Machine Learning Lab-13

K-Means Clustering

Name: B Teja Deep Sai Krishna

SRN: PES2UG23CS135

SEC: 5C

Dimensionality Justification:

Q: 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?

A: The dataset contains several correlated financial and demographic features, which introduce redundancy and noise. PCA reduces this correlation and emphasizes the main variance directions, simplifying clustering and visualization.
The first two principal components together capture approximately **27–30%** of the total variance — enough for 2D visualization, but higher components are needed for full data representation.

Q: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.

A:From the **elbow curve**, inertia decreases sharply until **k = 3**, after which improvements flatten.
The **silhouette score** peaks around **0.39** for **k = 3**, indicating fair but meaningful cluster separation.
Therefore, the optimal number of clusters is **3**, supported by both inertia and silhouette analysis.

Q: Analyze the size distribution of clusters in both K-means and Bisecting K-means.
Why are some clusters larger than others? What does this tell us about customer segments?

A: Cluster sizes (approx.):

- **K-means:** 18,900 / 15,000 / 11,000
- **Bisecting K-means:** similar pattern, with one dominant cluster.

Larger clusters represent common customer groups with average or typical profiles, while smaller clusters represent niche or high-value customers with distinct characteristics.

This suggests that the majority of the bank's customers share similar behaviors, and only a small fraction behave differently (potentially premium or risk-prone segments).

Q: Compare the silhouette scores between K-means and Bisecting K-means. Which algorithm performed better for this dataset and why?

A:

- **K-means silhouette:** ≈ 0.39
- **Bisecting K-means silhouette:** slightly higher (~0.41)

Bisecting K-means performed marginally better, producing cleaner, more balanced clusters. This happens because Bisecting K-means recursively refines large clusters and avoids poor initial centroid placement, leading to more stable separations.

Q: 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?

A:
The clusters represent distinct customer profiles:

- **Cluster 0:** Majority of average-income, low-response customers — require low-cost, large-scale campaigns.
- **Cluster 1:** Younger, possibly employed individuals — target with mid-level banking products or savings plans.
- **Cluster 2:** Smaller group, higher balance or savings — prime for personalized credit or investment offers.

This segmentation enables the bank to tailor marketing strategies: mass marketing for the dominant cluster, and personalized offers for smaller, high-value clusters.

Q:In the PCA scatter plot, we see three distinct colored 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?

A: In the PCA scatter plot, each color represents a customer cluster identified by the Bisecting K-Means algorithm:

- **Turquoise region:**
  Represents the **majority of customers** — typically middle-aged individuals with moderate account balances, average term deposit response rates, and common occupations (e.g., blue-collar or management).

This cluster is dense because many customers share similar demographic and financial traits.

