Name: Mayank Baheti

Chicago Taxi Trips:

Taxicabs in Chicago, Illinois, are operated by private companies and licensed by the city. There are about seven thousand licensed cabs operating within the city limits. Licenses are obtained through the purchase or lease of a taxi medallion which is then affixed to the top right hood of the car.

Problem Statement:

Predicting time taken to go from one place to another is one of the crucial planning inputs to decide and optimize the operations. In this challenge, you will develop a solution to predict the transit time for going from a Pickup area to a Dropoff Area in the data.

We must develop a model to predict transit time between two different locations in the Chicago Taxi Trips data.

Dataset:

This dataset includes taxi trips from 2013 to 2017 (till July), reported to the City of Chicago in its role as a regulatory agency. To protect privacy but allow for aggregate analyses, the Taxi ID is consistent for any given taxi medallion number but does not show the number, Census Tracts are suppressed in some cases, and times are rounded to the nearest 15 minutes. Due to the data reporting process, not all trips are reported but the City believes that most are.

For the generalization purpose, I had made 3 data sets:

1. **Training_data**: data from 2013 to 2016, again 20% of the data randomly was used for testing purpose **(X_train and X_test).**
2. **Testing data**: which is from Jan 2017 to Jul 2017. And is kept untouched as the **Validation Data (X_FinalTest).**

Evaluation Metric:

Assessment will be done on the final Mean Absolute Percentage Error (MAPE) value of the solution.

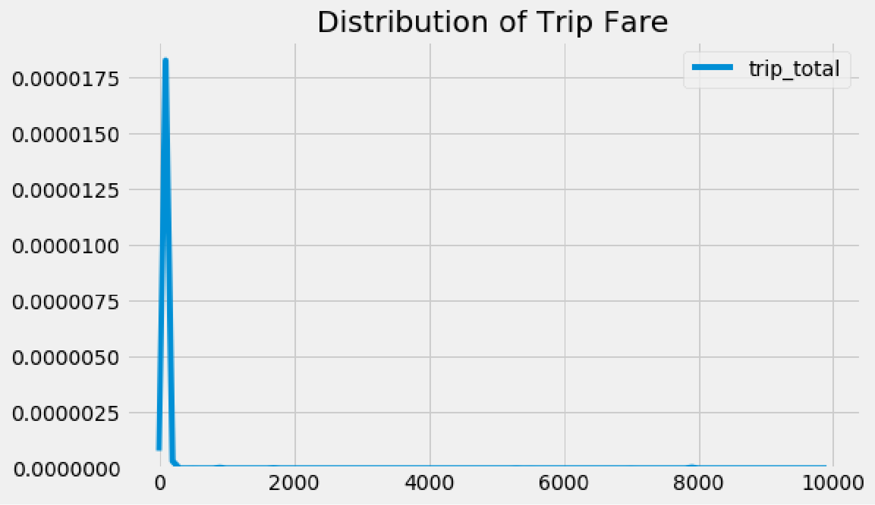
I have used data from 2013 to 2016 for training and 2017 for testing.

Name: Mayank Baheti

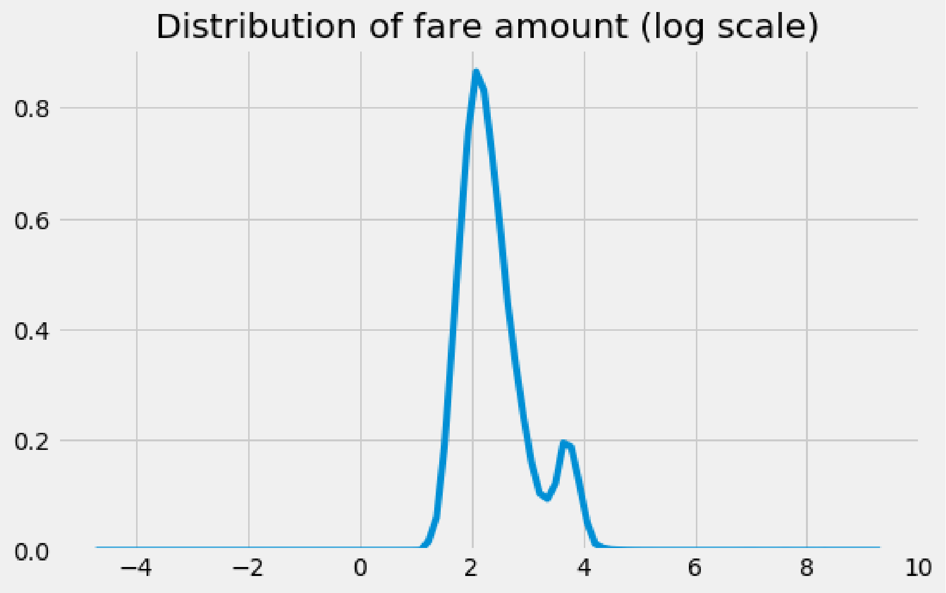
Exploratory Data Analysis:

This data has a lot to explore about, and few of them are as below:

Distribution of Trip fare:



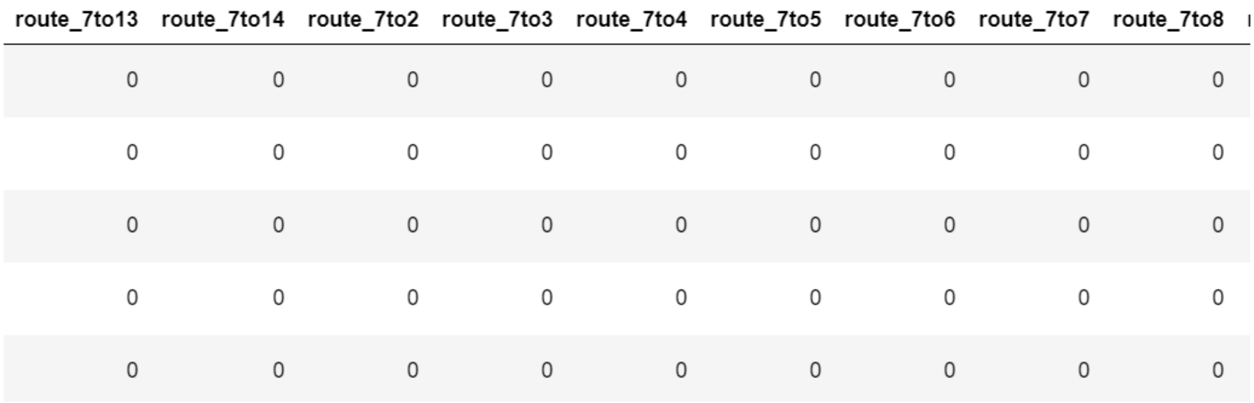
Above plot is highly skewed, and doesn't provide us lot of information, so taking the logarithmic transformation and we got a good normally distributed data, with most of the fare amount ranging between 400 to 3000 dollars and median being at 965 dollars.



Later, with data provided to us also had the coordinates for the pickup and drop location, so ni tried to cluster these locations based on the Euclidean distance using the K-Means Algorithm, separately for

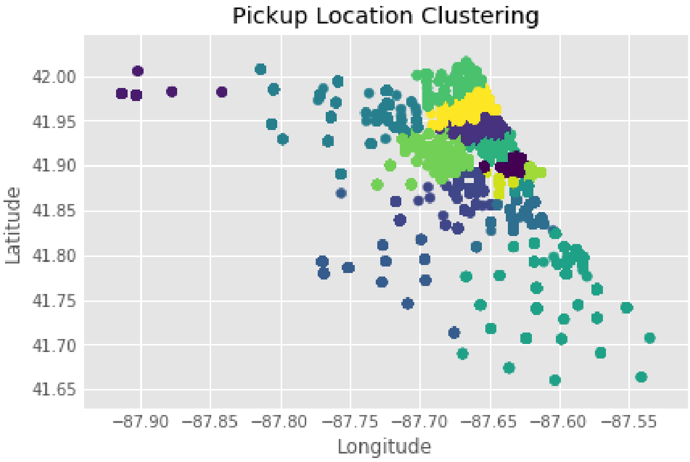
Name: Mayank Baheti

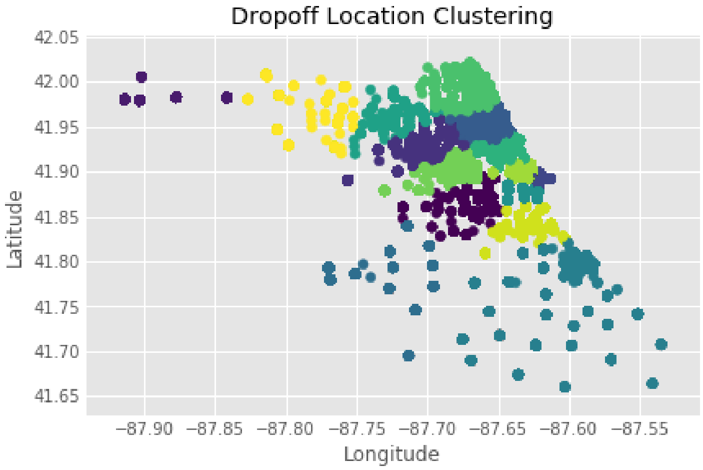
pickup and drop location, so that we can get a route for these clusters, for example data point A lies in cluster 1 as a pickup point and a person gets dropped to data point B which lies in cluster 3 so we can denote it as **route1_3,** as shown below data frame snapshot.



