Customer Profiling for Malaysia Online Retail Industry using K-Means Clustering and RM Model

Tan Chun Kit[1], Nurulhuda Firdaus Mohd Azmi[2]
Advanced Informatics Department
Razak Faculty of Technology and Informatics
Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia

***Abstract*—Malaysia's online retail industry is growing sophisticated for the past years and is not expected to stop growing in the following years. Meanwhile, customers are becoming smarter about buying. Online Retailers have to identify and understand their customer needs to provide appropriate services/products to the demanding customer and attracting new customers. Customer profiling is a method that helps retailers to understand their customers. This study examines the usefulness of the LRFMP model (Length, Recency, Frequency, Monetary, and Periodicity), the models that comprised part of its variables, and its predecessor RFM model using the Silhouette Index test. Furthermore, an automated Elbow Method was employed and its usefulness was compared against the conventional visual analytics. As result, the RM model was selected as the finest model in performing K-Means Clustering in the given context. Despite the unusefulness of the LRFMP model in K-Means Clustering, some of its variables remained useful in the customer profiling process by providing extra information on cluster characteristics. Moreover, the effect of sample size on cluster validity was investigated. Lastly, the limitations and future research recommendations are discussed alongside the discussion to bridge for future works.**

Keywords—Customer Profiling; LRFMP; RFM; Data Mining; K-Means Clustering

I. Introduction

The online retail industry in Malaysia is in its growth trajectory in the past decade and is likely to enjoy strong growth over the coming years [1-2]. While internet freeing up the geographical limitations, online retailer now has to compete with each other directly regardless of their location. To the local small and medium online retailers, the battlefield quickly levels to the international level when online retail giant with economies of scale such as Amazon and Taobao provides a wide range of product selections and reduced shipping fare. Furthermore, despite the benefits of being able to sell online, compared to traditional physical outlets, the lack of seller-buyer interaction leaving the seller confused about customer preference and trends. Moreover, customers are becoming smarter and selective with their spending given the wide range of selection [3].

Under such circumstances, one retailer could never fulfill every customer's needs and provide satisfiable service to every new customer. Hence it is important to identify the most profitable customer in the long-run, then provide tailored and customized service to these groups of customers to reduce the cost while maximizing profitability. Customer profiling is a potential method of achieving it and resolve the described issues. Customer profiling is technique retailers or service providers used in analyzing consumer characteristics and needs. It increases retailers' understanding of consumers and provides a foundation to retailers in making an informed decision in many business aspects such as product selection or marketing tactics.

To conduct a customer profiling, two aspect has to be considered: The theoretical guidance where a theory that guided which variables to be used to create customer profiles and the technical calculation aspect focuses on methods and formulas to calculate the scoring for each variable and aggregate the result in creating customer profiles.

Theoretical-wise, apart from some grounded theory development that proposed unique domain-specific variables to use in customer profiling [4], RFM (Recency, Frequency, Monetary) analysis is used widely in performing customer profiling in many domains [5-6]. In the past decade, many modifications of the original RFM model, either to create a domain-specific RFM variation or to optimize the RFM model in general have been done. Some examples are RFQ (Quality) [7], and LRFMP (Length and Periodicity) [3].

Technical-wise, conventionally, analyst convert the entire RFM model's raw data into Likert-scale as it not only eases understanding but also simplify calculation [8]. Researchers such as Peker, Kocygit, and Eren [3] and Palaniappan, Mustapha, and Mohd Foozy [9] had used data mining techniques while conducting customer profiling due to its powerful capability and the benefit of no need to skimming down the data.

The research objectives (RO) are as follows:

RO1: To identify the theoretical model and techniques that could be used in customer profiling within Malaysia's online retail industry.

RO2: To develop a customer profiling model based on RFM's variation model and K-Means clustering within Malaysia's online retail industry.

RO3: To compare and evaluate the developed customer profiling models to identify an optimal model for the given dataset.

The purpose of this paper is to examine the usefulness of the LRFMP model, and its predecessor RFM model using the

Silhouette Index (SI) test for customer profiling. In addition, the work explained in this paper examined the possibility of an automated customer profiling process where the K-value decision and optimal variables to be used could be decided without human intervention such as the conventional EM's visual analysis.

This paper is presented by the following sections: Section I will discuss the general context of the work explained in the paper. Next, in Section II, the literature review that presents the theoretical and conceptual of the study is explained. In Section III, the methodology of the experimentation and analysis is described. The experimental result is explained in Section IV and finally, Section V conclude the work discuss in the paper.

II. Literature Review

To identify the current method and technique of conducting customer profiling, a literature review was conducted and is summarized into the following subsections:

A. *Theoretical Model in Customer Profiling*

While some researcher employs grounded theory and proposed domain-specific variables in customer profiling [4] [10–13], most researchers favored the use of the RFM model. The popular usage of RFM or its various models covers many domains such as banking [14], Hotel industry [15], Small and Medium Enterprises (SME) [16], grocery retail industry [3], and online retail industry [5] [11] [17].

