NYC Green Taxi Exploratory Data Analysis

Homayoun Sadri

I. Objective

The objective of this project is to perform exploratory analysis over Green Taxi trip data collected from September 2015 to get some insights about the nature of the trips, average trip distance and other interesting observations.

II. Q1-Data:

The Green Taxi data is collected by the New York City Taxi and Limousine commission about "Green" Taxis. I will use the data from September 2015. The available data set consists of approximately 1.5M taxi data record. Each taxi data record contains of 21 different information fields.

The set of given information fields are[1]:

VendorID, lpep_pickup_datetime, Lpep_dropoff_datetime, Store_and_fwd_flag, RateCodeID, Pickup_longitude, Pickup_latitude, Dropoff_longitude, Dropoff_latitude, Passenger_count, Trip_distance, Fare_amount, Extra, MTA_tax, Tip_amount, Tolls_amount, Ehail_fee, improvement_surcharge, Total_amount, Payment_type, Trip_type

The data is programmatically downloaded and loaded into Pandas Data-Frame.

III. Q2-Trip distances:

The histogram and emprical cumulative distribution of the trip distance random variable are ploted in Figure 1. The histogram of the trip distance shows that the majority of trips are around one mile, and then the number of trips with higher trip distances decay drastically.

The histogram of the trip distance shows two different patterns:

- Number of trips with distance less than 0.9-1 mile which increases rapidly until the trip distance around one mile
- Number of trips with distance higher than 1-1.1 mile which decays in a light-tailed distribution

[1] The data dictionary can be found at:
http://www.nyc.gov/html/tlc/downloads/pdf/data_dictionary_trip_records_green.pdf

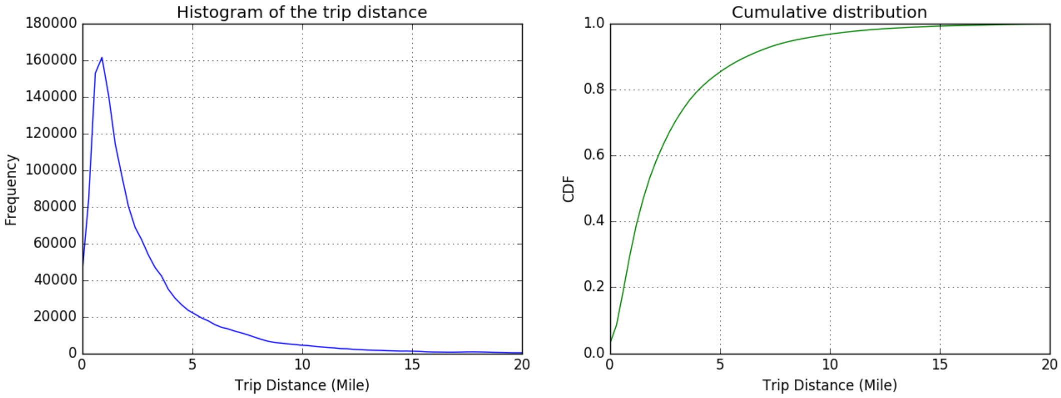

Figure 1: histogram and empirical cumulative distribution of trip distances

Empirical cumulative probability distribution shows that %2 of trips are zero-distance trips (probably due to computer or human error), %25 of trips are less than one mile, %52 of trips are less than 2 miles, %78 of trips are less than 4 miles, and %95 of trips are less than 9 miles.

Due to the continuous positive-valued nature of the trip distance, the Rayleigh distribution ($f(x;\sigma)=x/\sigma^2 \, exp(-x^2/2\sigma^2), \, x> 0$) could be a potential generative probability distribution model. However, simulation shows that the decay rate is less than $exp(-x^2)$ which rejects the Rayleigh distribution hypothesis.

Noticing that the distances are measured in New York City which has a grid layout structure, it is reasonable to think about Manhattan distance rather than Euclidean distance.

Therefore, one hypothesis is that the longitude and latitude could be generated from independent Poisson variables, X and Y with parameters λ_x and λ_y. Then the trip distance is $R^2= L_x X^2+L_y Y^2$, where L_x and L_y are the average horizontal and vertical distance of the grids.

