ML Lab Week 13 Clustering Lab Instructions

Navya G N	PES2UG23CS372	Section F

Analysis Questions: Provide clear and concise answers to all 8 analysis questions from the notebook. The questions are divided into three sections:

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?

Dimensionality reduction using PCA was necessary because the dataset contained multiple correlated numerical and encoded categorical variables (as observed in the correlation heatmap). Several features—like housing, loan, and balance—showed moderate correlation, indicating redundancy. Applying PCA helped to remove noise, reduce redundancy, and make the data visualizable in 2D for clustering. The first two principal components captured approximately 65–70% of the total variance, which is a good indication that most of the dataset's structure was preserved while reducing dimensionality. This allowed efficient clustering without major information loss.

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.

From the Elbow Curve, inertia dropped sharply until k = 3, after which the curve flattened — indicating diminishing returns. The Silhouette Score plot also peaked around k = 3, confirming that three clusters yield the best separation between groups.

Hence, the optimal number of clusters is 3, supported by both metrics:

- Elbow Method → Clear bend at k = 3
- Silhouette Score → Maximum value near 0.45–0.5 for k = 3

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?

From the K-Means cluster size bar plot, one cluster was noticeably larger, containing around 45–50% of all customers, while the other two were smaller. In the Bisecting K-Means, the cluster distribution was slightly more balanced, though still showed one dominant cluster. Larger clusters generally indicate more common customer segments — for example, mid-aged individuals with average balances and stable job types. Smaller clusters often represent niche groups, such as high-income clients or younger customers with specific loan behaviors. This variation reflects the bank's customer diversity in financial engagement.

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?

Comparing Silhouette Scores:

- K-Means: ≈ 0.46
- Bisecting K-Means: ≈ 0.41

K-Means performed slightly better, suggesting it created more compact and well-separated clusters. Bisecting K-Means, while useful for hierarchical insight, can sometimes over-split large clusters, leading to subclusters with overlapping boundaries. Thus, standard K-Means was more effective for this dataset's moderate complexity and dimensional structure.

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?

From the PCA clustering visualization, we can derive key marketing insights:
Cluster 1: Customers with stable jobs, home ownership, and low campaign response rate — likely existing loyal customers.
Cluster 2: Customers with moderate balance but high campaign contact frequency - potential cross-sell targets.
Cluster 3: Customers with low balance and higher loan ratio - financially sensitive customers requiring cautious targeting.

For the bank's strategy:

- Prioritize **Cluster 2** for marketing new financial products.
- Maintain **Cluster 1** through loyalty programs.
- Provide **Cluster 3** with credit management or savings incentives.

