I look at a Kaggle competition with data from the NYC Taxi and Limousine Commission, which asks competitors to predict the total ride time (trip\_duration) of taxi trips in New York City. The data provided by Kaggle is structured data provided as a CSV file. The data in the CSV file includes multiple formats: timestamps, text, and numerical data.I have not tried whole project yet but I am planning to do in near future using XGboost . I have used Python 3 and Scikit-learn: Python’s open source machine learning library.

The primary train dataset (train.csv) and test dataset (test.csv) is at the [Kaggle competition website](https://www.kaggle.com/c/nyc-taxi-trip-duration/data).

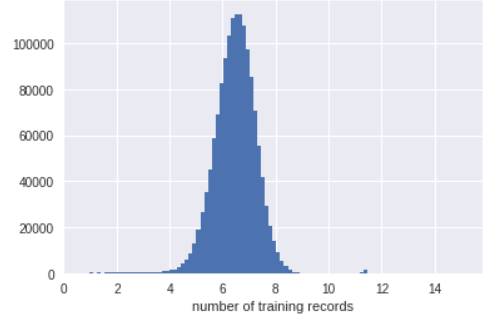
The weather dataset is at: [weather\_data\_nyc\_centralpark\_2016.csv](https://www.kaggle.com/mathijs/weather-data-in-new-york-city-2016).

The datasets for the fastest routes from OSRM can be found [here](https://www.kaggle.com/oscarleo/new-york-city-taxi-with-osrm). The files are: fastest\_routes\_train\_part\_1.csv, fastest\_routes\_train\_part\_2.csv, and fastest\_routes\_test.csv

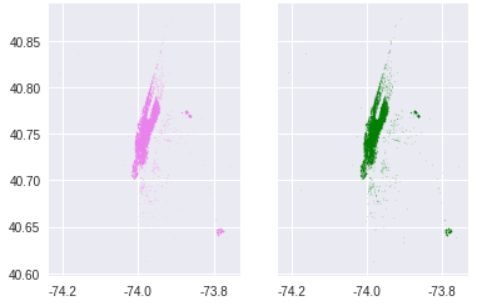
For the Kaggle competition, the data was cleaned and sampled, so that the training dataset contains 1,458,644 trip records, and the testing dataset contains 625,134 trip records. Each trip record contains the different attributes.

Additionally, I used a dataset that provides the fastest routes for each trip, based on maps from the Open Source Routing Machine, OSRM.

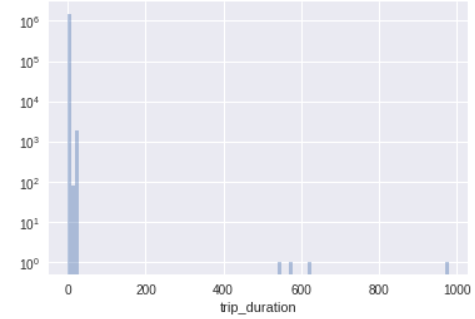
First I ploted how long is average trip



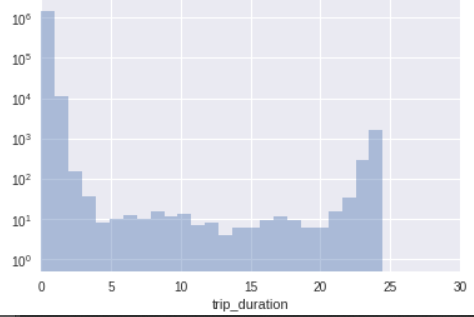
Next I plot training and testing data in terms of longitude and longitude.hich surprisingly showed the map of manhatten



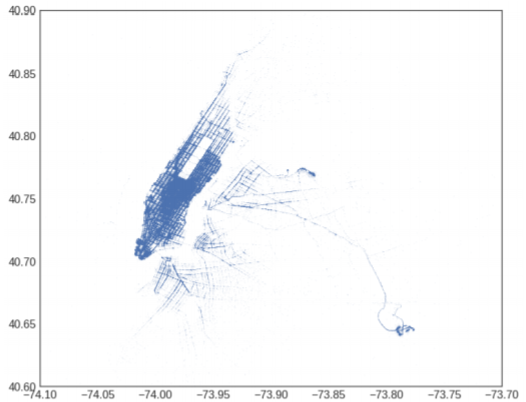
The distribution of trip\_duration in the training data set is shown below.



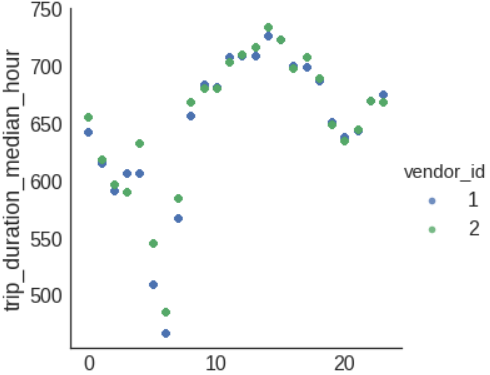
Restricting the range of trip\_duration from 0 to 30 hours, results in the following distribution that has a log-scale y-axis.



Plotting the coordinates of the pickup locations reveals the outline of Manhattan in New York City, with fewer pickups in the outer boroughs:



I calculate the median trip\_duration by hour for each vendor\_id and plot it versus the pickup hour in the following plot:



I calculate the distance of each airport from the pickup location and the dropoff location. The coordinates of JFK International Airport are 40.639722° N, -73.778889° W, and the coordinates of LaGuardia Airport are 40.77725° N, -73.872611° W.. where I have used a log-scale to see the shape of the features better. In the JFK and LGA pickup and dropoff distance distributions, I see that there are two peaks in the data. There is one peak around 20 miles, and another peak around 2 miles. The distributions of the distances are as follows:

