Capstone Project - 2

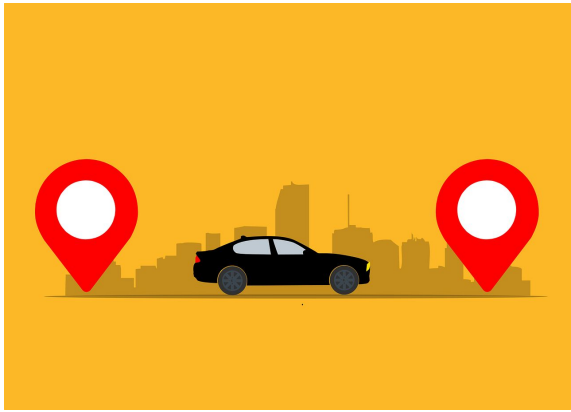
NYC Taxi Trip Time Prediction

Prepared By:-
Amlana Jyoti Biswal
Manas Ranjan Behera
Varnit Chauhan

Table of content:

Introduction:-

- New York, often called New York City (NYC) is the most populous city in the United States. Its also most densely populated city.
- With this large population comes traffic problem. The duration of travel increases as there are more number of vehicles on road.
- It is seen that most of the New Yorkers prefer Taxi services, as the popularity of Taxi increases, service companies want to inform their riders about their estimated arrival(ETA), trip distance, total fare etc.
- So, here we are going to apply various model to predict the taxi trip duration by considering various aspects such as distance, pickup time, traffic etc.



Problem Statement:-

Our main goal is to make a suitable machine learning model, so that we can predict the taxi trip duration for NYC Taxi.

The primary dataset is one released by the NYC Taxi and Limousine Commission, which includes pickup time, geo-coordinates, number of passengers, and several other variables.

Data Summary:-

Given NYC Taxi csv dataset contains 1458644 trip records.

- 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.

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.

- trip_duration - duration of the trip in seconds.

- **Number of rows in our dataset are 1458644 and the Number of columns in our dataset are 11.**
- To understand the problems with the data, such as missing values, inaccuracies, and Outliers we have checked the dataset.
- Luckily there are no NAN/NULL values in our dataset.

Feature Creation:-

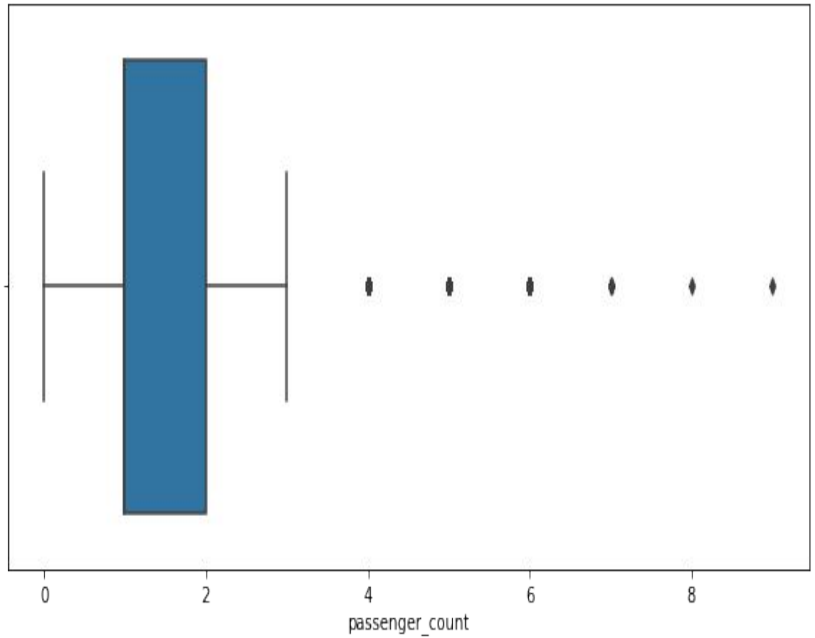
We have created the following features:

- pickup_weekday which contains the name of the day on which the ride was taken.
- pickup_weekday_num which contains the day number instead of characters with Monday = 0 and Sunday = 6.
- pickup_hour with an hour of the day in the 24 - hour format.
- pickup_month with month number as January = 1 and December = 12.
- Distance from geographical coordinates using Haversine.
- Calculated the Speed in km/h from above distance data.

