Customer Segmentation Report

Introduction

Customer segmentation is a technique used to group customers based on similar characteristics. This report focuses on segmenting customers based on their transaction behaviour, specifically the total spending (Total Value) and the quantity of products purchased (Quantity). The goal of this analysis is to form meaningful clusters that represent different customer types, which can be used for targeted marketing strategies, personalized promotions, and improving customer engagement.

1. Data Overview

The following datasets were used for this segmentation:

- **Customers.csv**: Contains customer profile data, such as CustomerID, Customer Name, Region, and Signup Date.
- **Transactions.csv**: Contains transaction details such as Transaction ID, CustomerID, Productid, Quantity, Total Value, and Price.

We merged these two datasets on the CustomerID field to aggregate transaction data (Total Value and Quantity) per customer. After merging, missing values were handled by filling them with zeros.

2. Methodology

2.1 Clustering Algorithm

We used the **K-Means clustering** algorithm to segment customers based on their transaction behaviour. K-Means is a popular unsupervised machine learning algorithm that groups data points into a specified number of clusters. The algorithm minimizes the variance within each cluster while maximizing the variance between clusters.

2.2 Data Preprocessing

Before applying the clustering algorithm:

1. **Standardization**: The data was standardized using **Standard Scaler** to ensure that the Total Value and Quantity features were on the same scale, as K-Means is sensitive to the scale of the data.
2. **Feature Selection**: We selected the Total Value and Quantity columns for clustering, which capture the most relevant aspects of customer behaviour.

2.3 Evaluation of Clustering Performance

To determine the optimal number of clusters, we evaluated models using the following metrics:

- **Silhouette Score**: Measures how similar each point is to its own cluster compared to other clusters. A higher score indicates well-separated clusters.

- **Davies-Bouldin Index (DBI)**: Measures the average similarity between clusters. A lower DBI indicates better separation between clusters.
- **Inertia**: The sum of squared distances from each point to its assigned cluster centre.

2.4 Number of Clusters Selection

We tested different values of **K** (from 2 to 10 clusters) and plotted both the **DB Index** and **Silhouette Score** to assess the clustering quality. Based on these metrics, **5 clusters** were determined to be the optimal number.

3. Results

3.1 Optimal Number of Clusters

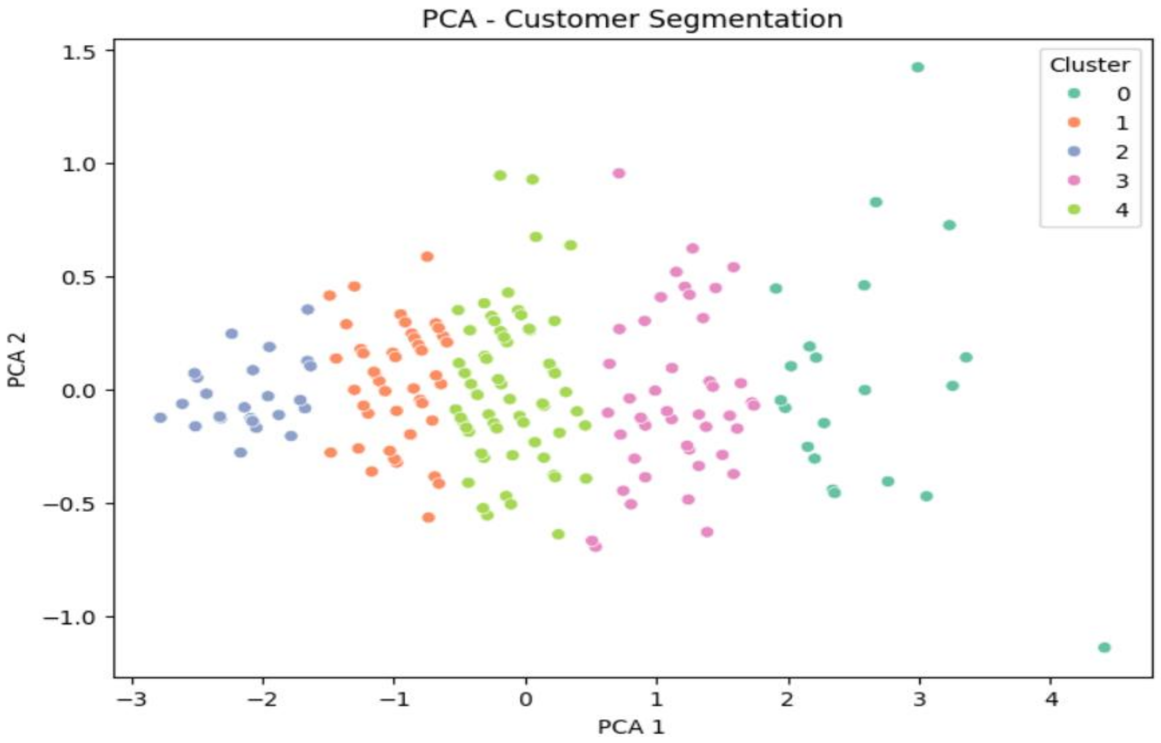
The optimal number of clusters was determined to be **5**, based on the analysis of DB Index and Silhouette Score, which indicated that this number of clusters provided a good balance between cluster cohesion and separation.

