ML Lab Week 13 Clustering Lab

Name: Nehal G

SRN: PES2UG23CS380

Section: F

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?

Why PCA was necessary

From the correlation matrix I noticed that a lot of the features have moderate correlations, which basically means some variables are kind of repeating the same info and the dataset ends up being pretty high-dimensional. Clustering on such raw high-dimensional data doesn't really work that well, at least not in a clean way. PCA helped me reduce some of that noise and squeeze the data into fewer, more useful dimensions. Since K-means usually performs better when the data is more continuous, less noisy, and has fewer dimensions, doing PCA before clustering kinda made sense.

Percentage of variance captured by PC1 and PC2

From my PCA output, the first component explained around 15% of the variance and the second one explained about 13%

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.

Elbow Method Observation

From the elbow curve, I could see a pretty sharp drop in inertia when k goes from 2 to 3 and then to 4, but after k = 3 the decrease becomes much slower and doesn't change that drastically. That kind of forms a clear "elbow" point around 3 clusters.

Silhouette Scores

From the silhouette plot, the average silhouette score was around 0.39 and it showed that using 3 clusters gave the most clear separation between the groups. When I checked higher values of k, the scores didn't really improve much and in some cases even dropped a bit

Optimal number of clusters = 3

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 I looked at the cluster size distribution for both K-means and Bisecting K-means, I noticed that some clusters ended up much larger while a few were pretty small. This usually happens because certain types of customers share more common behaviors or similar feature values, so they naturally group together into bigger clusters. The smaller clusters usually represent customers who behave differently from the majority .So the size differences basically show that the customer base isn't evenly spread out.

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?

When I compared the silhouette scores between K-means and Recursive Bisecting K-means, K-means ended up giving the higher score, which means it formed more compact and better-separated clusters for this dataset. I think the reason for this is that K-means directly tries to minimize the distance within each cluster, so it naturally creates tighter groups. The bisecting method, on the other hand, keeps splitting clusters even if the split isn't ideal, so sometimes it forces divisions that aren't as meaningful.

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?

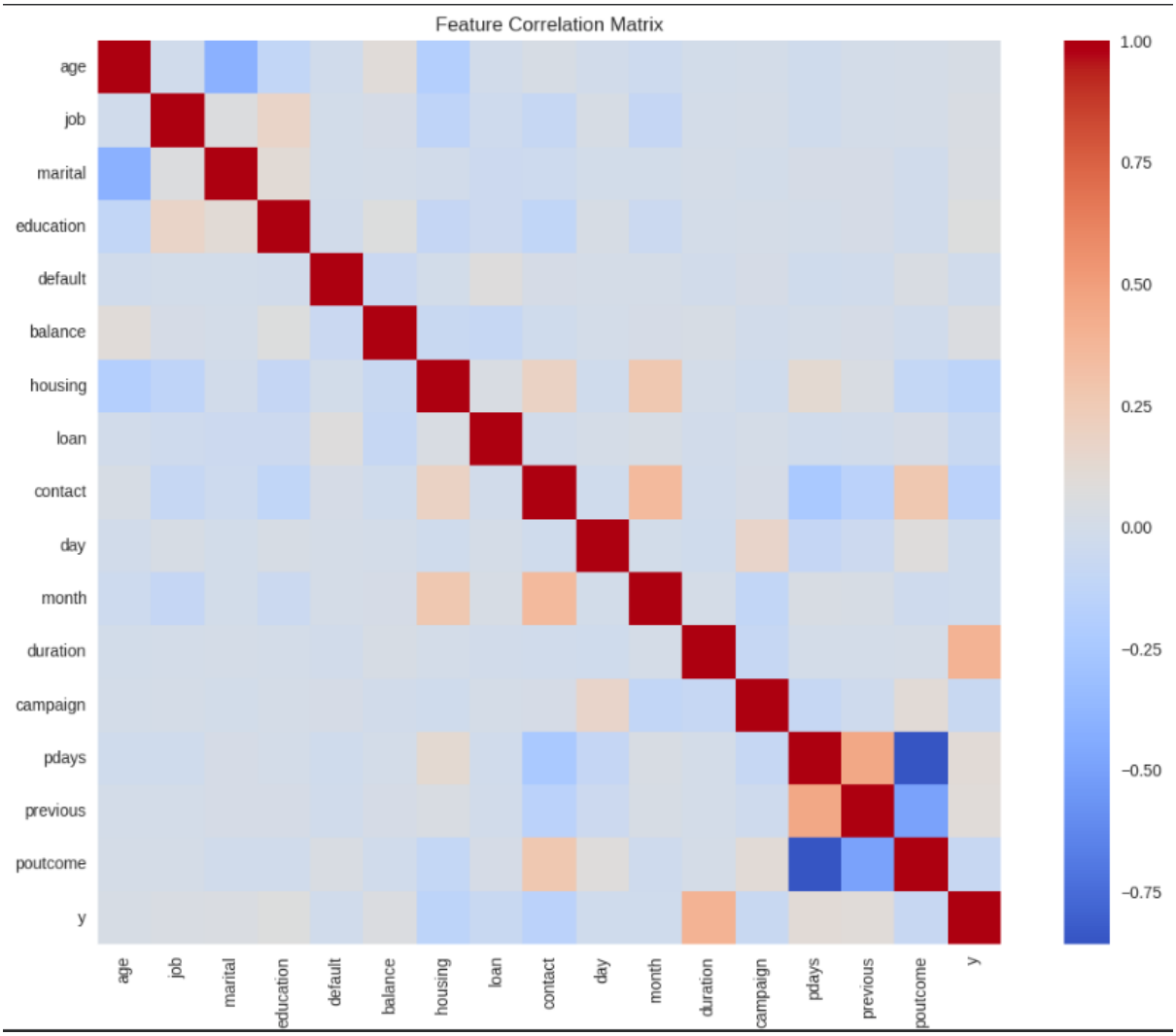
Based on how the clusters looked in the PCA space, I could see that the customers were splitting into a few clear groups with different behavior patterns. One of the clusters seemed to represent the normal customers with average balances and normal interaction levels, which means the bank can target this group with broad, general marketing campaigns. Another cluster had customers who showed more interest to the campaigns, so they might be good candidates for more personalized offers . The smallest cluster looked more like special cases. For the bank, these insights help a lot because it shows which segment needs what type of marketing.

