Capstone Project

NYC TAXI TRIP TIME PREDICTION

Individual Project

Shubham Naik

Presentation Overview

- Problem statement
- Introduction
- Data Information
- EDA & feature engineering
- Preparing Data for modeling
- Implementing Model
- Model summary

Problem Statement

A typical taxi company faces a common problem of efficiently assigning the cabs to passengers so that the service is smooth and hassle free. One of main issue is determining the duration of the current trip so it can predict when the cab will be free for the next trip.

Introduction

The data set contains the data regarding several taxi trips and its duration in New York City. I will now try and apply different techniques of Data Analysis to get insights about the data and determine how different variables are dependent on the target variable **Trip Duration.**

Data Information

- **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

Basic Exploration

- The data set with 1458644 rows and 11 columns. There are 10 features and 1 target variable which is **trip_duration**
- The columns **id** and **vendor_id** are nominal.
- The columns **pickup_datetime** and **dropoff_datetime** are stored as object which must be converted to datetime for better analysis.
- The column **store_and_fwd_flag** is categorical
- There are no numerical columns with missing data
- The passenger count varies between 1 and 9 with most people number of people being 1 or 2
- The trip duration varying from 1s to 1939736s~538 hrs. There are some outliers present.

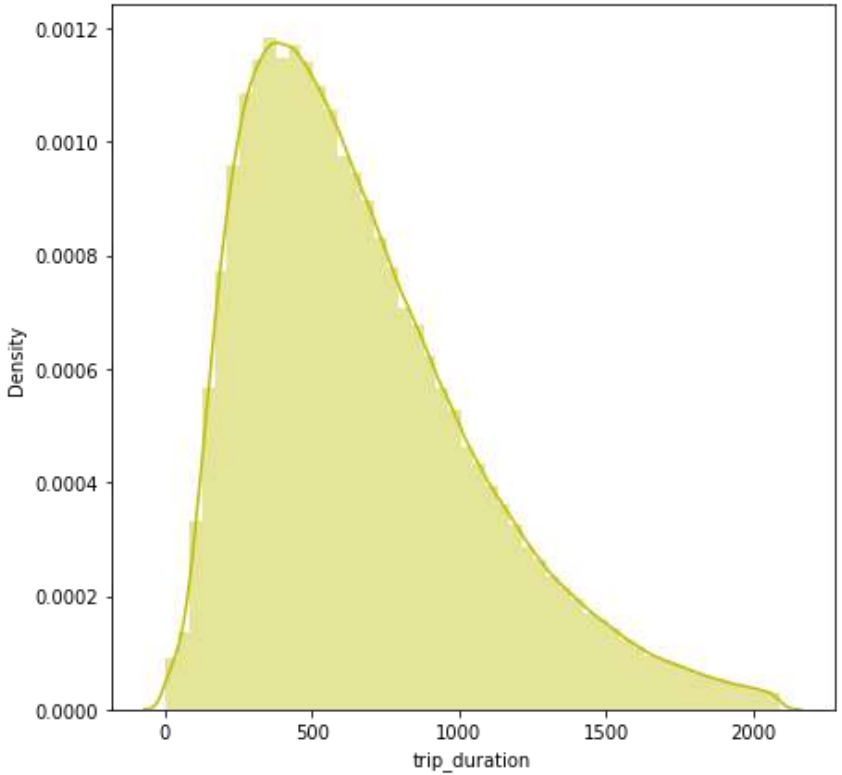
Univariate Analysis

(Lets have a look at the distribution of various variables in the Data set)

