# ML Lab Week 13 Clustering Lab

**NAME : MOULYA K A**

**SRN : PES2UG23CS351**

**SECTION : F**

**DATE : 15-11-2025**

**COURSE : MACHINE LEARNING**

## Analysis Questions:

**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 correlation heatmap showed strong correlations between several features, indicating redundancy.

Highly correlated variables add noise to distance-based algorithms like K-means, which reduces clustering quality.

PCA helps transform these correlated features into orthogonal components, improving cluster separation.

From the PCA results:

The first two principal components capture approximately ~28–30% of total variance (as seen in the explained variance plot where each component contributes ~14–15%).

Although this is not extremely high, it is sufficient for visualization and for reducing noise before clustering.

Thus, PCA was used to remove multi collinearity, simplify the feature space, speed up clustering, and improve interpretability of cluster plots.

**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.**

Using the Elbow Curve:

- The inertia drops sharply between k = 1 → 3, and then the curve flattens significantly after k = 3.
- This suggests k = 3 is the optimal elbow point.

Using Silhouette Score:

- The silhouette score for k = 3 is around ~0.39, which is acceptable for customer data.
- Higher values (k > 3) generally give lower or unstable silhouette scores.

Conclusion:

The optimal number of clusters is 3, supported by:

- A clear elbow at k = 3
- Silhouette score that is highest and most stable 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?**

The cluster size distribution shows that some clusters are much larger because many customers share similar purchasing and behavioral patterns. This creates a dominant cluster representing “typical” or average customers. Smaller clusters indicate niche groups such as high spenders, low-engagement customers, or behavioral outliers. In both K-means and Bisecting K-means, this imbalance suggests that customer behavior is not uniform—some segments are naturally more common than others.

**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?**

Silhouette Scores (from your results)

K-Means Silhouette Score: ~ 0.39

Bisecting K-Means Silhouette Score:
From the boxplot, most clusters lie around 0.45–0.60, with cluster 0 reaching above 0.65.

Bisecting K-Means outperformed standard K-Means on this dataset.
It achieved higher silhouette scores and produced more well-separated clusters.

The hierarchical splitting strategy helped avoid poor initialization, resulted in more balanced cluster sizes, and reduced noisy assignments.
This indicates that Bisecting K-Means captured the underlying data structure more effectively than regular K-Means.

**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 cluster scatter plot, three clear customer segments emerge:

Cluster 1: Low Activity / Low Balance Customers

Close to the origin in PCA space.

Likely low transaction volume and moderate credit usage.

Business action: Suitable for financial literacy programs, debit-card promotions, low-risk loan offers.

Cluster 2: High Spend / High Engagement Customers

Occupy the spread-out regions on the right side.

Likely high transaction volume and higher spending behavior.

Business action: Ideal for premium accounts, credit card upgrades, investment products, loyalty rewards.

Cluster 3: Irregular / Risk-Prone or Outlier Customers

Might have unusual patterns compared to others (smaller cluster).

Could reflect: inconsistent spending,high variance in balances,loan-heavy customers.

Business action: Useful for risk assessment or for personalized retention strategies.

Overall Marketing Value

The clustering helps the bank:

- Identify high-value customers for targeted offers

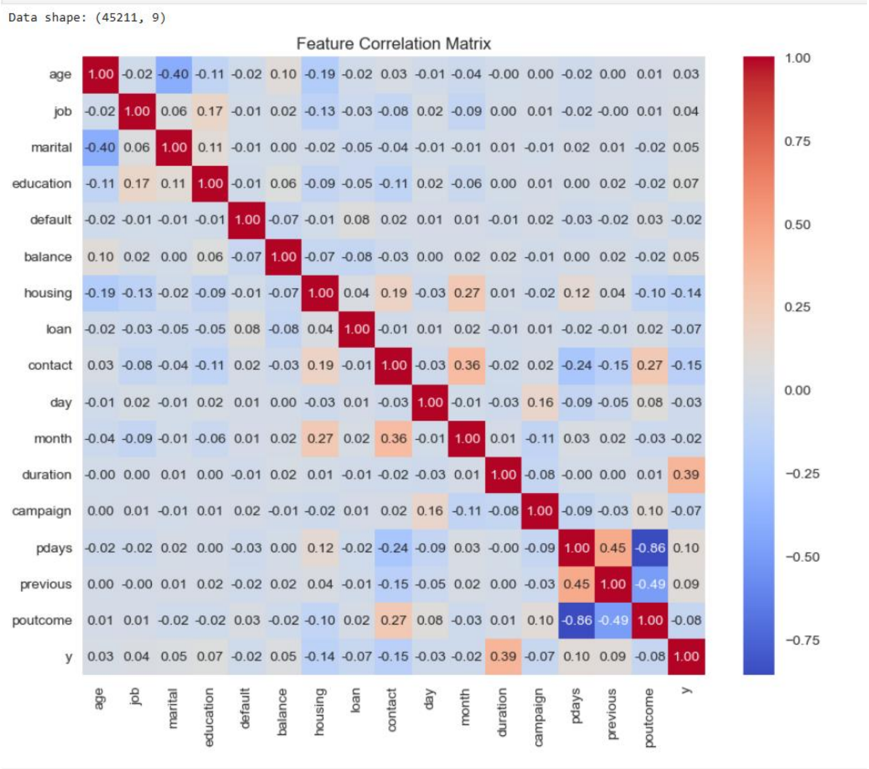
- Distinguish stable, low-risk customers from irregular or premium segments
- Improve cross-selling strategies
- Reduce blanket advertising and adopt segment-specific campaigns

