STA 380 Exercise 8

Ryota Y., Janie C., Neha B., Gayathree G.

2024-08-13

```
library(arules)
```

```
## Loading required package: Matrix

##
## Attaching package: 'arules'

## The following objects are masked from 'package:base':
##
##     abbreviate, write
```

```
library(arulesViz)
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x tidyr::expand() masks Matrix::expand()
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x tidyr::pack()   masks Matrix::pack()
## x dplyr::recode() masks arules::recode()
## x tidyr::unpack() masks Matrix::unpack()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(igraph)
```

```
##
## Attaching package: 'igraph'
##
## The following objects are masked from 'package:lubridate':
##
##     %--%, union
##
```

```
## The following objects are masked from ’package:dplyr’:
##
##     as_data_frame, groups, union
##
## The following objects are masked from ’package:purrr’:
##
##     compose, simplify
##
## The following object is masked from ’package:tidyr’:
##
##     crossing
##
## The following object is masked from ’package:tibble’:
##
##     as_data_frame
##
## The following object is masked from ’package:arules’:
##
##     union
##
## The following objects are masked from ’package:stats’:
##
##     decompose, spectrum
##
## The following object is masked from ’package:base’:
##
##     union
```

8. Association rule mining

Revisit the notes on association rule mining and the R example on music playlists: playlists.R and playlists.csv. Then use the data on grocery purchases in groceries.txt and find some interesting association rules for these shopping baskets. The data file is a list of shopping baskets: one person's basket for each row, with multiple items per row separated by commas. Pick your own thresholds for lift and confidence; just be clear what these thresholds are and say why you picked them. Do your discovered item sets make sense? Present your discoveries in an interesting and visually appealing way.

Notes:

- This is an exercise in visual and numerical story-telling. Do be clear in your description of what you've done, but keep the focus on the data, the figures, and the insights your analysis has drawn from the data, rather than technical details.
- The data file is a list of baskets: one row per basket, with multiple items per row separated by commas. You'll have to cobble together your own code for processing this into the format expected by the "arules" package. This is not intrinsically all that hard, but it is the kind of data-wrangling wrinkle you'll encounter frequently on real problems, where your software package expects data in one format and the data comes in a different format. Figuring out how to bridge that gap is part of the assignment, and so we won't be giving tips on this front.

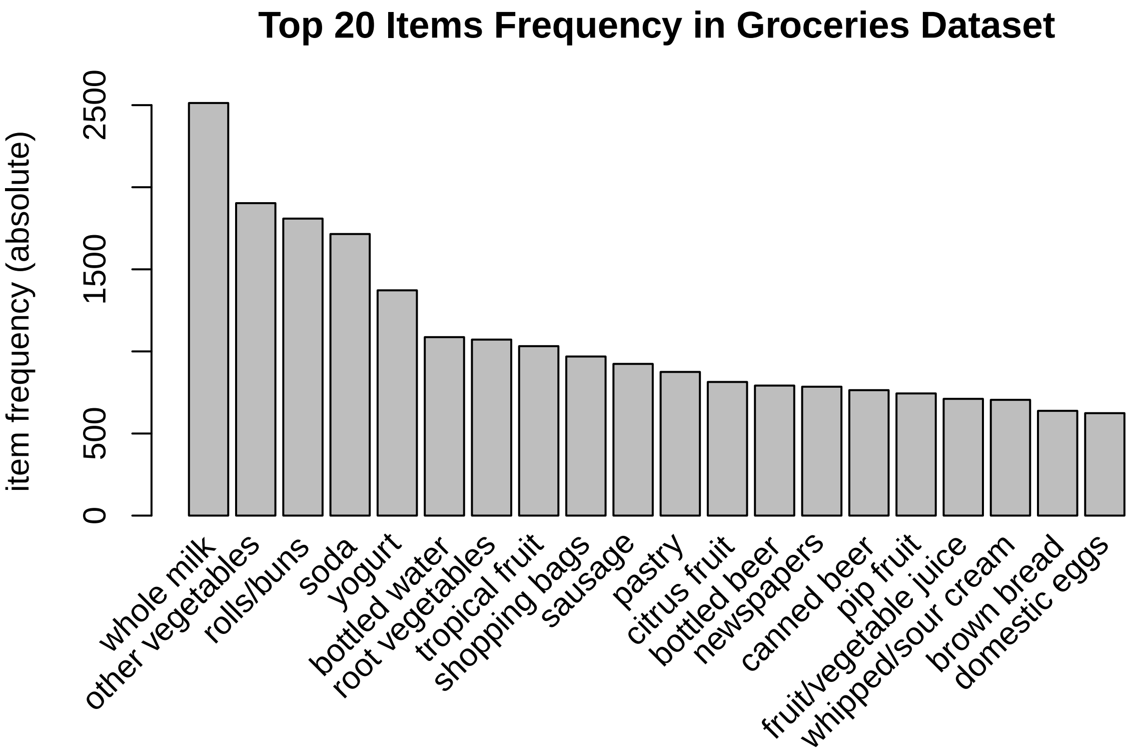
```r
groceries <- read.transactions(file="groceries.txt",
                               sep = ',',format="basket",rm.duplicates=TRUE)
head(groceries)
```

```
## transactions in sparse format with
##  6 transactions (rows) and
##  169 items (columns)
```

```r
summary(groceries)
```

```
## transactions as itemMatrix in sparse format with
##  9835 rows (elements/itemsets/transactions) and
##  169 columns (items) and a density of 0.02609146
##
## most frequent items:
##       whole milk other vegetables       rolls/buns             soda
##             2513             1903             1809             1715
##           yogurt          (Other)
##             1372            34055
##
## element (itemset/transaction) length distribution:
## sizes
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16
## 2159 1643 1299 1005  855  645  545  438  350  246  182  117   78   77   55   46
##   17   18   19   20   21   22   23   24   26   27   28   29   32
##   29   14   14    9   11    4    6    1    1    1    1    3    1
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   4.409   6.000  32.000
##
## includes extended item information - examples:
##             labels
## 1 abrasive cleaner
## 2 artif. sweetener
## 3   baby cosmetics
```

```r
# Plot the item frequency
itemFrequencyPlot(groceries, topN = 20, type = "absolute",
                  main = "Top 20 Items Frequency in Groceries Dataset")
```



