## **NYC Taxi Trip Time Prediction**

**Capstone Project Using Regression Machine Learning**

**Monica Patel**

**(Data science Enthusiast at Alma Better Pro Member Bangalore)**

**Introduction**

### **The Dataset Is Based On The 2016 NYC Yellow Cab Trip Record Data Made Available In Big Query On Google Cloud Platform. The Data Was Originally Published By The NYC Taxi And Limousine Commission (TLC). The Data Was Sampled And Cleaned For The Purposes Of This Project. Based On Individual Trip Attributes, You Should Predict The Duration Of Each Trip In The Test Set.**

**Our Experiment Can Help Understand The Reason For A Taxi Time Trip By Feature Selection, Data Analysis And Prediction With Python.**

**Problem Statement**

### **My task is to build a model that predicts the total ride duration of taxi trips in New York City. Mine's primary task dataset is one released by the NYC Taxi and Limousine Commission, which includes pickup time, geo-coordinates, number of passengers, and several other variables.**

**About data**

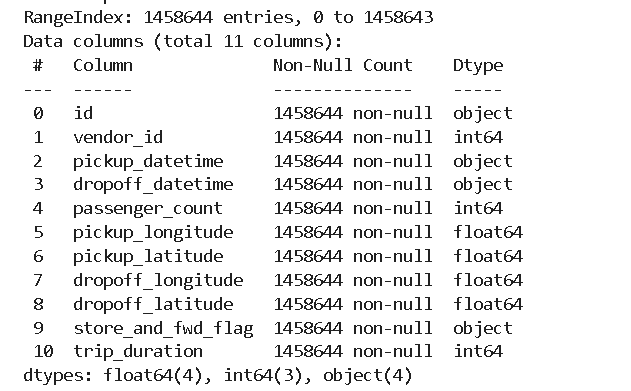
**So, in our dataset we have**

* **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.**

1. **Data Handling And Feature Engineering:**

**3.1. Data loading and Checking**

The data is published in 2016 by NYC Taxi and Limousine Commission and made available to the public via the Google clouds platform.



We loaded the data from the drive using pandas.read\_csv function. Our data had 1458664 rows and 11 Columns.

The dataset contains data of type object, int, and float.

**3.2. Checking for Null values and Duplicates**

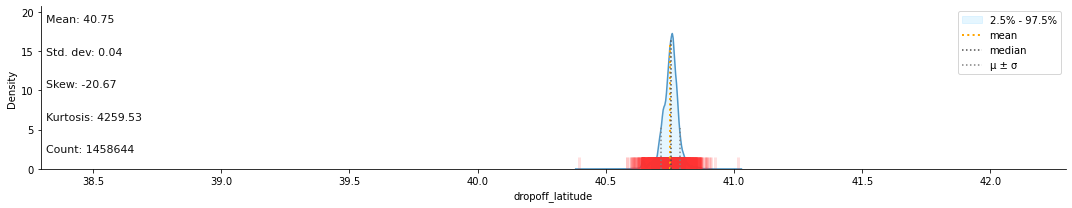
As we checked, it was noted that there were no NULL values throughout. Hence it was not needed to clean them.

There were no duplicate entries.

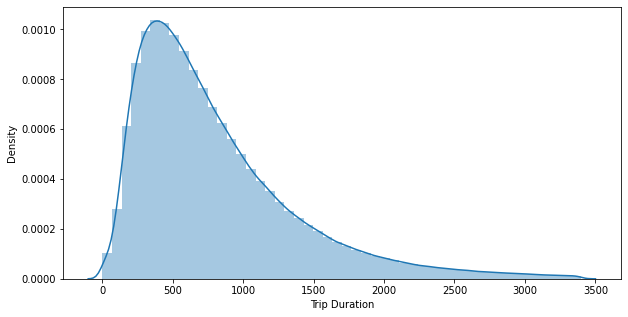
The pickup\_datetime and dropoff\_datetime columns were given in the object format. But it would be better to handle them if they are in date time format. Therefore, we have changed them to the appropriate format.

