Task 3: Customer Segmentation / Clustering

By Sarika Kushwaha

1. Introduction

The goal of this analysis was to segment customers based on their transactional behaviors and cluster them into meaningful groups. By understanding the spending and transaction patterns, I can provide insights into different customer segments, which can guide personalized marketing, better product offerings, and strategies for improving customer retention.

2. Data Preparation

<u>Data Loading and Merging:</u> I started by loading the two main datasets: *Customers.csv* and *Transactions.csv*. These datasets were merged based on CustomerID to create a unified view (*merged_df*), where I could see customer profiles alongside their transaction details.

<u>Missing Values Check:</u> After merging the data, I checked for missing values. Fortunately, there were no missing values in the dataset, so no further imputation was needed.

<u>Outlier Detection and Removal:</u> I used Z-scores to detect any outliers in numerical columns like Quantity, TotalValue, and Price. Any data points with Z-scores above 3 were considered outliers and were removed from the dataset. After this step, I was left with a cleaner dataset that I used for further analysis.

3. Transaction-Based Features

To create meaningful features for segmentation, I aggregated the transaction data by CustomerID to calculate the following metrics:

Total Spend: The total amount spent by a customer in all transactions.

Number of Transactions: The total number of transactions a customer has made.

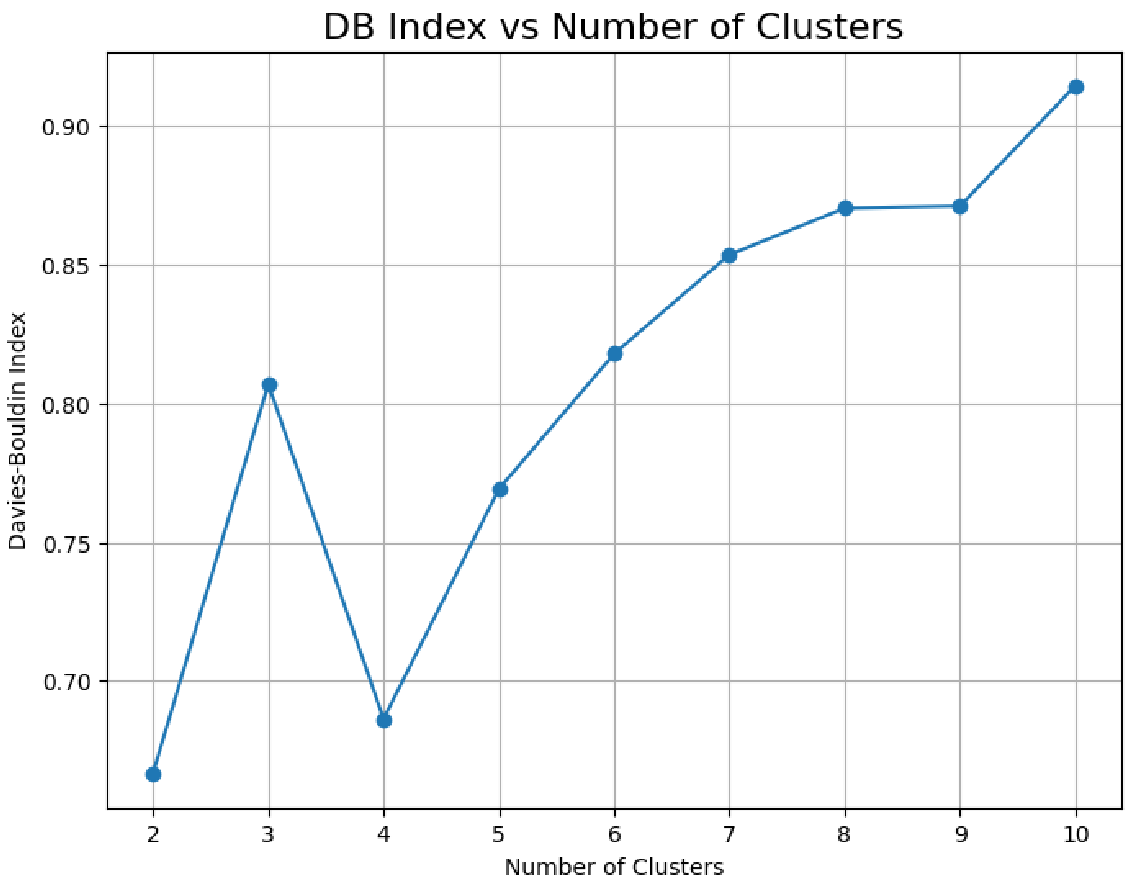
Average Transaction Value: The mean value of all transactions made by the customer.

Transaction Frequency: The number of unique transaction dates, reflecting how frequently the customer makes purchases.