We can expect rules with the above items to have higher supports. Confidence and lift will need to be investigated.

Support

Definition: Support is the proportion of transactions in the dataset that contain a particular itemset (both the antecedent and consequent of a rule).

Mathematical Expression: For a rule $A \to B$, the support is calculated as:

$$\text{Support}(A \to B) = \frac{\text{Number of transactions containing both } A \text{ and } B}{\text{Total number of transactions}}$$

Interpretation: Support measures how frequently the itemset (both items in the rule) occurs in the dataset. A higher support indicates that the itemset is common in the transactions, while a lower support indicates that the itemset is rare.

Confidence

Definition: Confidence is the proportion of transactions that contain the antecedent (e.g., itemset A) that also contain the consequent (e.g., itemset B).

Mathematical Expression: For a rule $A \to B$, confidence is calculated as:

$$\text{Confidence}(A \to B) = \frac{\text{Number of transactions containing both } A \text{ and } B}{\text{Number of transactions containing } A}$$

Interpretation: Confidence measures the likelihood that the consequent of the rule (itemset B) is also purchased when the antecedent (itemset A) is purchased. Higher confidence indicates that the rule $A \rightarrow B$ is more reliable, meaning that whenever A is bought, B is also likely to be bought.

Example in the Grocery Dataset

Suppose you have a rule Milk → Bread.

- **Support:** If the support for this rule is 0.2, it means that 20% of all transactions in the dataset contain both milk and bread.
- **Confidence:** If the confidence for this rule is 0.8, it means that 80% of the transactions that include milk also include bread.

```r
print(paste('Size of the dataset: ', dim(groceries)[1]))
```

```
## [1] "Size of the dataset:  9835"
```

Since the data set has around 10,000 rows, the support should be around 0.1% (containing at least 10 instances) to be worth investigation. The following thresholds were chosen for the interpretability of a graph later in this file.

```r
# Create association rules
support_threshold <- .001 #item set appears in at least 10 transactions
confidence_threshold <- 0.7 #more reliable rules
maxlength <- 5 #to filter for interpretable, actionable rules

groceries_rules = apriori(groceries,
                          parameter=list(support=support_threshold,
                                         confidence=confidence_threshold,
                                         maxlen=maxlength))
```

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##         0.7    0.1    1 none FALSE            TRUE       5   0.001      1
##  maxlen target  ext
##       5  rules TRUE
##
## 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 [0.00s].
## 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
## Warning in apriori(groceries, parameter = list(support = support_threshold, :
## Mining stopped (maxlen reached). Only patterns up to a length of 5 returned!
```

```
##  done [0.01s].
## writing ... [1255 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

```r
# inspect(groceries_rules)
length(groceries_rules)
```

```
## [1] 1255
```

Definition of Lift

Lift is the ratio of the observed support for the rule to the expected support if the items were independent. It tells you how many times more likely the consequent is to be found in transactions that contain the antecedent compared to transactions that do not contain the antecedent.

