MARKET BASKET

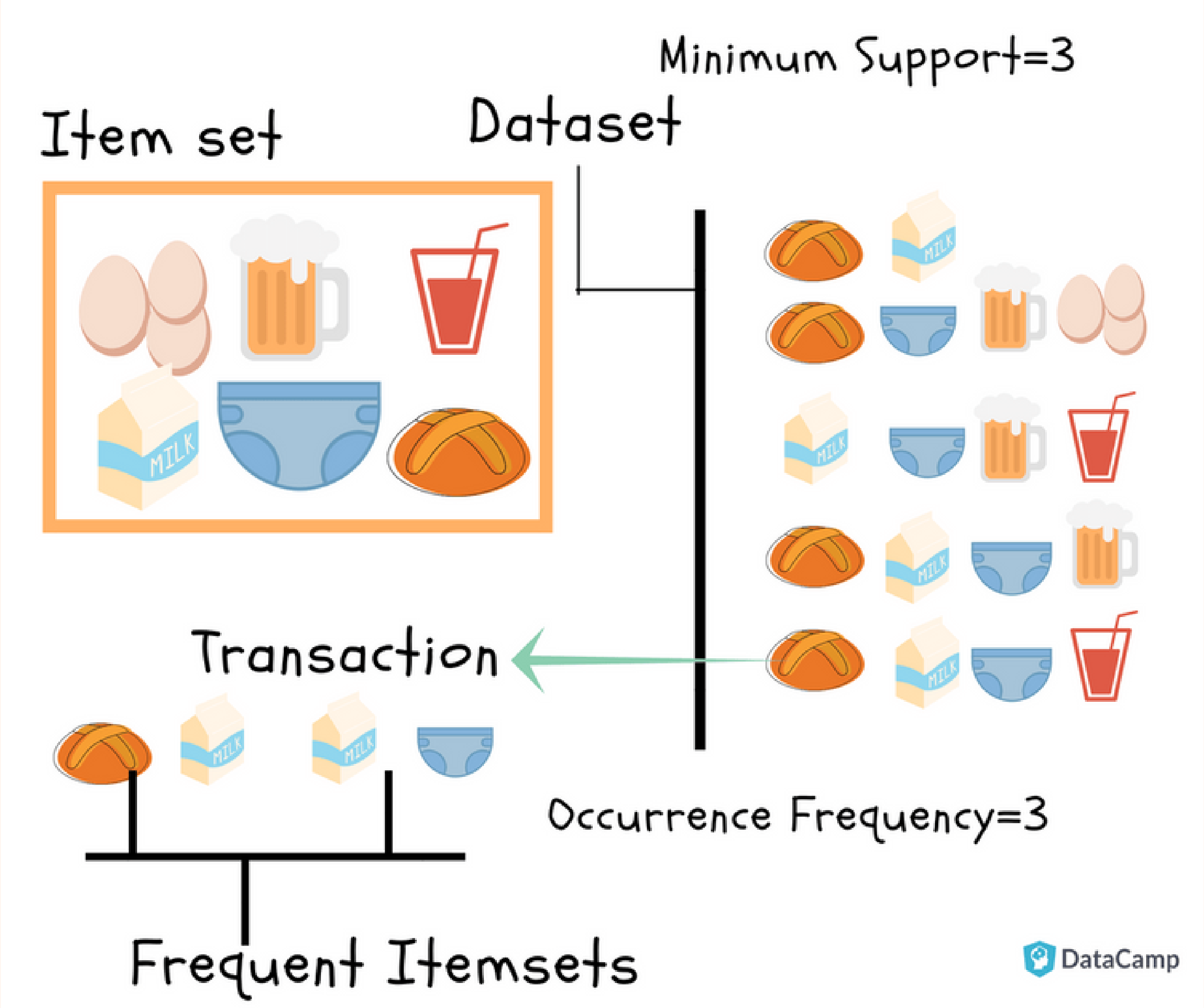
analysis

report by khanhuyenthai

1.Introduction

Market Basket Analysis is a method used by large retailers to discover product correlations. Purchase behavior can be well determined through constant checks on items that frequently appear together in transactions.

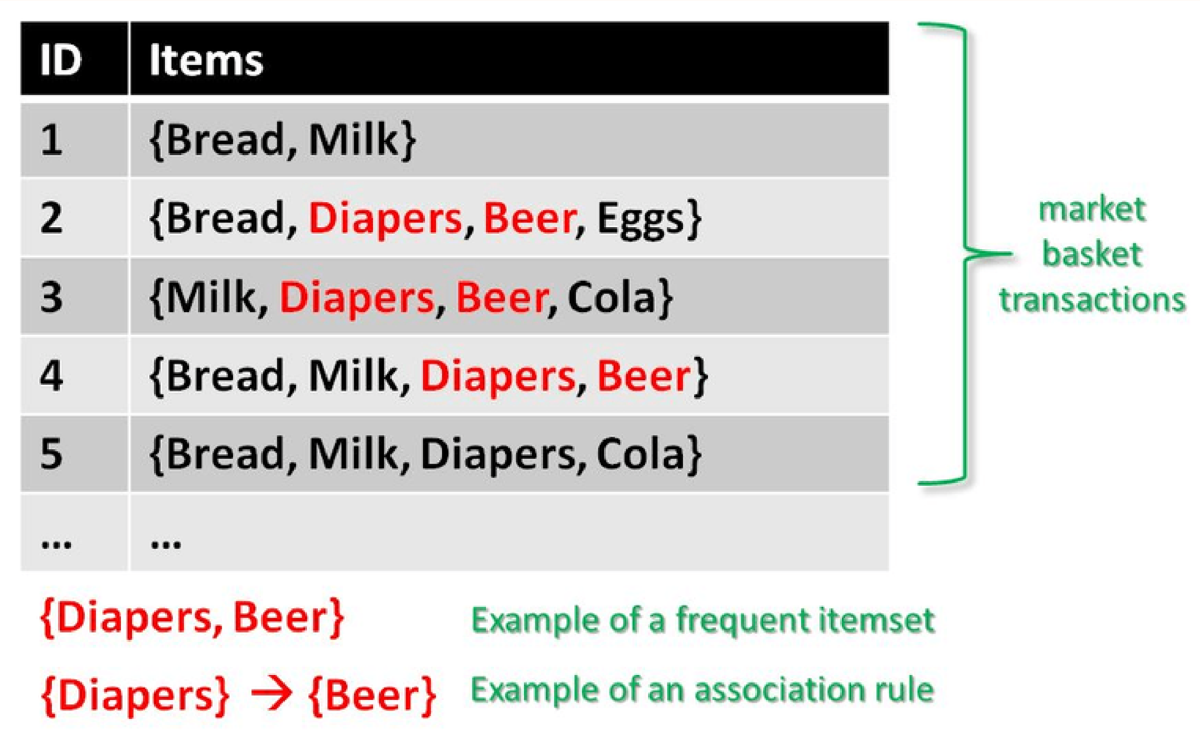
Association Rule Mining is used when you want to find an association between different objects in a set, find frequent patterns in a transaction database, relational databases or any other information repository.



