

# Capstone Project

## NYC TAXI TRIP TIME PREDICTION.

Individual Project

Uthaman A

# Content

- **Problem Statement**
- **Introduction**
- **Data Summary**
- **Data Analysis**
- **Analysis Details**
- **Challenges**
- **Conclusions**
- **Q&A**

# Problem Statement

**We have the data which was originally published by the NYC Taxi and limousine Commision (TLC), for the year 2016. This dataset consists of various trip related feature and our aim is to predict the trip duration based on these features.**

# Introduction

**In today's world it has become a race to gain more and more number of customers.**

**To gain more number of customers companies/vendors usually try to provide their customers with more comfort to attract them.**

**So here we will be predicting the time of trip duration our customers will take and which algorithm is best suited for that time prediction.**

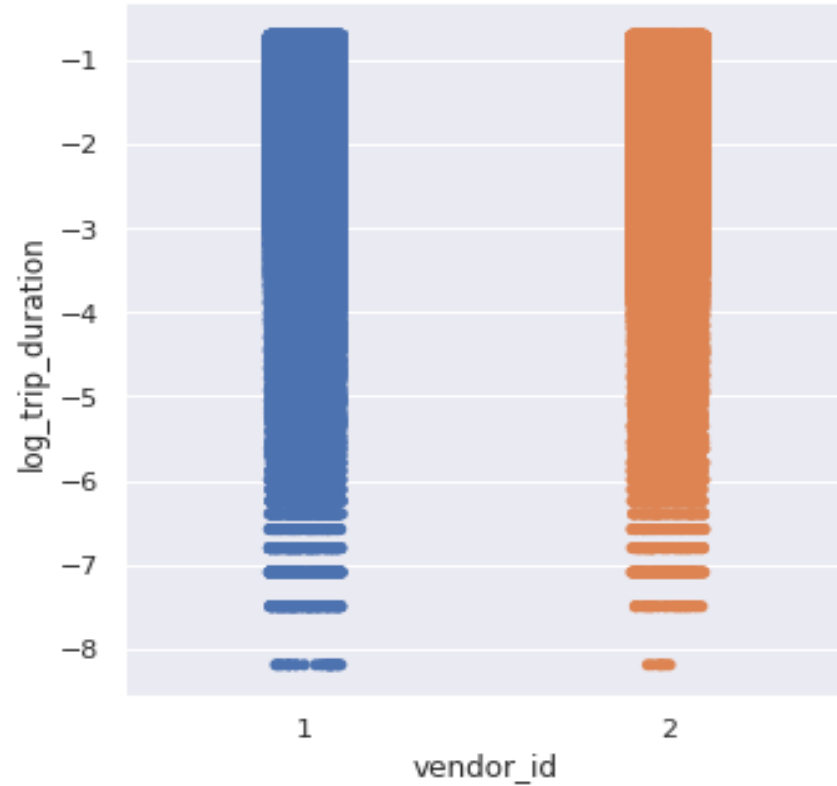
# Data Summary

- **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

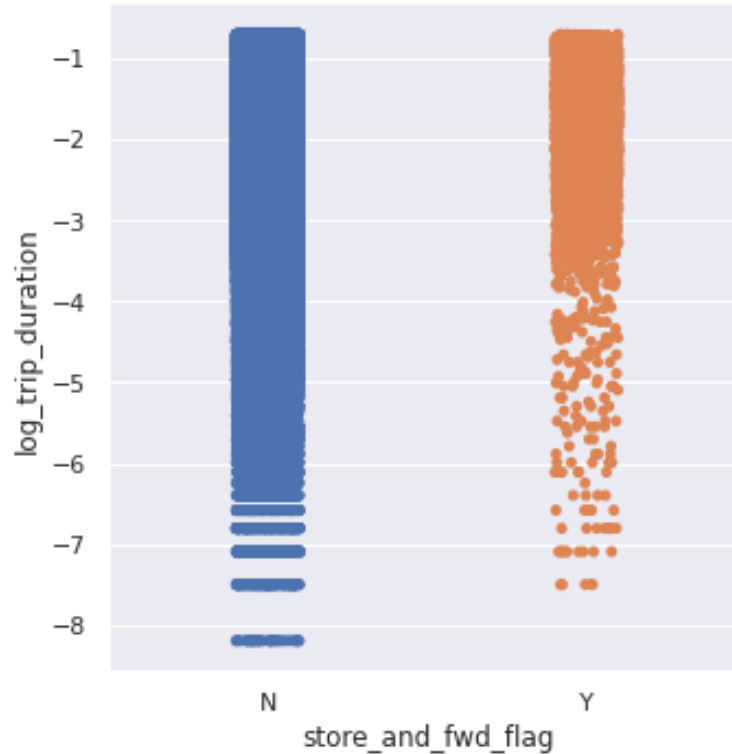
# Basic Exploration

- The dataset contains 1458644 rows and 11 features(columns)
- Two categorical features 'store\_and\_fwd\_flag' and 'vendor\_id'
- Outlier present in all numerical features
- Data cleaning steps required for datetime features
- No null values present

