# EDA Capstone Project-2

## New York Taxi Trip Time Prediction

Submitted by- Shreesha K

# A brief introduction

- Generally, the duration of a trip can be calculated by dividing the distance travelled by the average speed of the vehicle.
- But, this is not the case in larger cities where a lot of other factors affect the overall trip duration.
- According to the official source there were 13,587 taxis in the New York city (as per 2016) and nearly 55 percent of the people do not own a vehicle. Hence, a large population depends on other modes of transportation.
- A customer is always concerned about the time taken for his journey. Therefore, it becomes important to estimate the time taken for a journey considering the effect of all the relevant factors.

# Contents

- Describing the problem.
- About the dataset
- Exploratory Data Analysis
- Feature Engineering
- ML algorithms used
- Comparing the different algorithms
- Conclusion

# Problem statement

- We need to build a Machine Learning model which predicts the total travel duration of taxi trips in the New York City.
- The dataset available for this task is one released by New York city Limousine and Taxi commission in the year of 2016.
- It includes columns like pickup and drop-off datetime, coordinates of the pickup and drop-off location, number of passengers and some other variables as well.

# About the Dataset

- This dataset is based on the 2016 New York yellow cab trip record data. This was made available in Big Query on Google Clouds Platform.
- It consists of 1458644 rows and 11 columns.
- There are no NULL and duplicated values.

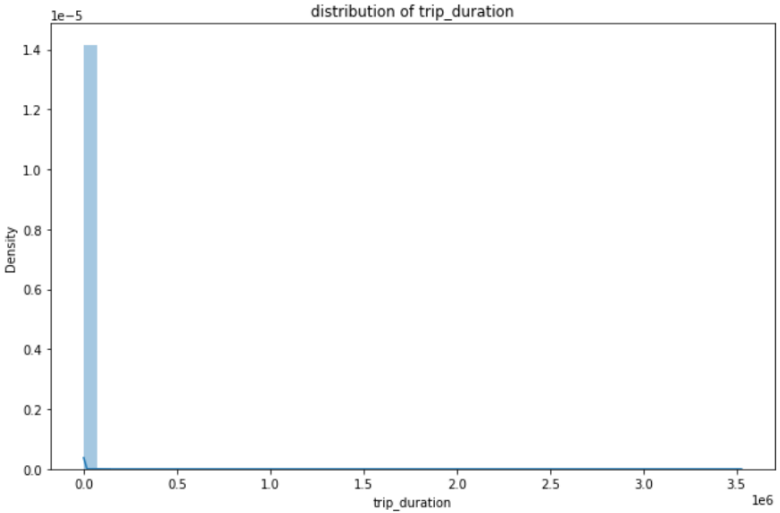
# Description of data

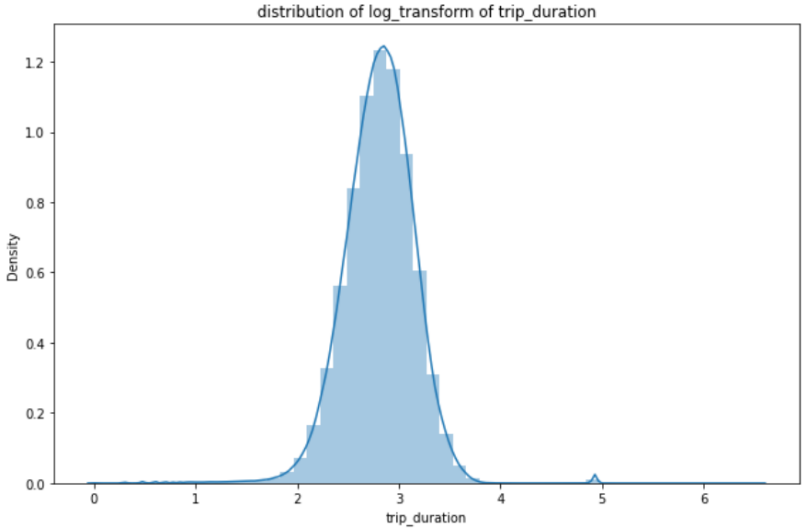
- Id: A unique ID for each trip.
- Vendor_id: An ID representing the service provider.
- Pickup_datetime: Indicates the time at which the meter was engaged.
- Dropoff_datetime: Indicates the time at which the trip was ended.
- Passenger_count: No. of passengers in each trip.
- Pickup_longitude and latitude: Corresponding to the location at which the meter was engaged.
- Dropoff_longitude and latitude: Corresponding to the location at which the trip was ended.

(Continued..)

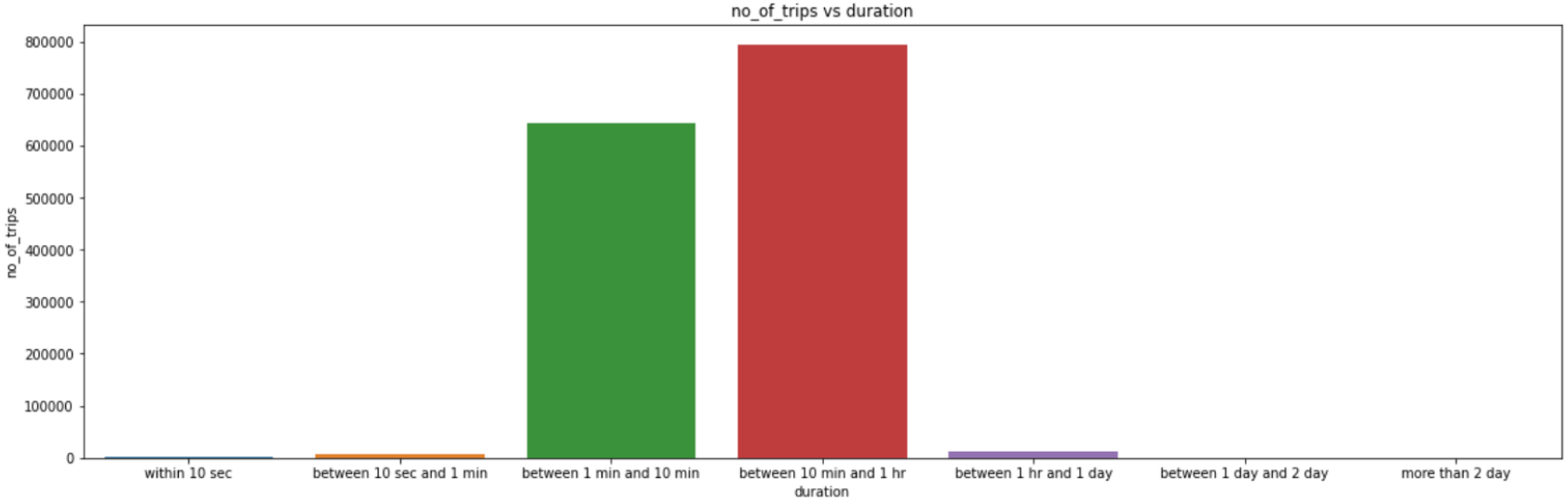
- Store_and_fwd_flag: This flag indicates whether the trip record was held in vehicle memory before sending to the vendor because the vehicle did not have a connection to the server - Y=store and forward; N=not a store and forward.
- Trip_duration: Trip duration in seconds. This is the target variable.

