Capstone Project 2

NYC Taxi Trip Time Prediction

Ashik Kumar

Content

1. Introduction
2. Problem statement
3. Data Summery
4. Approach Overview
5. Explanatory Data Analysis
6. Modelling Implementation
7. Model Comparison
8. Conclusion

Introduction

Earth is filled with an enormous population that tends to move from one place to another. Advancement in technologies had led to different ways of transportation. These include buses, autos and especially taxi services. New York City is one of the highly advanced cities of the world with extensive use of taxi services. Along with a vast population, the requirement of commonly available transportation serves the common purpose as it provides a very large transportation system. New York facilitates one of the largest subway systems in the world and comprises various green and yellow cabs which approximately count of around 13,000 taxis. Most of the population of New York depends upon public transport, and it has been estimated that 54 percent of the people do not own a car or a personal vehicle. As a matter of fact, it accounts for almost 200 million taxi trips per year.

Problem Statement

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

The dataset is based on the 2016 NYC Yellow Cab trip record data made available in Big Query on Google Cloud Platform. The data was originally published by the NYC Taxi and Limousine Commission (TLC). The data was sampled and cleaned for the purposes of this project. Based on individual trip attributes, you should predict the duration of each trip in the test set.

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

Data fields

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

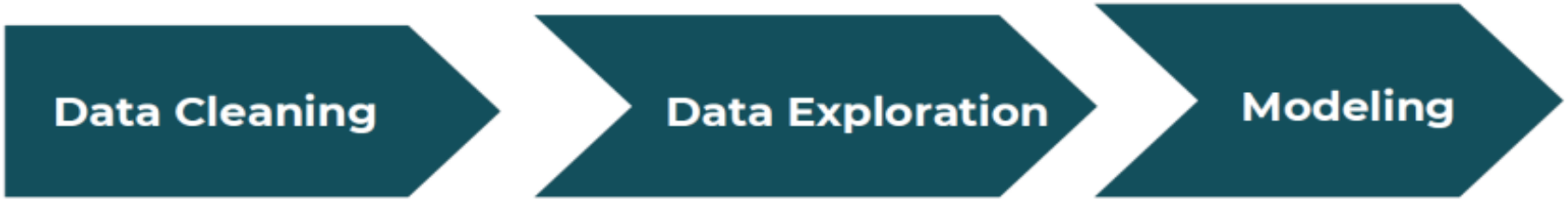
dropoff_longitude - the longitude where the meter was disengaged

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

trip_duration - duration of the trip in seconds

Approach Overview



Understanding and Cleaning

- Find information on documented columns values
- Clean data to get it ready for Analysis

