Recap on transactions

MARKET BASKET ANALYSIS IN R

Christopher Bruffaerts
Statistician

datacamp

Important points in market basket analysis

Market basket analysis

Focus on the **what**, not on the **how much**;
i.e. what do customers have in their baskets?

Main metrics

- Support
- Confidence
- Lift

A word of caution

The set of extracted rules can be very large! Do not inspect or display all rules in that case - always use a subset of rules or use the functions *head* or *tail*!

Groceries dataset

Let's go back to the Grocery store

Dataset from arules package

```
# Loading the arules package
library(arules)

# Loading the Groceries dataset
data(Groceries)
```

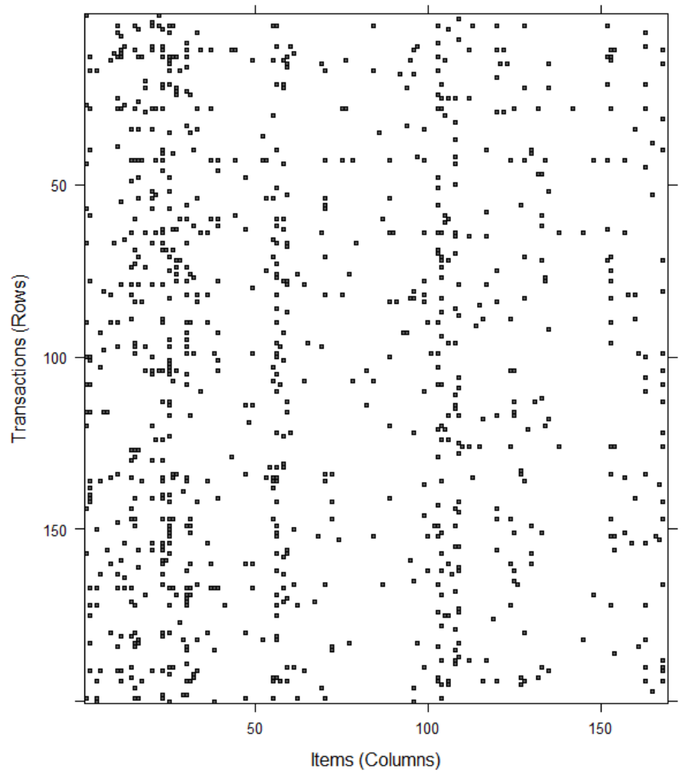
```
summary(Groceries)
```

Summary of Groceries

```
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           yogurt
            2513             1903             1809             1715             1372
         (Other)
           34055

element (itemset/transaction) length distribution:
sizes
   1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17
2159 1643 1299 1005  855  645  545  438  350  246  182  117   78   77   55   46   29
  18   19   20   21   22   23   24   26   27   28   29   32
  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  level2           level1
1 frankfurter sausage meat and sausage
2     sausage sausage meat and sausage
3  liver loaf sausage meat and sausage
```

Density of Groceries

```
# Plotting a sample of 200 transactions
image(sample(Groceries, 200))
```

