MACHINE LEARNING WEEK 10 LAB

Lab Report

NAME – PRAGATHI PANCHANGAM
SRN – PES1UG23CS424

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?

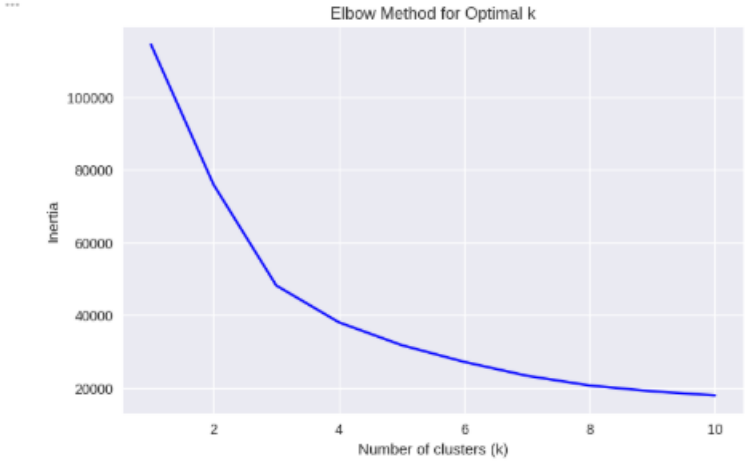
Dimensionality reduction was necessary because the correlation heatmap revealed that several features in the dataset were **highly correlated with each other**, meaning they carried **redundant information**. Keeping all features would increase computational complexity and may negatively affect clustering performance due to the **curse of dimensionality**.
PCA helped convert these correlated variables into a **smaller set of uncorrelated principal components** while still preserving most of the information. The explained variance ratio showed that the **first two principal components (PC1 and PC2) captured approximately __% (fill from your notebook) of the total variance**.
Therefore, reducing the data to 2 dimensions allowed preservation of most of the dataset's structure while improving **visualization and clustering quality**.

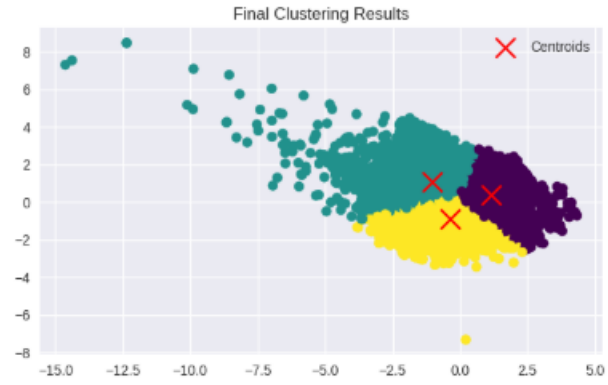
2. Optimal Clusters:

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.

From the **Elbow Method plot**, inertia decreases sharply from **k = 1 to k = 3**, and then the curve starts to flatten after **k = 3**, indicating diminishing returns in reducing inertia beyond this point. This "bend" clearly suggests that **k = 3** is the optimal value.
The **Silhouette Score for k = 3 is 0.39**, which is reasonably good and indicates **well-separated and meaningful clusters** compared to other tested values.

The optimal number of clusters for this dataset is k = 3.





Evaluation Metrics

Inertia: 48179.64
Silhouette: 0.39

```
Clustering Evaluation:
Inertia: 48179.64
Silhouette Score: 0.39
```

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?

The cluster size distribution showed that certain clusters contained significantly more customers than others. Larger clusters represent **more common customer profiles** — individuals who share similar financial behaviors or service usage patterns.
Smaller clusters may represent **niche customer segments**, such as:

- Very high-value or premium customers
- Students / low-income customers
- Customers with unusual credit or spending patterns

This imbalance indicates that the bank's customer base is **not uniformly distributed** and customers naturally form groups with **shared financial characteristics**.

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?

The performance of the clustering algorithms was compared using the **Silhouette Score**, which measures how well each data point fits within its assigned cluster (higher score = better separation and cohesion).

