Capstone Project – 2

Supervised ML - Regression NYC Taxi Trip Time Prediction

Team Members

1. 1 Arvind Kale
   2 Priyanka Jain
   . 3 Mohd Sakib Quraishi
   4 Nitish kumar

Presentation Outline

1 Problem Statement
2 Introduction
3 Exploring the dataset
4 Methodology
5 EDA and Data Processing
6 ML Model – Regression
7 Conclusion

Problem Statement:

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

Introduction:

The data is the travel information for the New York taxi.The prediction is using the regression method to predict the trip duration depending on the given variables. The variables contains the locations of pickup and drop-off presenting with latitude and longitude, pickup date/time, number of passenger etc.The design of the learning algorithm includes the preprocess of feature explanation and data selection, modeling and validation.To improve the prediction,we have done several test for modeling and feature extraction.

Exploring the Dataset

Data Summary:

Data Set Name -- NYC Taxi Data.csv - the training set

Statistics –

- Rows - 1458644
- Features - 11 (Including Target)
- Target – Trip Duration Important

Column -- ‘id’, 'vendor_id', 'pickup_datetime', 'dropoff_datetime', 'passenger_count', 'pickup_longitude', 'pickup_latitude', 'dropoff_longitude', 'dropoff_latitude', 'store_and_fwd_flag', 'trip_duration’.

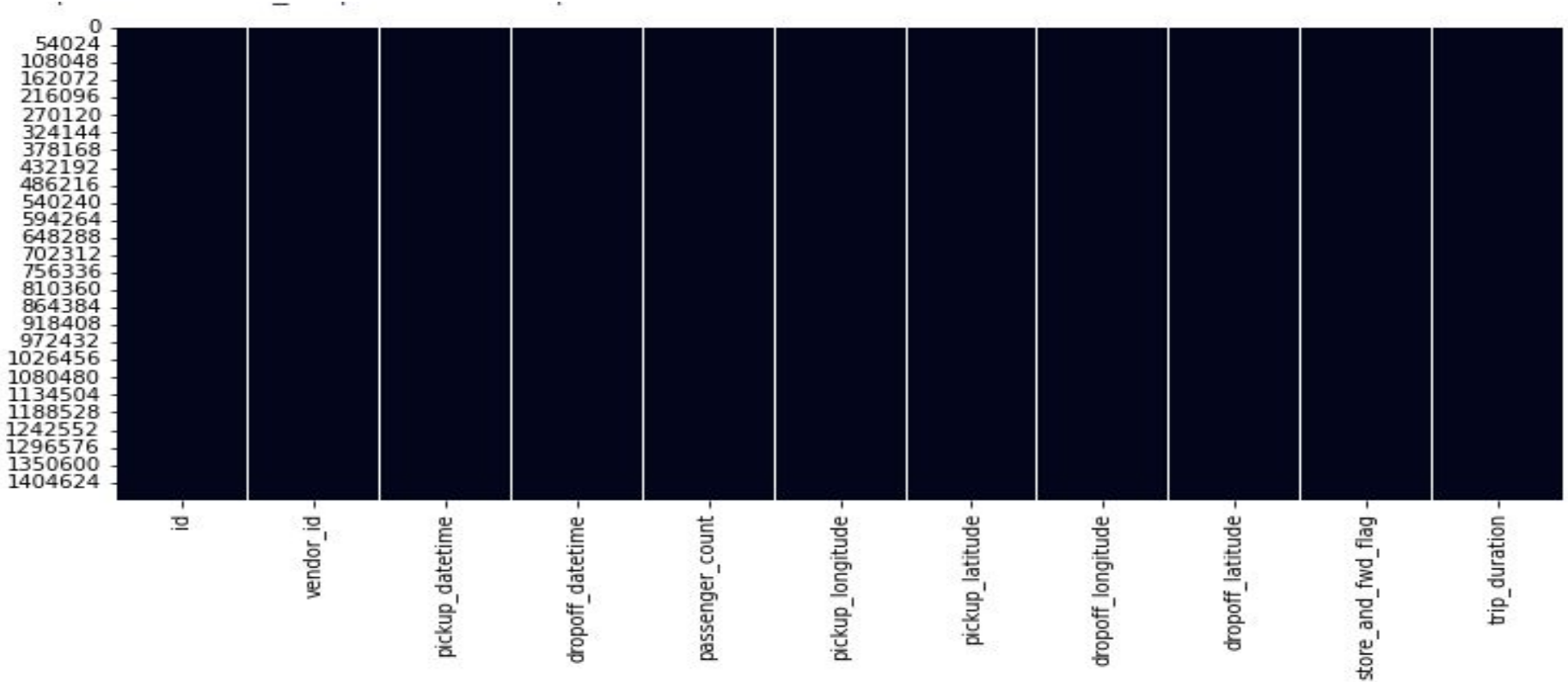
Data Menu:

Independent Variables –

- 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
- 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.

