International Journal of Scientific Research in Science and Technology
Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)
doi : https://doi.org/10.32628/IJSRST2183118

Customer Segmentation using RFM Model and K-Means Clustering

Rahul Shirole, Laxmiputra Salokhe, Saraswati Jadhav

Department of Computer Engineering, Vishwakarma Institute of Technology Pune, Maharashtra, India

ABSTRACT

Article Info
Volume 8, Issue 3
Page Number : 591-597

Publication Issue
May-June-2021

Article History
Accepted : 03 June 2021
Published : 10 June 2021

Today as the competition among marketing companies, retail stores, banks to attract newer customers and maintain the old ones is in its peak, every company is trying to have the customer segmentation approach in order to have upper hand in competition. So Our project is based on such customer clustering method where we have collected, analyzed, processed and visualized the customer's data and build a data science model which will help in forming clusters or segments of customers using the k-means clustering algorithm and RFM model (Recency Frequency Monetary) for already existing customers. The input dataset we used is UK's E-commerce dataset from UCI repository for Machine Learning which is based on customer's purchasing behavioral. At the very simple the customer clusters would be like super customer, intermediate customers, customers on the verge of churning out based on RFM score .Along with this we also have created a web model where an e-commerce startup or e-commerce business analyst can analyze their own customers based on model we created .So using this it will be easy to target customers accordingly and achieve business strength by maintaining good relationship with the customers .

Keywords - RFM, Clustering, Silhouette Index.

I. INTRODUCTION

The focus of many companies is to provide the best product along with quality service to stand out in the market. But along with their quality service and better product they also have to make sure that customer don't slip out in search of alternatives because company exists cause of them and they create every company market share which generates revenue and profits for the company.

Every customer is different, here different in the sense we are talking about their age, location, psychology but the most potential parameter is it's purchasing behavior .Therefore, every segment of customer requires different product at different price. Therefore, every group of customer requires different marketing strategy, to develop different we first have to group similar customer in one segments.

The use of data mining techniques is one of the way to solve the problem of customer segmentation for developing new market strategy. Data mining is extremely powerful tool and has change many companies' fortune. Identification of customers is based on data collect from customers and grouping them in some meaningful order. Clustering [1][2][3][4][5]is a task of diving or grouping customer on the basis of their interaction with the company either direct or indirect. Here customer data can be anything like time spend on social media platform, transaction data, time spend on particular post but this paper focuses on the transaction data of a customer of UK online retail ecommerce . There are several other attributes in the dataset but one need to make a good selection of these attribute to get optimal results. To overcome this problem many of the data scientists prefer to use K-means algorithm which is a (unsupervised learning) along with RFM model[1][2]. RFM stands for recency, frequency and monetary values of a given customer. Now, the data that has been collected can be a base for segmentation of customers.

The aim of this paper is to find out the type of customer (super customer, intermediate customers, base customers) and to determine the value of customers so that companies can decide which class of customer generate healthy revenue and which do not, and also what new market strategy they can apply to improve their revenue growth.

II. LITERATURE SURVEY

There are some previous studies related to the segmentation of customer some of them are listed below:

In some pervious studies published in 2018 [1] and was published by students of Bina Sarana University. This paper is to segment customer based on their credit taken from the company called Nine reload. The algorithm used here is k-means based on RFM model. They form 2 Clusters from dataset of 82648 entries.

Other research was done in PD karya Mulya [2]a local wood product company. The segmentation was done based on the sales transaction the same company as the company grew a lot faster than expected. This also uses k-means based on RMF model and form cluster of 3. Clustering was done to identify the profit-making costumer and develop the marketing strategy accordingly.

In this research they have investigated the scope of the customer value based on current value, cross-selling probability and customer loyalty, this paper user uses neural network approach which uses Self Organization Map (SOM) to form clusters for banking use.[3]

This paper is to segment customer based on transaction from a supermarket which include 200 data entries. Here the clustering was done to help retail industry to develop new market strategy [4]. This paper user Hierarchical clustering algorithm which does not prerequire information about no of clusters required.

