Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 1, 2021
DOI : https://doi.org/10.33395/sinkron.v6i1.11191
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Clustering algorithm for determining marketing targets based customer purchase patterns and behaviors

Amir Mahmud Husein [1)*], Februari Kurnia waruwu [2)], Yacobus M.T. Batu Bara [3)], Meleyaki Donpril[4)], Mawaddah Harahap[5)]
[1)2)3,4,5)]Universitas Prima Indonesia, Indonesia
[1)]amirmahmud@unprimdn.ac.id, [2)]arywaruwu142@gmail.com, [3)] yacobusbatubara5@gmail.com,
[4)]Meleyakidonpril95@yahoo.com, [5)]mawaddah@unprimdn.ac.id

Submitted : Oct 13, 2021 | **Accepted** : Oct 17, 2021 | **Published** : Oct 19, 2021

Abstract: Customer segmentation is one of the most important applications in the business world, specifically for marketing analysis, but since the Corona Virus (Covid-19) spread in Indonesia it has had a significant impact on the level of digital shopping activities because people prefer to buy their needs online, so It is very important to predict customer behavior in marketing strategy. In this study, the K-Means Clustering technique is proposed on the RFM (Recency, Frequency, Monetary) model for segmenting potential customers. The proposed model starts from the data cleaning stage, exploratory analysis to understand the data and finally applies K-Means Clustering to the RFM Model which produces three clusters based on the Elbow model. In cluster 0 there are 2,436 customers, in cluster1 1,880 and finally in cluster2 there are 18 customers. RFM analysis can segment customers into homogeneous groups quickly with a minimum set of variables. Good analysis can increase the effectiveness and efficiency of marketing plans, thereby increasing profitability with minimum costs.

Keywords: K-Means Clustering, Customer Segmentation, RFM Model, Marketing analysis, marketing.

INTRODUCTION

Indonesia is currently one of the countries in Southeast Asia that has a large market potential in the economic sector, especially the growth of online trade. This is related to the courier industry that delivers goods to consumers and the warehousing sector as a place to store goods, besides that since the Corona Virus (Covid-19) has spread in Indonesia, it has had a significant impact on the level of digital shopping activities because people prefer to buy their needs online. This is in line with the implementation of government policies, namely work from home or work from home (WFH) and the extension of the study period at home. The current increase in the growth of people's digital shopping will have an effect on dynamic changes in customer buying behavior, so it is very important to predict behavior. customers in marketing strategy.

Customer Segmentation is one of the most important applications in the business world, especially for Marketing analysis. By using clustering techniques, companies can identify several customer segments that allow them to target a potential user base (consumers). The application of data mining is one of the models proposed by many researchers because it has the ability to find hidden knowledge about the relevance of large amounts of online transaction data(J. Wu et al., 2020).

Cluster analysis is one of the data mining algorithms that is widely applied to group customer databases so that customers in clusters are similar, and as different as possible from customers in other clusters (S. G. Carbajal, 2021). Clustering algorithms such as K-Means are proposed by (Dedi, 2019) for segmenting customer data based on RFM (Recency, Frequency and Monetary) values and (O. Piskunova, 2020) on e-commerce datasets while work (E. Lee, 2020)] applies customer segmentation for an effective housing demand response program and (M. P. Fernandes, 2017) for electricity demand.

The application of customer segmentation has an important role in determining marketing targets, in addition to aiming to classify potential customers, customer segmentation can also be applied to investigate the impact of consumer preferences on the intensity of competition for companies (T.-Y. Ou, 2021) the application of analysis

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 1, 2021
DOI : https://doi.org/10.33395/sinkron.v6i1.11191
e-ISSN : 2541-2019
p-ISSN : 2541-044X

and visuals is also very useful in e-commerce business. for market segmentation based on geographic distribution of user behavior (D. Kamthania, 2018).

In this study, customer segmentation analysis is proposed to determine marketing targets based on the behavior of customers' purchasing transactions that can change dynamically. The analytical framework proposed the elbow method to determine the optimal K value, then the K-Means Clustering algorithm was applied to identify population groups in order to see that several clusters could represent the characteristics of the company's existing customer base.

