# STA 380 Homework 2

Emily Graves and Hillary Regan

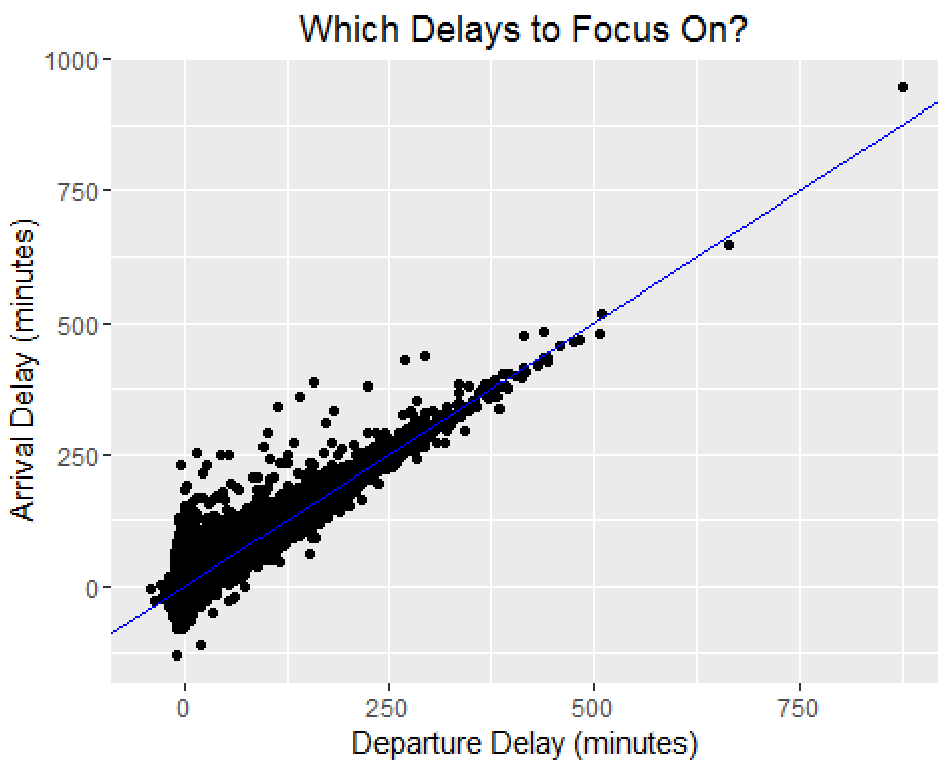
August 16, 2016

## Flights at ABIA

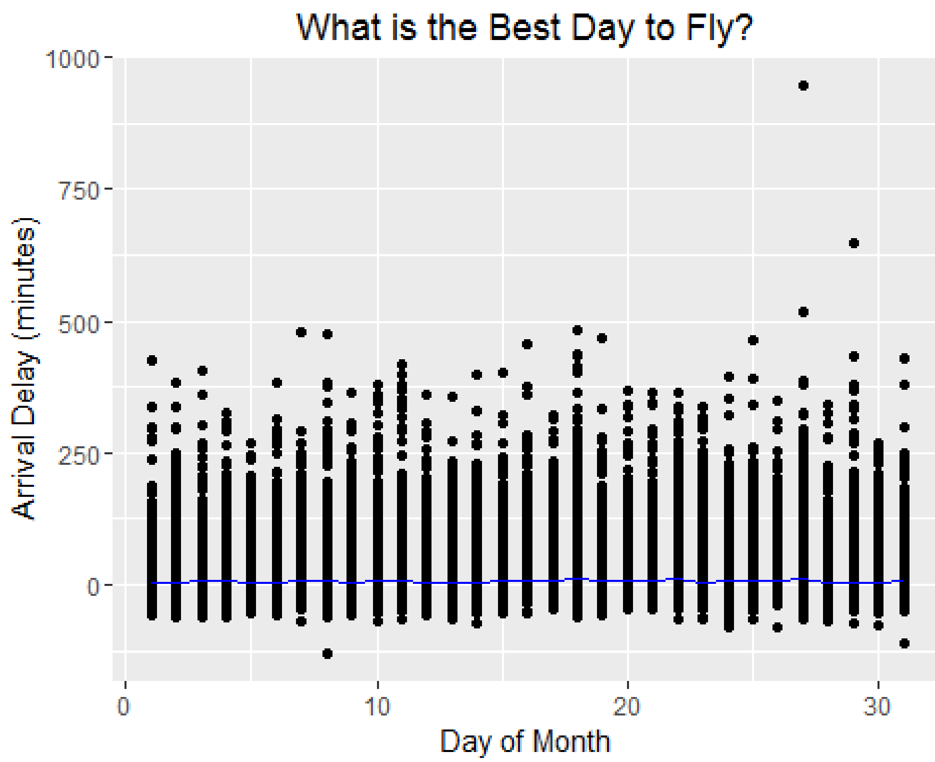
### Assignment:

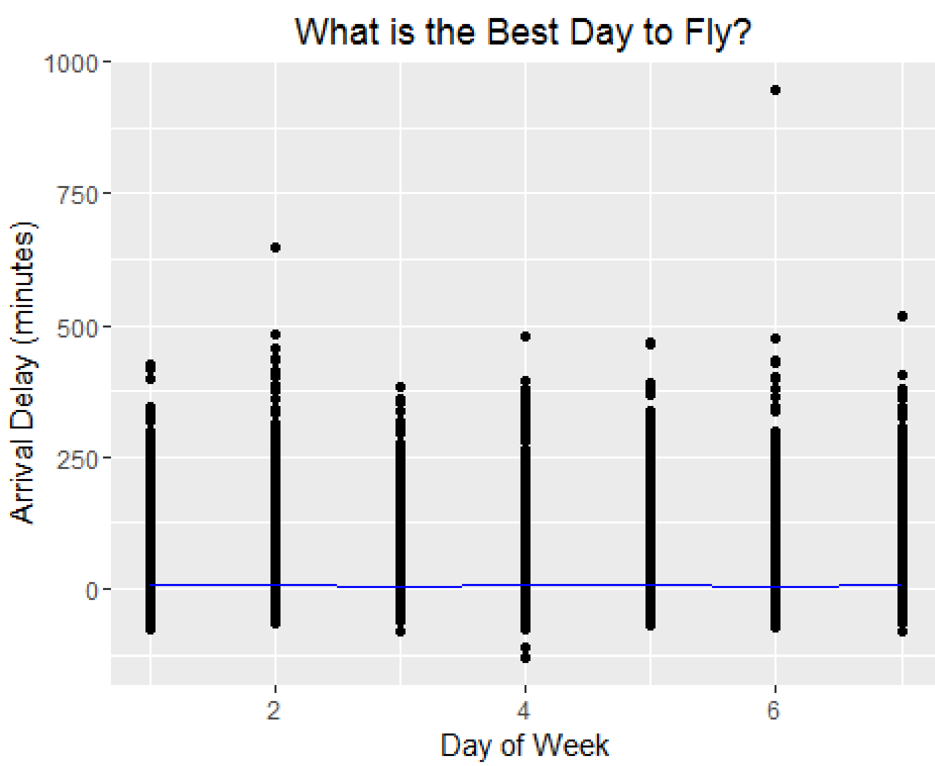
Tell an interesting story about flights into and out of Austin.

