

Customer segmentation using association rule mining on retail transaction data

Siriwan Kajornkasirat¹, Pattarawan Gunclin¹, Kritsada Puangsuwan¹, Nawapon Kaewsuwan²

¹Faculty of Science and Industrial Technology, Prince of Songkla University, Surat Thani Campus, Surat Thani, Thailand

²Faculty of Humanities and Social Sciences, Prince of Songkla University, Pattani Campus, Pattani, Thailand

Article Info

Article history:

Received Dec 11, 2023

Revised Nov 21, 2024

Accepted Jan 27, 2025

Keywords:

Association rules

Clustering algorithm

Customer segmentation

Frequent pattern-growth

Recency, frequency, and
monetary analysis

ABSTRACT

This research aimed to investigate a suitable algorithm for customer segmentation using as customer behavior indicators the recency, frequency, and monetary (RFM) values of the customers. The clustering algorithms K-means, fuzzy C-means, and self-organizing neural network (SONN) were compared for finding the most appropriate algorithm. The customer segmentation was analyzed using association rule mining with the frequent pattern algorithm (FP-Growth). Data on retail transactions during January 2021 - May 2023 were obtained from Tuenjai Company, Thailand, with a total of 202,469 records. The results from the three algorithms were compared by the silhouette coefficient (SC), Calinski-Harabasz (CH) index, Davies-Bouldin (DB) index, iteration count, and execution time. The results showed that the K-means algorithm was the most suitable algorithm for customer segmentation in this study. K-means clustering grouped the customers into three groups here labeled as “important value”, “general development”, and “lost”, based on the RFM values. There were 38 rules for the important value segment, and two rules each for the general development and the lost groups. These results could be useful to the business organization for improving the customer experiences, increasing sales, preparing or promoting products, and stock management efficiency.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Siriwan Kajornkasirat

Faculty of Science and Industrial Technology, Prince of Songkla University, Surat Thani Campus

31 Moo 6, Makhantia, Muang, Surat Thani 84000, Thailand

Email: siriwan.wo@psu.ac.th

1. INTRODUCTION

Online shopping has become one of the most popular online activities, offering customers the convenience of placing orders through their smartphones or other mobile devices, and eliminating the need to visit physical stores. According to e-commerce trends, in 2022 a significant 68.3% of Thai people engaged in online shopping every week [1]. However, the abundance of products available online often leads to decision-making challenges for customers, resulting in prolonged browsing and hesitation [2], [3].

To address this issue, recommendation systems play a crucial role in business activities by suggesting products to customers based on various factors such as purchase history, best-selling products, or customer interests [3]. Previous research has shown that recommendations significantly influence customers to make more purchases [4]. Effective customer segmentation is essential for recommendation systems to tailor products and services to suit the diverse needs of customers [5].

Customer segmentation, vital for understanding customer needs and preferences, is often based on various factors such as gender, age, lifestyle, or customer purchase behavior [5]. This technique has been used in various contexts, including automotive industry, travel business, and electricity market, to decipher

consumer preferences and behaviors [6]–[8]. This segmentation is facilitated by clustering algorithms and tools, which streamline the process and enhance accuracy [9]–[11]. In addition to segmentation, customer behavior analysis, particularly using the recency, frequency, and monetary (RFM) model, is crucial for retaining existing customers and attracting new ones. The RFM model helps identify valuable customer segments and devise targeted strategies for growth and retention [5], [12]. Association rule mining further enhances understanding of purchase behavior by uncovering patterns and relationships within transaction data [13]. This approach has been applied across various industries, including sports, food, and product design [14]–[16].

This study investigates the integration of customer segmentation and association rule mining for sales data using RFM analysis with the clustering algorithms K-means, fuzzy C-means, and self-organizing neural network (SONN) to identify the most effective approach for segmentation. The evaluation metrics silhouette coefficient (SC), Calinski-Harabasz (CH) index, Davies-Bouldin (DB) index, iteration count, and execution time were employed in comparisons. Subsequently, the frequent pattern algorithm (FP-Growth) algorithm was used to unveil association rules within each customer segment.

The subsequent sections of this paper are organized as follows: section 2 elucidates the methodology adopted in this study. Section 3 presents the results and discussion, including comparisons with prior research. Section 4 delineates the limitations, future research directions, and implications of our findings. Finally, section 5 concludes the paper by summarizing the key insights gleaned from this investigation.

2. METHOD

The three high level stages in this study were data collection, data preparation, and data analysis, the last one comprising customer segmentation, customer classification and association rules mining as shown in Figure 1. Subsections 2.1-2.3 elaborate on each step. These steps are important processes in selecting the best methods for customer segmentation.

2.1. Data collection

Data on retail transactions during January 2021 - May 2023 were obtained from Tuenjai Panit Group Company Limited (Co., Ltd.), Thailand for a total of 202,469 records covering 38 stores in the Surat Thani province. There were six attributes per customer transaction, namely member number, sale date, bill number, product name, the number of product sales, and sales (Baht). We contacted relevant stakeholders to request data for our research and the wait for responses took a significant time due to the workload of the stakeholders and the process of exporting data from the system.

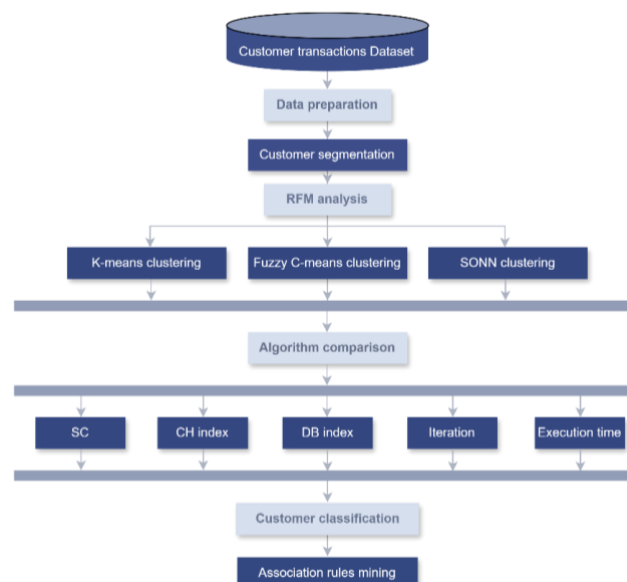


Figure 1. Schematic outline of the study approach

2.2. Data preparation

Data preparation ensures that raw data are accurate and complete, thereby suitable for analysis to find knowledge or insights [17]. The data preparation process in this study consisted of data cleansing and data

transformation to prepare raw data into a format that can be analyzed. The large dataset may contain errors or be inconsistent, so data cleansing and data transformation are necessary to avoid problems with the analysis.

