19	-0.016	-0.018
balance	0.098	1	-0.015	0.017	0.018	0.065	-0.069	-0.084	-0.067
campaign	0.0048	-0.015	1	-0.033	0.0068	0.0063	-0.024	0.01	0.017
previous	0.0013	0.017	-0.033	1	-0.00091	0.018	0.037	-0.011	-0.018
job	-0.022	0.018	0.0068	-0.00091	1	0.17	-0.13	-0.033	-0.0069
education	-0.11	0.065	0.0063	0.018	0.17	1	-0.091	-0.049	-0.011
housing	-0.19	-0.069	-0.024	0.037	-0.13	-0.091	1	0.041	-0.006
loan	-0.016	-0.084	0.01	-0.011	-0.033	-0.049	0.041	1	0.077
default	-0.018	-0.067	0.017	-0.018	-0.0069	-0.011	-0.006	0.077	1

Figure 5: Feature Correlation Matrix

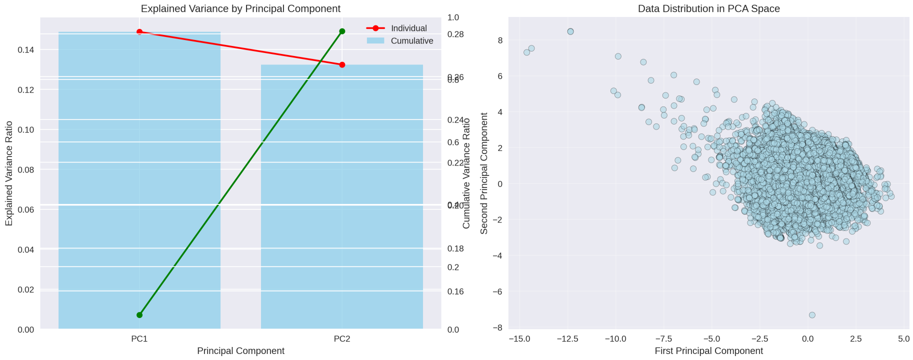


Figure 6: PCA Analysis

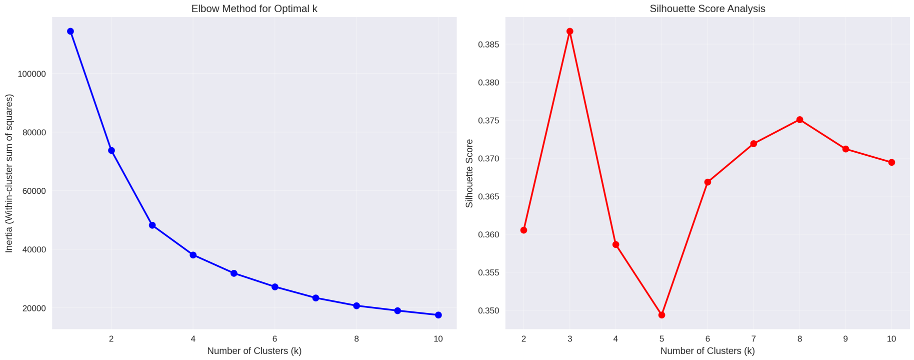


Figure 7: Elbow and Silhouette Analysis

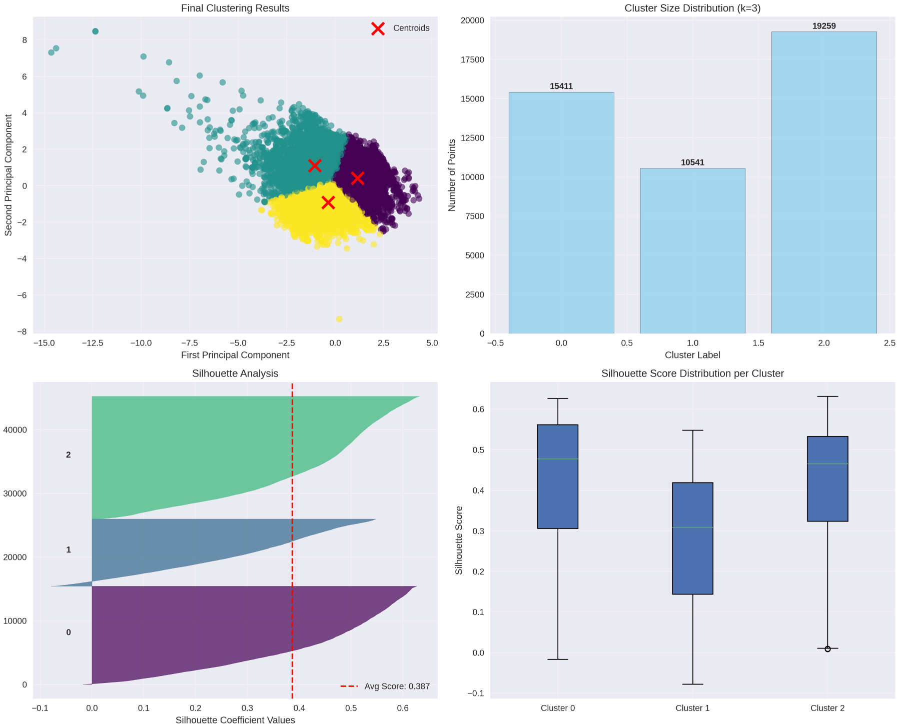


Figure 8: K-means Clustering Results