# Vendor ID Analysis

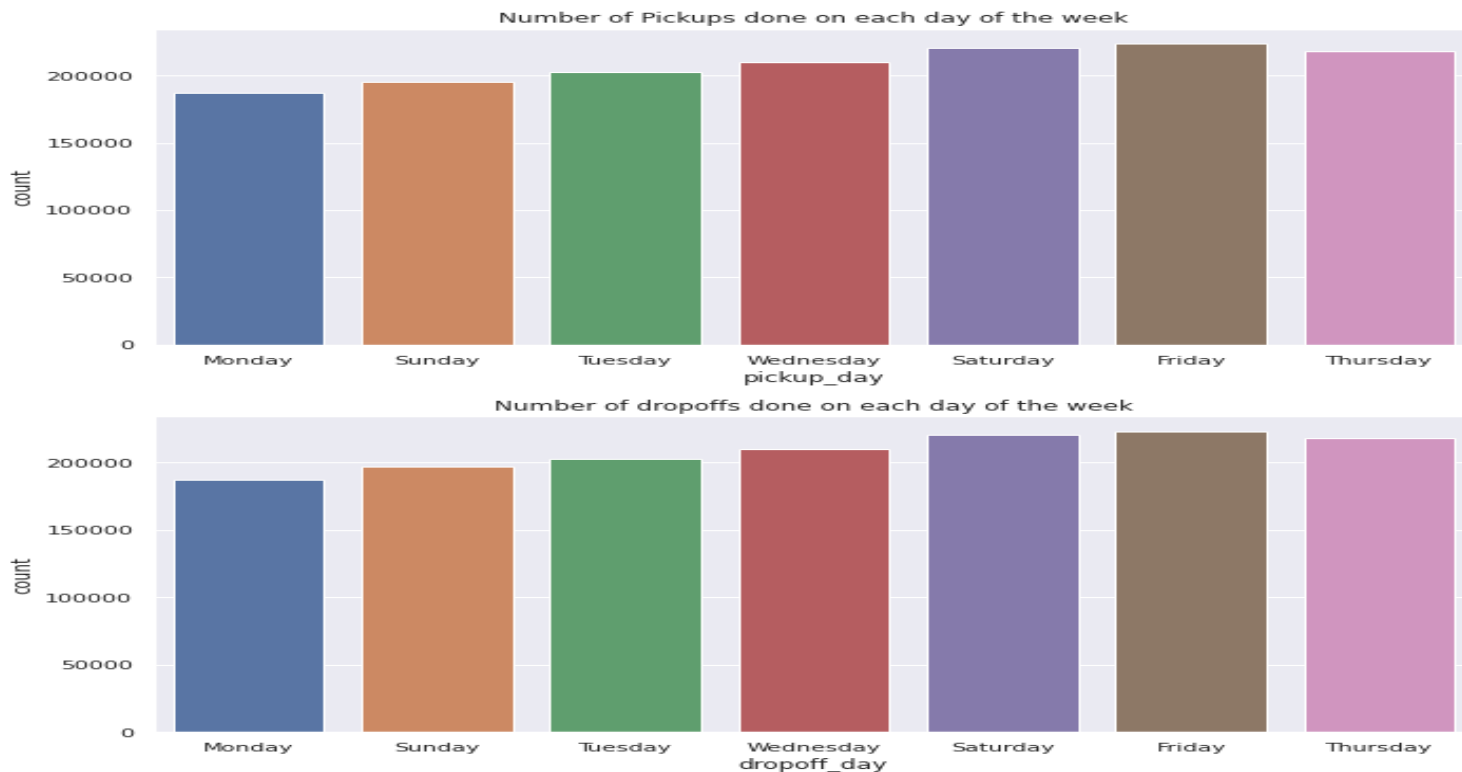


# Store and forward flag

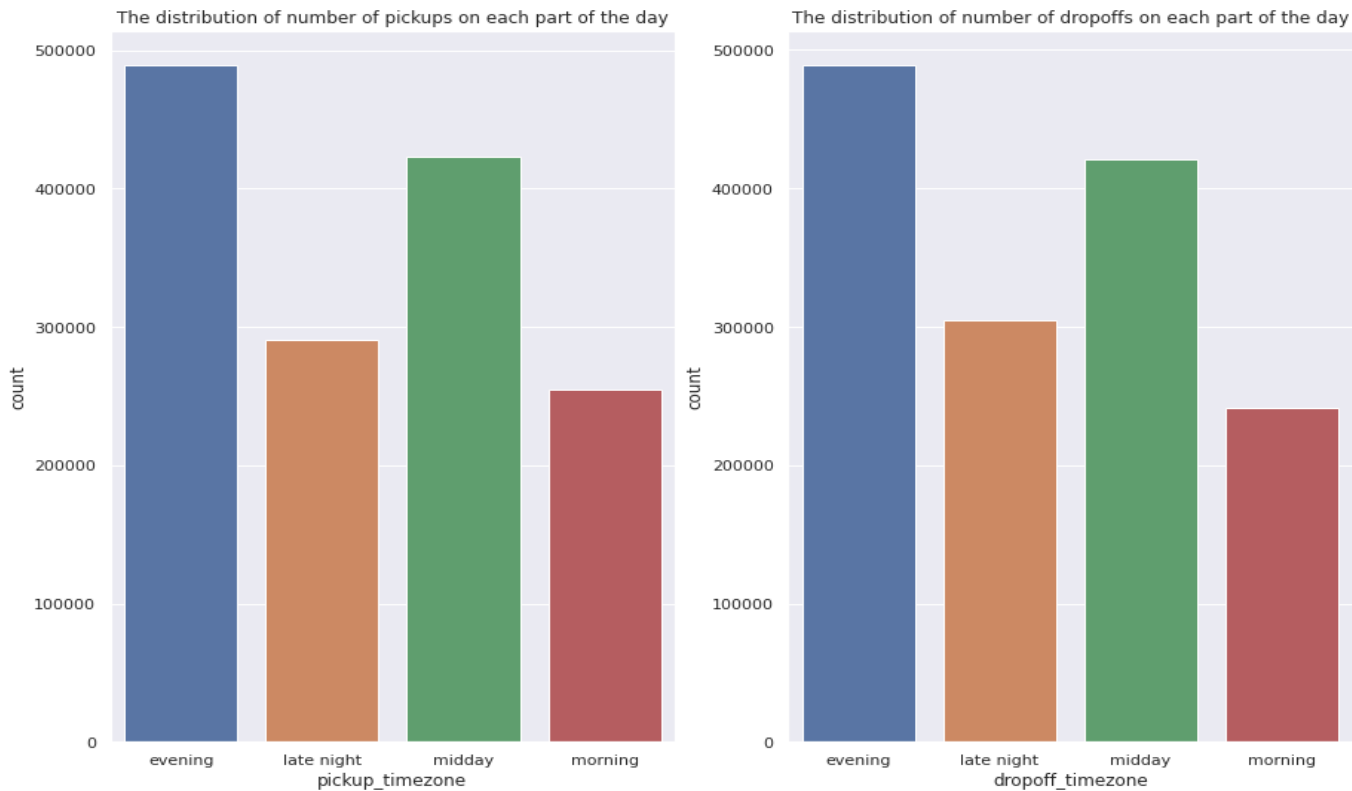




