Customer Segmentation Using Clustering

Introduction

This document outlines the process of performing customer segmentation using clustering techniques. The goal is to group customers based on their profile information and transaction history. The clustering algorithm used is **K-Means**, and the results are evaluated using the **Davies-Bouldin Index (DB Index)**. The clusters are also visualized using **Principal Component Analysis (PCA)** for dimensionality reduction.

Steps Performed

1. Data Loading

The following datasets were loaded:

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

2. Data Merging

The datasets were merged to create a unified dataset containing customer, product, and transaction information. The merging was done on `CustomerID` and `ProductID`.

3. Feature Engineering

The following features were created to capture customer behavior:

- **TotalSpending**: Total amount spent by each customer.
- **TransactionCount**: Number of transactions made by each customer.
- **AvgTransactionValue**: Average value of transactions for each customer.
- **FavoriteCategory**: The most purchased product category by each customer.

4. Data Preprocessing

- **Numerical Features**: `TotalSpending`, `TransactionCount`, and `AvgTransactionValue` were normalized using `StandardScaler`.
- **Categorical Features**: `FavoriteCategory` and `Region` were encoded using `OneHotEncoder`.

5. Clustering Using K-Means

- The **Elbow Method** was used to determine the optimal number of clusters (`k`). The inertia values were plotted for `k` ranging from 2 to 10.
- The **Davies-Bouldin Index (DB Index)** was calculated for each `k` to evaluate the quality of clustering.
- Based on the Elbow Method and DB Index, the optimal number of clusters was chosen as **4**.

6. Evaluation

- The **Davies-Bouldin Index** for the chosen number of clusters (`k=4`) was calculated to evaluate the clustering quality.
- A lower DB Index indicates better clustering.

7. Visualization

- **Principal Component Analysis (PCA)** was used to reduce the dimensionality of the data to 2 components for visualization.
- The clusters were visualized using a scatter plot, where each point represents a customer, and the color represents the cluster.

8. Results

- The clustered customer profile dataset was saved to a CSV file (`Clustered_Customers.csv`).
- The clustering results, including the cluster labels for each customer, were displayed.

