Capstone Project – 2

NYC Taxi Trip Time Prediction

By – Rajvee Sharma

Content :

> Introduction
> Project Description
> Exploring the dataset
> Problem Definition and Solution
> Techniques used
> Decomposition of data – PCA
> Conclusion

Introduction

1. Transportation plays a vital role in large cities.
2. Taxi mode of transportation has become a key player in United States and other countries.
3. In NYC there is large amount of taxi and taxi drivers are present.
4. Different variety of service provider are Uber, OLA, yellow taxi, green taxi and many more.
5. The data that contains ride details was made available by NYC taxi.
6. We use these details to perform analytics on ride data that would benefit for various business growth and also useful for government.

Project Description-

In this project I perform analytics on NYC data and made predictions. This predictions is using regression method to predict the trip duration depending on the given variables.

The variables contain the location of pickup and drop-off presenting with latitude and longitude, pickup date/time, number of passengers etc.

The design of learning algorithm includes the pre-process of data selection, feature extraction, validation, modelling. To improve the prediction I’ve done several tests for modelling and feature extraction.

Problem definition

The problem is divided in three stages :

1st stage : Data Cleaning and Analysis of data
2nd stage : Finding time duration between pickup and drop-off location.
3rd stage : Visualisation of results

Exploring the dataset

Data summary-

Data set name – NYC Taxi Data.csv – the training data

Statistics-

> No. of Rows – 1458644
> No. of Features -- 11 (Including Target)
> Target – Finding trip duration is important

Column name –

id, vendor_id, pickup_datetime, dropoff_datetime, pickup_latitude, pickup_longitude, passenger_count, dropoff_latitude, dropoff_longitude, store_and_fwd_flag, trip_duration

Data menu-

Independent variable-

> id – a unique identifier for each trip
> vendor_id – a code indicating the provider associated with the trip duration
> 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 passenger in a vehicles
> dropoff _latitude – the latitude where the meter was disengaged.
> dropoff _longitude – the longitude where the meter was disengaged
> pickup _longitude – the longitude where the meter was engaged
pickup latitude – the latitude where the meter was engaged
> store_and_fwd_flag – the flag indicates where the trip record was held in vehicle memory before sending to the vendor
the server ; Y= store and forward flag
N = not a store and forward flag

Target variable –

trip_duration – trip duration is seconds

Methodology

Machine learning algorithms-

Decomposition – PCA

Linear Regression

Decision Tree

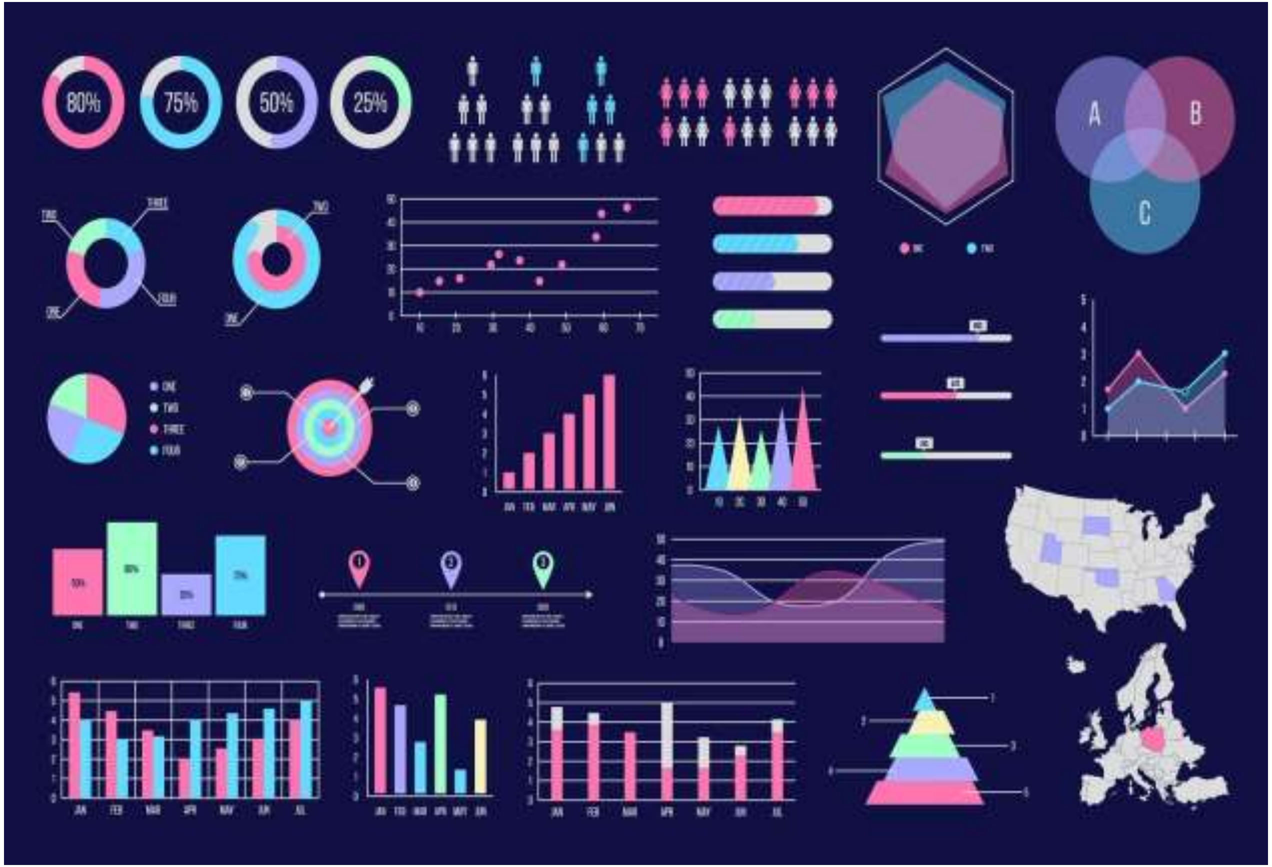
Random Forest

Tools used-

Google colab notebook

Python language

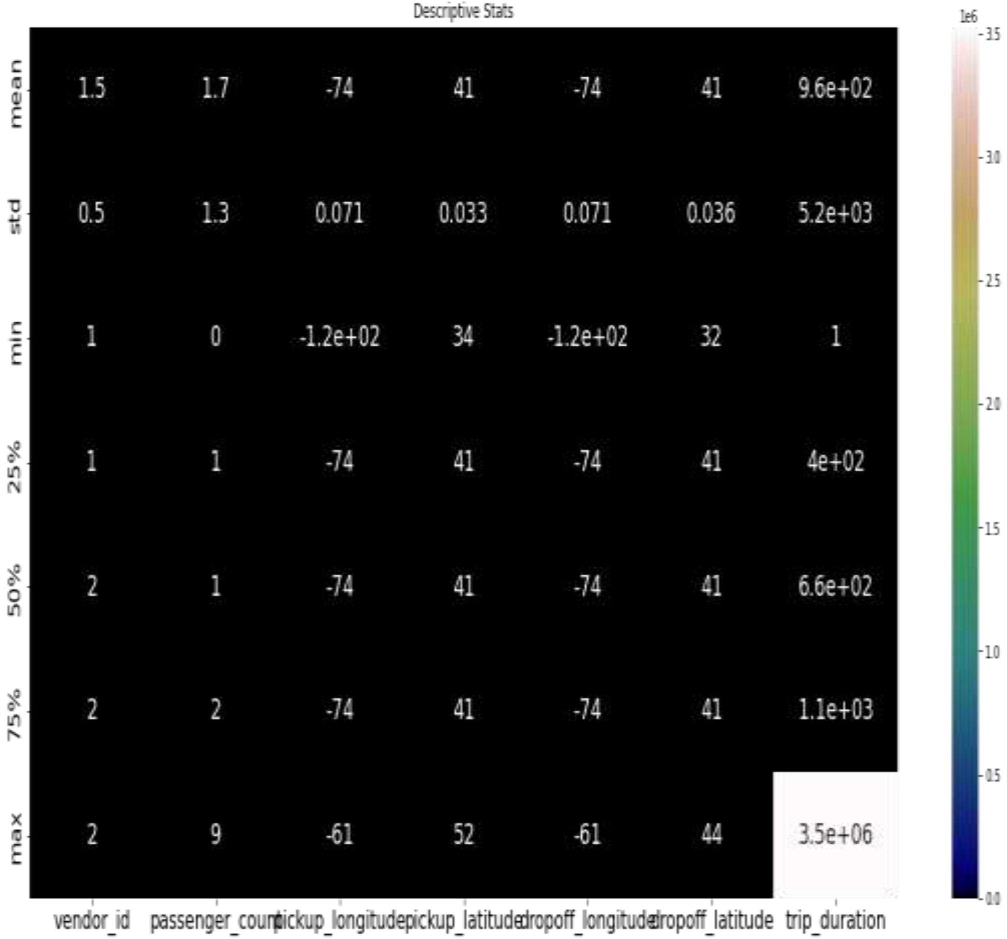
EDA and Data Processing



Descriptive Stats Visualisation

> In this graph I can observe that there were trips having 0 passengers can consider as false trips.

