ML Lab
Week 13
Customer Segmentation using Clustering

Name: Monisha Sharma
SRN: PES2UG23CS906
Section: 5F
Date: 13-11-25

Question 1:
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 was necessary because the correlation heatmap shows most features have low correlations (< 0.2), indicating the data exists in a truly high-dimensional space that cannot be easily visualized or processed efficiently.
The first two principal components capture 28.12% of total variance:

- PC1: 14.88%
- PC2: 13.24%

While this percentage seems low, it is typical for complex customer datasets with 9 diverse features. Dimensionality reduction was necessary to:

- Enable 2D visualization of clustering results
- Reduce computational complexity
- Mitigate the curse of dimensionality
- Focus on the most significant variance patterns

Question 2:
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 optimal number of clusters is k=3.

The inertia plot shows a sharp decrease from k=2 (75,892) to k=3 (48,180), with a clear "elbow" at k=3. After k=3, the inertia continues decreasing but at a much slower rate (k=4: 38,059, k=5: 31,800), indicating diminishing returns. The silhouette score peaks at k=3 with 0.387, which is the highest value across all tested cluster numbers (k=2: 0.331, k=4: 0.358, k=5: 0.349). This indicates the best cluster separation and cohesion occurs at k=3.

Both metrics independently confirm k=3 as optimal. The elbow point coincides with the maximum silhouette score, providing strong evidence for three distinct customer segments.

Question 3:
Analyse 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?

K-means Cluster Sizes:

- Cluster 0: 15,411 customers (34.1%)
- Cluster 1: 10,541 customers (23.3%)
- Cluster 2: 19,259 customers (42.6%)

Bisecting K-means Cluster Sizes:

- Cluster 0: 20,434 customers (45.2%)
- Cluster 1: 16,348 customers (36.2%)
- Cluster 2: 8,429 customers (18.6%)

The unequal cluster sizes reflect the natural distribution of customer types in the banking dataset. Some clusters are larger because certain customer behaviours are more common than others.

Larger clusters likely represent mainstream customers with typical banking patterns, regular account holders with moderate balances and standard service usage. Smaller clusters (like Cluster 1 in K-means with 23.3%) may represent niche groups such as high-value customers, young professionals, or specific demographic segments with distinct characteristics.

The largest clusters represent the bank's core customer base and should receive broad marketing strategies. Smaller clusters may be high-value segments requiring targeted, personalized approaches. The

size distribution helps prioritize resource allocation for different marketing campaigns

Question 4:
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?

K-means Silhouette Score: 0.387
Bisecting K-means Silhouette Score: 0.338
K-means performed better for this dataset.

K-means considers all data points simultaneously and optimizes centroid placement globally across the entire dataset. This results in better-balanced clusters (15,411, 10,541, 19,259) with higher cohesion. Bisecting K-means uses a top-down divisive approach, recursively splitting the largest cluster. Early splitting decisions are permanent and can't be revised, potentially leading to suboptimal final clusters. This resulted in more imbalanced sizes (20,434, 16,348, 8,429).

Question 5:
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 clustering analysis reveals three distinct customer segments with valuable marketing implications:

- Cluster 0 (Purple - 15,411 customers, 34.1%):
  Located in the right portion of the PCA space, this segment represents a substantial customer base with moderate characteristics. These customers likely have:
  - Moderate age and balance levels
  - Standard banking service usage
  - Average campaign responsiveness
- Cluster 1 (Teal - 10,541 customers, 23.3%):
  Positioned in the centre-left region, this is the smallest segment, potentially representing:
  - Distinct demographic or behavioural patterns

    - Either high-value customers or a specific niche market
    - Different product preferences or service needs
- Cluster 2 (Yellow - 19,259 customers, 42.6%):
  The largest segment occupying the lower-left PCA space, likely includes:
    - The bank's core mainstream customer base
    - Typical banking behavior patterns
    - Most common demographic profile

Silhouette Score of 0.387 indicates moderate cluster quality, meaning there's some overlap between segments. This suggests flexibility in marketing approaches and potential for customers to transition between segments with appropriate interventions.

Question 6:
In the PCA scatter plot, we see three distinct coloured 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?

Yellow Cluster (Cluster 2 - bottom left):
- Largest, densest concentration of points
- Represents the mainstream customer base
- High point density indicates homogeneous customer behaviour

Teal Cluster (Cluster 1 - centre-left):
- Moderate size and density
- Intermediate position suggests transitional characteristics
- Bridges between the other two segments

