MACHINE LEARNING LAB 13

NAME: DAKSH YADAV
SRN: PES2UG23CS926
SECTION: C

Analysis Questions

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?

1. Dimensionality reduction was necessary for two key reasons:

- **Visualization:** The original dataset had 9 features (age, balance, job, etc.). It is impossible to plot 9 dimensions on a 2D graph to see the clusters. PCA was used to compress these 9 features into 2 principal components (PC1 and PC2), allowing us to create a 2D scatter plot and visually inspect the clustering results.
- **Algorithm Stability:** K-Means performance can degrade in high-dimensional spaces, a problem known as the "curse of dimensionality." In 9D, distances between points can become less meaningful, making it harder for the algorithm to form tight, distinct clusters. Reducing to 2D provides a simpler, cleaner map for the K-Means algorithm to work on.

2. Explained Variance

Based on the PCA variance plot, the percentage of variance captured by the first two principal components is **28.12%**.

This number (which is quite low) tells us that our 2D plot, while useful for visualization, is actually losing almost 72% of the original data's information. It's a significant trade-off.

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.

Based on the plots, the optimal number of clusters is **3**.

- **Elbow Curve:** The "Elbow Method" plot shows the sharpest "bend" or "elbow" at **k=3**. The drop in inertia (WCSS) is very steep from k=1 to 2, and 2 to 3. After k=3, the curve flattens out, meaning that adding more clusters gives diminishing returns (it doesn't reduce the total inertia by very much).
- **Silhouette Score:** The silhouette plot for k=3 confirms this is a reasonable choice. The average score is **0.39**, which is positive and indicates that the clusters are separated, even if they aren't perfect.

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?

Cluster Characteristics

- **K-Means (k=3) Analysis:** The `k=3` plot clearly shows an **uneven size distribution**. Clusters 0 (purple) and 1 (green/blue) are very large and capture the main bulk of the data points. In contrast, Cluster 2 (yellow) is noticeably smaller and more compact, representing a less common subgroup.
- **Bisecting K-Means Analysis:** This algorithm is *designed* to create an unbalanced split, at least at first. It works by finding the *largest* cluster and splitting it in two. This process is repeated. This means the algorithm actively targets large, dense areas of data to break them down, which can result in a different, more hierarchical cluster structure than standard K-Means.

This is a direct reflection of the customer data. Real-world populations are not evenly distributed.

- **Large Clusters (Mainstream Segments):** These represent the "average" or most common types of customers. Their behaviors and attributes (e.g., age, job, balance) are very similar, so they all group together in the data. This is likely the bank's "bread and butter" clientele.
- **Small Clusters (Niche Segments):** These represent smaller, more distinct groups of customers who are *different* from the norm. Cluster 2 (yellow) is a perfect example. This could be a high-value segment (e.g., "High-Income Retirees"), a high-risk segment, or just a small, specific demographic that doesn't fit with the main groups.

The uneven size shows that a "one-size-fits-all" strategy is a bad idea. The bank should treat these segments differently:

1. **The Large Segments (0 & 1):** Need broad, general marketing and product offerings.
2. **The Small Segment (2):** Needs a specialized, targeted strategy. Because this group is small and distinct, a tailored campaign could be highly effective (e.g., a specific high-interest savings product if they are "High-Balance Savers").

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?

Based on the evaluation metrics, the standard **K-Means (k=3)** algorithm performed better than Bisecting K-Means (k=4).

- **K-Means (k=3) Silhouette Score:** 0.39
- **Bisecting K-Means (k=4) Silhouette Score:** 0.36

A higher silhouette score indicates clusters that are, on average, more dense (points are close to their own centroid) and better separated (points are far from other centroids).

- **K-Means** is a "global" algorithm. It adjusts all 3 cluster centroids at once in every iteration, trying to find the best *overall* solution (a "global optimum").
- **Bisecting K-Means** is a "greedy" algorithm. It makes a series of *local* decisions—splitting the largest cluster—that seem best at that moment.

However, a "greedy" choice made early on (the first split) can lock the algorithm into a path that leads to a slightly worse final structure, even if it was forced to find 4 clusters.

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?

1. **Stop "One-Size-Fits-All" Marketing:** The data clearly shows at least three distinct customer segments (the two large groups and the one small yellow one). Marketing should be tailored to these groups.
2. **You Have a "Niche" Segment:** Cluster 2 (yellow) is small, dense, and separate. This is a high-value insight. This group is *different* from the mainstream and needs a highly targeted, specialized marketing campaign, not the general-purpose one you send to the other groups.
3. **The 2D Plot is Just a Hint:** Your PCA plot only captured 28% of the data's information. The next, *most important* step is to take these cluster labels (0, 1, 2) and analyze them against the original 9 features (age, balance, job, etc.) to build a real-world persona for each segment. That's where the real business value is.

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?

