Task 3: Customer Segmentation / Clustering Report

Objective:

To perform customer segmentation using clustering techniques based on profile information (from Customers.csv) and transaction data (from Transactions.csv). The goal is to identify meaningful customer groups that can be targeted with specific business strategies. The evaluation metric is the Davies-Bouldin Index (DB Index).

Dataset Overview:

1. **Customers.csv**:

    o Contains customer demographic information, including CustomerID, CustomerName, Region, and SignupDate.

2. **Transactions.csv**:

    o Contains transaction details, including TransactionID, CustomerID, ProductID, TransactionDate, Quantity, TotalValue, and Price.

Methodology:

1. **Data Preprocessing:**
    o Converted SignupDate and TransactionDate to datetime format and handled invalid dates using pd.to_datetime().
    o Derived key customer metrics from transaction data, such as:
        ▪ **Transaction Count:** Total number of transactions for each customer.
        ▪ **Total Spend:** Total monetary value of transactions for each customer.
        ▪ **Average Purchase Value:** Average transaction value.
        ▪ **Total Quantity:** Total quantity purchased by each customer.
    o Calculated **Days Since Signup** using the reference date (pd.Timestamp.now()).
    o Applied one-hot encoding for the Region feature.
2. **Feature Selection:**
    o Selected key features for clustering:
        ▪ days_since_signup
        ▪ transaction_count
        ▪ total_spend
        ▪ avg_purchase_value
        ▪ total_quantity
3. **Data Scaling:**
    o Used StandardScaler to normalize all numerical features for clustering.

4. **Clustering Algorithm:**
   - o Evaluated different numbers of clusters (2 to 10) using the K-Means algorithm.
   - o Calculated the Davies-Bouldin Index (DB Index) for each configuration.
   - o Selected the optimal number of clusters corresponding to the minimum DB Index.
5. **Cluster Visualization:**
   - o Plotted the DB Index scores against the number of clusters.
   - o Visualized clusters based on Total Spend and Transaction Count.
6. **Cluster Profiling:**
   - o Analyzed each cluster's characteristics by aggregating key features.

Results:

1. **Optimal Number of Clusters:**
   - o The optimal number of clusters determined was **7**, based on the minimum DB Index value.
2. **Davies-Bouldin Index:**
   - o The lowest DB Index value achieved was **1.0265**, indicating good clustering quality.
3. **Cluster Profiles:**

Cluster	Days Since Signup	Transaction Count	Total Spend	Avg. Purchase Value	Total Quantity
0	860.84	8.53	6348.17	751.36	23.42
1	546.45	2.24	756.93	347.67	4.10
2	791.12	5.12	4743.41	955.19	14.54
3	292.93	5.37	3101.28	588.79	12.42
4	882.50	4.43	2550.37	599.61	10.35
5	253.00	2.45	2287.42	930.89	6.90
6	227.42	7.08	5508.56	800.36	19.58

Visualizations:

1. **DB Index vs Number of Clusters:**

    o   A line plot shows how the DB Index varies with the number of clusters. The minimum point indicates the optimal number of clusters.

Davies-Bouldin Index vs Number of Clusters
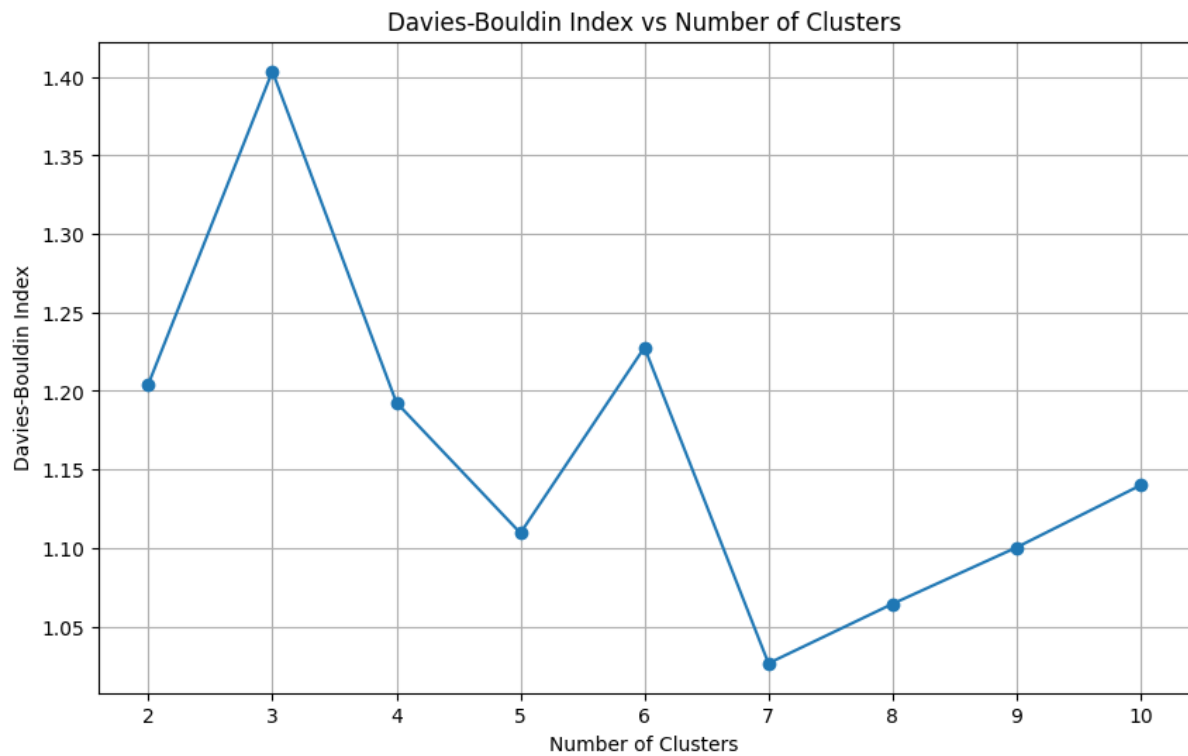
2. **Cluster Distribution:**

   o A scatter plot visualizes customer clusters based on Total Spend and Transaction Count, with color-coded clusters.

Customer Segments: Total Spend vs Transaction Count

Insights:

1. Customers in **Cluster 0** have the highest transaction count and spend, suggesting they are loyal and high-value customers.
2. **Cluster 1** represents low-spend, low-transaction customers, who may benefit from engagement campaigns.
3. Customers in **Cluster 2** have high average purchase values, indicating a preference for premium products.
4. **Cluster 5** includes relatively new customers with high average purchase values but low transaction counts, presenting opportunities for targeted promotions to increase loyalty.

Conclusion:

The clustering analysis successfully segmented customers into seven distinct groups. These clusters can inform targeted marketing strategies, such as rewarding high-value customers, re-engaging low-spend users, and promoting premium products to relevant segments. The Davies-Bouldin Index of 1.0265 confirms the clustering quality.