HW #2 - market basket analysis / recommender systems

## Introduction

I am assigning one simple problem on market basket analysis / recommender systems.

Imagine 10000 receipts sitting on your table. Each receipt represents a transaction with items that were purchased. The receipt is a representation of stuff that went into a customer’s basket – and therefore ‘Market Basket Analysis’.

That is exactly what the Groceries Data Set contains: a collection of receipts with each line representing 1 receipt and the items purchased. Each line is called a transaction and each column in a row represents an item.

Here is the dataset = GroceryDataSet.csv (comma separated file)

#### You assignment is to use R to mine the data for association rules. You should report support, confidence and lift and your top 10 rules by lift. Turn in as you would the other problems from HA and KJ. Due 07/18/20 with the packaged set, HW #2.

### MARKET BASKET ANALYSIS

## Data preprocessing and exploring

## citrus.fruit semi.finished.bread margarine  
## 1 tropical fruit yogurt coffee  
## 2 whole milk   
## 3 pip fruit yogurt cream cheese   
## 4 other vegetables whole milk condensed milk  
## 5 whole milk butter yogurt  
## 6 rolls/buns   
## ready.soups X X.1 X.2 X.3 X.4 X.5 X.6 X.7  
## 1   
## 2   
## 3 meat spreads   
## 4 long life bakery product   
## 5 rice abrasive cleaner   
## 6   
## X.8 X.9 X.10 X.11 X.12 X.13 X.14 X.15 X.16 X.17 X.18 X.19 X.20 X.21 X.22  
## 1   
## 2   
## 3   
## 4   
## 5   
## 6   
## X.23 X.24 X.25 X.26 X.27  
## 1   
## 2   
## 3   
## 4   
## 5   
## 6

## citrus.fruit semi.finished.bread margarine   
## sausage : 825 :2159 :3802   
## whole milk : 717 whole milk : 654 whole milk : 506   
## frankfurter : 580 other vegetables: 550 other vegetables: 415   
## tropical fruit : 482 root vegetables : 383 rolls/buns : 293   
## other vegetables: 460 rolls/buns : 378 yogurt : 289   
## citrus fruit : 452 tropical fruit : 355 soda : 229   
## (Other) :6318 (Other) :5355 (Other) :4300   
## ready.soups X X.1   
## :5101 :6105 :6960   
## whole milk : 315 rolls/buns : 176 soda : 150   
## other vegetables: 254 soda : 168 rolls/buns : 146   
## rolls/buns : 238 yogurt : 160 shopping bags: 107   
## soda : 211 whole milk : 149 bottled water: 95   
## yogurt : 202 shopping bags: 145 yogurt : 93   
## (Other) :3513 (Other) :2931 (Other) :2283   
## X.2 X.3 X.4   
## :7605 :8150 :8588   
## soda : 120 shopping bags: 76 soda : 61   
## shopping bags: 107 bottled water: 68 shopping bags : 56   
## rolls/buns : 92 newspapers : 66 fruit/vegetable juice: 55   
## newspapers : 68 rolls/buns : 59 bottled water : 54   
## domestic eggs: 57 soda : 59 newspapers : 51   
## (Other) :1785 (Other) :1356 (Other) : 969   
## X.5 X.6   
## :8938 :9184   
## shopping bags : 49 shopping bags: 40   
## soda : 39 newspapers : 36   
## fruit/vegetable juice: 34 pastry : 27   
## newspapers : 33 bottled water: 25   
## bottled water : 26 napkins : 23   
## (Other) : 715 (Other) : 499   
## X.7 X.8   
## :9366 :9483   
## soda : 30 soda : 24   
## shopping bags : 19 shopping bags : 18   
## chocolate : 17 fruit/vegetable juice: 16   
## fruit/vegetable juice: 17 napkins : 14   
## napkins : 17 newspapers : 14   
## (Other) : 368 (Other) : 265   
## X.9 X.10   
## :9561 :9638   
## shopping bags : 18 shopping bags : 16   
## fruit/vegetable juice: 17 napkins : 13   
## newspapers : 14 fruit/vegetable juice: 11   
## soda : 14 hygiene articles : 11   
## napkins : 11 candy : 9   
## (Other) : 199 (Other) : 136   
## X.11 X.12   
## :9693 :9739   
## shopping bags : 11 napkins : 8   
## napkins : 9 chocolate : 5   
## chocolate : 8 newspapers : 5   
## hygiene articles : 6 candy : 4   
## long life bakery product: 6 fruit/vegetable juice: 4   
## (Other) : 101 (Other) : 69   
## X.13 X.14 X.15   
## :9768 :9782 :9796   
## candy : 5 detergent : 4 bottled beer : 3   
## chocolate : 5 fruit/vegetable juice: 4 napkins : 3   
## napkins : 5 shopping bags : 4 pot plants : 3   
## newspapers : 3 chocolate : 3 candy : 2   
## bottled water: 2 bottled water : 2 hygiene articles: 2   
## (Other) : 46 (Other) : 35 (Other) : 25   
## X.16 X.17   
## :9805 :9816   
## napkins : 4 napkins : 2   
## fruit/vegetable juice : 2 baking powder : 1   
## house keeping products: 2 bottled beer : 1   
## hygiene articles : 2 cleaner : 1   
## candles : 1 cling film/bags: 1   
## (Other) : 18 (Other) : 12   
## X.18 X.19   
## :9820 :9826   
## waffles : 2 bottled beer : 2   
## chocolate marshmallow: 1 bottled water: 1   
## cling film/bags : 1 cake bar : 1   
## dental care : 1 coffee : 1   
## dog food : 1 flour : 1   
## (Other) : 8 (Other) : 2   
## X.20 X.21   
## :9827 :9827   
## chocolate : 2 chocolate : 1   
## fruit/vegetable juice : 1 female sanitary products: 1   
## liquor (appetizer) : 1 long life bakery product: 1   
## long life bakery product: 1 margarine : 1   
## pasta : 1 rum : 1   
## white wine : 1 (Other) : 2   
## X.22 X.23   
## :9828 :9829   
## abrasive cleaner : 1 chocolate : 1   
## chocolate : 1 hygiene articles: 1   
## hygiene articles : 2 napkins : 2   
## long life bakery product: 1 sugar : 1   
## specialty fat : 1   
##   
## X.24 X.25 X.26   
## :9830 :9833 :9833   
## cooking chocolate : 1 skin care: 1 hygiene articles: 1   
## house keeping products: 2   
## soups : 1   
##   
##   
##   
## X.27   
## :9833   
## candles: 1   
##   
##   
##   
##   
##

