Write-up

January 27, 2018

1 New York City Taxi Trip Duration Project

1.0.1 Summary

The project is inspired by Kaggle.com competition. The goal is to analyze 1.4 million taxi trips from Jan 2016 to Jun 2016, and explore any factor that affects the trip duration. The dataset I use for this project is the file train.csv provided within the competition.

1.0.2 Design

The raw dataset train.csv contains 1458644 trip records. After reading several public EDA kernels, I noticed there is a portion of unreasonable trips that should be removed from the dataset. I don't have enough information to tell why some extreme trips happen, but in my intuition, I created some criteria to restrict the trips speed distance etc. also use for cleaning the raw data (See the scripts). The cleaned dataset contains 1449431 trip records so I almost removed 10000 rows from the raw training data.

The Tableau story see below:

https://public.tableau.com/profile/pengchong.tang#!/vizhome/taxi2/Story1

My presentation consists of 6 dashboard. The first dashboard is introduction and the last dashboard is references and information. So I create 4 pages of dashboard to describe the story about NYC taxi trip duration with topics to be discussed: geographical view, weather and holidays, time distribution and miscellaneous factors.

1.0.3 References

Kaggle.com competition page:

https://www.kaggle.com/c/nyc-taxi-trip-duration

Weather Data:

https://www.kaggle.com/mathijs/weather-data-in-new-york-city-2016

Route Data Source:

http://project-osrm.org/

OSRM API documentation:

http://project-osrm.org/docs/v5.5.1/api/#general-options