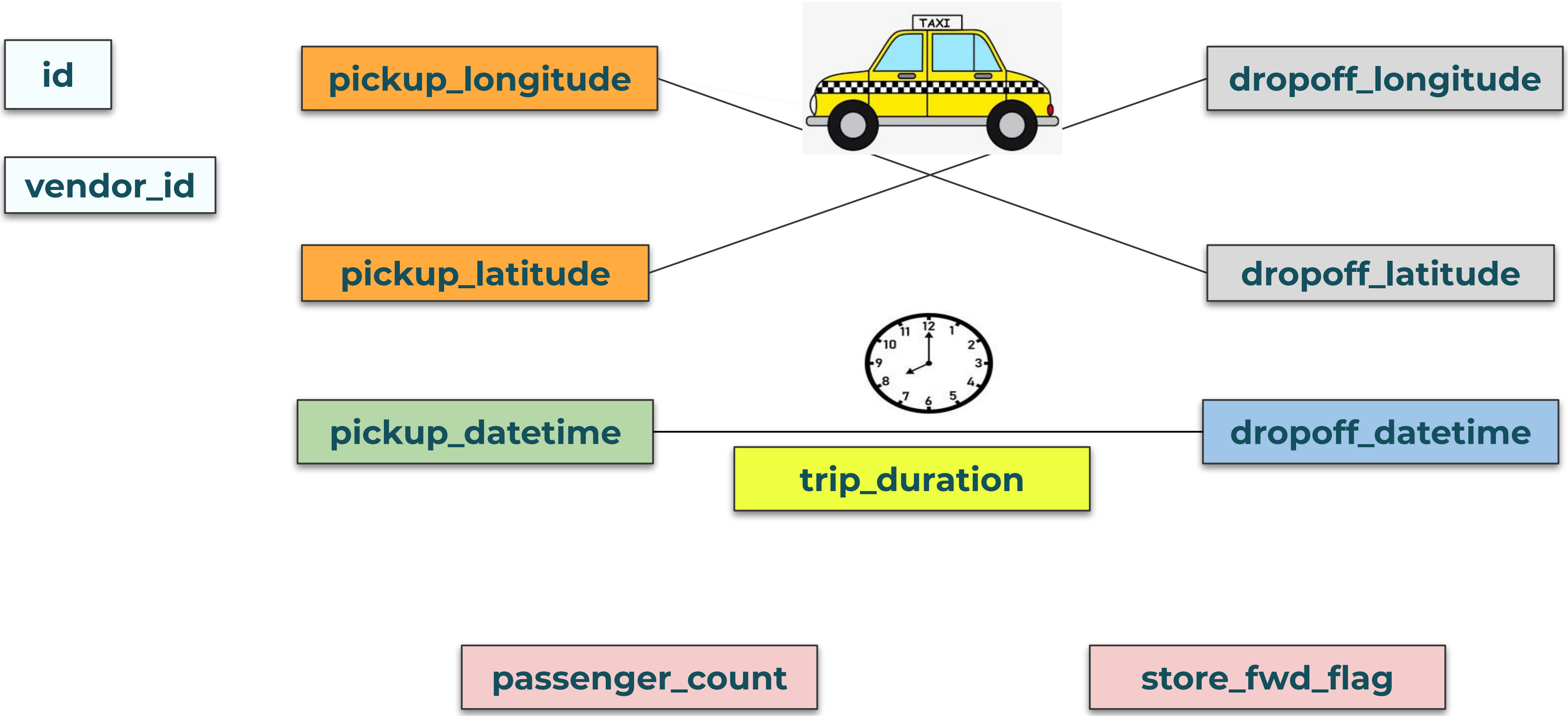
Capstone Project - 2
Supervised ML - Regression
NYC Taxi Trip Time Prediction

By
Subhajit Ganguly
Data Science Trainee, Almabetter

Problem in Hand

Our task is to build a model that predicts the total ride duration of taxi trips in New York City of NYC Taxi. Our 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 on a taxi trip.

Variables that Describe



Approach Discussion

Let's discuss, how we are going to solve the problem.

