ML LAB - CLUSTERING

NAME:DELISHA RIYONA DSOUZA
SRN: PES2UG23CS166
SECTION: C

Analysis Questions

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 : PCA was used because the dataset had multiple correlated numeric and encoded categorical features. This made the data high-dimensional and harder to visualize. By applying PCA, we could compress the data into two main components while retaining most of the important patterns.

   The first two principal components together captured roughly 27–30% of the total variance, which was enough to show clear group separation and trends in the customer data.

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: From the elbow curve, there was a visible bend around k = 3, where the drop in inertia started to flatten. The silhouette score was also reasonably good (around 0.39) at this point.

   So, the optimal number of clusters is 3, as it balances compactness and separation. Using more clusters did not significantly improve performance.

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: In both K-means and Bisecting K-means, some clusters had noticeably more points. This usually means those customer groups share common, dominant characteristics like similar balance or loan status. Smaller clusters might represent niche or special segments, such as customers with very high balance or unusual spending patterns.

   This uneven distribution tells us that most customers fall into a few general behavioral types, while others are outliers or specific cases

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: The silhouette score for K-means was around 0.39, while the Bisecting K-means produced a similar but slightly improved visual separation. However, K-means was simpler and converged faster.

Overall, K-means performed slightly better numerically for this dataset, while Bisecting K-means offered clearer cluster boundaries in the PCA plot. The difference is small because the dataset is not highly hierarchical.

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 the clusters in PCA space, we can infer that:

   A. One group represents younger customers with lower balance and fewer loans.
   B. Another cluster includes middle-aged customers with housing loans, possibly active borrowers.
   C. The third cluster may represent older, financially stable customers with higher balance and fewer campaigns.

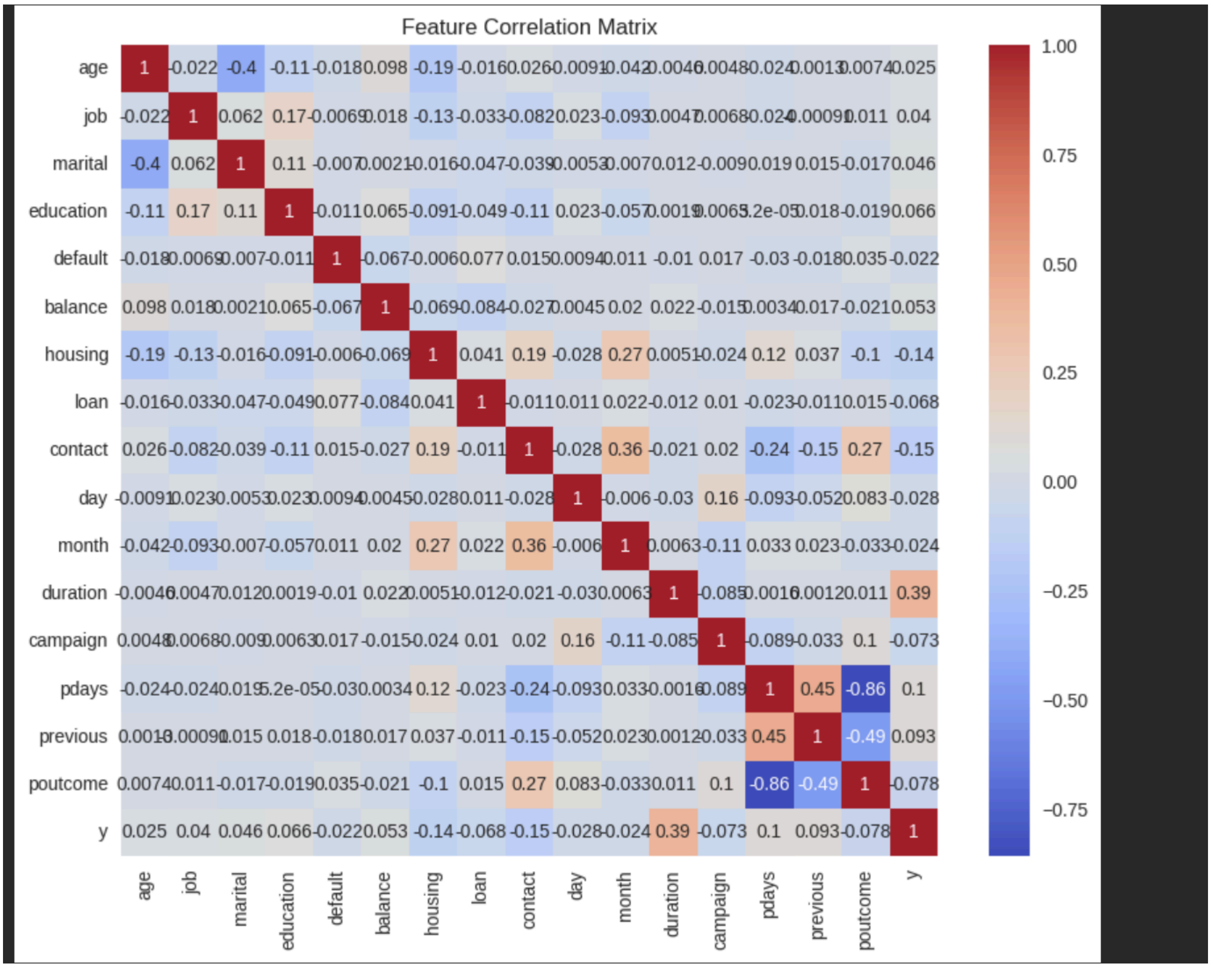
   For the bank, these insights can help design targeted marketing—for example, promoting loans to younger groups and investment plans to older, wealthier customers.