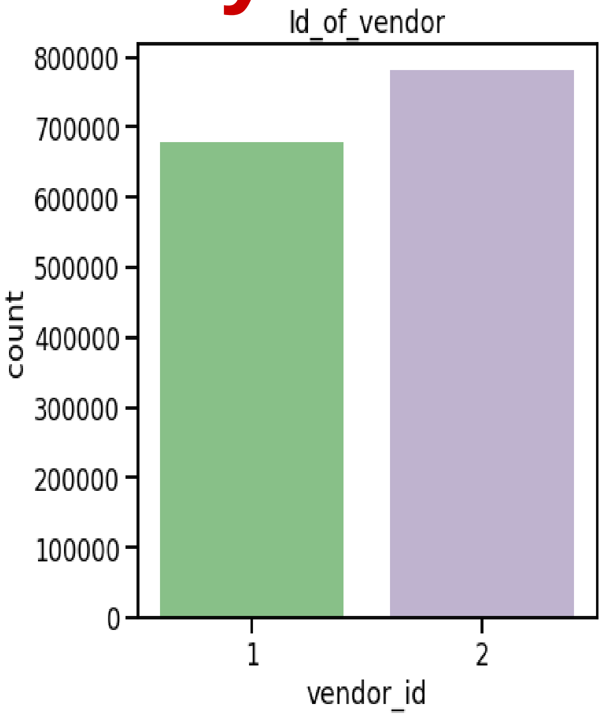
> Also there are trips having very large duration which is kind of impossible in a day.

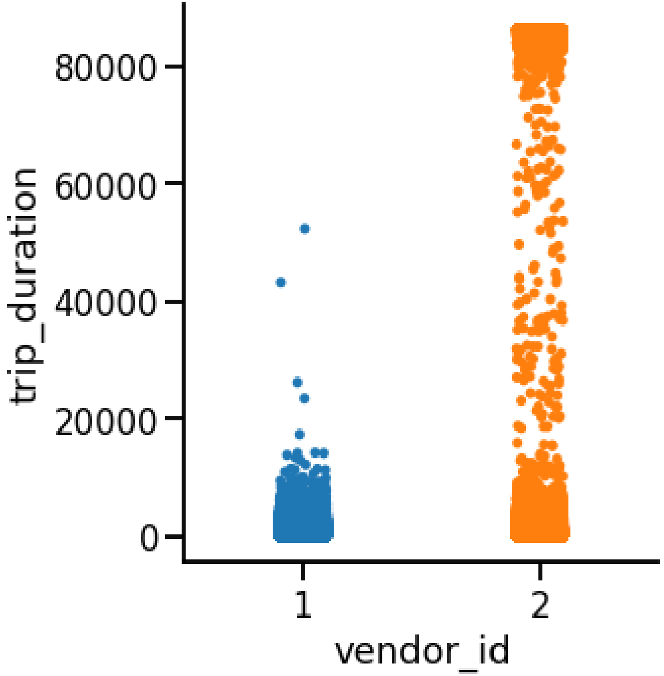


Descriptive Stats

	vendor_id	passenger_count	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	trip_duration
mean	1.5	1.7	-74	41	-74	41	9.6e+02
std	0.5	1.3	0.071	0.033	0.071	0.036	5.2e+03
min	1	0	-1.2e+02	34	-1.2e+02	32	1
25%	1	1	-74	41	-74	41	4e+02
50%	2	1	-74	41	-74	41	6.6e+02
75%	2	2	-74	41	-74	41	1.1e+03
max	2	9	-61	52	-61	44	3.5e+06

Analysis on : Vendor id

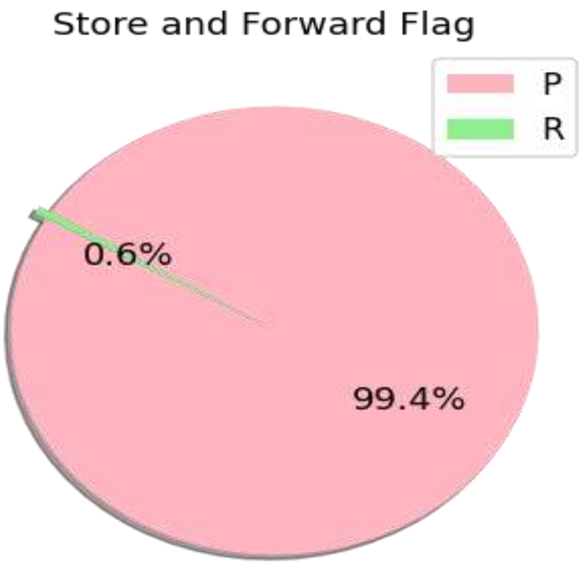




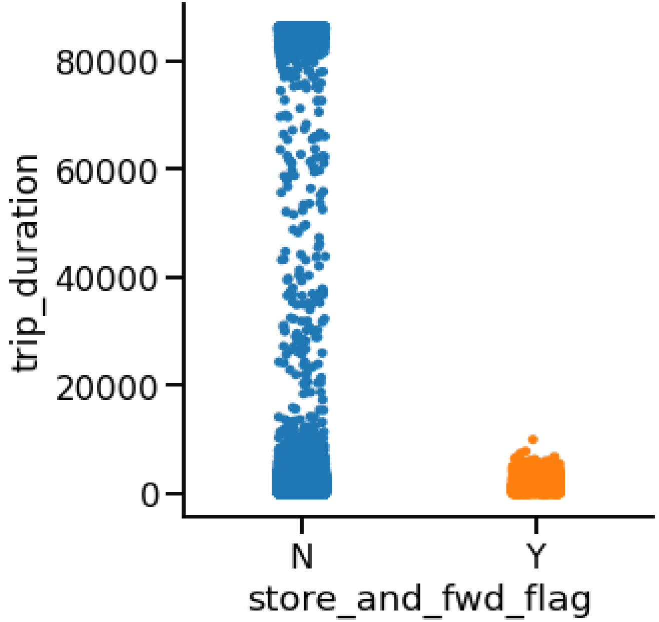
From above visualisation I can say that there are two vendors (service providers) and 2nd one is most opted by people of New York.

Vendor 2 has taken longer trips as compared to vendor 1

Analysis on : Store and fwd flag

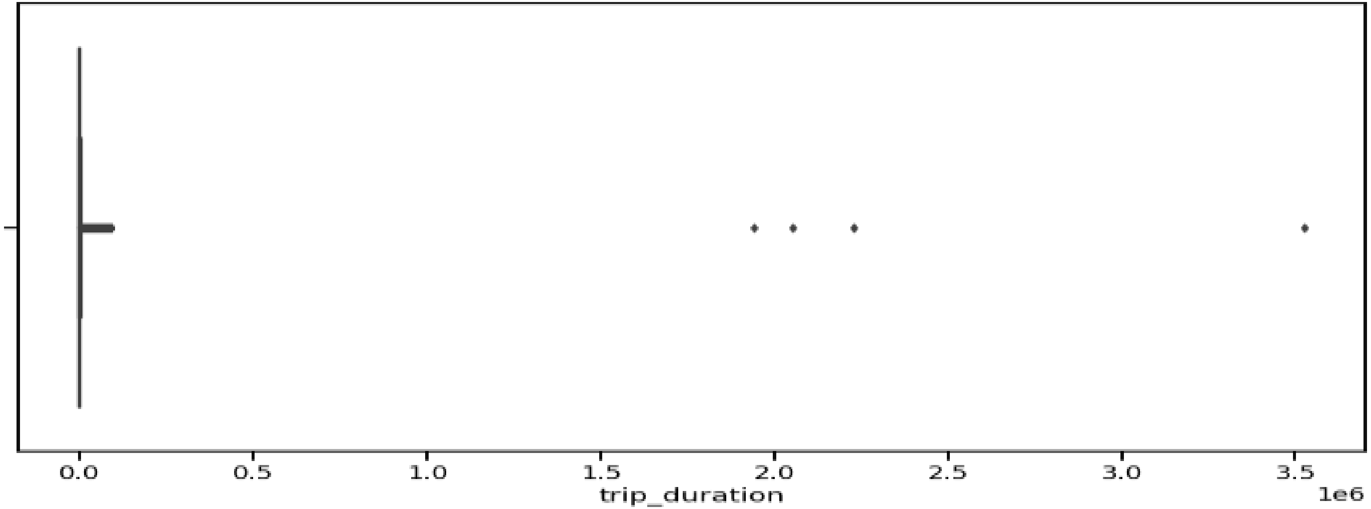


Trips duration is generally longer for those trips having whose flag was not sorted.



From above observation I can conclude that there are less than 1 % of trips that were sorted before forwarding. The number of P flag is much larger than R.