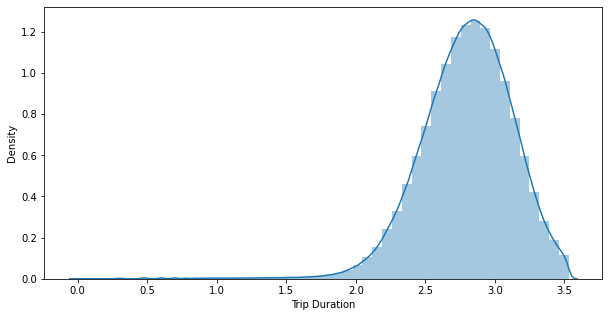


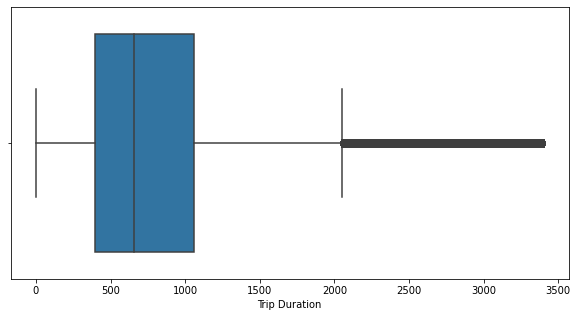


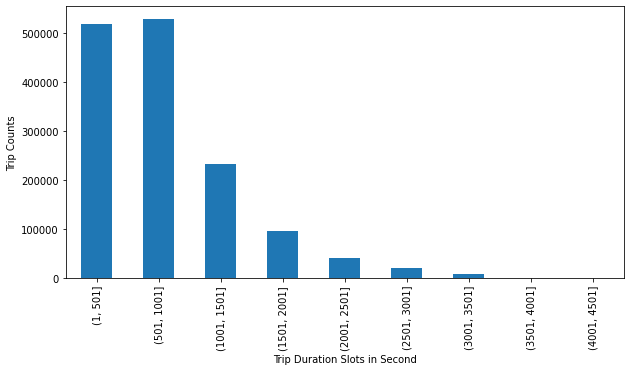


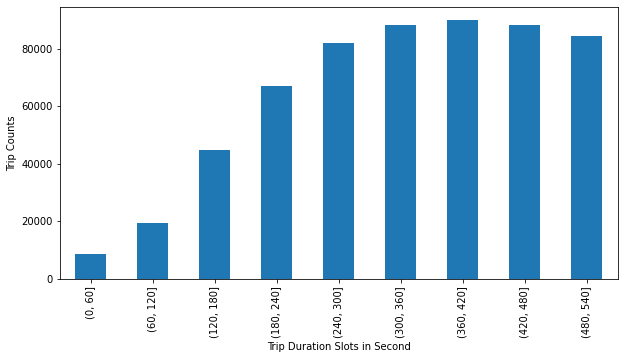


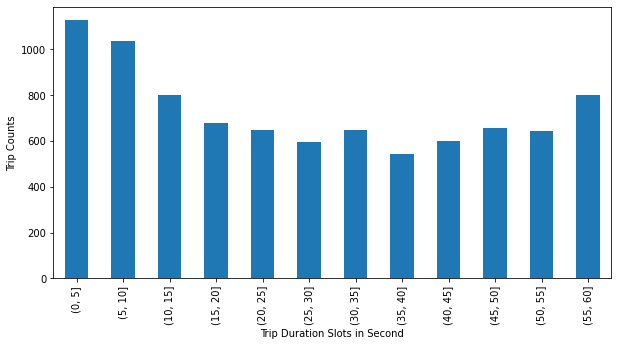


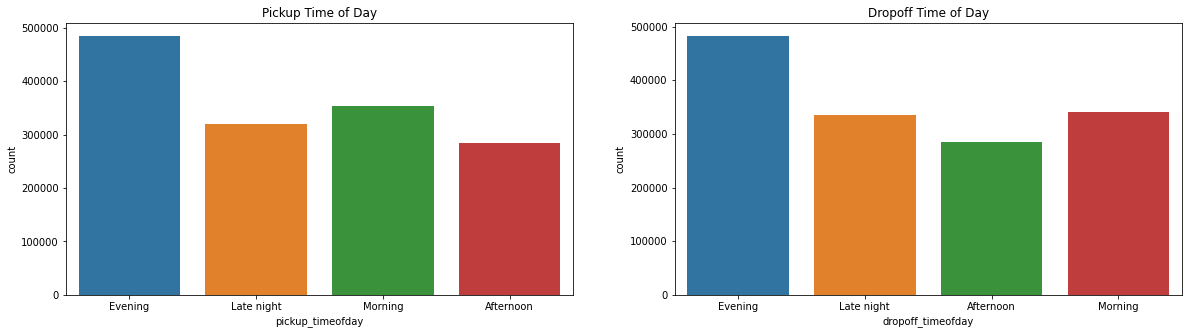


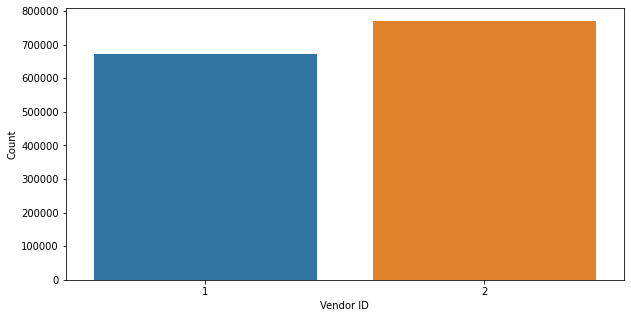


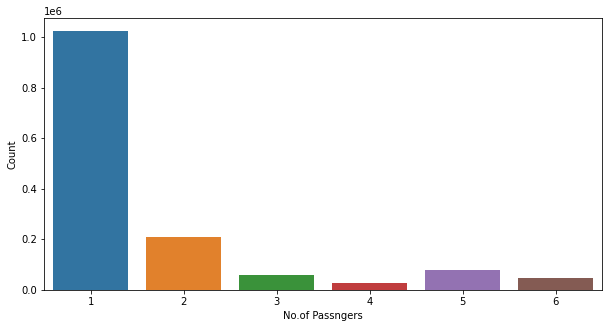




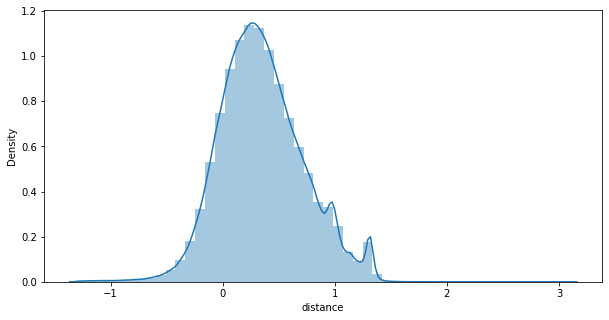




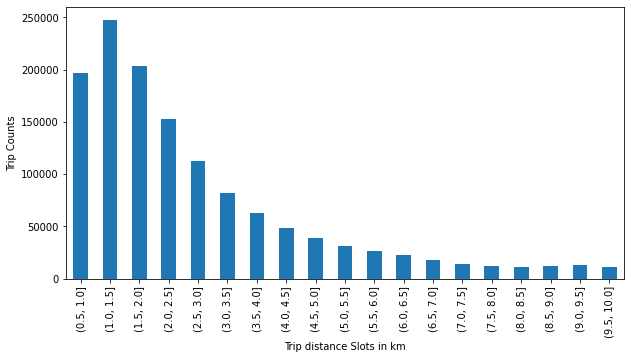


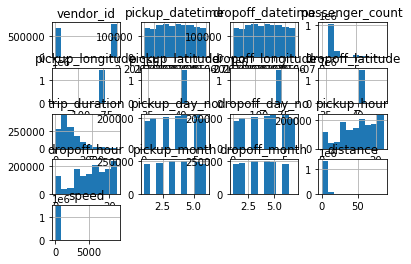


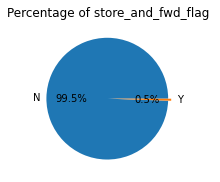


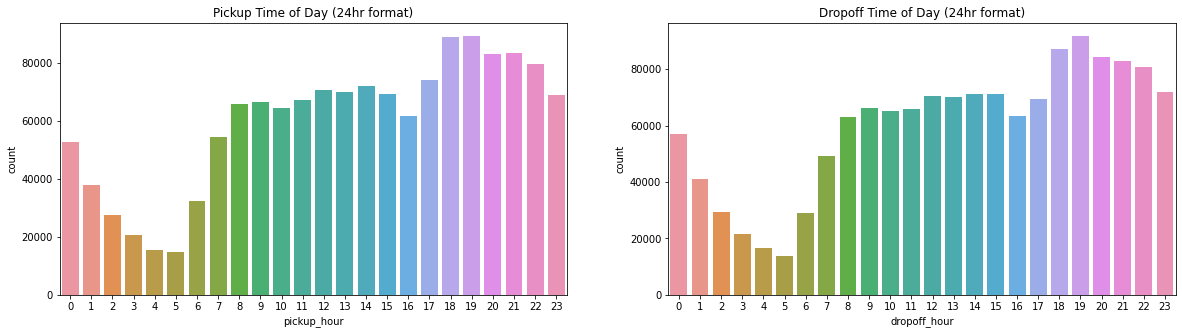


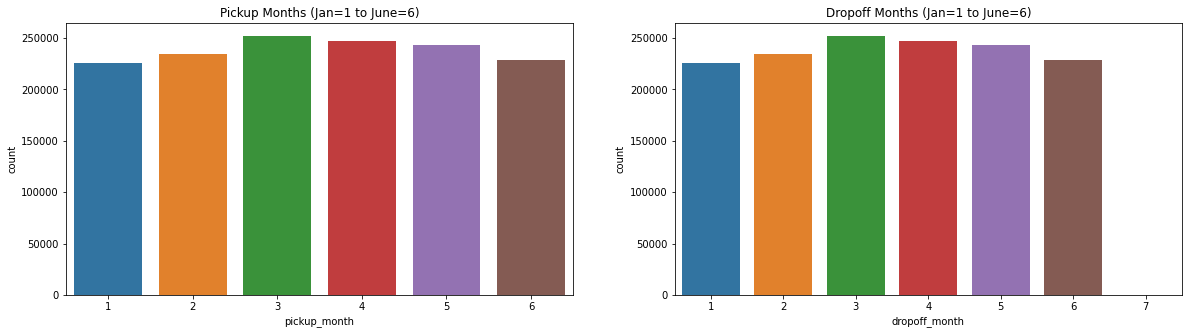


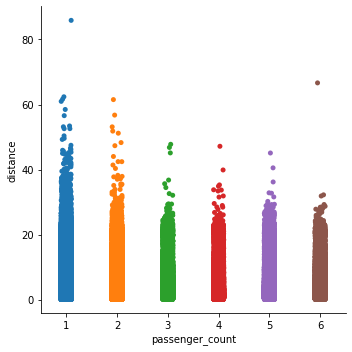


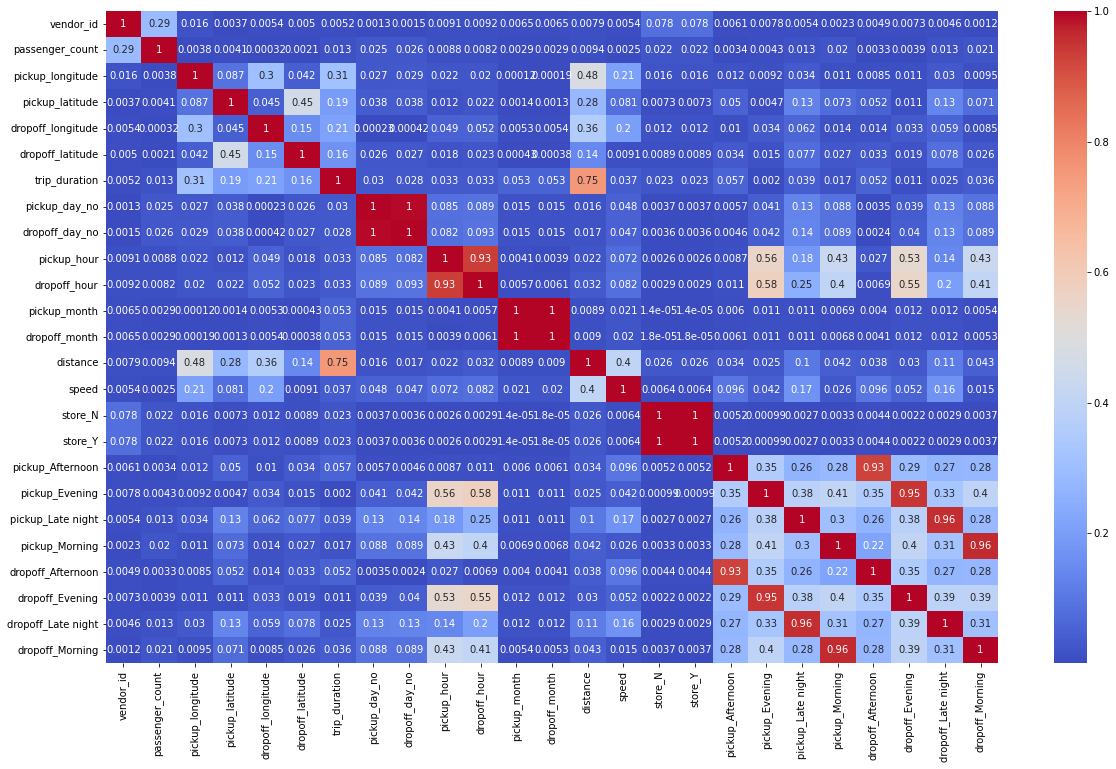


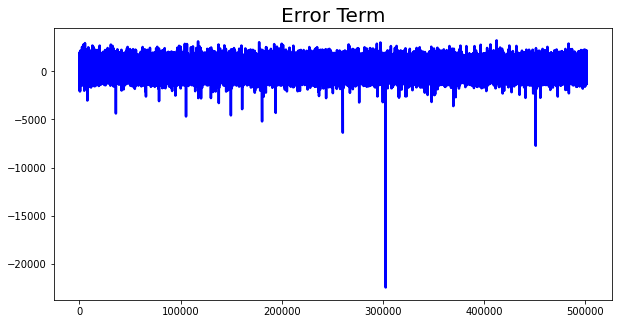


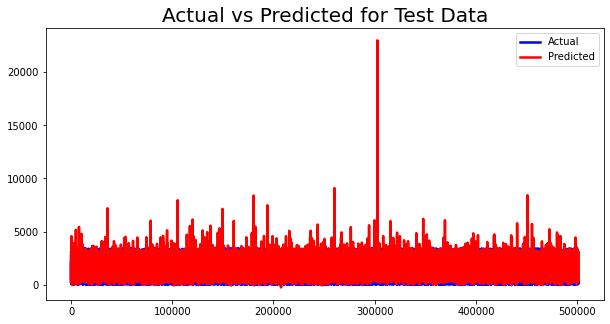




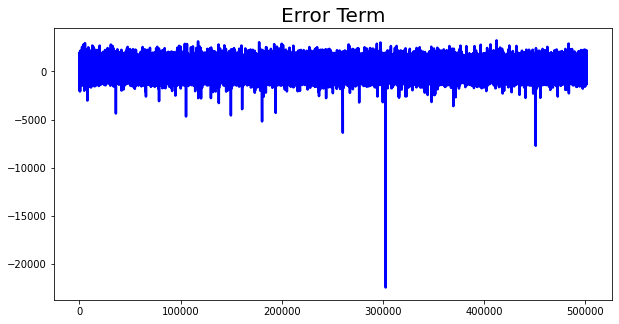


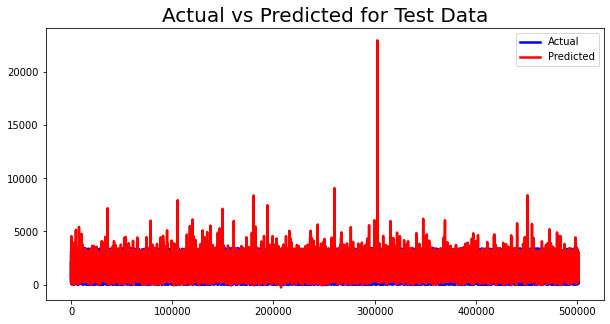


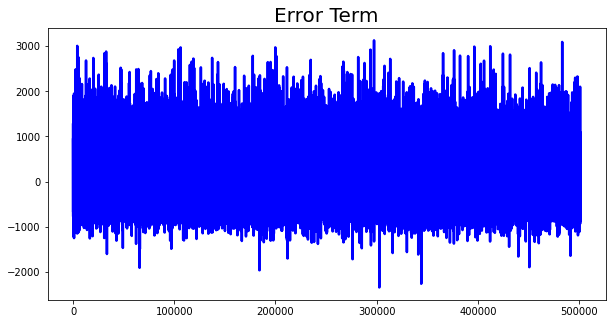


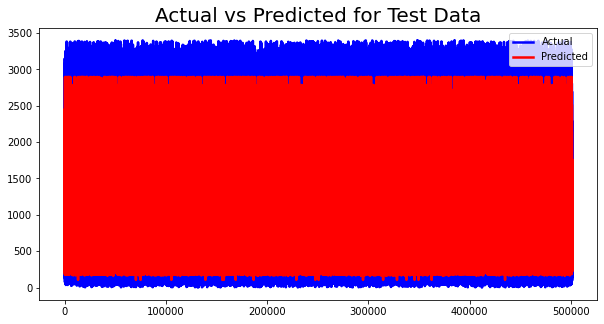


