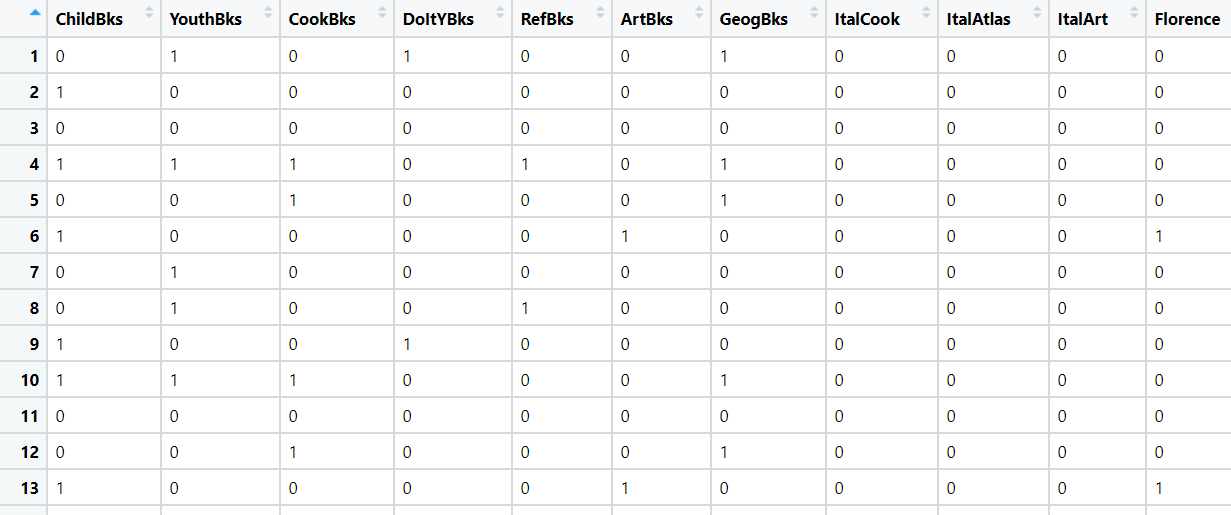
# Topic: Association Rules

## Prepare rules for the all the data sets

1. Try different values of support and confidence. Observe the change in number of rules for different support, confidence values
2. Change the minimum length in apriori algorithm
3. Visualize the obtained rules using different plots

**1.) Books.csv**



**Ans:**

**Data Preprocessing:**

* Loading the dataset.
* The data is already factorized and there no further feature engineering process is not required

**Making Rules using Apriori Algorithm:**

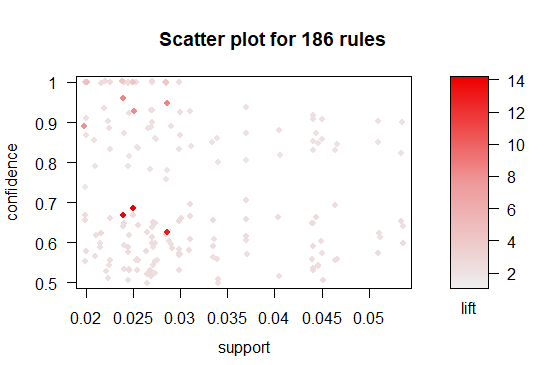
* Using the experimental value of support and confidence by taking min length = 5.
* Provided the rules with 2 % Support, 50 % Confidence and Minimum to purchase

5 books.

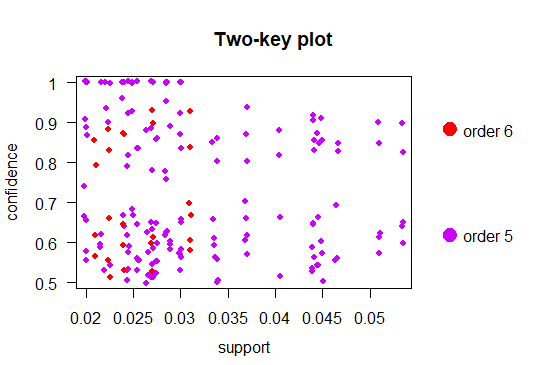
* By the experimental value 186 rules are formed.
* Using inspect function, based on the lift values below are the first 6 rules of the model

|  |
| --- |
| **lhs rhs support confidence coverage lift count**  [1] {CookBks,DoItYBks,ArtBks,ItalCook} => {ItalArt} 0.0250 0.6849315 0.0365 14.122299 50  [2] {CookBks,ArtBks,GeogBks,ItalCook} => {ItalArt} 0.0240 0.6666667 0.0360 13.745704 48  [3] {ChildBks,CookBks,ArtBks,ItalCook} => {ItalArt} 0.0285 0.6263736 0.0455 12.914920 57  [4] {CookBks,ArtBks,GeogBks,ItalArt} => {ItalCook} 0.0240 0.9600000 0.0250 8.458150 48  [5] {ChildBks,CookBks,ArtBks,ItalArt} => {ItalCook} 0.0285 0.9500000 0.0300 8.370044 57  [6] {CookBks,DoItYBks,ArtBks,ItalArt} => {ItalCook} 0.0250 0.9259259 0.0270 8.157938 50 |

