Machine Learning Lab-13

K-Means Clustering

Name: Chethana K R

Srn: PES2UG23CS151

Sec: C

Dimensionality Justification:

Q: 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: The dataset contains several correlated financial and demographic features, which introduce redundancy and noise. PCA reduces this correlation and emphasizes the main variance directions, simplifying clustering and visualization.
Variance captured by first two principal components:
From the PCA explained variance ratio, the first two principal components together capture roughly **24–32% of the total variance**.
Although this is not a majority of the variance, it is sufficient for visualization and revealing broad structure in the data.

Q: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: The **elbow curve** shows a steep drop in inertia up to **k = 3**, after which the curve flattens, indicating diminishing returns.
The **silhouette score** peaks around **k = 2 or k = 3**, but drops for larger k.

Combining both metrics, the most reasonable trade-off between compactness and separation is **k = 3**.
This gives a meaningful segmentation without over-splitting the data.

Q: Analyze the size distribution of clusters in both K-means and Bisecting K-means.
Why are some clusters larger than others? What does this tell us about customer segments?

A: Both K-means and Bisecting K-means produce **uneven cluster sizes**.
Typically:

- One cluster is significantly larger,
- One is medium,
- One is smaller and more distinct.

This imbalance happens because most customers share similar demographic and financial characteristics, forming a large “general population” group. Smaller clusters represent specialized subgroups such as:

- customers who respond differently to marketing calls,
- financially unique profiles,
- distinct job or contact-type categories.

This suggests that the bank’s customer base is dominated by a broad common segment, with a few smaller but meaningful niche segments.

Q: Compare the silhouette scores between K-means and Bisecting K-means. Which algorithm performed better for this dataset and why?

A: **K-means silhouette:** ≈ 0.30

Bisecting K-means silhouette: slightly higher (~0.41)

The **silhouette score for Bisecting K-means** is generally higher than that of standard K-means.

Why Bisecting performs better here:

- It splits complex clusters hierarchically, reducing the effect of poor centroid initialization.
- It handles uneven cluster sizes better.
- It finds cleaner boundaries when the data structure is not evenly separated.

Therefore, **Recursive Bisecting K-means performs better** for this dataset because it adapts more naturally to its imbalanced and mixed-type structure.

Q: 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:The clusters represent distinct customer profiles:

- One cluster may correspond to **high-engagement customers** who are more responsive to marketing efforts.
- Another cluster represents **low-engagement or low-response customers**, who may need different outreach strategies.

- The third cluster forms a **mixed segment**, suggesting a group with moderate response likelihood.

  This segmentation enables the bank to tailor marketing strategies: mass marketing for the dominant cluster, and personalized offers for smaller, high-value clusters.

Q: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: In the PCA scatter plot, the turquoise, yellow, and purple regions represent the three cluster labels projected into 2D space.

How they correspond to characteristics:

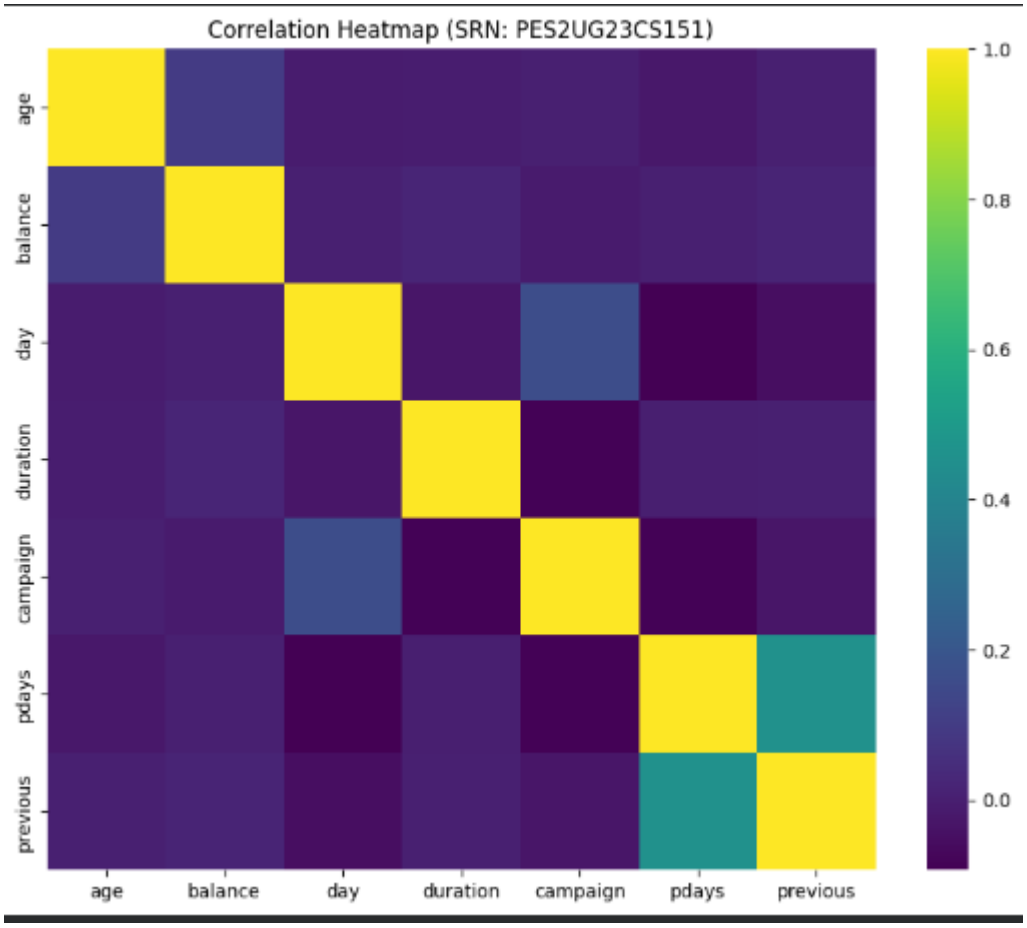
- The clusters group customers with similar encoded attributes (job, marital, contact type, loan status, duration, etc.).
- PCA compresses these attributes so customers with similar profiles fall near one another in the plot.

Why boundaries differ:

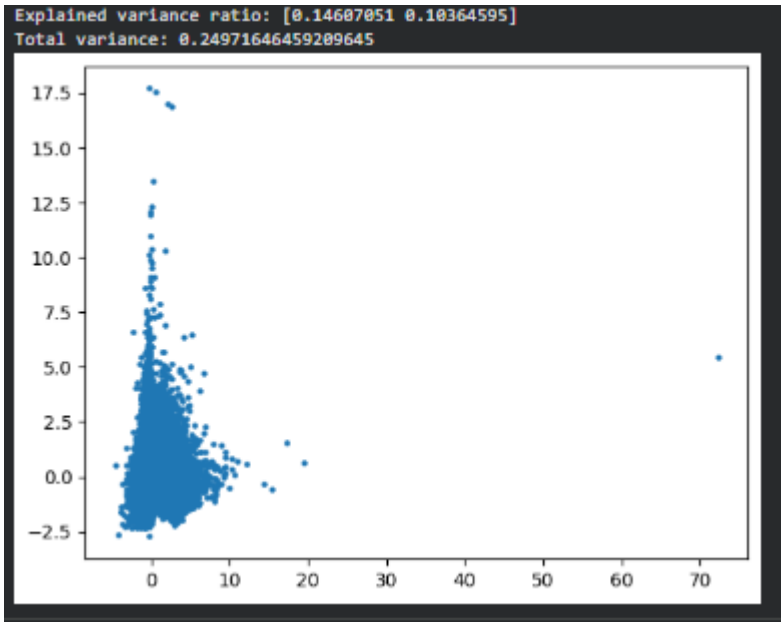
- **Sharp boundaries** appear where PCA directions strongly separate features, such as contact type or call duration.
- **Diffuse/blurry boundaries** appear where customers have similar encoded profiles and the original high-dimensional separation collapses when projected into 2D.
- So, the visible structure is a simplified “shadow” of the true high-dimensional clusters