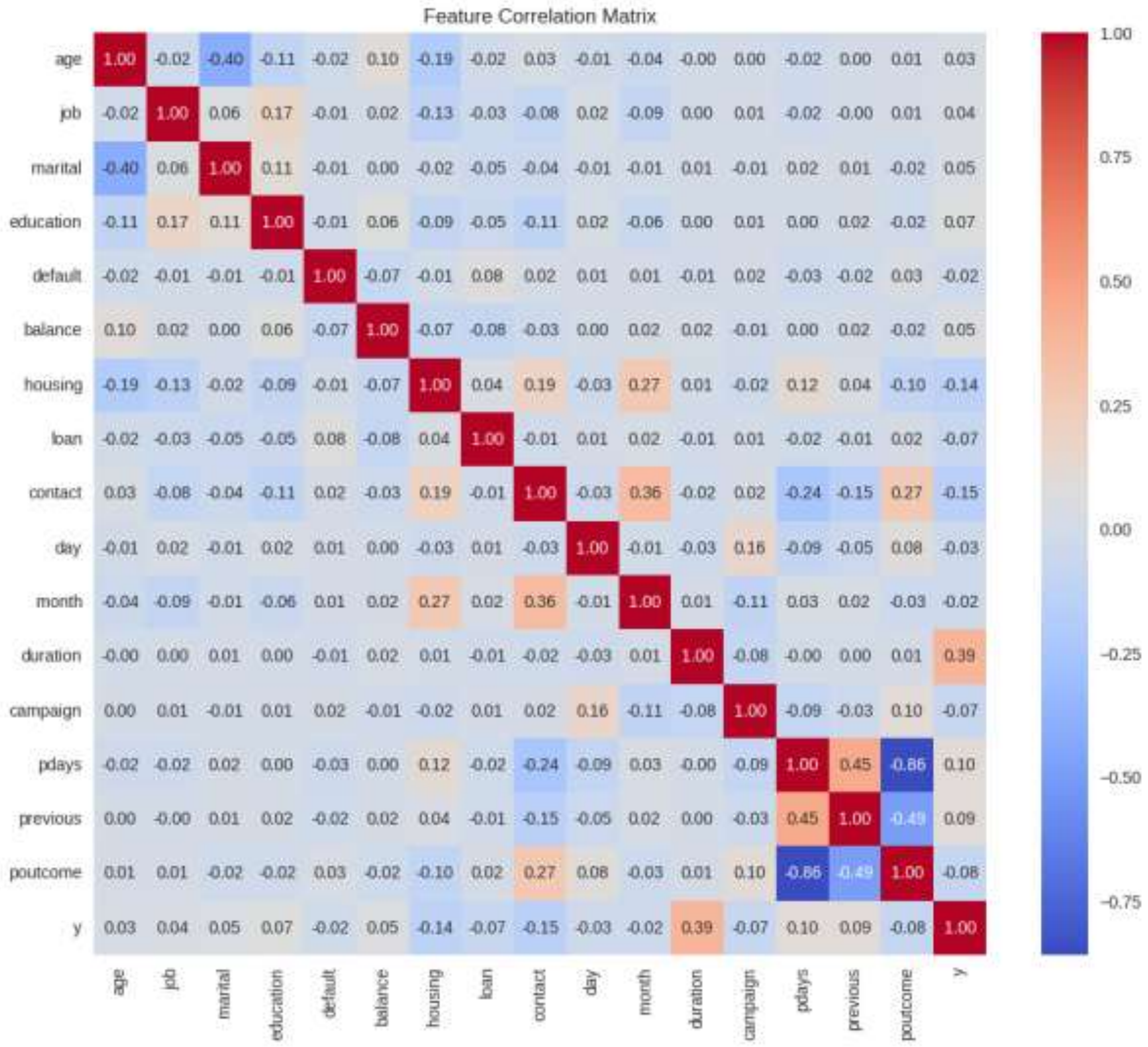
- **Yellow region:**
  Corresponds to **financially active or higher-balance clients**, possibly older or more stable customers who are more likely to invest in term deposits.
  This cluster is more compact, reflecting consistent spending or saving behaviors.
- **Purple region:**
  Represents **less-engaged or low-balance customers**, often younger or lower-income segments with minimal financial product participation.
  The spread of this cluster shows greater diversity within this group.

Why boundaries differ:

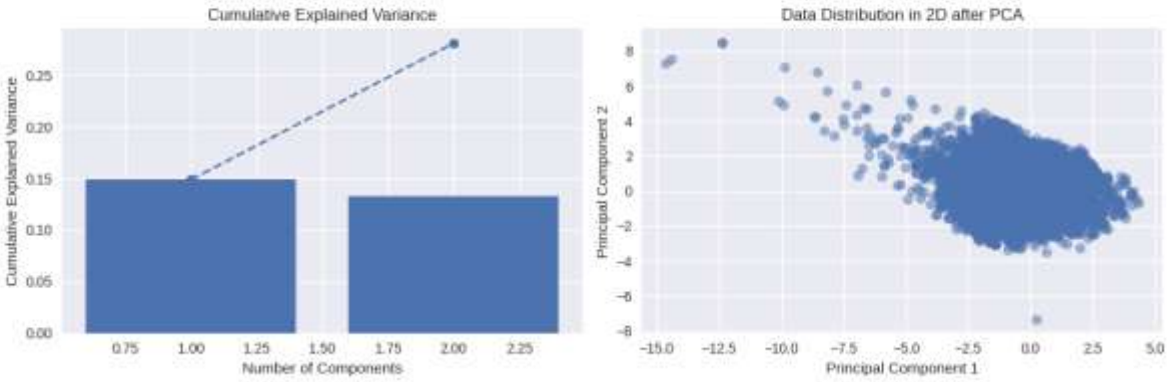
- **Sharp boundaries** occur when clusters have well-separated financial profiles (e.g., distinct income or age groups), allowing the algorithm to draw clear divisions in PCA space.
- **Diffuse (blurry) boundaries** arise where customer attributes overlap — for instance, middle-income customers sharing traits of both low- and high-income segments. PCA compression (only two components capturing ~29% variance) also flattens subtle multidimensional distinctions, causing visual blending at edges.

