Online Retail Customer Segmentation

Submitted By:

Divya Kumari

Institute: Intellipaat

Date:25th May 2023

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.

Dataset Information:

The online\_retail.csv contains 387961 rows and 8 columns.

Feature Name Description

Invoice Invoice number

StockCode Product ID

Description Product Description

Quantity Quantity of the product

InvoiceDate Date of the invoice

Price Price of the product per unit

CustomerID Customer ID

Country Region of Purchase

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

2. Segment the customers based on their purchasing behavior

Introduction:

In this article we are going to made a project on **Online Retail Customer Segmentation or Market Segmentation** in Python 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 3 terms.

But Before jumping in making the Project directly first know some basic terms-

**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 :-

1. Which group of customers are loyal.
2. Which group can spend more money.
3. Which group visit them infrequency.
4. 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, e.t.c.

In this project we are dividing our customers on the basis of 3 factors

1. **Recency**:- It represents how recently a customer purchased a product.
2. **Frequency**:- It represents how often a customer purchased a product. The more frequent will be the better score.
3. **Monetary**:- It represents how much an customer spends.

Now Lets start building our project

## **Supervised Machine Learning:**

Supervised learning is a machine learning method in which models are trained using labeled data. In supervised learning, models need to find the mapping function to map the input variable (X) with the output variable (Y).

Supervised Machine learning

Supervised learning needs supervision to train the model, which is similar to as a student learns things in the presence of a teacher. Supervised learning can be used for two types of problems: **Classification** and **Regression**.

**Learn more** [Supervised Machine Learning](https://www.javatpoint.com/supervised-machine-learning)

**Example:** Suppose we have an image of different types of fruits. The task of our supervised learning model is to identify the fruits and classify them accordingly. So to identify the image in supervised learning, we will give the input data as well as output for that, which means we will train the model by the shape, size, color, and taste of each fruit. Once the training is completed, we will test the model by giving the new set of fruit. The model will identify the fruit and predict the output using a suitable algorithm.

