ML Lab Week 13
Clustering Lab

Name: Cheruku Manas Ram
SRN: PES2UG23CS147
Section: 5 ‘C’
Date: 11/11/2025

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?

A. Dimensionality reduction using PCA was beneficial for this dataset because it helps to reduce the number of features while retaining most of the important information. This can be particularly useful when dealing with potentially correlated features, as PCA transforms them into a set of uncorrelated principal components. Reducing the dimensionality can also help with visualizing the data and can sometimes improve the performance of clustering algorithms by reducing noise and the effects of the "curse of dimensionality." The first two principal components capture approximately 30% of the total variance in the data.

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.

A. **Elbow Curve:** The elbow method suggests looking for the "elbow point" in the plot of inertia vs. the number of clusters (k). This point is where the rate of decrease in inertia slows down significantly, indicating that adding more clusters beyond this point doesn't provide much benefit in reducing within-cluster variance. Looking at the elbow curve, there is a bend around **3 or 4 clusters**.

   Silhouette Score: The silhouette score measures how similar an object is to its own cluster compared to other clusters. A higher silhouette score indicates better-defined clusters. The plot shows the silhouette score is approximately **0.39**. While not extremely high, it's a positive score, suggesting that the data points are reasonably well-assigned to their clusters.

Considering both metrics, a choice of **3 or 4 clusters** appears reasonable. The elbow curve shows a potential elbow around these values, and the silhouette score for 3 clusters is positive.

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?

A. The cluster size distributions for K-means and Bisecting K-means are different. In both cases, some clusters are larger than others because the underlying data has varying densities and structures. These differences in cluster sizes suggest that

certain customer segments are more prevalent than others in the dataset. Analyzing the features within each cluster would reveal the specific characteristics defining these larger and smaller customer groups.

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?

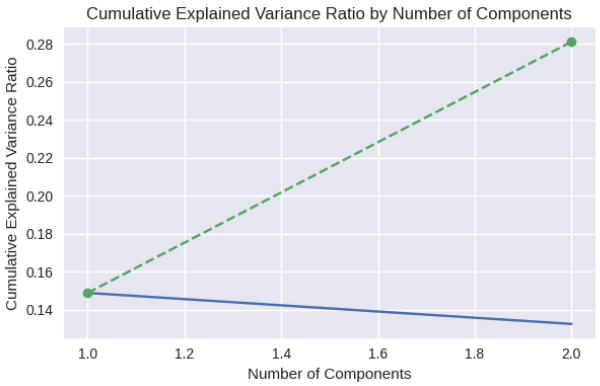
A. Based on the silhouette scores calculated, K-means with a score of 0.39 performed better than Bisecting K-means with a score of 0.34 for this dataset. A higher silhouette score indicates that the clusters are better separated and the data points are more tightly grouped within their assigned clusters. The standard K-means likely found a more optimal partitioning of the data in this specific case, potentially due to the nature of the clusters or the initialization method used in the Bisecting K-means implementation.

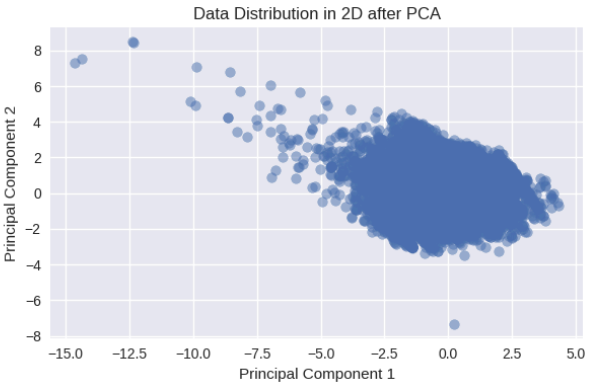
5. Business Insights: Based on the clustering results in the PCA space, what insights can you draw about customer segmentation that might be valuable for the bank's marketing strategy?

