**PROJECT 3 FINAL REPORT**

(min\_sup values: 0.006,009,0.014. and min\_conf values: 0.5,0.7,0.8)

**Submitted by:**

Team No. 6

Amrutha Kenchanagowdra -1001861876

Ajay Shivarudrappa - 1001720944

**PROJECT OVERVIEW**

* The dataset containing the retail data of store purchases of 541909 rows was imported into R studio as a .xlsx file using read\_excel function (Library: readxml).
* The “invoiceNo” values in the dataset which were starting with “C” were discarded (9288 records were removed)
* Records with Invalid item descriptions given below were carefully eliminated using **grepl** command so that other valid descriptions containing these below words were not removed.

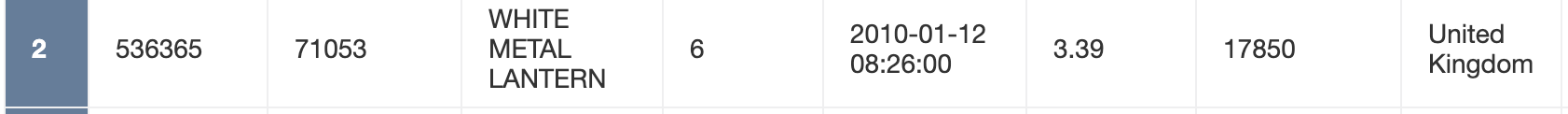
("WRONG","LOST", "CRUSHED", "SMASHED", "DAMAGED", "FOUND", "THROWN", "MISSING", "AWAY", "\\?", "CHECK", "POSTAGE", "MANUAL", "CHARGES", "AMAZON", "FEE", "FAULT", "SALES", "ADJUST", "COUNTED", "LABEL", "INCORRECT", "SOLD", "BROKEN", "BARCODE", "CRACKED", "RETURNED", "MAILOUT", "DELIVERY", "MIX UP", "MOULDY", "PUT ASIDE", "ERROR", "DESTROYED", "RUSTY" )

* Incomplete records with missing invoiceNo and Description were eliminated using **complete.cases** command and thus total of 13509 records were eliminated as part of preprocessing and **528400** records were selected for further process.
* The data in the dataframe in the below format was converted into baskets format where all the items of each transactions are grouped by invoiceNo and written to csv file

Xlsx file:

A picture containing table

Description automatically generated



Csv file:



* The csv file was read and converted into transactions using read.transactions command

Text

Description automatically generated

After converting the data into transactions, we did analyze the summary of the transactions generated:

There were 19846 transactions generated.

Total number of items sold were equal to 19899\*8747\*0.002368195 = 412199 (non-empty cells)

A picture containing text

Description automatically generated

**Frequent items analysis:**

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated

We have plotted the top 20 frequent items from the total number of transactions for both absolute and relative types. Absolute type plots the numeric frequencies of each item independently. Relative type plots how many times the items have appeared as compared to others.

This plot shows that 'WHITE HANGING HEART T-LIGHT HOLDER' and 'REGENCY CAKESTAND 3 TIER' have the most sales. So to increase the sale of 'SET OF 3 CAKE TINS PANTRY DESIGN' the retailer can put it near 'REGENCY CAKESTAND 3 TIER'.

**Candidate itemsets:**

Chart

Description automatically generated Chart, treemap chart

Description automatically generated Graphical user interface, chart

Description automatically generated

We have generated the candidate itemsets by minimizing the given support and confidence values to the system capacity and plotted the variation of candidate itemsets over 5 iterations. The number of candidate itemsets were increasing in the beginning iterations and are expected to decrease in the next iterations as the combination of items will be increased going further.

Below plot shows the candidate itemsets for all 3 minimum support and confidence values for all the iterations.

