ML LAB 13- CLUSTERING

NAME- NISSCHAY KHANDELWAL

SRN-PES2UG23CS394

SECTION-F

ANALYSIS FROM QUESTIONS

Question 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 because the correlation matrix shows weak correlations between most features (values close to 0), indicating that the features capture different aspects of customer behavior. The strongest correlations are between job-education (0.17) and loan-default (0.08), which are relatively weak. With 9 features, visualizing and computing clusters becomes computationally expensive. PCA reduced the data to 2 dimensions while retaining 28.12% of the total variance (PC1: 14.88%, PC2: 13.24%). Although this captures only about one-quarter of the variance, it provides sufficient information for meaningful clustering and enables 2D visualization, making the analysis more interpretable.

Question 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 is k=3. The elbow curve shows a clear "elbow" at k=3, where the inertia drops sharply from 75,000 to 48,000, then decreases more gradually afterward. The silhouette score analysis confirms this choice, showing the highest score of 0.3867 at k=3, compared to 0.3311 for k=2 and 0.3581 for k=4. While k=8 shows a slightly higher silhouette score (0.3751), the marginal improvement doesn't justify the added complexity. Both metrics converge on k=3 as the optimal balance between cluster cohesion, separation, and model simplicity.

Question 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:

K-means (k=3) produced unbalanced clusters: Cluster 2 (19,259 customers), Cluster 0 (15,411 customers), and Cluster 1 (10,541 customers). This suggests that bank customers naturally form different-sized segments. The largest cluster likely represents "standard" or "typical" customers with common banking behaviors. The medium-sized cluster may represent moderately active or mid-tier customers, while the smallest cluster could represent either high-value customers or high-risk customers who are less common in the population. Bisecting K-means (k=4) produced more balanced clusters (16,348, 12,795, 8,429, and 7,639), suggesting that hierarchical splitting creates more evenly distributed segments. The unbalanced distribution in K-means reflects real-world customer heterogeneity, where certain customer profiles are more prevalent than others.

Question 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:

K-means (k=3) achieved a silhouette score of 0.3867, while Bisecting K-means (k=4) achieved 0.3602. K-means performed better for this dataset. This is because K-means with k=3 naturally aligns with the inherent structure of the data, as evidenced by the elbow curve and silhouette analysis. Bisecting K-means forces a hierarchical splitting structure that may not match the natural cluster distribution. With k=4, Bisecting K-means over-segments the data, creating clusters that are less cohesive and well-separated. K-means benefits from simultaneous optimization of all cluster assignments, while Bisecting K-means makes irreversible binary splits that may not be globally optimal. For this bank dataset, the three-cluster structure identified by K-means provides the best balance of simplicity and cluster quality.

Question 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 three distinct customer segments suggest differentiated marketing strategies. Cluster 2 (largest, 42.6% of customers) likely represents the "mass market" segment with standard banking needs - these customers could be targeted with automated, cost-effective digital marketing campaigns. Cluster 0 (34.1% of customers) represents a significant "core customer" segment that may have moderate account balances and engagement - personalized offers for savings products or credit cards could be effective. Cluster 1 (smallest, 23.3% of customers) is the most interesting for targeted marketing - these customers may represent either high-value clients worth premium service investment, or high-risk customers requiring careful credit management. The clear separation in PCA space indicates these segments have genuinely different characteristics, making segment-specific product offerings and communication strategies worthwhile investments.

Question 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?