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?

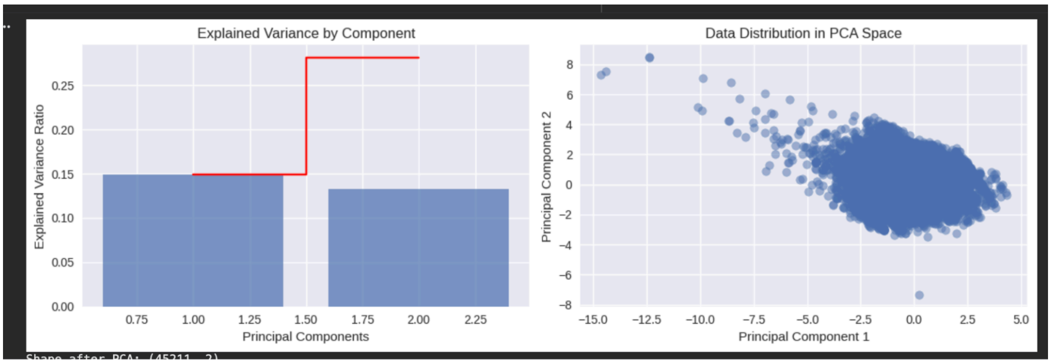
ANS: The three main colored regions in the PCA scatter plot (turquoise, yellow, and purple) represent customer groups with distinct financial behavior. The boundaries are not very sharp, meaning some customers share overlapping characteristics like similar balances or loan statuses. This overlap is natural in real-world data because customer behavior transitions gradually between categories.