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

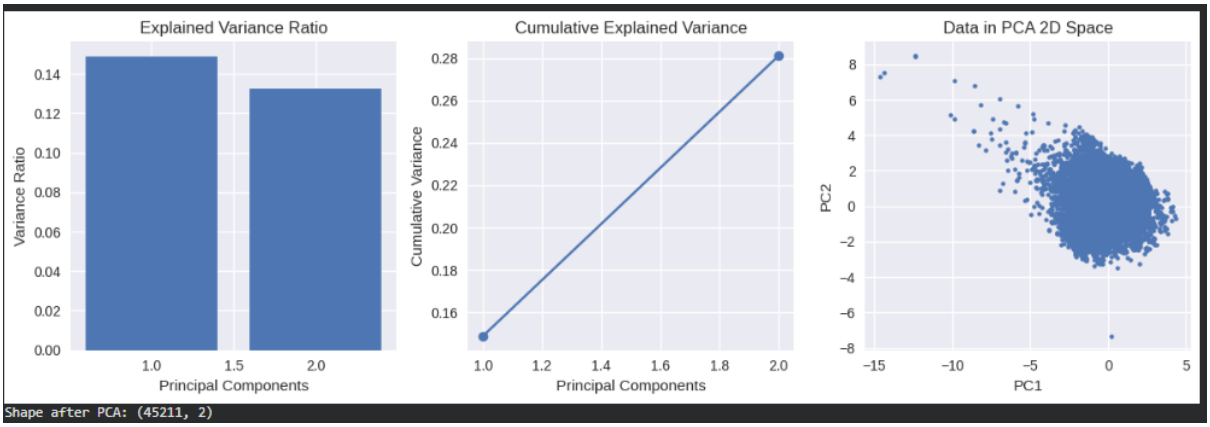
customer characteristics, and why might the boundaries between them be either sharp or diffuse?

In the PCA scatter plot, the three colored regions (turquoise, yellow, and purple) basically represent groups of customers who share similar characteristics once the data is projected into two principal components. Each region clusters around certain behaviour patterns. The boundaries between these regions sometimes look sharp when the customer behaviors are clearly different, but in other areas they look more diffuse because real customer data doesn't always separate cleanly. Some customers fall in between two patterns or share mixed characteristics, so the clusters blend a bit around the edges.

