ICISE2024
10th International conference on Industrial and Systems Engineering
18 & 19 Sep. 2024 • Ferdowsi University of Mashhad
دهمین کنفرانس بین‌المللی مهندسی صنایع و سیستم‌ها
۲۸ و ۲۹ شهریورماه ۱۴۰۳ • دانشگاه فردوسی مشهد

Comparison of Two Customer Segmentation Methods (Case Study: Customer Data from GreenWeb Company)

Mehdi Mohammadi
CEO of GreenWeb Co.
Mashhad, Iran
manager@greenweb.ir

Mehdi Jabbari Nooghabi*
Department of Data Analysis & BI
GreenWeb Co.
Mashhad, Iran,
Department of Statistics
Ferdowsi University of Mashhad
Mashhad, Iran
jabbarinm@um.ac.ir

Sanaz Nia
Department of Data Analysis & BI
GreenWeb Co.
Mashhad, Iran
s.nia@greenweb.ir

Rozhin Sharifi
Department of Data Analysis & BI
GreenWeb Co.
Mashhad, Iran
ro.sharifi@greenweb.ir

Ameneh Mohammadi
Department of Data Analysis & BI
GreenWeb Co.
Mashhad, Iran
a.mohammadi@greenweb.ir

***Abstract*—In the world of competition between companies, it is required to cluster the customers based on the behaviors and preferences and it is a critical strategic imperative in any company. In this research, the well-known methods of RFM and Clara clustering based on Manhattan and Euclidean measures are used to cluster the customers of GreenWeb company. The results show the frequency of customers in some of the clusters are slightly equal whereas the number of clustering in each three methods are not equal.**

Keywords—RFM, Clara, k -medoids, GreenWeb, Customer Segmentation

I. Introduction

In today's dynamic marketplace, where competition is fierce and consumer preferences are ever-evolving, businesses face the formidable challenge of not only attracting customers but also retaining them. In this pursuit, understanding the diverse needs, behaviors, and preferences of consumers has emerged as a critical strategic imperative. Customer segmentation, the process of dividing a heterogeneous market into distinct groups based on shared characteristics, lies at the heart of this endeavor. There are diverse clustering methods for customer segmentation one of which is RFM method.

The RFM (Recency, Frequency, Monetary) method is a customer segmentation technique used in marketing analysis, where customers are categorized based on their transactional behavior[1]. It evaluates three primary metrics: recency, which measures how recently a customer made a purchase; frequency, indicating how often they make purchases within a defined timeframe; and monetary, representing the amount of money spent. By assigning numerical scores to these metrics for each customer and segmenting them accordingly, businesses can identify distinct customer segments, such as high-value, loyal customers or those at risk of attrition. RFM analysis enables targeted marketing efforts, personalized customer experiences, and effective customer retention strategies by focusing on the specific needs and behaviors of different customer segments. In order to evaluate the effectiveness of RFM techniques, it's valuable to compare the clustering results obtained from the RFM method with those derived from other popular clustering methods such as Clara.

Clara [1] is a method tailored for clustering large datasets efficiently. It operates by first sampling subsets of the data, applying K-medoids clustering to each sample to select representative medoids. These medoids are then used to assign all data points to clusters. The process iteratively refines the medoids by swapping them with non-medoid points until no further improvement is possible. By working with smaller samples, Clara reduces computational complexity while still delivering robust clustering results. This approach is particularly useful for situations where

*Corresponding Author
[1] Clustering Large applications

computational resources are limited or datasets are too large to process using traditional clustering methods. By comparing the clustering outcomes of these two methods, businesses can gain insights into the effectiveness of different segmentation approaches. In this paper the aim is to compare the clustering results of RFM and Clara methods.

