Predicting Customers Behavior on an Online Retail System Using Association and Clustering Machine Learning Algorithms (Comparative Analysis using SAS and R Languages)

IBEKWE ABUNDANCE EMERIE[1], OGUOMAIKECHUKWU STANLEY[2], EZURIKE ONYEWUCHI[3], VICTORY CHIBUIKE ONUMAKU[4], EBELE PRECIOUS OKEMBA[5]

[1] Carribean and African Health Network (CAHN), United Kingdom
[2] Department of Computer Science, University of Agriculture and Environmental Science (UAES) Umuagwo, Nigeria
[3] Department of Computer Science, Fed Polytechnic Nekede Owerri, Nigeria
[4] Edge Hill University, Ormskirk, United Kingdom
[5] University of Kent, United Kingdom

Abstract- The aim of this research is to adopt Machine Learning and Data Mining tools in predicting Customers behaviour on an Online Retail System Using Association and Clustering Machine learning Algorithms (Comparative Analysis using SAS and R Languages). The objective of the research includes analyzing a dataset using Association and Clustering mining tool to predict customers behaviors' with likely products to purchase based on customers purchase records and to facilitate customer's product decision on an online retail systems. The research was motivated due to the high demand of online services and products, inadequate customer's product satisfaction and poor product response time and product decision. Two different data mining methodologies were applied differently in the study, association and clustering algorithm while apriori and k-mean modeling tools where adopted. The data was analyzed with R and SAS Enterprise Miner. The experiments are done on UK-based and registered non-store online retail dataset, sourced from UCI machine learning repository. The result after the experiment was able to provide a model that could facilitate customer product decision making and hence provide fast online retail activities to customers and after a comparative analysis on the two different approaches, it was proven that R is one of the best data modeling tools as it is easier to clean or explore a dataset accurately than in SAS.

Indexed Terms- Artificial Intelligence, Machine Learning, Online Retail, Association, Clustering Algorithm, K-Means Modeling, Apriori modeling and Data Mining tools

I. INTRODUCTION

Data science has set the way for better analysis of data, looking at the amount of data generated on daily basis in most online transactional systems, more especially marketing on ecommerce platforms. Online Retail services are taking over the internet as number of online marketing website has increased rapidly over the globe with advanced technological tools and devices. Ability for customers to search and take decision based on a particular product online is a very huge task that requires speed and accuracy. Providing an easy means of assistance towards these issues is one of the objectives of applying association and clustering rule in this paper. For example, a customer might visits an online shop in purchase of a product say "A" which maybe has been purchased by the customer earlier, therefore the system will provide a similar type of same product say "B" base on the previous purchase made by the customer. Here, the association data mining rule has been applied to the marketing activity in this process. Now let's look at what data mining is all about, data mining also sometimes referred as knowledge discovery, is a process of analyzing data from different perspectives and summarizing it into useful information [1]. More so [1] further stated that data mining deals with huge amount of data kept in the database, so as to locate required information and hence provides reliable facts, the birth of data mining tool towards data discovery, prediction and extraction has helped researcher and scientist in numerous ways to uncover hidden facts both in physical and biological sciences. Data mining (DM) is one of the hidden extractions of predictive tool in progression from a very big database, and is now a new great technology by means of large potential to provide huge support to companies looking into the direction of large data discovery from knowledge warehouses.[2].It is now clear that the introduction of DM tool in data processing has help to predict the future trends and behaviours of events, enabling various business owners create practical knowledge-driven selections in such that it could help to answer business queries that factually remained too time irresistible to resolve. On the other hand, Cluster analysis according to [1] stated that it is a procedure which learns the substructure of a data set by distributing it into several clusters. Clustering is defined by [3]as methods for grouping of unlabeled data. Nevertheless, the clustering algorithm was adopted so that it will be easy to group customer transaction on the online retail system based on their purchase behavior and history. Customer's records are being tracked based on purchasing behavior and designed strategic initiatives. The objectives of this research is to provide a more unique means of providing and predicting customers behaviors' with likely products to purchase based on customers purchase records and hence facilitate customers product decision on an online retail systems. This research is organized as follows: Introduction: presents general introduction of machine learning (ML) and its importance to online business, it gave meaning and definitions of association rule and clustering algorithms, it also looked at data mining (DM) and how it is an important tool for effective prediction and making discovery for future use, it also highlighted the objectives of the study and gave significant facts why adoption of machine learning and data mining tools are good for decision making more especially for an online businesses. Literature Review: looks at generally the literature review on related works, machine learning modeling tools and technique, association and clustering algorithm, types of association rule approaches, and different techniques of the clustering, Methodology: methodology adoption for the study, proposed system diagram, system algorithm while Results: present results, conclusion and recommendation of the study.

II. LITERATURE REVIEW

This is one of the most recommended and researched data mining technique, these technique is aimed at extracting correlations, relationships or associations on various objects (sets of items or transactions) in either relational or distributed databases or from other related repositories within a system. According [3] stated that association is an issue surrounding the process of finding relationships in different attributes in very large customers data center of database. The attributes here may be in zeros or once (0_s and 1_s) literals, or in a quantitative state. The problem in an association data mining rule is to identify the nature of the causalities between values of different attributes [4]. The researcher used an example to illustrate an association rule stating that a supermarket that is being maintained, containing customers information from different transactions are sets of items bought by each of the customers, hence finding the purchase behavior of any item would affect another customer's purchase behavior respectively. Therefore applying association rule in this regards would provide an easy process of finding the relationship between such transactions accurately without affecting other information in the database. The identified customers data would be used to make accurate decision in respect to the purchase behavior of the customer [4].ASSOCIATION RULE APPROACH: Let A = {A1, A2, A3, A4………..An}be a set of binary literals known as items. Each of the transactions (T) is a set of items, such that $T \subseteq A$. These correspond to the set of items which a customer may buy in a market transaction. Now the association rule here is a condition of the form $z \Rightarrow Y\ where\ Z \subseteq$ A and $Z \subseteq A$ are two sets of items. Looking at the above equations, the idea of a data mining association rule is to provide a systematic means by which a user can figure out how to infer the presence. Furthermore, there are different types of association algorithms as stated by [5], such as: Apriori algorithm, Eclat algorithm, FP growth algorithm, Node set based algorithm, GUHA procedure ASSOC, OPUS search and Context based association rule mining algorithm. On the Other hand, Clustering Rule:is a type of data mining technique that deals with the grouping of similar records together in a very large database of multidimensional records, these clustering rule creates segments of data with the same similarity within a given group of points, based on the type of application use for the analysis, each of these segments may be treated differently [4]. Figure 1 below illustrates the various phases of clustering rule algorithm.

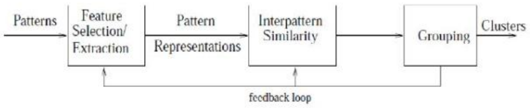


Figure 1: Diagram Showing Clustering Algorithm (Source: [1])

III. ADOPTED METHODOLOGY

- Association and Clustering Rule Algorithm

This study adopted two different data mining algorithm for an effective and more accurate result provision. The study implemented the both methodology on the topic differently (An online Retail Services) while each methodology is being adopted, a given model from the methodology was selected and apply on the online retail services dataset. The methodologies used are: Association Rule Algorithm and Clustering Rule Algorithm. The adopted algorithm for the study includes Apriori Rule Algorithm (Bottom up Approach) on dataset which is a known approach by any data scientist as it goes together with market basket analysis (MBA) in helping retailers boost business by proving accurate prediction of items to customers by using R language.

The variables in the dataset are:

1. InvoiceNo – number uniquely assign to each transaction in which some starts with "C" meaning cancelled transaction (6-digit integer)
2. StockCode– SKU number that uniquely differentiate a product from another product (5-digit integer)
3. Description—Product (item) name. Nominal
4. Quantity—The quantities of each product (item) per transaction. Numeric
5. InvoiceDate- Invoice Date and time. Numeric, the day and time when each transaction was generated.
6. UnitPrice- Unit price. Numeric, Product price per unit in pound sterling.
7. CustomerID—Customer identification number number. Nominal, a 5-digit integral number uniquely assigned to each customer
8. Country—Country name. Nominal, the name of the country where each customer resides.