LITERATURE REVIEW

Some of the latest research works in the field of market segmentation that should be considered to conduct an analysis of customer needs, including (J. Wu et al., 2020) conduct empirical research on customer segmentation based on Behavior Purchase. The author applies RFM and the K-Means Algorithm to the world's data real from a company in Beijing, China. Customers are classified into four groups based on buying behavior including total volume increase purchases and total consumption. Other works reported by (S. G. Carbajal, 2021) discuss automatic classification based on movement around the center of the sports shop. The authors defined variables that could be calculated from the data comes from the indoor positioning system, which has been carried out by the procedure automatic grouping so as to generate relevant knowledge from data customer positioning and (K. Baek, 2021) implement customer data segmentation requested Electric Vehicle Charger.

3. METHOD

This study uses data obtained from a UK-based registered non-store online retail e-commerce dataset, this dataset is provided by Dr. Daqing Chen, Director of the Public Analysis group. The dataset used is 541,910, below is a partial view of the dataset used.

Table 1. Dataset provided by Dr. Daqing Chen

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/01/2010 08:26	2.55	17850	United Kingdom
536365	71053	WHITE METAL LANTERN	6	12/01/2010 08:26	3.39	17850	United Kingdom
536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/01/2010 08:26	2.75	17850	United Kingdom
536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/01/2010 08:26	3.39	17850	United Kingdom
536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/01/2010 08:26	3.39	17850	United Kingdom
536365	22752	SET 7 BABUSHKA NESTING BOXES	2	12/01/2010 08:26	7.65	17850	United Kingdom
536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	12/01/2010 08:26	4.25	17850	United Kingdom
536366	22633	HAND WARMER UNION JACK	6	12/01/2010 08:28	1.85	17850	United Kingdom
536366	22632	HAND WARMER RED POLKA DOT	6	12/01/2010 08:28	1.85	17850	United Kingdom
536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	12/01/2010 08:34	1.69	13047	United Kingdom
536367	22745	POPPY'S PLAYHOUSE BEDROOM	6	12/01/2010 08:34	2.1	13047	United Kingdom
536367	22748	POPPY'S PLAYHOUSE KITCHEN	6	12/01/2010 08:34	2.1	13047	United Kingdom
536367	22749	FELTCRAFT PRINCESS CHARLOTTE DOLL	8	12/01/2010 08:34	3.75	13047	United Kingdom
536367	22310	IVORY KNITTED MUG COSY	6	12/01/2010 08:34	1.65	13047	United Kingdom
536367	84969	BOX OF 6 ASSORTED COLOUR TEASPOONS	6	12/01/2010 08:34	4.25	13047	United Kingdom
536367	22623	BOX OF VINTAGE JIGSAW BLOCKS	3	12/01/2010 08:34	4.95	13047	United Kingdom
536367	22622	BOX OF VINTAGE ALPHABET BLOCKS	2	12/01/2010 08:34	9.95	13047	United Kingdom
536367	21754	HOME BUILDING BLOCK WORD	3	12/01/2010 08:34	5.95	13047	United Kingdom
536367	21755	LOVE BUILDING BLOCK WORD	3	12/01/2010 08:34	5.95	13047	United Kingdom
536367	21777	RECIPE BOX WITH METAL HEART	4	12/01/2010 08:34	7.95	13047	United Kingdom

:

.

Procedures are a series of procedures for carrying out a study in which each stage is interconnected so that a research can be completed in accordance with the predetermined target. The following is an explanation of the work procedures in this study:

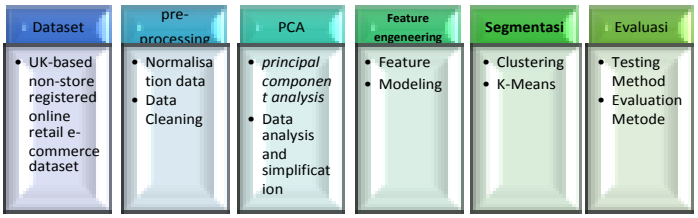


Fig. 1 Procedure Work

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 1, 2021
DOI : https://doi.org/10.33395/sinkron.v6i1.11191
e-ISSN : 2541-2019
p-ISSN : 2541-044X

RESULT

This research proposes a K-Means grouping technique that is applied to improve the results of customer segmentation on the RFM (Recency, Frequency, Monetary) model based on the elbow method to determine the optimal K value.

