Data Science Assignment: eCommerce Customer Segmentation and Clustering

This report presents the findings of the **Customer Segmentation** task, part of an eCommerce transaction analysis. The goal of this task is to identify distinct customer segments based on transaction data. Clustering techniques were applied to generate actionable insights regarding customer behavior, which could help in tailoring marketing strategies, improving customer experience, and optimizing product offerings.

1.Data Overview

Three datasets were provided for this assignment:

- **Customers.csv**: Contains information about customers, such as their unique ID, name, region, and signup date.
- **Products.csv**: Contains details about products, such as product ID, name, category, and price.
- **Transactions.csv**: Contains transaction data, including transaction ID, customer ID, product ID, transaction date, quantity, total value, and price.

These datasets were merged to combine customer, product, and transaction information for comprehensive analysis.

2. Exploratory Data Analysis (EDA)

Data Inspection

The first step was to load and inspect the datasets. Here are the first few rows of the datasets:

- **Customers Dataset**: Shows customer details like ID, name, region, and signup date.
- **Products Dataset**: Contains product-related information such as name, category, and price.
- **Transactions Dataset**: Includes transaction records with customer IDs, product IDs, and financial details.

Missing Values Analysis

We checked for missing values in each dataset to ensure the quality of data. No significant missing data was found, which is crucial for ensuring robust analysis.

Descriptive Statistics

Basic statistics were calculated for the **Transactions** dataset. This includes details such as total quantity, price, and total value of products sold, giving us a clearer picture of the eCommerce transactions.

3. Clustering: Customer Segmentation

Feature Engineering

We created new features to better represent customer behavior:

- **Total Spend**: Total amount spent by each customer.
- **Frequency of Purchase**: Number of transactions made by each customer.
- **Recency of Purchase**: The number of days since the customer made their most recent purchase.

These features are crucial for clustering, as they capture key aspects of customer behavior.

Data Preprocessing

We applied feature scaling using **StandardScaler** to standardize the features (Total Spend, Frequency, Recency) so that no feature dominates the clustering process due to scale differences.

Clustering with KMeans

We used **KMeans clustering** with **4 clusters** to segment the customers based on the engineered features. KMeans is chosen for its simplicity and efficiency in partitioning data into groups.

Clustering Evaluation

We evaluated the clustering result using the **Davies-Bouldin Index (DB Index)**, which is a metric used to assess the quality of clusters. A lower DB Index indicates better clustering performance. The obtained value for the DB Index was **1.013**, suggesting reasonable cluster separation.

Cluster Visualization

To visualize the customer segments, we performed **Principal Component Analysis (PCA)** for dimensionality reduction. The two principal components were plotted to visualize how the clusters are distributed in a 2D space.

The visualization showed distinct clusters representing different customer groups, which can be analyzed further.

4. Insights from Clustering

Average Spend by Cluster

We computed the average total spend for each customer cluster. The clusters varied in terms of total spend, with some clusters showing high spenders, while others were more budget-conscious.

Cluster-Specific Statistics

Further analysis of the clusters revealed the following key insights:

- Cluster 1: High spenders with frequent purchases and recent activity.
- Cluster 2: Low spenders with infrequent and older purchases.
- Cluster 3: Moderate spenders with a good mix of frequency and recency.
- Cluster 4: Budget-conscious customers with occasional purchases.

These insights are useful for tailoring marketing efforts to different customer segments.

5. Conclusion

The customer segmentation exercise provides valuable insights into the behavior of different customer groups. By understanding these segments, the business can focus on personalized marketing, improving customer retention, and offering products suited to specific customer needs.

Further clustering iterations and metrics could be applied for more refined insights.

7. Recommendations

Based on the clustering results, the following strategies are recommended:

- **High spenders**: Offer loyalty programs and premium services to keep them engaged.
- **Low spenders**: Target them with discounts or promotions to increase their purchase frequency.
- **Moderate spenders**: Upsell and cross-sell relevant products to increase spend.
- **Budget-conscious customers**: Provide affordable product options and bundle offers.

8. References

- Data Source: Customers.csv, Products.csv, Transactions.csv
- Clustering Algorithm: KMeans
- Evaluation Metric: Davies-Bouldin Index
- Visualization: PCA for dimensionality reduction