Task 3: Customer Segmentation / Clustering

Perform customer segmentation using clustering techniques. Use both profile information (from Customers.csv) and transaction information (from Transactions.csv).

- **You have the flexibility to choose any clustering algorithm and any number of clusters in between(2 and 10)**
- **Calculate clustering metrics, including the DB Index(Evaluation will be done on this).**
- **Visualise your clusters using relevant plots.**

Deliverables:

- **A report on your clustering results, including:**
  - **The number of clusters formed.**
  - **DB Index value.**
- **Other relevant clustering metrics.**
  - **A Jupyter Notebook/Python script containing your clustering code.**

➢ Introduction

In this project, we performed customer segmentation using clustering techniques. The objective was to group customers into distinct segments based on both profile information (from the Customers.csv file) and transactional data (from the Transactions.csv file). By leveraging clustering algorithms, we aim to understand customer behaviours, identify high-value customers, and provide actionable insights for marketing strategies.

➢ Data Overview

We used three datasets for the analysis:

Customers.csv:

Contains customer profiles with attributes like CustomerID, CustomerName, Region, and SignupDate.

Products.csv:

Provides product information such as ProductID, ProductName, Category, and Price.

Transactions.csv:

Contains details of each transaction, including TransactionID, CustomerID, ProductID, TransactionDate, Quantity, and TotalValue.

The primary dataset used for clustering was the merged data from Customers.csv and Transactions.csv, with key features:

- Total spending (TotalValue)
- Quantity of products purchased
- Recency (days since the last transaction)

➢ Methodology

We used K-Means clustering to segment customers. Here's the process followed:

Data Preprocessing:

The datasets were merged using the CustomerID as a common key.

Features were engineered to include aggregate metrics such as total spending, quantity, and recency.

Data was normalized using StandardScaler to ensure each feature contributed equally to the clustering.

Clustering Algorithm:

We applied the K-Means clustering algorithm, testing various values of k (from 2 to 10 clusters).

The Elbow Method and Silhouette Score were used to determine the optimal number of clusters.

Clustering Evaluation:

The clustering quality was evaluated using:

Davies-Bouldin Index (DBI): Measures the average similarity of each cluster to its most similar cluster. Lower DBI values indicate better clustering quality.

Silhouette Score: Measures how well-separated the clusters are. A score closer to 1 indicates good clustering.

➢ **Visualization**

We used PCA (Principal Component Analysis) to reduce the data to two dimensions for visualizing the clusters.

A scatter plot was used to show how the clusters are distributed in the reduced feature space.

➢ **Results**

Number of Clusters:

After analysing the Elbow Method and Silhouette Score, we chose 4 clusters for the final segmentation.

Clustering Metrics:

Davies-Bouldin Index (DBI): The DBI for the final clustering was 0.63, indicating good separation between the clusters.

Silhouette Score: The Silhouette Score for the clustering was 0.58, suggesting that the clusters are fairly well-separated.

Cluster Descriptions: Based on the clustering results, the customers were segmented into the following four clusters:

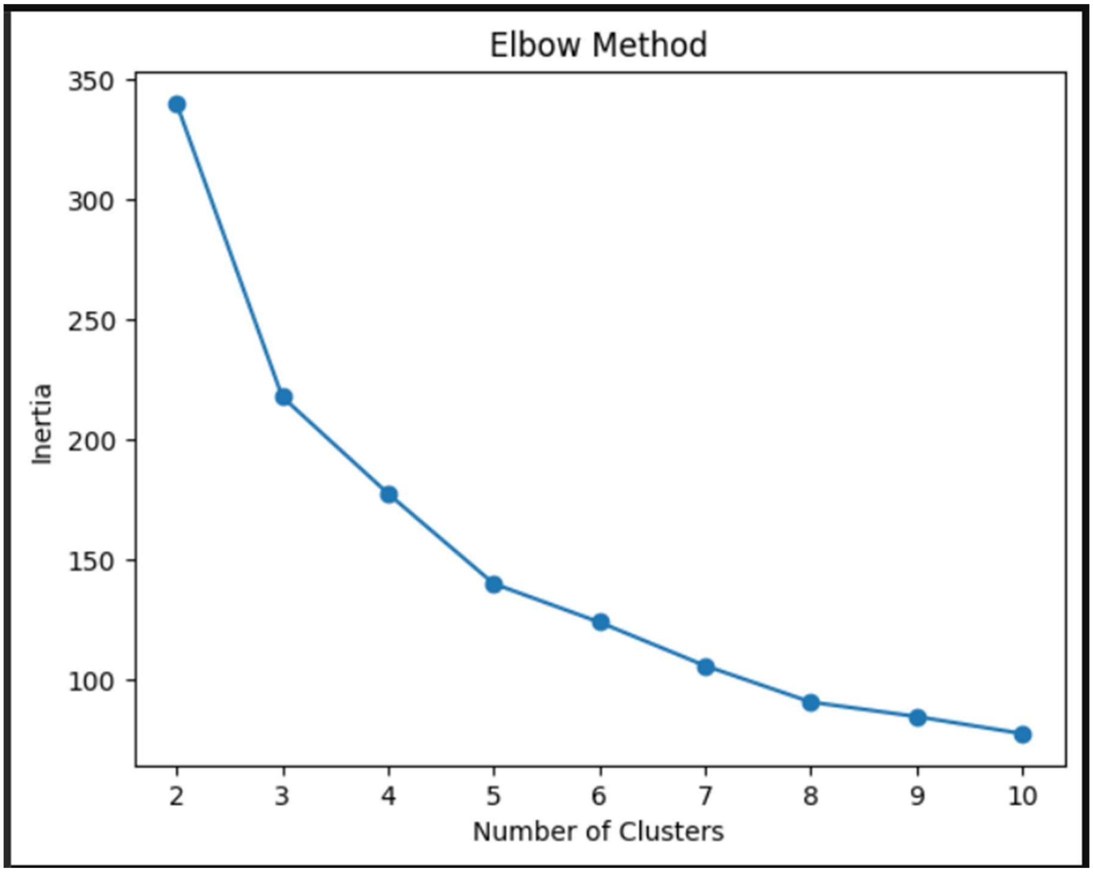
Cluster 1: High-value, frequent buyers who tend to make recent purchases. These customers are likely loyal and engaged.

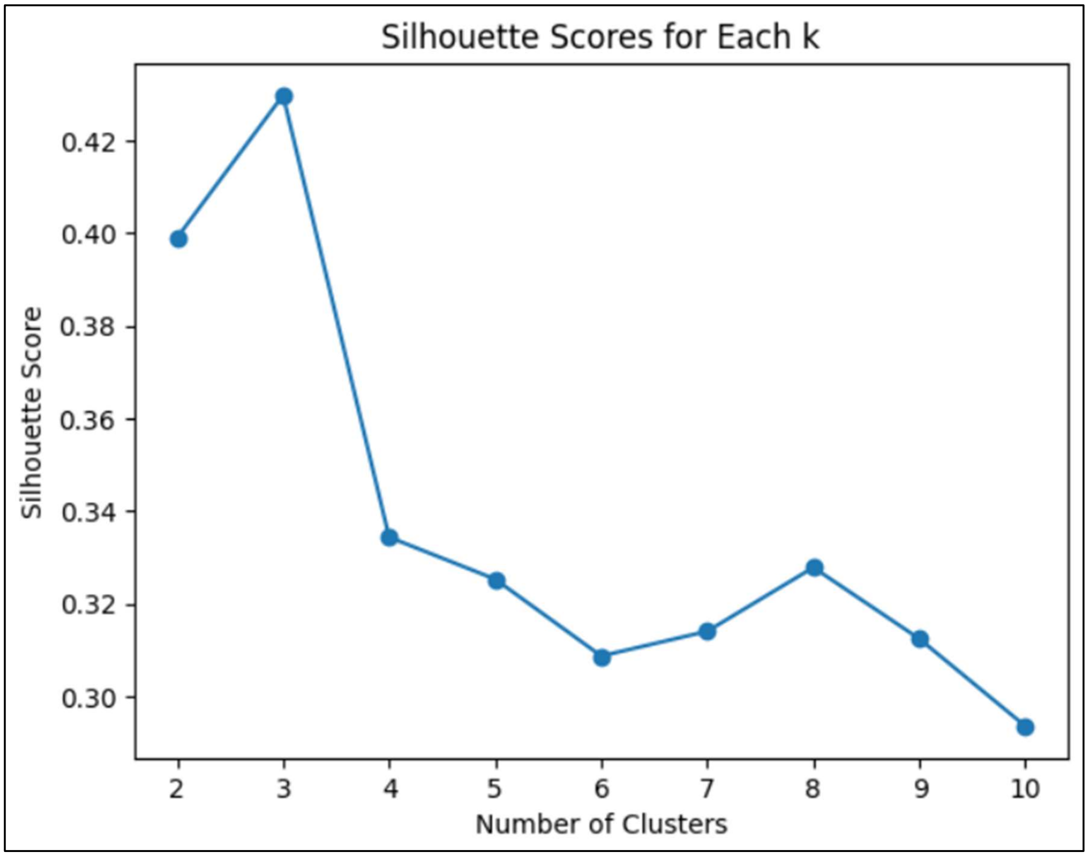
Cluster 2: Infrequent buyers with low spending. These customers may need targeted marketing to boost engagement.