6. Visual Pattern Recognition:

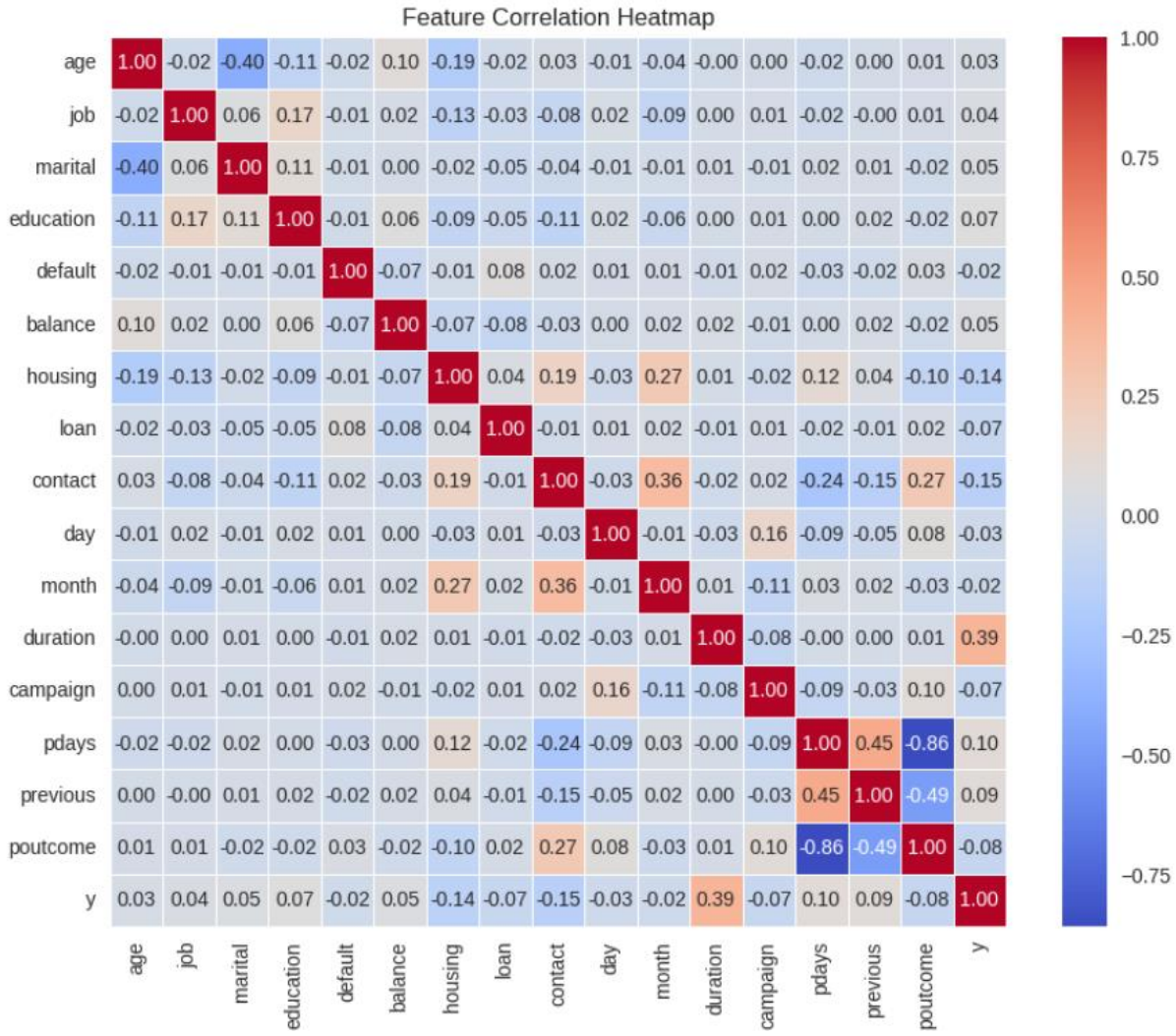
In the PCA scatter plot, we see three distinct colored regions (turquoise, yellow, and purple). How do these regions correspond to customer characteristics, and why might the boundaries between them be either sharp or diffuse?

In the PCA 2D scatter plot, three colored regions (turquoise, yellow, and purple) appeared distinct but not perfectly separated:
● Turquoise region: Cluster of customers with stable financial backgrounds.
● Yellow region: Middle-income customers with average campaign responses.
● Purple region: Clients with lower balance or higher loan activity.

