Capstone Project – 2
Supervised ML - Regression
NYC Taxi Trip Time Prediction

By-

Avinash Yadav

Deepika Yadav

Presentation Outline:

- ❖ Problem Statement
- ❖ Introduction
- ❖ Exploring the dataset
- ❖ Methodology
- ❖ EDA and Data Processing
- ❖ Decomposition of Data:PCA
- ❖ ML Model – Regression
- ❖ 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 : Dtype and Null Values

```
#Attribute information

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
```

```
#checking missing values

df.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
```

Attribute Information : Unique Values

```
# Let us check for unique values of all columns.

print(df.nunique().sort_values())
```

```
vendor_id                   2
store_and_fwd_flag          2
passenger_count            10
trip_duration            7417
pickup_longitude        23047
dropoff_longitude       33821
pickup_latitude         45245
dropoff_latitude        62519
pickup_datetime       1380222
dropoff_datetime      1380377
id                    1458644
dtype: int64
```

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:

- Decomposition: PCA
- Linear Regression
- Decision Tree
- Random Forest

Tools Used:

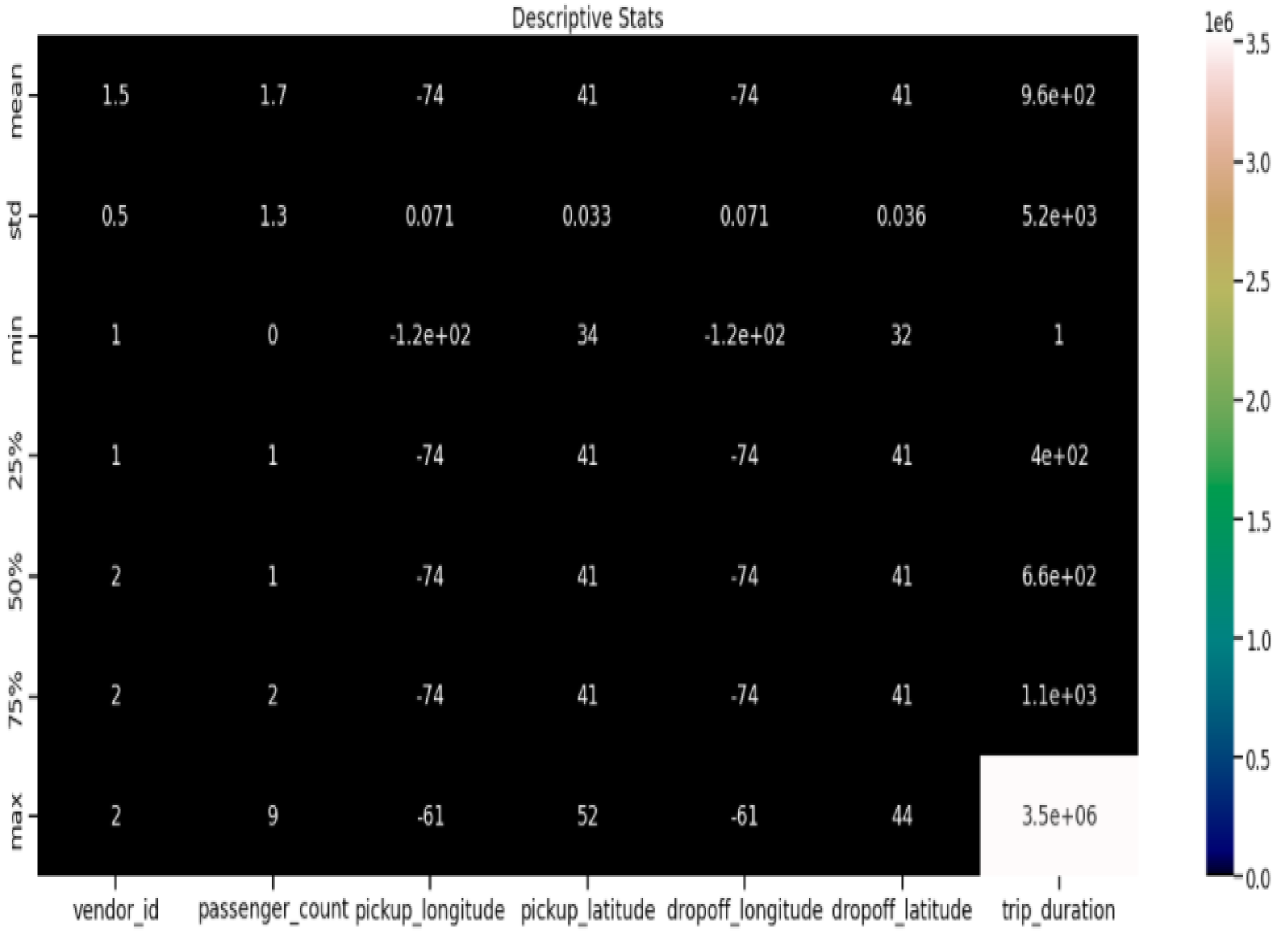
- Jupyter Notebook (Python)
- Google Colab Research

EDA AND DATA PROCESSING

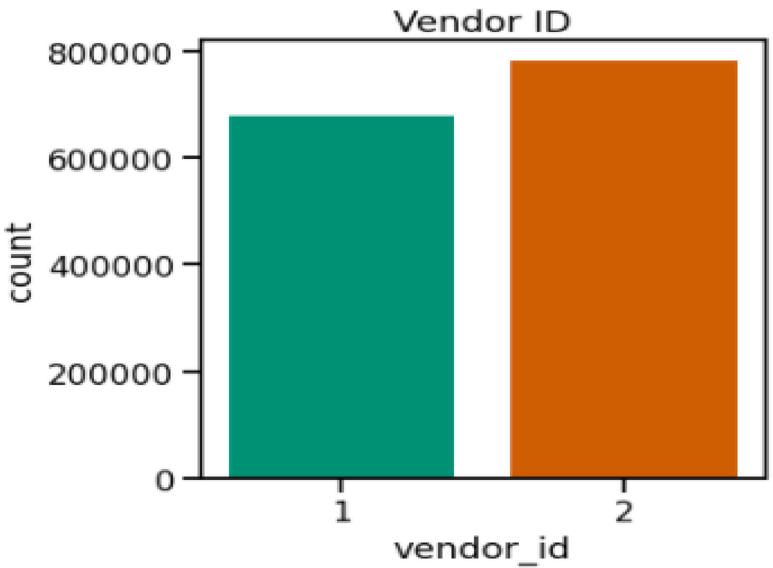
Descriptive Stats in Visual Form

- ❖ We can observe that There were trips having 0 passengers which we can consider as false trip.

- ❖ Also, there are trips having trip duration upto 3526282 seconds (Approx. 980 hours) which is kind of Impossible in a day.

