Christine Luongo

November 25, 2014

**Predicting Domestic Airline On-Time Performance**

1. Problem Statement and Hypothesis

In 2013, the average delay time for a flight was only 12 minutes. Given the volume of the number of flights that are flown each day however the total flight delay time added up to well over 13 million hours. Given that there are only 8,760 hours in a year – this statistic is particularly astounding. The Research and Innovative Technology Administration (RITA) provides detailed data ranging from high-level summary statistics to granular-level US domestic flight data points. Information is provided pertaining to average fare costs,[[1]](#footnote-1) carrier market share, and on-time performance[[2]](#footnote-2). The objective of this project is to survey flight information and determine which factors significantly influence flight delay and to further build upon this are there any possible identifiable solutions to ameliorate flight delay, particularly for the consumer.

1. Description of Dataset and how it was obtained

RITA works closing with the Department of Transportation (DOT) to document, research, and analyze the United States’ transportation systems. Among data pertaining to freight and passenger train, vehicular, border crossing, and international trade, RITA publishes data on on-time performance and documents the causes of airline delays.

The **On-Time Performance Database** reports on those carriers that comprise at least one percent of total domestic passenger-airline revenue on a monthly basis. A flight is considered delayed upon arrival if it arrives at its gate least 15 minutes later than its scheduled time of arrival. There is also a departure delay metric, however for the purposes of this analysis, arrival delay is studied.[[3]](#footnote-3)

RITA’s data goes back to 1987 and is as up-to-date as September of 2014. Due to the size of the data given, currently the project need only focus on a few months. In calendar year 2013 alone, there were over six million scheduled domestic flights. However, at this point in the project, I have considered only the months of November 2013 to February 2014. Winter months were chosen since it might make sense there would be greater delays at this season, and therefore tell a more interesting story. Each month can be pulled individually form RITA’s website.[[4]](#footnote-4) An abbreviated list of the fields selected is given in Table 1. A complete list is provided in the appendix.

Table 1. RITA On-Time Performance Data Dictionary

|  |  |
| --- | --- |
| Field | Description |
| FL Date | The date of the flight |
| Unique Carrier | The carrier of the flight. A crosswalk of these carriers is provided. |
| Tail Number | Aircraft registration number. Useful for tracking individual airplane movement. |
| Flight Number | A one to four character alpha-numeric code for a particular flight. |
| Origin/ Dest | The originating airport code of the flight. A crosswalk is provided of International Air Transport Association (IATA) airport codes. |
| Dep\_Time | The departure time of the flight, given in local time. Format is hhmm. |
| Dep\_Delay | Delay time of the departure given in minutes. |
| Arr\_Time | The arrival time of the flight, given in local time. Format is hhmm. |
| Arr\_Delay | Delay time of the arrival given in minutes. |
| Actual\_Elaped\_Time | The elapsed time of the flight, given in minutes |
| Air\_Time | Flight time in air, given in minutes. |
| Distance | Distance between airports, given in miles. |

Though not currently utilized in the analysis, airlines report the following flight delay reasons: air carrier, extreme weather, national aviation system, late-arriving aircraft, and security. This reporting is not mandated for all airports, as evident by the dataset which reports mostly null values for these columns. Twenty-nine airports that account for at least one percent of all enplanements must reports on these flight delay causes.

In addition to flight performance information, geographical airport information was collected through openflights.org, a tool that allows users to map and analyze airline information. The website provides geographical information (latitude and longitude) of airports.[[5]](#footnote-5) The following relevant fields in the dataset are provided in the table below:

Table 2. OpenFlights.org Airport Database

|  |  |
| --- | --- |
| Field | Description |
| Name | Name of the airport |
| City | Main city served by airport. |
| Country | Country of airport |
| IATA/FAA | IATA code used to identify the airport. This code is used to crosswalk with other on-time performance dataset.[[6]](#footnote-6) |
| Latitude | Decimal degrees, usually to six significant digits. Negative is South, positive is North. |
| Longitude | Decimal degrees, usually to six significant digits. Negative is West, positive is East. |
| Timezone | Hours offset from Coordinated Universal Time (UTC). |
| Tz | Another description of time zone, given by locations within that time zone. For example Eastern Standard Time is denoted by “America/New York”; Central Standard Time is denoted by “America/Chicago”; Mountain Standard Time is denoted by “America/Denver”; Pacific Standard Time by “America/Los Angeles” |

Geospatial coordinates and timezones provided by this dataset are used in data analysis discussed in future sections of the project.