Graphical

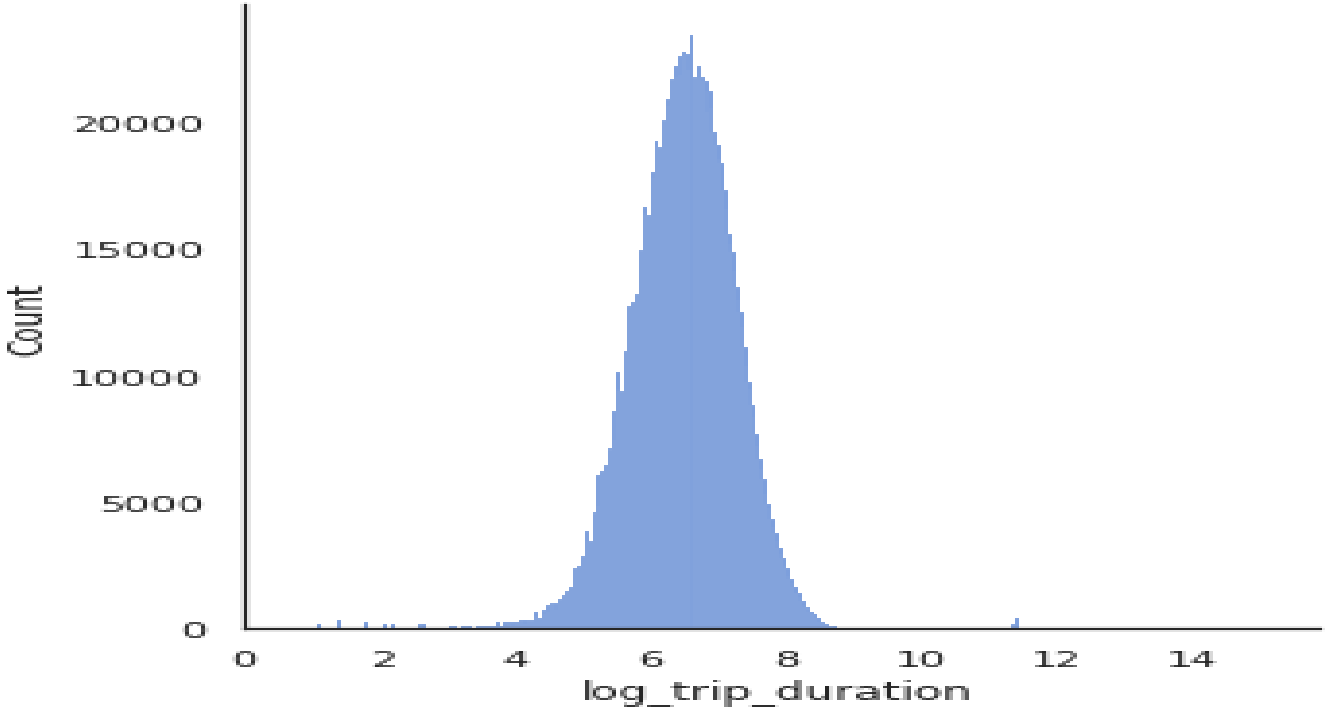
- Examining the data with visualization

Machine Learning

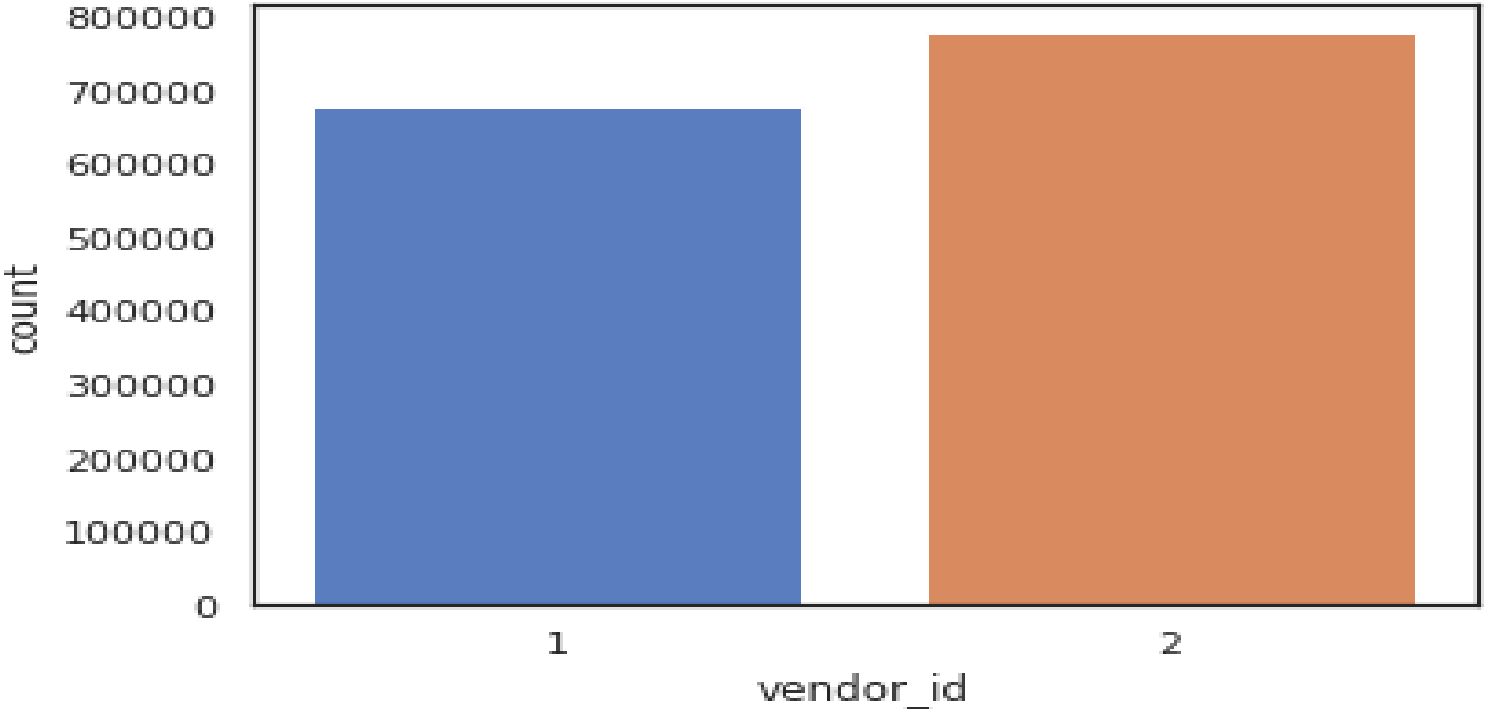
- Linear Regression
- Decision Tree

Exploratory Data Analysis

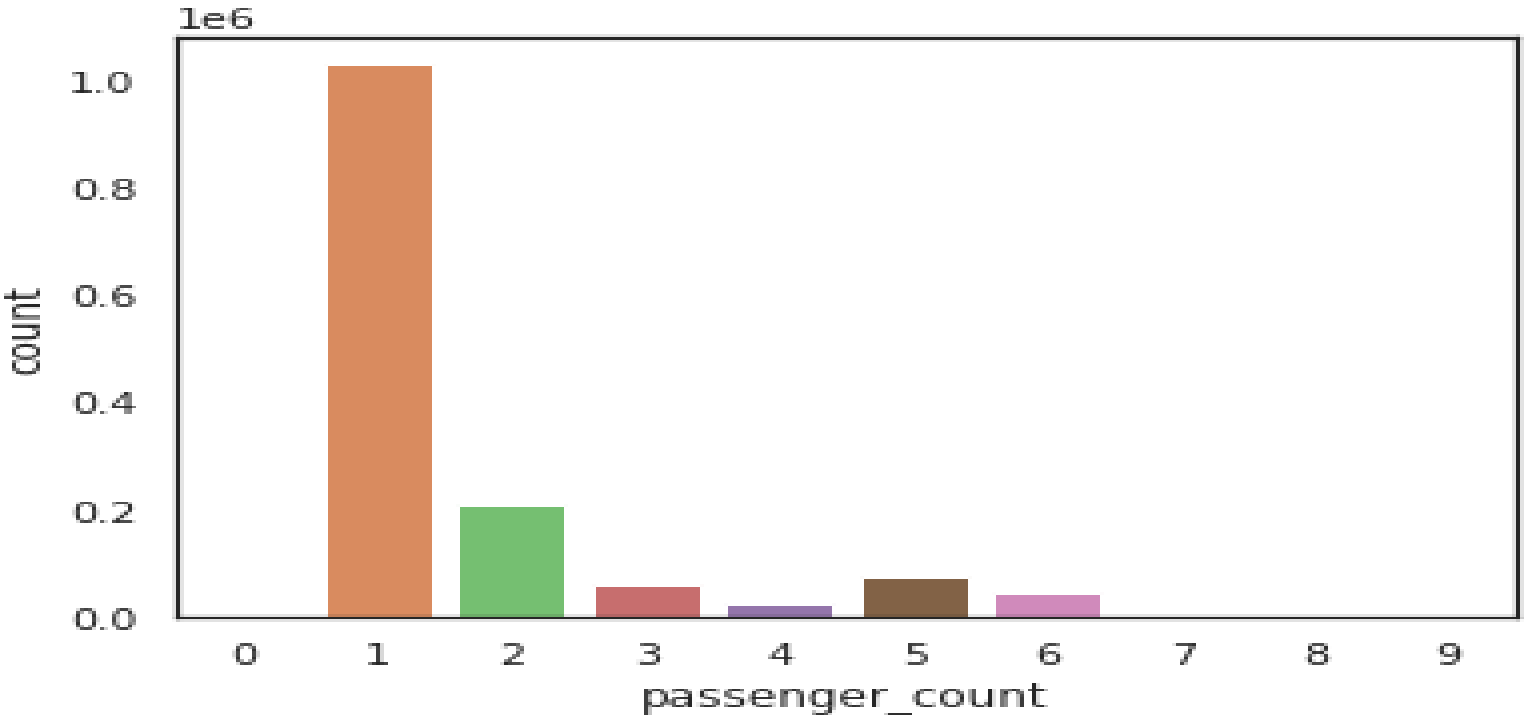
Univariate



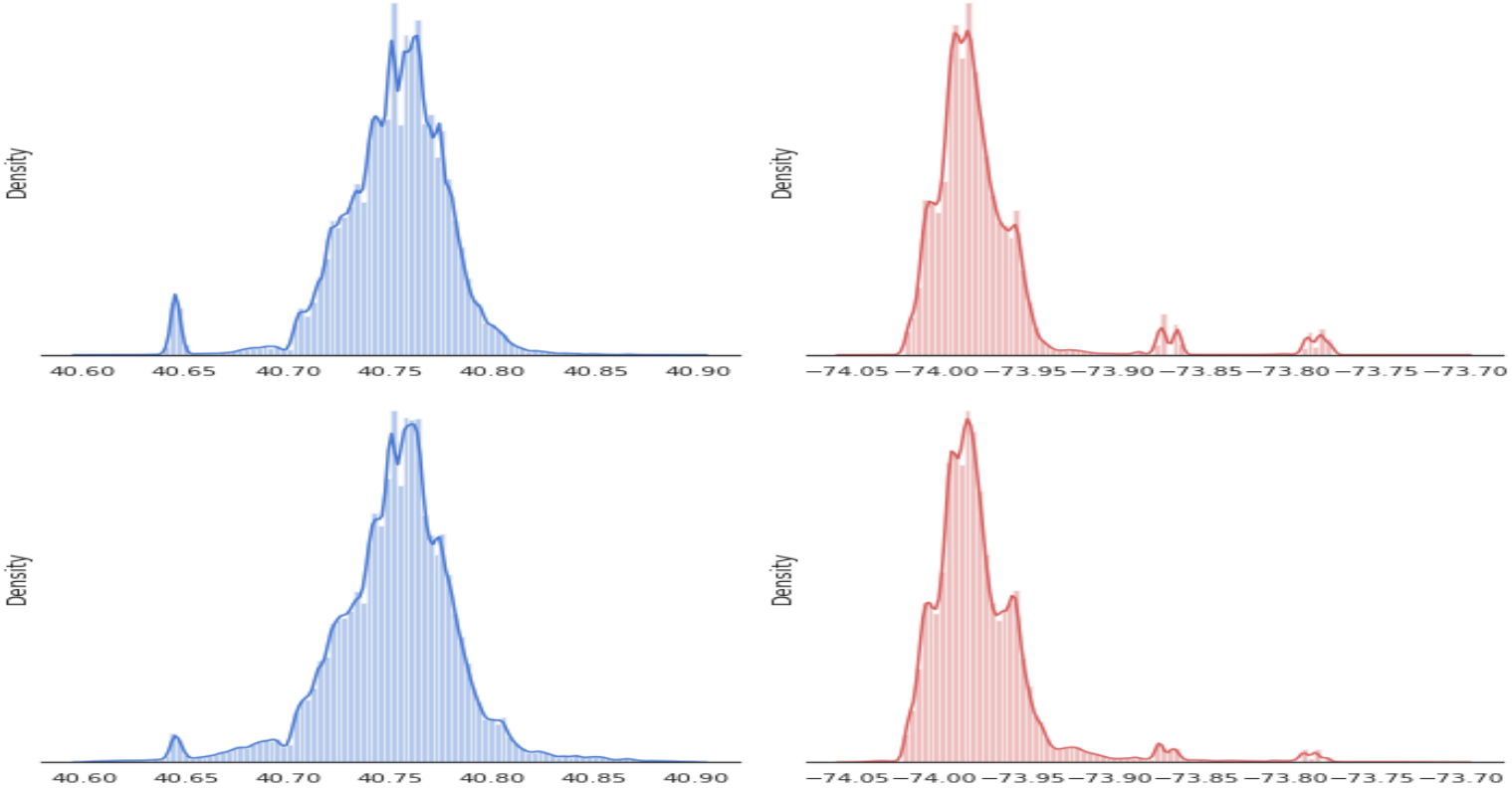
vendor_id


800000
700000
600000
500000
400000
300000
200000
100000
0
count
1
2
vendor_id

