Market basketball analysis algorithm for determining products association

Fadhil Muhammad Basysyar*, Gifthera Dwilestari, Agus Bahtiar, Martanto, Nisa Dienwati Nuris

STMIK IKMI Cirebon, Cirebon, Indonesia

*fadhil.m.basysyar@gmail.com

Abstract. Market basket analysis may be a technique identify the association between couple's products purchased jointly and identify cases. Scene an occurrence is when two or more things happening. Market basket analysis makes rules if-then, scenario for instance, if an item purchased a and b will likely. bought items rule is probability in nature or, in other words, they are available from the incident. In observation the frequency is that the proportion of a basket of things which will be employed interesting. Rule, price, product placement and various varieties of cross-selling strategy. In order to create it easier to grasp, think in terms of market research handbasket in a very supermarket. Market basket analysis taking data on the transactions containing an inventory of all items that bought by customers in a very purchase. The technique is ascertaining the relations which product bought on other products. This relationship then wont to construct the containing the if-then items purchased. Market basket analysis also called a rule association or analysis affinity, learning may be a technique data processing can be used in numerous fields, as education, marketing, informatics and science. The aim is to include retailer's awareness of buyers' actions in order for the retailers with knowledge about the actions of purchasers, which will allow the tailor to make the right choices. All of the numerous market basket algorithms available to try to do. Algorithms, whose current work is on statics and which did not catch data changes in time, but algorithms not only intended to mine static information, but also provided new ways of calculating modifications on paper-listed association processing rules, and provide new algorithms that could stimulate market engagement and encourage up-selling.

1. Introduction

Currently, the quantity of knowledge being maintained in massive databases, market sectors like retail the banking sector, life science, and so forth however ne'er mind that everyone helpful information for users of that's. Why is the retrieval of the valuable knowledge from a greater volume of data so relevant? The method of extracting knowledge is understood as helpful processing or a discovery data and knowledge process of locating and deciphering the data concerned several steps like, election, preprocessing, transformation processing and interpretation. [1,2,3]. The marketing and distribution business is promoted by data mining. The job of exploitation associate in nursing lysis market basket in research management is carried out by agonies whose market basket analysis additionally known as mining association law. It helps to consider the selling analyst customers, e.g. the commodity that they bought together. There are a number of methods and algorithms used for information analysis.[4] One of the problems for businesses that have produced a range of customer data is that, by collecting the required data from detailed customer knowledge and product attributes, in order to gain a competitive edge, certain forms of customer analysis of the business basket are discussed in academic literature, How to use profiles of buyer preference and interest in particular goods for one-to-one marketing[5], how to transaction trends in a multi-store environment[6] to increase sales. Basket research has been commonly used by many businesses as a way of identifying product partnerships and drawing on the retailer's promotion approach.

Retailers must consider and respond to the demands of shoppers. Business basket research is a possibility so they realize that things are moving together. Business basket research provides marketers with smart insights into related transactions around the core product category from shoppers who usually get bread to the top. Buy a bread-related food, such as milk, butter or jelly. It's good that this squad is terribly side-by-side in a retail hub for consumers to have easy access, and even similar product teams can stand side by side to educate customers.

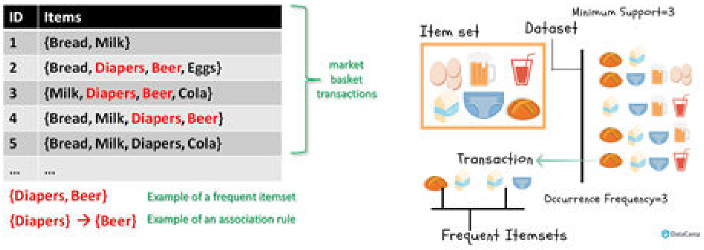
Market basket analysis is one of the techniques for data mining [7] that focuses and figure out what to buy habits by collecting connections or exchanging transaction data from stores. Buy together and customize the architecture of the supermarkets, and even for fashion marketing programs in such a manner that commodity sales can be enhanced, so that the actions of the consumer on the market can be measured by means of entirely different processing techniques.

