Customer Segmentation in e-Commerce using ML

Jenita J[1], Yash Agarwal[2], Yash Ashok Vishwakarma[3], Akhil Reddy[4], Arifulla[5]*
[1]Professor, [2,3,4,5]Student
Department of CSE, HKBKCE, Bangalore, India

***Corresponding Author**
E-mail Id:-Jenitam.ec@hkbk.edu.in*

ABSTRACT

Many small online outlets and newcomers to the online retail territories are keen to record and promote the shopper-centric stores, but technically lack the know-how and information needed to do so. doing. This article provides his case study on the use of information mining strategies in customer-centric business intelligence for an online retailer. The primary purpose of this assessment is to help the company better understand its customers and, as a result, make its customer-centric advertising and her marketing more efficient. Under the assumptions of recency, frequency, and economic version, the customer firms were divided into a number of significant firms using the ordering method clustering rule set and selection tree derivation, and the main features of the customer phase have been fully identified. For this reason, the company is provided with his customer-centric marketing tips.

Keywords:-*Clustering, K-means, Online, shopping, Retail industry, Customer segmentation*

INTRODUCTION

The significant increase in online sales indicates a fundamental shift in the way consumers purchase and use financial services. Online shopping has some unique features compared to traditional shopping at retail stores. Each customer has an online business account with vital contact and payment information, and the entire process of their purchasing can be tracked promptly and correctly.

Additionally, each client's order is often correlated with their shipping and billing addresses. These appealing and distinctive online shopping features give online merchants the ability to treat every consumer as an individual by getting to know them on a personal level and developing customer-centric business intelligence. When it comes to business information for their customers, online retailers typically address common business questions such as: How long did the customer stay on each website and in what order did the customer visit the series of product websites? Who are your company's most valuable/least valuable customers? What are their special features? Who are your most loyal/least loyal customers? What are the specifications?

What are your customers' buying patterns? What products/items do your customers often buy together? In what order did he purchase the products? What type of customer is most likely to respond to a particular sales letter? What is your sales model in terms of product/item, region and time (weekly, monthly, quarterly, yearly, seasonal) and different perspectives? It is deployed holistically and combined with many well-known business indicators of customer profitability and value.

Companies are using multiple segmentation criteria and ways to more accurately identify and comprehend consumer groups, and to provide prioritized goods and services that cater to their various wants and needs. In order to generate lucrative segments and serve specific segments based on competitive advantage, segmentation is crucial for businesses.

To arrange their marketing campaigns, many marketers find it challenging to choose the appropriate consumer categories (Mohammadian Makrani, 2016).

Therefore, the purpose of this paper is to investigate whether employing her RFM analysis along with cluster analysis can improve consumer segmentation. In particular, we want to a) identify sporting goods retail consumer groups, b) compare these segments to the company's current segmentation, and c) find out how RFM values cluster together.

LITERATURE SURVEY:

Decision makers use many variables to segment their customers. Demographic variables such as age, gender, family, education, and income are the simplest. to break down and are the most common.

