Market Basket Analysis for Corse Topics CSV File

Didem Aykurt

2022-12-24

Market Basket Analysis

Market basket analysis is an essential technique for market research that helps identify which products to purchase together for each transaction or, if the customer buys the product, identify which product strongly relates to purchasing or is recommended to buy. Market basket analysis, like recommended products, advertisements, shopping online, identifying the relational group of people who have identical product buy, marketing promotion, coupon offers, and so on. All this marketing strategy to increase the cross-selling and up-selling of products and decrease the time customers spend on product research also eliminates risk in a future business plan. There are three types of recommender systems:

The Knowledge-base Recommendations system helps to predict the group of products bought together. The collaborative Filtering technique uses the data for customer preferences and remanent products to a similar type of buyer.

The Content-based technique uses historical purchase data to predict and recommend a similar product.

Market basket analysis has a different model to analyze the relationship between products and one of the models is association rules the following step: IF [antecedent] THEN [consequent] these technics set of rules if “buy this” then “high relation product to buy that” Association rules have three types of measures:

Association rules have three types of measures:

Support helps to measure the frequency group of products bought.

$$Support = \hat{P}(antecent\ AND\ consequent)$$

Confidence measures the probability that the customer will buy an antecedent product rather than a consequent product as a conditional probability.

$$Confidence = \frac{\hat{P}(antecentANDconsequent)}{\hat{P}(antecedent)}$$

The Lift measures the ratio of confidence to expected confidence so helps to increase confidence.

$$Benchmarkconfidence = \frac{no.transaction\ with\ consequent\ itemset}{no.transections\ in\ database}$$

A result lift ratio is a considerable value that means a high probability of buying the consequent product.

Market Basket Analysis for CorseTopics CSV

The CorseTopics dataset included 365 observations with 8 variables intro, Datamining, Survey, Cat.Data, Regression, Forecast, DOE, and SW.Each that all the topics for one course.

```
# Install and Load Packages for market basket analysis
#install.packages("pacman")
library("pacman")
pacman::p_load(arules, arulesViz)
```

LOAD and prepare CourseTopics CSV data

```
marketBas.df<-read.csv("C:/Users/didem/OneDrive/Documents/CSUG Master
DA/MIS510-1 Data Mining_4 term/Module 6 Market basket
Analysis/Coursetopics.csv", header=TRUE)
```

Check to null object(result False means there isn't null)

```
is.null(marketBas.df)

## [1] FALSE
```

Print the list in a useful column format

```
t(t(names(marketBas.df)))

##      [,1]
## [1,] "Intro"
## [2,] "DataMining"
## [3,] "Survey"
## [4,] "Cat.Data"
## [5,] "Regression"
## [6,] "Forecast"
## [7,] "DOE"
## [8,] "SW"

#show the first nine rows
head(marketBas.df, 9)

##   Intro DataMining Survey Cat.Data Regression Forecast DOE SW
## 1     1          1      0        0          0        0   0  0
## 2     0          0      1        0          0        0   0  0
## 3     0          1      0        1          1        0   0  1
## 4     1          0      0        0          0        0   0  0
## 5     1          1      0        0          0        0   0  0
## 6     0          1      0        0          0        0   0  0
## 7     1          0      0        0          0        0   0  0
```

```
## 8     0          0      0        1          0        1   1  1
## 9     1          0      0        0          0        0   0  0

#find summary statistics for each column
summary(marketBas.df)

##      Intro          DataMining         Survey          Cat.Data
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.3945   Mean   :0.1781   Mean   :0.1863   Mean   :0.2082
##  3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##    Regression        Forecast           DOE               SW
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.2082   Mean   :0.1397   Mean   :0.1726   Mean   :0.2219
##  3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
```

Remove first column and convert to matrix

```
marketBas.mat<- as.matrix(marketBas.df[,-1])
```

