Clustering Results Report

1. Introduction

Objective:

The goal of this analysis is to segment customers into distinct clusters based on transaction and customer profile data in order to uncover meaningful insights into their behaviour.

Dataset Description:

- **Customers Dataset:** Includes customer information such as CustomerID, CustomerName, SignupDate and Region.
- **Transactions Dataset:** Includes transaction details like TransactionID, CustomerID, ProductID, TransactionDate, Quantity, TotalPrice and Value.

2. Methodology

Features used:

I. Recency (days since last purchase)
II. Frequency (Total number of transactions)
III. Monetary (total spend)
IV. Total unique Products Purchased

Clustering Algorithm:

For customer segmentation, we employed the KMeans clustering algorithm, a widely-used unsupervised learning technique that partitions the data into K clusters based on feature similarity.

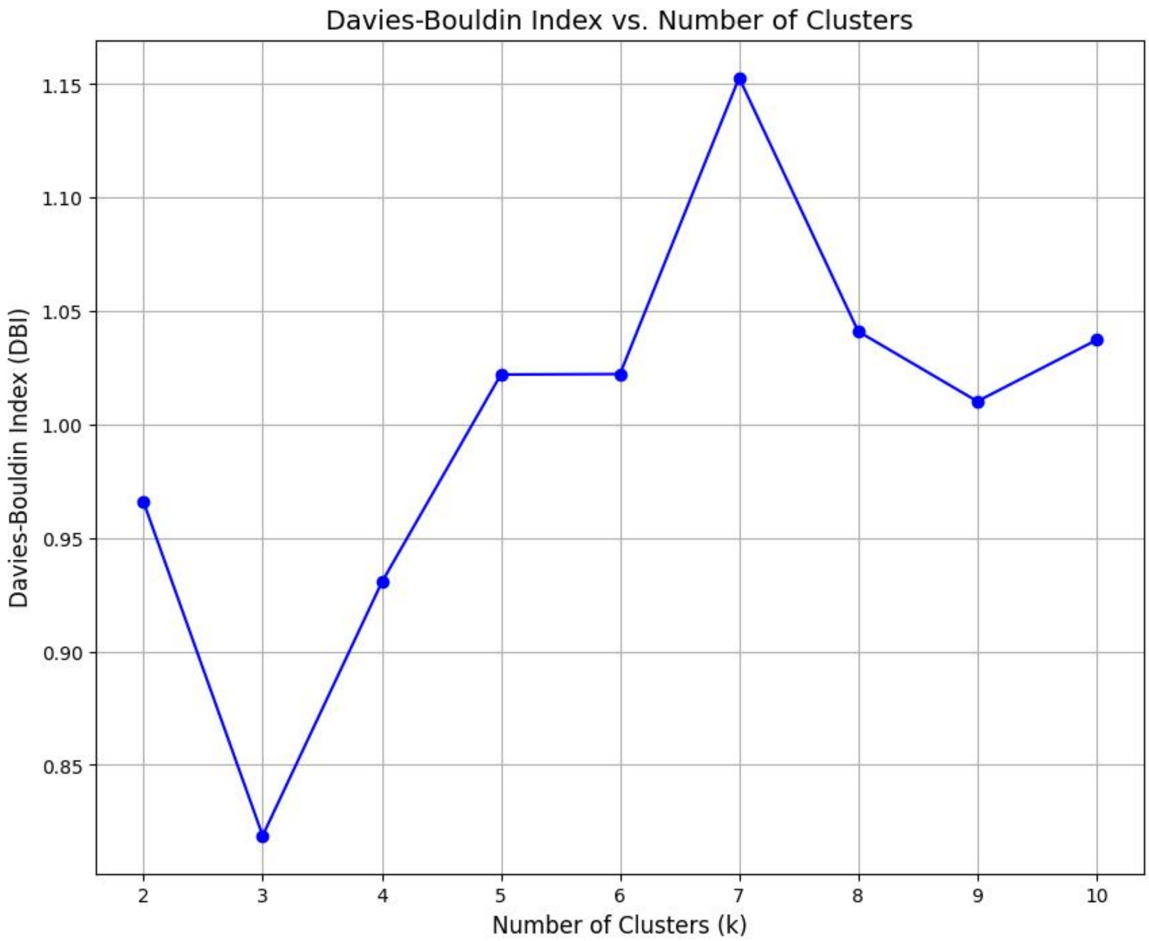
Why KMeans:

- The KMeans clustering algorithm was chosen for segmenting the customer data due to its simplicity, efficiency, and effectiveness in handling datasets with a clear structure.
- **k-means++** initialization was chosen to improve convergence speed and quality of clustering, reducing the likelihood of poor local optima.