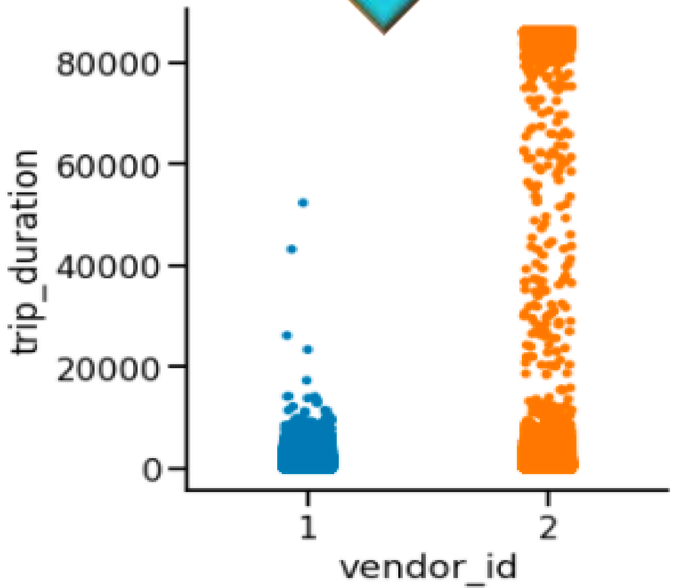


Analysis on : Vendor Id

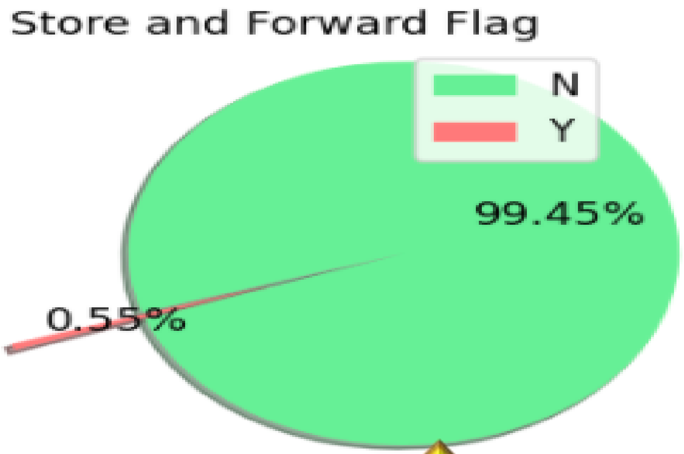


From Above Visualization, we can say that there are 2 vendors (Service Providers).
2nd Service provider is the most opted one by New Yorkers.

Vendor id 2 takes longer trips as compared to vendor 1

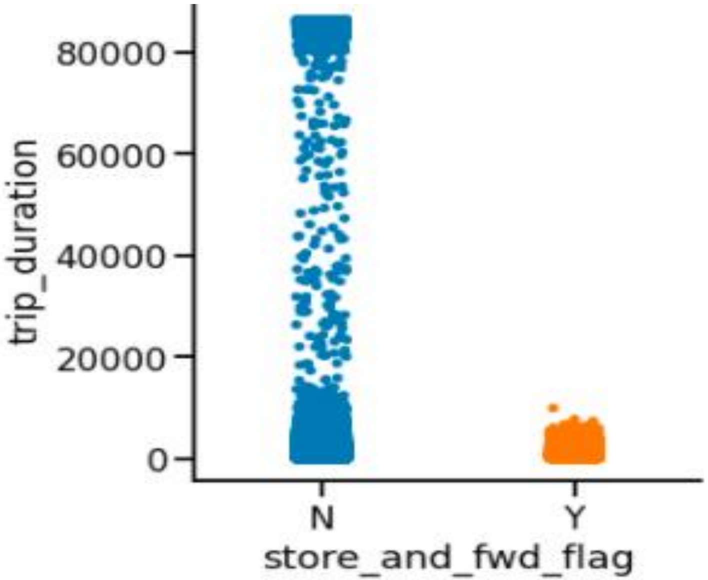


Analysis on : Store and Forward Flag

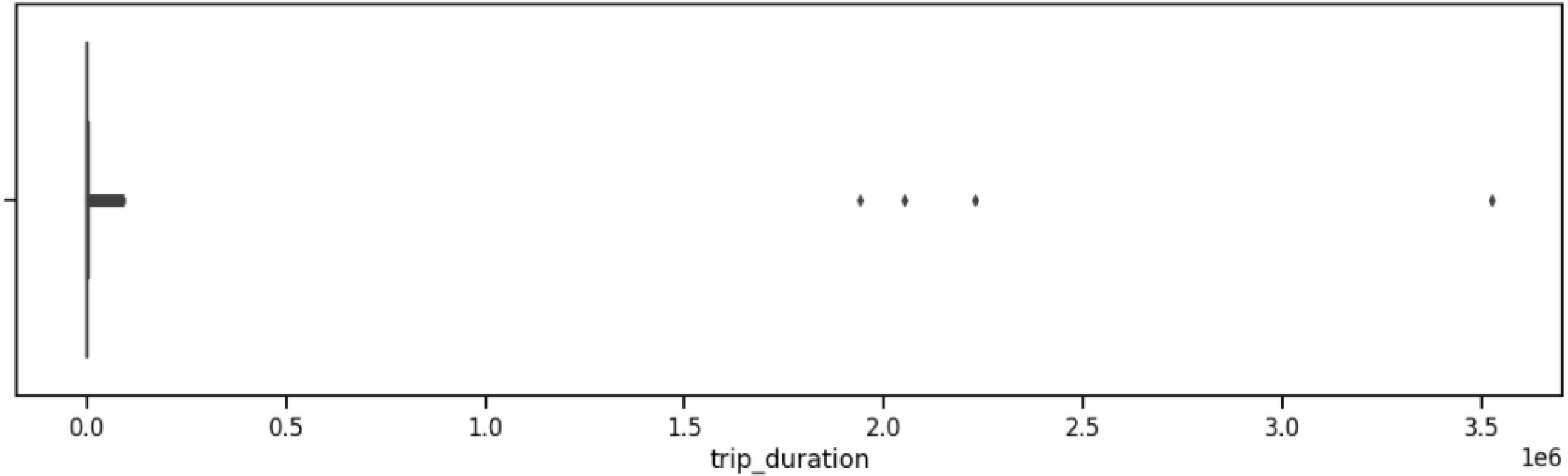


❖We see there are less than 1% of trips that were stored before forwarding
❖The number of N flag is much larger. We can later see whether they have any relation with the duration of the trip.

Trip duration is generally longer for trips whose flag was not stored.