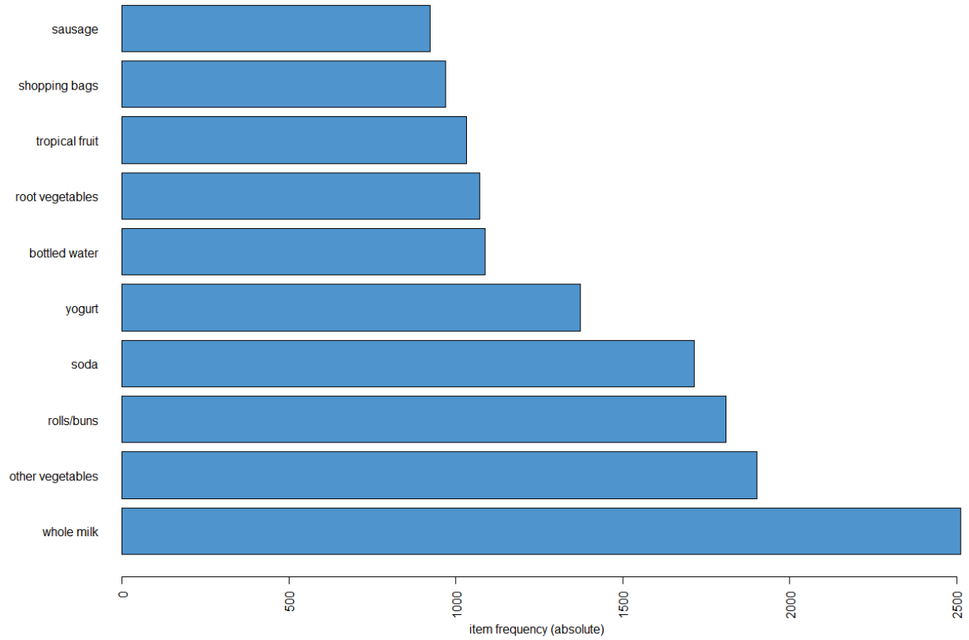


[1] The density of the item matrix is of 2.6%.

Most and least popular items

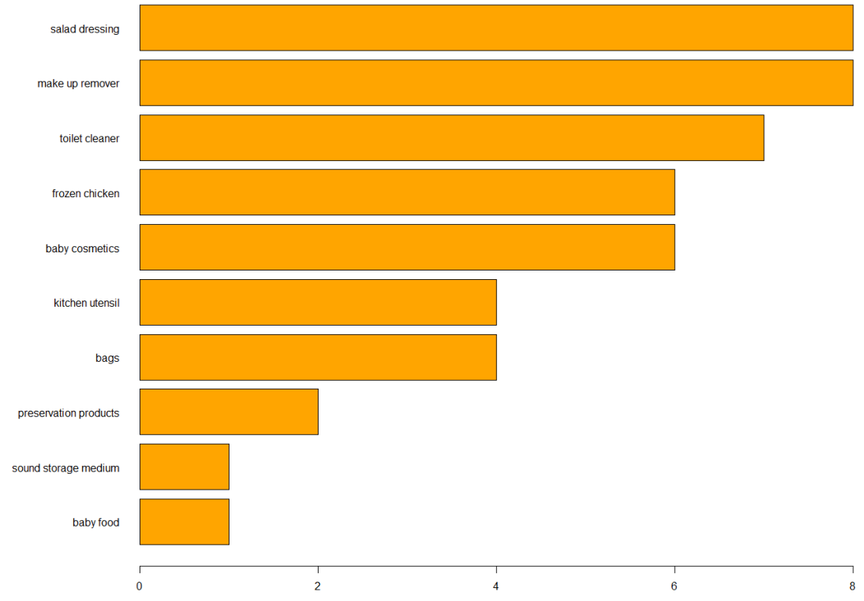
Most popular items

```
itemFrequencyPlot(Groceries,type="relative",
                  topN=10,horiz=TRUE,col='steelblue3')
```



Least popular items

```
par(mar=c(2,10,2,2), mfrow=c(1,1))
barplot(sort(table(unlist(LIST(Groceries))))[1:10],
        horiz = TRUE,las = 1,col='orange')
```



Cross tables by index

Contingency tables

```
# Contingency table
tbl = crossTable(Groceries)
tbl[1:4,1:4]
```

```
             frankfurter sausage liver loaf ham
frankfurter          580      99          7  25
sausage               99     924         10  49
liver loaf             7      10         50   3
ham                   25      49          3 256
```

Sorted contingency table

```
# Sorted contingency table
tbl = crossTable(Groceries, sort = TRUE)
tbl[1:4,1:4]
```

```
                 whole milk other vegetables rolls/buns s
whole milk             2513              736        557
other vegetables        736             1903        419
rolls/buns              557              419       1809
soda                    394              322        377
```