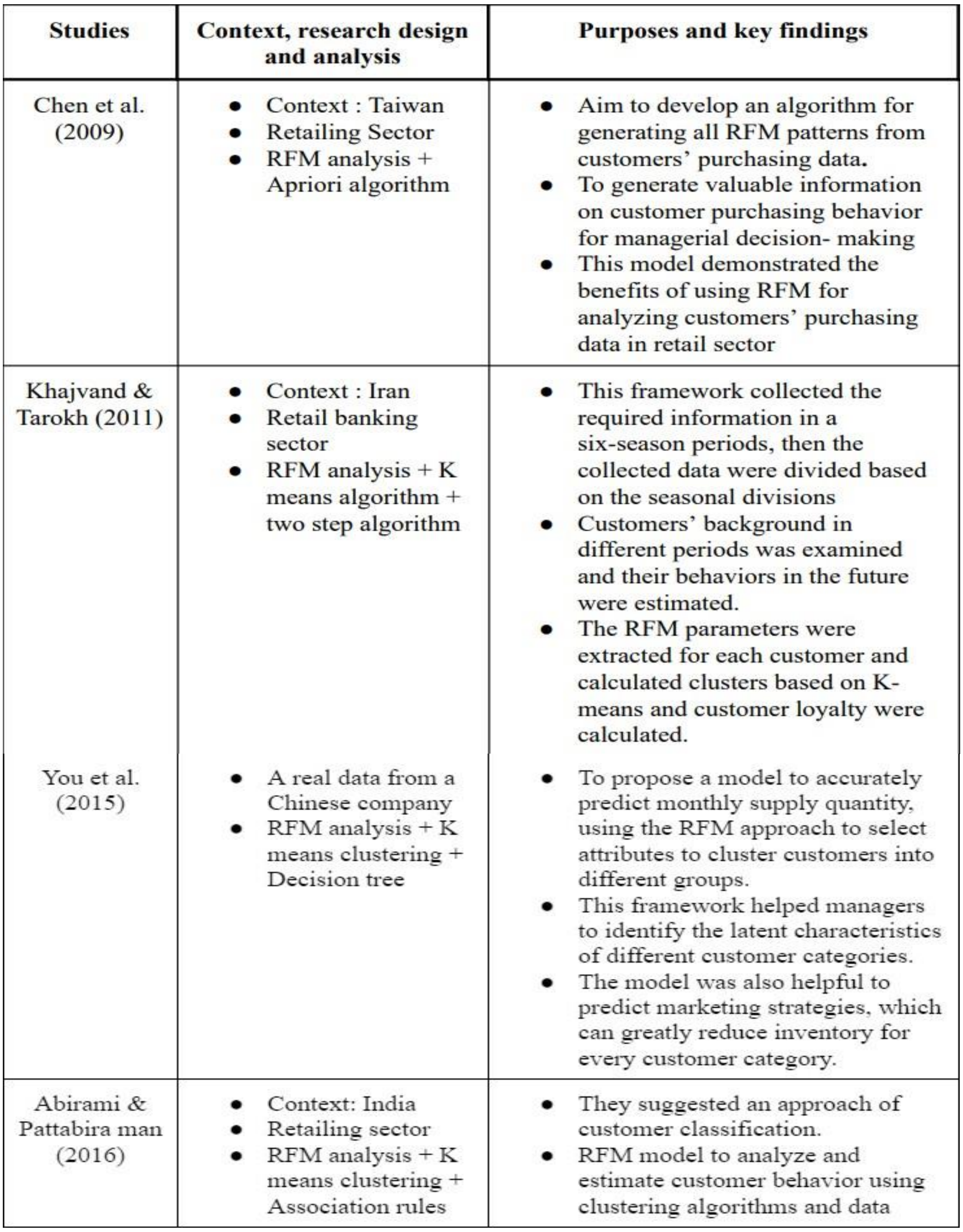
Sociocultural, geographic, psychological, and behavioral variables are other key variables used for segmentation. In recent years, many researchers have studied customer segmentation in the sports industry. Initially, the focus was on football and audience base segmentation according to the Russian dualistic type. Stewart et al. (2003) included types of Type 1 and Type 2 consumers, with Type 1 being loyal, traditional, expressive, irrational, symbolic, and persistent, and Type 2 being adventurous, modern, and dependent. . Secrecy, rationality, middle-class and low-loyalty...

Then multidimensional patterns emerged. Smith and Stewart (1999) divided 4444 sports consumers into 5 groups. RFM is a popular model for customer value analysis. It has been used by many researchers for customer segmentation (Spring et al., 1999; Jonker et al., 2006; Cheng & Chen, 2009; Khajvand & Tarokh, 2011).