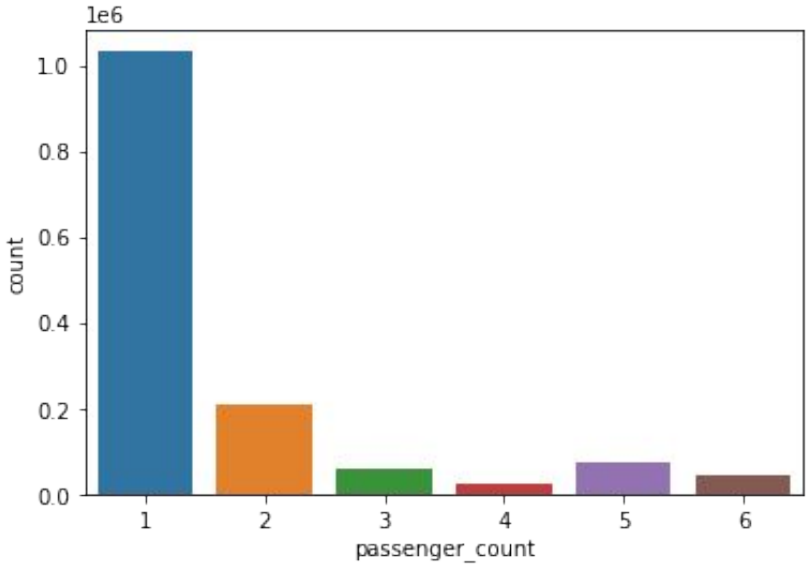
EDA Univariate Analysis

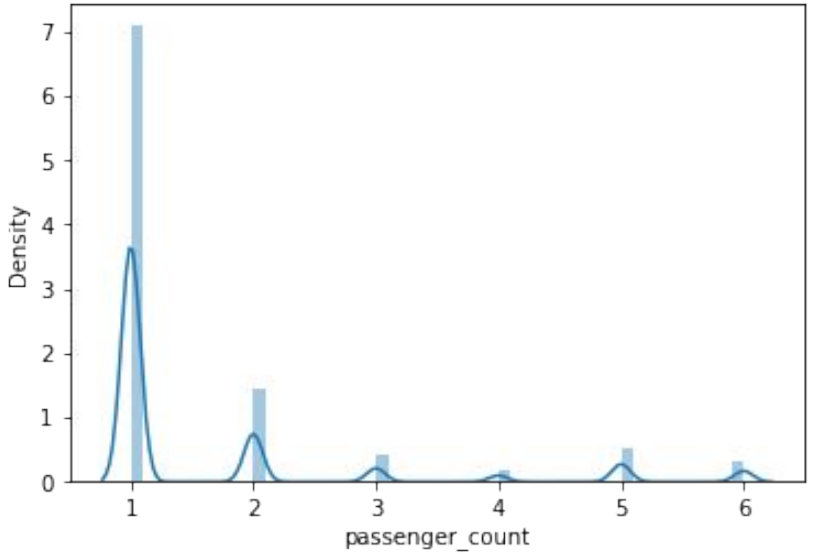
Number of Passengers per Taxi

- From the Box Plot we can clearly notice there are some outliers in our dataset. It shows number of passengers per taxi is more than 7.
- In most of the trips passenger number is between 1 or 2.
- Minimum number of passengers per taxi is 0.
- The minimum number is 0, it looks little concerning as a trip can not be done without a passenger.



Passenger count Vs. Number of Trips

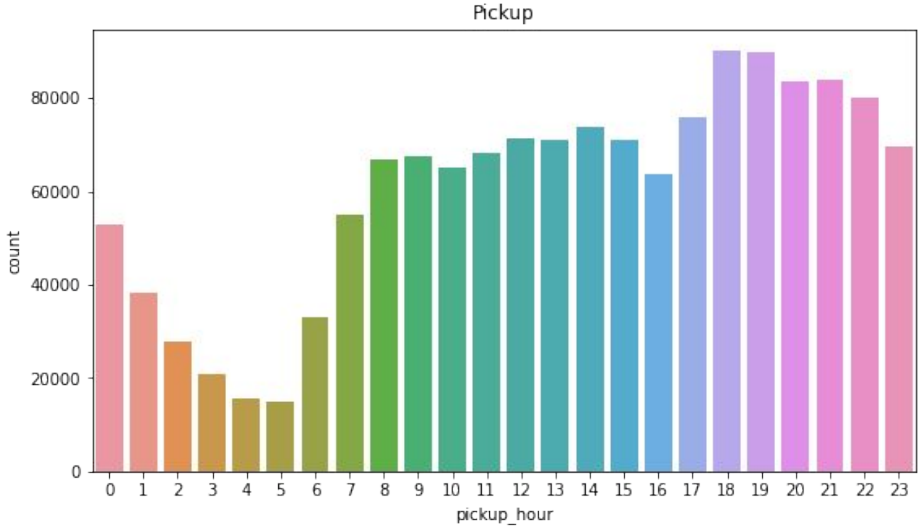


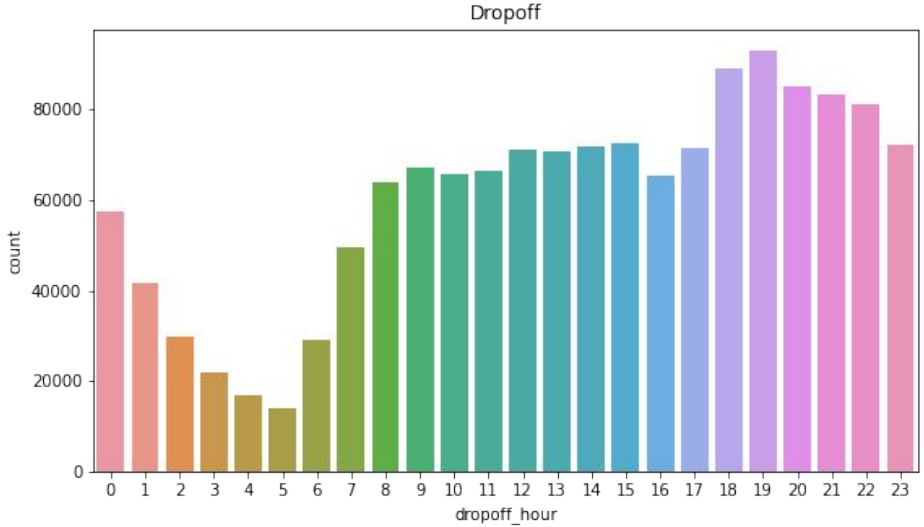


we can observe that most of the trips were taken by 1 passenger.

