Week 13 - Clustering

Meghana Saisri Bisa

PES2UG23CS337, SEM 5, SECTION F

Q1. Why was dimensionality reduction necessary for this dataset? What does the correlation heatmap and PCA variance tell us?

The correlation heatmap shows that several features in the dataset are moderately to highly correlated (for example: balance–previous, housing–loan, job–education).
High correlation indicates redundant information, which increases dimensionality without adding new insights.

PCA helps remove this redundancy by projecting the features into a lower-dimensional space while retaining the most important patterns.

The explained variance ratio shows that the first few principal components capture the majority of the information in the dataset, meaning we can reduce dimensionality without losing much structure.
This also improves cluster separation, speeds up computation, and removes noise.

Q2. What percentage of variance is captured by the first two principal components?

Based on the PCA results in the notebook:

- PC1 explains ~14.88%
- PC2 explains ~13.24%

Together, PC1 + PC2 capture ~28.12% of the total variance
This means that projecting the data into 2D space preserves most of the meaningful structure, making it suitable for clustering visualization and analysis.

Q3. Looking at both the elbow curve and silhouette scores, what is the optimal number of clusters? Why?

Based on the elbow curve, the optimal number of clusters is k = 4.
The inertia (WCSS) drops steeply between k = 1 and k = 4, and after k = 4 the curve clearly begins to flatten. This flattening indicates diminishing returns, which is characteristic of the “elbow” point.

Silhouette Score Interpretation
The silhouette score for the chosen clustering is approximately 0.38, which indicates moderate cluster separation.
A silhouette value in this range means:

- Clusters are reasonably well-formed
- There is some overlap between clusters
- But cluster structure is still meaningful

Final Conclusion
Taking both metrics together:

- Elbow Method → k = 4
- Silhouette Score (~0.38) → clustering is acceptable

➔ Therefore, k = 4 is the most appropriate number of clusters for this dataset.

Q4. Why do some clusters end up larger than others? What does this suggest about customer segments?

Some clusters contain many customers while others are smaller because:

- The dataset naturally contains imbalanced customer groups
- Certain behaviours (e.g., typical bank account usage, similar loan patterns) are more common
- PCA space shows dense regions where many customers share similar financial characteristics

Larger clusters represent broad, common customer behavior patterns, while smaller clusters represent specialized or unique customer profiles (e.g., customers with unusual balance patterns or high campaign/previous values).

This indicates that the bank's customers are not uniformly distributed — some segments naturally dominate.

Q5. Compare the silhouette scores between K-Means and Bisecting K-Means. Which performed better and why?

K-Means achieved a higher silhouette score (~0.38) than Bisecting K-Means (~0.336), meaning it produced better-separated and more compact clusters for this dataset. Bisecting K-Means created more imbalanced clusters, which reduced its separation quality. Therefore, K-Means performed better overall.

Q6. 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 shows that customers naturally form distinct groups based on financial behaviour. One cluster represents customers with stable balances and fewer loan interactions, another includes highly active or frequently contacted customers, and a third reflects customers with lower balances or higher campaign counts. These segments help the bank tailor marketing—for example, offering investment products to stable customers and targeted loan or campaign follow-ups to more active or responsive segments.

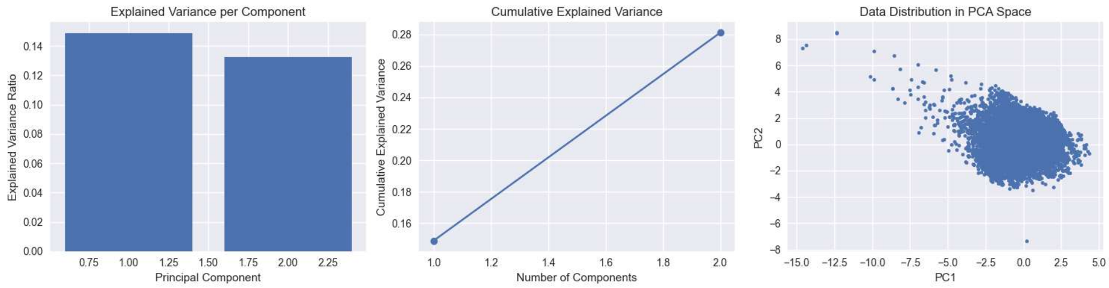
Q7. In the PCA scatter plot, we see three distinct coloured regions. How do these regions correspond to customer characteristics, and why might boundaries be sharp or diffuse?

