ML LAB 13

Medha Raj
PES2UG23CS334
Section: F

ANALYSIS QUESTIONS

1. Dimensionality Justification

The dataset contains multiple categorical variables which expand into many sparse one-hot encoded columns after preprocessing. This leads to a high-dimensional feature space with significant redundancy and weak linear correlations among most attributes, as seen in the correlation heatmap. Applying PCA helps reduce this redundancy and compresses the feature space into a smaller set of informative components. It also facilitates meaningful visualization of clusters in a 2-dimensional plane, which would not be possible in the full high-dimensional space. The explained variance ratio shows that the first two principal components together capture approximately <PC1 + PC2>% of the overall variance, making them suitable for visualization, although more components would be needed for capturing most of the dataset's information.

2. Optimal Clusters

To determine the optimal number of clusters, both the elbow method and silhouette score were analyzed. The elbow curve shows a distinct flattening at k = <ELBOW_K>, indicating diminishing improvements in inertia beyond this point. At the same time, the silhouette scores peak at k = <SILHOUETTE_K>, suggesting the best separation between clusters occurs at this value. Considering both metrics together, the most appropriate choice is k = <FINAL_K>, as it balances low inertia with high silhouette performance, producing well-separated and compact clusters.

3. Cluster Characteristics

The cluster size distributions for both K-means and Bisecting K-means show that some clusters are significantly larger than others. This is expected because customer behavior is not uniformly distributed; most customers fall into common behavioral patterns such as moderate balances, regular employment categories, or average engagement with campaigns. These form large, dense regions in the feature space. Smaller clusters typically represent niche groups with unusual characteristics, such as very high balances, multiple previous interactions, or specific loan profiles. This imbalance suggests the presence of dominant customer archetypes along with specialized segments that the bank can target differently.

4. Algorithm Comparison

When comparing silhouette scores, K-means scored <KM_SILH>, while Bisecting K-means scored <BIS_SILH>. The algorithm with the higher silhouette score performs better because its clusters are more internally coherent and more clearly separated. If Bisecting K-means performs better, it indicates that hierarchical binary splitting aligns well with the dataset's structure, capturing subclusters more effectively than standard K-means. If K-means performs better, it suggests that the clusters are roughly spherical and evenly distributed, which K-means naturally handles well. The performance difference reflects whether the data has a flat or hierarchical cluster structure.