Cross tables by item names

Contingency tables

```
# Counts
tbl['whole milk','flour']
```

```
[1] 83
```

```
# Chi-square test
crossTable(Groceries, measure='chi')['whole milk', 'flour']
```

```
[1] 0.003595389
```

Contingency tables with other metrics

```
crossTable(Groceries, measure='lift',sort=T)[1:4,1:4]
```

```
                 whole milk other vegetables rolls/buns      soda
whole milk               NA        1.5136341   1.205032 1.571735
other vegetables  1.5136341               NA   1.197047 0.9703476
rolls/buns        1.2050318        1.1970465         NA 1.1951242
soda              0.8991124        0.9703476   1.195124        NA
```

MovieLens dataset

MovieLens: Web-based recommender system that recommends movies for its users to watch.

Let's watch movies!

MARKET BASKET ANALYSIS IN R

Mining association rules

MARKET BASKET ANALYSIS IN R

Christopher Bruffaerts
Statistician

datacamp

Frequent itemsets with the apriori

Extracting frequent itemsets of min size 2

```
# Extract the set of most frequent itemsets
itemsets_freq2 =
  apriori(Groceries,
          parameter = list(supp = 0.01,
                           minlen = 2,
                           target = 'frequent'
                          ))
```

Sorting and inspecting frequent itemsets

```
inspect(head(sort(itemsets_freq2, by="support")))
```

```
   items                                  support    count
[1] {other vegetables,whole milk}      0.07483477 736
[2] {whole milk,rolls/buns}            0.05663447 557
[3] {whole milk,yogurt}                0.05602440 551
[4] {root vegetables,whole milk}       0.04890696 481
[5] {root vegetables,other vegetables} 0.04738180 466
[6] {other vegetables,yogurt}          0.04341637 427
```

Rules with the apriori

```
rules = apriori(Groceries, parameter = list(supp=.001,
                                            conf=.5,
                                            minlen=2,
                                            target='rules'
                ))
```

```
inspect(head(sort(rules, by="confidence")))
```

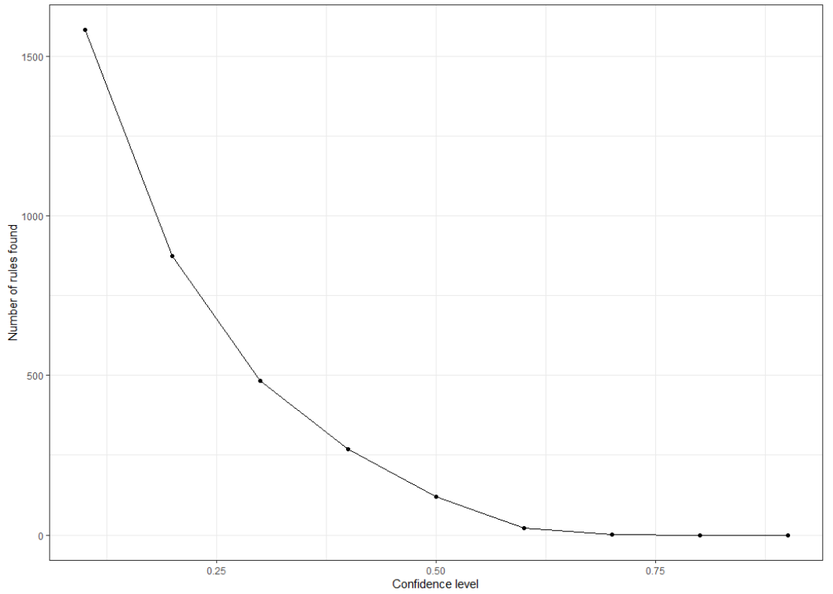
```
   lhs                                            rhs                support     confidence lift     count
[1] {rice,sugar}                                => {whole milk}       0.001220132 1          3.913649 12
[2] {canned fish,hygiene articles}              => {whole milk}       0.001118454 1 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

Choose parameters arules

Looping over different confidence values

```
# Set of confidence levels
confidenceLevels = seq(from=0.1, to=0.9, by =0.1)