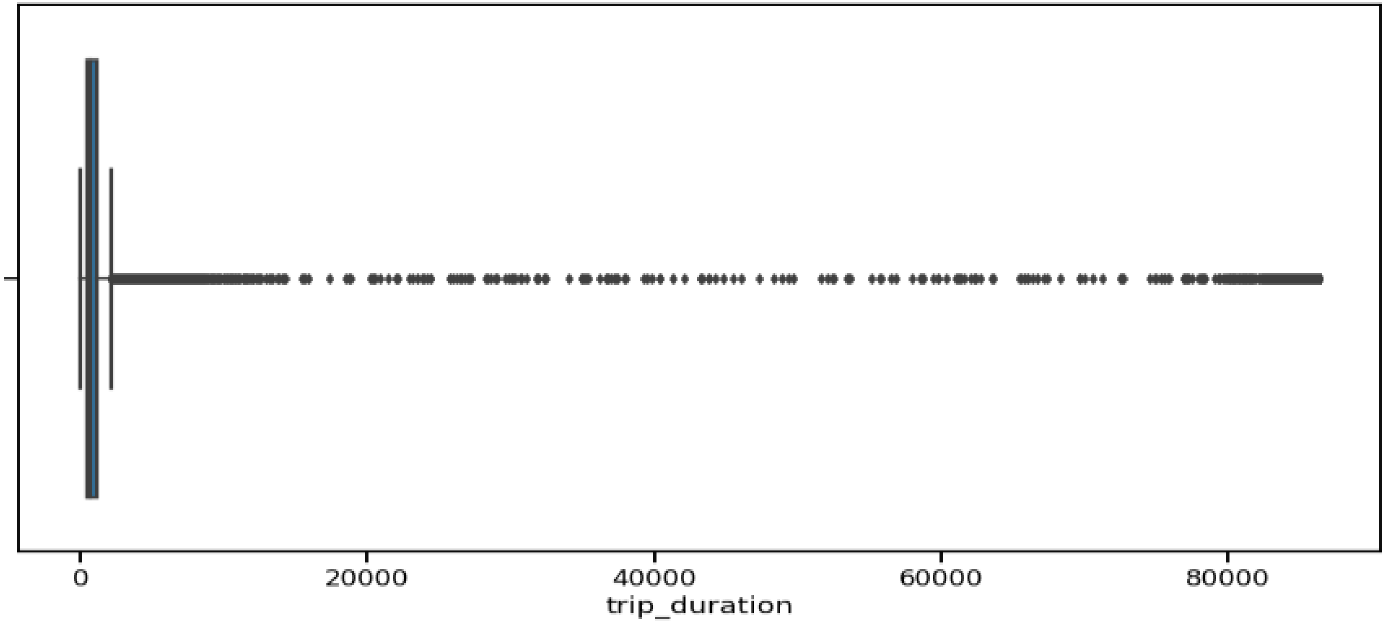
Analysis on: Target Variable



Probably is this visualisation one can clearly see some outliers and their trips are lasting for very long hours (approx. 972) which is impossible in case of Taxi Trips.

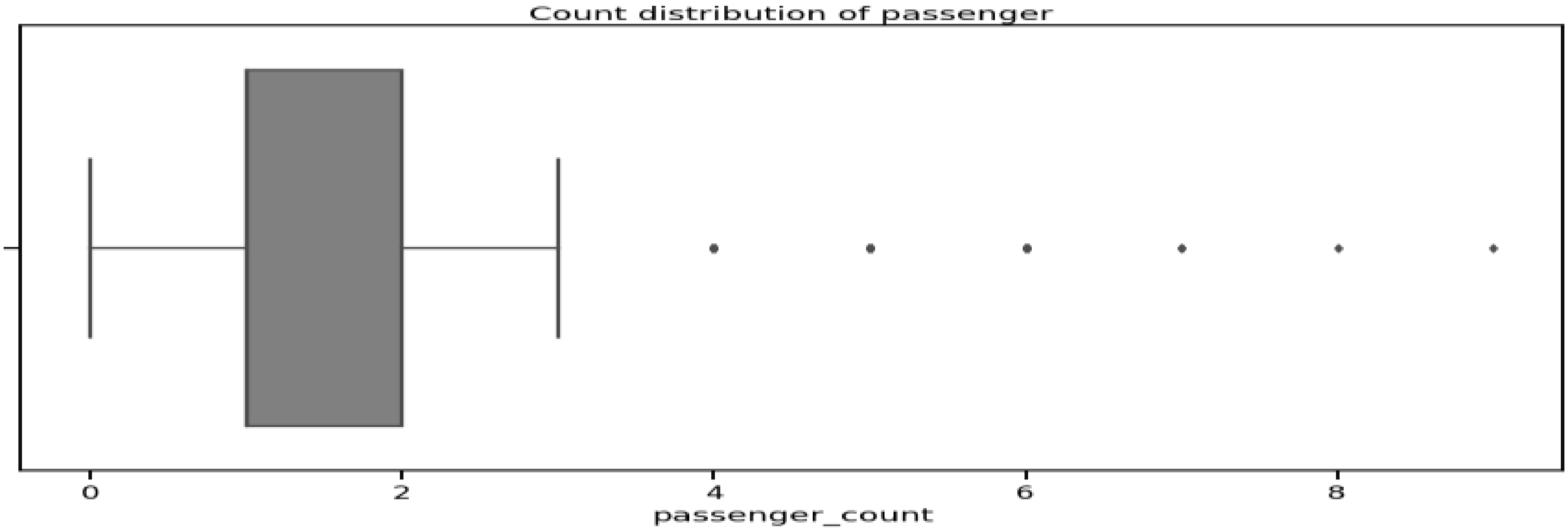
How can a taxi trip be that long ? It's quite suspicious.

Analysis on : Target Variable – Trip duration



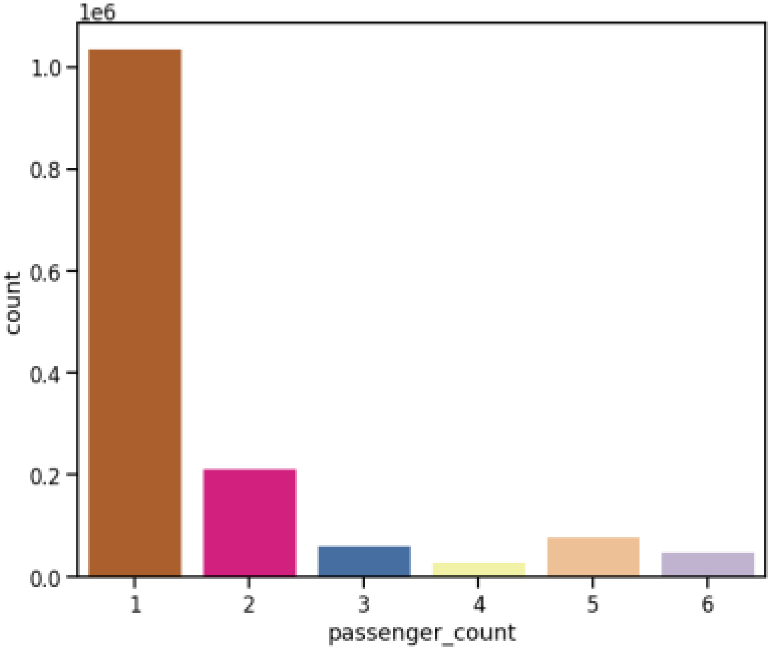
From above visualisation we can clearly observe that there is some entries which is significantly differ from others. As there is 4 rows only, want to drop these rows.

Analysis on : Passenger count



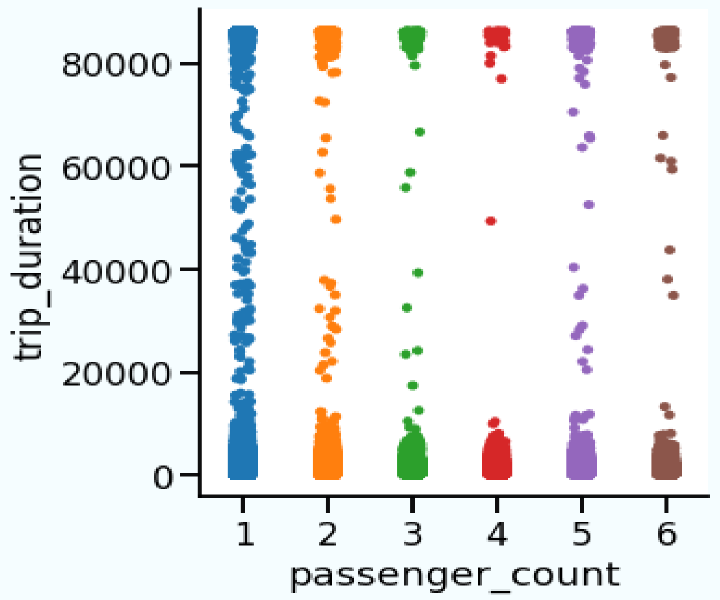
There are some trips with even 0 passenger count and 3 trips with 7 passenger and there is only 1 trip each for 8 -9 passengers.
Above visualisation depicts us that there were most number of trips done by 1 to 2 passenger(s). 5-9 passenger trip states us that cab is large vehicle.

