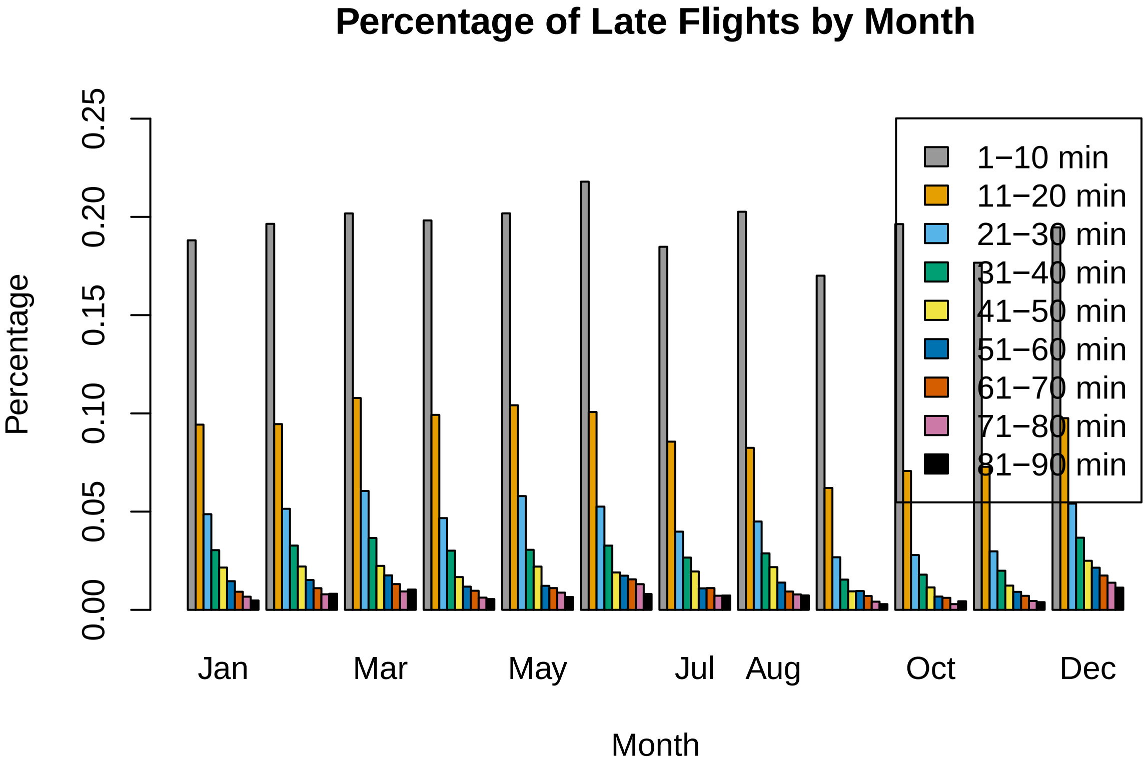
Statistics 380 Homework 2

Valerie Roth

Problem 1: Flights at ABIA

Introduction: I decided to answer the question "What is the best time of year to fly to minimize delays?" To do this, I created a plot that shows the percentage of flights that were late each month, categorized by how late they were. I did not include flights that were over 90 minutes late, because they were rare and I did not expect their occurrences to be statistically significant. I also only looked at arrival lateness, because this is what I expect matters most to passengers.



Discussion: I found that September had a particularly low percentage of late flights in all categories. This may be explained by the fact that September does not usually have weather that would slow a plane's arrival down. It is also a month without any major breaks from school or holidays, which would result in fewer passengers at the airport overall. Perhaps when the airport is less crowded fewer flights are late.

Alternatively, I found that June had a large percentage of flights that were 1-10 minutes late. Perhaps this is due to a more crowded airport. I would expect crowding to make flights slightly later, but not by a large amount of time, so these results make sense.