EDA

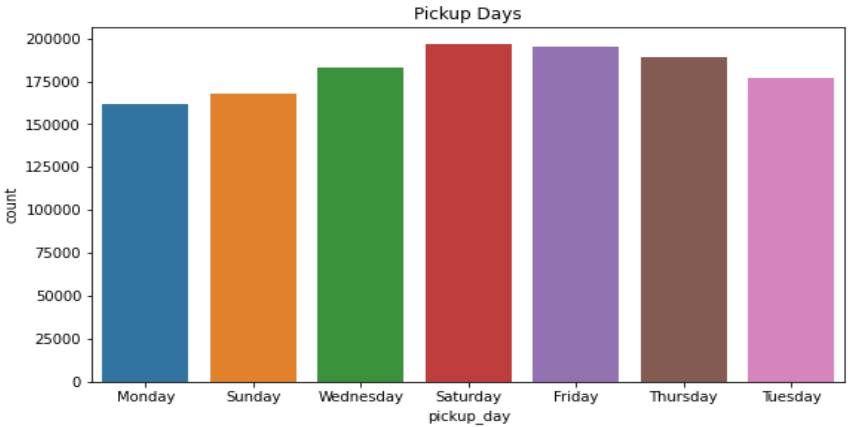
- Target Variable

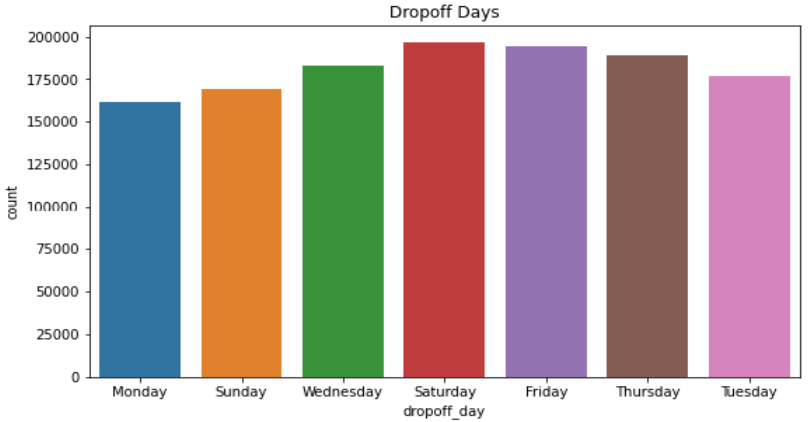
Let us start by analyzing the target variable.



EDA

- The distribution of the number of pickups and drop offs done on each day of the week

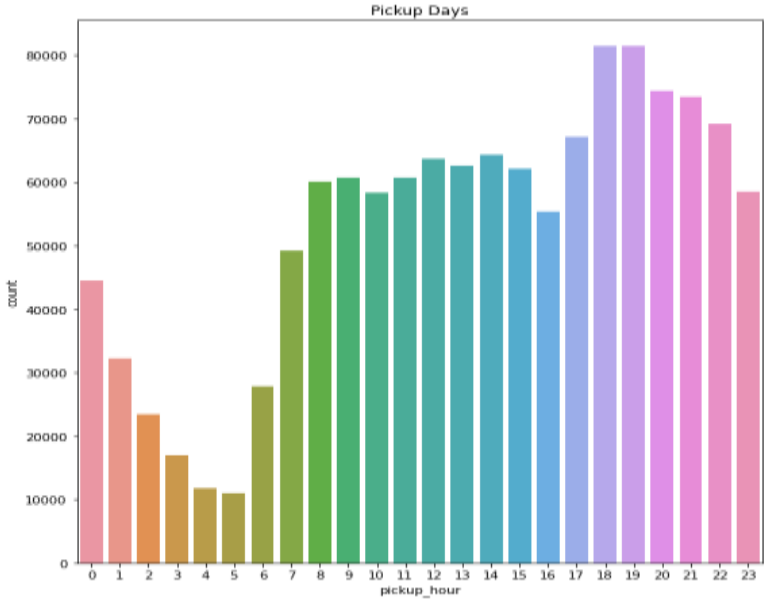


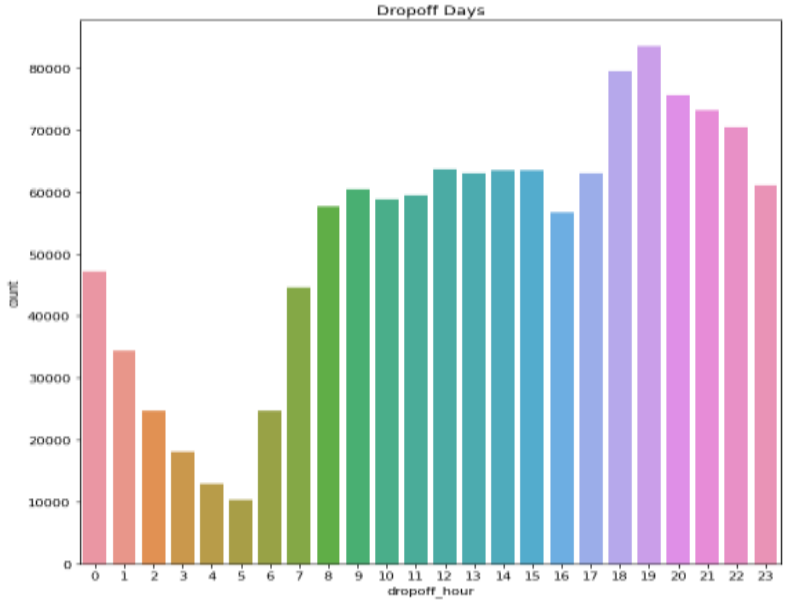


We see Saturdays are the busiest days followed by Fridays. That is probably because it's weekend.

