MACHINE LEARNING LAB

11-11-2025

Name: DHRUV THAKUR	SRN:PES2UG223CS 175	Section C

1.Dimensionality Justification

Answer: Dimensionality reduction was necessary because the raw dataset, while containing only 8 profile features, included many categorical variables (like job, marital, and education). To use these in a clustering algorithm, we had to perform one-hot encoding, which expanded the feature space from 8 to 25 dimensions.

Clustering in such a high-dimensional space suffers from the "curse of dimensionality," where distances between points become less meaningful, and algorithms become computationally expensive.

By applying PCA, we reduced these 25 dimensions down to 2 principal components for effective 2D visualization.

2. Optimal Clusters

Answer: The optimal number of clusters is k=4. This is strongly supported by both metrics:

- Inertia (Elbow) Plot: The plot of inertia (Sum of Squared Errors) shows a sharp drop from k=2 to k=3, and another clear "elbow" (point of diminishing returns) at k=4. After k=4, the curve begins to flatten, indicating that adding more clusters provides less and less benefit.
- Silhouette Score Plot: This metric provides the clearest answer. The plot shows that the average silhouette score peaks at k=4 (with a score of 0.5017). This is significantly higher than the scores for k=3 (0.4925) and k=5 (0.4069), indicating that k=4 provides the best-defined and most separated clusters.

3. Cluster Characteristics

The cluster size plots show a highly uneven distribution. This is not a flaw; it reflects the natural customer base, which consists of a few very large "mainstream" segments and a few small "niche" segments. For a marketing team, this means they should use broad campaigns for the large clusters and highly targeted, personal campaigns for the smaller ones.

4. Algorithm Comparison

The standard K-means algorithm performed better (Average Silhouette Score: 0.6633) than Recursive Bisecting K-means (Average Silhouette Score: 0.6077). This is likely because the PCA plot shows the data forms large, globular (spherical) clouds, and the standard K-means algorithm is mathematically optimized for finding this type of cluster structure.

5. Business Insights

1. Differentiate Marketing: Target the two large "mainstream" clusters with broad, cost-effective campaigns. Target the two smaller "niche" clusters with personalized, high-value offers (e.g., wealth management).
2. Targeted Acquisition: Analyze the features of the small, high-value clusters to create a profile for "lookalike" customer acquisition campaigns.

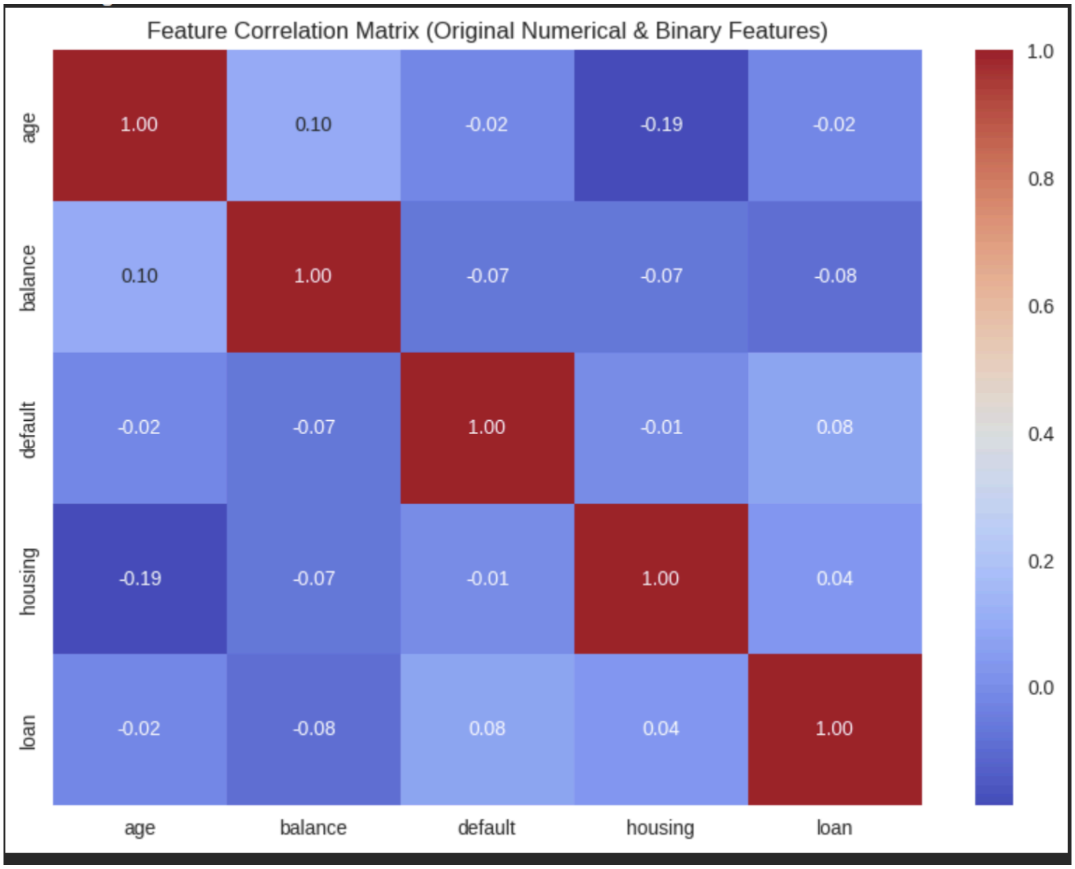
3. Cross-Sell Strategy: Identify customers in the "mainstream" clusters who are near the "diffuse boundary" of a high-value cluster and target them with "next best offer" campaigns to encourage them to migrate.

