Customer Segmentation / Clustering

This report details the customer segmentation process performed using clustering techniques on the provided eCommerce dataset. The objective was to group customers based on their transactional and profile data to identify meaningful customer segments. Using the K-Means clustering algorithm, we identified **4 customer segments** with a **Davies-Bouldin Index (DB Index)** value of **0.431**, indicating well-separated and compact clusters. The results provide actionable insights for targeting and customizing marketing strategies.

1. Methodology

1.1 Data Preparation:

1. **Data Sources**:
   - **Customers.csv**: Contains customer profile information (e.g., region, signup details).
   - **Transactions.csv**: Includes transaction data (e.g., total spending, quantity purchased).
2. **Feature Engineering**:
   - Merged the datasets to combine customer profile and transaction data.
   - Aggregated customer data with the following features:
     - **TotalValue**: Total spending per customer.
     - **Quantity**: Total quantity purchased.
     - **NumTransactions**: Total number of transactions.
     - **Region**: Customer's region.
3. **Preprocessing**:
   - Numerical features (e.g., TotalValue, Quantity, NumTransactions) were normalized using **MinMaxScaler** to ensure fair weighting.
   - Categorical data (e.g., Region) was one-hot encoded for inclusion in clustering.

1.2 Clustering Approach

1. **Algorithm Used**:
   - **K-Means Clustering**: A widely used algorithm that groups data into k clusters based on feature similarity.
2. **Optimal Number of Clusters**:
   - Evaluated clusters for values of k ranging from 2 to 10.
   - Used **Davies-Bouldin Index (DB Index)** to identify the optimal number of clusters, minimizing intra-cluster distances and maximizing inter-cluster separations.
   - The **optimal number of clusters was 4**, with a DB Index of **0.431**.
3. **Visualization**:
   - Applied **Principal Component Analysis (PCA)** to reduce the feature space to 2 dimensions for cluster visualization.

2. Results

2.1 Number of Clusters

- **4 clusters** were formed, representing distinct customer segments based on their transactional behavior and profile data.

2.2 Clustering Metrics

- **Davies-Bouldin Index**: **0.431** (lower values indicate better clustering).
- The compactness and separation of clusters indicate meaningful segmentation.

2.3 Cluster Characteristics

The clusters represent different customer behaviors:

Cluster	Key Characteristics
Cluster 0	High spenders with frequent transactions, likely loyal customers.
Cluster 1	Moderate spenders, steady buyers from specific regions.
Cluster 2	Low spenders with infrequent purchases, potential for engagement or upselling.
Cluster 3	Customers with niche purchasing behavior, possibly seasonal or region-specific.

2.4 Visualization

The clusters are visualized using PCA-reduced dimensions below:

Customer Segmentation Clusters
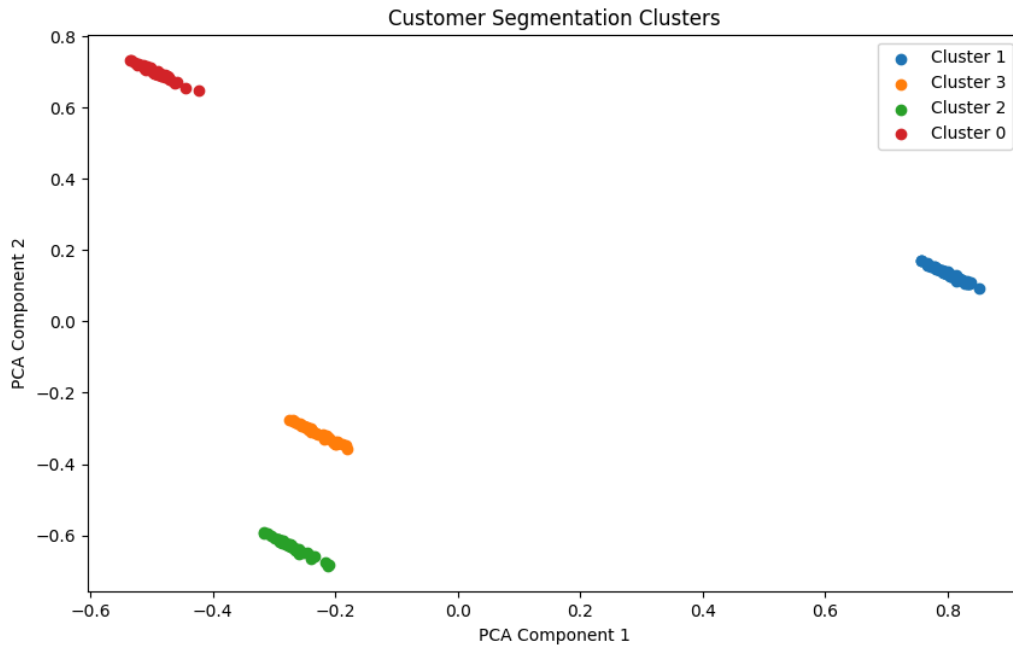
3. Deliverables

