ML Lab Week 13 Clustering Lab

NAME: NEHA HARISH

SRN: PES2UG23CS378

SECTION: F

1. ANALYSIS QUESTIONS:

1.

In the correlation heatmap, most of the features showed very weak relationships with each other almost all the cells were light blue or light red. That tells that the dataset has a lot of low-correlation, noisy features that don't contribute strong structure individually. Because of that, clustering in the original high-dimensional space becomes messy, since the algorithm is trying to separate points using many weakly informative dimensions. PCA helps by compressing these weak signals into a smaller number of components that capture the most important variation. In my PCA results, the first two principal components together captured around 28% of the total variance. That's not extremely high, but it's enough to give a much cleaner 2D representation of the data for visualization and clustering. So overall, dimensionality reduction was necessary to reduce noise, remove redundancy, and make the underlying cluster structure more visible.

2.

From the elbow curve, the inertia dropped sharply from k = 1 to k = 3, and after k = 3 the curve started to flatten out. That means that after 3 clusters, the reduction in inertia becomes marginal, which is the classic elbow point. For the silhouette scores, the value for k = 3 was reasonably higher compared to other k values, with an overall silhouette score of about 0.39, which is typical for noisy marketing datasets. The silhouette didn't improve significantly beyond 3 clusters in fact, using more clusters tends to break the natural structure and lowers cohesion. So, using both metrics together, k = 3 is the

optimal number of clusters for this dataset. The elbow curve shows diminishing returns after 3, and the silhouette score confirms that 3 clusters give the best balance between separation and compactness.

3.

In K-means, one cluster was much larger because most customers share similar financial and demographic patterns. This big cluster represents the dominant customer segment. The smaller clusters capture the more unusual or distinct customer profiles. In Bisecting K-means, the cluster sizes were more balanced because the algorithm keeps splitting the largest cluster into two, which naturally evens out the distribution. The size differences tell us that the dataset has one major customer group and a few smaller, more unique segments, which is useful for targeted marketing.

4.

K-means gave a slightly higher silhouette score (around 0.39) compared to the Recursive Bisecting K-means, which means K-means formed more compact and better-separated clusters for this dataset. Bisecting K-means still created reasonable clusters, but because it forces repeated binary splits, the boundaries are not always the most natural fit for the data. K-means directly optimizes cluster cohesion and separation, so it performed a bit better here.

5.

The PCA clusters show that customers naturally fall into a few distinct groups. One large cluster represents the bank's mainstream customers with typical balances and campaign responses. The smaller clusters represent more specialized groups—such as customers with higher balances, more loan dependencies, or different engagement levels in past campaigns. For marketing, this means the bank can target each segment differently. The large cluster may need broad, generic campaigns, while the smaller clusters are ideal for personalized offers, such as loan products, investment plans, or targeted follow-ups. Overall, the segmentation helps the bank focus its marketing strategy based on behaviour rather than treating all customers the same.