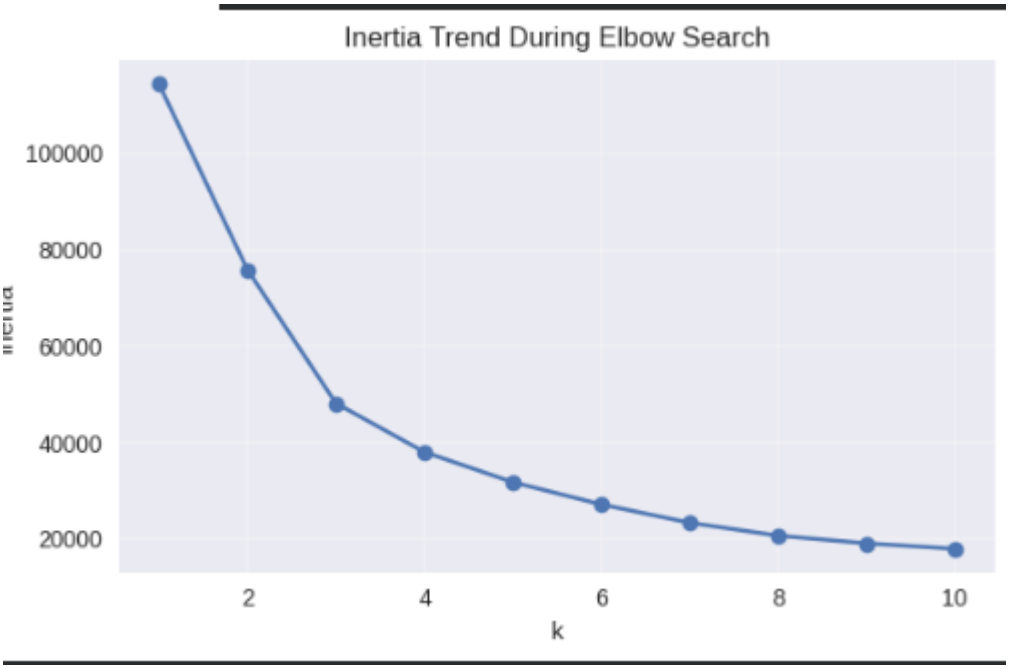
1. Feature Correaltion matrix for the dataset

