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

By-
PRINCE KUMAR JHA

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.

Presentation Outline:

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

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 – 1048575
- Features - 11 (Including Target)
- Target – Trip Duration Important

Columns Name -- ‘id’, 'vendor_id', 'pickup_datetime', 'dropoff_datetime','passenger_count', 'pickup_longitude‘,pickup_latitude', 'dropoff_longitude','dropoff_latitude', 'store_and_fwd_flag', 'trip_duration’.

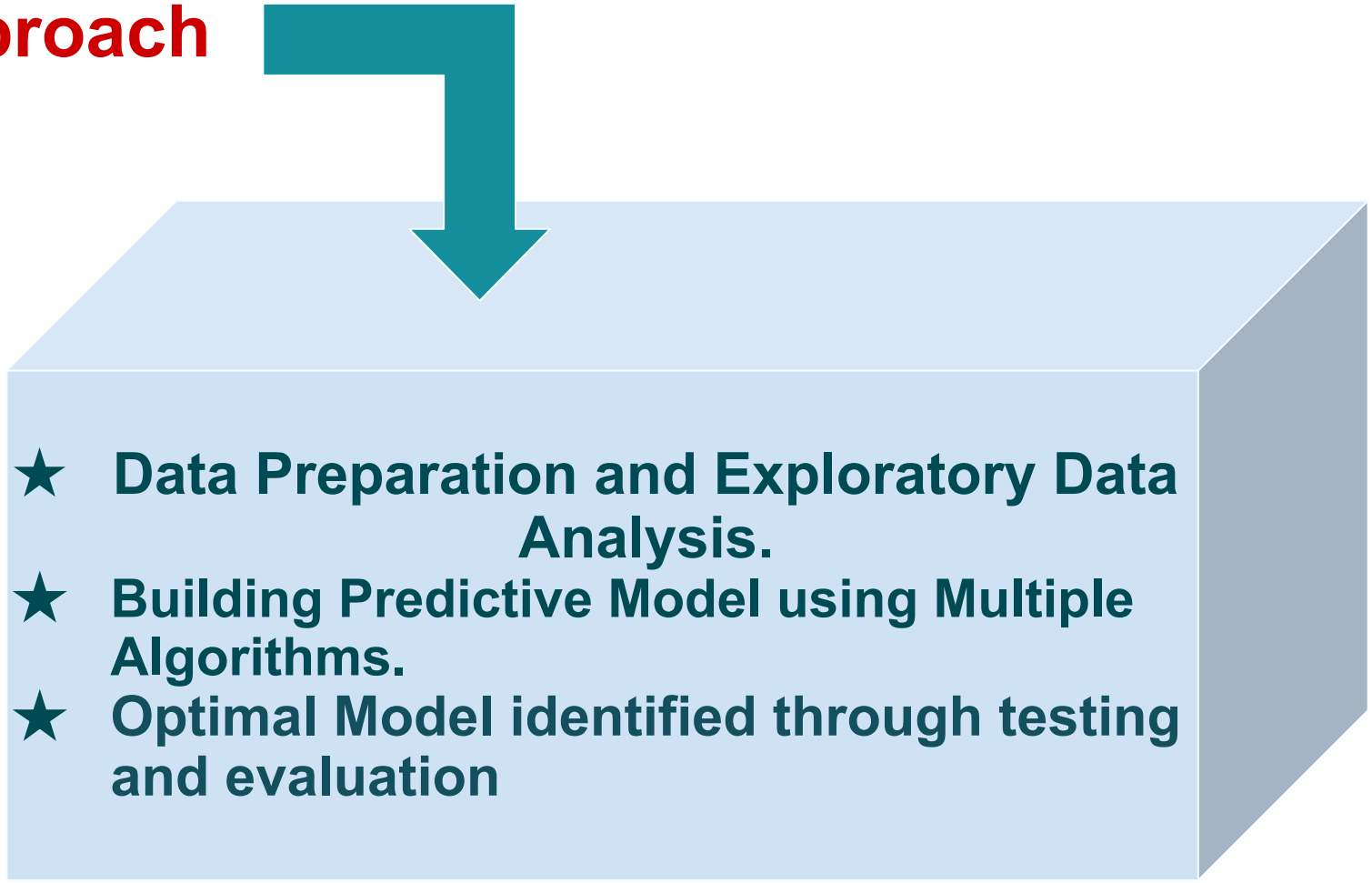
Data Menu:

Independent Variables –

1. id— unique identifier
2. vendor_id—a code indicating the provider associated with the trip record
3. pickup_datetime—date and time when the meter was engaged
4. dropoff_datetime—date and time when the meter was disengaged
5. passenger_count—the number of passengers in the vehicle (driver entered value)
6. pickup_longitude—the longitude where the meter was engaged
7. pickup_latitude—the latitude where the meter was engaged
8. dropoff_longitude—the longitude where the meter was disengaged
9. dropoff_latitude—the latitude where the meter was disengaged
10. 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.
11. **Target Variable –**
12. trip_duration—duration of the trip in seconds

