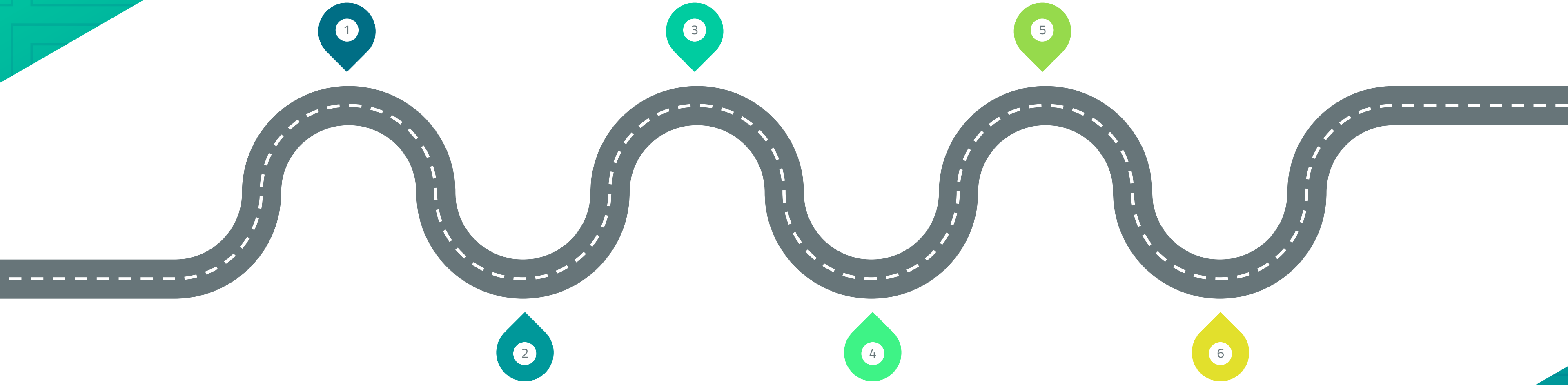
Capstone Project NYC taxi trip time Prediction

Supervised Learning Regression

ROADMAP



1 Introduction, Problem Statement & Understanding the Data

2 Data Wrangling , Data Manipulation & Data Visulization

3 Hypothesis Testing & Feature Engineering

4 Data Pre-Processing & Model Implementaion

5 Competitive Advantage & Future Work

6 Conclusion

1

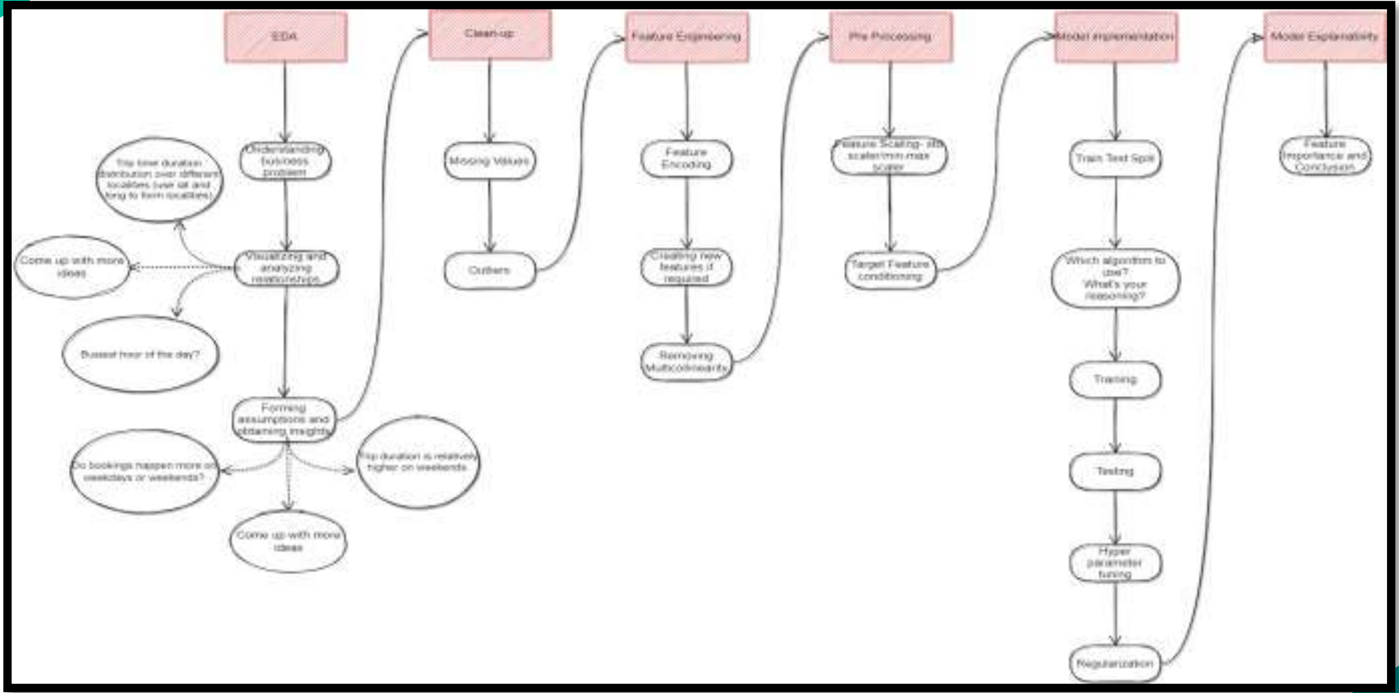
Introduction

Information about NYC taxi trips

INTRODUCTION

The NYC taxi trip time prediction project involves building a machine learning model to predict the time it takes for a taxi to travel from one location to another in New York City. The model takes into account various factors such as traffic conditions, weather time of day, and origin-destination pairs to make its predictions. To build the model, a large dataset of historical taxi trip records is used to train the model. This dataset includes information such as pickup and drop-off locations, time of day, and trip duration. Various machine learning algorithms, such as regression and RandomForest, can be applied to the data to build a predictive model. Once the model is trained, it can be used to predict the trip time for new, unseen taxi trips in NYC. This information can be useful for taxi drivers and passengers, as well as for transportation planning and optimization. The results of the project can also be used to understand the factors that influence taxi trip times in NYC, such as traffic patterns, weather conditions, and time of day. This information can be used to make improvements to the city's transportation infrastructure and to develop more efficient transportation systems.

Project Architecture



EDA

Understanding business problem

Trip time duration distribution over different localities (use lat and long to form localities)

Come up with more ideas

Visualizing and analyzing relationships

Busiest hour of the day?

Forming assumptions and obtaining insights

Do bookings happen more on weekdays or weekends?

Trip duration is relatively higher on weekends

Come up with more ideas

Clean-up

Missing Values

Outliers

Feature Engineering

Feature Encoding

Creating new features if required

Removing Multicolinearity

Pre Processing

Feature Scaling- std scaler/min max scaler

Target Feature conditioning

Model implementation

Train Test Split

Which algorithm to use?
What's your reasoning?

Training

Testing

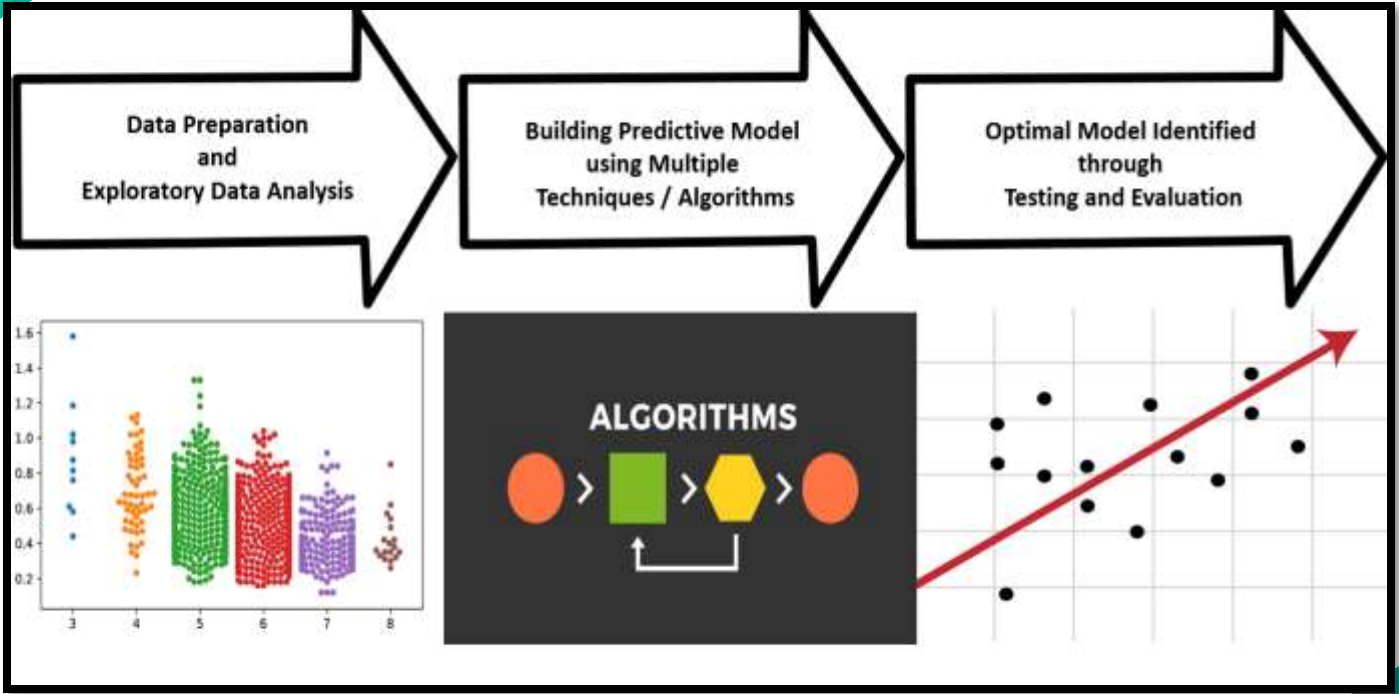
Hyper parameter tuning

Regularization

Model Explainability

Feature Importance and Conclusion

Approach



Data Preparation
and
Exploratory Data Analysis

Building Predictive Model
using Multiple
Techniques / Algorithms

Optimal Model Identified
through
Testing and Evaluation

2

Problem Statement

How we predict the target variable

BUSINESS PROBLEM OVERVIEW

The problem statement for the NYC taxi trip time prediction project is to accurately predict the time it takes for a taxi to travel from one location to another in New York City. The objective is to develop a machine learning model that takes into account various factors such as time of day, and origin-destination pairs to make its predictions.

The challenge lies in capturing the complex relationships between the various factors that influence taxi trip times and accurately predicting the trip duration for any given trip.

The solution to this problem will have practical applications for taxi drivers and passengers, as well as for transportation planning and optimization. Accurate taxi trip time predictions can help drivers plan their routes more effectively and reduce the time and cost of travel for passengers. It can also be used to improve the city's transportation infrastructure and to develop more efficient transportation systems.

