

Capstone Project

NYC Taxi Trip Time Prediction

by

Ritik Vaidande

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

trip_duration
duration of trip in seconds

dropoff_datetime
timestamp of drop-off

passenger_count
no of passenger riding

store_and_fwd_flag
store and forward record or not

pickup_latitude
latitude of pickup location

pickup_longitude
longitude of pickup location

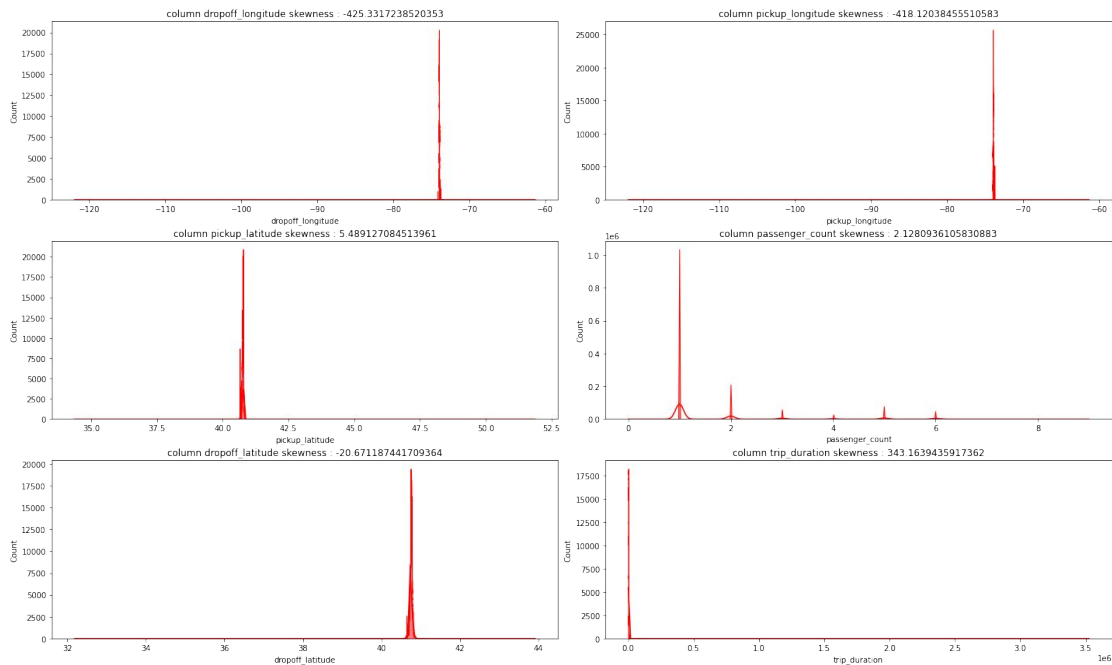
dropoff_latitude
latitude of drop-off location

dropoff_longitude
longitude of drop-off location

vendor_id
unique id of trip provider

DISTRIBUTION OF DATA

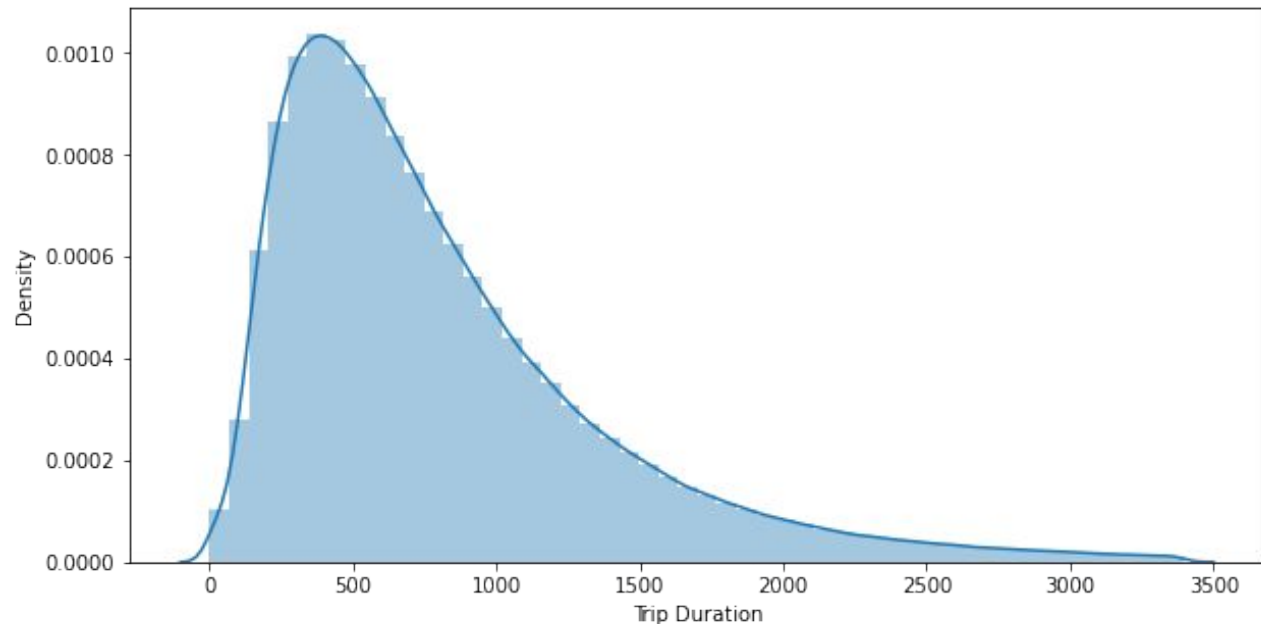
Majority Data
Distribution Type :
**Extremely
Skewed**



