Clustering Results Report

Overview

This report presents the results of customer segmentation performed using clustering techniques. The dataset used combines profile information from Customers.csv and transaction information from Transactions.csv. The clustering process included data preprocessing, feature engineering, clustering model evaluation, and visualization of clusters. The primary objective was to identify distinct customer groups to enable targeted marketing and improve customer understanding.

Key Results

1. **Number of Clusters Formed**

The optimal number of clusters was determined to be **4**, based on the lowest Davies-Bouldin Index value.

2. **Davies-Bouldin Index Value**

The Davies-Bouldin Index for the chosen number of clusters (**4**) was **0.352726**, indicating well-defined clusters with low intra-cluster variance and high inter-cluster separation.

3. **Other Relevant Metrics**

Number of Clusters	Davies-Bouldin Index	Silhouette Score
2	0.936131	0.410051
3	0.802377	0.598974
4	0.352726	0.773412
5	0.407570	0.739113
6	0.530661	0.676724
7	0.605089	0.584516
8	0.692452	0.504069
9	0.682267	0.501010
10	0.689968	0.483305

The **Silhouette Score** for the 4 clusters was **0.773412**, further confirming the quality of clustering, with well-separated and compact clusters.

Methodology

1. **Data Preprocessing**

- Aggregated transaction data (Transactions.csv) to calculate:
  - **Purchase Frequency**: Count of transactions per customer.
  - **Total Spend**: Sum of transaction values per customer.
  - **Most Purchased Product**: Mode of products purchased per customer.
- Merged transaction data with customer profile data (Customers.csv) using CustomerID.
- Applied feature engineering:
  - One-hot encoding for categorical variables (e.g., Region).
  - Min-Max normalization for numerical features (Purchase Frequency, Total Spend).

2. **Clustering Algorithm**

- Used **K-Means** clustering with k ranging from 2 to 10.
- Evaluated clustering performance using:
  - **Davies-Bouldin Index**: Measures cluster compactness and separation.
  - **Silhouette Score**: Measures cohesion and separation of clusters.

Conclusion

The optimal number of clusters for this dataset is **4**, with the lowest Davies-Bouldin Index value of **0.352726**. These clusters are well-separated and compact, as supported by the Silhouette Score of **0.773412**. This segmentation provides actionable insights into customer behavior and can be leveraged for targeted marketing and personalized services.

The identified clusters reflect distinct customer groups based on their purchase frequency, total spend, and regional distribution. These clusters can be further analyzed to derive specific behavioral patterns, such as high-value customers, frequent buyers, or customers from particular regions.