II. Literature review

A. *RFM*

Customer segmentation is crucial for businesses aiming to navigate the complexities of diverse consumer landscapes. For classification of customers, they are grouped according to any type of variable, which can generally be divided into two categories: general variables and product-specific variables. The first group include their characteristics (e.g. sex, age, income, education level, etc.) and lifestyles. The second group include customer behaviours of purchasing (e.g. frequency of purchase, consumption, spending, etc.) and intentions. The general variables are always can easily be obtained, but for the capturing purchase behaviours of the customers the product-specific variables are more important. These variables are mostly used to distinguish customer participations in a business [2].

We have different methods for customer segmentation, one of the simplest to implement is RFM (Recency, frequency, and monetary). There are several advantages for these models. For example, their results are quickly derived and can be easily explained for the managers and decision makers. These models depict the customer's characteristics by using only a relatively small number of features. In the last ten years, various types of RFM models have mainly a good performance in classification of customers in different industries, including health and beauty [2].

Although there are several questions related to the efficacy of RFM, quantitative research documents its superiority over newer statistical techniques. One reason is that RFM gives a general approach to data-mining; there are different ways to use recency, frequency, and monetary values. Research which is investigated on efficiency of RFM generally focuses on proprietary or judgmental models of RFM models rather than empirically based RFM ones [1]. Recently, research has moved away from RFM and instead focused on newer and more sophisticated approaches to data mining. [1].

Aryuni, Madyatmadja, and Miranda [3] conducted a study applying clustering models to characteristics of customer derived from Internet Banking usage at XYZ Bank. This research utilized both K-Means and K-Medoids clustering methods, focusing on the RFM scores of customers transactions of Internet Banking. The effectiveness of these methods was evaluated and compared. Results indicated that the K-Means outperformed the K-Medoids in terms of intra-cluster distance. Additionally, according to the Davies-Bouldin index, K-Means demonstrated a slight performance advantage over K-Medoids.

Rahim, Mushafiq, and Khan [4] utilized the RFM model along with various data modelling techniques to identify behaviour patterns. They tested their proposed approach on a publicly available real-world dataset, employing customized machine learning techniques such as decision tree classification (DTC), support vector machine (SVM), and multi-layer perceptron (MLP) methods. The results revealed a high rate of customer classification, exceeding 97 percent across various customer segments. Additionally, based on the empirical analysis to accurately classify a customer, it is enough to use a few as eight transactions.

Shirole, Madyatmadja, and Jadhav [5] used a transaction data of a UK online retail store and fit a RFM model according to the K-means algorithm. The customers characteristics made four clusters. In this classification, class A has the highest revenue and Class D has the least. For the given dataset, silhouette index is 0.442 and it is in a good level. According to the results of this study, one can develop market strategies and promotional medium to their loyal customers can be useful.

Tavakoli, Ghanavati-Nejad, and Tajally[6] applied a hybrid clustering-rules approach to create policies for managing patients throughout their treatment. Using the LRFM and K-means algorithms, they clustered patients into groups. To explore connections between procedure groups, they applied the APRIORI algorithm to identify and analyse the most frequent sequential rules, working closely with company experts. Based on the characteristics of the patient groups and the discovered rules, they developed several policy scenarios.

B. *Clara (Manhattan, Euclidean)*

Clara is a method for cluster analysis and much extended the original of earlier method which is based on Kaufman and Rousseeuw[7]. This method performs the metric to be used for calculating dissimilarities between observations. Some of the metrics are Manhattan, Eucliden, etc.

III. Methodology and data explanation

A. *RFM*

A table can be used to represent customer purchases, with columns detailing the customer's name, the purchase date, and the amount spent. Fader and Hardie [8] describe various methods for quantitatively defining RFM values. The most effective methods will vary based on the customer journey and the business model. One way to handle RFM is by assigning a score to each dimension using a binary scale (0, 1), where a score of 1 indicates the desired behavior. A formula can then be applied to compute the three scores for each customer. For instance, a service-oriented business might utilize the following calculations:

Recency = 1 if the number of days remaining until the end of the study period (12 months) is below the median for all customers; otherwise, it is set to 0.

Frequency = 1 if the total number of purchases made by the customer in the past 12 months exceeds the median for all customers; otherwise, it is marked as 0.

