Grocery Rule Mining

Eeshana Hamed, Anurag Arakala, Luis Villazon, Sardar Muhammad Ahmad Ali

2023-08-14

1. Convert txt format into data frame (saved as a csv)

2. Analyze dataset and determine support, confidence, and list

```
## 'data.frame':    43367 obs. of  2 variables:
##  $ customer: int  1 1 1 1 2 2 2 3 4 4 ...
##  $ purchase: chr  "citrus fruit" "semi-finished bread" "margarine" "ready soups" ...
```

```
##     customer      purchase
##  Min.   :   1   Length:43367
##  1st Qu.:2456   Class :character
##  Median :4828   Mode  :character
##  Mean   :4909
##  3rd Qu.:7380
##  Max.   :9835
```

Q3 Values (highest frequency):

```
##
##               whole milk         other vegetables               rolls/buns
##                     2513                     1903                     1809
##                     soda                   yogurt           bottled water
##                     1715                     1372                     1087
##          root vegetables           tropical fruit           shopping bags
##                     1072                     1032                      969
##                  sausage                  pastry            citrus fruit
##                      924                      875                      814
##             bottled beer              newspapers             canned beer
##                      792                      785                      764
##                pip fruit    fruit/vegetable juice       whipped/sour cream
##                      744                      711                      705
##              brown bread           domestic eggs              frankfurter
##                      638                      624                      580
##                margarine                  coffee                     pork
##                      576                      571                      567
##                   butter                    curd                     beef
##                      545                      524                      516
##                  napkins               chocolate       frozen vegetables
##                      515                      488                      473
##                  chicken             white bread           cream cheese
##                      422                      414                      390
```

```
##                   waffles              salty snack long life bakery product
##                       378                       372                       368
##                   dessert                     sugar                  UHT-milk
##                       365                       333                       329
##                   berries            hamburger meat          hygiene articles
##                       327                       327                       324
##                    onions
##                       305
```

Q1 Values (lowest frequency):

```
##
##                       tea            specialty fat       abrasive cleaner
##                        38                       36                       35
##               skin care              nuts/prunes       artif. sweetener
##                        35                       33                       32
##           canned fruit                     syrup              nut snack
##                        32                       32                       31
##         snack products                      fish        potato products
##                        30                       29                       28
##       bathroom cleaner                 cookware                     soap
##                        27                       27                       26
##      cooking chocolate           pudding powder                  tidbits
##                        25                       23                       23
##           cocoa drinks          organic sausage                 prosecco
##                        22                       22                       20
## flower soil/fertilizer             ready soups   specialty vegetables
##                        19                       18                       17
##       organic products              decalcifier                    honey
##                        16                       15                       15
##                  cream            frozen fruits              hair spray
##                        13                       12                       11
##        rubbing alcohol                  liqueur        make up remover
##                        10                        9                        8
##         salad dressing                   whisky         toilet cleaner
##                         8                        8                        7
##         baby cosmetics           frozen chicken                    bags
##                         6                        6                        4
##        kitchen utensil  preservation products              baby food
##                         4                        2                        1
##   sound storage medium
##                         1
```