The final portion of data merged with the On-Time Performance Dataset comes from the Iowa Environmental Mesonet (IEM). The IEM collects environmental data from Automated Surface Observing System (ASOS) units.[[7]](#footnote-7) There are more than 900 ASOS units in the United States that collect weather information hourly, and possibly more frequently for special observations.[[8]](#footnote-8) Query pulls based on stations with sensors and date ranges going back as far as 1995 are publically available at the IEM website. The data provides temperature, dew point, humidity, wind and precipitations metrics hourly. The website provides a python script to generate automatic pulls, for multiple timeframes and ASOS stations. I took the python script and amended it slightly to pull from those only the airports applicable to this analysis and the fields that seemed useful and printed out to a single txt file that was later manually manipulated in Microsoft Excel.

Table . Fields in Weather Dataset

|  |  |
| --- | --- |
| Field | Description |
| station | Three/four character site identifier |
| valid | Timestamp of the observation |
| tmpf | Air temperature in Fahrenheit |
| drct | Wind direction in degrees from north |
| p0li | One hour precipitation (inches) |
| vsby | Visibility in miles |

Once I had this dataset, some data cleaning was required. The timestamp had to be manipulated to so that it could be matched up to the On-Time Performance Dataset. With the timestamp delimited to Month, Day, Hour in conjunction with the airport code, the weather information for every single arrival and departure can be combined with the original dataset.

1. Description of the On-Time Performance Dataset

RITA provides fairly clean data although there are NaN values for those flights that are not originating or heading toward the top 29 airports. Data manipulation is described in the data exploration section below. Within python, I have appended different months of on-time performance databases.

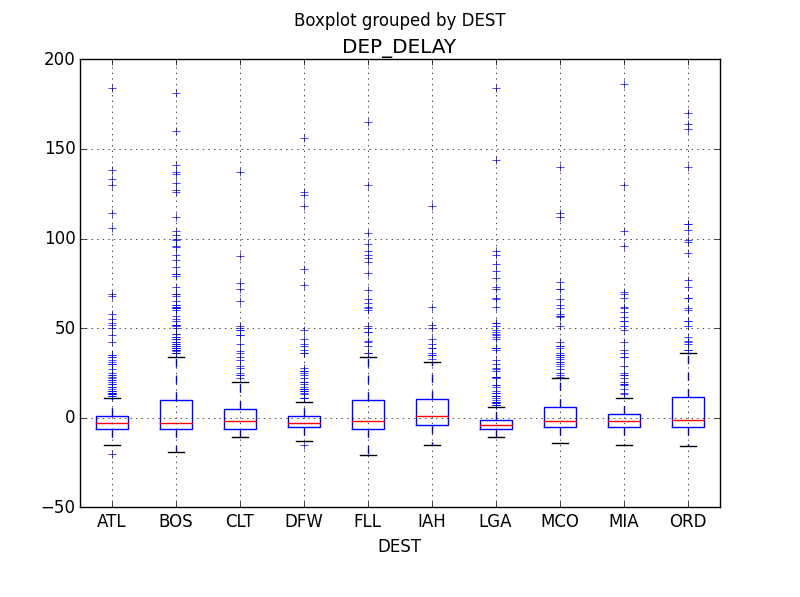
Every single flight in our time range is an observation. The selected predictors in the On-Time Performance dataset can include starting and ending airports of each flight leg, time of the flight, arrival times, and airport-to-airport distances. The response variable, or the outcome of the flight, is some type of description of the delay (actual delay time or a binary variable indicating a delay of greater than 15 minutes). Other predictors that can be merged, or cross-walked, with the On-Time Performance dataset include ASOS hourly weather metrics including temperature and precipitation for the initial and destination airports at the time of take-off and landing respectively.

1. Data Exploration

At this juncture of the analysis the data exploration has been broken out into two sections. The first subsets the dataset and dives deep into **DCA’s On-Time Performance** in February 2014. This airport was chosen randomly for the time being, although it might make sense to choose two airports that are have the best and worst on-time performance metrics and repeat the investigation.[[9]](#footnote-9) The following analyses were conducted:

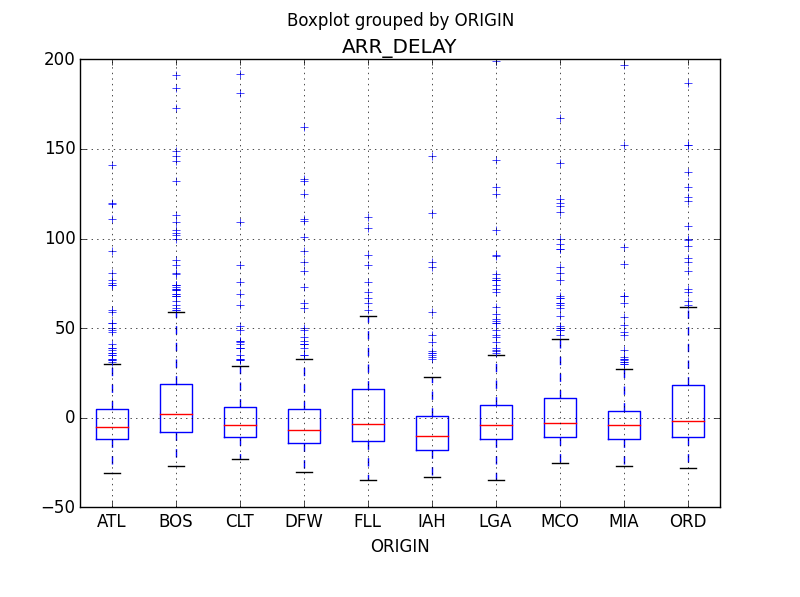
1. From what airports do most flights originating out of DCA arrive? The top flights are along the Eastern Seaboard: Boston, Atlanta, LaGuardia, Orlando and Miami.[[10]](#footnote-10)
2. Boxplots were created to see the distribution of the delays for a given set of airports – the top ten most traveled-to airports. On average it seems that flights to Boston and Orlando are usually delayed by approximately 10 minutes. Flights to Houston and Atlanta are usually a little early.

Figure . Boxplot Showing Distribution of Delay for Flights out of DCA



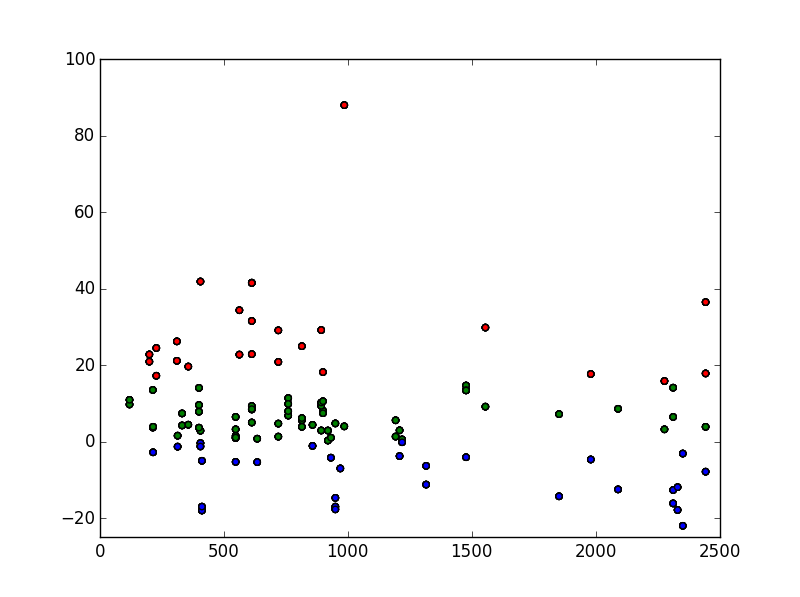
1. The top ten most frequent destinations of flights into DCA are all on average early, with the exception that flights originating from Boston usually late.

Figure Boxplot Showing Distribution of Delay for Flights into DCA



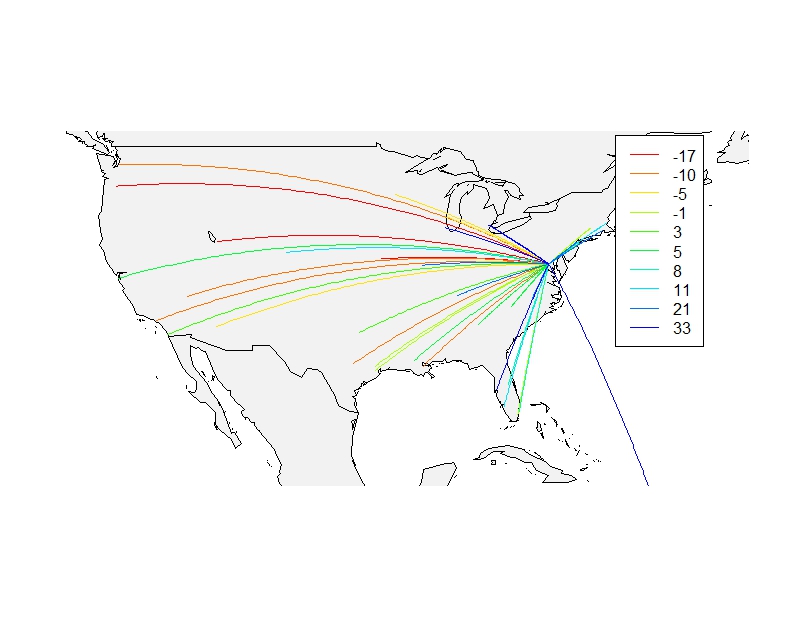
1. Interestingly the average delay out of DCA is almost seven and half minutes whereas the delay into DCA is less than five minutes. Therefore on average, flights in and out of DCA fall within the grace period of not being “late.” However, it does seem to beg the question is there a bottleneck within the DCA airport that planes are delayed getting out of DCA, as opposed to getting into DCA?
2. One of the few continuous variables in the flight data is distance between airports. The arrival delay has been color-coded based on whether or not the flight arrived on-time (blue), within the 15 minute grace period, between 0-15 minute late, or greater than 15 minutes late (red). The plot indicates that with the exception of a few plots, the majority of delay flights over shorter-haul missions.