IV. THE PROPOSED SYSTEM DIAGRAM

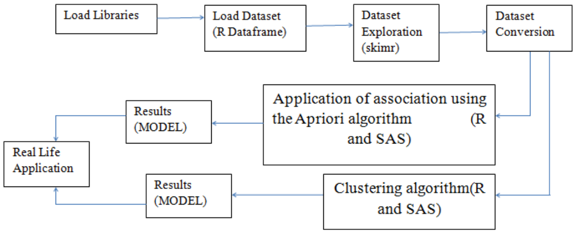
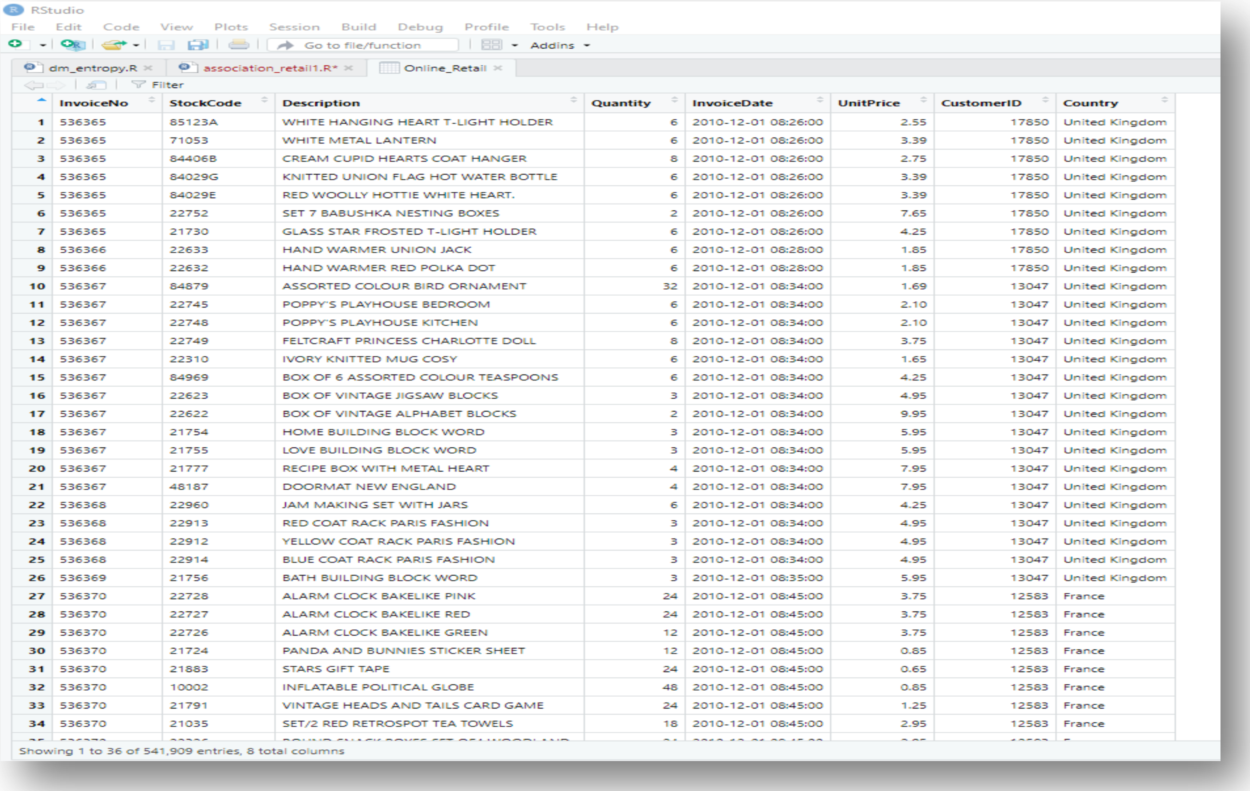

Load Libraries
Load Dataset (R Dataframe)
Dataset Exploration (skimr)
Dataset Conversion
Application of association using the Apriori algorithm (R and SAS)
Results (MODEL)
Real Life Application
Results (MODEL)
Clustering algorithm(R and SAS)

Figure 2: The Proposed System Diagram (Source: Fieldwork 2023)

V. ASSOCIATION RULE EXPERIMENTS ON THE DATASET USING R LANGUAGE

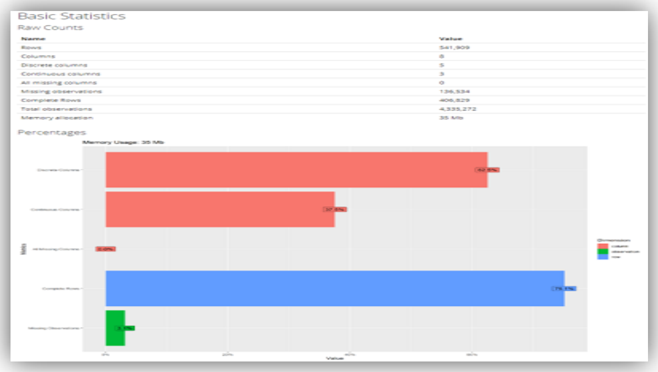
After proper packages installation, the Arule was used to initiate the association rules where the files are read by using the readxl script which was used to read or load the file named online_retail into the RStudio environment. Figure 3 shows the file after reading with 541909 observations of 8 variables. In trying to clean the data set so as to remove some values with the same invoiceN0, cancel transactions with c alphabet witnessed in the data set. Some of the cause of the double invoiceN0 might be because of some items were bought at the same time therefore cleaning the data set will help in correcting and arranging them as one transaction.



	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
1	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850	United Kingdom
2	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850	United Kingdom
3	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850	United Kingdom
4	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850	United Kingdom
5	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850	United Kingdom
6	536365	22752	SET 7 BABUSHKA NESTING BOXES	2	2010-12-01 08:26:00	7.65	17850	United Kingdom
7	536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	2010-12-01 08:26:00	4.25	17850	United Kingdom
8	536366	22633	HAND WARMER UNION JACK	6	2010-12-01 08:28:00	1.85	17850	United Kingdom
9	536366	22632	HAND WARMER RED POLKA DOT	6	2010-12-01 08:28:00	1.85	17850	United Kingdom
10	536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	2010-12-01 08:34:00	1.69	13047	United Kingdom
11	536367	22745	POPPY'S PLAYHOUSE BEDROOM	6	2010-12-01 08:34:00	2.10	13047	United Kingdom
12	536367	22748	POPPY'S PLAYHOUSE KITCHEN	6	2010-12-01 08:34:00	2.10	13047	United Kingdom
13	536367	22749	FELTCRAFT PRINCESS CHARLOTTE DOLL	8	2010-12-01 08:34:00	3.75	13047	United Kingdom
14	536367	22310	IVORY KNITTED MUG COSY	6	2010-12-01 08:34:00	1.65	13047	United Kingdom
15	536367	84969	BOX OF 6 ASSORTED COLOUR TEASPOONS	6	2010-12-01 08:34:00	4.25	13047	United Kingdom
16	536367	22623	BOX OF VINTAGE JIGSAW BLOCKS	3	2010-12-01 08:34:00	4.95	13047	United Kingdom
17	536367	22622	BOX OF VINTAGE ALPHABET BLOCKS	2	2010-12-01 08:34:00	9.95	13047	United Kingdom
18	536367	21754	HOME BUILDING BLOCK WORD	3	2010-12-01 08:34:00	5.95	13047	United Kingdom
19	536367	21755	LOVE BUILDING BLOCK WORD	3	2010-12-01 08:34:00	5.95	13047	United Kingdom
20	536367	21777	RECIPE BOX WITH METAL HEART	4	2010-12-01 08:34:00	7.95	13047	United Kingdom
21	536367	48187	DOORMAT NEW ENGLAND	4	2010-12-01 08:34:00	7.95	13047	United Kingdom
22	536368	22960	JAM MAKING SET WITH JARS	6	2010-12-01 08:34:00	4.25	13047	United Kingdom
23	536368	22913	RED COAT RACK PARIS FASHION	3	2010-12-01 08:34:00	4.95	13047	United Kingdom
24	536368	22912	YELLOW COAT RACK PARIS FASHION	3	2010-12-01 08:34:00	4.95	13047	United Kingdom
25	536368	22914	BLUE COAT RACK PARIS FASHION	3	2010-12-01 08:34:00	4.95	13047	United Kingdom
26	536369	21756	BATH BUILDING BLOCK WORD	3	2010-12-01 08:35:00	5.95	13047	United Kingdom
27	536370	22728	ALARM CLOCK BAKELIKE PINK	24	2010-12-01 08:45:00	3.75	12583	France
28	536370	22727	ALARM CLOCK BAKELIKE RED	24	2010-12-01 08:45:00	3.75	12583	France
29	536370	22726	ALARM CLOCK BAKELIKE GREEN	12	2010-12-01 08:45:00	3.75	12583	France
30	536370	21724	PANDA AND BUNNIES STICKER SHEET	12	2010-12-01 08:45:00	0.85	12583	France
31	536370	21883	STARS GIFT TAPE	24	2010-12-01 08:45:00	0.65	12583	France
32	536370	10002	INFLATABLE POLITICAL GLOBE	48	2010-12-01 08:45:00	0.85	12583	France
33	536370	21791	VINTAGE HEADS AND TAILS CARD GAME	24	2010-12-01 08:45:00	1.25	12583	France
34	536370	21035	SET/2 RED RETROSPOT TEA TOWELS	18	2010-12-01 08:45:00	2.95	12583	France

Showing 1 to 36 of 541,909 entries, 8 total columns

Figure 3: first 36 rows view of the dataset (Source: Fieldwork 2023)


Basic Statistics

Raw Counts

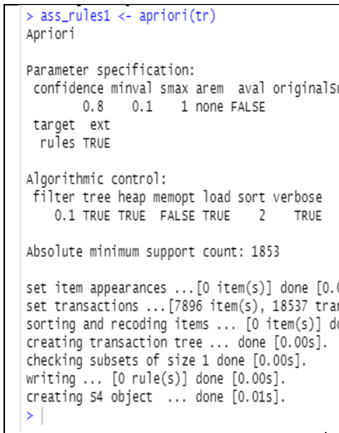
Name	Value
Rows	541,909
Columns	8
Discrete columns	5
Continuous columns	3
All missing columns	0
Missing observations	136,534
Complete Rows	406,829
Total observations	4,335,272
Memory allocation	35 Mb

