



Print ISSN: 1738-3110 / Online ISSN 2093-7717  
JDS website: <http://www.jds.or.kr/>  
<http://dx.doi.org/10.15722/jds.20.06.202206.1>

# Target Market Determination for Information Distribution and Student Recruitment Using an Extended RFM Model with Spatial Analysis

ERNAWATI<sup>1</sup>, Safiza Suhana Kamal BAHARIN<sup>2</sup>, Fauziah KASMIN<sup>3</sup>

Received: March 12, 2022. Revised: April 13, 2022. Accepted: June 05, 2022.

## Abstract

**Purpose:** This research proposes a new modified Recency-Frequency-Monetary (RFM) model by extending the model with spatial analysis for supporting decision-makers in discovering the promotional target market. **Research design, data and methodology:** This quantitative research utilizes data-mining techniques and the RFM model to cluster a university's provider schools. The RFM model was modified by adapting its variables to the university's marketing context and adding a district's potential (D) variable based on heatmap analysis using Geographic Information System (GIS) and K-means clustering. The K-prototype algorithm and the Elbow method were applied to find provider school clusters using the proposed RFM-D model. After profiling the clusters, the target segment was assigned. The model was validated using empirical data from an Indonesian university, and its performance was compared to the Customer Lifetime Value (CLV)-based RFM utilizing accuracy, precision, recall, and F1-score metrics. **Results:** This research identified five clusters. The target segment was chosen from the highest-value and high-value clusters that comprised 17.80% of provider schools but can contribute 75.77% of students. **Conclusions:** The proposed model recommended more targeted schools in higher-potential districts and predicted the target segment with 0.99 accuracies, outperforming the CLV-based model. The empirical findings help university management determine the promotion location and allocate resources for promotional information distribution and student recruitment.

**Keywords :** Information Distribution, RFM, Spatial Analysis, Student Recruitment, Target Market

**JEL Classification Code :** C38, C55, M31

## 1. Introduction

Besides collaborating with industry and research institutions, many universities build partnerships with high schools since their graduates are universities' primary customers. High schools can be considered student

providers for universities. In higher education marketing, the promotional effort is often not directly to target the students themselves but on key opinion formers (Tapp, Hicks, & Stone, 2004) such as high school teachers/counselors. Thus, to reach both students and teachers/counselors, higher education marketing can use a

1 First and Corresponding Author, Lecturer, Informatics Department, Universitas Atma Jaya Yogyakarta, Indonesia & Ph.D. Student, Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Malaysia. Email: [ernawati@uajy.ac.id](mailto:ernawati@uajy.ac.id)  
2 Senior Lecturer, Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Malaysia. Email: [safiza@utem.edu.my](mailto:safiza@utem.edu.my)

3 Senior Lecturer, Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Malaysia. Email: [fauziah@utem.edu.my](mailto:fauziah@utem.edu.my)

© Copyright: The Author(s)

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/4.0/>) which permits unrestricted noncommercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

direct marketing strategy for information distribution and student recruitment to the provider schools. Collaboration with high schools for promotion and student recruitment is common in Indonesia (Hidayat, Rismayati, Tajuddin, & Merawati, 2020). However, because a university has many provider schools with varying characteristics, segmenting and selecting the targeted schools is critical in a competitive environment for decision-makers to develop effective marketing strategies (Sukoroto, Haryono, & Kharisma, 2020).

Market segmentation and targeting are critical for any organization engaged in marketing and strategic planning (Peker, Kocyigit, & Eren, 2017; Simkin & Dibb, 1998) because they enable companies to target specific customers and customize the products/services to the customer preferences. Market segmentation is a popular approach for grouping customers based on customer features (Abbasimehr & Shabani, 2019). Market segmentation helps identify customer characteristics, define target customers, and develop marketing activities. Traditionally, customer behavior segmentation employs survey data. However, the adoption of information technology in business activities allows the collected data for analytics purposes. Data analytics using data mining techniques allow companies to mine the hidden knowledge and develop more efficient marketing strategies and personalize promotional offers (Roshan & Afsharinezhad, 2017).

The information used to reveal the target market's characteristics, known as the segmentation variable, is one major focus of market segmentation and targeting research. Customer segmentation variables are vital because proper segmentation variables will provide valuable customer segmentation results. The Recency-Frequency-Monetary (RFM) model is popular in customer segmentation and targeting studies (Christy, Umamakeswari, Priyatharsini, & Neyaa, 2021; Firdaus & Utama, 2021; Hwang & Lee, 2021; Kit, Firdaus, & Azmi, 2021). The RFM model can identify valuable customers and find target customers effectively based on purchase behavior (Wei, Lin, & Wu, 2010; Hwang & Lee, 2021). RFM stands for recency, frequency, and monetary, which refers to the novelty of customer relationships, the number of transactions, and the amount of money paid by customers in a specific period. A recent customer who frequently spends large sums of money is more valuable. However, RFM is based solely on customer transaction data. Some researchers modified RFM by adding new variables or redefining existing variables according to the context of the application.

Location issue is important in marketing (Libório, Bernardes, Ekel, Ramalho, & Santos, 2020) because knowing customers' geographical distribution helps businesses devise more effective marketing strategies to increase profits (Kamthania, Pahwa, & Madhavan, 2018).

Besides, the marketing strategies are also more targeted. Therefore, this study aims to extend the RFM model with geographical information for segmentation and discover the target segment to assist the decision-makers in developing a more comprehensive marketing strategy, especially in delivering promotion information and student recruitment and allocating resources more effectively.

