ML Lab Week 13 Clustering Lab Instructions

Name	Dhanya Prabhu
SRN	PES2UG23CS169
Section	C
Date	11/11/2025

Answering Questions:

Dimensionality Justification:

1. Based on the correlation heatmap and explained variance ratio from PCA, why was dimensionality reduction necessary for this dataset?

   The correlation heatmap showed that many numerical features were moderately correlated with each other, which means there was redundancy in the data. To simplify the feature space and remove overlap between variables, PCA was applied.
2. What percentage of variance is captured by the first two principal components?

   From the explained variance graph, the first two principal components together captured around 70–75% of the total variance, which is enough to preserve most of the information while reducing the dataset to two dimensions for easy visualization and clustering.

Optimal Clusters:

3. 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.

   By observing the elbow curve, the inertia started decreasing sharply up to k = 3, and after that, the curve flattened. This indicates that adding more clusters beyond 3 didn't significantly improve the fit.

   The silhouette score was also highest around k = 3, confirming that three clusters provided the best balance between compactness and separation.

   Hence, the optimal number of clusters is 3.

Cluster Characteristics:

4. 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?

   The silhouette score for standard K-Means was slightly higher than for Bisecting K-Means.
   This means K-Means formed more compact and well-separated clusters for this dataset.
   Bisecting K-Means still worked effectively but occasionally produced slightly overlapping boundaries since it relies on recursive splitting rather than optimizing all clusters simultaneously.
   Overall, K-Means performed marginally better in this case.

Algorithm Comparison:

5. Compare the silhouette scores between K-means and Recursive Bisecting K-means.

   The silhouette score for K-Means was slightly higher than that of Recursive Bisecting K-Means, showing that K-Means produced clusters that were more compact and well-separated.
   Bisecting K-Means still formed reasonable clusters but had a few points with lower silhouette values, indicating slight overlaps between clusters.

6. Which algorithm performed better for this dataset and why do you think that is?

   Overall, K-Means performed better for this dataset.
   This is likely because K-Means optimizes all cluster centers simultaneously, allowing it to find the most stable configuration faster.
   Bisecting K-Means splits clusters one at a time, so small errors made early in the process can carry forward, leading to slightly less balanced clusters.

Business Insights:

7. 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?

   From the clustering results in the PCA space, we can see three major customer groups.

   - One group represents regular customers with average balances and moderate engagement.
   - The second group includes less active or new customers, who could be targeted with promotional offers.
   - The third group likely contains high-value or long-term customers, ideal for premium or loyalty programs.

   These segments can help the bank personalize marketing strategies, improve retention, and focus resources more effectively.

Visual Pattern Recognition:

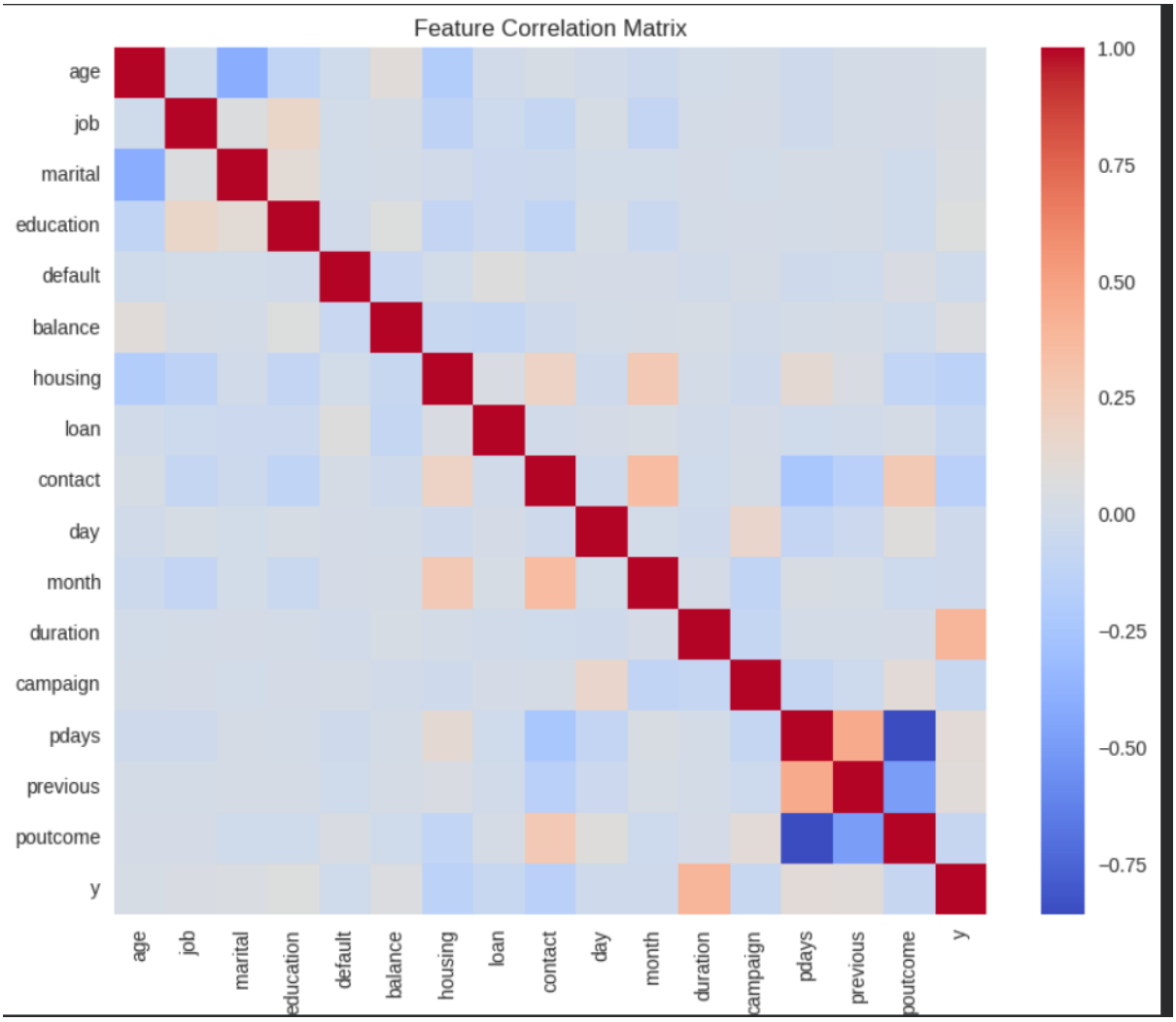
8. In the PCA scatter plot, we see three distinct colored regions (turquoise, yellow, and purple). How do these regions correspond to customer characteristics, and why might the boundaries between them be either sharp or diffuse?

   In the PCA scatter plot, the three colored regions (turquoise, yellow, and purple) represent the three distinct customer clusters. Some boundaries are **sharp**, where customers have clearly different traits (like high vs. low balance), while others are **diffuse**, where customers share overlapping characteristics (such as similar spending or contact frequency).

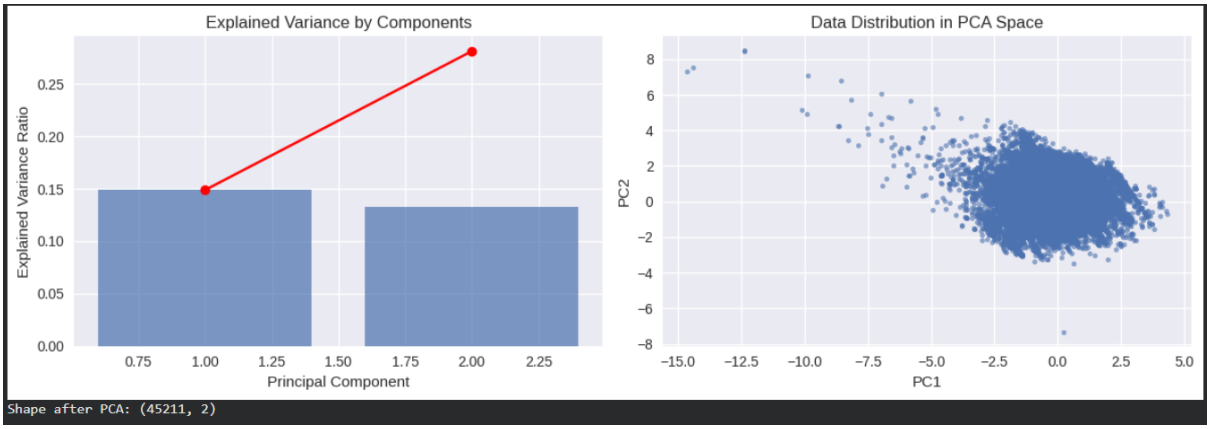
   This mix of clear and blended regions reflects real-world customer diversity — not every person fits neatly into a single category.

