**Market Basket Analysis**

Imagine 10000 receipts sitting on your table. Each receipt represents a transaction with items that were purchased. The receipt is a representation of stuff that went into a customer’s basket – and therefore ‘**Market Basket Analysis**’. That is exactly what the Groceries Data Set contains: a collection of receipts with each line representing 1 receipt and the items purchased. Each line is called a transaction and each column in a row represents an item.

Please download the dataset and perform Market Basket analysis and look at association rules. This will help you understand what goods can be bundled and what types of promotions or discounts can be offered on the products either online or at the retail store!

Dataset: [groceries-1.csv](https://olympus.greatlearning.in/courses/4750/files/466542/download?verifier=CN1GyBclSlVOfJexub6y0VLSDF2aFZk6TxnZqraQ&wrap=1)

**Group Members**

1. Aditya Godse
2. Akshay Surve
3. Amay Saxena

**Answer**

**1] Loading Dataset and Required packages**

library(datasets)

library(arules)

It is essential for mining Association rules, pruning redundant **rules** and visualizing association.

setwd("C:/Users/DELL/Desktop/Akshay/Group Assignments/Group Assignment Marketing and Retail Analytics")

getwd()

Groceries <- read.transactions(read\_excel("groceries xls.xlsx"), format = "basket", header = FALSE,

sep = "")

data("Groceries")

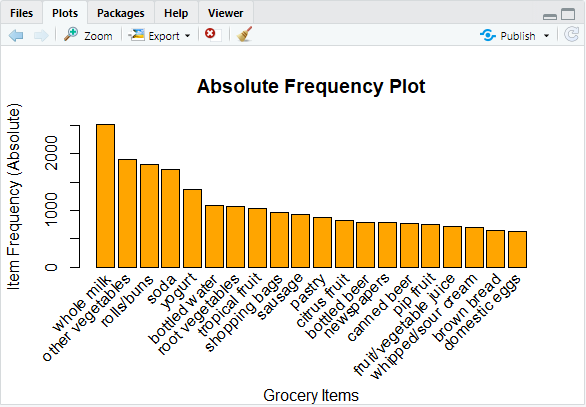
**2] Exploratory Data Analysis**

In order to make any association rules for mining, It is essential to know which grocery item was frequently bought.

Therefore, plotting Item Frequency Plot,

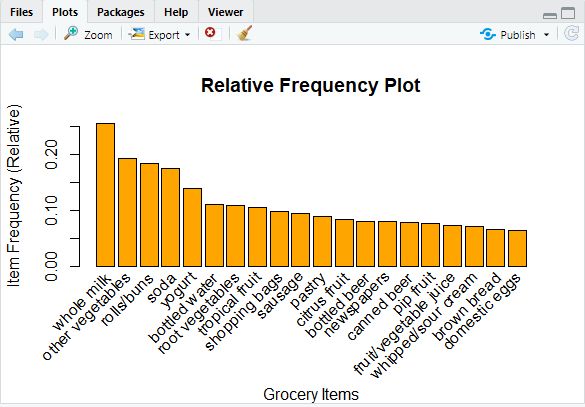
**Absolute Frequency Plot**

itemFrequencyPlot(Groceries,topN=20,type="absolute", xlab = 'Grocery Items', ylab = 'Item Frequency (Absolute)', col = "orange", cex.names = graphics::par("cex.axis"))



**Relative Frequency Plot**

itemFrequencyPlot(Groceries,topN=20,type="relative", xlab = 'Grocery Items', ylab = 'Item Frequency (Relative)', col = "orange", cex.names = graphics::par("cex.axis"))



Item Frequency plot shows that, Whole milk was the most frequently bought grocery item among customers followed by other vegetables, rolls/buns, soda and yogurt.

**3] Preparing Association Rules**

We will start with very restricted threshold and see if gets enough rules In order to get higher number of rules we will relax the threshold. We set the minimum support to 0.001 (This will set up threshold).

