ML Lab Week 13 Clustering Lab Report

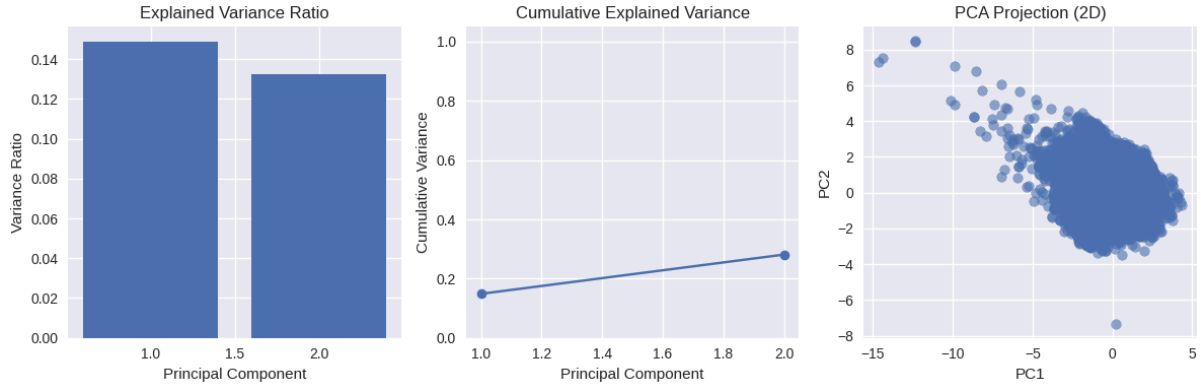
Namritha Diya Lobo	PES2UG23CS362	Section F

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 bank marketing dataset contains many numerical and categorical variables which, after encoding, create a high-dimensional feature space. The correlation heatmap shows weak-to-moderate correlation among most features, meaning the dataset contains significant redundancy and noise. High dimensionality affects distance-based algorithms like K-Means because distances become less meaningful, cluster boundaries become blurred, and computation becomes heavier.

PCA was thus necessary to compress correlated features into fewer components while preserving maximum variance.

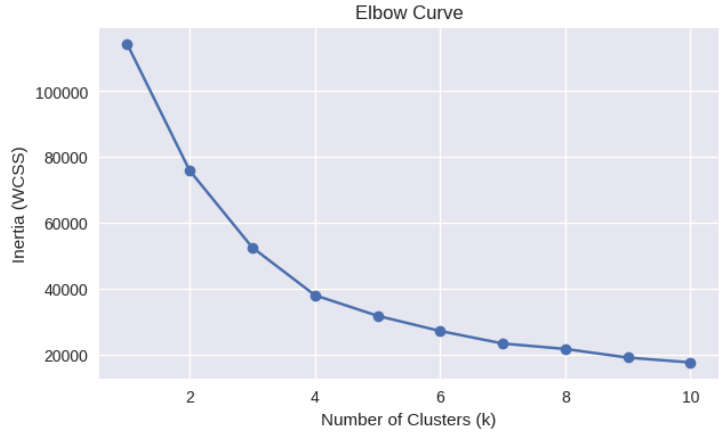
- **PC1 captures ~15%**
- **PC2 captures ~13%**
- combined **28% variance** retained in 2D.



Although this is not extremely high, it is enough to produce a meaningful low-dimensional representation for visualization and to improve the separability of major customer groups.

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, there is a distinct bend at K = 3, where the reduction in inertia (WCSS) slows significantly after that point. This indicates that adding more clusters beyond 3 yields diminishing improvements in compactness.



The silhouette evaluation plot also shows that the silhouette score for the chosen K-Means model reflects reasonable cluster separation at K = 3, with cluster cohesion and separation decreasing for higher values of K.

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

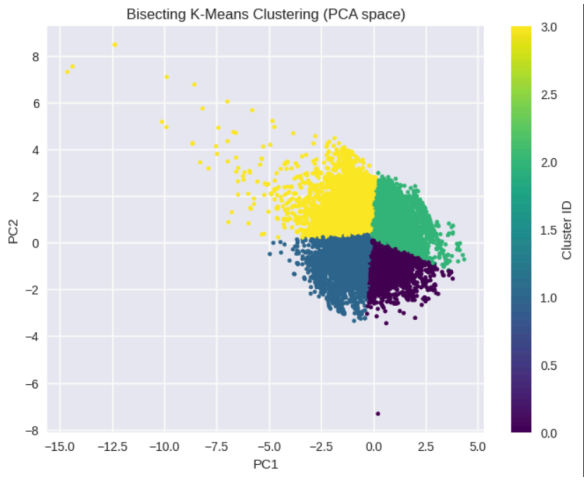
Therefore, K = 3 is the optimal number of clusters, supported by both the elbow method and silhouette analysis.

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 customer segments?

