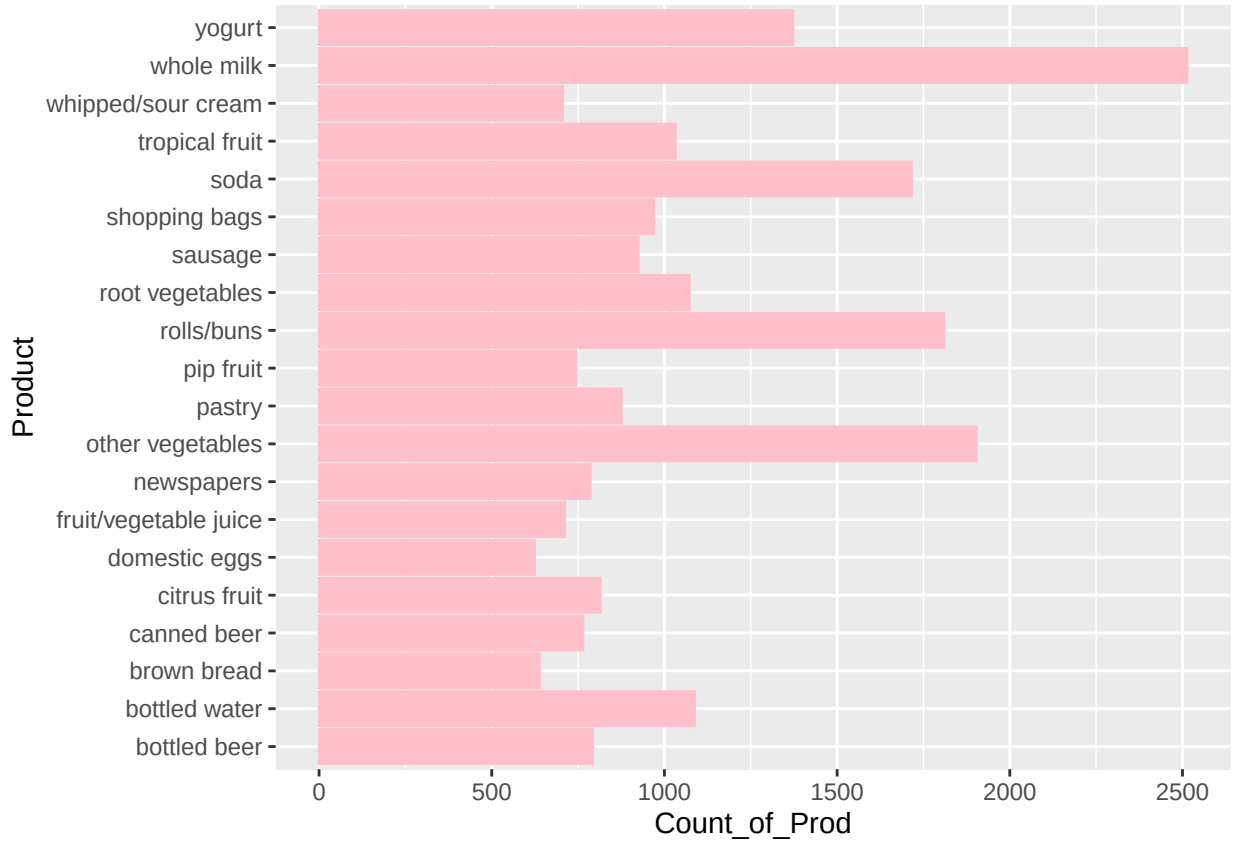
groceries_Arm

2023-08-10

Reading the data

Reading the data in long format

Identify the most bought items



We see that Whole milk, Other vegetable, Rolls/buns and Soda are the top 4 frequently bought products.

Splitting data to use the apriori algorithm

```
## 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
##
## includes extended transaction information - examples:
##   transactionID
## 1             1
## 2             2
## 3             3
```

Running the 'apriori' algorithm

Rule 1

Here we have a support of 0.005 and confidence of 0.1. This indicates that atleast 0.5% of all the transactions have that particular combination of products and out of all the antecedents at least 10% of them is a consequent. For example, if cereal is the antecedent, and milk is consequent, then of the orders containing cereals, 10% of them are likely to have whole milk.

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##         0.1    0.1    1 none FALSE            TRUE       5   0.005      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: 49
```

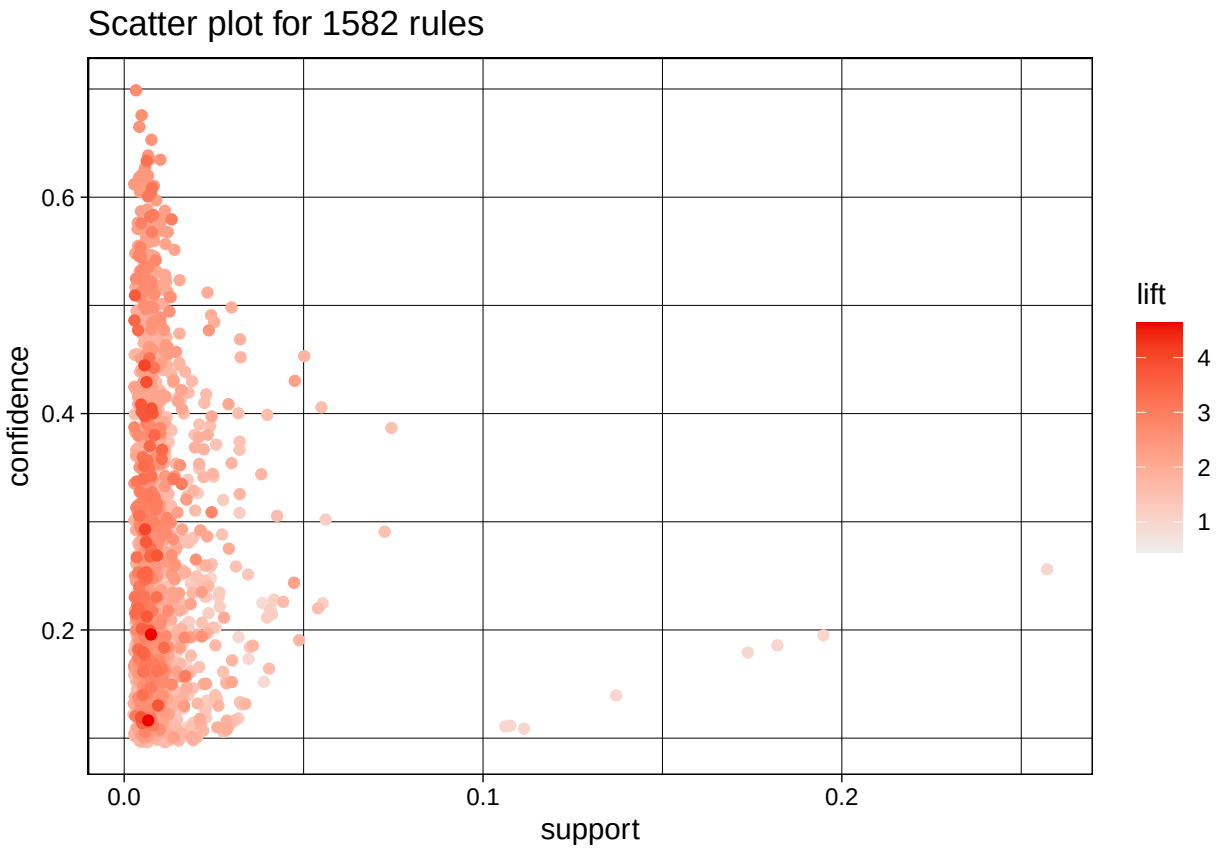
```
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[169 item(s), 9835 transaction(s)] done [0.00s].
## sorting and recoding items ... [120 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 4 done [0.00s].
## writing ... [1582 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

```
##      lhs                     rhs                      support confidence   coverage     lift count
## [1]  {ham}                => {white bread}        0.005083884  0.1953125 0.02602949 4.639851    50
## [2]  {white bread}        => {ham}                0.005083884  0.1207729 0.04209456 4.639851    50
## [3]  {citrus fruit,
##       other vegetables,
##       whole milk}         => {root vegetables}    0.005795628  0.4453125 0.01301474 4.085493    57
## [4]  {butter,
##       other vegetables}   => {whipped/sour cream} 0.005795628  0.2893401 0.02003050 4.036397    57
## [5]  {herbs}              => {root vegetables}    0.007015760  0.4312500 0.01626843 3.956477    69
## [6]  {other vegetables,
##       root vegetables}    => {onions}             0.005693950  0.1201717 0.04738180 3.875044    56
## [7]  {citrus fruit,
##       pip fruit}          => {tropical fruit}     0.005592272  0.4044118 0.01382816 3.854060    55
## [8]  {berries}            => {whipped/sour cream} 0.009049314  0.2721713 0.03324860 3.796886    89
## [9]  {whipped/sour cream} => {berries}            0.009049314  0.1262411 0.07168277 3.796886    89
## [10] {other vegetables,
##       tropical fruit,
##       whole milk}         => {root vegetables}    0.007015760  0.4107143 0.01708185 3.768074    69
```