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**



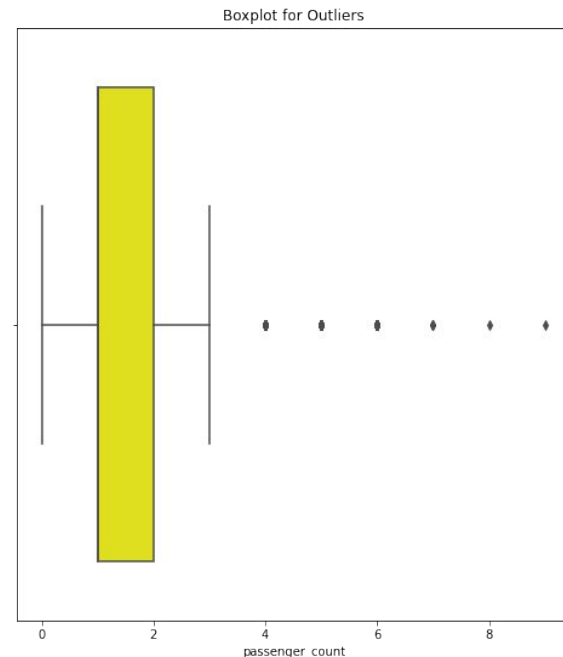
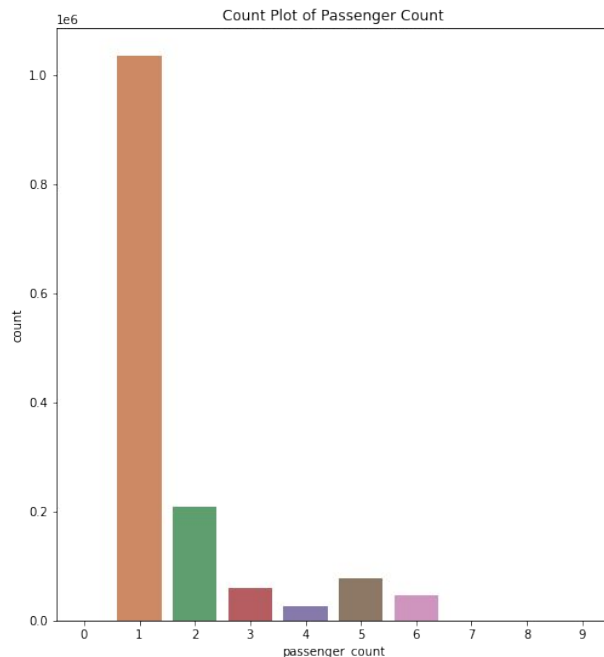
**99th percentile of trip duration is
completed under 3440 seconds i.e.
approx. 1 hour.**

PASSENGER COUNT DISTRIBUTION

Irrelevant Number of
Passenger:

0, 7, 8, 9

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

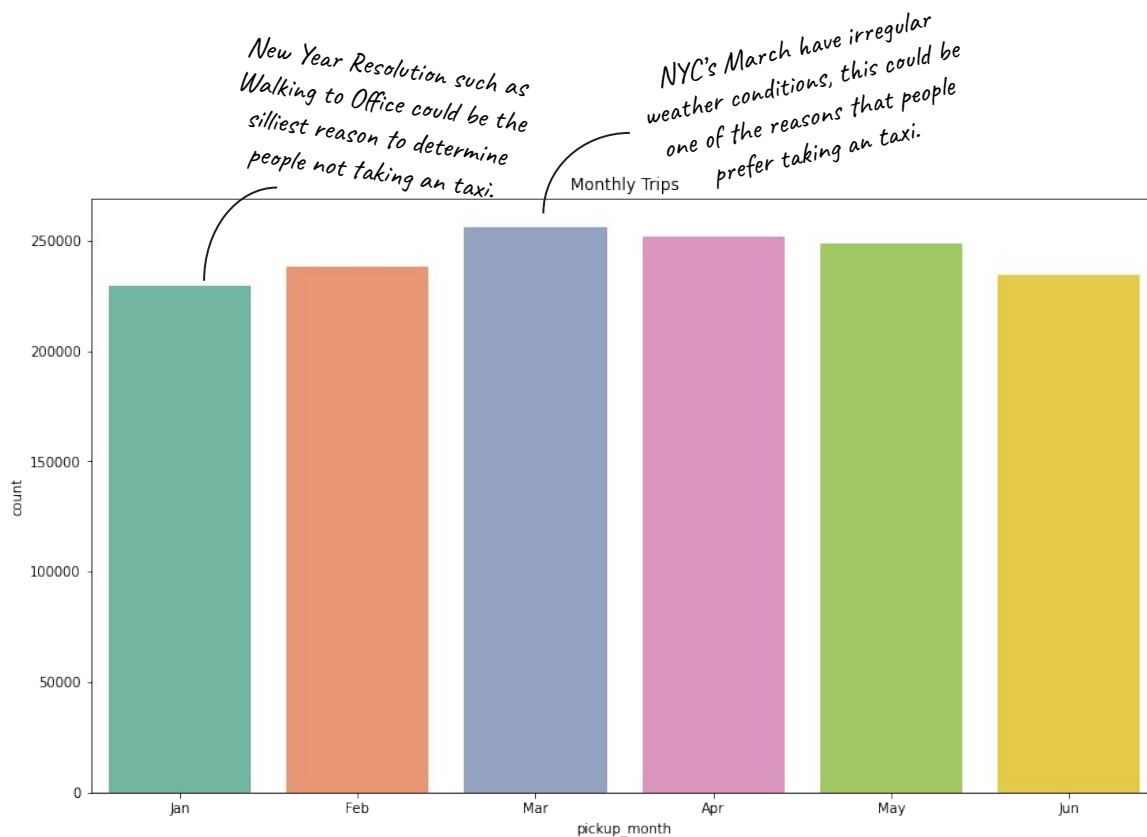


Single passenger trips holds the highest amount of Taxi trips. New Yorker's rarely travel in groups.

