Market Basket Analysis for FDMart Grocery

Code

Imali Nadya Wanigasundara

The below report illustrates and analyzes various customer purchasing patters at FDMart Grocery by performing Market Basket Analysis.

First we load Transactions List raw data and install relevant packages.

Hide

```
transaction_list <- read_csv("C:/Users/nadyaw/Downloads/TransactionList.csv", col_names = FALSE)
```

```
Parsed with column specification:
cols(
  X1 = col_integer(),
  X2 = col_character()
)
```

Hide

```
View(transaction_list)
```

In order to clean the dataset, we will add header names to loaded dataset and convert the dataframe into a transaction matrix.

Hide

```
colnames(transaction_list) <- c("transaction_id","item")
grocery_list <- as(split(transaction_list$item, transaction_list[,"transaction_id"]), "transacti
ons")
inspect(head(grocery_list,5))
```

Per below summary, there were 64,808 transactions with 106 distinct items. Most ferquently bought item is Fresh Vegetables with 20,001 purchases Median basket size was 5 items while mode was 4 items.

Hide

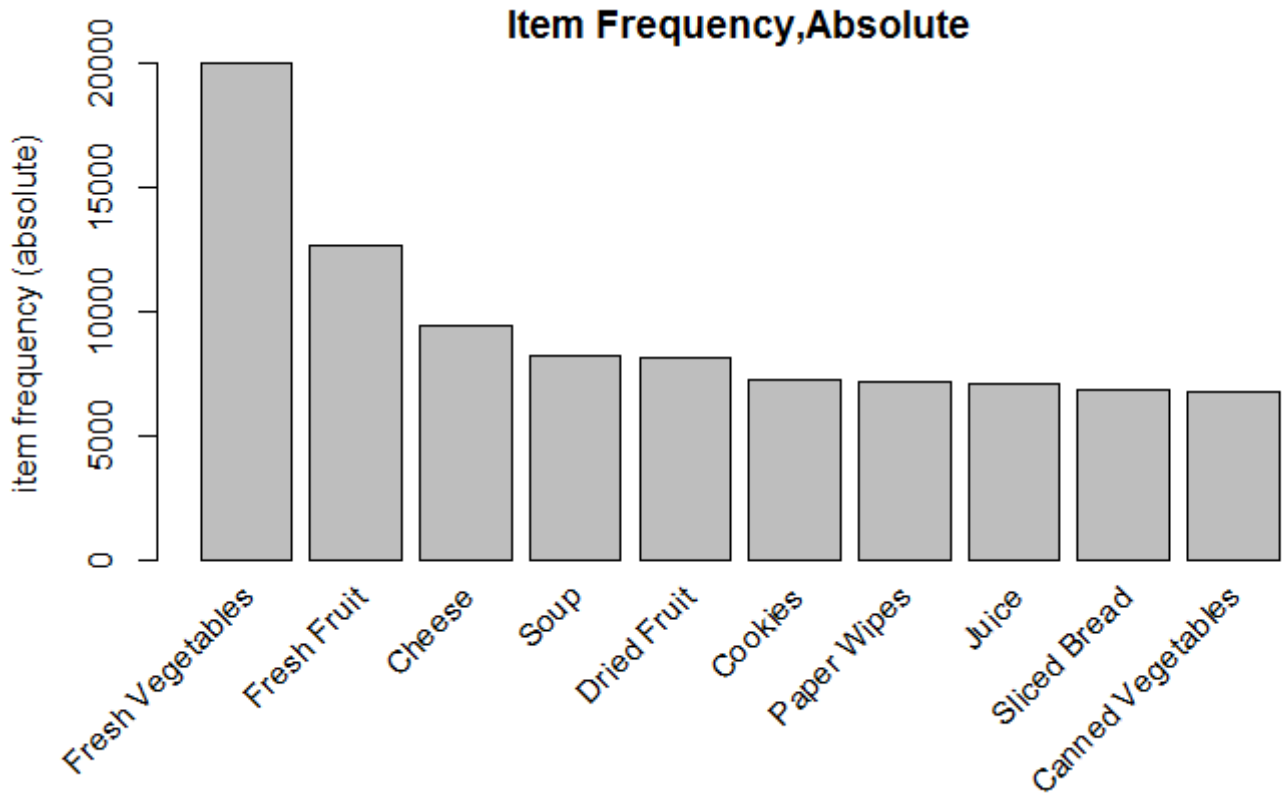
```
summary(grocery_list)
```

```
transactions as itemMatrix in sparse format with
 64808 rows (elements/itemsets/transactions) and
 106 columns (items) and a density of 0.054

most frequent items:
