Capstone Project - 2

NYC Taxi Trip Time Prediction

Project By :
Sanjay Yadav

CONTENT :

- Problem Statement
- Data Summary
- EDA
- Feature Engineering and Feature Selection
- Regression Models used
- Challenges
- Conclusion

Problem Statement :

New York City taxi rides form the core of the traffic in the city of New York. The many rides taken every day by New Yorkers in the busy city can give us a great idea of traffic times, road blockages, and so on.

Predicting the duration of a taxi trip is very important since a user would always like to know precisely how much time it would require him to travel from one place to another.

Our task is to build a model that predicts the total ride duration of taxi trips in New York City. The dataset is based on the 2016 NYC Yellow Cab trip record data made available in Big Query on Google Cloud Platform.

Data Summary :

NYC Taxi Data.csv - the training set
(contains 1458644 trip records)

Independent features :
id , vendor_id , pickup_datetime , dropoff_datetime , passenger_count , pickup_longitude , pickup_latitude , dropoff_longitude , dropoff_latitude , store_and_fwd_flag

Target Variable :
trip_duration

Exploratory Data Analysis

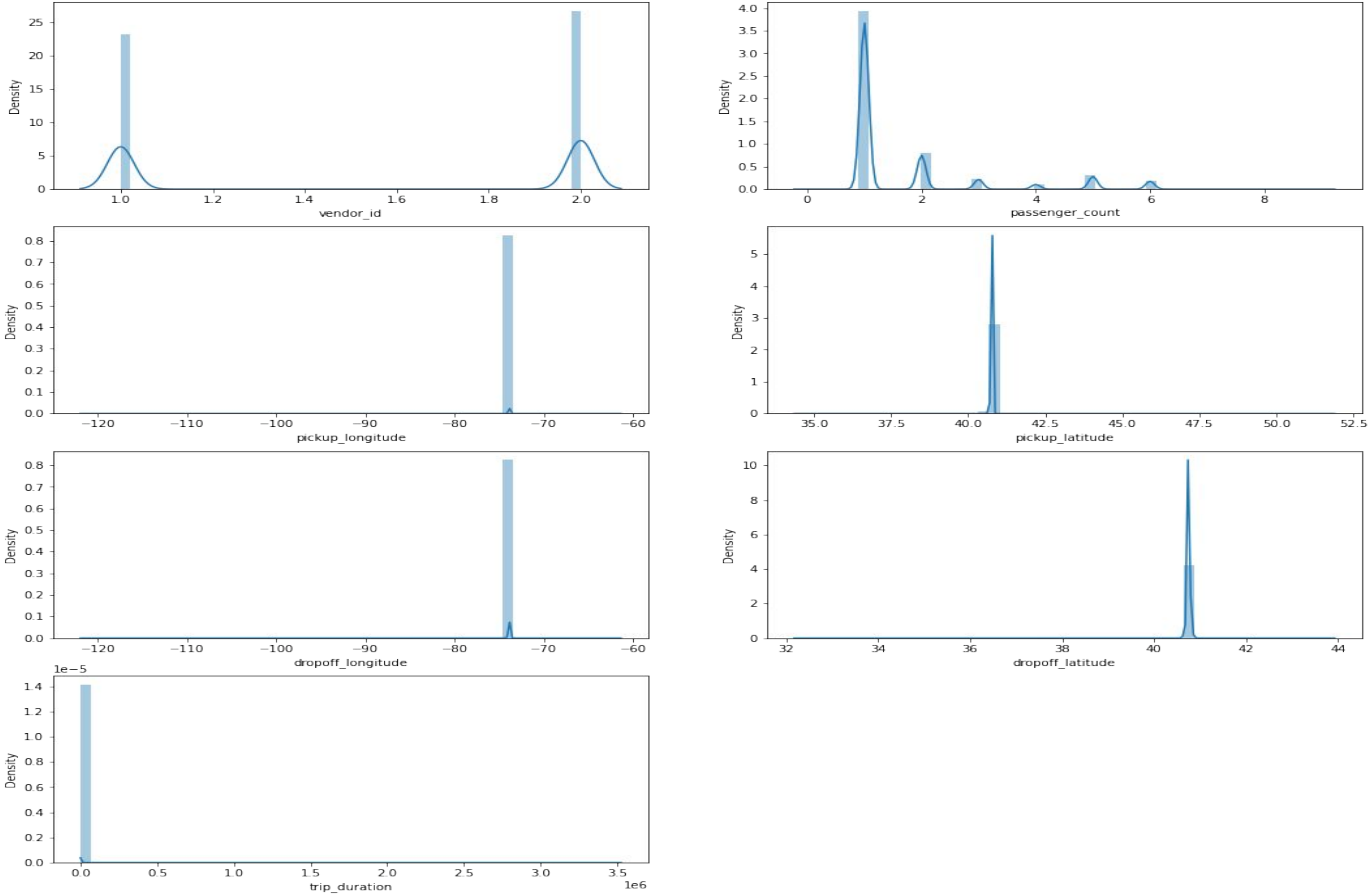
Missing and Duplicate Values:

```
# Checking null values
dataset.isnull().sum()
```

```
id                    0
vendor_id             0
pickup_datetime       0
dropoff_datetime      0
passenger_count       0
pickup_longitude      0
pickup_latitude       0
dropoff_longitude     0
dropoff_latitude      0
store_and_fwd_flag    0
trip_duration         0
dtype: int64
```

```
# Checking duplicated values
dataset.duplicated().sum()
```

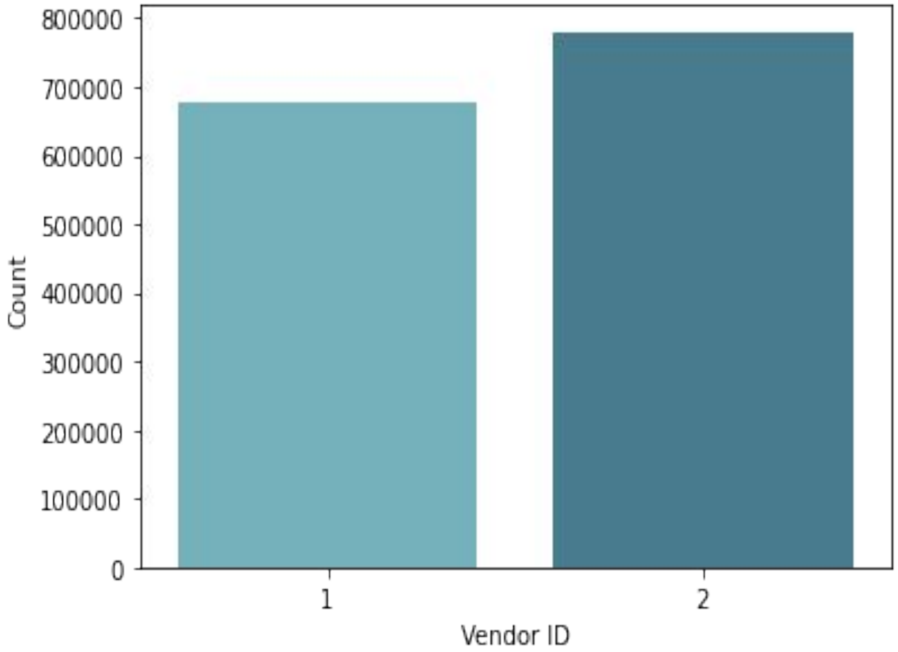
```
0
```



Density
vendor_id
passenger_count
pickup_longitude
pickup_latitude
dropoff_longitude
dropoff_latitude
1e−5
trip_duration
1e6

Vendor ID :

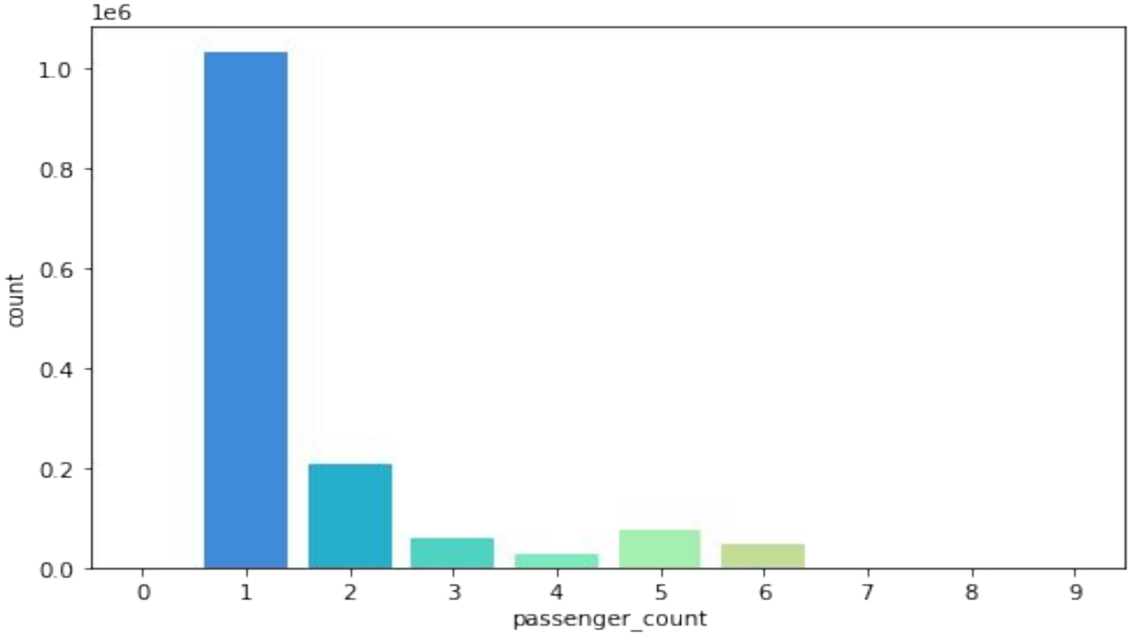
Both the vendors seems to have almost equal market share. But Vendor 2 is evidently more famous among the population as per the graph.