Highest Value in each Quartile:

```
## [1] 38
```

```
## [1] 103
```

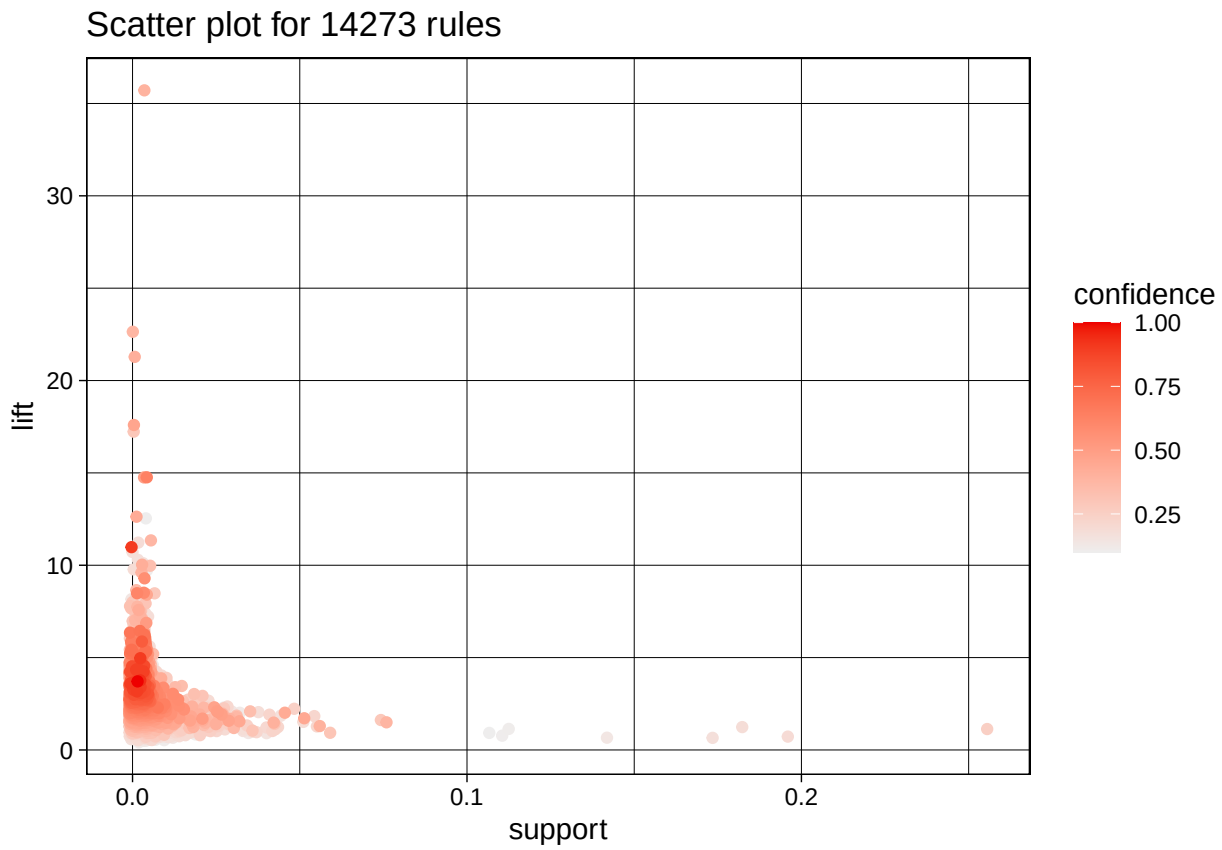
```
## [1] 2513
```

After looking at the frequency of each item and dividing them into groups of 25%, 50%, and 75% of data, looking at the highest and average values in each quartile gives me an idea of the range for support I could

