Machines and Algorithms
http://www.knovell.org/mna

*Research Article*

# Market Basket Data-Mining Analysis

**Sheikh Abdul Hannan**[1, *]

[1] Department of Computer Science and Information Technology, Virtual University, Lahore, 54000, Pakistan
[*]Corresponding Author: Sheikh Abdul Hannan. Email: ms090400008@vu.edu.pk

Received: 8 August 2024; Revised: 02 September 2024; Accepted: 01 October 2024; Published: 10 October 2024
AID: 003-03-000044

**Abstract:** The Market Basket analysis is the key factor for customer-centric marketing in this era. It strongly requires the data mining techniques on massive sales transaction data. The aim in this paper is to study and find the different data mining solutions for large and sparse sales transaction data. Here a real-world data set has been summarized and analyzed. In this paper the problems related to Association rule mining (ARM) on large and sparse data has been discussed. It has also shown that the application of association rule mining on sparse data is not easy if implemented directly, so there is a significant need to find some other mining technique and solutions like k-means clustering in this paper, to pre-process the data for ARM. Recency, Frequency and Monetary (RFM) model has been discussed and implemented in detail, so that K-Means algorithm can be applied easily. Additionally, this analysis will be helpful in the future research horizons like multi label classification of temporal data set and sequence to sequence neural network implementation for prediction.

**Keywords:** Customer-Centric Marketing; Online Retail; Data Mining Rules; *K*-Means Clustering; Apriori Algorithm;

## 1. Introduction

Businesses often utilize market basket analysis, which is a common analytical tool to better grasp buying behavior by spotting items that are routinely purchased in concert. It shapes tactics such targeted marketing, cross-selling, and product placement quite significantly. Companies use data-driven approaches that examine enormous amounts of transactional data in order to get such insights. Data mining is among the most often used techniques available for this aim. Data mining has evolved into a vital instrument for sales transaction analysis and pattern recognition in the competitive market of today. Particularly with the increased emphasis on customer-centric marketing and temporal buying patterns, big businesses are using several data-mining approaches to generate useful knowledge. Data-mining is the most wanted technique, now market is using for their sales transaction's analysis. Focusing on customer-centric marketing using temporal data, different data-mining techniques are being adopted by large scale companies.

There are three major data-mining techniques: Association, Classification and Clustering. Each of these categories have a range of algorithms for the analysis of marketing trends, but all of them are not applicable for every situation. So, it is pivotal to properly determine relevant algorithm on the basis of given scenario. Swee [1] mentioned that association rule mining is not necessarily the best strategy for analysing large market-basket temporal data.

Implementation of data-mining evolved multiple technologies and tool such as data management, data warehousing, machine and statistical analysis [2]. Association rule mining or affinity analysis is one of the

most commonly applied approach to discover the relationships among transactions. It helps to find the item-to-item relationship. In case of market baskets, it can be used to get frequent sales patterns. Association rule mining identifies all the rules in the database according to the predefined parameters like minimum support and minimum confidence factor etc. The most common algorithm is Apriori. Apriori algorithm is a classical algorithm, used for mining the frequent patterns and association rules in datasets. It is widely used in Market Basket analysis and healthcare sectors. It produces the association rules according to the minimum support and confidence, the pre-defined parameters. Apriori is built on the breadth-first search algorithm and a Hash tree data structure. It creates candidate item sets of length k from item sets with length k-1. Then it eliminates candidates with an infrequent sub pattern. According to the downward closure lemma, the candidate set includes all common k-length item sets. Following that, it analyses the transaction database to identify common item sets among the candidates. There are a lot of modifications in Apriori as per different requirements, as it is already mentioned that Association is not always suitable for large and sparse data, so there are different techniques to implement Association in this scenario. Apriori-Tid is one example [2], another approach is to integrate association rule mining with classification rule mining for this purpose.[3]

Classification rule mining is used to find a small set of rules in the data using an appropriate classification algorithm. In classification rule mining, there is only one pre-determined target we have. This target is known as the class. Same as classification, we have the clustering mining techniques. In this approach, data will be summarized and analysed in groups on basis of different models. Clustering rule mining is highly adopted for sparse and large data i.e. market basket time series data. In this paper, we have implemented clustering-based rule mining to find out different clusters in data using RFM model. The most common algorithm is, k-means clustering algorithm (which is being used for analysis).

