# Question 8

2024-08-15

# Q8 ~ Association rule mining

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'arules'
```

```
## The following objects are masked from 'package:base':
##
##     abbreviate, write
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ ggplot2   3.5.1     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.1
## ✔ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ tidyr::expand() masks Matrix::expand()
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ✖ tidyr::pack()   masks Matrix::pack()
## ✖ dplyr::recode() masks arules::recode()
## ✖ tidyr::unpack() masks Matrix::unpack()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflic
ts to become errors
##
## Attaching package: 'igraph'
##
##
## The following objects are masked from 'package:lubridate':
##
##     %--%, union
##
##
## The following objects are masked from 'package:dplyr':
##
##     as_data_frame, groups, union
```

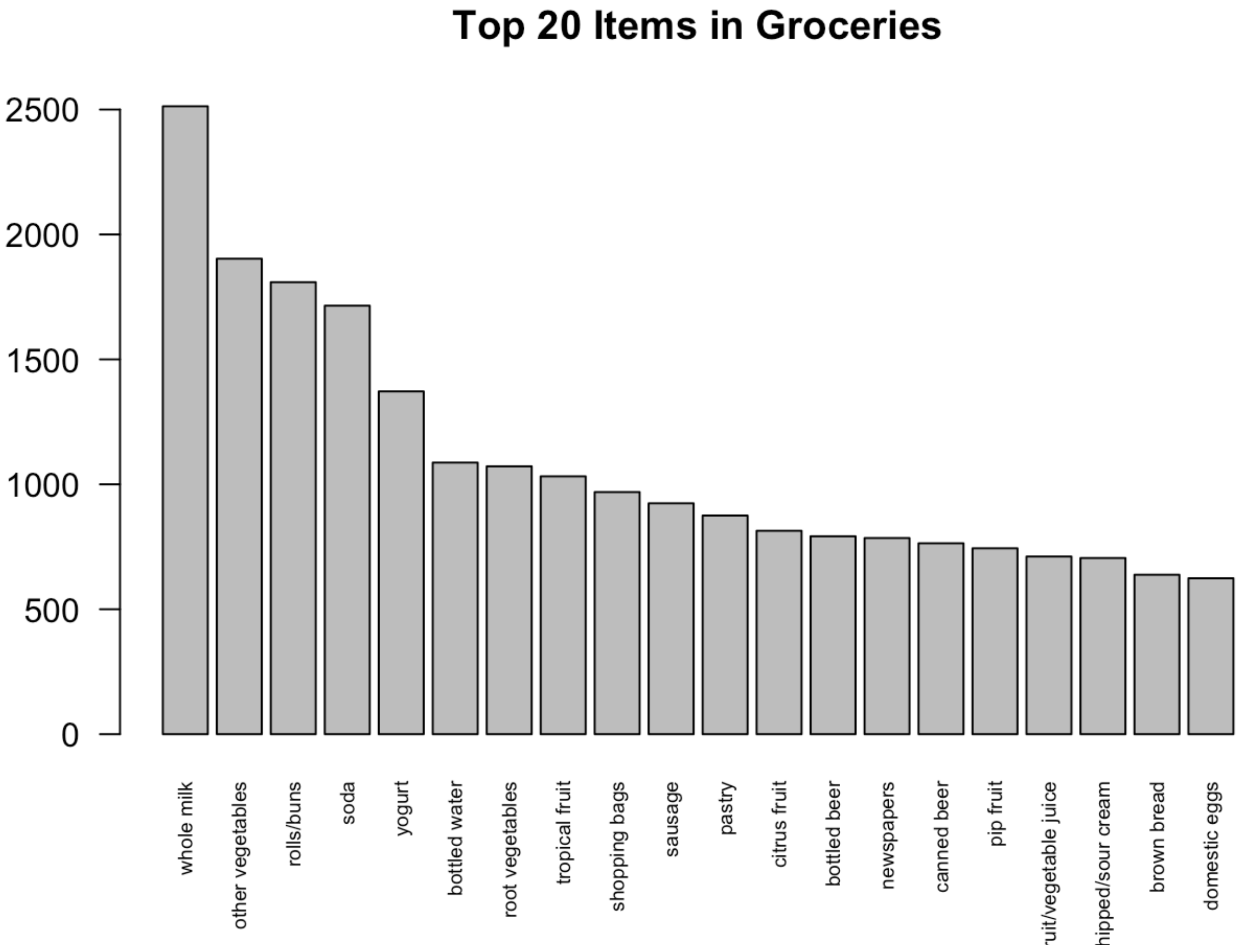
```
##
##
## The following objects are masked from 'package:purrr':
##
##     compose, simplify
##
##
## The following object is masked from 'package:tidyr':
##
##     crossing
##
##
## The following object is masked from 'package:tibble':
##
##     as_data_frame
##
##
## The following object is masked from 'package:arules':
##
##     union
##
##
## The following objects are masked from 'package:stats':
##
##     decompose, spectrum
##
##
## The following object is masked from 'package:base':
##
##     union
```

```
summary(groceries)
```

```
## transactions as itemMatrix in sparse format with
##  9835 rows (elements/itemsets/transactions) and
##  169 columns (items) and a density of 0.02609146 
## 
## most frequent items:
##       whole milk other vegetables       rolls/buns             soda 
##             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
```