# Exploratory Data Analysis



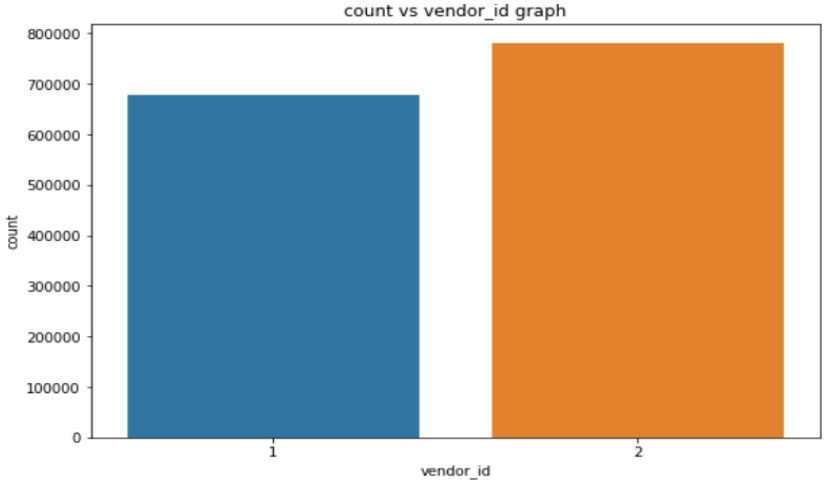



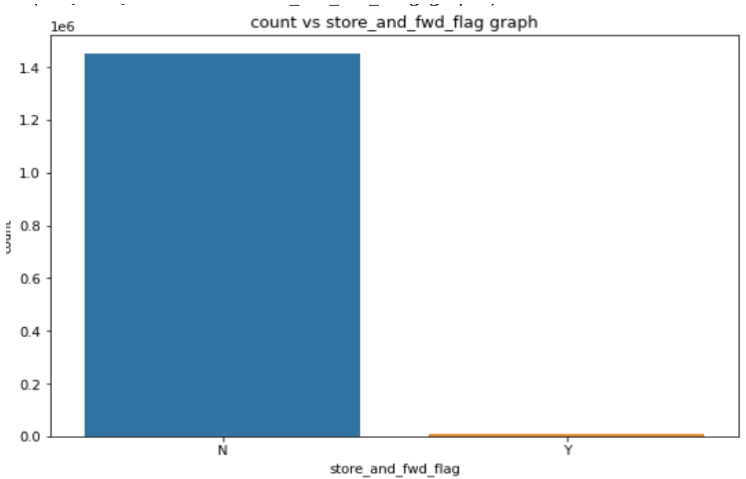

**The trip_duration is highly positively skewed with coefficient of skewness around 343. We took the logarithm of trip_duration and then plotted it's distribution and it was distributed almost normally.**



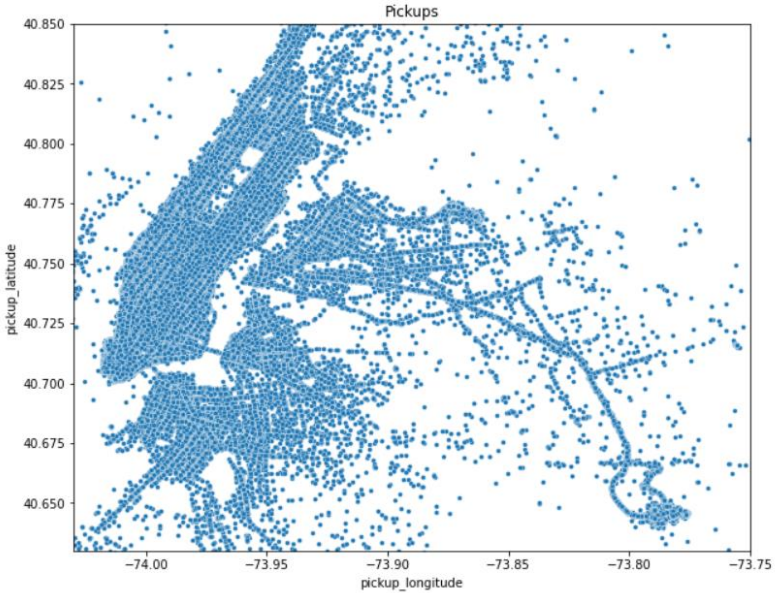

- **By dividing the trip duration into different parts it is seen that most trips were lying between 1 min and 1 hour duration.**
- **Some of the trips are even shorter than 10 seconds, indicating the possibility of recording a new trip unknowingly and then cancelling immediately.**

