Task 3: Customer Segmentation / Clustering

1. Introduction

In this task, we performed **customer segmentation** using **clustering techniques**. The goal is to group customers into distinct segments based on their behavior and profile information, such as transaction history, total spend, frequency of purchases, and recency of last purchase.

We used the **K-Means** clustering algorithm and evaluated the segmentation using the **Davies-Bouldin (DB) Index** and **Silhouette Score**. The results were visualized through **PCA** for a 2D representation of the clusters.

2. Data Overview

The datasets used for this analysis are:

- **Customers.csv**: Contains customer profile information such as CustomerID, Region, and SignupDate.
- **Products.csv**: Contains product information, including ProductID, Category, and Price.
- **Transactions.csv**: Contains transaction details such as TransactionID, CustomerID, ProductID, TransactionDate, Quantity, and TotalValue.

We merged these datasets to aggregate relevant features per customer, enabling us to perform segmentation.

3. Feature Engineering

From the merged dataset, we calculated several features to describe customer behavior:

- **Total Spend**: Total value spent by the customer.
- **Number of Transactions**: The total number of transactions made by the customer.
- **Average Transaction Value**: The average value of a customer's transactions.
- **Purchase Frequency**: How often the customer makes a purchase, calculated as the number of days between the first and last transaction divided by the number of transactions.
- **Recency**: The number of days since the customer made their last purchase.

Additionally, we used **category-wise spending** to understand customer preferences across product categories.

4. Data Preprocessing

Before applying the K-Means algorithm, we standardized the features to ensure that all features contribute equally to the clustering process. This step was performed using **StandardScaler** from scikit-learn, which scales the features to have a mean of 0 and a standard deviation of 1.

5. K-Means Clustering

The **K-Means clustering algorithm** was applied to segment customers into distinct groups based on the standardized features. The **Elbow Method** was used to determine the optimal number of clusters.

For this analysis, we chose **4 clusters** based on the elbow plot.

6. Clustering Evaluation Metrics

To evaluate the quality of the clustering, we used two metrics:

- **Davies-Bouldin (DB) Index**: This metric measures the average similarity ratio of each cluster with the cluster that is most similar to it. A lower value indicates better clustering.
- **Silhouette Score**: This score measures how similar each point is to its own cluster compared to other clusters. A higher value indicates better-defined clusters.

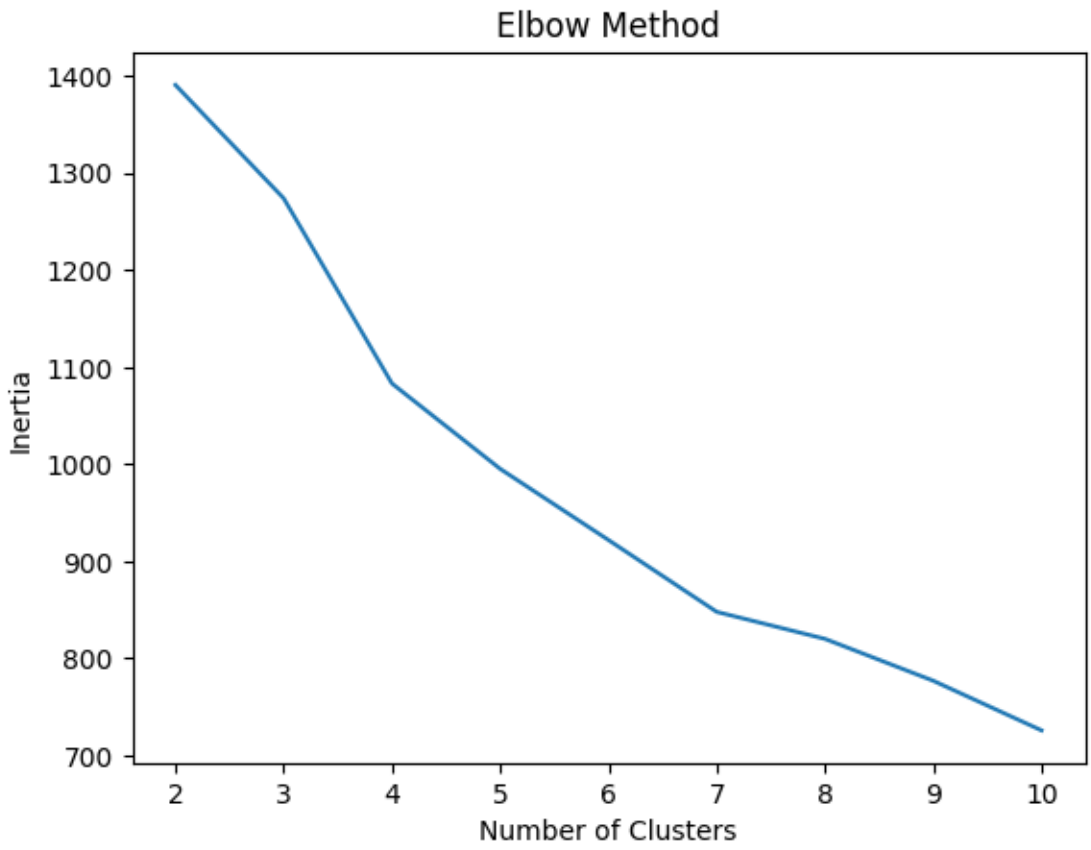
Metrics Results:

