# J-WS: A Hybrid Unsupervised Mining Approach for Customer Segmentation in B2C e-Commerce

Muhammad Azry Bin Ali, Fang-Fang Chua(✉), and Amy Hui Lan Lim

Faculty of Computing and Informatics, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Selangor, Malaysia
ffchua@mmu.edu.my

**Abstract.** Nowadays, with the use of technology and the Internet, it is easy to start a business, more specifically an e-commerce business. However, maintaining a consistent sale and having returning customers can prove a challenge as most businesses rely on new customers for profits and does not generate a reliable profit as compared to relying on old customers. One might resort to applying different kinds of marketing strategies but without understanding of their customer base and proper segmentation of customers, these efforts could result in waste of resources and low probability of success. Therefore, an approach named J-WS that can perform customer segmentation based on customer sales data and Recency, Frequency, and Monetary (RFM) model is proposed. Meaningful information from different groups of customers can later be utilized by target marketing strategy to improve customer retention and impactful marketing. The proposed work consists of 5 phases which include data cleaning, identifying the best clustering algorithm between K-Means and Hierarchical clustering in terms of execution time and Sum of Squared Error, applying association rule mining to generate sets of frequent association rules among the clusters. Conclusively, J-WS can be used by e-commerce businesses to segment their customers meaningfully and properly.

**Keywords:** Association rule mining · Customer segmentation · E-commerce · Hierarchical · K-Means · RFM model

## 1 Introduction

Due to the ease of creating an online business, a new form of retailing called E-commerce has been rising rapidly over the years. With the ease of creating an online shop, it is also much easier for consumers to shop online which contributes to an on-growing customer base. With that, businesses need to be able to handle and manage these new customers and retaining old ones too. Traditional method of business without proper planning would direct them to carry out non cost effective and subpar marketing plan targeting to all customers. CRM is encompassing of technologies, practices, and strategies used to manage and analyse customers with the purpose of improving customer relationship and thus helps in customer retention. Retaining current customers rather than relying


S.-C. Haw and K. Sonai Muthu (Eds.): CITIC 2022, 10, pp. 156–170, 2022.
https://doi.org/10.2991/978-94-6463-094-7_13

on new customers for revenue can be proven to be more beneficial. According to [1], the revenue of a company largely contributed by 20% of their customers compared to other revenue sources. CRM allows businesses to have customers' information such as purchases history and other data be accessed easily which then helps identifying insights and groups about their customers. Segmentation can be categorized into several categories whereby one of them is Behavioural Segmentation, which groups customers based on behaviours such as products purchased, when was it purchased and how much product bought alongside the proper use of machine learning algorithms and model which the CRM tool provides. With these segmented group of customers, businesses can then focus more directly on profitable customers and carry an appropriate marketing technique accordingly.

However, CRM is a software tool that require hefty subscription and is not suitable or justifiable for Small Medium Enterprises. Therefore, an approach namely J-WS is proposed whereby the idea of any size of a company can utilize this approach without spending a huge fee on a software but with free software and basic knowledge of programming. However, unlike a pre-build tool that suggest a machine learning algorithm, The proposed approach implore comparing between two know algorithms to justify which algorithm is more suitable for business use case.

The proposed approach takes the business's customers and divides them into groups based on different attributes using different data mining approaches and applying target marketing strategies to keep and maintain loyal customers. Segmenting the customers is not the main issue to be addressed and instead, identifying the clustering algorithm to be applied and creating a detailed approach to produce clusters of customers that are meaningful will be the focus of our proposed work.

For this paper, the chosen domain will be business-to-consumer (B2C) online retail businesses and behavioural attributes will be used to produce clusters that are insightful. Data that will be analysed will be the company's customer sales data. In terms of segmenting customers using behavioural attributes, the use of RFM models is quite popular in current trends.

In identifying the clustering algorithm to produce meaningful customer segmentation clusters, data will be analysed using two different clustering algorithms which are K-Means and Hierarchical clustering. The segmentation output of these two algorithms will be clusters. Apriori algorithm, an association rule mining algorithm will be used to identify the characteristics in each cluster. To determine which clustering algorithm is better, few criteria will be taken in consideration such as execution time, Sum of Square Error (SSE), and its meaningfulness. The expected output of this proposed work is an approach that can perform customer segmentation on E-commerce business so that target marketing can be applied and thus retain customer loyalty.

