Available at https://journalenrichment.com/index.php/jr/

Enrichment: Journal of Multidisciplinary Research and Development

Consumer segmentation using K-Medians algorithm on transaction data based on LRFMP (length, recency, frequency, monetary, periodecity)

Akbar Dena Maulana[1*], Ade Kania Ningsih[2], Gunawan Abdillah[3]

[1*,2,3] Universitas Jenderal Achmad Yani, Indonesia

*Email: akbardena12@gmail.com

ARTICLE INFO

Keywords:
Clustering, Consumer Segmentation, K-Medians, LRFMP

ABSTRACT

Consumer loyalty has a crucial role for companies, especially in conditions of competition between companies. Success in retaining loyal customers is crucial. For this reason, customer loyalty analysis is needed to identify the level of consumer compliance with the company. In this case, consumer segmentation is also an important step to group consumers with similar characteristics to facilitate the management process. One of the analysis methods used is the LRFMP (Length, Recency, Frequency, Monetary, Periodecity) model, which examines consumer purchasing patterns based on various factors such as relationship length, last transaction time span, number of transactions, total money spent, and purchase regularity. The K-Medians grouping method was also used in this study. The data used is the history of purchase transactions in e-commerce for 373 days. From the application of LRFMP analysis and the K-Medians method, 4 clusters were obtained. The number of consumers in cluster 1 is 1183, cluster 2 is 1221, cluster 3 is 1206, and cluster 4 is 1102. The interpretation of the LRFMP model shows that 25.1% of consumers have high loyalty potential, 25.9% of consumers have low loyalty potential, 25.6% of consumers have high loyalty potential, and 23.4% of consumers have medium loyalty potential.

INTRODUCTION

At this time, the business world is experiencing very rapid development. The Internet has become a very important tool in supporting the buying process through a new platform known as e-commerce or electronic commerce. Based on information provided by Bank Indonesia, there has been an almost twofold increase in the number of trade transactions in the E-commerce sector as a result of the Covid-19 Virus pandemic. In 2019, the number of transactions was recorded at 80 million, but increased to 140 million transactions in 2020. This increase in the number of transactions has an impact on the increase in e-commerce transaction data stored. One of the main assets that is important for the company is maintaining relationships with customers. One way to obtain information is to investigate the transaction data recorded from these e-commerce customers. One way to achieve this is to divide customers into different segments, with the aim of predicting and identifying potential customers so as to attract new customers. By analyzing customer segments, companies can implement marketing strategies that are appropriate for each of those segments. In addition, companies can understand how long customers remain loyal and know the pattern of relationships between customers and companies. By understanding this, companies can increase profitability by optimizing interactions with customers and providing services that meet their needs.

In E-Commerce businesses, transaction data is usually only used for weekly, monthly, or yearly sales reports (Siagian et al., 2022). Even though transaction data has the potential to be used to decide marketing strategies, which is consumer segmentation (Priati, 2019). Consumer segmentation needs to be done in order to group consumers who have the same level of characteristics to improve marketing strategies to be more focused. To select the variables used, the model that will be used to group consumers in this study is LRFMP (*Length, Recency, Frequency, Monetary, Periodecity).* (Sulistyawati & Sadikin, 2021) LRFMP is a numerical expression of customer behavior and is very effective in determining the most profitable customers.(Desember, 2017)

This study applies the K-Median algorithm because this algorithm is to generate consumer loyalty level groups based on purchase transactions until LRFMP analysis models can be carried out to determine the level of customer loyalty for e-commerce businesses, with the help of silhouette calculations to test clusters of consumer loyalty grouping results and to convert data into meaningful information.

METHOD

The method used in the application of the clustering method for customer segmentation consists of stages depicted in the form of a flowchart contained in Figure 1.

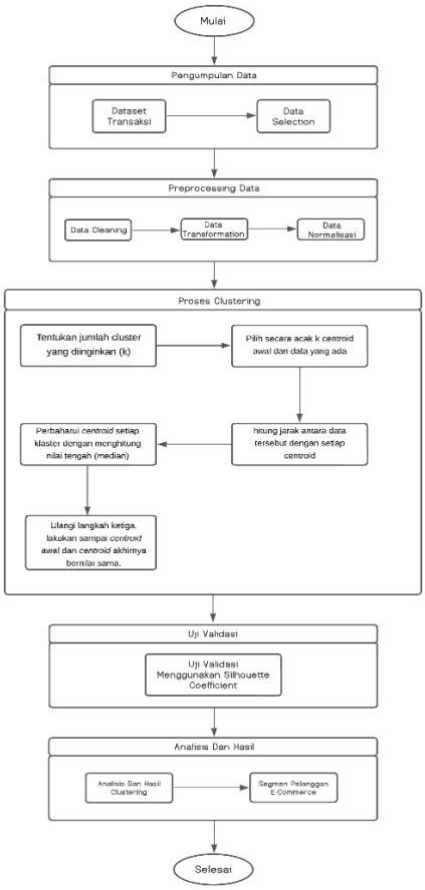


Figure 1. Stages of Research

1. Data Selection

   At the data selection stage, transaction data to be used in this study was selected. The selection of transaction data is based on LRFMP analysis. (Length, Recency, Frequency, Monetary, Periodecity). Data selection can be seen in table 1.