2.2.1. Data cleansing

The data cleansing process identifies and corrects errors or missing data to prepare it for analysis [17], [18]. The first step is to look for entries with missing data points, and to remove rows or columns with a high percentage of missing values, especially if the missing data are not systematic and removing them won't significantly impact the analysis. In this process, we used the “.isnull().sum()” command to check for all missing values in the data. When we find a missing value, we used the “.dropna(inplace=True)” command to delete the row with the missing value before performing RFM analysis. After that, we used the interquartile range (IQR) method to check for outliers and remove them from the results before clustering. We used the 75th and 25th quartile values as thresholds. In addition, we checked the member number to make sure it is valid, deleting rows where the member number had special characters and was not a 10-digit phone number. After the data cleansing, customer transaction data were reduced to 173,261 records, having removed 14.44% of the data due to duplicate records, missing values, special characters, incorrect data, and outliers.

2.2.2. Data transformation

Data transformation converts data into a format that is suitable for analytical purposes [18]. In this process the data types in the dataset were adjusted to prevent problems during data analysis. We used the “.to_datetime()” command with the sale date to convert the string data into a datetime object format. This was done if the column data was in string format before RFM analysis. We then used “StandardScaler()” with the RFM results to normalize the data before clustering.

2.3. Data analysis

In the data analysis process, if the computer used is not powerful enough, this may result in long waiting times for results, especially when dealing with large datasets, particularly in the customer segmentation process. Customer segmentation consists of three sub-steps. The first step is RFM analysis, followed by K-means, fuzzy C-means, and SONN clustering, which comprised the second step. Finally, the last step involved comparing the clustering algorithms based on SC, CH index, DB index, iteration count, and execution time to determine the optimal approach for customer segmentation. In the customer classification process, labels were assigned to customer groups based on their RFM values, which represent their purchasing behaviors. Association rule mining is the analysis of relationships between items that have been previously purchased by customers in each group. The customer classification process and association rule mining are the subsequent steps following customer segmentation.

2.3.1. Customer segmentation

This study used RFM analysis for customer segmentation together with the clustering algorithms K-means, SONN, and fuzzy C-means [19]–[21]. We analyzed preprocessed customer transaction data to identify the purchasing behavior of each customer, determining the recent purchase date (Recency: R), the frequency of purchases (Frequency: F), and the amount of money spent by the customer (Monetary: M) [5]. The results of the RFM analysis were used to segment customers using clustering algorithms in the algorithm comparison step (i.e., K-means, SONN, and fuzzy C-means).

In the algorithm comparison step, we used 5 criteria to compare the algorithms to find the most suitable algorithm. These criteria were SC, CH index, DB index, iteration count, and execution time [21], [22]. SC is based on the distance between groups and the distance between points in the group, giving values in the range [-1, 1]. If this value is high that indicates suitable clustering, but if this value is low or negative that indicates unsuitable clustering. CH index is calculated to assess dense and sparse parts in a group. If this value is high that group is dense and well-spaced: the higher, the better. This is the opposite of the DB index, for which a low value indicates that the group is dense and well-spaced [22]. We used 'sklearn.cluster' for K-means clustering and to find its iteration count. For fuzzy C-means clustering and its iteration count, we relied on 'fcmmeans'. Meanwhile, 'minisom' was used for SONN clustering and its iteration count. The execution time of each clustering comes from the time displayed on the code editor program. For computing SC, CH index, and DB index, we utilized 'sklearn.metrics'.

2.3.2. Customer classification

After selecting the most suitable clustering algorithm, we used RFM values in each cluster to name the customer segments according to the customer classification table [23] as shown in Table 1. Sun *et al.* [23] identified eight types of customers, which are important value, important development, important protection, important retention, general value, general development, general retention, and lost. These types exhibit different purchasing behaviors, as shown in Table 1.

Table 1. Cluster naming based on the RFM values

Customer type	Recency	Frequency	Monetary
Important value	High	High	High
Important development	High	Low	High
Important protection	Low	High	High
Important retention	Low	Low	High
General value	High	High	Low
General development	High	Low	Low
General retention	Low	High	Low
Lost	Low	Low	Low

From the customer classification as in Table 1, we can see that the RFM values were quantized too high or low. These values can be calculated using the RFM values of a cluster to determine whether they are high or low by comparing them to the averages in RFM values over that same cluster. The recency is high when that value is less than the average recency (R_{avg}). The frequency and monetary values are high when these exceed the averages of frequency (F_{avg}) and monetary (M_{avg}) as in (1), (2), and (3) [24]:

$$Recency\ value = \begin{cases} High, Recency\ value < R_{avg} \\ Low, Recency\ value \geq R_{avg} \end{cases} \quad (1)$$

$$Frequency\ value = \begin{cases} High, Frequency\ value > F_{avg} \\ Low, Frequency\ value \leq F_{avg} \end{cases} \quad (2)$$

$$Monetary\ value = \begin{cases} High, Monetary\ value > M_{avg} \\ Low, Monetary\ value \leq M_{avg} \end{cases} \quad (3)$$

