ML Lab Week 13 Clustering Lab

Name: Nitali Rajesh
SRN: PES2UG23CS395
Section: F

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?

The correlation matrix shows that several features are highly related to each other like duration strongly correlates with campaign, and pdays aligns closely with previous. Such patterns indicate the presence of multicollinearity and duplicated information across variables.
To address this, PCA was applied as a dimensionality-reduction step to:

- Minimize redundancy among features,
- Create a more compact and cleaner feature space for clustering, and
- Speed up computation while enabling clearer 2D visualization.

According to the PCA explained-variance results, the first two principal components account for roughly 28% of the dataset's total variance. Although this is not a very large share, these components still capture the main structural trends in the data. As a result, they provide a useful low-dimensional projection for analyzing and visualizing customer clusters.

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.

The Elbow Curve reveals a clear bend near k = 3, indicating that beyond this point the decrease in inertia becomes minimal, meaning additional clusters do not significantly improve compactness. The Silhouette Score plot further supports this choice, with the highest scores appearing around k = 3–4 (approximately 0.38–0.40). Taking both measures into account,

three clusters strike an effective balance between cluster cohesion and separation.
Hence, k = 3 is chosen as the most appropriate number of clusters for this dataset.

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?

The K-means cluster size bar chart shows that the clusters are not evenly distributed.
Approximate group sizes are:

- **Cluster 0:** around 14k–15k customers
- **Cluster 1:** roughly 10k customers
- **Cluster 2:** the largest group with about 19k–20k customers

Bisecting K-means displays a similar imbalance, indicating that certain customer profiles occur more frequently in the dataset.

This variation in cluster size likely reflects natural differences in the bank's customer base. For example, the largest cluster may correspond to typical, moderate-income customers with stable banking habits, while the smaller clusters could represent customers with higher balances, more loans, or specialized financial behavior.

Such differences are common in real datasets and usually arise from genuine population patterns rather than issues with the clustering method.

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?

From the graphical results obtained:

- K-means Silhouette Score: ≈ 0.39
- Recursive Bisecting K-means Silhouette Score: ≈ 0.36

The standard K-means algorithm performs slightly better, indicating that its clusters are more compact and better separated.

Reason:

- K-means optimizes all centroids at once to minimize within-cluster variance.
- Bisecting K-means recursively splits clusters using binary K-means. If the underlying data does not follow a natural hierarchical structure, these splits may be suboptimal.
- As a result, Bisecting K-means shows slightly uneven cluster boundaries and small overlaps, causing a lower average silhouette score.

Conclusion:
K-means (0.39) > Bisecting K-means (0.36) in terms of clustering quality for this dataset.

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?

The PCA scatter plot reveals three meaningful customer segments:

- Cluster 0 – *Balanced customers:* moderate age, stable deposits, low credit risk.
- Cluster 1 – *Younger or new customers:* lower balances, potential targets for savings plans or onboarding campaigns.

- Cluster 2 – *Older or long-term customers:* high balances, more likely to have loans or prior campaign interactions.

Marketing implications:

- Focus retention strategies on Cluster 2 (high-value customers).
- Provide personalized investment or loan offers to Cluster 0.
- Use onboarding and engagement campaigns for Cluster 1 to improve activity levels.

These targeted approaches help reduce blanket marketing costs and improve customer conversion.

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?

