Task 3 - Customer Segmentation / Clustering

Clustering Results Report

Introduction

Customer segmentation is essential for businesses to better understand their customers, provide personalized services, and improve profitability.

This report highlights the results of a clustering analysis using the K-Means algorithm on customer transaction and demographic data. The goal was to group customers into meaningful segments based on their behaviour and demographics, determine the ideal number of clusters, and evaluate cluster quality using the Davies-Bouldin Index (DBI).

The process involved cleaning and preparing the data, scaling features for consistency, and applying K-Means with different values of k. The clusters were visualized using Principal Component Analysis (PCA) to uncover patterns and actionable insights.

These insights can help businesses design targeted marketing strategies, optimize product offerings, and enhance customer satisfaction.

Data Overview

This analysis was based on two datasets:

1. **Customers.csv**: Included demographic details like CustomerID, Region, and SignupDate.
2. **Transactions.csv**: Contained transaction-level data, such as CustomerID, TransactionID, Price, and other attributes.

The datasets were merged using CustomerID to create a single, comprehensive dataset for clustering. Key features like total_spend (total amount spent), avg_spend (average spend per transaction), and num_transactions (number of transactions) were derived to represent customer behavior and used in the clustering process.

Data Preprocessing

To prepare the data for clustering, the following steps were taken:

1. **Feature Engineering**:
    - **total_spend**: Total amount spent by each customer.
    - **avg_spend**: Average transaction amount per customer.
    - **num_transactions**: Total number of transactions for each customer.
2. **Categorical Encoding**: The Region column was converted into numerical format using one-hot encoding.
3. **Feature Scaling**: Numerical features were standardized with StandardScaler to ensure they were on the same scale, which is critical for accurate K-Means clustering.

Clustering Analysis

The K-Means algorithm was applied to the prepared dataset with the number of clusters (k) ranging from 2 to 10. The clustering quality was evaluated using the Davies-Bouldin Index (DBI), which measures how well clusters are defined. Lower DBI values indicate better clustering.

The DBI scores for each value of k are as follows:

k=2: 1.756487729921195

k=3: 1.294450280679287

k=4: 1.2062663734927819

k=5: 1.113399776747031

k=6: 1.0111993648355604

k=7: 0.9765280137579607

k=8: 0.9226015406797958

k=9: 0.848424959966636

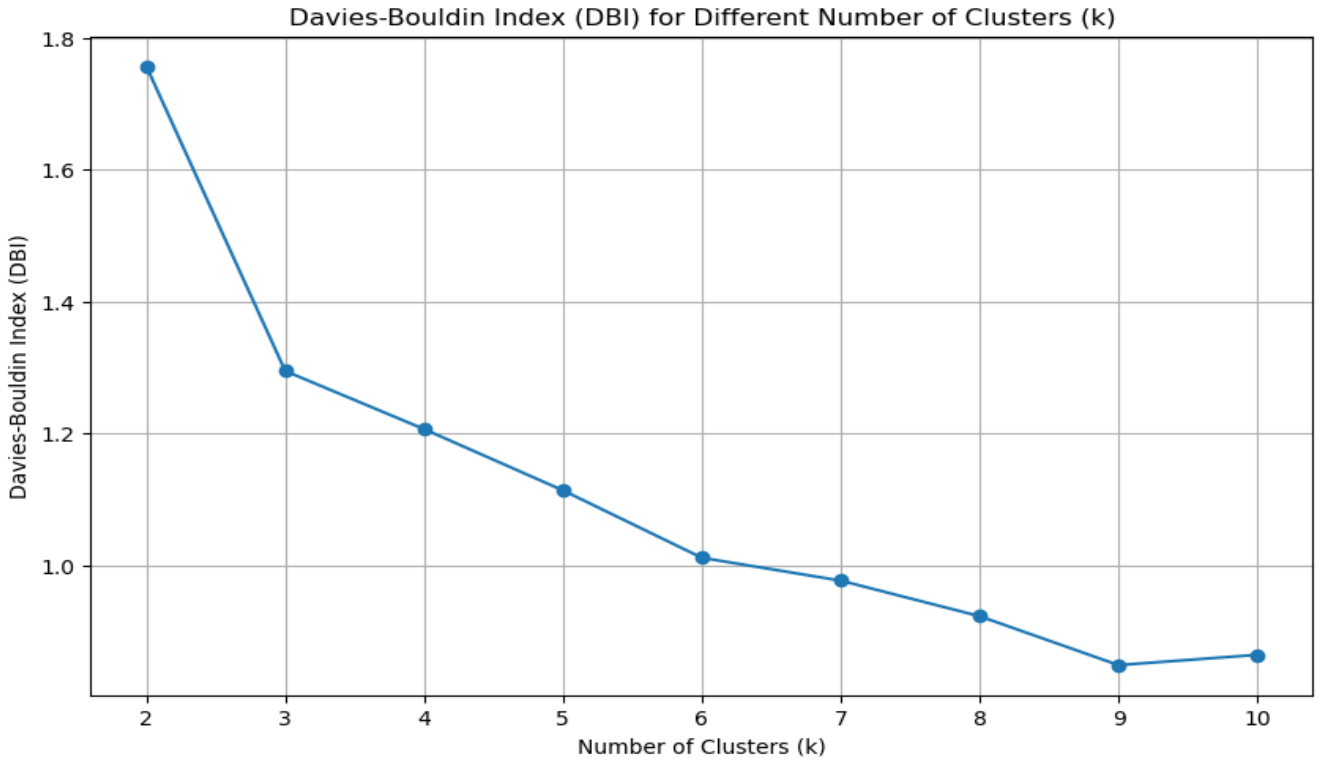
k=10: 0.8640054821990638

The optimal number of clusters was determined to be **k = 9**, as it had the lowest DBI score (0.85). This value was used for the final clustering model.

