Capstone Project - 2

NYC Taxi Trip Time Prediction

Team members :

Rohit Meshram

Rahul Gayakwad

Rutuja Hingankar

Prashik Ingle

Narayan Borde

Content

- **Introduction**
- **Problem statement**
- **Data summary**
- **Exploratory Data Analysis (EDA)**
- **Feature Engineering & Selection**
- **Building and Evaluating Model**
- **Conclusion**

Introduction:

- Transportation plays a vital role in large cities
- Taxi mode of transportation has become a key player in large cities of NewYork and other countries.
- Different variety of service providers are Uber, Yellow Taxi, Green Taxi etc.
- The data that contain ride details was made available by NYC taxi and Limousine commission.
- We use these details to perform analytics on ride data that would benefit businesses of various types and government.
- Taxi Drivers also have to choose best route having lesser trip time.
- So here we will be building a model which will be predicting the trip duration of taxies running in NewYork. This prediction will help customers to select the taxi based on trip duration and driver to select optimum route to their destination.

Problem Statement :

- To build a model that predicts the total ride duration of taxi trips in New York City.
- This will help in placing the taxis at the right places at the right time thereby reducing the waiting time for customers.

Dataset Description:

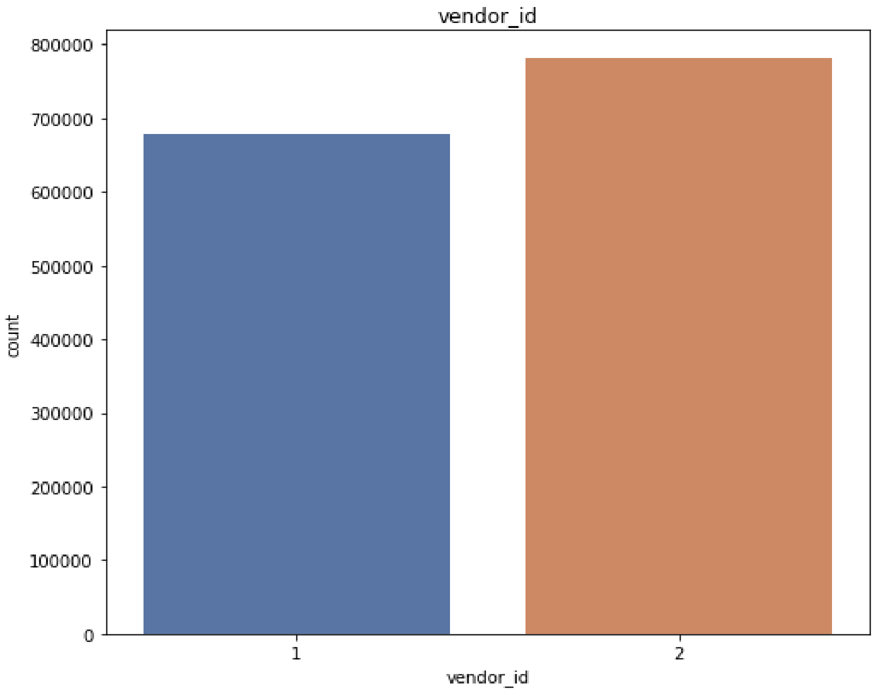
- id - a unique identifier for each trip
- vendor_id - a code indicating the provider associated with the trip record
- pickup_datetime - date and time when the meter was engaged
- dropoff_datetime - date and time when the meter was disengaged
- passenger_count - the number of passengers in the vehicle (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

BASIC EXPLORATION:

- The dataset contains 1458644 rows and 11 columns.
- Two categorical features ‘store_and_fwd_flag’ and ‘vendor_id’
- Outliers present in all numerical features
- Data formating steps required for datetime features
- No null values present
- Passenger_count, Vendor_id and trip_duration are having integer value.
- pickup_datetime,dropoff_datetime is a datetime variable
- pickup_longitude,pickup_latitude,dropoff_longitude,dropoff_latitude are real numbers having float as data type *store_and_fwd_flag and Id belongs to a string data type.

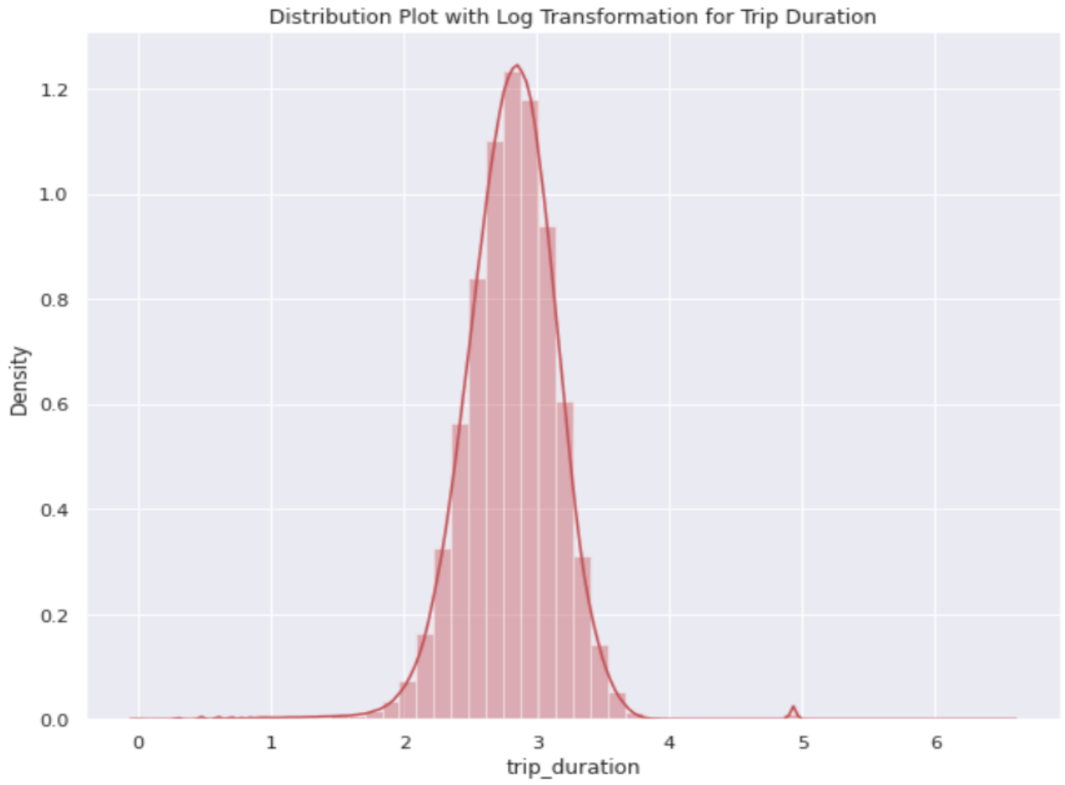
Data Analysis :

- Performed basic EDA to get some insights of the data.



