Journal of Information Systems and Informatics

Vol. 4, No. 3, September 2022 e-ISSN: **2656-4882** p-ISSN: **2656-5935**

http://journal-isi.org/index.php/isi Published By DRPM-UBD

Edutech Digital Start-Up Customer Profiling Based on RFM Data Model Using K-Means Clustering

Dedy Panji Agustino[1], I Gede Harsemadi[2], I Gede Bintang Arya Budaya[3]

[1,2] Information System Department, Institute of Technology and Business STIKOM Bali, Denpasar, Indonesia

[3] Information Technology Department, Institute of Technology and Business STIKOM Bali, Denpasar, Indonesia

Email: [1]panji@stikom-bali.ac.id, [2]harsemadi@stikom-bali.ac.id, [3]bintang@stikom-bali.ac.id

Abstract

Digital start-up is companies with a high risk because they are still looking for the most fitting business model and the right market. The company's growth is the primary goal of the start-up. As a newly established company, digital start-ups have one challenge, it is the ineffectiveness of the marketing process and strategic schemes in terms of maintaining customer loyalty, the same goes for edutech digital start-ups. Ineffective and inefficient plans can waste resources. Hence, a method is needed to find out the optimal solution to understanding the customer characteristic. Business Intelligence is needed, with the customer profiling process using transaction data based on the RFM (Retency, Frequency, Monetary) model using the K-Means algorithm. In this study, the transaction data comes from an education platform digital start-up assisted by the STIKOM Bali business incubator. Based on three metrics, namely the Elbow Method, Silhouette Scores, and Davis Bouldin Index, transaction data for sales retency, sales frequency, and sales monetary can be analyzed and can find the optimal solution. For this case, K = 2 is the optimum cluster solution, where the first cluster is the customer who needs more engagement, and the second cluster is the best customer.

Keywords: Customer Segmentation, Silhouette Coefficient, Elbow Method, Davies Bouldin Index, Business Intelligences

1. INTRODUCTION

A digital start-up is a collection of individuals who form an organization as a digital company that produces products in the field of technology [1]. Digital start-up is companies with a high risk, because they are still looking for the most fitting business model and the right market. The company's growth is the primary goal of the start-up even though the company still has to burn the investment money in the early period. However, it cannot be denied that the ultimate goal is still profit, the earlier start-ups can enter and understand the market, and also validate their products, the faster the goal of making a profit will occur, the same goes for education technology (edutech) digital start-ups. One of the first steps to quickly find out whether the digital start-up product is fit or not and also increase the benefits of digital start-up products to the market is to understand customer characteristics.

The increasingly complex business ecosystem also encourages intense business competition, including for digital startups. There is a need to maintain customer loyalty and long-term relationships between customers. At least customers who have shopped or at least know about the products from digital start-ups can help introduce them to others. They also have the potential to reuse and repurchase the products from these digital start-ups. Hence, the implementation of the customer profiling helps to find out which customers include in the profitable category. Customer profiling through segmenting customer information is one of the activities of the customer relationship management (CRM) series [2]–[4]. In terms of customer profiling with the customer data segmentation process, customers are divided based on individuals who have some information in common based on historical transactions that can lead to desired results, for example, increased sales and profits for the company [5].

As a newly established company, digital start-ups have one challenge, it is the ineffectiveness of the marketing process and strategic schemes in terms of maintaining customer loyalty. Ineffective and inefficient plans can wasted the resources. Hence, a method is needed to find out the similarity of customer profiles from digital start-up so they can plan appropriate strategies, especially for the marketing process [6], and maintain relationships with their customers to increase business profits as a company. The customer profiling process can be utilized using transaction data based on the RFM (Retency, Frequency, Monetary) model using the K-Means algorithm [5], [7].

