EXPLORE – Volume 12 No 1 Tahun 2022 p-ISSN : 2087-894X
Terakreditasi Sinta 5 SK No : 23/E/KPT/2019 e-ISSN : 2656-615X

IMPLEMENTATION RFM ANALYSIS MODEL FOR CUSTOMER SEGMENTATION USING THE K-MEANS ALGORITHM CASE STUDY XYZ ONLINE BOOKSTORE

Tri Juhari[1], Asep Juarna[2]
Universitas Gunadarma[1,2]
trijuhari0@gmail.com, ajuarna@staff.gunadarma.ac.id

Abstract – XYZ online bookstore is one of the companies engaged in the online book sales industry that located in Jakarta, Indonesia, but the marketing strategy given to customers has not been maximized, so it has not been able to increase book purchase transactions. Therefore a customer-centered marketing strategy is needed by implementing Customer Relationship Management, One of the methods that can be applied is customer segmentation. Customer segmentation can be done by implementing a data mining process which carried out by using the K-means clustering algorithm and based on the RFM (Recency, Frequency, Monetary) model. . Determining the number of clusters in the clustering process using the elbow method. Performance tests on cluster results using the silhouette method, and the Calinski-Harabasz index. The results of cluster analysis based on customer value using the RFM Combination and Customer Value Matrix methods show that based on the RFM Combination method produces 3 types of customer characteristics namely loyal customers, new customers, and lost customers. Meanwhile, based on the customer value matrix method, it produces 2 types of customer characteristics namely best customer and uncertain customer.

Keywords: Clustering, Customer Segmentation, Customer Value, K-means, RFM;

1. Introduction

The rapid development of information and technology has an impact on increasingly large data storage such as data warehouses. Every year XYZ's online bookstore generates a large volume of data, but this data will only take up storage memory if it is not processed for marketing or decision-making purposes. The utilization of the data warehouse has not been maximized so that the data used to analyze is only the total value of the resulting transactions. In fact, many interesting patterns can be analyzed further. This situation indicates that the data currently owned has not been fully utilized to analyze the characteristics of each XYZ Online Bookstore customer.

Based on sales transaction data on the XYZ Online Bookstore website in the period January 2019 to November 2020, the total customers who made transactions every month decreased by 2.7% or around 1311 customers, while the total value of book transactions decreased by 2.4% or around IDR 194.333.518 [15]

In addition to a decrease in the number of transactions and total transaction value, the growth of new customers during the period January 2019 to November 2020 also decreased by 4.9% or around 1144, however, the decrease in total customer growth indicates that there are customers who make repeat purchases, or it can be said that there are types of loyal customers.

Based on the conditions and problems described above, there are several factors that affect the decline in book shopping during the 2019-2020 period including (1) many pirated books circulating on the market (2) Comparison of book prices and book quality which are relatively the same as other bookstores (3) Many bookstores that sell printed books by online (4) haven't implemented technology to develop markets that adapt to current developments such as implementing business strategies obtained from analyzing customer needs (5) the outbreak of COVID19 during the 2020 period that caused transactions book sales to decline due to a decrease in the purchasing power of the community and Enforcement of Large-Scale Social Restrictions which caused the book order delivery process is hampered.

Based on several factors that affect the decline in book shopping, so that a segmentation approach is needed based on changing consumer trends such as a deeper understanding of customer preferences, customer habits so that it allows companies to create more targeted offers and campaigns that are more responsive to consumers. An understanding of customers is contained in customer relationship management which describes a comprehensive strategy in the process of acquiring, retaining, and partnering with customers. Therefore, an efficient

alternative is to segment customers

In this study, the customer segmentation process was carried out by exploring customer transaction history data at the XYZ Online Bookstore during the period January 2019 to November 2020 by implementing the K-Means clustering algorithm along with the adoption of the RFM (Recency, Frequency, Monetary) model. By examining customer consumption patterns, the RFM Model can help companies effectively identify valuable customers and then develop appropriate marketing strategies. In determining the optimal number of clusters is to use 3 methods including the Elbow Method, Silhouette Method, and Calinski- Harabasz. The result clustering is carried out based on the customer value analysis based on study by (Ha and Park, 1998)to develop an RFM combination based on compared the average RFM value for each cluster with the total average RFM value of all clusters to determine four types of customers and by (Marcus, 1998) to develop customer value matrix based combined the average value of frequency and monetary to determine four types of customers [9].

2. Literature Review

a. Related Research

First, research by Wei et al. (2016) conducted a study on implementing the RFM model to analyze customer value in a veterinary hospital located in Taiwan. The purpose of this study is to identify valuable customers based on the RFM analysis model and to develop a marketing strategy with case studies of customers who own dogs. This study applies the self-organizing maps (SOM) and K-means method along with the adoption of RFM (recency, frequency, and monetary).