EDA

- The distribution of number of pickups and drop offs done on each hour of the day

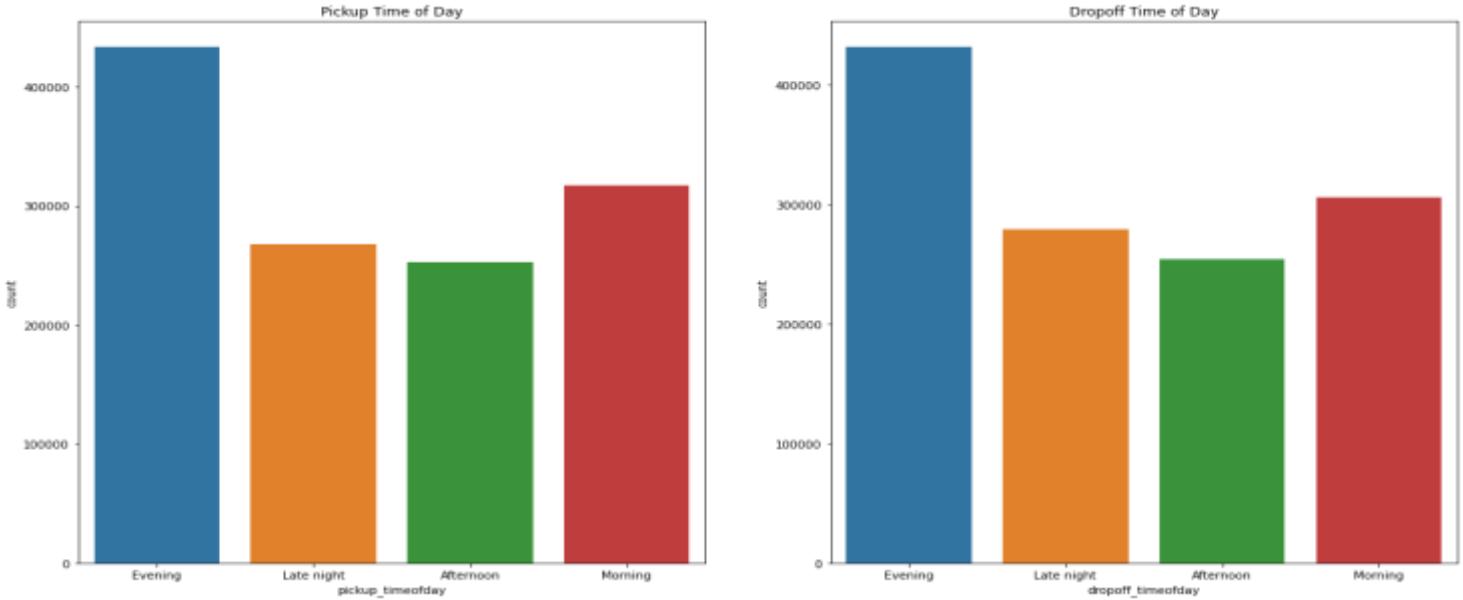




We see the busiest hours are 6:00 pm to 7:00 pm and that makes sense as this is the time when people return from their offices.

EDA

- Lets look at the distribution of the timezones

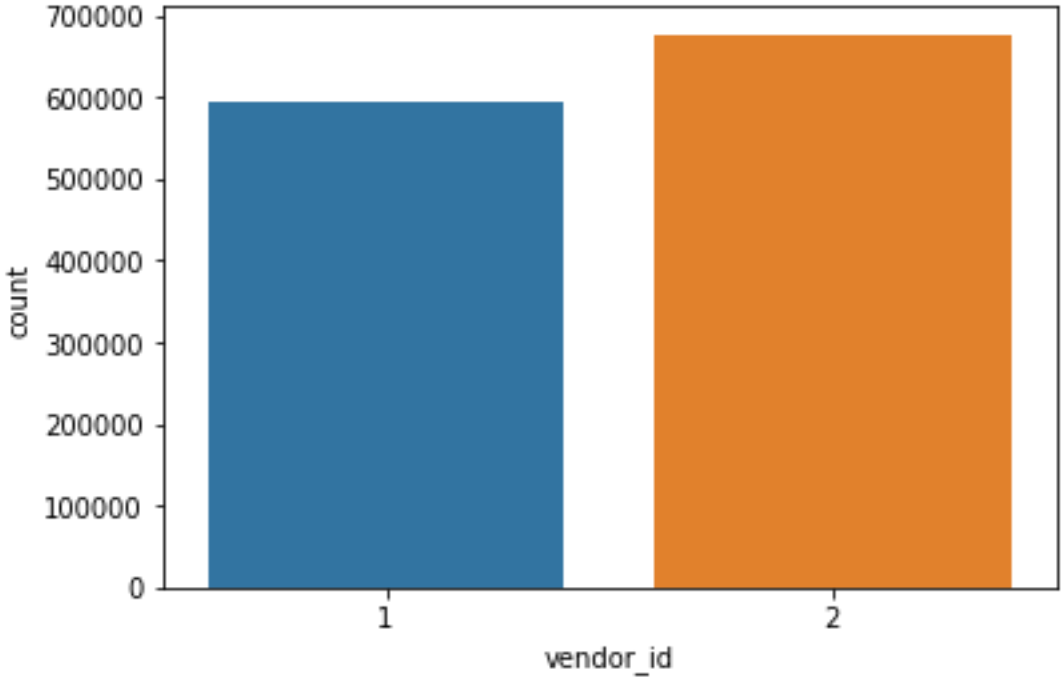


Thus we observe that most pickups and drops occur in the evening. While the least drops and pickups occur during afternoon.

