# DATA 624 Homework10

*Bin Lin*

*2018-5-17*

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. Here is the dataset = GroceryDataSet.csv (comma separated file)

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.

For this assignment, I will need the R package called arules. This package can be used forn mining association ruls and frequent itemsets. From the following summary statistics, we know this dataset contains 9835 transactions and 169 items.

```
#install.packages("arules")
library(arules)
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'arules'
```

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

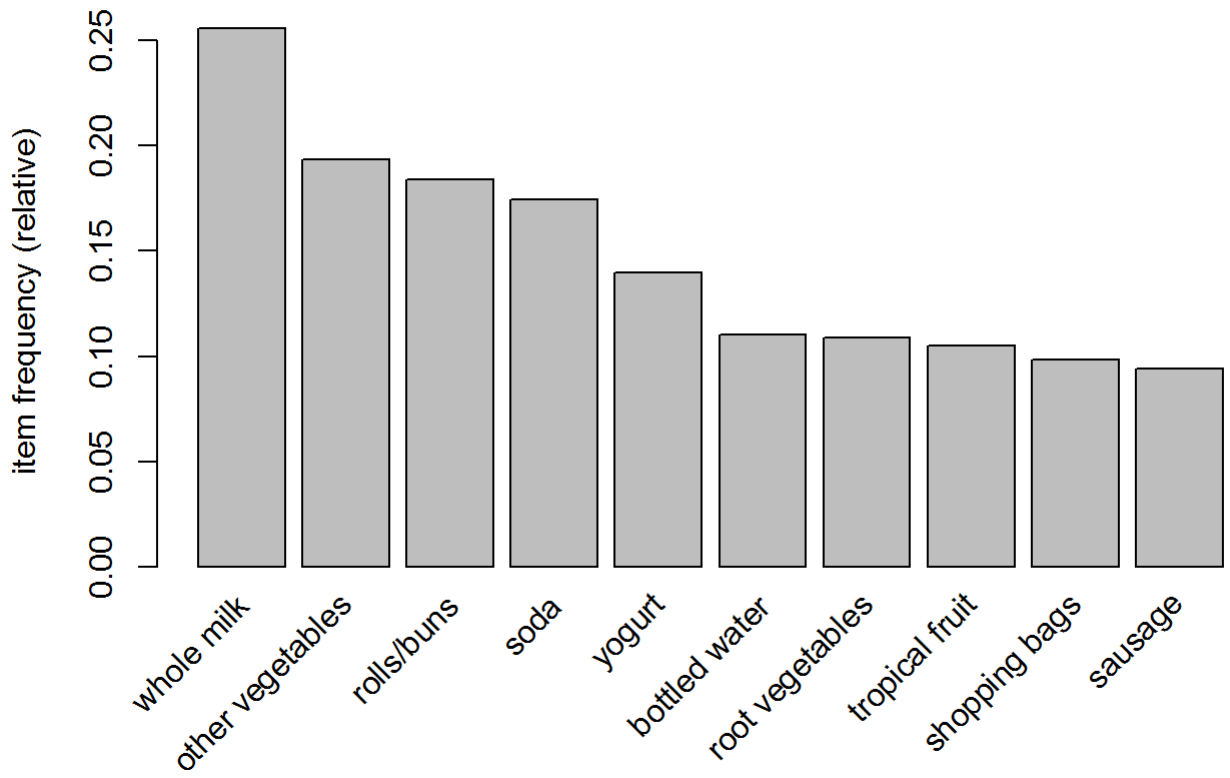
```
raw_data <- read.transactions("https://raw.githubusercontent.com/blin261/624/master/GroceryDataSet.csv", format = "basket", sep = ",")

summary(raw_data)
```

```
## 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
## 2159 1643 1299 1005  855  645  545  438  350  246  182  117   78   77   55
##   16   17   18   19   20   21   22   23   24   26   27   28   29   32
##   46   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
```

First of all, I would like to investigate the frequencies of each item and filter out the top 10 items using itemFrequencyPlot method.

```
itemFrequencyPlot(raw_data, topN = 10)
```




Association Rules: Support is an indication of how frequently the itemset appears in the dataset. Confidence is an indication of how often the rule has been found to be true. Lift is the ratio of the observed support to that expected

The following result shows the summary statistics of the apriori algorithms. There are total 410 rules with the length distributed from 3 to 6. The distribution for support, confidence, and lift are also shown as follows.

```
rules <- apriori(raw_data, parameter = list(supp = 0.001, confidence = 0.8, minlen = 2), control
 = list(verbose = FALSE))
summary(rules)
```

```
## set of 410 rules
##
## rule length distribution (lhs + rhs):sizes
##   3   4   5   6
##  29 229 140  12
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   4.000   4.000   4.329   5.000   6.000
##
## summary of quality measures:
##     support           confidence          lift            count
##  Min.   :0.001017   Min.   :0.8000   Min.   : 3.131   Min.   :10.00
##  1st Qu.:0.001017   1st Qu.:0.8333   1st Qu.: 3.312   1st Qu.:10.00
##  Median :0.001220   Median :0.8462   Median : 3.588   Median :12.00
##  Mean   :0.001247   Mean   :0.8663   Mean   : 3.951   Mean   :12.27
##  3rd Qu.:0.001322   3rd Qu.:0.9091   3rd Qu.: 4.341   3rd Qu.:13.00
##  Max.   :0.003152   Max.   :1.0000   Max.   :11.235   Max.   :31.00
##
## mining info:
##      data ntransactions support confidence
##  raw_data          9835   0.001        0.8
```

The following code will order all the rules by descending order based on lift. Then pick up the top 10 rules and investigate further.

```
inspect(head(rules, 10, by = "lift"))
```

```
##      lhs                          rhs                  support confidence      lift count
## [1]  {liquor,
##       red/blush wine}        => {bottled beer}    0.001931876  0.9047619 11.235269    19
## [2]  {citrus fruit,
##       fruit/vegetable juice,
##       other vegetables,
##       soda}                  => {root vegetables} 0.001016777  0.9090909  8.340400    10
## [3]  {oil,
##       other vegetables,
##       tropical fruit,
##       whole milk,
##       yogurt}                => {root vegetables} 0.001016777  0.9090909  8.340400    10
## [4]  {citrus fruit,
##       fruit/vegetable juice,
##       grapes}                => {tropical fruit}  0.001118454  0.8461538  8.063879    11
## [5]  {other vegetables,
##       rice,
##       whole milk,
##       yogurt}                => {root vegetables} 0.001321810  0.8666667  7.951182    13
## [6]  {oil,
##       other vegetables,
##       tropical fruit,
##       whole milk}            => {root vegetables} 0.001321810  0.8666667  7.951182    13
## [7]  {ham,
##       other vegetables,
##       pip fruit,
##       yogurt}                => {tropical fruit}  0.001016777  0.8333333  7.941699    10
## [8]  {beef,
##       citrus fruit,
##       other vegetables,
##       tropical fruit}        => {root vegetables} 0.001016777  0.8333333  7.645367    10
## [9]  {butter,
##       cream cheese,
##       root vegetables}       => {yogurt}          0.001016777  0.9090909  6.516698    10
## [10] {butter,
##       sliced cheese,
##       tropical fruit,
##       whole milk}            => {yogurt}          0.001016777  0.9090909  6.516698    10
```

Interpretation: The rule with the highest lift is for a purchase of bottled beer after purchase of liquor,red/blush wine. The rule has the lift value at 11.235. This pattern appears 19 times in the datasets.