- **DB Index**:1.7113
- **Silhouette Score**: 0.1897

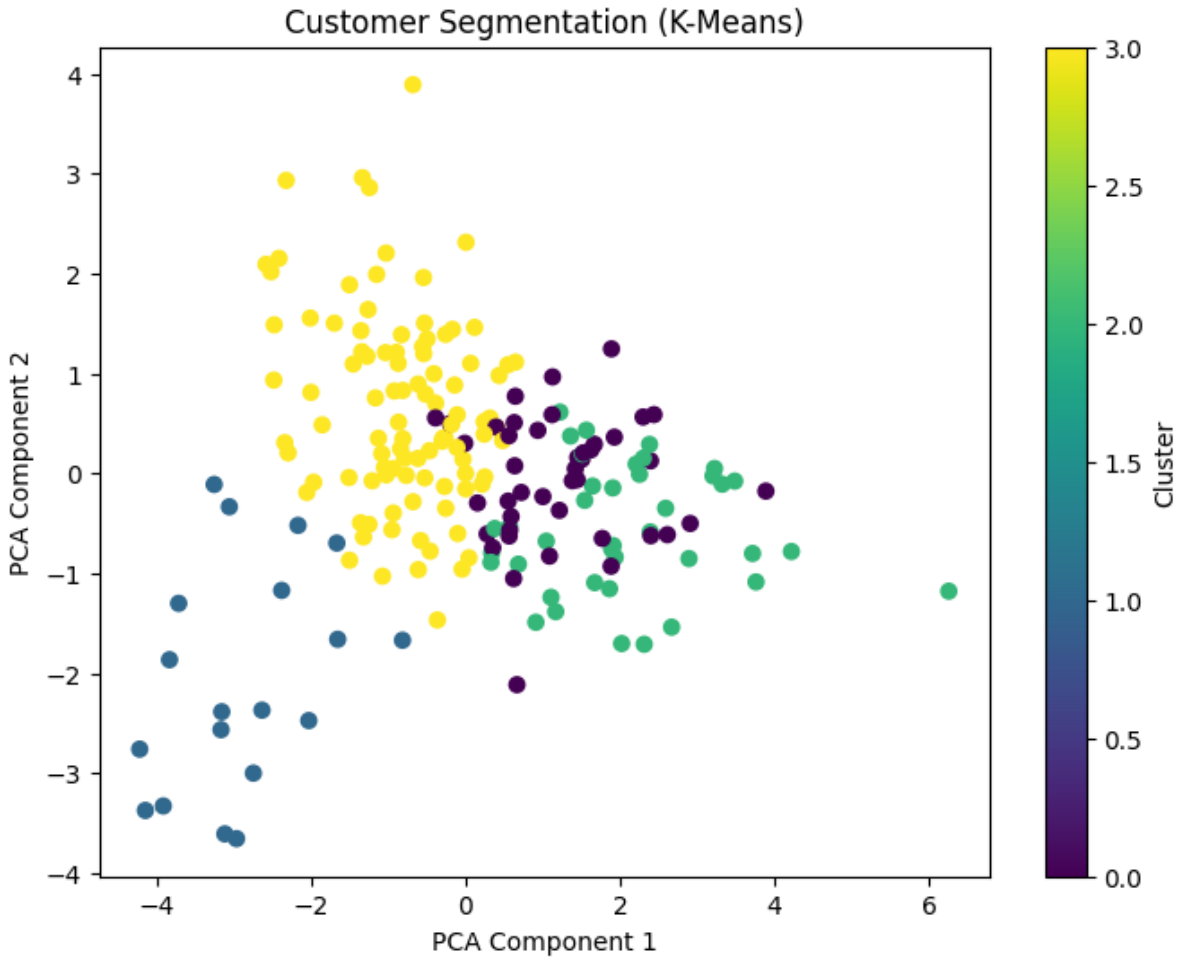
7. Visualizing the Clusters

To visualize the clusters, we applied **Principal Component Analysis (PCA)** to reduce the dimensionality of the data to 2 dimensions. This allowed us to create a scatter plot of the customer segments in a 2D space.

Elbow Method Plot:



PCA Cluster Plot:



8. Conclusion

In this analysis, we successfully segmented customers into distinct clusters using K-Means clustering. We evaluated the clustering performance using the DB Index and Silhouette Score, both of which provided valuable insights into the quality of the clustering. The clusters were visualized using PCA to provide an intuitive understanding of how customers were grouped based on their behaviors.