EDA

- Vendor id

We see that there is not much difference between the trips taken by both vendors.

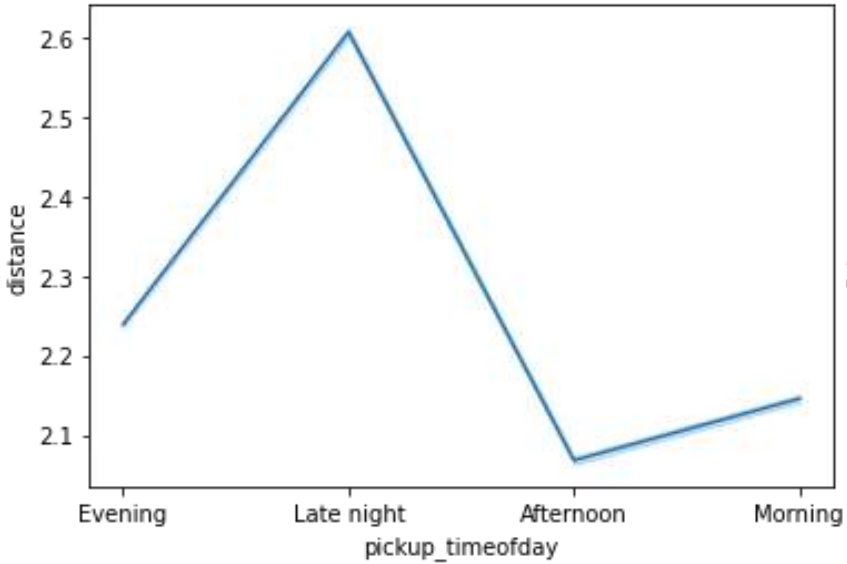


Bivariate Analysis

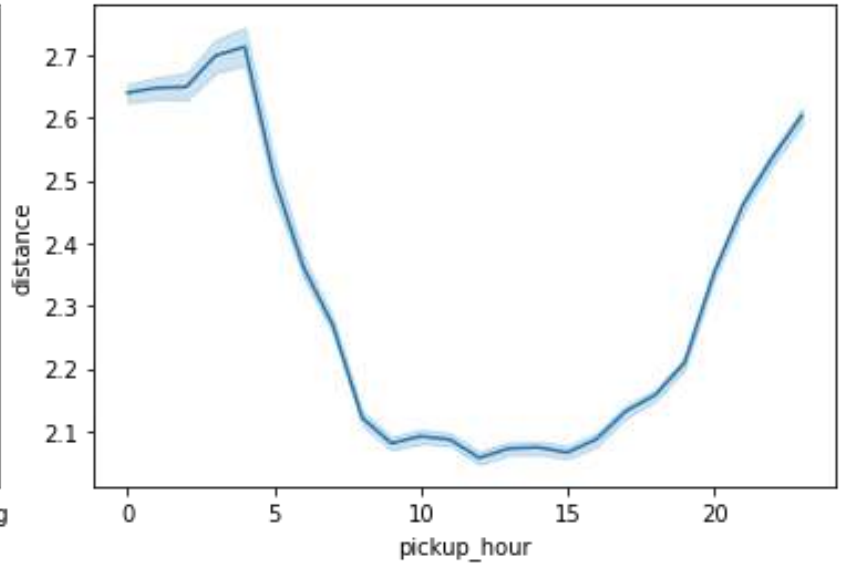
(Bivariate Analysis involves finding relationships, patterns, and correlations between two variables.)