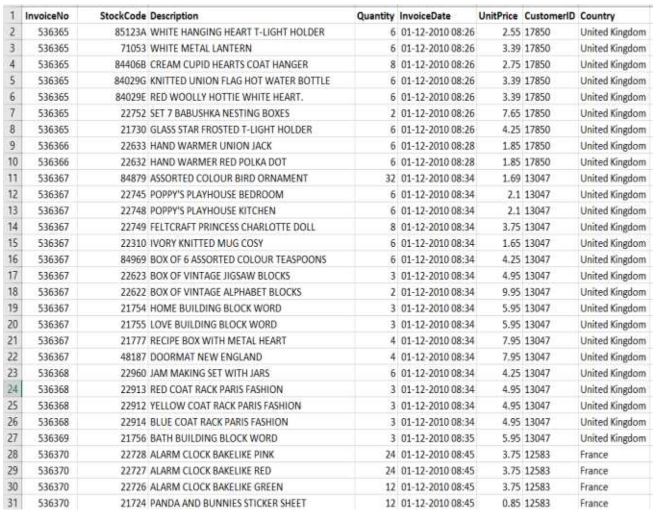
This paper user DBSCAN algorithm to segment customer. Which uses wholesale customers dataset of 440 entries to analyze the spending habit of customer. As we know K-means algorithm forms on well-spaced and circular shaped clusters, but in real life clusters can be arbitrary and k-means can give best possible clusters. So, the use of DBSCAN algorithm is made here.[5]

III. METHODOLOGY

STEP 1 : Business Understanding

Stages of business focus on understanding the purpose of needs based on business valuation. After understanding the business initial data mining plan is designed to reach the goal. The study of this paper is of an online retail E-commerce website of a UK retail. The transaction are of year 2010 and 2011.

Table 1. Sample Dataset



1	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
2	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	01-12-2010 08:26	2.55	17850	United Kingdom
3	536365	71053	WHITE METAL LANTERN	6	01-12-2010 08:26	3.39	17850	United Kingdom
4	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	01-12-2010 08:26	2.75	17850	United Kingdom
5	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	01-12-2010 08:26	3.39	17850	United Kingdom
6	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	01-12-2010 08:26	3.39	17850	United Kingdom
7	536365	22752	SET 7 BABUSHKA NESTING BOXES	2	01-12-2010 08:26	7.65	17850	United Kingdom
8	536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	01-12-2010 08:26	4.25	17850	United Kingdom
9	536366	22633	HAND WARMER UNION JACK	6	01-12-2010 08:28	1.85	17850	United Kingdom
10	536366	22632	HAND WARMER RED POLKA DOT	6	01-12-2010 08:28	1.85	17850	United Kingdom
11	536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	01-12-2010 08:34	1.69	13047	United Kingdom
12	536367	22745	POPPY'S PLAYHOUSE BEDROOM	6	01-12-2010 08:34	2.1	13047	United Kingdom
13	536367	22748	POPPY'S PLAYHOUSE KITCHEN	6	01-12-2010 08:34	2.1	13047	United Kingdom
14	536367	22749	FELTCRAFT PRINCESS CHARLOTTE DOLL	8	01-12-2010 08:34	3.75	13047	United Kingdom
15	536367	22310	IVORY KNITTED MUG COSY	6	01-12-2010 08:34	1.65	13047	United Kingdom
16	536367	84969	BOX OF 6 ASSORTED COLOUR TEASPOONS	6	01-12-2010 08:34	4.25	13047	United Kingdom
17	536367	22623	BOX OF VINTAGE JIGSAW BLOCKS	3	01-12-2010 08:34	4.95	13047	United Kingdom
18	536367	22622	BOX OF VINTAGE ALPHABET BLOCKS	2	01-12-2010 08:34	9.95	13047	United Kingdom
19	536367	21754	HOME BUILDING BLOCK WORD	3	01-12-2010 08:34	5.95	13047	United Kingdom
20	536367	21755	LOVE BUILDING BLOCK WORD	3	01-12-2010 08:34	5.95	13047	United Kingdom
21	536367	21777	RECIPE BOX WITH METAL HEART	4	01-12-2010 08:34	7.95	13047	United Kingdom
22	536367	48187	DOORMAT NEW ENGLAND	4	01-12-2010 08:34	7.95	13047	United Kingdom
23	536368	22960	JAM MAKING SET WITH JARS	6	01-12-2010 08:34	4.25	13047	United Kingdom
24	536368	22913	RED COAT RACK PARIS FASHION	3	01-12-2010 08:34	4.95	13047	United Kingdom
25	536368	22912	YELLOW COAT RACK PARIS FASHION	3	01-12-2010 08:34	4.95	13047	United Kingdom
26	536368	22914	BLUE COAT RACK PARIS FASHION	3	01-12-2010 08:34	4.95	13047	United Kingdom
27	536369	21756	BATH BUILDING BLOCK WORD	3	01-12-2010 08:35	5.95	13047	United Kingdom
28	536370	22728	ALARM CLOCK BAKELIKE PINK	24	01-12-2010 08:45	3.75	12583	France
29	536370	22727	ALARM CLOCK BAKELIKE RED	24	01-12-2010 08:45	3.75	12583	France
30	536370	22726	ALARM CLOCK BAKELIKE GREEN	12	01-12-2010 08:45	3.75	12583	France
31	536370	21724	PANDA AND BUNNIES STICKER SHEET	12	01-12-2010 08:45	0.85	12583	France