Mathematical Expression

For a rule $A \to B$, lift is calculated as:

$$\text{Lift}(A \to B) = \frac{\text{Support}(A \to B)}{\text{Support}(A) \times \text{Support}(B)}$$

Where: - **Support(A $\to$ B):** The proportion of transactions containing both A and B. - **Support(A):** The proportion of transactions containing A. - **Support(B):** The proportion of transactions containing B.

Interpretation

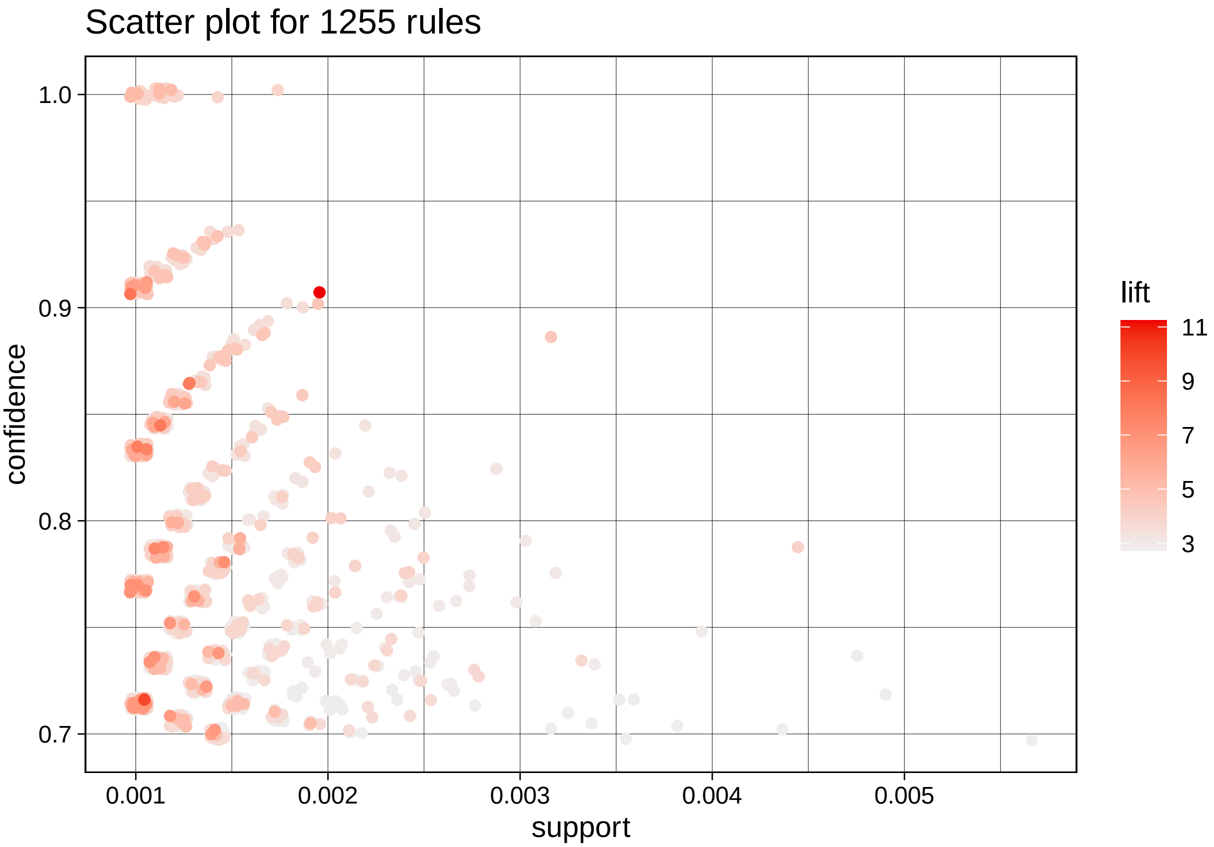
- **Lift $=$ 1:** The antecedent and consequent are independent of each other, meaning the presence of A does not influence the presence of B.
- **Lift $>$ 1:** There is a positive association between the antecedent and consequent, meaning that the presence of A increases the likelihood of B occurring. The higher the lift, the stronger the association.
- **Lift $<$ 1:** There is a negative association between the antecedent and consequent, meaning that the presence of A actually decreases the likelihood of B occurring.

Example in the Grocery Dataset

Suppose you have a rule Milk → Bread with a lift of 3.
This means that customers who buy milk are three times more likely to also buy bread compared to a random customer buying bread.
A lift greater than 1 indicates that the rule is meaningful and the items in the rule are associated in a way that is more than just by chance.

```r
plot(groceries_rules)
```

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



Takeaways: The above graph shows the relationship between confidence and support as colored by lift. As you can see from the graph, support and confidence are negatively correlated. This is because as the support of a rule increases, the rule becomes less specific, reducing its confidence. Higher confidence often occurs with rules that have lower support, as they apply to fewer, more targeted transactions where the antecedent strongly predicts the consequent. In this set, this is also such case with lift and support. The more that an item set appears, the more expected that item set is to appear randomly, and there is not much additional predictive power.