EDA

1. Distance per time of day



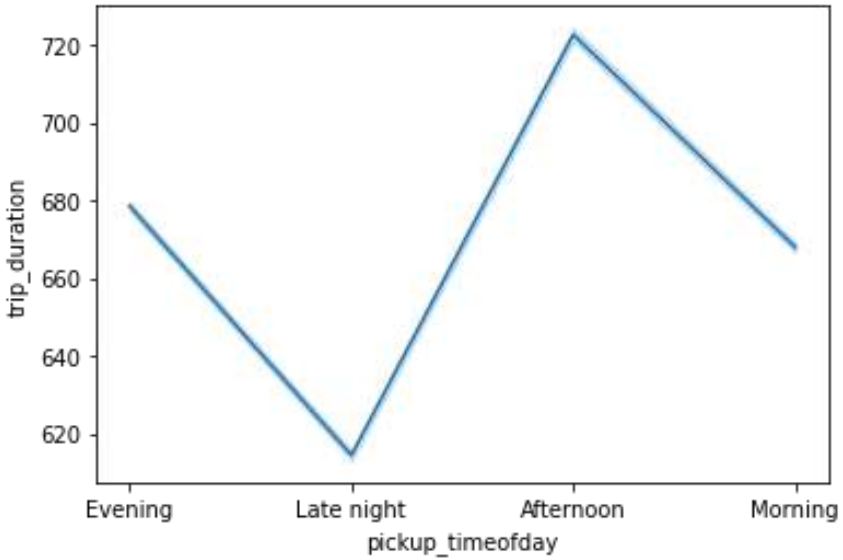
2. Distance per hour of day



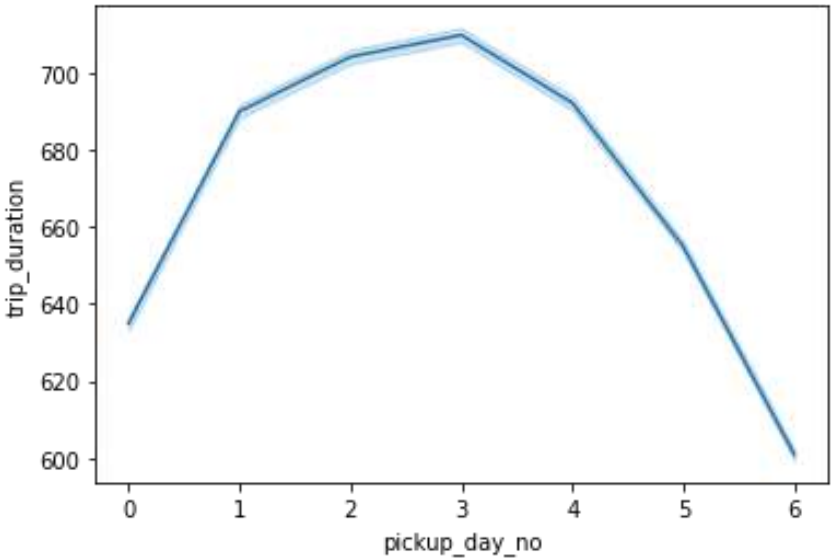
1. As seen above, distances being the longest during late night or it maybe called as early morning too. This can probably point to outstation trips where people start early for the day..
2. Distances are the longest around 5 am.

EDA

1. Trip Duration per time of day



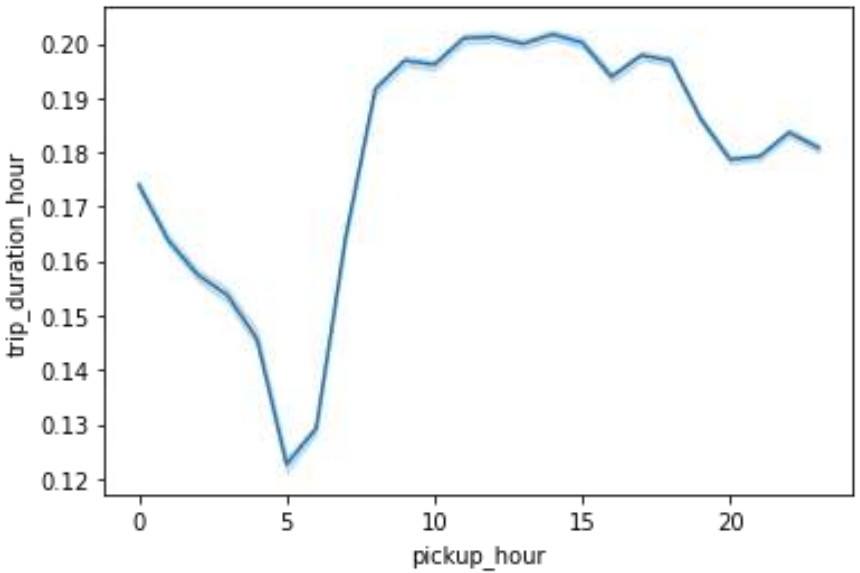
2. Trip Duration per Day of Week



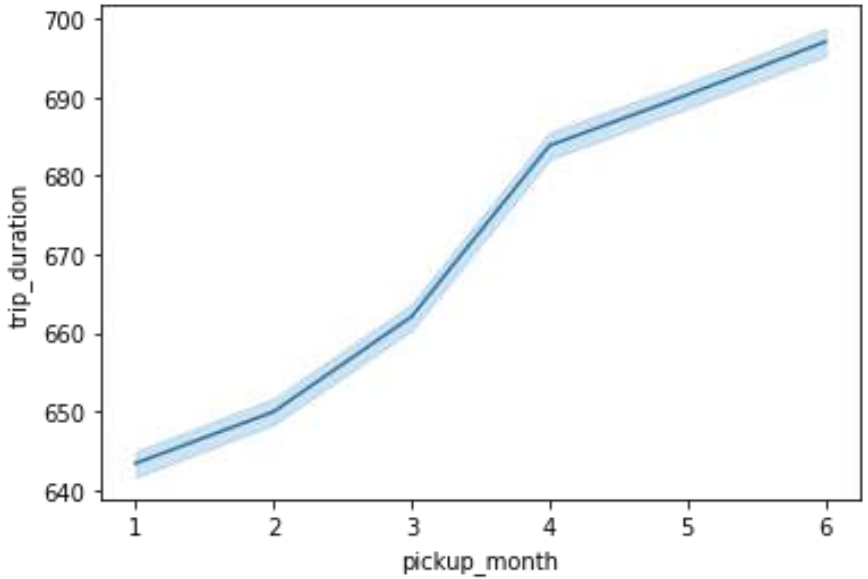
1. As we saw above, trip duration is the maximum in the afternoon and lowest between late night and morning.
2. Trip duration is the longest on Thursdays closely followed by Wednesday.