Percentages

By further exploring the dataset using the head script which shows the first 6 values with tail of first 6 set structure of the data set. Applying skinr enable the output of the statistical analysis of the data set shown in figure 4 above. Therefore trying to remove the canceled transactions, the script was used "length(which(substr(str_to_UPPER(Online_Retail& InvoiceN0)1,1)== "C"))" which shows a value of 9288 observations with cancelation, then the script to remove all observations with alphabet C and save it in a new variable name called onlineRetail_non_Cancel <Online_Retail(whic h(substr(str_to_UPPER(Online_Retail &Invoic eN0),1,1)=="C")). Now to totally roved all variable containing missing values, the script was used "Online_Retail <- Online_Retail_nonCance(Complete.case(Online Retail_nonCancel))" which gave values of observation of 397924 of 8 variables. Adding rows or editing some of the columns in the data set, a function was called "mutate" because there is need to carry out some conversions like the description variable to factor variable, the same goes to conversion of country variable to a factor and editing orseparation of date/time and set independently then lastly, the invoiceN0 variable was converted to numeric then bind the new columns together. Then after the various conversions, the data set was further converted into a transaction format where a script was written to check every transaction to find out if there is any transaction with the same invoiceN0, Country, date and description if there is, it will join them together and separate with a comma which shows 18536 observations of 4 variables, because to enable the creation of the data set to basket format requires removing of all variables and leave one, therefore there is need to rename the newly created variable to “Product” by using the script “colnames(transactionData<-c(“Product”)),” then save it with a new name called clrd_online_retail_mkt_bskt.csv in the directory. Then loading of the new dataset into an object of the transaction class and assign the name “tr” to it, this is done by using the R function read.transactions of the arules package which shows transactions in sparse format with 18537 transactions in rows and 7896 items in columns. Figure 7 shows the summary is of the most frequent items with (White hanging heart T-Light holder of 1729 times, Regency cakestand of 1611 times, Jumbo bag red retrospot 1363 times, Party bunting 1285 and Assorted color bird ornament 1180 times and others with 319810) while Figure 6: absolute frequency plot which shows how frequent they appeared independently while figure 6 below shows relative item frequency bar plot of the top 20 products/items after dataset cleaning and color with “RColorBrewer” package. The relative frequency plot shows how each item is related while absolute frequency bar plot shows the absolute state of the items.

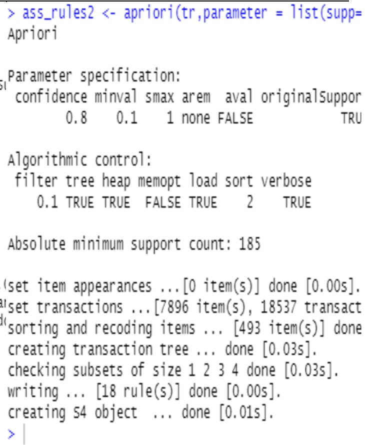
MODEL DEVELOPMENT USING APRIORI ALGORITHM

Themodel was achieved usingapriori algorithmwhich is a function in the arules package. After using the default parameters of rule which are minimum support of 0.1, minimum confidence of 0.8, maximum of 10 items (maxlen), and a maximal time for subsetchecking of 5 seconds (maxtime) and after a successive run, there was no result generated shown figure 8a.



```
> ass_rules1 <- apriori(tr)
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSu
        0.8    0.1    1 none FALSE
 target  ext
  rules TRUE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 1853

set item appearances ...[0 item(s)] done [0.(
set transactions ...[7896 item(s), 18537 tra
sorting and recoding items ... [0 item(s)] d
creating transaction tree ... done [0.00s].
checking subsets of size 1 done [0.00s].
writing ... [0 rule(s)] done [0.00s].
creating S4 object  ... done [0.01s].
> |
```

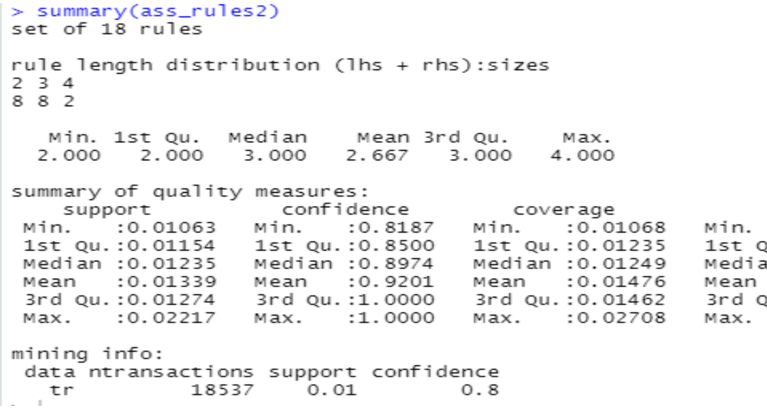
Figure 8a: The association rule using apriori algorithm



```
> ass_rules2 <- apriori(tr,parameter = list(supp=
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSuppor
        0.8    0.1    1 none FALSE             TRU

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 185

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[7896 item(s), 18537 transact
sorting and recoding items ... [493 item(s)] done
creating transaction tree ... done [0.03s].
checking subsets of size 1 2 3 4 done [0.03s].
writing ... [18 rule(s)] done [0.00s].
creating S4 object  ... done [0.01s].
> |
```

Figure 8b: The association rule using apriori algorithm

Therefore, there is need to reduce the parameters to a smaller state so as to generate a small rules. Because the researcher is interested in stronger rules in order to make the report and the visualization simple. Hence the parameters were set as follows (supp=0.01, conf=0.8, maxlen=10), and 18 rules were generated shown in figure 8b while Figure 9 below shows the summary of the apriori algorithm rules on the data set.



```
> summary(ass_rules2)
set of 18 rules

rule length distribution (lhs + rhs):sizes
2 3 4
8 8 2

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   2.000   3.000   2.667   3.000   4.000

summary of quality measures:
    support            confidence        coverage
 Min.   :0.01063   Min.   :0.8187   Min.   :0.01068   Min.
 1st Qu.:0.01154   1st Qu.:0.8500   1st Qu.:0.01235   1st Q
 Median :0.01235   Median :0.8974   Median :0.01249   Media
 Mean   :0.01339   Mean   :0.9201   Mean   :0.01476   Mean
 3rd Qu.:0.01274   3rd Qu.:1.0000   3rd Qu.:0.01462   3rd Q
 Max.   :0.02217   Max.   :1.0000   Max.   :0.02708   Max.

mining info:
 data ntransactions support confidence
   tr         18537    0.01        0.8
```

Figure 9: Summary of the apriori algorithm rules on the data set

In other to inspect the generated rules, figure 10 below inspection, shows that: 100% of the customer that purchased “SUGAR” also purchased “SET 3 RETROSPOT TEA”, 100% of the customer that purchased “COFFEE and SET 3 RETROSPOT TEA” also purchased “SUGAR” Also from bar plot; “WHITE HANGING LIGHT T-HEART HOLDER” was discovered to be the most frequent item. Some exploration was made on it to see rule where customers are purchasing more of the item, and this was carried out by reducing the support to 0.001, set the item at right hand side and maximum number of 3 item so as not to generate too many rules. Hence 36 rules were generated shown figure 10. Then lastly, carrying out data visualization, so many plot and graph were used which are made available by the arules package in R, the first graph was the scatter plot follows by the two key plot which shows a more way of using colors to add more visibility to items interactive plot which provides access to hover around the each rules and view all possible qualities, measures, support, confidence, lift of each item and attributes. Other visualization of the model plots carried out is shown from figure 11 to figure 16 respectively.

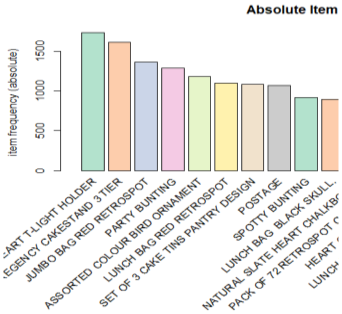
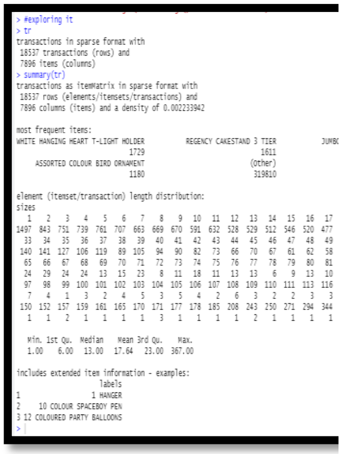
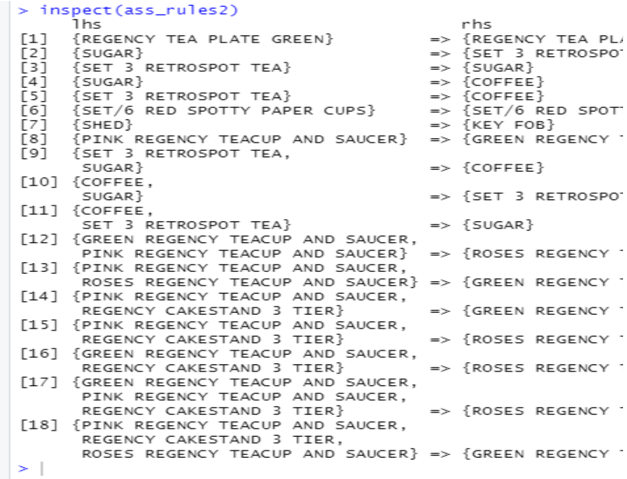


Figure 6: absolute frequency plot which shows how frequent they appeared independently



```
> #exploring it
> tr
transactions in sparse format with
 18537 transactions (rows) and
 7896 items (columns)
> summary(tr)
transactions as itemMatrix in sparse format with
 18537 rows (elements/itemsets/transactions) and
 7896 columns (items) and a density of 0.002233942

most frequent items:
WHITE HANGING HEART T-LIGHT HOLDER          REGENCY CAKESTAND 3 TIER           JUMBC
                              1729                              1611
    ASSORTED COLOUR BIRD ORNAMENT                            (Other)
                              1180                            319810

element (itemset/transaction) length distribution:
sizes
   1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17
1497  843  751  739  761  707  663  669  670  591  632  528  529  512  546  520  477
  33   34   35   36   37   38   39   40   41   42   43   44   45   46   47   48   49
 140  141  127  106  119   89  105   94   90   82   73   66   70   67   61   62   58
  65   66   67   68   69   70   71   72   73   74   75   76   77   78   79   80   81
  24   29   24   24   13   15   23    8   11   18   11   13   13    6    9   13   10
  97   98   99  100  101  102  103  104  105  106  107  108  109  110  111  113  116
   7    4    1    3    2    4    5    3    5    4    2    6    3    2    2    3    3
 150  152  157  159  161  165  170  171  177  178  185  208  243  250  271  294  344
   1    1    2    1    1    1    1    3    1    1    1    1    2    1    1    1    1

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    6.00   13.00   17.64   23.00  367.00

includes extended item information - examples:
                 labels
1              1 HANGER
2   10 COLOUR SPACEBOY PEN
3 12 COLOURED PARTY BALLOONS
> 
```