2. Methodology

Mining rule mappings are used when you want to search for mappings between different fields during a record and often want to find patterns in transactional databases, relational databases, or other information repositories. Applications of mining mapping principles for distribution, data processing baskets (or basketball market analysis) for shopping, clustering and grouping. It will tell you which things consumers sometimes purchase together by developing a set of rules called association rules. In basic terms, it typically offers you a form, such that consumers can extend these principles to a range of sales tactics.

a. Modification the layout of the search by trend
b. Client behavior analysis
c. Catalog style
d. Cross-marketing
e. Trending buying.

Examples of implementations:



ID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}
...	...

market basket transactions
{Diapers, Beer} Example of a frequent itemset
{Diapers} → {Beer} Example of an association rule
Minimum Support=3
Item set
Dataset
Transaction
Occurrence Frequency=3
Frequent Itemsets

Figure 1. Example of Implementations

You can see customer transactions with the numbers one to five. Each transaction shows the item purchased within the transaction, where diapers are purchased in three transactions with beer. Likewise, bread with milk is bought in three transactions, whereby both often appear as fixed items. Association rules are given in the following form:

$$A \Rightarrow B\ [Support\ and\ Confidence]$$

The section before => is given as (previous) and a part after => as (consistent). Where A and B are sets of things in the transaction, data; A and B are most sets. Here is the basic concept of the Association of Mining Rules,

a. Item set, a group of one or more items. K item set means a collection of k items.

b. Support Count, the frequency with which the items are configured.
a. Supports, Transaction fraction that contains the item-set 'X'

$$Support(X) = \frac{frequency(X)}{N}$$

Rule A => B the support is provided by, P (AUB) is expected to arise along with A and B. P indicates probability

$$Support(A => B) = \frac{frequency(A,B)}{N}$$

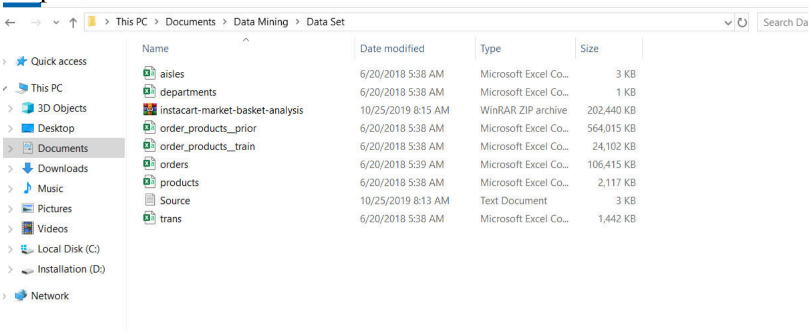
Confidence (c): for rule A => B the confidence indicates the proportion during which B buys with,

$$Confidence(A => B) = \frac{P(A \cap B)}{P(A)} = \frac{frequency(A,B)}{frequency(A)}$$

The purpose of the Association rules after applying the Association to the mining rules in a set selected from T transactions, your goal is to find all the bases using:
a. Major support or until min_support
b. Confidence higher or up to min_confidence

2.1 Dataset

This paper is based on instacart datasets from https://www.kaggle.com/. It consists of 131209 orders and 39123 different products.