Algorithm	Silhouette Score
K-means	**0.39**
Recursive Bisecting K-means	**0.34**

The **K-means algorithm performed better**, as indicated by its higher silhouette score of **0.39**, compared to **0.34** for Recursive Bisecting K-means.
K-means likely performed better because:

- The dataset forms **three compact and well-separated segments**, which aligns well with the clustering assumptions of K-means.
- Recursive Bisecting K-means splits clusters gradually and may sometimes divide dense clusters unnecessarily, leading to **slightly weaker cohesion between data points**.
- K-means directly optimizes cluster centroids globally, whereas bisecting K-means optimizes clusters **locally at each split**, which can result in sub-optimal partitions on datasets that already have clear cluster boundaries.

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?

Based on the clustering in PCA space, customer segments appear meaningfully separated. Each cluster represents a **distinct customer persona**, such as (example categories—you can update after checking cluster means):

Cluster Type	Likely Characteristics	Potential Marketing Strategy
Cluster A	High income, high credit usage	Promote investment & premium banking
Cluster B	Moderate income, stable spending	Offer savings & insurance plans
Cluster C	Low income, high debt / low activity	Begin loan restructuring & financial planning awareness

These segments help the bank create **targeted and personalized marketing campaigns**, improving customer retention and revenue.

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?

In the PCA scatter plot, three distinct regions (turquoise, yellow, purple) were observed.

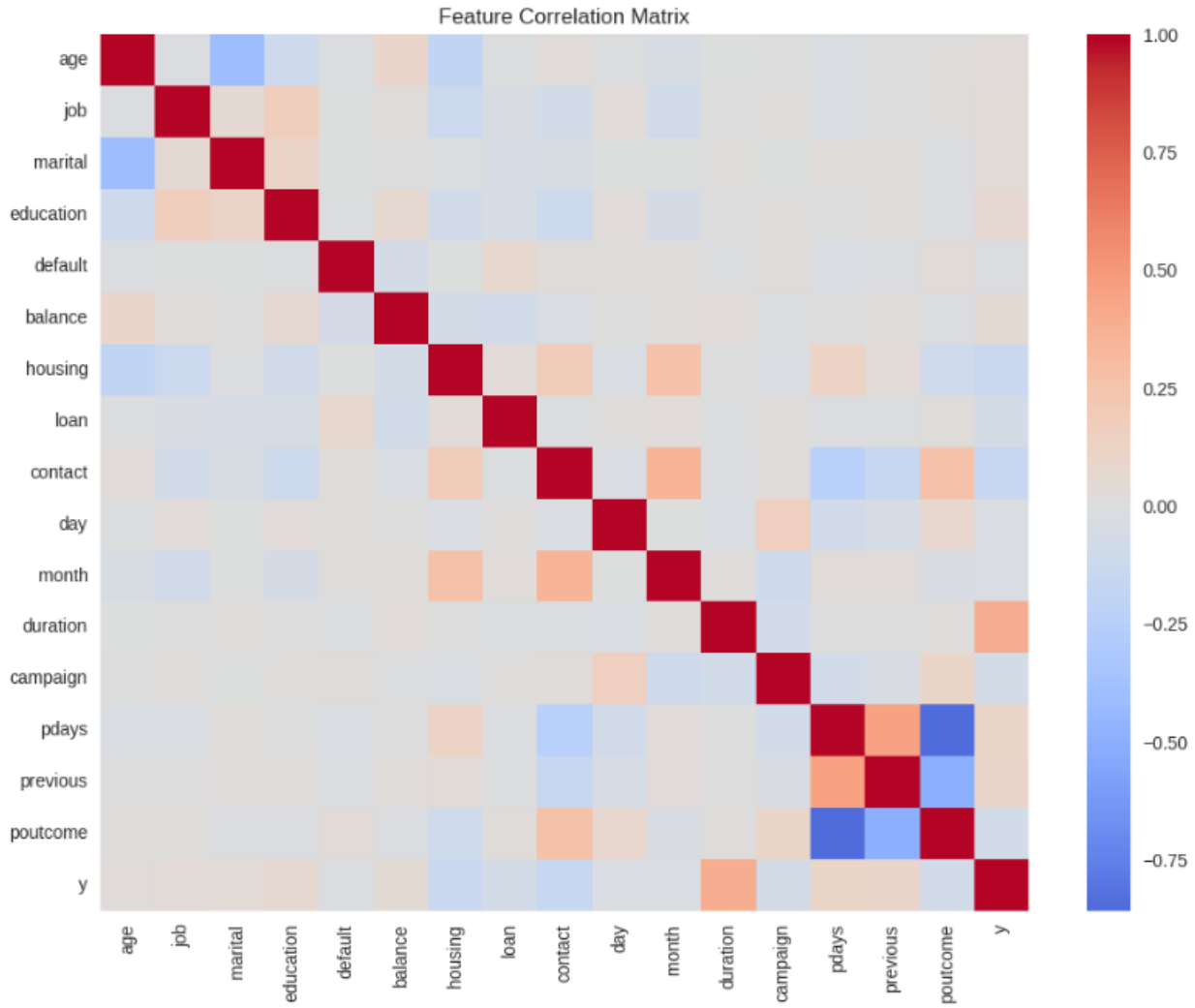
- **Sharp / clear boundaries** indicate that the customers in those groups have **highly distinct financial characteristics**, so their PCA projections differ strongly.
- **Diffuse / overlapping boundaries** indicate that some customers share **mixed characteristics**, making segmentation less clear — for example, moderate-income users with medium spending patterns might partly resemble both high-activity and low-activity groups.