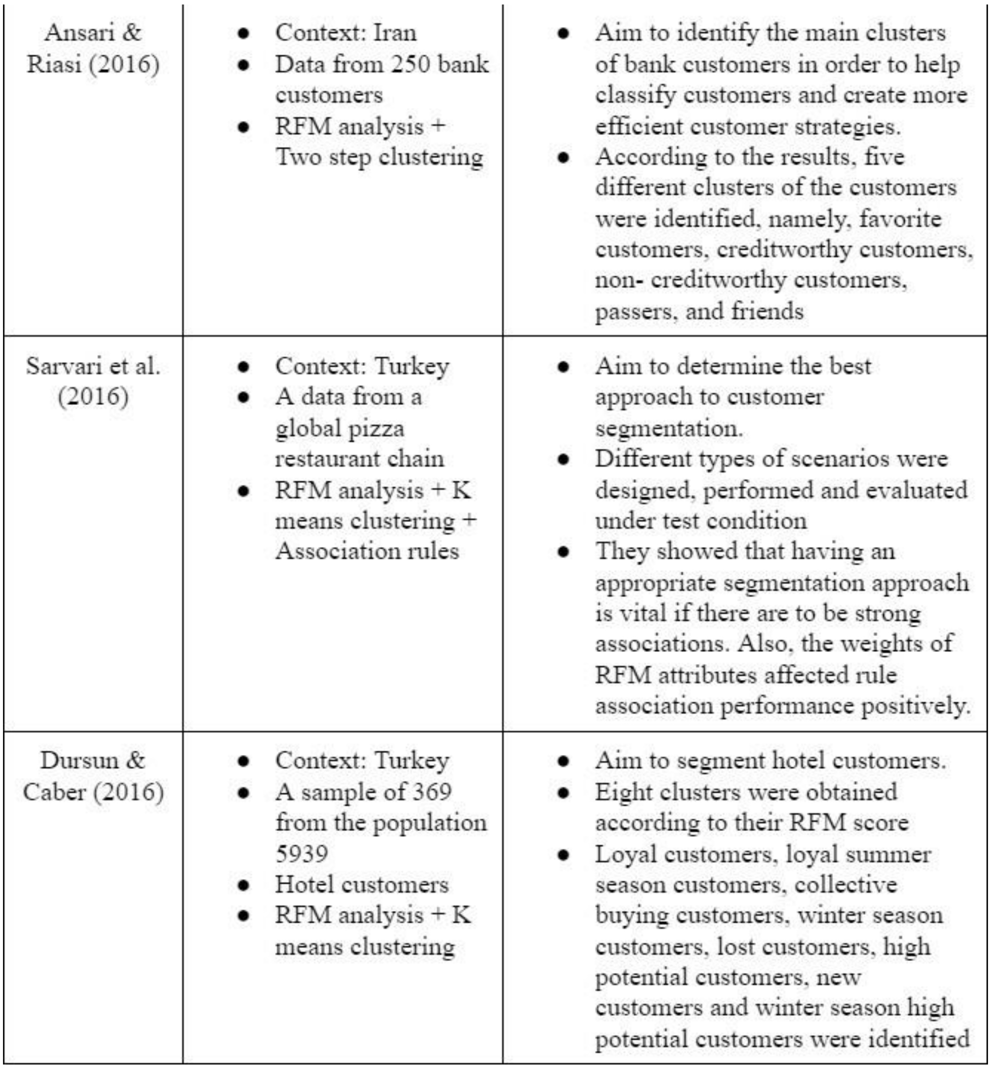
Because RFM analysis customer behavior. Over the past 20 years, several researchers have considered his RFM model when developing prediction and classification models. In the example of Etzion et al. (2004) classified customers according to profitability and customer value creation. Choi et al. (2006) proposed a model using RFM variables to evaluate customer response.

Cheng &amp; Chen (2009) presented a data mining model for predicting customer loyalty. Further reading includes RFM models of clustering algorithms. It actually has to do with the model used in this article. These studies are shown in Table 1.

***Table 1:**-Literature overview on researches includes **RFM (**Recency, Frequency and Monetary**)** models and clustering techniques*



Studies	Context, research design and analysis	Purposes and key findings
Chen et al. (2009)	• Context : Taiwan • Retailing Sector • RFM analysis + Apriori algorithm	• Aim to develop an algorithm for generating all RFM patterns from customers' purchasing data. • To generate valuable information on customer purchasing behavior for managerial decision- making • This model demonstrated the benefits of using RFM for analyzing customers' purchasing data in retail sector
Khajvand & Tarokh (2011)	• Context : Iran • Retail banking sector • RFM analysis + K means algorithm + two step algorithm	• This framework collected the required information in a six-season periods, then the collected data were divided based on the seasonal divisions • Customers' background in different periods was examined and their behaviors in the future were estimated. • The RFM parameters were extracted for each customer and calculated clusters based on K-means and customer loyalty were calculated.
You et al. (2015)	• A real data from a Chinese company • RFM analysis + K means clustering + Decision tree	• To propose a model to accurately predict monthly supply quantity, using the RFM approach to select attributes to cluster customers into different groups. • This framework helped managers to identify the latent characteristics of different customer categories. • The model was also helpful to predict marketing strategies, which can greatly reduce inventory for every customer category.
Abirami & Pattabira man (2016)	• Context: India • Retailing sector • RFM analysis + K means clustering + Association rules	• They suggested an approach of customer classification. • RFM model to analyze and estimate customer behavior using clustering algorithms and data



