Submitted by Team 2

Aparna Krishna Bhat | Surbhi Zambad

Data Mining: Project 3

Association Rule Mining

Association rule mining finds interesting association relationships among a large set of data items. With massive amounts of data continuously being collected and stored in databases, many industries are becoming interested in mining association rules from their databases. A typical example of association rule mining is market basket analysis. This process analyzes customer-buying habits by finding associations among the different items that customers place in their shopping baskets. The discovery of such associations can help retailers develop marketing strategies by gaining insight into which items are frequently purchased together by customers.

**Project Overview**

* The UK based retail store dataset consisting of **500 K transactions** each having several items was downloaded and imported to R studio as a xlsx file for analysis.
* The task was to perform market basket analysis on the transactions.
* Pre-processing was done using R (read\_xl, dplyr libraries).
* The data was preprocessed to extract basket for each transaction and items bought together in that transaction to perform the analysis.
* **Minimum support values provided are 0.003, 0.007, 0.012 and minimum confidence values provided are 0.5, 0.7, 0.8.**
* Rules were generated for the given min\_sup\_values and min\_conf\_value with minlen and maxlen using lift as the measure.
* Redundant rules were pruned.
* Rules are plotted using grouped, paracoord, two-key plot.
* Top 15 items (initial data exploration) are plotted using relative, absolute frequency type using itemFrequencyPlot().

**Libraries Used**

1. read\_xl
2. dplyr
3. tm
4. tidyverse
5. lubridate
6. plyr
7. arules
8. arulesviz

