Customer Segmentation / Clustering Report

Introduction:

Customer segmentation is essential for understanding customer behavior and targeting them with personalized offers. In this task, we used both customer profile data (from `Customers.csv`) and transaction data (from `Transactions.csv`) to segment customers. This segmentation helps in identifying distinct customer groups based on spending and transaction frequency.

Data Overview:

- **Customer Data (`Customers.csv`)**:
  - CustomerID, CustomerName, Region, SignupDate.
- **Transaction Data (`Transactions.csv`)**:
  - TransactionID, CustomerID, ProductID, TotalValue, Price, Quantity, TransactionDate.

The goal was to segment customers using transaction behavior (spending and frequency) and customer profile (region, signup date).

Methodology:

1. **Data Preprocessing:**
   - Merged customer and transaction data on `CustomerID`.
   - Aggregated data to compute `TotalSpending` and `TransactionFrequency` for each customer.
2. **Feature Scaling:**
   - Standardized features using **StandardScaler** to ensure equal importance across features.
3. **Clustering:**
   - Applied **KMeans clustering** with `n_clusters=4`, chosen based on the **Elbow Method**.
4. **Evaluation Metrics:**
   - **Davies-Bouldin Index (DB Index):** Measures clustering quality (lower value = better).
   - **Silhouette Score:** Indicates how well-separated the clusters are (higher value = better).
5. **Dimensionality Reduction:**
   - Used **PCA** to reduce the features to 2D for easy visualization.

Results:

- **Number of clusters formed:** 4
- **Davies-Bouldin Index:** 1.22 (indicating reasonably well-separated clusters)
- **Silhouette Score:** 0.51 (indicating reasonable separation between clusters)

Cluster Analysis:

- **Cluster 1 (Low Spending, Low Frequency):** Customers who spend less and transact infrequently.
- **Cluster 2 (High Spending, High Frequency):** High spenders who transact frequently.
- **Cluster 3 (Moderate Spending, Moderate Frequency):** Customers with moderate spending and frequency.
- **Cluster 4 (Low Spending, High Frequency):** Customers with frequent low-value transactions.

These clusters reveal different customer behaviors, useful for targeted marketing and personalized strategies.

Visualization:

PCA was used to reduce the features to 2D, visualizing the clusters where each point represents a customer and colors indicate the cluster.

Conclusion:

The KMeans clustering algorithm successfully segmented customers into 4 clusters, with reasonable separation as indicated by the Davies-Bouldin Index and Silhouette Score. The insights from these clusters can inform marketing strategies, loyalty programs, and product offerings.

Next Steps:

- Integrate additional customer profile data for deeper insights.
- Experiment with other clustering algorithms (e.g., DBSCAN, Agglomerative Clustering) for improved results.