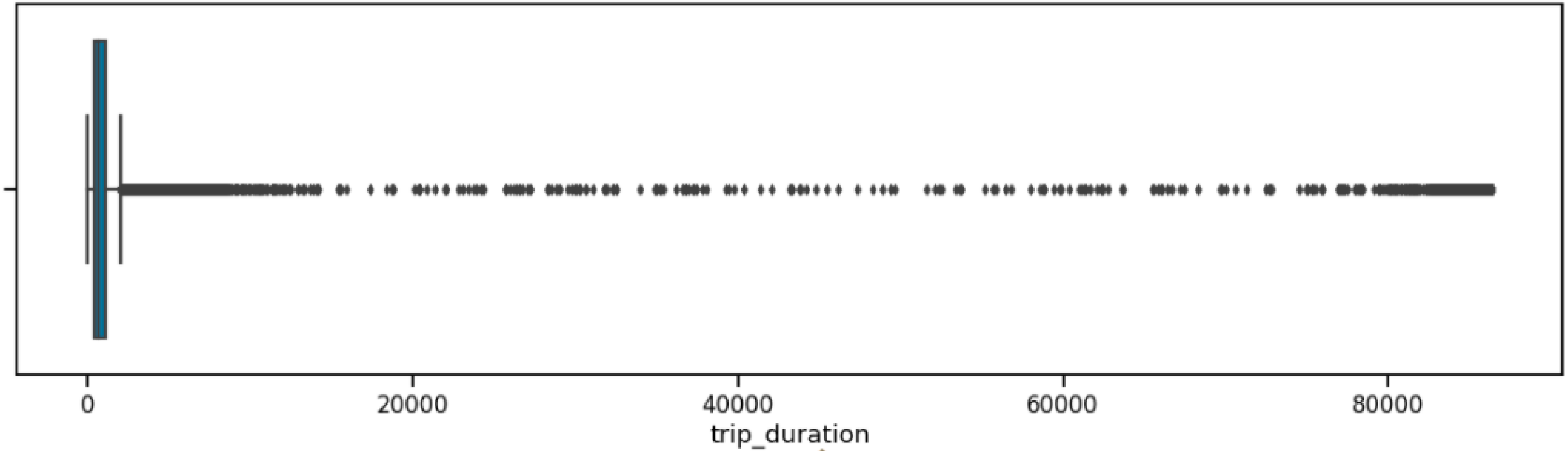


Analysis on : Target Variable – Trip Duration



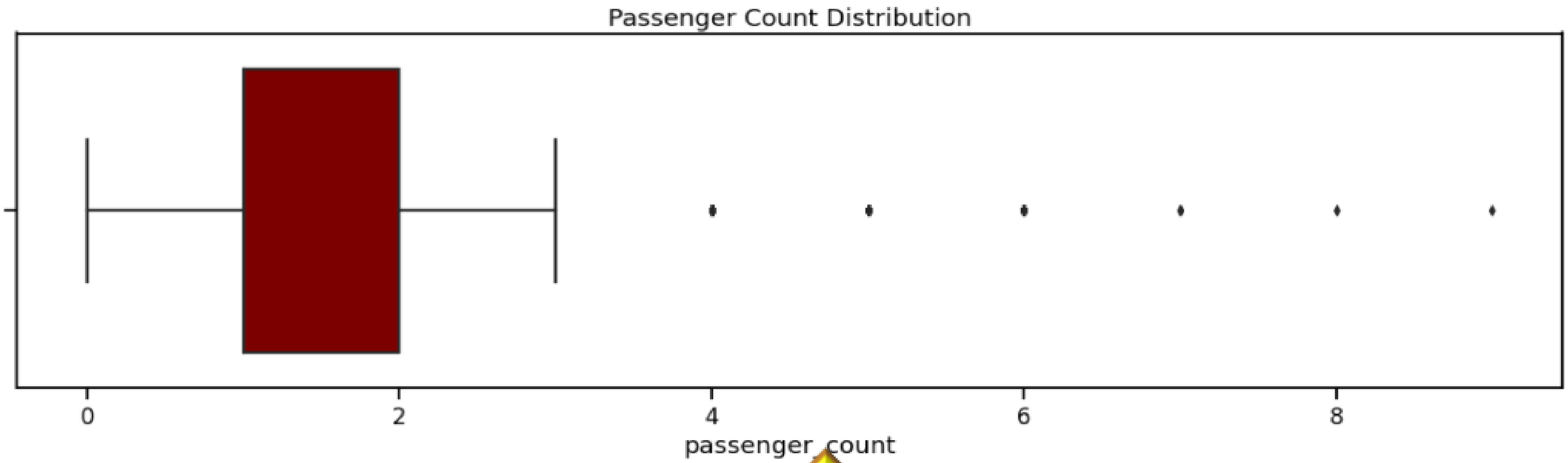
Probably in this visualization we can clearly see some outliers , their trips are lasting between 1900000 seconds (528 Hours) to somewhere around 3500000 (972 hours) seconds which is impossible in case of taxi trips , How can a taxi trip be that long ?It's Quite suspicious. We'll have to get rid of those Outliers.

Analysis on : Target Variable – Trip Duration (Contd.)



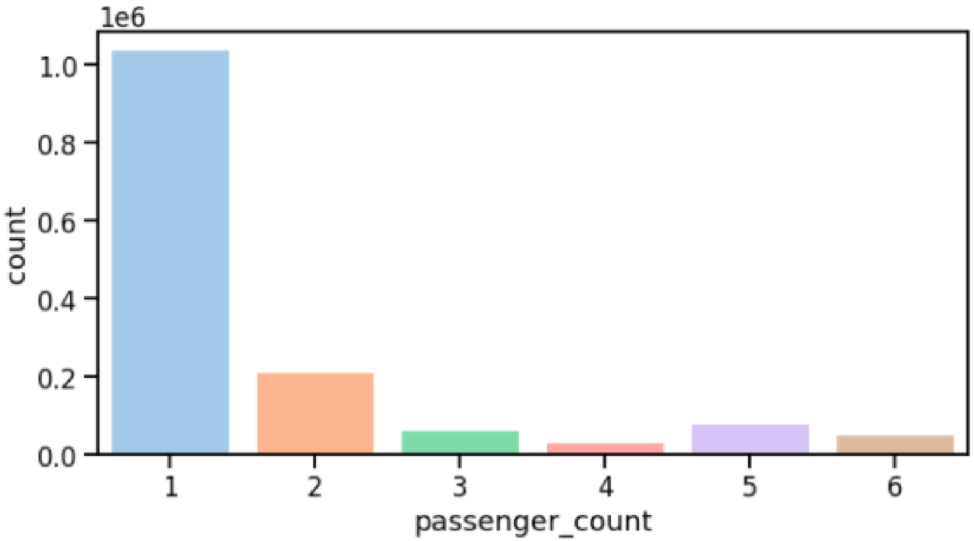
We can see that there is some entries which is significantly different from others. As there is this 4 rows only, we drop these rows.

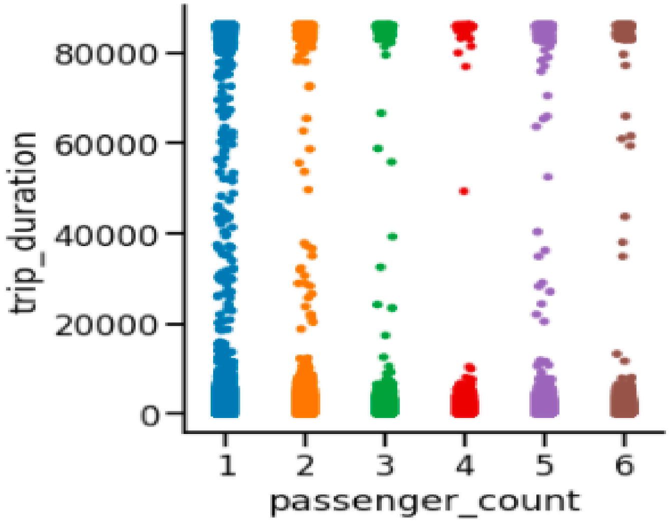
Analysis on : Passenger Count



❖There are some trips with even 0 passenger count. And 3 trips with 7 passengers. And there is only 1 trip each for 8 and 9 passengers.
❖Above visualization tells us that there were most number of trips are done by 1-2 passenger(s).
❖5 - 9 passengers trip states us that cab must be a Large vehicle.

Analysis on : Passenger Count (Contd.)