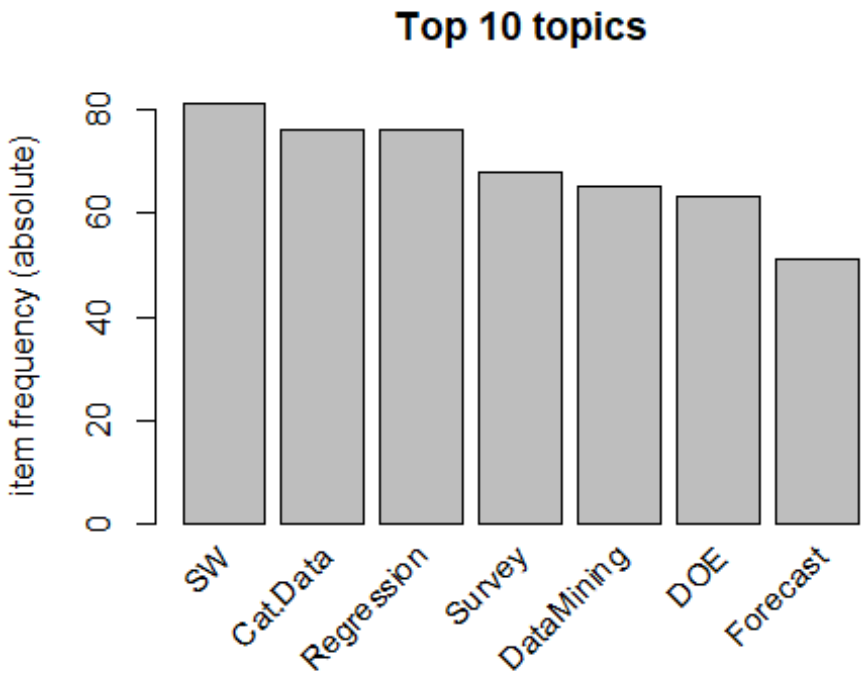
Convert the binary incidence matrix into a transactions database

```
marketBas.trans<- as(marketBas.mat,"transactions")
inspect(head(marketBas.trans))

##     items
## [1] {DataMining}
## [2] {Survey}
## [3] {DataMining, Cat.Data, Regression, SW}
## [4] {}
## [5] {DataMining}
## [6] {DataMining}
```

Plot the Sort topics

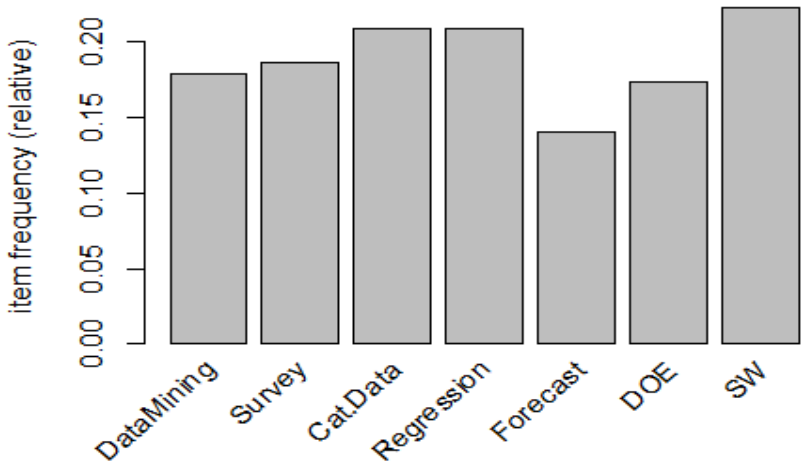
```
itemFrequencyPlot(marketBas.trans, topN=7, type="absolute", main="Top 10
topics")
```



The above graph tells highest to lowest frequency topics that helps to focus and structure as the higher to lower SW, Cat.Data, Regression, Survey, DataMining, DOE,and Forecast.

Topic Frequncy of 1%

```
itemFrequencyPlot(marketBas.trans, support=0.01)
```



The above chart shows all the topics having take frequency of 14% and above. And the top three topics are SW, Cat.Data , and Regression.

Association rules

when running the apriori() function in R, including the minimum support, minimum confidence, and target as arguments. And this case's minimum support is 0.01, with minimum confidence of 0.1. I will build the basket and find which topics significantly occur to understand which topics are preferable.

```
rules<- apriori(marketBas.trans, parameter = list( supp=0.01, conf=0.1,
target="rules"))

## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##         0.1    0.1    1 none FALSE            TRUE       5    0.01      1
##  maxlen target  ext
##      10  rules TRUE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 3
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[7 item(s), 365 transaction(s)] done [0.00s].
## sorting and recoding items ... [7 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 4 done [0.00s].
## writing ... [121 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].

summary(marketBas.trans)

## transactions as itemMatrix in sparse format with
##  365 rows (elements/itemsets/transactions) and
##  7 columns (items) and a density of 0.1878669
##
## most frequent items:
##         SW   Cat.Data Regression     Survey DataMining    (Other)
##         81         76         76         68         65        114
##
## element (itemset/transaction) length distribution:
## sizes
##   0   1   2   3   4   5   6
##  84 173  44  43  16   4   1
##
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   1.000   1.315   2.000   6.000
##
## includes extended item information - examples:
##       labels
## 1 DataMining
## 2     Survey
## 3   Cat.Data
```