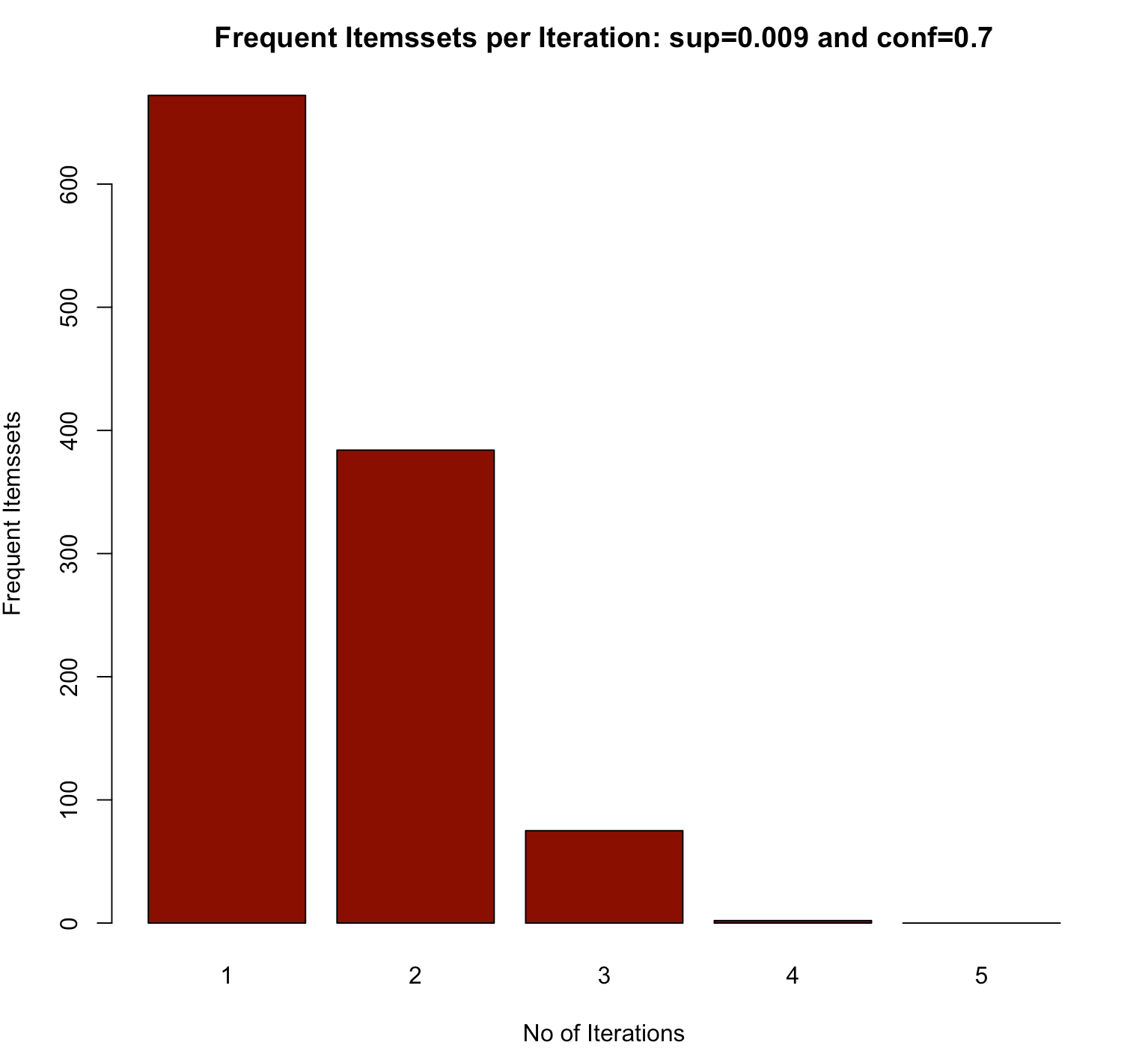
All in one:

A picture containing table

Description automatically generated

**Frequent itemsets:**

Chart, bar chart, histogram

Description automatically generated  Chart, bar chart

Description automatically generated

We have generated the Frequent itemsets by using the given mininum support and confidence values and plotted the variation of frequent itemsets over 5 iterations. The number of frequent itemsets were decreasing as expected as the number of iterations increase the itemset length also increases and chances of satisfying the minimum confidence value will become unlikely.