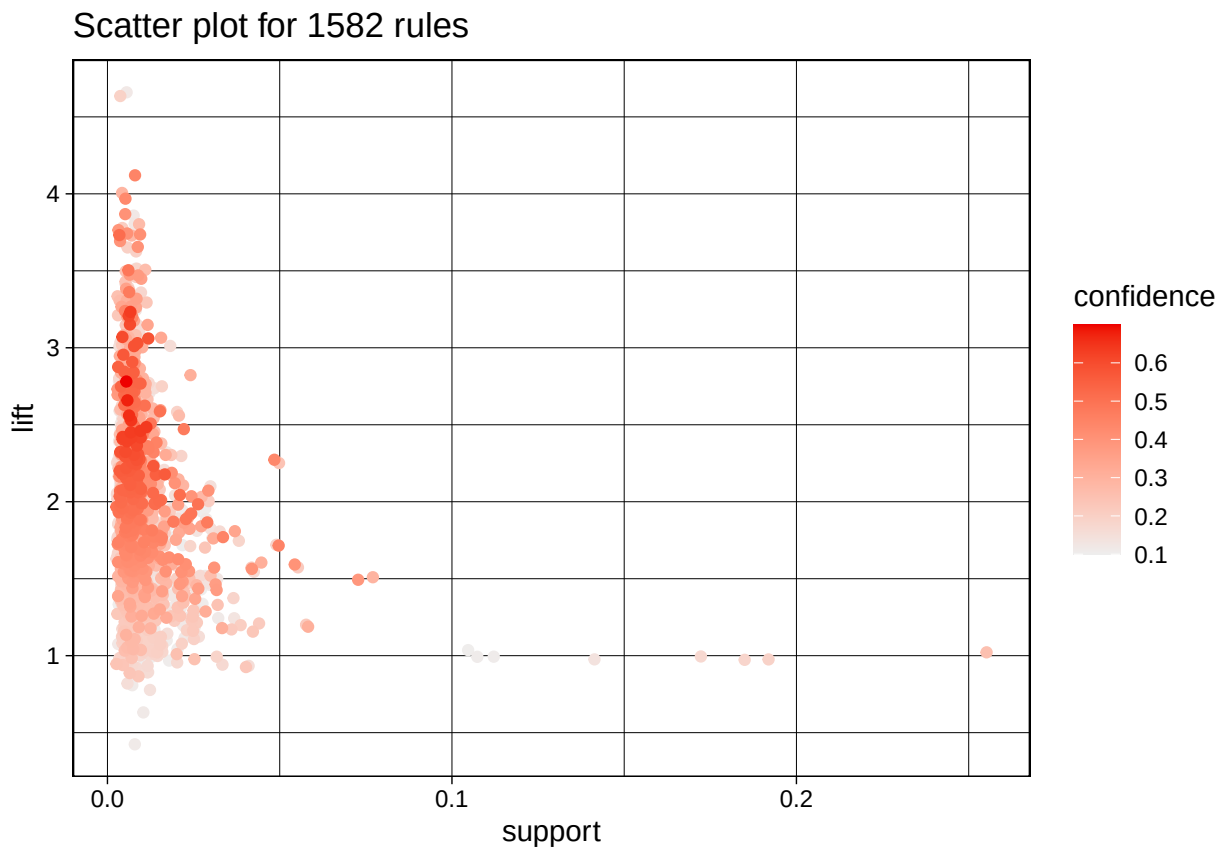
In the above table, I have taken a support of 0.005 and a confidence of 0.1.

Plotting the above rules

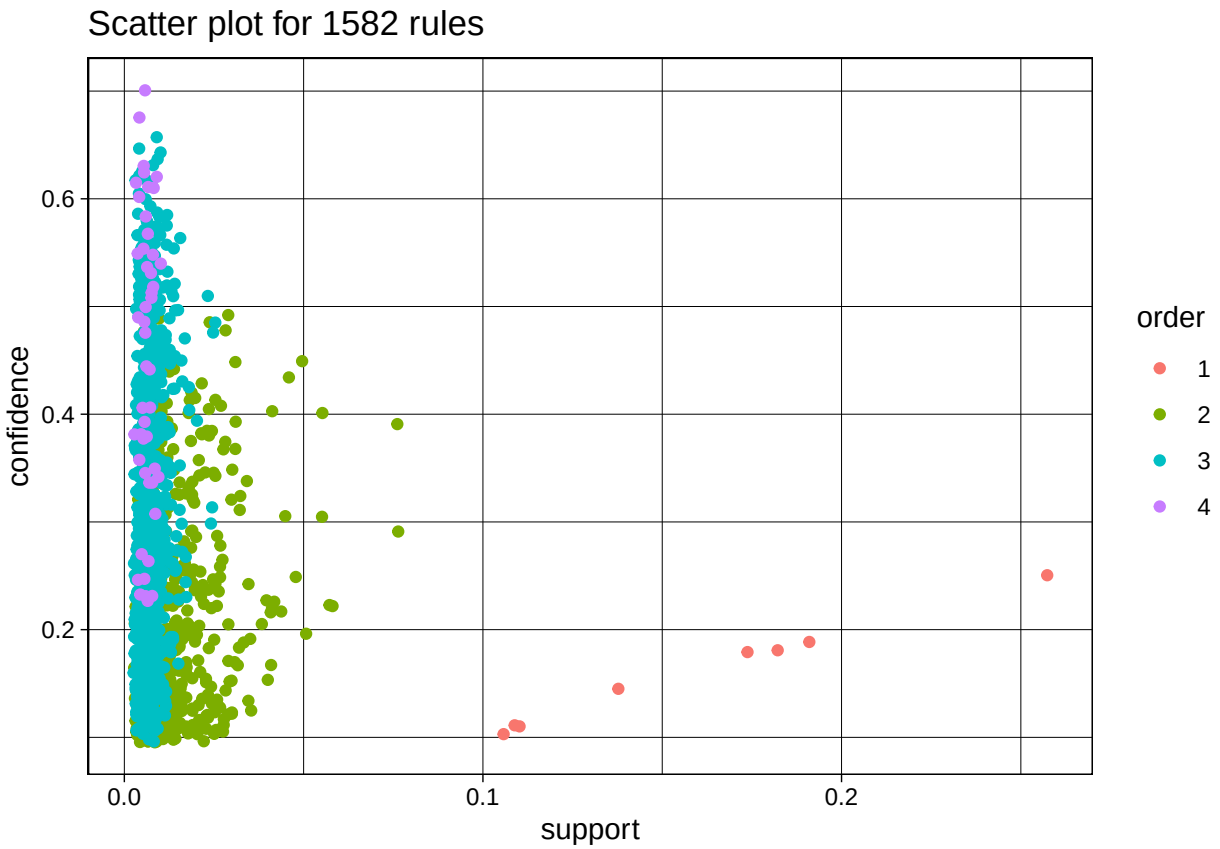
```
## To reduce overplotting, jitter is added! Use jitter = 0 to prevent jitter.
```




```
## To reduce overplotting, jitter is added! Use jitter = 0 to prevent jitter.
```




```
## To reduce overplotting, jitter is added! Use jitter = 0 to prevent jitter.
```


Scatter plot for 1582 rules
confidence
0.6
0.4
0.2
0.0
0.1
0.2
support
order
1
2
3
4

Items in LHS Groups
17 rules: {onions, oil, +15 items}
33 rules: {citrus fruit, other vegetables, +5 items}
38 rules: {soft cheese, cream cheese , +13 items}
39 rules: {grapes, pip fruit, +8 items}
6 rules: {ham}
45 rules: {yogurt, other vegetables, +1 items}
71 rules: {root vegetables, other vegetables, +1 items}
74 rules: {domestic eggs, tropical fruit, +3 items}
98 rules: {sliced cheese, hard cheese, +17 items}
41 rules: {herbs, oil, +11 items}
44 rules: {whipped/sour cream, whole milk, +1 items}
94 rules: {berries, onions, +9 items}
240 rules: {white bread, napkins, +13 items}
61 rules: {cake bar, mustard, +34 items}
29 rules: {UHT−milk, hygiene articles, +5 items}
40 rules: {butter milk, tropical fruit, +3 items}
197 rules: {pork, chicken, +8 items}
172 rules: {dessert, waffles, +12 items}
98 rules: {hamburger meat, baking powder, +27 items}
145 rules: {candy, canned beer, +28 items}
{ham}
{chicken}
RHS (+28 not shown)
support
0.007
0.009
0.011
lift
2.5 3.0 3.5 4.0 4.5

The first graph is a general plot of all the rules with support on the x axis and confidence on the y axis and lift being the indicator. We see that there is high lift for low support and slightlty low confidence values in general.

The second graph is a modification of the first where the y axis is replaced with lift instead of the confidence and the confidence is the indicator. We see that rules with high lift have a confidence of 40% or higher but a low support

The 3rd graph gives us the number of items in our rule. Most rules have 2 to 3 items

The 4th graph gives us a grouped data where the size of the circle represents the support and the color of the circle represents the lift

The 5th graph is a network chart showing the connections between the rules

Rule 2

Here we have a support of 0.01 and confidence of 0.5. This indicates that atleast 1% of all the transactions have that particular combination of products and out of all the antecedents at least 50% of them is a consequent. For example, if cereal is the antecedent, and milk is consequent, then of the orders containing cereals, 50% of them are likely to have whole milk. We have added a new parameter called minlen where the minimum length of the antecedents is 2 products.

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
```

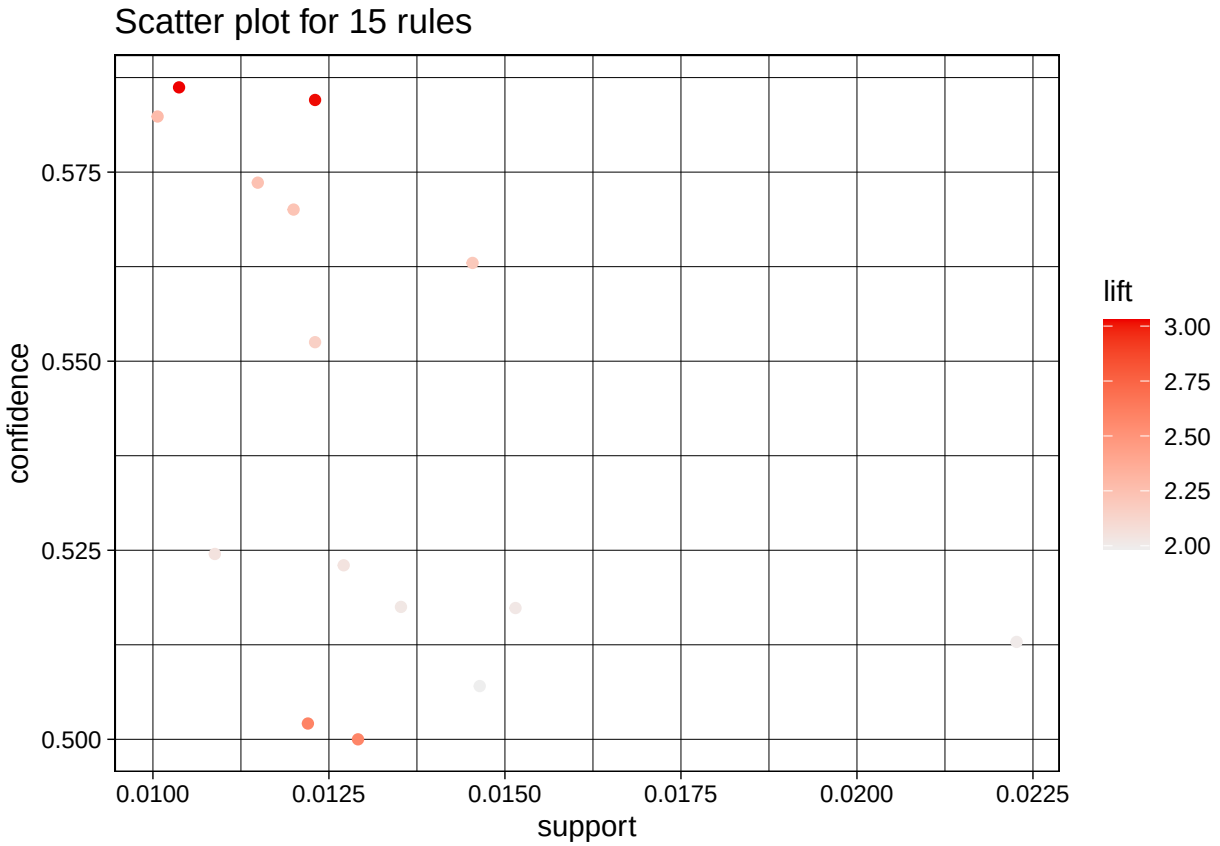
```
##         0.5    0.1    1 none FALSE            TRUE       5    0.01      2
##  maxlen target  ext
##      10  rules TRUE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 98
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[169 item(s), 9835 transaction(s)] done [0.00s].
## 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 [0.00s].
```

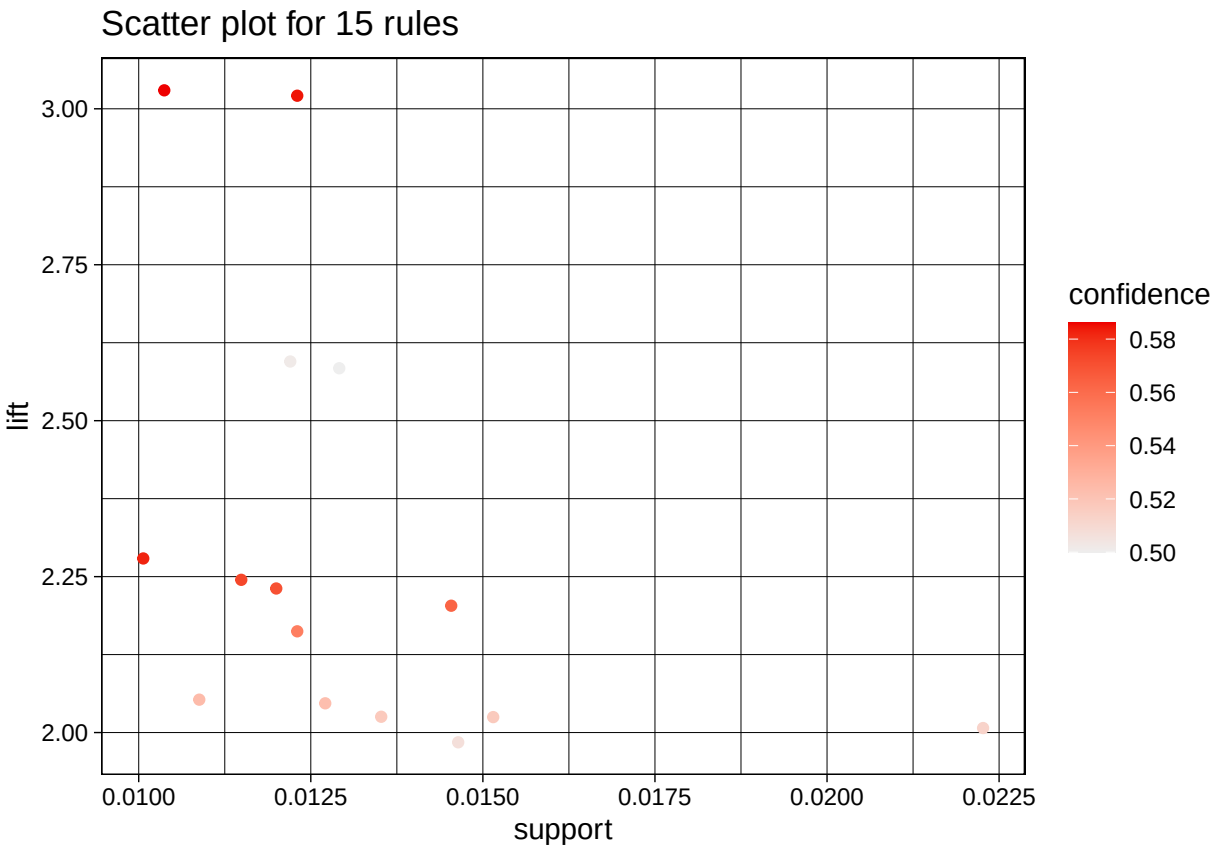
```
##      lhs                                    rhs                support
## [1]  {curd, yogurt}                      => {whole milk}       0.01006609
## [2]  {butter, other vegetables}          => {whole milk}       0.01148958
## [3]  {domestic eggs, other vegetables}   => {whole milk}       0.01230300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

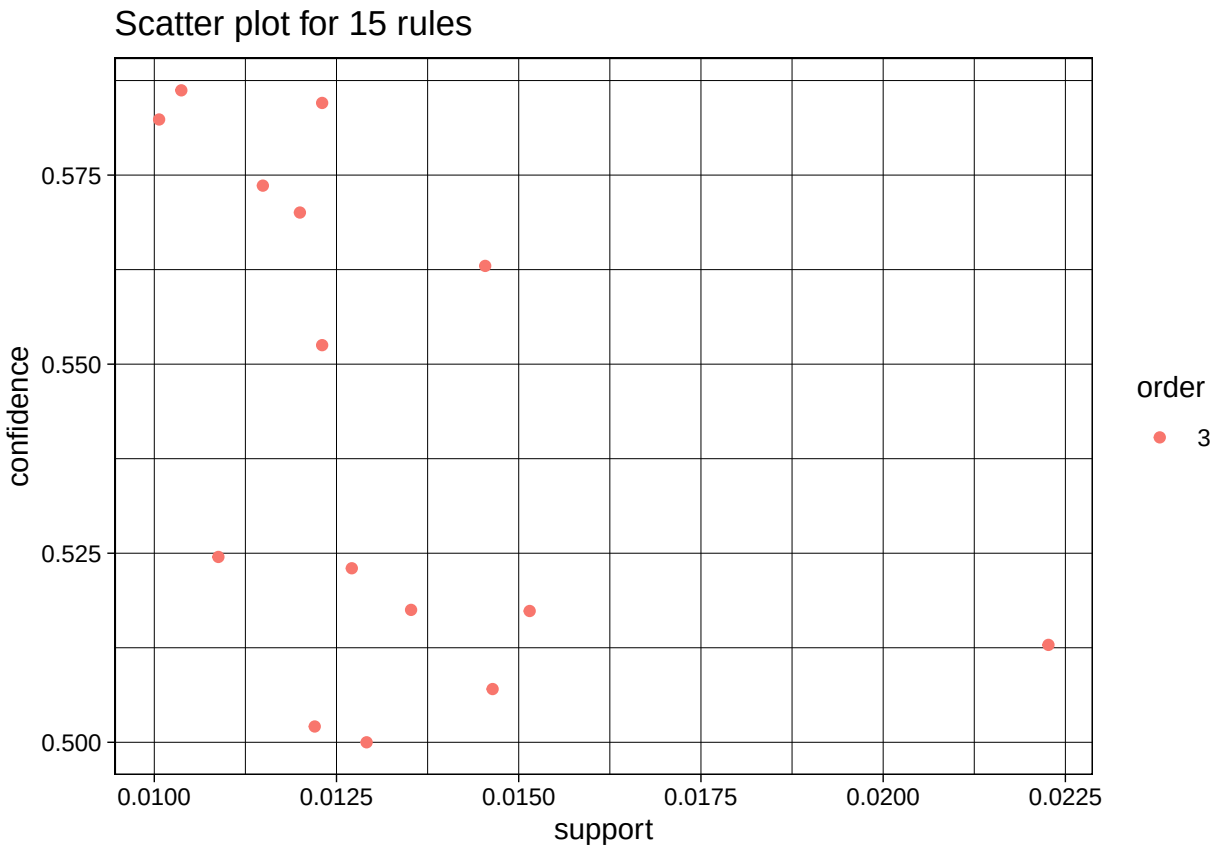
In the above table, we have a total of 15 rules. I have taken a support of 0.01 and a confidence of 0.5. we have a lift ranging from 2 to 3 indicating that customers buying the antecedent are highly likely to buy the