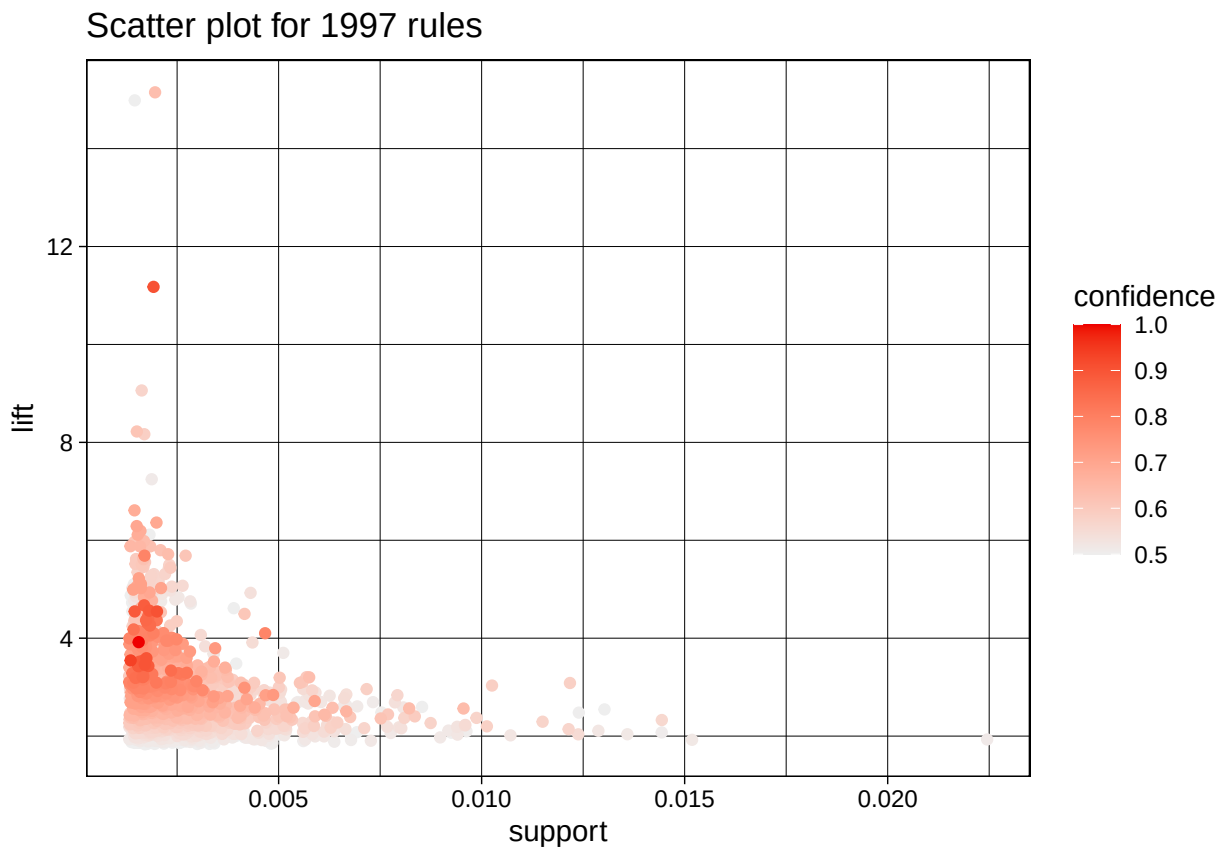
use to include most of the data set in determining rules or narrow it down to use only very high frequency items in the set. Looking at a support range of [0.0008(38/43367), 0.05(2513/43367)] gives a range of support that would include the most amount of data to the least. Looking at a median value of 103 in Q2, to analyze the items that occur a good amount, we started with a support of .002(103/43367). However, we were not getting high enough lift values (<8), so we decided to decrease it to .0015 and start with a confidence of .1.

3. Apply rule mining model and plot data points

Beginning with a support of .0015 and a confidence of .1, we get the following plot that gives an overall idea of confidence, lift, and support values:



Seeing the concentration of high confidence values around a lift of 4, we can adjust the plot to get a better idea of the values:



Looking at this scatter plot, we want to focus our analysis on high lift and confidence values:

- Our highest lift points (above 7) tend to have low to medium confidence and lower support -> this means these items don't appear in grocery baskets very often, but when it does, the presence of one item influences the occurrence of the other item. However, a lower confidence could question the accuracy of this association.