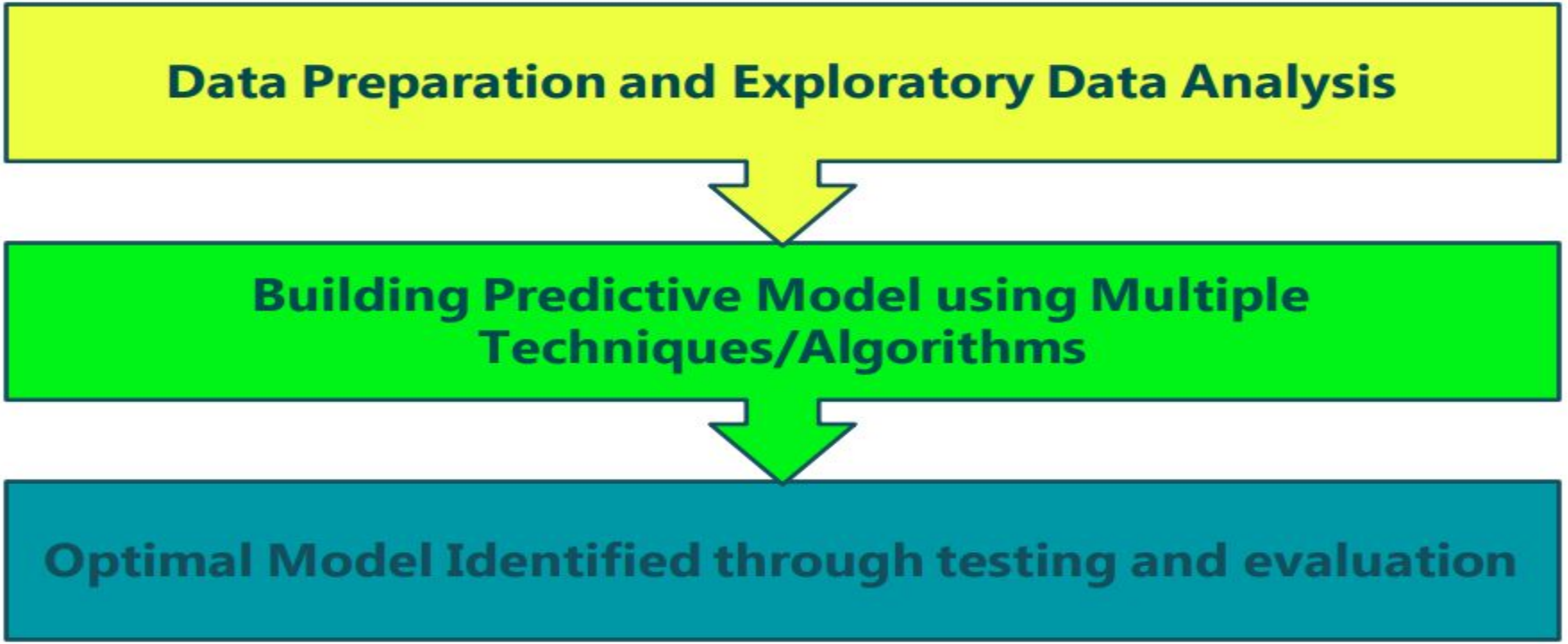
Target Variable – ❖ trip_duration—duration of the trip in seconds

Attribute Information :Null Value



METHODOLOGY

Approach:


Data Preparation and Exploratory Data Analysis
Building Predictive Model using Multiple Techniques/Algorithms
Optimal Model Identified through testing and evaluation

Machine Learning Algorithm:

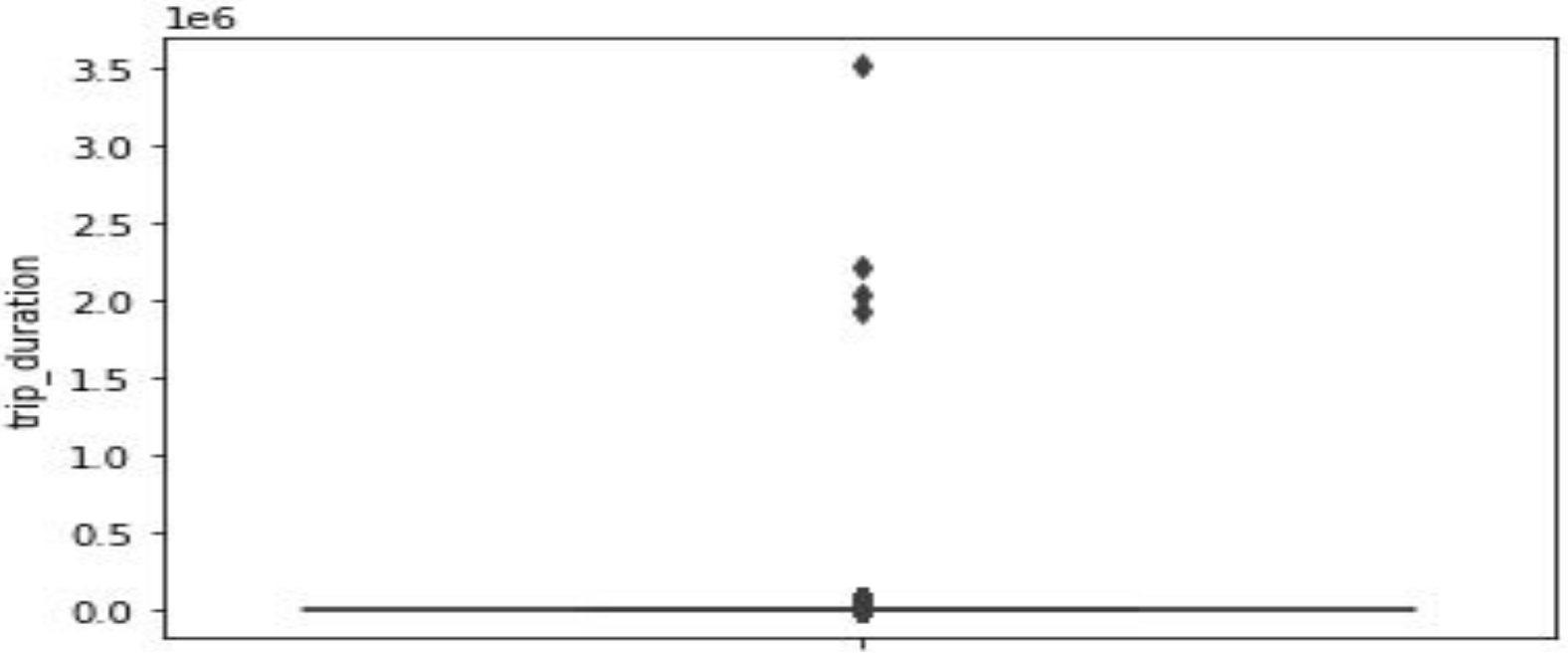
- Linear Regression
- Random Forest
- XGBoost

Tools Used:

- Google Colab Research
- python

DATA PREPROCESSING AND EDA

Outlier treatment: Outlier in column trip duration

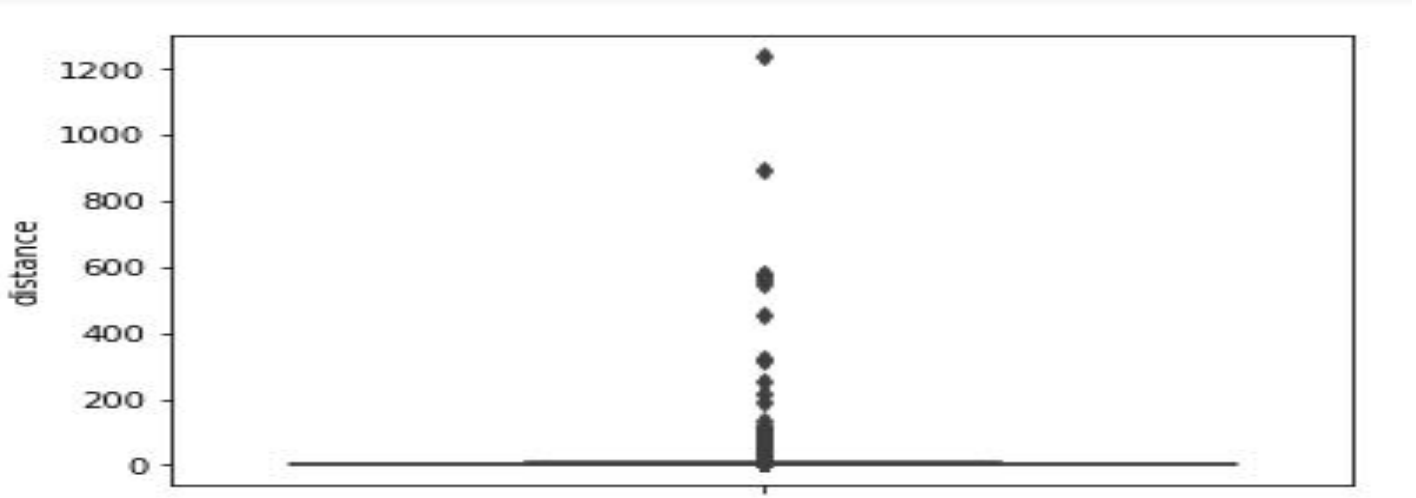


1.You can see that there are some outliers in dataset so we will try to find 0-100 percentile value to find a the correct percentile value for removal of outliers.
2.We can see that the value of outlier is in the range from 90 to 100 percentile but we do not know exactly so we do some more step from 90-100

```
trip_duration
(1, 3601]                 1446313
(3601, 7201]                10045
(7201, 10801]                 141
(10801, 14401]                 35
(14401, 18001]                  5
                              ...
(3506401, 3510001]              0
(3510001, 3513601]              0
(3513601, 3517201]              0
(3517201, 3520801]              0
(3520801, 3524401]              0
Name: trip_duration, Length: 979, dtype: int64
```

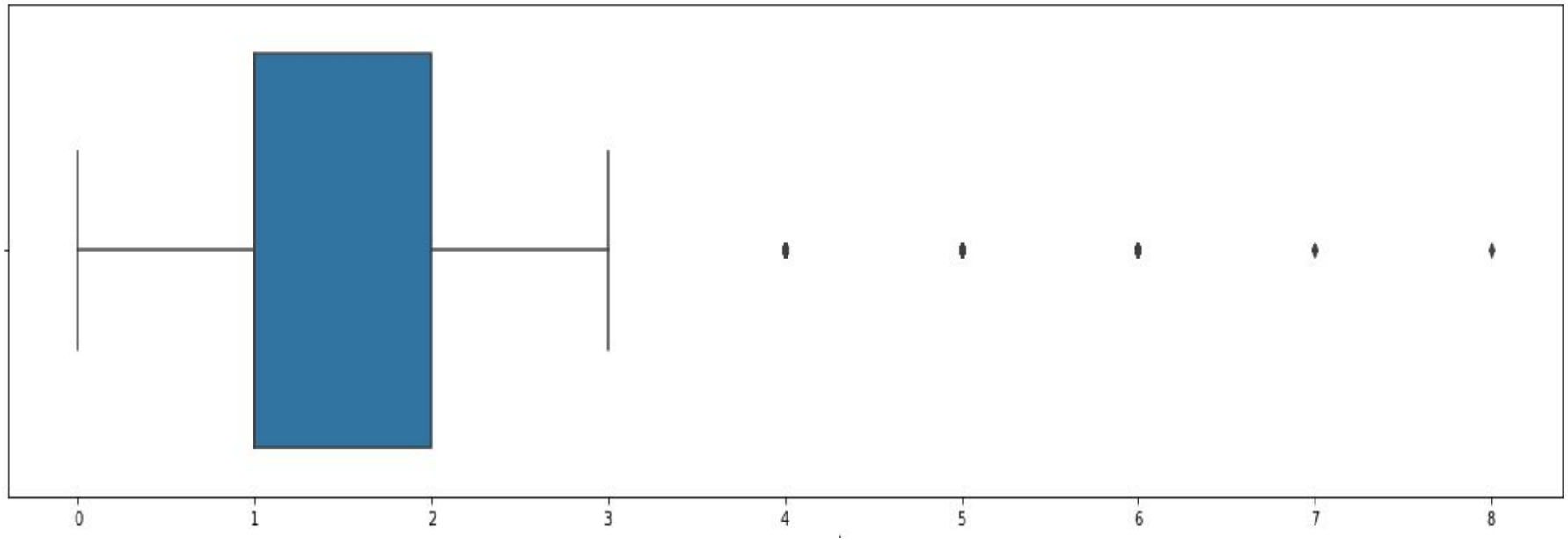
From here we can set the value for trip duration less than 18000 ,otherwise all the values>18000 will be treated as outlier because more than 5 hour of taxi trip is not feasible

Outlier in distance column



We can take value as 30 because 99.9 percentile value are less than it
We will only keep those observation whose distance is greater than 0 ,as equal to zero will be a false entry