Choosing the Optimal Number of Clusters (K): To determine the optimal number of clusters for our data, we used the Davies-Bouldin Index (DB Index).

Steps to Find the Optimal K:

I. We experimented with different values of K (ranging from 2 to 10) and computed the Davies-Bouldin Index for each K.
II. The value of K that yielded the minimum DB index was selected as the optimal number of clusters.



Optimal number of clusters = 3
DB Index value = 0.81

Final Model:

- Once the optimal K was determined, we applied the KMeans algorithm with k-means++ initialization and the optimal number of clusters.
- The cluster labels were added to the original dataset for further analysis.

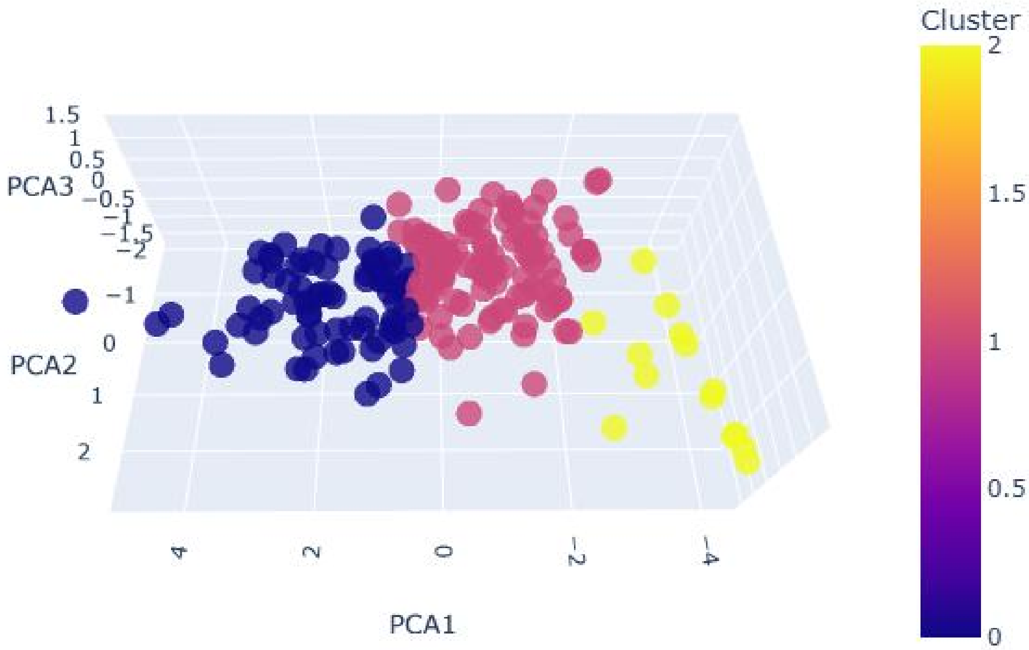
3. Results

- Using KMeans with the optimal number of clusters determined by the Davies-Bouldin Index (0.81), the clusters were moderately compact and well-separated.
- The Silhouette Score of 0.39 suggests that while the clusters exhibit some separation, there is room for improvement in terms of distinctiveness and cohesion.
- The Calinski-Harabasz Index of 161.80 indicates a moderate balance between within-cluster compactness and between-cluster separation.

Number of Clusters = 3
Davies-Bouldin Index (DBI) = 0.81
Silhouette Score = 0.39
Calinski-Harabasz Index = 161.80

Visualization:

To visualize the clusters in a 3D space, we used Principal Component Analysis (PCA) to reduce the high-dimensional data to three principal components.



Observations from the Visualization:

1. The clusters are well-separated in the PCA-reduced space, suggesting that the clustering algorithm successfully segmented the data into distinct groups.
2. Certain clusters appear to overlap slightly, indicating some similarity between customer segments. This could be an area for further analysis or refinement of features.
3. Outliers: A few points might lie far from their assigned clusters, indicating potential outliers or edge cases in the dataset.