Comparing Different Rule Subsets

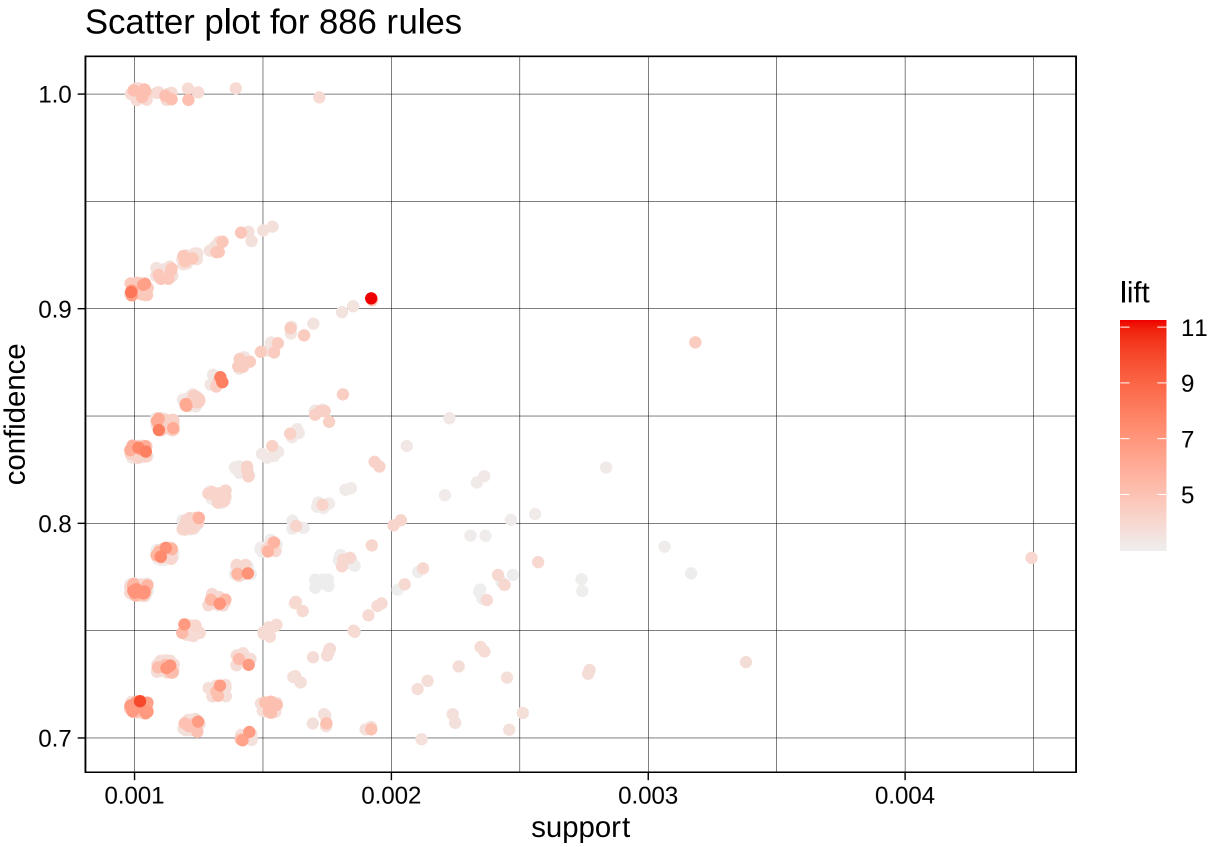
We create different groups of association rules by adjusting the thresholds.

1: Test a the Same Support Threshold and Higher Lift Threshold:

```
support_threshold1 <- 0.001 #item set appears in at least 10 people
lift_threshold1 <- 3 #antecedent leads to 2 times likelihood of consequent
inspect(subset(groceries_rules, lift > lift_threshold1 & support > support_threshold1))
length(subset(groceries_rules, lift > lift_threshold1 & support > support_threshold1))
```

```
plot(subset(groceries_rules, lift > lift_threshold1 & support > support_threshold1))
```