Table 1. Selection

CustomerNo	Date	Price	Quantity
12004	9/14/2019	18.75	4
12004	9/14/2019	11.92	4
12004	9/14/2019	13.62	4
12004	9/14/2019	12.38	5

2. Data Cleaning

   In this step, the process of deleting data that is not needed or not used in customer grouping is carried out. By cleaning data from duplication, missing value, and discrepancies, and ensuring that the data used in analysis is quality data.

3. Data Transformation

Performed on customerid, quantity, date, and price attributes to make them more measurable and appropriate in the LRFMP model so that they can be used as clustering attributes. [12] Then attribute construction is carried out according to LRFM attributes (length, recency, frequency, monetary, periodecity)[10] including:

a) Length (L): This attribute is the length of the relationship between the customer and the company.
b) Recency (R): This attribute is the distance between the last time of purchase and the current time.
c) Frequency (F): This attribute represents the number of transactions that have been carried out in a given period of time.
d) Monetary (M): This attribute is the amount of money that consumers have spent to make purchases in a certain period of time.
e) Periodecity (P): This attribute is the regularity of customers in making purchases.

Table 2. Transformation

Length (L)	Recency (R)	Frequency (F)	Monetary (M)	Periodecity (P)
0	229	1	842.39	373
0	220	1	12.38	373
0	278	1	2846.6	373
46	33	2	23.32	186

4. Normalization

Data normalization with the min-max method is done because this allows processing variables with different scales into comparable, using the equation:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

5. K-Medians Clustering

This K-medians clustering stage is a calculation stage carried out to group against LRFMP data on e-commerce sales that have previously been normalized, namely:

a) Specify the desired number of clusters (k).
b) Randomly select the initial k centroid from the existing data. Centroids can be actual data points or random points located within a range of data.
c) For each record in the data set, calculate the distance between that data and each centroid. Assign data to the nearest centroid based on distance. In k-medians clustering, the distance generally used is the Manhattan distance or L1-norm.

$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$ (2)

Information:

d : distance and xy
x : Data Center Cluster
y : Data attribute
y_i : cluster center data to i
x_i : cluster center data to i
n : amount of data
i : every data

d) Each observation groups based on the closest distance each data has to its center.
e) Once all data is assigned to the centroid, calculate the median of each data attribute in each cluster. Make this median value the new centroid for each cluster.
f) Do the third step repeatedly until the initial centroid and centroid values finally become the same.

6. Stages Silhouette Coeffient

Silhoutte Coeffient is a stage where the results of calculations carried out by k-medians are tested for accuracy using silhouette to see how accurate the grouping carried out by the k-medians method is.

$$S = \frac{b-a}{\max(a,b)} \quad (3)$$

Where b is the average of the median distance with the object outside the cluster, a is the average distance between the median and the object inside the cluster.

RESULTS AND DISCUSSION

The following transaction data has been altered in the LRFMP analysis in Table 3.

Table 3. Transaction Data

CustomerNo	Date	Price	Quantity
12004	9/14/2019	18.75	4
12004	9/14/2019	11.92	4
12004	9/14/2019	13.62	4
12004	9/14/2019	12.38	5
12004	9/14/2019	12.77	8
12004	9/14/2019	13.62	10

The normalization method transforms numerical attributes on a smaller scale, with the lowest limit of 0 and the highest limit of 1. Here is the calculation of the normalization of Min -Max as in Tables 5 and 6.

Table 1. LRFMP Data

Length (L)	Recency (R)	Frequency (F)	Monetary (M)	Periodecity (P)
0	229	1	842.39	373
0	220	1	12.38	373
0	278	1	2846.6	373
46	33	2	23.32	186
222	109	2	53.4	186
23	289	46	678.7	2
90	251	34	5487.31	11
71	72	76	49.52	41

Table 5. Min-Max Normalization

CustomerNo	Length (L)	Recency (R)	Frequency (F)	Monetary (M)	Periodecity (P)
12004	0.3780	0.2353	0.0072	0.0093	0.0226
12006	0.0000	0.5882	0.0000	0.0001	0.0000
12008	0.4129	0.3316	0.0263	0.0315	0.0113
12013	0.0000	0.9652	0.0000	0.0002	0.0000
12024	0.0000	0.4759	0.0005	0.0005	0.0052
12025	0.0617	0.7727	0.0059	0.0075	0.0095
12026	0.2413	0.6711	0.0532	0.0609	0.0079
12031	0.1903	0.1925	0.0008	0.0005	0.0710

The number of k is chosen randomly as the basis of the cluster center (initial centroid) as much as k (number of clusters). In determining the number of k, 4 test results have been determined according to the program in the Figure 2.