Following is the Dataset used in the study which is data on transactions that took place between 01/12/2010 and 09/12/2011 for a UK based and registered non-store online retailer selling gifts. Data obtained at https://www.kaggle.com/carrie1/ecommerce-data. Overall the dataset consists of 541,909 rows and 8 columns.

Table 2. Dataset transacion

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850	United Kingdom
...	...	...	...	...	...	...	...	...
541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	12/9/2011 12:50	0.85	12680	France
541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	12/9/2011 12:50	2.10	12680	France
541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	12/9/2011 12:50	4.15	12680	France
541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	12/9/2011 12:50	4.15	12680	France
541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	12/9/2011 12:50	4.95	12680	France

541909 rows × 8 columns

The first step is to clean the data so that it can be processed better so that it is more efficient for further analysis. In summary the data set has some odd and irregular values in the 'UnitPrice' and 'Quantity' columns. Negative values can be due to order cancellation or refund, this should be removed as it will negatively affect the analysis. After all the data analysis needs have been adjusted to the needs, the next step is transaction analysis to understand the data.

Table 2. Dataset final

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850	United Kingdom
...	...	...	...	...	...	...	...	...
406824	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	2011-12-09 12:50:00	0.85	12680	France
406825	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	2011-12-09 12:50:00	2.10	12680	France
406826	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	2011-12-09 12:50:00	4.15	12680	France
406827	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	2011-12-09 12:50:00	4.15	12680	France
406828	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	2011-12-09 12:50:00	4.95	12680	France

387883 rows × 8 columns

In table 2. it can be seen that the data quantity, unit price and date are in accordance with the needs of further analysis. Overall the dataset consists of 541,909 rows and 8 columns, then the results of cleaning the data are 387,883 rows and 8 columns so that there are 154,026 rows of data that have been deleted. For analysis needs, several columns were added, namely TotalPrice, Year, Quarter, Month, Week, Weekday, Day, Dayofyear and Date with Year, Month, Day format. The addition of a new column aims to make it easier to display visual data. The following is a visual data display of the 10 best customers.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 1, 2021
DOI : https://doi.org/10.33395/sinkron.v6i1.11191
e-ISSN : 2541-2019
p-ISSN : 2541-044X

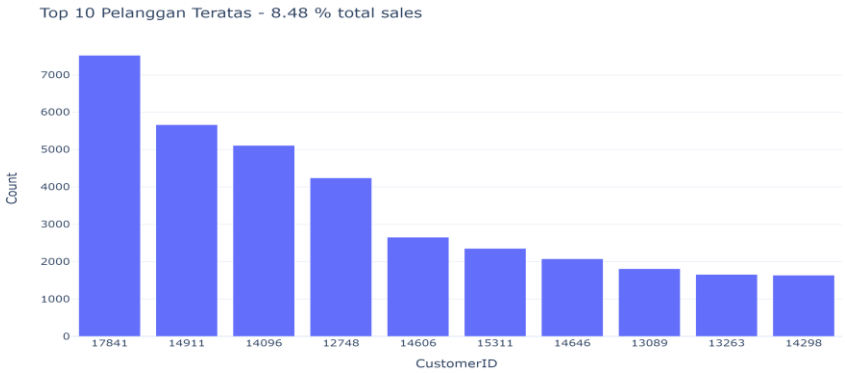


Fig. 2 The 10 best customers.