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

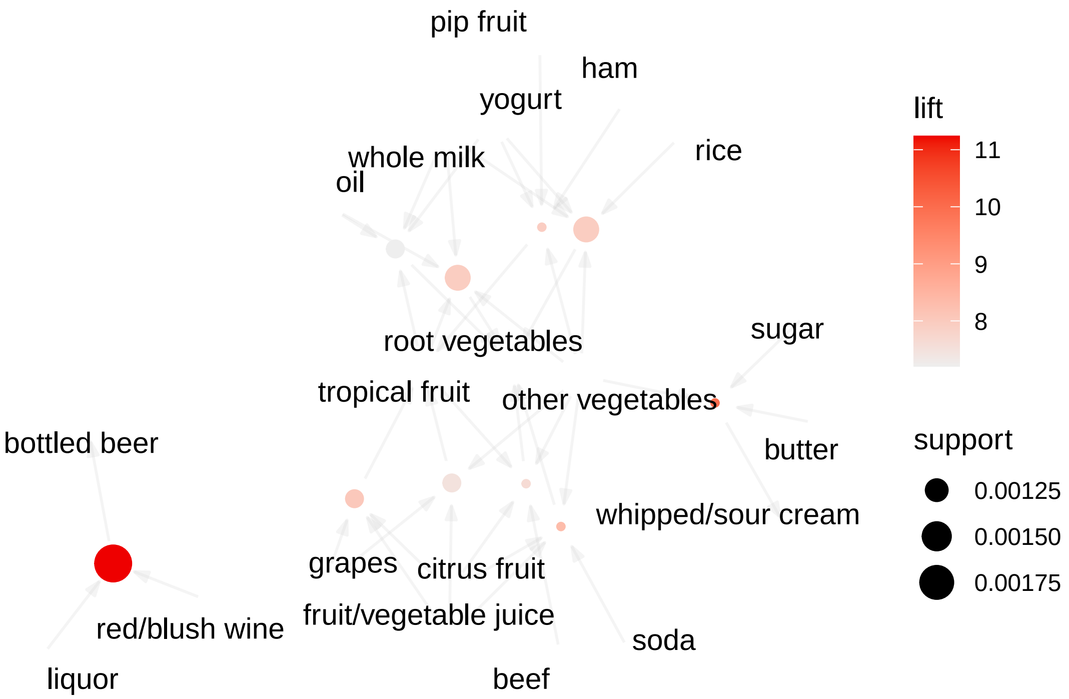


Insights: This is similar to the previous plot but with much less noise (only lift > 3). These are all rules which are powerful enough to be worth investment; however, many rules only appear in 10 to 20 people.

```
groceries_rules2 = subset(groceries_rules, lift > support_threshold1 & support > support_threshold1)
plot(head(sort(groceries_rules2, by="lift"), 10),
  method="graph", control=list(cex=.9), main = "Top 10 Groceries Rules (High Support, Low Lift)")
```

```
## Warning: Unknown control parameters: cex, main

## Available control parameters (with default values):
## layout    =  stress
## circular  =  FALSE
## ggraphdots    =  NULL
## edges     =  <environment>
## nodes     =  <environment>
## nodetext  =  <environment>
## colors    =  c("#EE0000FF", "#EEEEEEFF")
## engine    =  ggplot2
## max   =  100
## verbose   =  FALSE
```



Takeaways: The above graph represents with higher lifts. There seems to be two core clusters, with one as liquor and another as produce and dairy. This is a great graph showing the grocery items that are important as well as their corresponding rules, but let's see if we can extract rules that apply to more people.

2: Test a Higher Support Threshold and Higher Lift Threshold:

```
support_threshold2 <- 0.003 #item set appears in at least 30 people
lift_threshold2 <- 3 #antecedent leads to 3 times likelihood of consequent
inspect(subset(groceries_rules, lift > lift_threshold2 & support > support_threshold2))
```

