ASSOCIATION RULES

Bogdan Tanasa

1. DATA EXPLORATION

2. ASSOCIATION RULES

1. DATA EXPLORATION

We are using the data that we had from **UCI** a while ago in the file "Association_DataSet.csv".

```
options(warn=-1)
suppressPackageStartupMessages(library(ggplot2))
suppressPackageStartupMessages(library(reshape2))
suppressPackageStartupMessages(library(arules))
suppressPackageStartupMessages(library(arulesViz))

######################################################
######################################################

FILE1="Association_DataSet.csv"

######################################################
######################################################

file = read.delim("Association_DataSet.csv", sep = ",", header=TRUE, stringsAsFactors=F)

######################################################
######################################################

str(file)
```

```
## 'data.frame':    3483 obs. of  12 variables:
##  $ Elapsed_Time     : num  8.71 5.24 4.22 4.81 3.95 9.35 2.91 4.54 4.79 3.07 ...
##  $ Time_in_Community: chr  "Short" "Medium" "Medium" "Long" ...
##  $ Gender           : chr  "M" "F" "M" "F" ...
##  $ Working          : chr  "No" "No" "No" "No" ...
##  $ Age              : int  53 31 42 30 29 40 33 27 50 28 ...
##  $ Family           : int  1 0 1 0 0 0 0 1 1 0 ...
##  $ Hobbies          : int  0 0 1 0 0 0 0 1 1 0 ...
##  $ Social_Club      : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ Political        : int  0 0 0 0 1 0 0 0 0 0 ...
##  $ Professional     : int  0 0 1 0 1 1 0 0 1 0 ...
##  $ Religious        : int  0 1 0 0 0 0 0 1 1 1 ...
##  $ Support_Group    : int  0 1 0 0 1 0 1 0 0 1 ...
```

```
class(file)
```

```
## [1] "data.frame"
```

```
summary(file)
```

```
##   Elapsed_Time    Time_in_Community     Gender             Working
##  Min.   : 2.010   Length:3483        Length:3483        Length:3483
##  1st Qu.: 3.875   Class :character   Class :character   Class :character
##  Median : 5.930   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 5.922
##  3rd Qu.: 7.840
##  Max.   :10.150
##       Age            Family          Hobbies      Social_Club    
##  Min.   :17.00   Min.   :0.0000   Min.   :0.0   Min.   :0.0000  
##  1st Qu.:27.00   1st Qu.:0.0000   1st Qu.:0.0   1st Qu.:0.0000  
##  Median :36.00   Median :0.0000   Median :0.0   Median :0.0000  
##  Mean   :36.73   Mean   :0.3899   Mean   :0.3   Mean   :0.1881  
##  3rd Qu.:46.00   3rd Qu.:1.0000   3rd Qu.:1.0   3rd Qu.:0.0000  
##  Max.   :57.00   Max.   :1.0000   Max.   :1.0   Max.   :1.0000  
##    Political        Professional      Religious      Support_Group   
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000  
##  Median :0.00000   Median :0.0000   Median :0.0000   Median :0.0000  
##  Mean   :0.09388   Mean   :0.3244   Mean   :0.4186   Mean   :0.1588  
##  3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000  
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000  
```

```
######################################################
######################################################

file$Family = as.factor(file$Family)
file$Hobbies  = as.factor(file$Hobbies)
file$Social_Club = as.factor(file$Social_Club)
file$Political  = as.factor(file$Political)
file$Professional = as.factor(file$Professional)
file$Religious  = as.factor(file$Religious)
file$Support_Group = as.factor(file$Support_Group)

summary(file)
```

```
##   Elapsed_Time    Time_in_Community     Gender            Working         
##  Min.   : 2.010   Length:3483        Length:3483        Length:3483       
##  1st Qu.: 3.875   Class :character   Class :character   Class :character  
##  Median : 5.930   Mode  :character   Mode  :character   Mode  :character  
##  Mean   : 5.922                                                           
##  3rd Qu.: 7.840                                                           
##  Max.   :10.150                                                           
##       Age        Family   Hobbies  Social_Club Political Professional Religious
##  Min.   :17.00   0:2125   0:2438   0:2828      0:3156    0:2353       0:2025   
##  1st Qu.:27.00   1:1358   1:1045   1: 655      1: 327    1:1130       1:1458   
##  Median :36.00                                                                 
##  Mean   :36.73                                                                 
##  3rd Qu.:46.00                                                                 
##  Max.   :57.00                                                                 
##  Support_Group
##  0:2930       
##  1: 553       
##               
##               
##               
##               
```

```
######################################################
######################################################
# to exclude Elapsed_Time and Age
# to transform 0 and 1 into YES or NO
```

```
a_file <- file[, -which(names(file) %in% c("Elapsed_Time", "Age"))]

summary(a_file)
```

```
##  Time_in_Community     Gender            Working          Family   Hobbies
##  Length:3483        Length:3483        Length:3483        0:2125   0:2438
##  Class :character   Class :character   Class :character   1:1358   1:1045
##  Mode  :character   Mode  :character   Mode  :character
##  Social_Club Political Professional Religious Support_Group
##  0:2828      0:3156    0:2353       0:2025    0:2930
##  1: 655      1: 327    1:1130       1:1458    1: 553
##
```

```
##  Time_in_Community     Gender            Working          Family   Hobbies
##  Length:3483        Length:3483        Length:3483        0:2125   0:2438
##  Class :character   Class :character   Class :character   1:1358   1:1045
##  Mode  :character   Mode  :character   Mode  :character
##  Social_Club Political Professional Religious Support_Group
##  0:2828      0:3156    0:2353       0:2025    0:2930
##  1: 655      1: 327    1:1130       1:1458    1: 553

a_file$Family = ifelse(a_file$Family == "0", "Family_No", "Family_Yes")
a_file$Hobbies = ifelse(a_file$Hobbies == "0", "Hobbies_No", "Hobbies_Yes")
a_file$Social_Club = ifelse(a_file$Social_Club == "0", "Social_Club_No", "Social_Club_Yes")
a_file$Political = ifelse(a_file$Political == "0", "Political_No", "Political_Yes")
a_file$Professional  = ifelse(a_file$Professional == "0", "Professional_No", "Professional_Yes")
a_file$Religious = ifelse(a_file$Religious == "0", "Religious_No", "Religious_Yes")
a_file$Support_Group = ifelse(a_file$Support_Group == "0", "Support_Group_No", "Support_Group_Yes")

summary(a_file)
```

```
##  Time_in_Community     Gender            Working             Family
##  Length:3483        Length:3483        Length:3483        Length:3483
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##    Hobbies          Social_Club         Political         Professional
##  Length:3483        Length:3483        Length:3483        Length:3483
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##   Religious         Support_Group
##  Length:3483        Length:3483
##  Class :character   Class :character
##  Mode  :character   Mode  :character
```

```
write.csv(a_file, file = "the_dataset.csv", row.names = FALSE)
```