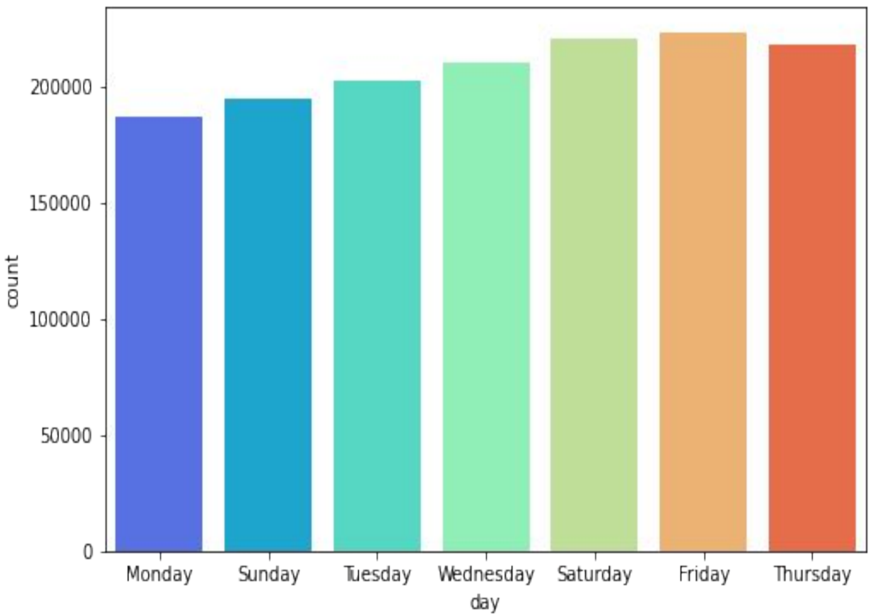
Passenger count :

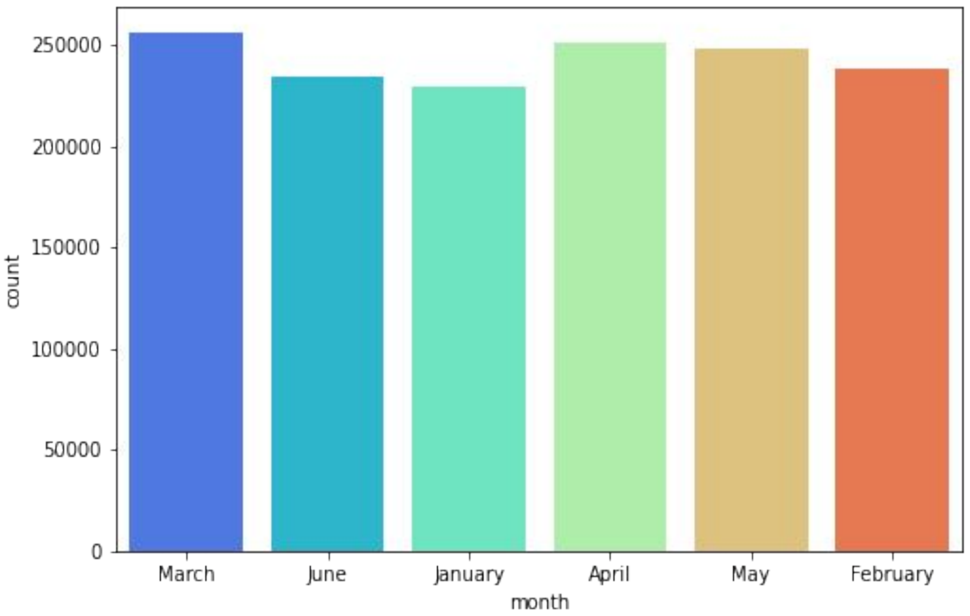
The passenger_count variable has a minimum value of 0 passengers. These observations are most likely errors and will need to be removed from the dataset.

According to the NYC Taxi & Limousine Commission, the maximum number of people allowed in a yellow taxicab, by law, is 5 passengers and one child .The observations more than 6 are likely an error and will also need to be removed from the dataset.