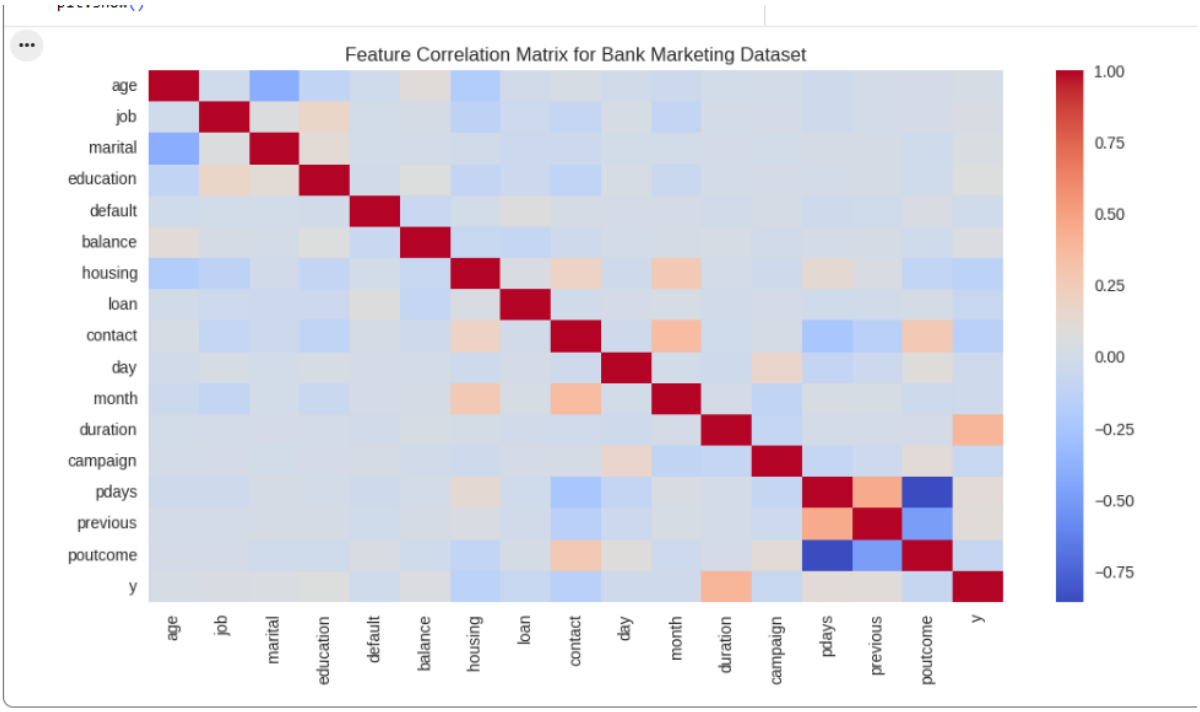
6.

In the PCA scatter plot, the turquoise, yellow, and purple regions represent groups of customers who have similar patterns in features like balance, loan status, campaign response, and age. Some boundaries look sharp because those customers differ more clearly on certain features, so PCA separates them well. Other boundaries look diffuse

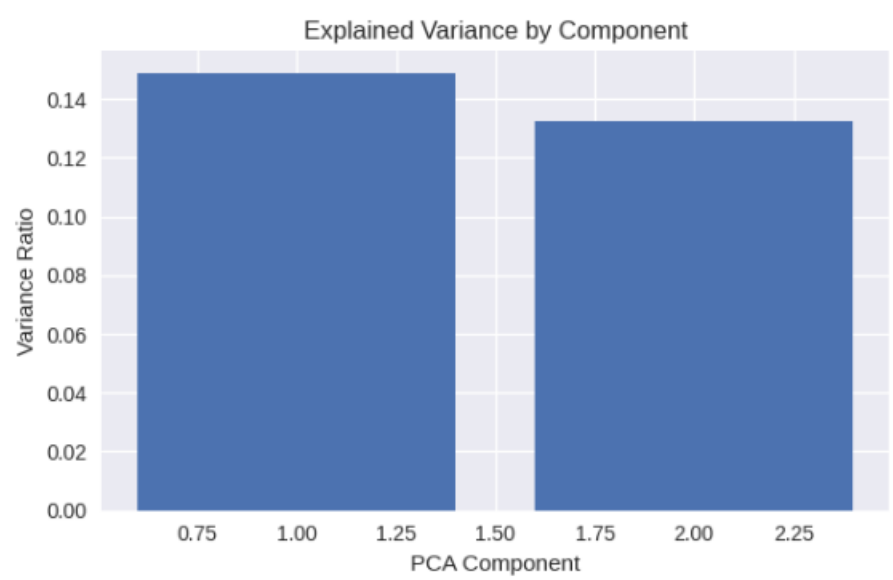
because many customers share overlapping behaviours, making their characteristics blend together in the PCA space. So the sharp regions show well-defined customer types, while the fuzzy regions reflect customers with mixed or similar behaviours.

