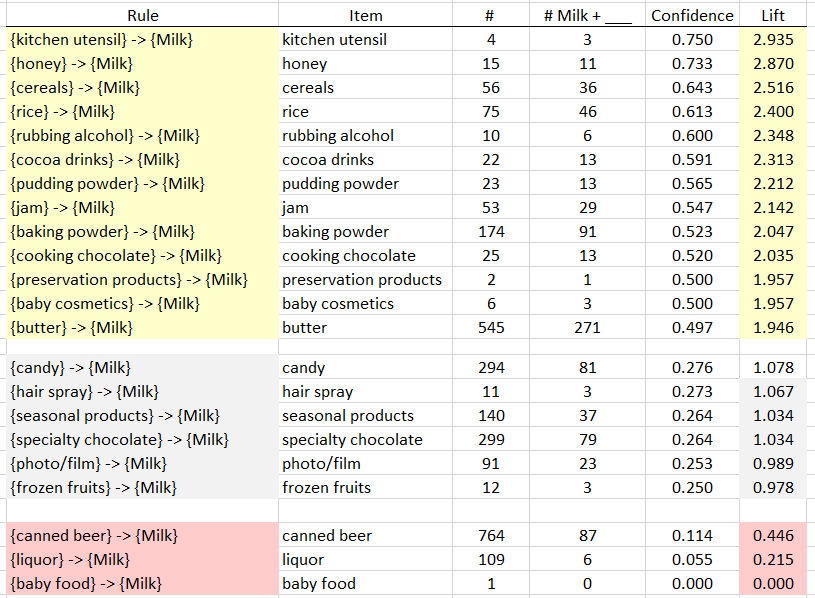
**7 - Association Rules** (also known as *Market Basket Analysis*)

**7.1 – Introduction to Association Rules**

Market Basket Analyses are a common application of association rules. One goal of a market basket analysis is to understand the association between items purchases. The relationship between items purchased at a grocery store will be considered in this handout.

|  |  |
| --- | --- |
| http://imagesus.homeaway.com/mda01/f4e2379a-d720-459f-b9c0-9999fc559022.1.6 |  |

An association rule highlights the fact that some items are more (or less) indicative of the purchase of others. For example, purchasing cereal increases the likelihood of purchasing milk. These types of analyses may also reveals that liquor and milk are rarely purchased together.



**Association Rules** are used to uncover associations or relationships that exist between items. Often these rules are constructed to identify relationships between items purchased, i.e. Market Basket Analysis.

Procedural Steps

* 1. Determine how often items are purchased.
  2. Determine how often items are purchased in conjunction with other items.
  3. Identify which purchased items are indicative of others being purchased.

Data Technologies

* 1. Filtering/Subsets
  2. Creating Tables
  3. Data Summaries

Example 7.1: Grocery Store Transactions (small subset)

Consider the following subset of data from a collection of transactions from a grocery store.

|  |  |
| --- | --- |
| Transaction ID | Items Purchased |
| 1 | {Bread, Milk} |
| 2 | {Eggs, Ham} |
| 3 | {Bread, Fruit, Milk} |
| 4 | {Beer, Bread, Butter, Fruit, Soda} |
| 5 | {Bread, Fruit, Milk, Soda} |

Association rules are developed under the following guiding principles.

|  |  |  |
| --- | --- | --- |
| 1. | Items should be purchased somewhat often. | **Support** |
| 2. | Reliability, i.e. the degree to which one set of items predicts the purchase of another set of items. | **Confidence** |

**7.2 – Itemsets, Rules, Support, Confidence, and Lift**

An ***itemset*** is a collection of items that could potentially be in the same market basket/transaction. An itemset can consist of a single item.

Examples: {Bread, Fruit, Milk}

A ***rule*** is simply put says that buying itemset implies they will also buy itemset . Are goal in conducting a market basket analysis is to identify rules where buying itemset also implies you are “likely” to buy itemset . Rules are denoted using the notation: .

Examples: {Bread} 🡪 {Milk}

{Bread,Ham} 🡪 {Cheese, Mayo}

The ***support*** of a set of items, or rule involving items, (e.g. milk AND bread) is the estimated probability of simultaneously observing both itemsets in a randomly selected market basket/transaction. If and are itemsets, in standard probability notation the support = . The support of a rule can be denoted .

The ***confidence***of a rule is the probability of buying item set given the basket already contains itemset . In standard probability notation . It can also be defined in terms of support as follows:

Example 7.1: Grocery Store Transactions (small subset)

Consider the following association rule – the purchase of Bread indicates the purchase of Milk.

|  |  |
| --- | --- |
| Rule #1 |  |

Compute the support and confidence for this rule.