Ansari & Riasi (2016)	• Context: Iran • Data from 250 bank customers • RFM analysis + Two step clustering	• Aim to identify the main clusters of bank customers in order to help classify customers and create more efficient customer strategies. • According to the results, five different clusters of the customers were identified, namely, favorite customers, creditworthy customers, non- creditworthy customers, passers, and friends
Sarvari et al. (2016)	• Context: Turkey • A data from a global pizza restaurant chain • RFM analysis + K means clustering + Association rules	• Aim to determine the best approach to customer segmentation. • Different types of scenarios were designed, performed and evaluated under test condition • They showed that having an appropriate segmentation approach is vital if there are to be strong associations. Also, the weights of RFM attributes affected rule association performance positively.
Dursun & Caber (2016)	• Context: Turkey • A sample of 369 from the population 5939 • Hotel customers • RFM analysis + K means clustering	• Aim to segment hotel customers. • Eight clusters were obtained according to their RFM score • Loyal customers, loyal summer season customers, collective buying customers, winter season customers, lost customers, high potential customers, new customers and winter season high potential customers were identified

Leenheer and Bijmolt (2008) define a loyalty program as: Demoulin & Zidda, 2009), collecting data on shoppers and shopping habits (Liu, 2007; Sands & Ferraro, 2010), increasing customer retention and sales (Liu et al., 2011), rewarding loyal customers Courtesy of (Jere et al Posthumous, 2014).) encourage individual suggestions. In recent years, advanced data mining techniques have made customer segmentation and implementing more effective loyalty programs easier and more valuable. Gomez et al. (2006). Kandampulli and Suhartando (2000) and Bulut (2015) also list repeat purchase behavior and repeat purchase frequency of customers as components of customer loyalty. There are also numerous studies in the existing marketing literature confirming the association between participation in loyalty programs and brand his loyalty (Sharp and Sharps, 1997; Bolton et al., 2000; Maity & Gupta, 2016).) indicates a significant result. Customers participating in loyalty programs exhibit higher behavioral and loyalty attitudes, visit retail stores more frequently and purchase more than non-members (Ha & Stoel, 2014; Melnyk & Bijmolt, 2015). Lu, 2007). Kimet al. (2009) found that the higher the loyalty, the higher the position in the corporate loyalty program.

METHODOLOGY

The three key steps of the suggested research approach are as follows. Data transformation and cleansing preanalytical operations are part of the first step. RFM analysis, two-level cluster analysis, Fuzzy-C clustering, and k-means clustering were all used to examine the data after that. The results have now been made public. Figure 1 depicts the full methodological approach. Huway (2014). Retail network data were used in this investigation. One of Turkey's largest sports retail chains is this one. The company sells items including shoes, shirts, hoodies, accessories, and sporting goods, much like all other sports merchants. In order to categorize clients and establish a customer loyalty programme, the manager made the decision in 2010 to develop a customer loyalty card system. There are three tiers of cards in the loyalty card programme. Silver, Gold, and Premium. Clients who are Customers who are members of our loyalty program earn more points based on their spending during the year. A Bronze Card customer is a member whose annual spending is less than her 2,000 Turkish Lira (TL) (≈520$).