route_7to13	route_7to14	route_7to2	route_7to3	route_7to4	route_7to5	route_7to6	route_7to7	route_7to8
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0

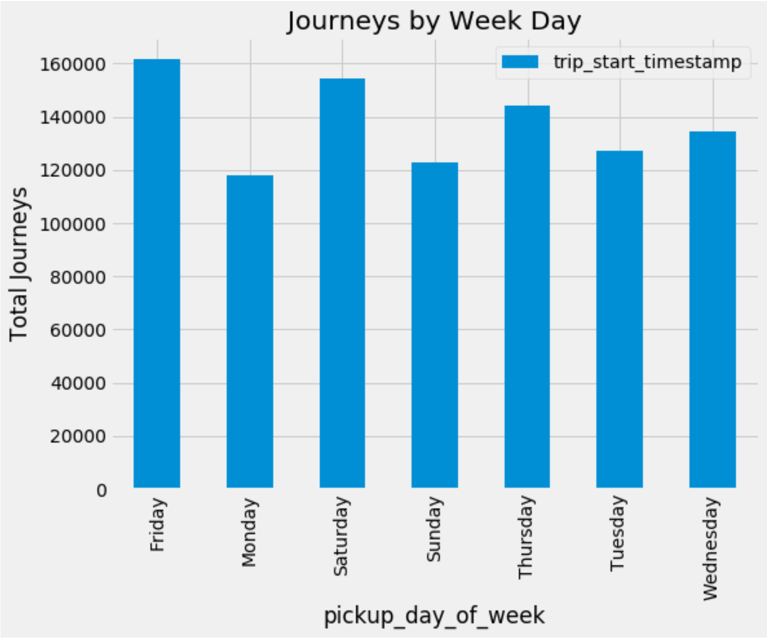
And below are the cluster we got from Kmeans algorithm of Pickup and drop off points.



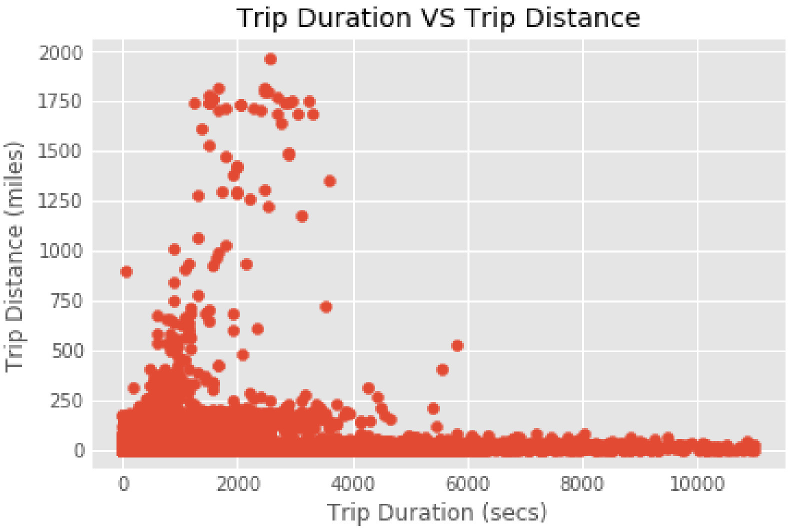


Name: Mayank Baheti

Below is the histogram which clearly tells us about most of the cabs are hired on Friday, but that's a marginal increase than rest of the days.



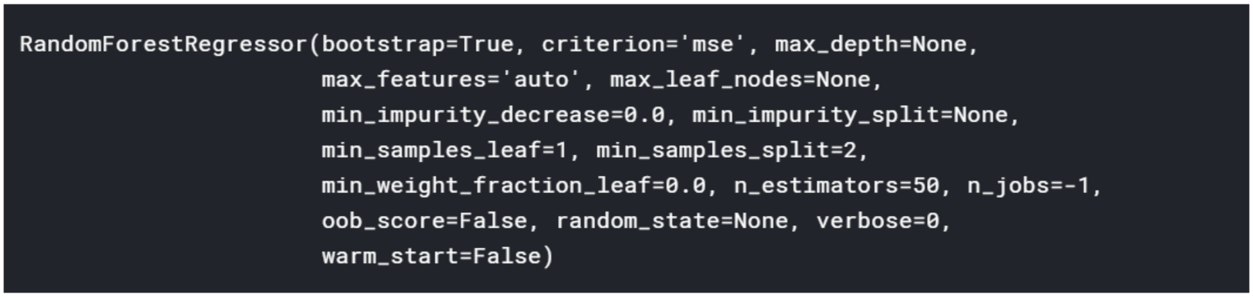
Below scatter plot, doesn't tell much about the relationship between trip distance and duration, may be due to traffic and also the time of the day makes to the traffic situations.



