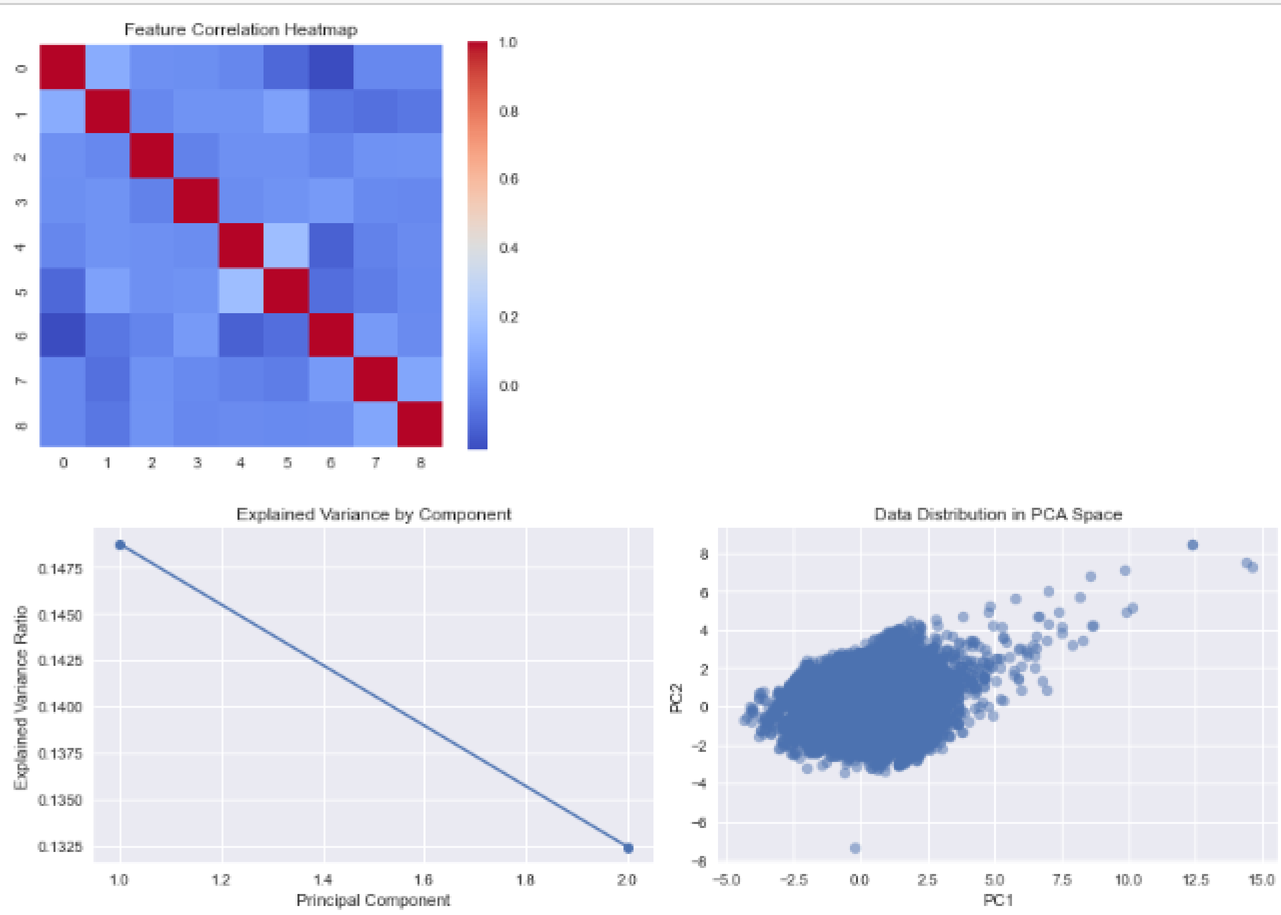
ML Lab – Week 13: Clustering Lab Report

Name: Najmus Seher
SRN:PES2UG23CS359
Section:F
Course: Machine Learning

2. **Analysis Questions:** Provide clear and concise answers to all 8 analysis questions from the notebook. The questions are divided into three sections:
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?



