Association Rule Mining using Groceries Dataset

Sahitya Sundar Raj Vijayanagar

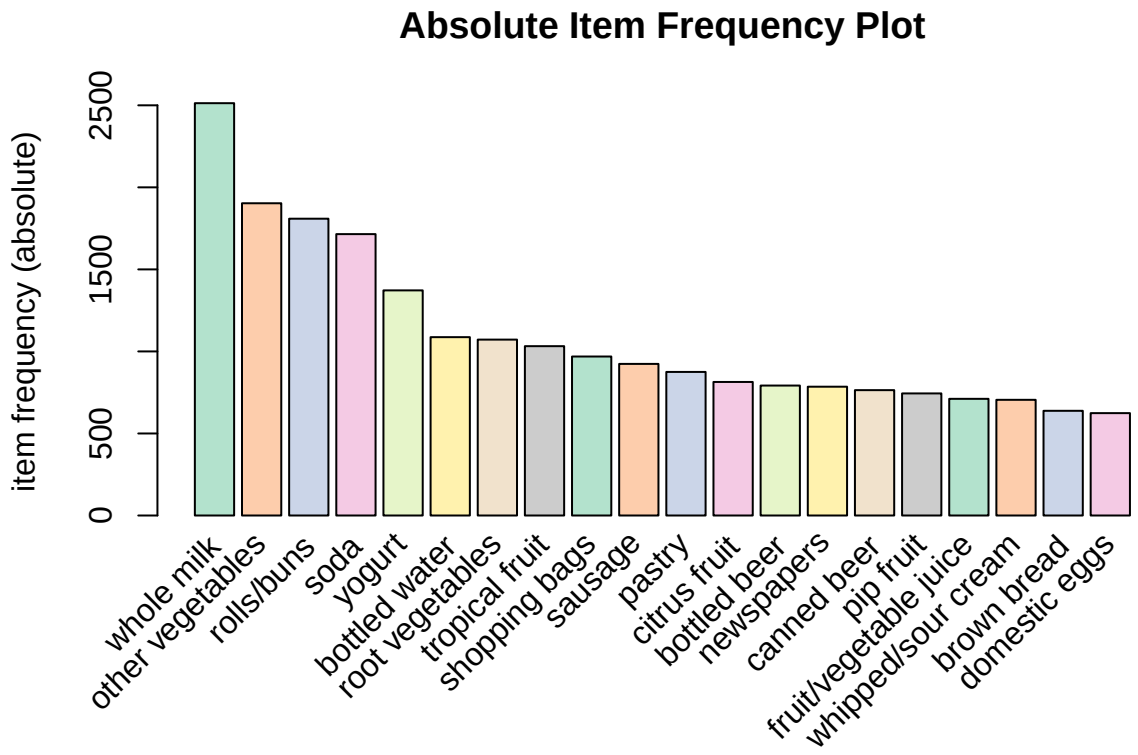
11/08/2021

Load the dataset from 'groceries.txt' as transactions of format basket. As shown, there are 2159 transactions with just one item in the basket, 1643 transactions with two items in the basket, and so on.

```
## Formal class 'transactions' [package "arules"] with 3 slots
##   ..@ data       :Formal class 'ngCMatrix' [package "Matrix"] with 5 slots
##   .. .. ..@ i       : int [1:43367] 29 88 118 132 33 157 167 166 38 91 ...
##   .. .. ..@ p       : int [1:9836] 0 4 7 8 12 16 21 22 27 28 ...
##   .. .. ..@ Dim     : int [1:2] 169 9835
##   .. .. ..@ Dimnames:List of 2
##   .. .. .. ..$ : NULL
##   .. .. .. ..$ : NULL
##   .. .. ..@ factors : list()
##   ..@ itemInfo   :'data.frame':  169 obs. of  1 variable:
##   .. ..$ labels: chr [1:169] "abrasive cleaner" "artif. sweetener" "baby cosmetics" "baby food" ...
##   ..@ itemsetInfo:'data.frame':  0 obs. of  0 variables
```

```
## 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   16
## 2159 1643 1299 1005  855  645  545  438  350  246  182  117   78   77   55   46
##   17   18   19   20   21   22   23   24   26   27   28   29   32
##   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
```

Create the Item frequency plot to indicate the frequencies for different bought items. As shown, whole milk is the most frequently bought grocery item, followed by other vegetables and rolls/buns.