Questions

1. What is the interpretation of the ?

60% of the costumer bought Bread and Milk

1. What is the interpretation of Confidence of this rule? Discuss.  
   Note: Remember Confidence is simply a conditional probability, i.e P(Milk | Bread).

Knowing that a person had Bread in their cart, what is the chance they also have milk? 75%

Consider a second association rule for the purchase of Milk.

|  |  |
| --- | --- |
| Rule #2 |  |

Compute the support and confidence for this rule.

Question

1. Why might Rule #1 be considered “better” than Rule #2 when interest lies in the purchase of Milk?

Rule 1 has more support,

Consider a third association rule for the purchase of Milk.

|  |  |
| --- | --- |
| Rule #3 |  |

Compute the support and confidence for this rule.

2/5

2/5 / 3/5 = 66%

***Lift*** is another measure often considered when evaluating rules of association. It measures how many times more likely you are to buy itemset given you have bought itemset to the probability you buy without any knowledge of what else is in the market basket/transaction.

Interpretation of Lift

* implies positive association between items
* implies no association between items
* implies negative association between items

We will now compute the lift of the three rules developed above.

Rule #1:

Rule #2:

Rule #3:

For our examples, realize that the support for Milk is fairly large, i.e. Milk was purchased in 60% of all transactions. This provides a baseline value for confidence. That is, rules that exceed this value indicate gains when considering the association provided by the rule. When the lift of a rule is near 1, then the rule provides little information to understanding the purchase of the item.

|  |  |  |  |
| --- | --- | --- | --- |
| **Rule** | **Support** | **Confidence** | **Lift** |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Some Comments

* Association rules with no support have zero confidence. E.g. Beer is never purchased with Milk, so the rule should not be considered.
* The confidence of a rule should not be considered independent of it’s support. For example, the rule has Confidence = 1. That is, 100% of the time eggs were purchased, so was Ham. However, this rule has very low support as Eggs and Ham were only purchased together once.
* Association rules are not invariant. For example, the confidence for the rule is different than the confidence of the rule . In association rule terminology the itemset before the arrow is called the *antecedent* and the tiem following the arrow is called the *consequent*. Which itemset plays which role matters!

**7.3 – Transactional Databases**

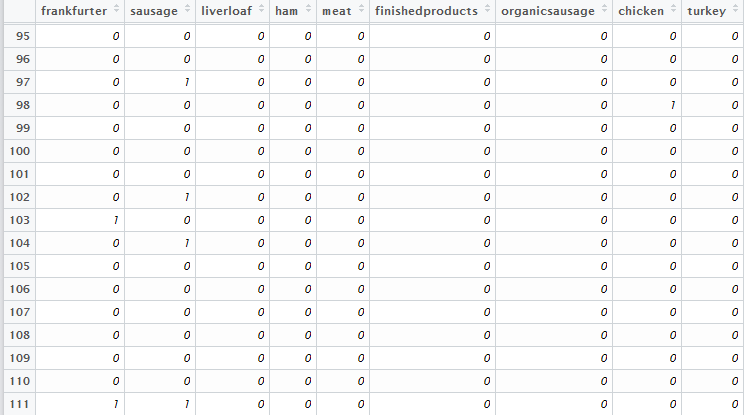
A transaction will consist of a list of all items purchased in that transaction. For purposes of finding “good” or “meaningful” association rules however we need to create a binary representation matrix for the set of all transactions. The matrix will have one column for each possible item purchased and each row of the matrix will correspond to a single transaction. For each item in the transaction we place a 1 in appropriate column and a zero in all columns corresponding to items not purchased in the transaction.

