Assignment 2

Marissa Hausman

August 16, 2016

I. Flights at ABIA

1. Overview

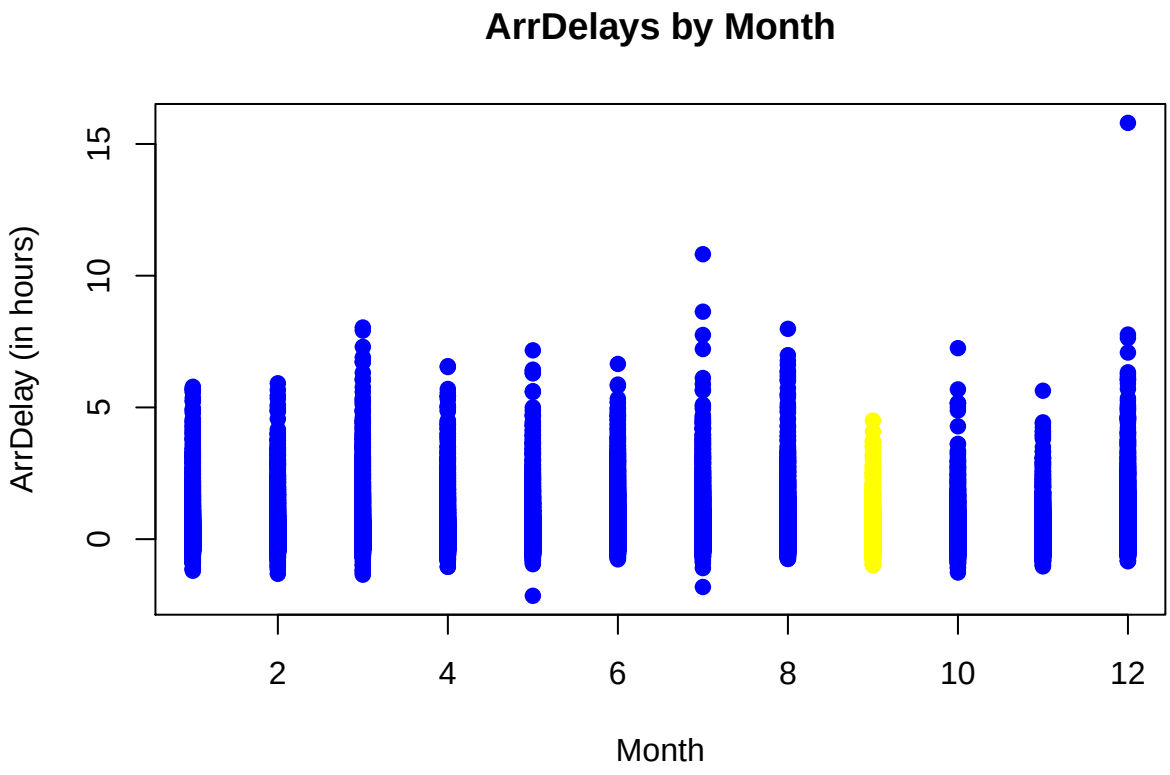
Create a figure, or set of related figures, that tell an interesting story about flights into and out of Austin. What is the best time to fly to minimize delays?

