Association rule mining

Load and Process the Data

```
## Summary of the structure of the data after processing


## 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
```

Prior to performing any analysis on the data, there was the challenge of loading and processing the data, especially since the data was unstructured in a .txt file. Therefore, after reading in the lines of the data, I split each string by commas into a list. Afterwards, I converted the list of baskets into a "transactions" format which is used by the arules package for association rule mining.

Run the Apriori Algorithm

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##         0.2    0.1    1 none FALSE            TRUE       5    0.01      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: 98
```

```
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[169 item(s), 9835 transaction(s)] done [0.00s].
## sorting and recoding items ... [88 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 4 done [0.00s].
## writing ... [232 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].

## 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
```
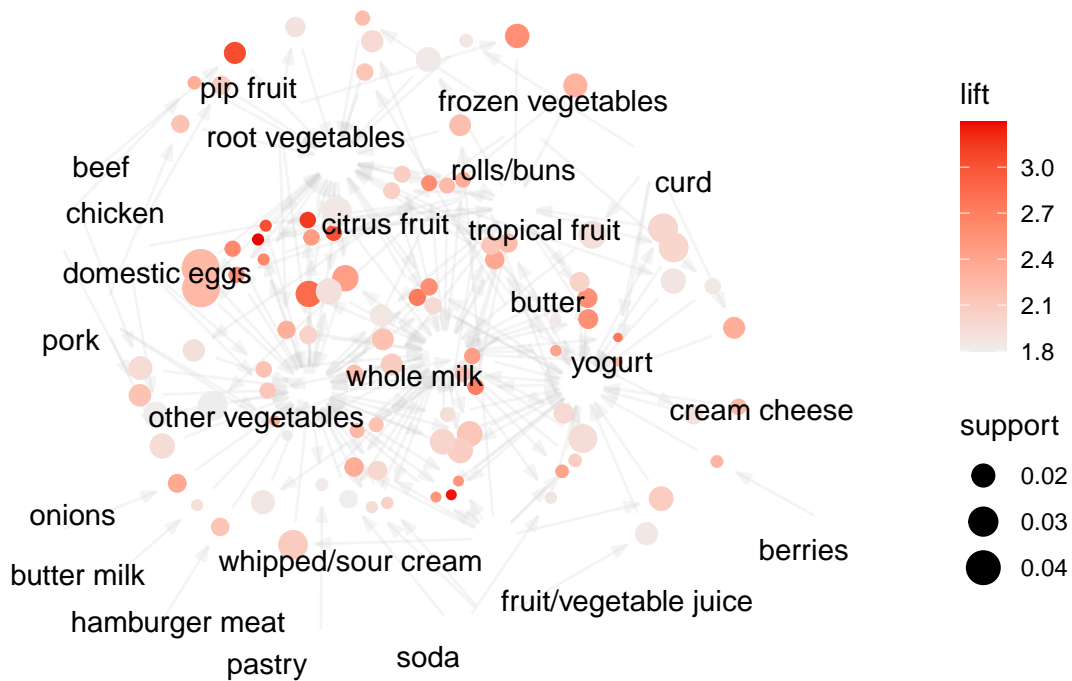


```
## Plot of associations, where lift >3 & confidence > 0.3

## 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
```



```
## Table of associations, where lift >3 & confidence > 0.3

##     lhs                               rhs                 support
## [1] {beef}                         => {root vegetables}  0.01738688
## [2] {citrus fruit, root vegetables}    => {other vegetables} 0.01037112
## [3] {citrus fruit, other vegetables}   => {root vegetables}  0.01037112
## [4] {root vegetables, tropical fruit}  => {other vegetables} 0.01230300
## [5] {other vegetables, tropical fruit} => {root vegetables}  0.01230300
##     confidence coverage   lift     count
## [1] 0.3313953  0.05246568 3.040367 171
## [2] 0.5862069  0.01769192 3.029608 102
## [3] 0.3591549  0.02887646 3.295045 102
## [4] 0.5845411  0.02104728 3.020999 121
## [5] 0.3427762  0.03589222 3.144780 121
```

The table above focuses on rules where:

- **Lift > 3**: The items in the rule are at least three times more likely to be bought together than by random chance.

- **Confidence > 0.3**: There's at least a 30% chance that when the items on the left-hand side (LHS) are bought, the items on the right-hand side (RHS) are also bought.

I picked these thresholds because I wanted to focus on the more interesting, meaningful rules. For example, rules with high support might just reflect common items (like whole milk), but rules with high lift and confidence can reveal stronger and more unexpected associations between items. Also, this helps reduce complexity in the number of rules, allowing us to focus on the most insightful rules.

Association Rules Breakdown:

Rule 1: {beef} => {root vegetables}

- **Support: 0.0174 (1.74%)**

  This means that 1.74% of all transactions in the dataset include both "beef" and "root vegetables."

- **Confidence: 0.3314 (33.14%)**

  This indicates that 33.14% of the transactions that include "beef" also include "root vegetables."

- **Lift: 3.04**

  The lift value of 3.04 suggests that customers who buy "beef" are 3.04 times more likely to also buy "root vegetables" compared to the likelihood of buying "root vegetables" independent of "beef."