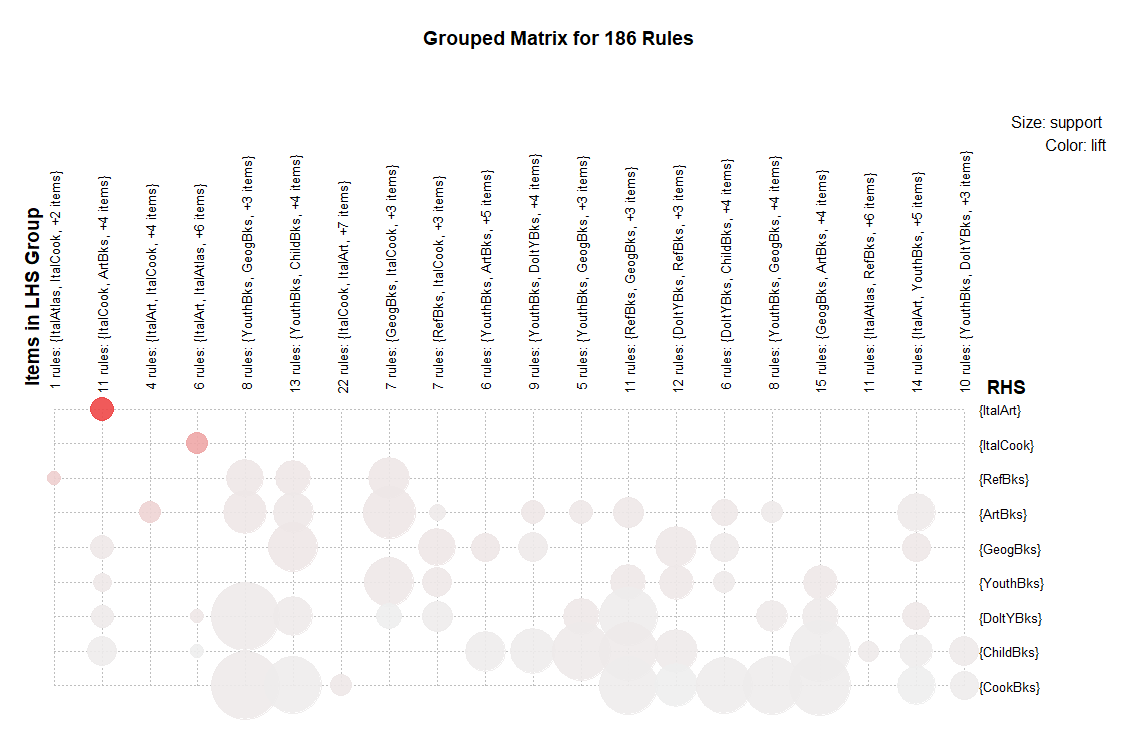
* Checking the over all quality of the rules which gives support, confidence, coverage, lift and count values for all the 186 rules set.
* Using library(“arulesViz”), we can do the visualization of the rules



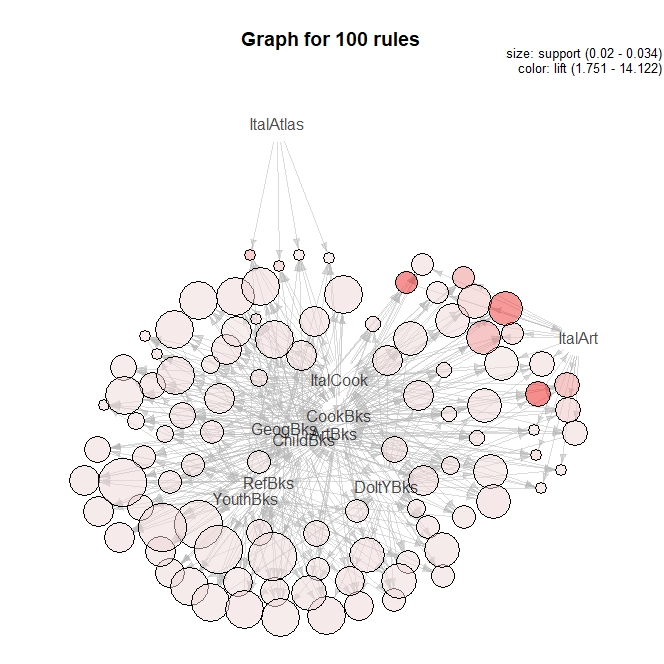
* As there are no label represented in the scatter plot its very difficult to interpret the analysis of visualization