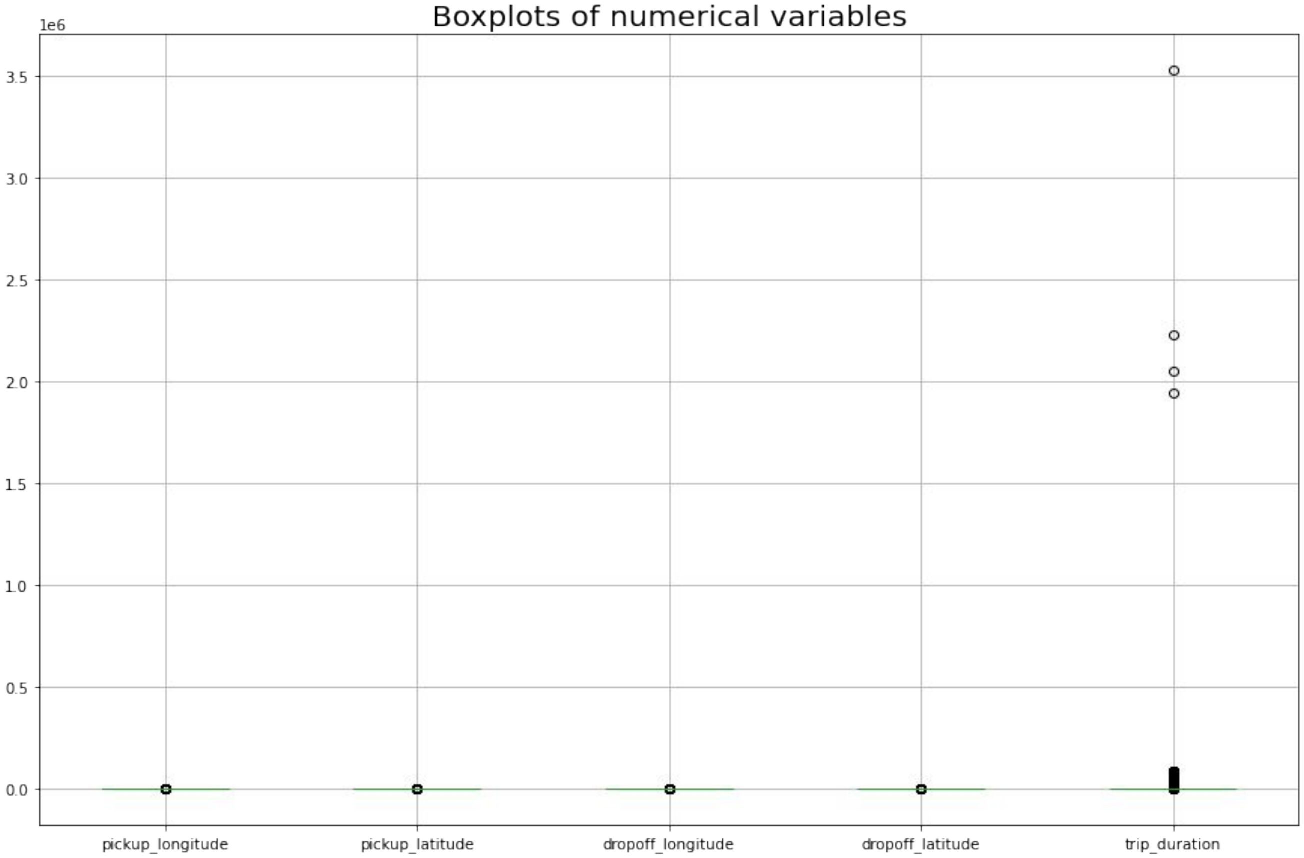
1. <u>Basic EDA :</u> Like any ML task, I shall first perform basic EDA to see what the data has in it to give me. Here, I shall perform some data cleaning tasks too.

2. <u>Feature Engineering :</u> In this step, I will try to come up with new features that might be better in describing and predicting trip durations. In this part, I shall also do some data cleaning too.

3. <u>Model Training and Testing :</u> In this section, I've tried to fit, train and test 4 regression models using a defined function. I've tried Linear Regression, Random Forest, Gradient Boosting Machine and Support Vector Machine.