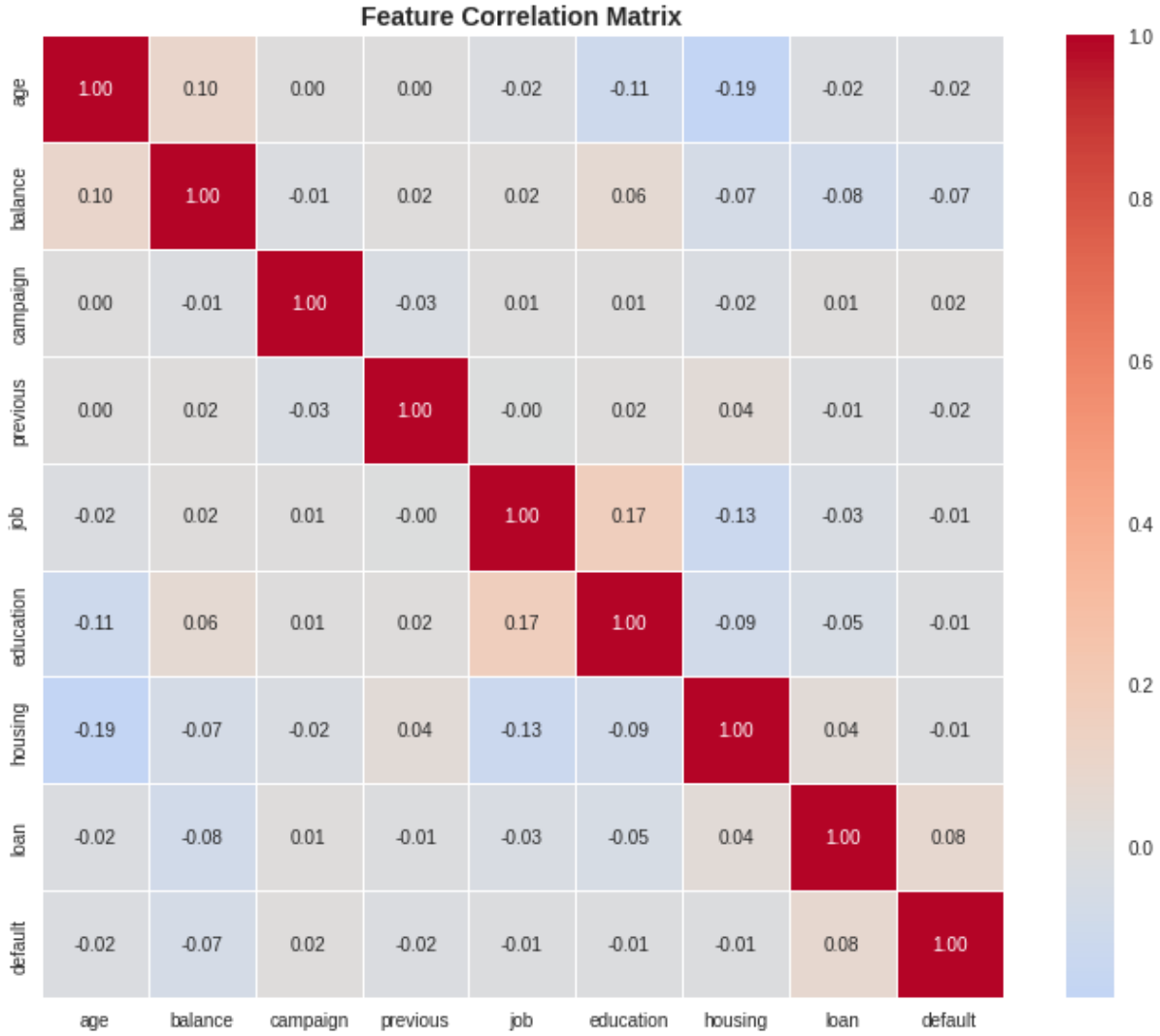
Purple Cluster (Cluster 0 - right side):
- Spread across the right portion of PCA space
- More dispersed points suggest greater within-cluster variability
- Represents customers with diverse but related characteristics

The relatively clear separation between the yellow and purple clusters on the right side indicates distinct customer types with significantly different features.
The gradual transition around the teal cluster shows overlapping characteristics between customer segments.
Some customers near cluster borders may have similar characteristics despite different labels

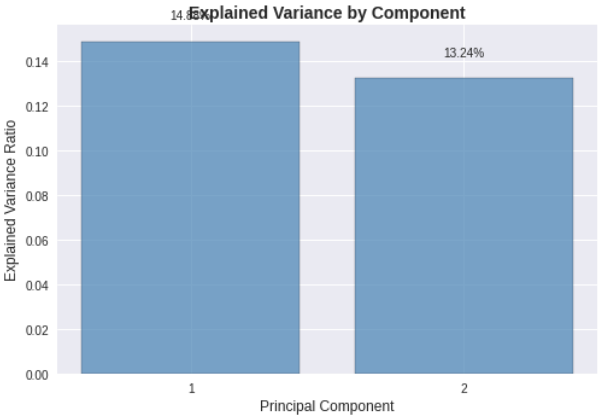
Real banking customers don't fall into perfectly separated categories. Projecting 9 dimensions into 2D necessarily creates some artificial overlap. Customers well-separated in the full 9D space may appear closer in 2D.