**Data Pre-Processing**

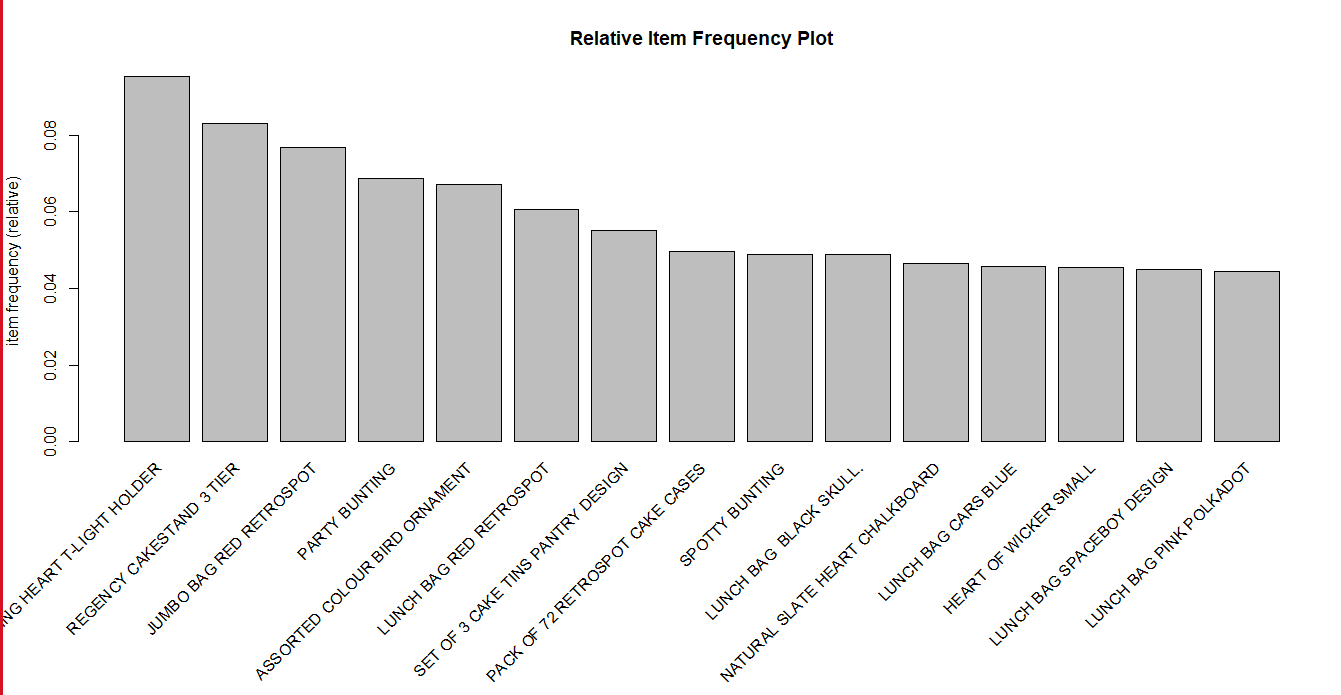
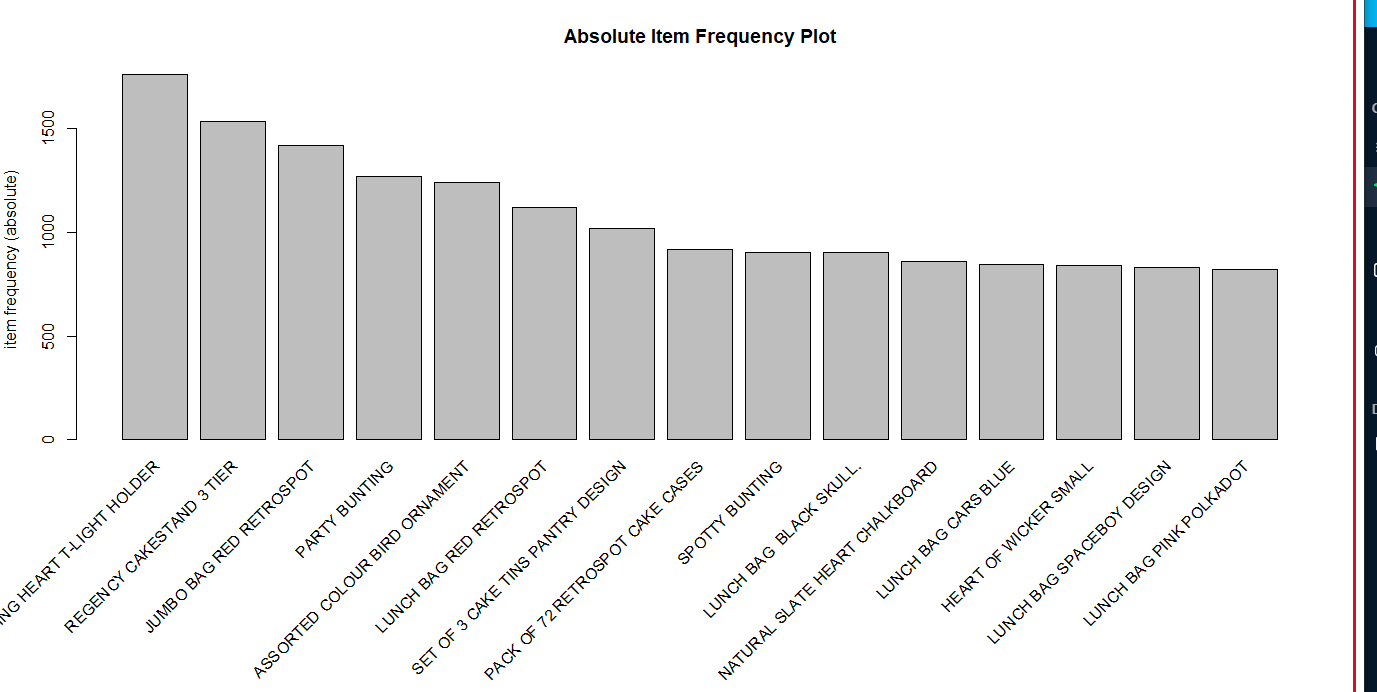
1. read\_xl() to read the data set that was provided with 8 attributes.
2. Used **complete.cases(data)** function to return rows which have no missing values.
3. Discarded InvoiceNo that starts with a C using **grepl().**
4. Used the stop words list to discard the transactions that contained any of the stop words in the description indicating that they are NOT items bought. E.g., of **stop words used are** **"WRONG","LOST", "CRUSHED", "SMASHED", "DAMAGED"** etc.
5. Used Omit function to omit the NAs caused due to coercion.
6. Used Mutate function to add new columns into the data-frame. Converted Description, country columns to factor columns.
7. From InvoiceDate column date and time were separated using format().
8. Used cbind to bind TransTime and InvoiceNo into a data-frame.
9. For Txs combined all products from one InvoiceNo and date and combined all products from that InvoiceNo and date as one row, with each item, separated by a comma(,).
10. InvoiceNo and Date are not of any use in the rule mining, thus set them to NULL.
11. Storing basket format transaction data into a .csv file for further analysis.

