# NYC Flights

The first data set records over 330 000 flights that departed from New York in 2013 for 105 different destinations in the USA. There are also three metadata sets that contain the full names of the airlines, detailed information about the destinations and the hourly weather data for New York.

### 1) Are there any anomalies in the data? If so, how did you handle them?

Yes, there are anomalies.

First, there is **missing data** . This was dealt with by simply removing the rows with NA values because there was enough data to analyse nonetheless. The wind_gust column of the weather.csv file was primarily NA filled, and so this column was removed. This was seen as OK since a gust by definition is temporary and is assumed not have a significant effect on delays. Second, there were an **inconsistent number of departures and arrivals** . This either implied that many flights crashed or some data was missing. Inconsistencies between scheduled arrivals and departures and actual arrivals and departures also implied that some flights were cancelled. The rows with incomplete entries were also left out in the analysis in distance-delay analysis.Third, an **non-existant/unknown airport with FAA code ERW** was reported to be in charge of some flights. It is suspected that these entries were meant to be from FAA code airport EWR, but corrective measures were taken as there was enough data to ignore it.

### 2) How does the distance of a flight influence its delay? What might be the reason for this relationship?

The **distance did not have a linear correlation or a large influence on flight delay** . A quick regression model showed that for every mile flown, there is roughly going to be 0.003min less of a delay ~0.18s. This could be because a delay is caused primarily by other factors such as rain, adverse weather and visibility. It could also be caused by connection times and cascading effects from previous flights, which are unobservable in the given data.

### 3) Which New Yorker airport is the best and why (according to this data only)?

'Best Airport' awards are awarded primarily on customer experience, but given the data, this kind of evaluation is not feasible. Thus the airports were examined from 3 more functional, efficiency oriented metrics which were weighted equally. Final winner was calculated based on best performance across all 3 metrics. The metrics were as follows.

Metric 1: How many flights were not cancelled at each airport?
Metric 2: Mean delay per flight
Metric 3: Total flights processed

This method suggests that **La Guardia Airport is the best New York Airport** . Please refer to the submitted Jupyter notebook for specifics regarding the metric calculations.