## 2 Literature Review

This section elaborates the recent works related to customer segmentation. The researchers [2] segment customers using customer transaction sale data from a beauty clinic shop by fitting the data into the RFM model. The researchers propose a modification of K-Means which is Repetitive Median K-Means that replaces the randomly

assigned initial centroids with the median values that are iteratively calculated from sorted R, F and M in ascending order. The Repetitive Median K-Means is benchmarked with traditional K-Means and Fuzzy C-Means in terms of iterations, compactness and execution time. The proposed Repetitive Median K-Means outperforms the other two models in its performance to segment customers.

In another paper, the researchers [3] perform customer segmentation on a credit provider business to promote a way to retain customers rather than getting new customers by identifying the profitable customers. The researchers propose the use of RFM model and K-Means clustering on Nine Reload Credit Server. K-Means clustering in RapidMiner 5.2 is applied and two clusters are formed with one cluster consisting of 63 customers and cluster two consisting of 39 members. The researchers conclude that Cluster 1 has a higher RFM average value and therefore suggest that marketing to be targeted on customers in Cluster 1.

Moreover, in another paper [4], the researchers argue that using machine learning hierarchical agglomerative clustering (HAC), will produce appropriate segmented customers. The motivation is to segment the customers without relying only on demographics data. By using HAC on a dataset that records the transactions belonging to 9000 credit card holders, the researchers can identify three clusters with Cluster 3 being the targeted customers. However, the researchers stated that this procedure is slow and hardware dependent. The entire experiment is written in the R programming language.

Another group of researchers [5] are proposing a model to identify profitable customers by applying RFM model and K-Means clustering on customers' sales transactions belonging to a small and medium-sized enterprise (SME) named PD Karya Mulya. The entire research experiment is implemented step-by-step using Cross Industry Standard Process for Data Mining (CRISP-DM). The customers are clustered as to whether they are everyday shoppers, dormant or golden customers. The analysis reveals that PD Karya Mulya should focus on golden customers.

In another related work, the researchers [6] apply the RFM model with HAC and association rules on transactional data belonging to an actual Taiwan-based apparel store. The total transactions are 49,040 transactions. The objective of the experiment is to rank the customers into one of the following four categories which are Champions, Loyalty Customers, General Customers and Indifferent Customers. The dataset is split into 80/20 and 5-fold cross validation is applied. The researchers have recorded high accuracy in their proposed model by measuring the framework using F1 scores, Mean Absolute Error (MAE) and average MAE. The proposed framework has been practically applied to the actual apparel store. The researchers have highlighted a few directions for future work and one of the researchers' suggestions is to incorporate different sources and format of data for analysis.

As part of marketing strategy, the researchers [7] apply a combination of RFM model with K-means on XYZ online bookstore that has a company in Jakarta, Indonesia [9] to determine three categories of customers which are loyal customers, new customers and lost customers. When these three types of customers are mapped to the customer value matrix, the authors conclude that loyal customers belong to the "best" customer group while new customers and lost customers belong to "uncertain" customer groups. The authors have reported that inclusion of socio-demographic data and more data description

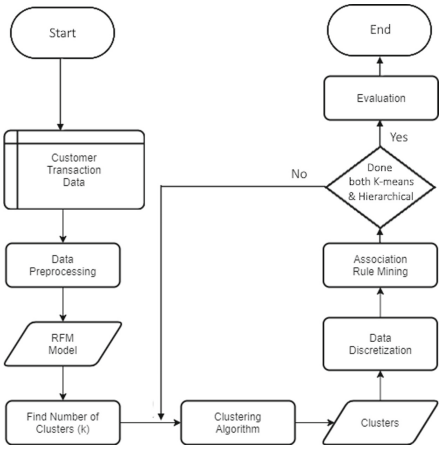



**Fig. 1.** Research Methodology.

about the transaction can help to improve the framework in providing insights which can help to develop a more specific marketing strategy.

Lastly, the researchers [8] have applied machine learning approach on synthetic data that is generated based on the model that has been verified by a telecom company in Malaysia. Principal Component Analysis (PCA) is used to pre-process the data and the Elbow method is used to determine the optimal to cluster the synthetic data.

The result is presented at INSIGHT, a dashboard that projects the customer segments based on demographic, behavioural, and regional traits.

After reviewing existing methods in recent papers, the most common clustering algorithm used is K-Means clustering to segment customers. The use of RFM models to select appropriate attributes are also commonly used as this model provides an easy way to segment customers based on customer behaviour. However, there is a lack of published work on the use of hierarchical clustering as benchmarking with K-Means to segment customers. Therefore, the proposed approach, J-WS will use the RFM model followed by clustering approach to segment the customers. In the clustering approach, both hierarchical and K-Means will be compared to identify the best method for segmenting customers. Finally, association rule mining will be applied to know the characteristics of each cluster that is formed.

## 3 Research Methodology

The research methodology is divided into 5 main phases and has been detailed in Fig. 1.

### 3.1 Phase 1: Data Collection

The data that will be used is obtained from online sources and is a company's customer transaction sale data which have variables that can fit the RFM model.

**Table 1.** RFM model properties

| Attribute | Data type | Description |
|---|---|---|
| Customer ID | Number/String | This attribute is a unique data that each customer identifies. |
| Purchased date | Date | These attributes will be used to calculate the value for both attribute Recency and Frequency. |
| Item price | Number | These attributes will be used to calculate the value for attribute Monetary. |

### 3.1.1 Recency, Frequency, Monetary (RFM)

RFM model is a commonly used model in segmenting customers where the raw customers' transactional sales data are fit into this model for a compact and easier analysis. The premise of the model is differentiating the transactional sales data into 3 distinct customer behaviours, how recently a customer made a purchase, how frequent a made a purchase, and how much a customer has spent. The attributes of the model can be further details as below:

Recency (R), defined as the customer last purchase. The dataset will require a date variable showing the customers purchase date which is then later used to find out the customer's last purchase date for the given time. Frequency (F), defined as the number of purchased made. The date variable will be used. The total frequency of a customer's purchases will be counted for the given time. Monetary (M), defined as accumulated money spent. The dataset will require a variable showing the amount paid by the customers for each of their transactions. The amount paid will be accumulated for each of the customers for the given time.

According to [9], businesses commonly use this RFM model in combination with their data to analyse the customers behaviour in terms of purchase history to identify profitable customers. To use this model, the raw data must have attributes that include unique customer ID, date, and amount paid. Table 1 describes the relevant attributes in the dataset that fits the RFM model's requirements.

## 3.2 Phase 2: Data Pre-processing

The need for data pre-processing is common, generally due to the nature of real-world data, which is incomplete, noisy, or inconsistent. In this paper, data cleaning, data transformation, and data discretization will be applied. Some data cleaning activities involve filling in missing values, identifying, or removing outliers, fixing inconsistencies, and smooth noisy data. During this stage too, appropriate attributes will be chosen to fit into the RFM model which include date variable, price variable, and unique customer ID variable. Data transformation involves two different kinds of transformations which are normalization and aggregation. For aggregation, unique customer in the dataset will need to be fit into the RFM model and combined with a unique customer ID. This new dataset will then need to be normalized. Min-max normalization is chosen. It transforms the data into values ranging from 0 to 1 where the minimum value is 0, the maximum

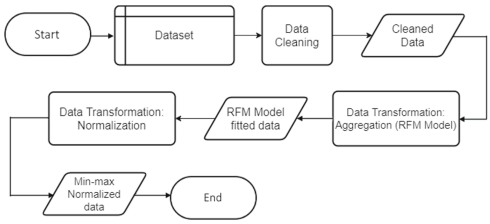



**Fig. 2.** Phase 2 process flow.

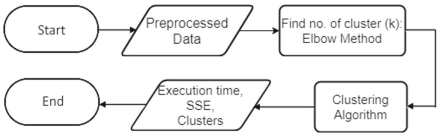



**Fig. 3.** Phase 3 process flow.

value is 1, and every other value is a decimal between 0 and 1. The process flow for Phase 2 is depicted in Fig. 2.

### 3.3 Phase 3: Customer Segmentation

In this phase, the segmentation of customers will be done using clustering algorithms namely K-Means and Hierarchical clustering and their performance will be analysed. Elbow method will be used to identify the appropriate number of clusters and Euclidean distance is used to measure distance. After each clustering, the execution time and sum of squared error (SSE) need to be observed as part of segmentation evaluation.

K-Means clustering is an unsupervised learning used for unlabelled data. The use of this algorithm is to find clusters with similar features with the number of clusters represented by variable K. It will iteratively assign each data point to one of the K clusters centroids until there are no changes. Similar to K-means clustering, hierarchical clustering groups data with similar characteristics. In terms of hierarchical clustering however, the process of clustering goes by treating each data point as its own cluster which is then repeatedly followed by these two steps. First, identify two of the closest clusters and second, merge the two clusters until all clusters are merged. The output of this algorithm is called a dendrogram. The process flow for Phase 3 is depicted in Fig. 3.

### 3.4 Phase 4: Association Rule Mining (Apriori)

After clustering is done and the clusters are obtained, association rule mining can be applied to these clusters to better understand the patterns or characteristics among the clusters and establish the meaningfulness of the clusters. Association rule mining are if-then statements that can help to uncover relationships among items within the transactional datasets [10] with Apriori algorithm being the first and still the commonly used algorithm in association rule mining. The association rules are measured according to the following metrics:

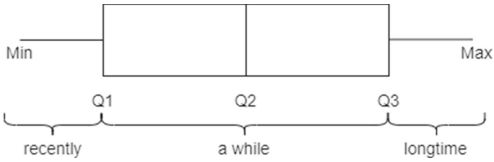



**Fig. 4.** The boxplot and bins' labels for Recency attribute.


**Fig. 5.** The boxplot and bins' labels for Frequency attribute.

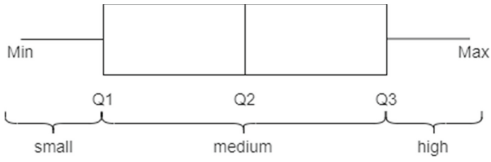



**Fig. 6.** The boxplot and bins' labels for Monetary attribute.

- Support: Indicates how frequently the itemset appears in the transactional dataset. The higher the support, the more frequent the itemset occur in the transactional dataset.
- Confidence: Indicates the conditional probability which is probability that the itemset as stated in the consequent part of the rule occur given the occurrence of itemset as stated in the antecedent part of the rule.
- Lift: Indicates the degree of relations or dependency among the itemset. If the value of lift is $>1$, the items in the itemset are likely to occur together, thus potentially useful while the value of lift $<1$ are the opposite. The value of lift $= 1$ indicates that there is no association among the itemset in the association rule.

However, to apply association rule mining, the data needs to be discretized which means changing the data from numerical to nominal.

### 3.4.1 Data Discretization

A method called binning using boxplot will be used where the continuous values are grouped into bins labelled as shown below each boxplot. Since this research uses the RFM model, three different bins are needed as depicted in Figs. 4, 5, and 6.

After the data of each cluster have been discretized, then association rule mining can be applied. For this research, the minimum Support is set at 10%, Confidence is set at 60%, and Lift value is more than 1. The result of this phase is a list of frequent itemsets within the given constraints that explains the characteristics of the cluster which then

the cluster can be categorized as valuable customers or not for the purpose of cluster meaningfulness.

### 3.5 Phase 5: Evaluation

After applying association rule mining on all the clusters for both K-means and Hierarchical clustering algorithms, evaluation will be done on the clustering algorithms' performance. To compare between the two clustering algorithms, several components are taken into consideration which are execution time, how long does each algorithm take to segment the customers. Sum of Squared Error (SSE), the sum of the squared differences between each observation and its group's mean where the closer the value to 0, the higher the probability the data are clustered correctly [11]. Another component is cluster meaningfulness. By applying association rule mining, each cluster can be analysed and categorized to either valuable customers or not. With categorizing the customers, target marketing can be applied in retaining old customers.

## 4 J-WS Approach

Jupyter-WEKA Segmentation or J-WS is named because the approach was developed using the Python programming language on a web-based platform named Jupyter with Waikato Environment for Knowledge Analysis (WEKA), a machine learning software developed at the University of Waikato, New Zealand. Listed below are J-WS in full details in terms of activities and tasks that revolved around the main 5 phases mentioned in research methodology. This is a step-by-step process on how to efficiently and meaningfully segment customers for businesses to use. These 12 activities include processes that are needed to satisfy the project objectives in terms of identifying the better algorithm between K-Means and Hierarchical.

*Activity 1: Data Cleaning*

- Export raw data into CSV file
- Using Python, the codes are written in Jupyter notebook. The codes include apply necessary data cleaning techniques (remove empty values, duplicates, unused columns, reconfigure date column, etc.).

*Activity 2: Transform Data into RFM Model*

- Calculate values for Recency column
- Calculate values for Frequency column
- Calculate values for Monetary column
- Combine columns based on the unique customer ID and then download them.

*Activity 3: Data Normalization*

- In WEKA, load the RFM transformed data
- Choose normalize attribute under unsupervised filters
- Select class attribute "CustomerID" and check "ignore class: False"
- Download the CSV file.

*Activity 4: Identify no. Of k*

- Import data into Jupyter notebook (RFM Normalized CSV)
- Fit the data into K-Mean with "n_clusters = k" specified
- Plot the result and select k using elbow method.

*Activity 5: K-Mean Clustering*

- In WEKA, load the RFM Normalized data
- Under cluster tab, choose "SimpleKMeans" and set the number of clusters
- Select the unique customer ID attribute to ignore
- Click start and the execution time and SSE are displayed.

*Activity 6: Download K-Mean Clusters*

- In Pre-process tab, choose "AddCluster" under filters, unsupervised and attribute
- In "AddCluster" setting, choose "SimpleKMeans" and set the number of clusters
- Write 'first' in "ignoreAttributeIndices" and select the unique customer ID as class attribute
- Click apply and open 'edit' and click on 'cluster' tab to arrange the data by cluster and save it as CSV.

*Activity 7: Hierarchical Clustering*

- In WEKA, load the RFM Normalized data
- Under cluster tab, choose "HierarchicalClusterer" and set the "linkType" to 'WARD' and the no. of clusters
- Select the unique customer ID attribute to ignore
- Click start and the execution time are displayed.

*Activity 8: Download Hierarchical Clusters*

- In "Preprocess" tab, choose "AddCluster" under filters, unsupervised and attribute
- In "AddCluster" setting, choose "HierarchicalClusterer" and set the "linkType" to 'WARD' and the no. of clusters
- Write 'first' in "ignoreAttributeIndices" and select the unique customer ID as class attribute

- Click apply and open 'edit' and click on 'cluster' tab to arrange the data by cluster and save it as CSV.

*Activity 9: Calculate the Hierarchical SSE*

- In Jupyter notebook, load the Hierarchical clustered data
- Split the data by cluster and calculate the SSE of each cluster
- Add the SSE value of each cluster to identify the Hierarchical SSE.

*Activity 10: Data Discretization*

- In Jupyter notebook, load the RFM Normalized data and apply ".describe()" function to get boxplot value for each R, F, and M attributes (min, max, Q1, Q3)
- Label the value according to the boxplot value for each R, F, and M attributes
- Load K-Mean cluster and Hierarchical cluster data and map the value with the set labels
- Split the data according to clusters and save each cluster as a CSV file.

*Activity 11: Apply Apriori Algorithm*

- In WEKA, load the discretize cluster file one by one
- Under Associate tab, under Apriori setting, set "lowerBoundMinSupport" to 0.1, "metricType" to Confidence, and "minMetric" to 0.6
- Click start and record the frequent item set rules generated and sort it by Lift value in Descending order
- Evaluate the rules by mapping it into a table and color-coded it based on labels.

*Activity 12: Evaluation*

- Identify the most profitable cluster of customers (for target marketing)
- Compare the profitable cluster of K-Mean and Hierarchical by checking their similarities
- Check K-Mean and Hierarchical execution time and SSE to identify the better clustering algorithm.

## 5 Experiment Analysis

Data is obtained from University of California, Irvine (UCI) Machine Learning Repository. It is a transactional data with approximately 541 thousand instances consisting of transaction history from 01/12/2010 to 09/12/2011 for a UK-based and registered non-store online retail that fits the RFM model requirements. Attributes of this data that satisfy the RFM model are 'CustomerID', 'InvoiceData', and 'UnitPrice'. Missing and duplicate values are deleted. For Monetary, for each unique customer ID, the sum of all

| | CustomerID | Recency | Frequency | Monetary |
|---|---|---|---|---|
| 0 | 12346.0 | 326.0 | 2 | 2.08 |
| 1 | 12347.0 | 3.0 | 7 | 481.21 |
| 2 | 12348.0 | 76.0 | 4 | 178.71 |
| 3 | 12349.0 | 19.0 | 1 | 605.10 |
| 4 | 12350.0 | 311.0 | 1 | 65.30 |
| ... | ... | ... | ... | ... |
| 4367 | 18280.0 | 278.0 | 1 | 47.65 |
| 4368 | 18281.0 | 181.0 | 1 | 39.36 |
| 4369 | 18282.0 | 8.0 | 3 | 62.68 |
| 4370 | 18283.0 | 4.0 | 16 | 1174.33 |
| 4371 | 18287.0 | 43.0 | 3 | 104.55 |