Fresh Vegetables      Fresh Fruit           Cheese             Soup
           20001             12641              9380             8209
     Dried Fruit          (Other)
            8140           312839

element (itemset/transaction) length distribution:
sizes
    1     2     3     4     5     6     7     8     9    10    11    12    13    14
 4489  8628  8522 10010  8344  9013  6075  2247   997  1024   999   672   436   249
   15    16    17    18    19    20    21    22    23    24    25    26    27    28
  235   226   149    96    80    94    91    77    85    91    92   123   162   207
   29    30    31    32    33    34    35    36    37    38    39    40    41    42
  226   216   174   152   124   115    79    63    62    28    26    14     8     6
   43    44
    1     1

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      1       3       5       6       6      44

includes extended item information - examples:
        labels
1 Acetominifen
2    Anchovies
3      Aspirin

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

Below is an item frequency plot for top 10 most frequently purchased items.

Hide

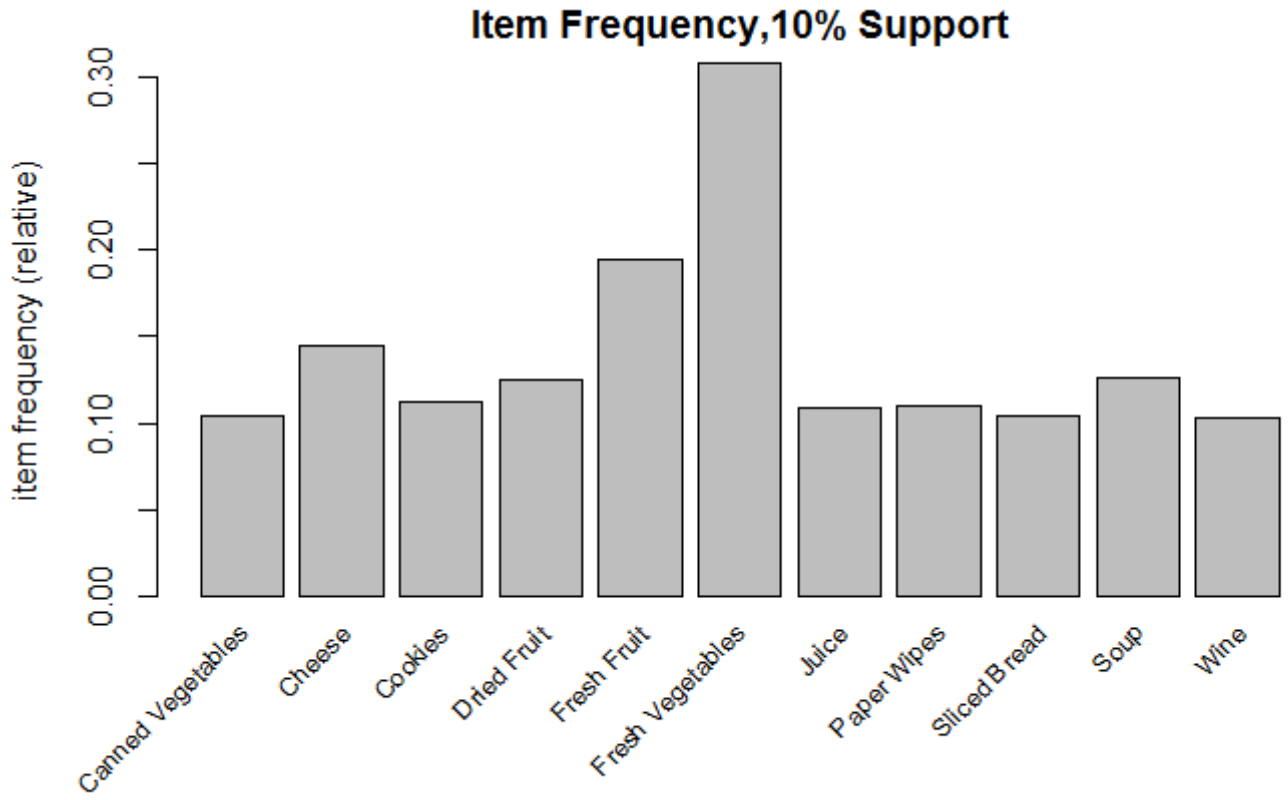
```
itemFrequencyPlot(grocery_list,topN=10,type="absolute", main="Item Frequency,Absolute")
```



Here is the plot for items that occured in 10% of transactions.

Hide