**Analysis**

There are **18485 transactions (rows), 7793 items (columns),** and a **density of 0.002277027** i.e. 7793 is the product descriptions involved in the dataset and 18485 transactions are collections of these items.

Density tells the percentage of non-zero cells in a sparse matrix [A sparse matrix or sparse array is a matrix in which most of the elements are zero].

An initial exploration of the data (Using top 15 items) we can see that the most purchased items are T-Light Holder, followed by Cake stand 3 Tier, while the least is Lunch bag.

**Fig:- Relative Frequency(items) plot Fig:- Absolute Frequency (items) plot**

Item frequency (relative) shows how many times these items have appeared as compared to others. Item frequency (absolute) shows the numeric frequencies of each item independently.

The Apriori algorithm was used to detect association rules. In the arules package, apriori takes two parameters: minimal support and minimal confidence. The default values for support and confidence are 0.1 and 0.8, respectively. The apriori method returns a set of rules. The number of rules produced depends on the minimal support count and minimal confidence specified.

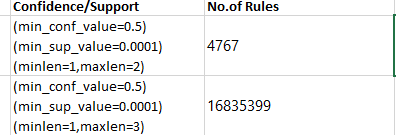
**Support**: This measure gives an idea of how frequent an itemset is in all the transactions.

**Support({X} -> {Y}) = Transaction containing both X and Y / Total number of transactions**. Value of support helps us identify the rules worth considering for further analysis

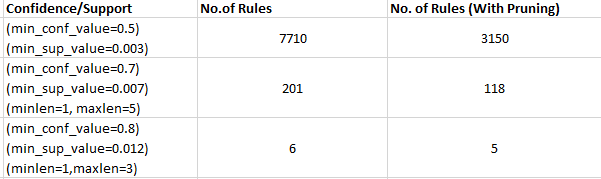
**Confidence**: This measure defines the likeliness of occurrence of consequent on the cart given that the cart already has the antecedents.

**Confidence({X} -> {Y}) = Transaction containing both X and Y / Total number of transactions containing.**

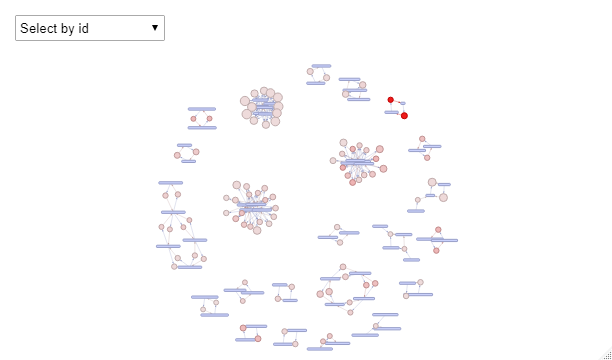
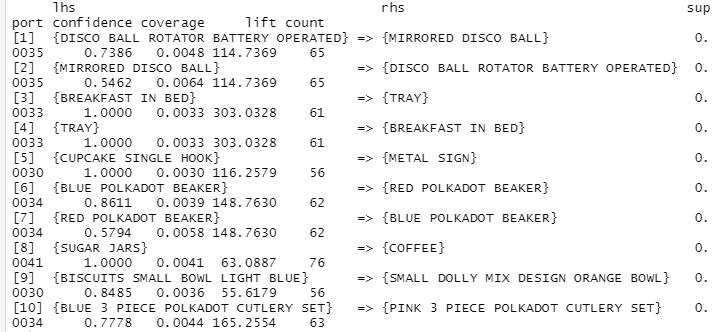
**Lift {support/support(x) \* support(y), range: [0, inf]}** :- indicates the strength of a rule over the random occurrence of x and y. It explains the strength of a rule and more the Lift more is the strength.