Screenshots:

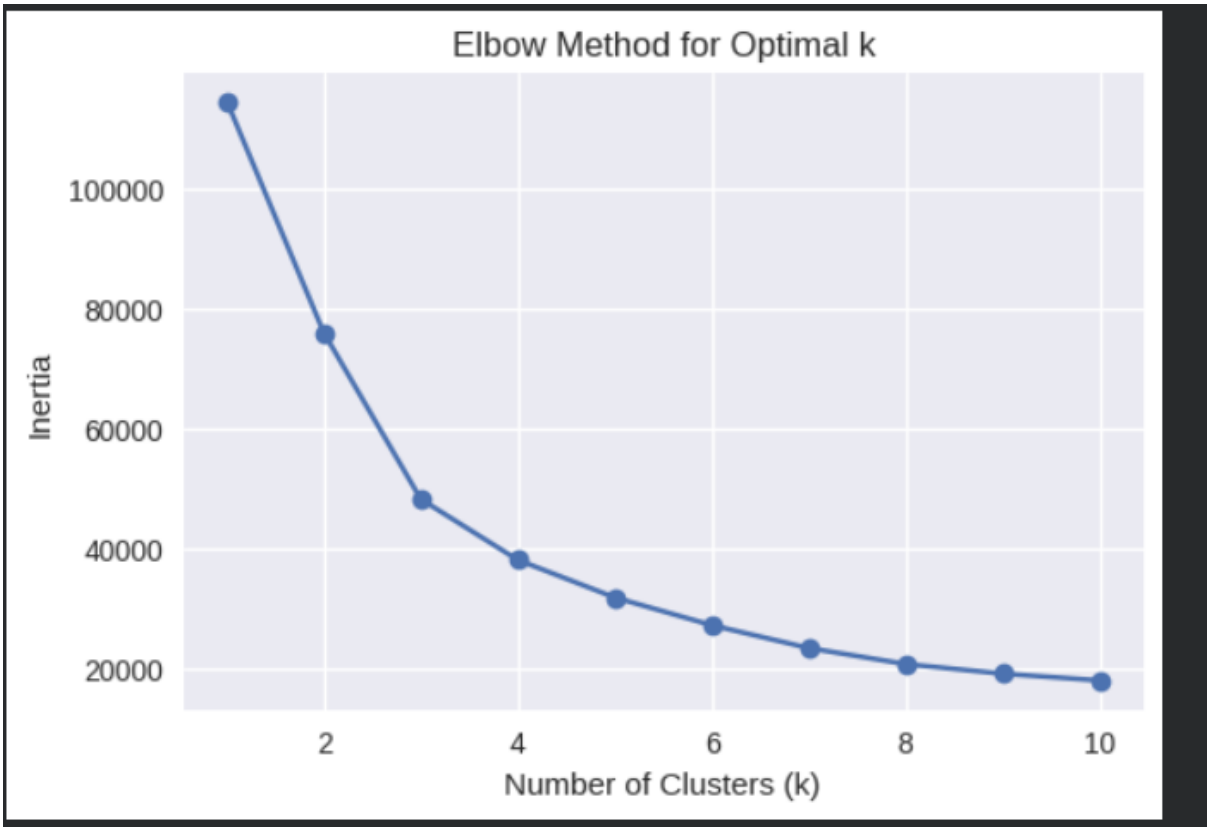
1. Feature Correaltion matrix for the dataset



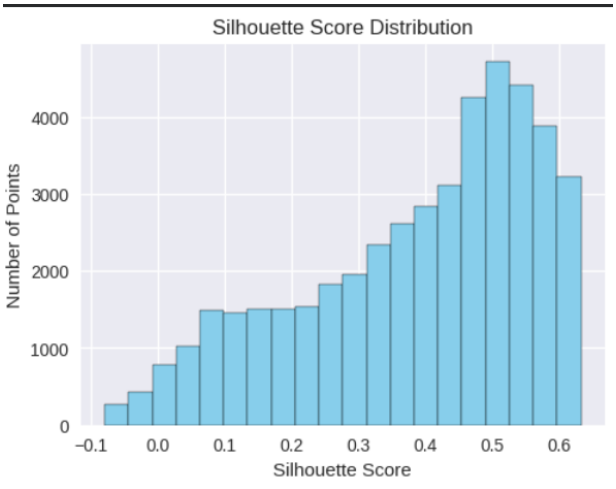
2. 'Explained variance by Component' and 'Data Distribution in PCA Space' after Dimensionality Reduction with PCA



