Customer Segmentation and Product Enhancement

B.V. Sathish.[1], L. Jyothi.[2], L. Siva kumara.[2], M. Sanjana.[2*]
[1]Assistant professor, Department of Computer Science and Engineering.
[1,2]Andhra Loyola Institute of Engineering And Technology, Vijayawada, Andhra Pradesh.
*Corresponding Author's email: sanjanamyna9@gmail.com

Abstract:
In this evolving era of constant growth and development business sector is one such important sector that plays it's role very well in the up lifting of the economy in any country. Considering this there are many businesses budding and growing everyday. Customers and products are the two main things in any business.

So this article about the "CUSTOMER SEGMENTATION AND PRODUCT ENHANCEMENT" proposes the effective ways for the companies to know what set of customers they had to target and increase their sales and it helps customer to deal with such a large raft of products and makes it easy to know what they had to purchase. It segments the customers on the basis of their purchasing habits and gives the them a list of products according to their interests. We start the process by gathering the information of the customers when ever they make a purchase and the outcome would be the segmented cluster results by implementing different methods and algorithms. This includes different set of information regarding the customers or the people interested with the products of this business. Different kinds of strategies, demographics, lifestyles and usage patterns of customers are used to segment the customers and to identify the potential target. These methods are useful for both communication and product development of a company. In this project we identify common characteristics that define good customers. To define what a good customer means we clearly need to define quality score that we can use objectively rank the customer. We also use different visualization techniques to compare different methods and find the accuracy between the methods.

By the end of this project, we can identify the targeted customers and help the businesses to improve their products and services.

Keywords: RFM ,k-means,RFM quartiles, RFM score,segmented graphs,clusters.

[1] Introduction

Segmentation is not the process of targeting the customers but dividing them into groups based on different traits they share. Direct-to-consumer brands and business to business companies are at a great advantage because of the amount of information they can obtain from their customers at a purchase from their transaction data alone.

Customer Segmentation is the process of dividing customers into groups based on certain similar characteristics shared among the customers. For a product from a company all customers share the common need of product or service.

The information obtained include data types such as their location based on their shipping info or browser info and such also the type of device based on device used, promo codes and also type of payment method used for their transaction. Also, as a checkout process, they can also obtain additional information such as gender, job title, purpose of purchase such as business or personal or gift or self-consumption etc...But beyond that there are demographic differences like age, gender, marital status, education, location and all and they may also have additional socio-economic, lifestyle or other behavioural differences that can be useful for an organization.

Customer segmentation models range from simple to complex and can be used for different business reasons. There are actually different types of segmentation process.

In the present project we brought machine learning into action and various algorithms are applied for revealing the hidden patterns of the data and help in making better decisions. We focused on calculating the RFM values which means recency, frequency and monetary values of the customer data and obtain the clustering results in a graphical form. We use RFM algorithm to calculate RFM values and K- means algorithm to get the clustering results. We also describe on how we actually calculate these mathematically.

To any person that's handling the business it gets easy to understand the segmented and clustered data of customers rather understanding the raw data.

[2] Literature Survey

A. In the process of working with the customer segmentation we came across various project papers which we used as our references. By these references we came to know that this is already in use in many of the larger countries where business sector is the most prioritised sector. Here we have different types of clustering process to define what a good customer means we clearly need to define quality score that we can use objectively rank the customer. We also use different visualization techniques to compare different methods and find the accuracy between the methods like –

-> Customer Segmentation using K-means

-> Clustering Customer Segmentation Analysis Based on SOM

-> Clustering Analysis of Customer Segmentation Based on Broad Learning System Data Mining Approach for Customer Segmentation in B2B Settings using Centroid Based Clustering.

-> Customer Segmentation in a Travel Agency Dataset using Clustering Algorithms

-> Customer Segmentation based on RFM model and Clustering Techniques with K-Means Algorithm

-> Customer Segmentation Based on RFM Value Using K-Means Algorithm

B. PROPOSED WORK

In the current project we implement different algorithms on the data set of retail business customer information that we have taken from an online data providing website which is actually a real time data set.