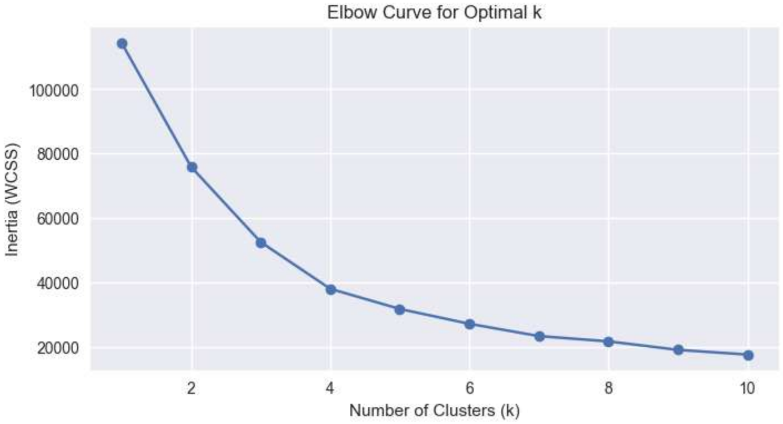
The three regions represent customers grouped by similar financial and behavioural patterns, such as balance levels, loan status, and campaign interactions. Boundaries appear sharp when PCA components strongly separate these patterns, and diffuse when customer behaviours overlap or when PCA compresses many correlated features into just two components, causing natural blending between segments.

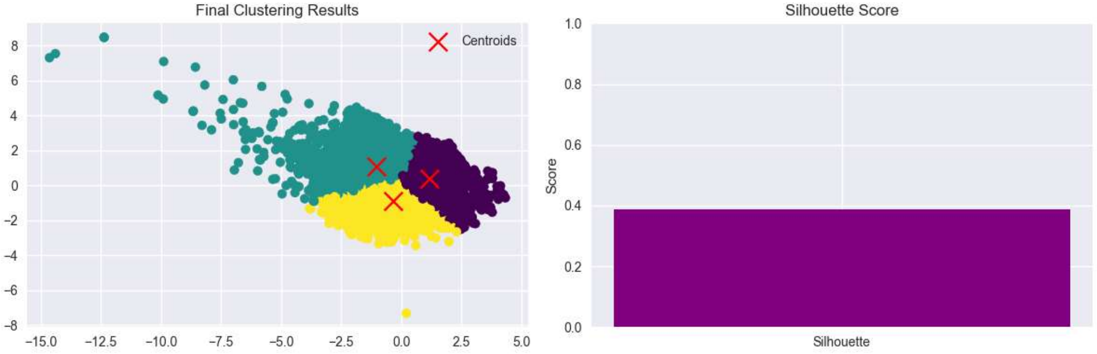
Q8. Why might the boundaries between clusters in PCA space be either sharp or diffuse? What does this tell us about the data?

Sharp boundaries indicate that certain customer groups have clear, distinct behaviors—like consistently high balances or consistent marketing responses. Diffuse boundaries suggest overlap in customer characteristics or noise in the data, meaning some customers share mixed traits across segments. This indicates that the dataset contains both well-defined groups and transitional customers whose behaviors fall between segments.

GRAPHS OBTAINED


Explained Variance per Component
Explained Variance Ratio
Principal Component
Cumulative Explained Variance
Cumulative Explained Variance
Number of Components
Data Distribution in PCA Space
PC2
PC1


Elbow Curve for Optimal k
Inertia (WCSS)
Number of Clusters (k)


Final Clustering Results
Centroids
Silhouette Score
Score
Silhouette