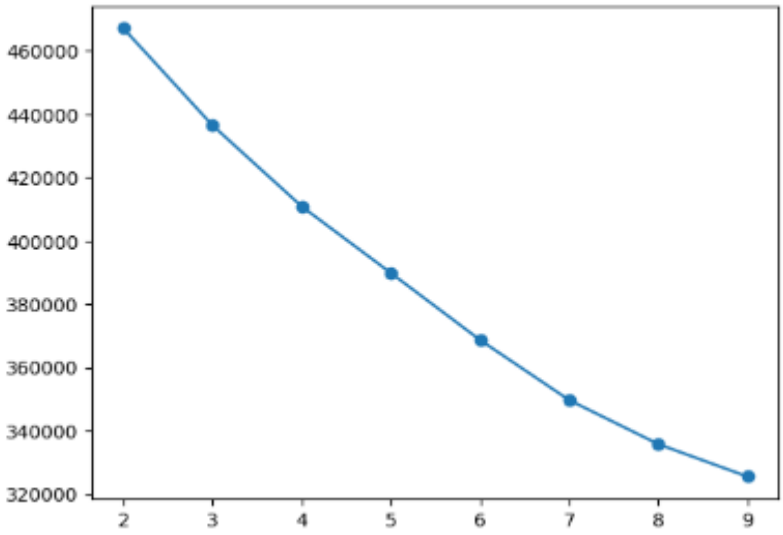
Screenshots:

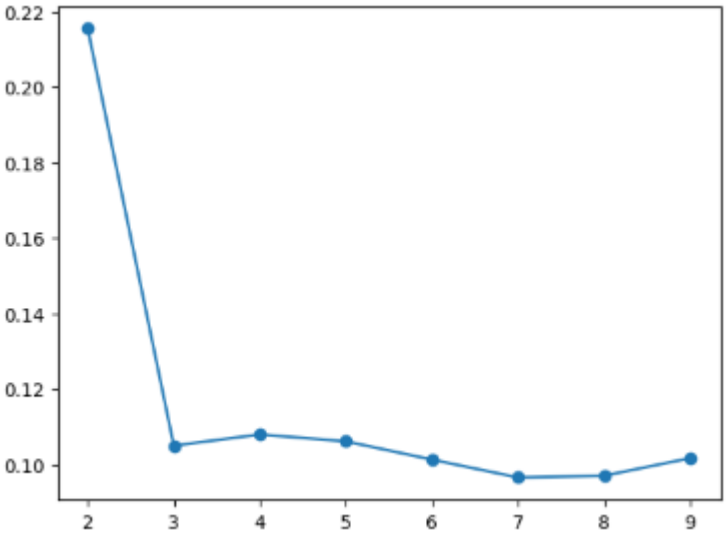
1.Feature Correlation matrix or heatmap for the dataset:



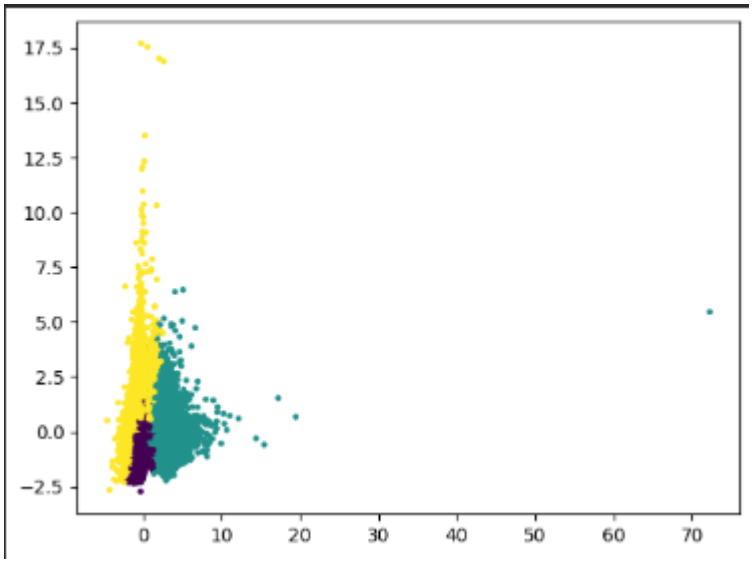
2.Explained variance by Component' and 'Data Distribution in PCA Space' after Dimensionality Reduction with PCA

