Accredited Ranking SINTA 2
Decree of the Director General of Higher Education, Research and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online on: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 7 No. 6 (2023) 1430 - 1438 ISSN Media Electronic: 2580-0760

Comparison of the RFM Model's Actual Value and Score Value for Clustering

Samidi[1], Ronal Yulyanto Suladi[2], Dewi Kusumaningsih[3]
[1,2,3]Master of Computer Science, Faculty of Information Technology, Universitas Budi Luhur, Jakarta, Indonesia
[1]samidi@budiluhur.ac.id, [2]surel.ronal@gmail.com, [3]dewi.kusumaningsih@budiluhur.ac.id

Abstract

Clustering algorithms and Recency-Frequency-Monetery (RFM) models are widely implemented in various sectors of e-commerce, banking, telecommunications, and other industries to obtain customer segmentation. The RFM model will assess a line of data which includes the recency and frequency of data appearance as well as the monetary value of a transaction made by a customer. Choosing the right RFM model also influences the analysis of cluster results, the output of cluster results is more compact for the same clusters (inter-cluster) and separate for other clusters (intra-cluster). Through an experimental approach, this research aims to find the best dataset transformation model between actual RFM values and RFM scores. The method used is to compare the actual RFM value model and the RFM score and use the silhouette score value as an indicator to get the best clustering results using the K-Means algorithm. The subject of this research is a stall-based e-commerce application, where data was taken in the Wiradesa area, Central Java. The resulting dataset consisted of 273,454 rows with 18 attributes from January 2022 to December 2022 through collecting historical data from shopping outlets to wholesalers. Analysis of the dataset was carried out by transforming the dataset using the RFM method into actual values and score values, then the dataset was used to obtain the best cluster data. The results of this research show that transaction data based on time (time series) can be transformed into data in the RFM model where the RFM model's actual value is better than the RFM score model with a silhouette score = 0.624646 and the number of clusters (K) =3. The results of the clustering process also form a series of data with a cluster label, thus forming supervised learning data.

Keywords: RFM model; RFM actual value; RFM core value; clustering

1. Introduction

The commercial industry has a goal to optimize return on investment in several ways, such as through acquisitions by influencing and attracting new customers or by retaining existing customers by providing new offers and products to increase the revenue [1]. According to Pareto, of all customers owned by a company, only 20% (one-fifth) of the total number of customers contribute more to the company's revenue than other customers [1]. The customers have diverse and different priority tendencies, for instance the customer grouping or segmentation is considered one of the best ways to manage and understand customers [2], [3]. On top of that, the customers have diverse and different priority tendencies; therefore, customer grouping or segmentation is considered one of the best ways to manage and understand customers [4], [5].

Various studies have been carried out on customer groups known as customer segmentation, where this segmentation tries to group customers based on certain similar characteristics. Grouping or segmenting customers using data mining is one of the things that can provide an advantage for an organization to analyze customer behavior and other matters related to relationships [6].

The RFM model is a behavior-based model that is used to analyze customer behavior and then make predictions based on the behavior database [7], The RFM model classifies customer segmentation based on recency (when was the last transaction made?), frequency (how often did the customer make a transaction?), and monetary (the value of transactions made) [8], and the ability of the RFM model has been widely used to analyze customer values combined with clustering techniques [9].

The application of the RFM model with a score model and actual value is used in various industrial sectors as a combination of clustering techniques and CLV (customer lifetime value) analysis. The RFM score

Accepted: 06-09-2023 | Received in revised: 25-12-2023 | Published: 28-12-2023

model has several score calculation techniques, for example, the customer quintile method and the behavior quintile method [7]. At the same time, the actual RFM uses the technique of combining the total value (sum/count), average (mean), min, max, and median [10], which is then analyzed with RFM based on the average for each attribute R, F, and M, so that each attribute can be marked with a symbol (↑) when the attribute value is above the average value (high) and marked with a symbol (↓) when the attribute value is below the average (low) [11]. While the RFM actual value model generally carries out the normalization process with the standard scaler/z-score technique in scaling the R, F, and M attribute values, replacing the scoring technique carried out by the RFM model score [12].

The clustering technique that is commonly used to obtain customer segmentation or grouping uses a clustering algorithm. Clustering algorithms such as K-MEANS, Agglomerative, and DBSCAN are algorithms that group data into several groups based on the similarity of the data, so that data with similar attribute characteristics are grouped in one cluster (homogeneous), while data with different attribute characteristics (heterogeneous) are grouped in another different cluster. The application of clustering with various comparisons of cluster algorithms and RFM models has been widely carried out in various fields, for example, online retail data [13], data e-commerce [14], banking transaction data [15], and telecommunication company transaction data [16].

The previous research stated that the RFM model used or selected as input in the clustering algorithm process has an influence on the quality of the cluster results [7],[8],[9]. The quality of the cluster results is calculated based on one of the cluster validation methods, sum square error [17]. In addition, the selection of the right RFM model also influences the analysis of cluster results; the output of cluster results is more compact for fellow clusters (inter-cluster) and separate for other clusters (intra-cluster) [18].