MONTHLY TRIPS COUNT

Most Busiest Month :
March

Least Busiest Month :
January



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

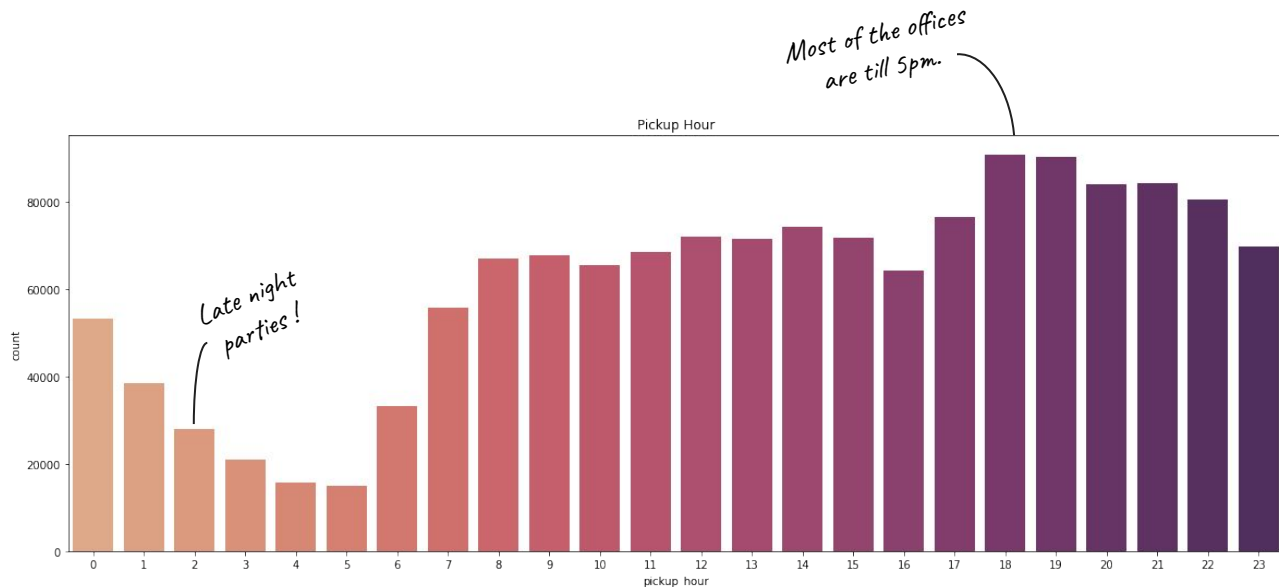
HOURLY PICKUP COUNT

Rush Hours:

6 to 7

No Rush Hours :

3 to 5

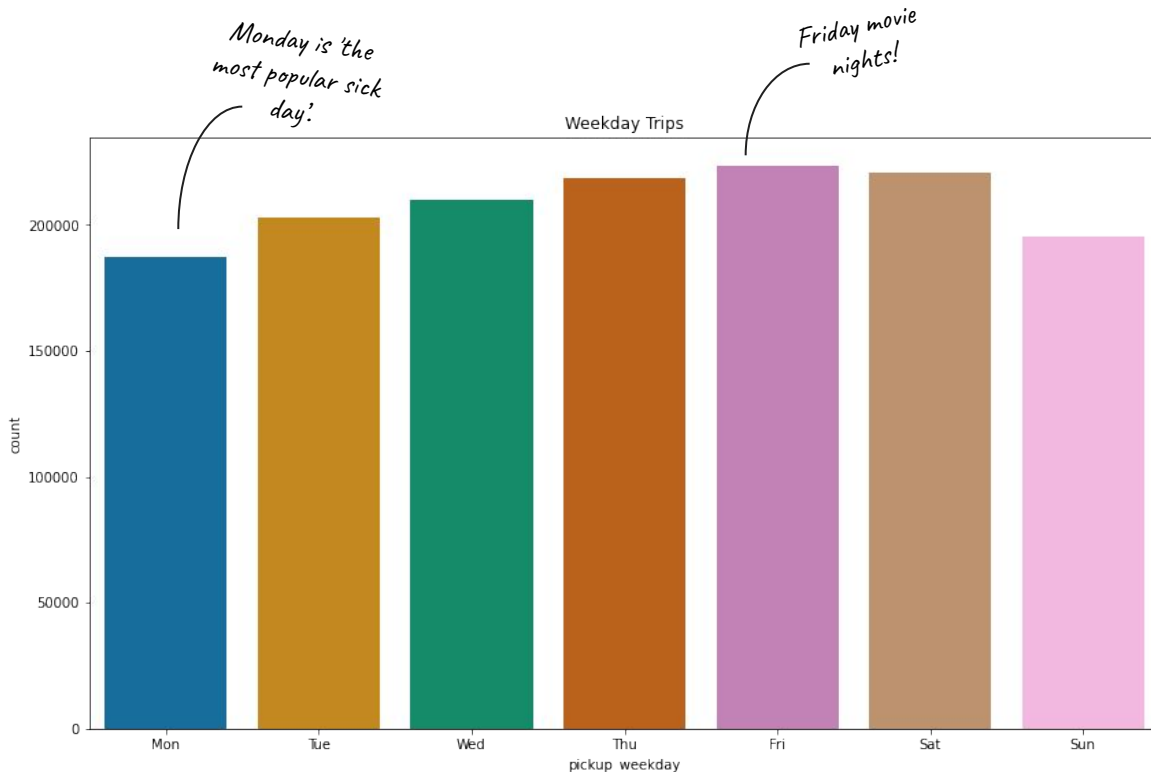


People prefer taking an taxi more after daylight.