Monetary = 1 if the customer's average purchase value surpasses the median for all customers; otherwise, it is assigned a 0.

Based on this approach we have eight class of customers, Best, Valuable, Churn, Shopper, Spender, Frequent, First time and Uncertain. In Table 1 the definition of each class is described.

Table I. RFM Labels definition

	Best	Valuable	Shopper	First Time	Churn	Frequent	Spender	Uncertain
Recency	1	1	0	0	1	0	1	0
Frequency	1	0	1	0	1	1	0	0
Monetary	1	1	1	1	0	0	0	0

B. Clara (Manhattan, Euclidean)

The Clara algorithm, developed by Kaufmann and Rousseeuw [7], is a partitioning technique that involves taking several samples from the dataset and applying the k-medoids algorithm to each sample. The best clustering result is then selected as the final output. The k-medoids algorithm follows these steps:

1. Randomly choose k out of the n data points to serve as the medoids.
2. Assign each data point to the nearest medoid.
3. For every pair of non-selected object h and selected object i, determine the total swapping cost TC_{ih}. This cost is calculated using the Minkowski distance, with Euclidean and Manhattan distances being the most commonly used. For every pair of i and h,
    1. If $\mathrm{TC}_{ih} < 0$, replace i with h,
    2. Next assign each object that wasn't selected to the representative object it most closely resembles,
    3. Repeat steps 2 and 3 until no further changes occur,

where

$$\mathrm{TC}_{ih} = \sum_{i=1}^{k}(\sum_{h=1}^{n}|m_{ih} - x_{ih}|^{q})^{1/q}. \quad (1)$$

One should note that for Euclidean distance, q is equal to 2, and when $q = 1$, Manhattan distance was considered.

C. Data (case study)

The dataset is taken form an IT-Based company, GreenWeb, customers. This company mainly offers services such as Cloud Hosting, Web Applications, UI/UX Design, Mobile Applications and online integrative services. The dataset contains customer transactions in 2023 and has three features recency, frequency and monetary

IV. Result

A. Data summary

The dataset is composed of customer transactions of GreenWeb company. The features are recency (R), frequency (F) and monetary (M) of customers in 2023. In Figure 1 and Figure 2, histogram and boxplot of customer's recency is depicted, respectively. Also, histogram and boxplot of customer's frequency are respectively shown in Figure 3 and Figure 4. Considering the state of monetary feature, histogram and boxplot of the customer's monetary are brought in Figures 5 and 6, respectively. Summary of descriptive statistics of the dataset including minimum, maximum, median, mean, first, and third quartiles are shown in Table 2.

Table II. Descriptive statistics of R, F, and M

Summary	R (Day)	F	M (Rial)
Minimum	0	1	0.000e+00
1st Quartile	43	1	2.499e+06
Median	134.0	2	5.875e+06
Mean	152.5	3.82	8.248e+06
3rd Quartile	256.0	3	8.066e+06
Maximum	365	2646	1.626e+09