```
item_freq <- itemFrequency(groceries, type = "absolute")

top_items <- sort(item_freq, decreasing = TRUE)[1:20]
barplot(top_items, las = 2, cex.names = 0.6, main = "Top 20 Items in Groceries")
```



### From the summary of groceries, we can see that the most frequent items bought are:

| Item | No. of times bought |
|---|---|
| Whole Milk | 2513 |
| Other Vegetables | 1903 |
| Rolls/buns | 1809 |
| Soda | 1715 |

# I have picked following thresholds:

## Support = 0.5 percent

To identify commonly purchased combinations of items, support value of 0.5% can be a good fit

## Confidence = 0.20

Confidence value should be a good mix of neither too small and neither too high as then it might not truly capture correlated baskets or may limit us to a very narrow scope. Therefore, confidence of 20% can be a good fit here

### MaxLength = 3

Most of the dishes have nearly 3 core ingredients so maxlength of 3 can be a good fit as we should have some correlation in the baskets of such size.

## Applying Apriori to find frequent item sets.

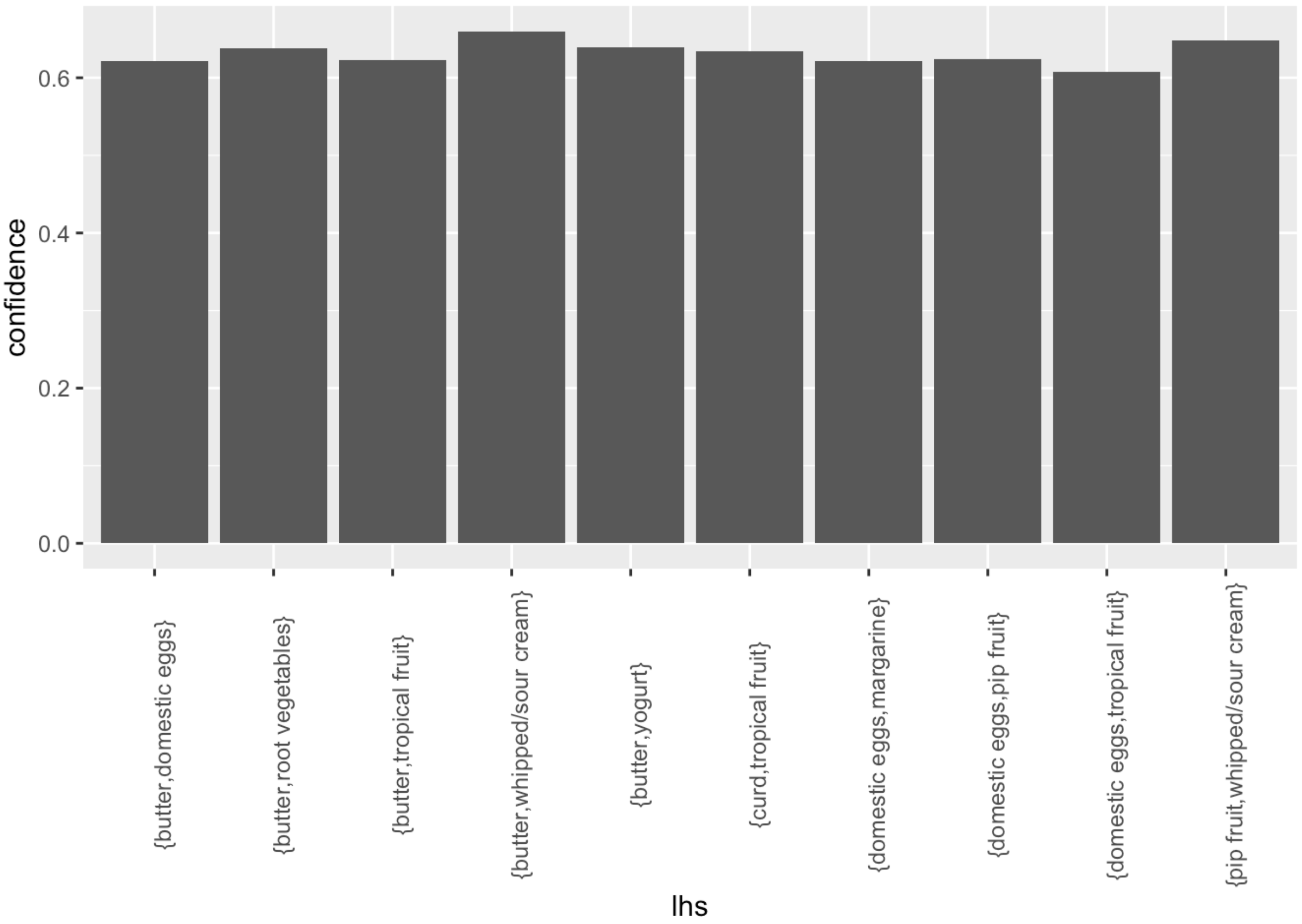
```
top_30_rules <- grocery_rule[1:30]
inspect(top_30_rules)
```