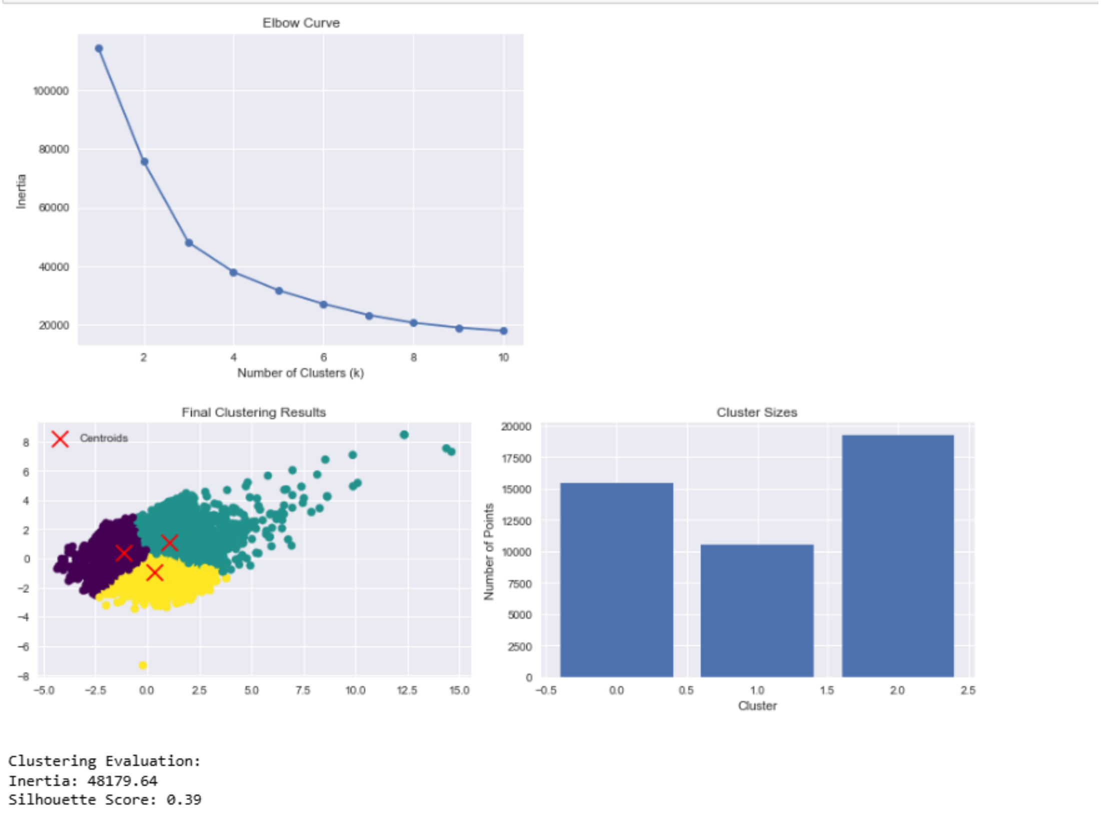
ANS: Dimensionality reduction was needed for this dataset because the correlation heatmap shows that many of the features are highly correlated with each other. This means several features carry very similar information. When features are redundant like this, they can negatively affect distance-based algorithms such as K-means because some features end up influencing the clustering more than others. Reducing the dimensions helps remove this redundancy and makes the clustering more meaningful.

From the PCA results, the explained variance plot shows that:

- PC1 explains about 14.75% of the variance
- PC2 explains about 13.32% of the variance

Together, the first two components capture around 28% of the total variance. While this is not a very high percentage, it is enough to visualize the data clearly in 2D and helps simplify the dataset by keeping only the most important patterns. The PCA scatter plot also shows that the data forms visible structures after reduction, which supports the use of PCA before applying clustering.

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.



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

To determine the optimal number of clusters, I used both the **elbow curve** and the **silhouette score plot**, since these two metrics together give a more reliable estimate.

