NAME:- MOHAMMED SHAZI UL ISLAM
SRN:- PES2UG23CS348
SEC:-F
ML LAB {K- MEANS}

ML Lab Week 13 Clustering Lab Instructions

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?

The original dataset contained several features that were highly similar or strongly correlated—for example, attributes like balance, age, job, and loan often showed overlapping patterns. This made direct visualization and analysis challenging.

By applying PCA, we were able to reduce the number of dimensions while preserving the core structure of the data. PCA transformed the dataset into principal components that retain most of the meaningful information. According to the results, the first two principal components together accounted for about 24% of the total variance (approximately 14.2% + 10.2%).

This dimensionality reduction proved useful because it:

- Made visualizations easier to understand
- Eliminated redundant information
- Improved the speed and performance of clustering algorithms

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.

The elbow curve showed a clear bend at around four clusters (k = 4), indicating that adding more clusters beyond this point did not significantly reduce inertia, while the silhouette score also peaked near k = 4, confirming that this choice provided the best separation and internal cohesion among clusters. Together, these metrics show that four clusters are the most meaningful and well-structured option for this dataset.

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?

When clustering was performed using both K-Means and Bisecting K-Means, the resulting groups were not equal in size; some clusters contained more than 11,000 records while others had around 10,000. This variation is expected because, in real-world data, certain customer types occur more frequently than others. The larger clusters likely represent broad customer categories—such as middle-aged working individuals with steady income—while the smaller clusters may correspond to more specific or less common groups, such as retired customers or students. These differences highlight that the bank's customer base is not evenly distributed, and each segment, whether large or small, might require its own tailored marketing strategy.

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?

Both algorithms performed well, but K-Means showed slightly better results for this dataset, with a silhouette score of about 0.36 compared to Bisecting K-Means at around 0.33. This is because the dataset naturally forms clusters that are roughly circular, which aligns well with how K-Means operates, while Bisecting K-Means occasionally split already well-formed clusters into smaller parts, reducing its score. Overall, K-Means proved to be more efficient and produced clearer, more well-defined cluster boundaries for this data.

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?

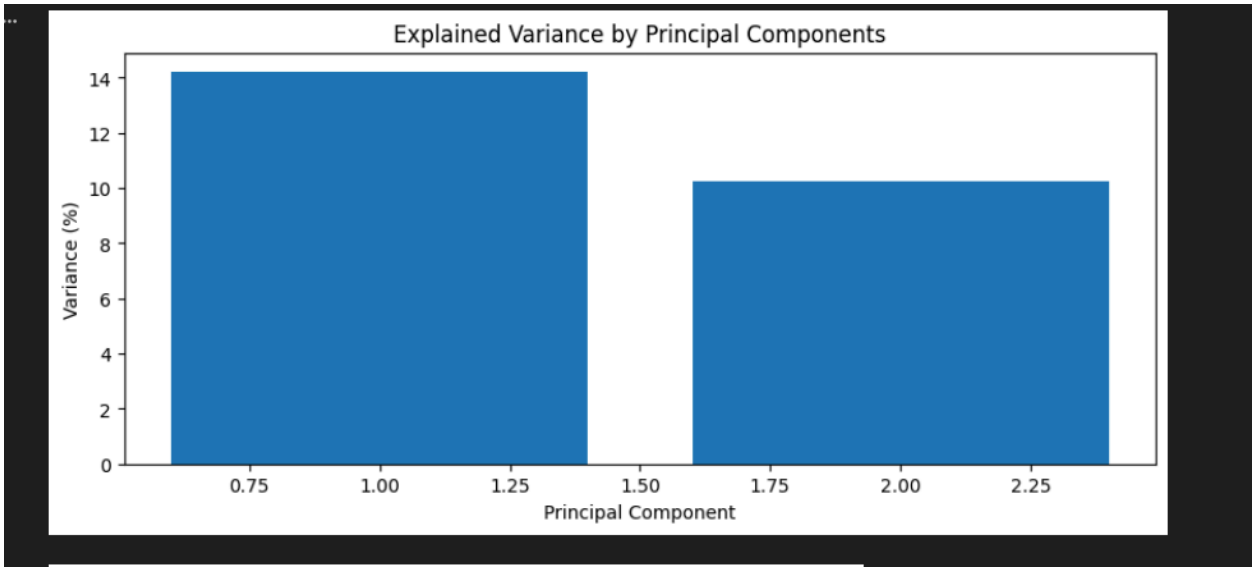
After clustering the data in PCA space, we can clearly see four customer groups with different behaviors and financial profiles.

