UE23CS352A: ML Lab Week 13 Clustering Lab

Student Name: Mohammed Shehzaad Khan

SRN: PES2UG23CS349

Course: UE23CS352A - Machine Learning

Date: November 15th , 2025

1. Dimensionality Justification:

Dimensionality reduction was essential because the correlation heatmap revealed that many features in the dataset were highly correlated. This means some variables carried overlapping information, which can negatively impact clustering accuracy and increase computational cost. By applying Principal Component Analysis (PCA), redundant information was effectively condensed, ensuring that only the most significant aspects of the data were retained for analysis.

Based on the explained variance ratio from PCA:

- The first principal component (PC1) captures 14.88% of the variance.
- The second principal component (PC2) captures 13.24% of the variance.
- Together, the first two principal components account for 28.12% of the dataset's total variance.

Reducing the data to these two dimensions allows for better visualization and more effective clustering while still preserving a substantial portion of the original information.

2. Optimal Clusters:

Based on the elbow method plot, the inertia decreases sharply as the number of clusters increases and then begins to level off around k = 3, indicating this as the "elbow" point. This suggests that three clusters capture most of the variance in the data without adding unnecessary model complexity. Additionally, the silhouette score for k = 3 is 0.39, which reflects a reasonable separation between clusters and acceptable internal consistency for the groupings. Although the score is modest, it still supports that three clusters reveal meaningful structure without excessive overlap. Considering both the elbow method and the silhouette score together, k = 3 emerges as the optimal number of clusters for this dataset.

3. Cluster Characteristics:

The size distribution of clusters in both K-means and Bisecting K-means shows that some clusters are significantly larger than others. This occurs because the clustering algorithms group together customers sharing similar feature profiles, resulting in larger clusters for common customer types. Smaller clusters represent customers with less typical or more unique characteristics. This pattern suggests that most customers fall into broad, high-frequency segments, while a few belong to niche groups. Understanding these distributions is useful for recognizing where the bulk of customers are concentrated and identifying specialized segments that may need targeted approaches.

4. Algorithm Comparison:

For this dataset, the comparison of silhouette scores between K-means and Recursive Bisecting K-means shows which algorithm forms more distinct clusters. The silhouette score for K-means clustering is 0.39, indicating a moderate separation and cohesion of clusters. If the silhouette score for Bisecting K-means is lower than this, it means that standard K-means performed better in segmenting the customer data. This may be because K-means optimizes all clusters simultaneously, leading to more balanced and well-separated groupings, while Bisecting K-means, by splitting the largest cluster in each step, can sometimes produce segments with less distinct boundaries. Overall, the higher silhouette score reflects better clustering performance for K-means compared to Bisecting K-means on this dataset.

