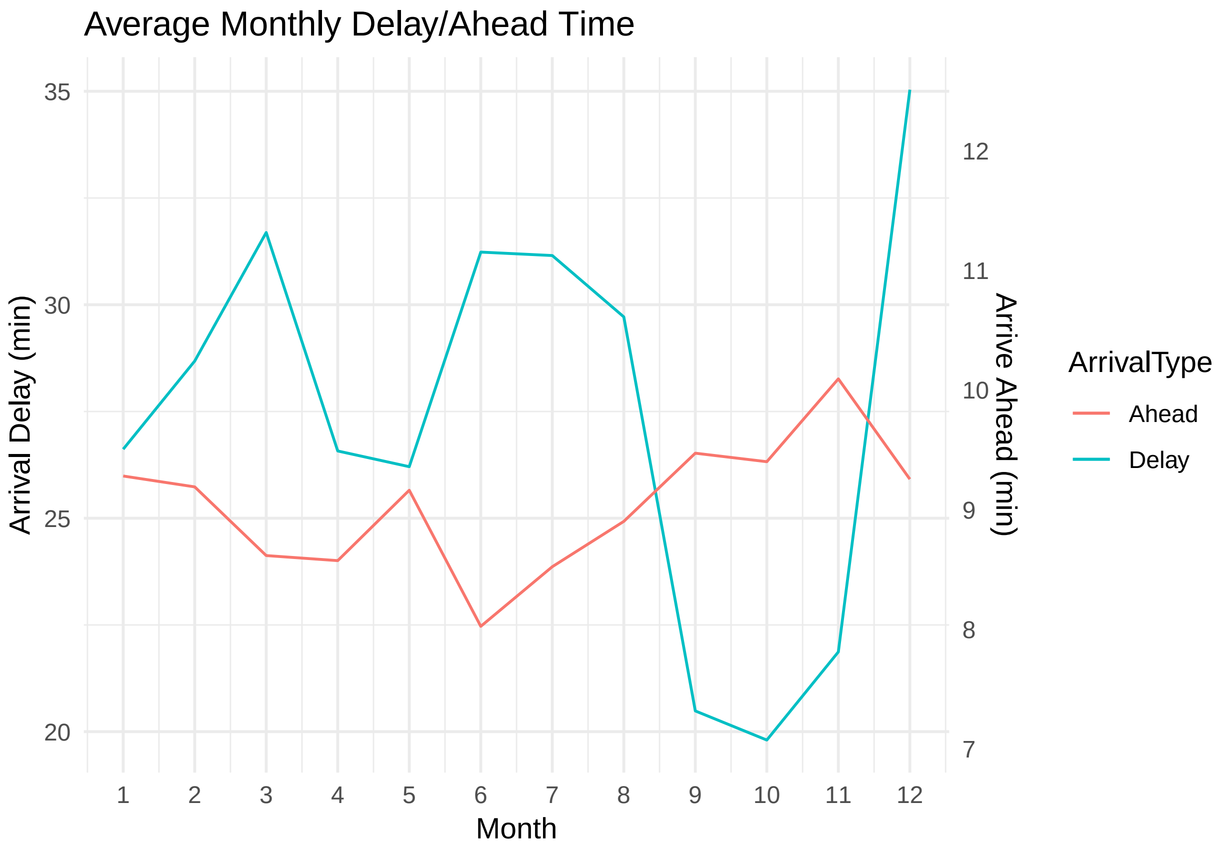
STA 380, Part 2: Exercises 2

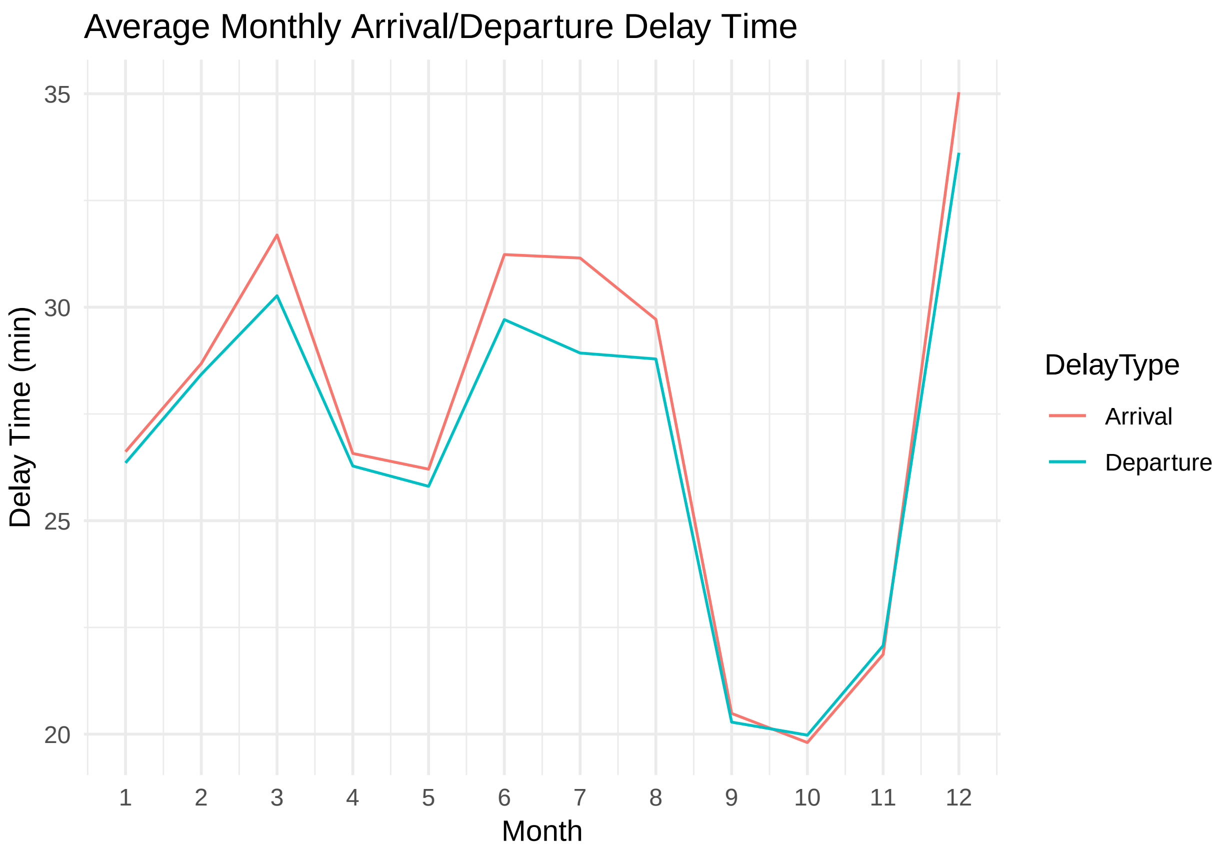
Mengying Yu, Lining Jiang, Shuyuan Sun, Cuiting Zhong

August 19, 2018

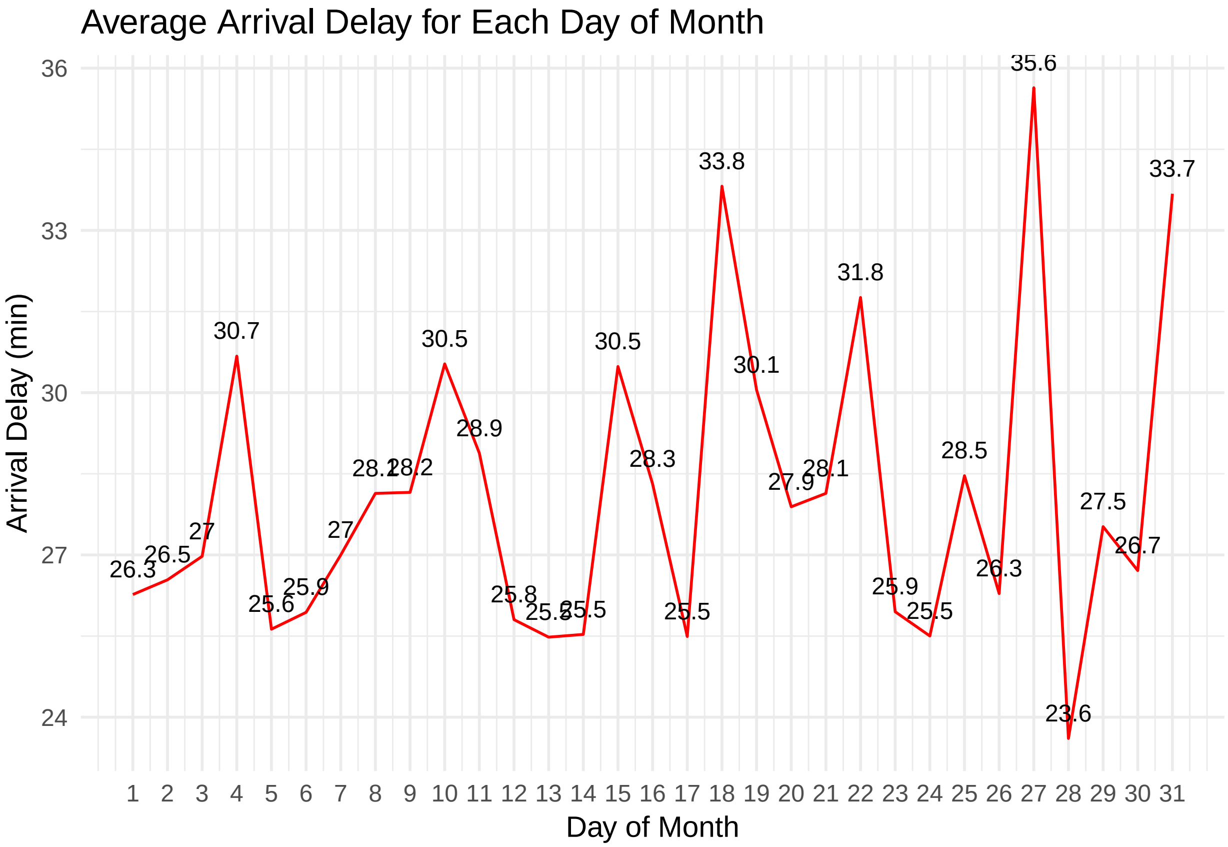
Flights at ABIA



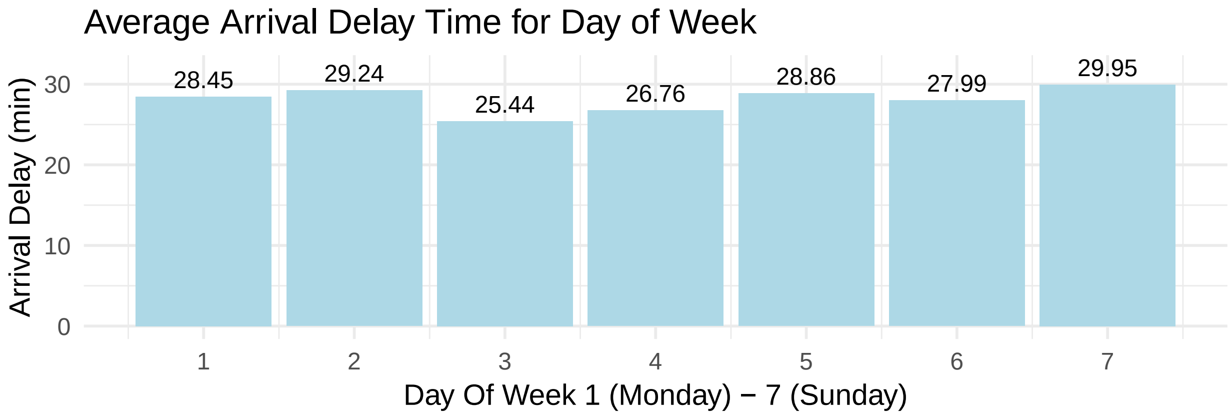
We can see that December is the month with the highest average delay, following by March, June and July. September, October and November seem to be most 'Punctuate'. Typically, the month with a higher average delay tends