To test this hypothesis, I can find the maximum likelihood estimation of the latent parameters L_x, L_y and Poisson distributions, $\lambda_x$and λ_y using Expectation-Maximization method and check how well the Poisson distribution hypothesis fits the underlying distribution for the trip distance random variable R.

IV. **Q3-Trip Patterns:**

Figure 2 shows the mean and median trip distance grouped by hour of day, indicating that the average value is higher than the median which in turn acknowledge the observation in Figure 1 that, the trip distance distribution is unbalenced and more streched toward the end tail of distribution. From Figure 2 we can see that the long trips happen at early morning (5-6 AM) and, during afternoon rush hour (4-7 PM) the average trips are the shortest.

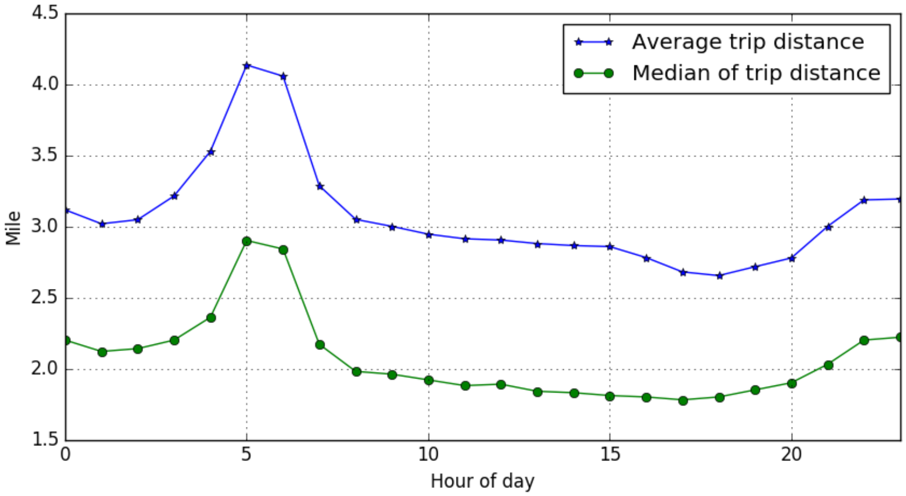

Figure 2: mean and median trip distance grouped by hour of day

V. Q3-Trips originating or terminating at one of the NYC area airports

To study the behavior of trips originating or terminating at one of the NYC area airports, I have used two approaches:

- Extracting trips that originate or terminate at one of the NYC area airports from geo-location data for JFK, EWR and LGA airports
- Extracting trips that originate or terminate at one of the NYC area airports from RateCodeID data

A. Airports from Geo-Location Data:

I retrieved the latitude and longitude information for JFK, EWR and LGA airports using google map and, then checked if the pick-up or drop-off locations are within a given distance of those airports as it is shown below (Figure 3).