Passenger column outlier treatment

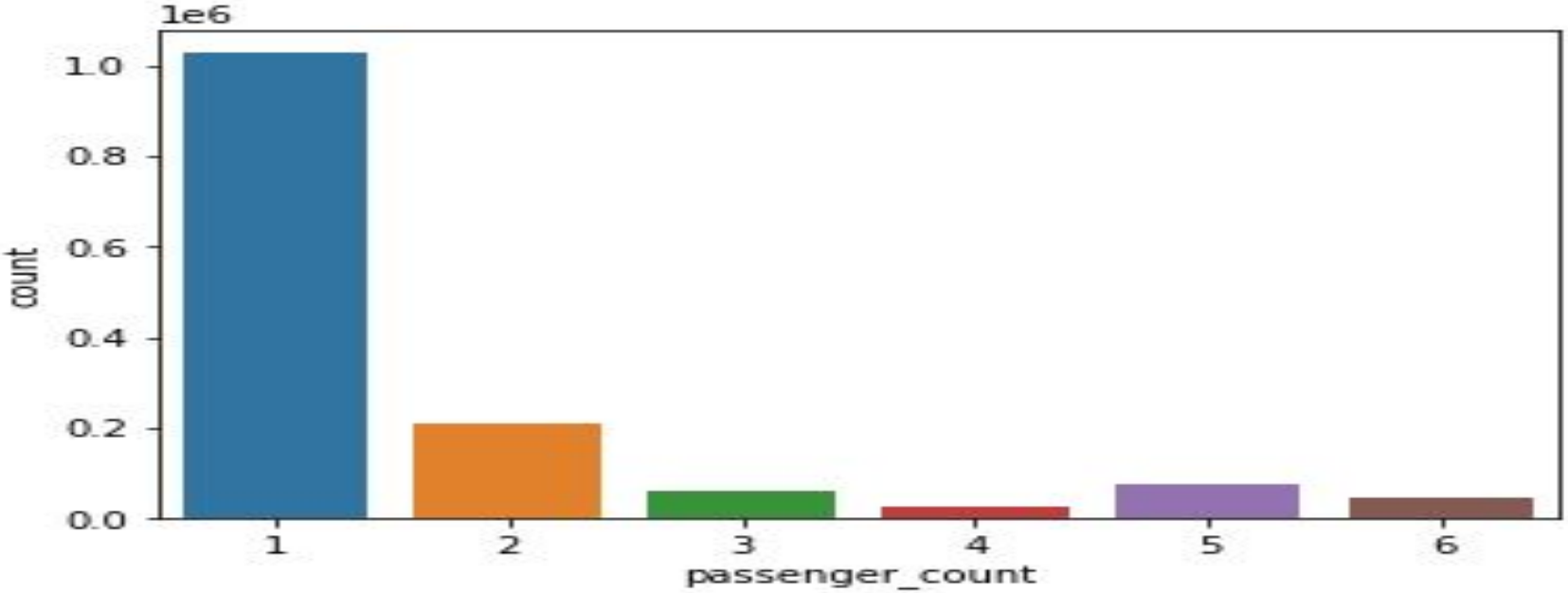


1 There are some trips with 0 passenger count.

2 Few trips consisted of even 7, 8 or 9 passengers. Also, we will remove the records with passenger count > 7, 8 or 9 as they are extreme values and looks very odd to be occupied in a taxi.

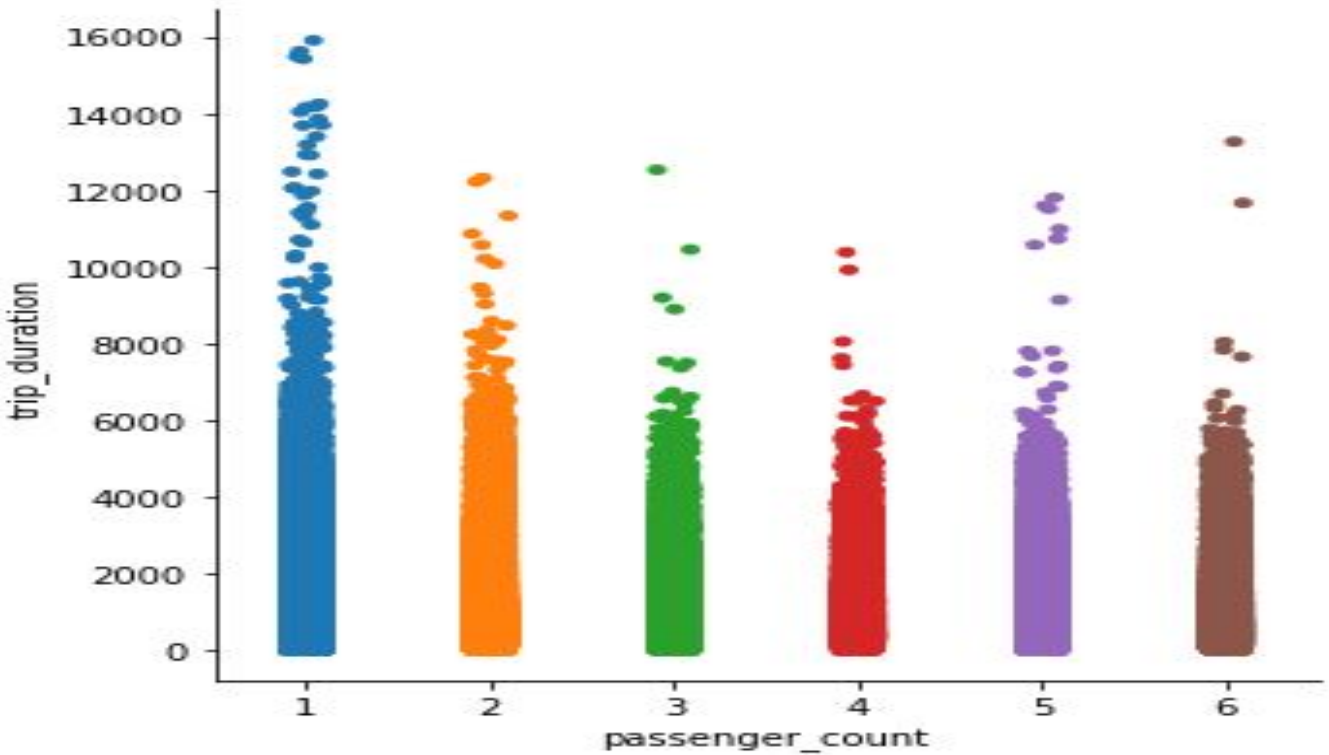
3 Most of trip consist of passenger either 1 or 2 So we would replace the 0 passenger count with 1.

Analysis on : Passenger Count (Contd.)