The result of rules gave 121 rules to create baskets, and the summary showed; 0 topic list has 84 rules, 1 topic list has 173, 2 topics list has 44 rules, 3 topics list has 43 rules, 4 topics list has 16, 5 topics list has 4 rules, and 6 topics list has 1.

Inspect the first six rules, sorted by their lift

```
inspect(head(sort(rules, by="lift"), n=6))
```

```
##     lhs                          rhs          support    confidence
coverage
## [1] {DataMining, DOE}        => {Cat.Data}   0.01643836 0.6666667
0.02465753
## [2] {Survey, Regression}     => {Cat.Data}   0.01643836 0.6666667
0.02465753
## [3] {DataMining, Regression} => {Forecast}   0.01917808 0.4375000
0.04383562
## [4] {DataMining, Regression} => {Cat.Data}   0.02739726 0.6250000
0.04383562
## [5] {Regression, DOE}        => {SW}         0.01917808 0.6363636
0.03013699
## [6] {Regression, Forecast}   => {DataMining} 0.01917808 0.5000000
0.03835616
##     lift     count
## [1] 3.201754  6
## [2] 3.201754  6
## [3] 3.131127  7
## [4] 3.001645 10
## [5] 2.867565  7
## [6] 2.807692  7
```

The results of the sort-by-lift statistics show how many times more likely this is to be in there as opposed to chance. The result of rule 1 shows a high probability of the topic group being DataMining and DOE who have more like to have Cat.Data topic as their result of lift ratio is 3.2 chance of having Cat.Data at the same time if the student has two other topics. Rule sixth shows that if students have Regression and Forecast together, they have a 2.8 chance to have a data DataMining topic. The top 6 lift score specify Cat.Data topic taken with Survey is 4 times, and DataMining is 3 times. The foremost lift is always superior for product assosiaction that needs more focus.

Rules by confidence

```
rulesConf<- sort(rules, by="confidence", decreasing = TRUE)
inspect(head(rulesConf))
```

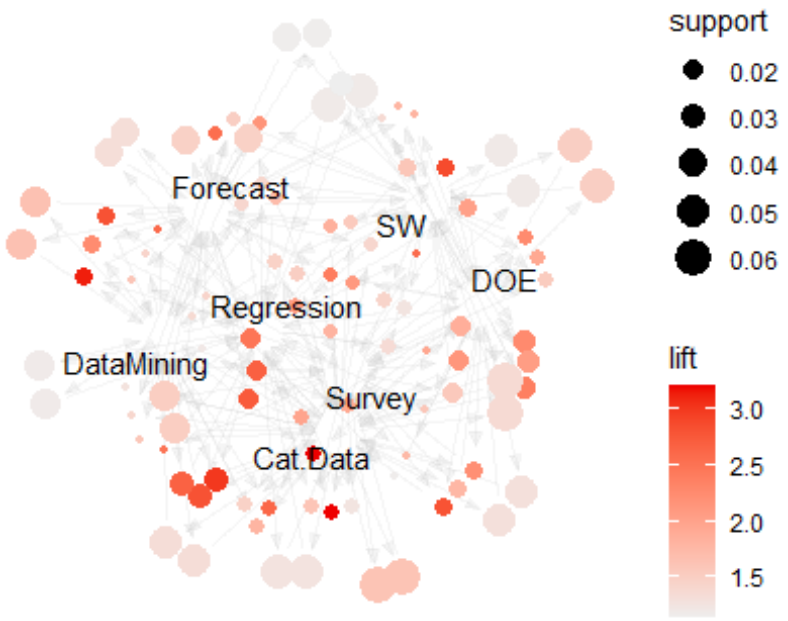
```
##     lhs                           rhs        support    confidence coverage
## [1] {DataMining, DOE}          => {Cat.Data} 0.01643836 0.6666667
0.02465753
## [2] {Survey, Regression}       => {Cat.Data} 0.01643836 0.6666667
0.02465753
## [3] {Regression, DOE}          => {SW}       0.01917808 0.6363636
0.03013699
## [4] {DataMining, Regression}   => {Cat.Data} 0.02739726 0.6250000
0.04383562
## [5] {Survey, DOE}              => {Cat.Data} 0.01917808 0.5833333
0.03287671
## [6] {Survey, Forecast}         => {Cat.Data} 0.02191781 0.5714286
0.03835616
##     lift     count
## [1] 3.201754  6
## [2] 3.201754  6
## [3] 2.867565  7
## [4] 3.001645 10
## [5] 2.801535  7
## [6] 2.744361  8
```