2. ASSOCIATION RULES

Here we are performing the association analysis and we display the data.

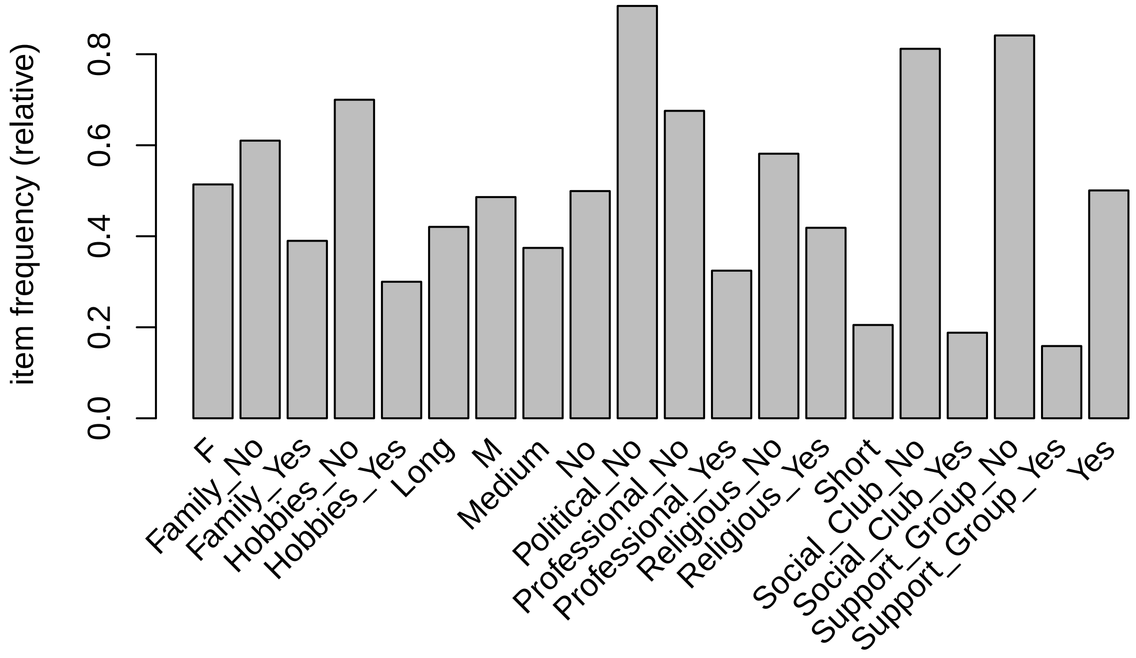
```
#######################################################################################

the_data <- read.transactions("the_dataset.csv", sep = ",", header=TRUE)

###################################################################
###################################################################

summary(the_data)
```

```
## transactions as itemMatrix in sparse format with
##  3483 rows (elements/itemsets/transactions) and
##  21 columns (items) and a density of 0.4761905
##
## most frequent items:
##     Political_No Support_Group_No   Social_Club_No       Hobbies_No
##             3156             2930             2828             2438
##  Professional_No          (Other)
##             2353            21125
##
## element (itemset/transaction) length distribution:
## sizes
##   10
## 3483
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      10      10      10      10      10      10
##
## includes extended item information - examples:
##       labels
## 1          F
## 2  Family_No
## 3 Family_Yes
```