The object of this research is to develop an e-commerce platform that can be used to accommodate the needs of the traditional retail (outlet) ecosystem. The platform connects retailers and outlets with wholesalers in the same sub-district area, where wholesalers register all the products, and then the outlets are used to carry out shopping transactions for their product by accessing this platform digitally. To increase salespersons' efficacy in visiting active merchants and meeting retail priorities and demands, this e-commerce platform must group the current retail environment. Currently, salespeople visit the location based solely on retail demand and without regard to priority, which prevents retailers from meeting their growth ambitions

This study conducted the comparison of the actual value of the RFM model and the RFM score. The value of the analysis of the comparison of the value of the RFM model is based on the cluster validation value, using one of the clustering algorithms to obtain the validation value of the cluster results. In contrast the elbow method is used in determining the best number of clusters [19]. The dataset used in the formation of the RFM model is the outlet to wholesale shopping transaction history dataset that is queried from the e-commerce platform, with a total of 273,454 transactions with 18 attributes from January 2022 to December 2022. This study retrieved transaction data from one district, namely Wiradesa in the district of Pekalongan, Central Java. The RFM model with the best cluster validation is used as an appropriate input for the clustering model. The cluster output is then interpreted based on RFM segmentation analysis to get more interesting information and knowledge compared to just using cluster parameters [20]. With the aim of making the interpretation of clustering deeper and more varied as suggestions and recommendations for the business domain.

2. Research Methods

This research goes through the stages shown in figure 1.

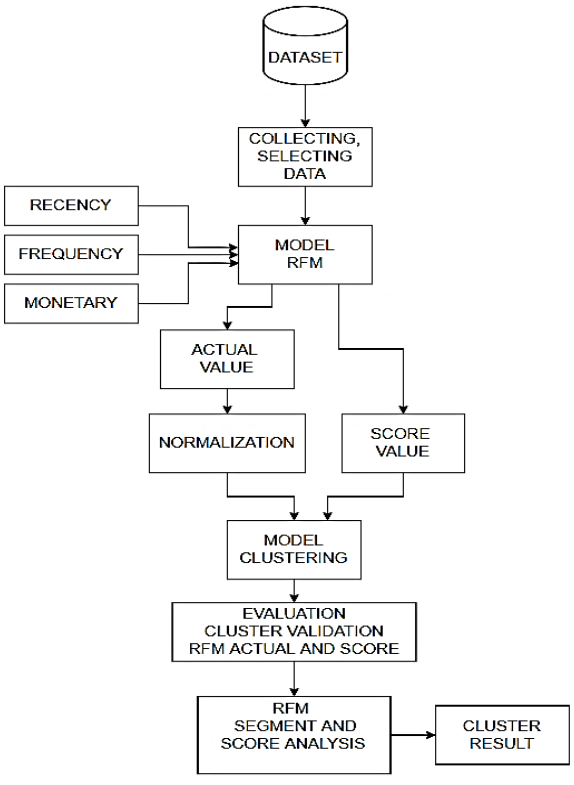

Figure 1. Research Framework

Collect and Select Data, collecting and selecting data and information sourced from literature studies, reading

DOI: https://doi.org/10.29207/resti.v7i6.5416

and studying research related to research topics, observing research objects, viewing and understanding outlet shopping transaction data by querying databases, and systematically recording and observing problems that are examined regarding the research object with the aim of obtaining data as input in the RFM model process.

Formation of the RFM Model, historical transaction data serves as a data source for the RFM model, which is based on earlier research by [2], [10], [17] and others. This research uses historical outlet shopping transaction data for 12 months (January–December) in 2022.

RFM Actual Value, the RFM model describes customer consumption behavior based on past transaction databases in a simplified form into three attributes [2] namely Recency (R), Frequency (F) and Monetary Value (M). Recency (R), also known as the range of one transaction at a specific time in the past, is what it stands for. The shorter the interval, the greater the R value. Frequency (F) represents frequency, namely the number of transactions in a certain period at a certain period, for example, twice in one year or twice in one month. The higher the frequency, the greater the F value. Monetary Value (M) represents monetary value, namely the value of the product in the form of money in a certain period. The greater the amount of money in that period, the higher the value of M.