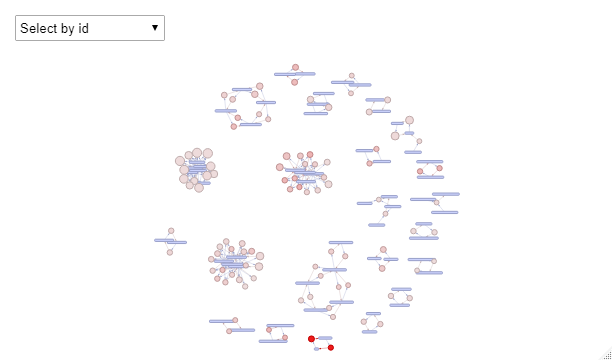
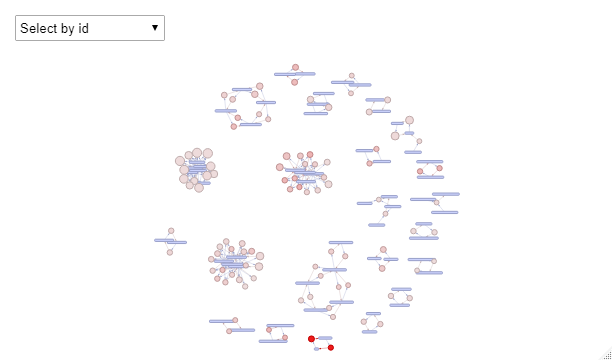
Candidate itemset(s) are generated using very low support 0.001 and taking different maxlen value. For minlen and maxlen both set to 1, 0 rules were obtained and then maxlen value were incremented in step of 1. For maxlen=1 and maxlen=3, number of rules obtained are tabulated as above.



Since lift is a good interest measure for correlation, the rules are sorted according to lift. Redundant rules that provide no additional knowledge are pruned That is, if a rule A is a super rule of another rule B and rule A has a lower lift, then the former rule, rule A, is considered to be redundant. This helps in reducing the size of the subset and improves the quality of the results.

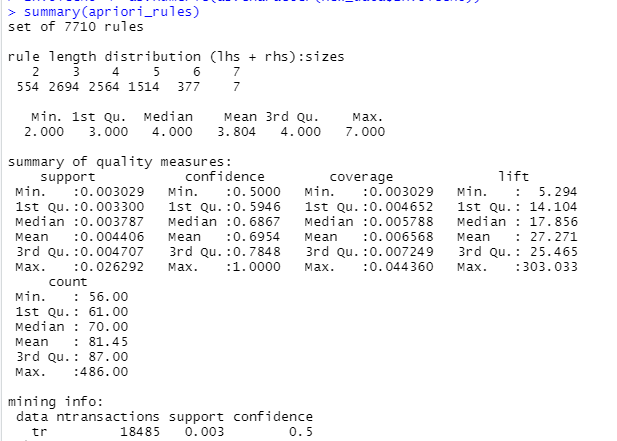
**Fig:- Rules without pruning for min\_ conf=0.5, min\_sup =0.003**