## Association rules for the grocery store

The grocery dataset shows the items that have been bought togheter. Now we going to see how often they are bought together , and rules.

We can then see how many transactions we have and what they are.

## [1] "Description of the transactions"

## transactions in sparse format with  
## 9835 transactions (rows) and  
## 169 items (columns)

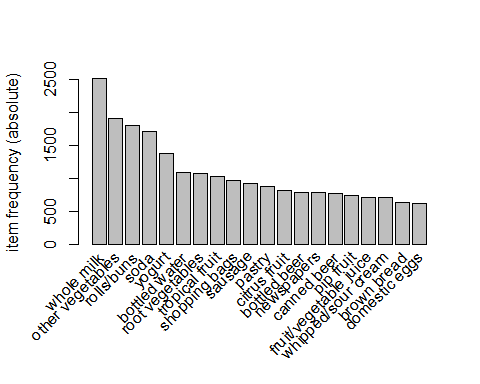
## transactions as itemMatrix in sparse format with  
## 9835 rows (elements/itemsets/transactions) and  
## 169 columns (items) and a density of 0.02609146   
##   
## most frequent items:  
## whole milk other vegetables rolls/buns soda   
## 2513 1903 1809 1715   
## yogurt (Other)   
## 1372 34055   
##   
## element (itemset/transaction) length distribution:  
## sizes  
## 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15   
## 2159 1643 1299 1005 855 645 545 438 350 246 182 117 78 77 55   
## 16 17 18 19 20 21 22 23 24 26 27 28 29 32   
## 46 29 14 14 9 11 4 6 1 1 1 1 3 1   
##   
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 2.000 3.000 4.409 6.000 32.000   
##   
## includes extended item information - examples:  
## labels  
## 1 abrasive cleaner  
## 2 artif. sweetener  
## 3 baby cosmetics

We see 9835 transactions, this is the number of rows as well, and 169 items.

The summary gives us some useful information:

* density: The percentage of non-empty cells in the sparse matrix. In another word, the total number of items that was purchased divided by the total number of possible items in that matrix.
* Transactions sizes There are 2159 transactions with one item, 1643 with 2 items, 1299 with 3 items, 1299 with items. This indicates that most customers buy small number of items on each purchase.
* The data distribution is right skewed.

Let’s have a look item freqnency plot.



## Create some rules

* We use the Apriori algorithm in arules library to mine frequent itemsets and association rules. The algorithm employs level-wise search for frequent itemsets.
* We pass supp=0.001 and conf=0.8 to return all the rules have a support of at least 0.1% and confidence of at least 80%.
* We sort the rules by decreasing confidence.
* Have a look the summary of the rules.

## 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 FALSE  
##   
## Algorithmic control:  
## filter tree heap memopt load sort verbose  
## 0.1 TRUE TRUE FALSE TRUE 2 TRUE  
##   
## Absolute minimum support count: 9   
##   
## set item appearances ...[0 item(s)] done [0.00s].  
## set transactions ...[169 item(s), 9835 transaction(s)] done [0.00s].  
## sorting and recoding items ... [157 item(s)] done [0.00s].  
## creating transaction tree ... done [0.00s].  
## checking subsets of size 1 2 3 4 5 6 done [0.02s].  
## writing ... [410 rule(s)] done [0.00s].  
## creating S4 object ... done [0.00s].