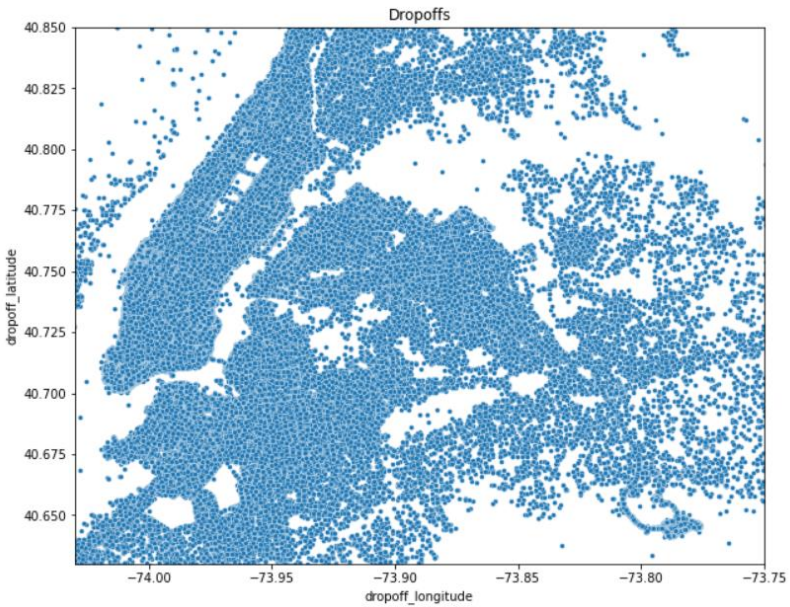
## Categorical features



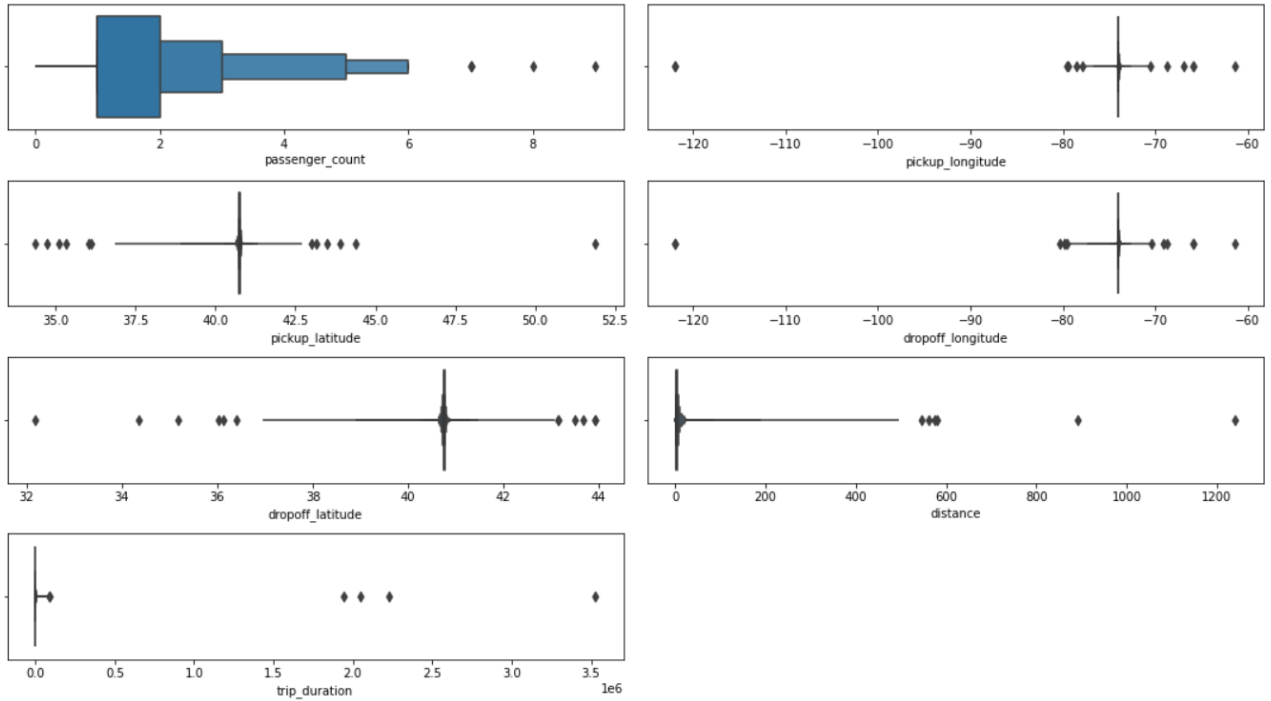




- It is seen that from the first graph, vendor with id 2 has performed more number of trips.
- From the second graph, it is observed that most trips were not saved in the memory of vehicle during the commencement of tip.



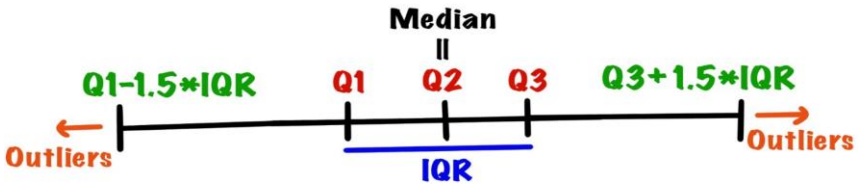



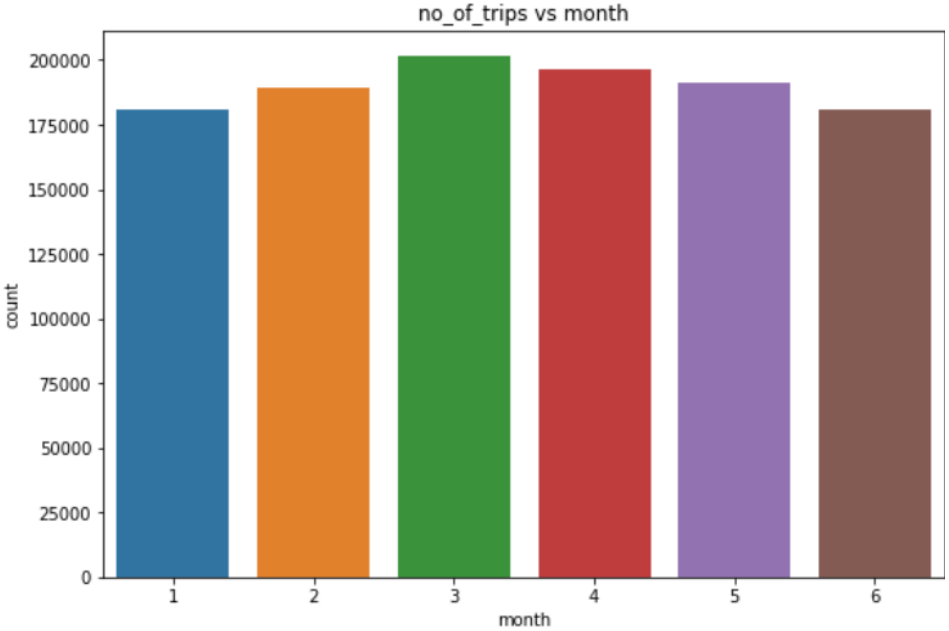

- **These plots represent the points of pickup and drop-off.**
- **The New York city lies within -74.3 and -73.75 longitudes and Latitudes are 40.63 N and 40.85 N. We considered only those trips whose starting and ending points were within this range.**