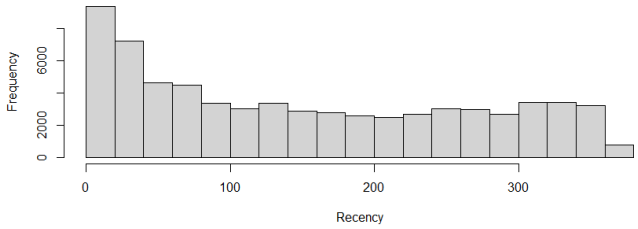

Fig. 1. Frequency distribution of R

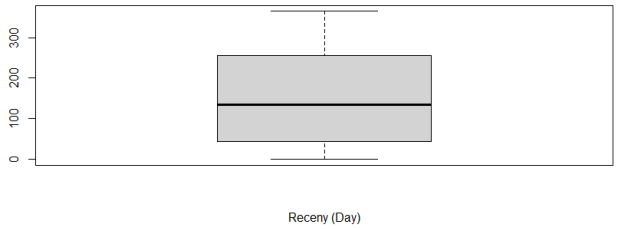

Fig. 1. Boxplot of R

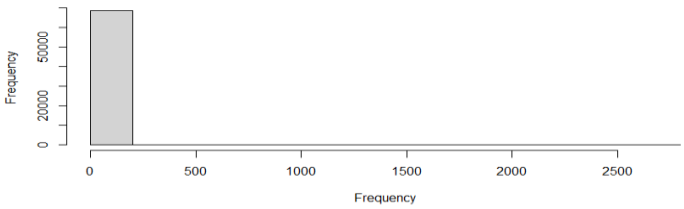

Fig. 2. Frequency distribution F

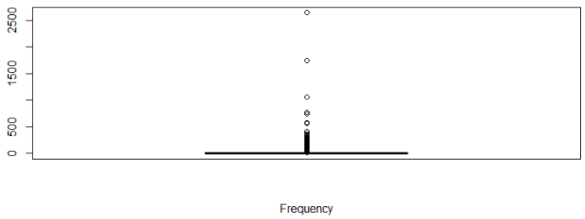

Fig. 3. Boxplot of F

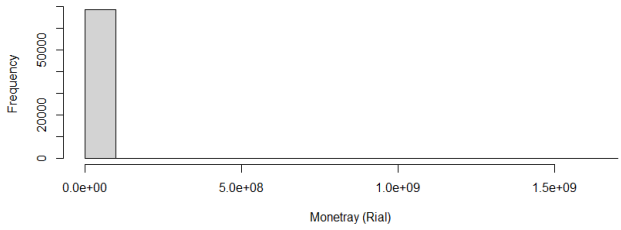

Fig. 4. Frequency distribution M

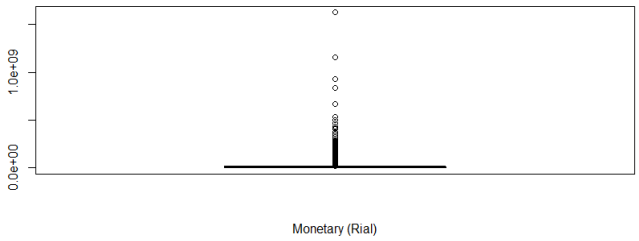

Fig 5. Boxplot of M

Tables 3 and 4 show the result of clustering based on the clara methods for Euclidean and Manhattan distances, respectively. According to the Tables, it has been seen that, there are only 5 clusters when distance is Euclidean and there are only 6 clusters for Manhattan distance. It means that some of the cases are not distributed in all 8 clusters which are define by the RFM methods.

Table III . Results of the centres of clustering based on Euclidean metric

Clara-Euclidean	N	RDate	FCount	MAmount	RFM-Class	Percent
1	25435	116	1	6202100	Valuable	36.977%
2	2615	116	1	24143500	Valuable	3.802%
3	1950	178	4	38741300	Churn	2.835%
4	14752	149	5	2779500	Frequent	21.446%
5	3745	37	2	17843300	Valuable	5.444%
6	7062	119	3	11205200	Best	10.267%
7	12623	239	1	861100	Uncertain	18.351%
8	604	14	6	70922704	Best	0.878%

Table IV. Results of the centres of clustering based on Manhattan metric

Clara-Manhattan	N	RDate	FCount	MAmount	RFM-Class	Percent
1	16166	89	1	6420100	Valuable	23.502%
2	10844	24	9	5499660	Shopper	15.765%
3	2167	178	4	38741300	Churn	3.150%
4	5160	70	1	20764500	Valuable	7.502%
5	7701	108	1	11325100	Valuable	11.196%
6	13521	149	5	2779500	Frequent	19.657%
7	12623	239	1	861100	Uncertain	18.351%
8	604	14	6	70922704	Best	0.878%

Table V. Frequency distribution of clusters in three methods

	Best	Valuable	Shopper	First Time	Churn	Frequent	Spender	Uncertain
RFM	11.809%	13.534%	12.465%	12.440%	3.031%	5.860%	18.238%	22.622%
Clara (Euclidean)	11.145%	46.223%	0.000%	0.000%	2.835%	21.446%	0.000%	18.351%
Clara (Manhattan)	0.878%	42.199%	15.765%	0.000%	3.150%	19.657%	0.000%	18.351%

To compare the association and agreement of the Clara method with the RFM, results of Pearson's Chi-squared test in crosstabulation as well as the Kappa measures of agreement are shown in Tables 6-9.

