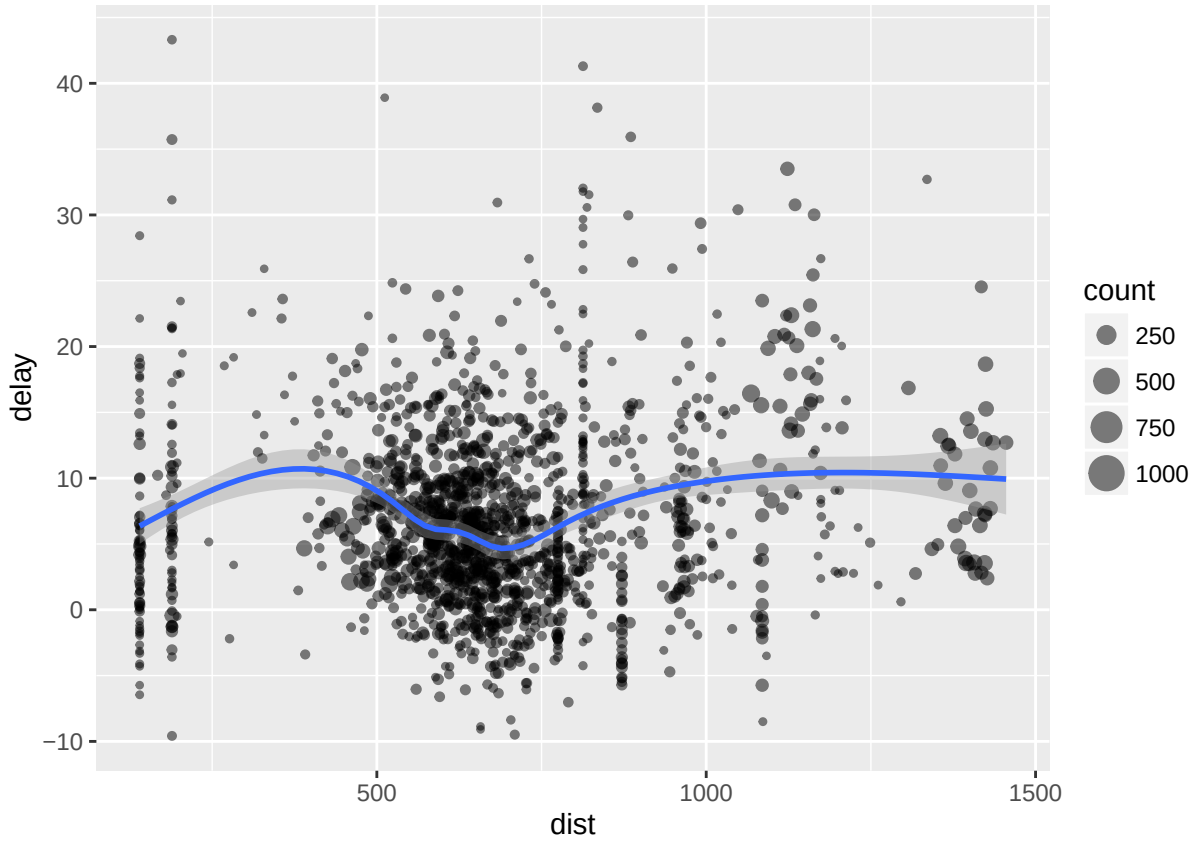
STA 380, Part 2: Exercises 2

Flights at ABIA

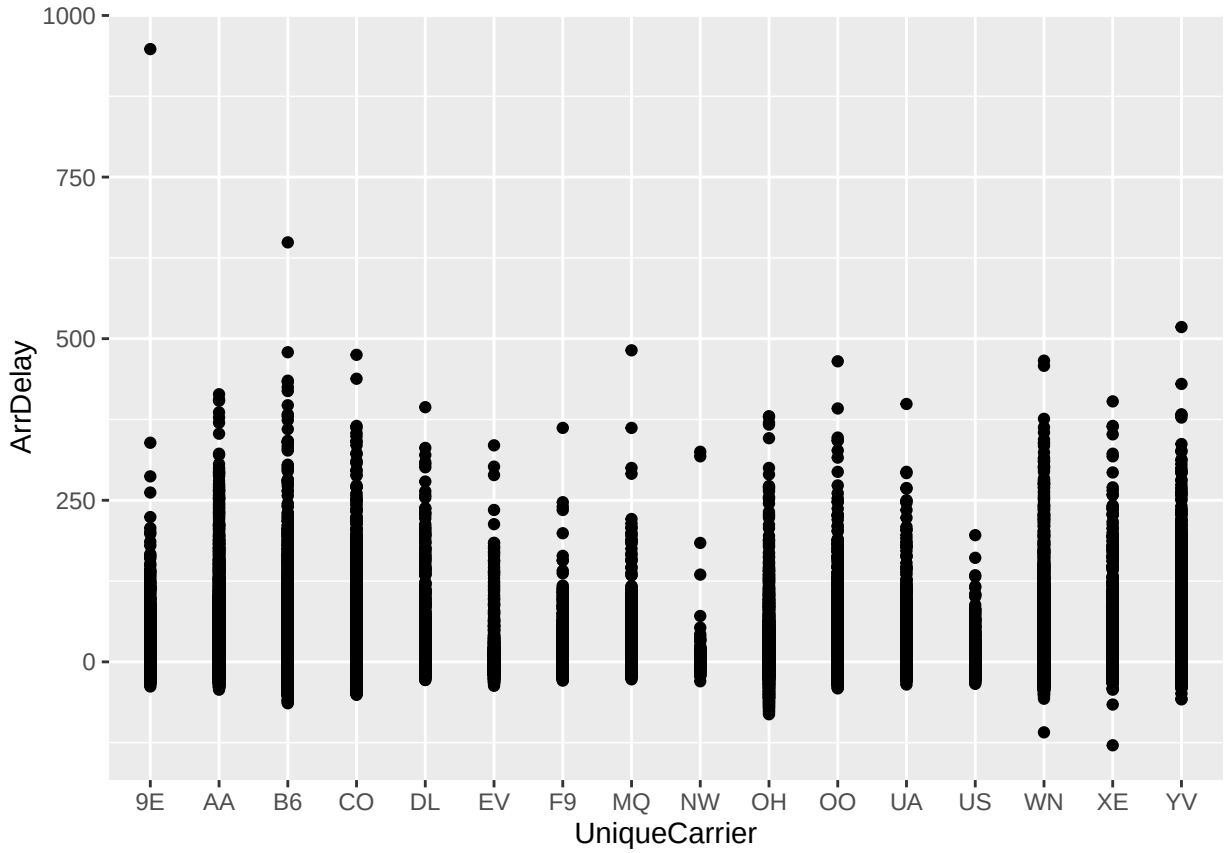
Your task is to create a figure, or set of related figures, that tell an interesting story about flights into and out of Austin. You can annotate the figure and briefly describe it, but strive to make it as stand-alone as possible. It shouldn't need many, many paragraphs to convey its meaning. Rather, the figure should speak for itself as far as possible.

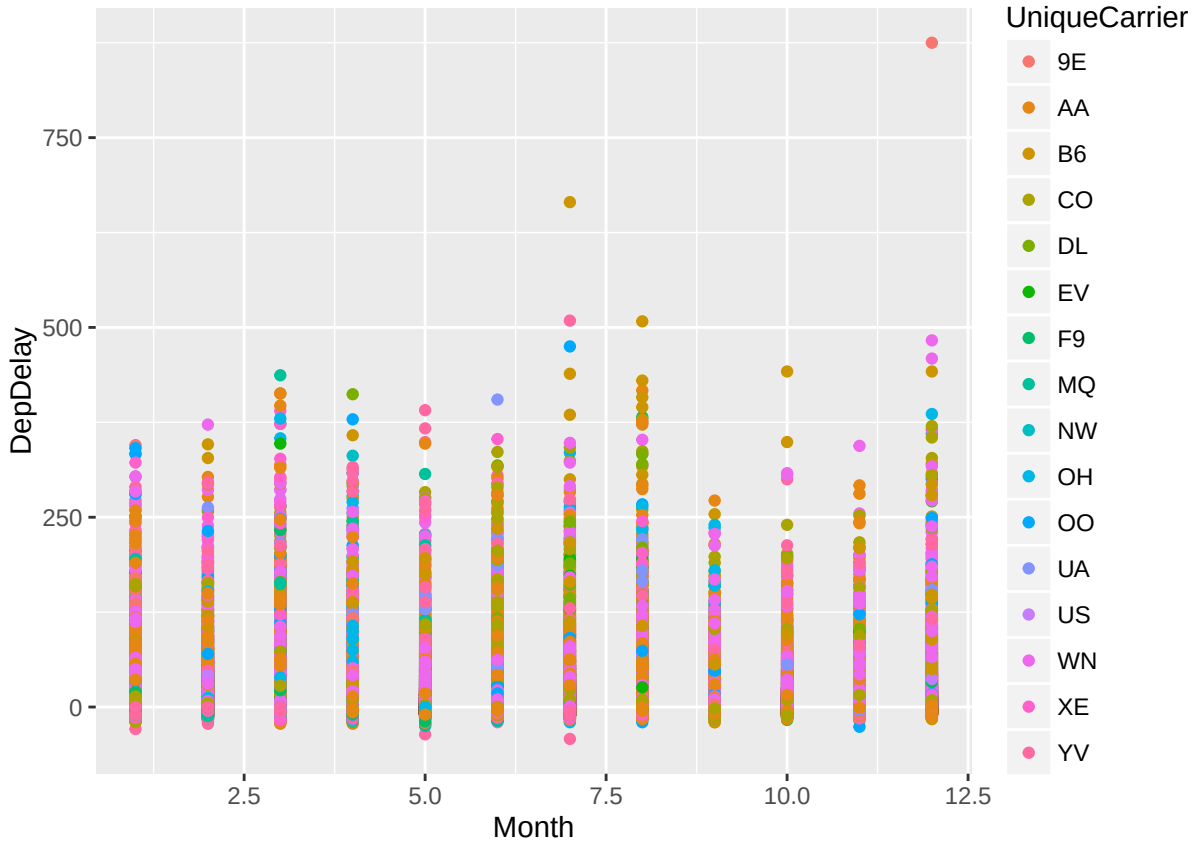
Flights in and out of Austin in 2008!

