Zeotap – Data Scientist Intern Clustering assignment

Submission By:- Ansh Phutela (IIT Patna)

Contact:- anshphutela.2012@gmail.com

My approach:

1. **Data Preprocessing**: The data was cleaned and transformed, including handling missing values, creating new features (e.g., customer lifetime and recency), and encoding categorical variables. The data was scaled using StandardScaler for better clustering performance.

2. **Clustering Techniques**: K-Means and DBSCAN were used for customer segmentation. K-Means was initialized with 4 clusters, while DBSCAN parameters (eps and min_samples) were optimized to identify meaningful clusters and outliers.

3. **Evaluation Metrics**: The Davies-Bouldin Index and silhouette score were calculated for both clustering methods.

4. **Visualization**: PCA and t-SNE were employed to reduce dimensionality and visualize clusters in a 2D space, providing a clear view of customer groupings for both K-Means and DBSCAN.

5. **Results**: The final clustered dataset was saved for further analysis, with DBSCAN highlighting noise and outliers, while K-Means provided more structured groupings.

Results and Scope of Improvements:

1. Best Davies-Bouldin Index was: 0.9257393236490291

2. Best Silhouette Score was: 0.29302331815607724

Main Reasons for not so good results could be:

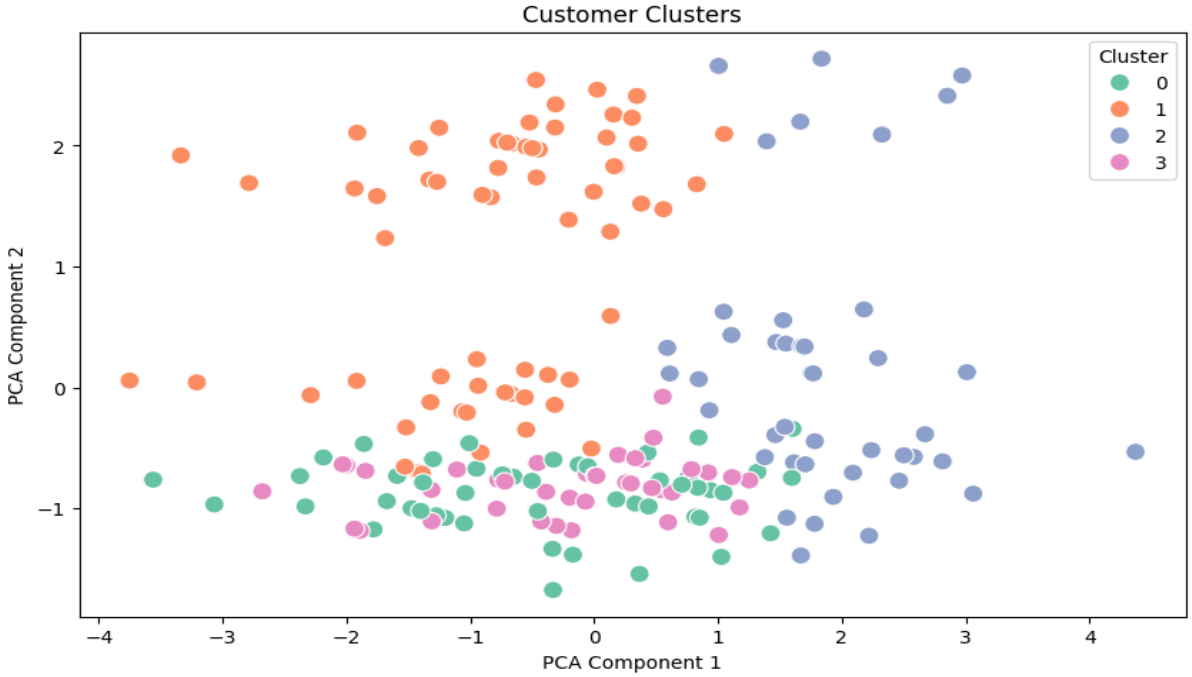
1. **Overlapping Clusters**: The features used for clustering, such as transactional data and customer demographics, might not have well-separated boundaries. This causes clusters to overlap, reducing the clarity and compactness of the clusters, which negatively impacts both DB Index and Silhouette Score.

2. **High Dimensionality**: Even after scaling, the dataset's dimensionality can create challenges for clustering algorithms. While t-SNE helped with visualization, the curse of dimensionality can reduce the ability to form distinct, well-separated clusters.

3. **Feature Engineering Limitations**: The engineered features (e.g., recency, total spent) may not fully capture the true underlying customer behaviors or relationships, leading to clusters that are less compact or well-separated.

Visualizations:

1. K-means with



2. DBSCAN with PCA

3. K-Means with t-

4. DBCAN with t-sne