Capstone Project

NYC Taxi Trip Time Prediction

by

Mayank Sawant

Data Science Practitioner

AlmaBetter, Bengaluru

INTRODUCTION

A typical taxi company faces a common problem of efficiently assigning the cabs to passengers so that the service is smooth and hassle free. One of main issue is determining the duration of the current trip so it can predict when the cab will be free for the next trip.

Predicting Taxi Trip Time Using Machine Learning Regression

DATASET

Shape - (1458644, 11)

Our dataset is feature rich containing, pickup and drop-off datetime, passenger count, pickup and drop-off location co-ordinates and target feature as trip duration.

pickup_datetime
timestamp of pickup

id
unique id

dropoff_datetime
timestamp of drop-off

trip_duration
duration of trip in seconds

store_and_fwd_flag
store and forward record or not

passenger_count
no of passenger riding

pickup_latitude
latitude of pickup location

pickup_longitude
longitude of pickup location

dropoff_latitude
latitude of drop-off location

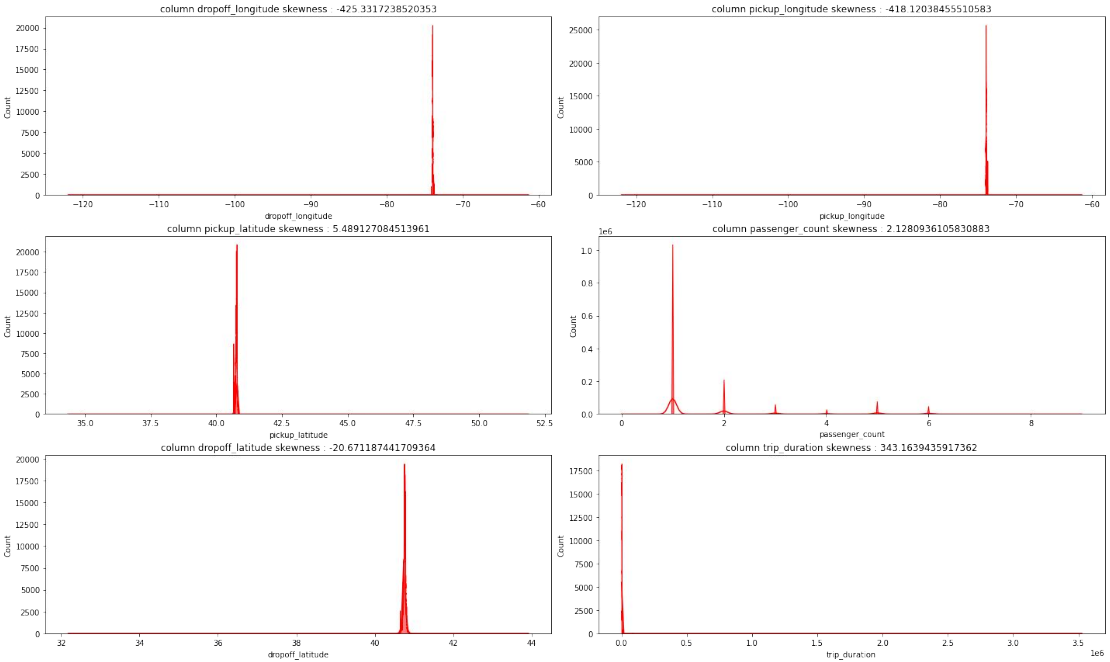
dropoff_longitude
longitude of drop-off location

vendor_id
unique id of trip provider

DISTRIBUTION OF DATA

Majority Data Distribution Type :

Extremely Skewed

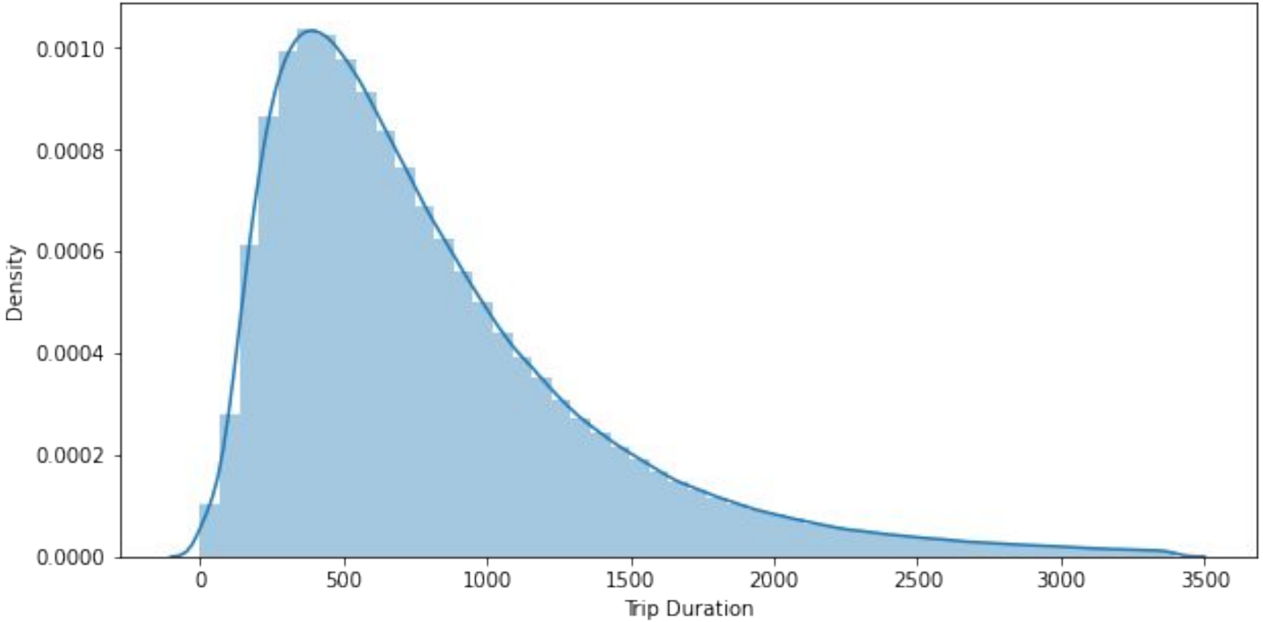


From the distribution plots of the numerical features, we can conclude that most of the data is extremely skewed including trip duration.