# Days of the week



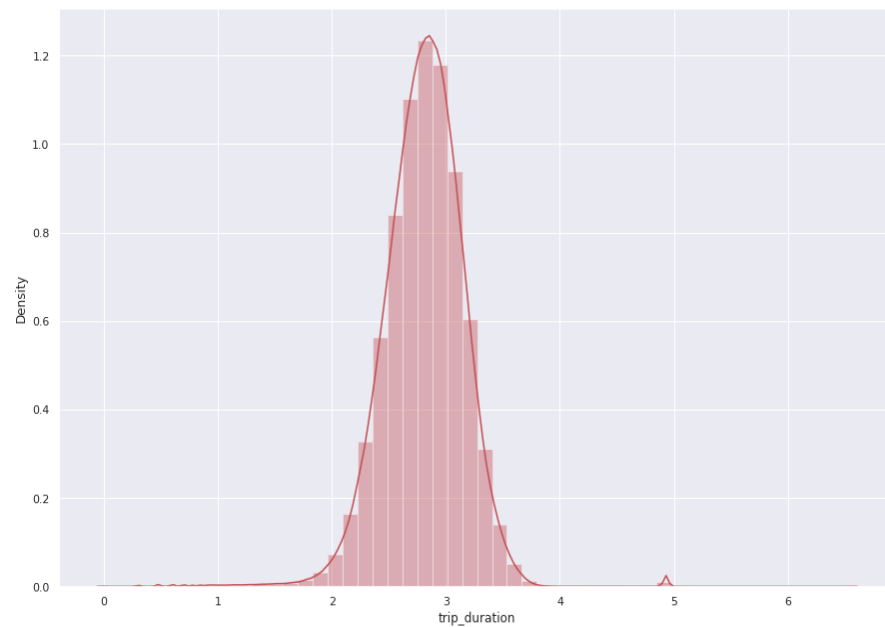
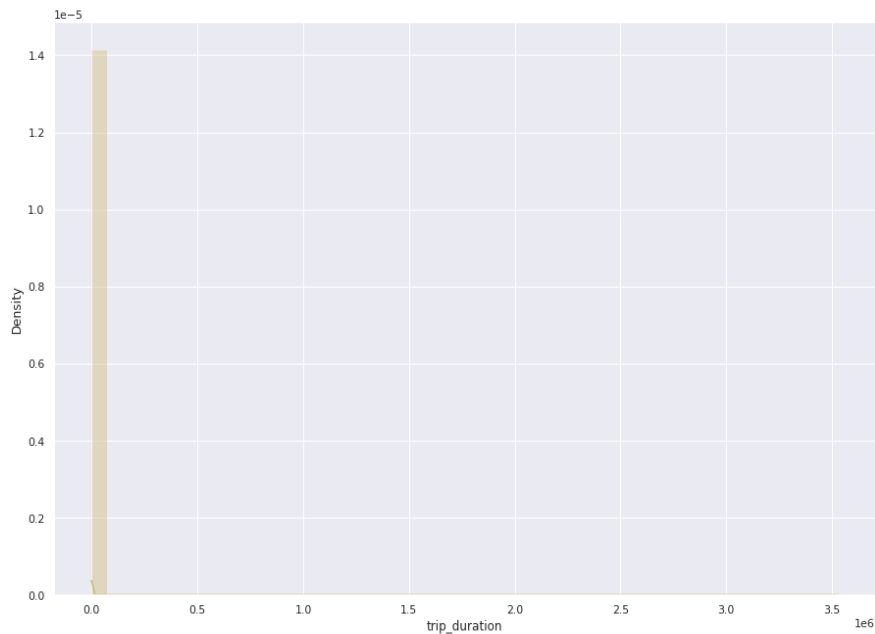
# Day Segmentation



