Association rules:

Association rules are statements that help to find patterns in seemingly unrelated data or a relational database (information repository). Easy example of such would be: If I buy milk, there is 80% probability that I will also buy yogurt"

An association rule has two parts, an antecedent (if) and a consequent (then).

An antecedent is an item found in the data.

A consequent is an item that is found in combination with the antecedent.

Association rules:

Association rules are created by analyzing data for frequent if/then patterns and using the criteria support and confidence to identify the most important relationships.

Support is an indication of how frequently the items appear in the database.

Confidence indicates the number of times the if/then statements have been found to be true.

In data mining, association rules are useful for analyzing and predicting customer behavior. They play an important part in shopping basket data analysis, product clustering, catalog design and store layout.

Groceries dataset provided with arules package in R :

REVA UNIVERSITY

```
> library(datasets)
> data(Groceries)
> transactions <- Groceries
> summary(transactions )
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  level2           level1
1 frankfurter sausage meat and sausage
      sausage sausage meat and sausage
   liver loaf sausage meat and sausage
```

Groceries dataset provided with arules package in R :

REVA UNIVERSITY

I will be using free Groceries dataset provided with rules package. It contains 9835 transactions (rows) and 169 items (columns). The aim of the analysis is to show the methods that allow to obtain certain rules within the dataset.

The most commonly found items in the dataset are: whole milk, other vegetables, rolls/buns, soda, yogurt.

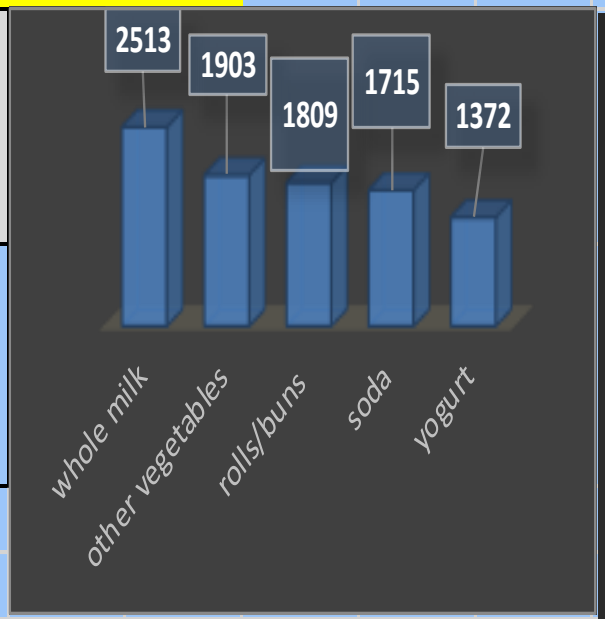
Density of 0.026 means that there are 2.6% non zero cells in the matrix. Matrix has 9835 times 169 = 1662115 cells. Since 2.6% of that are non-zero cells, so 43214.99 items were purchased.

Average transaction consisted of 4.409 items, whereas only one item have been bought in 2159 transactions. Maximum number of items bought was 32.

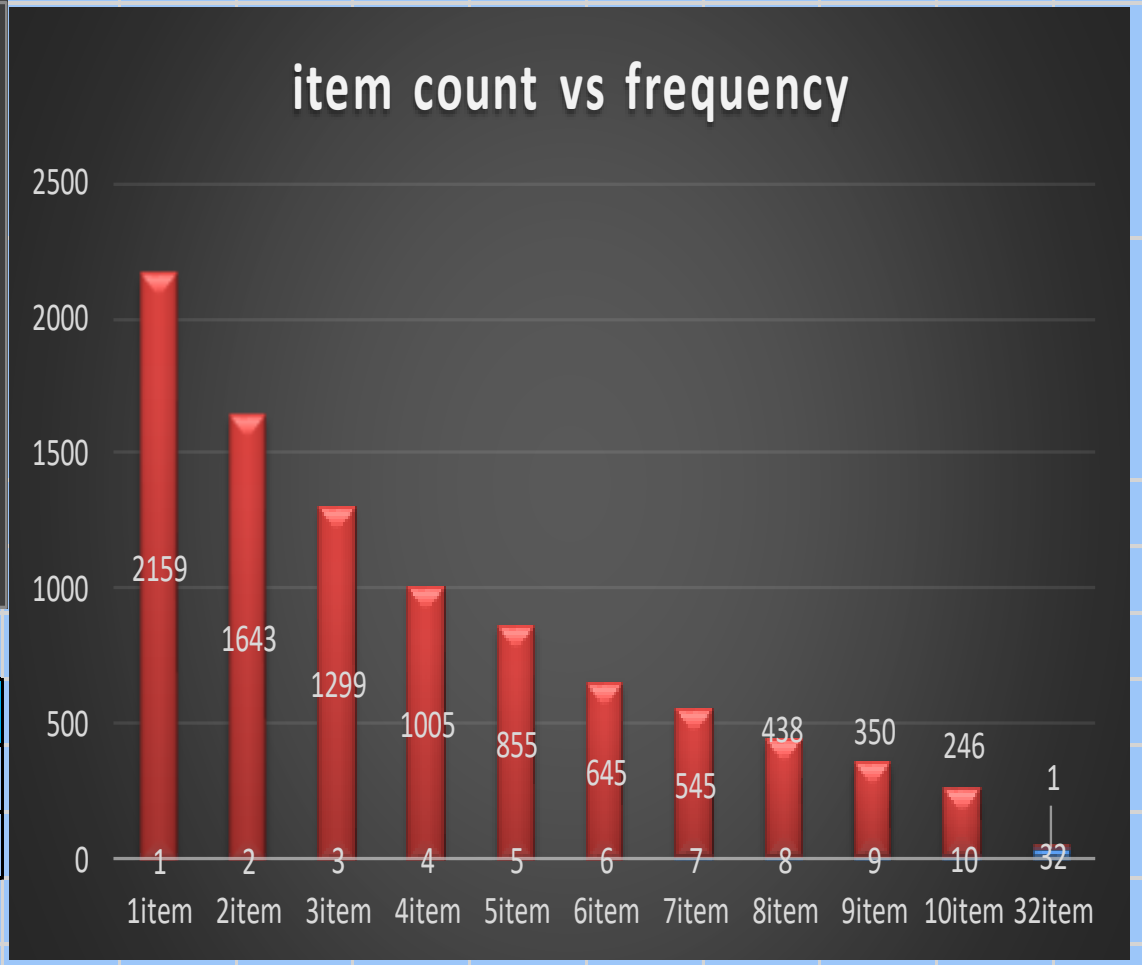
Groceries dataset provided with arules package in R :

most frequent items:

whole milk	other vegetables	rolls/buns	soda	yogurt	(Other)
2513	1903	1809	1715	1372	34055



1item	2item	3item	4item	5item	6item	7item	8item	9item	10item	32item
1	2	3	4	5	6	7	8	9	10	32
2159	1643	1299	1005	855	645	545	438	350	246	1

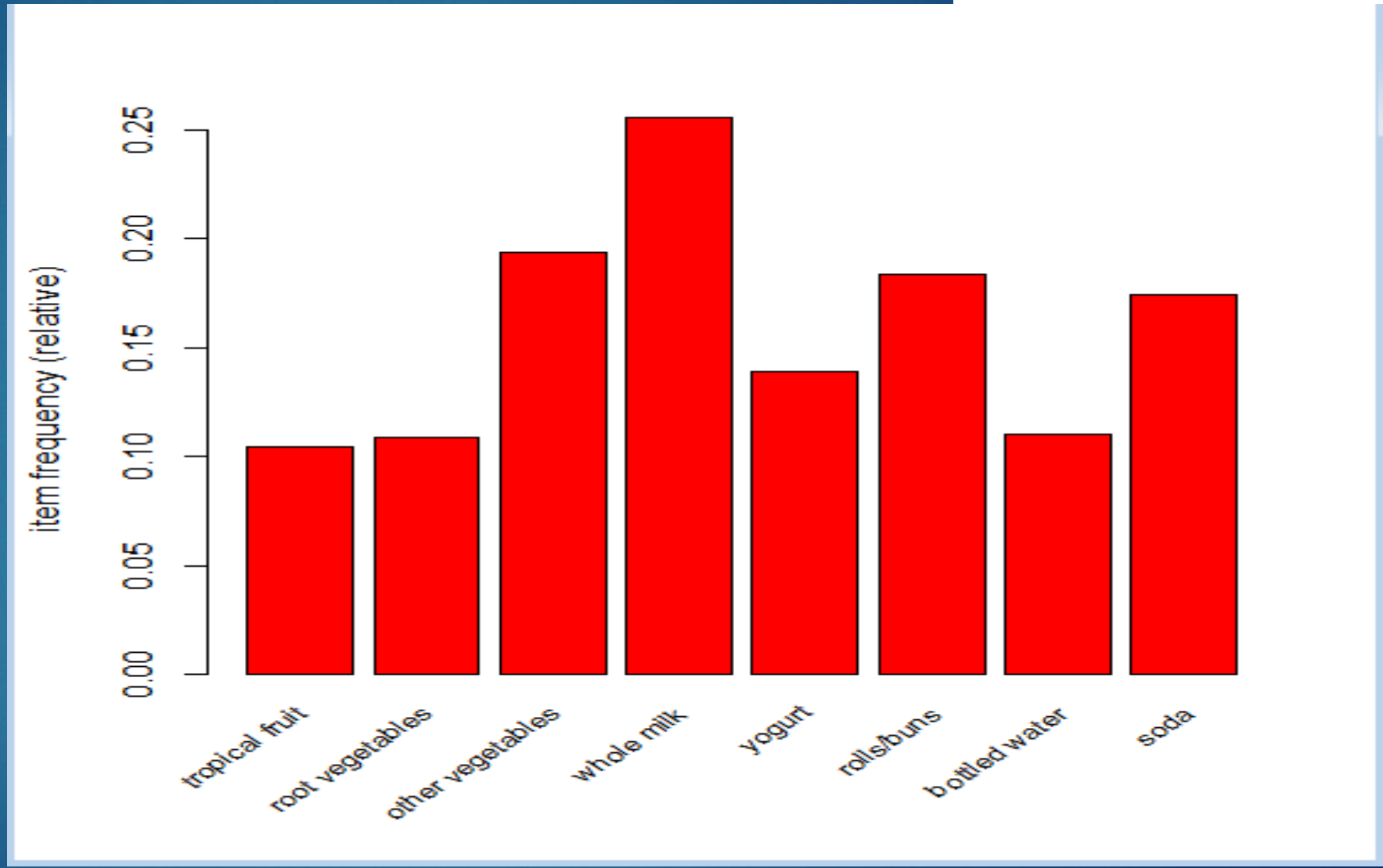


Groceries dataset provided with arules package in R :

REVA UNIVERSITY

Let us proceed to frequency plots. The more frequent the item will be in transaction the higher its bar. Moreover there are plots with different support levels. Support is the frequency of the pattern in the rule, therefore it being set to 0.1 means that the item must occur at least 10 times in 100 transactions.

```
itemFrequencyPlot(transactions,
support=0.1, cex.names=0.8,col=2)
```

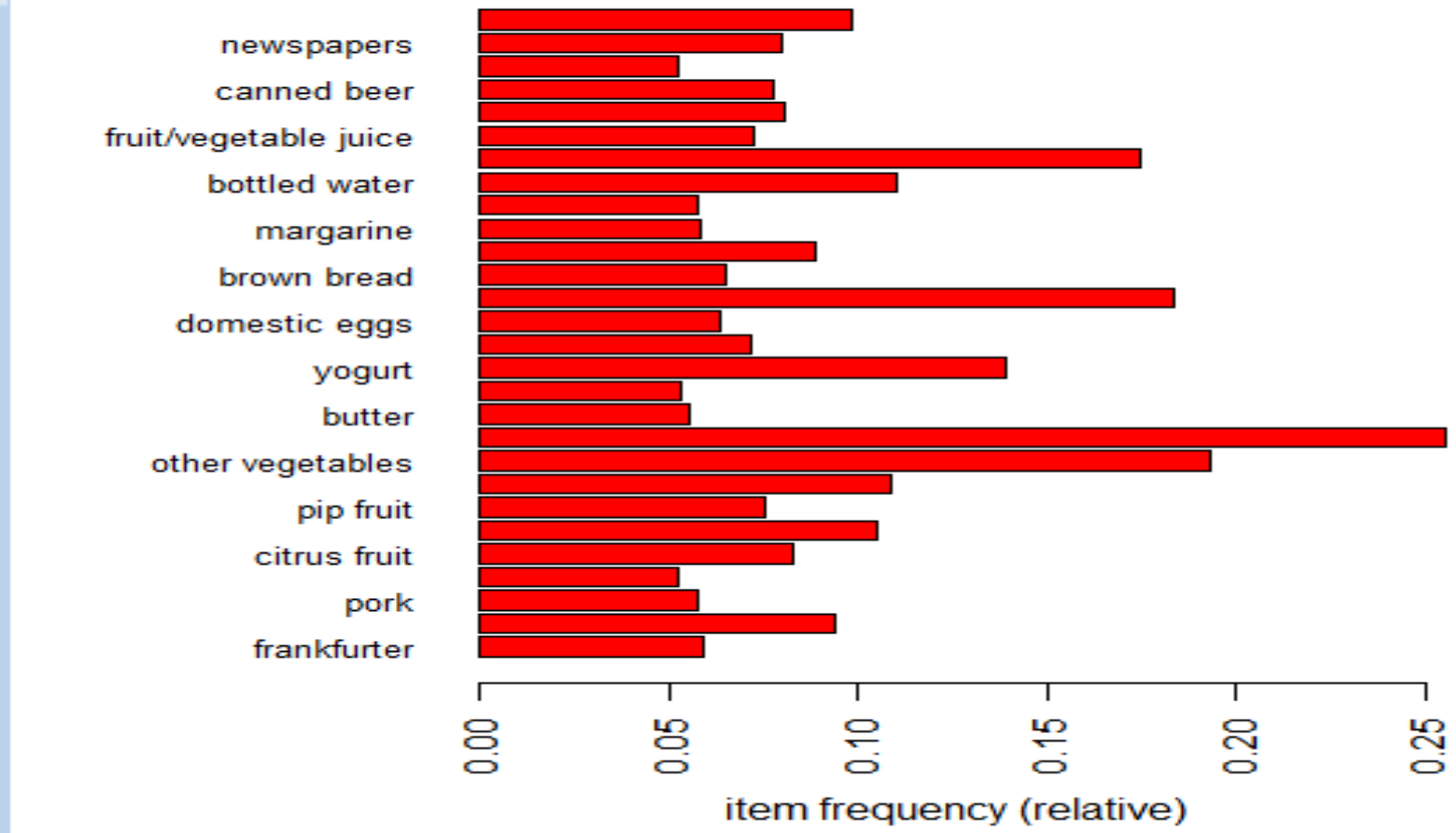


Groceries dataset provided with arules package in R :

REVA UNIVERSITY

Support is the frequency of the pattern in the rule, therefore it being set to 0.05 means that the item must occur at least 5 times in 100 transactions.

```
itemFrequencyPlot(transactions, support=0.05,cex.names=0.8,col=2,horiz=TRUE)
```