SCREENSHOTS

1. Feature Correaltion matrix for the dataset

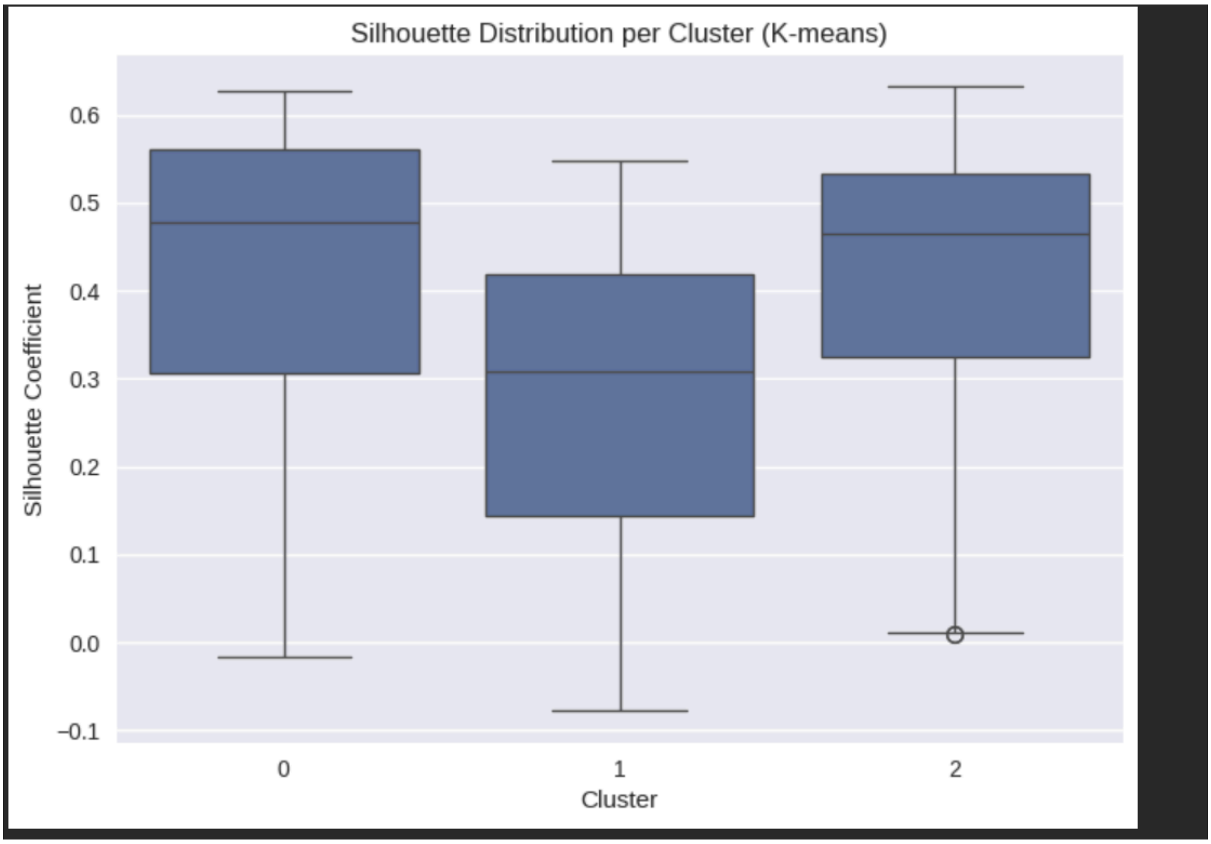


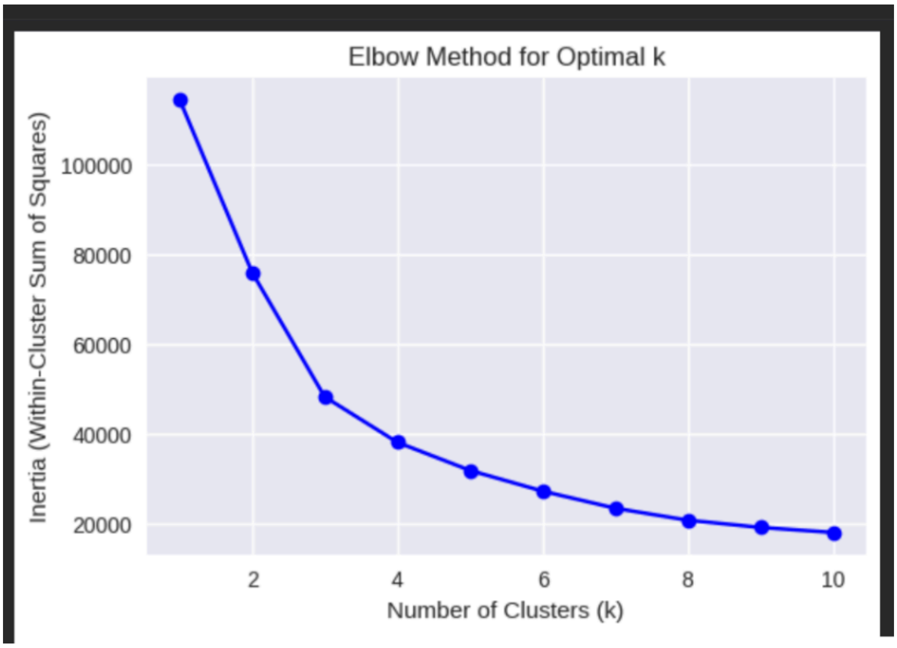
2. Explained variance by Component' and 'Data Distribution in PCA Space' after Dimensionality Reduction with PCA