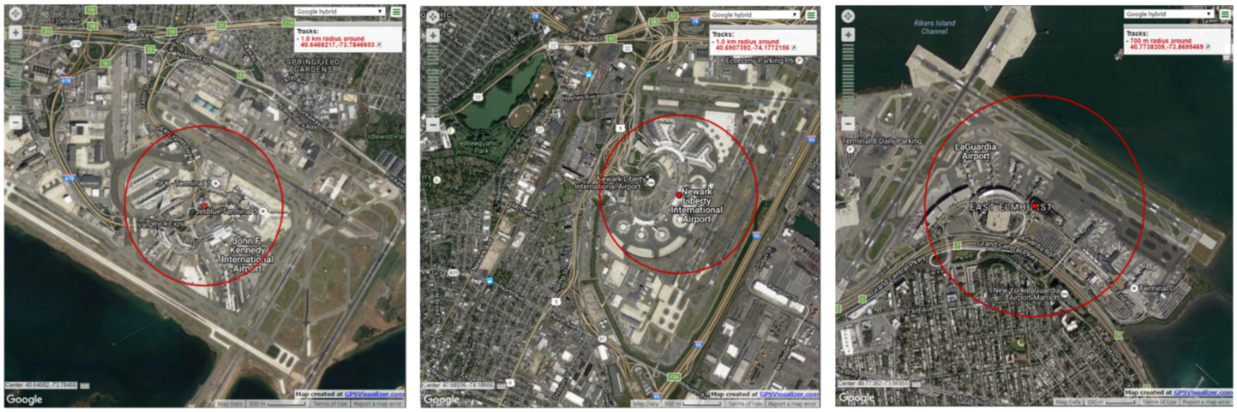
Google hybrid
Tracks:
- 1.0 km radius around 40.6468217,-73.7846603
SPRINGFIELD GARDENS
JFK - Terminal 7
JetBlue Terminal 5
Van Wyck Expy
John F. Kennedy International Airport
Google
Map created at GPSVisualizer.com
Google hybrid
Tracks:
- 1.0 km radius around 40.6907392,-74.1772156
Economy Parking P6
Weequahic Park
Newark Liberty International Airport
Newark Liberty International Airport
Google
Map created at GPSVisualizer.com
Google hybrid
Rikers Island Channel
Tracks:
- 700 m radius around 40.7738209,-73.8695469
Terminal B Daily Parking
LaGuardia Airport
EAST ELMHURST
Grand Central Pkwy
New York LaGuardia Airport Marriott
Terminal D
Google
Map created at GPSVisualizer.com

Figure 3: airport geo-location data retrieved from google map

The trips retrieved from geo-location data are shown at the following table:

Origin or destination	Count	% Correct RateCodeID
EWR destination	666	% 63
EWR origin	40	% 1
JFK destination	12651	% 18
JFK origin	283	% 7
LGA destination	17919	% 0
LGA origin	362	% 0
Town	1463005	-

As it can be seen from the table, the majority of trips that happen in an airport area don't have the correct airport RateCodeID. Therefore, the fare and accordingly, the tip are not calculated based on airport trips.

Thus, the geo-location information doesn't correctly identify the trips that originate or terminate at one of the NYC area airports

B. **Airports from RateCodeID Data**

In this scenario, I use RateCodeID data field to identify if a trip is originated or terminated at JFK or EWR airports.

The following table summarize the airport information for the latter scenario.

Airport	Count	Paid by credit Card	non-zero distance trip	Average distance	Average Fare
JFK	4435 (%0.3)	1588	1563	10.24	49
EWR	1117 (%0.07)	395	385	10.91	48.8

Here are some interesting characteristic of airport trips

- The data on the tip amount is available just for the credit card payment method (for any other type of payment, the tip value is zero).
- Though the average trip distance and average fare value for EWR and JFK are close to each other, but the trip distance and the fare value for EWR and JFK have totally different distributions (histograms).

The trip distance histogram in Figure 4 shows that the majority of trips to or from JFK airport are populated around 16-20 miles.

This observation makes sense since, using google map information, the distances from lower Manhattan, midtown Manhattan and upper Manhattan to JFK are around 16-18 miles.

However, trips to or from EWR are populated more uniformly around 0-30 miles, with higher density around 16 and 26 miles.

This observation also tallies data from google map information, where the distances from lower Manhattan, midtown Manhattan and upper Manhattan to EWR airport are around 15, 17 and 25 miles.