- K-means algorithm: K-means is an unsupervised learning algorithm used to cluster data into groups based on their similarity. It works by first initializing k centroids, where k is the number of clusters desired. Then, it calculates the distance between each data point and the centroids to assign the data points to the cluster with the minimum distance. The centroids are then updated, and the process is repeated until the centroids no longer change [19].
- SONN algorithm: SONN is a subtype of artificial neural networks (ANNs) called self-organizing map (SOM) or Kohonen Map. This is an unsupervised learning algorithm and uses competitive learning in training the adjustable parameters. A SOM shows data clustering by reducing the high-dimensional data to a low dimension. In the structure of this algorithm, each input neuron in the input layer is connected to all neurons in the output layer (Kohonen layer) through weighted connections. When data are fed to an input neuron, they are transmitted to all neurons in the output layer, essentially broadcasting the same information. Among the neurons in the Kohonen layer, the one with the highest weight is chosen as the output for customer segmentation. This means the neuron with the strongest connection to the input data is considered the winner and its activation determines the customer segment [20].
- Fuzzy C-means algorithm: fuzzy C-means is an unsupervised algorithm. When clustering, each object in the cluster has a weighting associated with a particular cluster. Therefore, some objects may be in cluster_i, but their position is not in this cluster_i. The principle of this algorithm is to emphasize the uncertainty of membership in a group by using likelihood to determine which group each object is likely to belong to. This is done by calculating the weight or probability of membership in each group, considering the distance between the object and the centroid of the group in each iteration of the algorithm. The use of weights allows specifying the degree to which each object belongs to a particular group, rather than forcing them to belong exclusively to one group [21].

2.3.3. Model evaluation

To identify the optimal clustering algorithm for customer segmentation, we compared several metrics including SC, CH index, DB index, iteration count, and execution time [19]. The values from these evaluation criteria help us make appropriate decisions when selecting clustering algorithms. We can obtain the SC, CH index, and DB index according to the calculations in (4) to (6):

$$SC = \frac{(b_i - a_i)}{\max(a_i, b_i)}, a_i > b_i \quad (4)$$

Where SC represents the SC value; a_i is the average distance between object_i and other objects in the cluster; b_i is the average distance between object_i and another cluster.

$$CH = \frac{\sum S_B}{\sum S_W} \cdot \frac{n_p - 1}{n_p - k} \quad (5)$$

Where CH represents the CH index value; S_B is the inter-cluster dispersion matrix; S_W is the intra-cluster dispersion matrix; n_p is the number of grouped samples; k is the cluster number.

$$DB = \frac{1}{c} \sum_{i=1}^c \max_{i \neq j} \left\{ \frac{d(x_i) + d(x_j)}{d(c_i, c_j)} \right\} \quad (6)$$

Where DB represents the DB index value; c is the cluster number; i and j are the cluster label; $d(x_i)$ and $d(x_j)$ are all samples in clusters i and cluster j to their respective cluster centroids; $d(c_i, c_j)$ is the distance between clusters i and cluster j .

2.3.4. Association rule mining

We utilized the 'mlxtend.frequent_patterns' module for conducting this process. After assigning names to the customer segments, we performed association rule mining using FP-Growth algorithm on the customer transaction data to find associations between the related products in each segment. Since the transaction data for each customer segment is very large, we selected a sample of 1,000 records with the highest customer spending to analyze. Products that were purchased were assigned a value of true, while products that were not purchased were assigned a value of false. We set the support and confidence thresholds to be at least 0.01 and 0.57, respectively. Confidence and support were calculated with the (7) and (8) [13].

$$\text{Confidence} = \frac{\text{The frequency of A,B in customer transactions}}{\text{The frequency of A in customer transactions}} \quad (7)$$

$$\text{Support}(A) = \frac{\text{The frequency of A in customer transactions}}{\text{Total of transactions}} \quad (8)$$

3. RESULTS AND DISCUSSION

3.1. Results

For the sales data from 2021-2023, the RFM indexes were used with K-means, fuzzy C-means, and SONN algorithms for customer segmentation by clustering. We found that the preferred number of clusters in these algorithms was 3 (Figure 2). The K-means did cluster by centroid points. The fuzzy C-means also clustered by centroid points like K-means, and calculated the probability for a cluster position. The SONN algorithm did clustering by mapping. From the scatter plots in Figure 2, it can be observed that there was no significant difference in the clustering results between K-means (Figure 2(a)) and SONN clustering (Figure 2(b)) algorithms. On the other hand, fuzzy C-means clustering (Figure 2(c)) shows different results compared to the previous two algorithms. The cluster positions of the customers were switched. However, the values of SC, CH index, DB index, iteration count, and execution time of all three algorithms clearly demonstrate some differences in clustering performance.

The comparison of clustering algorithms with the SC, CH index, and DB index shows that the K-means, fuzzy C-means, and SONN are suitable for segmentation to 3 groups by the count of wins across the comparison criteria (two clusters wins 3 times; three clusters win 6 times). The highest SC and CH indexes and the lowest DB index were achieved by the K-means algorithm as presented in Table 2. The iteration and execution time for K-means clustering was lower than for the fuzzy C-means and SONN (K-means: iteration count was 3 and execution time was 0.6076 seconds; fuzzy C-means: iteration count was 150 and execution time 2.5954 seconds; SONN: iteration count was 45 and execution time 3.2230 seconds).

Therefore, the K-means algorithm performed better than fuzzy C-means and SONN in customer segmentation of these data. The number of clusters based on the RFM in K-means clustering was k-3. The customer purchase behavior as presented in Figure 3 indicators recency (Figure 3(a)), frequency (Figure 3(b)), and monetary (Figure 3(c)) are graphically illustrated for each cluster.

We used RFM values from the customer purchase behavior of each cluster (Figure 3) to calculate the average RFM values for a comparison between groups. The average recency value is 223, the average frequency value is 7, and the average monetary value is 1,351 as presented in Table 3. We used comparisons between RFM value and average RFM value to label the customer groups. Cluster 1 is high in recency but low in frequency and monetary values, so it is the general development group. Cluster 2 is high in all RFM values and is the important value group. Cluster 3 is low in RFM values and is labeled the lost group.