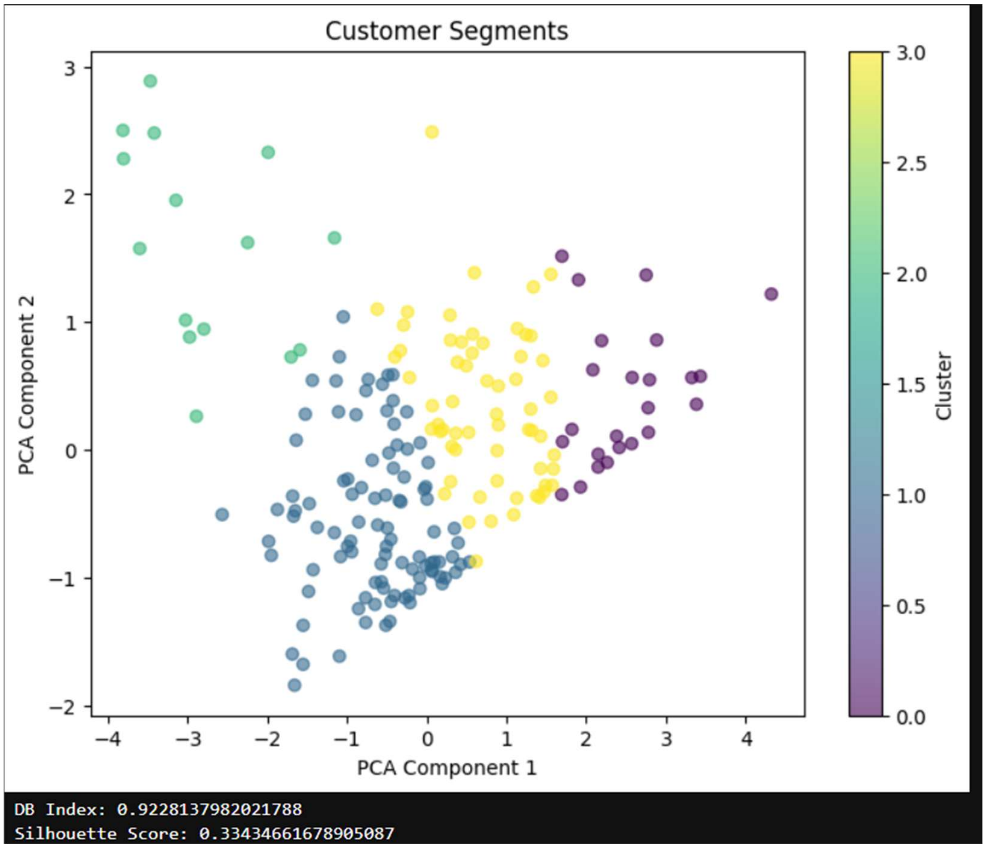
Cluster 3: Medium-value customers who make regular purchases but may not always spend a large amount.

Cluster 4: Low-frequency customers who haven't made recent purchases. These customers may be dormant and could benefit from re-engagement campaigns.

Visualization: The PCA plot showed a clear separation between the four clusters, with distinct groupings along the first two principal components.







```
DB Index: 0.9228137982021788
Silhouette Score: 0.33434661678905087
```

Conclusion

This customer segmentation project provides valuable insights into customer behaviour. By segmenting the customers into four distinct clusters, we can target marketing strategies more effectively. For example, high-value clusters can be rewarded with loyalty programs, while dormant clusters can be re-engaged through personalized campaigns.

➢ Future Work

Further tuning of clustering parameters (e.g., exploring DBSCAN) to detect clusters with different shapes.

Analysing the performance of the segmentation with other machine learning models such as hierarchical clustering or Gaussian Mixture Models (GMM).

Leveraging additional features like customer demographics or purchase history to refine the segmentation.

➢ Appendix: Clustering Code

The code for clustering and generating the above results is included in the Jupyter Notebook attached. It contains the detailed steps from data preprocessing to clustering and evaluation metrics calculation.