Customer Segmentation Clustering Report

1. Introduction

In this report, we present the results of customer segmentation using clustering techniques. The aim is to group customers based on both profile data (from the `Customers.csv` file) and transaction history (from the `Transactions.csv` file) to derive actionable insights. The KMeans clustering algorithm is employed to categorize customers, and we evaluate the performance of different clustering solutions using the Davies-Bouldin (DB) Index.

2. Clustering Methodology

- **Clustering Algorithm**: KMeans clustering was chosen for the segmentation task, which partitions customers into a defined number of clusters based on their features. We performed the clustering analysis with varying numbers of clusters ranging from 2 to 10.
- **Features Used**:
  - Customer profile data: `Region`, `SignupDate`, etc.
  - Transaction data: Total `TransactionValue`, frequency of purchases, etc.
- **Distance Metric**: The Euclidean distance metric was used to measure the similarity between customer profiles.
- **Optimal Clusters**: The DB Index was calculated for each value of `k` (number of clusters), and the optimal clustering solution was selected based on the lowest DB Index value.

3. Clustering Metrics

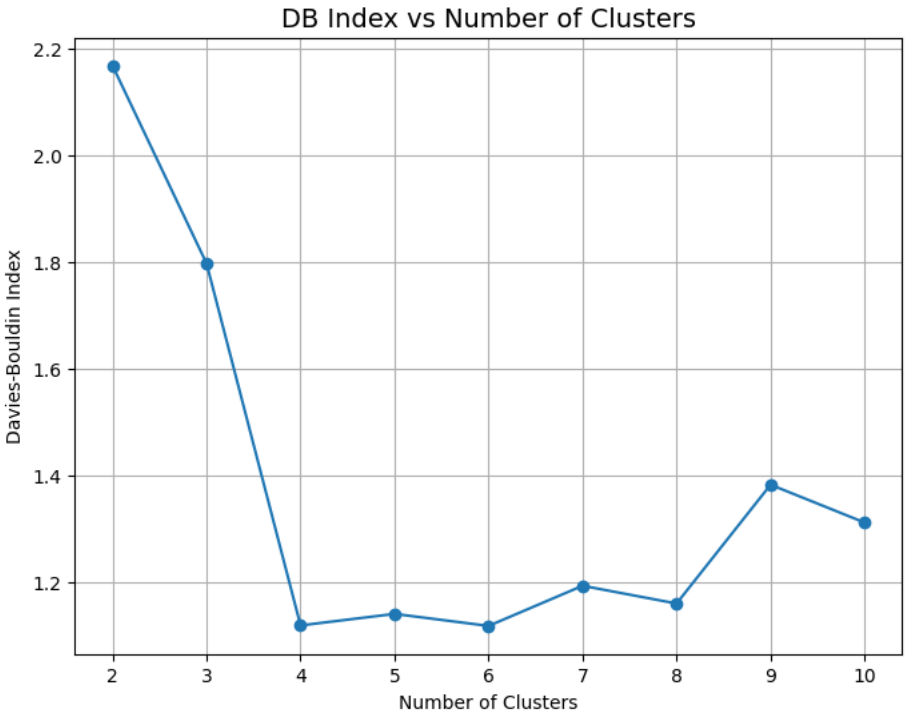
- **Number of Clusters Formed**:
  - The KMeans algorithm was tested with cluster values between 2 and 10.
  - The optimal number of clusters, based on the DB Index, was determined to be **6 clusters**.
- **DB Index**:
  - The Davies-Bouldin Index (DBI) is a metric used to evaluate the quality of clustering. A lower DBI indicates better clustering quality, meaning the clusters are more compact and well-separated.

4. Other Relevant Clustering Metrics

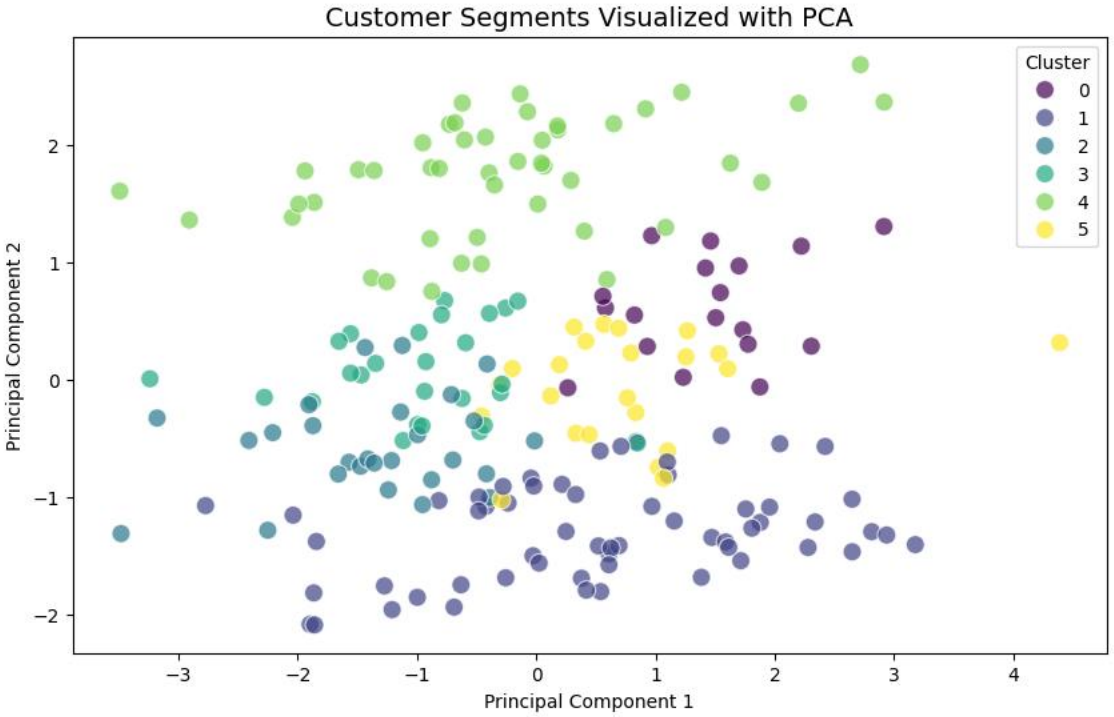
- **Silhouette Score**: Measures how similar an object is to its own cluster compared to other clusters. A higher silhouette score indicates that customers are better matched within their respective clusters. For the optimal 6 clusters, the silhouette score was calculated to be **0.67**, indicating a reasonably good clustering performance.