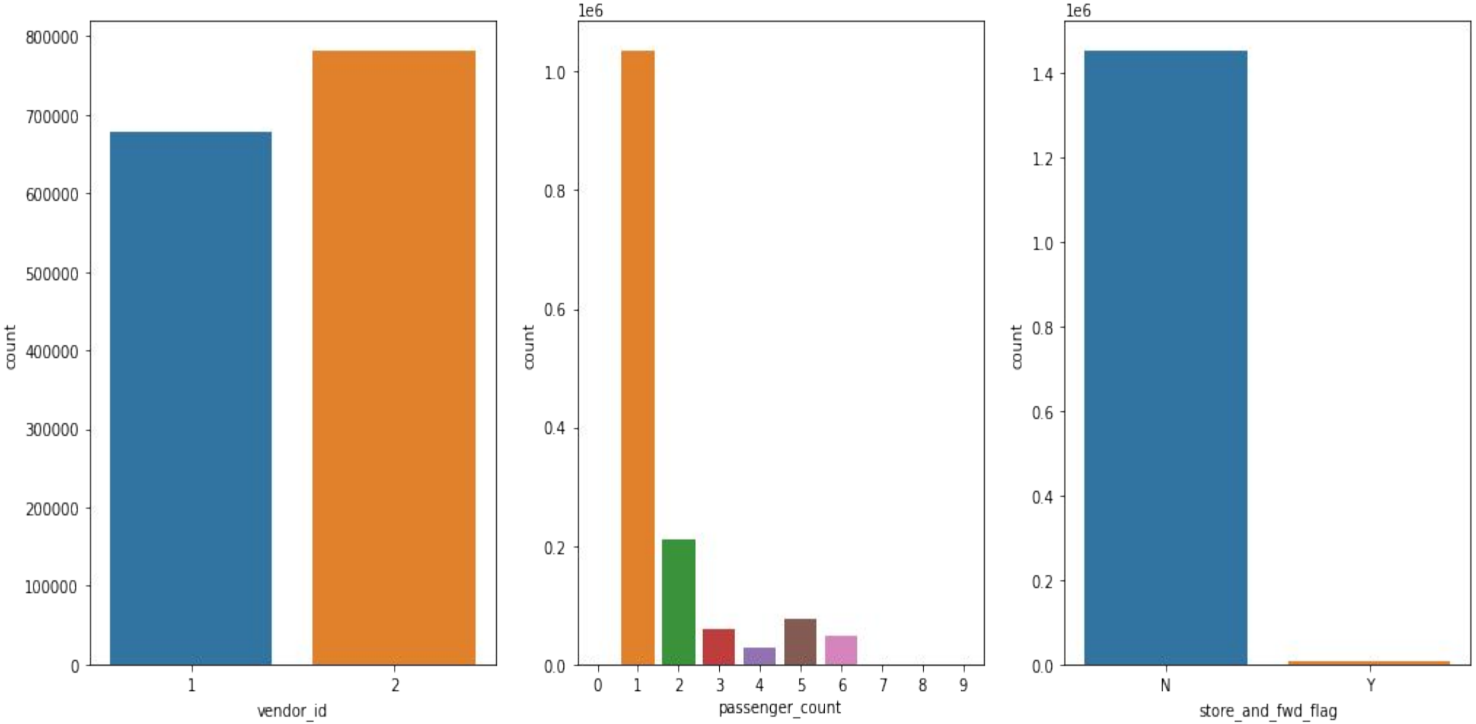
4. <u>Model Evaluation :</u> In this section, I've tried to score all the 4 models based on the training time and score and chosen an optimal model for this problem. I've also tried to find out the important features to train the model too.

Basic EDA - Are there any Outliers?



We can see that there are no visible and distant outliers in the dataset except for trip_duration which is our dependent variable.

Basic EDA - The Categoricals



There are a few conclusions to make here:

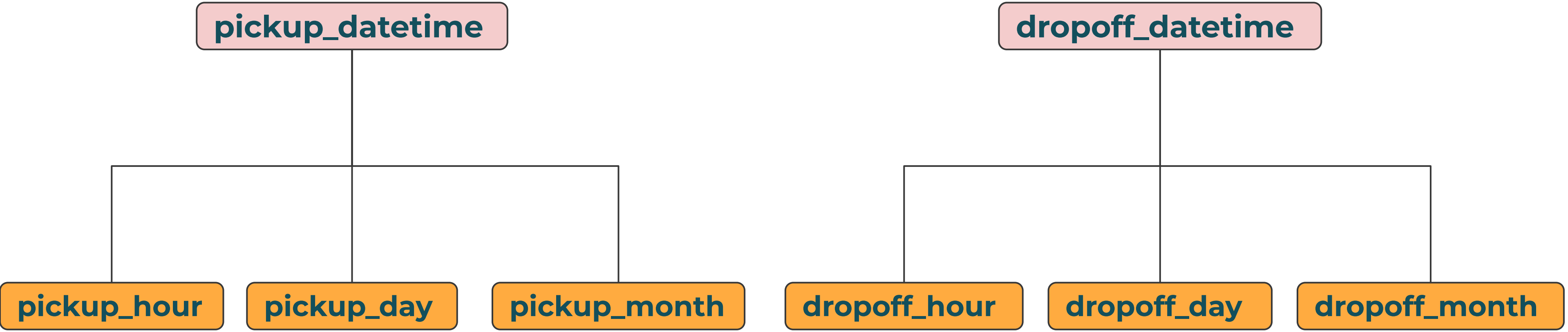
- Vendor id 2 gets most trips
- Passengers are more likely to travel solo.
- Taxis with more than 6 passengers are rare.
- There are some entries which have 0 passengers.
- Most of the trips were not held in vehicle memory.