WEEKDAY TRIPS COUNT

Most Busiest Day :
Friday

Least Busiest Day :
Monday

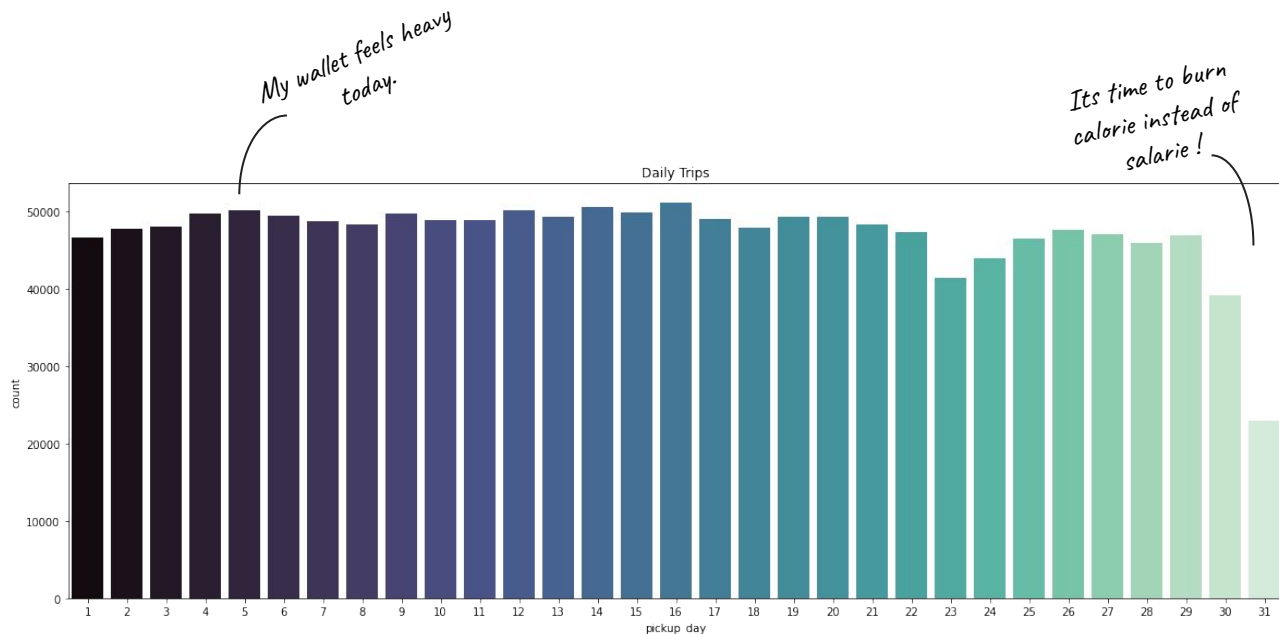


New Yorker's prefer going out on **Friday** and **Saturday**.