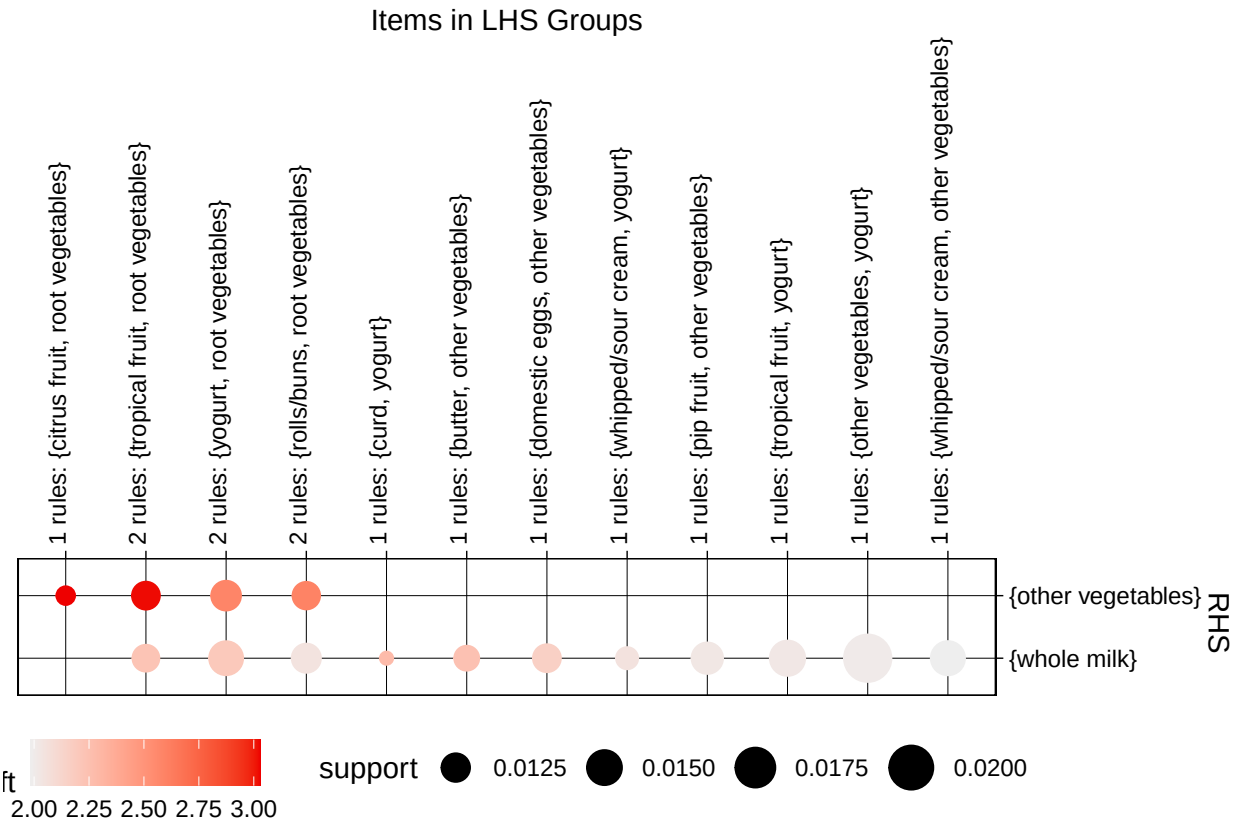
consequent. The confidence here is higher than 50%. We see that the consequent is whole milk and other vegetables which is a common and most frequent product.

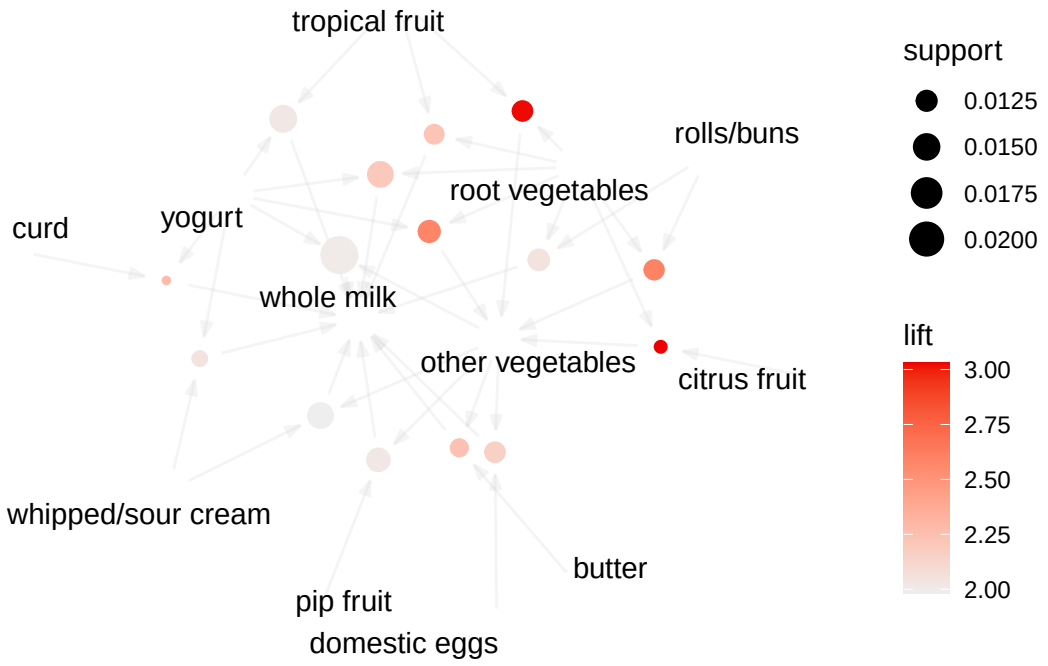
Plotting the above results we get the following:




Scatter plot for 15 rules
lift
3.00
2.75
2.50
2.25
2.00
0.0100
0.0125
0.0150
0.0175
0.0200
0.0225
support
confidence
0.58
0.56
0.54
0.52
0.50


Scatter plot for 15 rules
confidence
0.575
0.550
0.525
0.500
0.0100
0.0125
0.0150
0.0175
0.0200
0.0225
support
order
3


Items in LHS Groups
1 rules: {citrus fruit, root vegetables}
2 rules: {tropical fruit, root vegetables}
2 rules: {yogurt, root vegetables}
2 rules: {rolls/buns, root vegetables}
1 rules: {curd, yogurt}
1 rules: {butter, other vegetables}
1 rules: {domestic eggs, other vegetables}
1 rules: {whipped/sour cream, yogurt}
1 rules: {pip fruit, other vegetables}
1 rules: {tropical fruit, yogurt}
1 rules: {other vegetables, yogurt}
1 rules: {whipped/sour cream, other vegetables}
{other vegetables}
{whole milk}
RHS
ft
2.00 2.25 2.50 2.75 3.00
support
0.0125
0.0150
0.0175
0.0200



In the first graph we see that there is high lift for low support and high confidence values in general.

The second graph we see that rules with high lift have a confidence of 50% or higher but a low support

The 3rd graph gives us the number of items in our rule. All rules have 2 items

Focusing on the largest circle and the darkest circle, the 4th graph tells us that the customers who buy other vegetables and yogurt are more likely to buy whole milk whereas customers who buy tropical fruits and root vegetables are most likely to buy other vegetables.

The 5th graph is a network chart showing the connections between the rules which is similar to the 4th graph

Rule 3

Here we have a support of 0.005 and confidence of 0.1. This indicates that atleast 0.5% of all the transactions have that particular combination of products and out of all the antecedents at least 80% of them is a consequent. For example, if cereal is the antecedent, and milk is consequent, then of the orders containing cereals, 80% of them are likely to have whole milk. We have added a new parameter called maxlen where the maximum length of the antecedents is 5 products.

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##         0.8    0.1    1 none FALSE            TRUE       5   0.008      1
##  maxlen target  ext
```

```
##       5  rules TRUE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 78
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[169 item(s), 9835 transaction(s)] done [0.00s].
## sorting and recoding items ... [100 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 4 done [0.00s].
## writing ... [0 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

With the above criteria, we did not generate any results implying that no rule was 80% strong with a max length of 5 items in the basket.

Rule 4

Here we have a support of 0.008 and confidence of 0.5. This indicates that atleast 0.8% of all the transactions have that particular combination of products and out of all the antecedents at least 50% of them is a consequent. For example, if cereal is the antecedent, and milk is consequent, then of the orders containing cereals, 50% of them are likely to have whole milk. We have set the max length of the products to 5.

