Christine Luongo
December 18, 2014

Predicting Domestic Airline On-Time Performance

I. 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 especially 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] carrier market share, and on-time performance[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.

II. 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 airline 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]

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] An abbreviated list of the fields selected is given in Table 1.

[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
[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
[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.
[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

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 report 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] 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

[5] The airport database, airport.dat can be found at http://openflights.org/data.html

IATA/FAA	IATA code used to identify the airport. This code is used to crosswalk with other on-time performance dataset.[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 time zones 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] There are more than 900 ASOS units in the United States that collect weather information hourly, and possibly more frequently for special observations.[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 only those airports applicable to this analysis and the fields that seemed useful. The ASOS data was printed out to a single txt file that was later manually manipulated in Microsoft Excel. The relevant data fields are provide din the table below:

Table 3. 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

[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.

[7] https://mesonet.agron.iastate.edu/

[8] http://www.ncdc.noaa.gov/data-access/land-based-station-data/land-based-datasets/automated-surface-observing-system-asos

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 and in conjunction with the airport code, the weather information for every single arrival and departure location can be combined with the original dataset.

III. Description of the On-Time Performance Dataset

RITA provides clean data with few missing or null values. The data manipulation process is described in the data exploration section below. Within python, I have appended three different months of on-time performance databases: November 2013, December 2013, and January 2014.

Every single scheduled domestic flight in our dataset is an observation, or a row. The selected predictors in the On-Time Performance dataset include originating and destination airports of each flight leg, flight time, arrival time, and airport-to-airport distance, departure time, and the carrier. The response variable, or the outcome of the flight, is some type of description of delay outcome (for example, the actual delay time or a binary variable indicating a delay of greater than 15 minutes). This analysis assumes that the Arrival Delay is the most delay outcome. The arrival delay is the ultimate indication of delay – a departure delay can be ameliorated or corrected over the course of the flight.

Predictors from other data sources that can be merged, or cross-walked, with the On-Time Performance dataset include the following fields:

Table 4 Added and Relevant Data Fields

Field	Source	Description/Purpose
Latitude (Airport)	Airports Dataset	This field can inform us the distance and bearing between arrival and destination airports.
Longitude (Airport)	Airports Dataset	This field can inform us the distance and bearing between arrival and destination airports.
Timezone	Airports Dataset	This field can be used to indicate inter-timezone flights.
Temperature	Weather Dataset	This field can be used to gauge weather factors at both arrival and destination airports.
Wind Direction	Weather Dataset	This field was deemed potentially useful in determining effects of wind direction on both arrival and destination airports.
Wind Speed	Weather Dataset	This field was deemed potentially useful in determining effects of wind speed on both arrival and destination airports.
Precipitation	Weather Dataset	This field was deemed potentially useful in determining effects of the amount of hourly

		precipitation on both arrival and destination airports.
Visibility	Weather Dataset	This field was deemed potentially useful in determining effects of cloudiness on both arrival and destination airports.

IV. Data Exploration

The initial data exploration was broken out into two sections. The first portion of the data exploration subsets the dataset and dives deep into **DCA's On-Time Performance** in February 2014. This airport was arbitrarily chosen, 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] The following analyses were conducted:

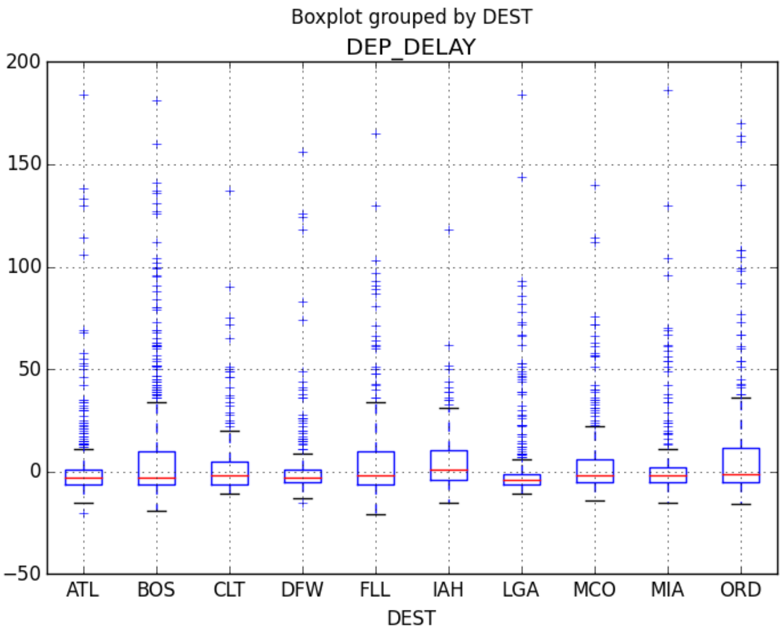
1. From what airports do most flights originating out of DCA arrive? The top flights are along the Eastern Seaboard: Boston, Atlanta, LaGuardia, Orlando and Miami.[10]
2. Boxplots were created to see the distribution of the delays for the top ten most traveled-to airports. On average it seems that flights to Boston and Orlando are usually delayed by approximately 10 minutes.[11] Flights to Houston and Atlanta are usually a little early.

