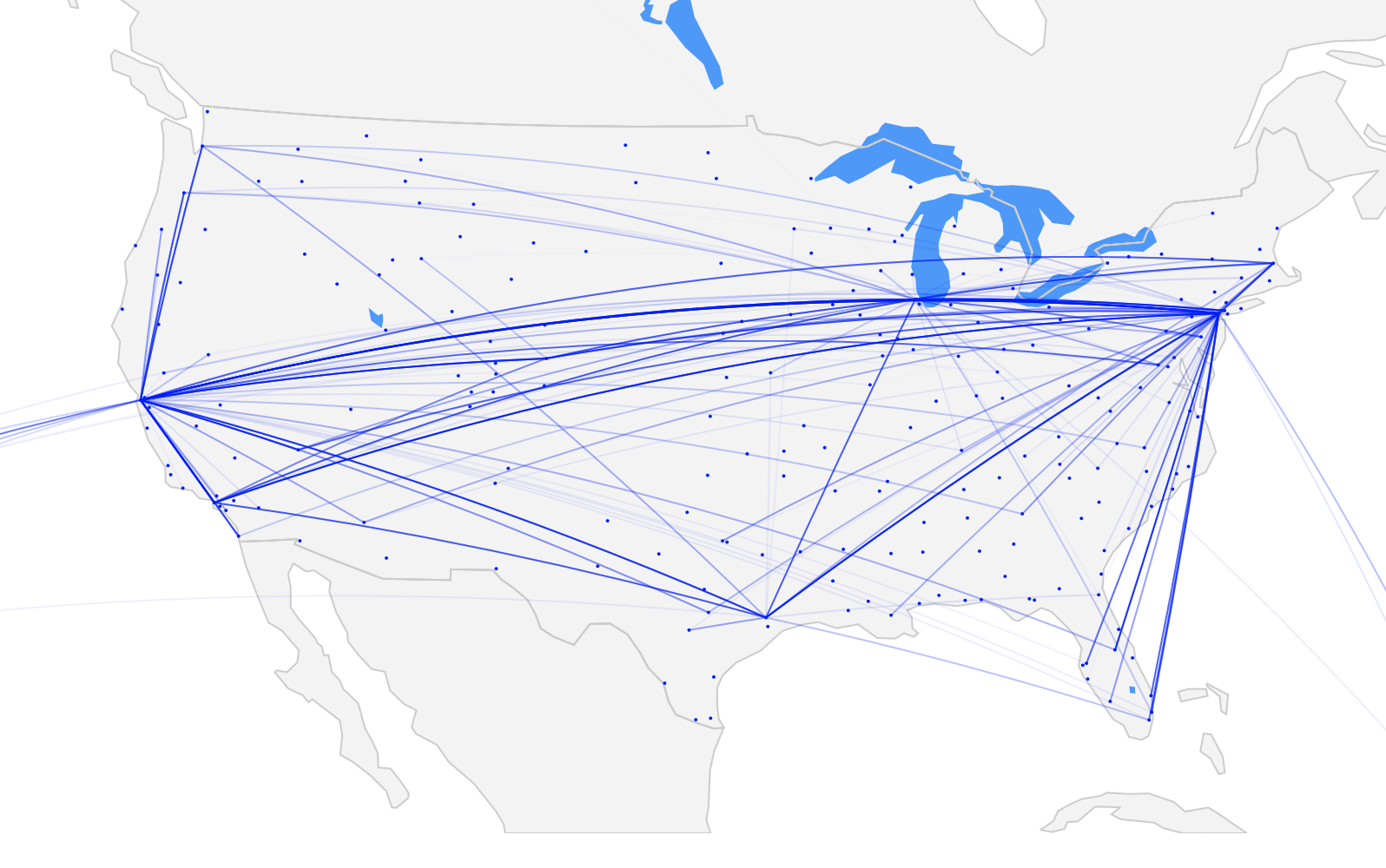


US AIRLINE PERFOMANCE

 2017 United Airlines Most Delayed Routes

Dmitriy Kats

May 2019

Table of Contents

[1. Introduction and Executive Summary 3](#_Toc7514338)

[2. Acquiring and aggregating the data 3](#_Toc7514339)

[2.1. Adding Weather Data 3](#_Toc7514340)

[2.2. Adding Airport Coordinates 4](#_Toc7514341)

[3. Cleaning the data 5](#_Toc7514342)

[4. Exploratory Data Analysis (EDA) 6](#_Toc7514343)

[4.1. Distribution of Arrival Delays 6](#_Toc7514344)

[4.2. Airport and Airline On-Time Performance 8](#_Toc7514345)

[4.3. Regional Analysis 10](#_Toc7514346)

[4.4. Regional Traffic Volume and Delays 13](#_Toc7514347)

[4.5. Individual Route Analysis 14](#_Toc7514348)

[5. EDA Conclusions 16](#_Toc7514349)

[6. Modeling 17](#_Toc7514350)

[7. Feature Engineering 18](#_Toc7514351)

1. **Introduction and Executive Summary**

Air travel has always been and still is a headache for many travelers. The unknowns of delays and cancellations are some of the biggest contributors to the stress. In this analysis we’ll attempt to shine a light on the unknowns and try to predict the probability of delay and even the delay length of a given future flight. We’ll take a look at 15 years of airline performance data, containing over 75 million flights in a dataset available from US Department of Transportation.

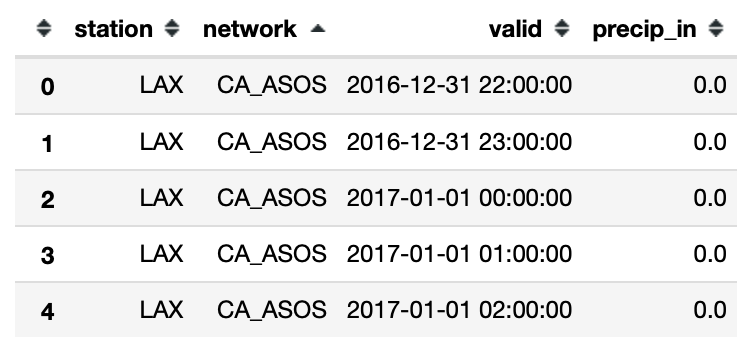
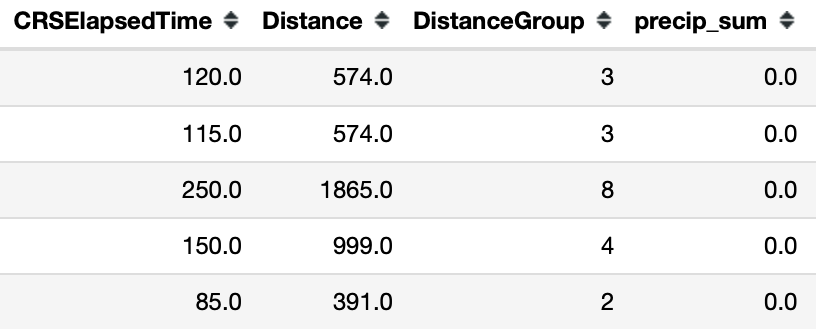
1. **Acquiring and aggregating the data**

The data was obtained from United States Department of Transportation. Unfortunately, the data are only available by month in a zipped file, which presents a challenge when it comes to downloading and concatenating. Additionally, the website utilizes a webapp to click on checkboxes and buttons to activate the download, which makes scraping impossible. As such, the data was acquired manually. Due to limitations in available data in years prior to 2004, only data for 2004 - 2018 was obtained for a total of 12 zipped files per year (for a total of 12 files x 15 years = 180 files).The data was then combined into a set of files organized by year and then into one complete csv file. The dataset will be uploaded to Google BigQuery for further analysis. For this analysis we will be using the 2017 Dataset. The code to combine all the files is available on github.

