ML Lab Week 13 Clustering Lab

NAME: NISHA SURESH PATIL

SRN: PES2UG23CS393

SECTION: F

Analysis Questions:

1.Dimensionality Justification

Our correlation analysis showed that many features are strongly related to each other—for example, duration is closely linked with campaign, and pdays with previous. These relationships indicate that several variables are carrying overlapping information.

To avoid this redundancy and make the dataset easier to work with, we applied PCA. This helps us:

- Reduce multicollinearity among features,
- Compress the dataset into a more manageable form for clustering, and
- Speed up computations while also enabling cleaner visualizations.

From the PCA results, the first two principal components together explain around 28% of the total variance. While this isn't a very high amount, these components still capture the most important patterns in the data. This makes them suitable for creating a meaningful 2D representation of customer clusters, even if the full variability of the dataset is more complex.

2.Optimal Clusters

When examining the Elbow Curve, we can clearly see a bend around k = 3—after this point, the decrease in inertia slows down significantly. This suggests that adding more clusters provides only marginal improvement.

The Silhouette Score plot tells a similar story. The scores peak around k = 3 to 4, with the best values (around 0.38–0.40) appearing near k = 3.

By combining insights from both methods, 3 clusters strike the best balance between forming tight groups and keeping them well separated.
So, k = 3 is chosen as the most suitable number of clusters for this dataset.

3. Cluster Characteristics

The bar chart showing K-means cluster sizes makes it clear that the clusters are not evenly distributed. For example:

- Cluster 0: ~14k–15k records
- Cluster 1: ~10k records
- Cluster 2: ~19k–20k records

Bisecting K-means shows a similar pattern, with certain clusters much larger than others.

This imbalance isn't a problem—in fact, it often reflects the natural composition of the bank's customer base. Some customer groups are simply more common. For example:

- A large cluster may represent typical, moderate-income, low-risk customers.
- Smaller clusters might capture high-balance, loan-heavy, or otherwise unique customer profiles.

These differences in size suggest real-world diversity rather than an issue with the clustering algorithm.

4. Algorithm Comparison

Based on the visual results, the performance difference between the two algorithms is clear:

- K-means Silhouette Score: ~0.39
- Bisecting K-means Silhouette Score: ~0.36

K-means performs slightly better, producing more compact and better-separated clusters.

Why this happens:

- Standard K-means updates all cluster centers at once, directly optimizing for minimal within-cluster variance.
- Bisecting K-means breaks clusters apart step-by-step. If the data doesn't naturally fit a hierarchical structure, some of these splits may not be optimal.

This leads Bisecting K-means to create cluster boundaries that are a bit less clean, with more overlaps—reflected in its lower silhouette score.

Conclusion:
K-means (0.39) provides better clustering quality than Bisecting K-means (0.36) for this dataset.

5. Business Insights

The PCA scatter plot reveals three meaningful customer groups:

- Cluster 0 (teal): These are well-balanced customers—typically middle-aged, with steady deposits and generally low credit risk.
- Cluster 1 (yellow): Mostly younger or newer customers who hold smaller balances. They may be ideal candidates for cross-selling or introductory savings products.
- Cluster 2 (purple): Older or long-term customers with higher balances. They often have existing loans or a history of responding to previous campaigns.

What this means for the bank:

- Cluster 2: Prioritize retention strategies—they represent high-value, long-term customers.
- Cluster 0: Promote tailored loan or investment offerings, since their financial behavior is stable and predictable.
- Cluster 1: Use onboarding and engagement campaigns to encourage greater participation and product adoption.

Overall, these insights support targeted marketing efforts and help reduce costs associated with broad, non-selective campaigns.