Figure 2 shows the RFM actual value model diagram:

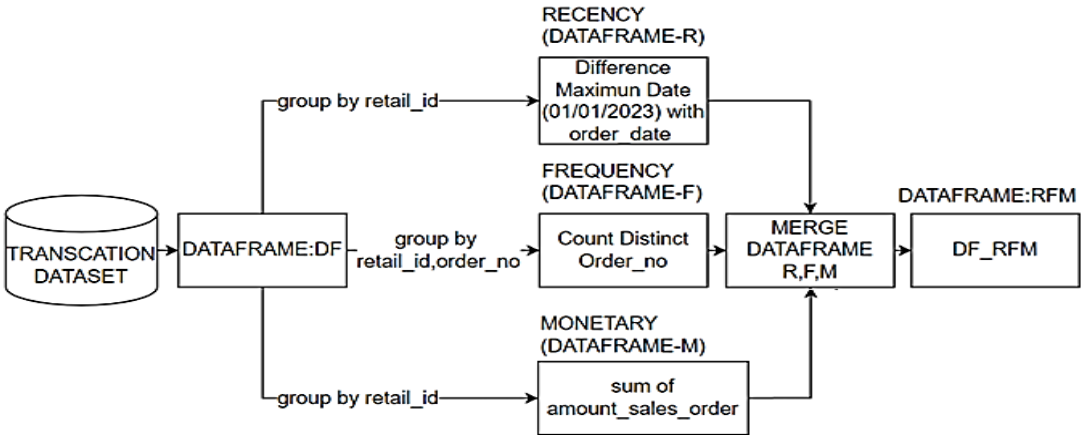

Figure 2. RFM Actual Value Model Diagram

The results of the process of forming a dataset into the RFM model are stored in a data frame with the name DF_RFM.

RFM Score Value, it is an RFM model that transforms RFM values into a quantitative score; the steps are [17]: Sort the dataset descending by attribute R from the earliest date to the oldest; Divide the dataset into 5 quartiles and give a value of 5 for the first 20% of the dataset, a value of 4 for the second 20% of the dataset, and so on until a value of 1; Repeat steps a and b for attributes F and M by sorting F and M in descending order and assigning values; Sort F in each category R and sort M in each combination of categories R and F.

This model will produce RFM segmentation with the criteria and scoring [2], [20], which are then used in the RFM analysis as shown in Table 1.

Standard Scaler Normalization, normalization is carried out so that the range (scale) of recency, frequency, and monetary data values do not differ much. In this study, normalization uses standardization or z-score normalization, where the normalization process is based on the mean and standard deviation as shown in Formula 1[21].

Table 1. RFM Scoring

Criteria	Recency Score	Frequency and Monetary Score
Champions	5	4-5
Loyal customers	3-4	4-5
Potential loyalists	4-5	2-3
Promising	4	1
Can't lose them	1-2	5
At risk	1-2	3-4
About to sleep	3	1-2
Hibernating	1-2	1-2
New customers	5	1
Need attention	3	3

$$v' = \frac{v - \mu A}{S} \quad (1)$$

μ is the mean, v is the values, s is the standard deviation. For example: What is the z-score of 73600 if μ = 54000 and s = 16000? Then *v'*: (73600-54000)/16000 = 1.255.

Each attribute R, F, and M with actual values will be normalized using the standard scaler technique; the mean is point 0, and the maximum value is the standard deviation value.

The K-Means algorithm is a clustering algorithm that is most widely used in data grouping processes in various industrial and scientific fields such as in marketing, computer vision, and geo-statistics. The advantages of

DOI: https://doi.org/10.29207/resti.v7i6.5416

K-Means are that , the K-Means simple and easy to implement, but has a relatively fast processing speed. On top of that the algorithm very good in processing quantitative data with numerical attributes and efficient use of computing resources [19], [22], [23].

K-Means Clustering Algorithm, the K-Means algorithm is used to cluster or segment outlet shopping transaction data based on the RFM model. In this research, the clustering process was carried out seven times (2–8 clusters). The steps taken in the clustering process were: Determine the number of clusters, which will make it easier to define shopping transaction patterns in outlet segmentation; Determine the initial centroid value by taking random data objects as shown in Formula 2.

$$V_{ij} = \frac{1}{N_i}\sum_{k=0}^{N_i} X_{kj} \quad (2)$$

Vij is The i^{th} cluster centroid for the j^{th} variable, *Ni* is the amount of data that is a member of the *i* cluster, *i* and *k* is the index of the cluster, *j* is the index of the variable, *Xkj* is The k^{th} data value in the cluster for the j^{th} variable

Calculate the distance between the centroid point and each object point as shown in Formula 3

$$D_e = \sqrt{(x_i - s_i)^2 + (y_t - t_t)^2} \quad (3)$$

D_e as euclidean distance, *i* as the amount of data, *(x,y)* as data coordinates and *(s,t)* as centroid coordinates

The closeness of two objects is determined based on the distance between the two objects. Likewise, the proximity of data to a particular cluster is determined by the distance between the data and the center of the cluster. In this stage, it is necessary to calculate the distance of each data point to each cluster center. To calculate the distance from the object to the cluster at this stage, use the Euclidean distance formula [22]. The closest distance between one piece of data and one particular cluster will determine which piece of data belongs to which cluster. Which are cluster or to the new centroid and allocate all objects to the closest cluster to the new centroid. If there are objects that move clusters, repeat step 2 again and if no objects move clusters, then the clustering process is complete.