```
itemFrequencyPlot(grocery_list, support=0.1, cex.names=0.8,main="Item Frequency,10% Support")
```



This table shows that if we sort frequency of pruchase, fresh fruits and fresh vegetables were pruchased together in 4766 times, along with other top 5 most frequently purchased items.

Hide

```
#crossTable(grocery_list)['Fresh Vegetables','Canned Vegetables']
#crossTable(grocery_list)['Beer','Wine']
tbl<-crossTable(grocery_list, sort=TRUE)
tbl[1:5,1:5]
```

```
                Fresh Vegetables Fresh Fruit Cheese Soup Dried Fruit
Fresh Vegetables            20001        4766   3211 2347        2730
Fresh Fruit                  4766       12641   2209 1430        1489
Cheese                       3211        2209   9380 1398        1277
Soup                         2347        1430   1398 8209         860
Dried Fruit                  2730        1489   1277  860        8140
```

Based on lift values, it seems that Fresh Vegetables are purchased together with all items below except Soup. A lift value of 0.93 indicates that these products may be substitute.

Hide

```
crossTable(grocery_list, measure='lift',sort=T)[1:5,1:5]
```

```
                Fresh Vegetables Fresh Fruit   Cheese      Soup Dried Fruit
Fresh Vegetables               NA   1.2216576 1.109212 0.9264026   1.0867137
Fresh Fruit             1.2216576          NA 1.207369 0.8930854   0.9378157
Cheese                  1.1092115   1.2073691       NA 1.1766375   1.0839076
Soup                    0.9264026   0.8930854 1.176638        NA   0.8340890
Dried Fruit             1.0867137   0.9378157 1.083908 0.8340890          NA
```

However, calculating the ChiSquared value enables us to confirm if soup and fresh vegetables are a substitute by chance.

Hide

```
crossTable(grocery_list, measure='chi')['Fresh Vegetables', 'Soup']
```

```
[1] 0.0002117436
```

Thus, the low ChiSquared p value of 0.0002 indicates that Fresh Vegetables and Soup may indeed be subsititues and this is not a coincidence.

No we will use arules and apriori function to examine more complex purchase patterns. First we will look at purchases based on wine and beer.

Hide

```
inspect(winerules[1:5])
```

```
    lhs                                rhs    support confidence lift
[1] {Fresh Vegetables,Sauces}       => {Wine} 0.015   0.67       6.6
[2] {Fresh Chicken,Fresh Vegetables} => {Wine} 0.010   0.64       6.2
[3] {Candles,Fresh Vegetables}      => {Wine} 0.010   0.62       6.1
[4] {Sauces}                        => {Wine} 0.016   0.53       5.1
[5] {Candles}                       => {Wine} 0.012   0.46       4.5
```

Based on the output above, there are 21 purchases involed with Wine and other basket of items. For instance, when customers purchase Candles and Fresh Vegetables, they are 62% likely to purchase Wine in the same grocery trip. Similarly, a lift of 5 shows that customers are about 5 times likely to purchase Sauces and Wine together compared to purchases that are assumed to be unrelated.

On the other hand, below output shows what customers are most likely to buy after buying beer based on the same minimum support and 0.2 confidence threshold. Accordingly, we can see that a customer who purchases Beer is likely to purchase Chips about 32% of the time and this is likely to happen about 3 times. However, this is not a storng correlation.

Hide