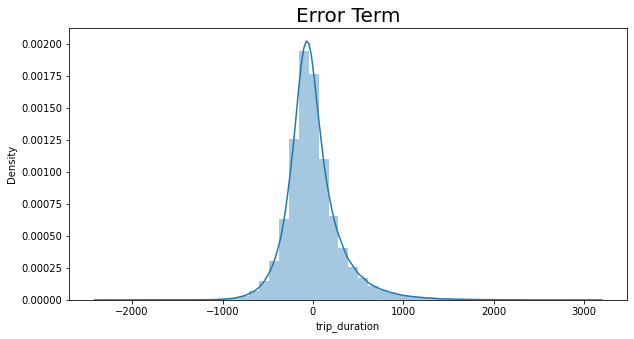












* 1. **. EDA**

As the first step, we checked the relation between variables vendor id, (stored and fwd flag) and no of trips. It was found that the vendor with id 2 has taken more trips. There was a very less number of trips that were stored in the memory of the vehicle before the commencement of the trip. Then we analyzed the number of passengers in a trip. It was found that most people travel alone. By using the Great Circle method, we found out the distances. We plotted boxplots for numerical features and identified the outliers. By segregating the duration into different classes it was found that most of the trips were of duration between one min and one hour. By using Interquartile ranges and the concept of outliers we tried to remove outliers. The latitudes and longitudes outside the city limit were not considered for further analysis. We created some new features such as pickup hour, pickup minute, and pick up