Due to the popular uses of the RFM model, many pieces of research had proposed a variated RFM model with additional or removal of certain variables, which claimed to improve its relatedness to a particular industry and better performance in customer profiling. For instance, Li [18] proposed the uses of the FM model with recency removed, the model remained effective in performing customer profiling in the retail industry. Li argues the uses of the only 2 important variables could effectively create a matrix of 4 distinctive groups of customer profiles. On the other hand, Liu, Zhao, and Li [7] argues the replacement of Quality over the original Monetary variable (RFQ model) is effective and more relevant in the mobile apps domain as most mobile apps are freemium oriented therefore no monetary aspect was involved. Moreover, Wei, Lin, and Weng [19] suggested the additional variable of Length (LRFM model) in the dental industry to improve the quality of customer profiling. They suggested the length in time a customer has visited since the first visitation indicates customer loyalty and is an important indicator in customer profiling. Recently, Peker, Kocyigit, and Eren [3] proposed the Periodicity variable on top of Wei, Lin, and Weng [19]'s model LRFM, creating a LRFMP model and was used in the grocery retail industry in Turkey.

B. *Technical aspect in Customer Profiling*

The technical aspect of customer profiling refers to the method and calculation used to aggregate data into useful insight. Conventionally, analyst tends to convert the RFM model's raw data into Likert-scale following by assigning customer into a certain group (for instance group that has high in R, low in F, and high in M, etc.) [8]. However, this is not the case for the past decade. Recent research shows a trend of adopting a data mining technique in aggregating and clustering the consumer group [3] [19]. Given the nature of customer profiling is to cluster consumers into several comprehensible groups, the clustering technique from the data mining domain seems to be the best method of aggregating the customer data to provide useful insight.

Hung, Yen and Wang [10] employs a Decision Tree and Neural Network in conducting customer profiling in Taiwan's Telecom industry. Similarly, Sankar [11] also employed the Neural Clustering technique in the USA's Online retail industry. Another example of data mining used in customer profiling has been mentioned by Hu and Yeh [20], which used constraint-based mining in the food and beverage industry. Apart from that, the majority of the research had adopted the K-Means clustering technique in the conjunction with RFM model or its variation [3] [5-6] [14-15] [17] [18] [21–23].

Among the researchers used K-Means Clustering in customer profiling, Christy et. al. [23] had attempted to compare the performance of K-Means clustering with other clustering techniques. They compare the effect of K-Means Clustering and Fuzzy C-Means using the RFM model. Interestingly, they also attempted to include the RM model in the comparison. The result suggested Fuzzy C-Means gains better Silhouette width performance at the cost of runtime while the RM model K-Means Clustering achieved the lowest runtime and highest Silhouette width, indicating a proper implementation of K-Means Clustering can be effective in both clustering quality and runtime. However, their work did not include the RM model's Fuzzy C-Means analysis.

Evaluation on K-Means clustering customer profiling is another aspect that need to be addressed. While classification-related methods could be benchmarked through the use of confusion matrix, the K-Means clustering evaluation is benchmarked through the manipulation of K to measuring the cluster distances. The cluster distance indicates unique cluster characteristics and therefore, unique customer profiling result. Studies regarding validation on cluster distance was performed in many studies [6][13-14][21][23]. For instance, Dong, Zhang, and Ye [13] evaluated the consistency of cluster distances among iterations to justify validity. Another literature suggested the use of Self Organizing Maps analysis in identifying the best number of K [15]. Besides, more researches had employs a cluster distances-related analysis in justifying cluster validity [6][14][21][23]. For instance, Maryani and Riana [6] measures the Euclidean distance among clusters to justify each cluster contains unique characteristics and is not overlapping with other clusters. Similarly, both Walters and Bekker [21] and Christy et. al. [23] used SI in measuring inter-cluster distances while the former used it to identify the best number of K but the latter used it in inter-model comparison (FM vs. RFM model). In addition, Aryuni, Madyatmadja, and Miranda [14] used both Average Within Cluster (AWC) and Davies-Bouldin Index (DBI) in measuring cluster distances. Lastly, instead of employing cluster distance-related formulas in validating clusters, research such as Chen, Sain, and Guo [5] evaluate the characteristics of each cluster to justify its uniqueness and meaningful clustering result. Among the distance-based analysis, SI is much more popular compared with other methods, mainly due to its better accuracy [24]. However, it is important to note that one of the disadvantages of using the SI is its complex and long computational runtime [24].

C. RFM Model for Customer Profiling

Theoretical wise, the uses of the RFM model is currently a popular method of conducting customer profiling. While some studies proposed the removal of some certain variables will not have a significant impact on cluster quality, some proposed number of new variables which were believed to be able to improve the quality of customer profiling. However, these researches are mostly containing little to no replica study. Furthermore, most of these studies did not compare the newly proposed model against its original RFM model, leaving its improvement in terms of quality beyond the original RFM model questionable.

Technical wise, K-Means clustering were used frequently among other data mining technique when it comes to customer profiling. While most of the studies did not tackle the evaluation of the clustering result, limited studies suggested the uses of techniques such as the SI, Davies-Bouldin Index (DBI) to validate the clustering results [14] [21]. Furthermore, it is possible to evaluate the quality and usefulness of the clustering result by directly evaluating the unique characteristics of each cluster, which is the combination of RFM variables in the given context, where unique clusters indicating successful clustering while similar cluster characteristics indicating low clustering quality.

To identify the optimal model among the vast proposed RFM variations in the online retail industry, RO2 and RO3 were coined. K-Means Clustering was implemented guided by the LRFMP model as proposed by Peker, Kocyigit, and Eren [3] due to its relatedness and the promising result beyond the original RFM model. Nevertheless, the LRFMP model contains all newly proposed variables as reviewed in the literature review except for Quality, which was created specifically for the mobile apps industry.

