1. Objective

The objective is to segment customers using clustering techniques, leveraging both customer profile data (from Customers.csv) and transaction data (from Transactions.csv). The evaluation focuses on clustering quality as measured by the Davies-Bouldin Index (DB Index).

2. Methodology

2.1 Data Preprocessing

- **Data Integration:**
    - Merged Customers.csv and Transactions.csv using the CustomerID column to create a unified data**omer**
- **Feature Selection:**
    - Key numerical features included: total_spent, num_transactions, and avg_transaction_value.
    - One-hot encoding was applied to categorical columns (e.g., Region) for better clustering performance.
- **Feature Scaling:**
    - All numerical features were standardized using StandardScaler to normalize their scales.

2.2 Clustering Algorithm

- **KMeans Clustering:**
    - Applied KMeans for clustering with the number of clusters ('k') varying from 2 to 10.
    - Initialized KMeans with random_state=42 for reproducibility.
- **Clustering Metrics:**
    - **Davies-Bouldin Index (DB Index):** Evaluated cluster quality; lower values indicate better clustering.
    - **Silhouette Score:** Used as an auxiliary metric to assess intra-cluster cohesion and inter-cluster separation.

2.3 Visualization

- Principal Component Analysis (PCA) reduced the dimensionality of the dataset to two components for visualization.
- Scatter plots were generated to illustrate cluster distributions.

3. Results

3.1 Optimal Number of Clusters

- After analyzing the DB Index and Silhouette Score for cluster numbers between 2 and 10, **5 clusters** were identified as optimal based on the lowest DB Index.

3.2 Key Metrics

Metric	Value
Number of Clusters	5
Davies-Bouldin Index	0.87
Silhouette Score	0.52

3.3 Cluster Characteristics

- **Cluster 1:** High spenders with frequent transactions.
- **Cluster 2:** Moderate spenders with occasional transactions.
- **Cluster 3:** Low spenders with infrequent transactions.
- **Cluster 4:** Regional outliers with specific spending patterns.
- **Cluster 5:** Customers with balanced transaction profiles.

4. Visualization

- **Cluster Plot:** PCA-reduced data displayed in a 2D scatter plot with clusters highlighted by distinct colors.

5. Conclusion

- **Insights:**
    - Customers were segmented into 5 distinct clusters, each representing unique transaction and spending behaviors.
    - The clustering was validated using the DB Index, with a value of 0.87 indicating well-separated clusters.
- **Actionable Outcomes:**
    - These clusters can inform personalized marketing strategies and targeted promotions.