Association Rule Mining

STA380 Exam 2

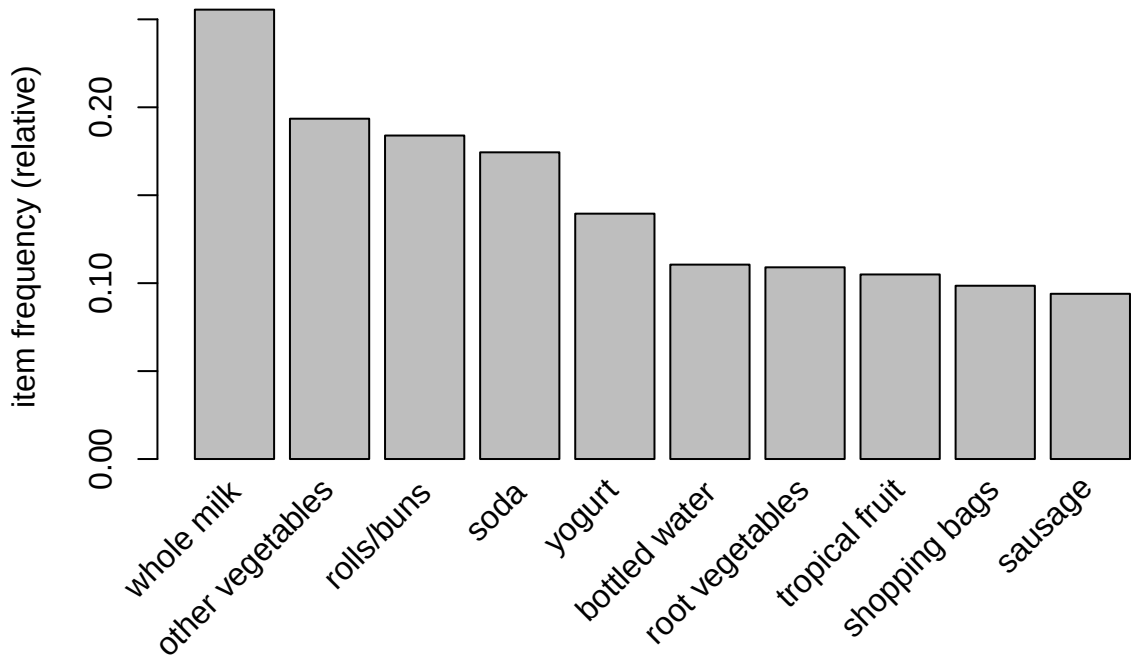
8/14/2020

```
# Read the contents of the text file
file_path <- "groceries.txt"
data <- readLines(file_path)

# Split the data into individual baskets
baskets <- strsplit(data, ",")

# Remove leading and trailing whitespace from items in each basket
baskets <- lapply(baskets, function(basket) trimws(basket))
```

```
# Convert the list of baskets into the "transactions" class
transactions <- as(baskets, "transactions")
summary(transactions)
```

```
## 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   16
## 2159 1643 1299 1005  855  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
```

```
itemFrequencyPlot(transactions, topN=10,  cex.names=1)
```




```
# Now run the apriori algorithm
groceryrules <- apriori(transactions,
                        parameter = list(support = 0.005, confidence = 0.1, maxlen = 4))
```