TARGET FEATURE DISTRIBUTION – TRIP DURATION

Most of the Trips Duration :

4 to 12 minutes

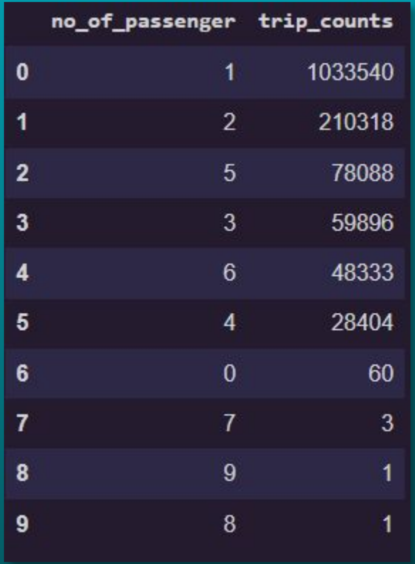


99th percentile of trip duration is completed under 3440 seconds i.e. approx. 1 hour.

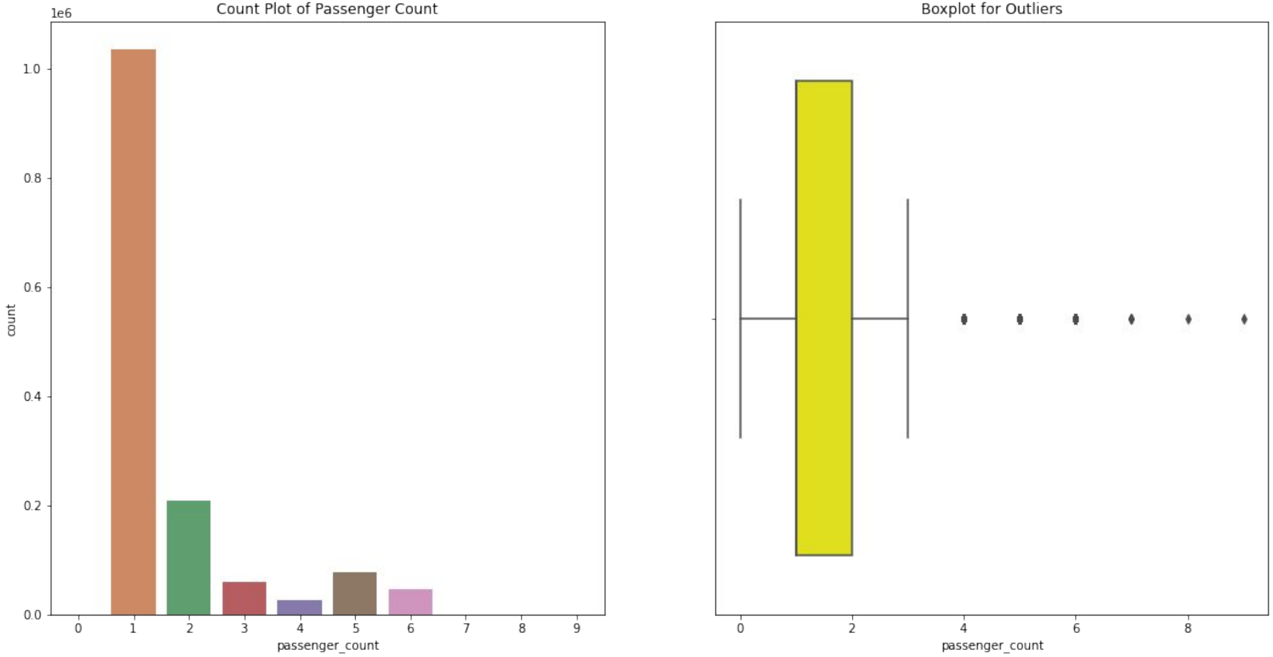
PASSENGER COUNT DISTRIBUTION

Irrelevant Number of Passenger:

0, 7, 8, 9



	no_of_passenger	trip_counts
0	1	1033540
1	2	210318
2	5	78088
3	3	59896
4	6	48333
5	4	28404
6	0	60
7	7	3
8	9	1
9	8	1



Single passenger trips holds the highest amount of Taxi trips. New Yorker's rarely travel in groups.

MONTHLY TRIPS COUNT

Most Busiest Month :

