Data Science Assignment Task - 3

Objective:
The goal of this analysis is to segment customers based on their transactional behavior, using features such as total spend, transaction count, and average transaction value, in combination with customer demographic details like age and income. The analysis includes data preprocessing, clustering, and evaluation of clustering performance.

---

1. Data Preprocessing

To begin, the data was preprocessed by combining two datasets: Customers.csv and Transactions.csv. The following steps were taken:

- **Aggregating Transaction Data**: Transaction data was aggregated by CustomerID to compute key metrics for each customer:
    - TotalSpend: The total money spent by each customer.
    - TransactionCount: The number of transactions per customer.
    - AvgTransactionValue: The average value of each transaction.
- **Merging with Customer Profiles**: Customer details, such as age and income, were merged with the aggregated transaction data.
- **Handling Missing Data**: Any missing values in the dataset were filled with zero to ensure consistency during the clustering process.

---

2. Feature Selection and Scaling

The following features were selected for clustering:

- **Age**
- **Income**
- **TotalSpend**
- **TransactionCount**
- **AvgTransactionValue**

These features were then standardized using StandardScaler to normalize them, ensuring all variables are on the same scale, as K-Means clustering is sensitive to feature scaling.

---

3. Clustering

The next step involved clustering the customers using the **K-Means** algorithm. Before applying the clustering, the optimal number of clusters was determined using the **Elbow Method**. The following steps were taken:

- **Elbow Method**: A plot was generated to visualize the inertia (sum of squared distances from each point to its assigned cluster centroid) for different values of k. The elbow point suggested that k=4 was the optimal number of clusters.
- **K-Means Clustering**: With k=4, the K-Means algorithm was applied, resulting in four distinct customer clusters.

---

4. Evaluation of Clustering Performance

The clustering performance was evaluated using the following metrics:

- **Davies-Bouldin Index**: This metric measures the average similarity ratio of each cluster with its most similar cluster. A lower Davies-Bouldin Index indicates better clustering. The obtained score was **0.92**, indicating a relatively good clustering solution.
- **Silhouette Score**: The silhouette score quantifies how well each point lies within its cluster. A score close to +1 indicates that the points are well-clustered. The obtained score was **0.58**, which suggests that the clusters are reasonably distinct but could benefit from further refinement.

---

5. Cluster Visualization

To visualize the clusters, Principal Component Analysis (PCA) was applied to reduce the data to two dimensions. A scatter plot was then created to visualize how the clusters are distributed in the reduced 2D space. Each cluster is represented by a distinct color, making it easier to identify the groupings.