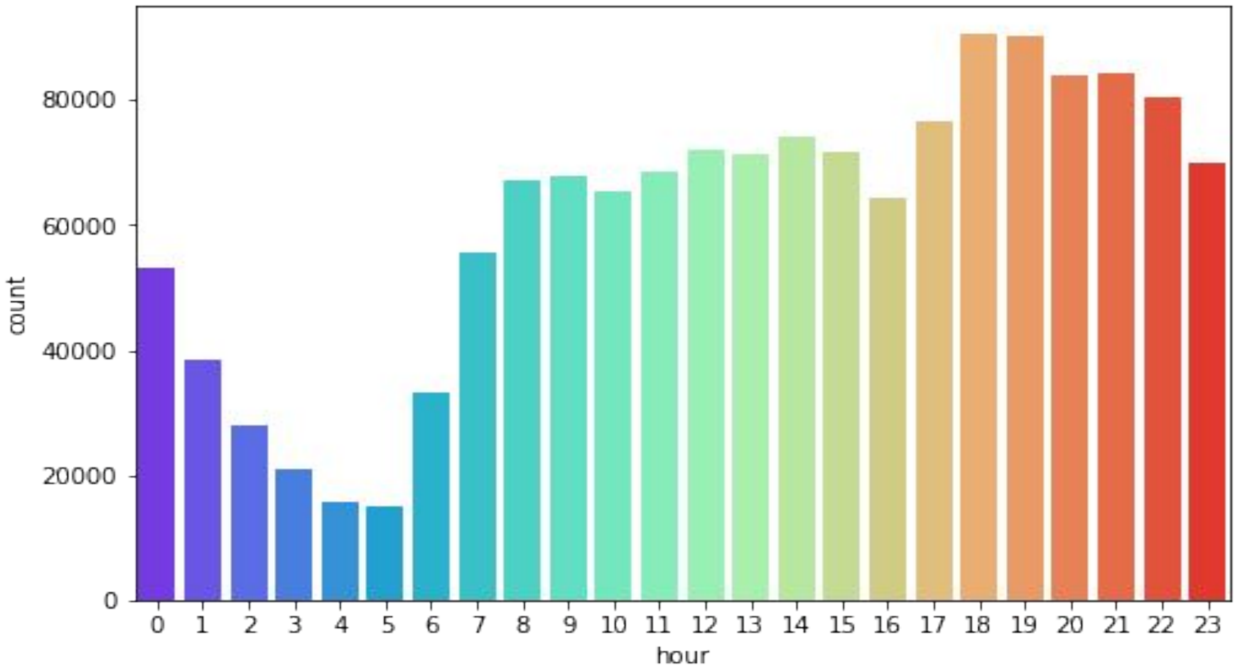


New features by decomposing datetime





Distribution of trips throughout the day

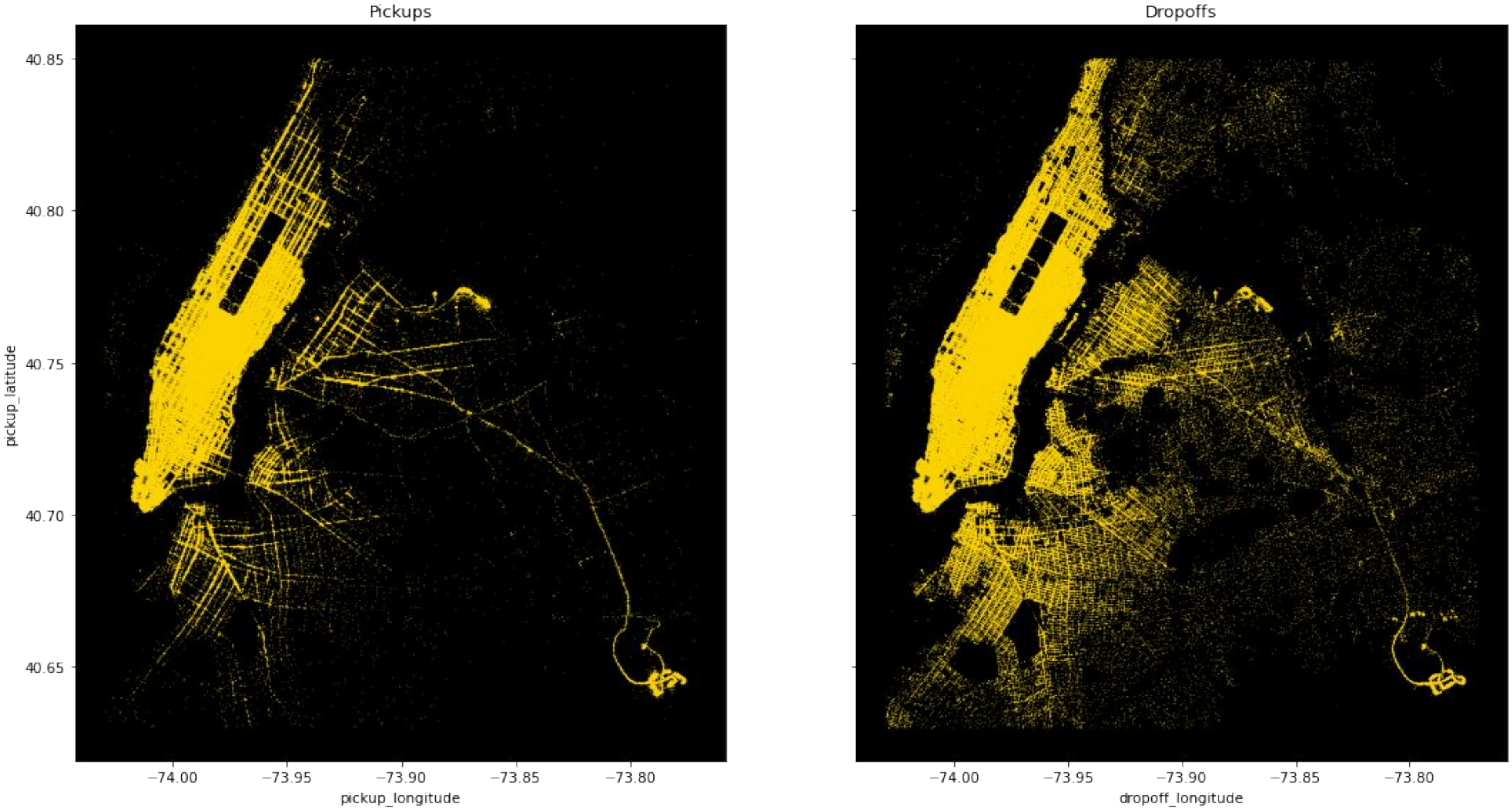


Longitude and Latitude

Looking into it, the borders of NY City coordinates comes out to be:

longitude = (-74.03, -73.77) , latitude = (40.63, 40.85)

Any coordinates outside will be outliers.


Pickups
Dropoffs
40.85
40.80
40.75
40.70
40.65
pickup_latitude
−74.00
−73.95
−73.90
−73.85
−73.80
pickup_longitude
dropoff_longitude

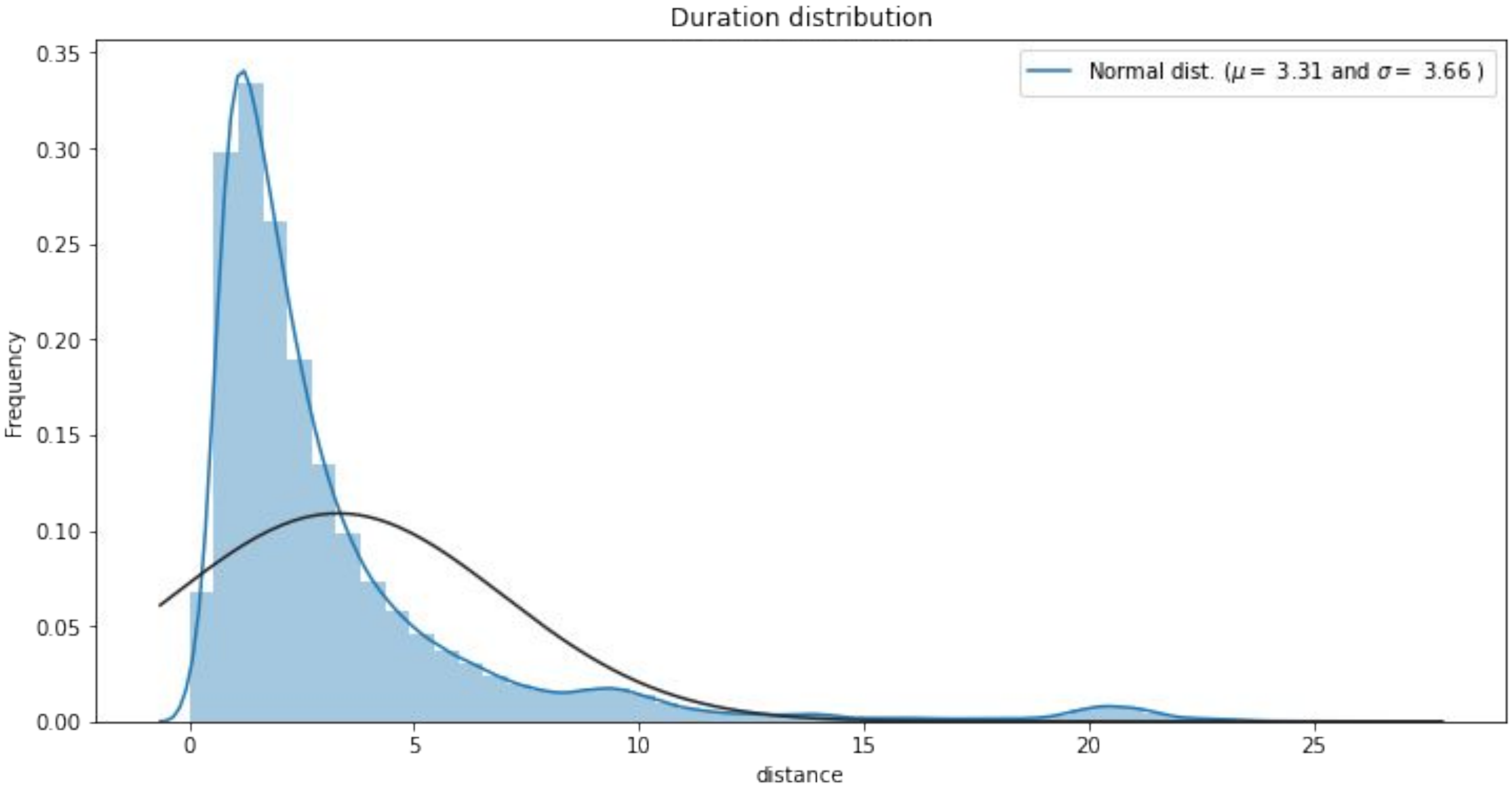
Haversine Distance

$$D = 2rsin^{-1}\left(\sqrt{sin^2\left(\frac{\varphi_2-\varphi_2}{2}\right) + cos(\varphi_1)cos(\varphi_2)sin^2\left(\frac{\lambda_2-\lambda_2}{2}\right)}\right)$$

```
df.distance.describe()
```

```
count    1.438573e+06
mean     3.292866e+00
std      3.662317e+00
min      0.000000e+00
25%      1.224953e+00
50%      2.068546e+00
75%      3.767414e+00
max      2.720017e+01
Name: distance, dtype: float64
```