### Solution:

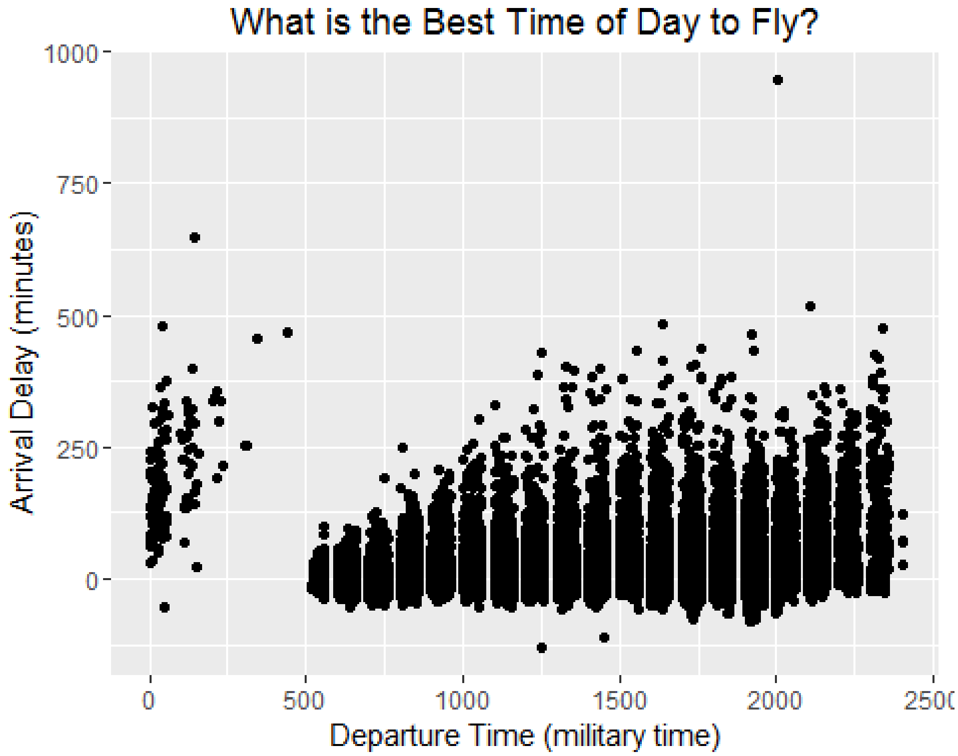



To determine which delay variable to focus on, the relationship between departure delays and arrival delays was plotted. The result shows that while there is essentially a 1 to 1 relationship, Arrival Delays have a tendency to be average on higher. Additionally, Arrival Delays have greater influence of flying experience, so Arrival Delays were used in the analysis.

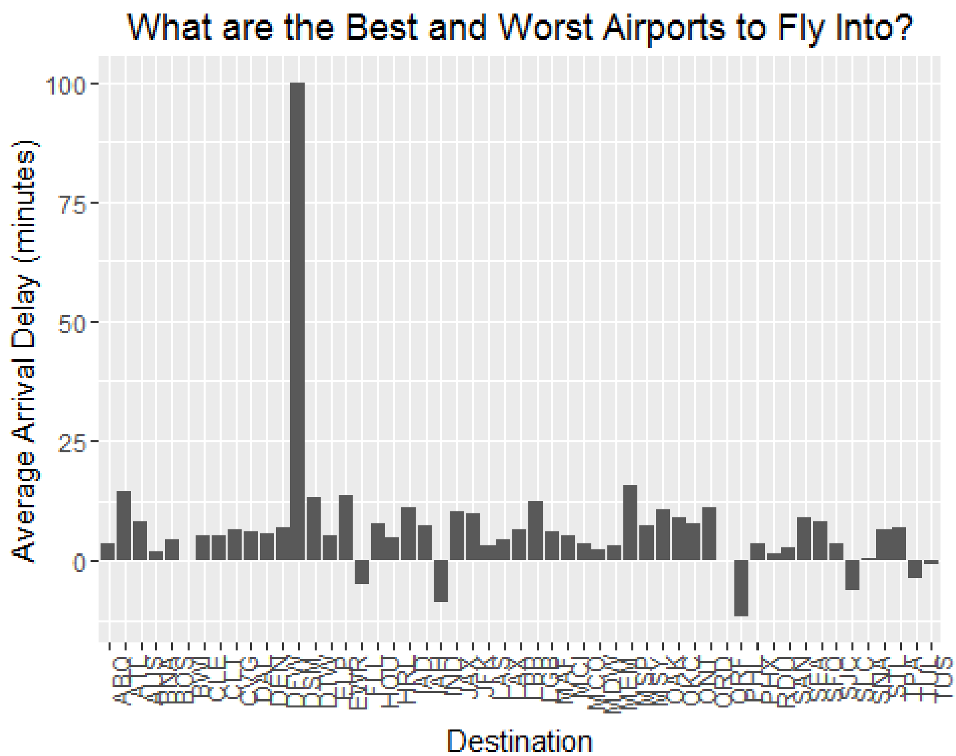



To analyze the best day to fly, arrival delay was plotted against day of the month and day of the week. While there is broad range of delays, the average remains relatively stable around zero minutes for both plots. Tuesday and Saturday tend to have slightly higher delays, with drastic outliers.




Next, the best time of day to fly was researched. Departure time was plotted against Arrival Delay, and a "fanning" effect was noticed as the day went on. This corresponds to previous beliefs that if flights in the morning are delayed, it will create a domino effect on later flights. If hoping to avoid delays, the best time to schedule a flight is between 5 and 8 am.