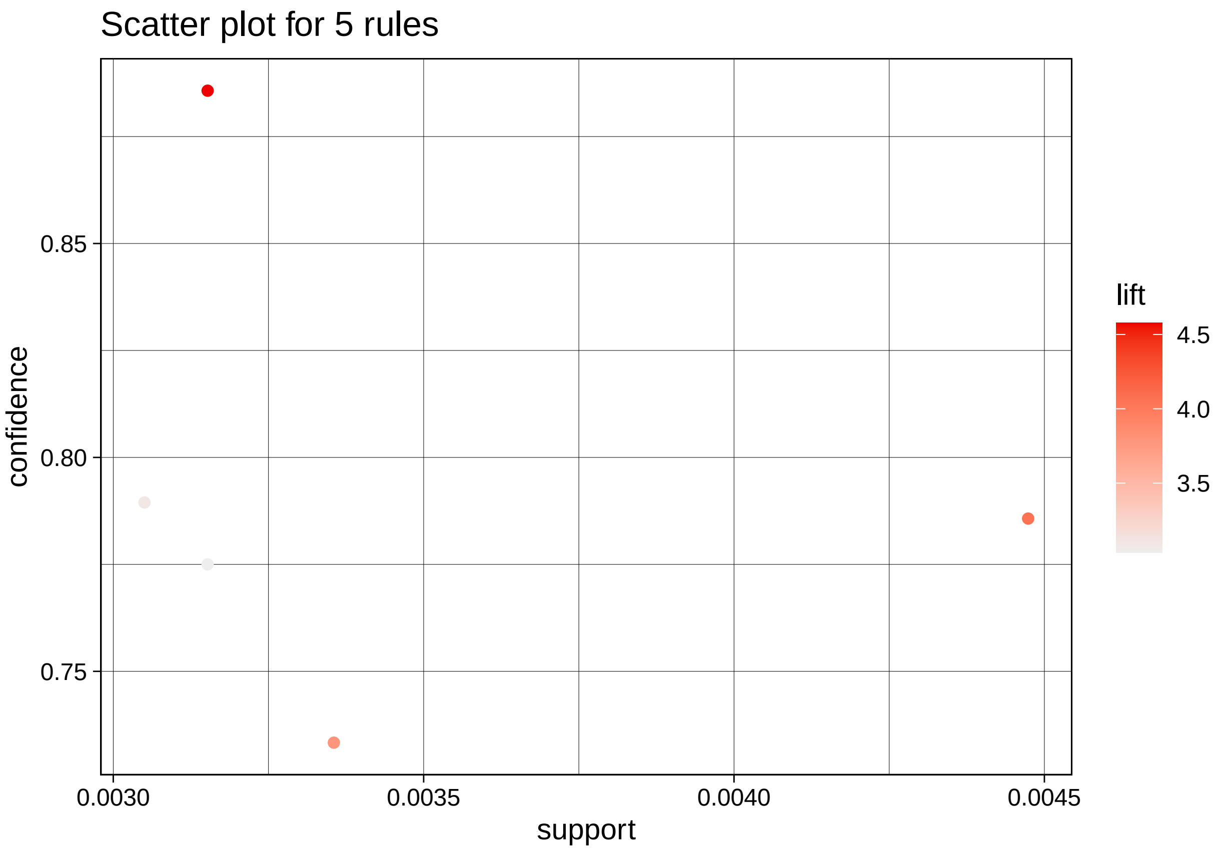
```
##     lhs                        rhs                    support confidence    coverage     lift count
## [1] {brown bread,
##      other vegetables,
##      root vegetables}    => {whole milk}       0.003152008  0.7750000 0.004067107 3.033078    31
## [2] {butter,
##      root vegetables,
##      yogurt}             => {whole milk}       0.003050330  0.7894737 0.003863752 3.089723    30
## [3] {root vegetables,
##      tropical fruit,
##      whipped/sour cream} => {other vegetables} 0.003355363  0.7333333 0.004575496 3.789981    33
## [4] {citrus fruit,
##      root vegetables,
##      tropical fruit}     => {other vegetables} 0.004473818  0.7857143 0.005693950 4.060694    44
## [5] {citrus fruit,
##      root vegetables,
##      tropical fruit,
##      whole milk}         => {other vegetables} 0.003152008  0.8857143 0.003558719 4.577509    31
```

```
length(subset(groceries_rules, lift > lift_threshold2 & support > support_threshold2))
```

```
## [1] 5
```

The two consequents are whole milk and other vegetables. This is not surprising since these are core to the american diet. The antecedents, however, are rather interesting, with the inclusion of whipped/sour cream together with vegetables. Yogurt with butter and root vegetables leading to whole milk is also rather interesting in this specific combination. These rules reveal great opportunities for cross-merchandising.