The analysis results of our group members show that Tuenjai Company has the highest number of customers in the general development category, ranking first among all customers. Lost ranks second,

followed by important value as third. The labels given to each cluster summarize the customer purchasing behaviors based on Figure 3. Detailed information on each group can be seen in Table 4.

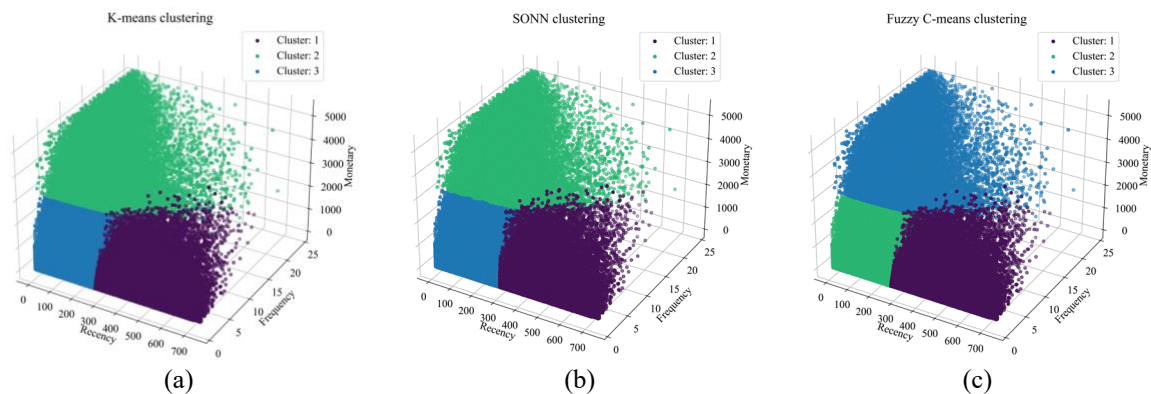


Figure 2. Customer segmentation by (a) K-means, (b) SONN, and (c) fuzzy C-means clustering algorithms

Table 2. A comparison of clustering algorithms with the SC, CH index, and DB index

The number of clusters	K-means			Fuzzy C-means			SONN		
	SC	CH index	DB index	SC	CH index	DB index	SC	CH index	DB index
2	0.4625	156318	0.8914	0.4494	155082	0.9205	0.4804	149416	0.8410
3	0.4364	184977	0.7913	0.4355	184920	0.7943	0.4361	181914	0.7931
4	0.4008	173952	0.9287	0.3947	173595	0.9253	0.3856	169578	0.9287
5	0.3495	157078	0.9691	0.3150	147978	1.1130	0.3489	149227	0.9839
6	0.3502	150646	1.0086	0.3276	144799	1.1020	0.3432	136502	0.9891

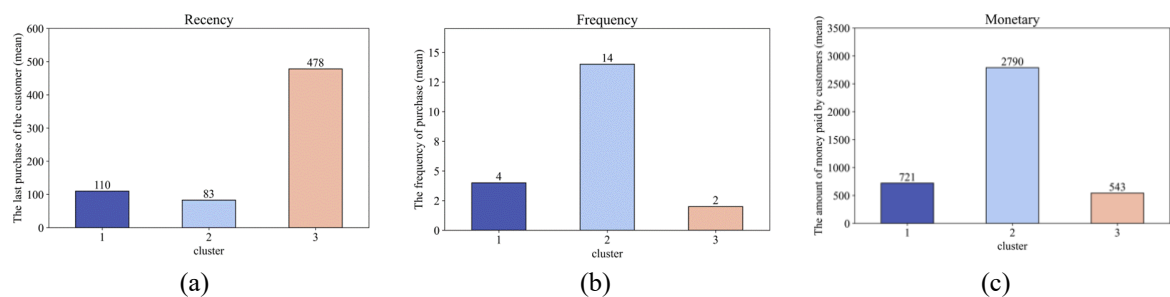


Figure 3. Customer purchase behaviors by cluster (a) recency, (b) frequency, and (c) monetary

Table 3. A comparison between RFM values in each cluster and average RFM values

Cluster	Recency vs average recency	Frequency vs average frequency	Monetary vs average monetary
1	High (110<223)	Low (4<7)	Low (721<1,351)
2	High (83<223)	High (14>7)	High (2,790>1,351)
3	Low (478>223)	Low (2<7)	Low (543<1,351)

Table 4. Customers in each group

Group	Number of people (%)	Average last purchase	Average purchase frequency (times)	Expenses average (THB)
Important value	39,440 (23)	About 2-3 months ago	14	2,790
Lost	47,205 (27)	About 1-2 years ago	2	543
General development	86,616 (50)	About 3-4 months ago	4	721

For the association rule mining, we selected 1,000 samples from each cluster based on top monetary value. We set the minimum support to 0.01 and confidence at 0.57. After association rule analysis of each cluster, there were 38 rules for the important value group as shown in Table 5, two rules for the general development group as explained in Table 6, and two rules for the lost group as presented in Table 7.