report by khanhhuyenthai

Source: Market Basket Analysis in R

Association Rule Mining



ID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}
…	…

market basket transactions

{Diapers, Beer} Example of a frequent itemset

{Diapers} → {Beer} Example of an association rule

There are some important indicators:

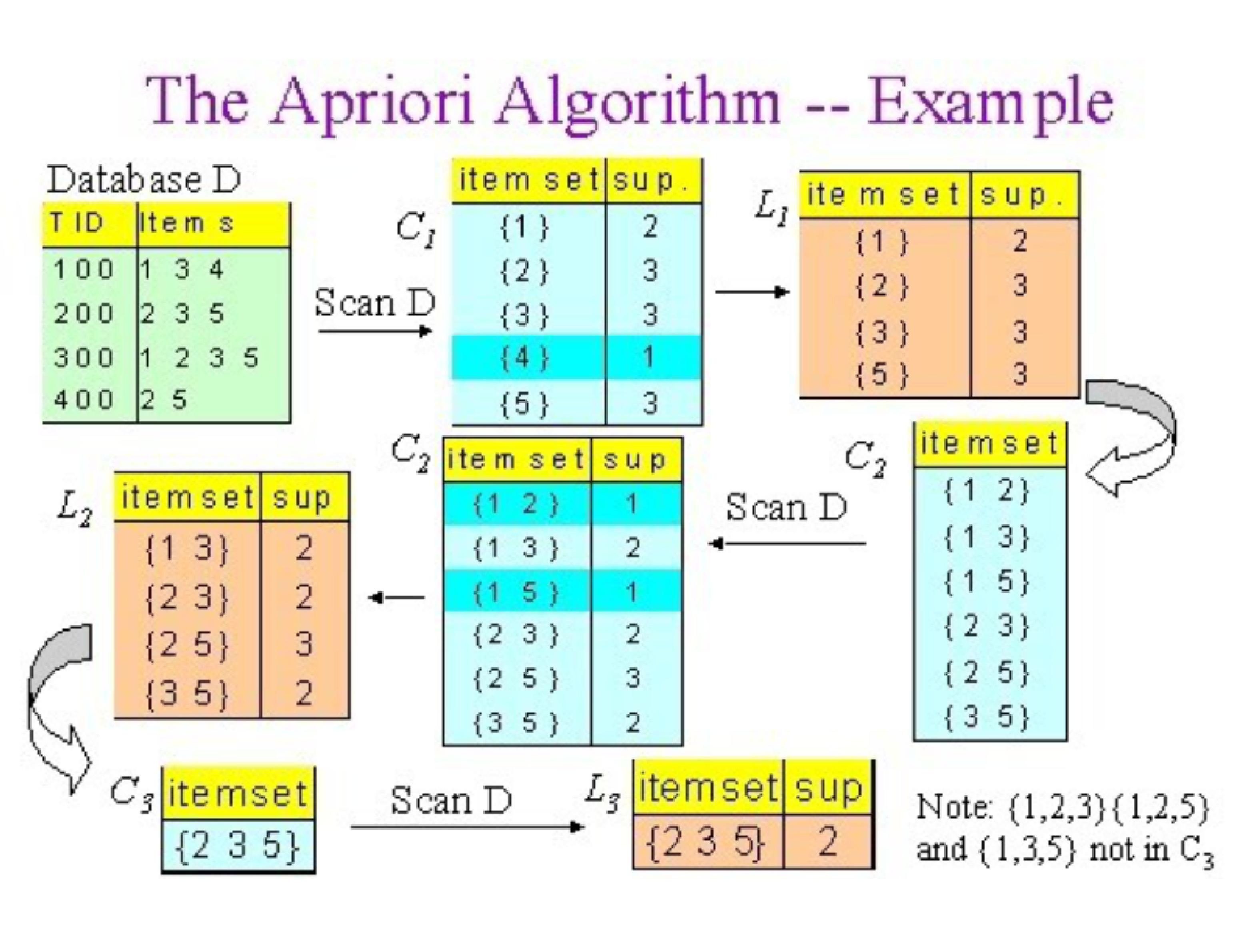
support: number of transactions which contain “item sets” or “antecedent and consequent”.
confidence: the co-occurrence of the “item sets” / the occurrence of antecedent.
Benchmark confidence: the occurrence of consequent / all baskets
lift: confidence/ benchmark confidence

- If the rule had a lift of 1,then A and B are independent and no rule can be derived from them.
- If the lift is > 1, then A and B are dependent on each other, and the degree of which is given by ift value.
- If the lift is < 1, then presence of A will have negative effect on B.