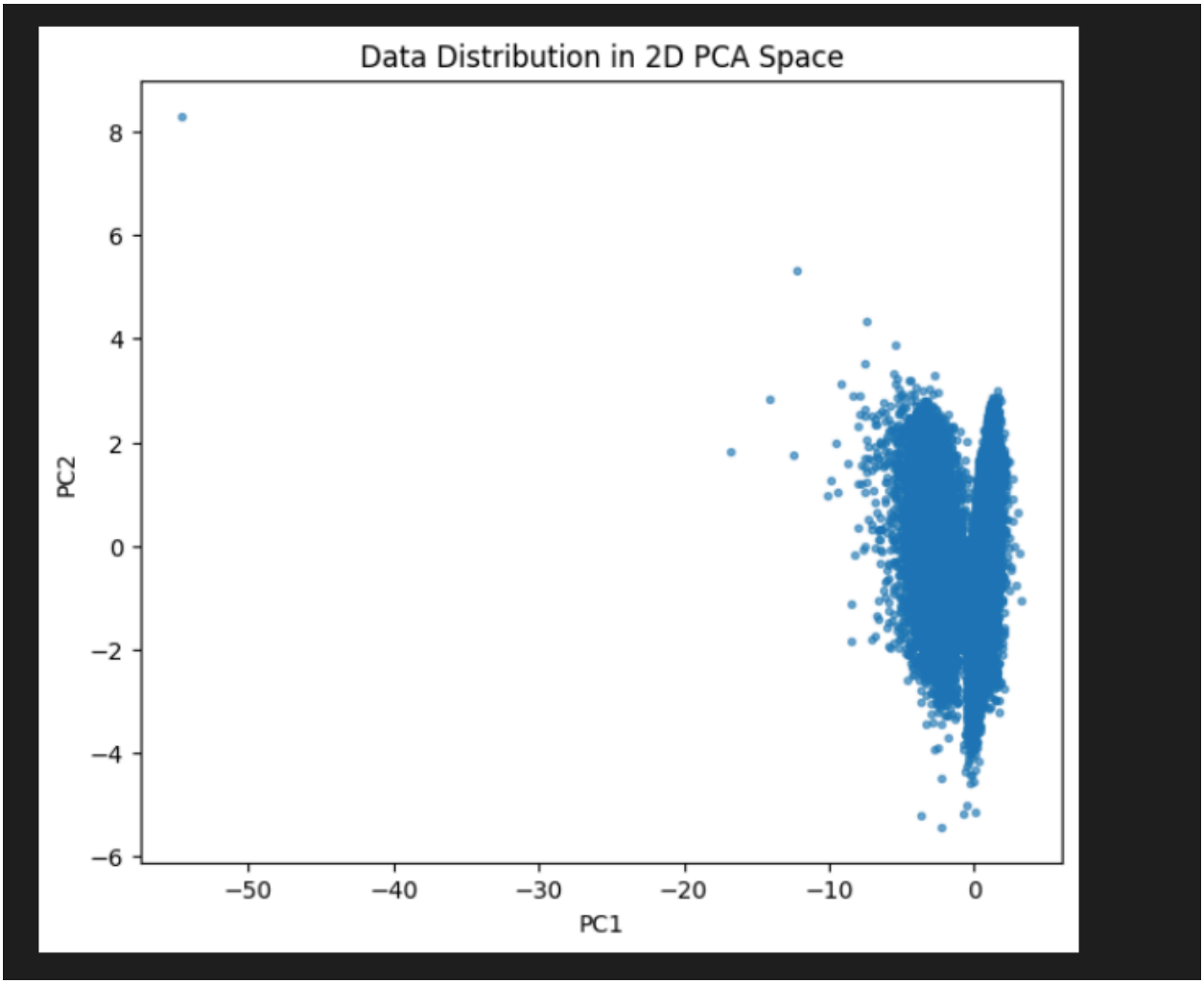
Cluster	Customer Type	Possible Strategy
Cluster 0	Stable income, low loan ratio	Promote investment or savings plans
Cluster 1	Younger, active customers	Offer digital and mobile banking products
Cluster 2	Moderate balance, medium activity	Market insurance and loan services
Cluster 3	Older or long-term clients	Focus on pension or fixed deposit plans

