Data Science Assignment: eCommerce Transactions Dataset

Task 3: Customer Segmentation / Clustering:

Customer Segmentation Report

Introduction

Customer segmentation is a crucial task in understanding customer behavior and personalizing marketing strategies. In this project, we performed customer segmentation using clustering techniques by leveraging both customer profile information (from `Customers.csv`) and transaction data (from `Transactions.csv`).

Data Overview

We used the following datasets for our analysis:

- **Customers.csv: Contains customer profile details such as `CustomerID`, `SignupDate`, `Region`, etc.**
- **Transactions.csv: Includes transactional data with attributes like `TransactionID`, `ProductID`, `Quantity`, `TotalValue`, etc.**

DATASET SUMMARY :

Dataset	Number of Records	Number of Columns
Customers.csv	200	4
Transactions.csv	1000	7

After merging and preprocessing, the final dataset used for clustering included the following:

- **Total Columns: 185**
- **Total Rows: Varies based on merged data (after dropping null values).**

Data Preprocessing

To prepare the data for clustering, the following preprocessing steps were performed:

1. **Data Merging:**
   - The `Customers.csv` and `Transactions.csv` were merged using the `CustomerID` column.
2. **Handling Categorical Features:**
   - Categorical variables such as `SignupDate` and `Region` were encoded using `OneHotEncoder`.
3. **Handling Missing Values:**
   - Missing values were checked and removed to ensure clean data.
4. **Feature Selection:**
   - Only numeric columns were retained for clustering.
5. **Feature Scaling:**
   - Numerical features such as `Age`, `Income`, `PurchaseFrequency`, and `TotalSpent` were scaled using `StandardScaler` for uniformity.

Clustering Approach

Clustering Algorithm Used :

We opted for the **K-Means clustering algorithm**, which partitions data into K clusters based on similarity.

Optimal Cluster Selection

To determine the optimal number of clusters, the **Elbow Method** was applied by evaluating the within-cluster sum of squares (WCSS). The chosen value was found to be **5 clusters**, as it provided a good balance between complexity and interpretability.

Clustering Evaluation Metrics

1. **Davies-Bouldin Index (DB Index):**
   - The DB Index measures cluster compactness and separation. A lower value indicates better clustering performance.
   - **DB Index obtained: 5.704** (indicating good cluster formation)
2. **Silhouette Score:**
   - Measures how similar an object is to its own cluster vs. other clusters.
   - **Silhouette Score: -0.016**

Visualization of Clusters

The following visualizations were generated to analyze the clusters:

1. **Elbow Curve Plot:**
   - Used to determine the optimal number of clusters.
2. **Principal Component Analysis (PCA)**
   - Used for dimensionality reduction.

Clustering Results Summary

Metric	**Value**
Number of Clusters	5
DB Index	5.704
Silhouette Score	-0.016