- Vendor_id stating the provider associated with trip preferable different taxi companies.
- Analysis tells us that second service provider has been most frequently opted by people over first service provider over the period of time.

Data Analysis :

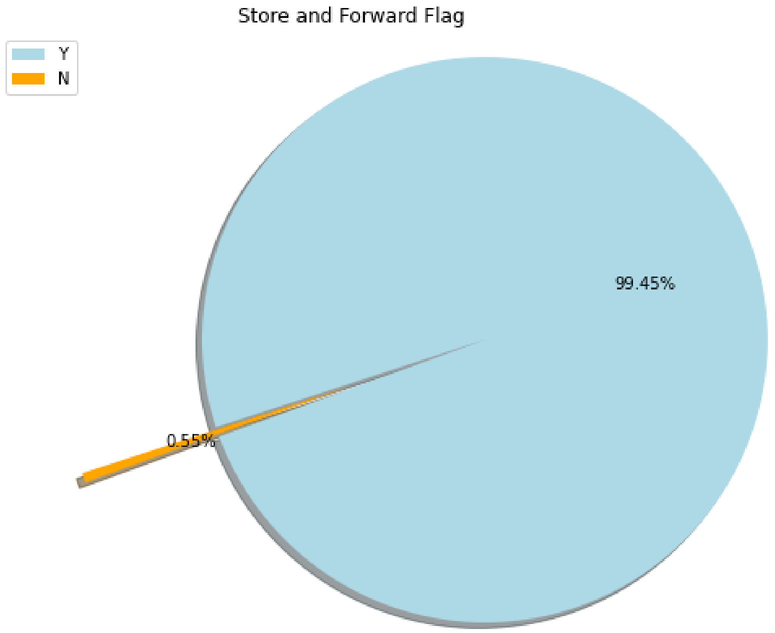


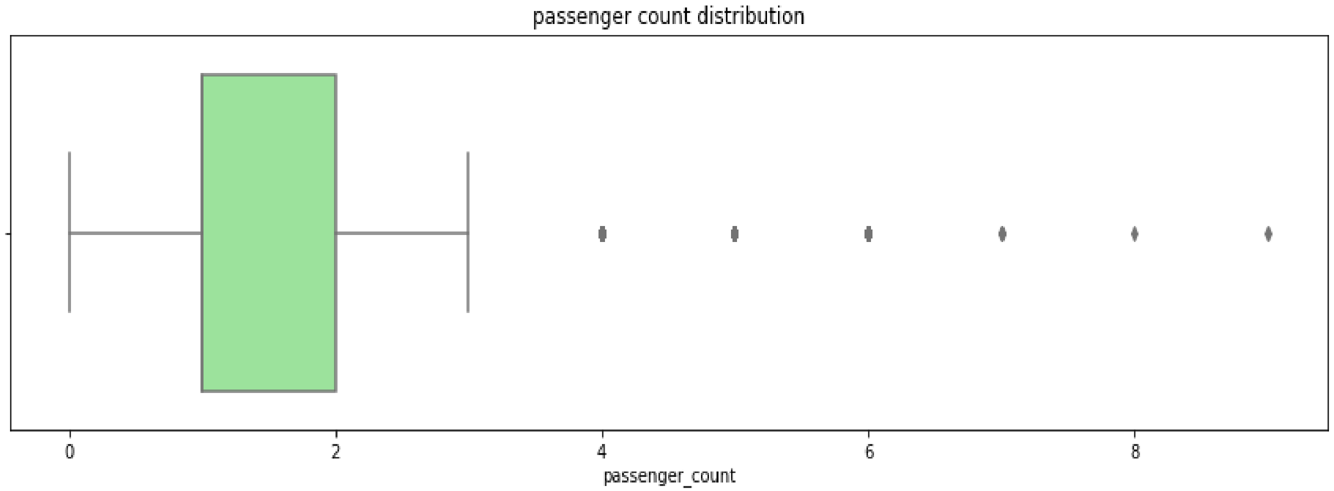
- The log transformation made the data conform more closely to a normal distribution.

- By taking logarithm of trip duration we can smoothen those variables.

Data Analysis :

- We have almost all the trips records send to the servers.





- Most number of trips are done by 1-3 passenger(s).

Feature Engineering:

DISTANCE

AVG_SPEED

PICKUP_DAY_NAME

DROPOFF_DAY_NAME

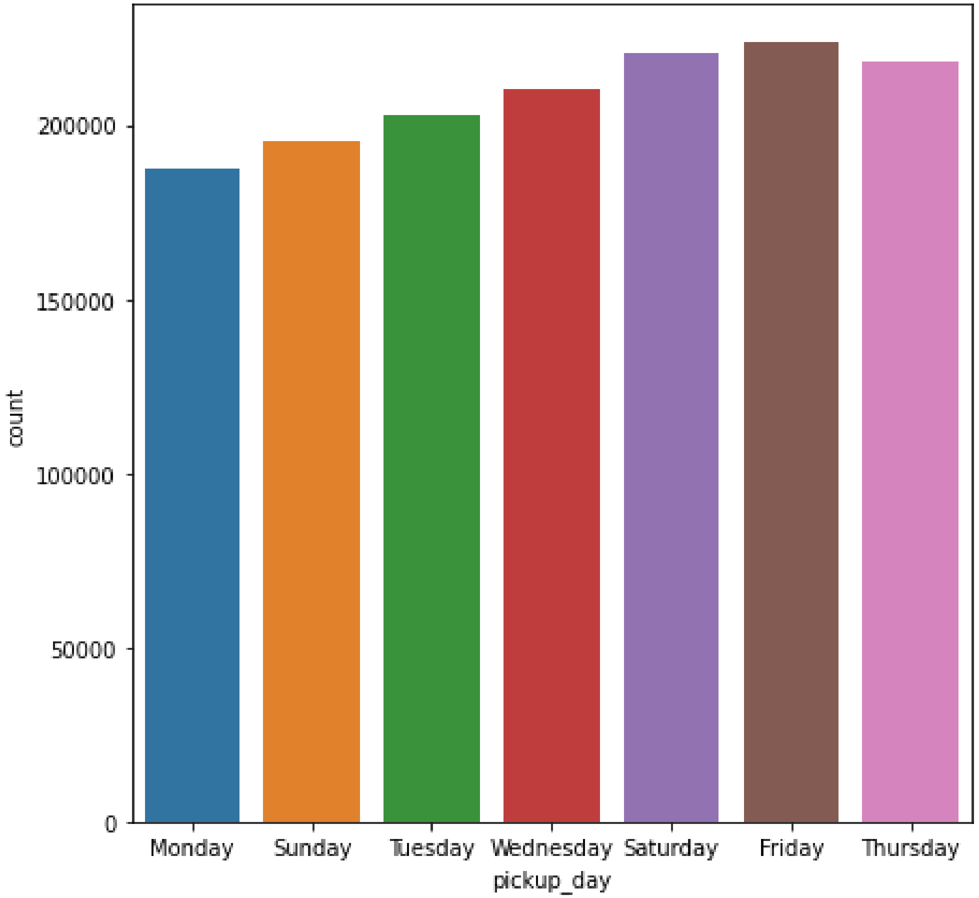
PICKUP_TIME

DROPOFF_TIME

PICK_TIME_FRAME

DROP_TIME_FRAME

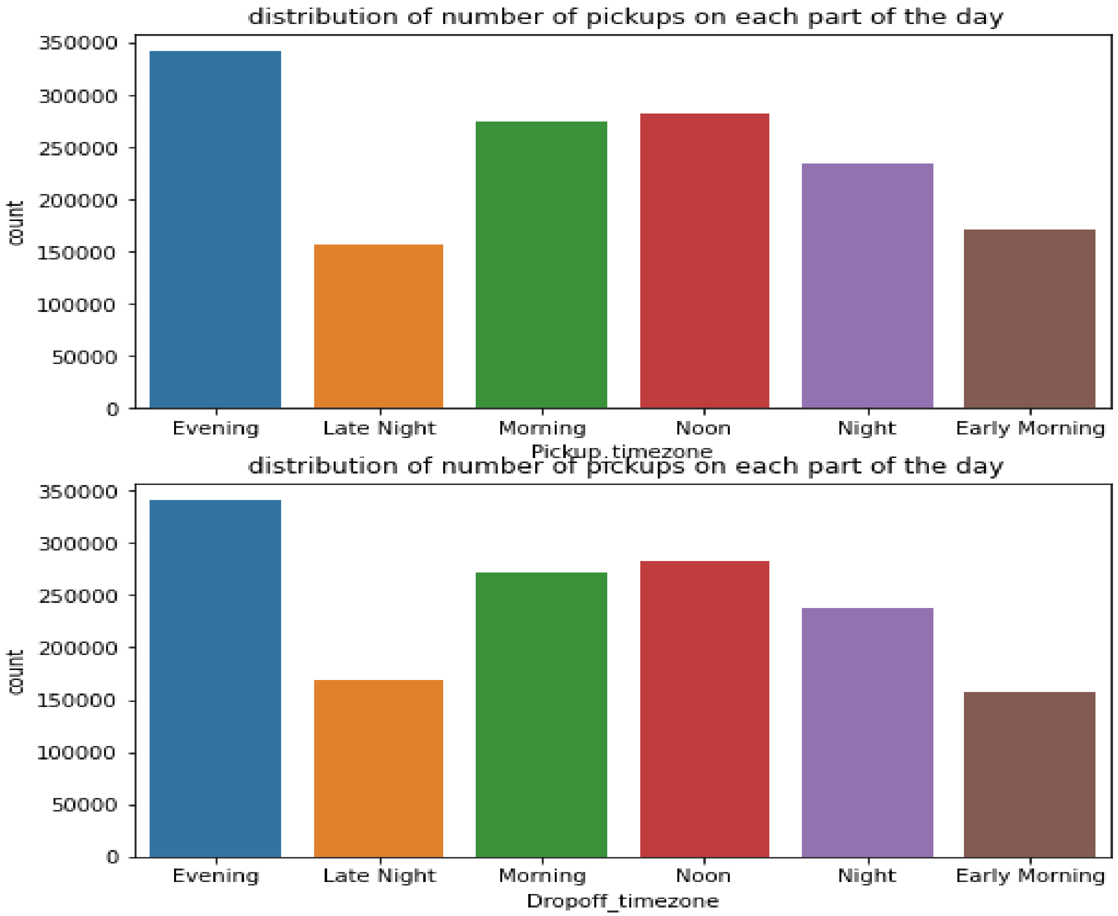
Data Analysis :



- Observations tells us that Fridays and Saturdays are those days in a week when New Yorkers prefer to rome in the city.
- Increasing trend is observed in the number of trips from Monday to Friday and it decreases on the weekends.

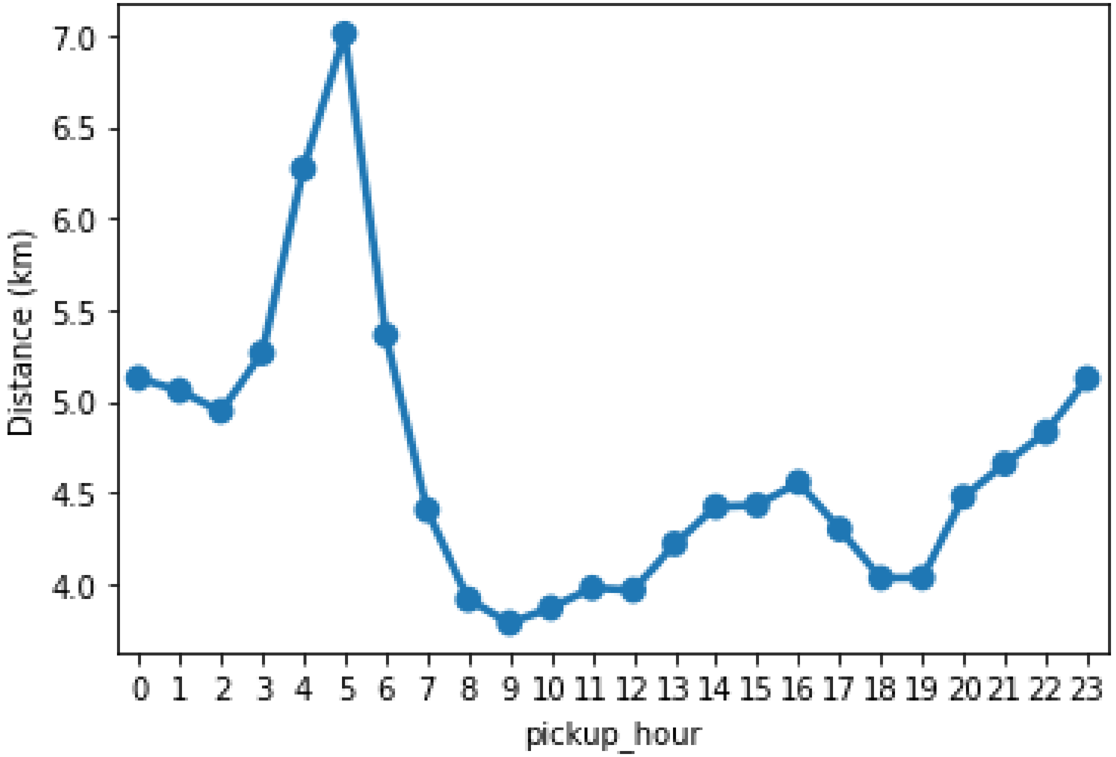
Data Analysis :

- Most number of trips are during the evening time and the least number of trips is between late night to early morning



Data Analysis :

DISTANCE



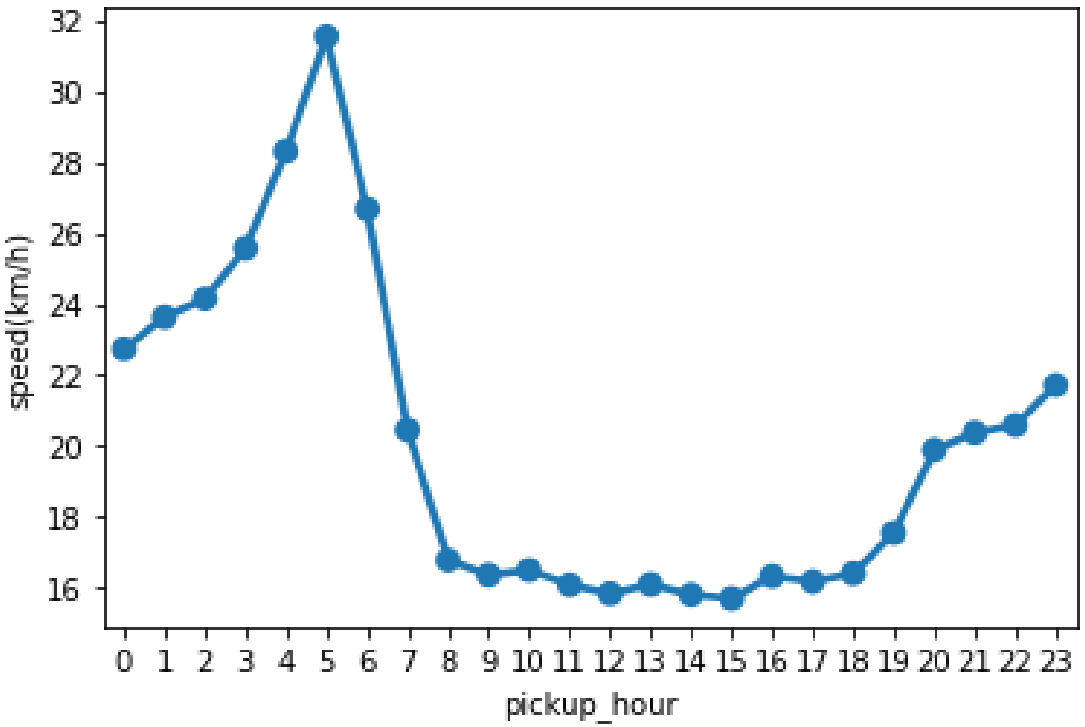
- Trip distance is fairly equal from morning till the evening varying around 4 - 4.5 kms.

- It starts increasing gradually towards the late night hours starting from evening till 5 AM and decrease steeply towards morning.

- Trip distance is highest during early morning hours which can account for some things like: Outstation trips taken during the weekends. Longer trips towards the city airport which is located in the outskirts of the city.

Data Analysis :