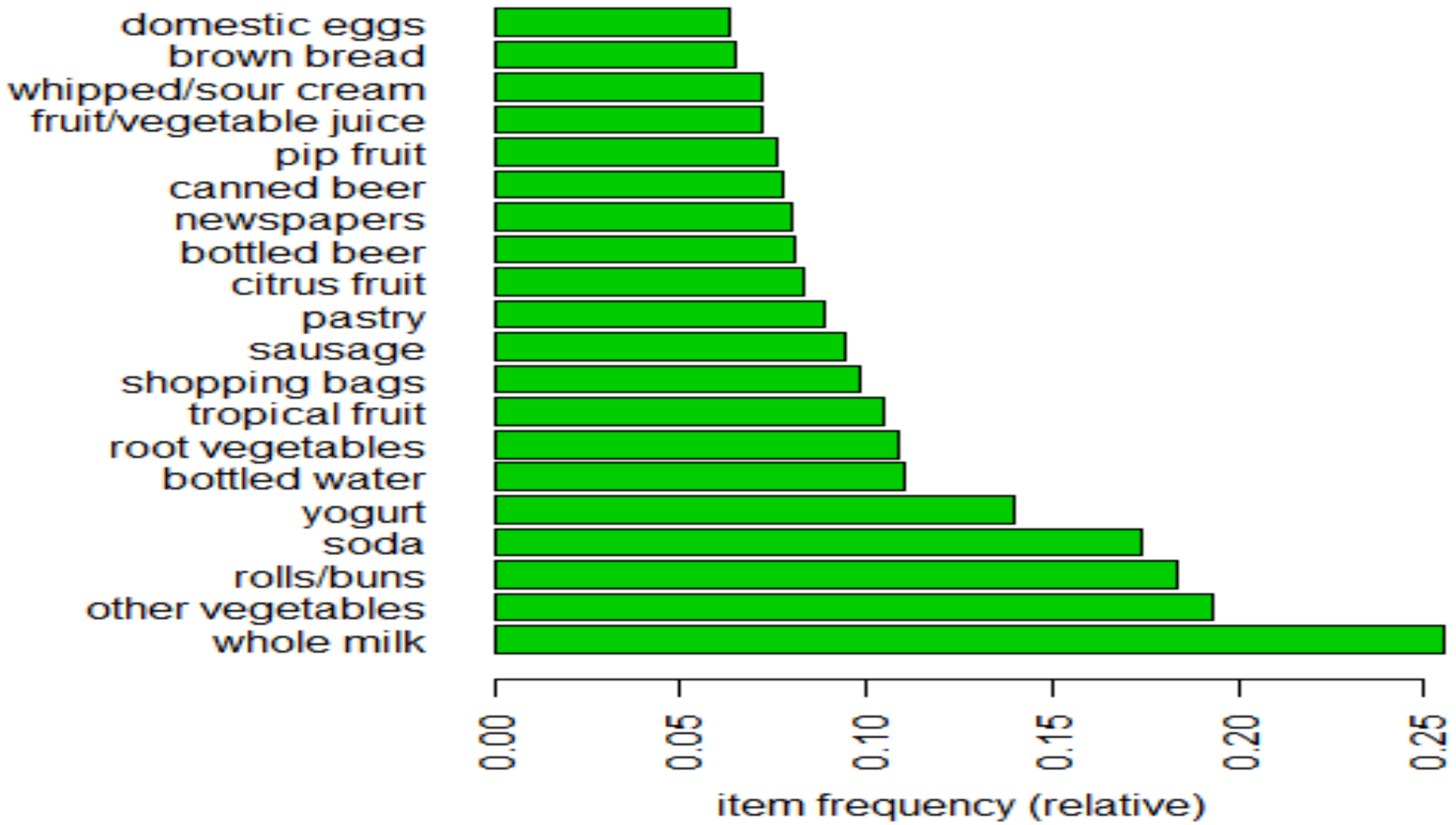


Groceries dataset provided with arules package in R :

REVA UNIVERSITY

Other way of selecting desired number of elements is to provide not support, but just the desired number. This is as presented on the third graph.

```
itemFrequencyPlot(transactions,
topN=20,col=3,horiz=TRUE)
```



Groceries dataset provided with arules package in R :

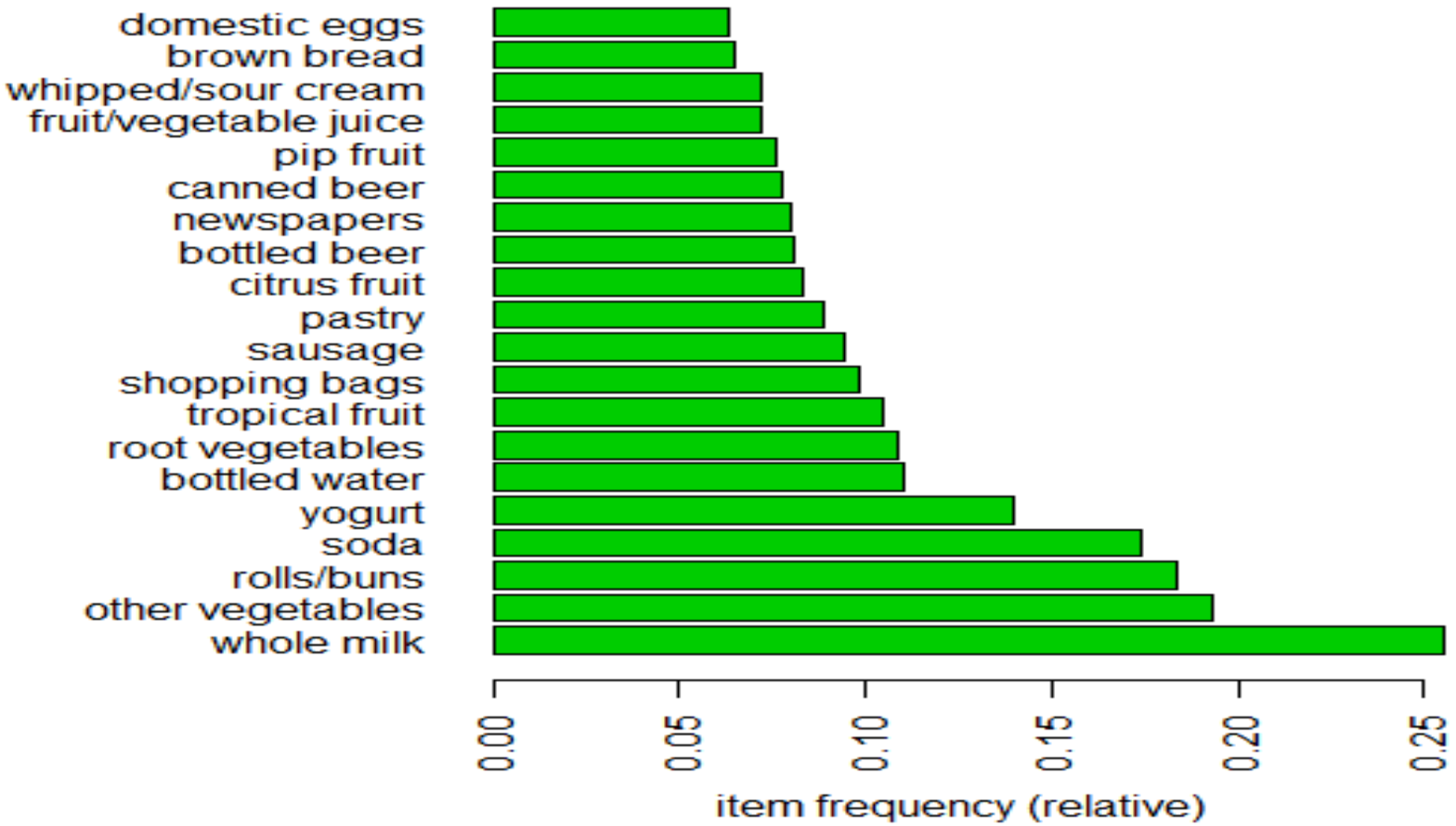
REVA UNIVERSITY

On an average, each item set or basket contains 4 to 5 items.

In other words, basket having less than 5 items is more frequent as compare to baskets having more than 15 items.

Buyers generally come to purchase fewer items from the shop.

Support being set to .01 means that plot only includes item set having more than 1 repetition in each 100 transactions. Support less than 0.05 won't generate significant rules and needed to be ignored for the study.



Business rules:

Support shows the frequency of the patterns in the rule; it is the percentage of transactions that contain both A and B,

Support = Probability (A and B) Support = (# of transactions involving A and B) / (total # of transactions).

Confidence is the strength of implication of a rule; it is the percentage of transactions that contain B if they contain A,

Confidence = Probability (A and B) = P(A) Confidence = (# of transactions involving A and B) / (total # of transactions that have A).

Expected confidence is the percentage of transactions that contain B to all transactions, i.e. Expected confidence = Probability (B)

Correlation analysis

REVA UNIVERSITY

The lift score meanings:

Lift = 1 meaning A and B are independent .There is no relation between A and B as far as buying pattern is concerned.

Lift > 1 meaning A and B are positively correlated .correlated.so if A Is bought then we can be positively confident that B will be bought by as many times as the lift value.

Lift < 1 meaning A and B are negatively correlated.so if A Is bought then we can be negatively confident that B will not be bought by as many times as the lift value.

eclat algorithm - to see most frequent item sets

Firstly let us try the éclat algorithm - to see most frequent item sets. Below we will see the list of the most common items together with their individual support.

freq.itemsets <eclat(transactions, parameter=list(supp=0.075, maxlen=15))

We can conclude that Most frequent item sets correspond to the most frequent items (as there are no more than 2 items item sets.)

```
Eclat

parameter specification:
 tidLists support minlen maxlen            target   ext
    FALSE   0.075      1     15 frequent itemsets FALSE

algorithmic control:
 sparse sort verbose
      7   -2    TRUE

Absolute minimum support count: 737

create itemset ...
set transactions ...[169 item(s), 9835 transaction(s)] done [0.01s].
sorting and recoding items ... [16 item(s)] done [0.00s].
creating sparse bit matrix ... [16 row(s), 9835 column(s)] done [0.00s].
writing  ... [16 set(s)] done [0.00s].
Creating S4 object  ... done [0.00s].
> inspect(freq.itemsets)
     items              support    count
[1]  {whole milk}       0.25551601 2513
[2]  {other vegetables} 0.19349263 1903
[3]  {rolls/buns}       0.18393493 1809
[4]  {yogurt}           0.13950178 1372
[5]  {soda}             0.17437722 1715
[6]  {root vegetables}  0.10899847 1072
[7]  {tropical fruit}   0.10493137 1032
[8]  {bottled water}    0.11052364 1087
[9]  {sausage}          0.09395018  924
[10] {shopping bags}    0.09852567  969
[11] {citrus fruit}     0.08276563  814
[12] {pastry}           0.08896797  875
[13] {pip fruit}        0.07564820  744
[14] {newspapers}       0.07981698  785
[15] {bottled beer}     0.08052872  792
[16] {canned beer}      0.07768175  764
```

. Rules created using apriori algorithm:

REVA UNIVERSITY

Let us create rules then. Rules are created using apriori algorithm and giving minimal support and confidence of a rule.

rules <- apriori(Groceries, parameter = list(support = 0.009, confidence = 0.25, minlen = 2))**summary(rules)**

Summary of rules will provide us with statistical information about support, confidence, lift and count of items.

We obtained a set of 224 rules, where mean support is equal to 16% and mean confidence is 37%. These are not bad values. It means that mean rule occurs in 16% transactions and its implication has 37% power. Inspect top.

```
Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen
       0.25    0.1    1 none FALSE            TRUE       5   0.009      2
 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: 88
set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[169 item(s), 9835 transaction(s)] done [0.02s].
sorting and recoding items ... [93 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 done [0.02s].
writing ... [224 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
> summary(rules)
set of 224 rules

rule length distribution (lhs + rhs):sizes
  2   3
111 113
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   2.000   3.000   2.504   3.000   3.000
summary of quality measures:
     support         confidence          lift            count
 Min.   :0.009049   Min.   :0.2513   Min.   :0.9932   Min.   : 89.0
 1st Qu.:0.010066   1st Qu.:0.2974   1st Qu.:1.5767   1st Qu.: 99.0
 Median :0.012303   Median :0.3603   Median :1.8592   Median :121.0
 Mean   :0.016111   Mean   :0.3730   Mean   :1.9402   Mean   :158.5
 3rd Qu.:0.018480   3rd Qu.:0.4349   3rd Qu.:2.2038   3rd Qu.:181.8
 Max.   :0.074835   Max.   :0.6389   Max.   :3.7969   Max.   :736.0

mining info:
      data ntransactions support confidence
 Groceries          9835   0.009       0.25
```

Inspect top 5 rules sorted by lift from high to low:

REVA UNIVERSITY

Above rules (sorted by lift - preference of buying B if A was bought) can be interpreted as such:

Anyone who buys citruses/tropical fruits is 3.29 and 3.14 times more likely respectively to buy root vegetables than any other client.

Anyone who buys berries is 3.7 times more likely to buy whipped/sour cream than any other client.

People like to buy berries and eat them with cream.

```
> inspect(head(sort(rules, by ="lift"),5))
    lhs                                   rhs                    support
[1] {berries}                          => {whipped/sour cream}   0.009049314
[2] {tropical fruit,other vegetables}  => {pip fruit}            0.009456024
[3] {pip fruit,other vegetables}       => {tropical fruit}       0.009456024
[4] {citrus fruit,other vegetables}    => {root vegetables}      0.010371124
[5] {tropical fruit,other vegetables}  => {root vegetables}      0.012302999
    confidence lift     count
[1] 0.2721713  3.796886  89
[2] 0.2634561  3.482649  93
[3] 0.3618677  3.448613  93
[4] 0.3591549  3.295045 102
[5] 0.3427762  3.144780 121
>
```

Inspect top 5 rules sorted by lift from high to low:

REVA UNIVERSITY

Let us see rules that have high support and high confidence.

There is new rule (very strong) that says that buying milk is associated with buying curd, yoghurt or butter.

```
> inspect(sort(sort(rules, by ="support"),by ="confidence")[1:5])
    lhs                                  rhs                  support
[1] {butter,yogurt}                   => {whole milk}         0.009354347
[2] {citrus fruit,root vegetables}    => {other vegetables}   0.010371124
[3] {tropical fruit,root vegetables}  => {other vegetables}   0.012302999
[4] {curd,yogurt}                     => {whole milk}         0.010066090
[5] {other vegetables,curd}           => {whole milk}         0.009862735
    confidence lift     count
[1] 0.6388889  2.500387  92
[2] 0.5862069  3.029608 102
[3] 0.5845411  3.020999 121
[4] 0.5823529  2.279125  99
[5] 0.5739645  2.246296  97
> milk.rules <- sort(subset(rules, subset = rhs %in% "whole milk"), by = "confi$
> summary(milk.rules)
set of 85 rules

rule length distribution (lhs + rhs):sizes
 2  3
46 39

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   2.000   2.000   2.459   3.000   3.000

summary of quality measures:
    support           confidence          lift            count
 Min.   :0.009049   Min.   :0.2538   Min.   :0.9932   Min.   : 89.0
 1st Qu.:0.010269   1st Qu.:0.3845   1st Qu.:1.5047   1st Qu.:101.0
 Median :0.013523   Median :0.4344   Median :1.7002   Median :133.0
 Mean   :0.018057   Mean   :0.4374   Mean   :1.7116   Mean   :177.6
 3rd Qu.:0.021251   3rd Qu.:0.4976   3rd Qu.:1.9474   3rd Qu.:209.0
 Max.   :0.074835   Max.   :0.6389   Max.   :2.5004   Max.   :736.0

mining info:
      data ntransactions support confidence
 Groceries          9835   0.009       0.25
```