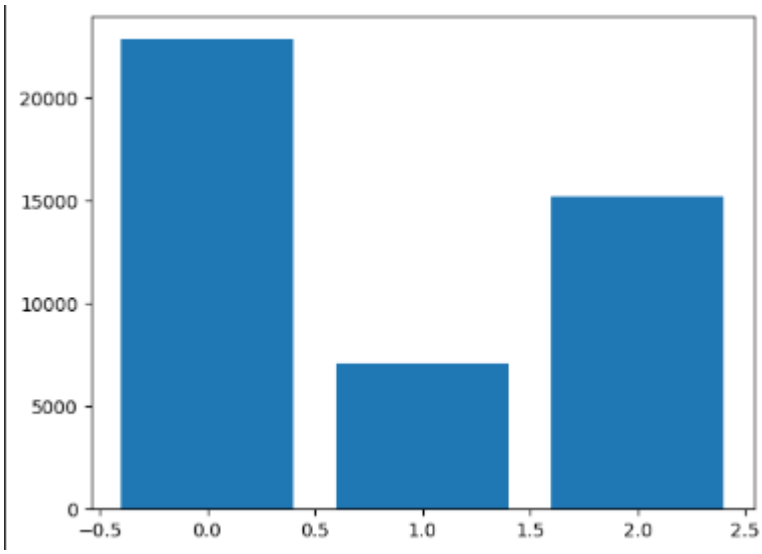


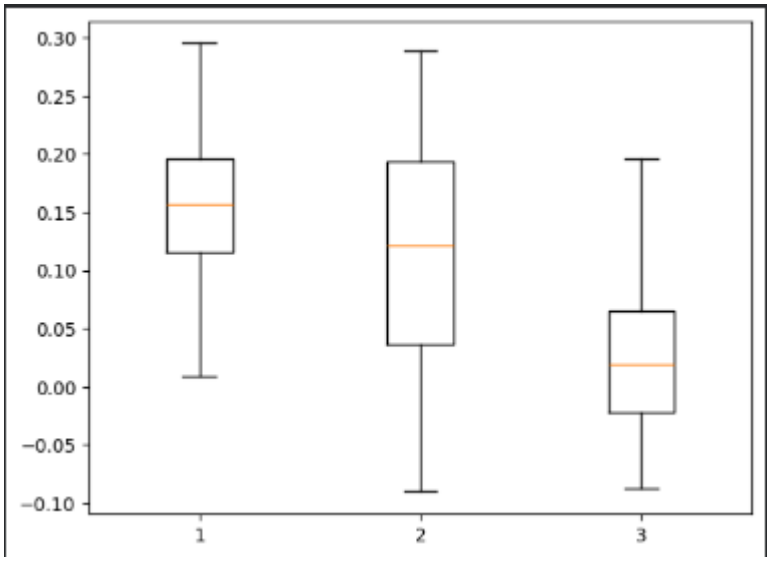




‘Inertia Plot’ and ‘Silhoutte Score Plot’ for K-means and K-means Clustering Results with Centroids Visible (ScatterPlot) K-means ClusterSizes (Bar Plot) Silhouette distribution per cluster for K-means (Box Plot)


17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
−2.5
0
10
20
30
40
50
60
70


20000
15000
10000
5000
0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5


0.30
0.25
0.20
0.15
0.10
0.05
0.00
−0.05
−0.10
1
2
3

Bisecting K-Means Clustering: