Creating Association Rules for Supermarket Dataset

by Kennedy Muriuki

17/09/2020

Loading the csv file

```
# the csv file will be loaded from a local repository and will be loaded as class transactions.
# loading the required library
library(arules)
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'arules'
```

```
## The following objects are masked from 'package:base':
##
##     abbreviate, write
```

```
# loading the dataset
trans <- read.transactions(file.choose(), sep = ",", rm.duplicates = TRUE)
```

```
## distribution of transactions with duplicates:
## 1
## 5
```

```
# inspecting the class
class(trans)
```

```
## [1] "transactions"
## attr(,"package")
## [1] "arules"
```

```
# checking the head of the dataset
inspect(trans[1:5])
```

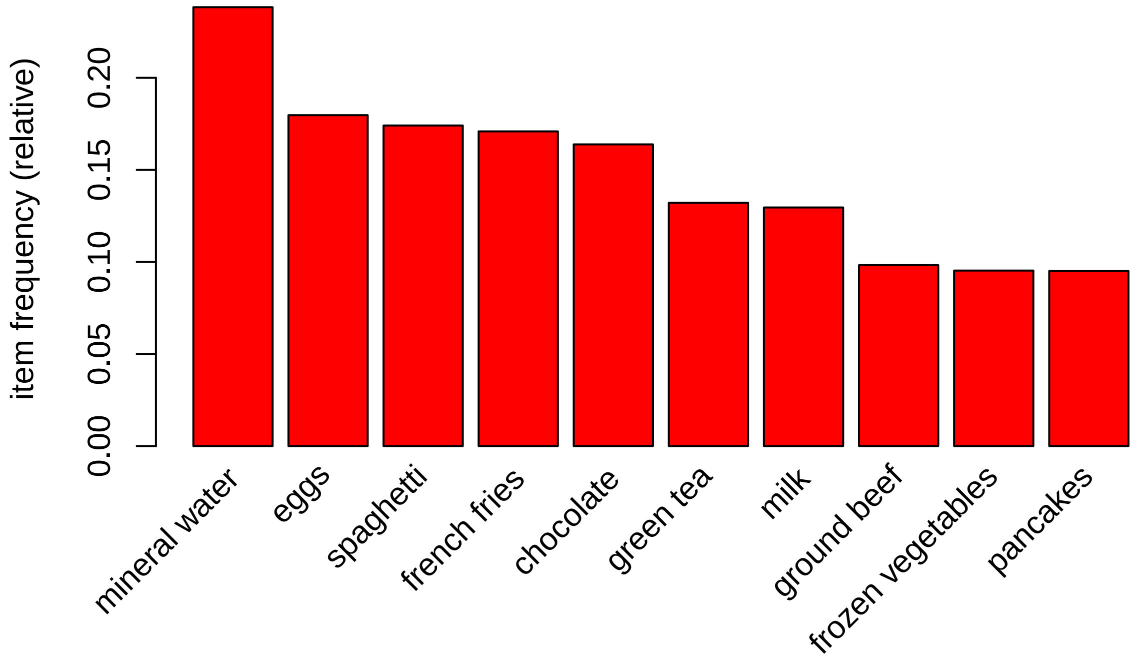
```
##     items
## [1] {almonds,
##      antioxydant juice,
##      avocado,
##      cottage cheese,
##      energy drink,
##      frozen smoothie,
##      green grapes,
##      green tea,
##      honey,
##      low fat yogurt,
##      mineral water,
##      olive oil,
##      salad,
##      salmon,
##      shrimp,
##      spinach,
##      tomato juice,
##      vegetables mix,
##      whole weat flour,
##      yams}
## [2] {burgers,
##      eggs,
##      meatballs}
## [3] {chutney}
## [4] {avocado,
##      turkey}
## [5] {energy bar,
##      green tea,
##      milk,
##      mineral water,
##      whole wheat rice}
```

```r
# generating a summary of the dataset
summary(trans)
```