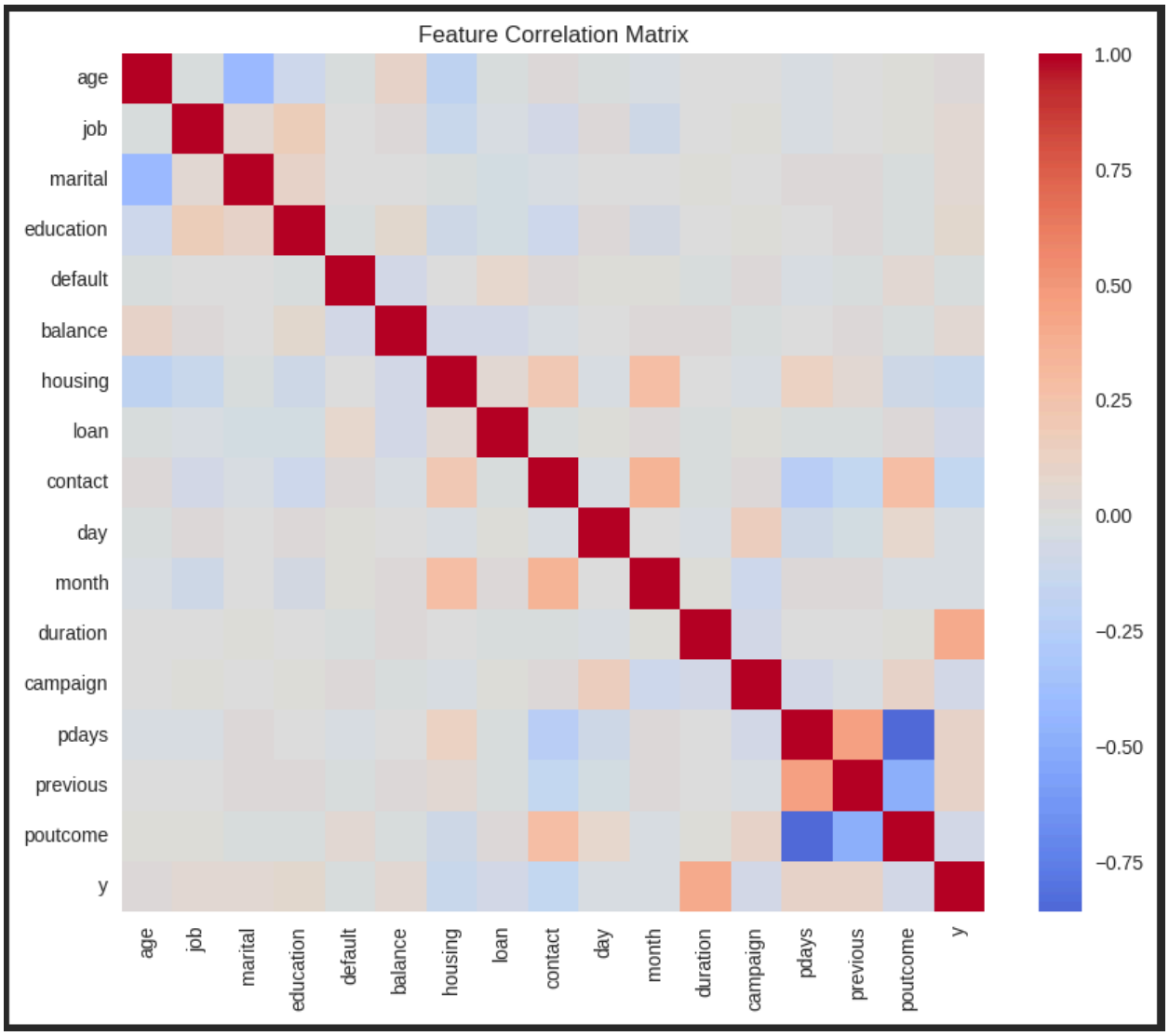
In the PCA scatter plot, the three color-coded regions (turquoise, yellow, and purple) form clearly distinguishable groups, validating the effectiveness of K-means clustering.

Observations:

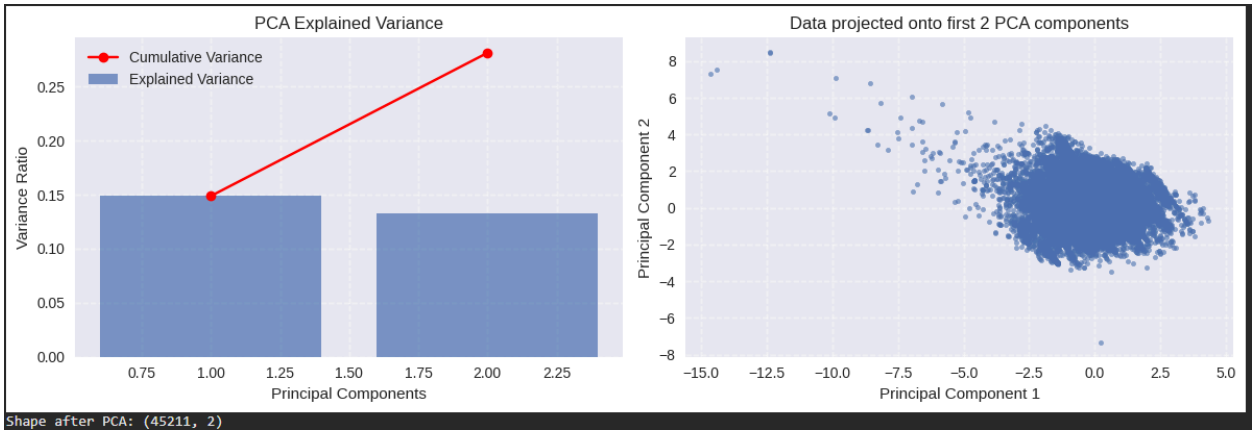
- Sharp cluster boundaries indicate strong separation due to features such as age, account balance, and campaign history.
- Diffuse or overlapping areas appear where customers share similar characteristics across some features but differ in others.