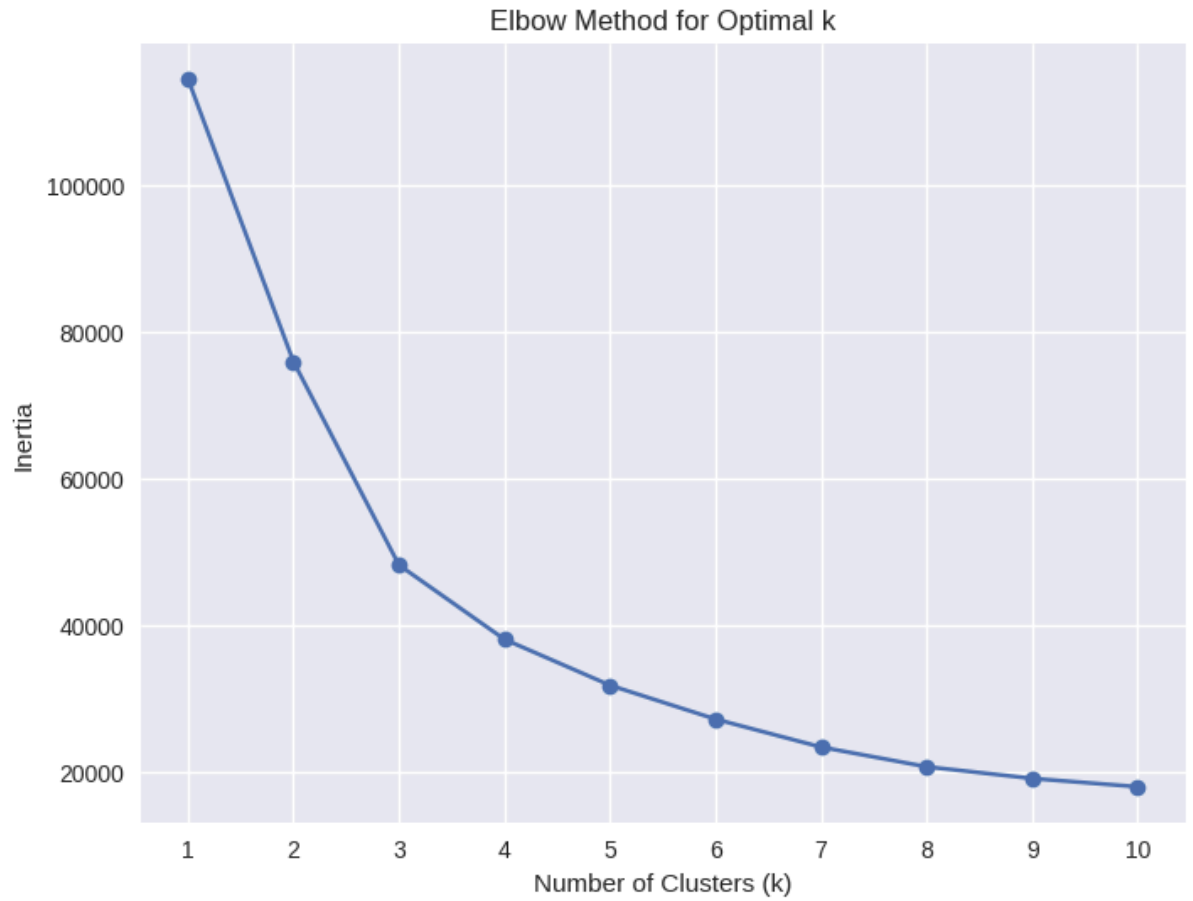
A. Based on the clustering results visualized in the PCA space, we can observe distinct groupings of customers. These clusters represent different customer segments with unique characteristics. To draw valuable insights for the bank's marketing strategy, we would need to further analyse the original features (like age, balance, job, etc.) of the customers within each cluster. For example, one cluster might represent younger customers with lower balances, while another could consist of older, wealthier individuals. Understanding these differences will allow the bank to tailor marketing campaigns, products, and services to the specific needs and preferences of each segment, potentially leading to increased engagement and profitability.