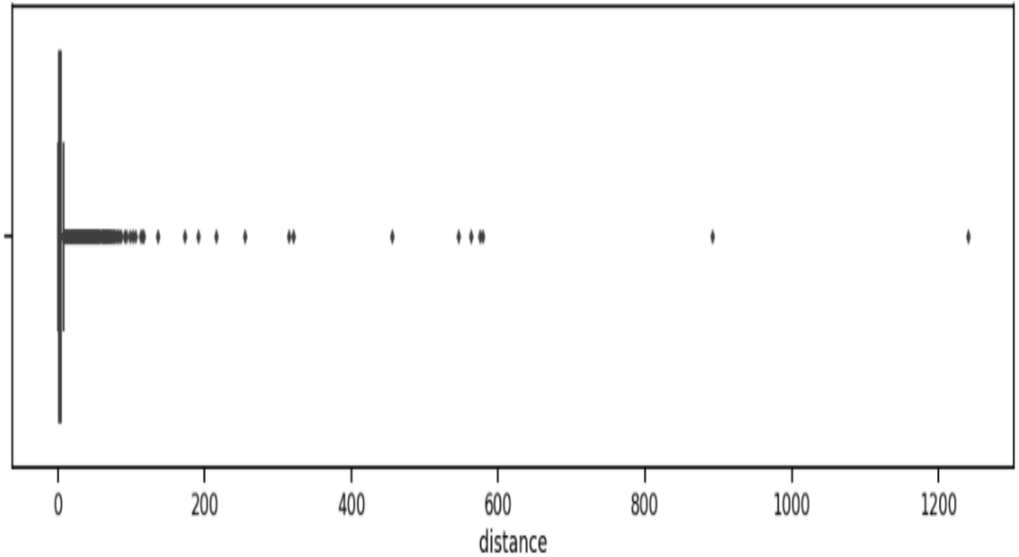


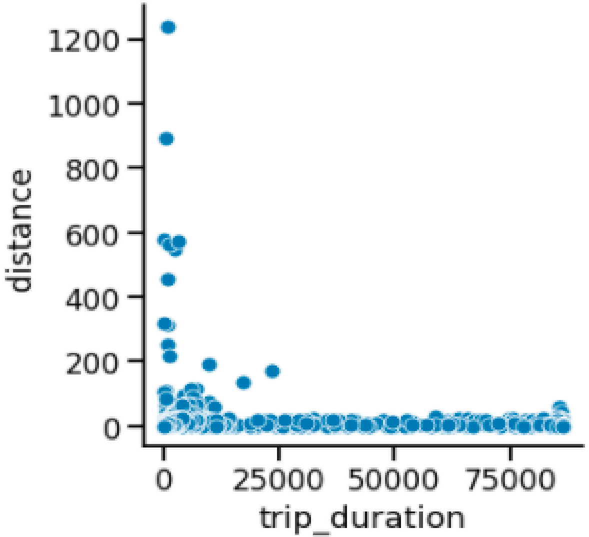


Now, that seems like a fair distribution. We see the highest amount of trips are with 1 passenger.

There is no visible relation between trip duration and passenger count

Analysis on : Distance





❖We can see there are trips which trip duration is as short as 0 seconds and yet covering a large distance. And, trips with 0 km distance and long trip durations.

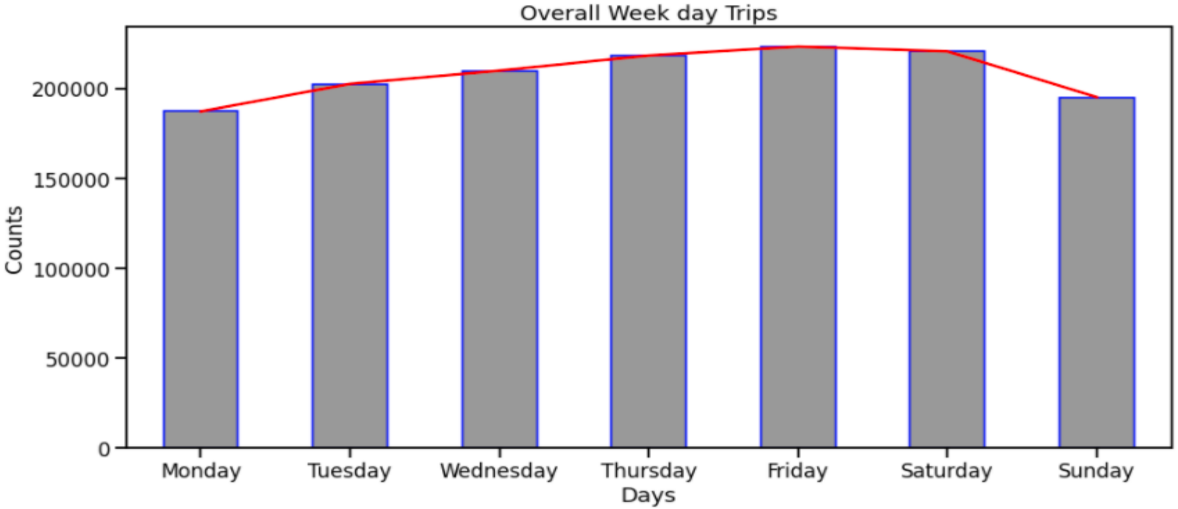
❖The reasons for 0 km distance can be:

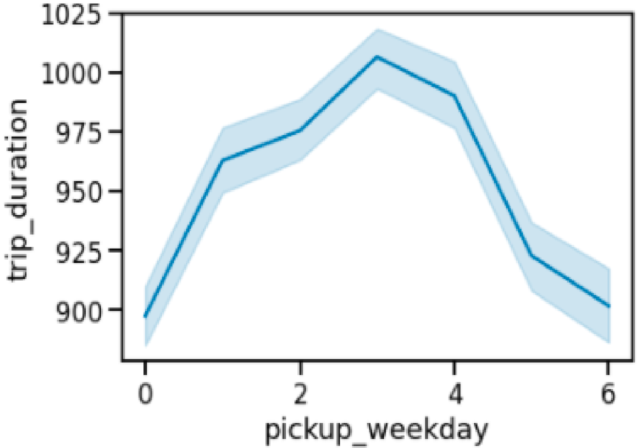
i. The drop off location couldn't be tracked.

ii. The driver deliberately took this ride to complete a target ride number.

iii. The passengers canceled the trip.