We read the data file after uploading it and read the head values from the data file. And for that data we start considering the date and time, country and etc. We starting calculating the recency,frequency and monetary values for the customer ID's respectively. Then we calculate the quartiles and on a whole we calculate RFM quartile which helps in calculating the RFM score.

By this we get the list of best customers, loyal customers, big spenders, almost lost, lost customers, lost cheap customers. RFM is a data-driven customer segmentation technique that allows marketers to take tactical decisions. It empowers marketers to quickly identify and segment users into homogeneous groups and target them with differentiated and personalized marketing strategies. This in turn improves user engagement and retention.

• The more recent the purchase, the more responsive the customer is to promotions

• The more frequently the customer buys, the more engaged and satisfied they are

• Monetary value differentiates heavy spenders from low-value purchasers.

The other data mining technique is the K- means algorithm. This algorithm works with the machine learning technique it gives the clear graphs of the clusters. K-Means Clustering Algorithm Specify number of clusters K. Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement. Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn't changing.

• Compute the sum of the squared distance between data points and all centroids.

• Assign each data point to the closest cluster (centroid)

• Compute the centroids for the clusters by taking the average of the all-data points that belong to each cluster.The approach K means follows to solve the problem is called Expectation- Maximization. The E-step is assigning the data points to the closest cluster. The M-step is computing the centroid of each cluster. Below is a breakdown of how we can solve it mathematically is

$$J = \sum_{i=1}^{m} \sum_{k=1}^{K} w_{ik} \|x^i - \mu_k\|^2 \qquad (1)$$

$$\frac{\partial J}{\partial w_{ik}} = \sum_{i=1}^{m} \sum_{k=1}^{K} \|x^i - \mu_k\|^2$$
$$\Rightarrow w_{ik} = \begin{cases} 1 & \text{if } k = argmin_j \|x^i - \mu_j\|^2 \\ 0 & \text{otherwise.} \end{cases} \qquad (2)$$

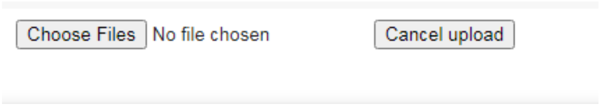
$$\frac{\partial J}{\partial \mu_k} = 2 \sum_{i=1}^{m} w_{ik}(x^i - \mu_k) = 0$$
$$\Rightarrow \mu_k = \frac{\sum_{i=1}^{m} w_{ik} x^i}{\sum_{i=1}^{m} w_{ik}} \qquad (3)$$

$$\frac{1}{m_k} \sum_{i=1}^{m_k} \|x^i - \mu_{c^i}\|^2 \qquad (4)$$

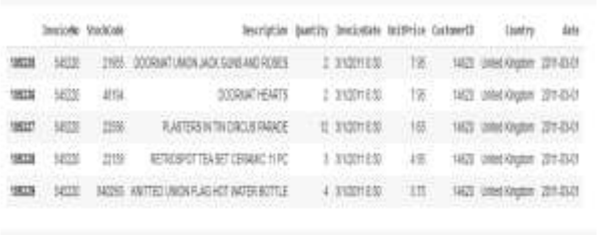
[3] Work flow

A. RFM ALGORITHM

In this paper- The very first step in the working of this segmentation process would be uploading the data file of customer information we have. This file uploading process starts right after we import all the required libraries and data files in our python program. The template for uploading the file will be something like this:

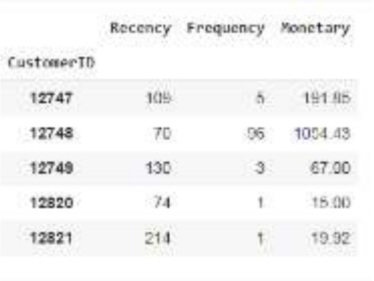


After uploading the file we start reading all the rows and columns from the data set. We read the customer id, invoice no, stock code, description, quantity, invoice date, unit price, country, date and time. The following table represents the data read:



	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	date
[illegible]	545220	21955	DOORMAT UNION JACK GUNS AND ROSES	2	3/1/2011 8:30	7.95	14620	United Kingdom	2011-03-01
[illegible]	545220	48194	DOORMAT HEARTS	2	3/1/2011 8:30	7.95	14620	United Kingdom	2011-03-01
[illegible]	545220	22556	PLASTERS IN TIN CIRCUS PARADE	12	3/1/2011 8:30	1.65	14620	United Kingdom	2011-03-01
[illegible]	545220	22139	RETROSPOT TEA SET CERAMIC 11 PC	3	3/1/2011 8:30	4.95	14620	United Kingdom	2011-03-01
[illegible]	545220	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	4	3/1/2011 8:30	3.75	14620	United Kingdom	2011-03-01

We take the segregate the customers who are from a particular region , who spent more amount on the purchases, who are frequently purchasing, who most visited. This is nothing but the recency, frequency and monetary values. We calculate all the three values respectively step by step using customer id and the last purchase date. And finally the values of recency, frequency and monetary are dispkayed together in a single table as :

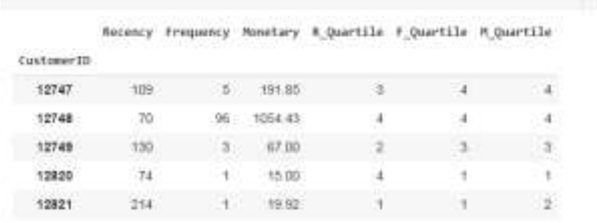


	Recency	Frequency	Monetary
CustomerID			
12747	109	5	191.85
12748	70	96	1054.43
12749	130	3	67.00
12820	74	1	15.00
12821	214	1	19.92

The next step is the pre last for the rfm algorithm process. We calculate the quartiles. That is r_quartile, f_quartile, m_quartile. These quartiles will help you increase conversion rates, it also reduces the number of customers in each cell. For most online marketers, quartiles will be sufficient. With customers now in quartiles, it's time to group them into RFM segments.

The quartile values are calculated like the top quartile for Recency is called R-1, the second quartile is called R-2, and so on. Dividing into quartiles will create 64 RFM segments: 4 Recency groups x 4 Frequency groups x 4 Monetary Value groups.

The final caluculated quartile values are re- checked and the table is printed as follows:



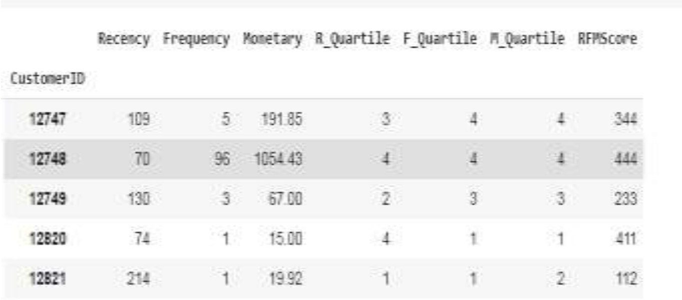
	Recency	Frequency	Monetary	R_Quartile	F_Quartile	M_Quartile
CustomerID						
12747	109	5	191.85	3	4	4
12748	70	96	1054.43	4	4	4
12749	130	3	67.00	2	3	3
12820	74	1	15.00	4	1	1
12821	214	1	19.92	1	1	2

The later step will be calculating the rfm score.

Once we have RFM values from the purchase history, we assign a score from one to five to recency, frequency and monetary values individually for each customer . Five is the best/highest value, and one is the lowest/worst value. A final RFM score is calculated simply by combining individual RFM score numbers.

Both the quartile values and the score values are re-checked and the table is printed.

The final table containing the rfm quartile values and the rfm score is something like this:

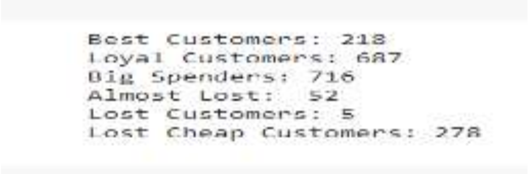


	Recency	Frequency	Monetary	R_Quartile	F_Quartile	M_Quartile	RFMScore
CustomerID							
12747	109	5	191.85	3	4	4	344
12748	70	96	1054.43	4	4	4	444
12749	130	3	67.00	2	3	3	233
12820	74	1	15.00	4	1	1	411
12821	214	1	19.92	1	1	2	112

Nevertheless we are at the end of our process. The rfm algorithm process ends by knowing the list of best customers, loyal customers, big spenders, almost lost, lost customers, lost cheap customers.

