Results

1. **Number of Clusters**:
    - The analysis was performed with **5 clusters**.
    - Each cluster represents customers with distinct transactional and profile patterns.
2. **Davies-Bouldin Index**:
    - **DB Index**: **1.1575385765743047**
3. **Cluster Distribution**:
    - **Cluster Sizes**:
        - Cluster 0: 70 customers.
        - Cluster 1: 38 customers.
        - Cluster 2: 75 customers.
        - Cluster 3: 16 customers.
        - Cluster 4: 32 customers.
    - Cluster sizes vary, reflecting diversity in customer behaviors.
4. **Key Metrics per Cluster**:
    - The clustering process grouped customers based on the following aggregated metrics:
        - Total spent by customers (total_spent).
        - Average transaction value (avg_transaction_value).
        - Total number of transactions (total_transactions).

Methodology

1. **Data Preparation**:
    - **Datasets Used**: Customers.csv and Transactions.csv.
    - The customer data includes profile attributes (Region, Signup Date), and the transaction data provides details about purchase behavior.
    - Aggregated transaction data into metrics such as:
        - Total number of transactions.
        - Total spending.
        - Average transaction value.
    - Merged customer and transaction data to prepare features for clustering.

2. **Feature Selection and Scaling**:
    - Selected features:
        - Region: Encoded as numeric using LabelEncoder.
        - total_transactions: Total number of purchases.
        - total_spent: Total value of all purchases.
        - avg_transaction_value: Average value per transaction.
    - Standardized the selected features using StandardScaler to ensure equal contribution of each feature to clustering.
3. **Clustering Algorithm**:
    - **K-Means** clustering was applied to segment the customers.
    - The number of clusters was set to **5**, chosen arbitrarily but tested for evaluation metrics.

Visualizations

1. **Bubble Plot**:
    - Visualized clusters based on total_spent and avg_transaction_value.
    - Bubble size represents the relative number of customers in each cluster.
2. **PCA-Based Decision Boundaries**:
    - Applied Principal Component Analysis (PCA) to reduce feature dimensions for visualization.
    - Plotted decision boundaries for the K-Means clusters in a 2D space.
3. **3D Plot**:
    - Visualized clusters in a 3D space using total_spent, avg_transaction_value, and total_transactions.