Table 5. Association rules in the important value group

Antecedents	Consequents	Support	Confidence
Single items			
- Pressure adjusting head No. R326 (SCG)	- Gas cable, white/orange, 50m.	0.010	1.000
- Glue gun No. 9049/ST001	- Transparent glue sticks 1*12 (Soji)	0.012	0.857
- Free wrapping service	- Gift wrapping paper-100 sheets	0.012	0.857
- Assorted pull-bow ribbons	- Gift wrapping paper-100 sheets	0.031	0.795
- Gas hose clip No. 18, thin type	- Gas cable, white/orange, 50m.	0.011	0.786
- Gas regulator No. 339/889 (LGA)	- Gas cable, white/orange, 50m.	0.010	0.769
- Thick gas hose clip	- Gas cable, white/orange, 50m.	0.021	0.750
- Assorted pull bow ribbons	- Gift wrapping paper-100 sheets	0.028	0.718
- Box with locking lid 60 (55) l., black (CNN)	- Box with locking lid 60 (55) l., assorted colors (CNN)	0.020	0.606
- Children's clothes hanger winner 1*6	- Clothes hanger No.3 (Eagle brand)	0.013	0.591
- Diagonal cutting knife 6.5 in. No.171P (Kiwi brand)	- Big spoon (zebra head)	0.011	0.579
Double items			
- Tissue paper 1*6 (Vivy)	- Gift wrapping paper-100 sheets	0.010	1.000
- Assorted pull bow ribbons			
- Vipada orange soap mixed with collagen	- Vipada carrot soap 65g.	0.010	0.909
- Leo soft fabric softener, touch of love scent, blue, 500ml.			
- Assorted pull bow ribbons	- Gift wrapping paper-100 sheets	0.015	0.882
- Large spoon (Zebra brand)			
- De Paree fabric softener, joy scent, refill bag 540ml.	- De Paree fabric softener, blooming scent, refill bag type, 540ml.	0.011	0.846
- De Paree fabric softener, blue paradise scent, refill bag type, 540ml.			
- Box with locking lid 60 (55) l., assorted colors (CNN)	- Gift wrapping paper-100 sheets (5B)	0.010	0.833
- Assorted pull bow ribbons			
- Famony fabric softener, sweet beautiful scent (purple) 300ml.	- Farmony fabric softener, romance scent (red), 300ml.	0.010	0.769
- Palm oil 1l. (Flower brand)			
- 108 Shop detergent 1000g.	- Palm oil 1l. (Flower brand)	0.010	0.769
- Pinto dishwashing liquid, lemon scent 450/430/420ml.			
- Pinto dishwashing liquid, lemon scent 450/430/420ml.	- Palm oil 1l. (Flower brand)	0.012	0.750
- Leo soft fabric softener, touch of love scent, blue, 500ml.			
- Assorted pull bow ribbons	- Gift wrapping paper-100 sheets (5B)	0.012	0.750
- Large spoon (Zebra brand)			
- Tissue paper-180 sheets (Solly brand)	- Small piece of linoleum 2.50 m.	0.011	0.733
- Large piece of linoleum 2.70m.			
- Leo soft fabric softener, touch of love scent, blue, 500ml.	- Leo Soft fabric softener, purple passion scent, purple, 500ml.	0.019	0.731
- Leo soft fabric softener, sweet floral scent, pink, 500ml.			
- Box with locking lid 60 (55) l., assorted colors (CNN)	- Small piece of linoleum 2.50m.	0.019	0.679
- Large piece of linoleum 2.70m.			
- Pinto dishwashing liquid, strawberry scent 450/400ml.	- Palm oil 1l. (Flower brand)	0.010	0.667
- Gas gun (Top Light)			
- Pinto dishwashing liquid, kiwi scent, 450/400ml.	- Cook Leo dishwashing liquid, lemon scent, 450/420ml.	0.014	0.667
- Palm oil 1l. (Flower brand)			
- Palm olein oil 1l. (Bee brand)	- Palm oil 1l. (Flower brand)	0.015	0.652
- Lighter 1*3 (MOTO)			
- Vivy tissue paper 1*6	- Vipada orange soap mixed with collagen	0.011	0.647
- Vipada carrot soap 65g.			
- Gas gun (Top Light)	- Palm oil 1l. (Flower brand)	0.014	0.636
- Cook Leo dishwashing liquid, lemon scent, 450/420ml.			
- Palm olein oil 1l. (Bee brand)	- Palm oil 1l. (Flower brand)	0.010	0.625
- Feather razor blade (Feather)			
- Palm olein oil 1l. (Bee brand)	- Palm oil 1l. (Flower brand)	0.010	0.625
- Professional Green Forest 2700g.			
- Wet wipes (HAPPYHIPPO)	- Clothes hanger No.3 (Eagle brand)	0.011	0.611
- Large spoon (zebra head)			
- Aluminum teaspoon (Flower brand)	- Palm oil 1l. (Flower brand)	0.011	0.611
- Vivy tissue paper 1*6.			
- Pinto dishwashing liquid, strawberry scent 450/400ml.	- Big spoon (zebra head)	0.011	0.611
- Clothes hanger No.3 (Eagle brand)			
- Professional Green Forest 2700g.	- Palm oil 1l. (Flower brand)	0.011	0.611
- Cook Leo dishwashing liquid, lemon scent 450/420ml.			
- Leo Soft fabric softener, sweet floral scent, pink, 500ml.	- Palm oil 1l. (Flower brand)	0.015	0.600
- Cook Leo dishwashing liquid, lemon scent, 450/420ml.			
- Cook Leo dishwashing liquid, lemon scent, 450/420ml.	- Palm oil 1l. (Flower brand)	0.010	0.588
- Cook Leo dishwashing liquid, strawberry kiwi scent 450ml.			
- Aluminum teaspoon (Flower brand)	- Box with locking lid 60(55) l., assorted colors (CNN)	0.011	0.579
- Clothes hanger No.3 (Eagle brand)			
- Pinto dishwashing liquid, lemon scent, 450/430/420ml.	- Palm oil 1l. (Flower brand)	0.011	0.579
- Leo soft fabric softener, purple passion scent, purple, 500ml.			

Table 6. Association rules in the general development group

Antecedents	Consequents	Support	Confidence
- Thick gas hose clip	- Gas cable, white/orange, 50m.	0.012	0.706
- 6 pairs of spoons and forks, frangipani pattern No.100030 (Zebra brand)	- Big spoon (Zebra brand)	0.012	0.571

Table 7. Association rules in the lost group

Antecedents	Consequents	Support	Confidence
- Assorted pull bow ribbons	- Gift wrapping paper-100 sheets	0.016	0.941
- Gas cable, white/orange, 50m.	- Thick gas hose clip	0.010	0.625

Important value: the association rules show that the customers in this cluster bought a pressure adjusting head No. R326 (SCG) and then bought a gas cable, white/orange, 50m repeatedly. There is 100% confidence in this rule (Table 5). This means that every time customers in this group purchase a pressure adjusting head No. R326 (SCG), they always buy a gas cable, white/orange, 50m.

