Data Analysis Demonstration with NYC Flight Data

Kaan Aksoy

2023-09-04

Summary

This is a demonstration of exploratory data analysis (EDA) and predictive statistics using R. For the purposes of this demonstration, I use the data provided by the package `nycflights13` in order to ensure high-quality data. In reality, data may require additional cleaning, tidying, and general wrangling. I use this data in order to focus on the predictive statistics aspect of R without having to worry about data quality affecting model results.

The Data

First, I look at the general characteristics of the data. These can give us useful insights.

```
## [1] 336776     19
```

From here, we understand that there are 336,776 observations and 19 variables in the data. This means that we are dealing with data from 336,776 flights. A next step would be to understand which time frame these flights belong to.

```
## [1] "2013-01-01 05:00:00 EST" "2013-12-31 23:00:00 EST"
```

The variable `time_hour` is provided in a year-month-day, hour-minute-second format (more technically, as a POSIXct vector). In real analyses, this data may not be as precise. Using the `range` command, we make R give us the minimum and the maximum value of the supplied vector. As a result, we now know that we are looking at flights in 2013, starting from January 1 and ending on December 31.

We might be interested in other aspects of an airport, such as delays, the variation in how much time a flight spends in the air, or distance between airports.

Table 1: Descriptive statistics for NYC flights

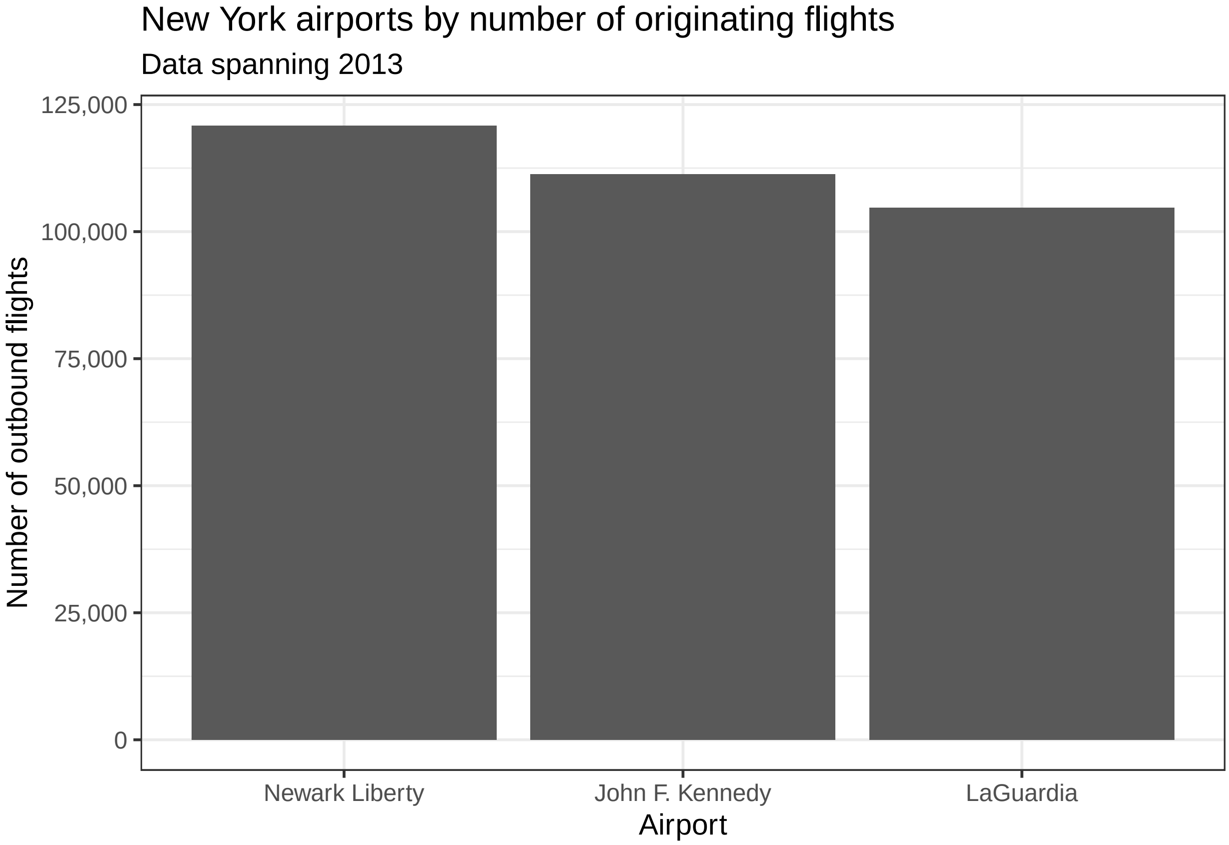
	Mean	Median	SD	Max	Min	N
Distance	1039.91	872.00	733.23	4983.00	17.00	336776
Air Time	150.69	129.00	93.69	695.00	20.00	327346
Departure delay	12.64	−2.00	40.21	1301.00	−43.00	328521
Arrival delay	6.90	−5.00	44.63	1272.00	−86.00	327346

Data from 2013.