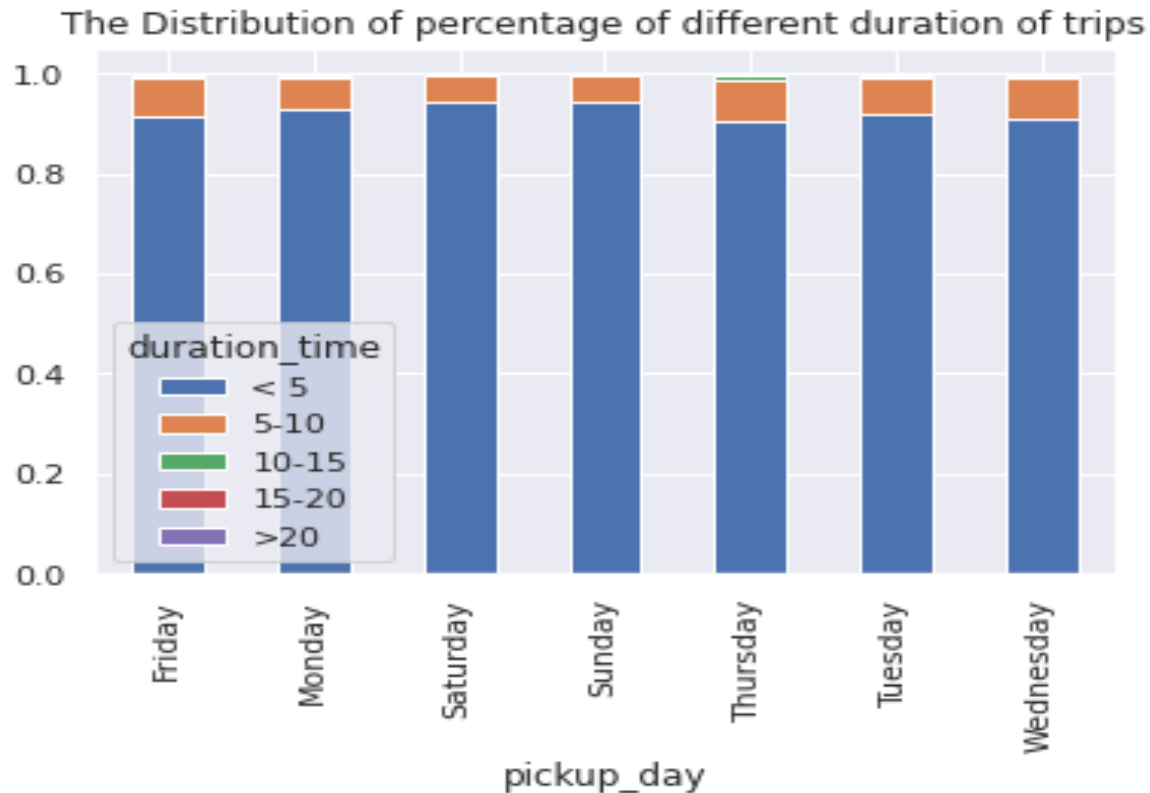
# Plotting longitude and latitude



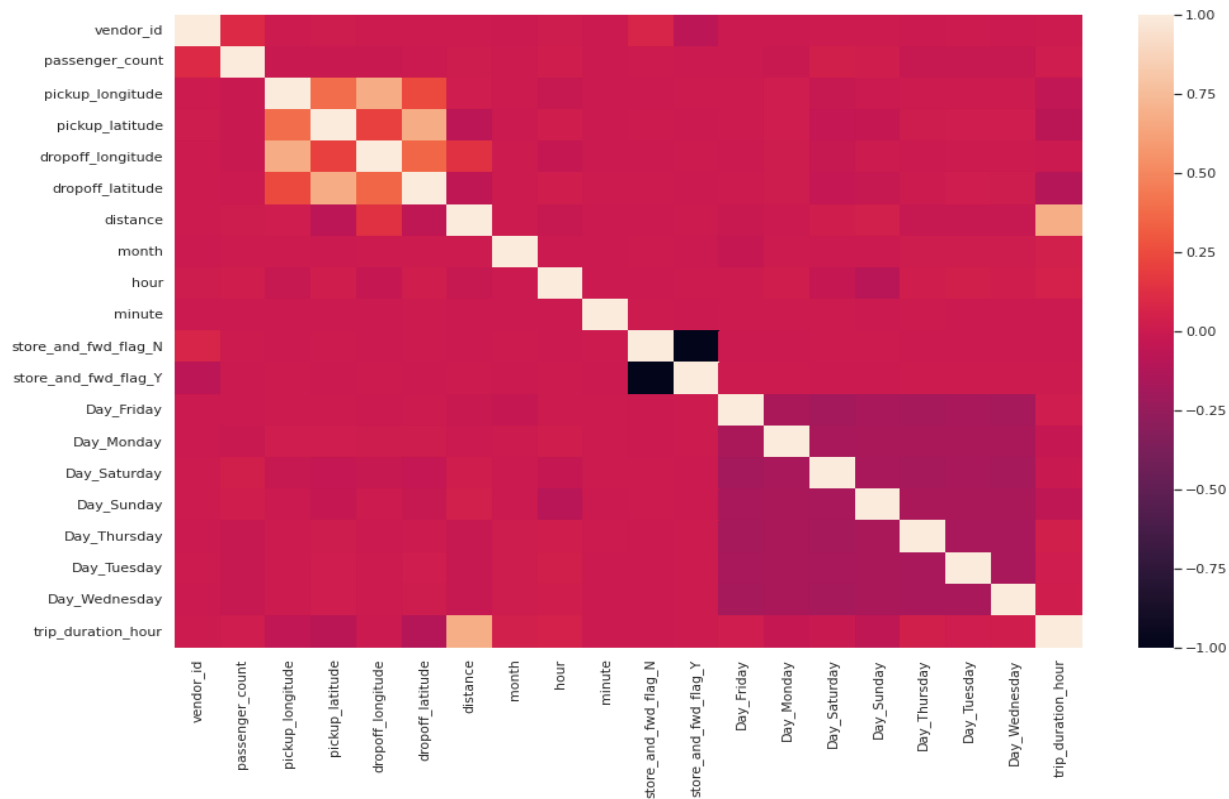
# Trip Duration (dependent variable) Data Analysis