Source: Market Basket Analysis using R (DataCapm)

Apriori Algorithm

It is an algorithm developed to extract the relationship between data in machine learning. The algorithm uses a bottom-up approach, examines one data at a time and seeks a relationship between this data and others.



The Apriori Algorithm -- Example

Database D

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

C_1

itemset	sup.
{1}	2
{2}	3
{3}	3
{4}	1
{5}	3

L_1

itemset	sup.
{1}	2
{2}	3
{3}	3
{5}	3

C_2

itemset
{1 2}
{1 3}
{1 5}
{2 3}
{2 5}
{3 5}

C_2

itemset	sup
{1 2}	1
{1 3}	2
{1 5}	1
{2 3}	2
{2 5}	3
{3 5}	2

L_2

itemset	sup
{1 3}	2
{2 3}	2
{2 5}	3
{3 5}	2

C_3

itemset
{2 3 5}

L_3

itemset	sup
{2 3 5}	2

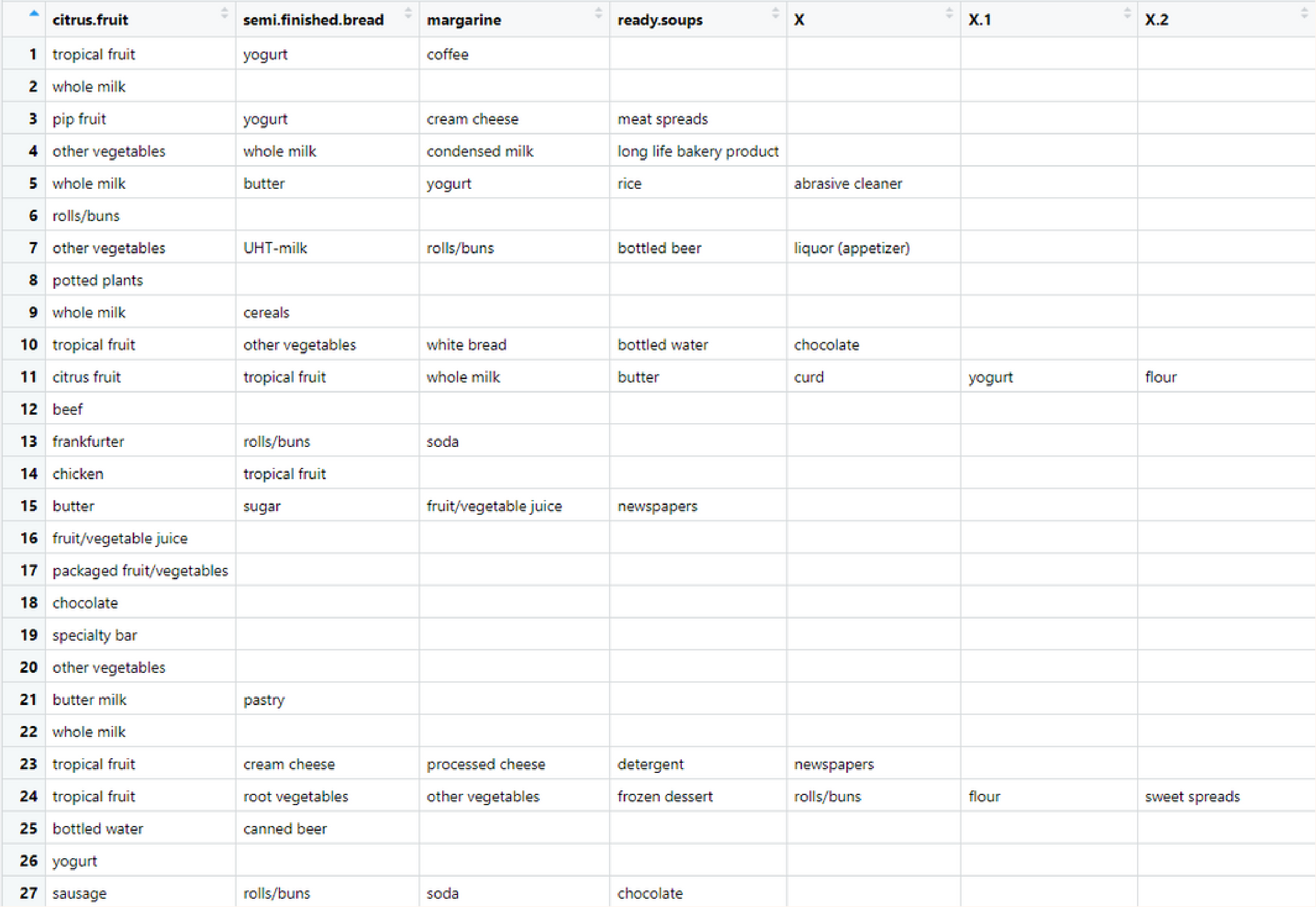
Note: {1,2,3}{1,2,5} and {1,3,5} not in C_3

Source: Market Basket Analysis on Groceries With Association Rules

For example, suppose the above figure is the shopping baskets of customers in a market. When we look at the first table, we see the products received. (1 3 4–2 3 4 etc.) The algorithm first finds the frequency of these products, ie the total number of intakes. (1st product was bought 2 times, 3rd product is 3 times etc.)

After finding these values, it gets the minimum support value of the highest frequency (50–3 * 50/100 = 1.5%) and those whose frequency is less than this value are eliminated. By combining the remaining values, the same process is repeated and the table is further reduced. This continues until a relationship is found.