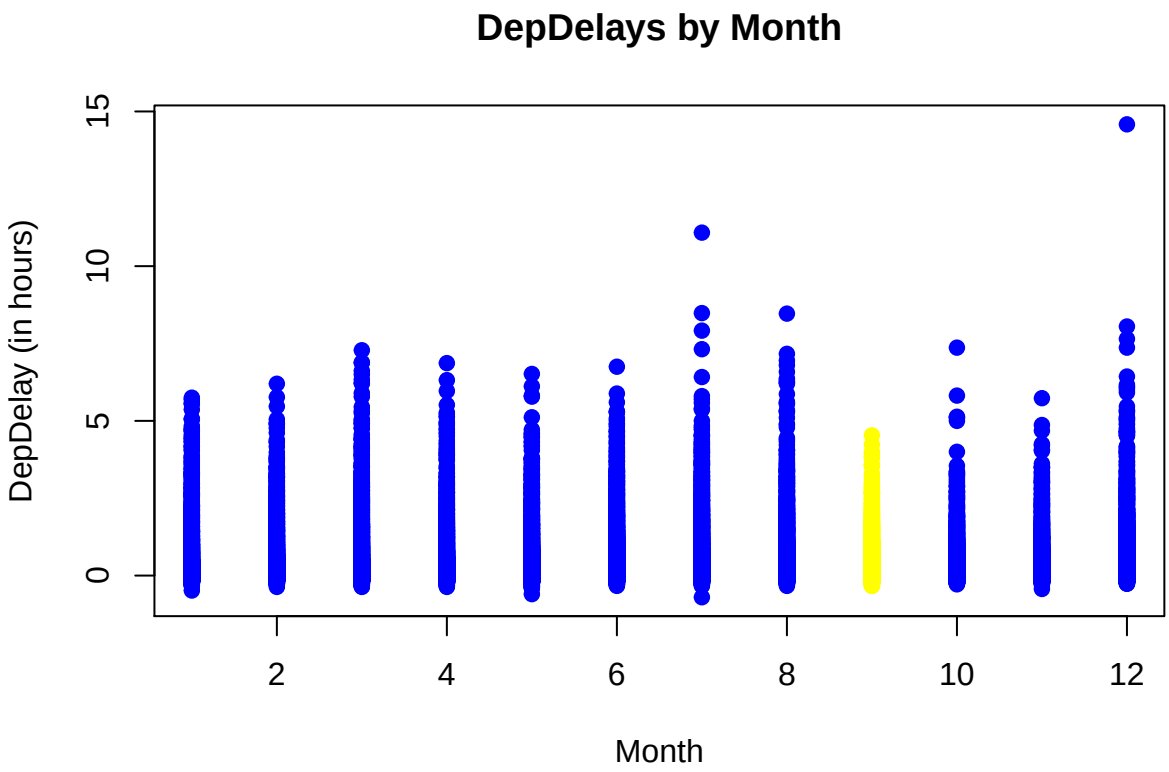
2. Data and Model

In order to address this question, the Month, DayOfWeek, DepTime, ArrDelay, and DepDelay columns of the dataset, ABIA, will be used. First plot ArrDelay and DepDelay by Month to find the Month with the shortest delays. Next subset the data to only that Month and find the DayOfWeek with the shortest delays. Next subset the data to only that DayOfWeek of that Month and find the group of DepTime with the shortest delays. The results of this analysis provide the time to fly that has been, in the past, the time to fly with the shortest delays. This is the optimal time to fly to minimize delays.