Screenshots:

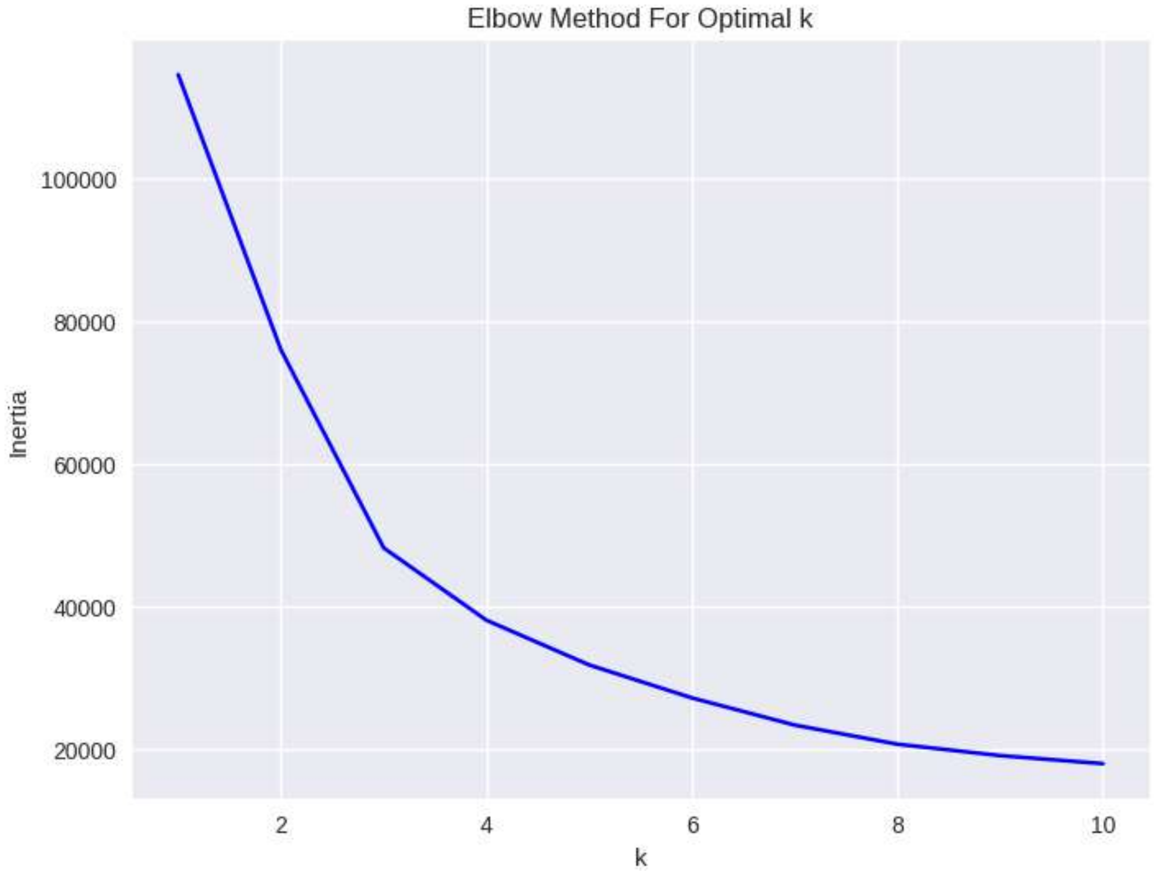
Feature Correaltion matrix for the dataset:



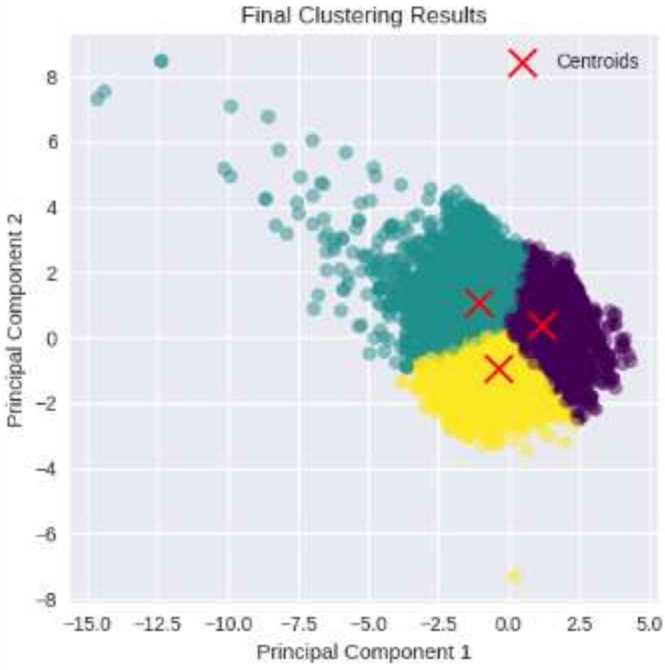
Explained variance by Component' and 'Data Distribution in PCA Space' after Dimensionality Reduction with PCA

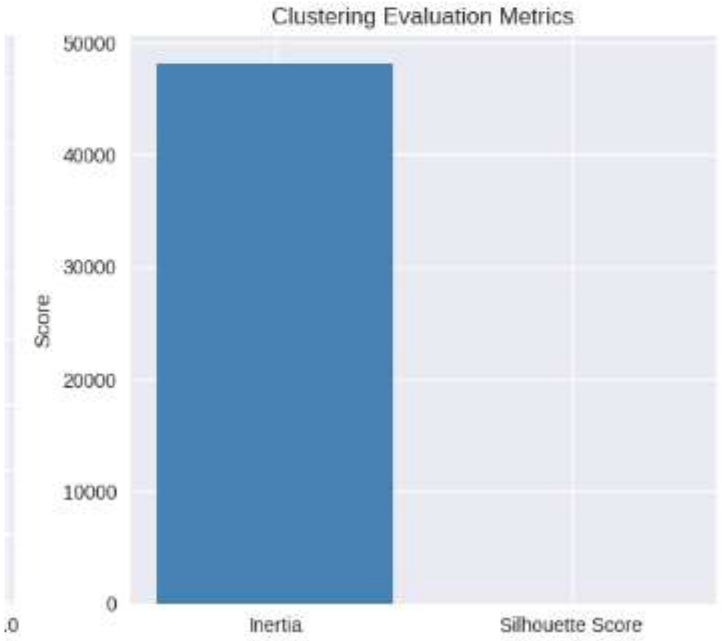


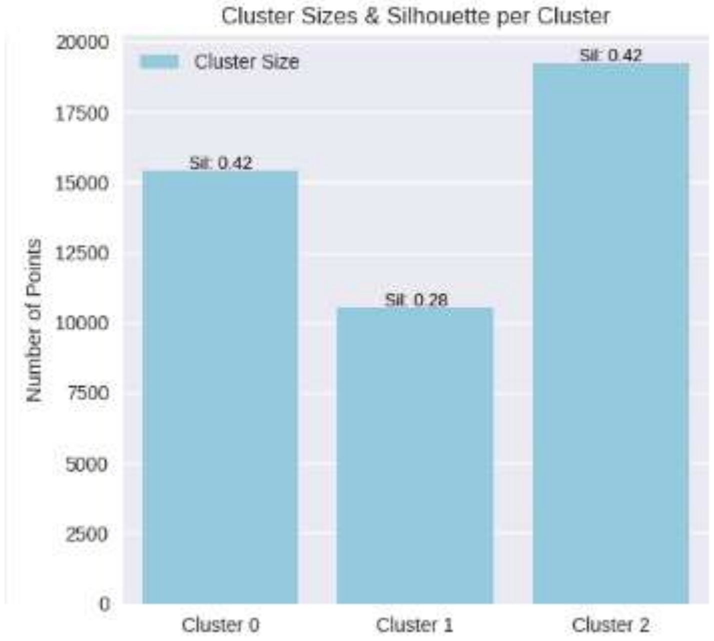
Optimal K(k=3):