Evaluation of Cluster, evaluation of K-MEANS cluster results using the silhouette index (SI). This method is a validity criterion based on geometric considerations of cohesion, which functions to measure how close the relations are between objects in a cluster, and the separation method, which functions to measure how far a cluster is separated from the cluster. others[23]. The formula used to obtain the silhouette index value is shown in Formula 4

$$S_i = \frac{b_i - a_i}{\max\{a_i, b_i\}} \quad (4)$$

S_i as the silhouette coefficient value, a_i as the average distance between point i and all points in a (the cluster where point a is), b_i as the average distance between point *i* and all points in the cluster other than *a*.

RFM Score Analysis, perform an analysis based on the score that has been given, assigning a score to each retail_id for the recency, frequency, and monetary attributes. The score is worth a scale between 5 and 1. The highest value is 5, and the next is 4, 3, 2, 1 [2]. In Table 2, the RFM analysis segments are shown [20]:

Table 2. RFM Segment Analysis

Criteria	Description
Champions	Active customers have recently made transactions, buy frequently, and spend the most.
Loyal customers	Customers who make regular purchases and are responsive to promotions
Potential loyalist	New customers with average frequency
Promising	Customers with recent purchases but who didn't spend a lot of money
Needs attention	Customers with above-average scores for recency, frequency, and monetary
About to sleep	Customers with recency and frequency below average may be hibernating.
At risk	Customers who shopped some time ago and need to be reactivated
Can't lose them	Customers with characteristics in the past frequently made transactions but currently have not made transactions for a long time.
Hibernating	Customers with high recency and low shopping value are likely to become lost customers (inactive customers).

Each attribute R, F, and M will be changed to a value with a range of 1 to 5, according to the table in Table 2.

3. Results and Discussions

In the process of collecting and selecting data, information is needed regarding understanding the running business.

In Figure 3, there is a form of shopping transaction dataset, and in Table 3, there is an explanation:

Table 3. Expenditure Transaction Data Structure

No	Field Name	Description
1	Region	Region Name
2	Subdist_nm	District name (sub-distribution)
3	Retail_id	Outlet ID
4	Retail_name	Outlet Name
5	Wholesaler_id	Wholesale ID
6	Wholesaler_name	Wholesale Name
7	Order_date	Order date
8	Order_no	Order Number
9	Pcode	Product ID
10	Category	Product category
11	Principal	Principal product name
12	Qty_sales_order	Number of transactions per transaction
13	Amount_sales_order	Value-for-money transactions

DOI: https://doi.org/10.29207/resti.v7i6.5416

	region	subdist_nm	rsm	asm	retail_id	retail_name	wholesaler_id	wholesaler_name	order_date	order_no	pcode	pcode_name
0	JAWA TENGAH	Area Pekalongan	Yusuf V	Aryanto	C100179805	HIDAYAH	G00110	BAROKAH-1	2022-04-18	SO2204000002906930	000379	POCARI SWEAT BTL 550ML
1	JAWA TENGAH	Area Pekalongan	Yusuf V	Aryanto	C100179805	HIDAYAH	G00110	BAROKAH-1	2022-04-18	SO2204000002906930	000854	GUDANG GARAM SIGNATURE KRETEK 12
2	JAWA TENGAH	Area Pekalongan	Yusuf V	Aryanto	C100179805	HIDAYAH	G00110	BAROKAH-1	2022-04-18	SO2204000002906930	001169	INDOMILK KID CHOCOLATE 115 ML
3	JAWA TENGAH	Area Pekalongan	Yusuf V	Aryanto	C100179805	HIDAYAH	G00110	BAROKAH-1	2022-04-18	SO2204000002906930	001550	HEMART MINYAK BOTOL 1000 ML
4	JAWA TENGAH	Area Pekalongan	Yusuf V	Aryanto	C100179805	HIDAYAH	G00110	BAROKAH-1	2022-04-18	SO2204000002906930	001944	BALSEM GELIGA 20 GR

Figure 3. Sample of historical shopping transaction data in 2022

In Formation of the RFM model of actual value and score value, attributes are selected for the transaction dataset formed into data aggregation to obtain the RFM value, so that the number of outlets becomes 280 outlets. The RFM value is formed from recency, frequency, and currency. Recency is formed by calculating the difference between the outlet's last transaction time for 12 months and the specified time, namely January 1, 2023.

Frequency is formed by the number of transactions carried out by the outlet, while monetary is formed by the nominal amount spent by the outlet to buy products at wholesalers. Table 4 explains the attributes selected for the transaction dataset:

Table 4. Transaction Data Structure After Attribute Selection

No	Field Name	Description
1	Retail_id	Outlet ID
2	Order_date	Order date
3	Order_no	Order Number
4	Qty_sales_order	Number of transactions per sales order
5	Amount_sales_order	Transaction monetary value

Table 5 is an example of data in the form of RFM actual value and RFM score.

The RFM frame data in Table 5 is normalized using a standard scaler/z-score transformation for each outlet. Meanwhile, Table 6 shows the form of the data frame that has been transformed.

