Marketing and Retail Analytics

Contents

[1 Introduction 1](#_Toc52132444)

[2 Exploratory Data Analysis 2](#_Toc52132445)

[2.1 Converting single format data into transactions 2](#_Toc52132446)

[2.2 Checking frequency of items 4](#_Toc52132447)

[3 Association Rules – Apriori Recommendation Algorithm 7](#_Toc52132448)

[3.1 Generating Association Rules 7](#_Toc52132449)

[3.2 Pruning the generated Association Rules 10](#_Toc52132450)

[3.3 Analysing top rules by lift 11](#_Toc52132451)

[4 Visualizing Association Rules 12](#_Toc52132452)

[4.1 Scatter Plot of pruned association rules 12](#_Toc52132453)

[4.2 Top rules based on Lift 13](#_Toc52132454)

[4.3 Top rules based on Confidence 15](#_Toc52132455)

[4.4 Top rules based on Support 16](#_Toc52132456)

[5 Observation based on Market Basket Analysis 17](#_Toc52132457)

[6 Actionable Insights and Recommendations 19](#_Toc52132458)

# Introduction

The Groceries Dataset is provided which contains the items purchased in each transaction of a store. The problem requires us to create association rules and perform Market Basket Analysis on the given dataset in order to recommend the goods that can be bundled together.

The R Code can be accessed [here](https://drive.google.com/file/d/1w93Nlf6MIxeh4i_yorBA5_j6aiVJI7q8/view?usp=sharing).

# Exploratory Data Analysis

## Converting single format data into transactions

The dataset which is provided is in basket format where each row represents a transaction with a list of items purchased in that transaction. We need to convert it into an object of transaction class before we proceed to use the data for creating association rules.

The items will be separated by delimiter ‘,’

# load the data:

#1.1 Converting single format data set into transactions after cleaning data

#following function reads the basket format data

groceries\_1 <- read.transactions('groceries.csv',rm.duplicates= TRUE,format ="basket", sep = ",")

#Performing EDA to acertian rules:

length(groceries\_1)

# View doesnt work in this case and hence we use inspect to view the data

inspect(groceries\_1[1])

# Checking the first 10 transaction to check for format issues or check for data loss

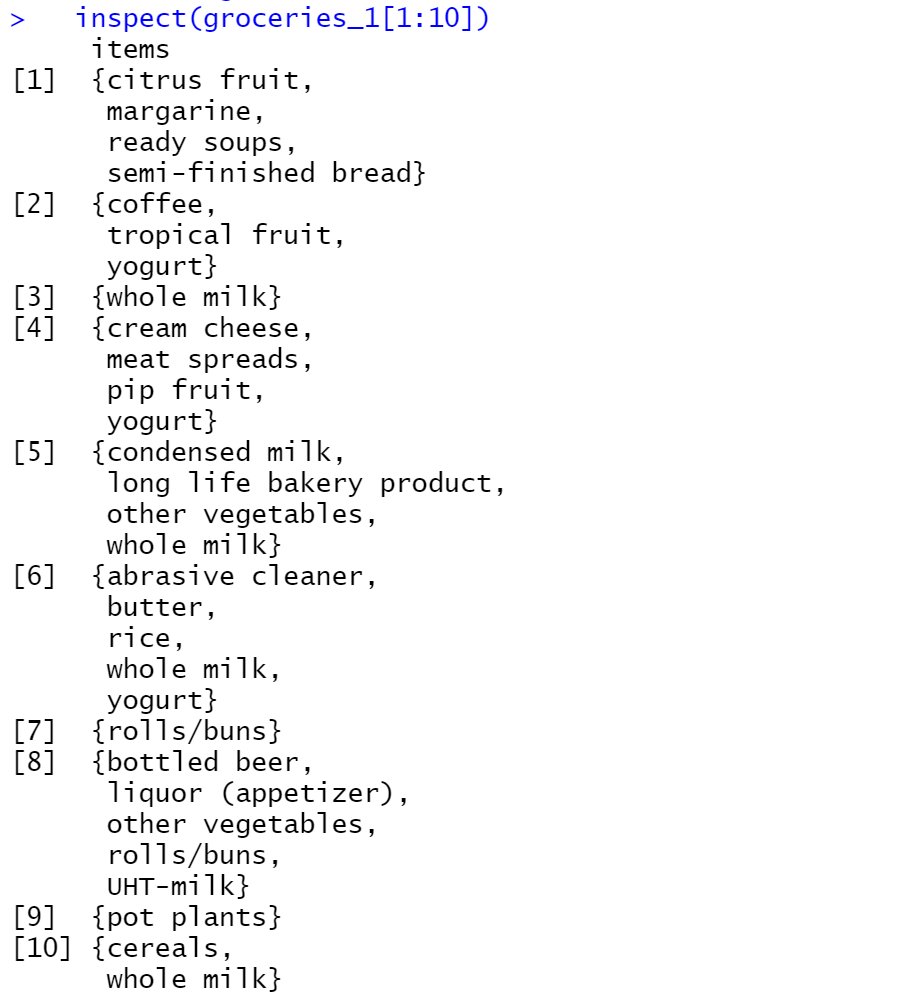
inspect(groceries\_1[1:10])

#indicates the number of items in each basket referred to as length

size(groceries\_1)

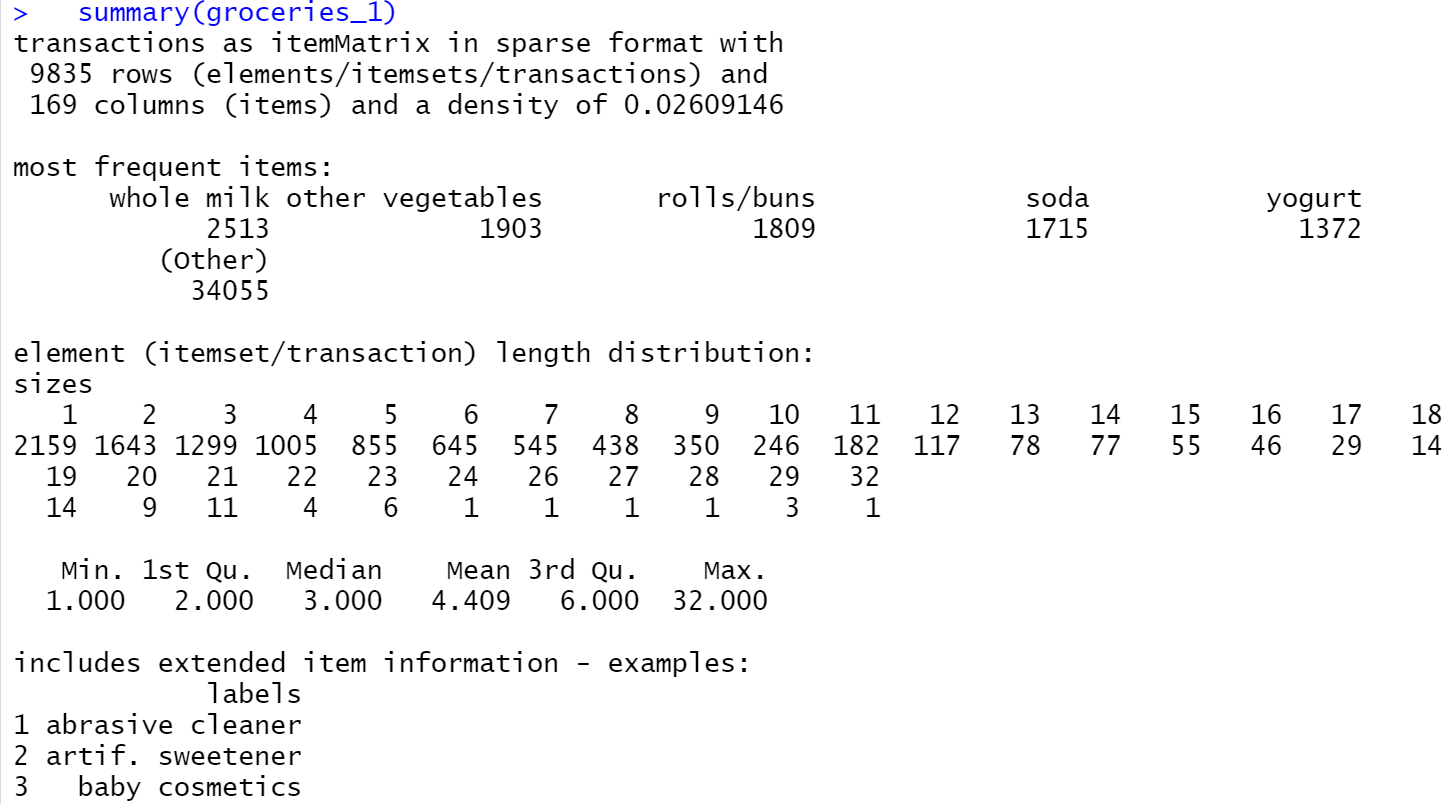
The dataset consists of 9835 transactions. The data is in the form of transaction object, hence “inspect” function is being used instead of “view” function to display first 10 transactions and check for format issues or data loss.

