retail.R

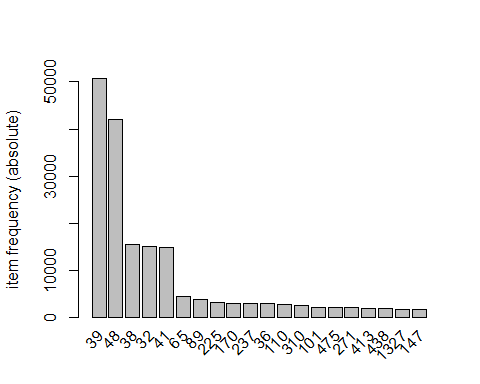
Administrator

2019-10-23

# About: We want to analyze which items across a retail data set are more likely to be purchased together. For this, we'll explore the retail.csv file which, though might give us insight into items purchased as a whole, will not give us an idea towards customer preference as we do not know what items they are purchasing.   
# Load required packages  
require(pacman)

## Loading required package: pacman

p\_load(arules, arulesViz)  
#```  
  
## Loading the data   
#We will use the arules and arulesViz package to get a better idea as to which group of items generate the highest lift and confidence.   
  
# Now to the same with the retail.csv file  
retail <- read.transactions('C:/Users/Administrator/Documents/MSDatascience/IST707/Week3/retail.csv', format = 'basket', sep = ',')  
  
  
itemFrequencyPlot(retail, topN = 20, type = 'absolute')



## Initial analysis  
#We can see from the previous plot that the most frequently purchased items are items 39 and 48. After those two items, items 38, 32, and 41 were the most frequently bought, though at a much more reduced rate.   
  
## Getting the most common three-item rules.   
#Setting support to 0.001, confidence to at least 0.80, and max length of baskets to three items, we'll explore those baskets with first a high lift followed by high confidence.   
  
rules <- apriori(retail, parameter = list(supp = 0.001, conf = 0.8, maxlen = 3))

## Apriori  
##   
## Parameter specification:  
## confidence minval smax arem aval originalSupport maxtime support minlen  
## 0.8 0.1 1 none FALSE TRUE 5 0.001 1  
## maxlen target ext  
## 3 rules FALSE  
##   
## Algorithmic control:  
## filter tree heap memopt load sort verbose  
## 0.1 TRUE TRUE FALSE TRUE 2 TRUE  
##   
## Absolute minimum support count: 88   
##   
## set item appearances ...[0 item(s)] done [0.00s].  
## set transactions ...[16470 item(s), 88162 transaction(s)] done [0.60s].  
## sorting and recoding items ... [2117 item(s)] done [0.01s].  
## creating transaction tree ... done [0.05s].  
## checking subsets of size 1 2 3

## Warning in apriori(retail, parameter = list(supp = 0.001, conf = 0.8,  
## maxlen = 3)): Mining stopped (maxlen reached). Only patterns up to a length  
## of 3 returned!

## done [0.07s].  
## writing ... [355 rule(s)] done [0.01s].  
## creating S4 object ... done [0.02s].

rules <- sort(rules, decreasing = TRUE, by = c('lift', 'confidence'))  
inspect(rules[1:5])

## lhs rhs support confidence lift count  
## [1] {3311,795} => {1819} 0.001406502 0.8435374 296.2866 124   
## [2] {1818,3311} => {1819} 0.001293074 0.8142857 286.0122 114   
## [3] {1818,1819} => {795} 0.001315760 0.8000000 280.9944 116   
## [4] {1080,1378} => {1379} 0.001077562 0.8119658 252.9489 95   
## [5] {1379,1380} => {309} 0.001043533 0.8214286 234.3650 92

#We can see that items 1819, 795, 3311, and 1818 are closely related as a combination of these three items are three of the top baskets with both a high lift and confidence - though their count is not among the highest.   
  
#To finalize, we'll take a look at these rules visually. We'll limit this graph to the top 5 rules.   
plot(rules[1:5], method = 'graph')