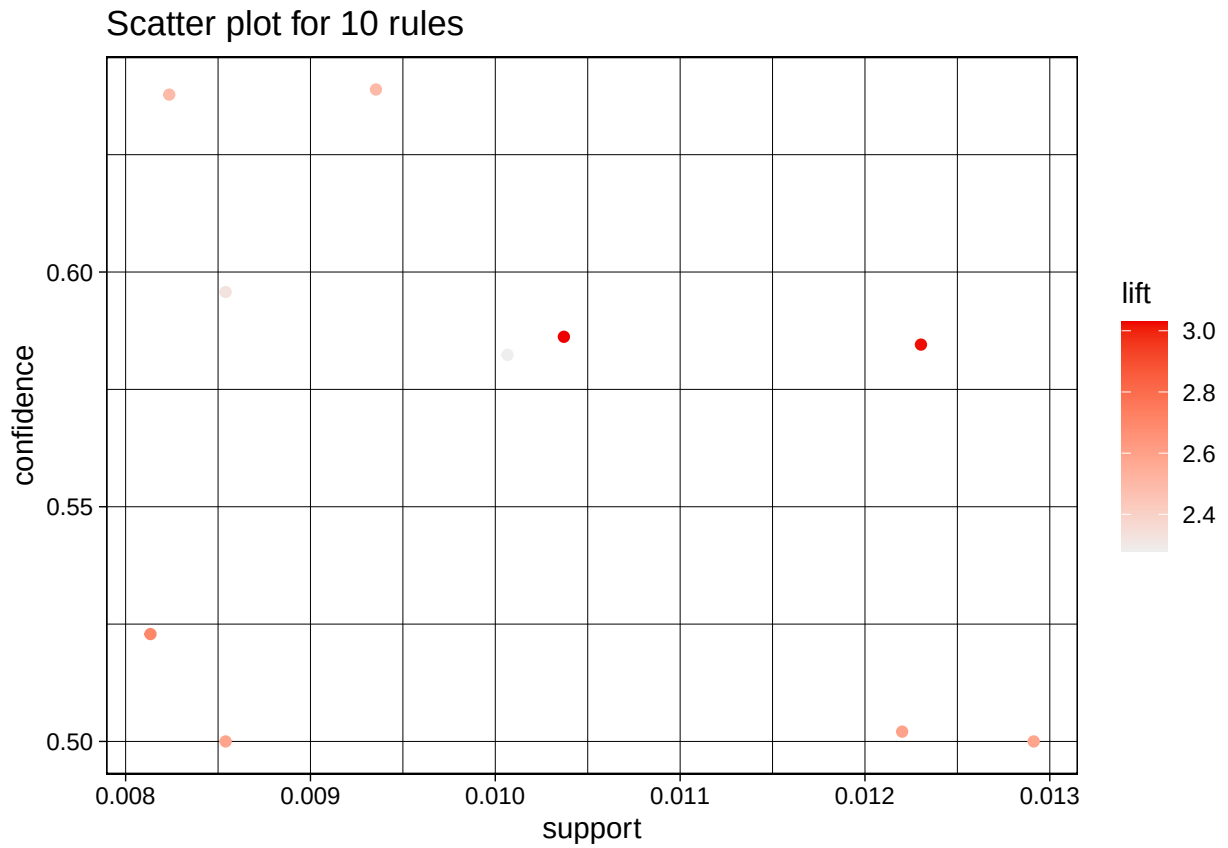
```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##         0.5    0.1    1 none FALSE            TRUE       5   0.008      1
##  maxlen target  ext
##       5  rules TRUE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 78
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[169 item(s), 9835 transaction(s)] done [0.00s].
## sorting and recoding items ... [100 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 4 done [0.00s].
## writing ... [30 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

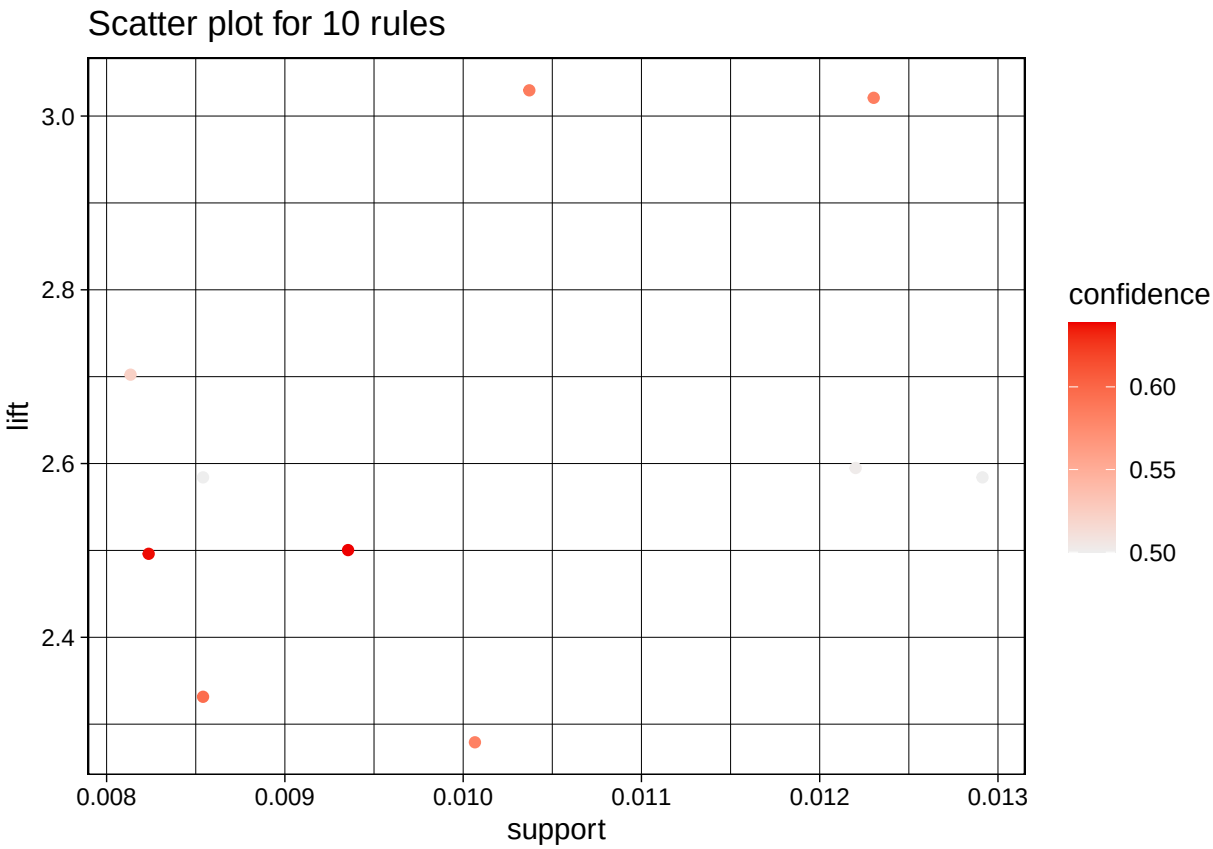
```
##      lhs                     rhs                    support confidence   coverage     lift count
## [1]  {citrus fruit,
##       root vegetables}    => {other vegetables} 0.010371124  0.5862069 0.01769192 3.029608   102
## [2]  {root vegetables,
```

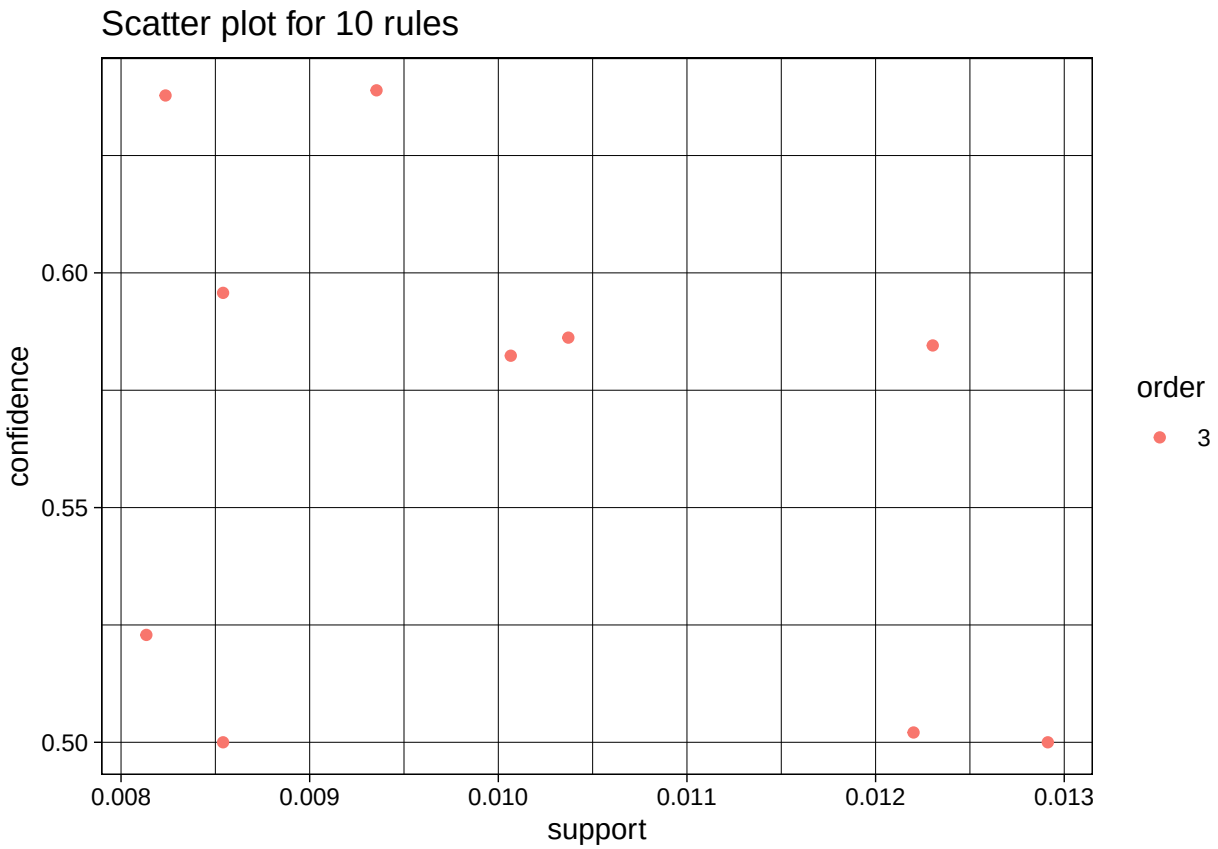
```
##       tropical fruit}     => {other vegetables} 0.012302999  0.5845411 0.02104728 3.020999   121
## [3]  {pip fruit,
##       root vegetables}    => {other vegetables} 0.008134215  0.5228758 0.01555669 2.702304    80
## [4]  {rolls/buns,
##       root vegetables}    => {other vegetables} 0.012201322  0.5020921 0.02430097 2.594890   120
## [5]  {root vegetables,
##       whipped/sour cream} => {other vegetables} 0.008540925  0.5000000 0.01708185 2.584078    84
## [6]  {root vegetables,
##       yogurt}             => {other vegetables} 0.012913066  0.5000000 0.02582613 2.584078   127
## [7]  {butter,
##       yogurt}             => {whole milk}       0.009354347  0.6388889 0.01464159 2.500387    92
## [8]  {butter,
##       root vegetables}    => {whole milk}       0.008235892  0.6377953 0.01291307 2.496107    81
## [9]  {domestic eggs,
##       root vegetables}    => {whole milk}       0.008540925  0.5957447 0.01433655 2.331536    84
## [10] {curd,
##       yogurt}             => {whole milk}       0.010066090  0.5823529 0.01728521 2.279125    99
```

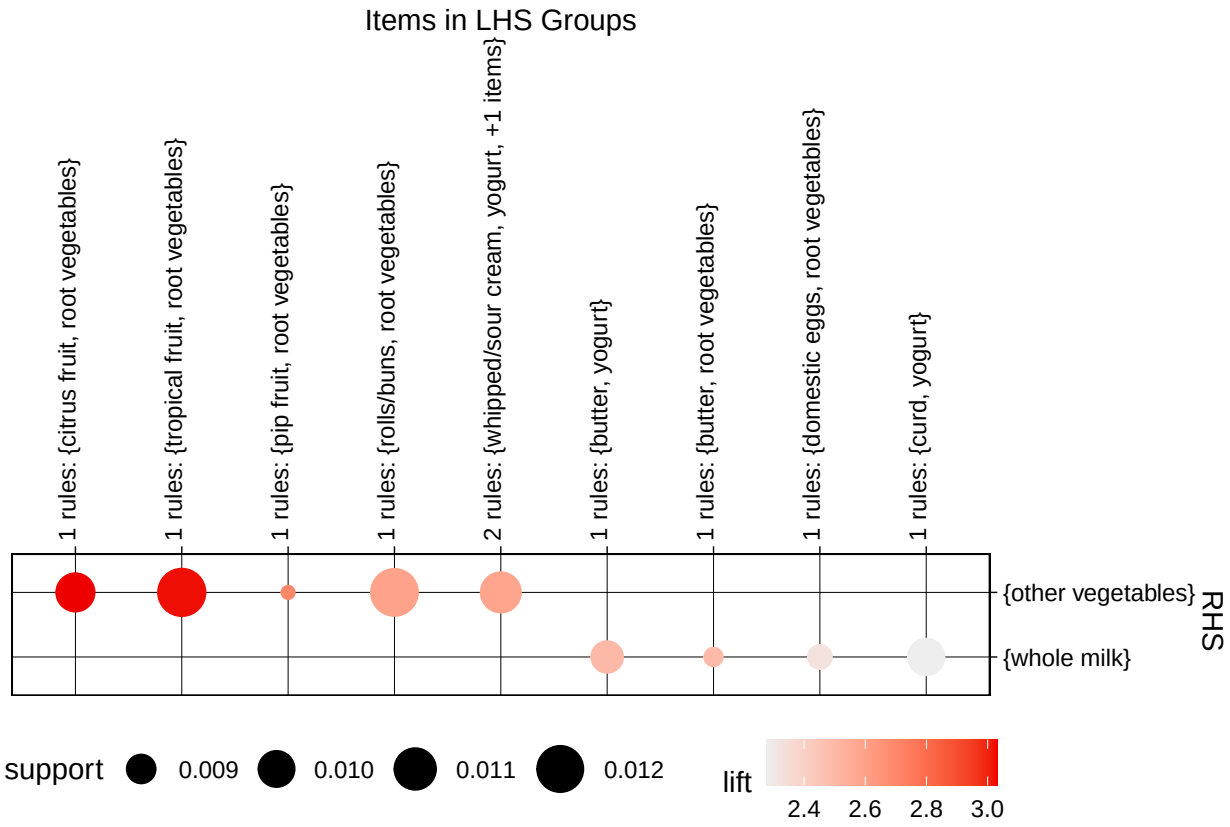
In the above set of rules, we have a total of 30 rules. We will be investigating the first 10 based on lift for ease of interpretation. I have taken a support of 0.008 and a confidence of 0.5. we have a lift ranging from 2 to 3 indicating that customers buying the antecedent are highly likely to buy the consequent. The confidence here is higher than 50%. We see that the consequent is whole milk and other vegetables which is a common and most frequent product.

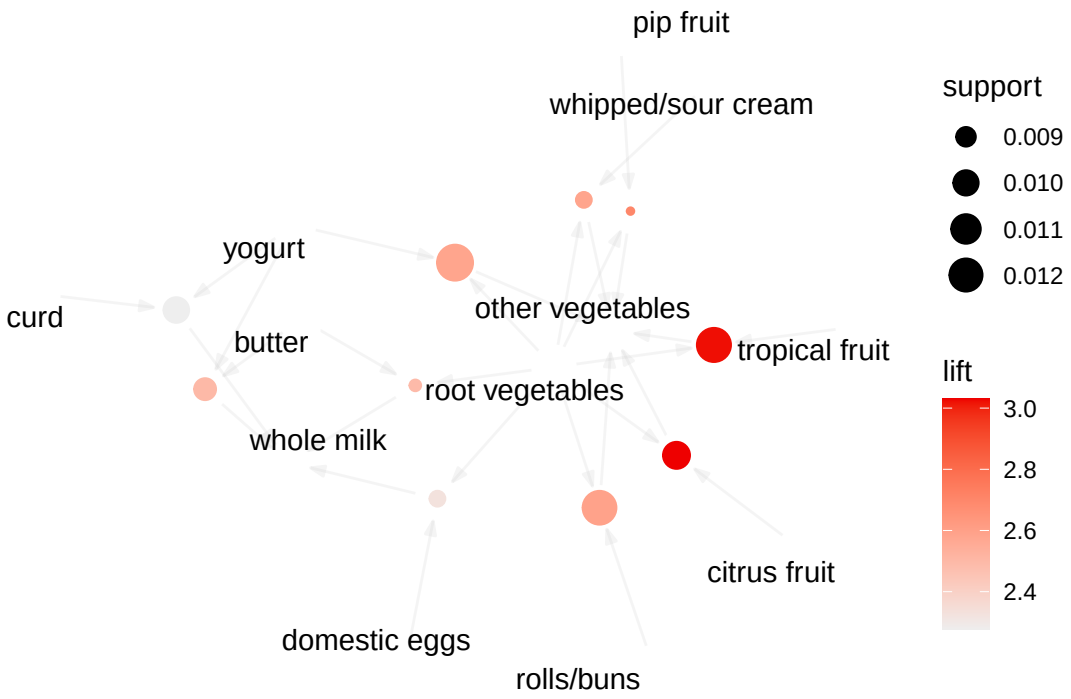
Plotting the above results we get the following:




Scatter plot for 10 rules
lift
3.0
2.8
2.6
2.4
0.008
0.009
0.010
0.011
0.012
0.013
support
confidence
0.60
0.55
0.50


Scatter plot for 10 rules
confidence
0.60
0.55
0.50
0.008
0.009
0.010
0.011
0.012
0.013
support
order
3


Items in LHS Groups
1 rules: {citrus fruit, root vegetables}
1 rules: {tropical fruit, root vegetables}
1 rules: {pip fruit, root vegetables}
1 rules: {rolls/buns, root vegetables}
2 rules: {whipped/sour cream, yogurt, +1 items}
1 rules: {butter, yogurt}
1 rules: {butter, root vegetables}
1 rules: {domestic eggs, root vegetables}
1 rules: {curd, yogurt}
{other vegetables}
{whole milk}
RHS
support
0.009
0.010
0.011
0.012
lift
2.4
2.6
2.8
3.0



In the first graph we see that there is high lift for avg to high support and avg confidence values in general.

The second graph we see that rules with low lift have a confidence of 55% or higher but a low support

The 3rd graph gives us the number of items in our rule. All rules have 3 items

Focusing on the largest circle and the darkest circle (which happen to be the same here), the 4th graph tells us that the customers who buy tropical fruits and root vegetables are most likely to buy other vegetables.

The 5th graph is a network chart showing the connections between the rules which is similar to the 4th graph.Here we see that all the vegetables are grouped closer to each other

Rule 5

Here we have a support of 0.005 and confidence of 0.6. This indicates that atleast 0.5% of all the transactions have that particular combination of products and out of all the antecedents at least 65% of them is a consequent. For example, if cereal is the antecedent, and milk is consequent, then of the orders containing cereals, 65 of them are likely to have whole milk. We have set the max length of the products to 5 and sorted the results by confidence.

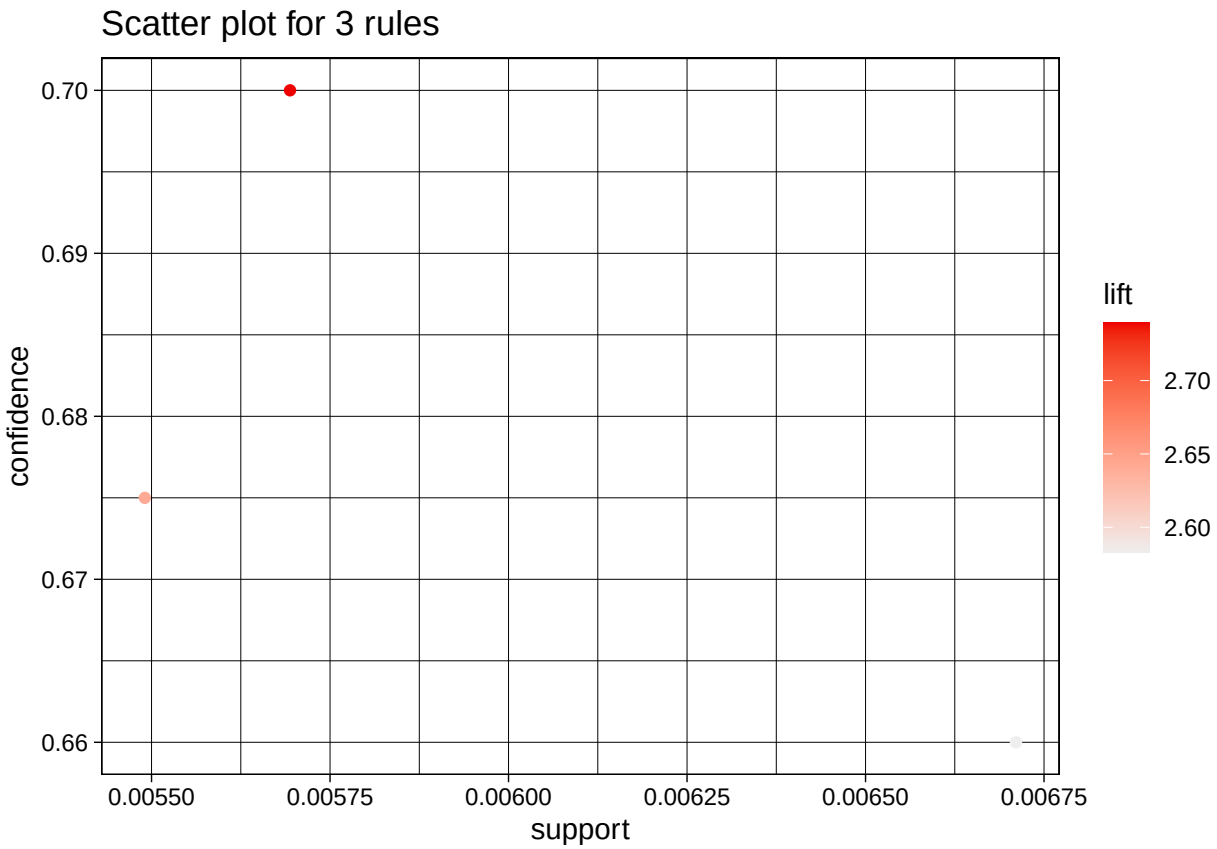
```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##        0.65    0.1    1 none FALSE            TRUE       5   0.005      1
##  maxlen target  ext
##       5  rules TRUE
```

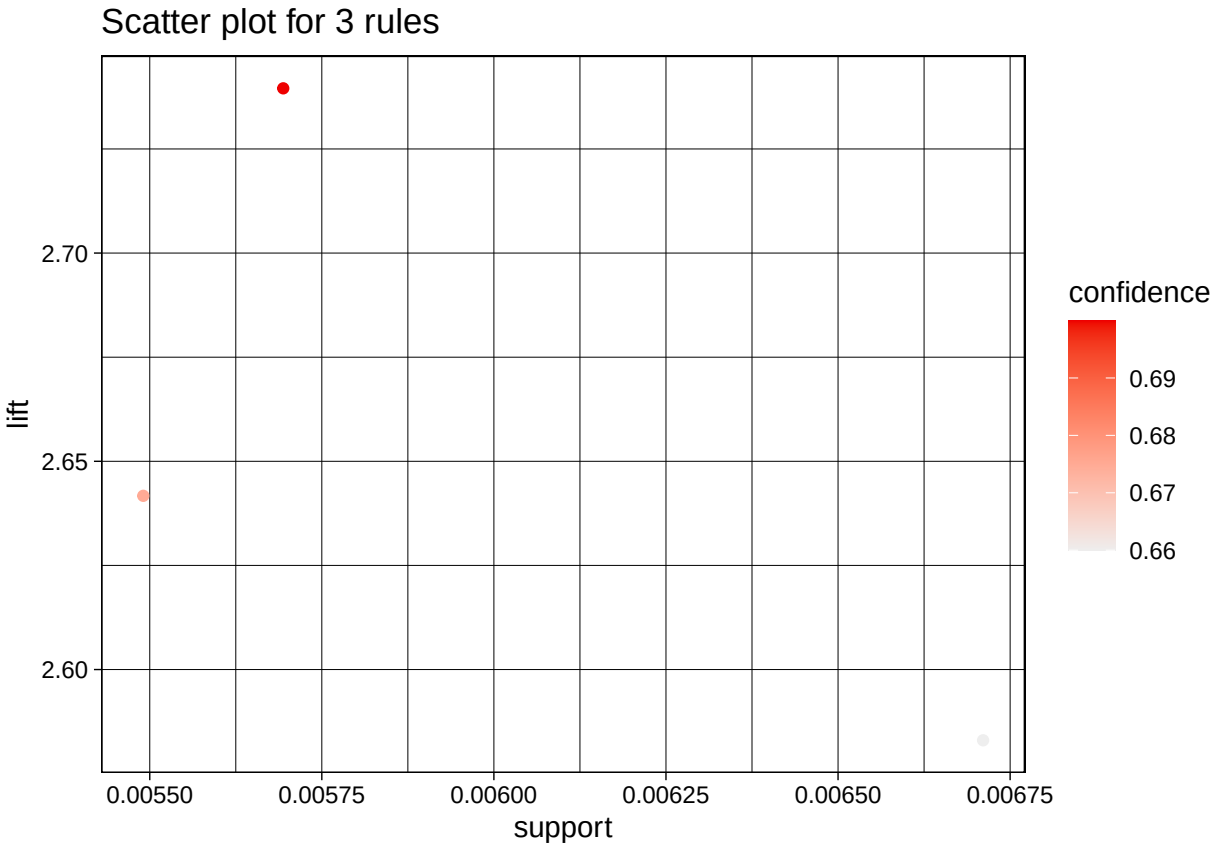
```
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 49
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[169 item(s), 9835 transaction(s)] done [0.00s].
## sorting and recoding items ... [120 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 4 done [0.00s].
## writing ... [3 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