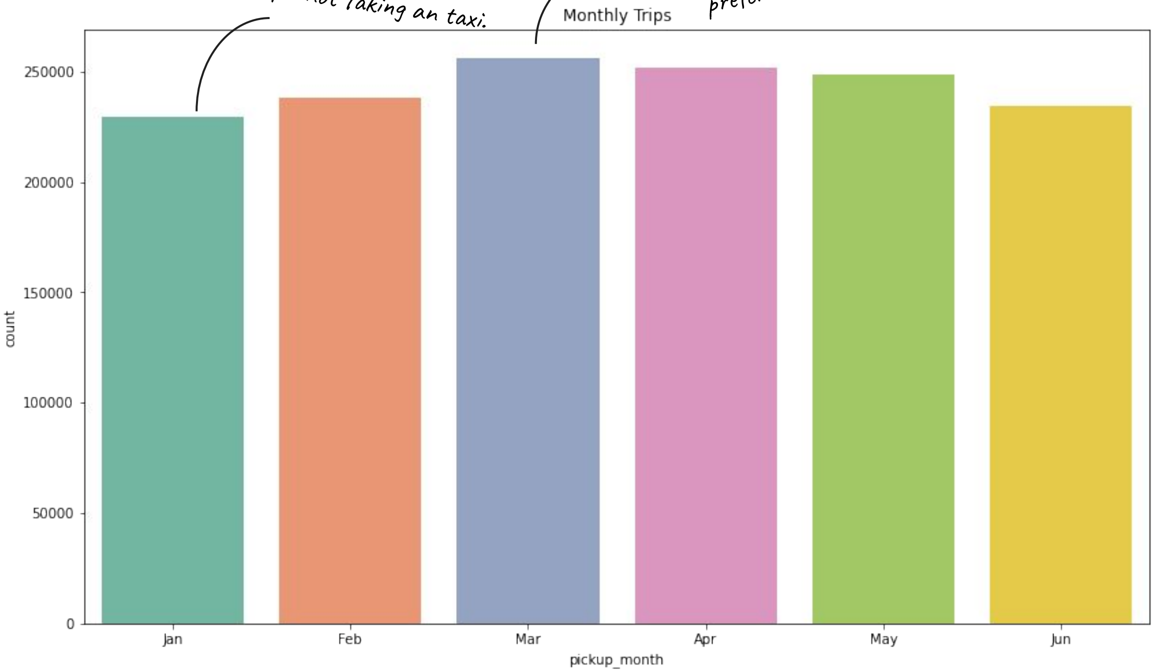
March

Least Busiest Month :

January

New Year Resolution such as Walking to Office could be the silliest reason to determine people not taking an taxi.

NYC's March have irregular weather conditions, this could be one of the reasons that people prefer taking an taxi.



Month March crosses the 25k mark with the most number of trips in first-half of the year.

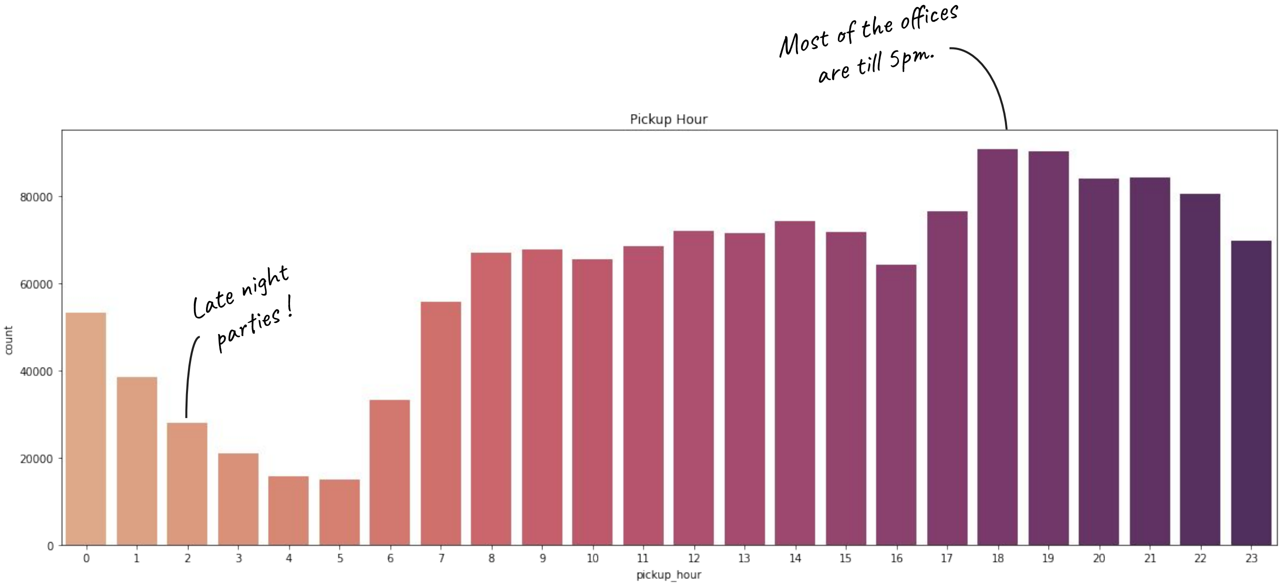
HOURLY PICKUP COUNT

Rush Hours:

6 to 7

No Rush Hours :

3 to 5



People prefer taking an taxi more after daylight.

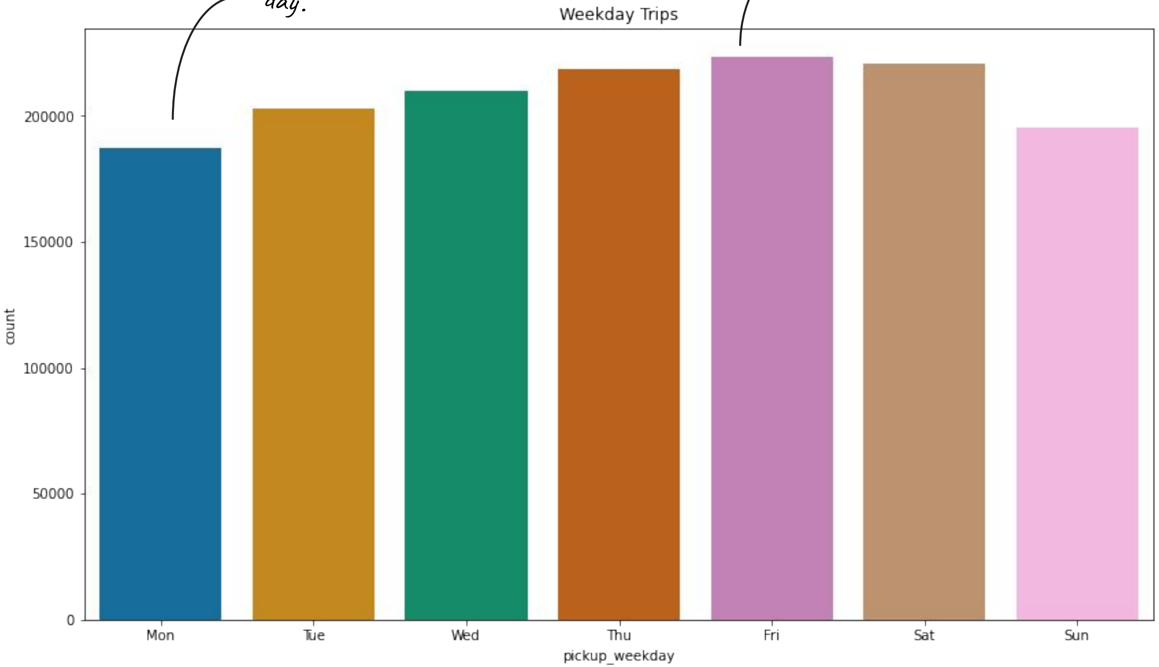
WEEKDAY TRIPS COUNT

Most Busiest Day :

Friday

Least Busiest Day :

Monday



Monday is 'the most popular sick day'

Friday movie nights!