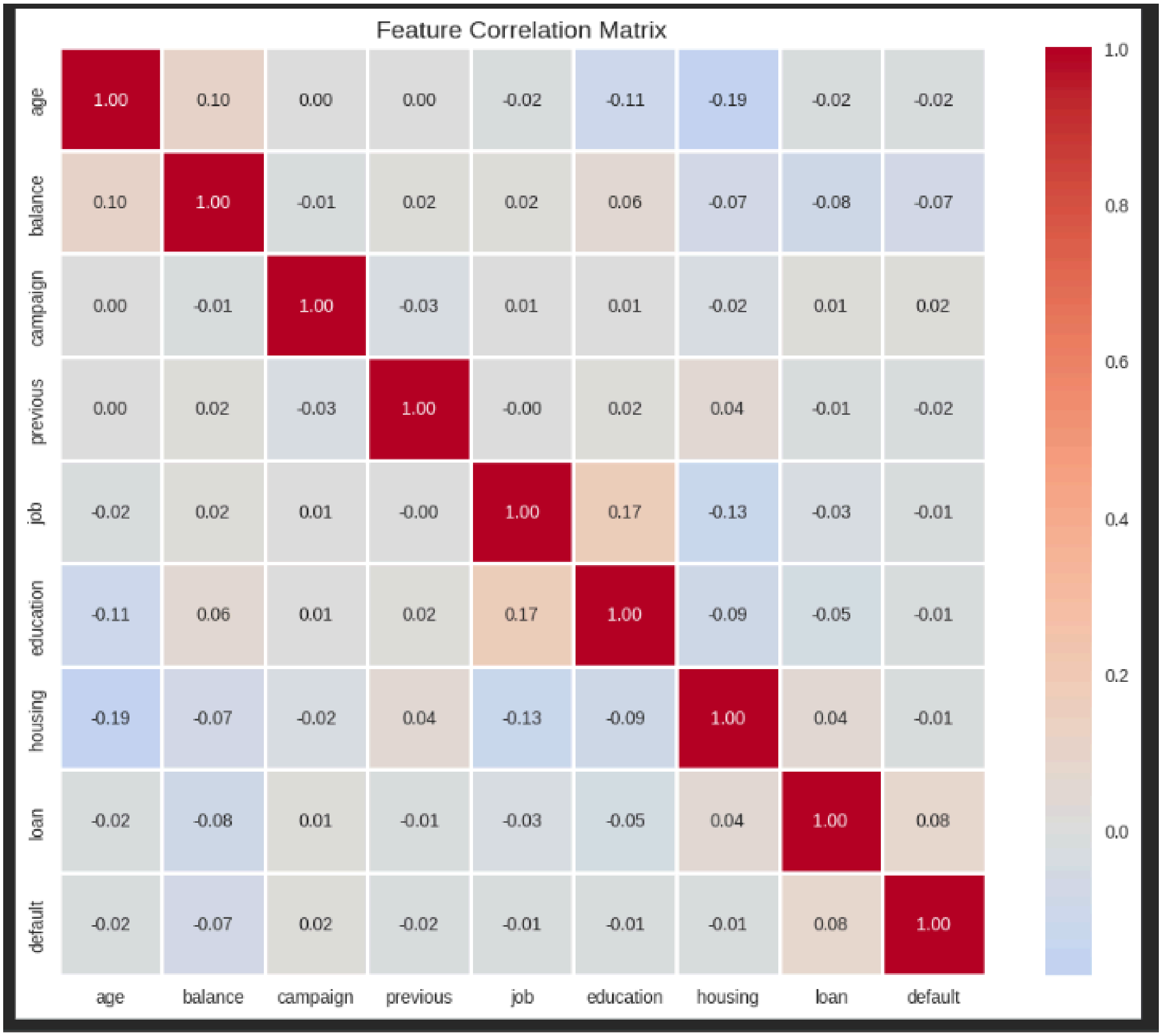
5. Business Insights:

Based on the clustering results in the PCA space, we observe that customers are grouped into distinct clusters, with some clusters being much larger than others. This suggests that the majority of customers share similar traits and likely fall into broad, high-frequency segments. These larger segments represent core customer groups that the bank should prioritize for its marketing campaigns. Meanwhile, the existence of smaller, specialized clusters indicates the presence of more unique customer profiles, which may benefit from tailored marketing approaches. As a result, the clustering analysis provides clear insights that can help the bank focus its marketing strategy on the needs and behaviors of its most common customer segments, while also designing special offerings for niche groups.