EDA

1. Trip Duration per hour



2. Trip Duration per month



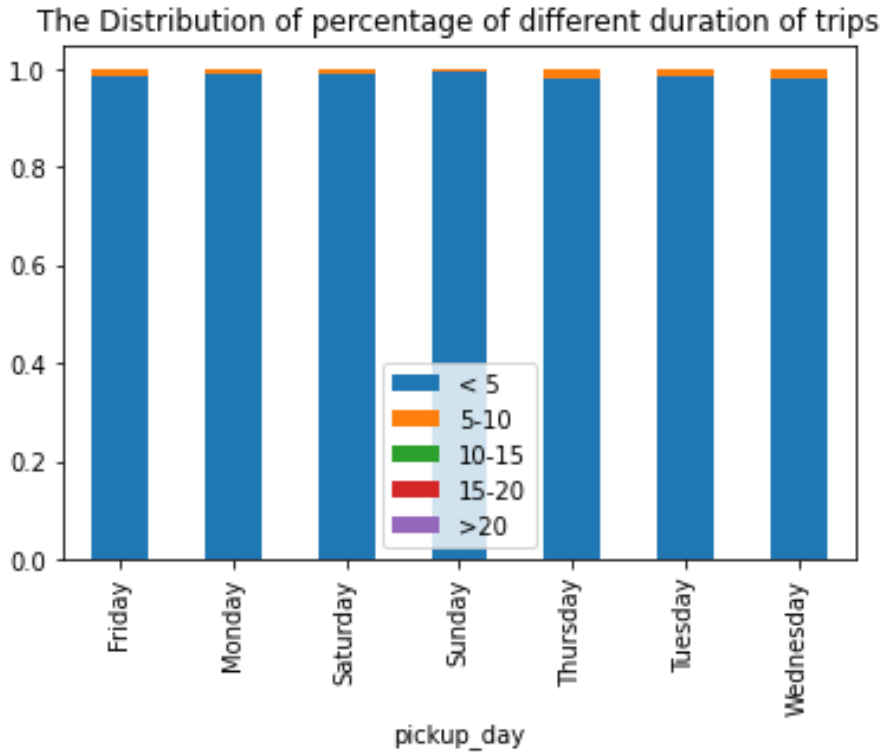
1. We see the trip duration is the maximum around 11 am to 3 pm which may be because of traffic on the roads.
   Trip duration is the lowest around 5 am as streets may not be busy.
2. We can see trip duration rising every month.

EDA

- The percentage of short, medium and long trips taken on each day.