This research adopted the RFM model for a university marketing context. Since provider schools are spread out over a vast area, spatial analysis is applied to understand the provider school districts' characteristics. The district's potential categories are assigned based on heatmap analysis and K-means clustering. The district potential variable is added to the modified RFM model. Then, the K-prototype algorithm and the Elbow method reveal the provider school segments. This work uses Geographic Information System (GIS) for mapping, analysis, and visualization. This study validates the proposed model using empirical data from an Indonesian university, and the results are compared to the CLV-based RFM model. Based on the provider schools' segmentation results, the classification technique is used to build a decision tree, then is used to predict model performance using accuracy, precision, recall, and F1-score metrics. This study contributes by redefining the RFM model to meet the university marketing context and extending the RFM model with spatial analysis. The suggested model finds the university's target market segment using empirical data. In the rest of this paper, Section 2 reviews the associated literature. Section 3 introduces the Recency-Frequency-Monetary-District's potential (RFM-D) model and its proposed framework for the application. Section 4 presents the results and discussion. Finally, Section 5 presents a brief conclusion.

## 2. Literature Review

Differentiated marketing applies different marketing actions to different market segments. A market segment is a group of people with similar characteristics and interests. Market segmentation divides customers into groups based on similar characteristics. First introduced in 1956 by Smith, it is a vital and popular marketing activity that many organizations use to understand their customers better (Abbasimehr & Shabani, 2019; Peker et al., 2017). Businesses can focus on a certain group of customers and establish long-term relationships with them when they have clearly defined customer segments. Studies on market segmentation have been widely applied in various fields, for example, e-commerce (Wu, Shi, Lin, Tsai, Li, Yang, & Xu, 2020), retail (Kit et al., 2021), banking (Abbasimehr & Shabani, 2019; Firdaus & Utama, 2021), telecommunication (Babaiyan & Sarfarazi, 2019; Hwang & Lee, 2021),

education (Davari, Noursalehi, & Keramati, 2019; Hidayat et al., 2020; Sulastri, Usman, & Syafitri, 2021).

Segmentation variables are attributes considered essential for classifying individuals, groups, or organizations into segments (Dolnicar, Grün, & Leisch, 2018). Traditional segmentation bases include demographic, geographic, behavioral, and psychographic (Hemsley-brown, 2017; Wei et al., 2010). In recent customer segmentation studies, the RFM model included in behavioral segmentation is a frequently adopted segmentation model (Christy et al., 2021; Kit et al., 2021; Wu et al., 2020). The RFM model is widely used because it requires fewer segmentation variables and is easily interpreted by managers and decision-makers (Peker et al., 2017; Rezaenia & Rahmani, 2016). However, the RFM model is incomplete (Moghaddam, Abdolvand, & Harandi, 2017) and solely based on customer transaction data. So many studies have tried to improve and modify the RFM model by adding new variables, removing others, or redefining some of the RFM variables (Peker et al., 2017; Hosseini & Mohammadzadeh, 2016).

Customer data usually contains geographic location information, and when used, it can add insight for decision-makers in developing marketing strategies. Davari et al. (2019) used market segmentation to identify segments in the professional education industry by grouping customers based on the nine most appropriate variables based on expert interviews, one of which is the customer's geographical location. But the study didn't use RFM. Another RFM researcher suggested conducting additional analysis using geographical information (Babaiyan & Sarfarazi, 2019). Several studies have included location factors such as city and area but have not been analyzed spatially (Beheshtian-Ardakani, Fathian, & Gholamian, 2018; Hidayat et al., 2020). Meanwhile, to treat their customers differently by customer locations, after clustering based on the RFM model, Rahadian & Syairudin (2020) mapped the customers' locations to see the customer segments distributions. However, to the best of our knowledge, studies that incorporate variables related to geographic information into the RFM model for customer segmentation are still infrequent, especially in the education domain.

The RFM model has been applied in the education domain. Hidayat et al. (2020) have performed schools segmentation using the RFM model. The study redefined the RFM model to segment high schools by loyalty to the university. The research used the Fuzzy C-means algorithm to cluster high schools which bound partnerships with the university and found four potential categories of partners. Although paying attention to the school area in presenting the clustering results in the form of crosstabulation, the study did not integrate spatial analysis into the RFM model. Therefore, this current study proposes a modified RFM

enriched with spatial analysis for provider schools' grouping and targeting, which is useful in supporting university decision-makers in developing marketing strategies.

### 3. Methodology

This section describes the proposed approach for discovering the university's provider school segments using the RFM-D model. This study employed enrollment data from an Indonesian university as an empirical study. The shapefiles for Indonesia's districts and provinces were obtained from Indonesia Geospatial Portal, and high school locations were collected using Google Maps for spatial analysis. This research used the R programming language for data processing and Quantum GIS software for mapping. Figure 1 presents the research framework conducted. First, data were pre-processed by identifying the necessary features and then integrating and cleaning them. After that, data transformation and aggregation were performed. The RFM variables and spatial data were then extracted for analysis using the proposed data-mining-based RFM-D approach, following these steps:

**Schools mapping and heatmap analysis.** This step conducts geocoding of provider schools' latitude and longitude data, followed by heatmap analysis. The heatmap analysis shows provider schools' density using a color gradient. This study used the Kernel Density Estimation algorithm, the most popular spatial technique for identifying hotspots, to create a heatmap. The heatmap visualizes the concentration of provider schools in the Indonesian districts.