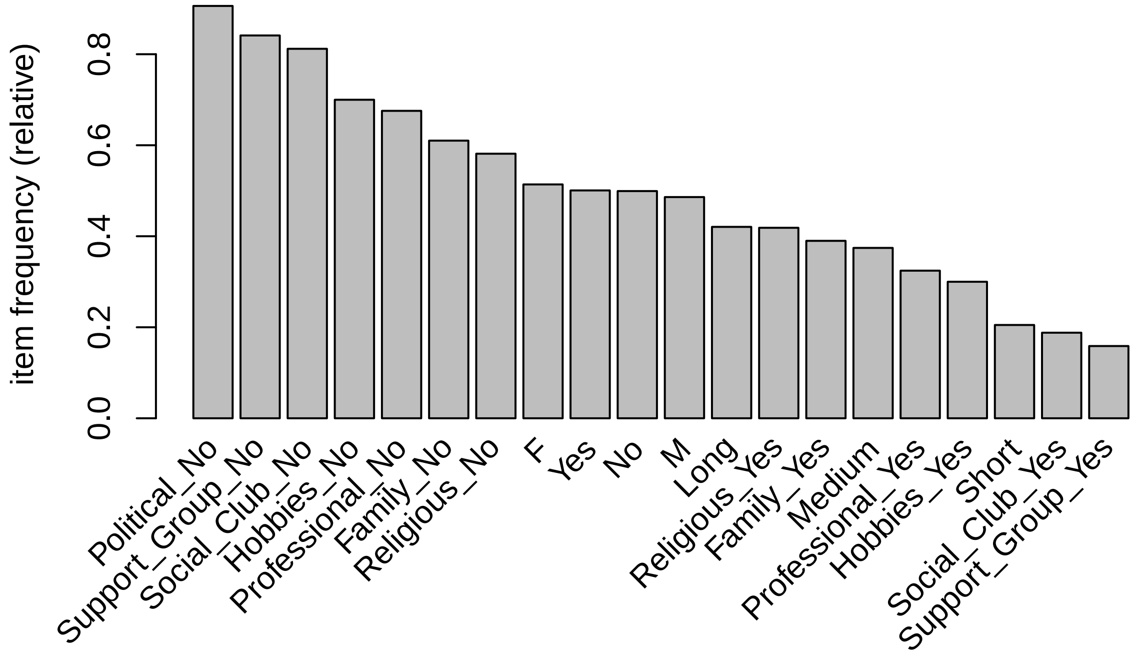
```
# inspect(the_data)
inspect(the_data[1:5])
```

```
##     items
## [1] {Family_Yes,
##      Hobbies_No,
##      M,
##      No,
##      Political_No,
##      Professional_No,
##      Religious_No,
##      Short,
##      Social_Club_No,
##      Support_Group_No}
## [2] {F,
##      Family_No,
##      Hobbies_No,
##      Medium,
##      No,
##      Political_No,
##      Professional_No,
##      Religious_Yes,
##      Social_Club_No,
##      Support_Group_Yes}
## [3] {Family_Yes,
##      Hobbies_Yes,
##      M,
##      Medium,
##      No,
##      Political_No,
##      Professional_Yes,
##      Religious_No,
##      Social_Club_No,
##      Support_Group_No}
## [4] {F,
##      Family_No,
##      Hobbies_No,
##      Long,
##      No,
##      Political_No,
##      Professional_No,
##      Religious_No,
##      Social_Club_No,
##      Support_Group_No}
## [5] {Family_No,
##      Hobbies_No,
##      Long,
##      M,
##      Political_Yes,
##      Professional_Yes,
##      Religious_No,
##      Social_Club_No,
##      Support_Group_Yes,
##      Yes}
```

