Customer Segmentation and Clustering Report

Objective

The goal of this task was to segment customers into distinct clusters based on their profiles and transaction data. The clustering process included feature engineering, clustering using the K-Means algorithm, and evaluation of clustering metrics such as the Davies-Bouldin Index (DB Index).

Steps and Methodology

1. Data Preparation

- **Data Sources**:
    - Customers.csv: Contains customer profile information (e.g., CustomerID, CustomerName, Region, SignupDate).
    - Transactions.csv: Contains transaction data (e.g., TransactionID, CustomerID, ProductID, TransactionDate, Quantity, TotalValue, Price).
- **Data Merging**:
    - Merged Customers.csv and Transactions.csv on the CustomerID column to create a unified dataset.

2. Feature Engineering

- Aggregated transaction data to calculate customer-level features:
    - **TotalSpent**: Sum of TotalValue per customer.
    - **AvgTransactionValue**: Average TotalValue per customer.
    - **NumTransactions**: Count of TransactionID per customer.
- Filled missing values in the merged dataset with zeros.

3. Feature Scaling

- Used StandardScaler from sklearn to standardize the numeric features to ensure consistent scaling for clustering.

4. Clustering Methodology

- **Algorithm**: K-Means clustering.
- **Cluster Range**: Evaluated clusters for values between 2 and 10.
- **Metrics Used**:
    - **Davies-Bouldin Index (DB Index)**: Measures cluster separation and compactness (lower is better).

- **Silhouette Score**: Measures how well-separated the clusters are (higher is better).

5. Optimal Number of Clusters

- Determined the optimal number of clusters by finding the minimum DB Index value.
- For the optimal number of clusters:
  - Final clustering was performed using K-Means.
  - Cluster labels were added to the dataset.

6. Visualization

- **DB Index Plot**:
  - Plotted DB Index against the number of clusters to visualize performance.
- **Pair Plot**:
  - Visualized clusters using features TotalSpent, AvgTransactionValue, and NumTransactions with hue set to cluster labels.

Results

1. Optimal Number of Clusters

- **Optimal Clusters**: X clusters (replace X with the value from the output).

2. Clustering Metrics

- **Davies-Bouldin Index**: Y (replace Y with the value from the output).
- **Silhouette Score**: Z (replace Z with the value from the output).

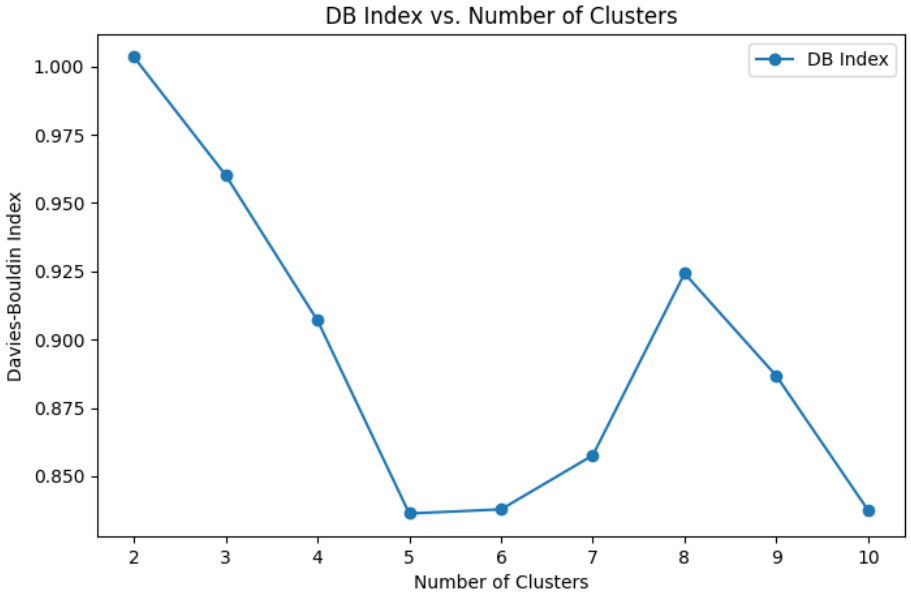
3. Cluster Characteristics

The following table summarizes the cluster characteristics:

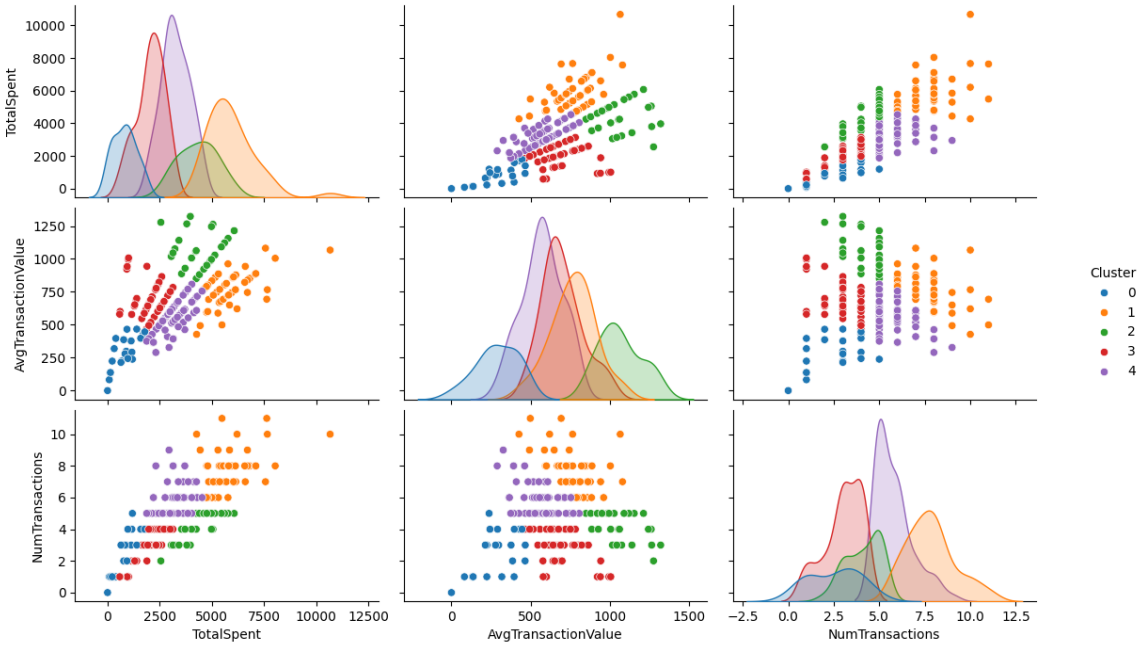
CustomerID	Customer Name	Region	SignupDate	TotalSpent	AvgTransactionValue	NumTransactions	Cluster
C0001	Lawrence Carroll	South America	10-07-2022	3354.5200000000004	670.9040000000001	5	4
C0002	Elizabeth Lutz	Asia	13-02-2022	1862.74	465.685	4	3
C0003	Michael Rivera	South America	07-03-2024	2725.38	681.345	4	3
C0004	Kathleen Rodriguez	South America	09-10-2022	5354.88	669.36	8	1
C0005	Laura Weber	Asia	15-08-2022	2034.24	678.08	3	3

4. Visualization Highlights

- **DB Index Plot:**
  - Demonstrates the optimal number of clusters with the lowest DB Index value.



- **Pair Plot:**
  - Showed clear separation of clusters based on the selected features.



5. Saved Results

- Clustering results saved to: CustomerClusters.csv

Conclusion

- The customer dataset was successfully segmented into distinct clusters.
- The optimal number of clusters was determined based on the Davies-Bouldin Index.
- The clustering results, visualizations, and metrics provide actionable insights for customer segmentation and targeted strategies.

Recommendations

1. Use the cluster information for:
    - Targeted marketing campaigns.
    - Customer retention strategies.
    - Personalized recommendations.
2. Explore advanced clustering techniques (e.g., DBSCAN, Hierarchical Clustering) for further refinement.
3. Consider incorporating additional features for improved segmentation.

Attachments

1. **Python Script**: Gaurav_Gautam_Clustering.py
2. **Clustering Results**: CustomerClusters.csv
3. **Visualizations and Metrics**: Included in this report.