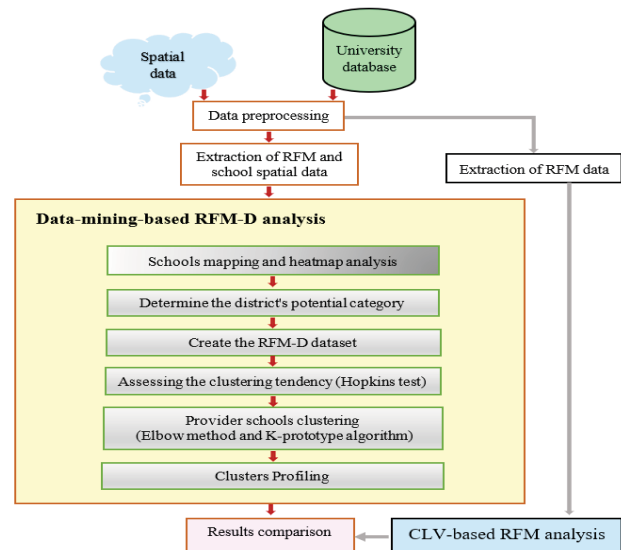


Figure 1: The research framework (Source: compiled by authors)

**Determine the potential category of a district.** The number of provider schools and the number of enrolled students from each district were computed, normalized, and grouped to determine the district's potential. The Elbow method and K-means clustering were used for grouping (Firdaus & Utama, 2021; Wu et al., 2020). The Elbow method determined the optimal number of clusters (k), and then the K-means algorithm grouped the schools' districts. Furthermore, each cluster profile was analyzed, resulting in k district potential categories.

**Create the RFM-D dataset.** This step determines the value of each provider school's RFM-D variables. The proposed RFM-D model has four variables. It enhanced the RFM model by adapting the RFM variables in the university marketing context and adding one variable indicating the potential category of the school district. Table 1 shows the definition of the RFM-D variables. The novelty and number of times a provider school makes a relationship with the university, as indicated by its alumni enrolling, respectively, are defined as the recency and frequency variables. The monetary variable is redefined by assuming that all students at the university pay the same amount of tuition fees; thus, the monetary factor is proportional to the number of students. The greater the number of enrolled students from a provider school, the greater the monetary contribution of the provider school to the university. At the same time, the district's potential category depends on the number of provider schools and enrolled students at the university from the district.

**Table 1:** The RFM-D variables definition

Variable	Definition
Recency (R)	The novelty of a provider school's most recent alumnus enrolled at the university during the analysis period.
Frequency (F)	The number of times a provider school alumni enrolled at the university during the analysis period.
Monetary (M)	The total number of alumni from a provider school enrolled at the university during the analysis period.
District's potential (D)	The potential of a district where a provider school is located.

**Assessing the clustering tendency.** This study uses the Hopkins statistic to evaluate the dataset's clustering tendency. An evaluation of clustering tendency was done to see if a data set tends to split into two or more clusters. The Hopkins tested hypothesis as follows:

**H0:** the dataset is uniformly distributed; there are no meaningful clusters in the dataset.

**H1:** the dataset is not uniformly distributed; meaningful clusters exist in the dataset.

Referring to (Lawson & Jurs, 1990), before applying the Hopkins test, the dataset value should be normalized using Z-score to have a mean of zero and a standard deviation of one. The Hopkins statistic is computed as the difference between a real point's distance from its nearest neighbor, U, and a random point in the data space's distance from the nearest actual data point, W. The Hopkins statistic, for an m sample data, is calculated using the following equation (Lawson & Jurs, 1990):

$$H = \frac{\sum_{i=1}^m u_i}{\sum_{i=1}^m u_i + \sum_{i=1}^m w_i} \quad (1)$$

Hopkins' statistic ranges from 0.5 for unclustered extreme to 1.0 for clustered extreme. If  $H > 0.75$ ,  $H_0$  is rejected with greater than 90% confidence (Lawson & Jurs, 1990).

**Provider Schools Clustering.** The K-prototype algorithm, a mixed-data type clustering, was used to obtain the provider schools segments (Huang, 1998; Sulastri et al., 2021). Because, in this context, the R, F, and M variables are numeric data types, and in contrast, the district's potential variable is a categorical data type. Before applying the K-prototype algorithm, the Elbow method was used to determine the optimum number of clusters (k).

**Clusters profiling.** The characteristics of each cluster were analyzed based on the output of the K-prototype algorithm. Knowing the clusters profile helps the decision-makers set the target segment and use them to develop marketing strategies to benefit the organization.

In the result comparison step, the RFM-D model was compared to the CLV-based RFM model (Hosseini & Mohammadzadeh, 2016). Unlike the proposed RFM-D model, the CLV-based RFM does not use clustering for provider schools grouping. After extracting RFM variables, CLV-based RFM sorts each variable's values descendingly and divides them into five quantile groups. Then, each group was assigned a score ranging from 5 to 1, yielding 125 different groups based on RFM score combinations. Subsequently, the CLV was calculated for each school using the following equation:

$$CLV_i = W_R \cdot R_i + W_F \cdot F_i + W_M \cdot M_i \quad (2)$$

$W_R$ ,  $W_F$ ,  $W_M$  are the relative weights of the RFM variables, while  $R_i$ ,  $F_i$ ,  $M_i$  are the values of the RFM variables for each school. In this case, we used  $W_R=W_F=W_M$ . The provider schools were ranked from highest to lowest CLV value. The comparison was conducted based on the target segment's characteristics and model performance. The model's performance was examined by building a predictive decision tree model (Bunnak, Thammaboosadee, & Kiattisun, 2015; Hosseini & Mohammadzadeh, 2016) using the Classification and Regression Tree (CART) algorithm.