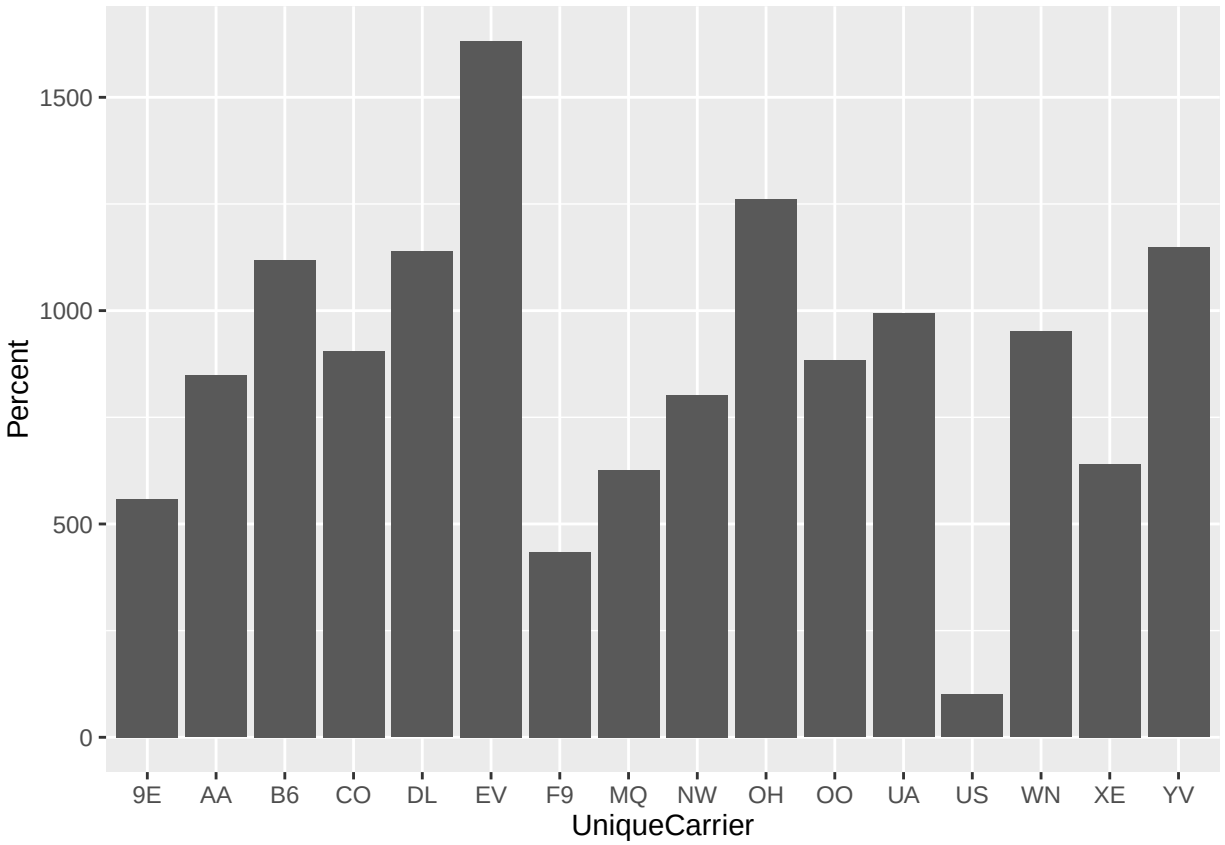


We can see that the planes that fly from 550 to 770 miles are more prone to delays.This can be explained,since the median distance for a plane is 775 miles and there are less flights involving shorter/longer routes. Overall however the average delay is slightly related to the average distance flown by a plane.

In case some of us would like to see the arrival delays per carrier can see the following plots:

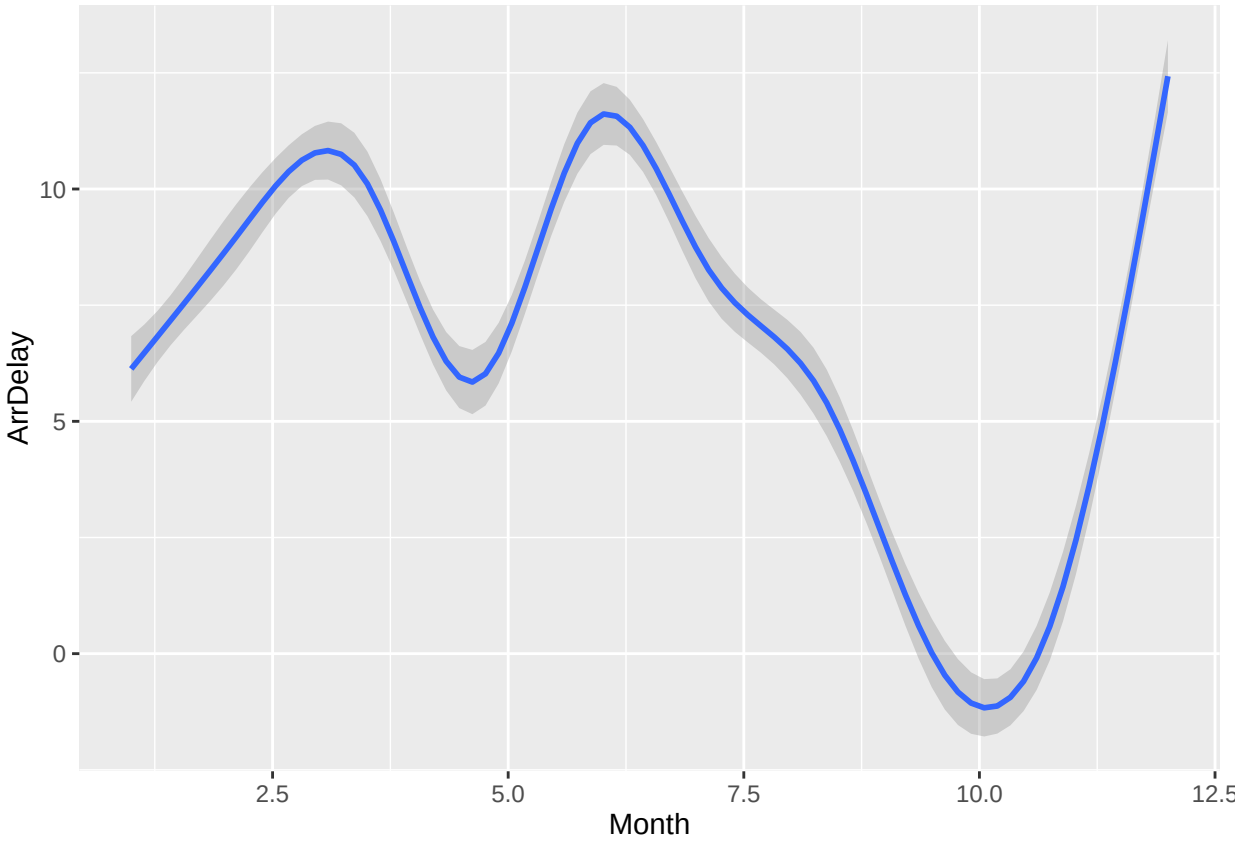

1000
750
500
250
0
ArrDelay
9E
AA
B6
CO
DL
EV
F9
MQ
NW
OH
OO
UA
US
WN
XE
YV
UniqueCarrier

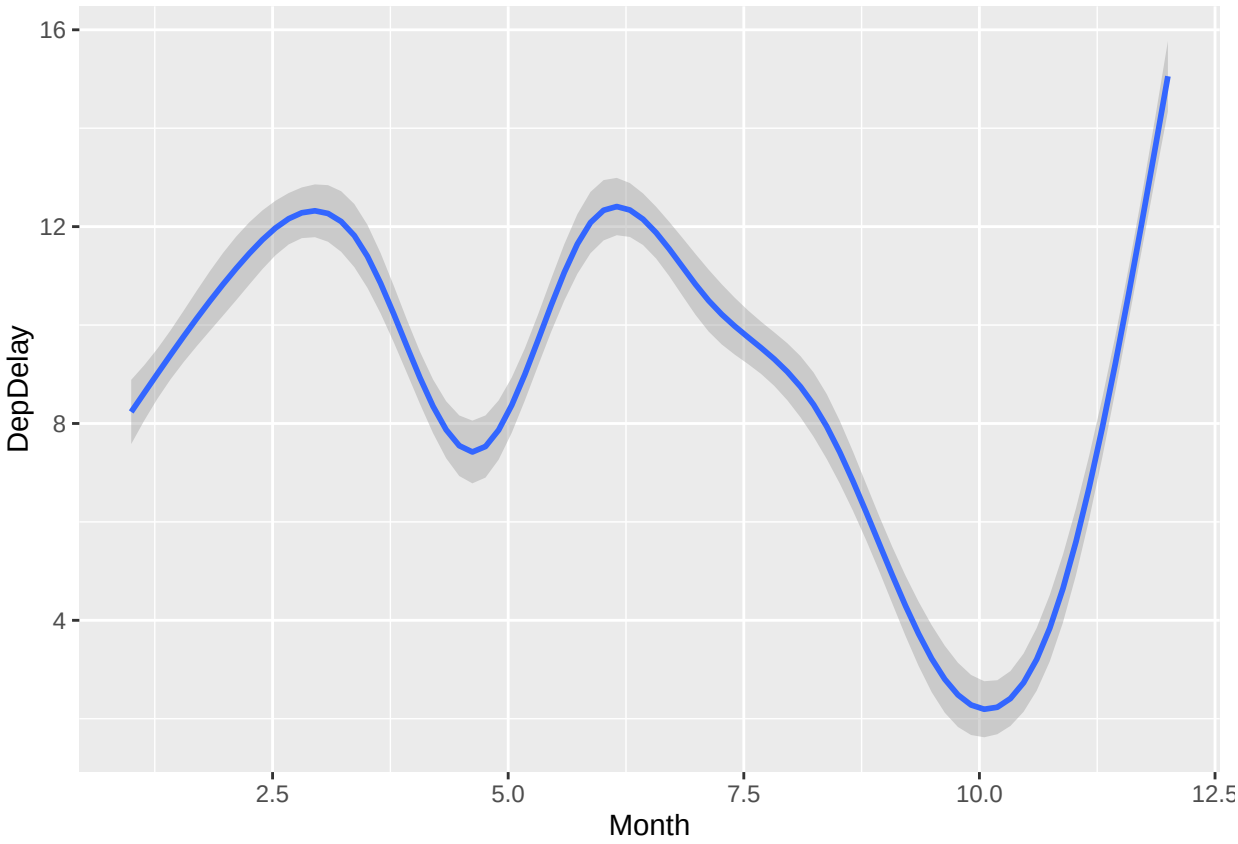

750
500
250
0
DepDelay
2.5
5.0
7.5
10.0
12.5
Month
UniqueCarrier
9E
AA
B6
CO
DL
EV
F9
MQ
NW
OH
OO
UA
US
WN
XE
YV



We can see that the "safest" option delay-wise is to fly with US airlines since they have the smallest delay percentage.

And if we wish to plan ahead our trip based on the delays per month in 2008 under the assumption that there is a pattern, we can consult the graphs below.





Based on these I would choose to fly sometime in October! We can see that during the Christmas holidays and the summer period where the flight "load" is increased, the most delays occur.

Author attribution

Revisit the Reuters C50 corpus that we explored in class. Your task is to build two separate models (using any combination of tools you see fit) for predicting the author of an article on the basis of that article's textual content. Describe clearly what models you are using, how you constructed features, and so forth. (Yes, this is a supervised learning task, but it potentially draws on a lot of what you know about unsupervised learning!)

In the C50train directory, you have ~50 articles from each of 50 different authors (one author per directory). Use this training data (and this data alone) to build the two models. Then apply your model to the articles by the same authors in the C50test directory, which is about the same size as the training set. How well do your models do at predicting the author identities in this out-of-sample setting? Are there any sets of authors whose articles seem difficult to distinguish from one another? Which model do you prefer?

