Association

Amey Ghate

2023-08-10

Association rule mining

Problem

For data on grocery purchases in groceries.txt, 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.

Presenting the structure of the raw dataset:

```
## [1] "citrus fruit,semi-finished bread,margarine,ready soups"
## [2] "tropical fruit,yogurt,coffee"
## [3] "whole milk"
## [4] "pip fruit,yogurt,cream cheese ,meat spreads"
## [5] "other vegetables,whole milk,condensed milk,long life bakery product"
## [6] "whole milk,butter,yogurt,rice,abrasive cleaner"
```

We transform the data into a "transactions" class before applying the apriori algorithm. The summary of the dataset reveals the following:

1. There are total of 9835 transactions in our dataset
2. Whole milk is the present in 2513 baskets and is the most frequently bought item
3. More than half of the transactions have 4 or lesser items per basket
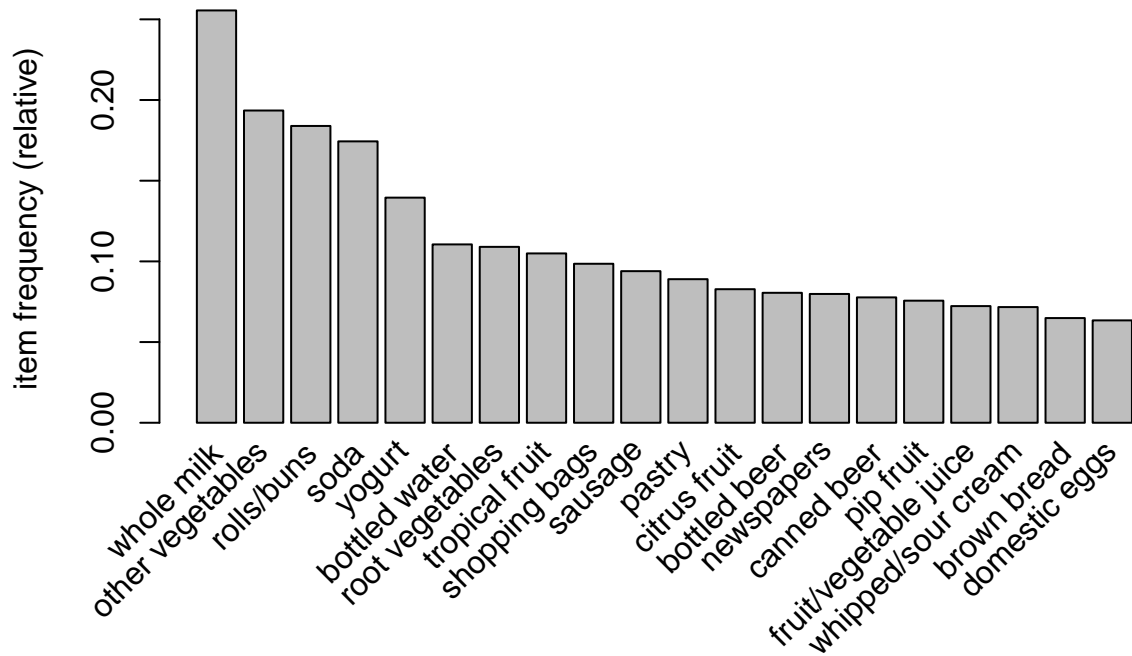
Let's explore rules with support > 0.05, confidence > 0.1 and length <= 2 using the 'apriori' algorithm There are only 6 rules generated because of the high support and low confidence level. We also notice that most relationships in this item set include whole milk, yogurt and rolls/buns which is in accordance with the transaction frequency plot we saw earlier. These are some of the most frequently bought items.

