International Journal of Advance and Innovative Research
Volume 9, Issue 4 October - December 2022
ISSN 2394 - 7780

RFM TECHNIQUE FOR CUSTOMER SEGMENTATION: REALIZING THROUGH PYTHON CODE

Naresh Chandra[1], Dr. Arvind Kumar Shukla[2] and Subhash Chand[3]
[1,3]Research Scholar, IFTM University, Moradabad, India-244001
[2]Associate Professor, IFTM, University, Moradabad, India-244001

1. ABSTRACT

Customer segmentation plays very vital role for taking the decisions of optimising the Return on Investment (RoI) of any business. On the basis of customer segmentation an e-commerce business company frames their strategy to make of the most profit according to those segments. Those customers who are recognized as a high-value and frequent purchasers can be targeted with loyalty programs or special discounts. RFM analysis-based customer segmentation is an inordinate way to targeting the marketing.

In RFM analysis, a score for recency, frequency, and monetary value is assigned to each customer, and then a final RFM score is evaluated.

Most recent purchase is the criterion for Recency score, frequency score is based upon how many numbers of times the customers purchased. Higher score reflects the higher frequency.

Finally, an amount spent by a customer on the purchase is considered as a monetary and assigned a monetary score. Combining all these three scores, a final RFM score is calculated.

In this paper, analysis and customer segmentation are based upon a UK based e-commerce retailer companies' online transaction data from 01.12.2009 to 09.12.2011.

Keywords: RFM, Customer segmentation, k-means, Python, e-commerce

2. INTRODUCTION

Recency Frequency Monetary (RFM) model is the most widely used behaviour segmentation. All customers are presented by 555, 554, 553, ...,112, 111. The most beneficial customer group is assigned a value 555, whereas the worst customer group is assigned a value 111.

Due to fewer segmentation variables, this model is used extensively. Also, it is easy and simple to implement, and straightforward to understand for decision makers.

RFM model was proposed by Arthur M. Hughes. RFM model is based on the most common marketing axiom, the Pareto principle, which states that **"80% of your business comes from 20% of your customers".**

Hughes (1994) presented that the importance (weight) of the three variables R, F and M is equal while Stone (1995) treated different weights for the RFM variables. The weight of each RFM variable depends on the characteristics of the industry.

K-means clustering algorithm is extensively used algorithm in CRM and marketing. This algorithm introduced by MacQueen (1967), can process large amounts of data quickly.

The primary objective of the study was to implement a segmentation strategy to identify different customer's groups with similar purchase patterns shown by customers. RFM-based customer segmentation technique is utilized for segmentation through the study using the python codes. The data/figures shown in tables in this paper are the output of the python code used in analysis of the study.

3. DATASET

The Online Retail data set includes the sales of an UK based online retail store of the period from 1/12/2009 to 09/12/2011 freely available on https://www.kaggle.com/. The "Online Retail" dataset is characterized by the following 08 attributes:

Invoice No: Invoice number is a unique number for every transaction occurred. Invoice number starts with C is a cancelled operation.

StockCode: Product code is a unique number for every product exist in store.

Description: Product name.

Quantity: Number of the products in the invoices have been sold is referred by Quantity.

InvoiceDate: Ttransaction's Invoice date.

UnitPrice: Product price.

CustomerID: Unique customer number.

Country: The name of the country where the customer lives.

4. Data Preparation and Pre-Processing

First ten rows of the data using "head" function:

Table-01

	Invoice	StockCode	...	Customer ID	Country
0	536365	85123A	...	17850.0	United Kingdom
1	536365	71053	...	17850.0	United Kingdom
2	536365	84406B	...	17850.0	United Kingdom
3	536365	84029G	...	17850.0	United Kingdom
4	536365	84029E	...	17850.0	United Kingdom
5	536365	22752	...	17850.0	United Kingdom
6	536365	21730	...	17850.0	United Kingdom
7	536366	22633	...	17850.0	United Kingdom
8	536366	22632	...	17850.0	United Kingdom
9	536368	22960	...	13047.0	United Kingdom

Shape function tell us that there are 54190 rows and 08 columns in the dataset. Describe function tell us some basic statistics as count, mean, std, min, 25%, 50%, 75% and max, shown in following table-03:

Table-03

	Quantity	Price	Customer ID
Count	541910.000000	541910.000000	406830.000000
Mean	9.552234	4.611138	15287.684160
Std	218.080957	96.759765	1713.603074
Min	-80995.000000	-11062.060000	12346.000000
25%	1.000000	1.250000	13953.000000
50%	3.000000	2.080000	15152.000000
75%	10.000000	4.130000	16791.000000
Max	80995.000000	38970.000000	18287.000000

