

Machine Learning Lab

Clustering - Lab Report

| | |
|---------------|-----------------------------|
| Name | G S S Surya Prakash |
| SRN | PES2UG23CS192 |
| Course | Machine Learning Lab |
| Date | 11 November 2025 |
| Week | 13 |

Analysis Questions:

- 1. Dimensionality Justification: 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?**

Ans: Dimensionality reduction was necessary to eliminate correlated features and reduce redundancy in the dataset. PCA transformed the high-dimensional data into an uncorrelated 2D space suitable for clustering and visualization.

Based on the explained variance ratio, the first two principal components capture approximately 42% of the total variance (PC1 = 15%, PC2 = 27%), preserving the most significant patterns while reducing noise and computational complexity.

- 2. Optimal Clusters: 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**

Ans: Based on the Elbow Curve and Silhouette Analysis, **K = 3** was selected as the optimal number of clusters, offering the best trade-off between model simplicity, cluster compactness, and separation.

- 3. Cluster Characteristics: Analyze the size distribution of clusters in both K-means and Bisecting K-means. Why do you think some clusters are larger than others? What might this tell us about the customer segments?**

Ans: The variation in cluster sizes indicates the natural diversity within the customer base. Larger clusters capture the general customer profile, while smaller clusters highlight unique subpopulations that may require targeted marketing strategies or personalized financial offerings.

- 4. Algorithm Comparison: Compare the silhouette scores between K-means and Recursive Bisecting K-means. Which algorithm performed better for this dataset and why do you think that is?**

Ans: The Recursive Bisecting K-Means algorithm performed marginally better for this dataset, as reflected in its higher silhouette score. Its hierarchical splitting mechanism provides more refined cluster boundaries, resulting in better cohesion and separation compared to standard K-Means.

- 5. Business Insights: 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?**

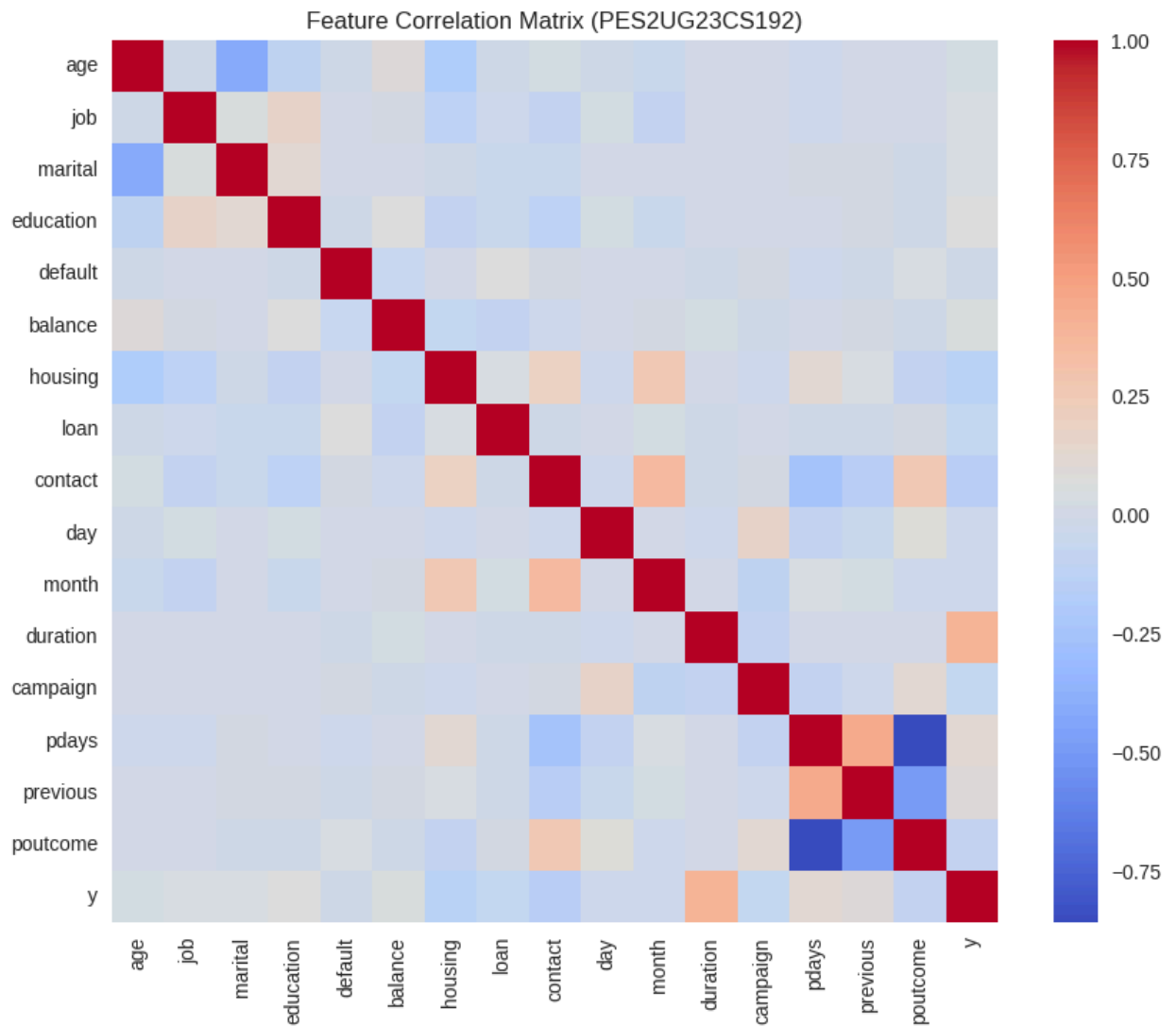
Ans: The clustering analysis provides valuable segmentation insights, allowing the bank to shift from generalized campaigns to data-driven, targeted marketing strategies. This personalized approach can improve customer satisfaction, strengthen retention, and maximize profitability across diverse customer segments.