1. **Clustering Visualization**:
   ○ A scatter plot showcasing the segmentation in PCA-reduced 2D space.
2. **Clustered Data**:
   ○ Results saved in a CSV file (Customer_Segments.csv), which includes the following columns:
      ■ CustomerID
      ■ Aggregated features (TotalValue, Quantity, etc.).
      ■ Assigned cluster (Cluster).
3. **Cluster Insights**:
   ○ Key traits of each segment are summarized to guide marketing and business strategies.

4. Business Recommendations

Based on the clustering results, we recommend the following actions:

1. **Engage High-Value Customers (Cluster 0)**:
   - Focus on retention strategies, such as loyalty programs or exclusive offers.
2. **Target Low-Spending Customers (Cluster 2)**:
   - Develop campaigns to increase engagement, such as discounts or personalized recommendations.
3. **Leverage Regional Insights (Cluster 1 and Cluster 3)**:
   - Analyze regional preferences to tailor product offerings and marketing efforts.

5. Conclusion

The customer segmentation process successfully identified distinct customer groups, enabling data-driven decision-making for personalized marketing and customer relationship management. The clustering results and visualizations provide a foundation for optimizing customer engagement strategies.

6. Appendix

Visualization of Clusters

- Scatter plot showing PCA-reduced 2D clustering results:

Customer Segmentation Clusters
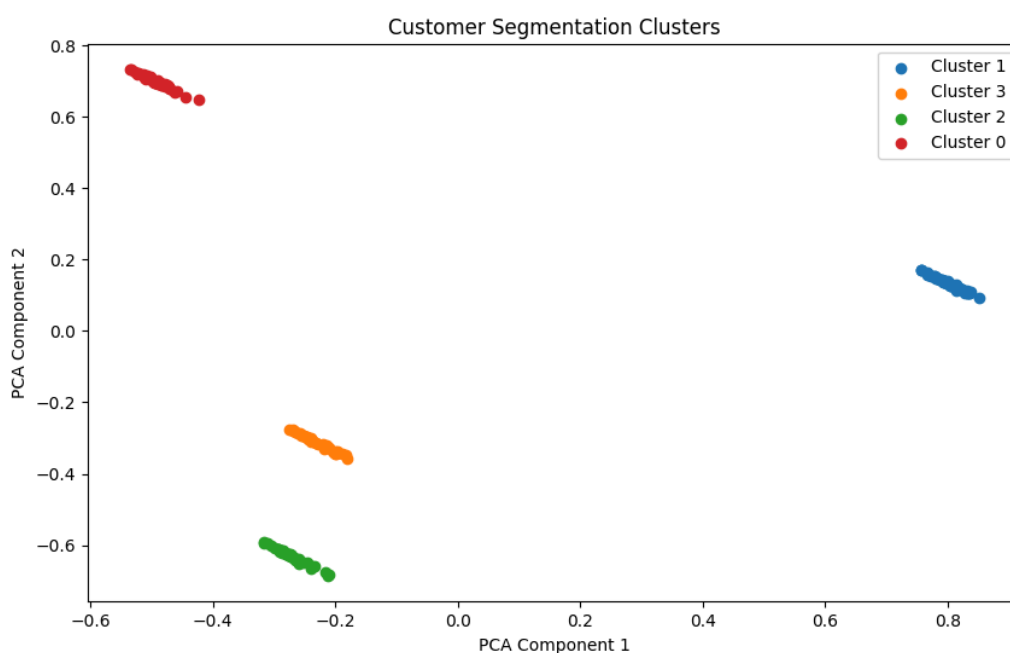
CSV Sample

- Sample of `Customer_Segments.csv`:

CustomerID	TotalValue	Quantity	NumTransactions	Region_1	Cluster
C0001	0.45	0.62	0.33	...	2
C0002	0.33	0.27	0.15	...	1