## set of 410 rules  
##   
## rule length distribution (lhs + rhs):sizes  
## 3 4 5 6   
## 29 229 140 12   
##   
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 3.000 4.000 4.000 4.329 5.000 6.000   
##   
## summary of quality measures:  
## support confidence lift count   
## Min. :0.001017 Min. :0.8000 Min. : 3.131 Min. :10.00   
## 1st Qu.:0.001017 1st Qu.:0.8333 1st Qu.: 3.312 1st Qu.:10.00   
## Median :0.001220 Median :0.8462 Median : 3.588 Median :12.00   
## Mean :0.001247 Mean :0.8663 Mean : 3.951 Mean :12.27   
## 3rd Qu.:0.001322 3rd Qu.:0.9091 3rd Qu.: 4.341 3rd Qu.:13.00   
## Max. :0.003152 Max. :1.0000 Max. :11.235 Max. :31.00   
##   
## mining info:  
## data ntransactions support confidence  
## transactions 9835 0.001 0.8

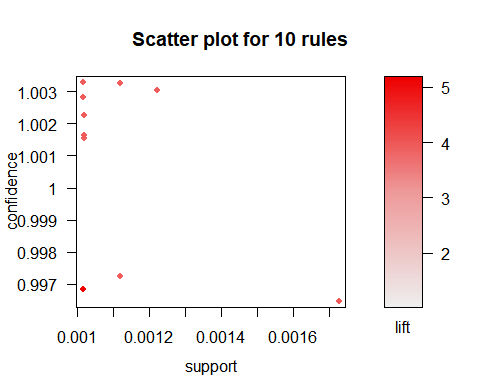
* The number of rules: 9835.
* The distribution of rules by length: Most rules are 4 items long.
* The summary of quality measures: ranges of support, confidence, and lift.
* The information on the data mining: total data mined, and minimum parameters we set earlier.

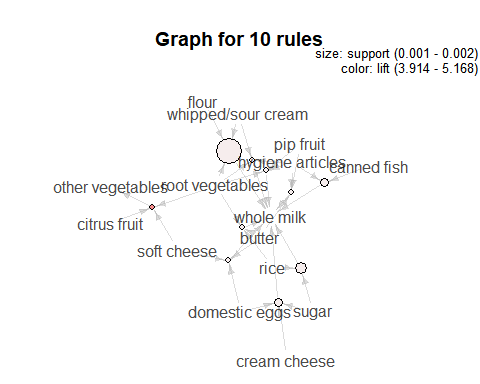
let’s inspect top 10.

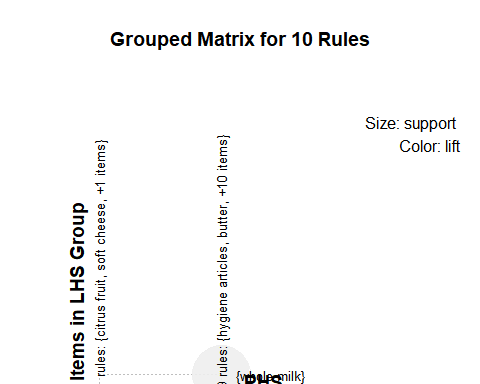
## lhs rhs support confidence lift count  
## [1] {rice,   
## sugar} => {whole milk} 0.001220132 1 3.913649 12  
## [2] {canned fish,   
## hygiene articles} => {whole milk} 0.001118454 1 3.913649 11  
## [3] {butter,   
## rice,   
## root vegetables} => {whole milk} 0.001016777 1 3.913649 10  
## [4] {flour,   
## root vegetables,   
## whipped/sour cream} => {whole milk} 0.001728521 1 3.913649 17  
## [5] {butter,   
## domestic eggs,   
## soft cheese} => {whole milk} 0.001016777 1 3.913649 10  
## [6] {citrus fruit,   
## root vegetables,   
## soft cheese} => {other vegetables} 0.001016777 1 5.168156 10  
## [7] {butter,   
## hygiene articles,   
## pip fruit} => {whole milk} 0.001016777 1 3.913649 10  
## [8] {hygiene articles,   
## root vegetables,   
## whipped/sour cream} => {whole milk} 0.001016777 1 3.913649 10  
## [9] {hygiene articles,   
## pip fruit,   
## root vegetables} => {whole milk} 0.001016777 1 3.913649 10  
## [10] {cream cheese,   
## domestic eggs,   
## sugar} => {whole milk} 0.001118454 1 3.913649 11

* 100% customers who bought {rice, sugar} end up bought {whole milk} as well.
* 100% customers who bought {canned fish,hygiene articles} end up bought {whole milk} as well.

And plot these top 10 rules.







In this post, we have learned how to Perform Market Basket Analysis in R and how to interpret the results.

reference: [R and Data Mining](http://www.rdatamining.com/examples/association-rules)