Association Rules Project

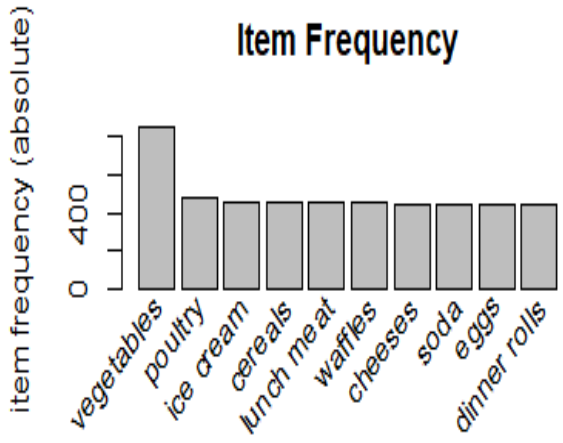
Rezart Abazi & Shamxal Haciyev

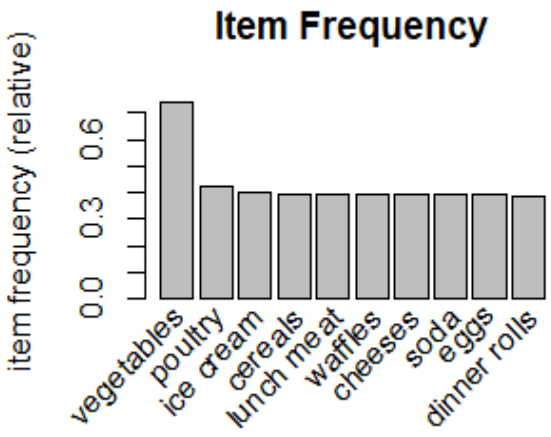
This data set is created by recording the transactions made. These shopping carts were randomly generated. This data set contains 3 columns: “Date to add register”, “Id transaction”, “Product for id transaction”.
In total are 38 products considered in this research:

Yogurt	Pork	Sandwich bags	Lunch-Meat
All- purpose	Flour	Soda	Butter
Vegetables	Beef	Aluminum Foil	Dinner rolls
Shampoo	Mixes	Soap	Laundry detergent
Ice cream	Toilet paper	Waffles	Cheeses
Milk	Individual meals	Hand soap	Dishwashing
Poultry	Cereals	Milk	Ketchup
Spaghetti sauce	Eggs	Juice	Pasta
Tortillas	Fruit	Coffee/Tea	Bagels
Sugar	Paper towels		

Just from a preliminary overview, we can say that vegetables represent 8% of all purchases, poultry 3% and the rest is contained in the remaining 89%

We start with some simple statistics on data to understand the frequency of purchases. On the left we have the histogram of item frequency in absolute terms and in the right one in relative terms. As we can see vegetables have the frequency that is more that twice





Another statistic test we can apply to this data to gain more information about the relation between them is chi-squared. This test measures the relations of data and the significativity level.
We first created a cross table between data and after we applied the chi-squared test as shown in the table below.
For this test the null hypothesis H_0: independent rows and columns
and to reject or accept this hypothesis we have to observe p-value .
From the whole output of R we just chose a small sample just to make it visible for the concept.

```
> round(chi2tab,3)
                    vegetables poultry ice cream cereals lunch meat waffles cheeses  soda  eggs
vegetables                  NA   0.001     0.000   0.001      0.001   0.002   0.001 0.001 0.005
poultry                  0.001      NA     0.001   0.001      0.002   0.000   0.002 0.000 0.001
ice cream                0.000   0.001        NA   0.000      0.002   0.002   0.004 0.001 0.001
cereals                  0.001   0.001     0.000      NA      0.001   0.000   0.003 0.000 0.003
lunch meat               0.001   0.002     0.002   0.001         NA   0.005   0.000 0.000 0.000
waffles                  0.002   0.000     0.002   0.000      0.005      NA   0.002 0.004 0.002
cheeses                  0.001   0.002     0.004   0.003      0.000   0.002      NA 0.002 0.002
soda                     0.001   0.000     0.001   0.000      0.000   0.004   0.002    NA 0.008
eggs                     0.005   0.001     0.001   0.003      0.000   0.002   0.002 0.008    NA
```

Generating Rules

There are three parameters controlling the number of rules to be generated **Support and Confidence**. Another parameter **Lift** is generated using Support and Confidence and is one of the major parameters to filter the generated rules.