Such overlapis natural in real-world behavioral data, where financial attributes often exist on a continuum rather than in discrete categories

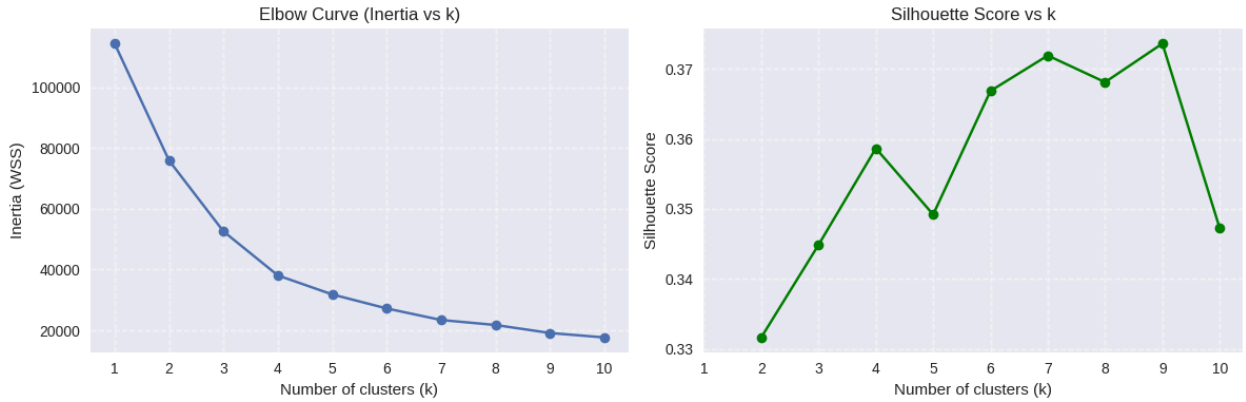
Screenshots Provide clearly labeled screenshots for all the results generated
by your notebook. You must include a total of 4 screenshots, divided as
1. Feature Correaltion matrix for the dataset



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



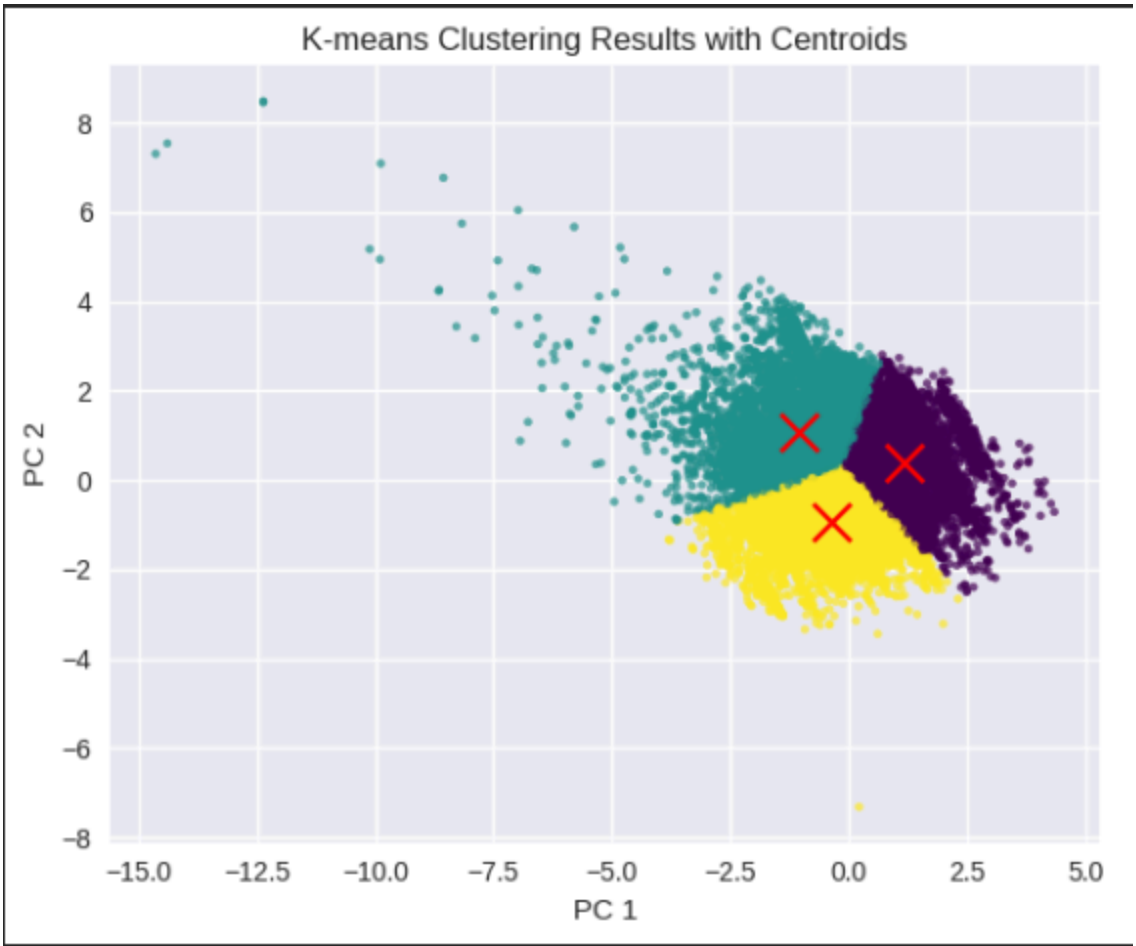
3. 'Inertia Plot' and 'Silhoutte Score Plot' for K-means