K-Means is an algorithm which partitioned the data in simple clusters using K-group technique, where K is the number of clusters as given by the user. The K-Means method assigns each entity to its nearest cluster. When the value of K is unknown in advance, it is important to construct several clustering solutions with different values of K. Cluster quality measurements may be used to determine which clustering solution better represents the actual clustering pattern in the data. The Silhouette coefficient is one of the most popular measurements. This is an excellent indication of cluster quality since it provides an objective assessment of cluster coherence and separation in the clustering solution.[1]

These all approaches and techniques will help in the next research areas, where is the need to predict the future behaviour of customers. Different neural network algorithms are being used for these purposes. Long Short-Term Memory (LSTM) models, Recurrent Neural Networks (RNNs) [4] and sequence-to-sequence neural networks are some most common types, used for temporal large datasets.

The rest of the paper is arranged as follows. The next part discusses the background and relevant work details. In the later section, the details about dataset and its preparation have been provided i.e., steps and tasks for data pre-processing and preparation are explained in detail. In the next part, k-means clustering analysis is used to discover the right data clusters. Each cluster is explained and the association rule mining approach is discussed further. The subsequent section, summarizes and concludes the paper along with future research horizons.

## 2. Related Work

As per study of different work done already, referenced in the end, Cumby has prototyped a shopping assistant that predicts the shopping list – comprising of 12 items – for the customer's current trip based on the past 4 instances of behaviour [5]. The hybrid approach comprised of decision tress and linear methods (Perceptron, Winnow and Naïve) which yielded a prediction accuracy of 50% [6]. The linear methods are known to ignore the data sparsity problem inherent in big sets of data, which could be the reason for such a low prediction accuracy. Kooti [7] measured the effects of consumer age, location and gender to classify the consumer behaviour using Bayesian Network Classification. He then uses these measurements to predict the price and time of the next online purchase. Lee [8] extracts the behaviour features of online shoppers – cart usage, source of item access (site itself or external sources), thinking time, putting the item

in cart etc., without considering gender or age. He then builds a model, based on support vector machine (SVM) classifier and radial basic function (RBF), to predict whether a customer will purchase an item or not based on the extracted behaviour features. He concludes that item browsing patterns are the important predictors of the actual purchases. Shangguan [9] proposes to attach an RFID device to all the entities in a physical clothing store. The RFID device is then used to detect the shopping behaviour of customers. The behaviour is defined as frequently viewing the popular clothes, picking up the hot items and excavating correlated items. However, the proposed strategy has the limitation of working in a self-service clothing store where customers are free to pick up and try clothing items as desired. In [10] author proposes a multi-task recurrent neural network learning architecture that predicts the clinical time series for patients. The parameters predicted are either binary (mortality, diagnosis, deteriorating conditions) or regressive (length of stay in hospital). The joint prediction of the four tasks leads to overfitting at different rates, which remains a challenge to address. In the case of diagnosis, the problem exacerbates because a specific patient can have multiple diagnosis, which are not mutually exclusive (multi-label classification). This problem has been tackled in [4] where a time series of 13 variables is given to the same RNN with 2 hidden layers with Long Short-Term Memory (LSTM) hidden neurons. The results – average predicted labels (diagnosis) of 2.281 per patient out of 128 labels – in the designed multi label classified RNN achieves faster training. However, for absent values of time series variables, it is assumed that doctors believed it to be normal and chose not to measure it, thereby filling the void with normal values. However, this ignores the distinction between truly normal and missing measurements.

Though a variety of approaches i.e., from decision trees and Bayesian classifiers to deep learning architectures, there is still a dearth of attention paid to tackling data sparsity and the dynamic character of customer behaviour in large-scale market basket datasets. Many current methods limit their relevance to real-world, sparse transactional data by either depending mostly on demographic characteristics or assuming regularity in data collecting. Furthermore, unexplored in the framework of temporal customer purchasing patterns is the merging of clustering methods with association rule mining. This work intends to close that gap by using k-means algorithm and clustering-based rule mining over the RFM model, so providing a scalable method for exposing latent behavioural patterns in sparse and high-dimensional retail data.