[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.

[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.

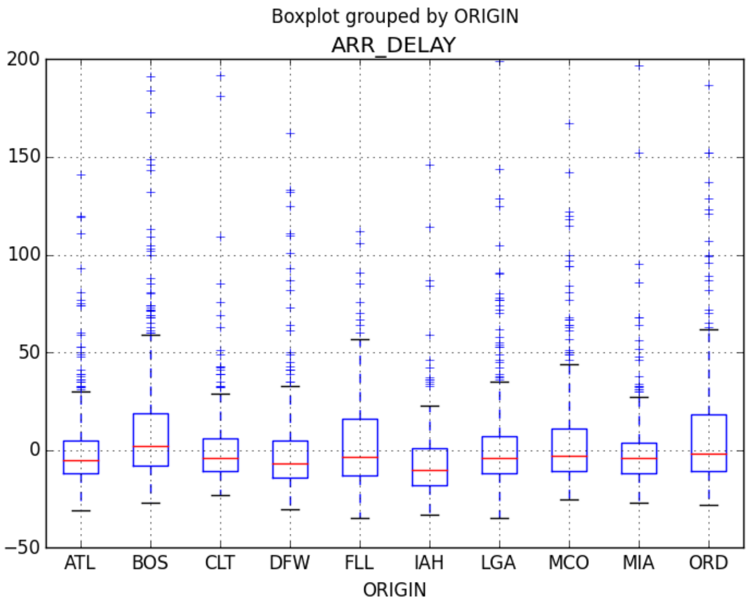
[11] These airports also happen to be frequently flown via Jet Blue.

Figure 1. Boxplot Showing Distribution of Delay for Flights out of DCA



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

Figure 2 Boxplot Showing Distribution of Delay for Flights into DCA

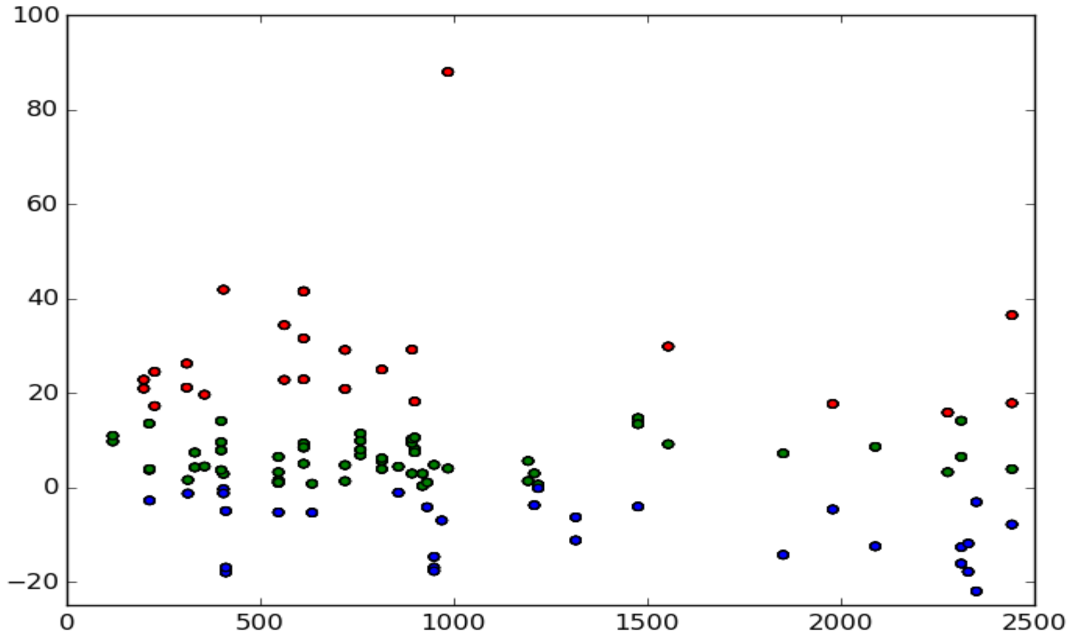


4. Interestingly the average delay out of DCA is almost seven and half minutes whereas the delay into DCA is less than five minutes. Therefore on average, flights in and out of DCA

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?