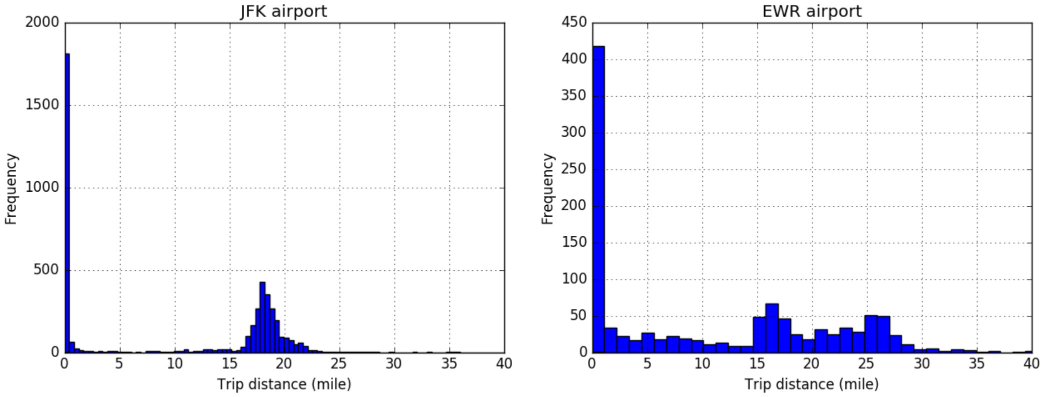

Figure 4: the trip histograms to or from JFK and EWR airports

On the other hand, the fare distributions in Figure 5 show that trips to or from JFK airport have flat rate of $52 and trips to EWR airport have a minimum charge of $20.

It seems that the fixed rate of $52 makes the trip fare from (to) JFK airport happens to (from) Manhattan reasonable but not very appealing for other origin-destination pairs.

However, the variable rate to EWR airport makes the fare more reasonable for both passengers located in Manhattan and off-Manhattan and probably, that’s the reason that we see more uniform trip distance distribution for EWR airport data.

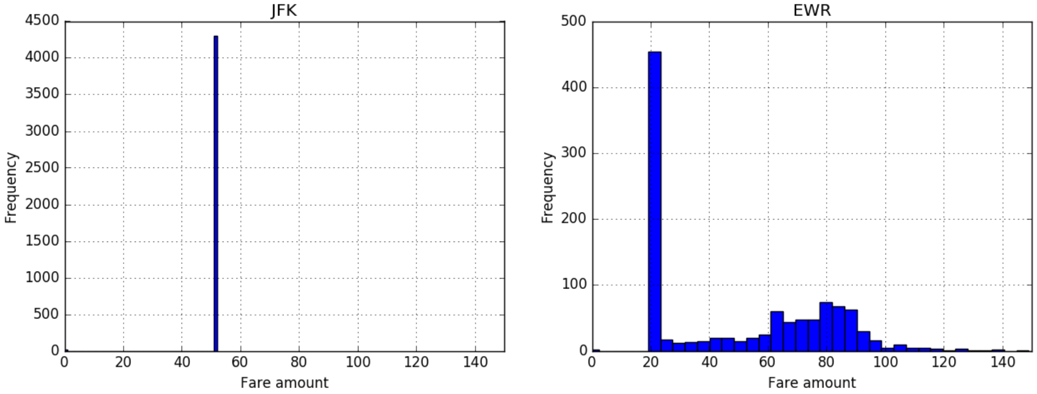

Figure 5: the fare histogram for trip originated or terminated at JFK and EWR airports

VI. **Q4- Tip Percentage Predictive Model**

The tip information is just available for credit card transitions which are %46.91 of all data records. The histogram for credit card transactions is plotted in Figure 6, which shows that the tip percentage is populated around %0, %16, %20 and %23.

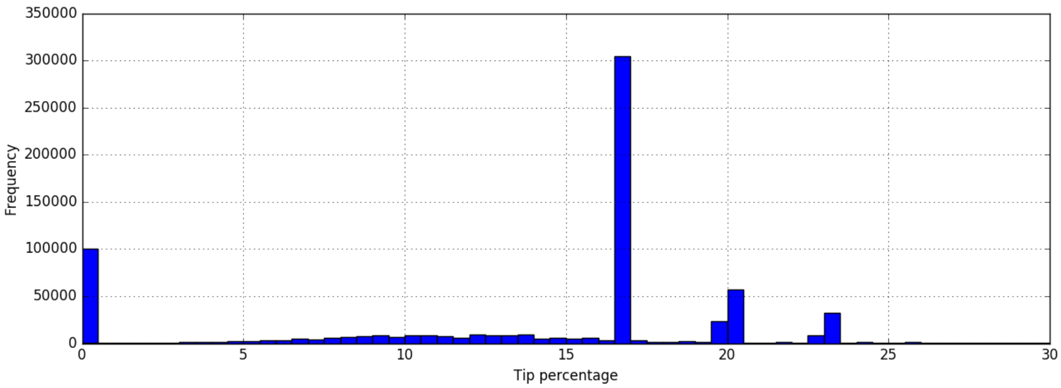

Figure 6: the tip percentage histogram for trips paid by credit card

C. **Predictive Model for the Tip Percentage**

The ability to have precise prediction relies on the nature of the features that we select for generating a model. Therefore, features should be selected such that they carry maximum information about the target parameter.