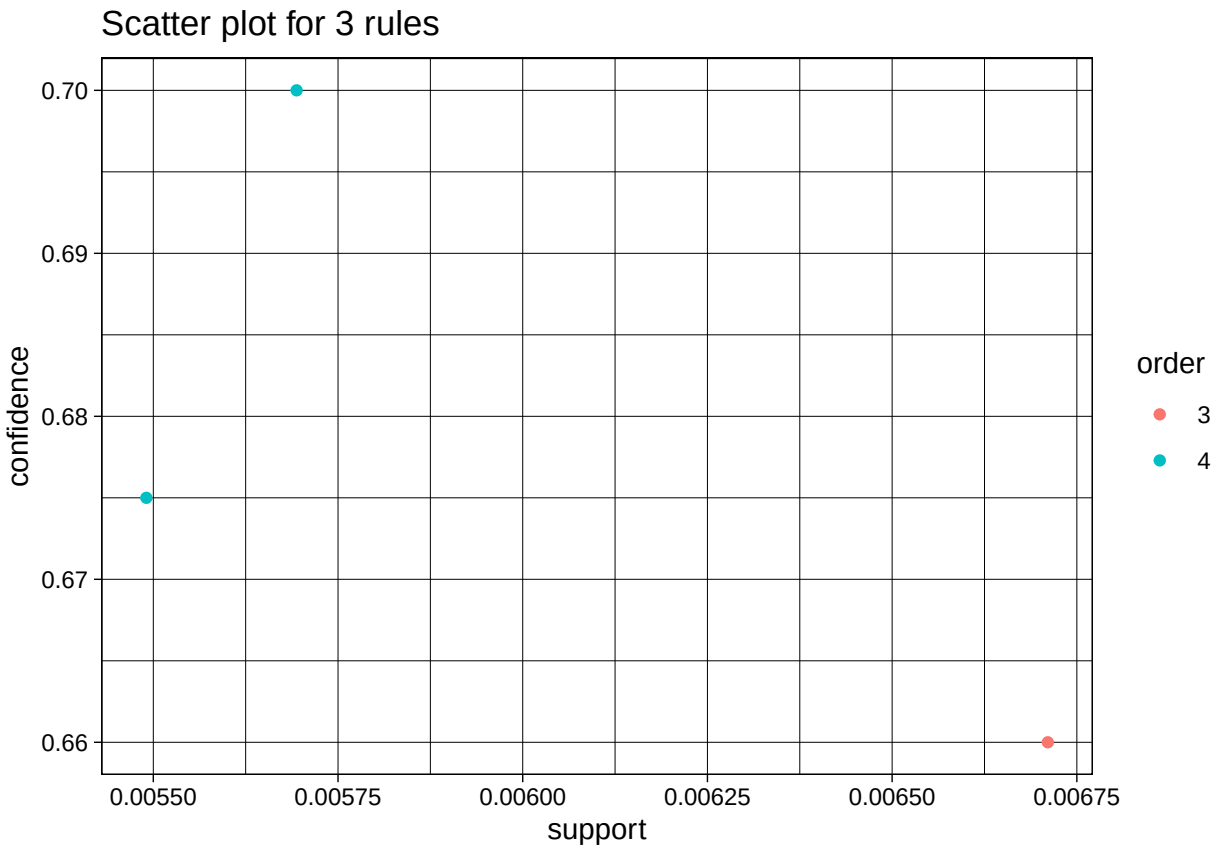
```
##     lhs                     rhs              support confidence    coverage     lift count
## [1] {root vegetables,
##      tropical fruit,
##      yogurt}             => {whole milk} 0.005693950      0.700 0.008134215 2.739554    56
## [2] {other vegetables,
##      pip fruit,
##      root vegetables}    => {whole milk} 0.005490595      0.675 0.008134215 2.641713    54
## [3] {butter,
##      whipped/sour cream} => {whole milk} 0.006710727      0.660 0.010167768 2.583008    66
```