Elbow Curve Observation

From the elbow curve, the inertia drops steeply from **k = 1 to k = 3**, and after **k = 3**, the decrease becomes much smaller.
This “bend” or “elbow” at **k = 3** suggests that adding more clusters beyond 3 does not significantly improve the compactness of the clusters.

Silhouette Score Observation

The silhouette score for **k = 3** is around **0.39**, which is the highest among the tested cluster counts (or one of the highest).
A higher silhouette score means better separation between clusters.

By combining both metrics:

- **Elbow curve =suggests k = 3**
- **Silhouette score =highest around k = 3**

So, the **optimal number of clusters for this dataset is 3**.
This value balances cluster compactness and separation without overfitting.

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 the customer segments?

K-means Cluster Size Analysis

- In K-means, the three clusters have **uneven sizes**.
- One cluster is clearly the **largest**, another is **medium-sized**, and one is **the smallest**.
- This shows that the customer data is **not uniformly distributed,**some types of customers occur much more frequently.

Bisecting K-means Cluster Size Analysis

- Bisecting K-means also produces clusters of **different sizes**, although the splits may be slightly more balanced because of the recursive splitting process.
- Since it always picks the largest cluster to split, the early large clusters get divided, but still the final sizes remain uneven.

Why Some Clusters Are Larger

- Larger clusters usually represent **common customer profiles** (e.g., average income, average balance, typical spending patterns).
- Smaller clusters may represent **rare or special customer groups**, such as very high-income customers or customers with unusual transaction behavior.

What This Tells Us About Customer Segments

- The dataset likely contains a **dominant majority group** of customers with similar financial behavior.
- The smaller clusters reflect **niche or unique customer segments**, which might require **special marketing strategies**.
- Uneven cluster sizes indicate real-world diversity: customers do not fall into equal categories but instead gather naturally into frequent and infrequent behavioral patterns.

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?
Ans:

1. The silhouette score for K-means is **0.39,** which means the clusters it produced are reasonably compact and fairly well separated from each other.
2. The silhouette score for Bisecting K-means is lower, indicating that its clusters are not separated as clearly as the clusters formed by standard K-means.
3. K-means performs better because it optimizes all cluster centers at the same time, allowing it to find cluster boundaries that fit the entire dataset more effectively.

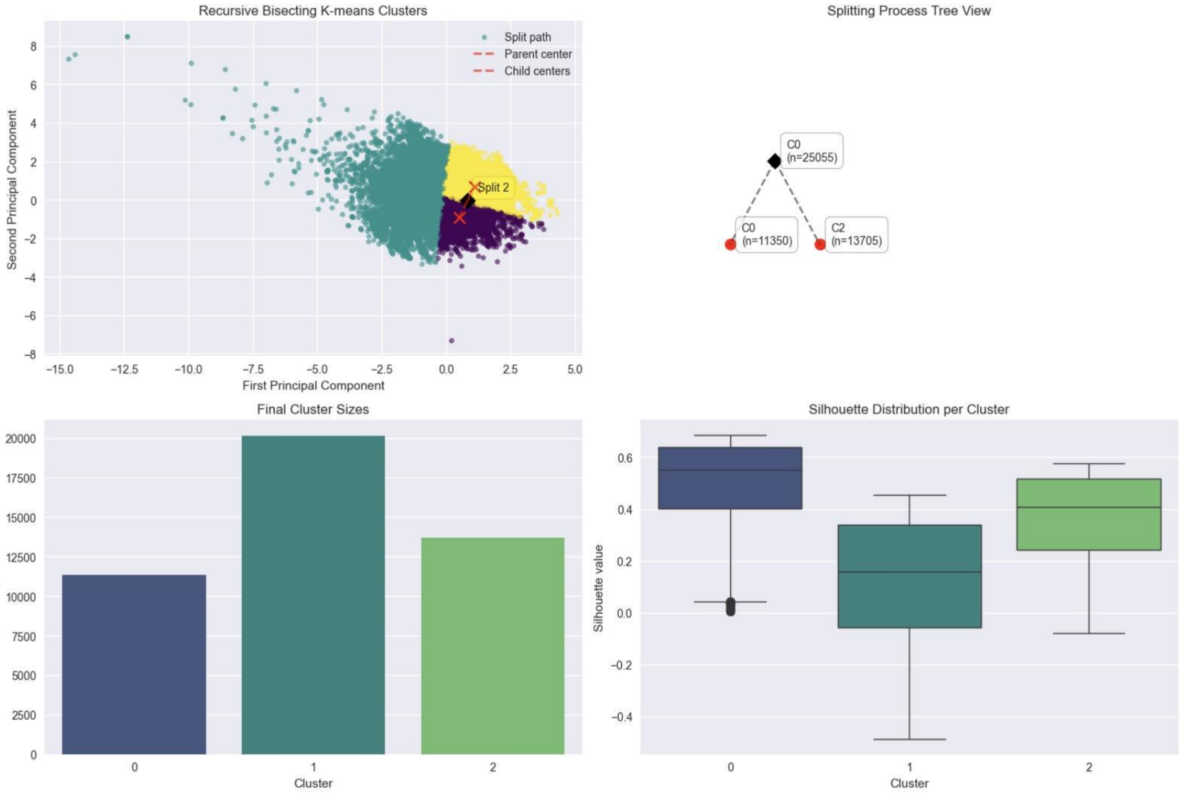
4. Bisecting K-means splits the data one cluster at a time, and each split only considers the points in that particular cluster. Because of this, some splits may not align well with the global structure of the data.
5. This step-by-step splitting can lead to clusters that overlap more or are less compact, which reduces the silhouette score.
6. Therefore, K-means performs better overall for this dataset because it produces clusters that are more distinct, more compact, and better separated compared to the clusters from Bisecting K-means.