Rules Graph

```
#install.packages("arulesViz")
library(arulesViz)
plot(rules, method="graph", control=list(type="items"))
```

```
## Warning: Unknown control parameters: type
```

```
## Available control parameters (with default values):
## layout    =  stress
## circular  =  FALSE
## ggraphdots    =  NULL
## edges     =  <environment>
## nodes     =  <environment>
## nodetext  =  <environment>
## colors    =  c("#EE0000FF", "#EEEEEEFF")
## engine    =  ggplot2
## max   =  100
## verbose   =  FALSE
## 'lift' (change control parameter max if needed).
```



The visualization shows three dots' support values are 0.05 and lift over 2.5. Those stay middle on DataMining, Regression, and Cat.Data.

Which topics led to Cat.Data

```
rules<-apriori(marketBas.trans,   parameter   =   list(minlen=2,   supp=0.01,
conf=0.1, target="rules"), appearance = list(default="lhs", rhs="Cat.Data"),
control=list(verbose=F))
inspect(rules, ruleSep=  "--->", itemSep="+")
```

```
##      lhs                          rhs        support    confidence
coverage
## [1]  {Forecast}              ---> {Cat.Data} 0.04383562 0.3137255
0.13972603
## [2]  {DOE}                   ---> {Cat.Data} 0.04657534 0.2698413
0.17260274
## [3]  {DataMining}            ---> {Cat.Data} 0.04931507 0.2769231
0.17808219
## [4]  {Survey}                ---> {Cat.Data} 0.06301370 0.3382353
0.18630137
## [5]  {Regression}            ---> {Cat.Data} 0.05479452 0.2631579
0.20821918
## [6]  {SW}                    ---> {Cat.Data} 0.06301370 0.2839506
0.22191781
## [7]  {Forecast+DOE}          ---> {Cat.Data} 0.01369863 0.5000000
0.02739726
## [8]  {DataMining+Forecast}   ---> {Cat.Data} 0.01095890 0.2666667
```

```
0.04109589
## [9]  {Survey+Forecast}       ---> {Cat.Data} 0.02191781 0.5714286
0.03835616
## [10] {Regression+Forecast}   ---> {Cat.Data} 0.01095890 0.2857143
0.03835616
## [11] {Forecast+SW}           ---> {Cat.Data} 0.01369863 0.3846154
0.03561644
## [12] {DataMining+DOE}        ---> {Cat.Data} 0.01643836 0.6666667
0.02465753
## [13] {Survey+DOE}            ---> {Cat.Data} 0.01917808 0.5833333
0.03287671
## [14] {Regression+DOE}        ---> {Cat.Data} 0.01095890 0.3636364
0.03013699
## [15] {DOE+SW}                ---> {Cat.Data} 0.02465753 0.4285714
0.05753425
## [16] {DataMining+Survey}     ---> {Cat.Data} 0.01643836 0.5454545
0.03013699
## [17] {DataMining+Regression} ---> {Cat.Data} 0.02739726 0.6250000
0.04383562
## [18] {DataMining+SW}         ---> {Cat.Data} 0.01643836 0.4285714
0.03835616
## [19] {Survey+Regression}     ---> {Cat.Data} 0.01643836 0.6666667
0.02465753
## [20] {Survey+SW}             ---> {Cat.Data} 0.02191781 0.4444444
0.04931507
## [21] {Regression+SW}         ---> {Cat.Data} 0.01643836 0.3000000
0.05479452
##      lift     count
## [1]  1.506708 16
## [2]  1.295948 17
## [3]  1.329960 18
## [4]  1.624420 23
## [5]  1.263850 20
## [6]  1.363710 23
## [7]  2.401316  5
## [8]  1.280702  4
## [9]  2.744361  8
## [10] 1.372180  4
## [11] 1.847166  5
## [12] 3.201754  6
## [13] 2.801535  7
## [14] 1.746411  4
## [15] 2.058271  9
## [16] 2.619617  6
## [17] 3.001645 10
## [18] 2.058271  6
## [19] 3.201754  6
## [20] 2.134503  8
## [21] 1.440789  6
```