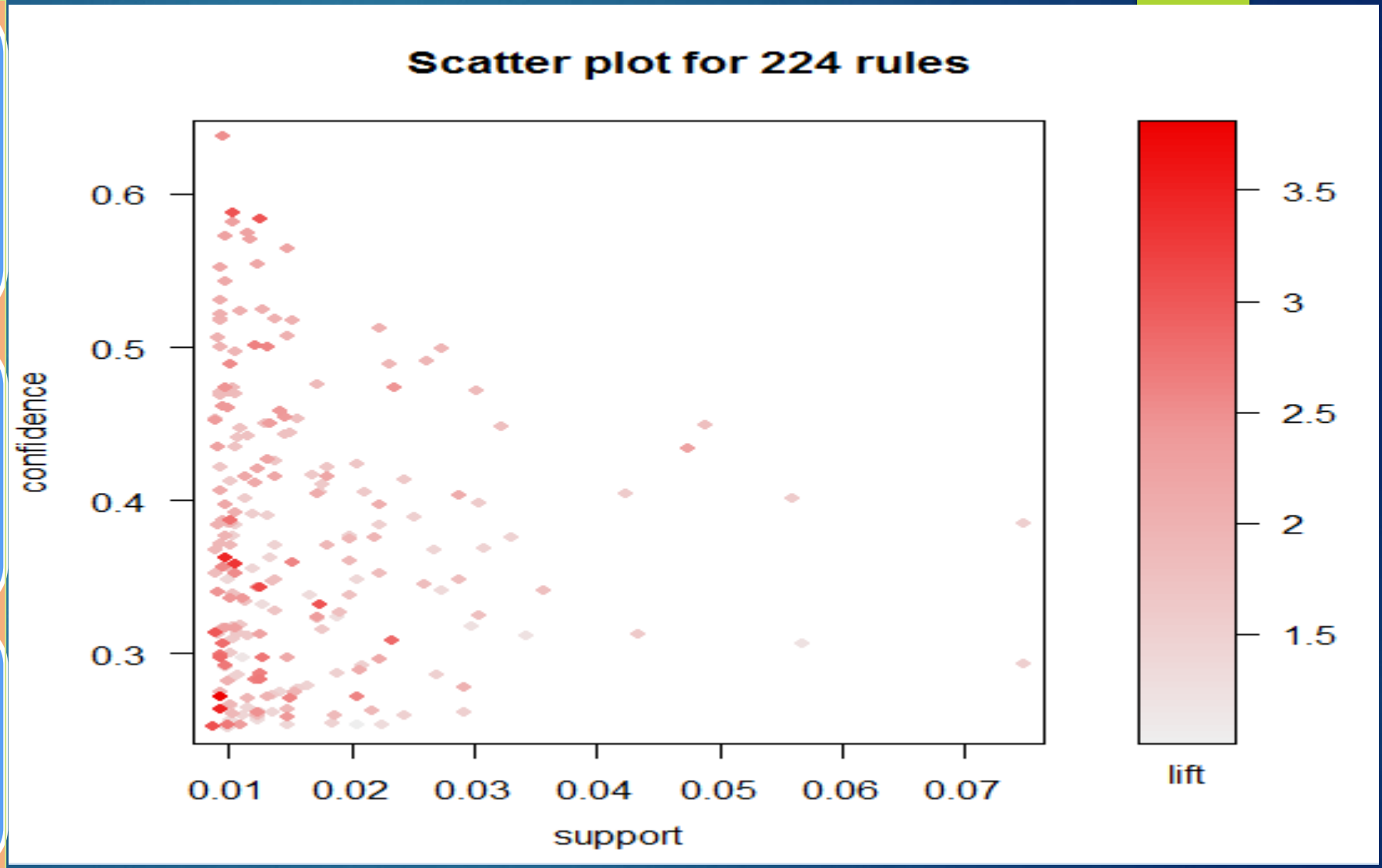
Scatter plot for 224 rules:

REVA UNIVERSITY

we can plot the rules in support and confidence axes and colour them with lift values.

Most of the rules have small values of support i.e. around 0.01, but confidence varies from 0.6 up to 0.3.

The reddier the point meaning higher is the lift and so then more likely is the rule to happen.



Induction:

REVA UNIVERSITY

Below analyses depend on choosing one product and checking which products it implies or by which products it is implied.

Right hand side we have milk and we can see that if a customer buys butter, yoghurt or curd, yoghurt or other vegetables, curd or other vegetables, butter or tropical fruits and root vegetables then we are more confident that the customer will buy whole milk as well.

The figure shown adjacent shows the top five lift given the condition that if a customer buy items on the l.h.s we are confident that customer will then buy whole milk.

```
> milk.rules <- sort(subset(rules, subset = rhs %in% "whole milk"), by = "confidence")
> summary(milk.rules)
set of 85 rules

rule length distribution (lhs + rhs):sizes
 2  3
46 39

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   2.000   2.000   2.459   3.000   3.000

summary of quality measures:
    support           confidence          lift            count
 Min.   :0.009049   Min.   :0.2538   Min.   :0.9932   Min.   : 89.0
 1st Qu.:0.010269   1st Qu.:0.3845   1st Qu.:1.5047   1st Qu.:101.0
 Median :0.013523   Median :0.4344   Median :1.7002   Median :133.0
 Mean   :0.018057   Mean   :0.4374   Mean   :1.7116   Mean   :177.6
 3rd Qu.:0.021251   3rd Qu.:0.4976   3rd Qu.:1.9474   3rd Qu.:209.0
 Max.   :0.074835   Max.   :0.6389   Max.   :2.5004   Max.   :736.0

mining info:
      data ntransactions support confidence
 Groceries          9835   0.009       0.25
> inspect(head(sort(milk.rules, by ="lift"),5))
    lhs                                rhs           support     confidence lift
[1] {butter,yogurt}                 => {whole milk} 0.009354347 0.6388889  2.500387
[2] {curd,yogurt}                   => {whole milk} 0.010066090 0.5823529  2.279125
[3] {other vegetables,curd}         => {whole milk} 0.009862735 0.5739645  2.246296
[4] {other vegetables,butter}       => {whole milk} 0.011489578 0.5736041  2.244885
[5] {tropical fruit,root vegetables} => {whole milk} 0.011997966 0.5700483  2.230969
    count
[1]  92
[2]  99
[3]  97
[4] 113
[5] 118
```

Milk.rules:

We can find that most of the milk.rules which is subset of the 224 rules which comes to 85 rules are significant.

also maximal condition getting satisfied for milk.rules.

Most of the subset of the rules so formed are non redundant as well.

```
> is.significant(milk.rules, transactions)
 [1]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[14]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[27]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[40]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[53]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[66]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE
[79] FALSE  TRUE  TRUE  TRUE FALSE FALSE FALSE
> is.maximal(milk.rules)
 [1]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[14]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE
[27]  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE
[40]  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE
[53]  TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE  TRUE FALSE FALSE  TRUE FALSE  TRUE
[66]  TRUE  TRUE FALSE  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[79]  TRUE FALSE FALSE FALSE  TRUE  TRUE  TRUE
> is.redundant(milk.rules)
 [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[14] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[27] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[40] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[53] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
[66]  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[79] FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
>
```

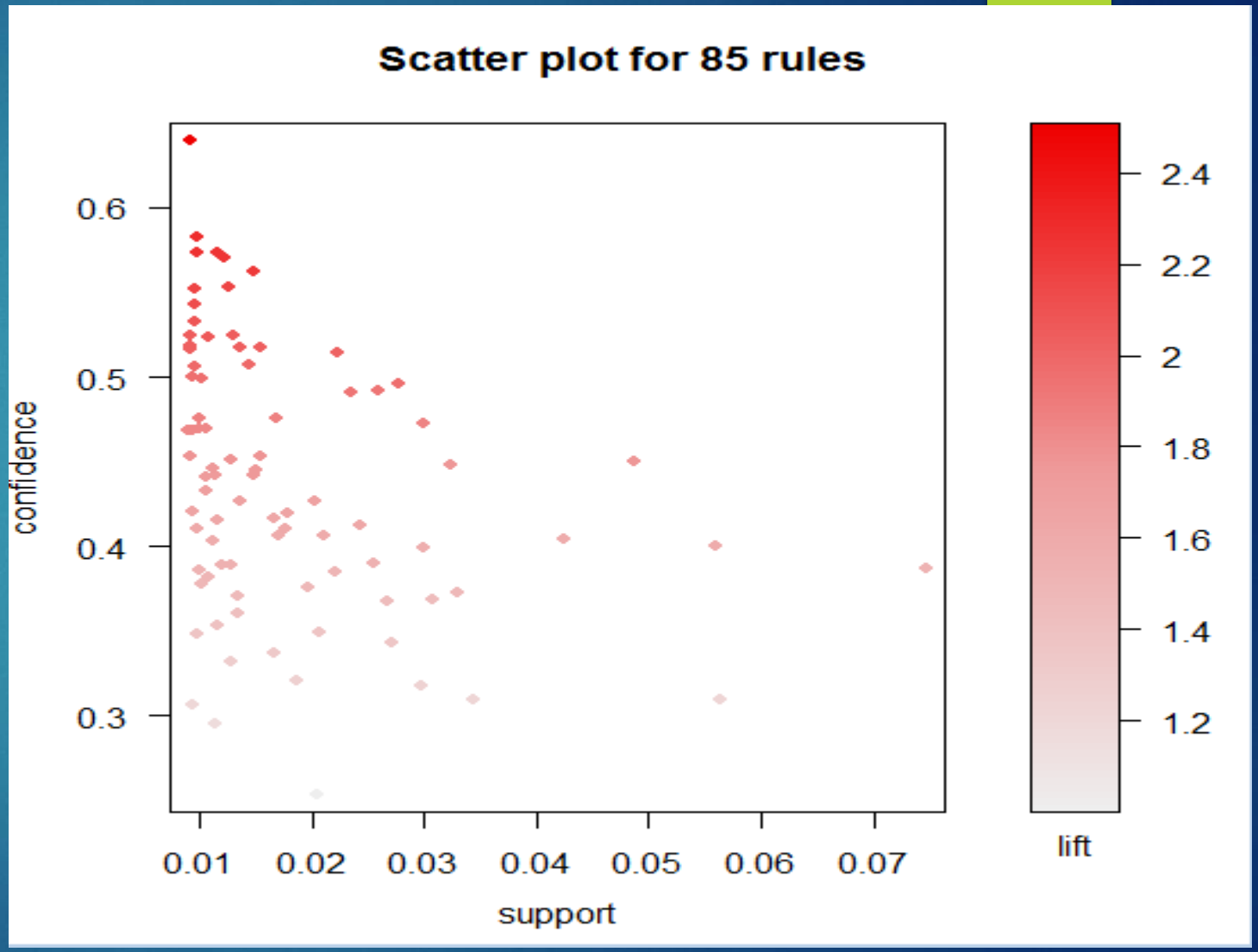
Scatter plot for Milk.rules:

REVA UNIVERSITY

we can plot the rules in support and confidence axes and colour them with lift values.

Most of the rules have small values of support i.e. around 0.01, but confidence varies from 0.6 up to 0.4 for most of the rules. Here we have framed subset of the rules under the condition that we are confident that customer will then buy whole milk if say customer buys certain predefined item set.

The reddier the point meaning higher is the lift and so then more likely is the rule to happen.



Visualization for milk. Rules using method="graph“:

REVA UNIVERSITY

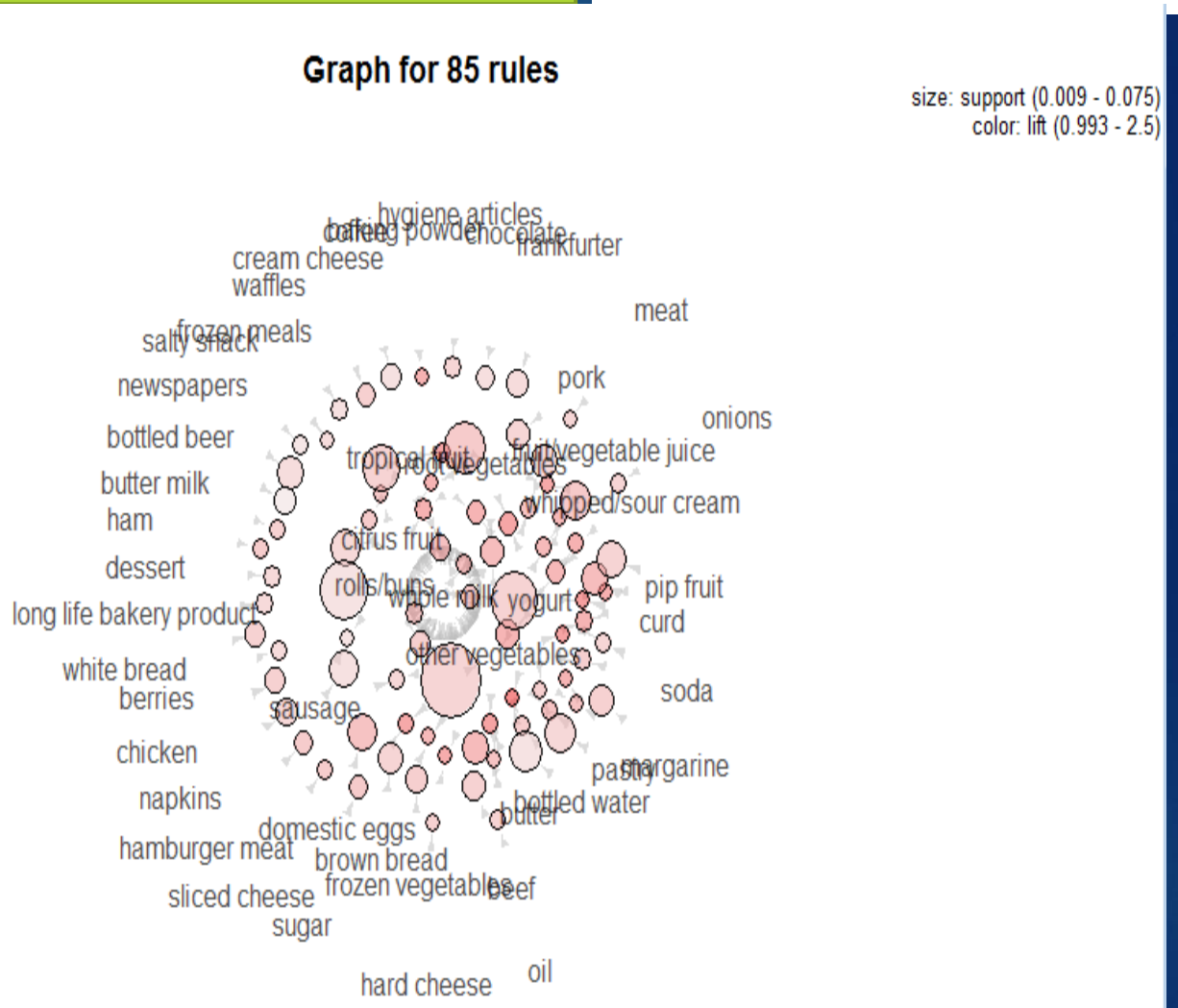
The reddier the circle the more probable is the client to buy two of those items than any other items

The bigger the circle the more probable is the client to buy two of those items.

Moreover the arrow points to the direction of a possible basket rule.

More complicated conclusions can be drawn from the milk.rules plot.

yoghurt ,other vegetables, tropical fruits, root vegetables are the mostly supported additional product for milk.



coke.rules:

Below analyses depend on choosing one product and checking which products it implies or by which products it is implied.

Right hand side we have soda and we can see that if a customer buys sausage, rolls/bun, chocolate, or dessert, or bottled water or sausage only or fruit/vegetable juice then we are more confident that the customer will buy soda as well.

