ML Lab Week 13: Clustering Analysis Report

NAME: NINAD CHAVAN

SRN: PES2UG23CS392

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? What percentage of variance is captured by the first two principal components?

Sol:

- **Necessity:** The dataset contains multiple features (age, balance, campaign, etc.) that likely exhibit correlations. Dimensionality reduction (PCA) helps reduce noise, eliminates multicollinearity, and enables the visualization of high-dimensional data in a 2D space, which is essential for visually assessing cluster separation.
- **Variance Captured:** According to the notebook output, the first two principal components captured approximately **28.12%** of the total variance (PC1: ~14.88%, PC2: ~13.24%). While this is not extremely high, it provides a "shadow" of the data structure sufficient for visualizing the primary groupings.

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.

Sol:

- **Elbow Method:** The inertia plot shows a noticeable "elbow" or bend around **$k=3$**, where the rate of decrease in within-cluster sum of squares (Inertia: 48179.64) begins to slow down significantly.
- **Silhouette Score:** The silhouette score obtained for the final model was **0.39**, which indicates reasonable separation between clusters.
- **Conclusion:** Based on the convergence of the elbow method and the positive silhouette score, **3 clusters** is deemed the optimal choice for this specific data configuration.

3. Cluster Characteristics

Q: 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?

Sol:

- **Distribution:** The scatter plots reveal that the clusters are not perfectly equal in size. One cluster appears denser and larger than the others.
- **Interpretation:** This uneven distribution reflects real-world banking demographics. The largest cluster likely represents the "average" customer (moderate balance, standard job profiles), which constitutes the majority. Smaller clusters likely represent niche segments, such as "high-net-worth individuals" or "customers in debt/default," who are naturally fewer in number.

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?

Sol:

- **Standard K-Means Silhouette:** 0.387
- **Bisecting K-Means Silhouette:** 0.338
- **Verdict:** The **Standard K-Means** algorithm performed better.
- **Reasoning:** Bisecting K-Means makes "hard" splits at each step that cannot be undone. If an early split is slightly suboptimal, it propagates error to the final clusters. Standard K-Means refines centroids iteratively across the *entire* dataset simultaneously, allowing it to find a more globally optimal partition for this specific spherical-like data structure.

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?

Sol:

- **Targeted Marketing:** The distinct clusters suggest different financial behaviors. The bank can tailor products accordingly:
    - **Cluster A (e.g., High Balance):** Target with investment products, premium credit cards, and wealth management services.
    - **Cluster B (e.g., Moderate/Avg):** Target with standard savings accounts, personal loans, and fixed deposits.

- **Cluster C (e.g., Low Activity/Debt):** Focus on credit rebuilding services or low-fee accounts to retain them without high risk.