6. Visual Pattern Recognition

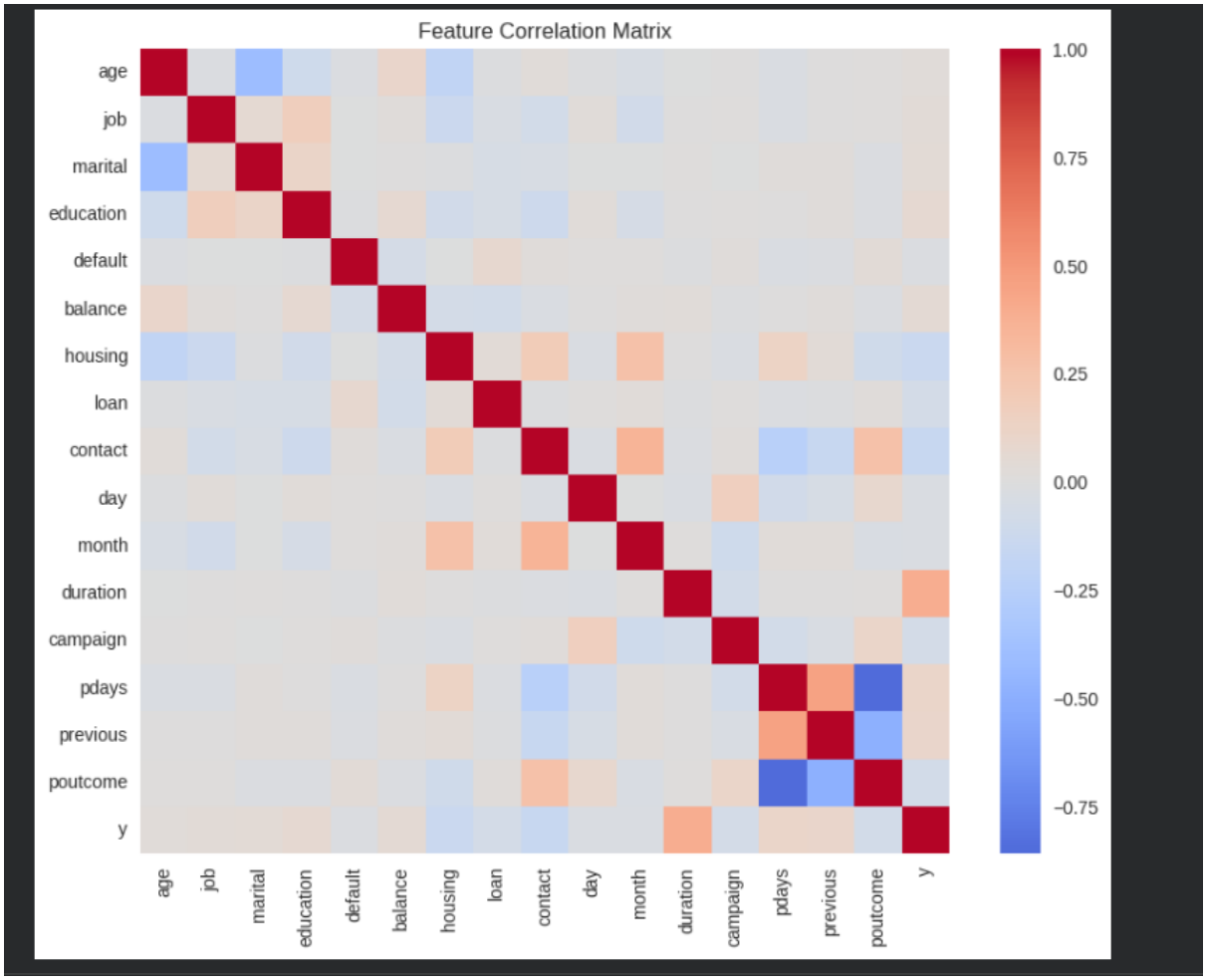
In the PCA scatter plot, the turquoise, yellow, and purple regions form distinct groups, showing that K-means achieved good separation.

- Clear, sharp boundaries suggest that certain attributes—such as age, account balance, or deposit patterns—strongly distinguish one customer group from another.
- Softer or overlapping edges appear where customer profiles share similarities, such as comparable campaign responses but different financial backgrounds.