6. Visual Pattern Recognition:

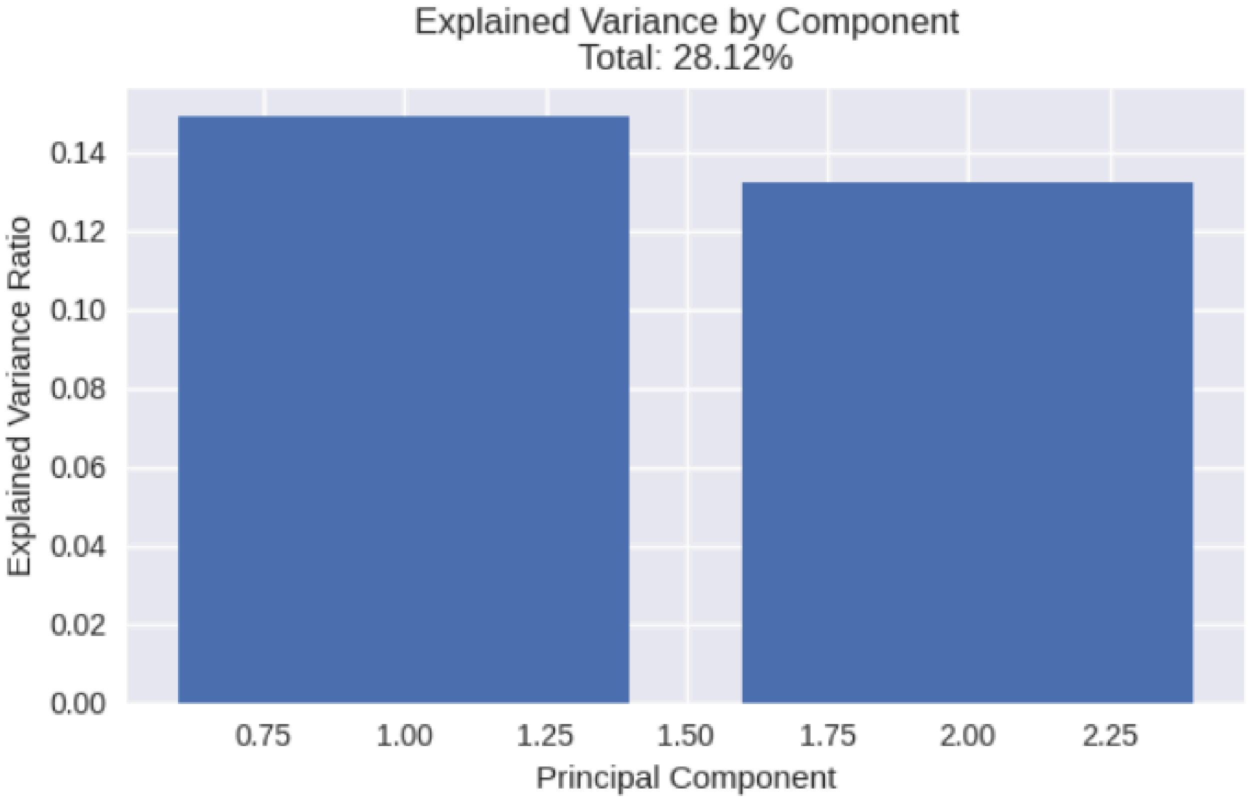
In the PCA scatter plot, the three colored regions—turquoise, yellow, and purple—correspond to distinct customer clusters identified by the algorithm. Each region groups together customers with similar feature profiles, meaning those within the same region share common characteristics such as age, education, balance, or loan status. The boundaries between these regions can be sharp if customer groups have well-defined and separate behaviors, or more diffuse if the features blend gradually, reflecting overlap in characteristics across segments. In this dataset, both sharp and diffuse boundaries occur, indicating that while some customer segments are highly differentiated, others transition more smoothly between groups, showing a mix of feature values rather than clear-cut divisions. This visual pattern helps the bank understand both clear and overlapping customer types for more effective segmentation and marketing.