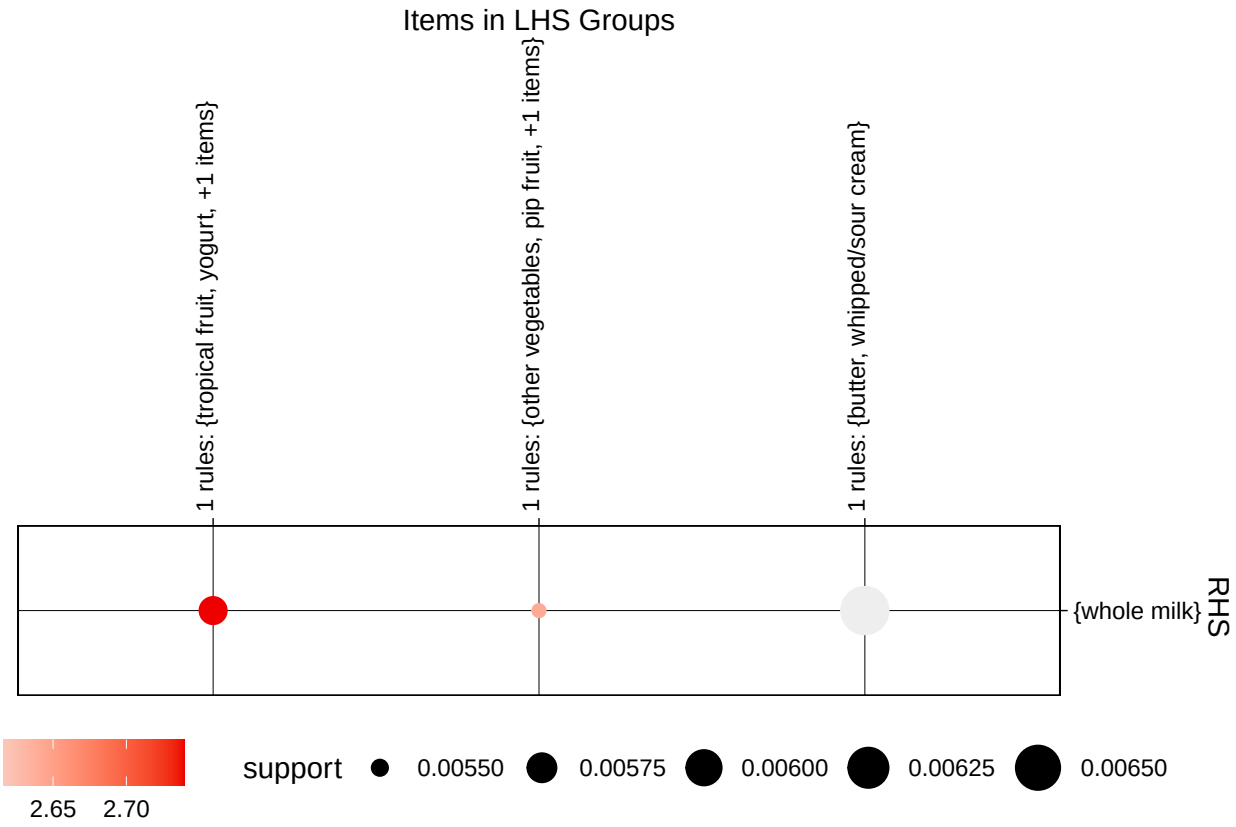
In the above set of rules, we have a total of 3 rules. I have taken a support of 0.008 and a confidence of 0.65. we have a lift ranging from 2.55 to 2.75 indicating that customers buying the antecedent are highly likely to buy the consequent. The confidence here is higher than 65%. We see that the consequent is whole milk.