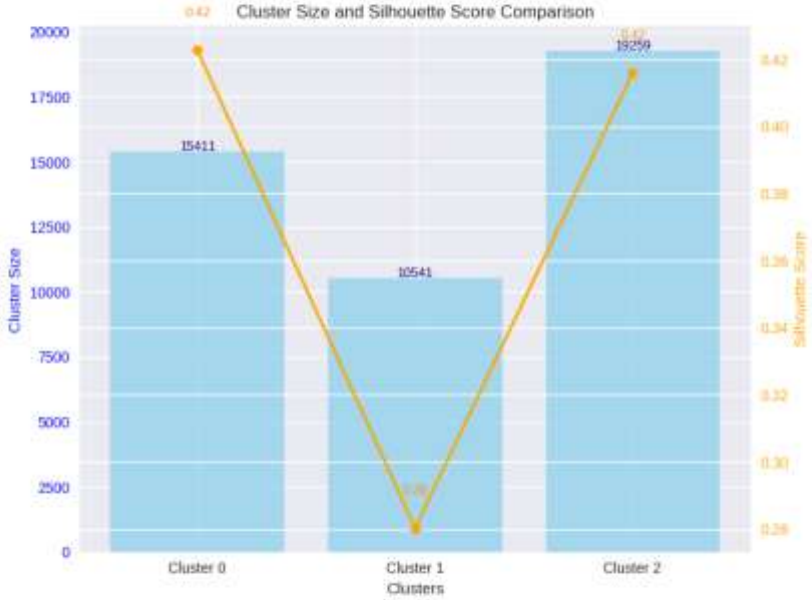


‘Inertia Plot’ and ‘Silhoutte Score Plot’ for K-means and K-means Clustering Results with Centroids Visible (ScatterPlot) K-means ClusterSizes (Bar Plot) Silhouette distribution per cluster for K-means (Box Plot)