In summary, the following steps were taken:

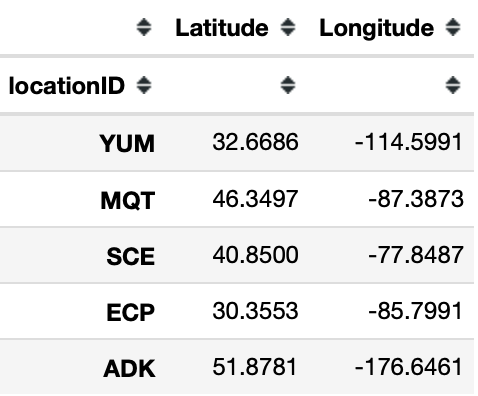
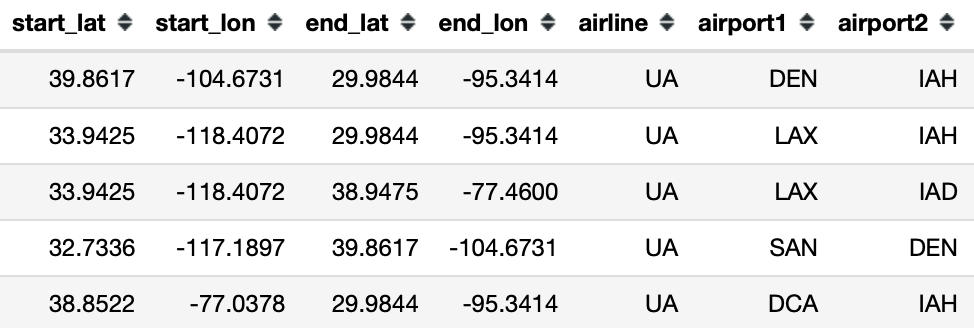
1. Create a list of files in each year directory via glob method.
2. Loop through each file, unzip, and read into a Pandas Dataframe via pd.read\_csv, utilizing the “compression=’zip’” parameter.
3. Since each year’s file will be ~3GB in memory, convert the datatypes in each column to reduce size.
4. Repeat the process for each year by looping through year folders.
5. Concatenate all csv files into one via shell and upload to Google BigQuery.
   1. Adding Weather Data

Additional data were obtained during the feature engineering stage in order to compute precipitation information for a specific airport and departure time. The precip\_sum column was created in the flight data frame. This column includes the sum of four hours’ worth of precipitation data prior to the departure hour.

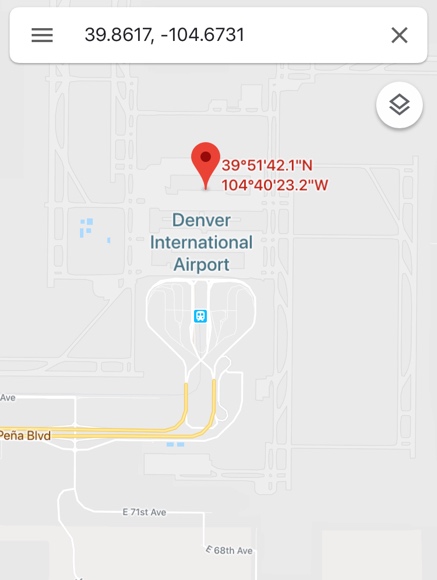
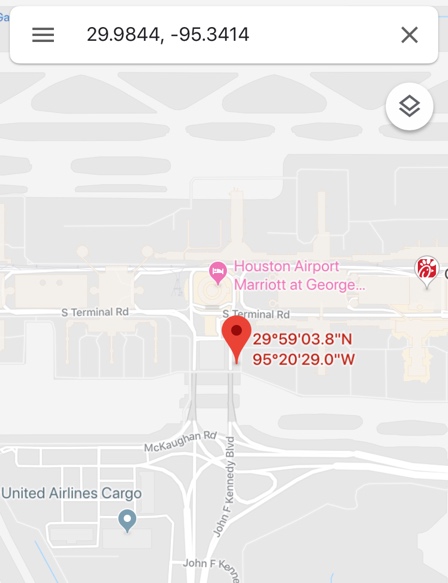
 

* 1. Adding Airport Coordinates

We’re also going to need airport coordinates in order to create extra features for the modeling portion of the project. We’ll first import a csv file with all US airport latitude and longitude data and join it with our flights data frame.

Spot checking DEN & IAH (Denver International & Houston International):

1. **Cleaning the data**

The dataset is relatively clean from the perspective of missing data. Only 1.5% of rows have missing values, which can be removed without significant impact on the analysis. As we go through EDA and ML portions of this analysis, the data will need to be formatted to fit the needs of the approach. As an example, time of day is presented in the dataset as a float-type number, ranging from 00 to 59 mins, which leaves 60 to 99 blank. This can be problematic when visualizing the data or converting to a timestamp. This will be addressed through a custom method which will convert the float-type number to a proper format.

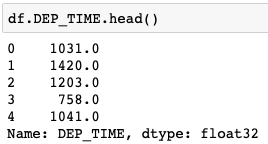


Table Example of Departure Time representation in the dataset (e.g. 1031.0 should be 10.52 as float)

Furthermore, we are going to be dealing with a lot of categorical data, such as airline names and airport codes, which will have an impact on predicting performance. In order apply modeling techniques these data will need to be converted into numerical values, as the solvers rely on mathematical algorithms. When encoding these values, we’ll have to be mindful of memory space required as there are over 300 unique origins and destinations and over 10 unique airlines. A possible solution to this is the use of a sparse matrix.

1. **Exploratory Data Analysis (EDA)**

In this section we’ll explore the data statistically and visually to understand any trends or features we should focus on during the modeling stage. The code for this section is available on github with a sample data set on Google drive:

Code:

<https://github.com/dmitriykats1/Springboard/blob/master/Capstone1/EDA-2017.ipynb>

Data:

<https://drive.google.com/file/d/15PXxxTY9X4w0exxu6vMEReJm3Pwcn4TE/view?usp=sharing>

* 1. Distribution of Arrival Delays

Let’s begin with the most important feature, delay time at the destination, represented as ArrDelay in our data set. This feature set has both negative and positive values, representing early and late arrival times, respectively. Let’s take a look at the distribution:

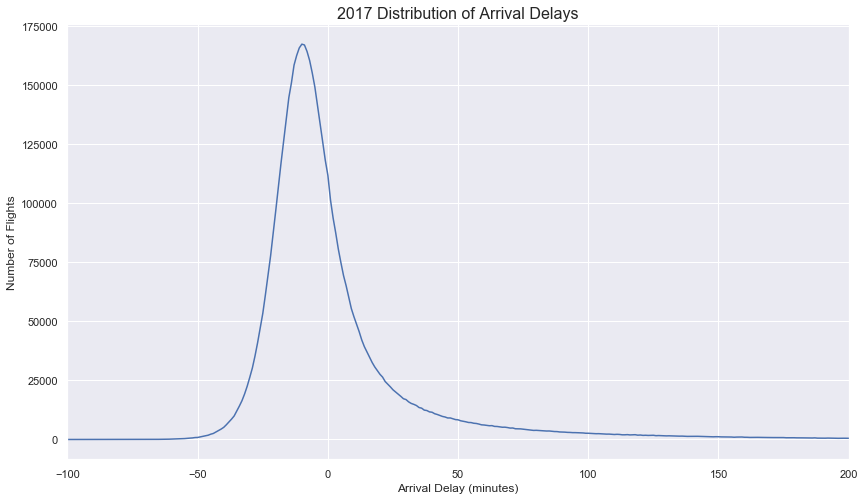


Figure Distribution of delays for all US flights in 2007

From the above we see that majority of flights being on time or early, and the distribution is also skewed and non-normal. With the mean arrival delay of 4.3mins., median of -6.0mins, and a standard deviation of 45.5mins. The next obvious questions we can ask is: Do long and short flights have same distributions? To answer this, we’ll split the dataset into long flights with durations longer than 3hrs and remainder will be short flights.

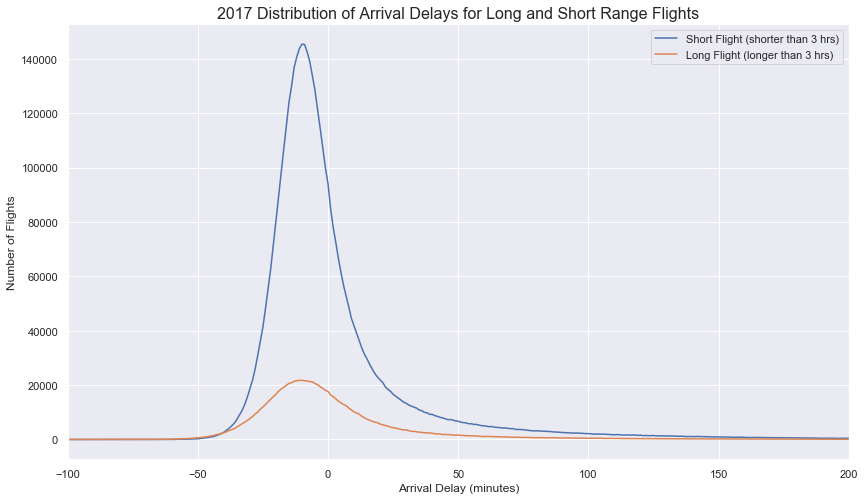


Figure Distribution of delays for all US flights in 2007 for short and long rage flights