second to draw some more inferences. We found out the peak time of travel, the month with the maximum number of trips, the part of the day in which demand for the taxis is high, variation of trip duration along with weekdays and weekends.

**3.4. Handling categorical features**

It is mandatory to encode the categorical features into numerical values otherwise the scikit library won’t recognize them. By using dictionary mapping, we successfully converted those variables into the appropriate format.

**3.5. Feature Standardization**

In the dataset, there is a vast difference between the range of independent features. To get a better performing model we need to

**3.6. Fitting Different Models**

We used many ML models to get better results. We used Linear Regression along with regularization methods Lasso and Ridge, Decision Trees, and Xgboost. To get the best parameters in each model, Cross-validation is performed. It ensures that we use the best hyperparameters in the model to get desired results.

1. **Algorithms:**

In this dataset, we have one dependent variable whose value shall be predicted and many independent input features. The dependent variable is continuous, so this is the case of a Supervised Regression Model. Hence, we have used some of the most commonly used Regression algorithms in this project.

**4.1. Linear Regression**

This is the algorithm which works on the assumption that the relationship between the input and output variables in linear.

**4.3. XGBoost**

It stands for Extreme Gradient boosting. Here, several decision trees are trained sequentially. The weights are assigned to independent variables and decision trees are constructed. Weights of values predicted wrongly are increased and the next decision tree is built. It results in a reliable, low bias and low-variance model.

**Conclusion**

That's it! We reached the end of our exercise.

Starting with loading the data so far we have done EDA , an important library and visualization.

* Most of the trip consists of 1 or 2 passengers.
* Vendor 2 is evidently more famous among the population.
* Most pickups are between 5 PM to 9 PM.
* Fridays and Saturdays are those days in a week when people prefer to roam in the city.
* Average distance traveled is approx 3.5 kms.
* Most of the trips are limited to the range of 1-10 kms.
* Most of the rides are 10 min.
* There were very few trips (0.55%) of which the records were stored in memory.
* Average trip duration is lowest at 6 AM when there is minimal traffic on the roads.
* Most of the taxi pickups were done in the Manhattan area as compared to the other areas in NYC.
* Trip distance is highest during early morning hours.
* Possible reason behind this is Outstation trips taken during the weekends.
* Trip distance is fairly equal from morning till the evening
* Sunday being at the top may be due to outstation trips or night trips.