III. Methodology

This research is quantitatively oriented that employs a case study and experimental approach through data analysis. The data was provided by an anonymous online retailer located in Malaysia. The following descriptions record the methodology conducted in this study.

Step 1: Data Collection: Utilizing convenient sampling techniques and a list of data requirements as shown in Table I, a totally of 60 data acquisition approaches were done and one had accepted to participate in the study anonymously and the customer and company-related information have to be protected. Any data that could possibly reveals the company had been shielded by converting this information into unique ID prior to any data processing. The participating company is an online retailer performs sales solely on Facebook and Instagram and communication between salespersons and buyers was done through WeChat, WhatsApp, calling, or direct messenger within the platform. The data provider's company focuses on cosmetic product sales with minor branches on clothing with about four years of company history.

TABLE I. Data requirement

Mandatory	
Data	Details
Customer ID	An identifier for a particular 1 customer, some alternative includes IC, numbers, name or mailing address
Date	Date of Purchase OR date of the recorded order
Expenses (MYR)	Total expenses in 1 receipt
Optional	
Demographical Information	Mailing Address, Gender, Age, etc.

Step 2: Data Pre-processing: Data aggregation was done to pivot the raw data table as in Table I into the new LRFMP table using the following formulas:

Length:

$$L=lv-fv \tag{1}$$

Where *lv* represents the date of the last visit, and *fv* represents the date of the first visit, and Length (L) measures in days. This variable reveals the length of history a customer spent with the company, measuring in day.

Recency:

$$R=od-lv \tag{2}$$

Where *od* represents observation date, and *lv* represents average days of all visit dates and observation day. This variable reflects whether a particular customer remained active recently.

Frequency: Counted in times, the total visit of a particular customer.

Monetary: The sum of all spending, which is different from [3] uses of average spending. While not affecting the pattern and distribution, using sum could directly reflect a particular customer's contribution to the company total revenue.

Periodicity:

$$P=stdev(IVT_1 +IVT_2 +\dots IVT_n) \tag{3}$$

Where:

$$IVT_i=date\_diff(t_{i+1},t_i) \tag{4}$$

Where $I \geq 1$ and t_i refers to the date corresponding to the ith visit of a particular customer. This variable reflects the customer consistency in visiting the shop.

Then, both the standardized z-score and normalized value is calculated based on the output of the 5 formulas. While the standardized z-score is used to feed into the data mining pipeline, a normalized score is used in descriptive statistics to compare dispersion of each variable within the LRFMP model.

Step 3: Model Development and Evaluation Phase: Models are proposed based on the dispersion measurement. Using a normalized LRFMP score which all variables will have the same range between 0 and 1, each variable is compared using standard deviation and the one with the lowest dispersion is removed one by one. This results in a list of models starting from LRFMP, following by a model with 1 less variable, and so on, down to when only 1 variable has remained. RFM model is added exclusively for the baseline comparison purposes. Then, K-Means Clustering is implemented on each model using the Scikit-Learn package on Python 3.5 with the settings as shown in Table II. The results are evaluated through an evaluation matrix as shown in Table III.

As shown in Table III, Elbow Method (EM) serves as the most important rule of thumb in deciding the best number of K, however, its method tends to be too naïve as stated by Buitinck et. al. [25]. Hence, further validity mechanism has been set to triangulate the results of the EM. Therefore, SI was used to justify the EM's results. The EM's formula is stated in the below cited from [26] :

$$W_k = \sum_{r=1}^{k} \frac{1}{n_r} D_r \tag{5}$$

Where *k* represents the number of clusters, n_r represents the total number of points in the cluster *r* and D_r is the sum of distances between all points in a particular cluster.

$$D_r = \sum_{i=1}^{n_r-1} \sum_{j=i}^{n_r} ||d_i - d_j||_2 \tag{6}$$

The conventional EM requires human intervention in deciding the knee jerk through visual analysis of the graph. However, this can be sometimes confusing and inconsistent. Therefore, an automated method was suggested by Bertagnolli [27] and is adopted in this study. It involves calculating the closest distance of each point of scores to an imaginary straight line between the first and the last k value as shown in Fig. 1.

Furthermore, the SI was adopted and the formula is listed below as discussed by Perera [28]:

$$s(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))} \tag{7}$$

TABLE II. CLUSTERING SETTINGS

Setting	Detail
No. of Cluster	2 - 10
Method for initialization	K-Means++
Max iteration	300
Max centroid iteration	10
Random seed	0

TABLE III. EVALUATION MATRIX

Validation technique	Range	Usage	Usage
EM	0 – ∞	K-value Selection	Decisional method
SI	-1 – 1	K-value Validation & Inter-model Comparison	Supportive method

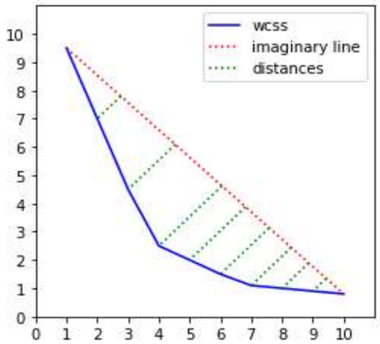


Fig. 1. Automated EM Illustration.

where *s(i)* refers to the SI test score, *b(i)* and *a(i)* refer to 2 different clusters. The score of the SI test can range between -1 and 1 where -1 indicates the clusters are overlapping and 1 indicates the boundaries and distances between each cluster are clear and distant [28]. While similar to the EM, the SI test is known to provide a better result with its complex calculation. Furthermore, unlike the EM, the purpose of utilizing the SI is its advantage of extracting a standardized score (-1 to +1), which is comparable among models, while the EM could only be used for intra-cluster comparison which is related to the selection of the best number of K.