SPEED

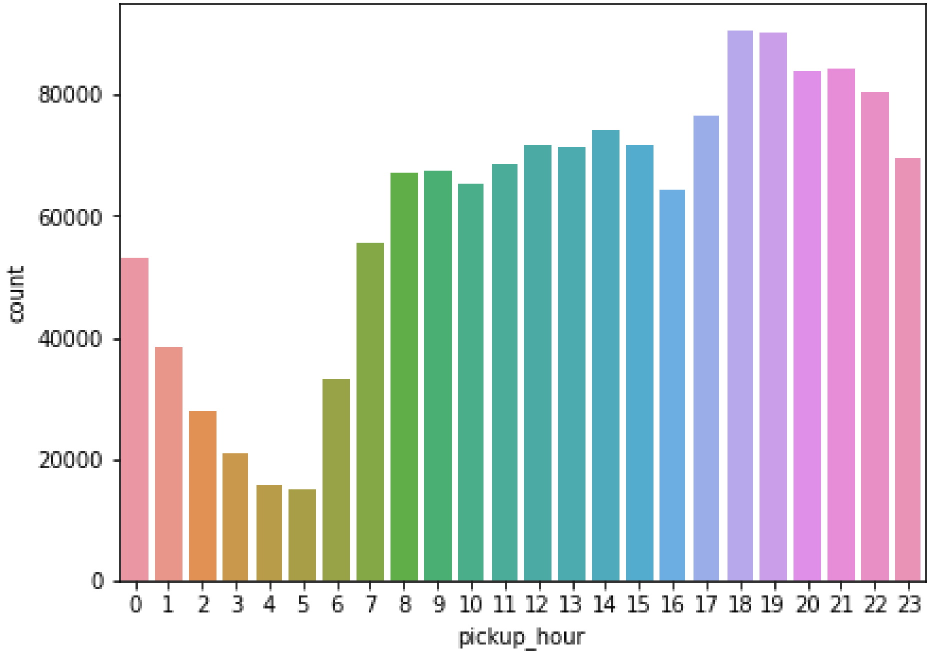


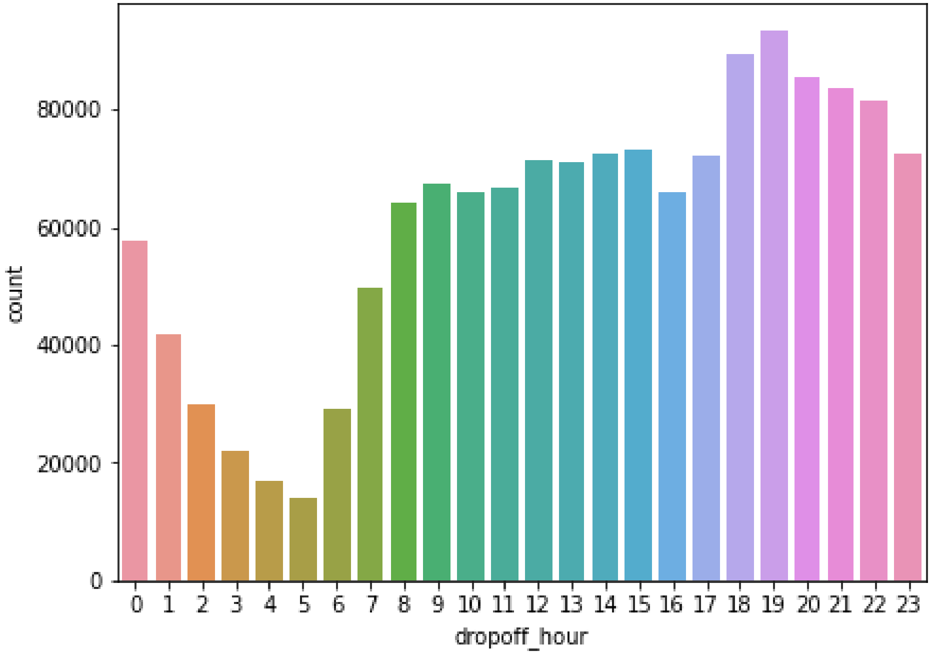
- Average speed tend to increase after late evening and continues to increase gradually till the late early morning hours.

- Average taxi speed is highest at 5 AM in the morning, then it declines steeply as the office hours approaches.

- Average taxi speed is more or less same during the office hours i.e. from 8 AM till 6PM in the evening.

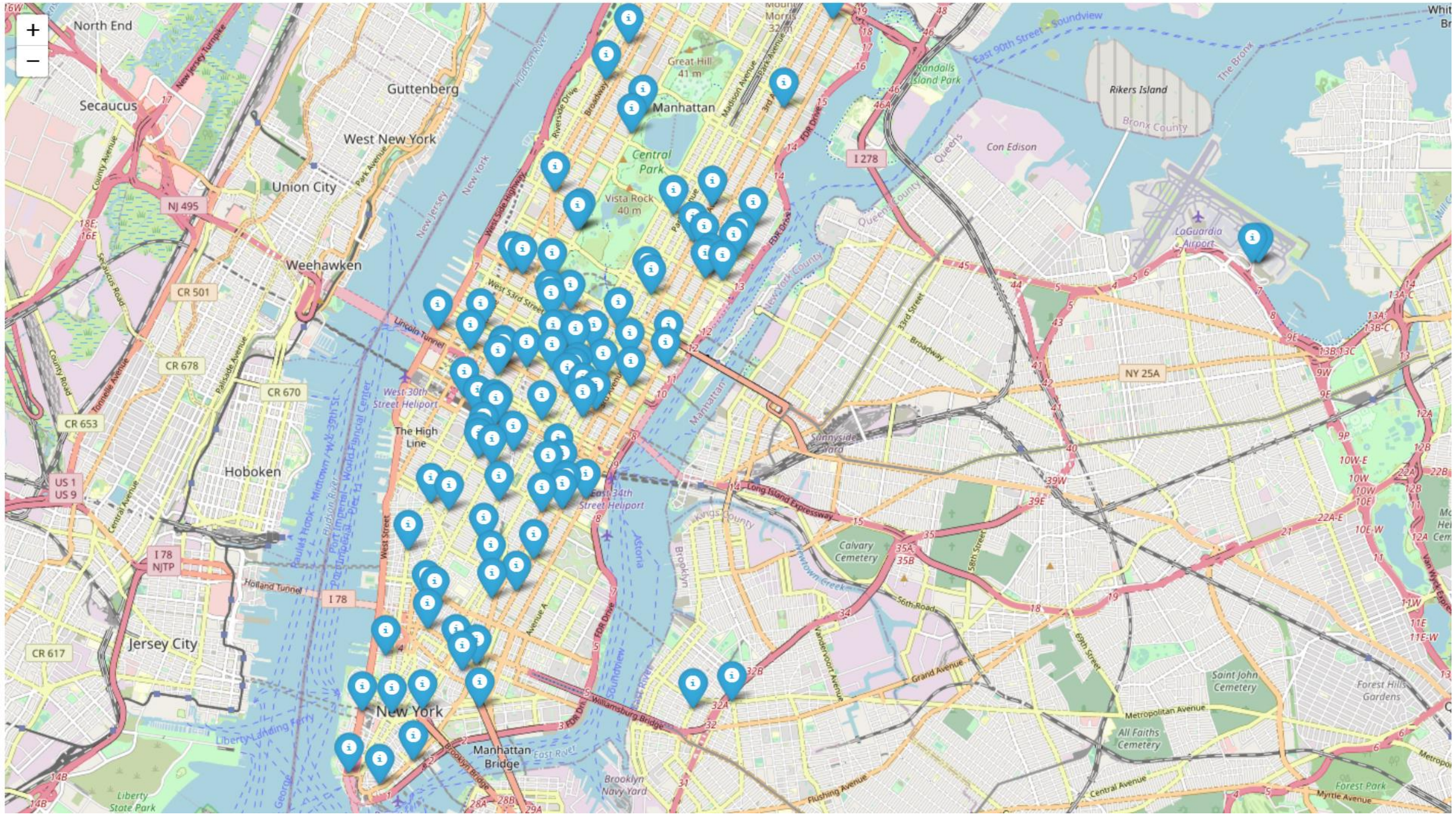
Data Analysis :