```
##      lhs                        rhs                 support     confidence
## [1]  {}                      => {whole milk}        0.255516014 0.2555160
## [2]  {cake bar}              => {whole milk}        0.005592272 0.4230769
## [3]  {dishes}                => {other vegetables}  0.005998983 0.3410405
## [4]  {dishes}                => {whole milk}        0.005287239 0.3005780
## [5]  {mustard}               => {whole milk}        0.005185562 0.4322034
## [6]  {pot plants}            => {whole milk}        0.006914082 0.4000000
## [7]  {chewing gum}           => {soda}              0.005388917 0.2560386
## [8]  {chewing gum}           => {whole milk}        0.005083884 0.2415459
## [9]  {canned fish}           => {other vegetables}  0.005083884 0.3378378
## [10] {pasta}                 => {whole milk}        0.006100661 0.4054054
## [11] {herbs}                 => {root vegetables}   0.007015760 0.4312500
## [12] {herbs}                 => {other vegetables}  0.007727504 0.4750000
## [13] {herbs}                 => {whole milk}        0.007727504 0.4750000
## [14] {processed cheese}      => {soda}              0.005287239 0.3190184
## [15] {processed cheese}      => {other vegetables}  0.005490595 0.3312883
## [16] {processed cheese}      => {whole milk}        0.007015760 0.4233129
## [17] {semi-finished bread}   => {other vegetables}  0.005185562 0.2931034
## [18] {semi-finished bread}   => {whole milk}        0.007117438 0.4022989
## [19] {beverages}             => {yogurt}            0.005490595 0.2109375
## [20] {beverages}             => {rolls/buns}        0.005388917 0.2070312
## [21] {beverages}             => {whole milk}        0.006812405 0.2617188
## [22] {ice cream}             => {soda}              0.006100661 0.2439024
## [23] {ice cream}             => {other vegetables}  0.005083884 0.2032520
## [24] {ice cream}             => {whole milk}        0.005897306 0.2357724
## [25] {detergent}             => {other vegetables}  0.006405694 0.3333333
## [26] {detergent}             => {whole milk}        0.008947636 0.4656085
## [27] {pickled vegetables}    => {other vegetables}  0.006405694 0.3579545
## [28] {pickled vegetables}    => {whole milk}        0.007117438 0.3977273
## [29] {baking powder}         => {other vegetables}  0.007320793 0.4137931
## [30] {baking powder}         => {whole milk}        0.009252669 0.5229885
##      coverage   lift      count
## [1]  1.00000000 1.0000000 2513
## [2]  0.01321810 1.6557746   55
## [3]  0.01759024 1.7625502   59
## [4]  0.01759024 1.1763569   52
## [5]  0.01199797 1.6914924   51
## [6]  0.01728521 1.5654596   68
## [7]  0.02104728 1.4683033   53
## [8]  0.02104728 0.9453259   50
## [9]  0.01504830 1.7459985   50
## [10] 0.01504830 1.5866145   60
## [11] 0.01626843 3.9564774   69
## [12] 0.01626843 2.4548739   76
## [13] 0.01626843 1.8589833   76
## [14] 0.01657346 1.8294729   52
## [15] 0.01657346 1.7121497   54
## [16] 0.01657346 1.6566981   69
## [17] 0.01769192 1.5148042   51
## [18] 0.01769192 1.5744565   70
## [19] 0.02602949 1.5120775   54
## [20] 0.02602949 1.1255679   53
## [21] 0.02602949 1.0242753   67
## [22] 0.02501271 1.3987058   60
## [23] 0.02501271 1.0504381   50
## [24] 0.02501271 0.9227303   58
## [25] 0.01921708 1.7227185   63
## [26] 0.01921708 1.8222281   88
## [27] 0.01789527 1.8499648   63
## [28] 0.01789527 1.5565650   70
## [29] 0.01769192 2.1385471   72
## [30] 0.01769192 2.0467935   91
```

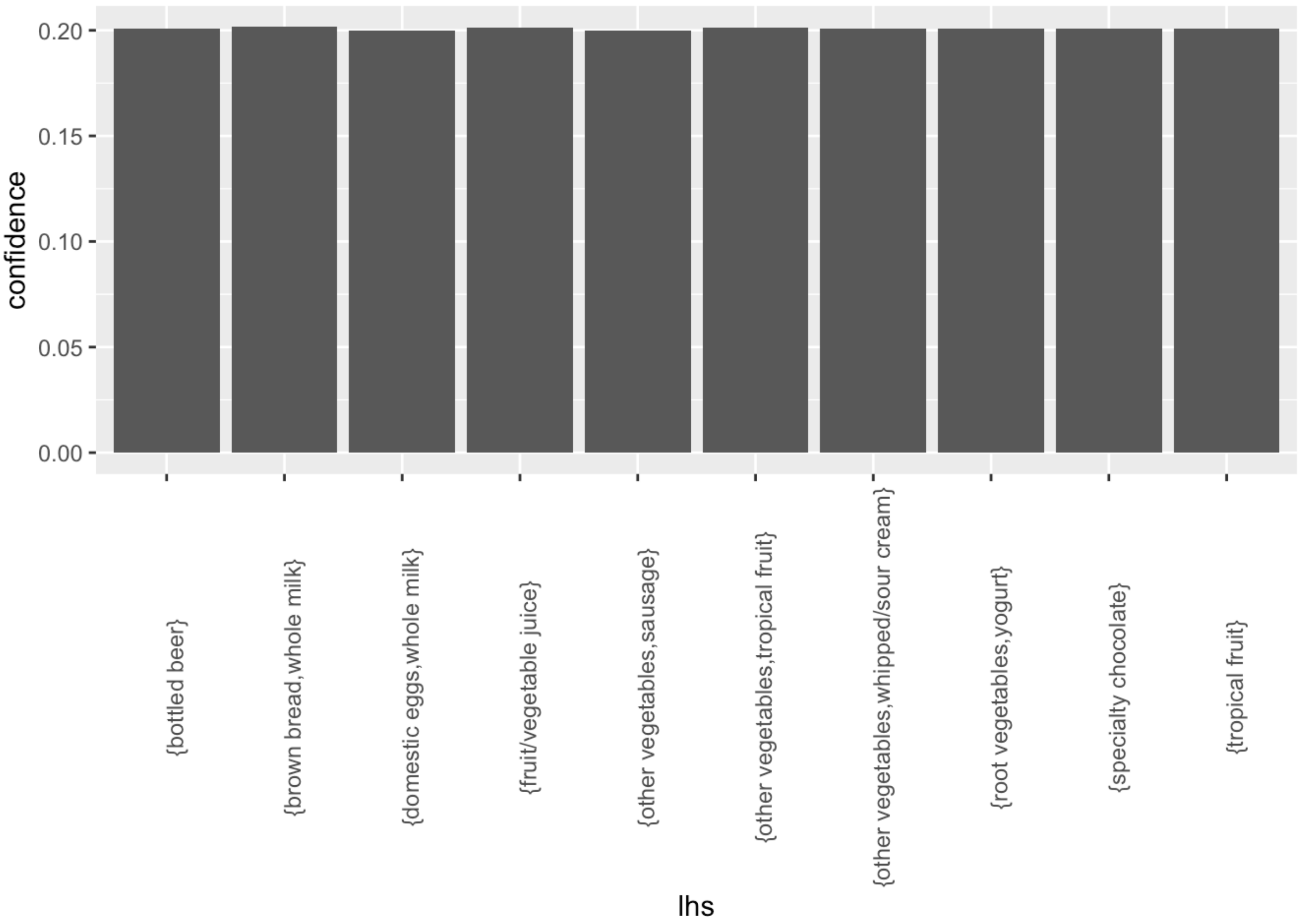
```
summary(grocery_rule)
```