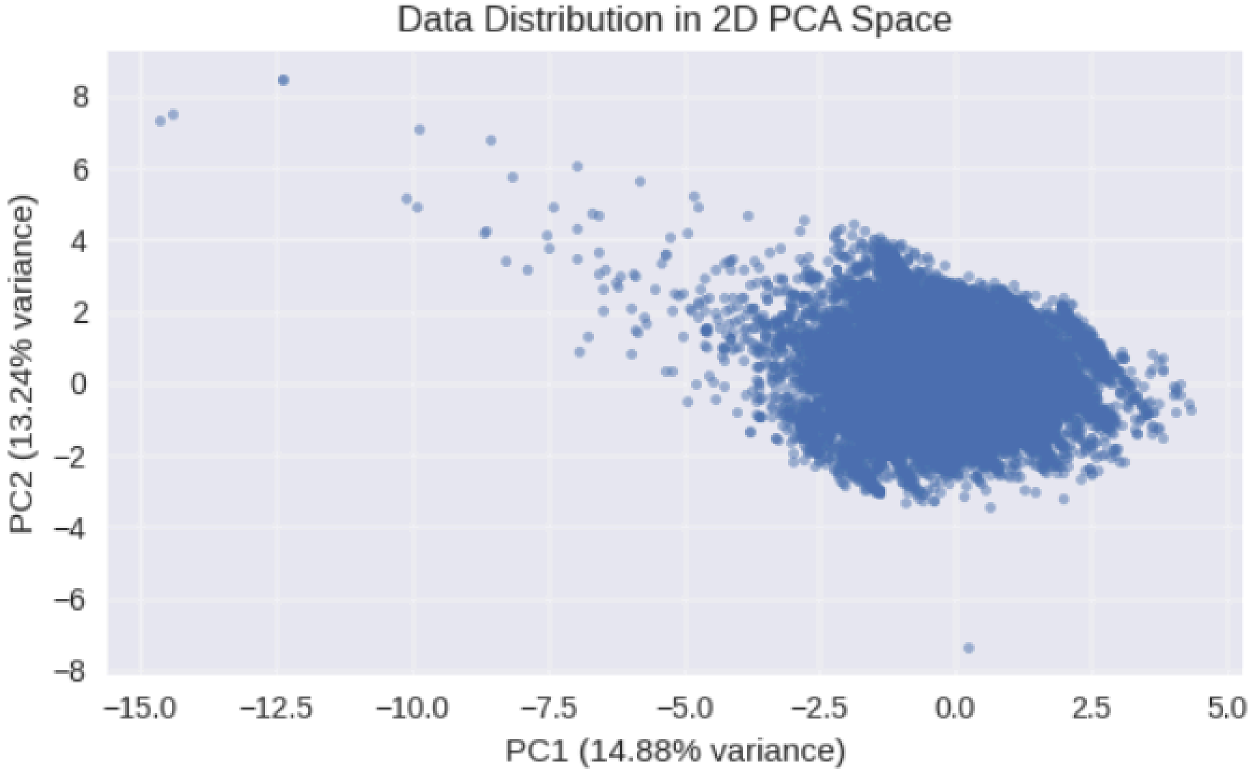
1. Feature Correlation matrix for the dataset



