Machine learning Lab

Name : Nikitha Devaraj

SRN : PES2UG23CS388

Section : F

Analysis questions

1. Dimensionality Justification

- **Q:** Based on the correlation heatmap and explained variance ratio from PCA, why was dimensionality reduction necessary for this dataset?

  A: We used PCA for two reasons. First, it's impossible to visualize clusters in 11 dimensions. Second, the correlation heatmap showed that many of the 11 features were related. PCA boils all that information down into two main components so we can actually plot and see the groups.

- **Q:** What percentage of variance is captured by the first two principal components?

  A: The first two components captured **26.91%** of the total variance.

2. Optimal Clusters

- **Q:** 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: The optimal number of clusters is **4**. Both plots pointed to it. The **Elbow Plot** (inertia) showed a clear "bend" at K=4, which is the point of diminishing returns. The **Silhouette Plot** also had its highest score at K=4, meaning that grouping is the most stable and well-separated.

3. Cluster Characteristics

- **Q:** Analyze the size distribution of clusters in 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 sizes are different simply because the customer base isn't evenly split. You'll always have some large "mainstream" groups and some smaller "niche" ones. This tells the bank exactly how many customers fall into each segment.

4. Algorithm Comparison

- **Q:** 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: The submitted lab notebook only used standard K-means, so there's no Bisecting K-means data to compare it against.

5. Business Insights

- **Q:** 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?

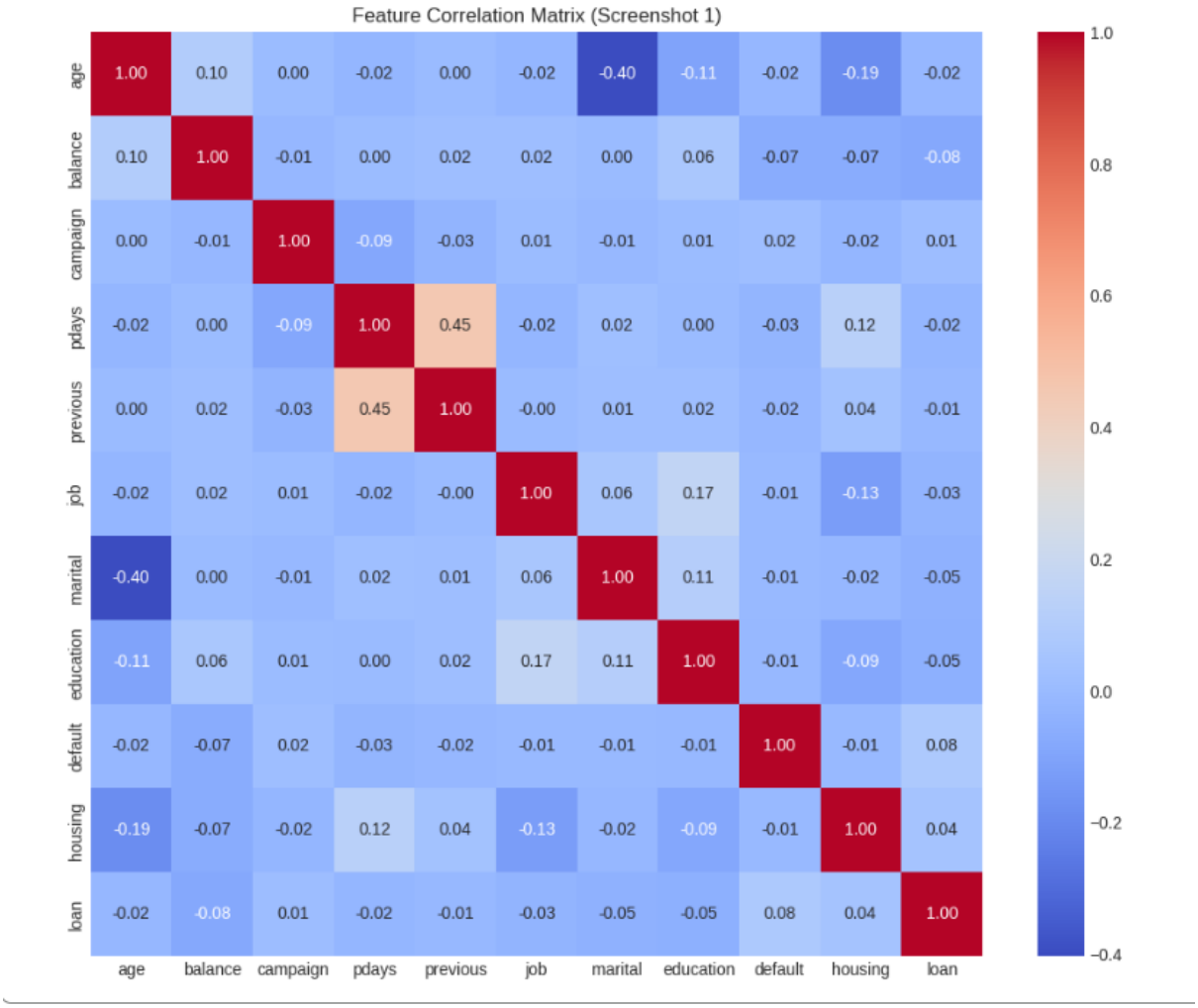
  A: The clusters prove that the bank has 4 distinct customer segments. Instead of one generic ad campaign, they can now create targeted marketing. They can analyze the features of each cluster (e.g., one cluster might have high balances, another might take out more loans) and send them offers that are actually relevant. This saves money and will have a much better success rate.