Figure 7:Summary showing the most frequent items



```
> inspect(ass_rules2)
     lhs                                     rhs
[1]  {REGENCY TEA PLATE GREEN}            => {REGENCY TEA PL
[2]  {SUGAR}                              => {SET 3 RETROSPO
[3]  {SET 3 RETROSPOT TEA}                => {SUGAR}
[4]  {SUGAR}                              => {COFFEE}
[5]  {SET 3 RETROSPOT TEA}                => {COFFEE}
[6]  {SET/6 RED SPOTTY PAPER CUPS}        => {SET/6 RED SPOT
[7]  {SHED}                               => {KEY FOB}
[8]  {PINK REGENCY TEACUP AND SAUCER}     => {GREEN REGENCY
[9]  {SET 3 RETROSPOT TEA,
      SUGAR}                              => {COFFEE}
[10] {COFFEE,
      SUGAR}                              => {SET 3 RETROSPO
[11] {COFFEE,
      SET 3 RETROSPOT TEA}                => {SUGAR}
[12] {GREEN REGENCY TEACUP AND SAUCER,
      PINK REGENCY TEACUP AND SAUCER}     => {ROSES REGENCY
[13] {PINK REGENCY TEACUP AND SAUCER,
      ROSES REGENCY TEACUP AND SAUCER}    => {GREEN REGENCY
[14] {PINK REGENCY TEACUP AND SAUCER,
      REGENCY CAKESTAND 3 TIER}           => {GREEN REGENCY
[15] {PINK REGENCY TEACUP AND SAUCER,
      REGENCY CAKESTAND 3 TIER}           => {ROSES REGENCY
[16] {GREEN REGENCY TEACUP AND SAUCER,
      REGENCY CAKESTAND 3 TIER}           => {ROSES REGENCY
[17] {GREEN REGENCY TEACUP AND SAUCER,
      PINK REGENCY TEACUP AND SAUCER,
      REGENCY CAKESTAND 3 TIER}           => {ROSES REGENCY
[18] {PINK REGENCY TEACUP AND SAUCER,
      REGENCY CAKESTAND 3 TIER,
      ROSES REGENCY TEACUP AND SAUCER}    => {GREEN REGENCY
> 
```

Figure 10: Generated Rules

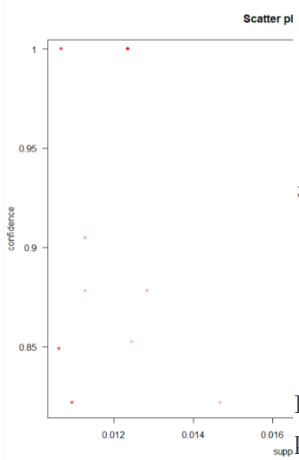
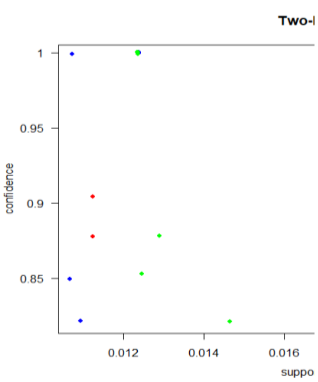


Figure 11: showing scatter plot for 18 rules, showing rules that have minimum support have higher confidence



Figuren12: The two-key plot uses support and confidence on x and y-axis, respectively. It uses *order* for colouring. The order is the number of items in the rule.

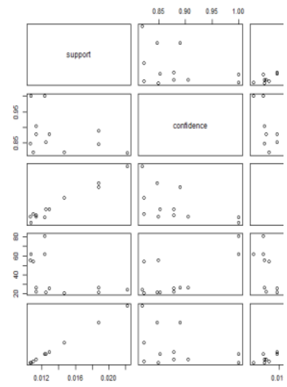


Figure 13: above is a scatterplot matrix to compare the support, confidence, and lift

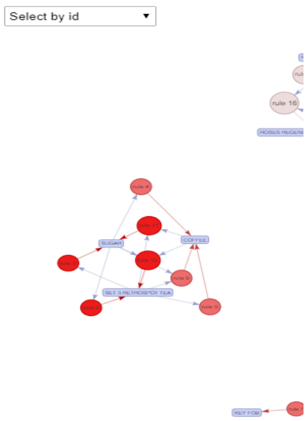
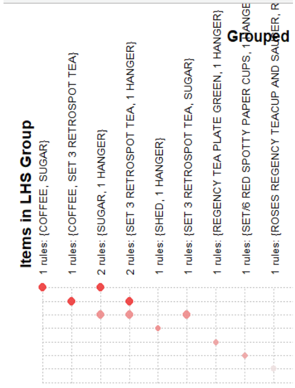


Figure 14: Showing the Graph-Based Visualizations



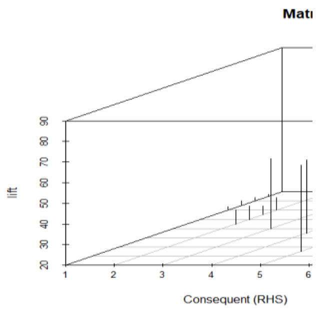
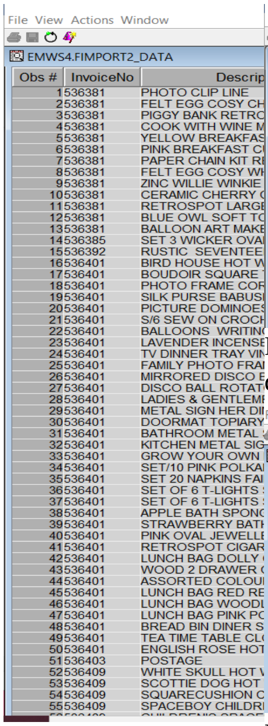


Figure 16: Showing Matrix 3 Dimension

Figure 15: Showing Grouped Matrix	Plot/Graph

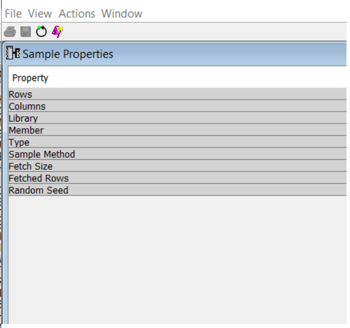
ASSOCIATION RULE EXPERIMENTS ON THE DATASET USING SAS ENTERPRISE MINER

This is summary on how the experiment was done usingSaS enterprise miner software. The enterprise miner software was launched after which a project was created. Then the cleaned data set (Online_Retail) was imported by dragging and dropping from the import node from sample menu bar of the software then exploration of the dataset was done which shows the first 55 rows shown in figure 17 below whilethe propertiesand Sample statistics of the data set was shown below in figure 18 and 19 respectively.



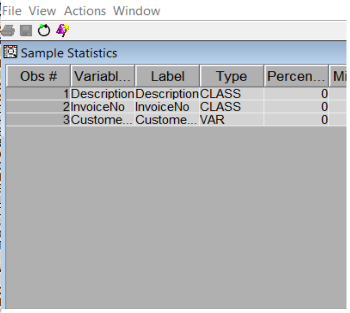
Obs #	InvoiceNo	Descrip
1	536381	PHOTO CLIP LINE
2	536381	FELT EGG COSY CH
3	536381	PIGGY BANK RETRO
4	536381	COOK WITH WINE M
5	536381	YELLOW BREAKFAS
6	536381	PINK BREAKFAST CU
7	536381	PAPER CHAIN KIT RE
8	536381	FELT EGG COSY WH
9	536381	ZINC WILLIE WINKIE
10	536381	CERAMIC CHERRY C
11	536381	RETROSPOT LARGE
12	536381	BLUE OWL SOFT TO
13	536381	BALLOON ART MAKE
14	536385	SET 3 WICKER OVAL
15	536392	RUSTIC SEVENTEE
16	536401	BIRD HOUSE HOT W
17	536401	BOUDOIR SQUARE
18	536401	PHOTO FRAME COR
19	536401	SILK PURSE BABUSI
20	536401	PICTURE DOMINOES
21	536401	S/6 SEW ON CROCH
22	536401	BALLOONS WRITING
23	536401	LAVENDER INCENSE
24	536401	TV DINNER TRAY VIN
25	536401	FAMILY PHOTO FRAI
26	536401	MIRRORED DISCO E
27	536401	DISCO BALL ROTATO
28	536401	LADIES & GENTLEME
29	536401	METAL SIGN HER DI
30	536401	DOORMAT TOPIARY
31	536401	BATHROOM METAL
32	536401	KITCHEN METAL SIG
33	536401	GROW YOUR OWN
34	536401	SET/10 PINK POLKA
35	536401	SET 20 NAPKINS FAI
36	536401	SET OF 6 T-LIGHTS
37	536401	SET OF 6 T-LIGHTS
38	536401	APPLE BATH SPONG
39	536401	STRAWBERRY BATH
40	536401	PINK OVAL JEWELLE
41	536401	RETROSPOT CIGAR
42	536401	LUNCH BAG DOLLY
43	536401	WOOD 2 DRAWER C
44	536401	ASSORTED COLOUI
45	536401	LUNCH BAG RED RE
46	536401	LUNCH BAG WOODL
47	536401	LUNCH BAG PINK PO
48	536401	BREAD BIN DINER S
49	536401	TEA TIME TABLE CLO
50	536401	ENGLISH ROSE HOT
51	536403	POSTAGE
52	536409	WHITE SKULL HOT W
53	536409	SCOTTIE DOG HOT
54	536409	SQUARECUSHION C
55	536409	SPACEBOY CHILDRI

Figure 17: view of dataset showing the first 55 rows.