Table 5. RFM Dataframe Model: Actual Values and Scores

retail_id	recency	frequency	monetary	R	F	M	RFM_Segment	RFM_Score
C100000641	3	13	47527750	5	2	3	523	10
C100000953	5	7	986650	4	1	1	411	6
C100002548	2	189	464154545	5	5	5	555	15
C100003179	5	26	24206851	4	3	2	432	9
C100003361	87	30	59794160	1	3	3	133	7
...	...	...	...	...	...	...	...	...
C100324412	25	6	17949450	2	1	2	212	5
C100324446	6	13	49567970	3	2	3	323	8
C100326000	50	3	13341500	2	1	2	212	5
C100327075	62	1	1744600	2	1	1	211	4
C100327998	55	1	4110300	2	1	1	211	4

Table 6. Normalized RFM Frame Data

retail_id	recency_standarscale	Frequency_standarscale	monetary_standarscale
C100000641	-0.552550	-0.536811	-0.344879
C100000953	-0.530903	-0.694918	-0.597927
...	...	...	...
C100327075	0.086048	-0.853025	-0.593806
C100327998	0.010283	-0.853025	-0.580943

In Table 6, we can see that the RFM frame data, which initially had actual values, was normalized using the standard scaler transformation.

Modeling in this research using the K-Means clustering algorithm and Jupyter Notebook tools with parameters and commands as shown in Figure 4.

DOI: https://doi.org/10.29207/resti.v7i6.5416

```
model = KMeans(n_clusters=n_clust, random_state=42)
model.fit(RFM_)
model.labels_.shape
```

Figure 4. K-Means Modeling

The commands and parameters in Figure 4 explain, that the random state = 42 parameter is used as a control generator so that the initial centroid initiation process is always fixed (not random) [26]. To get the optimal number of clusters, the *n*_cluster parameter is the desired number of clusters variable. To get the optimal number of clusters, one can use the elbow and silhouette score methods [6], [27], the elbow method is formed from the results of the difference in SSE values for each number of clusters (2-8). By default, the K-Means model uses Euclidian distance calculations for each cluster, as shown in Figure 5. Based on the elbow method in Figure 5, it can be seen that the sharpest elbow image [9] is at value = 3 (x axis), which means that the number of clusters selected is 3 clusters.

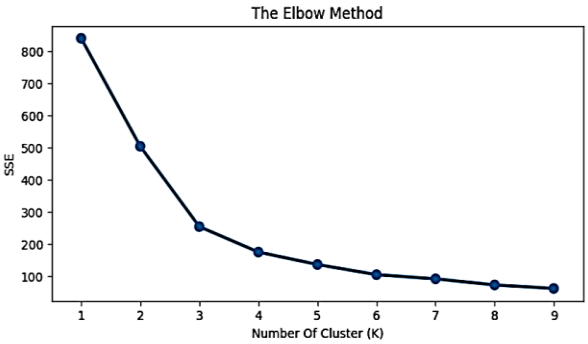

Figure 5. Elbow Method for Determining the Number of Clusters

The K-Means model is run with *n*_cluster = 3, resulting in a centroid value in the last iteration as shown in Table 7.

Table 7. Final Centroid Results

Cluster	Members	Centroid R	Centroid F	Centroid M
0	200	-0,3929005	-0,181334	-0,248403
1	35	-0,5386339	1,895777	2,114288
2	45	2,1651619	-0,668566	-0,540429

A scatter plot graph in Figure 6, can be seen the distribution of data for the monetary recency attribute:

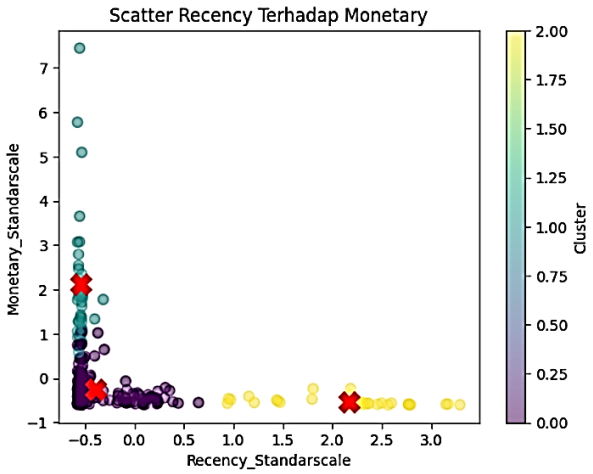

Figure 6. Scatter Recency Against Monatery

The green dots, which are data groups with cluster 1, the purple dots are cluster 0 data points, and the yellow g dots are cluster 2 data points. Each piece of data is spread by grouping similarities in R and M values that have been normalized. In addition, in the Figure 6, it can be seen that each cluster group has a centroid point, which is marked with a red X for each cluster.