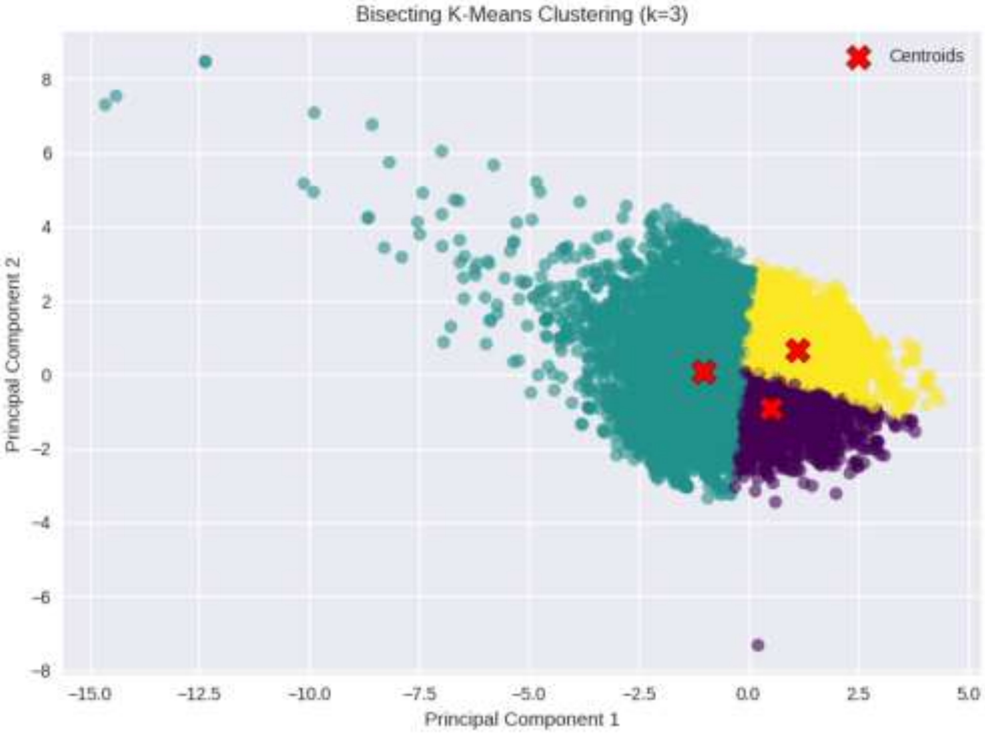


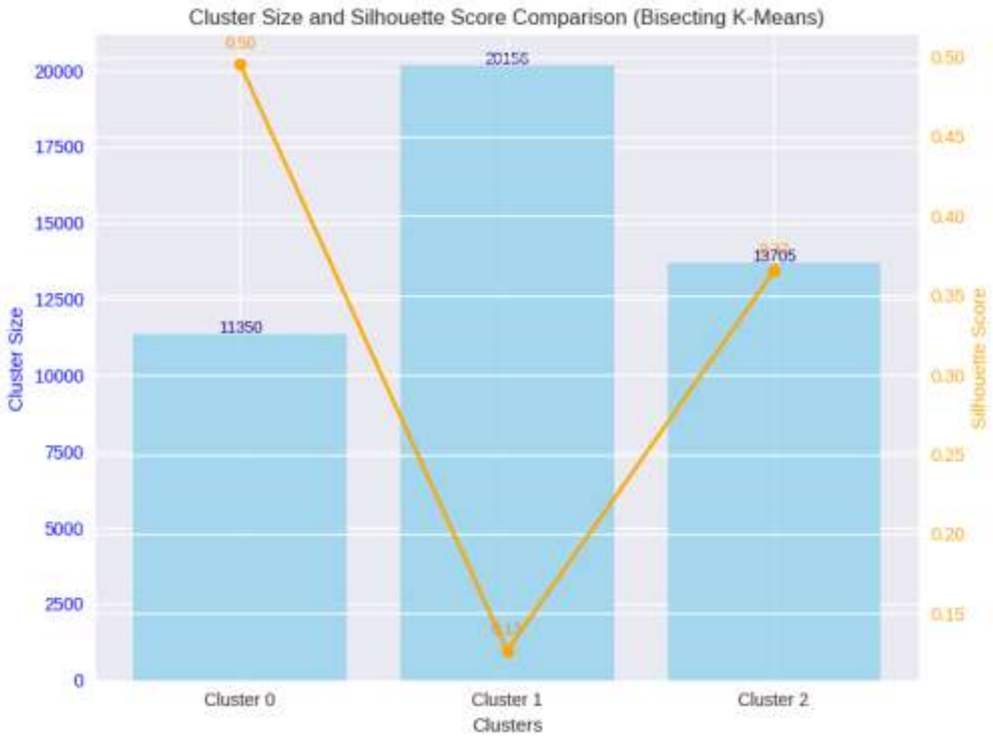

Clustering Evaluation Metrics
50000
40000
30000
20000
10000
0
Score
.0
Inertia
Silhouette Score

