---

# Model Development Summary

## 1 Feature Creation Logic

I combined customer and transaction data to build a **feature matrix** capturing each customer's behavior and profile. The key features include:

1. **Spending by Category**
    - Created a pivot table where each row (customer) includes total spending for each product category.
    - This helps identify customers with similar spending patterns across categories such as Electronics, Books, Clothing, etc.
2. **Purchase Count**
    - Calculated the number of distinct transactions per customer.
    - Frequent buyers with similar purchase frequencies may respond similarly to promotions.
3. **Total Spend**
    - Summed the total value of purchases per customer.
    - High-spending customers tend to cluster together, indicating similar levels of brand engagement.
4. **Days Since Signup** (Optional)
    - Derived by calculating the time difference between signup date and a reference date (e.g., current date).
    - Customers in similar lifecycle stages may exhibit comparable shopping patterns.

## 2 Choice of Similarity Metric

I used **cosine similarity** to gauge how alike two customers are based on their feature vectors. Cosine similarity focuses on the "direction" of each vector rather than its magnitude, which is particularly useful for emphasizing **relative spending patterns** rather than absolute differences.

## 3 Data Preparation

Before computing similarities, I:

1. **Ensured Data Quality**: Handled missing values by assigning zeros for any unrecorded purchases.
2. **Converted Data Types**: Confirmed that columns for Price, Quantity, and TotalValue were numeric.

3. **Constructed the Pivot Table**: Created a unified table where each row represents a unique customer and each column represents a selected feature.
4. **Considered Normalization**: Evaluated the option to apply scaling (e.g., `StandardScaler`) to ensure that no single feature—like Total Spend—would dominate the similarity metric.

## 4 Evaluation

After building the similarity matrix, I performed a qualitative check to see if the nearest neighbors made sense. For instance, high-spending customers generally showed other high spenders as top recommendations, and category specialists (e.g., mostly Books) often had similar category enthusiasts in their top matches. A more rigorous evaluation could involve A/B testing or comparisons with real-world conversion metrics if available.

## 5 Output Format

I generated a **Lookalike.csv** file containing the top 3 most similar customers for the first 20 customers (`C0001 - C0020`). Each record in the CSV contains:

- **CustomerID** : The focal customer.
- **LookalikeMap** : A compressed string listing the three closest CustomerIDs and their cosine similarity scores.

This format streamlines further integration with other tools or dashboards, enabling targeted marketing, cross-selling, or personalized engagement strategies.