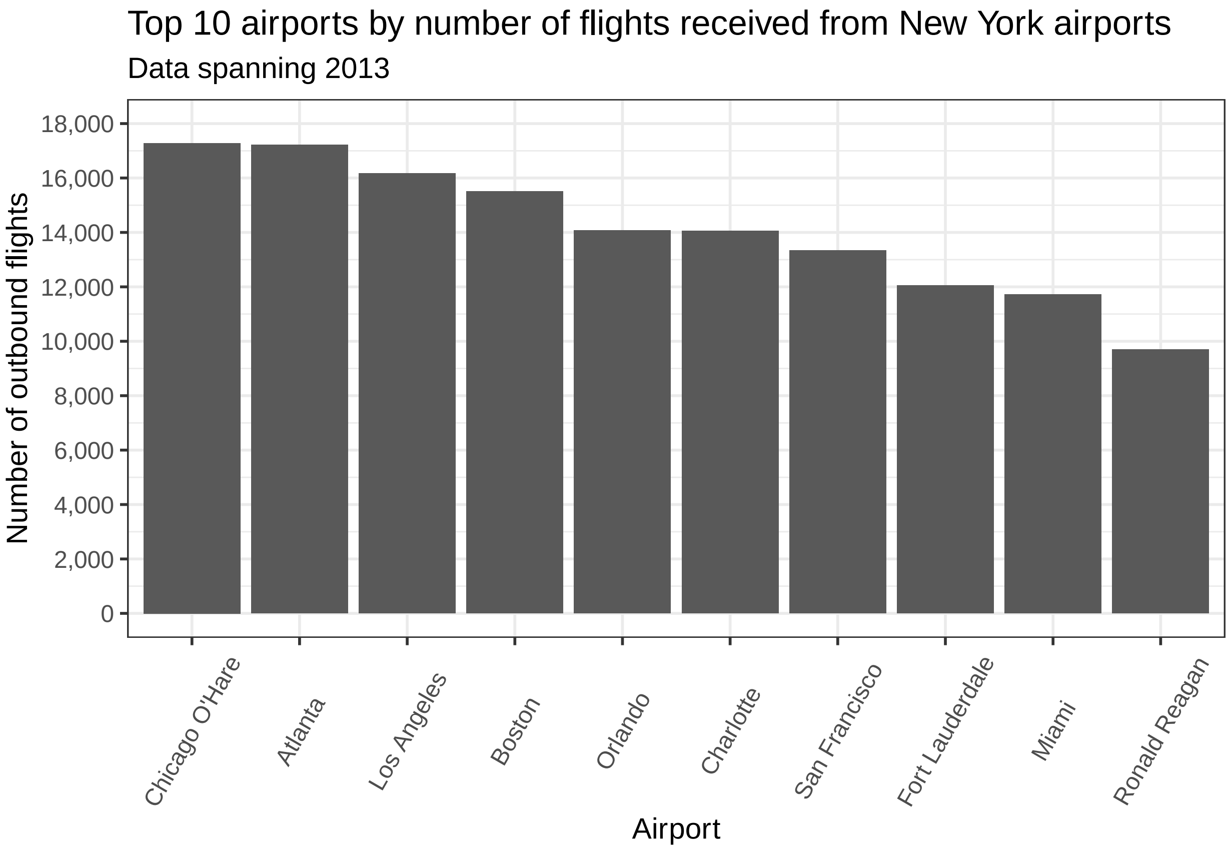
Negative values for departure and arrival delays indicate early departures/arrivals (in minutes).

Immediately, we see that the median flight flies 872 miles and spends 129 minutes in the air, departs two minutes early and arrives five minutes early. Meanwhile, the highest distance flown is 4,983 miles, and the most time spent in the air is 695 minutes, the longest departure delay is 1,301 minutes (or about 22 hours), while the longest arrival delay is 1,272 minutes (approximately 21 hours).

We might also be curious about which airports have the most inbound and outgoing flights:



Interestingly enough, we see that in 2013, Newark Liberty International Airport had more outgoing flights than John F. Kennedy International Airport. We can also look at which airports receive the most flights from NYC airports.