* The order 5 has been scatter more than order 6

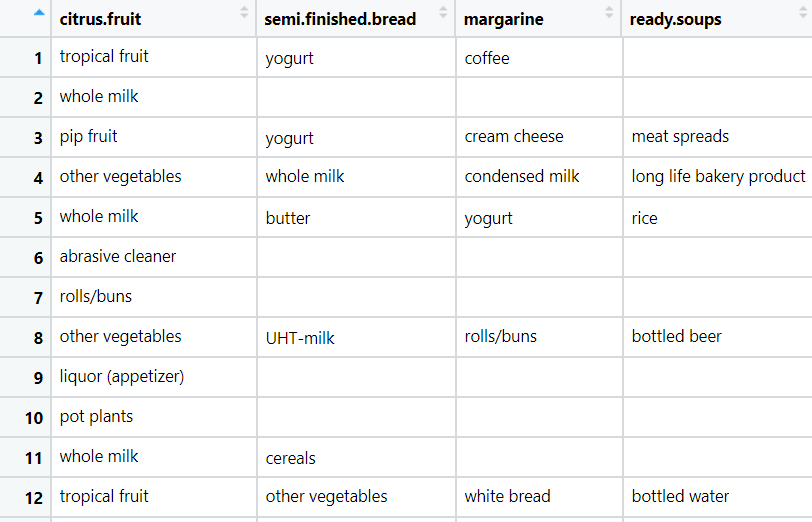


* The Art books are being sold at a larger extent along with other Cook, art, geo, child books
* Cook books are also being sold at a larger extent along with other chld, art, geo, Doit books)



* The above graph is created with 100 rules with the experimental values at 2% support and 50% confidence and no of books sold is 5.

**2.) Groceries.csv**



**Ans:**

**Data Preprocessing:**

* Data set is imported from Text(base), so the data will convert to factors
* The data is already factorized and there no further feature engineering process is not required
* There are no NA values exsits

**Making Rules using Apriori Algorithm:**

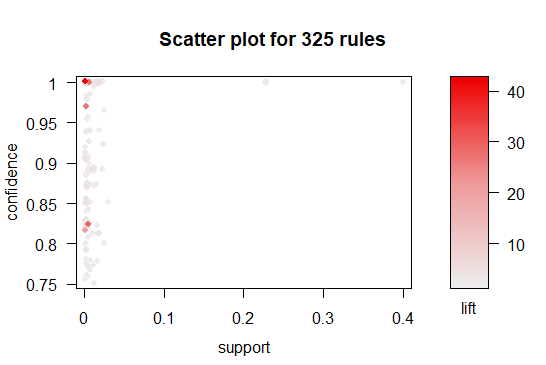
* Using the experimental value of support and confidence by taking min length = 5.
* Provided the rules with 0.02, Support, 75 % Confidence and Minimum to purchase

Is 2 products

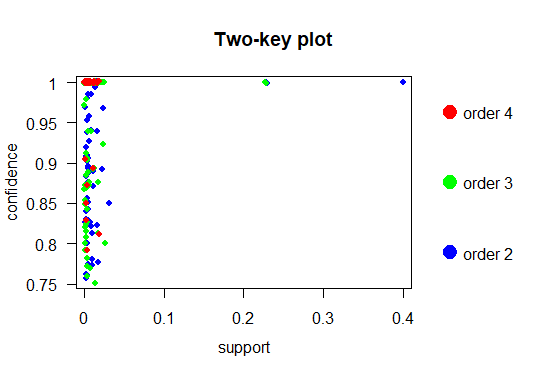
* By the experimental value 325 rules are formed.
* Using inspect function, based on the lift values below are the first 6 rules of the model
* Checking the overall quality of the rules which gives support, confidence, coverage, lift and count values for all the 325 rules set.

|  |
| --- |
| **Support confidence coverage lift count**  1 0.002026806 0.9687500 0.002092187 1.803217 31  2 0.002222949 0.8717949 0.002549853 1.622746 34  3 0.002222949 0.8717949 0.002549853 1.622746 34  4 0.002222949 0.8500000 0.002615234 3.730488 34  5 0.002615234 1.0000000 0.002615234 2.501226 40  6 0.002615234 1.0000000 0.002615234 1.861385 40 |