## 3. Dataset and data preparation

The data, considered in this paper, is data of a multiple store business with distributed database system [11]. This firm, founded in 1981, sells all-occasion presents. Initially, this organization conducted business by direct mailing catalogues and took orders over the phone. The firm also opened an internet store only two years ago. The corporation has a wealth of data about its goods and devoted clients from all across the United Kingdom and Europe, as well as a massive amount of data on sales transactions. The firm also markets and sells its items through Amazon.co.uk.

The sale transactions dataset, as shown in Table 1, consists of 8 parameters and includes all transactions from December 2010 to December 2011.

There are 3958 distinct products [StockCode] and 4372 distinct Customers. Total sales invoices over the time span mentioned above, are 22062. Total number of sales transactions in simple excel sheet are 541909 (about half million).

The data is just raw data in excel form, as shown in Figure 1 and it needs to be processed for further analysis. Firstly, the work done on the data such that each row will show one basket containing Invoice Number with all its respective StockCodes, shown in Figure 2.

**Table 1:** Attributes in dataset

| # | Name | Data Type | Description |
|---|---|---|---|
| **1** | InvoiceNo | Number | Invoice number; a 6-digit number |
| **2** | StockCode | Text | Unique ID for products |
| **3** | Description | Text | Description of the product |
| **4** | Quantity | Number | Quantity which sold out |
| **5** | InvoiceDate | Date/Time | Date and Time of transaction made |
| **6** | UnitPrice | Number | Price of the product |
| **7** | CustomerID | Number | Customer who bought the item |
| **8** | Country | Text | Country in which transaction made |

| InvoiceNo | StockCode | Description | Quantity | InvoiceDate | UnitPrice | CustomerID | Country |
|---|---|---|---|---|---|---|---|
| 536365 | 85123A | WHITE HANGING HEART T-LIGHT HOLDER | 6 | 12/1/2010 8:26 | 2.55 | 17850 | United Kingdom |
| 536365 | 71053 | WHITE METAL LANTERN | 6 | 12/1/2010 8:26 | 3.39 | 17850 | United Kingdom |
| 536365 | 84406B | CREAM CUPID HEARTS COAT HANGER | 8 | 12/1/2010 8:26 | 2.75 | 17850 | United Kingdom |
| 536365 | 84029G | KNITTED UNION FLAG HOT WATER BOTTLE | 6 | 12/1/2010 8:26 | 3.39 | 17850 | United Kingdom |
| 536365 | 84029E | RED WOOLLY HOTTIE WHITE HEART. | 6 | 12/1/2010 8:26 | 3.39 | 17850 | United Kingdom |
| 536365 | 22752 | SET 7 BABUSHKA NESTING BOXES | 2 | 12/1/2010 8:26 | 7.65 | 17850 | United Kingdom |
| 536365 | 21730 | GLASS STAR FROSTED T-LIGHT HOLDER | 6 | 12/1/2010 8:26 | 4.25 | 17850 | United Kingdom |
| 536366 | 22633 | HAND WARMER UNION JACK | 6 | 12/1/2010 8:28 | 1.85 | 17850 | United Kingdom |
| 536366 | 22632 | HAND WARMER RED POLKA DOT | 6 | 12/1/2010 8:28 | 1.85 | 17850 | United Kingdom |
| 536367 | 84879 | ASSORTED COLOUR BIRD ORNAMENT | 32 | 12/1/2010 8:34 | 1.69 | 13047 | United Kingdom |
| 536367 | 22745 | POPPY'S PLAYHOUSE BEDROOM | 6 | 12/1/2010 8:34 | 2.1 | 13047 | United Kingdom |
| 536367 | 22748 | POPPY'S PLAYHOUSE KITCHEN | 6 | 12/1/2010 8:34 | 2.1 | 13047 | United Kingdom |
| 536367 | 22749 | FELTCRAFT PRINCESS CHARLOTTE DOLL | 8 | 12/1/2010 8:34 | 3.75 | 13047 | United Kingdom |
| 536367 | 22310 | IVORY KNITTED MUG COSY | 6 | 12/1/2010 8:34 | 1.65 | 13047 | United Kingdom |