In Cluster Evaluation phase, we compare the Silhouette Index RFM Actual Value and the RFM Score Value

To obtain the best model that will be used to interpret the cluster and prototype clustering results, a comparison test was carried out by running the actual RFM value model and the RFM score model into the K-Means model to obtain the Silhouette Index value. Table 8 shows the results of the comparison of the two models:

Table 8. Comparison Results of Actual RFM and RFM Score Based on Silhouette

Number of Clusters	Actual Value RFM	RFM Value Score
2	0.561928	0.485876
3	**0.624646**	0.479203
4	0.581366	0.493759
5	0.568411	0.434975
6	0.498001	0.432770
7	0.475831	0.391266
8	0.482351	0.402628

From the Table 8, a silhouette index or score value that is close to 1 means the cluster quality is relatively good and ideal. For the RFM score value model, the best cluster quality is at the number of clusters (K) = 4 with a silhouette value of 0.493759, while for the RFM actual value model, the best cluster quality is at the number of clusters (K) = 3 with a silhouette value of 0.624646. From the comparison of the two RFM models, the actual value model has a higher silhouette value than the score model, so the actual value model is considered better than the score value model in this study.

Apart from that, the comparison Table 8 also shows that the number of clusters (K) produced between the elbow and silhouette index methods has comparable or harmonious values, namely K = 3.*RFM Analysis of Cluster Results*

DOI: https://doi.org/10.29207/resti.v7i6.5416

The K-Means modeling is processed using the actual value RFM dataset because the actual value RFM model has better silhouette values based on the comparative evaluation stage of silhouette values. However, in this research, the RFM score model analysis is also used to add to the RFM segmentation analysis rules, which can provide information and knowledge in interpreting and understanding outlet segmentation. Table 9 shows a sample of cluster results after adding segment and score attributes.

The data frame in Table 9 displays information on grouping outlets based on clusters and other information, where outlets are also divided based on segment and score criteria. Retail_id C100000641 is a member of cluster 0 with the Potential Loyalist and Gold criteria, and C100007252 is a member of cluster 2 with the Hibernating and Green criteria. The results will add value to cluster interpretation analysis, which is useful for the business domain.

Table 9. Cluster Results Data Frame with Segments and Scores

retail_id	recency	frequency	monetary	cluster	RFM_Segment	RFM_Score	segment	score
C100000641	3	13	47527750	0	523	10	Potential loyalists	Gold
C100000953	5	7	986650	0	411	6	Promising	Bronze
C100002548	2	189	4,64E+08	1	555	15	Champions	Platinum
C100003179	5	26	24206851	0	432	9	Potential loyalists	Silver
C100003361	87	30	59794160	0	133	7	At risk	Bronze
C100006134	13	35	67717750	0	244	10	At risk	Gold
C100006393	44	3	13927600	0	212	5	Hibernating	Green
C100006680	5	8	1016650	0	421	7	Potential loyalists	Bronze
C100006808	2	54	4,51E+08	1	555	15	Champions	Platinum
C100007252	258	6	3329000	2	111	3	Hibernating	Green

Apart from that, this research provides information on outlet mapping based on existing clusters and score criteria, as can be seen in Figure 7.

Figure 7. Map Outlet RFM Segment

The outlet segment map in Figure 7 provides outlet information with the composition of segment criteria: hibernating (84%), potential loyalists (22%), loyal outlets (18%), champions (17%), at risk (8%), new outlets (2%), need attention (0.7%), can't lose them (1.1%), and promising (0.7%).

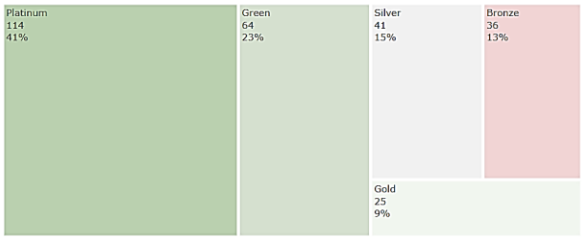

Figure 8. Map Outlet RFM Score

The outlet score map in Figure 8 provides outlet information with the composition of the score criteria: platinum (41%), green (23%), silver (15%), bronze (13%), and gold (9%).

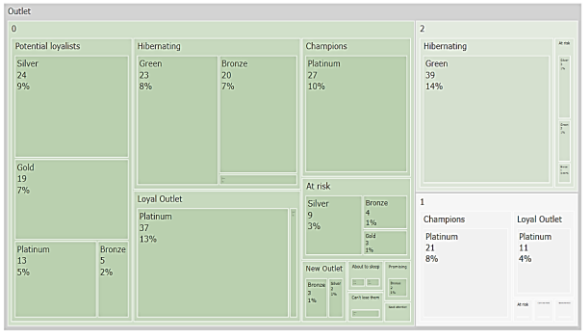

Figure 9. Map Outlet Cluster, Segment, and Score

The map image in Figure 9 is a combination of clusters, segments, and scores, which can be seen in the Table10:

Table 10. Cluster Map, Segment, and Score

Cluster	Segment	Members	Percentage
0	About to sleep	2	1
0	At risk	16	8
0	Can't lose them	2	1
0	Champions	27	13,5
0	Hibernating	45	22,5
0	Loyal customers	39	19,5
0	Need attention	1	0,5
0	New Outlet	5	2,5
0	Potential loyalists	61	30,5
0	Promising	2	1
1	At risk	1	2,9
1	Can't lose them	1	2,9
1	Champions	21	60
1	Loyal Outlet	11	31,4
1	Need attention	1	2,9
2	At risk	6	13,3
2	Hibernating	39	86,7