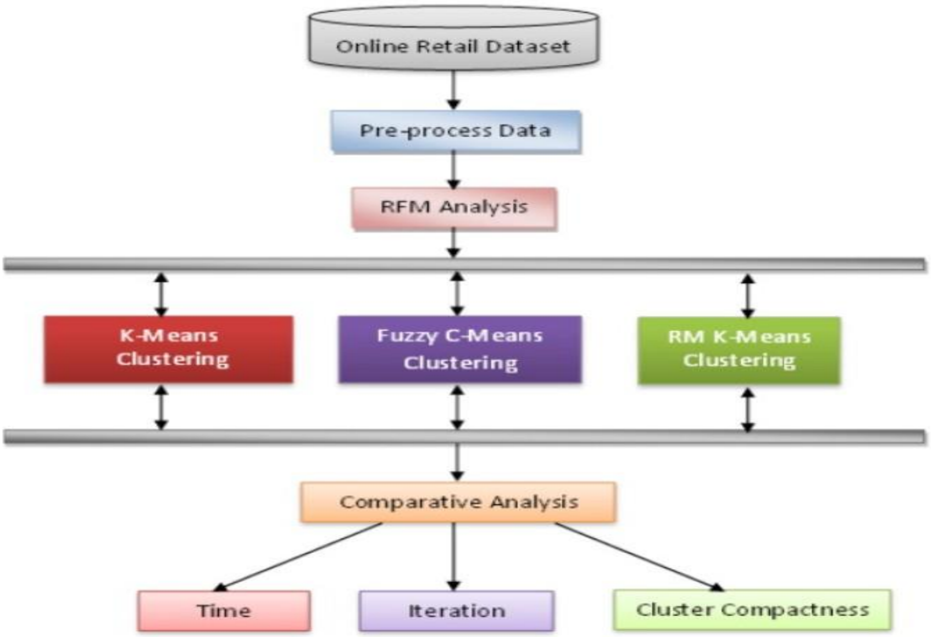


***Fig 1**: methodology flow chart*

The data set contains customer variables related to the period from 1/1/2016 to 12/31/2016. The data set consists of 715328 registers belonging to both clients and e-clients. A data cleaning process was run to remove some missing and invalid values from the data set. In 2016, 700,032 registries were extracted and used for analysis, including customers who made purchases in physical stores and online. The entire trained population is used for analysis. Therefore, no sampling method was applied. According to the company's current segmentation; Of the 700,033 customers, 694,647 have Bronze cards, 4,469 Gold cards, and 916 Premium cards.00

RFM analysis was performed to determine the customer's R, F, and M values, and these metrics were used to determine the clusters proposed by the company. Table 2 shows the indicators R, F and M. (P). A capital letter on the novelty indicates the time of the last purchase. Frequency (F) represents the total number of purchases, and amount (M) represents the total value of a customer.

DATA PRE-PROCESSING

To perform the required model-based cluster analysis according to RFM, the original dataset must be preprocessed. The main steps and related tasks involved in data preparation are:

1. Select an appropriate target variable from the given dataset. In this example, six variables were selected: Invoice, StockCode, Quantity, Price, InvoiceDate, and PostCode.

2. Create an aggregate variable called Amount by multiplying Quantity by Price to find the total amount spent per product/item in each transaction.

3. Divide the InvoiceDate variable into the two variables Date and Time. This enables you to apply different policies to transactions made by the same customer at various times on the same day.

4. Excludes all transactions not associated with a zip code of. This solves all missing values in the 's relationship to the PostCode variable. It also filters out all transactions that are not linked to the postal code.

5. Sort the dataset by postal code and create three main summary variables for recency, frequency and monetary

RFM MODEL-BASED CLUSTERING ANALYSIS

Clustering

Using the processed target data set, we wanted to determine whether we could segment consumers in meaningful segments in terms of topicality, frequency, and monetary value. For this purpose, the k-means clustering algorithm was used. As we know, the k-means clustering algorithm is very sensitive to data sets containing outliers (abnormalities) and variables with unique scales and magnitudes. As shown in Figure 2, histograms of the recency, frequency, some instances with currency and frequency values that differ significantly from the majority of the dataset. becomes clear. instance in the dataset. These instances are valid from a business perspective as they are actual transaction records. However, from a data analysis perspective, they are outliers. Therefore, these cases should be separated from the majority and handled individually. Furthermore, the three variables are not of equal scale and have completely different value ranges.