Figure Plot of Average Delay v. the Distance between Airports and DCA



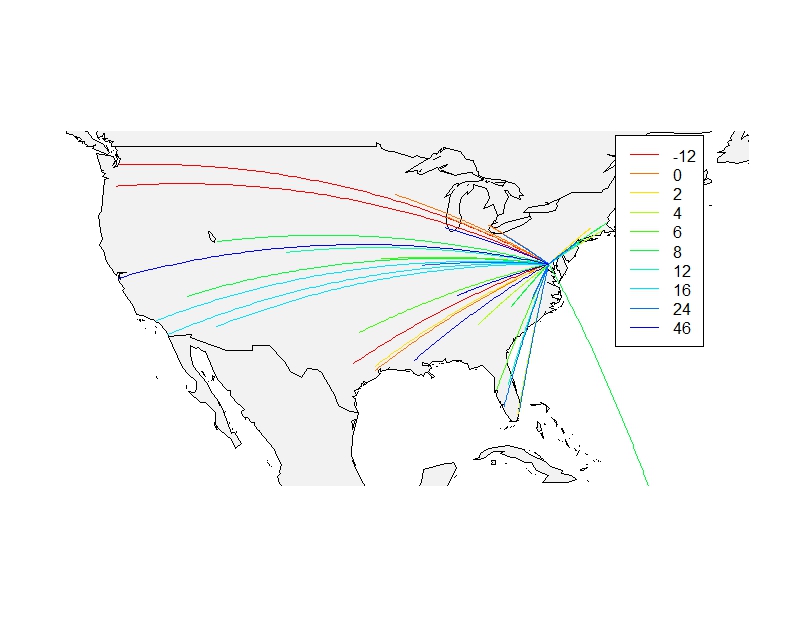
I used an R-script to plot flight patterns in and out of DCA. The first plot below indicates the Arrival Delay for flights arriving into DCA in February 2013. The flights with the greatest on-time performance originate from Portland, Seattle, Saint Paul, and Salt Lake City. These findings align with findings reported from *Travel and Leisure*. There does not seem to be a geographical area within the United States that consistently has delayed flights. I initially wanted to infer that east-bound flights have the best performance. However since DCA is located on the Eastern Seaboard, there are few, long-haul, domestic flights flying into DCA originating from the East.[[11]](#footnote-11) One might infer that the real reason that these flights from the West have the best on-time performance is not related to geographical origin but rather the sheer distance the flight has traveled. However it may be conjectured that shorter flights do indeed have worst performance times, since many of these flights have less air time, and therefore less buffer time to make up for any delay.

Figure 4 Arrival Delay of Flights Into DCA in February 2013



Flights out of DCA also tell an interesting story. Again we see that flights to the Northwest have good on-time performance metrics. The theory that long-haul flights correspond to better on-time performance does not correspond with the story told in this graphic. Rather flight headed towards the Southwest are on average delayed.

Figure 5 Arrival Delay of Flights Out of DCA in February 2013



A high-level view of the data has also been conducted, looking at aggregated view of delayed flights from November 2013 to January 2014. The first step in the aggregation process was to take the **Arrival Times** of each flight and bucket them to the nearest hour. This simplifies the dataset’s temporal component. Next, an average of **Arrival Delay** is taken for all unique combinations carrier, hour, originating airport and departure airport to obtain average flight delay. This variable **Arrival Delay** is a continuous variable, however a discrete variable **Avg\_Delay** is created to determine if on average the flight is delayed. The DOT defines a flight as delayed if the delay exceeds 15 minutes.