For the predictive variable, I won't use the Tip_amount, Total_amount and Ehail_fee. The Tip_amount is partially what we want to predict and the Ehail_fee is missed for all data records.

Before running any model, I do exploratory analysis to look at scatter plots of different variables vs. the target variable (Figure 7).

From Figure 7, it can be seen that the set of given parameters are combination of continuous and categorical variables. Also, the selection of transformation functions for given variables doesn't look straight forward.

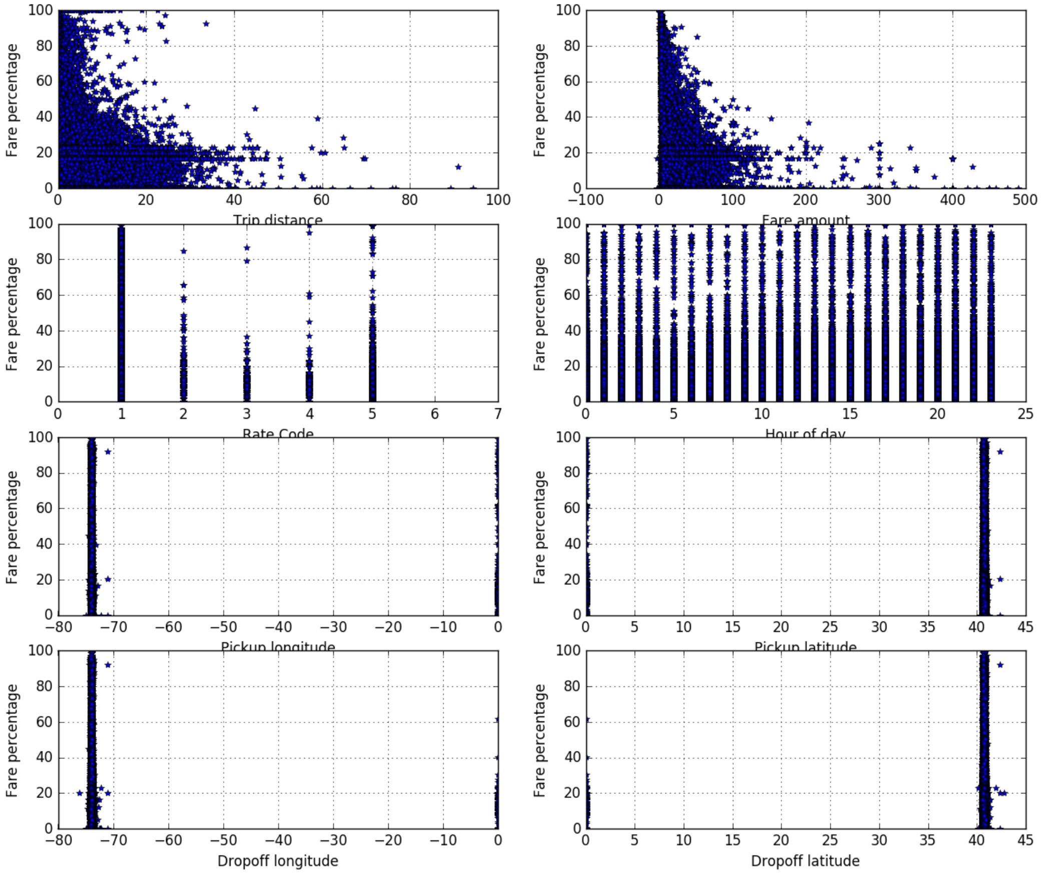

Figure 7: scatter plots of different variables vs. the target variable

Due to the regression nature of the problem, first I consider linear regression model, including Least Squares Linear Regression with Ridge Regularization and Cross Validation.