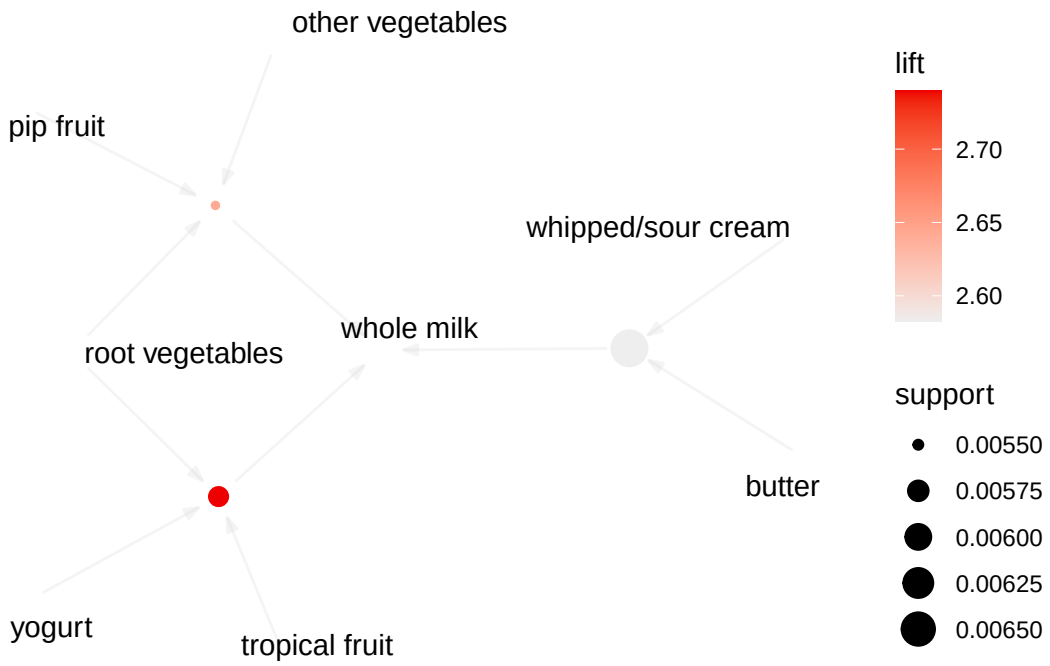
Plotting the above results we get the following:


Scatter plot for 3 rules
confidence
0.70
0.69
0.68
0.67
0.66
0.00550
0.00575
0.00600
0.00625
0.00650
0.00675
support
lift
2.70
2.65
2.60


Scatter plot for 3 rules
lift
2.70
2.65
2.60
0.00550
0.00575
0.00600
0.00625
0.00650
0.00675
support
confidence
0.69
0.68
0.67
0.66


Scatter plot for 3 rules
confidence
0.70
0.69
0.68
0.67
0.66
0.00550
0.00575
0.00600
0.00625
0.00650
0.00675
support
order
3
4


Items in LHS Groups
1 rules: {tropical fruit, yogurt, +1 items}
1 rules: {other vegetables, pip fruit, +1 items}
1 rules: {butter, whipped/sour cream}
{whole milk}
RHS
2.65
2.70
support
0.00550
0.00575
0.00600
0.00625
0.00650



In the first graph we see that there is high lift for a confidence value of 70%.

The second graph tells a similar story as the first where we see a high lift and confidence but low support

The 3rd graph gives us the number of items in our rule. 2 of the 3 rules have 4 items and one of them has 3 items

Focusing on the darkest circle, the 4th graph tells us that the customers who buy tropical fruits, yogurt and another item tend to buy whole milk

The 5th graph is a network chart showing the connections between the rules which is similar to the 4th graph.

Thus we have seen examples of 5 different cases of association rule mining. Based on our use case and requirements, these rules can be modified to our benefit as required.