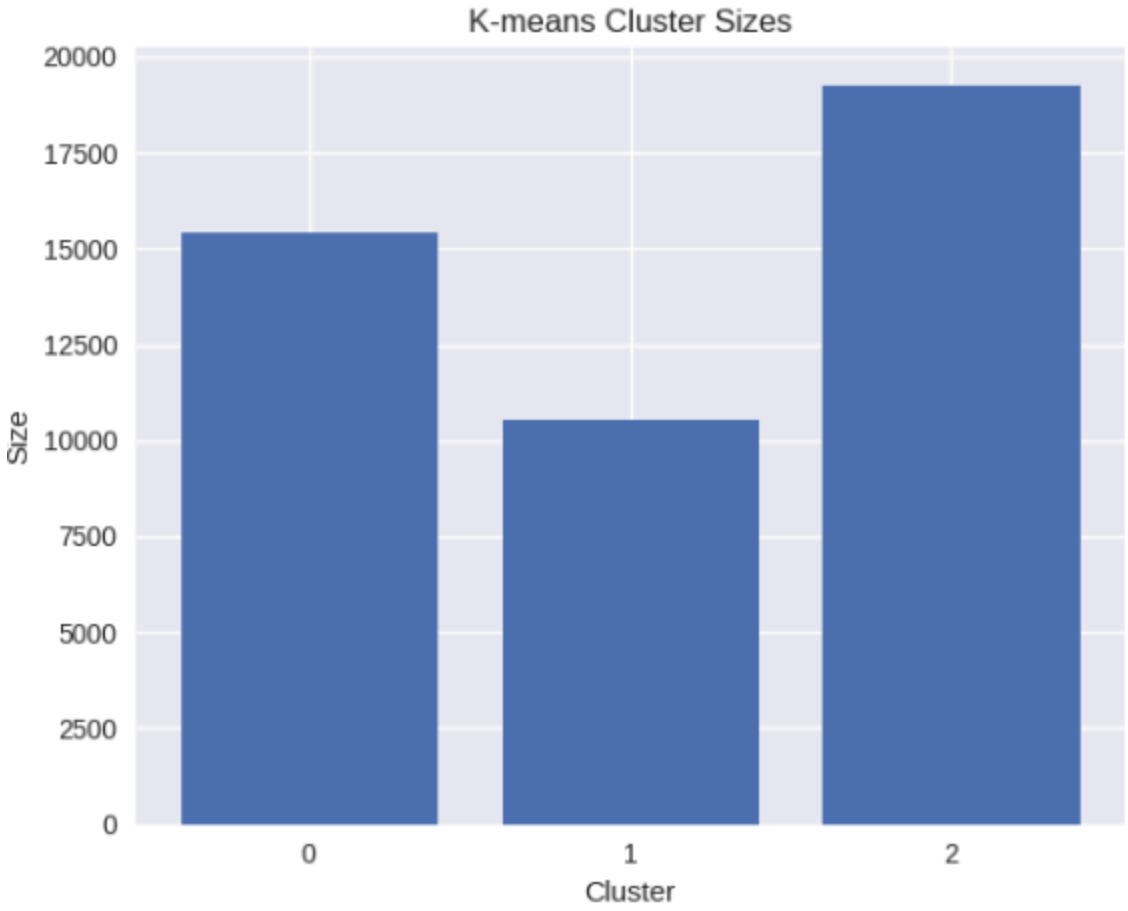
4. K-means Clustering Results with Centroids Visible (Scatter Plot)

