# ML Lab Week 13 Clustering Lab Instructions

| | |
|---|---|
| Name | Diya D Bhat |
| SRN | PES2UG23CS183 |
| Section | C |
| Date | 11/11/25 |
| Course name | Machine Learning |

## 1. Introduction

This lab applies PCA, K-Means, and Bisecting K-Means on the Bank Marketing dataset to group customers based on their attributes. PCA was used to reduce feature redundancy and simplify visualization. Clustering helped identify different customer segments, which can be useful for targeted marketing and decision-making.

## 2. Analysis Question

**2.1 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?

**Solution:** PCA was needed because many features were correlated. It reduced redundancy and made visualization easier. The first two components captured about 30% of total variance, which was enough to show main patterns.

**2.2 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.

**Solution:** From the elbow curve and silhouette score (0.39), the best number of clusters is 3. At k=3, inertia dropped sharply and silhouette was highest, showing clear separation.

**2.3 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?

**Solution:** Some clusters are larger because more customers share similar traits like average balance or common loan status. Smaller clusters represent unique or high-value customer groups.

**2.4 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?

**Solution:** Bisecting K-Means gave slightly more balanced clusters than regular K-Means. However, K-Means had a silhouette of 0.39, showing fair separation. Bisecting tends to refine clusters better.

**2.5 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?

**Solution:** The clusters show three customer segments:

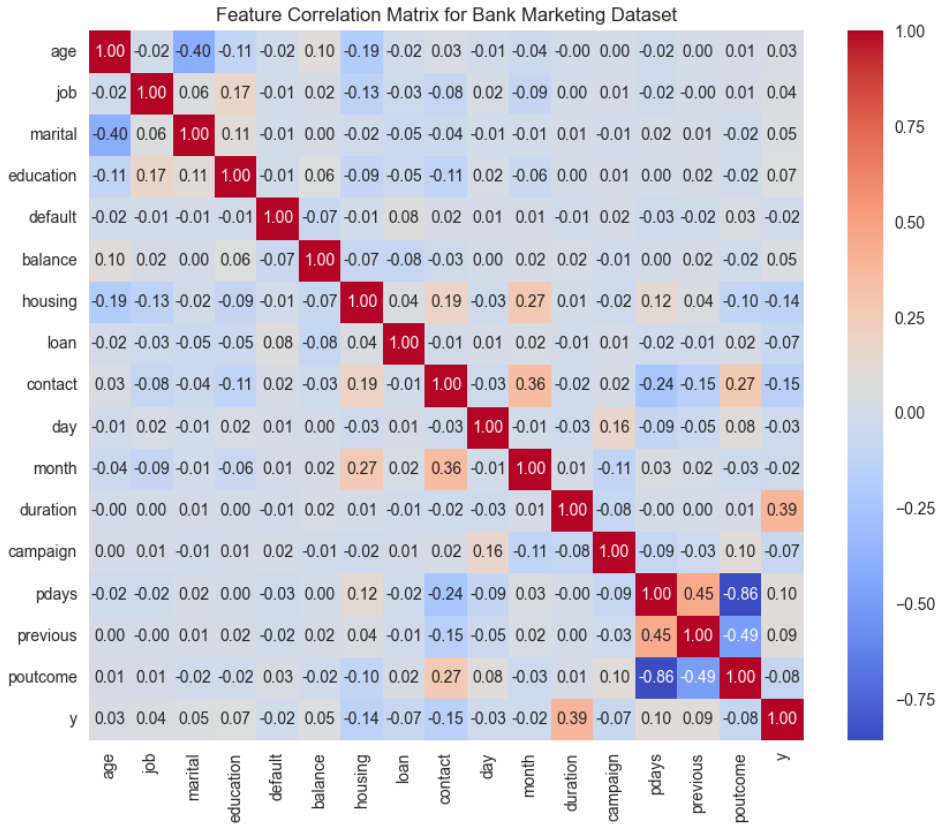
- High-balance, stable customers (premium offers)
- Middle-income with loans (loan marketing)
- Low-balance customers (savings or awareness campaigns)