Step 4: Customer Profiling: Based on the result of K-Means Clustering, customer profiling was conducted and descriptive statistics of each cluster were extracted. Lastly, a theme was given to each cluster and cluster characteristics were discussed.

IV. Finding and Result

A. Descriptive Statistics

The result of the descriptive statistics serves the purpose of data exploration and suggesting models for the following K-Means clustering analysis. Table IV records both the summary statistics and the trimmed version of it on the right. It shows that with the outliers removed, Length, Frequency, and Periodicity presents a mean score of 0, 1, and 0 respectively with little to no deviation.

To investigate this in detail, the count of identical scores was conducted and is recorded in Table V, showing that 87%, 87%, and 96% of the variable Length, Frequency, and Periodicity are the same value (0 for Length and Periodicity, 1 for Frequency). This hinted the 3 variables might not be useful in K-Means Clustering analysis due to the lack of heterogeneity.

Furthermore, the standard deviation of normalized scores was calculated and recorded in Table VI, and model proposal for K-Means clustering analysis is created based on the standard deviation scoring, such that: variable with lower standard deviation values is excluded 1 by 1 in the next model, starting from the complete LRFMP model.

The complete model proposal is recorded in Table VII. Apart from that, model 6 which contains the original RFM model was added.

TABLE IV. SUMMARY STATISTICS

Summary Statistics								**Summary Statistics (trimmed)**					
Var.	**Mean**	**Median**	**Range**	**SD**	**Variance**	**Skewness**	**Kurtosis**	**Mean**	**Range**	**SD**	**Variance**	**Skewness**	**Kurtosis**
L	8.06	0	0-340	33.19	1101.65	5.83	40.00	0	0-0	0	0	0	0
R	109.15	89	32-389	70.11	4915.80	1.75	3.20	105.38	32-389	68.51	4693.38	1.89	3.96
F	1.43	1	1-74	3.48	12.11	16.50	306.01	1	1-2	0.06	0	17.58	307.99
M	151.76	110	1-4793	272.79	74413.42	11.78	173.44	116.99	0-1135	100.53	10105.54	4.65	36.06
P	0.81	0	0-76	5.47	29.94	8.95	91.22	0	0-0	0	0	0	0

TABLE V. DATA HOMOGENEITY STATISTICS

	Count (= 0)	**Count (= 1)**	**Related Proportion**
Length	621	N/A	0.87
Frequency	N/A	619	0.87
Periodicity	680	N/A	0.96

TABLE VI. STANDARD DEVIATION. OF NORMALIZED VARIABLES

Variable	**S.D.**
nfrequency	0.003790
nperiodicity	0.013200
nlength	0.085361
nmonetary	0.253688
nrecency	0.255671

TABLE VII. PROPOSED MODELS

Count	**Model**
Model 1	LRFMP
Model 2	LRMP
Model 3	LRM
Model 4	RM
Model 5	R
Model 6	RFM

B. K-Means Clustering – Cluster Evaluation

To identify the optimal K value, the proposed automated EM was implemented and Fig. 3, 4, and 5 were constructed where the green line indicating the optimal k-value as recommended by the automated EM while the red line/shade indicating the potential K values based on conventional visual analytics. These models with their respective best K value based on the automated EM are then compared with other models to identify the optimal model for the given dataset and is recorded in Table VIII. The comparison as shown in Table VIII shows that the RM model provides the highest SI scores among all following by the RFM model. Therefore, the RM model is selected as the most relevant model among all and was used to conduct customer profiling.

TABLE VIII. INTER-MODEL COMPARISON

	LRFMP	LRMP	LRM	RM	R	RFM
K value	4	4	4	**3**	3	3
SI	0.60	0.61	0.61	**0.63**	0.59	0.62

C. The Effect of Sample Size toward Cluster Validity

It is expected that as the business grows, the sample size could increase dramatically. To investigate its effect towards the proposed cluster validity, a t-test was carried out based on the SI scoring with the following Hypothesis:

H_0: There is no significant difference between the full sample and half sample group in terms of the Silhouette Index scoring at a 95% confidence interval.

H_1: There is a significant difference between the full sample and half sample group in terms of the Silhouette Index scoring at a 95% confidence interval.

The data was then split into half and only half of it was used in the testing. There were two methods used in splitting the dataset which is the random selection among all samples and only selects the first half of the dataset based on time order. The result in Table IX records the result where both sampling method returns the scoring of p>0.05, indicates sample size could have a significant effect on the SI test scores. The result suggested SI test must be conducted periodically to ensure cluster validity as the business grow.

D. Customer Profiling based on RM Model

Customer profiling was conducted using the selected model RM due to its highest SI score among all models. The mean scores of the remaining variables: Length, Frequency, and Periodicity were recorded too for discussion purposes and is shown in Table X.

