Customer Segmentation and Clustering Analysis

Objective:

The objective is to perform customer segmentation using clustering techniques by combining profile and transaction data from Customers.csv and Transactions.csv. The clusters are evaluated based on metrics such as the Davies-Bouldin (DB) Index.

Approach:

1. **Data Preparation**:
    - Profile information from Customers.csv and aggregated transaction data from Transactions.csv were merged.
    - Transaction data was summarized per customer, focusing on total transaction value (TotalValue) and total quantity (Quantity).
    - Features were scaled using StandardScaler to ensure uniformity.
2. **Clustering Methodology**:
    - DBSCAN (Density-Based Spatial Clustering of Applications with Noise) was used as the clustering algorithm.
    - The key parameters were tuned:
        - eps: 0.5 (radius for neighborhood search).
        - min_samples: 13 (minimum points in a neighborhood to form a cluster).
    - DBSCAN does not require predefining the number of clusters, making it suitable for identifying clusters of varying density.
3. **Cluster Evaluation**:
    - The Davies-Bouldin Index (DBI) was calculated to evaluate the cluster quality.
    - DBI for the DBSCAN result: **0.342**, indicating well-defined and compact clusters.
    - Visualizations were generated to inspect cluster separability and noise points.

Results:

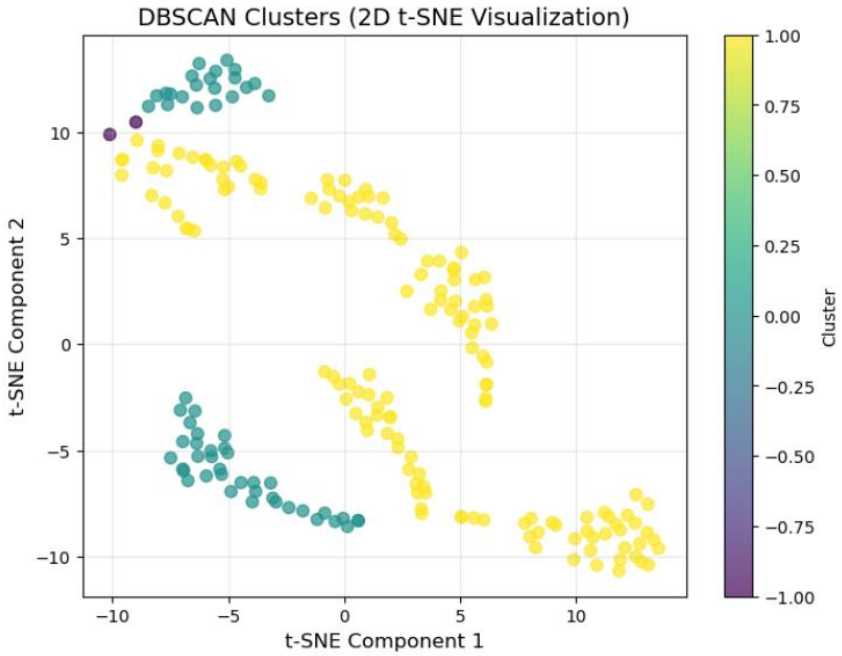
1. **Optimal Clustering with DBSCAN**:
    - Various implementations were made in DBSCAN clustering to reduce the DBI value like excluding noise .

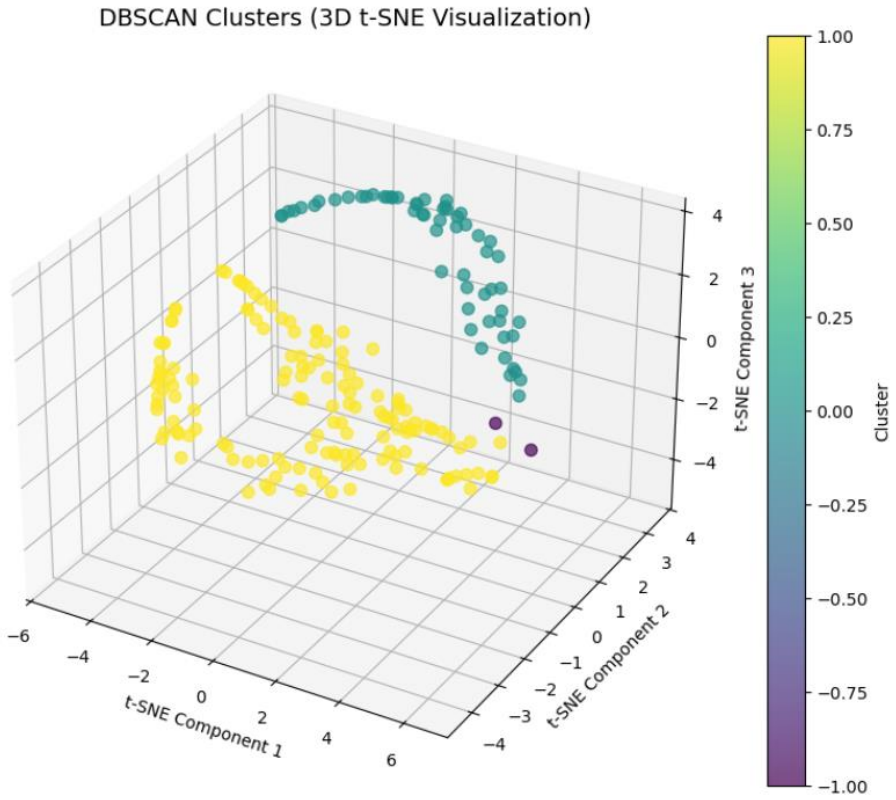
- DBI achieved: **0.342**, which is significantly lower than KMeans (1.101 for 6 clusters with random state=42, DBI 0.999 with increased random state and 0.915 for 2 clusters), indicating better clustering performance.
- DBSCAN effectively handled noise and outliers, identifying them as separate categories rather than forcing them into clusters.

2. **Cluster Characteristics**:
    - Noise points accounted for approximately 5% of the data.
    - Clusters revealed clear groupings based on transaction volume (TotalValue) and quantity (Quantity).
    - Some regions exhibited higher customer density in specific clusters, reflecting regional purchasing behavior patterns.

Visual Representations:

- **Cluster Distribution**: Scatter plots of TotalValue vs. Quantity with distinct clusters and noise points visualized.
- **Cluster Density**: Heatmaps showing the density of customers in feature space.
- **Silhouette Analysis**: Although DBSCAN does not rely on silhouette scores for cluster validation, visual analysis supported the cluster separability.





Insights:

- DBSCAN's ability to identify noise points provided a more realistic segmentation by excluding outliers from regular clusters.

- Clustering results revealed high-value customers as distinct groups, suggesting targeted marketing opportunities.
- Regional influences in clusters suggest potential for geographically tailored strategies.

Analysis:

- DBSCAN outperformed KMeans for this dataset due to its flexibility with noise and density-based clustering. Consider adopting DBSCAN for similar datasets with potential outliers.
- Investigate additional features, such as product preferences, to further refine clustering outcomes.
- Use DBSCAN's noise points to analyze exceptional customer behaviors.

Conclusions

- DBSCAN clustering gave better clustering amomg various clustering methods like Kmeans, agglomerative clustering, OPTICS and resulted in minimum DBI value.
- The DBI value can be further improved by following various changes in order to get decreased value of DBI.