To generate the rules we have to use the Apriori with a confidence factor of 80% and a support rank of 15%. We can notice from the output that are 22 generated rules from 38 items and 1139 transactions.
Also is shown the minimum support count means the minimum number of appears of the itemset is 170.
'Minlen' and 'Maxlen' has value 1 like minimum and maximum number of items required in the rule
'Maxtime' is 5 and stands for the maximum amount of time allowed to check for subsets
We have 'originalSupport' indicator that is "TRUE" because the traditional support value only considers both LHS and RHS items for calculating support.
'Aval' is logical indicator whether to return the additional rule evaluation measure selected with 'arem' and in our case is "FALSE".

```
> rules <- apriori(mba, parameter = list(support = 0.15, confidence = 0.8))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxle
n target   ext
        0.8    0.1    1 none FALSE            TRUE       5    0.15      1     1
0  rules FALSE

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

Absolute minimum support count: 170

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[38 item(s), 1139 transaction(s)] done [0.00s].
sorting and recoding items ... [38 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 done [0.00s].
writing ... [22 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
```

Association Rules Project

Rezart Abazi & Shamxal Haciyev

1) Support: It is calculated to check how much popular a given item is. It is measured by the proportion of transactions in which an itemset appears. For example, there are 100 people who bought something from grocery store today, amoung those 100 people, there are 10 people who bought bananas. Hence, the support of people who bought bananas will be (10/100 = 10%).

A rule with a lift of 1.5 imply that, the items in LHS and RHS (X & Y) are 1.5 times more likely to be purchased together compared to the purchases when they are assumed to be unrelated.
Lift is symmetric for the relations of X and Y Lift (X->Y)=Lift (Y->X)

```
> inspect(sort(optimal_value_rules, by="support", decreasing = T))

lhs                                        rhs          support   confidence lift       count
{eggs}                                  => {vegetables} 0.3266023 0.8378378  1.133370   372
{yogurt}                                => {vegetables} 0.3195786 0.8310502  1.124188   364
{aluminum foil}                         => {vegetables} 0.3107989 0.8082192  1.093304  354
{laundry detergent}                     => {vegetables} 0.3090430 0.8167053  1.104783   352
{sugar}                                 => {vegetables} 0.2976295 0.8248175  1.115757   339
{sandwich loaves}                       => {vegetables} 0.2827041 0.8090452  1.094421   332
{dinner rolls,poultry}                  => {vegetables} 0.1615452 0.8288288  1.121183   184
{dishwashing liquid/detergent,poultry}=> {vegetables} 0.1597893 0.8544601  1.155855   182
{eggs,soda}                             => {vegetables} 0.1580334 0.8450704  1.143153   180
{lunch meat,poultry}                    => {vegetables} 0.1580334 0.8490566  1.148546   180
{eggs,yogurt}                           => {vegetables} 0.1571554 0.8994975  1.216779   179
{lunch meat,waffles}                    => {vegetables} 0.1571554 0.8523810  1.153043   179
{mixes,poultry}                         => {vegetables} 0.1562774 0.8599034  1.163218   178
{dinner rolls,eggs}                     => {vegetables} 0.1562774 0.8989899  1.216092   178
{eggs,poultry}                          => {vegetables} 0.1553995 0.8805970  1.191211   177
{dishwashing liquid/detergent,eggs}     => {vegetables} 0.1536435 0.8974359  1.213990   175
{aluminum foil,yogurt}                  => {vegetables} 0.1527656 0.8613861  1.165224   174
{poultry,yogurt}                        => {vegetables} 0.1527656 0.8446602  1.142599   174
{poultry,sugar}                         => {vegetables} 0.1518876 0.8480392  1.147169   173
{cereals,laundry detergent}             => {vegetables} 0.1510097 0.8911917  1.205543   172
{cereals,eggs}                          => {vegetables} 0.1510097 0.8643216  1.169195   172
{cheeses,eggs}                          => {vegetables} 0.1501317 0.8860104  1.198534   171
```