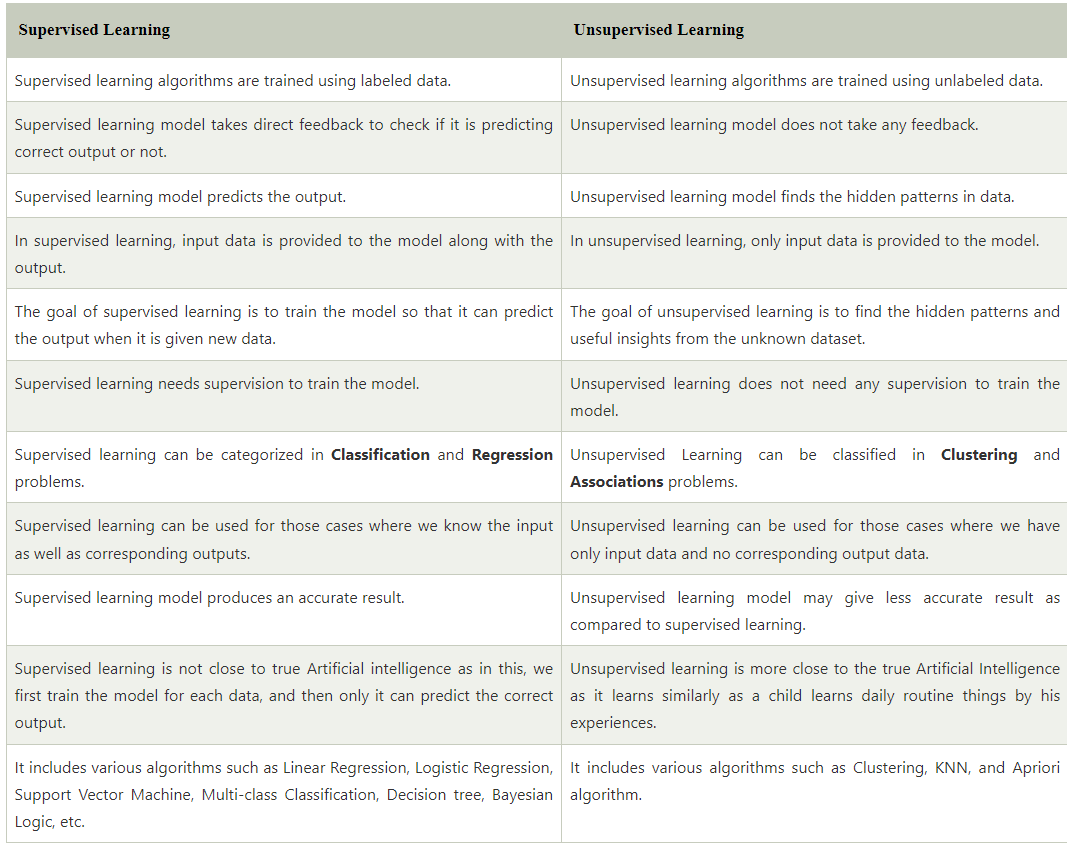
## **Unsupervised Machine Learning:**

Unsupervised learning is another machine learning method in which patterns inferred from the unlabeled input data. The goal of unsupervised learning is to find the structure and patterns from the input data. Unsupervised learning does not need any supervision. Instead, it finds patterns from the data by its own.

**Learn more** [Unsupervised Machine Learning](https://www.javatpoint.com/unsupervised-machine-learning)

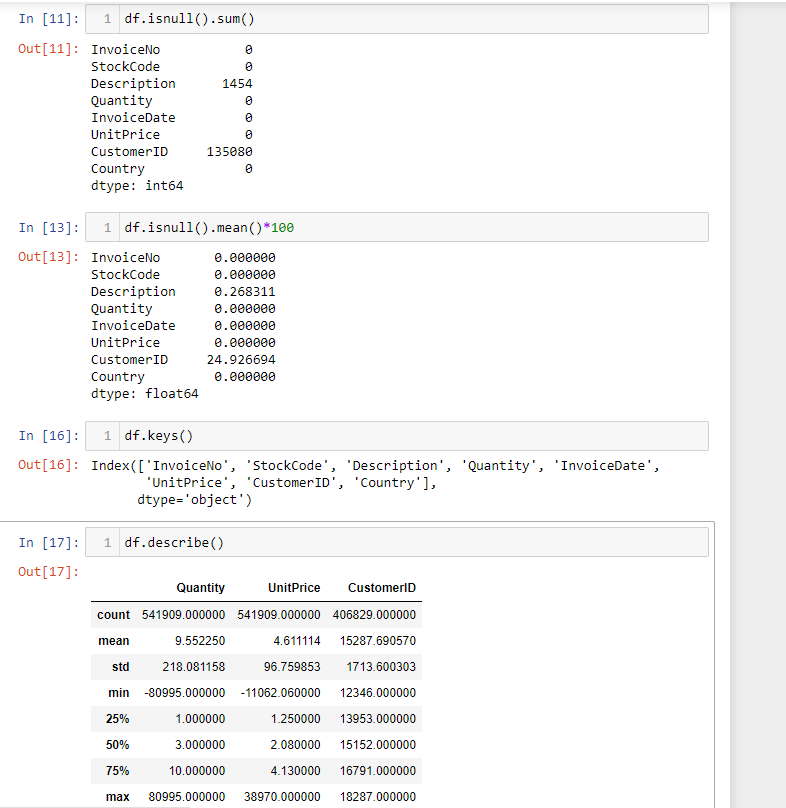
Unsupervised learning can be used for two types of problems: **Clustering** and **Association**.

**Example:** To understand the unsupervised learning, we will use the example given above. So unlike supervised learning, here we will not provide any supervision to the model. We will just provide the input dataset to the model and allow the model to find the patterns from the data. With the help of a suitable algorithm, the model will train itself and divide the fruits into different groups according to the most similar features between them.



Hence the problem statement which we have been given is of unsupervised learning.

So will use Kmeans algorithm to get the behaviour of customer and will use python lib like pandas, numpy, matplotlib to extract the useful insights.

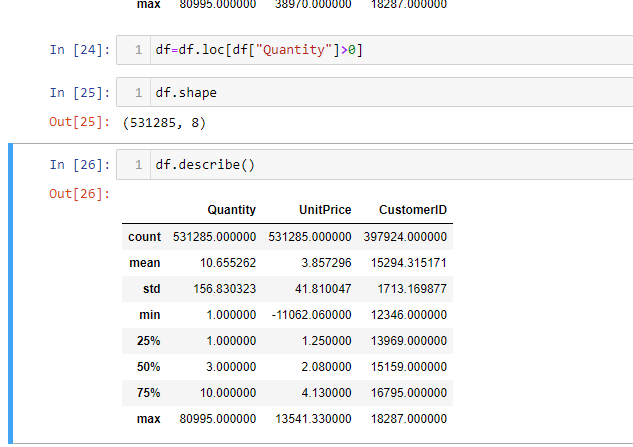


Here we have null values in two columns Description and customer\_id.