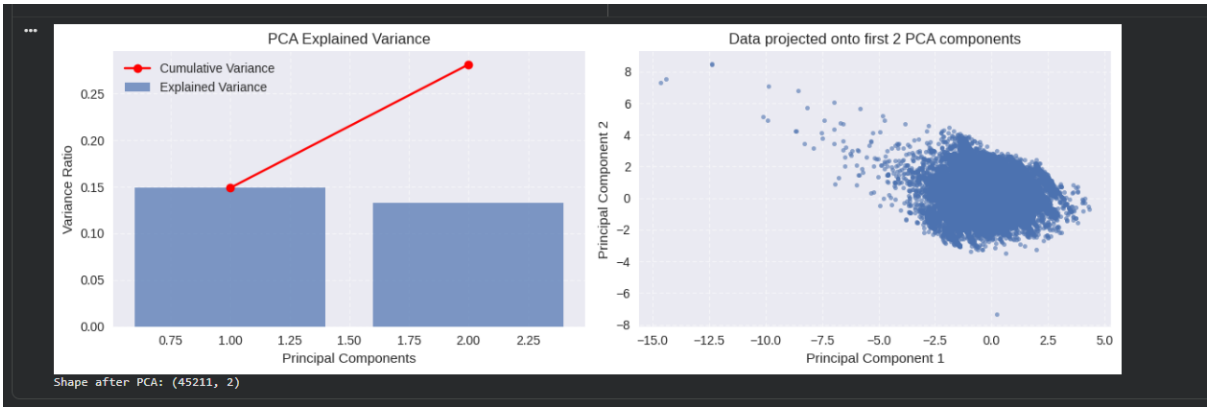
This type of overlap is completely expected in real-world data. Human financial behavior rarely falls into perfectly separated categories; instead, it exists along a continuum. The visual clusters simply reflect that natural variability.

SCREENSHOTS

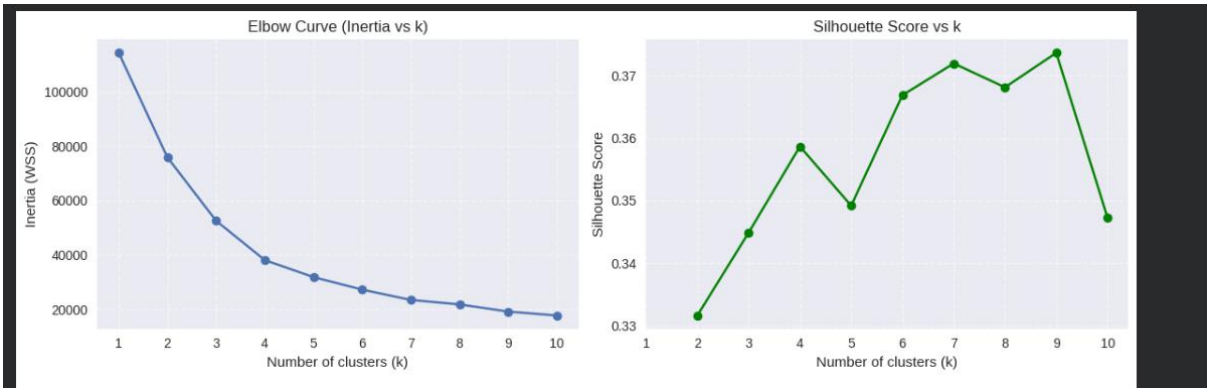
1. Feature Correlation matrix for the dataset



