Roy Chan
chanroyc@sfu.ca
301202770

CMPT 459 Fall 2017

DataMining

Martin Ester

TA: Zhilin Zhang

Programming Assignment 4

1. The following histogram was produced by using itemFrequencyPlot() function. The function produces a histogram of the number of items per transaction. In the following histogram, I have used the parameters topN=20 to only show the top 20 most frequent items per transaction with whole milk being the most frequent item purchased.
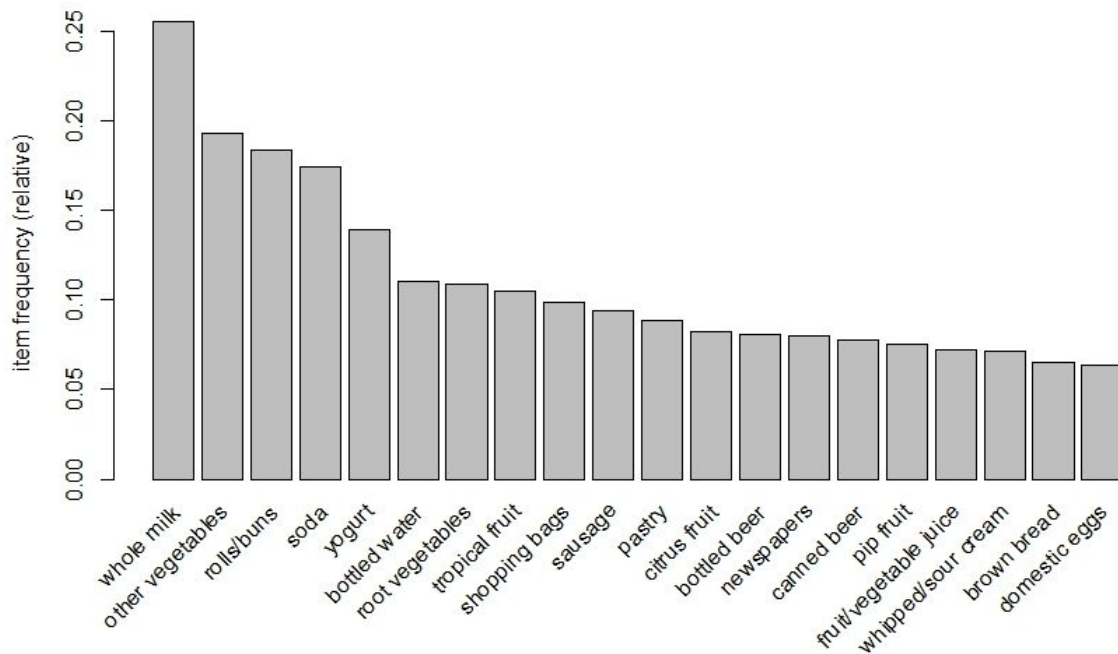
2. Using the apriori algorithm with minimum support of 0.001 produces the following results below. There are a total of 410 rules, some of which are shown in question 4.

```
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
 Groceries          9835   0.001        0.8
```

3. Using the apriori algorithm with minimum support of 0.01 produces no frequent item sets.

4. The 10 itemsets with the highest support and their support are shown in the figure below.

```
> inspect(rules2[1:10])
     lhs                                                 rhs                  support     confidence lift      count
[1]  {liquor,red/blush wine}                          => {bottled beer}       0.001931876 0.9047619  11.235269 19
[2]  {curd,cereals}                                   => {whole milk}         0.001016777 0.9090909   3.557863 10
[3]  {yogurt,cereals}                                 => {whole milk}         0.001728521 0.8095238   3.168192 17
[4]  {butter,jam}                                     => {whole milk}         0.001016777 0.8333333   3.261374 10
[5]  {soups,bottled beer}                             => {whole milk}         0.001118454 0.9166667   3.587512 11
[6]  {napkins,house keeping products}                 => {whole milk}         0.001321810 0.8125000   3.179840 13
[7]  {whipped/sour cream,house keeping products}      => {whole milk}         0.001220132 0.9230769   3.612599 12
[8]  {pastry,sweet spreads}                           => {whole milk}         0.001016777 0.9090909   3.557863 10
[9]  {turkey,curd}                                    => {other vegetables}   0.001220132 0.8000000   4.134524 12
[10] {rice,sugar}                                     => {whole milk}         0.001220132 1.0000000   3.913649 12
```

5. Support is an indication of how frequently the itemset appears in the dataset, however, a relatively small number of frequent items with the minimum support of 0.01 is not uncommon due to the context of the dataset. Groceries sell a variety of items, which explains why such a low minimum support would produce 0 rules. The frequent itemsets, closed frequent itemsets, and maximal frequent itemsets are similar, again, due to the context of the dataset. Customers often buy a specific collection of items together at groceries.