This PC > Documents > Data Mining > Data Set

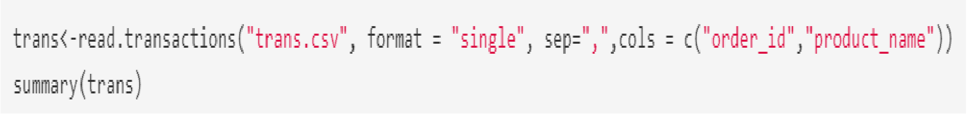
Name	Date modified	Type	Size
aisles	6/20/2018 5:38 AM	Microsoft Excel Co...	3 KB
departments	6/20/2018 5:38 AM	Microsoft Excel Co...	1 KB
instacart-market-basket-analysis	10/25/2019 8:15 AM	WinRAR ZIP archive	202,440 KB
order_products__prior	6/20/2018 5:38 AM	Microsoft Excel Co...	564,015 KB
order_products__train	6/20/2018 5:38 AM	Microsoft Excel Co...	24,102 KB
orders	6/20/2018 5:39 AM	Microsoft Excel Co...	106,415 KB
products	6/20/2018 5:38 AM	Microsoft Excel Co...	2,117 KB
Source	10/25/2019 8:13 AM	Text Document	3 KB
trans	6/20/2018 5:38 AM	Microsoft Excel Co...	1,442 KB

Figure 2. Data Set

a. Analysis of the Instacart market basket. What items the Instacart customer will purchase again?
b. Definition Instacart, a grocery buying and delivery service, is designed to make it convenient to fill refrigerators and pantry with your personal favorites and essentials when you need them. Personal shoppers can review your purchases and shop with you for things from the Instacart App. Instacart's data science unit plays a vital role in creating this enjoyable shopping experience. They are using purchasing data to create a model that the customers can purchase goods again, for the first time or for the first time., after the session, add them to their next basket. Instacart recently opened this information check for 3,000,000 Instacart commands in their web registry article, open source.

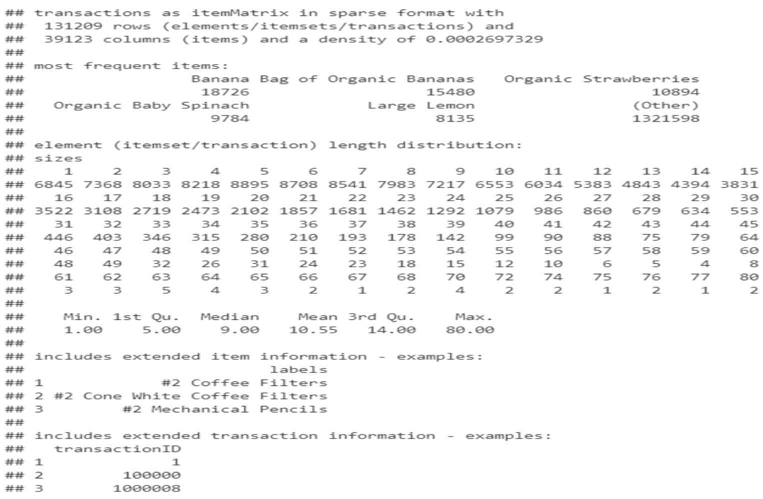
3. Result and Discussion

3.1 Discussion



```
trans<-read.transactions("trans.csv", format = "single", sep=",",cols = c("order_id","product_name"))
summary(trans)
```

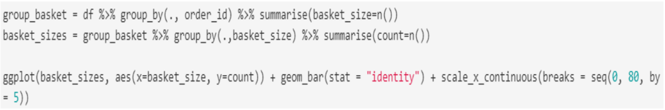
From the dataset summary, basic information can be obtained. The first most frequent item is,