```
inspect(beer_rules[1:5])
```

```
    lhs       rhs                 support confidence lift
[1] {Beer} => {Fresh Vegetables}  0.018   0.36       1.2
[2] {Beer} => {Chips}             0.016   0.32       3.3
[3] {Beer} => {Cheese}            0.015   0.30       2.1
[4] {Beer} => {Canned Vegetables} 0.015   0.30       2.9
[5] {Beer} => {Eggs}              0.014   0.28       3.1
```

Lets consider the case of Canned vs Fresh vegetables. Intuitively it may appear to be subsititute products.

Hide

```
inspect(fresh_canned[1:10])
```

```
     lhs                                    rhs          support confidence lift
[1]  {Canned Vegetables,Jelly,Juice}        => {Sour Cream} 0.010   0.82       17
[2]  {Canned Vegetables,Jelly,Pancake Mix}  => {Sour Cream} 0.010   0.82       17
[3]  {Canned Vegetables,Cereal,Jelly}       => {Sour Cream} 0.010   0.82       17
[4]  {Fresh Vegetables,Pancake Mix,Waffles} => {Sour Cream} 0.011   0.81       17
[5]  {Deodorizers,Fresh Vegetables,Rice}    => {Sour Cream} 0.011   0.81       17
[6]  {Cottage Cheese,Fresh Vegetables,Rice} => {Sour Cream} 0.011   0.80       17
[7]  {Fresh Vegetables,Pancake Mix,Rice}    => {Sour Cream} 0.011   0.80       17
[8]  {Fresh Vegetables,Rice,Waffles}        => {Sour Cream} 0.011   0.80       17
[9]  {Deodorizers,Fresh Vegetables,Waffles} => {Sour Cream} 0.011   0.80       17
[10] {Fresh Vegetables,Jam,Rice}            => {Sour Cream} 0.011   0.80       17
```

It seems from the above output that customers do not have a preference between canned vegetables and fresh vegetables when they are buying sour cream in the same basket. This can be expected since if sour cream and other ingredients such as Pancake Mix or Waffles are involved then customers are likely to purchase either canned or fresh vegetables for a potential desert or breakfast.The confidence percent is about the same for those transactions.

Hide