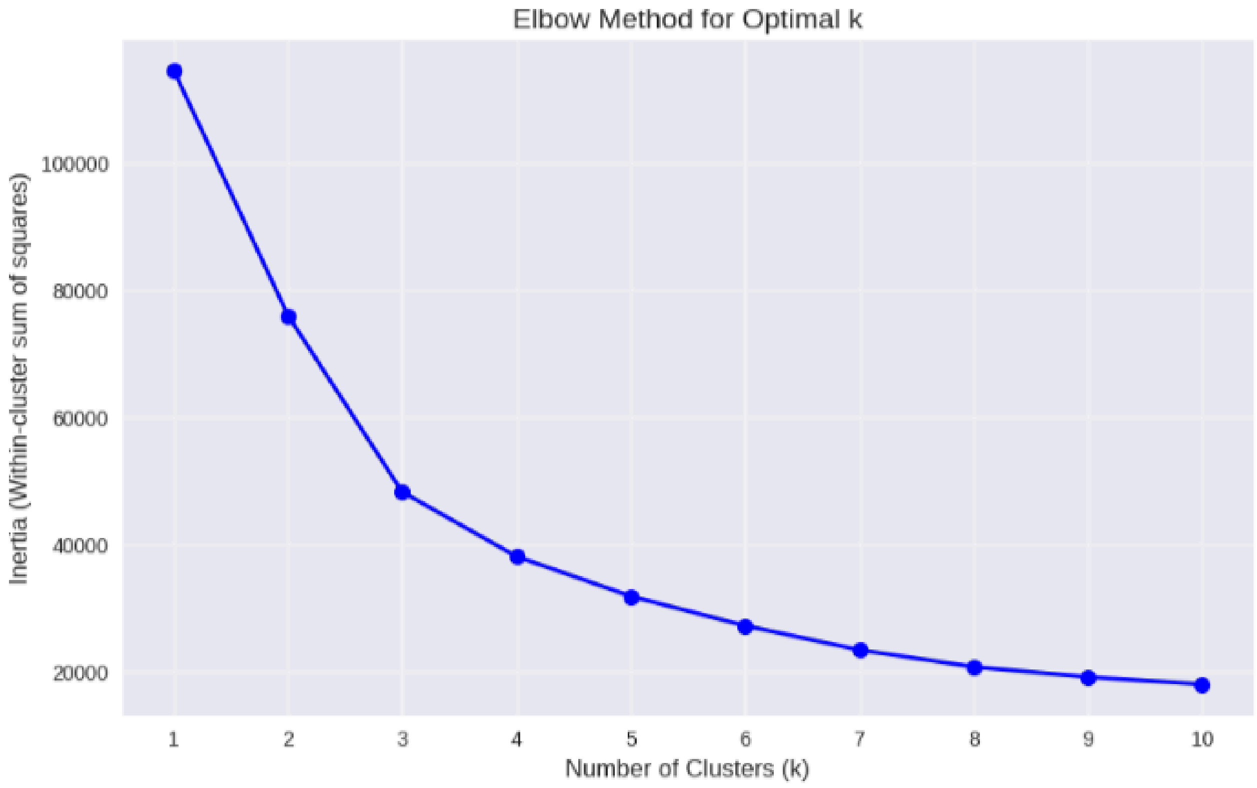
	age	balance	campaign	previous	job	education	housing	loan	default
age	1.00	0.10	0.00	0.00	-0.02	-0.11	-0.19	-0.02	-0.02
balance	0.10	1.00	-0.01	0.02	0.02	0.06	-0.07	-0.08	-0.07
campaign	0.00	-0.01	1.00	-0.03	0.01	0.01	-0.02	0.01	0.02
previous	0.00	0.02	-0.03	1.00	-0.00	0.02	0.04	-0.01	-0.02
job	-0.02	0.02	0.01	-0.00	1.00	0.17	-0.13	-0.03	-0.01
education	-0.11	0.06	0.01	0.02	0.17	1.00	-0.09	-0.05	-0.01
housing	-0.19	-0.07	-0.02	0.04	-0.13	-0.09	1.00	0.04	-0.01
loan	-0.02	-0.08	0.01	-0.01	-0.03	-0.05	0.04	1.00	0.08
default	-0.02	-0.07	0.02	-0.02	-0.01	-0.01	-0.01	0.08	1.00

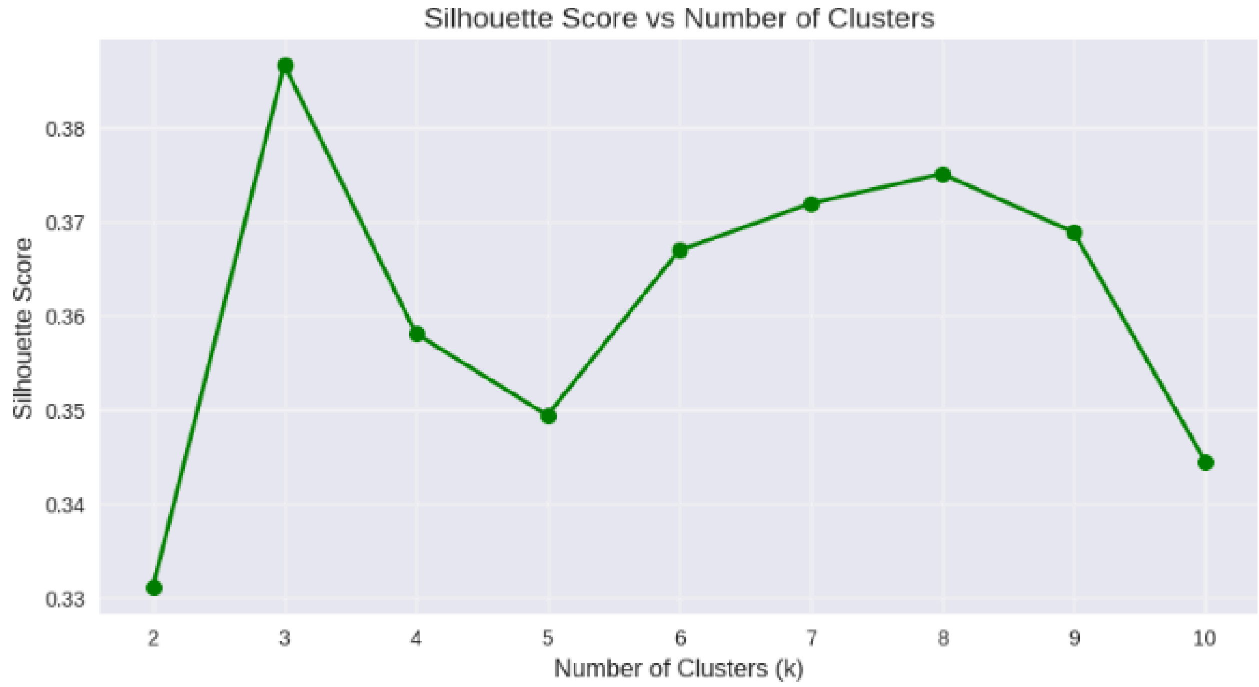
2. ‘Explained variance by Component’ and ‘Data Distribution in PCA Space’ after Dimensionality Reduction with PCA