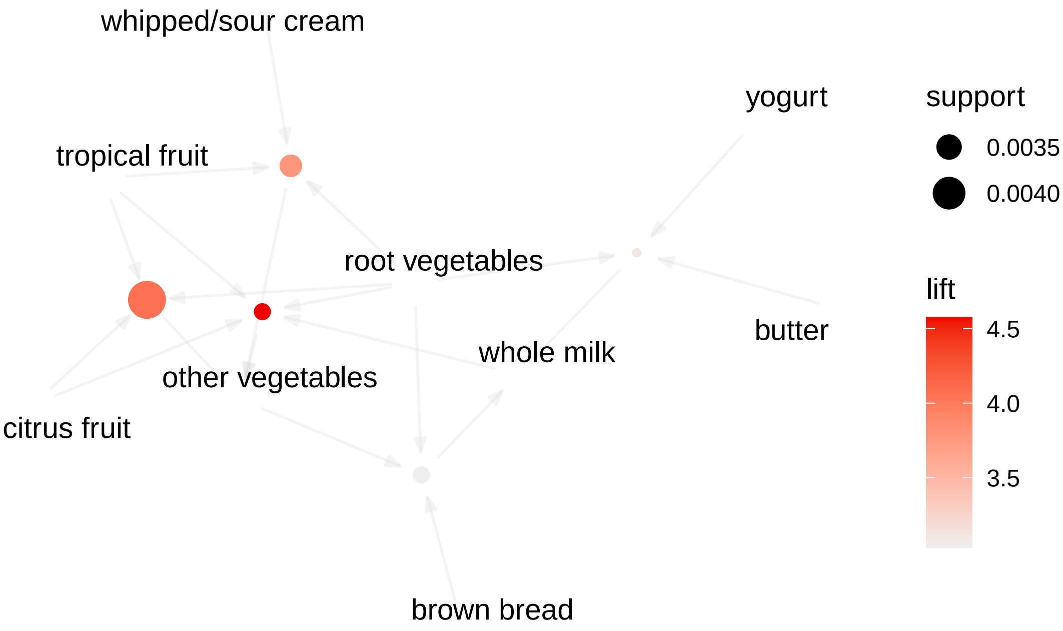
```
plot(subset(groceries_rules, lift > lift_threshold2 & support > support_threshold2))
```



Insights: We can see that increasing the support threshold has removed many rules with low lifts, but also removed the rules with higher lifts than 4.5. This means that although these removed rules are very strong (lift > 4.5), these rules only appear in less than 30 transactions (support < 0.003). From a business perspective, these are especially relevant to target when breaking into niche markets. However, for the sake of this analysis, we will focus on higher support rules to understand general trends that are also strong.

```
groceries_rules2 = subset(groceries_rules, lift > lift_threshold2 & support > support_threshold2)
plot(groceries_rules2,
  method="graph", control=list(cex=.9), main = "Top 5 Groceries Rules (High Support, High Lift)")
```

```
## Warning: Unknown control parameters: cex, main
## Available control parameters (with default values):
## layout    =  stress
## circular  =  FALSE
## ggraphdots    =  NULL
## edges     =  <environment>
## nodes     =  <environment>
## nodetext  =  <environment>
## colors    =  c("#EE0000FF", "#EEEEEEFF")
## engine    =  ggplot2
## max   =  100
## verbose   =  FALSE
```



Description: The above graph represents the top 5 rules when filtered for a support of 0.3% (30 instances) and a lift of 3 (rule is 3 times more likely than a random buyer). As expected, the grocery items included are core to the american diet (bread, fruit, milk, etc.), although yogurt is an interesting inclusion.

Insights: As compared to the previous graph with a lower support threshold, the liquor relationship has disappeared. This rule is important for wine and liquor companies; however, the grocery stores themselves can focus on the other rules that have survived these higher thresholds. Ideally, after considering profit margins, these will be the rules that lead to the highest ROI. Of course, these patterns will change seasonally and geographically, as we can expect that strong rules with turkey will emerge around November each year in the United States.

```
groceries_graph = associations2igraph(groceries_rules2, associationsAsNodes = FALSE)
igraph::write_graph(groceries_graph, file='groceriesrules.graphml', format = "graphml")
```