To check the missing observation of the dataset using the "is null function", shown in following table:

Table-04

Invoice	0
StockCode	0
Description	1454
Quantity	0
InvoiceDate	0
Price	0
Customer ID	135080
Country	0

First, we clean dataset by removing the missing observations from the dataset, using the function "dropna". Now we have 406830 rows in dataset. By the function nunique we came to know that there are 3896 unique items available in the dataset. Item wise in descending order are shown as follows, using the [Description"].Value-counts ()

Table-05

White Hanging Heart T-Light Holder	2070
Regency Cakestand 3 Tier	1905
Jumbo Bag Red Retrospot	1662
Assorted Colour Bird Ornament	1418
Party Bunting	1416
Antique Raspberry Flower Earrings	1
Wall Art,Only One Person	1

Gold/Amber Drop Earrings W Leaf	1
Incense Bazaar Peach	1
Pink Baroque Flock Candle Holder	1

Descending order of items sold quantity wise are shown as:

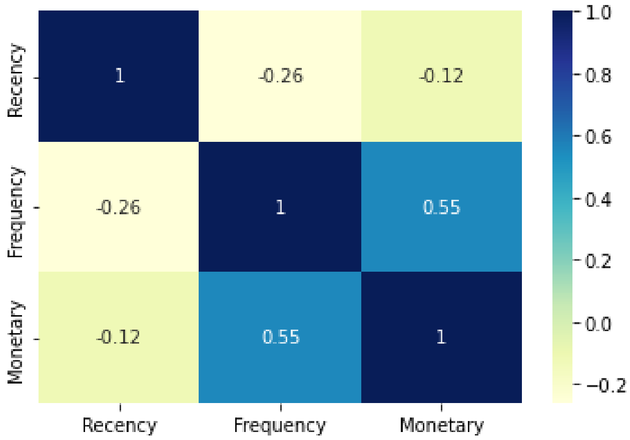
Table-06	
Description	**Quantity**
World War 2 Gliders Asstd Designs	53215
Jumbo Bag Red Retrospot	45066
Assorted Colour Bird Ornament	35314
White Hanging Heart T-Light Holder	34147
Pack Of 72 Retrospot Cake Cases	33409
Popcorn Holder	30504
Rabbit Night Light	27094
Mini Paint Set Vintage	25880
Pack Of 12 London Tissues	25321
Pack Of 60 Pink Paisley Cake Cases	24163

Removing the rows from dataset for those which transactions are cancelled.Now we have remains 397925 rows. Also removing the rows for which columns "Quantity" and "Price" have the -ve values. Now we have remains 397885 rows in dataset.

We have created a new columns "Total Price" for the purpose of monetary value, by multiply the "Price and Quantity". Now we have 397885 rows and 09 columns. 09 columns.

Calculating of RFM Metrics
Since maximum date of transaction is 09-12-2011, therefore, we calculate recency from 10-12-2011. Correlation and heatmap of the dataset are shown in following figure-01 and table-09.

Figure-01



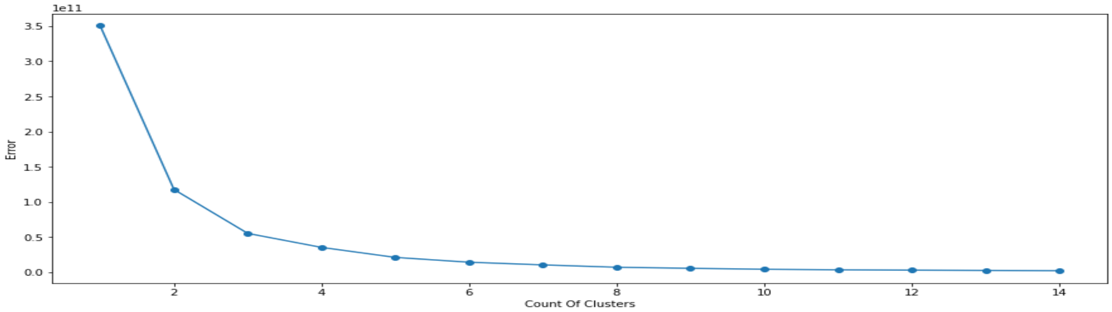
Capturing the cluster labels and cluster centroids, we get the following result.