Before forming a decision tree, the target and non-target segments were set as the class labels, and the RFM-D variables were discretized (Hosseini & Mohammadzadeh, 2016). The performance evaluation was carried out using 10-fold cross-validation methods (Bunnak et al., 2015) by measuring the accuracy, precision, recall, and F1-score metrics.

## 4. Results and Discussion

### 4.1. Data preprocessing

This study examines 18,657 student enrollment records during the analysis period (eight academic years) from a university in Daerah Istimewa Yogyakarta Province, Indonesia. After preprocessing, this study used 17,034 enrolled students' data. When students were aggregated by origin school, 2,252 provider schools were obtained. The schools were across all 34 provinces.

### 4.2. Extraction of RFM and School Spatial Data

In this step, R, F, and M values were calculated for each provider school, and recorded the school location. The high schools' latitude and longitude were obtained from Google Maps, while the district and province shapefile were obtained from the Indonesian Geospatial Portal. Table 2 shows examples of provider school data in this study. The first school (11.04.AN.001) is in Aceh Tengah (1104) with a coordinate point (4.622920, 96.846777). This school only sent one time (F=1) and one student (M=1) in the second year of the analysis period (R=2) (or six-year before the last year of the analysis period).

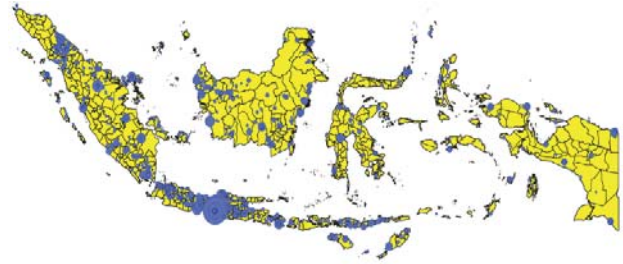
**Table 2:** Examples of provider school RFM and spatial data

No	School Code	District code	Latitude	Longitude	R	F	M
1	11.04.AN.001	1104	4.622920	96.846777	2	1	1
2	34.71.AS.029	3471	-7.784089	124.84746	8	8	695
...	...	...	...	...	...	...	...
2252	71.71.AS.005	7171	1.471357	131.253900	7	5	10

### 4.3. Data Mining-Based RFM-D Analysis

#### 4.3.1. School mapping and heatmap analysis

The school locations were geocoded on the Indonesian shapefiles. Figure 2 depicts the distribution of provider schools in 364 from 514 Indonesian districts. The dot represents the provider school, and the dot size is proportional to the number of students enrolled at the university from the school. The map shows that the provider schools are spread out across Indonesia but not evenly.



**Figure 2:** The distribution of the provider schools and students enrolled (Source: data processing result)

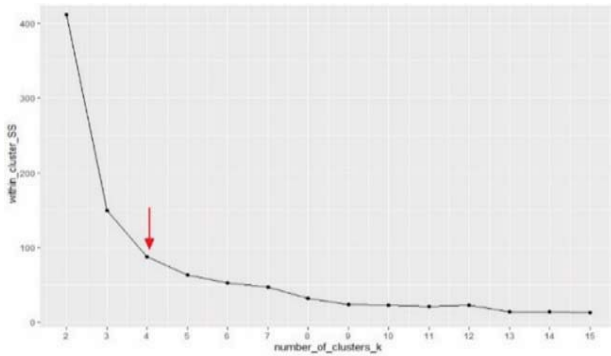
From the heatmap analysis as presented in Figure 3, it appears that most provider schools were concentrated on Java Island. They are in the districts surrounding Daerah Istimewa Yogyakarta Province (where this study's university is located), Jawa Tengah Province (the closest province to Daerah Istimewa Yogyakarta Province), and DKI Jakarta Province (the capital city of Indonesia). Outside of Java, they were distributed in several Sumatera, Bali, and Kalimantan districts, though less densely. Based on Figures 2 and 3, each district had a different number of provider schools and the number of enrolled students, implying that each district's contribution to the university varies. Thus, it is necessary to understand the potential of each district for decision-making.



**Figure 3:** The provider schools' heatmap (Source: data processing result)

#### 4.3.2. The district's potential determination

According to the Elbow method for determining the optimum number of clusters, as shown in Figure 4, before  $k=4$ , there was a high decrease in the sum of squared errors (SSE), but after  $k=4$ , the SSE decreased slowly. So, we chose  $k=4$  as the optimum number of clusters. As a result, four categories of the district potential were yielded after applying K-means for grouping districts. Table 3 displays the statistics of the four categories of the district potential, including the average number of provider schools, the average number of students enrolled at the university from each district in the cluster, and the number of districts contained in each cluster.



**Figure 4:** The Elbow method to find the optimum number of districts' potential categories (Source: data processing result)

According to Table 3, the district potential profiles are described below. While the distribution of the districts based on their potential is displayed in Figure 5. The higher the district's potential, the more the provider schools and the enrolled students at the university from the district.

**D1: the lowest-potential districts.** This cluster contains 295 districts, and each district has an average of three provider schools and 13 students enrolled during the analysis period. Most of the university provider schools are in the lowest-potential districts.

**D2: moderately-potential districts.** This cluster consists of 56 districts. Most of these districts are located on Java Island (54%), Sumatra Island (18%), and Kalimantan Island (14%). Each district sends an average of 100 students and has about 14 provider schools.

**D3: high-potential districts.** There are 12 districts in this cluster, with an average of 34 provider schools and 360 enrolled students from each district. Three districts are on Sumatra Island, eight on Java Island, and one on Bali Island.