Below plot shows the frequent itemsets for all 3 minimum support and confidence values for all the iterations. We can see that as the min\_sup increases the number of frequent itemsets are decreasing for every iteration

Chart

Description automatically generated

**Rule generation:**

Association rules for the 9 combinations of all the 3 min\_sup and min\_conf values were generated and the scatter plots were made by taking min\_sup on xaxis and min\_conf on yaxis and the shading of the points tells us which rules have maximum lift value.

We can see that the number of rules are drastically reduced with increase in minimum support and confidence value.

supp = 0.006, conf = 0.5. supp = 0.009, conf = 0.7. supp = 0.014, conf = 0.8

(total: 1389) (total: 326) (total: 77)

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

Here are some of the sample rules generated:

A picture containing text

Description automatically generated

Chart, waterfall chart

Description automatically generated

The graph above represents all the rules generated for all 9 combinations of support and confidence in one context

**Filter 10 rules each for lift > 10, Lift < 10**

Rules generated in the above step for all 9 combinations of the min\_sup and min\_conf values are combined and unique rules from that combination were selected and the rules were sorted by lift value.

Then the we filtered the top 10 rules whose lift is >10 and also filtered the top 10 rules whose lift < 10

**lift > 10**

We have plotted an interactive scatter plot which has all the 10 points representing the top 10 rules which has lift>10, we can click on each point to check the rule and its lift, confidence and support values as shown below. The points which are darker represents the rules with high lift value and the points which are lighter represents the rules with low lift value.

Chart, scatter chart

Description automatically generated

**lift < 10**

We have plotted an interactive scatter plot which has all the 10 points representing the top 10 rules which has lift<10, we can click on each point to check the rule and its lift, confidence and support values as shown below

Chart, scatter chart

Description automatically generated

Lift>10. Lift<10

Chart, line chart

Description automatically generatedChart

Description automatically generated

We can see from the above graph that the rules with lift >10 are more clearly established than the rules with lift<10 as the lift measure for any rule ensures the high strength of association, hence we can consider that the rules with high lift value as shown in the graph are more likely to be bought together than the items in the rules which have lift <10.

Below are the rules generated for lift>10

Table

Description automatically generated

We can see that the items HERB MARKER BASIL,MINT,PARSLEY,ROSEMARY,THYME are occurring in many transactions together that means that we can suggest another value in this family such as HERB MARKER CHIVES to the customers and we can keep them together in the store.

Below are the rules generated for lift<10

A picture containing text, newspaper, document

Description automatically generated

Rules with lift<10 might also suggest some good associations as shown above where Lunch bags of different designs are often bought together by customers , in such cases we can suggest another lunch bag of another design to the customer and keep them together in the store but we need to note that the rules or associations with lift<10 does not exhibit strong associations and thus the items are less likely to be bought together compared to the items in the rules with higher lift value, in such cases strong domain knowledge and analysis efforts are needed.