STEP 2 : Data Understanding

After the data is collected, we have be familiar with the data what actually the data. In this paper as said earlier we are using an online retail E-commerce website of a UK retail which consists of sales transaction from December 2010 to December 2011[1]. The dataset has 7 attributes which are listed below in table 1:

Table 1. Attributes during Data Understanding

No	Attribute	Description
1	Invoice No	Invoice No which is auto generated by Software, and Invoice no preceding C indicate Cancelled order
2	Stock Code	A Stock code definition an items of stock on stock quotation scheme. Made of Alphanumeric characters
3	Description	Description of item purchased by customer
4	Quantity	Quantity of purchased item on an single order
5	Invoice Date	Date and time of purchase of an item by Customer
6	Unit Price	Unit product price
7	Country	Country to which product is to be shipped

STEP 3 : Data Preparation

Data preparation consists of Data cleansing/preprocessing, Data visualization[2]. Removing errors, filling missing values, dropping negative transaction are all done in this phase of data science lifecycle.

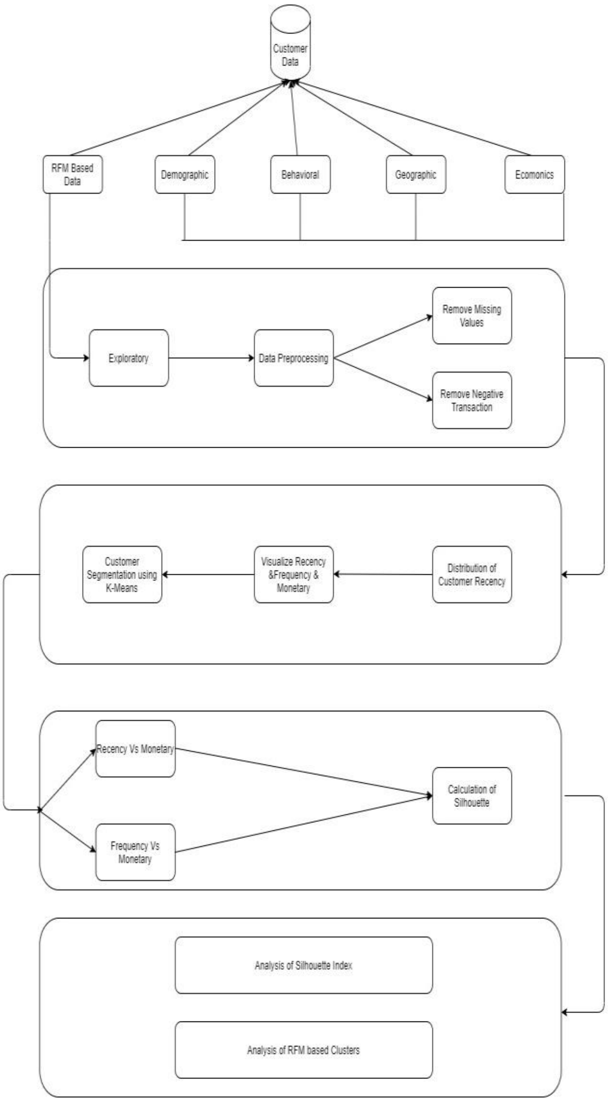

Customer Data
RFM Based Data
Demographic
Behavioral
Geographic
Ecomonics
Exploratory
Data Preprocessing
Remove Missing Values
Remove Negative Transaction
Customer Segmentation using K-Means
Visualize Recency &Frequency & Monetary
Distribution of Customer Recency
Recency Vs Monetary
Frequency Vs Monetary
Calculation of Silhouette
Analysis of Silhouette Index
Analysis of RFM based Clusters

Figure 1. Block Diagram

Data Pre-processing

- Generally In real world data is not proper like it may be incomplete, inconsistent and contains errors .Also if there is missing attributes, attribute values or contains duplicate and wrong data then it is called as unclean data.[3]

- So this behaviours degrade the quality of the result
- Hence we have to pre-process the data to use it more efficiently. We need to transform raw data into understandable format and this technique is called as Data Pre-processing..