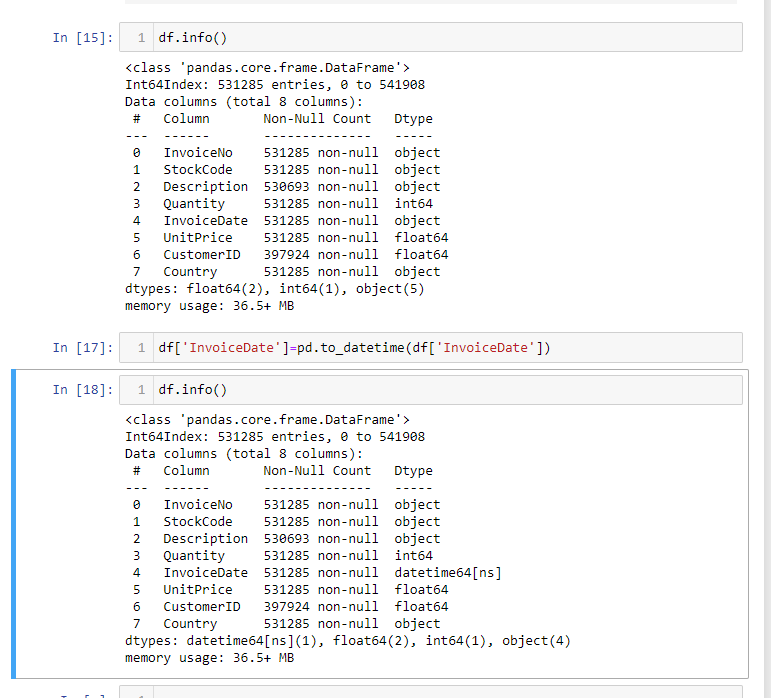
By using the df.describe() function we came to know that our minimum quantity is in negative and we all know that a quantity will never be in Negative. So, we have to remove this redundancy in order to get better accuracy because redundancy can cause miss grouping of data.

Data Pre-processing:

#removing the redundancy



Now here we can see that Invoice date is object type now we have to convert this into datetime for calculating all the values.



from here we now calculate our **Monetary Value**

#calculating our monetary value

df["Sale"] =df.Quantity \* df.UnitPrice

#created a column of sale

df.head()

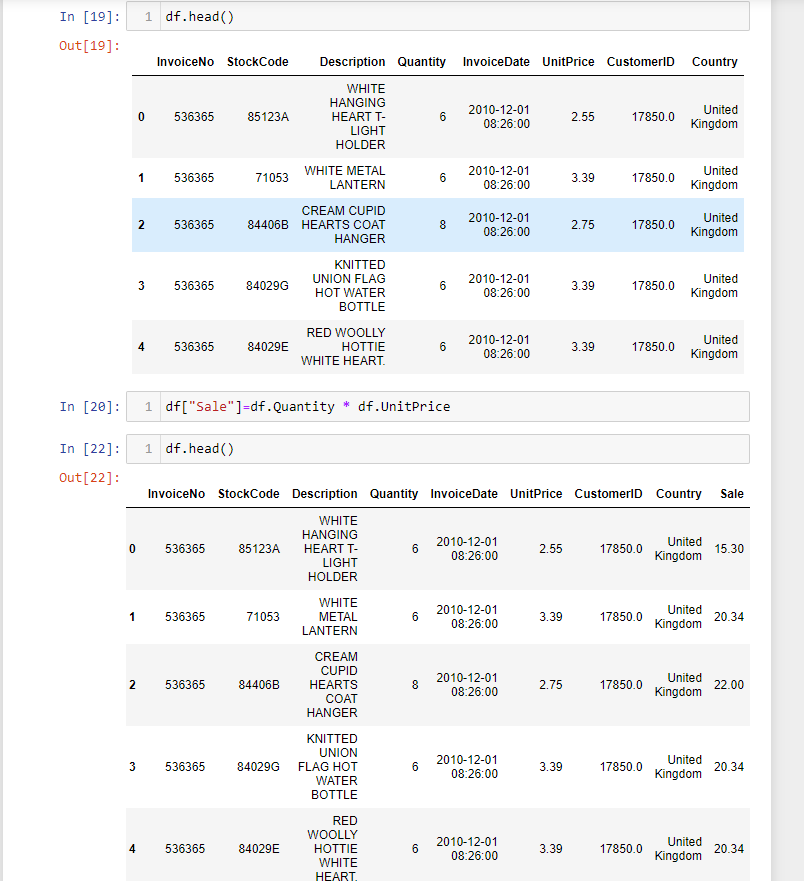
monetary =df.groupby("CustomerID").Sale.sum().reset\_index()