In addition, 23 times, Survey and SW had students will likely have Cat.Data. Rules 19, 12, and 17 have the highest lift and a high probability of having Cat.Data.

I will create a basket with minimum support of 0.01 and minimum confidence of 0.5 with R's apriori() function.

```
rules1<- apriori(marketBas.trans, parameter = list(supp=0.01, conf=0.5,
target="rules"))
```

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##         0.5    0.1    1 none FALSE            TRUE       5    0.01      1
##  maxlen target  ext
##      10  rules TRUE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 3
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[7 item(s), 365 transaction(s)] done [0.00s].
## sorting and recoding items ... [7 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 4 done [0.00s].
## writing ... [14 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

Inspect the first six rules, sorted by their lift

```
inspect(head(sort(rules1, by="lift"), n=6))
```

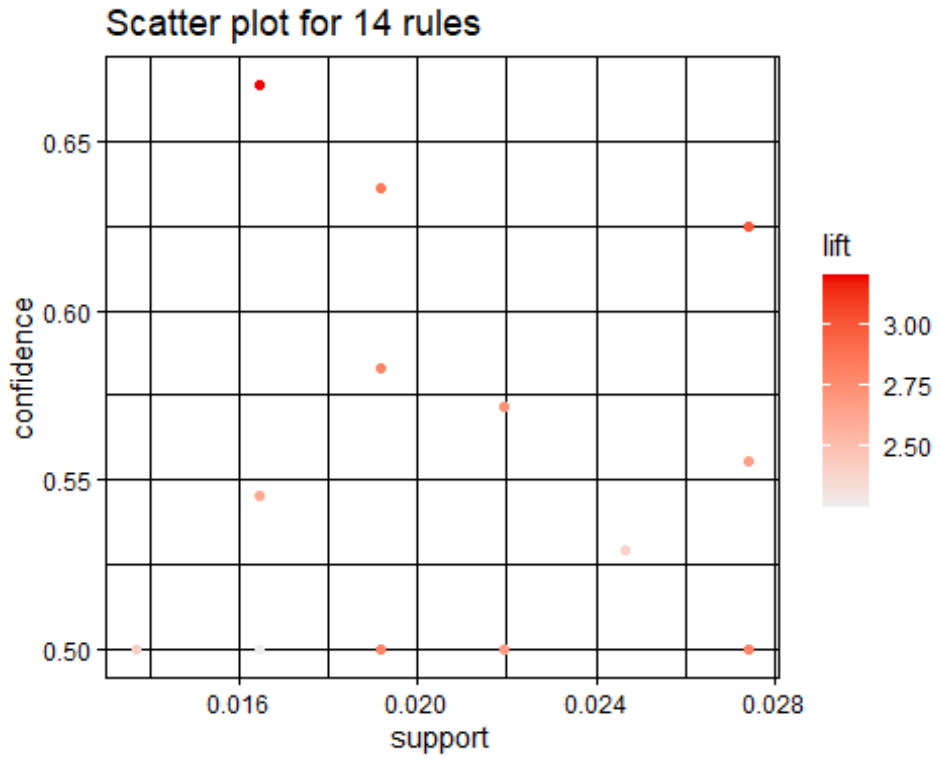
```
##     lhs                          rhs          support    confidence
coverage
## [1] {DataMining, DOE}        => {Cat.Data}   0.01643836 0.6666667
0.02465753
## [2] {Survey, Regression}     => {Cat.Data}   0.01643836 0.6666667
0.02465753
## [3] {DataMining, Regression} => {Cat.Data}   0.02739726 0.6250000
0.04383562
## [4] {Regression, DOE}        => {SW}         0.01917808 0.6363636
0.03013699
## [5] {Regression, Forecast}   => {DataMining} 0.01917808 0.5000000
0.03835616
## [6] {Cat.Data, Regression}   => {DataMining} 0.02739726 0.5000000
0.05479452
##     lift     count
## [1] 3.201754  6
```

```
## [2] 3.201754  6
## [3] 3.001645 10
## [4] 2.867565  7
## [5] 2.807692  7
## [6] 2.807692 10
```

Lift ratio for confidence 0.5 show high value of group is DataMining and DOE with Cat.Data topic same result of confidence 0.1's lift ratio. Although sixth line is group different if student assign Cat.Data and Regression topics 2.8 chance to have DataMining topic.