Distribution of trips across Day

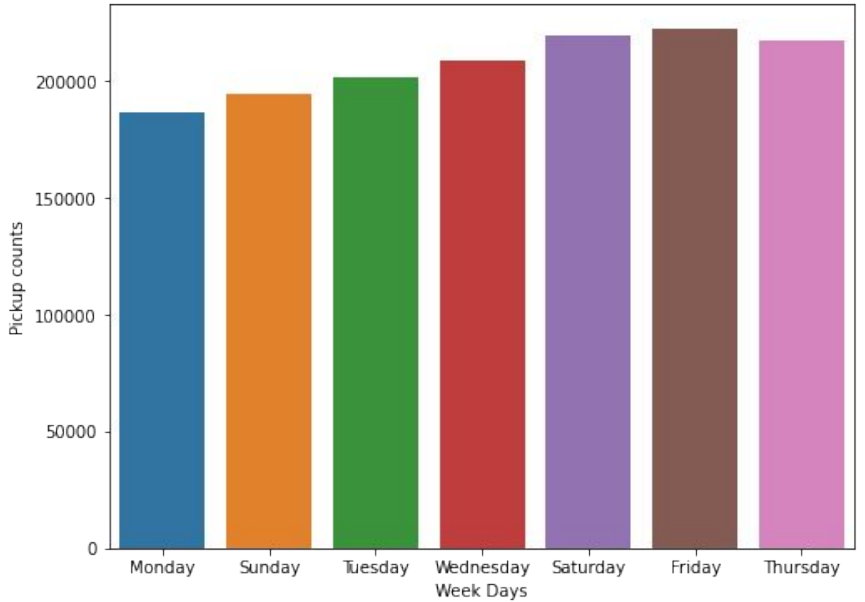
- Total trips Per Hour : The plot consist of the distribution of the pickups and dropoffs across the 24 hour time scale.
- Most number of pickups are at 6 PM to 7 PM. Least number of pickups are at 4-5 AM.
- This coincides with the general trend of taxi pickups which starts increasing from 6AM in the morning and then declines from late evening i.e. around 8 PM. There is no unusual behavior here.

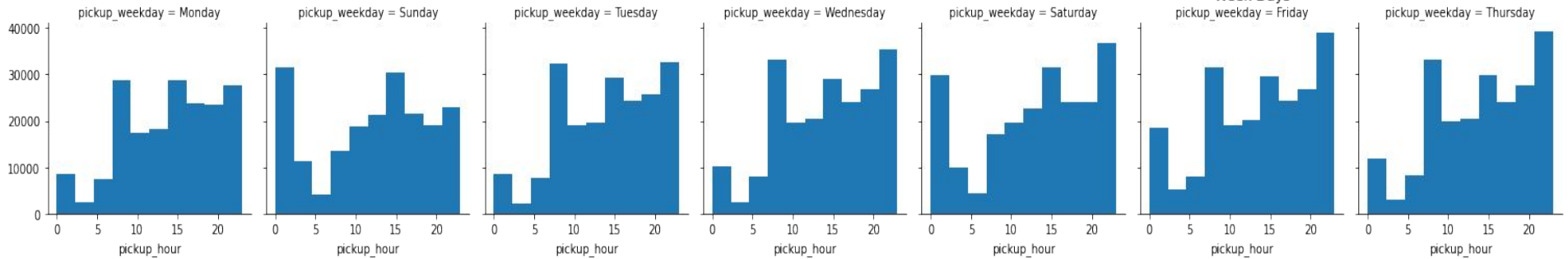




Trips Per Day

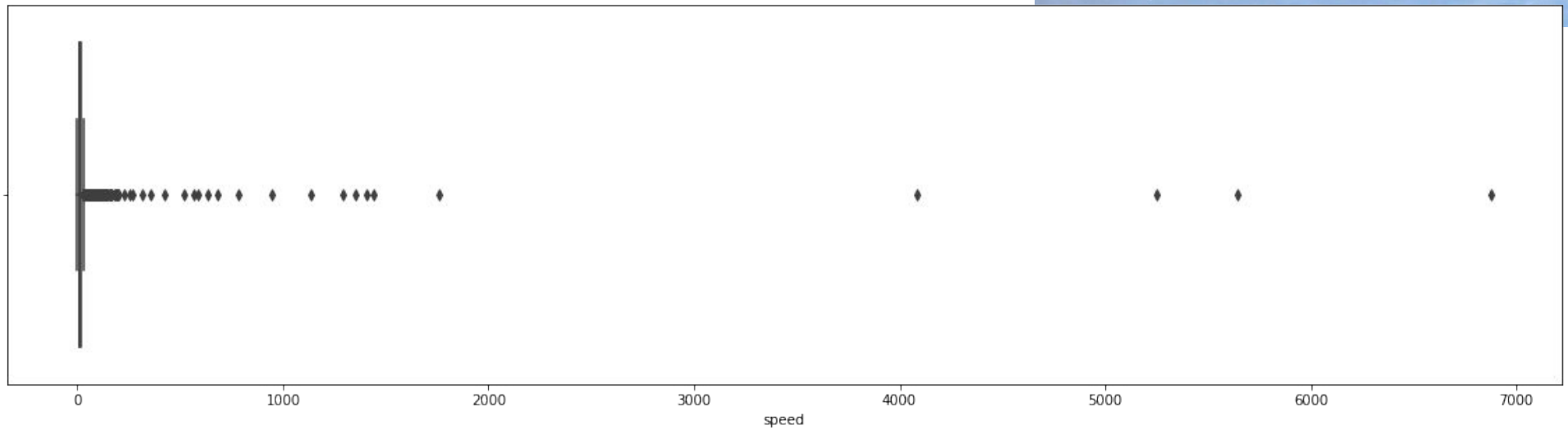
- We can see the number of pickups starts increasing from monday to friday, then starts decreasing on weekends. Because offices and other establishments are closed on weekends.
- Pickups are maximum on friday most probably because of the weekend.
- During weekends late night pickups are more than weekdays.
- There is consistent high pickups at 7 AM during weekdays, most probably because of office hours.