Table-09		
	Num_Clusters	**Cluster_Errors**
0	1	3.505004e+11
1	2	1.176336e+11
2	3	5.512823e+10
3	4	3.522170e+10
4	5	2.098160e+10
5	6	1.412109e+10
6	7	1.031624e+10
7	8	6.970139e+09
8	9	5.450355e+09
9	10	4.086092e+09

 ISSN 2394 - 7780

Plotting the graphs between error and count of cluster shown in following figure-02.

Figure-02



Calculating silhouette scores for 02 to 15 cluster we get the following values:

For 2 Clusters, The silhouette score is 0.9844890477261269

For 3 Clusters, The silhouette score is 0.9579526616883844

For 4 Clusters, The silhouette score is 0.9543151351092204

For 5 Clusters, The silhouette score is 0.8373987537332015

For 6 Clusters, The silhouette score is 0.7750376900030161

For 7 Clusters, The silhouette score is 0.73332141890852

For 8 Clusters, The silhouette score is 0.7321294945237423

For 9 Clusters, The silhouette score is 0.6735679186569941

For 10 Clusters, The silhouette score is 0.673992521720467

For 11 Clusters, The silhouette score is 0.6359349509744601

For 12 Clusters, The silhouette score is 0.6264706845195438

For 13 Clusters, The silhouette score is 0.6180098735304097

For 14 Clusters, The silhouette score is 0.6106586300038647

For 15 Clusters, The silhouette score is 0.6115826706947971

Using no of clusters as 04 we get the centroid as and the cluster centers as the tabel.

[2 0 0 ... 0 0 0]

Table-10		
[[9.26465116e+01	3.89488372e+00	1.43819457e+03]
[8.40000000e+00	6.50000000e+01	1.49828502e+05]
[3.00322581e+01	4.27741935e+01	4.63930139e+04]
[5.00000000e-01	6.65000000e+01	2.69931660e+05]]

First 10 rows of Clusterwise Recency, frequency and monetary are shown as:

Table-11

Customer ID	Recency	Frequency	Monetary	Clusters
12347.0	2	7	4310.00	0
12348.0	75	4	1797.24	0
12349.0	18	1	1757.55	0
12350.0	310	1	334.40	0
12352.0	36	8	2506.04	0
12353.0	204	1	89.00	0
12354.0	232	1	1079.40	0
12355.0	214	1	459.40	0
12356.0	22	3	2811.43	0

Clusterwise number of counts are as follows:

Table-12	
0	4300
2	31
1	5
3	2

Basic Statistics of Recency, Frequency and Monetary are as:

Table-13

	Count	Mean	Std	Min	25%	50%	75%	Max
Recency	4338.0	92.059474	100.012264	0.00	17.000	50.000	141.75	373.00
Frequency	4338.0	4.272015	7.697998	7.697998	1.000	2.000	5.00	209.00
Monetary	4338.0	2054.270609	8989.229895	3.75	307.415	674.485	1661.74	280206.02

On the scale of 01 to 05, we assign a value for each transaction to Recency, Frequency and Monetary like below.

Converting RFM Scores to Single Variable

On the basis of these scale, we get the recency score, frequency score and monitory score as in the following table:

Table-14

Customer ID	Recency	Frequency	Monetary	Recency_Score	Frequency_Score	Monetary_Score
12346.0	325	1	77183.60	1	1	5
12347.0	2	7	4310.00	5	5	5
12348.0	75	4	1797.24	2	4	4
12349.0	18	1	1757.55	4	1	4
12350.0	310	1	334.40	1	1	2
12352.0	36	8	2506.04	3	5	5
12353.0	204	1	89.00	1	1	1
12354.0	232	1	1079.40	1	1	4
12355.0	214	1	459.40	1	1	2
12356.0	22	3	2811.43	4	3	5

Combining the value of Recency. Frequency and Monetary to form a single value of RFM Score.