Analysis on : Passenger count (cont)

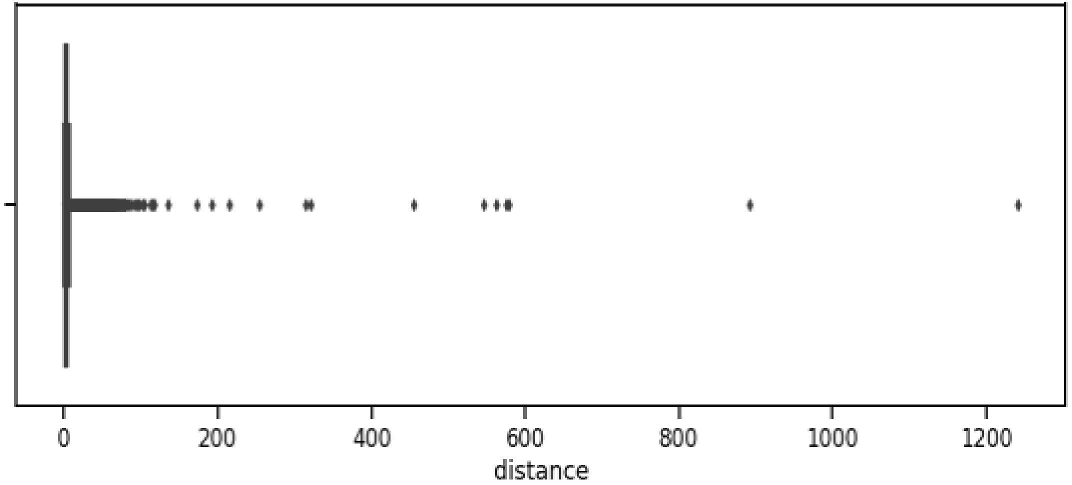


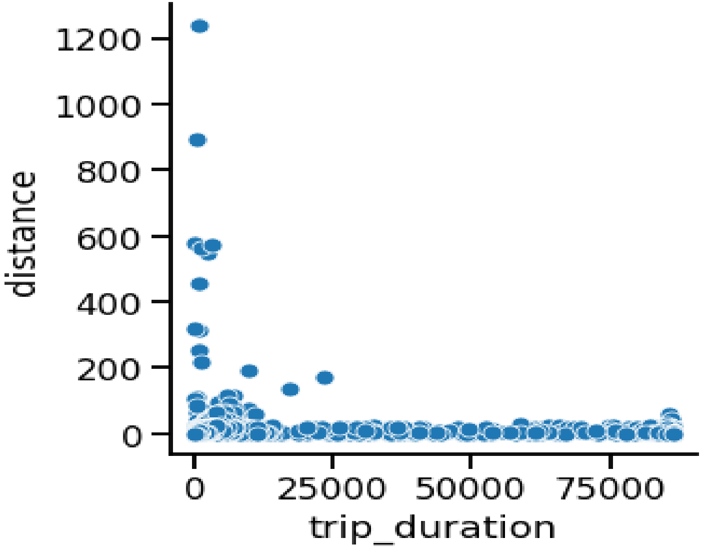
Nothing that seems like a fair distribution.
Here I can see large amount of trips are with one passenger.

There is no visible relationship between Trip duration and passenger count



Analysis on : Distance



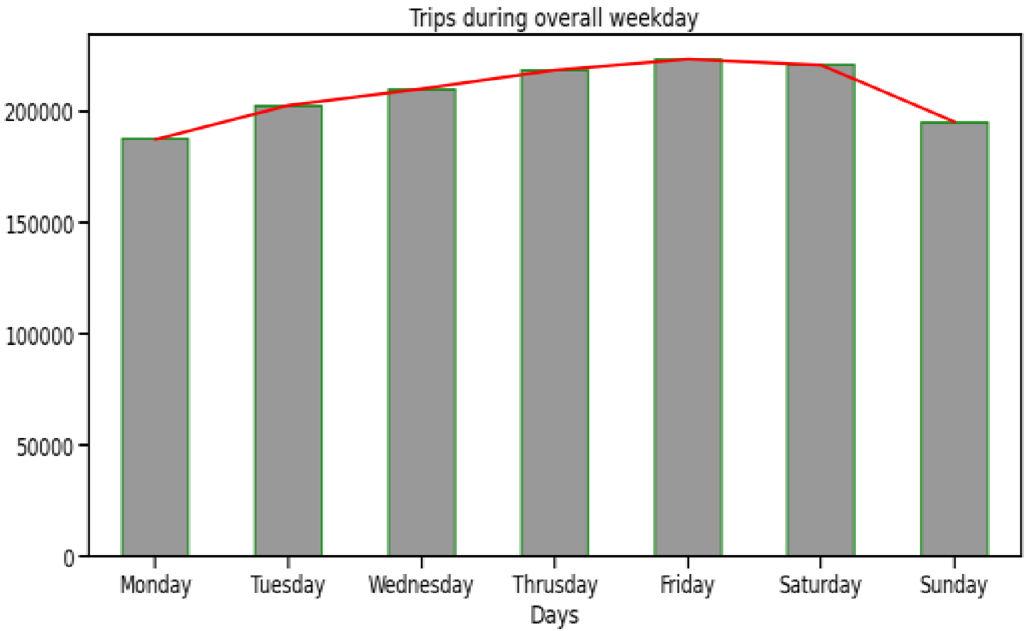


We see there are trips which trip duration is as short as 0 sec and yet covering a large distance . And some trips with 0 km distance and longer trip duration.

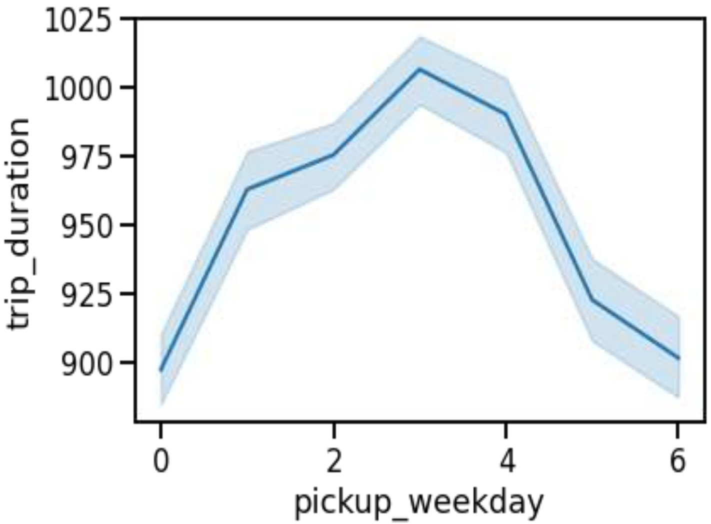
The reasons for 0 km distance can be :

1. **The dropoff location couldn’t be traced.**
2. **The driver deliberately took this ride just to cover the target ride number.**
3. **The passengers cancelled the trip.**

Analysis on : Trip duration on weekday

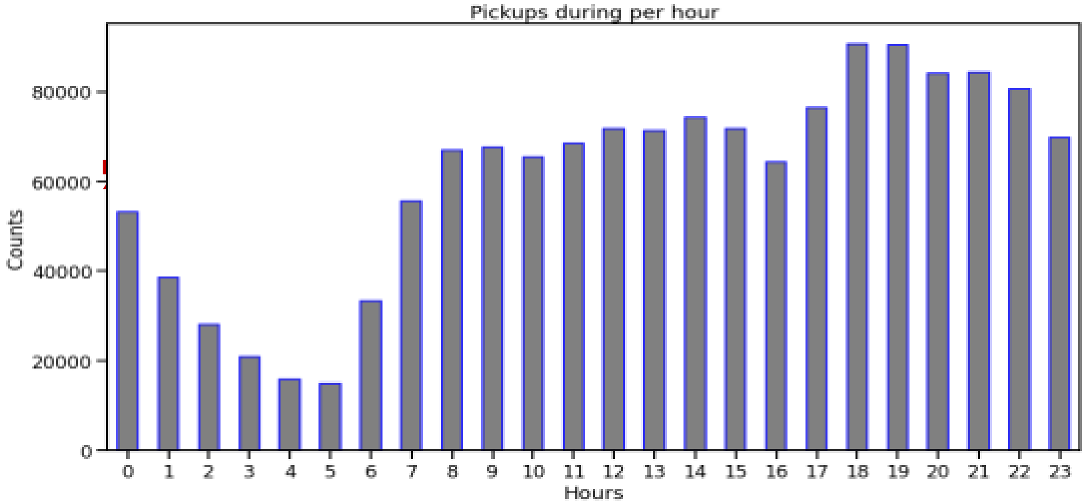


Trip duration is longest on Thrusdays and closely followed by Fridays.

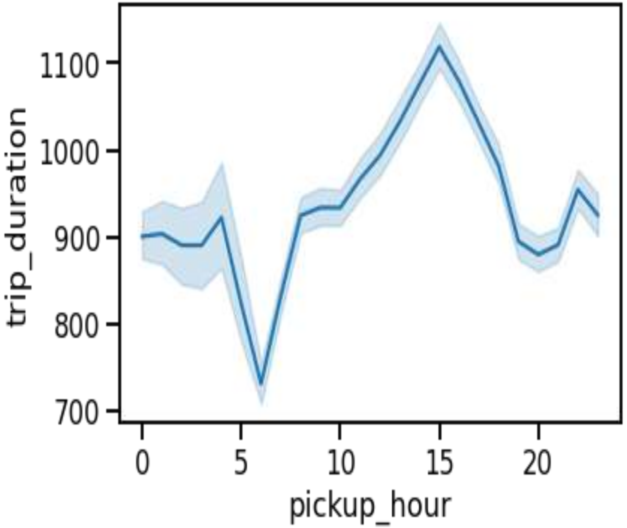


Above observation tells us most of people in NYC comes out from their home on Fridays and Saturdays.