Creating association mining rules by randomly selecting minimum support as 0.001, confidence as 0.8, and max length=10, results in 410 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 TRUE
##
## 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 [0.01s].
## 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        coverage             lift
##  Min.   :0.001017   Min.   :0.8000   Min.   :0.001017   Min.   : 3.131
##  1st Qu.:0.001017   1st Qu.:0.8333   1st Qu.:0.001220   1st Qu.: 3.312
##  Median :0.001220   Median :0.8462   Median :0.001322   Median : 3.588
##  Mean   :0.001247   Mean   :0.8663   Mean   :0.001449   Mean   : 3.951
##  3rd Qu.:0.001322   3rd Qu.:0.9091   3rd Qu.:0.001627   3rd Qu.: 4.341
##  Max.   :0.003152   Max.   :1.0000   Max.   :0.003559   Max.   :11.235
##      count
##  Min.   :10.00
##  1st Qu.:10.00
##  Median :12.00
##  Mean   :12.27
##  3rd Qu.:13.00
##  Max.   :31.00
##
## mining info:
##       data ntransactions support confidence
##  groceries          9835   0.001        0.8
```

```
##      lhs                          rhs                    support confidence    coverage      lift cou
## [1]  {liquor,
##       red/blush wine}         => {bottled beer}     0.001931876  0.9047619 0.002135231 11.235269    1
## [2]  {cereals,
##       curd}                   => {whole milk}       0.001016777  0.9090909 0.001118454  3.557863    1
## [3]  {cereals,
##       yogurt}                 => {whole milk}       0.001728521  0.8095238 0.002135231  3.168192    1
## [4]  {butter,
##       jam}                    => {whole milk}       0.001016777  0.8333333 0.001220132  3.261374    1
## [5]  {bottled beer,
##       soups}                  => {whole milk}       0.001118454  0.9166667 0.001220132  3.587512    1
## [6]  {house keeping products,
##       napkins}                => {whole milk}       0.001321810  0.8125000 0.001626843  3.179840    1
## [7]  {house keeping products,
##       whipped/sour cream}     => {whole milk}       0.001220132  0.9230769 0.001321810  3.612599    1
## [8]  {pastry,
##       sweet spreads}          => {whole milk}       0.001016777  0.9090909 0.001118454  3.557863    1
## [9]  {curd,
##       turkey}                 => {other vegetables} 0.001220132  0.8000000 0.001525165  4.134524    1
## [10] {rice,
##       sugar}                  => {whole milk}       0.001220132  1.0000000 0.001220132  3.913649    1
```

Looking at the above result, 100% of customers who bought {rice,sugar} also bought whole milk. Similarly, 90.48% of customers who bought {liquor, red/blush wine} also bought bottled beer.

Trying more stricter rules with conf=0.9 and shorter rules with maxlen=3, results in only 10 rules as shown below:

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##         0.9    0.1    1 none FALSE            TRUE       5   0.001      1
##  maxlen target  ext
##       3  rules TRUE
##
## 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 done [0.00s].
## writing ... [10 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].

## set of 10 rules
##
## rule length distribution (lhs + rhs):sizes
##  3
## 10
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       3       3       3       3       3       3
##
## summary of quality measures:
##     support           confidence        coverage             lift
##  Min.   :0.001017   Min.   :0.9048   Min.   :0.001118   Min.   : 3.558
##  1st Qu.:0.001118   1st Qu.:0.9110   1st Qu.:0.001144   1st Qu.: 3.594
##  Median :0.001118   Median :0.9167   Median :0.001220   Median : 3.763
##  Mean   :0.001210   Mean   :0.9319   Mean   :0.001301   Mean   : 4.647
##  3rd Qu.:0.001220   3rd Qu.:0.9231   3rd Qu.:0.001296   3rd Qu.: 4.532
##  Max.   :0.001932   Max.   :1.0000   Max.   :0.002135   Max.   :11.235
##      count
##  Min.   :10.0
##  1st Qu.:11.0
##  Median :11.0
##  Mean   :11.9
##  3rd Qu.:12.0
##  Max.   :19.0
##
## mining info:
##       data ntransactions support confidence
##  groceries          9835   0.001        0.9
```

```
##      lhs                          rhs                    support confidence    coverage      lift cou
## [1]  {liquor,
##       red/blush wine}         => {bottled beer}     0.001931876  0.9047619 0.002135231 11.235269    1
## [2]  {cereals,
##       curd}                   => {whole milk}       0.001016777  0.9090909 0.001118454  3.557863    1
## [3]  {bottled beer,
##       soups}                  => {whole milk}       0.001118454  0.9166667 0.001220132  3.587512    1
## [4]  {house keeping products,
##       whipped/sour cream}     => {whole milk}       0.001220132  0.9230769 0.001321810  3.612599    1
## [5]  {pastry,
##       sweet spreads}          => {whole milk}       0.001016777  0.9090909 0.001118454  3.557863    1
## [6]  {rice,
##       sugar}                  => {whole milk}       0.001220132  1.0000000 0.001220132  3.913649    1
## [7]  {bottled water,
##       rice}                   => {whole milk}       0.001220132  0.9230769 0.001321810  3.612599    1
## [8]  {canned fish,
##       hygiene articles}       => {whole milk}       0.001118454  1.0000000 0.001118454  3.913649    1
## [9]  {grapes,
##       onions}                 => {other vegetables} 0.001118454  0.9166667 0.001220132  4.737476    1
## [10] {hard cheese,
##       oil}                    => {other vegetables} 0.001118454  0.9166667 0.001220132  4.737476    1
```

