Taxi Duration NYC

IST 718 Big Data Analytics​

Rajalakshmi Venkateswaran, Jeffery Kao, Jay Mahapatra

**Goal:**

Can we predict the the trip duration of taxi rides in New York city based upon a variety of variables including pick-up and drop-off time, pick-up longitude and latitude, the number of passengers, trip duration, collision data, as well as potentially traffic data. If we can accurately predict ride duration then we can find the best time to take a taxi thus optimizing route and someone’s day.

**Hypothesis:**

As a team we suspect that the collision and weather variable will have the largest impact on the duration variable as longer travel duration is generally caused by traffic collisions and inclimate weather.

**Data Sources:**

Taxi Trip Duration- <https://www.kaggle.com/c/nyc-taxi-trip-duration>

*Data Sets:*

* Train.csv- 1,048,575 rows X 11 columns
* Test.csv- 625,135 rows X 9 columns

*Metrics:*

* Unique identifier
* Pick-up/drop-off time
* Passenger count
* Pick-up/drop-off longitude and latitude
* Trip duration

Vehicle collisions in NYC, 2015- <https://www.kaggle.com/nypd/vehicle-collisions>

*Data Sets:*

* Database.csv- 477,732 rows X 29 columns

*Metrics:*

* Longitude/latitude
* Location
* Persons injured/killed
  + Pedestrians, cyclists, motorists
* Vehicle types

OSRM-<https://github.com/PaulEmmanuelSotir/NYC_TaxiTripDuration/tree/master/NYC_taxi_data_2016/osrm>

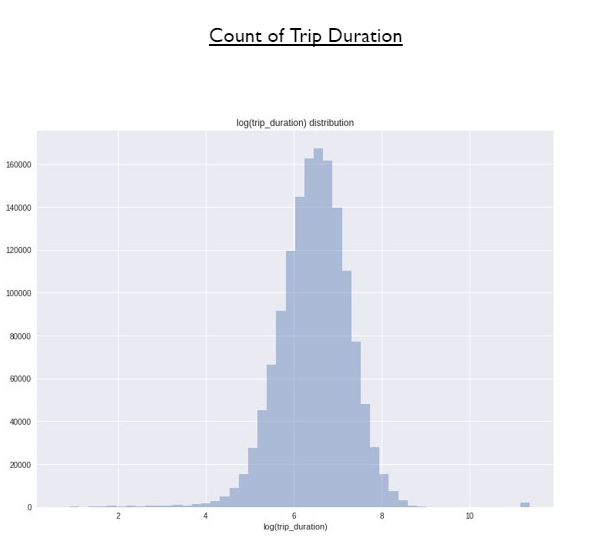
*Data Sets:*

* fastest\_route\_train\_ part 1. csv- 700,000 rows X 12 columns
* fastest\_route\_train\_ part 2. csv- 758,643 X12 columns
* fastest\_rout\_test.csv-625,135X12 columns

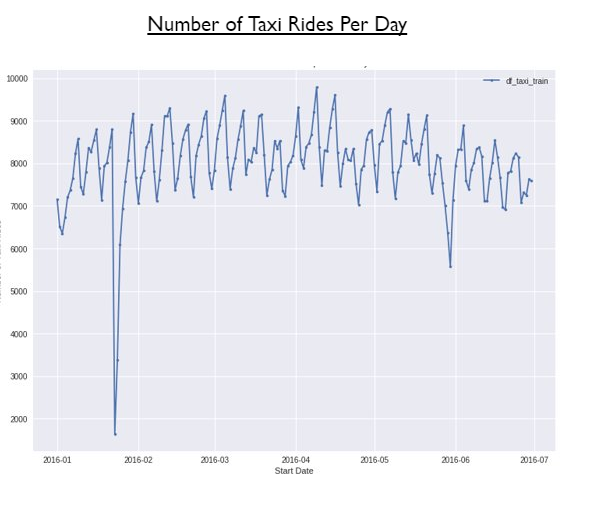
*Metrics:*

* Trip Duration
* Number of Steps- The number of stops it took to get from point a to point b.