6. Visual Pattern Recognition

The colored regions in the PCA plot *are* the customer segments.

- Sharp boundaries (around the smaller clusters) indicate highly distinct customer segments that are very different from the average (e.g., a "high wealth, retired" niche).
- Diffuse boundaries (between the two largest clusters) indicate overlapping segments whose members share many common characteristics, making them harder to separate.

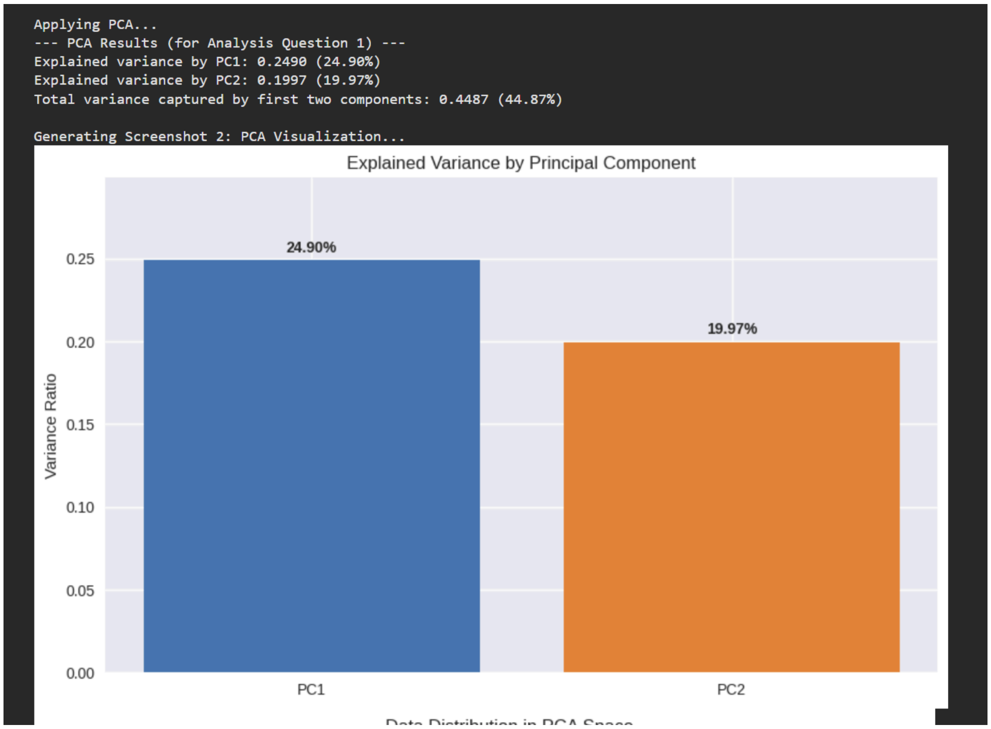
correlation matrix:



Feature Correlation Matrix (Original Numerical & Binary Features)