2. Data Preparation



	citrus.fruit	semi.finished.bread	margarine	ready.soups	X	X.1	X.2
1	tropical fruit	yogurt	coffee				
2	whole milk						
3	pip fruit	yogurt	cream cheese	meat spreads			
4	other vegetables	whole milk	condensed milk	long life bakery product			
5	whole milk	butter	yogurt	rice	abrasive cleaner		
6	rolls/buns						
7	other vegetables	UHT-milk	rolls/buns	bottled beer	liquor (appetizer)		
8	potted plants						
9	whole milk	cereals					
10	tropical fruit	other vegetables	white bread	bottled water	chocolate		
11	citrus fruit	tropical fruit	whole milk	butter	curd	yogurt	flour
12	beef						
13	frankfurter	rolls/buns	soda				
14	chicken	tropical fruit					
15	butter	sugar	fruit/vegetable juice	newspapers			
16	fruit/vegetable juice						
17	packaged fruit/vegetables						
18	chocolate						
19	specialty bar						
20	other vegetables						
21	butter milk	pastry					
22	whole milk						
23	tropical fruit	cream cheese	processed cheese	detergent	newspapers		
24	tropical fruit	root vegetables	other vegetables	frozen dessert	rolls/buns	flour	sweet spreads
25	bottled water	canned beer					
26	yogurt						
27	sausage	rolls/buns	soda	chocolate			