Finally, what are the best and worst airports to fly into? Using the average arrival delay for each destination, it can clearly be seen that Des Moines, Iowa (DSM) is the airport with the highest average delays. In contrast, flights going into Fort Lauderdale, Florida (FLL), Indianapolis, Indiana (IND), Philadelphia, Pennsylvania (PHL), Salt Lake City, Utah (SLC), and Tulsa, Oklahoma (TUL), on average, arrive ahead of schedule - therefore being considered the best airports to fly into based off of this criteria.

## Author Attribution

## Assignment:

Build two separate models (using any combination of tools) to predicting the author of an article on the basis of that article's textual content.

## Solution:

A Logistic Regression model was applied to the training dataset, with Author as the response variable. Based on the results, using Logistic Regression to predict the Authors in the test set is not reccommended. An acurracy of only **0.0008** was produced. Therefore, Logistic Regression is not a good method for text classificitation for this data set.

**Only 10 Correct Predictions using Naive Baye's** Because only 10 out of 50 authors were predicted correctly, Naive Baye's is not a successful method. The most common predictions were Eric Auchard and Sarah Davidson.

| Training Column/Author | Predicted Column/Author | Correct? |
|---|---|---|
| 1 | 43 | No |
| 2 | 43 | No |
| 3 | 3 | Yes |
| 4 | 14 | No |
| 5 | 25 | No |
| 6 | 10 | No |
| 7 | 25 | No |
| 8 | 50 | No |
| 9 | 42 | No |
| 10 | 10 | Yes |
| 11 | 43 | No |
| 12 | 50 | No |
| 13 | 42 | No |
| 14 | 14 | Yes |
| 15 | 14 | No |
| 16 | 16 | Yes |
| 17 | 25 | No |
| 18 | 25 | No |
| 19 | 43 | No |
| 20 | 25 | No |
| 21 | 43 | No |
| 22 | 25 | No |
| 23 | 23 | Yes |
| 24 | 25 | No |
| 25 | 25 | Yes |
| 26 | 10 | No |
| 27 | 10 | No |
| 28 | 43 | No |
| 29 | 43 | No |
| 30 | 23 | No |
| 31 | 25 | No |
| 32 | 32 | Yes |

| | | |
|---|---|---|
| 33 | 43 | No |
| 34 | 43 | No |
| 35 | 14 | No |
| 36 | 42 | No |
| 37 | 10 | No |
| 38 | 35 | No |
| 39 | 32 | No |
| 40 | 42 | No |
| 41 | 42 | No |
| 42 | 42 | Yes |
| 43 | 43 | Yes |
| 44 | 43 | No |
| 45 | 25 | No |
| 46 | 35 | No |
| 47 | 10 | No |
| 48 | 25 | No |
| 49 | 23 | No |
| 50 | 50 | Yes |

**Key:** 1 - Aaron Pressman, 2 - Alan Crosby, 3 - Alexander Smith, 4 - Benjamin Kang Lim, 5 - Bernard Hickey, 6 - Brad Dorfman, 7 - Darren Schuettler, 8 - David Lawder, 9 - Edna Fernandes, 10 - Eric Auchard, 11 - Fumiko Fujisaki, 12 - Graham Earnshaw, 13 - Heather Scoffield, 14 - Jane Macartney, 15 - Jan Lopatka, 16 - Jim Gilchrist, 17 - Joe Ortiz, 18 - John Mastrini, 19 - Jonathan Birt, 20 - Jo Winterbottom, 21 - Karl Penhaul, 22 - Keith Weir, 23 - Kevin Drawbaugh, 24 - Kevin Morrison, 25 - Kristin Ridley, 26 - Kourosh Karimkhany, 27 - Lydia Zajc, 28 - Lynne O'Donnell, 29 - Lynnley Browning, 30 - Marcell Michelson, 31 - Mark Bendeich, 32 - Martin Wolk, 33 - Matthew Bunce, 34 - Michael Connor, 35 - Mure Dickie, 36 - Nick Louth, 37 - Patricia Commins, 38 - Peter Humphrey, 39 - Pierre Tran, 40 - Robin Sidel, 41 - Roger Fillion, 42 - Samuel Perry, 43 - Sarah Davidson, 44 - Scott Hillis, 45 - Simon Cowell, 46 - Tan Ee Lyn, 47 - Therese Poletti, 48 - Tim Farrand, 49 - Todd Nissen, 50 - William Kazer

## Conclusion:

The Naive Bayes is preferred over Logistic Regression. The Naive Bayes model predicted the author to be Sarah Davidson 10 times incorrectly, suggesting she has a writing style similar to 10 other authors.