Common Data Structure

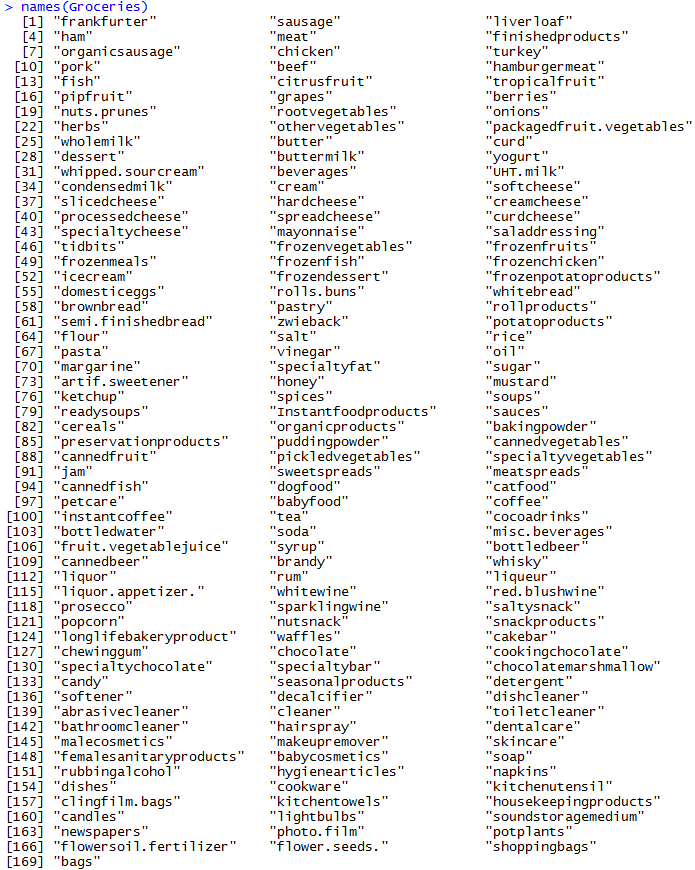
|  |  |  |
| --- | --- | --- |
| List |  | Binary Representation (Matrix) |

Example 7.2: Grocery Store Transactions (full transactional database)

Next, let us consider the complete grocery store transactional database. This dataset contains one month (30 days) of real-world point-of-sale transaction data from grocery store. These data are contained in the data frame Groceries in the arules package. This transactional database contains transactions and there were a total of 169 unique items purchased amongst all of these transactions.

The Groceries data frame consists of the binary representation of this market basket data. Re  


> Groceries.csv = read.table(file.choose(),header=T,sep=”,”) 🡨 read in the file **Groceries.txt**



The column names of the Groceries data frame denote the 169 unique items that were purchased amongst the 9835 transactions. We will now look for rules using a single item antecedent that have as the consequent.

e.g.

Using simple logic statements on the columns of Whole Milk and Butter allows one to easy compute the support and confidence for the rule: .

|  |  |  |  |
| --- | --- | --- | --- |
| |  |  | | --- | --- | | Rule |  |   > sum(Groceries$butter==1 & Groceries$wholemilk==1)  [1] 271  > nrow(Groceries)  [1] 9835  > sum(Groceries$butter==1)  [1] 545  > sum(Groceries$wholemilk==1)  [1] 2513 | |
|  |  |
|  | |

Evaluating All Potential Single Item Rules for Whole Milk  
The procedure provided above lack efficiencies and does not scale well when several rules need to be evaluated. For example, there are 169 items in this analysis, a more efficient process would be to write a loop to cycle through all columns automatically. This is the purpose of the code provided below.

|  |  |
| --- | --- |
| Rule #1 |  |
| Rule #2 |  |
| : | : |
| Rule #169 |  |

First we set up a data frame called wholemilk.rules to store the results in a table with meaningful column headings.

> wholemilk.rules = data.frame(Item=rep(NA,169),Numberi=rep(NA,169),

Numberi\_Wholemilk=rep(NA,169),Support=rep(NA,169),Confidence=rep(NA,169),

Lift=rep(NA,169))

Compute the support of the itemset which will be the consequent of the rules we are constructing.

> support.wholemilk = sum(Groceries[,25]==1)/length(Groceries[,25])

> support.wholemilk

[1] 0.255516 🡨 chance they unconditionally   
 purchase whole milk.

The loop below will extract the item name of the antecedent, compute the support of the rule , compute the support of the antecedent item, compute the confidence of the rule, i.e. , and finally compute the lift of the rule, i.e. .

> for (i in 1:169){

wholemilk.rules[i,1] = colnames(Groceries)[i]

if (i != 25){

support\_i\_wholemilk=sum(Groceries[,i]==1 & Groceries[,25]==1)/length(Groceries[,i])

support\_i = sum(Groceries[,i]==1)/length(Groceries[,i])

confidence = support\_i\_wholemilk/support\_i

lift = confidence/support.wholemilk

wholemilk.rules[i,2]=sum(Groceries[,i]==1)

wholemilk.rules[i,3]=sum(Groceries[,i]==1 & Groceries[,25]==1)