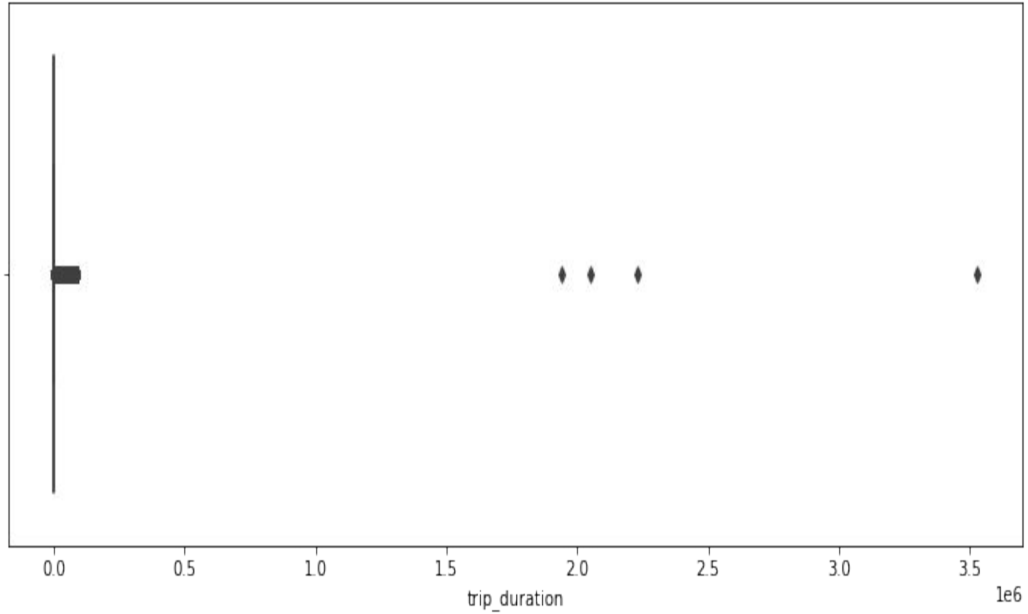
Distribution of distance

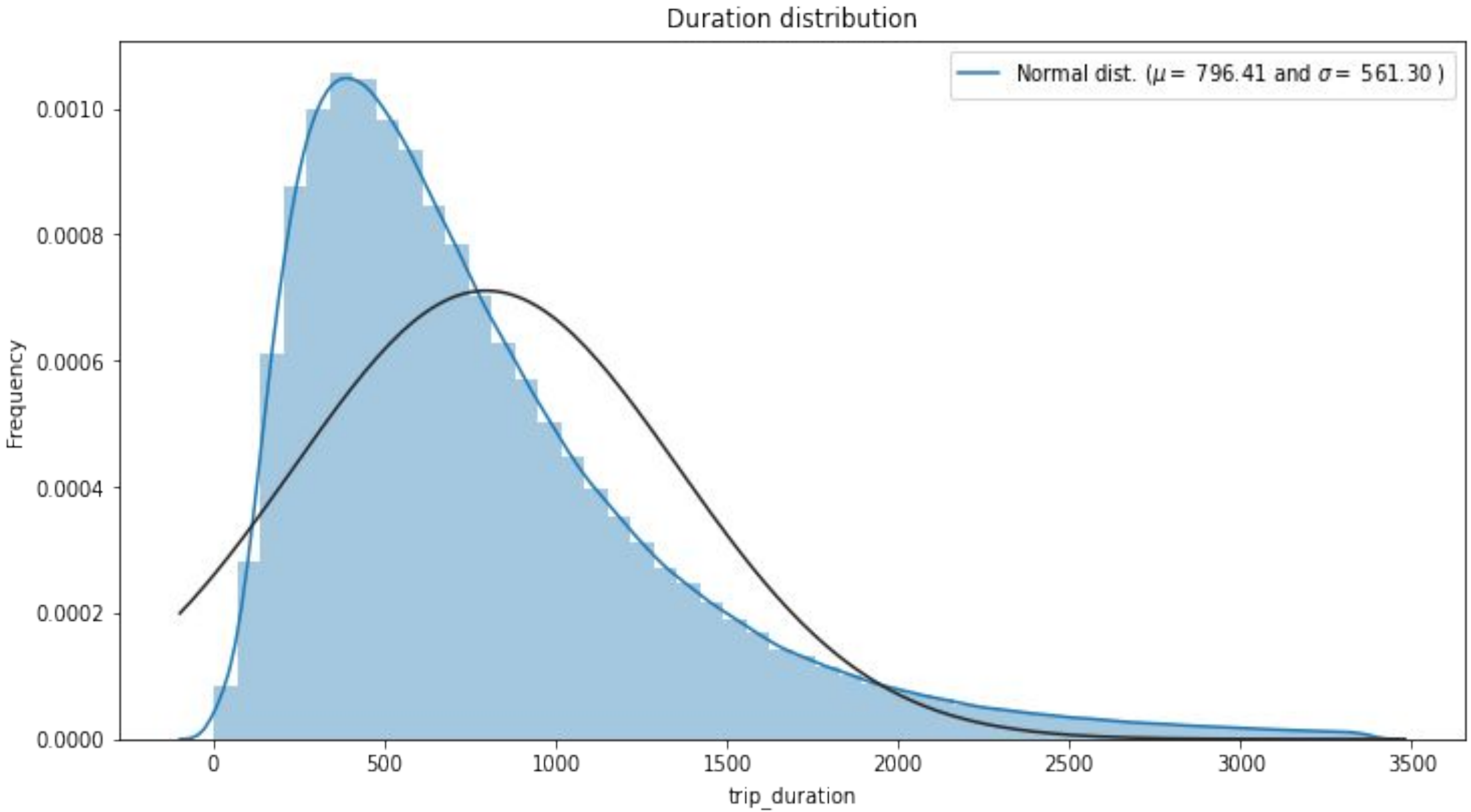


Distribution of trip duration

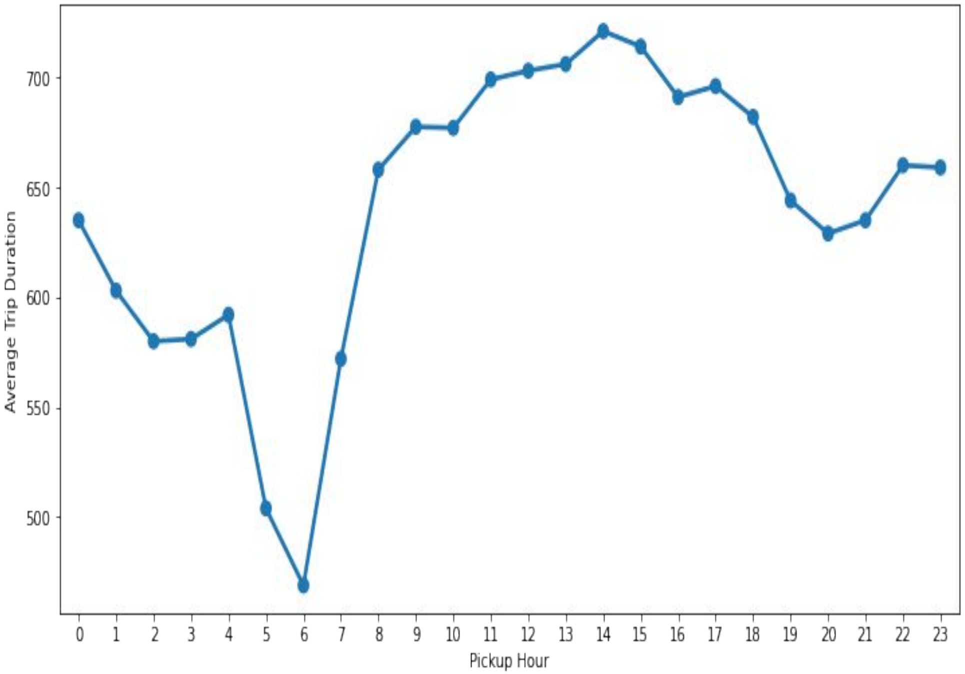
```
df.trip_duration.describe()
```

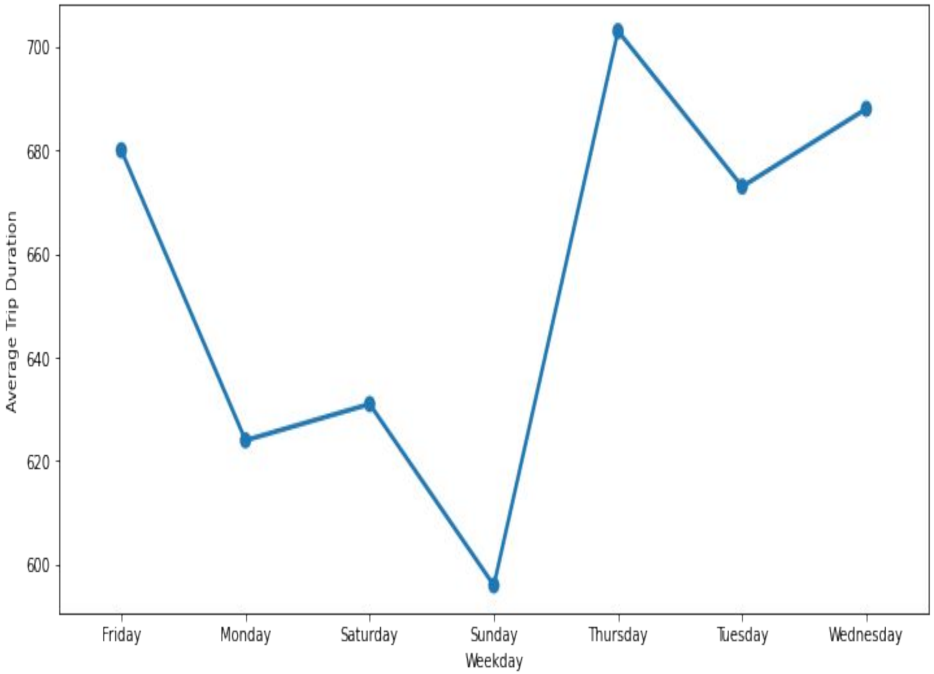
```
count    1.433206e+06
mean     9.474388e+02
std      5.260065e+03
min      1.000000e+00
25%      3.950000e+02
50%      6.570000e+02
75%      1.060000e+03
max      3.526282e+06
Name: trip_duration, dtype: float64
```