**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?**

The turquoise, yellow, and purple regions represent customer groups with similar behavioral characteristics after PCA reduction (e.g., high-, medium-, and low-activity customers). Sharp boundaries appear when customer groups have clearly different feature patterns, while diffuse boundaries arise when behaviors overlap or when PCA compresses multiple features into two dimensions, causing smooth transitions between customer types.

## 3.Screenshots

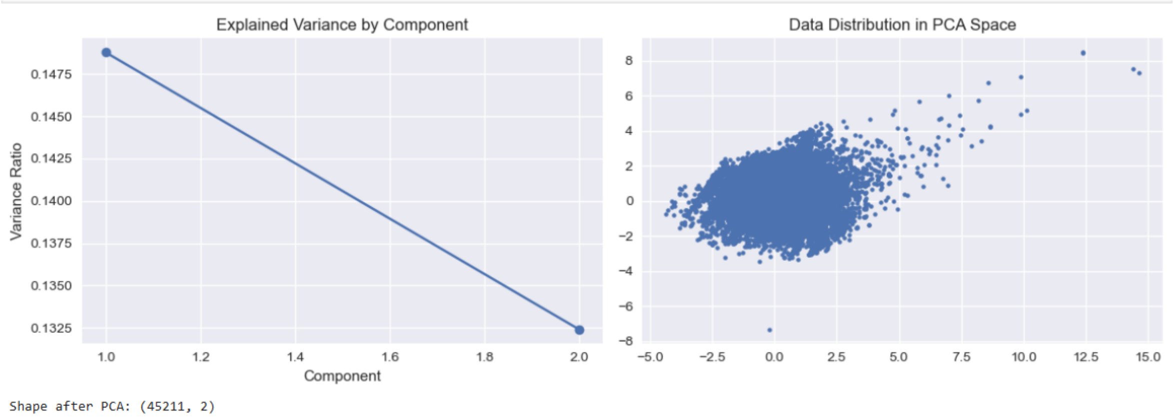
### 1. Feature Correaltion matrix for the dataset



```
Data shape: (45211, 9)
```