**Figure 1:** A simple view of raw data

Further this data has been arranged to implement RFM (Recency, Frequency and Monetary) model on it which is in Figure 3. It is also tried to implement Apriori (using Weka 3.8) on this data. For this purpose, the data should be transformed in True False relations for each Invoice and all Products, as in fig 4. For all these purposes some code and algorithms were implemented so that the data can be generated automatically as per requirement.

| InvoiceNo | Col 0 | Col 1 | Col 2 | Col 3 | Col 4 | Col 5 | Col 6 | Col 7 | Col 8 | Col 9 | Col 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 536370 | 22728 | 22727 | 22726 | 21724 | 21883 | 10002 | 21791 | 21035 | 22326 | 22629 | 22659 |
| 536371 | 22086 | | | | | | | | | | |
| 536372 | 22632 | 22633 | | | | | | | | | |
| 536373 | 85123A | 71053 | 84406B | 20679 | 37370 | 21871 | 21071 | 21068 | 82483 | 82486 | 82482 |
| 536374 | 21258 | | | | | | | | | | |
| 536375 | 85123A | 71053 | 84406B | 20679 | 37370 | 21871 | 21071 | 21068 | 82483 | 82486 | 82482 |
| 536376 | 22114 | 21733 | | | | | | | | | |
| 536377 | 22632 | 22633 | | | | | | | | | |
| 536378 | 22386 | 85099C | 21033 | 20723 | 84997B | 84997C | 21094 | 20725 | 21559 | 22352 | 21212 |
| 536380 | 22961 | | | | | | | | | | |
| 536381 | 22139 | 84854 | 22411 | 82567 | 21672 | 22774 | 22771 | 71270 | 22262 | 22637 | 21934 |
| 536382 | 10002 | 21912 | 21832 | 22411 | 22379 | 22381 | 22798 | 22726 | 22926 | 22839 | 22838 |
| 536384 | 82484 | 84755 | 22464 | 21324 | 22457 | 22469 | 22470 | 22224 | 21340 | 22189 | 22427 |
| 536385 | 22783 | 22961 | 22960 | 22663 | 85049A | 22168 | 22662 | | | | |
| 536386 | 84880 | 85099C | 85099B | | | | | | | | |

**Figure 2:** Converted data for Invoice and its products like a basket

| CustomerID | Recency | Freuen | Monetary | FirstPurchase |
|---|---|---|---|---|
| 12346 | 1/18/2011 10:17:00 AM | 2 | 0 | 1/18/2011 10:01:00 AM |
| 12347 | 12/7/2011 3:52:00 PM | 7 | 4310 | 12/7/2010 2:57:00 PM |
| 12348 | 9/25/2011 1:13:00 PM | 4 | 1797.24 | 12/16/2010 7:09:00 PM |
| 12349 | 11/21/2011 9:51:00 AM | 1 | 1757.55 | 11/21/2011 9:51:00 AM |
| 12350 | 2/2/2011 4:01:00 PM | 1 | 334.4 | 2/2/2011 4:01:00 PM |
| 12352 | 11/3/2011 2:37:00 PM | 11 | 1545.41 | 2/16/2011 12:33:00 PM |
| 12353 | 5/19/2011 5:47:00 PM | 1 | 89 | 5/19/2011 5:47:00 PM |
| 12354 | 4/21/2011 1:11:00 PM | 1 | 1079.4 | 4/21/2011 1:11:00 PM |
| 12355 | 5/9/2011 1:49:00 PM | 1 | 459.4 | 5/9/2011 1:49:00 PM |
| 12356 | 11/17/2011 8:40:00 AM | 3 | 2811.43 | 1/18/2011 9:50:00 AM |
| 12357 | 11/6/2011 4:07:00 PM | 1 | 6207.67 | 11/6/2011 4:07:00 PM |