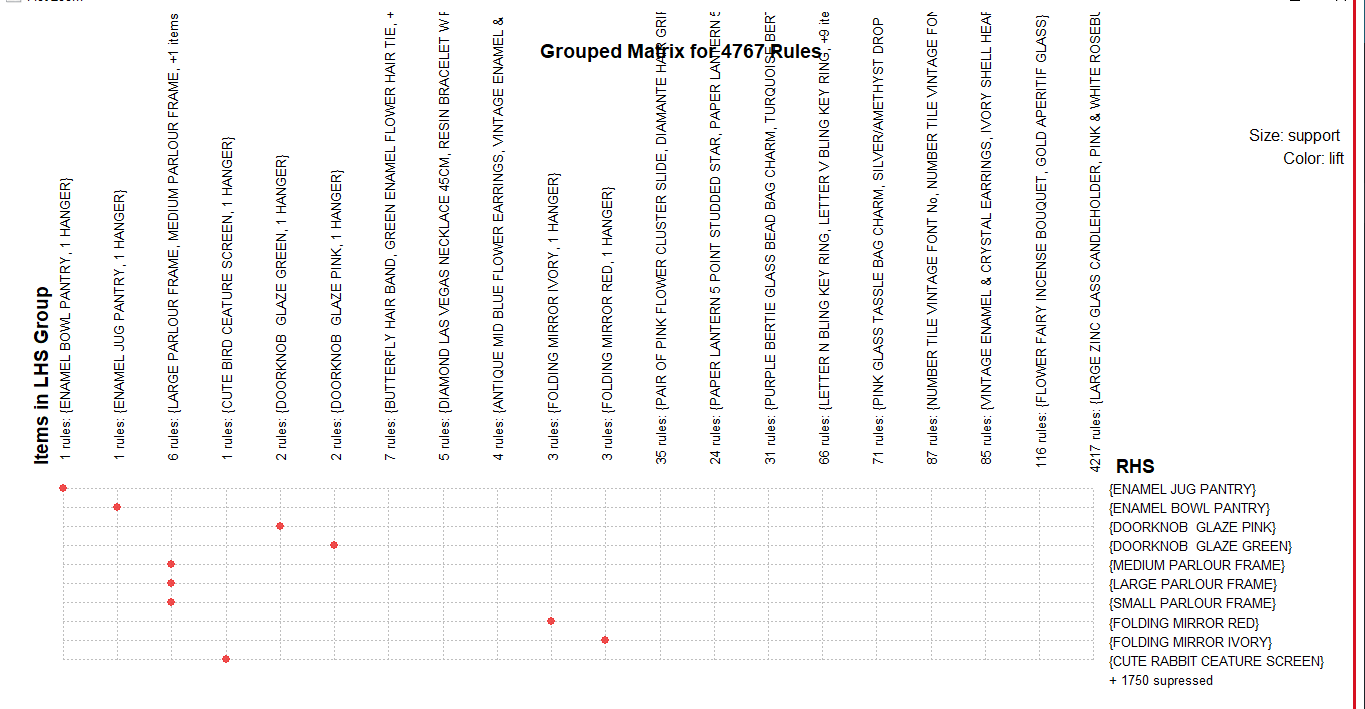
**Fig:- Rules with pruning Fig:- Rules without pruning by confidence**

The first trial produced a set of 7710 association rules sorted according to lift values. Pruning redundant rules is done in the second trial which reduces the number of rules from 7710 to 3150. Pruning provides better quality of results which can be easily seen by comparing the graphs in the above figures. It is obvious that the figure of trial 2 (with pruning) is less cluttered and provides a better visual representation of the rules.

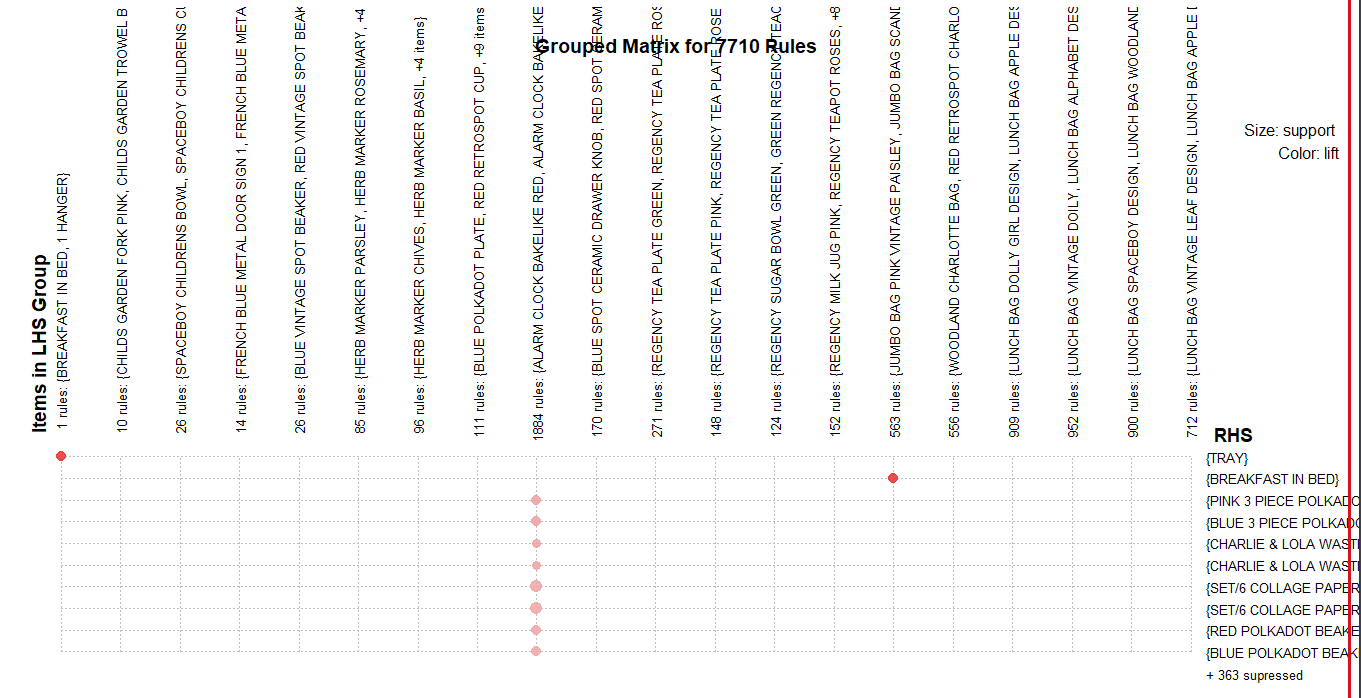
Summary for apriori\_rules are shown below



