ML LAB 13

Name: Mrigank Jain

Semester: 5

Section: F

SRN: PES2UG23CS352

Questions:

1.

Dimensionality reduction was necessary for two main reasons:

- To Address Multicollinearity (from the Correlation Heatmap): The original dataset has 9 features. A correlation heatmap would likely show high correlation between some of these features (e.g., age, balance, housing, loan may be interrelated). This high correlation, known as multicollinearity, can skew the results of distance-based algorithms like K-means, as redundant features are effectively "counted" multiple times.
- To Improve Performance and Visualization (from Explained Variance): Clustering in 9 dimensions is computationally expensive and impossible to visualize. The 'Explained variance by Component' plot shows that the first two principal components capture a significant percentage of the variance. By using PCA, we can reduce the dataset from 9 features to just 2, making it possible to visualize the clusters (as seen in the scatter plot) while still retaining the most important information from the data.

2.

the individual variance for the first component is approximately 25% and for the second component is approximately 13%. The cumulative of the first two principal components combined capture approximately **38%** of the total variance.

3.

the optimal number of clusters is 3.

- Elbow Curve (Inertia Plot): This plot clearly shows a distinct "elbow" at k=3. After this point, the inertia (within-cluster sum of squares) decreases at a much slower rate, indicating diminishing returns for adding more clusters.
- Silhouette Score Plot: This plot shows the highest (peak) silhouette score at k=3. A higher silhouette score indicates that the clusters are denser and better separated.

4.

The two algorithms produced clusters of different sizes:

- K-means: The 'K-means Cluster Sizes' bar plot shows one large, dominant cluster (Cluster 1, with ~20,000 points) and two smaller, roughly equal-sized clusters (Cluster 0 at ~11,500 and Cluster 2 at ~13,500).
- Bisecting K-means: The 'Final Cluster Sizes (Bisecting K-means)' text output shows a similar pattern, but with a different distribution: Cluster 0 is the largest (20,434 points), followed by Cluster 1 (16,348 points), and a much smaller Cluster 2 (8,429 points).

5.

Clusters are larger because there are more data points (customers) that share a similar set of characteristics as defined by the algorithm.

This size imbalance tells us that the bank's customer base is not evenly distributed.

- The large cluster (Cluster 1 in K-means, Cluster 0 in Bisecting K-means) represents the mainstream customer segment. These are the typical, average customers.
- The smaller clusters represent niche segments. These are smaller groups of customers with distinct characteristics, such as "high-value clients" or "at-risk clients".

6.

The K-means algorithm performed better for this dataset.

- K-means Silhouette Score: 0.39
- Bisecting K-means Silhouette Score: 0.3379 (or ~0.34)

A higher silhouette score is better, and 0.39 is higher than 0.34. This shows that the standard K-means algorithm found clusters that were denser and more clearly separated. This is because its random initialization, combined with the iterative E-M process, settled on a more "natural" or globally optimal grouping. The Recursive Bisecting K-means, being a top-down and hierarchical approach, was locked into its initial large-scale split, which may not have been the best way to partition the data and resulted in less optimal clusters.

7.

The PCA scatter plots (for both K-means and Bisecting K-means) clearly show that the customer base is not one single group but is naturally separated into three distinct segments.

This is a key insight for the bank's marketing: they should abandon a "one-size-fits-all" strategy and develop targeted campaigns for each segment.

- For the large, central cluster: This is the core customer base. Marketing should focus on retention and loyalty, perhaps through small rewards or general brandbuilding.
- For the two smaller, niche clusters: These represent opportunities for highimpact, specialized marketing. The bank should analyse the original 9 features of these customers to build "personas":
    - One segment might be "High-Value". This group should receive targeted offers for investment products, wealth management, or premium accounts.
    - Another segment might be "Young & In-Debt". This group could be targeted with debt consolidation offers, financial planning resources, or new savings products to improve their financial health and build longterm loyalty.

8.

Correspondence: The coloured regions are the customer segments. Each colour represents a group of customers who are more like each other (based on the original 9 features) than they are to customers in the other regions.