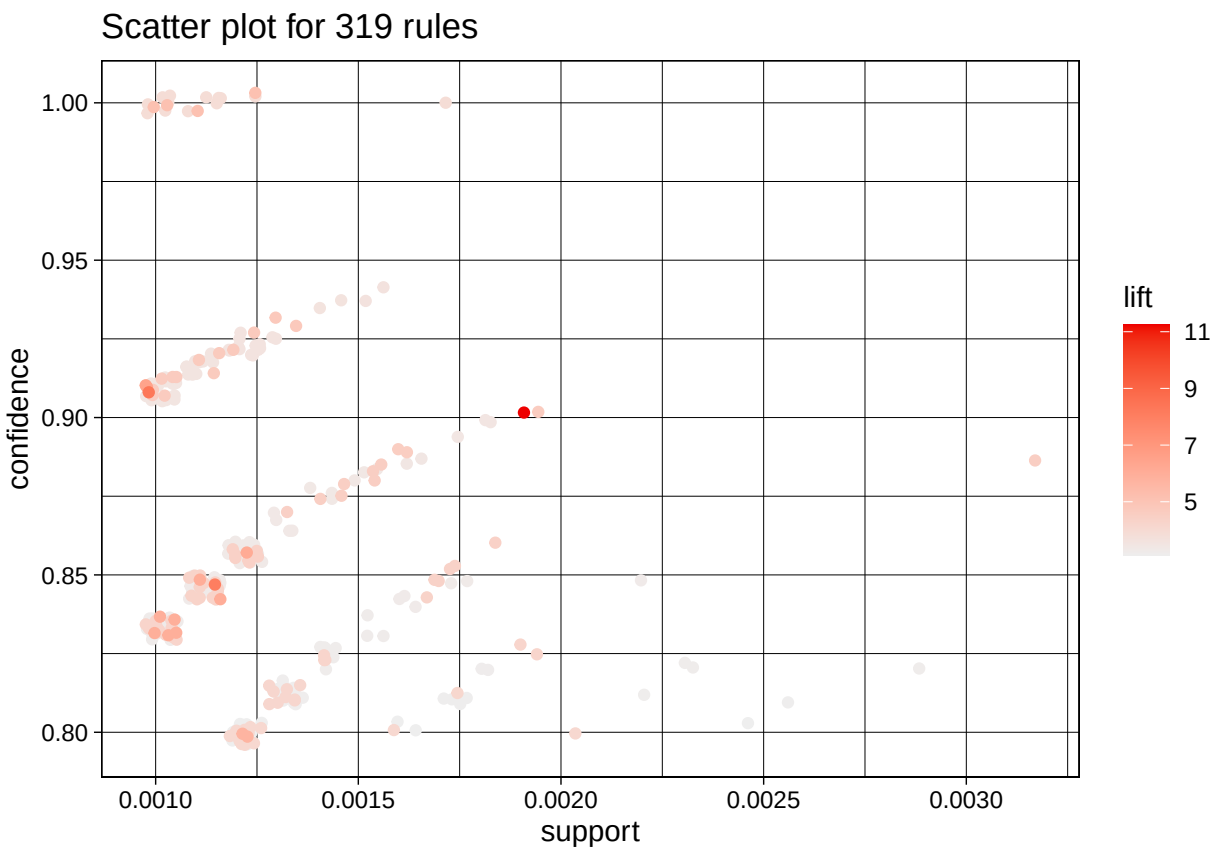
Considering the 410 association rules created by minimum support as 0.001, confidence as 0.8, and max length=10, the next step included removing subsets of larger rules, which resulted in a total of 319 rules.