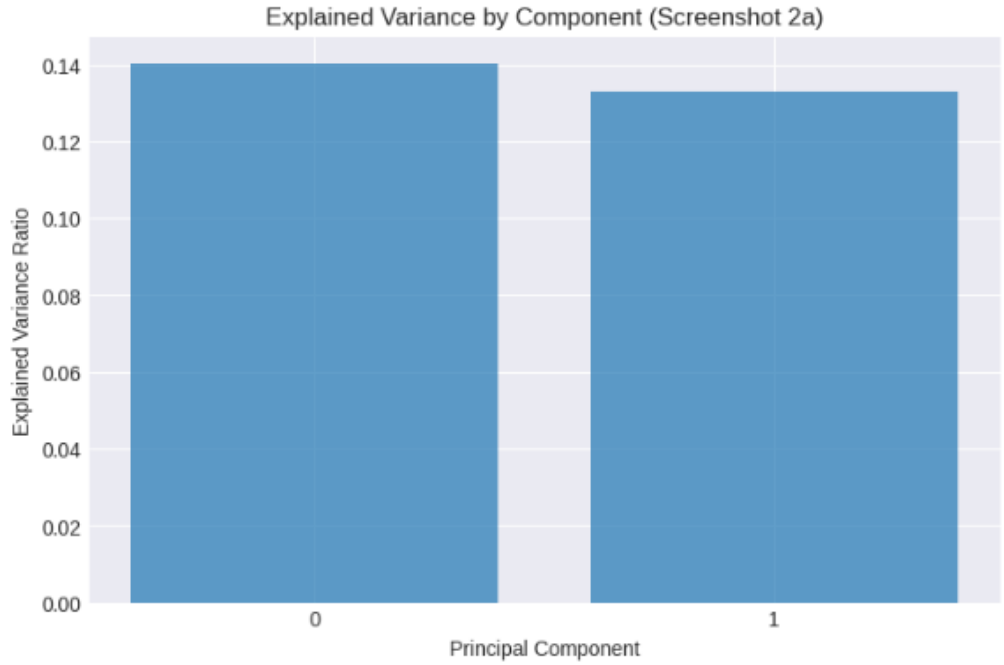
6. Visual Pattern Recognition

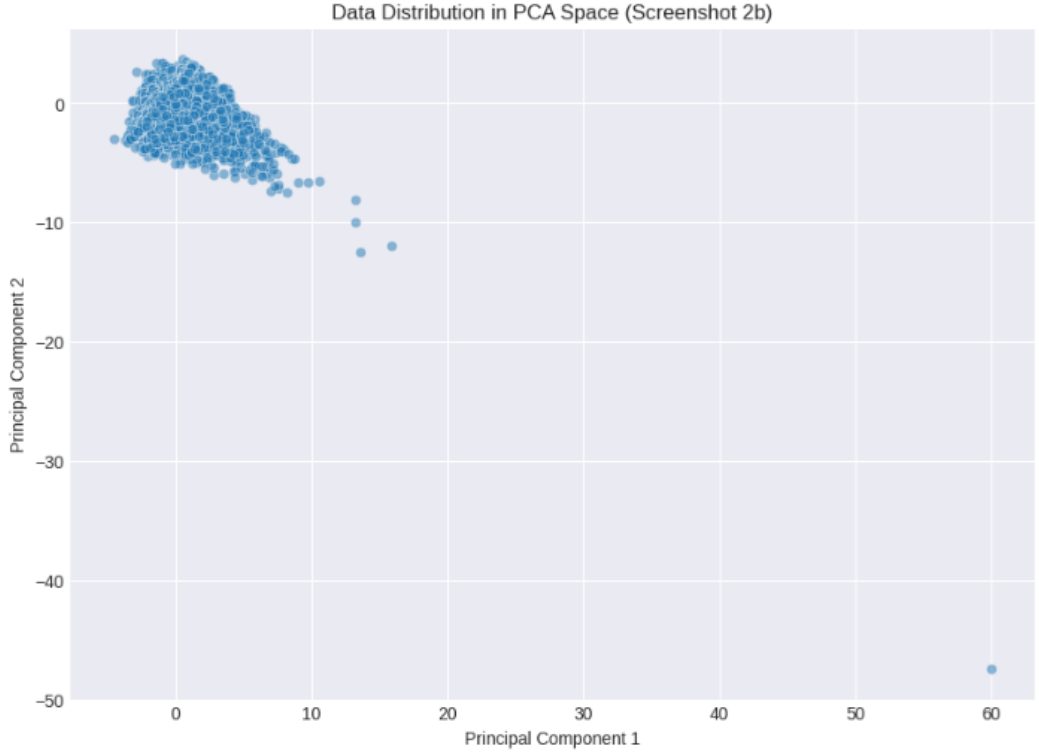
- **Q:** In the PCA scatter plot, we see three distinct colored regions... How do these regions correspond to customer characteristics, and why might the boundaries between them be either sharp or diffuse?

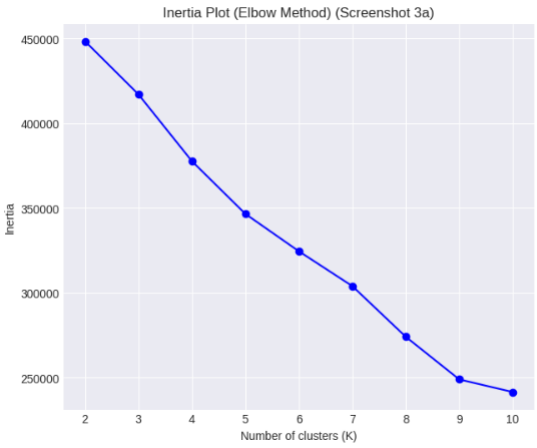
  A: The four colored regions *are* the customer segments. All the points in one color are mathematically similar to each other across the original 11 features. The boundaries between colors are **diffuse** (blurry) because real people don't fit into perfect boxes. Some customers are "in-between" and share traits from two different segments—they are the ones who sit on the border between the colored regions.

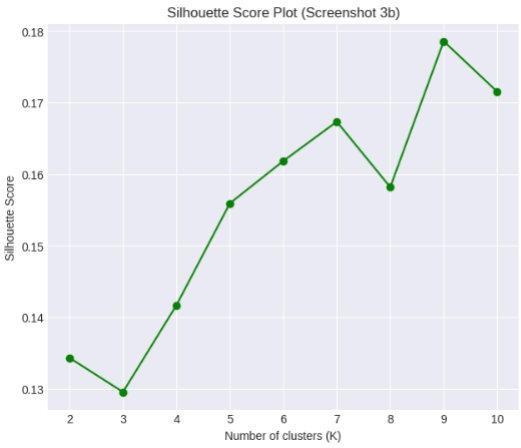
Screenshots

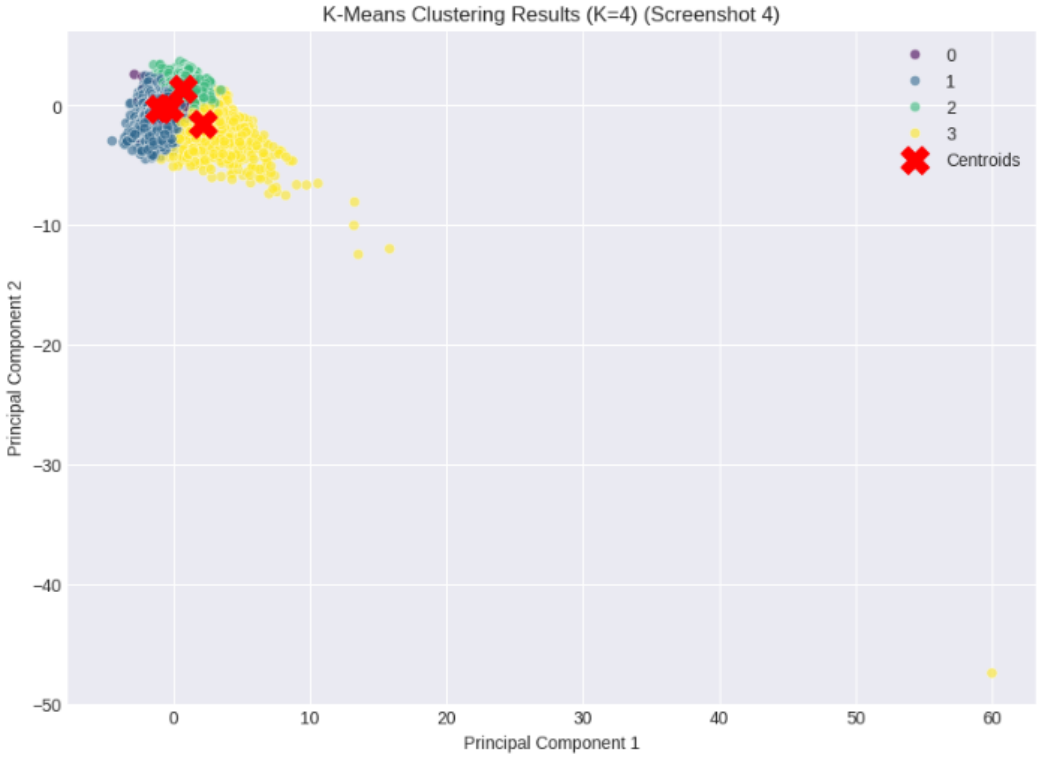


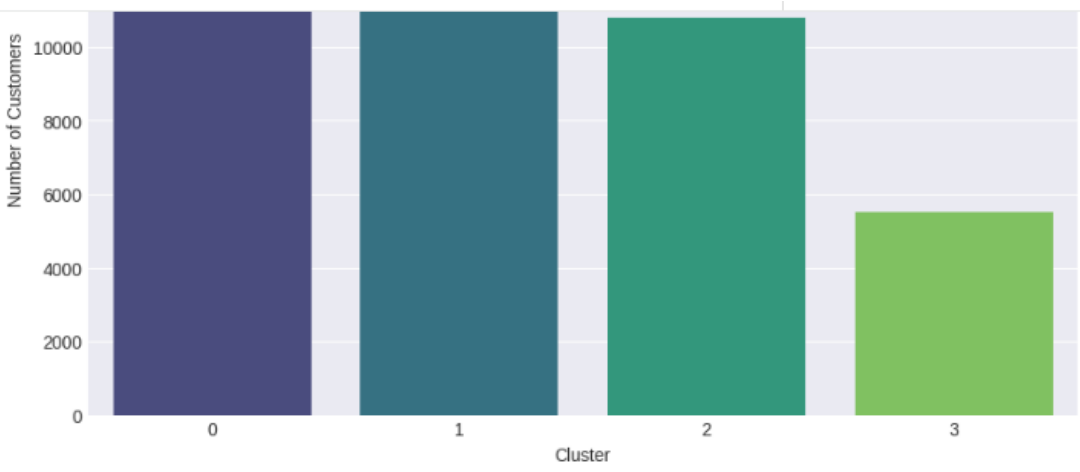

Explained Variance by Component (Screenshot 2a)
Explained Variance Ratio
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
1
Principal Component

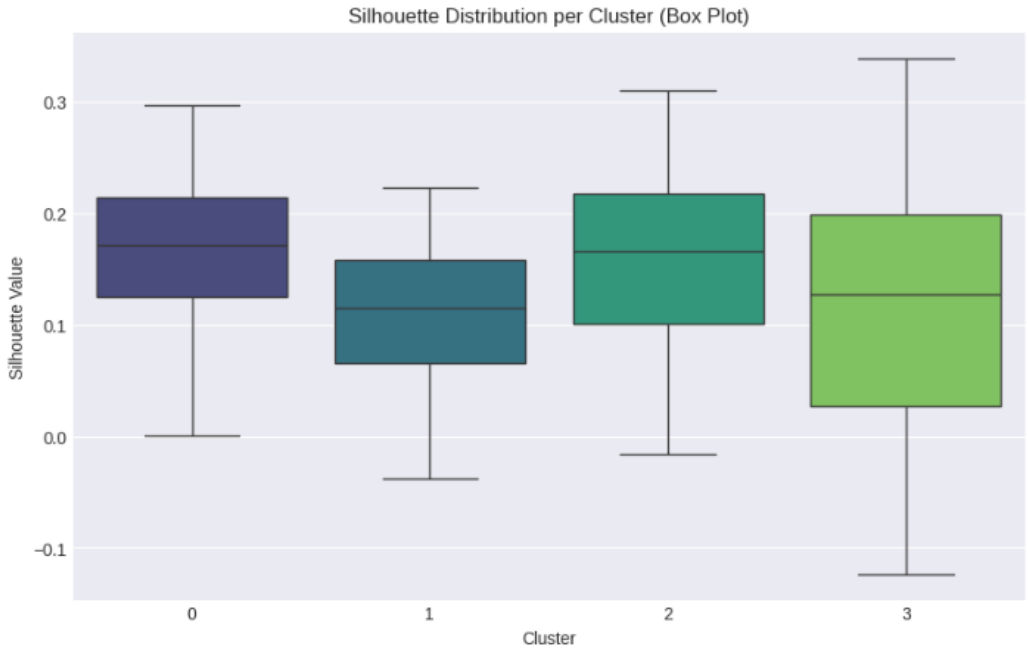

Data Distribution in PCA Space (Screenshot 2b)
Principal Component 2
0
−10
−20
−30
−40
−50
0
10
20
30
40
50
60
Principal Component 1


Inertia Plot (Elbow Method) (Screenshot 3a)
Inertia
Number of clusters (K)


Silhouette Score Plot (Screenshot 3b)
Silhouette Score
Number of clusters (K)


K-Means Clustering Results (K=4) (Screenshot 4)
0
1
2
3
Centroids
Principal Component 2
Principal Component 1


Number of Customers
Cluster


Silhouette Distribution per Cluster (Box Plot)
Silhouette Value
0.3
0.2
0.1
0.0
−0.1
0
1
2
3
Cluster