Customer Segmentation Report: K-Means Clustering

1. Introduction

This report presents the findings of customer segmentation using K-Means clustering based on transactional data. The goal was to group customers into distinct segments based on their purchasing behavior, which includes total spend, purchase count, average purchase value, diversity of categories purchased, and recency of transactions. These clusters can aid in understanding customer groups for targeted marketing strategies.

2. Data Preparation

We used three datasets for this analysis:

- **Customers.csv**: Contains customer information.
- **Products.csv**: Contains product details, including categories.
- **Transactions.csv**: Contains transaction details, including ProductID, CustomerID, TransactionID, and the transaction's total value.

The following steps were carried out during data preprocessing:

- **Data Merging**: The Transactions dataset was merged with the Products dataset to obtain product category information for each transaction.

- **Customer Profile Creation**: A profile was generated for each customer by aggregating the transaction data, resulting in the following features:
  - total_spend: Total money spent by the customer.
  - purchase_count: Total number of transactions by the customer.
  - avg_purchase_value: Average transaction value.
  - diversity_of_categories: Number of unique product categories purchased.
  - recency_days: Days since the most recent transaction.

3. Preprocessing and Transformation

- **Categorical Data Encoding**: The Region column was encoded using OneHotEncoder to convert categorical data into numerical form.
- **Feature Scaling**: All the customer profile features were standardized using StandardScaler to ensure that each feature contributed equally to the clustering process.

4. Clustering Approach

K-Means clustering was applied to the preprocessed data to group customers into clusters. To determine the optimal number of clusters, we evaluated the following metrics:

- **Davies-Bouldin Index (DB Index)**: This index measures how well-separated the clusters are. A lower value indicates better clustering.
- **Silhouette Score**: This score measures how similar a sample is to its own cluster compared to other clusters. A higher silhouette score indicates well-separated and distinct clusters.

5. Optimal Number of Clusters

We evaluated K-Means clustering for cluster sizes ranging from 2 to 10. The results for both DB Index and Silhouette Score were plotted to visually inspect which cluster size yields the best results.

- **Optimal Number of Clusters**: The analysis revealed that the optimal number of clusters is **7**, as it resulted in the lowest DB Index and a reasonable Silhouette Score.

6. Clustering Metrics

- **Davies-Bouldin Index (DB Index)**: The final DB Index value was **1.0642**, indicating relatively good separation between clusters.
- **Silhouette Score**: The silhouette scores for the clusters ranged from 0.3 to 0.6, which also suggested well-separated clusters.

7. Final Cluster Visualization

The data was reduced to two dimensions using **Principal Component Analysis (PCA)** for easier visualization. A scatter plot was generated, where each customer was plotted based on the two principal components, and the clusters were color-coded.

8. Conclusion

The customer segmentation process successfully identified 7 distinct customer clusters, each representing unique purchasing behaviors. The DB Index and Silhouette Score confirmed that the clustering is of reasonable quality, with well-separated clusters.
These customer segments can be further analyzed for targeted marketing strategies, product recommendations, and personalized customer experiences.

Key Findings

- **Optimal Number of Clusters**: 7
- **Final Davies-Bouldin Index (DB Index)**: 1.0642