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 arrival delays. I will begin by analyzing the data, followed by making a model to predict arrival delays.

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

- `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

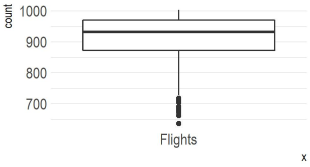
Before I begin, it's also worth noting that all departure times are in the US Eastern timezone and the arrival times are in the timezone of the local airports.

What can you use this for?

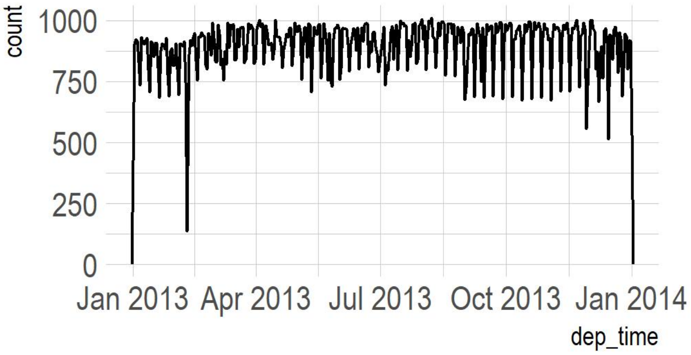
The model I built has a variety of possible uses. It can give us an idea as to what variables affect the arrival delays the most. It could be used as part of an in-flight system to give a better idea of arrival times at the start of the start of the flight, although I believe GPS-based estimates will likely be more accurate.

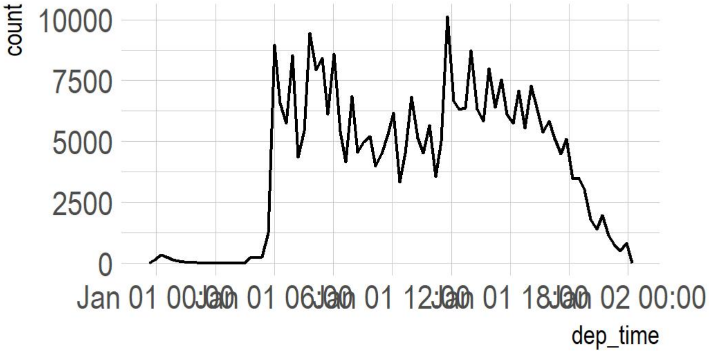
How people fly?

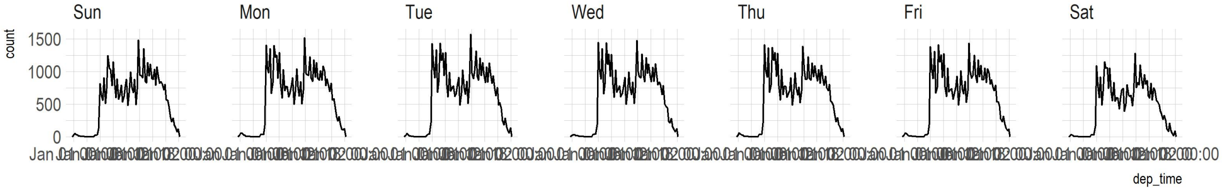
From the chart below, we can see that flights throughout the year are quite consistent. Additionally, in the boxplot to the right, we see that the distribution of flights appears close to normal.



What time do people fly?