***Table.2:**-RFM Statistics*

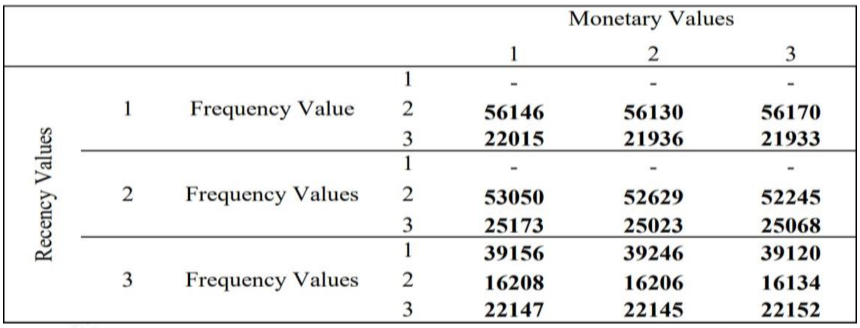
	R	F	M
N	700032	700032	700032
Mean	119,46	1,9	336,67
Std. Deviation	71,635	2,336	498,8
Minimum	1	1	1
Maximum	261	489	130103

RESULT

RFM result:

Three different levels were defined for each index and RFM values were evaluated for each index (1, 2, or 3) prior to RFM analysis. The values of R, F, M and number of customers are as shown in Table 3.

Table.3: *Cross tabulation of RFM indicators*



				Monetary Values		
				1	2	3
Recency Values	1	Frequency Value	1	-	-	-
			2	56146	56130	56170
			3	22015	21936	21933
	2	Frequency Values	1	-	-	-
			2	53050	52629	52245
			3	25173	25023	25068
	3	Frequency Values	1	39156	39246	39120
			2	16208	16206	16134
			3	22147	22145	22152

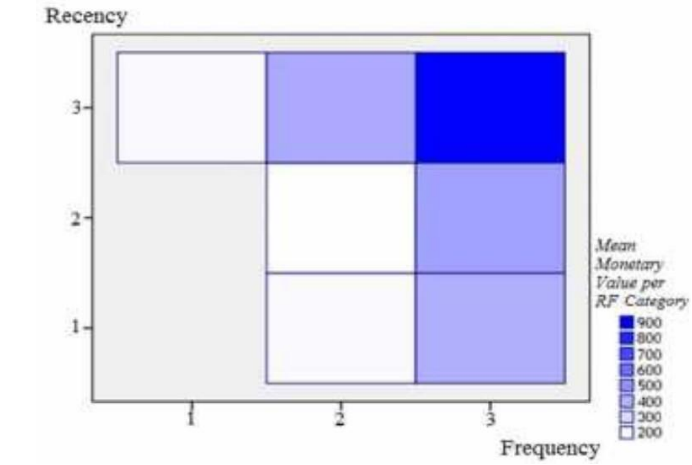


Fig.2: *RFM heat map*

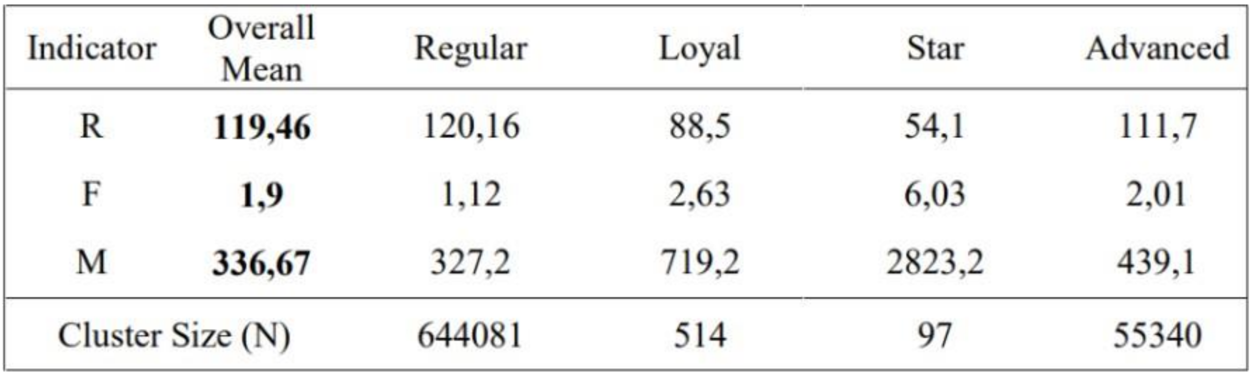
Represent RFM heat map which is creating categories according to R and F scores

According to the heat map (Figure 2), several subgroups with different R, F and M values are distinguished. He has customers with high R and F values and high cost (top is defined as left box/darkest). There is also another group with an F score of 3, but relatively low R scores of 1 and 2, but above average cost. In this figure, it is easy to observe the traces for several clusters related to shoppers' buying behavior.