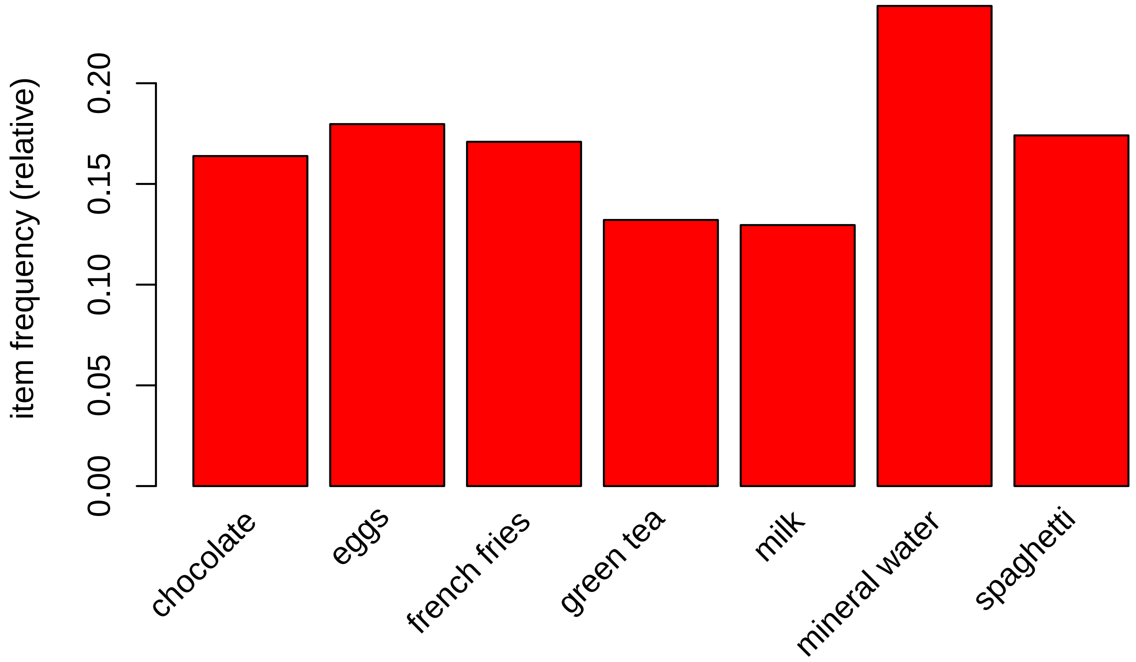
```
## transactions as itemMatrix in sparse format with
##  7501 rows (elements/itemsets/transactions) and
##  119 columns (items) and a density of 0.03288973
##
## most frequent items:
## mineral water          eggs     spaghetti  french fries     chocolate
##          1788          1348          1306          1282          1229
##       (Other)
##         22405
##
## element (itemset/transaction) length distribution:
## sizes
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16
## 1754 1358 1044  816  667  493  391  324  259  139  102   67   40   22   17    4
##   18   19   20
##    1    2    1
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   3.914   5.000  20.000
##
## includes extended item information - examples:
##               labels
## 1            almonds
## 2 antioxydant juice
## 3         asparagus
```

```r
# displaying the top 10 most common items in the dataset and items whose relative importance is atleast

# showing the top 10 common items
itemFrequencyPlot(trans, topN=10, col="red", type="relative")
```



```r
# showing the items whose importance is at least 10%
itemFrequencyPlot(trans, support=0.1, col="red", type="relative")
```



The top most most frequent item in the data set is mineral water followed by eggs and spaghetti while the 10th most frequent item was pancakes

The items whose popularity was at least 10% were 7 with mineral water having the highest popularity followed by eggs.

creating the model

```
# building the model based on apriori rules of association
rules <- apriori(trans, parameter = list(supp = 0.001, conf = 0.8))
```

```
## 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
##      10  rules TRUE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 7
```

```
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[119 item(s), 7501 transaction(s)] done [0.00s].
## sorting and recoding items ... [116 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 4 5 6 done [0.01s].
## writing ... [74 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