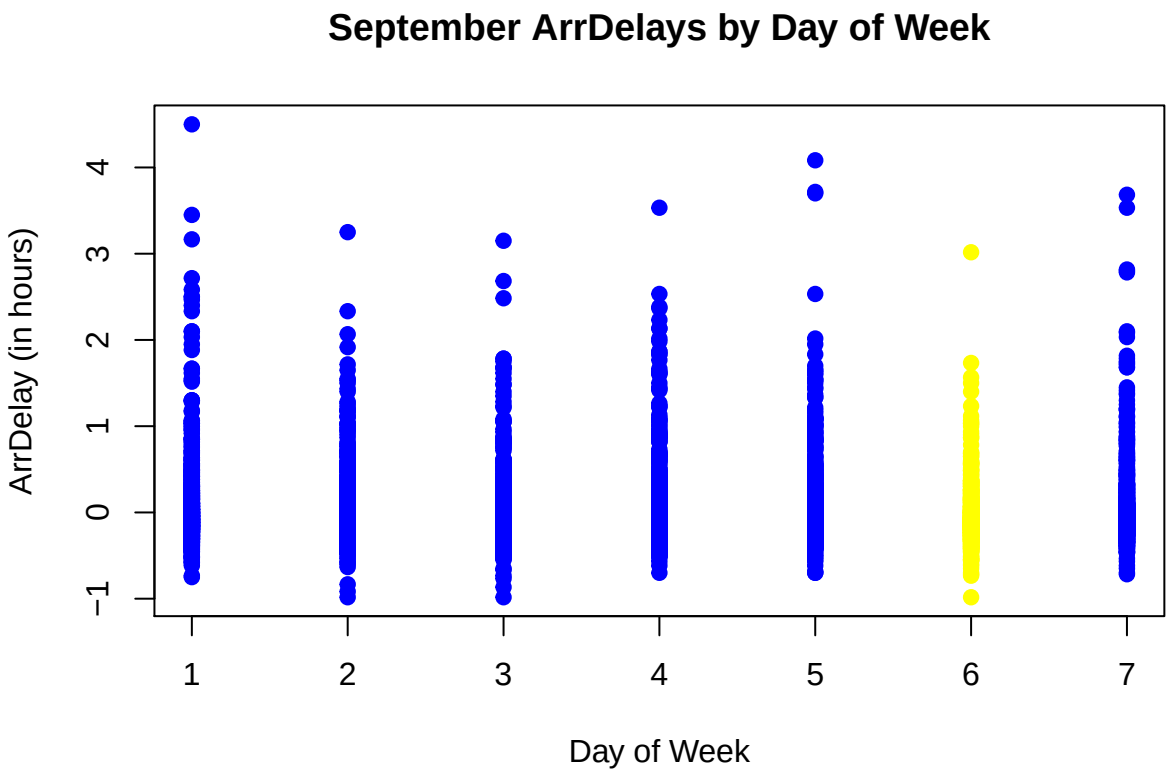
3. Results

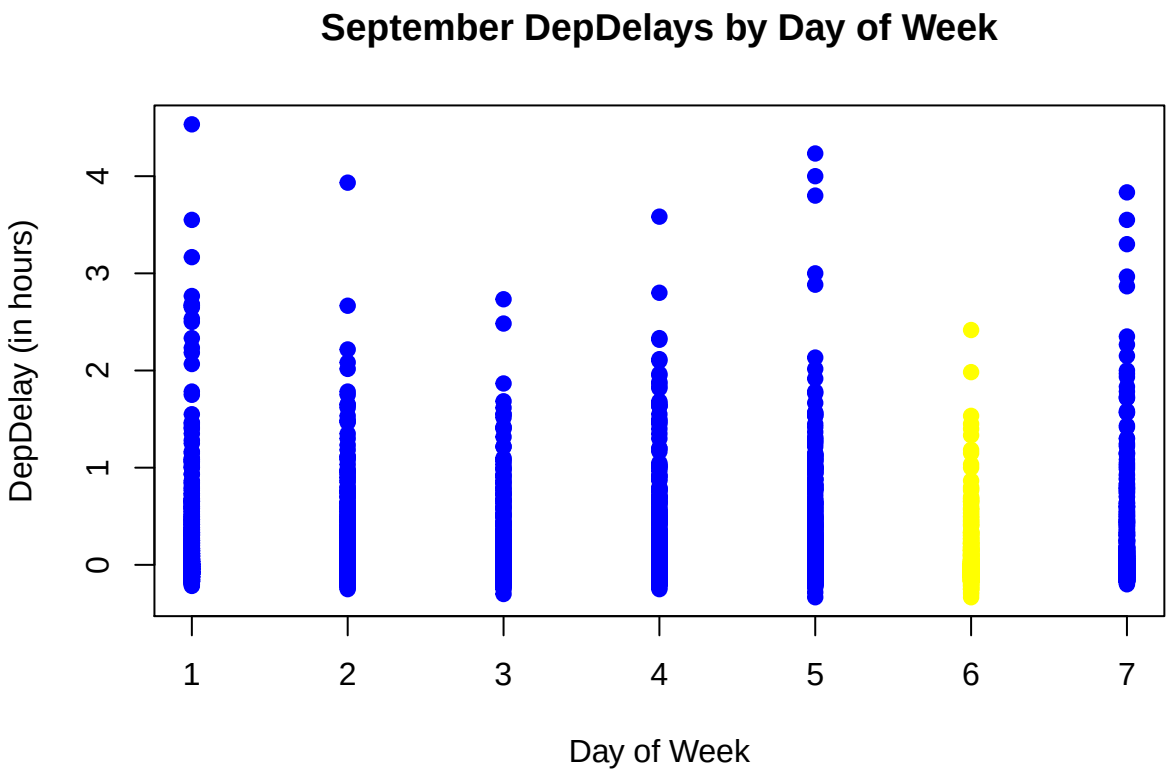
September has the shortest delays, on average:



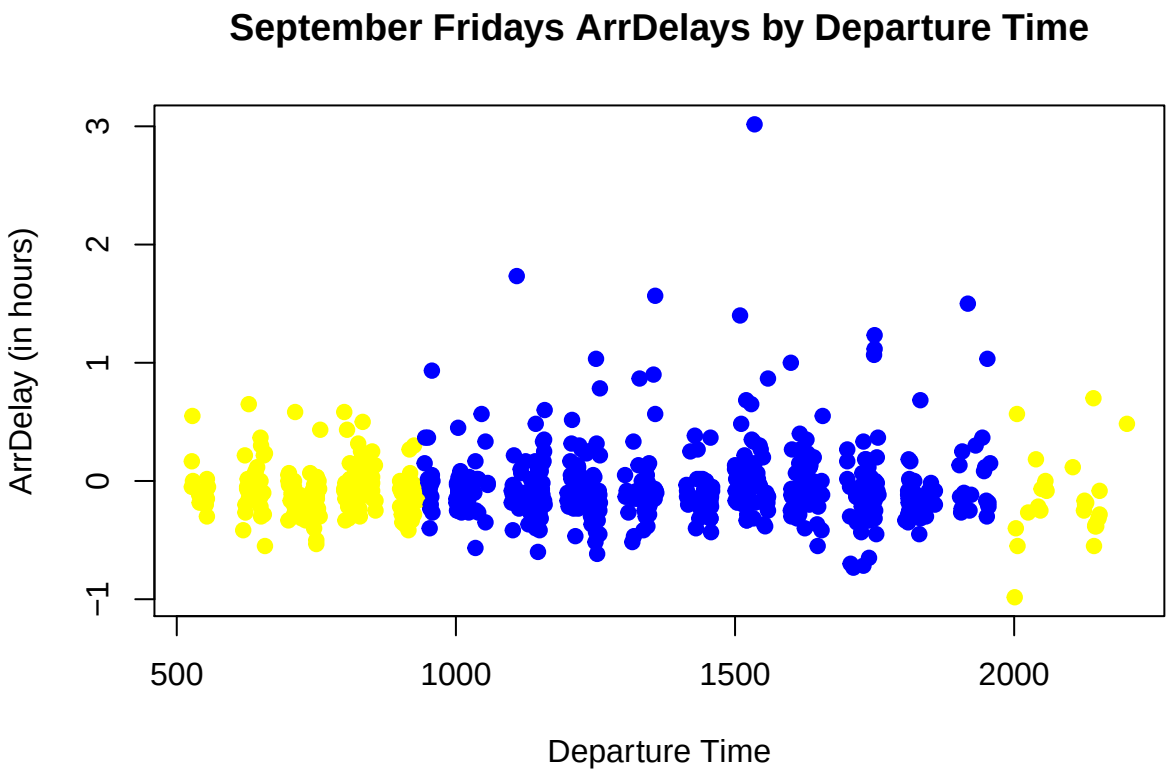


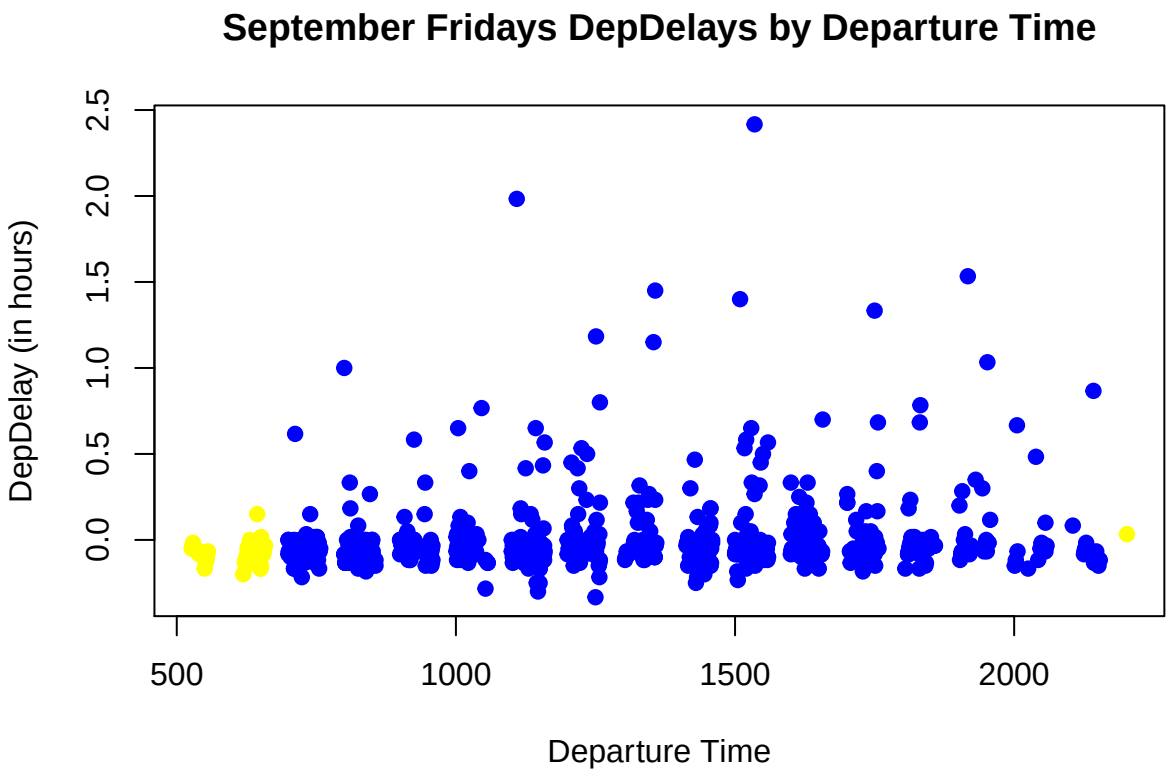
Fridays in September have the shortest delays, on average:


September ArrDelays by Day of Week
ArrDelay (in hours)
−1
0
1
2
3
4
1
2
3
4
5
6
7
Day of Week



Departure times in the early morning and late at night on Fridays in September have the shortest delays, on average:


September Fridays ArrDelays by Departure Time
ArrDelay (in hours)
−1
0
1
2
3
500
1000
1500
2000
Departure Time



4. Conclusion

The best Month to fly to minimize delays is September. The best DayOfWeek to fly in September is Friday. The best DepTime to fly on Fridays in September is very early in the morning or very late at night.

II. Author Attribution

I. Overview

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?

2. Data and Model

Naive Bayes will be used to detect the authors of articles based on those articles' textual content. A document term matrix will be created for both the training and test sets and the log probabilities will be compared.

For each test document, the author corresponding to the largest log probability is that document's predicted author. The accuracy of these results will be tested using a confusion matrix.

3. Results

```
## 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  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  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
##          47  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  8  1  0  4  0
##          49  0  0  0  0  0  1  0 23  6  0  0  0  0  0  0  0  0  0  1  0  0
##          50  0  0  0  9  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
```

Classes with low (<0.4) sensitivity (pos pred value): 4, 7, 8, 9, 13, 15, 35, 44, 50

4. Conclusion

Overall, **Naive Bayes is 60.36% accurate** in determining the author identities in this out-of-sample setting. Authors whose articles seem difficult to identify are: Benjamin Kang Lim, Darren Schuettler, David Lawder, Edna Fernandes, Heather Scoffield, Jan Lopatka, Mure Dickie, Scott Hillis, and William Kazer.