```
inspect(fresh_can[1:5])
```

```
    lhs                  rhs                 support confidence lift
[1] {Fresh Vegetables,
     Pancake Mix,
     Sour Cream}       => {Canned Vegetables}   0.011       0.81  7.8
[2] {Cottage Cheese,
     Fresh Vegetables,
     Jelly}            => {Canned Vegetables}   0.011       0.80  7.7
[3] {Fresh Vegetables,
     Juice,
     Sour Cream}       => {Canned Vegetables}   0.011       0.80  7.6
[4] {Cheese,
     Fresh Vegetables,
     Pancake Mix}      => {Canned Vegetables}   0.011       0.80  7.6
[5] {Fresh Vegetables,
     Pancake Mix,
     Waffles}          => {Canned Vegetables}   0.011       0.80  7.6
```

This result shows that perhaps Fresh vegetables and canned vegetables are not substitutes since they seem to be purchased at the same time as bunch of other dessert/ breakfast related ingredients. This could also imply that it is a coincidence that fresh vegetables are purchased at the same time as canned vegetables in some transactions.

Hide

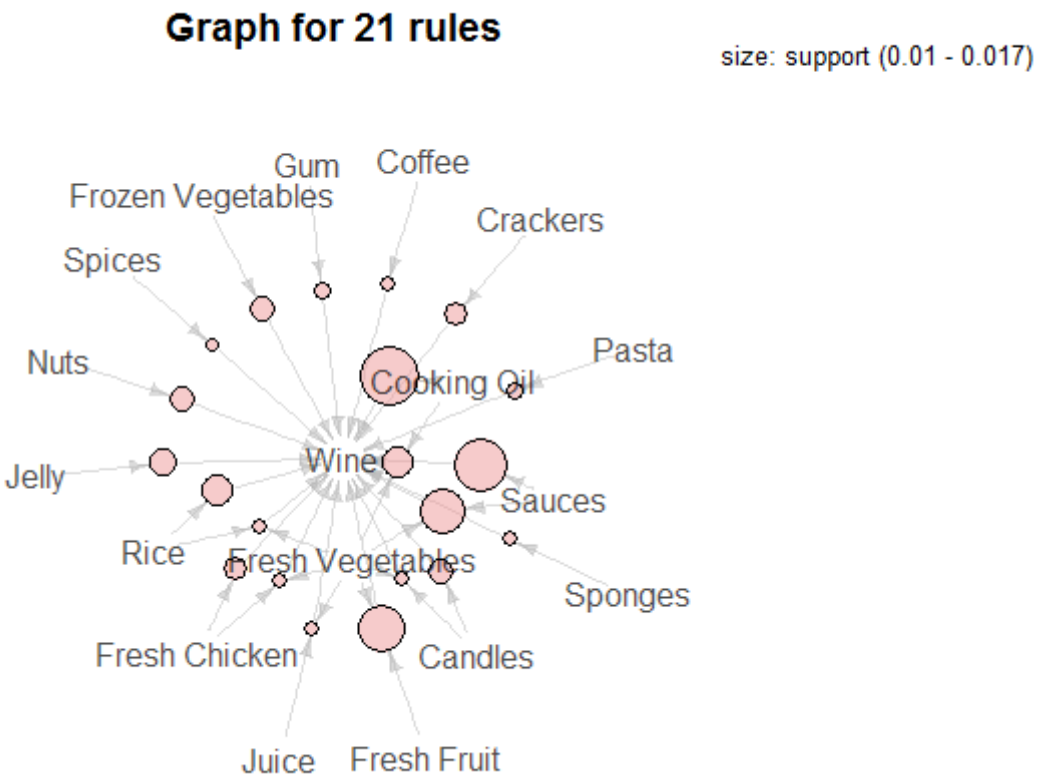
```
vegetables <- subset(rules, lhs %pin% 'Vegetables' & rhs %pin% 'Vegetables')
inspect(vegetables[1:10])
```

```
     lhs                                  rhs                 support confidence
[1]  {Frozen Vegetables}               => {Canned Vegetables} 0.011   0.14
[2]  {Canned Vegetables}               => {Frozen Vegetables} 0.011   0.10
[3]  {Frozen Vegetables}               => {Fresh Vegetables}  0.032   0.40
[4]  {Fresh Vegetables}                => {Frozen Vegetables} 0.032   0.10
[5]  {Canned Vegetables}               => {Fresh Vegetables}  0.040   0.38
[6]  {Fresh Vegetables}                => {Canned Vegetables} 0.040   0.13
[7]  {Canned Vegetables,Shrimp}        => {Fresh Vegetables}  0.010   0.69
[8]  {Fresh Vegetables,Shrimp}         => {Canned Vegetables} 0.010   0.65
[9]  {Canned Vegetables,Peanut Butter} => {Fresh Vegetables}  0.011   0.78
[10] {Fresh Vegetables,Peanut Butter}  => {Canned Vegetables} 0.011   0.45
     lift
[1]  1.3
[2]  1.3
[3]  1.3
[4]  1.3
[5]  1.2
[6]  1.2
[7]  2.3
[8]  6.2
[9]  2.5
[10] 4.3
```

Above output confirms that fresh and canned vegetables are indeed substitutes as they are only purchased together about 10% of the time. Similarly, frozen and canned vegetables seem to be substitutes as well due to the low confidence level.

Hide