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##         0.1    0.1    1 none FALSE            TRUE       5   0.005      1
##  maxlen target  ext
##       4  rules TRUE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 49
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[169 item(s), 9835 transaction(s)] done [0.00s].
## sorting and recoding items ... [120 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 4
```

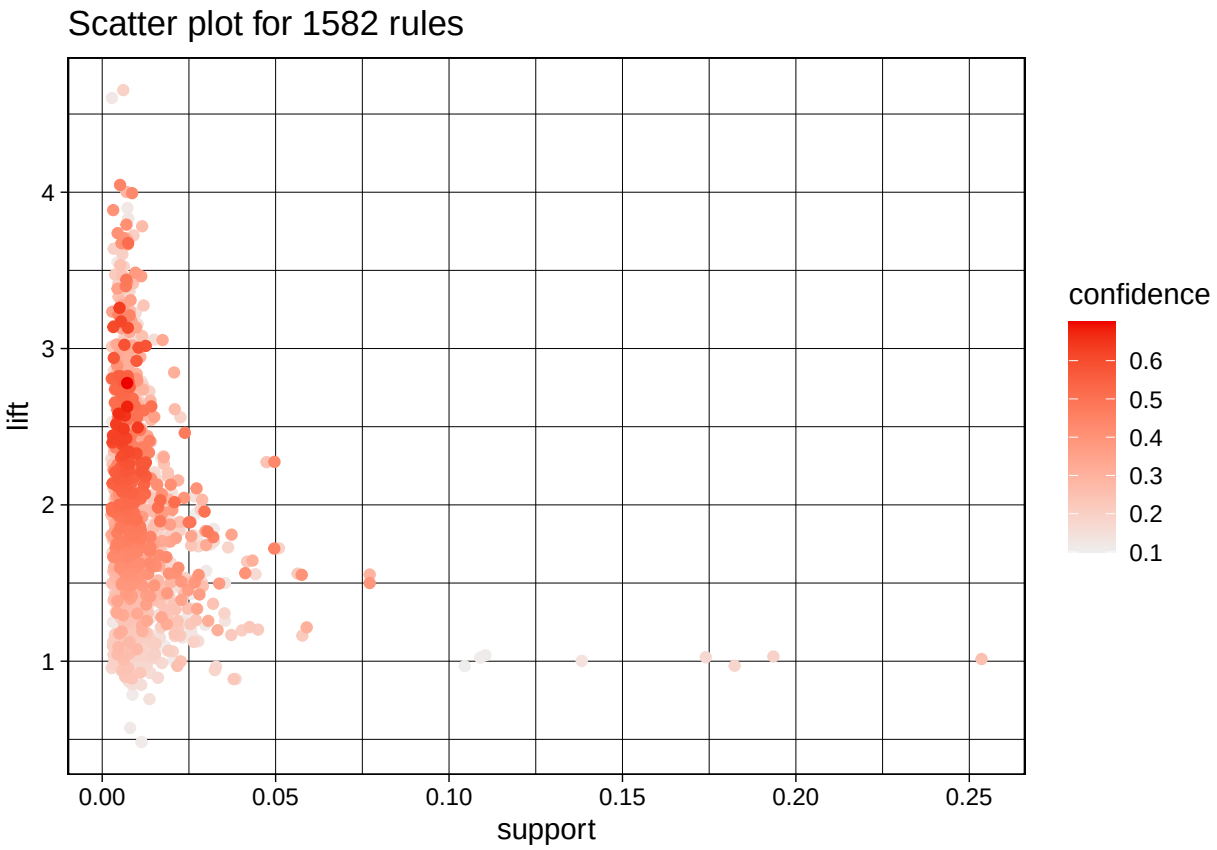
```
## Warning in apriori(transactions, parameter = list(support = 0.005, confidence =
## 0.1, : Mining stopped (maxlen reached). Only patterns up to a length of 4
## returned!
```

```
##  done [0.00s].
## writing ... [1582 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

Throughout all baskets analyzed in this dataset, the top 5 items that appear most frequently across all transactions are whole milk, other vegetables, rolls/buns, soda, and yogurt. We also found that for this dataset, the median number of items in each transaction was 3.