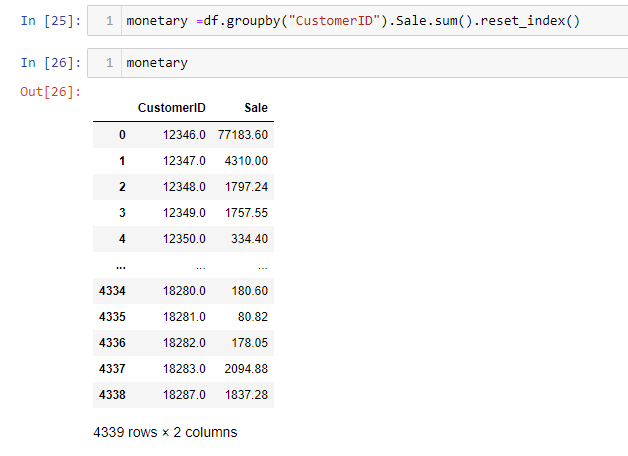
"""

Here we are getting our monetary value by grouping customer with their customer id and total no. of sales.

"""

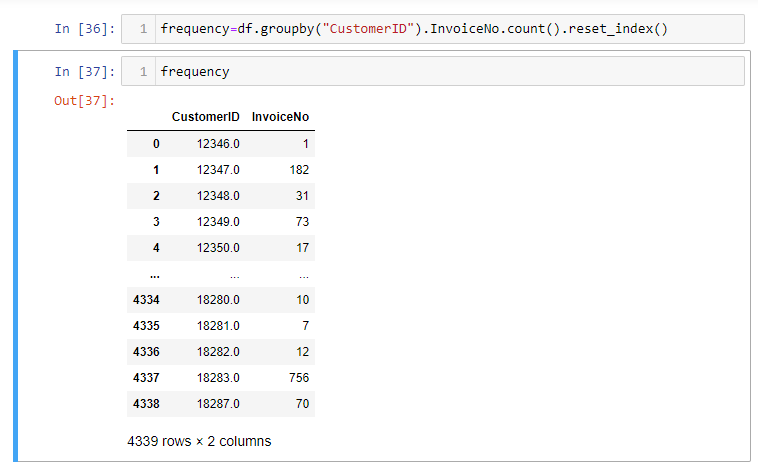
#resetting our index,our monetary has multiindex so we are removing it





Now we will calculate **frequency**of our dataset:

This will tell us, How frequent a customer is Purchasing products.



Now we will calculate our **recency**value

#calculating our recency value

LastDate=max(df.InvoiceDate) #calculating the last date of InvoiceDate

LastDate

LastDate = LastDate + pd.DateOffset(days=1)

#adding one to LastDate

LastDate

df["Diff"] = LastDate - df.InvoiceDate

#Diff is the difference between our Lastate and InvoiceData

recency = df.groupby("CustomerID").Diff.min()

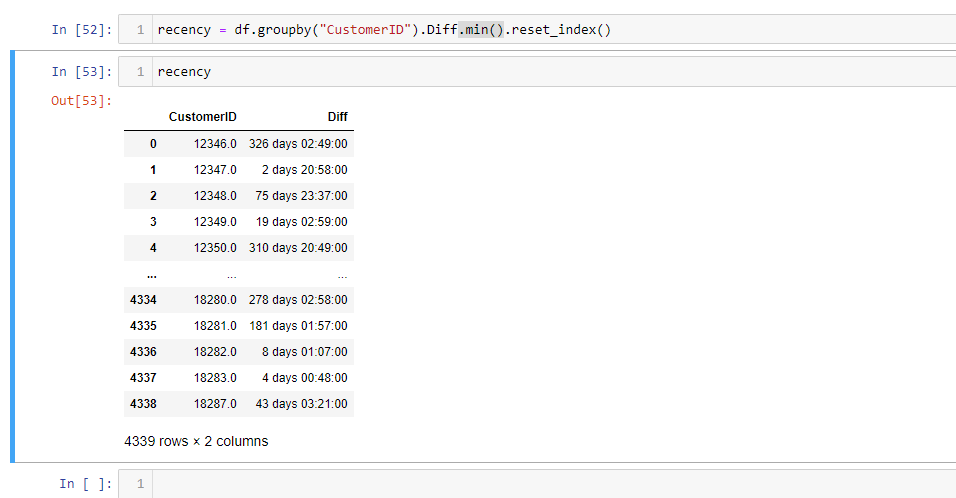
"""

here we get our recency value using group by

"""

recency = recency.reset\_index()