wholemilk.rules[i,4]=support\_i\_wholemilk

wholemilk.rules[i,5]=confidence

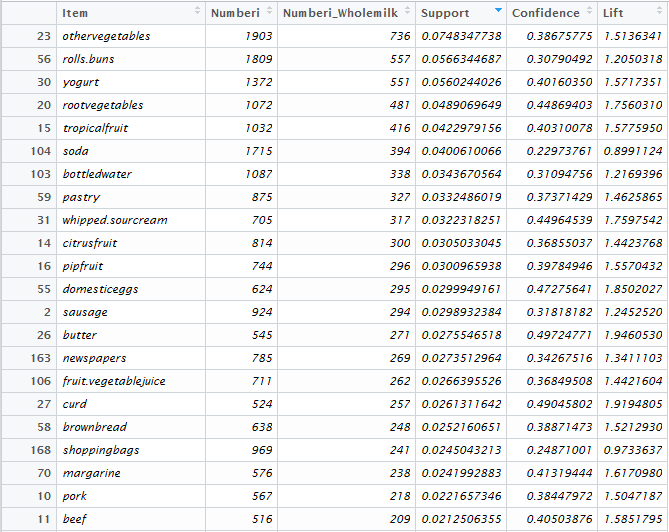
wholemilk.rules[i,6]=lift

}

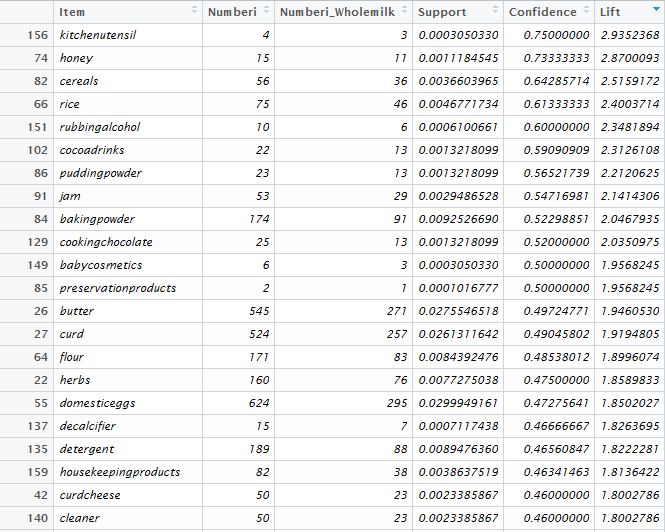
}

> View(wholemilk.rules)

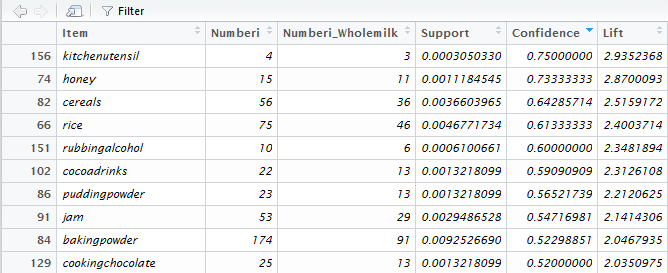
Rules sorted by Support (decreasing)



Rules sorted by Confidence/Lift (decreasing)

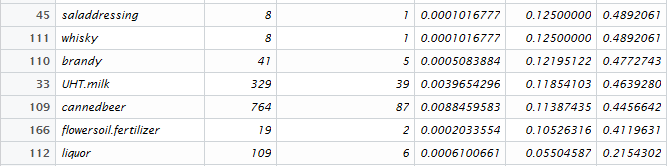


The following items have high lift, , – indicating if these items were purchased they are at least twice as likely to buy milk than if we did not know they had purchased these items.



For example, given that we know a customer has purchased cereals the probability they bought whole milk is .6429 or a 64.29% chance. Compare this to probability that a randomly selected customer buys whole milk which is 25.55% chance. The lift is the ratio of these two probabilities which is 2.52 for cereals. Thus buying cereals increases a customer’s likelihood of buying whole milk a 2.5 times.

The following items have a lift value much smaller than 1, i.e. – indicating that customers purchasing items are less likely to buy whole milk.



For example, customers that purchase liquor are about one-fifth as likely to also buy whole milk when compared to the general population of customers. If a customer purchases liquor the probability they also buy milk is .055 or a 5.5% chance.

Questions

1. As stated above the Lift for is about 2.5 which is fairly high. Thus, given that the transaction includes cereal, there is 2.5 fold increase in the likelihood of whole milk being purchased.
   1. What is the ?
   2. This value seems fairly low. Why does a low support value negate the usefulness of a rule?
   3. What rules seems to have a fairly high lift AND have a relatively large support?
2. The Lift value for the rule is amongst the lowest on this list. What can be said about the purchase of Canned Beer AND Whole Milk?

Clearly the approach above will not work in general as there are far too many rule combinations that we may wish to explore!! Thus we will use two R packages that have been developed to sort through “all” possible rules and identify those that have desirable properties, i.e. a balance between a reasonable support and confidence/lift. In the next section we will begin using functions in the arules and arulesViz packages for this purpose.