The short- and long-range flights seem to have the same distributions with slightly different means which we can investigate use to investigate any statistical differences, with long range flights having a mean delay of 3.8mins and short-range flights having a mean delay of 4.4mins. Since majority of flights are on time or early, taking an average of arrival delay does not paint a full picture due to the skew. This can be better visualized below:

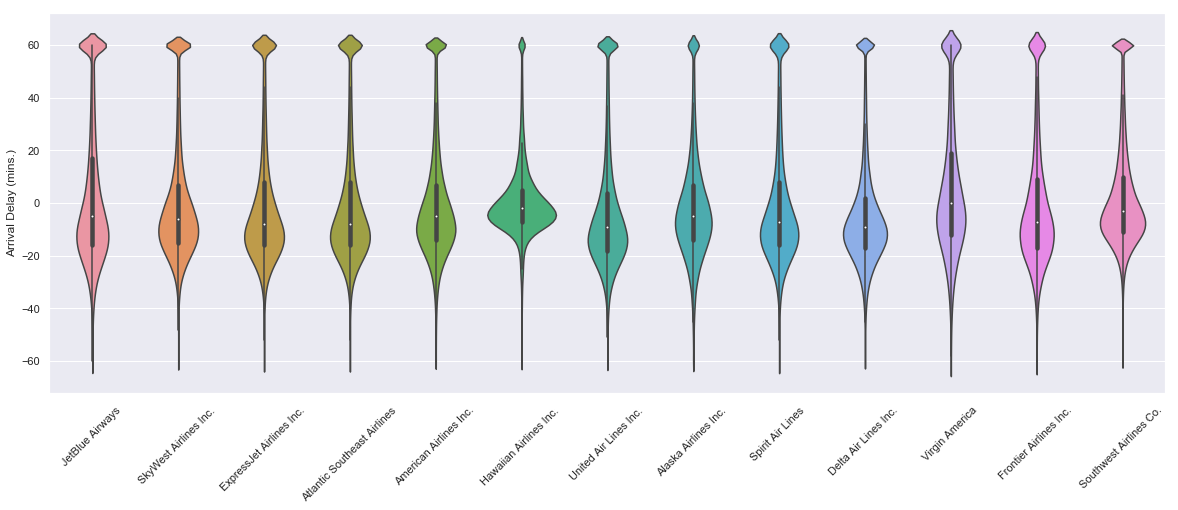


Figure Arrival delay distribution split by airlines

Next, we can take a look at ONLY delayed flights, or flights that are delayed 15 mins or more as defined by the FAA.

One of the questions we can ask is if there is any correlation between arrival delays and other features. First let’s look at departure delays and arrival delays:

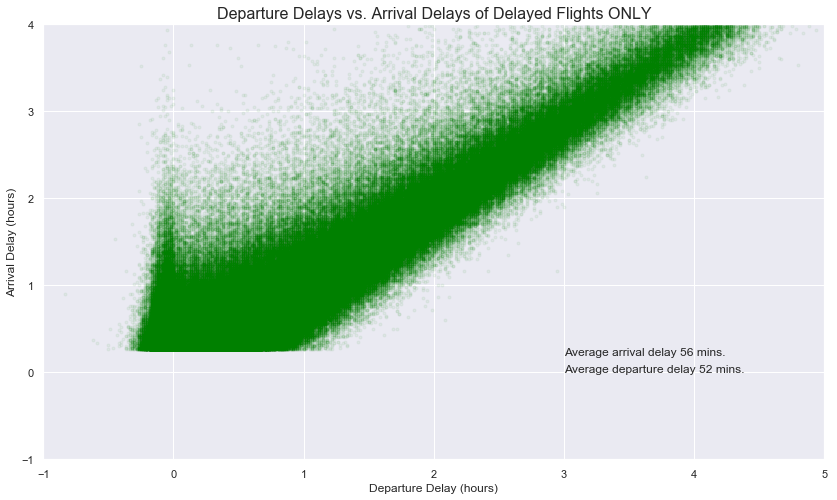
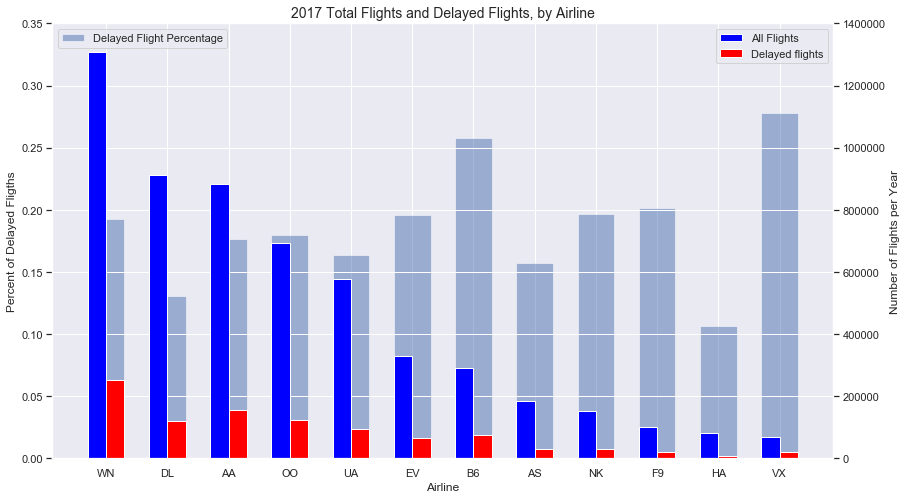


Figure Correlation of departure and arrival times for all delayed flights in 2017

There is a positive linear relationship between departure delays and arrival delays, this makes sense intuitively and we’ll need to keep this in mind when modeling our data. We do notice a small spike in arrival delays around the 0-minute mark for departure delays. This may be due to airlines trying to get out on time and closing the aircraft door in order to have an on-time departure. Only finding themselves waiting in taxi lines to depart and subsequently be delayed on arrival.

* 1. Airport and Airline On-Time Performance

Another set of features that may impact on-time performance are the airlines and airports. Let’s take a look at the summary of top airlines and on-time performance:



Delay Percentage

VX 0.277877

B6 0.257856

F9 0.201811

NK 0.196220

EV 0.195932

WN 0.192491

OO 0.179436

AA 0.176565

UA 0.163890

AS 0.157340

DL 0.130984

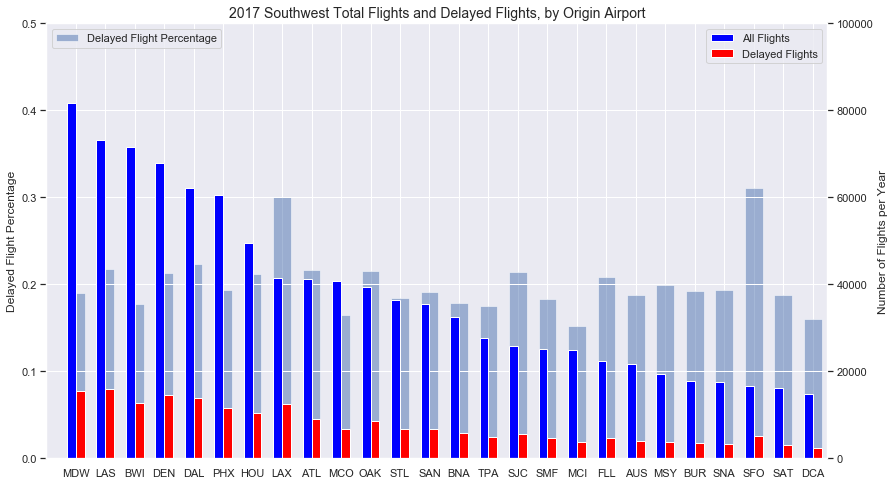
HA 0.106147

Figure Summary of Airline Performance for 2017

Above figure gives us a glimpse of how each airline performed in 2017, the data above is sorted by flight volume. It’s interesting that airlines with most flights don’t necessarily have the worst delay record, as we can see from the delay percentage bars. Delay percentages will certainly change as we break this data down by origin airport. Let’s see if there is a list of airports that have consistently high delays when broken down by airline.

Initial analysis looked at Southwest, American, and United Airlines (which make up 50% of all US flights) and delay percentages as a function of origin airport.