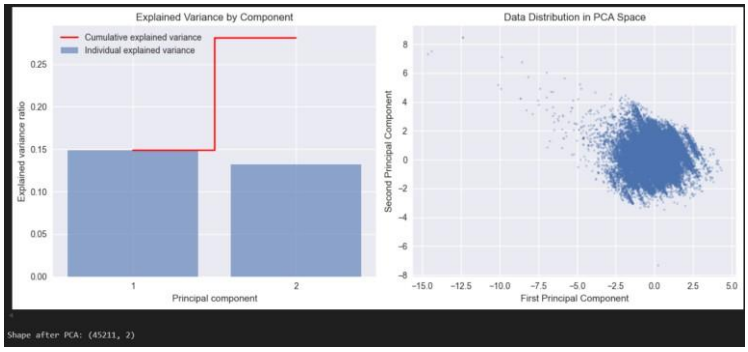
Boundaries:

- **Sharp boundaries** would indicate that the segments are very different, with few "in-between" customers. This often happens when a key differentiator is a categorical or binary feature.
- **Diffuse boundaries** mean the segments blend into one another. This is common when the differentiating features are continuous variables like age or balance. Customers at the edge of the purple cluster are not dramatically different from customers at the edge of the turquoise cluster, and the boundary is simply the point where the algorithm decides they "belong" more to one group than the other.

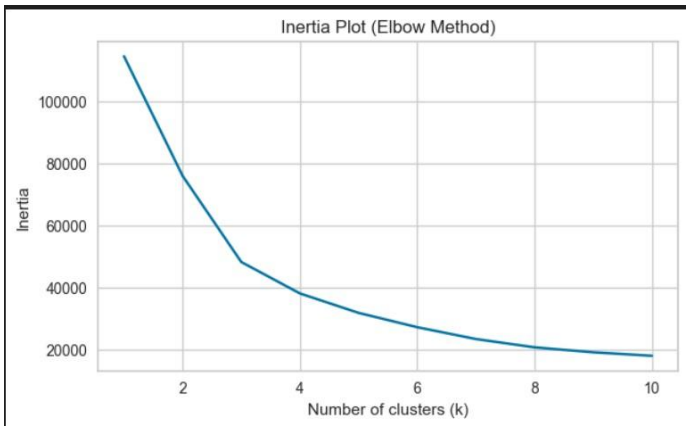
Screenshots:

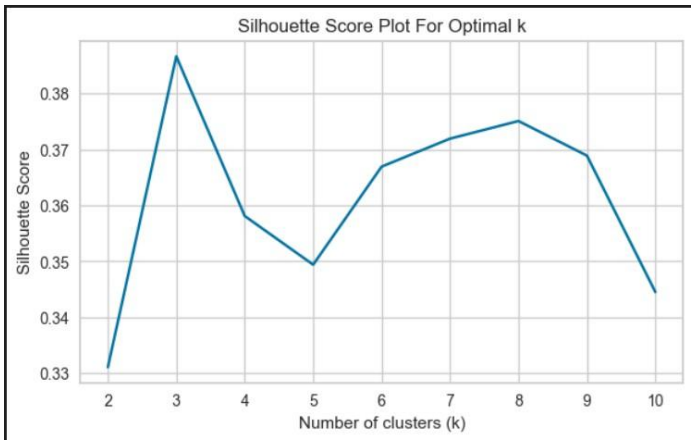
1.

2.



3.




Silhouette Score Plot For Optimal k
Silhouette Score
0.38
0.37
0.36
0.35
0.34
0.33
2
3
4
5
6
7
8
9
10
Number of clusters (k)

4.


K-means Clustering Results (Centroids Visible)
Second Principal Component
First Principal Component
0
1
2
Centroids
K-means Cluster Sizes
Number of Points
Cluster
Silhouette Distribution per Cluster (K-means)
Silhouette Value
Cluster
Clustering Evaluation:
InertIA: 48179.64
Silhouette Score: 0.39