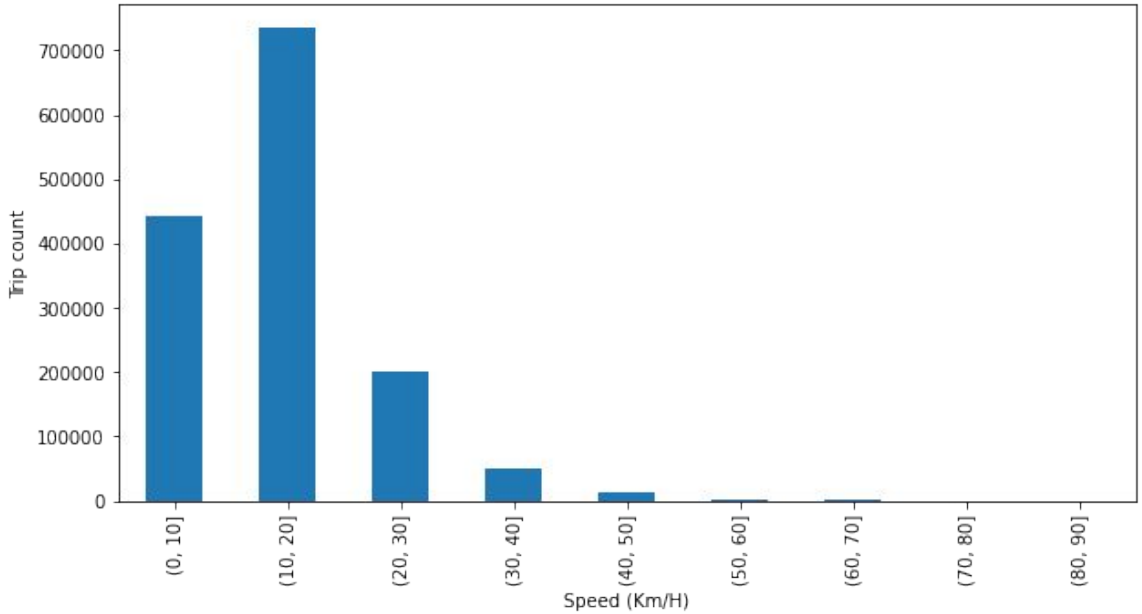
Speed

- As we can see there are some trips with speed more than 1000 km/h. Maybe they are trying to break the sound barrier.
- These are some outliers, so we are limiting the speed to somewhere around 100 km/hr.



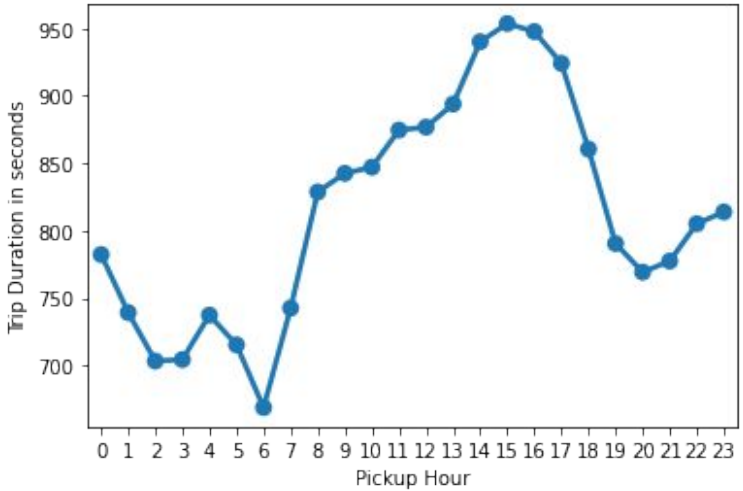
Speed

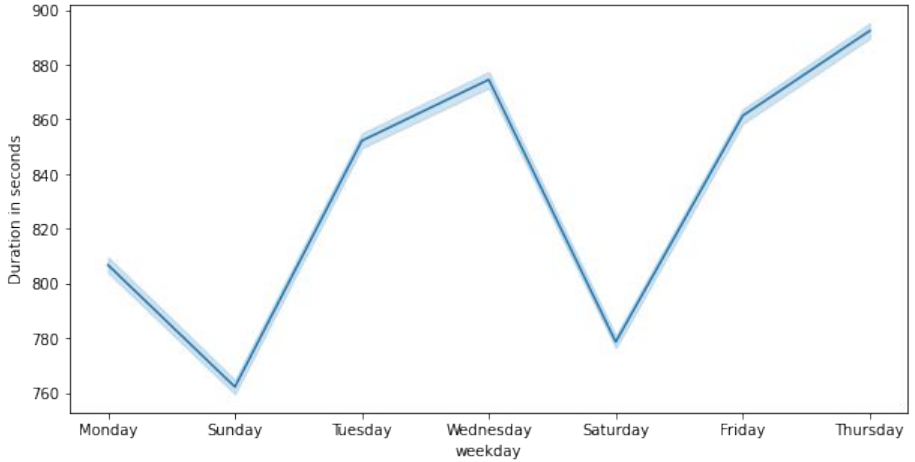
- After limiting speed to 100 km/hr.
- We can see that most of the trips are done at speed 10-20 km/hr.
- It's obvious that in a densely populated city like NYC, speed will be limited because of traffic.



