Report on eCommerce Transactions Data Analysis

This report presents the steps taken to perform **Exploratory Data Analysis (EDA)**, build a **Lookalike Model**, and apply **Customer Segmentation** using KMeans clustering on a given eCommerce dataset consisting of customer, product, and transaction data. The goal is to derive actionable insights, identify similar customers, and segment customers for targeted marketing.

---

1. Exploratory Data Analysis (EDA)

The first step in the analysis was to load the datasets: Customers.csv, Products.csv, and Transactions.csv. Each dataset was explored to understand its structure and potential issues.

- **Missing Values and Duplicates**: Initially, we checked for missing values and removed any duplicates. Columns like CustomerID and ProductID were crucial identifiers, and we ensured no missing data was present in essential fields like TotalValue and Quantity.
- **Summary Statistics**: We examined the distribution of numerical features such as Price, Quantity, and TotalValue. The distribution of Price indicated a skewed distribution, while Quantity revealed some outliers, which were removed using Z-scores.
- **Univariate and Bivariate Analysis**: We visualized distributions using histograms and boxplots. For instance, TotalValue was highly skewed, suggesting a few high-value transactions. The relationship between Region and TotalValue was explored using a bar plot to understand how sales were distributed across regions. We also analyzed the relationship between product Category and the quantity sold, revealing that some categories performed significantly better than others.
- **Feature Engineering**: We created a new feature, CustomerTenure, by calculating the difference between the latest transaction date and the SignupDate. This provided insights into the lifecycle of customers and helped in further profiling.

---

2. Lookalike Model

To create a **Lookalike Model**, the goal was to recommend similar customers based on both their transaction history and profile information.

- **Customer Profile Creation**: We aggregated each customer's transaction data, computing their total spending (TotalValue) and product preferences (using a combination of product categories bought). These preferences were transformed using **One-Hot Encoding** to convert categorical data into numeric form.

- **Cosine Similarity**: Using the customer profiles, we computed the **Cosine Similarity** between customers. This metric measures the similarity of two customers' profiles based on their total spending and product preferences. We then identified the top 3 most similar customers for each of the first 20 customers.
- **Results**: A map of lookalikes was created where each customer had a list of 3 similar customers along with their similarity scores. This model can be used to identify potential high-value customers who share similar characteristics.

---

3. Customer Segmentation / Clustering

In this step, the objective was to segment customers into groups with similar behaviors for targeted marketing using **KMeans clustering**.

- **Data Preprocessing**: We first prepared the customer data by removing unnecessary columns (like CustomerID) and performing **One-Hot Encoding** on the Category column to convert product preferences into numerical features. This step ensured that all features were numeric and could be processed by clustering algorithms.
- **Scaling**: We applied **StandardScaler** to normalize the numerical features, ensuring that all variables (e.g., TotalValue and the one-hot encoded categories) were on the same scale. This is crucial for clustering algorithms like KMeans, which are sensitive to the magnitude of features.
- **Clustering**: We performed **KMeans clustering** with 5 clusters (chosen arbitrarily). Each customer was assigned to a cluster based on their profile, which included both spending behavior and product preferences.
- **Evaluation**: The **Davies-Bouldin Index** (DB Index) was used to evaluate the quality of the clustering. A lower DB Index indicates well-separated clusters. The resulting DB Index was calculated to assess the effectiveness of the clustering.
- **Cluster Insights**: Each cluster likely represents customers with distinct purchasing behaviors and preferences. For example, one cluster could represent high-value customers who frequently buy products from a specific category, while another might consist of customers with sporadic purchases.