# Customer Lookalike Recommendation Analysis Report

## Executive Summary

This report outlines the methodology and results of generating customer lookalike recommendations based on transactional and profile data. Using advanced data processing techniques, we identified customers with similar purchasing behaviors and characteristics to the target customers, which can be utilized for enhancing marketing strategies, customer engagement, and segmentation.

## Objectives

The primary objective of this analysis is to identify **lookalike customers** for a specified group of target customers (C0001 to C0020). By leveraging aggregated transaction data and customer profile information, this analysis aims to provide actionable insights that can guide personalized marketing and customer retention strategies.

## Data Overview

The analysis is based on two key datasets:

### Transactional Data

This dataset contains detailed transaction information, including:

- `CustomerID`: Unique identifier for each customer.
- `TotalValue`: Total monetary value of the transaction.
- `Quantity`: The number of products purchased in a given transaction.
- `ProductID`: Unique identifier for each product purchased.
- `Category`: Product category.
- `Month`: The month during which the transaction occurred.

### Customer Profile Data

This dataset includes the following customer-specific details:

- `CustomerID`: Unique identifier for each customer.
- `Region`: Geographic location of the customer.
- `SignupDate`: Date when the customer signed up.

## Methodology

The methodology for generating the lookalike recommendations involves the following steps:

### Step 1: Feature Engineering

Customer-level features were derived from the transactional data using aggregation functions:

- **Total Spend (TotalValue)**:
$$\text{Total Spend} = \sum(\text{TotalValue})$$
- **Average Transaction Value**:
$$\text{Average Transaction Value} = \frac{\sum(\text{TotalValue})}{\text{Number of Transactions}}$$
- **Transaction Count**:
$$\text{Transaction Count} = \text{Count of Transactions}$$
- **Total Quantity**:
$$\text{Total Quantity} = \sum(\text{Quantity})$$
- **Average Quantity per Transaction**:
$$\text{Average Quantity} = \frac{\sum(\text{Quantity})}{\text{Number of Transactions}}$$
- **Unique Categories**:
$$\text{Unique Categories} = \text{Number of Unique Product Categories}$$
- **Unique Products**:
$$\text{Unique Products} = \text{Number of Unique Products}$$
- **Active Months**:
$$\text{Active Months} = \text{Number of Unique Months with Transactions}$$

## Step 2: Customer Profile Augmentation

Additional features were derived from the customer profile dataset:

- **Customer Age** (in days) was calculated as the difference between the current date (2024-08-25) and the `SignupDate`:

$$\text{Customer Age} = \text{Current Date} - \text{SignupDate}$$

- **Region**: One-hot encoding was applied to the `Region` column, where each region is represented as a separate binary feature:

$$\text{Region}_i = \begin{cases} 1 & \text{if customer is from region i} \\ 0 & \text{otherwise} \end{cases}$$

## Step 3: Normalization

To ensure comparability across customers, all customer features were scaled using `StandardScaler` from `sklearn`. This step standardizes the data, removing any bias due to differing scales of measurement between features. The formula for standardization is:

$$\text{Standardized Feature} = \frac{X - \mu}{\sigma}$$

Where:

- $X$ is the raw feature value,
- $\mu$ is the mean of the feature, and
- $\sigma$ is the standard deviation of the feature.

## Step 4: Cosine Similarity Calculation

Cosine similarity measures the cosine of the angle between two vectors. Given two vectors $A$ and $B$, the formula is:

$$\text{Cosine Similarity}(A, B) = \frac{A \cdot B}{\|A\| \|B\|}$$

Where:

- $A \cdot B$ is the dot product of the two vectors,
- $\|A\|$ is the magnitude (Euclidean norm) of vector $A$, and
- $\|B\|$ is the magnitude of vector $B$.

### Step 5: Identification of Lookalike Customers

For each target customer, the cosine similarity between their feature vector and those of all other customers was computed. The top N most similar customers were identified, where $N = 3$ in this analysis.

## Results

The analysis successfully identified the most similar customers for each target customer based on their transaction behaviors and profile characteristics. The following table summarizes the top 3 lookalikes for a selection of target customers:

| **Target Customer** | **Lookalike 1 (CustomerID)** | **Similarity Score** | **Lookalike 2 (CustomerID)** |
|---|---|---|---|
| C0001 | C0174 | 0.9654 | C0011 |
| C0002 | C0159 | 0.8898 | C0005 |
| C0003 | C0129 | 0.8642 | C0190 |

## CSV Output

The lookalike recommendations were saved in a CSV file titled `Lookalike.csv`. The file contains the following structure:

```
CustomerID,Lookalikes
C0001,"[('C0174', 0.9654), ('C0011', 0.9474), ('C0152', 0.9126)]"
C0002,"[('C0159', 0.8898), ('C0005', 0.8885), ('C0134', 0.8414)]"
C0003,"[('C0129', 0.8642), ('C0190', 0.7735), ('C0006', 0.7373)]"
...
```