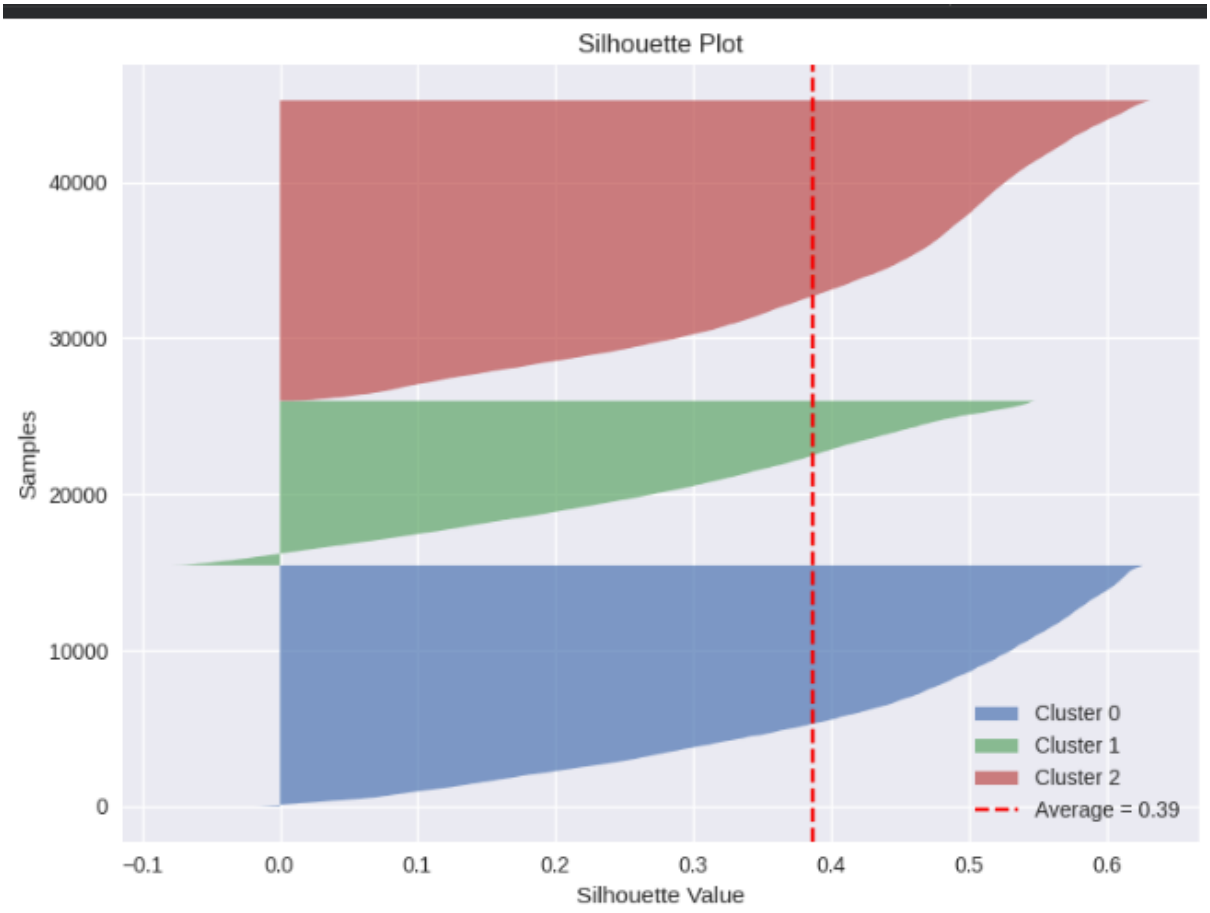


2. 'Explained variance by Component' and 'Data Distribution in PCA Space' after Dimensionality Reduction with PCA