DAILY TRIP COUNT

Rush Days:
4 to 20

No Rush Days :
23 to 31

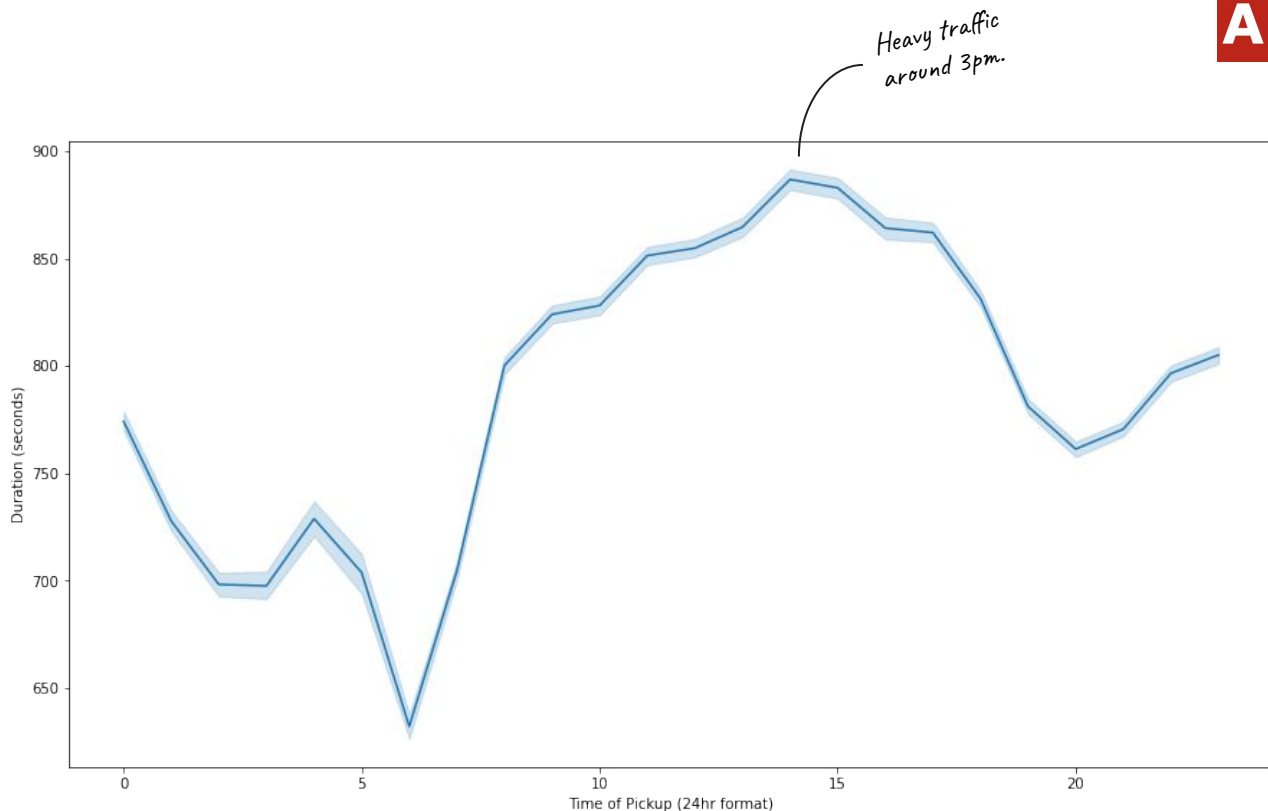


Seems like New Yorker's do not prefer to get a Taxi on Month-Ends.

TRIP DURATION PER HOUR

Time When Trip
Durations are Less :
5am to 7 am

Time When Trip
Durations are High:
2pm to 5pm

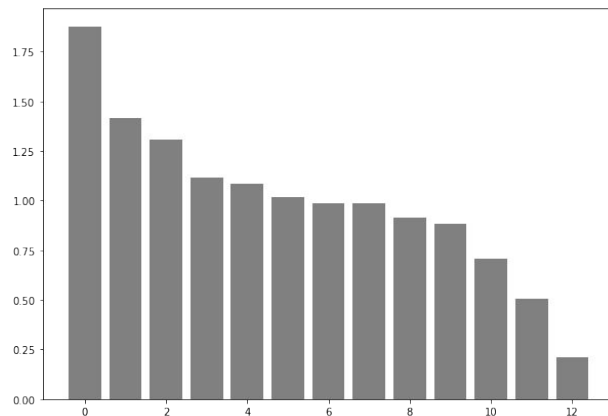


Around 3pm trip duration is high, NYC's heavy traffic could be the reason.

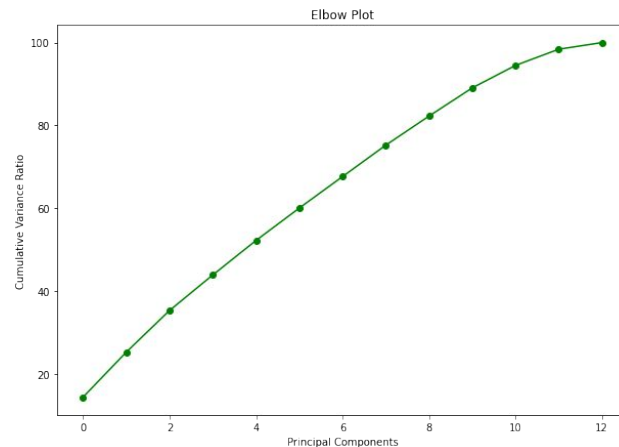
Principal Component Analysis

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.

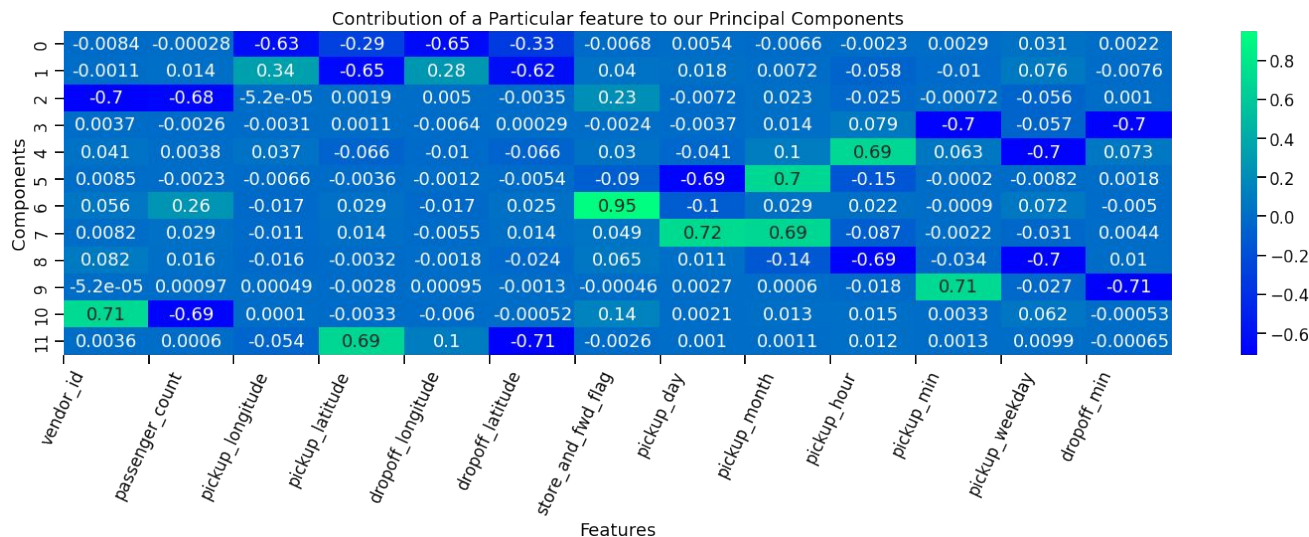


At 12th component our PCA model seems to go Flat without explaining much of a Variance.