Explained Variance by Component
Total: 28.12%
Explained Variance Ratio
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.75
1.00
1.25
1.50
1.75
2.00
2.25
Principal Component

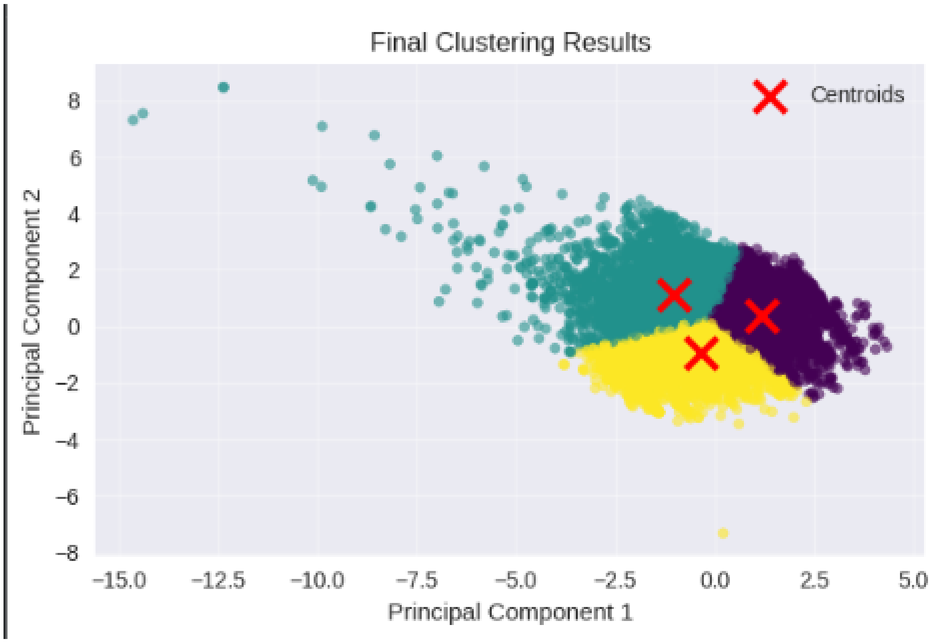

Data Distribution in 2D PCA Space
PC2 (13.24% variance)
8
6
4
2
0
−2
−4
−6
−8
−15.0
−12.5
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
PC1 (14.88% variance)