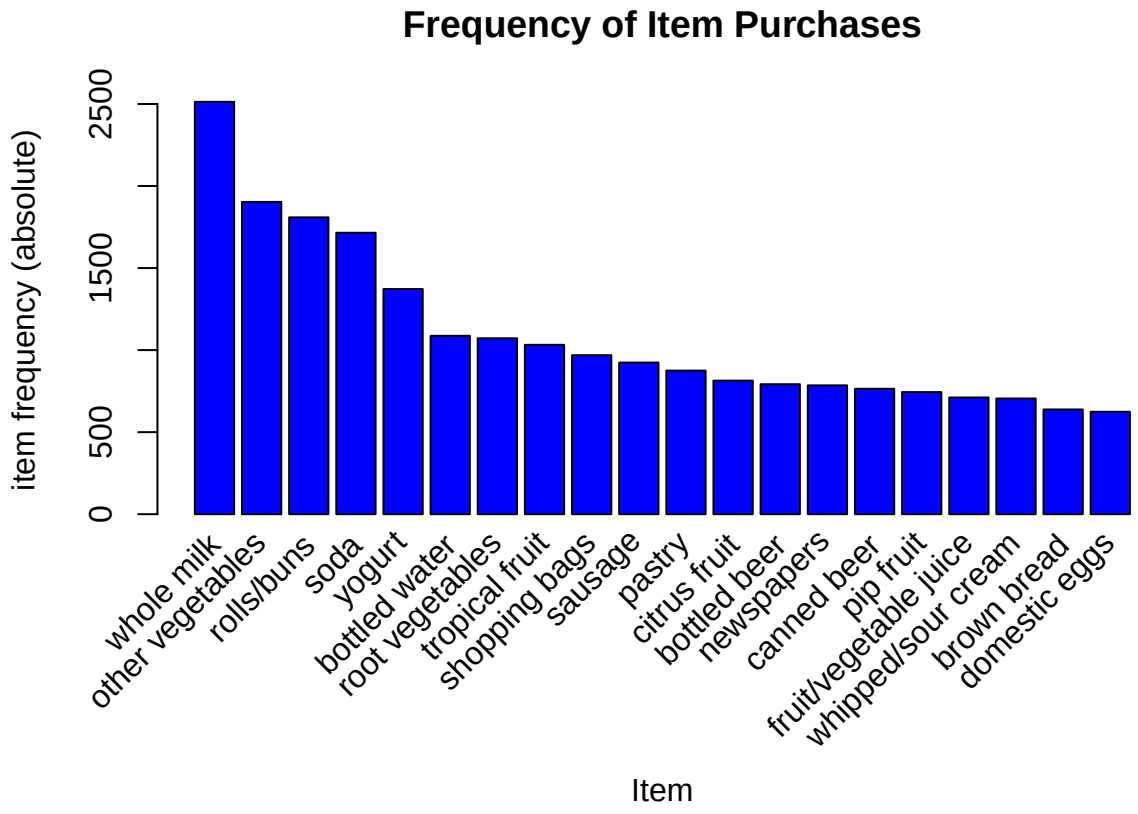
III. Practice with Association Rule Mining

1. Overview

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.

2. Data and Model

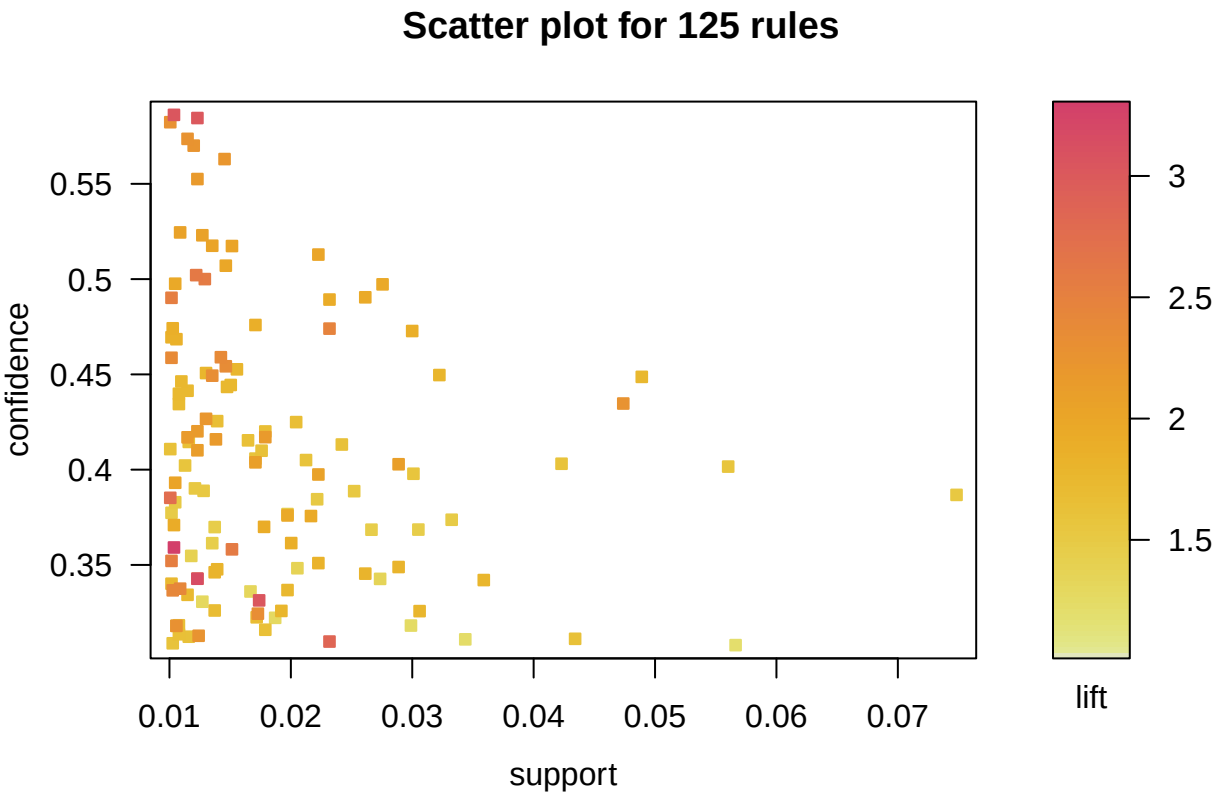
The following table shows the top 20 items in the groceries.txt dataset:



The 'arules' package will be used to perform a market basket analysis on the groceries.txt dataset. The thresholds for support, confidence, and lift will be determined based on the number of rules generated by different combinations.

3. Results

With support=.01 and confidence=.3, there are 125 rules:



Choice of Thresholds:

- Support=.01. The percent of transactions with the item set is at least 1%. Items must meet this minimum sales threshold.
- Confidence=.3. Item Y must appear in baskets that contain X at least 30% of the time.
- Lift=2.3. A lift of 2.3 provides 21 remaining rules to work with. For each of these rules, people are 2.3 times more likely to have item Y in their basket than random.

The following are the 21 association rules:

```
##    lhs                      rhs                    support confidence     lift
## 1  {onions}              => {other vegetables} 0.01423488  0.4590164 2.372268
## 2  {beef}                => {root vegetables}  0.01738688  0.3313953 3.040367
## 3  {curd}                => {yogurt}           0.01728521  0.3244275 2.325615
## 4  {curd,
##     whole milk}          => {yogurt}           0.01006609  0.3852140 2.761356
## 5  {pork,
##     whole milk}          => {other vegetables} 0.01016777  0.4587156 2.370714
## 6  {whipped/sour cream,
##     yogurt}              => {other vegetables} 0.01016777  0.4901961 2.533410
## 7  {other vegetables,
##     whipped/sour cream}  => {yogurt}           0.01016777  0.3521127 2.524073
## 8  {whipped/sour cream,
##     whole milk}          => {yogurt}           0.01087951  0.3375394 2.419607
## 9  {whipped/sour cream,
##     whole milk}          => {other vegetables} 0.01464159  0.4542587 2.347679
## 10 {pip fruit,
##     whole milk}          => {other vegetables} 0.01352313  0.4493243 2.322178
```

```
## 11 {citrus fruit,
##     root vegetables}    => {other vegetables} 0.01037112  0.5862069 3.029608
## 12 {citrus fruit,
##     other vegetables}   => {root vegetables}  0.01037112  0.3591549 3.295045
## 13 {citrus fruit,
##     whole milk}         => {yogurt}           0.01026945  0.3366667 2.413350
## 14 {root vegetables,
##     tropical fruit}     => {other vegetables} 0.01230300  0.5845411 3.020999
## 15 {other vegetables,
##     tropical fruit}     => {root vegetables}  0.01230300  0.3427762 3.144780
## 16 {other vegetables,
##     tropical fruit}     => {yogurt}           0.01230300  0.3427762 2.457146
## 17 {tropical fruit,
##     whole milk}         => {yogurt}           0.01514997  0.3581731 2.567516
## 18 {root vegetables,
##     yogurt}             => {other vegetables} 0.01291307  0.5000000 2.584078
## 19 {rolls/buns,
##     root vegetables}    => {other vegetables} 0.01220132  0.5020921 2.594890
## 20 {root vegetables,
##     whole milk}         => {other vegetables} 0.02318251  0.4740125 2.449770
## 21 {other vegetables,
##     whole milk}         => {root vegetables}  0.02318251  0.3097826 2.842082
```

4. Conclusion

From the results, it is apparent that other vegetables, root vegetables, and yogurt are the items most likely to be purchased based on various item sets. All three of those items fall in the top 20 most frequently purchased items (other vegetables ranks #2, yogurt ranks #5, and root vegetables ranks #7). The discovered item sets make sense: in looking at specific item sets, it is apparent that people who buy other dairy products (whipped/sour cream, whole milk) are likely to buy yogurt, people who buy other produce (other vegetables, tropical fruit) are likely to buy root vegetables, and people who buy specific vegetables (onions, root vegetables) are likely to buy other vegetables.