Bivariate Analysis:-
Trip Duration Per Hour

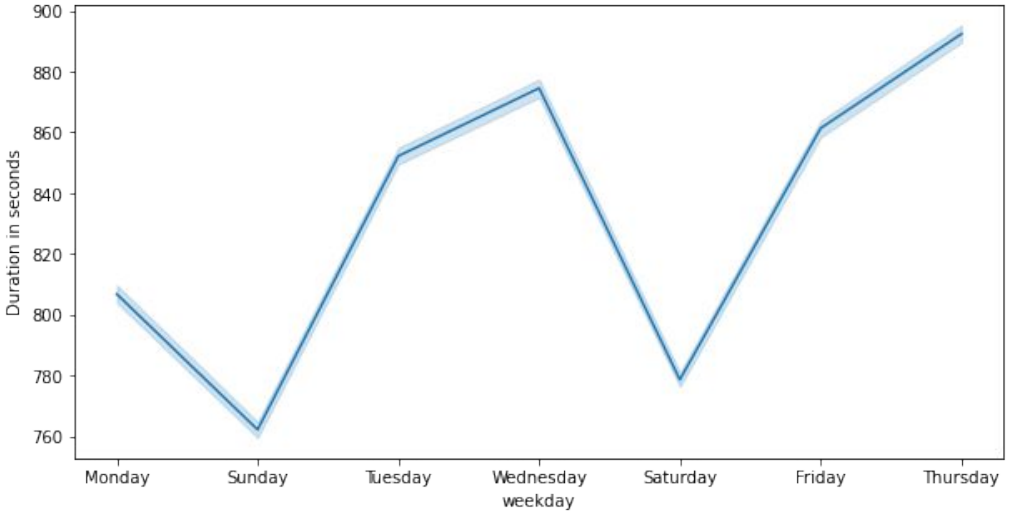
- From below plot we can see lowest mean trip duration is at 6 AM.
- Highest mean trip duration is at 3 PM. Most probably due to high traffic, as its closing hour for offices and schools.
- Trip duration is highest on thursday.
- Trip duration is lowest on weekends, because of holiday.

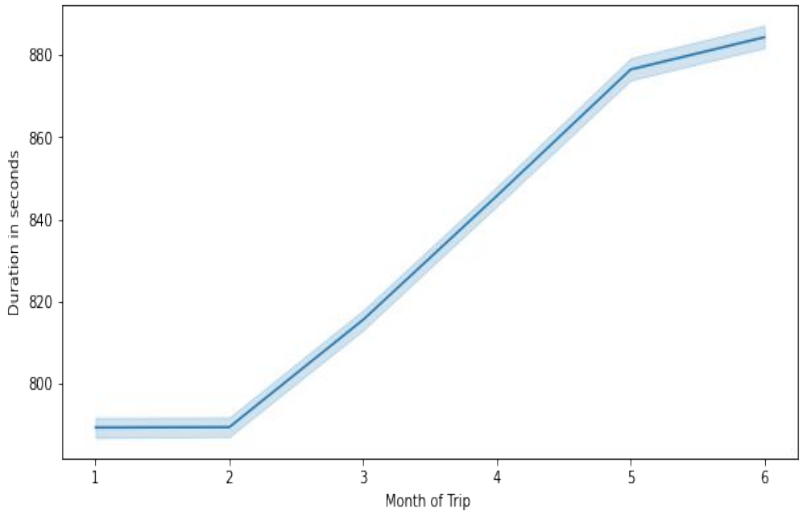




Trip Duration per Weekday and per Month

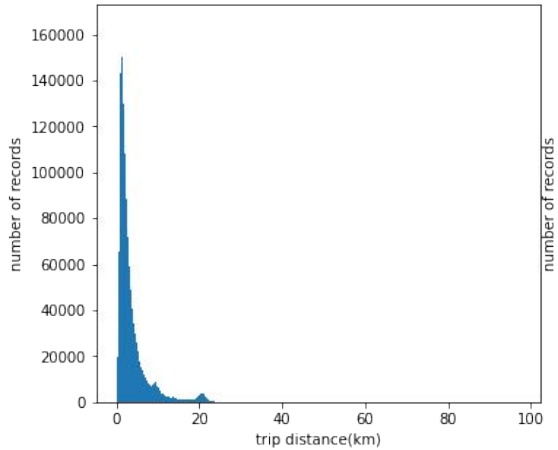
- Trip duration is highest on thursday. Can't predict the underlying cause from the data.
- Trip duration is lowest on weekends, because of holiday.
- There is an increasing trend of trip duration as month increases.
- This might be due to some climatic conditions like rain and snow.
- As in rainy season roads are more blocked and trip duration will increase.