to have lower average time arrive in advance (but this is not so obvious).

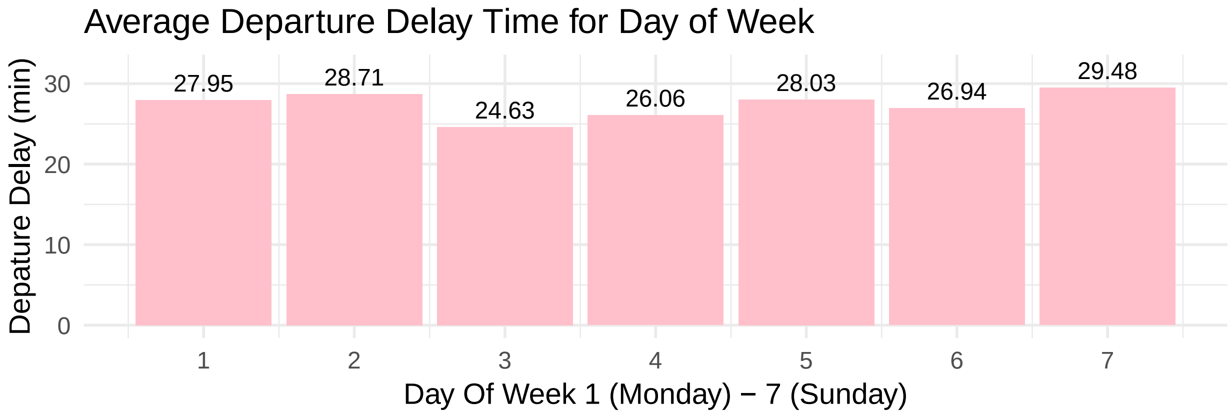


Apparently, months with a higher average arrival delay tend to have a higher average departure delay. This may be because that given flight time period stays the same, departure delays will always lead to arrival delays.