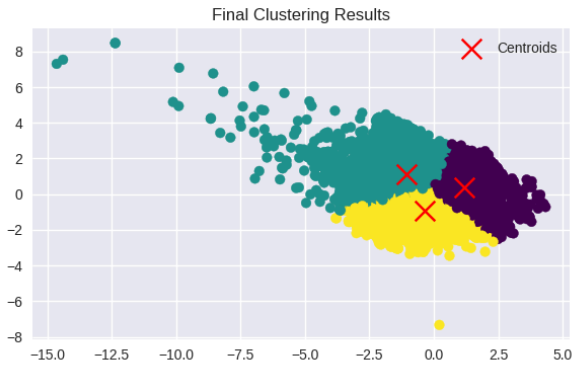
6. Visual Pattern Recognition: In the PCA scatter plot, we see three distinct colored regions (turquoise, yellow, and purple). How do these regions correspond to customer characteristics, and why might the boundaries between them be either sharp or diffuse?

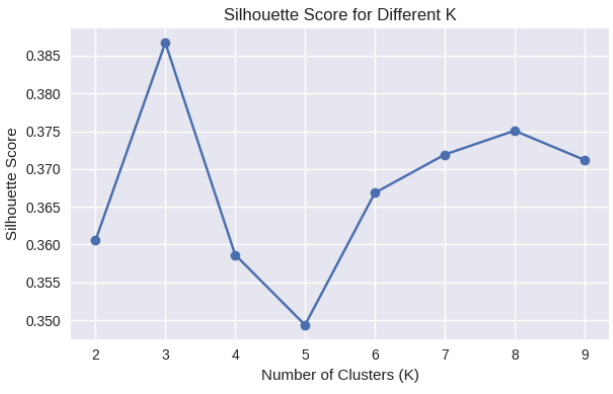
A. In the PCA scatter plot, the three distinct colored regions (turquoise, yellow, and purple) represent the customer clusters identified by the K-means algorithm. Each color corresponds to a different cluster. To understand how these regions correspond to specific customer characteristics, you would need to analyze the original features of the data points within each cluster (e.g., what are the typical age, job, and balance of customers in the turquoise cluster?). The boundaries between these clusters might appear sharp if there are clear distinctions in characteristics between groups. The boundaries might be diffuse if there is overlap in the feature space, indicating that some customers share characteristics across different segments, making the separation less distinct.