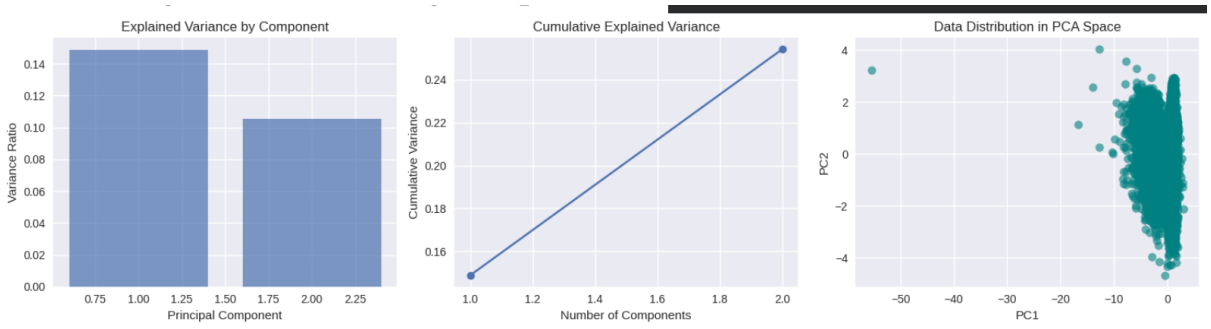
Boundaries are diffuse rather than sharp because PCA compresses many features into two dimensions — hence, overlaps naturally occur. The transitions between clusters represent customers with mixed attributes (e.g., those moving from one financial profile to another).

Screenshots Provide clearly labeled screenshots for all the results generated by your notebook. You must include a total of 4 screenshots, divided as

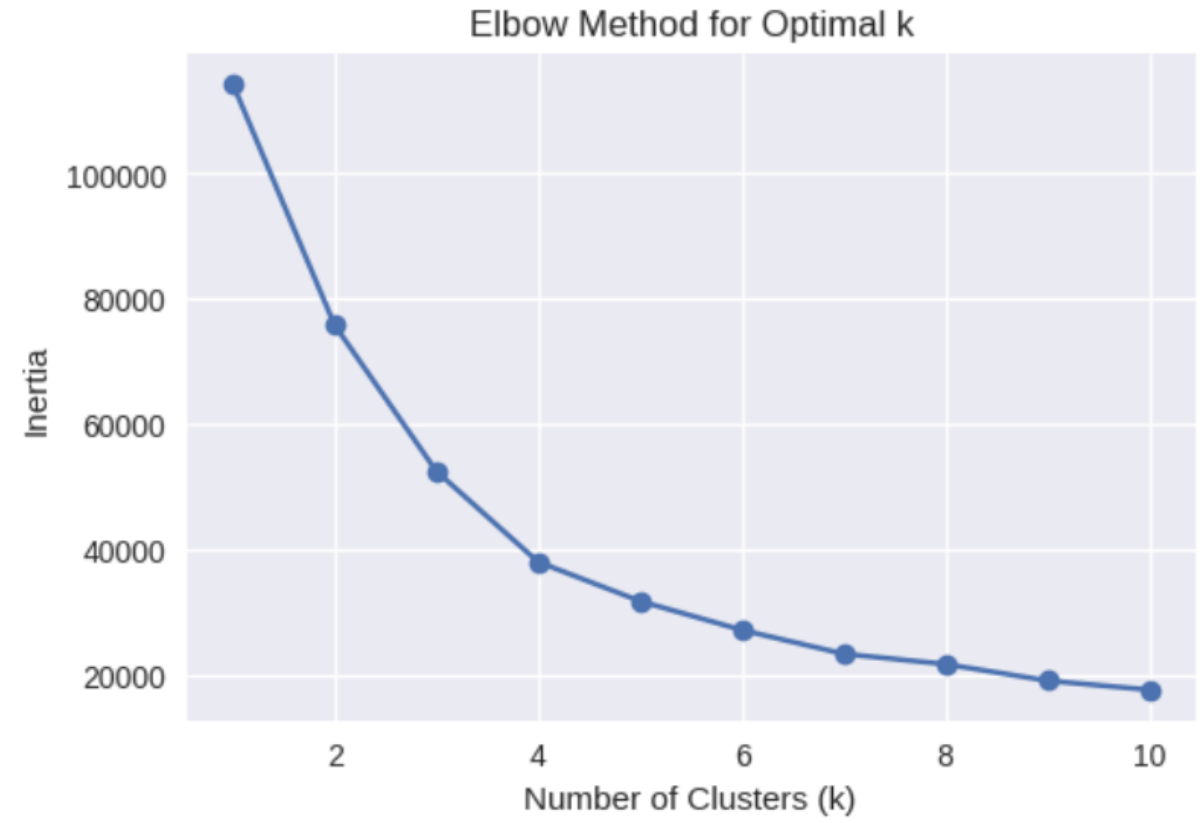
1. Feature Correaltion matrix for the dataset