Basic EDA - Data Handlings performed

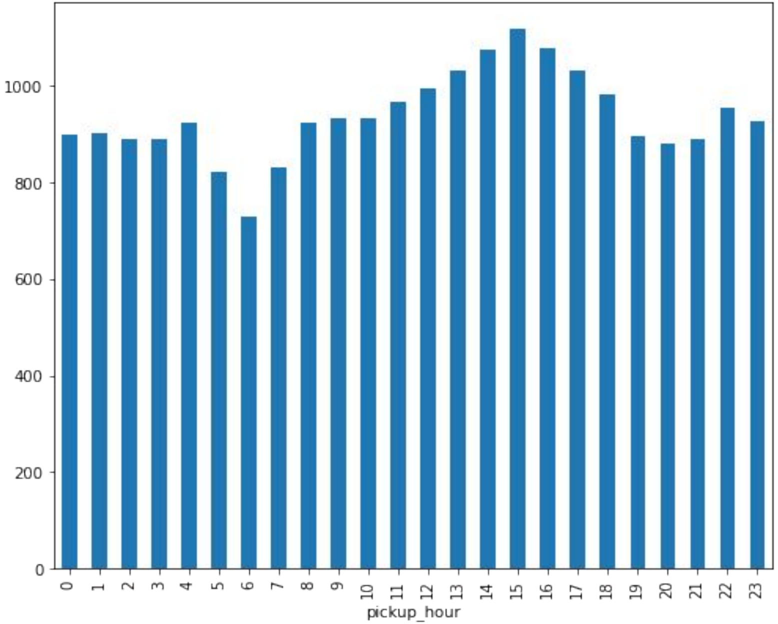
1. Removed 4 high values of trip_duration.
2. I’ve also found some very low trip durations (<60 seconds).
3. Removed entries with 0,7,8,9 passenger counts as they are minimal in numbers.

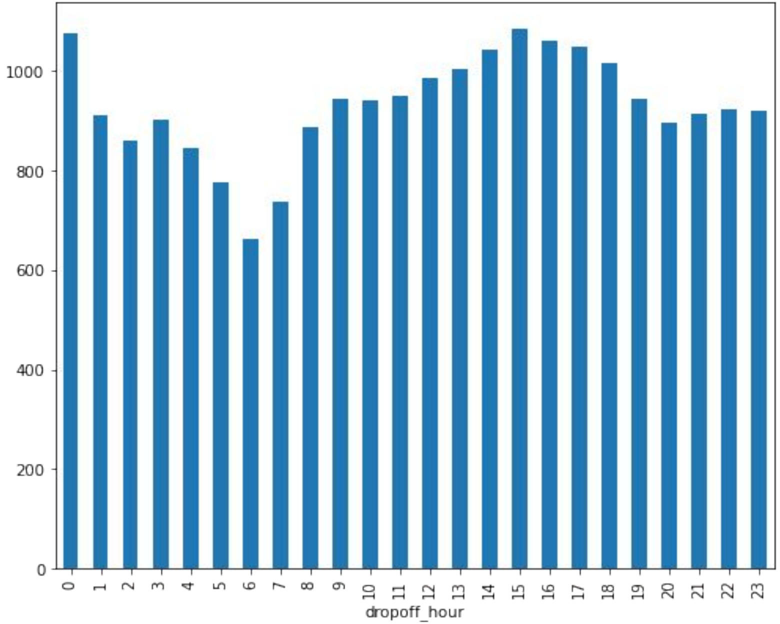
Feature Engineering - Pickup-Dropoff Times

The first feature breakdown I've performed is to get hours, day name and month from pickup and dropoff times.