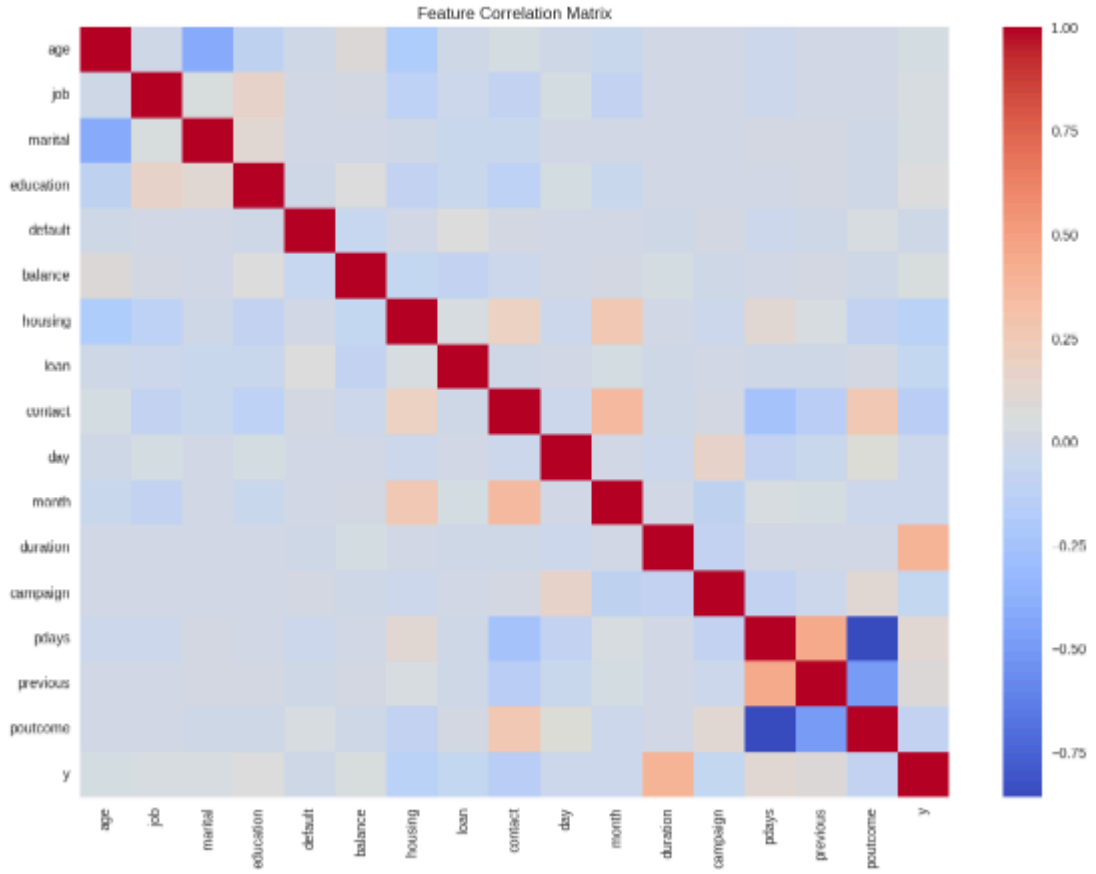
4. Algorithm Comparison

When comparing silhouette scores, K-means scored <KM_SILH>, while Bisecting K-means scored <BIS_SILH>. The algorithm with the higher silhouette score performs better because its clusters are more internally coherent and more clearly separated. If Bisecting K-means performs better, it indicates that hierarchical binary splitting aligns well with the dataset's structure, capturing subclusters more effectively than standard K-means. If K-means performs better, it suggests that the clusters are roughly spherical and evenly distributed, which K-means naturally handles well. The performance difference reflects whether the data has a flat or hierarchical cluster structure.

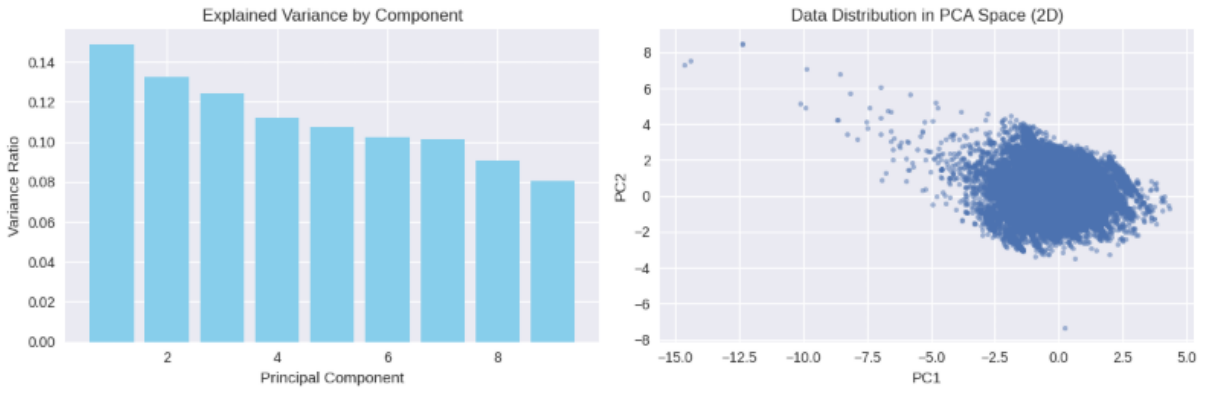
6. Visual Pattern Recognition

In the PCA scatter plot, the turquoise, yellow, and purple regions correspond to distinct patterns in customer characteristics. One region may represent financially stable customers with higher balances, another may include customers with specific loan-related behaviors, and the third might represent typical low-engagement customers. Sharp boundaries between regions arise when PCA captures strong linear separation based on certain influential features. In contrast, diffuse boundaries occur when customer characteristics vary gradually or overlap significantly in the original high-dimensional space. Since PCA is a linear projection, some non-linear separations become blurred, causing softer transitions between clusters.