**Visualization of top 100 Rules: Visualization of all the Rules:**

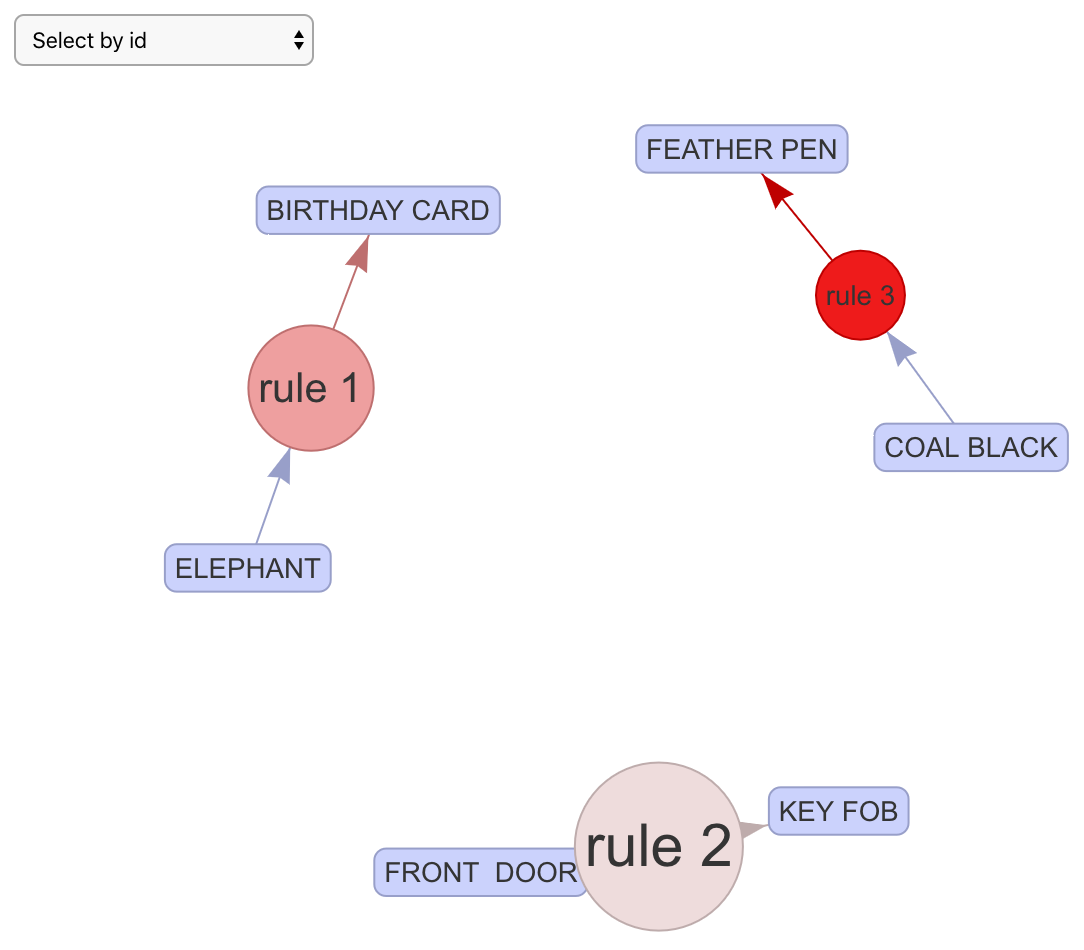
**Chart, scatter chart

Description automatically generated** **Chart, scatter chart

Description automatically generated**

In the above two maps we can see the rules of different orders in different colurs verying with respect to different confidence and support values. Rules(TOP 100) in the LHS are sorted in the descending order of min\_conf

**Interactive visualization: Order 2 rules (Plotting 3 rules)**

 A picture containing text

Description automatically generated

We can click on each rule to see the details of the rule such as confidence, we can see that the confidence value of all the 3 above rules are high because we have sorted the rules in the decreasing order of confidence and selected top 100 rules.

**Interactive visualization: Order 4 rules (Plotting 2 rules)**

Text

Description automatically generated Graphical user interface, text, application, chat or text message

Description automatically generated

Here we are plotting and visualizing 99th and 100th rules which have low confidence value as the rules are sorted in decreasing order of confidence, we can see that the rules have low confidence value

**CHALLENGES**

* Understanding the generation of candidate itemsets was challenging
* We faced below error while converting the data frame which is in single format to transactions

> txn1 = read.transactions(file="ItemList.csv",format="single",header = TRUE,rm.duplicates= FALSE,sep=",",cols=c(1,3))

Error in validObject(.Object) :

invalid class “ngTMatrix” object: all row indices (slot 'i') must be between 0 and nrow-1 in a TsparseMatrix

In addition: Warning message:

In scan(file = file, sep = sep, quote = quote, what = what, flush = TRUE, :

EOF within quoted string

Fix:

There were incomplete records present in the data which we eliminated using complete.cases command and used different approach to convert the dataframe into transactions.

**FILE NAMES**

* Association\_rules.r : This r file contains the complete code for this project
* mb\_transactions.csv: Intermediate CSV file containing basket format data of each transaction

**Division of Labor**

Amrutha: I have handled the preprocessing, Itemset generation and rule generation. I have done the analysis part and prepared the report for final submission.

Ajay: Worked on preprocessing and handled visualization part for top 100 rules and rules based on lift.