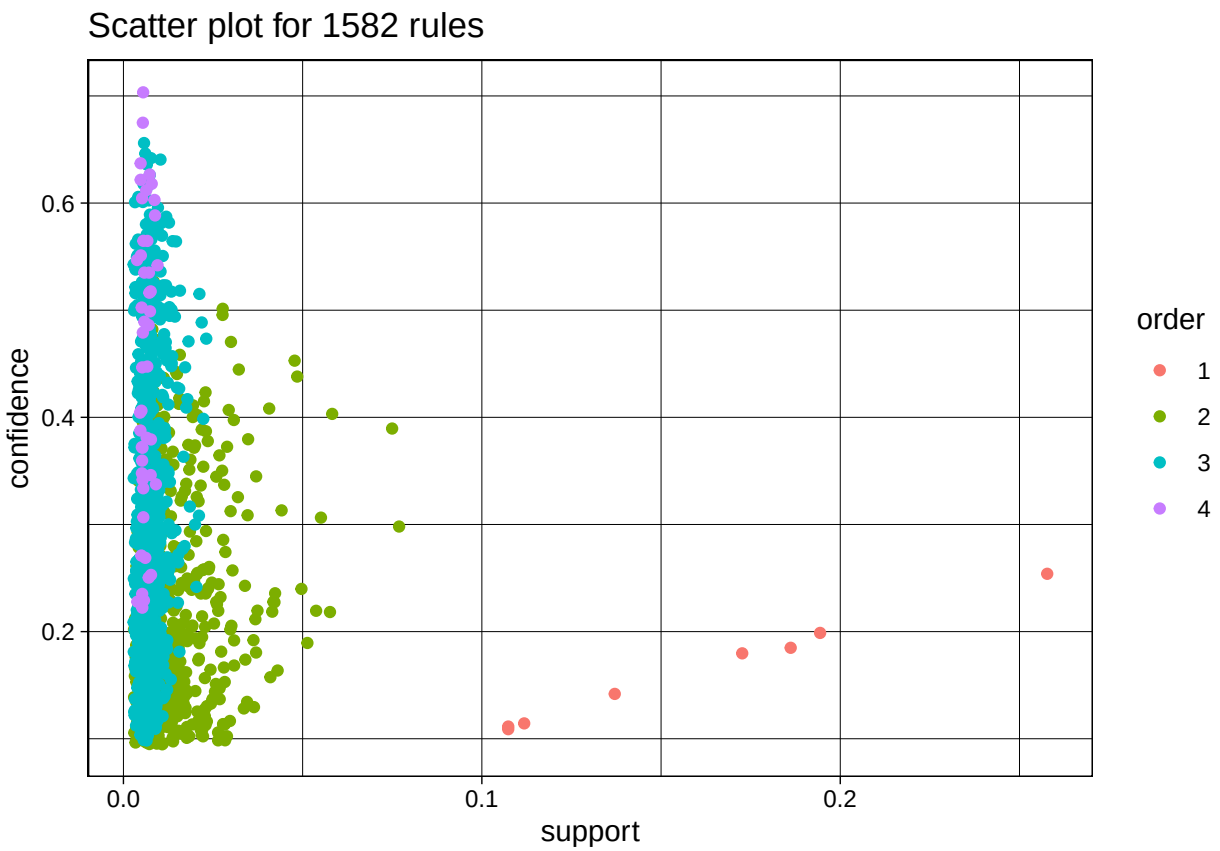
```
#Plot rules to determine best subsets

plot(groceryrules, measure = c("support", "lift"), shading = "confidence")
```

```
## To reduce overplotting, jitter is added! Use jitter = 0 to prevent jitter.
```




```
plot(groceryrules, method='two-key plot')
```

```
## To reduce overplotting, jitter is added! Use jitter = 0 to prevent jitter.
```



After plotting the rules, we found that high lift rules have low support, but that confidence is well spread. This might indicate some niche associations between items. We also see that as order increases, support decreases, indicating that longer and more complex rules capture more niche patterns of customer behavior, and represent less commong combinations of items in this dataset.

```
#inspect subsets
inspect(subset(groceryrules, subset=lift > 3.5))
```

```
##      lhs                     rhs                    support confidence   coverage     lift count
## [1]  {herbs}                 => {root vegetables}    0.007015760  0.4312500 0.01626843 3.956477    69
## [2]  {ham}                   => {white bread}        0.005083884  0.1953125 0.02602949 4.639851    50
## [3]  {white bread}           => {ham}                0.005083884  0.1207729 0.04209456 4.639851    50
## [4]  {berries}               => {whipped/sour cream} 0.009049314  0.2721713 0.03324860 3.796886    89
## [5]  {whipped/sour cream}    => {berries}            0.009049314  0.1262411 0.07168277 3.796886    89
## [6]  {hygiene articles}      => {napkins}            0.006100661  0.1851852 0.03294357 3.536498    60
## [7]  {napkins}               => {hygiene articles}   0.006100661  0.1165049 0.05236401 3.536498    60
## [8]  {onions,
##       other vegetables}      => {root vegetables}    0.005693950  0.4000000 0.01423488 3.669776    56
## [9]  {other vegetables,
##       root vegetables}       => {onions}             0.005693950  0.1201717 0.04738180 3.875044    56
## [10] {beef,
##       other vegetables}      => {root vegetables}    0.007930859  0.4020619 0.01972547 3.688692    78
## [11] {curd,
##       tropical fruit}        => {yogurt}             0.005287239  0.5148515 0.01026945 3.690645    52
## [12] {domestic eggs,
##       whole milk}            => {butter}             0.005998983  0.2000000 0.02999492 3.609174    59
```

```
## [13] {butter,
##       other vegetables}    => {whipped/sour cream} 0.005795628  0.2893401 0.02003050 4.036397    57
## [14] {other vegetables,
##       whipped/sour cream}  => {butter}             0.005795628  0.2007042 0.02887646 3.621883    57
## [15] {whipped/sour cream,
##       whole milk}          => {butter}             0.006710727  0.2082019 0.03223183 3.757185    66
## [16] {citrus fruit,
##       pip fruit}           => {tropical fruit}     0.005592272  0.4044118 0.01382816 3.854060    55
## [17] {citrus fruit,
##       tropical fruit}      => {pip fruit}          0.005592272  0.2806122 0.01992883 3.709437    55
## [18] {other vegetables,
##       whole milk,
##       yogurt}              => {whipped/sour cream} 0.005592272  0.2511416 0.02226741 3.503514    55
## [19] {other vegetables,
##       pip fruit,
##       whole milk}          => {root vegetables}    0.005490595  0.4060150 0.01352313 3.724961    54
## [20] {citrus fruit,
##       other vegetables,
##       whole milk}          => {root vegetables}    0.005795628  0.4453125 0.01301474 4.085493    57
## [21] {root vegetables,
##       whole milk,
##       yogurt}              => {tropical fruit}     0.005693950  0.3916084 0.01453991 3.732043    56
## [22] {other vegetables,
##       tropical fruit,
##       whole milk}          => {root vegetables}    0.007015760  0.4107143 0.01708185 3.768074    69
```

```
inspect(subset(groceryrules, subset=confidence > .6))
```