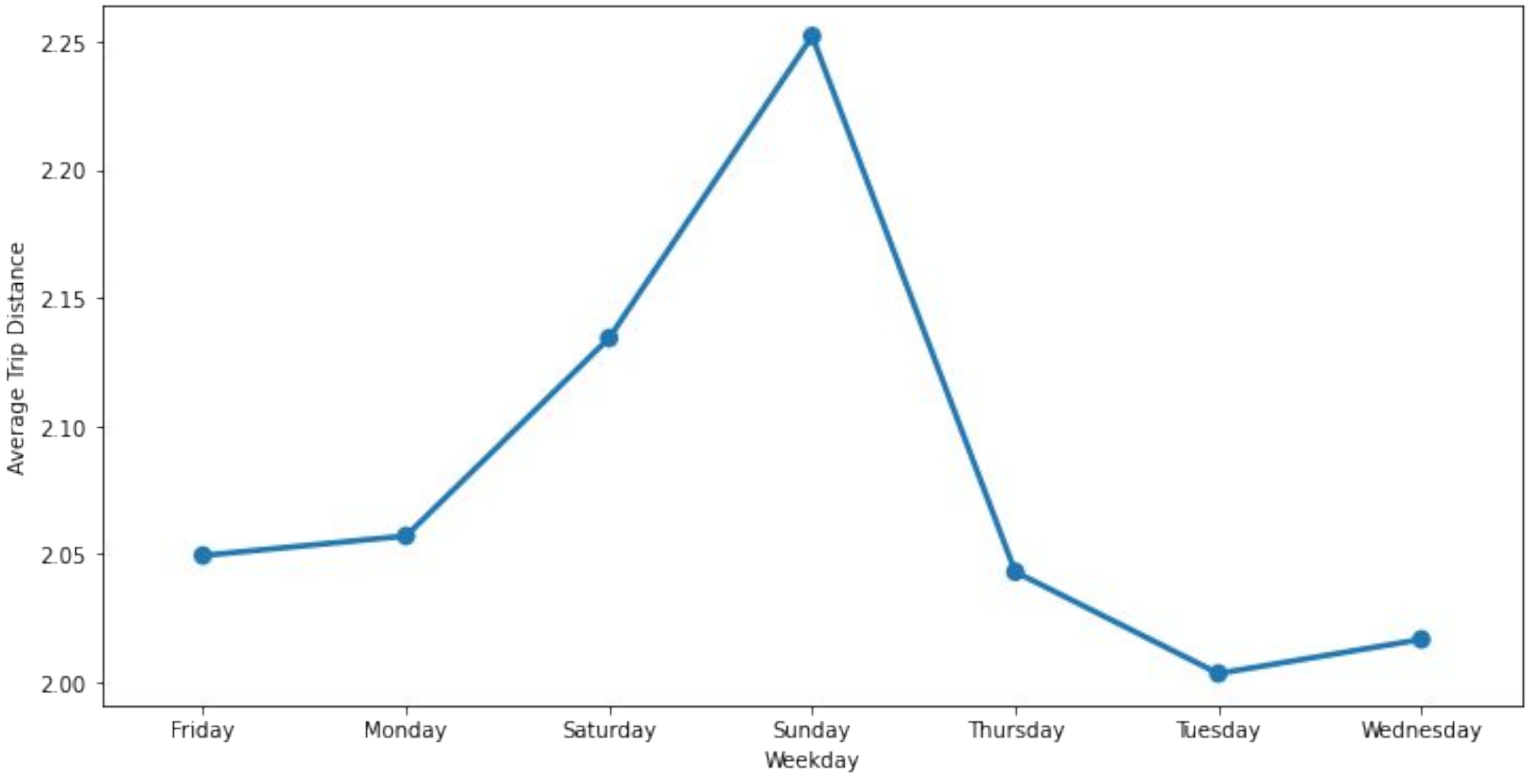

Duration distribution
Normal dist. (μ = 796.41 and σ = 561.30)
Frequency
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
0
500
1000
1500
2000
2500
3000
3500
trip_duration

Average trip duration on hourly and weekly basis

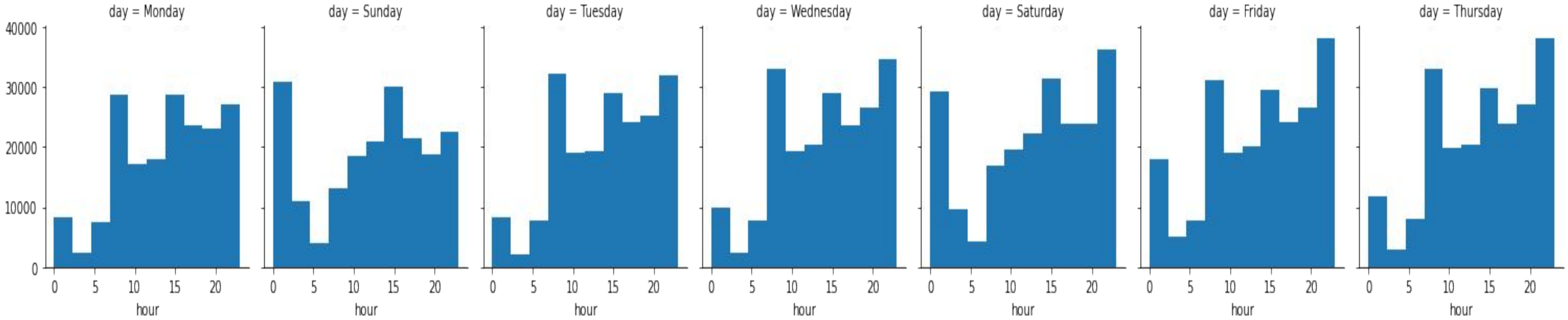


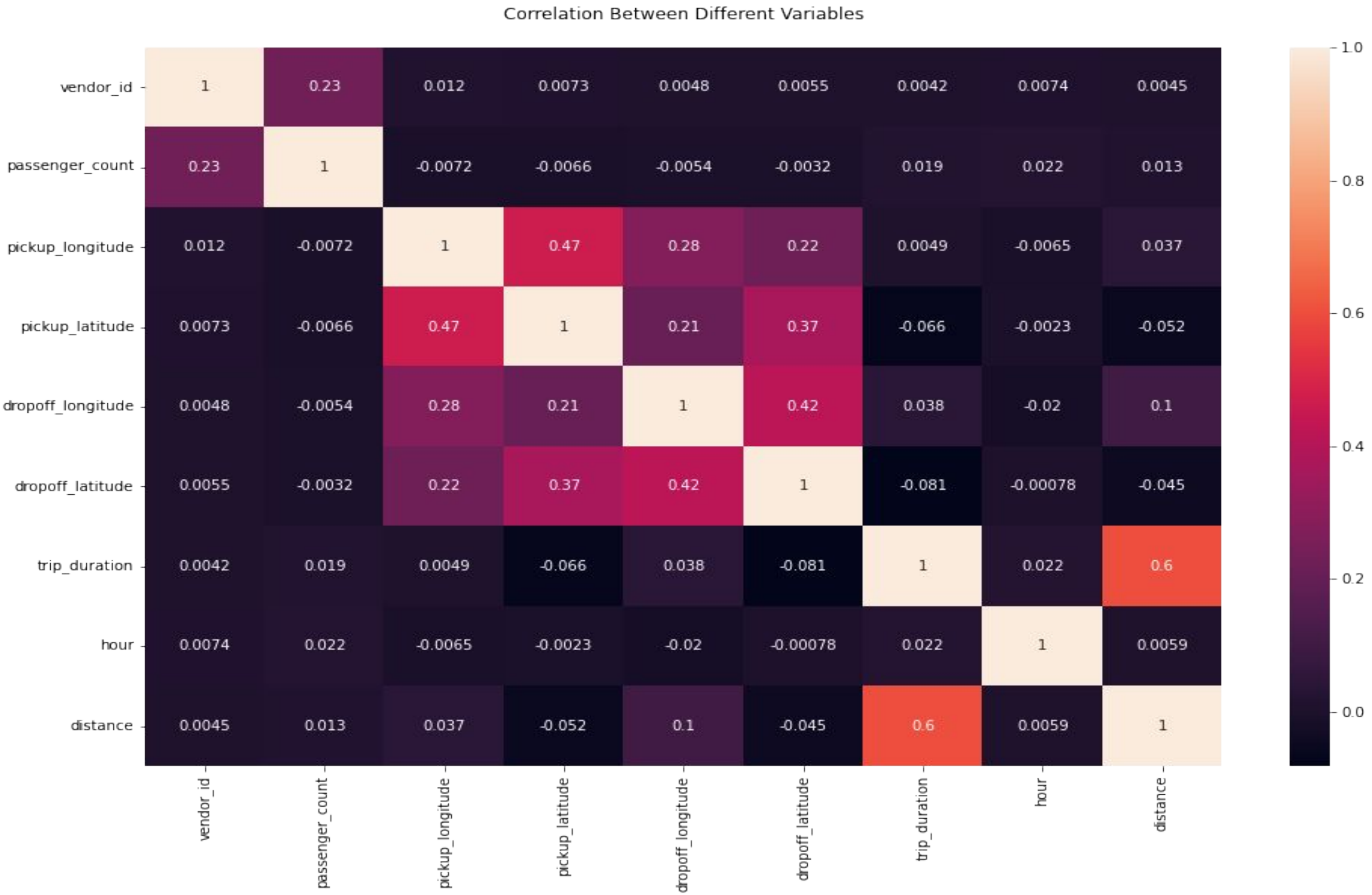


Average trip distance vs. Weekday



Hourly trip distribution across the week





Correlation Between Different Variables
vendor_id
passenger_count
pickup_longitude
pickup_latitude
dropoff_longitude
dropoff_latitude
trip_duration
hour
distance
1 0.23 0.012 0.0073 0.0048 0.0055 0.0042 0.0074 0.0045
0.23 1 -0.0072 -0.0066 -0.0054 -0.0032 0.019 0.022 0.013
0.012 -0.0072 1 0.47 0.28 0.22 0.0049 -0.0065 0.037
0.0073 -0.0066 0.47 1 0.21 0.37 -0.066 -0.0023 -0.052
0.0048 -0.0054 0.28 0.21 1 0.42 0.038 -0.02 0.1
0.0055 -0.0032 0.22 0.37 0.42 1 -0.081 -0.00078 -0.045
0.0042 0.019 0.0049 -0.066 0.038 -0.081 1 0.022 0.6
0.0074 0.022 -0.0065 -0.0023 -0.02 -0.00078 0.022 1 0.0059
0.0045 0.013 0.037 -0.052 0.1 -0.045 0.6 0.0059 1
1.0
0.8
0.6
0.4
0.2
0.0