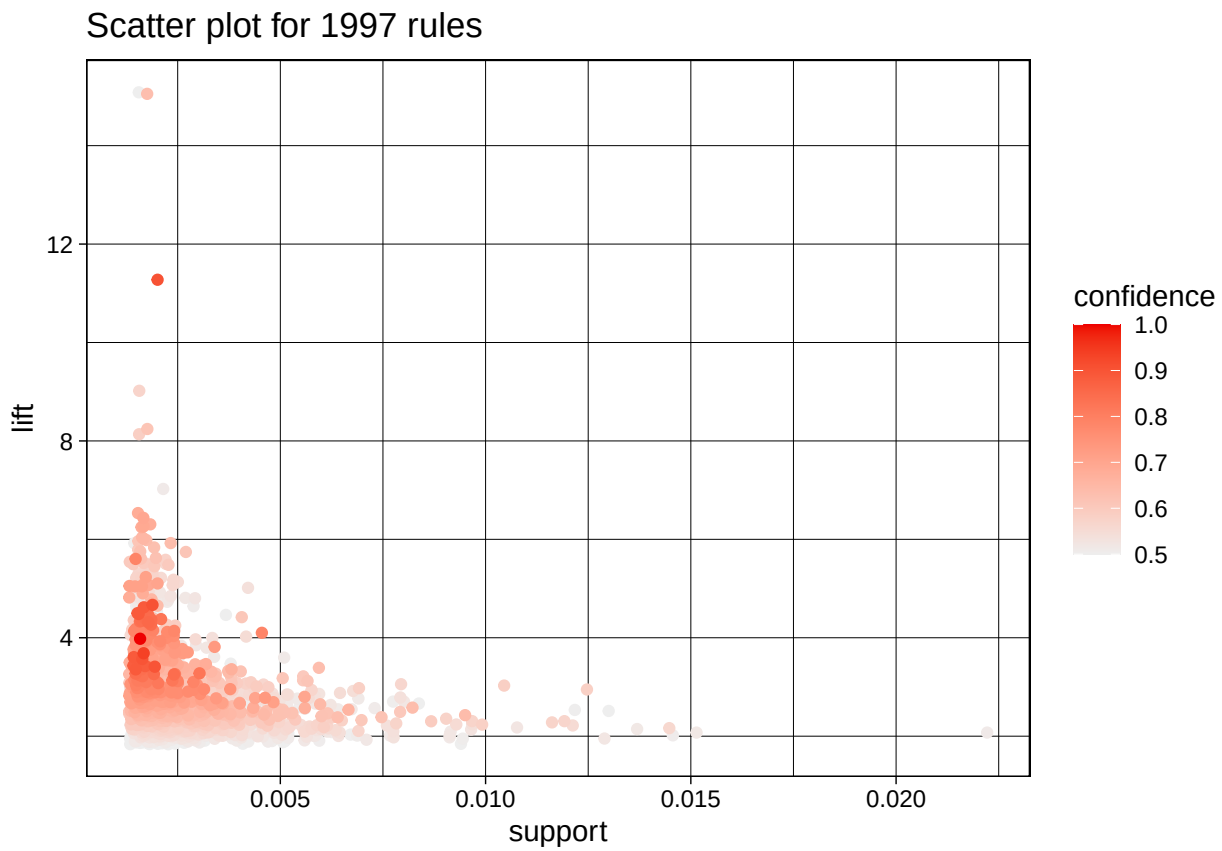
Here are some items that show this relation:

```
##      lhs                          rhs                        support confidence    coverage     lift
## [1]  {liquor}                  => {red/blush wine}        0.002135231  0.1926606 0.011082867 10.02548
## [2]  {red/blush wine}          => {liquor}                0.002135231  0.1111111 0.019217082 10.02548
## [3]  {Instant food products}   => {hamburger meat}        0.003050330  0.3797468 0.008032537 11.42144
## [4]  {liquor,
##       red/blush wine}          => {bottled beer}          0.001931876  0.9047619 0.002135231 11.23527
## [5]  {bottled beer,
##       liquor}                  => {red/blush wine}        0.001931876  0.4130435 0.004677173 21.49356
## [6]  {bottled beer,
##       red/blush wine}          => {liquor}                0.001931876  0.3958333 0.004880529 35.71579
## [7]  {Instant food products,
##       whole milk}              => {hamburger meat}        0.001525165  0.5000000 0.003050330 15.03823
## [8]  {hamburger meat,
##       whole milk}              => {Instant food products} 0.001525165  0.1034483 0.014743264 12.87866
## [9]  {ham,
##       processed cheese}        => {white bread}           0.001931876  0.6333333 0.003050330 15.04549
## [10] {processed cheese,
```

```
##       white bread}           => {ham}                  0.001931876  0.4634146 0.004168785 17.80345
## [11] {ham,
##       white bread}           => {processed cheese}     0.001931876  0.3800000 0.005083884 22.92822
## [12] {tropical fruit,
##       white bread}           => {processed cheese}     0.001525165  0.1744186 0.008744281 10.52397
## [13] {soda,
##       white bread}           => {processed cheese}     0.001728521  0.1683168 0.010269446 10.15580
## [14] {flour,
##       margarine}             => {sugar}                0.001626843  0.4324324 0.003762074 12.77169
## [15] {margarine,
##       sugar}                 => {flour}                0.001626843  0.2962963 0.005490595 17.04137
## [16] {sugar,
##       whole milk}            => {flour}                0.002846975  0.1891892 0.015048297 10.88114
## [17] {fruit/vegetable juice,
##       ham}                   => {white bread}          0.001626843  0.4210526 0.003863752 10.00254
## [18] {root vegetables,
##       whipped/sour cream,
##       whole milk}            => {flour}                0.001728521  0.1827957 0.009456024 10.51343
```