TABLE IX. HALF SAMPLE TESTS RESULT

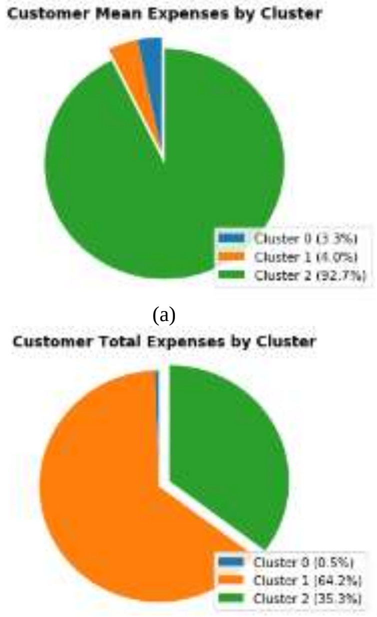
Half Sample Method	**p-value**
Random	0.548
Time-based	0.573

TABLE X. CLUSTER CHARACTERISTICS AND THEME

	Count	**L**	**R***	**F**	**M***	**P**	**M(Sum)**	**Customer Profiling Tag**
Cluster 0	581	4.26	81.73	1.17	131.35	0.27	76314.00	The One-Time buyer
Cluster 1	127	21.21	233.08	1.47	160.57	3.19	20393.00	The Loosen one
Cluster 2	3	187.33	175.00	50.67	3730.33	4.05	11191.00	The Loyal Buyer

Note: * are variables used in K-Means Clustering.

Based on the result of the K-Means Clustering, 3 unique clusters (cluster 0, 1, and 2) were identified, and themes were given to each cluster: The one-time buyer, the loosen one, and the loyal buyer after analyzing the characteristics of each cluster. It can be seen that Cluster 0 and 1 are the typical one-time buyer, having a mean Frequency around 1 (Cluster 0 = 1.17 and Cluster 1 = 1.47). The obvious difference that distinguishes Cluster 1 from Cluster 0 is the high level of recency, showing that Cluster 1 may be the customers that are losing interest in the shop in the past 1 year. Cluster 2 is a very unique, yet important cluster to the shop, consisting of only 3 customers, but contributed a significant amount of income toward the shop. Fig. 2a and Fig. 2b compares the cluster's expenses on average and total levels. On average as in Fig. 2a, customers in Cluster 2 spent extremely more amount of money when compares with Cluster 0 and 1. On total level as in Fig. 2b, Cluster 2 contributed 35.3% sales toward the shop's total income. Furthermore, customers from Cluster 2 are themed as the loyal buyers due to its high Frequency and Length, indicating frequent visitations in a long period.



(a)

(b)

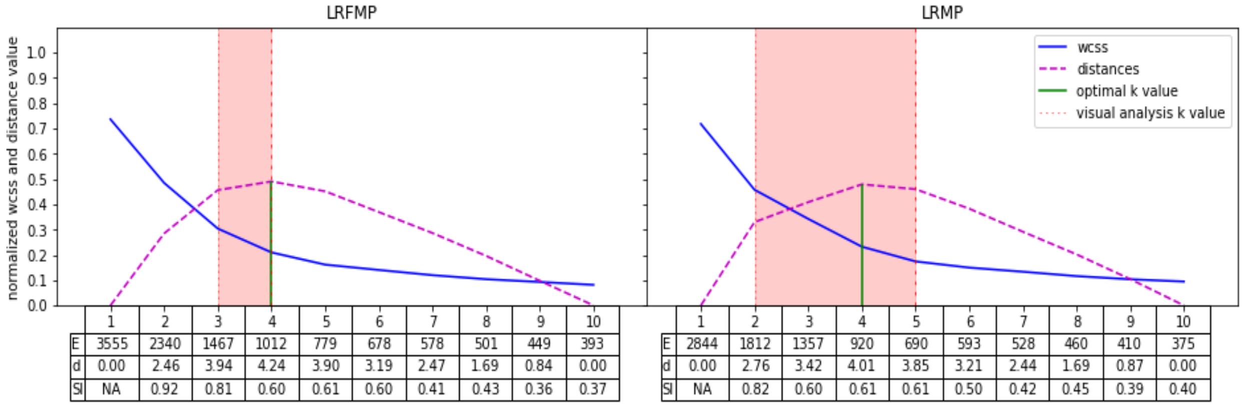
Fig. 2. Mean Expenses Pie Chart, (b) Total Expenses Pie Chart.

V. Discussions

A. Automated Elbow Method vs Silhouette Index

Through the process of K value selection of all proposed models, the employed automated EM techniques seem to select a K-value with only acceptable SI scorings instead of the one with the highest SI scorings. This could be due to the fact that the EM seeks for maximum possible K-value by selecting the maximum number of K before the diminishing mean squared distance flattens as the K-value increases. SI, on the other hand, did not perceive such biases, each SI score for every K value was calculated independently, resulting in extremely high runtime when compare to the automated EM but much objective cluster evaluation [29]. Despite being not able to select the K-value with the highest SI scores, it is important to note that the SI scorings generally served as the validation and supportive technique in K-Means clustering to triangulate EM's decision. On a larger sample size, the benefit of using the EM quickly outrun the SI due to the runtime advantage [29]. However, due to the identified effect of sample size toward the SI scores, the customer profiling based on the automated EM approach has to be validate periodically using the SI tests to ensures cluster validity.