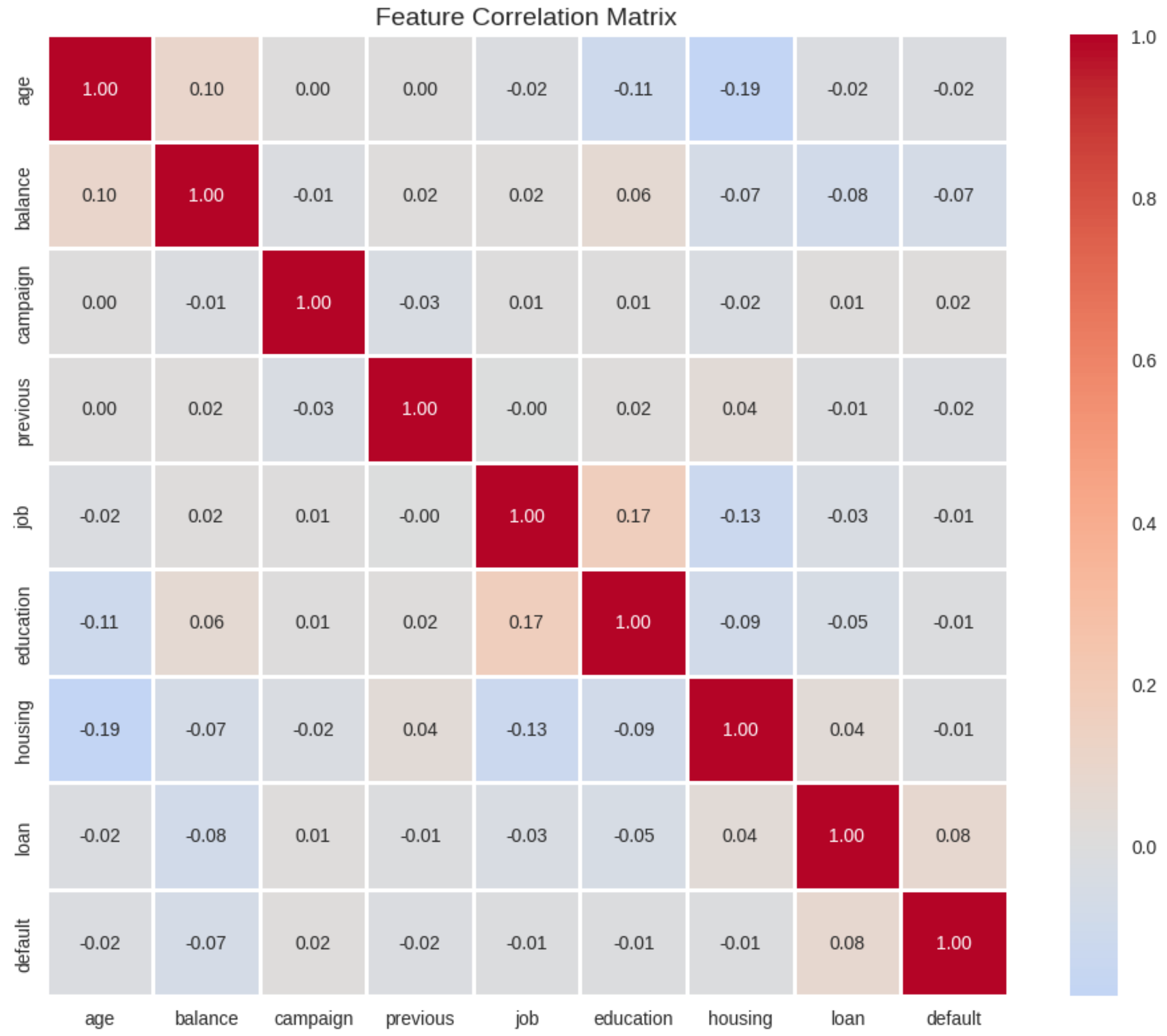
Answer:

The three colored regions in the PCA scatter plot represent different customer profiles along the principal components. PC1 (14.88% variance) likely captures primary behavioral differences such as account balance or transaction frequency, while PC2 (13.24% variance) captures secondary characteristics like loan status or campaign responsiveness. The boundaries between clusters show mixed characteristics: some regions have sharp boundaries indicating clear distinctions between customer types (e.g., customers with/without loans or housing), while other regions show diffuse boundaries suggesting gradual transitions between segments. The diffuse boundaries are natural because customer behavior exists on a continuum - many customers share characteristics across segments. The overlap regions represent customers who are transitional or hybrid between segments, perhaps moderate users who could shift segments with targeted marketing. This visual pattern confirms that while three clusters provide useful segmentation, real customer behavior is more nuanced than strict categorical divisions.