Fig. 2 is the data for the top 10 customers who make transactions with a total transaction value of 8.48%, the next analysis of daily product sales is presented in Fig. 3

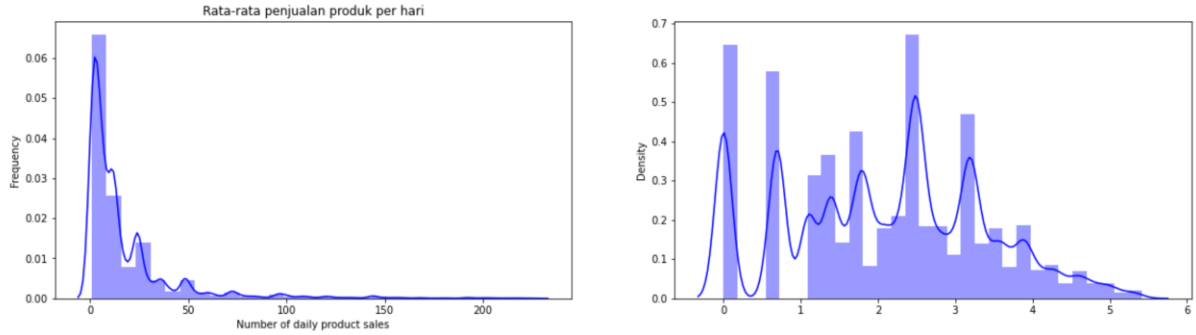
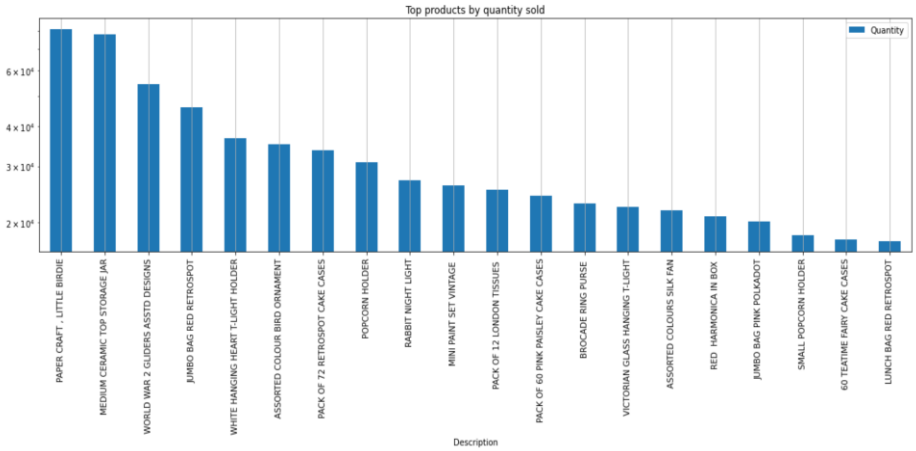


Fig. 3 Average daily product sales.

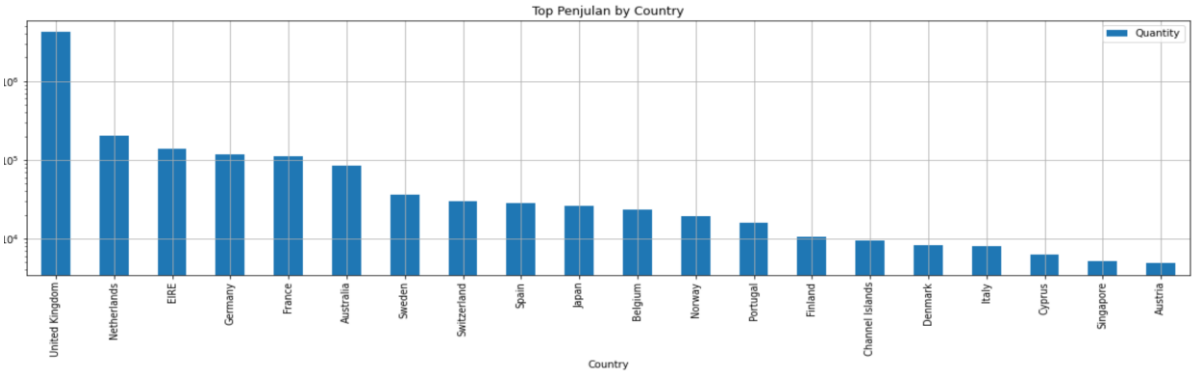
In Fig. 3 it can be seen that the distribution is skewed to the right. Lower values are more common. In addition, the number of daily sales appears to be multimodal. Daily sales of 1 are common as well as quantities of 12 and 24. This pattern is very interesting and leads to the conclusion that quantities are often divisible by 2 or 3. In short it can be said that certain products are often purchased as single quantites or in small groups. The last data analysis is which product is the best-selling and which country has the highest transaction value, this information is shown in Fig. 4.



