Clustering Analysis Report

Overview

This report outlines the clustering analysis performed on customer transaction data, aiming to segment customers into distinct and actionable groups. The analysis involved data preparation, feature engineering, applying both KMeans and DBSCAN clustering methods, and visualizing and evaluating the results to gain insights for personalized business strategies.

Steps Performed

1. **Data Preparation**
    - **Datasets Merged**:
        - Two datasets, Customers.csv and Transactions.csv, were merged on the CustomerID column.
    - **Datetime Conversion**:
        - Columns such as TransactionDate and SignupDate were converted into datetime formats to enable time-based feature calculations.
    - **Feature Engineering**:
        - Key customer-level metrics were aggregated:
            - **TotalValueSpent**: The total transaction value for each customer.
            - **TotalQuantityPurchased**: Total items purchased by the customer.
            - **NumTransactions**: The number of transactions completed.
            - **AveragePrice**: Average price of items purchased.
            - **AvgGapBetweenTransactions**: The average time gap between consecutive transactions.
            - **TimeSinceSignup**: The total number of days since the customer signed up.
            - **Region**: A categorical feature representing the geographic region of the customer.
2. **Data Preprocessing**
    - **Encoding**:
        - The categorical feature Region was one-hot encoded to create numerical columns for clustering.
    - **Scaling**:
        - Numerical features were standardized using the StandardScaler to ensure all features contributed equally to the clustering process.

3. **Clustering Methods**

    - **KMeans Clustering**:
        - KMeans was applied for different values of k ranging from 2 to 10.
        - The **Davies-Bouldin Score** was used to determine the optimal number of clusters.
        - Final clustering was performed using the optimal k, and cluster labels were assigned to each customer.
    - **DBSCAN Clustering**:
        - DBSCAN (Density-Based Spatial Clustering of Applications with Noise) was applied to identify clusters based on density.
        - Key parameters:
            - **Epsilon (eps)**: The maximum distance between two points to be considered in the same neighbourhood is 3.
            - **Minimum Samples**: The minimum number of points required to form a cluster are 5.
        - Outliers were identified and marked as noise points.
    - Comparison:
        - KMeans identified compact and well-separated clusters.
        - DBSCAN provided flexibility in identifying non-spherical clusters and outliers.
4. **Visualization**

    - **t-SNE Visualization**:
        - High-dimensional feature space was reduced to 2 dimensions using t-SNE for intuitive visualization.
        - Scatter plots were generated for both KMeans and DBSCAN to illustrate customer groupings, with clusters represented by distinct colors.

Results

The combination of KMeans and DBSCAN provided robust customer segmentation:

- **KNN**:

  K value: The best value of K is 5.

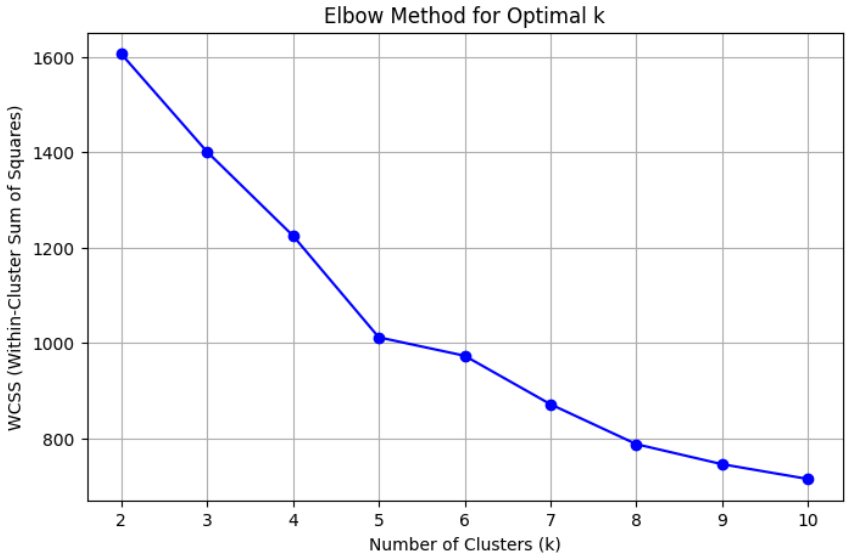
  DB Index: 1.2417022871412584

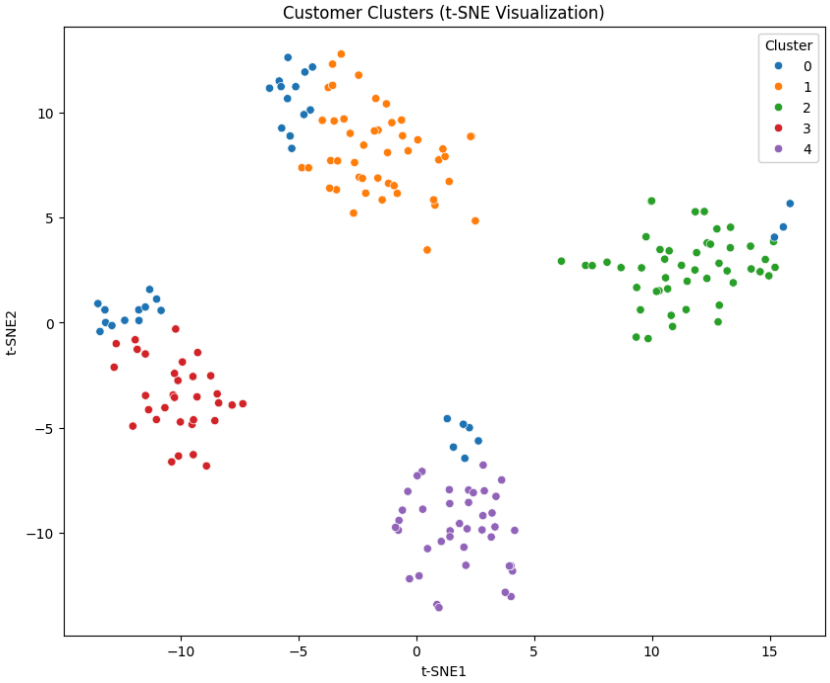
- **DBSCAN**:

  K value: found the number of clusters as 4.