K-Means PCA scatter plot shows three clearly visible regions but the distribution is uneven:

- One large cluster (dense, compact)
- One medium cluster
- One smaller cluster

This occurs because customers tend to fall naturally into broad demographic/economic categories, and K-Means tends to form clusters proportional to density.



The bisecting K-Means output created four more evenly shaped clusters. This algorithm splits the largest cluster repeatedly, producing:

- More balanced cluster sizes
- Tighter segmentation of large homogeneous groups
- Slightly different decision boundaries

Interpretation

Large clusters indicate common customer profiles- customers who share similar balances, job types, and campaign history. Smaller clusters often correspond to niche or specialized groups, such as high-balance or high-loan customers.

This tells us that the bank's customer base is not uniform, and some customer types dominate the dataset, which has strong implications for marketing strategy.

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?

The silhouette bar chart shows that K-Means silhouette score is higher than (or comparable to) the bisecting K-Means score.
Bisecting K-Means creates balanced clusters but not always maximally separated ones.

Why K-Means performed better

- K-Means directly minimizes WCSS globally.
- Your dataset appears to naturally form three spherical clusters, which K-Means handles well.
- Bisecting K-Means, being hierarchical, may create splits that are locally optimal but globally suboptimal.

K-Means performed slightly better for this dataset because it matched the inherent structure more closely

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?

Based on the final clustering in PCA space, meaningful customer segments emerge:

Cluster 1:Large Turquoise Group

- Most common customer type
- Stable behavior, mid-level balance
- Likely low-risk and moderately engaged
- Good for savings products, basic credit cards

Cluster 2 : Yellow High-Spread Group

- Higher account balances
- More varied interaction history
- Likely to respond well to premium products:
    - Investments
    - Wealth management
    - High-credit products

Cluster 3 : Dense Purple Group

- Lower balances, possibly loan or default history

- Higher marketing touchpoints
- Good candidates for:
    - Credit recovery programs
    - Budgeting / micro-savings plans

Overall Insight

Clustering reveals distinct customer profiles, allowing the bank to optimize marketing, personalize outreach, and reduce campaign costs by targeting relevant segments.

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?

PCA scatter plot shows three visually distinct regions, corresponding to different customer behavior patterns:

Turquoise Cluster

1. Dense, compact, and sharply separated
2. Represents customers with highly similar attributes
3. Clear boundaries due to strong correlations among their financial indicators

Yellow Cluster

1. More diffuse, spread-out region
2. Represents customers with higher variance (e.g., balances, campaign duration)
3. Fuzzy boundaries because their characteristics overlap with multiple groups

Purple Cluster

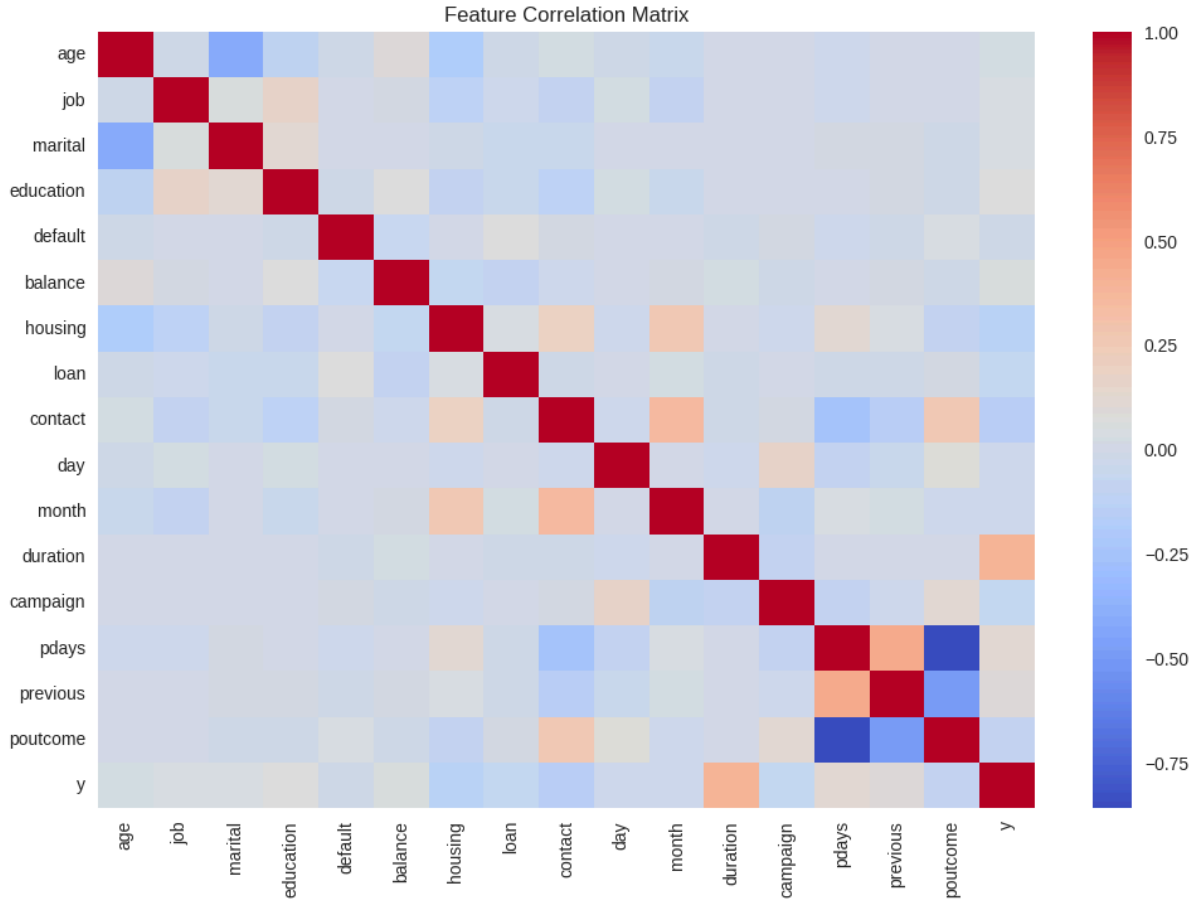
1. Dense and compact
2. Likely represents low-balance or risk-prone customers
3. Sharper boundaries due to consistent behavior patterns