On the other hand, when evaluating the result of the automated EM through visual analysis in Fig. 3, Fig. 4 and Fig. 5, the automated K-value selection seems to work properly on the obvious model such as RM and LRM. On the model with more confusing curves such as LRFMP and LRMP, the automated EM selected the K-value somewhat between the possible K value based on visual analysis, indicating its consistency beyond conventional visual analysis due to the well-defined criteria for the automation. Lastly, the SI scorings of all selected K-value are around 0.6, with model R being the lowest at 0.59, indicating the acceptable quality of the proposed automated EM. Combining the result with the sample size testing, the automated EM is an acceptable method of K-Means selection in the long run with the aid of SI validity test periodically, instead of purely relying on the SI test, which is known to be computationally intensive.



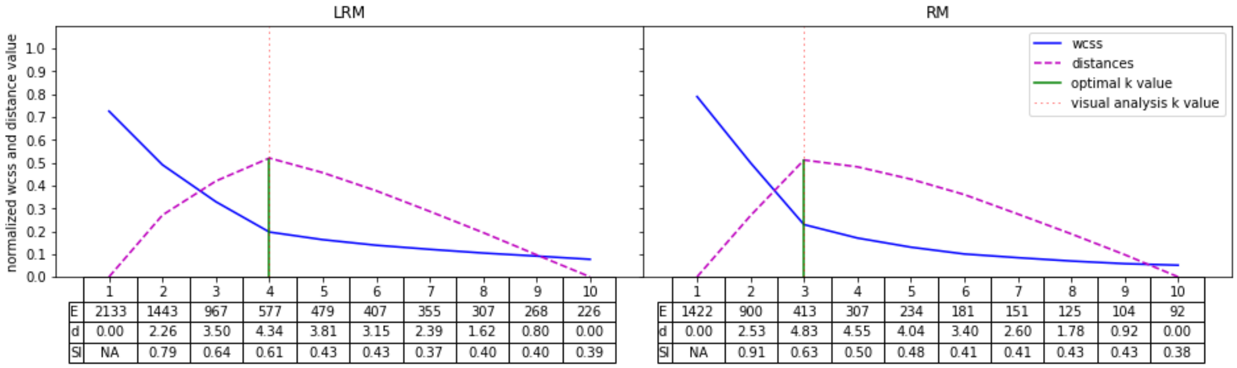
LRFMP

	1	2	3	4	5	6	7	8	9	10
E	3555	2340	1467	1012	779	678	578	501	449	393
d	0.00	2.46	3.94	4.24	3.90	3.19	2.47	1.69	0.84	0.00
SI	NA	0.92	0.81	0.60	0.61	0.60	0.41	0.43	0.36	0.37

LRMP

	1	2	3	4	5	6	7	8	9	10
E	2844	1812	1357	920	690	593	528	460	410	375
d	0.00	2.76	3.42	4.01	3.85	3.21	2.44	1.69	0.87	0.00
SI	NA	0.82	0.60	0.61	0.61	0.50	0.42	0.45	0.39	0.40

Fig. 3. K-Means Clustering Result for Model LRFMP and LRMP.



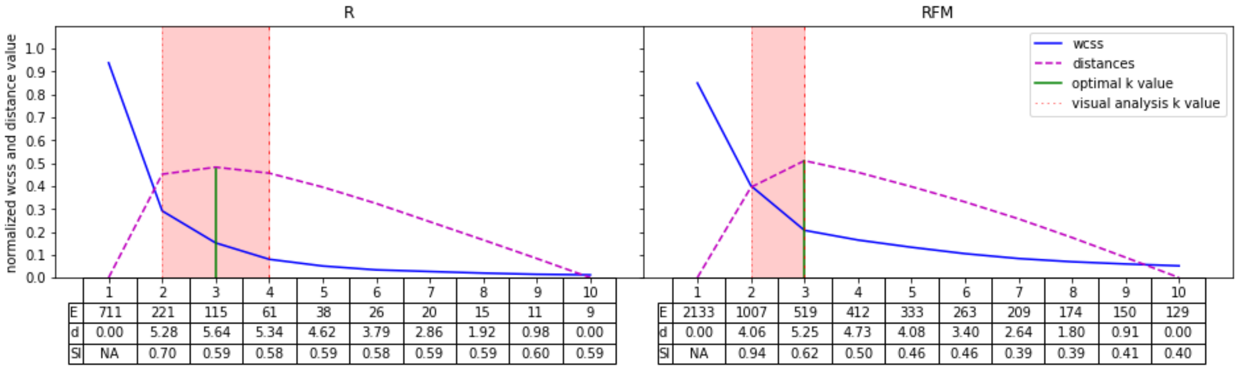
LRM

	1	2	3	4	5	6	7	8	9	10
E	2133	1443	967	577	479	407	355	307	268	226
d	0.00	2.26	3.50	4.34	3.81	3.15	2.39	1.62	0.80	0.00
SI	NA	0.79	0.64	0.61	0.43	0.43	0.37	0.40	0.40	0.39

RM

	1	2	3	4	5	6	7	8	9	10
E	1422	900	413	307	234	181	151	125	104	92
d	0.00	2.53	4.83	4.55	4.04	3.40	2.60	1.78	0.92	0.00
SI	NA	0.91	0.63	0.50	0.48	0.41	0.41	0.43	0.43	0.38

Fig. 4. K-Means Clustering Result for Model LRM and RM.



