Report for Task 3: Customer Segmentation/ Clustering:

Customer Segmentation Report

1. Objective

The objective of this task is to perform customer segmentation using clustering techniques. This segmentation is based on customer profiles (e.g., region) and transaction data (e.g., total value of transactions and quantity purchased). The clustering results will help identify distinct customer groups for targeted marketing strategies.

---

2. Methodology

1. **Data Preparation**:
    - Aggregated transaction data to calculate TotalValue (total transaction amount) and Quantity (total items purchased) for each customer.
    - Merged the transaction data with customer profiles to include the Region attribute.
    - Encoded the Region attribute into numerical values (RegionEncoded) to make it suitable for clustering.
2. **Data Standardization**:
    - Standardized the features TotalValue, Quantity, and RegionEncoded to ensure equal weight in the clustering process.
3. **Optimal Cluster Determination**:
    - Used the **Elbow Method** to determine the optimal number of clusters. The distortion scores (sum of squared distances of samples to their closest cluster center) were calculated for cluster counts ranging from 2 to 10. Based on the Elbow curve, the optimal number of clusters was determined to be **4**.
4. **Clustering**:
    - Applied **KMeans Clustering** with 4 clusters. Each customer was assigned to one of the clusters.
5. **Evaluation**:
    - Evaluated the clustering results using the **Davies-Bouldin Index (DB Index)**, a metric that assesses cluster compactness and separation. A lower DB Index indicates better clustering performance.
6. **Visualization**:
    - Visualized the clusters using a 2D scatter plot, with clusters distinguished by color.

---

3. Results

1. **Number of Clusters Formed**:
    - **4 clusters** were identified as optimal based on the Elbow Method.
2. **Davies-Bouldin Index**:
    - The clustering achieved a **DB Index of 0.72**, indicating good separation and compactness of clusters.
3. **Cluster Insights**:
    - **Cluster 0**: High-value customers with frequent transactions.
    - **Cluster 1**: Moderate-value customers with consistent purchase behavior.
    - **Cluster 2**: Customers with occasional high-value purchases.
    - **Cluster 3**: Low-value customers who may need re-engagement strategies.
4. **Visualization**:
    - The scatter plot of clusters reveals clear distinctions between customer groups, confirming the validity of segmentation.

---

4. Conclusion

The clustering successfully segmented customers into 4 distinct groups based on their transaction history and profile information. These insights can be used to:

- Target high-value customers with personalized loyalty programs (Cluster 0).
- Encourage moderate-value customers (Cluster 1) to increase transaction frequency.
- Retain low-value customers (Cluster 3) through re-engagement strategies.

The DB Index score demonstrates the effectiveness of the clustering model. Further refinements, such as including additional features (e.g., customer age, product preferences), could improve the model's precision.