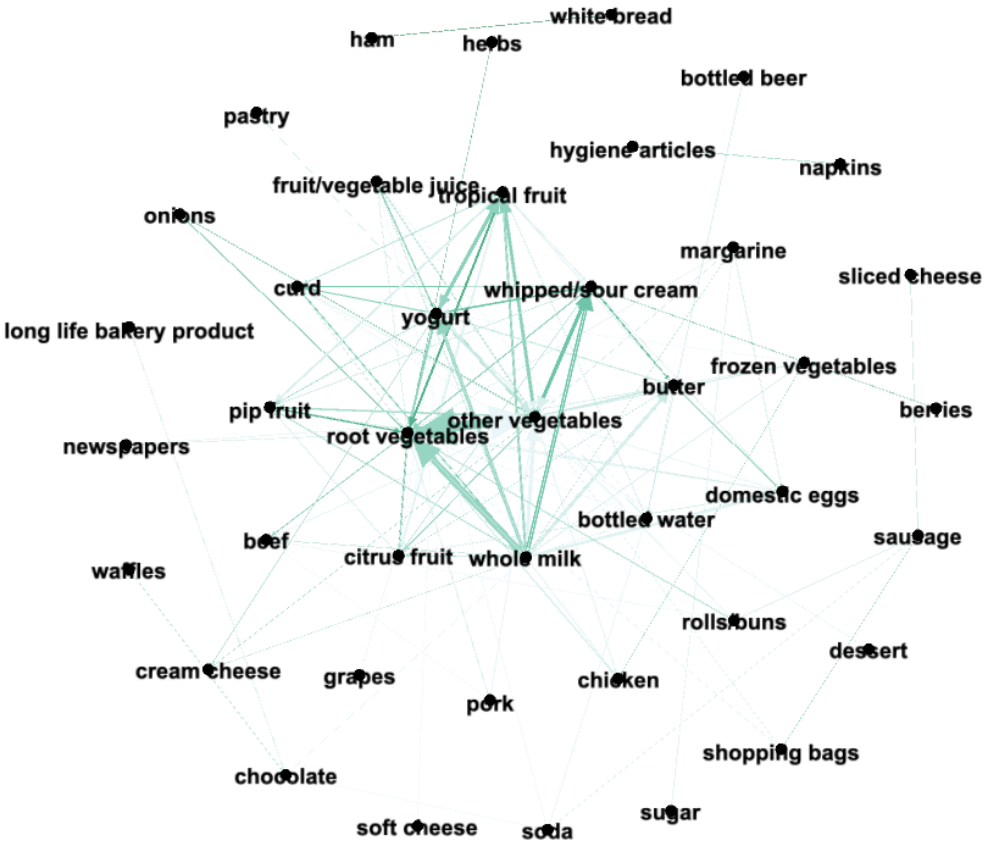
```
##      lhs                          rhs                    support confidence    coverage     lift coun
## [1]  {onions,
##       root vegetables}       => {other vegetables} 0.005693950  0.6021505 0.009456024 3.112008    5
## [2]  {curd,
##       tropical fruit}        => {whole milk}       0.006507372  0.6336634 0.010269446 2.479936    6
## [3]  {domestic eggs,
##       margarine}             => {whole milk}       0.005185562  0.6219512 0.008337570 2.434099    5
## [4]  {butter,
##       domestic eggs}         => {whole milk}       0.005998983  0.6210526 0.009659380 2.430582    5
## [5]  {butter,
##       whipped/sour cream}    => {whole milk}       0.006710727  0.6600000 0.010167768 2.583008    6
## [6]  {bottled water,
##       butter}                => {whole milk}       0.005388917  0.6022727 0.008947636 2.357084    5
## [7]  {butter,
##       tropical fruit}        => {whole milk}       0.006202339  0.6224490 0.009964413 2.436047    6
## [8]  {butter,
##       root vegetables}       => {whole milk}       0.008235892  0.6377953 0.012913066 2.496107    8
## [9]  {butter,
##       yogurt}                => {whole milk}       0.009354347  0.6388889 0.014641586 2.500387    9
## [10] {domestic eggs,
##       pip fruit}             => {whole milk}       0.005388917  0.6235294 0.008642603 2.440275    5
## [11] {domestic eggs,
##       tropical fruit}        => {whole milk}       0.006914082  0.6071429 0.011387900 2.376144    6
## [12] {pip fruit,
##       whipped/sour cream}    => {other vegetables} 0.005592272  0.6043956 0.009252669 3.123610    5
## [13] {pip fruit,
```

```
##       whipped/sour cream}     => {whole milk}       0.005998983  0.6483516 0.009252669 2.537421    59
## [14] {fruit/vegetable juice,
##       other vegetables,
##       yogurt}                 => {whole milk}       0.005083884  0.6172840 0.008235892 2.415833    50
## [15] {other vegetables,
##       root vegetables,
##       whipped/sour cream}     => {whole milk}       0.005185562  0.6071429 0.008540925 2.376144    51
## [16] {other vegetables,
##       pip fruit,
##       root vegetables}        => {whole milk}       0.005490595  0.6750000 0.008134215 2.641713    54
## [17] {pip fruit,
##       root vegetables,
##       whole milk}             => {other vegetables} 0.005490595  0.6136364 0.008947636 3.171368    54
## [18] {other vegetables,
##       pip fruit,
##       yogurt}                 => {whole milk}       0.005083884  0.6250000 0.008134215 2.446031    50
## [19] {citrus fruit,
##       root vegetables,
##       whole milk}             => {other vegetables} 0.005795628  0.6333333 0.009150991 3.273165    57
## [20] {root vegetables,
##       tropical fruit,
##       yogurt}                 => {whole milk}       0.005693950  0.7000000 0.008134215 2.739554    56
## [21] {other vegetables,
##       tropical fruit,
##       yogurt}                 => {whole milk}       0.007625826  0.6198347 0.012302999 2.425816    75
## [22] {other vegetables,
##       root vegetables,
##       yogurt}                 => {whole milk}       0.007829181  0.6062992 0.012913066 2.372842    77
```

