R Notebook

Gordon Wall (gwall2)

loading relevant packages. . .

import and examine data. . .

```
groceries.df <- read.transactions("~/r-directory/grocery-basket/groceries_v2.csv", sep = ",")

summary(groceries.df)
```

```
## transactions as itemMatrix in sparse format with
##  9834 rows (elements/itemsets/transactions) and
##  169 columns (items) and a density of 0.0260911
##
## most frequent items:
##       whole milk other vegetables       rolls/buns             soda
##             2513             1902             1809             1715
##           yogurt          (Other)
##             1372            34051
##
## 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  854  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
```

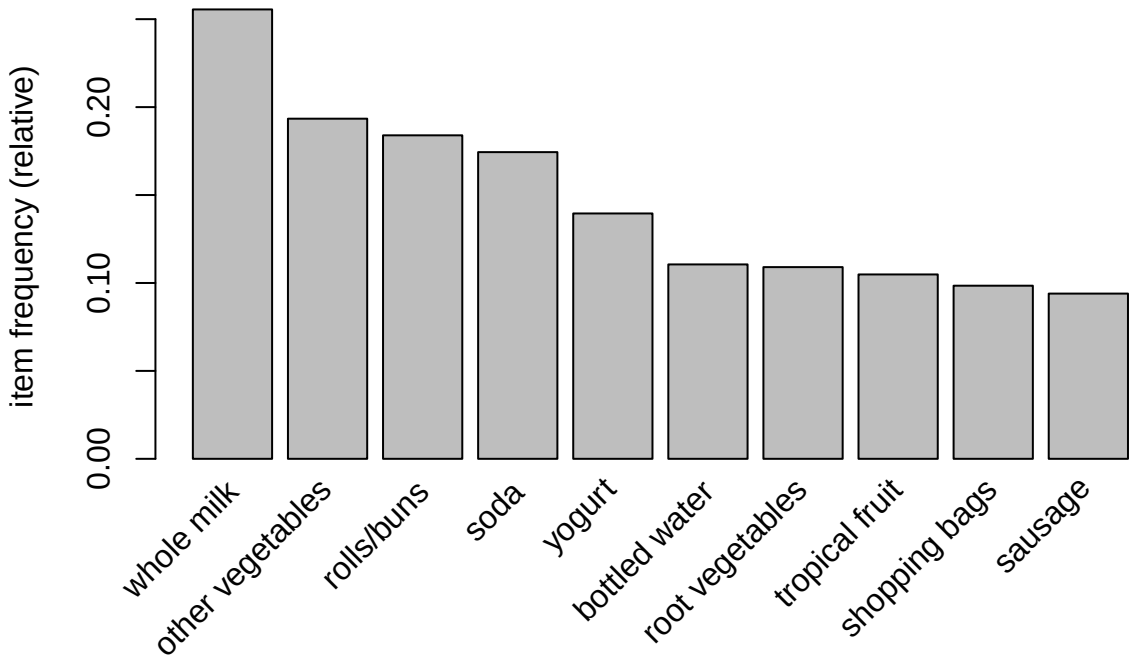
confirm that file was read properly by examining first 10 baskets

```
inspect(groceries.df[1:10])
```

```
##      items
## [1]  {citrus fruit,
##       margarine,
##       ready soups,
##       semi-finished bread}
## [2]  {coffee,
##       tropical fruit,
##       yogurt}
## [3]  {whole milk}
## [4]  {cream cheese,
##       meat spreads,
##       pip fruit,
##       yogurt}
## [5]  {condensed milk,
##       long life bakery product,
##       other vegetables,
##       whole milk}
## [6]  {abrasive cleaner,
##       butter,
##       rice,
##       whole milk,
##       yogurt}
## [7]  {rolls/buns}
## [8]  {bottled beer,
##       liquor (appetizer),
##       other vegetables,
##       rolls/buns,
##       UHT-milk}
## [9]  {pot plants}
## [10] {cereals,
##       whole milk}
```

further examine data; plot of top ten most frequent items in dataset

```
itemFrequencyPlot(groceries.df, topN = 10)
```



the transactions in this dataset most frequently sport purchases

of whole milk, followed by other vegetables and buns

train model and extract association rules

```
rules <- apriori(groceries.df, parameter = list(support =0.01, confidence = 0.5))
```

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##         0.5    0.1    1 none FALSE            TRUE       5    0.01      1
##  maxlen target   ext
##      10  rules FALSE
##
## 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), 9834 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 ... [15 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

evaluate model performance

```
summary(rules)
```

```
## set of 15 rules
##
## rule length distribution (lhs + rhs):sizes
##  3
## 15
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       3       3       3       3       3       3
##
## summary of quality measures:
##     support          confidence          lift           count
##  Min.   :0.01007   Min.   :0.5000   Min.   :1.984   Min.   : 99.0
##  1st Qu.:0.01174   1st Qu.:0.5151   1st Qu.:2.036   1st Qu.:115.5
##  Median :0.01230   Median :0.5245   Median :2.203   Median :121.0
##  Mean   :0.01316   Mean   :0.5411   Mean   :2.300   Mean   :129.4
##  3rd Qu.:0.01403   3rd Qu.:0.5718   3rd Qu.:2.432   3rd Qu.:138.0
##  Max.   :0.02227   Max.   :0.5862   Max.   :3.031   Max.   :219.0
##
## mining info:
##          data ntransactions support confidence
##  groceries.df          9834    0.01        0.5
```

there have been 15 rules extracted from the data set at 0.01 support and 0.5 confidence

the highest lift in this rule set is 3.031

check top 3 rules by lift

```
inspect(sort(rules, by = "lift")[1:3])
```

```
##     lhs                                rhs                support
## [1] {citrus fruit,root vegetables}   => {other vegetables} 0.01037218
## [2] {root vegetables,tropical fruit} => {other vegetables} 0.01230425
```

```
## [3] {rolls/buns,root vegetables}     => {other vegetables} 0.01220256
##     confidence lift     count
## [1] 0.5862069  3.030893 102
## [2] 0.5845411  3.022280 121
## [3] 0.5020921  2.595990 120
```

the number one rule suggests that people who buy citrus fruit and/or root vegetables

will also buy an item from the category "other vegetables" as well