

Customer Segmentation Report

1. Objective:

The goal of this task was to perform **customer segmentation** using clustering techniques. The segmentation helps in identifying customer groups with similar behaviors, preferences, and purchase patterns. This can be beneficial for targeted marketing, personalised recommendations, and strategic decision-making.

2. Data Preprocessing:

We used two datasets:

- **Customers.csv**: Contains customer profile information such as **CustomerID**, **Region**, and **SignupDate**.
- **Transactions.csv**: Contains transaction history, including **TotalValue**, **Quantity**, and product-related details.

Steps Performed:

- **Merging Datasets**: We merged the customer profiles with their corresponding transaction data to create a comprehensive dataset.
- **Feature Engineering**:
 - Aggregated **Total Spend**, **Quantity Purchased**, and **Average Product Price** per customer.
 - Categorical data like **Region** was encoded using **one-hot encoding** for compatibility with clustering algorithms.
- **Scaling Features**: We used **StandardScaler** to normalise the numerical features. This step ensures that features with larger scales don't dominate the clustering process.

3. Clustering Approach:

- **Algorithm Used**: We applied the **K-Means Clustering Algorithm**, which groups data points into **k** clusters based on feature similarity.
- **Choosing Optimal Clusters**:
 - We evaluated cluster numbers from **2 to 10** using the **Davies-Bouldin (DB) Index**.
 - The **DB Index** measures the quality of clustering based on intra-cluster cohesion and inter-cluster separation. A **lower DB Index** indicates better clustering.
- **Optimal Number of Clusters**: The optimal number of clusters identified was **10**, with the **lowest DB Index of 0.898**.

4. Visualization:

To visualise high-dimensional data:

- We used **Principal Component Analysis (PCA)** to reduce the data to **2 dimensions**.
- A **scatter plot** was created to display the clusters, with each cluster represented by a different color. The separation between clusters indicates meaningful segmentation.

5. Clustering Metrics:

- **Optimal Number of Clusters:** 10
- **Davies-Bouldin Index:** 0.898

This indicates reasonably well-defined clusters with good separation.

Conclusion:

This customer segmentation provides meaningful insights into customer behaviour. The clustering can now be used for:

- **Targeted marketing campaigns**
- **Personalised recommendations**
- **Customer retention strategies**