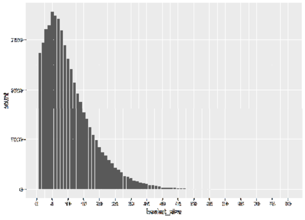
```
## transactions as itemMatrix in sparse format with
##  131209 rows (elements/itemsets/transactions) and
##  39123 columns (items) and a density of 0.0002697329
##
## most frequent items:
##                 Banana Bag of Organic Bananas   Organic Strawberries
##                  18726                  15480                  10894
##   Organic Baby Spinach            Large Lemon                (Other)
##                   9784                   8135                1321598
##
## element (itemset/transaction) length distribution:
## sizes
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15
## 6845 7368 8033 8218 8895 8708 8541 7983 7217 6553 6034 5383 4843 4394 3831
##   16   17   18   19   20   21   22   23   24   25   26   27   28   29   30
## 3522 3108 2719 2473 2102 1857 1681 1462 1292 1079  986  860  679  634  553
##   31   32   33   34   35   36   37   38   39   40   41   42   43   44   45
##  446  403  346  315  280  210  193  178  142   99   90   88   75   79   64
##   46   47   48   49   50   51   52   53   54   55   56   57   58   59   60
##   48   49   32   26   31   24   23   18   15   12   10    6    5    4    8
##   61   62   63   64   65   66   67   68   70   72   74   75   76   77   80
##    3    3    5    4    3    2    1    2    4    2    2    1    2    1    2
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    5.00    9.00   10.55   14.00   80.00
##
## includes extended item information - examples:
##                        labels
## 1            #2 Coffee Filters
## 2 #2 Cone White Coffee Filters
## 3        #2 Mechanical Pencils
##
## includes extended transaction information - examples:
##   transactionID
## 1             1
## 2        100000
## 3       1000008
```

a. Bananas
b. Bags from Organic Bananas,
c. Organic Strawberries,
d. Organic Baby Spinach,
e. Big Lemon.

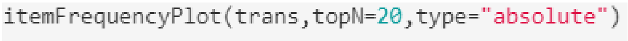
The distribution of basket sizes can be analyzed. In addition, that distribution plot can be useful. The most frequent basket is a size 5 and the average size equals 10.6.



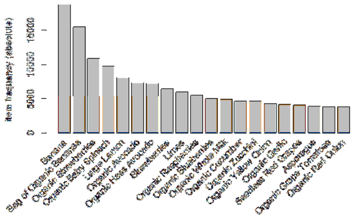
```
group_basket = df %>% group_by(., order_id) %>% summarise(basket_size=n())
basket_sizes = group_basket %>% group_by(.,basket_size) %>% summarise(count=n())

ggplot(basket_sizes, aes(x=basket_size, y=count)) + geom_bar(stat = "identity") + scale_x_continuous(breaks = seq(0, 80, by
= 5))
```

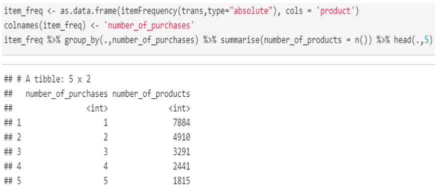


As it says 'banana' is the product most frequently purchased by customers. It is possible to plot topN items most often.



```
itemFrequencyPlot(trans,topN=20,type="absolute")
```





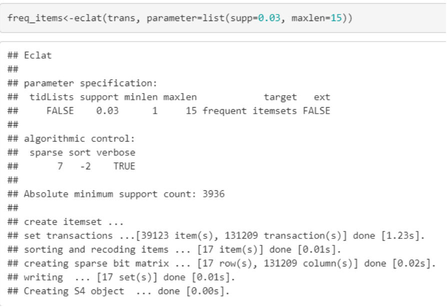
```
item_freq <- as.data.frame(itemFrequency(trans,type="absolute"), cols = 'product')
colnames(item_freq) <- 'number_of_purchases'
item_freq %>% group_by(.,number_of_purchases) %>% summarise(number_of_products = n()) %>% head(.,5)
```

```
## # A tibble: 5 x 2
##   number_of_purchases number_of_products
##                 <int>              <int>
## 1                   1               7884
## 2                   2               4910
## 3                   3               3291
## 4                   4               2441
## 5                   5               1815
```

Can be seen the most popular products, but to know all the products that are rarely purchased is difficult. The table shows that there are 7884 products purchased only once, 4910 products purchased twice and so on.