==

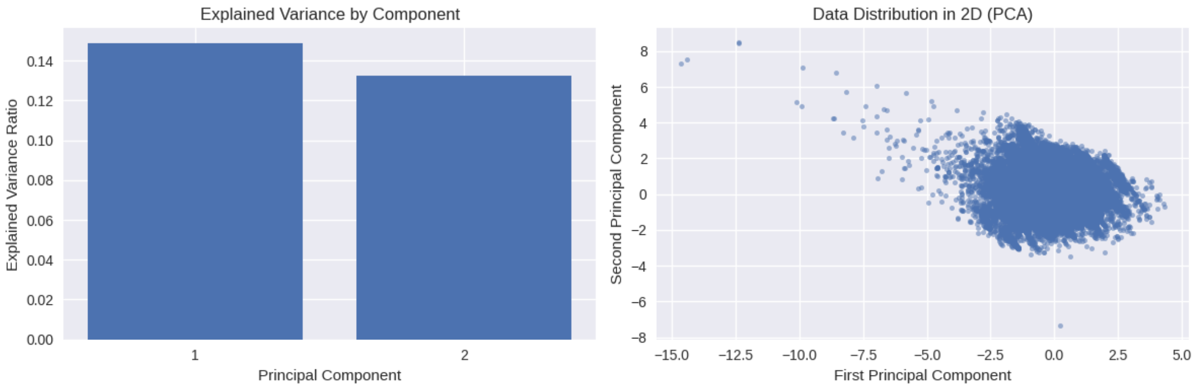
SCREENSHOTS

1. Feature Correaltion matrix for the dataset

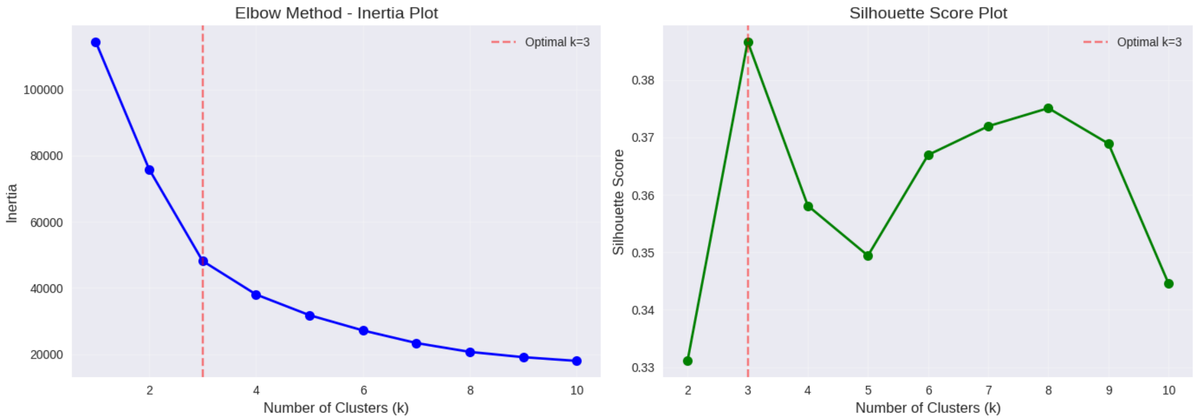


	age	balance	campaign	previous	job	education	housing	loan	default
age	1.00	0.10	0.00	0.00	-0.02	-0.11	-0.19	-0.02	-0.02
balance	0.10	1.00	-0.01	0.02	0.02	0.06	-0.07	-0.08	-0.07
campaign	0.00	-0.01	1.00	-0.03	0.01	0.01	-0.02	0.01	0.02
previous	0.00	0.02	-0.03	1.00	-0.00	0.02	0.04	-0.01	-0.02
job	-0.02	0.02	0.01	-0.00	1.00	0.17	-0.13	-0.03	-0.01
education	-0.11	0.06	0.01	0.02	0.17	1.00	-0.09	-0.05	-0.01
housing	-0.19	-0.07	-0.02	0.04	-0.13	-0.09	1.00	0.04	-0.01
loan	-0.02	-0.08	0.01	-0.01	-0.03	-0.05	0.04	1.00	0.08
default	-0.02	-0.07	0.02	-0.02	-0.01	-0.01	-0.01	0.08	1.00

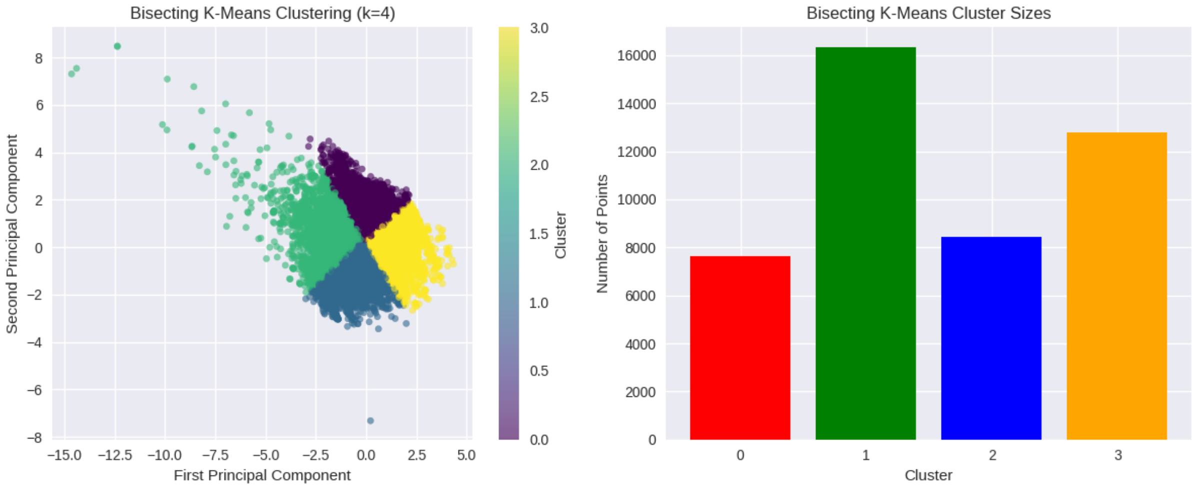
2. ‘Explained variance by Component’ and ‘Data Distribution in PCA Space’ after Dimensionality Reduction with PCA



3. ‘Inertia Plot’ and ‘Silhoutte 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)


Bisecting K-Means Clustering (k=4)
Second Principal Component
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
First Principal Component
Cluster
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Bisecting K-Means Cluster Sizes
Number of Points
16000
14000
12000
10000
8000
6000
4000
2000
0
0
1
2
3
Cluster