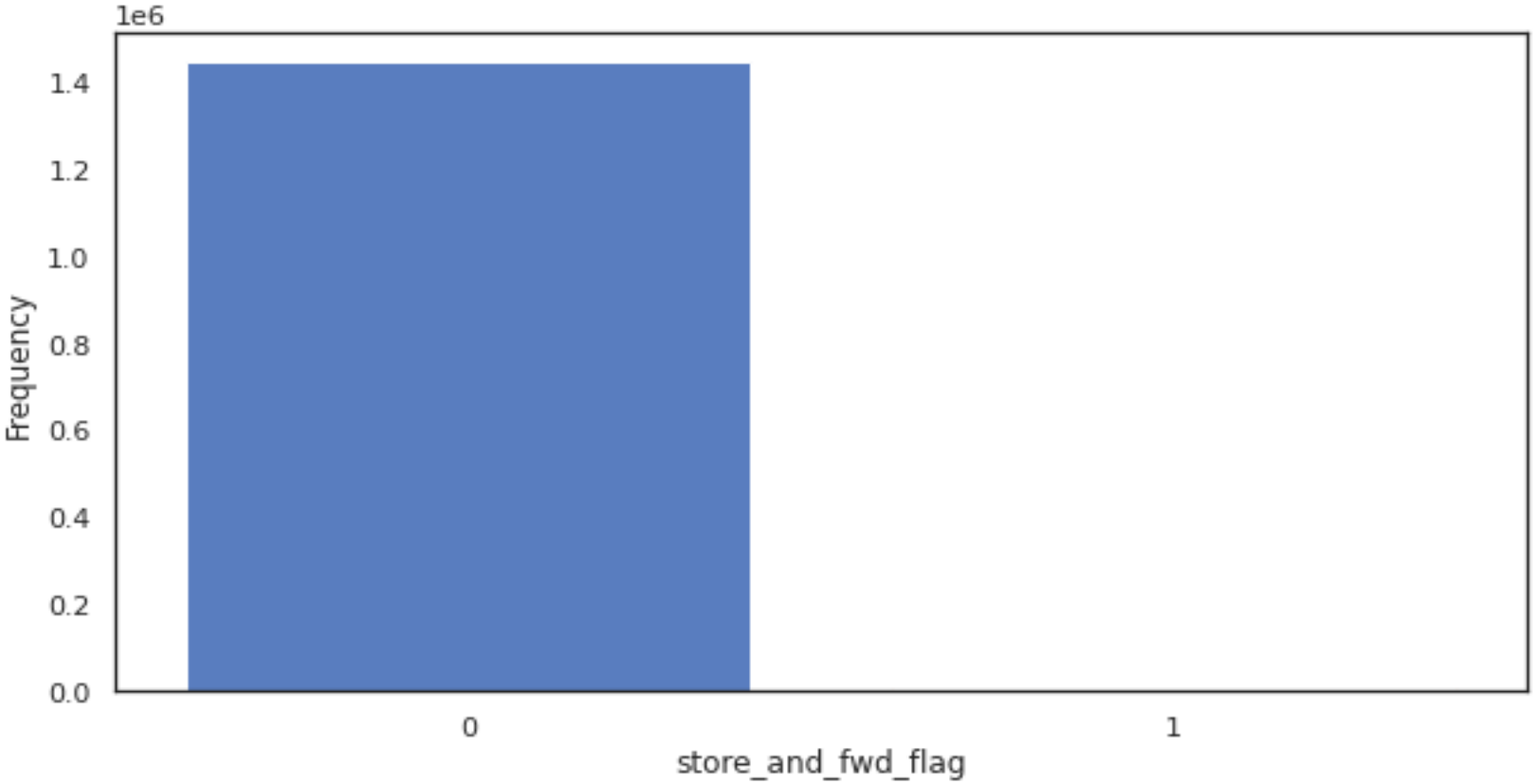
passenger_count



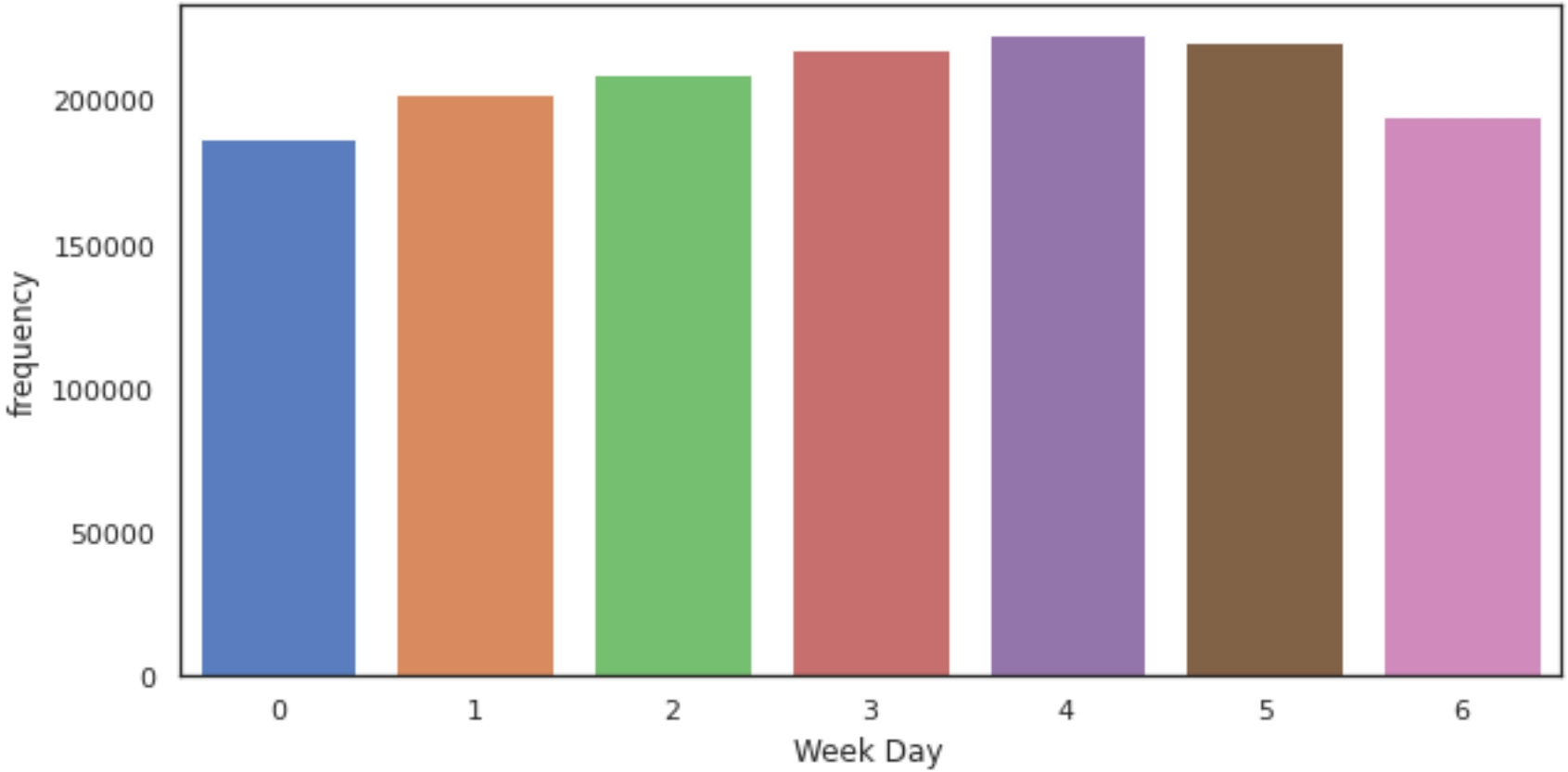
Longitude & Latitude



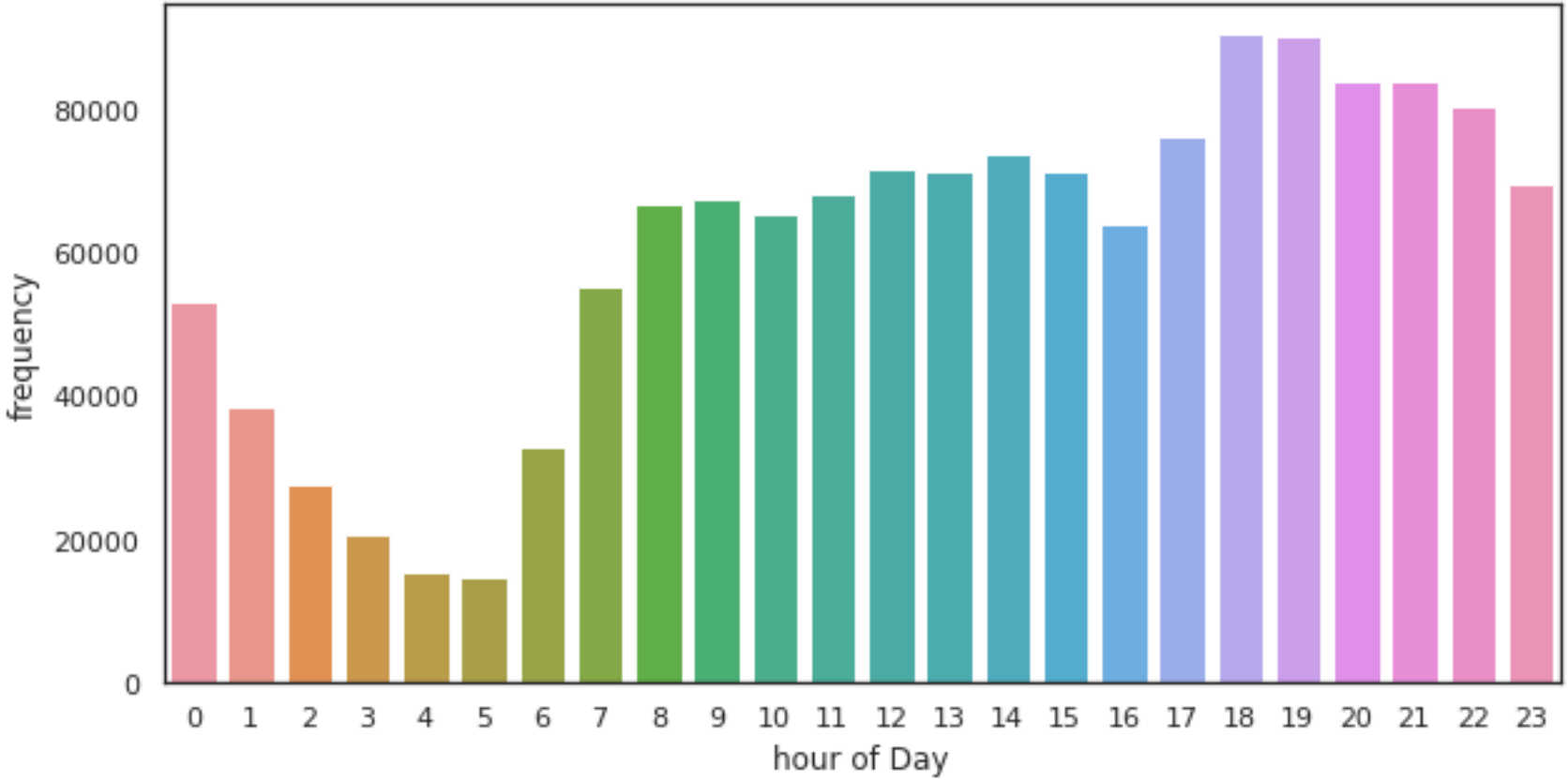
Store_and_fwd_flag



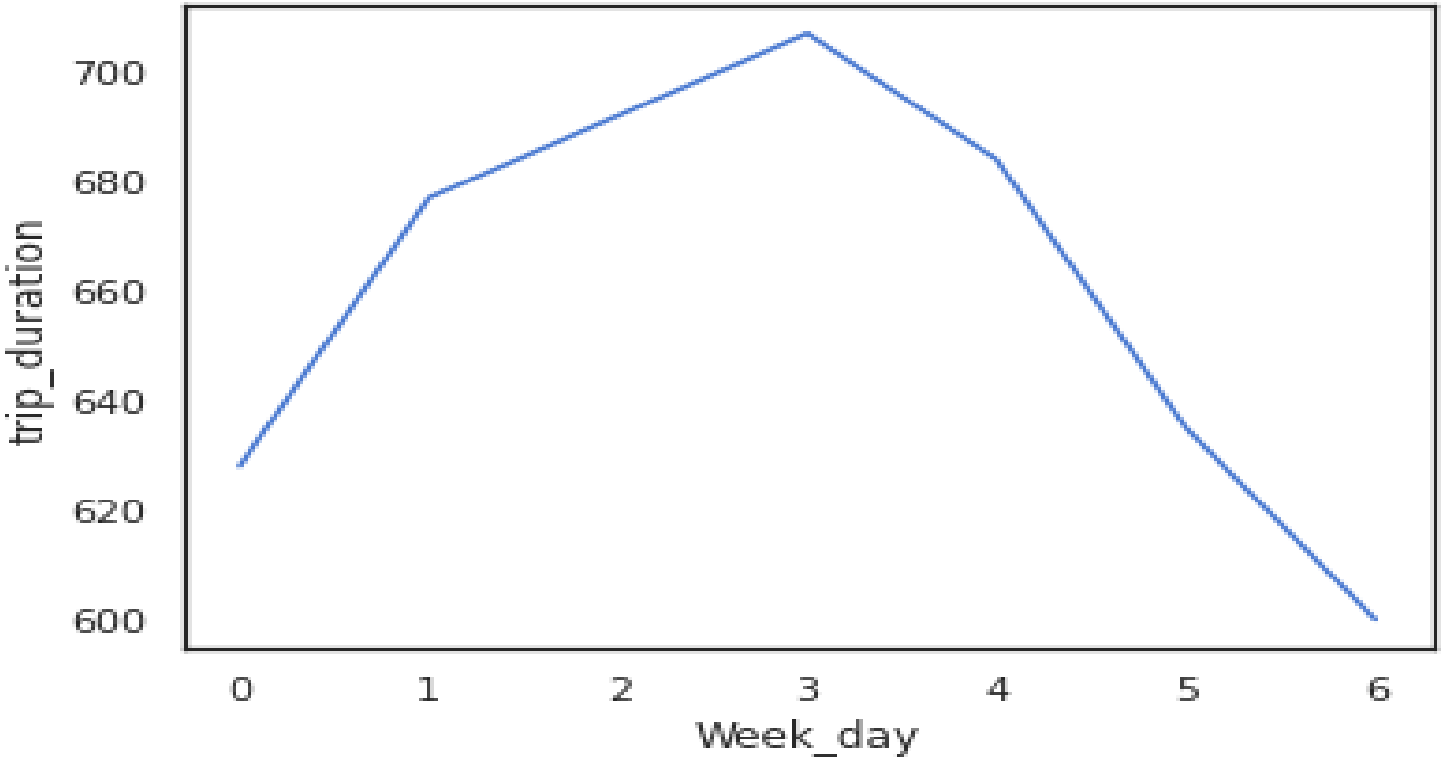
Week_day Vs. passenger



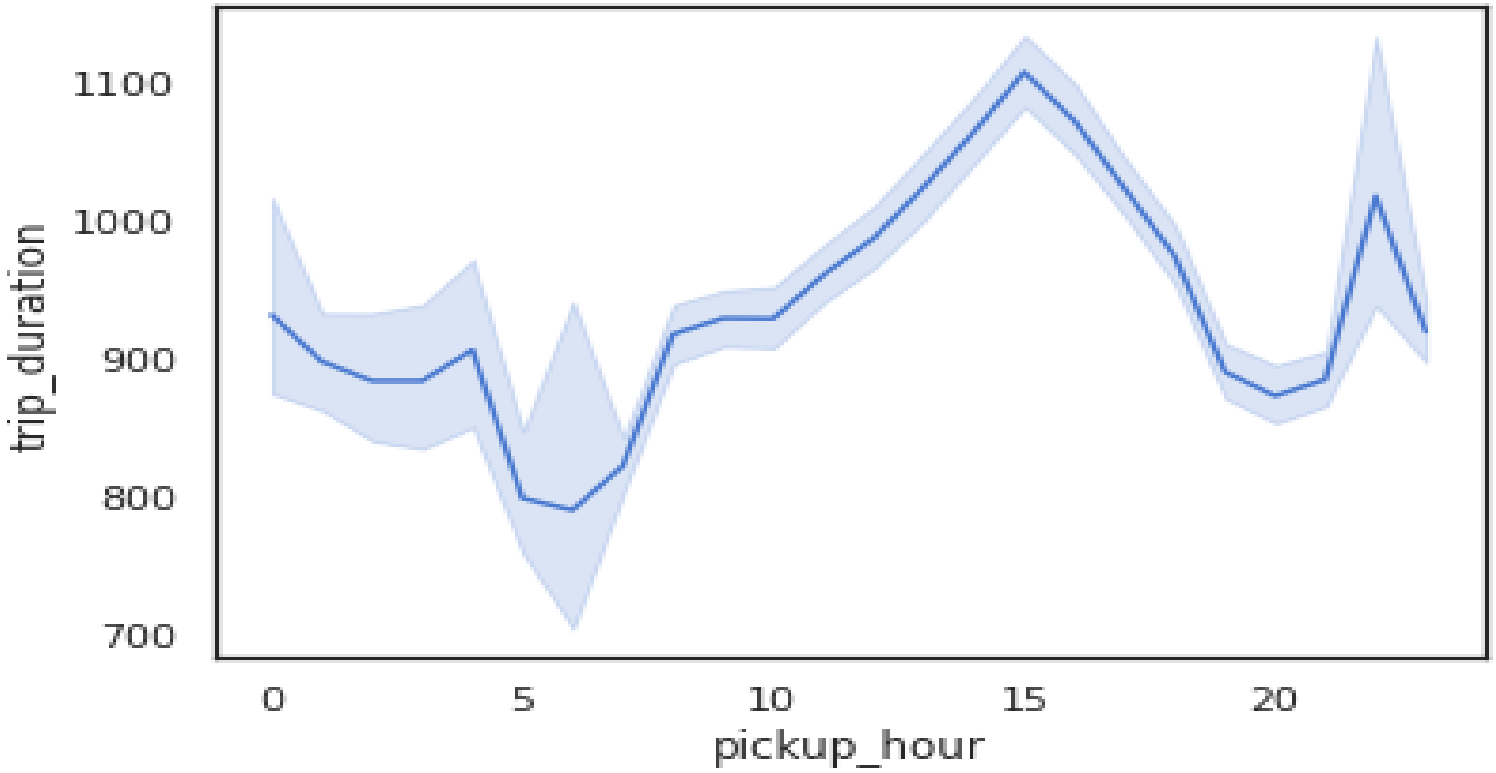
Hour of day Vs. passenger



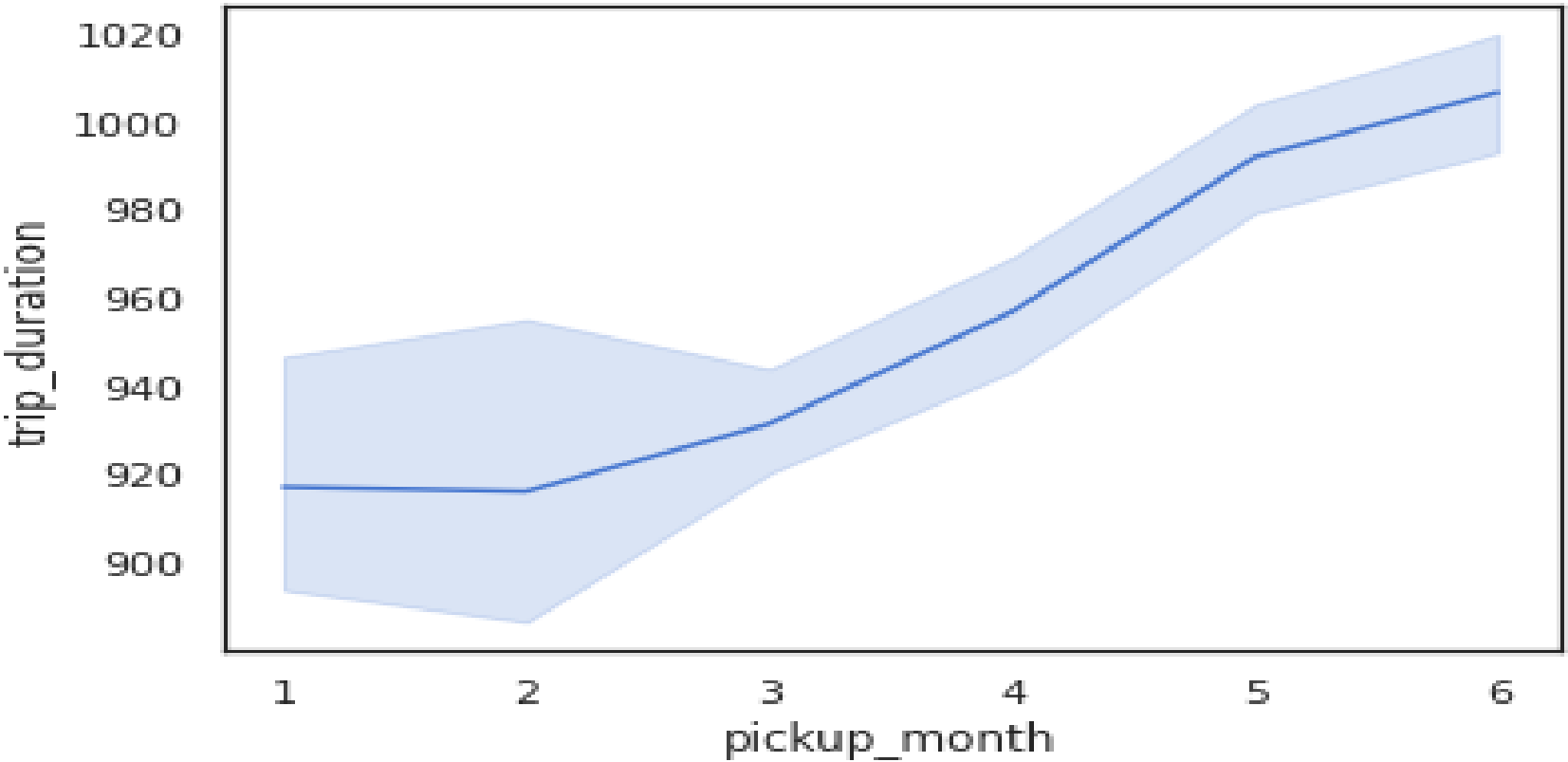
Bivariate Analysis



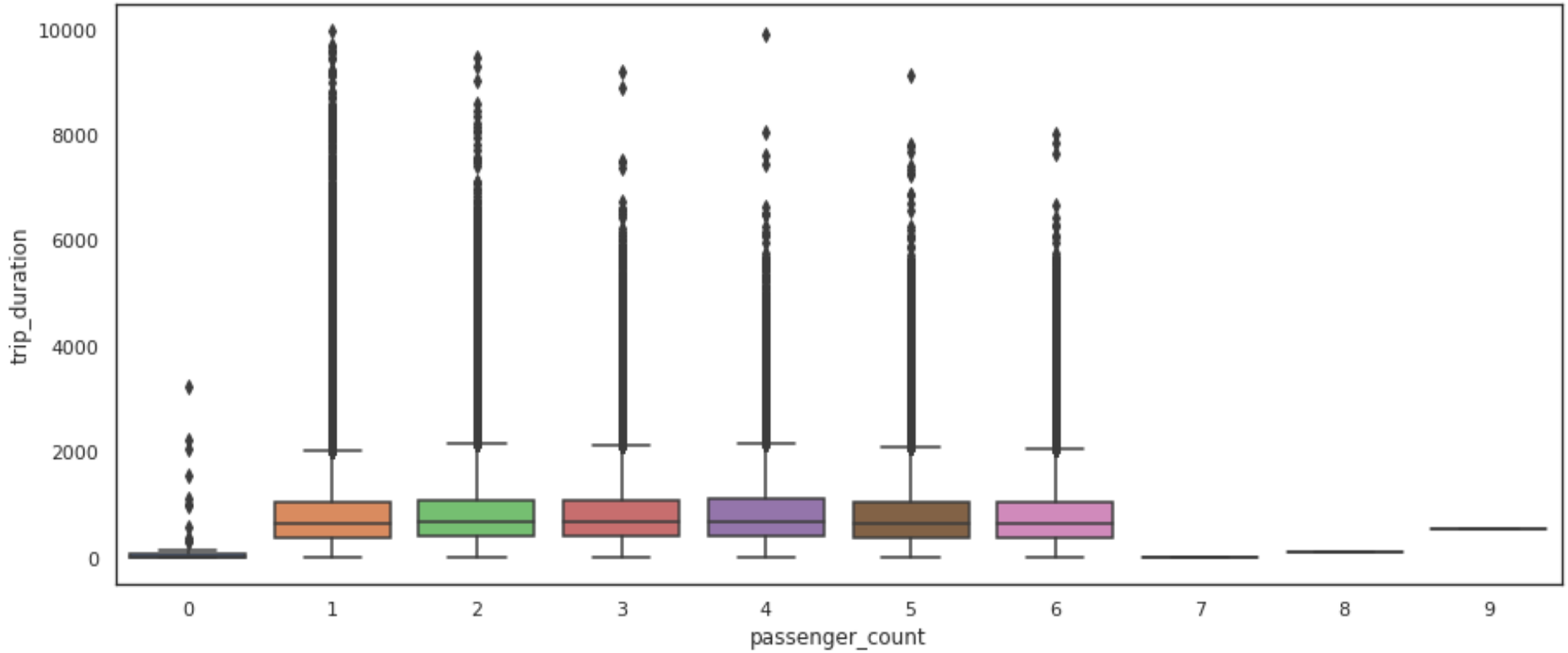