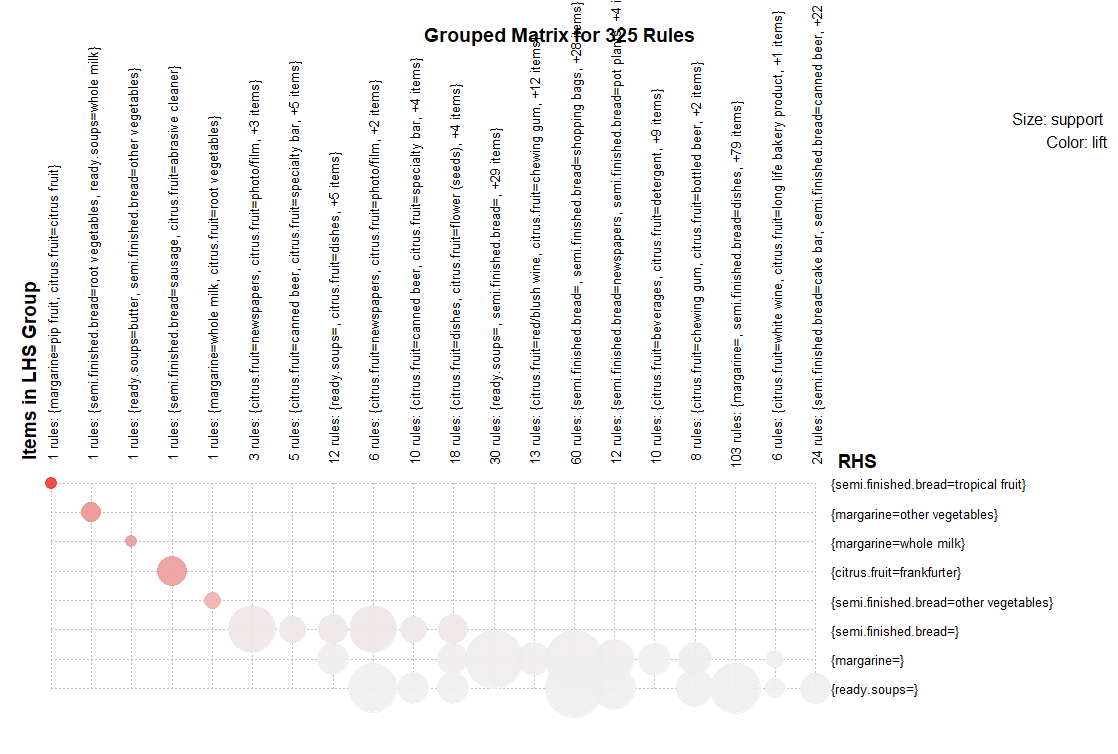
* Using library(“arulesViz”), we can do the visualization of the rules



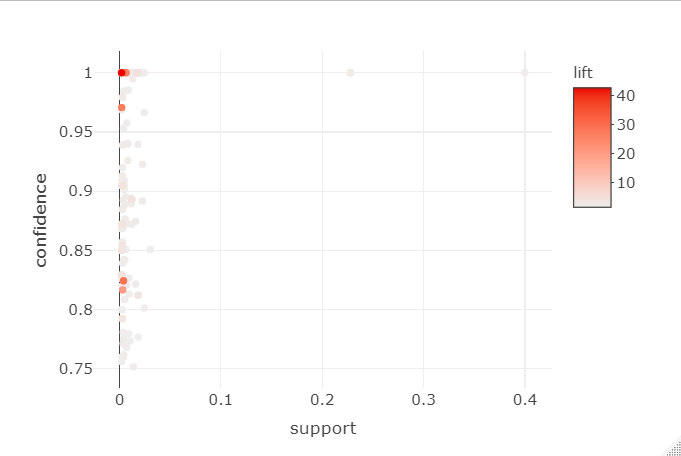
* As there are no label represented in the scatter plot its very difficult to interpret the analysis of visualization

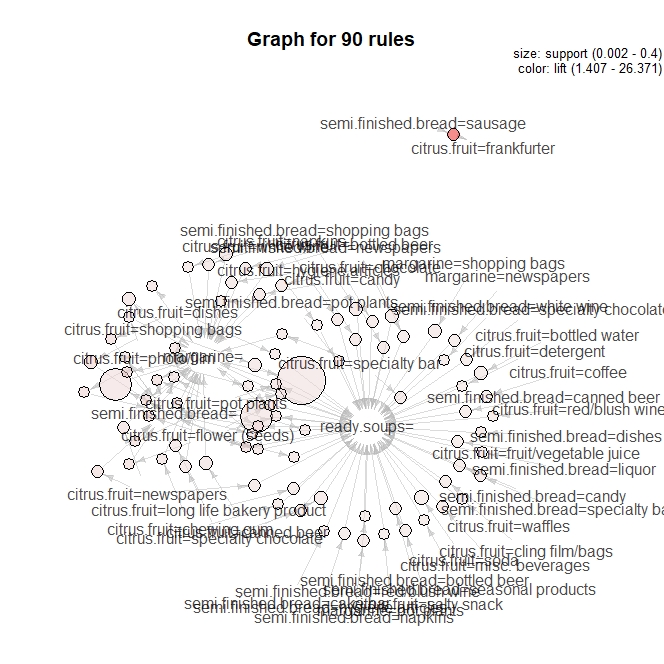
****

* The graph shows the order 2 and 3 are more scattered then order 4.