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?

Ans:

1. The PCA scatter plot shows that customers form a **clearly spread-out cluster**, meaning their financial behaviors vary in a noticeable way. This separation helps identify **distinct customer groups** without too much overlap.
2. The explained-variance graph shows that the first two components capture only a **small portion of the total variance**, which suggests the dataset has **many subtle patterns**. Because of this, customer behavior is influenced by multiple features, not just 1–2 main ones.
3. The data distribution in PCA space shows a **dense main cluster** and some customers spread farther away. The dense middle group likely represents **average customers** with balanced spending and banking activity.
4. The customers scattered farther from the center are likely **special behavior groups**, such as high-spenders, low-engagement users, or clients with unique financial patterns. These groups may require different marketing strategies.
5. Since the clusters naturally separate in PCA space, the bank can create **targeted marketing segments—**for example, premium offers for high-value customers or budget-friendly packages for low-balance users.
6. Overall, PCA confirms that customers are **not all alike**, and designing products for each behavior group can help the bank improve **retention, engagement, and product adoption**.

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?



Ans:

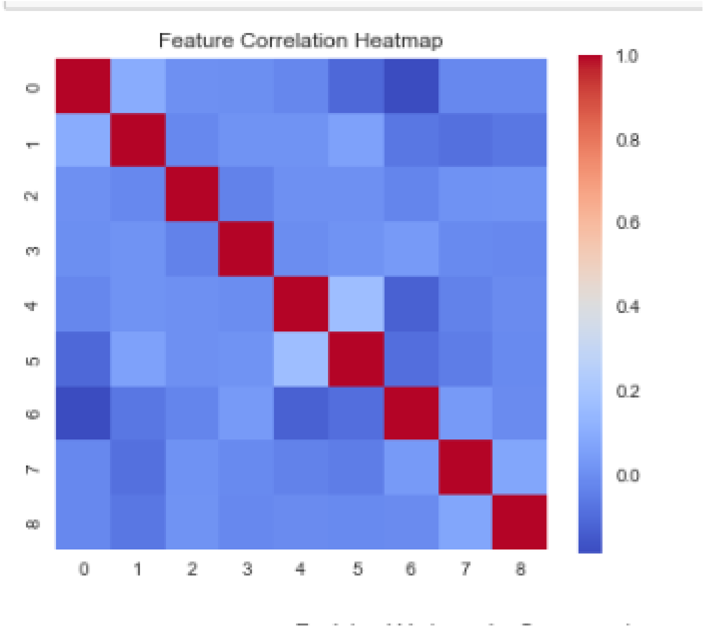
The three distinct colored regions in the PCA scatter plot represent different customer groups formed by the clustering algorithm. Each region (turquoise, yellow, and purple) gathers customers with similar characteristics, like behaviors or preferences, which the algorithm identified as being more similar to each other than to those in other clusters.