K-Means is an algorithm in machine learning that performs the modeling process without supervision (unsupervised) and is one of the algorithms that performs data grouping with a partition system. The K-Means algorithm tries to group the existing data into several groups, where the data in one group have the same characteristics and has different characteristics from the data in other groups. In other words, this algorithm seeks to make variations between data in one cluster and maximize variation with data in other clusters [8]. In this study, it is stated that K-Means is the most widely used algorithm, this is based on efficiency, ease of application, ease of use, and empirical success of the clustering result [9], [10].

In this study, the transaction data comes from a digital start-up assisted by the STIKOM Bali business incubator, *Visual Learning*. *Visual Learning* is an educational platform that provides teaching about design and videography. The digital start-up is using the concept of a one-time purchasing or subscription selling model for each learning video. The prices used are flat with different discount values. This study also wants to find how the RFM model with K-Means can identify the most widely used digital start-up business model - the subscription/one time purchasing model. The aim is to understand the characteristic of this specific business customer.

2. METHODS

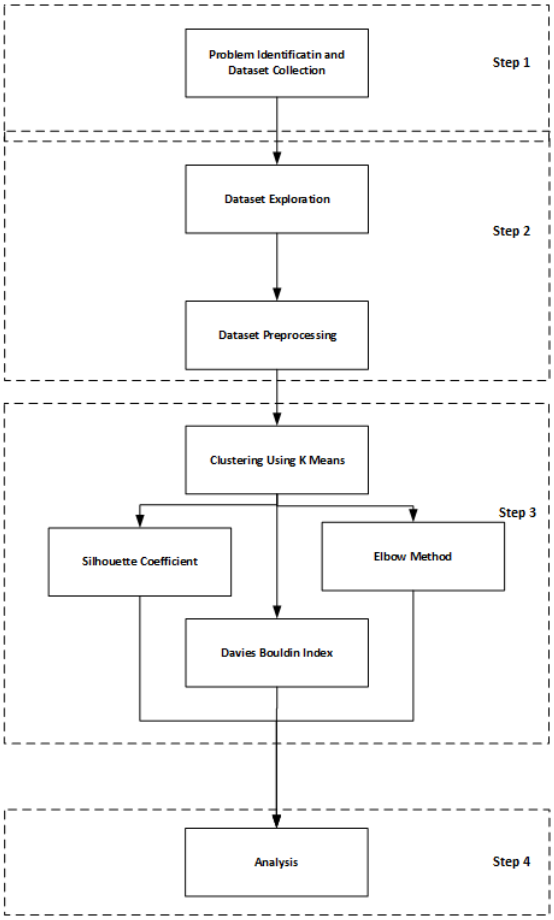

Figure 1. Research method.

The research method consists of four steps as shown in Figure 1. The first step of this research is to carry out the problem identification process and dataset collection, this aims to find out more about the profile of digital start-ups and find out information that can be a variable, related to the business activities of the tenant concerned.

The second steps are dataset exploration and pre-processing. The dataset used is a dataset of transaction records owned by digital start-ups for 1 year, namely from July 2021 – June 2022. Next, the data exploration and selection process, which in this process aims to form a dataset according to the needs of the RFM (Recency, Frequency, Monetary) [11], because not all variables were used in this study. Furthermore, the dataset that has been selected enters the dataset pre-processing stage, where the normalized process and dataset cleaning process is carried out [12].

For the third step, the cleaned dataset is used for the clustering process using the K-Means method [5] with a scenario of 2 to 10 clusters which is in. The third step is the process followed by carrying out the validity test using the Elbow method [13], Silhouette coefficient [14] and Davis Bouldin index [15] to find the optimal cluster. The fourth step is an analysis of the results for determining the validity of clustering and determining which cluster number is the best in terms of the customer profiling process by utilizing K-Means.

3. RESULTS AND DISCUSSION

3.1 Dataset Collection

The dataset comes from the business transaction by strat-up *Visual Learning*, an educational platform start-up for learning visual design and videography. *Visual Learning* is tenant that assisted by STIKOM Bali business incubator. The digital start-up is using the concept of a one-time purchasing or subscription selling model for each learning video. The prices used are flat with different discount values at a time.