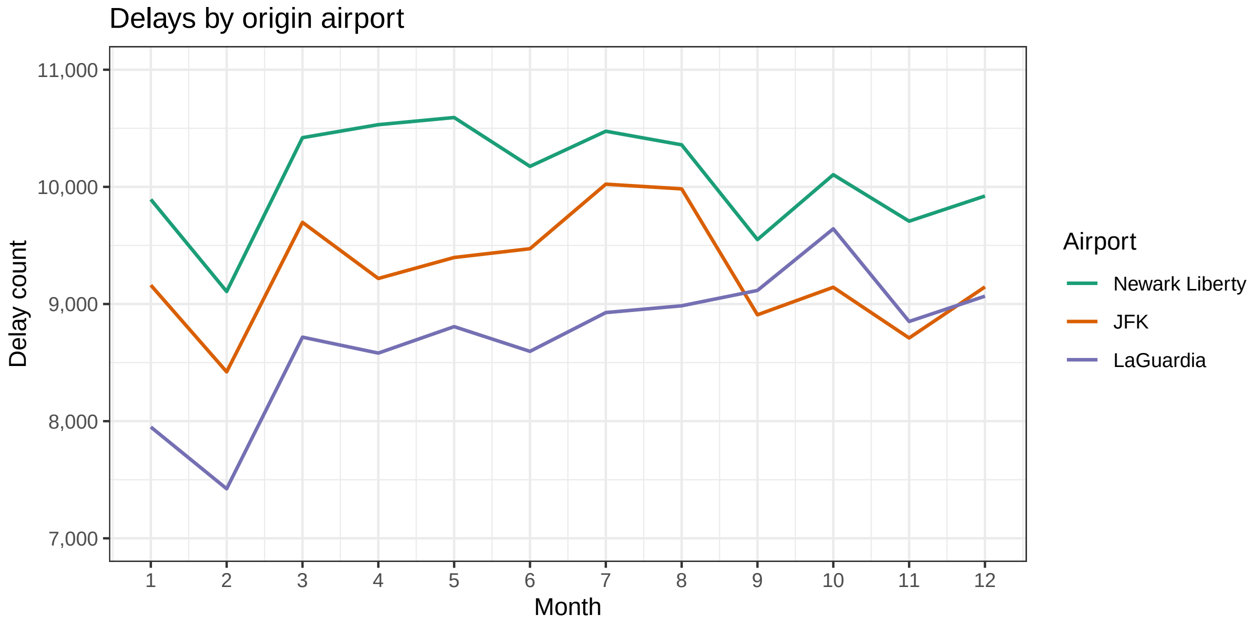


We see that Chicago's O'Hare International Airport receives the most flights from NYC airports, followed closely by Atlanta International Airport in Georgia. The graph is restricted to the top 10 recipients of flights from NYC airports for readability purposes.