First we read half of our data which consist of the training set. The we modify the R.script ,that we studied in class, with a for loop so that it then contains all 50 authors. Afterwards we strip our corpus of empty spaces and numbers, we make everything lowercase, we remove the stopwords and we print here the term document matrix:

```
## <<DocumentTermMatrix (documents: 2500, terms: 31423)>>
## Non-/sparse entries: 425955/78131545
## Sparsity           : 99%
## Maximal term length: 36
## Weighting          : term frequency (tf)
```

Then we will use the Naïve Bayes algorithm to the training set for all the authors while we put the 2 directories ensemble into a single corpus (we load the test data). We repeat again the processing procedure we did for the training set before and we prind once more the term document matrix:

```
## <<DocumentTermMatrix (documents: 2500, terms: 32264)>>
## Non-/sparse entries: 432766/80227234
## Sparsity           : 99%
## Maximal term length: 45
## Weighting          : term frequency (tf)
```

We modify to keep only those words that we used in the training set and we print the DTM:

```
## <<DocumentTermMatrix (documents: 2500, terms: 1389)>>
## Non-/sparse entries: 246565/3225935
## Sparsity           : 93%
## Maximal term length: 18
## Weighting          : term frequency (tf)
```

Next step is to calculate the log probabilities (since the probabilities aone are very small and the computer cannot make comparisons of very small things),

and find the document which corresponds to the maximum log-probability which in reality is the author.

Finaly, our predictions can be summed into a (very big) confusion matrix and we see our predictions accuracy:

```
## Confusion Matrix and Statistics
##
##            Reference
## Prediction  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21
##          1 40  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##          2  0 25  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0
##          3  0  0 21  0  2  0  0  0  4  0  0  0  0  0  0  0  5  7  0  2  0
##          4  0  0  0 10  0  0  0  0  0  0  0  0  0  0  6  0  0  0  0  0  0
##          5  0  0  0  0 27  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##          6  1  0  0  0  0 45  0 11  0  1  0  0  0  0  0  0  0  0  0  0  0
##          7  2  0  0  0  0  0 11  0  0  0  0  0  5  0  0  0  0  0  0  0  0
##          8  0  0  0  0  0  0  0  7  0  0  0  0  0  0  0  0  0  0  0  0  0
##          9  0  0  0  0  0  0  0  0 18  0  0  0  0  0  0  0  0  1  0  1  0
##          10 0  0  0  0  0  1  0  0  0 26  0  0  0  0  0  0  0  0  0  0  0
##          11 0  0  0  0  0  0  0  0  0  0 49  0  0  0  0  0  0  0  0  0  0
##          12 0  0  0  2  0  0  0  0  0  0  0 40  0  0  1  0  0  0  0  0  0
##          13 0  0  0  0  2  0 36  0  0  0  0  0 16  0  0  0  0  0  0  0  0
##          14 0  8  0  0  0  0  0  0  0  0  0  0  1 25  0  0  0  0  9  0  0
##          15 0  0  0 14  0  0  0  1  0  0  0  2 11  0 19  0  0  0  0  0  0
##          16 0  0  0  0  0  0  0  0  0  0  0  1  0  0  0 50  0  0  0  0  0
##          17 0  0  0  0  0  0  0  0  3  0  0  0  0  0  0  0 33  0  0  2  0
##          18 1  0 27  0  1  0  0  1  0  0  0  0  0  1  0  0  1 37  0  0  0
##          19 0 16  0  0  0  0  0  0  0  0  0  0  3 21  0  0  0  0 31  0  0
##          20 0  0  1  0  0  0  0  0  3  0  0  0  0  0  0  0  3  0  1 36  0
##          21 0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  4  0 47
##          22 0  1  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  1  0
##          23 0  0  0  0  0  1  0  3  0  1  0  0  0  0  0  0  0  0  0  0  0
##          24 0  0  0  0  5  0  0  0  0  0  0  0  2  0  0  0  0  0  0  0  0
##          25 0  0  0  0  0  0  0  0  2  0  0  0  0  1  0  0  0  0  0  1  0
```

```
##          26  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0
##          27  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##          28  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##          29  0  0  0  0  1  0  0  0  0  0  0  0  1  0  0  0  0  0  0  1  0
##          30  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##          31  0  0  0  0  3  0  1  0  1  0  0  0  3  0  0  0  0  0  0  0  0
##          32  0  0  0  0  0  1  0  2  0  1  0  0  0  0  0  0  0  0  0  0  0
##          33  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  2
##          34  0  0  0  0  2  0  1  0  0  0  0  0  0  1  0  0  0  0  0  0  0
##          35  0  0  0  2  0  0  0  0  0  0  0  0  0  0  3  0  0  0  0  0  0
##          36  0  0  0  0  0  0  0  0  0 10  0  0  0  0  0  0  0  0  0  0  0
##          37  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##          38  0  0  0  3  0  0  0  0  0  0  0  0  2  0  0  0  0  0  1  0  0
##          39  0  0  0  0  0  0  0  0  7  0  0  0  0  0  0  0  0  0  1  0  0
##          40  0  0  0  0  0  1  0  1  0  1  0  0  0  0  0  0  0  0  0  0  0
##          41  0  0  0  0  0  0  0  0  0  1  0  0  1  0  0  0  0  0  0  0  0
##          42  6  0  1  0  1  0  0  0  0  4  0  0  2  0  0  0  0  0  0  0  0
##          43  0  0  0  0  4  0  1  0  0  0  1  3  0  0  0  0  0  0  0  0  0
##          44  0  0  0  9  0  0  0  0  0  0  0  0  1  0  1 18  0  0  0  1  0  1
##          45  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  4  0  2  0
##          46  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##          47  0  0  0  0  0  0  0  0  0  0  3  0  0  0  0  0  0  0  0  0  0  0
##          48  0  0  0  0  2  0  0  0  6  0  0  0  0  0  0  0  0  8  1  0  4  0
##          49  0  0  0  0  0  1  0 23  6  0  0  0  0  0  0  0  0  0  0  1  0  0
##          50  0  0  0  9  0  0  0  0  0  0  0  0  3  3  0  3  0  0  0  0  0  0
##            Reference
## Prediction 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42
##          1   0  0  0  2  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  2  1
##          2   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##          3   0  0  0  0  0  0  0  0  0  0  0  0  2  0  0  0  0  0  0  0  0
##          4   0  0  0  0  0  0  0  0  0  0  1  0  0  1  0  0  0  0  0  0  0
##          5   0  0  5  0  0  0  0  0  0  5  0  0  0  0  0  0  0  0  0  0  0
##          6   0  5  0  0  0  5  0  0  0  0  3  0  1  0  1  6  0  0  2  0  0
##          7   0  0  0  0  0  7  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##          8   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##          9   4  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##          10  0  1  0  0 12  1  0  0  0  0  1  0  0  0  3  2  0  0  0  1  5
##          11  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##          12  0  0  0  0  0  0  1  0  0  0  0  0  0  5  0  0  0  0  0  0  0
##          13  0  0  0  0  0  1  0  0  0  0  0  2  0  0  0  0  0  0  0  0  0
##          14  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##          15  0  0  2  0  0  0  4  0  0  1  0  3  0 10  0  0  1  0  0  0  0
##          16  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0
##          17  1  0  0  2  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##          18  0  0  0  0  0  0  0  0  1  1  0  0  4  0  0  0  0  0  0  0  0
##          19  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##          20  1  0  0  2  0  0  0  0  5  0  0  0  0  0  0  1  0  1  2  0  0
##          21  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##          22 38  1  1  7  0  0  0  0  1  0  0  0  2  0  0  0  0  0  0  2  0
##          23  1 29  0  0  1  1  0  0  2  0  2  0  0  0  0  7  0  0  2  0  2
##          24  0  0 28  0  0  0  0  0  0  7  0  0  0  0  0  0  0  0  1  0  0
##          25  3  0  2 34  0  0  0  0  1  4  0  0  0  0  2  0  0  0  0  0  0
##          26  0  0  0  0 30  0  0  0  0  0  2  0  0  0  0  0  0  0  0  0  1
##          27  0  0  0  0  0 31  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
```

```
##         28  0  0  0  0  0  0 40  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##         29  0  0  0  0  0  0  0 50  4  1  0  2  0  1  0  0  0  0  0  0  0
##         30  0  0  0  0  0  0  0  0 32  0  0  0  0  0  0  0  0 13  1  0  0
##         31  0  0  5  0  0  0  0  0  0 21  0  0  1  0  0  0  0  0  0  0  0
##         32  0  0  0  0  0  0  0  0  0  0 23  0  0  0  0  0  0  0  1  0  2
##         33  0  0  0  0  0  0  0  0  0  0  0 42  0  0  0  0  0  0  0  0  0
##         34  0  1  0  0  0  1  0  0  0  0  0  0 37  0  0  1  0  0  1  1  1
##         35  0  0  0  0  0  0  2  0  0  1  0  0  0 15  0  0  0  0  0  0  0
##         36  0  0  1  0  1  0  0  0  0  0  0  0  0  0 42  3  0  0  0  0  0
##         37  0  9  0  0  1  0  0  0  0  0  1  0  1  0  1 28  0  0  0  0  0
##         38  0  0  0  0  0  0  0  0  0  0  0  0  0  2  0  0 36  0  0  0  0
##         39  0  0  0  0  0  0  0  0  1  1  1  0  0  1  0  0  0 35  0  0  0
##         40  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 40  0  0
##         41  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 39  0
##         42  0  3  0  0  1  2  0  0  1  0  6  0  1  0  1  0  0  0  0  2 30
##         43  0  0  3  1  0  1  0  0  0  5  1  0  0  0  0  0  2  0  0  0  0
##         44  0  0  0  0  0  0  0  0  0  1  0  0  0  5  0  0  1  0  0  3  0
##         45  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0
##         46  1  0  0  0  0  0  1  0  0  0  0  1  0  2  0  0  9  0  0  0  0
##         47  0  1  0  0  4  0  0  0  2  0  8  0  1  0  0  1  0  0  0  0  8
##         48  0  0  1  1  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0
##         49  0  0  0  0  0  0  1  0  0  2  1  0  0  0  0  0  0  0  0  0  0
##         50  0  0  2  0  0  0  1  0  0  0  0  0  0  8  0  0  0  0  0  0  0
##           Reference
## Prediction 43 44 45 46 47 48 49 50
##         1   0  0  0  0  3  0  0  0
##         2   0  0  0  0  0  0  0  0
##         3   0  0 16  0  0  0  0  0
##         4   0  1  0  0  0  0  0  4
##         5   0  0  0  1  0  0  0  0
##         6   0  0  0  0  0  0  5  0
##         7   0  0  0  0  0  0  0  0
##         8   0  0  0  0  0  0 13  0
##         9   0  0  0  0  0  1  0  0
##         10  0  0  0  0  9  0  0  0
##         11  0  0  1  0  0  0  0  2
##         12  0  1  0  1  0  0  0  2
##         13  0  0  0  0  0  0  0  0
##         14  0  0  0  0  0  0  0  0
##         15  0 25  0  0  0  0  0  8
##         16  1  0  0  0  0  0  1  0
##         17  0  0  0  0  0  0  0  0
##         18  0  0  4  0  0  1  0  0
##         19  0  0  0  0  0  0  0  0
##         20  0  0  0  0  0  2  0  0
##         21  0  0  0  0  0  0  6  0
##         22  0  0  0  0  0  2  0  0
##         23  0  0  0  0  4  0  3  0
##         24  0  0  0  0  0  0  0  0
##         25  0  0  0  0  0  0  0  0
##         26  0  0  0  0  1  0  0  0
##         27  0  0  0  0  0  0  0  0
##         28  0  1  0  0  0  0  0  1
##         29  0  0  0  0  0  0  0  0
```

```
##         30  0  0  0  0  0  0  0  0
##         31  0  0  0  0  0  1  0  0
##         32  0  0  0  0  2  0  0  0
##         33  0  0  0  0  0  0  0  0
##         34  0  0  0  0  0  1  1  0
##         35  1  3  0  0  0  0  0  2
##         36  0  0  0  0  0  0  0  0
##         37  0  0  0  0  0  0  0  0
##         38 10  4  0 15  0  0  0  2
##         39  0  0  0  0  0  0  0  0
##         40  0  0  0  0  0  0  0  0
##         41  0  0  0  0  0  0  0  0
##         42  0  0  0  0  6  0  0  0
##         43 28  0  0  8  0  0  0  1
##         44  1 13  0  1  0  0  0  6
##         45  0  0 29  0  0  4  0  0
##         46  9  1  0 23  0  0  0  3
##         47  0  0  0  0 25  0  0  0
##         48  0  0  0  0  0 38  0  0
##         49  0  0  0  0  0  0 21  0
##         50  0  1  0  1  0  0  0 19
##
## Overall Statistics
##
##                Accuracy : 0.6036
##                  95% CI : (0.5841, 0.6228)
##     No Information Rate : 0.02
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5955
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: 1 Class: 2 Class: 3 Class: 4 Class: 5 Class: 6
## Sensitivity            0.8000   0.5000   0.4200   0.2000   0.5400   0.9000
## Specificity            0.9967   0.9996   0.9845   0.9947   0.9955   0.9833
## Pos Pred Value         0.8333   0.9615   0.3559   0.4348   0.7105   0.5233
## Neg Pred Value         0.9959   0.9899   0.9881   0.9839   0.9907   0.9979
## Prevalence             0.0200   0.0200   0.0200   0.0200   0.0200   0.0200
## Detection Rate         0.0160   0.0100   0.0084   0.0040   0.0108   0.0180
## Detection Prevalence   0.0192   0.0104   0.0236   0.0092   0.0152   0.0344
## Balanced Accuracy      0.8984   0.7498   0.7022   0.5973   0.7678   0.9416
##                      Class: 7 Class: 8 Class: 9 Class: 10 Class: 11
## Sensitivity            0.2200   0.1400   0.3600    0.5200    0.9800
## Specificity            0.9943   0.9947   0.9971    0.9853    0.9988
## Pos Pred Value         0.4400   0.3500   0.7200    0.4194    0.9423
## Neg Pred Value         0.9842   0.9827   0.9871    0.9902    0.9996
## Prevalence             0.0200   0.0200   0.0200    0.0200    0.0200
## Detection Rate         0.0044   0.0028   0.0072    0.0104    0.0196
## Detection Prevalence   0.0100   0.0080   0.0100    0.0248    0.0208
## Balanced Accuracy      0.6071   0.5673   0.6786    0.7527    0.9894
##                      Class: 12 Class: 13 Class: 14 Class: 15 Class: 16
## Sensitivity             0.8000    0.3200    0.5000    0.3800    1.0000
```

```
## Specificity             0.9947    0.9833    0.9927    0.9665    0.9984
## Pos Pred Value          0.7547    0.2807    0.5814    0.1881    0.9259
## Neg Pred Value          0.9959    0.9861    0.9898    0.9871    1.0000
## Prevalence              0.0200    0.0200    0.0200    0.0200    0.0200
## Detection Rate          0.0160    0.0064    0.0100    0.0076    0.0200
## Detection Prevalence    0.0212    0.0228    0.0172    0.0404    0.0216
## Balanced Accuracy       0.8973    0.6516    0.7463    0.6733    0.9992
##                      Class: 17 Class: 18 Class: 19 Class: 20 Class: 21
## Sensitivity             0.6600    0.7400    0.6200    0.7200    0.9400
## Specificity             0.9967    0.9824    0.9837    0.9910    0.9955
## Pos Pred Value          0.8049    0.4625    0.4366    0.6207    0.8103
## Neg Pred Value          0.9931    0.9946    0.9922    0.9943    0.9988
## Prevalence              0.0200    0.0200    0.0200    0.0200    0.0200
## Detection Rate          0.0132    0.0148    0.0124    0.0144    0.0188
## Detection Prevalence    0.0164    0.0320    0.0284    0.0232    0.0232
## Balanced Accuracy       0.8284    0.8612    0.8018    0.8555    0.9678
##                      Class: 22 Class: 23 Class: 24 Class: 25 Class: 26
## Sensitivity             0.7600    0.5800    0.5600    0.6800    0.6000
## Specificity             0.9922    0.9878    0.9939    0.9935    0.9980
## Pos Pred Value          0.6667    0.4915    0.6512    0.6800    0.8571
## Neg Pred Value          0.9951    0.9914    0.9910    0.9935    0.9919
## Prevalence              0.0200    0.0200    0.0200    0.0200    0.0200
## Detection Rate          0.0152    0.0116    0.0112    0.0136    0.0120
## Detection Prevalence    0.0228    0.0236    0.0172    0.0200    0.0140
## Balanced Accuracy       0.8761    0.7839    0.7769    0.8367    0.7990
##                      Class: 27 Class: 28 Class: 29 Class: 30 Class: 31
## Sensitivity             0.6200    0.8000    1.0000    0.6400    0.4200
## Specificity             1.0000    0.9992    0.9955    0.9943    0.9939
## Pos Pred Value          1.0000    0.9524    0.8197    0.6957    0.5833
## Neg Pred Value          0.9923    0.9959    1.0000    0.9927    0.9882
## Prevalence              0.0200    0.0200    0.0200    0.0200    0.0200
## Detection Rate          0.0124    0.0160    0.0200    0.0128    0.0084
## Detection Prevalence    0.0124    0.0168    0.0244    0.0184    0.0144
## Balanced Accuracy       0.8100    0.8996    0.9978    0.8171    0.7069
##                      Class: 32 Class: 33 Class: 34 Class: 35 Class: 36
## Sensitivity             0.4600    0.8400    0.7400    0.3000    0.8400
## Specificity             0.9963    0.9988    0.9951    0.9943    0.9939
## Pos Pred Value          0.7187    0.9333    0.7551    0.5172    0.7368
## Neg Pred Value          0.9891    0.9967    0.9947    0.9858    0.9967
## Prevalence              0.0200    0.0200    0.0200    0.0200    0.0200
## Detection Rate          0.0092    0.0168    0.0148    0.0060    0.0168
## Detection Prevalence    0.0128    0.0180    0.0196    0.0116    0.0228
## Balanced Accuracy       0.7282    0.9194    0.8676    0.6471    0.9169
##                      Class: 37 Class: 38 Class: 39 Class: 40 Class: 41
## Sensitivity             0.5600    0.7200    0.7000    0.8000    0.7800
## Specificity             0.9947    0.9841    0.9951    0.9984    0.9992
## Pos Pred Value          0.6829    0.4800    0.7447    0.9091    0.9512
## Neg Pred Value          0.9911    0.9942    0.9939    0.9959    0.9955
## Prevalence              0.0200    0.0200    0.0200    0.0200    0.0200
## Detection Rate          0.0112    0.0144    0.0140    0.0160    0.0156
## Detection Prevalence    0.0164    0.0300    0.0188    0.0176    0.0164
## Balanced Accuracy       0.7773    0.8520    0.8476    0.8992    0.8896
##                      Class: 42 Class: 43 Class: 44 Class: 45 Class: 46
## Sensitivity             0.6000    0.5600    0.2600    0.5800    0.4600
```

```
## Specificity             0.9849    0.9873    0.9800    0.9951    0.9890
## Pos Pred Value          0.4478    0.4746    0.2097    0.7073    0.4600
## Neg Pred Value          0.9918    0.9910    0.9848    0.9915    0.9890
## Prevalence              0.0200    0.0200    0.0200    0.0200    0.0200
## Detection Rate          0.0120    0.0112    0.0052    0.0116    0.0092
## Detection Prevalence    0.0268    0.0236    0.0248    0.0164    0.0200
## Balanced Accuracy       0.7924    0.7737    0.6200    0.7876    0.7245
##                      Class: 47 Class: 48 Class: 49 Class: 50
## Sensitivity             0.5000    0.7600    0.4200    0.3800
## Specificity             0.9886    0.9902    0.9857    0.9873
## Pos Pred Value          0.4717    0.6129    0.3750    0.3800
## Neg Pred Value          0.9898    0.9951    0.9881    0.9873
## Prevalence              0.0200    0.0200    0.0200    0.0200
## Detection Rate          0.0100    0.0152    0.0084    0.0076
## Detection Prevalence    0.0212    0.0248    0.0224    0.0200
## Balanced Accuracy       0.7443    0.8751    0.7029    0.6837

## Accuracy
##   0.6036
```