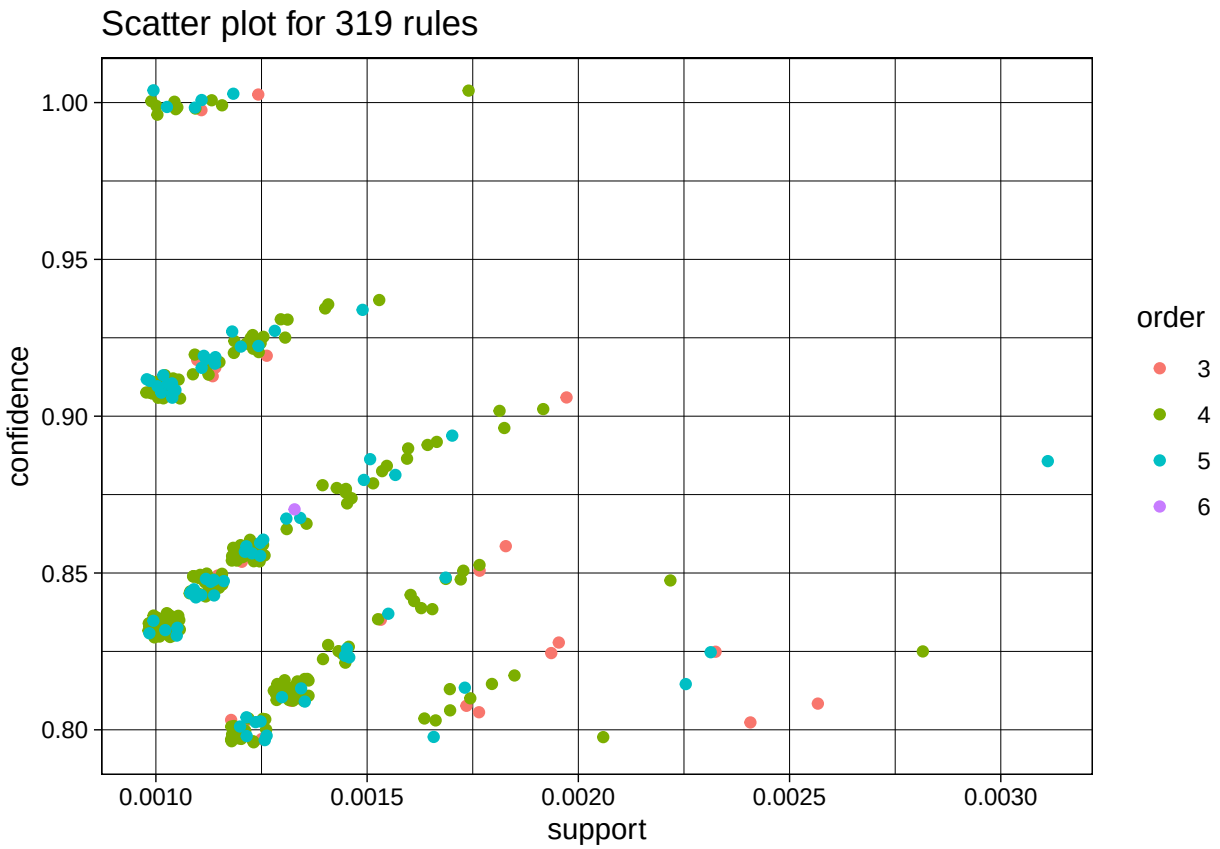
```
## [1] 91
```

```
## set of 319 rules
##
## rule length distribution (lhs + rhs):sizes
##   3   4   5   6
##  29 216  73   1
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   4.000   4.000   4.144   4.000   6.000
##
## summary of quality measures:
##     support           confidence        coverage             lift
##  Min.   :0.001017   Min.   :0.8000   Min.   :0.001017   Min.   : 3.131
##  1st Qu.:0.001017   1st Qu.:0.8235   1st Qu.:0.001220   1st Qu.: 3.261
##  Median :0.001220   Median :0.8462   Median :0.001423   Median : 3.558
##  Mean   :0.001273   Mean   :0.8615   Mean   :0.001486   Mean   : 3.858
##  3rd Qu.:0.001322   3rd Qu.:0.9091   3rd Qu.:0.001627   3rd Qu.: 4.307
##  Max.   :0.003152   Max.   :1.0000   Max.   :0.003559   Max.   :11.235
##      count
##  Min.   :10.00
##  1st Qu.:10.00
##  Median :12.00
##  Mean   :12.52
##  3rd Qu.:13.00
##  Max.   :31.00
##
## mining info:
```

```
##       data ntransactions support confidence
##  groceries          9835   0.001        0.8
```

```