Analysis using stats OLS model

OLS Regression Results

Dep. Variable:	trip_duration	**R-squared:**	0.624
Model:	OLS	**Adj. R-squared:**	0.624
Method:	Least Squares	**F-statistic:**	5.235e+04
Date:	Fri, 18 Mar 2022	**Prob (F-statistic):**	0.00
Time:	15:20:27	**Log-Likelihood:**	-1.0301e+07
No. Observations:	1418859	**AIC:**	2.060e+07
Df Residuals:	1418813	**BIC:**	2.060e+07
Df Model:	45		
Covariance Type:	nonrobust		

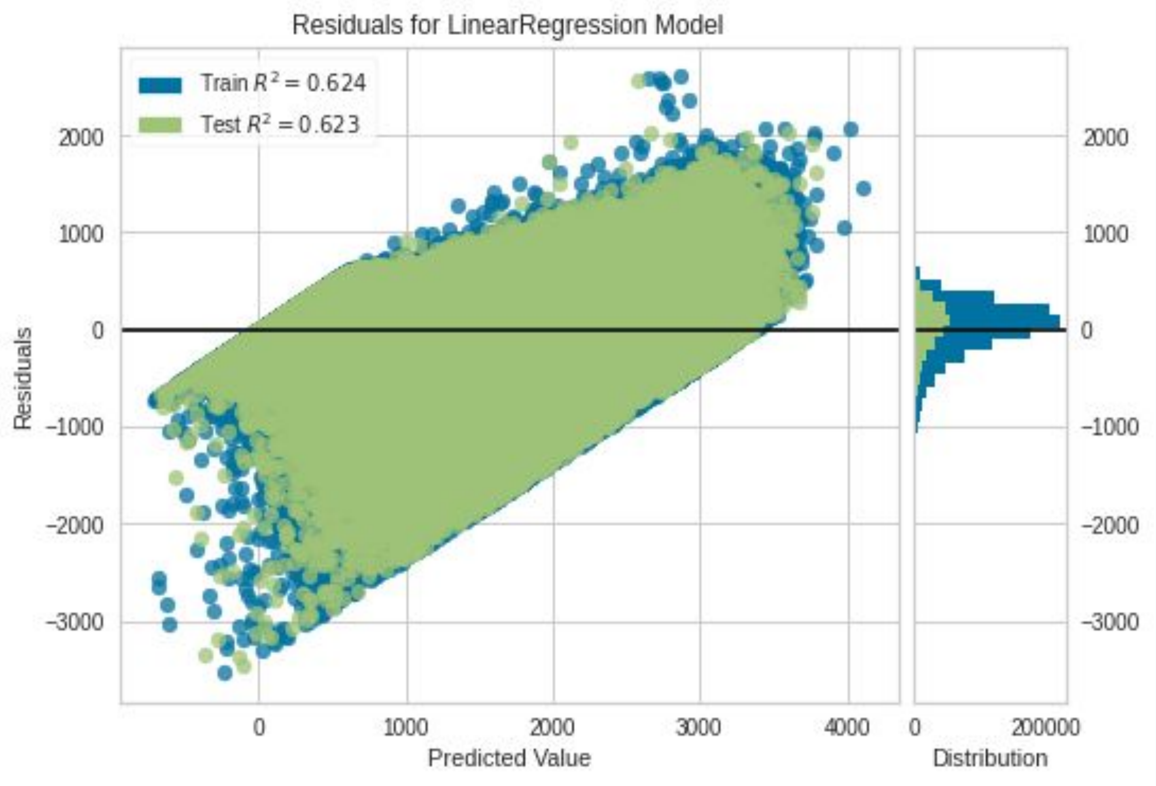
Linear Regression

Train metrics :

- ❑ MSE : 118225.054
- ❑ RMSE : 343.8387
- ❑ R2 : 0.6244
- ❑ Adjusted R2 : 0.6244

Test metrics :

- ❑ MSE : 119221.0466
- ❑ RMSE : 345.284
- ❑ R2 : 0.6230
- ❑ Adjusted R2 : 0.62295



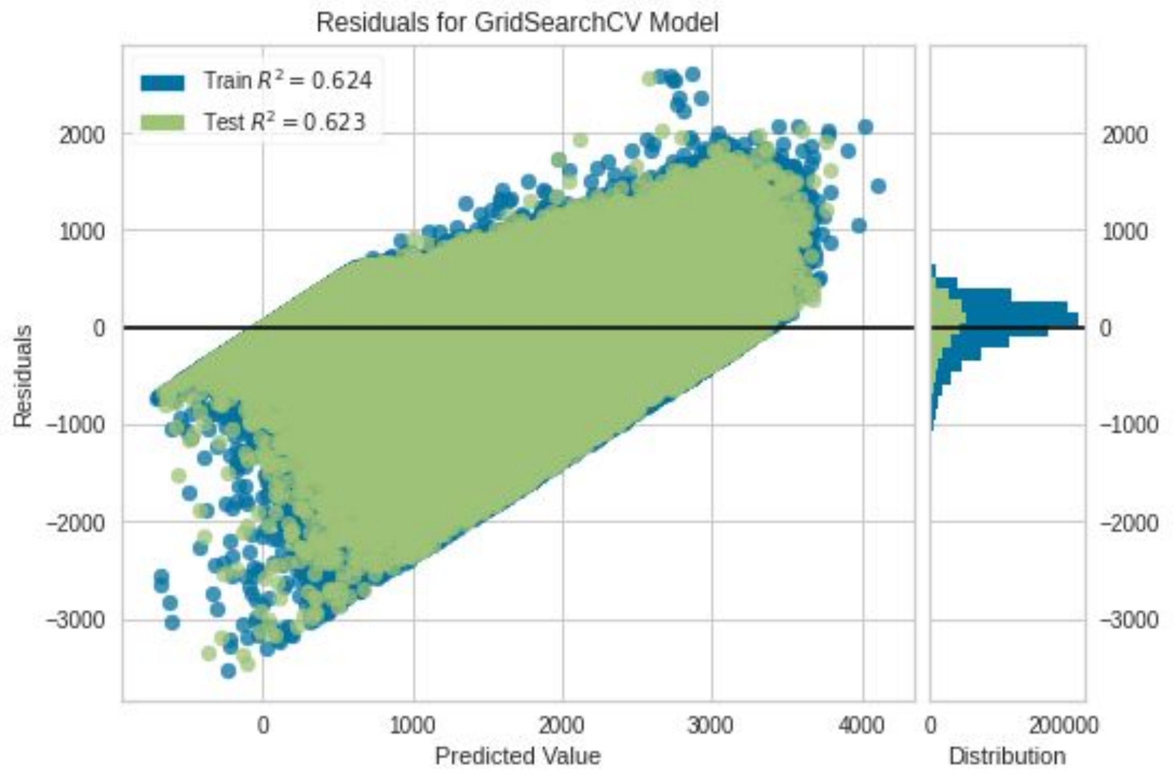
Ridge Regression

Train metrics :

- MSE : 118225.0541
- RMSE : 343.8387
- R2 : 0.6244
- Adjusted R2 : 0.62439

Test metrics :