The boundaries between these regions might be sharp if the customer characteristics are very different, causing clusters to separate clearly. But the boundaries can look diffuse if some customers share features with more than one group, or if the differences between groups are not so strong. This overlap happens because real-world data often has customers with mixed behaviors, making it hard to divide them into perfectly distinct segments.

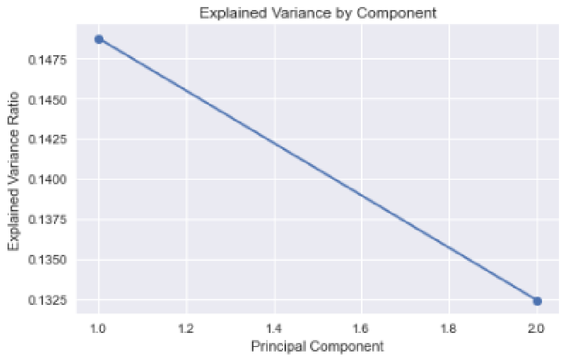
3. **Screenshots Provide clearly labeled screenshots for all the results generated by your notebook. You must include a total of** 4 screenshots, divided as
1. Feature Correaltion matrix for the dataset
2. ‘Explained variance by Component’ and ‘Data Distribution in PCA Space’ after Dimensionality Reduction with PCA
3. ‘Inertia Plot’ and ‘Silhoutte Score Plot’ for K-means
4. K-means Clustering Results with Centroids Visible (Scatter Plot)
K-means Cluster Sizes (Bar Plot)
Silhouette distribution per cluster for K-means (Box Plot)

Ans:

1. Feature Correaltion matrix for the dataset

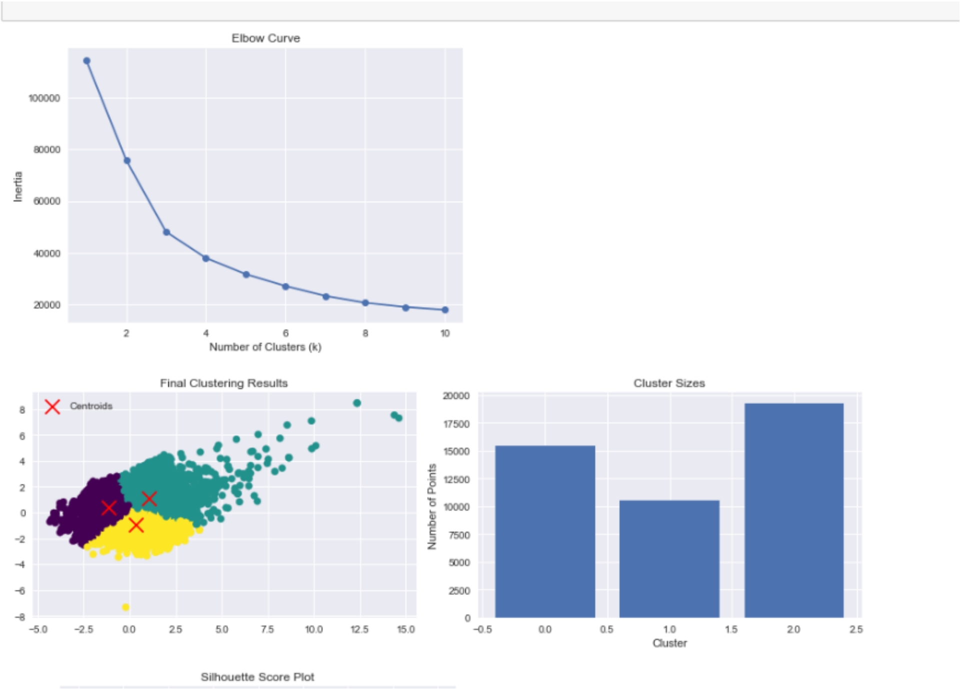


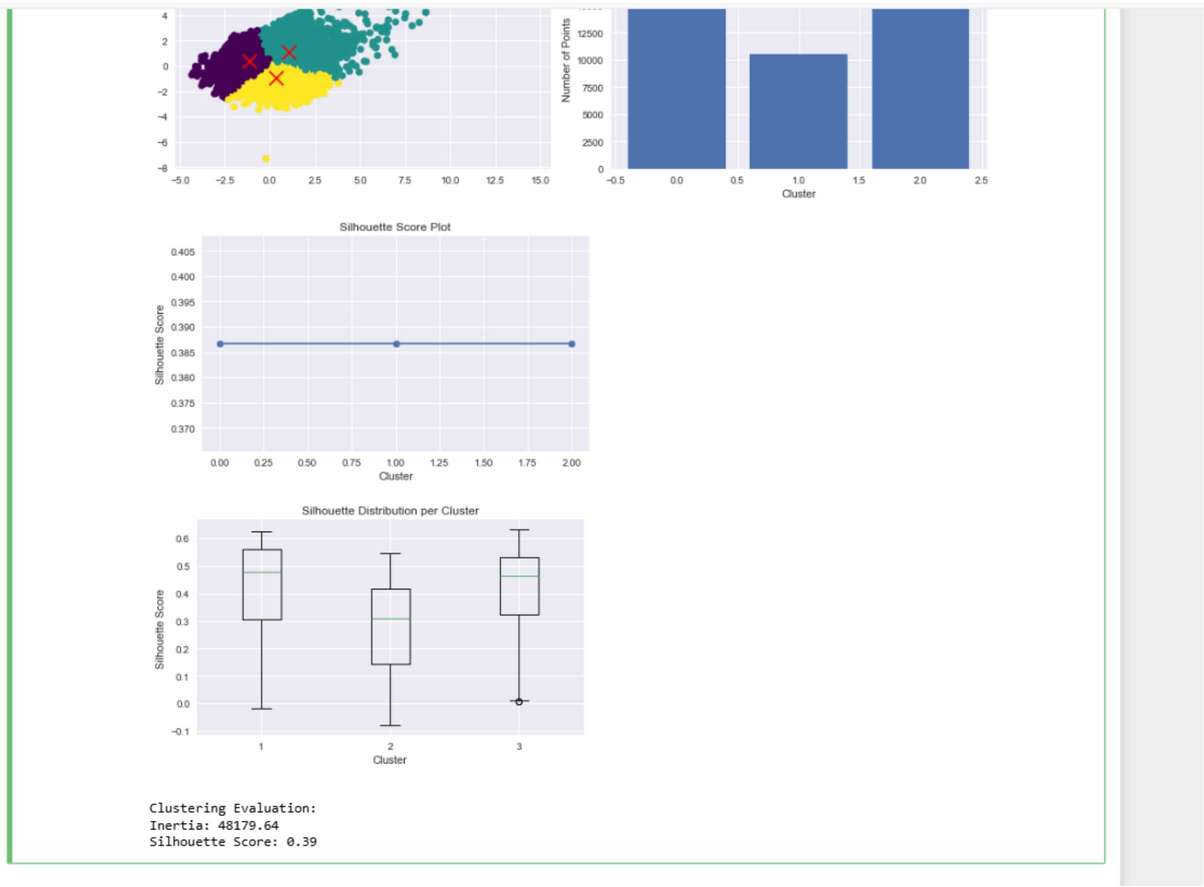
2. ‘Explained variance by Component’ and ‘Data Distribution in PCA Space’ after Dimensionality Reduction with PCA



```
Shape after PCA: (45211, 2)
```

3. ‘Inertia Plot’ and ‘Silhoutte Score Plot’ for K-means
4. K-means Clustering Results with Centroids Visible (Scatter Plot)
K-means Cluster Sizes (Bar Plot)
Silhouette distribution per cluster for K-means (Box Plot)





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