The results from implementing clustering along with the adoption of RFM, there are 12 clusters that are divided into 2 labels namely Best Customer and Uncertain Customer. Best customers consist of clusters 1, 3, 5, 7, 8, 10, and 12 while uncertain customers consist of clusters 2, 4, 6, 9, and 11.

Second, research by Dursun and Caber (2016) [4] conducted a study on investigating favorable customer profiles at hotels located in Antalya, Turkey. The purpose of this study is to identify a favorable customer based RFM analysis model with the segmentation pro- cess involving customer demographic characteristics. This study applies the self-organizing maps (SOM) and K-means method along with the adoption of RFM (recency, frequency, and monetary).

The results from implementing clustering customer based RFM analysis model with the segmentation process involving customer demographic characteristics, there are 8 clus- ters that divided into 8 labels namely 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.

Third, research Tavakoli et al. (2018) conducted a study on the implementation of cus- tomer segmentation using the development of RFM model namely R+ FM. The purpose of this study is to classify customers into several groups based on their purchasing behav- ior, demographic and geographic information, and psychographic attributes case study at Digikala company which is engaged in online retail.

The results from implementing clustering customer based R+FM, there are 2 segmenta- tion, first segmentation by recency and second, segmentation by customer value including frequency, monetary and weight frequency and monetary. Recency segmentation produces 3 customer characteristics namely active, lapsing, and lapsed while customer value segmen- tation produced 4 clusters namely High Value, Medium with High Monetary, Medium with High Frequency, and Low Value. The results combination of segments based on the R + FM model, there are 11 label segment namely Active High Value, Active Medium with High Monetary, Active Medium with High Frequency, Active Low Value, Lapsing High Value, Lapsing Medium Value, Lapsing Low Value, Lapsed High Value, Lapsed Medium Value, Lapsed Low Value, and Lapsed Low Value.

Fourth, research by Peker et al. (2017) conducted a study on the implementation of customer segmentation using a modified RFM model called the LRFMP model case study in the wholesale retail industry in Antalya, Turkey. The purpose of this study is to classify customers into several groups based on the LRFMP model and K-means algorithm clustering model. The difference between the LRFMP model and the RFM model is the addition of variables L and P. Variable P shows periodicity, which is the periodicity of customer visits which aims to characterize customer behavior and measure customer regularity, while variable L shows Length, which is the time interval in days between the first and last visit customer. This research used 3 cluster validation including the Silhouette index, Calinski

Harabasz index, and Davies Bouldin index that used to find the number clusters optimal.

The results from implementing clustering customer-based LRFMP, there is 5 cluster that divided into 5 label high contribution loyal customers, low-contribution loyal customers, uncertain customers, high spending lost customers and low spending lost customers.

Fifth, research by Dogan et al. (2018) conducted a study on implementing customer segmentation using the RFM model and the K-Means clustering algorithm. The purpose of this study is to classify customers into several groups based on two different models, that is two-step clustering model and the RFM analysis model. Two-step clustering uses the log-likelihood method to measure the centroid distance and Shwarz's Bayesian Criterion (BIC) is used as the clustering criteria. RFM analysis model uses K-means clustering algorithm and RFM (Recency, Frequency, Monetary) Variable.

The results from implementing clustering customers using two different models, that is two-step clustering produces 3 clusters with labels for each cluster including Bronze, Gold, and Premium. RFM analysis with K-means algorithm clustering produces 4 clusters namely Regular, Loyal, Star, and Advanced.
merupakan suatu bentuk penerapan teknologi elektronik untuk berbagai kegiatan pemerintahan dalam cakupan internal dan eksternal (pelayanan umum)[1]. Untuk pencapaian kinerja yang efektif, efisien, cepat dan transparan [2]
. Untuk di sektor pemerintah dan sektor publik, masing – masing dari kedua sektor tersebut

b. Customer Segmentation

Customer segmentation is the process of dividing the customer base into distinct and homogeneous groups to develop different marketing strategies according to the characteristics possessed by customers.

There are many different types of customer segmentation according to the specific trait criteria used for customer segmentation such as the following: [8]

The purpose of segmentation is to tailor products, services, and marketing messages for each segment. Another important benefit of customer segmentation is that it enables company management to understand customer behavior and preferences and acquire knowledge about different groups of customers [13]. With this opportunity, the organization can target high-value customer groups, and thus the organization can target specifically the high-value segments.

c. RFM Analysis

RFM analysis is a common approach for understanding customer purchase behavior. It is quite popular, especially in the retail industry. As its name implies, it involves the calculation and the examination of three KPIs – recency, frequency, and monetary that summarize the corresponding dimensions of the customer relationship with the organization (Tsiptsis and Chorianopoulos, 2010).