- In Data Pre-processing we remove the null values , missing values and the outliers from given data to make it a clean data.[3]

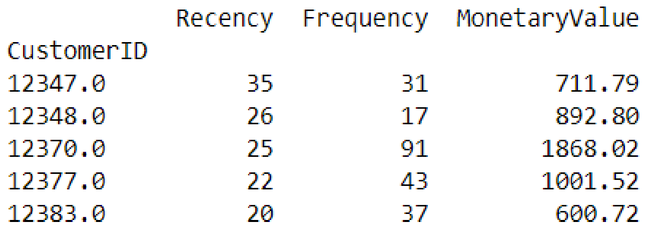
Data Visualization

- The graphical representation of data is Data Visualization. Visual elements such as graphs, maps and charts are used for data visualization.
- It provides us an accessible way to understand and analyze trends, patterns in data and outliers too.
- When we have such big data, data visualization technologies and tools are necessary to analyze huge amount of information. And thus, making data driven decisions.

RFM Model

- Recency, frequency and monetary values are analysis tools being used to identify any organisation's best customers[1][2]. RFM Model has 3 factors:
- Recency: How recently customer made a buy.
- Frequency: How often customer buys.
- Monetary: how much amount customer buys.
- Model ranks customers in each of these categories.

Table 3. Snapshot of Customer Transaction Data



```
            Recency  Frequency  MonetaryValue
CustomerID
12347.0          35         31         711.79
12348.0          26         17         892.80
12370.0          25         91        1868.02
12377.0          22         43        1001.52
12383.0          20         37         600.72
```

STEP 4 : Modelling

In this research paper we are using K -means algorithm for clustering because of it simplicity and advantageous over other algorithm.

K-Means Algorithm

K-means is the mostly popular and widely used algorithm for grouping data into groups to get right number of clusters.[1][2]

K-means is an iterative Algorithm which try to partition the data into k distinct groups. Here K is the number of clusters to be formed which is predetermined by Elbow method which we will discuss further in the paper.

Clustering Steps to follow while using K -means Algorithm:

- Predetermine Number of clusters K.
- Initialize Centroid by randomly selecting K data points.
- Compute the distance of the next data points with all centroids.
- Assign the data point to the nearest cluster
- Repeat this step until all data points converges to a cluster.