Table VI. Pearson Chi-squared test results and accuracy measure of comparing Clara (Euclidean) and RFM

Clara (Euclidean) and RFM	X-squared	df	p-value	Total Accuracy
Pearson's Chi-squared test	7501.8	28	< 2.2e-16	0.1255633

Table VII. Kappa measures of agreement for Clara (Euclidean) and RFM

Kappa: Clara (Euclidean) and RFM	value	ASE	z	Pr(>|z|)
Unweighted	-0.005868	0.001328	-4.417	9.996e-06
Weighted	0.054517	0.002517	21.656	5.381e-104

Table VIII. Pearson Chi-squared test results and accuracy measure of comparing Clara (Manhattan) and RFM

Clara (Manhattan) and RFM	X-squared	df	p-value	Total Accuracy
Pearson's Chi-squared test	8094.8	35	< 2.2e-16	0.1152124

Table IX. Kappa measures of agreement for Clara (Manhattan) and RFM

Kappa: Clara (Euclidean) and RFM	value	ASE	z	Pr (>|z|)
Unweighted	-0.01910	0.001311	-14.57	4.684e-48
Weighted	0.03743	0.002184	17.14	7.505e-66

V. Conclusion

Understanding the diverse needs, behaviors, and preferences of consumers is a compulsory approach as a critical strategic imperative in any company. So, classification and clustering of the customers was considered based on the well-known methods of RFM and Clara clustering based on Manhattan and Euclidean measures. It is concluded that the number of clustering in each three methods are not equal, but the frequency of customers in some of the clusters are slightly equal. Also, there are no association and agreement between the RFM method and each of two clustering methods of Clara. So, based on the exact object of clustering of the customers, one should select the better method to classify the customers.

References

[1] J. A. McCarty and M. Hastak, "Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression," *Journal of Business Research*, vol. 60, no. 6, pp. 656–662, 2007.

[2] S. Peker, A. Kocyigit, and P. E. Eren, "LRFMP model for customer segmentation in the grocery retail industry: A case study," *Marketing Intelligence & Planning*, vol. 35, no. 4, pp. 544–559, 2017.I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[3] M. Aryuni, E. D. Madyatmadja, and E. Miranda, "Customer segmentation in XYZ bank using K-means and K-medoids clustering," in *2018 International Conference on Information Management and Technology (ICIMTech)*, SepR. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[4] M. A. Rahim, M. Mushafiq, S. Khan, and Z. A. Arain, "RFM-based repurchase behavior for customer classification and segmentation," *Journal of Retailing and Consumer Services*, in 2021.M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[5] R. Shirole, L. Salokhe, and S. Jadhav, "Customer segmentation using RFM model and k-means clustering," *Int. J. Sci. Res. Sci. Technol.*, vol. 8, pp. 591–597, 2021.

[6] M. Tavakoli, M. Ghanavati-Nejad, A. Tajally, and M. Sheikhalishahi, "LRFM-based association rule mining for dentistry services patterns identification (case study: a dental center in Iran)," *Soft Computing*, vol. 28, no. M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[7] L. K. P. J. Rdusseeun and P. Kaufman, "Clustering by means of medoids," in *Proceedings of the Statistical Data Analysis Based on the L1 Norm Conference*, Neuchatel, Switzerland, vol. 31, Aug. 1987.

[8] P. S. Fader, B. G. Hardie, and K. L. Lee, "RFM and CLV: Using iso-value curves for customer base analysis," Journal of Marketing Research, vol. 42, no. 4, pp. 415-430, 2005.