# Create empty vector
rules_sup0005 = NULL

# Apriori algorithm with a support level of 0.5%
for (i in 1:length(confidenceLevels)) {
  rules_sup0005[i] =
    length(apriori(Groceries,
                   parameter=list(supp=0.005,
                                  conf=confidenceLevels[
                                  target="rules")))
}
```

```
library(ggplot2)
# Number of rules found with a support level of 0.5%
qplot(confidenceLevels, rules_sup0005,
      geom=c("point", "line"),xlab="Confidence level",
      ylab="Number of rules found") +
  theme_bw()
```



Subsetting rules

```
# Subsetting rules
inspect(subset(rules, subset =
               items %in% c("soft cheese","whole milk") &
               confidence >.95))
```

```
     lhs                                    rhs                support     confidence lift
[1]  {rice,sugar}                        => {whole milk}       0.001220132 1          3.9136
[2]  {canned fish,hygiene articles}      => {whole milk}       0.001118454 1          3.9136
[3]  {root vegetables,butter,rice}       => {whole milk}       0.001016777 1          3.9136
```

Flexibility of subsetting

```
inspect(subset(rules, subset=items %ain% c("soft cheese","whole milk") & confidence >.95))
```

```
inspect(subset(rules, subset=rhs %in% "whole milk" & lift >3 & confidence >0.95))
```

Let's mine the movie dataset!

MARKET BASKET ANALYSIS IN R

Visualizing transactions and rules

MARKET BASKET ANALYSIS IN R

Christopher Bruffaerts
Statistician

datacamp

Interactive inspection

Rule extraction

```
rules = apriori(Groceries,
                parameter = list(
                  supp=.001,
                  conf=.5,
                  minlen=2,
                  target='rules'
                 ))
```