The figure shown adjacent shows all of the coke.rules given the condition that if a customer buy items on the l.h.s we are confident that customer will then buy soda.

```
> coke.rules <- sort(subset(rules, subset = rhs %in% "soda"), by = "confidence")
> summary(coke.rules)
set of 6 rules

rule length distribution (lhs + rhs):sizes
2 3
5 1

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   2.000   2.000   2.167   2.000   3.000

summary of quality measures:
     support           confidence          lift            count
 Min.   :0.009659   Min.   :0.2546   Min.   :1.460   Min.   : 95.0
 1st Qu.:0.010778   1st Qu.:0.2595   1st Qu.:1.488   1st Qu.:106.0
 Median :0.015963   Median :0.2640   Median :1.514   Median :157.0
 Mean   :0.017455   Mean   :0.2716   Mean   :1.557   Mean   :171.7
 3rd Qu.:0.022827   3rd Qu.:0.2708   3rd Qu.:1.553   3rd Qu.:224.5
 Max.   :0.028978   Max.   :0.3156   Max.   :1.810   Max.   :285.0

mining info:
      data ntransactions support confidence
 Groceries          9835   0.009       0.25
> inspect(coke.rules)
    lhs                      rhs     support     confidence lift     count
[1] {sausage,rolls/buns}  => {soda} 0.009659380 0.3156146  1.809953  95
[2] {chocolate}           => {soda} 0.013523132 0.2725410  1.562939 133
[3] {dessert}             => {soda} 0.009862735 0.2657534  1.524015  97
[4] {bottled water}       => {soda} 0.028978139 0.2621895  1.503577 285
[5] {sausage}             => {soda} 0.024300966 0.2586580  1.483324 239
[6] {fruit/vegetable juice} => {soda} 0.018403660 0.2545710  1.459887 181
> is.significant(coke.rules, transactions)
[1] TRUE TRUE TRUE TRUE TRUE TRUE
>
```

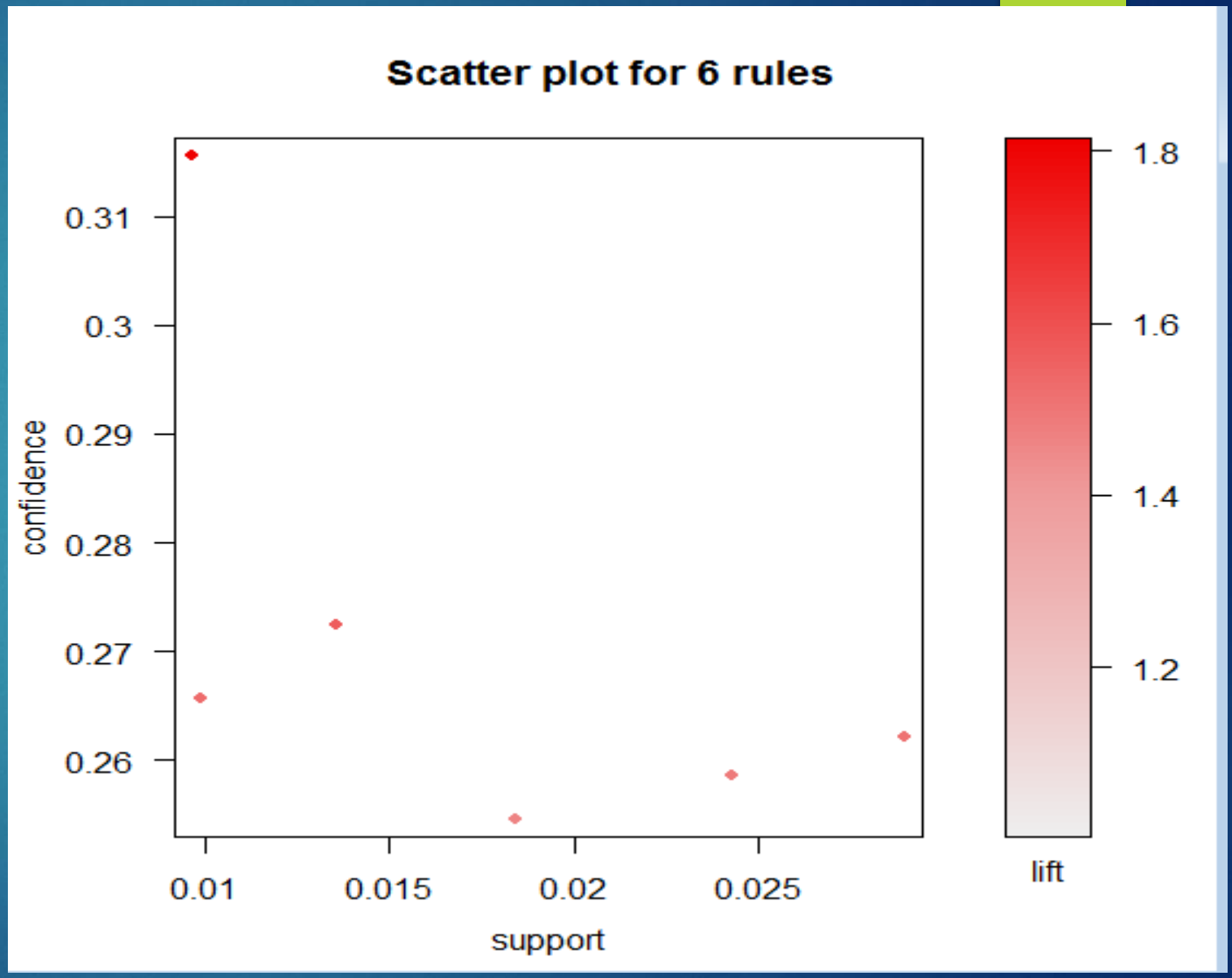
Scatter plot for coke.rules:

REVA UNIVERSITY

we can plot the rules in support and confidence axes and colour them with lift values.

Two of the six rules have small values of support i.e. around 0.01, but confidence varies from 0.31 up to 0.26. Here we have framed subset of the rules under the condition that we are confident that customer will then buy soda if say customer buys certain predefined item set.

The reddier the point meaning higher is the lift and so then more likely is the rule to happen.



Visualization for coke.rules using method="graph“:

REVA UNIVERSITY

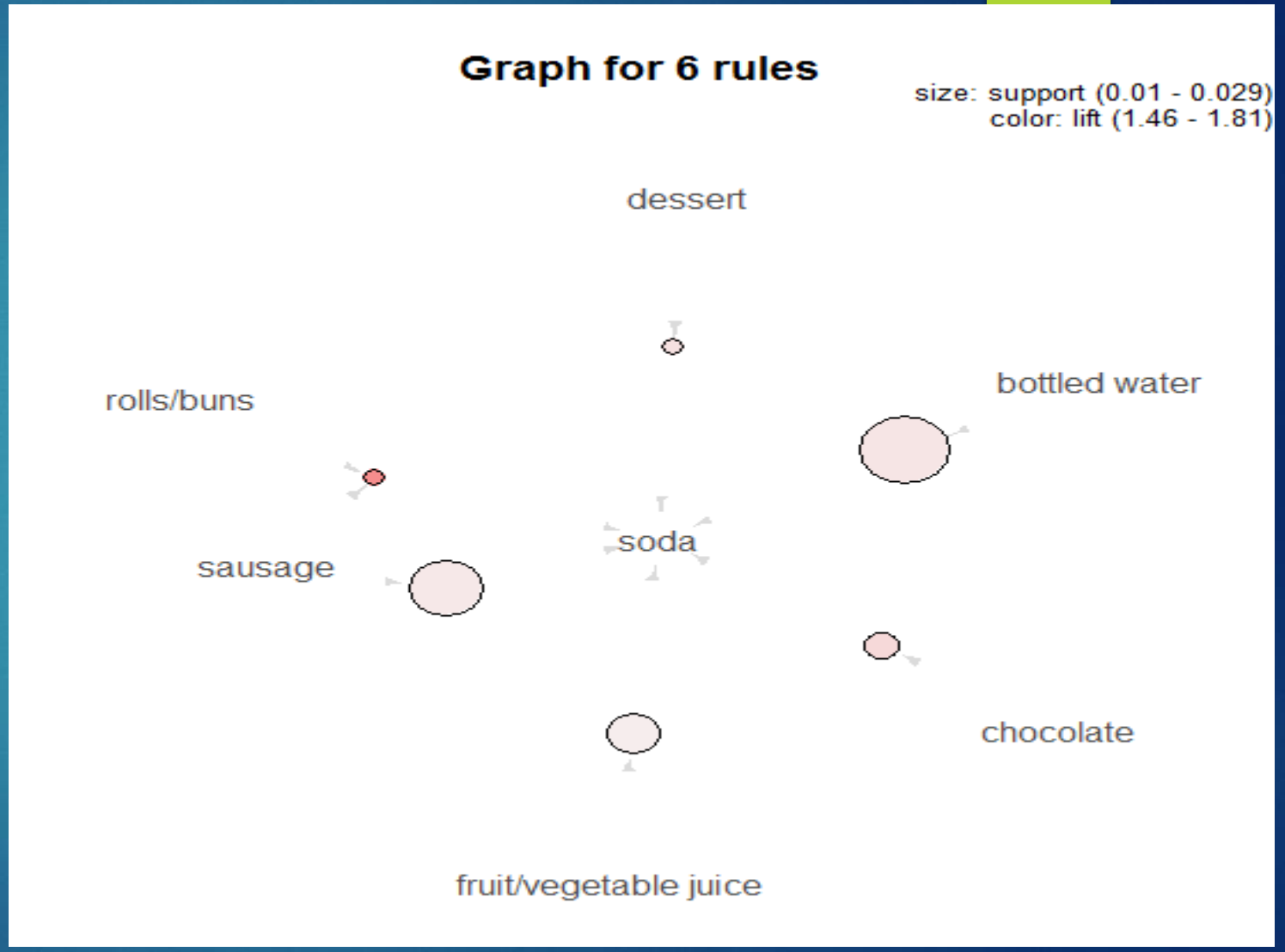
The reddier the circle the more probable is the client to buy two of those items than any other items

The bigger the circle the more probable is the client to buy two of those items.

Moreover the arrow points to the direction of a possible basket rule.

Therefore in case of Coke, we can notice bottled water and soda as the rule with highest support.

Sausage with soda is the second most import rule that can be formed.



Summary for milk.rules and coke.rules:

REVA UNIVERSITY

Analysis was aimed to see what makes people buy milk (what products to be exact).

To do se we should choose subset of rules that has whole milk (or soda) in right hand side of a rule.

It turns out that most popular baskets are curd, yoghurt or fruits and vegetables.

Seems like the most popular one-week ahead groceries we do.

On the other hand it seems that soda is mostly bought with either sweets (chocolate) or with beverages/meat.

Most of the rules are significant (Fisher's exact test) apart from some of the least confident rules of milk buying.

We can also see on the scatter plot of rules for milk that the higher the confidence the higher lift, which was not observed before. It also occurs on Coke rules plot, but is not that visible.

meat. Rules:

Below analyses depend on choosing one product and checking which products it implies or by which products it is implied.

left hand side we have meat and we can see that if a customer buys beef or chicken we are 3.04 times 2.32 times confident respectively that he will buy root vegetables. We can draw such similar inferences such as that on the l.h.s. customer always buys meat as an item.

The figure shown adjacent shows the top five lift given the condition that if a customer buy meat on the l.h.s we are confident that customer will then buy items as shown on R.H.S.

```
meat.rules <- sort(subset(rules, subset = lhs %in% "beef"|lhs %in% "sausage" |lhs %in% "chicken"), by = "confidence")
> summary(meat.rules)
set of 19 rules

rule length distribution (lhs + rhs):sizes
 2  3
11  8

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   2.000   2.000   2.421   3.000   3.000

summary of quality measures:
    support           confidence          lift           count
 Min.   :0.009253   Min.   :0.2536   Min.   :1.196   Min.   : 91.0
 1st Qu.:0.009659   1st Qu.:0.3093   1st Qu.:1.483   1st Qu.: 95.0
 Median :0.013625   Median :0.3314   Median :1.758   Median :134.0
 Mean   :0.016156   Mean   :0.3471   Mean   :1.802   Mean   :158.9
 3rd Qu.:0.020488   3rd Qu.:0.4013   3rd Qu.:2.049   3rd Qu.:201.5
 Max.   :0.030605   Max.   :0.4691   Max.   :3.040   Max.   :301.0

mining info:
      data ntransactions support confidence
 Groceries          9835   0.009       0.25
> is.significant(meat.rules, transactions)
 [1]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
[13]  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE
> inspect(head(sort(meat.rules, by ="lift"),5))
    lhs                    rhs                support     confidence lift
[1] {beef}              => {root vegetables}  0.017386884 0.3313953  3.040367
[2] {chicken}           => {root vegetables}  0.010879512 0.2535545  2.326221
[3] {beef,whole milk} => {other vegetables} 0.009252669 0.4354067  2.250250
[4] {sausage,soda}      => {rolls/buns}       0.009659380 0.3974895  2.161034
[5] {chicken}           => {other vegetables} 0.017895272 0.4170616  2.155439
    count
[1] 171
[2] 107
[3]  91
```

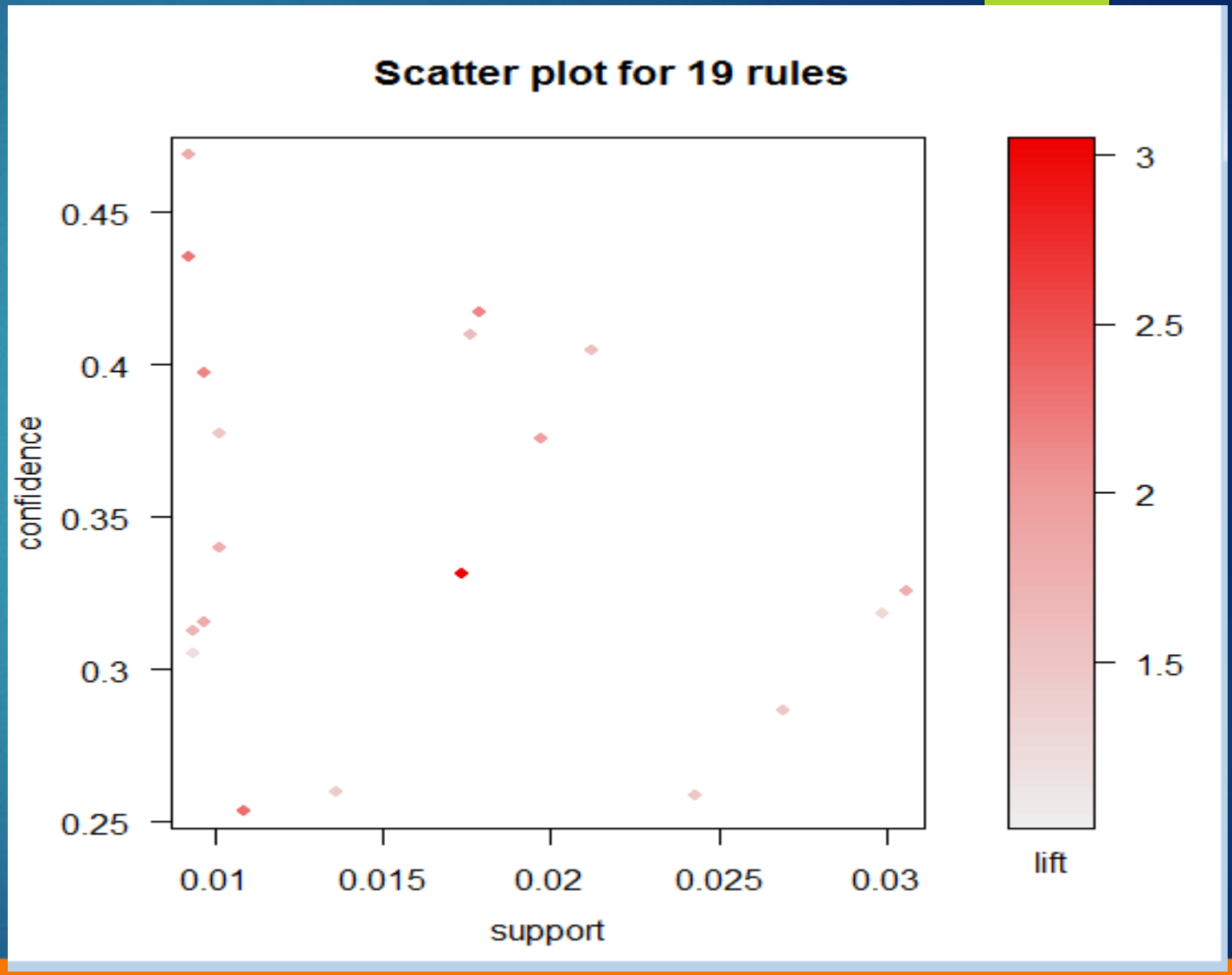
Scatter plot for meat. Rules:

REVA UNIVERSITY

we can plot the rules in support and confidence axes and colour them with lift values.