Proposed model: K-means clustering analysis

The analysis of K-means aims to build clusters considering the R, F and M indicators of the proposed model. As already mentioned, the number of clusters should be defined as k-means. Many values of k were tested (between 2 and 8) and the best solutions were evaluated for k=4. Clusters derived from k-means cluster analysis were named according to their RFM scores. The first cluster, marked "normal", consisted of 644,081 customers, 92% of Russia's total population. Clients in this cluster scored below the overall average across all metrics. Members of this cluster appear likely to be one-time buyers. The F-score is almost 1. Her second cluster, labeled Loyal, contains 514 customers. Her RFM scores for customers in her cluster were better than overall. Clients in the third cluster, called "stars", scored excellent in all indicators. Your company only has customers with this RFM rating. This cluster contains a total of 97 clients. His fourth group, labeled "Advanced" (55,340 clients), also had better RFM scores than the entire group. However, their RFM scores are lower than regular and senior clients, so on average they are very close to his scores. The results are summarized in Table 4.

***Table.4:**-Result of RFM*



Indicator	Overall Mean	Regular	Loyal	Star	Advanced
R	**119,46**	120,16	88,5	54,1	111,7
F	**1,9**	1,12	2,63	6,03	2,01
M	**336,67**	327,2	719,2	2823,2	439,1
Cluster Size (N)		644081	514	97	55340

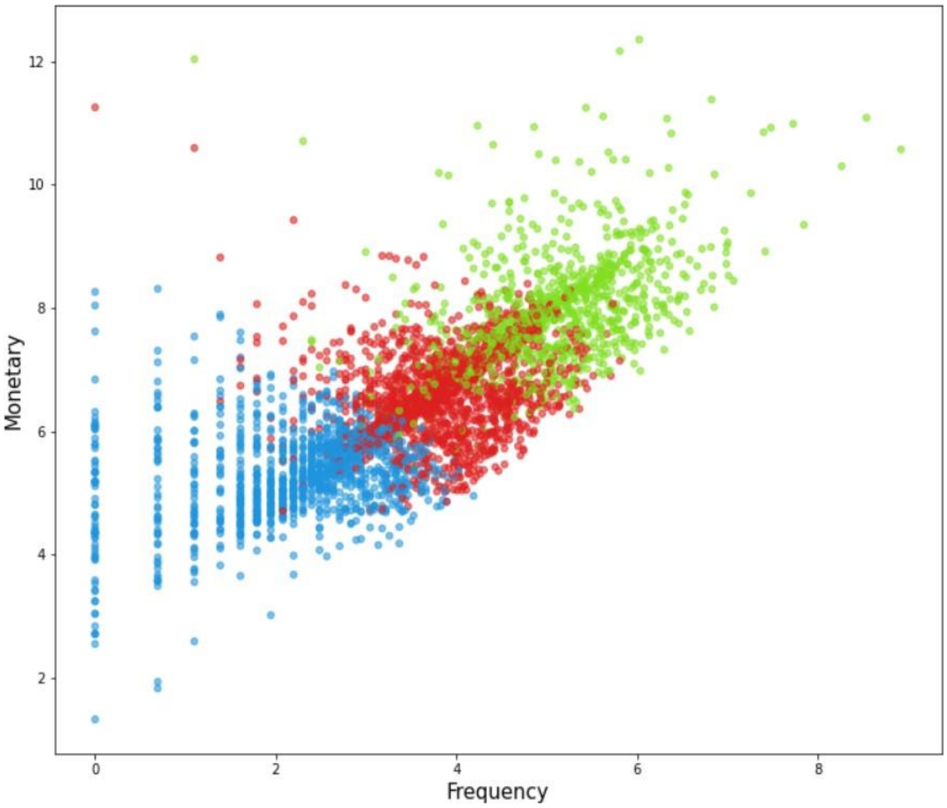
CONCLUSION

Organizations need to better understand the components. It is especially important for businesses to have detailed information about their customers' characteristics, behaviors, demographics, and more. A number of methods have been developed in this context. Several models and algorithms have been used to classify clients. Using these models and algorithms, companies can get a complete picture of their customers. By grouping customers based on data, companies can develop specific strategies that are right for them.