2. SCREENSHOTS

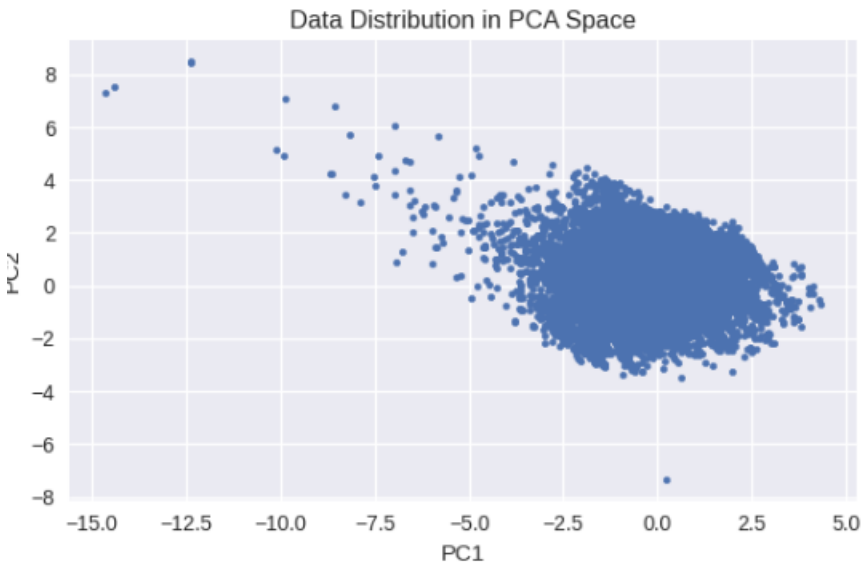
1. FEATURE CORRELATION MATRIX



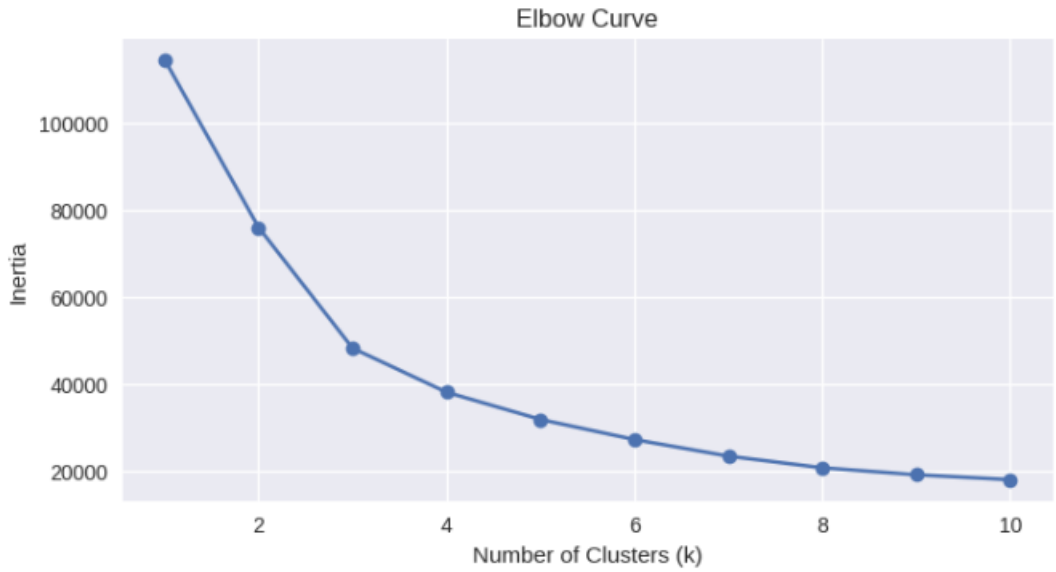
2. Explained variance by Component and Data Distribution in PCA Space after Dimensionality Reduction with PCA

