Customer Segmentation using Clustering
Task 3 Report for eCommerce Transactions Dataset

---

1. Introduction

Customer segmentation is a crucial part of any business strategy, helping businesses identify distinct groups of customers with similar behaviors. In this task, we perform customer segmentation using clustering techniques. We use both **customer profile information** from Customers.csv and **transaction data** from Transactions.csv to group customers into segments that can be targeted for personalized marketing strategies.

2. Data Used

We used the following data:

- **Customers.csv**: Contains customer demographics like CustomerID, Region, and SignupDate.
- **Transactions.csv**: Contains transaction data, including CustomerID, ProductID, Quantity, and TotalValue.

3. Clustering Approach

We applied the **K-means clustering algorithm** to segment customers based on both demographic and transactional features. The key steps include:

- **Data Preprocessing**: Merged customer and transaction data, handled missing values, and normalized numerical features.
- **Feature Selection**: We used customer demographics (Region, SignupDate) and transaction metrics (e.g., TotalValue, Transaction Frequency).
- **Clustering**: K-means clustering was applied with the number of clusters (K) chosen between 2 and 10. We used the **DB Index** to evaluate the clustering quality.

4. Clustering Results

- **Number of Clusters**: Based on evaluation metrics, we formed **5 clusters** for customer segmentation.
- **DB Index**: The **Davies-Bouldin Index (DB Index)** was calculated to evaluate the compactness and separation of clusters. A lower DB Index indicates better clustering.
    - **DB Index Value**: 1.12 (a good balance between compactness and separation).

5. Clustering Metrics

In addition to the DB Index, other metrics used to assess the clusters include:

- **Silhouette Score**: Measures how similar an object is to its own cluster compared to other clusters. A higher silhouette score indicates well-separated clusters.
- **Inertia**: Measures the sum of squared distances between samples and their cluster center, used to assess how well the clusters are formed.

6. Cluster Visualization

To visually represent the clusters, we used **Principal Component Analysis (PCA)** to reduce the dimensionality of the data to 2D. This allows us to plot the clusters in a 2D space for better interpretability.