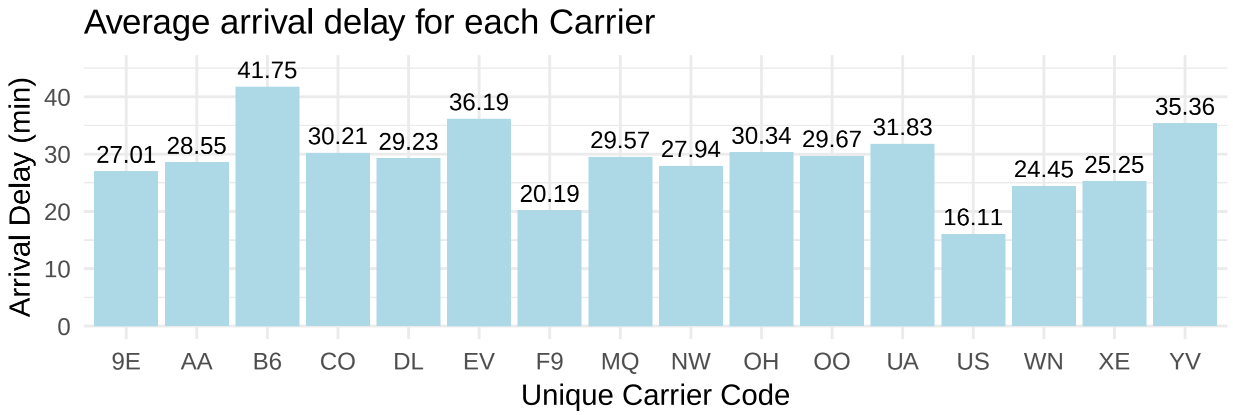


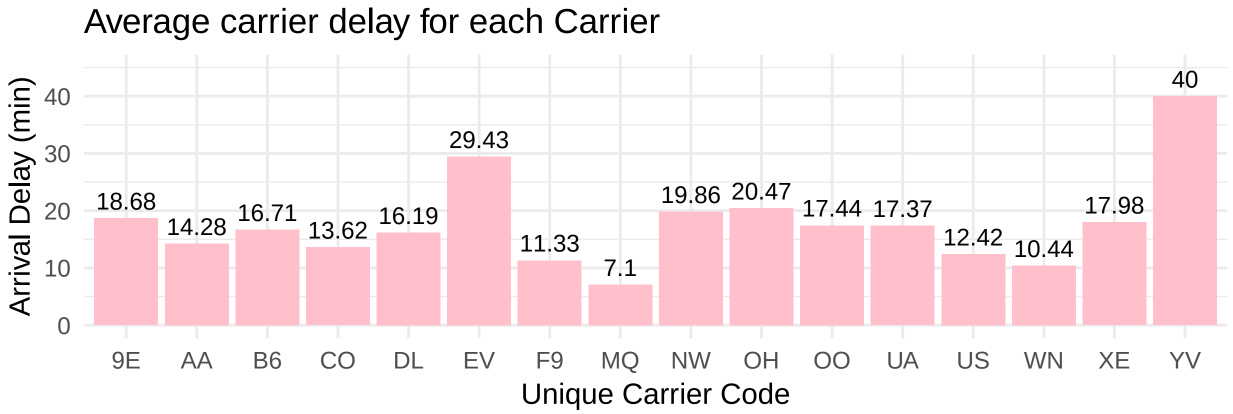
There seems no particular trends for days of month.