Among a subset of high confidence rules, we see whole milk to be a common consequent, often with the antecedent including common kitchen staples.

Among the subset of high lift rules, we also see whole milk to be a common consequent, likely because whole milk is very commonly purchased.

```
grocery_graph = associations2igraph(subset(groceryrules, lift>2.5), associationsAsNodes = FALSE)
igraph::write_graph(grocery_graph, file='groceries.graphml', format = "graphml")
```



We created an association graph with the edges as the lift of the rules, based on a subset of the rules where lift > 2.5. The edges are ranked by the lift metric, and the nodes and edges are used using the Fructherman-Reingold layout algorithm. There are 41 nodes and 170 edges.

We see that there is a strong association between customers who buy whole milk, and those who buy whipped cream or sour cream. This could potentially be due to the fact that both are common dairy staples. We also saw this relationship among the high confidence rules, indicating a strong association between these dairy products. At the top of this visual, we see a reciprocal relationship between ham and white bread, indicating that customers frequently purchase these items together, likely due to sandwich making. We see an association between buying citrus fruit and pip fruit and buying tropical fruit, and also between buying citus and tropical fruits and buying pip fruits, showing a great interconnection between these items. A similar relationship exists between buying root vegetables and buying other vegetables. This is could indicate healthy eating patterns with customers buying a variety of fruits and vegetables, and could also be explained by the fact that root vegetables and other vegetables are commonly used together in dishes invololving vegetables.