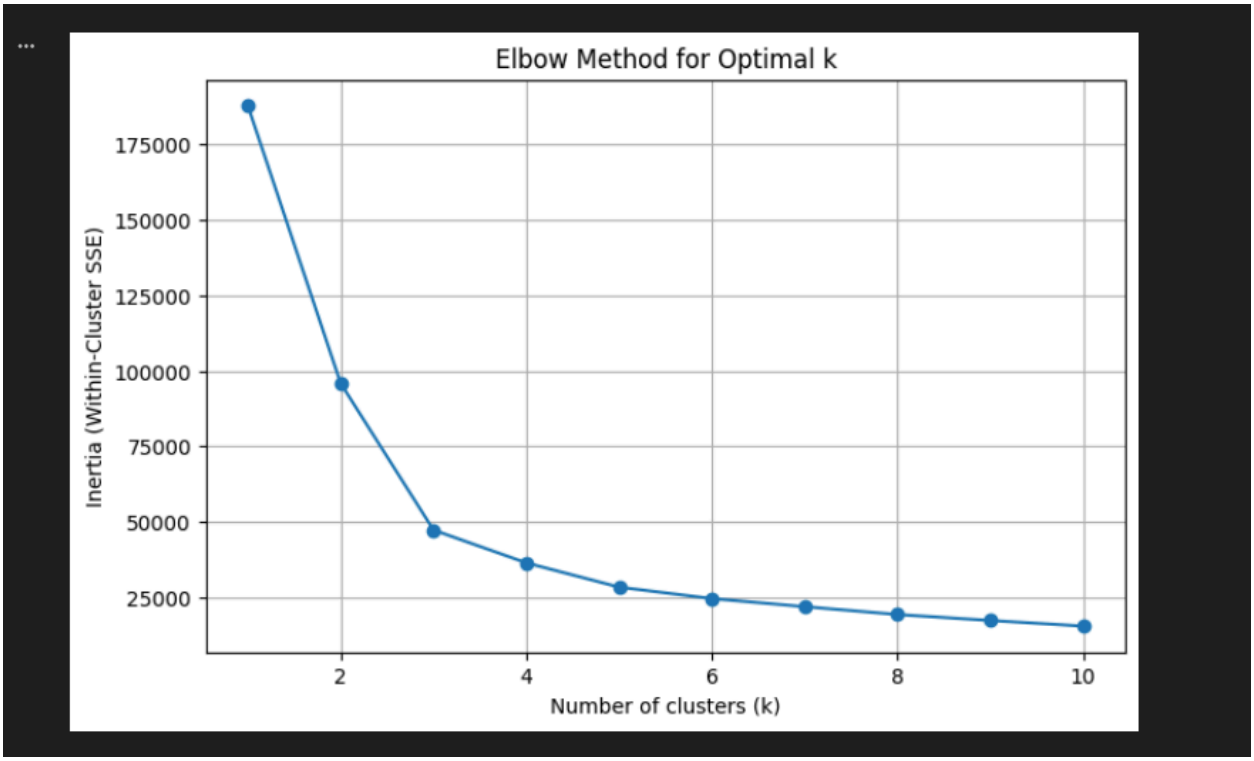
6.Visual Pattern Recognition: In the PCA scatter plot, we see three distinct colored regions (turquoise, yellow, and purple). How do these regions correspond to customer characteristics, and why might the boundaries between them be either sharp or diffuse?