1. Recency (R), the recency value shows the time since the last transaction of the customer's purchase. The smaller the range, the greater the R value.
2. Frequency (F), the frequency value shows the number of transactions in one period. The more frequency, the greater the F value.
3. Monetary (M), the monetary value shows the customer value in the form of money spent during the transaction.

d. K-means Clustering Algorithm

The K Means algorithm is a non-hierarchical method that initially takes most of the population components to become the center of the initial cluster. The K-Means algorithm basically carries out two processes, namely the process of detecting the central location of each cluster and the process of searching for members of each cluster. The workings of the K-Means clustering algorithm are as follows : [11]

1. Determine the value of k as the number of clusters formed.
2. Determine the initial value of the centroid or cluster center point. At this stage, the centroid value is determined randomly, but for the next stage using the formula below:

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

3. Calculate the distance between the centroid point and the point of each object using Euclidean Distance as shown in the formula below:

$$D_e = \sqrt{(x_i - s_i)^2 + (y_i - t_i)^2}$$

4. Grouping data to form clusters with the centroid point of each cluster being the closest

centroid point. The determination of cluster members is to take into account the minimum distance of objects
5. Update the centroid value for each cluster.
6. Repeating from step 2 to the end until the value of the centroid point is no longer changed.

e. Customer Value

Customer value is a general approach used to identify profitable customers to develop a marketing strategy. Research conducted by (Ha and Park, 1998) combined the RFM variable by comparing the average RFM value for each cluster with the total average RFM value of all clusters. The up arrow (↑) symbolizes if the average value is greater than the total average, while the down arrow symbol (↓) indicates if the average value is less than the total average.

Based on research conducted by(Wei et al., 2016)[14] that recency is defined as the number of days since the last visit when the first day of the specified period is set to one. A large recency value indicates that the customer has visited the organization more recently So that the variable recency arrows up (↑) symbolizes if the average value is smaller than the total average value, which means that shortly the customer has made a transaction, while the down arrow (↓) symbolizes if the average value is greater than the total average value, which means that the customer has not made a transaction for a long time.

By further taking into consideration the strategic positioning of customer clusters, four major types of customers can be grouped in terms of RFM symbols including types of lost customers with the RFM symbol (R ↓ F ↓ M ↓) and (R ↓ F ↑ M ↑), new customers with the RFM symbol (R ↑ F ↓ M ↓), loyal customers with the RFM symbol (R ↑ F ↑ M ↑), and promising customers with symbols (R ↑ F ↓ M ↑).

Another customer value method that can implemented such as research by (Marcus, 1998) [9] that customer value can be formed into a customer value matrix. This method is particularly suitable for analyzing customer values in small retail and service businesses. The customer value matrix combines the average value of frequency and monetary. The customer value matrix produces four customer type scenarios as shown in figure 1.

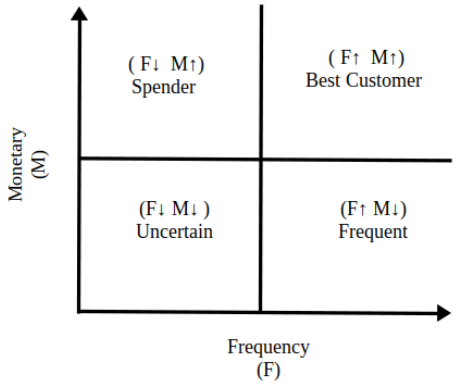


Figure 1. Customer Value Matrix

3. Research Methodology

a. Research Object

The object of this research is a companies that focus on selling books online through websites. The company already has an adequate datawarehouse. The time of research is between November 2020 and January 2021. The dataset to be used is customer transaction history data at XYZ Online Bookstore from January 2019 to November 2020.

b. Research Methodology

The research stages are as shown in the figure 2.

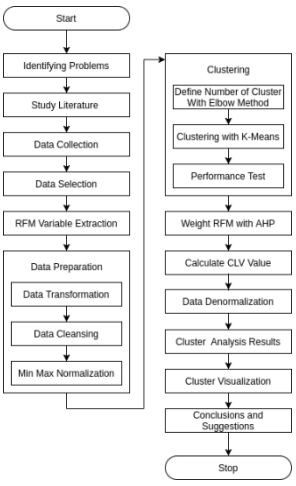


Figure 2. Research Methodology

1. Identifying Problem

In this section, an analysis of the object is carried out. The object that is selected and becomes an input in the identification process is XYZ online bookstore. The output of this stage is the problem raised, namely customer segmentation for input in planning a customer relationship management strategy, especially retaining customers.

2. Study Literature

In this section search for relevant references that are related to customer segmentation. These references can be obtained from interviews with resource persons, books, e-books, previous research studies such as journals and papers, articles, and documents. In performing customer segmentation, a data mining process is needed, so the references needed in the excavation process are those related to clustering, the K-Means algorithm, the RFM model, data mining, customer relationship management, the Elbow method, Min-Max normalization, Calinski-Harabasz index, and the Silhouette method.