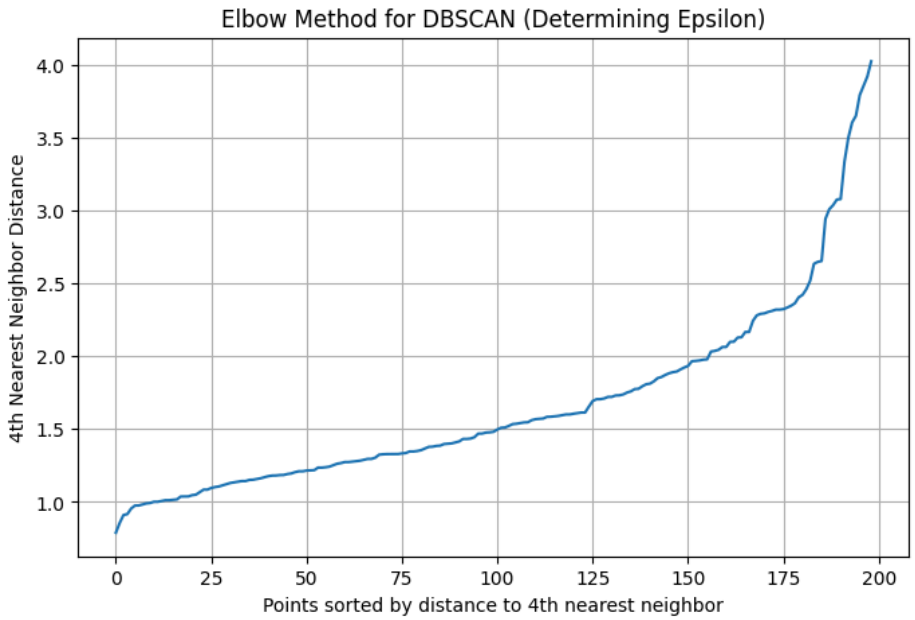
DB Index: 1.1474339017394706

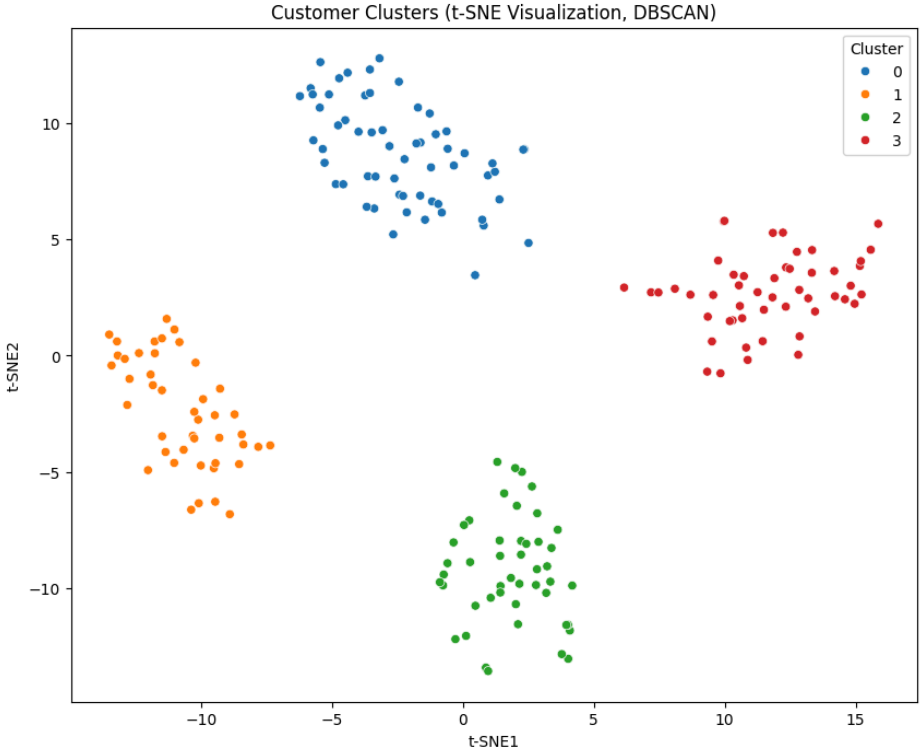
KNN





DBSCAN


Elbow Method for DBSCAN (Determining Epsilon)
4th Nearest Neighbor Distance
Points sorted by distance to 4th nearest neighbor
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0
25
50
75
100
125
150
175
200


Customer Clusters (t-SNE Visualization, DBSCAN)
Cluster
0
1
2
3
t-SNE2
t-SNE1
10
5
0
−5
−10
−10
−5
0
5
10
15