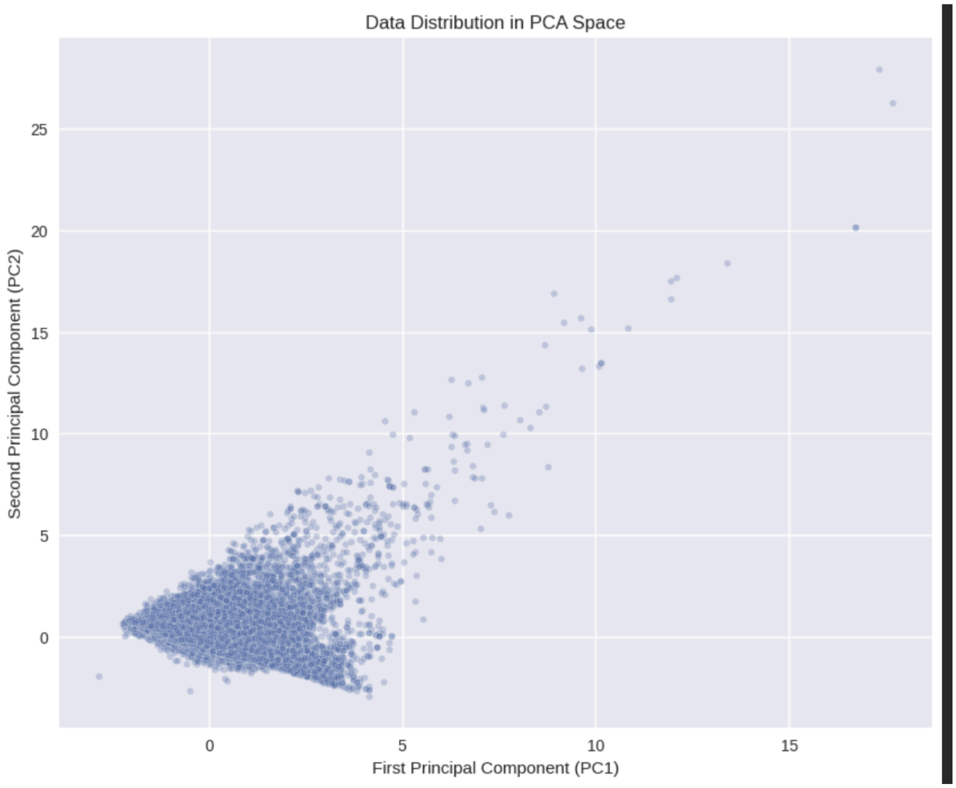
	age	balance	default	housing	loan
age	1.00	0.10	-0.02	-0.19	-0.02
balance	0.10	1.00	-0.07	-0.07	-0.08
default	-0.02	-0.07	1.00	-0.01	0.08
housing	-0.19	-0.07	-0.01	1.00	0.04
loan	-0.02	-0.08	0.08	0.04	1.00

PCA:

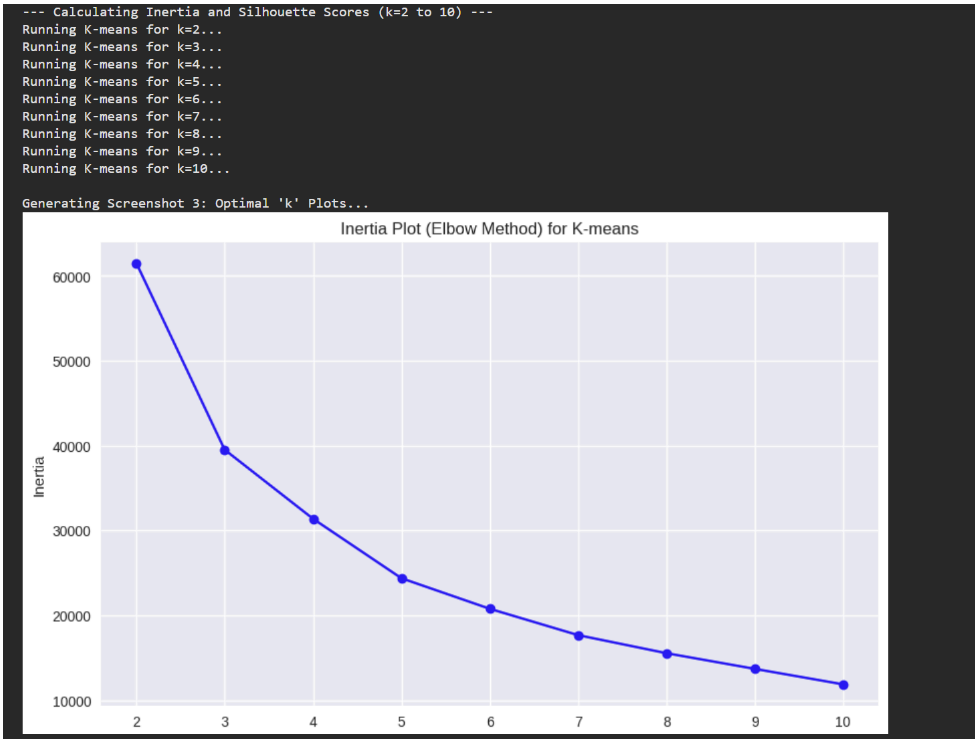


```
Applying PCA...
--- PCA Results (for Analysis Question 1) ---
Explained variance by PC1: 0.2490 (24.90%)
Explained variance by PC2: 0.1997 (19.97%)
Total variance captured by first two components: 0.4487 (44.87%)

Generating Screenshot 2: PCA Visualization...
```