3. Data Collection

In this section, data collection is needed as the main support in the customer segmentation process. The data collected is transaction history data from January 2019 to November 2020. Data obtained in excel format (.xlsx) and consists of 50025 data and 12 variables namely Invoice ID, Member ID, Birthday, Gender, Invoice Date, Invoice Time, Districts, City, Province, Shipment Charged, Quantity and Transaction Amount.

4. Data Selection

In this section, data selection is carried out to adjust the attributes used based on the needs in the clustering process. The data that has been collected will be selected based on variables that are related to the clustering process and RFM analysis.

The data selection process is carried out to select the RFM preference variable. Variable Recency is obtained by calculating the difference in the last customer transaction with analysis time is 01 December 2020. Variable Frequency is calculated by adding up customer transactions in one period. Meanwhile, the Monetary variable is calculated by adding up the amount of IDR issued in the transaction.

Based on the variables in the transaction data, it turns out that there is no total value variable for each transaction, so it is necessary to establish a new variable, namely the total transaction value variable based on the sum of the variable shipping charges and the transaction amount variable.

After the total transaction variable is generated, then removing some unused variables such as Districts, City, Quantity, Shipping Charges, and Transaction Amounts so that the variables used include Invoice id, Member id, Birthday, Gender, Invoice date, Invoice time, Districts, Province, and Total transactions.

5. RFM Variable Extraction

The RFM extraction stages are carried out by changing the values into variables of Recency, Frequency, and Monetary. the input of this process is data that has gone through the data selection process and the output is the RFM variable. RFM variable extraction using the R programming language.Transaction date data will be converted into recency and frequency variables. Meanwhile, the total transaction will be converted into the monetary variable. The output of this stage is the variable recency, frequency, and monetary.

In this section, a mapping of RFM data is carried out to determine the current state of the data. Then, the data is transformed using an algorithmic method. Furthermore, data cleaning is performed by removing outliers and data rows that have blank values and finally normalizing using the min-max method.

6. Data Transformation

At this stage, the data change process is carried out using the Logarithmic method, the goal is that the data distribution is not biased. The Logarithmic method was chosen because the raw data used has positive skewness.

7. Data Cleansing

The data cleansing process is the removal of empty or null data and the elimination of outlier data. The process of removing outliers uses the Interquartile Ranges (IQR) method to find the upper and lower bounds.

$$\mathrm{IQR} = Q_3 - Q_1$$
$$\mathrm{L}owerbound = Q_1 - 1.5 \times IQR$$
$$\mathrm{Upper}bound = Q_3 - 1.5 \times IQR$$

8. Min-Max Normalization

The three RFM variables certainly have different ranges, where the monetary variable has a value of up to millions while the frequency and frequency variables have values that only range in the tens or hundreds. At this stage the aim is that the three RFM variables have the same range of values with a certain scale, the aim is that the data distribution is not biased. The normalization method used is the minimum method, where all variables will be converted into values from zero to one. After normalizing the data will be integrated into one table. The input of this process is the RFM variable and the output is the RFM variable that has the same range of values.

9. Clustering

Clustering is the process of grouping a group of data objects into several groups or clusters so that the objects in the cluster have high similarities, but have significant differences, which are different from objects in other clusters. The similarity is assessed based on the attribute value that describes the data object and often involves measuring distance.

The technique in clustering is one of the techniques used to find knowledge from a data set. Clustering is one of the techniques included in the unsupervised model because there are no attributes used as a guide or there are no labels on the data in the learning process (Han et al., 2012).

10. Clustering Analysis

In this section is the main process of implementing customer segmentation. The clustering process consists of three stages namely determining the number of clusters, conducting the clustering process, and lastly conducting the cluster performance test. The input of this process is data that already has the RFM variable and has been carried out by the normalization process. and the output is in the form of an optimum cluster.

11. Clustering Visualization

In this section, the results of the cluster analysis will be converted into a dashboard which aims to make it easier to read the results of the clustering, so that it can help plan customer relationship management to be more precise and fast. The results of the visualization are a web-based dashboard interface including pie charts, bar charts, line charts, and data tables. In addition, members per cluster are displayed along with information on the results of the analysis.

4. Hasil dan Pembahasan

a. Data Collection

The data collected is transaction history data from January 2019 to November 2020. Data obtained in excel format (.xlsx) and consists of 50025 data and 12 variables as shown in Table 1.

Table 1. Raw Data Variable of Book Transaction History

Variable	Data Types	Information
Invoice ID	Varchar	Invoice ID Consists of a unique combination of numbers and characters for each transaction
Member ID	Varchar	Member ID is a unique email address
Birthday	Text	Customer birth date
Gender	Text	Customer gender
Invoice Date	Text	Transaction date
Invoice Time	Text	Transaction time
Districts	Text	Sub-district where the book is shipping
City	Text	The city where the book is shipping
Province	Number	The province where the book is shipping
Shipment Charged	Number	Total Shipping Costs
Quantity	Number	The number of types of books purchased per transaction

Transaction Amount	Number	The total value of the Book purchase transaction

b. Data Selection

The data that has been collected will be selected based on variables that are related to the clustering process and RFM analysis. Based on the variables in the transaction data, that there is no total value variable for each transaction, so it is necessary to establish a new variable, namely the total transaction value variable based on the sum of the variable shipping charges and the transaction amount variable. The variable result selection data as shown in Table 2.