```
##      lhs                   rhs                support    confidence coverage
## [1] {yogurt}           => {whole milk}        0.05602440 0.4016035  0.1395018
## [2] {whole milk}       => {yogurt}            0.05602440 0.2192598  0.2555160
## [3] {rolls/buns}       => {whole milk}        0.05663447 0.3079049  0.1839349
## [4] {whole milk}       => {rolls/buns}        0.05663447 0.2216474  0.2555160
## [5] {other vegetables} => {whole milk}        0.07483477 0.3867578  0.1934926
## [6] {whole milk}       => {other vegetables}  0.07483477 0.2928770  0.2555160
##      lift      count
## [1] 1.571735  551
## [2] 1.571735  551
## [3] 1.205032  557
## [4] 1.205032  557
## [5] 1.513634  736
## [6] 1.513634  736
```
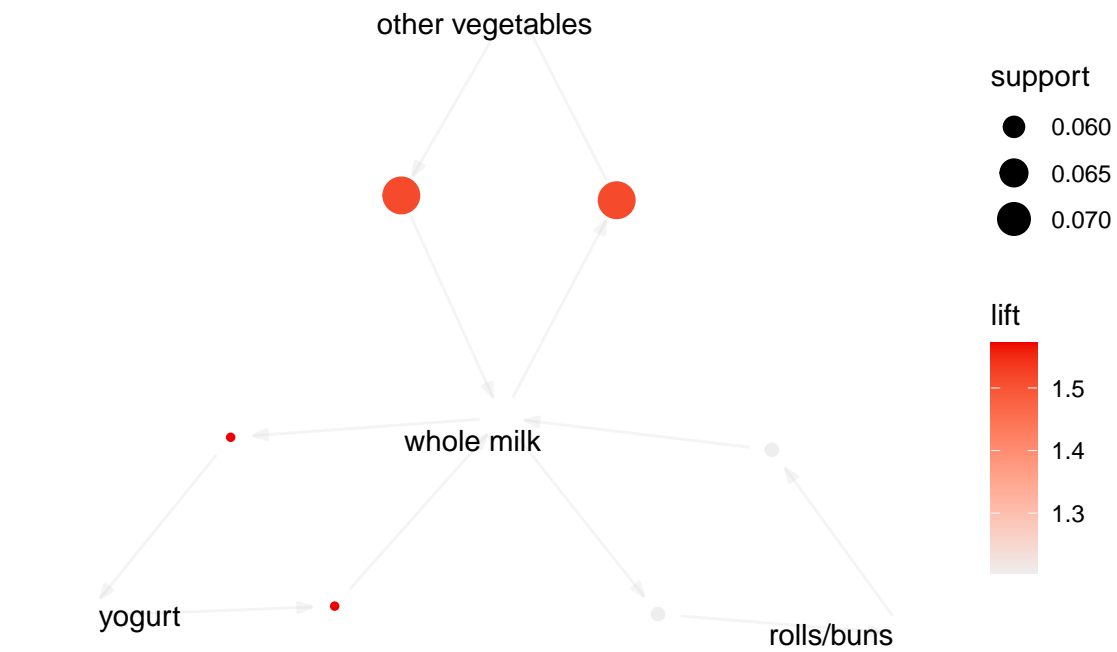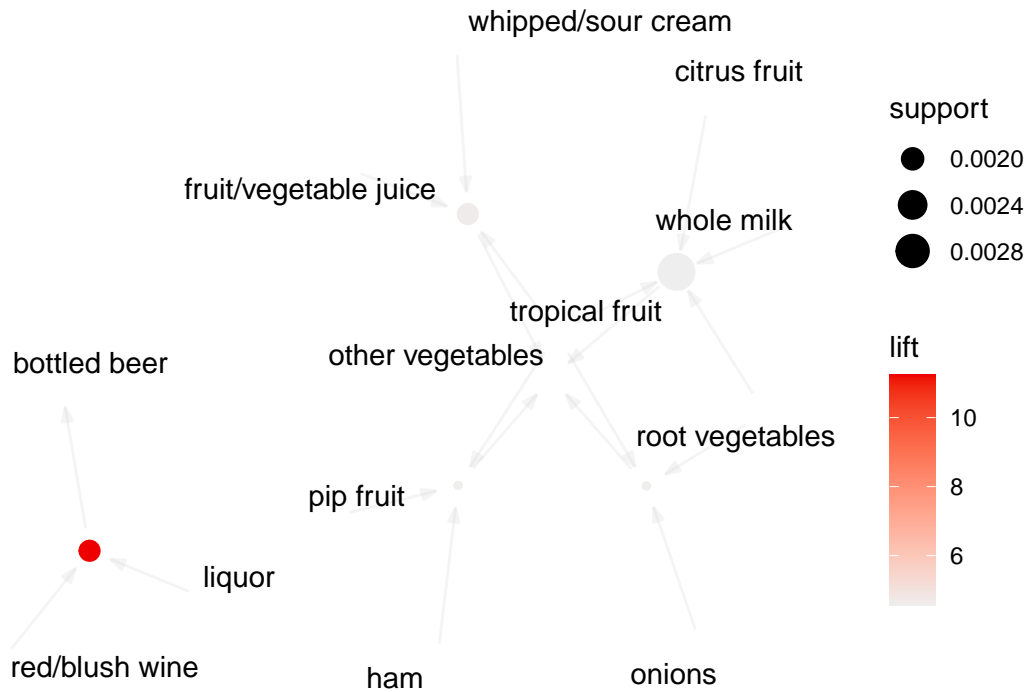
Let's decrease support further and increase confidence slightly with support > 0.02, confidence > 0.2 and length <= 2 This item set contains 72 rules and includes a lot more items. However, whole milk still seems to be a common occurence.