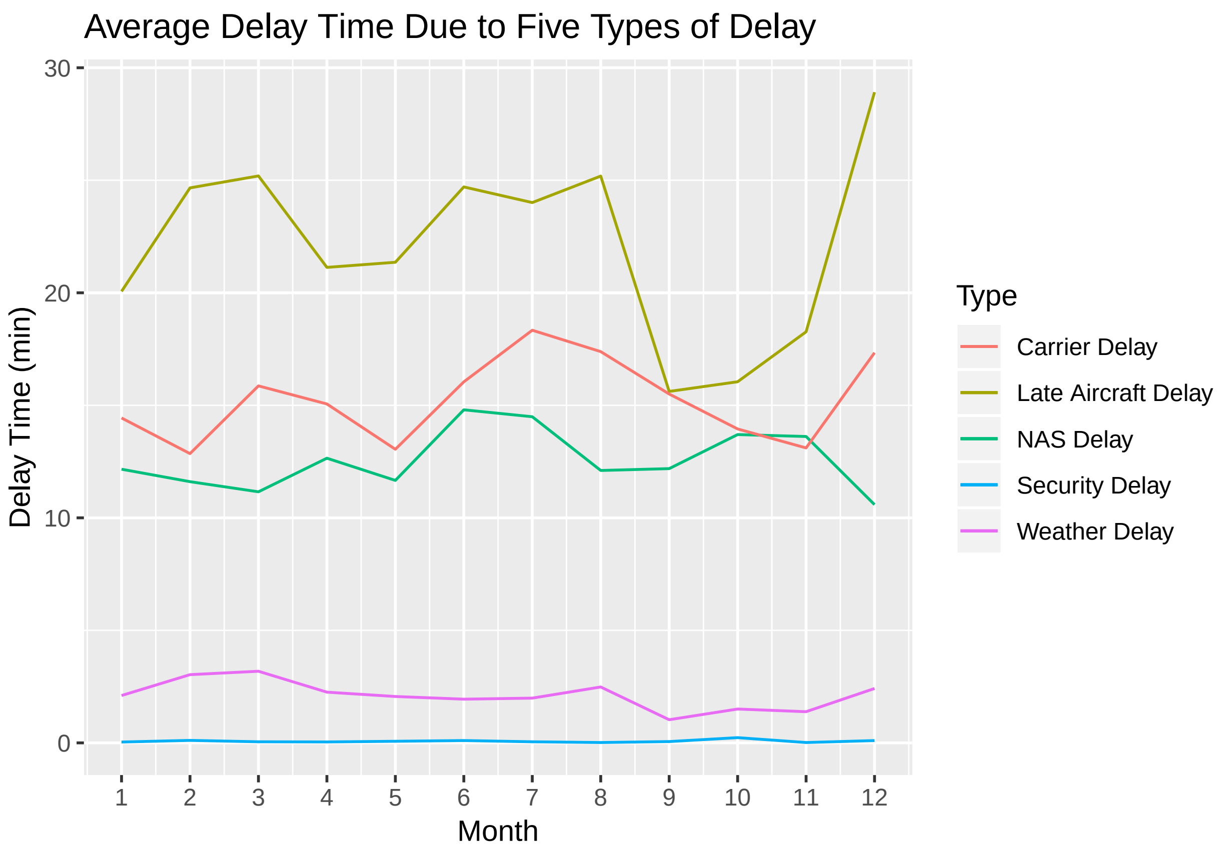
In the scope of week, we can't identify any specific pattern of which day having extremely high or low arrive delay time. However, the trend of these plots are totally identical.





From these plots we can see that the higher average arrival delay for each carrier may not be due to their own fault. Airline YV and EV have both higher average arrival delay and average carrier delay, so it may be unwise to choose from these two carriers.

Based on the five plots above, we can have a general idea of how these five major delay reasons contributed to the overall delay time.



According to this plot, we can see late aircraft delay is much higher than any other delay types except for September and October. Overall, there is one delay time drop on May, and another one on September to November expect for NAS delay.

Author attribution

1. Read in text data:
2. Use Naive Bayes to do classification:

```
##
## p                  AaronPressman AlanCrosby AlexanderSmith BenjaminKangLim
##   AaronPressman               44          0              0               0
##   AlanCrosby                   0         37              0               0
##   AlexanderSmith               0          0              7               0
##   BenjaminKangLim              0          0              0              21
##   BernardHickey                1          0              0               0
##   BradDorfman                  1          0              0               0
##   DarrenSchuettler             0          0              0               0
##   DavidLawder                  0          0              0               0
##   EdnaFernandes                0          0              2               0
##   EricAuchard                  0          0              0               0
##
## p                  BernardHickey BradDorfman DarrenSchuettler DavidLawder
##   AaronPressman                1           0                0           0
##   AlanCrosby                   0           0                0           0
##   AlexanderSmith               0           0                0           0
##   BenjaminKangLim              0           0                0           0
##   BernardHickey               37           0                0           0
##   BradDorfman                  0          41                0           6
##   DarrenSchuettler             0           0               10           0
##   DavidLawder                  0           0                0           7
##   EdnaFernandes                1           0                1           1
##   EricAuchard                  0           0                0           0
##
## p                  EdnaFernandes EricAuchard
##   AaronPressman                0           0
##   AlanCrosby                   0           0
##   AlexanderSmith               0           0
##   BenjaminKangLim              0           0
##   BernardHickey                0           0
##   BradDorfman                  0           1
##   DarrenSchuettler             0           0
##   DavidLawder                  0           0
##   EdnaFernandes               26           0
##   EricAuchard                  0          22
```

```
## [1] 0.6408
```

We try to use term frequency of the training set to built a Naive Bayes classifier and then make predictions for the testing set. The accuracy of Naive Bayes classifier is 64.08%.