Table 2. Raw Data Variable of Book Transaction History After Selection Data

Variable	Data Types	Information
Invoice ID	Varchar	Invoice ID Consists of a unique combination of numbers and characters for each transaction
Member ID	Varchar	Member ID is a unique email address
Birthday	Text	Customer birth date
Gender	Text	Customer gender
Invoice Date	Text	Transaction date
Invoice Time	Text	Transaction time
Districts	Text	Sub-district where the book is shipping
City	Text	The city where the book is shipping
Province	Number	The province where the book is shipping
Transaction Amount	Number	The total value of the Book purchase transaction

c. RFM Analysis

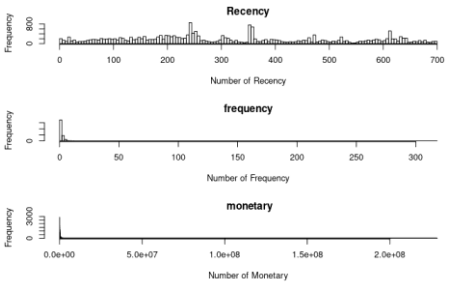
RFM variable extraction will produce three variables including the variable recency, frequency, and monetary. Recency and frequency variables are formed using the invoice date variable, while monetary variables are formed using the total transaction variable. Each customer accumulated their respective RFM values in the period January 2019 to November 2020. Total customers who made transactions during the period January 2019 to November 2020 were 23152 customers can be seen in Table 3.

Table 3. Result RFM Variables

Member Id	Recency (Days)	Frequency (Time)	Monetary (IDR)
****_01@***il.com	258	1	103500
****_01@***oo.com	59	3	503410
****_0401@***oo.com	360	3	780900
****_0576@***oo.co.id	262	2	131050
****_0704puspita@***oo.com	269	41	922534

d. Data Transformation

Based on checking the distribution of the RFM variable as shown in Figure 3 , show that the data on the variable frequency and monetary has a very large positive skew (positive skewed), According (Zumel and Mount, 2019), the distribution of highly skewed positive data, such as customer value, revenue, sales, and stock prices can be modeled as a lognormal distribution by using natural logarithm, which is log base 10.



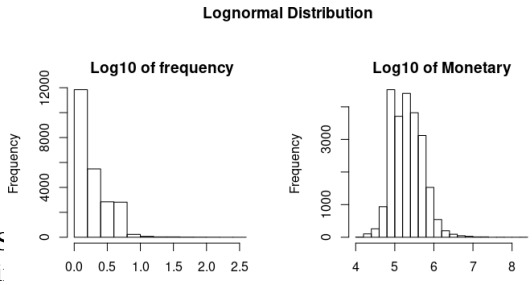
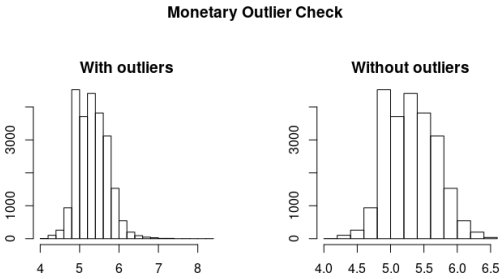


Figure 3. Data Transformation

Member Id	Norm R	Norm F	Norm *M*
****_01@***il.com	0.848	0.00	0.500
****_01@***oo.com	0.622	0.564	0.849
****_0401@***oo.com	0.898	0.416	0.755
****_0576@***oo.co.id	0.850	0.263	0.420
****_0704puspita@***oo.com	0.854	0.525	0.786

e. Data Cleansing

Based on checking the missing value



there are 46048 rows of null data on the birthday variable and 47768 rows of null data on the gender variable. Because the gender and birthday variables have more null values than the filled data, therefore the gender and birthday variables can be deleted. After removing outlier using Interquartile Range (IQR), data distribution on the frequency and monetary variable not very skewed as shown in Figure 4.

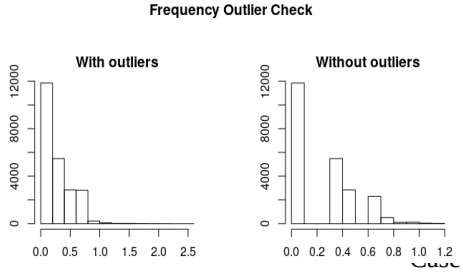


Figure 4. Outlier Checking

f. Min-max Normalization

Data normalization process uses the min-max method, which is the process of changing the data value to 0 to 1. .Min-Max normalization can be calculated using the equation below [6] .The result of data normalization can be seen in Table 4.

$$x' = \frac{x - min_a}{max_a - min_a}(newmax_a - newmin_a) + newmin_a$$

Table 4. Min-max Normalization Result

g. Determination Number of Cluster

In determining the value of k (number of clusters) using the elbow method. The Elbow method is used to determine the optimal number of clusters in K-means clustering [2].