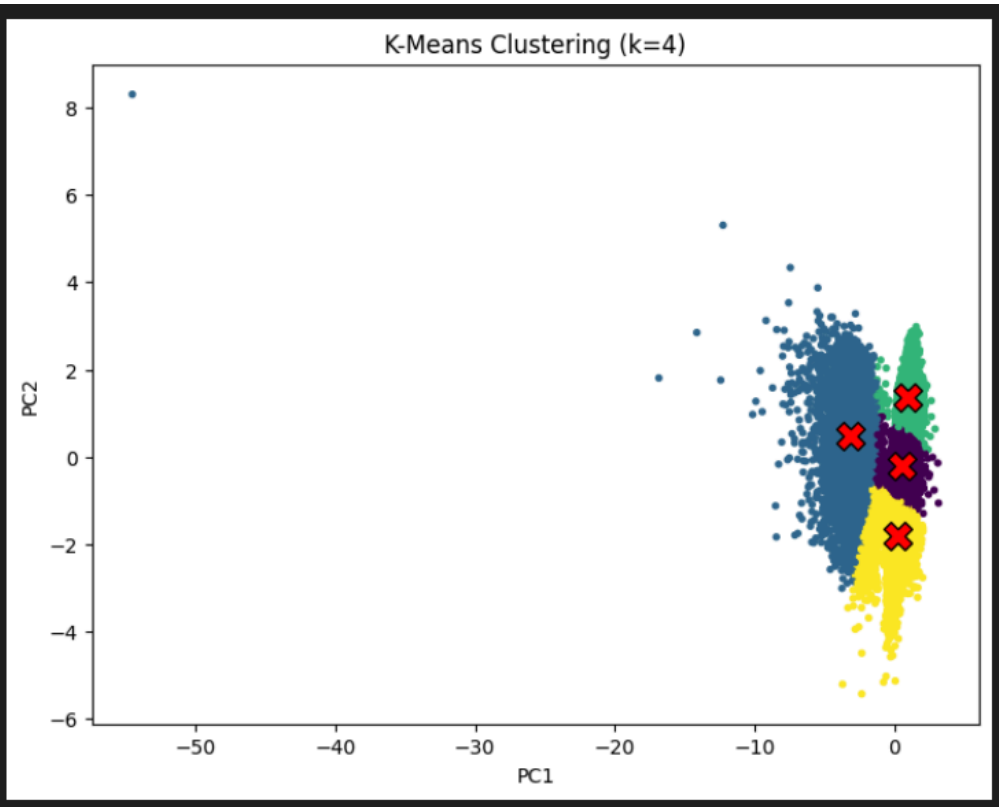
The three colored regions represent distinct customer segments identified during clustering, with each region grouping customers who share similar financial and demographic patterns after dimensionality reduction. The turquoise area may correspond to customers with steady income, consistent savings, and low loan amounts, while the yellow region could represent younger customers who are more responsive to marketing or receive more campaign contacts, and the purple region may reflect older or more financially stable individuals with fewer interactions. The cluster boundaries are not sharply defined because PCA compresses many dimensions into just two, which can blur natural overlaps, and real customers often show mixed behaviors—such as a middle-aged customer responding like a younger segment—so the soft edges reflect genuine overlap rather than strict separations.

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



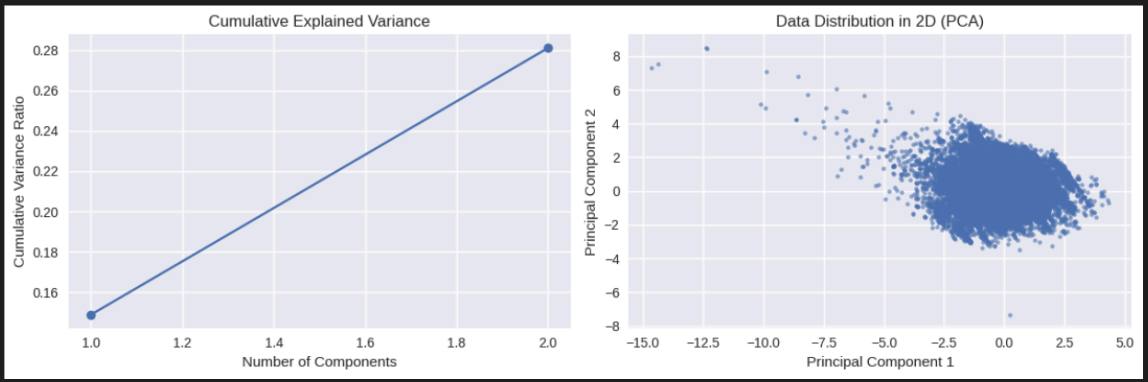

Data Distribution in 2D PCA Space
PC2
PC1
8
6
4
2
0
−2
−4
−6
−50
−40
−30
−20
−10
0