```
## set of 825 rules
##
## rule length distribution (lhs + rhs):sizes
##   1   2   3
##   1 265 559
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   2.676   3.000   3.000
##
## summary of quality measures:
##     support           confidence        coverage             lift
##  Min.   :0.005084   Min.   :0.2000   Min.   :0.008338   Min.   :0.8991
##  1st Qu.:0.005897   1st Qu.:0.2500   1st Qu.:0.017692   1st Qu.:1.5270
##  Median :0.007321   Median :0.3112   Median :0.024606   Median :1.9050
##  Mean   :0.010499   Mean   :0.3399   Mean   :0.034455   Mean   :1.9666
##  3rd Qu.:0.010574   3rd Qu.:0.4170   3rd Qu.:0.034367   3rd Qu.:2.2875
##  Max.   :0.255516   Max.   :0.6600   Max.   :1.000000   Max.   :4.0364
##      count
##  Min.   :  50.0
##  1st Qu.:  58.0
##  Median :  72.0
##  Mean   : 103.3
##  3rd Qu.: 104.0
##  Max.   :2513.0
##
## mining info:
##       data ntransactions support confidence
##  groceries          9835   0.005        0.2
##
call
##  apriori(data = groceries, parameter = list(support = 0.005, confidence = 0.2, max
len = 3))
```

```
##                                  lhs            rhs     support confidence
## 1     {butter,whipped/sour cream} {whole milk} 0.006710727  0.6600000
## 2  {pip fruit,whipped/sour cream} {whole milk} 0.005998983  0.6483516
## 3                 {butter,yogurt} {whole milk} 0.009354347  0.6388889
## 4        {butter,root vegetables} {whole milk} 0.008235892  0.6377953
## 5           {curd,tropical fruit} {whole milk} 0.006507372  0.6336634
## 6       {domestic eggs,pip fruit} {whole milk} 0.005388917  0.6235294
## 7         {butter,tropical fruit} {whole milk} 0.006202339  0.6224490
## 8       {domestic eggs,margarine} {whole milk} 0.005185562  0.6219512
## 9          {butter,domestic eggs} {whole milk} 0.005998983  0.6210526
## 10 {domestic eggs,tropical fruit} {whole milk} 0.006914082  0.6071429
##       coverage     lift count
## 1  0.010167768 2.583008    66
## 2  0.009252669 2.537421    59
## 3  0.014641586 2.500387    92
## 4  0.012913066 2.496107    81
## 5  0.010269446 2.479936    64
## 6  0.008642603 2.440275    53
## 7  0.009964413 2.436047    61
## 8  0.008337570 2.434099    51
## 9  0.009659380 2.430582    59
## 10 0.011387900 2.376144    68
```


confidence
0.6
0.4
0.2
0.0
{butter,domestic eggs}
{butter,root vegetables}
{butter,tropical fruit}
{butter,whipped/sour cream}
{butter,yogurt}
{curd,tropical fruit}
{domestic eggs,margarine}
{domestic eggs,pip fruit}
{domestic eggs,tropical fruit}
{pip fruit,whipped/sour cream}
lhs

```
##                                       lhs                rhs     support
## 1               {brown bread,whole milk}             {soda} 0.005083884
## 2      {other vegetables,tropical fruit}             {soda} 0.007219115
## 3                {fruit/vegetable juice}       {rolls/buns} 0.014539908
## 4               {root vegetables,yogurt}          {sausage} 0.005185562
## 5                         {bottled beer} {other vegetables} 0.016166751
## 6  {other vegetables,whipped/sour cream}           {butter} 0.005795628
## 7                  {specialty chocolate} {other vegetables} 0.006100661
## 8                       {tropical fruit}  {root vegetables} 0.021047280
## 9             {domestic eggs,whole milk}           {butter} 0.005998983
## 10            {other vegetables,sausage}    {shopping bags} 0.005388917
##    confidence   coverage     lift count
## 1   0.2016129 0.02521607 1.156188    50
## 2   0.2011331 0.03589222 1.153437    71
## 3   0.2011252 0.07229283 1.093458   143
## 4   0.2007874 0.02582613 2.137169    51
## 5   0.2007576 0.08052872 1.037546   159
## 6   0.2007042 0.02887646 3.621883    57
## 7   0.2006689 0.03040163 1.037088    60
## 8   0.2005814 0.10493137 1.840222   207
## 9   0.2000000 0.02999492 3.609174    59
## 10  0.2000000 0.02694459 2.029928    53
```



The most frequent individual items is whole milk and have high support values, indicating its prevalence in the transactions. The lift value of 1.0 indicates that whole milk is purchased independently, as its presence or absence doesn’t influence each other.

**{butter,yogurt} => {whole milk}:** Customers buying “butter” and “yogurt” are likely to buy “whole milk” with a confidence of 63.88%, suggesting a common combination, while the support for {whole milk} individually is just 25.5%

Milk shows up the most on rhs with high confidence meaning shoppers will tend to buy milk the most and with many items

*Related Categories:*

**{butter,whipped/sour cream}=> {whole milk}:** Customers buying “butter” and “whipped/sour cream” are likely to buy “whole milk” with a confidence of 66%. This suggests an association between dairy products.

*Unrelated Categories:*

Rules like **{brown bread,whole milk}=> {soda}** and **{bottled beer} => {other vegetables}** shows that customers buying brown bread/whole milk and bottled beer might not buy soda and other vegetables respectively

**Q) Pick your own thresholds for lift and confidence; just be clear what these thresholds are and say why you picked them**

- **lift > 3.5 :** A high lift value indicates a strong association between items. Generally, a lift value greater than 1 signifies a positive association. Choosing a threshold like 3.5 filters out rules that are significantly stronger than random chance.
- **confidence > 0.6 :** By setting a confidence threshold of 60%, we can filter out weaker associations and focus on the more significant and reliable relationships between items.
- **lift > 3 & confidence > 0.6 :** This subset is chosen based on two criteria: rules with a “lift” value greater than 3, indicating strong associations, and rules with a “confidence” value higher than 0.6, indicating reliable predictions. By applying these filters, we can extract rules that not only represent strong connections between items but also provide dependable insights. This approach is useful for uncovering significant and practically relevant patterns in transaction data.