Analysis on : Trip Duration on a weekday

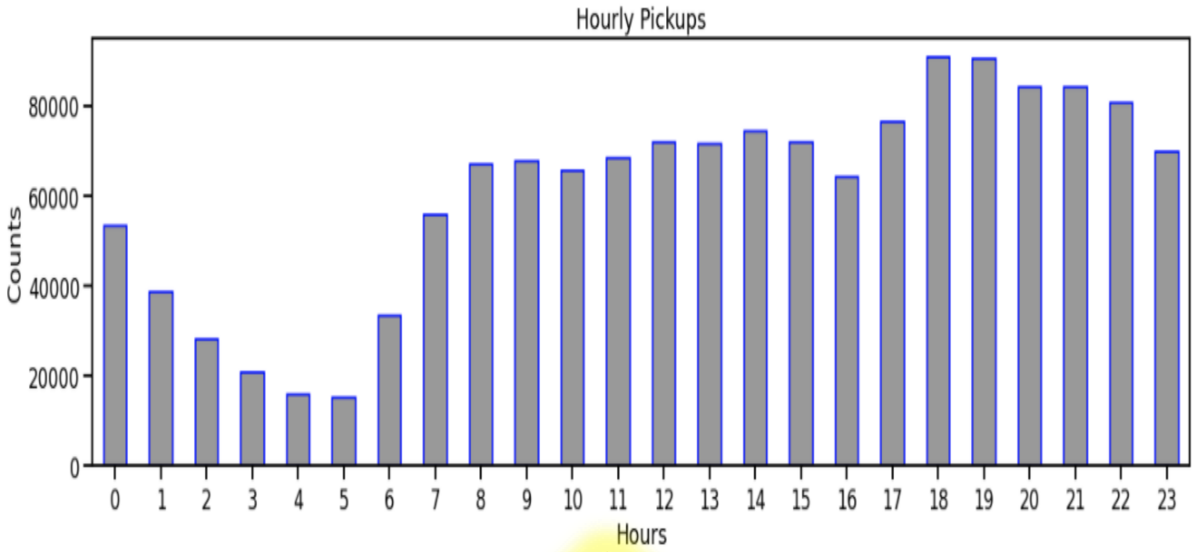


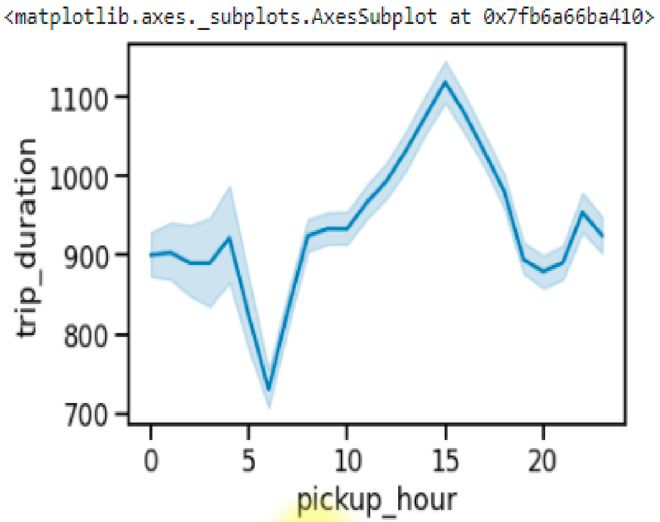


Observations tells us that Fridays and Saturdays are those days in a week when New Yorkers prefer to roam in the city.

Trip duration is the longest on Thursdays closely followed by Fridays.

Analysis on : Trip Duration per hour

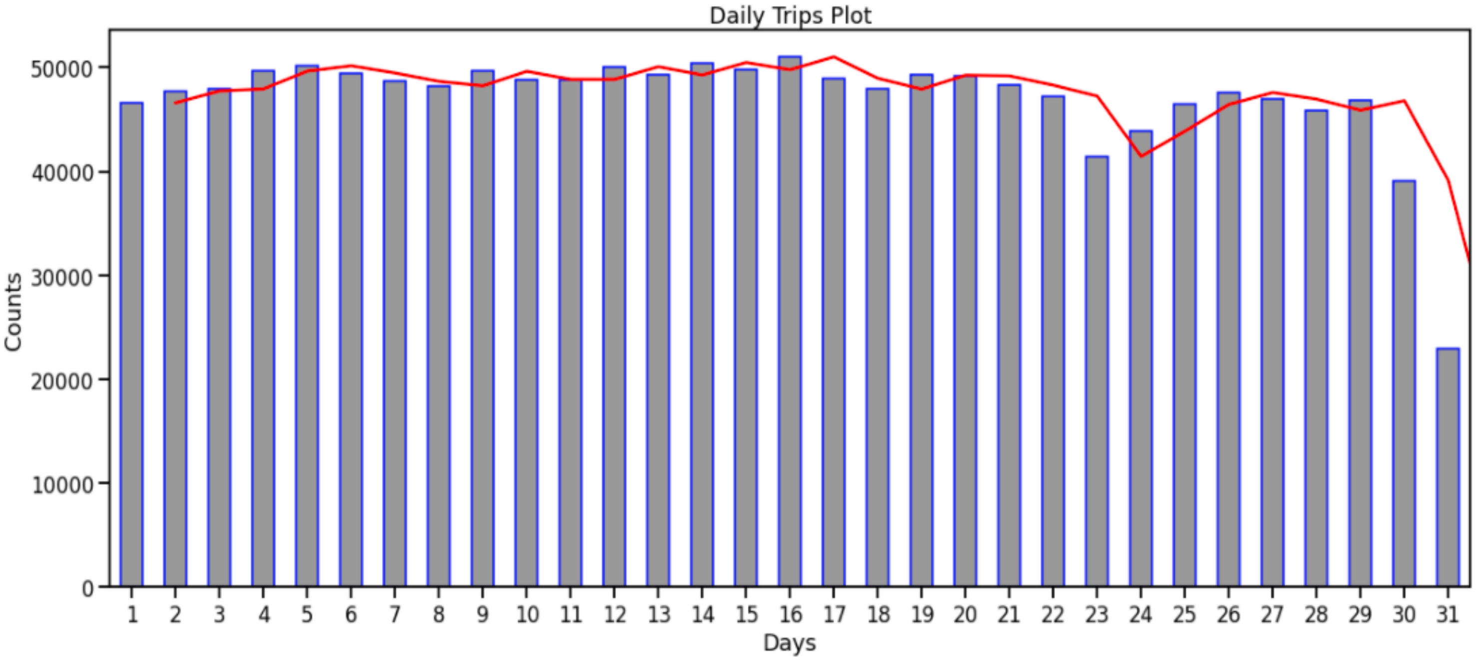




In which hour we get to see maximum pickups? - Rush hours (5 pm to 10 pm), probably office leaving time.
Thus we observe that most pickups and drops occur in the evening. While the least drops and pickups occur during midday.

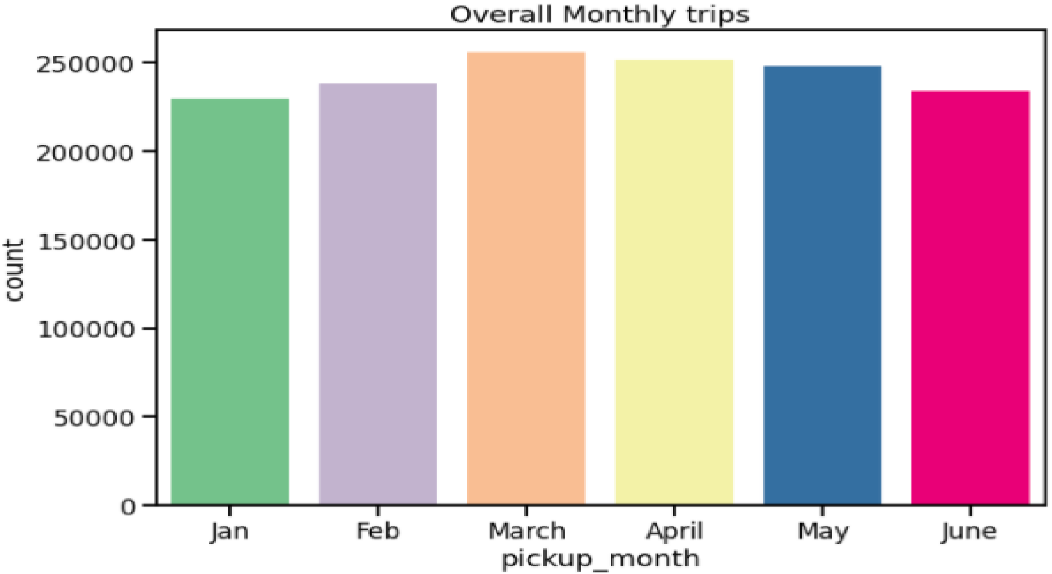
We see the trip duration is the maximum around 3 pm which may be because of traffic on the roads. Trip duration is the lowest around 6 am as streets may not be busy.