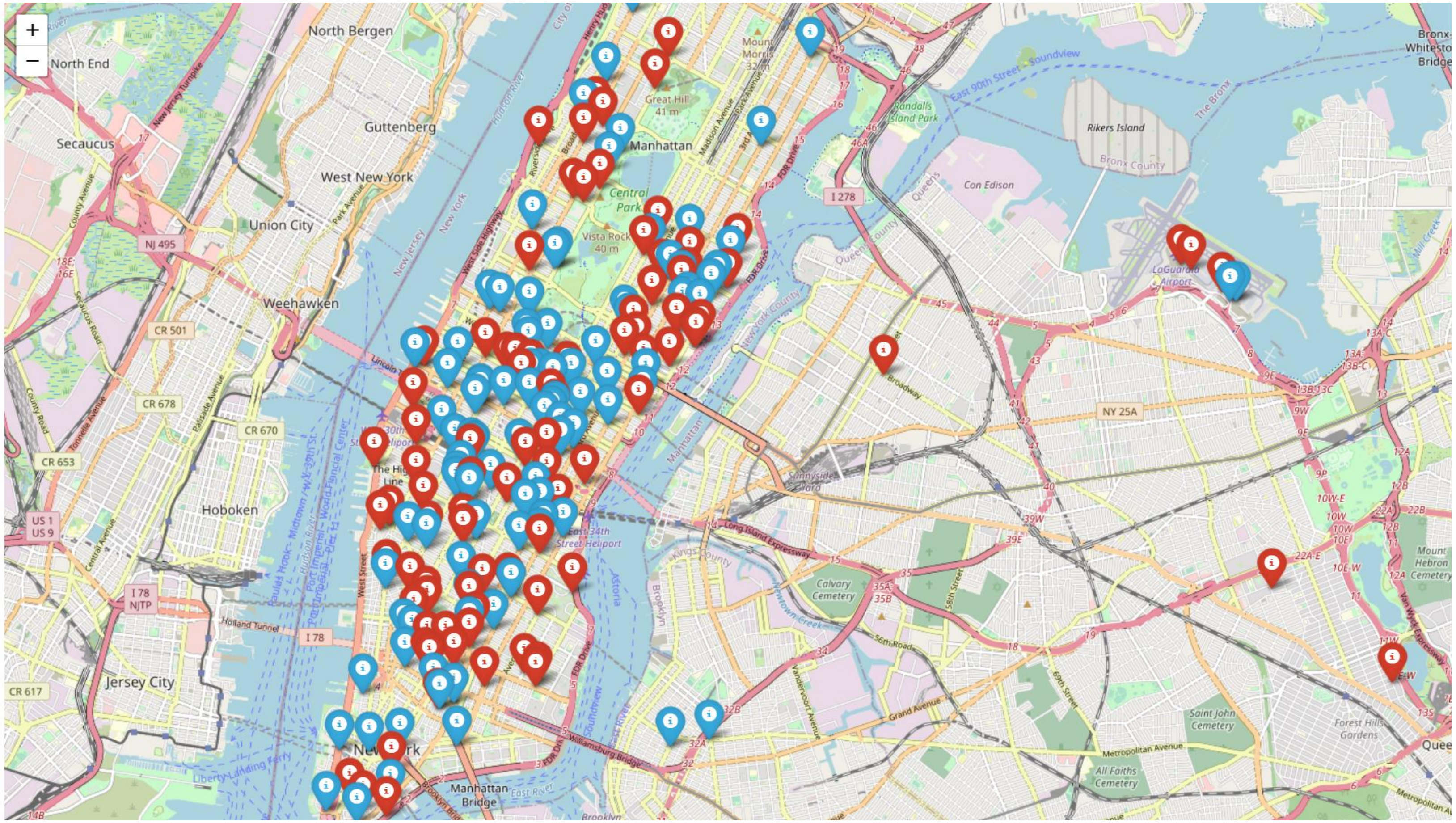


- We see the busiest hours are 6:00 pm to 7:00 pm and that makes sense as this is the time when people return from their offices.

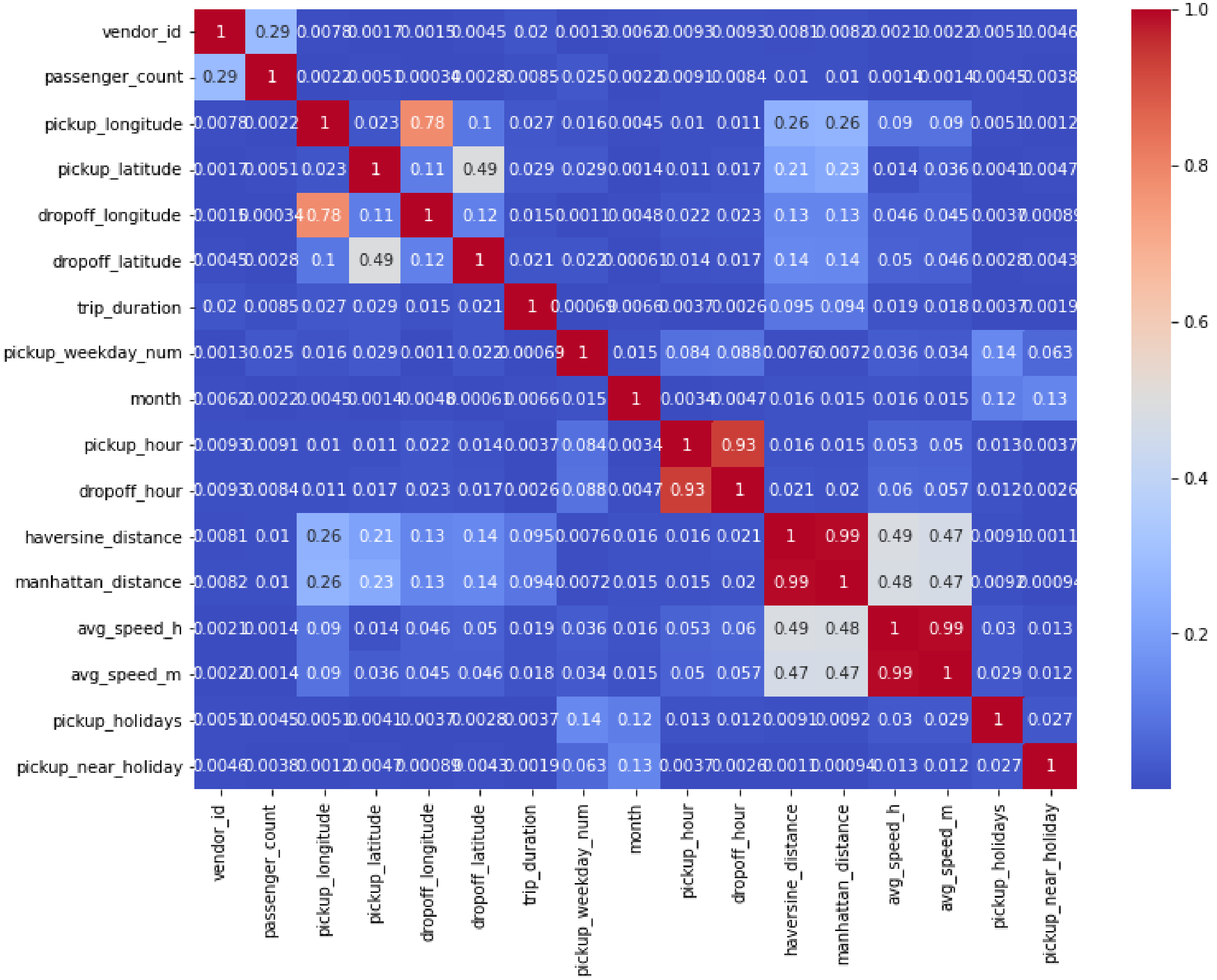
Distribution of Pickup Latitude and longitude