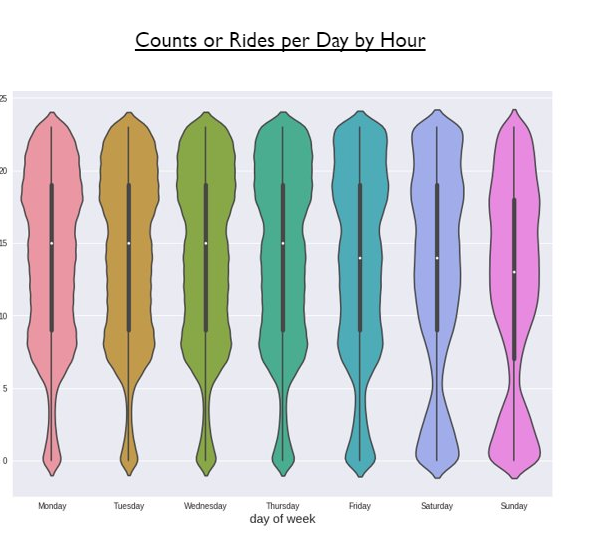
**Summary Statistics:**

****

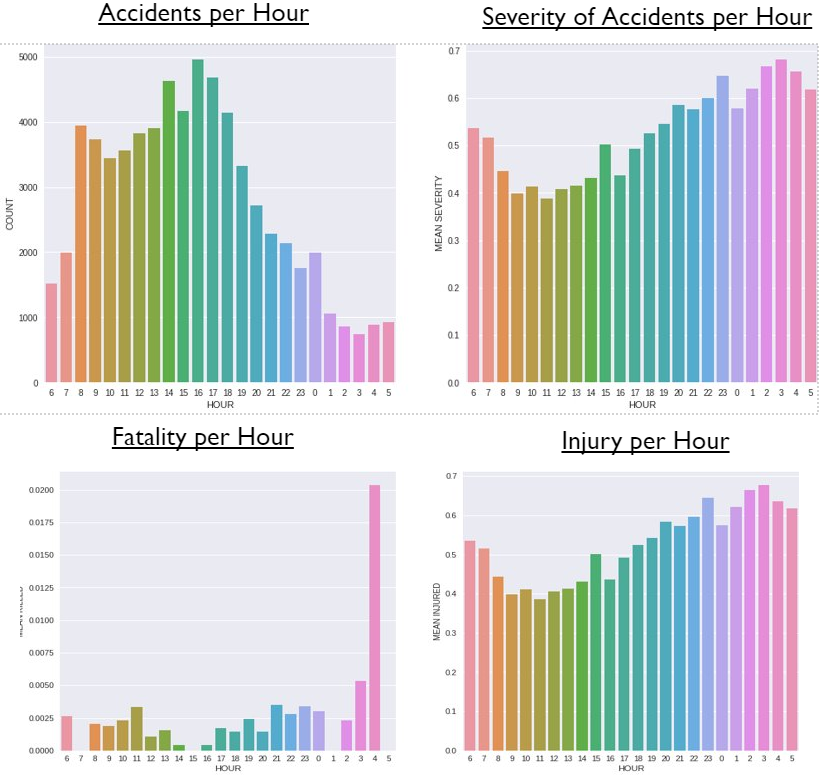
* Most of the taxi trips in New York City are between 10 and 13 minutes long.



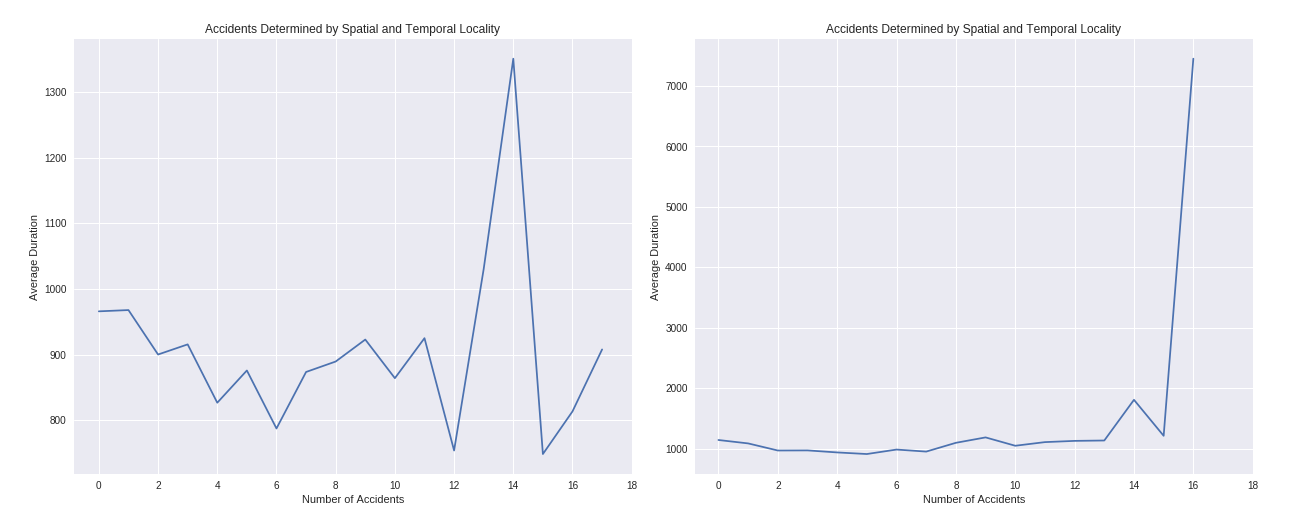
* The number of taxi rides ranges between 6.5k and 9.5k rides per day.
* There is also a regular increase-decrease pattern to the rides every 7 days as rides tend to increase during the weekdays and then decrease during the weekends.
* Outliers in the data can be attributed to bad weather days.
* The large dip in early 2016 can be attributed to the “snowpacalypse” when weather was extremely cold and hard to traverse in a car.