Summing up, the Naive Bayes model has an accuracy of approximately 60% at predicting the author identities out of sample. Also by looking the sensitivity at every class we can say that the authors whose articles seem difficult to distinguish from one another are for example author 4 with a sensisitivity of 0.2, author 8 with a sensitivity of 0.14 and author 44 with a sensitivity of 0.26.

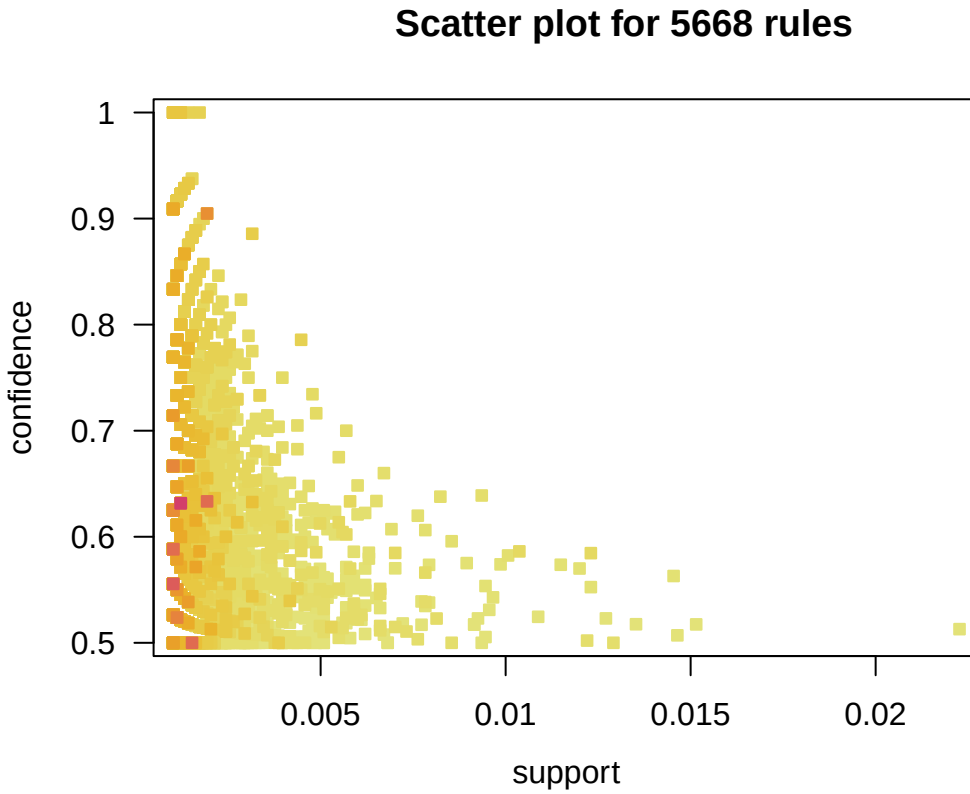
Practice with association rule mining

Use the data on grocery purchases in groceries.txt and find some interesting association rules for these shopping baskets. The data file is a list of baskets: one row per basket, with multiple items per row separated by commas – you'll have to cobble together a few utilities for processing this into the format expected by the "arules" package. Pick your own thresholds for lift and confidence; just be clear what these thresholds are and how you picked them. Do your discovered item sets make sense? Present your discoveries in an interesting and concise way.

