Customer Segmentation / Clustering Report

Introduction

This report presents the results of the customer segmentation analysis performed using K-Means clustering. The analysis utilizes customer profile information from the Customers.csv dataset and transaction data from the Transactions.csv dataset. The goal is to identify distinct customer segments based on their purchasing behavior and demographic characteristics.

Data Preparation

Merging Datasets

The analysis begins by merging the Transactions dataset with the Customers dataset on the CustomerID field. This allows for the aggregation of transaction data alongside customer demographic information.

Feature Engineering

Customer features were created by aggregating transaction data to derive meaningful metrics:

TotalSpent: The total amount spent by each customer.

TransactionCount: The total number of transactions made by each customer.

UniqueProducts: The number of unique products purchased by each customer.

Region: The geographical region of the customer.

One-hot encoding was applied to the Region column to facilitate clustering.

Customer Features DataFrame

The resulting DataFrame contains the following columns:

CustomerID

TotalSpent

TransactionCount

UniqueProducts

One-hot encoded region columns

Clustering Process

Standardization

To ensure that all features contribute equally to the distance calculations in K-Means, the features were standardized using StandardScaler.

K-Means Clustering

The K-Means algorithm was applied to the standardized features. The range of clusters tested was from 2 to 10. For each cluster size, the inertia and Davies-Bouldin (DB) Index were calculated to evaluate clustering performance.

Clustering Metrics

Inertia: Measures how tightly the clusters are packed. Lower values indicate better clustering.

Davies-Bouldin Index: A lower DB Index indicates better clustering quality, as it measures the average similarity ratio of each cluster with its most similar cluster.

The following results were obtained for different cluster sizes:

Clusters	Inertia	DB Index
2	837.70	1.49
3	646.74	1.25
4	522.03	1.06
5	384.20	0.97
6	327.83	0.97
7	281.34	0.86
8	255.84	0.83
9	234.26	0.90
10	190.10	0.85

Optimal Number of Clusters

Based on the analysis of inertia and the DB Index, the optimal number of clusters was determined to be 4. This choice balances compactness and separation of clusters effectively.

Clustering Results

The final K-Means model was fitted with the optimal number of clusters, and each customer was assigned to a cluster. The resulting DataFrame includes the following columns:

CustomerID

TotalSpent

TransactionCount

UniqueProducts

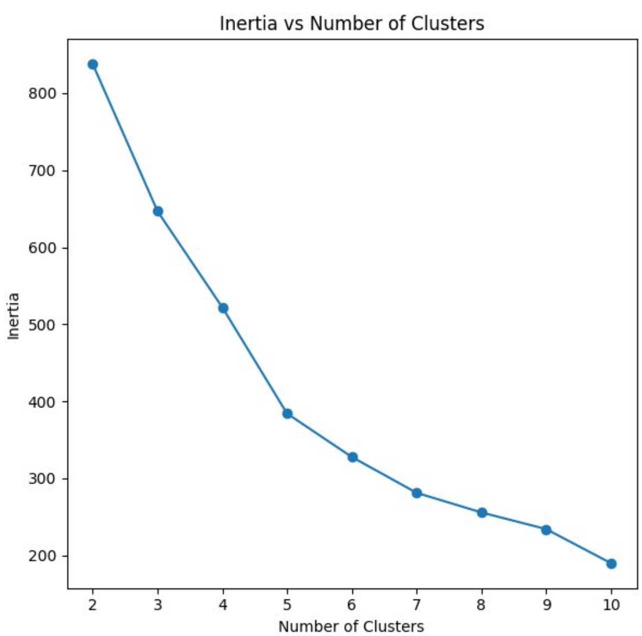
Cluster

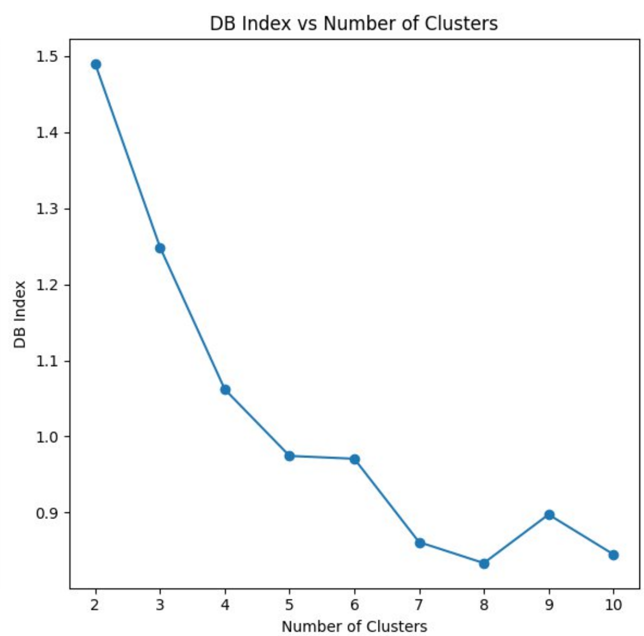
Visualization of Clusters

To visualize the clusters, Principal Component Analysis (PCA) was employed to reduce the dimensionality of the data to two components. The clusters were then plotted in a 2D space.