New Yorker's prefer going out on Friday and Saturday.

DAILY TRIP COUNT

Rush Days:

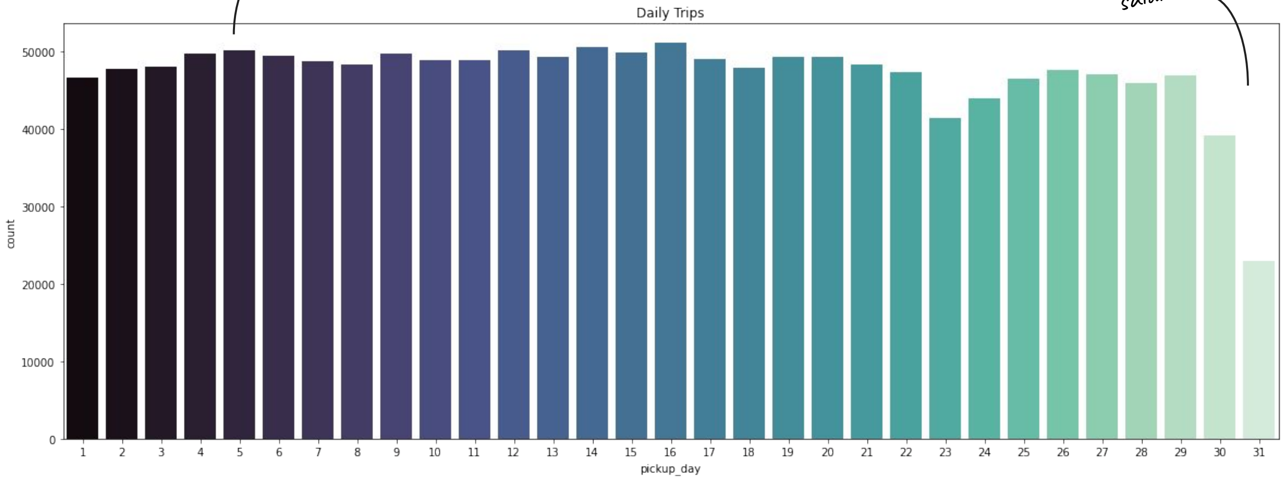
4 to 20

No Rush Days :

23 to 31

My wallet feels heavy today.

Its time to burn calorie instead of salarie !



Seems like New Yorker’s do not prefer to get a Taxi on Month-Ends.

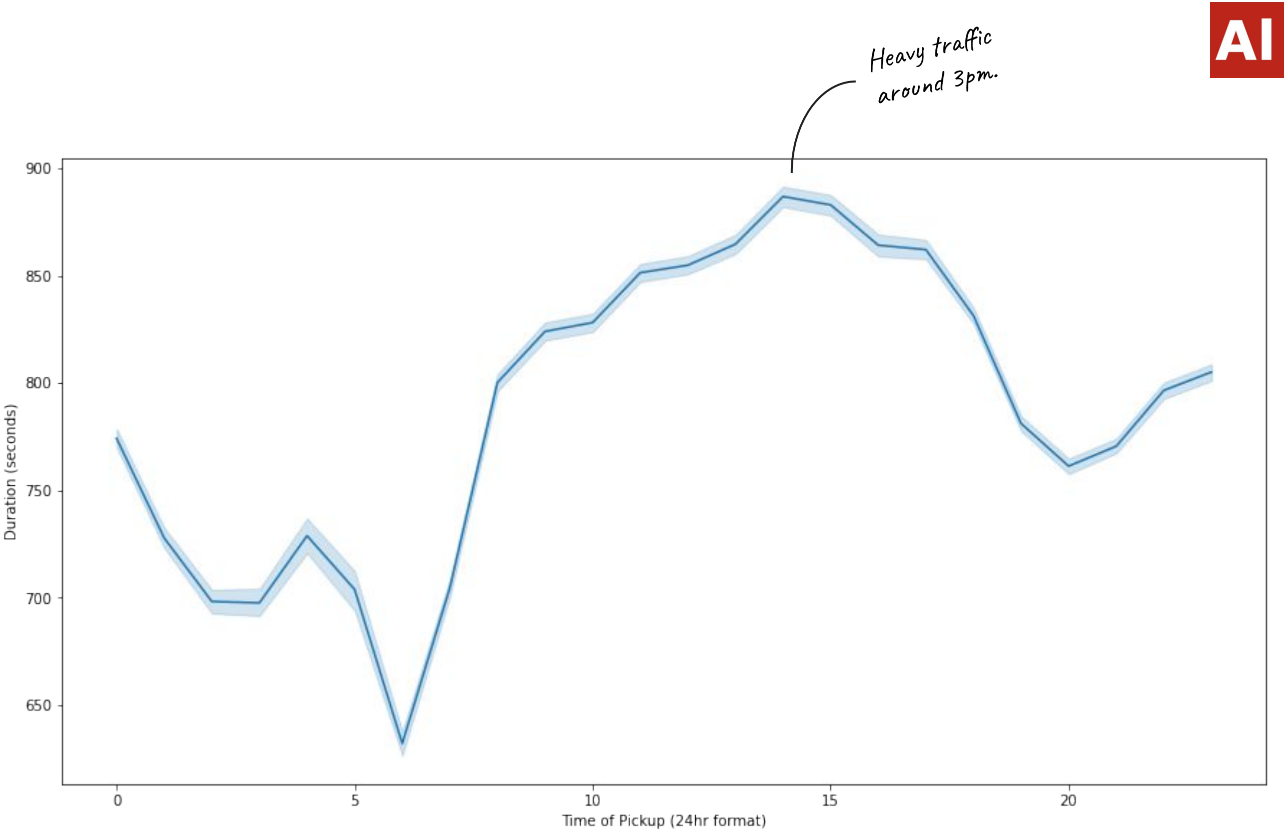
TRIP DURATION PER HOUR

Time When Trip Durations are Less :