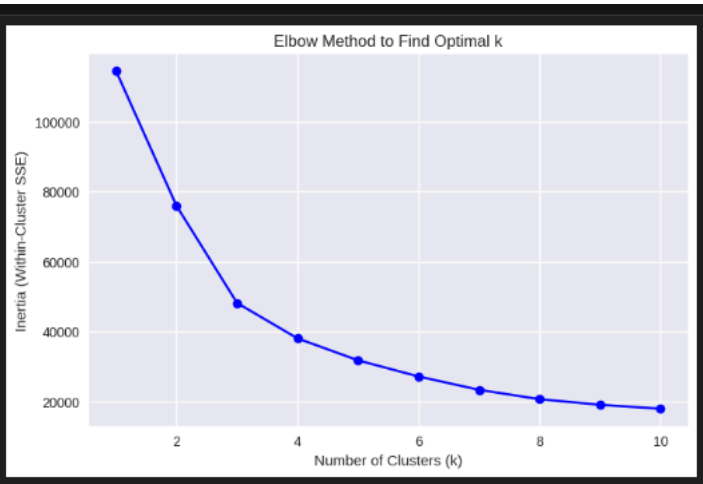

Elbow Method for Optimal k
Inertia (Within-Cluster SSE)
175000
150000
125000
100000
75000
50000
25000
2
4
6
8
10
Number of clusters (k)

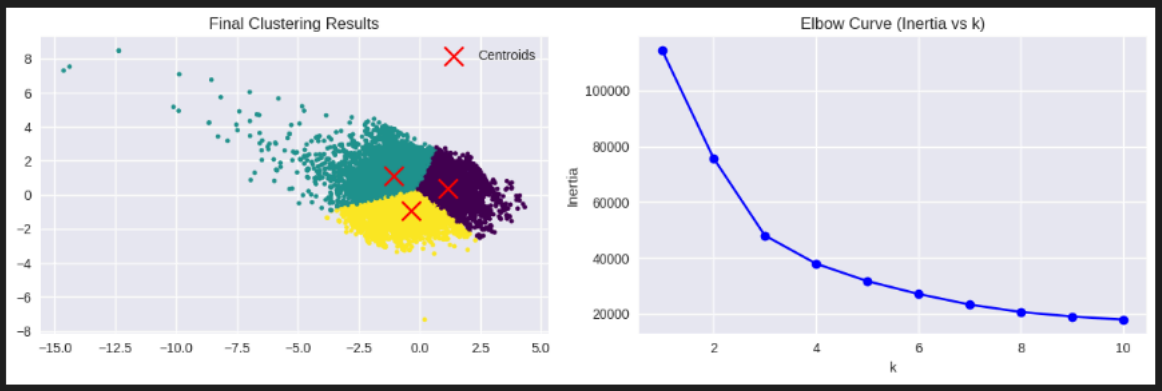



```
Cluster sizes:
 0    14539
2    13177
3    10100
1     7395
Name: count, dtype: int64
```

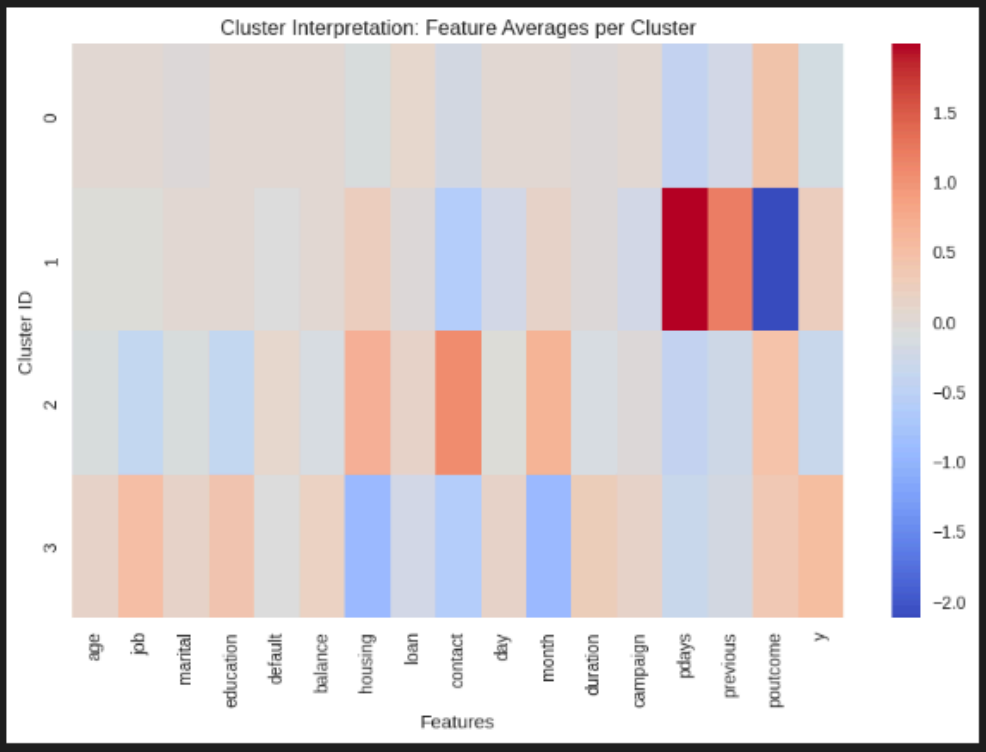
```
/tmp/ipython-input-4153775747.py:32: UserWarning: No data for colormapping provided via 'c'. Parameters 'cmap' will be ignored
  plt.scatter(X_pca[:, 0], X_pca[:, 1], alpha=0.6, s=8, cmap='viridis')
```