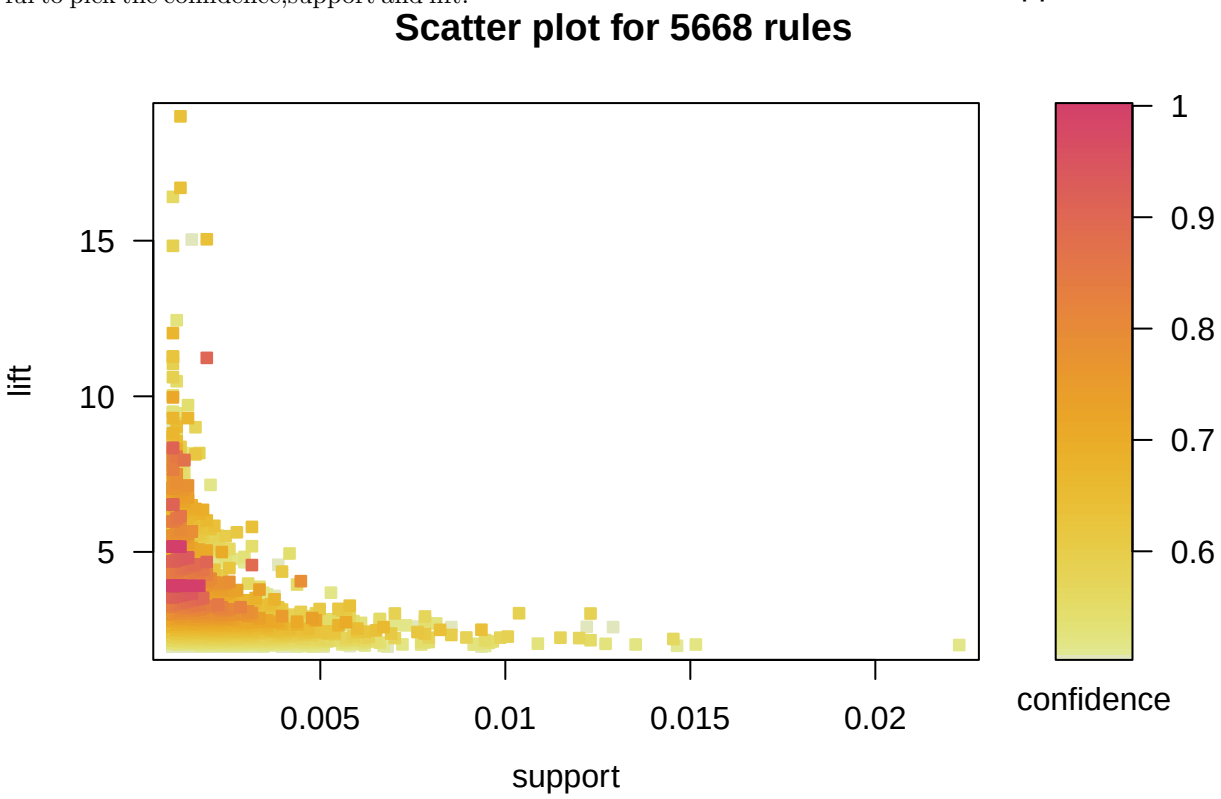
We begin with a set of 5668 rules.

We can see that the 3 most popular rules acording to lift are:

53 {Instant food products,soda} => {hamburger meat}
37 {soda,popcorn} => {salty snack}
444 {flour,baking powder} => {sugar}
Next we can see these very two informative scatterplots to visualise the different set of rules. It can also be help-



ful to pick the confidence,support and lift.



By seting the confidence to be 0.8 or greater we see that we are left with a set of 371 rules.

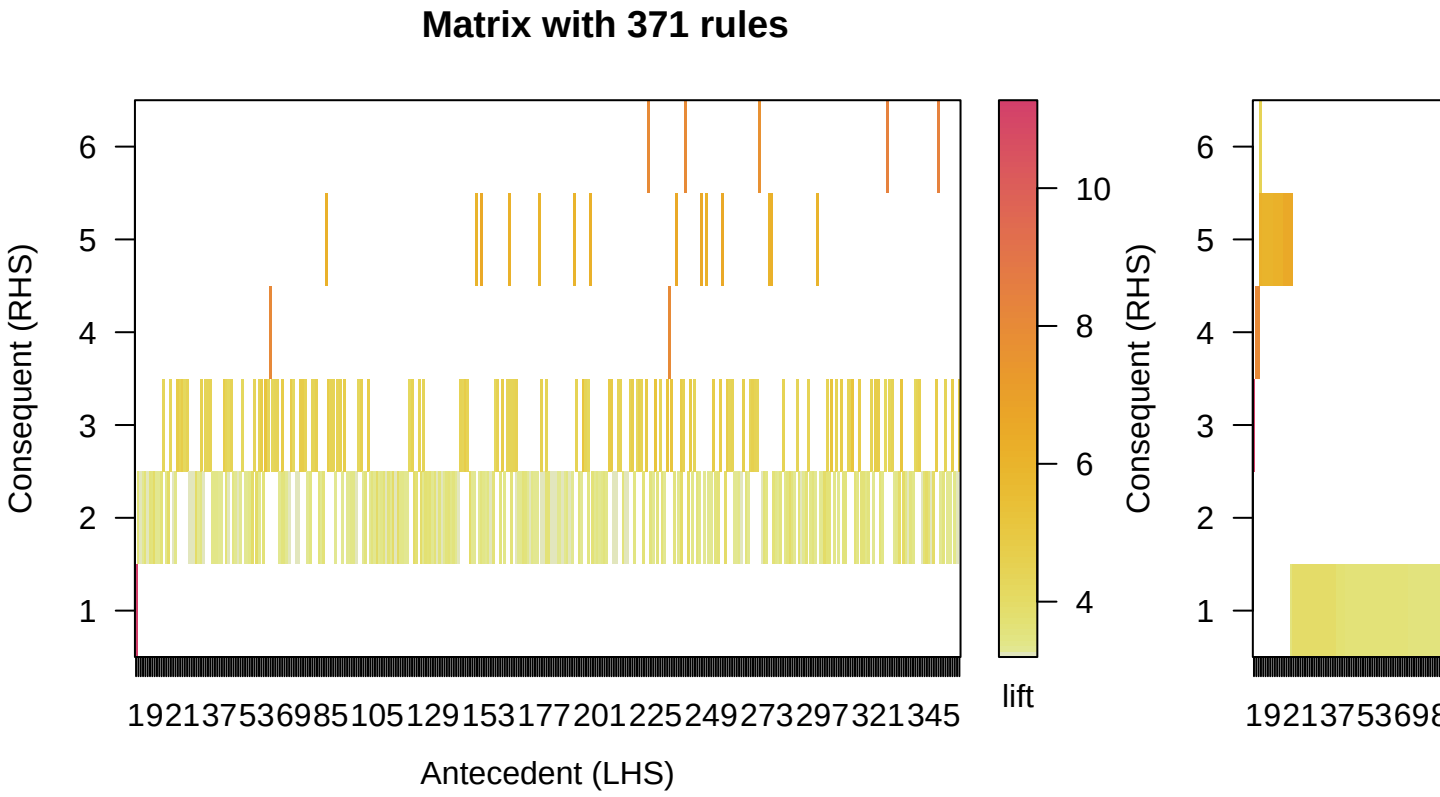
The confidence of a rule is the likelihood that it is true for a new transaction that contains the items on the LHS of the rule. (I.e. it is the probability that the transaction also contains the item(s) on the RHS.)

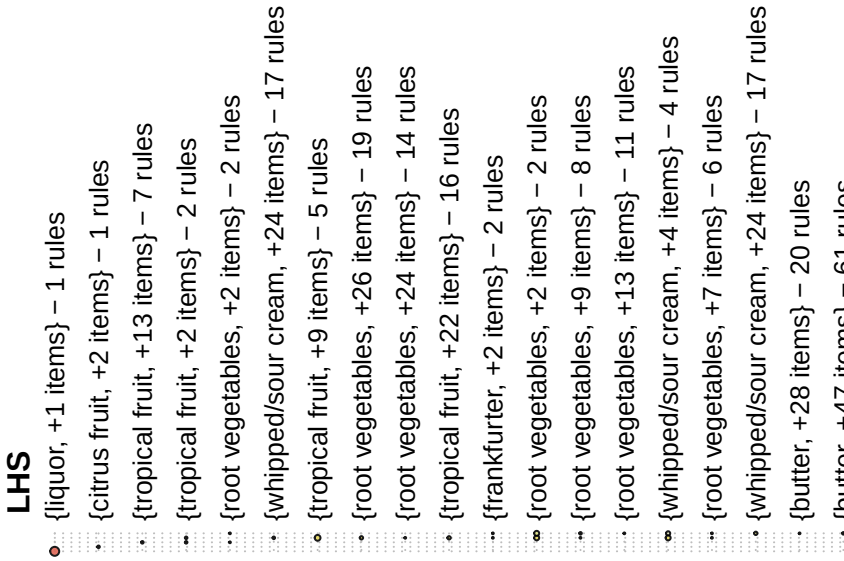
Formally:

The lift of a rule is the ratio of the support of the items on the LHS of the rule co-occuring with items on the RHS divided by probability that the LHS and RHS co-occur if the two are independent.

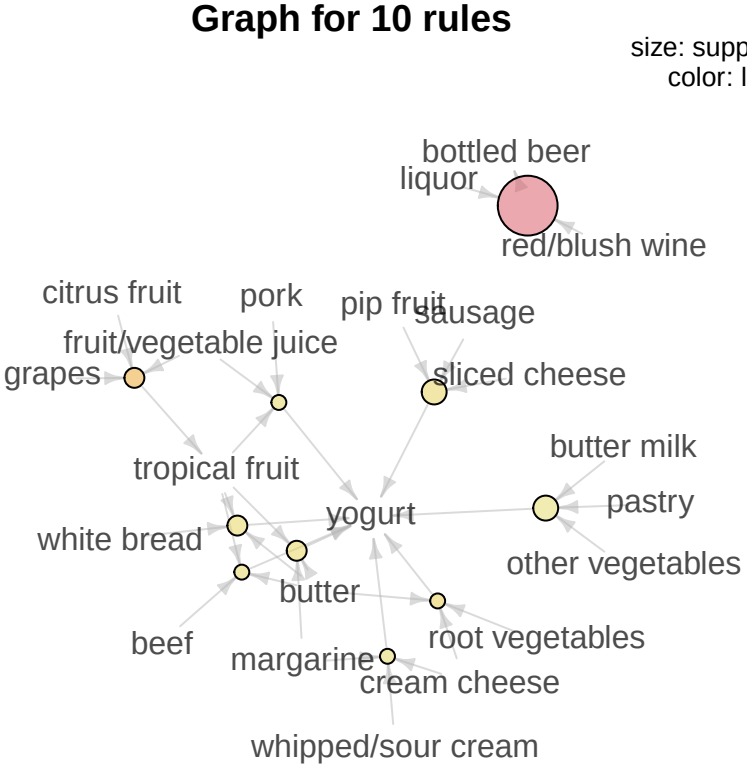
If lift is greater than 1, it suggests that the precense of the items on the LHS has increased the probability that the items on the right hand side will occur on this transaction. If the lift is below 1, it suggests that the presence of the items on the LHS make the probability that the items on the RHS will be part of the transaction lower. If the lift is 1, it suggests that the presence of items on the LHS and RHS really are independent: knowing that the items on the LHS are present makes no difference to the probability that items will occur on the RHS.

