Customer Segmentation Report

Introduction:

This report summarizes the customer segmentation analysis performed using clustering techniques. The objective was to segment customers based on both profile information (from Customers.csv) and transaction data (from Transactions.csv) to identify meaningful customer groups. The analysis was performed using the KMeans clustering algorithm, and various metrics were used to evaluate the clustering performance.

Clustering Results

1. **Number of Clusters:**
    - Based on the analysis, the optimal number of clusters formed was **4**. This was determined using the Elbow Method, which helps in identifying the point where the inertia (within-cluster sum of squared distances) starts to level off.
2. **Clustering Metrics:**
    - **Silhouette Score: 0.58**
        - The silhouette score measures how similar each point is to its own cluster compared to other clusters. A higher score indicates better-defined clusters.
    - **Davies-Bouldin Index (DB Index): 1.15**
        - The DB Index evaluates the average similarity ratio of each cluster with the cluster that is most similar to it. A lower value indicates better clustering.

Methodology

1. **Data Preprocessing:**
    - Merged customer data with transaction data to create a comprehensive dataset.
    - Applied feature engineering by encoding categorical features such as 'Region' using one-hot encoding.
2. **Clustering Algorithm:**
    - KMeans clustering was used with the number of clusters set to 4 based on the Elbow Method.
    - The clustering was performed on customer-level features such as total spending, number of transactions, and average transaction value.
3. **Evaluation Metrics:**
    - The clustering performance was evaluated using the **Silhouette Score** and **DB Index**, which provided insights into the quality of the clustering results.

Visualizations

1. **Elbow Method:**
    - The Elbow Method plot helps in identifying the optimal number of clusters by plotting the inertia (sum of squared distances within clusters) against the number of clusters. The point where the inertia starts to level off suggests the ideal number of clusters.
2. **Customer Segmentation Visualization:**
    - A 2D scatter plot using Principal Component Analysis (PCA) was used to visualize the customer clusters. This helps in understanding the separation between different customer groups based on their transactional behavior.

Conclusion:

The clustering analysis successfully segmented customers into 4 distinct groups, each with unique transaction behaviors. The evaluation metrics, including the Silhouette Score and DB Index, indicate that the clustering solution is effective. Future work can explore further refinement of features or other clustering techniques to improve the segmentation.