Trip_duration Vs pickup_hour



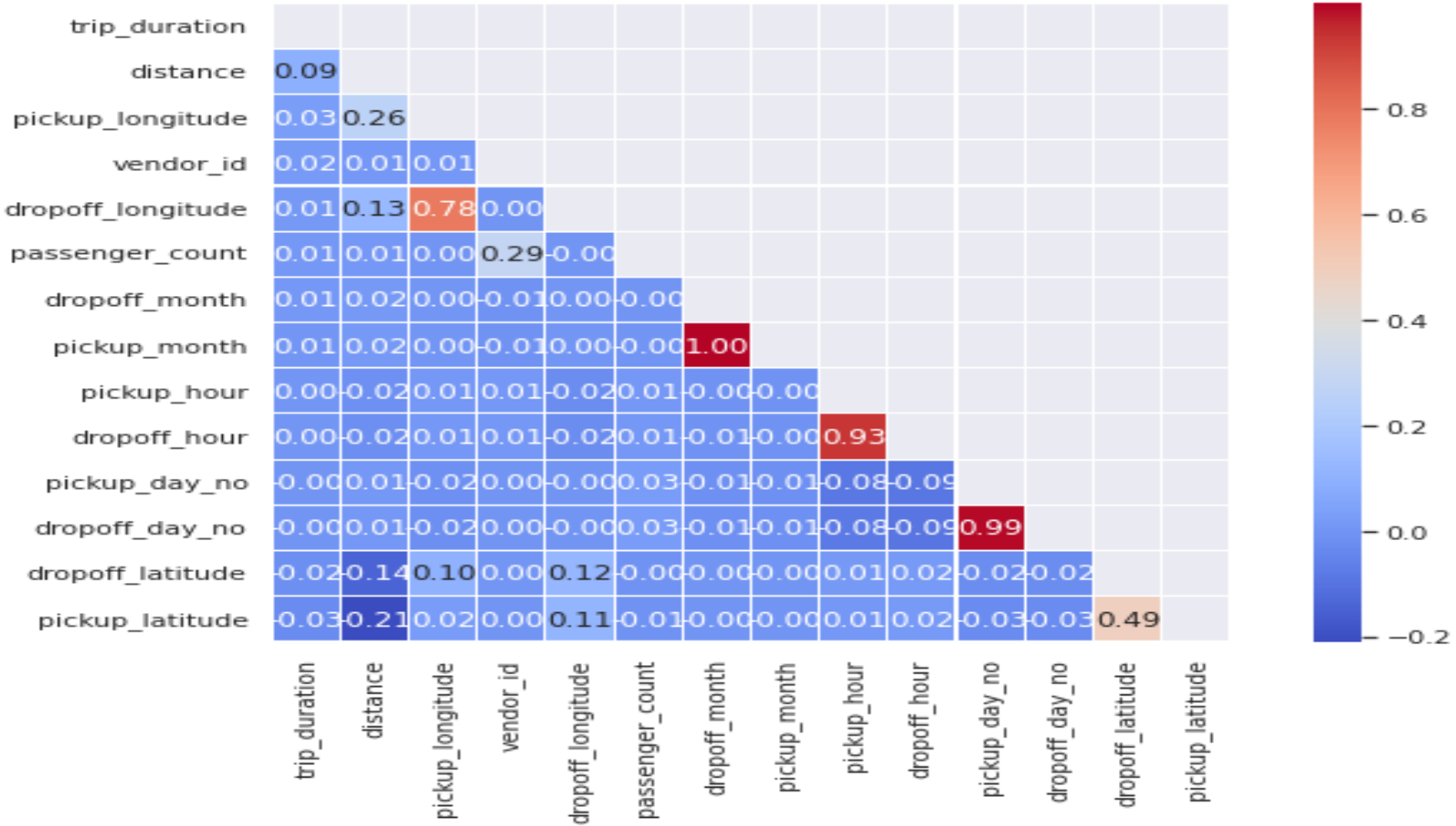
Trip_duration Vs. pickup_month



Trip_duration Vs. passenger_count



Heat Map


trip_duration
distance
pickup_longitude
vendor_id
dropoff_longitude
passenger_count
dropoff_month
pickup_month
pickup_hour
dropoff_hour
pickup_day_no
dropoff_day_no
dropoff_latitude
pickup_latitude
0.8
0.6
0.4
0.2
0.0
−0.2