General development: the association rules show that customers in this cluster bought a thick gas hose clip and then bought a gas cable, white/orange, 50m, too. There is 70.6% confidence in this rule (Table 6). This means that every time customers in this group purchase a thick gas hose clip, there is a 70.6% chance that they will also buy a gas cable, white/orange, 50m.

Lost: the association rules show that customers in this cluster bought assorted pull bow ribbons and then bought gift wrapping paper-100 sheets, too. There is 94.1% confidence in this rule (Table 7). This means that every time customers in this group purchase assorted pull bow ribbons, there is a 94.1% chance that they will also buy gift wrapping paper -100 sheets.

3.2. Discussion

In today's fast-paced business landscape, effectively segmenting customers plays a crucial role in optimizing collaboration with recommendation systems. Customer segmentation provides deep insights into purchasing behaviors, enabling organizations to customize their offerings and services to better meet customer needs, ultimately resulting in heightened satisfaction [25]. The process involves employing clustering algorithms for customer segmentation. For instance, banking sector uses behavior-based segmentation, restaurants use demographic-based segmentation, and the same principle applies to segmentation in supplement product markets, and to other contexts [26]–[28]. Even though we may study previous research on the workings of clustering algorithms, we cannot be certain whether those algorithms are suitable for the sales dataset we are using or not.

This study meticulously compared various clustering algorithms to determine the most suitable approach for customer segmentation, utilizing transaction data from Tuenjai Panit Group Co., Ltd., Thailand, spanning from January 2021 to May 2023. Through rigorous RFM analysis, we evaluated the performance of three prominent clustering algorithms: K-means, fuzzy C-means, and SONN, using key metrics such as SC, CH index, DB index, iteration count, and execution time [21], [22]. Notably, our findings strongly favored the K-means algorithm due to its exceptional speed and accuracy, particularly evident when handling large datasets [19].

Our results are consistent with prior research, affirming the widespread applicability of the K-means algorithm across diverse industries such as e-commerce, restaurants, and marketing [17], [27], [29]. The robust performance metrics of K-means, including high SC and CH indices coupled with lower DB index, iteration count, and execution time, further validate its effectiveness in customer segmentation (SC :0.4364; CH index :184977.0535; DB index :0.7913; iterations :3 times; execution time :0.6076 seconds). Three segments were identified by the customer segmentation, as follows. Important value group: customers with high RFM values. General development group: customers with high recency values but low frequency and monetary values. Lost group: customers with low RFM values. The general development group was the most prevalent customer type in this current study, comprising 50% of all customers.

Additionally, we conducted an analysis to uncover associations between previously purchased products within each cluster. In previous research, association rule mining was employed to analyze customer transactions following customer segmentation. This approach aids in uncovering hidden customer purchase patterns and generating association rules, which reveal connections between previously purchased products [14]. Our results demonstrated associations between previously purchased products for customers in each cluster (important value group: 38 rules; general development group: 2 rules; lost group: 2 rules). This outcome validates the use of association rule mining in the analysis of customer transactions.

However, it's important to acknowledge the limitations of our study. Future research could explore alternative clustering techniques or incorporate additional variables to enhance the accuracy and depth of

customer profiling. Moving forward, future studies could delve into the integration of advanced data mining techniques or machine learning algorithms to refine customer segmentation methodologies further.

Additionally, longitudinal analyses assessing the impacts of segmentation strategies on customer retention and profitability could yield valuable strategic insights for businesses. In summary, our study highlights the efficacy of the K-means algorithm for customer segmentation, particularly in analyzing extensive datasets. The associations revealed between previously purchased products within each cluster underscore the utility of association rule mining in deciphering customer behaviors. These segmentation findings have significant implications for business strategies, ranging from demand forecasting to targeted marketing efforts and inventory management [30].

4. LIMITATIONS, FUTURE RESEARCH, AND IMPLICATIONS

Since we utilize actual sales data from the company in this study, certain information cannot be disclosed. Member numbers, for instance, are considered personal data of customers as they are phone numbers. In the clustering process, we need to determine the number of clusters or the k-value each time we perform clustering. Furthermore, in the section concerning the analysis of the relationship rules, we did not select the time frame for the data. This results in a lack of knowledge regarding when the products displayed on those rules were purchased by customers. Such lack of information could potentially lead to issues in sales planning, promotions, and inventory management.

In future research, we aim to develop a recommendation system in the form of a web application that performs automatic customer segmentation. Users won't need to input the k-value themselves; the system will determine it automatically, including naming the customer groups. This system will assist in recommending products and promotions to retail stores. In addition to selecting customer groups for data viewing, users will also be able to choose specific time intervals for displaying information. This will enable organizations to identify which products certain customer segments tend to purchase together during specific time periods, facilitating inventory management and promotion planning. This research may contribute to enhancing business efficiency and effectively meeting the needs and importance of customers. Moreover, research in this area can serve as a guideline for developing more efficient business practices in the future by introducing strategies and marketing approaches based on solid and effective data.

5. CONCLUSION

This study performed a comparison of clustering algorithms, specifically K-means, fuzzy C-means, and SONN when applied alongside the RFM method, to determine the most suitable approach for customer segmentation. Five evaluation criteria were utilized: SC, CH index, DB index, iteration count, and execution time. Additionally, we employed the FP-Growth algorithm for association rule mining to uncover relationships among previously purchased products within each cluster. The findings from both customer segmentation and association rule mining offer valuable insights that can greatly inform product stock planning and promotional strategies within the organization. Moreover, these insights have the potential to advance the development of more effective product recommendation systems. By presenting a clear understanding of the efficiency of each clustering algorithm and its implications for practical applications, this study lays a solid foundation for informed decision making in marketing and customer relationship management.