**Figure 3:** Pre-processed data in RFM model

| InvoiceNo | 21730 | 22752 | 71053 | 84029E | 84029G | 84406B | 85123A | 22632 | 22633 | 21754 | 21755 | 21777 | 22310 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 536365 | t | t | t | t | t | t | t | ? | ? | ? | ? | ? | ? |
| 536366 | ? | ? | ? | ? | ? | ? | ? | t | t | ? | ? | ? | ? |
| 536367 | ? | ? | ? | ? | ? | ? | ? | ? | ? | t | t | t | t |
| 536368 | ? | ? | ? | ? | ? | ? | ? | ? | ? | ? | ? | ? | ? |
| 536369 | ? | ? | ? | ? | ? | ? | ? | ? | ? | ? | ? | ? | ? |
| 536370 | ? | ? | ? | ? | ? | ? | ? | ? | ? | ? | ? | ? | ? |
| 536371 | ? | ? | ? | ? | ? | ? | ? | ? | ? | ? | ? | ? | ? |
| 536372 | ? | ? | ? | ? | ? | ? | ? | t | t | ? | ? | ? | ? |
| 536373 | t | t | t | t | t | t | t | ? | ? | ? | ? | ? | ? |
| 536374 | ? | ? | ? | ? | ? | ? | ? | ? | ? | ? | ? | ? | ? |
| 536375 | t | t | t | t | t | t | t | ? | ? | ? | ? | ? | ? |

**Figure 4:** Data for Apriori implementation in True/False relations for each invoice and all products

## 4. K-Means Clustering Implementation

### *4.1. RFM Modeling and Clustering*

The data has been prepared in RFM model. Now the factors recency, frequency and monetary can be seen for each customer. Recency value tells how recent the customer visited the store and had some purchases. Frequency value shows how often customer visits the store. Monetary value is also important key that tells the value contributed by the customer in business generation. The resultant dataset consists of CustomerID, Recency, Frequency, Monetary and FirstPurchase (Just to calculate the recency value). Shown in Table 2.

**Table 2:** Attributes in resultant dataset

| # | Name | Data Type | Description |
|---|---|---|---|
| **1** | CustomerID | Number | The unique ID of customer |
| **2** | Recency | Date/Time | Most recent date/time of transaction for the respective customer |
| **3** | Frequency | Number | Number of visits of customer in the defined time span |
| **4** | Monetary | Number | Total amount of customer spent during time span |
| **5** | FirstPurchase | Date/Time | Date/Time of customer's first purchase |

K-Means clustering was used to develop a series of clustering solutions with varying cluster counts based on this RFM model. It is simple to implement using the Cluster approach in Weka 3.8. Various Ks (numbers of clusters) have been formed, including K=2, K=3, and K=5. The solution with 5 clusters has a decent score based on the Silhouette coefficient, hence it is chosen for further study.

Using Weka 3.8, the following clusters have been generated for K = 5.

**Table 3:** Statistics of K-Means Algorithm

| Statistics | Value |
|---|---|
| **Instances** | 4732 |
| **Attributes** | 5 |
| **Number of Iteration** | 9 |
| **Sum of Squared Errors (within cluster)** | 9631.591035729285 |
| **Missing Values Replacement** | Global mean / mode |
| **Initial Starting Points (k-means++)** | |
| **Cluster 0** | 13617,'10/30/2011 1:50:00 PM',3,544.18,'6/12/2011 12:55:00 PM', cluster3 |
| **Cluster 1** | 15163,'11/21/2011 1:10:00 PM',2,304.47,'5/31/2011 3:22:00 PM', cluster1 |
| **Cluster 2** | 14868,'12/6/2011 2:49:00 PM',9,2939.64,'4/20/2011 1:36:00 PM', cluster1 |
| **Cluster 3** | 16484,'6/19/2011 12:06:00 PM',3,379.4,'6/2/2011 10:39:00 AM', cluster2 |
| **Cluster 4** | 14451,'10/10/2011 1:38:00 PM',2,662.59,'5/29/2011 12:35:00 PM', cluster3 |