##      lhs                           rhs                    support confidence    coverage      lift cou
## [1]  {liquor,
##       red/blush wine}         => {bottled beer}     0.001931876  0.9047619 0.002135231 11.235269
## [2]  {cereals,
##       curd}                   => {whole milk}       0.001016777  0.9090909 0.001118454  3.557863
## [3]  {cereals,
##       yogurt}                 => {whole milk}       0.001728521  0.8095238 0.002135231  3.168192
## [4]  {butter,
##       jam}                    => {whole milk}       0.001016777  0.8333333 0.001220132  3.261374
## [5]  {bottled beer,
##       soups}                  => {whole milk}       0.001118454  0.9166667 0.001220132  3.587512
## [6]  {house keeping products,
##       napkins}                => {whole milk}       0.001321810  0.8125000 0.001626843  3.179840
## [7]  {house keeping products,
##       whipped/sour cream}     => {whole milk}       0.001220132  0.9230769 0.001321810  3.612599
## [8]  {pastry,
##       sweet spreads}          => {whole milk}       0.001016777  0.9090909 0.001118454  3.557863
## [9]  {curd,
##       turkey}                 => {other vegetables} 0.001220132  0.8000000 0.001525165  4.134524
## [10] {rice,
##       sugar}                  => {whole milk}       0.001220132  1.0000000 0.001220132  3.913649
```