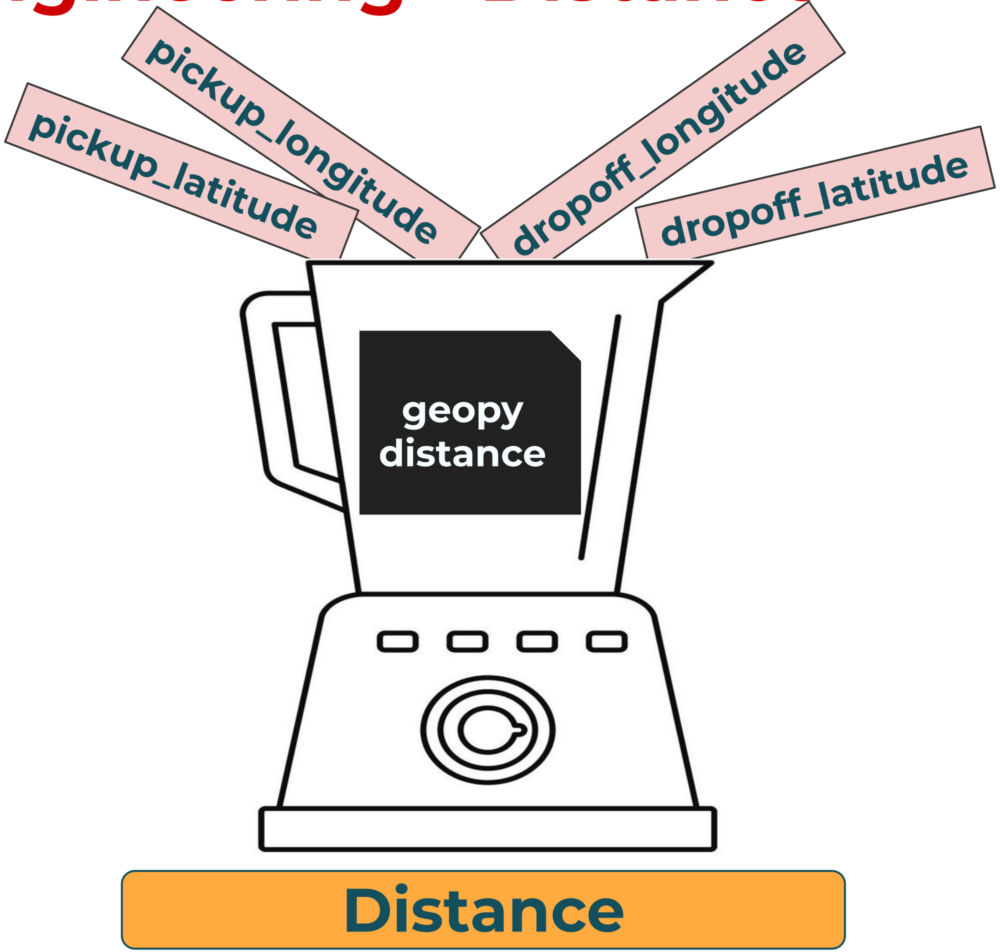
EDA on new features - Busy Times?



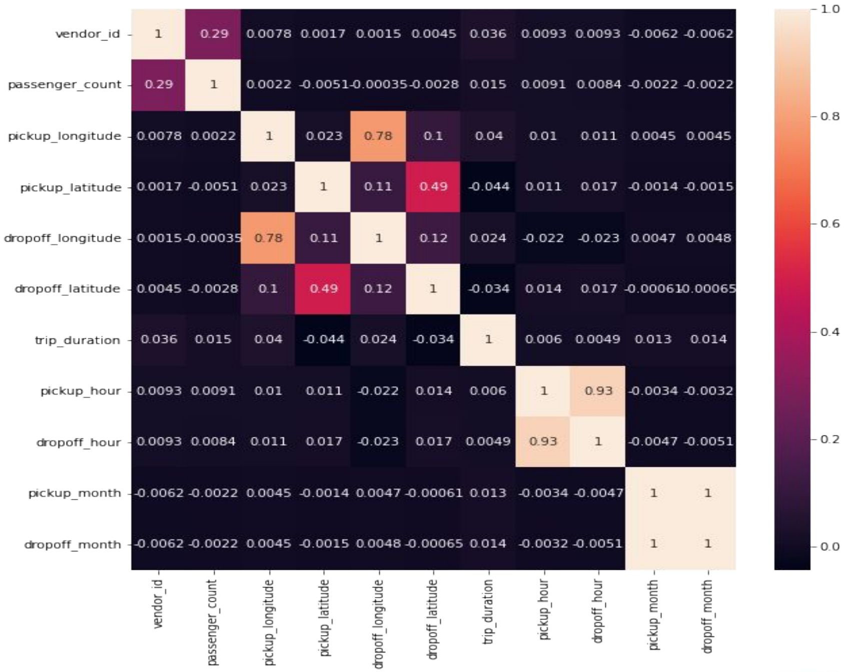


- **We can see that average trip durations are higher during 10AM-7PM. That's the rush hour in any city and it is obvious.**
- **Also, there is a little peak between 10PM-12AM.**