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

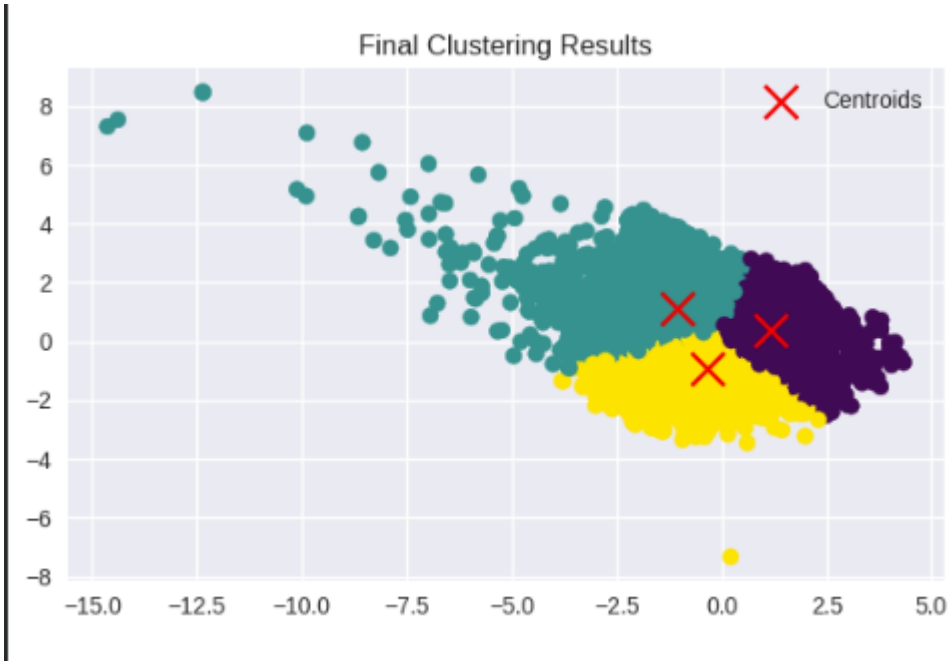


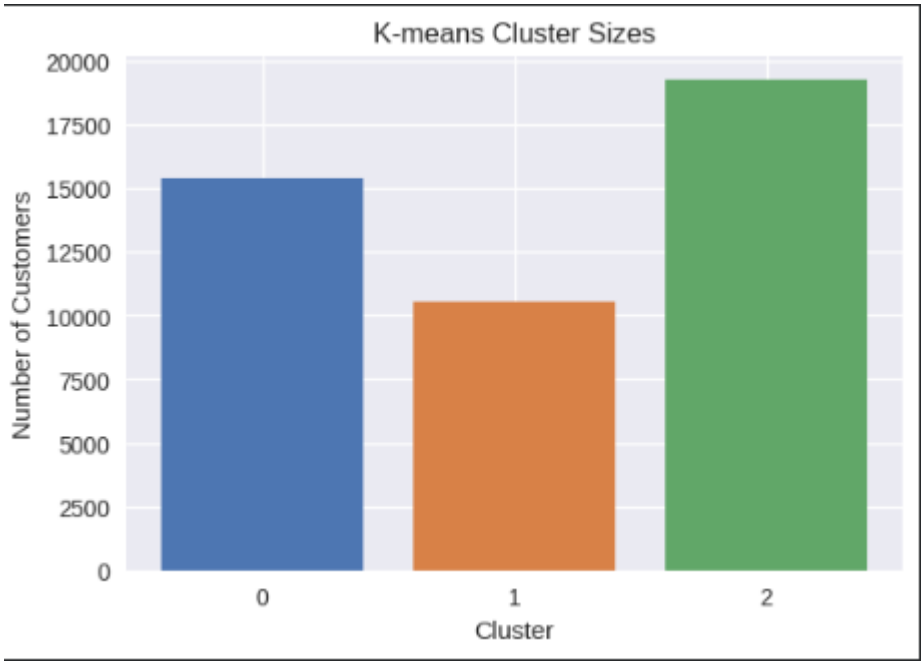


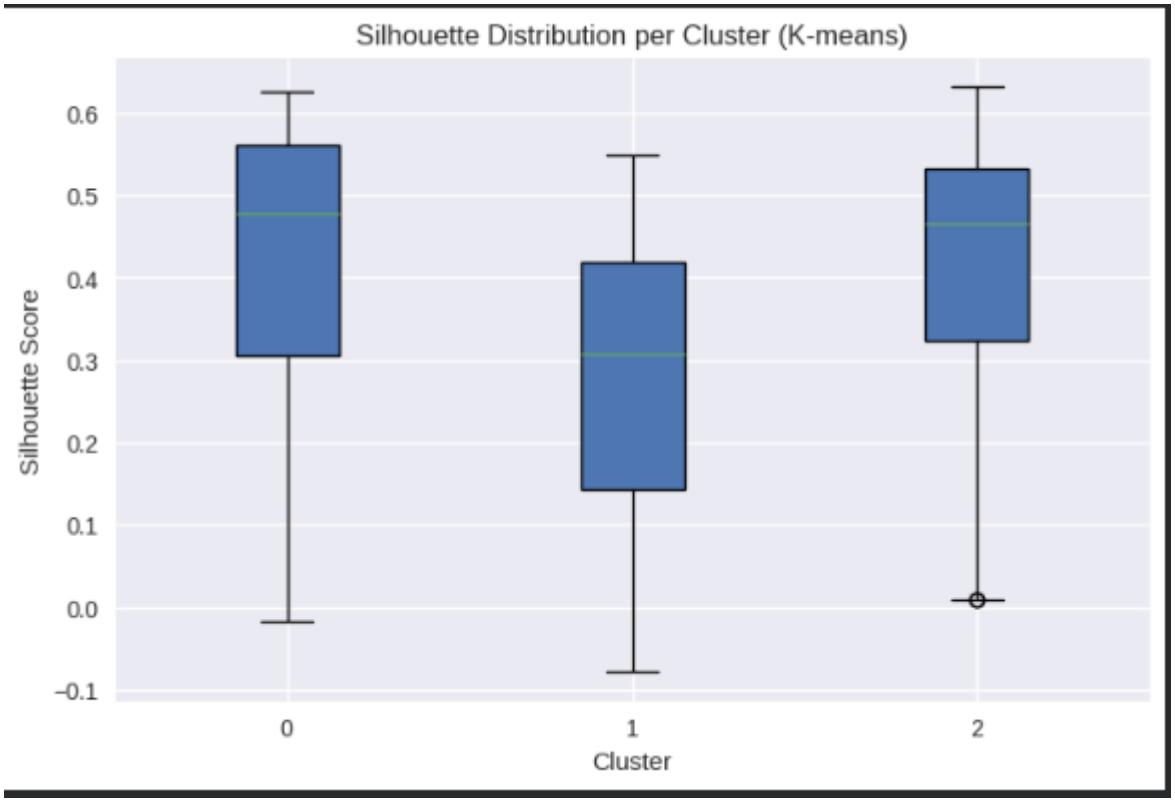
4. K-means Clustering Results with Centroids Visible (Scatter Plot)

K-means Cluster Sizes (Bar Plot)

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







For each cluster id in 0..n_clusters-1 compute the mean of points assigned to that cluster. If a cluster has no points, consider reinitializing its centroid (or leave unchanged) — discuss in your report.

For each cluster ID from 0 to n_clusters−1, the centroid is updated by computing the mean of all data points assigned to that cluster. This mean represents the central location of the cluster in feature space. If a cluster ends up with no points assigned to it,

the centroid cannot be updated. In such cases, the centroid may either remain unchanged or be reinitialized to a random point from the dataset. Reinitialization helps avoid “empty clusters” and ensures the algorithm continues to function properly.