5am to 7 am

Time When Trip Durations are High:

2pm to 5pm

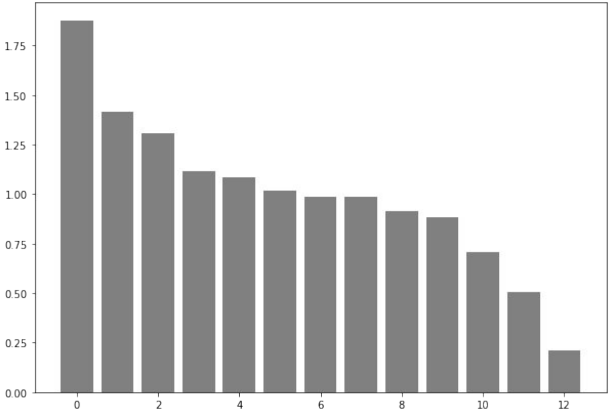


Around 3pm trip duration is high, NYC's heavy traffic could be the reason.

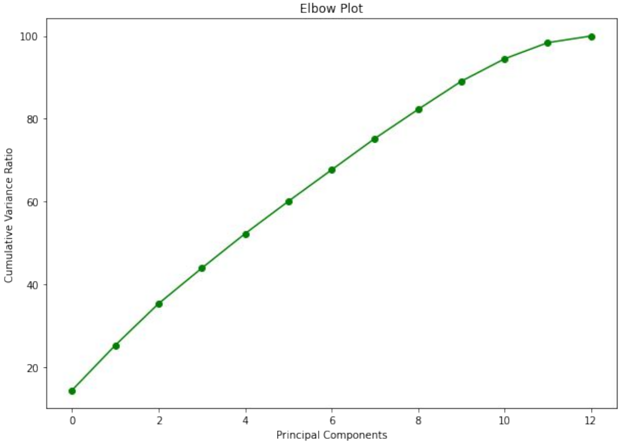
Principal Component Analysis

Why PCA ?

It's a Dimensionality Reduction Technique. It is also a Feature extraction Technique. By PCA we create new features from old (Original) Features but the new features will always be independent of each other. So, its not just Dimensionality Reduction Process, we are even eliminating Correlation between the Variables.

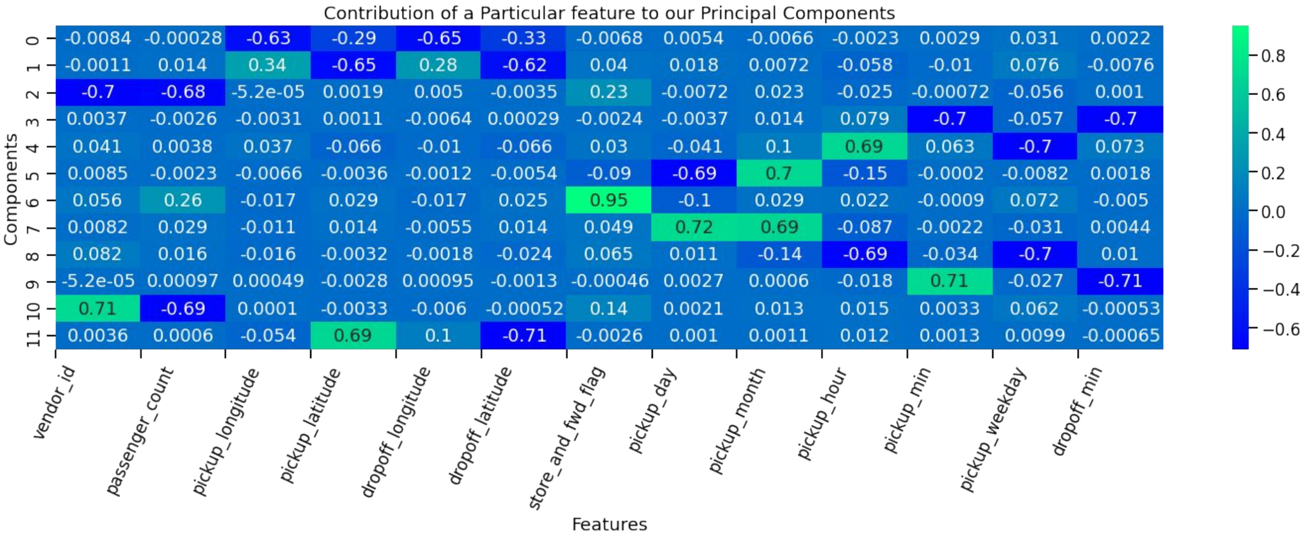


At 12th component our PCA model seems to go Flat without explaining much of a Variance.



By looking at the Elbow plot, 12 is likely to be the required number of components.