3.2 Dataset Exploration

The datasets that have been collected are then explored and pre-processed base on the methodology. The dataset consists of 283 rows of transaction data from the period 15-07-2021 to 27-06-2022 with five attributes. Table 1 displays the dataset description.

Table 1. Dataset description

No	Name of Attribute	Type of Attribute	Description of the Attribute
1	Customer Email	Character	The email which the customer has registered in the platform during account registration as an uniqe ID.
2	Transaction ID	Nominal	The ID of transaction

3	Transaction Date	Numeric	Date and time of each transaction generated
4	Type of Transfer	Character	The chosen media for transferring the subscription price money
5	Amount	Numeric	The price money for the product subcription

The dataset is then processed to make an RFM data frame model. The first is recency data. The recency data about how many days the last purchase of the customer. The recency data is calculated by subtracting the recent date from the last transaction date by the customers. The recent date was set exactly a day after the last transaction, for this case 28-06-2022. Table 2 shows the result of the recency calculation for each transaction.

Table 2. Recency calculation result

No	Customer Email	Recency
1	a.prasetyo.b86@gmail.com	41
2	abdulrajabrjr@gmail.com	20
3	abilliondesain11@gmail.com	118
4	activespin7@gmail.com	133
…	…	…
283	adamnvt35@gmail.com	211

The second is frequency data. The calculation of frequency by counting the transaction number of each customer, the higher the number it means the higher the customer spends their money or uses the product of the start-up. Table 2 shows the frequency data for the customer.

Table 3. Frequency calculation result

No	Customer Email	Frequency
1	a.prasetyo.b86@gmail.com	1
2	abdulrajabrjr@gmail.com	1
3	abilliondesain11@gmail.com	1
4	activespin7@gmail.com	1
…	…	…
283	adamnvt35@gmail.com	1

The third is monetary data. The calculation of monetary data by summing up every customer transaction amount. Table 4 shows the monetary data of the customer. Either recency, frequency, or monetary, all the calculations are grouped by the customer email as a unique ID. Then each column data merged become one RFM data frame to become a new dataset or RFM dataset.

Table 4. Monetary calculation result

No	Customer Email	Monetary

1	a.prasetyo.b86@gmail.com	99.000
2	abdulrajabrjr@gmail.com	125.000
3	abilliondesain11@gmail.com	99.000
4	activespin7@gmail.com	99.000
…	…	…
283	adamnvt35@gmail.com	99.000

The last step for dataset pre – processing is normalizing the dataset values. The dataset normalization using *sklearn* python library *Standard Scaler*. Table 5 shows the normalized values of RFM dataset.

Table 5. Normalized RFM dataset

No	Customer Email	Recency	Frequency	Monetary
1	a.prasetyo.b86@gmail.com	-1,58289	-0,10369	-0,39165
2	abdulrajabrjr@gmail.com	-1,87152	-0,10369	0,47303
3	abilliondesain11@gmail.com	-0,52461	-0,10369	-0,39165
4	activespin7@gmail.com	-0,31845	-0,10369	-0,39165
…	…			
283	adamnvt35@gmail.com	0,75357	-0,10369	-0,39165

3.3 Customer Profiling Process

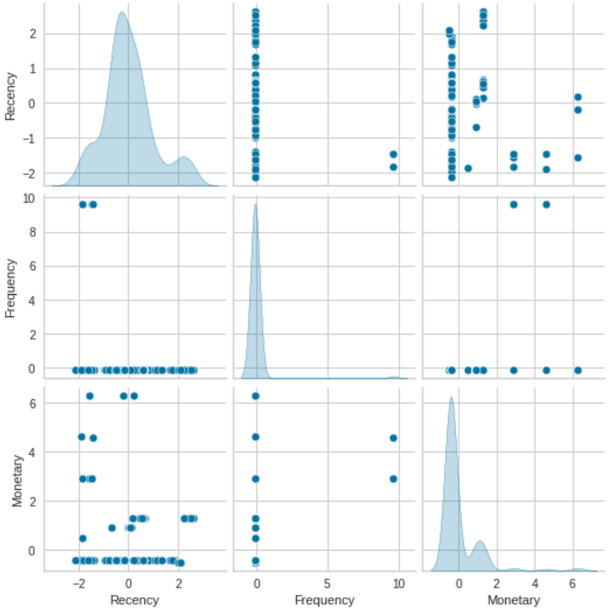

Figure 2. Data distribution.