**Table 4:** Final Cluster Centroids

| Attribute | C1 **(890.0)** (20%) | C2 **(446.0)** (10%) | C3 **(431.0)** (10%) | C4 **(1738.0)** (40%) | C5 **(867.0)** (20%) |
|---|---|---|---|---|---|
| **CustomerID** | 12922.564 | 15606.1839 | 15014.5267 | 17092.6191 | 14129.7785 |
| **Recency** | 12/6/2011 9:56:00 AM | 11/24/2011 12:48:00 PM | 1/31/2011 3:27:00 PM | 12/1/2011 1:47:00 PM | 12/2/2011 11:21:00 AM |
| **Frequency** | 5.3202 | 4.2466 | 5.6404 | 4.8205 | 5.301 |
| **Monetary** | 2075.6793 | 1420.2908 | 2021.7693 | 1774.929 | 2148.8492 |
| **FirstPurchase** | 11/28/2011 1:26:00 PM | 1/7/2011 12:44:00 PM | 1/31/2011 1:17:00 PM | 12/6/2010 12:55:00 PM | 4/7/2011 12:04:00 PM |

### *4.2. Clusters Explanation*

Understanding every cluster in details, is crucially required for customer-centric business intelligence. So now examining the dataset and results of K-Mean algorithm using Weka 3.8, it is observed that each cluster have a group of customers with certain features and parameter values.

In Cluster 1, there are 890 Customers involved. It is composed of 20% of the data. It seems the data in this cluster is just in the last quarter, as the first purchase date is 11/28/2011. Frequency is 5.32, which is above average and the monetary 2075.67, is also above average. This cluster will not be considered for further

analysis as the time span of transaction is very small, just 2 months. It means that the customers in this cluster didn't shop too much. The customers in this span may be newly registered customers and started shopping recently.

In Cluster 2, there are 446 customers involved. It is composed of 10% of whole data. It is observed that the data here is for a long time period from 1/7/2011 to 11/24/2011, which is sufficient to be considered. Still there are some factors such as the frequency here is very low which is 4.24 and monetary is also the lowest among all clusters. So, it is not a feasible cluster for further analysis, as the average case is always selected for analysis. This group of customers, is not profitable because of lowest monetary value and the customers also visit the stores seldom.

In Cluster 3, there are 431 customers involved. It is also composed of 10% of total data. It is clearly observed that the data in this cluster is just first quarter of time span. Monetary value 2021.76 is feasible here, as it is above average. Although the frequency 5.64, is the highest frequency than all of rest clusters, but still the time span is very short. It shows that the customers in this group, were very frequent customers and generated a feasible business value, stopped shopping from this store further or might be just occasionally customers, who needs to shop on some occasion for specific time period.

In Cluster 5, there are 867 customers involved. The time period here is a good long span starts from 4/7/2011 to 12/2/2011. Monetary value 2148.8492 is the best. Frequency 5.3 is also above average. On the other hand, the observations for Cluster 4 are also sufficient. In cluster 4, there are maximum number (1738[40% of data]) of customers involved. The period starts from 12/6/2010 until 12/1/2011, it means this cluster contains the transactions from whole time span. Although the frequency 4.82 and the monetary 1774.92, are just average factors as compared to Cluster 5, but still this is the best option to be considered.

The Cluster 4 has average cases and maximum number of customers within maximum time span. Overall, the firm appears to be fairly profitable.

All things considered, the clusters found using K-Means provide insightful analysis of consumer segmentation. For long-term retention plans, Cluster 4 offers a consistent and devoted clientele, perfect for Given its great financial worth, Cluster 5 might be focused on with luxury product offers or loyalty incentives. Cluster 2 consists of low-spending, infrequent consumers suggesting a chance for re-engagement campaigns or promotional offers. Being connected to shorter activity periods, clusters 1 and 3 could represent seasonal or freshly acquired clients and should be kept under close observation for possible retention or turnover. These divisions let marketing decisions and client relationship management be more targeted and successful.

### *4.3. Further Analysis*

As explained above, the cluster 4 is most feasible and diverse cluster among all these 5 identified clusters, as it contains the maximum number of newly registered and old customers within the maximum time span for dataset, having average and suitable frequency and monetary values. It can be further analysed using other data-mining techniques. One technique is Classification rule mining; it can be implemented on this result set and the customers can be classified into sub categories using some classification algorithm. Another approach to implement the decision tree for further analysis [12]. In decision tree method, Customers can be divided in sub clusters using some parameters e.g., on basis of frequency data can be further divided into sub-categories.