Analysis on : Trip duration per hour

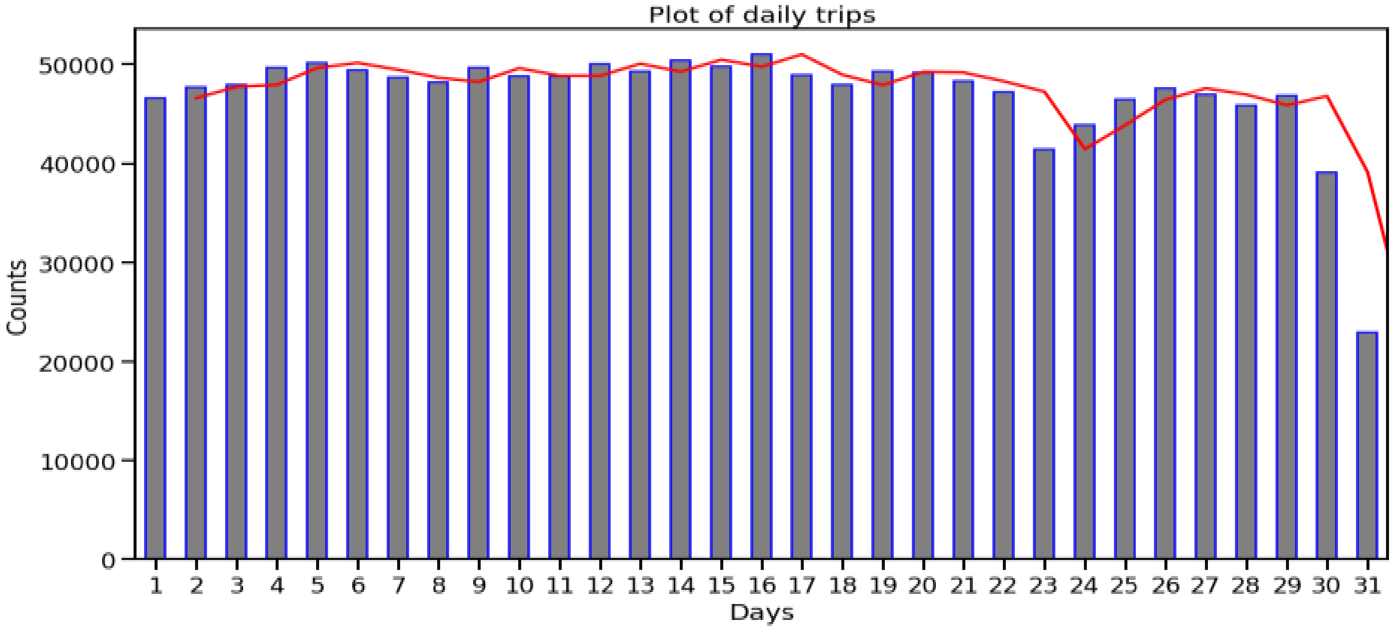


Trip duration is maximum around 3 pm may be because of traffic on the roads and lowest around 6 am as there roads are empty at this time.



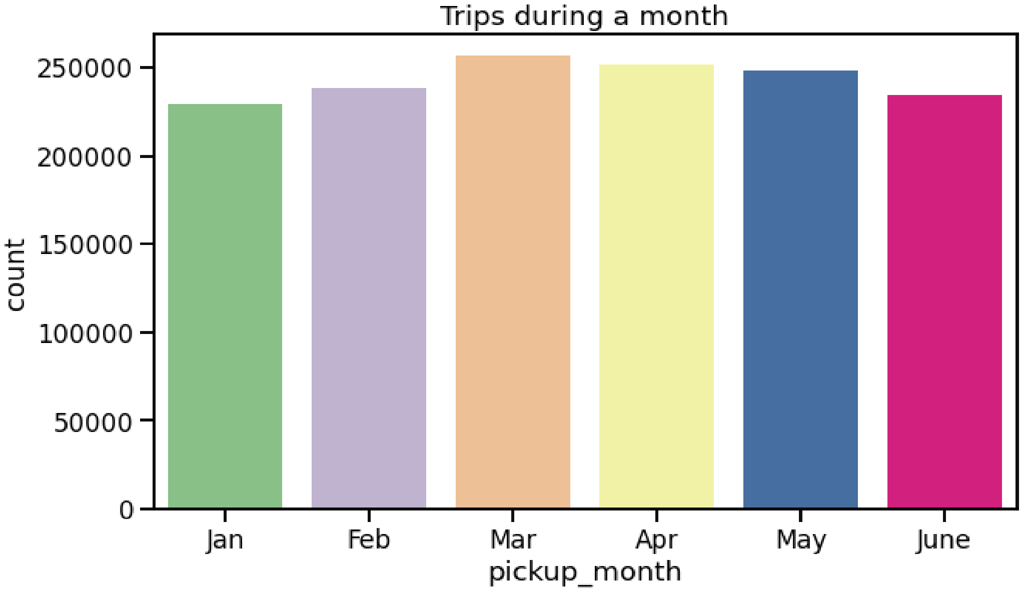
Maximum pickups occuring between rush hours (5pm-10 pm) prolly it's office leaving time. Thus I can see that most pickups and dropoffs happen in the evening time while least occur in midday time.

Analysis on: Trip duration in a month

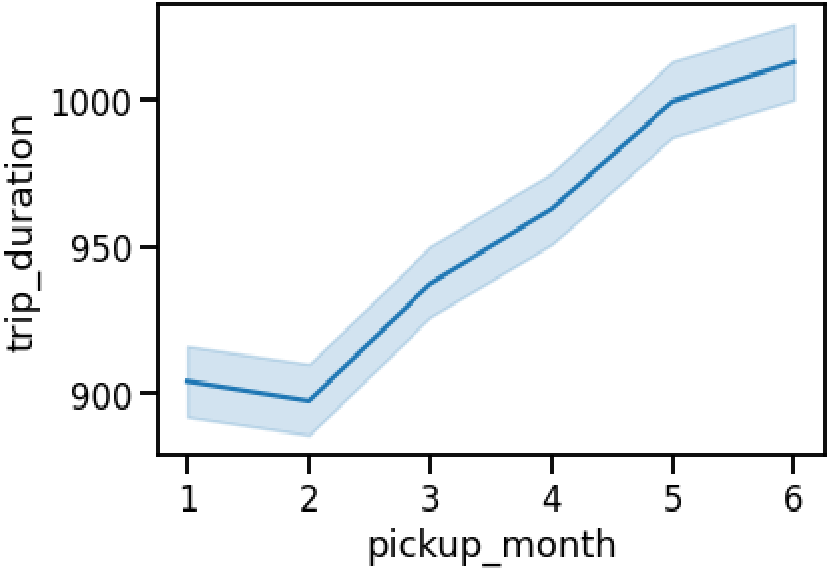


There is a significant drop in Taxi Trip count as the month end approaches and seems like people of NYC doesn’t prefer taxi in month end.