Property
Rows
Columns
Library
Member
Type
Sample Method
Fetch Size
Fetched Rows
Random Seed

Figure 18: Properties of the data set in SAS



Obs #	Variabl...	Label	Type	Percen...	Mi
1	Description	Description	CLASS	0	
2	InvoiceNo	InvoiceNo	CLASS	0	
3	Custome...	Custome...	VAR	0	

Figure 19: Sample statistics of the data set in SAS

The data exploration further shows thefrequency histogram with the scale of 0 to 1500 of customerID, where customerID was segmented in range, figure 21 shows frequency Bar plot with the scale of 0 to 60 of description, showing "Rex cash+carry jumbo shopper" and "Paper chain kit 50's Christmas" as the two most frequent items and lastly figure 20 showing frequency Bar plot with the scale of 0 to 150 of Invoice number, showing invoice "537781", "536796" as the two most invoice that contains the highest number of items in a transaction respectively.

MODEL BUILDING USING SAS: Now to set target before application of the association rule algorithm, an editing of the variable to select the variables needed for specification of role (description = target, invoiceNO = ID, and customer id rejected) is required which will enable applying association rule algorithm on the data set by dragging association node from the sample menu bar on the enterprise miner and then connect the association node to the file import node, then proceed to generation of the rule, because using the entire data set to generate rules will give huge one, therefore there is need to set parameters as follows (supp(count)=100, conf=80, maxlen=2), and 49 rules was generatedwith the rule description while figure 22 shown below shows the rule table, with the confidence of 90.08%, support of 1.09%, and analysis can be made such as: 100% of the customer that purchased "REGENCY TEA PLATE PINK" also purchased "REGENCY TEA PLATE GREEN ", also with the confidence of 82.81%, support of 1.27%, analysis can be made such as: 100% of the customer that purchased "SET/6 RED SPOTTY PAPER CUPS " also purchased "SET/6 RED SPOTTY PAPER PLATES". By applying visualization, various graph plots were achieved such as Figure 23: statistics plot showing confidence against support, figure 24 showing rule matrix, RHS against LHS, figure 25 showing link graph visualizations and figure 26 showing the statistic line-plot of the rule. The achieved design model was shown in figure 27 below.

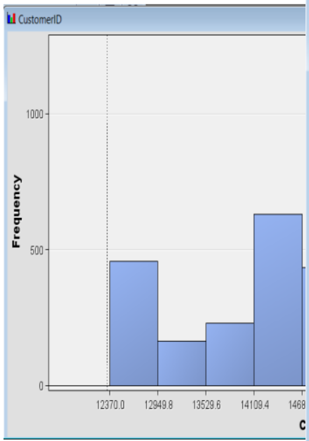


Figure 20: frequency

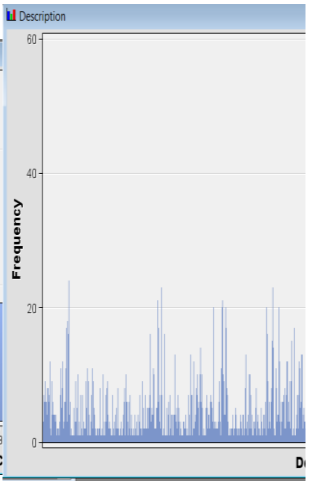
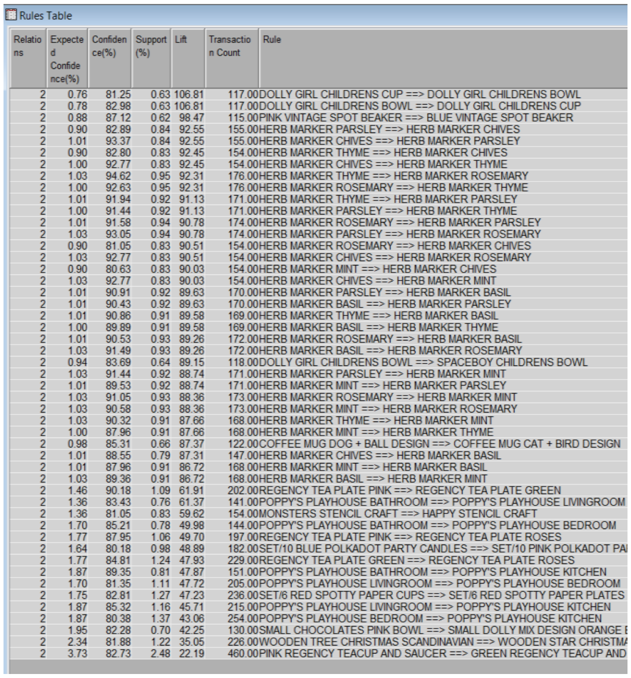


Figure 21: frequency Bar

histogram with the scale of 0 to 1500 of customerID, where customerID was segmented in range

plot with the scale of 0 to 60 of Description, showing "Rex cash+carry jumbo shopper" and "Paper chain kit 50's Christmas" as the two most frequent items respectively



Rules Table

Relations	Expected Confidence(%)	Confidence(%)	Support (%)	Lift	Transaction Count	Rule
2	0.76	81.25	0.63	106.81	117.00	DOLLY GIRL CHILDRENS CUP ==> DOLLY GIRL CHILDRENS BOWL
2	0.78	82.98	0.63	106.81	117.00	DOLLY GIRL CHILDRENS BOWL ==> DOLLY GIRL CHILDRENS CUP
2	0.88	87.12	0.62	98.47	115.00	PINK VINTAGE SPOT BEAKER ==> BLUE VINTAGE SPOT BEAKER
2	0.90	82.89	0.84	92.55	155.00	HERB MARKER PARSLEY ==> HERB MARKER CHIVES
2	1.01	93.37	0.84	92.55	155.00	HERB MARKER CHIVES ==> HERB MARKER PARSLEY
2	0.90	82.80	0.83	92.45	154.00	HERB MARKER THYME ==> HERB MARKER CHIVES
2	1.00	92.77	0.83	92.45	154.00	HERB MARKER CHIVES ==> HERB MARKER THYME
2	1.03	94.62	0.95	92.31	176.00	HERB MARKER THYME ==> HERB MARKER ROSEMARY
2	1.00	92.63	0.95	92.31	176.00	HERB MARKER ROSEMARY ==> HERB MARKER THYME
2	1.01	91.94	0.92	91.13	171.00	HERB MARKER THYME ==> HERB MARKER PARSLEY
2	1.00	91.44	0.92	91.13	171.00	HERB MARKER PARSLEY ==> HERB MARKER THYME
2	1.01	91.58	0.94	90.78	174.00	HERB MARKER ROSEMARY ==> HERB MARKER PARSLEY
2	1.03	93.05	0.94	90.78	174.00	HERB MARKER PARSLEY ==> HERB MARKER ROSEMARY
2	0.90	81.05	0.83	90.51	154.00	HERB MARKER ROSEMARY ==> HERB MARKER CHIVES
2	1.03	92.77	0.83	90.51	154.00	HERB MARKER CHIVES ==> HERB MARKER ROSEMARY
2	0.90	80.63	0.83	90.03	154.00	HERB MARKER MINT ==> HERB MARKER CHIVES
2	1.03	92.77	0.83	90.03	154.00	HERB MARKER CHIVES ==> HERB MARKER MINT
2	1.01	90.91	0.92	89.63	170.00	HERB MARKER PARSLEY ==> HERB MARKER BASIL
2	1.01	90.43	0.92	89.63	170.00	HERB MARKER BASIL ==> HERB MARKER PARSLEY
2	1.01	90.86	0.91	89.58	169.00	HERB MARKER THYME ==> HERB MARKER BASIL
2	1.00	89.89	0.91	89.58	169.00	HERB MARKER BASIL ==> HERB MARKER THYME
2	1.01	90.53	0.93	89.26	172.00	HERB MARKER ROSEMARY ==> HERB MARKER BASIL
2	1.03	91.49	0.93	89.26	172.00	HERB MARKER BASIL ==> HERB MARKER ROSEMARY
2	0.94	83.69	0.64	89.15	118.00	DOLLY GIRL CHILDRENS BOWL ==> SPACEBOY CHILDRENS BOWL
2	1.03	91.44	0.92	88.74	171.00	HERB MARKER PARSLEY ==> HERB MARKER MINT
2	1.01	89.53	0.92	88.74	171.00	HERB MARKER MINT ==> HERB MARKER PARSLEY
2	1.03	91.05	0.93	88.36	173.00	HERB MARKER ROSEMARY ==> HERB MARKER MINT
2	1.03	90.58	0.93	88.36	173.00	HERB MARKER MINT ==> HERB MARKER ROSEMARY
2	1.03	90.32	0.91	87.66	168.00	HERB MARKER THYME ==> HERB MARKER MINT
2	1.00	87.96	0.91	87.66	168.00	HERB MARKER MINT ==> HERB MARKER THYME
2	0.98	85.31	0.66	87.37	122.00	COFFEE MUG DOG + BALL DESIGN ==> COFFEE MUG CAT + BIRD DESIGN
2	1.01	88.55	0.79	87.31	147.00	HERB MARKER CHIVES ==> HERB MARKER BASIL
2	1.01	87.96	0.91	86.72	168.00	HERB MARKER MINT ==> HERB MARKER BASIL
2	1.03	89.36	0.91	86.72	168.00	HERB MARKER BASIL ==> HERB MARKER MINT
2	1.46	90.18	1.09	61.91	202.00	REGENCY TEA PLATE PINK ==> REGENCY TEA PLATE GREEN
2	1.36	83.43	0.76	61.37	141.00	POPPY'S PLAYHOUSE BATHROOM ==> POPPY'S PLAYHOUSE LIVINGROOM
2	1.36	81.05	0.83	59.62	154.00	MONSTERS STENCIL CRAFT ==> HAPPY STENCIL CRAFT
2	1.70	85.21	0.78	49.98	144.00	POPPY'S PLAYHOUSE BATHROOM ==> POPPY'S PLAYHOUSE BEDROOM
2	1.77	87.95	1.06	49.70	197.00	REGENCY TEA PLATE PINK ==> REGENCY TEA PLATE ROSES
2	1.64	80.18	0.98	48.89	182.00	SET/10 BLUE POLKADOT PARTY CANDLES ==> SET/10 PINK POLKADOT PA
2	1.77	84.81	1.24	47.93	229.00	REGENCY TEA PLATE GREEN ==> REGENCY TEA PLATE ROSES
2	1.87	89.35	0.81	47.87	151.00	POPPY'S PLAYHOUSE BATHROOM ==> POPPY'S PLAYHOUSE KITCHEN
2	1.70	81.35	1.11	47.72	205.00	POPPY'S PLAYHOUSE LIVINGROOM ==> POPPY'S PLAYHOUSE BEDROOM
2	1.75	82.81	1.27	47.23	236.00	SET/6 RED SPOTTY PAPER CUPS ==> SET/6 RED SPOTTY PAPER PLATES
2	1.87	85.32	1.16	45.71	215.00	POPPY'S PLAYHOUSE LIVINGROOM ==> POPPY'S PLAYHOUSE KITCHEN
2	1.87	80.38	1.37	43.06	254.00	POPPY'S PLAYHOUSE BEDROOM ==> POPPY'S PLAYHOUSE KITCHEN
2	1.95	82.28	0.70	42.25	130.00	SMALL CHOCOLATES PINK BOWL ==> SMALL DOLLY MIX DESIGN ORANGE E
2	2.34	81.88	1.22	35.05	226.00	WOODEN TREE CHRISTMAS SCANDINAVIAN ==> WOODEN STAR CHRISTMA
2	3.73	82.73	2.48	22.19	460.00	PINK REGENCY TEACUP AND SAUCER ==> GREEN REGENCY TEACUP AND