Looking at the most bought item in the item frequency list, i.e.m, 'whole milk', it is possible to find the items that are most likely to be bought before buying whole milk by using appearance. Further,using conf=1, will indicate the items where 100% of customers bought whole milk, after buying these items.

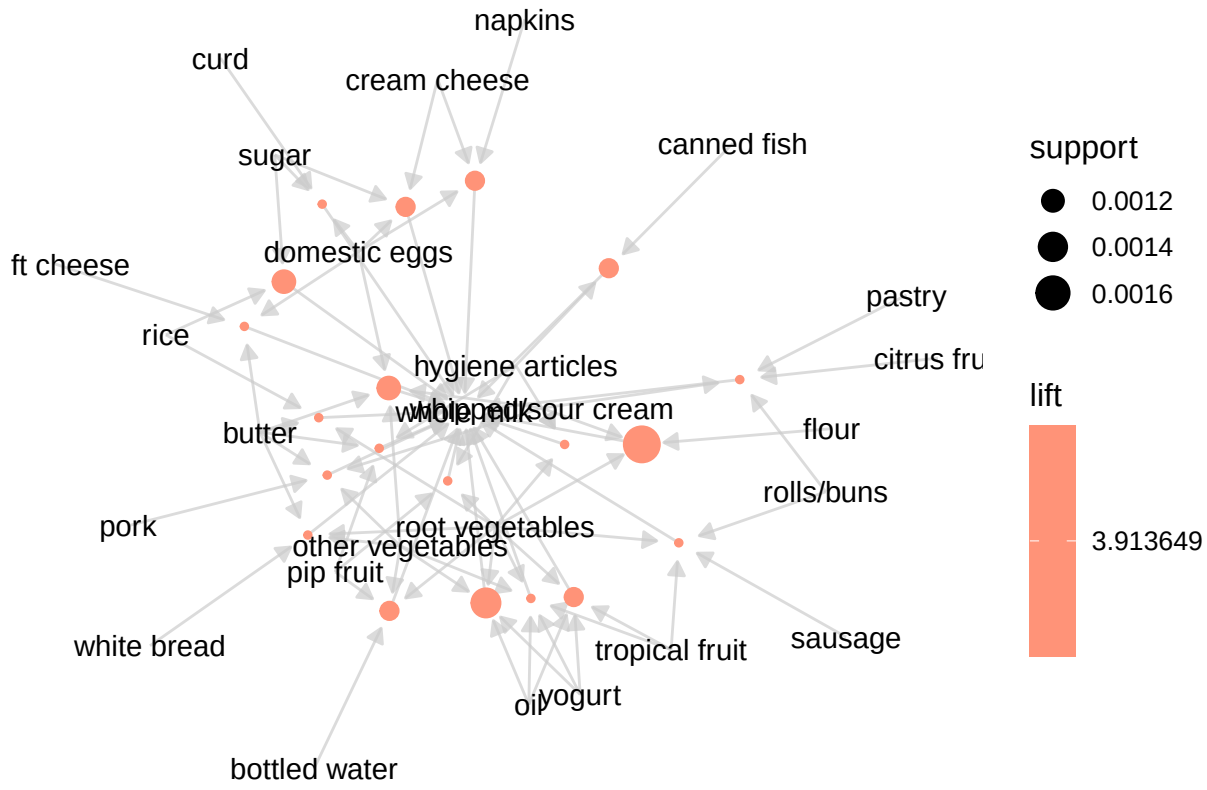
```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##           1    0.1    1 none FALSE            TRUE       5   0.001      1
##  maxlen target  ext
##      10  rules TRUE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 9
##
## set item appearances ...[1 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 [0.01s].
## writing ... [20 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].

## set of 20 rules
```

```
##
## rule length distribution (lhs + rhs):sizes
## 3 4 5 6
## 2 9 8 1
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     3.0     4.0     4.0     4.4     5.0     6.0
##
## summary of quality measures:
##     support           confidence    coverage             lift
##  Min.   :0.001017   Min.   :1    Min.   :0.001017   Min.   :3.914
##  1st Qu.:0.001017   1st Qu.:1    1st Qu.:0.001017   1st Qu.:3.914
##  Median :0.001017   Median :1    Median :0.001017   Median :3.914
##  Mean   :0.001118   Mean   :1    Mean   :0.001118   Mean   :3.914
##  3rd Qu.:0.001118   3rd Qu.:1    3rd Qu.:0.001118   3rd Qu.:3.914
##  Max.   :0.001729   Max.   :1    Max.   :0.001729   Max.   :3.914
##      count
##  Min.   :10
##  1st Qu.:10
##  Median :10
##  Mean   :11
##  3rd Qu.:11
##  Max.   :17
##
## mining info:
##       data ntransactions support confidence
##  groceries          9835   0.001          1
```