Roy Chan
chanroyc@sfu.ca
301202770

6. With a minimum confidence of 0.9, 0 association rules are obtained. Minimum confidence must be lowered to 0.5174 in order to obtain a set of 10 rules. Confidence must be strictly less than 0.5174 in order to obtain a set more than 10 rules as shown in the figure below.

```
> summary(rules6) #10 rules
set of 10 rules

rule length distribution (lhs + rhs):sizes
 3
10

   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.    :0.5175   Min.    :2.025   Min.    : 99.0
 1st Qu.:0.01103    1st Qu.:0.5315    1st Qu.:2.080    1st Qu.:108.5
 Median :0.01215    Median :0.5665    Median :2.217    Median :119.5
 Mean    :0.01202   Mean    :0.5577   Mean    :2.330   Mean    :118.2
 3rd Qu.:0.01261    3rd Qu.:0.5802    3rd Qu.:2.271    3rd Qu.:124.0
 Max.    :0.01454   Max.    :0.5862   Max.    :3.030   Max.    :143.0

mining info:
     data ntransactions support confidence
 Groceries          9835    0.01    0.5174
```

7. With minimum support of 0.01 and minimum confidence of 0.5, we get the following results and plot shown in the figures below.

```
> summary(rules7)
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    :2.299   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.    :3.030   Max.    :219.0

mining info:
     data ntransactions support confidence
 Groceries          9835    0.01    0.5
```

Roy Chan
chanroyc@sfu.ca
301202770

Scatter plot for 11 rules
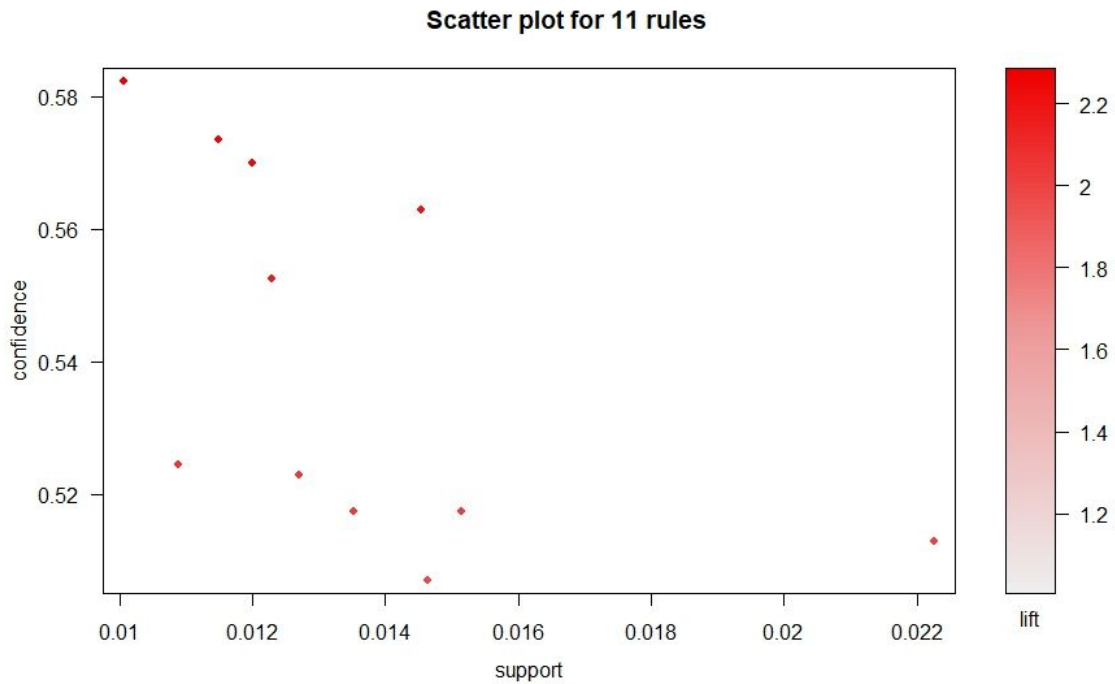
8. The figure below shows the rules produced in question 7. The rule with the highest lift is {curd, yogurt} at 2.279125. Lift values which are greater than 1 indicates that the discovered association rule is interesting. We can assume that these items are often purchased together.

```
> inspect(wholemilkrules)
     lhs                                        rhs              support    confidence lift     count
[1]  {curd,yogurt}                           => {whole milk} 0.01006609 0.5823529 2.279125  99
[2]  {other vegetables,butter}               => {whole milk} 0.01148958 0.5736041 2.244885 113
[3]  {other vegetables,domestic eggs}        => {whole milk} 0.01230300 0.5525114 2.162336 121
[4]  {yogurt,whipped/sour cream}             => {whole milk} 0.01087951 0.5245098 2.052747 107
[5]  {other vegetables,whipped/sour cream}   => {whole milk} 0.01464159 0.5070423 1.984385 144
[6]  {pip fruit,other vegetables}            => {whole milk} 0.01352313 0.5175097 2.025351 133
[7]  {tropical fruit,root vegetables}        => {whole milk} 0.01199797 0.5700483 2.230969 118
[8]  {tropical fruit,yogurt}                 => {whole milk} 0.01514997 0.5173611 2.024770 149
[9]  {root vegetables,yogurt}                => {whole milk} 0.01453991 0.5629921 2.203354 143
[10] {root vegetables,rolls/buns}            => {whole milk} 0.01270971 0.5230126 2.046888 125
[11] {other vegetables,yogurt}               => {whole milk} 0.02226741 0.5128806 2.007235 219
```