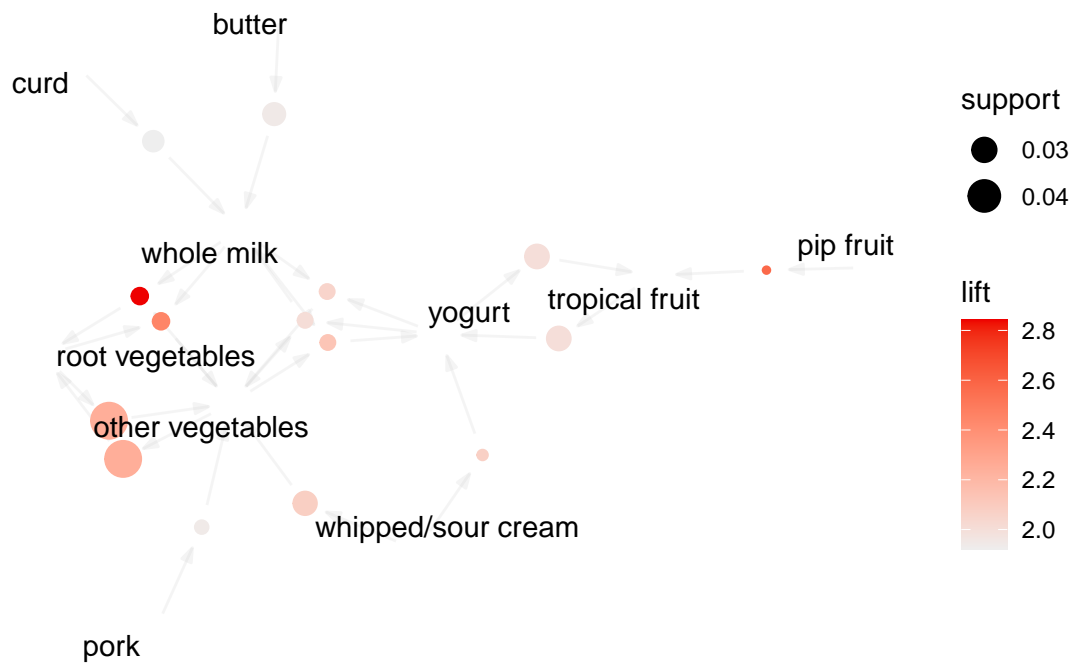
Let us increase the confidence level and decrease the support further. Let's explore rules with support > 0.0015, confidence > 0.8 and length <= 2

Summary From the derived association rules and presented graphs, we can infer the following:

1. Customers purchasing red wine or liquor often gravitate towards bottled beer.
2. Individuals who purchase vegetables like root vegetables are more inclined to buy additional vegetables, much like those who buy vegetable/fruit juice.
3. Whole milk stands out as the most frequently bought item among customers. Additionally, when individuals purchase dairy products such as curd, yogurt or butter, they are more likely to also buy whole milk.

The aforementioned association rules are logical when considering consumer behavior. Typically, shoppers exhibit patterns in their purchasing habits, often buying complementary or similar items during the same shopping trip. For instance, someone buying certain types of vegetables might be inclined to purchase other related vegetables or products that fit into their meal plans or dietary preferences.