# Trip Duration/Day of the week



# Analysis Details



# Lasso Regression

## Train set metrics

Train MSE : 0.00552025815855849  
Train RMSE : 0.0742984398124112  
Train R2 : 0.49716557121777893  
Train Adjusted R2 : 0.4970461181776832

## Test set metrics

Test MSE : 0.005547254775022398  
Test RMSE : 0.07447989510614525  
Test R2 : 0.4983934824098727  
Test Adjusted R2 : 0.4979164792149672

# Ridge Regression

## Train set metrics

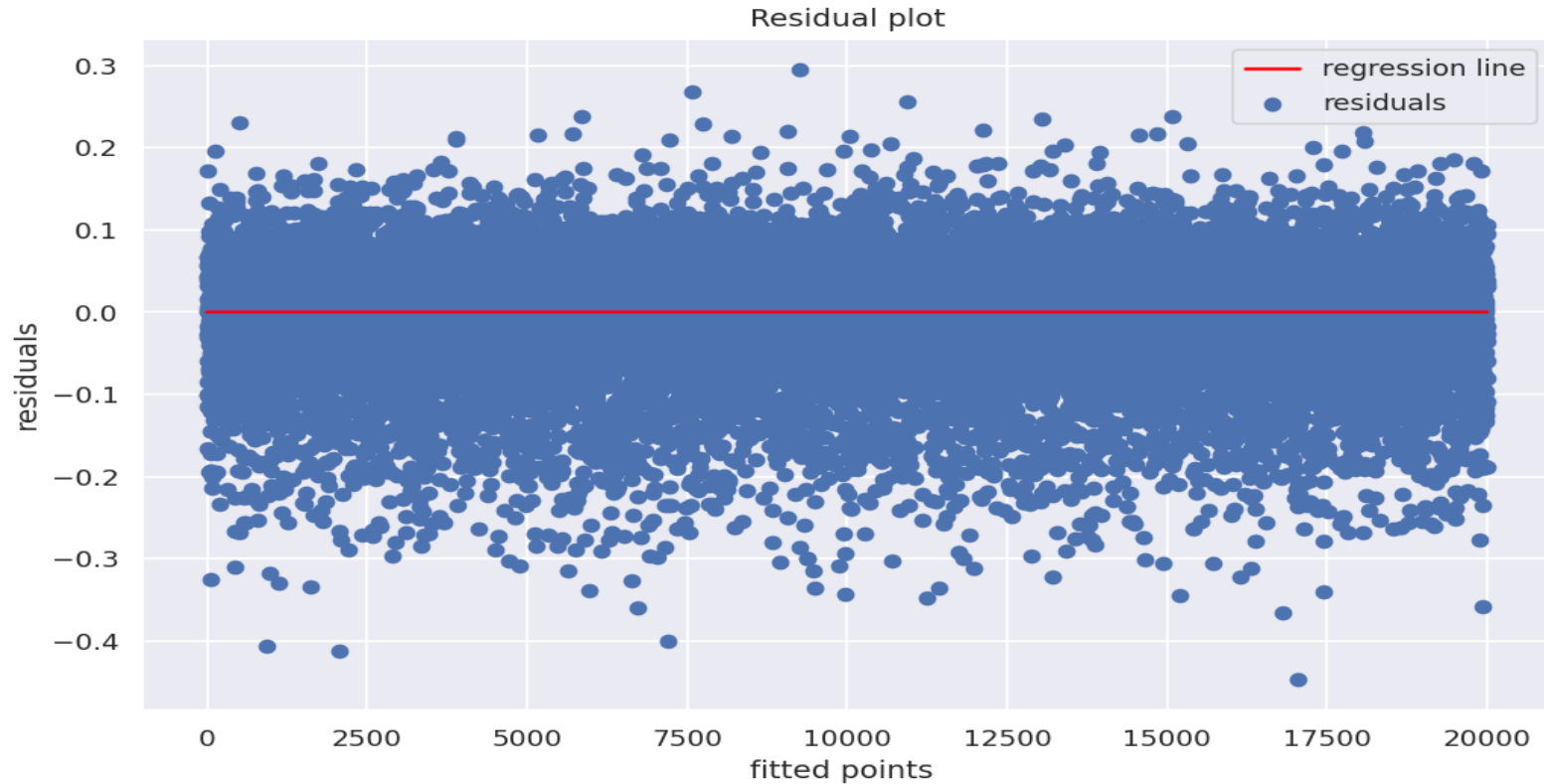
Train MSE : 0.005520256908751504  
Train RMSE : 0.07429843140168912  
Train R2 : 0.4971656850613645  
Train Adjusted R2 : 0.49704623204831333