5. 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 3 Plot of Average Delay v. the Distance between Airports and DCA

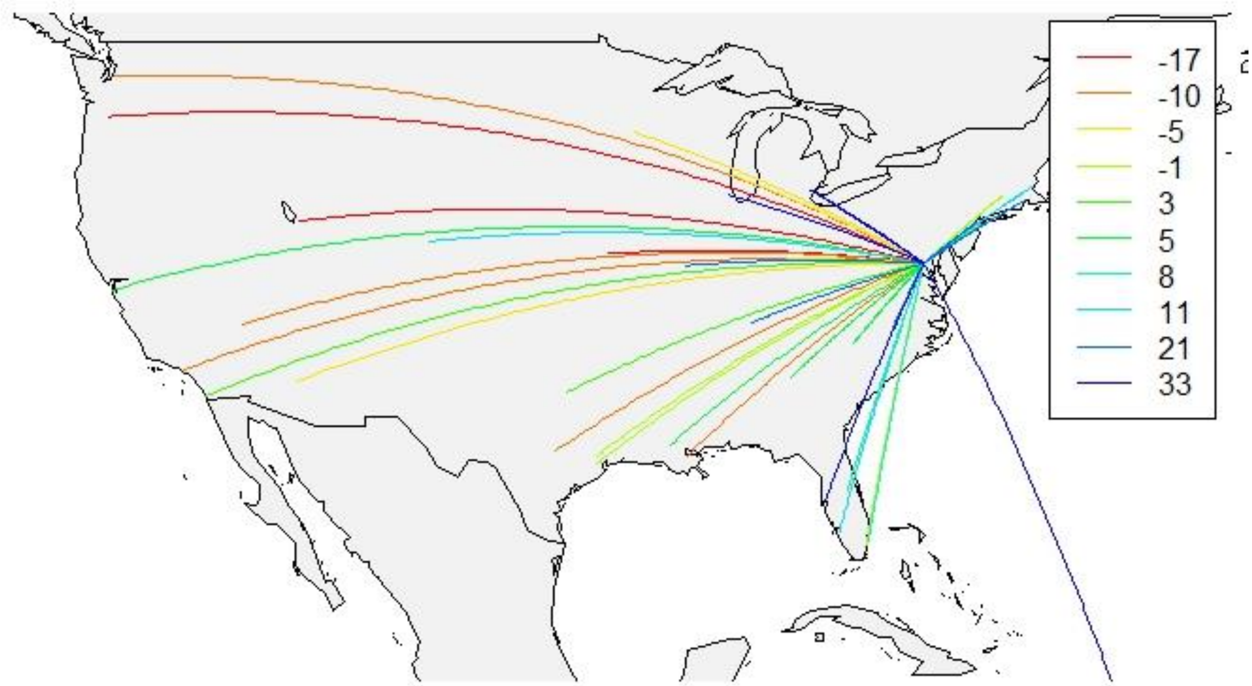


I used an R-script to plot flight patterns in and out of DCA. The first plot below indicates the Arrival Delay for flights arriving into DCA in February 2013. The flights with the greatest on-time performance originate from Portland, Seattle, Saint Paul, and Salt Lake City. These findings align with 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.[12] 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

[12] 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.