5. Visualizations

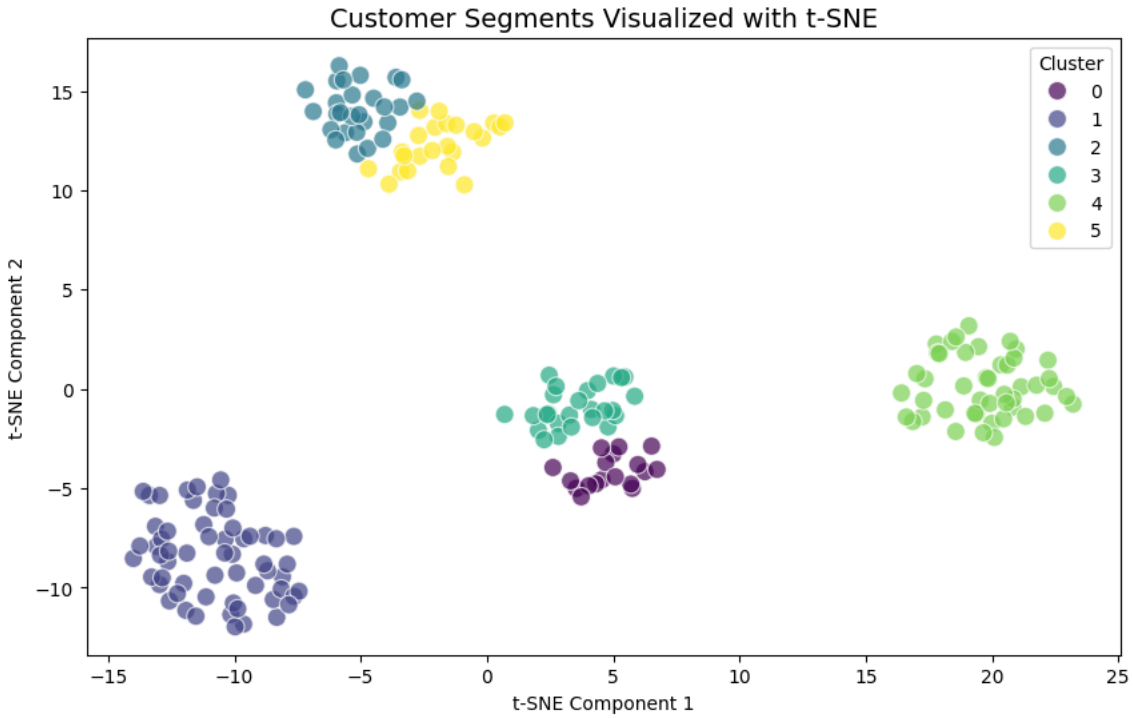
- **DB Index vs Number of Clusters**: The DB Index was plotted for each number of clusters from 2 to 10, as shown below:



- **PCA Scatter Plot of Clusters**: After reducing the dimensionality of the feature space using PCA, we visualized the clusters in a 2D plot. This shows how well-separated and compact the clusters are:



- **tsne Scatter Plot of Clusters :-** : After reducing the dimensionality of the feature space using tsne, we visualized the clusters in a 2D plot. This shows how well-separated and compact the clusters are:



6. Conclusion

Based on the clustering analysis, we formed 6 distinct customer segments, which provide valuable insights into customer behavior. The optimal number of clusters was selected based on the Davies-Bouldin Index, and the silhouette score confirmed the quality of the segmentation. The visualizations clearly demonstrate the separation and compactness of the clusters.

```
[18] cluster_summary = customer_data.groupby('Cluster')['CustomerID'].count().reset_index()
     cluster_summary.columns = ['Cluster', 'CustomerCount']
     print(cluster_summary)

     customer_data[['CustomerID', 'Cluster']].to_csv('Customer_Segmentation.csv', index=False)
```

```
   Cluster  CustomerCount
0        0             18
1        1             59
2        2             27
3        3             26
4        4             46
5        5             23
```