In the least square linear regression (LR) model, the feature weight vector is found by minimizing the least square measure. In the Ridge regression, the cost function has an L2 regularization term where the regularization coefficient is found using cross-validation.

In this problem since we don't have many features, Lasso regularization or any hybrid model such as Elastic Net are not very different in terms of regression coefficients.

To evaluate different models, I use the Score function which is the coefficient R^2 defined as (*1 - u/v*), where *u* is the regression sum of squares ***sum***((*y_true* - *y_pred*) ^{2}) and *v* is the residual sum of squares ***sum***((*y_true* – ***mean***(*y_true*))^2). The best possible score is 1.0 and it can be negative (because the model can be arbitrarily worse). A constant model that always predicts the expected value of y, disregarding the input features, would get an R^2 score of 0.

To evaluate the performance of the regression technique, I split the data into two random subsets, train and test sets with %80 and %20 ratio, respectively. For the model cross-validation and optimization over model parameters, I use10-fold evaluation method.

Data	LR Ridge Reg. score
Training	0.032
Test	0.030

The linear model doesn't show a good performance in predicting the target parameter. The exploratery analysis shows that the raw features are highly categorical. My analysis also show that using dummy variables is not very effective.

D. Random Forest Regressor

To take into account both continuous and categorical variables, another possible approach is that we can change our regression problem by dividing the target parameter to different intervals (classes) to train classifiers, and then find different regression models for each class. In this method we can use classifiers like Random Forest to first classify and then predict the value using regression models.

To find the best parameters based on cross-validation, I use SigOpt Bayesian optimizer package to find all hyper-parameters. The Bayesian optimizer suggests the max depth of 13 and the number of estimators of 11 for the random forest model.

In this case, to avoid overfitting, I split the data into two random subsets, train and test sets with %80 and %20 ratio, respectively, where the test data is just used for final evaluation. I use 3-fold cross-validation using training data.

Data	Random Forest Regressor score
Training	0.248
Cross-validation	0.167
Test	0.157

The random forest regression with Bayesian optimizer shows an order of magnitude improvement compared to the linear model regression model.

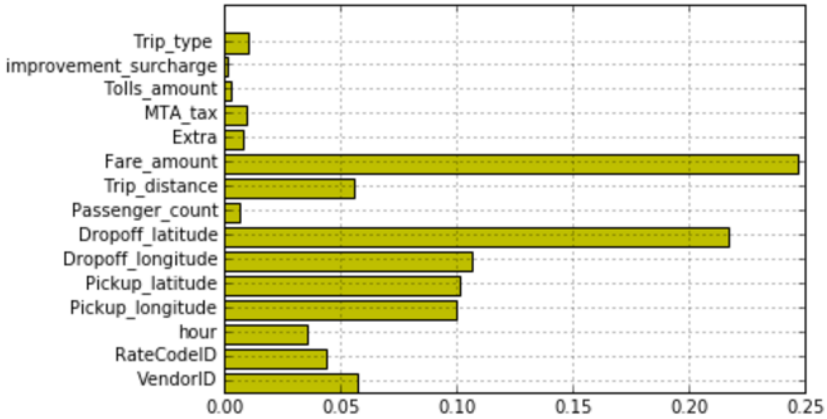

Figure 8: feature importance in random forest regression model

The feature importance diagram in Figure 8 shows that the Fare_amount is the most important parameter and, the drop-off coordinate are the second most important features in determining the tip percentage using random forest regression model.

VII. **Q5-Distribution**

I find the average trip speed by dividing the trip distance by the average trip duration calculated from pick-up and drop-off time. The speed histogram during September is shown at Figure 9.