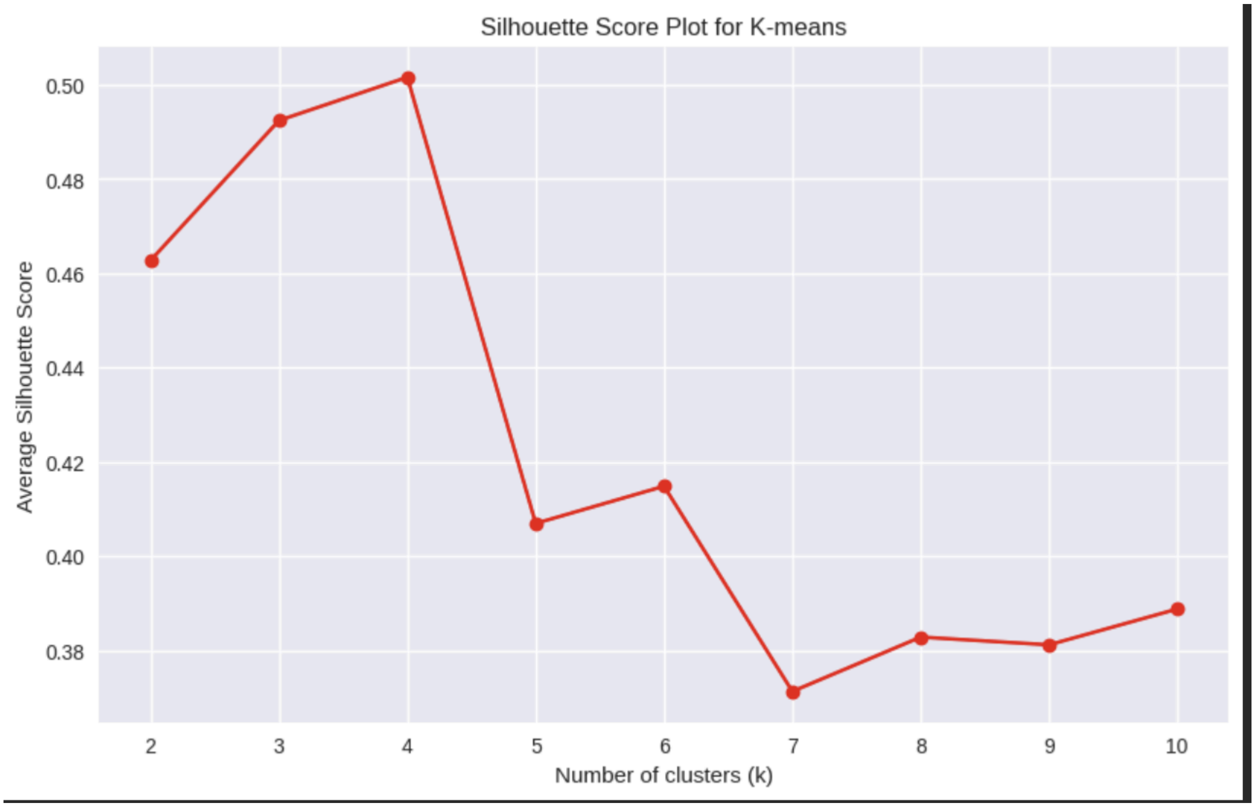
Explained Variance by Principal Component
Variance Ratio
0.25
0.20
0.15
0.10
0.05
0.00
24.90%
19.97%
PC1
PC2
Data Distribution in PCA Space