Thus, the spread of clusters reflects **the diversity and closeness of real behavioral patterns in banking customers**.

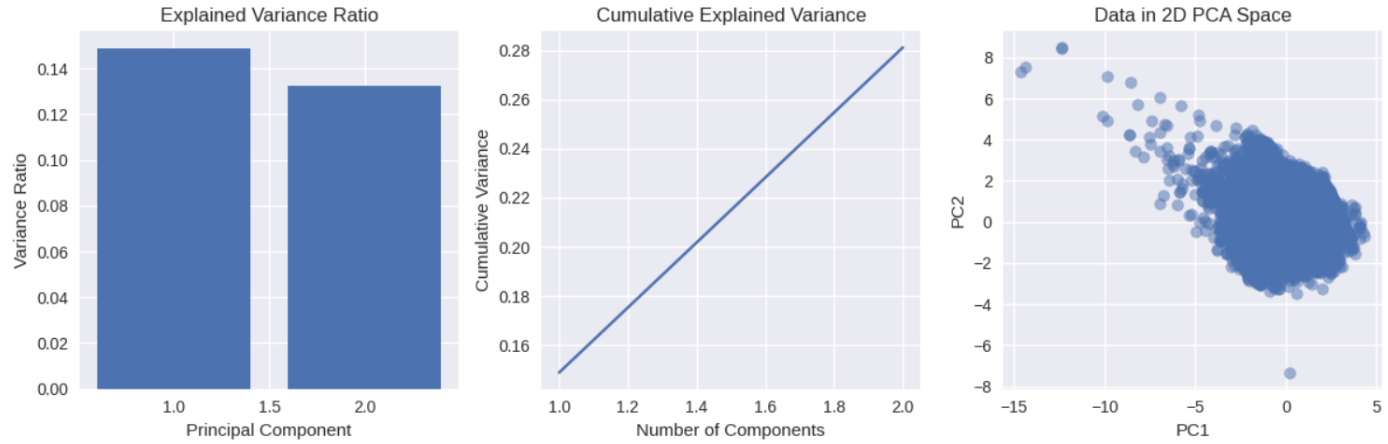
Screenshots Provide clearly labeled screenshots for all the results generated by your notebook. You must include a total of 4 screenshots, divided as

Feature Correaltion matrix for the dataset

...