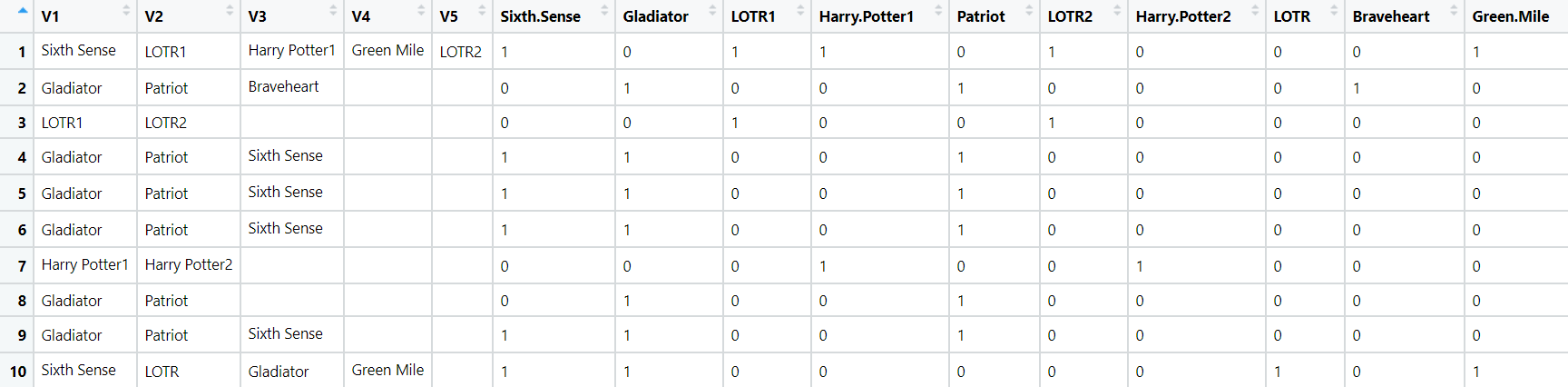
* The semi furnished bread are being sold at a larger extent along with other tropical fruits, citrus fruits.

****

****

* The above graph is created with 90 rules with the experimental values at 0.02 support and 75% confidence and no of products sold is 2 .

**3.) my\_movies.csv**



**Ans:**

**Data Preprocessing:**

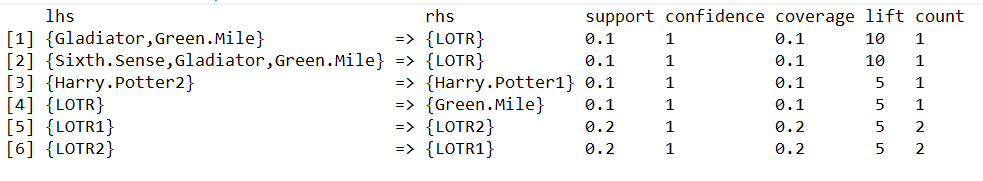
* Loading the dataset.
* The data is already factorized and there no further feature engineering process is not required

**Making Rules using Apriori Algorithm:**

* Using the experimental value of support and confidence by taking min length = 5.
* Only 6th to 15th columns are considered as they are results of factorized values of first five columns
* Provided the rules with 2 % Support, 50 % Confidence and Minimum to purchase

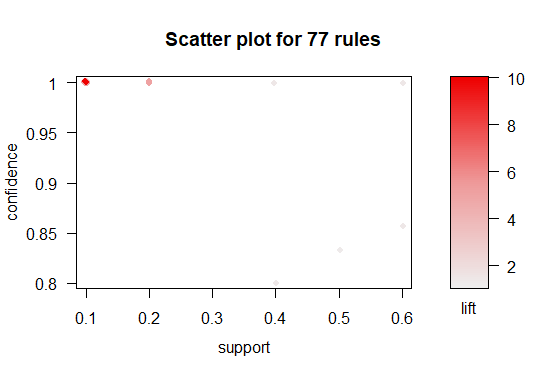
5 movies.

* By the experimental value 77 rules are formed.
* Using inspect function, based on the lift values below are the first 6 rules of the model

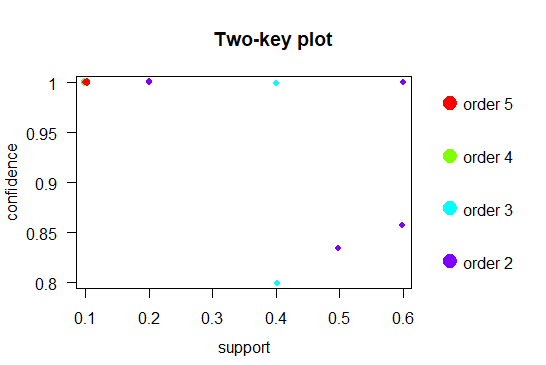


* Checking the overall quality of the rules which gives support, confidence, coverage, lift and count values for all the 77 rules set