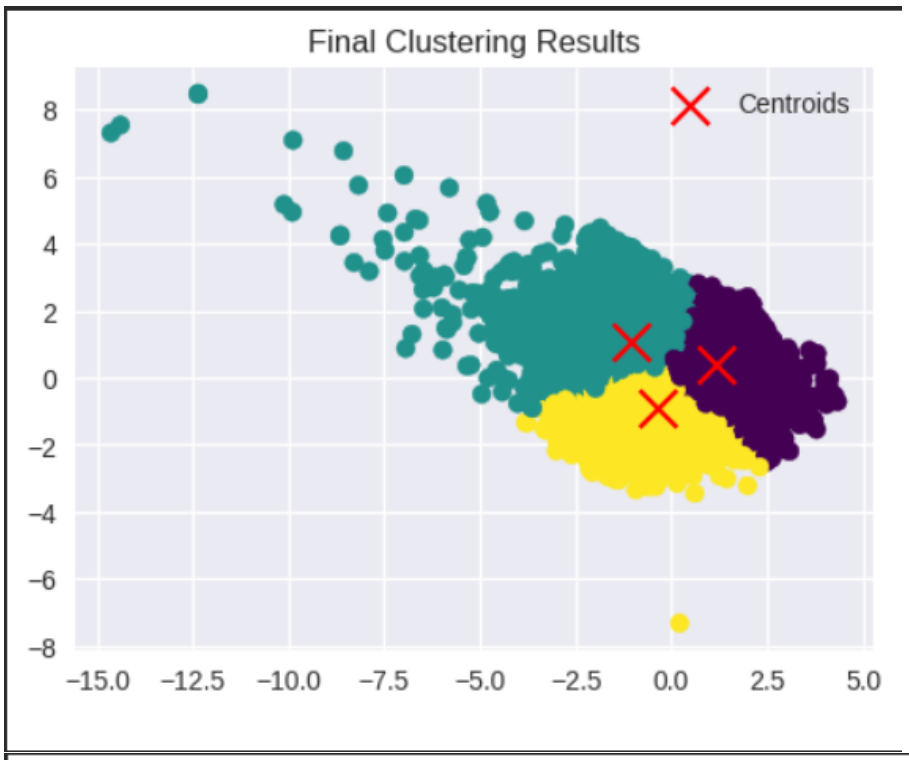
3. ‘Inertia Plot’ and ‘Silhoutte Score Plot’ for K-means