**D4: the highest-potential district.** There is only one district in this cluster, the nearest city to the university location. Fifty-eight provider schools from this city sent 3,226 students during the analysis period.

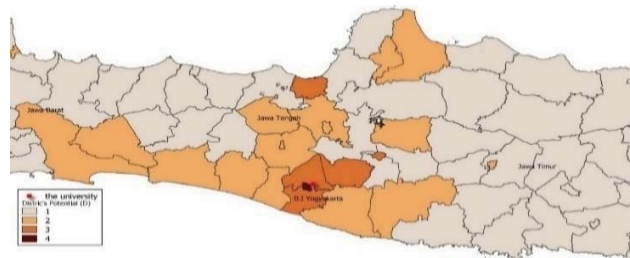
**Table 3:** Statistics of the four clusters of the district's potential

Cluster	The average number of schools	The average number of students	Total Districts
ClusterD1	3.29	13.03	295
ClusterD2	14.38	100.75	56
ClusterD3	34.75	360.08	12
ClusterD4	58	3226	1



**Figure 5:** The districts' potential distribution (Source: data processing result)

Although several districts outside Java Island have moderate or high potential, most are on Java Island, where the university is located. Figure 6 shows districts nearby the university have higher potential categories.



**Figure 6:** The districts' potential around the university location (Source: data processing result)

#### 4.3.3. Create the RFM-D dataset

Table 4 shows examples of provider schools' RFM-D values in the dataset. R and F values in the 1 to 8 range, M in the 1 to 695 range, and D in the 1 to 4 range. For the district's potential (D) variable, 1 represents the lowest-potential district, 2 for the moderately-potential district, 3 for the high-potential district, and 4 for the highest-potential district. The first school in Table 4 was located in the lowest-potential district, the second in the highest-potential district, and the last in the moderately-potential district.

**Table 4:** Examples of school's RFM-D values

No	School code	R	F	M	D
1	11.04.AN.001	2	1	1	1
2	34.71.AS.029	8	8	695	4
...	...	...	...	...	...
2252	71.71.AS.005	7	5	10	2

#### 4.3.4. Assessing the clustering tendency

According to the Hopkins test, the Hopkins statistic was 0.9766, greater than 0.75. Thus, the null hypothesis was rejected. It concluded that the dataset has a high clustering tendency, so it has meaningful clusters.

### 4.3.5. Provider Schools Clustering

After finding  $k=5$  as the optimal number of clusters using the Elbow method, as shown in Figure 7, the normalized data of R, F, and M variables and the categorical data of the district's potential (D) were clustered using the K-prototype algorithm reveals the provider schools segments. The visualization of the cluster features resulting from K-prototype clustering is illustrated in Figure 8. The upper left corner image shows a boxplot for the recency value of each cluster, and it appears that the median recency value is highest in Cluster 1 and Cluster 2. The upper right corner and lower-left corner graphics show boxplots for frequency and monetary values, with Cluster 1 having the highest median value. At the same time, the bottom right image shows the proportion of provider schools based on each district's potential in the five clusters. For example, Cluster 1 contains 40% of provider schools of high-potential districts ( $D=3$ ) and 60% of the highest-potential districts ( $D=4$ ).

The average R, F, M real and normalized values and the modus of the district potential were obtained by examining the cluster members. Table 5 summarizes each cluster's characteristics based on the RFM-D model and the number of provider schools and enrolled students at the university from the districts of each cluster.

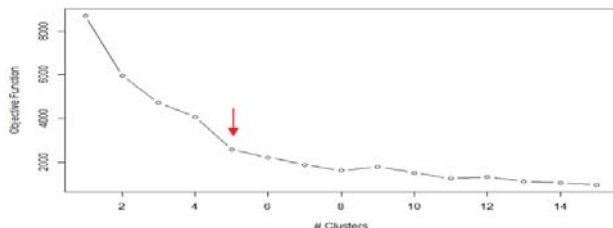


Figure 7: The Elbow method to find the optimum number of provider schools groups (Source: data processing result)

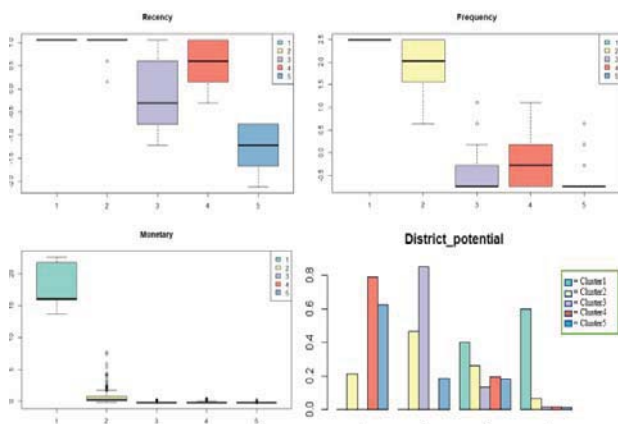


Figure 8: The characteristics of the K-prototype results clusters (Source: data processing result)

According to Table 5, Cluster 1 and Cluster 2 have R, F, and M values higher than the average, as indicated by their positive normalized values. It implies Cluster 1 and Cluster 2 have a higher contribution to the university. On the other hand, Clusters 3 and Cluster 5 have R, F, and M values lower than the average, indicating a minimal contribution to the university. Whereas Cluster 4 has a positive normalized recency value, the normalized frequency and monetary values are negative, showing that these schools are new provider schools. Thus, their contribution to the university is still low.