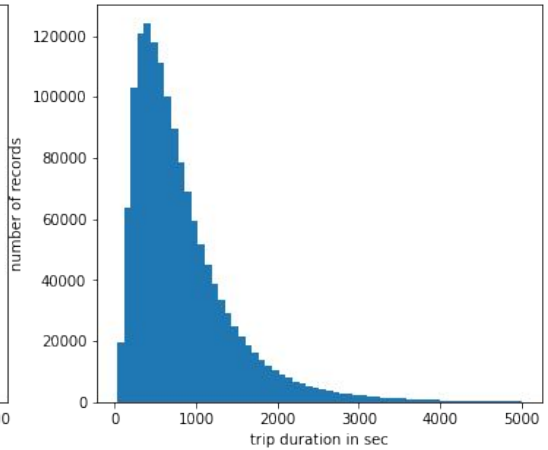


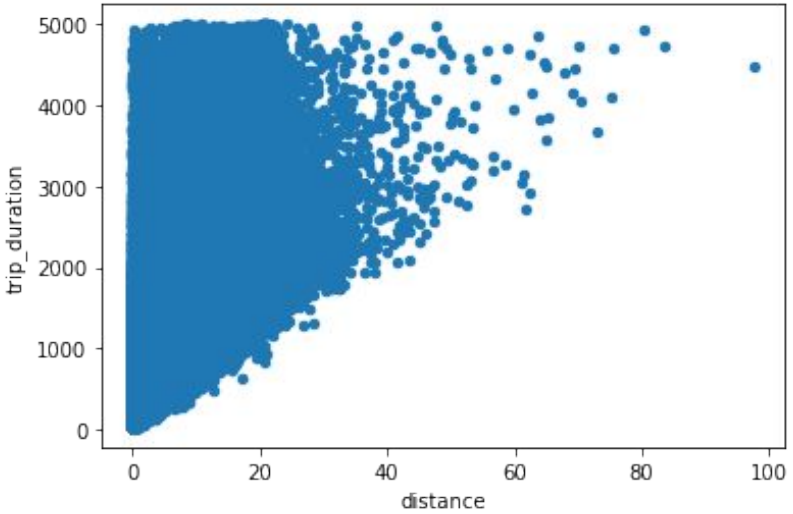


Trip Distance and Duration Distribution

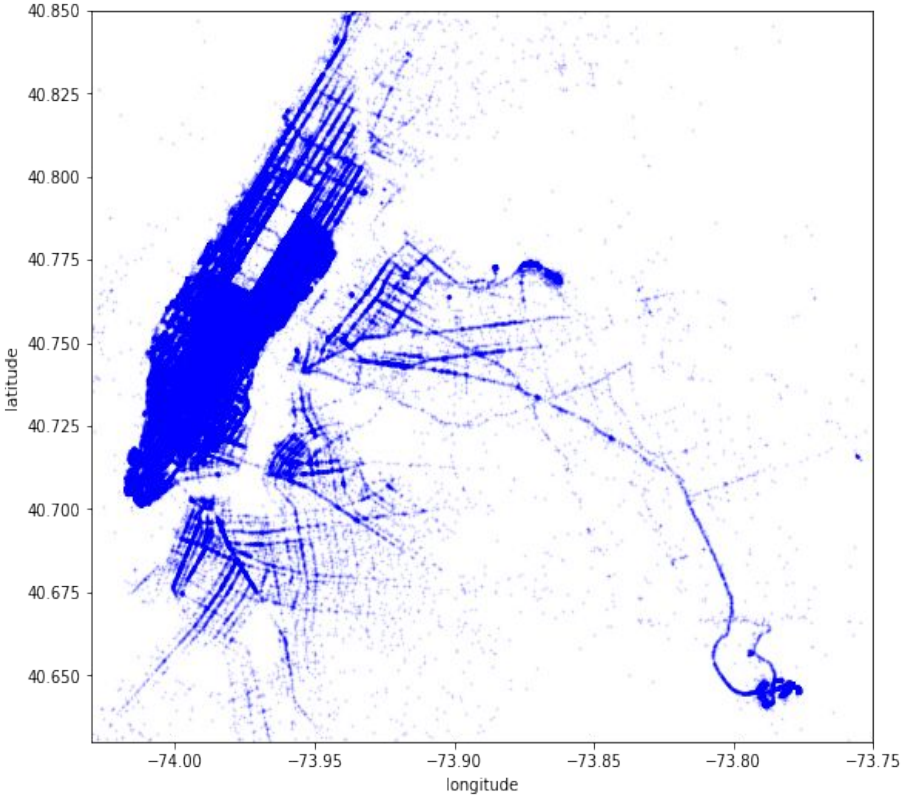
- Both trip distance and trip duration are right skewed.
- Most of the trip distance is distributed in between 0-5 Km and most of the trip duration is distributed between 0-500 sec.
- There should have been a linear relationship between the distance covered and trip duration on an average but we can see dense collection of the trips in the lower right corner which showcase many trips with the inconsistent readings.