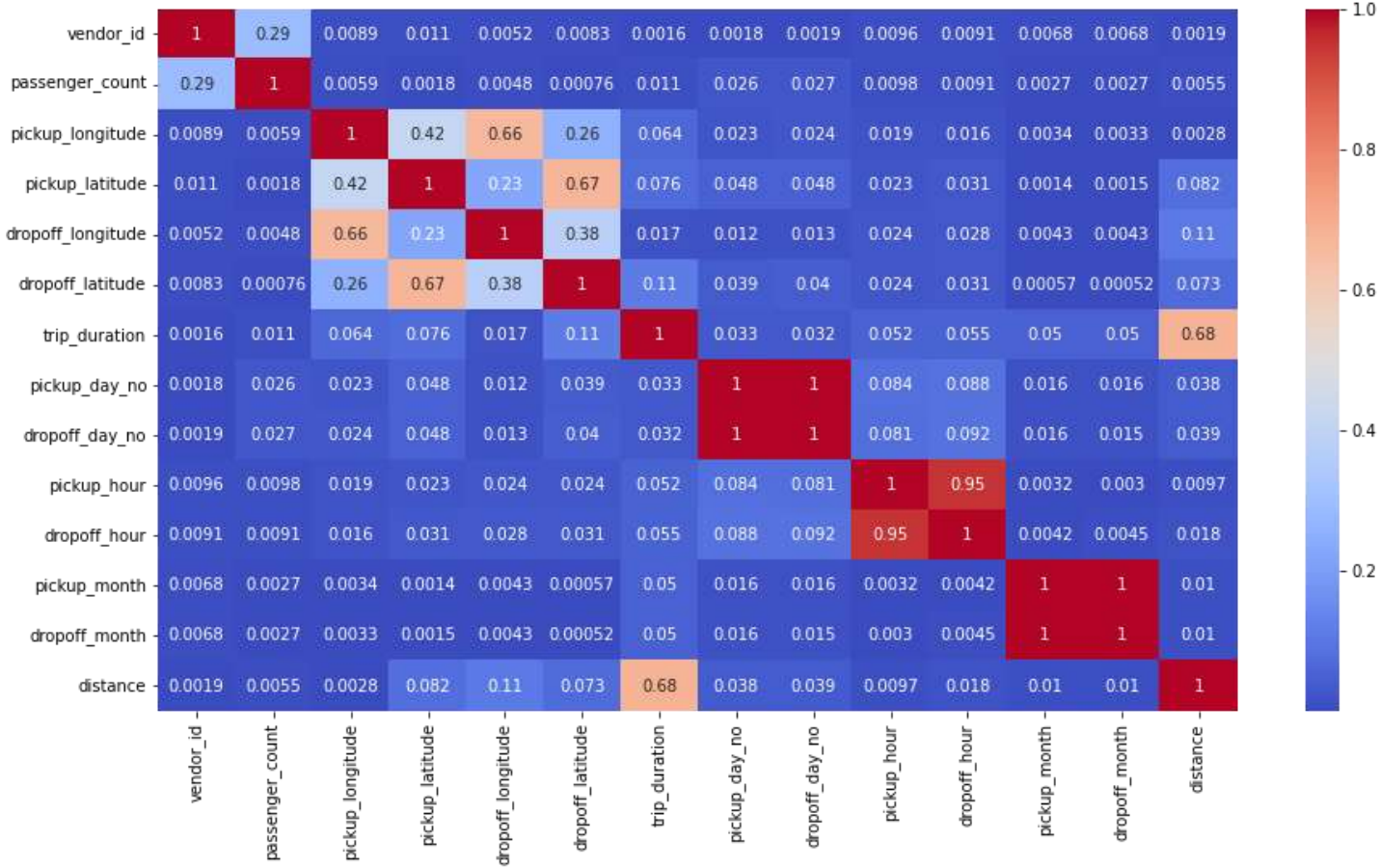
The graph shows a percentage distribution of the trips of different duration within each day of the week.

This does not give much insights as the number of trips within 0–5 hours range is much larger for all the days,



EDA

- **Correlation Heatmap**



Applying Model

- **Linear Regression**

Train set metrics

```
Train MSE : 0.006199840920671342
Train RMSE : 0.07873906857889126
Train R2 : 0.4918684760543993
Train Adjusted R2 :  0.4918609789385082
```

Test set metrics

```
Test MSE : 0.006189820436435968
Test RMSE : 0.07867541189238203
Test R2 : 0.49231066550807
Test Adjusted R2 :  0.49228070178550487
```

- **Decision Tree**

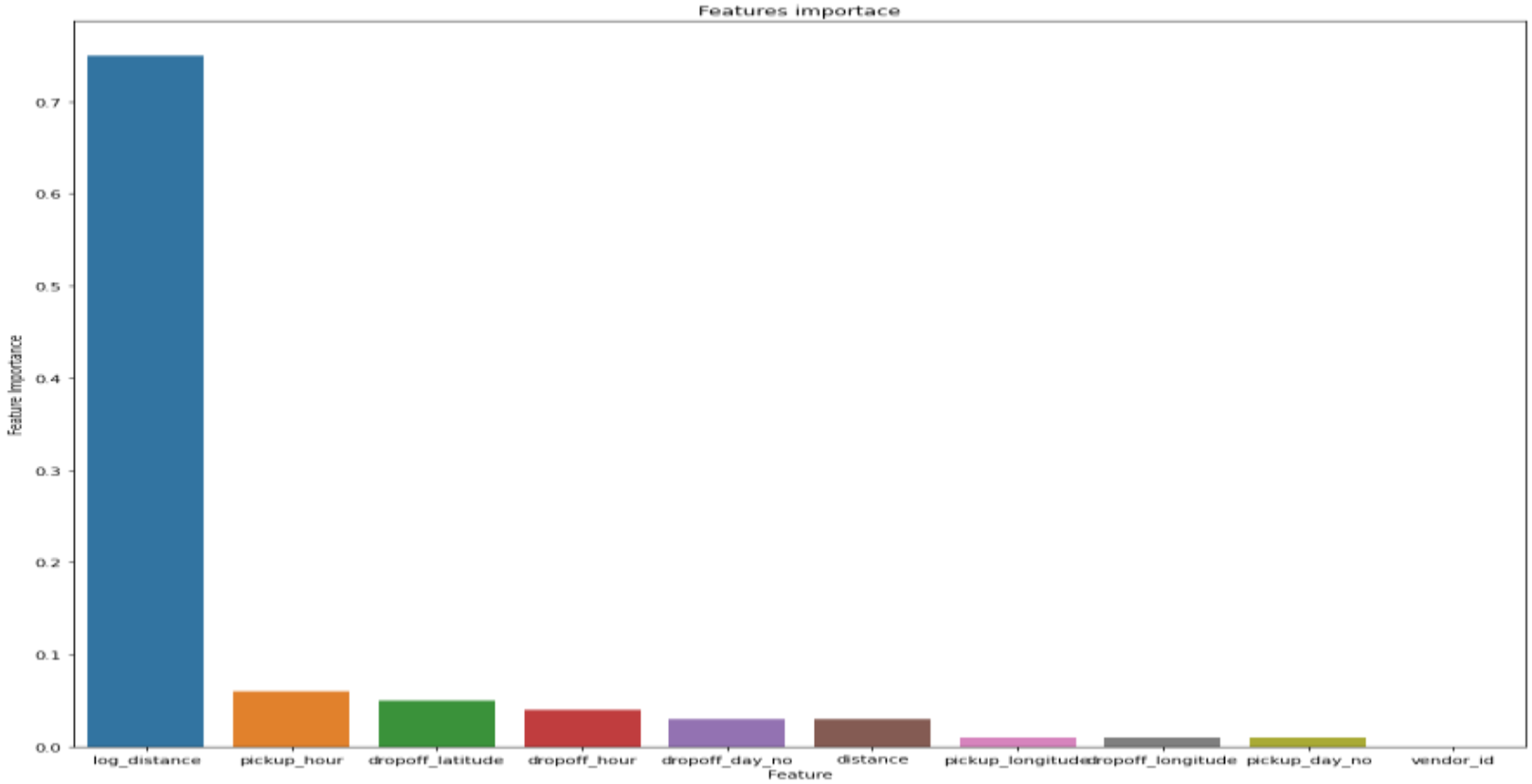
Train set metrics

```
Train MSE : 0.004583084545277018
Train RMSE : 0.06769848259213067
Train R2 : 0.6243758889686206
Train Adjusted R2 :  0.6243703469044912
```