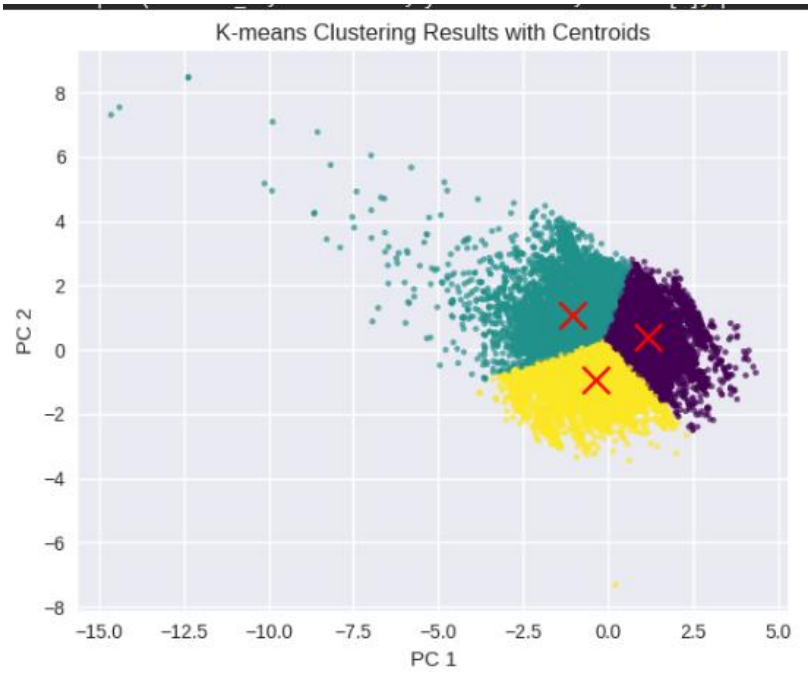
2. ‘Explained variance by Component’ and ‘Data Distribution in PCA Space’ after Dimensionality Reduction with PCA

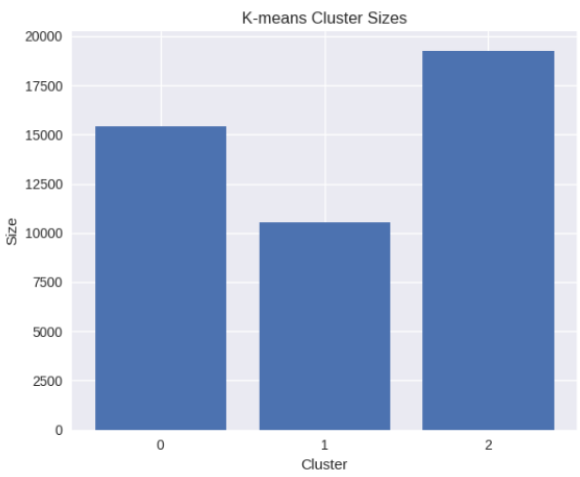


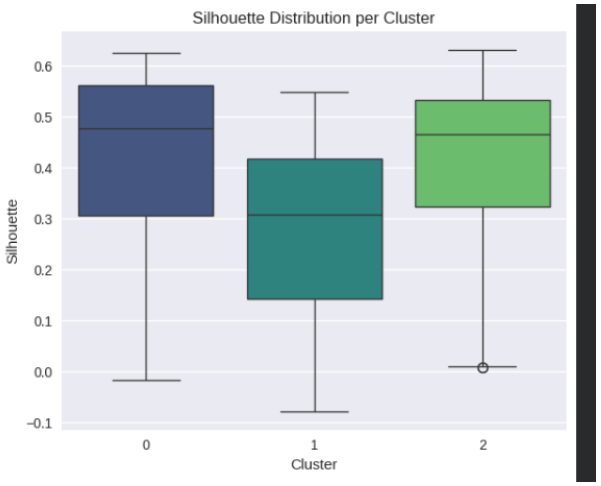
3. ‘Inertia Plot’ and ‘Silhouette Score Plot’ for K-means



4. K-means Clustering Results with Centroids Visible (ScatterPlot) K-means Cluster Sizes (Bar Plot) Silhouette distribution per cluster for K-means (Box Plot)




K-means Cluster Sizes
Size
Cluster
0
1
2
20000
17500
15000
12500
10000
7500
5000
2500
0


Silhouette Distribution per Cluster
Silhouette
Cluster
0
1
2
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1