For length of 3 items has the highest number of rules: **2694** whereas length of 7 items has the least number of rules : **7.** Using the inspect functionwe analyzed that 100% of customers who bought {BREAKFAST IN BED} also bought {TRAY} and customers who bought {CUPCAKE SINGLE HOOK} also purchased {METAL SIGN}



**Fig:- Group Matrix for 4767 Rules for min\_conf\_value=0.5, min\_sup\_value=0.00001**



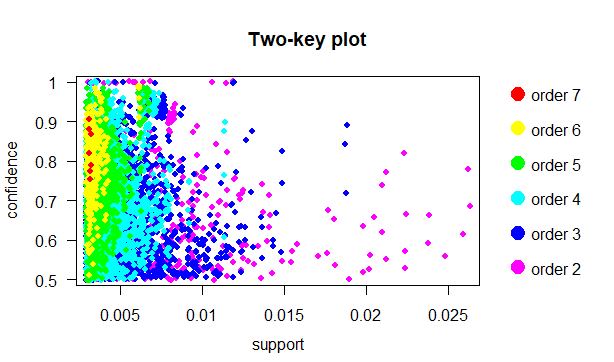
**Fig:- Group Matrix for 7710 Rules for min\_conf\_value=0.5, min\_sup\_value=0.003**

The groups of Matrix show the association's rules found order by lifts. The color of the lift bubble represents the interestingness of the rule.

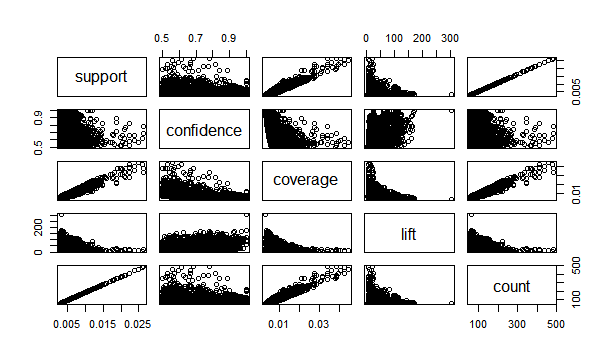


**Fig:- parallel co-ordinate for 7710 Rules for min\_conf\_value=0.5, min\_sup\_value=0.003**

The above fig shows the association rules using vertices and edges where vertices are labeled with item names, and item sets or rules are represented as a second set of vertices. Items relate to item-sets/rules using directed arrows. Arrows pointing from items to rule vertices indicate LHS items and an arrow from a rule to an item indicates the RHS.



The two-key plot uses support and confidence on x and y-axis, respectively. It uses order for coloring. The order is the number of items in the rule. It maps the relation between confidence and support.



The association's rules parameters matrix shows an overview of the relationship among all the parameters such as Support, Confidence, lift, coverage, and count.