We set the minimum confidence of 0.9

Apr\_rule <- apriori(Groceries, parameter = list(supp = 0.001, conf = 0.9))

Apriori

Parameter specification:

confidence minval smax arem aval originalSupport maxtime support minlen maxlen target

0.9 0.1 1 none FALSE TRUE 5 0.001 1 10 rules

ext

FALSE

Algorithmic control:

filter tree heap memopt load sort verbose

0.1 TRUE TRUE FALSE TRUE 2 TRUE

Absolute minimum support count: 9

set item appearances ...[0 item(s)] done [0.00s].

set transactions ...[169 item(s), 9835 transaction(s)] done [0.01s].

sorting and recoding items ... [157 item(s)] done [0.00s].

creating transaction tree ... done [0.00s].

checking subsets of size 1 2 3 4 5 6 done [0.02s].

writing ... [129 rule(s)] done [0.00s].

creating S4 object ... done [0.00s].

Showing the top 20 rules with 3 digits,

options(digits=3)

inspect(Apr\_rule[1:20])

**lhs rhs support confidence lift count**

[1] {liquor,

red/blush wine} => {bottled beer} 0.00193 0.905 11.24 19

[2] {curd,

cereals} => {whole milk} 0.00102 0.909 3.56 10

[3] {soups,

bottled beer} => {whole milk} 0.00112 0.917 3.59 11

[4] {whipped/sour cream,

house keeping products} => {whole milk} 0.00122 0.923 3.61 12

[5] {pastry,

sweet spreads} => {whole milk} 0.00102 0.909 3.56 10

[6] {rice,

sugar} => {whole milk} 0.00122 1.000 3.91 12

[7] {rice,

bottled water} => {whole milk} 0.00122 0.923 3.61 12

[8] {canned fish,

hygiene articles} => {whole milk} 0.00112 1.000 3.91 11

[9] {grapes,

onions} => {other vegetables} 0.00112 0.917 4.74 11

[10] {hard cheese,

oil} => {other vegetables} 0.00112 0.917 4.74 11

[11] {root vegetables,

butter,

rice} => {whole milk} 0.00102 1.000 3.91 10

[12] {herbs,

whole milk,

fruit/vegetable juice} => {other vegetables} 0.00102 0.909 4.70 10

[13] {citrus fruit,

tropical fruit,

herbs} => {whole milk} 0.00112 0.917 3.59 11

[14] {root vegetables,

whipped/sour cream,

flour} => {whole milk} 0.00173 1.000 3.91 17

[15] {butter,

soft cheese,

domestic eggs} => {whole milk} 0.00102 1.000 3.91 10

[16] {tropical fruit,

whipped/sour cream,

soft cheese} => {other vegetables} 0.00122 0.923 4.77 12

[17] {root vegetables,

whipped/sour cream,

soft cheese} => {whole milk} 0.00122 0.923 3.61 12

[18] {citrus fruit,

root vegetables,

soft cheese} => {other vegetables} 0.00102 1.000 5.17 10

[19] {frankfurter,

tropical fruit,

frozen meals} => {other vegetables} 0.00102 0.909 4.70 10

[20] {frankfurter,

tropical fruit,

frozen meals} => {whole milk} 0.00102 0.909 3.56 10

**Reading the chart**

**lhs rhs support confidence lift count**

[1] {liquor,

red/blush wine} => {bottled beer} 0.00193 0.905 11.24 19

Here if someone buys liquor and red/blush wine, they are 90.5% likely to buy bottled water too.

summary(Apr\_rule)

set of 129 rules

rule length distribution (lhs + rhs):sizes

3 4 5 6

10 57 56 6

Min. 1st Qu. Median Mean 3rd Qu. Max.

3.00 4.00 4.00 4.45 5.00 6.00

summary of quality measures:

support confidence lift count

Min. :0.00102 Min. :0.900 Min. : 3.52 Min. :10.0

1st Qu.:0.00102 1st Qu.:0.909 1st Qu.: 3.59 1st Qu.:10.0

Median :0.00102 Median :0.917 Median : 3.67 Median :10.0

Mean :0.00114 Mean :0.934 Mean : 4.22 Mean :11.2

3rd Qu.:0.00122 3rd Qu.:0.933 3rd Qu.: 4.70 3rd Qu.:12.0

Max. :0.00193 Max. :1.000 Max. :11.24 Max. :19.0

mining info:

data ntransactions support confidence

Groceries 9835 0.001 0.9

**Summary indicates**

The number of rules generated: 129

The distribution of rules by length: Most rules are 4 items long

The summary of quality measures: interesting to see ranges of support, lift, and confidence.