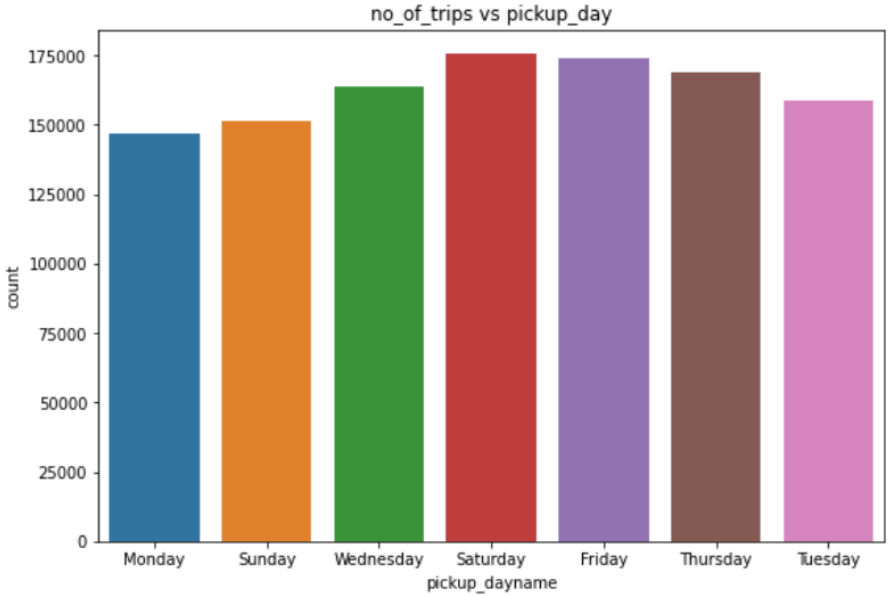



**After plotting these boxenplots, we can see that every feature has outliers.**

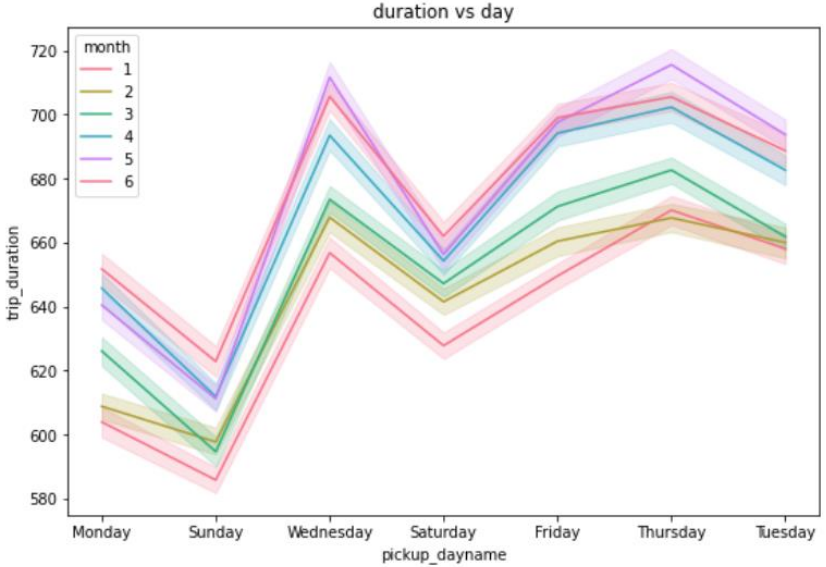
**In order to remove them, we used the concept of Inter quartile range.**