Figure 22: diagram showing the rules table;

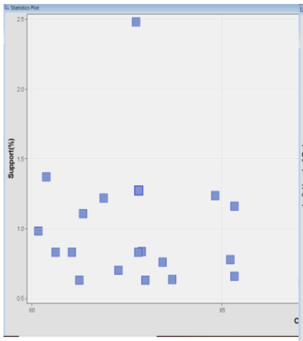


Figure 23: statistics plot showing confidence against support.

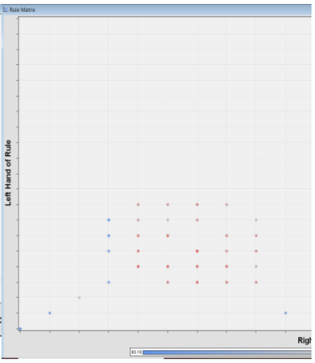
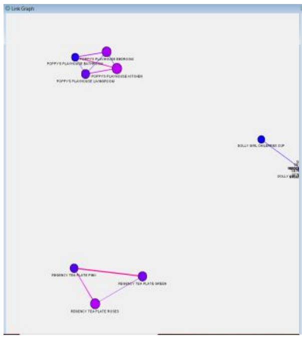


Figure 24: showing rule matrix, RHS against LHS



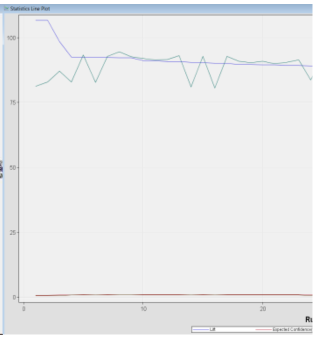
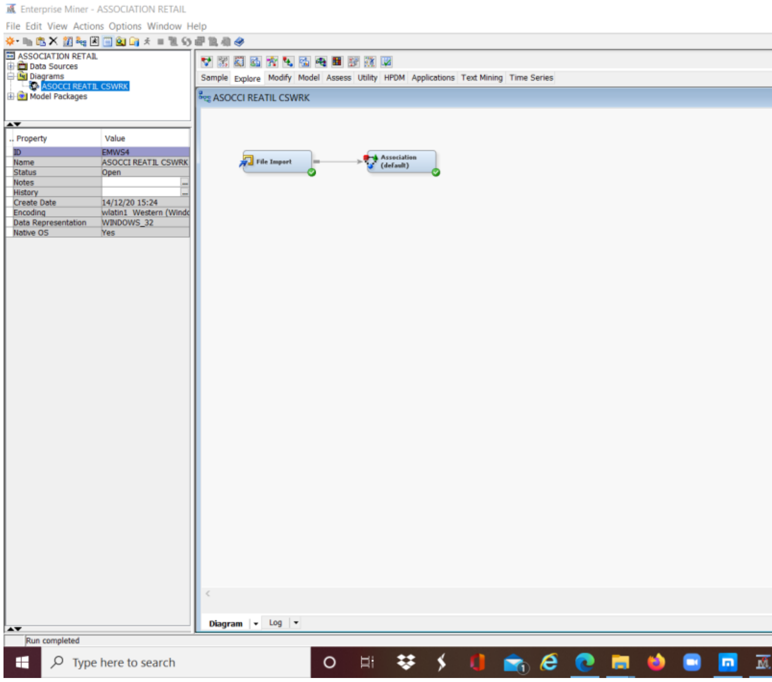


Figure 25: Showing Link Graph Visualizations

Figure 26: Showing the Statistic Line-Plot of the Rule



Enterprise Miner - ASSOCIATION RETAIL

Property	Value
ID	EMWS4
Name	ASOCCI REATIL CSWRK
Status	Open
Notes	
History	
Create Date	14/12/20 15:24
Encoding	wlatin1 Western (Windo
Data Representation	WINDOWS_32
Native OS	Yes

Figure 27: Showing the designed model on the SaS enterprise miner Software

RESULT COMPARISM OF ASSOCIATION RULE BETWEEN R AND SAS

CRITICAL FINDINGS

1. The both tools predicted (PINK REGENCY TEACUP AND SAUCER) => (GREEN REGENCY TEACUP AND SAUCER) as their highest support with

   R= 0.02217187 (2.22%)

   SAS= 2.48%
2. In R, 7rules with confidence of 1 =100% was generated while

   SAS has the highest confidence of 94.62%
3. In R, highest lift was 80.94760 with 4 rules while

SAS was 106.81 with 2 rules

4. Comparing parameter in both tools.
   R has parameter set with maxlen=10, Min_supp=, MIN_conf=, where R generated only (8 rules with maxlen=3) and 2 rules were generated with maxlen=4. While,
   SAS generated 49 rule parameter = maxlen=2, min_support_count = 100, min_confidence=80%
5. R is a very interesting language which has the ability for data manipulation. Hence it is easy to clean a data set in R while in SaS enterprise miner is not which is one of the disadvantages of using SaS in data prediction.
6. Results generated with R is more accurate than in SaS because R provided more ability to explore data to any point in other to achieve a perfect and more accurate results than in SaS.
7. Parameters can be sets for more clarifications of results in R than in SaS

APPLYING CLUSTERING RULE ALGORITHM USING R AND SAS

Clustering algorithm uses an approach where customer's records in the database are grouped and hence creates segments of the data within the group based on data similarity, then after which each segment can be treated differently depending. Figure 1 above illustrates the clustering rule algorithm which is used to analyze customer's records in the database.

EXPERIMENTS ON THE DATASET USING R

The main grail of this in R is to find the clusters of customers with the amount likely to be spent, how frequent and recent they are likely to buy based on customers purchase records, so phasing the clients based on RFM (Recency, Frequent, Monetary) which have 4339 rows and 4 column in order that the business

enterprise can target its customers efficaciously. The experiments where done and a model was developed. The approach towards the experiment includes, loading of the required packages, loading the dataset to R dataframe, exploring the data which when the first six rows) and the last six rows are viewed shows the data-types of the variables, the dimension of the dataset, and also used data-explorer package to see the frequency bar plot of the "country" variable shown in figure 28 below, data cleaning is another stage, data preparationvisual exploration of prepared dataset. After a successive exploration of the dataset, figure 28, 29, 30, 31, 32, 33 and 34below shows the result after the data exploration.