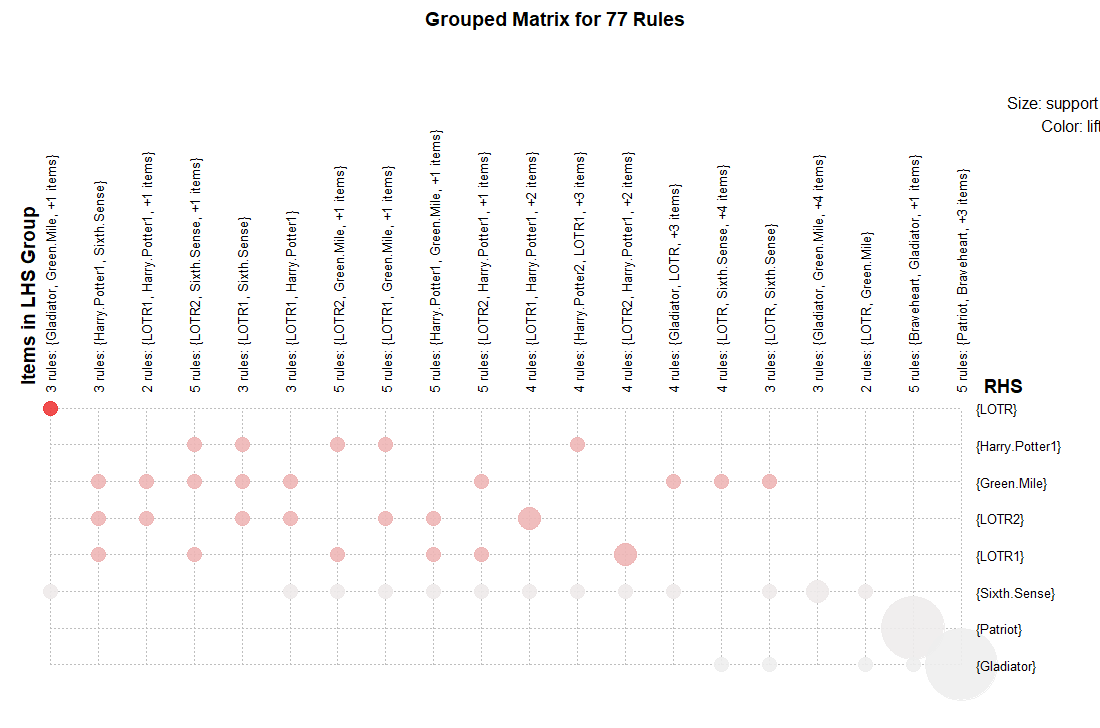
**Graphical Visualization:**



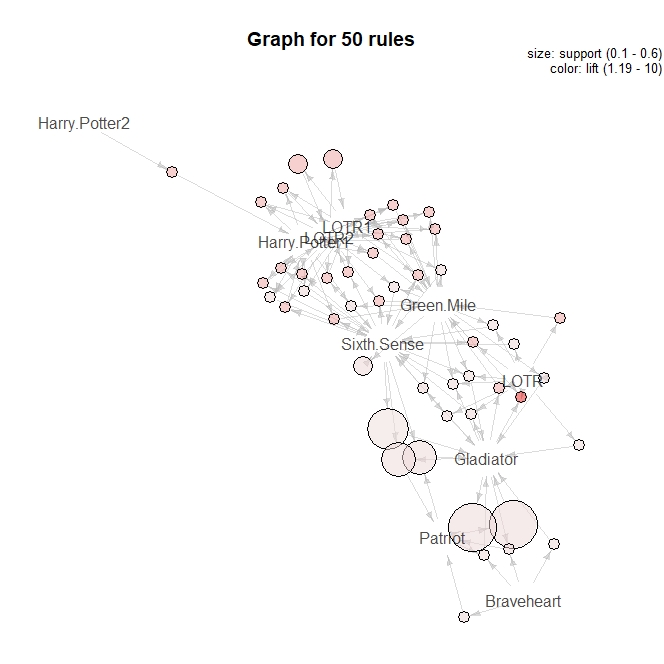
* As there are no label represented in the scatter plot it’s very difficult to interpret the analysis of visualization



* The plots show that all orders are scattered in same frequency



* The matrix shows the most watched movies are LOTR along with Gladiator and Green mile
* The next most watched is Harry potter 1 along with sixth sense.



**4.) myphonedata.csv**



**Ans:**

**Data Preprocessing:**

* Loading the dataset.
* The data is already factorized and there no further feature engineering process is not required

**Making Rules using Apriori Algorithm:**

* Using the experimental value of support and confidence by taking min length = 2.
* Only 4th to 9th columns are considered as they are results of factorized values of first three columns
* Provided the rules with Support 0.02, 75 % Confidence and Minimum to purchase

2 phones.

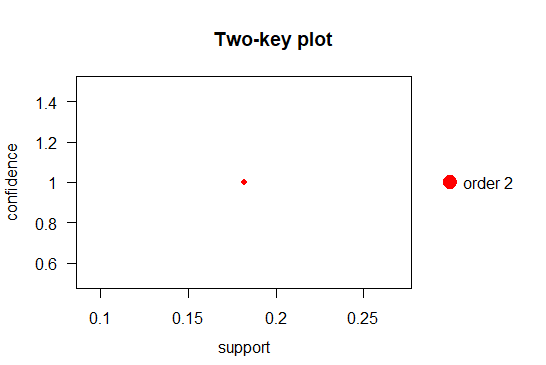
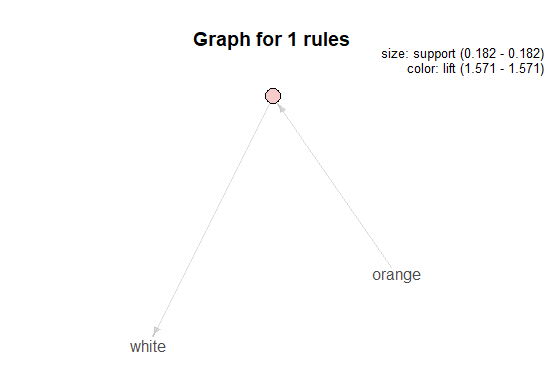
* By the experimental value 1 rules are formed.
* Using inspect function, based on the lift values below are the first 6 rules of the model