Based on the explanation Table 9, the cluster results can be interpreted by looking at the data distribution in the graph in Figure 10:

DOI: https://doi.org/10.29207/resti.v7i6.5416

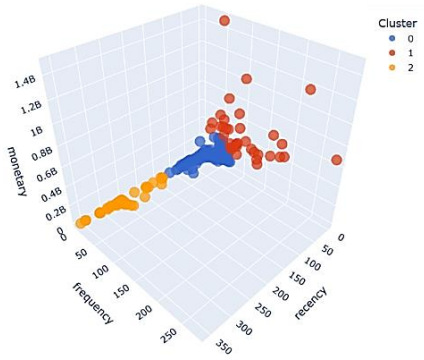

Figure 10. Actual Model RFM Distribution Graph

In the image in Figure 10, the red data point is cluster 1, with high monetary value, low recency, and high frequency. The blue data point is cluster 0, with medium recency, medium frequency, and medium currency.

The yellow data point is cluster 2, with high recency, low shopping frequency, and low monetary value. In addition to the interpretation based on Figure 10, this research explains the cluster results using the RFM segment and score interpretation, which can be seen in Table 11:

Table 11. Interpretation of Cluster Results Based on RFM Analysis

Cluster	Outlet Type	%	RFM Label	RFM Intepretation	Suggest
0	New Outlet	71%	30.5% Potential Loyalists 19.5% Loyal Outlet	Outlets with low recency, medium shopping frequency, and medium shopping value Based on RFM segmentation, 30% of this cluster are potential loyalists, and 19.5% are outlet loyalists. There are also hibernating outlets, which can cause outlets to be lost if not handled properly.	Product promotions, shopping balance credits, and other features were launched for these outlets to increase shopping interest and turn them into champions. Special attention to this cluster is important because there is a potential for hibernating outlets that need to be reviewed per period.
1	Loyal Outlet	13%	31.4% Loyal Outlet 60% Champions	The outlet that transacts most frequently with the highest amount of shopping value (monetary) and transacts with the lowest frequency	Management should provide high-value information and products and solicit reviews from these outlets regarding improved service and better products.
2	Lost Outlet	16%	86.7% Hibernating	Outlets with high recency (long time without shopping transactions), low shopping frequency, and low shopping value RFM segmentation provides information that shows that most outlets in this cluster are hibernating outlets.	It is necessary to survey the condition of the outlet to determine whether it is still actively operating or not. If they are still active, they will be directed to become potential loyal outlets; if they are not removed from the customer base, this will increase salesman productivity by looking for new outlets.

4. Conclusion

Based on the discussion and research results, the conclusions that can be drawn are: Transaction data based on time (time series data) can be transformed into data in the Recency, Frequent, and Monetary (RFM) model; The cluster quality of the RFM model's actual value is better than the RFM model's score, based on a comparison of calculations using the Silhouette index or score; The K-means algorithm can carry out the outlet clustering process, with the number of clusters (K) equal to 3 outlet clusters based on the elbow and silhouette score methods; The results of this outlet clustering process create a series of data that has a cluster label and forms supervised learning data, so that it can be used to analyze patterns or trends using other data mining models such as classification, estimation, and prediction; Business actors (business domain) can plan marketing strategies and how to treat customers appropriately based on the results of outlet cluster interpretation.

The following are suggestions for further research: The research uses historical outlet shopping transaction data over a wider area, for example, district, city, and even provincial transaction data.; Further research uses other cluster algorithms such as agglomerative, DBSCAN, GMM, and others; Further research can be carried out using the same data sources to segment products and relationships (associations) of outlet behavior in carrying out transactions, so that combining these (outlet segmentation and product associations) will provide deeper and more accurate information.

References

[1] M. Y. Smaili and H. Hachimi, "Hybridization of improved binary bat algorithm for optimizing targeted offers problem in direct marketing campaigns," Adv. Sci. Technol. Eng. Syst., vol. 5, no. 6, pp. 239–246, 2020, doi: 10.25046/aj050628.

[2] A. J. Christy, A. Umamakeswari, L. Priyatharsini, and A. Neyaa, "RFM ranking – An effective approach to customer segmentation," J. King Saud Univ. - Comput. Inf. Sci., vol. 33, no. 10, pp. 1251–1257, 2021, doi: 10.1016/j.jksuci.2018.09.004.

[3] R. Srivastava, A. Parvaneh, and H. Abbasimehr, "Identification of Customer Clusters using RFM Model: A Case of Diverse Purchaser Classification," Int. J. Bus. Anal. Intell., vol. 4, no. 2, p. 6, 2016.

[4] S. Dibb, "Market segmentation: Strategies for success," Mark. Intell. Plan., vol. 16, no. 7, pp. 394–406, 1998, doi: 10.1108/02634509810244390.