(a)

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 1, 2021
DOI : https://doi.org/10.33395/sinkron.v6i1.11191
e-ISSN : 2541-2019
p-ISSN : 2541-044X



(b)

Fig. 4 (a) Best Selling Products, (b) Top Selling by country.

In Fig. 4. it can be seen that in picture (a) Paper Craft products, Little Birdie with Medium Ceramic Top Storage Jar products are the two best-selling products with almost the same level of sales, while United Kingdom is one of the highest countries.

The next step is to analyze the RFM model using K-Means Clustering, the first step is to make data transformations to determine the features used. Therefore it is necessary to calculate the total number of different products that customers buy in a year from the store, calculate the value of Retention and Revenue, Total Transactions and Products that are different each time a transaction occurs. The overall data structure for the analysis of the RFM model can be seen in table 3

Table 3. Overall data for RFM model model analysis.

	CustomerID	Revenue	Tot_Trans	uniqueproducts	Quantity	Retention
0	12346	77183.60	1	1	74215	1
1	12347	4310.00	7	103	2458	7
2	12348	1437.24	4	21	2332	4
3	12349	1457.55	1	72	630	1
4	12350	294.40	1	16	196	1
...	...	...	...	...	...	...
4329	18280	180.60	1	10	45	1
4330	18281	80.82	1	7	54	1
4331	18282	178.05	2	12	103	2
4332	18283	1992.73	16	258	1317	16
4333	18287	1837.28	3	59	1586	3

4334 rows × 6 columns

In table 3 it can be seen that a total of 4,334 rows need to be analyzed to obtain a potential customer model, this is because some transaction data are not included in the criteria. The results of the correlation plot between variables are shown in Fig. 5.

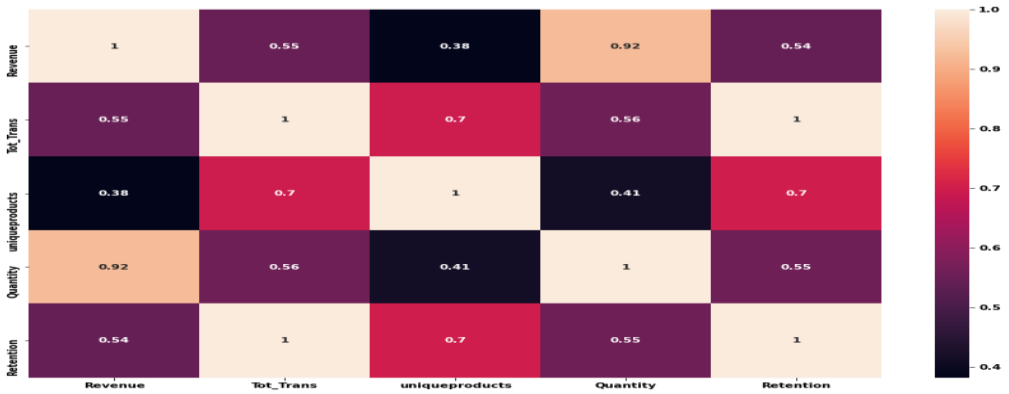


Fig. 5 Variable Correlation.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 6, Number 1, 2021
DOI : https://doi.org/10.33395/sinkron.v6i1.11191
e-ISSN : 2541-2019
p-ISSN : 2541-044X

In Fig. 6 it can be seen that Customer Retention is highly correlated with Total Transactions and Unique Products purchased. The more unique the product and the customer buys, the sooner he will return to the store. The next step is to change the data in such a way that the distribution will have an average value of 0 and a standard deviation of 1, then look for the optimum k value in the data, the search results are shown in Fig. 6.