Analysis on: Trip duration in 6 months

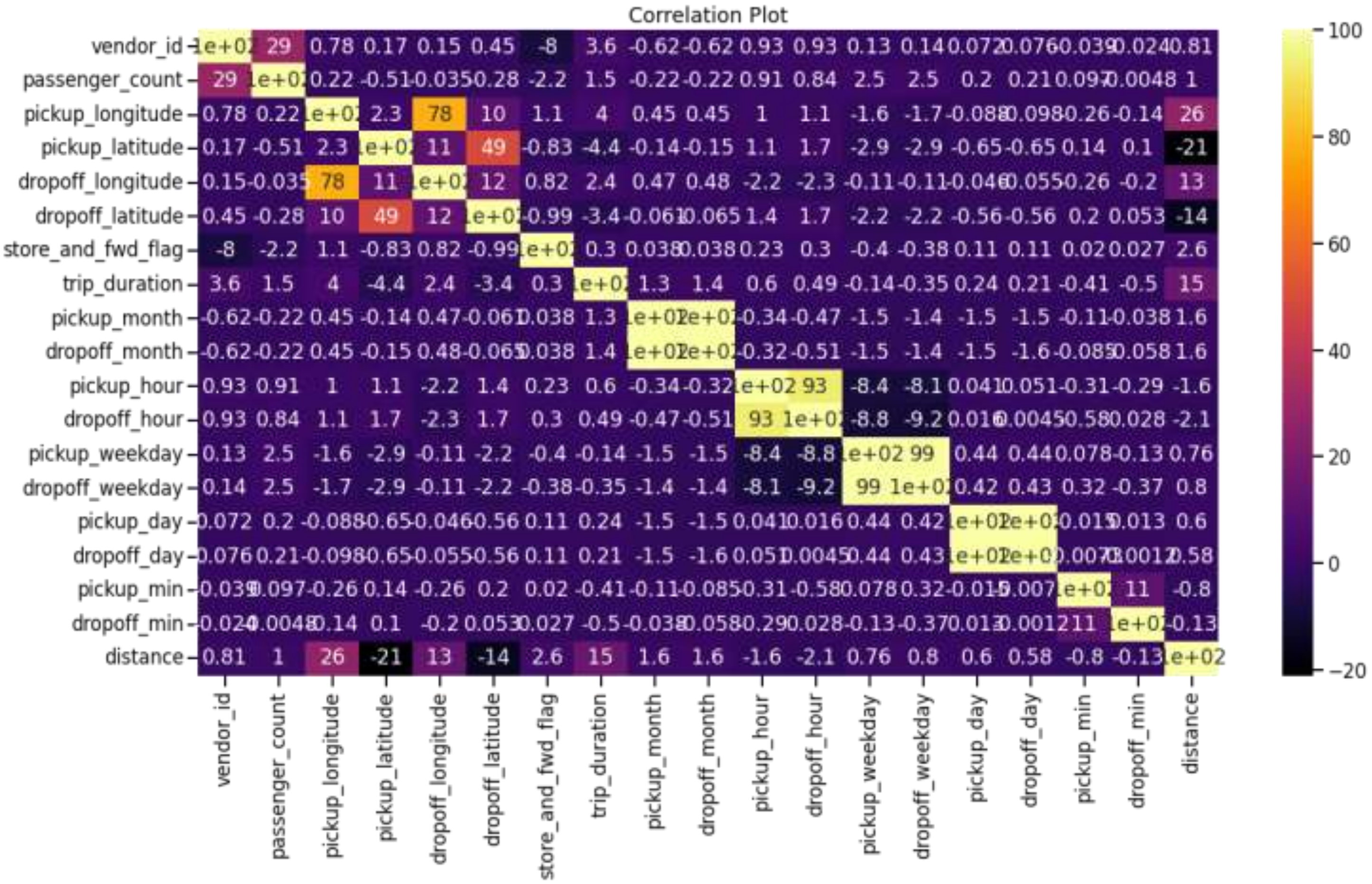


From February trip duration rising every month



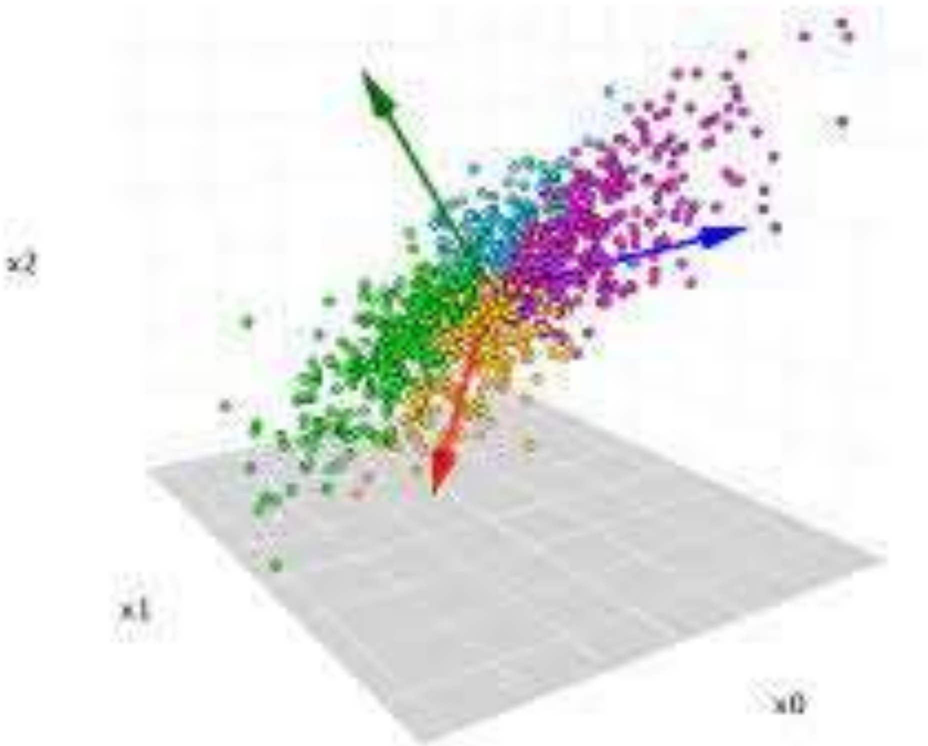
Here, the data of 6 months.
Number of trips in particular month – March and April is marking the highest.
January being the lowest due to extreme snowfall .

Analysis on: Correlation heat map

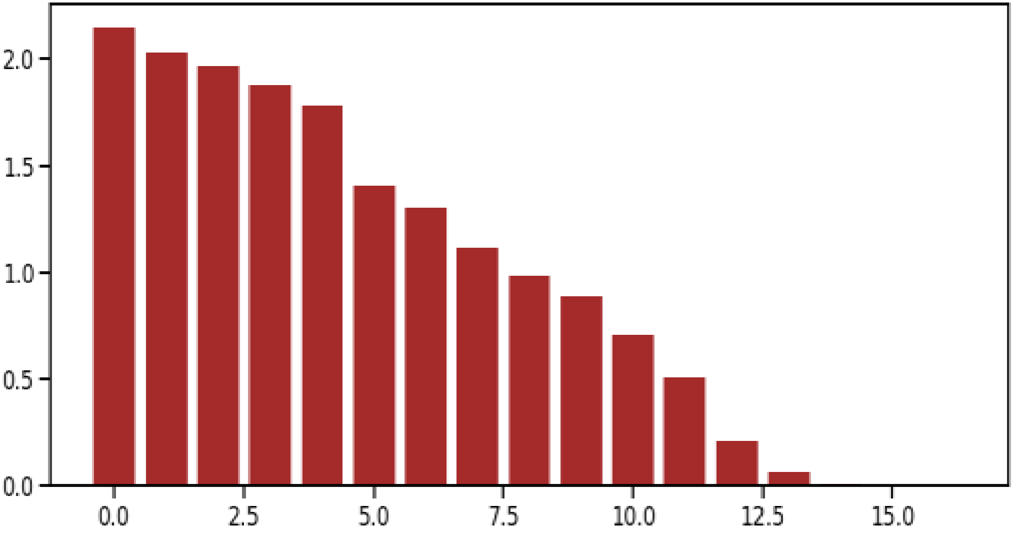


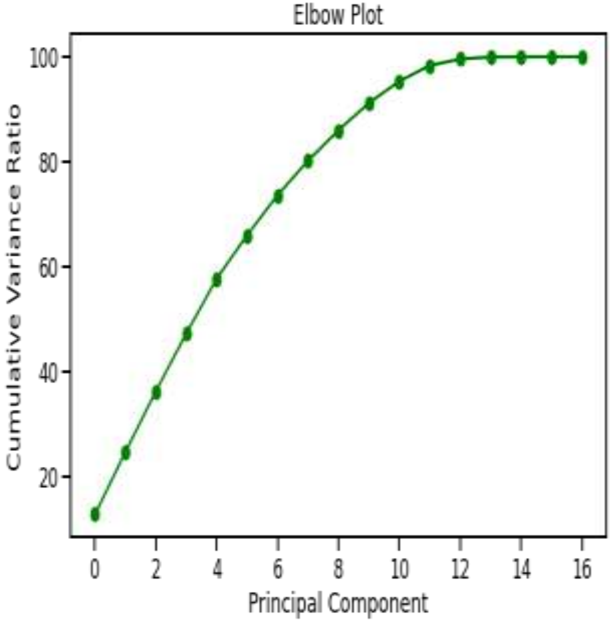
	vendor_id	passenger_count	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	store_and_fwd_flag	trip_duration	pickup_month	dropoff_month	pickup_hour	dropoff_hour	pickup_weekday	dropoff_weekday	pickup_day	dropoff_day	pickup_min	dropoff_min	distance
vendor_id	1e+02	29	0.78	0.17	0.15	0.45	-8	3.6	-0.62	-0.62	0.93	0.93	0.13	0.14	0.072	0.076	-0.039	-0.024	0.81
passenger_count	29	1e+02	0.22	-0.51	-0.035	0.28	-2.2	1.5	-0.22	-0.22	0.91	0.84	2.5	2.5	0.2	0.21	0.097	0.0048	1
pickup_longitude	0.78	0.22	1e+02	2.3	78	10	1.1	4	0.45	0.45	1	1.1	-1.6	-1.7	-0.088	-0.098	-0.26	-0.14	26
pickup_latitude	0.17	-0.51	2.3	1e+02	11	49	-0.83	-4.4	-0.14	-0.15	1.1	1.7	-2.9	-2.9	-0.65	-0.65	0.14	0.1	-21
dropoff_longitude	0.15	-0.035	78	11	1e+02	12	0.82	2.4	0.47	0.48	-2.2	-2.3	-0.11	-0.11	-0.046	-0.055	-0.26	-0.2	13
dropoff_latitude	0.45	-0.28	10	49	12	1e+02	-0.99	-3.4	-0.061	-0.065	1.4	1.7	-2.2	-2.2	-0.56	-0.56	0.2	0.053	-14
store_and_fwd_flag	-8	-2.2	1.1	-0.83	0.82	-0.99	1e+02	0.3	0.038	0.038	0.23	0.3	-0.4	-0.38	0.11	0.11	0.02	0.027	2.6
trip_duration	3.6	1.5	4	-4.4	2.4	-3.4	0.3	1e+02	1.3	1.4	0.6	0.49	-0.14	-0.35	0.24	0.21	-0.41	-0.5	15
pickup_month	-0.62	-0.22	0.45	-0.14	0.47	-0.061	0.038	1.3	1e+02	1e+02	-0.34	-0.47	-1.5	-1.4	-1.5	-1.5	-0.11	-0.038	1.6
dropoff_month	-0.62	-0.22	0.45	-0.15	0.48	-0.065	0.038	1.4	1e+02	1e+02	-0.32	-0.51	-1.5	-1.4	-1.5	-1.6	-0.085	-0.058	1.6
pickup_hour	0.93	0.91	1	1.1	-2.2	1.4	0.23	0.6	-0.34	-0.32	1e+02	93	-8.4	-8.1	0.041	0.051	-0.31	-0.29	-1.6
dropoff_hour	0.93	0.84	1.1	1.7	-2.3	1.7	0.3	0.49	-0.47	-0.51	93	1e+02	-8.8	-9.2	0.016	0.0045	-0.58	0.028	-2.1
pickup_weekday	0.13	2.5	-1.6	-2.9	-0.11	-2.2	-0.4	-0.14	-1.5	-1.5	-8.4	-8.8	1e+02	99	0.44	0.44	0.078	-0.13	0.76
dropoff_weekday	0.14	2.5	-1.7	-2.9	-0.11	-2.2	-0.38	-0.35	-1.4	-1.4	-8.1	-9.2	99	1e+02	0.42	0.43	0.32	-0.37	0.8
pickup_day	0.072	0.2	-0.088	-0.65	-0.046	-0.56	0.11	0.24	-1.5	-1.5	0.041	0.016	0.44	0.42	1e+02	1e+02	-0.015	0.013	0.6
dropoff_day	0.076	0.21	-0.098	-0.65	-0.055	-0.56	0.11	0.21	-1.5	-1.6	0.051	0.0045	0.44	0.43	1e+02	1e+02	0.007	0.0012	0.58
pickup_min	-0.039	0.097	-0.26	0.14	-0.26	0.2	0.02	-0.41	-0.11	-0.085	-0.31	-0.58	0.078	0.32	-0.015	0.007	1e+02	11	-0.8
dropoff_min	-0.024	0.0048	-0.14	0.1	-0.2	0.053	0.027	-0.5	-0.038	-0.058	-0.29	0.028	-0.13	-0.37	0.013	0.0012	11	1e+02	-0.13
distance	0.81	1	26	-21	13	-14	2.6	15	1.6	1.6	-1.6	-2.1	0.76	0.8	0.6	0.58	-0.8	-0.13	1e+02