Combining all dataframes:

rmf = monetary.merge(frequency, on = "CustomerID")

rmf = rmf.merge(recency, on = "CustomerID")

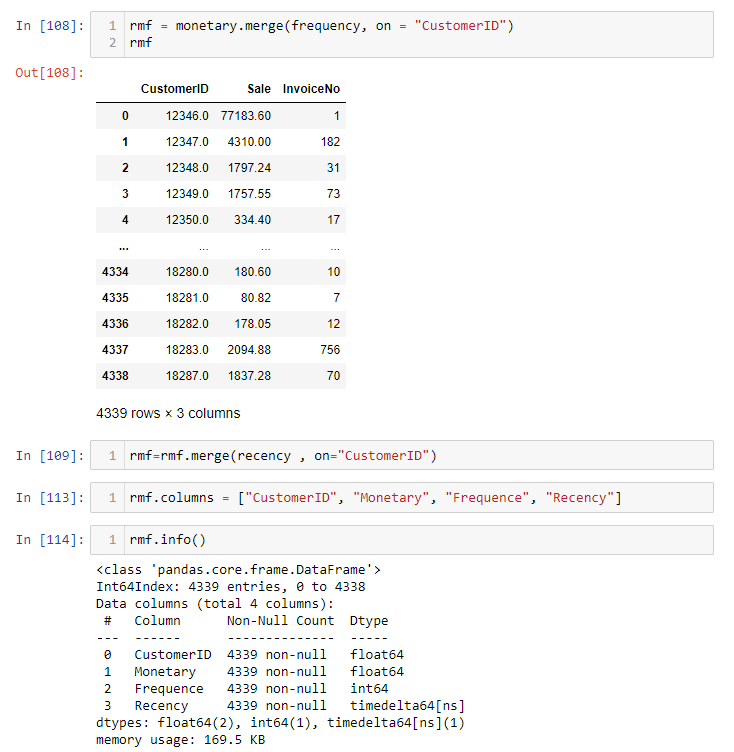
rmf.columns = ["CustomerID", "Monetary", "Frequence", "Recency"]

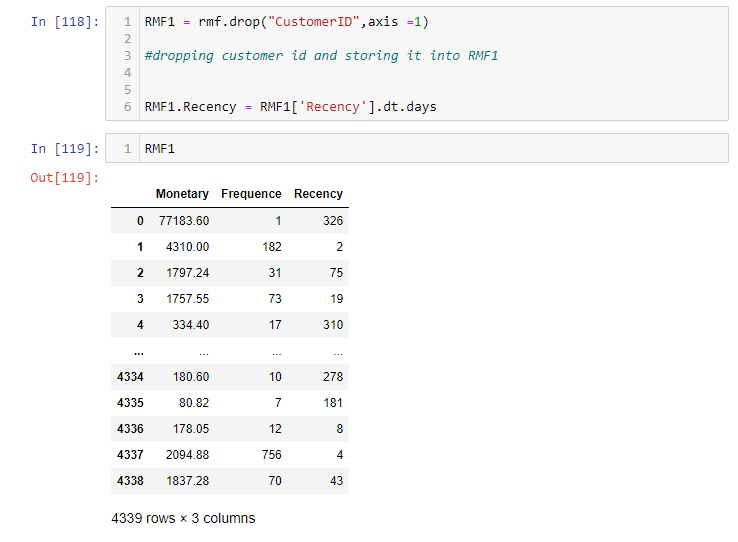
rmf

RMF1 = rmf.drop("CustomerID",axis =1)