In terms of high confidence, we can see that the purchase of a lot of alcoholic drinks indicates the likelihood of buying another type of alcoholic drink. For example, {liquor, red/blush wine} and {bottled beer} has a confidence of 90%, indicating many people who purchased liquor and wine also purchased bottled beer. When we look at someone who purchased {bottled beer, red/blush wine} , they also had {liquor} in their cart 40% of the time. This is less than the previous association, however it has a much higher lift, indicating that when some buys {bottled beer, red/blush wine} they are much more likely to also purchase {liquor} than some who has {liquor, red/blush wine} and is likely to buy {bottled beer}. Furthermore, we also see high lift, medium confidence associations between instant products and hamburger meat, ham, processed cheese and white bread (for sandwiches?), and baking ingredients (sugar, flour, margarine). These points indicate interesting facts about eating habits of consumers.

Returning to this plot, we can now look at some more higher confidence but lower lift points:



Here are some items that show this relation:

```
##      lhs                      rhs                    support confidence    coverage      lift coun
## [1]  {liquor,
##       red/blush wine}      => {bottled beer}       0.001931876  0.9047619 0.002135231 11.235269    1
## [2]  {rice,
##       yogurt}              => {root vegetables}    0.001626843  0.6956522 0.002338587  6.382219    1
## [3]  {ham,
##       processed cheese}    => {white bread}        0.001931876  0.6333333 0.003050330 15.045491    1
## [4]  {butter,
##       hard cheese}         => {whipped/sour cream} 0.002033554  0.5128205 0.003965430  7.154028    2
## [5]  {other vegetables,
##       rice,
##       whole milk}          => {root vegetables}    0.001830198  0.6923077 0.002643620  6.351536    1
## [6]  {flour,
##       whipped/sour cream,
##       whole milk}          => {root vegetables}    0.001728521  0.6800000 0.002541942  6.238619    1
## [7]  {flour,
##       root vegetables,
##       whole milk}          => {whipped/sour cream} 0.001728521  0.5862069 0.002948653  8.177794    1
## [8]  {frozen meals,
##       pip fruit,
##       whole milk}          => {tropical fruit}     0.001525165  0.6818182 0.002236909  6.497754    1
## [9]  {ham,
##       other vegetables,
##       tropical fruit}      => {pip fruit}          0.001626843  0.6153846 0.002643620  8.134822    1
## [10] {oil,
```

```
##       other vegetables,
##       tropical fruit}      => {root vegetables}    0.001728521  0.6800000 0.002541942  6.238619    1
## [11] {citrus fruit,
##       cream cheese ,
##       whole milk}          => {domestic eggs}      0.001626843  0.5714286 0.002846975  9.006410    1
```

Here we can see that these items occur more often in baskets together, however buying one does not always indicate that they will buy the other. These items are more indicative of purchases that would be deemed basic household items typically bought together when consumers go for their larger monthly purchases of common necessities (milk, rice, yogurt, vegetables, etc.)