Many of the rules have small values of support i.e. around 0.01, but confidence varies from 0.45 up to 0.25 for many of the rules. Here we have framed subset of the rules under the condition that if customer buys meat then we are confident that customer will buy certain item set as listed in summary(meat. Rules).

The reddier the point meaning higher is the lift and so then more likely is the rule to happen.



Visualization for meat. Rules using method="graph“:

REVA UNIVERSITY

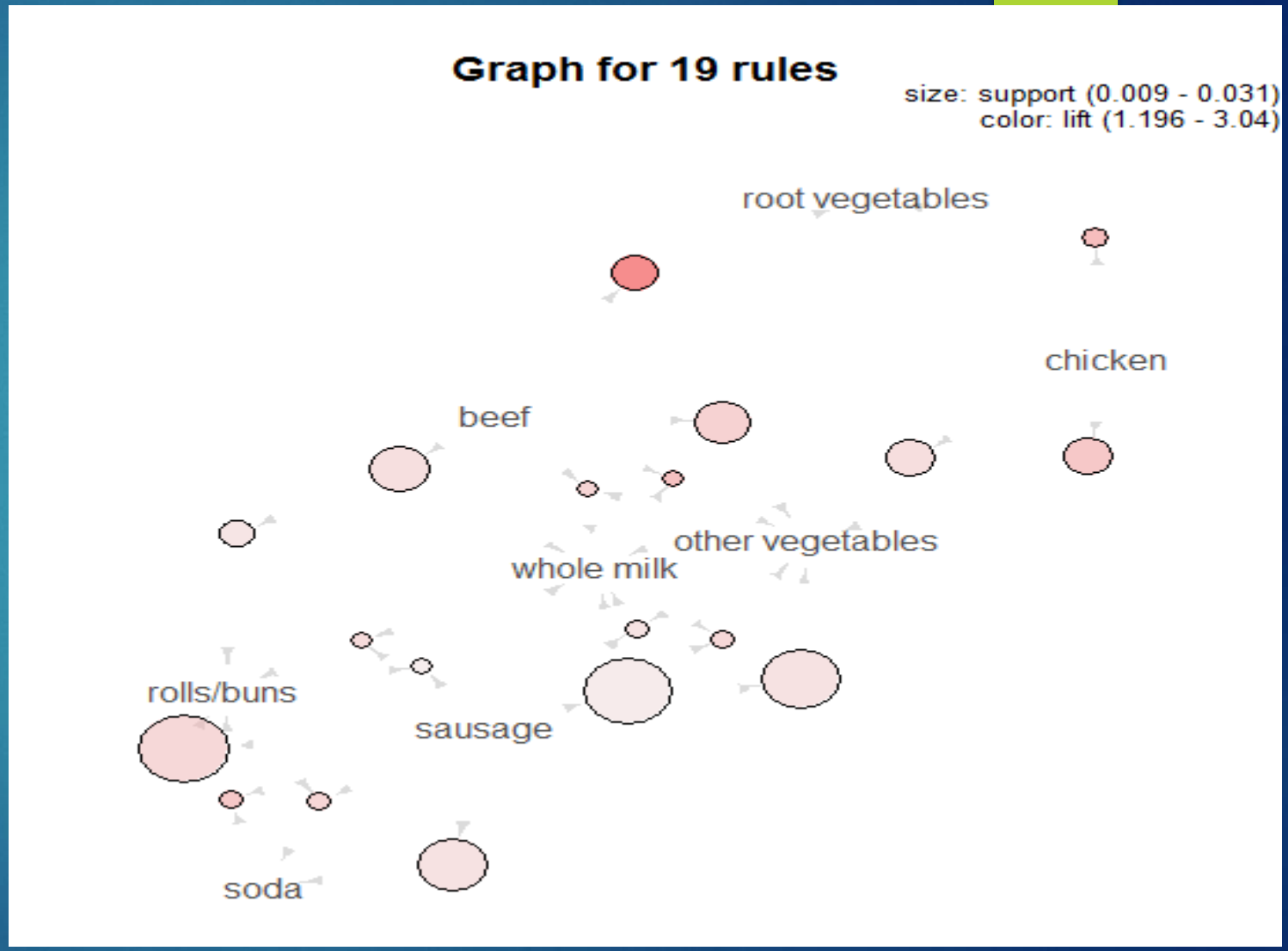
The reddier the circle the more probable is the client to buy two of those items than any other items.

The bigger the circle the more probable is the client to buy two of those items.

Moreover the arrow points to the direction of a possible basket rule.

Sausage, rolls/buns, soda are the mostly supported additional product for beef and chicken.

Root vegetable is the one with the highest lift values.however,the support for it is comparatively smaller. Other vegetables also seems to be there in contention though it is not as strong as root vegetables.



Summary for meat. Rules :

REVA UNIVERSITY

In case of meat, we search whether meats like: beef, chicken (poultry) or sausage show up in the left hand sides of rules.

Let's see what people buy after they have put meat (sausage or beef) to the basket.

It turns out that the most popular option associated with meat is milk!

It is a little bit confusing, because only in lift column we see how popular option is.

The real winner here are root vegetables that are 3 times more likely to be put into the basket than other products.

Rest of the products are just regular grocery stuff.

yogurt. Rules:

Below analyses depend on choosing one product and checking which products it implies or by which products it is implied.

left hand side we have yoghurt and we can see that if a customer buys yoghurt in combination with some other item in L.H.S. We are confident that he will buy item set as depicted in summary(yog.rules)

The figure shown adjacent shows the top five lift given the condition that if a customer buy yoghurt on the l.h.s we are confident that customer will then buy items as shown on R.H.S.

Most of the times someone buys yogurt he will also put milk or vegetables into his basket - with greater correlation to 'other vegetables'. There is not much variation, nothing changes with the lowering confidence.

```
> yog.rules <- sort(subset(rules, subset = lhs %in% "yogurt"), by = "confidence")
> summary(yog.rules)
set of 26 rules
rule length distribution (lhs + rhs):sizes
 2  3
 2 24
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   3.000   3.000   2.923   3.000   3.000

summary of quality measures:
    support           confidence          lift            count
 Min.   :0.009151   Min.   :0.2595   Min.   :1.439   Min.   : 90.0
 1st Qu.:0.010193   1st Qu.:0.3170   1st Qu.:1.739   1st Qu.:100.2
 Median :0.012303   Median :0.4365   Median :2.039   Median :121.0
 Mean   :0.015651   Mean   :0.4281   Mean   :2.058   Mean   :153.9
 3rd Qu.:0.015150   3rd Qu.:0.5162   3rd Qu.:2.356   3rd Qu.:149.0
 Max.   :0.056024   Max.   :0.6389   Max.   :2.729   Max.   :551.0

mining info:
      data ntransactions support confidence
 Groceries          9835   0.009       0.25
> is.significant(yog.rules, transactions)
 [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
[16] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
> inspect(head(sort(yog.rules, by ="lift"),5))
    lhs                          rhs                  support    confidence
[1] {other vegetables,yogurt}    => {root vegetables}  0.01291307 0.2974239
[2] {other vegetables,yogurt}    => {tropical fruit}   0.01230300 0.2833724
[3] {root vegetables,yogurt}     => {other vegetables} 0.01291307 0.5000000
[4] {whole milk,yogurt}          => {tropical fruit}   0.01514997 0.2704174
[5] {yogurt,whipped/sour cream} => {other vegetables} 0.01016777 0.4901961
    lift     count
[1] 2.728698 127
[2] 2.700550 121
[3] 2.584078 127
[4] 2.577089 149
[5] 2.533410 100
```

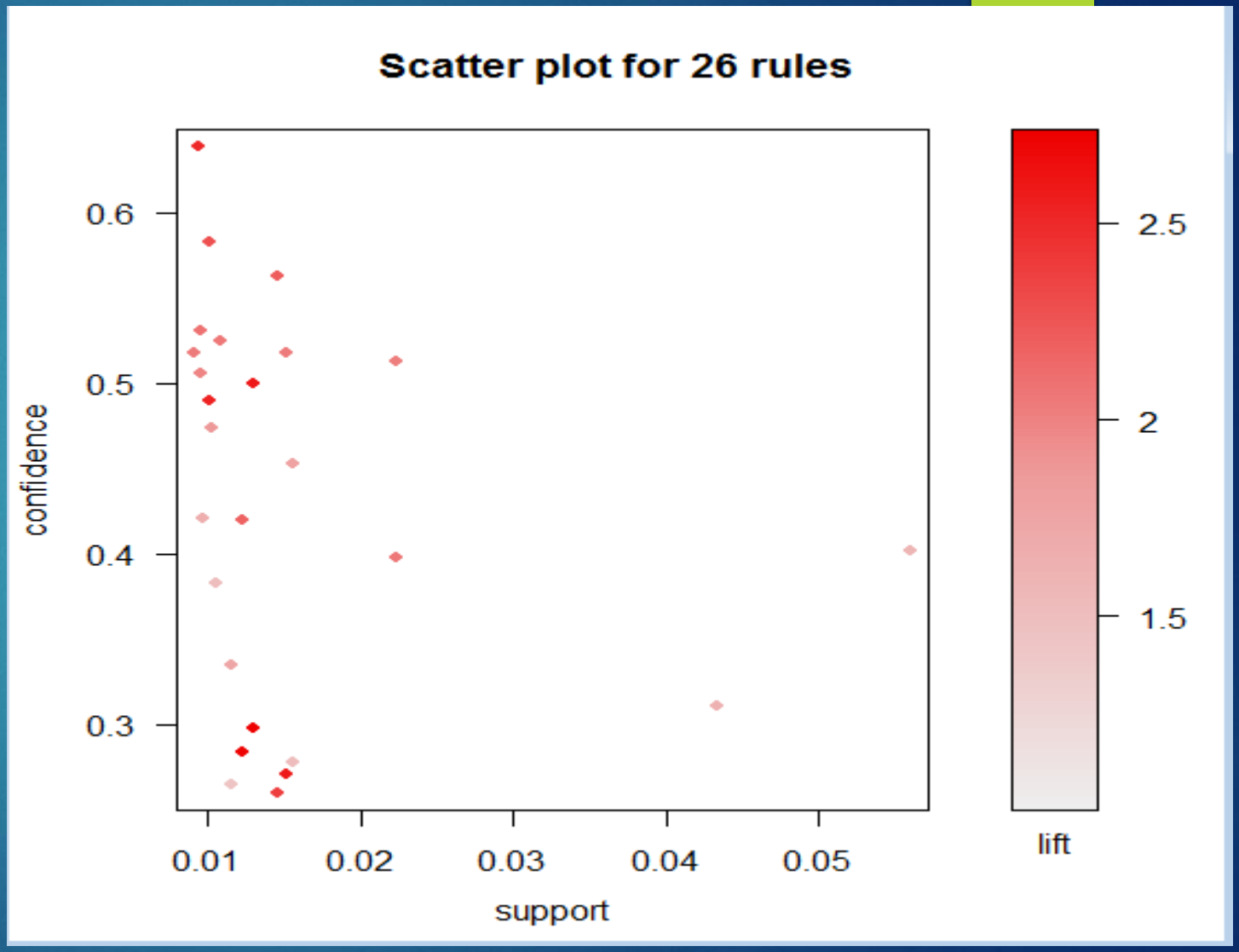
Scatter plot for yogurt. Rules:

REVA UNIVERSITY

we can plot the rules in support and confidence axes and colour them with lift values.

Many of the rules have small values of support i.e. around 0.01, but confidence varies from 0.6 up to 0.3 for many of the rules. Here we have framed subset of the rules under the condition that if customer buys yoghurt then we are confident that customer will buy certain item set as listed in summary(yog. Rules).

The reddier the point meaning higher is the lift and so then more likely is the rule to happen.

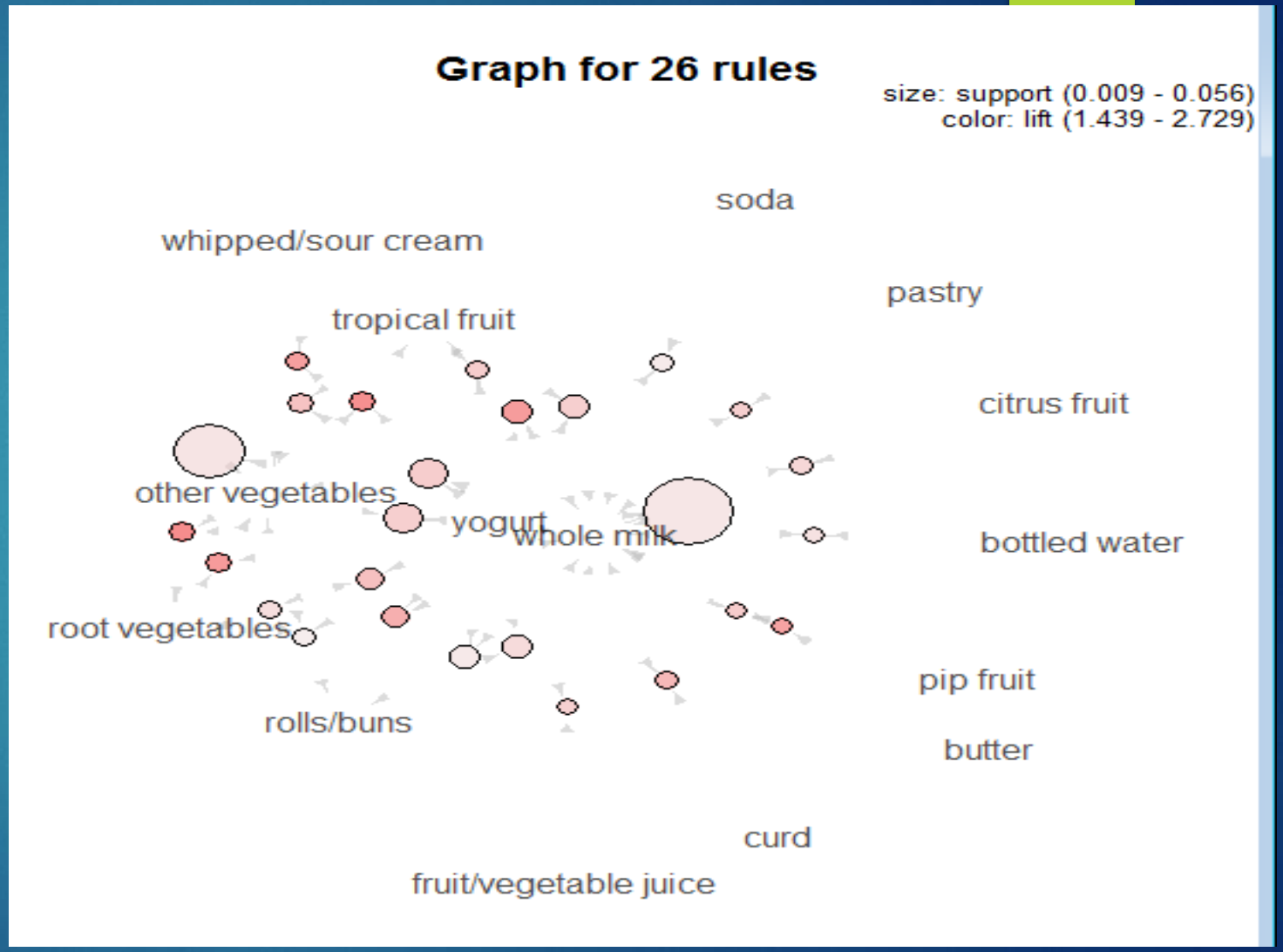


Visualization for yoghurt. Rules using method="graph“:

REVA UNIVERSITY

Whole milk ,other vegetables, are the mostly supported additional product for yoghurt.

Tropical fruits and root vegetables also seem to be in contention though their support size is needed to improve.



Jaccard Index:

REVA UNIVERSITY

Jaccard Index: It is the representation of how much likely are two items to be bought together.

Because I have picked such high minimal frequency we have not much items, but moreover Jaccard Index seems to have high values telling us that most of those products do not overlap.