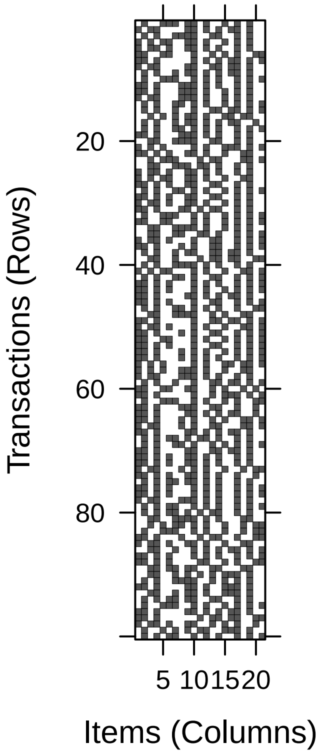
```
##################################################################
##################################################################

itemFrequencyPlot(the_data, support = 0.1)
```



```r
itemFrequencyPlot(the_data, topN = 20)
```



```
##################################################################
##################################################################

# to visualize the data :

image(sample(the_data, 100))
```



```
######################################################################
######################################################################

# if we attempt to use the default settings of support = 0.1 and confidence = 0.8,
# find a set of 2918 rules:

the_rules = apriori(the_data)
```

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##         0.8    0.1    1 none FALSE            TRUE       5     0.1      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: 348
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[21 item(s), 3483 transaction(s)] done [0.00s].
## sorting and recoding items ... [20 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 4 5 6 7 8 done [0.00s].
## writing ... [2918 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

```
summary(the_rules)
```

```
## set of 2918 rules
##
## rule length distribution (lhs + rhs):sizes
##   1   2   3   4   5   6   7   8
##   3  49 332 857 979 554 135   9
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   4.000   5.000   4.716   5.000   8.000
##
## summary of quality measures:
##     support         confidence        coverage           lift
##  Min.   :0.1002   Min.   :0.8000   Min.   :0.1034   Min.   :0.9164
##  1st Qu.:0.1183   1st Qu.:0.8636   1st Qu.:0.1326   1st Qu.:1.0210
##  Median :0.1490   Median :0.9042   Median :0.1662   Median :1.0439
##  Mean   :0.1787   Mean   :0.8975   Mean   :0.1997   Mean   :1.0947
##  3rd Qu.:0.2038   3rd Qu.:0.9342   3rd Qu.:0.2283   3rd Qu.:1.1558
##  Max.   :0.9061   Max.   :0.9867   Max.   :1.0000   Max.   :2.1222
##       count
##  Min.   : 349.0
##  1st Qu.: 412.0
##  Median : 519.0
##  Mean   : 622.5
##  3rd Qu.: 709.8
##  Max.   :3156.0
##
## mining info:
##      data ntransactions support confidence                     call
##  the_data          3483     0.1        0.8 apriori(data = the_data)
```

```
# if we change the settings we may find less rules:

# the_rules = apriori(the_data, parameter = list(support = 0.1,
#                                               confidence = 0.8,
#                                               minlen = 2))

# summary(the_rules)
```