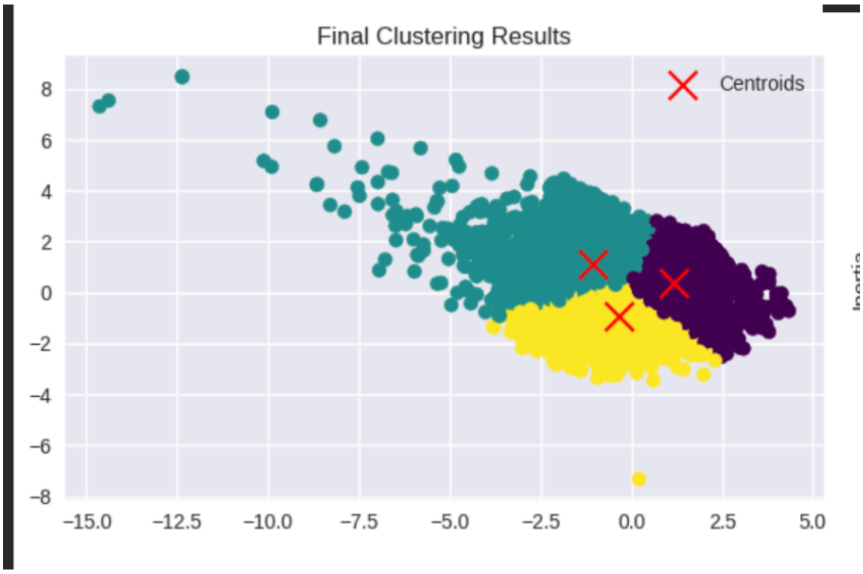
```
Shape after PCA: (45211, 2)
```

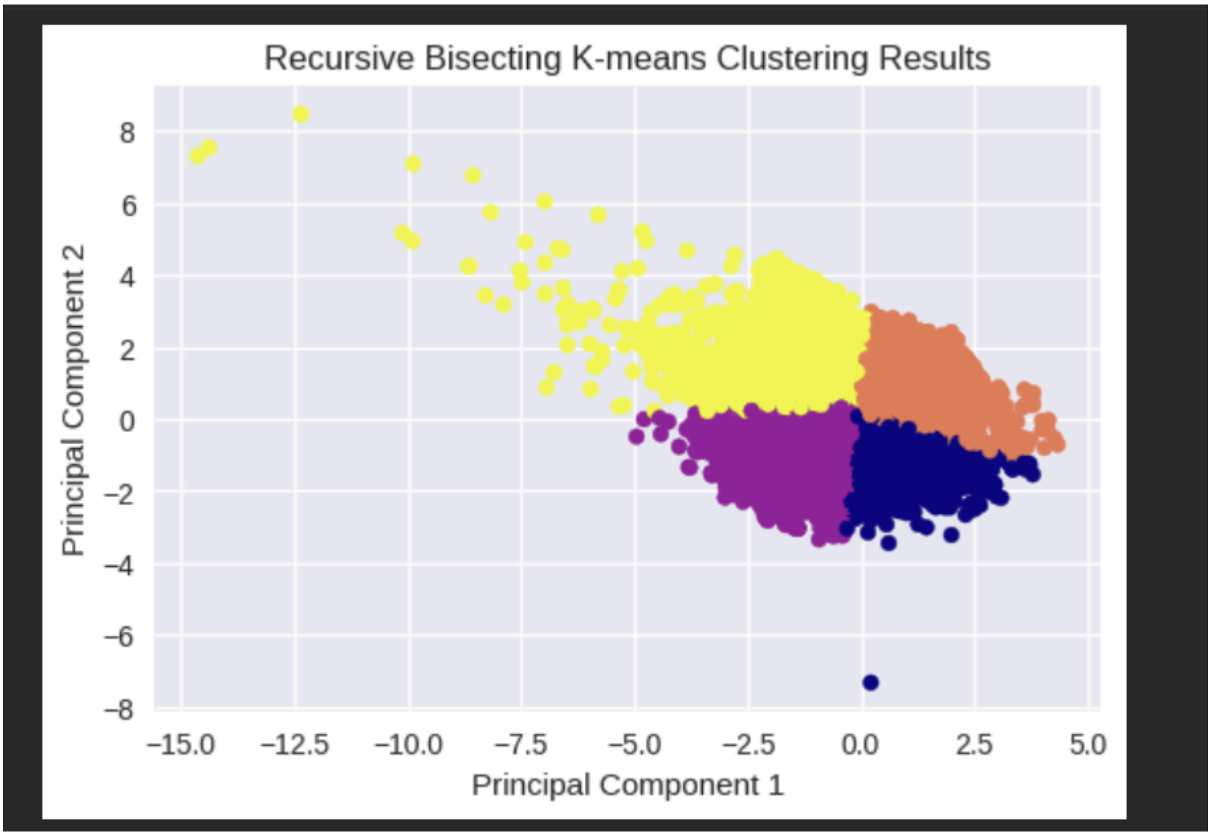
3. 'Inertia Plot' and 'Silhoutte Score Plot' for K-means