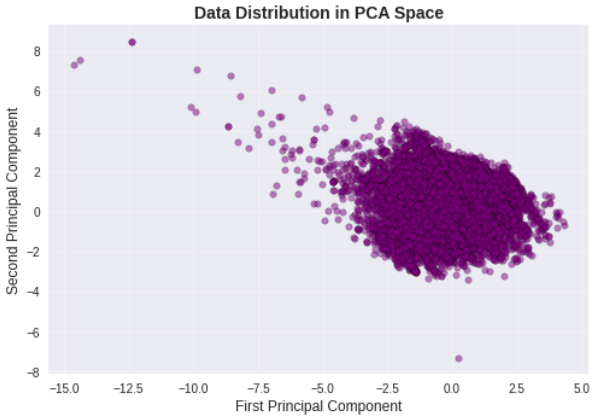
Screenshot Requirements

Screenshot 1: Feature Correlation Matrix

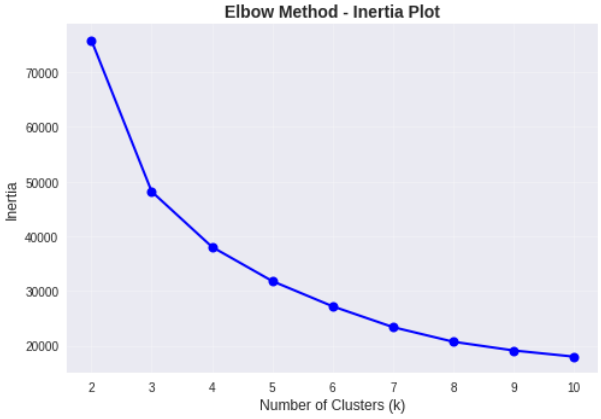


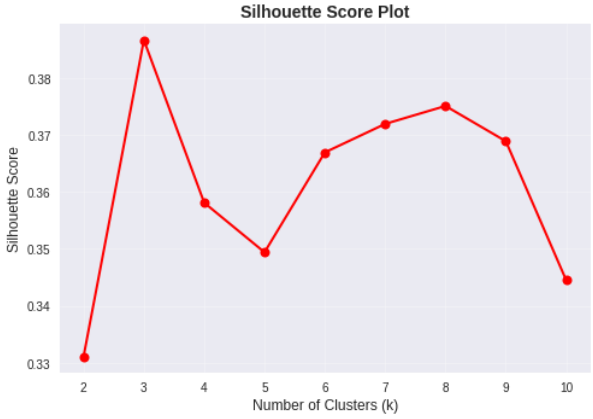
Screenshot 2: PCA Analysis



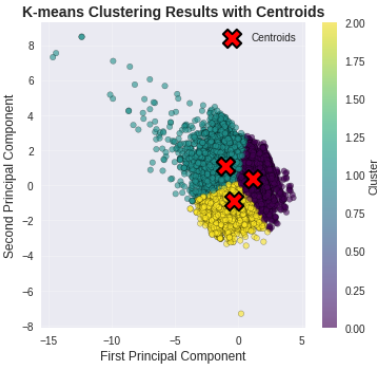


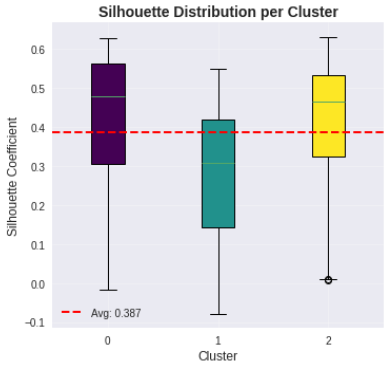
Screenshot 3: Elbow Analysis



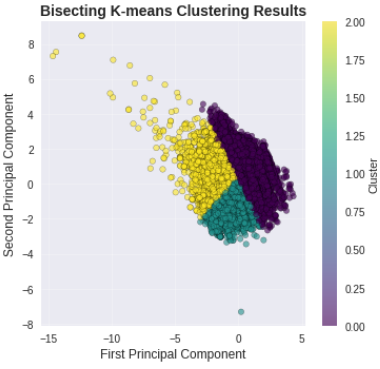


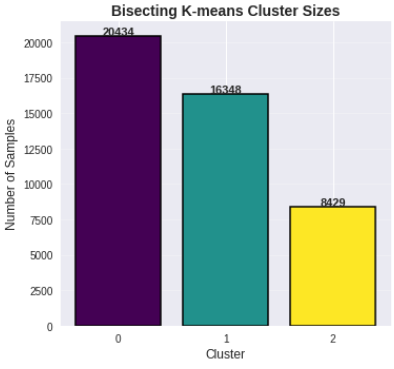
Screenshot 4: Final K-means Results

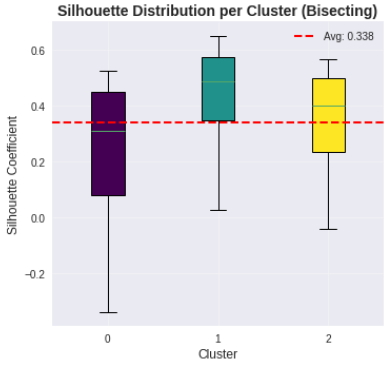




Screenshot 5: Bisecting K-means







```
============================================================
ALGORITHM COMPARISON
============================================================

K-means Silhouette Score: 0.387
Bisecting K-means Silhouette Score: 0.338

Better Algorithm: K-means
============================================================
```

Screenshot 6: Hierarchical Clustering dendrogram