```
inspect(subset(grocery_rule, subset=lift > 3.5))
```

```
##      lhs                                     rhs                   support
## [1]  {herbs}                              => {root vegetables}    0.007015760
## [2]  {berries}                            => {whipped/sour cream} 0.009049314
## [3]  {onions, other vegetables}           => {root vegetables}    0.005693950
## [4]  {beef, other vegetables}             => {root vegetables}    0.007930859
## [5]  {curd, tropical fruit}               => {yogurt}             0.005287239
## [6]  {domestic eggs, whole milk}          => {butter}             0.005998983
## [7]  {butter, other vegetables}           => {whipped/sour cream} 0.005795628
## [8]  {other vegetables, whipped/sour cream} => {butter}           0.005795628
## [9]  {whipped/sour cream, whole milk}     => {butter}             0.006710727
## [10] {citrus fruit, pip fruit}            => {tropical fruit}     0.005592272
## [11] {citrus fruit, tropical fruit}       => {pip fruit}          0.005592272
##      confidence coverage   lift     count
## [1]  0.4312500  0.01626843 3.956477 69
## [2]  0.2721713  0.03324860 3.796886 89
## [3]  0.4000000  0.01423488 3.669776 56
## [4]  0.4020619  0.01972547 3.688692 78
## [5]  0.5148515  0.01026945 3.690645 52
## [6]  0.2000000  0.02999492 3.609174 59
## [7]  0.2893401  0.02003050 4.036397 57
## [8]  0.2007042  0.02887646 3.621883 57
## [9]  0.2082019  0.03223183 3.757185 66
## [10] 0.4044118  0.01382816 3.854060 55
## [11] 0.2806122  0.01992883 3.709437 55
```

**{herbs} => {root vegetables}:** Customers who buy both “herbs” are highly likely to also purchase “root vegetables” The high lift value of 3.95 indicates a strong association between these green food items.

