Customer Segmentation with Clustering: Report

1. Clustering Algorithm Used:

• K-Means Clustering Was Selected As The Clustering Method For Customers Segmentation.

• The number of clusters was selected based on the Davies-Bouldin (DB) index, which helps in assessing the compactness and separation of the clusters. The lower the DB index, the better the clustering.2. Steps Taken:

2. Steps Taken:

Data Preprocessing

Customer profile data (including SignupDate, Region) was merged with transaction data (Quantity, TotalValue, TransactionCount).

Categorical variables, like Region, were one-hot encoded.

Numeric features were standardized using StandardScaler.

3. Clustering Evaluation:

- The **Davies-Bouldin Index** was calculated for each cluster size (K=2 to K=10).
- The optimal number of clusters is chosen based on the **lowest DB Index** value.

4. Highest Clustering Performance:

• Optimal K: K=4 clusters (optimized as the smallest DB Index)

• DB Index for K=4: 1.25 (Lower indicates the better clustering performance)

• Other Important Metrics:

thas produced 4 different customer segments with transactional and profiling feature using K-means algorithm

o\thas observed that these clusters have good separability and the cluster boundaries are also close to each other; this is shown as a relatively low value of DB Index indicating boundary clarity between the clusters.

5. Visualization:

- o DB Index vs. Number of Clusters:
- o A line plot was created to represent the DB Index for each cluster size. The optimal number of clusters is where the DB Index is minimized.

- 2D Visualization of Clusters:
- PCA (Principal Component Analysis) was used to reduce the high-dimensional feature space to 2D for visualization.
- A scatter plot was constructed where each point represents a customer and is colored according to the assigned cluster.

6. Cluster Profile and Insights:

- **Cluster 1**: Typically customers with low transaction volume and low total value, possibly new or infrequent shoppers.
- **Cluster 2**: Customers with high transaction volume and moderate total value, likely frequent buyers.
- **Cluster 3**: Customers with low transaction count but high total value, potentially loyal high-spending customers.
- **Cluster 4**: Customers with moderate transaction volume and high value, balancing between frequent buyers and high spenders.

7. Final Output:

- The customer segments and their associated cluster labels were saved in a new CSV file (`Customer_Segments.csv`).

Clustering Results Summary:

- **Number of Clusters Formed**: 4 clusters
- **Best DB Index Value**: 1.25
- **Cluster Distribution**: Customers were segmented into 4 distinct groups based on both their profile and transaction data.