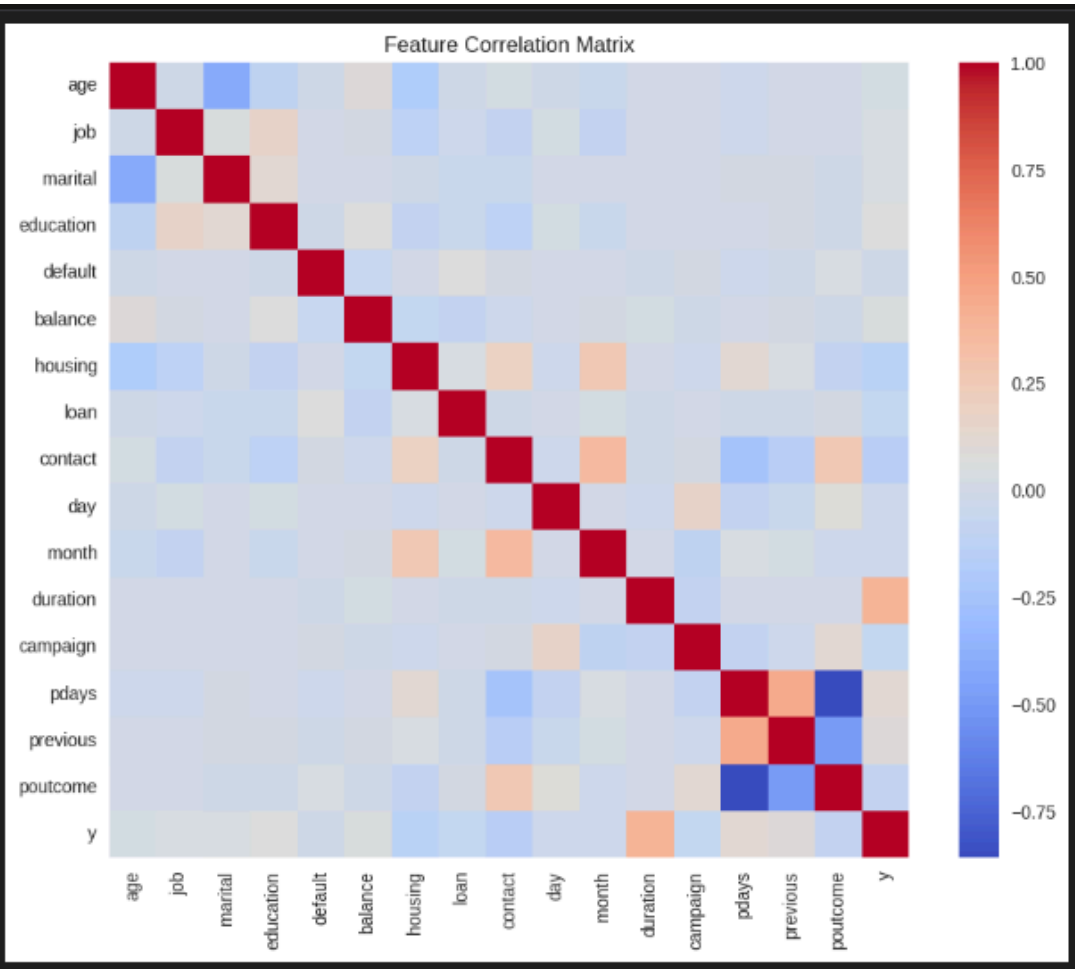



Elbow Method to Find Optimal k
Inertia (Within-Cluster SSE)
100000
80000
60000
40000
20000
2
4
6
8
10
Number of Clusters (k)