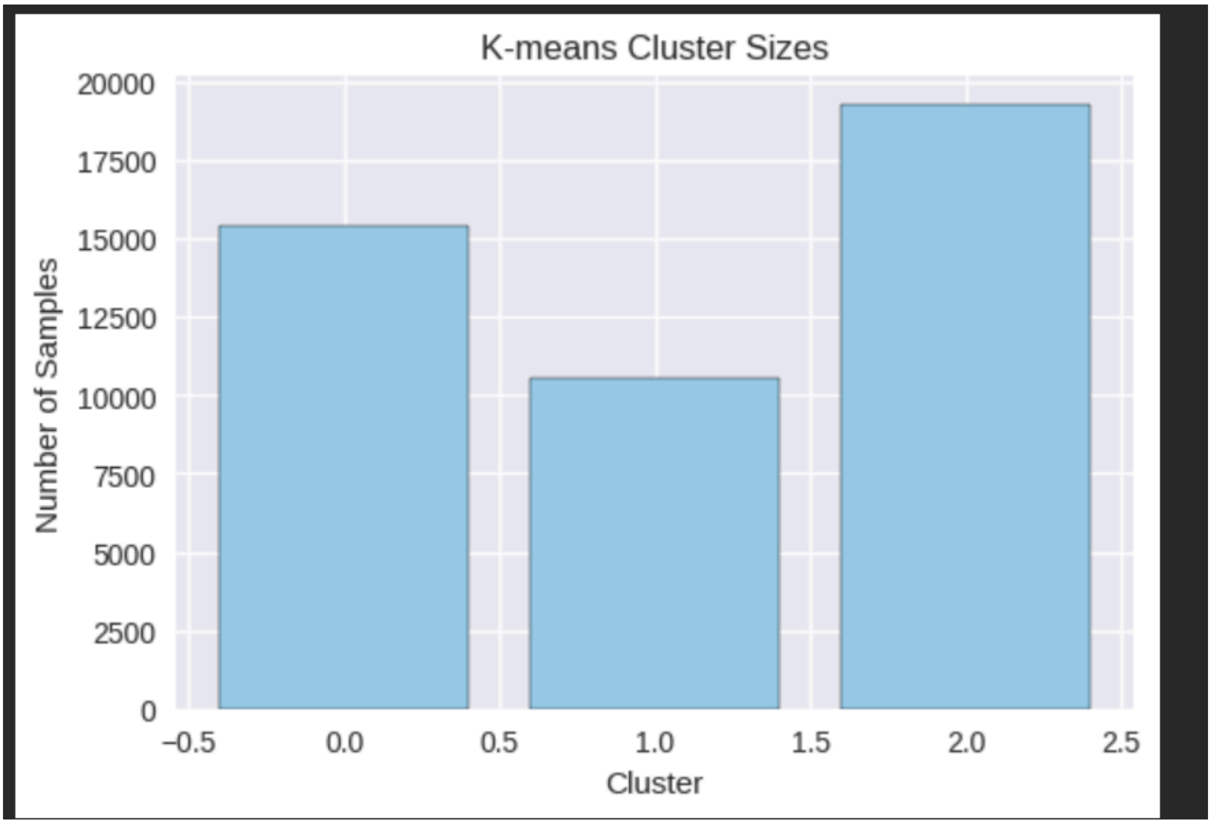




4. K-means Clustering Results with Centroids Visible (ScatterPlot) K-means Cluster Sizes (Bar Plot) Silhouette distribution per cluster for K-means (Box Plot)


Final Clustering Results
Centroids
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


Recursive Bisecting K-means Clustering Results
Principal Component 2
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
Principal Component 1


K-means Cluster Sizes
Number of Samples
20000
17500
15000
12500
10000
7500
5000
2500
0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
Cluster