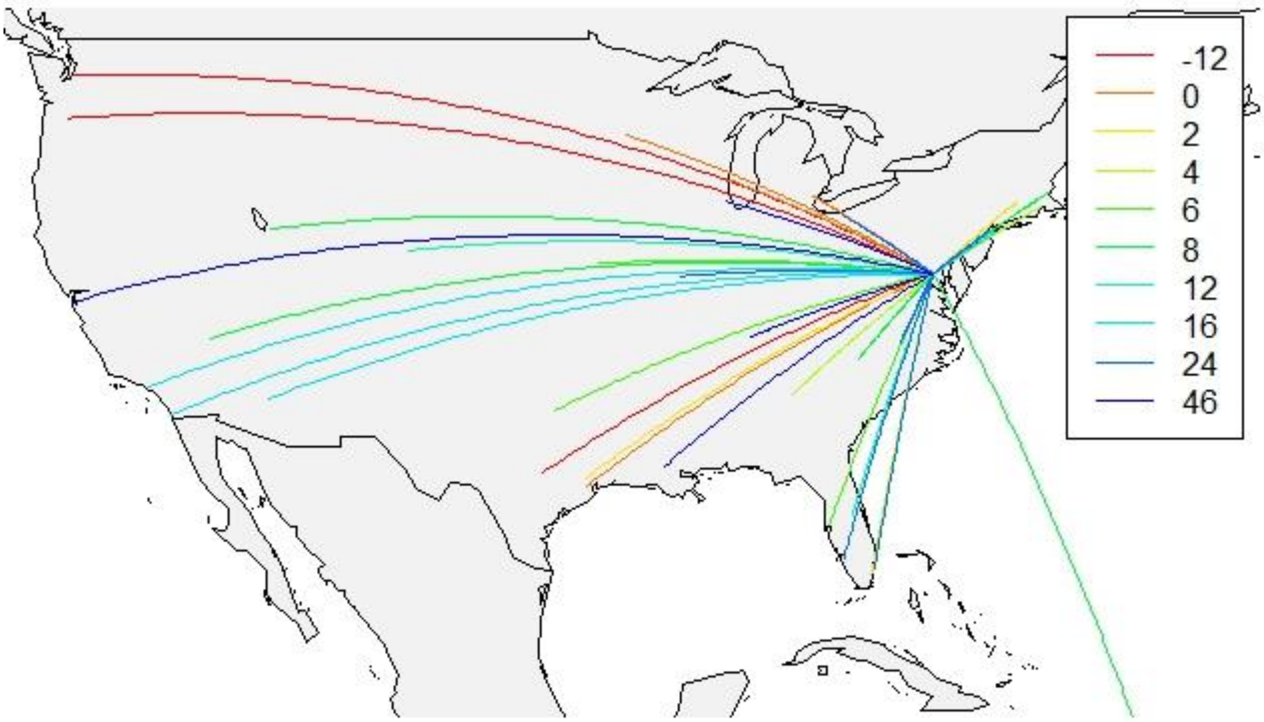
times, since many of these flights have less air time, and therefore less buffer time to make up for any delay.

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, flights headed towards the Southwest are on average delayed.

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



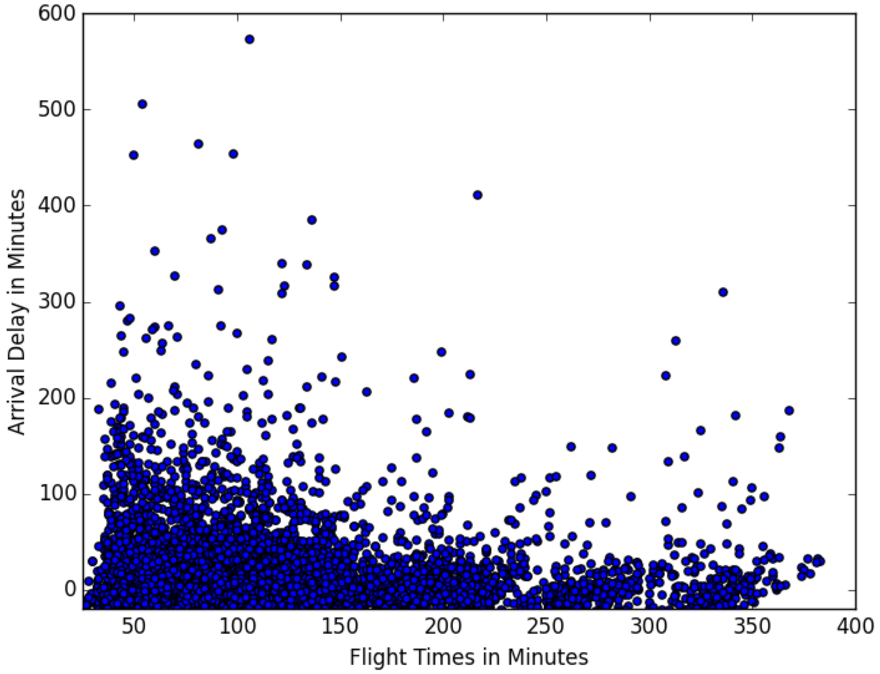
A high-level view of the data was also conducted, looking at aggregated view of delayed flights from November 2013 to January 2014. The first step in the aggregation process was to take the **Arrival Times** of each flight and bucket them to the nearest hour. This simplifies the dataset's temporal component. 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, two insights can be inferred. Flight distance and flight arrival time are potentially significant factors in the arrival delay.