Another set of plots that gives us information about the set of rules at hand can be seen below:





Here is a balloon plot with our grouped relationships:

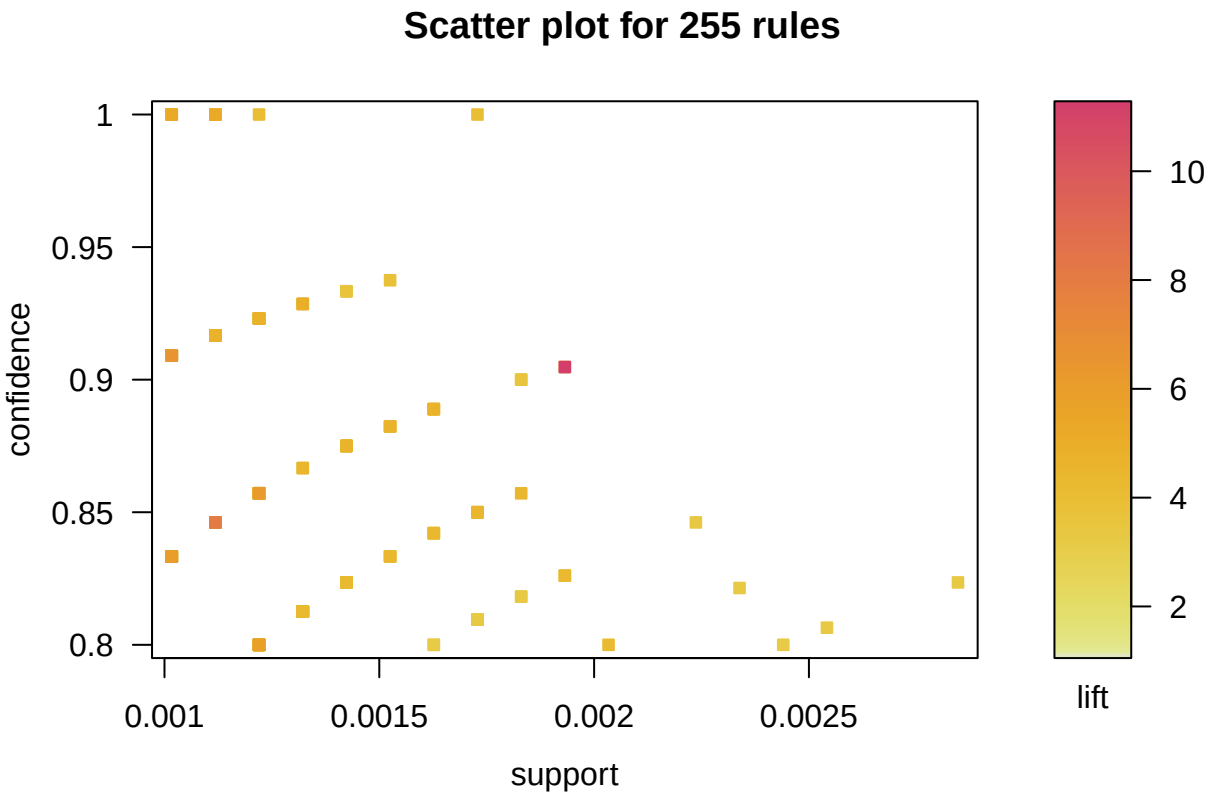


And another one sorted by lift:

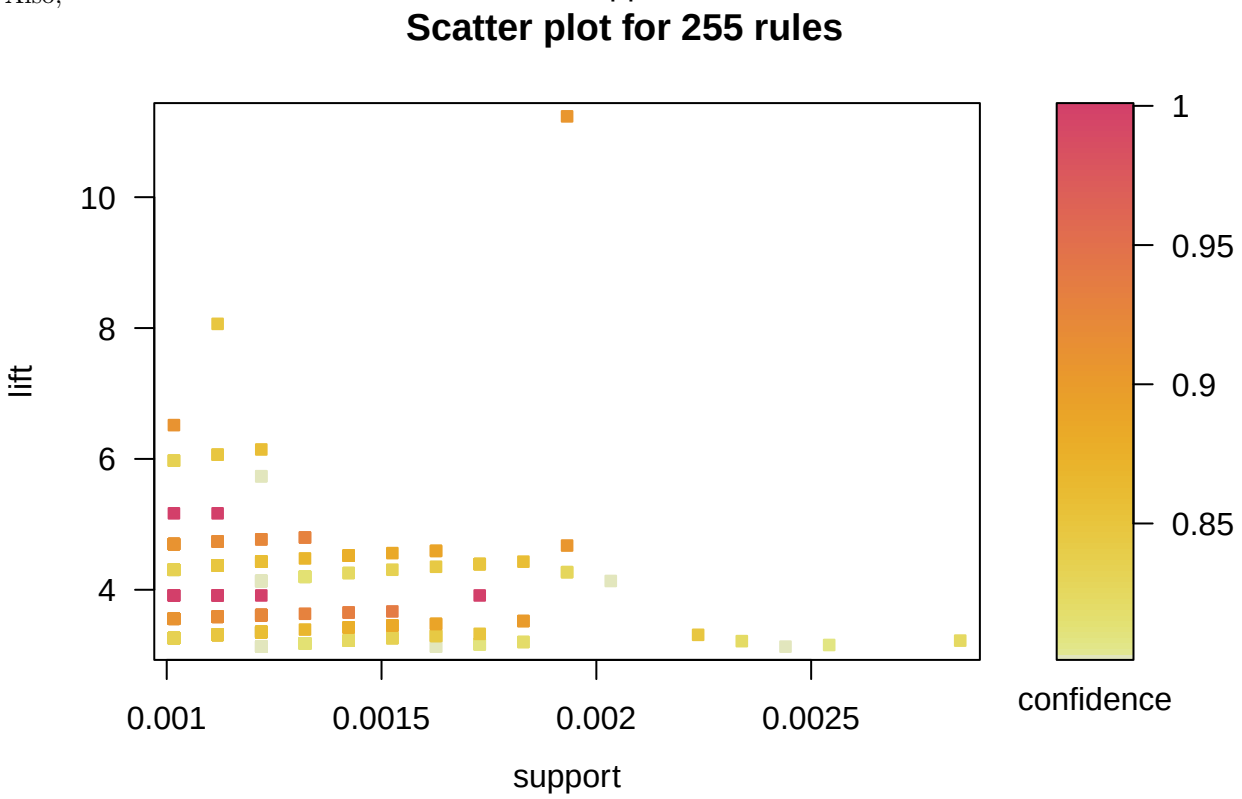
Given that, I will use a support of 0.001 and confidence of 0.8. The summary of our rules then is:

Next we can see what is most probable to hapen if we list the rules by confidence.

1 {rice,sugar} => {whole milk} 2 {canned fish,hygiene articles} => {whole milk} 3 {root vegetables,butter,rice} => {whole milk}
4 {root vegetables,whipped/sour cream,flour} => {whole milk}
5 {butter,soft cheese,domestic eggs} => {whole milk}



Also,

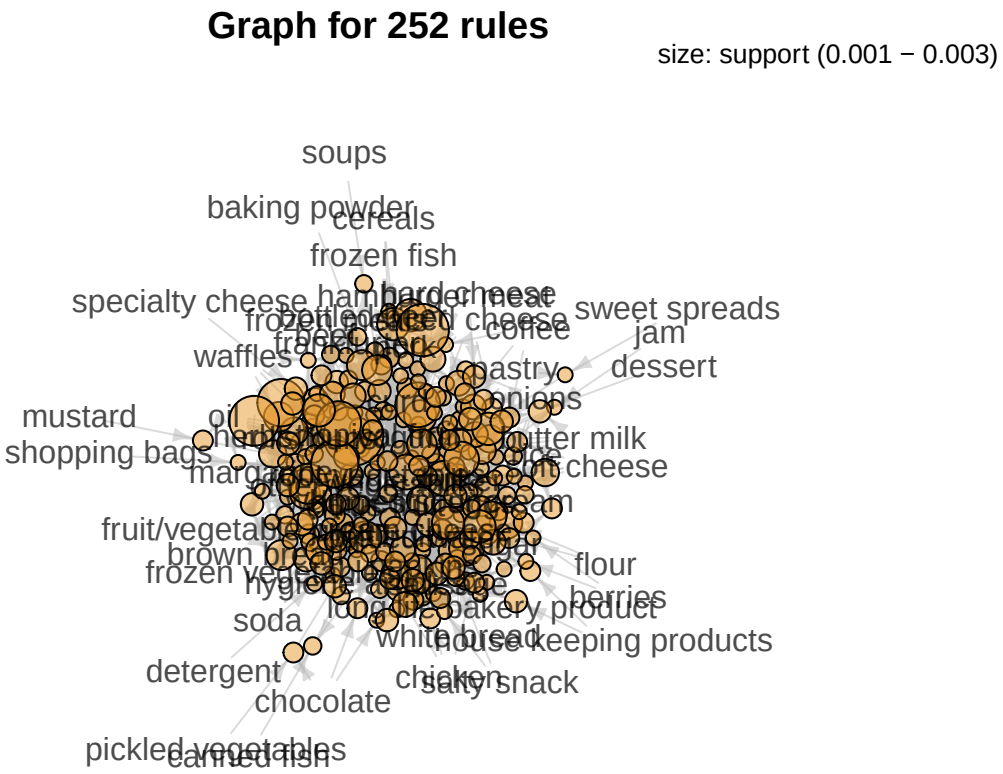


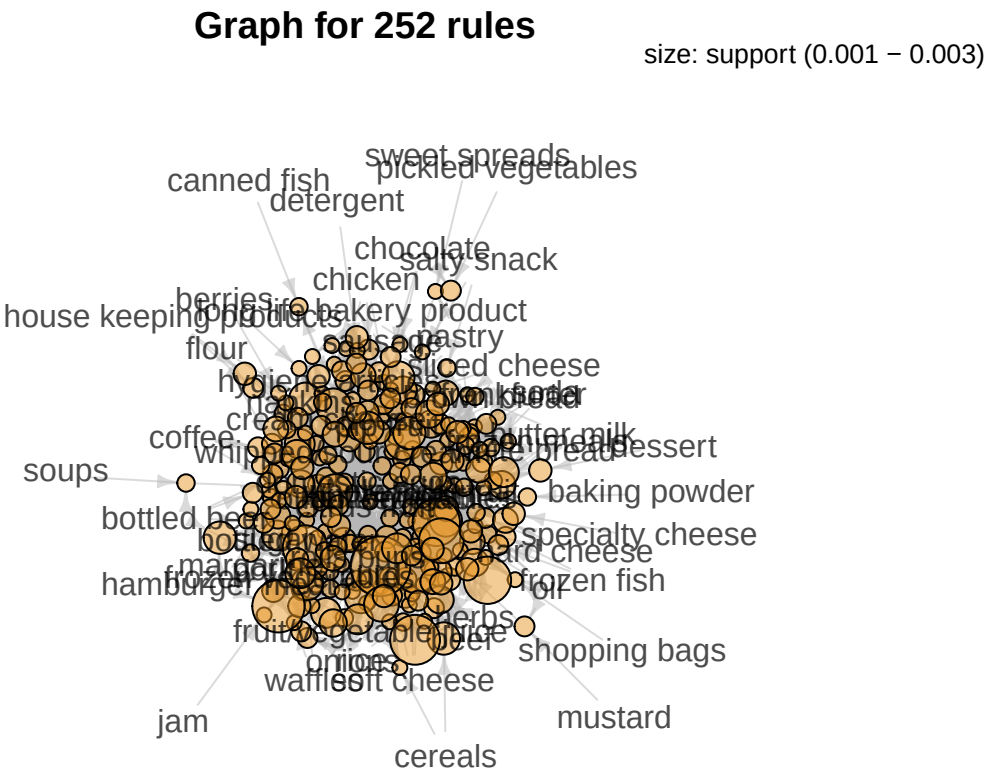

```
##        support confidence     lift
## 10 0.001220132          1 3.913649
## 16 0.001118454          1 3.913649
```

```
## 33 0.001016777          1 3.913649
## 60 0.001728521          1 3.913649
## 62 0.001016777          1 3.913649
## 68 0.001016777          1 5.168156
```