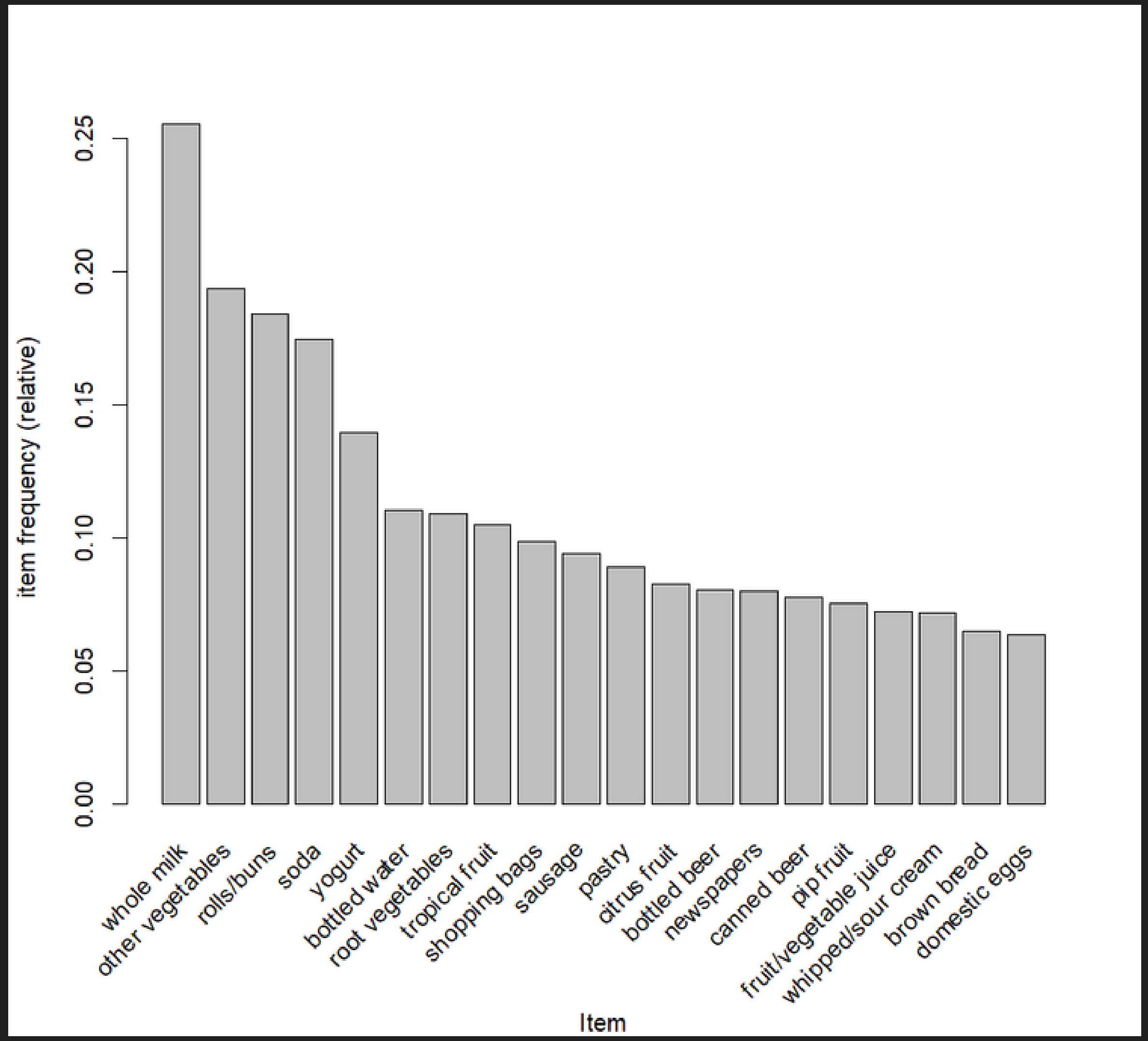
dataset groceries

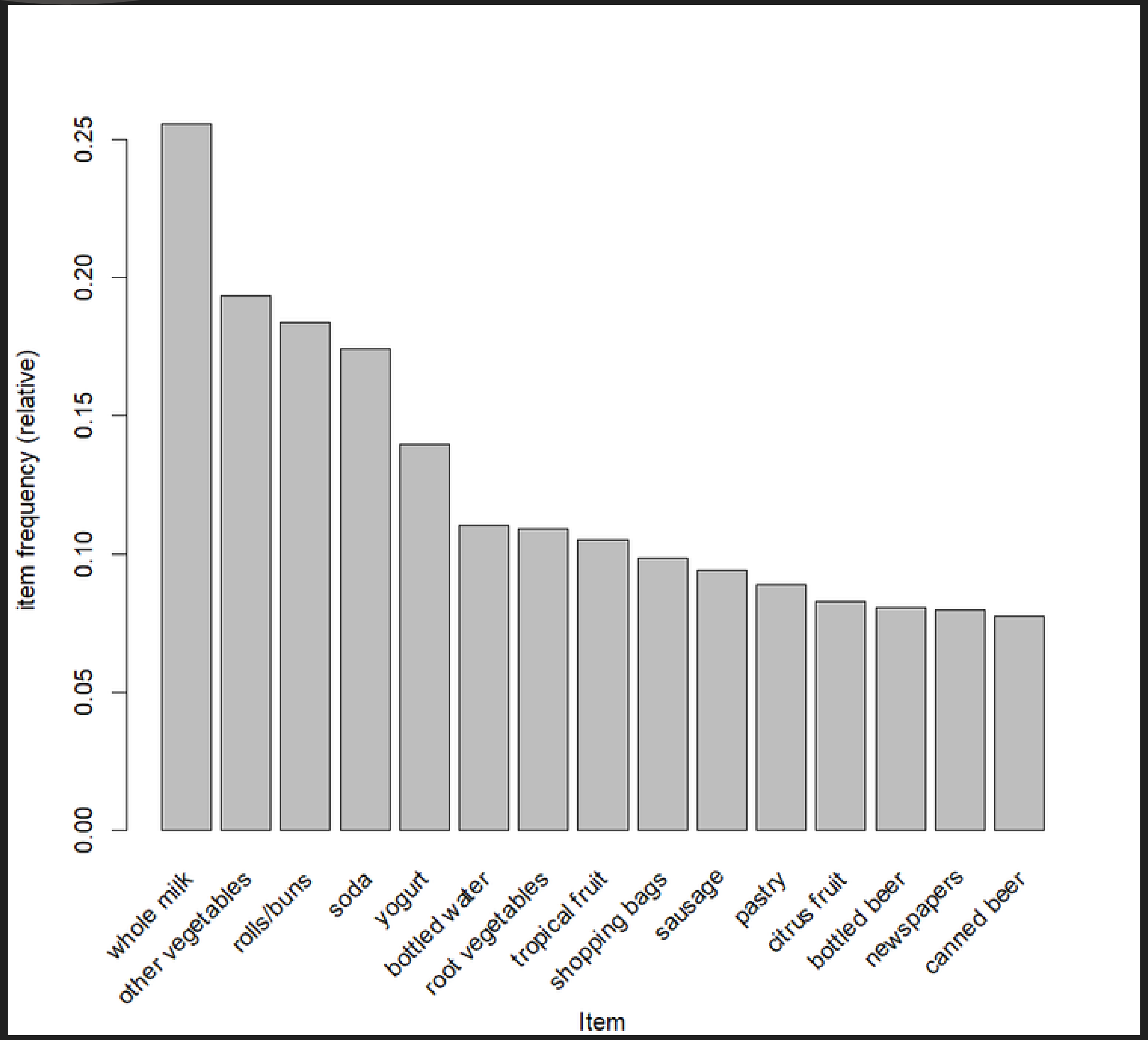
9835 obs 32 variables

The dataset describes the shopping lists of supermarket customers. The data contains no NA values and is categorical. It includes 169 unique items.

Following applications done by the data owner. The csv file was read transaction by transaction and each transaction was saved as a list. A mapping was created from the unique items in the dataset to integers so that each item

EXPLORATORY DATA





In the TOP 20 or TOP 15 products, whole milk is purchased with the highest frequency. We can leverage this to boost the sales of other product lines by promoting products (bundled with whole milk) or vice versa.