[5] F. Safari, N. Safari, and G. A. Montazer, "Customer lifetime value determination based on RFM model," Mark. Intell. Plan., vol. 34, no. 4, pp. 446–461, 2016, doi: 10.1108/MIP-03-2015-0060.

DOI: https://doi.org/10.29207/resti.v7i6.5416

[6] Y. Huang, M. Zhang, and Y. He, "Research on improved RFM customer segmentation model based on K-Means algorithm," Proc. - 2020 5th Int. Conf. Comput. Intell. Appl. ICCIA 2020, pp. 24–27, 2020, doi: 10.1109/ICCIA49625.2020.00012.

[7] J. Wei, S. Lin, and H. Wu, "A review of the application of RFM model," African J. Bus. Manag., vol. 4, no. 19, pp. 4199–4206, 2010.

[8] P. D. Bangsa and I. Hermawan, "Jurnal Teknologi Terpadu," J. Teknol. Terpadu, vol. 7, no. 1, pp. 15–22, 2021.[9] J. Wu et al., "An Empirical Study on Customer Segmentation by Purchase Behaviors Using a RFM Model and K -Means Algorithm," Math. Probl. Eng., vol. 2020, no. April 2019, 2020, doi: 10.1155/2020/8884227.

[10] D. Chen, S. L. Sain, and K. Guo, "Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining," J. Database Mark. Cust. Strateg. Manag., vol. 19, no. 3, pp. 197–208, 2012, doi: 10.1057/dbm.2012.17.

[11] B. Sohrabi and A. Khanlari, "Customer lifetime value determination based on RFM model," Mark. Intell. Plan., vol. 14, 2007, doi: 10.1108/MIP-03-2015-0060.

[12] C. Y. Tsai and C. C. Chiu, "A purchase-based market segmentation methodology," Expert Syst. Appl., vol. 27, no. 2, pp. 265–276, 2004, doi: 10.1016/j.eswa.2004.02.005.

[13] S. H. Shihab, S. Afroge, and S. Z. Mishu, "RFM Based Market Segmentation Approach Using Advanced K-means and Agglomerative Clustering: A Comparative Study," 2019 Int. Conf. Electr. Comput. Commun. Eng., pp. 1–4, 2019.

[14] D. Devarapalli, S. Veera, V. Satya, S. Geddam, A. S. Sravya, and A. P. Devi, "Analysis of RFM Customer Segmentation Using Clustering Algorithms," Int. J. Mech. Eng. Vol., vol. 7, no. February, 2022.

[15] M. Aliyev, E. Ahmadov, H. Gadirli, A. Mammadova, and E. Alasgarov, "Segmenting Bank Customers via RFM Model and Unsupervised Machine Learning," 2020.

[16] B. Arivazhagan and G. Vijaiprabhu, "An Enhanced Hierarchical Model for Customer Segmentation in Customer Relationship Management with Demographic , Recency , Frequency and Monetary Values," Int. J. Mech. Eng., vol. 7, no. 2, pp. 1878–1886, 2022.

[17] D. Elzanfaly and S. Salama, "Investigation in Customer Value Segmentation Quality under Different Preprocessing Types of RFM Attributes," vol. 4, no. 4, pp. 5–10, 2016.

[18] A. Gülcü and S. Çalişkan, "Clustering electricity market participants via FRM models," Intell. Decis. Technol., vol. 14, no. 4, pp. 481–492, 2020, doi: 10.3233/IDT-200092.

[19] C. Yuan and H. Yang, "Research on K-Value Selection Method of K-Means Clustering Algorithm," J, vol. 2, no. 2, pp. 226–235, 2019, doi: 10.3390/j2020016.

[20] I. Karacan, I. Erdogan, and U. Cebeci, "A Comprehensive Integration of RFM Analysis, Cluster Analysis, and Classification for B2B Customer Relationship Management," Proc. Int. Conf. Ind. Eng. Oper. Manag., pp. 497–508, 2021.

[21] D. A. Nasution, H. H. Khotimah, and N. Chamidah, "Perbandingan Normalisasi Data untuk Klasifikasi Wine Menggunakan Algoritma K-NN," Comput. Eng. Sci. Syst. J., vol. 4, no. 1, p. 78, 2019, doi: 10.24114/cess.v4i1.11458.

[22] B. Rizki, N. G. Ginasta, M. A. Tamrin, and A. Rahman, "Customer Loyality Segmentation on Point of Sale System Using Recency-Frequency-Monetary (RFM) and K-Means," J. Online Inform., vol. 5, no. 2, p. 130, 2020, doi: 10.15575/join.v5i2.511.

[23] A. Nowak-Brzezinska and C. Horyn, "ScienceDirect Outliers Outliers in in rules - the the comparision comparision of of LOF , LOF , COF COF and and K-MEANS K-MEANS," vol. 00, 2020, doi: 10.1016/j.procs.2020.09.152.

DOI: https://doi.org/10.29207/resti.v7i6.5416