- MSE : 119221.0552
- RMSE : 343.8387
- R2 : 0.6230
- Adjusted R2 : 0.62295



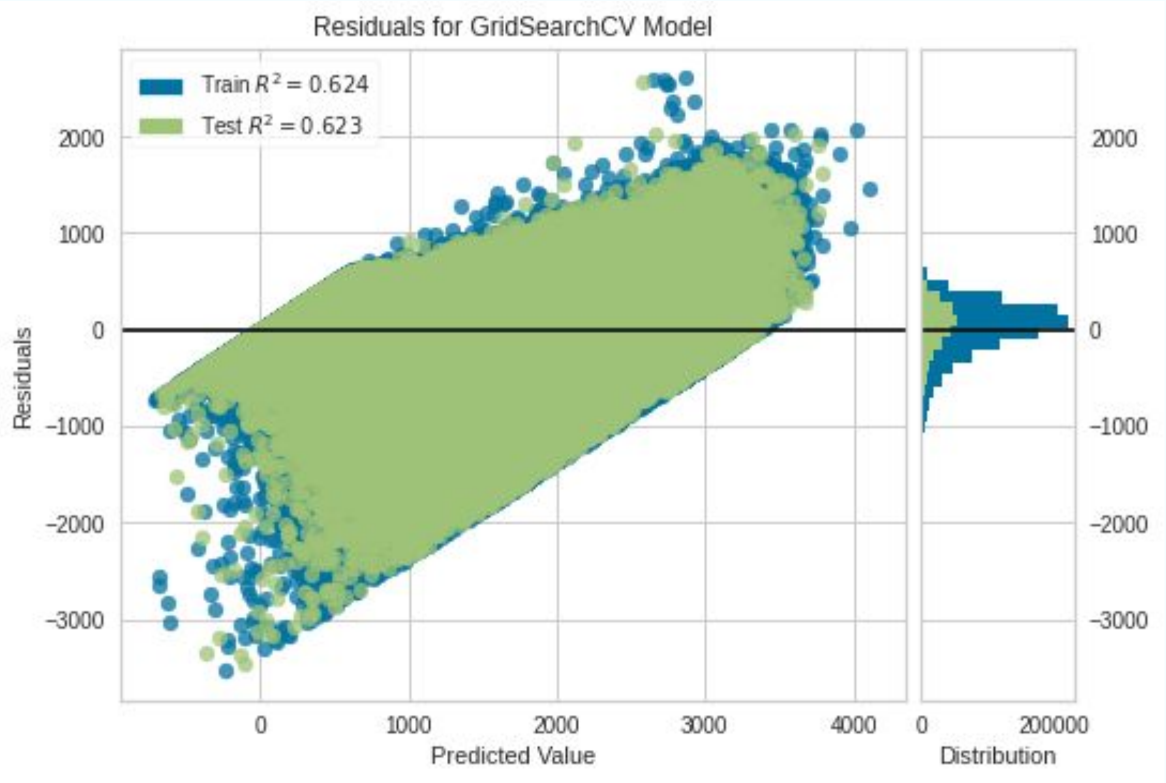
Lasso Regression

Train metrics :

- MSE : 118225.0540
- RMSE : 343.8387
- R2 : 0.624411
- Adjusted R2 : 0.62439

Test metrics :

- MSE : 119221.04665
- RMSE : 345.2840
- R2 : 0.6230
- Adjusted R2 : 0.62295



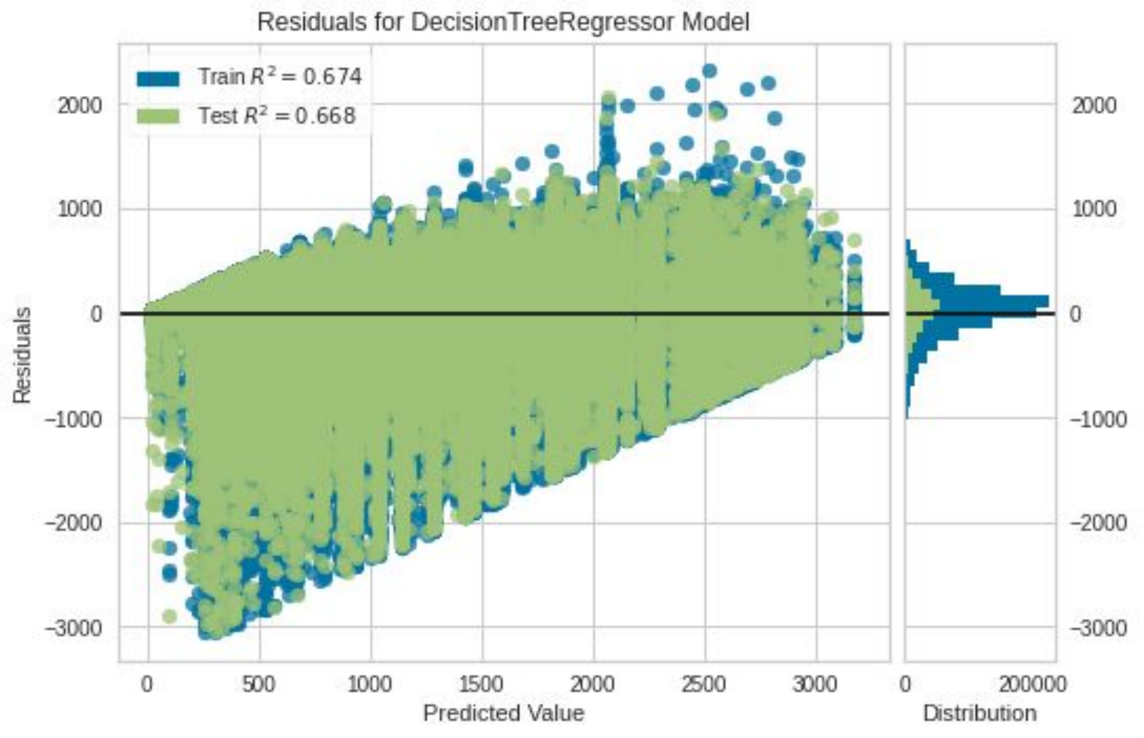
Decision Tree Regression

Train metrics :

- ❑ **MSE : 102637.2283**
- ❑ **RMSE : 320.37045**
- ❑ **R2 : 0.6739**
- ❑ **Adjusted R2 : 0.67391**

Test metrics :

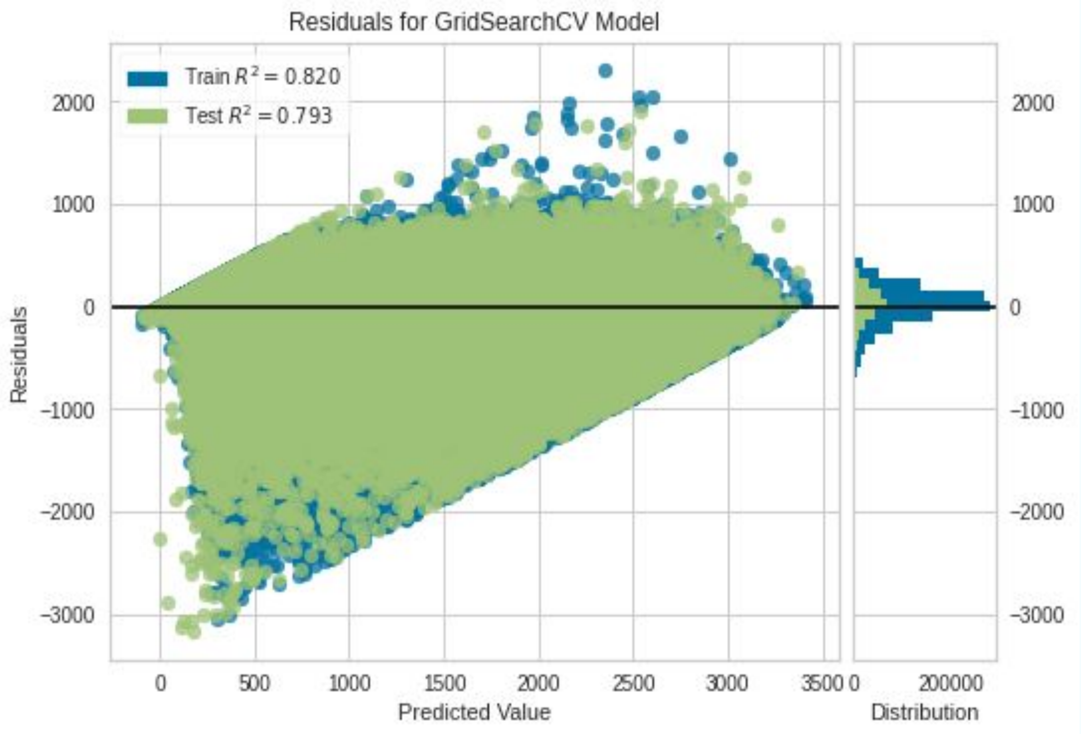
- ❑ **MSE : 105111.8452**
- ❑ **RMSE : 324.20957**
- ❑ **R2 : 0.6739274269631621**
- ❑ **Adjusted R2 : 0.6676**



XGBoost Regression

- Train MSE : 56502.1899
- Train RMSE : 237.7019
- Train R2 : 0.8205
- Train Adjusted R2 : 0.82049

