Data Science Assignment eCommerce Transactions

Objective

To Perform customer segmentation using clustering techniques. Using both profile information (from Customers.csv) and transaction information (from Transactions.csv).

Dataset Description

Files Description:

1. Customers.csv
   - CustomerID: Unique identifier for each customer.
   - CustomerName: Name of the customer.
   - Region: Continent where the customer resides.
   - SignupDate: Date when the customer signed up.
2. Products.csv
   - ProductID: Unique identifier for each product.
   - ProductName: Name of the product.
   - Category: Product category.
   - Price: Product price in USD.
3. Transactions.csv
   - TransactionID: Unique identifier for each transaction.
   - CustomerID: ID of the customer who made the transaction.
   - ProductID: ID of the product sold.
   - TransactionDate: Date of the transaction.
   - Quantity: Quantity of the product purchased.
   - TotalValue: Total value of the transaction.
   - Price: Price of the product sold.

Procedures

1.Data Collection

The first step is collecting the data. The code assumes that the data is available in CSV format for three tables:

- **Customers.csv**: Contains customer-specific information such as CustomerID, SignupDate, and Region.
- **Products.csv**: Contains details about the products sold.
- **Transactions.csv**: Contains transactional data such as TransactionID, CustomerID, TotalValue, and TransactionDate.

2. Importing Libraries

The necessary libraries are imported to handle various tasks:

- **Pandas**: Used for data manipulation and handling.
- **NumPy**: Used for numerical operations.
- **Matplotlib and Seaborn**: For visualization, though these are not used in the final code snippet.
- **Scikit-learn**: For machine learning and clustering, particularly KMeans, NearestNeighbors, StandardScaler, etc.
- **Datetime**: For working with dates.

3. Data Preprocessing

a. Datetime Conversion

The SignupDate and TransactionDate columns are converted to datetime objects to facilitate date-based calculations:

b. Customer Aggregation

Customer-level aggregates are computed from the transactional data:

- **TotalValue**: Total money spent by each customer.
- **Quantity**: Total number of items bought by each customer.

c. Recency, Frequency, and Monetary (RFM) Calculation

The **RFM** values are calculated to represent each customer's activity:

- **Recency**: The number of days since the customer's last purchase.
- **Frequency**: The number of transactions made by the customer.
- **Monetary**: The total value of all transactions by the customer.

d. Normalization of RFM Values

The RFM values are normalized to ensure they are on the same scale.

4. Feature Engineering

The **LabelEncoder** is used to convert categorical variables (such as Region) into numerical values, which makes it suitable for machine learning models.

5. Clustering (KMeans)

a. Selecting Features for Clustering

The features selected for clustering include Region, Recency, Frequency, and Monetary

b. Determining Optimal Number of Clusters (Elbow Method)

To find the optimal number of clusters, the **elbow method** is applied. This involves running KMeans clustering for a range of cluster numbers (1 to 10), and plotting the **Within-Cluster Sum of Squares (WCSS)** to visualize the elbow point:

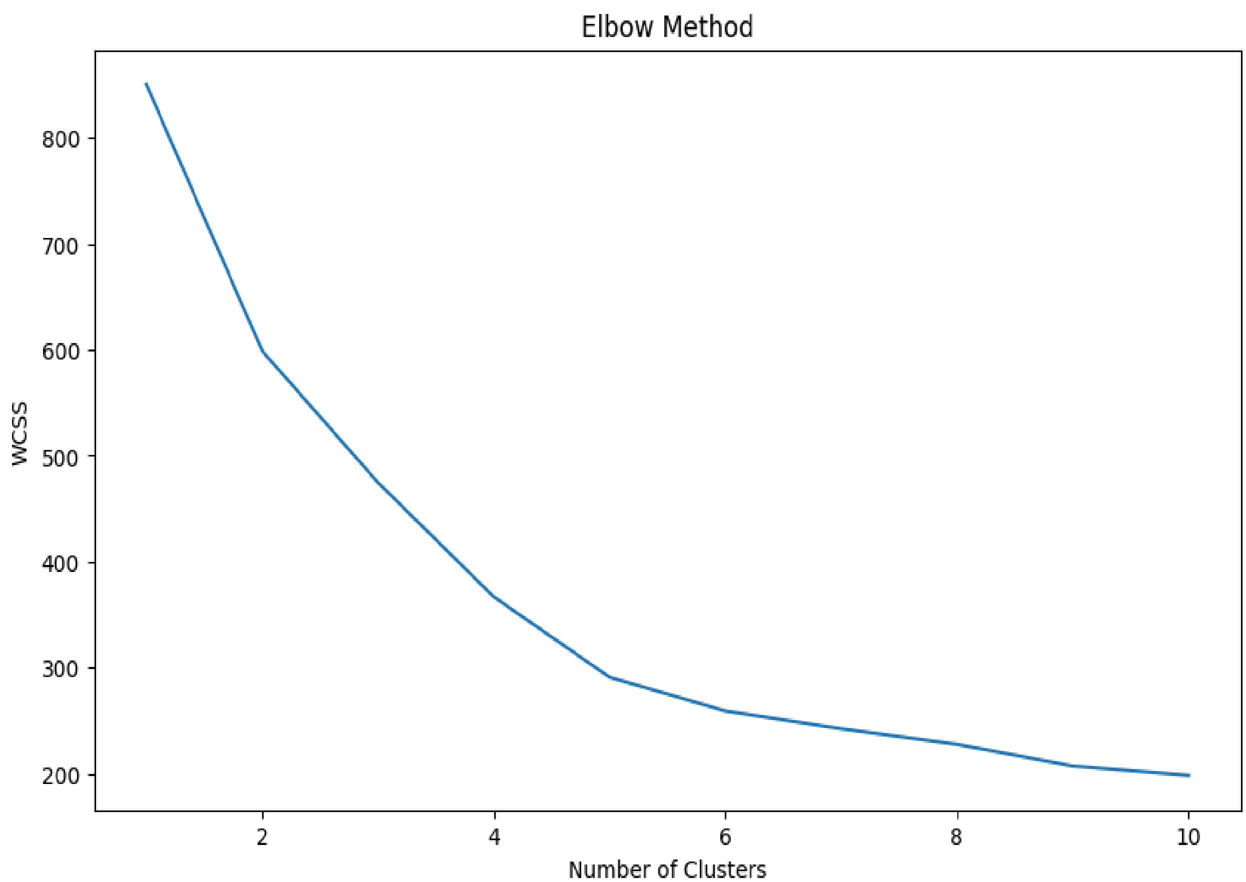

Figure: To determine number of clusters using elbow method

c. KMeans Clustering

A KMeans model is created with the chosen number of clusters, 9 number was chosen for this assignment

6. Model Evaluation

a. Davies-Bouldin Index

The **Davies-Bouldin Index** is computed to evaluate the quality of the clustering. A lower DB index indicates better-defined clusters:

DB Index value:1.14

b. Visualization

A **3D scatter plot** is created to visualize the clusters. The axes represent Recency, Frequency, and Monetary:

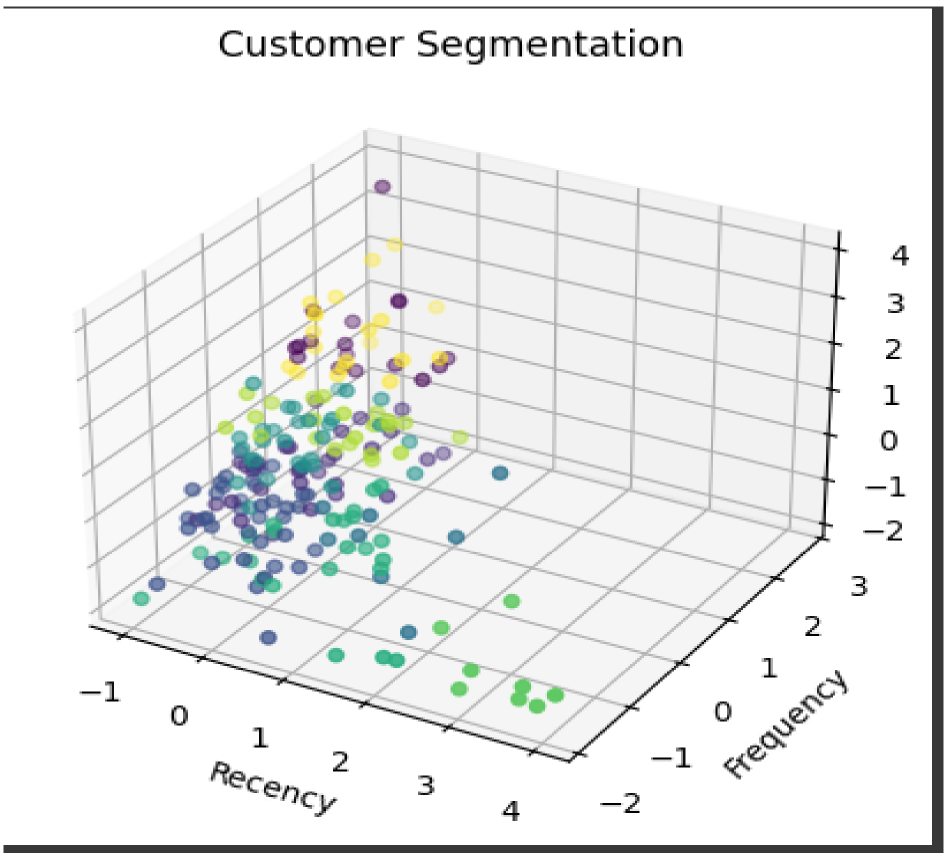

Figure:Customer Segmentation

7. Cluster Profiling

After clustering, an analysis of the clusters is performed to understand their characteristics:

Cluster	Region	Recency	Frequency	Monetary
0	0	-0.44	1.43	1.43
1	2	-0.25	0.145	-0.33
2	0	-0.41	-0.73	-0.76
3	[0,1]	1.41	-0.81	-0.26
4	3	-0.35	0.16	0.23
5	3	0.60	-0.98	-0.87
6	[0,2]	3.31	-1.66	-1.41
7	1	0.02	0.32	0.57
8	3	-0.54	1.45	1.46

Table: Cluster profiling

Conclusion

In this analysis, we used **KMeans clustering** to segment customers based on their **Recency**, **Frequency**, and **Monetary** (RFM) values, along with their **Region**. The key steps involved data preprocessing (such as calculating RFM metrics), feature engineering (encoding categorical variables), and clustering customers using the **Elbow Method** to determine the optimal number of clusters.

Key findings include:

- **Customer Segments**: Different customer groups were identified, each with unique purchasing behaviors.
- **Business Implications**: These segments can guide targeted marketing strategies. High-value customers can be offered loyalty programs, while low-value ones may benefit from personalized promotions.