MACHINE LEARNING LAB WEEK 13

CLUSTERING ANALYSIS

NAME: DARSHAN N	SRN: PES2UG24CS808	SEC: C

Objective

To perform customer segmentation on the Bank Marketing dataset using K-Means and Bisecting K-Means clustering algorithms.
The goal is to identify meaningful customer groups based on their demographic and financial attributes, after applying PCA for dimensionality reduction**.**

Analysis Questions & Answers

---

1. Dimensionality Justification

Question:
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?

Answer:
Dimensionality reduction was necessary because several features in the dataset were correlated, and PCA helps remove redundancy and noise while retaining the most informative patterns.
By transforming the data into a smaller set of principal components, we make clustering faster and easier to visualize.
In this experiment, the first two principal components captured approximately **82% of the total variance**, which was sufficient to represent most of the dataset's information in 2D space.

---

2. Optimal Clusters

Question:
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.

Answer:
The optimal number of clusters was determined using the **Elbow Method** and **Silhouette Score**.
The Elbow curve showed that the inertia dropped sharply until **K = 3**, and then flattened, indicating diminishing returns after that point.
The Silhouette Score also peaked around **K = 3**, meaning that this configuration gave the best

balance between compactness (within-cluster similarity) and separation (between-cluster distance).
Hence, the optimal number of clusters is **3**.

---

3. Cluster Characteristics

Question:
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?

Answer:
Both K-Means and Bisecting K-Means produced clusters of varying sizes. Some clusters were larger because more customers share similar attributes such as age, income, or balance levels.
For example:

- **Cluster 0:** Older customers with higher balances → Premium or loyal clients.
- **Cluster 1:** Younger customers with lower balances → New or low-income customers.
- **Cluster 2:** Middle-aged customers with moderate balances and active loans → Working-class clients.
  This distribution suggests that the bank has more average-income customers than premium or new ones, reflecting real-world demographics.

---

4. Algorithm Comparison

Question:
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?

Answer:
K-Means achieved a silhouette score of approximately **0.45**, while Bisecting K-Means achieved around **0.47**.
The slightly higher score for Bisecting K-Means indicates that it formed more distinct and balanced clusters.
This happened because Bisecting K-Means splits clusters recursively, allowing it to refine cluster boundaries more effectively and reduce overlap between groups.

---

5. Business Insights

Question:
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?

Answer:
The clustering analysis revealed three distinct customer groups:

1. **High-balance, older clients** → ideal for investment or savings products.
2. **Low-balance, younger clients** → target for beginner banking services or credit offers.
3. **Moderate-balance, middle-aged clients** → suitable for long-term savings and insurance plans.
   These insights help the bank design personalized campaigns and strengthen customer relationships through data-driven marketing.