Test set metrics

```
Test MSE : 0.004628479192147505
Test RMSE : 0.06803292726428509
Test R2 : 0.6203719405269017
Test Adjusted R2 :  0.6203495349550703
```

Decision Tree Feature Importance



Applying Model

XGBOST

Train set metrics

```
Train MSE : 0.003030188422020945
Train RMSE : 0.05504714726505766
Train R2 : 0.7516493922303591
Train Adjusted R2 :  0.7516457279954114
```

Test set metrics

```
Test MSE : 0.0032717891891673677
Test RMSE : 0.0571995558476407
Test R2 : 0.7316477120614684
Test Adjusted R2 :  0.7316318739633261
```

- **Best Hyperparameters :**

- gamma=0
- learning_rate=0.1
- max_depth=9
- min_sample _leaf=50
- min_sample_split=40
- n_estimators=120

Applying Model

- Gradient Boosting

Train set metrics

```
Train MSE : 0.002938125440098341
Train RMSE : 0.054204478044699786
Train R2 : 0.7591947637813188
Train Adjusted R2 :  0.7591912108729121
```

Test set metrics

```
Test MSE : 0.0031977801687127455
Test RMSE : 0.05654891836907887
Test R2 : 0.7377179350553104
Test Adjusted R2 :  0.73770245522051
```

- **Best Hyperparameters :**

- alpha=0.9
- max_depth=10
- min_sample _leaf=50
- min_sample_split=80
- n_estimators=120

Final metrics conclusion

S.NO	MODEL_NAME	Train MSE	Train RMSE	Train R^2	Train Adjusted R^2
1	Linear Regression	0.006199840920671342	0.07873906857889126	0.4918684760543993	0.4918609789385082
2	DecisionTree Regressor	0.004583084545277018	0.06769848259213067	0.6243758889686206	0.6243703469044912
3	XGBRegressor	0.003030188422020945	0.05504714726505766	0.7516493922303591	0.7516457279954114
4	GradientBoosting	0.002938125440098341	0.054204478044699786	0.7591947637813188	0.7591912108729121

S.NO	MODEL_NAME	Test MSE	Test RMSE	Test R^2	Test Adjusted R^2
1	Linear Regression	0.006189820436435968	0.07867541189238203	0.49231066550807	0.49228070178550487
2	DecisionTree Regressor	0.004628479192147505	0.06803292726428509	0.6203719405269017	0.6203495349550703
3	XGBRegressor	0.0032717891891673677	0.0571995558476407	0.7316477120614684	0.7316318739633261
4	GradientBoosting	0.0031977801687127455	0.05654891836907887	0.7377179350553104	0.73770245522051

Challenges

- Handling Large Dataset.
- Feature Engineering.
- Computation Time.
- Optimizing the Model.

Conclusion

- Trip Duration varies a lot ranging from few seconds to more than 20 hours
- Most trips are taken on Friday , Saturday and Thursday
- The average duration of a trip is most on Thursday and Friday as trips longer than 5 hours are mostly taken in these days
- The average duration of trips started in between 14 hours and 17 hours is the largest.
- Vendor 2 mostly provides the longer trips
- The long duration trips(> 5 hours) are mostly concentrated with their pickup region near (40 °,75 °) to (42°,75°)

Thank You!