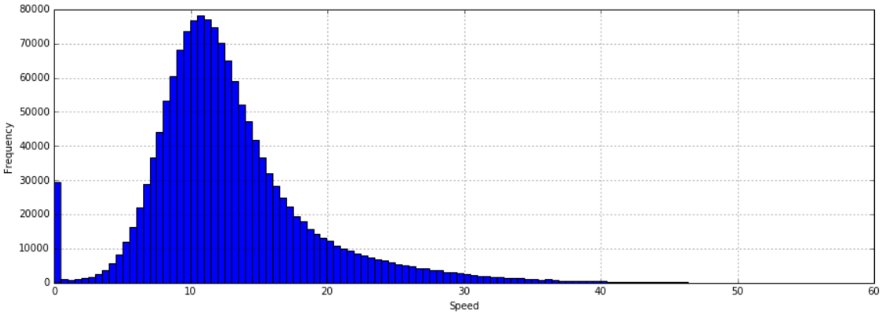

Figure 9: the speed histogram for the trips in September 2015

Interestingly, the speed histogram at Figure 9 is similar to a horizontally shifted version of trip distance histogram, which indicates that average trip duration should be constant. To test this hypothesis, I plot the average trip duration per hour of day (Figure 10). Figure 10 acknowledges this hypothesis which indicates that passengers spend in average 20-25 minutes in a trip regardless of the hour of the day (probably they use other public transportations if the expected trip duration is more than 20-25 min).

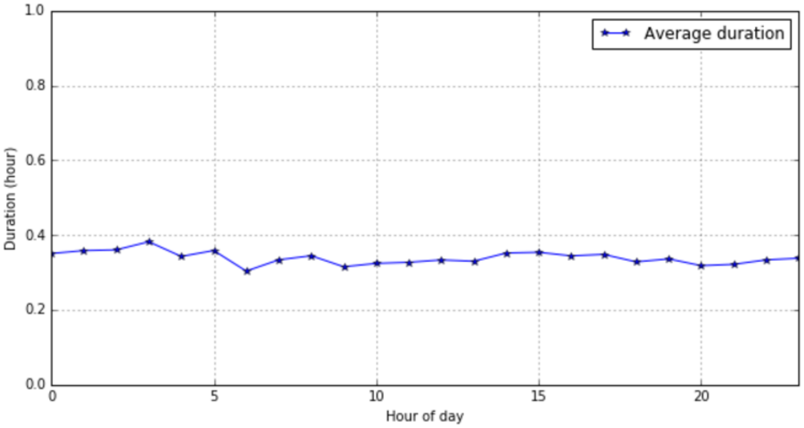

Figure 10: average trip duration per hour of day

E. Average speed in weeks of September:

The average speed per week shows a gradual increase toward the end of the month (Figure 11). One hypothesis about this observation is that schools start in September, therefore, visitors start leaving New York City toward the end of the month. The smaller number of visitors leaves the city with less traffic and consequently larger average speed.

One other hypothesis could be that the New York City is under construction all the time, and maybe one impactful construction is over during September. However, this hypothesis is less likely as the increase is gradual rather than abrupt.