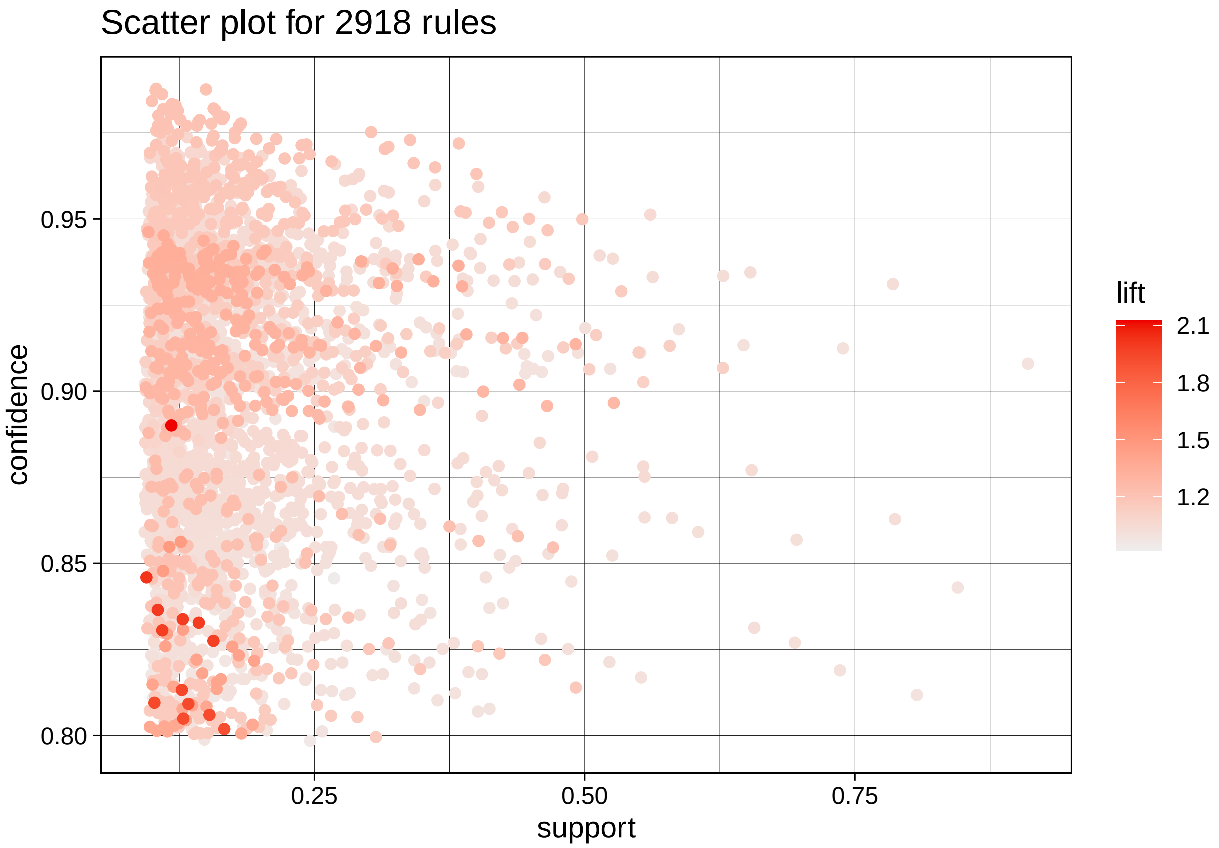
```
################################################################## DISPLAYING strongly
################################################################## SUPPORTED RULES

rules.sorted <- sort(the_rules, by="lift")

# inspect(rules.sorted)

inspect(rules.sorted[1:5])
```

```
##     lhs                   rhs               support confidence  coverage     lift count
## [1] {Hobbies_Yes,
##      Social_Club_Yes}  => {Religious_Yes} 0.1096756  0.8883721 0.1234568 2.122222   382
## [2] {Family_Yes,
##      Hobbies_Yes,
##      Political_No,
##      Professional_No}  => {Religious_Yes} 0.1016365  0.8448687 0.1202986 2.018298   354
## [3] {Family_Yes,
##      Hobbies_Yes,
##      Professional_No}  => {Religious_Yes} 0.1076658  0.8370536 0.1286247 1.999628   375
## [4] {Family_Yes,
##      Hobbies_Yes,
##      Political_No}     => {Religious_Yes} 0.1401091  0.8341880 0.1679587 1.992783   488
## [5] {Family_Yes,
##      Hobbies_Yes,
##      Support_Group_No} => {Religious_Yes} 0.1231697  0.8330097 0.1478610 1.989968   429
```

```
## saving the data :

write(the_rules, file = "the_rules.csv",
                 sep = ",",
                 quote = TRUE,
                 row.names = FALSE)
```