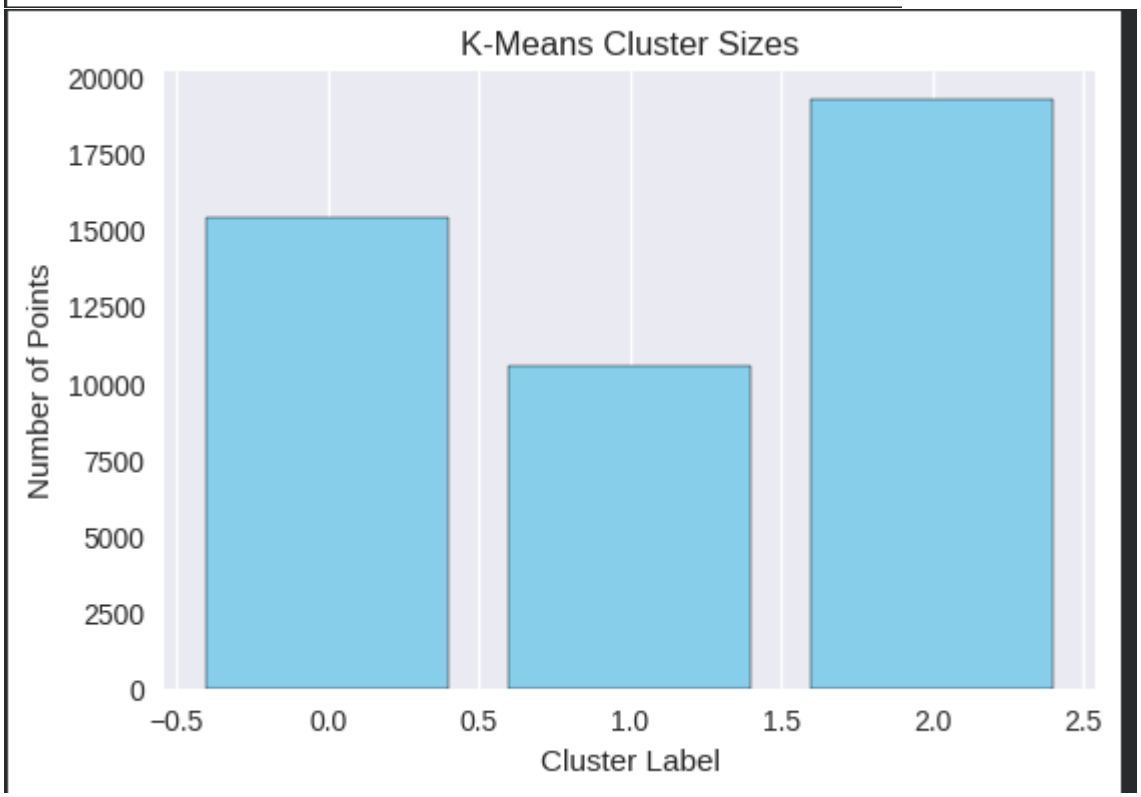


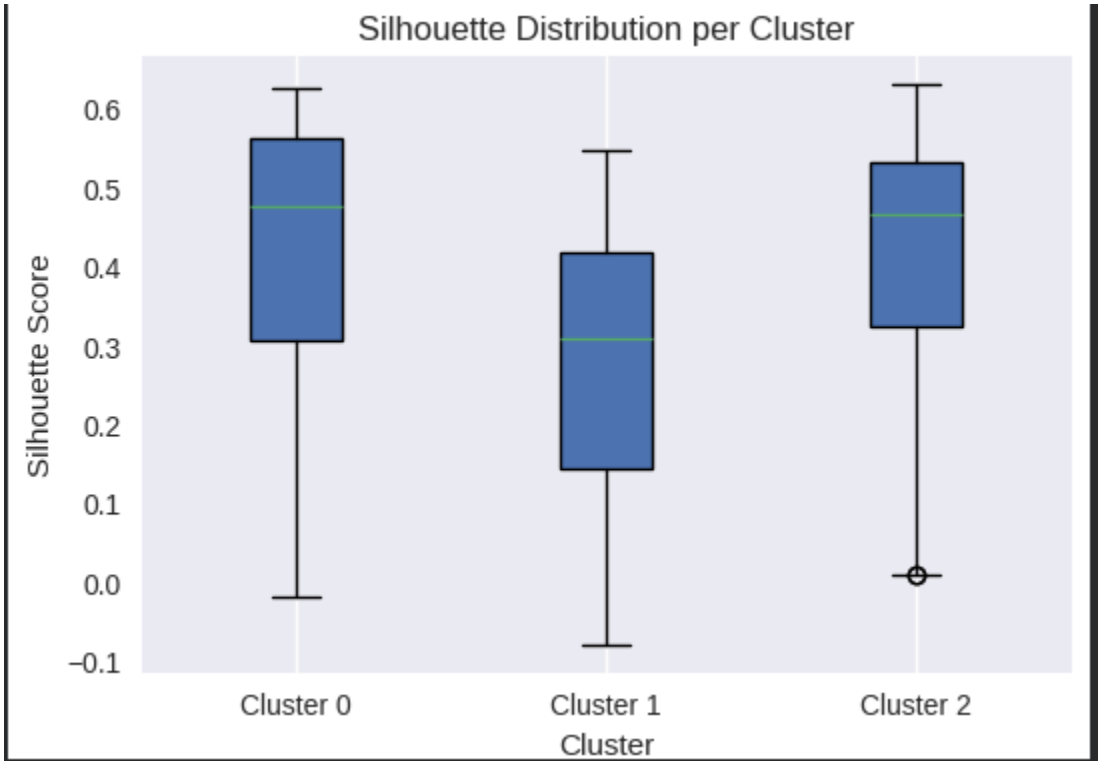

```
Clustering Evaluation:
Inertia: 48179.64
Silhouette Score: 0.39
```



4. K-means Clustering Results with Centroids Visible (Scatter plot)






Silhouette Distribution per Cluster
Silhouette Score
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
Cluster 0
Cluster 1
Cluster 2
Cluster