The first basket contains Citrus fruit, margarine, Ready Soups & Semi-finished bread.



The size function displays the number of items in each basket referred to as length, which ranges from 1 to 32 items in 1 basket.

Let us check the summary to get further understanding of the data.



As seen from the summary, there are 9835 transactions and 169 products involved in the dataset. The most frequently purchased products are:

* Whole Milk
* Other vegetables
* Rolls/Buns
* Soda
* Yogurt

The element length distribution demonstrates how many transactions had 1 item in the basket, 2 items in the basket and so on. In the given dataset, there are 2159 transactions which have only one item in the basket. The average basket size is of 4 items, whereas some baskets have as high as 32 items.

## Checking frequency of items

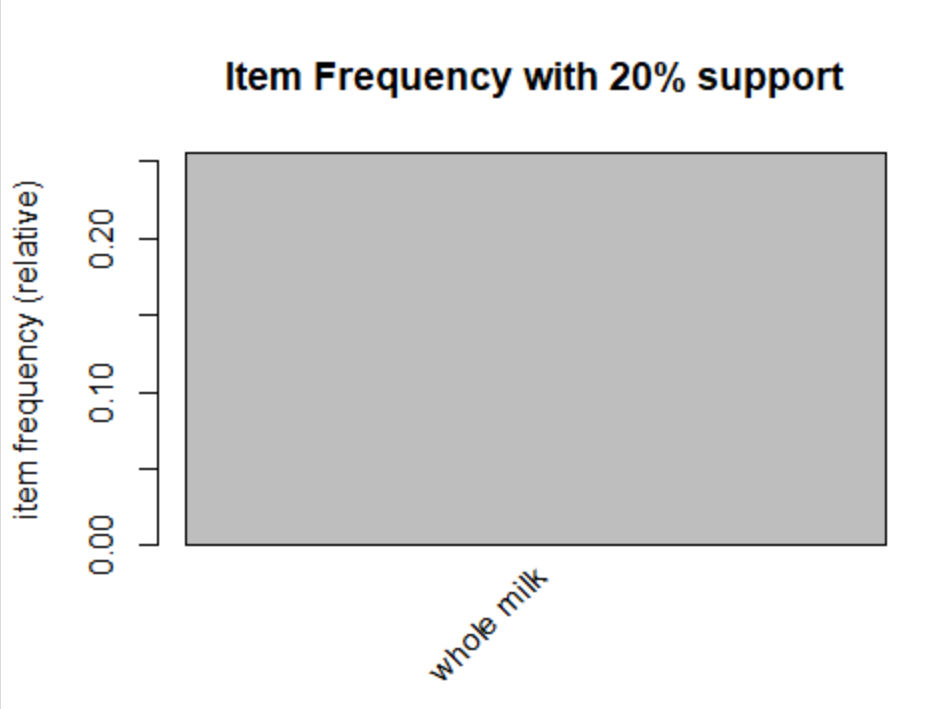
Let us generate an Item Frequency Plot to understand the distribution of items.

#1.2 Check the frequency of items purchased by plotting Item frequency graphs

#Check the frequency of items purchased by plotting Item frequency graphs

itemFrequencyPlot(groceries\_1, support = 0.2, main="Item Frequency with 20% support")

Support is given as 0.2 which means the item appears in 20% of the total transactions.

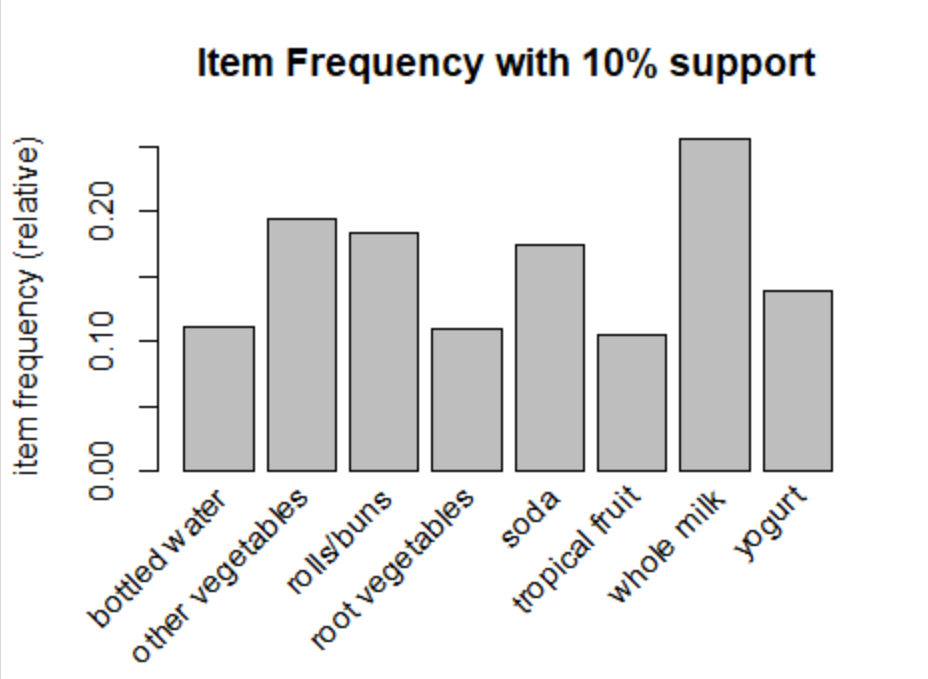


There is only one item, Whole Milk, which occurs for 20% of the total transactions.

Since there was only 1 item, let us relax the support to 0.1 to check for items with 10% in total transactions.