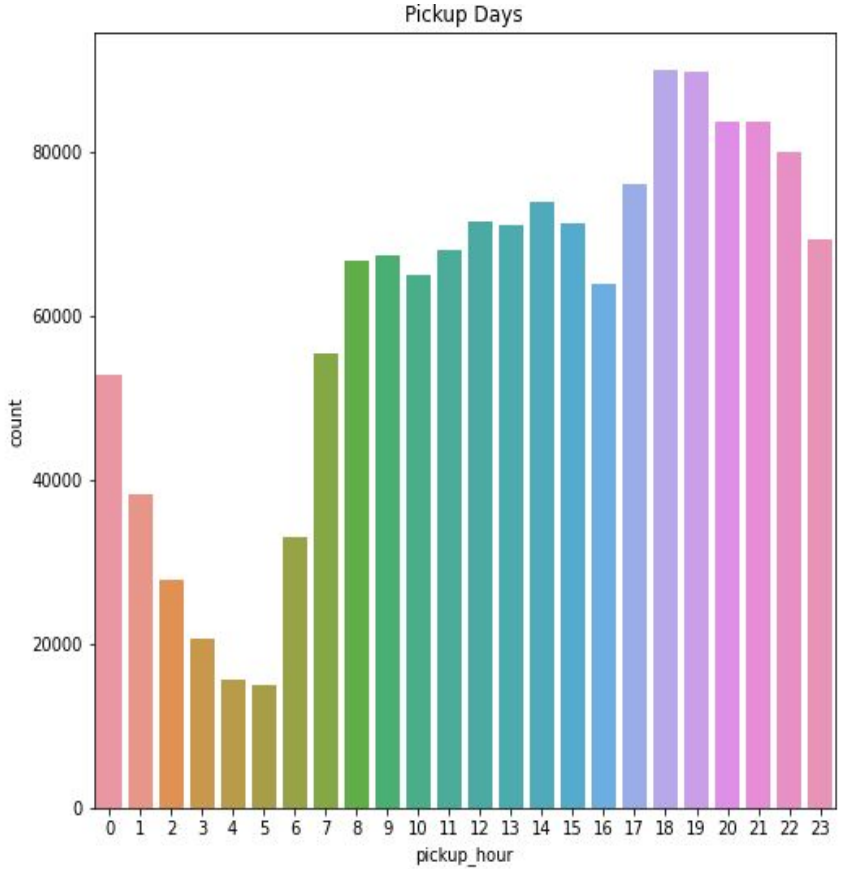


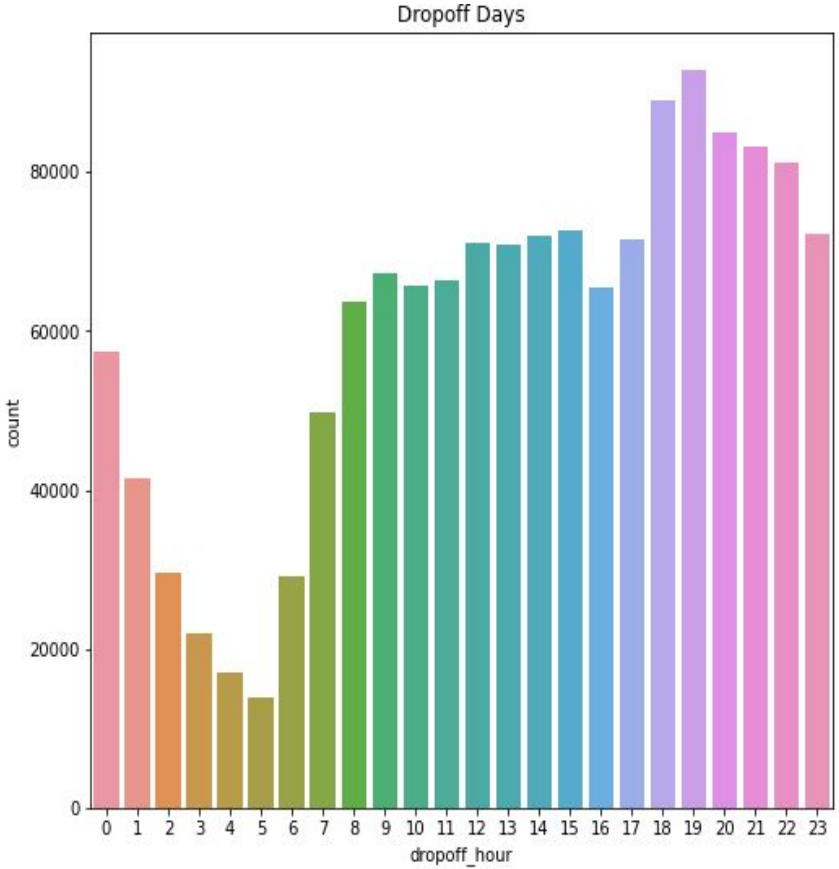
Now, that seems like a fair distribution.We see the highest amount of trips are With 1 passenger