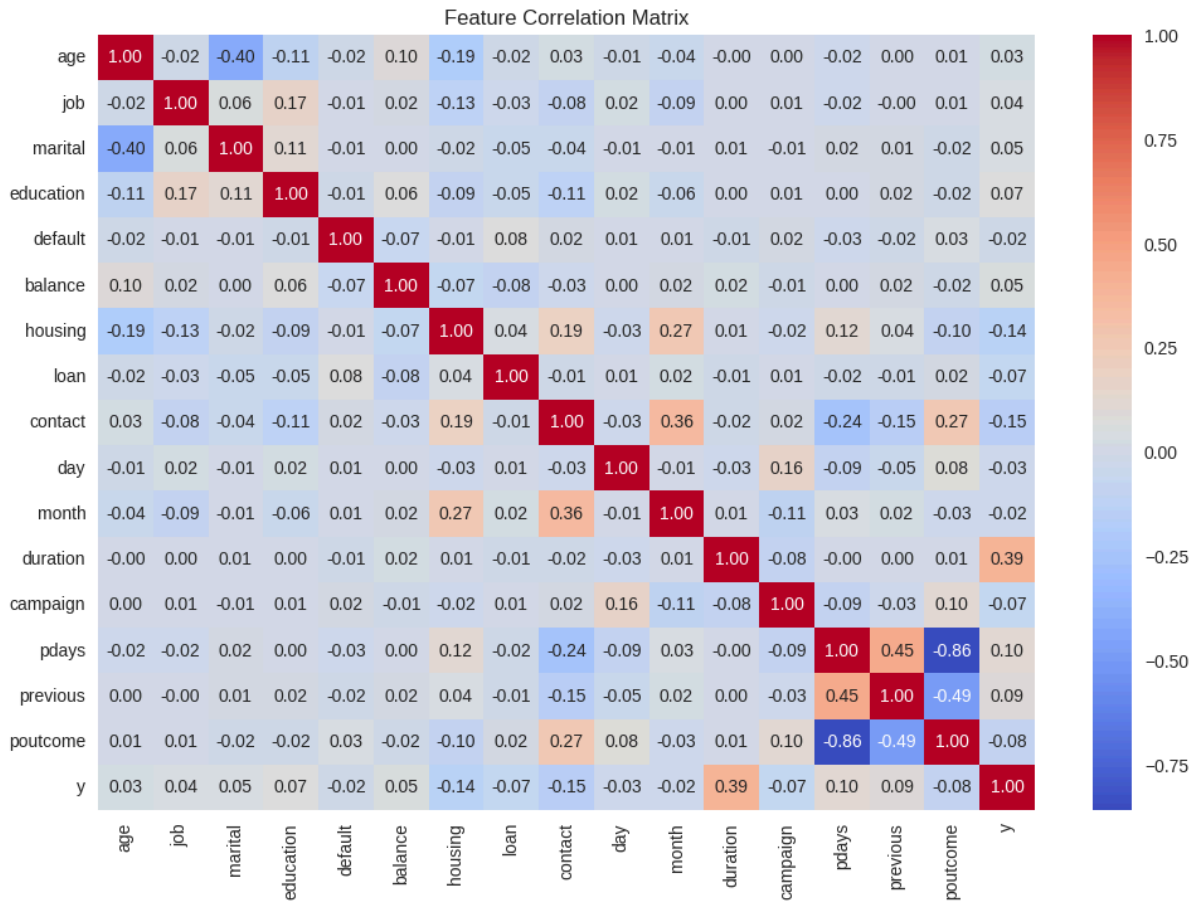

Cumulative Explained Variance Ratio by Number of Components
Cumulative Explained Variance Ratio
Number of Components