# Practice with Association Rule Mining

## Assignment:

Find some interesting association rules for a list of shopping baskets.

## Solution:

Once the data set was in the format expected by the arules package, association rule mining was conducted.

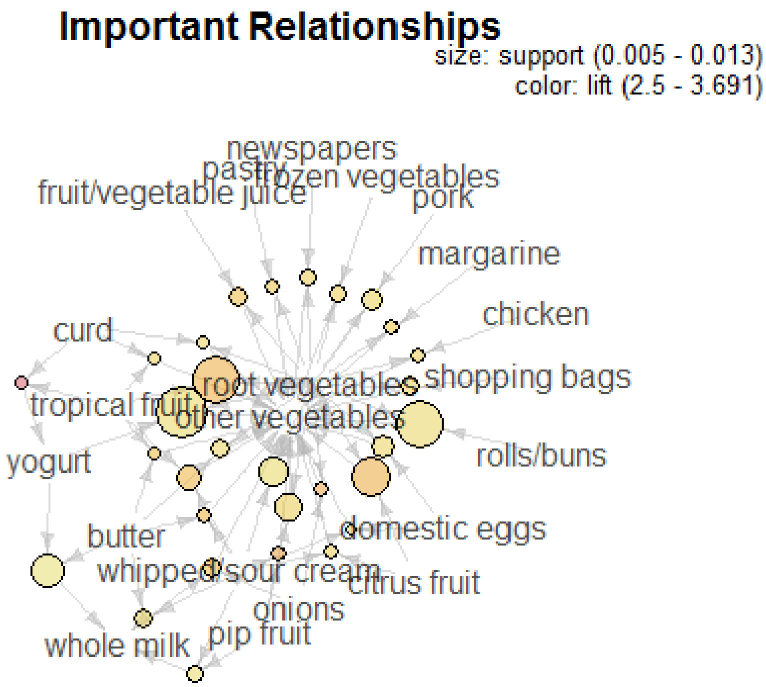
Various threshold levels were tested to see what combination gave the most meaningful results. The maximum number of items was set to 3. Increasing this value above 3 did not change the results significantly holding the other threshold levels constant. Additionally, the confidence threshold was set to 0.5. Setting this value any lower than 50% doesn't provide a trustworthy prediction because you don't have majority confidence in the results. Setting the confidence higher than 50% didn't produce any results, forcing this threshold to remain at 50%. Similarly, the support level was tested at high and low values. Higher values (0.01) produced very few to no results, while low values (0.001) produced far too many results to be meaningful. The support threshold was decided as 0.005.

The final mining resulted in 99 rules. Looking at rules with high lift (or higher odds of containing this subset of items), it can be seen that purchasing fruits and vegetables is likely to result in purchasing the set of items labeled "other vegetables". Additionally, yogurt is 3.69 times more likely to be purchased if fruit and curd is in the shopping basket.

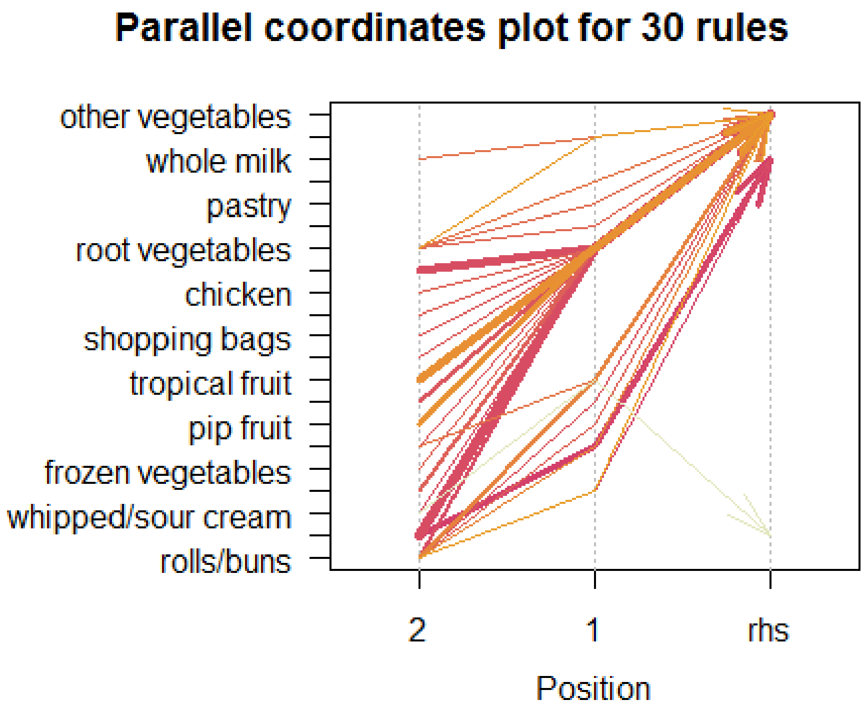
```
##   lhs                     rhs                     support confidence     li
ft
## 1 {onions,
##    root vegetables}    => {other vegetables} 0.005693950  0.6021505 3.1120
08
## 2 {curd,
##    tropical fruit}     => {yogurt}           0.005287239  0.5148515 3.6906
45
## 3 {pip fruit,
##    whipped/sour cream} => {other vegetables} 0.005592272  0.6043956 3.1236
10
## 4 {citrus fruit,
##    root vegetables}    => {other vegetables} 0.010371124  0.5862069 3.0296
08
## 5 {root vegetables,
##    tropical fruit}     => {other vegetables} 0.012302999  0.5845411 3.0209
99
```