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

PCA FEATURE IMPORTANCE MATRIX



- 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 **XGBoost**



Linear Regression

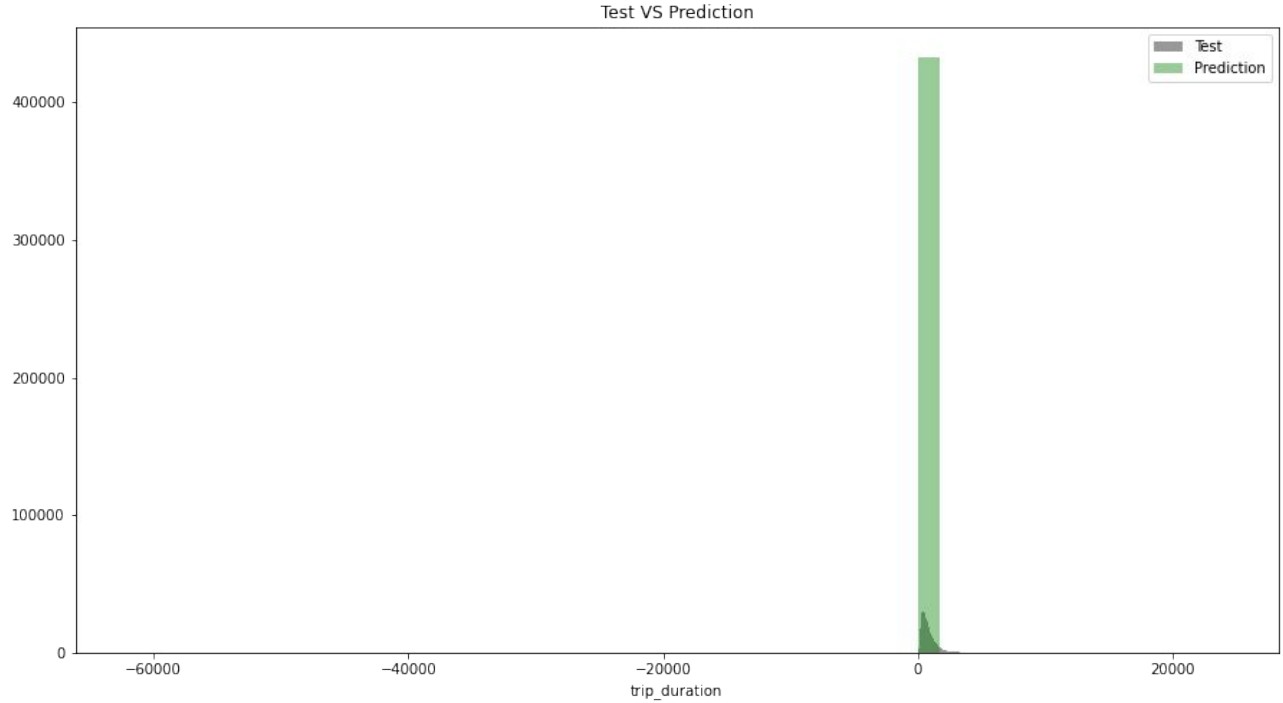


**Extra Trees
Regressor**



Decision Trees

LINEAR REGRESSION



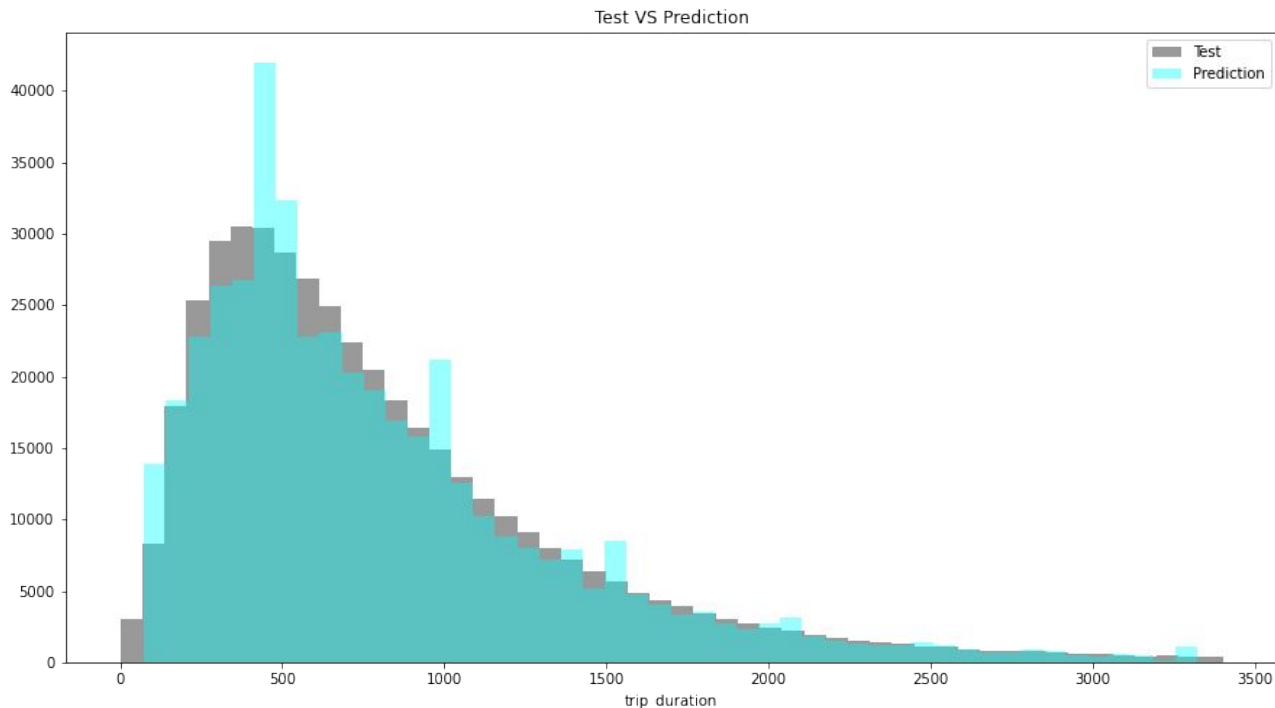
We can clearly see Linear Regression is not well suited for this data.

