PATTERN DISCOVERY USING CLUSTERING AND ASSOCIATION RULE ALGORITHM IN RETAIL TRANSACTIONAL DATA

Contents

1. **ABSTRACT**

This study investigates the application of appropriate machine learning algorithms on online retail transactional data. The volume of transactional data collected by businesses have grown tremendously over the past decade and many businesses seek to analyse their data to generate insight that may lead to increased profit. In the domain of machine learning, algorithm selection and data pre-processing are crucial steps which is dependant on the analysis goal. Improper applications may lead to misleading results, hence there is a need to understand suitable applications of machine learning methods. This study employs association rule and clustering algorithm on a UK-based real online dataset using the SEMMA methodology to compare the algorithm's use case and evaluate their performance. It was discovered that the association rule algorithm is able to generate rules and group frequent items effectively, while the clustering algorithm is able to segment the customers into distinct groups. These findings have implications for businesses that seek to utilize their transactional data to implement effective sales and marketing strategies based on relations and patterns observed in items and customers and ultimately increase profit.

2. INTRODUCTION

In the age of the industry 4.0, data is being generated and collected at unprecedented rates. Organizations across different sectors have the option to collect and analyse data through a wide variety of technologies such as the Internet-of-Things (IoT), SQL and NoSQL databases. This is the result of the rapid expansion in the field of data analytics where new methods, technologies and tools are constantly being developed to address challenges or improve on existing methods.

Institutions that utilize data their data effectively benefit in a multitude of ways such as improved decision-making, automation and complex problem solving. In the retail domain, vast amounts of transactional data are often collected and stored in databases (Oliveira & Sousa, 2021). A common analysis performed on the data is association rule mining where purchasing patterns for items can be identified. This insight is then utilized in sales and marketing tactics which may increase profit, customer engagement and customer satisfaction.

Apart from association analysis, another widely used pattern discovery algorithm is clustering. Clustering is an unsupervised learning method where samples are grouped based on their similarity. Clustering is normally performed in the exploration stage of the analysis on unlabelled data which results in the data being grouped according to their clusters. In the retail transaction domain, clustering can be used to groups customers based on their similarities. There are many variations of clustering algorithm these include distribution based, density based, hierarchical based and centroid based. The centroid variation of the clustering algorithm is a popular and well-known variation.

Each of the algorithms mentioned have their own use cases as their operating principle and output in terms of result is different. In this study, a comparative analysis will be performed between the clustering and association rule algorithms in the domain of retail sales.

Problem Statement

Transactional data is being collected at increasing rates and volume over time (Anitha & Patil, 2022). A significant portion of businesses also have online shops in addition to their physical retail stores, while some businesses only operate online. As a unique identifier such as name or email are required for online purchases, businesses are able to collect purchasing data for specific users as opposed to physical retail stores where keeping track of a unique customer purchase is difficult (Dongre et al., 2014). While businesses tend to have more data than ever before, if this data is not utilized effectively in analytics, this will lead to a missed opportunity to possibly increase revenue and profit. In addition, if the wrong algorithm and inadequate data pre-processing steps are taken to address a business objective, the results of the analysis is likely to be misleading or wrong.

Research Questions

Based on the problem statement the following research questions are formulated.

1. In the domain of retail sales, how is clustering and association rule algorithm useful in addressing a business problem?

2. What are the data pre-processing steps required to conduct clustering and association analysis effectively?
3. How well does association rule and clustering algorithm perform in addressing business goals?

Research Objective

1. To investigate how clustering and association rule algorithm can be applied to solve a business problem.
2. To determine best practices of data pre-processing methods for clustering and association analysis.
3. To evaluate the performance of association rule and clustering algorithm when applied to retail transactional data.

3. REVIEW OF RELEVANT LITERATURE

Pattern discovery has the objective of discovering hidden patterns or relationships among features in data and is one of the major objectives in data mining, subject to the analysis goal. In the retail domain, it is crucial for business to understand customer behaviour and purchasing trends so that appropriate marketing and sales strategy can be implemented to increase revenue and profit. The 2 algorithms examined in this study includes association rule and clustering.

Association Rule

One method of analysing relationships among items is called market basket analysis. When market basket analysis is performed on a transactional database, a rule is formed between items such as $X \rightarrow Y$, which implies that customers who purchase item X are also likely to purchase item Y. Given that association rule mining is an active area of research, a wide variety of algorithms have been developed, examples of these include Apriori, tree-based algorithms and other algorithms (Ahn, 2012). Ahn (2012) attempted to address the limitations of association rule mining related to product assignment by incorporating nearest neighbour assignments and updating assignments of the original rules generated by association mining. He discovered that product assignment was improved, which had the potential to increase revenue for retailers by cross-selling.

In a more recent study, Ogurtsov & Dorrer (2019) investigated the application of association rule mining based on Apriori algorithm to rearrange products of an actual retail store based on transactional data. They compared popular algorithms for association rule generation which includes Apriori, Eclat and FP-Growth and discovered that choice of algorithm is not important if the count of frequent transactions does not exceed 20% and length of transactions is less than 20. They discovered that clusters with a smaller number of products are less susceptible to noise which are inclusions of random items in the rules. The results of their analysis were utilized to plan procurement of goods and the arrangement of items in the store.

In a similar study (Dongre et al., 2014) investigated the application of the Apriori algorithm to discover association rules for E-Commerce applications on a simulated dataset. The purpose of their analysis is to discover frequent item sets which will enhance business strategy. The methodology employed consists of 3 main steps where in the first step, all individual elements (*i*) are search that have a minimum support of *s*. In

the second step, from the list of items in the step *i*, a *i + 1* item list is then generated. The third step involves repeating step 2 until the maximum item list is generated. They conclude that the Apriori algorithm is useful to discover hidden patterns in large transactional databases because it can discover customer behaviour based on frequent item sets.

Clustering

Clustering is an unsupervised learning algorithm that attempts to categorize unlabelled data in groups based on their measures of similarity. The algorithm attempts to maximize the similarity in a cluster while minimizing similarity between clusters to distinguish among hidden groups in the data. The performance of the clustering algorithm is highly dependent on a variety of factors such as the characteristics of the data, the pre-processing steps performed and the specific type of clustering algorithm used (Lico et al., 2021).

Clustering essentially labels unlabelled data based on measures of similarity. The algorithm has the potential to be utilized in the retail domain to identify and segment groups of customers. Parikh & Abdelfattah (2020) tested a variety of clustering algorithms along with Recency, Frequency and Monetary (RFM) analysis on online transactions to determine appropriate business strategies based on customer behaviour. The number and characteristics of clusters generated by each algorithm is different, however, the clustering algorithms successfully identified customers that yielded high profit and customers that require attention such as slippage.

A similar study was conducted by Anitha & Patil (2022) where a RFM model was utilized along with K-Means algorithm to identify new customers by analysing sales history and purchasing behaviour. Their study differs from Parikh & Abdelfattah (2020) as they only use 1 clustering algorithm, however, they tested it on online retail data and real time data. They discovered that the K-Means clustering algorithm along with RFM feature transformation technique produced clusters that had distinct groupings of recency, frequency and monetary values, enabling segmentation of customers and targeted marketing and promotions.

Lico et al. (2021) also investigated the application of various clustering algorithms on an actual retail dataset to discover groups of customers with similar purchasing patterns. In their study, they tested K-Means, K-Medoids, Agglomerative clustering and DBSCAN. Their study differs from the previously mentioned literature as they do not generate RFM features, instead they grouped purchased items of customers into 3 distinct categories. They discovered that K-Medoids is robust to outliers and is the recommended clustering algorithm, while K-Means and DBSCAN have the smallest processing time due to simpler calculations.

4. DESCRIPTION OF SOLUTION

Application of Clustering and Association Rule Algorithm in Transactional Data

Based on the method and output of the association rule and clustering algorithm, the former is primarily used to create association rules for items in a transactional dataset based on frequency of observations. The output of the algorithm, which are rules for items, can be incorporated into sales and marketing tactics to boost sales. An example of this is bundling items together, arranging items on the shop floor based on the rules and creating discounts or promotions for items frequently bought together.