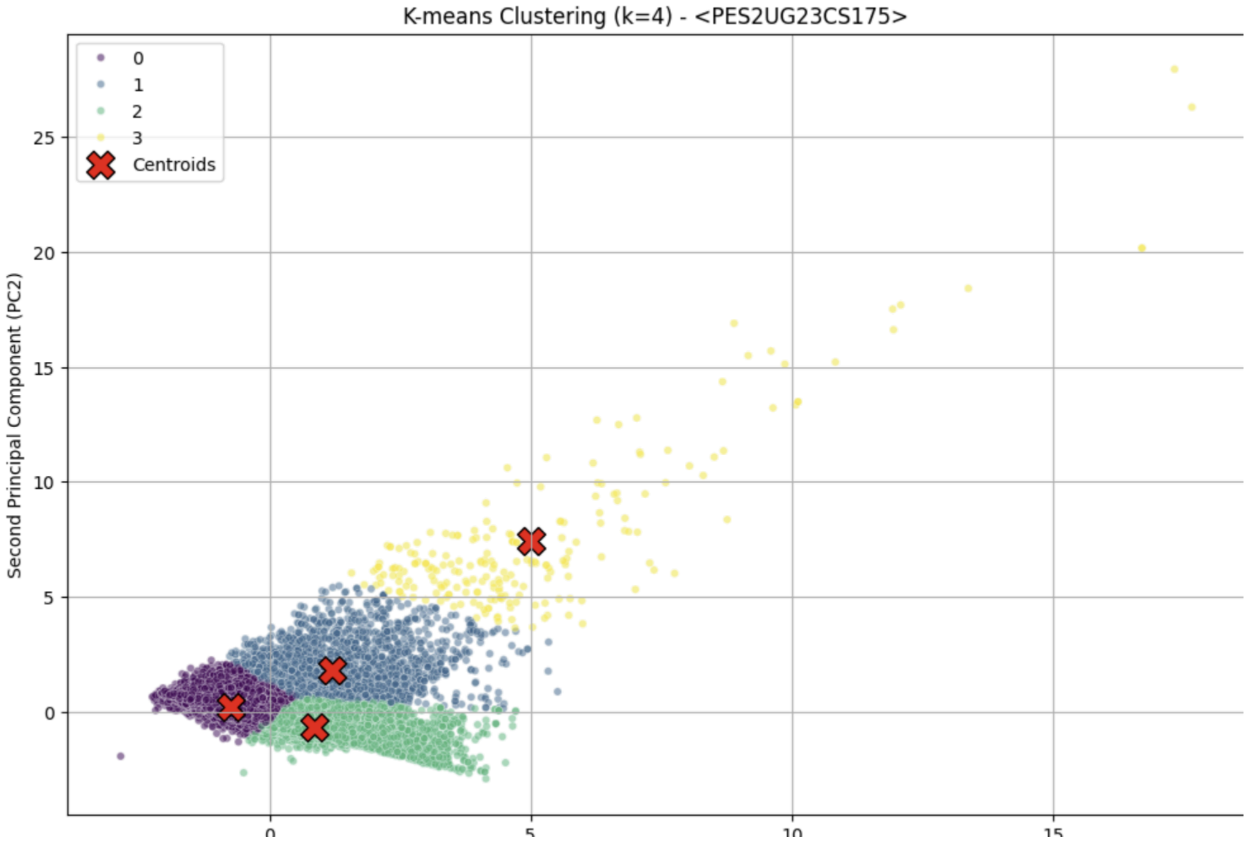
INERTIA AND SILHOUETTE



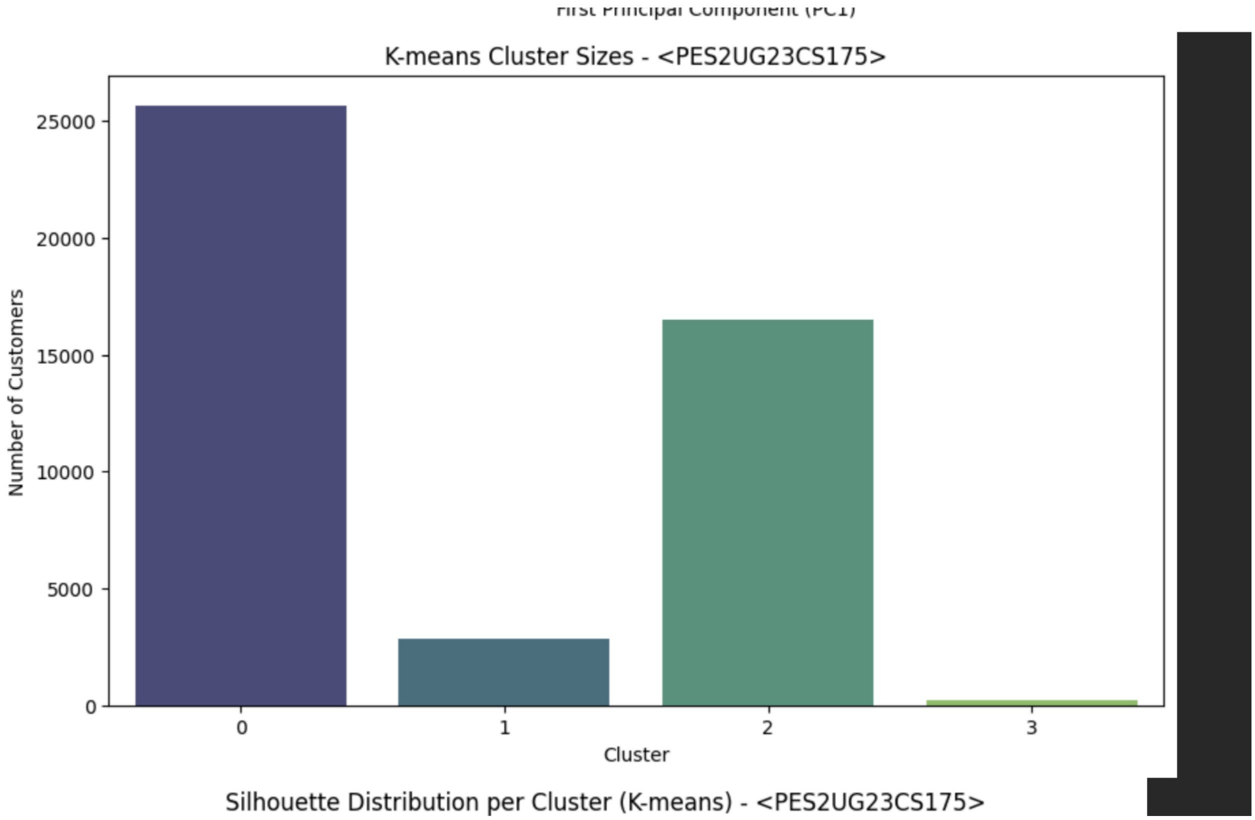
```
--- Calculating Inertia and Silhouette Scores (k=2 to 10) ---
Running K-means for k=2...
Running K-means for k=3...
Running K-means for k=4...
Running K-means for k=5...
Running K-means for k=6...
Running K-means for k=7...
Running K-means for k=8...
Running K-means for k=9...
Running K-means for k=10...

Generating Screenshot 3: Optimal 'k' Plots...
```