```
inspect(subset(grocery_rule, subset=confidence > 0.6))
```

```
##      lhs                                 rhs                support
## [1]  {onions, root vegetables}       => {other vegetables} 0.005693950
## [2]  {curd, tropical fruit}          => {whole milk}       0.006507372
## [3]  {domestic eggs, margarine}      => {whole milk}       0.005185562
## [4]  {butter, domestic eggs}         => {whole milk}       0.005998983
## [5]  {butter, whipped/sour cream}    => {whole milk}       0.006710727
## [6]  {bottled water, butter}         => {whole milk}       0.005388917
## [7]  {butter, tropical fruit}        => {whole milk}       0.006202339
## [8]  {butter, root vegetables}       => {whole milk}       0.008235892
## [9]  {butter, yogurt}                => {whole milk}       0.009354347
## [10] {domestic eggs, pip fruit}      => {whole milk}       0.005388917
## [11] {domestic eggs, tropical fruit} => {whole milk}       0.006914082
## [12] {pip fruit, whipped/sour cream} => {other vegetables} 0.005592272
## [13] {pip fruit, whipped/sour cream} => {whole milk}       0.005998983
##      confidence coverage    lift     count
## [1]  0.6021505  0.009456024 3.112008 56
## [2]  0.6336634  0.010269446 2.479936 64
## [3]  0.6219512  0.008337570 2.434099 51
## [4]  0.6210526  0.009659380 2.430582 59
## [5]  0.6600000  0.010167768 2.583008 66
## [6]  0.6022727  0.008947636 2.357084 53
## [7]  0.6224490  0.009964413 2.436047 61
## [8]  0.6377953  0.012913066 2.496107 81
## [9]  0.6388889  0.014641586 2.500387 92
## [10] 0.6235294  0.008642603 2.440275 53
## [11] 0.6071429  0.011387900 2.376144 68
## [12] 0.6043956  0.009252669 3.123610 55
## [13] 0.6483516  0.009252669 2.537421 59
```

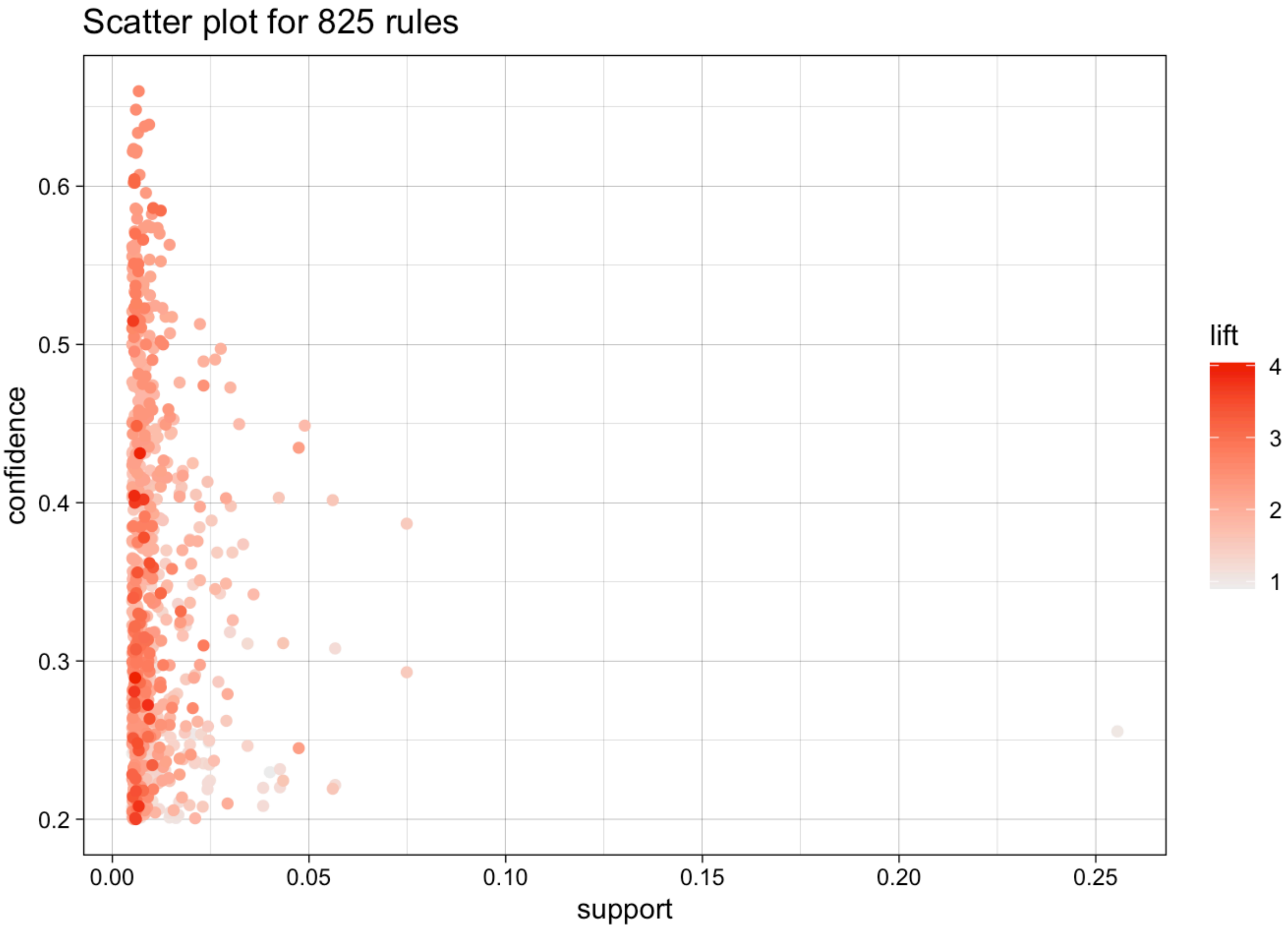
**{whipped/sour cream, butter} => {whole milk}:** Similarly, customers who buy “whipped/sour cream” and “butter” have a confidence of 60.2% to buy “whole milk” These associations might be due to dairy products