In this study, after the dataset has gone through the pre-processing stage, the dataset is then used for the profiling process, where the profiling process uses the

K-Means algorithm for clustering. The scenario for the number of clusters is from 2 to 10 clusters. An overview of the data distribution based on the RFM model before the cluster labels are available can be seen in Figure 2.

The results of clustering using K-Means were then validated to find the optimum cluster or value of K. The validation process used three metrics, Elbow Method, Silhouette Coefficient, and Davis Bouldin index. Figure 3 shows the results of the Elbow method, Figure 4 shows the results of the Silhouette Coefficient, and Figure 5 shows the results of the Davis Bouldin Index.

The result of the Elbow Method shows that the value of cluster inertia or SSE decreases with increasing clusters. Based on the graph, the possible values for the optimal K value are between K = 2, K = 4, and K = 8. It is based on how the graphs look like they form an elbow and start to level off. To find the optimal cluster, it still needs to validate with the other two metrics.

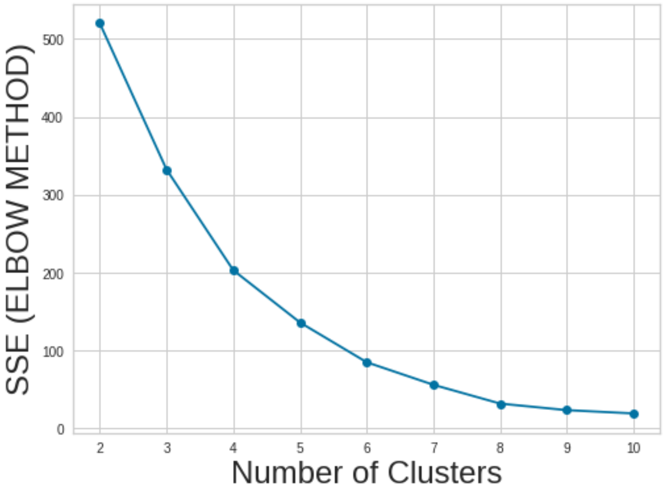

Figure 3. Elbow method result.

The second validation uses the Silhouette coefficient method with the output of silhouette score for each cluster. At K = 2, the silhouette score is 0.87. At K = 3, the silhouette score is 0.52. At K = 4, the silhouette score is 0.22. At K = 5, the silhouette score is 0.56. At K = 6, the silhouette score is 0.62. At K = 7, the silhouette score is 0.70. At K = 8, the silhouette score is 0.70. At K = 9, the silhouette score is 0.71. At K = 10, the silhouette score is 0.65. The higher the silhouette score, the better. Based on the silhouette scores, the most optimal cluster is with K = 2.

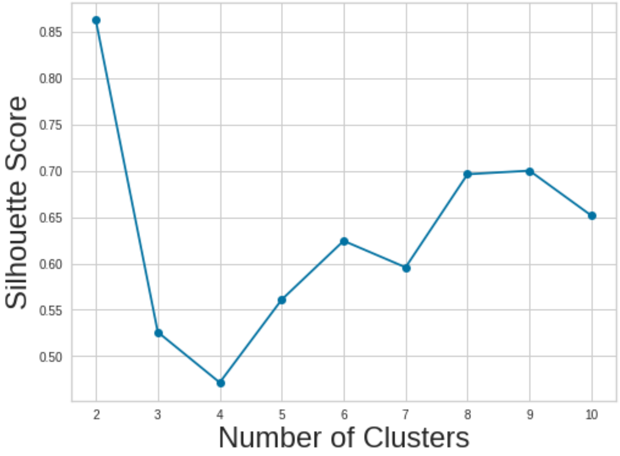

Figure 4. Silhouette scores result.