Preparing dataset for modelling

	id	vendor_id	pickup_datetime	dropoff_datetime	passenger_count	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	store_and_fwd_flag	trip_duration
0	id2875421	2	2016-03-14 17:24:55	2016-03-14 17:32:30	1	-73.982155	40.767937	-73.964630	40.765602	N	455
1	id2377394	1	2016-06-12 00:43:35	2016-06-12 00:54:38	1	-73.980415	40.738564	-73.999481	40.731152	N	663
2	id3858529	2	2016-01-19 11:35:24	2016-01-19 12:10:48	1	-73.979027	40.763939	-74.005333	40.710087	N	2124
3	id3504673	2	2016-04-06 19:32:31	2016-04-06 19:39:40	1	-74.010040	40.719971	-74.012268	40.706718	N	429
4	id2181028	2	2016-03-26 13:30:55	2016-03-26 13:38:10	1	-73.973053	40.793209	-73.972923	40.782520	N	435
5	id0801584	2	2016-01-30 22:01:40	2016-01-30 22:09:03	6	-73.982857	40.742195	-73.992081	40.749184	N	443
6	id1813257	1	2016-06-17 22:34:59	2016-06-17 22:40:40	4	-73.969017	40.757839	-73.957405	40.765896	N	341
7	id1324603	2	2016-05-21 07:54:58	2016-05-21 08:20:49	1	-73.969276	40.797779	-73.922470	40.760559	N	1551

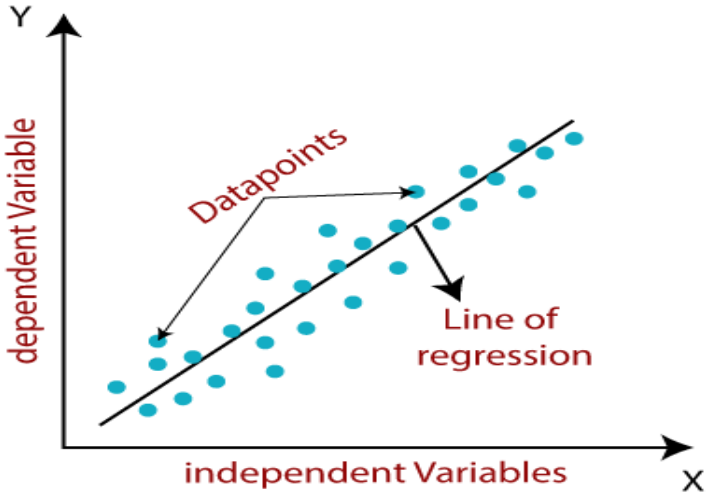
Train Set (972332,11)

Test Set (486312,11)

Model Building

Linear Regression

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as **sales, salary, age, product price,** etc.