PCA FEATURE IMPORTANCE MATRIX



Contribution of a Particular feature to our Principal Components

	vendor_id	passenger_count	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	store_and_fwd_flag	pickup_day	pickup_month	pickup_hour	pickup_min	pickup_weekday	dropoff_min
0	-0.0084	-0.00028	-0.63	-0.29	-0.65	-0.33	-0.0068	0.0054	-0.0066	-0.0023	0.0029	0.031	0.0022
1	-0.0011	0.014	0.34	-0.65	0.28	-0.62	0.04	0.018	0.0072	-0.058	-0.01	0.076	-0.0076
2	-0.7	-0.68	-5.2e-05	0.0019	0.005	-0.0035	0.23	-0.0072	0.023	-0.025	-0.00072	-0.056	0.001
3	0.0037	-0.0026	-0.0031	0.0011	-0.0064	0.00029	-0.0024	-0.0037	0.014	0.079	-0.7	-0.057	-0.7
4	0.041	0.0038	0.037	-0.066	-0.01	-0.066	0.03	-0.041	0.1	0.69	0.063	-0.7	0.073
5	0.0085	-0.0023	-0.0066	-0.0036	-0.0012	-0.0054	-0.09	-0.69	0.7	-0.15	-0.0002	-0.0082	0.0018
6	0.056	0.26	-0.017	0.029	-0.017	0.025	0.95	-0.1	0.029	0.022	-0.0009	0.072	-0.005
7	0.0082	0.029	-0.011	0.014	-0.0055	0.014	0.049	0.72	0.69	-0.087	-0.0022	-0.031	0.0044
8	0.082	0.016	-0.016	-0.0032	-0.0018	-0.024	0.065	0.011	-0.14	-0.69	-0.034	-0.7	0.01
9	-5.2e-05	0.00097	0.00049	-0.0028	0.00095	-0.0013	-0.00046	0.0027	0.0006	-0.018	0.71	-0.027	-0.71
10	0.71	-0.69	0.0001	-0.0033	-0.006	-0.00052	0.14	0.0021	0.013	0.015	0.0033	0.062	-0.00053
11	0.0036	0.0006	-0.054	0.69	0.1	-0.71	-0.0026	0.001	0.0011	0.012	0.0013	0.0099	-0.00065

- Above plot gives us detailed ideology of which feature has contributed more or less to our each Principal Component.
- Principal Components are our new features which consists of Information from every other original Feature we have.
- We reduce the Dimensions using PCA by retaining as much as Information possible.

MACHINE LEARNING REGRESSION MODELS IMPLEMENTED

Random Forest

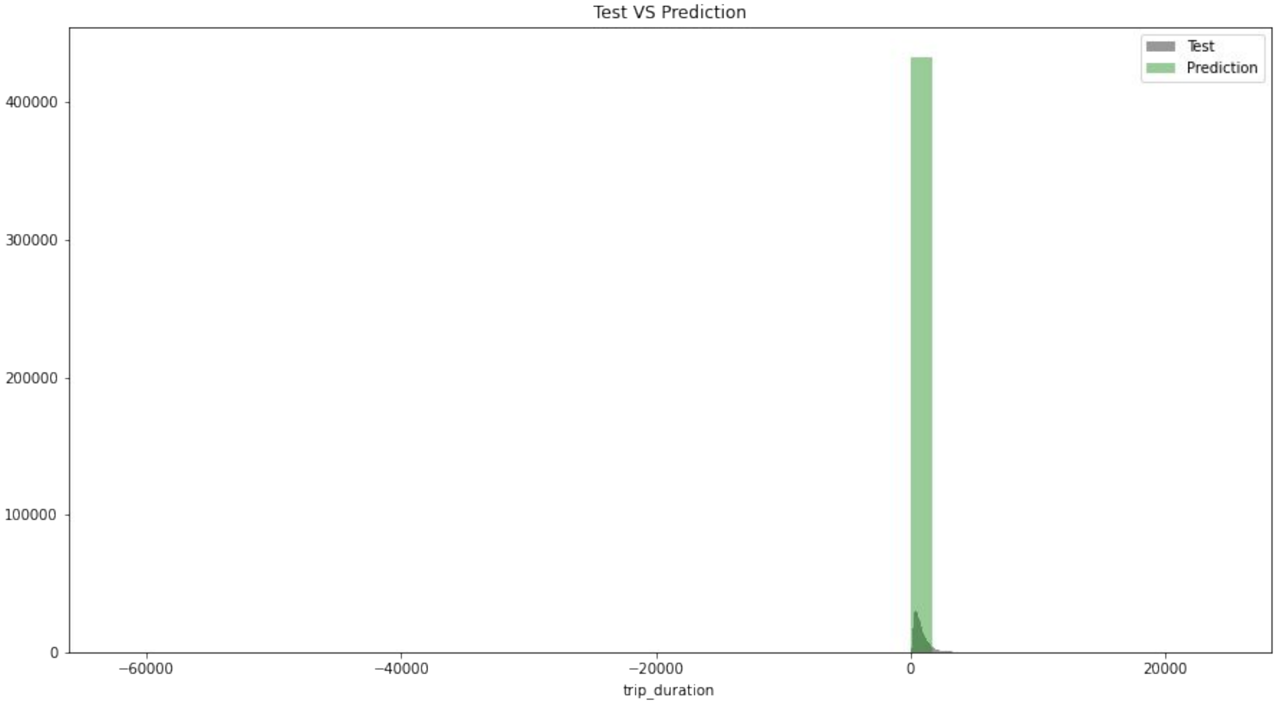
XGBoost

Linear Regression

Extra Trees Regressor

Decision Trees

LINEAR REGRESSION



We can clearly see Linear Regression is not well suited for this data.

DECISION TREES

MSE :