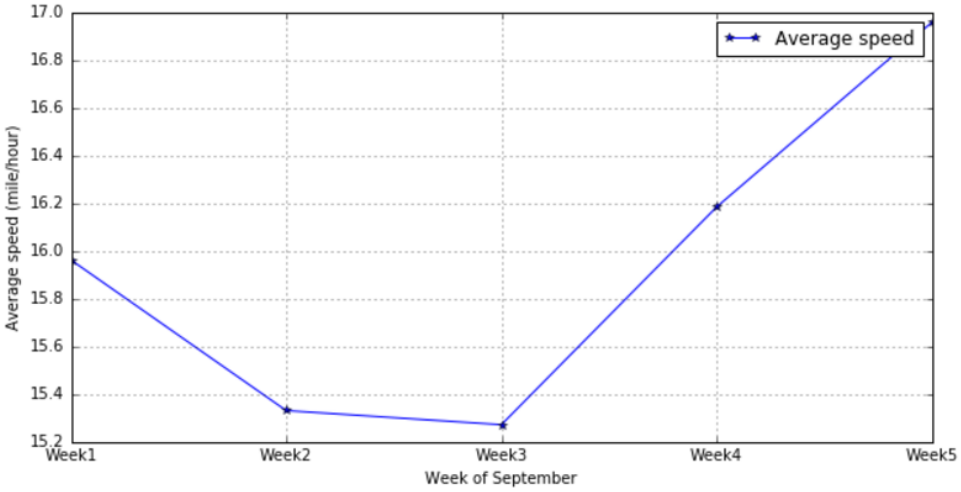

Figure 11: the average trip speed for weeks of September

The average trip speed per hour of day is plotted in Figure 12. Interestingly, the average speed per hour of day shows the same trend as the average and the median of trip distances in Figure 2. This observation implies the hypothesis that, the average trip duration per hour of day should be constant. We had this hypothesis before, which was concluded from the similarity of distance and speed histograms.

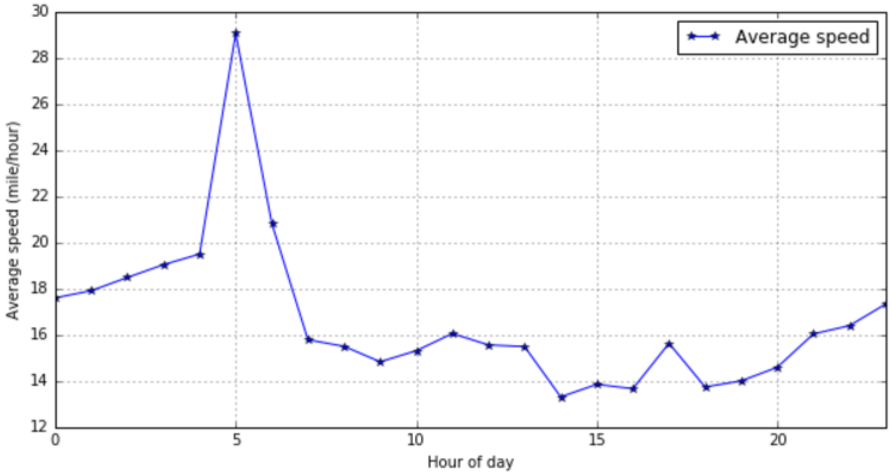

Figure 12: average trip speed per hour of day

The hypothesis that the average trip duration during a day should be constant can be verified using Figure 10 which indicates that cab passengers spend in average a constant time (20-25 minutes) in a trip regardless of the hour of the day.

VIII. **Summary and Observation Highlights**

I studied the Green Taxi trip data for month of September 2015. The histogram of the trip distances shows two different patterns: number of trips with distance less than 0.9-1 mile which increase rapidly until around one mile and, number of trips with distance higher than 1-1.1 mile which decay in a light-tailed distribution.

Due to the continuous positive-valued nature of the trip distance, the Rayleigh distribution could be a potential underlying probability distribution, however, the decay rate study rejects this hypothesis. On the other hand, it seems that due to the Manhattan grid structure of New York City, the trip distance random variable is function of two independent Poisson variables, in X and Y coordination. This hypothesis needs further study.

Also, it can be seen that long trips happen at early morning (5-6 AM), while during afternoon rush hour (4-7 PM), trips have the shortest average distance.

The data for airport trips reveal that the majority of trips to or from JFK airport are populated around 16-20 miles. This observation matches the data from google map where the distances from lower Manhattan, midtown Manhattan and upper Manhattan to JFK are around 16-18 miles. Also, the trips to or from EWR are populated more uniformly around 0-30 miles, with higher density around 16 and 26 miles.

The fare distributions show that trips to or from JFK airport have flat rate of $52 and trips to EWR airport have a minimum charge of $20.

The tip percentage histogram shows that the tip percentage is populated around %0, %16, %20 and %23. To build a predictive model for the tip percentage information, using the random forest regression along with Bayesian optimization for the model hyper-parameters, I could improve the predication score in order of magnitude compared to the linear model regression model.

Despite different traffic pattern on different hour of day, the trip duration data reveals an interesting observation which indicates that passengers spend in average a fixed amount of time (20-25 minutes) in a trip as probably they use other transportation systems if the expected trip duration is high.