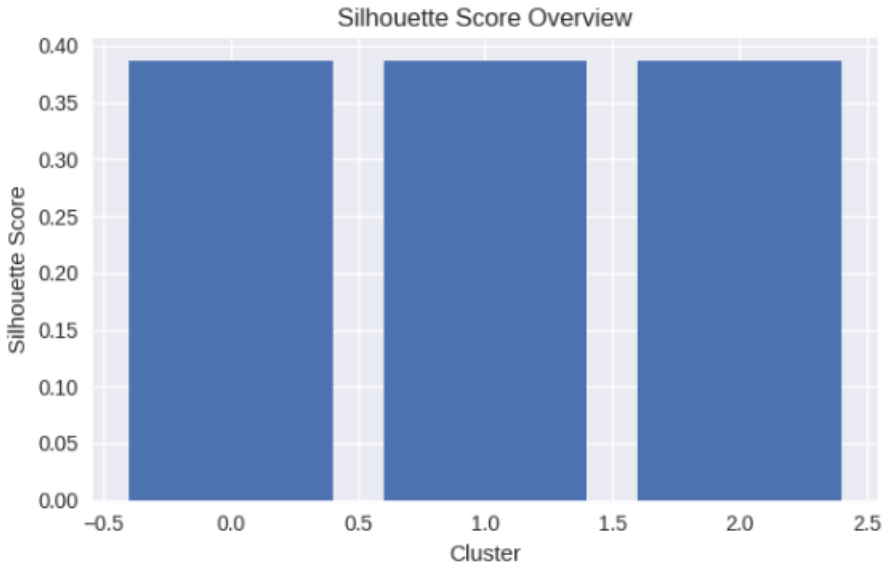



```
Shape after PCA: (45211, 2)
```

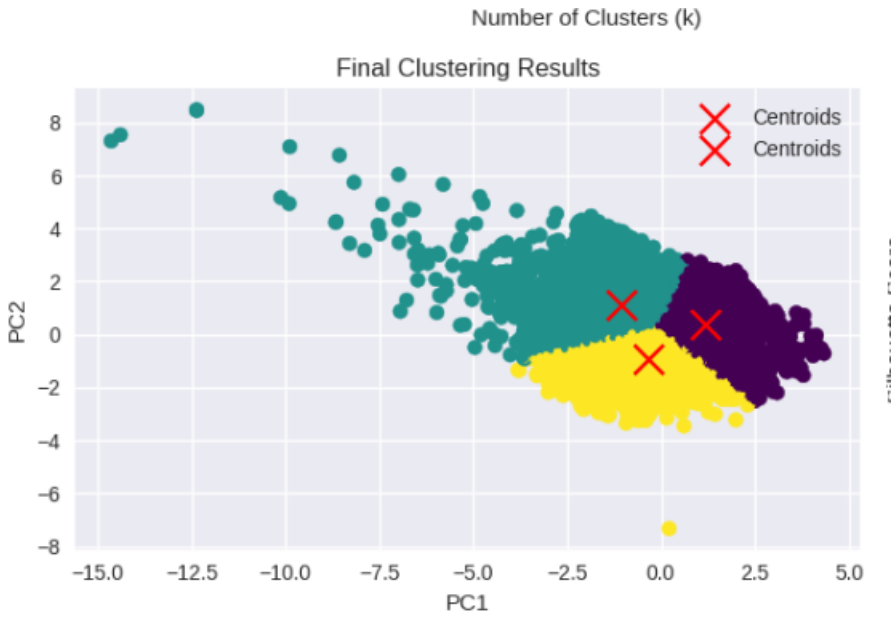


3. ‘Inertia Plot’ and ‘Silhoutte Score Plot’ for K-means

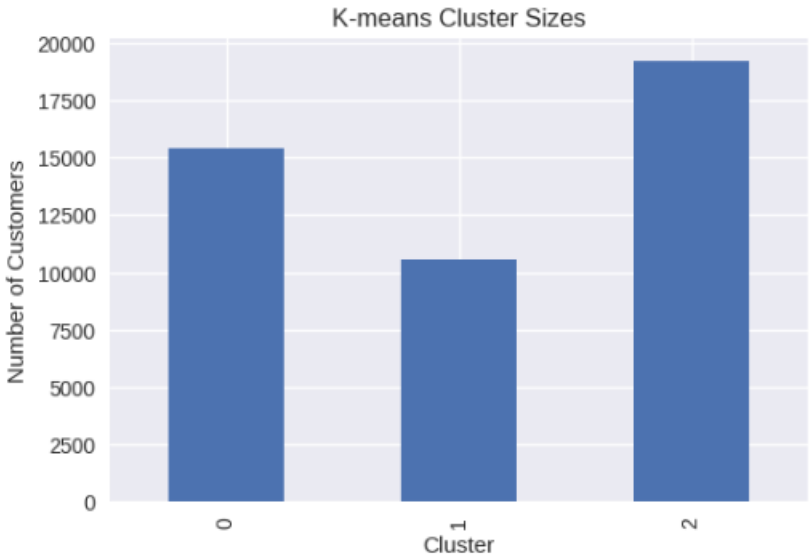




4.K-means Clustering Results with Centroids Visible (Scatter Plot)



K-means Cluster Sizes (Bar Plot)



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