Customer Segmentation / Clustering Report

1. Introduction:

In this task, we applied customer segmentation using clustering techniques to group customers based on their profile information and transaction behavior. We used the **K-Means clustering algorithm** to segment the customers into distinct groups, with clustering evaluated using the **Davies-Bouldin Index** (DB Index).

2. Methodology:

- **Data Source**: We used two datasets: `Customers.csv` and `Transactions.csv`. The customer dataset contains demographic information such as region and signup date, while the transaction dataset contains transactional data like total spending (`TotalValue`) and quantity purchased.
- **Clustering Algorithm**: We applied **K-Means clustering** due to its simplicity and effectiveness for customer segmentation tasks.
- **Features Used**:
  - **Customer Profile Features**: Region (encoded as one-hot), Signup Date (converted to days since signup).
  - **Transaction Features**: Total spending (`TotalValue`), Quantity purchased, Transaction frequency.
- **Feature Scaling**: The features were standardized using **StandardScaler** to ensure equal contribution from each feature.
- **Clustering Evaluation**: The **Davies-Bouldin Index (DB Index)** was calculated to assess the clustering quality. A lower DB Index indicates better-defined and well-separated clusters.

3. Results:

- **Number of Clusters**: We performed clustering with **4 clusters** based on the K-Means algorithm.
- **DB Index**: The DB Index value was calculated to evaluate the clustering performance. A lower value indicates better clustering.
- **Visualization**: A scatter plot was generated to visualize the clusters, where the x-axis represents `TotalValue`, and the y-axis represents `Quantity`. Each customer is assigned a color corresponding to their cluster.

4. Clustering Metrics:

- **Davies-Bouldin Index**: The DB Index value is a measure of the cluster separation and compactness. The lower the DB Index, the better the clustering.
- **Other Metrics**:
  - **Silhouette Score**: An additional metric could be calculated to assess how well-separated the clusters are. It quantifies how similar a point is to its own cluster compared to other clusters.

5. Visual Representation:

The clusters were visualized using a scatter plot, where each customer is represented by a point on the plot. The points are color-coded based on their assigned cluster label.