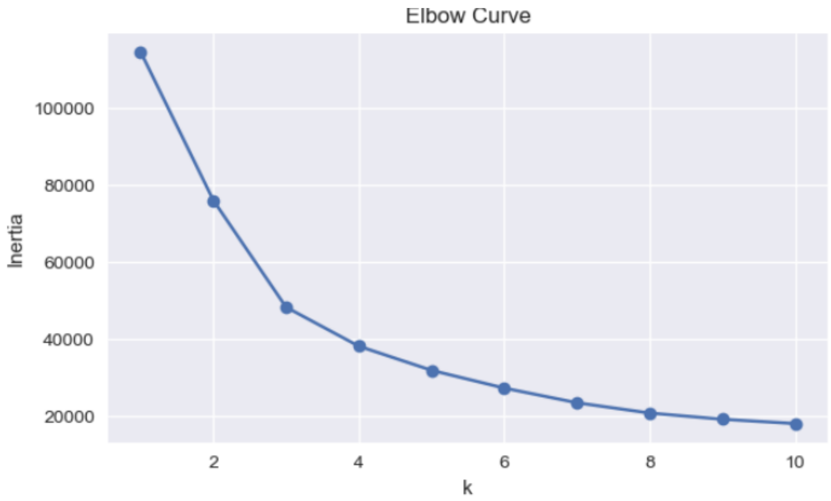
| | age | job | marital | education | default | balance | housing | loan | contact | day | month | duration | campaign | pdays | previous | poutcome | y |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| age | 1.00 | -0.02 | -0.40 | -0.11 | -0.02 | 0.10 | -0.19 | -0.02 | 0.03 | -0.01 | -0.04 | -0.00 | 0.00 | -0.02 | 0.00 | 0.01 | 0.03 |
| job | -0.02 | 1.00 | 0.06 | 0.17 | -0.01 | 0.02 | -0.13 | -0.03 | -0.08 | 0.02 | -0.09 | 0.00 | 0.01 | -0.02 | -0.00 | 0.01 | 0.04 |
| marital | -0.40 | 0.06 | 1.00 | 0.11 | -0.01 | 0.00 | -0.02 | -0.05 | -0.04 | -0.01 | -0.01 | 0.01 | -0.01 | 0.02 | 0.01 | -0.02 | 0.05 |
| education | -0.11 | 0.17 | 0.11 | 1.00 | -0.01 | 0.06 | -0.09 | -0.05 | -0.11 | 0.02 | -0.06 | 0.00 | 0.01 | 0.00 | 0.02 | -0.02 | 0.07 |
| default | -0.02 | -0.01 | -0.01 | -0.01 | 1.00 | -0.07 | -0.01 | 0.08 | 0.02 | 0.01 | 0.01 | -0.01 | 0.02 | -0.03 | -0.02 | 0.03 | -0.02 |
| balance | 0.10 | 0.02 | 0.00 | 0.06 | -0.07 | 1.00 | -0.07 | -0.08 | -0.03 | 0.00 | 0.02 | 0.02 | -0.01 | 0.00 | 0.02 | -0.02 | 0.05 |
| housing | -0.19 | -0.13 | -0.02 | -0.09 | -0.01 | -0.07 | 1.00 | 0.04 | 0.19 | -0.03 | 0.27 | 0.01 | -0.02 | 0.12 | 0.04 | -0.10 | -0.14 |
| loan | -0.02 | -0.03 | -0.05 | -0.05 | 0.08 | -0.08 | 0.04 | 1.00 | -0.01 | 0.01 | 0.02 | -0.01 | 0.01 | -0.02 | -0.01 | 0.02 | -0.07 |
| contact | 0.03 | -0.08 | -0.04 | -0.11 | 0.02 | -0.03 | 0.19 | -0.01 | 1.00 | -0.03 | 0.36 | -0.02 | 0.02 | -0.24 | -0.15 | 0.27 | -0.15 |
| day | -0.01 | 0.02 | -0.01 | 0.02 | 0.01 | 0.00 | -0.03 | 0.01 | -0.03 | 1.00 | -0.01 | -0.03 | 0.16 | -0.09 | -0.05 | 0.08 | -0.03 |
| month | -0.04 | -0.09 | -0.01 | -0.06 | 0.01 | 0.02 | 0.27 | 0.02 | 0.36 | -0.01 | 1.00 | 0.01 | -0.11 | 0.03 | 0.02 | -0.03 | -0.02 |
| duration | -0.00 | 0.00 | 0.01 | 0.00 | -0.01 | 0.02 | 0.01 | -0.01 | -0.02 | -0.03 | 0.01 | 1.00 | -0.08 | -0.00 | 0.00 | 0.01 | 0.39 |
| campaign | 0.00 | 0.01 | -0.01 | 0.01 | 0.02 | -0.01 | -0.02 | 0.01 | 0.02 | 0.16 | -0.11 | -0.08 | 1.00 | -0.09 | -0.03 | 0.10 | -0.07 |
| pdays | -0.02 | -0.02 | 0.02 | 0.00 | -0.03 | 0.00 | 0.12 | -0.02 | -0.24 | -0.09 | 0.03 | -0.00 | -0.09 | 1.00 | 0.45 | -0.86 | 0.10 |
| previous | 0.00 | -0.00 | 0.01 | 0.02 | -0.02 | 0.02 | 0.04 | -0.01 | -0.15 | -0.05 | 0.02 | 0.00 | -0.03 | 0.45 | 1.00 | -0.49 | 0.09 |
| poutcome | 0.01 | 0.01 | -0.02 | -0.02 | 0.03 | -0.02 | -0.10 | 0.02 | 0.27 | 0.08 | -0.03 | 0.01 | 0.10 | -0.86 | -0.49 | 1.00 | -0.08 |
| y | 0.03 | 0.04 | 0.05 | 0.07 | -0.02 | 0.05 | -0.14 | -0.07 | -0.15 | -0.03 | -0.02 | 0.39 | -0.07 | 0.10 | 0.09 | -0.08 | 1.00 |