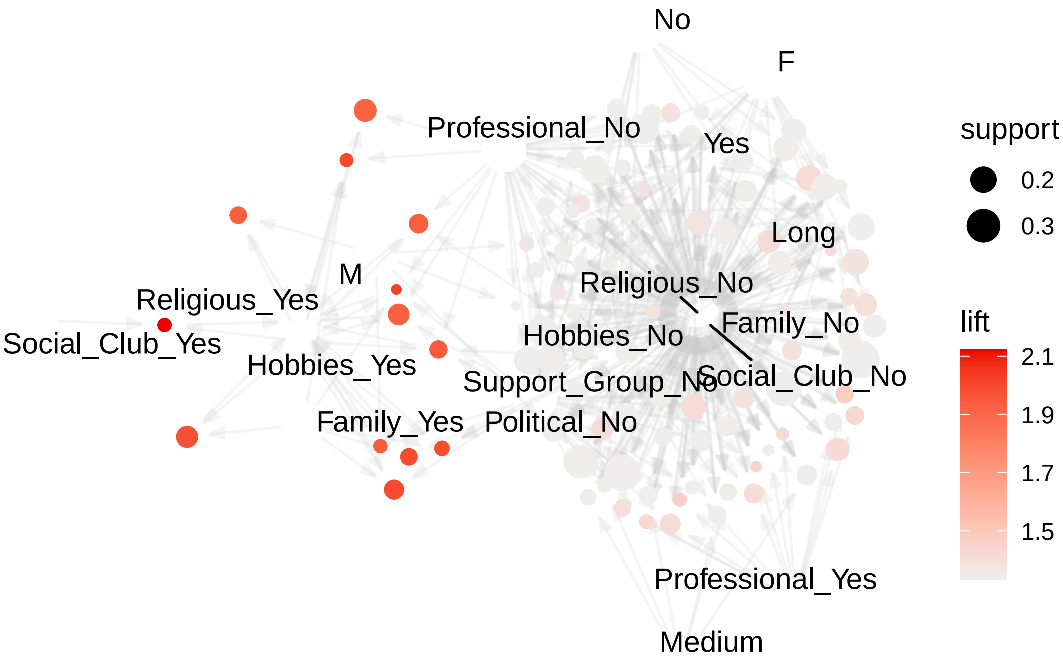
```r
## visualizing the data :

plot(the_rules)
```

```
## To reduce overplotting, jitter is added! Use jitter = 0 to prevent jitter.
```



```r
plot(the_rules, method="graph", control=list(type="items"))
```

```
## 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
```



```
# plot(the_rules, method="paracoord", control=list(reorder=TRUE))

############################################################################
############################################################################
################################ a piece of R code if we may have to use
################################ depending on the context

# inspect(rules.sorted)

# prune redundant rules.
# subset.matrix <- is.subset(rules.sorted, rules.sorted)
# subset.matrix[lower.tri(subset.matrix, diag=T)] <- NA
# redundant <- colSums(subset.matrix, na.rm=T) >= 1
# which(redundant)

# remove redundant rules.
# rules.pruned <- rules.sorted[!redundant]
# inspect(rules.pruned)
```