Decomposition of data : PCA



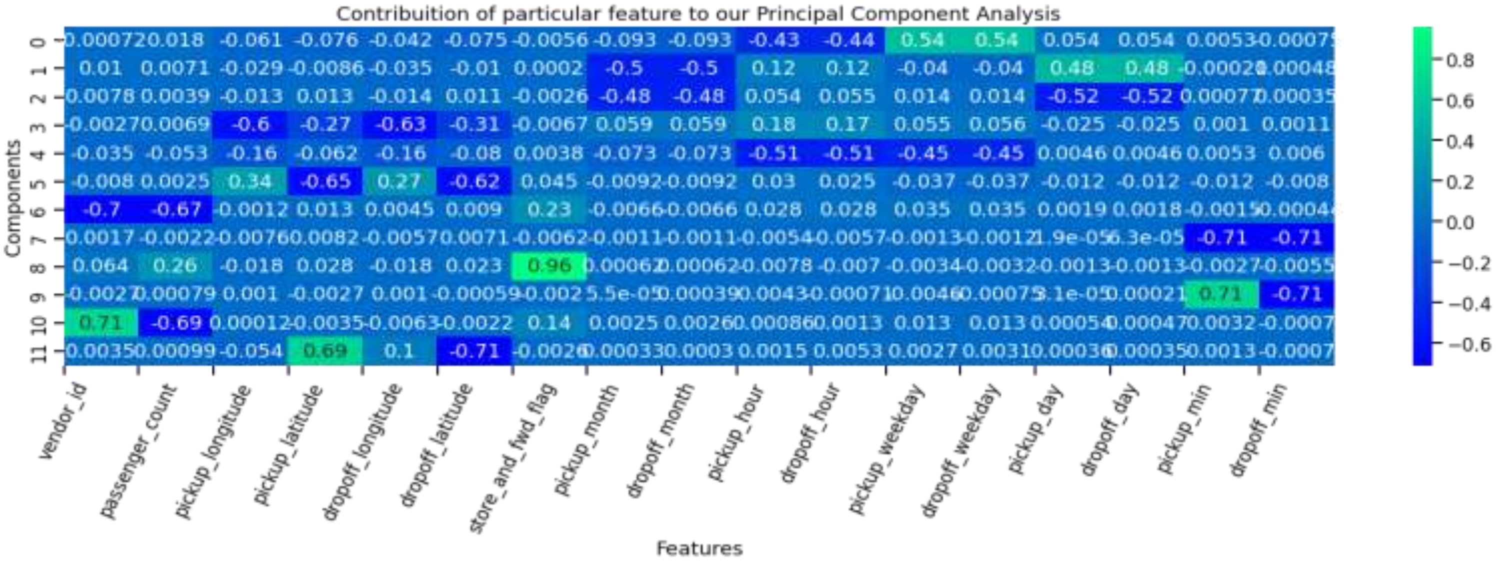
Analysis on: Principal Component Analysis





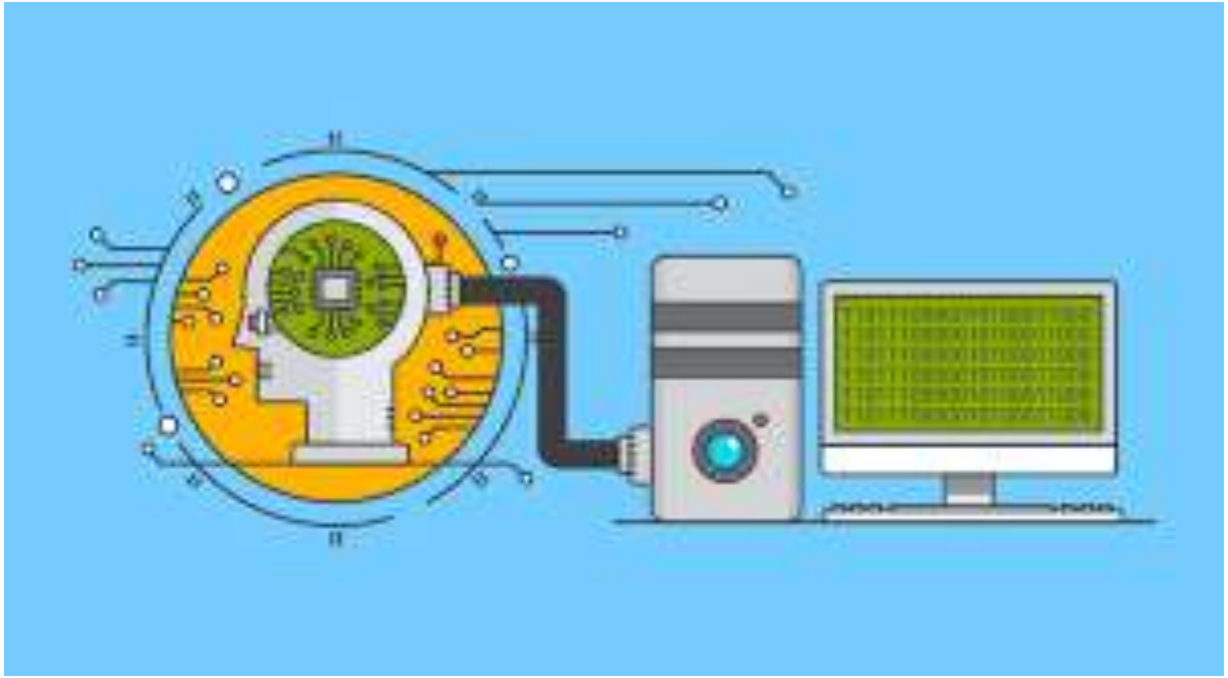
Now , we have to pass our scaled DataFrame in our PCA model and observe the elbow plot to get better idea about explained variance. At 12th model of PCA component our PCA model seems to go flat without explaining much of variance.