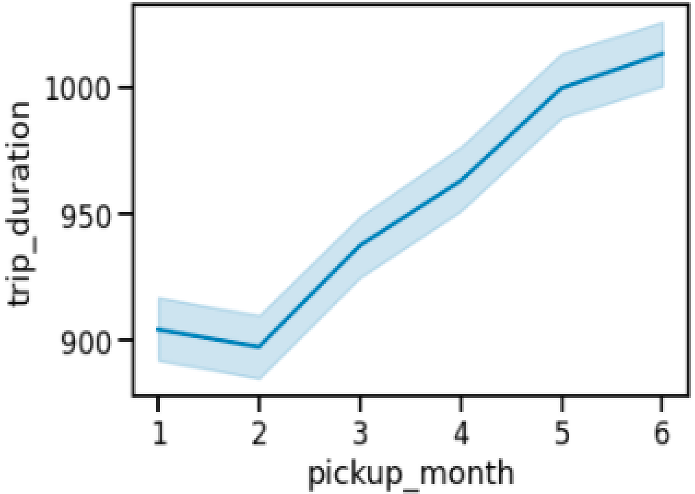
Analysis on : Trip Duration in a month



Seem like New Yorker's do not prefer to get a Taxi on Month end's , there is a significant drop in the Taxi trip count as month end's approach.

Analysis on : Trip Duration in 6 months

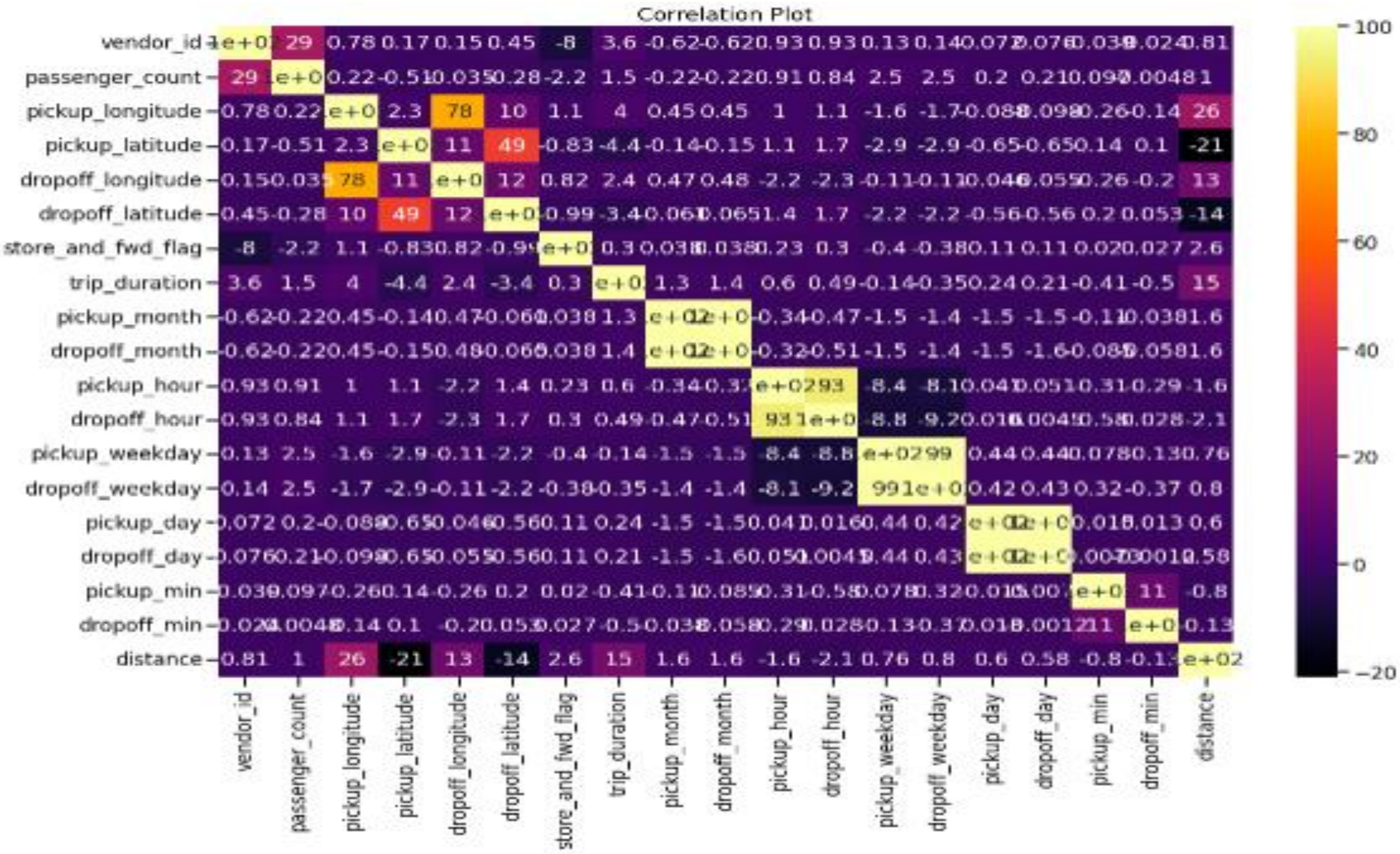




- ❖We've data of 6 months.
- ❖Number of trips in a particular month - March and April marking the highest.
- ❖January being lowest probably due to extreme SnowFall NYC.

From February, we can see trip duration rising every month.