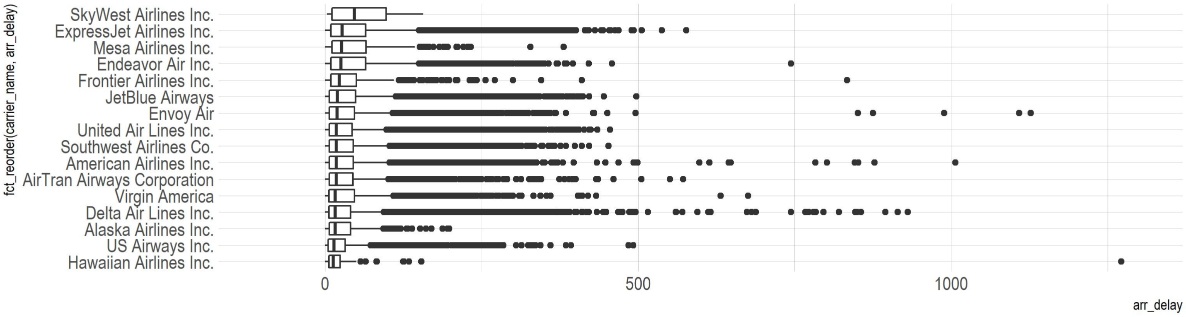


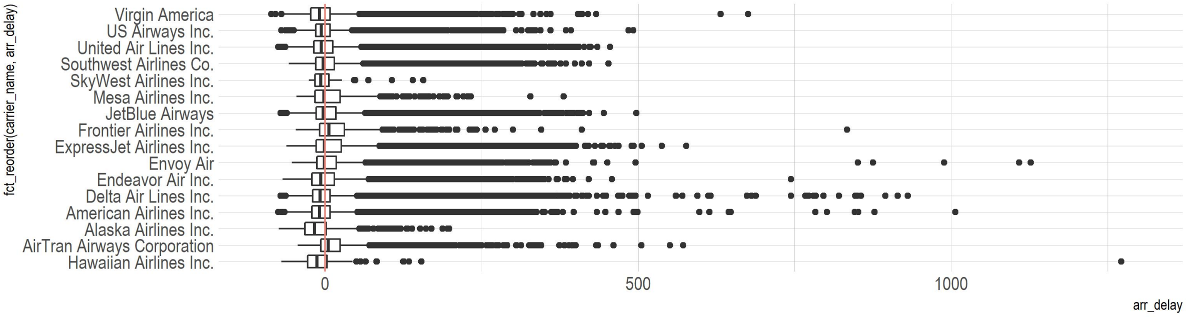




Airlines

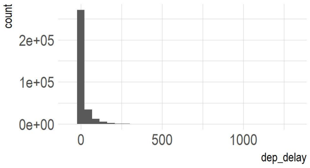
Below you can see the distribution of positive arrival delays by airline sorted by median.





Arrival delays

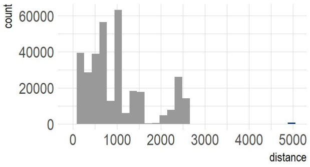
Distribution



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

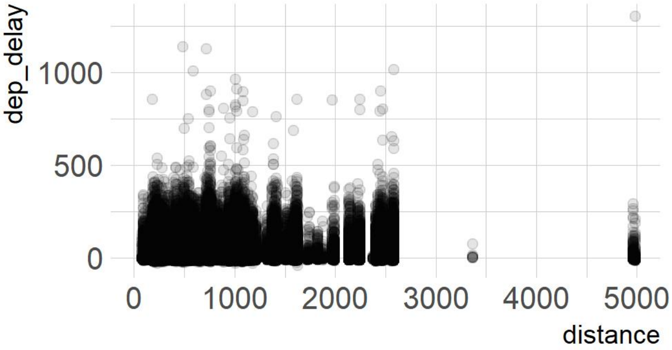
Most departure delays appear to be relatively short, with relatively few long delays.

Distances of flights



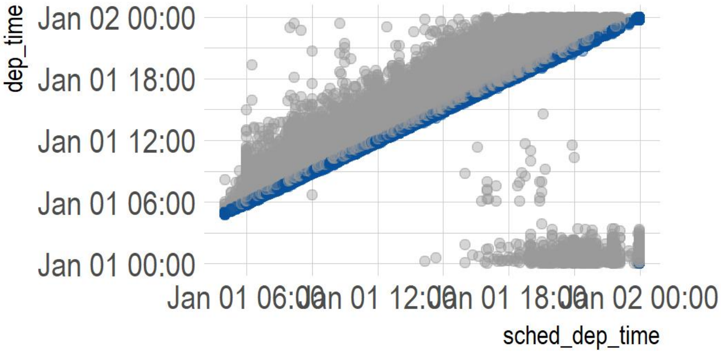
As you can see in the histogram, most flights are under 2,000 miles away. Additionally, all of the furthest flights are to Honolulu, Hawaii and they are colored in blue.

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.



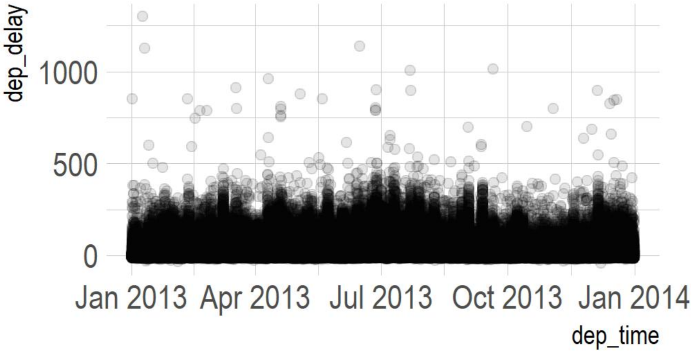
Scheduled times and actual times of departure

Below, we can see a scatter plot of scheduled departure times versus actual departue 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.



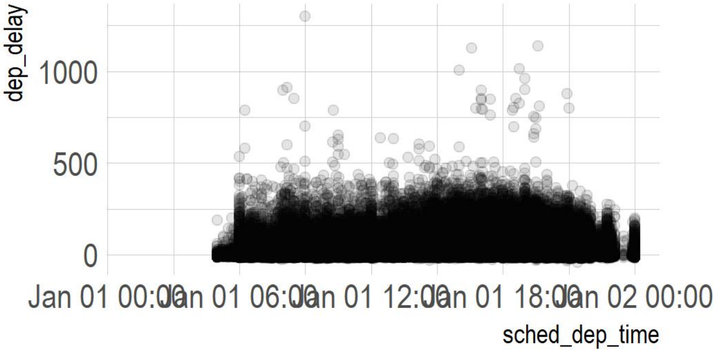
Delays with respect to time

Below, we can see departure delays throughout the year of 2013. There seems to be a dip in departure delays between September and December. This leads me to believe that the delays vary with season.