### 4.3.6. Clusters profiling

The profiles of the five provider schools clusters are described as follows:

**Cluster 1: the highest-value provider schools.** The schools in this cluster regularly sent their students every year ( $R=8$ ,  $F=8$ ). They sent the most students ( $M=554.60$ ). This cluster contains only five (0.22%) provider schools but sent 2,773 (16.28%) students. Three schools are from the highest-potential district, and two are from the high-potential districts.

**Cluster 2: high-value provider schools.** This cluster includes 396 (17.58%) provider schools from all potential district categories, mostly moderately-potential districts. This cluster of schools contributes the most significant number of students (59.49%). Until recently ( $R=7.82$ ), the schools in this cluster very often sent students ( $F=6.61$ ) in a relatively large number ( $M=25.59$ ).

**Cluster 3: low-value provider schools.** This cluster consists of 617 (27.40%) provider schools from the lowest-potential, moderately-potential, and high-potential districts categories, with the majority coming from moderately-potential districts. Since these schools only sent students on rare occasions ( $F=1.68$ ), their M value is small (2.04), so the provider schools in this cluster contribute only 7.40% of students. Their alumni were not enrolled at the university in the last few years ( $R=5.57$ ).

**Cluster 4: moderately-value provider schools.** This cluster contains the most schools (31.71%), but the frequency and monetary values are small ( $F=2.21$ ,  $M=3.04$ ), so the contribution of the schools to the university in sending students is not high (12.72%). The recency value remains quite high ( $R=6.79$ ), which means they are still engaged with the university. Most of the schools are from the lowest-potential districts.

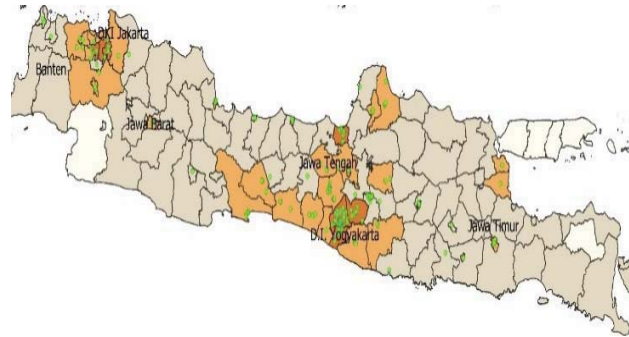
**Cluster 5: the lowest-value provider schools.** The majority of schools in this cluster are from the lowest-potential districts. These schools are one-time providers ( $F=1.17$ ) and only sent one or two students ( $M=1.34$ ) during the analysis period. They have the smallest recency value ( $R=2.59$ ), showing that the alumni have not enrolled at the university for a long time. Despite having many schools (23.09%), this cluster contributes the least students (4.10%).

**Table 5:** The RFM-D clusters' characteristics

Cluster	R		F		M		D (modus)	School		Student	
	real	scaled	real	scaled	real	scaled		amount	%	amount	%
Cluster 1	8	1.06	8	2.49	554.60	18.06	4	5	0.22	2,773	16.28
Cluster 2	7.82	0.98	6.61	1.85	25.59	0.60	2	396	17.58	10,134	59.49
Cluster 3	5.57	-0.05	1.68	-0.43	2.04	-0.18	2	617	27.40	1261	7.40
Cluster 4	6.79	0.51	2.21	-0.19	3.04	-0.15	1	714	31.71	2167	12.72
Cluster 5	2.59	-1.40	1.17	-0.67	1.34	-0.21	1	520	23.09	699	4.10

According to the clusters profiles, this study concluded that Cluster 1 and Cluster 2 contain valuable provider schools to the university and should be designated as the target segment to establish cooperation with the provider schools in marketing activity and student recruitment. Based on the historical data, the target segment contains 401 (17.80%) provider schools with 12,907 (75.77%) students enrolled. On average, a school sent 32 students six to seven times (F=6.63) until the last year of analysis (R=7.82). These schools were across 29 of Indonesia's 34 provinces and 130 of its 514 districts. Figure 9 depicts the distribution of the targeted provider schools across different potential areas. There were 29 targeted schools in the highest-potential districts, 105 schools in high-potential districts, 184 in moderately-potential districts, and 83 in the lowest-potential districts.

schools in the higher potential areas and design different strategies based on the characteristics of the districts.



**Figure 10:** The distribution of the targeted provider schools on Java Island (Source: data processing result)



**Figure 9:** The distribution of the targeted provider schools (Source: data processing result)

Figure 10 shows the distribution of the targeted provider schools on Java Island. Most were in the higher potential districts concentrated near DKI Jakarta, Central Java, and Daerah Istimewa Yogyakarta Provinces. The knowledge gained from the analysis of target schools and their locations can help the university decision-makers to develop marketing strategies that will benefit the organization. Knowing the targeted schools' distribution can provide decision-makers with insight to help them determine promotion locations, allowing useful information about the university brand to be delivered to prospective students and their teachers and conducting on-site recruitment. The management might focus its resources on the provider

#### 4.4. Results Comparison

For evaluation, this study compared the results of the RFM-D model with the CLV-based RFM model. The 401 top-ranked schools from the CLV-based RFM model were chosen as the target schools and compared to the RFM-D results. Both models' average R, F, and M values are similar. As shown in Table 6, CLV-based RFM produces schools with slightly higher recency and monetary value, whereas RFM-D produces schools with a slightly higher frequency. However, the RFM-D model, on the other hand, has the advantage of providing the district's potential information.

Although the CLV-based RFM model does not include the district's potential as a segmentation variable, for comparison purposes, the district's potential variable was analyzed further. Both models suggest the same 370 schools (92.27%). Whereas for the different 31 schools suggested, the RFM-D recommends schools in higher potential districts than CLV-based RFM, as shown in Table 6.