Formula for Centroid Determination:

$$C_i = {}^1/_M \sum_{j=1}^{m} X_j$$

Formula for Euclidean Distance

$$d\,(p,\, q) = \sqrt{}(p1\text{-}q1)^2 + (p2\text{-}q2)^2$$

Elbow Method

Elbow method is used to determine the optimal number of clusters based on the dataset. The idea is simple behind it, i.e., plotting the SSE (Sum squared Error) against suitable no of cluster value. Then we will select the value at which there is maximum curve in the graph[2].

STEP 5: Evaluation

Evaluation is a major aspects of any machine learning model. To know your model produces the right output with correct accuracy.

Silhouette Index

Silhouette Index is a value which is used to check the interpretation and validation consistency within clusters of data. This method/technique a brief graphical Representation of how well each object in a dataset is classified. Silhouette Index measures how similar is the object to its own cluster compared to other clusters. Silhouette value ranges between -1 to +1, high value indicates object is well matched and vice versa. Silhouette Index is useful to determine the right cluster configuration, i.e., if many points have low or negative then clustering configuration may have many or few clusters[1].

The Formula is :

$S(i) = bi-ai / \max (bi, ai)$

where:

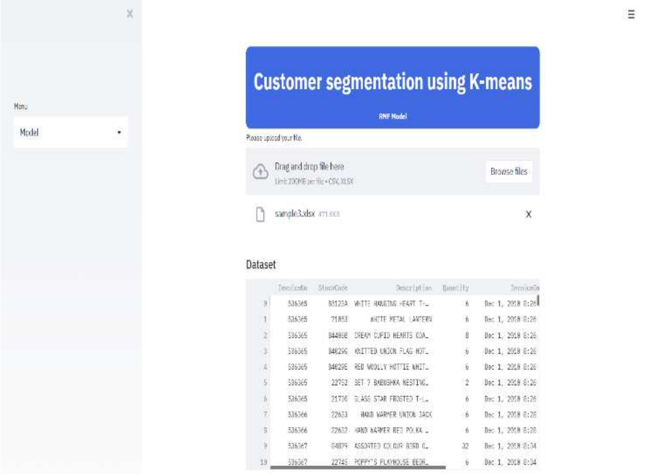
bi = The average distance between i and the same cluster.

ai = The average distance between i with different clusters.

Max (bi, ai) = Average distance between bi with ai.

I. STEP 6: Web Model

So based on this data science model we have also created a web model where any E-commerce based analyst, start-up can analyse it's own customers by just feeding it's customer dataset to our website and the detailed analysis of those customers will be made. So that will help them to know more about their customers and they can make business strategies based on that thereby avoiding future churn of less responsive customers, giving different attention to potential customers etc.



Menu

Model

Customer segmentation using K-means

RMF Model

Please upload your file.

Drag and drop file here

Limit 200MB per file • CSV, XLSX

Browse files

sample3.xlsx 471.0KB

Dataset

	InvoiceNo	StockCode	Description	Quantity	InvoiceDa
0	536365	85123A	WHITE HANGING HEART T-…	6	Dec 1, 2010 8:26
1	536365	71053	WHITE METAL LANTERN	6	Dec 1, 2010 8:26
2	536365	84406B	CREAM CUPID HEARTS COA…	8	Dec 1, 2010 8:26
3	536365	84029G	KNITTED UNION FLAG HOT…	6	Dec 1, 2010 8:26
4	536365	84029E	RED WOOLLY HOTTIE WHIT…	6	Dec 1, 2010 8:26
5	536365	22752	SET 7 BABUSHKA NESTING…	2	Dec 1, 2010 8:26
6	536365	21730	GLASS STAR FROSTED T-L…	6	Dec 1, 2010 8:26
7	536366	22633	HAND WARMER UNION JACK	6	Dec 1, 2010 8:28
8	536366	22632	HAND WARMER RED POLKA …	6	Dec 1, 2010 8:28
9	536367	84879	ASSORTED COLOUR BIRD O…	32	Dec 1, 2010 8:34
10	536367	22745	POPPY'S PLAYHOUSE BEDR…	6	Dec 1, 2010 8:34

Figure 2 Web Model using Streamlit

IV. RESULT

Based on the results of our model are from the data of UK online retail store from December 2010 to dec 2011 which consists of total 13 months. The data consists of total 4 attributes which are listed above in Table 3.