2) Confidence: It is calculated to check how likely if item X is purchased when item Y is purchased. This is measured by the proportion of transactions with item X, in which item Y also appears. Suppose, there are 10 people who bought apple(out of 100), and from those 10 people, 6 people also bought yogurt. Hence, the confidence of (apple -> yogurt) is: (apple -> yogurt) / apple [i.e. 6/10 = 0.6].

```
> inspect(sort(optimal_value_rules, by="confidence", decreasing = T))
    lhs                                rhs          support   confidence lift       Count
{eggs,yogurt}                       => {vegetables} 0.1571554 0.8994975  1.216779    179
{dinner rolls,eggs}                 => {vegetables} 0.1562774 0.8989899  1.216092    178
{dishwashing liquid/detergent,eggs} => {vegetables} 0.1536435 0.8974359  1.213990    175
{cereals,laundry detergent}         => {vegetables} 0.1510097 0.8911917  1.205543    172
{cheeses,eggs}                      => {vegetables} 0.1501317 0.8860104  1.198534    171
{eggs,poultry}                      => {vegetables} 0.1553995 0.8805970  1.191211    177
{cereals,eggs}                      => {vegetables} 0.1510097 0.8643216  1.169195    172
{aluminum foil,yogurt}              => {vegetables} 0.1527656 0.8613861  1.165224    174
{mixes,poultry}                     => {vegetables} 0.1562774 0.8599034  1.163218    178
{dishwashing liquid/detergent,poultry} => {vegetables} 0.1597893 0.8544601  1.155855  182
{lunch meat,waffles}                => {vegetables} 0.1571554 0.8523810  1.153043    179
{lunch meat,poultry}                => {vegetables} 0.1580334 0.8490566  1.148546    180
{poultry,sugar}                     => {vegetables} 0.1518876 0.8480392  1.147169    173
{eggs,soda}                         => {vegetables} 0.1580334 0.8450704  1.143153    180
{poultry,yogurt}                    => {vegetables} 0.1527656 0.8446602  1.142599    174
{eggs}                              => {vegetables} 0.3266023 0.8378378  1.133370    372
{yogurt}                            => {vegetables} 0.3195786 0.8310502  1.124188    364
{dinner rolls,poultry}              => {vegetables} 0.1615452 0.8288288  1.121183    184
{sugar}                             => {vegetables} 0.2976295 0.8248175  1.115757    339
{laundry detergent}                 => {vegetables} 0.3090430 0.8167053  1.104783    352
{sandwich loaves}                   => {vegetables} 0.2827041 0.8090452  1.094421    322
{aluminum foil}                     => {vegetables} 0.3107989 0.8082192  1.093304    354
```