R

	1	2	3	4	5	6	7	8	9	10
E	711	221	115	61	38	26	20	15	11	9
d	0.00	5.28	5.64	5.34	4.62	3.79	2.86	1.92	0.98	0.00
SI	NA	0.70	0.59	0.58	0.59	0.58	0.59	0.59	0.60	0.59

RFM

	1	2	3	4	5	6	7	8	9	10
E	2133	1007	519	412	333	263	209	174	150	129
d	0.00	4.06	5.25	4.73	4.08	3.40	2.64	1.80	0.91	0.00
SI	NA	0.94	0.62	0.50	0.46	0.46	0.39	0.39	0.41	0.40

Fig. 5. K-Means Clustering Result for Model R and RFM.

B. *Customer Profiling*

In the process of conducting the analysis, due to the lack of dispersion on variable Length, Frequency, and Periodicity, it was expected these variables will not provide useful information in customer profiling due to the almost universal scorings as shown in Table V. This assumption was further supported by the fact that the SI scorings of the model including these variables were lower when comparing to RM model. However, during the customer profiling process, variables such as Length and Frequency were able to provide supportive numbers in justifying the differences between some clusters and providing insight regarding cluster characteristics.

C. *LRFMP Model in Malaysia's Online Retail Industry*

Based on the given dataset and testing, the finding indicates variable Periodicity may not be the best variable to be used in customer profiling in the online retail industry. This could be due to fact that most buyers are one-time buyers, resulting in identical scorings among customers. This is especially problematic as most clustering technique such as K-Means clustering requires dispersion to work with when creating clusters. Due to the same issue of one-time buyers, both Length and Frequency which utilizes date in calculation also results in high universal scoring, and only provided a supportive reference in the customer profiling process. Variables Recency and Monetary are the two variables that remained useful in both K-Means clustering and the customer profiling process. When comparing the SI scoring for both the LRFMP model and its original predecessor RFM model, the latter scores slightly higher than the LRFMP model, indicating the LRFMP model may not provide extra information beyond the RFM model in the given dataset and assumingly, the online retail industry domain, but extensive research has to be done to test this assumption.

VI. Conclusion

In the study, we employed K-Means Clustering and LRFMP model in conducting customer profiling after the literature review. The usefulness of each variable in K-Means clustering was evaluated through descriptive statistics, following by some test model proposal. These models were then compared using the SI scoring, resulting in the RM model being the finest model in the given dataset.

The study demonstrated the automated customer profiling process where the K-value decision and optimal variables to be used could be decided without human intervention such as the conventional EM's visual analysis. Moreover, the outlined process is able to compare and evaluate if the variables are useful in any new dataset, and automatically select the model that fits the dataset. Furthermore, the process is expandable to cover more variables as variables are proposed in the future.

Acknowledgment

The authors would like to thank Ministry of Education, Malaysia and Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia for the support of the resources. This work is currently funded with vot. R.K130000.7856.5F229.

References

[1] S. Kemp and S. Moey, "Digital 2019 Spotlight: Ecommerce In Malaysia," Datareportal, 2019. https://datareportal.com/reports/digital-2019-ecommerce-in-malaysia.

[2] Department of Statistics Malaysia, “Information and Communication Technology Satellite Account 2017,” 2018.

[3] S. Peker, A. Kocyigit, and P. E. Eren, “LRFMP model for customer segmentation in the grocery retail industry: a case study,” Mark. Intell. Plan., vol. 35, no. 4, pp. 544–559, 2017, doi: 10.1108/MIP-11-2016-0210.

[4] W. M. Lim, “How can challenger marketers target the right customer organization? The A-C-O-W customer organization profiling matrix for challenger marketing,” J. Bus. Ind. Mark., 2018, doi: 10.1108/JBIM-02-2017-0039.

[5] D. Chen, S. L. Sain, and K. Guo, “Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining,” J. Database Mark. Cust. Strateg. Manag., vol. 19, no. 3, pp. 197–208, 2012, doi: 10.1057/dbm.2012.17.

[6] I. Maryani and D. Riana, “Clustering and profiling of customers using RFM for customer relationship management recommendations,” 2017 5th Int. Conf. Cyber IT Serv. Manag. CITSM 2017, pp. 2–7, 2017, doi: 10.1109/CITSM.2017.8089258.

[7] F. Liu, S. Zhao, and Y. Li, “How many, how often, and how new? A multivariate profiling of mobile app users,” J. Retail. Consum. Serv., vol. 38, no. December 2016, pp. 71–80, 2017, doi: 10.1016/j.jretconser.2017.05.008.

[8] P. Makhija, “RFM Analysis for Customer Segmentation,” CleverTap, 2020. https://clevertap.com/blog/rfm-analysis/ (accessed Aug. 25, 2020).

[9] S. Palaniappan, A. Mustapha, C. F. Mohd Foozy, and R. Atan, “Customer Profiling using Classification Approach for Bank Telemarketing,” JOIV Int. J. Informatics Vis., vol. 1, no. 4–2, p. 214, 2018, doi: 10.30630/joiv.1.4-2.68.