For clustering, I removed irrelevant columns such as *CustomerID*, *CustomerName*, *SignupDate*, *TransactionID*, *ProductID*, *TransactionDate*, and *Region*, as they didn't contribute to segmenting customers based on their transactional behavior. The remaining features, such as total spend, transaction count, and transaction frequency, were used to build the dataset (*clustering_data*) for clustering.

4. KMeans Clustering and Davies-Bouldin Index

For segmentation, I applied KMeans clustering on the customer data. To decide the optimal number of clusters, I ran the clustering for different values of k (from 2 to 10) and calculated the Davies-Bouldin Index (DB Index) for each. The DB Index is a measure of cluster quality, where a lower value indicates better cluster separation.



The result showed that 2 clusters gave the best separation, with the lowest DB Index (0.6666983068858876). This means that 2 clusters provide the most meaningful segmentation based on the data.

5. Silhouette Score Evaluation

To further assess the quality of the clustering, I calculated the Silhouette Score, which measures how well-separated the clusters are. A higher score indicates better-defined clusters.

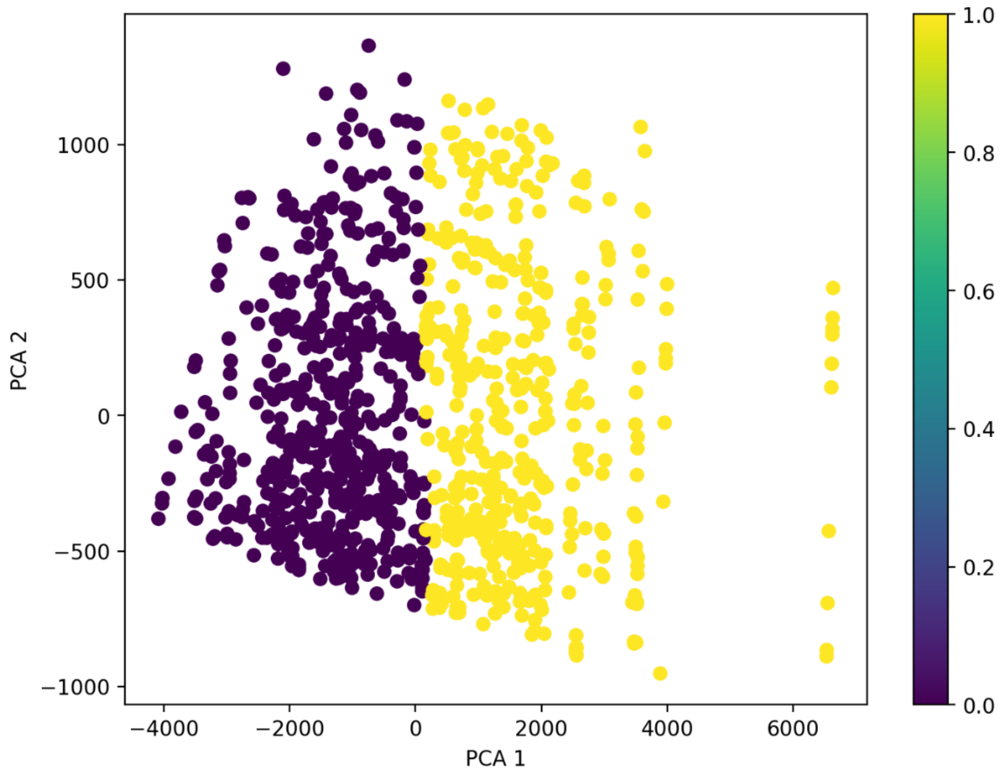
Silhouette Score for **optimal_cluster=2 is 0.5228**

This score suggests that the two clusters are reasonably well-separated and that the clustering model is effective.

6. PCA for Visualization

To help visualize the clustering results, I used Principal Component Analysis (PCA) to reduce the dimensionality of the data to 2 principal components. This allowed me to plot the customers in a 2D space and see the separation between the clusters.

The PCA plot showed that the clusters were separated, confirming that 2 clusters is an appropriate choice for this segmentation.



7. Cluster Profiles

Cluster 0 (Low Spend, Low Frequency):

Average Spend: $593.97

Total Spend: $2,761.88

Transaction Frequency: 4.88 transactions per customer

Cluster 1 (High Spend, High Frequency):

Average Spend: $809.28

Total Spend: $5,764.99

Transaction Frequency: 7.35 transactions per customer

8. Conclusion and Insights

The two clusters I identified represent distinct customer segments: low spenders with fewer transactions (Cluster 0) and high spenders with more frequent transactions (Cluster 1).

Customers in Cluster 1 are the most valuable, as they spend more and engage more frequently. They should be the focus of loyalty programs, special offers, and personalized marketing campaigns.

Cluster 0 customers could be targeted with strategies designed to increase their transaction frequency and spending, such as offering discounts on future purchases or recommending products based on their past behavior.