Cluster_Summary.csv
https://drive.google.com/file/d/1vYyPBwCW5z-jEmMTtOaHBCco7mFrxy4X/view?usp=sharing

Clustered_Customers.csv
https://drive.google.com/file/d/1Y49WpAuLV4MggJm1-bpBLihcrs6osFRx/view?usp=sharing

Clustering_metrics.csv
https://drive.google.com/file/d/12doSV2vdFNjlgzSGTLzr8SVZmQsV8Cdb/view?usp=sharing

```
Davies-Bouldin Index: 0.807
Silhouette Score: 0.361
```

Data Preparation Report

Step 1: Data Preparation

1.1 Load the Data

Two datasets, Customers.csv and Transactions.csv, were loaded into pandas dataframes for analysis.

- **Customers Data**
  - Contains columns: CustomerID, CustomerName, Region, SignupDate.
  - Sample data shows various customers from regions like South America and Asia.
- **Transactions Data**
  - Contains columns: TransactionID, CustomerID, ProductID, TransactionDate, Quantity, TotalValue, Price.
  - Sample data shows transaction records for customers, including product IDs, quantities, and values.

1.2 Merge Customer and Transaction Data

- The datasets were merged using the CustomerID column to provide comprehensive customer transaction data.
- The resulting merged_data contains the transaction details along with customer information, such as Name, Region, and Signup Date.

1.3 Handle Missing Values

- No missing values were found in the merged dataset after inspection.

Step 2: Feature Engineering

2.1 Aggregate Transactional Data

Aggregated transaction data for each customer was calculated, including:

- Total Purchase Value (sum of TotalValue).
- Average Purchase Value (mean of TotalValue).

- Frequency of Purchases (count of transactions).
- Number of Transactions (count of TransactionDate).

2.2 Add Time-Based Features

- Extracted the day of the week and month from TransactionDate.
- Aggregated these features to find the most frequent transaction day and month for each customer.
- Combined these time features with the previously aggregated transactional data.

2.3 Normalize or Scale Features

- Numerical features (TotalValue, AveragePurchase, Frequency, TransactionCount) were normalized using MinMaxScaler to scale them to a range between 0 and 1.

2.4 Select Features for Clustering

- Categorical feature (TransactionDay) was one-hot encoded.
- Final features selected for clustering included:
    - Numerical features: TotalValue, AveragePurchase, Frequency, TransactionCount, TransactionMonth.
    - One-hot encoded TransactionDay.

Step 3: Clustering

3.1 Choose Clustering Algorithm

- KMeans and DBSCAN were selected as potential clustering algorithms.

3.2 Determine the Optimal Number of Clusters

- **Elbow Method** and **Silhouette Scores** were plotted to determine the optimal number of clusters.
    - Elbow method suggested that 4 clusters might be optimal based on inertia.

- Silhouette scores also indicated 4 as a reasonable choice for optimal clustering.

3.3 Implement KMeans Clustering

- KMeans with n_clusters=4 was applied to the selected features.
- The resulting clusters were added to the customer_features dataframe.

3.4 Implement DBSCAN Clustering

- DBSCAN was applied with eps=0.5 and min_samples=5, resulting in a different set of clusters.

Cluster Summary

- KMeans clustering grouped customers into 4 clusters based on purchasing behavior, frequency, and transaction patterns.
- DBSCAN clustering provided a density-based clustering that could potentially identify outliers and dense customer groups.

Step 4: Exploratory Data Analysis (EDA)

1. **Distribution of Key Features**
   - Plots were generated to visualize the distribution of key numerical features like TotalValue, AveragePurchase, and Frequency.
   - Histograms and box plots showed the spread of values, helping to identify outliers and trends in customer spending patterns.
2. **Correlation Analysis**
   - A heatmap of correlations between numerical features was created to understand relationships.
   - Features like Frequency and TotalValue showed a strong positive correlation, indicating that more frequent purchases are linked to higher spending.
3. **Customer Segmentation Visualization**
   - The customer segmentation results from K-Means clustering were visualized using scatter plots and pair plots.

    - Different clusters exhibited distinct patterns in terms of spending, frequency, and product preferences.

Step 5: Analysis of Clusters

1. **Cluster Summary**
    - **Cluster 1**: High spending, frequent purchases, and premium products.
    - **Cluster 2**: Moderate spending and occasional purchases.
    - **Cluster 3**: Low spending, less frequent purchases, and budget products.
    - **Cluster 4**: Sporadic transactions, high-value but low-frequency purchases.
2. **Cluster Profiling**
    - Each cluster was profiled based on key features like TotalValue, Frequency, and product preferences. This provides insights into customer behavior for targeted marketing.
3. **Customer Lifetime Value (CLV) Analysis**
    - The CLV of each cluster was estimated to determine which segments are most profitable in the long term.
    - Cluster 1 showed the highest CLV, while Cluster 3 exhibited the lowest CLV.

Step 6: Insights for Strategy Development

1. **Targeted Marketing Campaigns**
    - **Cluster 1**: High-value customers; campaigns should focus on loyalty programs and premium offerings.
    - **Cluster 2**: Moderate value; offer personalized discounts to encourage higher frequency.
    - **Cluster 3**: Low-value customers; offer bundle deals and more affordable products to increase purchase volume.
    - **Cluster 4**: Engage with targeted promotional offers to drive repeat purchases.
2. **Product Recommendations**
    - Based on the clustering, product recommendations were personalized for each segment.

- High-value clusters (Cluster 1) received recommendations for premium products, while low-value clusters (Cluster 3) were offered budget-friendly alternatives.

3. **Customer Retention Strategies**
    - For high-value customers, retention strategies like exclusive membership benefits were suggested.
    - For low-value customers, strategies focused on increasing purchase frequency and product variety were recommended.