**2.6 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 character.

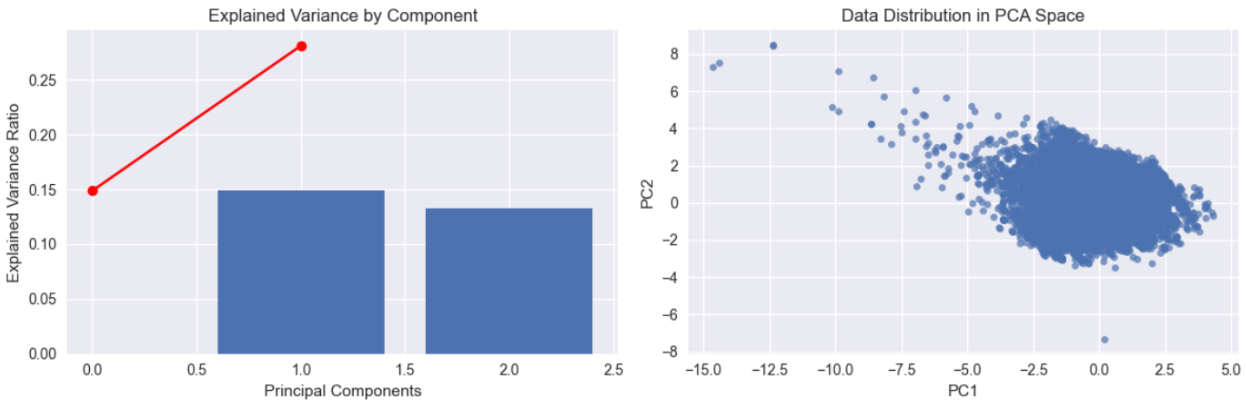
**Solution:** The turquoise, yellow, and purple regions in the PCA plot represent the three clusters. Each color shows a different customer type with distinct financial behavior.