There is only one "rule" with confidence higher than 60% and support just 0.01. This rule states that you are 2.5 times more likely to purchase whole milk if you have butter and yogurt in your shopping basket.

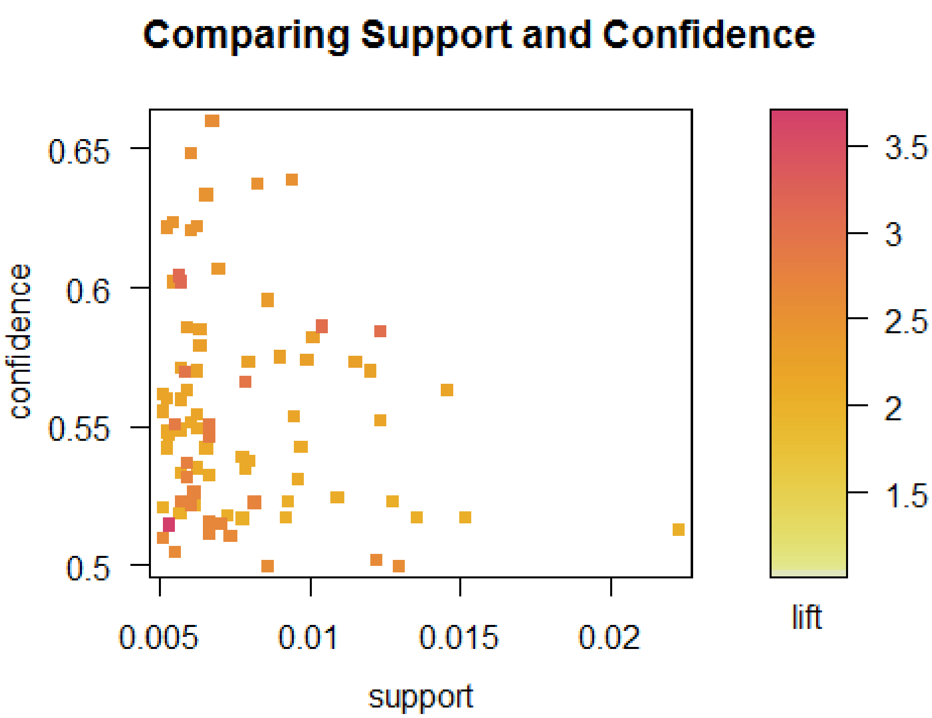
```
##    lhs                rhs          support     confidence lift
## 50 {butter,yogurt} => {whole milk} 0.009354347 0.6388889  2.500387
```




Viewing the plot above, it can be noted that other vegetables and root vegetables provide the highest support levels for the various basket items. Additionally, larger circles represent higher support.




The plot above shows combination of basket items that predict more common resulting items. Most of the lines converge to other vegetables, fruit/vegetable juice, and onions.



**Conclusion:** Comparing support against confidence, there are few rules that have support higher than 0.01. Additionally, most have confidence below 60%. This tells us that our mining rules might not be the most reliable and that we might not be able to make meaningful conclusions about shopping patterns.