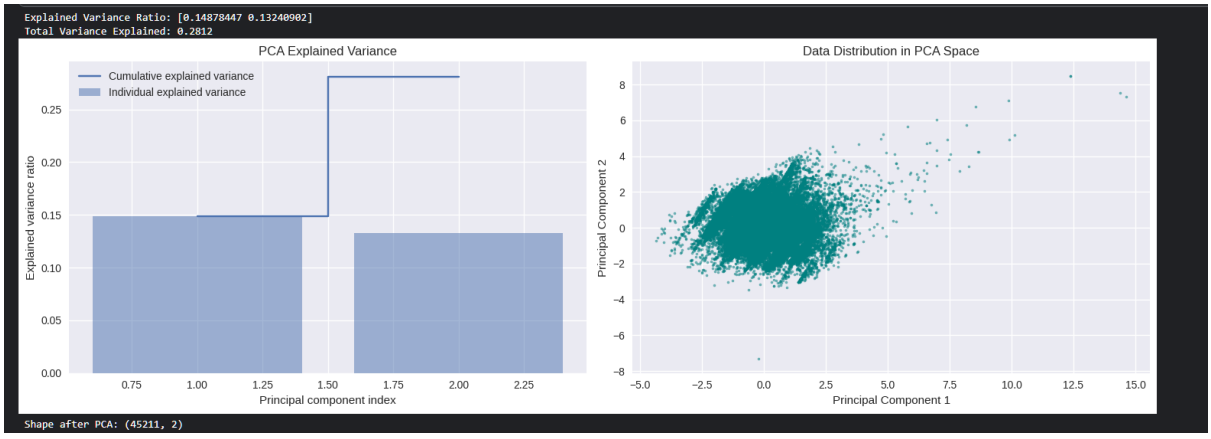
6. Visual Pattern Recognition

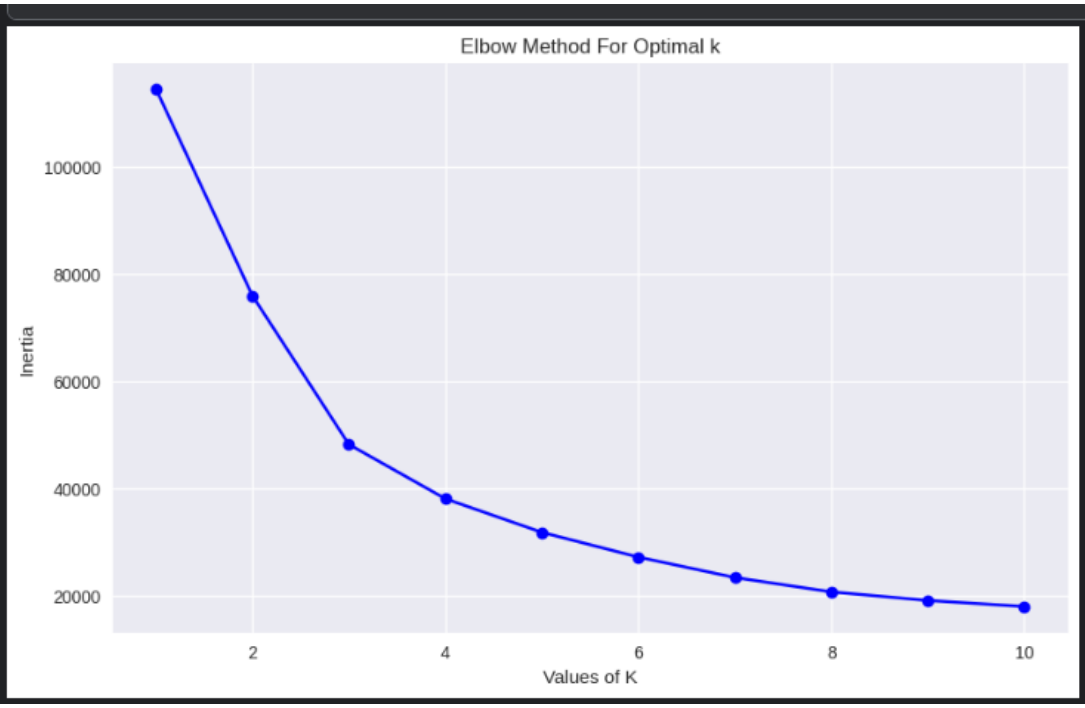
Q: In the PCA scatter plot, we see three distinct colored regions. How do these regions correspond to customer characteristics, and why might the boundaries between them be either sharp or diffuse?