The end resulting values of the best customers, loyal customers, big spenders, almost lost, lost customers, lost cheap customers will be calculated on the bases of the quartile and score values.

The end results will be something like this:



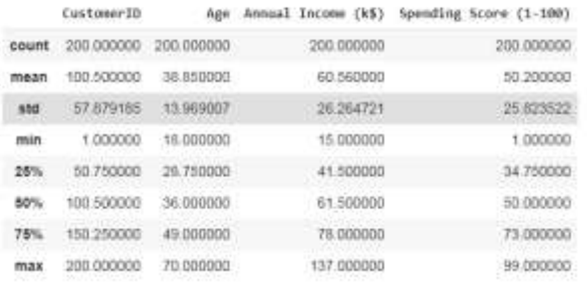
```
Best Customers:  218
Loyal Customers:  687
Big Spenders:  716
Almost Lost:   52
Lost Customers:  5
Lost Cheap Customers:  278
```

B. WORK FLOW OF K MEANS ALGORITHM.

In this paper for the implementation of the next data mining process- Take another data set containing the information about the annual income and the spending score of the top customers which we get from the rfm algorithm execution. The process for uploading the following data file will be same as in the rfm algorithm execution.

Display the table considering the information about the age, customer id , annual income, spending score .

It is as follows:



	CustomerID	Age	Annual Income (k$)	Spending Score (1-100)
count	200.000000	200.000000	200.000000	200.000000
mean	100.500000	38.850000	60.560000	50.200000
std	57.879185	13.969007	26.264721	25.823522
min	1.000000	18.000000	15.000000	1.000000
25%	50.750000	28.750000	41.500000	34.750000
50%	100.500000	36.000000	61.500000	50.000000
75%	150.250000	49.000000	78.000000	73.000000
max	200.000000	70.000000	137.000000	99.000000

We have to follow certain steps for going through the k means clustering. They are:

Step 1: Choose the number of clusters based on the raw information we have.

Step 2: Select few random points from the data as centroids.

Step 3: Assign all the points to the closest cluster centroid.

Step 4: Recompute the centroids of newly formed clusters.

Step 5: Repeat steps 3 and 4. Until you complete clustering for the entire data set.

The clustering results includes clusters tables, plotting graphs of different types like the boxplot graph, swarmplot graph, clustered scatter plotting graphs, etc…

The annual income is considered as (k$) and the spending score is considered as (1-100)

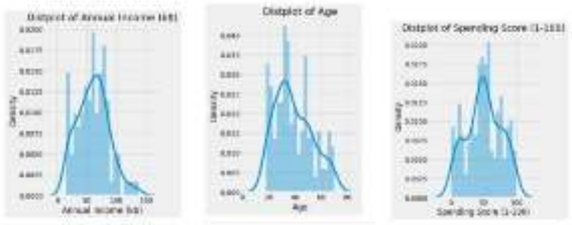
The displot graphs are considered as

Density and age

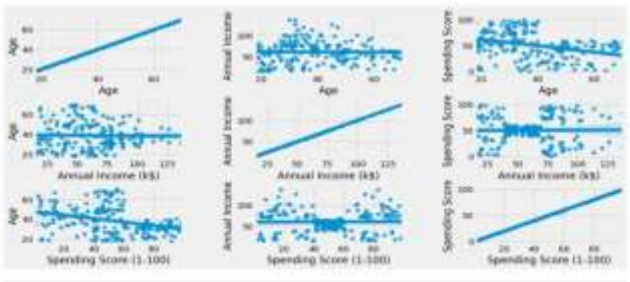
Density and annual income

Density and spending score

They come out something like this:

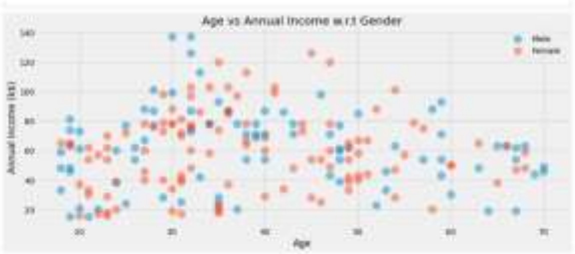


The straight and the combined cluster graphs are displayed at a stretch easily after the displot graphs. They will be something like this:

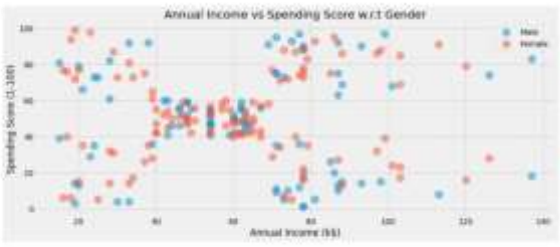


Every boxplot , cluster scatter plotting graphs has the scatter graph of them in this paper.

It will be something like this:



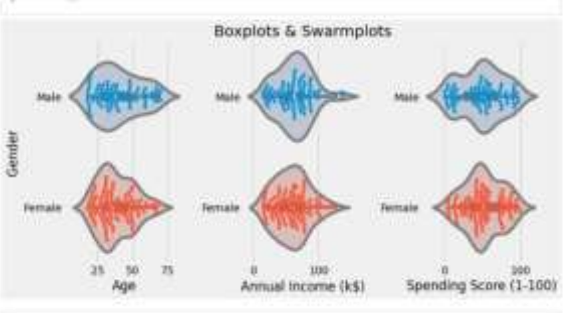
The following represents the scatter plotting for the annual income and gender combined against age.



The following represents the scatter plot graph for the annual income vs spending score .

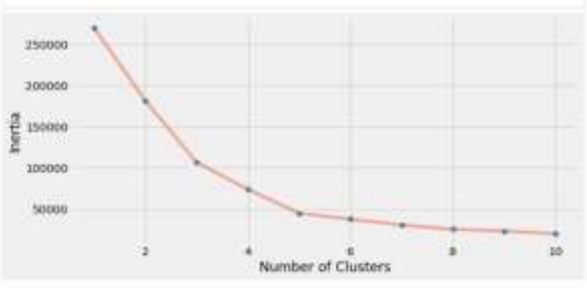
This paper provides the swarmp plot graphs where Seaborn swarmplot is probably similar to stripplot, only the points are adjusted so it won't get overlap to each other as it helps to represent the better representation of the distribution of values. The process to make this swarmp plot graphs is we first draw a single horizontal swarm plot using only one axis: If we use only one data variable instead of two data variables then it means that the axis denotes each of these data variables as an axis. X denotes an x-axis and y denote a y-axis.

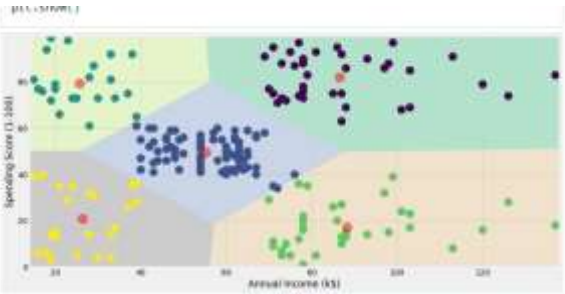
The outcome of the swarmp plot graph for the following data that this paper publishes is something like this:



After all these graphs the next step will be for producing the clustering graphs and the straight graphs for the age,annual income and the spending score now in the next step we calculate the annual income and the spending score alone on the basis of the customer and the age information from the previous graphs results.

It will be something like this:





By this we come to the end of the paper after getting the results of segmented information of the customers that we have in the data set.