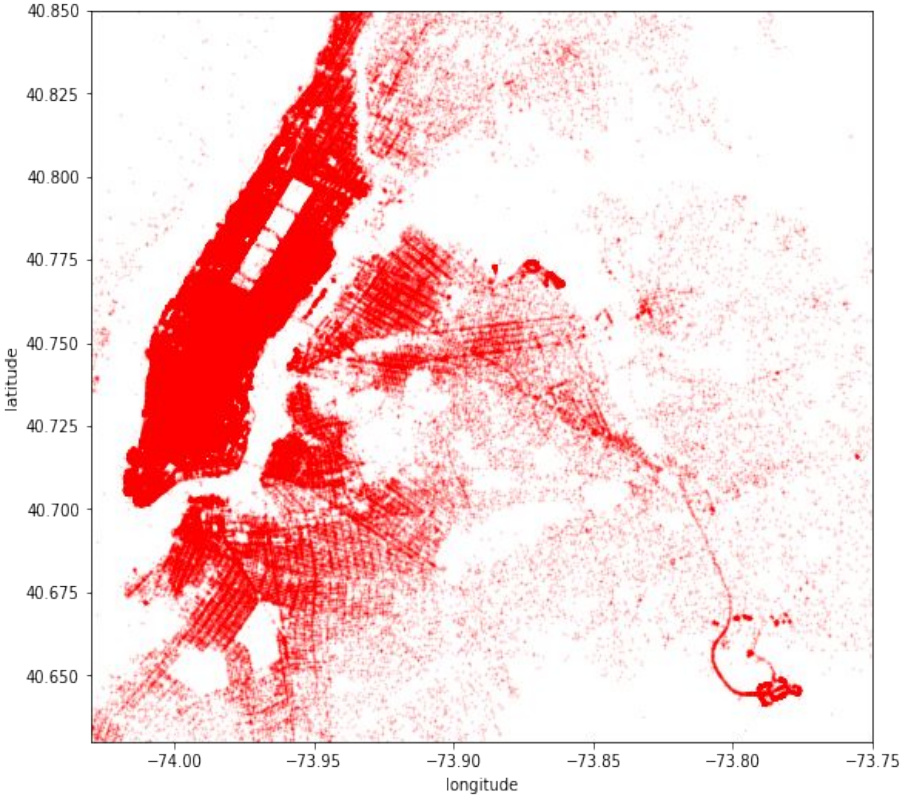






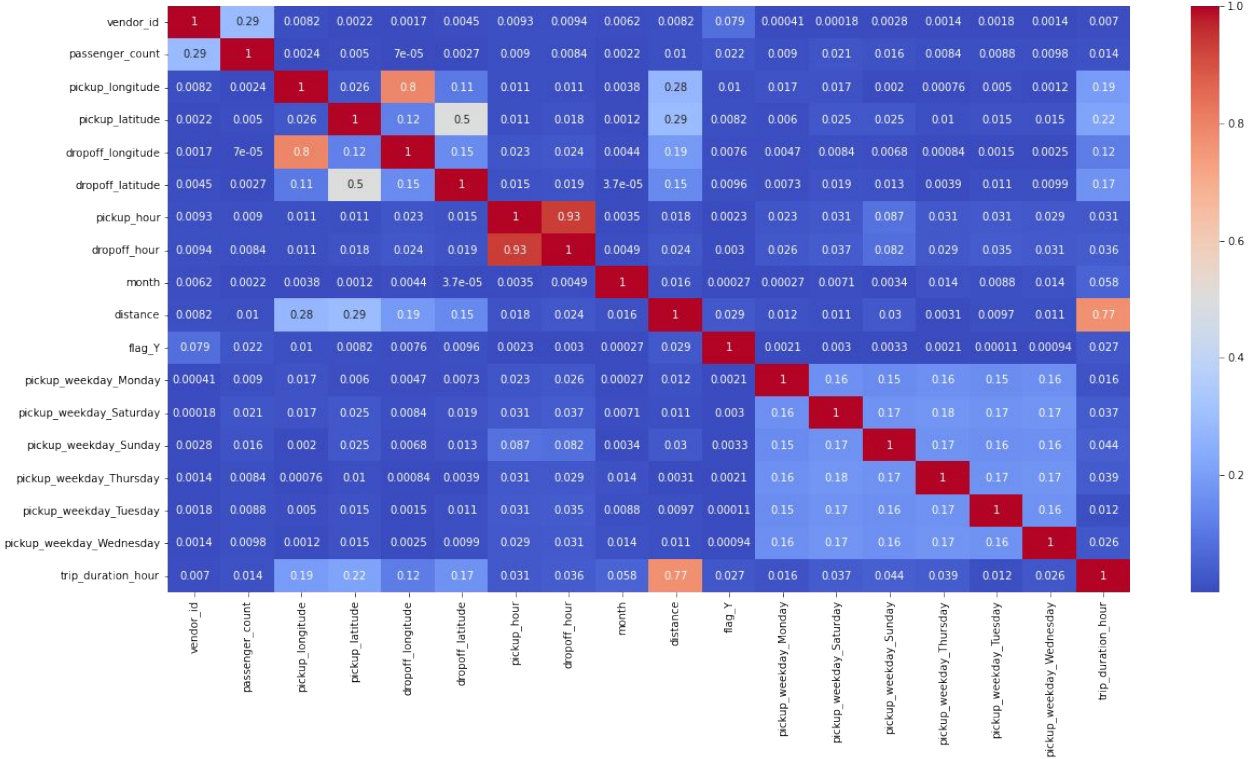
Pickup and Dropoff Locations on Map





Correlation HeatMap

Correlation Analysis: We draw heatmap to find correlation between different independent variables and dependent variables. If correlation between independent features are high and has very less relation with dependent feature, we ll remove them.



Feature Engineering:-

One Hot Encoding :

We used one hot encoding to create new features such as pickup weekdays and represented them with 1 and 0. So that it can be interpreted by the machine more easily.

Dummify features like store_and_fwd_flag and pickup_weekday.

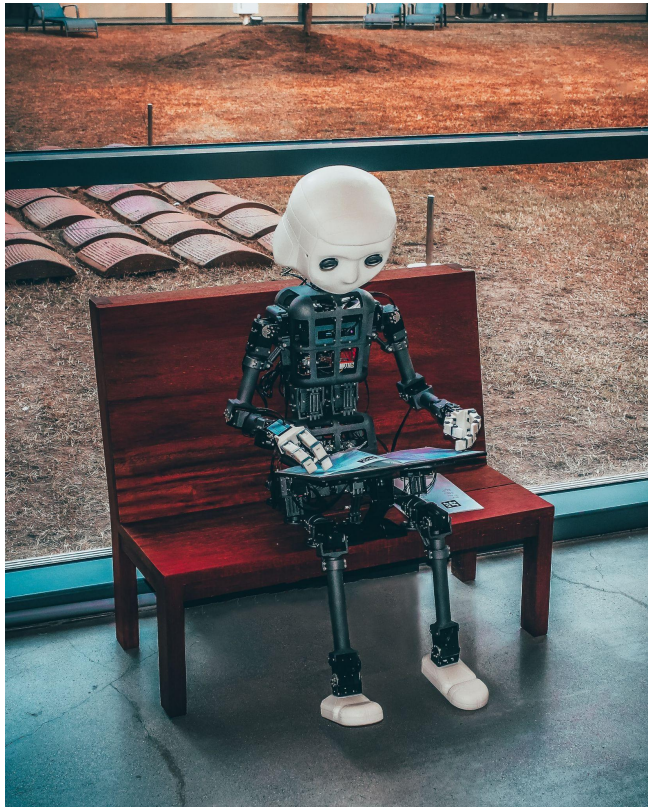
Feature Selection: We removed the columns which are not important for further analysis such as id, pickup_datetime,dropoff_datetime, store_and_fwd_flag, pickup_weekday, dropoff_weekday,pickup_weekday_num, pickup_timeofday, trip_duration,speed.