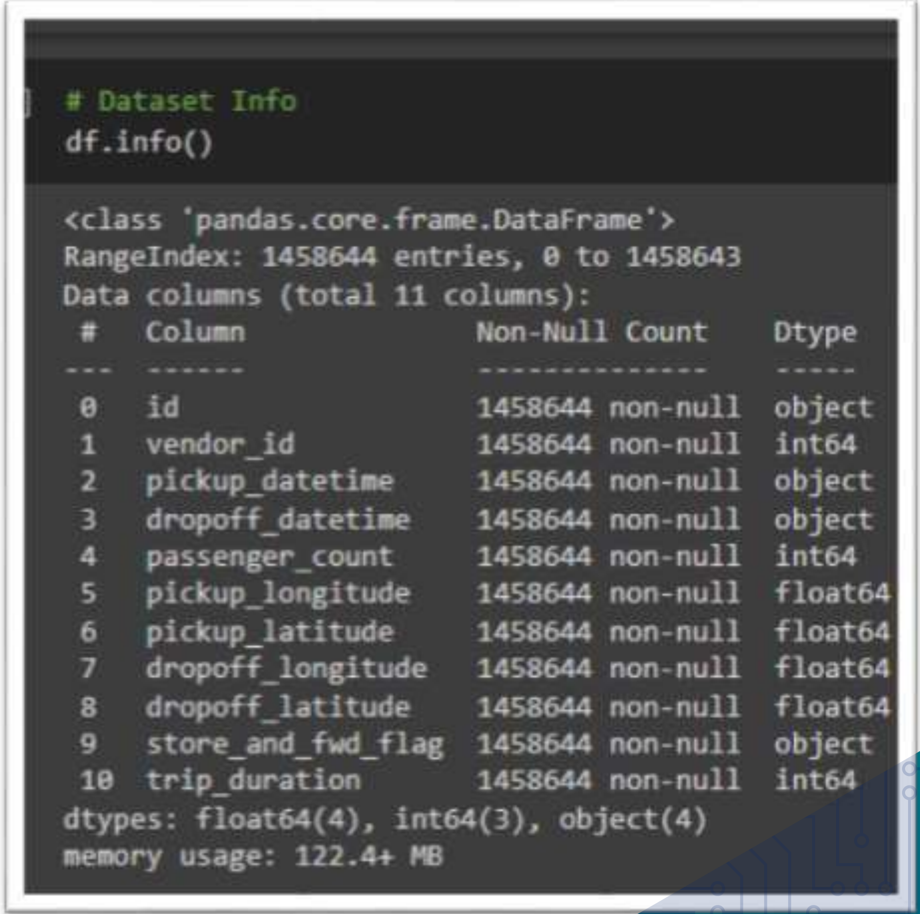
3

Understanding Data

What are the feature and labels in our data

Understanding Data

In our dataset of NYC taxi trip time prediction, there are **1458644 rows** and **11 columns** with **zero null values** and **zero duplicate value**



```
# Dataset Info
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1458644 entries, 0 to 1458643
Data columns (total 11 columns):
 #   Column              Non-Null Count    Dtype
---  ------              --------------    -----
 0   id                  1458644 non-null  object
 1   vendor_id           1458644 non-null  int64
 2   pickup_datetime     1458644 non-null  object
 3   dropoff_datetime    1458644 non-null  object
 4   passenger_count     1458644 non-null  int64
 5   pickup_longitude    1458644 non-null  float64
 6   pickup_latitude     1458644 non-null  float64
 7   dropoff_longitude   1458644 non-null  float64
 8   dropoff_latitude    1458644 non-null  float64
 9   store_and_fwd_flag  1458644 non-null  object
 10  trip_duration       1458644 non-null  int64
dtypes: float64(4), int64(3), object(4)
memory usage: 122.4+ MB
```

Variables Description

Continuous numerical: id, pickuplongitude, pickuplatitude, dropofflongitude, dropofflatitude, trip_duration
Discrete numerical: vendorid, passenger_count, pickup_datetime_hour, dropoff_datetime_hour
Datetime: pickupdatetime, dropoffdatetime
Nominal categorical: store_and_fwd_flag, pickup_datetime_day, pickup_datetime_month, dropoff_datetime_day, dropoff_datetime_month

Fields	Description
id	A unique identifier for each trip
vendor_id	A code indicating the provider associated with the trip rec
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 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

4

Data Wrangling (EDA)

Finding meaningful insights

Overall Insights

- There are only 2 vendors , vendor 1 and 2, vendor 2 has **780302** while vendor 1 has **678342** trip respectively
- 10 unique number of passenger count. with passenger count 1 has most trips **1033540**
- There are 1450599 times trip data Store and forwarded and 8045 time trip data not stored and forwarded
- we convert the datatype of pickup_datetime and drop-off datetime from object to datetime
- Also we mapped the stored and forwarded flag from Y/N to 1/0 which is now converted into int64 i.e. numeric data that we require further
- Our dependent variable is trip duration
- We introduce a new column say **trip distance** which will give us the distance covered during a trip
- Longest trip distance in km is **1240.91 km**

Overall Insights

- **7935 trips** where distance covered by taxi is 0.00 km
- **240459 trips** where distance covered by taxi is between zero and 1 km
- **450488 trips** where distance covered by taxi is between 1 and 2 km
- **497469 trips** where distance covered by taxi is between 2 and 5 km
- **262293 trips** where distance covered by taxi is greater 5 km
- We introduce a new column say Average speed of a journey in kmph
- In peak hours (between 6 to 10 pm) traffic/ trips are **77.18%** more than off peak hours (between 2 am to 6 am)
- Busiest hour in a day is between **6 to 7 pm**

Overall Insights

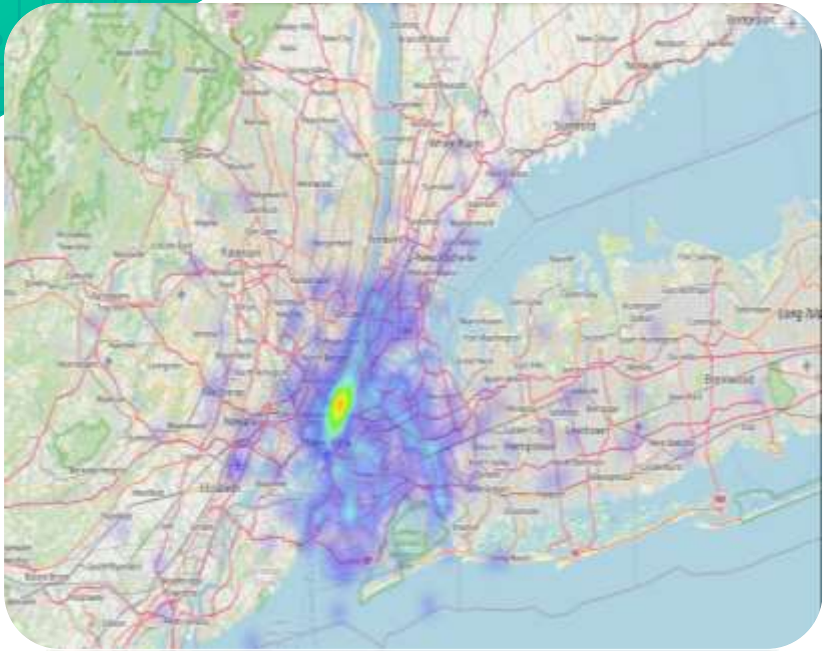
- The **busiest day** for taxi trips is: **2016-04-09** i.e. **9th April 2016** The number of trips on that day is: **9796**
- The number of trips on busiest day in peak_hour is **2095**
- The **PEAK hours** is between 6 pm to 10 pm where on an average **1918** number of ongoing trips per day in Network City
- The **OFF-PEAK hours** is between 2 to 6 am where on an average **438** number of ongoing trips per day in New York City
- during peak hours trips are **67 % less on weekends** as compare to weekdays
- during off-peak hours trips are **12 % more on weekends** as compare to weekdays
- Average trip duration during weekdays :**16 minutes 12 seconds**
- Average trip duration during weekends :**15 minutes 26 seconds**
- Average speed during weekdays :**13.87 kmph**
- Average speed during weekends :**15.81 kmph**

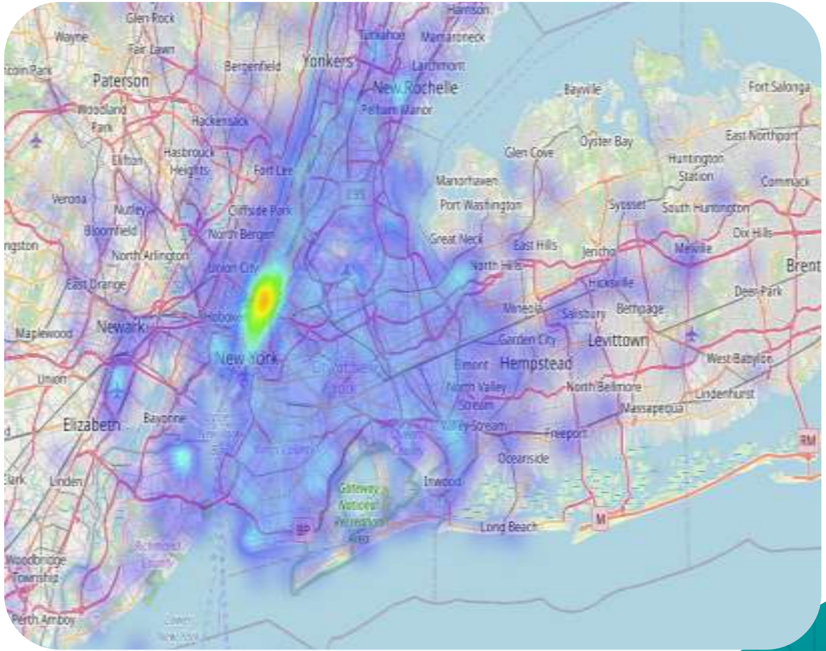
5

Data Visualization & Business Insights

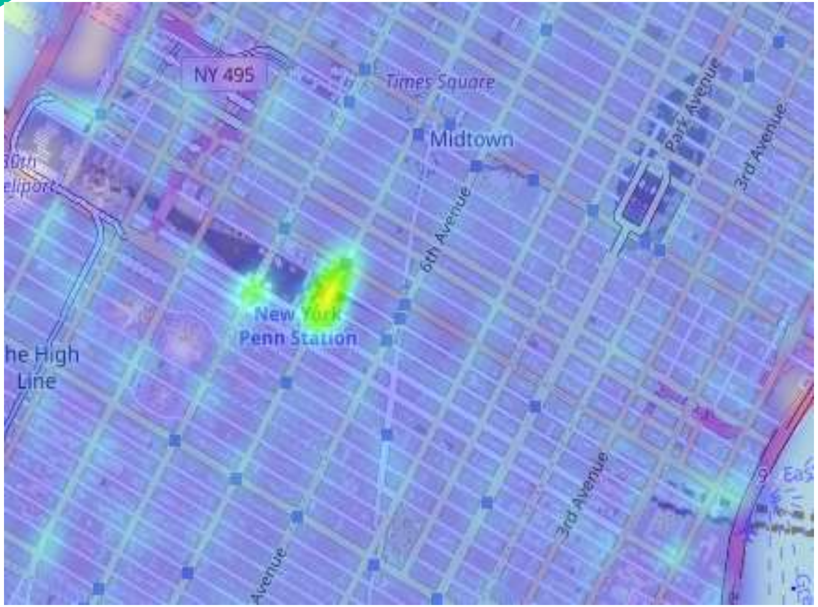
Data Visualization using different charts & find meaningful insights

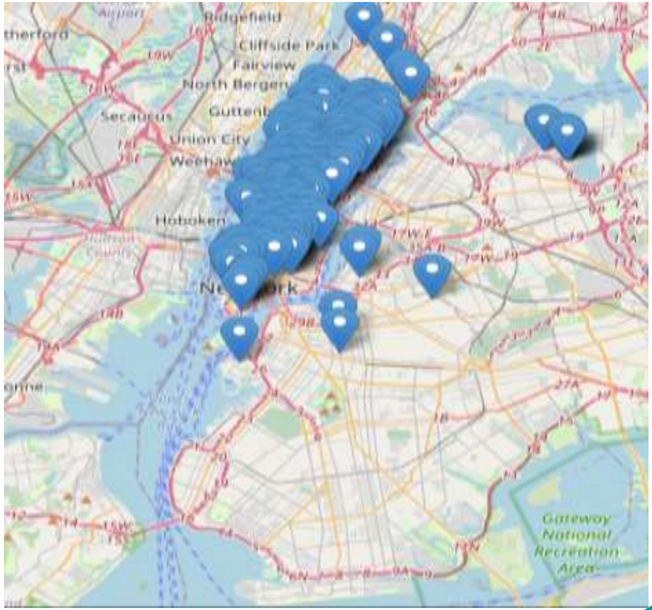
Heatmap of pickup and droppoff locations



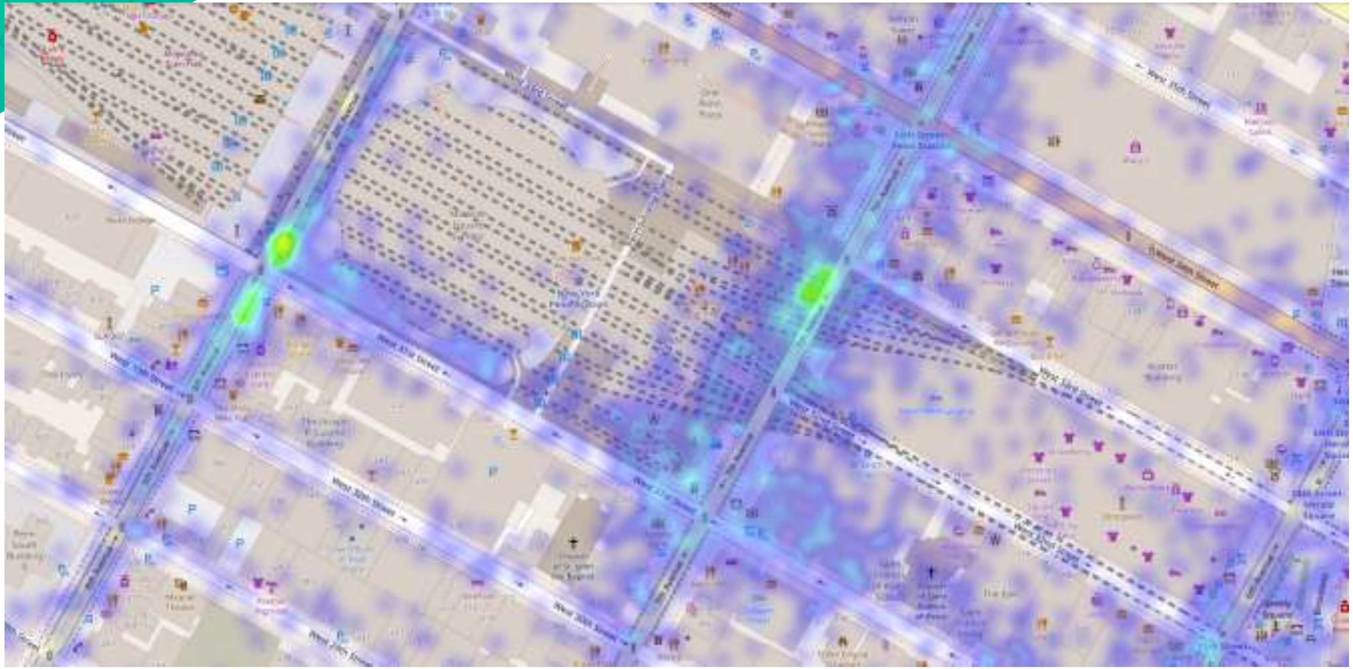


Pickup markers & hottest place for pickup





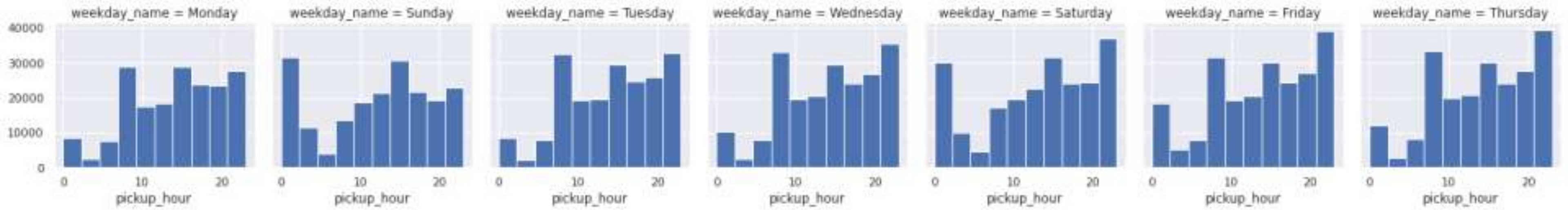
Most pickup Locations in NYC



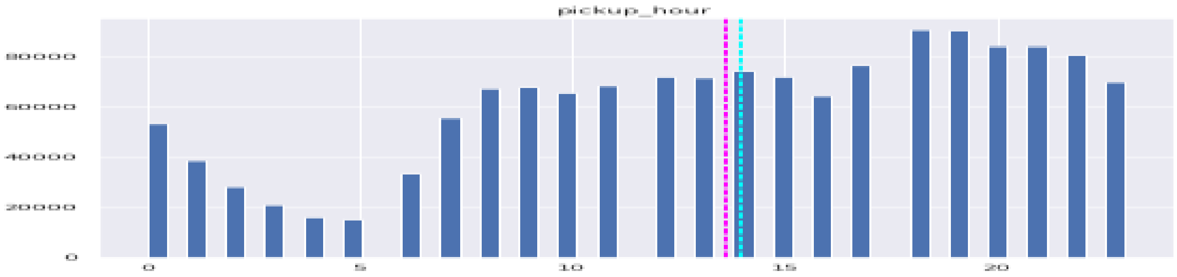
pickups are generally from NYC penn station (outstation) mainly on 7th & 8th Avenue street & intersection between 8th avenue & west 31st Street, for 250K samples of our data

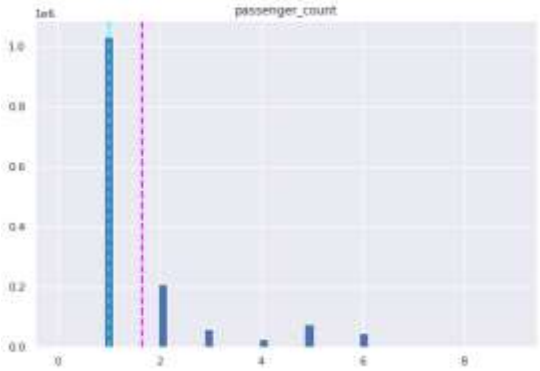
Pickup Hours

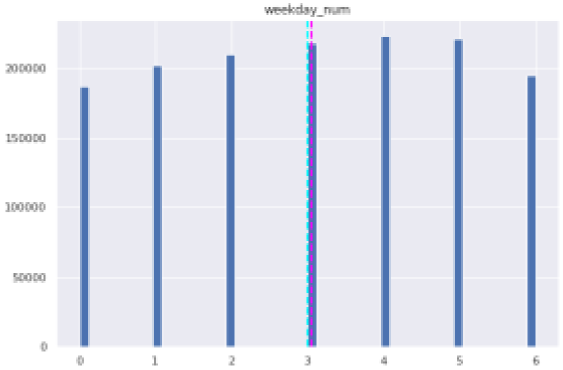
- Taxi pickups increased in the late night hours over the weekend possibly due to more outstation rides or for the late night leisure's nearby activities.
- Early morning pickups i.e. before 5 AM have increased over the weekend in comparison to the office hours pickups i.e. after 7 AM which have decreased due to obvious reasons.
- Taxi pickups seems to be consistent across the week at 15 Hours i.e. at 3 PM.



Bar Plots


pickup_hour
80000
60000
40000
20000
0
0
5
10
15
20

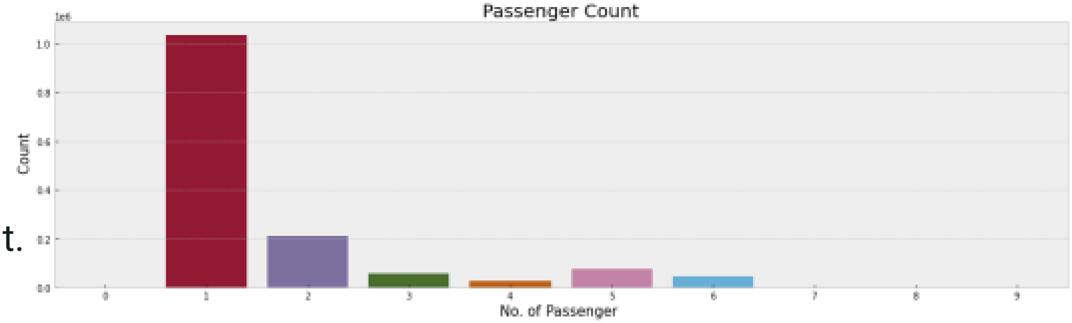

1e6
passenger_count
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8

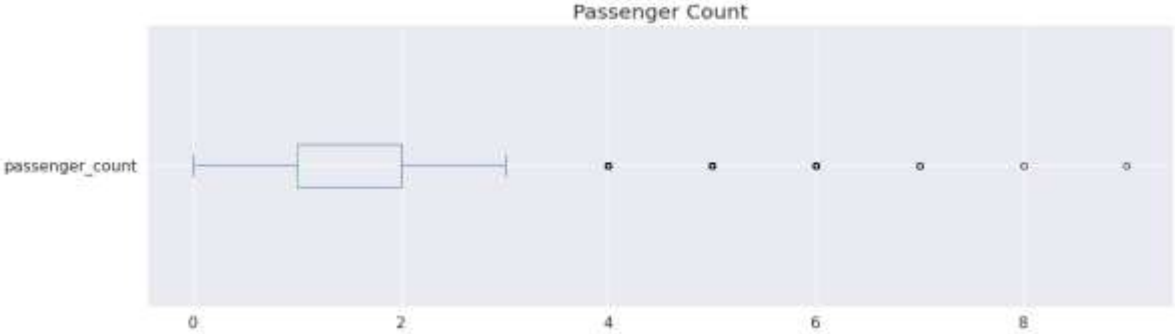

weekday_num
200000
150000
100000
50000
0
0
1
2
3
4
5
6

Passenger Counts

Number of passengers per trip

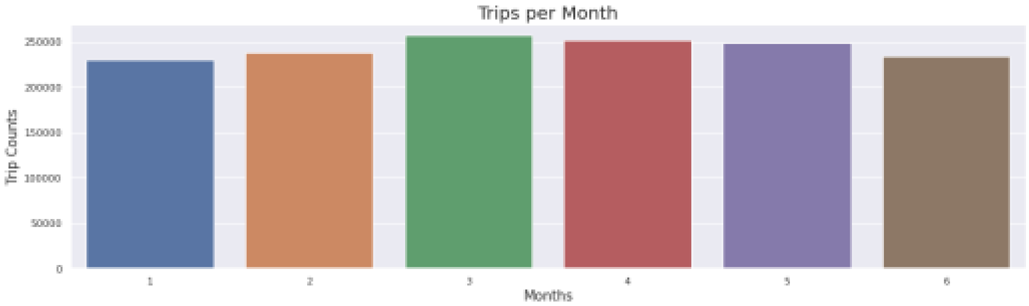
- There are some trips with 0 passenger count.
- Few trips consisted of even 6, 7, 8 or 9 passengers. Clear outliers and pointers to data inconsistency
- Most of trip consist of passenger either 1 or 2.



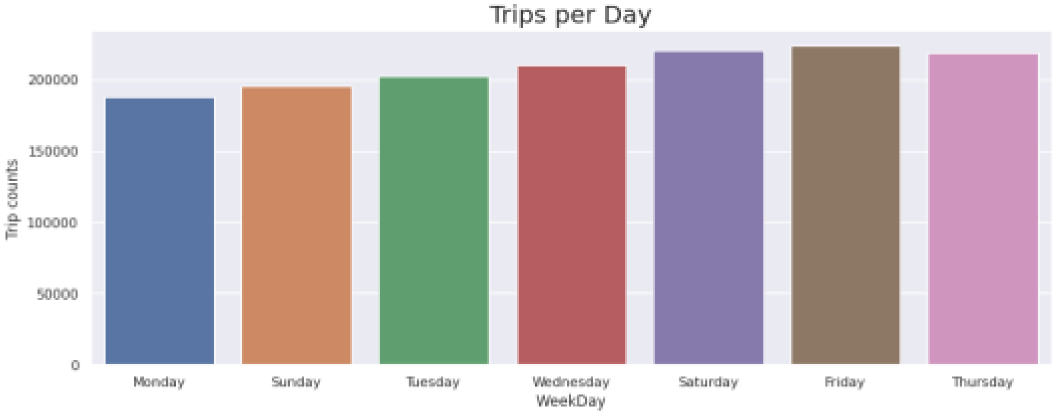


Number of trips

Number of trip per month
Most number of trips are in march and then after April and may

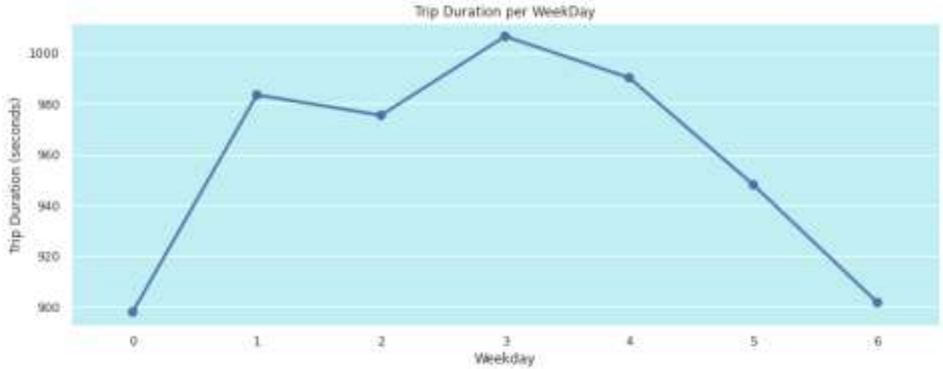


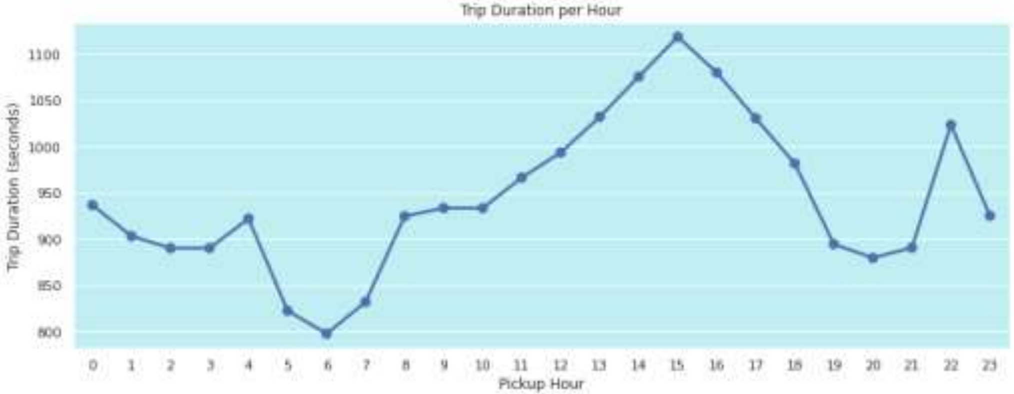
Number of trip per day of week
Here we can see an increasing trend of taxi pickups starting from Monday till Friday.
 The trend starts declining from Saturday till Monday which is normal where some office going people likes to stay at home for rest on the weekends.



Trip Duration

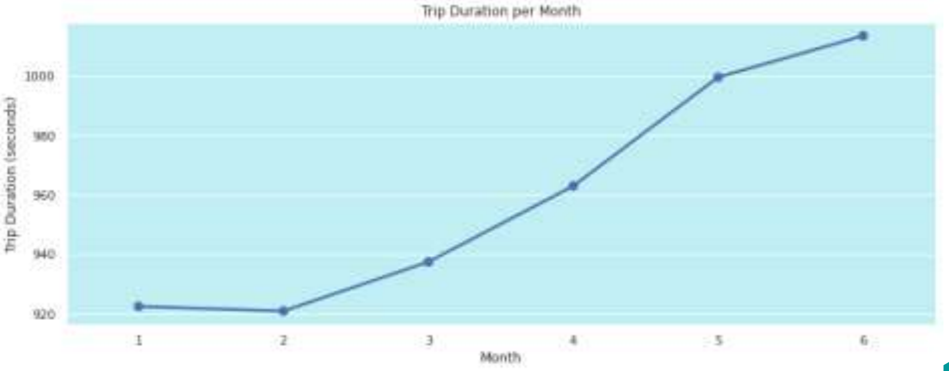
- We can see that trip duration is almost equally distributed across the week on a scale of 0-1000 minutes with minimal difference in the duration times. Also, it is observed that trip duration on thursday is longest among all days
- Average trip duration is lowest at 6 AM when there is minimal traffic on the roads.
- Average trip duration is generally highest around 3 PM during the busy streets.
- Trip duration on an average is similar during early morning hours i.e. before 6 AM & late evening hours i.e. after 6 PM





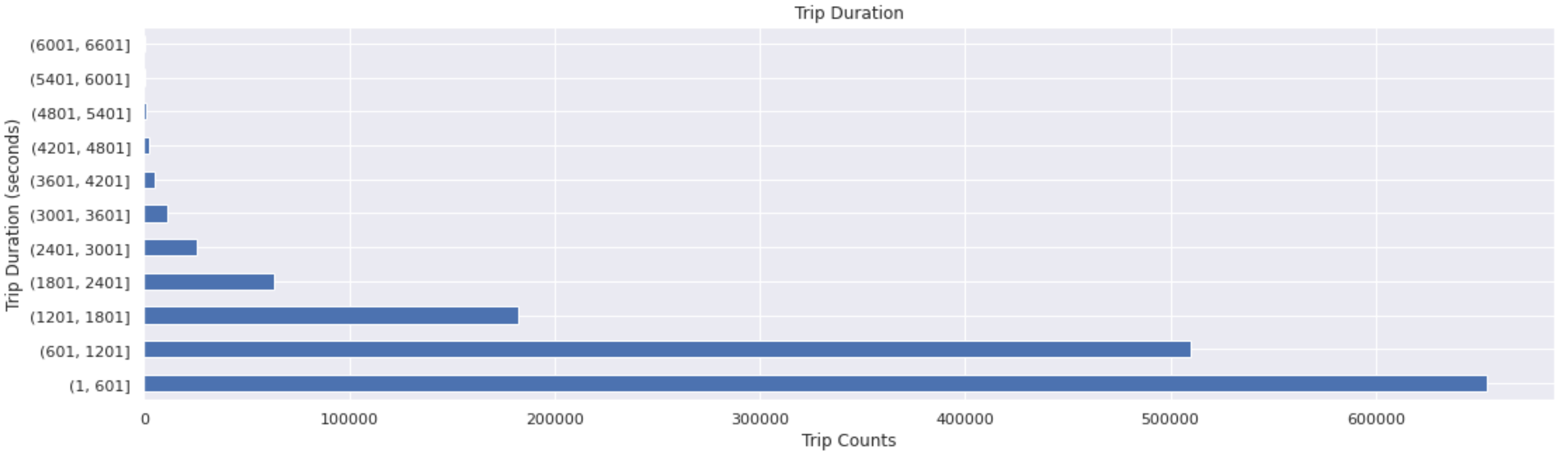
Trip Duration

There is an increasing linear relationship be tween the months and the average value of the duration of trips.



Trip Duration in 10 minute slab

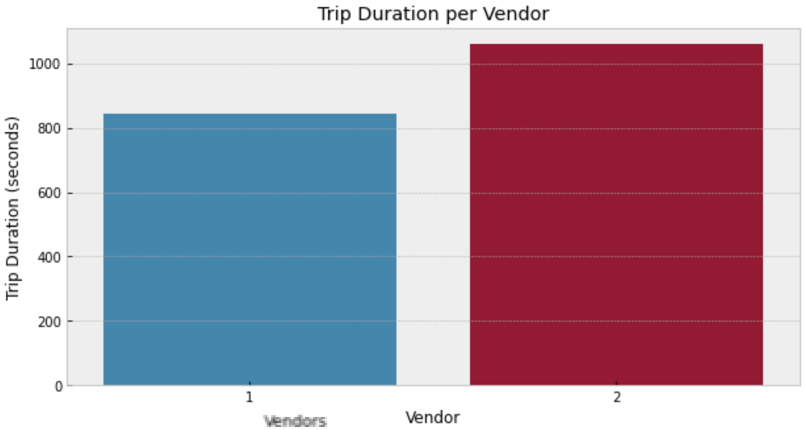
- Some trip durations are over 100000 seconds which are clear outliers and should be removed.
- There are some durations with as low as 1 second. which points towards trips with 0 km distance.
- Major trip durations took between 10-20 mins to complete.
- Mean and mode are not same which shows that trip duration distribution is skewed towards right

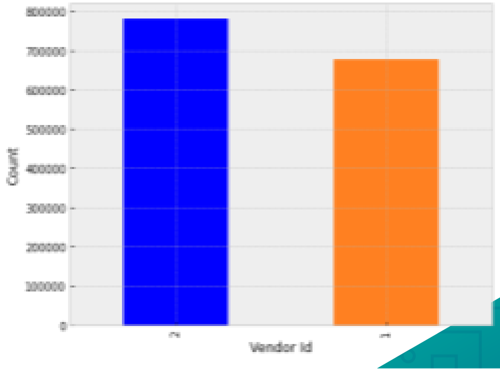


All about Vendor

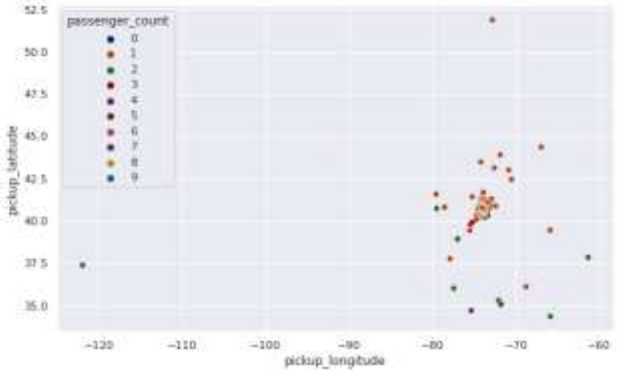
•Vendor 2 takes the crown. Average trip duration for vendor 2 is higher than vendor 1 by a quite low margin

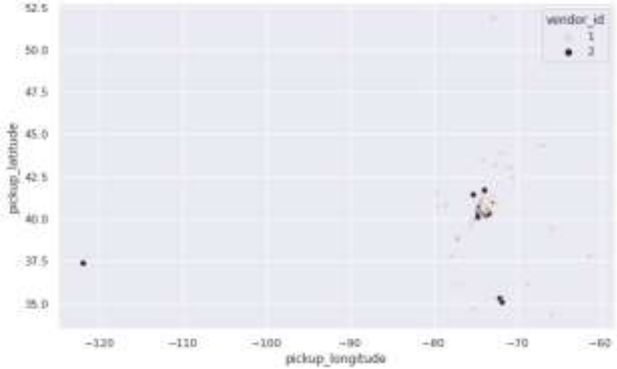
•Both the venders share almost equal amount of trips, the difference is quite low between two venders
•But Vendor 2 is evidently more famous among the population as per the above graphs

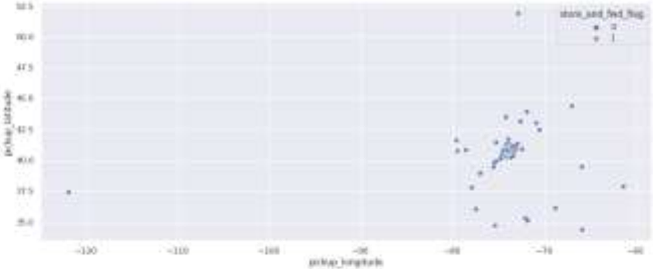


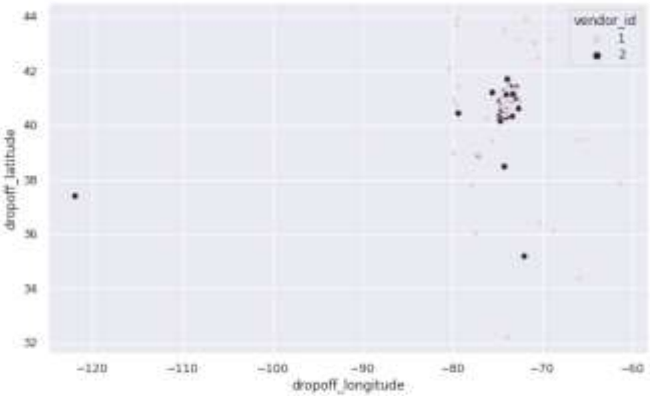


Scatter Plots









6

Hypothesis Testing

Using different scenarios and assumptions validates claim about a population based on sample data.