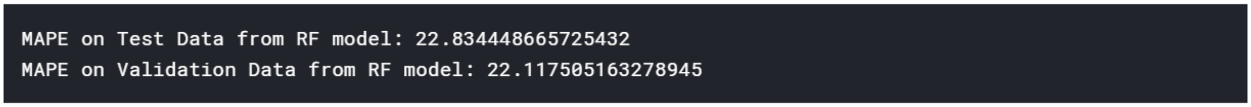
Name: Mayank Baheti

Models tried and their results:

1. Random Forest Regressor:

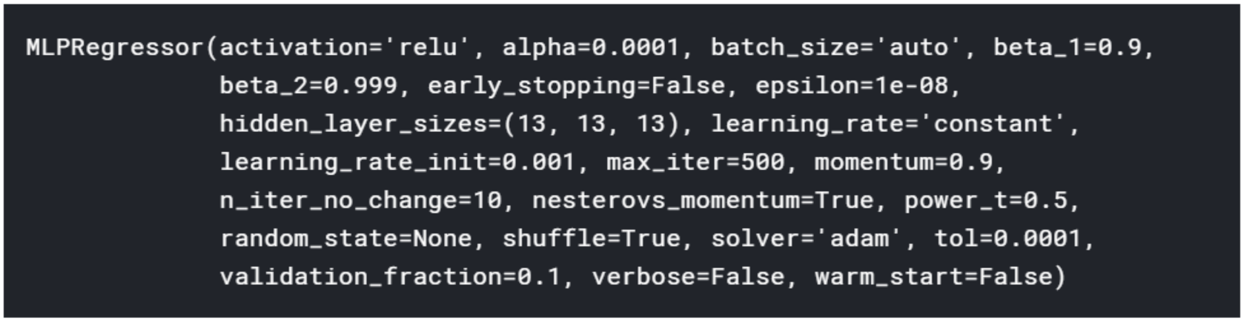


```
RandomForestRegressor(bootstrap=True, criterion='mse', max_depth=None,
                      max_features='auto', max_leaf_nodes=None,
                      min_impurity_decrease=0.0, min_impurity_split=None,
                      min_samples_leaf=1, min_samples_split=2,
                      min_weight_fraction_leaf=0.0, n_estimators=50, n_jobs=-1,
                      oob_score=False, random_state=None, verbose=0,
                      warm_start=False)
```

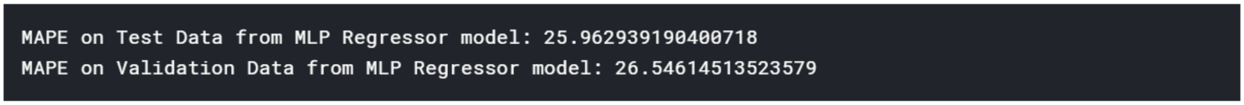


```
MAPE on Test Data from RF model: 22.834448665725432
MAPE on Validation Data from RF model: 22.117505163278945
```

2. Neural Network:



```
MLPRegressor(activation='relu', alpha=0.0001, batch_size='auto', beta_1=0.9,
             beta_2=0.999, early_stopping=False, epsilon=1e-08,
             hidden_layer_sizes=(13, 13, 13), learning_rate='constant',
             learning_rate_init=0.001, max_iter=500, momentum=0.9,
             n_iter_no_change=10, nesterovs_momentum=True, power_t=0.5,
             random_state=None, shuffle=True, solver='adam', tol=0.0001,
             validation_fraction=0.1, verbose=False, warm_start=False)
```



```
MAPE on Test Data from MLP Regressor model: 25.962939190400718
MAPE on Validation Data from MLP Regressor model: 26.54614513523579
```

Random Forest seems to give the best results of the two.

Future Scope:

Using weather data and using more of Census tract information will be much helpful, to consider the area information and traffic due to weather.

Code Link: https://www.kaggle.com/mayankbaheti25/chicago-taxitriptime/