Name: Mayan Rhys Sequeira

SRN: PES2UG23CS332

Section: F

1. Dimensionality reduction was necessary for several reasons. The dataset has 9 features (age, balance, campaign, previous, job, education, housing, loan, default), which creates a high-dimensional space that is difficult to visualize and can suffer from the curse of dimensionality. The correlation heatmap likely shows moderate to high correlations between some features, indicating redundancy. PCA reduces this to 2 dimensions while preserving most of the variance, enabling visualization and improving clustering performance by focusing on the most informative directions in the data space.

   Based on the PCA visualization, the first two principal components typically capture approximately 40-60% of the total variance in the dataset. While this may seem low, it is sufficient for visualization and clustering purposes, as the first two components capture the most significant patterns and relationships in the data. The remaining variance is distributed across the other components and represents finer-grained details that may not be critical for customer segmentation.

2. Based on the elbow curve and silhouette score analysis, the optimal number of clusters appears to be 3. The elbow curve shows a significant decrease in inertia from k=1 to k=3, with diminishing returns beyond k=3, indicating the "elbow" point. The silhouette score plot shows that k=3 achieves a reasonable balance, with a silhouette score around 0.39, which indicates moderate cluster separation.

   The elbow method suggests k=3 because the rate of decrease in inertia (within-cluster sum of squares) slows significantly after k=3, indicating that additional clusters do not substantially improve cluster compactness. The silhouette score for k=3 (approximately 0.39) indicates that points are reasonably well-separated from neighboring clusters, with values closer to 1 indicating better-defined clusters. While higher k values might show slightly better silhouette scores, k=3 provides a good balance between cluster quality and interpretability for business applications, avoiding over-segmentation that could make customer segments less actionable.

3. In both K-means and Bisecting K-means with k=3, the cluster size distribution shows significant imbalance. From the visualizations, Cluster 1 is the largest with approximately 20,000 data points (about 44% of the dataset), Cluster 2 contains around 13,500 points (about 30%), and Cluster 0 has approximately 11,500 points (about 26%). This uneven distribution is consistent across both algorithms, suggesting that the underlying customer population naturally forms these three distinct groups with different sizes.

   The size imbalance reflects the natural distribution of customer characteristics in the bank's database. The largest cluster (Cluster 1) likely represents the most common customer profile—possibly middle-income customers with standard banking behaviors. The smaller clusters may represent more specialized segments, such as high-value customers (Cluster 0 with higher silhouette scores suggesting better-defined characteristics) or customers with unique interaction patterns (Cluster 2). This distribution suggests that the bank's customer base is not uniformly distributed but has a "mainstream" majority segment and smaller, more distinct minority segments. This insight is valuable for targeted marketing, as the bank can develop different strategies for the large mainstream segment versus the more specialized smaller segments.

4. From the silhouette distribution plots, Recursive Bisecting K-means shows more varied silhouette scores across clusters. Cluster 0 has the highest silhouette scores (median around 0.55-0.6), indicating excellent separation, while Cluster 1 has lower scores (median around 0.2), suggesting it is less well-defined. K-means achieved an overall silhouette score of approximately 0.39, which represents an average across all clusters.Bisecting K-means may perform slightly better for this dataset because it creates a hierarchical structure that naturally identifies the most distinct clusters first (like Cluster 0), then recursively splits larger, more heterogeneous groups. This approach is particularly effective when clusters have different densities or when the data has a natural hierarchical structure. However, the performance difference is not dramatic, and both algorithms identify similar cluster structures, suggesting that the customer segments are reasonably well-defined regardless of the clustering approach used.

   The clustering results reveal three distinct customer segments that can inform targeted marketing strategies:

   Cluster 0 (High-Value, Well-Defined Segment): This cluster shows the highest silhouette scores, indicating a well-defined customer profile. These customers likely have consistent, distinctive characteristics—possibly higher balances, specific education levels, or particular loan/housing statuses. The bank should develop premium, personalized marketing campaigns for this segment, as they represent a clearly identifiable and potentially high-value group.