```
# Datatable inspection
inspectDT(rules)
```

HTML table

Show 10 entries Search:

	LHS	RHS	support	confidence	lift	count
	All	All	All	All	All	All
[1]	{honey}	{whole milk}	0.001	0.733	2.870	11.000
[2]	{tidbits}	{rolls/buns}	0.001	0.522	2.837	12.000
[3]	{cocoa drinks}	{whole milk}	0.001	0.591	2.313	13.000
[4]	{pudding powder}	{whole milk}	0.001	0.565	2.212	13.000
[5]	{cooking chocolate}	{whole milk}	0.001	0.520	2.035	13.000
[6]	{cereals}	{whole milk}	0.004	0.643	2.516	36.000
[7]	{jam}	{whole milk}	0.003	0.547	2.141	29.000
[8]	{specialty cheese}	{other vegetables}	0.004	0.500	2.584	42.000
[9]	{rice}	{other vegetables}	0.004	0.520	2.687	39.000
[10]	{rice}	{whole milk}	0.005	0.613	2.400	46.000

Interactive scatterplots

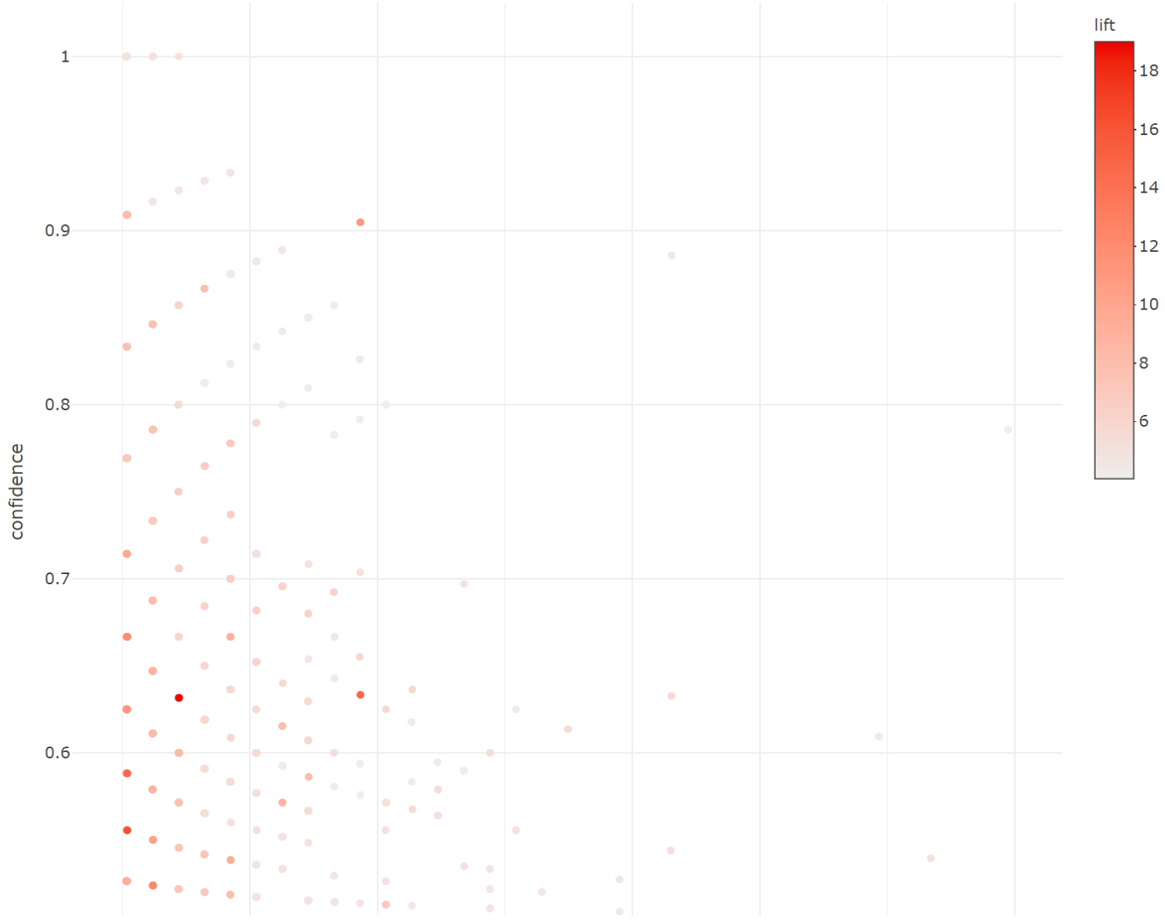
Plot from arulesViz