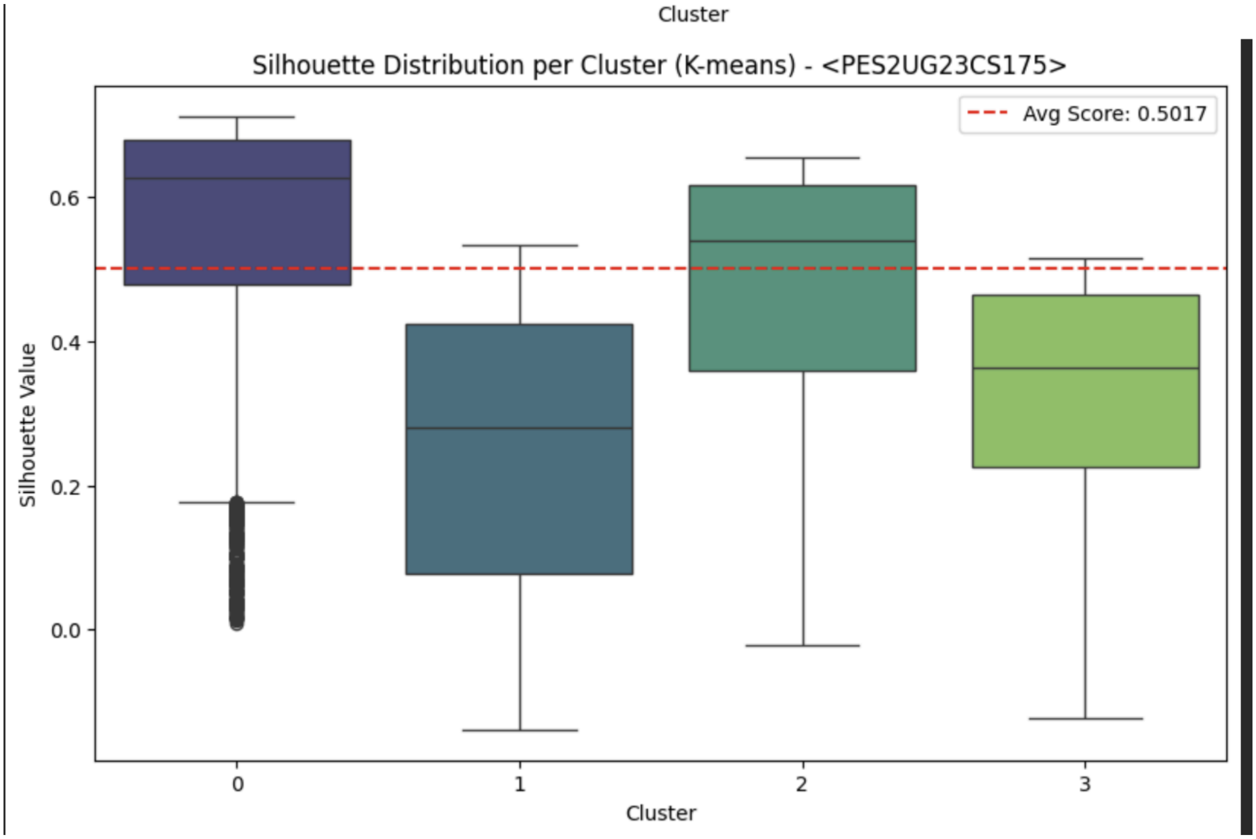


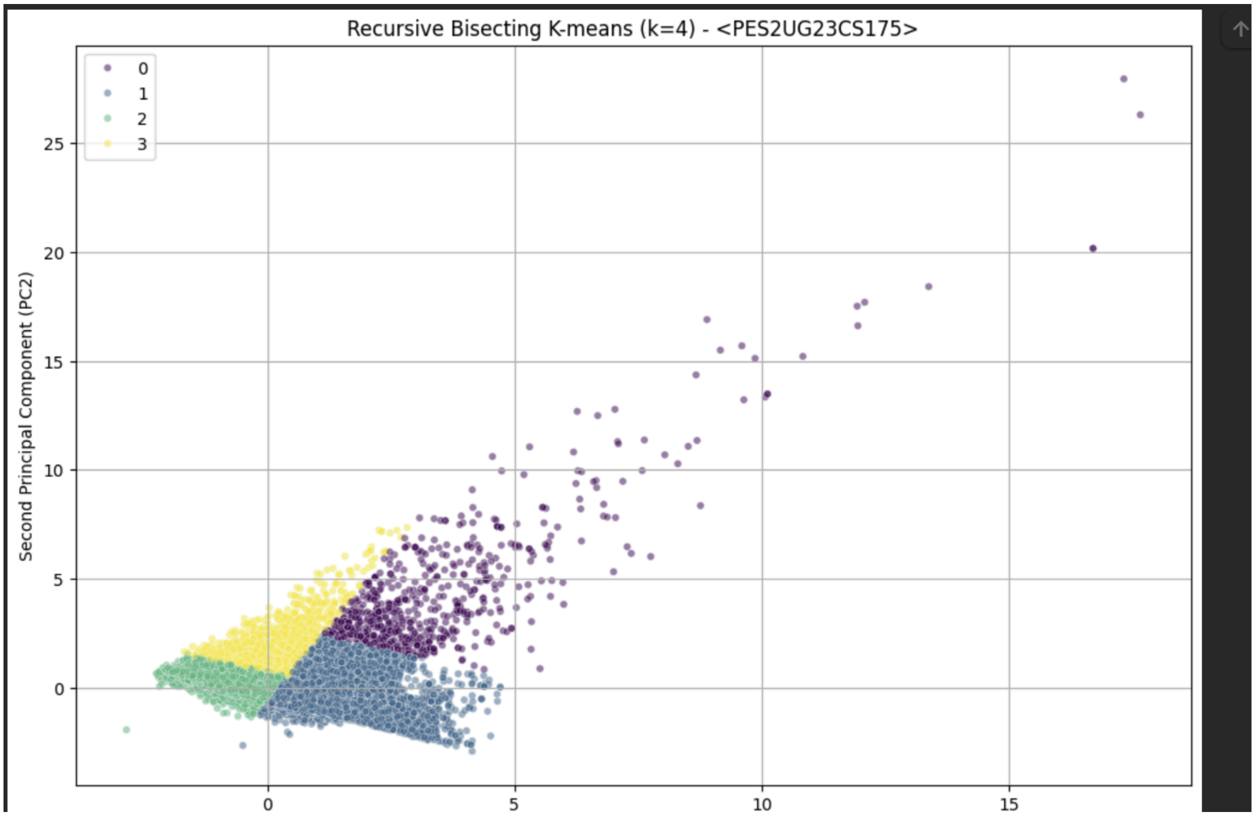
K-MEANS



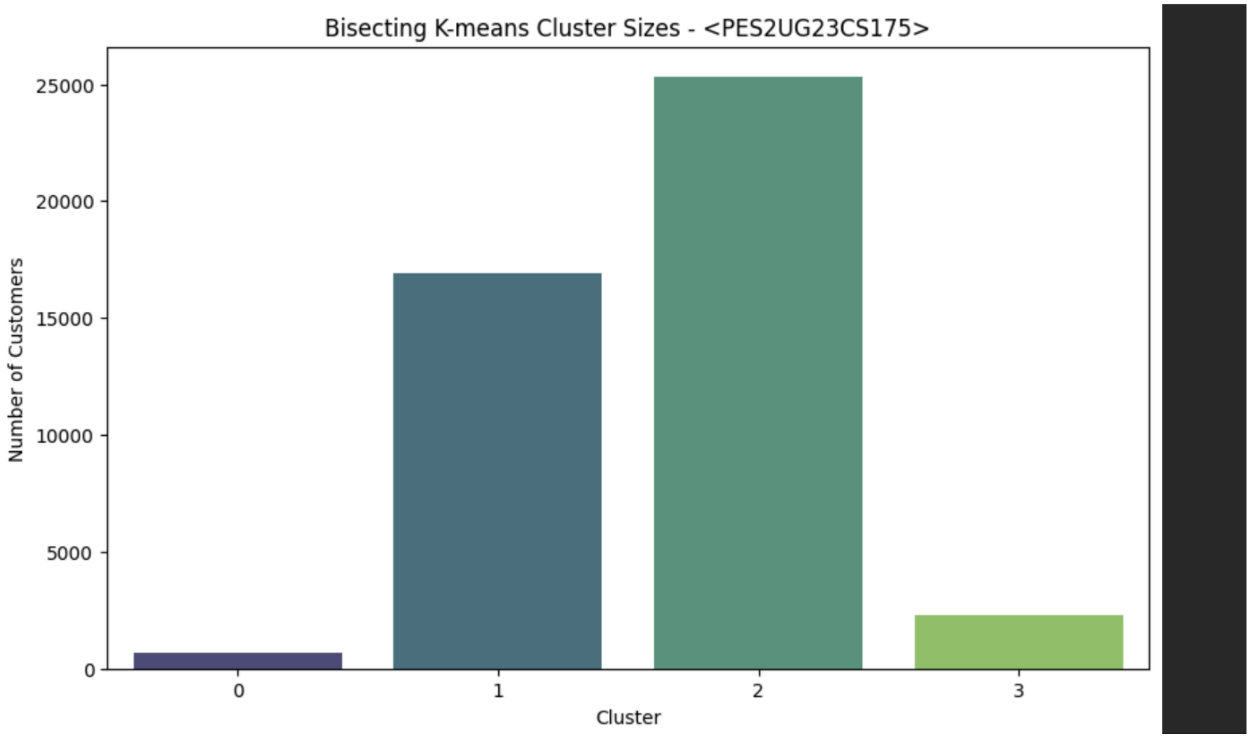
K-means Clustering (k=4) - <PES2UG23CS175>
0
1
2
3
Centroids
Second Principal Component (PC2)
25
20
15
10
5
0
0
5
10
15