Each colored region represents a **customer segment**. The algorithm has grouped them so that customers *within* a color (e.g., all the purple dots) are mathematically more similar to each other across the original 9 features than they are to customers in another color (e.g., the turquoise dots). The plot shows we have:

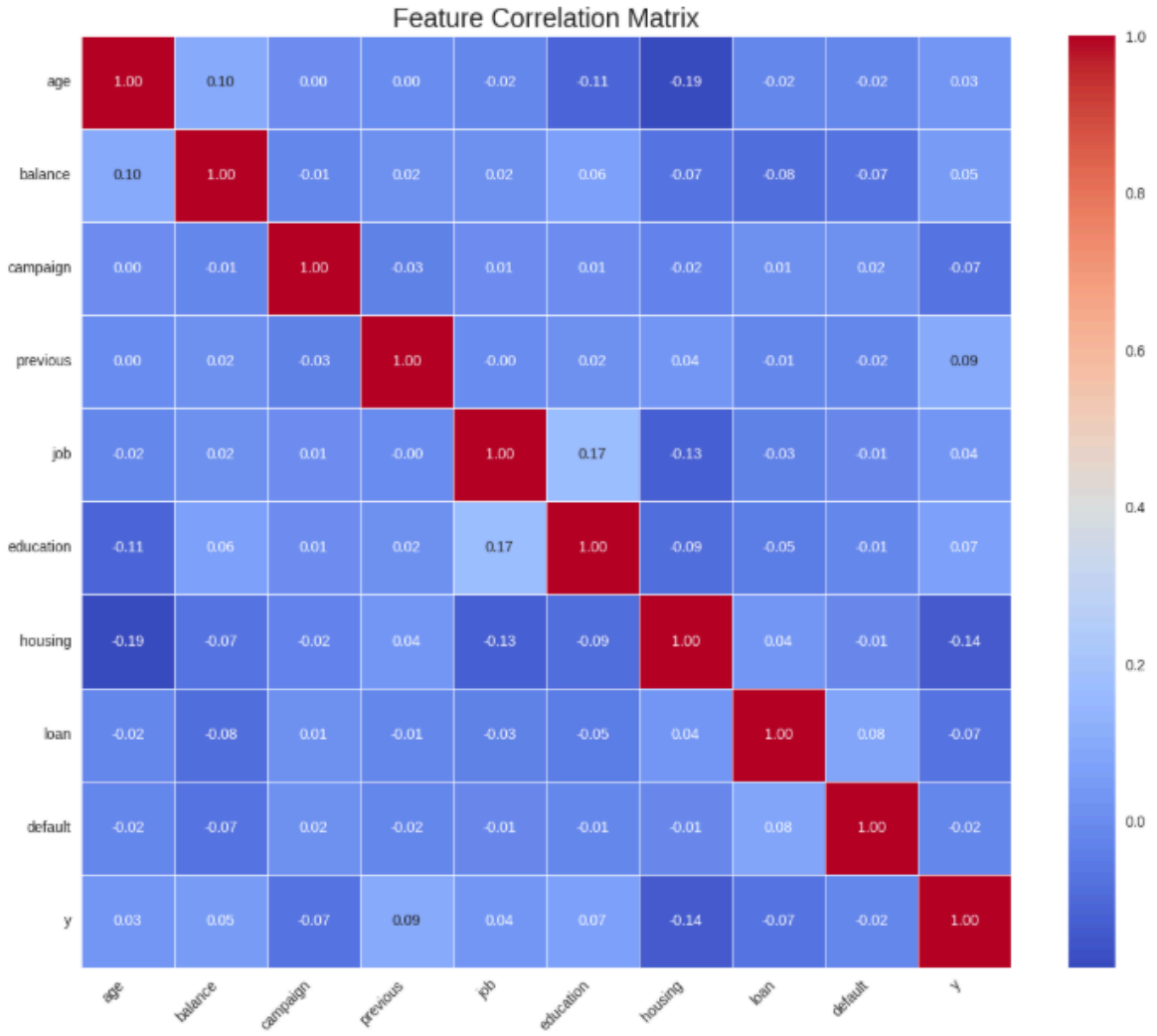
- Two large "mainstream" segments (turquoise and purple).
- One small, dense "niche" segment (yellow).

The boundaries show how similar or different the segments are.