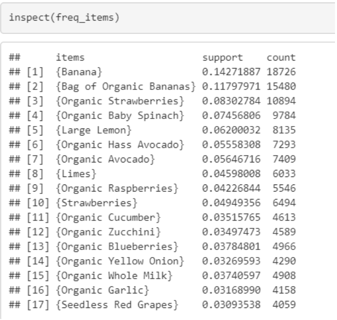
3.2 Rules

The rules are based on support and confidence levels, so we have to determine the level of those Statistics. We need to do it to be able to analyze most often the rules/patterns. First, by using the Eclat algorithm the most frequent set items will be displayed. The default support is 0.1 but in this dataset a lower value is required to get any results.



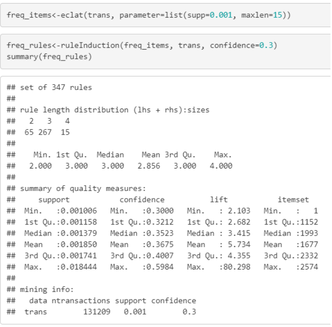
```
freq_items<-eclat(trans, parameter=list(supp=0.03, maxlen=15))
```

```
## Eclat
##
## parameter specification:
##  tidLists support minlen maxlen            target   ext
##     FALSE    0.03      1     15 frequent itemsets FALSE
##
## algorithmic control:
##  sparse sort verbose
##       7   -2    TRUE
##
## Absolute minimum support count: 3936
##
## create itemset ...
## set transactions ...[39123 item(s), 131209 transaction(s)] done [1.23s].
## sorting and recoding items ... [17 item(s)] done [0.01s].
## creating sparse bit matrix ... [17 row(s), 131209 column(s)] done [0.02s].
## writing  ... [17 set(s)] done [0.01s].
## Creating S4 object  ... done [0.00s].
```



```
inspect(freq_items)
```

```
##      items                      support    count
## [1]  {Banana}                   0.14271887 18726
## [2]  {Bag of Organic Bananas} 0.11797971 15480
## [3]  {Organic Strawberries}     0.08302784 10894
## [4]  {Organic Baby Spinach}     0.07456806  9784
## [5]  {Large Lemon}              0.06200032  8135
## [6]  {Organic Hass Avocado}     0.05558308  7293
## [7]  {Organic Avocado}          0.05646716  7409
## [8]  {Limes}                    0.04598008  6033
## [9]  {Organic Raspberries}      0.04226844  5546
## [10] {Strawberries}             0.04949356  6494
## [11] {Organic Cucumber}         0.03515765  4613
## [12] {Organic Zucchini}         0.03497473  4589
## [13] {Organic Blueberries}      0.03784801  4966
## [14] {Organic Yellow Onion}     0.03269593  4290
## [15] {Organic Whole Milk}       0.03740597  4908
## [16] {Organic Garlic}           0.03168990  4158
## [17] {Seedless Red Grapes}      0.03093538  4059
```



```
freq_items<-eclat(trans, parameter=list(supp=0.001, maxlen=15))
```

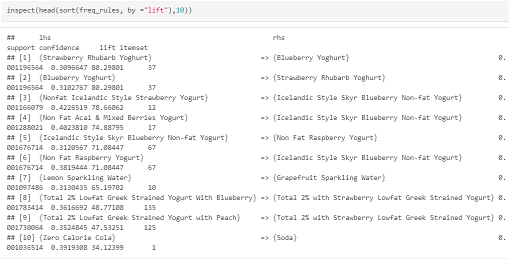
```
freq_rules<-ruleInduction(freq_items, trans, confidence=0.3)
summary(freq_rules)
```

```
## set of 347 rules
##
## rule length distribution (lhs + rhs):sizes
##   2   3   4
##  65 267  15
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.000   3.000   3.000   2.856   3.000   4.000
##
## summary of quality measures:
##     support            confidence          lift            itemset
##  Min.   :0.001006   Min.   :0.3000   Min.   : 2.103   Min.   :   1
##  1st Qu.:0.001158   1st Qu.:0.3212   1st Qu.: 2.682   1st Qu.:1152
##  Median :0.001379   Median :0.3523   Median : 3.415   Median :1993
##  Mean   :0.001850   Mean   :0.3675   Mean   : 5.734   Mean   :1677
##  3rd Qu.:0.001741   3rd Qu.:0.4007   3rd Qu.: 4.355   3rd Qu.:2332
##  Max.   :0.018444   Max.   :0.5984   Max.   :80.298   Max.   :2574
##
## mining info:
##   data ntransactions support confidence
##  trans        131209   0.001        0.3
```

