DATA_624_HW2

Market Basket Analysis

Imagine 10000 receipts sitting on your table. Each receipt represents a transaction with items that were purchased. The receipt is a representation of stuff that went into a customer's basket – and therefore 'Market Basket Analysis'.

That is exactly what the Groceries Data Set contains: a collection of receipts with each line representing 1 receipt and the items purchased. Each line is called a **transaction** and each column in a row represents an **item**.

You assignment is to use R to mine the data for association rules. You should report support, confidence and lift and your top 10 rules by lift.

Association Rules in **R**

```
library(arules)
library(arulesViz)
```

First, let's load the transactions data.

```
grocery_data <- read.transactions("GroceryDataSet.csv",format="basket", sep=",")
```

An association rule is a specific pattern derived from existing observations that an antecedent action leads to a consequent action. In this case, we would like to find rules that people who purchase a given set of items also purchase another set of items.

We will use three measures to evaluate our association rules:

- Support: the percentage of the transactions that contain all items in an association rule.
- Confidence: given the transactions that a specific set of items is purchased, the probability that another set of items is also included in the transactions.
- Lift: a measure of the strength of the causation effect. A lift = 1 implies the presence of the two sets of items are independent. A lift < 1 implies that the presence of one set has negative effect on the presence of another set. A lift > 1 shows a good positive causation effect.

We have a total of 9835 transactions.

```
length(grocery_data)
```

```
## [1] 9835
```

We would like to filter out some rules that are not significantly important. Hence, we would capture only the rules occur in 10 or more transactions. That is, support must be equal or higher than 0.001.

```
10/9835
```

```
## [1] 0.001016777
```

A grocery store has limited resources to improve their business. We may find hundreds of rules but the store can only manage some items. Hence, would focus on the rules with high confidence (equal or greater than 0.8). We would also add a restriction to generate rules with three or less items.

Setting all the parameter, we can generate the association rules using the apriori function.

```
rules <- apriori(grocery_data, parameter = list(supp = 10/9835, conf = 0.8, minlen = 2, maxlen= 3))
```

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime     support minlen
##         0.8    0.1    1 none FALSE           TRUE       5 0.001016777      2
##  maxlen target  ext
##       3  rules TRUE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 10
##
## 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 done [0.00s].
## writing ... [29 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

```
summary(rules)
```

```
## set of 29 rules
##
## rule length distribution (lhs + rhs):sizes
##  3
## 29
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       3       3       3       3       3       3
##
## summary of quality measures:
##     support         confidence        coverage            lift
##  Min.   :0.001017   Min.   :0.8000   Min.   :0.001118   Min.   : 3.131
##  1st Qu.:0.001118   1st Qu.:0.8125   1st Qu.:0.001220   1st Qu.: 3.261
##  Median :0.001220   Median :0.8462   Median :0.001525   Median : 3.613
##  Mean   :0.001473   Mean   :0.8613   Mean   :0.001732   Mean   : 4.000
##  3rd Qu.:0.001729   3rd Qu.:0.9091   3rd Qu.:0.002135   3rd Qu.: 4.199
##  Max.   :0.002542   Max.   :1.0000   Max.   :0.003152   Max.   :11.235
##      count
##  Min.   :10.00
##  1st Qu.:11.00
##  Median :12.00
##  Mean   :14.48
##  3rd Qu.:17.00
##  Max.   :25.00
##
## mining info:
##          data ntransactions     support confidence
##  grocery_data          9835 0.001016777        0.8
##                                                                                                call
##  apriori(data = grocery_data, parameter = list(supp = 10/9835, conf = 0.8, minlen = 2, maxlen
= 3))
```

We can show top 10 rules with the highest lift but we find it more interesting to view the top 20 rules.

```
kable(DATAFRAME(head(rules, n = 20, by = "lift", decreasing=T)), row.names = FALSE)
```

LHS	RHS	support	confidence	coverage	lift	count
{liquor,red/blush wine}	{bottled beer}	0.0019319	0.9047619	0.0021352	11.235269	19
{grapes,onions}	{other vegetables}	0.0011185	0.9166667	0.0012201	4.737476	11
{hard cheese,oil}	{other vegetables}	0.0011185	0.9166667	0.0012201	4.737476	11
{butter milk,pork}	{other vegetables}	0.0018302	0.8571429	0.0021352	4.429848	18
{margarine,meat}	{other vegetables}	0.0017285	0.8500000	0.0020336	4.392932	17
{rice,yogurt}	{other vegetables}	0.0019319	0.8260870	0.0023386	4.269346	19
{herbs,shopping bags}	{other vegetables}	0.0019319	0.8260870	0.0023386	4.269346	19
{butter milk,onions}	{other vegetables}	0.0013218	0.8125000	0.0016268	4.199126	13
{curd,turkey}	{other vegetables}	0.0012201	0.8000000	0.0015252	4.134524	12
{fruit/vegetable juice,herbs}	{other vegetables}	0.0012201	0.8000000	0.0015252	4.134524	12

LHS	RHS	support	confidence	coverage	lift	count
{onions,waffles}	{other vegetables}	0.0012201	0.8000000	0.0015252	4.134524	12
{rice,sugar}	{whole milk}	0.0012201	1.0000000	0.0012201	3.913649	12
{canned fish,hygiene articles}	{whole milk}	0.0011185	1.0000000	0.0011185	3.913649	11
{house keeping products,whipped/sour cream}	{whole milk}	0.0012201	0.9230769	0.0013218	3.612599	12
{bottled water,rice}	{whole milk}	0.0012201	0.9230769	0.0013218	3.612599	12
{bottled beer,soups}	{whole milk}	0.0011185	0.9166667	0.0012201	3.587512	11
{cereals,curd}	{whole milk}	0.0010168	0.9090909	0.0011185	3.557863	10
{pastry,sweet spreads}	{whole milk}	0.0010168	0.9090909	0.0011185	3.557863	10
{mustard,oil}	{whole milk}	0.0012201	0.8571429	0.0014235	3.354556	12
{chocolate,pickled vegetables}	{whole milk}	0.0012201	0.8571429	0.0014235	3.354556	12

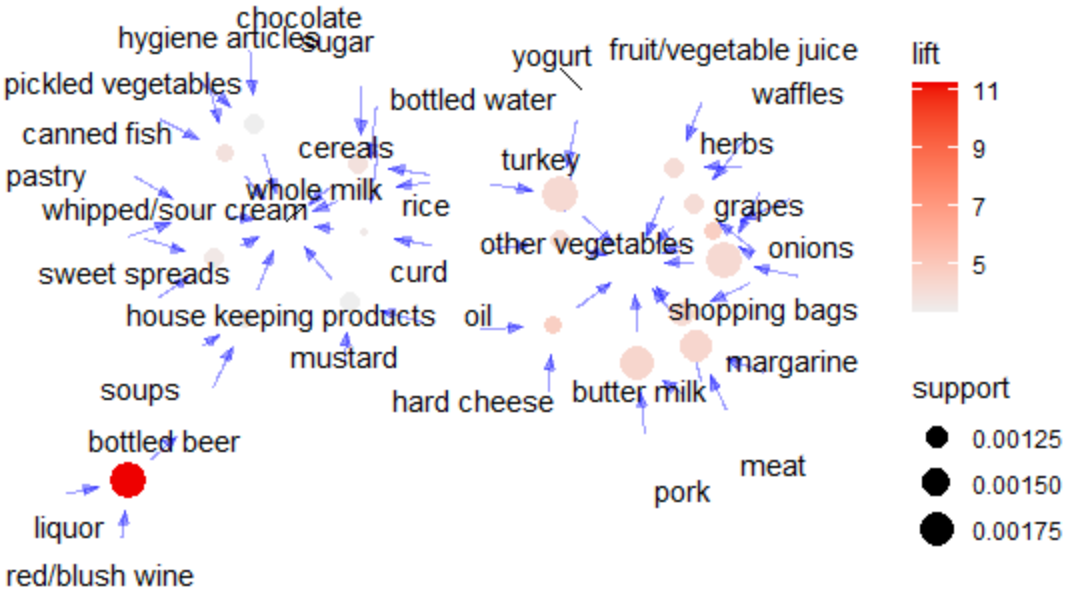
- The first rule shows the purchases of alcohol drinkers. People who purchase **liquor and wines** are very likely to purchase **beers** also. The lift is significantly higher than any other rule. It has a confidence of 0.9 and the highest support among the top 20 rules. We can do a sales promotion on liquor and wines and increase the price of beers to increase our revenue. However, we should not push the sales of beers to increase the sales of liquor and wines since the rule is not one of the top rules.

- Top 2-11 have the same consequent item - **other vegetables**. Looking at the antecedent items of the rules, **other vegetables** may be an ingredient of a recipe with the other items. Also, the term **other vegetables** is vague as it can be a lot of different items. As **other vegetables** is needed by many customers, it would worth time to find out which specific vegetables are mostly needed and make sure we have enough in the inventory.
- Top 12-20 all point to the same item - **whole milk**. One may think that **whole milk** is an essential item that appears on most transactions just by chance. But the lift score of higher than 3 implies that there is a causation effect of the purchases. Unlike **other vegetables**, **whole milk** is a well defined single item that we can manage. The items on the left of the rules would be the good candidates on the our weekly advertisements.
- There is one interesting rule that shows people who purchase **bottled beer** and **soups** also purchase **whole milk**. We may utilize this chain effect to sell **liquor and wines**, **beers**, **soups**, and **whole milk** in a bundle.

We can visualize the top 20 rules use the arulesViz package.

We can confirm from the following plot that **other vegetables** and **whole milk** are the centers of the grocery items and the **alcohols** stick together.

```
plot(rules,method="graph", control=list(max=20,
                                edges = ggraph::geom_edge_link(
                                   end_cap = ggraph::circle(4, "mm"),
                                   start_cap = ggraph::circle(4, "mm"),
                                   color = "blue",
                                   arrow = arrow(length = unit(2, "mm"), angle = 20, type = "closed"),
                                   alpha = 0.4
                                  )
                                ))
```


chocolate
hygiene articles
sugar
yogurt
fruit/vegetable juice
pickled vegetables
bottled water
waffles
canned fish
cereals
herbs
turkey
pastry
whole milk
rice
grapes
whipped/sour cream
other vegetables
onions
sweet spreads
curd
house keeping products
oil
shopping bags
mustard
margarine
soups
hard cheese
butter milk
bottled beer
meat
pork
liquor
red/blush wine
lift
11
9
7
5
support
0.00125
0.00150
0.00175