Southwest Airlines



SFO 0.311330

LAX 0.300265

EWR 0.271363

DAL 0.223535

LAS 0.218197

ATL 0.216150

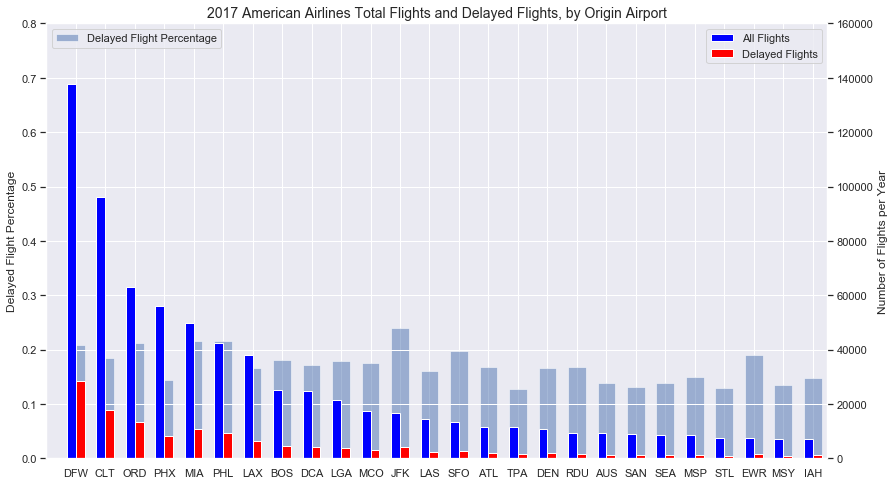
OAK 0.215234

SJC 0.214712

DEN 0.213428

HOU 0.212105

American Airlines



JFK 0.239166

MIA 0.215648

PHL 0.215173

ORD 0.212090

DFW 0.207660

IAD 0.203976

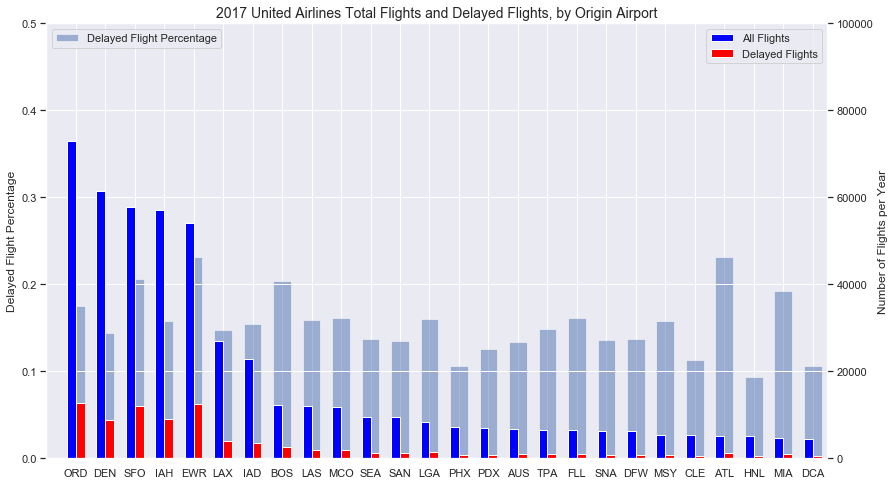
SFO 0.197801

EWR 0.189142

CLT 0.184354

BOS 0.180293

United Airlines



PBI 0.263673

EWR 0.231172

ATL 0.230904

BNA 0.219697

SFO 0.206219

BOS 0.204175

EUG 0.198057

SJU 0.197797

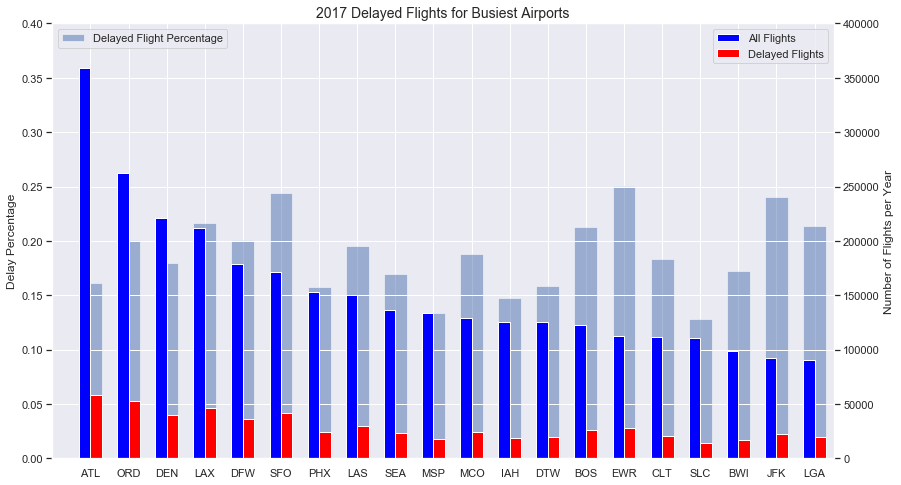
MIA 0.192012

ORD 0.174702

Figure Top Airlines by flight volume and origin airport on-time performance

We see EWR and SFO appear on the list for all three airlines. We can further investigate these airports and surrounding airports to see if regional air congestion impacts other airports in the area. But first let’s confirm these airports appear on the top delayed airports list.

Analyzing flight volume / delay volume percentages can give us a better picture of how the delays are weighted.



EWR 0.250590

SFO 0.244134

JFK 0.240389

LAX 0.216559

LGA 0.213504

BOS 0.212972

DFW 0.201238

ORD 0.200680

LAS 0.195448

MCO 0.187614

Figure Top Airlines by flight volume and origin airport on-time performance

We can see, in fact that EWR and SFO are top offending airports in 2017. But this list is sorted by flight volume, so we don’t get the full picture. Surrounding airports may be impacted. We can also note that all three major airports in the NYC area appear on the list: EWR, JFK, and LGA. Let’s look at EWR and airports within 30-mile radius and compare the performance to the national average.