The most frequent set of items is just one basket item. In this dataset with a minimum support value of 0.03 no basket contains at least two different items. The next step is to recognize the rules most often. To get a rule, the support value must be lower to get a set item of at least two items. There are 347 rules, from which 65 is a size 2 (LHS is one product and RHS is one product), 267 is a size 3 (LHS is two items) and 15 out of a size 4 (LHS is three items). Support means equal to 0.0018 and means confidence for 0.36. Lift average equals 5.73.

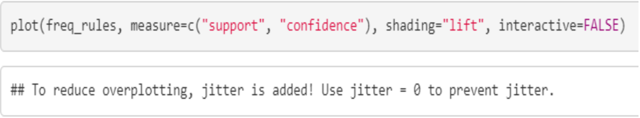
3.3 Top Rules

Rules with the highest lift value will be evaluated.



```
inspect(head(sort(freq_rules, by ="lift"),10))
```

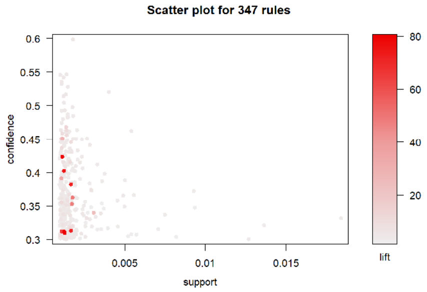
```
##      lhs                                                   rhs
support confidence     lift itemset
## [1]  {Strawberry Rhubarb Yoghurt}                       => {Blueberry Yoghurt}                                0.
001196564  0.3096647 80.29801      37
## [2]  {Blueberry Yoghurt}                                => {Strawberry Rhubarb Yoghurt}                       0.
001196564  0.3102767 80.29801      37
## [3]  {Nonfat Icelandic Style Strawberry Yogurt}         => {Icelandic Style Skyr Blueberry Non-fat Yogurt}    0.
001166079  0.4226519 78.66062      12
## [4]  {Non Fat Acai & Mixed Berries Yogurt}              => {Icelandic Style Skyr Blueberry Non-fat Yogurt}    0.
001288021  0.4023810 74.88795      17
## [5]  {Icelandic Style Skyr Blueberry Non-fat Yogurt}    => {Non Fat Raspberry Yogurt}                         0.
001676714  0.3120567 71.08447      67
## [6]  {Non Fat Raspberry Yogurt}                         => {Icelandic Style Skyr Blueberry Non-fat Yogurt}    0.
001676714  0.3819444 71.08447      67
## [7]  {Lemon Sparkling Water}                            => {Grapefruit Sparkling Water}                       0.
001097486  0.3130435 65.19702      10
## [8]  {Total 2% Lowfat Greek Strained Yogurt With Blueberry} => {Total 2% with Strawberry Lowfat Greek Strained Yogurt} 0.
001783414  0.3616692 48.77108     135
## [9]  {Total 2% Lowfat Greek Strained Yogurt with Peach}     => {Total 2% with Strawberry Lowfat Greek Strained Yogurt} 0.
001730064  0.3524845 47.53251     125
## [10] {Zero Calorie Cola}                                => {Soda}                                             0.
001036514  0.3919308 34.12399       1
```