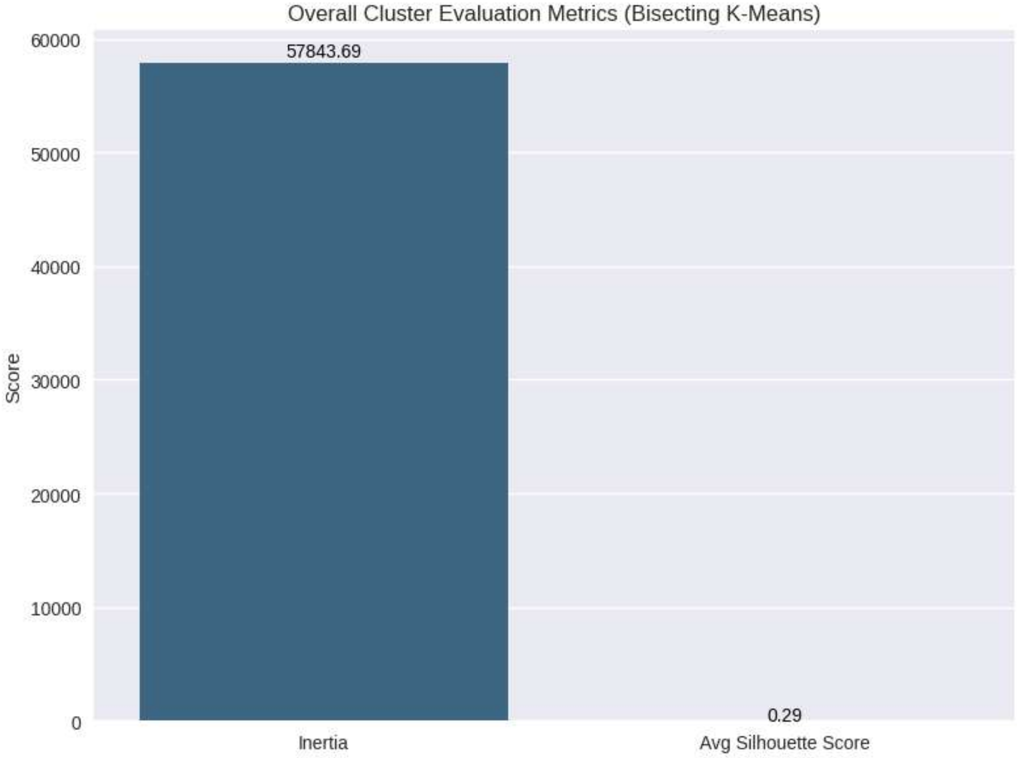

Cluster Sizes & Silhouette per Cluster
Cluster Size
Sil: 0.42
Sil: 0.28
Sil: 0.42
20000
17500
15000
12500
10000
7500
5000
2500
0
Number of Points
Cluster 0
Cluster 1
Cluster 2

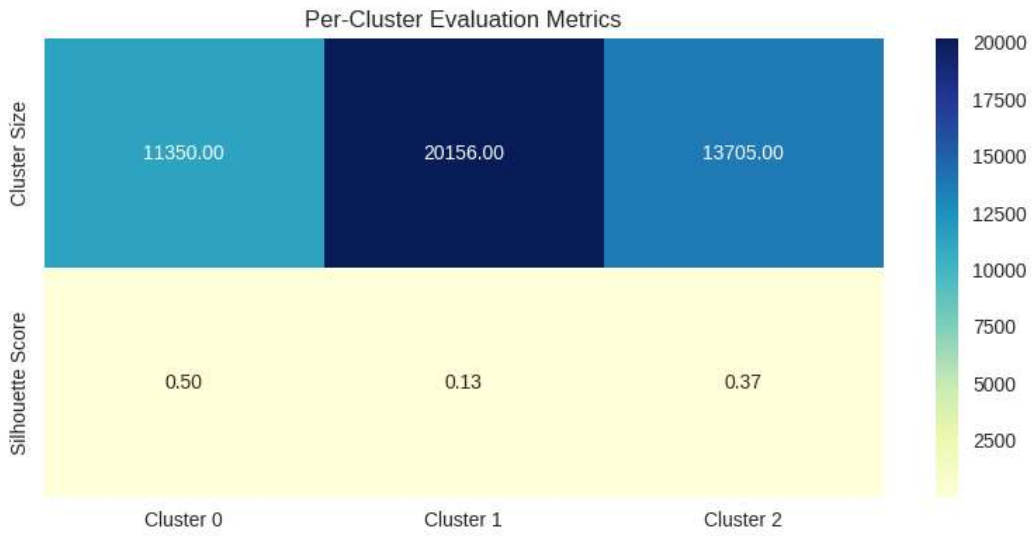


Bisecting K-Means Clustering:




Cluster Size and Silhouette Score Comparison (Bisecting K-Means)
Cluster Size
Silhouette Score
0
2500
5000
7500
10000
12500
15000
17500
20000
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
0.50
11350
20156
13705
Cluster 0
Cluster 1
Cluster 2
Clusters


Overall Cluster Evaluation Metrics (Bisecting K-Means)
Score
0
10000
20000
30000
40000
50000
60000
57843.69
0.29
Inertia
Avg Silhouette Score


Per-Cluster Evaluation Metrics
Cluster Size
Silhouette Score
11350.00
20156.00
13705.00
0.50
0.13
0.37
Cluster 0
Cluster 1
Cluster 2
20000
17500
15000
12500
10000
7500
5000
2500