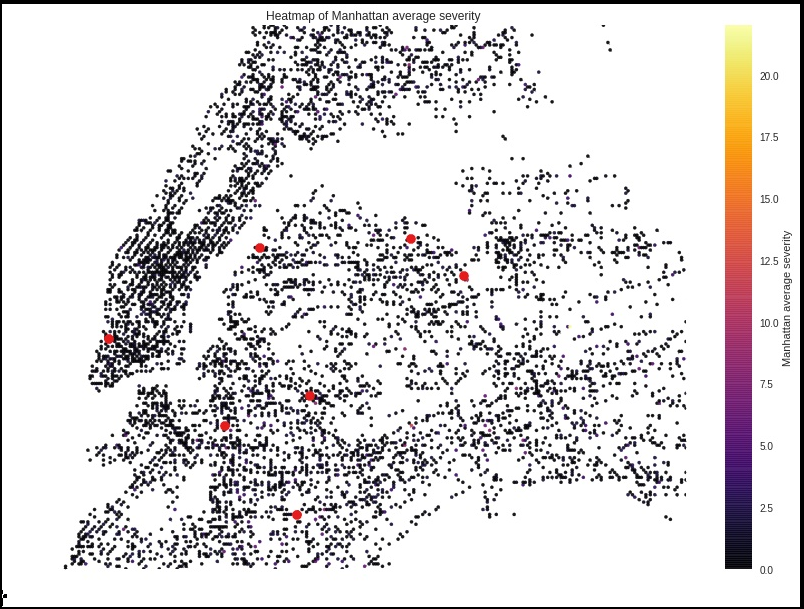
* A further dive into duration by day and time shows the following:
  + Most rides on weekdays occur between the hours of 5pm-10pm.
  + On weekends most rides occur between 11pm-and 2am as people return from bars.



* Most accidents tend to occur at rush hour during the weekdays.
  + This includes 8a-9am and 4pm- 6pm.
  + Most of these accidents are lower on the severity scale.
* The severity of accidents increases between the hours of 1am-5am coupled with the data above this is most likely to occur on the weekend.

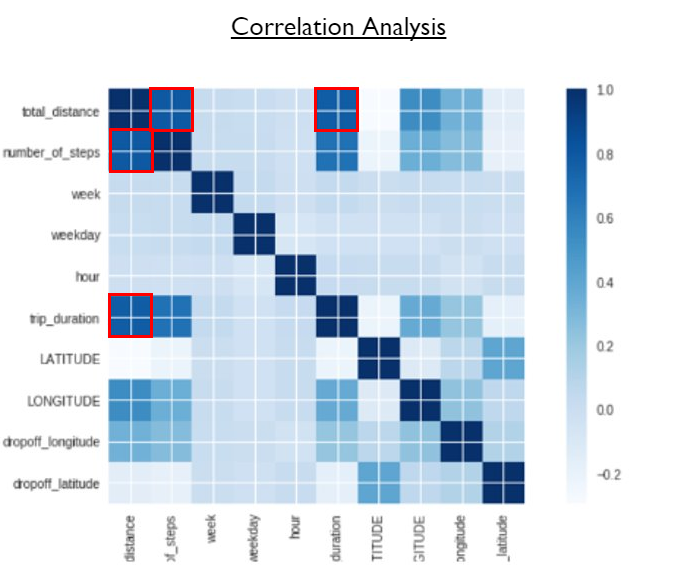


* As the proximity to the number of accidents increases the duration of taxi rides increases.



* The severity of accidents all tend occur at the junction points to different bridges and tunnels such as Manhattan to Brooklyn, Manhattan to Queens, and Manhattan to New Jersey.

**Model:**

****

* Prior to running our model we found a strong correlation with a couple variables:
  + Total distance is related to duration.
  + The steps it takes to complete a trip affects both distance and total duration.
* Collision was not a strong variable to duration.



* In running a linear model on collision data surprisingly we found the variable to be non-influential in predicting duration.
* However, OSRM, and particularly the variable “steps”(stops taken from point a to point b) was a major indicator of trip duration.

**Conclusion:**

* Based on the models and the correlation matrix the number of stops and the route a taxi takes are the most important factors in determining the duration of taxi ride. ​
* In addition, surprisingly, collision data is not a significant factor in affecting taxi duration.​
* To maximize the efficiency of taking a taxi, consider taking a taxi on the weekend between the hours and 5am and 9am and consider planning the route you want to take from point a to point b. In addition, try avoiding having others get in the taxi cab and going over bridges.

**Next Steps:**

* The OSRM data set offered many other variables that contribute to trip duration efficiency. We would like to dive into some of these variables further to see if going in a particular direction, the number of turns taken, and if there longitude or latitude areas in particular that affect duration.

**References:**

* <https://www.kaggle.com/c/nyc-taxi-trip-duration>
* <https://www.kaggle.com/nypd/vehicle-collisions>
* <https://github.com/PaulEmmanuelSotir/NYC_TaxiTripDuration/tree/master/NYC_taxi_data_2016/osrm>