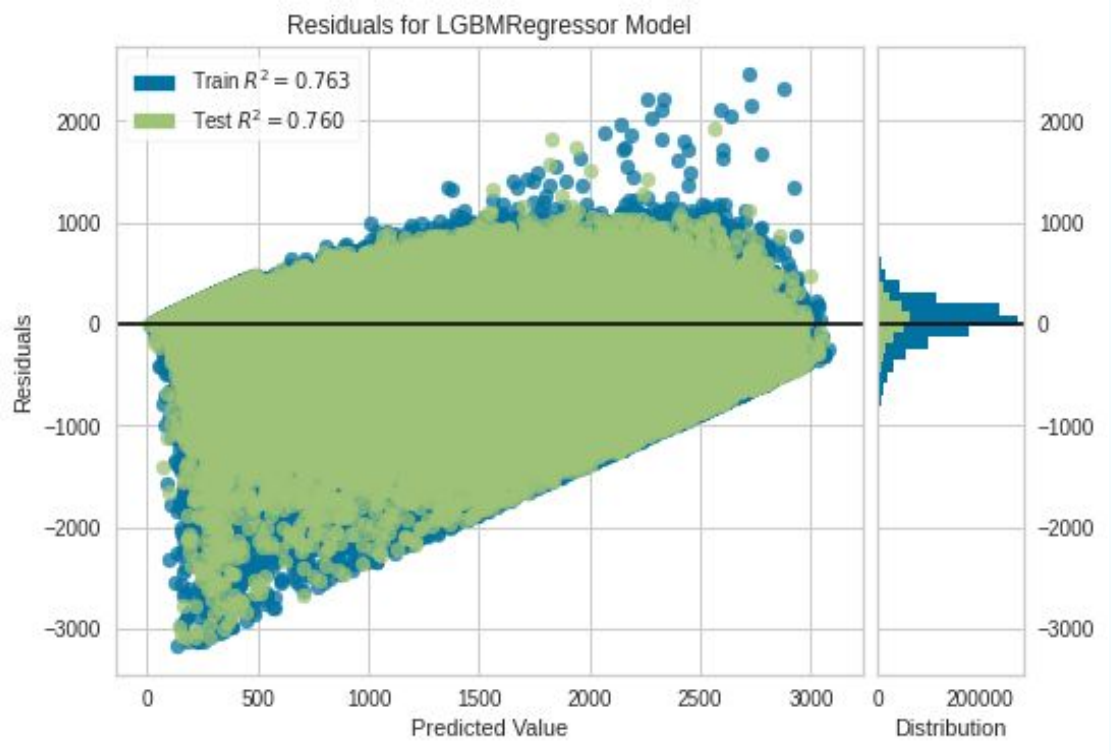
- Test MSE : 65551.7555
- Test RMSE : 256.0308
- Test R2 : 0.7927
- Test Adjusted R2 : 0.79268



LightGBM Regression

- Train MSE : 74751.0686
- Train RMSE : 273.4064
- Train R2 : 0.76252
- Train Adjusted R2 : 0.7625

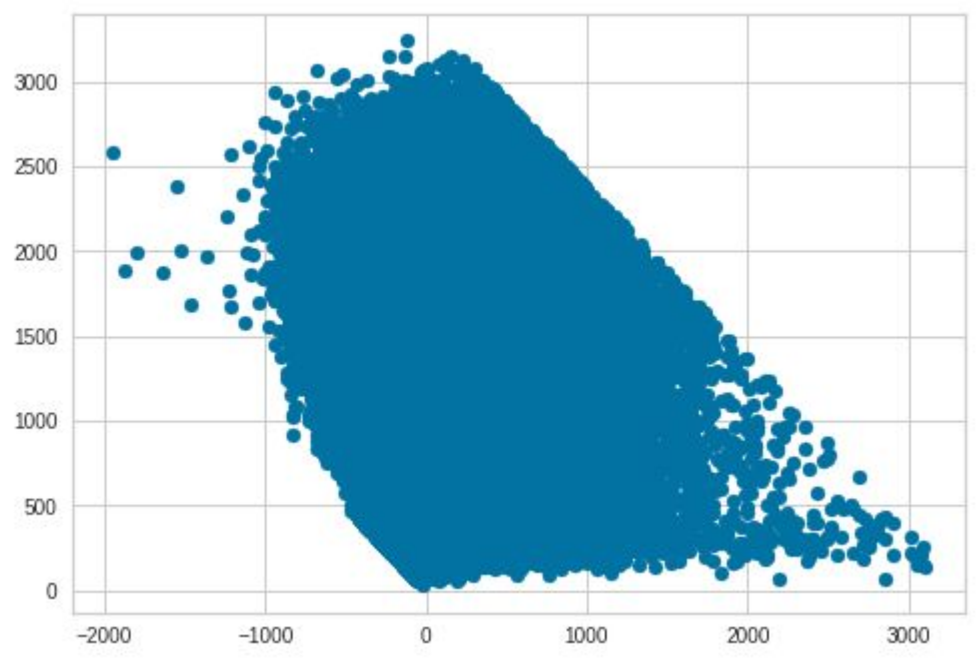
- Test MSE : 75925.3053
- Test RMSE : 275.5455
- Test R2 : 0.7599
- Test Adjusted R2 : 0.7599



Catboost Regression

- Train MSE : 71542.6924
- Train RMSE : 267.4747
- Train R2 : 0.7727
- Train Adjusted R2 : 0.7727

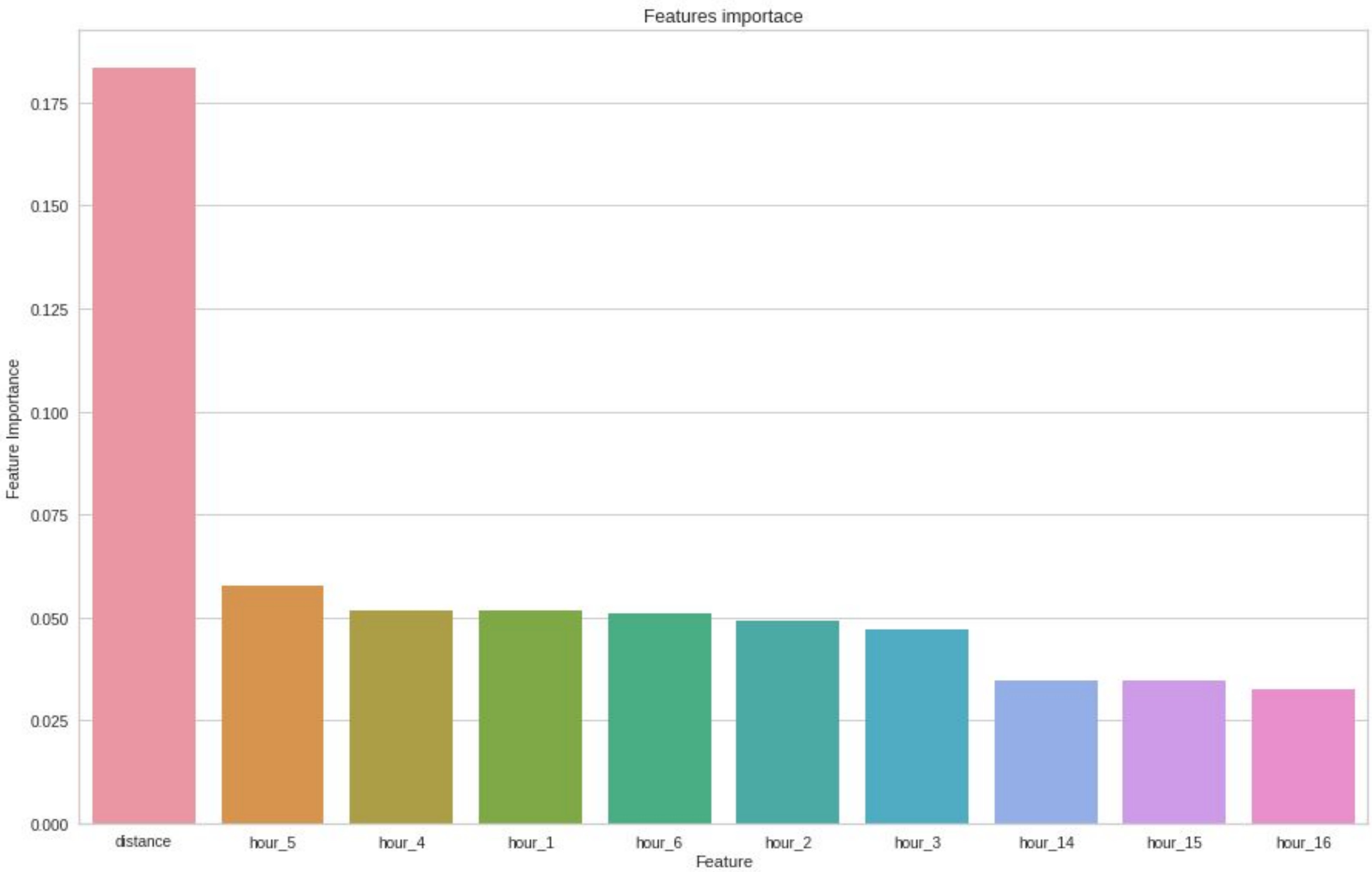
- Train MSE : 72864.7323
- Train RMSE : 269.9347
- Train R2 : 0.7696
- Train Adjusted R2 : 0.76959



Model Selection

Boosting models performed really well with the test and training dataset and no overfitting was observed.

XGBoost Regressor is the best performing model with an Adjusted R2 score of around 80% on test data and 82% with training data.


Features importace
Feature Importance
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
distance
hour_5
hour_4
hour_1
hour_6
hour_2
hour_3
hour_14
hour_15
hour_16
Feature

Challenges :

- Large dataset (around 1.4 million records).
- Properly extracting information from coordinate features and datetime feature.
- Generation of new features which needs to be added in the model.
- Treating the outliers in independent as well as target variable.
- Choosing the right features for modelling.
- Choosing the right models to get the best scores.

Conclusion

In this project we covered various aspects of the Machine learning development cycle. We observed that the data exploration and variable analysis is a very important aspect of the whole cycle and should be done for thorough understanding of the data.

We also cleaned the data while exploring as there were some outliers which should be treated before feature engineering. Further we did feature engineering to filter and gather only the optimal features which are more significant and covered most of the variance in the dataset. Then finally we trained the models on the optimum featureset to get the results

THANK YOU !!