Clustering on the other hand groups observations based on the measure of their similarity or dissimilarity, when applied to pre-processed transactional data, the customers can be labelled into distinct groups so that appropriation retention or loyalty strategies can be implemented based on the characteristics of the customers.

The two algorithms can be applied to a transactional dataset to for analysis goals which includes customer segmentation and frequent item rule with the main objective to boost sales.

Based on the findings in the literature review, this study will apply association rule and clustering algorithm to a real online transactional dataset for a UK-based retailer from the year 2009 to 2010. The source of the raw data is described in appendix.

Methodology SEMMA

The methodology used in this study is SEMMA, which is the short form for Sample, Explore, Modify, Model and Assess. The software used in this study includes SAS Enterprise Miner (SAS EM) and Python. Most of the analysis and preprocessing will be conducted in SAS EM, while a portion of the data pre-processing for clustering will be performed in Python. The following section describes the SEMMA process for each of the algorithms implemented in this study.

Analysis Goal

A UK based retailer has collected vast amounts of transactional data throughout its operation and would like to identify frequent items and segment its customers so that appropriate business strategies such as marketing and sale tactics can be implemented to increase revenue.

Analysis Data

The dataset contains 525461 records of online transactions of a UK based online retailer from the year 2009 to 2010. The dataset contains 8 features which includes invoice number, stock code, item description, quantity, invoice date, price, customer ID and country

Sample

In the sample phase, the data will be extracted from the source, imported into SAS EM and converted into the SAS EM file format and stored in the cloud server. This data is then ready to be used in the create data source part of the analysis. It is important to note that the data source for the clustering and association rules are different as they require different variables for analysis. This is described in table 1 and 2.

Variable	Variable Description	Role – Association Rule
Invoice	Invoice number	Rejected
StockCode	Stock number	Rejected
Description	Item name	Target
Quantity	Quantity of transaction	Rejected
Invoice Date	Invoice date	Rejected
Price	Price of item	Rejected
Customer ID	Customer ID	ID
County	Customer Country	Rejected

Table 1: Data Source Description Association Rule

For the clustering algorithm, feature transformation will be performed on the quantity, invoice, invoice date and price to generate 3 features which will be grouped according to the customer. This pre-processing will be performed in Python. Table 2 describes the features generated.

Variable	Variable Description	Role - Clustering
Customer ID	Customer ID	ID
Recency	Number of days before reference date customer made a transaction	Input
Frequency	Number of times customer made transaction	Input
Monetary	Monetary sum of all transactions made by customer	Input

Table 2: Data Source Description Clustering

Explore

In the explore phase, the variables will be explored to understand its characteristics. Special attention is paid to signs of dirty data which includes missing data, noise, intentionally bad data and inconsistent data. Depending on the investigation of the discovered dirty data, an appropriate data cleaning or transformation step will be taken.

Modify

In the modify stage, data that has been identified as dirty will be managed accordingly based on the circumstances. For example, if missing data is discovered, if the percentage of missing data is low relative to the whole dataset, the samples will be removed.

There are specific steps required in the pre-processing stage of the clustering algorithm. 3 features will be generated from the data which is described as the below. This is done as clustering on the raw transactional data will only segment the price and quantity of the transactions. The other variables are ID related variables, while the Invoice Date variable is not particularly useful in clustering in its raw form.

Feature Generated	Description
Recency	Calculated by finding the number of days between the last purchase date and a set date which is 1st Jan 2011
Frequency	Calculated by counting the number of transactions for each customer
Monetary	The calculated total amount of transaction for a customer

Table 3: Feature Generation for Clustering

The generated features will then be standardized so that the clustering algorithm places equal importance for each. This is important as the clustering algorithm is sensitive to different data scales, which if not conducted, may lead to misleading results or poor grouping of data.

Model

The models utilized in this study includes the association rule algorithm based on Apriori and the K-Means clustering algorithm which is described in detail below.

Apriori Algorithm

The Apriori algorithm involves a sequence of step to discover the most frequent item set. It consists of 2 main steps which is the Join step followed by the Prune step. An important concept of the Apriori algorithm is that all subsets of a frequent itemset must be frequent, this reduces the search space. The algorithm is represented as the following.

C_k: Candidate itemset of size k

F_k: Frequent itemset of size k

F_1 = [frequent items]

For ($k = 1$; $F_{k\ !=} \emptyset$; k++) do begin

    C_{k+1} candidates generated from L_k

        For each transaction t in database do

            increment the count of all candidates in C_{k+1} that is contained in t

    F_{k+1} = candidates in C_{k+1} with minimum support

    end

return $u_k L_k$

K-Means Clustering

K-means clustering is an unsupervised learning algorithm that groups samples on measures of similarity. The algorithm only accepts numerical values and is sensitive to different feature scales. The algorithm functions by initializing *k* number of centroids and assigning each sample to a centroid based on its distance to it. This distance can be based on a variety of measures such as Euclidean distance or Manhattan distance. A new centroid is then calculated based on the assigned clusters and the process is repeated until there are no changes in cluster assignment or the maximum number of iterations are achieved.

Assess

The results of the algorithms will be explored to discover the patterns identified. As the algorithms are a type of unsupervised learning, evaluation metrics such as RMSE and accuracy which are used for regression and classification are not employed.

5. RESULT AND DISCUSSION

This section describes the findings for explore, modify, model and assess

Explore

The variables of particular interest in explore includes country, quantity, invoice date, description and price.

There are 40 levels for the country feature, implying that the customers in the dataset were spread in 40 different nations. This suggests that this retailer has international presence and is quite well known, however, approximately 92% of the customers were from the UK. This suggests that there might be opportunities to increase the customer base overseas through targeted online advertising or investigating matters like shipping cost that may dissuade potential customers.

The quantity feature surprisingly has a minimum value of -9600 and a maximum value of 19152. As this is a transactional database for customers, it is not possible to have a negative quantity hence the quantity feature will be filtered using the filter node to remove values less than 1. Quantity is also significantly skewed at 36 and has a kurtosis value of 6277.

The most frequent single item in the dataset is 'WHITE HANGING HEART T-LIGHT HOLD' that has mode of 0.77% followed by 'REGENCY CAKESTAND 3 TIER with a mode of 0.42%. There are 4429 different levels for the description variable, meaning there are 4429 different items in the dataset.

The price variable also has negative values similar to quantity, as they are not relevant to a transactional dataset for customers observations that have a negative price will also be removed. The maximum price is 25111.09 while the feature is severely negatively skewed at -140.

There are also missing data for features Customer ID and Description, they have 3095 and 1 missing data respectively. These samples will be removed from the dataset.

Modify

Based on the findings in the explore stage the following summarizes issues that we discovered and actions to be performed on the dataset.

Feature	Issue	Pre-Processing
Quantity	Negative value	Filter data so that there are no more negative values
Price	Negative value	Filter data so that there are no more negative values
Customer ID	Missing values	Remove samples with missing values
Description	Missing Values	Remove samples with missing values

Table 4: Dirty Data

There is no additional pre-processing required for the Apriori algorithm as it only requires the Customer ID and Description variable.

3 features were created as part of modify for clustering which includes Recency, Frequency and Monetary. It was discovered that all 3 features had very heavy tails as shown in the figure below. The clustering algorithm is sensitive to skewed data hence a log transformation will be performed to make the data more normally distributed. This data will then be standardized prior clustering.