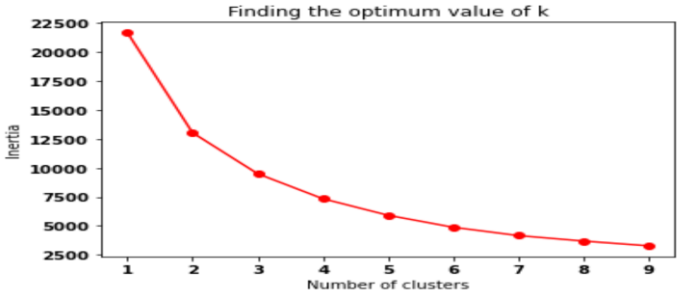


Fig. 6 Finding K Value in K-Means Clustering.

Fig. 7. is a graph of finding the optimal K value using the elbow method, it can be seen that the best K value is in the 4th cluster, but the sample value used is very small and will be analyzed by applying k=3 and k=4. Then apply K-Means Clustering to group the RFM model, the results of which can be seen in Fig. 7.

	Revenue	Tot_Trans	uniqueproducts	Quantity	Retention	labels
1	0.257708	0.360789	0.498900	0.253086	0.366611	0
2	-0.064506	-0.032189	-0.470663	0.228075	-0.029654	0
3	-0.062228	-0.425167	0.132358	-0.109763	-0.425920	0
5	-0.070282	0.360789	-0.045001	-0.130407	0.366611	0
6	-0.215727	-0.425167	-0.671670	-0.230846	-0.425920	0
...	...	...	...	...	...	...
4327	-0.213329	-0.425167	-0.624374	-0.221318	-0.425920	0
4328	-0.206204	-0.425167	-0.612550	-0.221715	-0.425920	0
4329	-0.205453	-0.425167	-0.600727	-0.225883	-0.425920	0
4330	-0.216644	-0.425167	-0.636198	-0.224097	-0.425920	0
4331	-0.205739	-0.294175	-0.577079	-0.214370	-0.293831	0

2436 rows × 6 columns

	Revenue	Tot_Trans	uniqueproducts	Quantity	Retention	labels
4	-0.192689	-0.425167	-0.529783	-0.195910	-0.425920	1
7	-0.104642	-0.425167	-0.033177	-0.129613	-0.425920	1
8	-0.174182	-0.425167	-0.565255	-0.187177	-0.425920	1
9	0.053285	-0.163182	-0.104121	0.077417	-0.161743	1
10	0.470554	-0.425167	0.829970	0.302710	-0.425920	1
...	...	...	...	...	...	...
4324	-0.202828	-0.163182	-0.707142	-0.218936	-0.161743	1
4325	-0.205978	-0.425167	-0.588903	-0.217348	-0.425920	1
4326	-0.188039	-0.425167	-0.553431	-0.197895	-0.425920	1
4332	-0.002201	1.539724	2.331611	0.026603	1.555406	1
4333	-0.019637	-0.163182	-0.021353	0.079998	-0.161743	1

1880 rows × 6 columns

(a) (b)

	Revenue	Tot_Trans	uniqueproducts	Quantity	Retention	labels
0	8.431341	-0.425167	-0.707142	14.496505	-0.425920	2
55	13.745671	2.063695	4.519040	15.123353	1.951672	2
326	3.256633	26.428349	19.972929	4.654320	26.784290	2
562	6.358000	12.150138	6.789236	5.914565	11.726213	2
996	7.069251	5.993478	3.608596	12.332312	6.046412	2
1333	12.901766	6.517449	7.687855	11.206050	6.442677	2
1434	5.553685	5.207521	9.733397	11.345989	5.253882	2
1661	1.102920	11.233188	8.917545	0.983945	11.594125	2
1689	31.082909	8.875319	7.545968	38.837771	8.952357	2
1879	15.501946	25.380407	20.386767	15.670405	25.595495	2
2176	6.559810	11.364181	5.973384	7.327851	11.462036	2
2700	7.909844	7.565391	-0.210537	7.706382	7.367296	2
3006	18.670467	-0.294175	-0.683494	15.842699	-0.293831	2
3174	7.250273	3.111637	0.688083	9.740560	3.008379	2
3725	21.559563	5.469507	0.747203	13.650518	5.518058	2
3768	9.988010	3.504616	4.637279	12.577850	3.668821	2
4007	4.269777	15.686942	14.876811	4.250580	15.820953	2
4197	28.897917	7.303406	1.054625	12.493490	6.706854	2

(c)

Fig. 9. Cluster Results (a) Cluster 0, (b) Cluster 1, (c) Cluster 2.