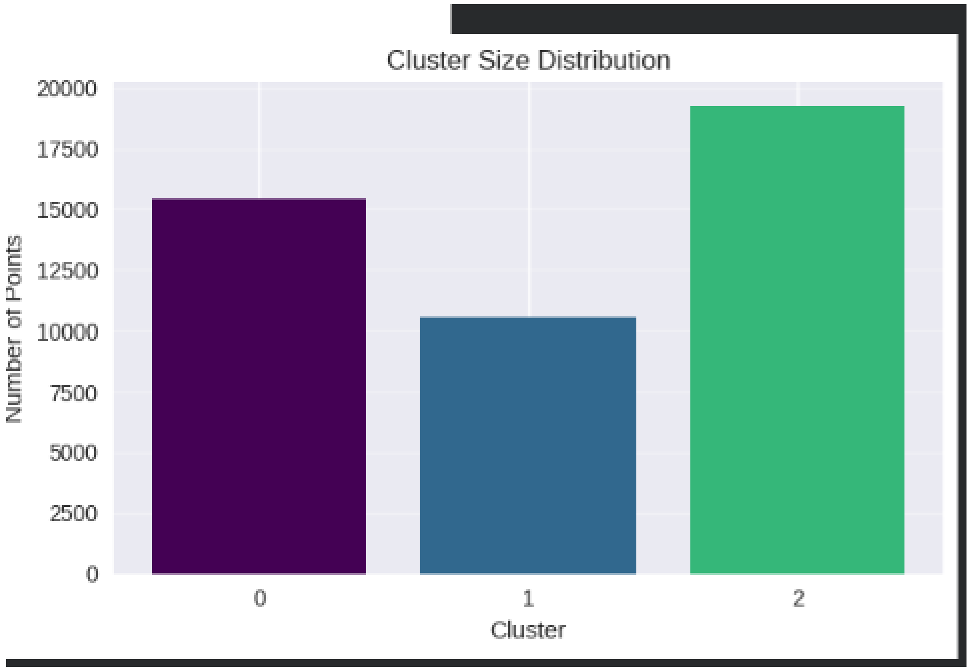
3. ‘Inertia Plot’ and ‘Silhouette Score Plot’ for K-means

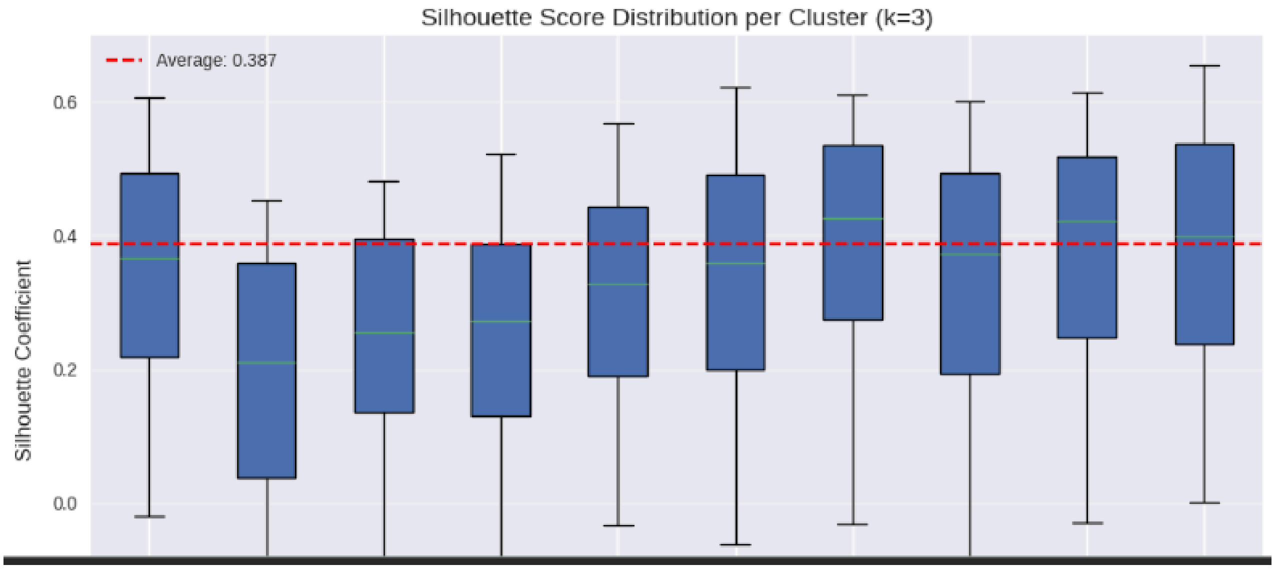




4. K-means Clustering Results with Centroids Visible (Scatter Plot) K-means Cluster Sizes (Bar Plot) Silhouette distribution per cluster for K-means (Box Plot)






Silhouette Score Distribution per Cluster (k=3)
Average: 0.387
Silhouette Coefficient
0.6
0.4
0.2
0.0