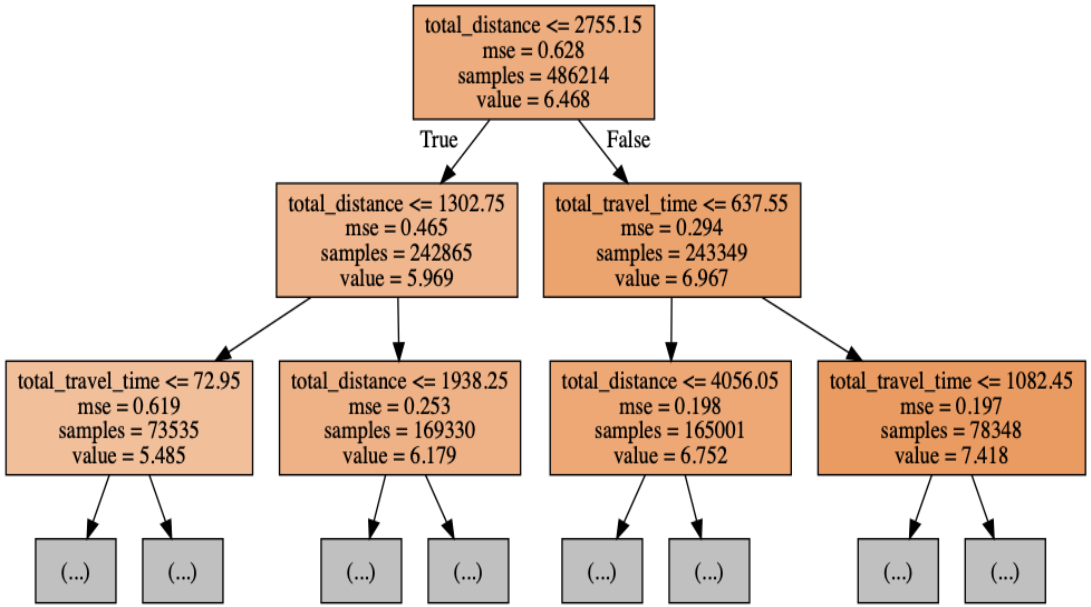



```
linreg_scores = cv_score(LinearRegression())
```

```
1 of kfold 5
Valid RMSE: 0.61369

2 of kfold 5
Valid RMSE: 0.61481

3 of kfold 5
Valid RMSE: 0.61963

4 of kfold 5
Valid RMSE: 0.61161

5 of kfold 5
Valid RMSE: 0.61786
```

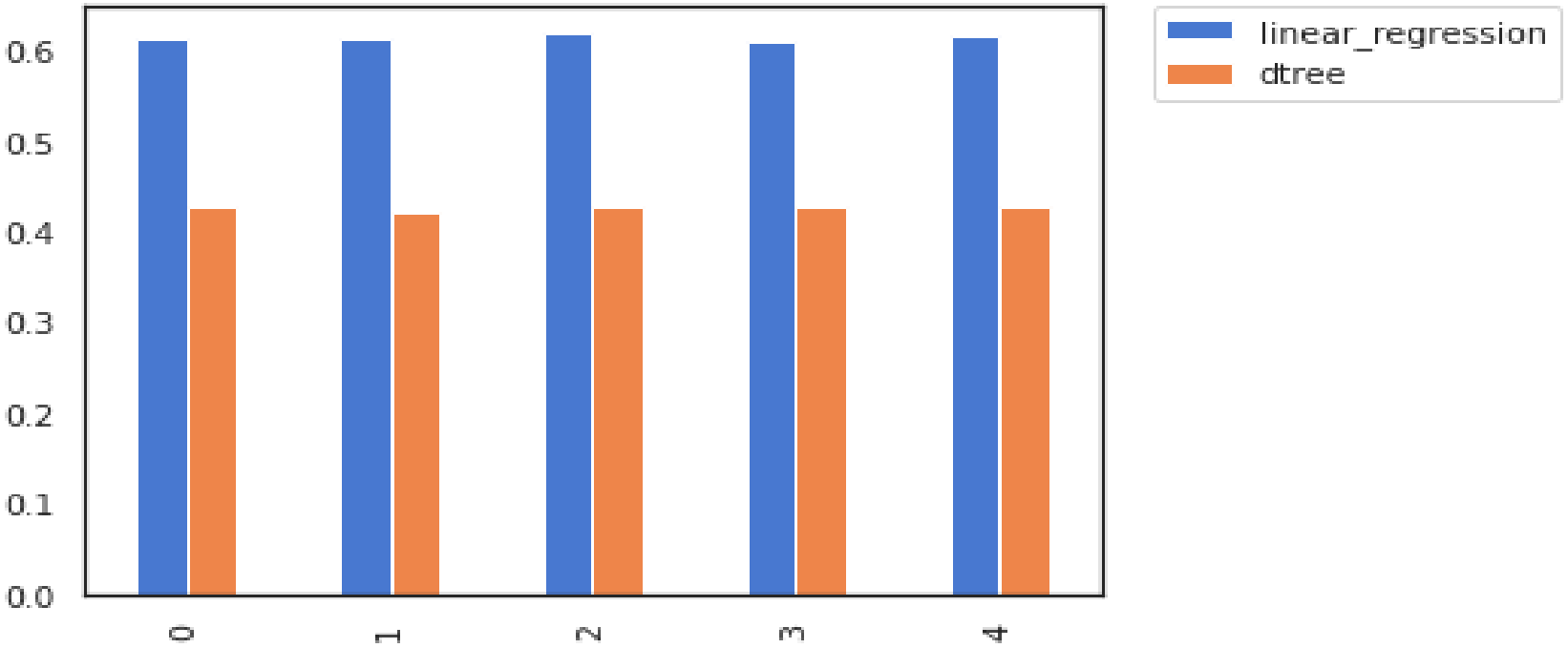
Decision Tree

Decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.




```
1 of kfold 5
Valid RMSE: 0.42935

2 of kfold 5
Valid RMSE: 0.42351

3 of kfold 5
Valid RMSE: 0.42965

4 of kfold 5
Valid RMSE: 0.42823

5 of kfold 5
Valid RMSE: 0.42852
```

Comparison Between Linear Regression And Decision Tree



Conclusions

1. Mostly 1 or 2 passengers avail the cab. The instance of a large group of people travelling together is rare.
2. Most trips were taken on Friday and Monday being the least.
3. We observe that most pickups and drops occur in the evening. While the least drops and pickups occur during morning.
4. The highest average time taken to complete a trip are for trips started in midday(between 14 and 17 hours) and the least are the ones taken in the early morning(between 6–7 hours)
5. vendor 1 mostly provides short trip duration cabs while vendor 2 provides cab for both short and long trips.
6. The flag was stored only for short duration trips and for long duration trips the flag was never stored.
7. Decision tree model gives more accurate value on data set than linear regression model.

Thank You....