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                      air_time        arr_time   
#>  Min.   :2013-01-01 05:00:00   Min.   : 20.0   Min.   :   1  
#>  1st Qu.:2013-04-04 13:00:00   1st Qu.: 82.0   1st Qu.:1104  
#>  Median :2013-07-03 10:00:00   Median :129.0   Median :1535  
#>  Mean   :2013-07-03 05:02:36   Mean   :150.7   Mean   :1502  
#>  3rd Qu.:2013-10-01 07:00:00   3rd Qu.:192.0   3rd Qu.:1940  
#>  Max.   :2013-12-31 23:00:00   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  
```

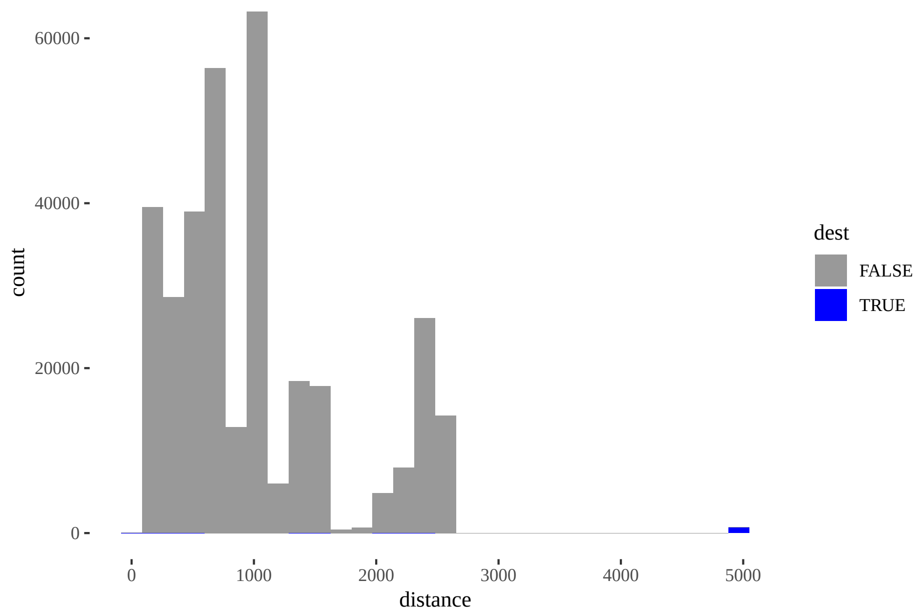
There are a few interesting observations above: - `time_hour` has no clear meaning (it contains times rounded down to the nearest hour for joining the `weather` data with `flights`) - `air_time` aappears to be in minutes - Arrival and departure times are given in a 4-digit format

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

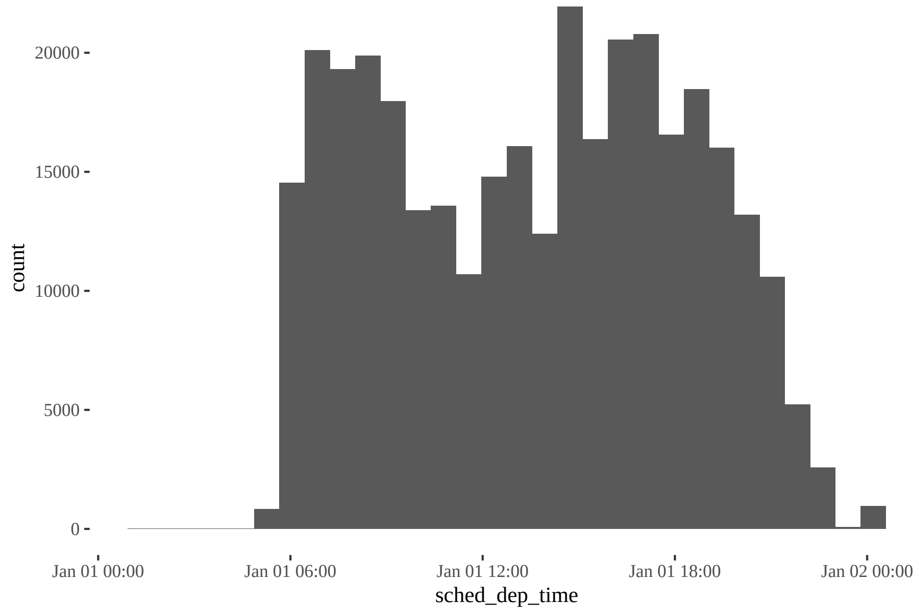
As you can see below, the longest flights are to Honolulu, Hawaii