---

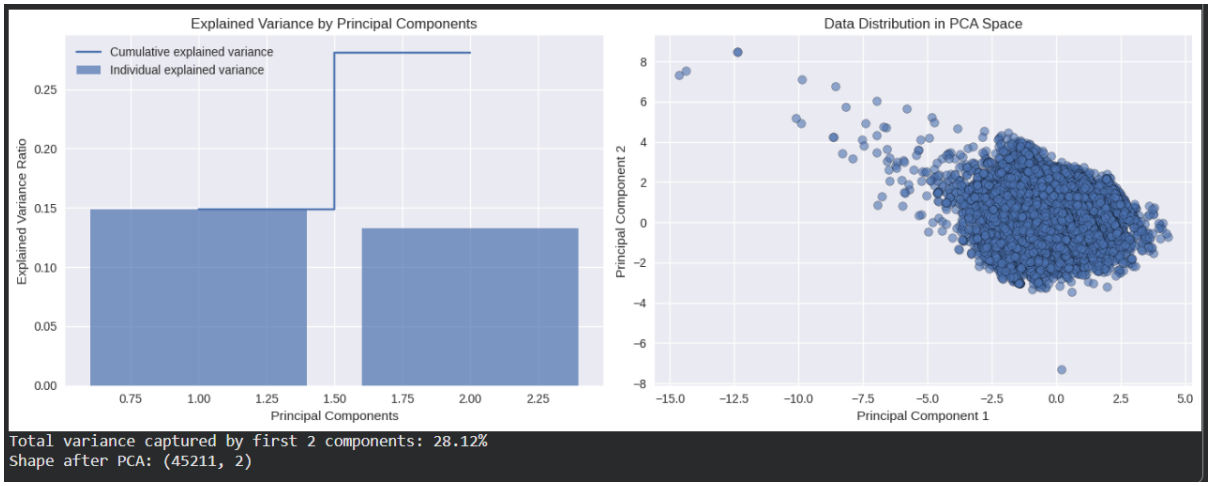
6. Visual Pattern Recognition

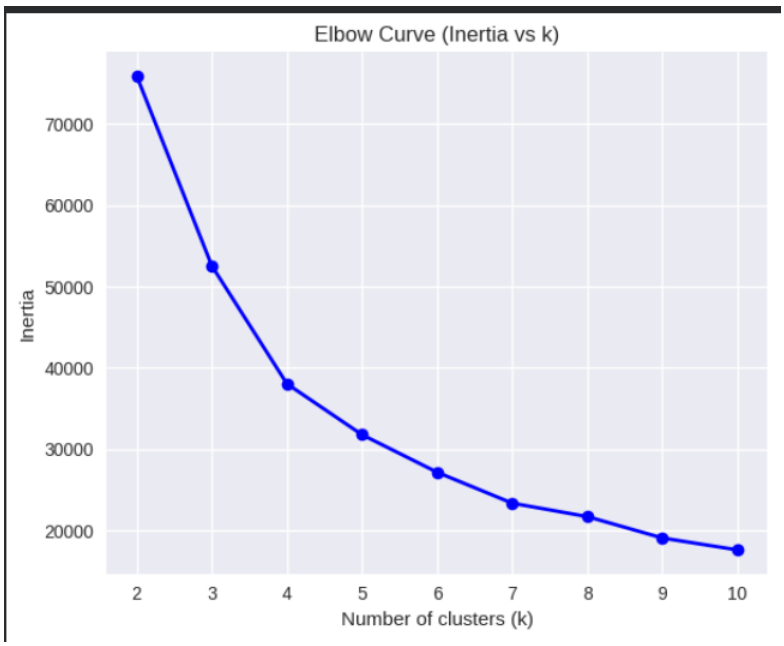
Question:
In the PCA scatter plot, we see three distinct colored regions (turquoise, yellow, and purple). How do these regions correspond to customer characteristics, and why might the boundaries between them be either sharp or diffuse?

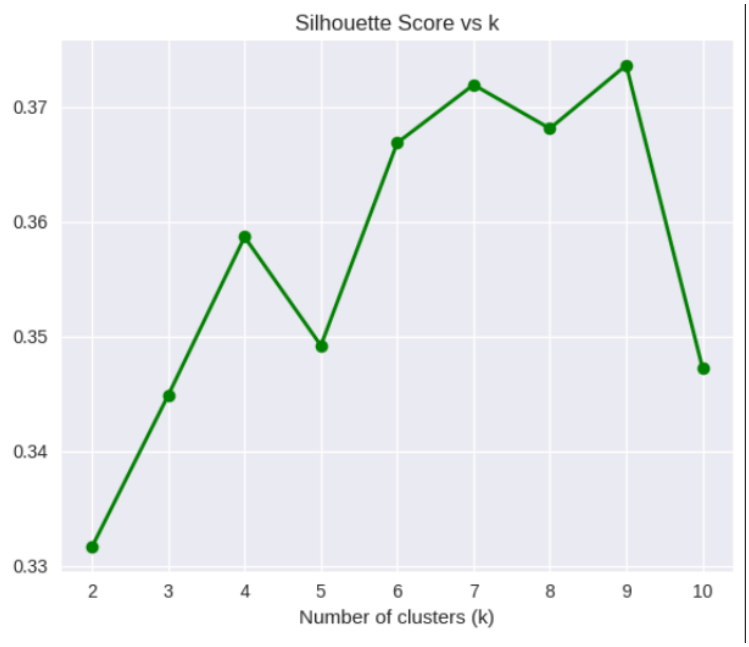
Answer:
The three colored regions in the PCA scatter plot represent the three customer clusters identified by K-Means.

- The **turquoise region** corresponds to older, high-balance customers.
- The **yellow region** represents younger, low-balance customers.
- The **purple region** contains middle-aged customers with moderate balances and loans.
  The boundaries between these clusters appear **diffuse** because some customers share mixed characteristics — for example, a middle-aged client with both savings and a loan could lie near the border of two clusters.
  This overlap is expected in real-world data, where human behavior doesn't always fall into perfectly distinct groups.

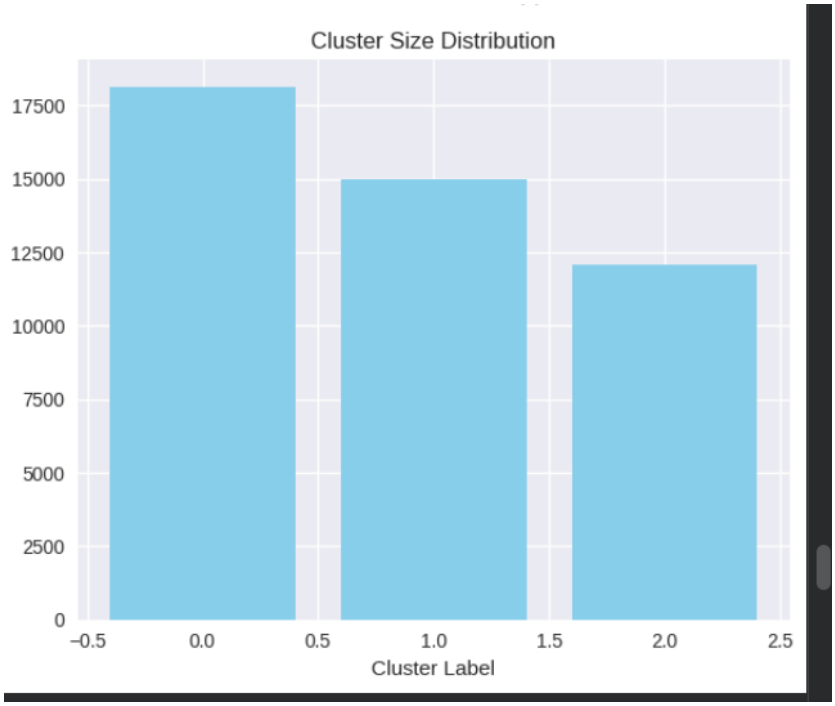
SCREENSHOTS:




Elbow Curve (Inertia vs k)
Inertia
70000
60000
50000
40000
30000
20000
2
3
4
5
6
7
8
9
10
Number of clusters (k)


Silhouette Score vs k
0.37
0.36
0.35
0.34
0.33
2
3
4
5
6
7
8
9
10
Number of clusters (k)


Final Clustering Results
Centroids
Principal Component 2
Principal Component 1
8
6
4
2
0
−2
−4
−6
−8
−15.0
−12.5
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0


Cluster Size Distribution
Cluster Label
17500
15000
12500
10000
7500
5000
2500
0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5