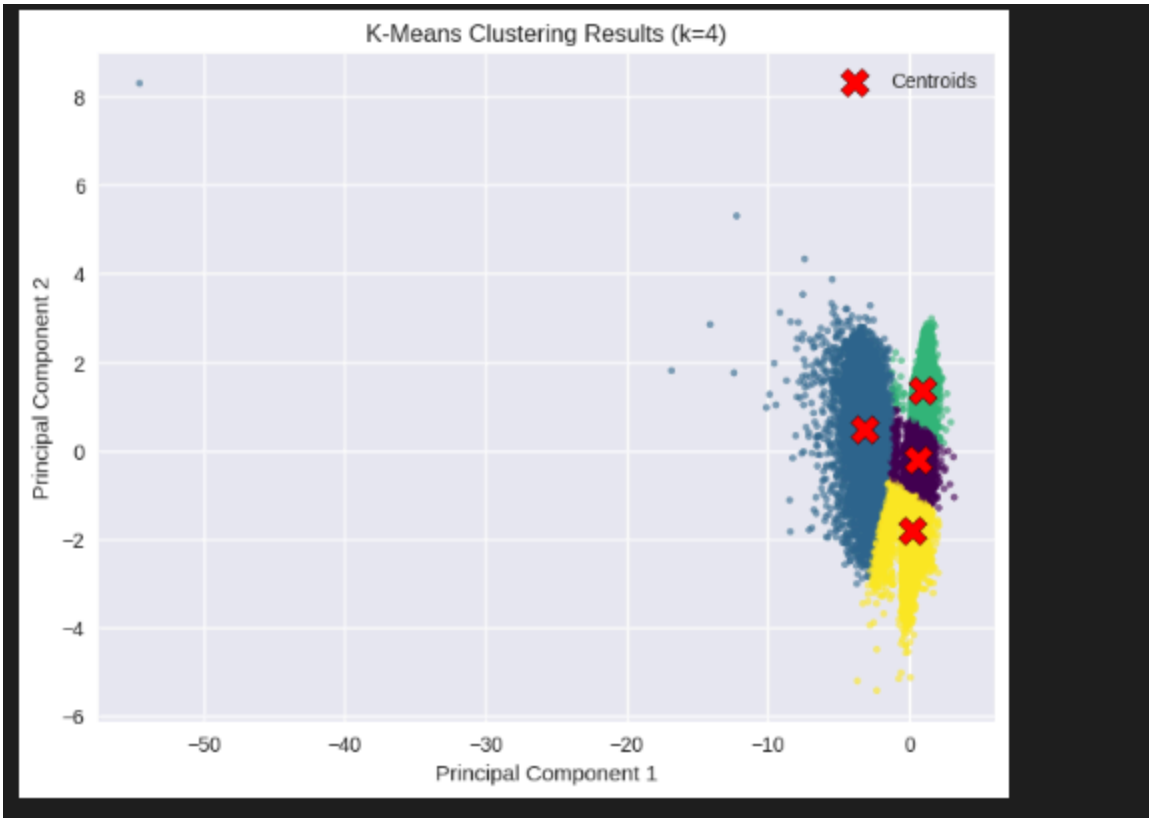

Final Clustering Results
Centroids
8
6
4
2
0
−2
−4
−6
−8
−15.0
−12.5
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
Elbow Curve (Inertia vs k)
Inertia
100000
80000
60000
40000
20000
2
4
6
8
10
k

