

NYC 2013 Flight Analysis

Christopher Peralta

September 15, 2018

In this project, I will analyze the 2013 flight data for New York City and build a model for predicting flight delays. I will begin by getting all of the time data into a more usable format, then I will begin analyzing the data, and finally I'll build the model.

Before I begin, I'll tell you a little about the data.

- airlines is a list of airlines and their abbreviations
- airports is a list of airports with their locations, timezones, and faa codes
- flights is a list of all flights that departed NYC in 2013 with other related data
- planes is a dataset of all of the planes that went on the flights above
- weather is a dataset of the weather conditions by hour and airport for the year of 2013

Below are some summary statistics for the flights dataset.

```
#>    time_hour
#> Min.      :2013-01-01 05:00:00
#> 1st Qu.:2013-04-04 13:00:00
#> Median :2013-07-03 10:00:00
#> Mean    :2013-07-03 05:02:36
#> 3rd Qu.:2013-10-01 07:00:00
#> Max.    :2013-12-31 23:00:00
#>
#>    air_time    arr_time
#> Min.   : 20.0   Min.    :  1
#> 1st Qu.: 82.0   1st Qu.:1104
#> Median :129.0   Median :1535
#> Mean   :150.7   Mean    :1502
#> 3rd Qu.:192.0   3rd Qu.:1940
#> Max.   :695.0   Max.    :2400
#> NA's   :9430    NA's    :8713
#>    dep_time
#> Min.   :  1
#> 1st Qu.: 907
#> Median :1401
#> Mean   :1349
#> 3rd Qu.:1744
#> Max.   :2400
#> NA's   :8255
```

There are a few interesting observations above:

- `time_hour` has no clear meaning (it contains times rounded down to the nearest hour for joining the weather data with flights)
- `air_time` appears to be in minutes
- Arrival and departure times are given in a 4-digit format

It's also worth noting that all departure times are in the US Eastern timezone and the arrival times are in the timezone of the local airports.

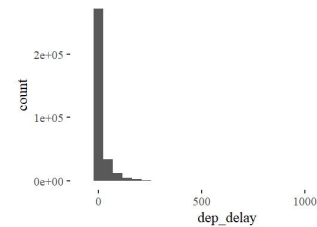
model: `dep_delay = season + weekday + month + sched_dep_time + carrier + tailnum + dest + distance`

Departure delays

Distribution

Let's start by looking at the distribution of departure delays.

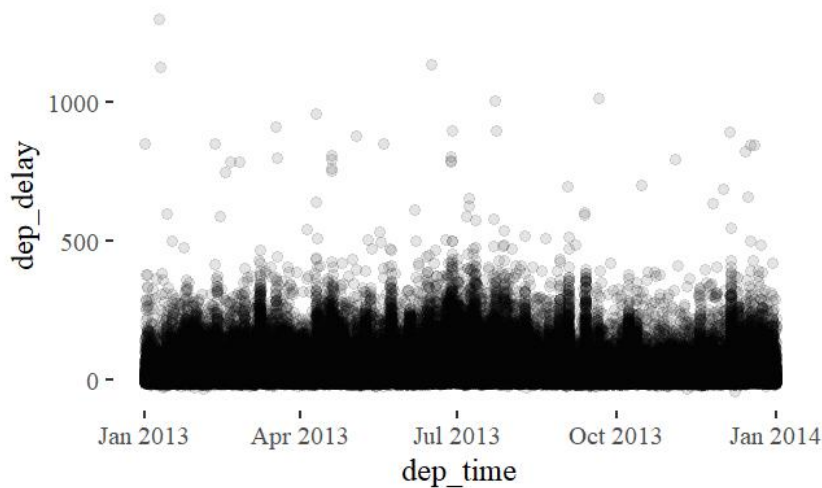
Most departure delays appear to be relatively short, with relatively few long delays.



Delays through the year

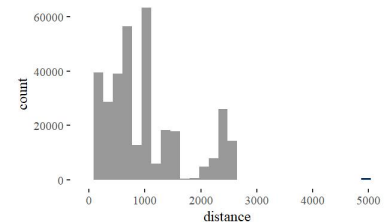
Below, we can see departure delays throughout the year of 2013.

There seems to be a dip in departure delays between September and December. It appears that the delays vary with season.

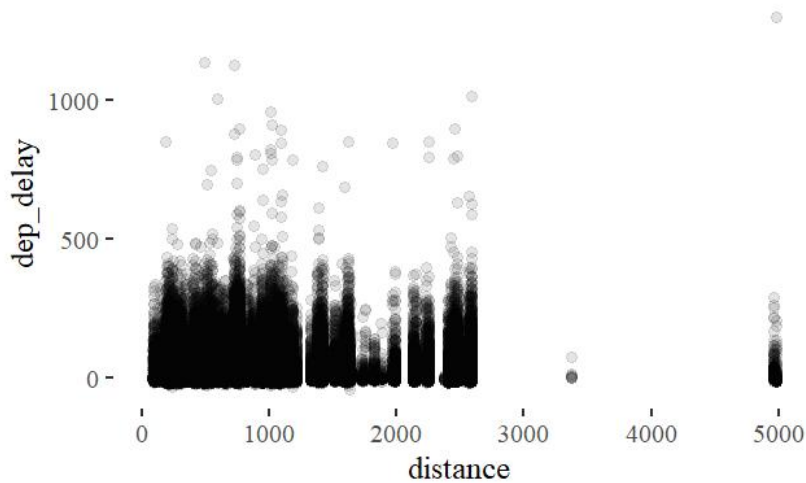


Distance of flights

As you can see in the histogram, most flights are under 2,000 miles away. Additionally, all of the furthest flights are to Honolulu, Hawaii and they are colored in blue.

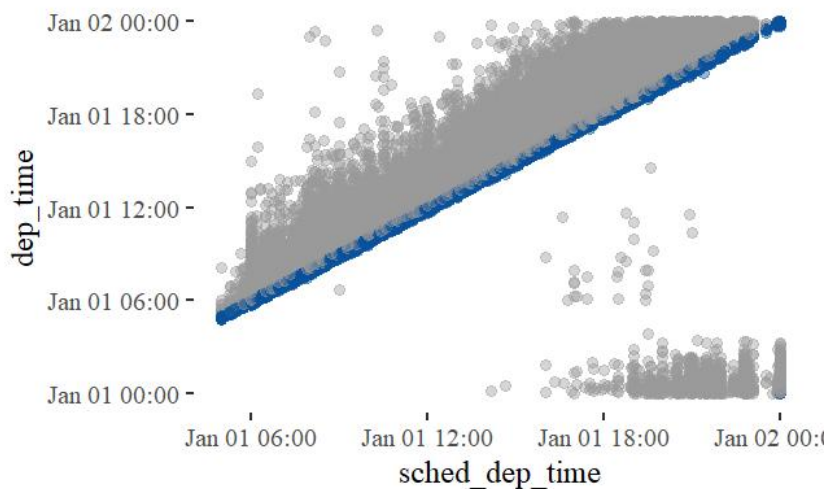


Are departure delays affected by the distance of the flight? According to the plot below, it seems quite unlikely that distance significantly affects the departure delays.



Scheduled times and actual times of departure

Below, we can see a scatter plot of scheduled departure times versus actual departure times. Most flights appear to leave New York on time, or with slight delays. The flights in the bottom right corner are flights that left the day after they were scheduled. All flights with under 2 minute delays are colored in blue.



Delays if all flights left on the same day

Below, is a plot of the scheduled departure times versus the departure delays. It shows that flights that leave later usually have longer

delays, where earlier flights typically have shorter delays.