This rule indicates a strong association between "beef" and "root vegetables." If a customer buys beef, there's a significant likelihood (33.14%) that they will also buy root vegetables, and this combination is more likely to occur than if the two items were bought independently. This item set makes sense since many cooking recipes use both beef and root vegetables together, like beef stew.

Rule 2: {citrus fruit, root vegetables} => {other vegetables}

- **Support: 0.0104 (1.04%)**

  1.04% of all transactions include the combination of "citrus fruit," "root vegetables," and "other vegetables."

- **Confidence: 0.5862 (58.62%)**

  58.62% of the transactions that include both "citrus fruit" and "root vegetables" also include "other vegetables."

- **Lift: 3.03**

  The lift value of 3.03 means that customers who buy "citrus fruit" and "root vegetables" are 3.03 times more likely to also buy "other vegetables" compared to the general likelihood of buying "other vegetables."

This rule highlights a strong association between these three items. If a customer buys citrus fruit and root vegetables, there's a high chance (58.62%) that they will also purchase other vegetables.

Rule 3: {citrus fruit, other vegetables} => {root vegetables}

- **Support: 0.0104 (1.04%)**

  1.04% of transactions include "citrus fruit," "other vegetables," and "root vegetables."

- **Confidence: 0.3592 (35.92%)**

  35.92% of the transactions that include both "citrus fruit" and "other vegetables" also include "root vegetables."

- **Lift: 3.30** The lift value of 3.30 indicates that customers who buy "citrus fruit" and "other vegetables" are 3.30 times more likely to also buy "root vegetables" compared to the general likelihood of buying "root vegetables."

This rule shows a moderately strong association. If a customer buys citrus fruit and other vegetables, there's a 35.92% chance they will also buy root vegetables.

Rule 4: {root vegetables, tropical fruit} => {other vegetables}

- **Support: 0.0123 (1.23%)**

  1.23% of all transactions include "root vegetables," "tropical fruit," and "other vegetables."

- **Confidence: 0.5845 (58.45%)**

  58.45% of the transactions that include both "root vegetables" and "tropical fruit" also include "other vegetables."

- **Lift: 3.02**

  The lift value of 3.02 suggests that customers who buy "root vegetables" and "tropical fruit" are 3.02 times more likely to also buy "other vegetables" than the average likelihood of buying "other vegetables."

There is a strong association between these items. If a customer buys root vegetables and tropical fruit, they are highly likely (58.45%) to also purchase other vegetables.

Rule 5: {other vegetables, tropical fruit} => {root vegetables}

- **Support: 0.0123 (1.23%)**

  1.23% of transactions include "other vegetables," "tropical fruit," and "root vegetables."

- **Confidence: 0.3428 (34.28%)**

  34.28% of the transactions that include both "other vegetables" and "tropical fruit" also include "root vegetables."

- **Lift: 3.14** The lift value of 3.14 means that customers who buy "other vegetables" and "tropical fruit" are 3.14 times more likely to also buy "root vegetables" than the average likelihood of buying "root vegetables."

This rule shows a moderate association. Customers buying other vegetables and tropical fruit have a 34.28% chance of also buying root vegetables.

Rules 2-5 center around similar food items: fruit and vegetables. It makes sense to me that they have a strong association with each other, both being plant-based and generally considered healthy staples of a person's diet.