Customer Segmentation and Clustering

Introduction

The goal of this task is to segment customers using clustering techniques based on both their profile information (from **Customers.csv**) and transaction history (from **Transactions.csv**). The clustering will help identify groups of customers with similar behaviors and characteristics. This can provide insights into different customer segments for personalized marketing, product recommendations, or targeted offers.

Data Description

The dataset used for this task consists of two files:

1. **Customers.csv**: Contains customer profile data such as CustomerID, Name, Region, and SignupDate.
2. **Transactions.csv**: Contains transaction data, including CustomerID, ProductID, Quantity, and TotalValue.

These datasets will be merged and preprocessed to create features that will be used for clustering.

Clustering Approach

- We applied **K-Means clustering** to segment customers into distinct groups.
- The number of clusters was chosen based on the **Davies-Bouldin Index** and **Silhouette Score**.
- The features used for clustering were:
    - **Profile Features**: Region (converted to numerical), SignupDate (converted to the number of days since signup).
    - **Transaction Features**: TotalTransactionValue (total amount spent by the customer), TotalQuantity (total quantity of products bought).

Clustering Results

- **Number of Clusters**: 4
  We decided to segment the customers into 4 clusters, as this number yielded the most meaningful separation based on our evaluation metrics.
- **Davies-Bouldin Index**: 1.17
  The Davies-Bouldin Index is a metric used to evaluate the separation of clusters. A lower value indicates better clustering, and 1.17 suggests that while the clusters are somewhat well-separated, there is room for improvement.
- **Silhouette Score**: 0.34
  The Silhouette Score helps to measure how well-separated the clusters are. A score closer to 1 indicates better clustering. With a score of 0.34, the clustering is somewhat effective, though there is overlap between some clusters.

Visualizations

We visualized the clusters using:

- **Scatter plots**: Representing customer segmentation on two main axes (e.g., TotalValue vs. TotalQuantity).
- **Cluster Centers**: Showing the central point of each cluster.

Code Implementation

The code for this task is implemented in a Jupyter Notebook, which includes:

1. Data preprocessing and merging.
2. Feature extraction and scaling.
3. Clustering with K-Means.
4. Evaluation metrics calculation (Davies-Bouldin Index, Silhouette Score).
5. Visualizations of the clusters.

Conclusion and Next Steps

The clustering results provide valuable insights into customer segmentation. However, further refinements could be made:

- **Optimizing the number of clusters**: Trying values between 2 and 10 to find the most optimal segmentation.
- **Feature Engineering**: Introducing additional features (e.g., customer activity, frequency of transactions) might improve the clustering.
- **Trying alternative algorithms**: Exploring other clustering methods like DBSCAN or hierarchical clustering could yield different segmentation patterns.