The last validation to find optimal cluster using the Davis Bouldin Index. At K = 2, the Davies Bouldin Index value is 0.11. At K = 3, the Davies Bouldin Index value is 0.78. At K = 4, the Davies Bouldin Index value is 0.61. At K = 5, the Davies Bouldin Index value is 0.63. At K = 6, the Davies Bouldin Index value is 0.42. At K = 7, the Davies Bouldin Index value is 0.42. At K = 8, the Davies Bouldin Index value is 0.35. At K = 9, the value of the Davies Bouldin Index is 0.41. At K = 10, the Davies Bouldin Index value is 0.45. The lower the Davis Bouldin Index value, the farther apart the clusters are, which means the better. Based on the Davis Bouldin Index, the optimal cluster is K = 2.

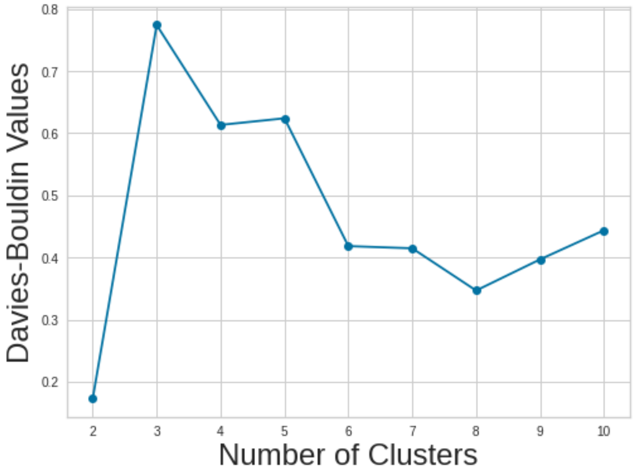

Figure 5. Davies-Bouldin index result.

Based on the results of the three metrics, K = 2 can be the optimal cluster solution. Next is to generate a new cluster using K = 2 and assign a cluster label to each row of the dataset. Table 7 shows the results of labeling with the appropriate cluster in the dataset, the cluster with code 0 is the first cluster, and the cluster with code 1 is the second cluster. Figure 6 shows a pair plot graph after the labeling process on the dataset occurs.

Table 6. Dataset cluster labeling

No	Customer Email	Recency	Frequency	Monetary	Cluster
1	a.prasetyo.b86@gmail.com	-1,58289	-0,10369	-0,39165	0
2	abdulrajabrjr@gmail.com	-1,87152	-0,10369	0,47303	0
3	abilliondesain11@gmail.com	-0,52461	-0,10369	-0,39165	0
4	activespin7@gmail.com	-0,31845	-0,10369	-0,39165	0
…	…				0
283	adamnvt35@gmail.com	0,75357	-0,10369	-0,39165	

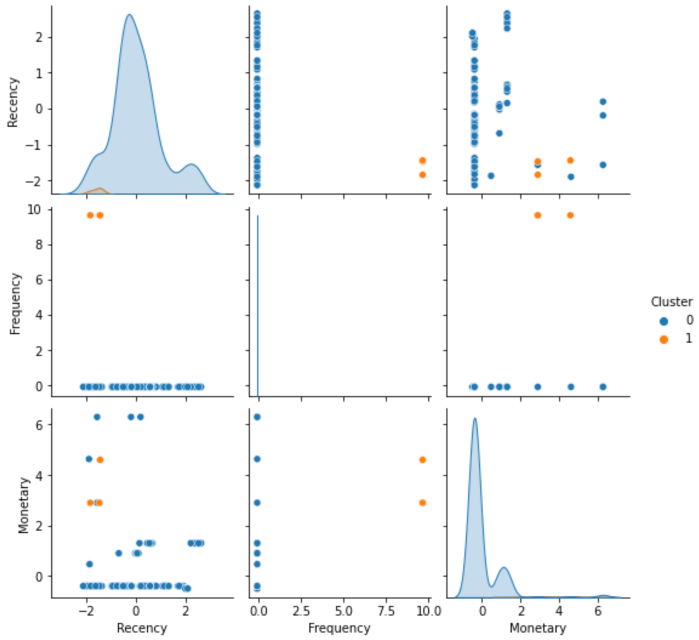

Figure 6. Labeled dataset cluster distribution.