## Test set metrics

Test MSE : 0.005547259274275785  
Test RMSE : 0.07447992531062168  
Test R2 : 0.49839307556812584  
Test Adjusted R2 : 0.4979160719863338



# Homoscedasticity check



# Decision Tree

## Train set metrics

Train MSE : 0.003903026546923643

Train RMSE : 0.0624742070531803

Train R2 : 0.6444774740830809

Train Adjusted R2 : 0.644393016368747

## Test set metrics

Test MSE : 0.004354489573938102

Test RMSE : 0.0659885563862258

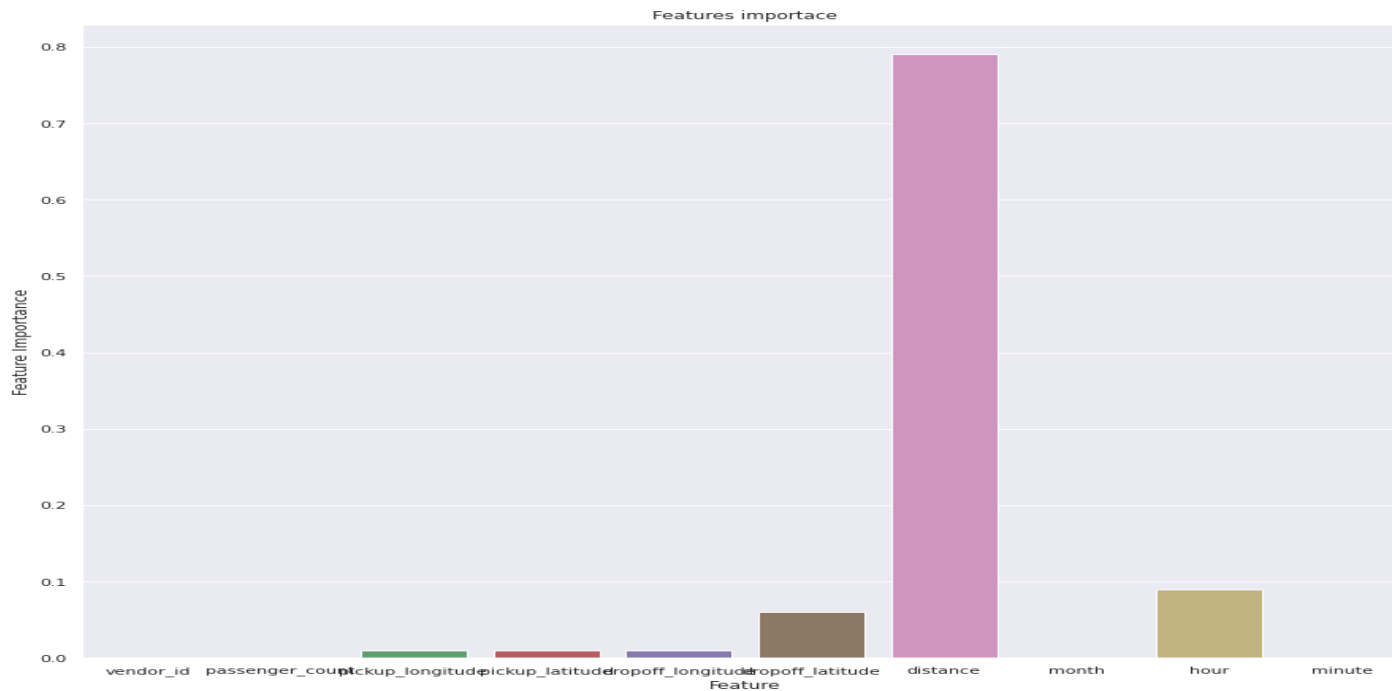
Test R2 : 0.6062484166221138

Test Adjusted R2 : 0.6058739781794622

## Parameters:

- Criterion=mse
- Max\_depth=10