3. Combination of glmnet model and PCA:

```
## [1] 0.6048
```

First, perform PCA on the training set and extract the first princial components to fit glmnet models with 100 different lambda. Then we calculate the first 100 PC scores of the test set, and use those fitted models to do classification. We find that the highest accuracy we can get from a glmnet model is 60.48%.

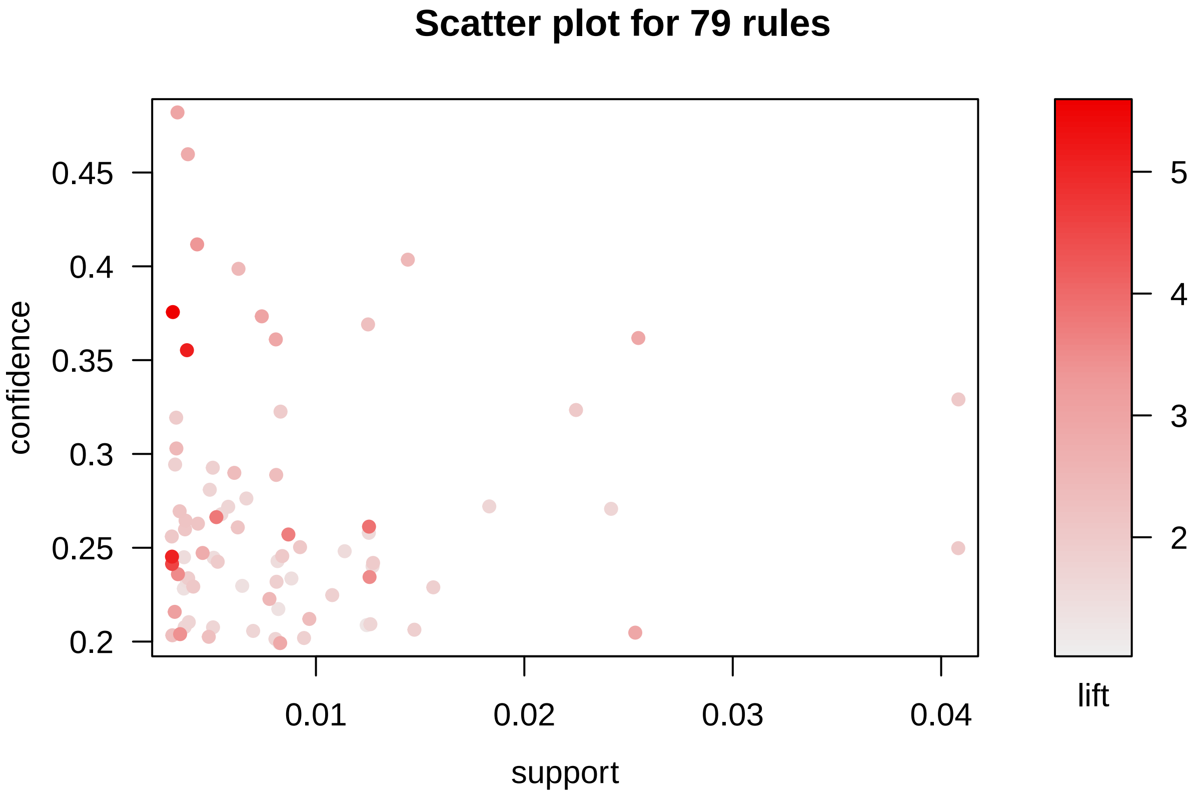
4. Similarity:

```
##             authors
## 1     AaronPressman
## 2        AlanCrosby
## 3    AlexanderSmith
## 4   BenjaminKangLim
## 5     BernardHickey
## 6       BradDorfman
## 7  DarrenSchuettler
## 8       DavidLawder
## 9     EdnaFernandes
## 10      EricAuchard
##                                                               similar
## 1
## 2                                         JanLopatka , JoWinterbottom
## 3                                                        JonathanBirt
## 4                             KarlPenhaul , ScottHillis , WilliamKazer
## 5                                                       KevinMorrison
## 6                                         KevinDrawbaugh , RobinSidel
## 7                                     HeatherScoffield , MarkBendeich
## 8                            BradDorfman , KevinDrawbaugh , ToddNissen
## 9                                           JimGilchrist , TimFarrand
## 10 KevinDrawbaugh , KouroshKarimkhany , SamuelPerry , TheresePoletti
```