I also found that December had a large percentage of very late flights. I expect that this is because December has particularly bad weather and a large number of travelers (people flying home for the holidays). I expect that bad weather could significantly slow a plane down, so these results also seem to make sense.

Nonetheless, there is not a huge amount of variation in lateness among the months of the year. Therefore, if I were to advise a passenger about the best time of year to fly based on their odds of making it to their destination, I would suggest that they pick a date that appealed to them for other reasons. The slight increase in lateness in certain months is simply not significant enough to plan a trip around.

Problem 2: Author Attribution

Naive Bayes:

```
## <<DocumentTermMatrix (documents: 2500, terms: 1389)>>
## Non-/sparse entries: 246803/3225697
## Sparsity           : 93%
## Maximal term length: 18
## Weighting          : term frequency (tf)
```

```
## [1] "Testing document 49..."
```

```
## [1] "Author 1 (Aaron Pressman): -1142.624"
```

```
## [1] "Author 2 (Alan Crosby): -1167.921"
```

```
## [1] "Author 3 (Alexander Smith): -1040.603"
```

```
## [1] "Author 4 (Benjamin Kang Lim): -1200.099"
```

```
## [1] "Author 5 (BernardHickey): -1046.860"
```

```
## [1] "Author 6 (BradDorfman): -976.723"
```

```
## [1] "Author 7 (Darren Schuettler): -995.090"
```

```
## [1] "Author 8 (David Lawder): -1155.157"
```

```
## [1] "Author 9 (Edna Fernandes): -1066.853"
```

```
## [1] "Author 10 (Eric Auchard): -1036.474"
```

```
## [1] "Author 11 (Fumiko Fujisaki): -1089.143"
```

```
## [1] "Author 12 (Graham Earnshaw): -1120.176"
```

```
## [1] "Author 13 (Heather Scoffield): -1122.969"
```

```
## [1] "Author 14 (Jane Macartney): -1135.441"
```

```
## [1] "Author 15 (Jan Lopatka): -1145.447"
```

```
## [1] "Author 16 (Jim Gilchrist): -1205.779"
```

```
## [1] "Author 17 (Joe Ortiz): -1122.668"
```

```
## [1] "Author 18 (John Mastrini): -1064.830"
```

```
## [1] "Author 19 (Johnathan Birt): -2348.817"
## [1] "Author 20 (Jo Winterbottom): -1109.705"

## [1] "Author 21 (Karl Penhaul): -1121.449"

## [1] "Author 22 (Keith Weir): -1073.487"

## [1] "Author 23 (Kevin Drawbaugh): -1019.371"

## [1] "Author 24 (Kevin Morrison): -1078.595"

## [1] "Author 25 (Kirstin Ridley): -1090.469"

## [1] "Author 26 (Kourosh Karimkhany): -991.275"

## [1] "Author 27 (Lydia Zajc): -1084.357"

## [1] "Author 28 (Lynne O'Donnel): -1191.685"

## [1] "Author 29 (Lynnley Browning): -1190.729"

## [1] "Author 30 (Marcel Michelson): -1141.159"

## [1] "Author 31 (Mark Bendeich): -1059.597"

## [1] "Author 32 (Martin Wolk): -1060.731"

## [1] "Author 33 (Matthew Bunce): -1150.590"

## [1] "Author 34 (Michael Connor): -1098.207"

## [1] "Author 35 (Mure Dickie): -1092.873"

## [1] "Author 36 (Nick Louth): -1020.162"

## [1] "Author 37 (Patricia Commins): -1072.570"

## [1] "Author 38 (Peter Humphrey): -1328.943"

## [1] "Author 39 (Pierre Tran): -1120.085"

## [1] "Author 40 (Robin Sidel): -1176.775"

## [1] "Author 41 (Roger Fillion): -1157.121"

## [1] "Author 42 (Samuel Perry): -1036.752"

## [1] "Author 43 (Sarah Davison): -1049.258"
## [1] "Author 44 (Scott Hillis): -1182.743"

## [1] "Author 45 (Simon Cowell): -1160.370"

## [1] "Author 46 (Tan Ee Lyn): -1219.564"

## [1] "Author 47 (Therese Poletti): -1078.543"

## [1] "Author 48 (Tim Farrand): -1055.048"

## [1] "Author 49 (Todd Nissen): -1053.901"

## [1] "Author 50 (William Kazer): -1149.697"
```

Linear Regression

Discussion: I used a naive bayes and a linear regression model. Features for both models were simply the words (besides stopwords) found in the documents.