Findings of the preliminary data exploration lead me to conduct the rest of my analyses on a single dataset, of those flights in and out of DCA from November 1, 2013 to January 31, 2014. The Weather Dataset, in addition to the Airport Dataset, had been added to this subset of the On-Time Performance Dataset. This lead to the following findings:

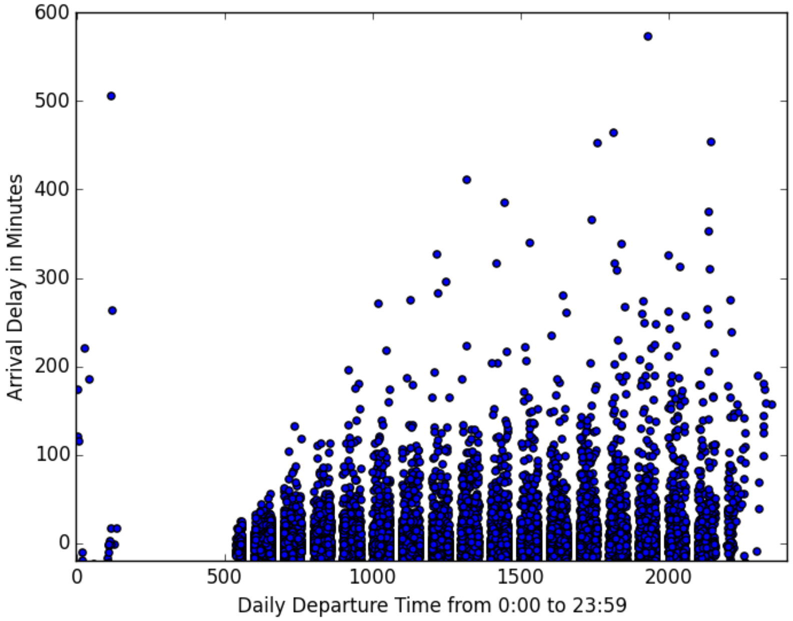
The relationship between airtime and the arrival delay. Shorter airtimes seem to indicate the greater likelihood of a delay.

Figure 6 Arrival Delay v. Flight Time

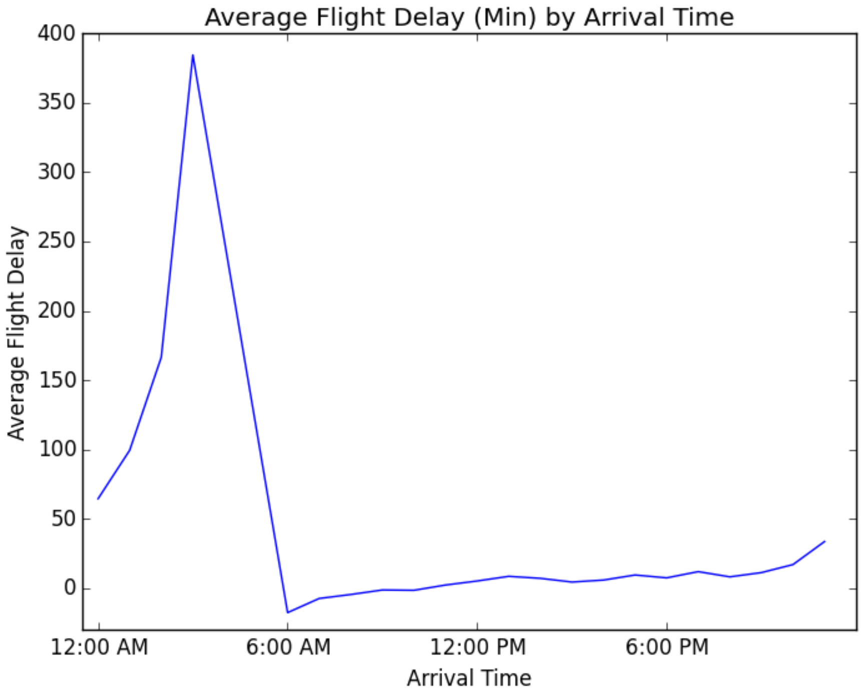


The relationship between departure time and arrival delay indicate that later in the day, the more likely the flight is going to be delayed.

Figure 7 Arrival Delay v. Departure Time



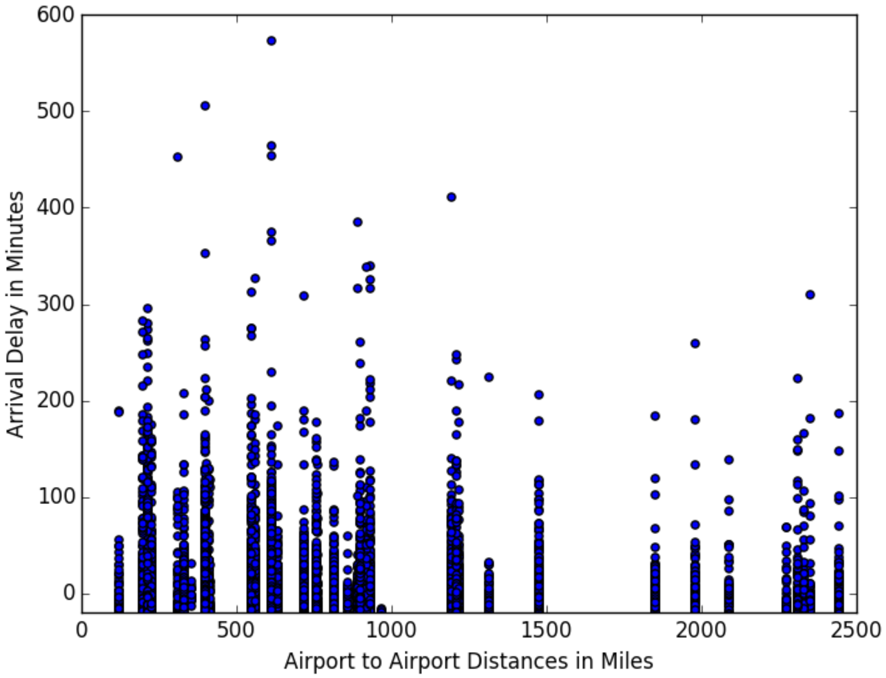
Another visualization showing the relationship between arrival times and average flight delay is shown below:



A variable **night_hour** was created to indicate whether the arrival time falls between 7:00PM and 5:00 AM.

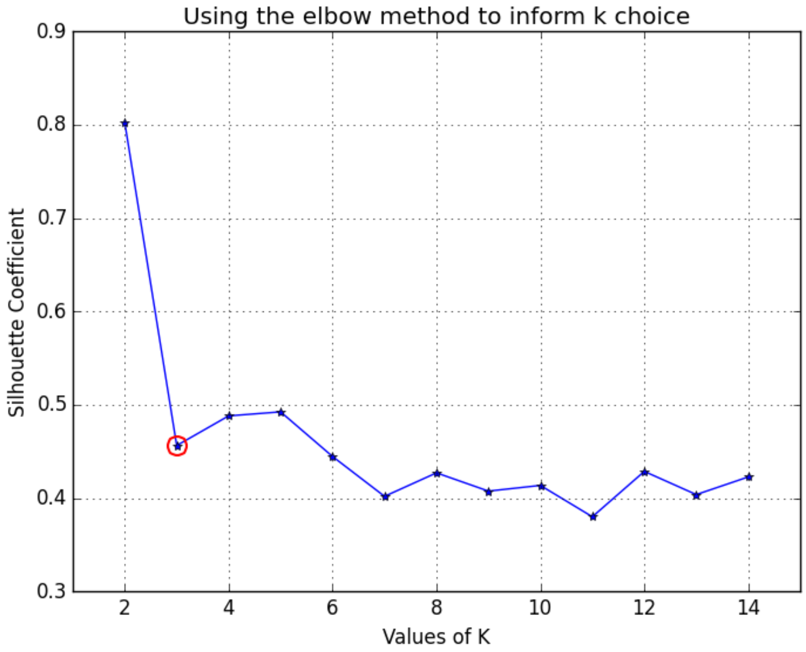
The relationship between arrival time and arrival delay mimics very closely to departure time and arrival delay and is not included. We might also infer that the relationship between airtime to delay and distance to delay be similar. The relationship between distance and arrival delay is shown below:

Figure 8 Arrival Delay v. Distances between Airports

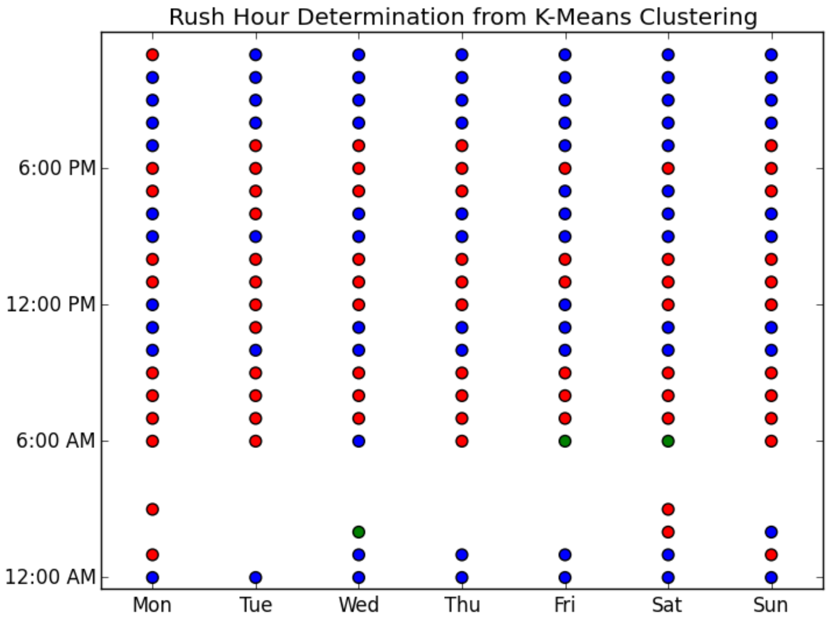


Since departure time seems to be relevant to the delay, a few variables were included:

- **Weekend_YN**: a binary variable that indicates whether or not it is a weekend. Weekend is defined as either Saturday or Sunday.
- **Rush_Hour**: a binary variable that indicates whether or not the flight was perceived to be a commuter flight. Taking the hour in the week (from Sunday midnight at 0 to Saturday 11 PM at 167) and the number of flights in those hours and the average distances of those flights a mini k-means analysis was carried out. This mini-analysis segmented flights by their departure time, size, and average distance (these were the variables in our dataset I deemed could indicate whether or not a flight was a commuter flight. Originating and departing airports like New York and Boston seemed like they also might be good indicators of whether or not the flight was a commuter flight, however this was not included).



Using the k-means cluster with k=3, an assignment was created indicating whether or it is "rush hour." The figure below shows the clustering assignment based on day of the week and hour of the day. It is possible that the red clusters could be indicative of "rush hour" or "commuter flights." This field was added to the model to see if it brought any additional information to the model.



Distance definitely plays a role in arrival delay and from the Figures showing flights in and out of DCA, it appears that the direction, or bearing, of flights might also play a role in the delay outcome. Two fields were created to represent the direction of the flight: **bearing** was calculated using the Haversine Formula and a categorical variable **bearing_bucket**. The categorical variable buckets the bearing into 45 degree increments.

Flights in and out of DCA, those heading north had less chances of delay. Flights heading W-NW and SW-W are generally more likely to be delayed.

V. 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 **Delayed_YN** 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.

The first model attempted was a Linear Regression. This model was not able to be solved since there were more unique possible combinations of independent variables than dependent variables. Since the dependent variable, the arrival delay, only takes on whole number values and there are over 11,000 observations. From what I could infer, there were more unique observations of the independent variables than the dependent variables.

I inferred that the arrival delay, though given in minutes, were actually rounded to the nearest integer value. So I induced some artificial randomness into the arrival delay. This way there would be more unique values for the independent variable. I did so by creating an array of random numbers between -.5 and .5 and adding that to the arrival delay. This field is called **ARR_Delay_Random.** However the runtime of the model proved to be to be too computationally taxing and I refrained from continuing to analyze it.

Next I tried a Logistic Regression, using the Arr_Delay_YN field. I started with all possible independent variables and removing those variables that are the least significant. The initial pseudo R-Squared Value was 0.15. This is the best possible R^2 we can get from this group of variates.

I tried replacing the dummy variables generated for the Carrier with those binary variables specific to the carriers I found that deviated from the average delay rate the most: JetBlue, Envoy, American Airlines, US Airways and Delta. One-by-one I removed variables that still seemed insignificant. Resulting in the following model:

Variable	Coefficient	Standard Error	Z	P >|Z|	95% Confidence Interval	
Intercept	0.2102	1.498	0.14	0.888	-2.726	3.147
AIR_TIME	0.072	0.004	18.185	0	0.064	0.08
DISTANCE	-0.0086	0.001	-15.82	0	-0.01	-0.008
ORIGIN_lat	-0.0808	0.012	-6.95	0	-0.104	-0.058
ORIGIN_long	-0.0719	0.011	-6.451	0	-0.094	-0.05
DEST_long	0.0506	0.01	4.897	0	0.03	0.071
HOUR_ARR	-0.0446	0.011	-4.098	0	-0.066	-0.023
HOUR_DEP	0.1684	0.012	14.197	0	0.145	0.192
DEST_temp	-0.0055	0.002	-2.437	0.015	-0.01	-0.001
DEST_wind_direct	-0.0007	0	-2.138	0.033	-0.001	-5.70E-05
DEST_wind_speed	0.0181	0.007	2.5	0.012	0.004	0.032
DEST_visibility	-0.1012	0.011	-9.395	0	-0.122	-0.08
ORIGIN_temp	-0.0247	0.003	-8.028	0	-0.031	-0.019
ORIGIN_wind_direct	-0.0008	0	-2.369	0.018	-0.001	0
ORIGIN_wind_speed	0.0308	0.007	4.303	0	0.017	0.045
ORIGIN_visibility	-0.1335	0.011	-11.964	0	-0.155	-0.112
weekend_YN	0.4422	0.063	6.983	0	0.318	0.566
JETBLUE	0.4311	0.14	3.083	0.002	0.157	0.705
ENVOY	0.5357	0.108	4.952	0	0.324	0.748
AMER	-0.3204	0.107	-3.005	0.003	-0.529	-0.111
USAir	-0.322	0.102	-3.164	0.002	-0.521	-0.123
DELTA	-0.4564	0.109	-4.197	0	-0.67	-0.243

I removed many variables that I thought would have been significant: bearing, precipitation in both originating and destination airports, and the K-means rush-hour variable.