#Relaxing the support to 0.1 - to find out itemset which occurs 10% in transcations

itemFrequencyPlot(groceries\_1, support = 0.1, main="Item Frequency with 10% support")



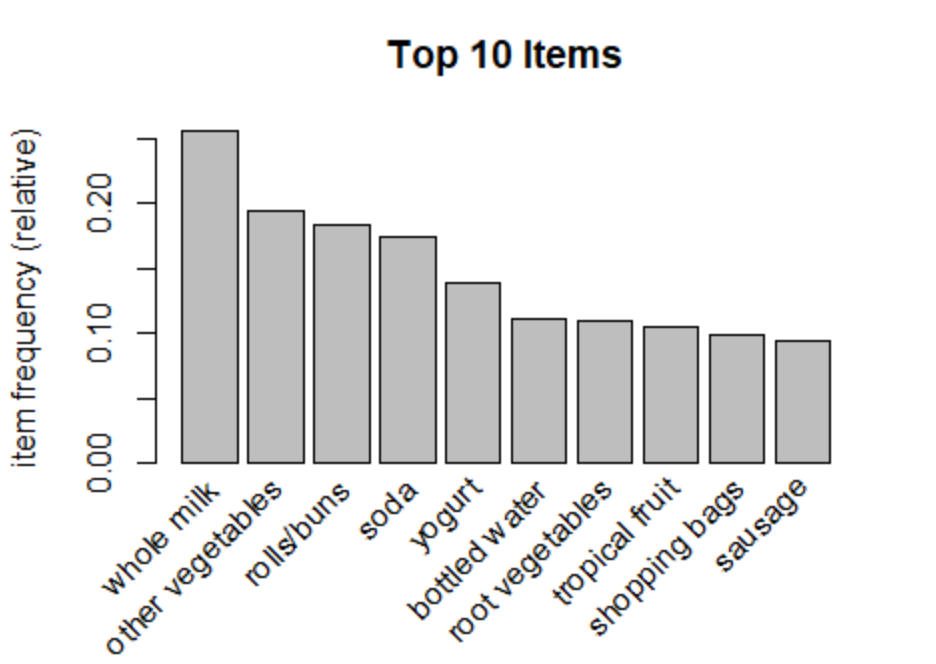
We can observe from the plot, the items occurring for atleast 10% of the total transactions are:

* Bottled Water
* Rolls/Buns
* Root vegetables
* Other vegetables
* Whole Milk
* Yogurt
* Tropical fuit
* Soda

Let us also find out top 10 items by frequencies.

# Top N graphs - arranged by support

itemFrequencyPlot(groceries\_1, topN = 10,main="Top 10 Items")



The above plot demonstrates the top 10 items in the order of their frequency.

We can observe, Shopping bags and Sausage, are the two items which are very close to 10% support.

# Association Rules – Apriori Recommendation Algorithm

## Generating Association Rules

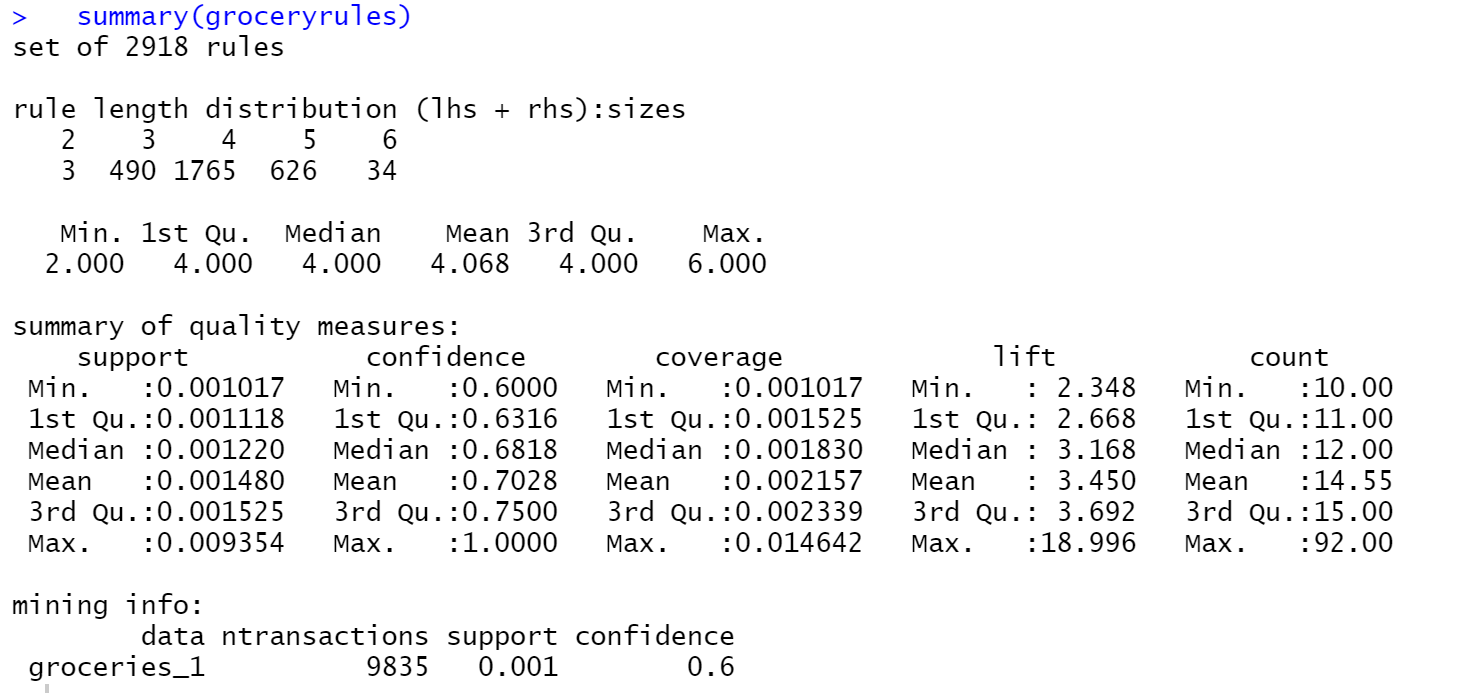
Let us mine for Association Rules by using Apriori Algorithm on the given set of transactions. Here we set the parameters for the rules.

First, we will find all the rules with minimum support as 0.001, that the item set appears at least in 0.1% of the total transactions and a minimum confidence of 0.6, that is the likelihood that the item will be bought along with other itemset is 60% . Only the baskets with two or more items will be considered.

Depending on the purpose of analysis the threshold are set, in the later section there is an example explained in which the rule looks for low confidence value.

#Association Rules

groceryrules <- apriori(groceries\_1, parameter = **list**(support = 0.001, confidence = 0.60, minlen = 2))



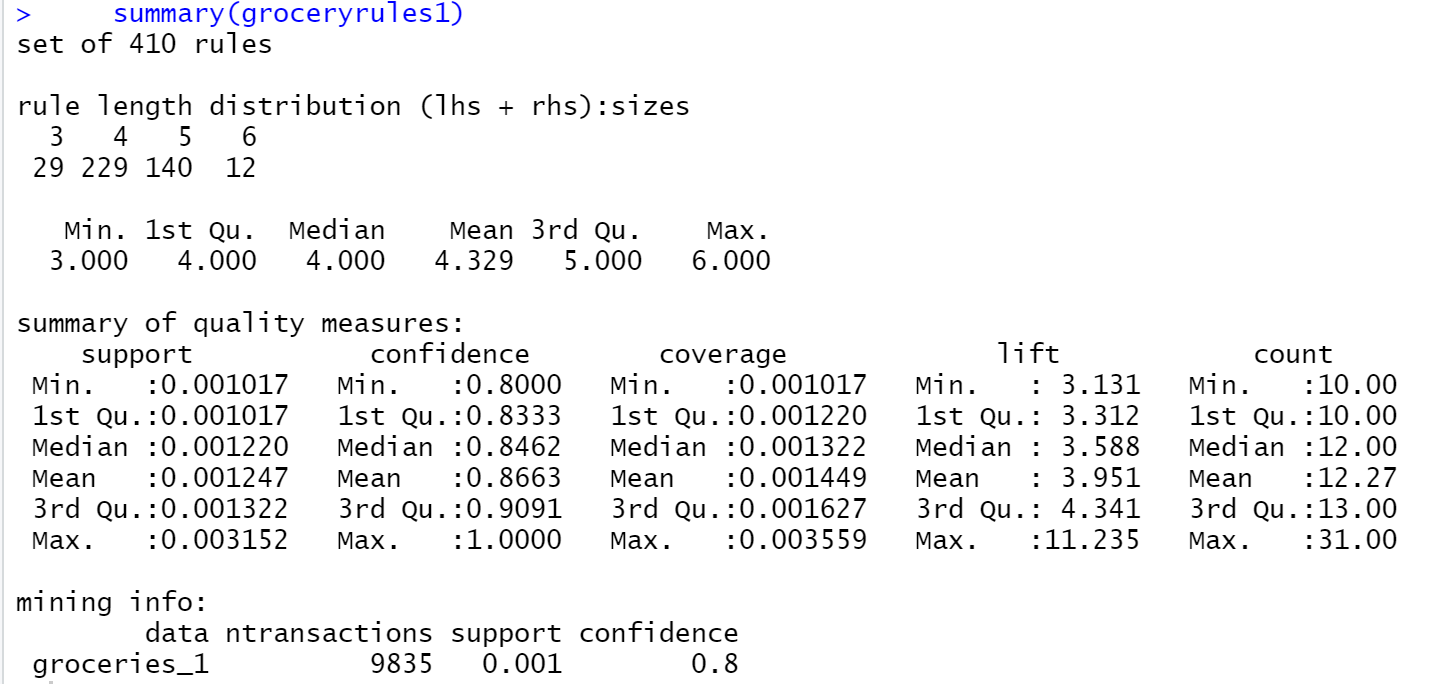
Total 2918 association rules are generated. The basket size of 4 has the highest number of rules, that is 1765.

Since the given parameters resulted in large number of rules, let us attempt to obtain stronger rules by increasing the confidence to 80%, while keeping support and basket length same.

#Association Rules

Groceryrules1 <- apriori(groceries\_1, parameter = **list**(support = 0.001, confidence = 0.80, minlen = 2))

We can check the summary information of the generated association rules:

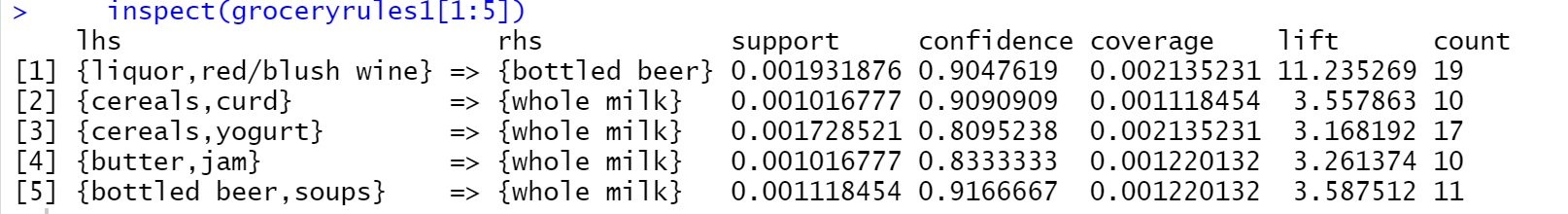


With 80% confidence, the number of association rules has reduced to 410.

It can be observed that the basket size of only 3,4,5,6 has been considered where minimum size was given as 2. Most of the rules consist of baskets containing 4 items (229) and 5 items (140).

The confidence of the generated rules has a median of 84%, whereas the lift ranges from 3 and goes on upto 11, but their median remains 3.5.

Let us display first 5 rules and summaries them:



* 90% of the customers who bought Liquor and Red/Blush Wine also bought Bottled Beer. This rule gives a significant lift of 11.
* Whole Milk seems to be paired up with lot of other itemset.
* If a customer buys Cereals and curd, they are 90% likely to buy Whole Milk.
* 83% of customers who bought Butter and jam also bought Whole Milk.

Please note that the rules have not been sorted yet, this was just to understand the rules generated by the algorithm.

## Pruning the generated Association Rules

The association rules that are generated gives us baskets of size 3 to 6. There can be some rules that are subsets of larger rules. Such redundant rules have to be eliminated from our set so that the most relevant rules can be considered.

Let us prune the association rules by removing the redundant rules.

# 2.3 Obtaining Association rules and pruning them as needed to arrive at a final set of rules

#removing reduandant rules

subset.matrix <- **is**.subset(groceryrules1, groceryrules1)

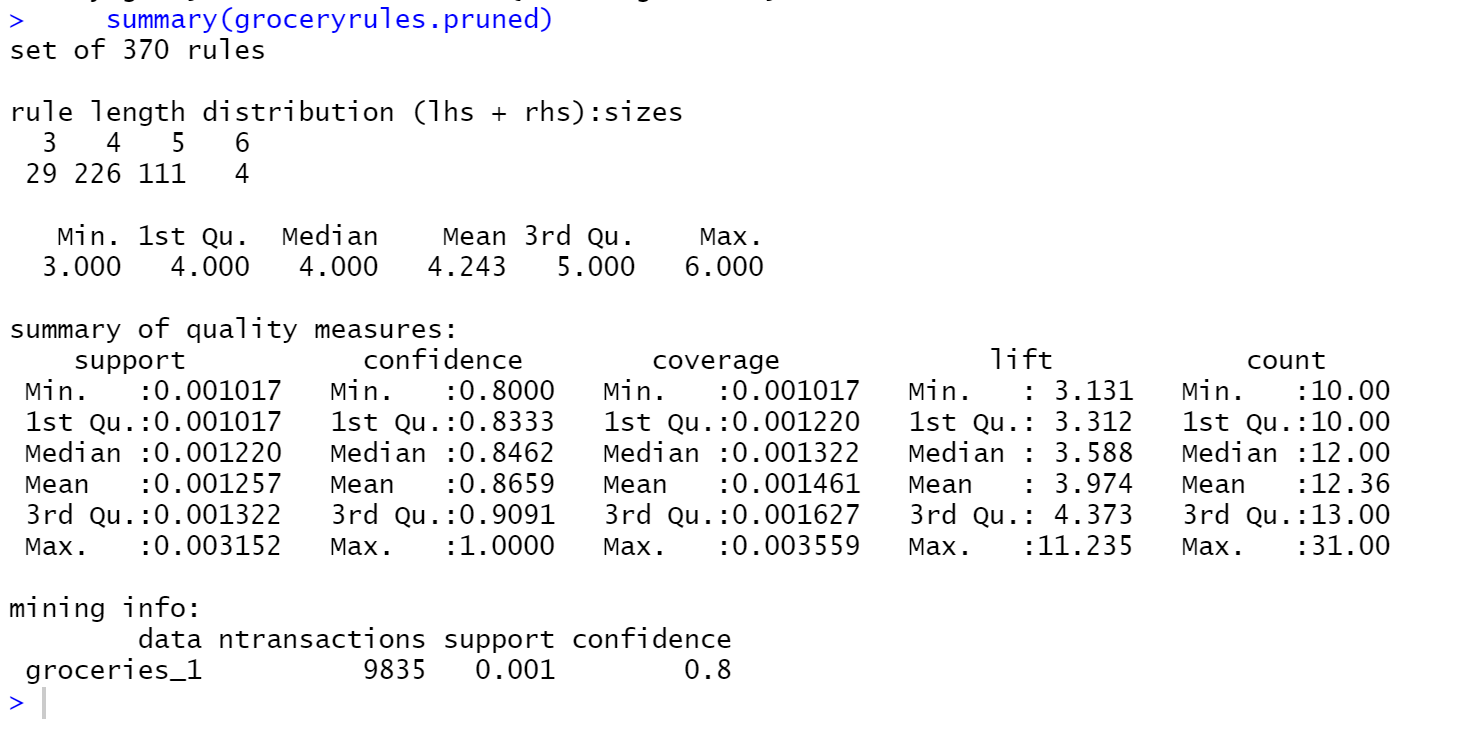
subset.matrix[lower.tri(subset.matrix, diag=T)] <- F