Hypothesis Test 1

STATEMENT

There is a significant difference in the trip duration between trips with a single passenger and trips with multiple passengers.

Null hypothesis: There is no significant difference in the trip duration between trips with a single passenger and trips with multiple passengers.

Alternative hypothesis: There is a significant difference in the trip duration between trips with a single passenger and trips with multiple passengers.

Result

The **p-value of 9.89e-28** signifies that the probability of observing such extreme differences in trip duration between single passenger trips and multi-passenger trips by chance is very low. Therefore, **we can reject the null hypothesis** and conclude that **there is a significant difference in trip duration between these two groups.**

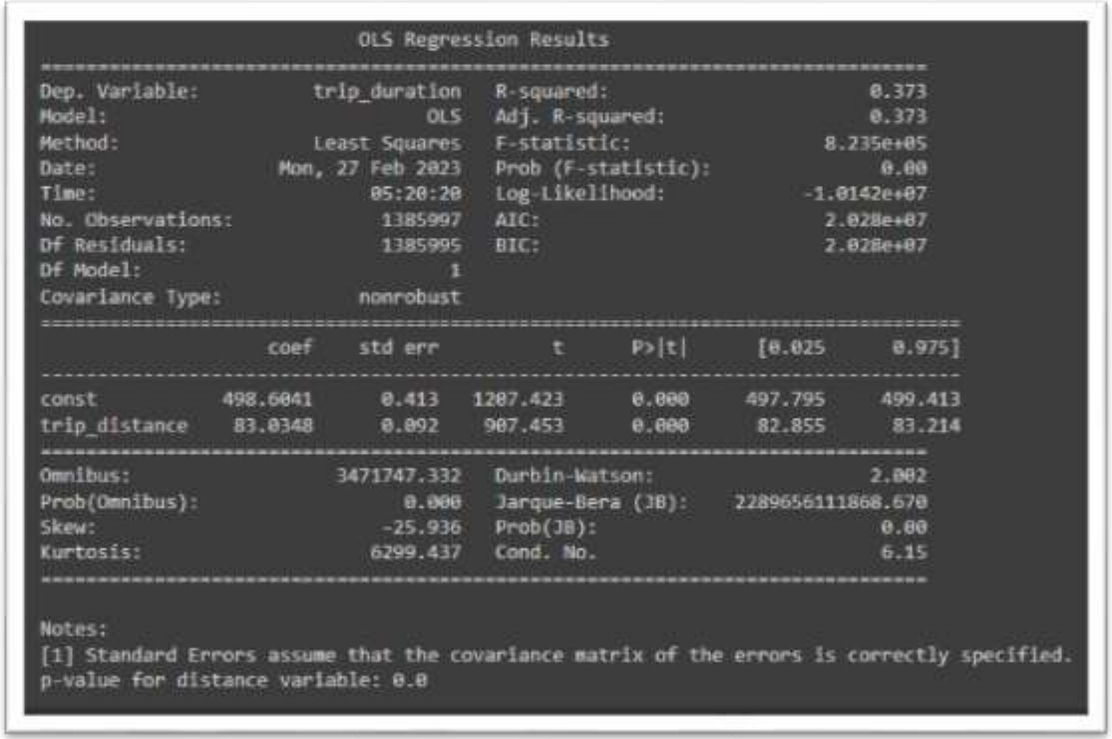
Hypothesis Test 2

STATEMENT
Trips with a higher distance have a longer duration

Null hypothesis: The distance of the trip does not significantly affect the duration of the trip.

Alternative hypothesis: The distance of the trip significantly affects the duration of the trip.

Result



```
                            OLS Regression Results
==============================================================================
Dep. Variable:          trip_duration   R-squared:                       0.373
Model:                            OLS   Adj. R-squared:                  0.373
Method:                 Least Squares   F-statistic:                 8.235e+05
Date:                Mon, 27 Feb 2023   Prob (F-statistic):               0.00
Time:                        05:20:20   Log-Likelihood:            -1.0142e+07
No. Observations:             1385997   AIC:                         2.028e+07
Df Residuals:                 1385995   BIC:                         2.028e+07
Df Model:                           1
Covariance Type:            nonrobust
=================================================================================
                    coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
const           498.6041      0.413   1207.423      0.000     497.795     499.413
trip_distance    83.0348      0.092    907.453      0.000      82.855      83.214
==============================================================================
Omnibus:                  3471747.332   Durbin-Watson:                   2.002
Prob(Omnibus):                  0.000   Jarque-Bera (JB):   2289656111868.670
Skew:                         -25.936   Prob(JB):                         0.00
Kurtosis:                    6299.437   Cond. No.                         6.15
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
p-value for distance variable: 0.0
```

Result

P_value of distance variable 0.0 indicates that the probability of obtaining the observed sample result assuming that null hypothesis is true is extremely small (essentially zero).
Therefore , we **reject Null Hypothesis** and conclude that **there is a significant relationship between the two variables.**

Hypothesis Test 3

STATEMENT:

Trips during weekdays have a longer duration compared to weekends

Null hypothesis: The distance of the trip does not significantly affect the duration of the trip.

Alternative hypothesis: The distance of the trip significantly affects the duration of the trip.

Result

The **t-statistic of 3.9008** indicates that the difference between the means of the two groups (weekday and weekend trip durations) **is 3.9008 times greater than the standard error of the difference between the means.**

The **p-value of 9.586970042130944e-05** (0.00009587) is less than the commonly used significance level of 0.05, which suggests **strong evidence against the null hypothesis.**

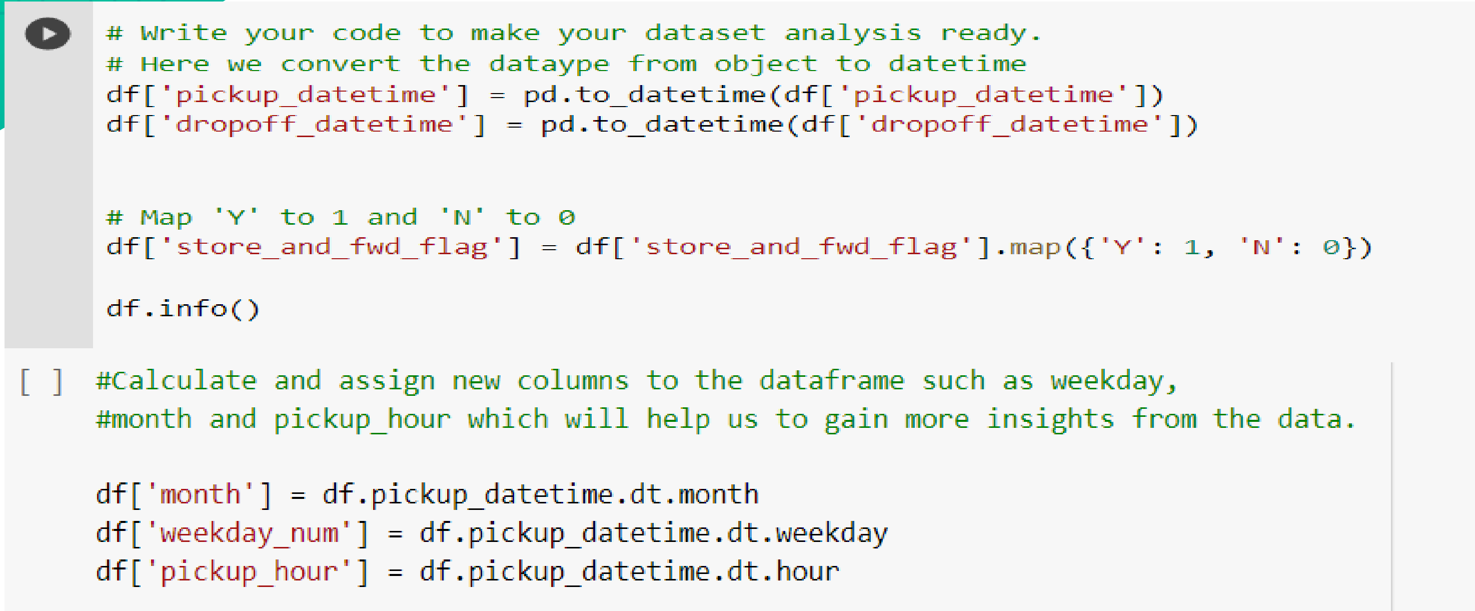
Therefore, **we reject Null hypothesis** and conclude that **there is a statistically significant difference between the means of weekday and weekend trip durations**

7

Data Prepocessing & Feature Engineering

Handling Outliers , Feature manipulation and Selection

Feature Engineering

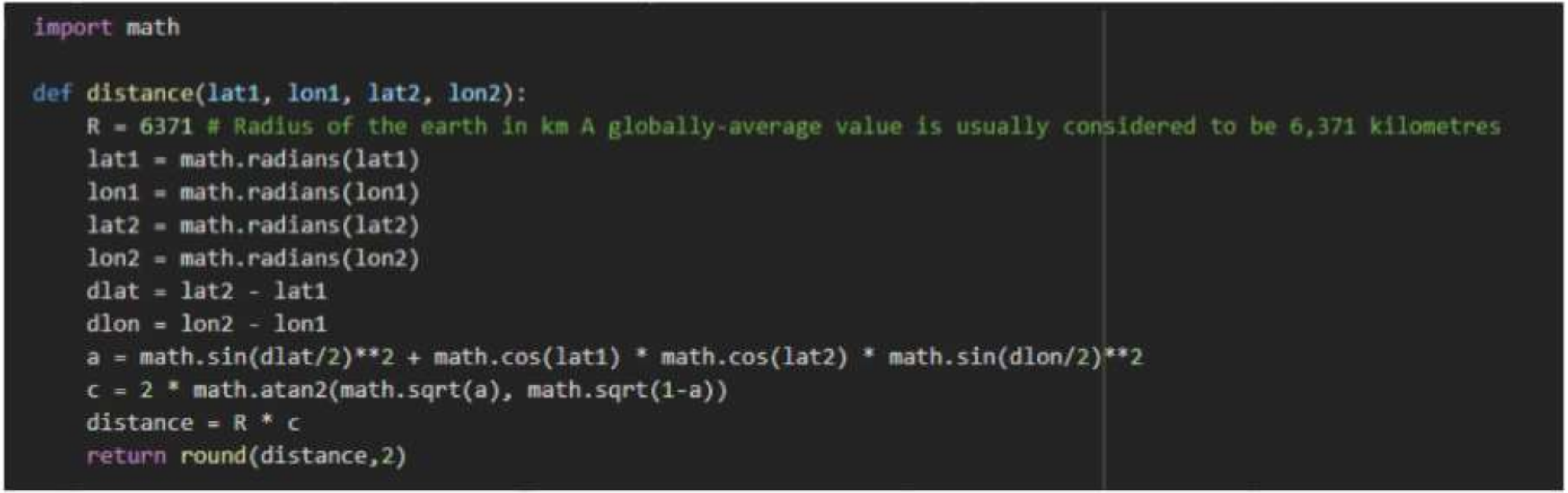


```
# Write your code to make your dataset analysis ready.
# Here we convert the dataype from object to datetime
df['pickup_datetime'] = pd.to_datetime(df['pickup_datetime'])
df['dropoff_datetime'] = pd.to_datetime(df['dropoff_datetime'])


# Map 'Y' to 1 and 'N' to 0
df['store_and_fwd_flag'] = df['store_and_fwd_flag'].map({'Y': 1, 'N': 0})

df.info()
```

```
#Calculate and assign new columns to the dataframe such as weekday,
#month and pickup_hour which will help us to gain more insights from the data.

df['month'] = df.pickup_datetime.dt.month
df['weekday_num'] = df.pickup_datetime.dt.weekday
df['pickup_hour'] = df.pickup_datetime.dt.hour
```

The Date and time columns in the Dataset has whole lot story to tell, we have to fetch them as separate columns. We do not have to fetch pickup and dropoff time both, as they may lead to strong positive correlation in the respective fetched features. Further we can use these columns for Analysis.