Distribution of Dropoff Latitude and longitude



Correlation Heatmap:

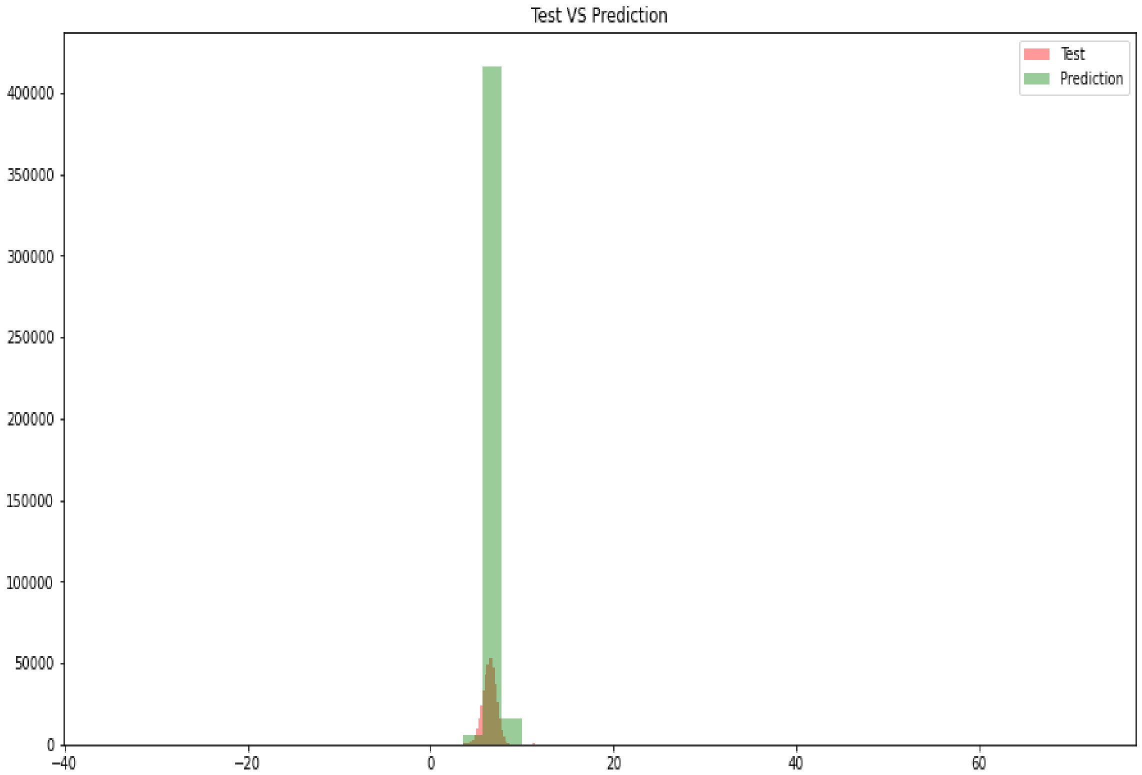


Building Model

We need a model to train on our dataset to serve our purpose of prediciting the NYC taxi trip duration given the other features as training and test set. Since our dependent variable contains continous values so we will use regression technique to predict our output

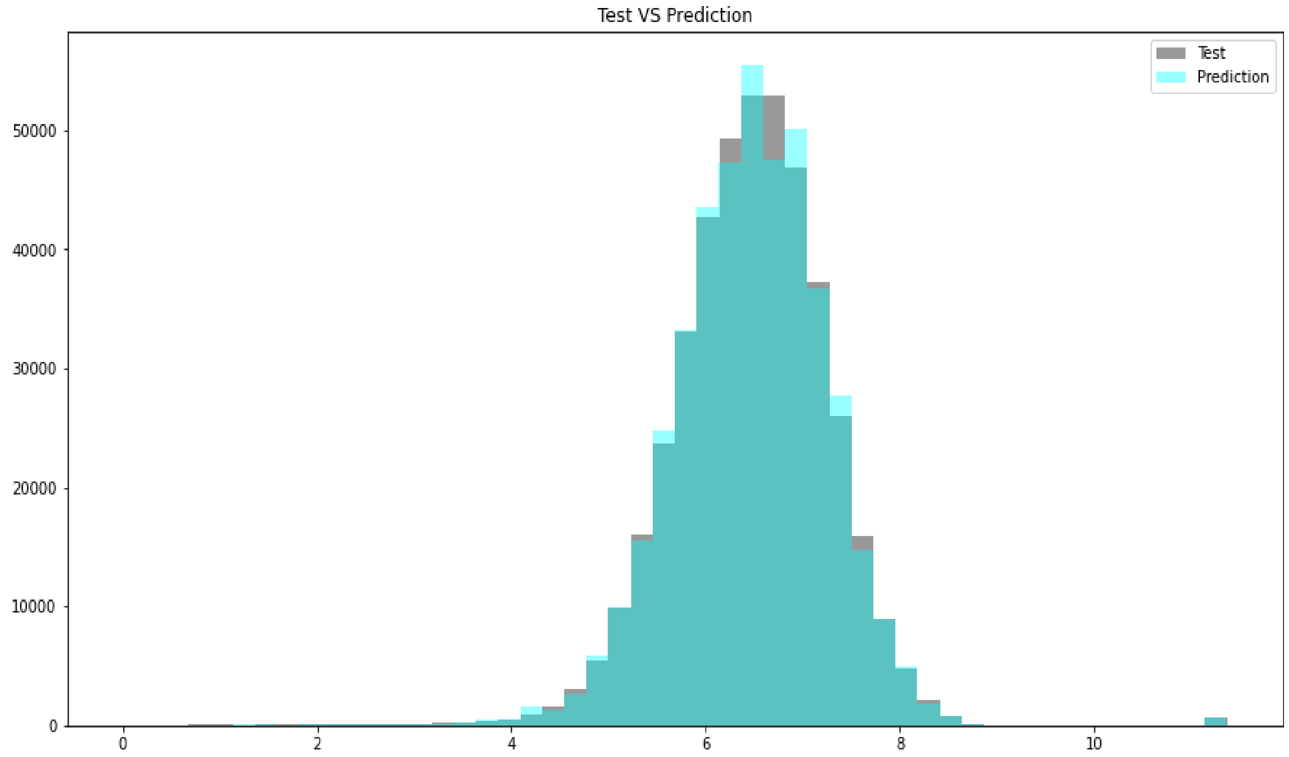
- Linear Regression
- Random Forest
- Decision Tree

Linear Regression:



- From the Viz. we can clearly identify that the Linear Regression isn't performing good. The Actual Data (in red) and Predicted values (in green) are so much differing. We can conclude that Linear Regression doesn't seem like a right choice for Trip duration prediction.