Feature Engineering - Distance



Feature Engineering - Speed



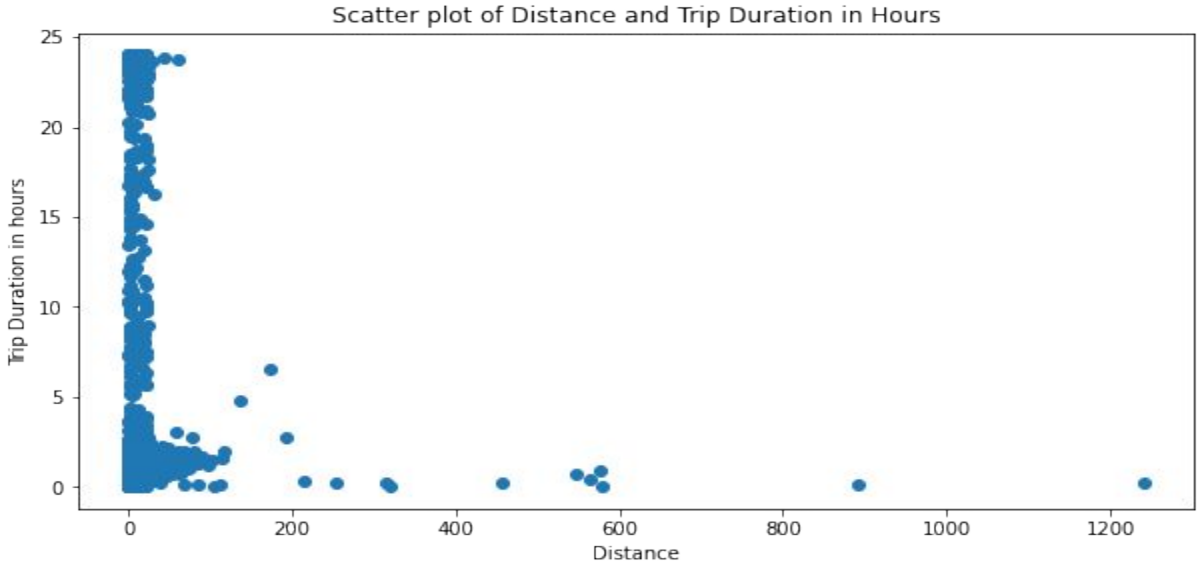
It is evident from the above heatmap that pickup and dropoff longitude & pickup and dropoff latitude are highly correlated. Here, we can combine them by calculating the distance between those points and introduce a new variable.

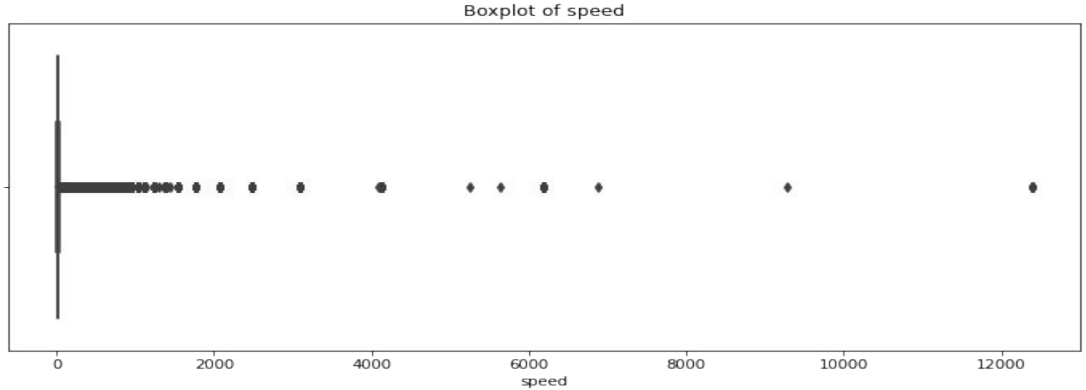
trip_duration_hour

distance

speed

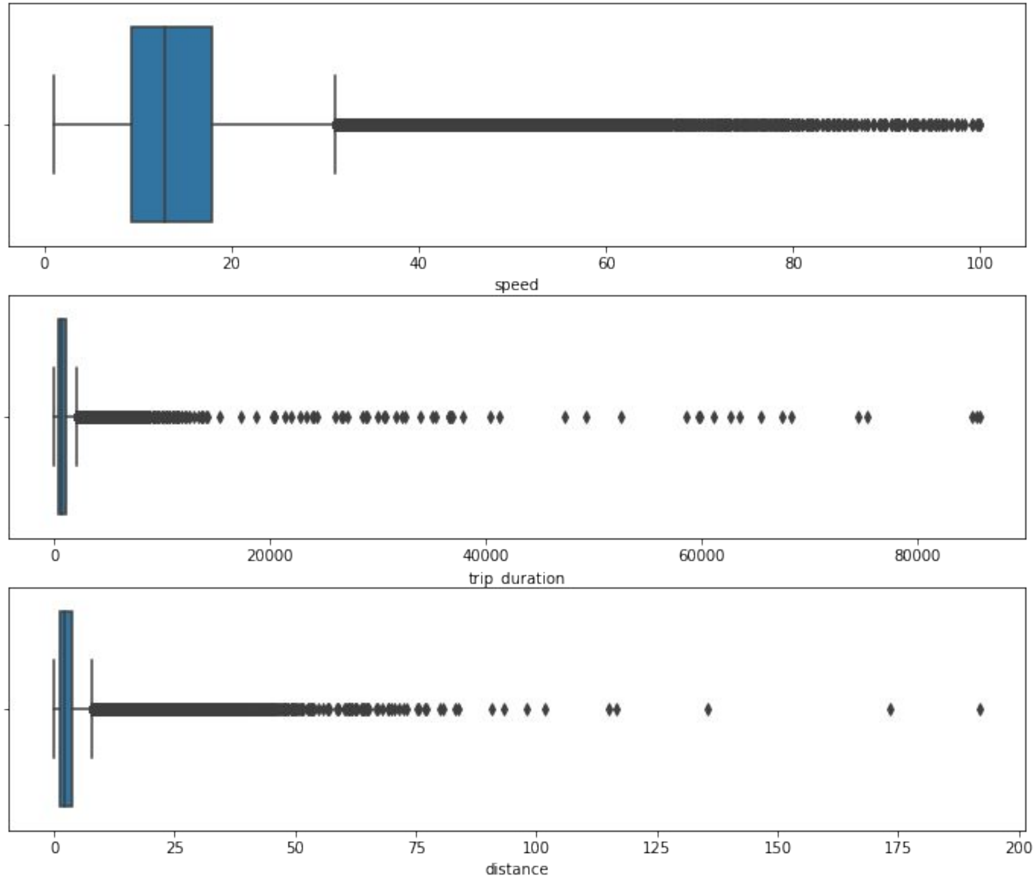
EDA on new features - Speed, Distance and Trip Duration