redundant <- colSums(subset.matrix, na.rm=T) >= 1

which(redundant)

groceryrules.pruned <- groceryrules1[!redundant]

The result obtained is as follows:



The number of rules has reduced to 370 after the removal of redundant rules. We can observe the basket length distribution and other ranges of lift, confidence remains the same even after pruning.

## Analysing top rules by lift

The lift value is a measure of importance of a rule. The lift value of an association rule is the ratio of the confidence of the rule and the expected confidence of the rule. The Lift tells us how much better a rule is at predicting the result than just assuming the result in the first place. Greater lift values indicate stronger associations.

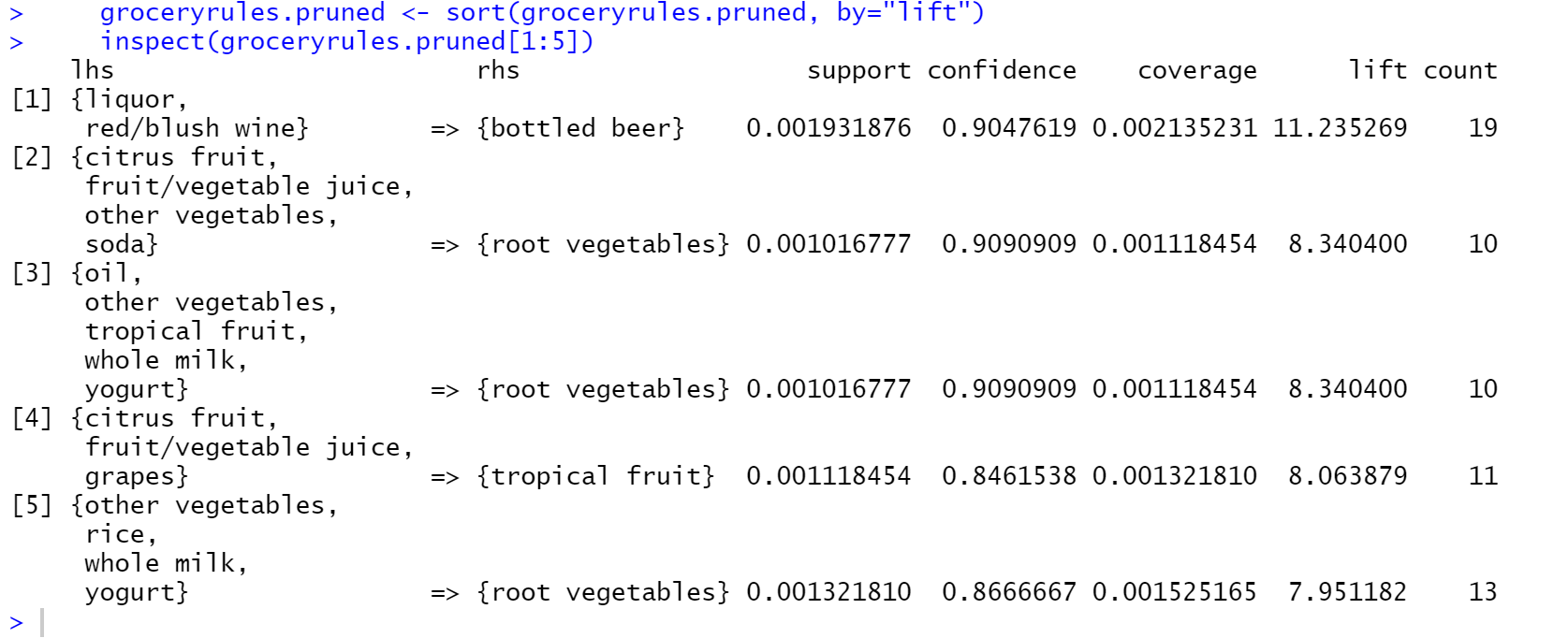
The association rules have been sorted based on their lift, to check for the rules that give maximum lift.

#2.4 Analysing the top rules based on lift value

#3.1 Interpret the top 5 rules

groceryrules.pruned <- sort(groceryrules.pruned, by="lift")

Displaying the top 5 rules with maximum lift, as higher the lift, higher is the association between the item sets.



* There is high association of Bottled Beer with the itemset, Liquor and Red/Blush Wine. The lift is as high as 11, there is 90% probability that the customer will also buy Bottled Beer if he has already bought Liquor and Red wine.
* The second rule demonstrates fruits, vegetables, soda are likely to be bought together. If a customer already has Citrus fruits, fruit/vegetable juice, soda, other vegetables in the basket, they will most likely also buy Root vegetables.
* Root vegetables has another itemset of products, that is oil, other vegetables, tropical fruits, whole milk, yogurt, which can be bought also with 90% of confidence.
* The fourth rule is a mix of fruits and juices, where tropical fruit has 84% probability of being added to the basket of the customer has already picked up Citrus fruits, fruit/vegetable juice and grapes.
* 86% of the customers that buy Other vegetables, rice, whole milk, yogurt are also likely to buy Root vegetables, with lift of close to 8.

The high value of lift signifies how impactful the association rule is, if the product placement is done accordingly, thereby increasing the probability of sales.