[4] Conclusion

In this project we targeted to group or divide the customers based on their age, gender, education qualifications and purchases. For achieving this we have used two of the algorithms namely K means and RFM (Recency, Frequency and Monetary value). We have successfully implemented both the algorithms and plotted various graphs for better visualizations. In RFM algorithm eventually we had to use the concept of K means algorithm. To say precisely the RFM algorithm is advisable to be used for small or medium scaled business.Because as the large business have greater factors and criteria to consider which this particular algorithm cannot handle and give accurate results.Whereas the K means algorithm is very common but most effective. It can be used for a larger business or bigger marketing company too as it has the capacity to withstand the more number of factors and gives the accurate results as well.

[5]Acknowledgments

This research project was partially supported by the Department of Computer Science and Engineering, Andhra Loyola Institute of Engineering and Technology, Jawaharlal Nehru Technological University. We are grateful to Associate Professor Mrs.B. Venkata Sathish for leading us to develop and contribute a paper to the conference.

[6] References

[1]. Sabri Serkan Güllüoğlu “SEGMENTING CUSTOMERS WITH DATA MINING TECHNIQUES”

[2]. Hua Ertian, Lü Huanhuan, Chen Daqiang, Fei Yulian “A Method for Customer Demands Groups Segmentation in Product Design Based on Fuzzy Clustering and Trigonometric Functions”

[3]. Anu Gupta Aggarwal, Sweta Yadav “Customer Segmentation Using Fuzzy-AHP and RFM Model”

[4]. Ina Maryani, Dwiza Riana, Rachmawati Darma Astuti, Ahmad Ishaq, Sutrisno, Eva Argarini Pratama “Customer Segmentation based on RFM model and Clustering Techniques

[5]. With K-Means Algorithm” Dedi, Muhammad Iqbal Dzulhaq, Kartika Wulan Sari, Syaipul Ramdhan, Rahmat Tullah, Sutarman “Customer Segmentation Based on RFM Value Using K-Means Algorithm”

[6]. Yong Huang, Mingzhen Zhang “Research on improved RFM customer segmentation model based on K-Means algorithm”

[7]. Tushal Kansal, Suraj Bahuguna, Vishal Singh, Tanupriya Choudhury “Customer Segmentation using K-means Clustering”

[8]. Ying Li, Feng Lin “Customer Segmentation Analysis Based on SOM Clustering”

[9]. Zhenyu Wang, Yi Zuo, Tieshan Li, “Analysis of Customer Segmentation Based on Broad Learning System”

[10]. Nadhira Riska, Maulina, Isti Surjandari, Annisa Marlin Masbar Rus “Data Mining Approach for Customer Segmentation in B2B Settings using Centroid Based Clustering”

[11]. Clodomir J. Santana Jr., Pedro Aguiar and Carmelo J. A. Bastos-Filho “Customer Segmentation in a Travel Agency Dataset using Clustering Algorithms”

[12]. Sukru Ozan, Ph.D. “A Case Study on Customer Segmentation by using Machine Learning Methods”

[13]. Pāvels Gončarovs, "Using Data Analytics for Customers Segmentation: Experimental Study at a Financial Institution"

[14]. SK Kotamraju, PG Arepalli, SS Kanumalli (2021), Implementation patterns of secured internet of things environment using advanced blockchain technologies, Materials Today: Proceedings, 2021.

[15]. Gopi A.P., Patibandla R.S.M. (2021) An Efficient Methodology for Avoiding Threats in Smart Homes with Low Power Consumption in IoT Environment Using Blockchain Technology. In: Choudhury T., Khanna A., Toe T.T., Khurana M., Gia Nhu N. (eds) Blockchain Applications in IoT Ecosystem. EAI/Springer Innovations in Communication and Computing. Springer, Cham. https://doi.org/10.1007/978-3-030-65691-1_16.

[16]. RSML Patibandla, SN Mohanty (2021), Need of Improving the Emotional Intelligence of Employees in an Organization for Better Outcomes,Decision Making And Problem Solving: A Practical guide, 2021.

[17]. Lu Siyue, Zhang Baoqun, Zhang Lu, Xu Hui, Zhang Jianxi, Ma Longfei, Wang Peiyi "Prediction of Business User Segmentation Model Based on Customer Value"

[18]. K. Torizuka1, H. Oi, F. Saitoh, S. Ishizu "Benefit Segmentation of Online Customer Reviews Using Random Forest"