Association rule mining algorithm such as Apriori, can be implemented for these customers and the shopping behaviour can be identified. Weka 3.8 conducted the Apriori algorithm at a minimum confidence level of 0.5 and a minimum support threshold of 0.01—that is, 1% of all transactions. These values were selected to provide a balance between rule relevance and computing practicality, therefore enabling the identification of significant yet non-trivial association rules. Apriori can be implemented to find out the relationship between purchased products. It identifies the association rule between products and it can be easily identified that which product was purchased with other product.

As it is known that there are too many issues with association rule mining implementation. For example, the primary issue is the creation of an excessive number of repetitive regulations. While the data mining community has attempted to solve this issue, research is currently continuing. The second problem is about the "interestingness" of regulations. Association rule mining generates a large number of trivial rules that the user already knows.

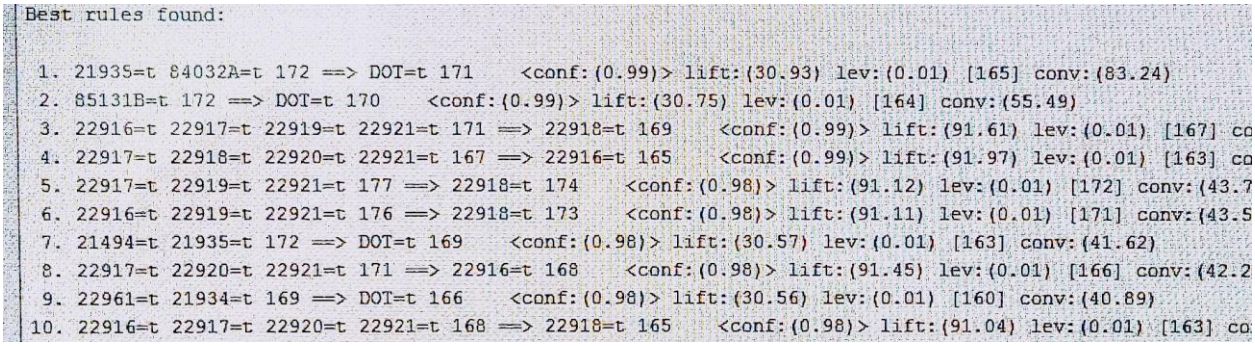


**Figure 5:** Apriori Implementation using Weka 3.8

These rules frequently distract people from recognizing rules that are both intriguing and beneficial. Finding intriguing association rules is a popular issue in data mining research. The third difficulty is the lengthy computation time necessary to identify huge item-set patterns. To address this issue, several attempts have been made to design more efficient algorithms or use sampling techniques to limit the quantity of data that must be processed.[1]

So, this cluster is further analysed and Apriori using Weka 3.8, has been implemented on the data which this cluster showed. In figure 5, some of the identified Association rules are shown. For example, the association rule # 3 is showing that the Product ID (StockCode) 22981 was purchases 169 times when the customer purchased 22916,22917,22919 and 22921 together for 171 times. Another example is the association rule # 8, where the Product 22916 was purchased for 168 times when customer bought 22917, 22920 and 22921 altogether for 171 times. There are number of rules are identified and it is observed that these rules are very useful for future sales forecasting and basket prediction for customers.

## 5. Conclusions

This paper shows different data-mining techniques and algorithms. Once these clustering and apriori techniques have been applied to sales data, they can help disclose intriguing information about customers, goods, and sales trends, so contributing to competitive business intelligence. For example, the discovery of association rules can result in higher-level sales forecasting and more cautious inventory control. The data may also be utilized to optimize pricing.

This article demonstrates how to develop customer-centric business information for the market using data mining techniques. The unique customer groups described in this article can assist businesses in better understanding their customers' profitability and, as a result, developing suitable marketing tactics for different consumers. Association rule mining improves sales forecasting and reveals consumer buying patterns and interests, which may be useful in customer-centric marketing and advertising.

This investigation has demonstrated that the two most important and time-consuming processes in the data mining process are data preparation and model interpretation and assessment.