DECISION TREES

MSE :
1352.143

R2 SCORE:
0.99587

ADJUSTED R2
SCORE:
0.99588



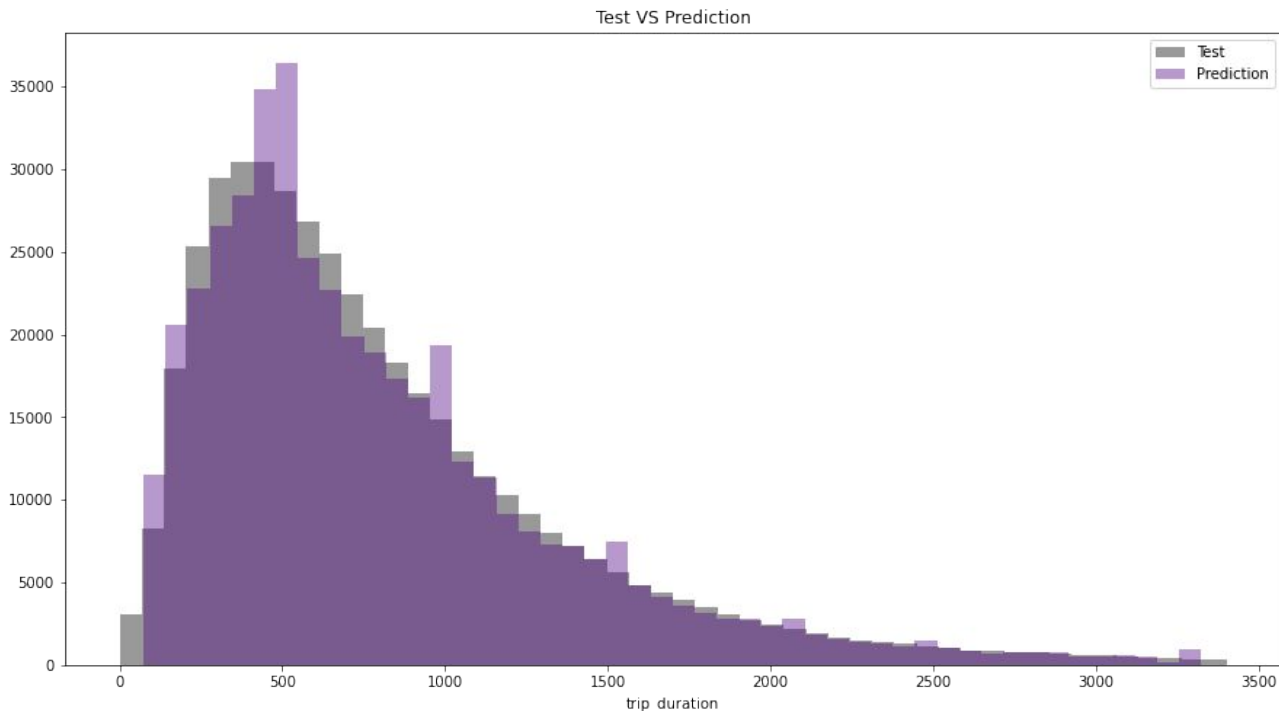
Decision Tree has performed well has compared to Linear Regression.