```
plot(winerules,method="graph",interactive=FALSE,shading=NA)
```



Now we will compare some small item sets and large item sets. First we will only look at 3 items in the basket ordered by lift. Based on data below, it seems that 75% of customers would purchase Pots and Pans when they buy Cooking Oil and Rice and they are 28 times likely to purchase these 3 items together.

Hide

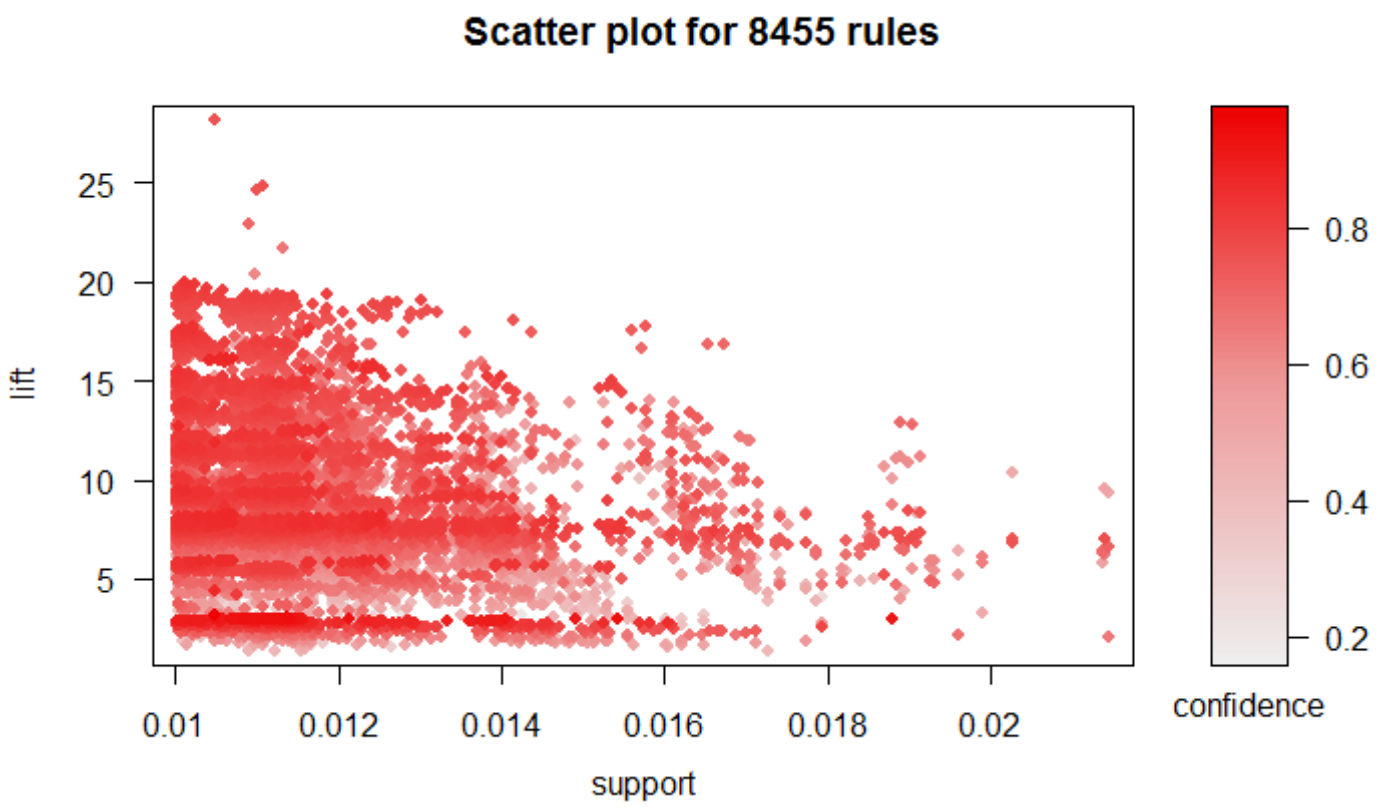
```
inspect(head(sort(basket_rules, by = "lift",5)))
```

```
    lhs                          rhs             support    confidence lift
[1] {Cooking Oil,Rice}        => {Pots and Pans} 0.01047710 0.7502762  28.18777
[2] {Chips,Deodorizers}       => {Shrimp}        0.01106345 0.7563291  24.86868
[3] {Chips,Pancake Mix}       => {Shrimp}        0.01100173 0.7489496  24.62604
[4] {Chips,Frozen Chicken}    => {Shrimp}        0.01089372 0.6983185  22.96125
[5] {Chips,Waffles}           => {Shrimp}        0.01131033 0.6597660  21.69361
[6] {Bagels,Fresh Vegetables} => {Conditioner}   0.01098630 0.6111588  20.34308
```

The below scatterplot based on 3 itemssets show the area of transactions that lie within the confidence level and lift. Rules with high lift generally have a lower support as evident from the plot.

Hide