**Table 6:** The comparison of the targeted schools' characteristics

Model	R	F	M	Distribution of schools in district categories			
				D=1	D=2	D=3	D=4
CLV-based RFM	7.89	6.62	32.35	108	159	106	28
RFM-D	7.82	6.63	32.19	83	184	105	29

The targeted schools generated by the CLV-based RFM model are distributed across 139 districts in 30 provinces, whereas the RFM-D model generates target schools in 130 districts from 29 provinces. Figure 11 shows the comparison of the number of targeted schools by province. The figure shows that the RFM-D model targets more provider schools in Daerah Istimewa Yogyakarta, Central Java, North Sumatra, and DKI Jakarta, which are hotspots based on the results of the heatmap analysis shown in Figure 3.

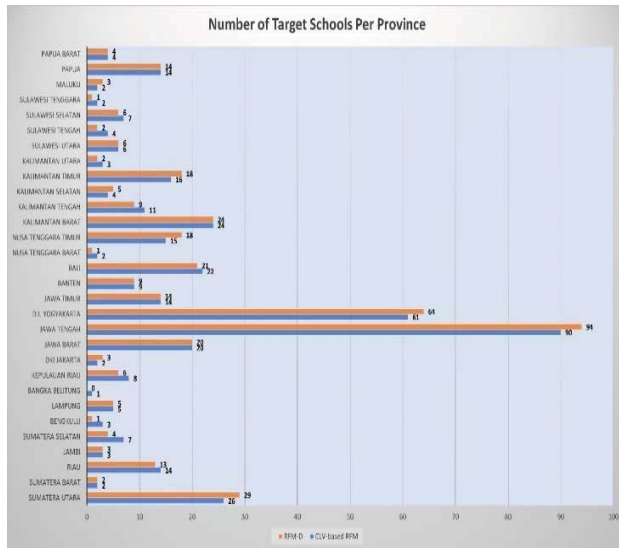


Figure 11: The comparison of the number of targeted schools in each province (Source: data processing result)

Additionally, the performance of the RFM-D model was compared to that of the CLV-based RFM model. This study performed 30 times 10-fold cross-validation and recorded macro values of accuracy, precision, recall, and F1-score. Then, an independent t-test was used to see if the RFM-D model outperformed the CLV-based RFM in terms of accuracy, precision, recall, and F1-score. Table 7 shows the RFM-D model's accuracy, precision, recall, and F1-score were higher than the CLV-based RFM models with a p-value of 0.00, less than a significance value of 0.05, implying that the RFM-D model outperforms the CLV-based RFM models.

Table 7: The comparison of the models' performance

Model	Accuracy		Precision		Recall		F1-score	
	Mean	St. dev.	Mean	St. dev.	Mean	St. dev.	Mean	St. dev.
CLV-based RFM	0.9877	0.0010	0.9758	0.0023	0.9829	0.0014	0.9793	0.0017
RFM-D	0.9995	0.0003	0.9985	0.0007	0.9997	0.0002	0.9991	0.0004

### 5. Conclusions and Future Works

This study adopted the RFM model for provider schools segmentation and extended it with spatial analysis. The spatial analysis of the provider school districts was conducted by geocoding the provider schools and employing heatmap analysis. The district potential category is added to enhance the RFM model. The number of provider schools and students enrolled at the university determined the district's potential. This study used enrollment data from a university in Indonesia for empirical study. It successfully identified five clusters of provider schools using the proposed RFM-D model and K-prototype clustering algorithm. The two clusters with the best value for the university are designated as the target segment. Compared to the CLV-based RFM model, the RFM-D model recommends schools from districts with higher potential categories. The model can help the university management reveal the characteristics of the provider schools, determine the target market, and provide the decision-makers with insight into the district's potential of the provider school location, enabling them to allocate promotional resources, distribute the university information, and conduct student recruitment effectively. In addition, by using the RFM-D model, better model performance is obtained, which is indicated by higher accuracy, precision, recall, and F1-score values. The limitation of this paper has not yet considered the academic performance of the enrolled students. Also, this research focuses on the cumulative number of students enrolled during the analysis period. This study can be developed for future work by analyzing the enrollment patterns that changed over time.

### References

Abbasimehr, H., & Shabani, M. (2019). A new methodology for customer behavior analysis using time series clustering: A case study on a bank's customers. *Kybernetes*, 50(2), 221–242. doi: 10.1108/K-09-2018-0506

Babaiyan, V., & Sarfarazi, S. A. (2019). Analyzing customers of South Khorasan telecommunication company with expansion of RFM to LRFM model. *Journal of AI and Data Mining*, 7(2), 331–340. doi: org/10.22044/JADM.2018.6035.1715

Beheshtian-Ardakani, A., Fathian, M., & Gholamian, M. (2018). A novel model for product bundling and direct marketing in e-commerce based on market segmentation. *Decision Science Letters*, 7(1), 39–54. doi: org/10.5267/j.dsl.2017.4.005

Bunnak, P., Thammaboosadee, S., & Kiattisin, S. (2015). Applying data mining techniques and extended RFM model in customer loyalty measurement. *Journal of Advances in Information Technology*, 6(4), 238–242. doi: 10.12720/jait.6.4.238-242

Christy, A. J., Umamakeswari, A., Priyatharsini, L., & Neyaa, A. (2021). RFM ranking – an effective approach to customer segmentation. *Journal of King Saud University - Computer and*