```
#> `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

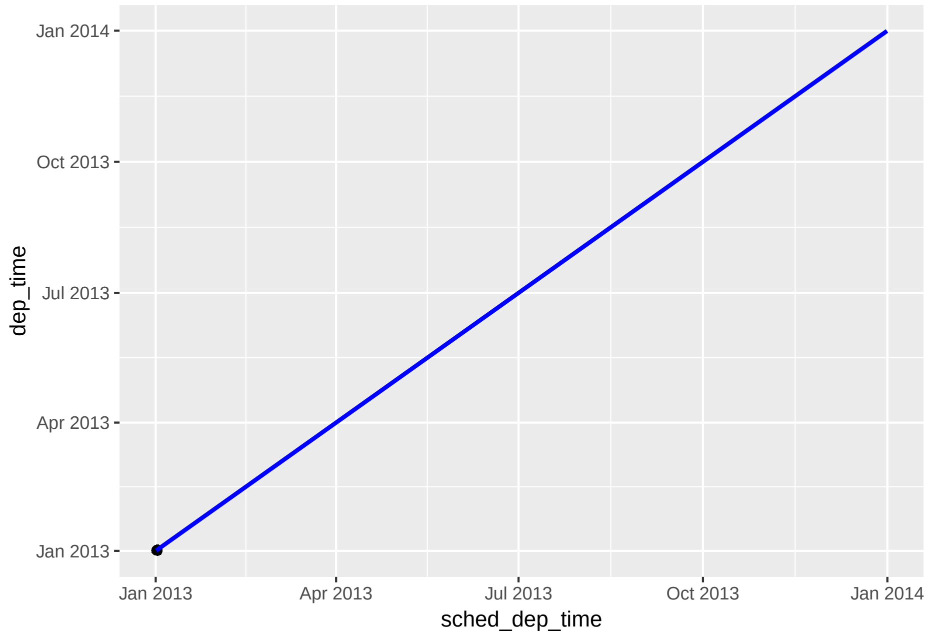


Above we can see the distances between flight origin and destination. It's worth mentioning that the longest flights (around 5000 miles) are all to Hawaii.

```
#> `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
#> `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

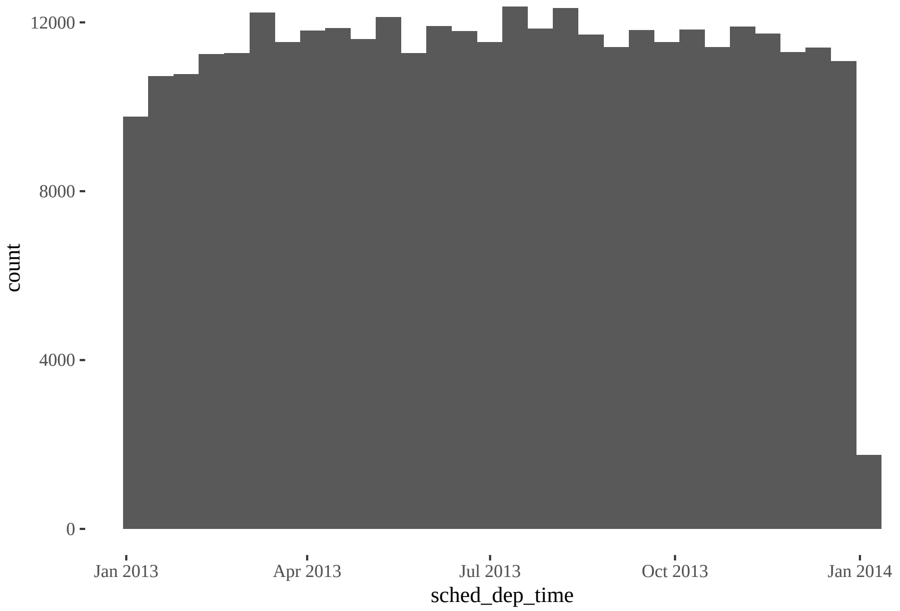


Above we can see histograms of the actual departure time and the scheduled departure time.



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.

```
#> `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
#> `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


count
Jan 2013
Apr 2013
Jul 2013
Oct 2013
Jan 2014
sched_dep_time