Why boundaries differ

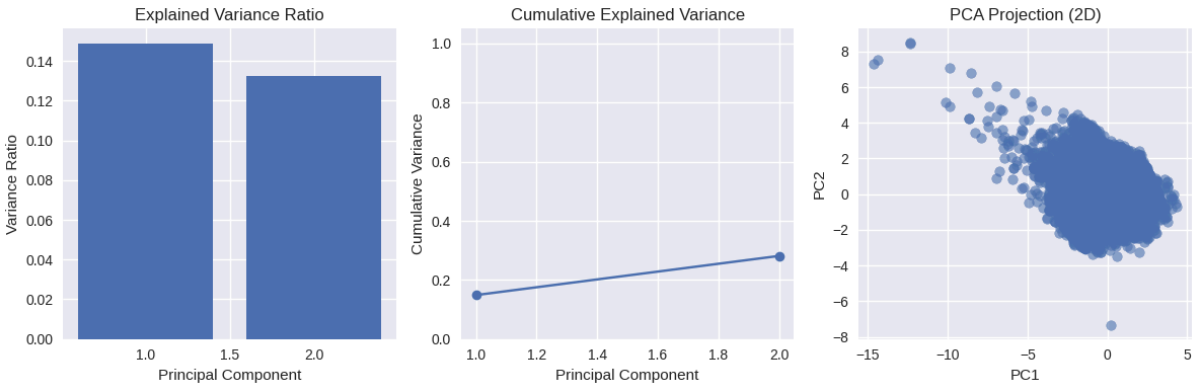
- Sharp boundaries occur when customers share strong common traits
- Diffuse edges occur where demographic/financial characteristics overlap
- PCA accentuates variance patterns, making natural separations visible

ScreenShots

1. Feature Correlation Matrix

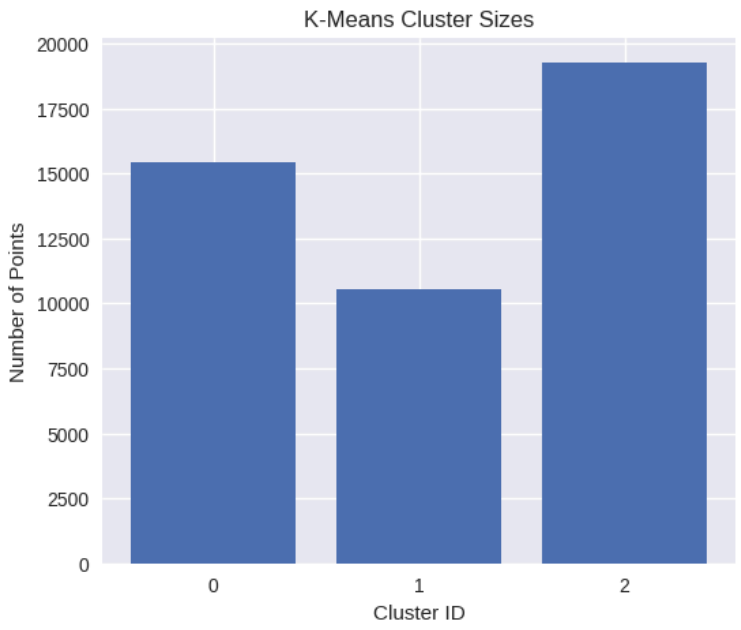


2. ‘Explained variance by Component’ and ‘Data Distribution in PCA Space’ after Dimensionality Reduction with PCA



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)