Plotting the time of arrival against the arrival delay paints an interesting picture. Arrival Delay is significantly worse during night hours and appears to be prevalent across all major Airlines. A possible explanation for this delay may be due to a cascading effect of flight delays building up over the course of the day. Alternatively, the delay could be due to the type of flights that arrive early in the morning, i.e. Red-Eye flights and night-loops. In the process of exploring and manipulating the data, a two insights can be inferred. Flight distance and flight arrival time are potentially significant factors in the arrival delay.

1. Features to use in the Analysis

The process of data manipulation, cleaning, and exploration is the most time consuming portion of any analysis (and usually the most tedious), however it is the most imperative step in any analysis. The insights and lessons learned from this process often determine the scope and focus of the model selection. Since RITA provides labeled outcome information, whether or not the flight was delayed, the supervised analysis indicates that a regression or classification is appropriate. However we need not limit ourselves to whether or not our predicted outcome is either continuous or categorical. We can use the categorical variable **Avg\_Delay** as a predictor variable for a classification model, K Nearest Neighbors (KNN), or Logistic Regression to determine the probability of delay. The continuous variable **Arrival\_Delay** can be used to create a Linear Regression. Finally a Decision Tree can be used for both discrete and continuous output.

1. Create a categorical variable that indicates the time of the week and time of the day to fly. Will have to determine when this categorical variable will defined as “commuter flight.”
2. Use the ASOS data to match up weather information: temperature and precipitation data points.
3. The dataset is heavy with categorical variables. Merging geospatial coordinates of origin and departure airports and can replace values the categorical variables defined by the Airport codes.
4. Try the following model types:
   * 1. KNN
     2. Logistic Regression
     3. Linear Regression
     4. Decision Tree
5. Analyze the breakdown the delays of the Top 29 Airports with delay causation details.

In all models created, both training and test data sets will be split and K-fold cross-validation. Validation will include Confusion Matrix and ROC Curve for the classification models that may be produced.

1. Appendix

Walking through the code:

**firstDataExplorationSteps.py :** walks through a deep-dive into a subset of the Data – all flight in and out of DCA. Flights are aggregated, split, merged, and appended. The script is eventually written out a csv file where it can then be pushed to an R-script for visualization purposes. The data is not provided due to the size. It can be pulled from here:

Filter Geography: All

Filter year: 2014

Filter Period: February

With the following fields selected are necessary for the analysis : Carrier, Origin, Dest, Departure Time, Arrival Time, Departure Delay, Arrival Delay, Air\_Time, Distance, Cancelled, Arr\_Del15

A list of all fields pulled are included in this csv file: **heading\_of\_RITA\_pulls.csv**

**KNN\_logisitic\_analyses.py** : appends November 2013, December 2013, and January 2014 data pulls from RITA. Plots each carrier’s delay based on time of arrival. Then investigate deeper into flights that arrive between 12AM and 6AM. The file begins to build out some models.

|  |  |  |
| --- | --- | --- |
| Filter Geography: All  Filter year: 2013  Filter Period: November | Filter Geography: All  Filter year: 2013  Filter Period: December | Filter Geography: All  Filter year: 2014  Filter Period: January |

1. 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-1)
2. 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-2)
3. The Arrival Delay metric makes more sense than a departure delay since often time airplanes can make for lost time on the ground with inflated scheduled flight times. [↑](#footnote-ref-3)
4. On-Time Performance queries can be pulled from http://www.transtats.bts.gov/DL\_SelectFields.asp?Table\_ID=236&DB\_Short\_Name=On-Time [↑](#footnote-ref-4)
5. The airport database, airport.dat can be found at http://openflights.org/data.html [↑](#footnote-ref-5)
6. Other airport codes are given in the dataset, the International Civil Aviation Organization (ICAO) code. However it is unnecessary for the purposes of the project, at this point of the project. [↑](#footnote-ref-6)
7. https://mesonet.agron.iastate.edu/ [↑](#footnote-ref-7)
8. http://www.ncdc.noaa.gov/data-access/land-based-station-data/land-based-datasets/automated-surface-observing-system-asos [↑](#footnote-ref-8)
9. In November 2014, *Travel and Leisure* provided a list of best and worst airports for flight delay. The top offender was Midway International Airport and the least offender was Salt Lake City International Airport. [↑](#footnote-ref-9)
10. A future analyses within this project might be to bucket flights – not only on what time of the day but whether or not the flight is during a commuter time or during hours of high demand. [↑](#footnote-ref-10)
11. This is also probably a case for why DCA is not the best airport to conduct a deep-dive. A more suitable airport could be in O’Hare or Midway, which are also notorious of delays. [↑](#footnote-ref-11)