In Fig. 9. is the result of grouping K-Means based on the Elbow model for customer segmentation of the RFM model, it can be seen that for cluster0 there are 2,436 customers who have the highest transaction potential, while

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

in cluster1 1,880 and finally in cluster2 there are 18 customers who have the lowest transaction value, so this information is very important to be improved in marketing.

DISCUSSIONS

How RFM analysis can segment customers into homogeneous groups quickly with a minimum set of variables. Scoring systems can be defined and ranged differently. We get better results for the clustering step by applying the assessments than using the raw calculated RFM values. Therefore, segmentation should be carried out by RFM assessment and further analysis of spending behavior should be carried out on raw values for targeted clusters to expose more insights and characteristics. RFM analysis relies solely on buying behavior and history, the analysis can be further improved by exploring weighted composite scores or incorporating customer demographic information and product information. A good analysis can increase the effectiveness and efficiency of marketing plans, thereby increasing profitability with a minimum cost.

CONCLUSION

Based on the results of testing the application of the K-Means Clustering algorithm on the RFM Model for potential customer segmentation, the data draws several conclusions, namely:

a. Determination of value Optimal K with the Elbow method can improve better optimization on the K-Means Clustering algorithm.
b. The application of the K-Means Clustering algorithm on the RFM model for customer segmentation produces 3 cluster groups, namely cluster0 totaling 2,436 customers who have the highest transaction potential, while in cluster1 1,880 and finally in cluster2 there are 18 customers who have the lowest transaction value
c. RFM analysis can segment customers into homogeneous groups quickly with a minimum set of variables
d. RFM analysis solely depends on behavior and purchase history, analysis can be further improved n by exploring a weighted composite score or incorporating customer demographic information and product information.

REFERENCES

D. Kamthania, A. Pahwa, and S. S. Madhavan, “Market segmentation analysis and visualization using K-mode clustering algorithm for E-commerce business,” J. Comput. Inf. Technol., vol. 26, no. 1, pp. 57–68, 2018, doi: 10.20532/cit.2018.1003863.

Dedi, M. I. Dzulhaq, K. W. Sari, S. Ramdhan, R. Tullah, and Sutarman, “Customer Segmentation Based on RFM Value Using K-Means Algorithm,” Proc. 2019 4th Int. Conf. Informatics Comput. ICIC 2019, 2019, doi: 10.1109/ICIC47613.2019.8985726.

E. Lee, J. Kim, and D. Jang, “Load profile segmentation for effective residential demand response program: Method and evidence from Korean pilot study,” Energies, vol. 16, no. 3, p. 1348, Mar. 2020, doi: 10.3390/en13061348.

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. November 2017, 2020, doi: 10.1155/2020/8884227.

K. Baek, S. Kim, E. Lee, Y. Cho, and J. Kim, “Data-Driven Evaluation for Demand Flexibility of Segmented Electric Vehicle Chargers in the Korean Residential Sector,” Energies, vol. 14, no. 4, p. 866, Feb. 2021, doi: 10.3390/

M. P. Fernandes, J. L. Viegas, S. M. Vieira, and J. M. C. Sousa, “Segmentation of residential gas consumers using clustering analysis,” Energies, vol. 10, no. 12, p. 2047, Dec. 2017, doi: 10.3390/en10122047.

O. Piskunova and R. Klochko, “Classification of e-commerce customers based on Data Science techniques,” CEUR Workshop Proc., vol. 2649, pp. 6–20, 2020.

S. G. Carbajal, “Customer segmentation through path reconstruction,” Sensors, vol. 21, no. 6, pp. 1–17, Mar. 2021, doi: 10.3390/s21062007.

T.-Y. Ou and Y. J. Chen, “Optimal Segmentation over a Generalized Customer Distribution,” Axioms, vol. 10, no. 2, p. 98, May 2021, doi: 10.3390/axioms10020098.

*name of corresponding author