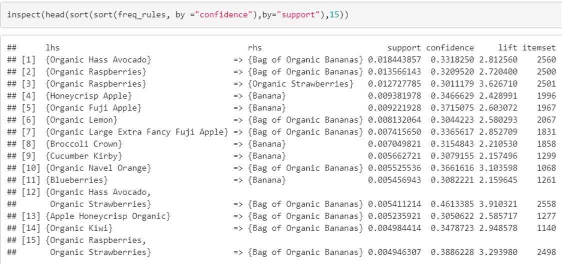


```
plot(freq_rules, measure=c("support", "confidence"), shading="lift", interactive=FALSE)
```

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

In the top 10 the lift value rule is huge but support for each rule is well below 1%. This means that baskets consisting of products within certain rules are rare cases. Trust 0.3-0.4 indicates that such products in LHS are often purchased with those in the RHS. In this case the support and value of trust is very important, the rules with the highest lift are some rare combinations of products. Planning all the rules in terms of support, confidence and lift is possible.

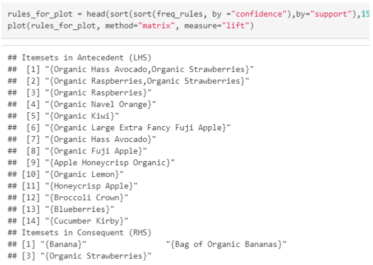




```
inspect(head(sort(sort(freq_rules, by ="confidence"),by="support"),15))
```

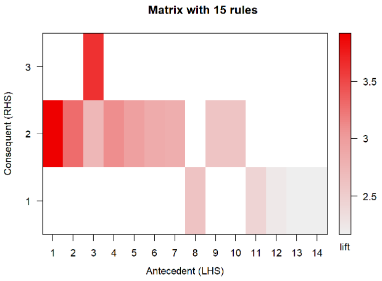
```
##      lhs                                     rhs                      support confidence     lift itemset
## [1]  {Organic Hass Avocado}               => {Bag of Organic Bananas} 0.018443857  0.3318250 2.812560    2560
## [2]  {Organic Raspberries}                => {Bag of Organic Bananas} 0.013566143  0.3209520 2.720400    2500
## [3]  {Organic Raspberries}                => {Organic Strawberries}   0.012727785  0.3011179 3.626710    2501
## [4]  {Honeycrisp Apple}                   => {Banana}                 0.009381978  0.3466629 2.428991    1996
## [5]  {Organic Fuji Apple}                 => {Banana}                 0.009221928  0.3715075 2.603072    1967
## [6]  {Organic Lemon}                      => {Bag of Organic Bananas} 0.008132064  0.3044223 2.580293    2067
## [7]  {Organic Large Extra Fancy Fuji Apple} => {Bag of Organic Bananas} 0.007415650  0.3365617 2.852709    1831
## [8]  {Broccoli Crown}                     => {Banana}                 0.007049821  0.3154843 2.210530    1858
## [9]  {Cucumber Kirby}                     => {Banana}                 0.005662721  0.3079155 2.157496    1299
## [10] {Organic Navel Orange}               => {Bag of Organic Bananas} 0.005525536  0.3661616 3.103598    1068
## [11] {Blueberries}                        => {Banana}                 0.005456943  0.3082221 2.159645    1261
## [12] {Organic Hass Avocado,
##       Organic Strawberries}               => {Bag of Organic Bananas} 0.005411214  0.4613385 3.910321    2558
## [13] {Apple Honeycrisp Organic}           => {Bag of Organic Bananas} 0.005235921  0.3050622 2.585717    1277
## [14] {Organic Kiwi}                       => {Bag of Organic Bananas} 0.004984414  0.3478723 2.948578    1140
## [15] {Organic Raspberries,
##       Organic Strawberries}               => {Bag of Organic Bananas} 0.004946307  0.3886228 3.293980    2498
```