Table-15

Customer ID	
12346.0	115
12347.0	555
12348.0	244
12349.0	414
12350.0	112
12352.0	355
12353.0	111
12354.0	114
12355.0	112
12356.0	435

On the basis of descending order of monetary we have first ten rows as:

Table-16

Customer ID	Recency	Frequency	Monetary	Recency_ Score	Frequency_ Score	Monetary_ Score	RFM_ SCORE
14646.0	1	73	280206.02	5	5	5	555
18102	0	60	25965	5	5	5	555

.0			7.30				
17450.0	8	46	194550.79	5	5	5	555
14911.0	1	201	143825.06	5	5	5	555
14156.0	9	55	117379.63	5	5	5	555
17511.0	2	31	91062.38	5	5	5	555
16684.0	4	28	66653.56	5	5	5	555
14096.0	4	17	65164.79	5	5	5	555
13694.0	3	50	65039.62	5	5	5	555
15311.0	0	91	60767.90	5	5	5	555

Segmenting Customers Using RFM Score

Now we segment the customers data using RFM scores. First 10 rows looks line as:

Table-17	
Customer ID	
12346.0	7
12347.0	15
12348.0	10
12349.0	9
12350.0	4
12352.0	13
12353.0	3
12354.0	6
12355.0	4
12356.0	12

Now we categorise the Unbeaten, Champs, Trustworthy, Prospective, Optimistic, Needs Heedfulness and Require Stimulating segment as per table-17 below, the first 10 rows looks like as in table-18.

Table-17	
Rfm_Score_S >= 9	Unbeaten
Rfm_Score_S >= 8<9	Champs
Rfm_Score_S >= 7<8	Trustworthy
Rfm_Score_S >= 6<7	Prospective
Rfm_Score_S >= 5<6	Optimistic
Rfm_Score_S >= 4<5	Needs Heedfulness
else	Require Stimulating

Table-18

Customer Id	Recency	Frequency	Monetary	Recency_Score	Frequency_Score	Monetary_Score	Rfm_Score	Rfm_Score_S	Rfm_Level
12346.0	325	1	77183.60	1	1	5	115	7	Trustworthy
12347.0	2	7	4310.00	5	5	5	555	15	Unbeaten
12348.0	75	4	1797.24	2	4	4	244	10	Unbeaten
12349.0	18	1	1757.55	4	1	4	414	9	Unbeaten
12350.0	310	1	334.40	1	1	2	112	4	Needs Heedfulness

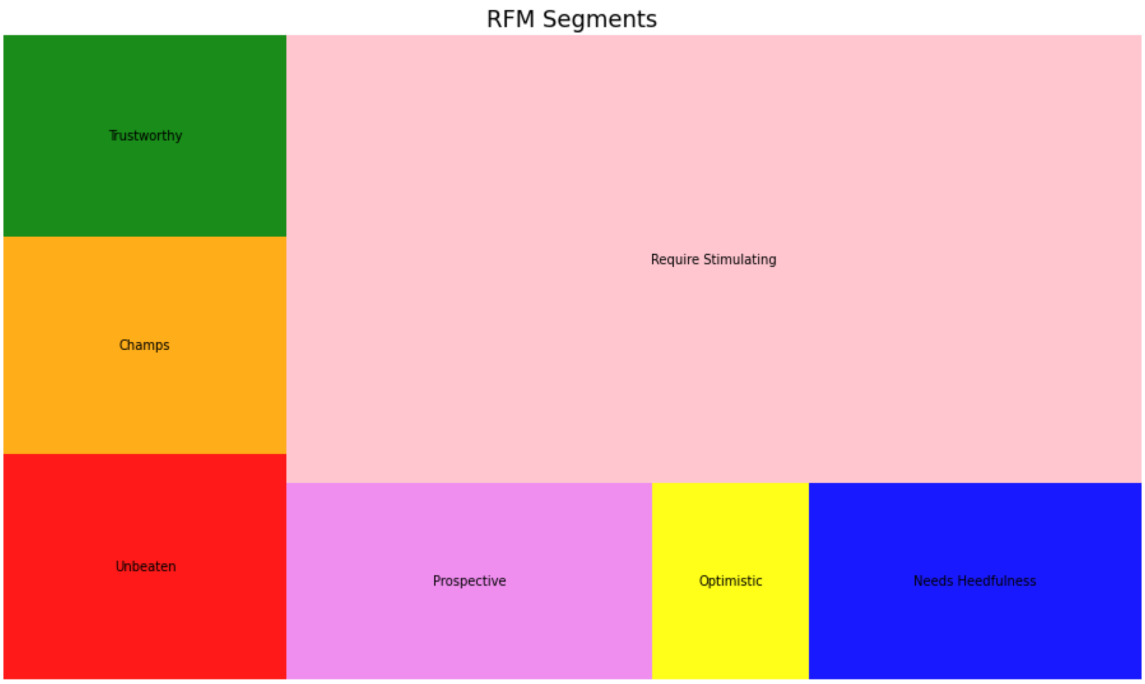
12352.0	36	8	2506.04	3	5	5	355	13	Unbeaten
12353.0	204	1	89.00	1	1	1	111	3	Require Stimulating
12354.0	232	1	1079.40	1	1	4	114	6	Potential
12355.0	214	1	459.40	1	1	2	112	4	Needs Heedfulness
12356.0	22	3	2811.43	4	3	5	435	12	Unbeaten

Now we calculate the average values for each RFM-Level and categorical data can be represented/visualised via tree map.