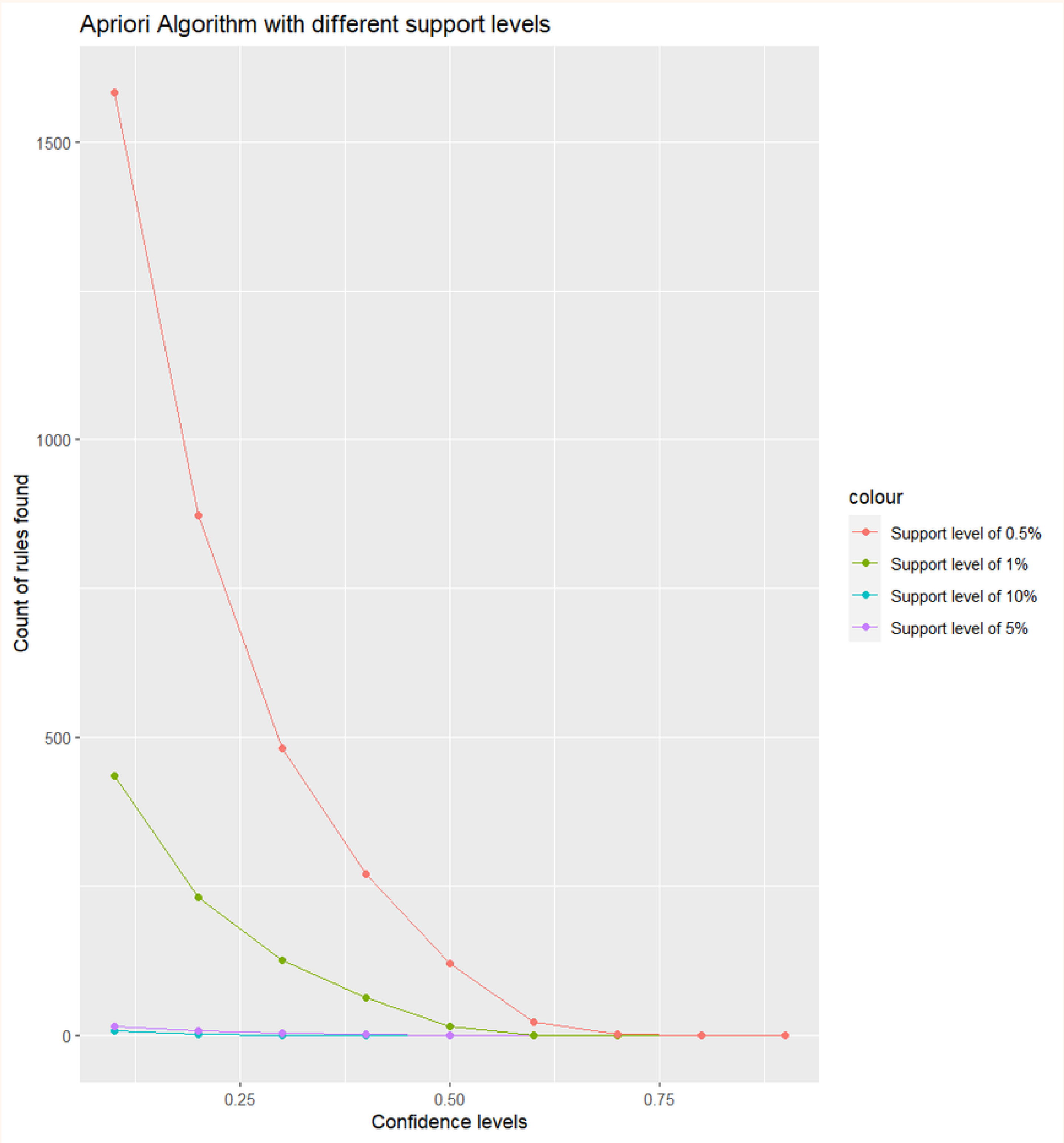
Data Modeling

The number of rules generated with a support level of 10%, 5%, 1% and 0.5% are shown as below
Code

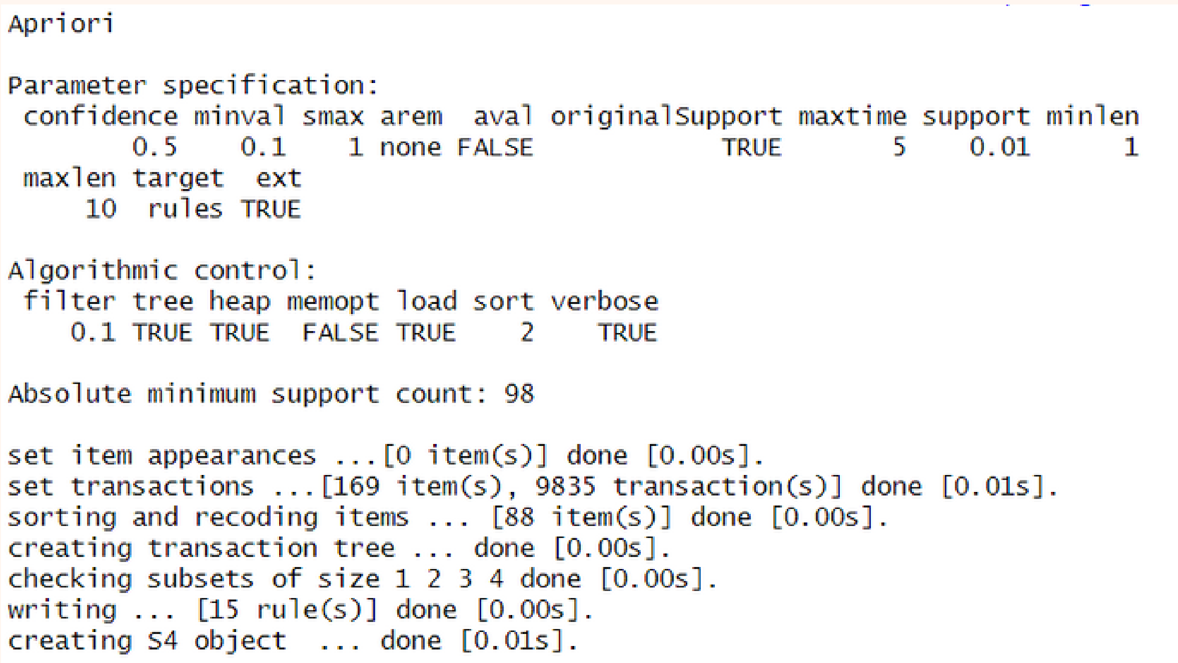
The results are as follows:

- **Support level of 10%**. The rules are generated with very low confidence level.
- **Support level of 5%**. We must look for support levels below 5% for rules with reasonable level of confidence.
- **Support level of 1%**. 13 rules have a confidence of at least 50%.
- **Support level of 0.5%**. Too many rules to analyze!

Thus, support level of 1% and a confidence level of 50%.



When selecting support = 1% and confidence = 50%, the algorithm will generate 15 rules.