[10] S. Y. Hung, D. C. Yen, and H. Y. Wang, “Applying data mining to telecom churn management,” Expert Syst. Appl., vol. 31, no. 3, pp. 515–524, 2006, doi: 10.1016/j.eswa.2005.09.080.

[11] R. Sankar, “Customer Data Clustering Using Data Mining Technique,” Int. J. Database Manag. Syst., vol. 3, no. 4, pp. 1–11, 2011, doi: 10.5121/ijdms.2011.3401.

[12] H. Ma and D. Gang, “The customer relationship management based on data mining,” WIT Trans. Eng. Sci., vol. 80, pp. 287–294, 2013, doi: 10.2495/aie120341.

[13] D. Dong, J. Zhang, and J. Ye, “Research on Customer Segmentation Method of Commercial Bank Based on Data Mining,” 2017 3rd Int. Conf. Innov. Dev. E-commerce Logist., no. Icidel, pp. 62–65, 2017.

[14] M. Aryuni, E. Didik Madyatmadja, and E. Miranda, “Customer Segmentation in XYZ Bank Using K-Means and K-Medoids Clustering,” Proc. 2018 Int. Conf. Inf. Manag. Technol. ICIMTech 2018, no. September, pp. 412–416, 2018, doi: 10.1109/ICIMTech.2018.8528086.

[15] A. Dursun and M. Caber, “Using data mining techniques for profiling profitable hotel customers: An application of RFM analysis,” Tour. Manag. Perspect., vol. 18, pp. 153–160, 2016, doi: 10.1016/j.tmp.2016.03.001.

[16] J. Silva, N. Varela, L. A. B. López, and R. H. R. Millán, “Association rules extraction for customer segmentation in the SMES sector using the apriori algorithm,” Procedia Comput. Sci., vol. 151, no. 2018, pp. 1207–1212, 2019, doi: 10.1016/j.procs.2019.04.173.

[17] O. Dogan, E. Aycin, and Z. A. Bulut, “Customer Segmentation By Using Rfm Model and Clustering Methods: a Case Study in Retail Industry,” Int. J. Contemp. Econ. Adm. Sci., vol. 8, no. 1, pp. 1–19, 2018.

[18] Z. Li, “Research on customer segmentation in retailing based on clustering model,” 2011 Int. Conf. Comput. Sci. Serv. Syst. CSSS 2011 - Proc., pp. 3437–3440, 2011, doi: 10.1109/CSSS.2011.5974496.

[19] J. T. Wei, S. Y. Lin, C. C. Weng, and H. H. Wu, “A case study of applying LRFM model in market segmentation of a children’s dental clinic,” Expert Syst. Appl., vol. 39, no. 5, pp. 5529–5533, 2012, doi: 10.1016/j.eswa.2011.11.066.

[20] Y. H. Hu and T. W. Yeh, “Discovering valuable frequent patterns based on RFM analysis without customer identification information,” Knowledge-Based Syst., vol. 61, pp. 76–88, 2014, doi: 10.1016/j.knosys. 2014.02.009.

[21] M. Walters and J. Bekker, “Customer Super-Profiling Demonstrator To Enable Efficient Targeting in Marketing Campaigns,” South African J. Ind. Eng., vol. 28, no. 3, pp. 113–127, 2017, doi: 10.7166/28-3-1846.

[22] H. Takci, P. A. Sarvari, and A. Ustundag, “Performance evaluation of different customer segmentation approaches based on RFM and demographics analysis,” Kybernetes, vol. 45, no. 7, pp. 1129–1157, 2016, doi: 10.1108/K-07-2015-0180.

[23] A. J. Christy, A. Umamakeswari, L. Priyatharsini, and A. Neyaa, “RFM ranking – An effective approach to customer segmentation,” J. King Saud Univ. - Comput. Inf. Sci., 2018, doi: 10.1016/j.jksuci.2018.09.004.

[24] S. Petrovic, “A Comparison Between the Silhouette Index and the Davies-Bouldin Index in Labelling IDS Clusters,” 11th Nord. Work. Secur. IT-systems, pp. 53–64, 2006, [Online]. Available: https://xp-dev.com/svn/b_frydrych.../silhuetteIndexRegulaStopu.pdf.

[25] K. Mahendru, “How to Determine the Optimal K for K-Means?,” Analytics Vidhya, 2019. https://medium.com/analytics-vidhya/how-to-determine-the-optimal-k-for-k-means-708505d204eb.

[26] N. Bertagnolli, “Elbow Method and Finding the Right Number of Clusters,” 2015. http://www.nbertagnolli.com/jekyll/update/2015/12/10/ Elbow.html.

[27] A. Perera, “Finding the optimal number of clusters for K-Means through Elbow method using a mathematical approach compared to graphical approach,” 2017. https://www.linkedin.com/pulse/finding-optimal-number-clusters-k-means-through-elbow-asanka-perera#:~:text=So the most easiest way,the use of Elbow method.&text=Just like the name suggests,points against the cluster centroid.

[28] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez, and I. Perona, “An extensive comparative study of cluster validity indices,” Pattern Recognit., vol. 46, no. 1, pp. 243–256, 2013, doi: 10.1016/j.patcog. 2012.07.021.

[29] C. Yuan and H. Yang, “Research on K-Value Selection Method of K-Means Clustering Algorithm,” J, vol. 2, no. 2, pp. 226–235, 2019, doi: 10.3390/j2020016.