Association Rules Project

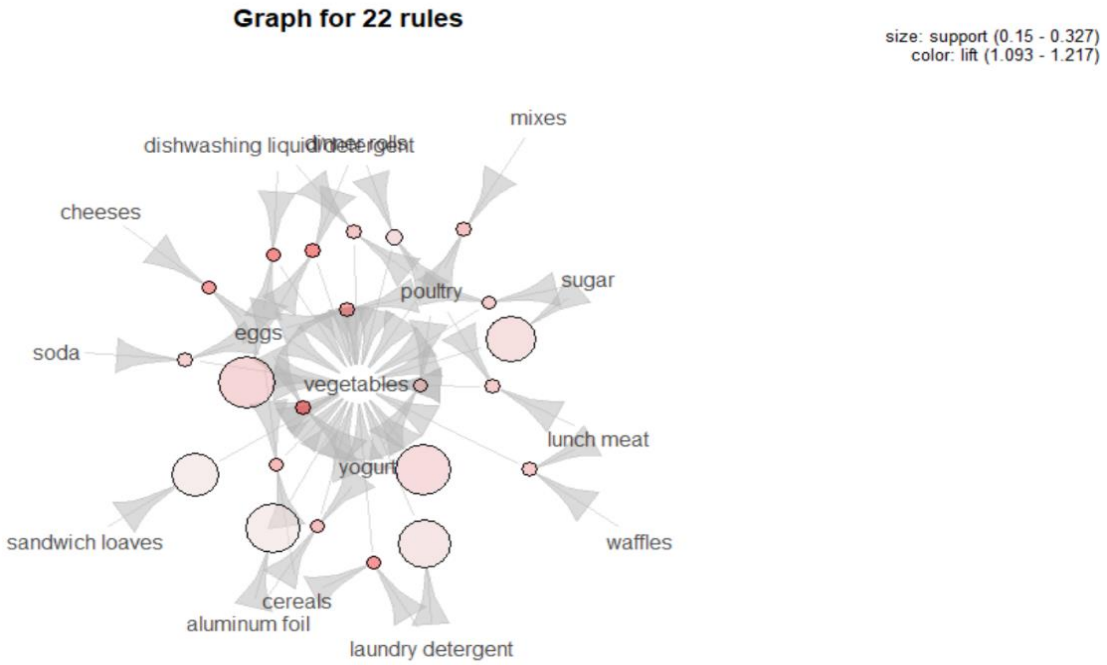
Rezart Abazi & Shamxal Haciyev

3) Lift: It is calculated to measure how likely item Y is purchased when item X is purchased, while controlling for how popular item Y is. The formula for lift is: (lift = support (X ->Y) / (support(X) * support(Y)).

Confidence can be max.1 for how much Y is linked with X, the higher confidence the stronger the rule

