Customer Segmentation / Clustering

1. Introduction

Objective:-

I have used the given datasets and merged to get the valuable insights and then performed the clustering algorithm.

Data Overview:-

The two datasets I have used are **Customers.csv** and **Transaction.csv** and eventually merged the datasets to calculate the valuable information to form clusters.

2. Methodology

Preprocessing:-

First I checked the missing values, every column was non null. Then convert the format of transactiondate and signupdate.

- **Recency:** Calculated how many days have passed since the customer have made a new transaction respective to the present date
- **Transaction Frequency:** Calculates the total number of transactions per customer has made.
- **Total Spend:** Calculated the sum of total value for each customer.
- **Region:** Then added the region column into the new feature dataframe, also performed the one hot encoding to convert the categorical data in binary form.

Standard scaling: Normalized the numeric columns in the dataframe (range -1 to 1). Dropped the the categorical columns as they were not required to form robust clusters

Clustering Algorithms:-

Applied Elbow method to determine the most relevant number of clusters. For this task I have applied **4 clusters** as there is a sudden drop, in the graph the WCSS value for 4 is 600. I had tried 10 as cluster value because WCSS value was the lowest there but when plotting the clusters, it was not a good idea for further exploration.

Then I implemented K-means algorithm compared to other algorithm it was an easy choice as there was no signs of outliers in large number which would affect the algorithm.

Then I plotted the 4 clusters by applying PCA1 and PCA2, after seeing the graph it was noted that all the four clusters were very well distributed and separated. Also visualized the clusters using TSNE

To get the summary of the clusters I calculated the mean of numeric columns. Below is the information i got:

Cluster 0:

Recency: Low (most recent activity); customers are relatively recent users.
Transaction Frequency: Below average.
Total Spend: Slightly below average.
Region: Strong presence in Europe; weak in other regions.

Interpretation:

Likely contains **moderately engaged** customers with **low spending** and activity in **Europe**.

Cluster 1:

Recency: Negative, indicating the oldest activity among clusters (least recent users).
Transaction Frequency: Slightly above average.
Total Spend: High spending.
Region: Primarily from South America.

Interpretation:

Represents **high-spending**, moderately **frequent customers** who are less recent users, concentrated in **South America**.

Cluster 2:

Recency: Positive, meaning the cluster has the least recent users.
Transaction Frequency: Highest frequency among clusters.
Total Spend: Low spending.
Region: Primarily from North America.

Interpretation:

Composed of **frequent shoppers** who spend less, predominantly from **North America**.

Cluster 3:

Recency: Neutral; activity is neither too recent nor too old.
Transaction Frequency: Slightly below average.
Total Spend: Near average.
Region: Dominantly from Asia.

Interpretation:

Reflects **average users** with balanced recency and spending patterns, mostly from **Asia**.

Metrics:

Davies-boudin-score: Got 0.95

Calinski -harabasz-score: Got 88.33

3. Summary

Completed customer segmentation analysis using **Customers.csv** and **Transactions.csv**, merged them to create features like recency, transaction frequency, total spend. After preprocessing and standard scaling, 4 clusters were chosen using the Elbow method. Visualizations showed well-separated clusters, with profiles such as high spenders from South America and frequent low spenders from North America. Evaluation was done using metrics like **Davies-boudin-score(0.95)** and **Calinski -harabasz-score: (88.33).**