Feature Manipulation



```
import math

def distance(lat1, lon1, lat2, lon2):
    R = 6371 # Radius of the earth in km A globally-average value is usually considered to be 6,371 kilometres
    lat1 = math.radians(lat1)
    lon1 = math.radians(lon1)
    lat2 = math.radians(lat2)
    lon2 = math.radians(lon2)
    dlat = lat2 - lat1
    dlon = lon2 - lon1
    a = math.sin(dlat/2)**2 + math.cos(lat1) * math.cos(lat2) * math.sin(dlon/2)**2
    c = 2 * math.atan2(math.sqrt(a), math.sqrt(1-a))
    distance = R * c
    return round(distance,2)
```

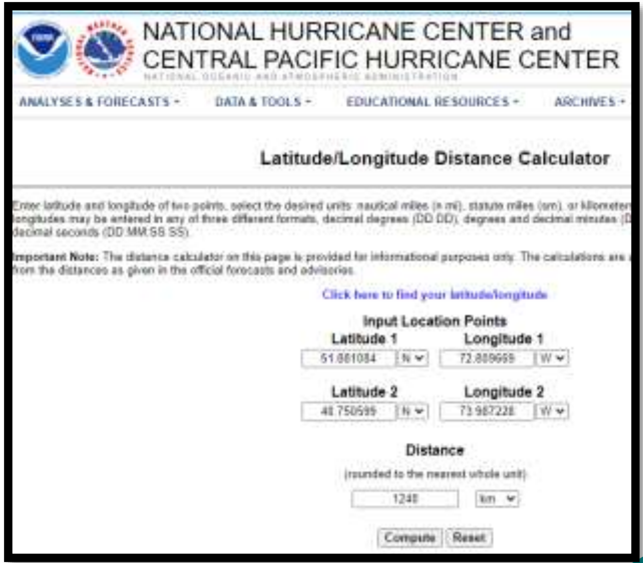
We use this distance function to calculate distance between two locations

Feature Manipulation

We use location base data that is latitude and longitude in our data and Find the distance between pickup and drop-off location

Also we cross-validate our answer with National Hurricane center and central pacific hurricane center website

We take one sample from our data and store the value into this website. The result is same


NATIONAL HURRICANE CENTER and
CENTRAL PACIFIC HURRICANE CENTER
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
ANALYSES & FORECASTS
DATA & TOOLS
EDUCATIONAL RESOURCES
ARCHIVES
Compute
Reset

Latitude/Longitude Distance Calculator

Enter latitude and longitude of two points, select the desired units: nautical miles (n mi), statute miles (sm), or kilometers longitudes may be entered in any of three different formats, decimal degrees (DD.DD), degrees and decimal minutes (D decimal seconds (DD:MM:SS.SS).

Important Note: The distance calculator on this page is provided for informational purposes only. The calculations are from the distances as given in the official forecasts and advisories.

Click here to find your latitude/longitude

Input Location Points

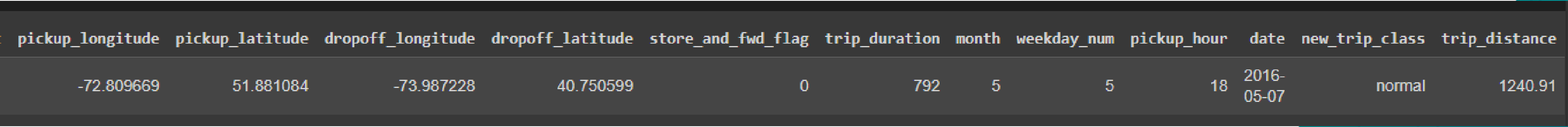
Latitude 1: 51.881084 N
Longitude 1: 72.809669 W

Latitude 2: 40.750599 N
Longitude 2: 73.987228 W

Distance

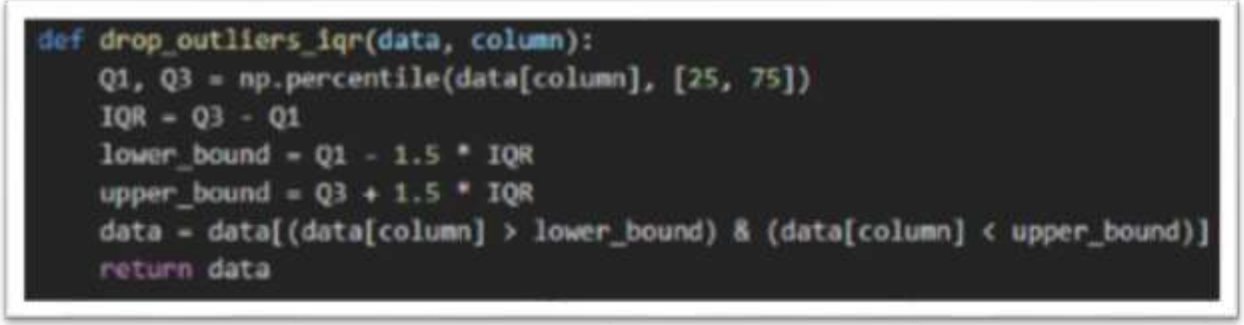
(rounded to the nearest whole unit)

1240 km



pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	store_and_fwd_flag	trip_duration	month	weekday_num	pickup_hour	date	new_trip_class	trip_distance
-72.809669	51.881084	-73.987228	40.750599	0	792	5	5	18	2016-05-07	normal	1240.91

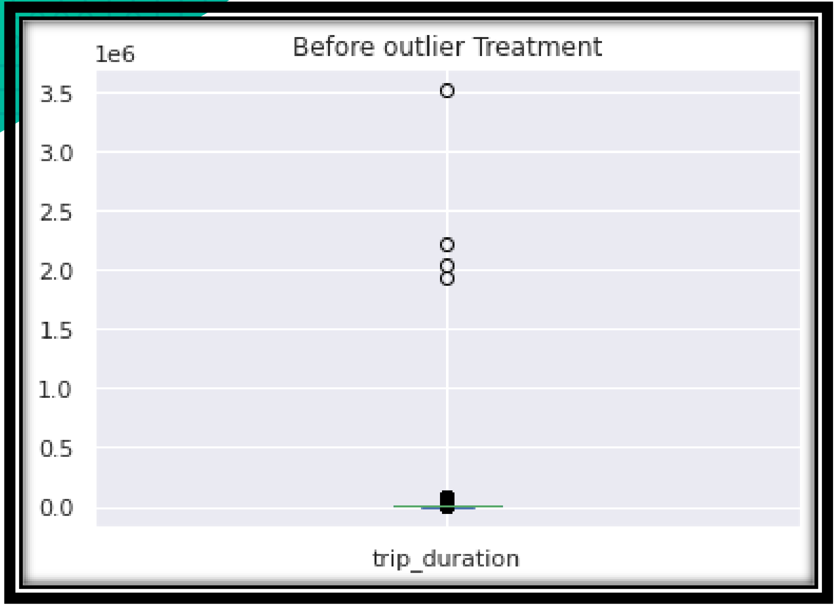
Handling Outliers

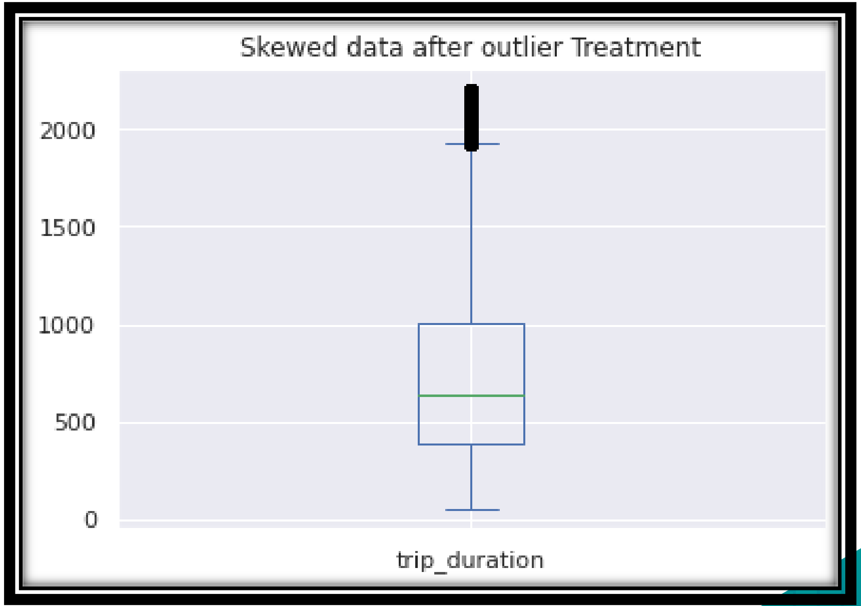


```
def drop_outliers_iqr(data, column):
    Q1, Q3 = np.percentile(data[column], [25, 75])
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    data = data[(data[column] > lower_bound) & (data[column] < upper_bound)]
    return data
```

We use this function to drop the outliers using Inter quartile Range of 25 and 75 Percentile