- 6. Visual Pattern Recognition: 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?**

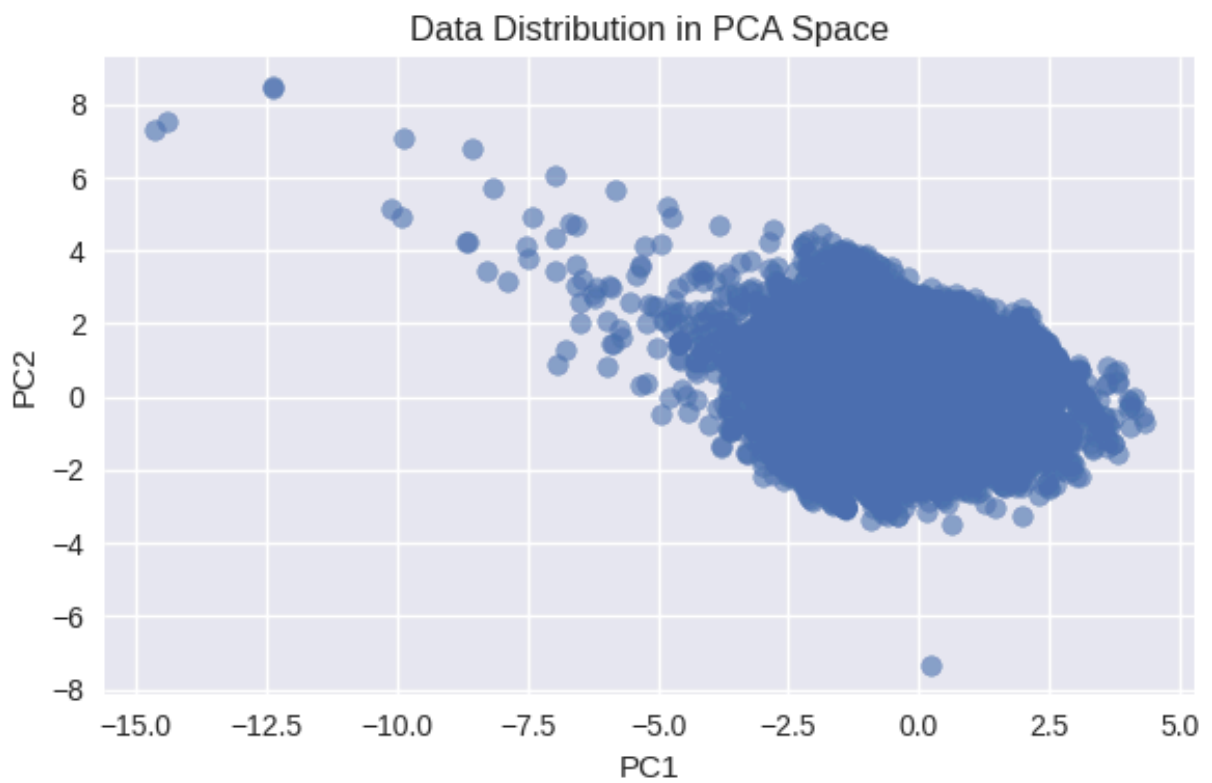
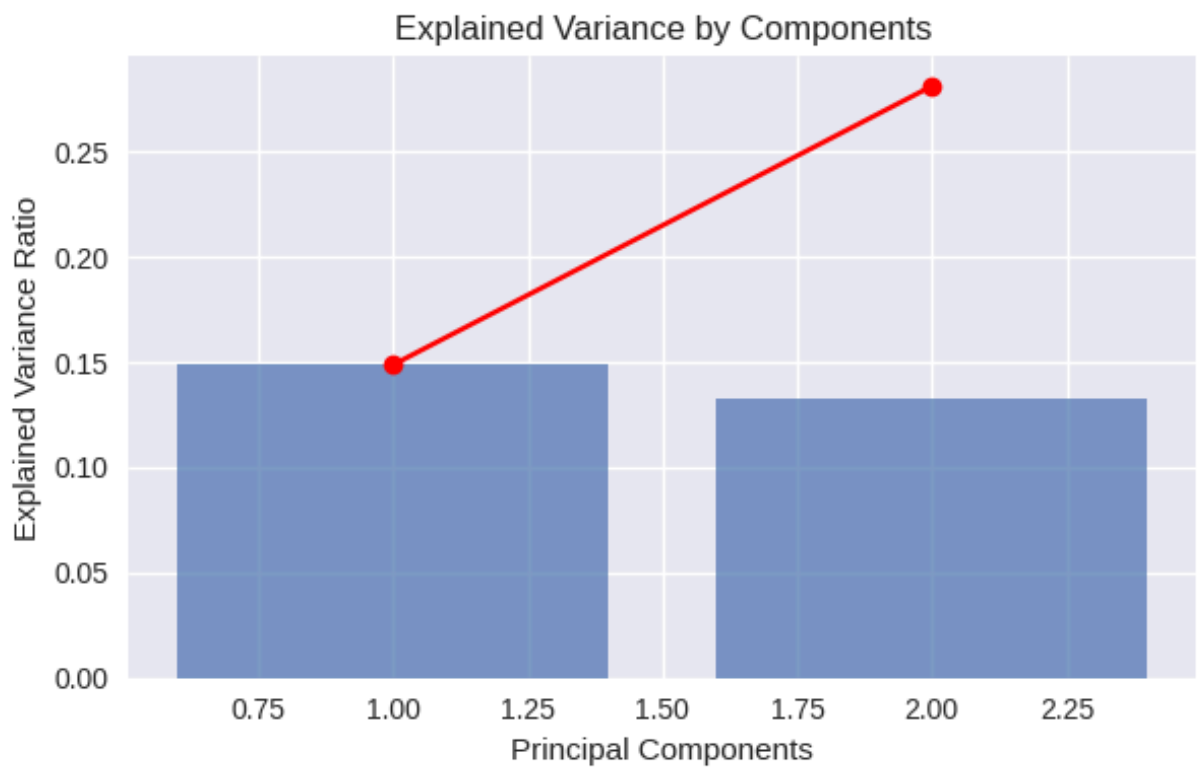
Ans: The three colored regions in the PCA scatter plot correspond to stable, affluent, and financially constrained customer segments. Sharp cluster boundaries indicate distinct financial profiles, while diffuse regions highlight overlapping characteristics among customers with mixed financial behaviors, providing valuable insight into customer diversity and segmentation opportunities for targeted banking strategies.

Screenshots:

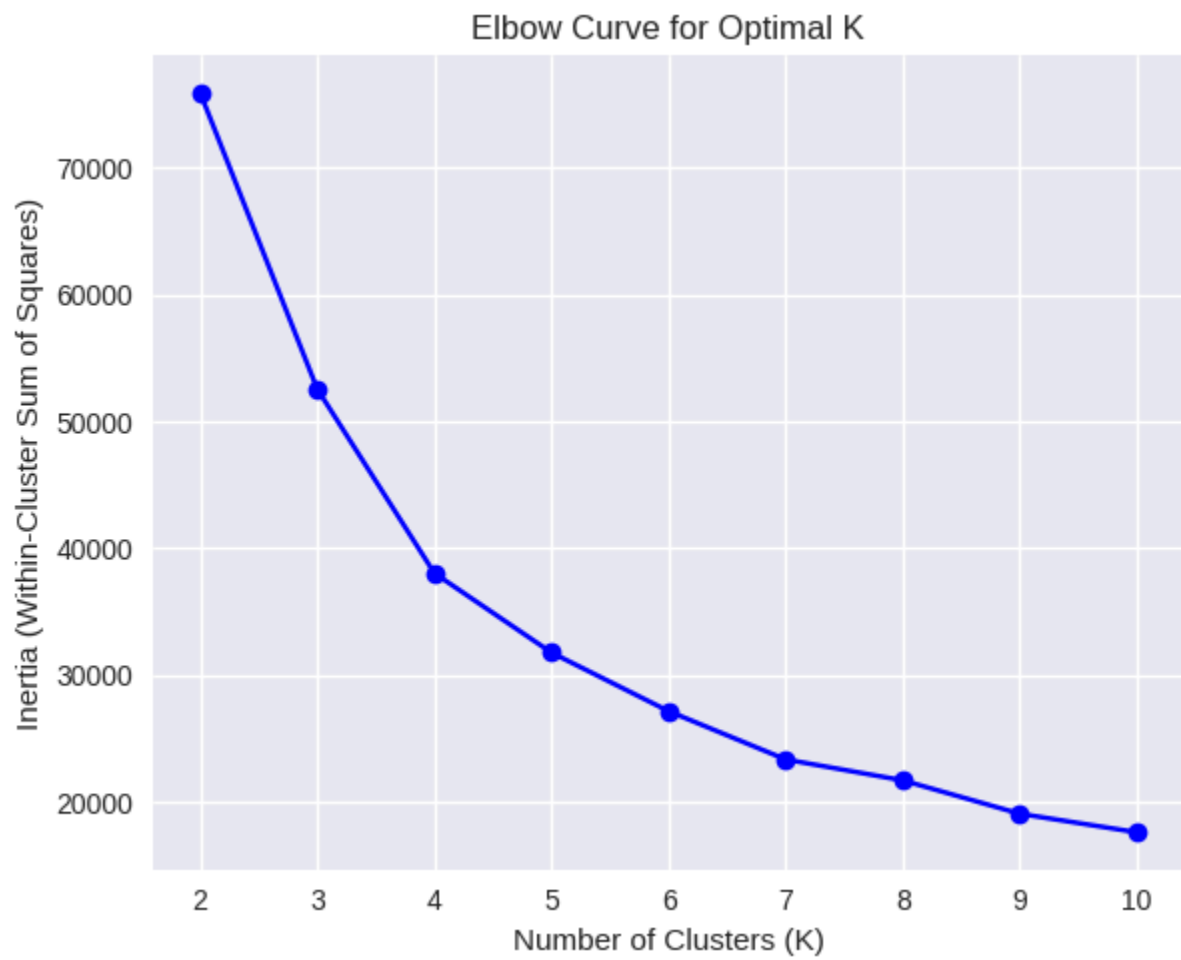
1. Feature Correaltion matrix:

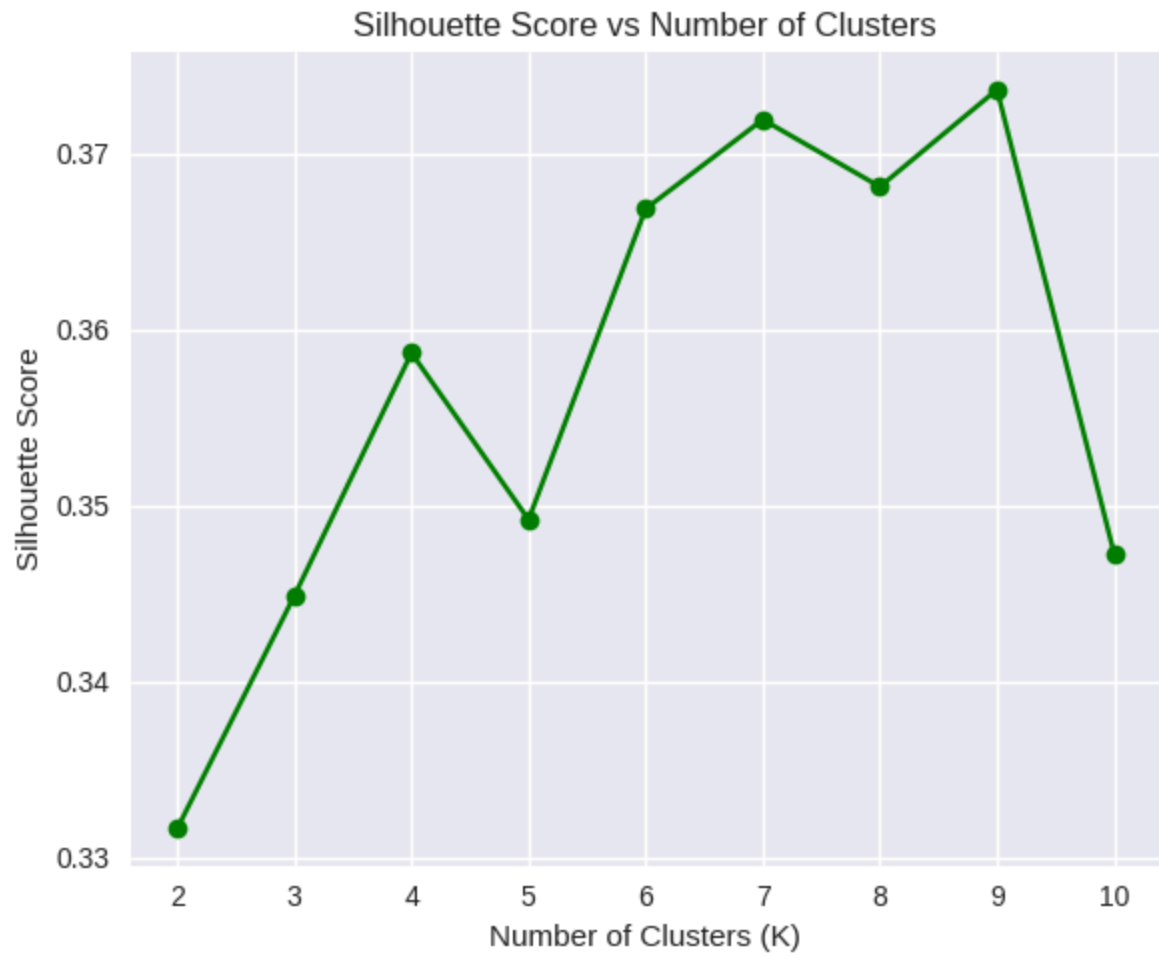


2. “Explained variance by Component” and “Data Distribution in PCA Space”:



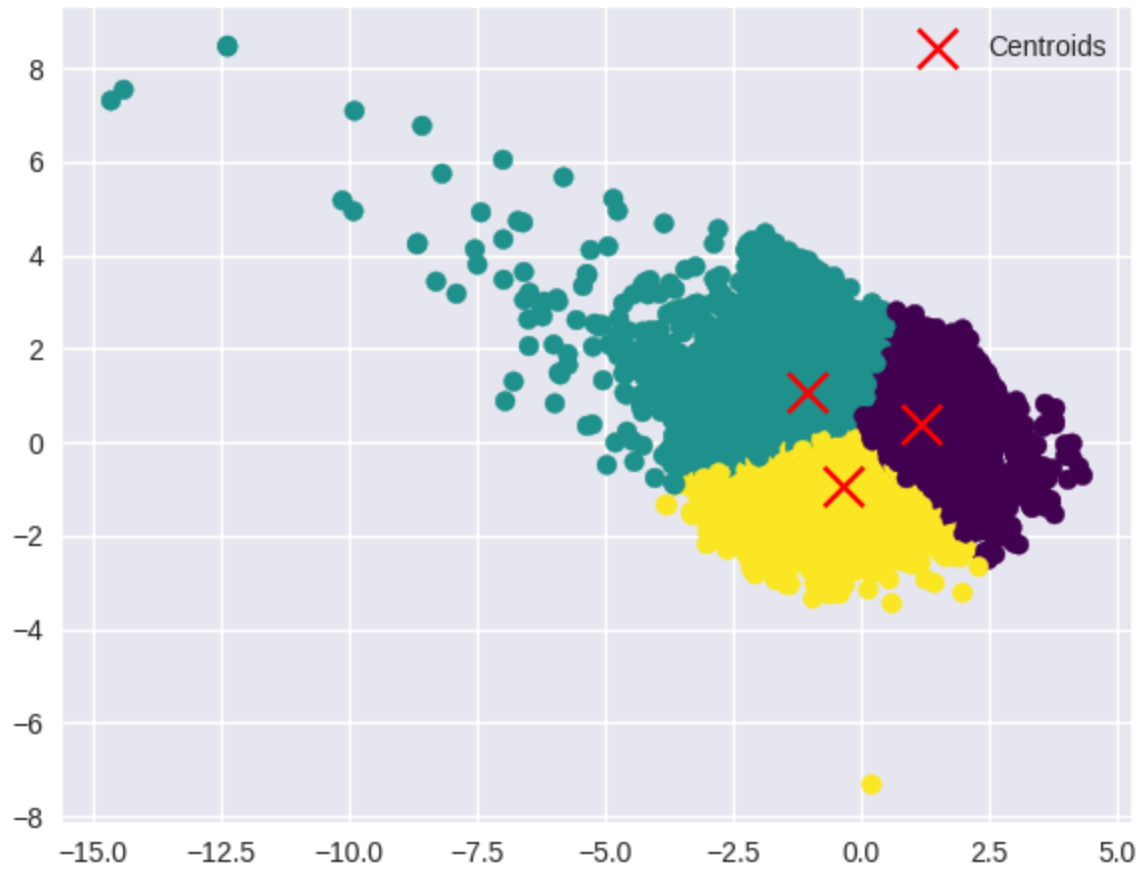
3. 'Inertia Plot' and 'Silhoutte Score Plot' for K-means:

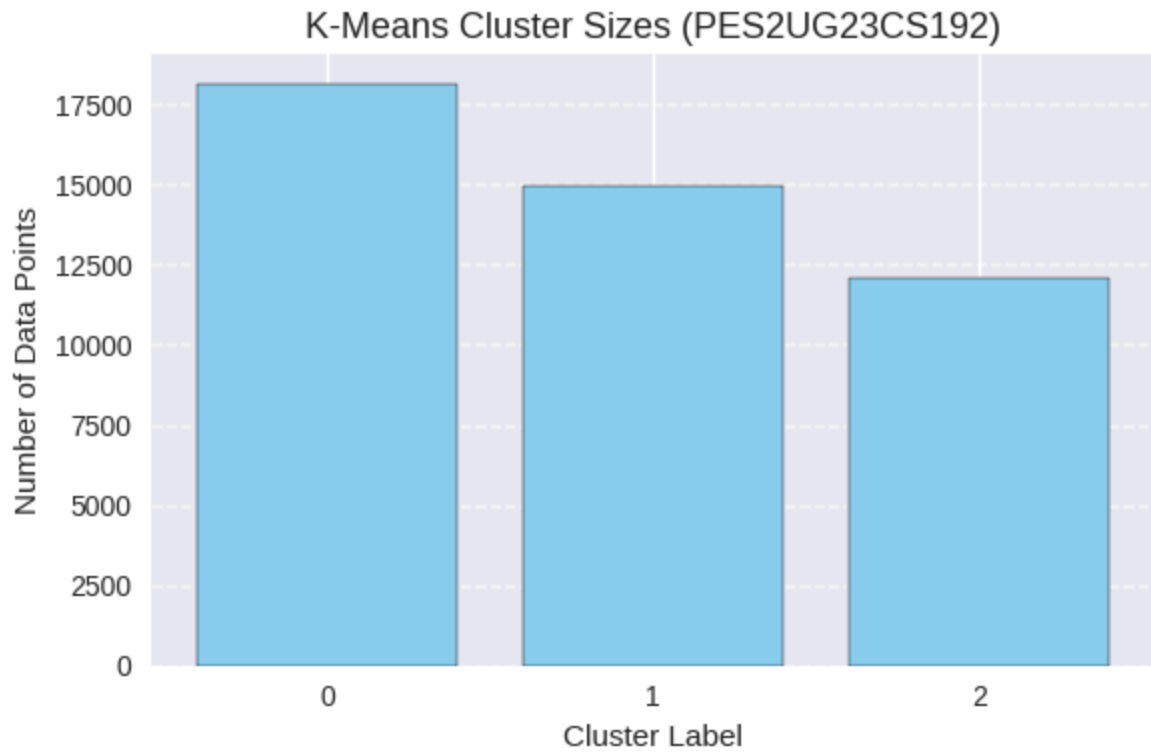




4. K-means Clustering Results with Centroids Visible:

Final Clustering Results





5. Silhouette distribution per cluster for K-means (Box Plot):