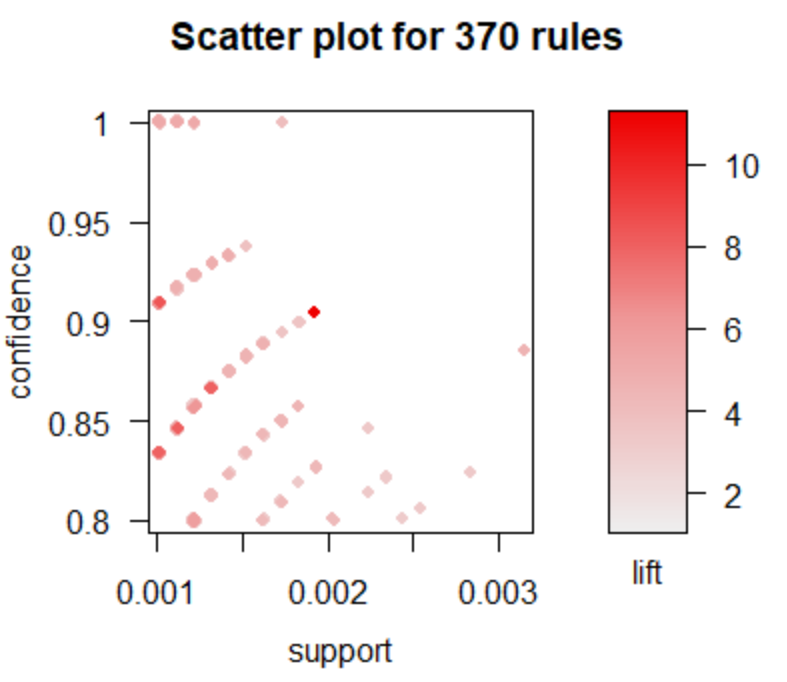
# Visualizing Association Rules

## Scatter Plot of pruned association rules

Let us create a visualization of pruned association rules using a scatter plot:

# 3.2 Plot the rules in a graph/scatter plot in R

plot(groceryrules.pruned,method="scatterplot")



The above scatterplot demonstrates support on X-axis, confidence on Y-axis, and the lift is indicated by density of the color.

The support is above 0.1% and confidence is above 80% as per the threshold specified while generating the association rules.

It can be observed that most of the rules have a support of 0.1% to 0.2%.

There are certain rules with a confidence of 100%, but the lift is not that high.

A lot of association rules are generated. In order to be able to make recommendations for product placement, the rules can be studied based on highest lift, confidence and support.

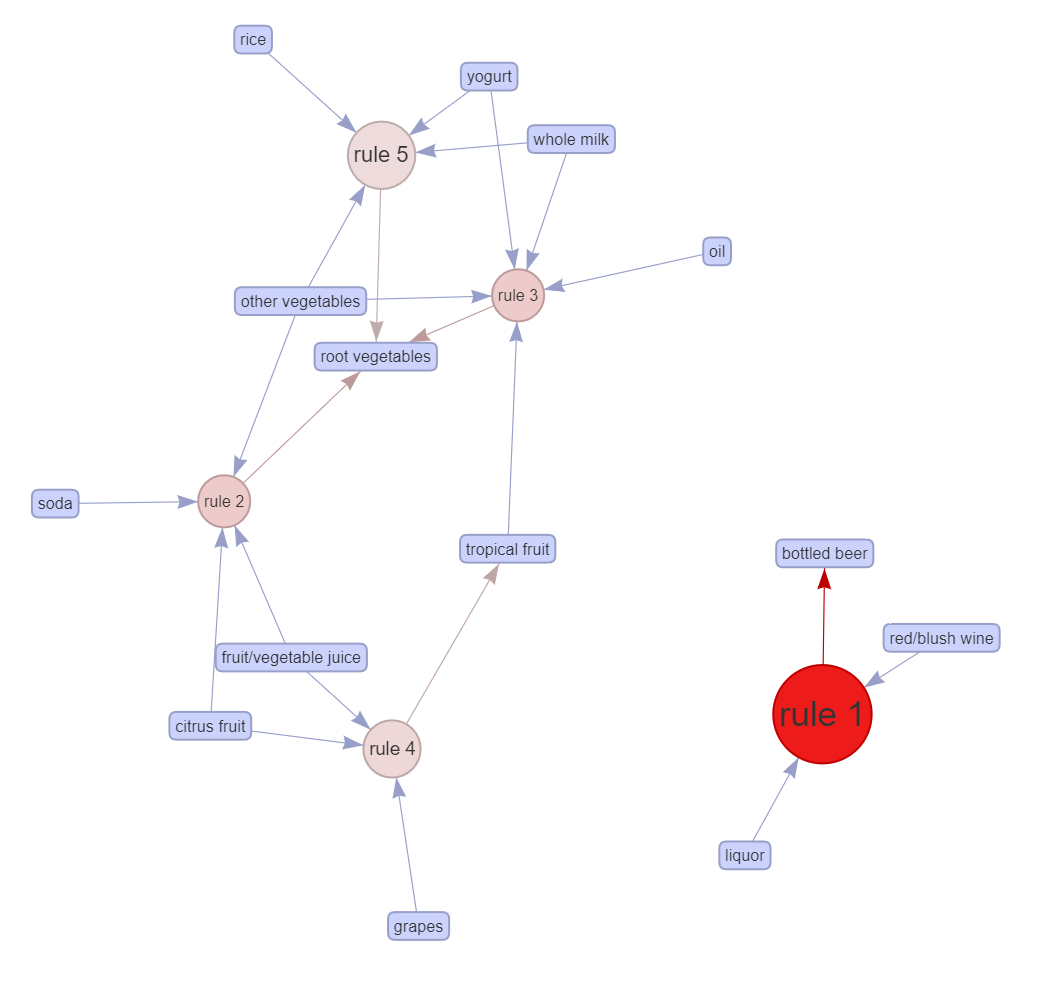
## Top rules based on Lift

Let us visualize the top 5 rules that have highest lift:

#top 5 rules by lift

top5subRules\_lift <- head(groceryrules.pruned, n = 5, by = "lift")

plot(top5subRules\_lift, method = "graph", engine = "htmlwidget")



The above graph helps us to visualize the products that can be grouped together.

* Rule 1 that consists of Bottled Beer, Liquor and Red/blush wine stands out significantly, giving the highest lift of 11.
* The other 4 rules are not so exclusive like Rule 1, they are somewhat related to each other. Fruits and vegetables form a subset, while also merging with juices and dairy products.
* Another significant observation from the graphs is that the product Root Vegetables is being derived from Rule 2, Rules 3 as well as Rule 5.