Among the several methods used, the K-Means clustering method based on RFM model turned out to be the most successful for grouping consumers according on behavioural trends. With Cluster 4 recognized as the most stable and varied group, with the biggest client base over the longest time span, the five cluster solution offered the most significant segmentation. The Apriori technique was then used to examine this cluster further and found product useful association rules. These techniques taken together provided a strong

means of revealing useful insights for strategies of client retention, inventory control, and focused marketing.

## 6. Future Concerns

Further research for the business includes: conducting association analysis to establish customer buying patterns in terms of which products have been purchased frequently by which customers and which customer groups; improving the merchant's stores to allow a consumer's shopping activities to be captured and tracked instantly and accurately; and predicting each customer's lifecycle value to quantify the level of diversity of each customer.

The results can help in future research areas where multi label classification is required or the prediction about future basket of each customer should be analyzed. These clustering and association rule mining can be helpful in implementation of different neural networks such as sequence to sequence or LSTM recurrent neural network approaches for prediction of baskets and sales forecasting.

This is the need of today's marketing and advertisements, that customer and their shopping behaviors should be focused and analyzed. Large scale business has already adopted many datamining approaches to achieve these marketing goals. Now small and medium sized organization are also focusing on these strategies. So, this is a large and rich research area.

**Funding Statement:** Author has received no funding.

**Conflicts of Interest:** Author has no conflicts of interest to declare.

**Data Availability:** The dataset used on this paper is available publicly and properly referenced.

## References

[1] Tan, Swee Chuan, and Jess Pei San Lau. "Time series clustering: A superior alternative for market basket analysis." In *Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013)*, pp. 241-248. Singapore: Springer Singapore, 2013.

[2] Sarma, Hiren Kumar Deva, and Swapnil Mishra. "Mining time series data with Apriori tid algorithm." In *2016 International Conference on Information Technology (ICIT)*, pp. 160-164. IEEE, 2016.

[3] Liu, Bing, Wynne Hsu, and Yiming Ma. "Integrating classification and association rule mining." In *Proceedings of the fourth international conference on knowledge discovery and data mining*, pp. 80-86. 1998.

[4] Lipton, Zachary C., David C. Kale, Charles Elkan, and Randall Wetzel. "Learning to diagnose with LSTM recurrent neural networks." *arXiv preprint arXiv:1511.03677* (2015).

[5] Cumby, Chad, Andrew Fano, Rayid Ghani, and Marko Krema. "Building intelligent shopping assistants using individual consumer models." In *Proceedings of the 10th international conference on Intelligent user interfaces*, pp. 323-325. 2005.

[6] Cumby, Chad, Andrew Fano, Rayid Ghani, and Marko Krema. "Predicting customer shopping lists from point-of-sale purchase data." In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 402-409. 2004.

[7]Kooti, Farshad, Kristina Lerman, Luca Maria Aiello, Mihajlo Grbovic, Nemanja Djuric, and Vladan Radosavljevic. "Portrait of an online shopper: Understanding and predicting consumer behavior." In *Proceedings of the ninth ACM international conference on web search and data mining*, pp. 205-214. 2016.

[8] Lee, Munyoung, Taehoon Ha, Jinyoung Han, Jong-Youn Rha, and Ted Taekyoung Kwon. "Online footsteps to purchase: Exploring consumer behaviors on online shopping sites." In *Proceedings of the ACM web science conference*, pp. 1-10. 2015.

[9] Shangguan, Longfei, Zimu Zhou, Xiaolong Zheng, Lei Yang, Yunhao Liu, and Jinsong Han. "ShopMiner: Mining customer shopping behavior in physical clothing stores with COTS RFID devices." In *Proceedings of the 13th ACM conference on embedded networked sensor systems*, pp. 113-125. 2015.

[10] Harutyunyan, Hrayr, Hrant Khachatrian, David C. Kale, Greg Ver Steeg, and Aram Galstyan. "Multitask learning and benchmarking with clinical time series data." *Scientific data* 6, no. 1 (2019): 96.

[11] Chen, D. *Online Retail II Data Set. UCI Machine Learning Repository*. 2019.

[12] Chen, Daqing, Sai Laing Sain, and Kun Guo. "Data mining for the online retail industry: A case study of RFM

model-based customer segmentation using data mining." *Journal of Database Marketing & Customer Strategy Management* 19 (2012): 197-208.