**Rules Filtering by Lift**

10 rules each for lift > 10 (min\_conf\_value=0.5, min\_sup\_value=0.003)

**303.03279, 303.03279, 172.75701, 168.02394, 165.25543, 165.25543, 163.64990, 162.90974, 161.99687, 156.4709**

10 rules each for lift <10 (min\_conf\_value=0.5, min\_sup\_value=0.003)

**6.517983, 6.517983, 6.500865, 6.473476, 6.454312, 6.376857, 6.335589, 6.282349, 5.387777, 5.29407758**

10 rules each for lift > 10 (min\_conf\_value=0.7, min\_sup\_value=0.007, minlen=1, maxlen=5)

**111.2111, 110.0197, 109.8788, 109.0709, 109.0016, 108.4018, 107.8145, 107.7214, 107.6939**

10 rules each for lift < 10 (min\_conf\_value=0.7, min\_sup\_value=0.007, minlen=1, maxlen=5)

**9.8144, 9.6481, 9.5950, 9.5939, 9.4916, 9.4251, 9.2437, 9.1570, 7.6160**

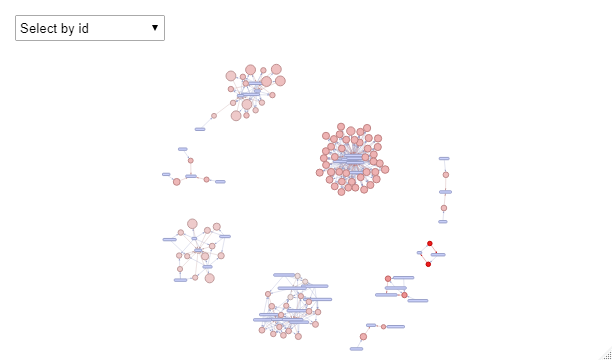
10 rules each for lift > 10 (min\_conf\_value=0.8, min\_sup\_value=0.012)

**26.36558 25.86681 24.23733 22.23028 21.94649 21.51050**

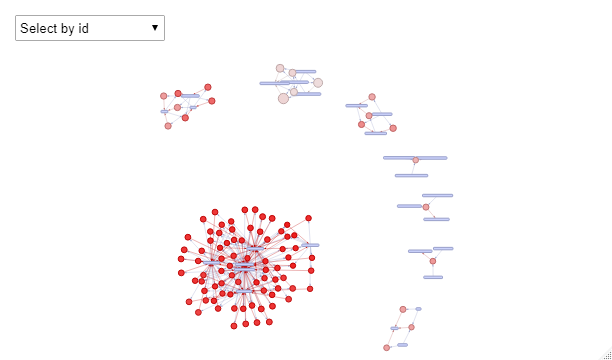
A value of lift greater than 1 vouches for high association between {Y} and {X}. The larger the lift ratio, the more significant the association. Rules with a high lift value, which means that it occurs more frequently than would be expected given the number of transaction and product combinations.

We also analyzed that as min\_conf\_value and min\_sup\_value increases more stronger rules gets generated.

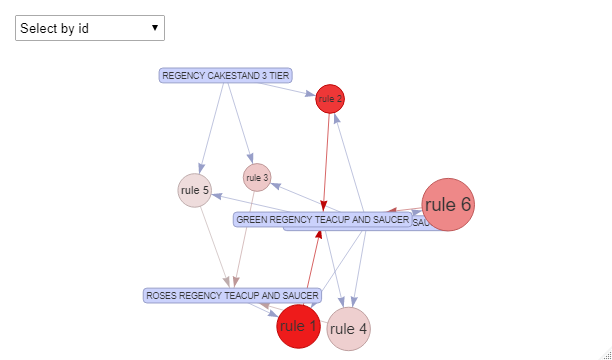
**Top Rules according to Minimum Confidence value**



**Fig:- Top 100 rules for min\_conf\_value=0.5, min\_sup\_value=0.003**



**Fig:- Top 100 rules for min\_conf\_value=0.7, min\_sup\_value=0.007**



**Fig:- Top rules for min\_conf\_value=0.8, min\_sup\_value=0.012**

**CHALLENGES**

1. Handling huge dataset.
2. Generating candidate itemset for very low support with higher maxlen.
3. Using different methods to plot and visualize various graphs.

**FILE NAMES**

1. R\_files\_team2 folder contains all the R source file used in the project.
2. visualization\_team2 folder has all the graph plots and analysis. (Many visualizations excluded from the report due to page limitation are present in this folder).
3. Main report file