Such an array as presented above tells that the higher the values of Jaccard Index the more likely are two products to be in the same transaction.

```
> trans.sel<-transactions[,itemFrequency(transactions)>0.1] # selected transactions
> dissimilarity(trans.sel, which="items")
                  tropical fruit root vegetables other vegetables whole milk
root vegetables        0.8908803
other vegetables       0.8632843       0.8142686
whole milk             0.8670502       0.8450387        0.8000000
yogurt                 0.8638941       0.8840183        0.8500702  0.8347331
rolls/buns             0.9068873       0.9095382        0.8727604  0.8520584
bottled water          0.9060403       0.9231920        0.9111435  0.8963826
soda                   0.9193548       0.9297235        0.9023058  0.8972353
                     yogurt rolls/buns bottled water
root vegetables
other vegetables
whole milk
yogurt
rolls/buns         0.8811115
bottled water      0.8987909  0.9104590
soda               0.9045422  0.8802034     0.8867700
```

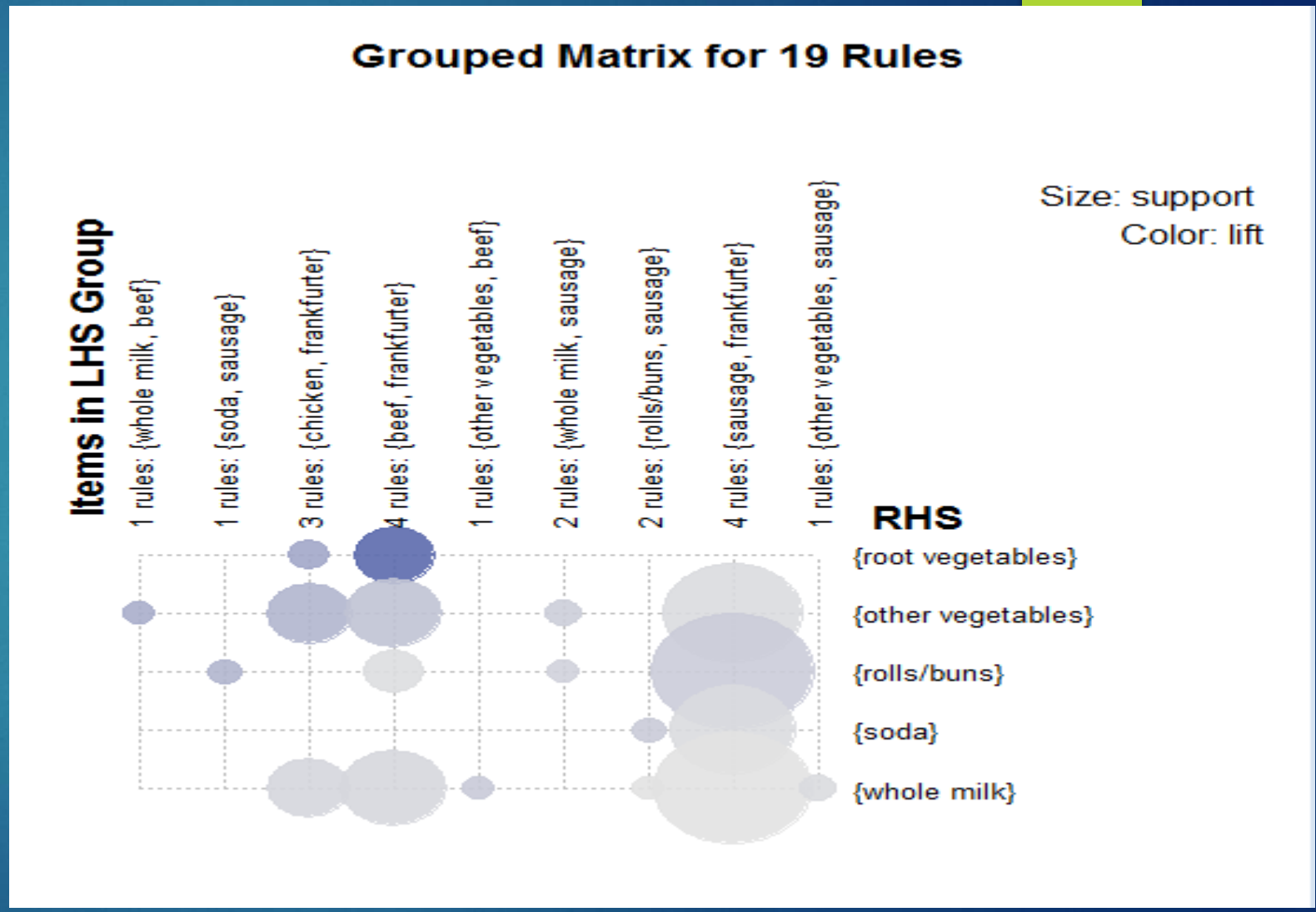
Advanced graphics:

REVA UNIVERSITY

```
plot(meat.rules, method="grouped",
measure="support",
control=list(col=sequential_hcl(100)))
```

can see that for beef and root vegetables, the lift is quite high and support is also quite significant.

For sausage support is quite significant for root vegetables, other vegetables,rools bun, soda and whole milk. However lift is not that good.



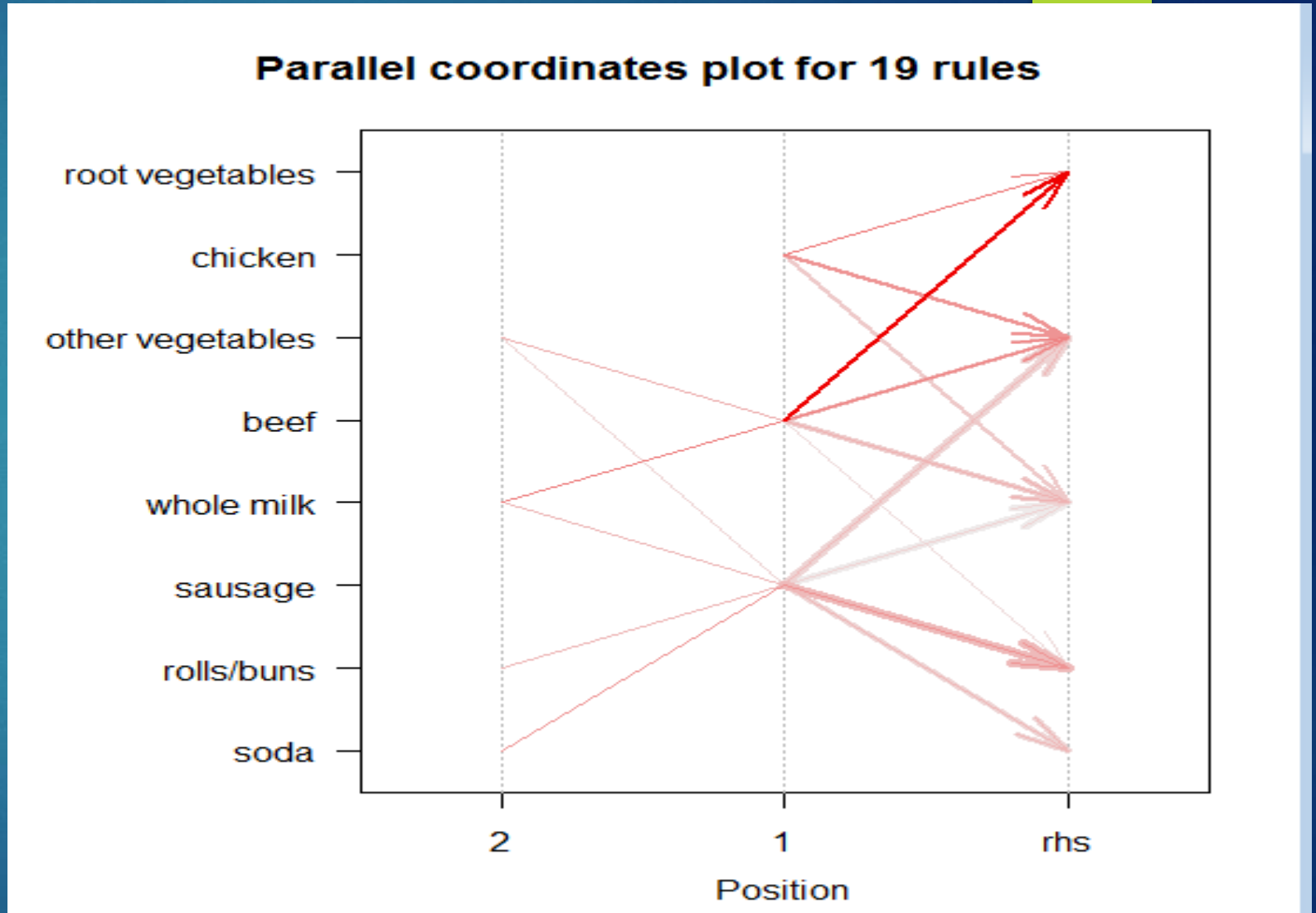
Advanced graphics:

REVA UNIVERSITY

```
plot(meat.rules,
method="paracoord",
control=list(reorder=TRUE))
```

We can even show dependencies with parallel coordinates plot. We can see that the mostly red arrow (each of them represents one rule) connects beef and root vegetables.

Moreover most of the arrows connect sausage on the first position, as previously stated.



Advanced graphics:

```
plot(meat.rules, method="matrix",
measure=c("support","confidence"),
control=list(col=sequential_hcl(200)))
```

Itemsets in Antecedent (LHS)

[1] "{beef,whole milk}" "{sausage,soda}" "{chicken}"

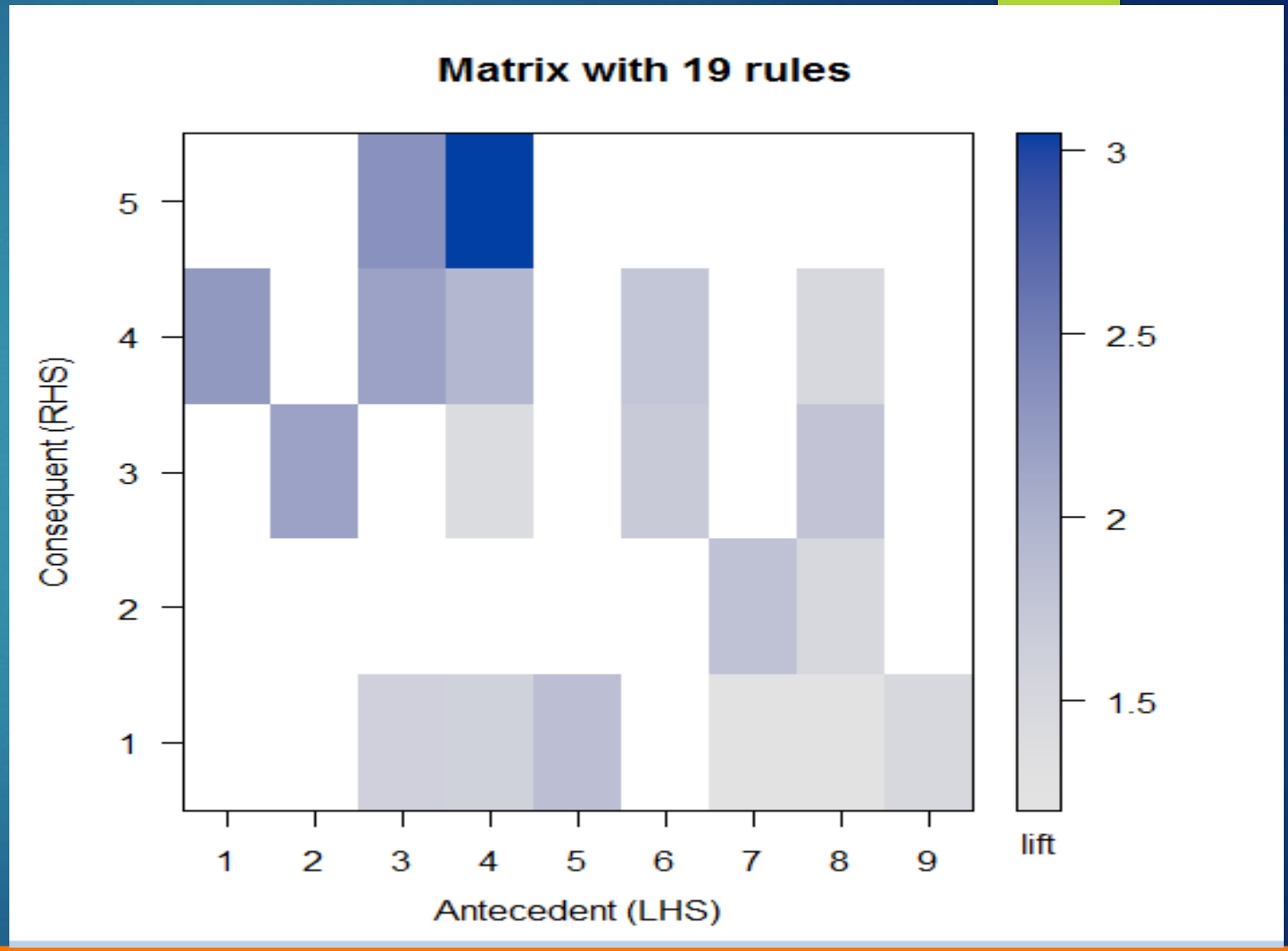
[4] "{beef}" "{beef,other vegetables}" "{sausage,whole milk}"

[7] "{sausage,rolls/buns}" "{sausage}" "{sausage,other vegetables}"

Itemsets in Consequent (RHS)

[1] "{whole milk}" "{soda}" "{rolls/buns}" "{other vegetables}"

[5] "{root vegetables}"



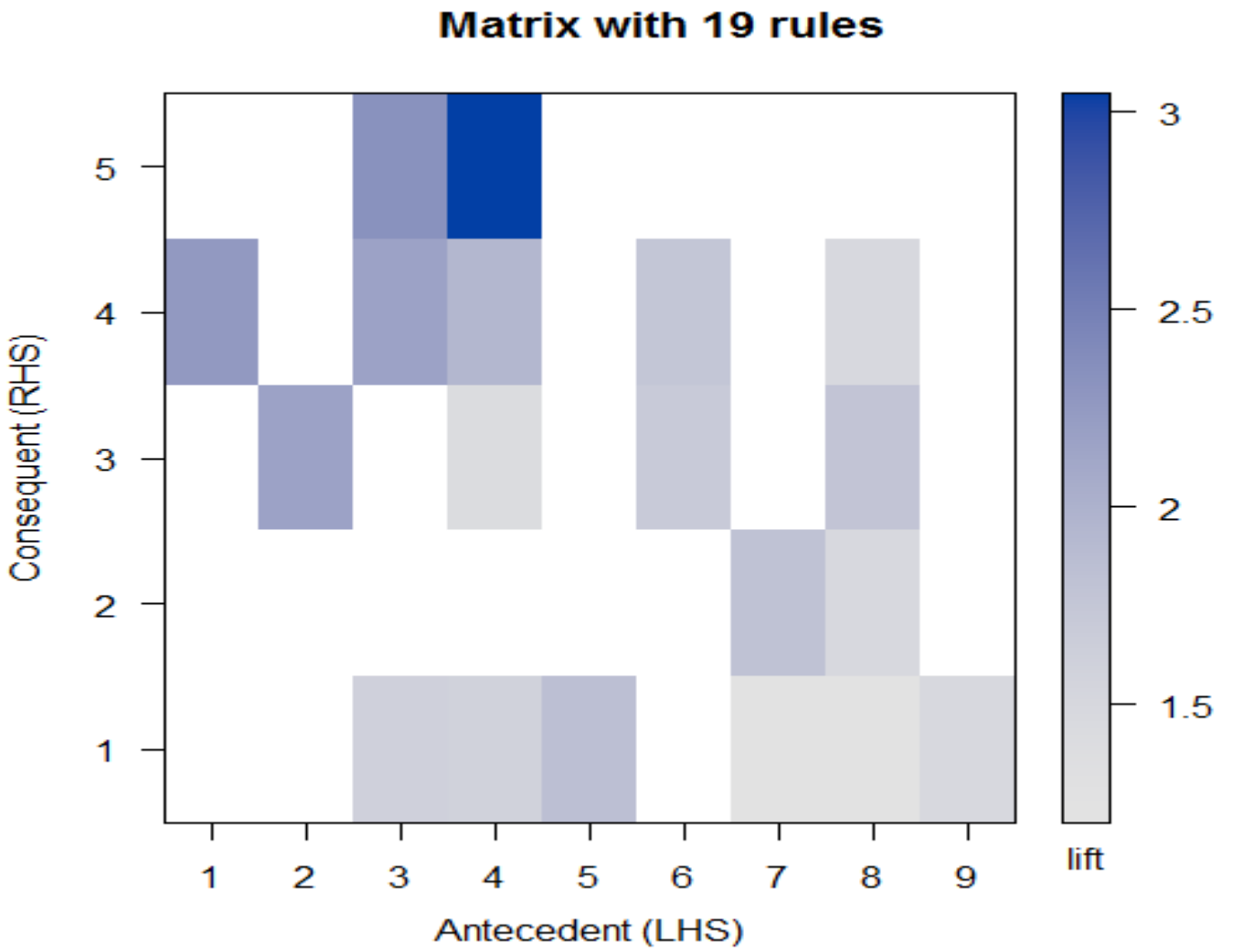
Advanced graphics:

Let's present the ruleset for meat but in a matrix form. Each of the matrix cells can have different blue shade depending on the lift value.