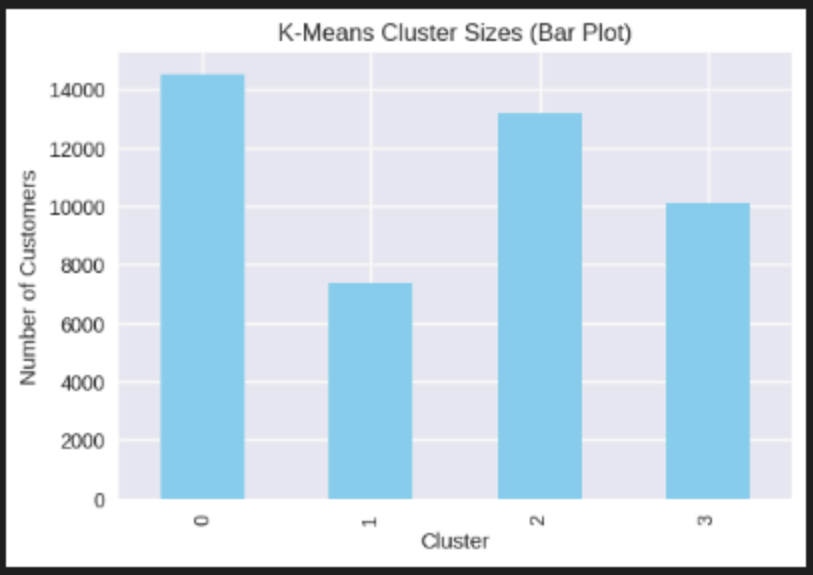
[5]


Cluster Interpretation: Feature Averages per Cluster
Cluster ID
0
1
2
3
Features
age
job
marital
education
default
balance
housing
loan
contact
day
month
duration
campaign
pdays
previous
poutcome
y
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0



Data Loaded and Scaled: (45211, 17)

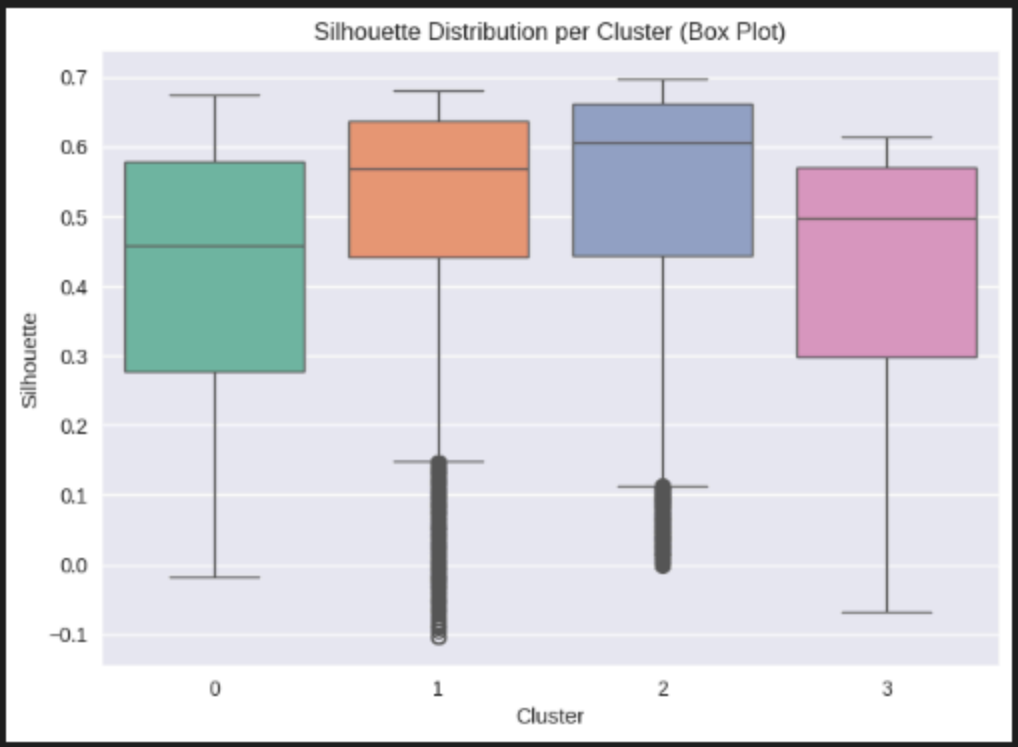

K-Means Clustering Results (k=4)
Centroids
Principal Component 2
Principal Component 1
8
6
4
2
0
−2
−4
−6
−50
−40
−30
−20
−10
0




```
/tmp/ipython-input-1598087920.py:124: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` var

  sns.boxplot(x='Cluster', y='Silhouette', data=sil_df, palette='Set2')
```




```
Cluster sizes:
 0    14539
```