Authors that the algorithm struggles to distinguish from each other are ones with similar values in the naive bayes algorithm. We can see that there are some examples of this. For instance, author 28 (Lynne O'Donnel) and author 29 (Lynnley Browning) must be similar, having values of -1191.685 and -1190.729, respectively.

I prefer naive bayes because I can immediately see which authors are similar. This similarity can help explain why a model would not predict certain articles well if there is another author that a document is very likely to be attributed to as well. Linear regression simply tells us which terms are important for prediction. This is neat, because it tells us which terms may be distinctive of some particular writers, however it does not allow us to attribute that style to a particular person as easily.

Since I had so many words (predictors) to look at, a summary of mylinear model is not shown. It is commented out, so the grader can comment it back in if they choose. Some particularly significant predictors include the word "users" (perhaps there are some writers distinguishable by the fact that they write about tech), thought, produce (economics writers, perhaps?), operations, online, link, japans, european, etc. This seems to indicate that the authors may have different focuses which helps to distinguish them.

Problem 3: Practice with Association Rule Mining

```
##   lhs                  rhs                   support confidence     lift
## 1 {citrus fruit,
##    root vegetables} => {other vegetables} 0.01037112  0.5862069 3.029608
## 2 {root vegetables,
##    tropical fruit}  => {other vegetables} 0.01230300  0.5845411 3.020999
```

```
##   lhs                  rhs                   support confidence     lift
## 1 {curd,
##    yogurt}          => {whole milk}       0.01006609  0.5823529 2.279125
## 2 {citrus fruit,
##    root vegetables} => {other vegetables} 0.01037112  0.5862069 3.029608
## 3 {root vegetables,
##    tropical fruit}  => {other vegetables} 0.01230300  0.5845411 3.020999
##   lhs                    rhs                  support confidence     lift
## 1 {domestic eggs,
##    other vegetables} => {whole milk}       0.01230300  0.5525114 2.162336
## 2 {root vegetables,
##    tropical fruit}   => {other vegetables} 0.01230300  0.5845411 3.020999
## 3 {root vegetables,
##    yogurt}           => {whole milk}       0.01453991  0.5629921 2.203354
```

Discoveries

- By looking at lift > 3, I found that people who purchased citrus fruit and root vegetables are approximately 3 times more likely to have purchased other vegetables as well. The same goes for people who have purchased root vegetables, and tropical fruit.
- By looking at confidence > .58, I found that whole milk appears in baskets containing curd and yogurt approximately 58.2% of the time. I also found that "other vegetables" appear in baskets with citrus fruit and root vegetables approximately 58.6% of the time. Additionally, people who purchase root vegetables and tropical fruit purchase other vegetables approximately 58.5% of the time.
- By looking at support > .012 & confidence > .55, I found that overall approximately 1.2% of purchases include domestic eggs, other vegetables and whole milk (can be seen by only looking at support). I found that whole milk appears in baskets with domestic eggs and other vegetables approximately 55% of the time. I also found that about 1.2% of purchases include root vegetables, tropical fruit, and other vegetables. I found that other vegetables appear in baskets with root vegetables and tropical fruit approximately 58% of the time. I also found that approximately 1.5% of purchases include root vegetables, yogurt and whole milk. Whole milk appears in purchases with root vegetables and yogurt approximately 56% of the time.