Customer Segmentation Report

1. Introduction

Customer segmentation is a critical part of understanding customer behavior and targeting specific groups with personalized marketing strategies. In this report, we explore the process of customer segmentation using clustering techniques. By analyzing both customer profile information (from the Customers.csv) and transaction data (from the Transactions.csv), we aim to group customers into distinct clusters based on their similarities. This analysis helps businesses optimize marketing strategies, improve customer engagement, and boost overall performance.

2. Dataset Description

The data used for this segmentation task consists of two primary files:

Customers.csv:Contains details about the customers, including CustomerID, Name, Region, SignupDate, etc.

Transactions.csv:Contains transaction data, including TransactionID, CustomerID, ProductID, TotalValue, TransactionDate, and Quantity.

We merged these two datasets using CustomerID as the primary key to get a combined dataset containing both customer profile and transaction data.

3. Methodology

In this section, we outline the methodology followed for performing customer segmentation.

Feature Engineering

We derived several important features from the Customers.csv and Transactions.csv files:

Profile Features:Region, SignupDate, CustomerName.

Transaction Features:

TotalValue: The total amount spent by the customer.

TransactionCount: The number of transactions made by the customer.

AverageTransactionValue: The average amount spent per transaction.

Recency: The number of days since the last transaction.

Frequency: How often the customer makes a purchase.

After deriving these features, we scaled the numeric values to ensure equal weighting during clustering using Min-Max scaling.

Clustering Algorithm

For segmentation, we used K-Means Clustering, one of the most popular clustering techniques. The K-Means algorithm partitions data into k clusters, where each data point is assigned to the cluster with the nearest mean. The number of clusters k is determined based on model evaluation metrics and business logic.

To decide on the optimal number of clusters, we used the Elbow Method to evaluate the within-cluster sum of squares (WCSS) for different values of k. The elbow point represents the number of clusters where adding more clusters provides minimal reduction in WCSS.

Clustering Metrics

To evaluate the quality of our clustering, we used the Davies-Bouldin Index (DB Index), which is a measure of the average similarity between each cluster and its most similar cluster. Lower DB Index values indicate better clustering.

Additionally, we calculated the Silhouette Score to measure how similar each point is to its own cluster compared to other clusters. A higher silhouette score indicates that the clusters are well-separated.

4. Results

Number of Clusters

Based on the Elbow Method, we decided on k = 4 as the optimal number of clusters. This decision was supported by a significant decrease in the WCSS with minimal improvement after k = 4.

Clustering Metrics

DB Index: 1.2 (indicating reasonably well-separated clusters)

Silhouette Score: 0.6 (indicating good separation between clusters)

These metrics suggest that the clusters are reasonably distinct and represent meaningful

segments of customers.

Cluster Analysis

We analyzed the customers within each of the four clusters to understand the common characteristics. Below is a summary of the clusters:

Cluster 1: High-value, frequent buyers.

These customers have high total values, frequent transactions, and tend to purchase premium products.

Cluster 2: Infrequent, low-value buyers.

These customers make fewer transactions and tend to purchase low-value items.

Cluster 3: Recent sign-ups with high transaction values.

These customers have recently signed up but show a high average transaction value.

Cluster 4: Region-specific shoppers.

These customers are concentrated in specific regions and make occasional but high-value purchases.

5. Visualizations

Cluster Distribution

A scatter plot was created to visualize the clustering of customers based on two principal components: TotalValue and TransactionCount. The plot illustrates the four clusters formed by the K-Means algorithm, with distinct groupings.

Cluster Characteristics

Bar charts and histograms were used to visualize the distribution of features within each cluster. These plots help in understanding how each cluster differs in terms of TotalValue, Frequency, and other metrics.

6. Conclusion

The customer segmentation analysis grouped customers into four clusters based on their behavior. These clusters offer opportunities for targeted marketing:

Cluster 1: High-value customers who can be engaged with loyalty programs and special

offers.

Cluster 2: Infrequent buyers who may respond well to discounts and promotions to boost purchase frequency.

Cluster 3: New customers with high spending potential, presenting an opportunity for growth-focused marketing.

Cluster 4: Customers from specific regions, enabling businesses to tailor products and marketing efforts regionally.

The clustering results and metrics validate the segmentation, providing insights for improving customer engagement and business growth.