This study recommended Two client segmentation models were suggested by this study for Turkish retailers. Based on client value, the corporation has already segmented its customer base. In case studies, this strategy has been seen in action. Customers are categorized based on their spending by Etzion et al. (2004) to assess their value. However, more recent research contends that categorizing consumers based solely on their spending is insufficient (e.g., Coussement et al., Sarvari et al., 2016; Ansari &amp; Riasi, 2016). We suggested consumer segmentation as a result, utilizing age, frequency, and currency as markers of client clustering. Models for client segmentation were suggested. the clustering technique k-means. clustering method.

The suggested model suggests four distinct groupings. One of them has 644081 customers. Because these consumers' RFM values are near to the average, the business can classify them as typical clients. Otherwise, the business can opt against providing cards or memberships to these clients, as the majority of them are one-time clients. But some research indicates that those with loyalty cards spend more than people without them (Benavent et al., 2000; Liu, 2007). Therefore, offering all card kinds to such clients or creating customer categories may be advantageous.

For ecommerce companies, better customer segmentation is very important. Grouping customers with similar needs, wants, and behaviors can help companies better understand their target market. In this way, businesses can: Set up marketing, price management and promotions.



REFERENCES

1. Abirami, M. & Pattabiraman, V. (2016). Data mining approach for intelligent customer behavior analysis for a retail store. In: proceedings of the 3rd international symposium on big data and cloud computing challenges (ISBCC–16). 283-291.
2. Ansari, A. & Riasi, A. (2016). Taxonomy of marketing strategies using bank customers' clustering. International Journal of Business and Management, 11(7), 106-119.
3. Bowen, J. T. and Chen, S. (2001). The relationship between customer loyalty and customer satisfaction. International Journal of Contemporary Hospitality Management, 13(5), 213-217.
4. Bulut, Z. A. (2015). Determinants of repurchase intention in online shopping: a Turkish consumers' perspective. International Journal of Business and Social Science, 6(10), 55-63.
5. Chen, Y. L., Kuo, M. H., Wu, S. Y. & Tang, K. (2009). Discovering recency, frequency, and monetary (RFM) sequential patterns from customers' purchasing data. Electronic Commerce Research and Applications, 8(5), 241-25.
6. Fotaki G, Gkerpini N & Triantou A I 2012 Online customer engagement management (Netherland: Utrecht University)
7. Muzumdar P 2012 Online bookstore - A new trend in textbook sales management for services
8. Deepak, N. R., & Balaji, S. (2016, April). Uplink Channel Performance and Implementation of Software for Image Communication in 4G Network. In *Computer Science On-*

line Conference (pp. 105-115). Springer, Cham.

9. Thiagarajan, R., Balajivijayan, V., Krishnamoorthy, R., & Mohan, I. (2022). A robust, scalable, and energy-efficient routing strategy for UWSN using a Novel Vector-based Forwarding routing protocol. *Journal of Circuits, Systems and Computers*.
10. NR, D., GK, S., & Kumar Pareek, D. (2022). A Framework for Food recognition and predicting its Nutritional value through Convolution neural network.
11. Thanuja, N., & Deepak, N. R. (2021, April). A convenient machine learning model for cyber security. In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 284-290). IEEE.
12. Shanmugam, P., Venkateswarulu, B., Dharmadurai, R., Ranganathan, T., Indiran, M., & Nanjappan, M. (2022). Electro search optimization based long short- term memory network for mobile malware detection. *Concurrency and Computation: Practice and Experience*, *34*(19), e7044.
13. Deepak, N. R., GK, S., & Bhagappa (2021, Nov). The Smart Sailing Robot for Navigational Investigation is Used to Explore all the Details on the Zone of the Water Pura. Indian Journal of Signal Processing (IJSP), 1(4).
14. Deepak, N. R., & Thanuja, N. Smart City for Future: Design of Data Acquisition Method using Threshold Concept Technique.
15. Kiran, M. P., & Deepak, N. R. (2021, May). Crop prediction based on influencing parameters for different states in india-the data mining approach. In *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 1785-1791). IEEE.
16. Deepak, N. R., & Balaji, S. (2015, December). Performance analysis of MIMO-based transmission techniques for image quality in 4G wireless network. In *2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)* (pp. 1-5). IEEE.