The purpose of the elbow method is to select a k value (number of clusters) that is small and still has a low Withinss value. Based on the elbow graph as shown in Figure 4, it is found the number of clusters optimal according to the Elbow method is 3 clusters.

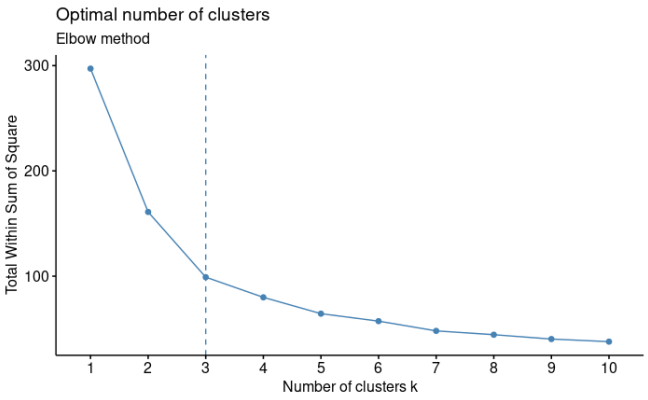


Figure 5. Determination Number of Clusters with Elbow Method

h. K-Means Clustering

Clustering is the process of grouping a group of data objects into several groups or clusters so that the objects in the cluster have high similarities, but have significant differences, which are different from objects

in other clusters [6].

In the clustering process, the RFM variables used in the clustering process are variables that have been normalized. The results of the K-Means clustering process are in the form of information showing the number of members of each cluster, the center point or centroid, and the cluster performance value based on within cluster sum of squares. The proportion of the number of clusters is 9142: 7045: 6965 with a value between_SS / total_SS = 66.7%. The result K-Means clustering as shown in Table 5

Table 5. Cluster Result

Member Id	Norm R	Norm F	Norm *M*	Cluster
****_01@***il.com	0.848	0.00	0.500	2
****_01@***oo.com	0.622	0.564	0.849	1
****_0401@***oo.com	0.898	0.416	0.755	1
****_0576@***oo.co.id	0.850	0.263	0.420	2
****_0704puspita@***oo.com	0.854	0.525	0.786	1

i. Cluster Performance

The cluster performance test or cluster evaluation was carried out using the Silhouette method and the Calinski-Harabasz Index method.Based on the Calinski-Harabasz Index method That he best number of clusters is indicated by the greater the CH value [1] and based on Silhouette that the results of clustering are said to be good if the value of the silhouette coefficient is positive, a positive value indicates good results [7] .The result cluster performance use 2 methods that 3 clusters are the optimal clusters as shown in Figure 5.

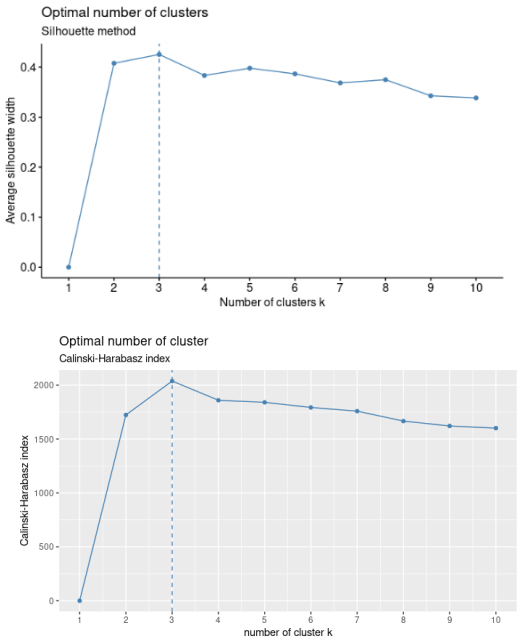


Figure 6. The Result Cluster Performance

j. Clustering Analysis

Based on the clustering process using the K-Means method, 23152 customers were divided into three customer groups as shown in Table 6 .

Table 6. The Clustering Analysis

Cluster	1	2	3
Total Customer	9142 (39.5%)	7045 (30.4%)	6965 (30.1%)
Average Recency	199	236	560
Average Frequency	3	1	1
Average Monetary	IDR 502.249	IDR 141.634	IDR 169.629
RFM Score	R ↑ F ↑ M↑	R ↑ F↓ M↓	R ↓ F↓ M↓

The results of clustering analysis based on customer distribution as shown pie chart in, that cluster 1 is the highest number of members with 9142 customer or 39.5% of the total customers. Cluster 2 with 7045 customers or 30.4% of the total customers and cluster 3 with 6965 members or 30.1% of the total customers.