Table-19

	Recency			**Frequency**			**Monetary**		
	Mean	**Count**	**Max**	**Mean**	**Count**	**Max**	**Mean**	**Count**	**Max**
RFM_Level									
Champs	85.1	376	313	2	376	6	676.9	376	9864.3
Needs Heedfulness	237.2	364	373	1	364	2	216.3	364	487.8
Optimistic	176.1	339	373	1.1	339	2	291.2	339	922.1
Prospective	122.8	422	373	1.3	422	3	383.4	422	1784.7
Require Stimulating	287.3	182	373	1	182	1	144.3	182	250
Trustworthy	97.2	384	372	1.6	384	5	705.4	384	77183.6
Unbeaten	35.2	2271	372	6.9	2271	209	3531.7	2271	280206

In case of large data, clean & informative insight from the data can be visualised using treemap. For treemap we have to install squarify, and using squarify we get the following graph:



5. CONCLUSION

The major goal of this study was to use the RFM model to segment customers from a total of 54,190 online transaction occurred from 1/12/2009 to 09/12/2011 at a UK based retailer. Customers are segmented as Unbeaten, champs, Trustworthy, Prospective, Optimistic, Need Heedfulness and Require Stimulating as per squarify graph above. Company can make the best and different marketing strategy as per this segmentation of the customers.

6. REFERENCES

[1] E Ernawati1 et al., “A review of data mining methods in RFM-based customer Segmentation”, Journal of Physics:Conference Series, 2021

[2] Nesma mahmoud Taher et al., “Investigation in Customer Value Segmentation Quality under Different Preprocessing Types of RFM Attributes”, IJES, Issue-4, 2016

[3] İnanç KABASAKAL et al., “Customer Segmentation Based on Recency Frequency Monetary Model: A Case Study in E-Retailing”, BILISIM TEKNOLOJIERI, CILT, OCAK, 2020

[4] Ina Maryani et al., “Customer Segmentation based on RFM model and Clustering Techniques With K-Means Algorithm”, Third International Conference on Informatics and Computing (ICIC), Oct, 2018

[5] Ms. Chaithra S et al., “Customer Segmentation using RFM analysis”, International Research Journal of Engineering and Technology (IRJET), eISSN:2395-0056, July 2021

[6] Dharmaiah Devarapalli et al., “Analysis of RFM Customer Segmentation Using Clustering Algorithms”, International Journal of Mechanical Engineering, ISSN:0974-5823, January, 2022

[7] Amin Parvaneha et al. “Integrating AHP and Data Mining for Effective Retailer Segmentation Based on Retailer Lifetime Value”, Journal of Optimization in Industrial Engineering, ISSN: 22519904, September, 2012

[8] Subhash Chand, A.K. Shukla and Naresh Chandra On-line Customers Buying Behaviour Prediction using XGBoost Algorithms in Python, Indian Journal of Natural Sciences (IJONS), ISSN 0976 – 0997, Issue 73, August, 2022

[9] Naresh Chandra, Arvind Kumar Shukla and Subhash Chand Customers’ Segmentation using RFM Model with k-means, Indian Journal of Natural Sciences (IJONS), ISSN 0976 – 0997, Issue 73, August, 2022

[10] Subhash Chand, A.K. Shukla and Naresh Chandra Prediction and Buying Behaviour of Customers using Machine Learning Technique, Indian Journal of Natural Sciences (IJONS), ISSN 0976 – 0997, Issue 70, Feb, 2022

[11] Naresh Chandra, Arvind Kumar Shukla and Subhash Chand Statistical Hypothesis Testing for Determining the Relationship of Categorical Variables using Python Code, Indian Journal of Natural Sciences (IJONS), ISSN 0976 – 0997, Issue 70, Feb, 2022

[12] https://gain-insights.com/resources/blogs/customer-segmentation-using-rfm-analysis/

[13] https://www.expressanalytics.com/blog/rfm-analysis-for-customer-segmentation/

[14] https://kaggle.com/