3.4 Analysis

Based on the results of the clustering, there are two categories of clusters that can be used as the basis for digital start-ups to perform customer profiling. The first category (cluster 0) can be categorized as new customers, where these new customers are those who make a purchase for the products/services belonging to digital start-ups for the first time. In the case of this research, new customer categories need more engagement based on their needs to increase the possibility of product/service repurchase. As a platform based Edutech start-up, it is important to add learning content to the platform, coupled with appropriate teacher support, so that customer relevance and interest can increase.

As seen in Figure 7, shows more clearly how the clustering results. All transactions in the first category depicted by the purple plot are customers who transact only once, and this number dominates the whole transaction data around 280 transactions. This can be the basic information for digital start-ups, that the product/services released to the market can attract new markets massively but have not been able to make old customers repurchase the product/services.

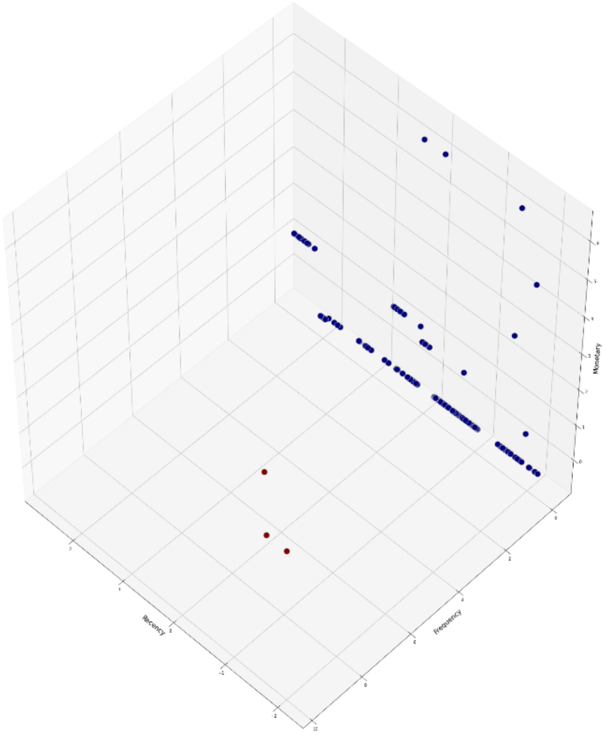

Figure 7. 3D Labeled dataset cluster distribution.

The second category (cluster 1) can be categorized as the best customers. This category is the customers who have purchased more than once and in terms of the most recent time in purchasing. Customers of this category have the potential to be offered every new product launch from a digital start-up. As well as to bind customers more strongly, special discounts can be given to customers in this category.

The second category shows in Figure 7 with a red plot. The customers who fall into this category are those who have made transactions more than once. The number is small, around 3 customers transaction, but this is still worthy to use as an information base to find out the customer motivation. For example, it can be used to design marketing strategies either for this second category customer to make them more loyal or for the first category to retain them and make them become the second category customer.

4. CONCLUSION

Customer profiling is one of the important strategies to understand the character of the customer. In this study, profiling was carried out using the concept of clustering on customer data related to the RFM data model. The results of this clustering process use to design strategies that can increase and retain customers. Based on three metrics, namely the Elbow Method, Silhouette Scores, and Davis Bouldin Index, transaction data for sales retency, sales frequency, and sales monetary can be analyzed and can find the optimal solution. Start-Ups that have a similar business model, especially an edutech, in providing products/services that are one-time purchases may experience the same challenge, namely the possibility of customers transacting again because they only need to buy the product/services in a single purchase for all the time uses. The solution is that start-ups can add new products/services that suit the customer's needs. So that customers can return to make transactions for the desired product/service. In the case of edutech, the package arrangements in the courses given should be arranged systematically and regularly. For future work, there is a need to increase the dataset size because this study only uses 283 rows of data, especially for early-stage digital startup transactions with the same business model. The aim is to understand the characteristic of this specific business customer. The clustering process can use other algorithms such as K-Medoids and Fuzzy C-Means to validate the optimum solution.