Analysis :Trip duration with passenger count



hourly Analysis

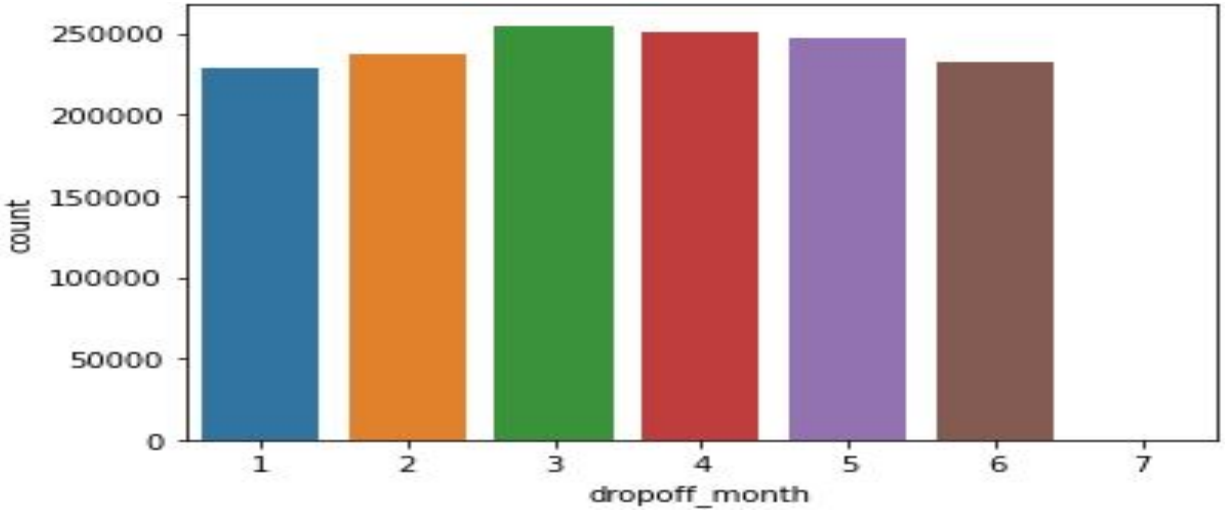




1 Evening hour is mostly busiest hour for trips in pickup days

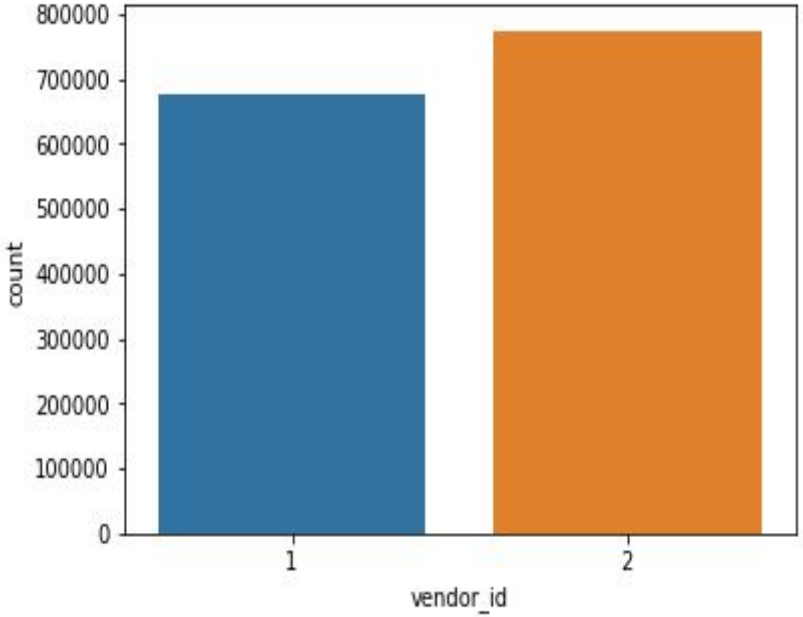
2 Evening hour is mostly busiest hour for trips in drop days

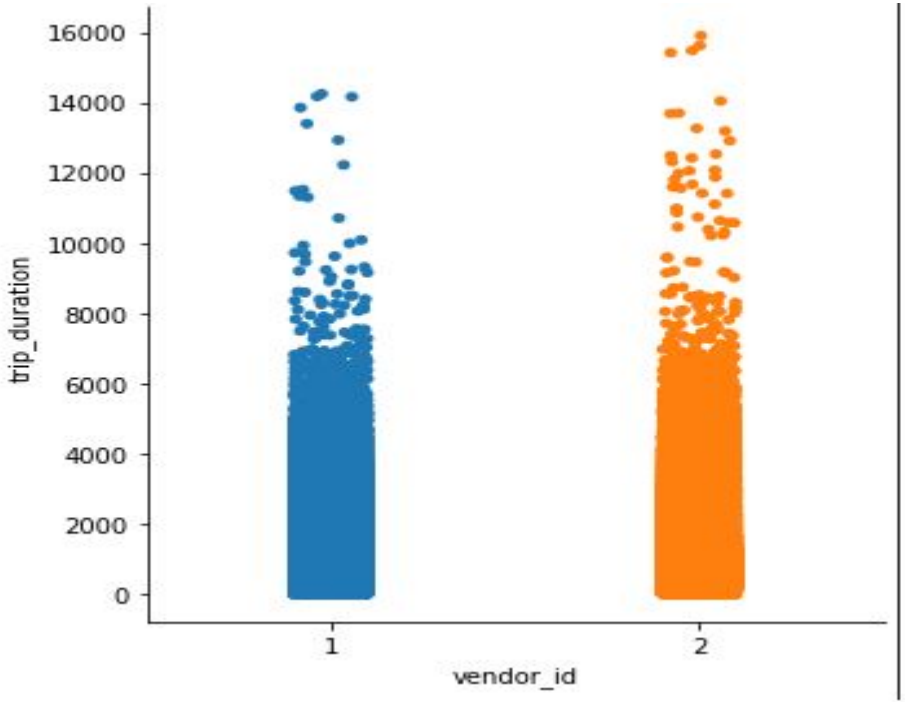
Monthly Analysis



There is not much difference in the number of trips across months.