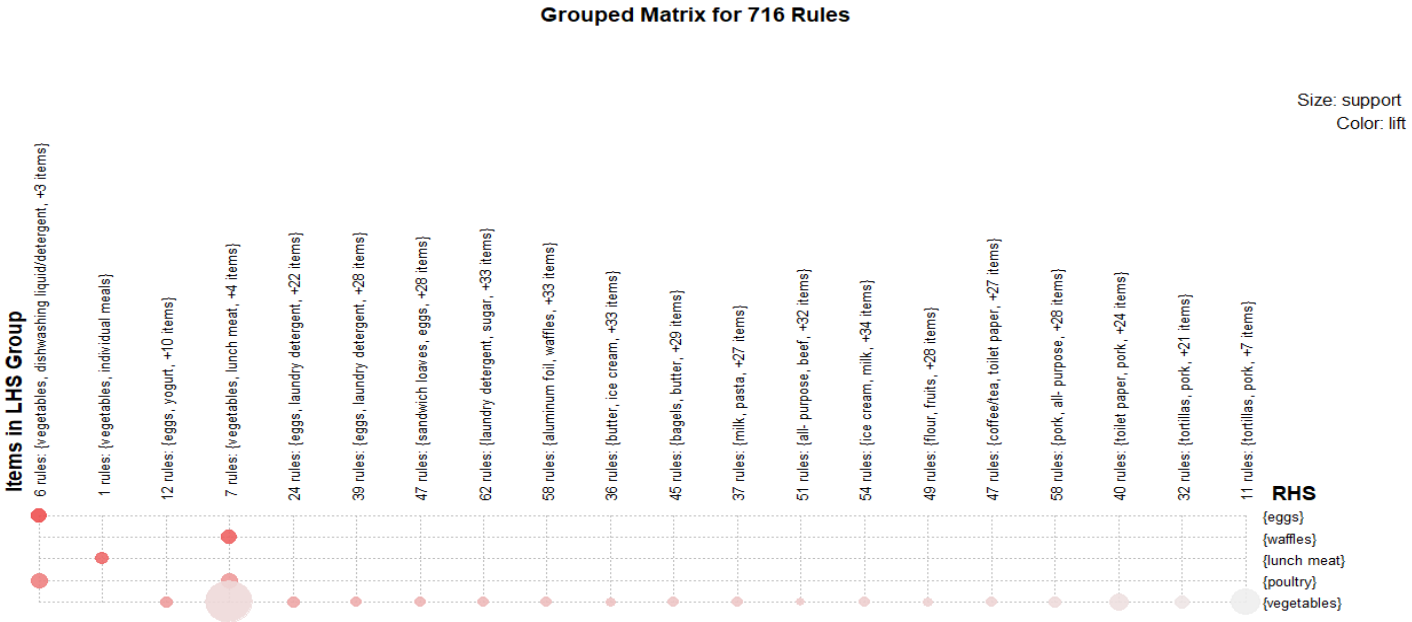
```
> inspect(sort(optimal_value_rules, by="lift", decreasing = T))
     lhs                                  rhs          support   confidence lift     Count
{eggs,yogurt}                          => {vegetables} 0.1571554 0.8994975  1.216779      179
{dinner rolls,eggs}                    => {vegetables} 0.1562774 0.8989899  1.216092      178
{dishwashing liquid/detergent,eggs}    => {vegetables} 0.1536435 0.8974359  1.213990      175
{cereals,laundry detergent}            => {vegetables} 0.1510097 0.8911917  1.205543      172
{cheeses,eggs}                         => {vegetables} 0.1501317 0.8860104  1.198534      171
{eggs,poultry}                         => {vegetables} 0.1553995 0.8805970  1.191211      177
{cereals,eggs}                         => {vegetables} 0.1510097 0.8643216  1.169195      172
{aluminum foil,yogurt}                 => {vegetables} 0.1527656 0.8613861  1.165224      174
{mixes,poultry}                        => {vegetables} 0.1562774 0.8599034  1.163218      178
{dishwashing liquid/detergent,poultry} => {vegetables} 0.1597893 0.8544601  1.155855      182
{lunch meat,waffles}                   => {vegetables} 0.1571554 0.8523810  1.153043      179
{lunch meat,poultry}                   => {vegetables} 0.1580334 0.8490566  1.148546      180
{poultry,sugar}                        => {vegetables} 0.1518876 0.8480392  1.147169      173
{eggs,soda}                            => {vegetables} 0.1580334 0.8450704  1.143153      180
{poultry,yogurt}                       => {vegetables} 0.1527656 0.8446602  1.142599      174
{eggs}                                 => {vegetables} 0.3266023 0.8378378  1.133370      372
{yogurt}                               => {vegetables} 0.3195786 0.8310502  1.124188      364
{dinner rolls,poultry}                 => {vegetables} 0.1615452 0.8288288  1.121183      184
{sugar}                                => {vegetables} 0.2976295 0.8248175  1.115757      339
{laundry detergent}                    => {vegetables} 0.3090430 0.8167053  1.104783      352
{sandwich loaves}                      => {vegetables} 0.2827041 0.8090452  1.094421      322
{aluminum foil}                        => {vegetables} 0.3107989 0.8082192  1.093304      354
```

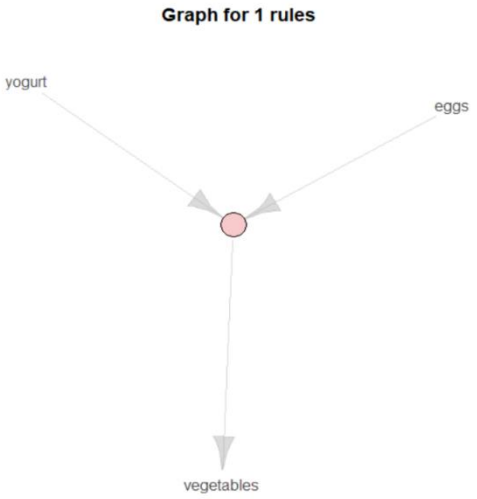
On the below graph we have the graphic representation of the 22 rules.
In the centre our rhs {vegetables} shown with arrows the association with the other items. We can notice that support is higher with items like yogurt, sugar ,eggs based on the size of support.
Low support we have on items like dishwashing , cereals, waffles , lunchmeat, soda.



Association Rules Project

Rezart Abazi & Shamxal Haciyev





Here we have a graph of one rule to understand better the concept.
We have yogurt and eggs associated together by a small support of 0,15 or 15% and a lift of 1,27 which means that there is 1,27 times likely that will be purchased together.