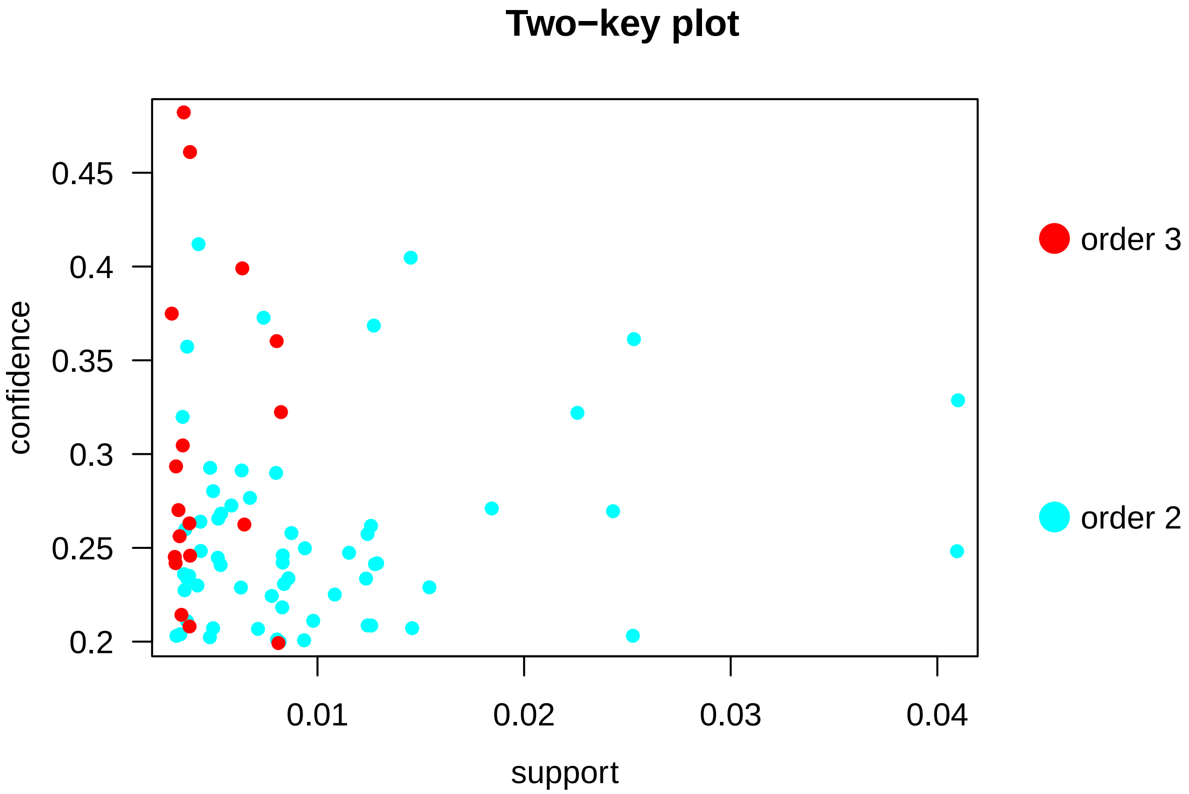
In order to find the sets of authors whose articles seem difficult to distinish from one another, we used the confusion table result of Naive Bayes Model. Then, we found the index of entries which the model predicted wrong more than twice. Finally, we constucted a table to conclude similar authors for each author. For example, If the origin author is AlanCrosby, the similar authors are JanLopatka and JoWinterbottom. The above table only shows 10 authors and their similar authors.