lhs rhs support confidence coverage lift count

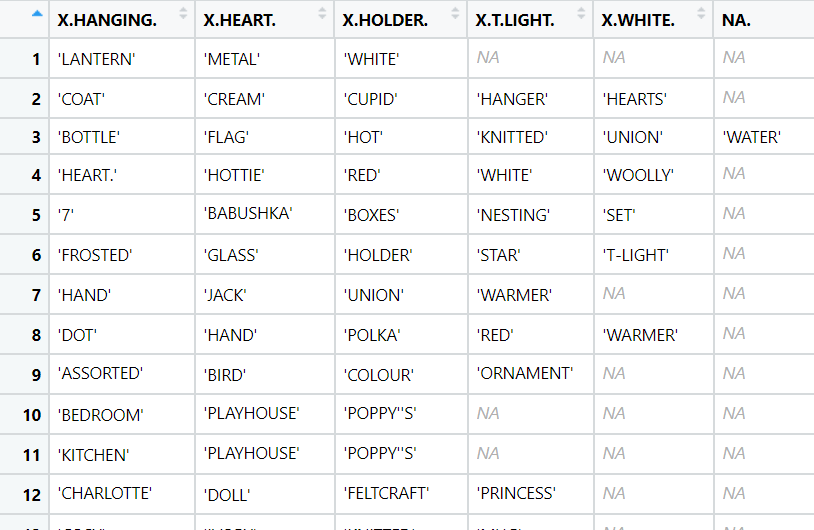
[1] {orange} => {white} 0.1818182 1 0.1818182 1.571429 2

* Checking the overall quality of the rules which gives support, confidence, coverage, lift and count values for all the 1 rule set

**Graphical Visualization:**

****

**5.) transaction\_retail.csv**



**Ans:**

**Data Preprocessing:**

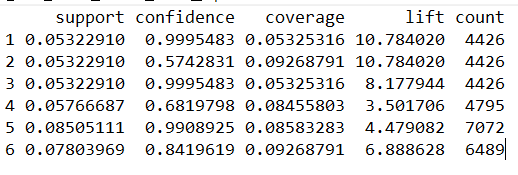
* Data set is imported from Text(base), so the data will convert to factors
* The data is already factorized and there no further feature engineering process is not required
* Using complete. Cases, we Cleans the non-missing values

**Making Rules using Apriori Algorithm:**

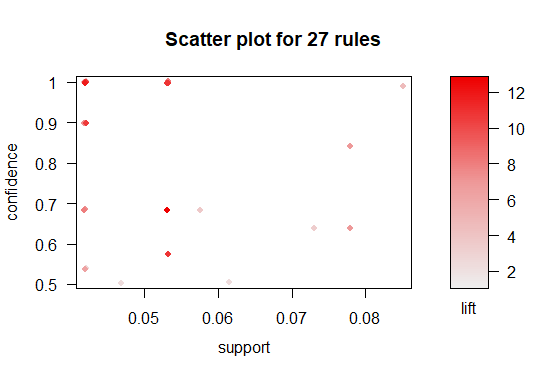
* Using the experimental value of support and confidence by taking min length = 2.
* Provided the rules with 0.04, Support, 50 % Confidence and Minimum to purchase

Is 2 products

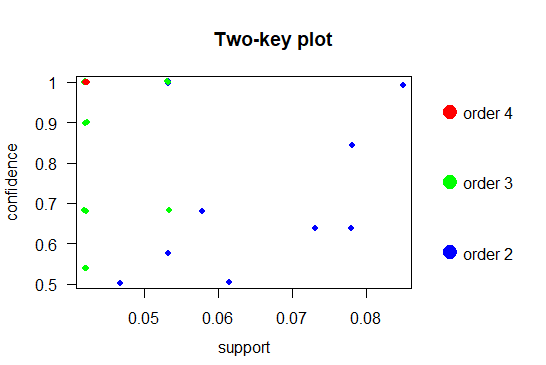
* By the experimental value 27 rules are formed.
* Using inspect function, based on the lift values below are the first 6 rules of the model
* Checking the overall quality of the rules which gives support, confidence, coverage, lift and count values for all the 27 rules set.