```
# Plot rules as scatterplot
plot(rules, method = "scatterplot",
     engine = "html")
```

Other types of plots using `method` :

- two-key plot
- grouped
- matrix

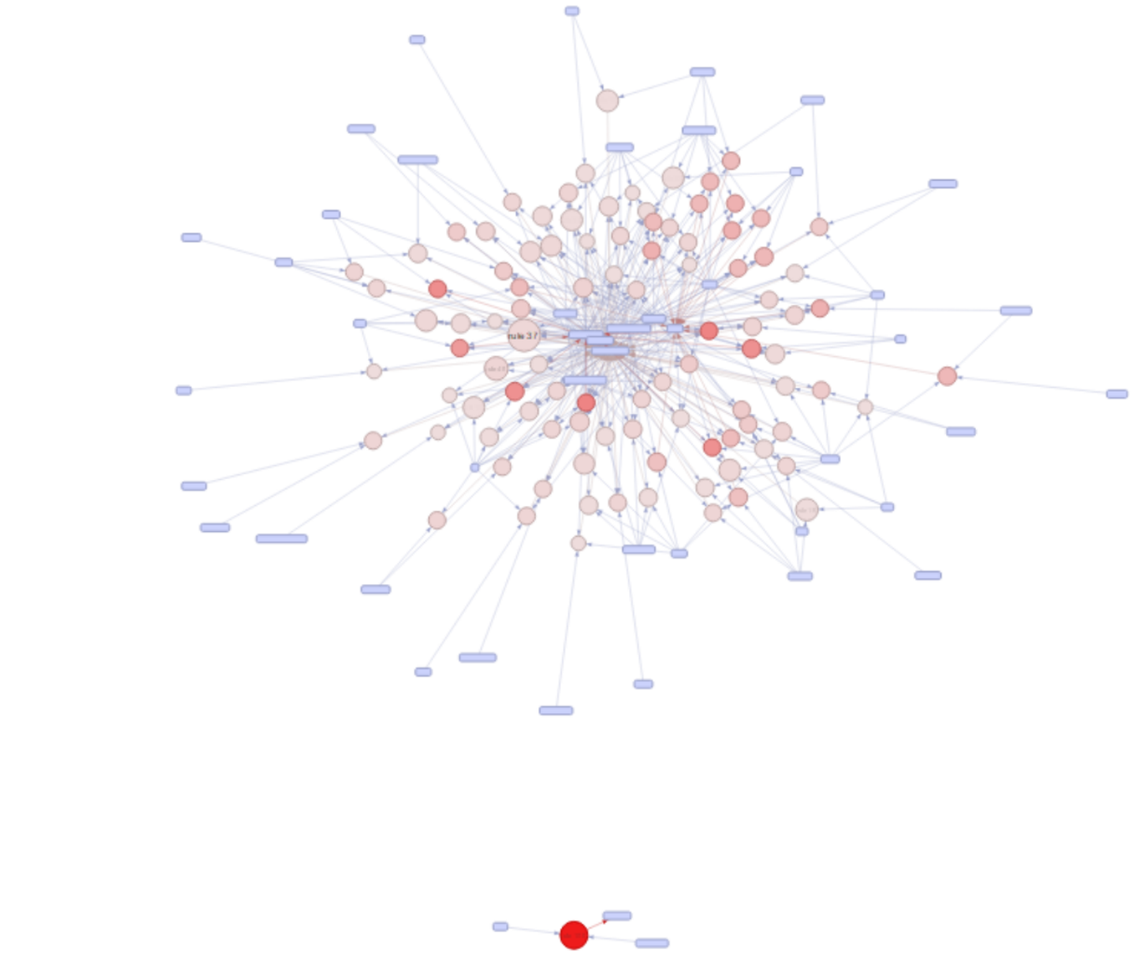
Scatterplots and others



Interactive graphs

The engine and the method