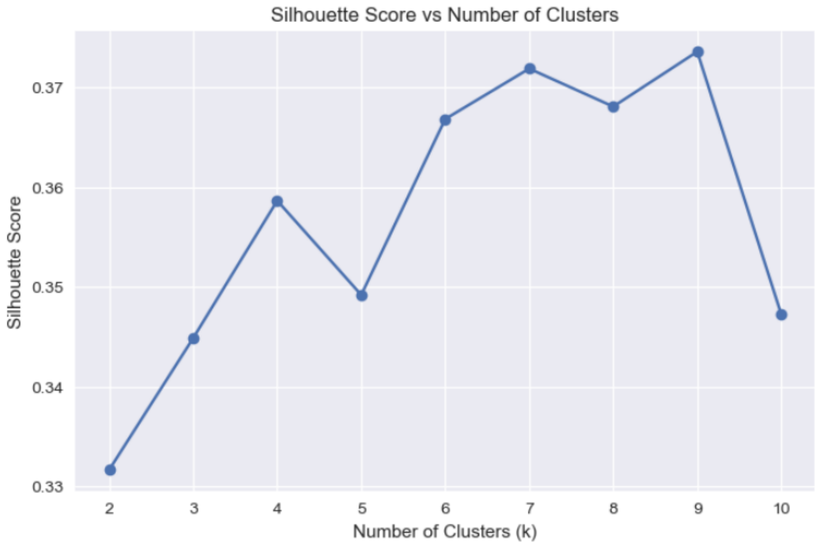
## 2. ‘Explained variance by Component’ and ‘Data Distribution in PCA Space’ after Dimensionality Reduction with PCA



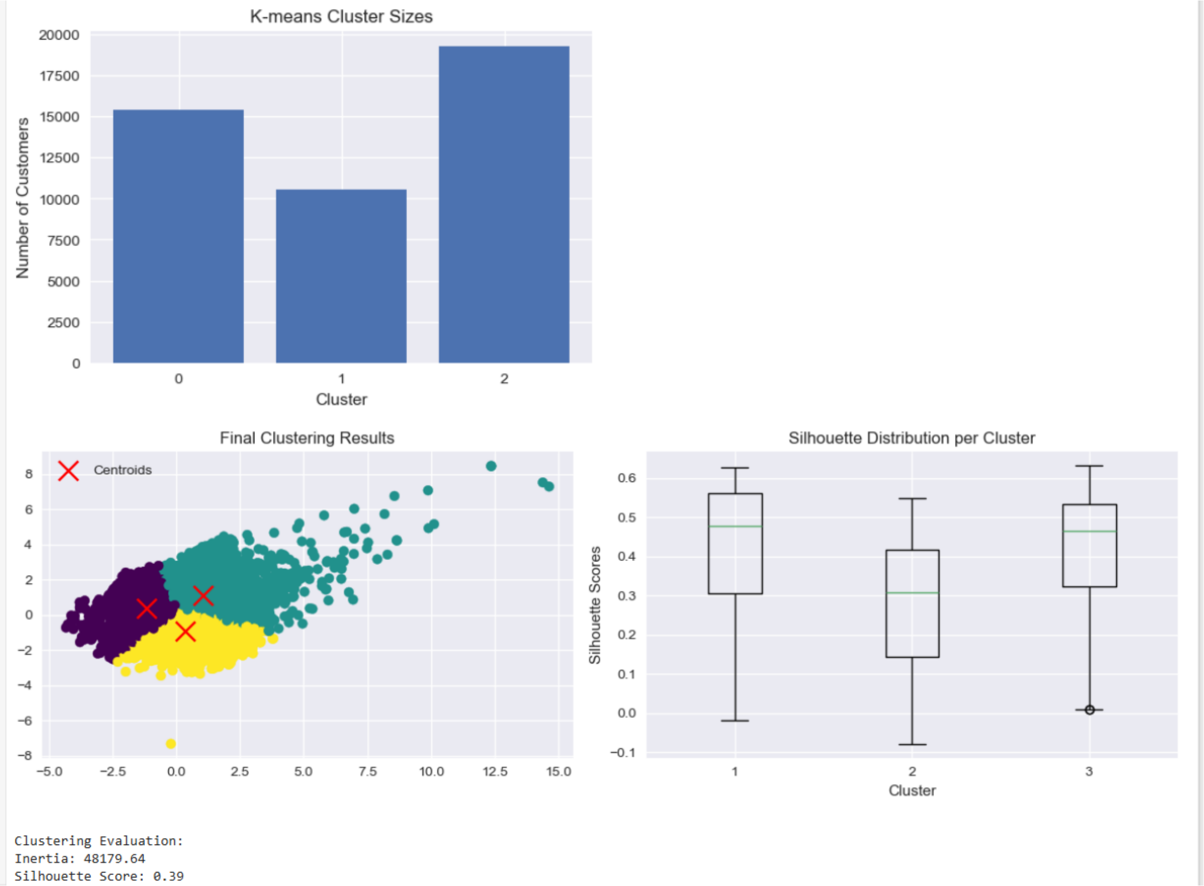
```
Shape after PCA: (45211, 2)
```

## 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
```