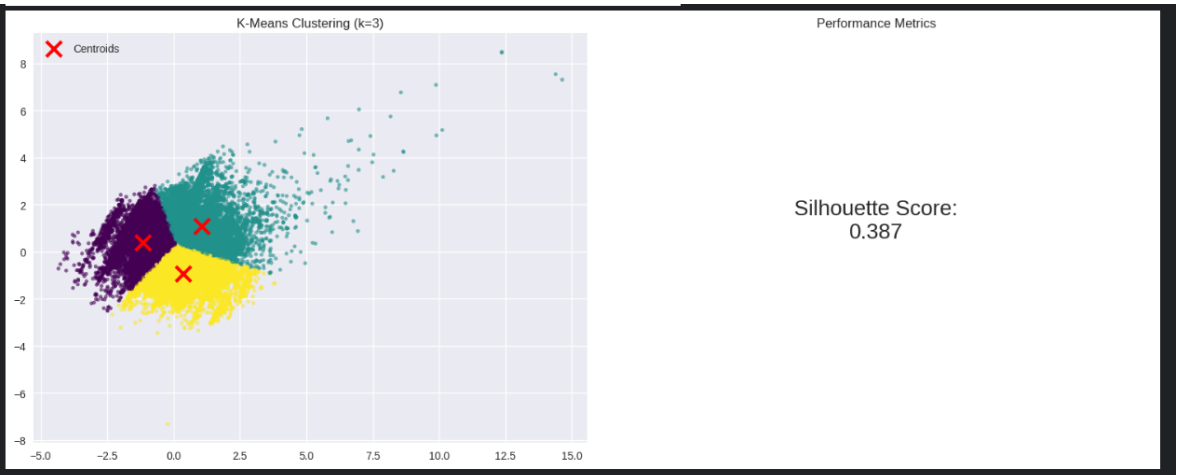
Sol:

- **Correspondence:** The regions correspond to the similarity in feature space (e.g., age + balance + job). Customers in the "Purple" region likely share demographic profiles distinct from those in the "Yellow" or "Green" regions.
- **Boundaries:** The boundaries are somewhat diffuse rather than sharp because customer behavior exists on a spectrum. There is no hard line where a "medium saver" suddenly becomes a "high saver"; there is a transition zone. The overlap in PCA space represents these transitional customers who share traits with multiple groups.

SCREENSHOTS:

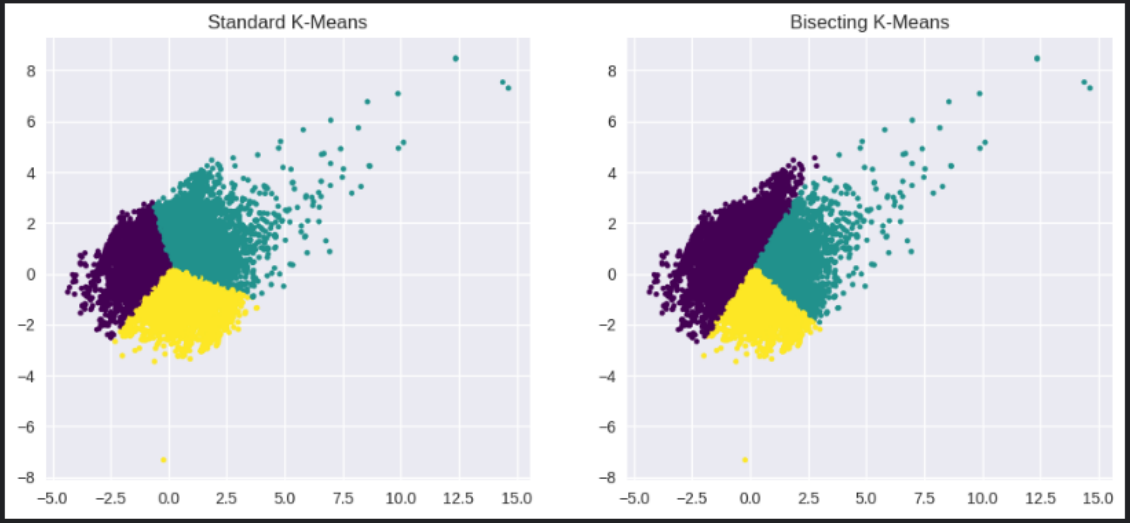







```
Clustering Evaluation:
Inertia: 48179.64
Silhouette Score: 0.39
```

```
Running Bisecting K-Means with 3 clusters...
Standard K-Means Silhouette: 0.387
```




```
Bisecting K-Means Silhouette: 0.338
```