Numbers on the axes are corresponding to the items listed before the matrix.

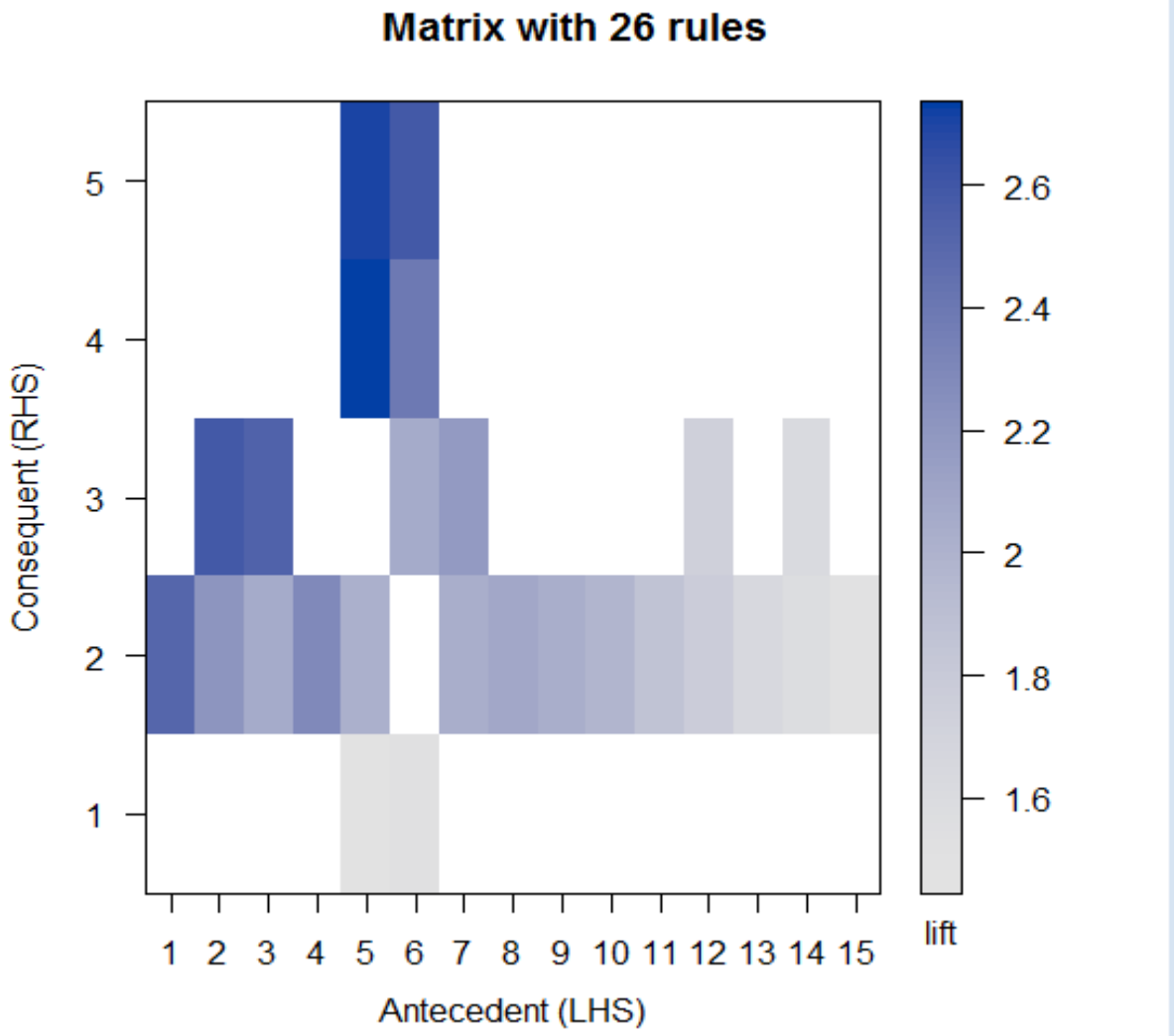
For example the most blue cell corresponds to the rule {beef} -> {root vegetables}, hence (as previously mentioned) root vegetables are most likely to be bought with beef. On the second place is the chicken and for the rest of antecedent items there is no significant lift at all (it is too small to be presented on the graph).

Such a graph is only confirmation fo the conclusions drawn before, but in a simplier form.



Advanced graphics:

```
...ot(yog.rules, method="matrix", measure=c("support","confidence"), control=list(col=sequential_...
(100)))
Itemsets in Antecedent (LHS)
 [1] "{butter,yogurt}"                "{root vegetables,yogurt}"
 [3] "{yogurt,whipped/sour cream}"    "{curd,yogurt}"
 [5] "{other vegetables,yogurt}"      "{whole milk,yogurt}"
 [7] "{tropical fruit,yogurt}"        "{pip fruit,yogurt}"
 [9] "{yogurt,pastry}"                "{yogurt,fruit/vegetable juice}"
[11] "{citrus fruit,yogurt}"          "{yogurt,rolls/buns}"
[13] "{yogurt,bottled water}"         "{yogurt}"
[15] "{yogurt,soda}"
Itemsets in Consequent (RHS)
[1] "{rolls/buns}"       "{whole milk}"       "{other vegetables}"
[4] "{root vegetables}"  "{tropical fruit}"
```

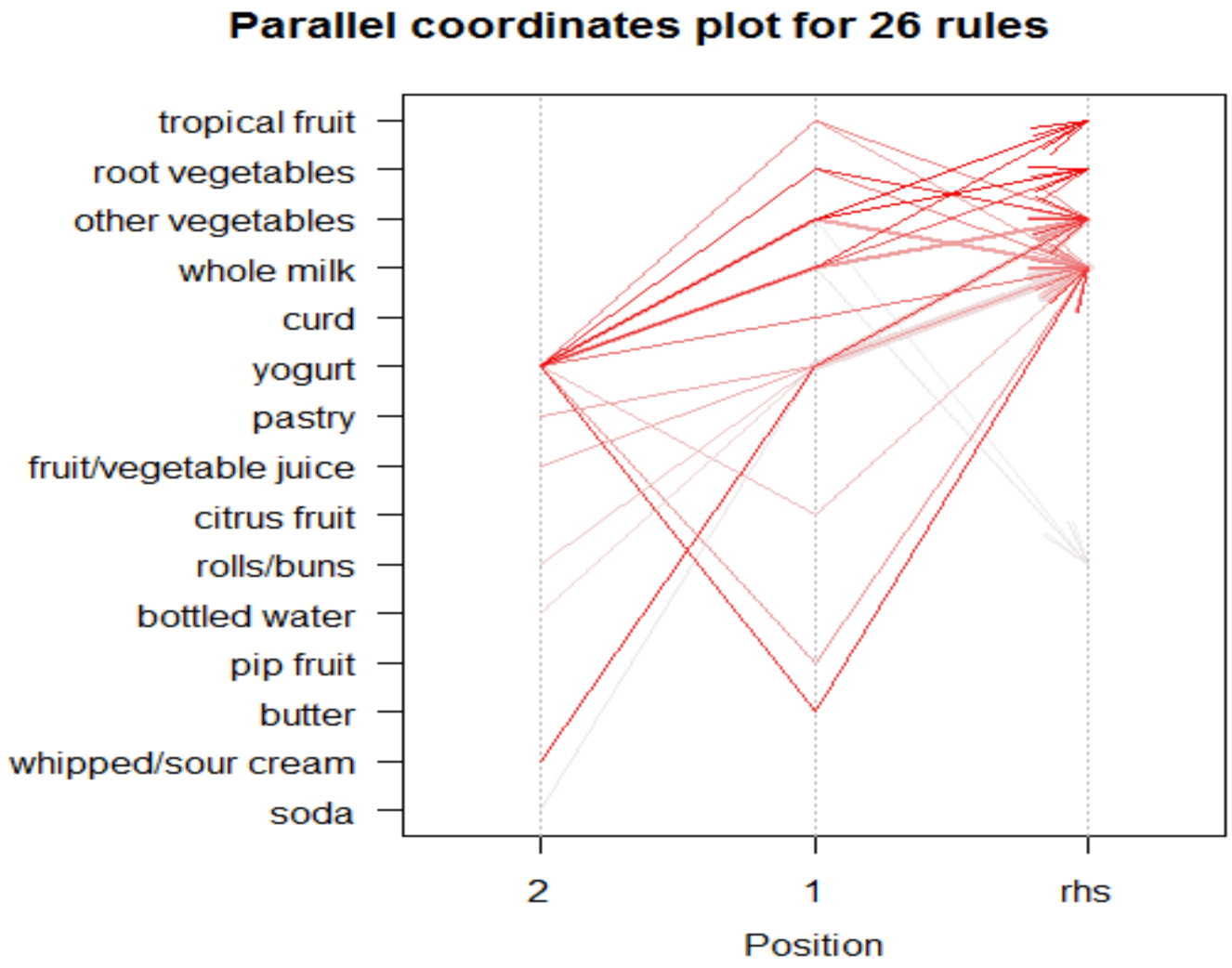


Advanced graphics:

REVA UNIVERSITY

Whole milk ,other vegetables, root vegetables,tropical fruits are the mostly supported additional product for yoghurt.

Moreover most of the arrows connect yoghurt on the first position, as previously stated. Also whipped/sour cream bought along with yoghurt for most of the times.



Advanced graphics:

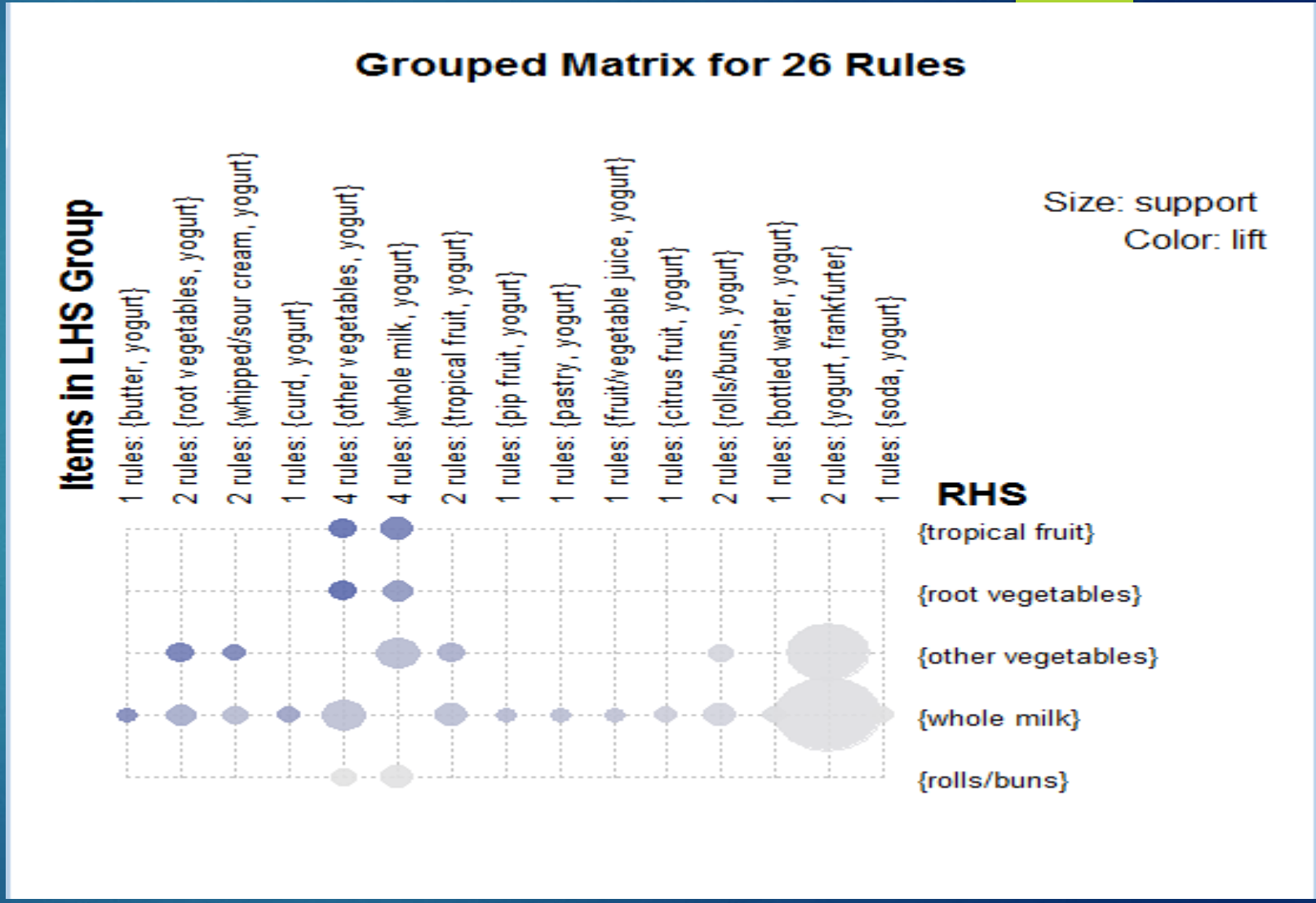
REVA UNIVERSITY

plot(meat.rules, method="grouped", measure="support", control=list(col=sequential_hcl(100)))

can see that for tropical fruits and other vegetables, yoghurt and whole milk, yoghurt the lift is quite high.

For root vegetable and other vegetables, yoghurt lift is quite high. For other vegetables, root vegetables, yoghurt lift is quite high.

For other items support is significant but lift is not very much notable.