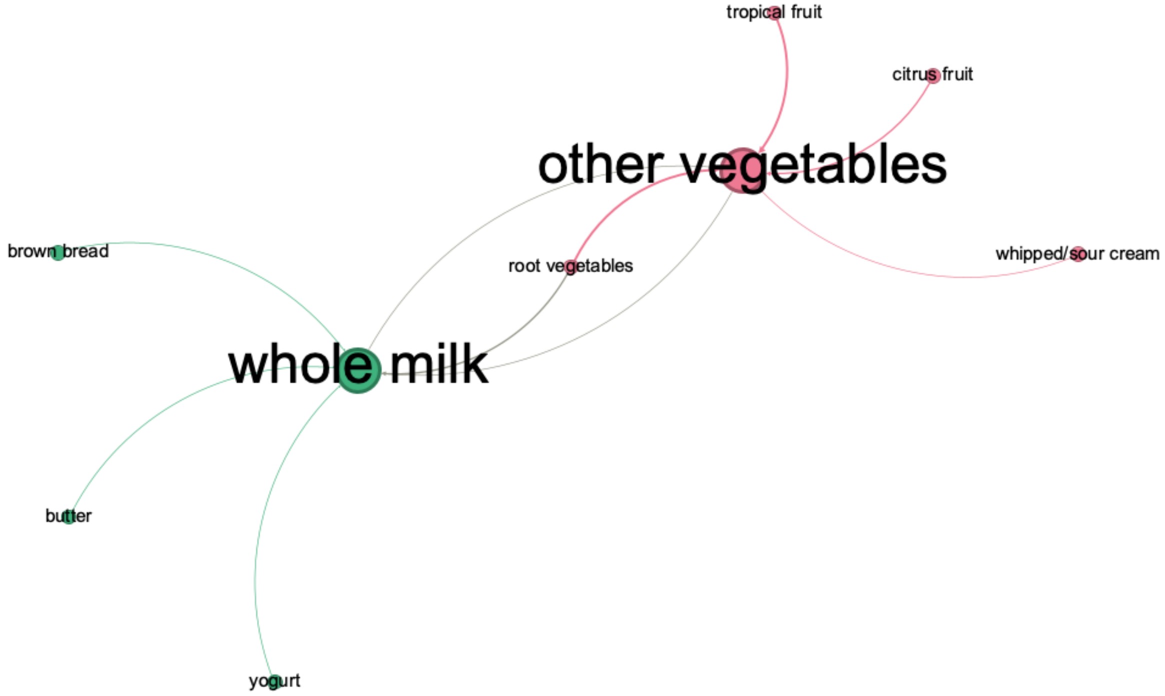


Figure 1: Gephi Groceries Rules (Lift > 3, Support > 0.3%)

```r
#extract the max lift
max_lift_rule <- groceries_rules[which.max(groceries_rules@quality$lift)]
inspect(max_lift_rule)
```

```
##     lhs                          rhs            support     confidence
## [1] {liquor, red/blush wine} => {bottled beer} 0.001931876 0.9047619
##     coverage    lift     count
## [1] 0.002135231 11.23527 19
```

```r
#extract the max support
max_support_rule <- groceries_rules[which.max(groceries_rules@quality$support)]
inspect(max_support_rule)
```

```
##     lhs                     rhs                support confidence    coverage     lift count
## [1] {root vegetables,
##      tropical fruit,
##      yogurt}           => {whole milk} 0.00569395        0.7 0.008134215 2.739554    56
```

```r
#extract the max confidence
max_confidence_rule <- groceries_rules[which.max(groceries_rules@quality$confidence)]
inspect(max_confidence_rule)
```

```
##     lhs               rhs          support     confidence coverage    lift
## [1] {rice, sugar} => {whole milk} 0.001220132 1          0.001220132 3.913649
##     count
## [1] 12
```

Notable Examples:

Best Balance of High Lift and High Support (citrus fruit, root vegetables, tropical fruit → other vegetables)

- Support = 0.0045. This item set appears in 0.45% of the transactions. (45 instances).
- Confidence = 0.79. If a customer purchases the antecedent, they also purchase other vegetables 79% of the time.
- Lift = 4.06. Buying citrus fruit, root vegetables, tropical fruit, makes the buying other vegetables 4 times more likely compared to the overall baseline probability.
- Interpretation: This insight makes sense because of the need for fresh produce in cooking.

Highest Lift: 11 (liquor, red/blush wine → bottled beer)

- Support = 0.0019 This item set appears in 0.19% of the transactions. (19 instances).
- Confidence = 0.91. If a customer purchases liquor or red/blush wine, they also purchase bottled beer 91% of the time.
- Lift = 11.24. Buying liquor and red/blush wine makes the buying bottled beer 11 times more likely compared to the overall baseline probability.
- Interpretation: This is not surprising when considering drinking culture in the United States.

Highest Support: 0.57% (root vegetables, tropical fruit, yogurt → whole milk)

- Support = 0.0057 This item set appears in 0.45% of the transactions. (45 instances).
- Confidence = 0.7. If a customer purchases the antecedent, they also purchase whole milk 70% of the time.
- Lift = 2.74. Buying root vegetables, tropical fruit, yogurt makes the buying whole milk 2.74 times more likely compared to the overall baseline probability.
- Interpretation: This rule is valuable because it has a relatively high support and lift, indicating that these items are commonly bought together in a significant number of transactions.

Highest Confidence: 1 (rice, sugar → whole milk)

- Support = 0.0012. This item set appears in 0.12% of the transactions (12 instances).
- Confidence = 1. If a customer purchases rice and sugar, they also purchase whole milk 100% of the time.
- Lift = 3.91. Buying rice and sugar makes the buying whole milk 3.91 times more likely compared to the overall baseline probability.
- Interpretation: This is an unexpected rule to have perfect confidence, and it is interesting for further investigation.

Takeaways:

- The above rules are all important to consider depending on the business goal. Whether it is to capture the greatest number of customers (highest support), be the most certain about their investment (highest confidence), have targeted marketing and upselling (highest lift), or get a good balance of both. Most of these above rules, however, have a strikingly good balance of all three.