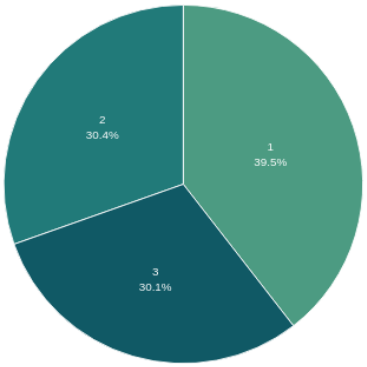


Figure 7. Percentage of Customer Distribution

Based on the average recency, cluster 1 has an average recency of 199 days, cluster 2 has an average recency of 236 days, and cluster 3 has an average recency of 560 days, If the three clusters are compared with each other, cluster 1 is the cluster that has the smallest recency, which means that the cluster 1 have made last transactions about 6 months ago, starting from the time of analysis or on December 1, 2020. Cluster 3 is the cluster that has the greatest recency which means that cluster 3 last made transactions about 18 months ago starting from the time of analysis that is December 1, 2020.

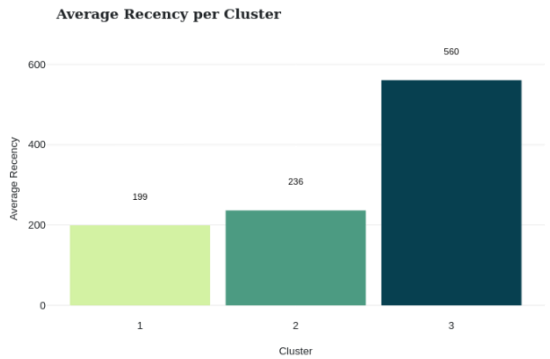


Figure 8. Average Recency Per cluster

Based on the average frequency recency, cluster 1 has an average frequency of 3 times, cluster 2 has an average frequency of 1 time, and cluster 3 has an average frequency of 1 time. If the three clusters are sorted from largest to smallest, cluster 1 is the cluster that has the most frequency, which means that cluster 1 has made repeated purchases, cluster 2 has made at least 1 transaction and cluster 3 has made at least 1 transaction.

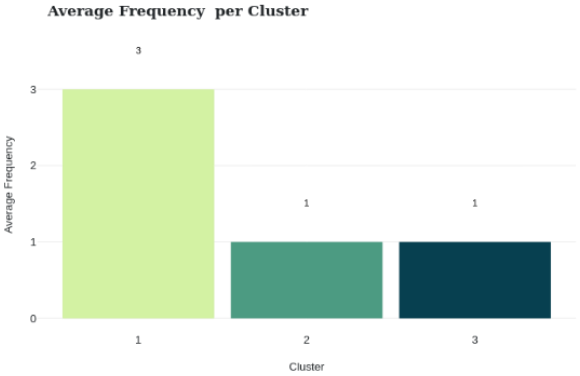


Figure 9. Average Frequency Per cluster

Based on the average monetary recency, cluster 1 has an average monetary value of IDR 502.249. Cluster 2 has an average monetary value of IDR 141.634 and cluster 3 has an average monetary value of IDR 169.629. If the three clusters are sorted from largest to smallest then cluster 1 has the largest monetary average, which means that customers in cluster 1 spend a lot of money to buy books at XYZ Online Bookstore, followed by cluster 3 by having the second-largest monetary average after cluster 1, and cluster 2 with the smallest monetary average.

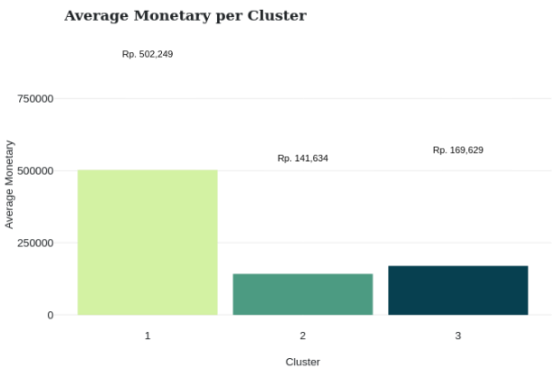


Figure 10. Average Monetary Per cluster

k. Customer Value Analysis with RFM Combination

The results of identification of customer groups based on the average value index of RFM (RFM Combination) that Cluster 1 can be identified as loyal customers with the symbol of (R ↑ F ↑ M ↑). In addition, cluster 2 with the symbol of (R ↑ F ↓ M ↓) is identified as a new customer. In contrast, Cluster 3 with the symbol of (R ↓ F ↓ M ↓) belong to lost customers. 3.5.2 Customer Value Analysis with Customer Value Matrix. The results of the customer value analysis with customer value matrix method, there are 2 types of customers identified as shown in

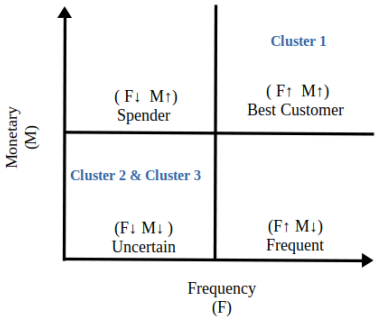


Figure 14, including best and uncertain customers.