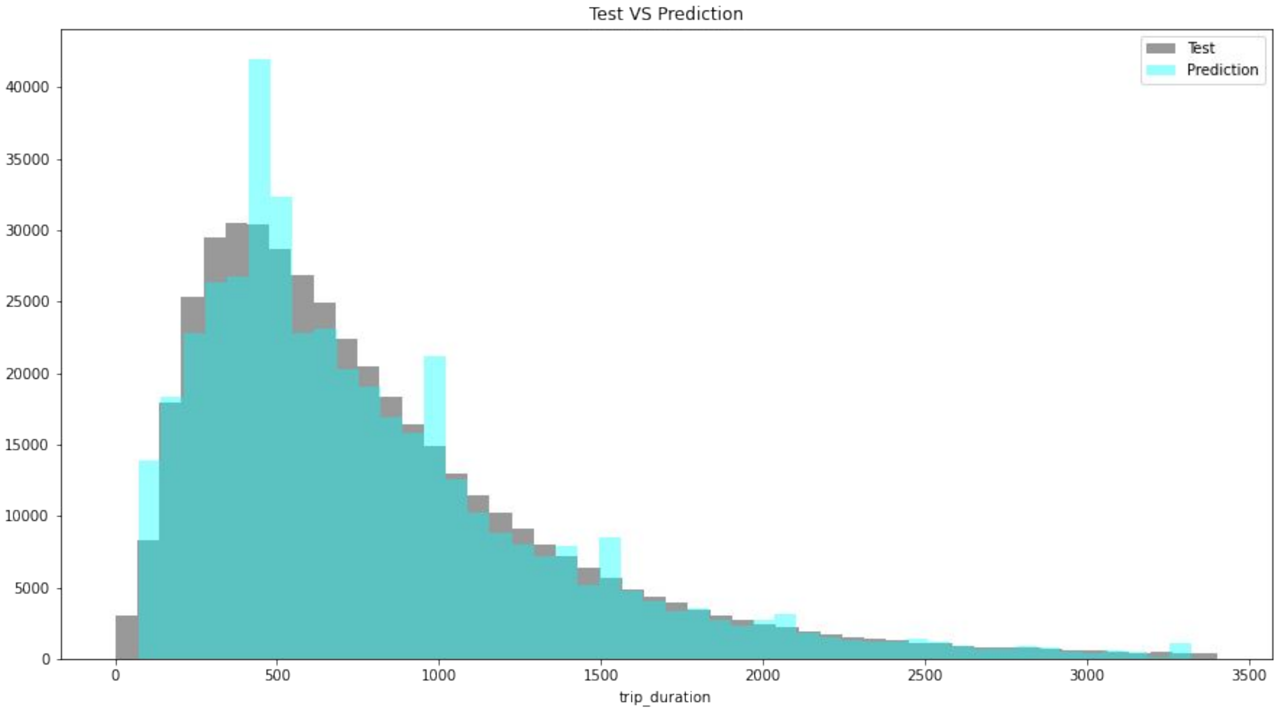
1352.143

R2 SCORE:

0.99587

ADJUSTED R2 SCORE:

0.99588



Decision Tree has performed well compared to Linear Regression.

RANDOM FOREST

MSE :

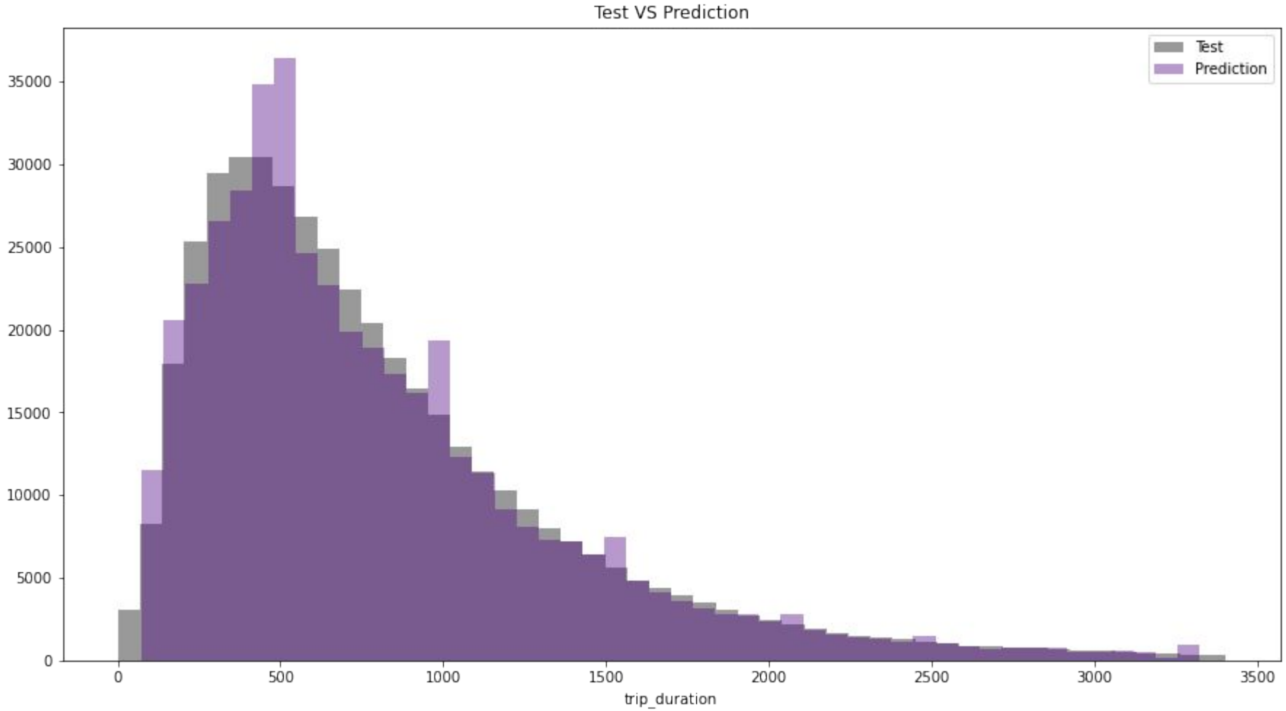
1186.074

R2 SCORE:

0.99637

ADJUSTED R2 SCORE:

0.99639



Random Forest has performed slightly better than Decision Trees.