The information on the data mined: total data mined, and minimum parameters

**4] Sorting Rules**

Result we got previously are not sorted according to the frequency and placement of relevant rules

Apr\_rule<-sort(Apr\_rule, by="confidence", decreasing=TRUE)

This will sort the rules based on confidence level. Thus, most relevant rules will appear first.

Showing top 20 rules with 3 digits,

options(digits=3)

inspect(Apr\_rule[1:20])

**lhs rhs support confidence lift count**

[1] {rice,

sugar} => {whole milk} 0.00122 1 3.91 12

[2] {canned fish,

hygiene articles} => {whole milk} 0.00112 1 3.91 11

[3] {root vegetables,

butter,

rice} => {whole milk} 0.00102 1 3.91 10

[4] {root vegetables,

whipped/sour cream,

flour} => {whole milk} 0.00173 1 3.91 17

[5] {butter,

soft cheese,

domestic eggs} => {whole milk} 0.00102 1 3.91 10

[6] {citrus fruit,

root vegetables,

soft cheese} => {other vegetables} 0.00102 1 5.17 10

[7] {pip fruit,

butter,

hygiene articles} => {whole milk} 0.00102 1 3.91 10

[8] {root vegetables,

whipped/sour cream,

hygiene articles} => {whole milk} 0.00102 1 3.91 10

[9] {pip fruit,

root vegetables,

hygiene articles} => {whole milk} 0.00102 1 3.91 10

[10] {cream cheese ,

domestic eggs,

sugar} => {whole milk} 0.00112 1 3.91 11

[11] {curd,

domestic eggs,

sugar} => {whole milk} 0.00102 1 3.91 10

[12] {cream cheese ,

domestic eggs,

napkins} => {whole milk} 0.00112 1 3.91 11

[13] {pip fruit,

whipped/sour cream,

brown bread} => {other vegetables} 0.00112 1 5.17 11

[14] {tropical fruit,

grapes,

whole milk,

yogurt} => {other vegetables} 0.00102 1 5.17 10

[15] {ham,

tropical fruit,

pip fruit,

yogurt} => {other vegetables} 0.00102 1 5.17 10

[16] {ham,

tropical fruit,

pip fruit,

whole milk} => {other vegetables} 0.00112 1 5.17 11

[17] {tropical fruit,

root vegetables,

yogurt,

oil} => {whole milk} 0.00112 1 3.91 11

[18] {root vegetables,

other vegetables,

yogurt,

oil} => {whole milk} 0.00142 1 3.91 14

[19] {root vegetables,

other vegetables,

butter,

white bread} => {whole milk} 0.00102 1 3.91 10

[20] {pork,

other vegetables,

butter,

whipped/sour cream} => {whole milk} 0.00102 1 3.91 10

Now, in the above table it can be seen that, some rules are excessively long. So in order to achieve more concise rules we can add “maxlen” parameter which defines the maximum number of items in each item set of frequent items.

Apr\_rule <- apriori(Groceries, parameter = list(supp = 0.001, conf = 0.8,maxlen=3))

Parameter specification:

confidence minval smax arem aval originalSupport maxtime support minlen maxlen target

0.8 0.1 1 none FALSE TRUE 5 0.001 1 3 rules ext

FALSE

Algorithmic control:

filter tree heap memopt load sort verbose

0.1 TRUE TRUE FALSE TRUE 2 TRUE

Absolute minimum support count: 9

set item appearances ...[0 item(s)] done [0.00s].

set transactions ...[169 item(s), 9835 transaction(s)] done [0.01s].

sorting and recoding items ... [157 item(s)] done [0.00s].

creating transaction tree ... done [0.00s].

checking subsets of size 1 2 3 done [0.01s].

writing ... [29 rule(s)] done [0.00s].

creating S4 object ... done [0.00s].

**5] Removing Repeating Rules**

While rules are being generated, it is essential to remove those which are of similar nature.