After the data was prepared there was still some skewness of data in RMF model which required further modification to reduce the skewness of the data.

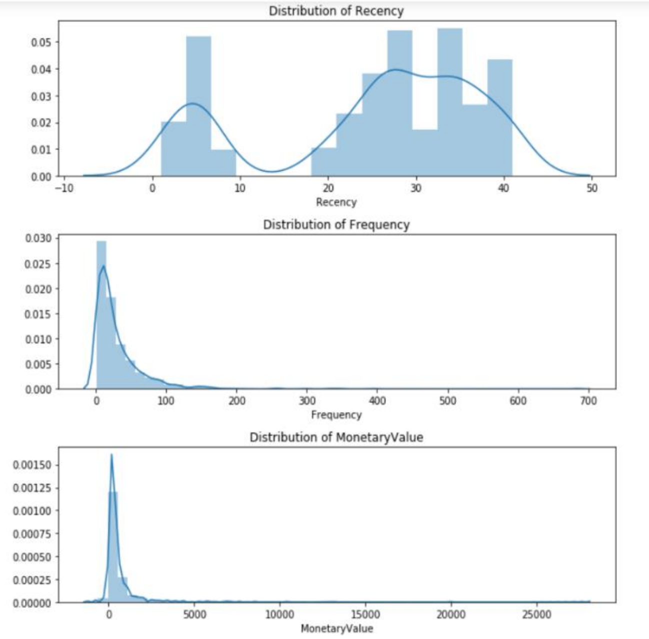


Figure 3 Skewness in RFM Model

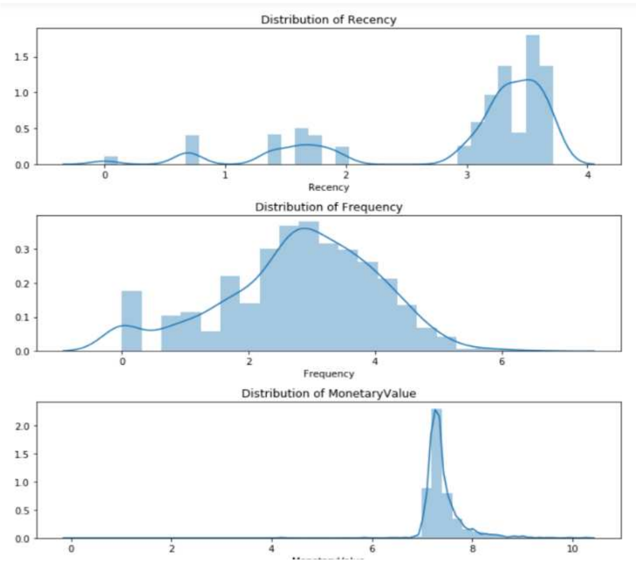


Figure 4 Normalized in RFM Model

After cluster testing the calculation of the silhouette index is continued by measuring the similarities of data points with its own cluster and other clusters as well.

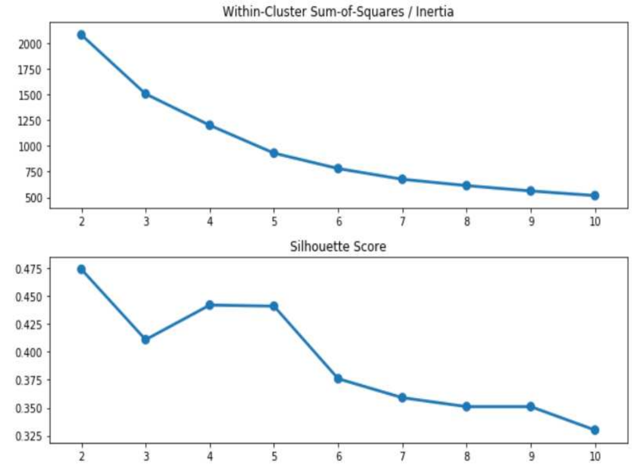


Figure 5 Elbow Graph

Based on the Elbow Method we can see that the right number of clusters is 4 which is showing the highest silhouette value.

