# Introduction

In this data set, we will build a model to accurately predict the trip duration of New York City(NYC) based on the historical data provided by NYC taxi committee. In many big cities such as New York, London, Tokyo, Hongkong, Shanghai, Beijing has enormous population and complicated transportation system. Public bus and taxi play a very important role in these giant cities. It is critical that city’s administrative system build an efficient managements system for transportation operation and optimization. In this technical article, we present a model that leverage machine learning and data analytical techniques to accurately predict the trip duration of a taxi trip given large volume of historical operation data.

In terms of data analysis for NYC, this city has 120,000 taxi and 302,500 taxi drivers. With the development of new economy such as Uber, DIDI, the operation of traditional taxi service has encountered big challenge. The NYC taxi committee urge that the operation system must be more efficient and optimal such that it can be competitive with other new business using big data and AI techniques. Generally, there are three aspects for estimation of cost for a taxi trip: flat fee, trip distance and trip duration, given departure time, start and destination geolocations (latitude and longitude) and other information. Flat fees are usually constant for each vendor and trip distance can be estimated by giving start and end location in GPS. The third factor, trip duration, however, is more difficult to estimate. It is because trip during from one location to the other location depends on many aspects except distance. It is a nonlinear factor which will be determined by distance, travelling time in a day, traffic jam and city road conditions in that period. Therefore, accurately estimating trip duration can be non-trivial task. In this task, you will build machine learning models to predict taxi trip duration.

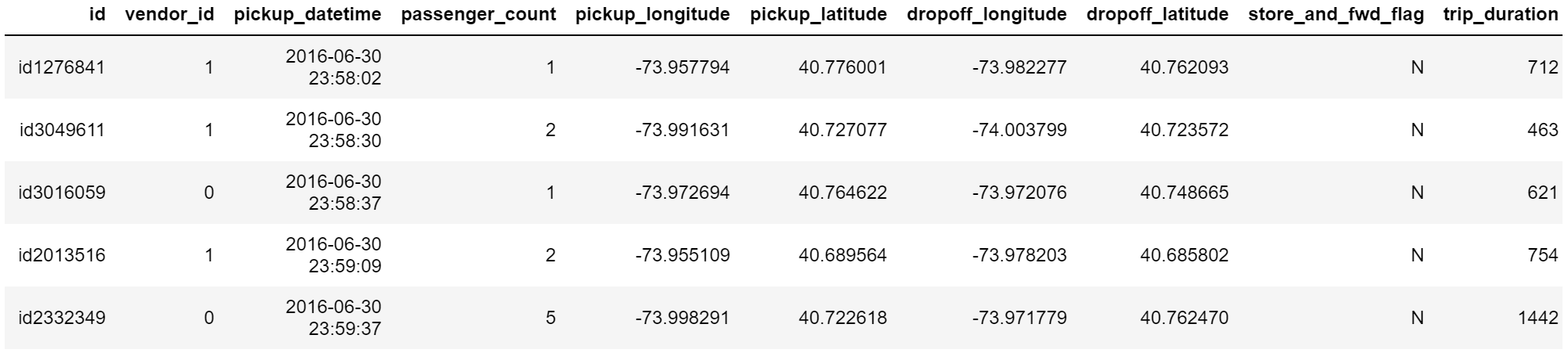
# Data description

NYC taxi operation committee releases 1.5 million data samples of one-year operation data ranging from 2016-01-01 to 2016-12-31. Among the whole data set, data of half a year from 2016-01-01 to 2016-06-30 is provided to be training set and the rest data is test set. Figure 1 shows the glimpse of data tables. As we can see, the data provides general information for each taxi order, which contains order ID, vendor ID (taxi firms), total passengers, pick-up timestamp, latitude and longitudes coordinates for both pick-up and drop-off. Drop-off timestamp is given in training set but not in test set. It is reasonable because when passenger order a taxi, he/she will not know the drop-off time. Therefore, drop-off timestamp cannot be used to build the predictive system. Pick-up and drop-off geolocation is given because passenger can set pickup and drop-off locations before ordering a taxi.

It is not enough to build accurate predictive system only based on these basic features. Feature engineering is the next step if one wants a wonderful predictive model. We performed various methods to extend our features. First, we show some visualization of statistical analysis for some useful features.

The features in training set is listed below:

* Id: order id for each taxi trip
* Vendor\_id: operation business firm id
* Pickup\_datetime: timestamp of pickup
* Pickup\_longitude: geolocation of pickup
* Pickip\_lattitude: geolocation of pickup
* Dropoff\_longtitude: geolocation of dropoff
* Dropoff\_lattitude: geolocation of dropoff
* Store\_and\_fwd\_flag: if the user’s information is recorded in system
* Trip\_duration(in seconds): time elapsed for this trip, this is the target we need to predict



**Your task is to build model(s) to predict trip duration**