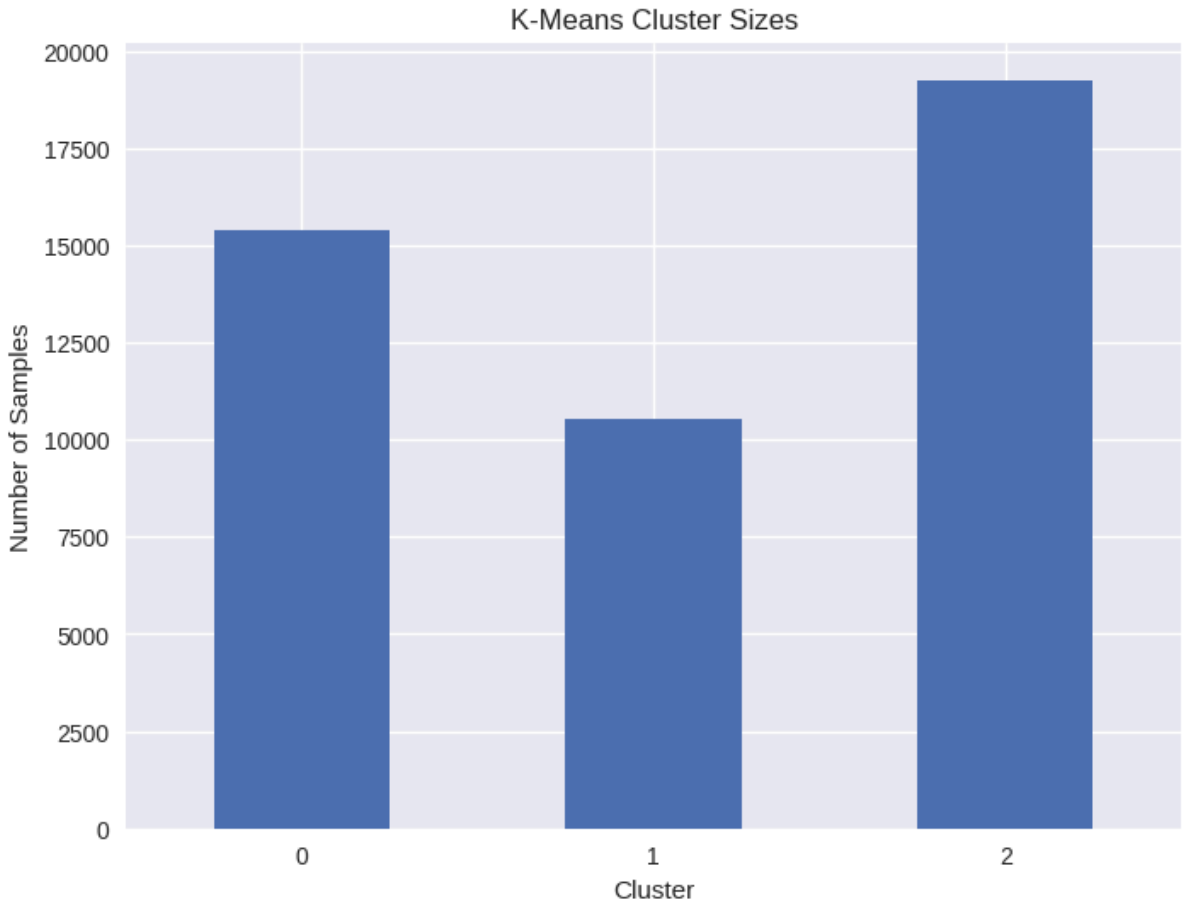

Data Distribution in 2D after PCA
Principal Component 2
Principal Component 1

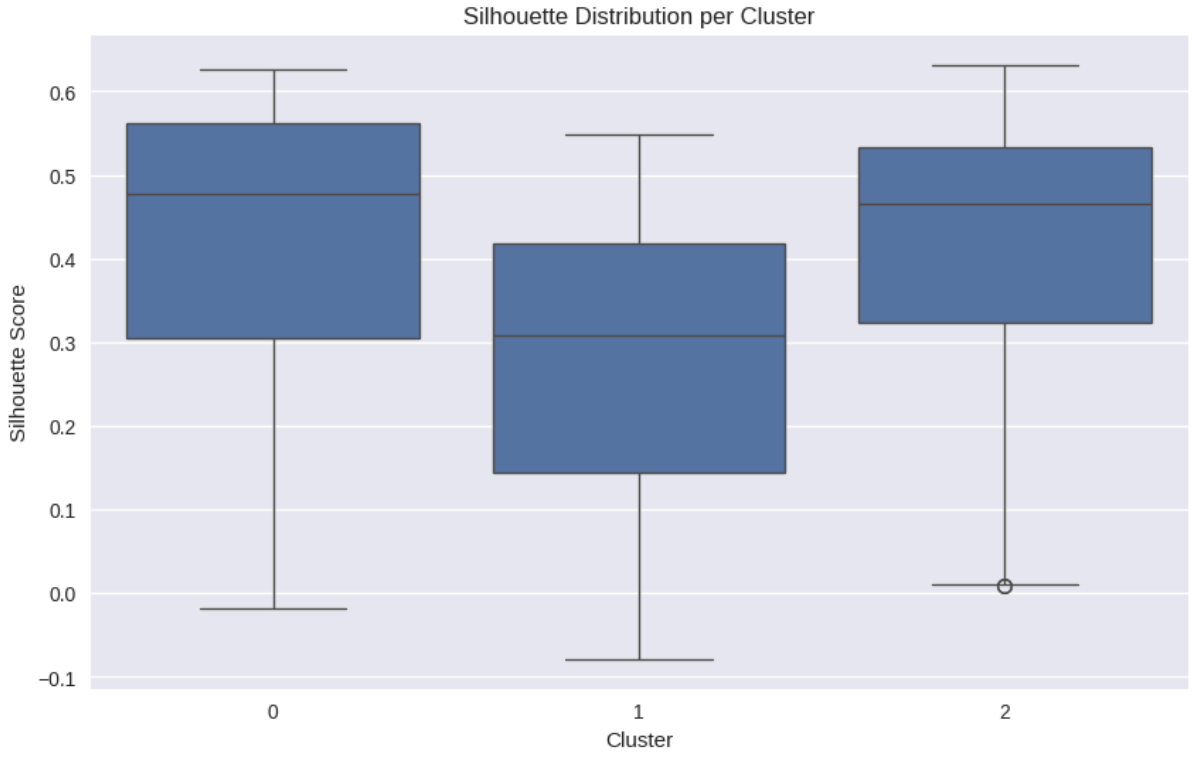

Elbow Method for Optimal k
Inertia
Number of Clusters (k)