Visualizing rules

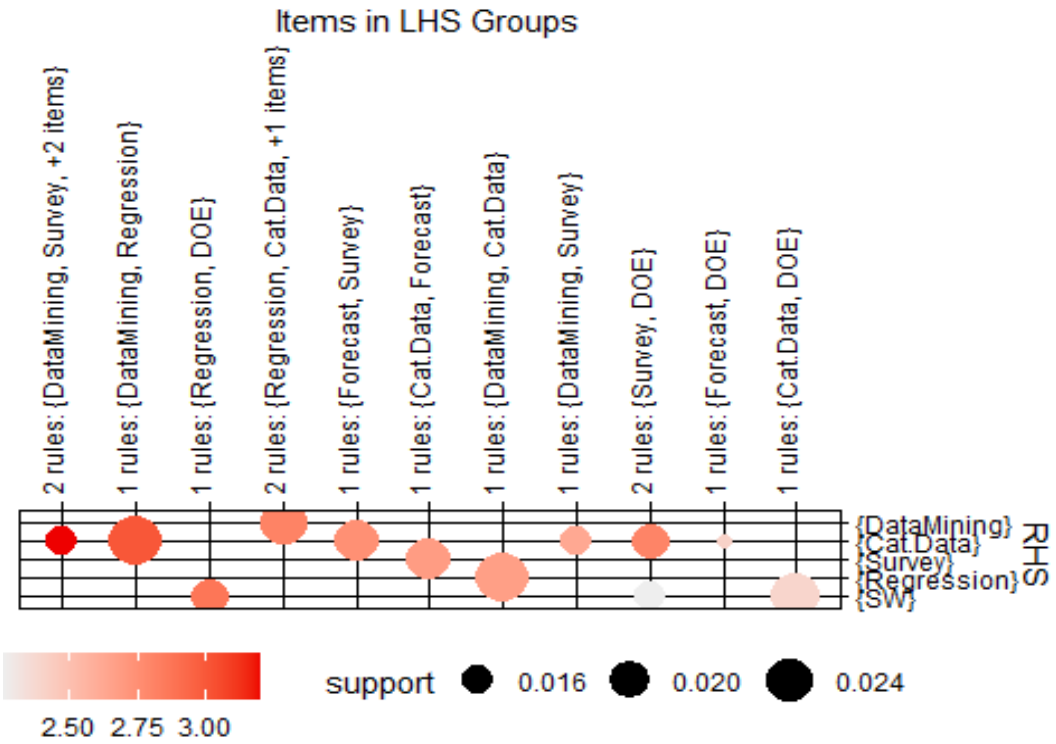
```
library(arulesViz)
plot(rules1, jitter=0)
```



Graph shows lift support and confidence values for all 14 rules. If the dot is darker color means lift value is high so one dot over the 0.65 point of confidence, support between 0.016 to 0.020 and lift over the 3.

Visual the group

```
library(arulesViz)
plot(rules1, method = "grouped")
```



The high probability topics are DataMining and Survey, and two plus topics with Cat.Data. The highest support group of topics is DataMining, Regression with Cat.Data topic.

Reference

Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., & Lichtendahl, K. C. (2018). Data mining for Business Analytics: Concepts, techniques, and applications in R (1st ed.). John Wiley & Sons.

Megha Goriya(2021), Published in Analytics Vidhya. Market Basket Analysis using Association Rules. https://medium.com/analytics-vidhya/market-basket-analysis-using-association-rules-2b0f3e2a897d

Ricardo Rodriguez ,Market Basket Analysis on Tech Business Guide. https://techbusinessguide.com/what-is-market-basket-analysis/

Rajarajachozhan(2020). Market basket analysis in e-commerce business explained (A Case study). https://gecdesigns.com/blog/how-can-market-basket-analysis-help-e-commerce-business