Analysis on: Feature Correlation

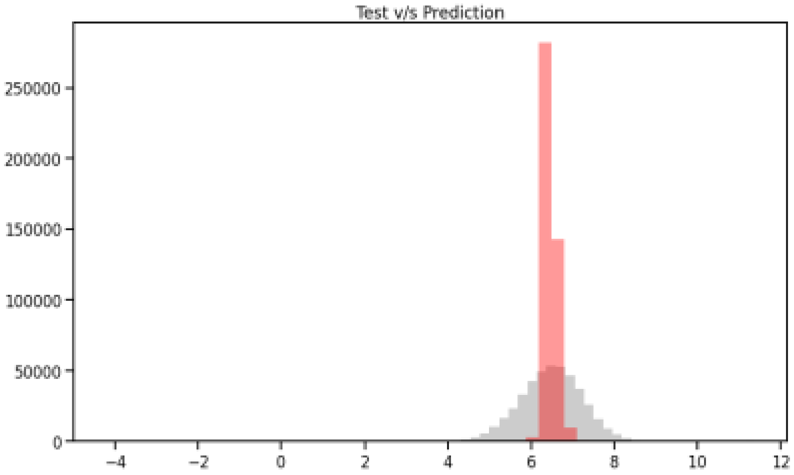


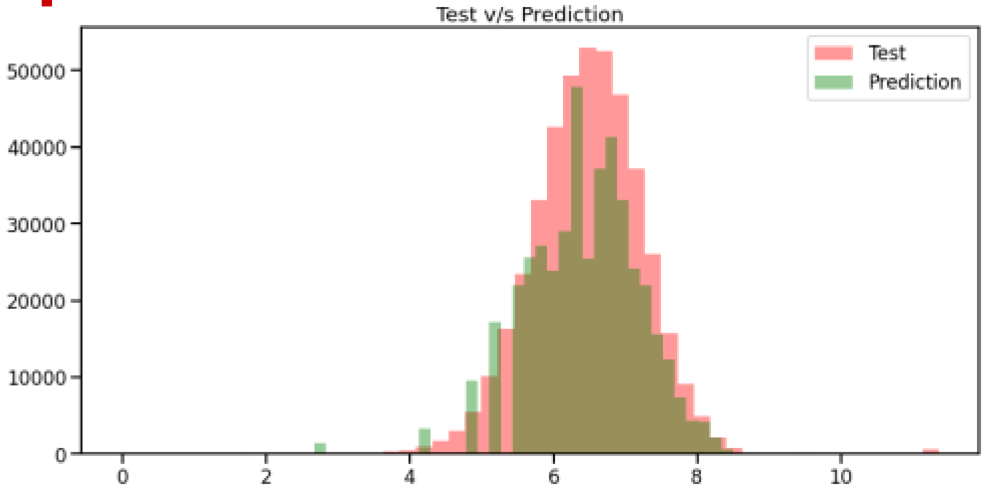
Above visualisation gives us brief idea of which feature contributed more or less to our Principal Component. These are new features which gives the information from every other original feature we have. This method is used for reducing the dimension and retaining the information as much as possible.

Machine learning model regression

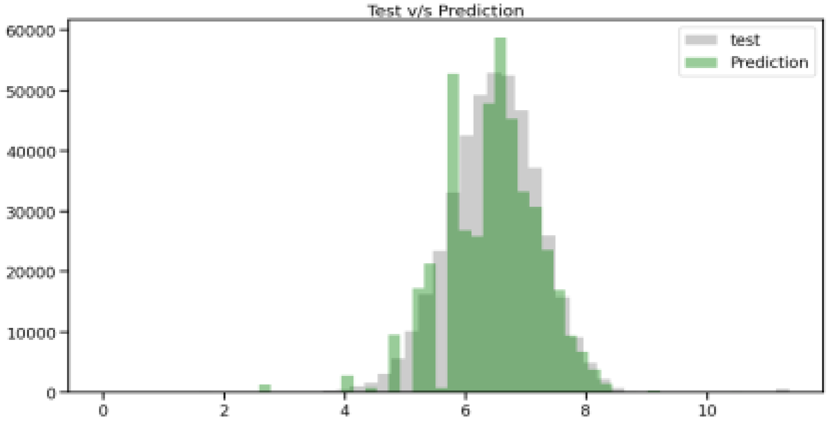


Analysis on: ML model prediction with PCA





These visualisations shows us that how our model predictions close to test data. It is evident that Decision Tree and Random Forest are performing well.



Analysis on: Model evaluation result with PCA

1. We can clearly see that Decision Tree and Random Forest are performing well.

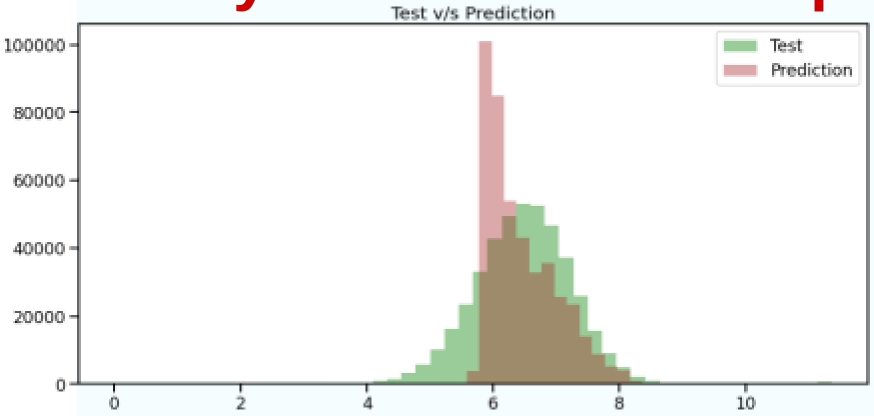
2. As, Random Forest is providing us reduced RMSE, we can say that it’s a model to opted for.

3. We are getting good score for Decision Forest and Random Tree i.e, close to 1.

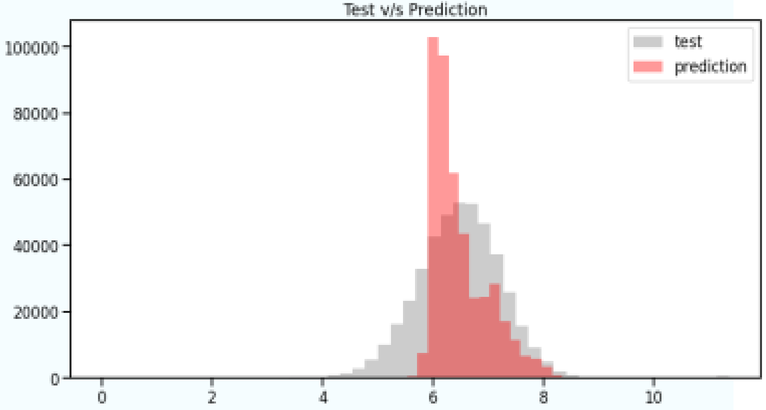
4. R2-score must be between 0 to 1. Score near to 1 considered as good fit score.

5. RMSE : Lesser is better

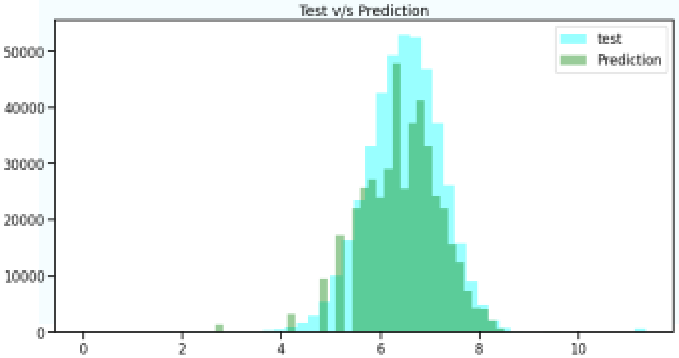
Analysis on: ML model prediction without PCA



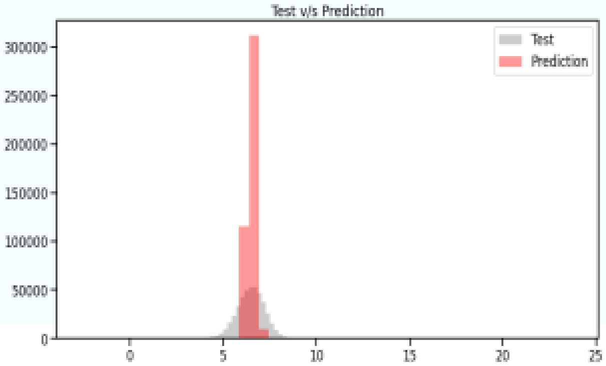
Decision Tree



Random Forest



Decision Tree with hyper parameter tuning



Linear Regression

Analysis on: Model Evaluation result without PCA

1. We can clearly see that Decision Tree with hyper parameter tuning model are good performers. As, it is providing a reduced RMSE we can say that it’s a model to opt for.

2. We are getting good fit score for Decision Tree with hyper parameter tuning i.e, close to 1

3. R2 score must be between 0 to 1, towards 1 is considered as good fit.

4. RMSE: lesser is better

Conclusion-

1. Observed which taxi service provider is mostly used by the people of New York.

2. Found out few trips which is of very larger duration, that is probably a outlier.

3. Passenger count analysis showed us that there few trips having 0 passenger and one trip with 7,8,9 passengers.

4. Monthly trip analysis gives us insight of month – March and April making the highest number of trip and January is making the lowest number of trip.

5. Taxi giant such as UBER and OLA can use the same data for analysing the trends throughout the day in city. This is not only helps in better transport analysis but also help in concerned authorities in planning traffic control and monitoring.

THANK YOU

By – Rajvee Sharma