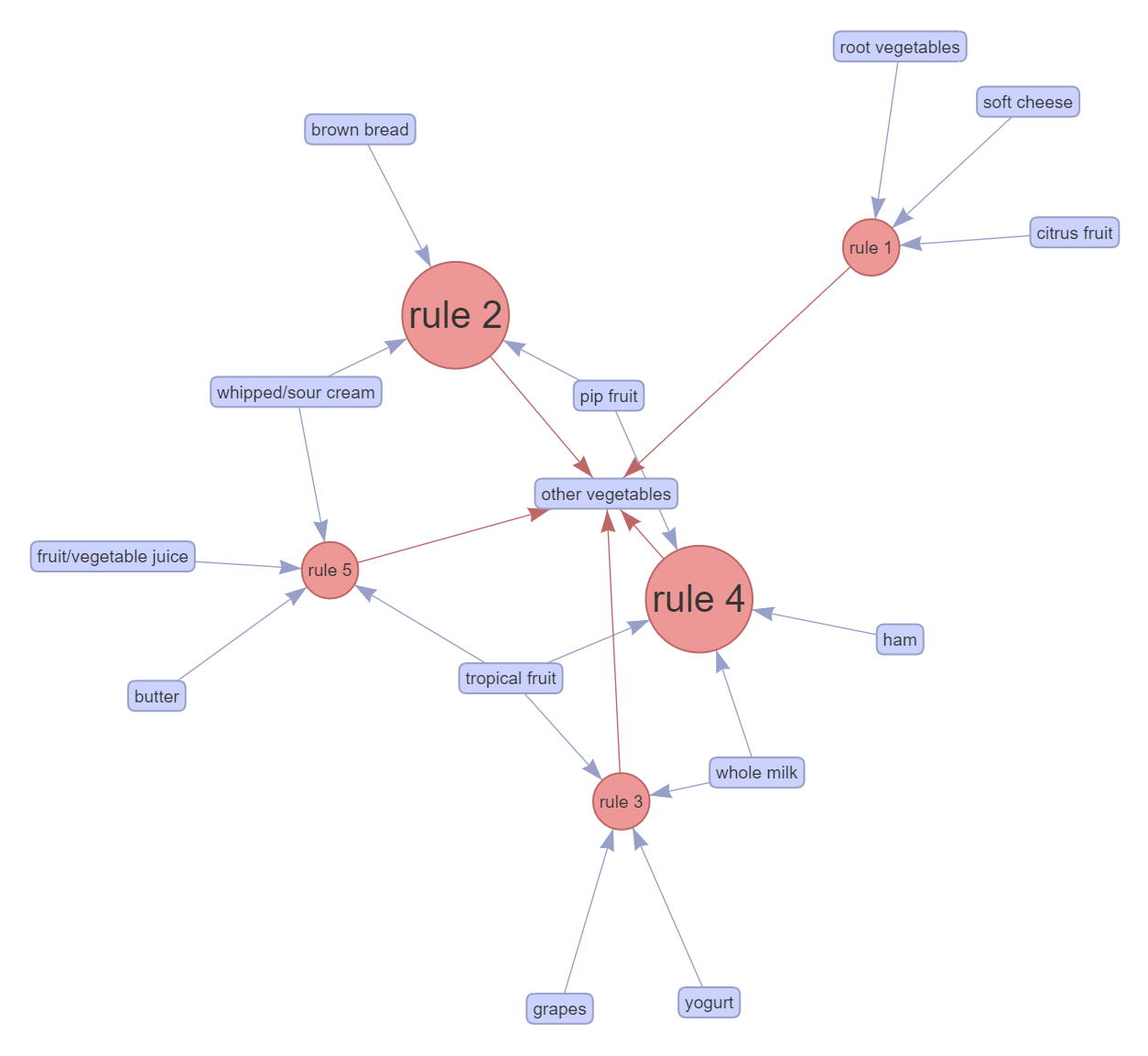
## Top rules based on Confidence

Let us visualize the top 5 rules that have highest confidence:

#top 5 rules by confidence

top5subRules\_confidence <- head(groceryrules.pruned, n = 5, by = "confidence")

plot(top5subRules\_confidence, method = "graph", engine = "htmlwidget")



When the rules that give highest confidence are visualized, it is observed that all the top 5 rules derive Other Vegetables as the product that will be purchased.

Even amongst those, this product is more likely to be bought along with 2 sets having more support than the rest, that is,

* + - 1. Brown Bread, whipped/sour cream, pip fruit
      2. Pip fruit, ham, whole milk, tropical fruit

All the rules demonstrated here have confidence of 100%.

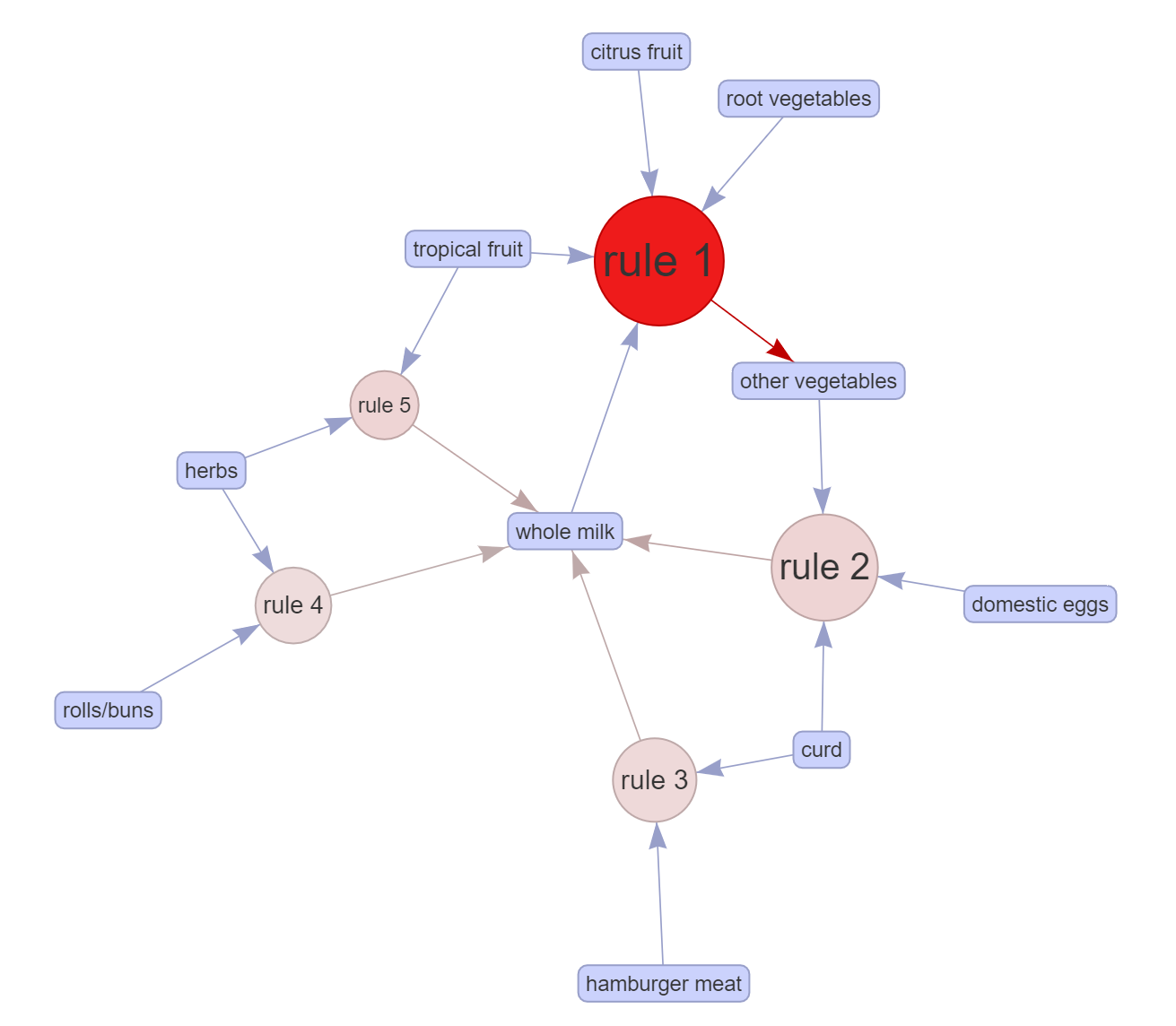
## Top rules based on Support

Let us visualize the association rules that contain most frequent itemsets of products:

#top 5 rules by support

top5subRules\_support <- head(groceryrules.pruned, n = 5, by = "support")

plot(top5subRules\_support, method = "graph", engine = "htmlwidget")



As observed from the graph,

* Most frequently bought itemset consists of Citrus fruit, Root vegetables, Tropical fruit and whole milk, which lead to the purchase of Other vegetables.
* This combination has 88% probability, giving a lift of 4.5
* The next itemset is Domestic Eggs, Curd, Other vegetables, which lead to the purchase of Whole milk.