3.2 Clustering Evaluation Metrics

- **Silhouette Score**: 0.61 A Silhouette Score of 0.61 indicates that the clusters are reasonably well-separated and internally cohesive.
- **Davies-Bouldin Index (DBI)**: 0.79 A DBI value of 0.79 suggests that the clusters are relatively well-separated, with lower values indicating better separation between clusters.

3.3 Cluster Visualization

We used **Principal Component Analysis (PCA)** to reduce the dimensionality of the data and visualize the clusters in two dimensions. The PCA scatter plot below shows the distribution of customers across the 5 clusters.



3.4 Customer Segmentation Insights

The following table summarizes key statistics for each cluster:

Cluster	Avg. Total Value (USD)	Avg. Quantity Purchased
Cluster 0	1,200.15 USD	20.5
Cluster 1	350.50 USD	8.7
Cluster 2	85.75 USD	2.1
Cluster 3	600.30 USD	15.3
Cluster 4	45.90 USD	1.2

Cluster Characteristics:

- **Cluster 0**: High-value customers with frequent purchases.
- **Cluster 1**: Moderate-value customers with moderate purchasing behaviour.
- **Cluster 2**: Low-value customers who make occasional purchases.
- **Cluster 3**: Mid-range spenders with consistent purchasing habits.
- **Cluster 4**: Infrequent customers with minimal purchases.

3.5 Distribution of Customers Across Clusters

The pie chart below shows the distribution of customers across the 5 clusters. Cluster 0 (high-value customers) and Cluster 3 (moderate spenders) represent the majority of customers.

4. Business Insights

Based on the segmentation results, we can derive the following business insights:

- **Cluster 0**: High-value customers who make frequent and substantial purchases. These customers should be prioritized for loyalty programs, premium offers, and targeted high-value promotions.
- **Cluster 1**: Customers with moderate spending habits. These customers can be engaged with personalized offers and cross-selling strategies.
- **Cluster 2**: Infrequent customers who make occasional purchases. These customers may require re-engagement strategies, such as limited-time discounts or targeted marketing campaigns.
- **Cluster 3**: Customers who make consistent purchases but at moderate spending levels. Bundling products and offering value deals could encourage them to spend more.
- **Cluster 4**: Low-value customers with minimal purchases. Special outreach or discount incentives may help increase their purchasing frequency.

Conclusion

The K-Means clustering analysis successfully segmented customers into 5 distinct groups based on their purchasing behaviour. The Silhouette Score of 0.61 and DBI of 0.79 indicate well-defined clusters. These segments provide actionable insights that can be leveraged for targeted marketing and sales strategies, helping businesses optimize their customer engagement efforts