- Information Sciences*, 33(10), 1251-1257. doi: 10.1016/j.jksuci.2018.09.004
- Davari, M., Noursalehi, P., & Keramati, A. (2019). Data mining approach to professional education market segmentation: a case study. *Journal of Marketing for Higher Education*, 29(1), 45-66. doi: 10.1080/08841241.2018.1545724
- Dolnicar, S., Grün, B., & Leisch, F. (2018). *Market segmentation analysis: understanding it, doing it, and making it useful*. Singapore: Springer Nature Singapore Pte Ltd. doi: 10.1007/978-981-10-8818-6\_1
- Firdaus, U., & Utama, D. N. (2021). Development of bank's customer segmentation model based on RFM+B approach. *ICIC Express Letters, Part B: Applications*, 12(1), 17-26. doi: 10.24507/icicelb.12.01.17
- Hemsley-brown, J. (2017). Higher Education Market Segmentation. In *Jung Cheol, Shin, Pedro, Teixeira, (Eds) Encyclopaedia of International Higher Education Systems and Institutions*. Dordrecht: Springer Netherlands. doi: 10.1007/978-94-017-9553-1
- Hidayat, S., Rismayati, R., Tajuddin, M., & Merawati, N. L. P. (2020). Segmentation of university customers loyalty based on RFM analysis using Fuzzy C-means clustering. *Jurnal Teknologi Dan Sistem Komputer*, 8(2), 133-139. doi: 10.14710/jtsiskom.8.2.2020.133-139
- Hosseini, Z. Z., & Mohammadzadeh, M. (2016). Knowledge discovery from patients' behavior via clustering-classification algorithms based on weighted eRFM and CLV model: An empirical study in public health care services. *Iranian Journal of Pharmaceutical Research*, 15(1), 355-367.
- Huang, Z. (1998). Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery*, 2(3), 283-304.
- Hwang, S., & Lee, Y. (2021). Identifying customer priority for new products in target marketing: using RFM model and TextRank. *Innovative Marketing*, 17(2), 125-136. doi: 10.21511/im.17(2).2021.12
- Kamthania, D., Pahwa, A., & Madhavan, S. S. (2018). Market segmentation analysis and visualization using K-mode clustering algorithm for E-commerce business. *Journal of Computing and Information Technology*, 26(1), 57-68. doi: g/10.20532/cit.2018.1003863
- Kit, T. C., Firdaus, N., & Azmi, M. (2021). Customer profiling for Malaysia online retail industry using K-Means clustering and RM model. *International Journal of Advanced Computer Science and Applications*, 12(1), 106-113. doi: 10.14569/IJACSA.2021.0120114
- Lawson, R. G., & Jurs, P. C. (1990). New index for clustering tendency and its application to chemical problems. *Journal of Chemical Information and Computer Sciences*, 30(1), 36-41. doi: 10.1021/ci00065a010
- Libório, M. P., Bernardes, P., Ekel, P. I., Ramalho, F. D., & Santos, A. C. G. dos. (2020). Geomarketing and the locational problem question in the marketing studies. *Brazilian Journal of Marketing*, 19(2), 448-469. doi: 10.5585/remark.v19i2.17777
- Moghaddam, S. Q., Abdolvand, N., & Harandi, S. R. (2017). A RFMV model and customer segmentation based on variety of products. *Journal of Information Systems & Telecommunication*, 5(3), 155-161.
- Peker, S., Kocyigit, A., & Eren, P. E. (2017). LRFMP model for customer segmentation in the grocery retail industry: a case study. *Marketing Intelligence and Planning*, 35(4), 544-559. doi: 10.1108/MIP-11-2016-0210
- Rahadian, Y. R., & Syairudin, B. (2020). Segmentation analysis of students in X course with RFM model and clustering. *Jurnal Sosial Humaniora (JSH), special ed*, 59-79. doi: 10.12962/j24433527.v0i1.6776
- Rezaeinia, S. M., & Rahmani, R. (2016). Recommender system based on customer segmentation (RSCS). *Kybernetes*, 45(6), 946-961. doi: 10.1108/K-07-2014-0130
- Roshan, H., & Afsharinezhad, M. (2017). The new approach in market segmentation by using RFM model. *Journal of Applied Research on Industrial Engineering*, 4(4), 259-267. doi: 10.22105/jarie.2017.91297.1011
- Simkin, L., & Dibb, S. (1998). Prioritising target markets. *Marketing Intelligence & Planning*, 16(7), 407-417.
- Sukoroto, Haryono, S., & Kharisma, B. (2020). Target market selection using MCDM Approach: A study of rolling stock manufacturer. *Journal of Distribution Science*, 18(7), 63-72. doi: 10.15722/jds.18.7.202007.63
- Sulastri, S., Usman, L., & Syafitri, U. D. (2021). K-prototypes algorithm for clustering schools based on the student admission data in IPB University. *Indonesian Journal of Statistics and Its Applications*, 5(2), 228-242. doi: 10.29244/ijsa.v5i2p228-242
- Tapp, A., Hicks, K., & Stone, M. (2004). Direct and database marketing and customer relationship management in recruiting students for higher education. *International Journal of Nonprofit and Voluntary Sector Marketing*, 9(4), 335-345. doi: 10.1002/nvsm.258
- Wei, J., Lin, S., & Wu, H. (2010). A review of the application of RFM model. *African Journal of Business Management*, 4(19), 4199-4206.
- Wu, J., Shi, L., Lin, W., Tsai, S., Li, Y., Yang, L., & Xu, G. (2020). An empirical study on customer segmentation by purchase behaviors using a RFM model and K-means algorithm. *Mathematical Problems in Engineering*, 2020, 1-7. doi: 10.1155/2020/8884227