Results and Visualization

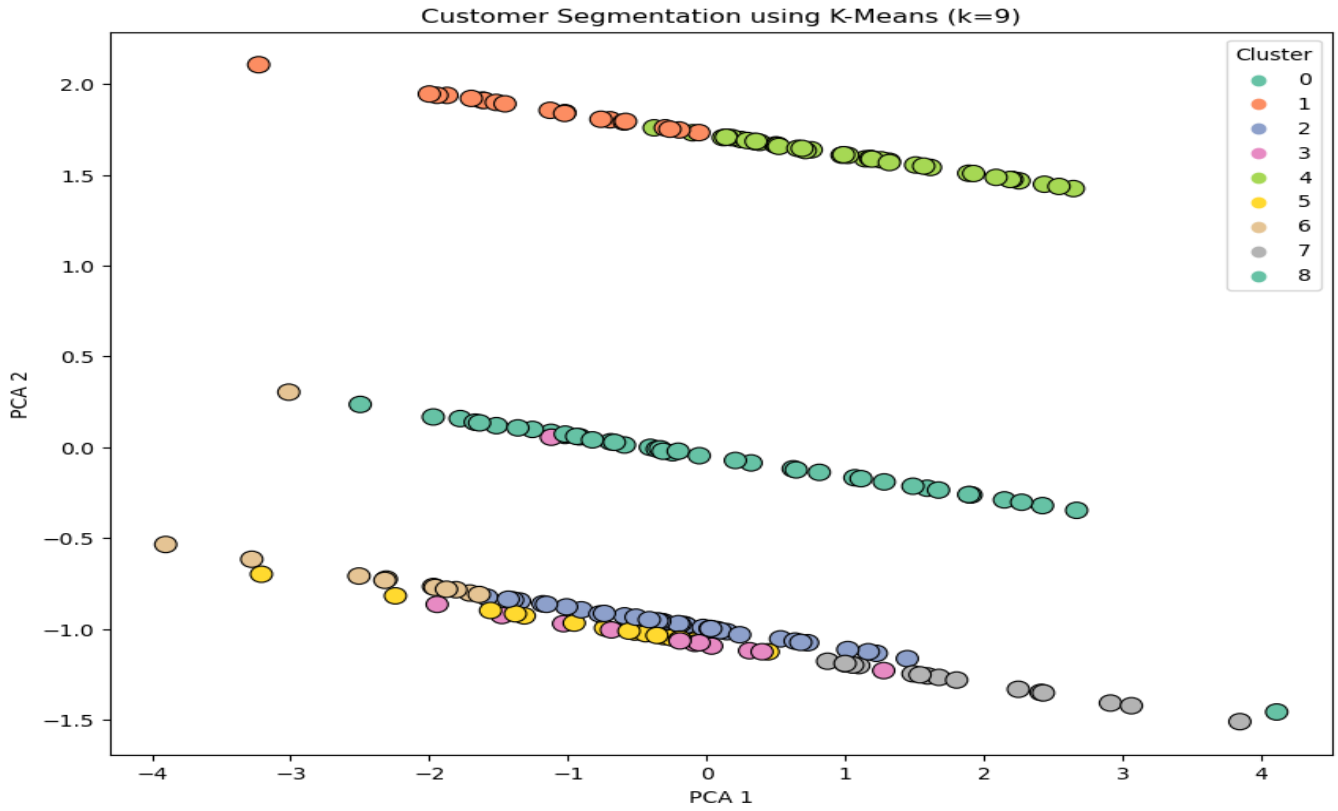
The final clustering model was trained with **k = 9**, assigning customers to one of the nine clusters.

1. **DBI vs. Number of Clusters**:
   A plot of DBI scores against the number of clusters was created to evaluate clustering quality. The lowest DBI score was observed at **k = 9**, confirming it as the optimal number of clusters for the analysis.



2. PCA Visualization:

The high-dimensional data was reduced to two dimensions using PCA, and the clusters were visualized in a 2D scatterplot. Each point represents a customer, and the points were color-coded based on their cluster assignment. This visualization demonstrated clear separation between clusters, highlighting the distinctiveness of customer segments.



Conclusion

The K-Means clustering analysis successfully grouped the customer base into **9 distinct clusters** using transactional and demographic data. The optimal number of clusters was determined using the Davies-Bouldin Index, with PCA used to visualize and interpret the results.

Key Outcomes:

- Identification of customer groups with similar spending behaviours and patterns.
- A framework for targeted marketing and improved customer relationship management.
- Insights to focus on high-value customers for retention and growth.

Future Improvements:

1. Testing other clustering methods like DBSCAN or hierarchical clustering.
2. Adding features such as customer lifetime value or purchase frequency trends.
3. Using additional metrics, like the silhouette score, to further assess clustering quality.