Clustering Logic and Metrics for Customer Segmentation

In this assignment, we perform customer segmentation using K-Means clustering based on customer transaction data. The steps for clustering are outlined below:

1. Data Preprocessing:

- The data is loaded from three CSV files: `Customers.csv`, `Transactions.csv`, and `Products.csv`.
- Missing values in all datasets are dropped using `dropna()`.
- Three key features are derived from the transaction data:
  - Total Spend per Customer: Sum of `TotalValue` for each customer.
  - Transaction Frequency: Number of transactions for each customer.
  - Average Transaction Value: Mean of `TotalValue` for each customer.
- These features are merged with the customer data to create a `customer_features` DataFrame.

2. Feature Engineering:

- A log transformation (np.log1p) is applied to the `TotalSpend` to normalize its distribution.
- Missing values in the derived features (`TotalSpend`, `TransactionFrequency`, and AvgTransactionValue) are filled with the column mean.

3. Feature Scaling:

- The three derived features are standardized using StandardScaler to ensure they are on the same scale, preventing any single feature from dominating the clustering process.

4. Clustering with K-Means:

- Choosing the Optimal Number of Clusters (k): The Silhouette Score is used to evaluate clustering quality for k values between 2 and 10. A higher Silhouette Score indicates better-defined clusters.
- Fitting K-Means: The K-Means algorithm is applied with the best k, and each customer is assigned to one of the clusters.

5. Clustering Evaluation Metrics:

- Silhouette Score: Measures the cohesion and separation of clusters. A higher score indicates well-separated clusters. In the code, a Silhouette Score of 0.378 suggests moderate clustering quality.

- Davies-Bouldin Index (DB Index): Measures cluster separation, with a lower value indicating better separation. The DB Index of 1.01 indicates moderate cluster distinctness.

6. Visualization:

- PCA (Principal Component Analysis)
  is used for dimensionality reduction to visualize the clusters in 2D. The clusters are visualized in a scatter plot with the first two principal components.

Conclusion:

The clustering process identifies customer segments based on their transaction behavior, with moderate separation and cohesion, as indicated by the Silhouette Score and DB Index.

Visual Representation of Clusters.