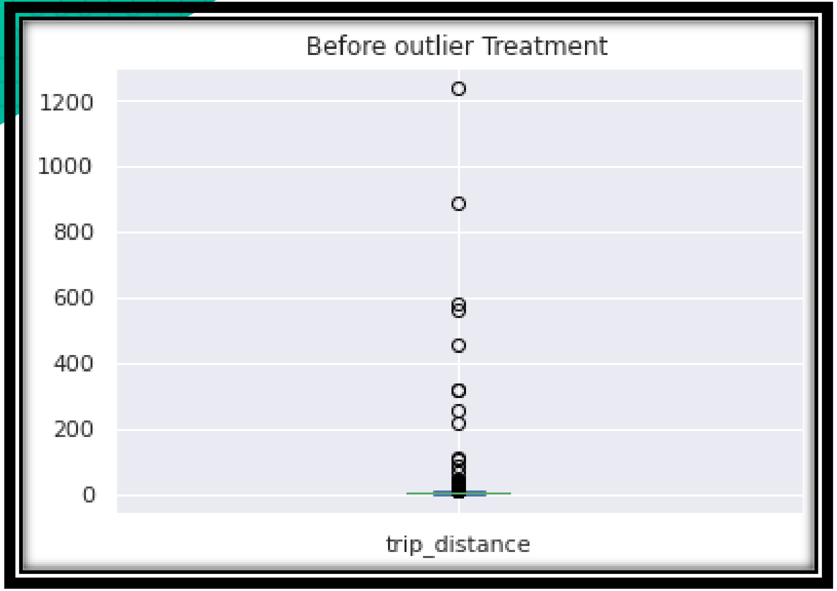
Outliers Treatment

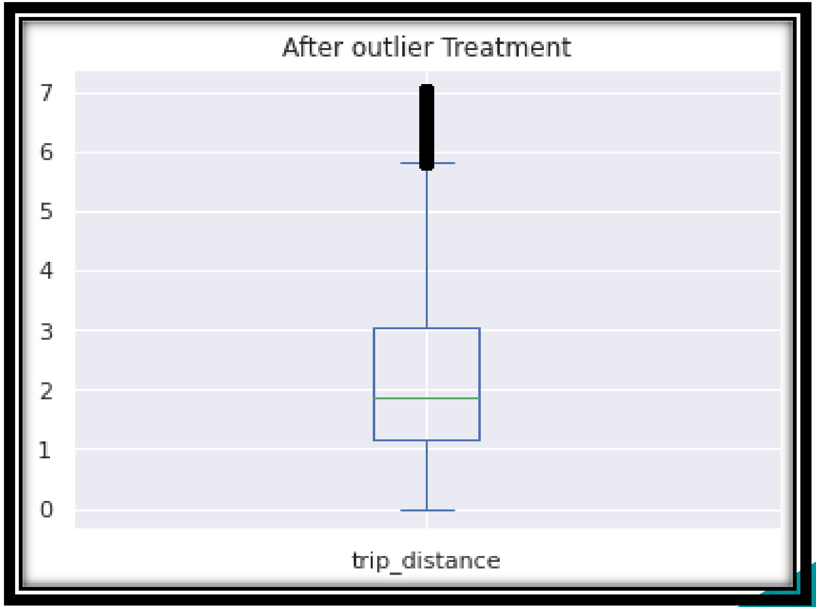




Trip Duration Before And After Outlier Treatment

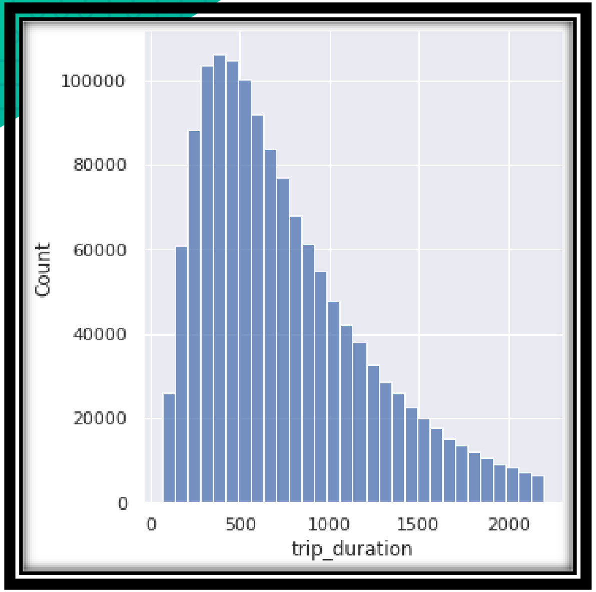
Outliers Treatment

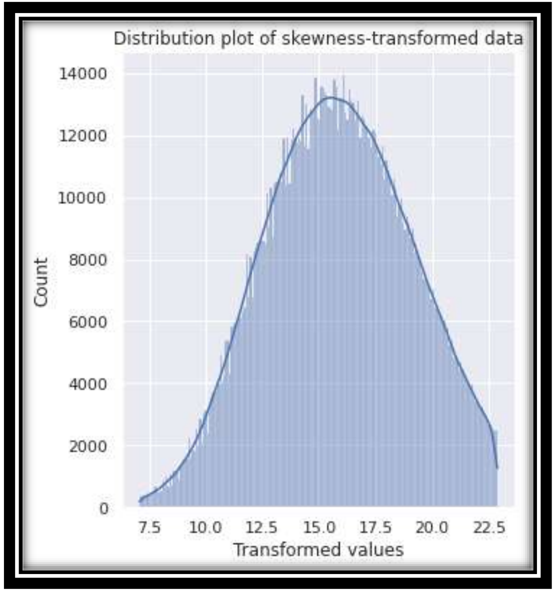


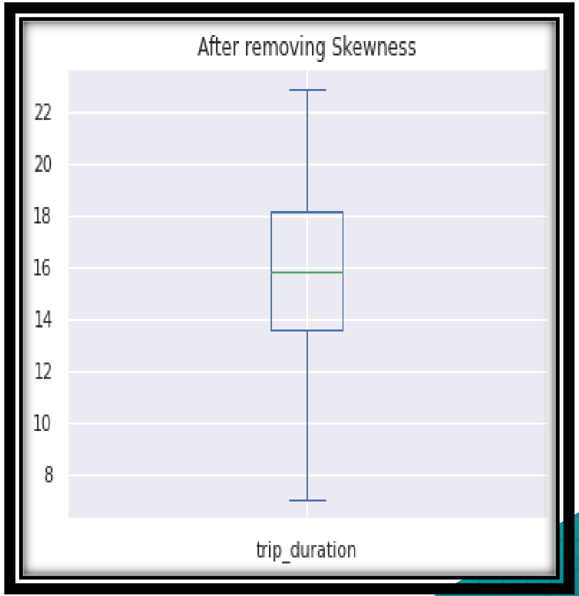


Trip Distance Before And After Outlier Treatment

Dealing with Skewness

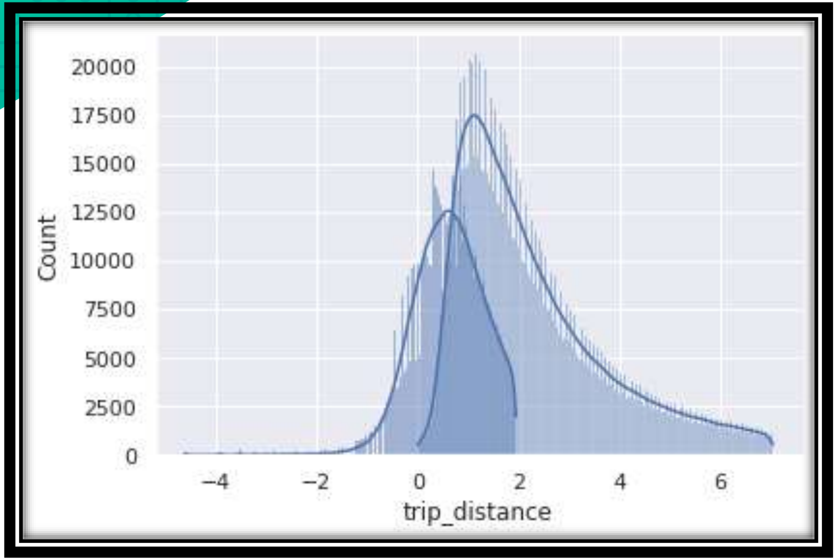






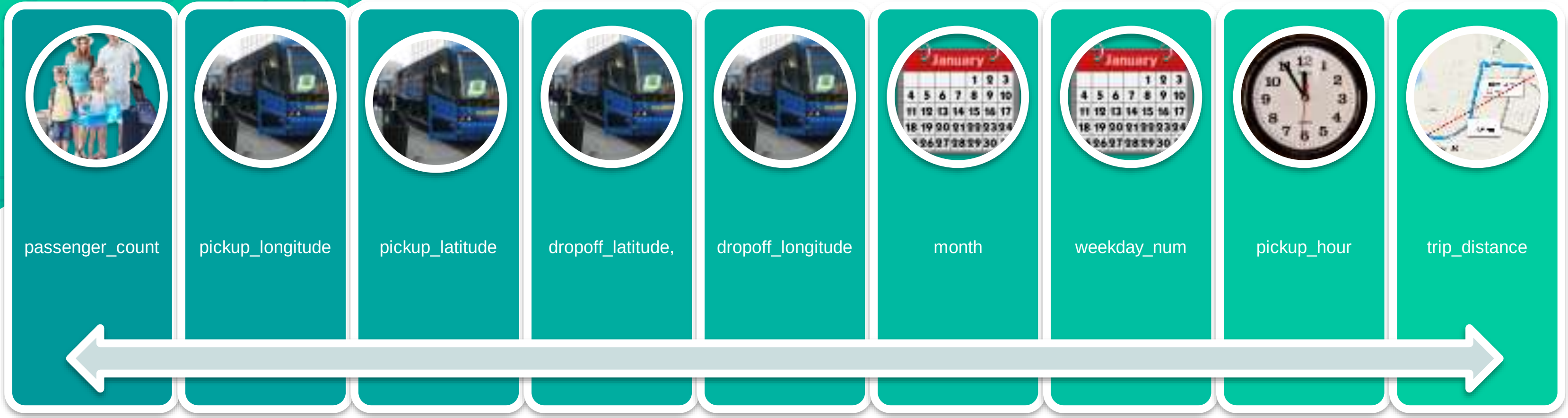
We use box-cox transformation to remove skewness but the thing is it reduces the trip duration to 7 – 24 from 0 to 2200, which causing large amount of data loss. Hence we are keeping the data as it is.

Dealing with Skewness



In this chart of trip distance count if we apply any sort of Transformation it leads to add negative values in our data

Feature Selection



passenger_count

pickup_longitude

pickup_latitude

dropoff_latitude,

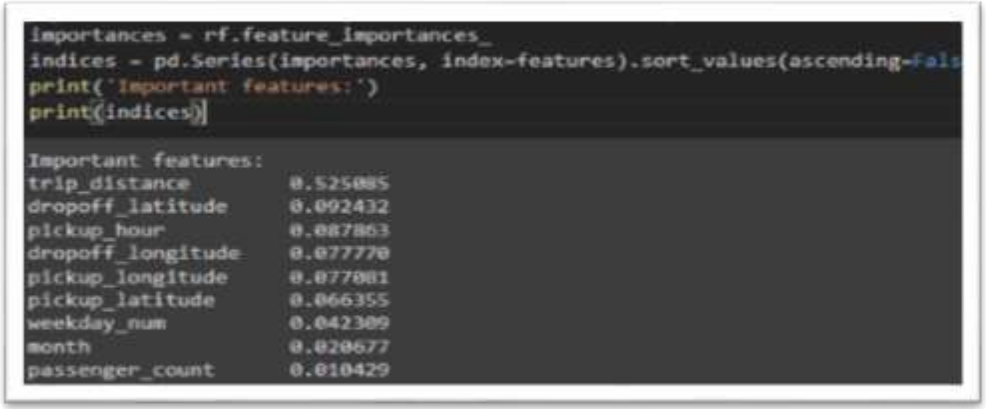
dropoff_longitude

month

weekday_num

pickup_hour

trip_distance



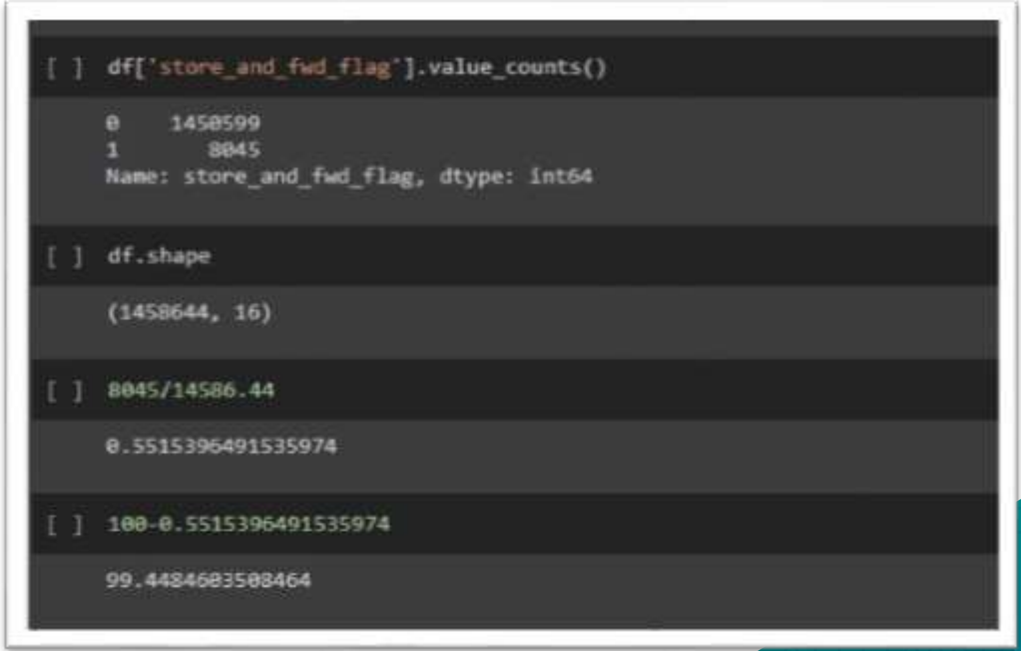
```
importances = rf.feature_importances_
indices = pd.Series(importances, index=features).sort_values(ascending=fals
print('Important features:')
print(indices)
```

```
Important features:
trip_distance        0.525085
dropoff_latitude     0.092432
pickup_hour          0.087863
dropoff_longitude    0.077770
pickup_longitude     0.077081
pickup_latitude      0.066355
weekday_num          0.042309
month                0.020677
passenger_count      0.010429
```