This can be done with,

Apr\_rule <- Apr\_rule[!is.redundant(Apr\_rule)]

inspect(head(Apr\_rule, 20))

**lhs rhs support**

[1] {liquor,red/blush wine} => {bottled beer} 0.00193

[2] {curd,cereals} => {whole milk} 0.00102

[3] {yogurt,cereals} => {whole milk} 0.00173

[4] {butter,jam} => {whole milk} 0.00102

[5] {soups,bottled beer} => {whole milk} 0.00112

[6] {napkins,house keeping products} => {whole milk} 0.00132

[7] {whipped/sour cream,house keeping products} => {whole milk} 0.00122

[8] {pastry,sweet spreads} => {whole milk} 0.00102

[9] {turkey,curd} => {other vegetables} 0.00122

[10] {rice,sugar} => {whole milk} 0.00122

[11] {butter,rice} => {whole milk} 0.00153

[12] {domestic eggs,rice} => {whole milk} 0.00112

[13] {rice,bottled water} => {whole milk} 0.00122

[14] {yogurt,rice} => {other vegetables} 0.00193

[15] {oil,mustard} => {whole milk} 0.00122

[16] {canned fish,hygiene articles} => {whole milk} 0.00112

[17] {herbs,fruit/vegetable juice} => {other vegetables} 0.00122

[18] {herbs,shopping bags} => {other vegetables} 0.00193

[19] {tropical fruit,herbs} => {whole milk} 0.00234

[20] {herbs,rolls/buns} => {whole milk} 0.00244

**confidence lift count**

[1] 0.905 11.24 19

[2] 0.909 3.56 10

[3] 0.810 3.17 17

[4] 0.833 3.26 10

[5] 0.917 3.59 11

[6] 0.812 3.18 13

[7] 0.923 3.61 12

[8] 0.909 3.56 10

[9] 0.800 4.13 12

[10] 1.000 3.91 12

[11] 0.833 3.26 15

[12] 0.846 3.31 11

[13] 0.923 3.61 12

[14] 0.826 4.27 19

[15] 0.857 3.35 12

[16] 1.000 3.91 11

[17] 0.800 4.13 12

[18] 0.826 4.27 19

[19] 0.821 3.21 23

[20] 0.800 3.13 24

**6] Targeting Items for Generating Rules**

Limits can be put on output by targeting items to generate rules.

For example:

1. What are customers likely to buy before buying whole milk?
2. What are customers likely to buy if they purchase whole milk?

This means setting up of either Left Hand Side or Right Hand Side.

For answering 1st question apriori() function can be modified as following:

1. **What are customers likely to buy before buying whole milk?**

Apr\_rule<-apriori(data=Groceries, parameter=list(supp=0.001,conf = 0.8),

appearance = list(default="lhs", rhs="whole milk"), control = list(verbose=F))

Apr\_rule<-sort(Apr\_rule, decreasing=TRUE, by="confidence")

inspect(Apr\_rule[1:5])

**lhs rhs support confidence lift**

[1] {rice,sugar} => {whole milk} 0.00122 1 3.91

[2] {canned fish,hygiene articles} => {whole milk} 0.00112 1 3.91

[3] {root vegetables,butter,rice} => {whole milk} 0.00102 1 3.91

[4] {root vegetables,whipped/sour cream,flour} => {whole milk} 0.00173 1 3.91

[5] {butter,soft cheese,domestic eggs} => {whole milk} 0.00102 1 3.91

**count**

[1] 12

[2] 11

[3] 10

[4] 17

[5] 10

Likewise, “whole milk” can be set to the left hand side in order to find Right Hand Side.

Since we are not getting significant number of rules with 0.8, setting the confidence level to 0.15.

Setting minimum length of 2 to avoid empty left hand side items.

For answering 2nd question apriori() function can be modified as following:

1. **What are customers likely to buy if they purchase whole milk?**

Apr\_rule<-apriori(data=Groceries, parameter=list(supp=0.001,conf = 0.15,minlen=2),

appearance = list(default="rhs",lhs="whole milk"), control = list(verbose=F))

Apr\_rule<-sort(Apr\_rule, decreasing=TRUE,by="confidence")

inspect(Apr\_rule[1:5])

**lhs rhs support confidence lift count**

[1] {whole milk} => {other vegetables} 0.0748 0.293 1.51 736

[2] {whole milk} => {rolls/buns} 0.0566 0.222 1.21 557

[3] {whole milk} => {yogurt} 0.0560 0.219 1.57 551

[4] {whole milk} => {root vegetables} 0.0489 0.191 1.76 481

[5] {whole milk} => {tropical fruit} 0.0423 0.166 1.58 416

summary(Apr\_rule)

set of 6 rules

rule length distribution (lhs + rhs):sizes

2

6

Min. 1st Qu. Median Mean 3rd Qu. Max.

