Springboard Capstone Project One
Predicting Taxi Ride Duration in NYC

Author: Anshul Dikshit

Date: September 16, 2018

Outline

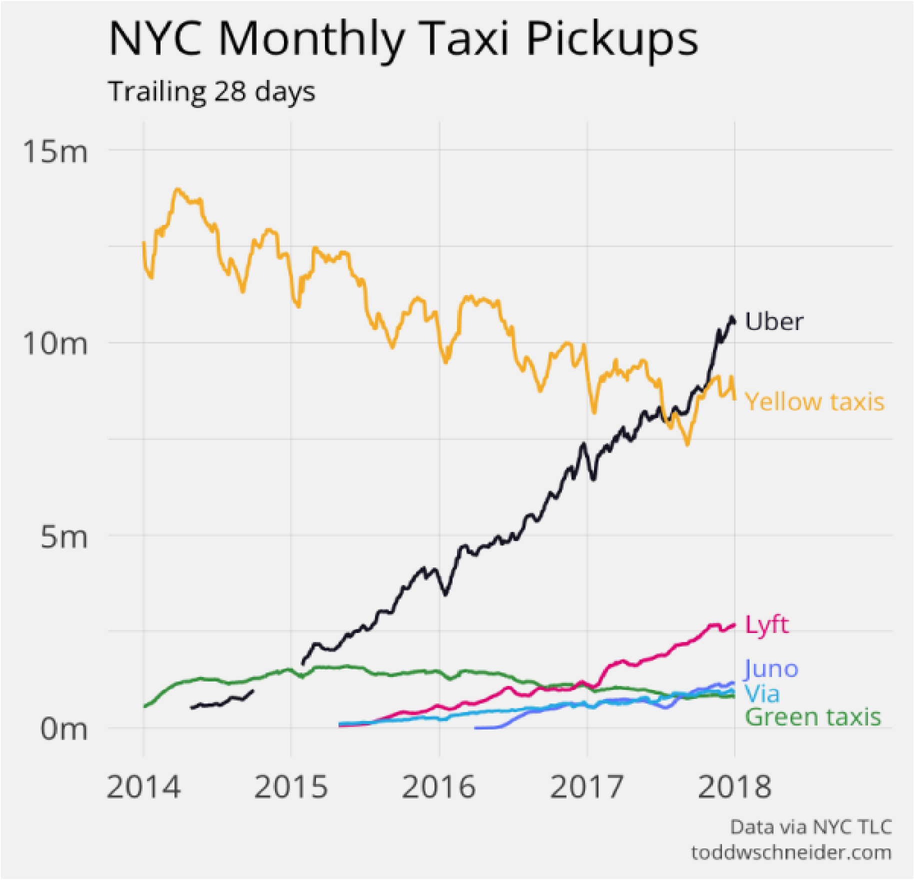
- Introduction & Problem Statement
- Dataset
- Analysis & Findings
- Statistical Inference
- Machine Learning
- Conclusions

Introduction & PROBLEM STATEMENT

With almost 1.5 million trips in 2016, riding through taxi at any time of the year across New York City is a big deal.

THE PROBLEM

Given the increase in number of taxi rides in NYC in recent years, there needs to be an estimates of taxi trip durations can improve the taxi utilization and the satisfaction of drivers and passengers



Data Set

The data has been taken from Kaggle which provides a starting point dataset consisting of the records of ~1.5 million taxi trips that took place in 2016. I’ve tried to go through the process of understanding the individual variables in the data by presenting beautiful, clear, and interactive data visualizations along with some approaches to their interpretation

Test dataset:

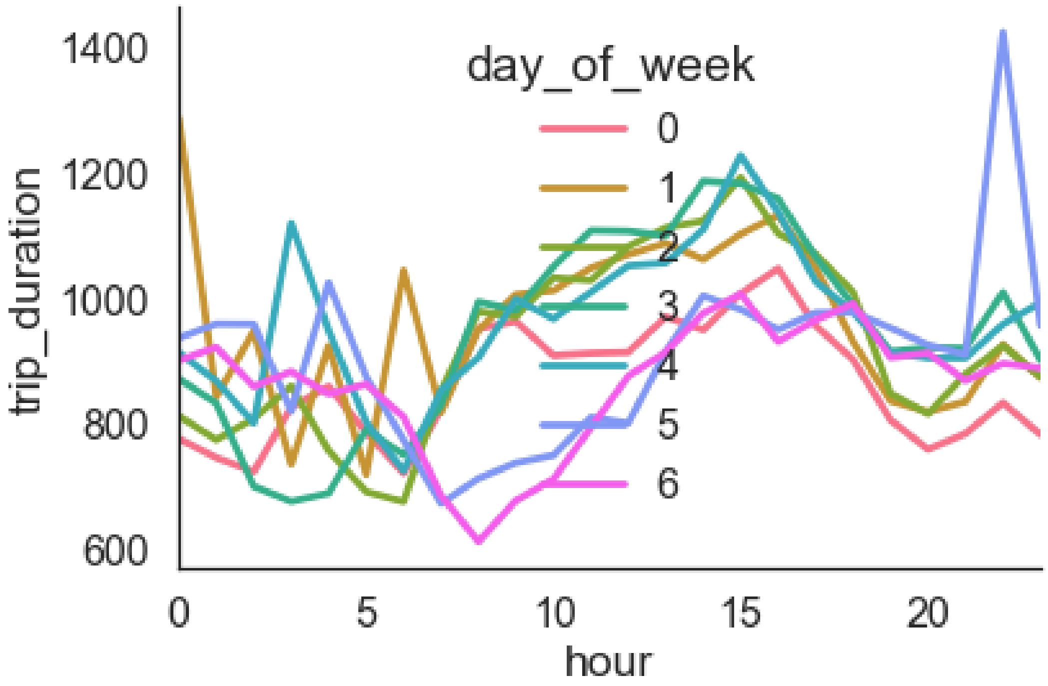
Variable name	Variable description
id	A unique identifier for each trip
vendor_id	A code indicating the provider associated with the trip record
pickup_datetime	Date and time when the meter was engaged
passenger_count	The number of passengers in the vehicle (driver entered value)
pickup_longitude	The longitude where the meter was engaged
pickup_latitude	The latitude where the meter was engaged
dropoff_longitude	The longitude where the meter was disengaged
dropoff_latitude	The latitude where the meter was disengaged
store_and_fwd_flag	This flag indicates whether the trip record was held in vehicle memory before sending to the vendor because the vehicle did not have a connection to the server. Y=store and forward; N=not a store and forward trip

Train dataset:

Variable name	Variable description
id	A unique identifier for each trip
vendor_id	A code indicating the provider associated with the trip record
pickup_datetime	Date and time when the meter was engaged
dropoff_datetime	Date and time when the meter was disengaged
passenger_count	The number of passengers in the vehicle (driver entered value)
pickup_longitude	The longitude where the meter was engaged
pickup_latitude	The latitude where the meter was engaged
dropoff_longitude	The longitude where the meter was disengaged
dropoff_latitude	The latitude where the meter was disengaged
store_and_fwd_flag	This flag indicates whether the trip record was held in vehicle memory before sending to the vendor because the vehicle did not have a connection to the server. Y=store and forward; N=not a store and forward trip
trip_duration	Duration of the trip in seconds