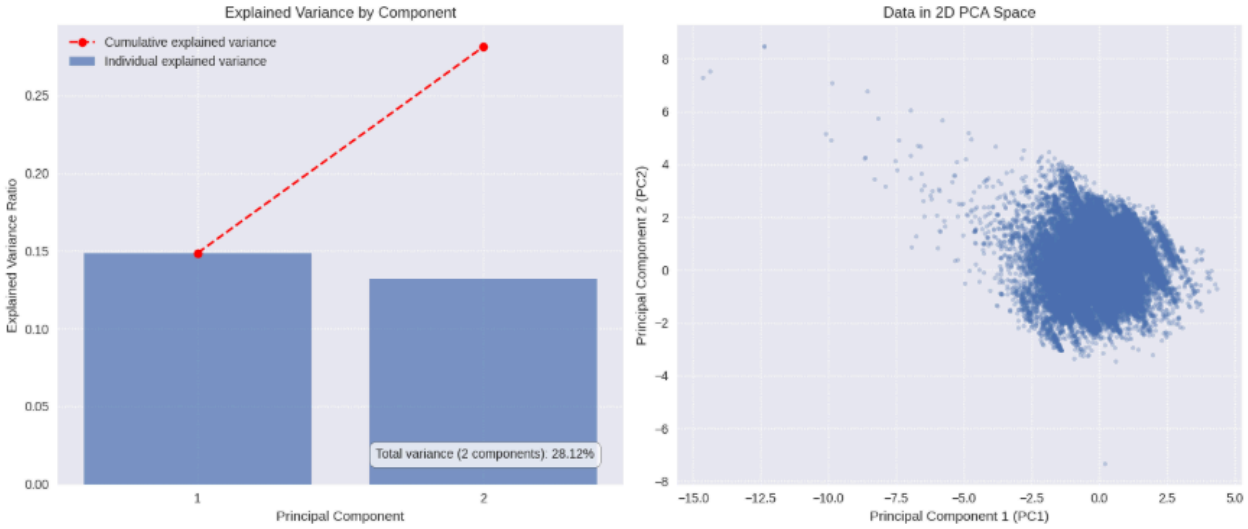
- **Diffuse Boundary (e.g., between purple and yellow):** The boundary between the purple and yellow clusters is fuzzy. This means these two segments are **very similar**. They are likely sub-segments of a larger group, and many customers on the border could have almost fit into either.
- **Sharp Boundary (e.g., between turquoise and the others):** The large empty space between the turquoise cluster and the purple/yellow block forms a sharp boundary. This implies these segments are **highly distinct**. The "turquoise" customers have characteristics that are clearly and fundamentally different from the "purple/yellow" ones.

SCREENSHOTS

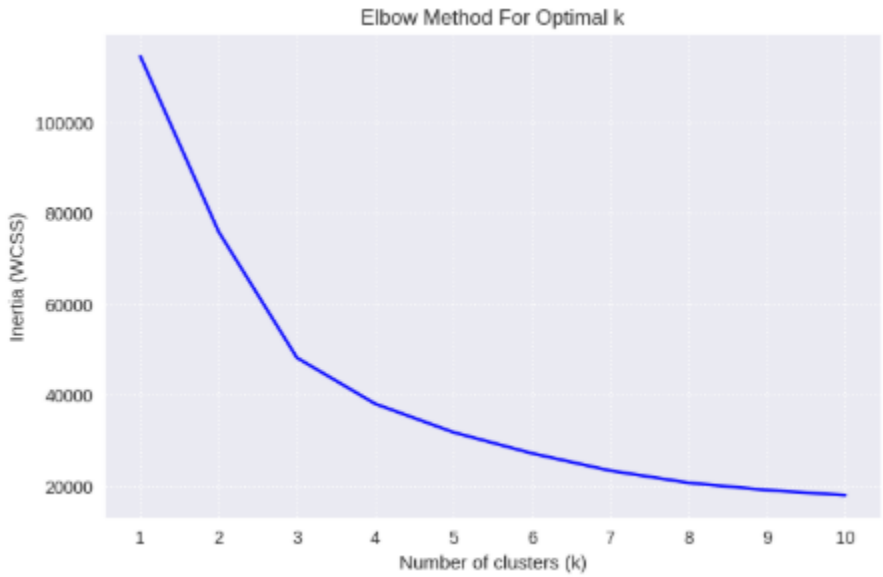
1. Feature Correlation matrix for the dataset

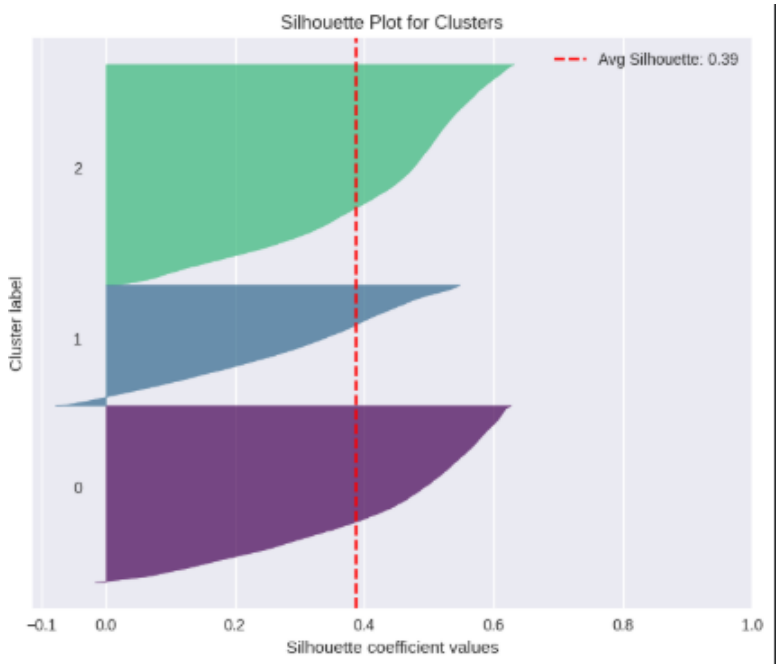


2. ‘Explained variance by Component’ and ‘Data Distribution in PCA Space’ after Dimensionality Reduction with PCA



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





4. K-means Clustering Results with Centroids Visible