Extra Trees Regressor

MSE :

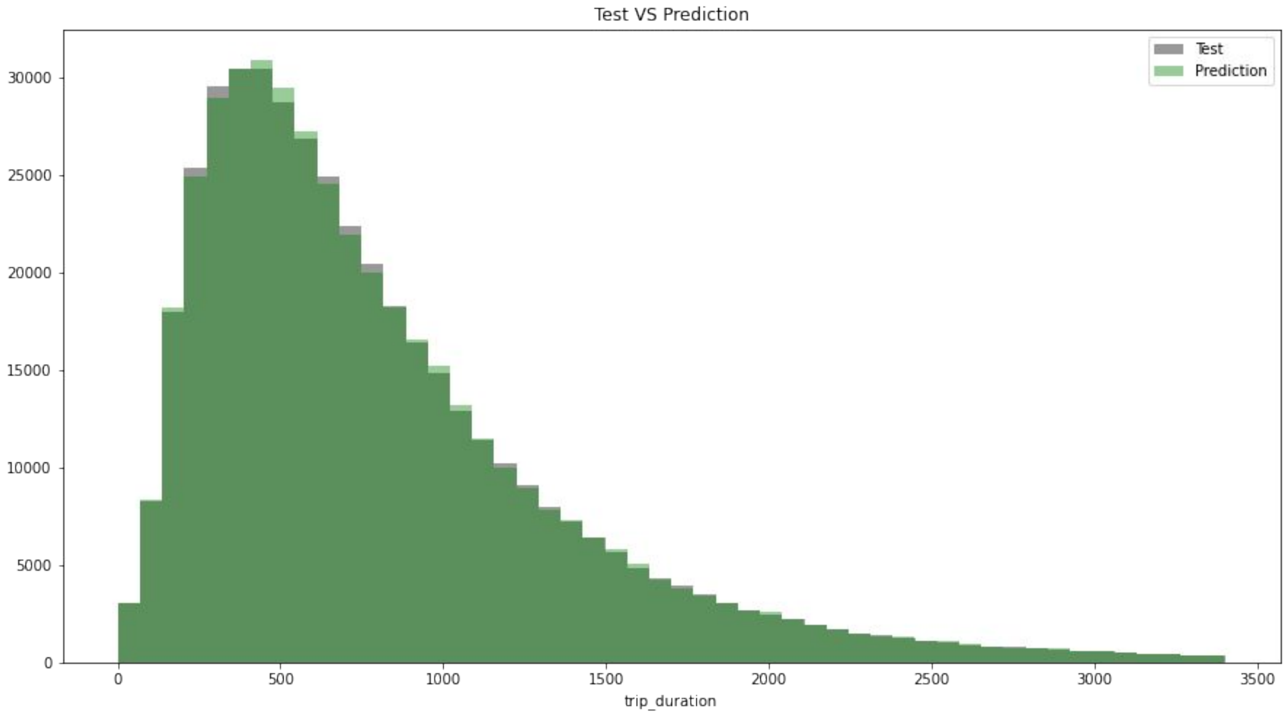
1007.620

R2 SCORE:

0.99693

ADJUSTED R2 SCORE:

0.99693



Extra Tress Regressor appears to be the optimal model.

XGBoost

MSE :

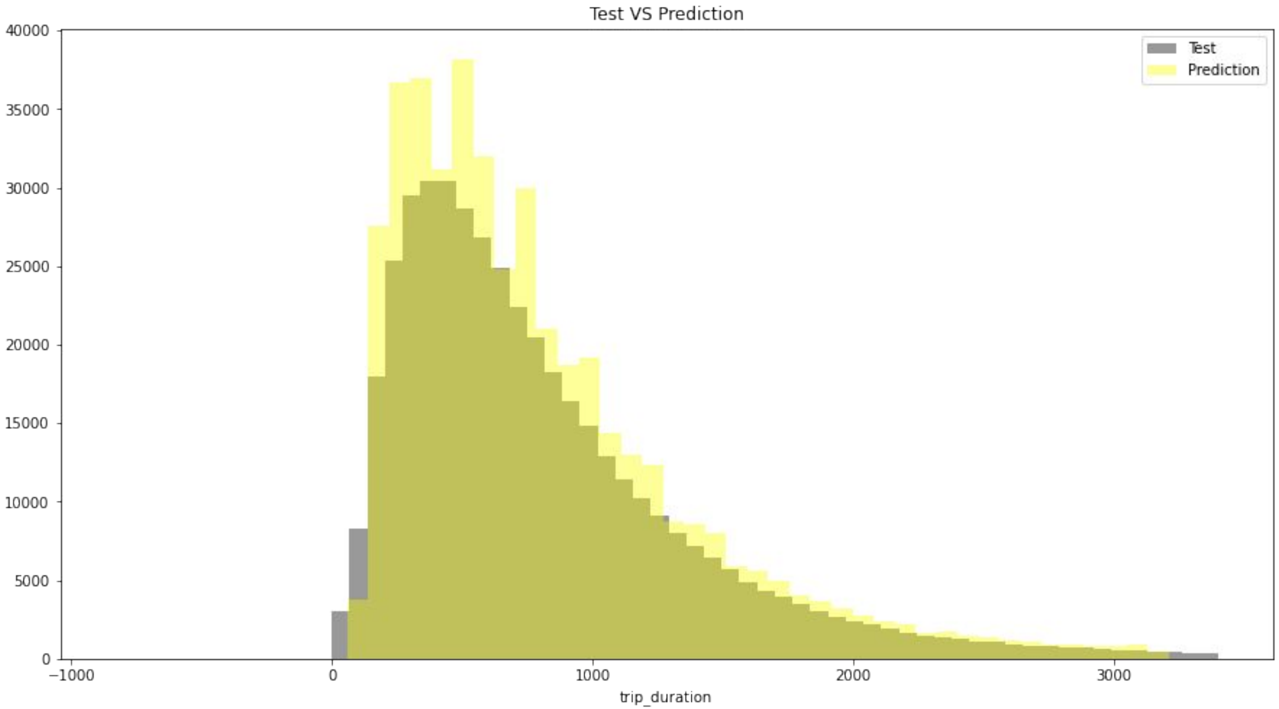
2192.367

R2 SCORE:

0.99309

ADJUSTED R2 SCORE:

0.99333



XGBoost seems slightly less effective than Tree Based Models.

CONCLUSION

Mostly 1 or 2 passengers avail the cab. The instance of large group of people travelling together is rare.

Most trips were taken on Friday and Monday being the least.
Fridays and Saturdays are those days in a week when peoples prefer to roam in the city.

The highest average time taken to complete a trip are for trips started in between 2 pm to 5 pm and the least are the ones taken between 5 am to 7 am.

Linear Regression doesn’t work well on this data.

The optimal model is Extra Trees Regressor.

Thank You