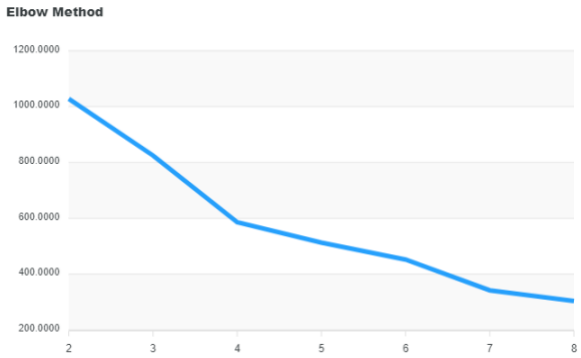


Figure 2. Determination K

Furthermore, the program results of k-median clustering calculations can be seen in Figure 3.

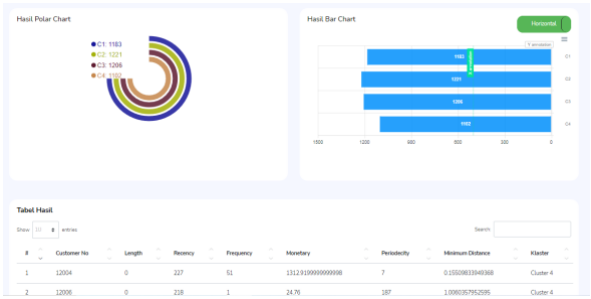


Figure 3. K-Medians Results

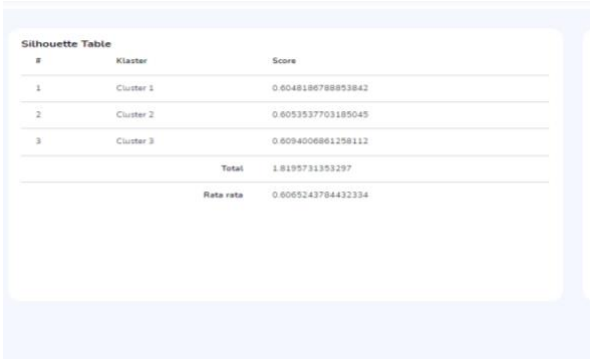
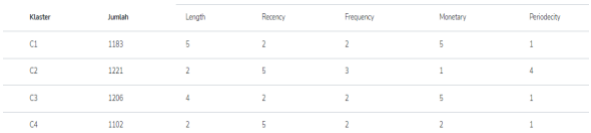


Figure 4. Cluster analysis

In the analysis, the results of LRFMP are obtained from the results of LRFMP interpretation of e-commerce sales that have been carried out.



For more detailed results can be seen in table 7 below.

Table 6. Analysis of cluster results

Cluster	Number of Customers	L	R	F	M	P
C1	1183	5	2	5	5	1
C2	1221	2	5	3	1	4
C3	1206	4	2	2	5	1
C4	1102	2	5	2	2	1

After obtaining the value of each Recency, Frequency and monetary attribute, the next thing that needs to be done is to provide a quantitative score for each value so that it can be processed in the next stage [11]. This quantitative value is used using a scale range of 1-5. Based on the score scale that has been made, the results of the evaluation of RFM attributes from the results of program interpretation using the K-means algorithm are shown as in Table 7.

Table 7. Clustering results and average score

Cluster	Amount of data on the cluster	LRFMP Analysis
1	1183	↑L ↓R ↑F ↑M ↓P
2	1221	↓L ↑R ↑F ↓M ↑P
3	1206	↑L ↓R ↓F ↑M ↓P
4	1102	↓L ↑R ↓F ↓M ↓P

By considering the results of scores and clustering that have been run using the developed software, it can describe customer characteristics based on the attributes of Length, Recency, Frequency, Monetary, and Periodecity. As presented on the table.