#dropping customer id and storing it into RMF1

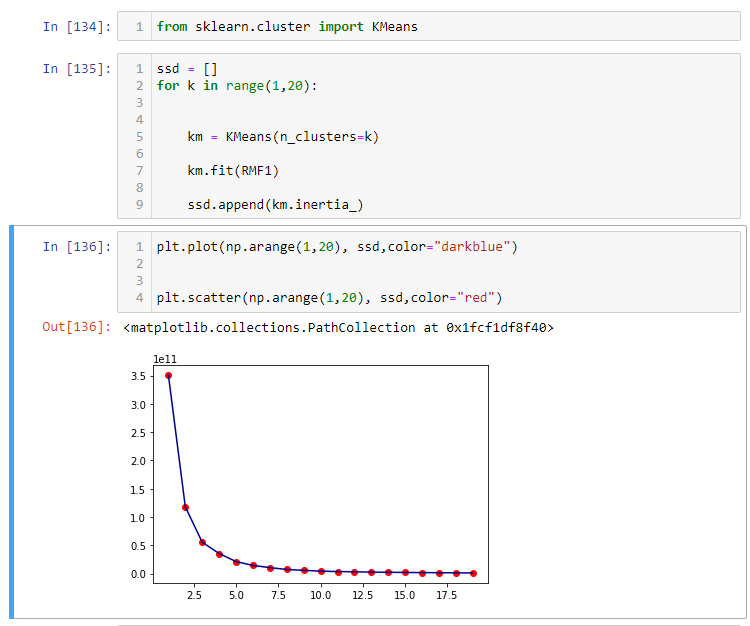
RMF1.Recency = RMF1.Recency.dt.days





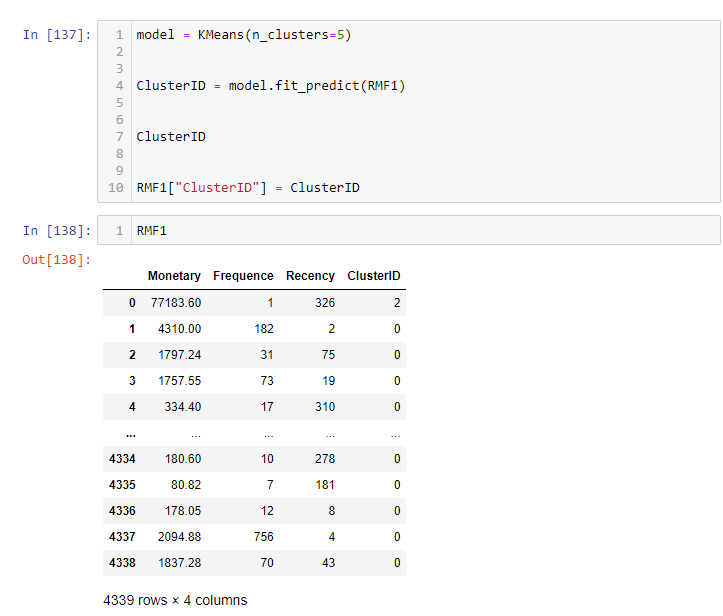
Our Data pre-processing part ends here now we will perform the analysis of or data.

we will perform analysis of data using KMeans algorithm.

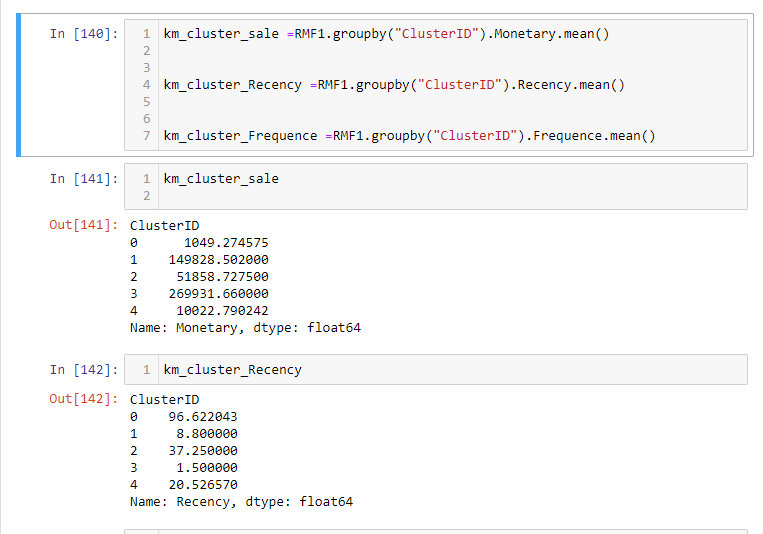


In the KMean algo we are using elbow method to find the no. of clustering groups.

We will perform clustering now onwards



Visualisation:







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

Group 1 is the group of customer who spends maximum amount of money and also has a good frequency and low recency rate.Group 4 are the customers whose frequency rate is maximum and monetary value is also good and recency rate is also quite good, whereas Group 0 is the group of customers who has a very high recency rate means they have not purchased anything from the past.

Thank you!!