STA 380, Part 2: Exercises 2

Turn these in by 12 PM on Tuesday, August 16th. As before, prepare your reports using RMarkdown so that they are fully reproducible, carefully integrating visual and quantitative evidence with prose.

As before, you should submit your work to statdropbox@gmail.com. Send a link to a GitHub page where the final report has been stored – preferably in Markdown (.md) format but PDF is OK too. Also include a link to the raw .Rmd file that can be used to reproduce your report from scratch. *Do not send a PDF or Rmd file as an attachment.*

Important: Use the subject line "STA 380 Homework 2: Name" so that I can sort my inbox easily. (Obviously use your own first and last names in the subject, or the names of all your group members if applicable.)

Flights at ABIA

Consider the data in ABIA.csv, which contains information on every commercial flight in 2008 that either departed from or landed at Austin-Bergstrom Interational Airport. The variable codebook is as follows:

- Year all 2008

- Month 1-12

- DayofMonth 1-31
- DayOfWeek 1 (Monday) - 7 (Sunday)
- DepTime actual departure time (local, hhmm)
- CRSDepTime scheduled departure time (local, hhmm)
- ArrTime actual arrival time (local, hhmm)
- CRSArrTime scheduled arrival time (local, hhmm)
- UniqueCarrier unique carrier code
- FlightNum flight number
- TailNum plane tail number
- ActualElapsedTime in minutes
- CRSElapsedTime in minutes
- AirTime in minutes
- ArrDelay arrival delay, in minutes
- DepDelay departure delay, in minutes
- Origin origin IATA airport code
- Dest destination IATA airport code
- Distance in miles
- TaxiIn taxi in time, in minutes
- TaxiOut taxi out time in minutes
- Cancelled was the flight cancelled?
- CancellationCode reason for cancellation (A = carrier, B = weather, C = NAS, D = security)
- Diverted 1 = yes, 0 = no
- CarrierDelay in minutes
- WeatherDelay in minutes
- NASDelay in minutes
- SecurityDelay in minutes

- LateAircraftDelay in minutes

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. For example, you might consider one of the following questions:

- What is the best time of day to fly to minimize delays?
- What is the best time of year to fly to minimize delays?
- How do patterns of flights to different destinations or parts of the country change over the course of the year?
- What are the bad airports to fly to?

But anything interesting will fly.

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?

Note: you will need to figure out a way to deal with words in the test set that you never saw in the training set. This is a nontrivial aspect of the modeling exercise.

Practice with association rule mining

Revisit the notes on association rule mining, and walk through the R example on music playlists: playlists.R and playlists.csv. Then 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.