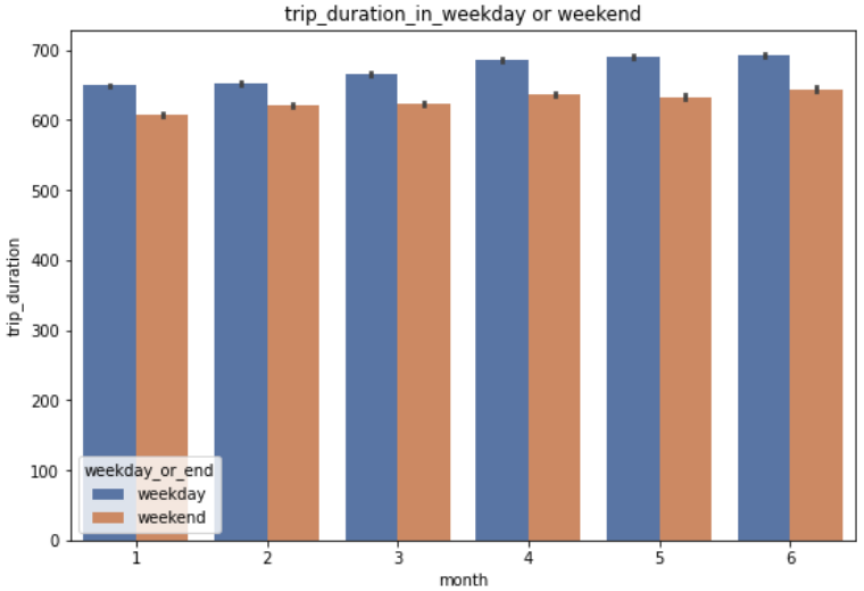




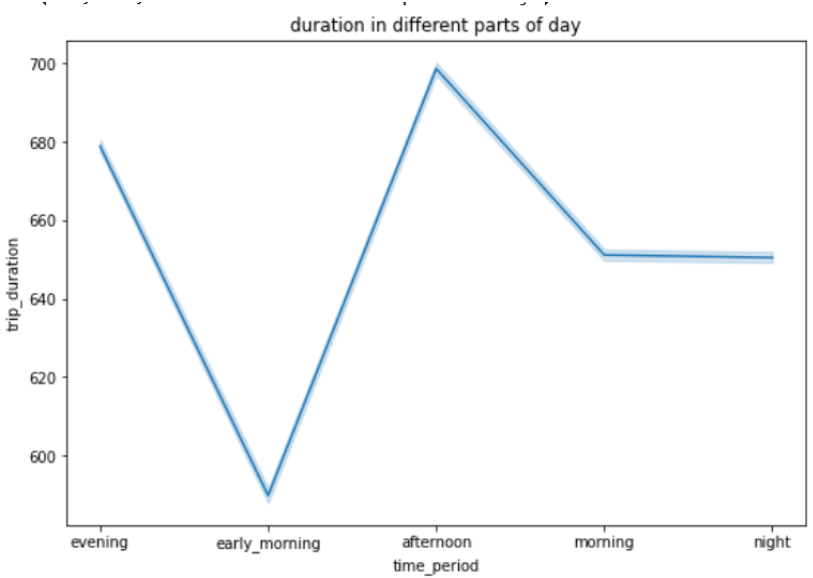



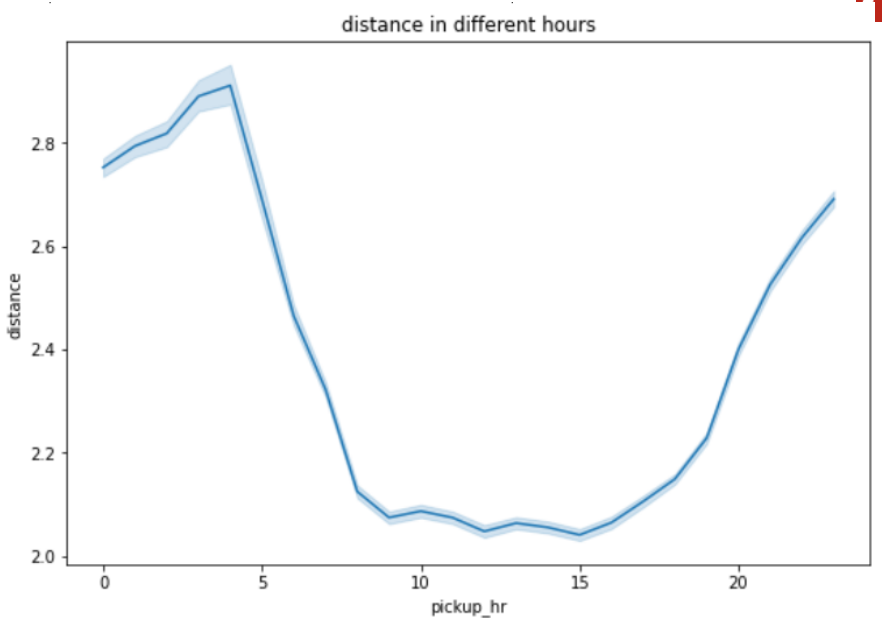

- **These graphs indicate the number of trips in different months and different days in a week. It is evident that in the month of March, most number of trips were occurred, with the least being in the month January.**
- **Saturday is the day with more number of trips and Monday has least number of trips.**



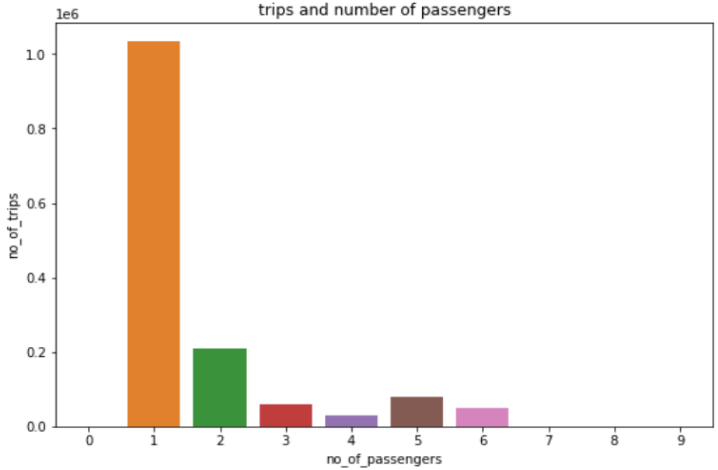

- From the first graph, we can see that January is the month with less duration, June being the month with maximum trip duration. It can also be seen that Wednesdays are the busier days in terms of travel duration.
- From the second graph, it is observed that during the weekends trip duration is lesser in comparison with weekdays.

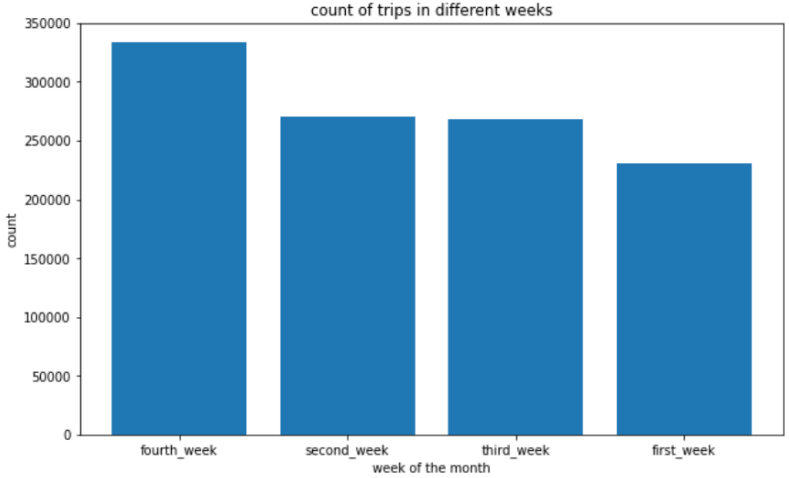



- **It can be observed that in the early morning trip duration is less and in the afternoon trip duration is maximum.**
- **But it can also be observed that distance travelled at 5 AM is the maximum in comparison with the distance travelled in afternoon hours.**

# Some other important observations




- Most passengers travel alone.
- On average, the trip duration is 10 minutes.
- The number of trips in the first week of a month are lesser, and maximum number of trips are done in the last week.
- Most longer trips begin at 5 AM. This can be due to outstation trips are usually started in the morning and these longer trips may be towards the airport located in the outskirts.

# Feature Engineering

- **Created Columns-**

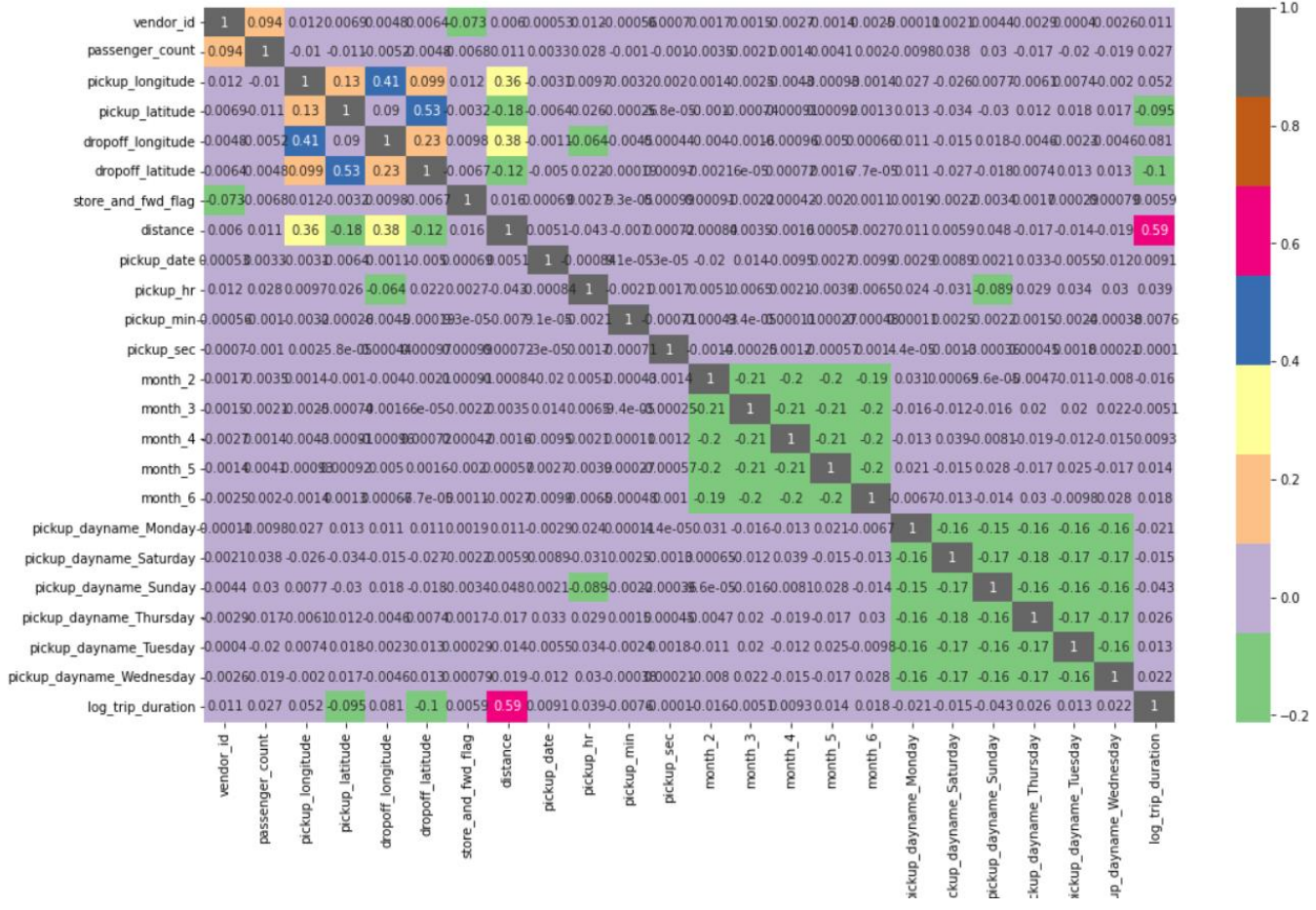
1. **pickup_date, month, hour, min , second from pickup_datetime.**
2. **Distance by using Great Circle method from geopy module.**
3. **Weekday_or_end, log of trip duration.**