```
# Plot rules as graph
plot(rules, method = "graph",
     engine = "html")
```

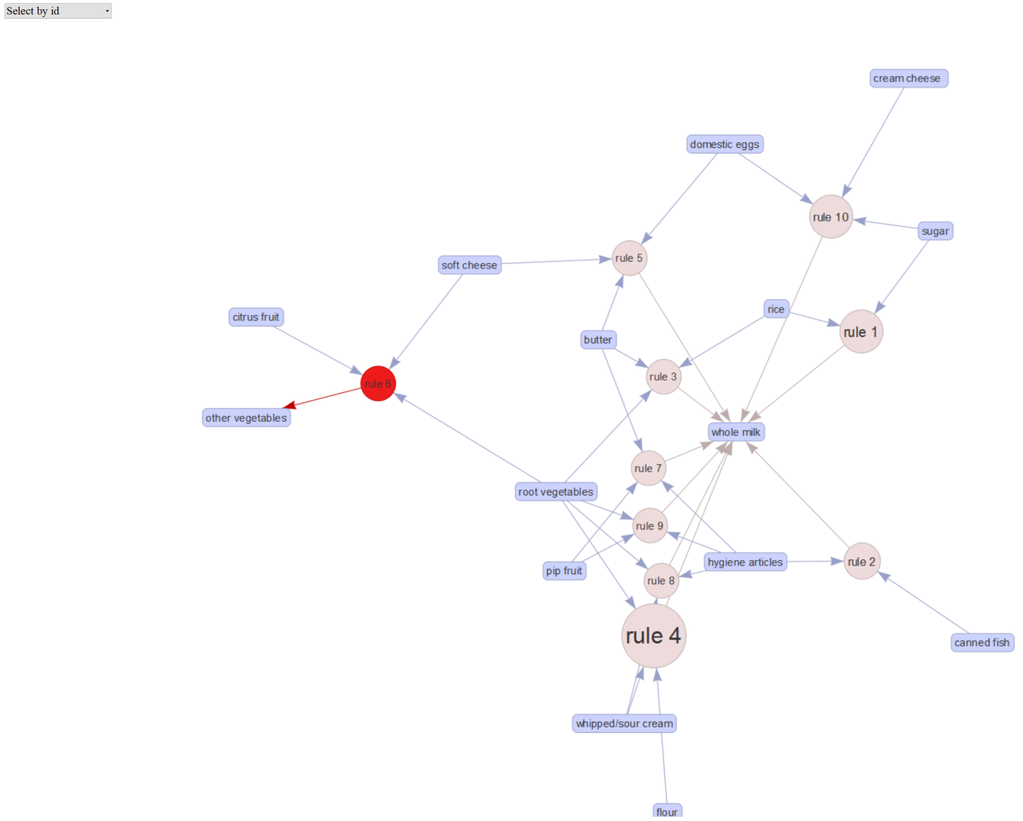


Interactive subgraphs

Sorting extracted rules

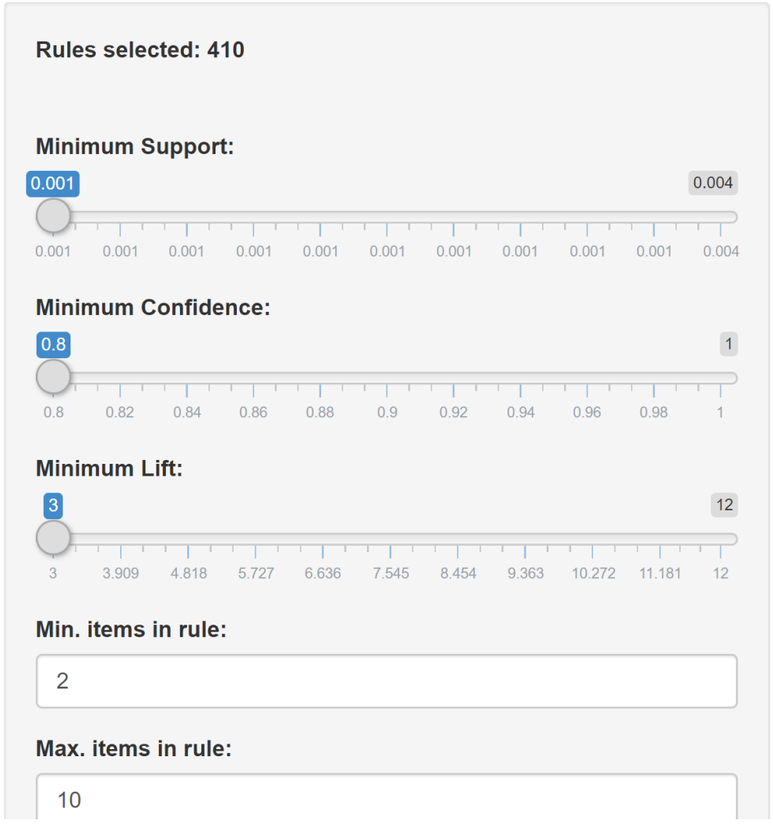
```
# Top 10 rules with highest confidence
top10_rules_Groceries =
    head(sort(rules,by = "confidence"), 10)
inspect(top10_rules_Groceries)
```

```
# Plot the top 10 rules
plot(top10_rules_Groceries,
     method = "graph",  engine = "html")
```



RuleExploring Groceries