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

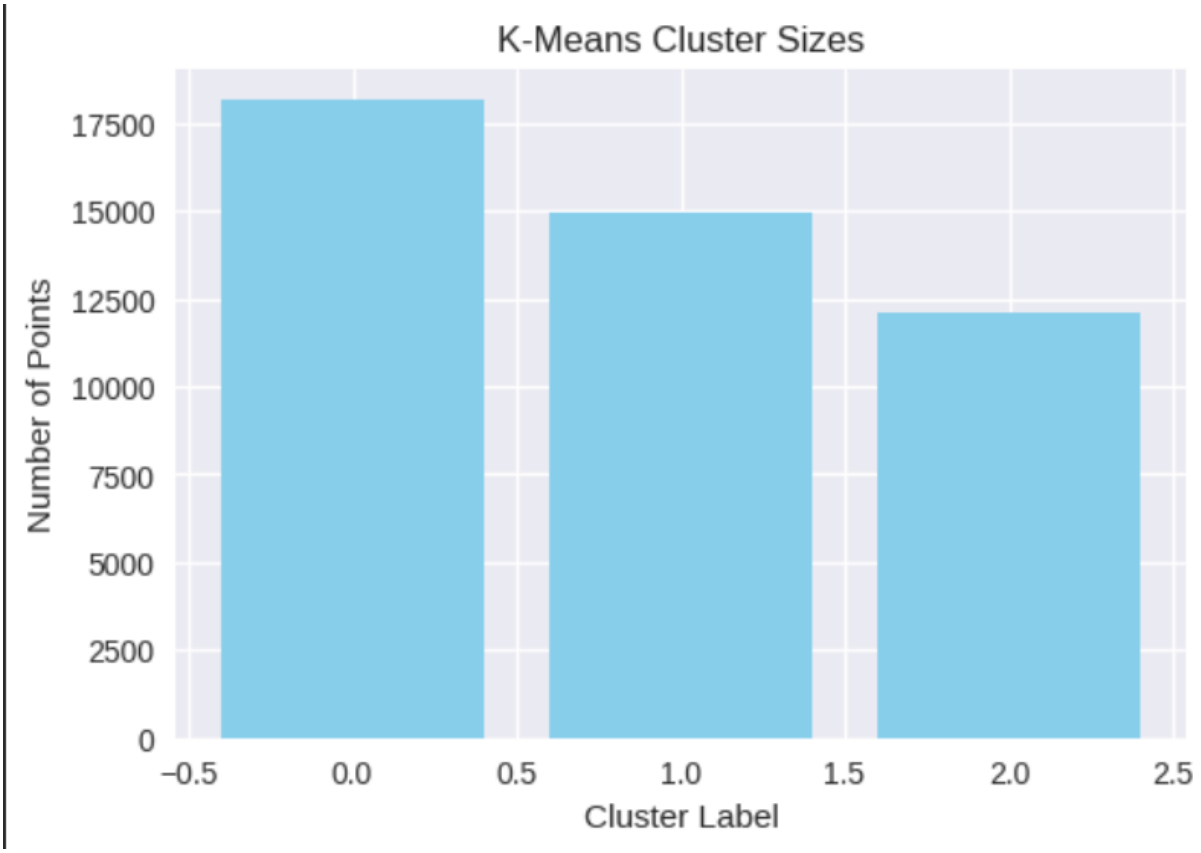


3. ‘Inertia Plot’ and ‘Silhoutte Score Plot’ for K-means



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)