```
plot(basket_rules, measure=c("support", "lift"), shading="confidence")
```



Below is a basket with 5 items illustrating what consumers are most likely to buy before Deodorizers. Since this analysis is not based on restricting Deodorizers, it is interesting that all top transaction sets with highest lift indicated the purchase connection of Deodorizers with other basket items.

Hide

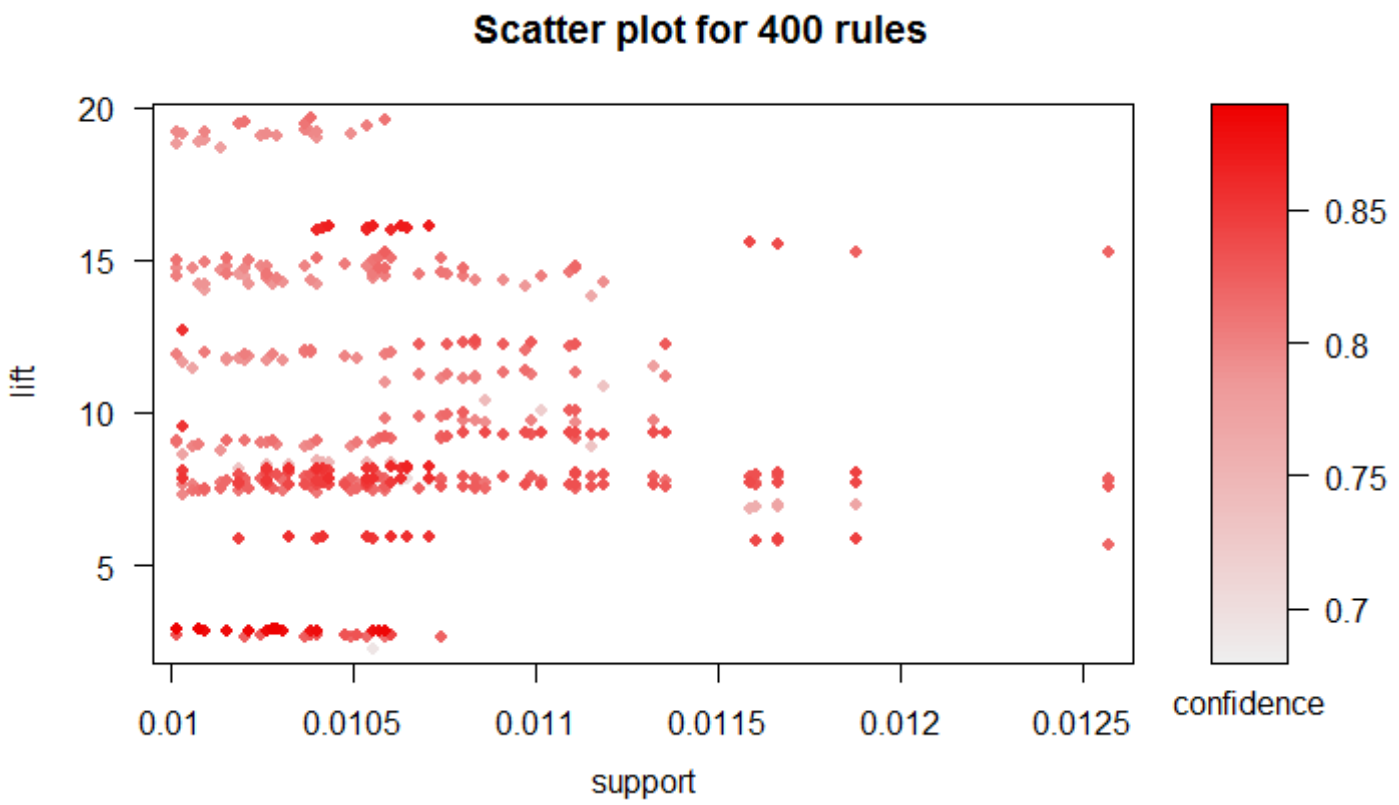
```
inspect(head(sort(basket_rules_large, by = "lift")))
```