**Transformed/ Scaled Columns-**

1. **Hot encoding of month and day columns**
2. **Store_and_fwd_flag column values were mapped to 0 and 1.**
3. **Minmax scaling of input columns.**

- **Dropped Columns-**
- **ID, pickup_datetime ,dropoff_datetime,**

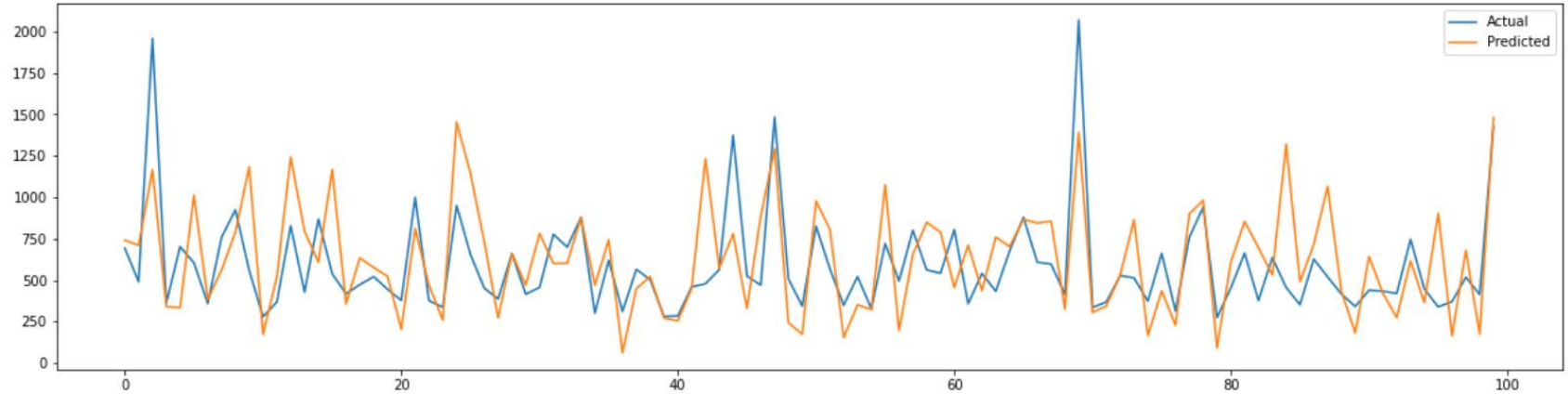
# Correlation Heatmap



# Models used

## 1. LINEAR REGRESSION:

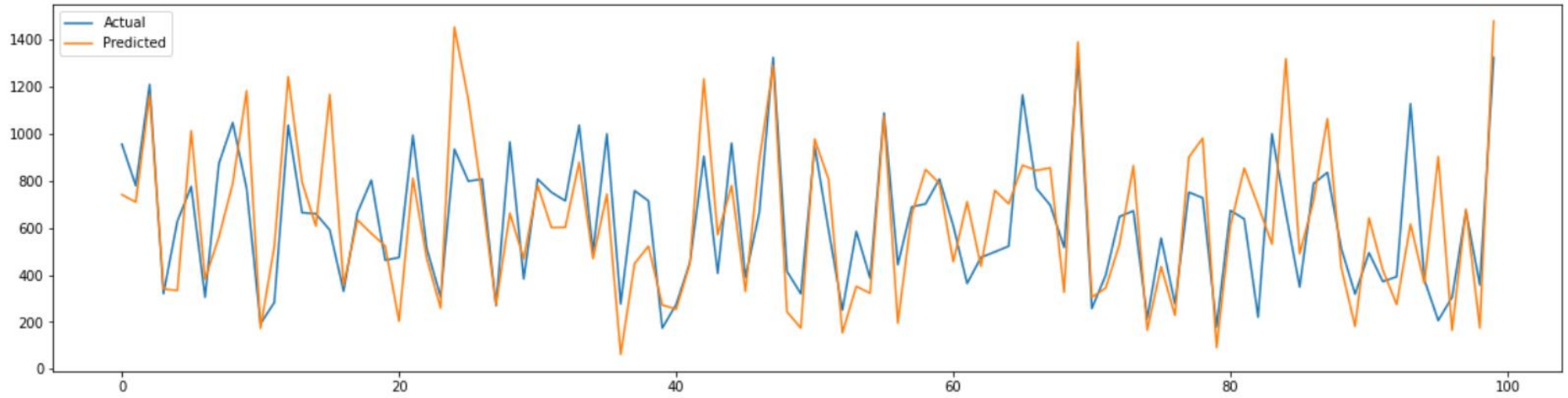
In linear regression, we assume a linear relationship between input and output variables. It finds the value of coefficients of the linear equation such that it minimizes the Sum of Squared Errors (Sum of squares of actual value- predicted value). A plot showing the actual and predicted values for first 100 samples is given below.