ACKNOWLEDGEMENTS

The authors thank Seppo J. Karrila for constructive comments on this manuscript. This research was funded by Prince of Songkla University, Surat Thani Campus. We also thank Tuenjai Panit Group Company Limited (Co., Ltd.), Thailand for the data in this study.

FUNDING INFORMATION

This research was supported by Prince of Songkla University, Surat Thani Campus.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Siriwan Kajornkasirat	✓	✓		✓	✓	✓	✓		✓	✓		✓	✓	✓
Pattarawan Gunglin		✓	✓		✓	✓		✓	✓	✓	✓			
Kritsada Puangsuwan	✓	✓			✓				✓	✓	✓			
Nawapon Kaewsuwan		✓		✓	✓				✓	✓	✓			

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project administration

Fu : Funding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The authors confirm that the data supporting the findings of this study are available within the article.




REFERENCES

- [1] S. Kemp, "Digital 2024: global overview report," *Data Reportal*, 2024. Accessed: Nov. 28, 2023. [Online]. Available: <https://datareportal.com/reports/digital-2024-global-overview-report>
- [2] X. Dong, H. Tu, H. Zhu, T. Liu, X. Zhao, and K. Xie, "Does diversity facilitate consumer decisions: a comparative perspective based on single-category versus multi-category products," *Asia Pacific Journal of Marketing and Logistics*, vol. 36, no. 4, pp. 936–956, 2024, doi: 10.1108/APJML-05-2023-0395.
- [3] C. P. Gupta and V. V. R. Kumar, "Recommendation system: a transformative artificial intelligence tool for e-commerce," *International Conference on Informatics and Computational Sciences*, pp. 60–65, 2024, doi: 10.1109/ICICoS62600.2024.10636825.
- [4] V. Venkatesh, H. Hoehle, J. A. Aloysius, and H. R. Nikkiah, "Being at the cutting edge of online shopping: Role of recommendations and discounts on privacy perceptions," *Computers in Human Behavior*, vol. 121, 2021, doi: 10.1016/j.chb.2021.106785.
- [5] M. A. Gomes and T. Meisen, "A review on customer segmentation methods for personalized customer targeting in e-commerce use cases," *Information Systems and e-Business Management*, vol. 21, no. 3, pp. 527–570, 2023, doi: 10.1007/s10257-023-00640-4.
- [6] H. Hartoyo, E. Manalu, U. Sumarwan, and P. Nurhayati, "Driving success: a segmentation of customer admiration in automotive industry," *Journal of Open Innovation: Technology, Market, and Complexity*, vol. 9, no. 2, 2023, doi: 10.1016/j.joitmc.2023.100031.
- [7] E. P. Vargas, C. de-Juan-Ripoll, M. B. Panadero, and M. Alcañiz, "Lifestyle segmentation of tourists: the role of personality," *Heliyon*, vol. 7, no. 7, 2021, doi: 10.1016/j.heliyon.2021.e07579.
- [8] F. Barjak, J. Lindeque, J. Koch, and M. Soland, "Segmenting household electricity customers with quantitative and qualitative approaches," *Renewable and Sustainable Energy Reviews*, vol. 157, 2022, doi: 10.1016/j.rser.2021.112014.
- [9] V. Duarte, S. Zuniga-Jara, and S. Contreras, "Machine learning and marketing: a systematic literature review," *IEEE Access*, vol. 10, pp. 93273–93288, 2022, doi: 10.1109/ACCESS.2022.3202896.
- [10] D. H. Khoiriyah and R. Ambarwati, "Dynamic segmentation analysis for expedition services: integrating k-means and decision tree," *Journal of Information Systems and Informatics*, vol. 6, no. 1, pp. 363–377, 2024, doi: 10.51519/journalisi.v6i1.666.
- [11] M. A. Uddin *et al.*, "Data-driven strategies for digital native market segmentation using clustering," *International Journal of Cognitive Computing in Engineering*, vol. 5, pp. 178–191, 2024, doi: 10.1016/j.ijcce.2024.04.002.
- [12] M. E. Jalal and A. Elmaghraby, "Analyzing the dynamics of customer behavior: a new perspective on personalized marketing through counterfactual analysis," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 19, no. 3, pp. 1660–1681, 2024, doi: 10.3390/jtaer19030081.
- [13] W. Wahyuningsih and P. T. Prasetyaningrum, "Enhancing sales determination for coffee shop packages through associated data mining: leveraging the FP-growth algorithm," *Journal of Information Systems and Informatics*, vol. 5, no. 2, pp. 758–770, 2023, doi: 10.51519/journalisi.v5i2.500.
- [14] S. H. Liao, R. Widowati, and K. C. Yang, "Investigating sports behaviors and market in Taiwan for sports leisure and entertainment marketing online recommendations," *Entertainment Computing*, vol. 39, 2021, doi: 10.1016/j.entcom.2021.100442.
- [15] M. Rostami, U. Muhammad, S. Forouzandeh, K. Berahmand, V. Farrahi, and M. Oussalah, "An effective explainable food recommendation using deep image clustering and community detection," *Intelligent Systems with Applications*, vol. 16, 2022, doi: 10.1016/j.iswa.2022.200157.
- [16] R. Dou, W. Li, G. Nan, X. Wang, and Y. Zhou, "How can manufacturers make decisions on product appearance design? A research on optimal design based on customers' emotional satisfaction," *Journal of Management Science and Engineering*, vol. 6, no. 2, pp. 177–196, 2021, doi: 10.1016/j.jmse.2021.02.010.
- [17] A. Griva, "'I can get no e-satisfaction'. What analytics say? Evidence using satisfaction data from e-commerce," *Journal of Retailing and Consumer Services*, vol. 66, 2022, doi: 10.1016/j.jretconser.2022.102954.
- [18] D. Varma, A. Nehansh, and P. Swathy, "Data preprocessing toolkit: an approach to automate data preprocessing," *International Journal of Scientific Research in Engineering and Management*, vol. 7, no. 3, 2023, doi: 10.55041/ijserm18270.
- [19] 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*, vol. 34, no. 5, pp. 1785–1792, 2022, doi: 10.1016/j.jksuci.2019.12.011.
- [20] C. Wang, "Efficient customer segmentation in digital marketing using deep learning with swarm intelligence approach," *Information Processing and Management*, vol. 59, no. 6, 2022, doi: 10.1016/j.ipm.2022.103085.
- [21] A. J. Christy, A. Umamakeswari, L. Priyatharsini, and A. Neyaa, "RFM ranking – an effective approach to customer segmentation," *Journal of King Saud University - Computer and Information Sciences*, vol. 33, no. 10, pp. 1251–1257, 2021, doi: 10.1016/j.jksuci.2018.09.004.