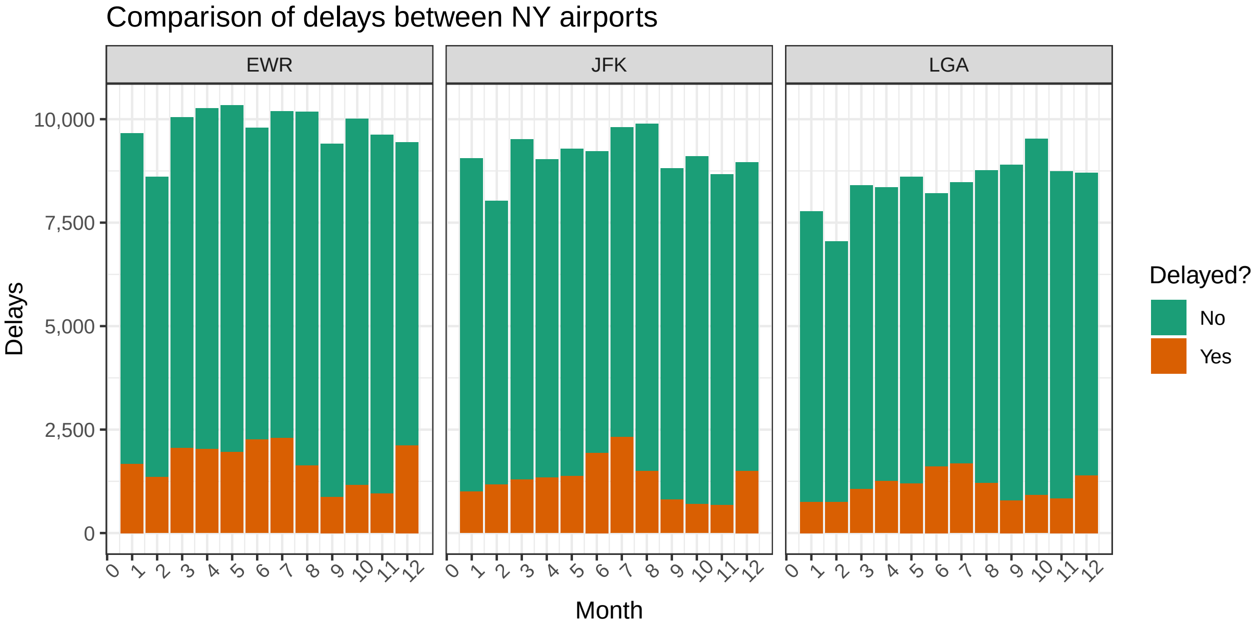
Understanding Delays

If we operated an airport, or were otherwise interested in airlines, we might be very curious about what causes delays and perhaps how to mitigate them, or at the very least be able to forecast them so that we can take pre-emptive action to mollify unhappy passengers. Considering that we are dealing with aircraft, we might imagine that weather conditions greatly impact how long flights take and whether they are substantially delayed.

First, we can see the number of delays per airport in New York:



We see that delays increase significantly in March for all three airports. We can also see what proportion of flights in each airport are substantially delayed. We might consider a delay of 30 minutes to be substantial, while a five-minute delay may be inconsequential for most situations. Therefore, let us assume that a delay of 30 minutes or more is relevant to us (and to passengers).



We can now begin to consider the determinants of delays. The `nycflights13` package also provides weather data. Naturally, if we were an airport, we would need to obtain this data independently (either from government agencies collecting such data, or on our own—most likely the former).

I now merge the datasets `flights` and `weather` in order to match weather observations with flight data. This is where data wrangling comes in, as there are an unequal number of observations and variables between the two separate dataframes:

```
## [1] 336776     19
```

```
## [1] 26115    15
```

The `flights` data has 336,776 observations while `weather` only has 26,115 observations, which will necessitate good use of join functions to merge them appropriately without creating duplicate or "rubbish" data. I name the new data frame `df1`. I also add in a new variable, `delay`, which is a binary (or dummy) variable indicating whether the flight was delayed or not. Consistent with the prior part of this demonstration, I refer to a flight as delayed only if the departure delay is 30 minutes or longer.

```
## [1] 336776     29
```

The resultant data frame has 336,776 observations and 29 variables. What is important here is that the number of observations is the same as the data frame `flights`, which indicates that the merging process has not created duplicate or rubbish data which could distort our prediction models.

Linear Regression: Predicting Delay Time

With the necessary data at hand, we can now start considering which factors influence departure delay times. We can imagine that wind speed, gust speed, precipitation, pressure, temperature, and visibility are all important factors in predicting delays. Therefore, we can model this, and receive the following output from a linear regression model:

	(1)
Wind speed (mph)	—0.109*
	(0.059)
Gust speed (mph)	0.327***
	(0.051)
Humidity	0.212***
	(0.010)
Precipitation (inches)	—4.661
	(11.665)
Pressure (mbar)	—0.210***
	(0.022)
Visibility (miles)	—1.828***
	(0.155)
Temperature (F°)	0.121***
	(0.008)
Intercept	220.527***
	(22.541)
Num.Obs.	73 233
R2	0.024
R2 Adj.	0.024
RMSE	37.83

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

The R^2 of 0.024 is not very encouraging. It tells us that our regression model explains only 2.4% of the variation in departure delays, which is a very small amount. However, we should also consider that many, many things can influence departure delays, and that our model accounts solely for weather conditions and absolutely nothing else.

Logistic Regression: Predicting Delays

We can also build a logistic regression (or logit) model. In this model, we try to understand whether a flight will be delayed or not depending on the factors which we think can lead to a delayed flight. In this case, we are not interested in the *number of minutes* by which a flight is delayed; we are only interested in whether a flight is delayed or not.

	(1)	(2)
Wind speed (mph)		0.001
		(0.004)
Gust speed (mph)		0.016***
		(0.004)
Humidity		0.013***
		(0.001)
Precipitation (inches)		0.222
		(0.652)
Pressure (mbar)		−0.018***
		(0.002)
Visibility (miles)	−0.097***	−0.063***
	(0.002)	(0.009)
Temperature (F°)		0.007***
		(0.001)
Intercept	−0.844***	16.001***
	(0.019)	(1.687)
Num.Obs.	326 993	73 233
AIC	274 864.9	60 810.8
BIC	274 886.3	60 884.4
Log.Lik.	−137 430.451	−30 397.379
RMSE	0.36	0.35

* p $<$ 0.1, ** p $<$ 0.05, *** p $<$ 0.01

It is important to note that the logistic regression model is a *likelihood* model, and unlike a linear regression, the outcome is restricted between 0 and 1 inclusive. Therefore, we can see that visibility actually has a substantial effect on whether or not a flight will be delayed: a one-mile loss in visibility can result in 6.3% more likelihood that a flight will be delayed by 30 minutes or more.