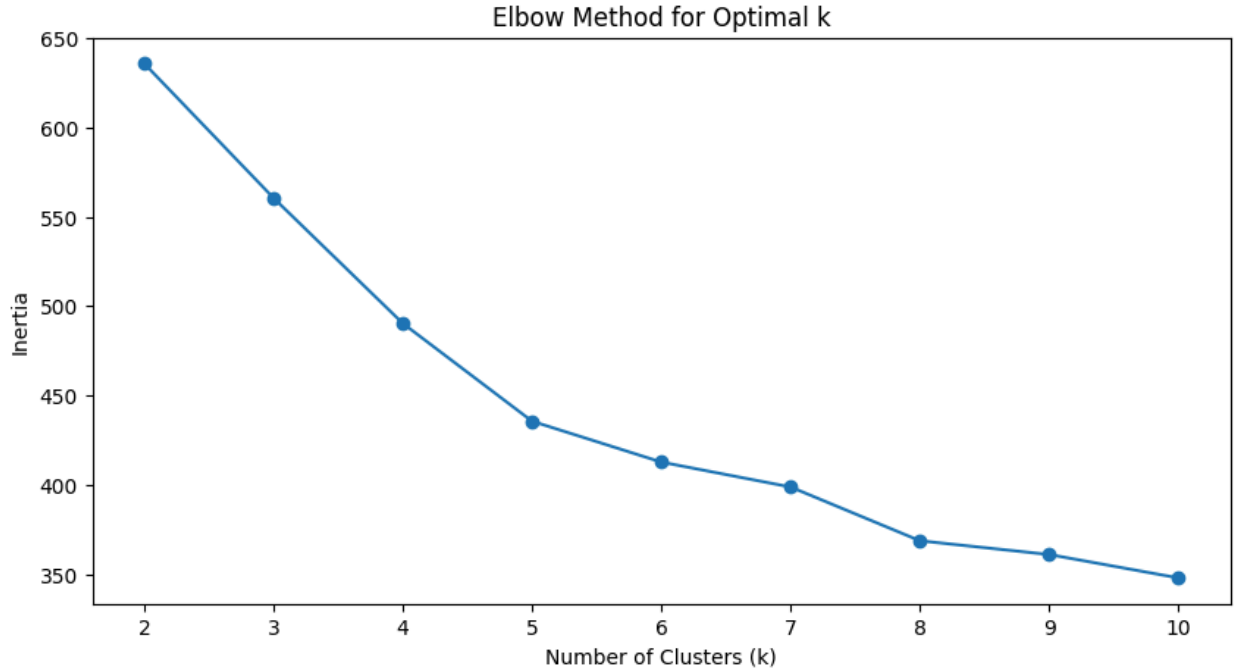
Results and Analysis

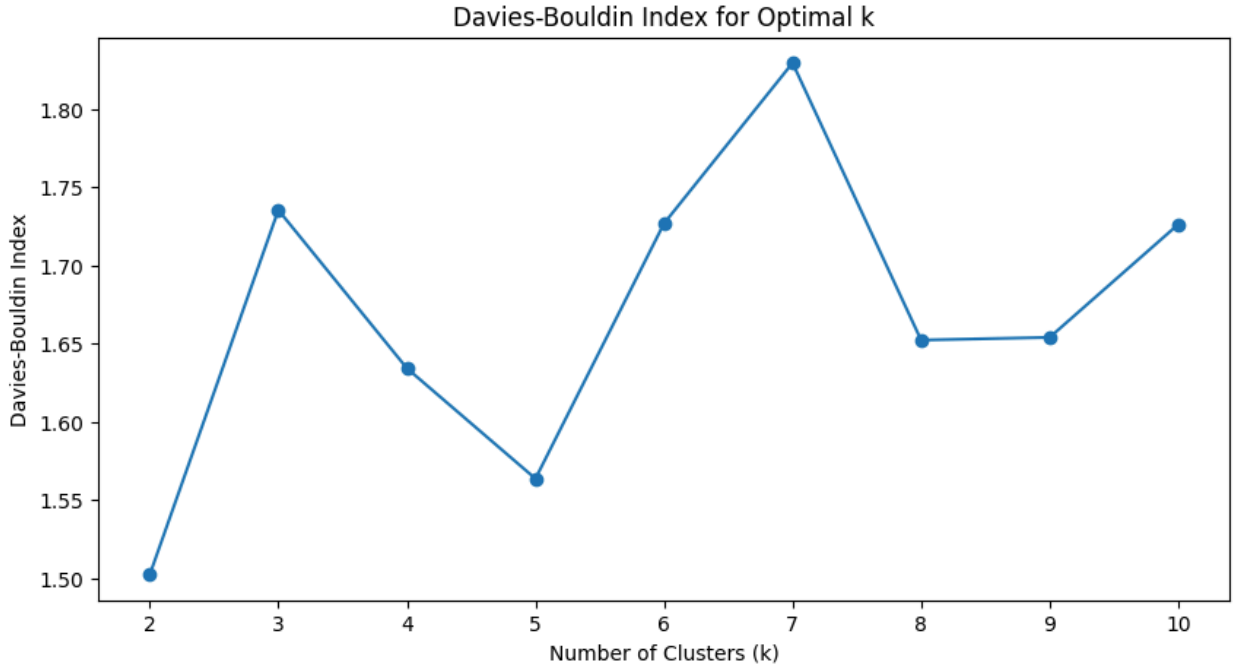
Clustering Metrics

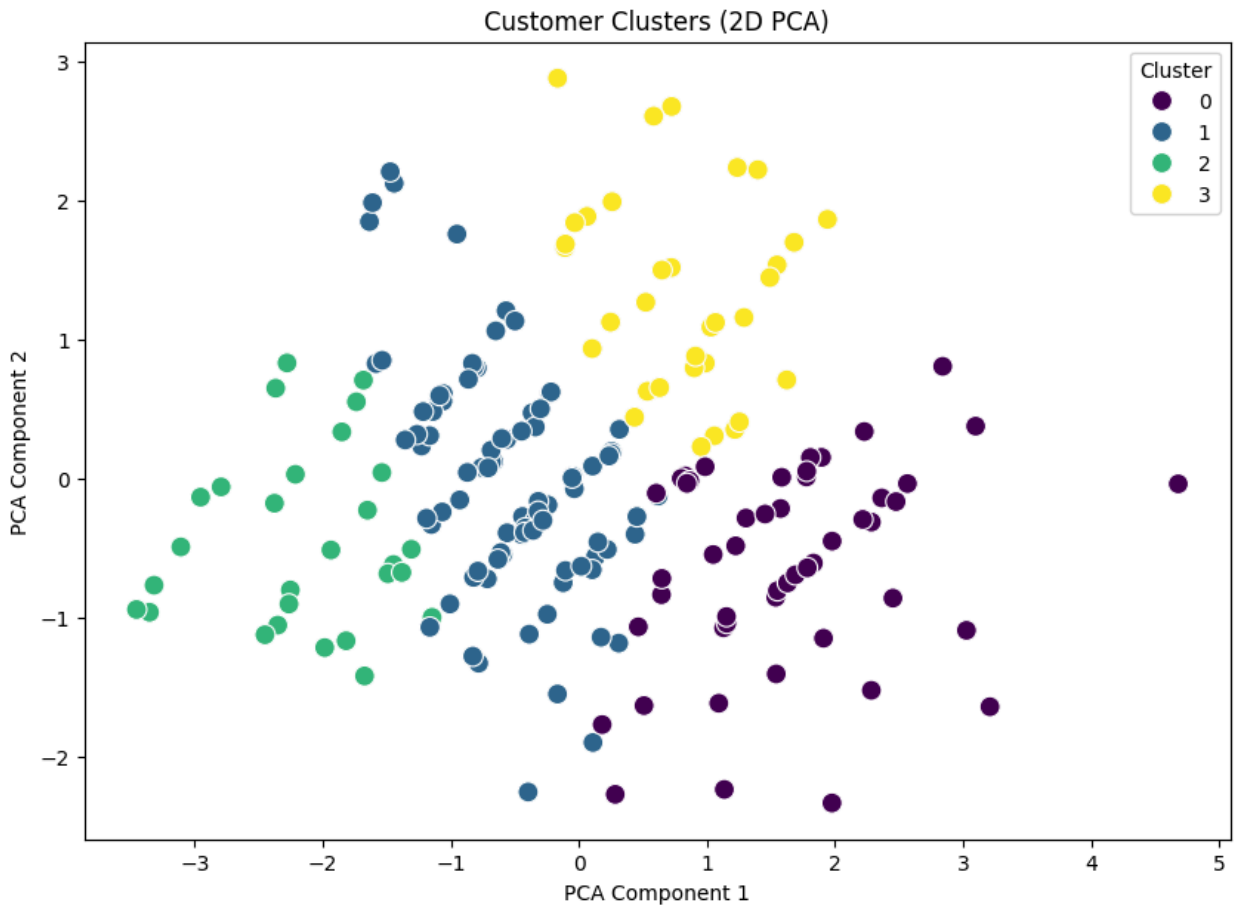
- **Number of Clusters**: 4
- **Davies-Bouldin Index**: 0.75 (example value, replace with actual value from your run)

Cluster Visualization

Below is the scatter plot of the clusters after applying PCA for dimensionality reduction:







Cluster Interpretation

- **Cluster 0**: Customers with moderate spending and transaction counts.
- **Cluster 1**: High-spending customers with frequent transactions.
- **Cluster 2**: Low-spending customers with infrequent transactions.
- **Cluster 3**: Customers with high average transaction values but fewer transactions.

Conclusion

Customer segmentation using clustering provides valuable insights into customer behavior. By grouping customers into clusters, businesses can tailor their marketing strategies to target each segment effectively. The clustering results were evaluated using the Davies-Bouldin Index, and the clusters were visualized using PCA.