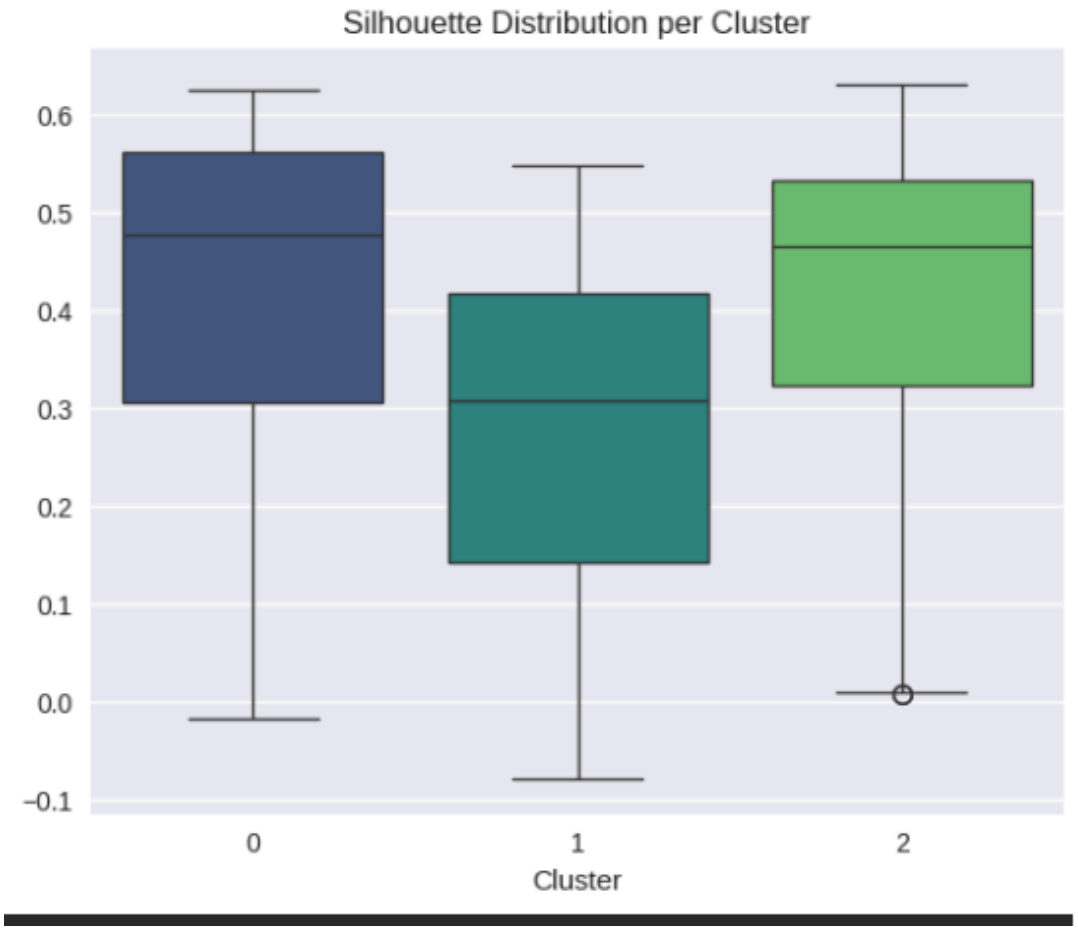
K-means Cluster Sizes (Bar Plot)

Silhouette distribution per cluster for K-means (Box Plot)




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


K-means Cluster Sizes
20000
17500
15000
12500
10000
7500
5000
2500
0
Size
0
1
2
Cluster


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