RANDOM FOREST

MSE :
1186.074

R2 SCORE:
0.99637

ADJUSTED R2
SCORE:
0.99639



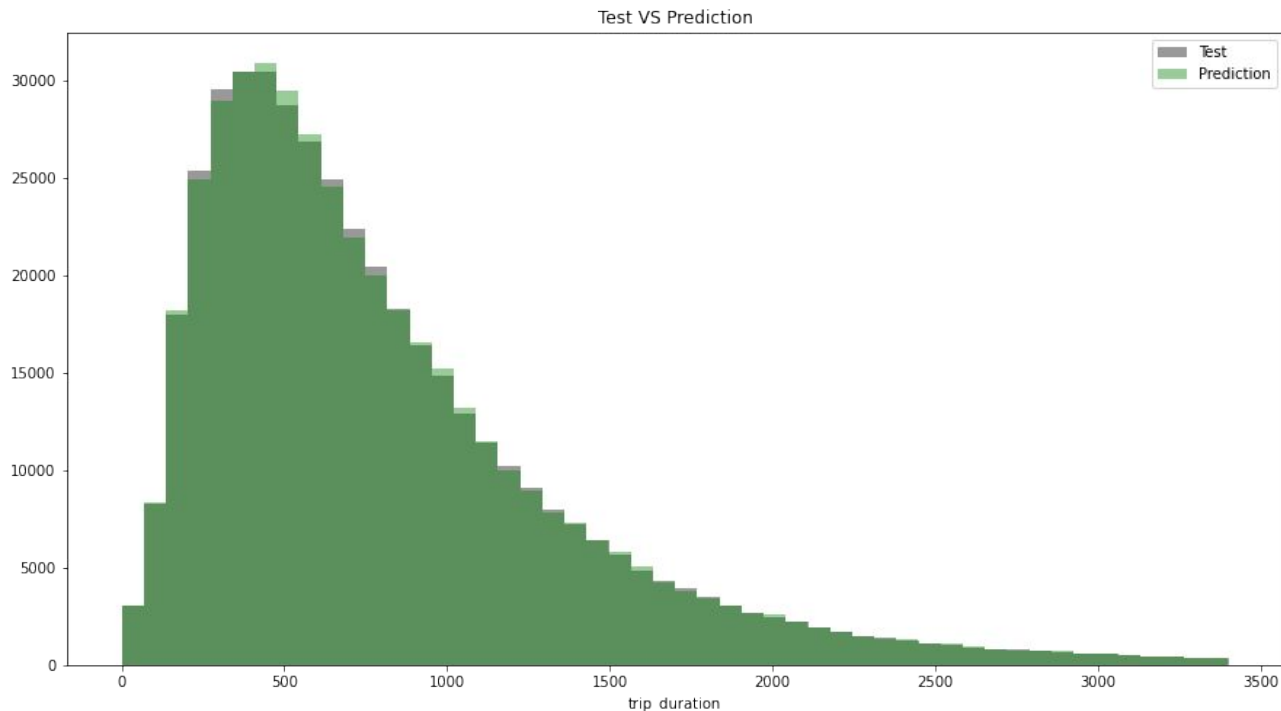
Random Forest has performed slightly better than Decision Trees.

Extra Trees Regressor

MSE :
1007.620

R2 SCORE:
0.99693

ADJUSTED R2
SCORE:
0.99693



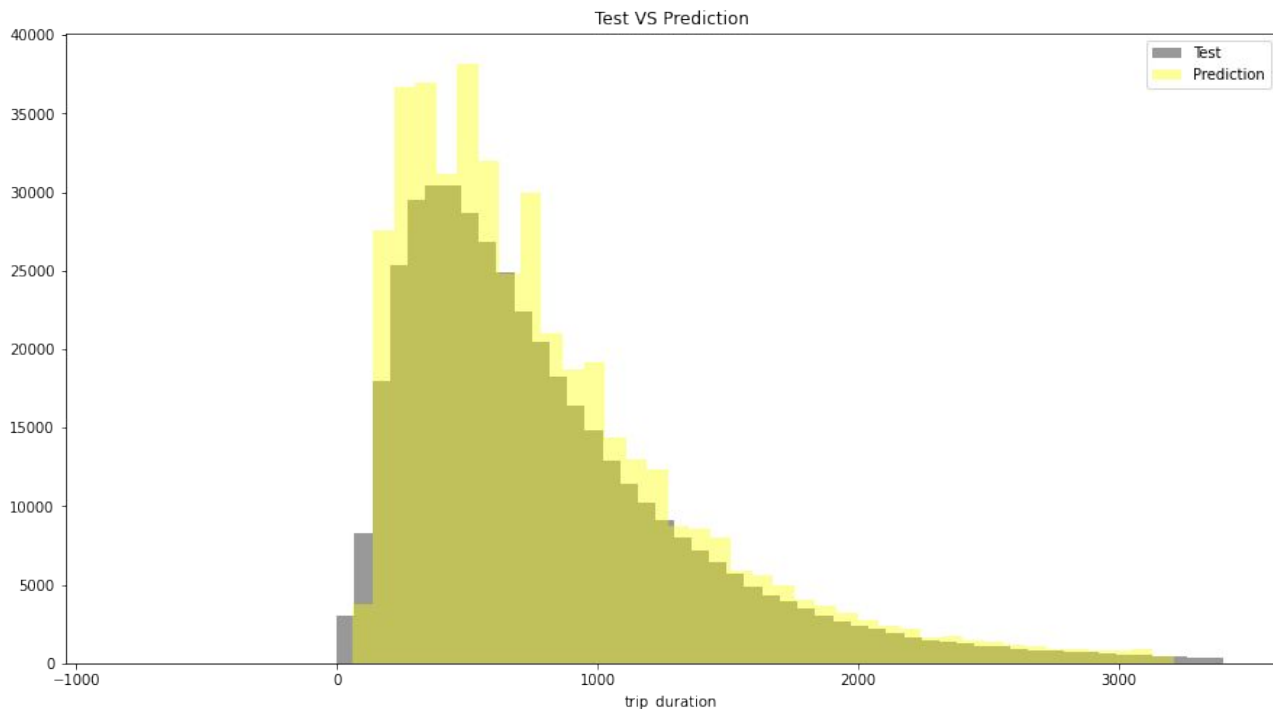
Extra Tress Regressor appears to be the optimal model.

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

This presentation is an part of the Regression Capstone Project by Ritik Vaidande,
AlmaBetter.