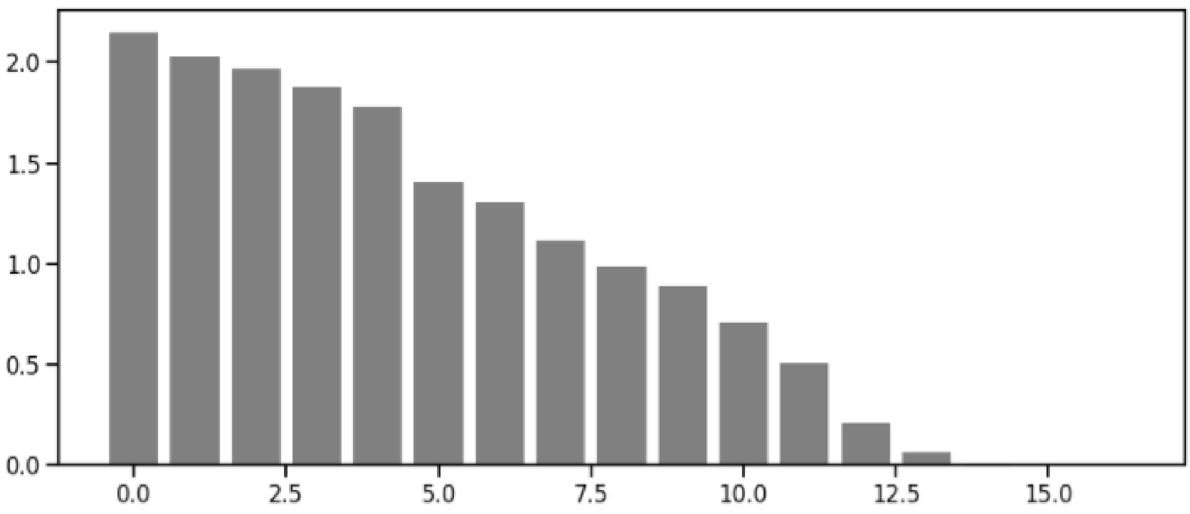
Analysis on : Correlation Heat map

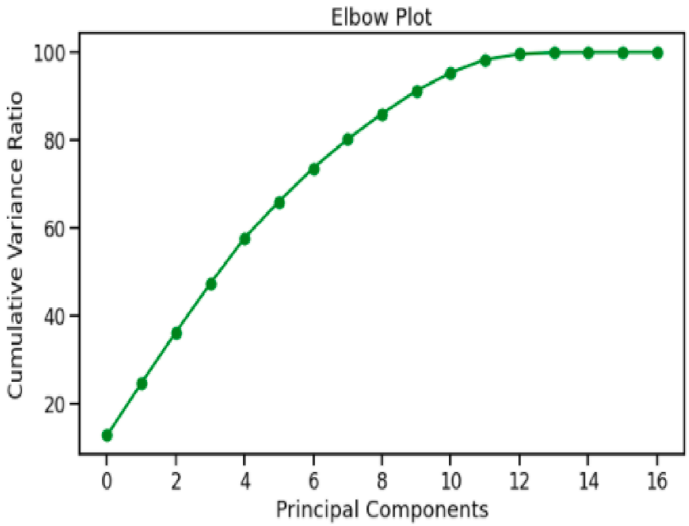


Decomposition of Data: PCA

Analysis on : Principal Component Analysis

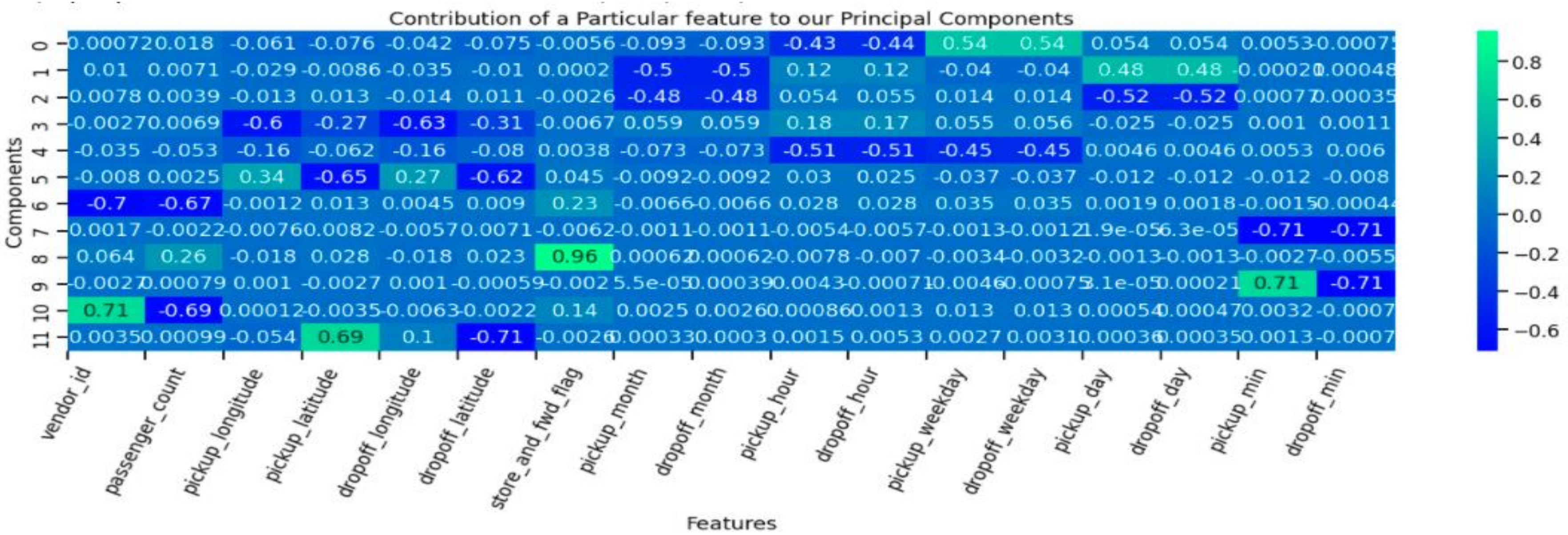
AI





Now that we're done, we have to pass our Scaled Dataframe in PCA model and observe the elbow plot to get better idea of explained variance. At 12th component our PCA model seems to go flat without explaining much of a variance.

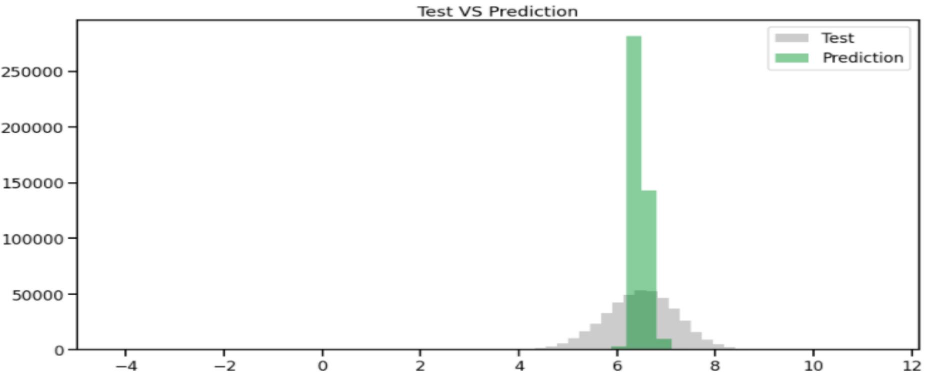
Analysis on: Feature Contribution



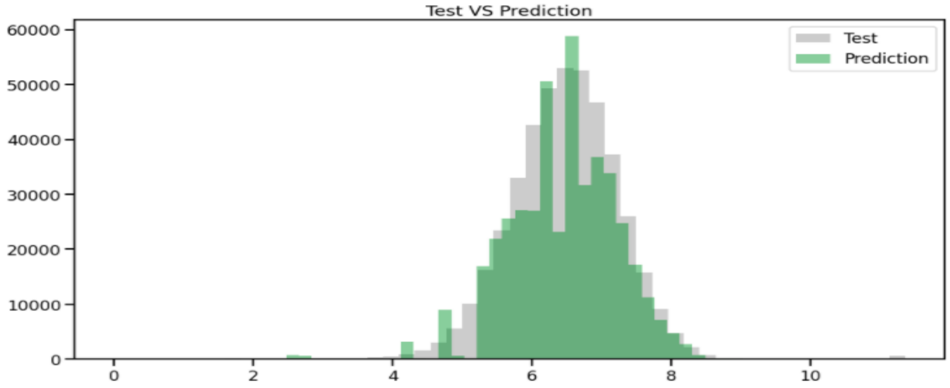
- ❖ Above plot gives us detailed ideology of which feature has contributed more or less to our each Principal Component.
- ❖ Principal Components are our new features which consists of Information from every other original Feature we have.
- ❖ We reduce the Dimensions using PCA by retaining as much as Information possible.