# Observation based on Market Basket Analysis

On further analysing the geenrated association rules,

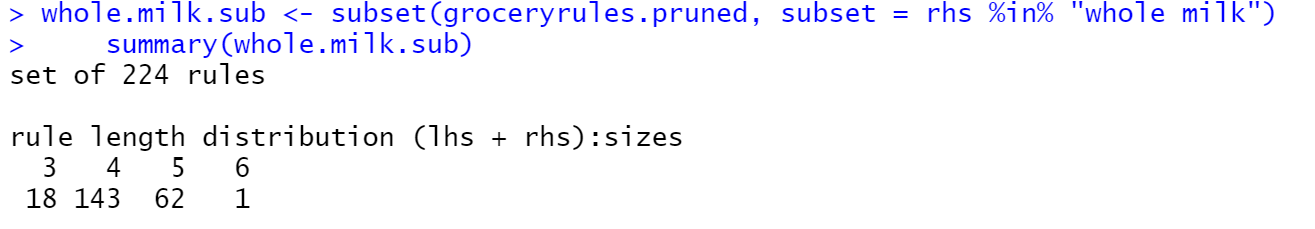
# Observations based on Market Basket analysis

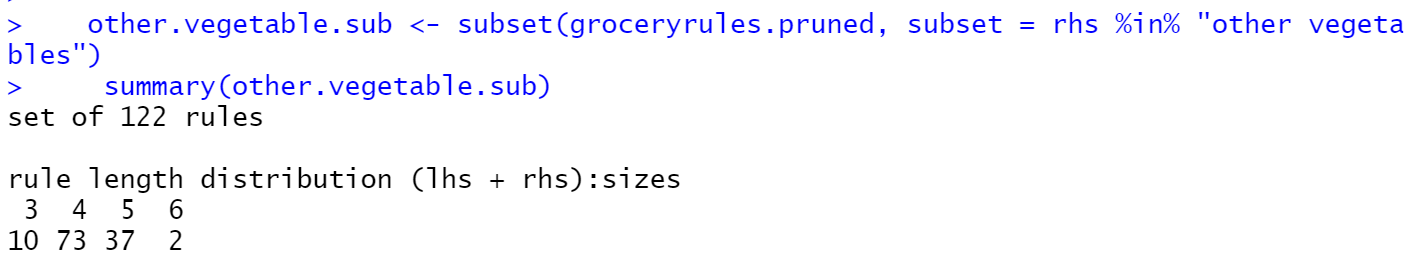
whole.milk.sub <- subset(groceryrules.pruned, subset = rhs %**in**% "whole milk")

summary(whole.milk.sub)

other.vegetable.sub <- subset(groceryrules.pruned, subset = rhs %**in**% "other vegetables")

summary(other.vegetable.sub)



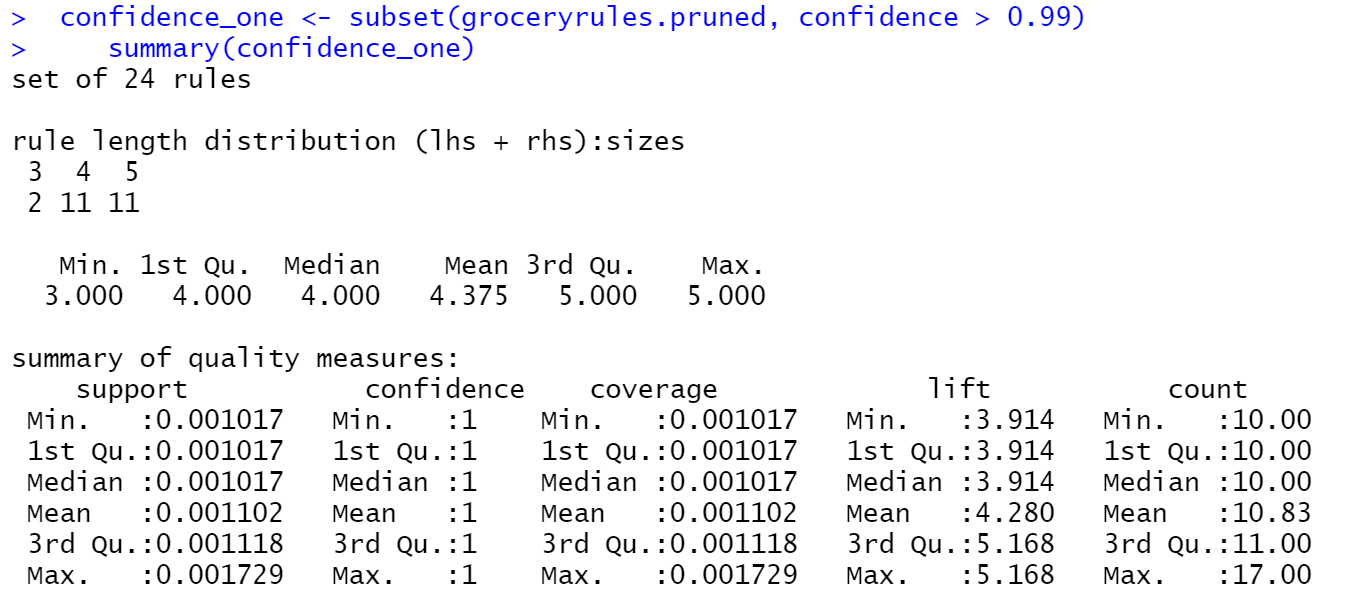


It can be observed that there are 224 out of 370 rules in which the recommendation is ‘whole milk’.

Whereas, 122 rules in which the recommendation is ‘Other vegetables’.

Thus, more than 90% of the rules recommend Whole Milk and Other Vegetables.

There are few association rules having confidence greater than 99%, which can certainly be implemented. Their lift ranges from 3.9 to 5.



The significant observations can be noted as follows:

* Whole Milk and Other Vegetables have high association with different combinations of products available in the store. Hence, most of the rules pertain to Whole Milk and Other Vegetables.
* 24 rules are obtained whose probability of purchase given the basket is above 99%. Such rules have lift value of around 4 to 5.
* The most important rule is the one which recommends Bottled Beer when Liquor and Red/Blush wine is added to the cart. It has 90% confidence and exceptionally high lift value of 11.
* All the rules with > 99% confidence either recommends Whole Milk or Other Vegetables.
* The product Root Vegetables also gives a very high lift around 8, with confidence of 90%.

# Actionable Insights and Recommendations

* The association rules can be used to plan the shelf place at a retail store:
  + Bottled Beer, Liquor and Red/blush wine can definitely be bundled together. This can be the Liquor Section.
  + Whole Milk and Other Vegetables can be kept at a prominent place, as they are the most frequently bought items.
  + Root Vegetables can be placed with Citrus fruits, Fruit/vegetable juice, soda, and other vegetables.
* Also, the rules can be used to build recommendation system for the online store, where the rules of higher lift can be considered.
* The association rules can be used to design promotional campaign, where we can identify the products with lower confidence value as they are not always picked up but with higher lift value. This will help to boost the sales.
  + One association rule specifies if one purchases Baking powder and Sugar, they are bound to purchase “flour”. The confidence is only 30% however the lift is 18 which seems good. Now these three products can be bundled and a promotional offer can be designed, from a common sense perspective also this makes sense as these ingredients are used for baking.
  + There is one more such rule which indicates the if one buys white bread and ham, they are bound to buy processed cheese, which can appeal to common sense. These rules having a low confidence may be ignore in the first place, but can be also be analysed.
* Association rules can be effective in targeting specific products, the rules can be filtered based on specific target product. For example, there are 25 rules which has Citrus Fruit as antecedent items which can be used for applying the above strategies such as product bundling or other discounts/offers.