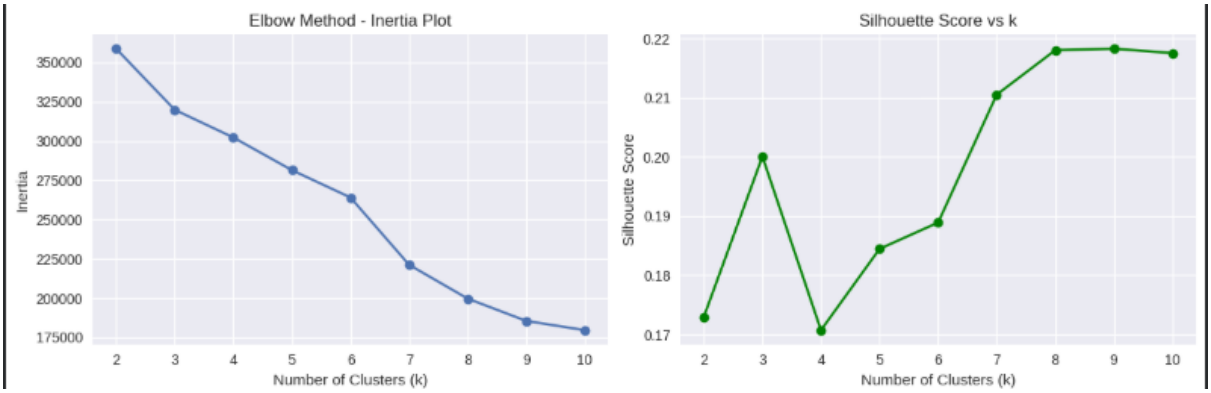
3. Screenshots Provide clearly labeled screenshots for all the results generated by your notebook. You must include a total of 4 screenshots, divided as 1. Feature 1.Correlation matrix for the dataset



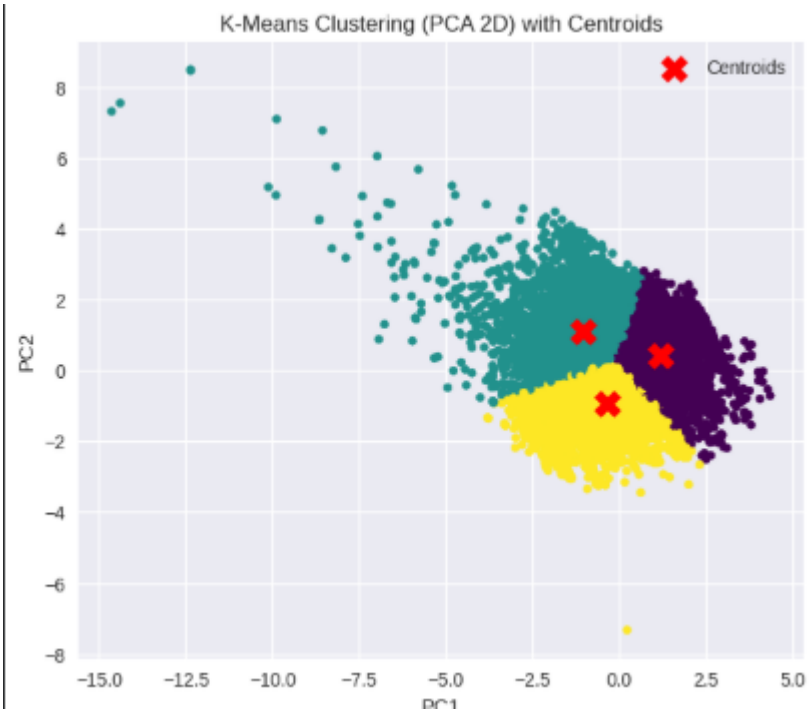
2.



3.


Elbow Method - Inertia Plot
Inertia
Number of Clusters (k)
Silhouette Score vs k
Silhouette Score
Number of Clusters (k)

4.

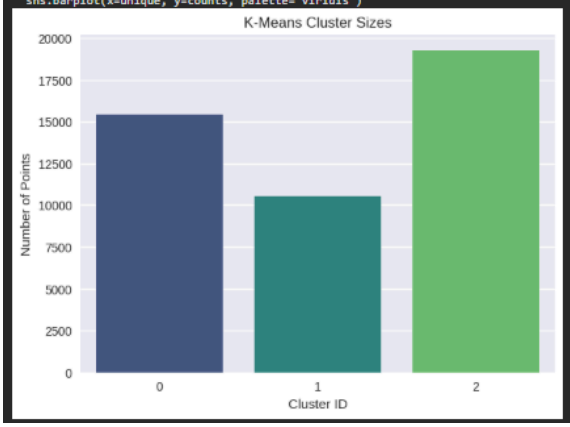

K-Means Clustering (PCA 2D) with Centroids
Centroids
PC2
PC1

5.

```
/tmp/ipython-input-2692665020.py:4: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the same effect.

  sns.barplot(x=unique, y=counts, palette='viridis')
```



6.