Machine Learning Model – Regression

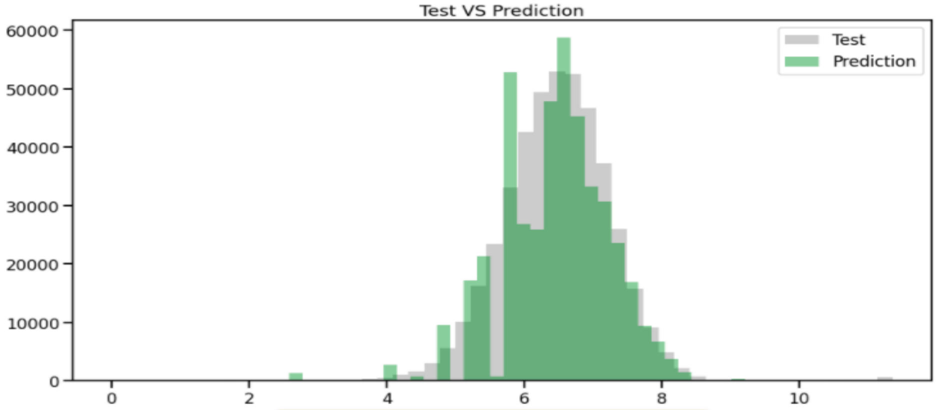
Analysis on: ML Model Prediction with PCA



Linear Regression



Decision Tree



Random Forest

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

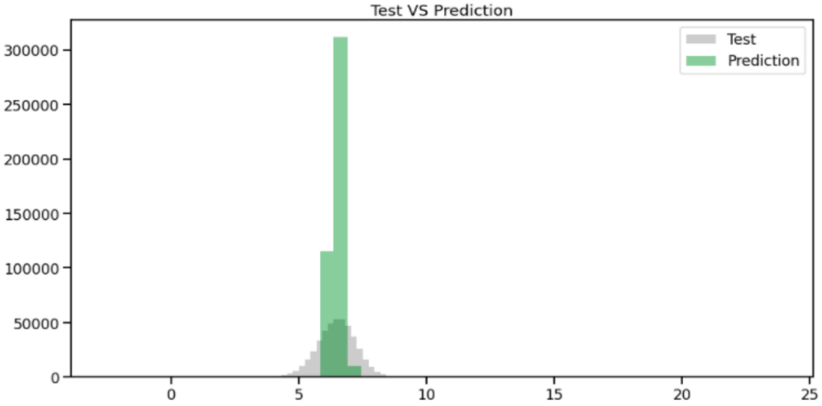
Analysis on : Model Evaluation Result with PCA

- We can clearly observe that our Decision Tree model and Random Forest model are good performers.
- As, Random Forest is providing us reduced RMSE, we can say that it's a model to Opted for.
- We're getting good fit score for Decision Tree and Random Forest , i.e., close to 1.0

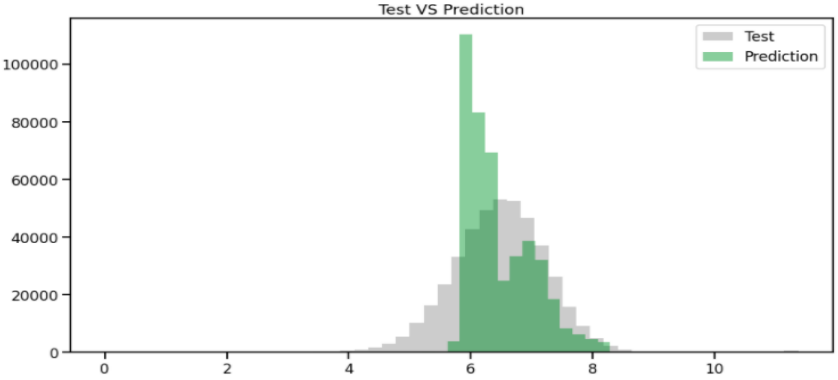
Algorithms	Training Score	Validation Score	Cross Validation Score	R2-Score	RMSE
Linear Regression	0.0389	0.0509	0.0329	-34.90	--
Decision Tree	0.9238	0.9149	0.9161	0.9076	0.038
Random Forest	0.9329	0.9260	0.9241	0.9191	0.036

- R2-score: Usually must be between 0 and 1, towards 1 considered as good fit.
- RMSE: Lesser is Better

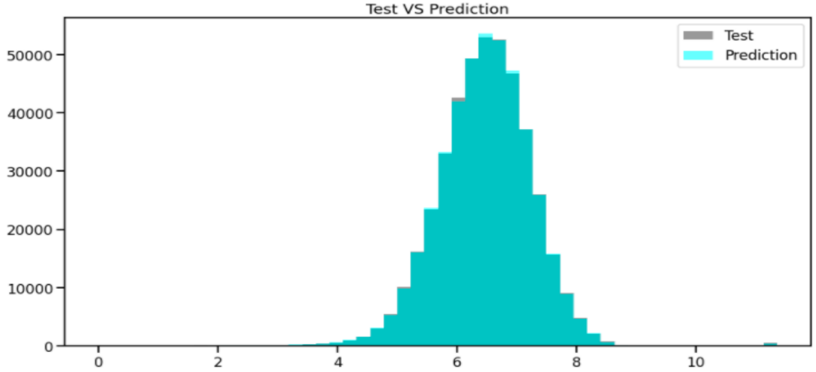
Analysis on : ML Model Prediction without PCA



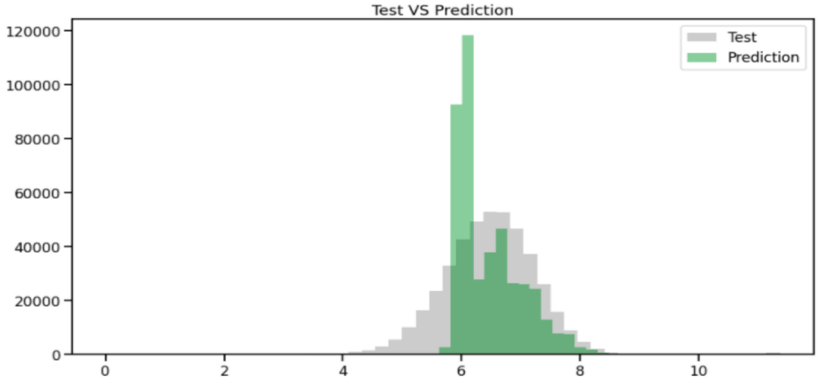
Linear Regression



Decision Tree



Decision Tree with GridsearchCV



Random Forest

Analysis on : Model Evaluation Result without PCA

- We can clearly observe that our Decision Tree with GridsearchCV model are good performers. As, It is providing us reduced RMSE, we can say that it's a model to Opt for.
- We're getting good fit score for Decision Tree with GridsearchCV , i.e, close to 1.0

Algorithms	Training Score	Validation Score	Cross Validation Score	R2-Score	RMSE
Linear Regression	0.0401	0.0517	0.0342	-32.90	--
Decision Tree	0.4649	0.4555	0.4550	-0.177	0.0885
Decision Tree with GridCV search	0.5135	0.4952	--	-0.0149	0.0857
Random Forest	0.4803	0.4720	0.4692	-0.231	0.0874

- R2-score: Usually must be between 0 and 1, towards 1 considered as good fit.
- RMSE: Lesser is Better

Conclusion

- Observed which taxi service provider is most Frequently used by New Yorkers.
- Found out few trips which were of duration 528 Hours to 972 Hours, possibly Outliers.
- Passenger count Analysis showed us that there were few trips with Zero Passengers and One trip with 7,8 and 9 passengers.
- Monthly trip analysis gives us a insight of Month – March and April marking the highest number of Trips while January marking lowest, possibly due to Snowfall.
- 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

Q & A