* 1. Regional Analysis

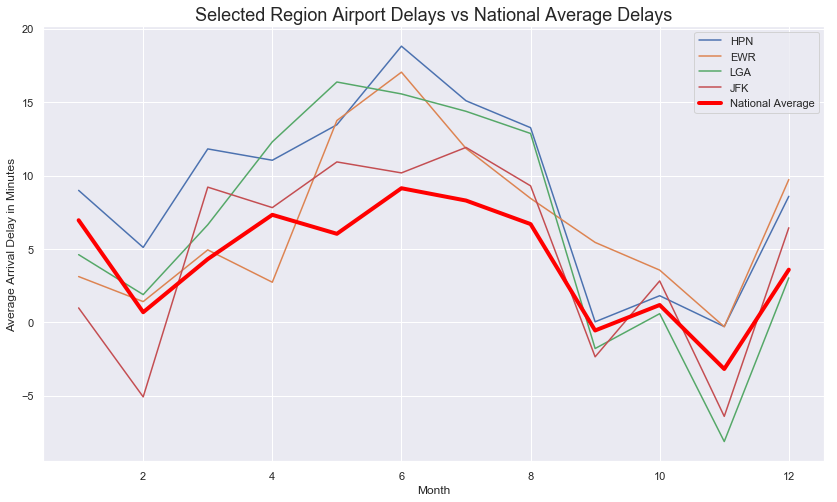


Figure NYC Area average delays compared to a national average, by month

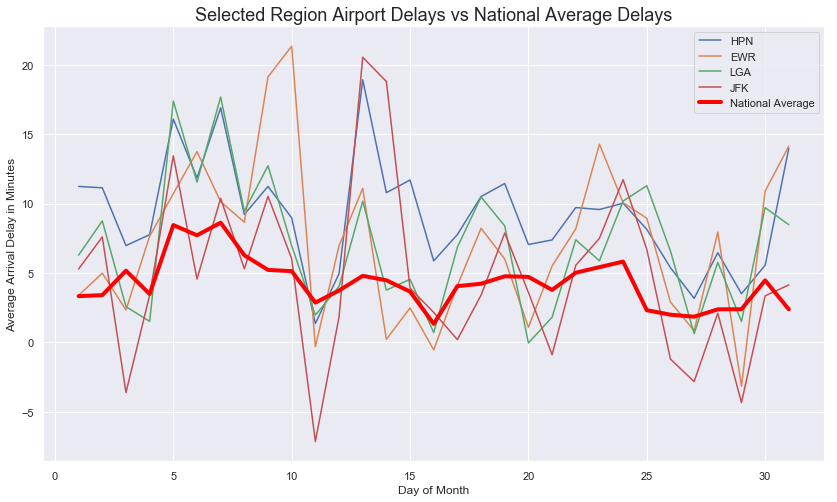


Figure NYC Area average delays compared to a national average, by day of month

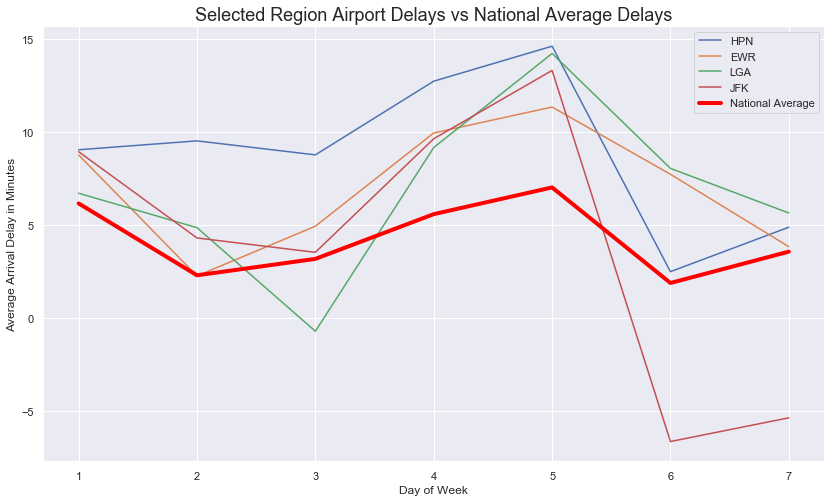


Figure NYC Area average delays compared to a national average, by day of week

We can see that all airports in the area tracks along with the national average on all charts, which was expected. However, regional airports tend to perform worse, on average. Note, that this analysis shows delays by minutes whereas last set of charts focused on percent of flights that were delayed. Next let’s take a look at how the delays break down, what is causing these delays?

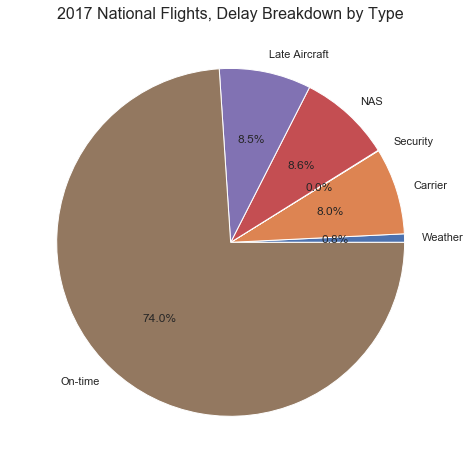
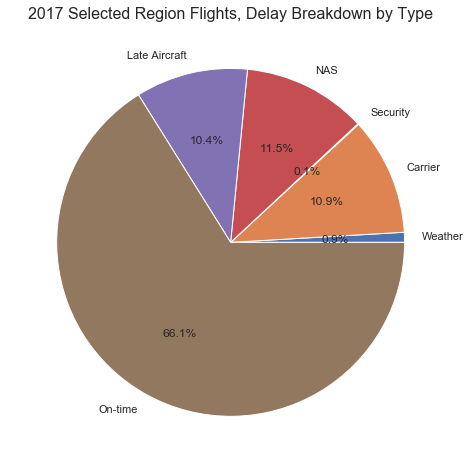
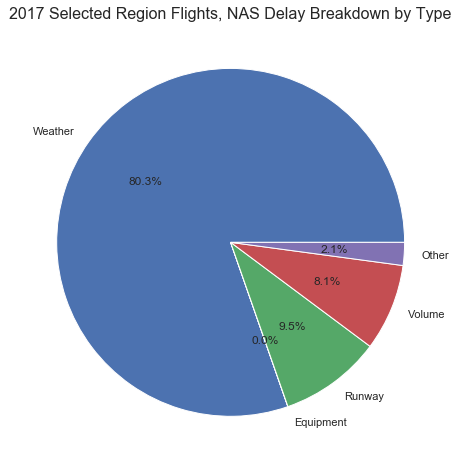


Figure NYC Area vs National delay breakdown by cause

Since NAS data is further broken down into sub-categories, we can download the actual data from FAA. Breaking down the FAA data (NAS delays section from above):



NAS - We can continue to look at patterns for heavy traffic or trends in increasing air traffic patterns areas and look for constancies. Additionally, extreme events are not considered here, normal weather patterns that cause air traffic slowdowns are. Looking at weather patterns may be beneficial. FAA has a database breaking down the NAS delays by cause. (FAA OPSNET)

Carrier - carrier performance can be further analyzed by location and see if there are patterns

Late Aircraft - This can be handled with arrival delay information. This field highly depends on the other delay causes.

Weather - only extreme weather events are considered here. These events are rare and would result in region-wide cancellations, as such, these can be ignored as outliers.

Figure NYC Area delay breakdown of NAS delay data

NYC Area had an on-time performance of only 66% in 2017 (as measured by arrival delay), which is lower than the national average of 74%. We can note that late aircraft, carrier, and NAS delays made up an overwhelming majority of delays. We’ll keep this in mind for further EDA and feature engineering stages of this project.

* 1. Regional Traffic Volume and Delays

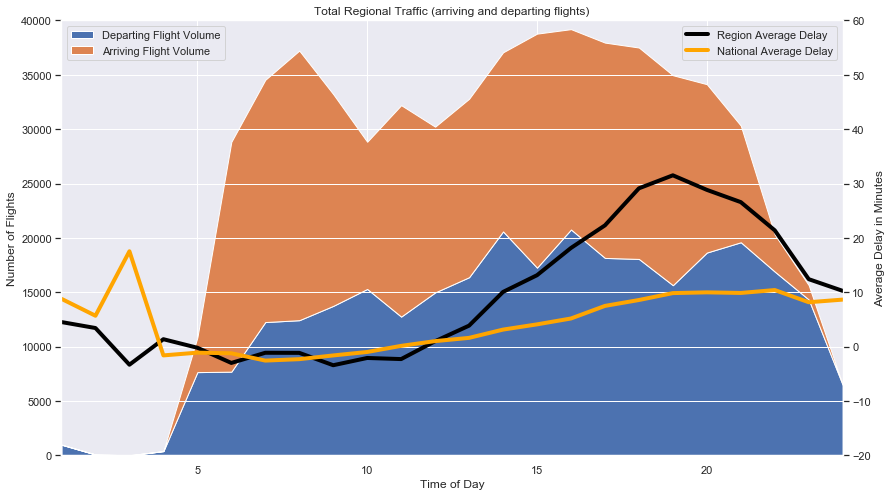


Figure Air traffic volume and average delay time, throughout the day

Taking into account arriving and departing flights in the NYC Area, we can see that air traffic is relatively flat throughout the day, however the average delay in minutes tends to increase towards the afternoon and evening hours. Although, the same can be seen on a national scale, the delays peak at ~30mins for NYC area.

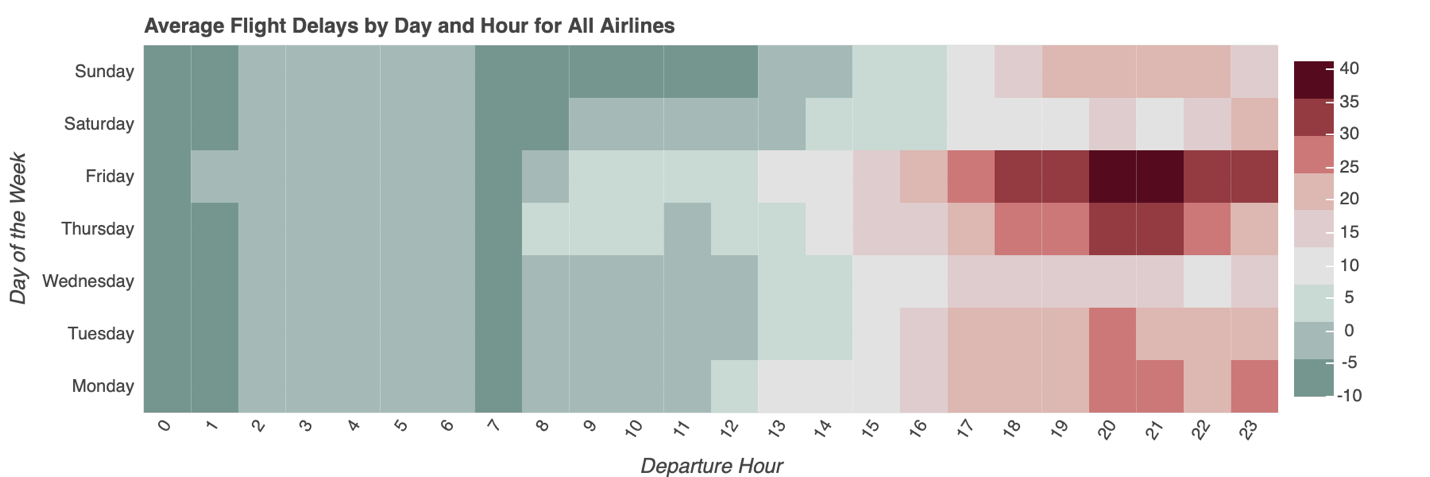
We can further breakdown the delay times by day and hour to get a full picture of performance of all airlines in the NYC Area: 

Figure Heatmap of arrival delays broken out by hour and day of week

It looks like NYC Area sees an increased delays during Thursday and Friday afternoons/evenings. This could be due to increase in passenger loading during those hours and days or more flights. We can look into this next.