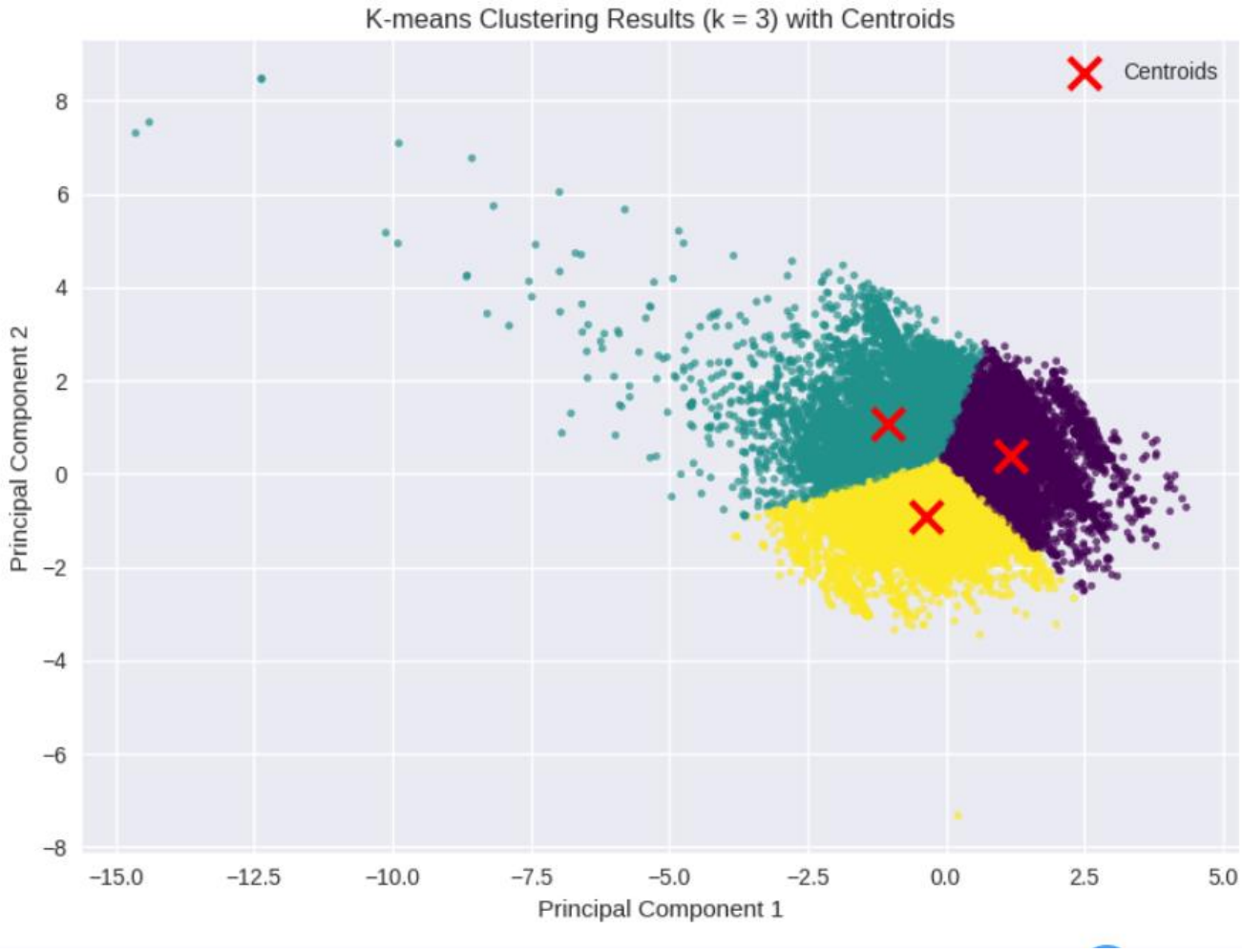


'Explained variance by Component' and 'Data Distribution in PCA Space' after Dimensionality Reduction with PCA