4372 rows × 4 columns

**Fig. 7.** Final output of RFM model.

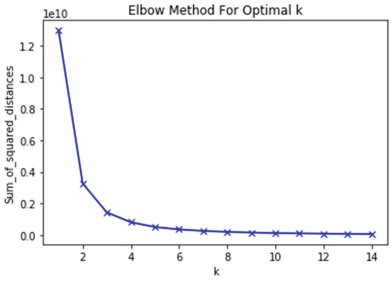



**Fig. 8.** Elbow method graph.

amounts paid by that customer in the given time frame is calculated. For Frequency, the number of invoices attached to each unique customer is calculated. Lastly for Recency, the date purchased is used and subtracted by the latest date which is 09/12/2011. The lowest number was chosen for each unique customer ID to symbolize how many days since they last made purchases. After the final output of the RFM model is obtained (as shown in Fig. 7), it will be normalized. To be able to do clustering, the number of k needed to be identified first. Using the Elbow Method, the number of k is set to 3 which is shown in Fig. 8.

As shown in Fig. 9, setting the distance as Euclidean Distance and number of k as 3, the execution time for K-Means is 0.17 s with Sum of Squared Error of 41.11. The cluster results are then saved as a CSV file.

Like K-Means, setting the distance as Euclidean Distance, number of k as 3, and link the type to WARD, the execution time for Hierarchical is 130.1 with Sum of Squared Error of 49.03 (as shown in Fig. 10). The calculation of Hierarchical clustering's SSE is computed via Jupyter notebook. The cluster results are then saved as a CSV file.

Next, the data will need to be discretized. To do this, normalized RFM data is loaded in Jupyter notebook and 'describe()' function is applied to identify the min, max, Q1, Q2, and Q3 value of each R, F, and M column. A bin and label list are also declared for easier mapping. The dataset will then be discretized using the bins and labels for

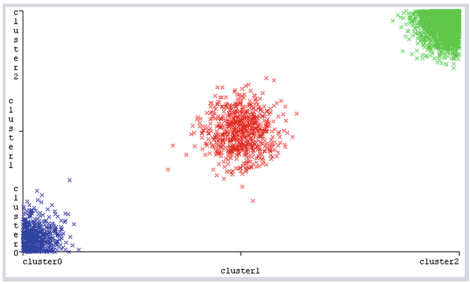



**Fig. 9.** K-means clusters

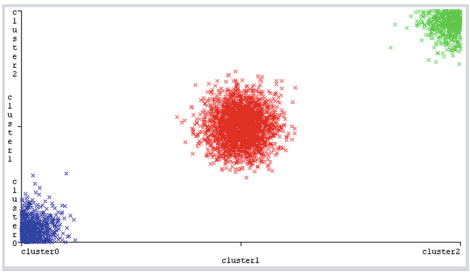



**Fig. 10.** Hierarchical clusters

K-means clustered data and Hierarchical clustered data saved earlier and be saved as a new CSV file. They are later then separated into their respective clusters for K-means and Hierarchical resulting in 6 different CSV files. Association Rule mining will be done 6 times, one for each cluster file. After running, a set of 10 association rules will be generated to show the common association within the cluster for each of the files. These are examples of association rules sorted by Lift:

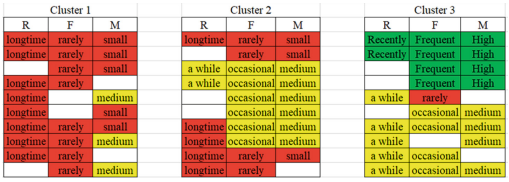
Monetary_Discretized = small 337 = = > Frequency_Discretized = rarely 274 < conf:(0.81) > lift:(1.21) lev:(0.07) [46] conv:(1.72).

Recency_Discretized = longtime Monetary_Discretized = small 337 = = > Frequency_Discretized = rarely 274 < conf:(0.81) > lift:(1.21) lev:(0.07) [46] conv:(1.72).

Monetary_Discretized = small 337 = = > Recency_Discretized = longtime Frequency_Discretized = rarely 274 < conf:(0.81) > lift:(1.21) lev:(0.07) [46] conv:(1.72).

The set of rules generated by these 6 clusters will be put into tabular format and highlighted to better understand the clusters. For context, red indicates undesirable, yellow indicates average, and green indicates good.