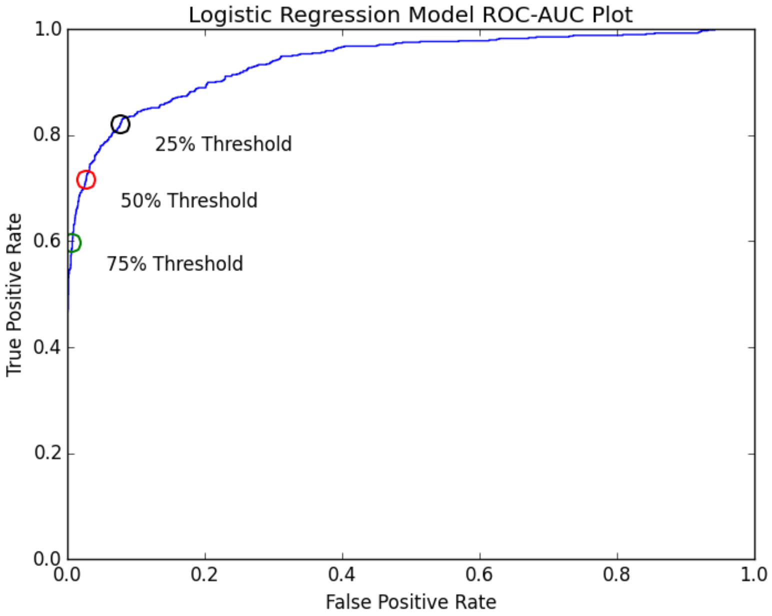
Next I considered a model with Departure Delay as an independent variable. Although some might have considered this a potential independent variable in these analyses, a delay taking off does not ultimately determine whether or not a flight was delayed upon arrival. I generated a second logistic regression that started with only a single dependent variable and added more to see how the adjusted R squared increased. This model produced a better Adjusted R squared, mostly due to the addition of the Departure Delay variable. Of the temporal variables I created, **rush_hour**, **night_hours**, and **time_of_day**, only **Weekend_YN** binary variable proved to be significant. Visibility did prove to be significant, as expected.

Variable	Coefficient	Standard Error	Z	P >|Z|	95% Confidence Interval
Intercept	-2.1072	0.215	-9.815	0	-2.528 -1.686

DEP_DELAY	0.1552	0.004	36.031	0	0.147 0.164
AIR_TIME	0.089	0.004	20.928	0	0.081 0.097
DISTANCE	-0.0112	0.001	-20.161	0	-0.012 -0.01
bearing	-0.005	0.001	-7.983	0	-0.006 -0.004
weekend_YN	0.2481	0.096	2.595	0.009	0.061 0.435
DEST_visibility	-0.0521	0.016	-3.223	0.001	-0.084 -0.02
ORIGIN_visibility	-0.1086	0.016	-6.795	0	-0.14 -0.077

Adding anymore variables to the model did not greatly improve the Pseudo R-Squared Value.

I split the dataset into training and testing sets. Using 50% likelihood as an indicator of delay the model produced an accuracy score of 91.7%, a ROC AUC score of 93.9%.



As a flyer, I would prefer to have a higher false negative rate – predict flights were delayed more often than they actually are.

Using 5-fold cross validation techniques, the mean score of my ROC AUC score was 93.7%.

VI. Challenges and Next Steps

In all, I believe that there are other variables not present in the data that could help tell more about predictors of delay. The type and size of the plane could potentially tell more about the likelihood of delay. Alone, the variables that RITA provides cannot accurately predict the likelihood of delay. Another miniature analysis that could shed more light on the causes of delay could originate from the top 29 airports, which break out delay time into four types security, weather . Additionally, continuing to delve into the iniated KNN and Decision Tree Models might also prove useful.

This analysis determines that there some ways the consumer can avoid delays.

- Booking flights that depart before 7:00PM, although this most certainly will be dependent on your departing airport.
- Flights during the weekend are more likely to be delayed as well. Flights on Tuesday and Wednesday had the least likeliness of delay.

Other factors, such as visibility and departure delay, the flyer will not have in advance of the trip.

VII. Appendix

Brief description of each python file.

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

Analysis_postcleaning.py: reads in winter months data, performs some more exploratory analysis on relevant variables and then performs splits the data into test-train subsections and eventually a 5-fold cross validation of the model. It also includes a Linear Regression, KNN evaluation and Decision Tree analysis.