Figure Number of Departing Flights in the NYC Area Throughout the Week

There does not seem to be any spikes during evening / night hours for Thursday and Friday. Scheduling seems relatively consistent.

Let’s take a look at the passenger loading data by month and see if there is a trend relative to delays:

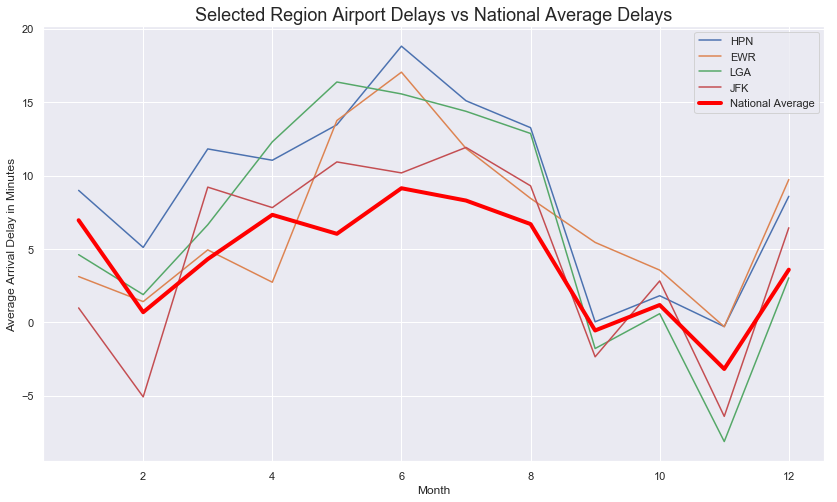
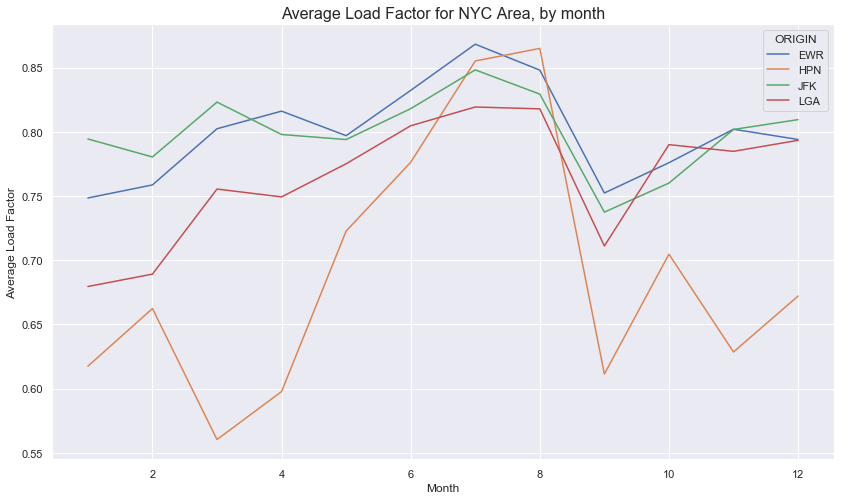
 

Figure Average Delay (left) and Passenger Load Factor (right), by month

There seems to be a trend between delay and passenger loading, the more passengers fly, the higher the average delay. We can break this down further by asking: Do the delays increase for specific flights with higher load factors?

* 1. Individual Route Analysis

Now we can take a look at specific routes and their historic performance throughout the year. Are there routes that are prone to delays and are there ones that are always on-time? We’ll first filter and create a data frame that we can use to answer our questions:

First we’ll group the data by airline, origin, and destination airport, to see what are the most flown routes for the specific airline and then add the average delay time for each route, we can do this for on-time flights, as well:

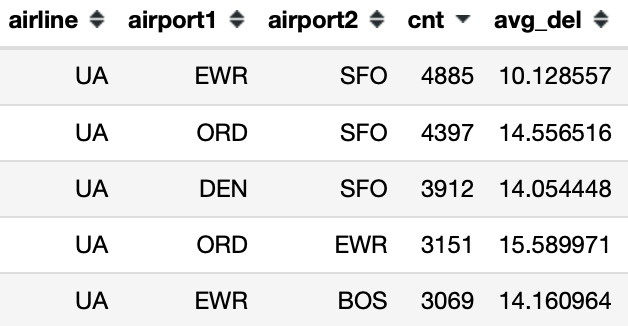
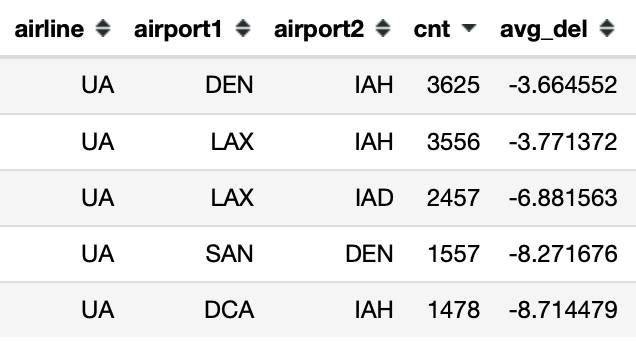
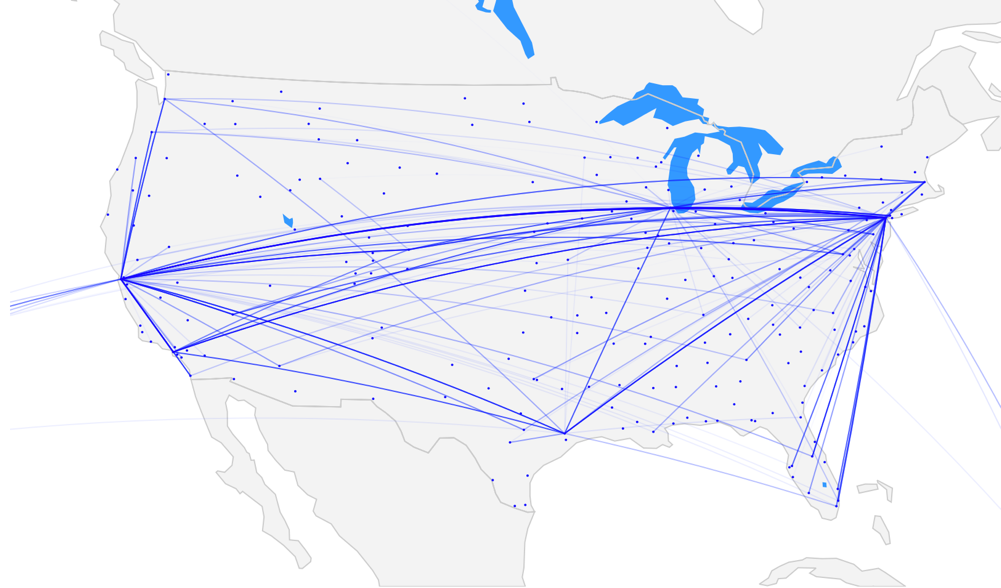
 

Figure Most Delayed and On-time Routes for United Airlines

Now we can merge the data with latitude and longitude information as described in section 2.2 and visualize what this looks like on a map.

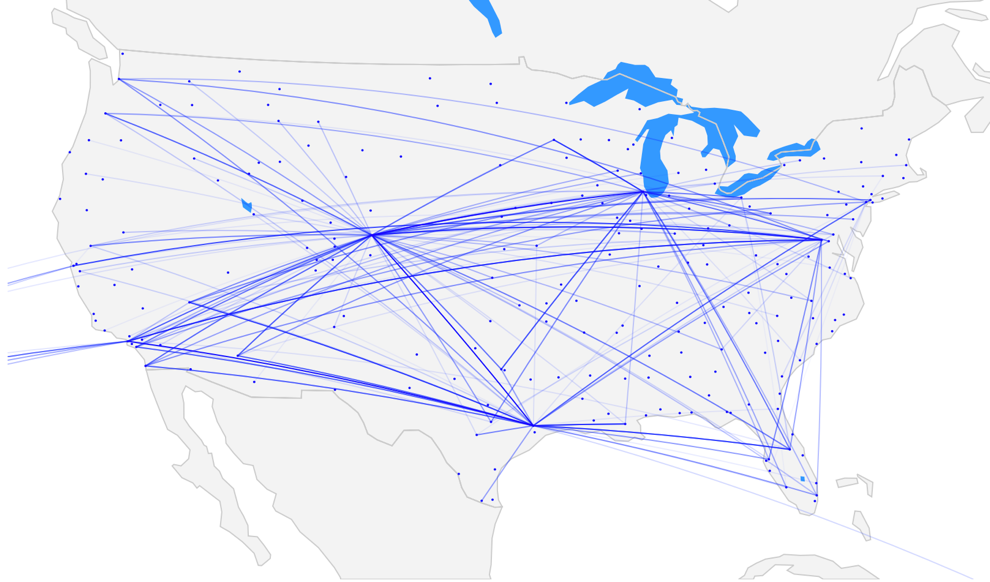


Most Frequently Delayed Routes

Figure: 3.16 United Airlines Most Delayed(top) and Most On-time(bottom) flight routes in 2017.