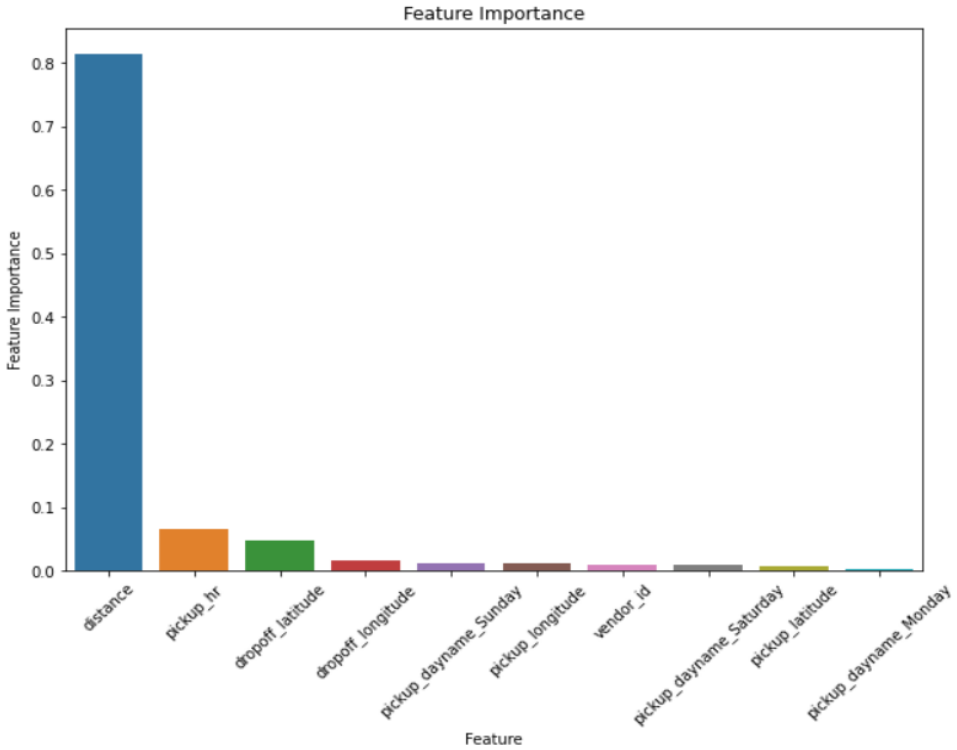


# Models used

## 2. Decision Tree

**It is an algorithm that divides the data into tree like structures based on some features. Generally, this model is highly prone to overfitting. In order to find the best hyper-parameters and to get best performance we performed Cross- Validation using GridSeachCV.**



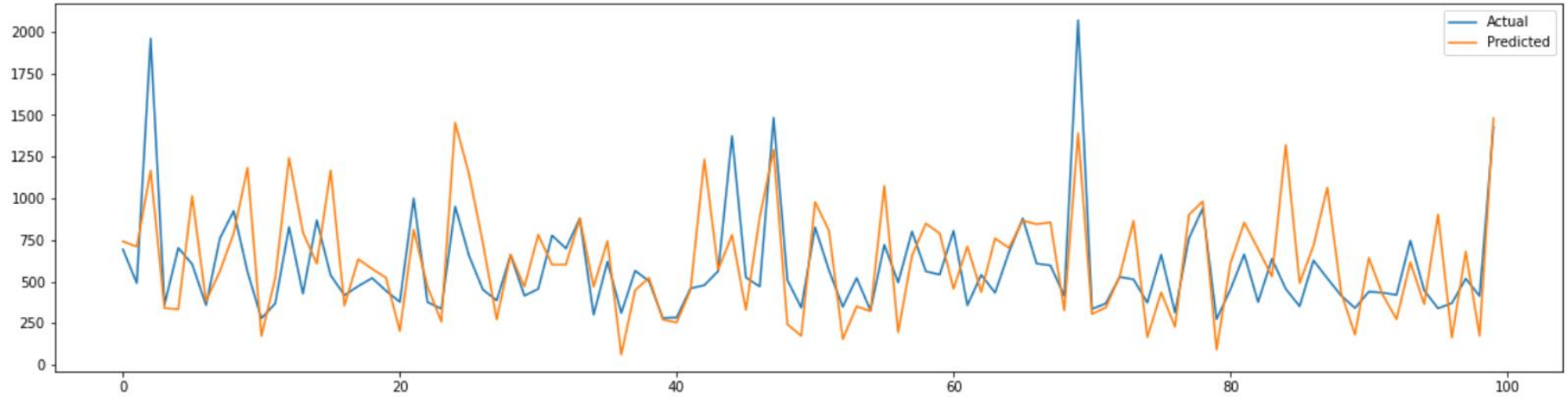



- **We plotted the above graph to find out the most important features according to decision tree. It can be observed that travel distance is the feature with maximum importance, followed by pickup_hr.**

# Models used

## 3. Lasso Regression

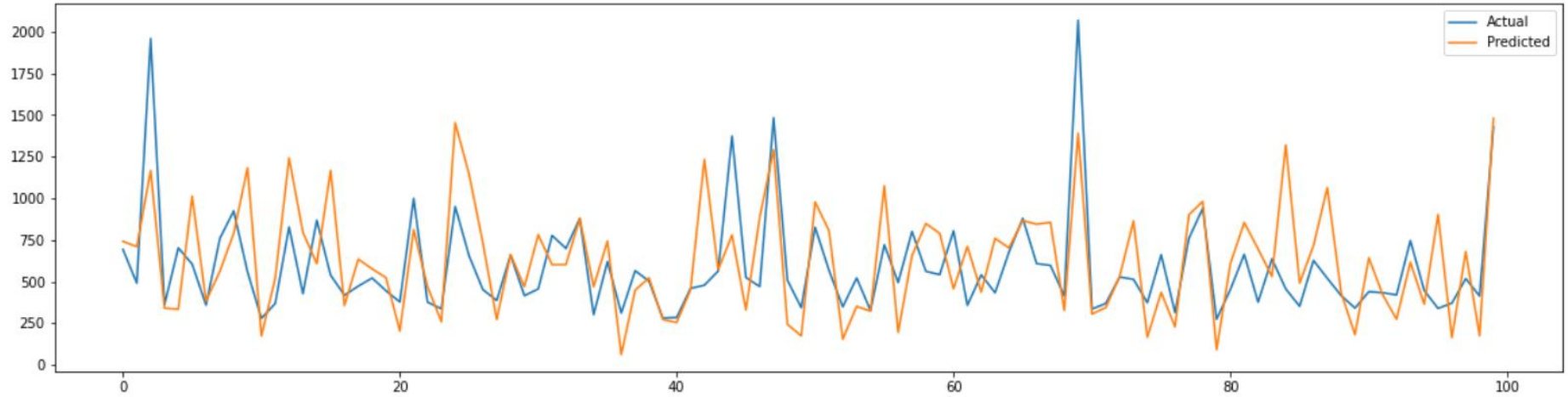
**It is an improvement over the Linear Regression which is used to reduce the overfitting tendency. It is done by shrinking the coefficients. By using cross-validation, we found the best penalizing factor which can minimize the error.**