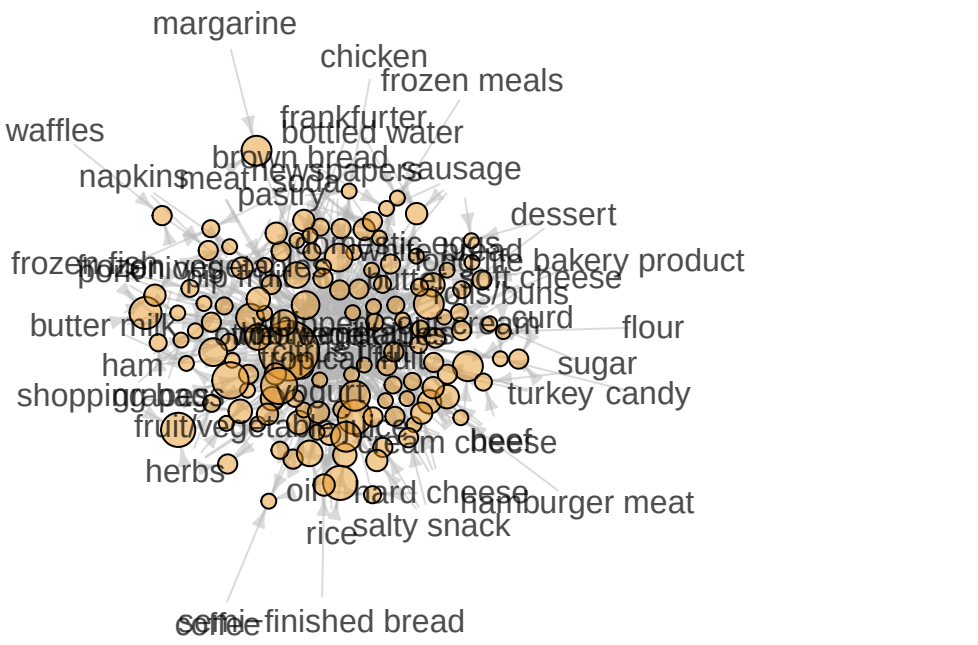
Now from our first frequency plot, we will emphasize on the first 5 selling items:

What are the customersmore likely to buy before buying whole milk?

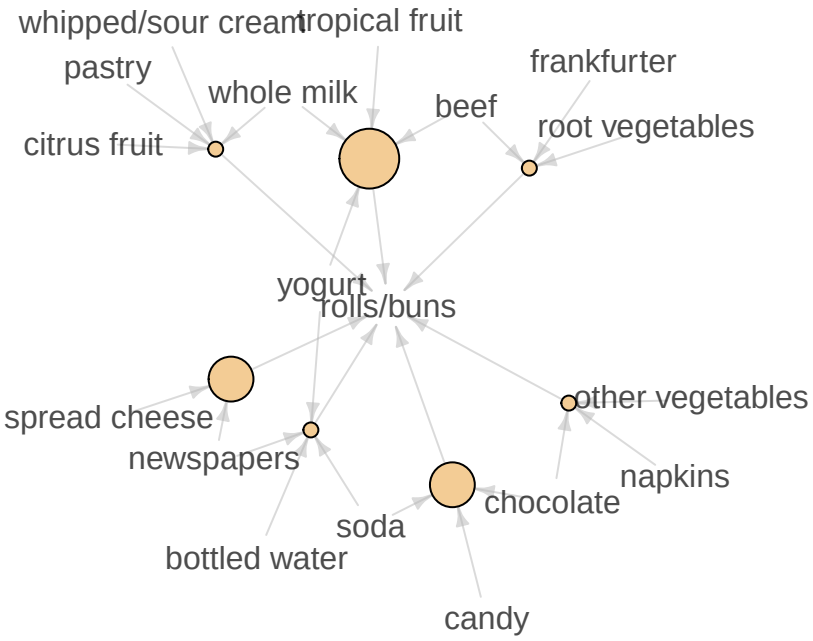




What are the customersmore likely to buy before buying other veggies?

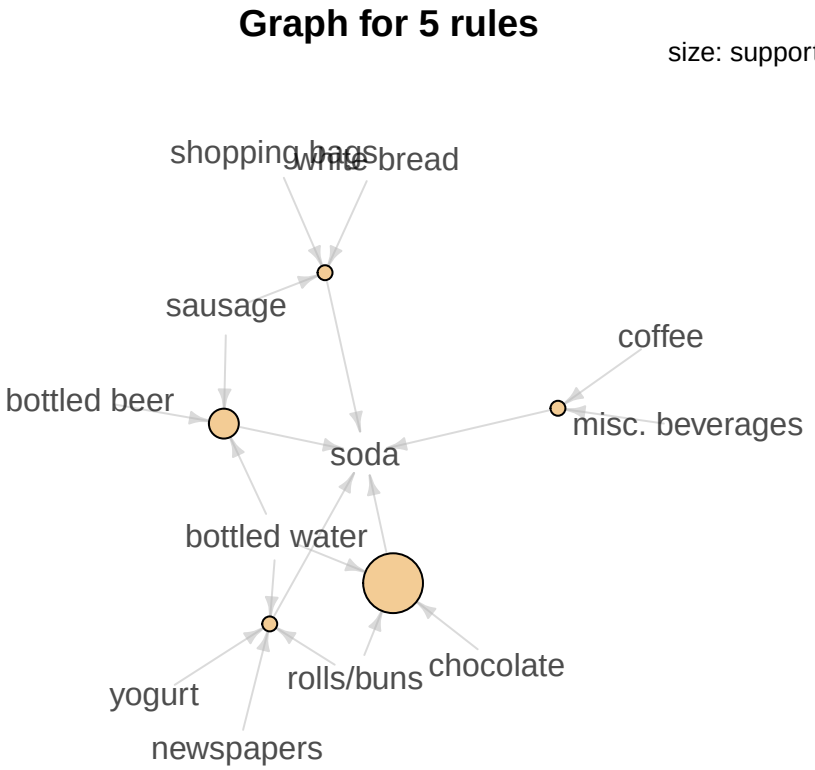


What are the customersmore likely to buy before buying rolls/buns?

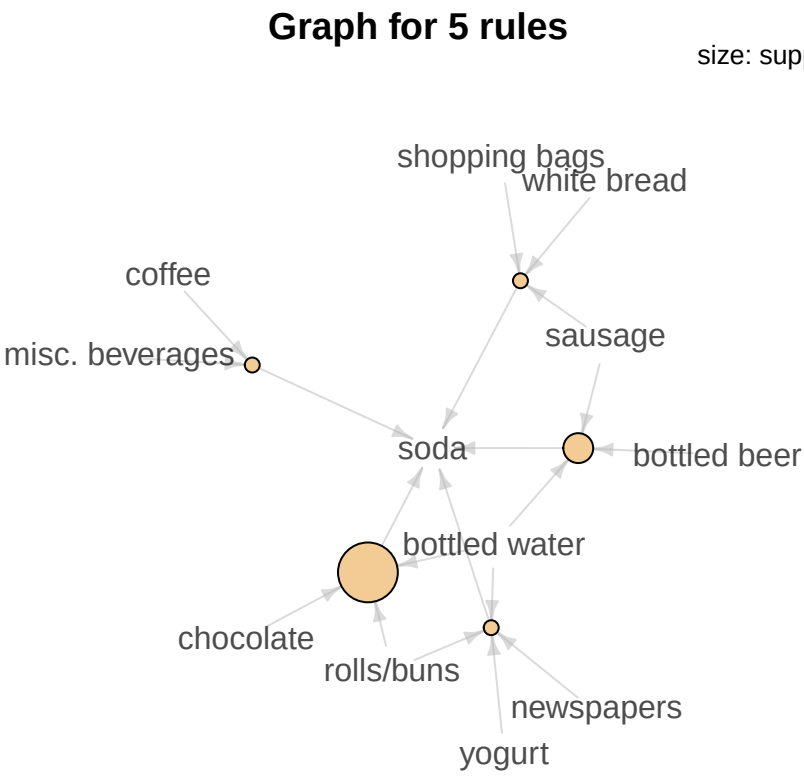


What are the customersmore likely to buy before buying soda?