- We can see that there are some outliers through the scatter plot, and also there are some 0 distances. I am going to replace the 0 distances with mean distances.
- We can see that there are some speed which are even in thousands. We are not driving planes on roads, right? So, I'm taking an upper limit of 100 Kmph and lower limit of 1 Kmph for speed.

EDA on new features - Get out liers!



- I am going to remove the outliers on the basis of congestion of data points as we can see that although some points are outside the boxplot, they are highly congested.

Final Data Cleanings performed

1. Performed Isolation Forest algorithm to remove 1% of anomalies.
2. Removed trip durations greater than 10000 seconds and less than 60 seconds.
3. Taken speeds only between 1 kmph and 100 kmph.
4. Removed distances more than 60 KMs.
5. Removed store_fwd_flag variable.

Total Data Lost

2.01%

Model Training and Testing - Linear Regression

Linear Regression

With Log-transformation

- MSE (Train) : 18645529.75571323
- RMSE (Train) : 4318.046983963147
- R2 Score (Train) : -46.233058168301
- MSE (Test) : 10112093.980670217
- RMSE (Test) : 3179.9518833891525
- R2 Score (Test) : -25.00514705602341

Without Log-transformation

- MSE (Train) : 77458.49870903775
- RMSE (Train) : 278.3136696409965
- R2 Score (Train) : 0.8037812911144445
- MSE (Test) : 76163.73736172303
- RMSE (Test) : 275.9777841814863
- R2 Score (Test) : 0.8041306583765871

Model Training and Testing - Regularized Linear Regression

Ridge Regression

alpha : 5

MSE (Train) : 77455.50312723164

RMSE (Train) : 278.30828792407823

R2 Score (Train) : 0.8037888795547602

MSE (Test) : 76162.22781073867

RMSE (Test) : 275.97504925398357

R2 Score (Test) : 0.8041345404701883

Lasso Regression

alpha : 0.0001

MSE (Train) : 77456.75289377451

RMSE (Train) : 278.3105332066584

R2 Score (Train) : 0.8037857136326014

MSE (Test) : 76161.68948971755

RMSE (Test) : 275.97407394485003

R2 Score (Test) : 0.8041359248637021

ElasticNet Regression

alpha : 0.01, l1_ratio : 0.9

MSE (Train) : 77462.20443832896

RMSE (Train) : 278.3203270304362

R2 Score (Train) : 0.8037719037208197

MSE (Test) : 76159.22534710358

RMSE (Test) : 275.9696094628964

R2 Score (Test) : 0.8041422618687938

Model Training and Testing - Random Forest

Random Forest

max_depth : 8
min_samples_split : 1000
n_estimators : 60

Training Set

MSE :
5158.998080607735

RMSE :
71.8261657100512

R2 Score :
0.986931170118305

Testing Set

MSE :
5231.406183289991

RMSE :
72.32846039623676

R2 Score :
0.9865464574037484

Model Training and Testing - Gradient Boosting Machine

Gradient Boosting Machine

max_depth : 8
min_samples_split : 1000
n_estimators : 60

Training Set

MSE : 93.18217322573778

RMSE : 9.653091381818458

R2 Score : 0.9997639499083221

Testing Set

MSE : 122.35652996423335

RMSE : 11.061488596216757

R2 Score : 0.9996853372248056

Model Training and Testing - Support Vector Machine

Support Vector Machine

C : 10000

Training Set

- MSE : 178145.77411115135
- RMSE : 422.07318572867354
- R2 Score : 0.5487191932183773