Figure 11. Customer Value Matrix Results

The first cluster is a type of best customer with an arrow symbol on the FM variable showing an up arrow (F ↑ M ↑) while the second and third clusters are types of uncertain customer with an arrow symbol on the FM variable showing a down arrow (F ↓ M ↓).

5. Conclusions

Based on the research that has been done, it can be concluded that 23152 customers made transactions during the period January 2019 to November 2020. Determination of the optimal number of clusters using the elbow method produces 3 clusters as the optimal number of clusters. The clustering process uses the K-Means algorithm and the number of clusters used is 3 resulting in a cluster proportion of 9142: 7045: 6965 with a value between_SS / total_SS = 66.7%. The performance test results based on the silhouette and the Calinski-harabasz Index method prove that the optimal number of clusters is 3 clusters.

Based on the customer value analysis with the combination RFM model, cluster 1 is identified in the loyal customer with symbol of (R ↑ F ↑ M ↑). Cluster 2 is identified in the new customer with symbol of (R ↑ F ↓ M ↓). Cluster 3 is identified in the lost customer with symbol of (R ↓ F ↓ M ↓).

Based on the customer value analysis with the customer value matrix, cluster 1 is identified in the best customer group with the FM score showing an up arrow (F ↑ M ↑), while cluster 2 and cluster 3 are identified in the uncertain customer group with the FM score showing a down arrow (F ↓ M ↓).

This research still has the weakness that can be improved in further research. Based on the research results and conclusions that have been made, the authors provide suggestions for further research. The suggestions that can be considered for the development of this research are adding data description of book transactions so that customer segmentation can be carried out based on book purchases. In addition to adding customer demographic data such as gender and age, can adding socio-economic characteristics, such as monthly income, level of education, and employment so as to produce a more specific marketing strategy.

6. References

[1] Baarsch, J., Celebi, M.E., 2012. Investigation of Internal Validity Measures forK-Means Clustering.
[2] Dangeti, P., 2017. Statictics for Machine Learning. Packt.
[3] Dogan, O., Aycin, E., Bulut, Z.A., 2018. Customer Segmentation by Using Rfm Model and Clustering Methods: A Casestudy in Retail Industry. Int. J. Contemp. Econ. Adm. Sci. 8, 1–19.
[4] Dursun, A., Caber, M., 2016. Using data mining techniques for profiling profitable hotel customers: An application of RFM analysis. Tour. Manag. Perspect. 18.
[5] Ha, S.H., Park, S.C., 1998. Application of data mining tools to hotel data mart on the Intranet fordatabase marketing. Expert Syst. Appl. 15, 1–13.
[6] Han, J., Kamber, M., Pei, J., 2012. Data Mining Concepts and Techniques Third Edition. Morgan Kaufmann.
[7] Larose, D.T., Larose, C.D., 2015. Data Mining and Predictive Analysis. Wiley.
[8] Linoff, G.S., Berry, M.J.A., 2015. Data Mining Techniques for Marketing, Sales, and Customer Relationship Management Third EDition, 3rd ed. Wiley.
[9] Marcus, C., 1998. A practical yet meaningful approach to customersegmentation. J. Consum. Mark. 15, 494–504.
[10] Peker, S., Kocyigit, A., Eren, P.E., 2017. LRFMP model for customer segmentation in the grocery retail industry: a case study.

Mark. Intell. Plan. 35, 544–559.
[11] Tan, P.-N., Steinbach, M., Karpatne, A., Kumar, V., 2019. Introduction to Data Mining Second Edition. Pearson.
[12] Tavakoli, M., Molavi, M., Masoumi, V., Mobini, M., Etemad, S., Rahmani, R., 2018. Customer Segmentation and Strategy Development Based on User Behavior Analysis, RFM Model and Data Mining Techniques: A Case Study. 2018 IEEE 15th Int. Conf. E-bus. Eng.
[13] Tsiptsis, K., Chorianopoulos, A., 2010. Data Mining Techniquesin CRM Inside Customer Segmentation. Wiley.
[14] Wei, J.T., Yang, Y.-Z., Lin, S.-Y., Wu, H.-H., 2016. Applying Data Mining and RFM Model to Analyze Customers Values of A Veterinary Hospital. 2016 Int. Symp. Comput. Consum. Control.
[15] XYZ, 2020. Data Warehouse.
[16] Zumel, N., Mount, J., 2019. Practical Data Science with R Second Edition. Manning Pu-blications Co.