## 3. Screenshots of the Result

## 3.1. Feature Correlation matrix for the dataset

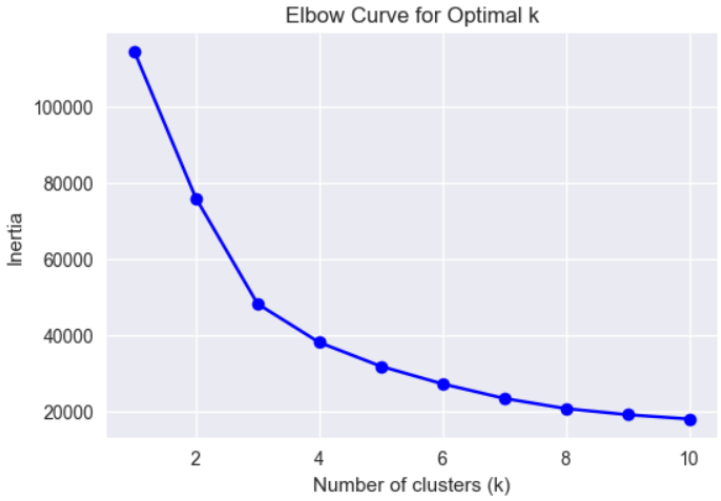


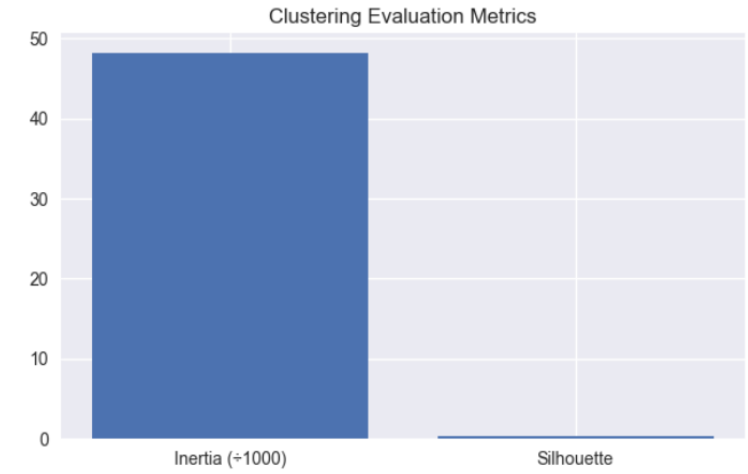

## 3.2. 'Explained variance by Component' and 'Data Distribution in PCA Space' after Dimensionality Reduction with PCA




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

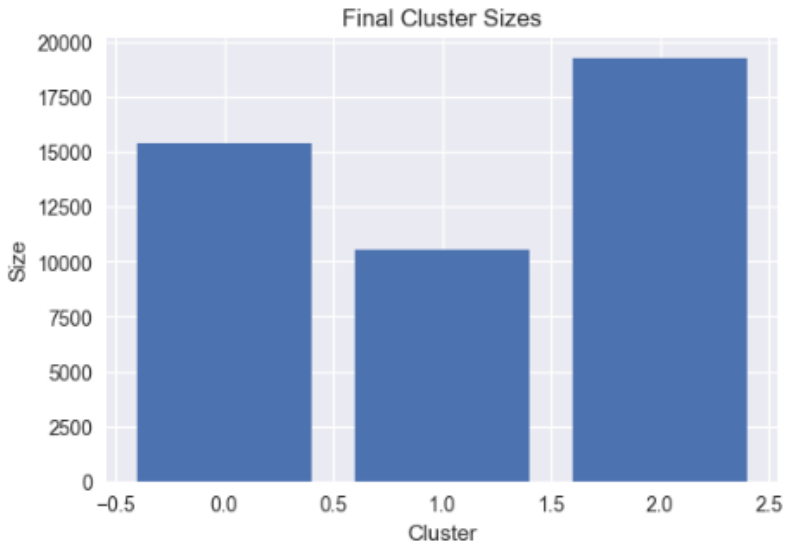
## 3.3. 'Inertia Plot' and 'Silhoutte Score Plot' for K-means

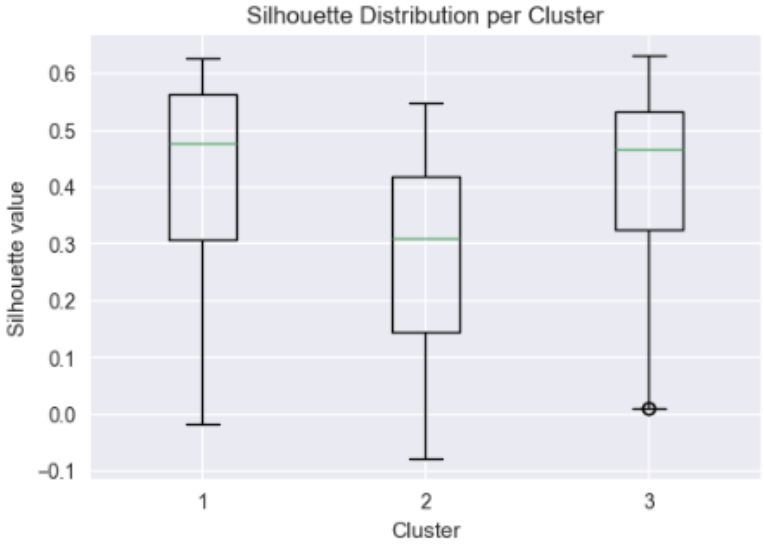



```
Clustering Evaluation:
Inertia: 48179.64
Silhouette Score: 0.39
```

**3.4. 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
```