The importance of each feature is then normalized such that the sum of all feature importance's is equal to 1.0. Therefore, a higher feature importance value indicates that the feature is more important for the model's prediction. In our case Most important feature is **trip distance.**

Imbalanced Data

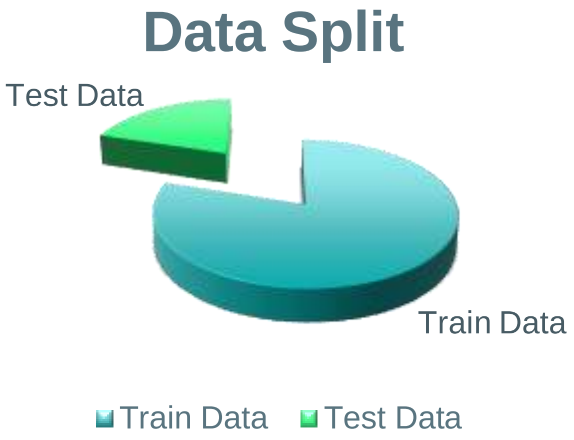
Handling Imbalance data
In our data set the feature Store and fwd flag
The value count of 1 Feature is 0.55 % almost nil
The value count of 0 Feature is 99.45% all data
Hence we drop this feature



```
[ ]  df['store_and_fwd_flag'].value_counts()

     0    1450599
     1       8045
     Name: store_and_fwd_flag, dtype: int64

[ ]  df.shape

     (1458644, 16)

[ ]  8045/14586.44

     0.5515396491535974

[ ]  100-0.5515396491535974

     99.4484603508464
```

Data Splitting

We have split the data into 80:20 ratio
Where 80 % data is for training and 20 % for test

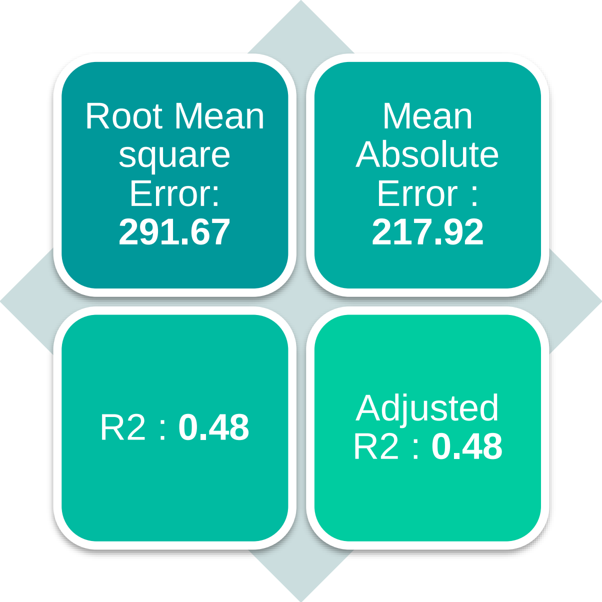


8

Model Implementation

Implementation of different Machine Learning models

Model 1 : Linear Regression



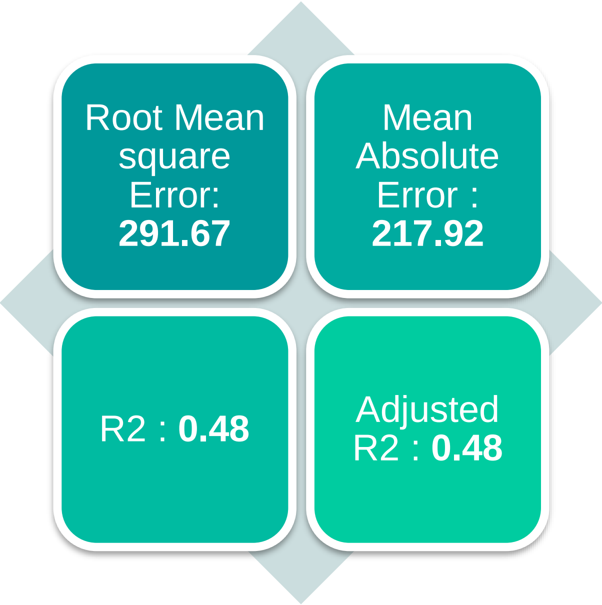
Root Mean square Error: **291.67**

Mean Absolute Error : **217.92**

R2 : **0.48**

Adjusted R2 : **0.48**

Model 2 : Ridge (L2 Regularization)



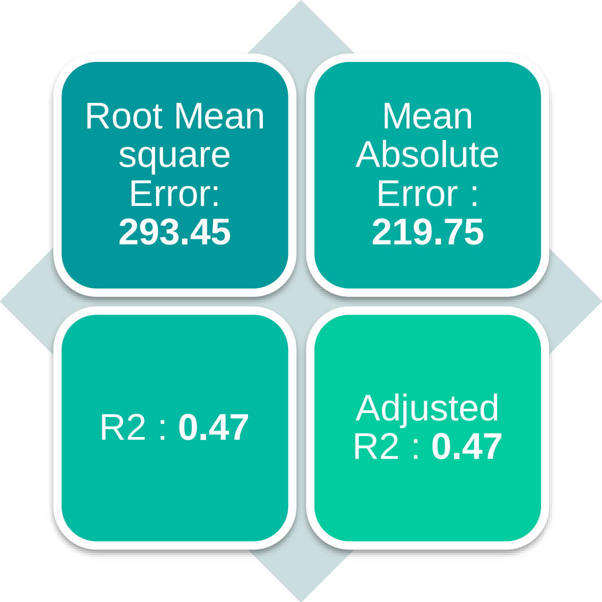
Root Mean square Error: **291.67**

Mean Absolute Error : **217.92**

R2 : **0.48**

Adjusted R2 : **0.48**

Model 3 : Lasso (L1 Regularization)



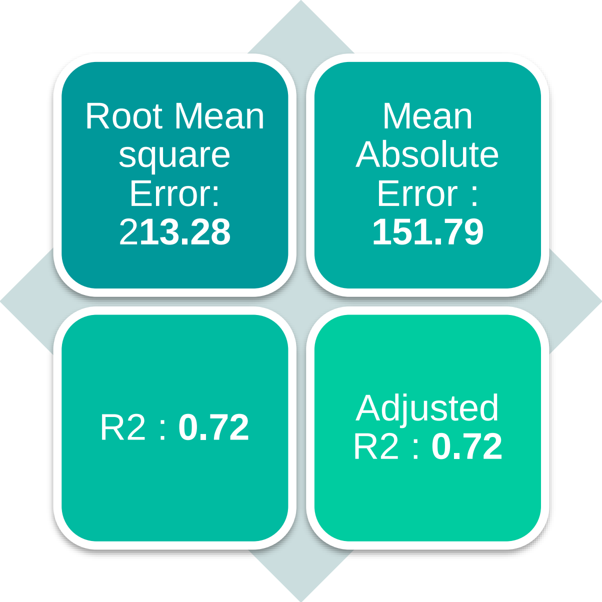
Root Mean square Error: **293.45**

Mean Absolute Error : **219.75**

R2 : **0.47**

Adjusted R2 : **0.47**

Model 4 : Random Forest Regressor



Root Mean square Error: **213.28**

Mean Absolute Error : **151.79**

R2 : **0.72**

Adjusted R2 : **0.72**

Best Model

From Above four Model we have chosen **Random Forest Regressor** Because

1. It reduces RMSE from 291.67 to 213.228
2. Reduces MAE from 217.92 to 151.78
3. Improves R2 from 0.48 to R2 score: 0.7208182963366805
4. Improve Adjusted r2 from 0.48 to 0.7208163191991594

9

Future Work

Future directions for improving taxi trip time prediction

Future Work

Incorporating more data
Taxi trip time prediction models can be improved by incorporating additional data sources, such as weather conditions, traffic patterns, and events happening in the city. This additional data can help the model better understand the factors that impact trip times.

Using advanced machine learning techniques
Advanced machine learning techniques, such as deep learning and ensemble learning, can be applied to taxi trip time prediction. These techniques can help capture complex patterns in the data that may be difficult to detect using traditional modeling approaches.

Developing real-time prediction models
Real-time prediction models can help provide more accurate estimates of trip times as conditions change. These models can be used to update taxi driver and passenger apps in real-time, providing more accurate and up-to-date information on trip times.

Improving the accuracy of pick-up and drop-off locations
Accurately predicting pick-up and drop-off locations can help improve the accuracy of trip time predictions. This can be accomplished through improved geocoding techniques or the use of more accurate location data.

Incorporating user feedback
User feedback can help improve the accuracy of trip time predictions over time. This feedback can be used to refine the model and identify areas where improvements can be made.

10

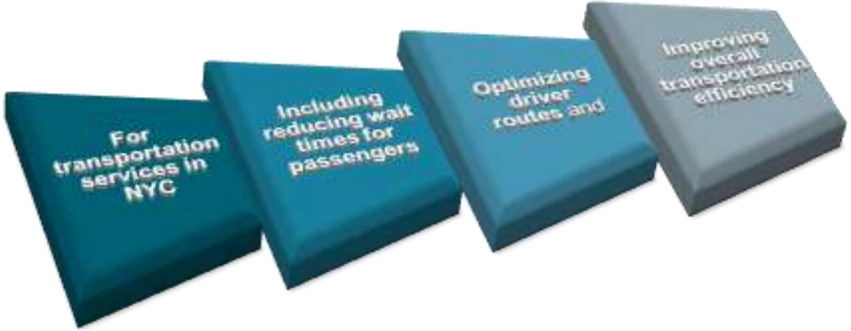
Conclusion

Final conclusion of our project

Conclusion

In conclusion, predicting taxi trip time accurately is an important task for optimizing transportation services in NYC. There have been many efforts to improve the accuracy of trip time predictions, including the use of advanced machine learning techniques, incorporating additional data sources, developing real-time prediction models, improving location accuracy, and incorporating user feedback.
Improving the accuracy of taxi trip time predictions has the potential to provide significant benefits:

As such, it is an important area of research and development that will likely continue to receive attention and investment in the years to come.



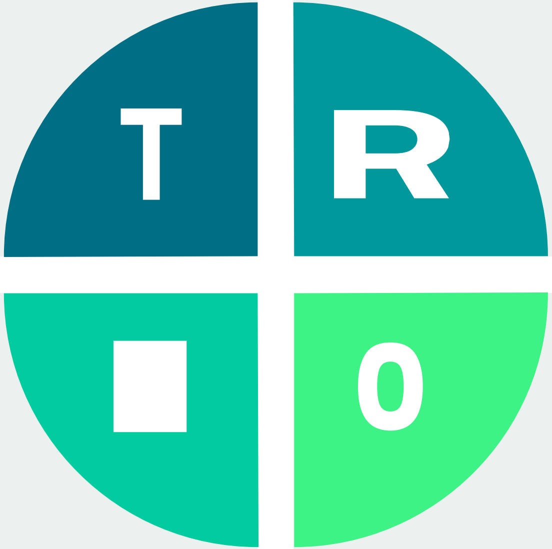
For transportation services in NYC

Including reducing wait times for passengers

Optimizing driver routes and

Improving overall transportation efficiency

Advantages TRIO



Transpotation

For transportation services in NYC

Reducing

Reducing wait times for passengers and increasing work efficiency

Improving overall transportation efficiency

Improving

Optimizing driver routes

Optimizing

THANKS!

Any questions?

You can find me at:

- https://github.com/riyapatelrp