Introduction

Customer segmentation is essential for understanding customer behavior, optimizing marketing strategies, and improving business decision-making. In this analysis, clustering was performed using KMeans on a dataset combining customer profiles (`Customers.csv`) and transactional data (`Transactions.csv`). The analysis aimed to identify distinct customer groups and evaluate the quality of clustering.

Clustering Process

1. **Data Preparation**:
   - Transactional features, such as total spent, total transactions, average spent per transaction, and unique products purchased, were derived from the `Transactions.csv` dataset.
   - These features were merged with customer demographic information from `Customers.csv`.
   - Categorical variables (e.g., `Region`) were one-hot encoded, and numerical features were standardized for uniform scaling.
2. **Clustering Algorithm**:
   - The KMeans algorithm was used due to its simplicity and effectiveness in segmenting numeric data.
   - Clustering was performed for a range of clusters (`k = 2 to 10`) to determine the optimal number of clusters.
3. **Evaluation Metrics**:
   - **Davies-Bouldin Index (DB Index)**: Measures clustering compactness and separation, where a lower score indicates better clustering.
   - **Silhouette Score**: Evaluates how well samples are assigned to their clusters, with higher scores indicating better-defined clusters.

Optimal Number of Clusters

- Based on the Davies-Bouldin Index and Silhouette Score plots, the optimal number of clusters was determined to be **4**.

Final Clustering Results

- **Number of Clusters Formed**: 4
- **Final Davies-Bouldin Index**: **0.9137** (indicating good compactness and separation).
- **Final Silhouette Score**: **0.3461** (indicating moderately well-defined clusters).

Cluster Profiles

Each cluster represents a distinct customer group based on transactional and demographic behavior. Below are high-level observations about the clusters:

1. **Cluster 0**: Customers with high spending and frequent transactions, representing high-value customers.
2. **Cluster 1**: Customers with average spending and moderate engagement levels.
3. **Cluster 2**: Low-spending customers who have fewer transactions and engagement.
4. **Cluster 3**: Customers with unique purchasing patterns, spending moderately but frequently buying diverse products.

Visualization

The clusters were visualized using a 2D scatter plot, where:

- **X-axis**: Standardized total spending.
- **Y-axis**: Standardized average spending per transaction.
- **Colors**: Represent different clusters.

This plot provides an intuitive understanding of how customer groups differ in spending behavior.

Conclusion

The clustering analysis successfully segmented customers into **4 distinct groups**, with notable differences in spending patterns and transactional behavior. The Davies-Bouldin Index (0.9137) suggests a good clustering structure, while the Silhouette Score (0.3461) indicates moderately defined boundaries.

These insights can aid businesses in targeted marketing, personalized promotions, and resource optimization.

Recommendations

1. **Marketing Strategies**:
   - Focus marketing efforts on high-value customers (Cluster 0) with loyalty programs and personalized offers.
   - Identify opportunities to increase engagement with low-spending customers (Cluster 2).
2. **Product Recommendations**:
   - Tailor product recommendations for customers in Cluster 3 based on their diverse purchasing behavior.
3. **Future Work**:

- Consider advanced clustering methods (e.g., DBSCAN or Gaussian Mixture Models) for improved boundary definition.
- Incorporate temporal data (e.g., seasonality) to further refine segmentation.