Christine Luongo

General Assembly: Data Science

Project Problem Statement

**Milestone 1:**

Transportation has always been a personal interest of mine and fortunately through my consultant work and a little of my classwork I have been able to come into some contact with transportation analyses. Commercial flights are used occasionally for removal of illegal aliens, however the capabilities of our Analysis team at Immigration and Customs Enforcement are limited. Furthermore, of the mode of transportation air transportation has better tractability. Due to the Law Enforcement Sensitivity our data, I cannot use any data from my consultant work.

The Research and Innovative Technology Administration (RITA) provides detailed data, both on summary statistics and down to the granularity of every US domestic flight.[[1]](#footnote-1) Information is provided pertaining to average fare costs,[[2]](#footnote-2) carrier market share, and on-time performance[[3]](#footnote-3). This data goes back several years but due to the size, only focusing on a single year should suffice.

Due to the size and scope of this data, an exploratory analysis and the tools developed in the course may result in answering the following:

* What factors determine or influence the on-time performance of a flight?
  + Potential variates include, but are not limited to, carriers, planes, planned duration of flight, airports, airport-carrier combinations, weather, time-of-year, time-of-day etc.
  + Using Tail Numbers and sorting by date, the data can be manipulated to see the effect of delay over an entire mission (flight loops). A delay in the first leg of a mission could have a cascading effect on the remainder of the legs. This will also be a good exercise in data manipulation and reshaping.
  + Do commuter routes have better on-time performance?
  + Is there any relationship between average fare and on-time performance?
  + Is there any way to cluster flight paths? Can we see if any characteristics are shared amount flights with similar outcomes?
  + Can we cluster airports or airport-path combinations and see which airports are most similar in characteristics and in on-time performance?
* Due to the sheer size of the dataset, it might be wise to subset the scope to only a few major airports: can we compare the operations of certain hub airports (perhaps the busiest ones) and determine which ones have smoothest operations and why?
* For my own personal interest, I’d like to explore some other questions for my travel. When’s the best time to fly around Christmas or Thanksgiving – in terms of best on-time performance.

**Milestone 2:**

For the beginning portion of my data exploration and analysis I’ve dived into RITA’s on-time performance of all flights. These files are pretty large, so for now I’ve been investigating a single month’s worth of data.[[4]](#footnote-4) This level of granularity is probably unnecessary, so firstly I need to aggregate this data.

**The data cleaning processes includes:**

1. Getting rid of extraneous column fields. Many of the fields are redundant.
2. Many flight numbers are flown almost daily (or at some frequency), we need only to keep track of the summary statistics of each flight – perhaps a five number summary of each flight (min, 25% percentile, median, 75% percentile, max) along with mean and count.
   1. If we need to, if a certain flight or route spikes some interest then we can dive deeper into certain flights routes.
3. After the file has been aggregated – we can do some more summary statistics.

For the most part the data seems pretty clean right now, not a lot of missing or nonsensical values.

**Interesting Deep Dive:**

Using the tail number and the flight number I would like to perform some linking logic between flight legs. At this point I’m not sure how to link these flights in Python (I’ve probably come up with logic in other languages) but aside from that I’m not sure what analyses I can do with this linkage. I could potentially create an variable to designate which leg this is in a flight plan and how this leg-variable correlates to delay time.

**Analyses Already Conducted:**

1. Looking at arrival delay for now, but eventually should be extended to departure delay.
2. Summary statistics, found the top twenty-five origin airports.
3. I might want to subset the dataset into only those flights that have a delay time – many flights have negative delay time, indicative of an early arrival.
4. Plotted some visualizations of delay’s relationship with other variates.
5. Describe some statistics of each delay’s relationship to carriers. Surprisingly I did not recognize a lot of carriers’ names.
6. Performed ordinary linear regression to see any relationships between variates and delay. Now I’m seeing that any regression gives very poor adjusted R squared, indicating that there may be other variates not included that better describe delays. I am somewhat concerned that the dataset wouldn’t be able to provide an interesting story.
7. Delay’s relationship to time (time of day, time of week) via ordinary linear regression and visualizations.
8. **Analysis to be conducted**: Investigate the delay cause time break out.

**Discerning Model-Type:**

Initially I was concerned due to the low rate of cancellation or diverted flights of the entire sample of flights. Will I be able to produce meaningful results if our sample of cancellations is so small in comparison the rest of the sample? Hopefully I have some flexibility in this analysis – I can use ordinal response by either bucketing, or binning, the delay time – perhaps in 15 minute increments. Or I can use binary responses since both cancelled and diverted fields are binary responses we can do KNN. In addition to these issues, I find that many of the variates seem to not be significant predictors. Either cancellation or diversion is more random than I thought or there are other variables I’m not including that influence that determination.

1. transtats.bts.gov [↑](#footnote-ref-1)
2. Tables of summary statistics are available at <http://www.rita.dot.gov/bts/airfares> along with more detailed tables: http://www.dot.gov/policy/aviation-policy/domestic-airline-fares-consumer-report [↑](#footnote-ref-2)
3. RITA’s website offers the ability to pull queries from their online database: http://www.transtats.bts.gov/DL\_SelectFields.asp?Table\_ID=236&DB\_Short\_Name=On-Time [↑](#footnote-ref-3)
4. For the time being I’ve been looking at January and February 2014. Aside from the conjecture that there may be more cancellations due to weather in the winter months, this decision was completely arbitrary, however the data is as up-to-date as August 2014 and goes back years. [↑](#footnote-ref-4)