Final Clustering Results
Centroids


Silhouette Score for Different K
Silhouette Score
Number of Clusters (K)


Feature Correlation Matrix
age 1.00 -0.02 -0.40 -0.11 -0.02 0.10 -0.19 -0.02 0.03 -0.01 -0.04 -0.00 0.00 -0.02 0.00 0.01 0.03
job -0.02 1.00 0.06 0.17 -0.01 0.02 -0.13 -0.03 -0.08 0.02 -0.09 0.00 0.01 -0.02 -0.00 0.01 0.04
marital -0.40 0.06 1.00 0.11 -0.01 0.00 -0.02 -0.05 -0.04 -0.01 -0.01 0.01 -0.01 0.02 0.01 -0.02 0.05
education -0.11 0.17 0.11 1.00 -0.01 0.06 -0.09 -0.05 -0.11 0.02 -0.06 0.00 0.01 0.00 0.02 -0.02 0.07
default -0.02 -0.01 -0.01 -0.01 1.00 -0.07 -0.01 0.08 0.02 0.01 0.01 -0.01 0.02 -0.03 -0.02 0.03 -0.02
balance 0.10 0.02 0.00 0.06 -0.07 1.00 -0.07 -0.08 -0.03 0.00 0.02 0.02 -0.01 0.00 0.02 -0.02 0.05
housing -0.19 -0.13 -0.02 -0.09 -0.01 -0.07 1.00 0.04 0.19 -0.03 0.27 0.01 -0.02 0.12 0.04 -0.10 -0.14
loan -0.02 -0.03 -0.05 -0.05 0.08 -0.08 0.04 1.00 -0.01 0.01 0.02 -0.01 0.01 -0.02 -0.01 0.02 -0.07
contact 0.03 -0.08 -0.04 -0.11 0.02 -0.03 0.19 -0.01 1.00 -0.03 0.36 -0.02 0.02 -0.24 -0.15 0.27 -0.15
day -0.01 0.02 -0.01 0.02 0.01 0.00 -0.03 0.01 -0.03 1.00 -0.01 -0.03 0.16 -0.09 -0.05 0.08 -0.03
month -0.04 -0.09 -0.01 -0.06 0.01 0.02 0.27 0.02 0.36 -0.01 1.00 0.01 -0.11 0.03 0.02 -0.03 -0.02
duration -0.00 0.00 0.01 0.00 -0.01 0.02 0.01 -0.01 -0.02 -0.03 0.01 1.00 -0.08 -0.00 0.00 0.01 0.39
campaign 0.00 0.01 -0.01 0.01 0.02 -0.01 -0.02 0.01 0.02 0.16 -0.11 -0.08 1.00 -0.09 -0.03 0.10 -0.07
pdays -0.02 -0.02 0.02 0.00 -0.03 0.00 0.12 -0.02 -0.24 -0.09 0.03 -0.00 -0.09 1.00 0.45 -0.86 0.10
previous 0.00 -0.00 0.01 0.02 -0.02 0.02 0.04 -0.01 -0.15 -0.05 0.02 0.00 -0.03 0.45 1.00 -0.49 0.09
poutcome 0.01 0.01 -0.02 -0.02 0.03 -0.02 -0.10 0.02 0.27 0.08 -0.03 0.01 0.10 -0.86 -0.49 1.00 -0.08
y 0.03 0.04 0.05 0.07 -0.02 0.05 -0.14 -0.07 -0.15 -0.03 -0.02 0.39 -0.07 0.10 0.09 -0.08 1.00
age job marital education default balance housing loan contact day month duration campaign pdays previous poutcome y
1.00 0.75 0.50 0.25 0.00 −0.25 −0.50 −0.75


K-Means Cluster Sizes
Number of Samples
20000
17500
15000
12500
10000
7500
5000
2500
0
0
1
2
Cluster


Silhouette Distribution per Cluster
Silhouette Score
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
0
1
2
Cluster

Recursive Bisecting K means