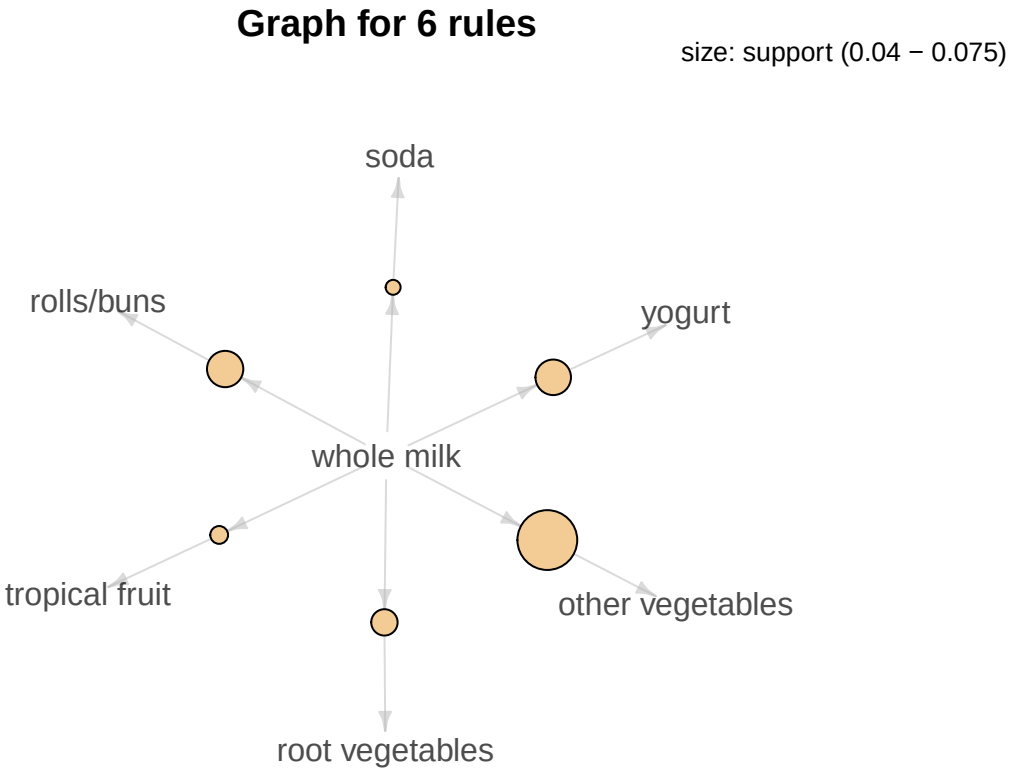
```
##   lhs                    rhs        support confidence     lift
## 1 {coffee,
##    misc. beverages} => {soda} 0.001016777  0.7692308 4.411303
## 2 {yogurt,
##    rolls/buns,
##    bottled water,
##    newspapers}      => {soda} 0.001016777  0.7692308 4.411303
## 3 {sausage,
##    bottled water,
##    bottled beer}    => {soda} 0.001118454  0.7333333 4.205442
## 4 {sausage,
##    white bread,
##    shopping bags}   => {soda} 0.001016777  0.6666667 3.823129
## 5 {rolls/buns,
##    bottled water,
##    chocolate}       => {soda} 0.001321810  0.6500000 3.727551
```



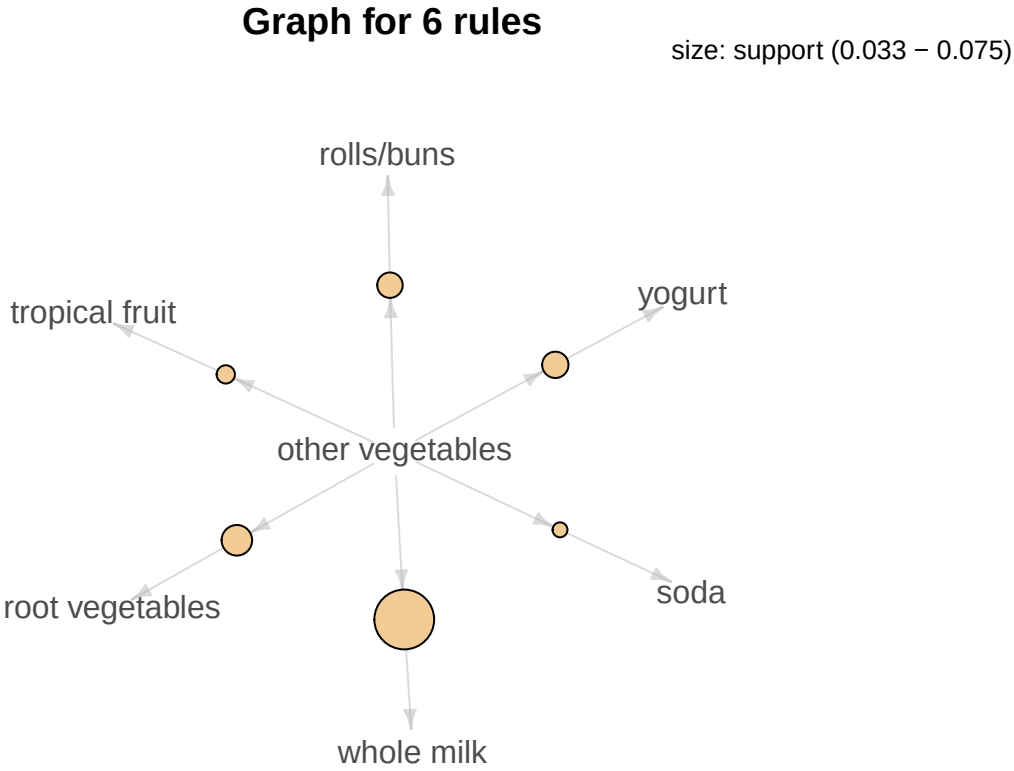
What are the customersmore likely to buy before buying yogurt?



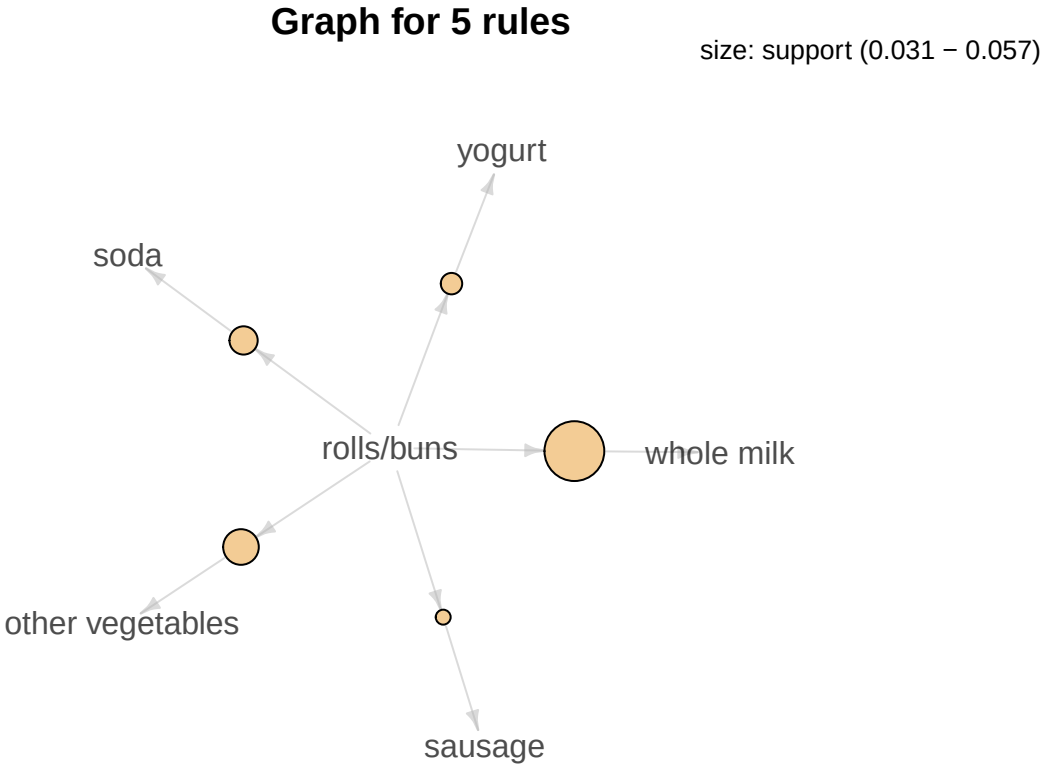
What are customers likely to buy if they purchase whole milk?



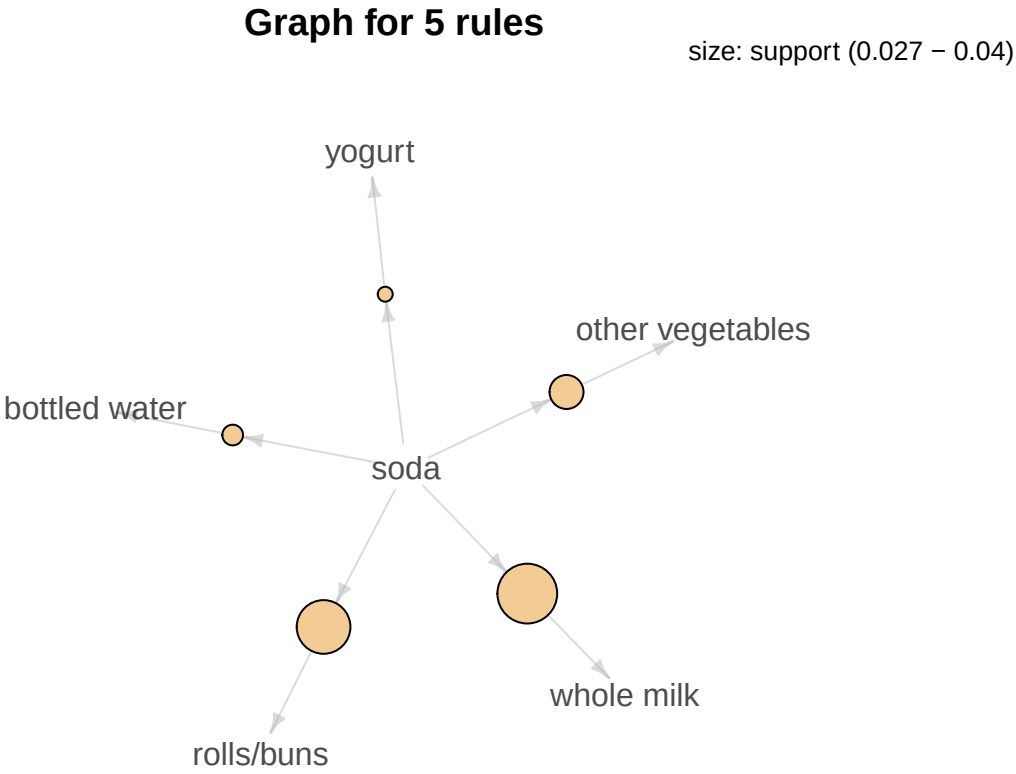
What are customers likely to buy if they purchase other vegetables?



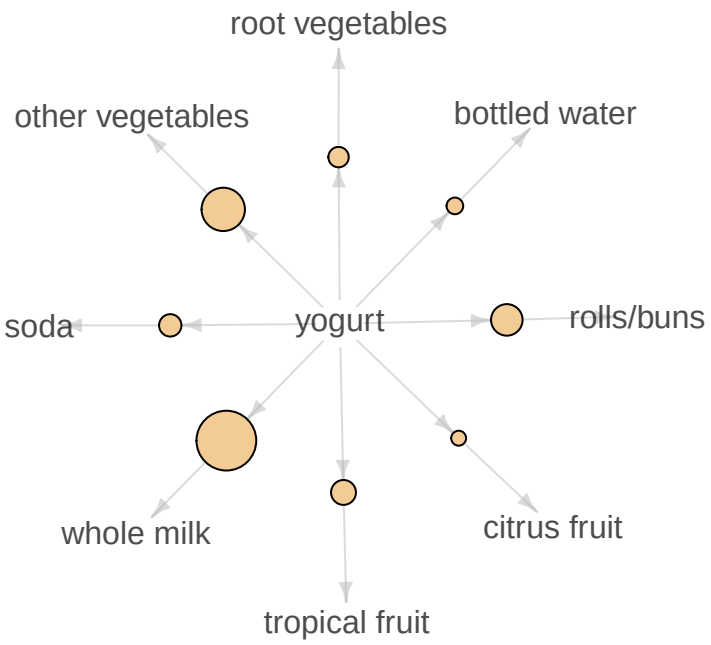
What are customers likely to buy if they purchase rolls/buns?



What are customers likely to buy if they purchase soda?



What are customers likely to buy if they purchase yogurt?


Graph for 8 rules
size: support (0.022 − 0.056)
root vegetables
other vegetables
bottled water
soda
yogurt
rolls/buns
whole milk
citrus fruit
tropical fruit