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

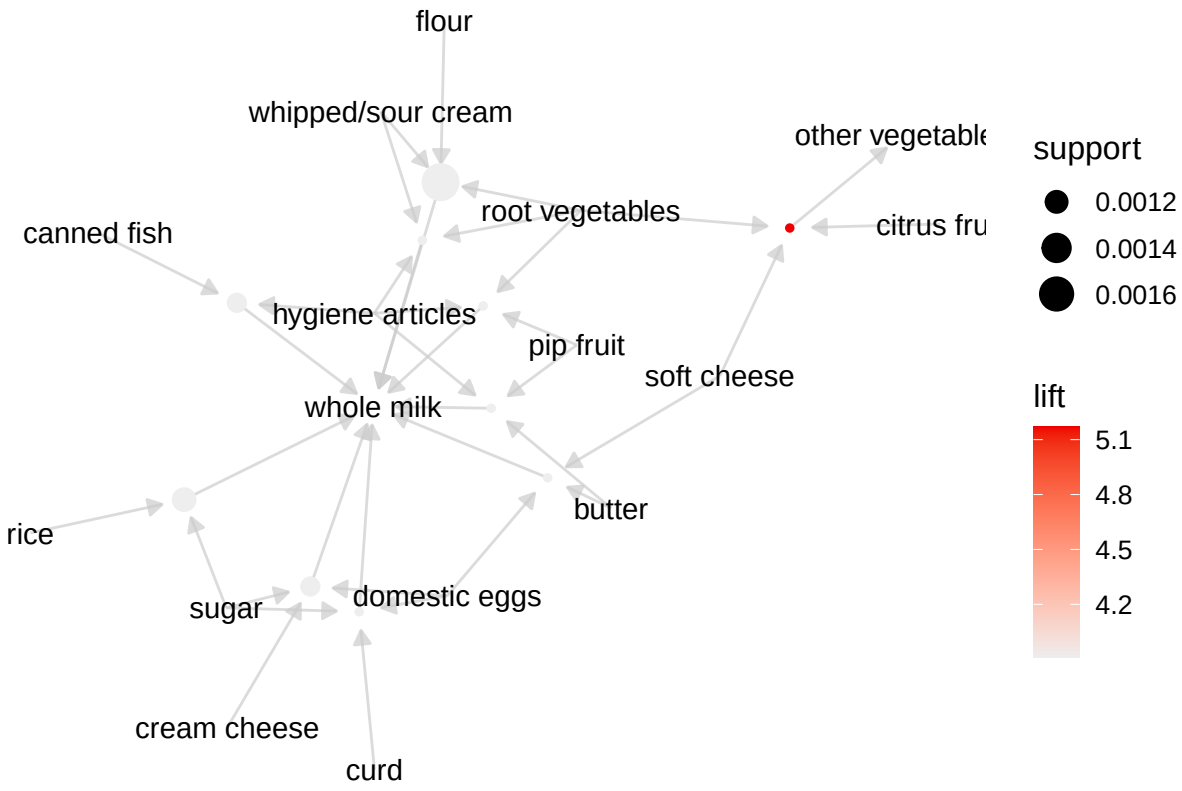
```
## [10] {curd,
##       domestic eggs,
##       sugar}              => {whole milk} 0.001016777          1 0.001016777 3.913649    10
```



As shown above, 100% of customers who bought {rice,sugar}, {canned fish,hygiene articles}, {butter,rice,root vegetables}, etc. have bought whole milk, Similarly, we find 20 such antecedents for whole milk.

Sorting rules based on confidence:

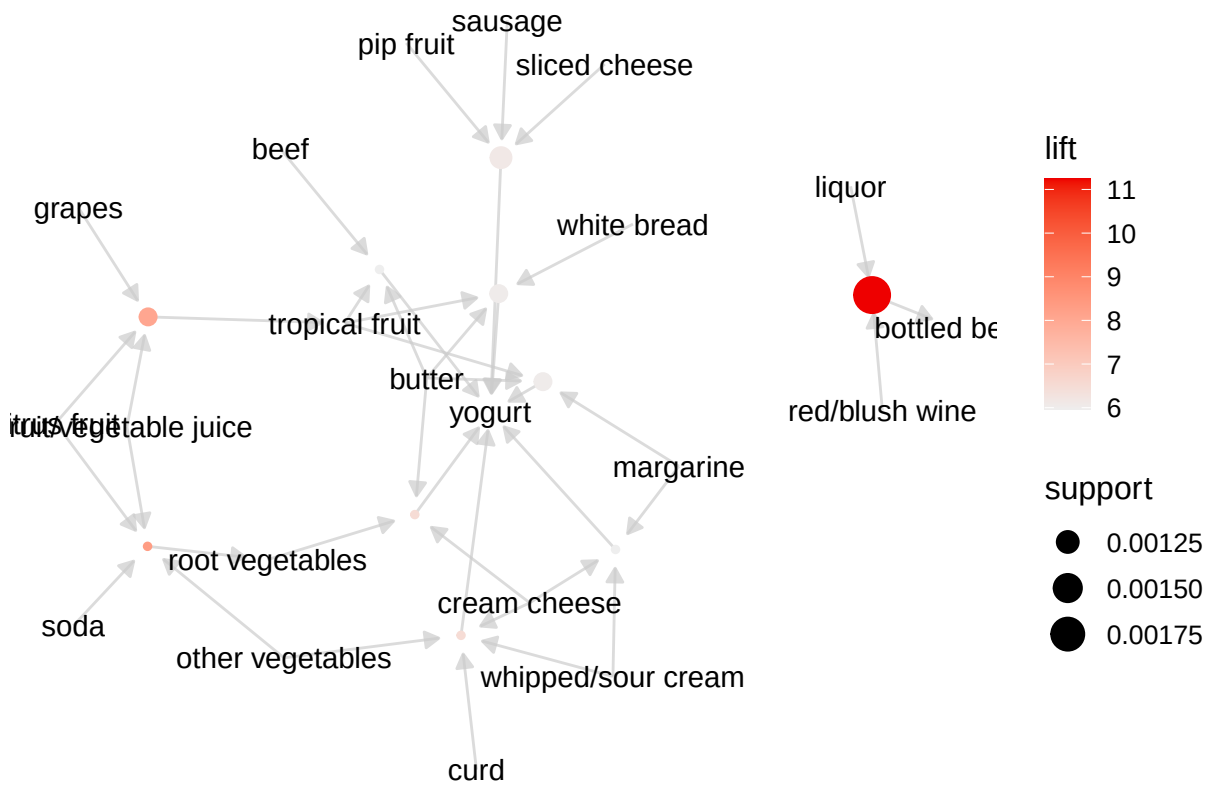
```
##     lhs                     rhs                    support confidence    coverage     lift count
## [1] {rice,
##      sugar}              => {whole milk}       0.001220132          1 0.001220132 3.913649    12
## [2] {canned fish,
##      hygiene articles}   => {whole milk}       0.001118454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



A confidence of 1 indicates that whenever the items on the antecedent are bought, 100% of customers bought the item(s) on the consequent.

Sorting rules based on lift:

```
##     lhs                          rhs                  support confidence    coverage      lift count
## [1] {liquor,
##      red/blush wine}          => {bottled beer}     0.001931876  0.9047619 0.002135231 11.235269    19
## [2] {citrus fruit,
##      fruit/vegetable juice,
##      other vegetables,
##      soda}                    => {root vegetables} 0.001016777  0.9090909 0.001118454  8.340400    10
## [3] {citrus fruit,
##      fruit/vegetable juice,
##      grapes}                  => {tropical fruit}  0.001118454  0.8461538 0.001321810  8.063879    11
## [4] {butter,
##      cream cheese,
##      root vegetables}         => {yogurt}          0.001016777  0.9090909 0.001118454  6.516698    10
## [5] {cream cheese,
##      curd,
##      other vegetables,
##      whipped/sour cream}      => {yogurt}          0.001016777  0.9090909 0.001118454  6.516698    10
## [6] {pip fruit,
##      sausage,
##      sliced cheese}           => {yogurt}          0.001220132  0.8571429 0.001423488  6.144315    12
```



A rule with lift of 11.23 for {liquore,red/blush wine}->{bottled beer} indicates that the items in the antecedent and consequent are ~11 times more likely to be bought together than bought individually.