```
    lhs                   rhs           support confidence lift
[1] {Cottage Cheese,
     Fresh Vegetables,
     Frozen Chicken,
     Sliced Bread}     => {Deodorizers}   0.010       0.81   20
[2] {Fresh Vegetables,
     Frozen Chicken,
     Juice,
     Sliced Bread}     => {Deodorizers}   0.011       0.81   20
[3] {Fresh Vegetables,
     Frozen Chicken,
     Pancake Mix,
     Sliced Bread}     => {Deodorizers}   0.010       0.81   20
[4] {Cereal,
     Fresh Vegetables,
     Frozen Chicken,
     Sliced Bread}     => {Deodorizers}   0.010       0.81   19
[5] {Frozen Chicken,
     Juice,
     Pancake Mix,
     Sliced Bread}     => {Deodorizers}   0.010       0.81   19
[6] {Fresh Vegetables,
     Juice,
     Pancake Mix,
     Sliced Bread}     => {Deodorizers}   0.011       0.80   19
```

In contrast to the above scatter plot, this is less clustered around a clear boundary of lift and confidence partly because we have restricted the number of rules here and increased the basked of items.

Hide

```
plot(basket_rules_large, measure=c("support", "lift"), shading="confidence")
```



Here we will look at another purchasing pattern: breakfast food items. Based on the below output we can see that consumers who purchased Bagels, Milk and Sliced Bread were about 84% likely to purchase Juice or Muffins. This can be further confirmed from high lift values.

Hide

```
inspect(breakfast_rules[1:10])
```

```
     lhs                            rhs                support confidence lift
[1]  {Bagels,Yogurt}             => {Fresh Vegetables} 0.011   0.86        2.8
[2]  {Bagels,Conditioner}        => {Fresh Vegetables} 0.011   0.85        2.8
[3]  {Bagels,Milk,Sliced Bread}  => {Muffins}          0.012   0.85       11.2
[4]  {Bagels,Juice,Muffins}      => {Milk}             0.012   0.84        9.4
[5]  {Bagels,Juice,Milk}         => {Muffins}          0.012   0.84       11.2
[6]  {Bagels,Milk,Sliced Bread}  => {Juice}            0.012   0.84        7.7
[7]  {Bagels,Milk,Muffins}       => {Juice}            0.012   0.84        7.7
[8]  {Bagels,Juice,Milk}         => {Sliced Bread}     0.012   0.84        8.0
[9]  {Bagels,Juice,Muffins}      => {Sliced Bread}     0.012   0.83        8.0
[10] {Bagels,Milk,Muffins}       => {Sliced Bread}     0.012   0.83        8.0
```

Below is an illustration of how purchase of Bagels is associated with other breakfast items based on top 10 confidence level. As evident from the graph, there is a strong connection between Bagels and basket items such as Muffins, Milk, Juice and Sliced Bread from the nodes and dark shades of colors in the circles represented.

Hide

```
rules <- apriori(grocery_list, parameter = list(support=.01,conf = .08))
```

```
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen
       0.08    0.1    1 none FALSE            TRUE       5    0.01      1     10
 target   ext
  rules FALSE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 648

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[106 item(s), 64808 transaction(s)] done [0.12s].
sorting and recoding items ... [104 item(s)] done [0.02s].
creating transaction tree ... done [0.08s].
checking subsets of size 1 2 3 4 5 6 done [0.26s].
writing ... [9956 rule(s)] done [0.01s].
creating S4 object  ... done [0.05s].
```

Hide

```
breakfast_rules <- subset(rules, subset = lhs %pin% "Bagel" & size(rules) > 2)
breakfast_rules <- head(sort(breakfast_rules, decreasing=TRUE,by="confidence"),10)
plot(breakfast_rules, method="graph",control=list(type="items",main=""))
```