Online Retail Customer Segmentation Analysis

Problem Statement

An online retail store is trying to understand the various customer purchase patterns for their firm, you are required to give enough evidence-based insights to provide the same.

Project Objective

a. Using the provided data, find useful insights about the customer purchasing history, that can be added advantage for the online retailer.

b. Segment the customers based on their purchasing behavior.

Data Description

Data File: Online Retail (3) (2).csv

Feature Name	Description	Data Type
InvoiceNo	Invoice Number	354345 non-null object
Stock Code	Product ID	354345 non-null object
Description	Product Description	354345 non-null object
Quantity	Quantity of the product	354345 non-null int64
Invoice Date	Date of the invoice	354345 non-null object
Price	Price of the product per unit	354345 non-null float64
Customer ID	Customer ID	354345 non-null float64
Country	Region of Purchase	354345 non-null object
Amount	Total amount	354345 non-null float64

Data Pre-processing Steps and Inspiration

- Df.head()→ This function is used to obtain a glimpse of data that the file constitutes and different columns with their content.
- Df.shape→ This is for checking dimension of the dataset.
- Df.info ()→This provides with the list of columns and the count of the non-null data within the column along with the data types. This helps to understand different column dtypes and take necessary steps to update the same.
- Df.describe()→This function is used to viewing some basic details of dataset.
- Df.isnull().sum()→ This functions provide with the sum of the missing values in each column if any, which can hint us to use different functions like apply and np.where to replace the blanks or nan data to replace the required column information with mean or median or mode as necessary.
- Df["InvoiceDate"] = pd.to_datetime(df["InvoiceDate"])→ To convert the data of the Invoice column into datetime in datasets, so that we can further process the data, for preprocessing and accurate evaluations.One of the columns in the dataset is the date column which is registered as object, when imported into the program. To change the data type, we can use 'astype' in most cases, but in this case, we have to utilize the pd.to_datetime function to achieve this convrsion.
- Df.nuninque()→ This is used to identify the number of unique values available in the dataset which helps us understand the number of different values within the column and if there are

any duplicate values (The duplicate value count can also be achieved by doing the df.duplicated() functions)

- df.Country.value_counts().reset_index()→ Getting a quick breakdown of the countries and the no.of.purchases in each country from the dataset.
- df.loc[df['Quantity'] >0]→This is used here to remove the redundancy of the dataset

Choosing the Algorithm for the Project:

Algorithms must be chosen depending on our problem and here we are going to segmented online retail customers or we should say we are going to segment Market on the basis of spending amount, frequency of visit,behaviour , age, gender etc. We are choosing the KMeans Clustering algorithm for this project.

K-Means Clustering

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms.

The algorithm works as follows:

- First we initialize k points, called means, randomly.
- We categorize each item to its closest mean and we update the mean's coordinates, which are the averages of the items categorized in that mean so far.
- We repeat the process for a given number of iterations and at the end, we have our clusters.

Motivation and Reasons for Choosing the Algorithm:

KMeans clustering algorithm is an unsupervised learning algorithm, which has a wide application in the segmentation, as this algorithm allows the analyst to review and segment the data into different segments based on the provided attributes.

This helps the companies to know Which group of customers are loyal, Which group can spend more money, Which group visit them infrequently, Which group of customers they are loosing. Through this companies tries to target the sub groups of customers in retaining them on the basis of their needs and desires by executing various marketing campaigns such as providing special offers, discounts etc.

Assumptions:

In this Online Retail Customer Segmentation or Market Segmentation project, by data pre-processing and KMeans Clustering technique ,we will divide the whole data of customers on the basis of RMF i.e. Recency, Monetary and Frequency and we will also visualize these groups on the basis of these terms.

What is Customer Segmentation or Market Segmentation:

We can say that Customer Segmentation or Market Segmentation is methodology or marketing practice through which we divide our customer group into various similar sub groups such as on the basis of spending amount, frequency of visit,behaviour , age, gender,e.t.c. This helps the companies

to know Which group of customers are loyal, Which group can spend more money, Which group visit them infrequently, Which group of customers they are loosing.

Through this companies tries to target the sub groups of customers in retaining them on the basis of their needs and desires by executing various marketing campaigns such as providing special offers, discounts etc.

What is RMF:

RMF is a simple statistical method for categorising customers based on their purchasing behaviour. The behaviour is identified by using only three customer data points: the *recency* of purchase (R), the *frequency* of purchases (F) and the mean *monetary value* of each purchase (M). After some calculations on the RFM data we can create customer segments that are actionable and easy to understand or we should say RFM segmentation will make you able to understand your customer base better.

Recency - It represents how recently a customer purchased a product.

Frequency - It represents how often a customer purchased a product. The more frequent will be the better score.

Monetary - It represents how much a customer spends.

steps involved in building the model:

- monetary =df.groupby("CustomerID").Sales.sum() used for calculating monetary value by grouping customer with their customer id and total no. of sales.
- frequency=df.groupby("CustomerID").InvoiceNo.count() used for calculating frequency value by count the invoice number grouping with respective customer id.
- Recency will get by deciding the reference date and it will define the reference date as one day before the last transaction.

```
ReferenceDate = max(df.InvoiceDate)
ReferenceDate = ReferenceDate + pd.DateOffset(days=1)
df["LastPurchaseDate"] = ReferenceDate - df.InvoiceDate
```

recency = df.groupby("CustomerID").LastPurchaseDate.min() used for calculating recency to find minimum Last purchase date groupby their respective customer id.

- Now combining all to make a RMF model

```
rmf = monetary.merge(frequency, on = "CustomerID")
rmf = rmf.merge(recency, on = "CustomerID")
rmf.columns = ["CustomerID", "Monetary", "Frequence", "Recency"]
RMF = rmf.drop("CustomerID",axis =1)
```

RMF Model is ready now for analysis .

Model Evaluation and Technique:

In this project we used KMean algorithm for model building and for evaluation we are using Elbow method to find the no. of clustering groups.And after performing this algo we came to know that

optimal number of cluster for this dataset here is 5. On this basis we can make model and will be able to make prediction and inferences of that and visualize it.

Inferences & Recommendation:

From the above bar chart and pie chart, we can easily understand our 5 groups according to Recency mean,Frequency mean and Monetary mean.

- Group 4 is the group of customers who spends maximufrm amount of money and also has a good frequency and low recency rate.
- Group 1 are the customers whose frequency rate is maximum and monetary value is also good and recency rate is also quite good.
- whereas Group 3 is the group of customers who has maximum recency rate means they have not purchased anything from the past,has low frequency also and didn't spend money(very low monetary rate)

Based on these results indicate that cluster 4 is a high value customers than cluster 1 customers taking place.Cluster 2 is middle value customers, and Cluster 0 and then cluster 3 are low value customers.

The smart marketing team understand the importance of ‘knowing your customer’ and using this successful customer segmented technique advertisers understand customer’s purchasing behaviour and engage them with various strategy.

*The program ipynb file is attached to this file is submitted with.

--END--