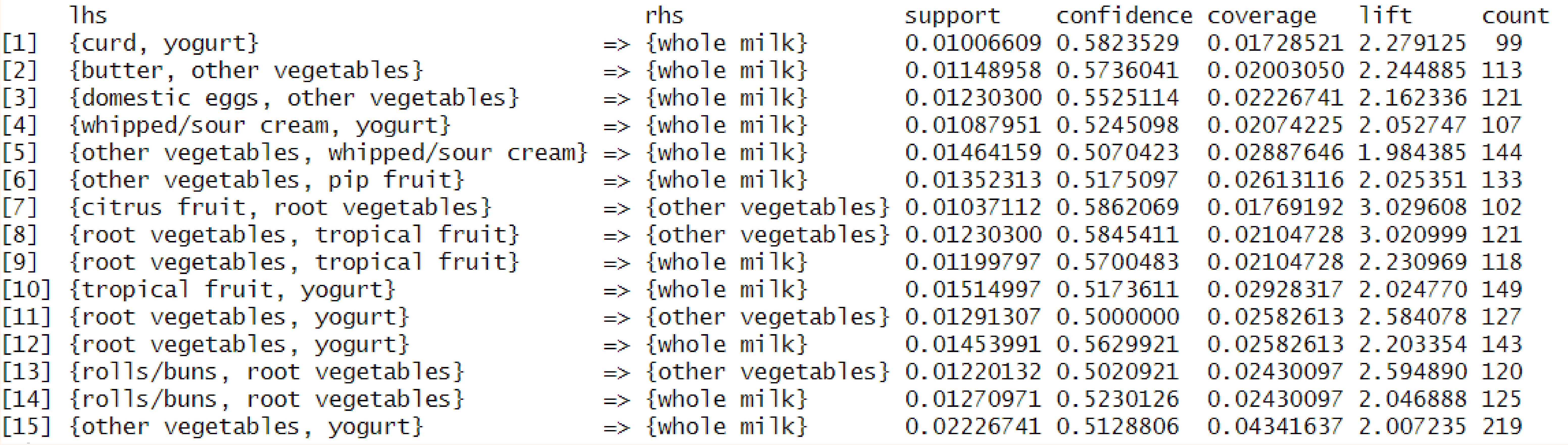


```
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen
        0.5    0.1    1 none FALSE            TRUE       5    0.01      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: 98

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[169 item(s), 9835 transaction(s)] done [0.01s].
sorting and recoding items ... [88 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 done [0.00s].
writing ... [15 rule(s)] done [0.00s].
creating S4 object  ... done [0.01s].
```

Finalization

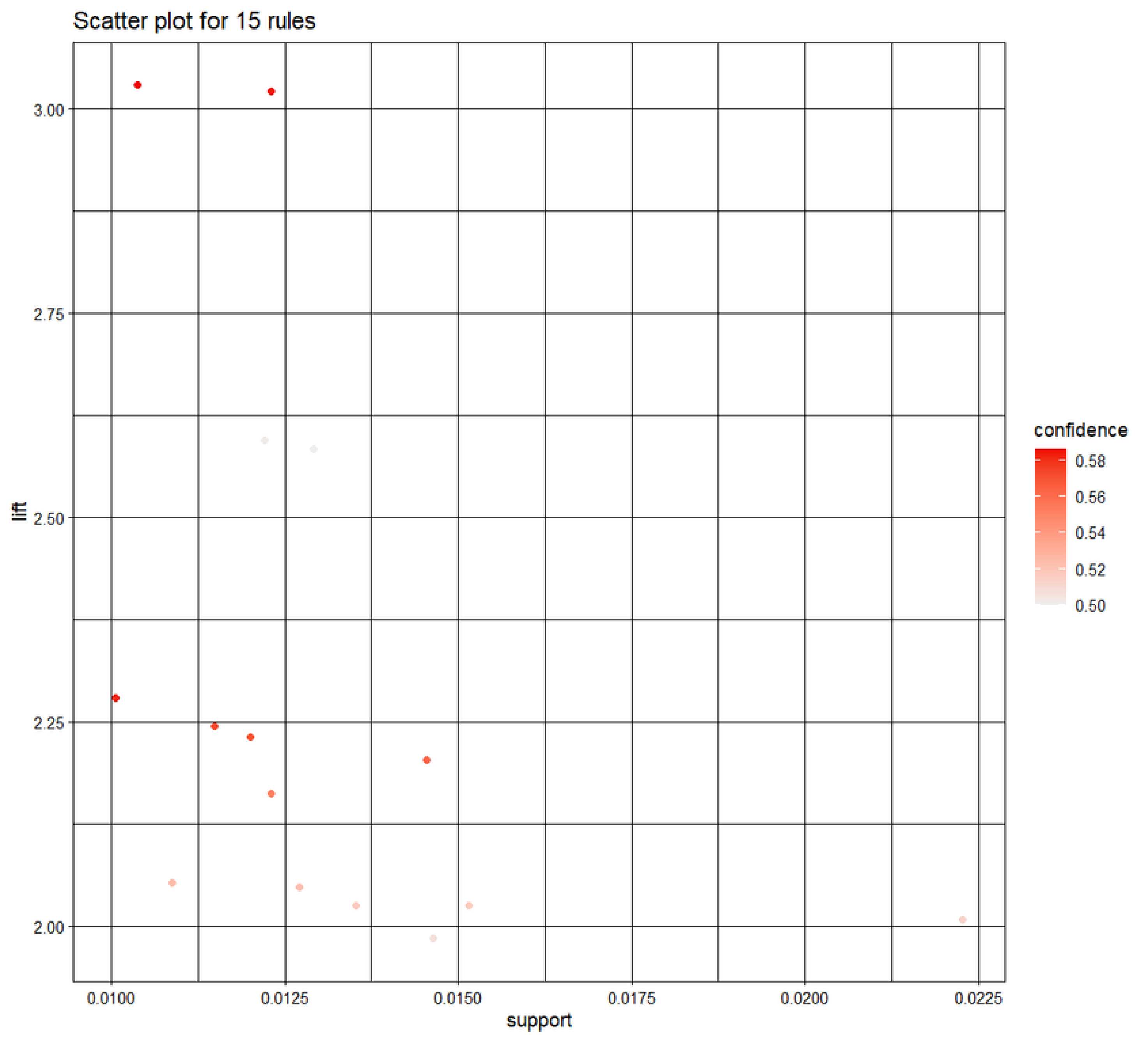
The rules can be interpreted as follows:

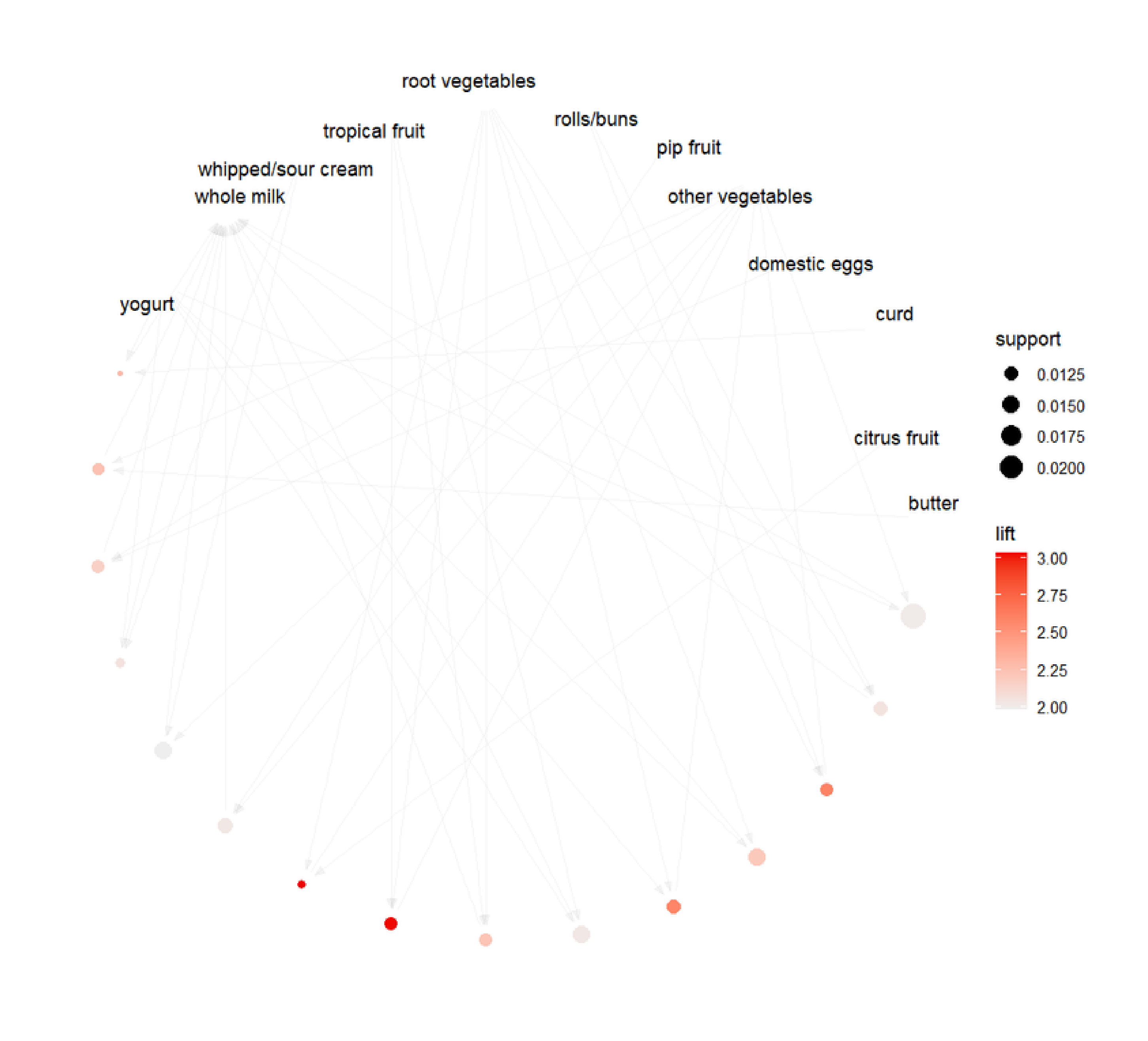
- 58,2% of the customers who bought curd, yogurt also bought a wholemilk.
- 59% of the customers who bought citrus fruit and root vegetables also bought a wholemilk.



```
     lhs                                        rhs                support    confidence coverage   lift     count
[1]  {curd, yogurt}                          => {whole milk}       0.01006609 0.5823529  0.01728521 2.279125  99
[2]  {butter, other vegetables}              => {whole milk}       0.01148958 0.5736041  0.02003050 2.244885 113
[3]  {domestic eggs, other vegetables}       => {whole milk}       0.01230300 0.5525114  0.02226741 2.162336 121
[4]  {whipped/sour cream, yogurt}            => {whole milk}       0.01087951 0.5245098  0.02074225 2.052747 107
[5]  {other vegetables, whipped/sour cream} => {whole milk}       0.01464159 0.5070423  0.02887646 1.984385 144
[6]  {other vegetables, pip fruit}           => {whole milk}       0.01352313 0.5175097  0.02613116 2.025351 133
[7]  {citrus fruit, root vegetables}         => {other vegetables} 0.01037112 0.5862069  0.01769192 3.029608 102
[8]  {root vegetables, tropical fruit}       => {other vegetables} 0.01230300 0.5845411  0.02104728 3.020999 121
[9]  {root vegetables, tropical fruit}       => {whole milk}       0.01199797 0.5700483  0.02104728 2.230969 118
[10] {tropical fruit, yogurt}                => {whole milk}       0.01514997 0.5173611  0.02928317 2.024770 149
[11] {root vegetables, yogurt}               => {other vegetables} 0.01291307 0.5000000  0.02582613 2.584078 127
[12] {root vegetables, yogurt}               => {whole milk}       0.01453991 0.5629921  0.02582613 2.203354 143
[13] {rolls/buns, root vegetables}           => {other vegetables} 0.01220132 0.5020921  0.02430097 2.594890 120
[14] {rolls/buns, root vegetables}           => {whole milk}       0.01270971 0.5230126  0.02430097 2.046888 125
[15] {other vegetables, yogurt}              => {whole milk}       0.02226741 0.5128806  0.04341637 2.007235 219
```

Visualizations of association rules


Scatter plot for 15 rules
lift
support
confidence


root vegetables
tropical fruit
rolls/buns
pip fruit
whipped/sour cream
whole milk
other vegetables
domestic eggs
yogurt
curd
citrus fruit
butter
support
lift