2 2 2 2 2 2

summary of quality measures:

support confidence lift count

Min. :0.0401 Min. :0.157 Min. :0.899 Min. :394

1st Qu.:0.0440 1st Qu.:0.172 1st Qu.:1.282 1st Qu.:432

Median :0.0525 Median :0.205 Median :1.543 Median :516

Mean :0.0531 Mean :0.208 Mean :1.421 Mean :522

3rd Qu.:0.0565 3rd Qu.:0.221 3rd Qu.:1.576 3rd Qu.:556

Max. :0.0748 Max. :0.293 Max. :1.756 Max. :736

mining info:

data ntransactions support confidence

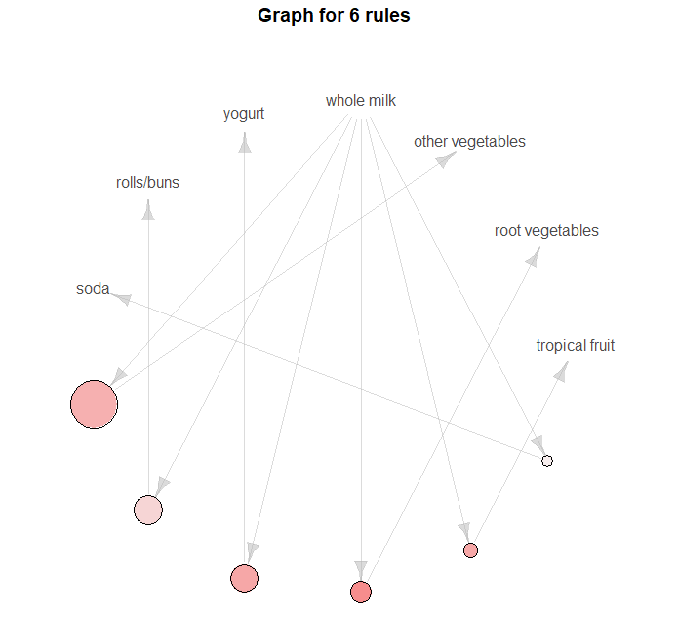
Groceries 9835 0.001 0.15

**7] Visualizing the Results**

To map out the rules in a graph, library called “arulesViz” is required.

library(arulesViz)

plot(Apr\_rule, method="graph", layout=igraph::in\_circle())





**Size**

Larger size indicates large value of support.

**Support** of a rule is a measure of how frequently the items involved bought together.

* Here, whole milk and other vegetables frequently bought together. Followed by whole milk and yogurt, whole milk and rolls/buns.
* While whole milk and soda are the least frequently bought by the customers.

**Lift**

Darker shade of color indicates higher lift and vice-a-versa.

**Lift is the** factor by which, the co-occurrence of A and B exceeds the expected probability of A and B co-occurring, had they been independent. So, higher the **lift**, higher the chance of A and B occurring together.

* The co-occurrence of whole milk and root vegetables exceeds the expected probability with lift of 1.756.
* The co-occurrence of whole milk and soda did not exceed the expected probability resulting in lift of 0.899

**Reading Chart**

* The frequency plot shows that in these data the five most purchased items in order of frequency are whole milk, other vegetables, rolls/buns, soda, and yogurt.
* The apriori() function returns 129 association rules for these data.
* Using the is.redundant() function reduces the number of rules from 129 to 6.
* The is.redundant() function pointless since there are no redundancies in the 6 rules.
* Inspecting the top five rules sorted by confidence shows that most of the associations are with whole milk and other vegetables which are the two most purchased items.

**8] Interpretation of Result**

1. It would be wise to give discount on rolls/buns in order to increase sale since the lift is only 1.21 which nearer to 1.
2. Since whole milk and root vegetables are mostly bought together there is no need to put discount on them, as they already giving profit.
3. Prices of other vegetables, yogurt and tropical fruit can be increased moderately. Since significant sales are occurring. Same can be said for root Vegetables.