Figure 5 Displays the final output of our model which has grouped the data in 3,4,5 clusters.

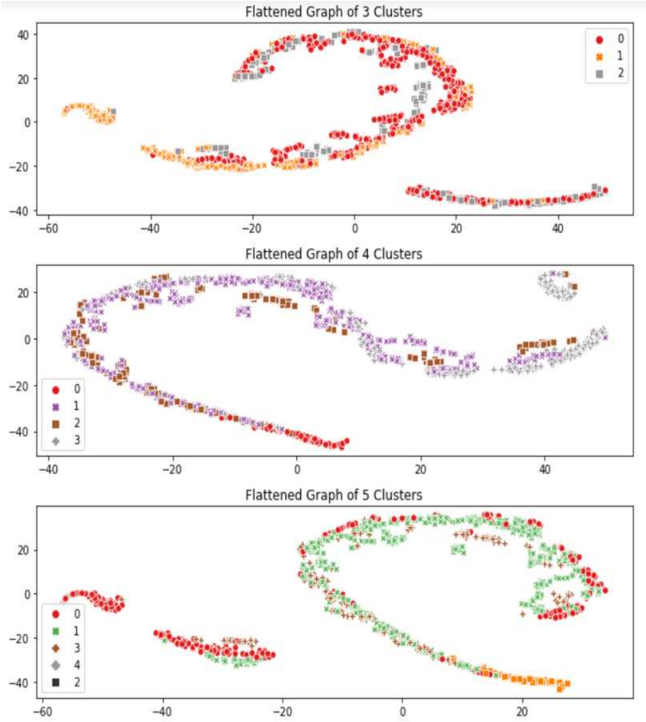


Figure 6 Clustering

V. CONCLUSION

Based on the process of making a customer segmentation which is based on RFM model using K-means algorithm on a transaction data of a UK online retail store, we categorised the customer into 4 clusters based on the characteristics. These 4 clusters are basically Class A, Class B, Class C and Class D. Where Class A generates the highest revenue and Class D least. Customer segmentation is a very powerful tool to get the business insights and on how the customer behave. The value of silhouette index is 0.442 which is considered as good for the given dataset. Based on this result we obtained, this can help the company to develop market strategies and also can use as a promotional medium to their loyal customers. There are many other tools and method which can be used as comparison to the system that are already developed.

VI. REFERENCES

[1]. Tushar Kansal; Suraj Bahuguna; Vishal Singh; Tanupriya ChoudhuryCustomer, " Segmentation based on RFM model and Clustering Techniques With K-Means Algorithm", IEEE 2018

[2]. Muhammad Iqbal Dzulhaq ,Kartika Wulan Sari, Syaipul Ramdhan, Rahmat Tullah ,Sutarman," Customer Segmentation Based on RFM Value Using K-Means Algorithm", ICIC 2019

[3]. Chaohua Liu,"Customer Segmentation and Evaluation Based On RFM, Cross- selling and Customer Loyalty", IEEE 2011

[4]. Shreya Tripathi1, Aditya Bhardwaj , Poovammal," Approaches to Clustering in Customer Segmentation", IJET 2018

[5]. A.S.M. Shahadat Hossain,"Customer Segmentation using Centroid Based and Density Based Clustering Algorithms", IEEE 2017

Cite this article as :

Rahul Shirole, Laxmiputra Salokhe, Saraswati Jadhav, "Customer Segmentation using RFM Model and K-Means Clustering", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 8 Issue 3, pp. 591-597, May-June 2021. Available at doi : https://doi.org/10.32628/IJSRST2183118
Journal URL : https://ijsrst.com/IJSRST2183118