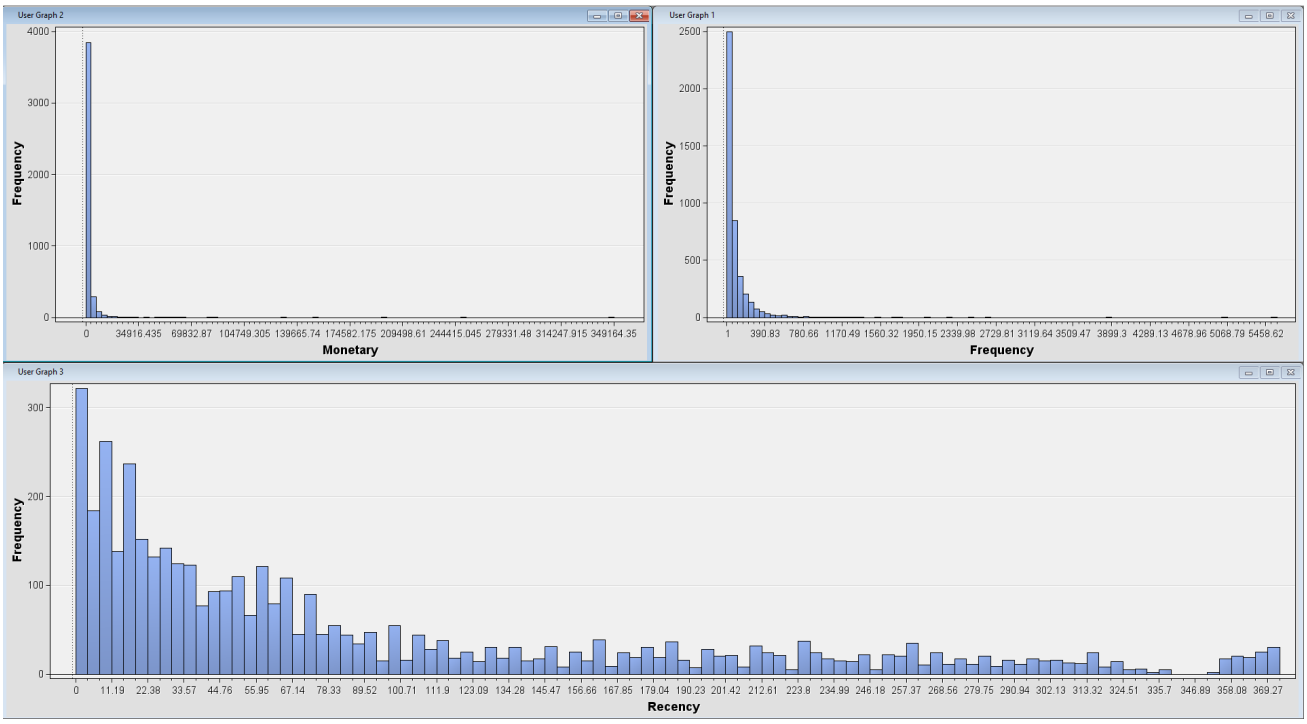


Figure 1: Distribution for RFM Variables

Assess

Apriori Algorithm

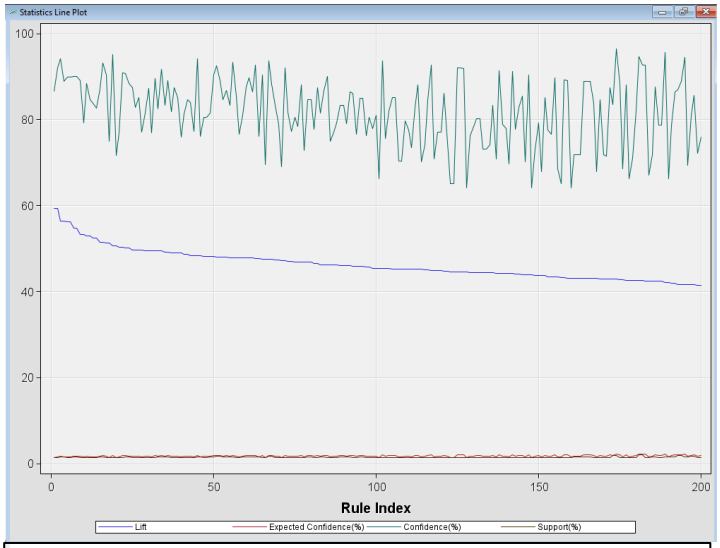
The rule with the highest lift is 'RED SPOTTY PLATE & GREEN SPOTTY BOWL (A) ➔ RED SPOTTY BOWL & GREEN SPOTTY PLATE (B)' with a lift of 59.28. This implies that customers that purchased items on the left-hand of the rule are 59 times more likely to purchase items on the right-hand side of the rule compared to a random customer. The highest rule also has a confidence of 86% and support of 1.34%. This implies that 86% of customers that have A also have B. In addition, 1.34% of the observations in the dataset have the items with the highest lift.

Association Report

Relations	Expected Confidence (%)	Confidence (%)	Support (%)	Lift	Transaction Count	Rule
4	1.46	86.57	1.34	59.28	58.00	RED SPOTTY PLATE & GREEN SPOTTY BOWL ==> RED SPOTTY BOWL & GREEN SPOTTY PLATE
4	1.55	92.06	1.34	59.28	58.00	RED SPOTTY BOWL & GREEN SPOTTY PLATE ==> RED SPOTTY PLATE & GREEN SPOTTY BOWL
4	1.67	94.12	1.48	56.39	64.00	PINK SPOTTY CUP & GREEN SPOTTY PLATE ==> PINK SPOTTY PLATE & GREEN SPOTTY CUP
4	1.58	88.89	1.48	56.39	64.00	PINK SPOTTY PLATE & GREEN SPOTTY CUP ==> PINK SPOTTY CUP & GREEN SPOTTY PLATE

Figure 2: Top 4 Rules of the Association Node

The statistics line plot of the node shows that confidence of the rules is at the higher range which varies around 80%, this suggest that for most rules, associations between items are quite high. The support however is low, which ranges between 1.3% to 2%. This implies that while confidence rules are high, the transactions do not appear frequently in the dataset in terms of percent of transactions.



Rule Statistics

The MEANS Procedure

Variable	Label	Minimum	Maximum	Mean
EXP_CONF	Expected Confidence(%)	1.4140009	2.2716736	1.7715577
CONF	Confidence(%)	64.0449438	96.4285714	81.8055246
SUPPORT	Support(%)	1.3212796	2.0630505	1.4397311
LIFT	Lift	41.5017722	59.2778962	46.3222488

Figure 3: Statistics Line Plot and Summary Statistics

The rule matrix in figure 4 highlights ‘GREEN SPOTTY PLATE’ which is an item found in the top rule. This set of rules also have high confidence.

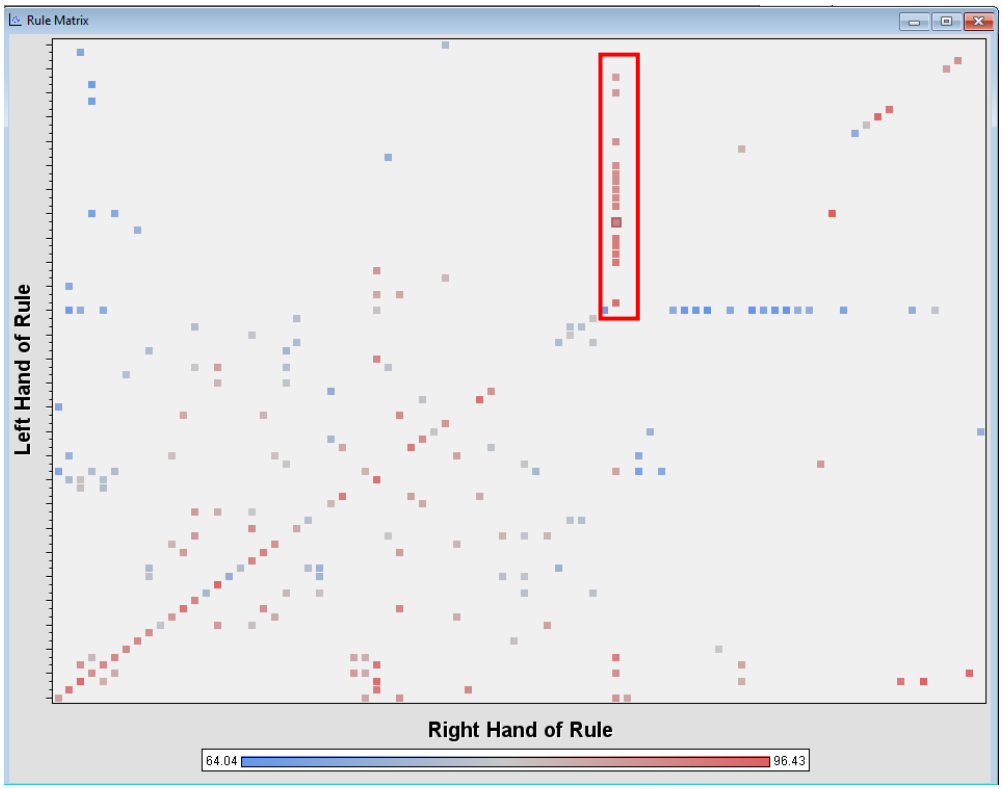


Figure 4: Rule Matrix

The rules which formed a column were highlighted and compared to a 3D plot which indicates that it has a lower support and lift but has high confidence.

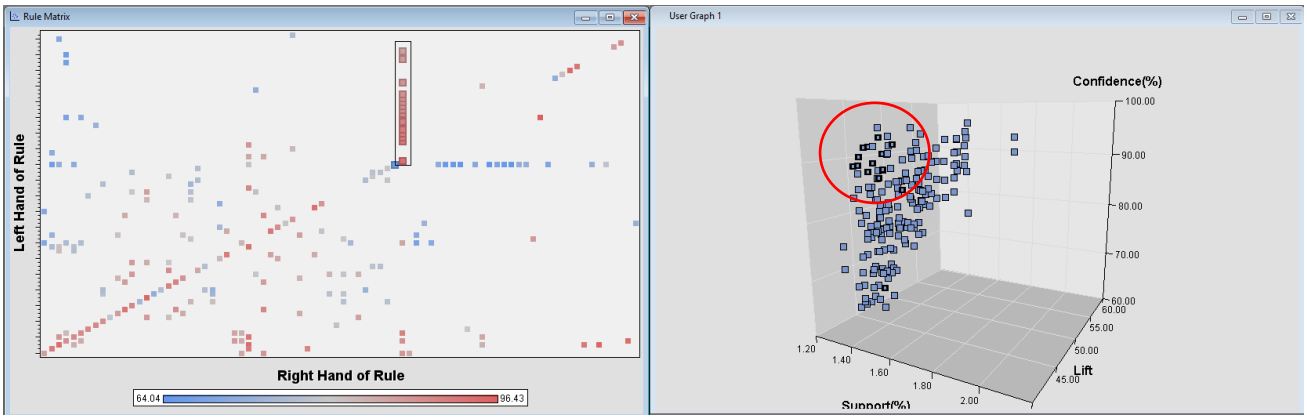


Figure 5: Rule Matrix

The association rule algorithm was successful in establishing rules based on the frequency of items and the customers that purchased them. When association rule analysis is applied on a transactional dataset a frequent item list can be generated which can affect promotional tactics such as recommendations to increase sales.

K-Means Clustering

The customers were grouped into 6 clusters as observed in figure 6. Cluster 5 has the highest number of observations with 1032 samples, while cluster 3 had the least number of samples at 383.

Clustering Criterion	Maximum Relative Change in Cluster Seeds	Improvement in Clustering Criterion	Segment Id	Frequency of Cluster	Root-Mean-Square Standard Deviation	Maximum Distance from Cluster Seed	Nearest Cluster	Distance to Nearest Cluster
0.511432	0.021872	.	1	594	0.530718	2.585221	3	1.547347
0.511432	0.021872	.	2	1013	0.471676	2.591556	1	1.609253
0.511432	0.021872	.	3	383	0.670124	3.574964	1	1.547347
0.511432	0.021872	.	4	804	0.508415	2.686408	5	1.397723
0.511432	0.021872	.	5	1032	0.429144	3.308877	4	1.397723
0.511432	0.021872	.	6	488	0.583	3.818214	5	1.535889

Figure 6: Cluster Summary

The segment plot shows the distribution of variables within the clusters. The monetary variable has the most distinct separation compared to frequency and recency. The recency variable on the other hand has a wider range data within its clusters

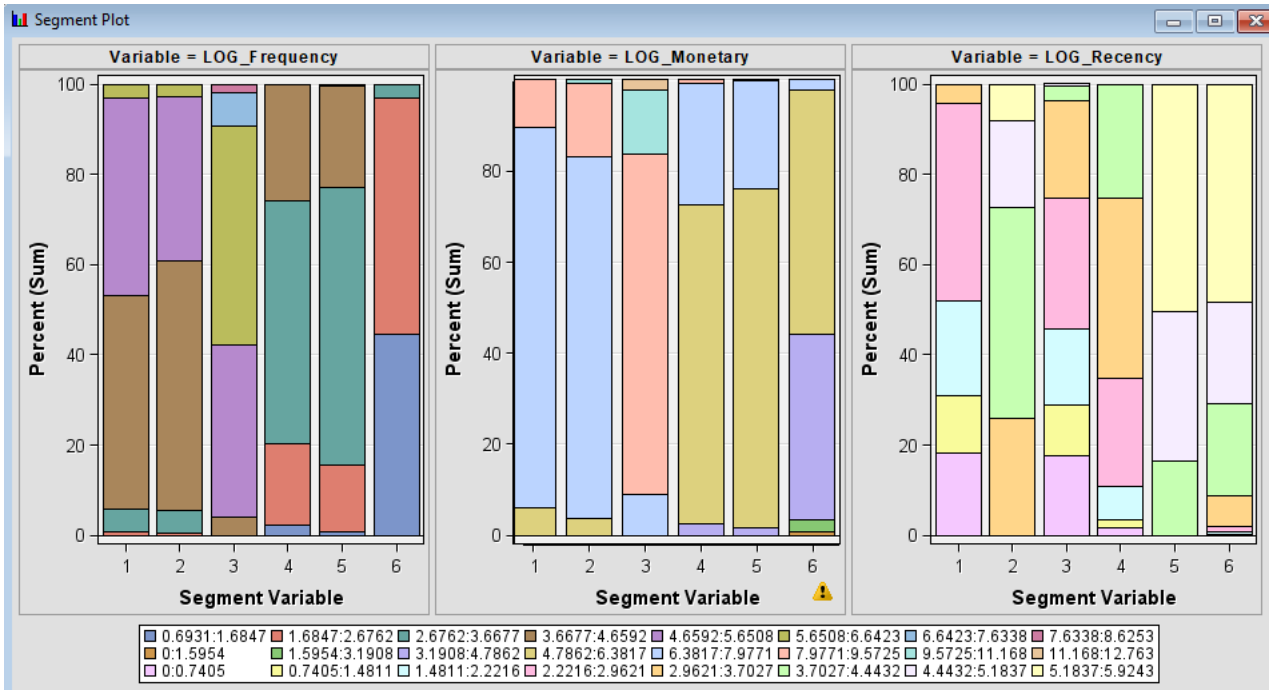


Figure 6: Cluster Segment for the 3 Variables

Through the use of the segment profile node, additional information was extracted regarding the clusters. This provided additional insight on the nature of clusters enabling implementation sales and marketing tactics.

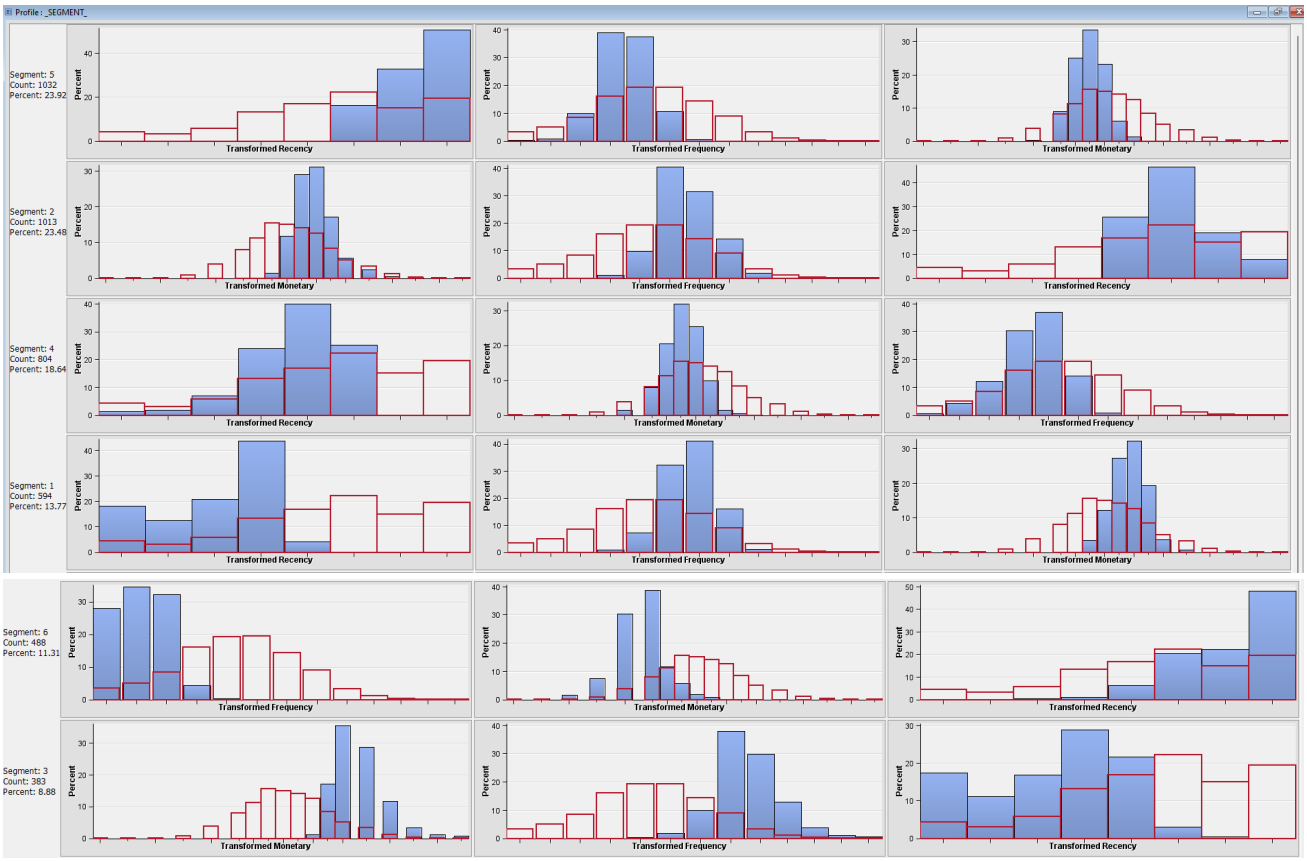


Figure 7: Segment Profile Chart

The findings of the cluster analysis are concluded in the table below.

Cluster	Characteristics
1	This cluster has high counts of low recency, high counts of average frequency and high counts of average monetary.
2	The customers have high counts of average monetary, high counts of average frequency and medium counts of recency
3	This cluster has high counts of high monetary, high counts of high frequency and high counts low of recency
4	This cluster has high counts of average recency, high counts of average monetary and low-medium high counts of frequency
5	The customers in this cluster have high recency, high counts of average frequency and high counts of average monetary
6	This cluster has high counts of low frequency, high counts of low-medium monetary and high counts of high recency.

Table 5: Cluster Summary

The clustering algorithm has segmented the customer groups according to their characteristics which can be utilized for business strategies. Cluster 3 are high value customers, they purchase frequently, have high value transactions and have low recency. A loyalty program can be implemented to these group of customers to maintain the relationship. Cluster 6 on the other hand are customers that have average monetary value but

has low frequency and high recency. These are customers upon which retaining programs might be implemented as to decrease recency and increase frequency of transactions.

6. CONCLUSION

The association rule and clustering algorithm have their own use cases in transactional data. Association rules can be used to identify patterns in transactional data to generate rules based on frequency, confidence and support. Clustering groups together samples based on measures of similarity. The data has to be pre-processed in order to conduct a cluster RFM analysis. While it is possible to cluster the transactional data directly without pre-processing, this will only result in segmentation of transactions based on price and quantity, which has less value compared to customer segmentation.

This study demonstrated the application of association rule and clustering on transactional data. The outcome of the analysis is valuable for businesses that seek to increase profit through marketing and sales tactics based on the findings of the analysis.

Data pre-processing is a crucial step especially for the clustering algorithm as it is sensitive to outliers, skewed data and different scales. If these steps are not performed, the results of the clustering are likely to be misleading.

Future studies can explore clustering on a different set of features. An example of this is to group data based on items instead of customers and cluster them according to frequency, quantity and value. For the association rule algorithm, different algorithm settings can be tested to generate rules based on different levels of support, confidence and the maximum item set.

7. BIBLIOGRAPHY

Ahn, K. il. (2012). Effective product assignment based on association rule mining in retail. *Expert Systems with Applications*, *39*(16), 12551–12556. https://doi.org/10.1016/j.eswa.2012.04.086

Anitha, P., & Patil, M. M. (2022). RFM model for customer purchase behavior using K-Means algorithm. *Journal of King Saud University - Computer and Information Sciences*, *34*(5), 1785–1792. https://doi.org/10.1016/j.jksuci.2019.12.011

Dongre, J., Prajapati, G. L., & Tokekar, S. v. (2014). The role of Apriori algorithm for finding the association rules in Data mining. *Proceedings of the 2014 International Conference on Issues and Challenges in Intelligent Computing Techniques, ICICT 2014*, 657–660. https://doi.org/10.1109/ICICICT.2014.6781357

Lico, L., Enesi, I., & Cico, B. (2021, September 16). Analyzing performance of clustering algorithms on a real retail dataset. *2021 35th International Conference on Information Technologies, InfoTech 2021 - Proceedings*. https://doi.org/10.1109/InfoTech52438.2021.9548359

Ogurtsov, D. A., & Dorrer, M. G. (2019). Application of association rules learning for studying the store history of a large retail chain. *Journal of Physics: Conference Series*, *1399*(3). https://doi.org/10.1088/1742-6596/1399/3/033114

Oliveira, J. P., & Sousa, R. D. (2021). Unsupervised anomaly detection of retail stores using predictive analysis library on SAP HANa XS advanced. *Procedia Computer Science*, *181*, 882–889. https://doi.org/10.1016/j.procs.2021.01.243

Parikh, Y., & Abdelfattah, E. (2020). Clustering Algorithms and RFM Analysis Performed on Retail Transactions. *2020 11th IEEE Annual Ubiquitous Computing, Electronics and Mobile Communication Conference, UEMCON 2020*, 0506–0511. https://doi.org/10.1109/UEMCON51285.2020.9298123

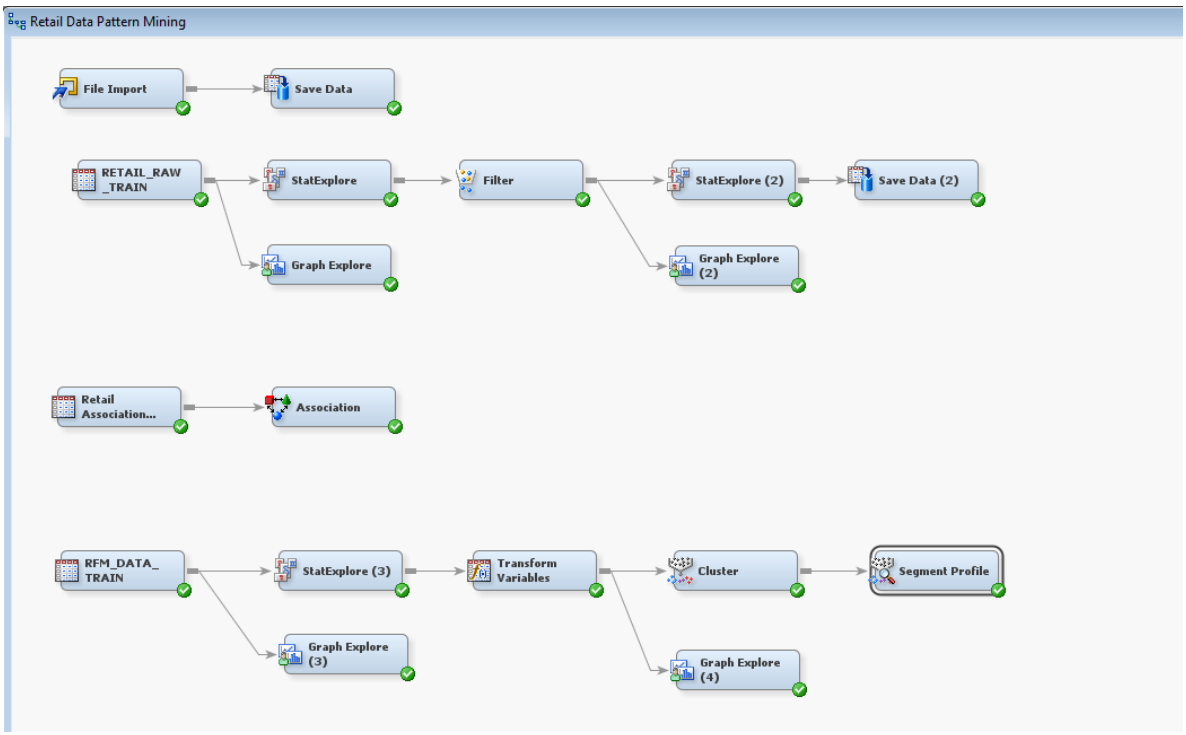
8. APPENDIX

The appendix depicts the data and analysis methods for reproducibility.

Data Source

https://www.kaggle.com/datasets/lakshmi25npathi/online-retail-dataset

Diagram of SAS EM Project for Online Retail Transaction



Sample

Raw Data - Metadata

Variables - RETAIL_RAW_TRAIN

(none) not Equal to

Columns: Label Mining

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Country	Input	Nominal	No		No	.	.
Customer_ID	Input	Nominal	No		No	.	.
Description	Input	Nominal	No		No	.	.
Invoice	Input	Nominal	No		No	.	.
InvoiceDate	Input	Interval	No		No	.	.
Price	Input	Interval	No		No	.	.
Quantity	Input	Interval	No		No	.	.
StockCode	Input	Nominal	No		No	.	.

Clean Data for Association Rule - Metadata

Variables - Retail Association Clean

(none) not Equal to

Columns: Label Mining

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Country	Rejected	Nominal	No		No	.	.
Customer_ID	ID	Nominal	No		No	.	.
Description	Target	Nominal	No		No	.	.
Invoice	Rejected	Nominal	No		No	.	.
InvoiceDate	Rejected	Interval	No		No	.	.
Price	Rejected	Interval	No		No	.	.
Quantity	Rejected	Interval	No		No	.	.
StockCode	Rejected	Nominal	No		No	.	.

RFM Data Metadata for Clustering - Metadata

Variables - RFM_DATA_TRAIN

(none) not Equal to

Columns: Label Mining

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Customer_ID	ID	Nominal	No		No	.	.
Frequency	Input	Interval	No		No	.	.
Monetary	Input	Interval	No		No	.	.
Recency	Input	Interval	No		No	.	.

Explore

The initial summary statistics depicting issues such as missing data, negative price and quantity

Variable Summary

Role	Measurement Level	Frequency Count
INPUT	INTERVAL	3
INPUT	NOMINAL	5

Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	Country	INPUT	40	0	United Kingdom	92.46	EIRE	1.84
TRAIN	Customer_ID	INPUT	513	3095	.	20.01	17850	1.57
TRAIN	Description	INPUT	513	1	PAPER CHAIN KIT 50'S CHRISTMAS	1.19	SCOTTIE DOG HOT WATER BOTTLE	1.06
TRAIN	Invoice	INPUT	513	0	490074	5.17	490149	4.99
TRAIN	StockCode	INPUT	513	0	22086	1.29	22111	1.03

Interval Variable Summary Statistics
(maximum 500 observations printed)

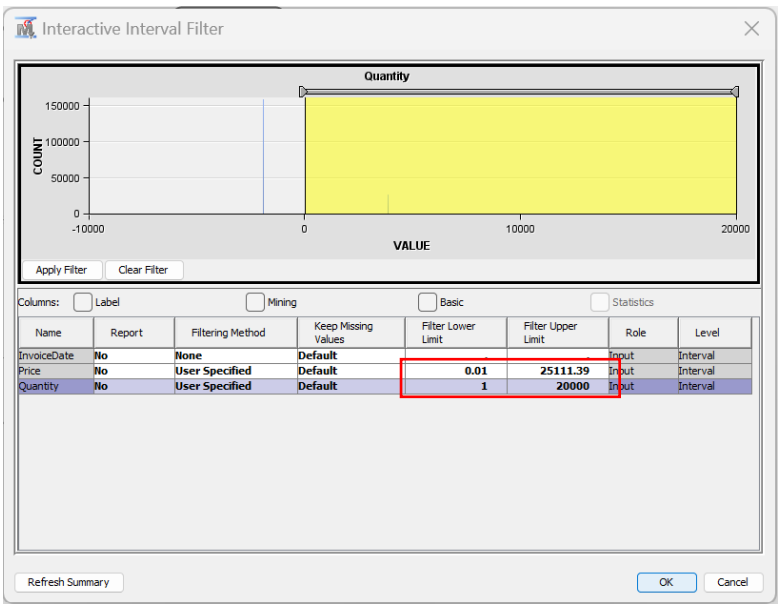
Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
InvoiceDate	INPUT	1.5933E9	10029105	525461	0	1.5753E9	1.594E9	1.6075E9	-0.24899	-1.25077
Price	INPUT	4.688834	146.1269	525461	0	-53594.4	2.1	25111.09	-140.768	64868.34
Quantity	INPUT	10.33767	107.4241	525461	0	-9600	3	19152	36.04462	6277.667

Modify

Filter node settings to remove missing data and remove negative values for quantity and price via interactive filtering. The Jupyter python notebook for creation of RFM dataset can be found in the earlier section of the appendix.

Property	Value
General	
Node ID	Filter
Imported Data	
Exported Data	
Notes	
Train	
Export Table	Filtered
Tables to Filter	All Data Sets
Distribution Data Sets	Yes
Class Variables	
Class Variables	
Default Filtering Method	None
Keep Missing Values	No
Normalized Values	No
Minimum Frequency Cutoff	10
Minimum Cutoff for Percentage	0.01
Maximum Number of Levels Cutoff	25
Interval Variables	
Interval Variables	
Default Filtering Method	User-Specified Limits
Keep Missing Values	No
Tuning Parameters	
Score	
Create Score Code	Yes
Update Measurement Level	No
Status	
Create Time	1/18/23 1:55 AM
Run ID	9d63150b-87ec-9843-96f3-bc79638f1345
Last Error	
Last Status	Complete
Last Run Time	1/18/23 4:35 AM
Run Duration	0 Hr. 0 Min. 7.13 Sec.
Grid Host	
User-Added Node	No



The below statistical summary after modify shows that the missing data and issues for quantity and price have been resolved.