Initial Analysis

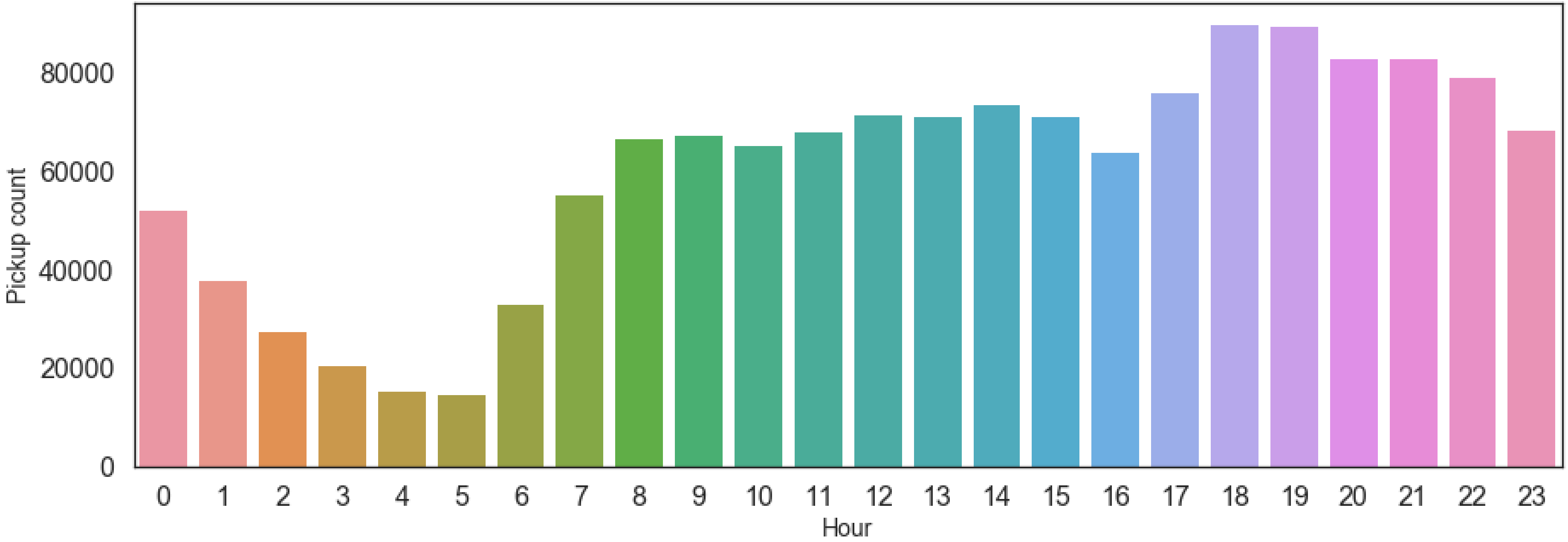
On day 0, that is Sunday and day 6 that is Saturday, the trip duration is very less that all the weekdays at 5 AM to 15 AM time. See this, on Saturday around midnight, the rides are taking far more than usual time, this is obvious through now verified using given data



Initial Analysis (cont.)