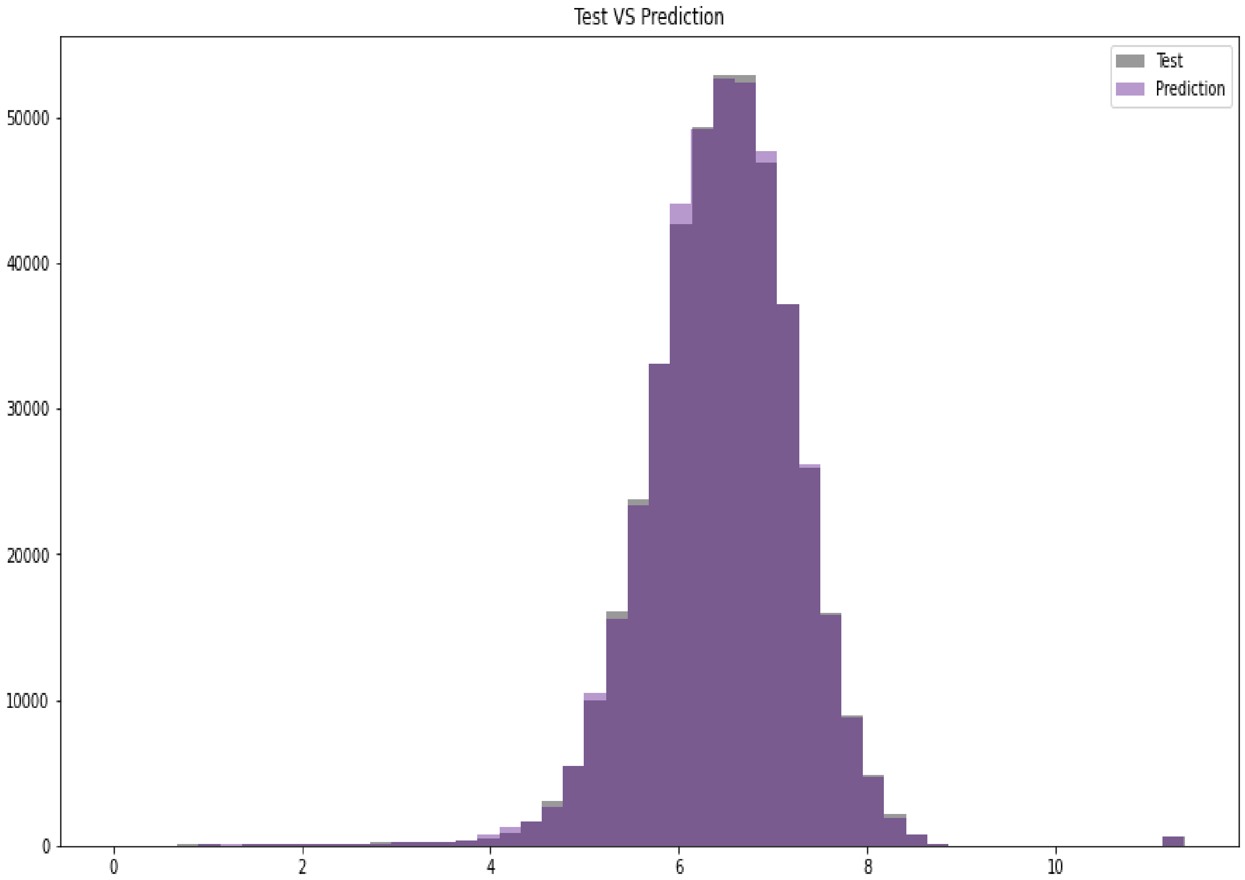
Decision Tree:



- Here is approx 200% improvement on the R2 score for the Decision tree regressor over the Linear regressor of the feature selection group

- From the above Viz. we can clearly identify that the Decision Tree Algorithm is performing good. The Actual Data (in Grey) and Predicted values (in Red) are as close as possible. We can conclude that Decision Tree could be a good choice for Trip duration prediction.

Random Forest:



- From the Viz. we can clearly identify that the Random Tree Algorithm is performing good. The Actual Data (in Grey) and Predicted values (in purple) are as close as possible. We can conclude that random forest could be a good choice for Trip duration prediction.

- The performance of Random Forest is also better than Linear regressor and somewhat similar to the decision tree but There is slight difference in r2 score and rmsle value.

Comparing the Model:

	Training Score	Validation Score	Cross Validation	R2_Score	RMSLE
• Linear Regression	0.485658	0.490973	0.408930	0.490973	_
• Decision Tree	0.976217	0.975402	0.975147	0.975402	0.024213
• Random Forest	0.979708	0.979119	0.979300	0.978531	0.023027

Conclusion on EDA:

- Vendor 2 has significantly more number of trips than Vendor 1.
- Around 73% of the trips have only one passenger with some anomalies of 0, 7, 9 passengers.
- Most of the trips were less than 10kms and were are of short duration as well (10-14 minutes)
- Evenings had the maximum number of taxi trips whereas it was the least during Early Mornings
- Increasing trend is observed in the number of trips from Monday to Friday and it decreases on the weekends.
- Average speed tend to increase after late evening and continues to increase gradually till the late early morning hours.
- The correlation heat map shows that there is not much correlation among the independent and target variables, except for slight correlation among latitude and longitudes.

CONCLUSION ON MODEL TRAINING

- At the end we conclude our project with 3 models namely Linear Regression, Decision Tree and Random Forest.
- Decision Tree and Random Forest both model giving us a good score but One problem that might occur with Decision Tree is that it can overfit.
- It gets overfitted on training data which couldn't predict well on unseen data.
- A random forest chooses few number of rows at random and interprets results from all the tress and combines it to get more accurate and stable final result.
- Among this Random Forest performs the best on our dataset as we saw that the RMSLE values came out to be the least for the same.

Thank you!