Variable Summary

Role	Measurement Level	Frequency Count
INPUT	INTERVAL	3
INPUT	NOMINAL	5

Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

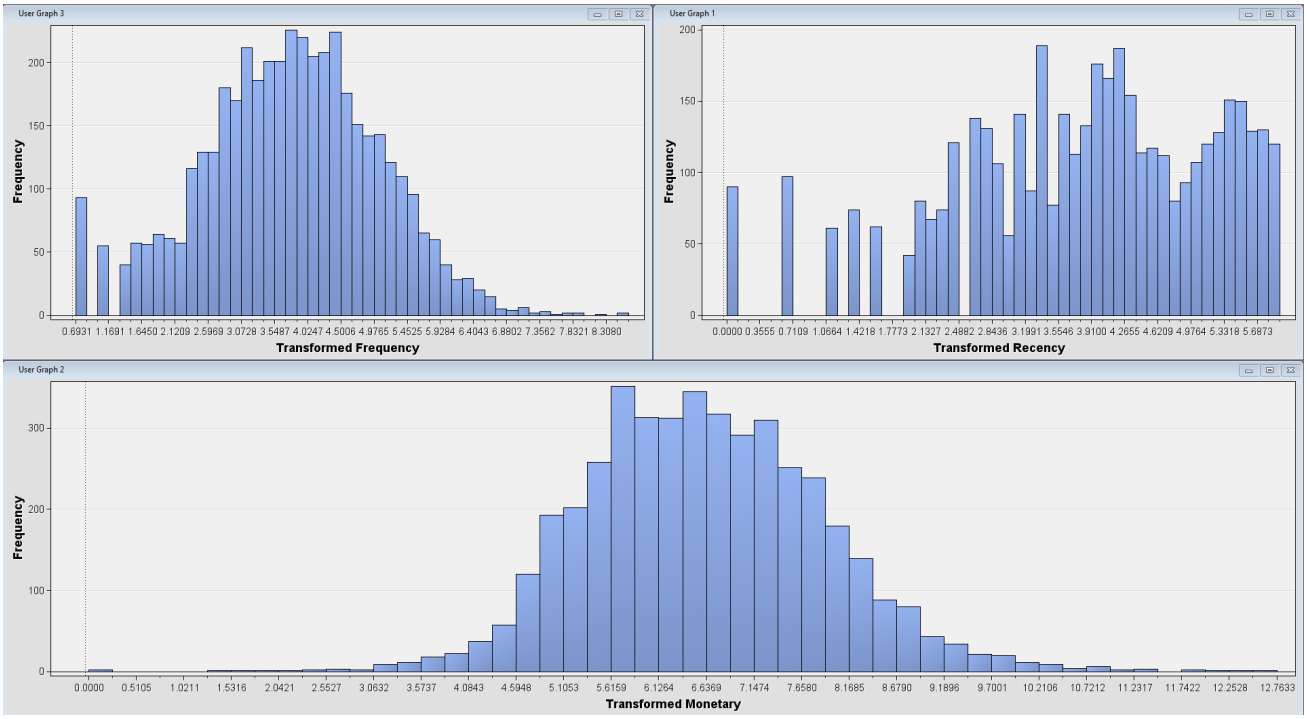
Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	Country	INPUT	37	0	United Kingdom	90.99	EIRE	2.09
TRAIN	Customer_ID	INPUT	4312	0	14911	1.37	17841	1.24
TRAIN	Description	INPUT	4429	0	WHITE HANGING HEART T-LIGHT HOLD	0.77	REGENCY CAKESTAND 3 TIER	0.42
TRAIN	Invoice	INPUT	9001	0	500356	0.14	511522	0.14
TRAIN	StockCode	INPUT	4016	0	85123A	0.77	85099B	0.43

Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

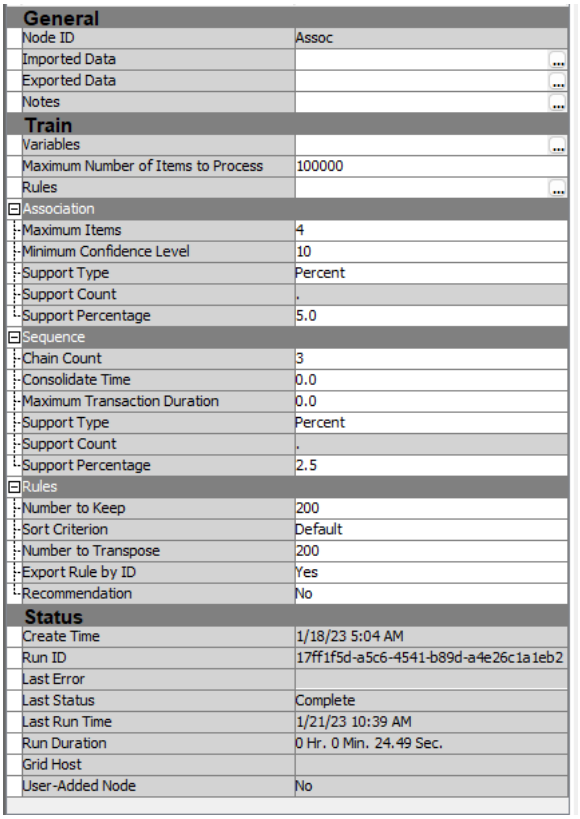
Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
InvoiceDate	INPUT	1.5936E9	9802794	407650	0	1.5753E9	1.5943E9	1.6075E9	-0.28477	-1.19795
Price	INPUT	3.294551	34.75856	407650	0	0.03	1.95	10953.5	235.0629	63863.17
Quantity	INPUT	13.58602	96.84238	407650	0	1	5	19152	79.91639	9571.746

The figure below depicts the log transformation of the RFM variables in SAS EM which resulted in a more normally distributed data.



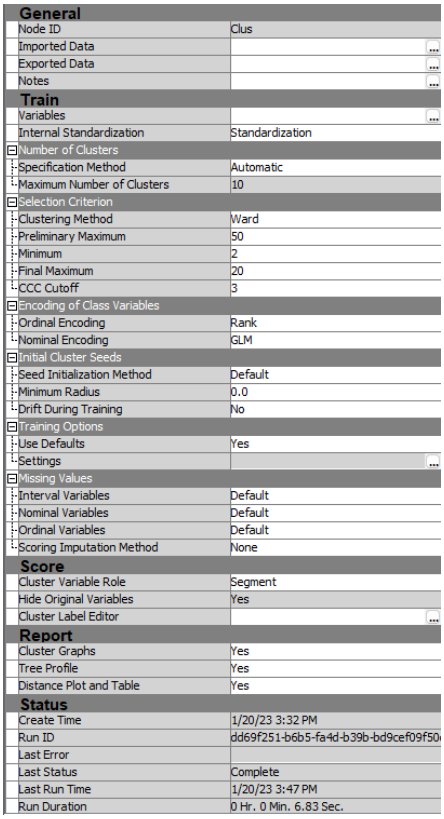
Model

Configuration for Association node and Clustering node



General	
Node ID	Assoc
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Maximum Number of Items to Process	100000
Rules	
Association	
Maximum Items	4
Minimum Confidence Level	10
Support Type	Percent
Support Count	.
Support Percentage	5.0
Sequence	
Chain Count	3
Consolidate Time	0.0
Maximum Transaction Duration	0.0
Support Type	Percent
Support Count	.
Support Percentage	2.5
Rules	
Number to Keep	200
Sort Criterion	Default
Number to Transpose	200
Export Rule by ID	Yes
Recommendation	No
Status	
Create Time	1/18/23 5:04 AM
Run ID	17ff1f5d-a5c6-4541-b89d-a4e26c1a1eb2
Last Error	
Last Status	Complete
Last Run Time	1/21/23 10:39 AM
Run Duration	0 Hr. 0 Min. 24.49 Sec.
Grid Host	
User-Added Node	No