Practice with association rule mining

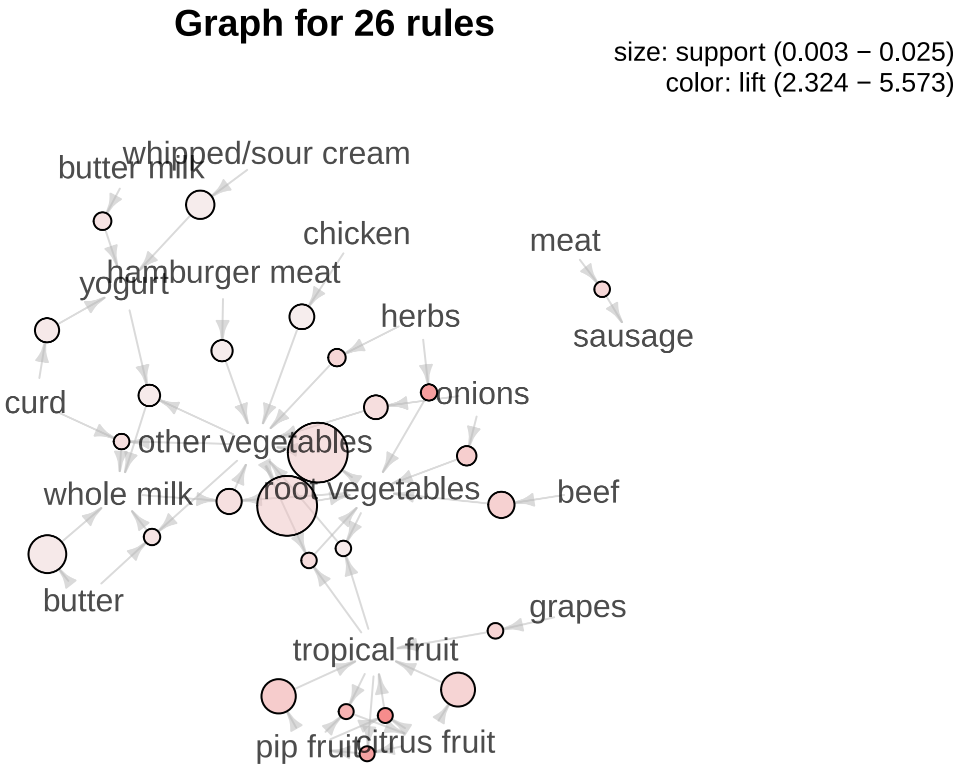
```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##         0.2    0.1    1 none FALSE            TRUE       5   0.003      1
##  maxlen target   ext
##       5  rules FALSE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 45
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[169 item(s), 15296 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 done [0.00s].
## writing ... [79 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```



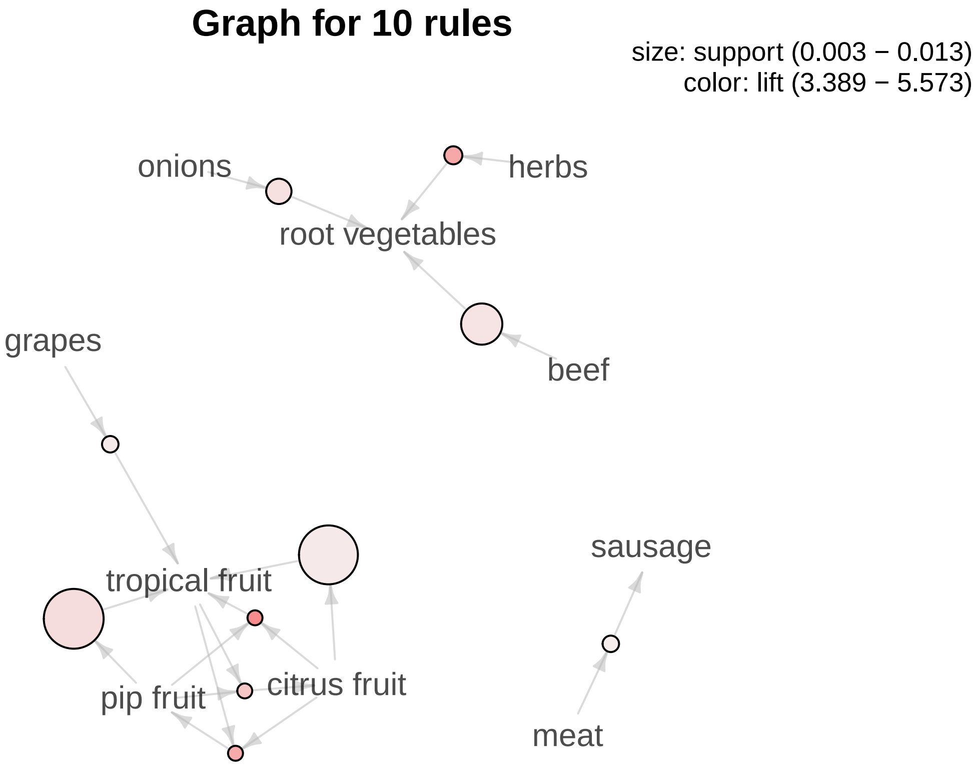
We examined several pairs of supports and confidences until we found sufficient number of rules. Since support stands for how popular a basket is, as measured by the proportion of transactions in which a basket appears; confidence says how likely grocery A is purchased when grocery B is purchased; lift shows how likely grocery A is purchased when grocery B is purchased, while controlling for how popular B is. Therefore, grocery with higher lift tends to have lower support.



Apparently, rule bodies with only two groceries tend to have higher support (frequency) than those with three groceries. This is obvious because itemset of only two can also appear in itemset of three.



We noticed something interesting from this plot. For example, the rules among tropical fruit, pip fruit and citrus fruit have higher lifts indicating that there exists a strong association among these fruits. In contrast, rules between root vegetables and other variables has higher support but lower lift. This implies a high frequency of this itemset but weak association between items in this itemset.



This plot selected top 10 rules based on the value of lifts. These rules show the strong associations between meat to sausage, onions to root vegetables, beef to root vegetables, herbs to vegetables, grapes to tropical fruit, pip fruit between citrus fruit and tropical fruit.