Report on Lookalike Customer Model

Objective:

The goal of this model is to find the most similar customers to a given set of customers based on transaction and demographic data. This is achieved by calculating the cosine similarity between the feature vectors of customers. The result is a set of lookalike customers who share similar purchasing behavior and demographic characteristics.

Data Overview:

- **Customers.csv**: Contains customer demographic information (CustomerID, Region, SignupDate).
- **Products.csv**: Contains information about products (ProductID, Category).
- **Transactions.csv**: Records transaction details for customers, including ProductID and TotalValue spent on products.

Preprocessing Steps:

1. **Data Cleaning**:
    - Stripped spaces from column names in the Transactions dataset.
    - Renamed columns for clarity (e.g., Category_x renamed to Category).
2. **Data Merging**:
    - Merged product information (ProductID, Category) with the transaction data to associate product categories with each transaction.
    - Aggregated transaction data by CustomerID and Product Category using TotalValue as the aggregation metric.
3. **Feature Engineering**:
    - Created a new feature, SignupDays, representing the number of days since a customer signed up.
    - Combined customer demographics (CustomerID, Region, SignupDays) with transaction-based features (aggregated values for each product category).
4. **Handling Missing Data**:
    - Missing values in the customer data were filled with zeros.
5. **Feature Normalization**:
    - Transaction columns (TotalValue for each product category) were standardized using StandardScaler.

Model Logic:

- **Cosine Similarity**: We used the cosine similarity metric to calculate the similarity between customers. This metric compares the angle between two vectors, with a smaller angle indicating higher similarity.
- **Lookalike Matching**: For the first 20 customers (C0001 to C0020), we calculated the similarity scores with all other customers and selected the top 3 most similar customers. The top similar customers were sorted based on the highest cosine similarity score.

Output:

A CSV file named Lookalike.csv was generated, containing the following columns:

- **CustomerID**: The ID of the customer whose lookalikes are being identified.
- **Lookalike_CustomerID**: The ID of a customer who is most similar to the given customer.
- **Similarity_Score**: The cosine similarity score between the two customers.

Example Output for Customers C0001 to C0020:

CustomerID	Lookalike_CustomerID	Similarity_Score
C0001	C0091	0.988881
C0001	C0069	0.984308
C0001	C0184	0.978609
C0002	C0159	0.979384
C0002	C0036	0.956507
C0002	C0134	0.907855
C0003	C0007	0.996860
C0003	C0085	0.964024
C0003	C0166	0.960495
C0004	C0075	0.983289
C0004	C0090	0.921588
C0004	C0065	0.885556
C0005	C0197	0.967718
C0005	C0085	0.963489
C0005	C0166	0.949384
C0006	C0169	0.970556
C0006	C0185	0.930166
C0006	C0081	0.927877
C0007	C0003	0.996860
C0007	C0085	0.979066

Clustering Metrics and Evaluation:

- **Cosine Similarity Score**: This is the primary metric used to evaluate the "closeness" between customers. A score close to 1 indicates high similarity.

Visual Representation:

- A visualization of the similarity matrix or a scatter plot showing customers clustered based on similarity scores could be used to visually represent the clusters.

Conclusion:

This model successfully identifies lookalike customers by using both demographic and transaction data. The cosine similarity metric is an effective tool for identifying customers with similar purchasing patterns and demographics. The output file provides a list of lookalikes, which could be used for targeted marketing or personalized product recommendations.