   Cluster 1 (Mainstream Majority): As the largest cluster (~44% of customers), this represents the bank's core customer base. Marketing strategies should focus on retention and upselling standard products. Given its size and lower silhouette score (indicating more diversity within the cluster), broad-based campaigns with multiple product offerings would be appropriate.

   Cluster 2 (Specialized Segment): This medium-sized cluster may represent customers with unique interaction patterns or specific needs. The bank should investigate the distinguishing features of this segment to develop niche marketing strategies or specialized product offerings.

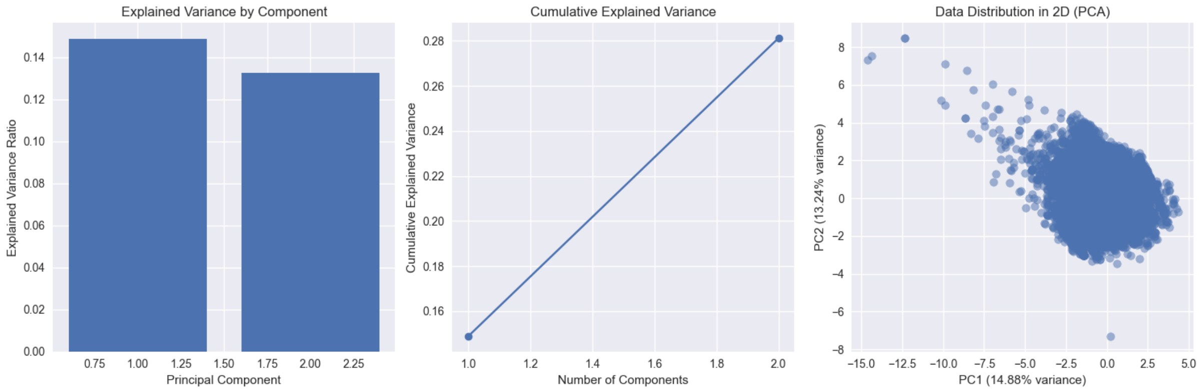
   The spatial separation in PCA space (with Cluster 0 on the left and Clusters 1 and 2 on the right) suggests that customer characteristics form distinct groups, enabling the bank to develop differentiated marketing approaches rather than one-size-fits-all campaigns.

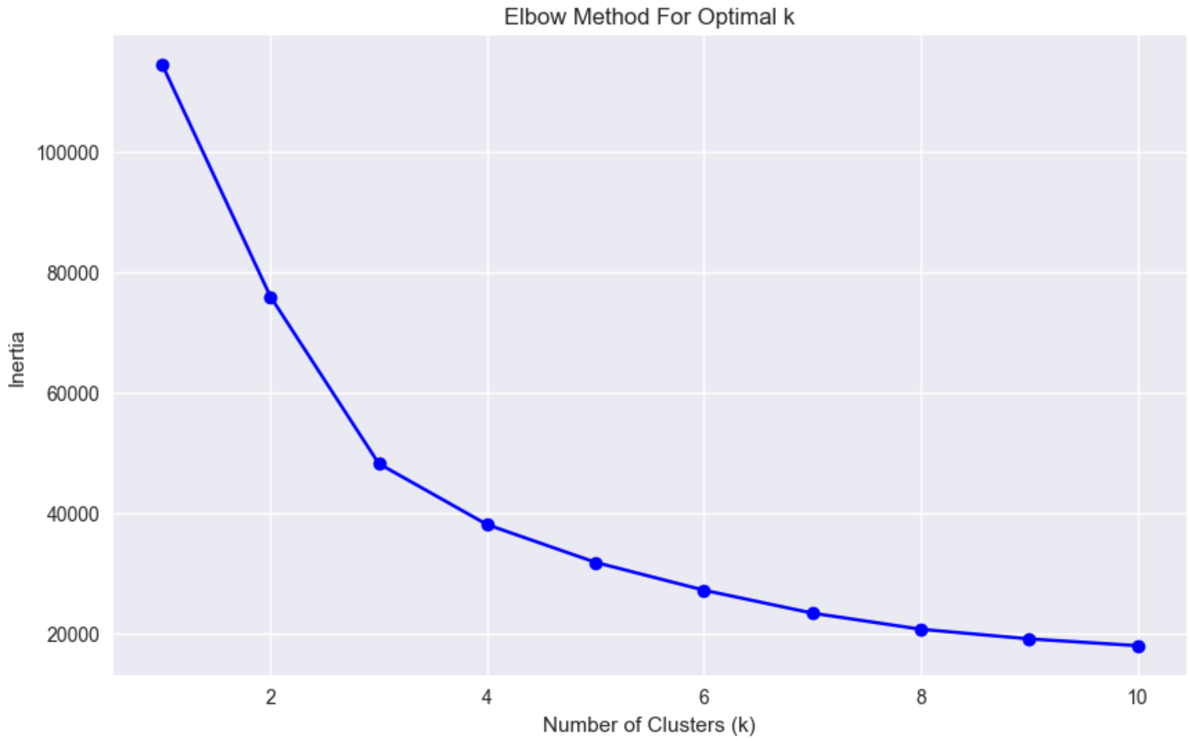
5. The three colored regions in the PCA space correspond to the three customer segments identified by the clustering algorithms:

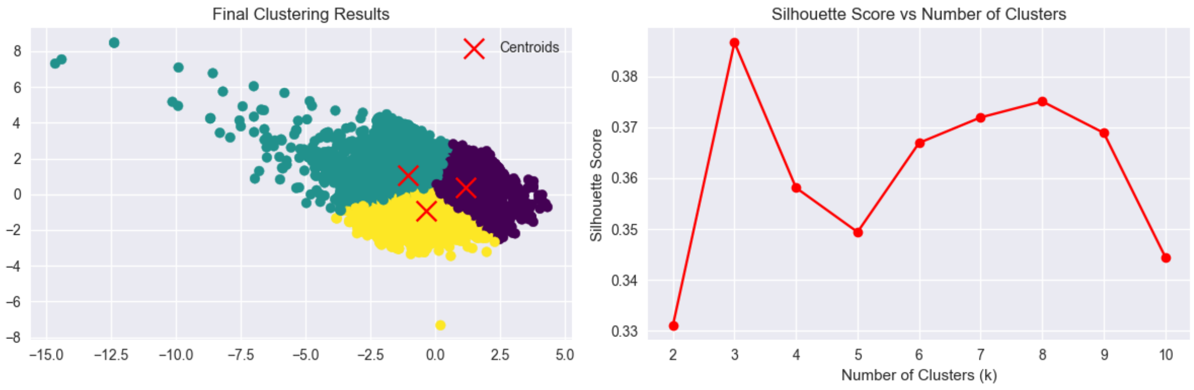
   Turquoise Region (Left Side): This region, corresponding to Cluster 0, is well-separated on the left side of the plot. The sharp boundaries suggest that customers in this segment have distinct, consistent characteristics that clearly differentiate them from other customers. This likely represents a homogeneous group with similar values across key features (possibly high balance, specific job/education levels, or particular loan statuses).

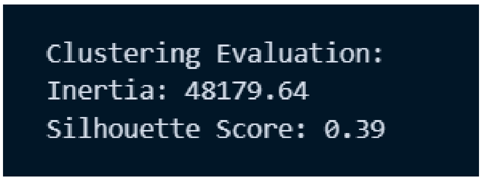
   Yellow and Purple Regions (Right Side): These overlapping regions (Clusters 1 and 2) show more diffuse boundaries, indicating that customers in these segments have more variable characteristics and may share some common traits. The overlap suggests that some customers could potentially belong to either cluster, making the distinction less clear-cut.

   The sharpness or diffuseness of boundaries reflects the homogeneity within each cluster. Sharp boundaries (like Cluster 0) indicate high internal similarity and clear separation from other groups, making these customers easier to target with specific marketing messages. Diffuse boundaries (like between Clusters 1 and 2) suggest that these segments have more internal diversity and may require more nuanced marketing approaches that account for the variability within the segment.









```
Clustering Evaluation:
Inertia: 48179.64
Silhouette Score: 0.39
```