5. ACKNOWLEDGEMENT

We would like to express our gratitude to the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia for funding this research through the novice lecturer research program *(Penelitian Dosen Pemula)* in 2022. We thank the Institute of Technology and Business STIKOM Bali for supporting the implementation of the research project. We also thank the tenants of the STIKOM Bali business incubator, *Visual Learning* for allowing us to use their business transaction data as research dataset.

REFERENCES

[1] Y. Brikman, *Hello, Startup: A Programmer's Guide to Building Products, Technologies, and Teams*. " O'Reilly Media, Inc.," 2015.

[2] M. Khajvand, K. Zolfaghar, S. Ashoori, and S. Alizadeh, "Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study," *Procedia Computer Science*, vol. 3, pp. 57–63, 2011.

[3] K. Khalili-Damghani, F. Abdi, and S. Abolmakarem, "Hybrid soft computing approach based on clustering, rule mining, and decision tree analysis for customer segmentation problem: Real case of customer-centric industries," *Applied Soft Computing*, vol. 73, pp. 816–828, 2018.

[4] P. Kolarovszki, J. Tengler, and M. Majerčáková, "The new model of customer segmentation in postal enterprises," *Procedia-Social and Behavioral Sciences*, vol. 230, pp. 121–127, 2016.

[5] P. Anitha and M. M. Patil, "RFM model for customer purchase behavior using K-Means algorithm," *Journal of King Saud University - Computer and Information Sciences*, no. xxxx, 2020, doi: 10.1016/j.jksuci.2019.12.011.

[6] S. Hwang and Y. Lee, "Identifying customer priority for new products in target marketing: Using RFM model and TextRank," *Marketing*, vol. 17, no. 2, pp. 125–136, 2021.

[7] J. Wu *et al.*, "An empirical study on customer segmentation by purchase behaviors using a RFM model and K-means algorithm," *Mathematical Problems in Engineering*, vol. 2020, 2020.

[8] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," *Journal of the royal statistical society. series c (applied statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[9] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognit Lett*, vol. 31, no. 8, pp. 651–666, 2010.

[10] J. Pérez-Ortega, N. N. Almanza-Ortega, A. Vega-Villalobos, R. Pazos-Rangel, C. Zavala-Díaz, and A. Martínez-Rebollar, "The K-means algorithm evolution," *Introduction to Data Science and Machine Learning*, 2019.

[11] J.-T. Wei, S.-Y. Lin, and H.-H. Wu, "A review of the application of RFM model," *African Journal of Business Management*, vol. 4, no. 19, pp. 4199–4206, 2010.

[12] K. Coussement, F. A. M. van den Bossche, and K. W. de Bock, "Data accuracy's impact on segmentation performance: Benchmarking RFM analysis, logistic regression, and decision trees," *Journal of Business Research*, vol. 67, no. 1, pp. 2751–2758, 2014.

[13] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman, and B. D. Satoto, "Integration k-means clustering method and elbow method for identification of the best customer profile cluster," in *IOP conference series: materials science and engineering*, 2018, vol. 336, no. 1, p. 12017.

[14] D.-T. Dinh, T. Fujinami, and V.-N. Huynh, "Estimating the optimal number of clusters in categorical data clustering by silhouette coefficient," in *International Symposium on Knowledge and Systems Sciences*, 2019, pp. 1–17.

[15] A. K. Singh, S. Mittal, P. Malhotra, and Y. V. Srivastava, "Clustering Evaluation by Davies-Bouldin Index (DBI) in Cereal data using K-Means," in *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, 2020, pp. 306–310.