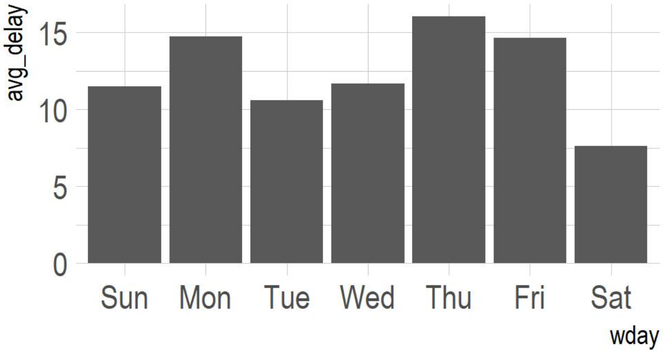
Delays if all flights left on the same day

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.

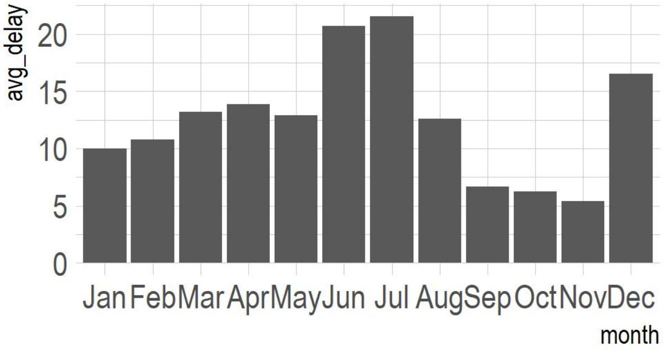


Delays if all flights happened in the same week

Here is a plot that shows the average delay on each day of the weak. The average delay on Saturday is quite low.



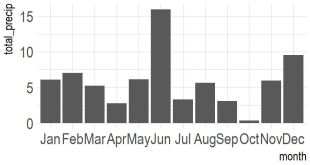
Delays by month

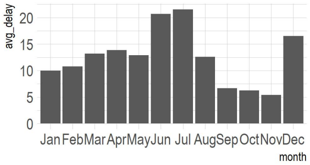


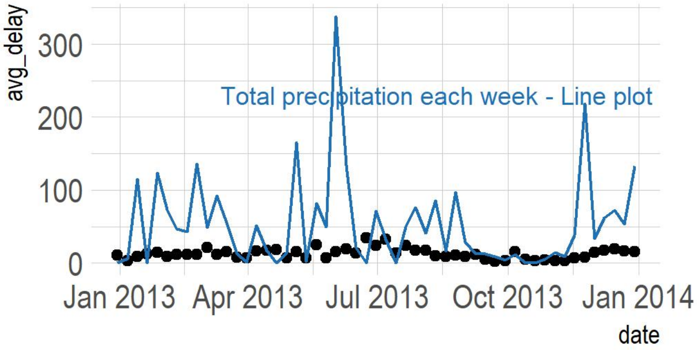
The seasonal trend seems to be quite significant. This implies that weather has a strong affect on departure delays. My guess is that June and July are the rainiest months in the year, and that December is the snowiest. In the next section, I will look at the weather for the year of 2013.

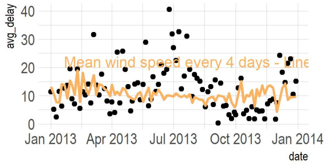
Weather

I started by checking if the months with the most precipitation do, in fact, have the longest departure delays. Which as shown in the figures to the right, is true.









It seems somewhat likely that precipitation has some sort of effect on departure delays. However, it appears that wind speed does not have any effect as shown in the plot to the right.

Is there a non meteorological seasonal effect?

It's possible that there is a seasonal effect that isn't based on the weather. People fly more often at certain times of the year and certain days. Reasons for this include holidays, work, and weather.

The model is quite bad, if we exclude arr_delay. But if we include arr_delay, then the model is useless in practical scenarios because anyone who is using it won't have access to the future information

Why the model is bad

https://www.bts.gov/topics/airlines-and-airports/understanding-reporting-causes-flight-delays-and-cancell
##

How the model could be improved.

Adding more data

One likely cause for delays is the airplane arriving after the scheduled arrival time on its previous flight causing a departure delay for its next flight. https://www.bts.gov/topics/airlines-and-airports/understanding-reporting-causes-flight-delays-and-cancellations

Using a more complex model

A more complex model could probably do a better job predicting arrival delays, but its uncertain how much it could improve the predictions with the current data.

Modifying the problem

The underlying problem could be modified. In other words, rather than predicting the duration of the flight delay, we could predict whether or not there will be a significant delay.