```
rules
```

```
## set of 74 rules
```

```
# obtaining the summary of the model
summary(rules)
```

```
## set of 74 rules
##
## rule length distribution (lhs + rhs):sizes
##  3  4  5  6
## 15 42 16  1
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   4.000   4.000   4.041   4.000   6.000
##
## summary of quality measures:
##     support           confidence        coverage             lift
##  Min.   :0.001067   Min.   :0.8000   Min.   :0.001067   Min.   : 3.356
##  1st Qu.:0.001067   1st Qu.:0.8000   1st Qu.:0.001333   1st Qu.: 3.432
##  Median :0.001133   Median :0.8333   Median :0.001333   Median : 3.795
##  Mean   :0.001256   Mean   :0.8504   Mean   :0.001479   Mean   : 4.823
##  3rd Qu.:0.001333   3rd Qu.:0.8889   3rd Qu.:0.001600   3rd Qu.: 4.877
##  Max.   :0.002533   Max.   :1.0000   Max.   :0.002666   Max.   :12.722
##      count
##  Min.   : 8.000
##  1st Qu.: 8.000
##  Median : 8.500
##  Mean   : 9.419
##  3rd Qu.:10.000
##  Max.   :19.000
##
## mining info:
##   data ntransactions support confidence
##  trans          7501   0.001        0.8
```

```
# inspecting the first five rules and sorting them with the level of confidence
rules <- sort(rules, by="confidence", decreasing = TRUE)
inspect(rules[1:5])
```

```
##     lhs                                            rhs             support
## [1] {french fries,mushroom cream sauce,pasta} => {escalope}      0.001066524
## [2] {ground beef,light cream,olive oil}       => {mineral water} 0.001199840
## [3] {cake,meatballs,mineral water}            => {milk}          0.001066524
## [4] {cake,olive oil,shrimp}                   => {mineral water} 0.001199840
## [5] {mushroom cream sauce,pasta}              => {escalope}      0.002532996
##     confidence coverage    lift      count
## [1] 1.00       0.001066524 12.606723  8
## [2] 1.00       0.001199840  4.195190  9
## [3] 1.00       0.001066524  7.717078  8
## [4] 1.00       0.001199840  4.195190  9
## [5] 0.95       0.002666311 11.976387 19
```

The first 4 rules had 100% confidence. For the first rule, this means that customers that bought french fries, mushroom cream sauce and pasta had a 100% chance to buy escalope.

A milk promotional case

Suppose Maziwa, milk company wanted to promote milk in Carrefour. The company would like to make a decision on product placement and would like to know which other products that customers bought before buying milk

```r
# generating the top five rules for milk
milk <- subset(rules, subset=rhs%pin% "milk")

# ordering the rules by confidence in descending order
milk <- sort(milk, by= "confidence", descending=TRUE)
inspect(milk[1:5])
```

```
##     lhs                                  rhs    support     confidence
## [1] {cake,meatballs,mineral water}    => {milk} 0.001066524 1.0000000
## [2] {escalope,hot dogs,mineral water} => {milk} 0.001066524 0.8888889
## [3] {meatballs,whole wheat pasta}     => {milk} 0.001333156 0.8333333
## [4] {black tea,frozen smoothie}       => {milk} 0.001199840 0.8181818
## [5] {burgers,ground beef,olive oil}   => {milk} 0.001066524 0.8000000
##     coverage    lift     count
## [1] 0.001066524 7.717078  8
## [2] 0.001199840 6.859625  8
## [3] 0.001599787 6.430898 10
## [4] 0.001466471 6.313973  9
## [5] 0.001333156 6.173663  8
```

Customers who bought cake meatballs and mineral water were 100% likely to buy milk. Therefore the milk company should place their product immediately after these products to guarantee a 100% likelihood of purchase.