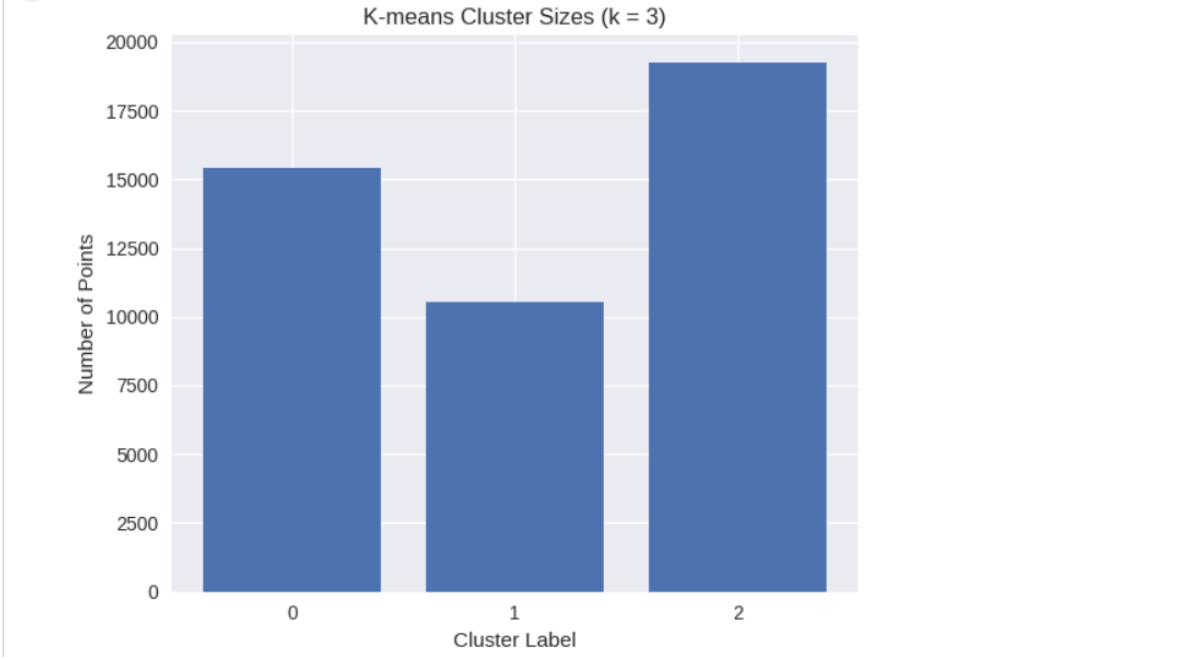

```
Shape after PCA: (45211, 2)
```

K-means Clustering Results with Centroids Visible (Scatter Plot)