```
rules = apriori(Groceries, parameter=list(supp=0.001, conf=0.8))
ruleExplorer(rules)
```



Data Table | Scatter | Matrix | Grouped | Graph

Show 25 entries Search:

LHS	RHS	support	confidence	lift	count
{liquor,red/blush wine}	{bottled beer}	0.001931876	0.9047619	11.235269	19
{curd,cereals}	{whole milk}	0.001016777	0.9090909	3.557863	10
{yogurt,cereals}	{whole milk}	0.001728521	0.8095238	3.168192	17
{butter,jam}	{whole milk}	0.001016777	0.8333333	3.261374	10
{soups,bottled beer}	{whole milk}	0.001118454	0.9166667	3.587512	11
{napkins,house keeping products}	{whole milk}	0.001321810	0.8125000	3.179840	13
{whipped/sour cream,house keeping products}	{whole milk}	0.001220132	0.9230769	3.612599	12
{pastry,sweet spreads}	{whole milk}	0.001016777	0.9090909	3.557863	10
{turkey,curd}	{other vegetables}	0.001220132	0.8000000	4.134524	12
{rice,sugar}	{whole milk}	0.001220132	1.0000000	3.913649	12

Let's visualize some movie rules!

MARKET BASKET ANALYSIS IN R

Making the most of market basket analysis

MARKET BASKET ANALYSIS IN R

Christopher Bruffaerts
Statistician

datacamp

Market basket in practice

Understanding customers/users

- Understand which items are purchased in combination
- Extract sets of rules
- Infer on the relationship between items

The extra mile to MBA

- Add customer/user information
- Segment (cluster) customers according to their preferences

Recommendations to customers/users

- **Offline world**: placing items strategically in the shop such that items often purchased together are close to each other.
- **Online world**: expose related items on the same page, just a click-away.

What influenced yogurt ?

Yogurt as a consequent

```
# Extract rules with Yogurt on the right side
yogurt_rules_rhs = apriori(Groceries, parameter = list(supp = 0.001,conf = 0.8),
                           appearance = list(default = "lhs",rhs = "yogurt"))
```

```
# Find first rules with highest lift
inspect(head(sort(yogurt_rules_rhs, by="lift")))
```

```
    lhs                                                rhs      support     confidence lift     count
[1] {root vegetables,butter,cream cheese }          => {yogurt} 0.001016777 0.9090909  6.516698 10
[2] {tropical fruit,whole milk,butter,sliced cheese} => {yogurt} 0.001016777 0.9090909  6.516698 10
[3] {other vegetables,curd,whipped/sour cream,cream cheese } => {yogurt} 0.001016777 0.9090909  6.516698 10
[4] {tropical fruit,other vegetables,butter,white bread} => {yogurt} 0.001016777 0.9090909  6.516698 10
[5] {sausage,pip fruit,sliced cheese}               => {yogurt} 0.001220132 0.8571429  6.144315 12
[6] {tropical fruit,whole milk,butter,curd}         => {yogurt} 0.001220132 0.8571429  6.144315 12
```

What did yogurt influence?

Yogurt as an antecedent

```
# Extract rules with Yogurt on the left side
yogurt_rules_lhs = apriori(Groceries, parameter = list(supp = 0.001, conf = 0.8, minlen = 2),
                                  appearance = list(default = "rhs",lhs = "yogurt"))
```

```
# Summary of rules
summary(yogurt_rules_lhs)
```

```
set of 0 rules
```

Let's find out recommendations for movies!

MARKET BASKET ANALYSIS IN R

Use your market basket skills!

MARKET BASKET ANALYSIS IN R

Christopher Bruffaerts
Statistician

datacamp

Recap of market basket analysis

- **Chapter 1**: Introduction to market basket analysis
- **Chapter 2**: Metrics and techniques in market basket analysis
- **Chapter 3**: Visualization in market basket analysis
- **Chapter 4**: Case study: Movie recommendations @ movieLens

Other points to consider with MBA

Not in the scope of this course

- Time dimension, *e.g.* when transactions were done, when a user watched a movie
- Qualitative assessment of transactions, *e.g.* movie ratings

Be careful when using the apriori() function

- Use sorting options, head and tail
- Do not print blindly rules
- Work with smaller subsets of rules

Congratulations!

MARKET BASKET ANALYSIS IN R