How many pickups/hr

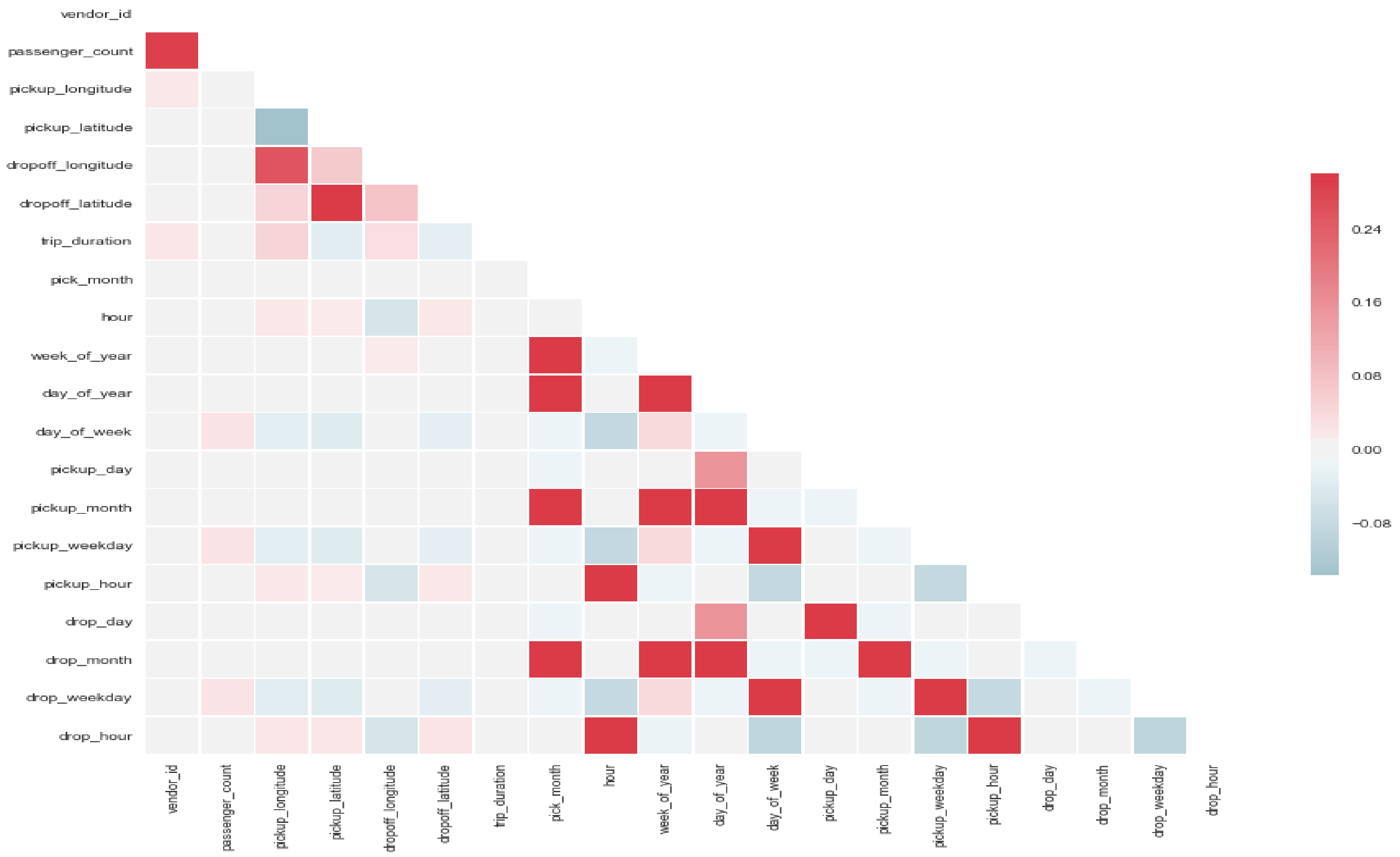
As expected, the number of pickups gradually decreases after mid-night. The highest number of pickups are around 6pm and 7pm in the evening which makes sense as many people are on their way to home from office.



Correlation between the different variables

The highest correlation is observed with the gollowing variables:

- `Week_of_year and pick_month`
- `day_of_year and pick_month`
- `trip_duration and pickup_longitude`



Training and testing the model using XGBoost

```
[0]	train-rmse:3.0237	valid-rmse:3.02175
Multiple eval metrics have been passed: 'valid-rmse' will be used for early stopping.

Will train until valid-rmse hasn't improved in 2 rounds.
[1]	train-rmse:1.55838	valid-rmse:1.55558
[2]	train-rmse:0.861315	valid-rmse:0.859258
[3]	train-rmse:0.563039	valid-rmse:0.564487
[4]	train-rmse:0.456612	valid-rmse:0.461999
[5]	train-rmse:0.423083	valid-rmse:0.430632
[6]	train-rmse:0.41285	valid-rmse:0.422282
[7]	train-rmse:0.407927	valid-rmse:0.419009
[8]	train-rmse:0.404263	valid-rmse:0.416805
[9]	train-rmse:0.401599	valid-rmse:0.415815
Modeling RMSLE 0.41581
```