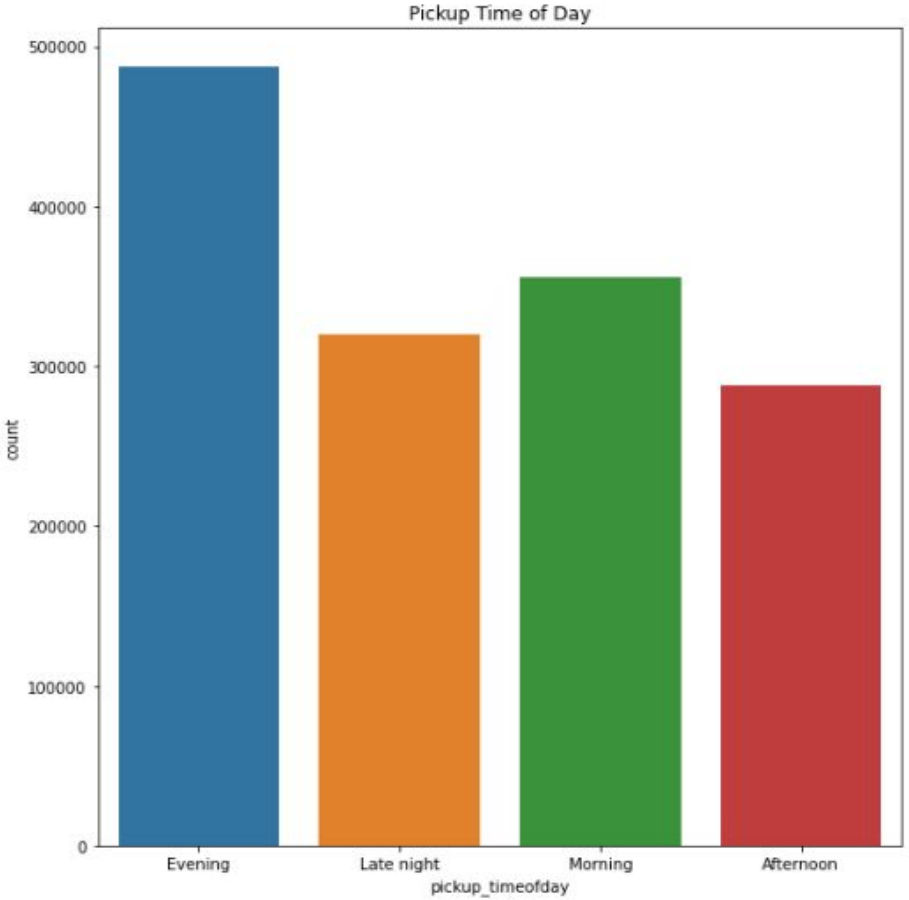
Analysis on : Vendor Id

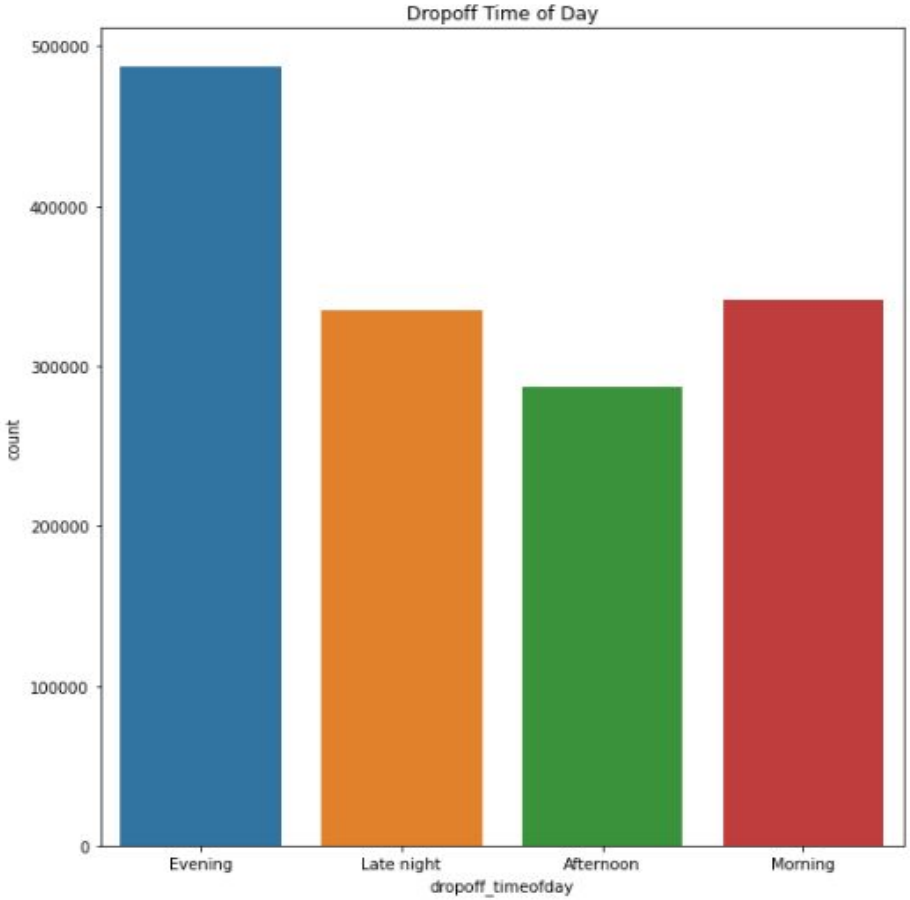




1.We can say that there are two vendors(Service).second service provider is the most opted one by new yorkers.
2.Vendor id 2 takes longer trips as compared to vendor 1