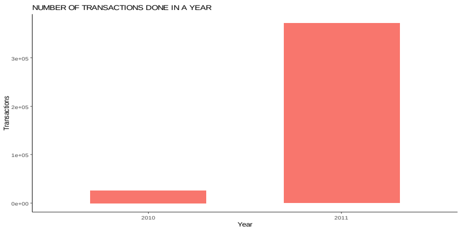


Figure 28: Bar graph showing plot of number of Transactions done in a year

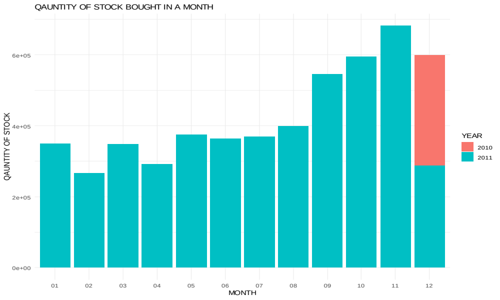
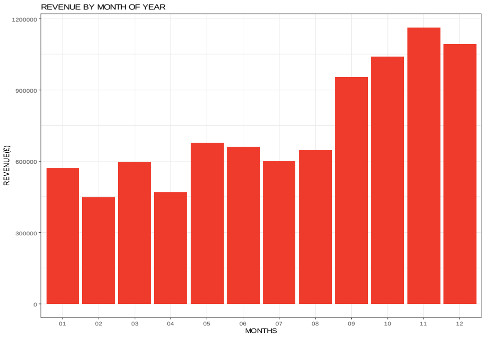


Figure 31: Plot showing the quantity of stock bought in a month within a year



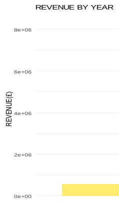


Figure 29:

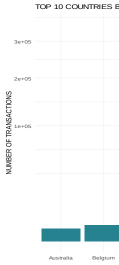


Figure 30:
highest nur

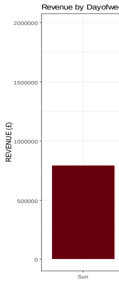


Figure 32:
basic in a v

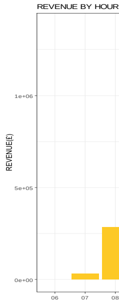
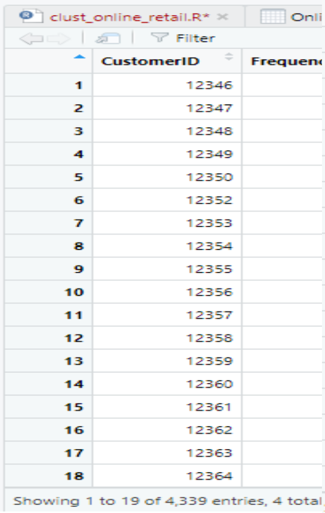


Figure 34:

Figure 33: Plot showing revenue generated in a month of a year

CALCULATING RFM

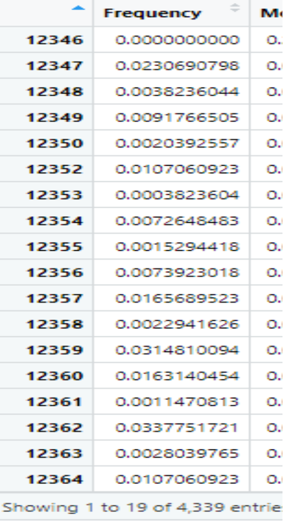
To get the RFM, where R is the closest time to the last day of transaction or the minimum difference of days calculated in STEP 14, F is the frequency of observation/row of customerID or how many times the customer had a transaction within, and M is the summation of the totalsales of a customer or adding up all the money generated by the customer and conversion of the recency to numeric by replacing all n/a observations with zero (0) shown in figure 35 and 36 respectively.



	CustomerID	Frequen
1	12346	
2	12347	
3	12348	
4	12349	
5	12350	
6	12352	
7	12353	
8	12354	
9	12355	
10	12356	
11	12357	
12	12358	
13	12359	
14	12360	
15	12361	
16	12362	
17	12363	
18	12364	

Showing 1 to 19 of 4,339 entries, 4 total

Figure 35: Showing the summarized group by customer id to get the RFM



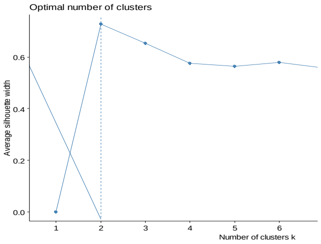
	Frequency	M
12346	0.0000000000	0.
12347	0.0230690798	0.
12348	0.0038236044	0.
12349	0.0091766505	0.
12350	0.0020392557	0.
12352	0.0107060923	0.
12353	0.0003823604	0.
12354	0.0072648483	0.
12355	0.0015294418	0.
12356	0.0073923018	0.
12357	0.0165689523	0.
12358	0.0022941626	0.
12359	0.0314810094	0.
12360	0.0163140454	0.
12361	0.0011470813	0.
12362	0.0337751721	0.
12363	0.0028039765	0.
12364	0.0107060923	0.

Showing 1 to 19 of 4,339 entrie

Figure 36: output after conversion of recency to numeric

APPLYING THE CLUSTERING ALGORITHM

In other to determine optimal number of cluster using three (3) methods namely, silhouette method, gap statistic method and elbow method shown in figure 37, 38 and 39 below.



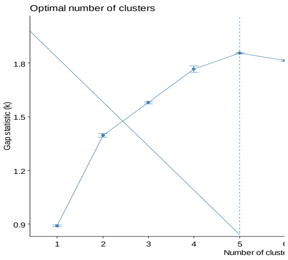


Figure 37: (SILHOUETTE METHOD) From above ofgraph it evidently shows that k=2 is the Optimal number of Cluster

Figure 38: showing OPTIMAL number of clusters with gap statistics method (which indicate k=5, hence we can determine that K=5 is the Optimal Cluster.)

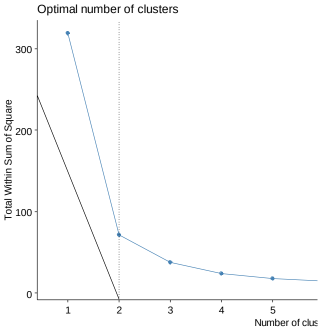


Figure 39: With the Elbow method, the graph starts to bend cluster 2, therefore we can ascertain that k=2 is the optimal cluster

Then the visualization of the clusters with the optimal clusters of the 3 methods was achieved using k-means clusters on a k=2, k=5 for better understanding shown in figure 40 and figure 41 below respectively.

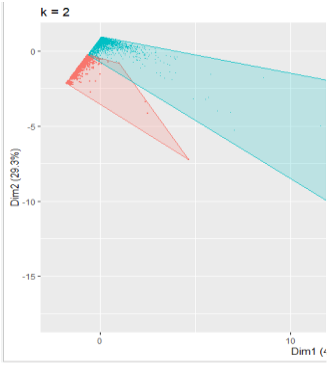


Figure 40: showing K-means cluster, where k=2

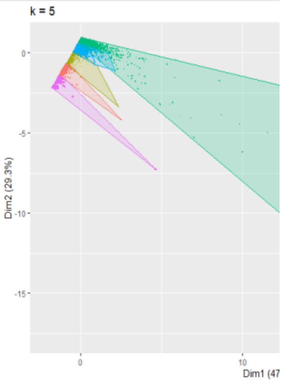


Figure 41: showing K-means, where k=5

EXPERIMENTS ON THE DATASET USING SAS

The data set was first cleaned in and then prepared using R by segmenting customers to determine clusters of customers with the amount likely to be spent, how frequent and recent they are likely to buy (RFM) and that is the objective. Next is to normalize and save it in the working directory. This is done because SAS cannot clean the data set before it could be used to carry out another experiment in SAS. After a successive data exploration, the clustering rule algorithm was done by dragging clustering node from the explore menu bar on the enterprise miner software. Then connect the clustering node to the file import node which shows the victuals on figure 42 while figure 43 shows the mean statistics of the dataset using SAS enterprise miner and figure 45 shows index screenshot of the generated model on the enterprise miner softwarebelow:

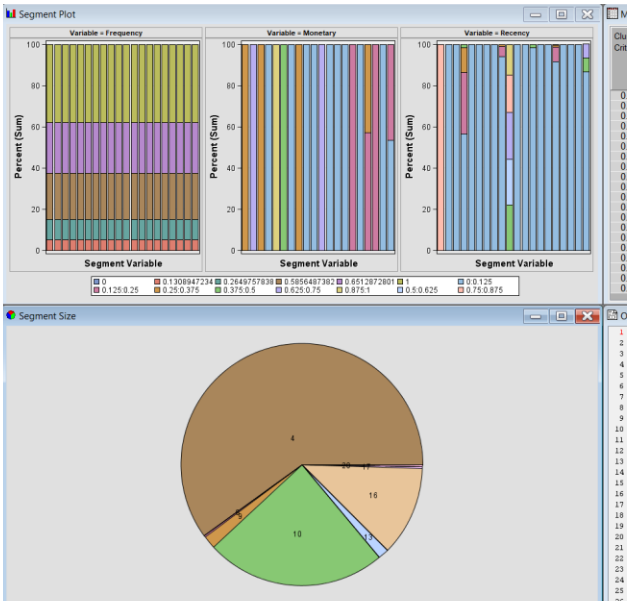
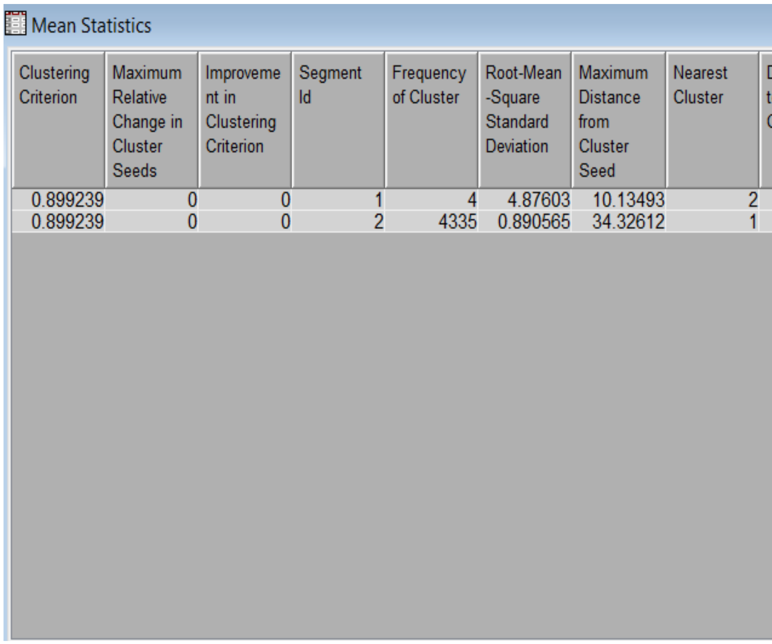
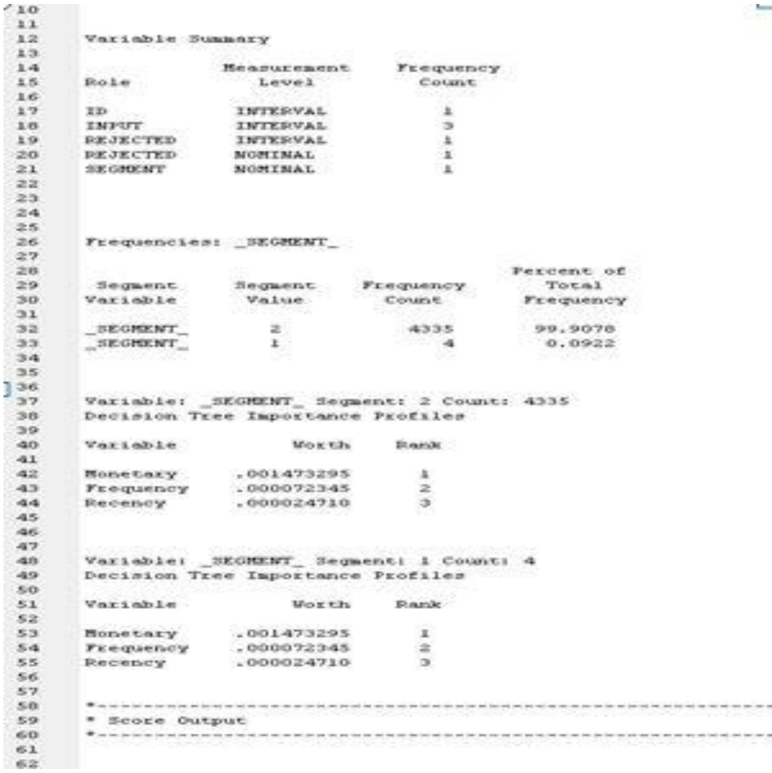


Figure 42: showing result generated at specification method set at automatic


Mean Statistics

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
0.899239	0	0	1	4	4.87603	10.13493	2
0.899239	0	0	2	4335	0.890565	34.32612	1

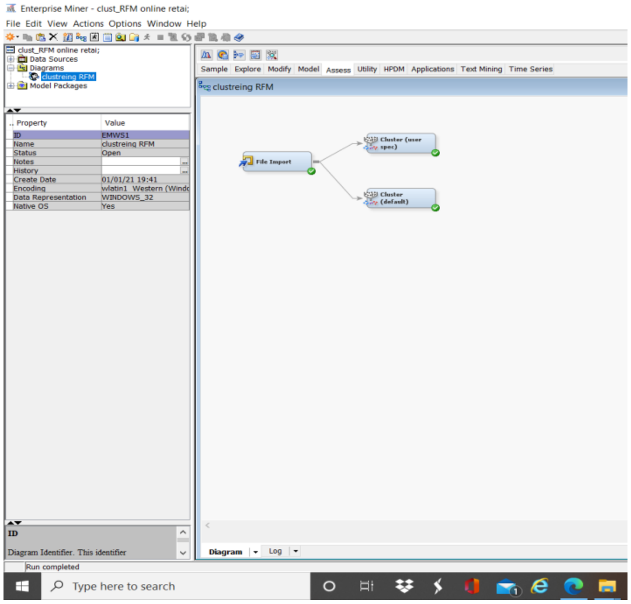
Figure 43: showing the mean statistics



```


Variable Summary

                 Measurement    Frequency
Role                Level         Count

ID                INTERVAL          1
INPUT             INTERVAL          3
REJECTED          INTERVAL          1
REJECTED          NOMINAL           1
SEGMENT           NOMINAL           1




Frequencies: _SEGMENT_

                                                   Percent of
 Segment        Segment      Frequency               Total
Variable         Value         Count               Frequency

_SEGMENT_          2            4335                99.9078
_SEGMENT_          1               4                 0.0922



Variable: _SEGMENT_ Segment: 2 Count: 4335
Decision Tree Importance Profiles

Variable                Worth       Rank

Monetary           .001473295         1
Frequency          .000072345         2
Recency            .000024710         3



Variable: _SEGMENT_ Segment: 1 Count: 4
Decision Tree Importance Profiles

Variable                Worth       Rank

Monetary           .001473295         1
Frequency          .000072345         2
Recency            .000024710         3


*------------------------------------------------------------
* Score Output
*------------------------------------------------------------


```

Figure 44: showing the result after the model was applied and summary of the variable

INDEX



Enterprise Miner - clust_RFM online retai;

Property	Value
ID	EMWS1
Name	clustreing RFM
Status	Open
Notes	
History	
Create Date	01/01/21 19:41
Encoding	wlatin1 Western (Windc
Data Representation	WINDOWS_32
Native OS	Yes

ID

Diagram Identifier. This identifier

Run completed

Figure 45: Index screenshot of the generated model on the enterprise miner software

CONCLUSION

Every business wants to satisfy its customers and also provide a free and easy access to her customers' needs remotely, therefore understanding your customers behaviors is one of the best approach in providing the goods and services they require and it is through their patronize the growth of the organization could be guaranteed. This study tends to predict customers behavior on an online retail system by using association and clustering machine learning algorithms a (comparative analysis using SAS and R languages). The analysis was first conducted using R language which showed values of observation of 397924 and 8 variables. The model was built by using apriori algorithm and it shows in figure 10 that: 100% of customer that purchased "SUGAR" also purchased "SET 3 RETROSPOT TEA", 100% of the customer that purchased "COFFEE and SET 3 RETROSPOT TEA" also purchased "SUGAR" and also "WHITE HANGING LIGHT T-HEART HOLDER" was discovered to be the most frequent item purchase by customers meaning that the organization can go ahead to produce more of the identified products than others. When the SaS enterprise miner software was used to perform the same association by using the same dataset, the result shown in figure 22 showsthe rule table, with the confidence of 90.08% and support of 1.09%, and analysis was made such as: 100% of the customer that purchased "REGENCY TEA PLATE PINK" also purchased "REGENCY TEA PLATE GREEN ", also with the confidence of 82.81%, support of 1.27%, analysis was made such as: 100% of the customer that purchased "SET/6 RED SPOTTY PAPER CUPS " also purchased "SET/6 RED SPOTTY PAPER PLATES".While figure 27: shows the designed model on the SaS enterprise miner Software whit the developed dashboard design. On the other hand, by application of clustering algorithm on both R and SaS revealed another predictive model that can help various business organization progress and understand their customers behavior and hence improve on their service delivery. To achieve this, R was first used to discover the clusters of customers with the amount likely to be spent on a product, how frequent and recent they are likely to purchased based on customers purchase records by calculating the customers RFM (Recency, Frequent, Monetary) status as the predictive formula and when the K-means model was applied on the dataset figure 35 shows the summarized group by customer id to get the RFM and figure 36 shows output after conversion of recency to numeric with number of customers recency, frequent and monetary on a particular product while application of SAS enterprise miner on the dataset shows a result on figure 44 after the model was applied and summary of the variable.

RECOMMENDATION

Machine learning models are part of artificial intelligence that helps developers predict or forecast an event and make intelligent decision base of knowledge discovery from data. This study after a comparative analysis on an online retail System as a dataset to predict customers market behavior therefore make the following recommendations:

1. Business owners should explore the intelligent information on this study for an effective understanding of their data to day business activities and also on how to handle their customers' needs always.
2. In other not to loss in business activities especially those involved with online transactions, customers comments/feedback is highly needed for effective and smooth running of the organization and hence

provide those required products and series needed by their customers.

3. It is very good for academic scholars to explore other data mining tools for more accurate prediction and development of intelligent models as this study has shown us the importance and huge difference between R language and SAS enterprise miner for data scientist and data exploration and visualization.

REFERENCE

1. MeenuSharma(2014)Clustering In Data Mining : A Brief Review;
International Journal Of Core Engineering & Management (IJCEM) Volume 1, Issue 5, accessed from file:///C:/Users/IPROSO~1/AppData/Local/Temp/Data_Clustering_Using_Data_Mining_Techni.pdf

2. Pravarti Jain AndSantosh Kr Vishwakarma (2017)A Case Study on Car
Evaluation and Prediction: Comparative Analysis using Data Mining Models, International Journal of Computer Applications (0975 – 8887) Volume 172 – No.9

3. Zhong and L. Zhou (1999): PAKDD'99, LNAI 1574, pp. 13–23, 1999. Springer-
Verlag Berlin Heidelberg.

4.Charu C. Aggarwal and Philip S. Yu (2012) Data Mining Techniques for
Associations, Clustering and Classification, IBM T. J. Watson Research Center, Yorktown Heights, NY 10598 accessed from https://link.springer.com/article/10.1057/dbm.2012.17

5.Aayushi Maheshwari, Garima Kharbanda and Harsh Patel(2015)Association Rules
in Data Mining, accessed from Data_mining_for_the_online_retail_industry_A_case_study_of_RFM_model-based_customer_segmentation_using_data_mining