```
inspect(subset(grocery_rule, subset=lift > 3 & confidence > 0.6))
```

```
##     lhs                                 rhs                support
## [1] {onions, root vegetables}       => {other vegetables} 0.005693950
## [2] {pip fruit, whipped/sour cream} => {other vegetables} 0.005592272
##     confidence coverage    lift     count
## [1] 0.6021505  0.009456024 3.112008 56
## [2] 0.6043956  0.009252669 3.123610 55
```

**{onions, root vegetables} => {other vegetables }:** With a confidence of 60.2% and a lift of 3.11, customers who buy “onions” and “root vegetables” are likely to purchase “other vegetables” as well. This rule could represent common vegetable cooking dishes.

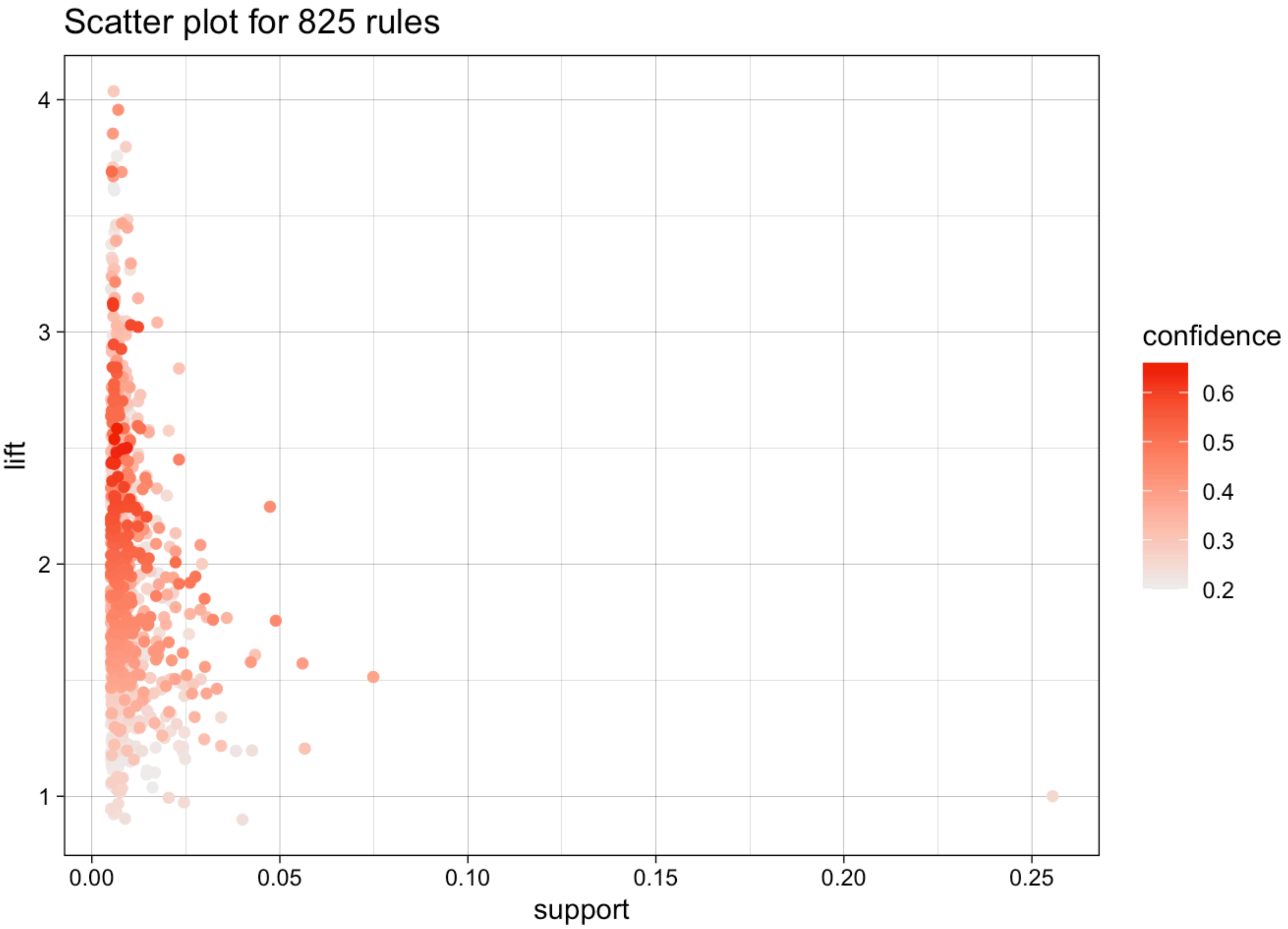
```
plot(grocery_rule, jitter =0)
```



- It can be observed that high lift rules tend to have low support. High lift rules with low support can provide valuable insights about specific interactions between items that might not be immediately obvious from looking at high-support items.

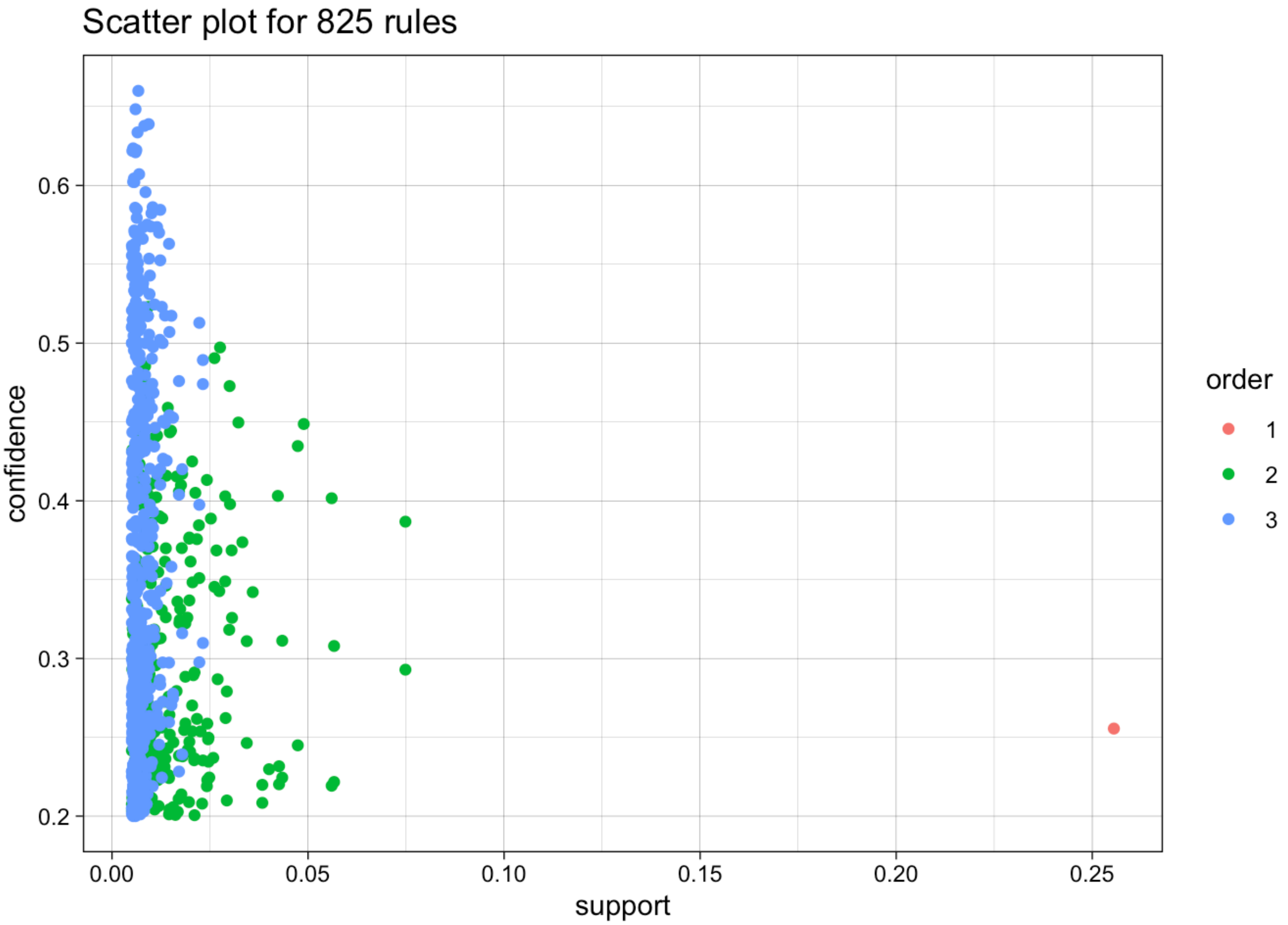
Swapping the axes and color scales:

```
plot(grocery_rule, jitter = 0, measure = c("support", "lift"), shading = "confidenc
e")
```



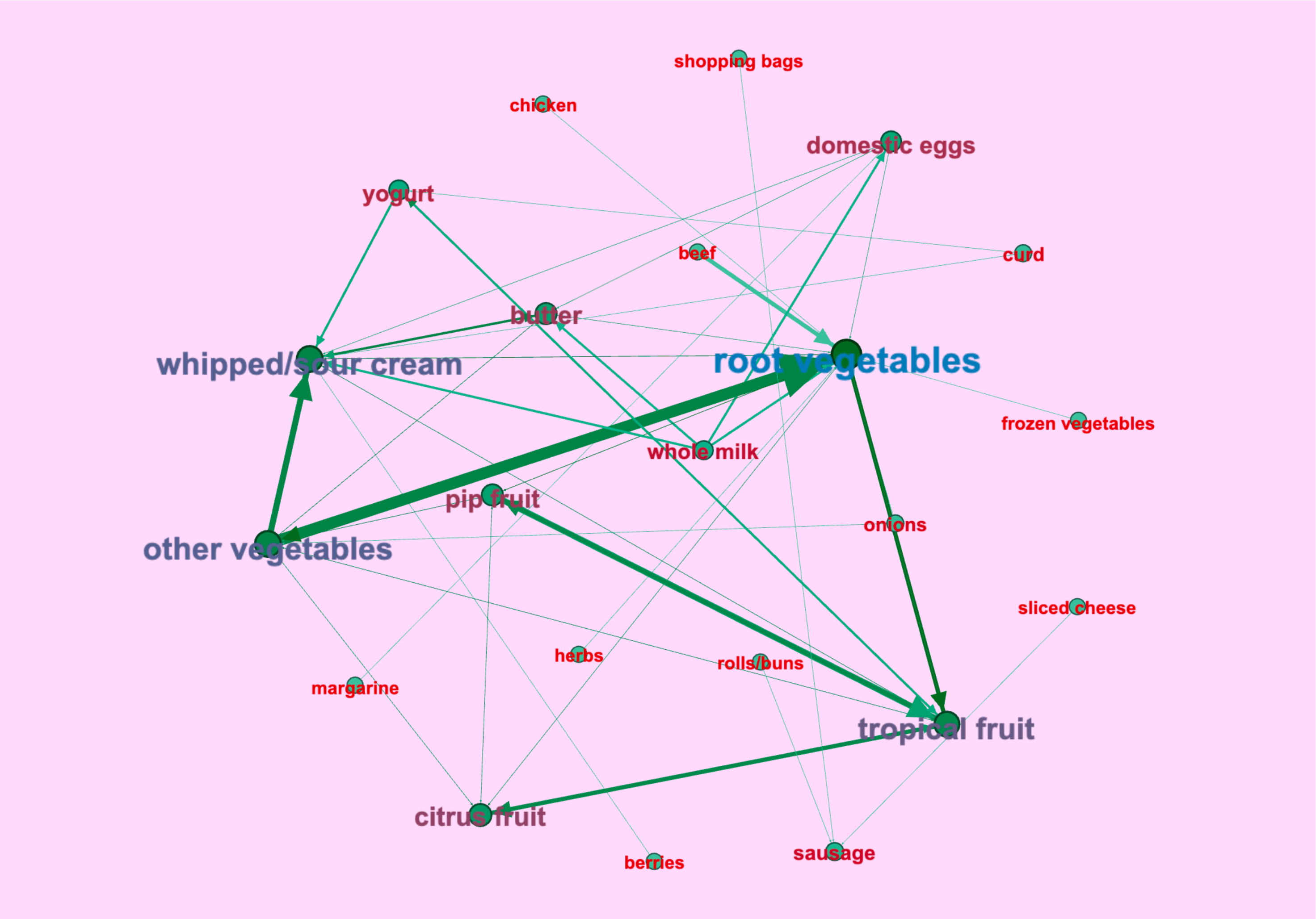
“Two key” plot: coloring is by size (order) of item set

```
plot(grocery_rule, method='two-key plot', jitter =0)
```



- The two key plot shows that most rules are lying in the lower support region, indicating that these item combinations are rare, but they have variations in their confidence
- Order Analysis: Majority of rules include combination of 3 items and 2 items where a single item association is very low. Most of them have low support and high variation in confidence

```
grocery_graph = associations2igraph(subset(grocery_rule, lift>3), associationsAsNodes
= FALSE)
igraph::write_graph(grocery_graph, file='groceries.graphml', format = "graphml")
```



Gephi Graph: Visualization of Grocery Item Associations

From the gephi graph, we can interpret that:

- Root vegetables, other vegetables, whipper/sour cream more frequently associated with others in the grocery list as they have larger nodes
- {other vegetables} => {root vegetables}: This connection is strongly associated amongst other connections
- Smaller, more isolated nodes like chicken and margarine indicates that these items have fewer or weaker associations with other items in the dataset