Cluster		Interpretasi	Result
1	Length	A score of 5 (↑L) means that customers in this group have been in a relationship with the company for a long time	The interpretation results show that the customer segment in this cluster can be indicated by high loyalty

	Recency	A score of 2 (↓R) means that the last transaction time has not been made for a long time, this gives the potential customer to make another transaction	
	Frequency	A score of 2 (↓F) means that this group of customers rarely makes purchase transactions at a certain period of time	
	Monetary	A score of 5 (↑M) means that this group of customers provides a high financial value for the company	
	Periodcity	A score of 1 (↓P) means that this group of customers has a time between transactions that tends to be short.	
2	Length	A score of 2 (↓L) means that customers in this group are new to the company	The interpretation results show that the customer segment in this cluster can be indicated by low loyalty
	Recency	A score of 5 (↑R) means that you have not made a transaction for a long time	
	Frequency	A score of 3 (↑F) means that this group of customers quite often make purchase transactions at a certain period of time	
	Monetary	A score of 1 (↓M) means that this group of customers provides less financial value for the company	
	Periodcity	A score of 4 (↑P) means that this group of customers has a long time between transactions	
3	Length	A score of 4 (↑L) means that customers in this group have been in a relationship with the company for a long time	The interpretation results show that the customer segment in this cluster can be indicated by high loyalty
	Recency	A score of 2 (↓R) means that the last transaction time has not been made for a long time, this gives the potential customer to make another transaction	
	Frequency	A score of 2 (↓F) means that this group of customers rarely makes purchase transactions at a certain period of time	
	Monetary	A score of 5 (↑M) means that this group of customers provides a high financial value for the company	
	Periodcity	A score of 1 (↓P) means that this group of customers has a time between transactions that tends to be short.	
4	Length	A score of 2 (↓L) means that customers in this group have not been in a relationship with the company for long	The interpretation results show that the customer segment in this cluster can be indicated as moderate loyalty
	Recency	A score of 5 (↓M) means that you have not made a transaction for a long time	
	Frequency	A score of 2 (↓F) means that this group of customers does not make frequent purchase transactions at a certain period of time	
	Monetary	A score of 2 (↓M) means that this group of customers provides less financial value for the company	
	Periodcity	A score of 1 (↓P) means that this group of customers has a time between transactions that tends to be short	

CONCLUSION

This research resulted in a system of grouping the level of customer loyalty using the K-Medians method with LRFMP parameters (Length, Recency, Frequency, Monetary, Periodecity). This system produces a grouping of customer loyalty levels as a reference in the analysis of customer loyalty levels in e-commerce. The results of the LRFMP cluster analysis conducted had 4 clusters with 1183 consumers in cluster 1 with high loyalty indications, 1221 consumers in cluster 2 with low loyalty indications, 1206 consumers in cluster 3 with high loyalty indications, and 1102 consumers in cluster 4 with medium loyalty indications. So that there are 1221 consumers who are potentially unloyal to this e-commerce business. The testing or evaluation of cluster quality contained in the system using silhouette with results reaching 0.608 close to number 1.

REFERENCES

Desember, V. N. (2017). *Jurnal Teknologi Pelita Bangsa*. *8*(4).

Priati, ahmad fauzi. (2019). *Penerapan Data Mining Dengan Teknik Clustering Menggunakan Algoritma K-Means Pada Data Transaksi Superst*. *September*, 15–19.

Siagian, R., Sirait, P., Halim, A., Komputer, P., Area, U. M., & Estate, M. (2022). *Penerapan Algoritma K-Means dan K-Medoids untuk Segmentasi Pelanggan pada Data Transaksi E-Commerce The Implementation of K-Means and K-Medoids Algorithm for Customer Segmentation on E-commerce Data Transactions*. *11*, 260–270.

Sulistyawati, A. A. D., & Sadikin, M. (2021). Penerapan Algoritma K-Medoids Untuk Menentukan Segmentasi Pelanggan. *Sistemasi*, *10*(3), 516. https://doi.org/10.32520/stmsi.v10i3.1332

Desember, V. N. (2017). *Jurnal Teknologi Pelita Bangsa*. *8*(4).

Priati, ahmad fauzi. (2019). *Penerapan Data Mining Dengan Teknik Clustering Menggunakan Algoritma K-Means Pada Data Transaksi Superst*. *September*, 15–19.

Siagian, R., Sirait, P., Halim, A., Komputer, P., Area, U. M., & Estate, M. (2022). *Penerapan Algoritma K-Means dan K-Medoids untuk Segmentasi Pelanggan pada Data Transaksi E-Commerce The Implementation of K-Means and K-Medoids Algorithm for Customer Segmentation on E-commerce Data Transactions*. *11*, 260–270.

Sulistyawati, A. A. D., & Sadikin, M. (2021). Penerapan Algoritma K-Medoids Untuk Menentukan Segmentasi Pelanggan. *Sistemasi*, *10*(3), 516. https://doi.org/10.32520/stmsi.v10i3.1332