The averages are weighted based on number of flights for the particular route. Darker lines represent higher delay frequency.

There is a clear difference in performance based on the route chosen. Coast to coast or long-haul flights tend to represent routes that are most delayed. This will be considered during the modeling stage. As simple average delay will not isolate problematic routes. This can be seen in Figure 3.1 and the distribution of arrival delays based on length of flight.



Most Frequently On-time Routes

Delta Airlines

Most Frequently Delayed Routes

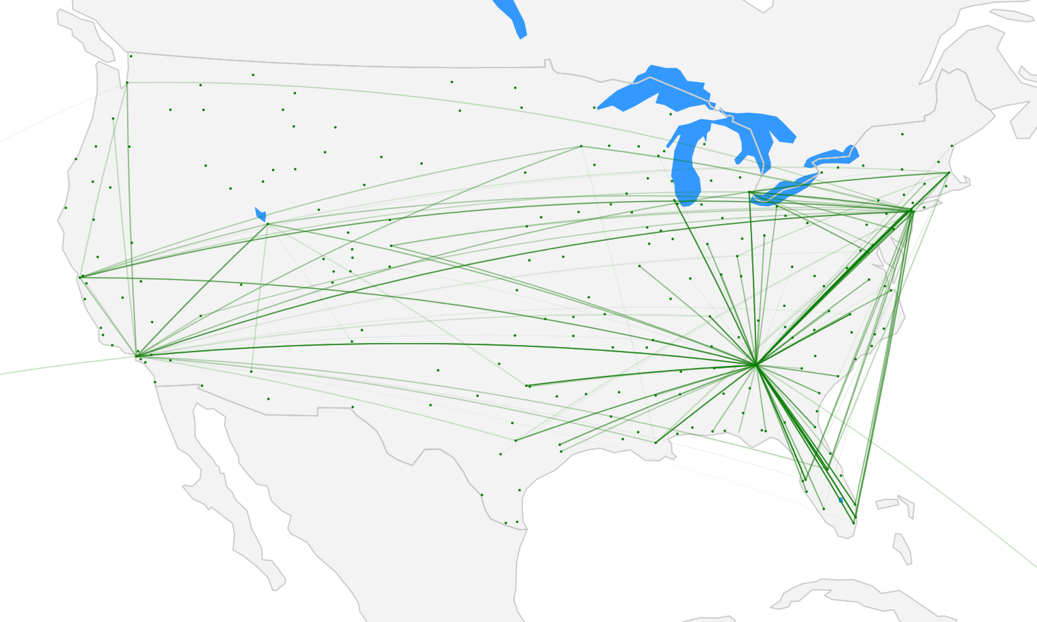
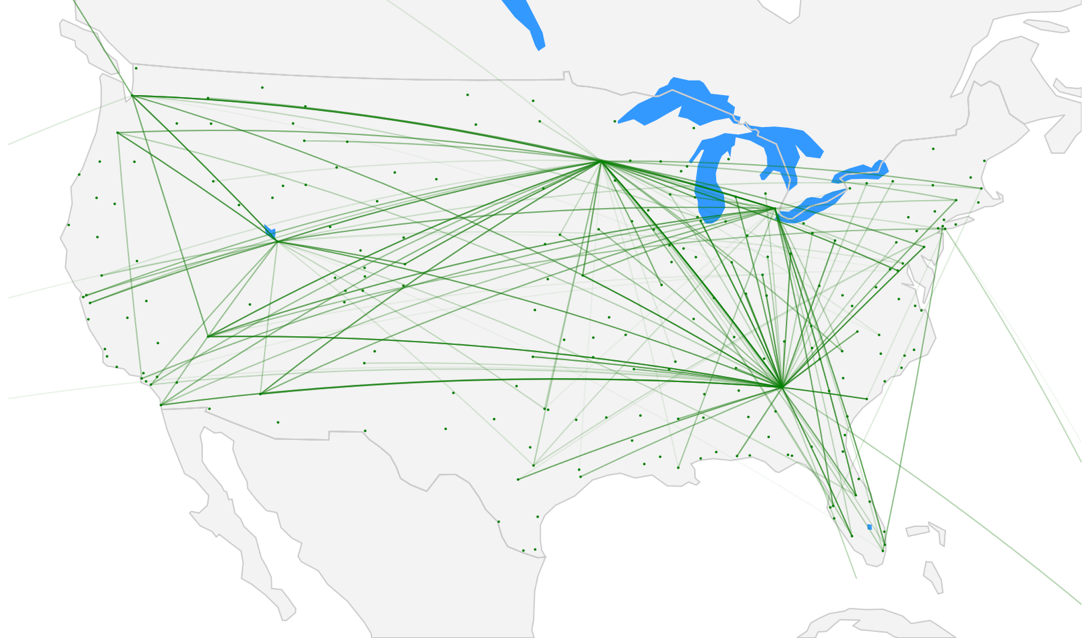


Figure: 3.17 Delta Airlines Most Delayed(top) and Most On-time(bottom) flight routes in 2017

We see the same pattern for Delta Airlines, long-haul flights seem to dominate most delayed routes. Additionally, Delta has three hubs with exceptional performance: MSP, DTW, and SLC. Which do not appear on the delayed route visualization.



Most Frequently On-time Routes

Is there a link between origin airports / destination airports and these delayed routes?

Do other airlines have similar performance for a given route?

Does a given dominate a given airport?

1. **EDA Conclusions**

The intent of the EDA was to find any patterns in the data by creating visualizations which can later be used to create additional features and consequently a more robust model. We found a lot of trends and insights during this EDA but when we find these insights it’s important to stop and ask why are we seeing these trends? What is the underlying cause? We can’t just apply this to future predicting without knowing the cause.

Possible questions to answer from EDA:

From route analysis:

1. Is there a link between origin airports / destination airports and these delayed routes?
2. Do other airlines have similar performance for a given route?
3. Does a given dominate a given airport?

From Day / Hour breakdown of delays heatmap:

1. Why are certain days / hours more prone to delays?
2. Is it congestion? Number of scheduled flights increased?
3. Number of travelers? Can we obtain passenger loading information?

From flight delay percentage comparison between Bay Area and National:

1. Why does the Bay Area have more delays than the national average?
2. Is there a statistically significant difference between number of delays in the Bay Area vs National average?
3. What region has lowest percentage of flight delays?

From the delayed percentage chart by airport:

1. Why do EWR, SFO, and JFK have the highest percentage of delayed flights?
2. Do these airports have most coast to coast flights? If so, why would these flights be delayed?
3. International traffic was not considered for this analysis which could have a huge impact on airport loading, especially major airport hubs such as EWR, SFO, and JFK.

Datasets

Loading Data

<https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=311>

Complete Datasets

<https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236>

Weather:

<https://mesonet.agron.iastate.edu/ASOS/>

Airline Info (automatic download)

<https://www.transtats.bts.gov/Download_Lookup.asp?Lookup=L_UNIQUE_CARRIERS>

Airport Coordinates:

FAA OPSNET:

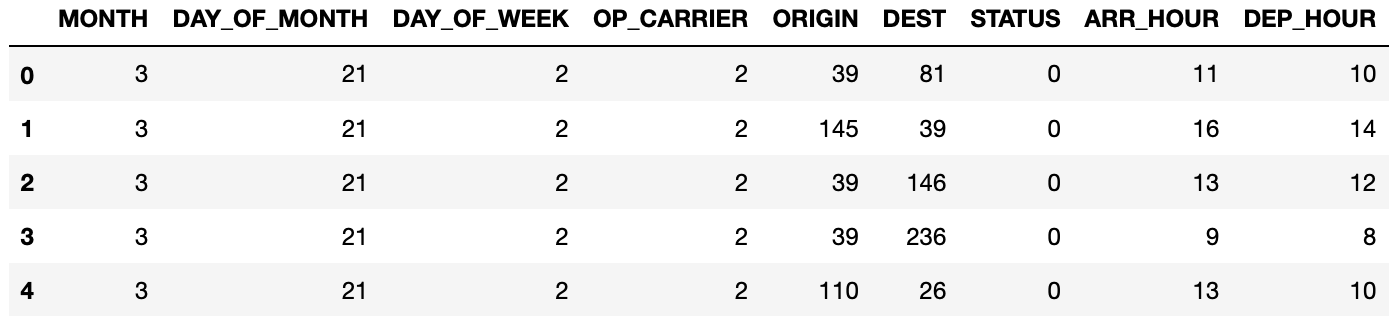
<https://aspm.faa.gov/opsnet/sys/Delays.asp>

1. **Modeling**

We’ll initially approach this problem using RandomForestClassifier with n\_estimators=25 and an 80/20 test/train split to predict if flight is delayed or not, along with the probability.