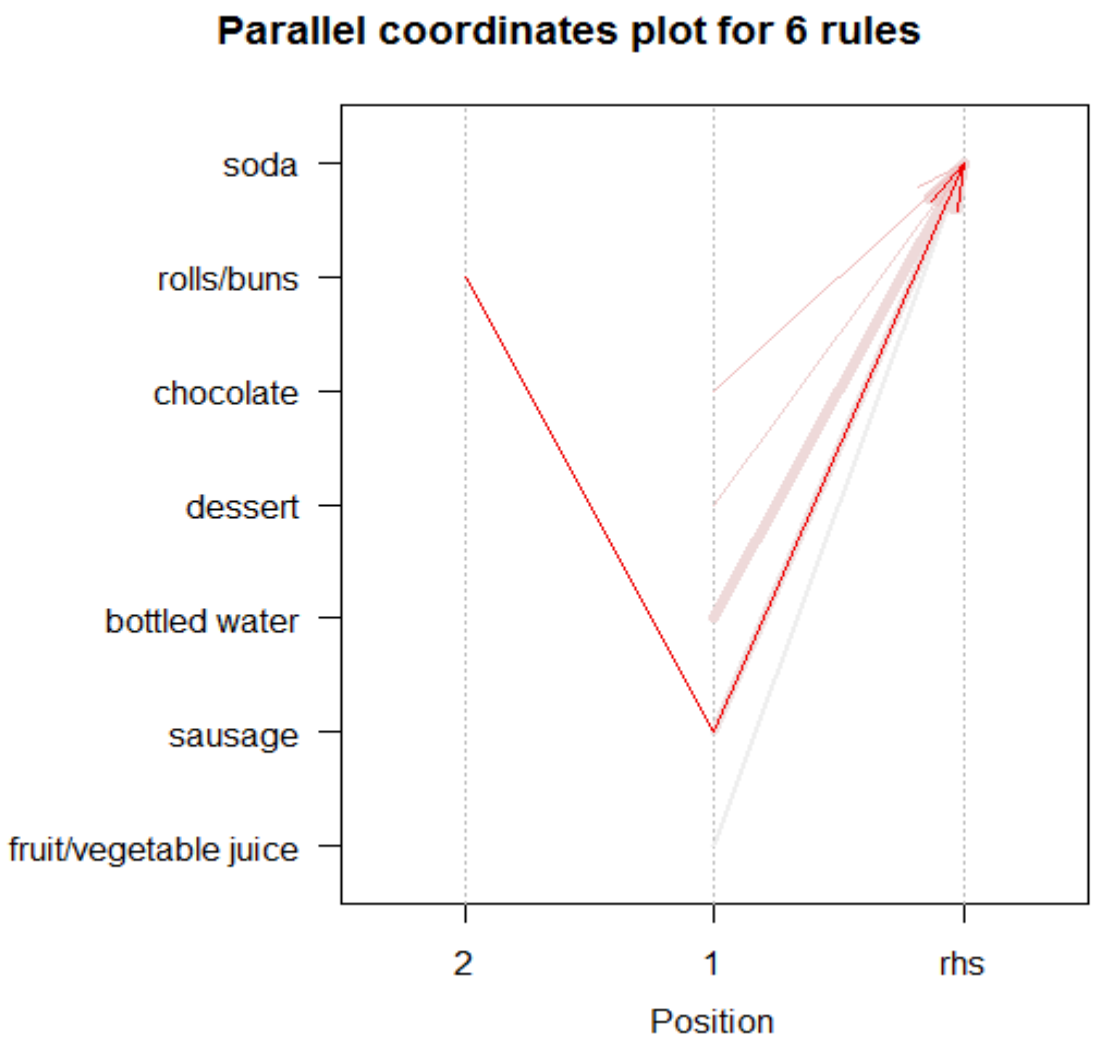


Advanced graphics:

REVA UNIVERSITY

rolls\bun bought along with soda with higher lift.

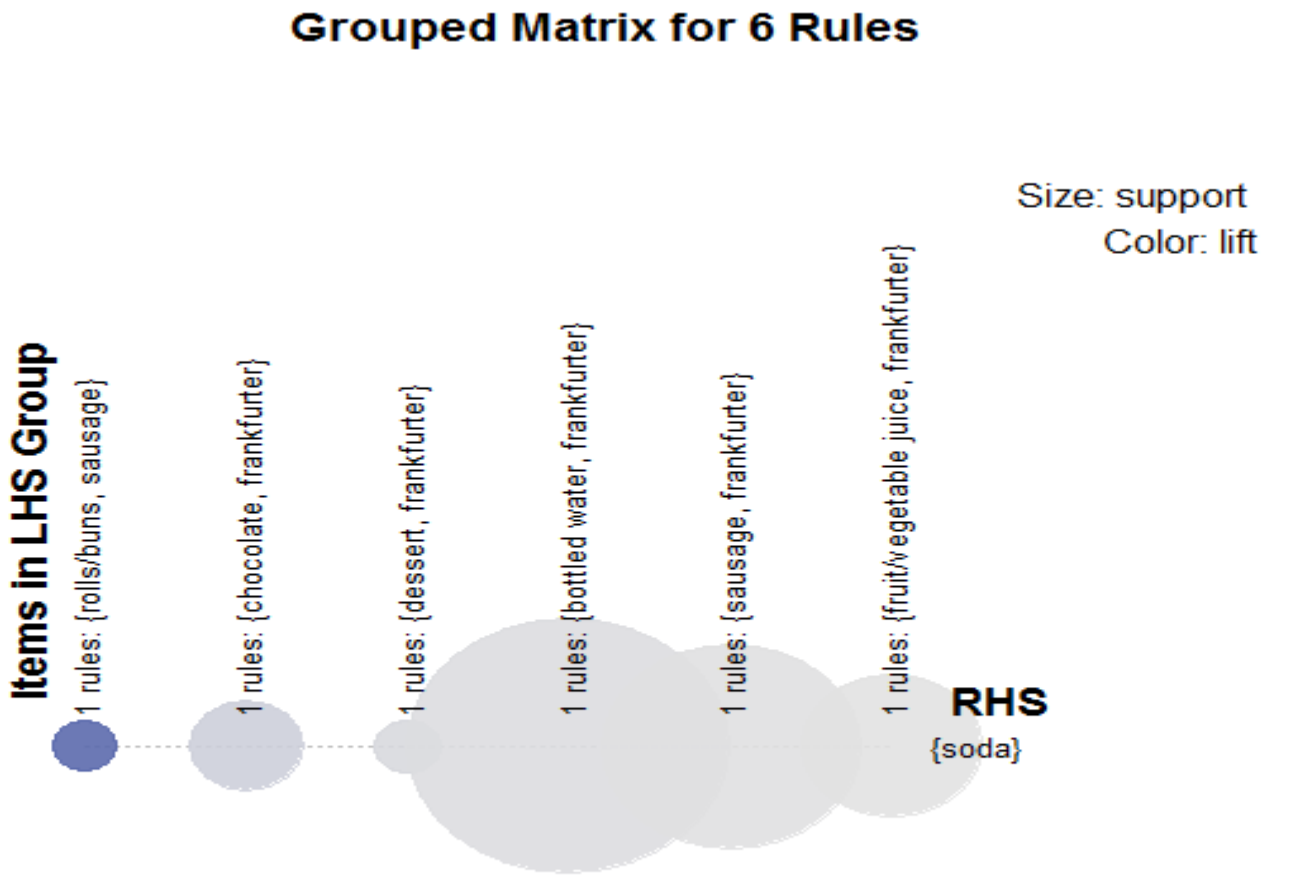
Moreover the arrows connect chocolate, dessert and boiled water on the first position.



Advanced graphics:

rolls\bun, sausage bought along with soda with higher lift.

There are other items in LHS whose support is bigger but lift is comparatively lower for them.



Tree map:

REVA UNIVERSITY

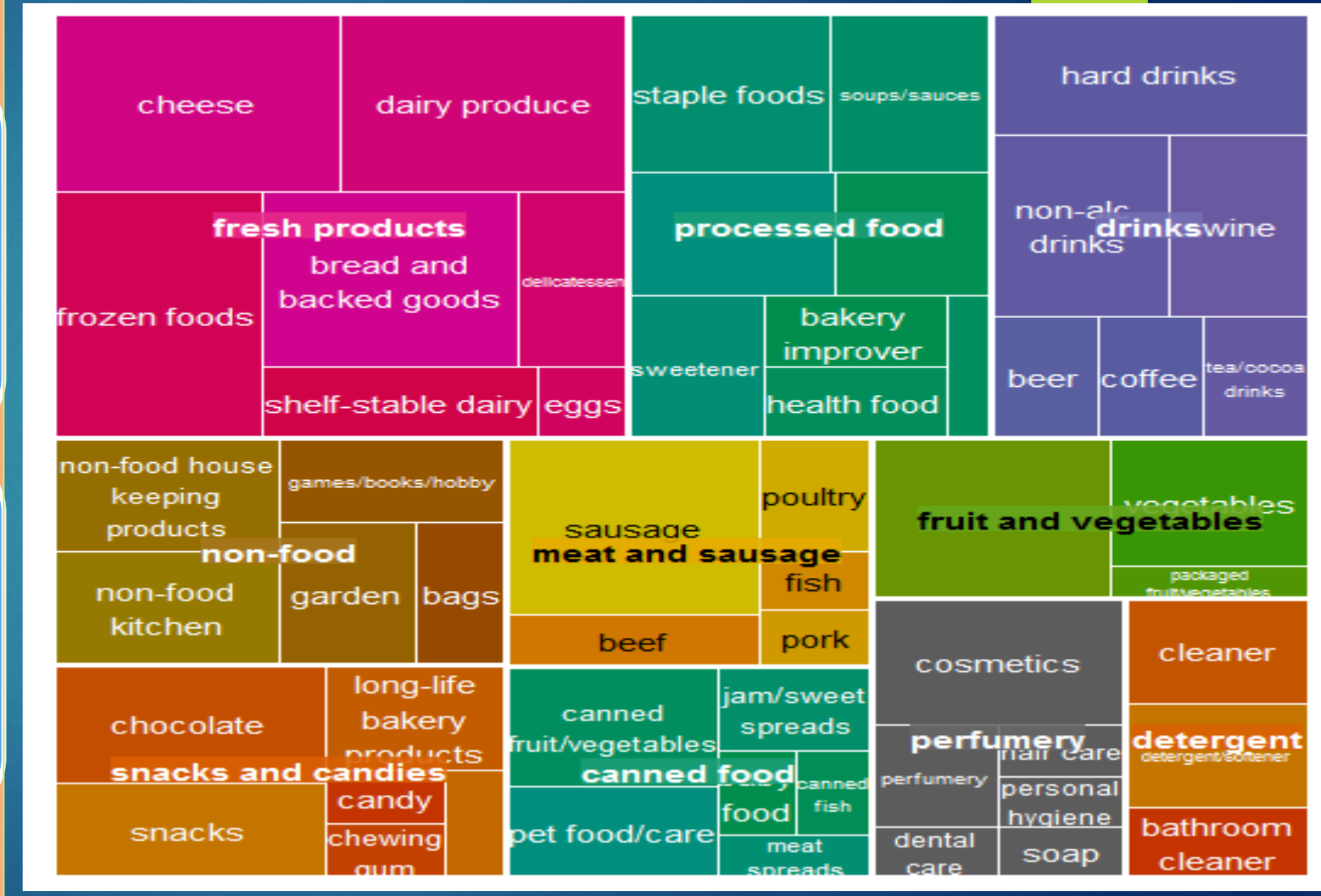
If we want to look into data deeper, we can create interesting plots that show us how many products of each type are available to buy in the grocery store.

Tree map:

REVA UNIVERSITY

Each of these charts have different level of depth. First only shows the bigger group names (like aisles in the shop).

Second shows deeper segmentation intro product types (for example within aisle).

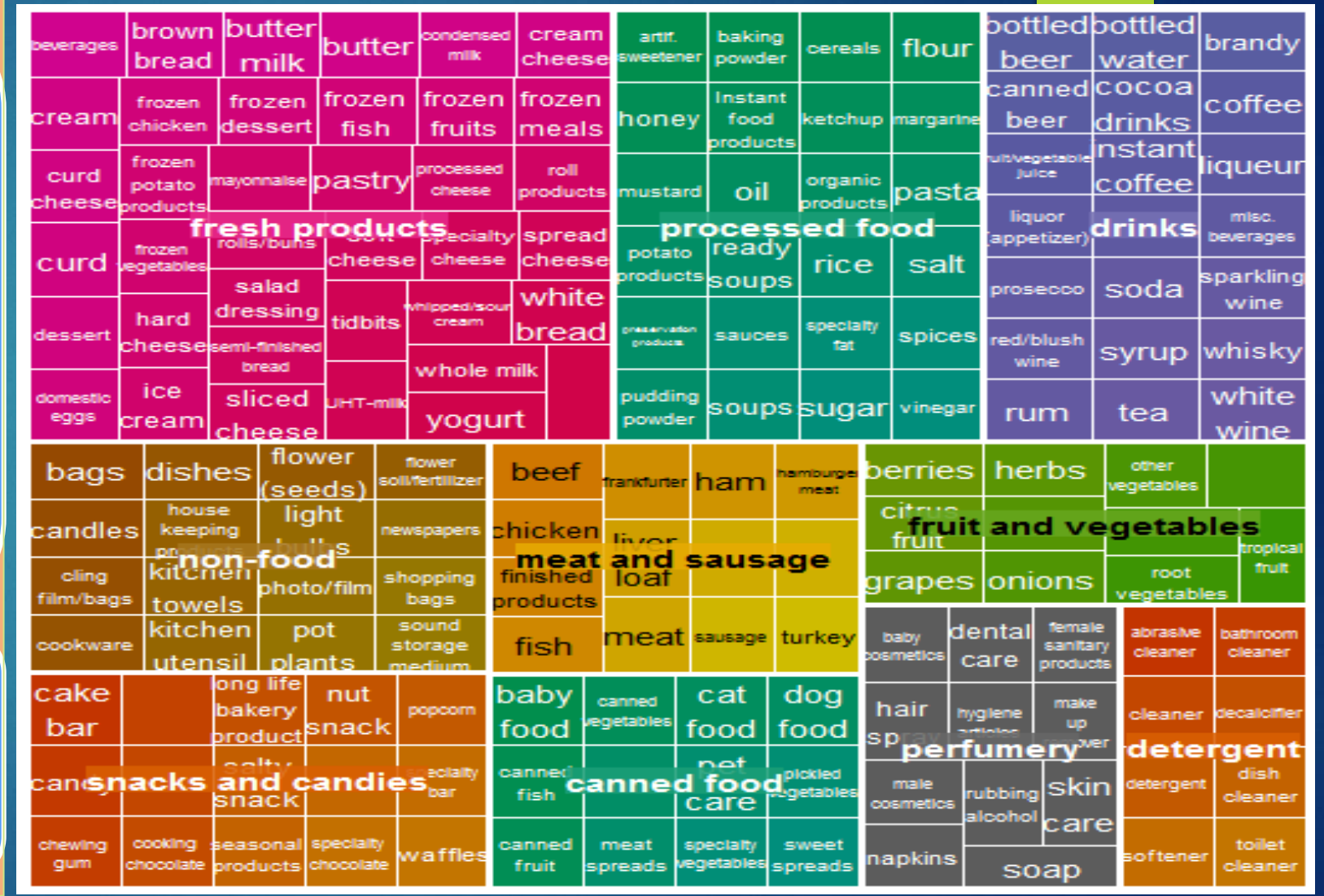


Tree map:

REVA UNIVERSITY

Here I made three tree maps, including one with deeper segmentation that present which products are on the lists of the shop. Moreover it can explain why there are so many connections with milk and fresh products, whereas just a little with coke.

The last chart presents each of the products available - it does give us less information than the second one.



Less than likely? - Lift < 1:

REVA UNIVERSITY

Interesting part of the study would be checking for items that are less than likely to be bought together. These would be described by lift < 1.

There is only one item in our rules set, that has lift less than 1. It is a connection between whole milk and bottled beer. It means that we are less likely to buy milk than any other product in dataset, while already having beer in basket.

```
> inspect(tail(sort(rules, by = "lift")))
    lhs                       rhs          support     confidence coverage
[1] {sausage}              => {whole milk} 0.029893238 0.3181818  0.09395018
[2] {bottled water}        => {whole milk} 0.034367056 0.3109476  0.11052364
[3] {rolls/buns}           => {whole milk} 0.056634469 0.3079049  0.18393493
[4] {sausage,rolls/buns}   => {whole milk} 0.009354347 0.3056478  0.03060498
[5] {salty snack}          => {whole milk} 0.011184545 0.2956989  0.03782410
[6] {bottled beer}         => {whole milk} 0.020437214 0.2537879  0.08052872
    lift      count
[1] 1.2452520 294
[2] 1.2169396 338
[3] 1.2050318 557
[4] 1.1961984  92
[5] 1.1572618 110
[6] 0.9932367 201
```