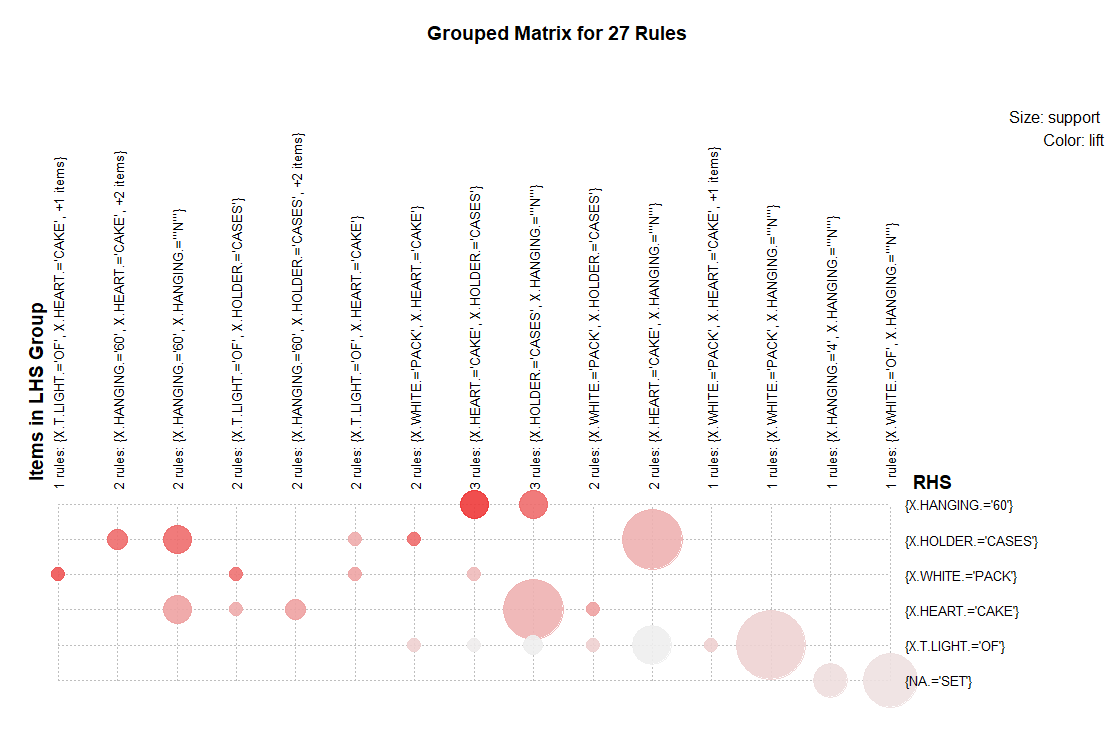
**Graphical Visualization:**

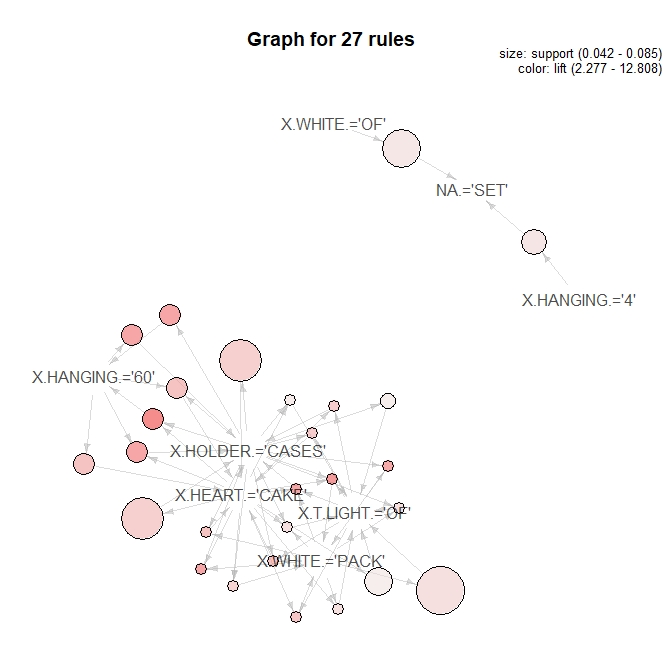


* As there are no label represented in the scatter plot it’s very difficult to interpret the analysis of visualization



* The plots show that all orders are scattered in same frequency





* The above graph is created with 90 rules with the experimental values at 0.04 support and 50% confidence and no of products sold is 2

# Hints:

1. Business Problem
   1. Objective
   2. Constraints (if any)
2. Data Pre-processing

2.1 Data cleaning, Feature Engineering, EDA etc.

1. Model Building
   1. Partition the dataset
   2. Model(s) - Reasons to choose any algorithm
   3. Model(s) Improvement steps
   4. Model Evaluation
   5. Python and R codes
2. Deployment

4.1 Deploy solutions using R shiny and Python Flask.

1. Result Share the benefits/impact of the solution - how or in what way the business (client) gets benefit from the solution provided.

**Note:**

1. For each assignment the solution should be submitted in the format
2. Research and Perform all possible steps for improving the model(s) accuracy Ex: Feature Engineering, Hyper Parameter tuning etc.
3. All the codes (executable programs) are running without errors
4. Documentation of the module should be submitted along with R & Python codes, elaborating on every step mentioned here