Association

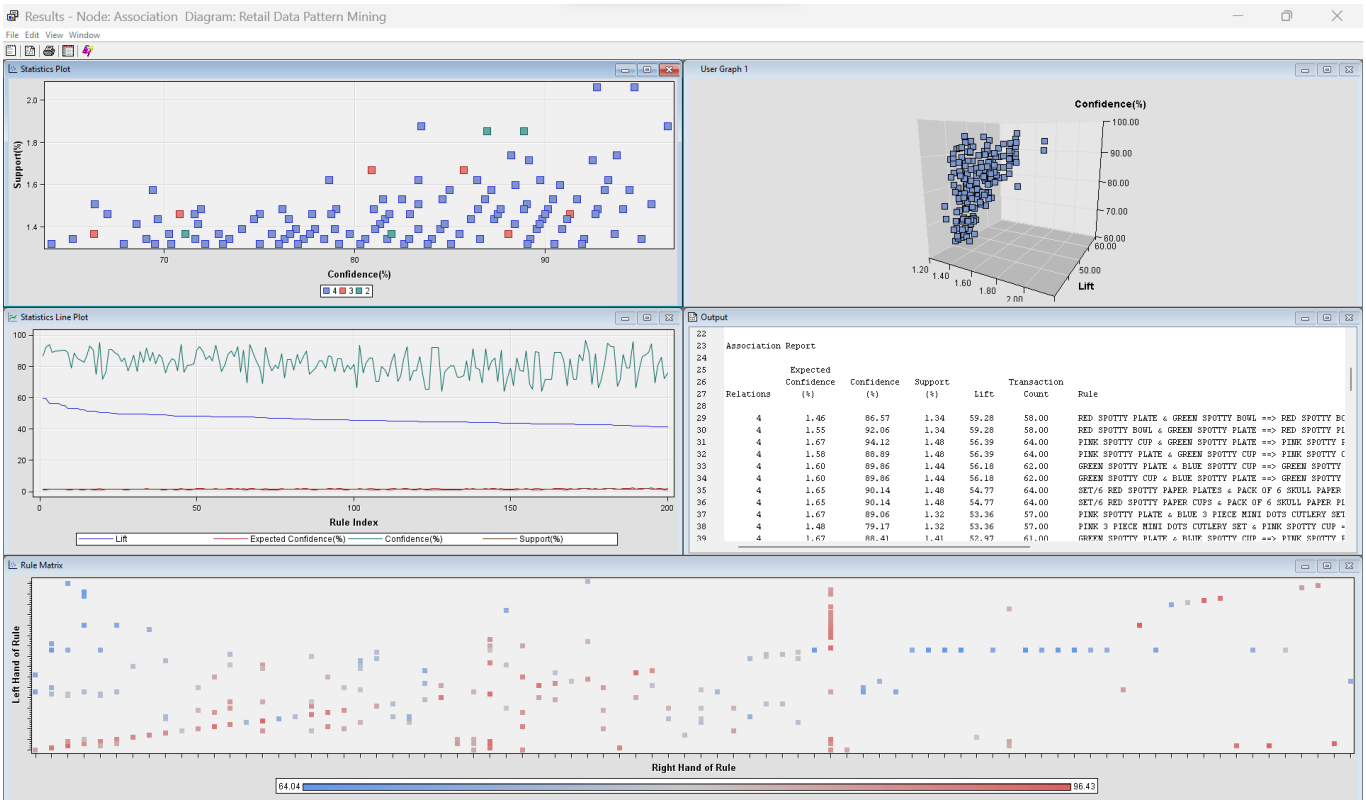


General	
Node ID	Clus
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Internal Standardization	Standardization
Number of Clusters	
Specification Method	Automatic
Maximum Number of Clusters	10
Selection Criterion	
Clustering Method	Ward
Preliminary Maximum	50
Minimum	2
Final Maximum	20
CCC Cutoff	3
Encoding of Class Variables	
Ordinal Encoding	Rank
Nominal Encoding	GLM
Initial Cluster Seeds	
Seed Initialization Method	Default
Minimum Radius	0.0
Drift During Training	No
Training Options	
Use Defaults	Yes
Settings	
Missing Values	
Interval Variables	Default
Nominal Variables	Default
Ordinal Variables	Default
Scoring Imputation Method	None
Score	
Cluster Variable Role	Segment
Hide Original Variables	Yes
Cluster Label Editor	
Report	
Cluster Graphs	Yes
Tree Profile	Yes
Distance Plot and Table	Yes
Status	
Create Time	1/20/23 3:32 PM
Run ID	dd69f251-b6b5-fa4d-b39b-bd9cef09f50
Last Error	
Last Status	Complete
Last Run Time	1/20/23 3:47 PM
Run Duration	0 Hr. 0 Min. 6.83 Sec.

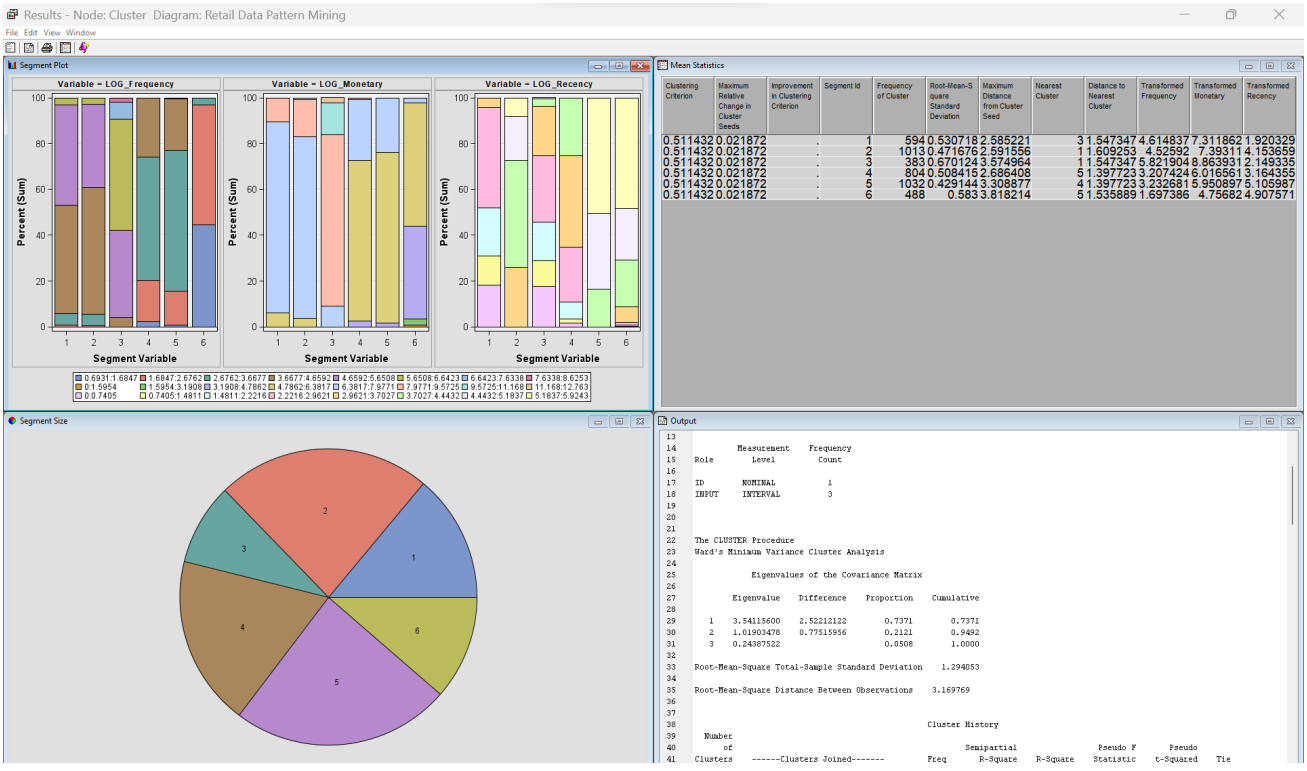
Cluster

Assess

Result of the Association Node



Result of the Clustering node



Result of the Segment Profile node