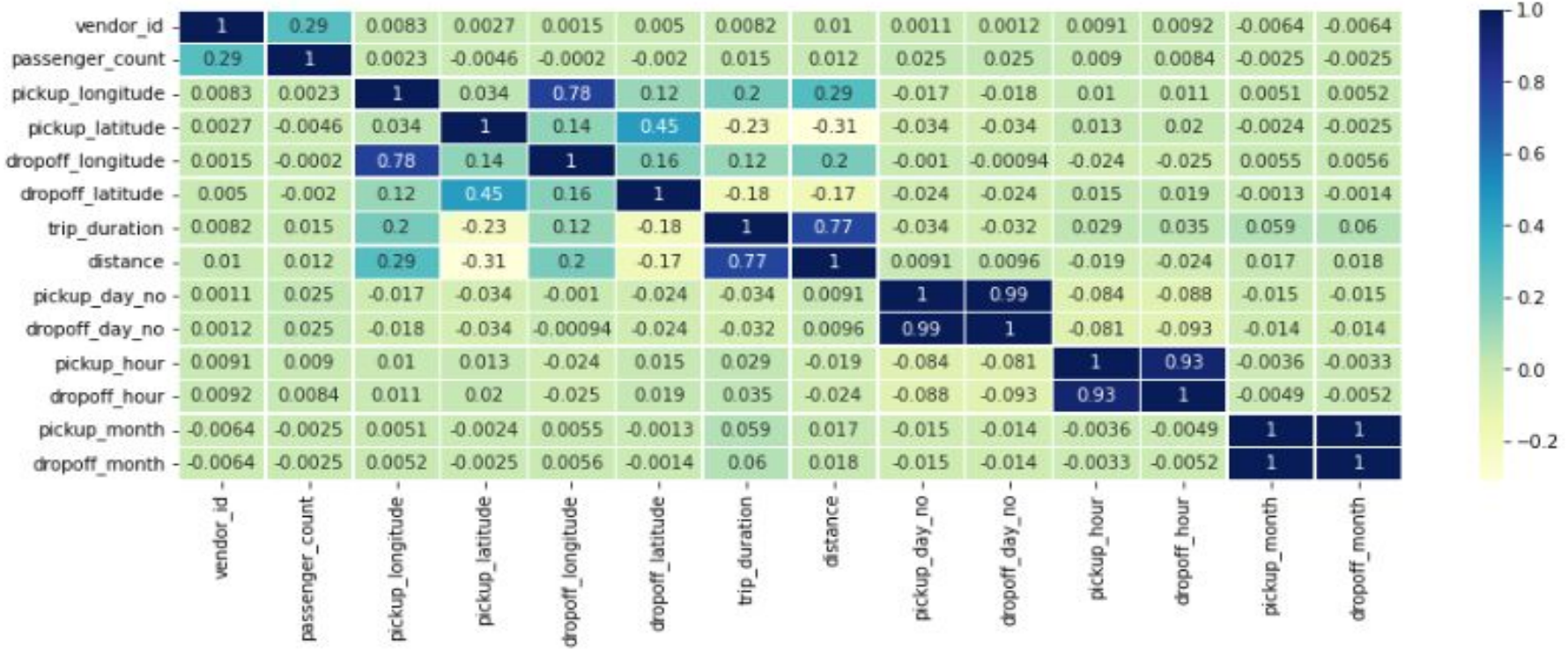
Trips per Time of Day



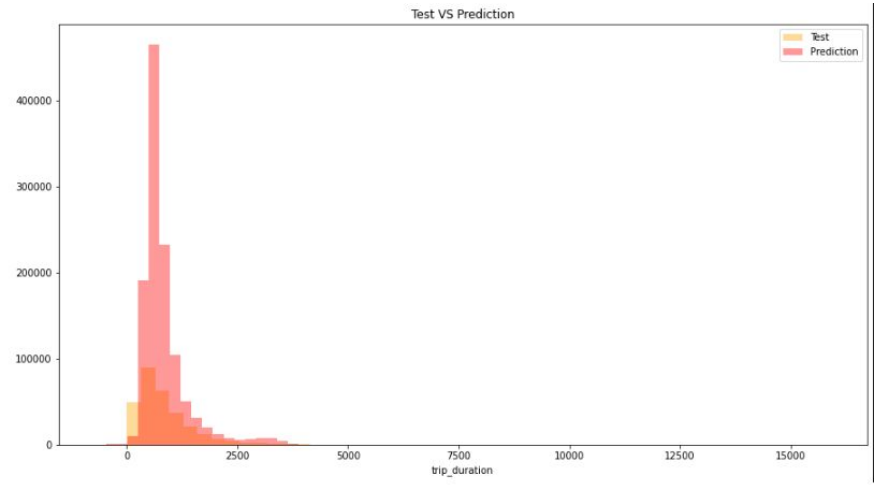


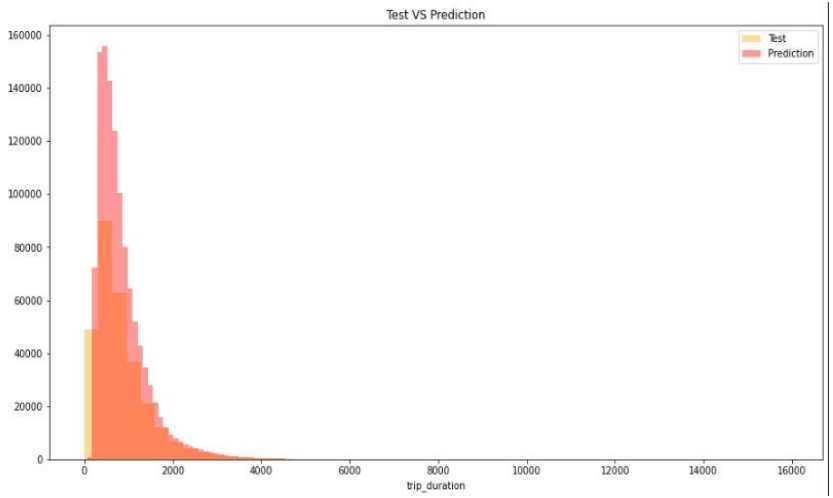
As we saw above, evenings are the busiest time for trips

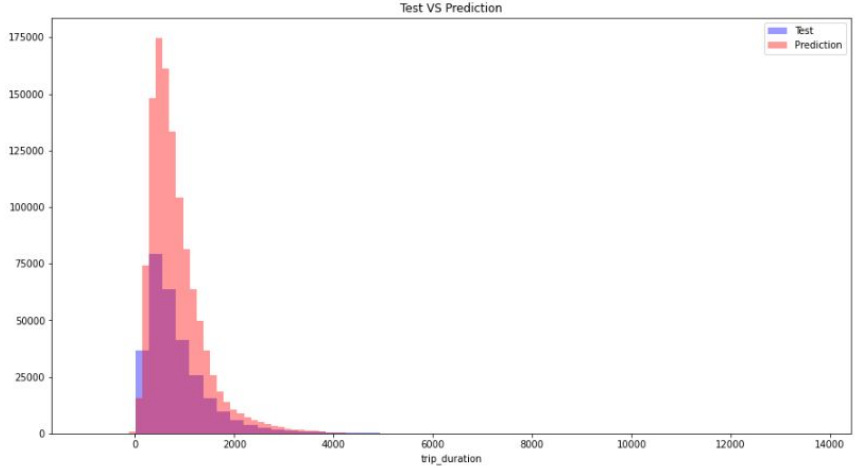
Correlation heatmap



Machine Learning Model – Regression







Visualizations show us how our model's predictions are close to Test Data. It is evident that Random forest regressor
and Xgboost are performing well.

Analysis on : Model Evaluation Result

Algorithms	R2-Score	MAPE
Linear regression	0.6556	0.3992
Random Forest Regressor (Bagging)	0.8482	0.2148
XGboost ML model	0.7758	-

Conclusion:

- Observed which taxi service provider is most Frequently used by New Yorkers.
- Passenger count Analysis showed us that there were few trips with Zero Passengers and One trip with 7,8 and 9 passengers.
- That now appears to be a reasonable distribution. We witness the most trips with only one passenger.
- Passenger count one have highest trip duration compared to the other passenger count.
- Evening hour is mostly busiest hour for trips in pickup day and drop days
- There isn't much of a variation between months in terms of the amount of trips taken. As we've seen, the busiest period for trips is in the nights.
- Taxi giants such as UBER and OLA can use the same data for analyzing the trends that vary throughout the day in the city. This not only helps in better transport analysis but also helps the concerned authorities in planning traffic control and monitoring

THANK YOU