Figure 11 of K-means clustered customers shows customers of Cluster 1 are considered undesirable customers as they are not frequent buyers that spend a small amount of money and the last time these customers bought something; it was a long time ago. Customers of Cluster 2 are considered an average day-to-day customer where these customers occasionally buy something with an average amount of spending. Customer of Cluster 3 however, is considered profitable customers as they frequently buy something



Cluster 1

| R | F | M |
|---|---|---|
| longtime | rarely | small |
| longtime | rarely | small |
| | rarely | small |
| longtime | rarely | |
| longtime | | medium |
| longtime | | small |
| longtime | rarely | small |
| longtime | rarely | medium |
| longtime | rarely | |
| | rarely | medium |

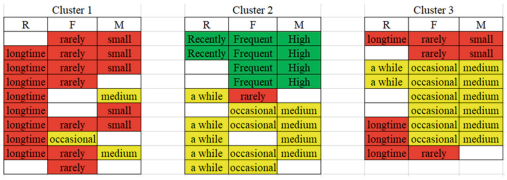
Cluster 2

| R | F | M |
|---|---|---|
| longtime | rarely | small |
| | rarely | small |
| a while | occasional | medium |
| a while | occasional | medium |
| | occasional | medium |
| | occasional | medium |
| longtime | occasional | medium |
| longtime | occasional | medium |
| longtime | rarely | small |
| longtime | rarely | |

Cluster 3

| R | F | M |
|---|---|---|
| Recently | Frequent | High |
| Recently | Frequent | High |
| | Frequent | High |
| | Frequent | High |
| a while | rarely | |
| | occasional | medium |
| a while | occasional | medium |
| a while | | medium |
| a while | occasional | |
| a while | occasional | medium |

**Fig. 11.** Result of K-means clustered customers.



Cluster 1

| R | F | M |
|---|---|---|
| | rarely | small |
| longtime | rarely | small |
| longtime | rarely | small |
| longtime | rarely | |
| longtime | | medium |
| longtime | | small |
| longtime | rarely | small |
| longtime | occasional | |
| longtime | rarely | medium |
| | rarely | |

Cluster 2

| R | F | M |
|---|---|---|
| Recently | Frequent | High |
| Recently | Frequent | High |
| | Frequent | High |
| | Frequent | High |
| a while | rarely | |
| | occasional | medium |
| a while | occasional | medium |
| a while | | medium |
| a while | occasional | medium |
| a while | occasional | |

Cluster 3

| R | F | M |
|---|---|---|
| longtime | rarely | small |
| | rarely | small |
| a while | occasional | medium |
| a while | occasional | medium |
| | occasional | medium |
| | occasional | medium |
| longtime | occasional | medium |
| longtime | occasional | medium |
| longtime | rarely | |

**Fig. 12.** Result of Hierarchical clustered customers.

with a high amount of spending and the last time these customers bought something, it was quite recently.

Figure 12 of clusters from Hierarchical algorithm shows that customers of Cluster 1 are considered undesirable customers as they are not frequent buyers that spend a small amount of money and the last time these customers bought something; it was a long time ago. Customers of Cluster 2 is considered as profitable customers as they frequently buy something with a high amount of spending and the last time these customers bought something, it was quite recently. Lastly, customers of Cluster 3 are considered an average day-to-day customer where these customers occasionally buy something with an average amount of spending.

With the resulted clusters, it can be concluded that proper customer segmentation has been achieved and therefore can be deduced that for clustering using K-means, customer segmentation of Cluster 3 is the one business should focus on while clustering using Hierarchical, it is customer segmentation of Cluster 2.

After association rule mining has been applied and thus the profitable cluster of customers can be identified for K-Means and Hierarchical clustering, it can then be compared. To help visualize this, the chosen clusters for maximum profitability and to apply target marketing to each algorithm are checked for similarity in Jupyter notebook by comparing if the same customer ID appears in both clusters. The result shows the chosen cluster has a 98.62% similarity. This means both K-Means and Hierarchical clustering segment the customers just about the same as shown in Fig. 11. Next, their execution time and SSE are checked as a final comparison to determine the better algorithm. For K-Means, the execution time is 0.23 s. with an SSE of 41.097 and for Hierarchical, the execution time is 113.03 s. with an SSE of 42.785. Thus, proving the K-Means algorithm is the better clustering algorithm in terms of speed and accuracy (Fig. 13).

```
K-Means cluster 3 length:  2927
Hierarchical cluster 2 length:  3009

The total difference between clusters are:  82
The total similarity between clusters are:  5854

Similarity between clusters in percentage:  98.61859838274933
```