Inertia vs Number of Clusters

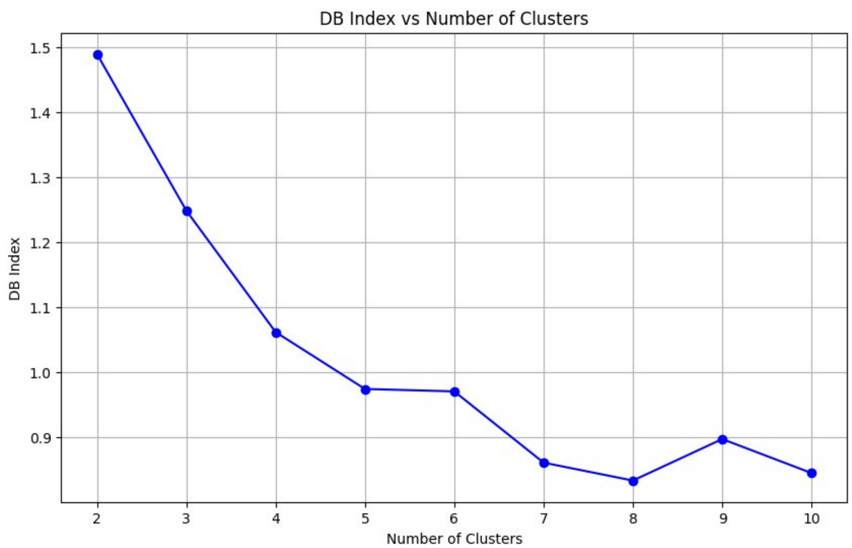
The plot below illustrates the relationship between the number of clusters and inertia, showing how inertia decreases as the number of clusters increases.





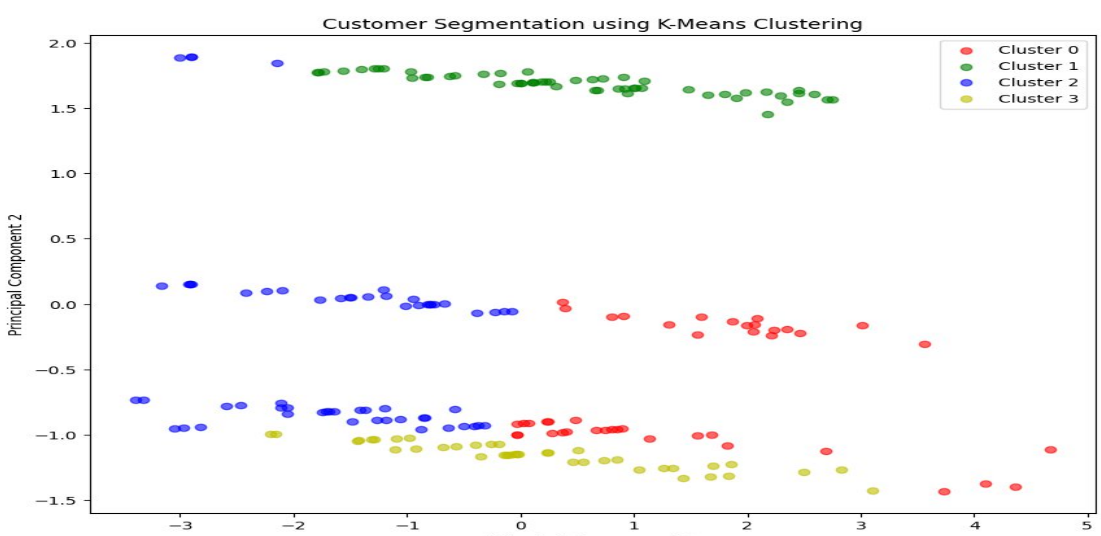
DB Index vs Number of Clusters

The plot below shows the Davies-Bouldin Index for different cluster sizes, indicating the quality of clustering.



Customer Segmentation Visualization

The following scatter plot visualizes the customer segments in PCA space, with each color representing a different cluster.



Insights Derived from Clustering

Distinct Customer Segments: The clustering analysis revealed four distinct customer segments, each characterized by different spending behaviors and transaction patterns. Understanding these segments allows for tailored marketing strategies that can enhance customer engagement and retention.

Targeted Marketing Opportunities: By identifying the characteristics of each cluster, the business can develop targeted marketing campaigns that resonate with specific customer groups. For example, high-spending customers may respond well to loyalty programs, while price-sensitive customers may benefit from promotional discounts.

Optimized Resource Allocation: The insights gained from the clustering can help the business allocate resources more effectively. For instance, marketing budgets can be directed towards segments that show the highest potential for growth or profitability.

Improved Customer Experience: Understanding the unique needs and preferences of each customer segment can lead to improved customer experiences. Tailoring product offerings and communication strategies to align with customer expectations can foster loyalty and increase satisfaction.

Future Growth Opportunities: The segmentation analysis provides a foundation for future growth strategies. By continuously monitoring customer behavior and preferences, the business can adapt its offerings and marketing strategies to meet evolving market demands.

Conclusion

The K-Means clustering analysis identified four distinct customer segments based on purchasing behavior. This optimal segmentation allows for targeted marketing, better resource allocation, and improved customer experiences. Overall, the insights gained provide a solid foundation for data-driven decision-making and future growth.