# Models used

## 4. Ridge Regression

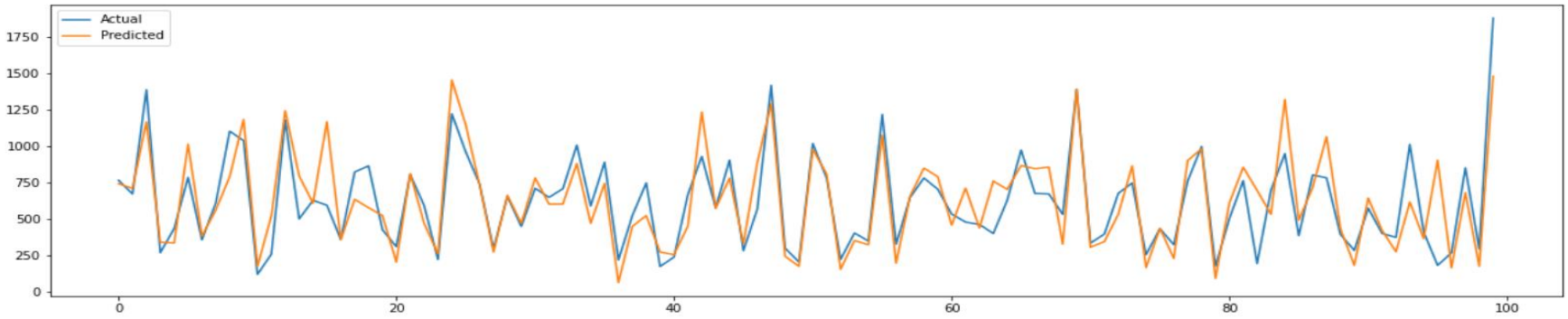
**It is another regularization technique used over the linear regression. It also helps to prevent the overfitting tendency by shrinking the coefficients. To find out the best penalizing factor, Cross validation was performed.**

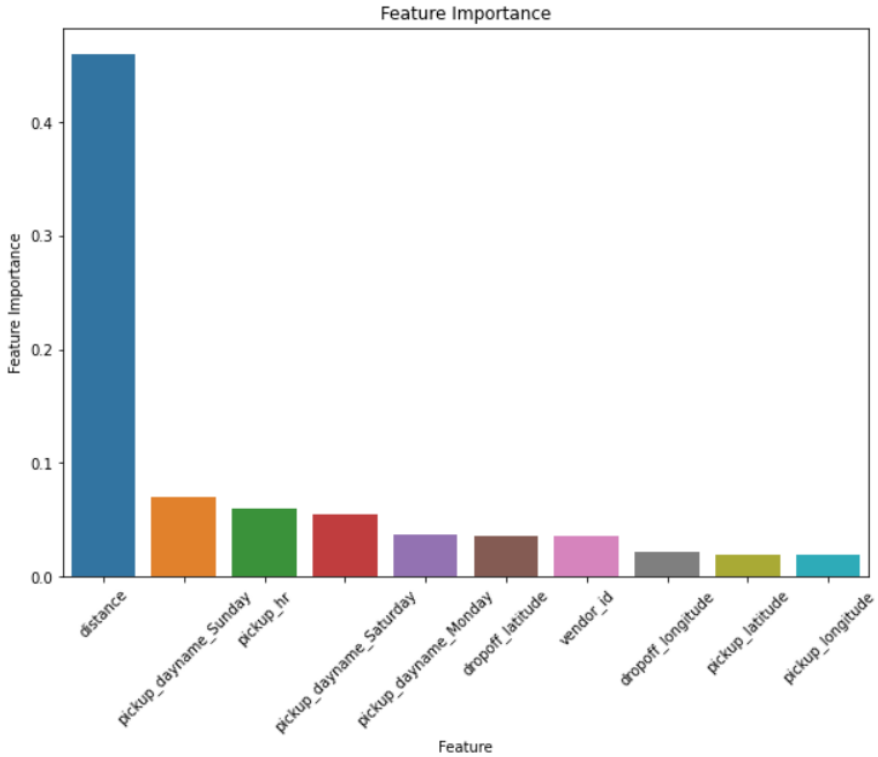


# Models used

## 5. XGBoost

It stands for extreme gradient boosting. It is a decision tree based ensemble ML model. In this method, initially some weight is given to all independent variables and a decision tree is grown. Based on the predictions of the decision tree, residuals were calculated and weight of variables predicted wrongly by the decision tree is increased. It performs well on larger dataset and provides a low bias- low variance model.






- **This graph represents the most important features according to XGBoost. It is evident that trip distance is the most important feature and pickup_dayname_Sunday is the one with second most importance.**

# Evaluation matrices used

1. Mean Absolute Error (MAE)

It is the mean of the absolute value of (actual value- predicted value).

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

2. Mean Squared Error (MSE)

It is the average of square of the difference between predicted and actual values.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

3. Root Mean Squared Error (RMSE) : Square root of MSE.

4. R2 Score

It indicates how much percentage deviation in the dependent variable can be expressed by the deviation in the input variables.

$$R2\,score = 1 - \frac{(SSE\,of\,Regressor)}{(SSE\,of\,Mean)}$$

5. Adjusted R2 Score

It is an improvement over R2 Score and imposes a penalty term.

$$R^2_{adjusted} = 1 - \frac{(1 - R^2)\,(N - 1)}{N - p - 1}$$

where

$R^2$ = sample R-square

p = Number of predictors

N = Total sample size.

# Comparing different models

| Algorithm | MAE | MSE | RMSE | R2 Score | Adjusted R2 Score |
|---|---|---|---|---|---|
| Linear Regression | 0.164 | 0.049 | 0.223 | 0.443 | 0.443 |
| Decision Tree | 0.127 | 0.031 | 0.176 | 0.653 | 0.653 |
| Lasso Regression | 0.164 | 0.049 | 0.223 | 0.443 | 0.443 |
| Ridge Regression | 0.164 | 0.049 | 0.223 | 0.443 | 0.443 |
| XGBoost | 0.103 | 0.023 | 0.154 | 0.734 | 0.734 |

# Conclusion

1. For this regression problem, linear models such as Linear Regression, Lasso Regression, Ridge Regression Performed poorly in comparison with other models.
2. The performance of Linear Regression has not improved when it is used with Ridge Regularization, and Lasso Regularization.
3. Decision Tree performed fairly better when compared with linear models.
4. XGBoost has the best values of MSE, MAE,R2Score for both training and testing data.
5. According to decision tree, Distance and pickup_hr are the most important features whereas Distance and Pickup_dayname_Sunday are the most important features as per XGBoost.
6. We can effectively use XGBoost model to predict the taxi trip time .

# Thank you