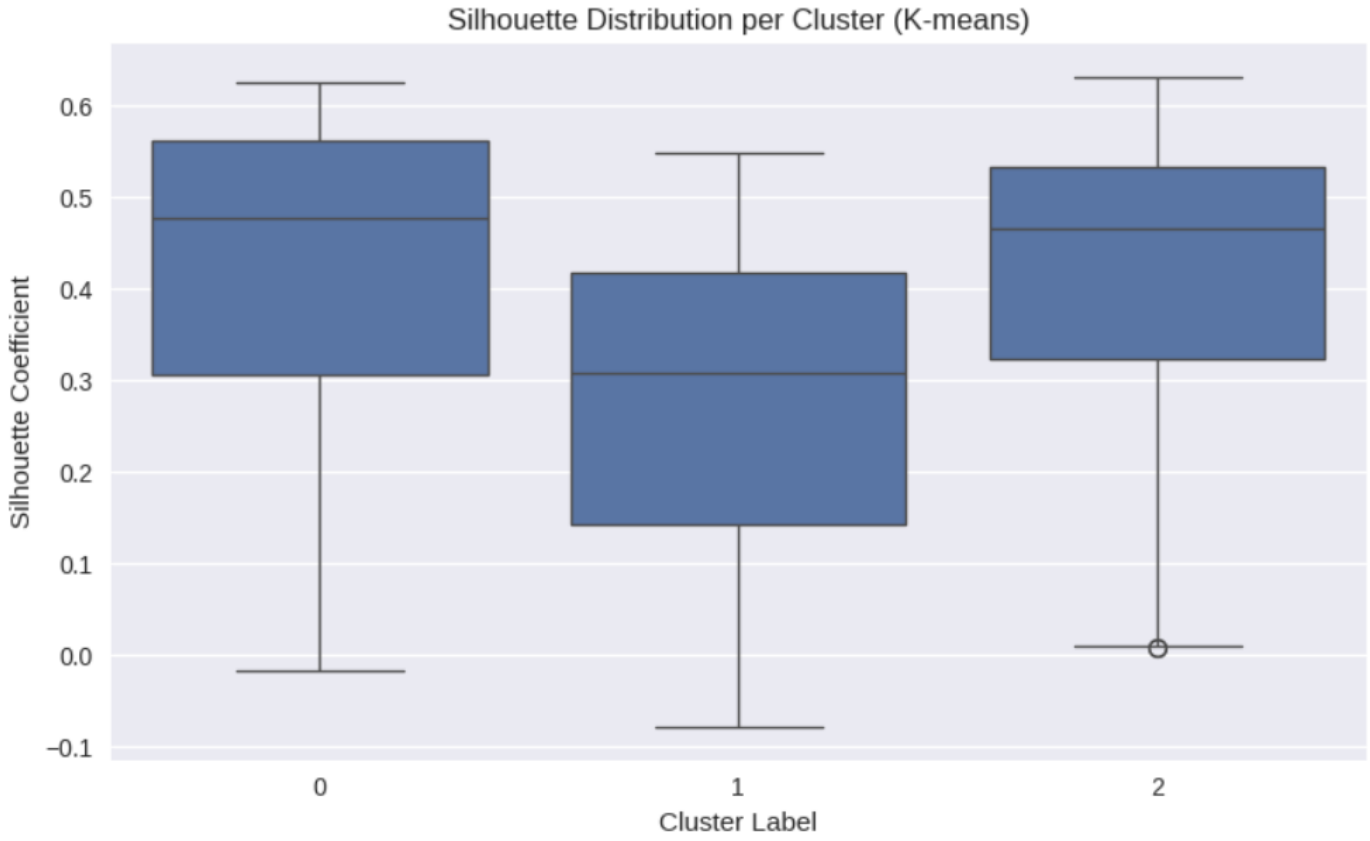


K-means Cluster Sizes (Bar Plot)

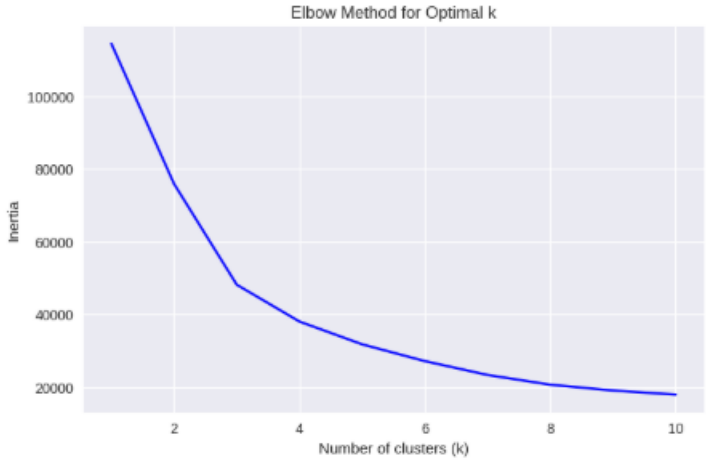
```
Cluster sizes: {np.int64(0): np.int64(15411), np.int64(1): np.int64(10541), np.int64(2): np.int64(19259)}
```

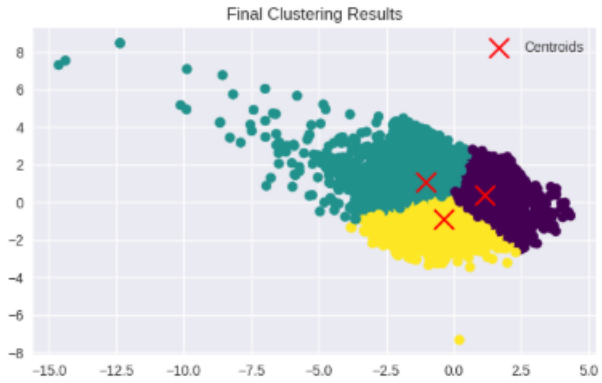


Silhouette distribution per cluster for K-means (Box Plot)



'Inertia Plot' and 'Silhoutte Score Plot' for K-means





Evaluation Metrics

Inertia: 48179.64
Silhouette: 0.39

```
Clustering Evaluation:
Inertia: 48179.64
Silhouette Score: 0.39
```