AI

Model Creation:-

- **Linear Regression :** This regression technique finds out a linear relationship between independent variable and dependent variable. It finds the set of θ coefficients that minimize the sum of squared errors.

- **Lasso Regression :** The lasso method was used to shrink coefficients. For duration prediction models, lasso was run using a range of values for the penalizing parameter, λ . Grid Search was used to find the lasso model with the lowest error and select the value of λ to use.



Model Creation(continued)..

- **XGBoost** was used for final prediction of the trip duration in the test dataset. The dataset was very large, as a result for this type of problem XGBoost was applied in which all the attributes were taken and parallel processing of boosting trees executed. Another aspect of XGBoost is that it keeps a nice check between bias and variance which helps in better prediction. The results were interpreted by using GridSearch, the XGBoost hyperparameters.

- **LightGBM** is just a gradient boosting framework based in Decision Tree modelling to increase the efficiency of the model and to reduce the memory usage. It splits the tree leaf_wise and chooses that leaf wise max dataloss to grow. Since the leaf is fixed, the leaf wise algorithm has low loss than compared to other models.

Model Evaluation:-

Training Model	Train MSE	Test MSE	Train RMSE	Test RMSE	Train R2	Test R2	Train Adjusted R2	Test Adjusted R2
Linear Regression	0.0568	0.0567	0.2383	0.2381	0.4477	0.4483	0.4476	0.4483
Lasso Regression	0.0613	0.0612	0.2476	0.2475	0.4040	0.4038	0.4040	0.4037
XGBoost	0.0172	0.0197	0.1312	0.1403	0.8325	0.8083	0.8325	0.8083
LightGBM	0.0108	0.0109	0.1042	0.1045	0.8943	0.8936	0.8943	0.8936

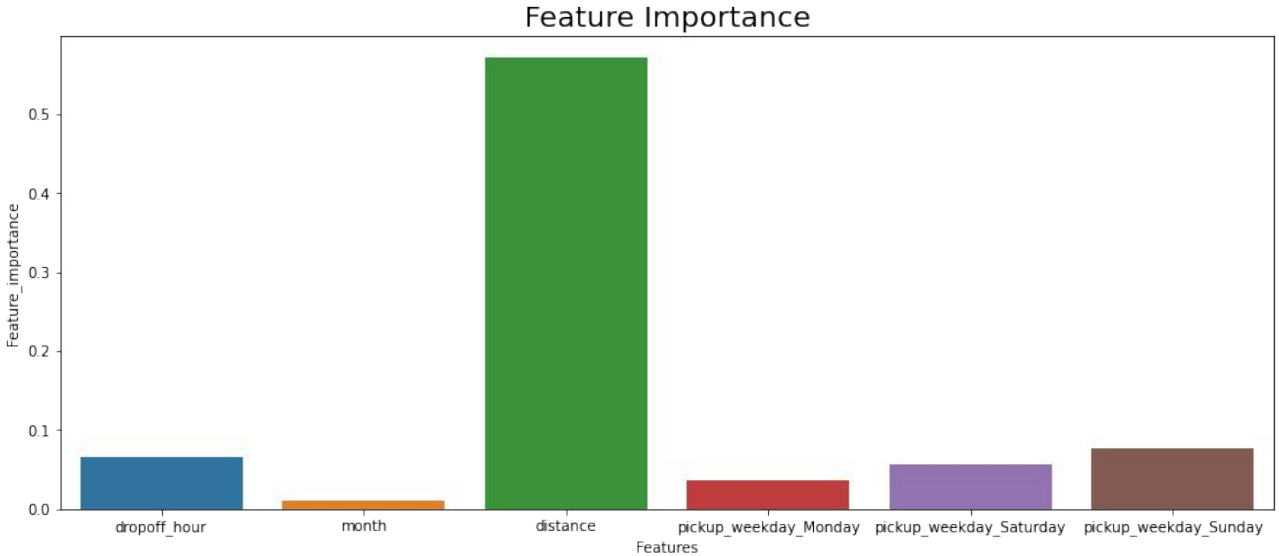
Conclusion

AI

- We compared MSE, RMSE and R2 for all four regression models, to find which is the best model to predict the NYC taxi trip duration.
- The Linear Regression and Lasso Regression didn't show any good prediction as compared to the other two.
- From above comparison table we can clearly see that **XGBoost** and **LightGBM** are the best models to predict trip duration of the NYC taxi. **LightGBM** is fastest and more accurate than **XGBoost**. So, in between these two **LightGBM** is the best model.
- **R Square:** R2 score represents the coefficient of how well the values fit compared to the original values. The value from 0 to 1 interpreted as percentages. The higher the value is, the better the model is.
- **RMSE (Root Mean Squared Error):** RMSE is the standard deviation of the residuals. Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it displays how concentrated the data is around the line of best fit.

Conclusion Continued…..

After doing different regression on the given dataset we found out that Distance is the independent variable which affects most the trip duration.



Challenges

- It’s no doubt this is a large dataset containing 1458644 rows and 11 columns.
- In order to handle the large dataset we took extra care in removing outliers.
- We took extreme care to handle feature selection part as it affect to the R2 score.
- Also focused on Hyperparameters as it also affects the R2 score.

DANKE!
THANK YOU!
MERCI!
GRAZIE!
GRACIAS!
DANK JE WEL!

. . . . . .