- Min\_sample\_leaf=20
- Min\_sample\_split=10

# Decision Tree Feature Importance



# Gradient Boosting

## Train set metrics

Train MSE : 0.002323670139456396

Train RMSE : 0.048204461821043036

Train R2 : 0.7883393649914061

Train Adjusted R2 : 0.7882890830200988

## Test set metrics

Test MSE : 0.003268332446055223

Test RMSE : 0.05716933134168374

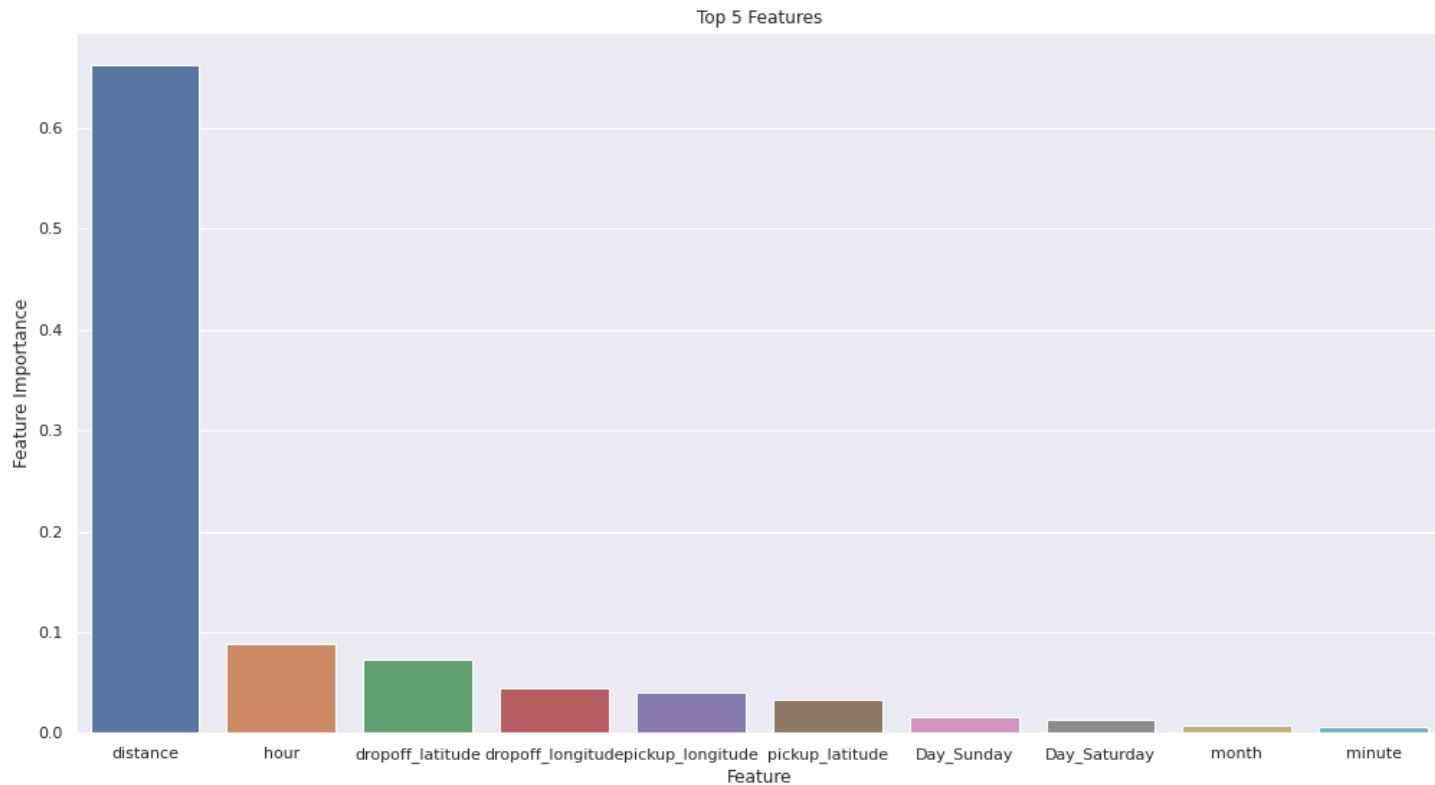
Test R2 : 0.704463392600178

Test Adjusted R2 : 0.7041823517823304

## Parameters:

- **alpha=0.9**
- **Max\_depth=10**
- **Min\_sample\_leaf=50**
- **Min\_sample\_split=80**
- **N\_estimators=120**