Original run is on the entire 2017 dataset, with no feature engineering, only binning departure and arrival hour blocks:

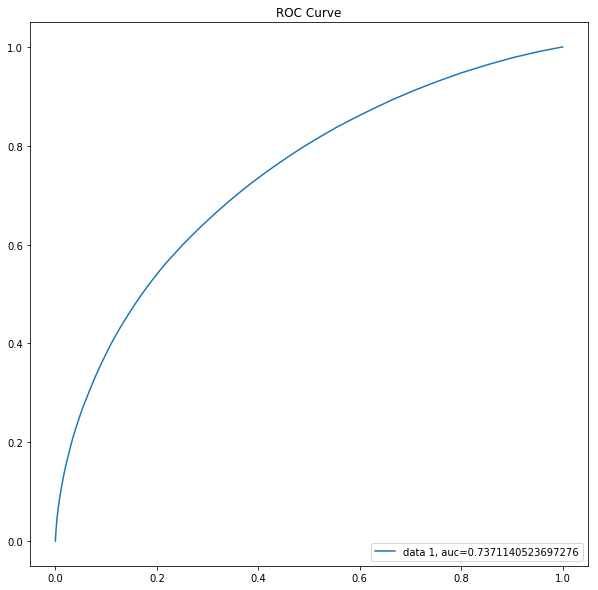
Accuracy=82.5% AUC=0.737



Confusion Matrix:

[874265, 36085]

[158869, 46663]



Feature Importance:

DEST 0.218760

ORIGIN 0.214175

DAY\_OF\_MONTH 0.193794

MONTH 0.086238

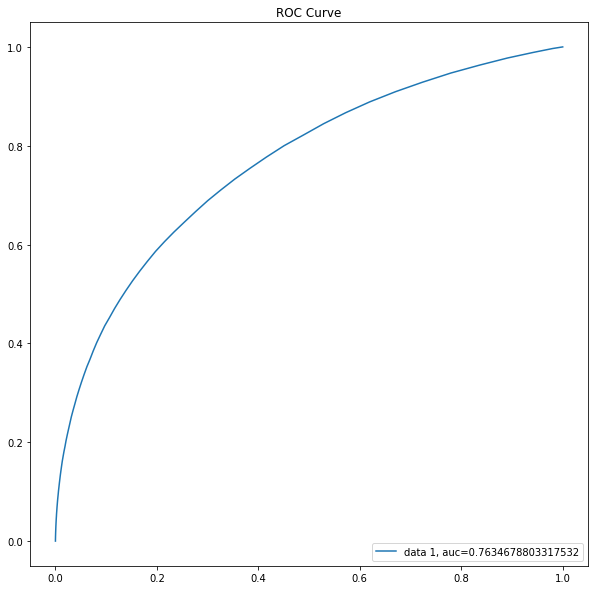
ARR\_HOUR 0.080972

STATUS 0.078484

DAY\_OF\_WEEK 0.076678

OP\_CARRIER 0.050899

1. **Feature Engineering**



Hour\_bins

Flight number

Plane tail number

Direction of flight

Narrow down to top 20 origin airports and top 5 airlines by volume

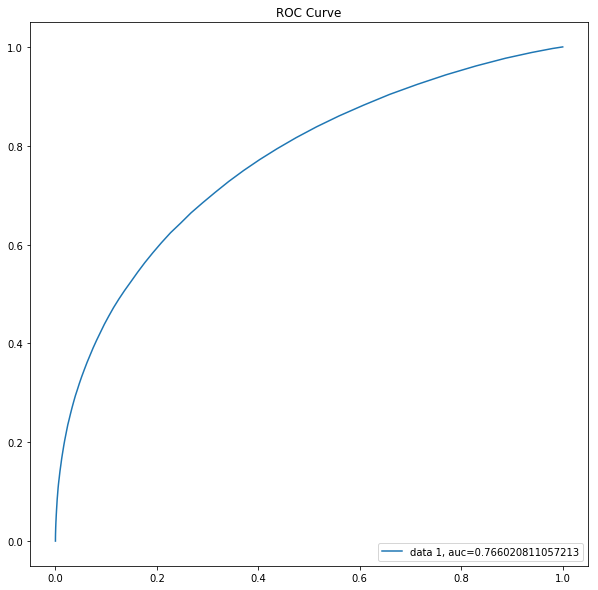
OneHotEncoder on date columns

n\_estimators=100: Accuracy=85.8% AUC=0.763

**Confusion Matrix:**

[409244, 5010]

[ 64382, 11527]



Precipitation data at origin for last 4 hrs

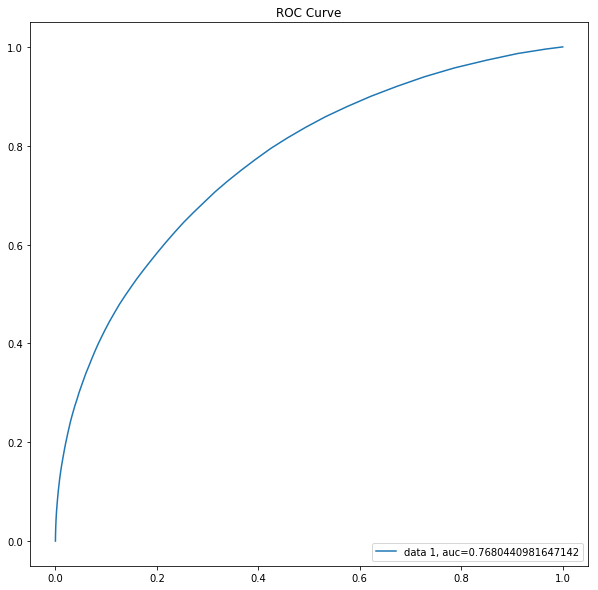
Number of flights scheduled for any given hour at the origin airport

n\_estimators=**100**: Accuracy=86.1% AUC=0.766

**Confusion Matrix:**

[408622, 5632],

[ 62604, 13305]



Monthly flight delay frequency aggregated by flight number.

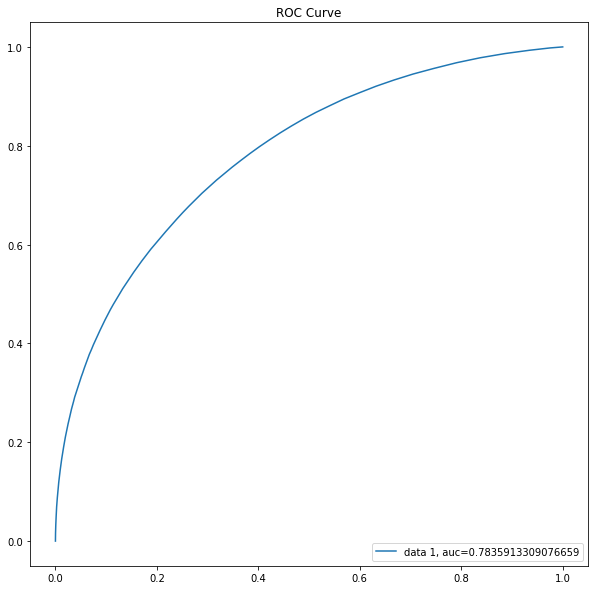
n\_estimators=**100**: Accuracy=85.9%

AUC=0.768

**Confusion Matrix:**

[407393, 6861],

[ 62341, 13568]



Monthly flight delay frequency aggregated by flight number (including origin and destination) in order to fine tune the avg. arr\_delay per flight number

\*Feature scaling

n\_estimators=**150**: Accuracy=86.1%

AUC=0.784

**Confusion Matrix:**

[406653, 7601],

[60396, 15513]