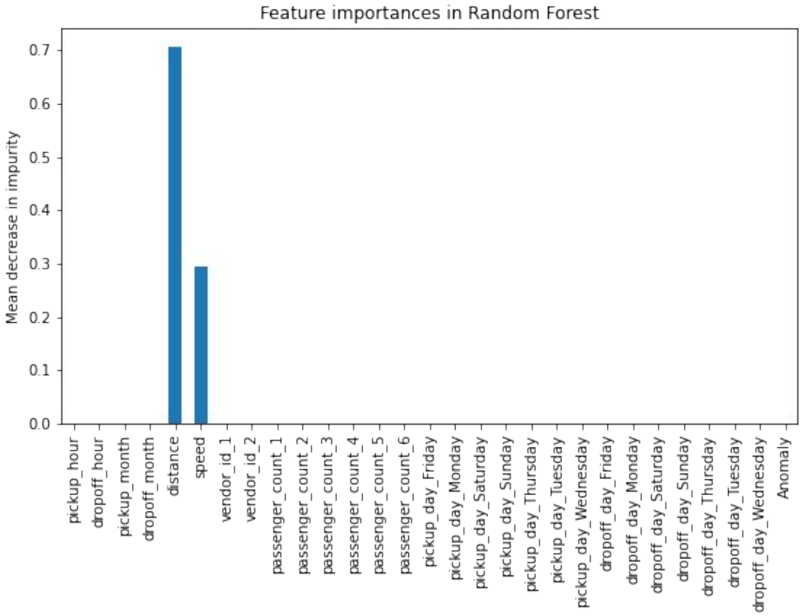
Testing Set

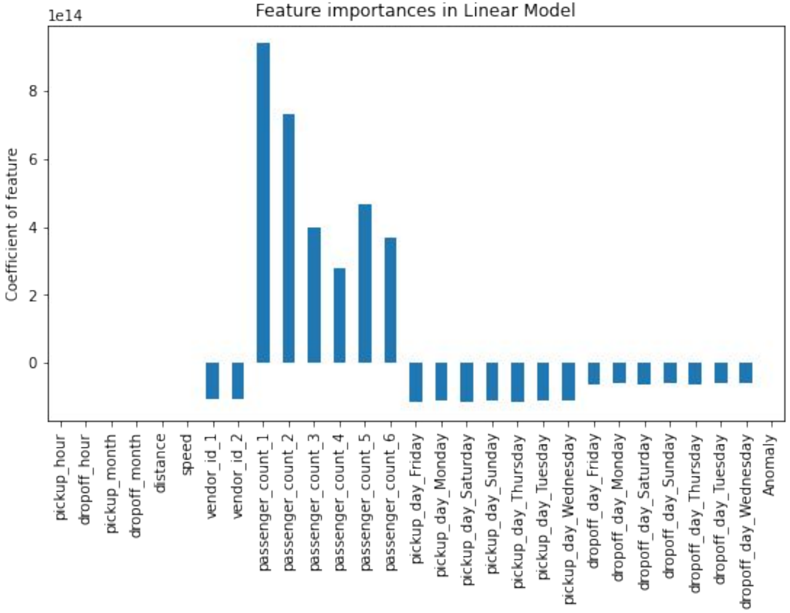
- MSE : 177285.86220306592
- RMSE : 421.05327715511953
- R2 Score : 0.5440761402774201

Model Evaluation - Best One?

Model Name	Performance Score	Speed Score	Final Score
Multiple Linear Regression Model	4	1	41
L1 Regularized(Lasso) Linear Regression Model	3	2	32
L2 Regularized(Ridge) Linear Regression Model	3	1	31
ElasticNet Regularized Linear Regression Model	3	2	32
Random Forest Regressor Model	2	4	24
Gradient Boosting Machine Regressor Model	1	5	15
Support Vector Machine Regressor Model	5	3	53

Model Evaluation - Which Features?





- **Now, this is an interesting picture. In Random Forest, distance and speed are the main features that are being used in estimating trip duration. But in the case of Linear Regression, almost all the other variables have an impact on estimating trip duration except for distance and speed. This might be the reason that Linear Models were so poor performance.**
- **But when I only took speed and distance in Linear Regression model, the model gave similar accuracy as it gave with all variables together.**

Final Verdicts

1. Important Variables :

When Random Forest used only speed and distance, it gave very high accuracy. But when Linear Regression used other variables except for speed and distance, the model couldn't get to a high accuracy.

2. Best Model :

Gradient Boosting Machine is the best choice here. If anyone has the resources to consume that much time, the model will predict trip durations with 99% accuracy.

3. Challenges faced :

I am listing some challenges faced by me :

- Huge data size.
- Getting new features which can predict trip duration more accurately.
- Too much training time for black box models.

4. Use cases :

With so much high accuracy across both train and test set, this model can be used for any intra-city journeys. But beware! As this model doesn't take account for long distance journeys, it might not be too accurate to predict inter-city trip durations. There might be cases when a cab might take a highway. Then that highway might be a variable that should be accounted for in predicting the trip duration. The high trip durations can be predicted by some other models and more data.