# GBoost feature importance



# XGBOOST

## Train set metrics

Train MSE : 0.001990850702399407

Train RMSE : 0.044618950036945146

Train R2 : 0.8186555325896028

Train Adjusted R2 : 0.8186124525085725

## Test set metrics

Test MSE : 0.0032639168965066688

Test RMSE : 0.05713070012267195

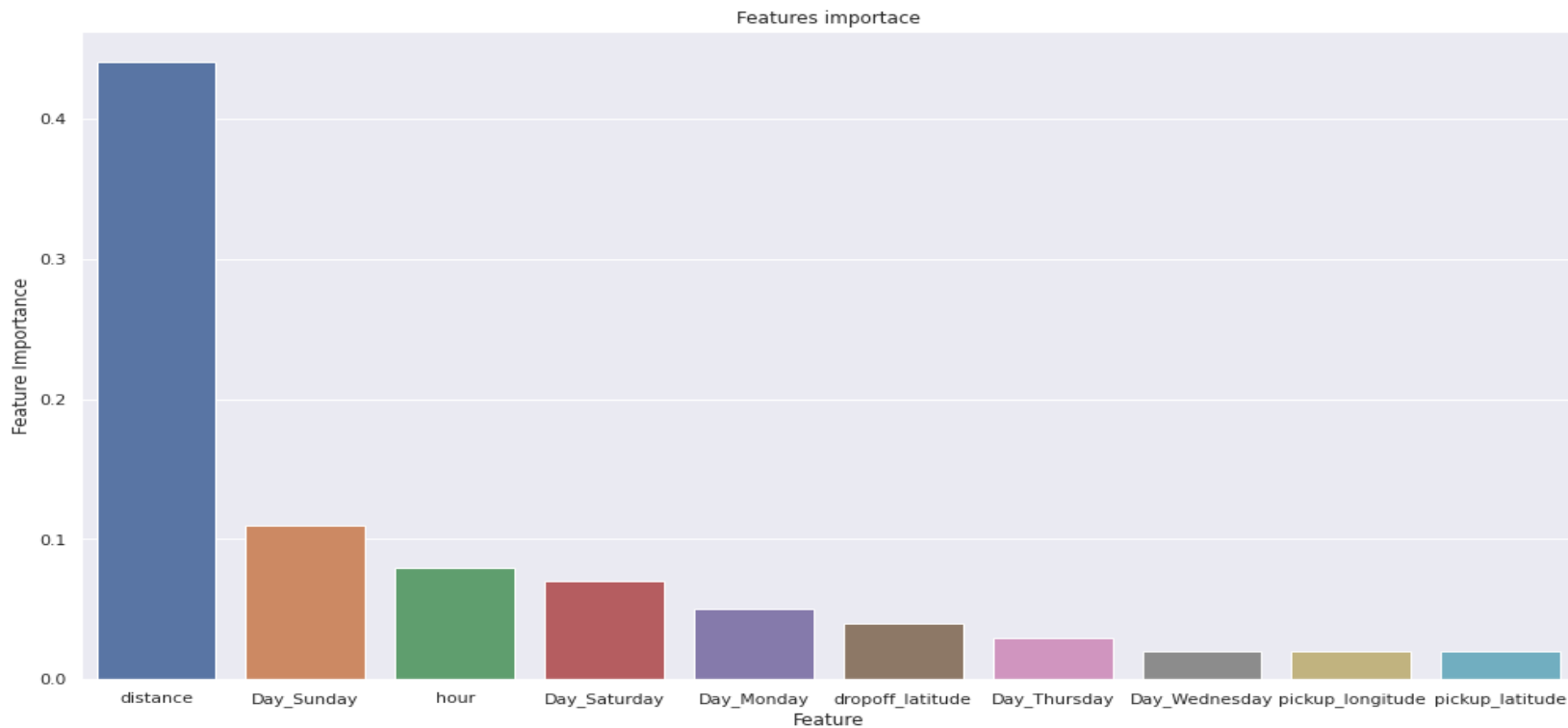
Test R2 : 0.7048626654877816

Test Adjusted R2 : 0.7045820043588662

## Parameters:

- **gamma=0**
- **Learning\_rate=0.1**
- **Max\_depth=9**
- **Min\_sample\_leaf=50**
- **Min\_sample\_split=40**
- **N\_estimators=120**

# XGBoost feature importance



# Final metric conclusion

SL NO	MODEL_NAME	Test MSE	Test RMSE	Test R^2	Test Adjusted R^2
1	Linear Regression	0.005539358995881834	0.07442687012015105	0.48551298500777995	0.48502373309162117
2	Lasso Regression	0.005448974432213879	0.07381716895285187	0.5030970331831028	0.502624502834278
3	Ridge Regression	0.005449008499105121	0.07381739970430495	0.5030939265545915	0.5026213932515153
4	DecisionTree Regressor	0.004203945325941736	0.06483783868962426	0.6166337480963826	0.6162691855945723
5	XGBRegressor	0.0031306995630522444	0.0559526546559879	0.7145051935101532	0.7142337019524301
6	GradientBoosting	0.00311712671404449	0.05583123421566543	0.7157429289820318	0.7154726144500327

SL NO	MODEL_NAME	Train MSE	Train RMSE	Train R^2	Train Adjusted R^2
1	Linear Regression	0.005467021181864388	0.07393930742077848	0.5042456435975543	0.5041278724951332
2	Lasso Regression	0.005494826110976565	0.07412709431089665	0.49994898506301866	0.4998301932490177
3	Ridge Regression	0.005494824127596807	0.07412708093265785	0.4999491655584595	0.49983037378733686
4	DecisionTree Regressor	0.003908733073695245	0.06251986143374956	0.6442897552818683	0.6442052529731706
5	XGBRegressor	0.001996779863964856	0.044685342831457114	0.8182850963041854	0.8182419282225373
6	GradientBoosting	0.002278863599313375	0.04773744441539969	0.7926143552635426	0.7925650888563159



# Challenges

- **Handling Large Dataset.**
- **Feature Engineering.**
- **Computation Time.**
- **Optimising the Model.**

# Conclusion

- In this project, we tried to predict the trip duration of a taxi in NYC.
- We are mostly concerned with the information of pick up latitude and longitude, to get the distance of the trip.
- Gradient Boosting will be the best model to predict the trip duration for a particular taxi.

**Thank You!**

# Q&A