- [22] A. Ullah *et al.*, "Customer analysis using machine learning-based classification algorithms for effective segmentation using recency, frequency, monetary, and time," *Sensors*, vol. 23, no. 6, 2023, doi: 10.3390/s23063180.
- [23] Y. Sun, H. Liu, and Y. Gao, "Research on customer lifetime value based on machine learning algorithms and customer relationship management analysis model," *Heliyon*, vol. 9, no. 2, 2023, doi: 10.1016/j.heliyon.2023.e13384.
- [24] H. Zuo *et al.*, "A data-driven customer profiling method for offline retailers," *Computational Intelligence and Neuroscience*, vol. 2022, 2022, doi: 10.1155/2022/8069007.
- [25] Y. Li, X. Chu, D. Tian, J. Feng, and W. Mu, "Customer segmentation using K-means clustering and the adaptive particle swarm optimization algorithm," *Applied Soft Computing*, vol. 113, 2021, doi: 10.1016/j.asoc.2021.107924.
- [26] M. Hosseini, N. Abdolvand, and S. R. Harandi, "Two-dimensional analysis of customer behavior in traditional and electronic banking," *Digital Business*, vol. 2, no. 2, 2022, doi: 10.1016/j.digbus.2022.100030.
- [27] N. Iofrida *et al.*, "Italians' behavior when dining out: Main drivers for restaurant selection and customers segmentation," *International Journal of Gastronomy and Food Science*, vol. 28, 2022, doi: 10.1016/j.ijgfs.2022.100518.
- [28] C. Kuesten, J. Dang, M. Nakagawa, J. Bi, and H. L. Meiselman, "Japanese consumer segmentation based on general self-efficacy psychographics data collected in a phytonutrient supplement study: Influence on health behaviors, well-being, product involvement and liking," *Food Quality and Preference*, vol. 99, 2022, doi: 10.1016/j.foodqual.2022.104545.
- [29] M. J. J. Gumasing, Y. T. Prasetyo, A. K. S. Ong, S. F. Persada, and R. Nadlifatin, "Factors influencing the perceived usability of wearable chair exoskeleton with market segmentation: a structural equation modeling and K-means clustering approach," *International Journal of Industrial Ergonomics*, vol. 93, 2023, doi: 10.1016/j.ergon.2022.103401.
- [30] M. Seyedan, F. Mafakheri, and C. Wang, "Cluster-based demand forecasting using Bayesian model averaging: an ensemble learning approach," *Decision Analytics Journal*, vol. 3, 2022, doi: 10.1016/j.dajour.2022.100033.

BIOGRAPHIES OF AUTHORS






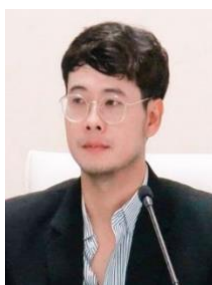
Siriwan Kajornkasirat    received the Ph.D. degree in Computational Science from Walailak University, Thailand, in 2011. She has participated in Ph.D. research experience in Deakin University, Australia funded by the Royal Golden Jubilee Ph.D. Program (RGJ-Ph.D. Program). In 2014, she was invited for STEM Education workshop under the International Visitor Leadership Program (IVLP). This is a program of the U.S. Department of State with funding provided by the U.S. Government. Currently, she is Assistant Professor at the Faculty of Science and Industrial Technology, Prince of Songkla University, Surat Thani, Thailand. Her research interests include data science, computing science, advanced analytics online, STEM education, smart farming, internet of things (IoT), smart health, digital marketing, e-marketing for tourism. She can be contacted at email: siriwan.wo@psu.ac.th.






Pattarawan Gunglin    received the Master of Science Program in Applied Mathematics and Computing Science, Faculty of Science and Industrial Technology, Prince of Songkla University, Surat Thani Campus, Thailand. Currently, she is a data scientist and her current research interests include data science and machine learning. She can be contacted at email: 6540320303@psu.ac.th.



Kritsada Puangsuwan    received the Ph.D. degree in computer engineering from Prince of Songkla University (PSU), Songkhla, in 2017. He joined the Department of Electrical Engineering, Faculty of Engineering, Rajamangala University of Technology Srivijaya, Songkhla, in 2016, as a lecturer. In 2020, he joined the Faculty of Science and Industrial Technology, Prince of Songkla University, Surat Thani Campus, Surat Thani, Thailand. His current research interests include electronics and computer technology for agriculture, microwave and high frequency heating, internet of things (IoT), image processing, and renewable energy. He can be contacted at email: kritsada.pu@psu.ac.th.



Nawapon Kaewsuwan    received the Ph.D. degree in information studies from Khon Kaen University, Thailand, in 2019. Currently, he is Asst. Prof. at the Department of Information Management, Faculty of Humanities and Social Sciences, Prince of Songkla University, Pattani Campus. His research interests include information science, information and knowledge management, technology administration, and educational technology adoption and utilization. He can be contacted at email: nawapon.k@psu.ac.th.