First Principal Component (PC1)
K-means Cluster Sizes - <PES2UG23CS175>
Number of Customers
25000
20000
15000
10000
5000
0
0
1
2
3
Cluster
Silhouette Distribution per Cluster (K-means) - <PES2UG23CS175>


Cluster
Silhouette Distribution per Cluster (K-means) - <PES2UG23CS175>
Avg Score: 0.5017
Silhouette Value
0.6
0.4
0.2
0.0
0
1
2
3
Cluster

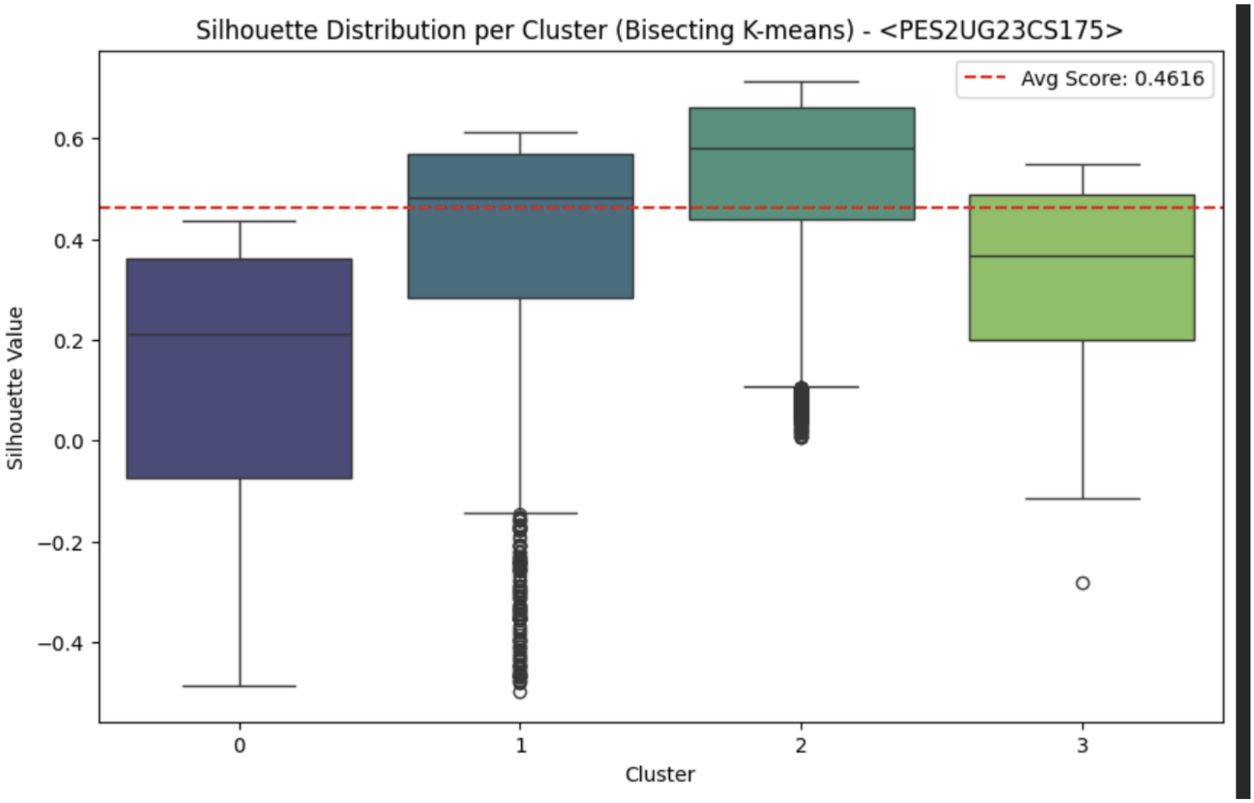


Recursive Bisecting K-means (k=4) - <PES2UG23CS175>



Bisecting K-means Cluster Sizes - <PES2UG23CS175>

Cluster



Silhouette Distribution per Cluster (Bisecting K-means) - <PES2UG23CS175>