**Fig. 13.** Similarity in percentage between the two clustering algorithms.

## 6 Conclusion

In conclusion, there are two main objectives that this research is trying to achieve in segmenting customers of E-commerce businesses which are to apply target marketing efficiently while retaining old customers. There is a lack of existing work on the use of hierarchical clustering as benchmarking with K-Means to segment customers. Therefore, the proposed approach, J-WS will use the RFM model followed by clustering approach to segment the customers. In the clustering approach, both hierarchical and K-Means will be compared to identify the best method for segmenting customers. Finally, association rule mining will be applied to know the characteristics of each cluster that is formed. J-WS is able to identify a group of customers that were more profitable to invest in thus making them a target for marketing. The proposed work shows that using tools that are easily accessible like WEKA and Jupyter notebook with understandable programming, any businesses can segment their customer base easily and meaningfully. As of now, J-WS rely mostly on the RFM model to segment the customers which is a behavioural type of segmentation. For future work, the proposed approach will be improved to segment customers using a combination of both behavioural and demographic segmentation in improving the quality of the segmented customers.

## References

1. Christy, A. J., Umamakeswari, A., Priyatharsini, L., & Neyaa, A. (2018). RFM ranking – An effective approach to customer segmentation. Journal of King Saud University - Computer and Information Sciences. https://doi.org/10.1016/j.jksuci.2018.09.004
2. Bachtiar, F. A. (2018). Customer Segmentation Using Two-Step Mining Method Based on RFM Model. 3rd International Conference on Sustainable Information Engineering and Technology, SIET 2018 - Proceedings, 10–15. https://doi.org/10.1109/SIET.2018.8693173
3. Maryani, I., Riana, D., Astuti, R. D., Ishaq, A., Sutrisno, & Pratama, E. A. (2018). Customer segmentation based on RFM model and clustering techniques with K-means algorithm. Proceedings of the 3rd International Conference on Informatics and Computing, ICIC 2018, 1–6. https://doi.org/10.1109/IAC.2018.8780570
4. Hung, P. D., Thuy Lien, N. T., & Ngoc, N. D. (2019). Customer segmentation using hierarchical agglomerative clustering. ACM International Conference Proceeding Series, Part F1483, 33–37. https://doi.org/10.1145/3322645.3322677
5. Dedi, Dzulhaq, M. I., Sari, K. W., Ramdhan, S., Tullah, R., & Sutarman. (2019). Customer Segmentation Based on RFM Value Using K-Means Algorithm. Proceedings of 2019 4th International Conference on Informatics and Computing, ICIC 2019. https://doi.org/10.1109/ICIC47613.2019.8985726
6. Chou, T., & Chang, S. (2022). The RFM Model Analysis for VIP Customer: A Case Study of Golf Clothing Brand. International Journal of Knowledge Management (IJKM), 18(1), 1-18. https://doi.org/10.4018/IJKM.290025

7. Juhari, T., & Juarna, A. (2022). Implementation RFM Analysis Model for Customer Segmentation using the K-Means algorithm Case Study XYZ Online Bookstore. EXPLORE, 12(1), 107–118. https://doi.org/10.35200/explore.v12i1.548
8. Sharaf Addin, E. H., Admodisastro, N., Mohd Ashri, S. N. S., Kamaruddin, A., & Chong, Y. C. (2022). Customer Mobile Behavioral Segmentation and Analysis in Telecom Using Machine Learning. Applied Artificial Intelligence, 1-21. https://doi.org/10.1080/08839514.2021.2009223
9. Huang, Y., Zhang, M., & He, Y. (2020). Research on improved RFM customer segmentation model based on K-Means algorithm. Proceedings - 2020 5th International Conference on Computational Intelligence and Applications, ICCIA 2020, 24–27. https://doi.org/10.1109/ICCIA49625.2020.00012
10. Rai, A. (2018, June 4). An Overview of Association Rule Mining and its Applications. UpGrad Blog. https://www.upgrad.com/blog/association-rule-mining-an-overview-and-its-applications/
11. Nainggolan, R., Perangin-angin, R., Simarmata, E., & Tarigan, A. F. (2019). Improved the performance of the K-means cluster using the sum of squared error (SSE) optimized by using the Elbow method. Journal of Physics: Conference Series, 1361, 012015. https://doi.org/10.1088/1742-6596/1361/1/012015