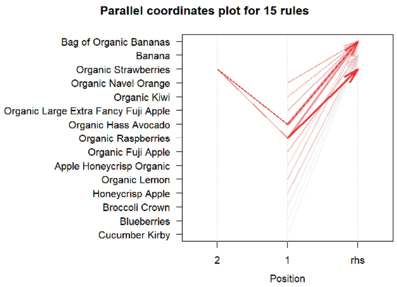
On the plot it is clearly seen that the majority of the rules have low support. As analyzed, the rule with the highest lift value has support well below 1%. To get rules that often appear in baskets and are purchased together, the top rules are sorted by support and confidence. There are some interesting rules. Basically, we can assume that buying one organic product leads to buying another organic product. In most rules the most frequent items appear. Rules can also be plotted as matrices, where we have LHS on x axes and RHS on the y-axis.



```
rules_for_plot = head(sort(sort(freq_rules, by ="confidence"),by="support"),15
plot(rules_for_plot, method="matrix", measure="lift")
```

```
## Itemsets in Antecedent (LHS)
##  [1] "{Organic Hass Avocado,Organic Strawberries}"
##  [2] "{Organic Raspberries,Organic Strawberries}"
##  [3] "{Organic Raspberries}"
##  [4] "{Organic Navel Orange}"
##  [5] "{Organic Kiwi}"
##  [6] "{Organic Large Extra Fancy Fuji Apple}"
##  [7] "{Organic Hass Avocado}"
##  [8] "{Organic Fuji Apple}"
##  [9] "{Apple Honeycrisp Organic}"
## [10] "{Organic Lemon}"
## [11] "{Honeycrisp Apple}"
## [12] "{Broccoli Crown}"
## [13] "{Blueberries}"
## [14] "{Cucumber Kirby}"
## Itemsets in Consequent (RHS)
## [1] "{Banana}"                 "{Bag of Organic Bananas}"
## [3] "{Organic Strawberries}"
```





Another way to plot rules and make them more affordable to analyze is Parallel coordinates plots. It shows for example that if the client has been in the basket of 'Organic strawberries' and 'Organic Hass avocados' he tends to buy 'Bag of Organic Bananas'.

4. Conclusion

Analysis can be used for better product placement. At first, the strongest rule is found, but it is rather unusual. Analysis is also carried out on three Products:

a. 'Bananas' and 'Bag of Organic Bananas' are the two most frequent items. The rules indicate that the two products are mostly purchased with other organic vegetables or fruits. This indicates that such products should be placed very close to other fruits and vegetables in the store.
b. 'Zero Calorie Cola' is mostly purchased with some kind of snack. This clearly shows that stores can offer for example Cola packages and chips to sell more of these products.
c. Market Basketball Analysis is a powerful tool to gain a better knowledge of customer behavior. This can help stores to increase cross-sell and become more profitable.

References

[1] Raorane AA, Kulkarni RV, Jitkar BD. Association Rule – Extracting Knowledge Using Market Basket Analysis.Research Journal of RecentSciences 2012:1(2):19-27.
[2] Majeed HEYDARI, Amir YOUSEFLI. A new optimization model for market basket analysis with allocation considerations: A genetic algorithm solution approach.Management & Marketing. Challenges for the Knowledge Society 2017:12(1).
[3] B.Suvarnamukhi, M.Seshashayee. Big Data Concepts and Techniques in Data Processing. International Journal of Computer Sciences and Engineering 2018:6(10).
[4] Herman A, Forcum LE, Joo Harry. Using Market Basket Analysis in Management Research.Journal of Management 2013:39(7):1799-1824.
[5] Weng, S.-S., Liu, J.-L.: Feature-based recommendations for one-to-one marketing, Expert Systems with Applications, Vol. 26, 2004, pp. 493-508.
[6] Chen, Y.-L., Tang, K., Shen, R.-J., Hu, Y.-H.: Market basket analysis in a very multiple store environment, Decision Support Systems, 2004.
[7] Berry, M.J.A., Linoff, G.S.: data processing Techniques: for Marketing, Sales and Customer Relationship Management (second edition), Hungry Minds Inc., 2004