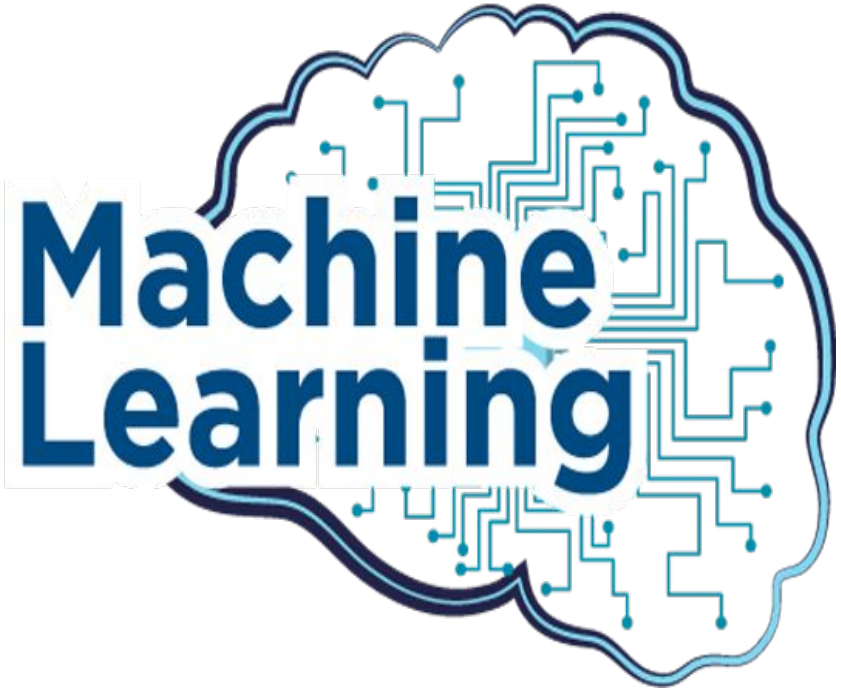
METHODOLOGY

Approach


★ Data Preparation and Exploratory Data Analysis.
★ Building Predictive Model using Multiple Algorithms.
★ Optimal Model identified through testing and evaluation

Machine Learning Algorithm:

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

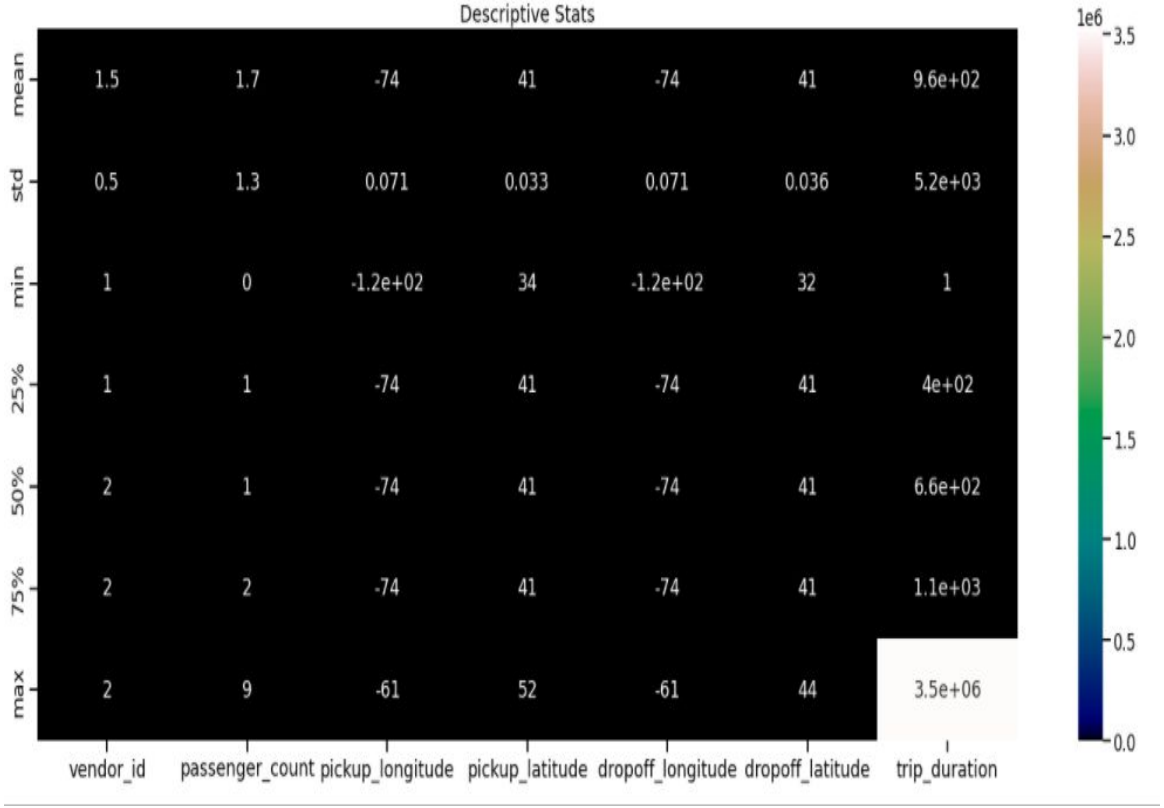


EDA AND DATA PROCESSING

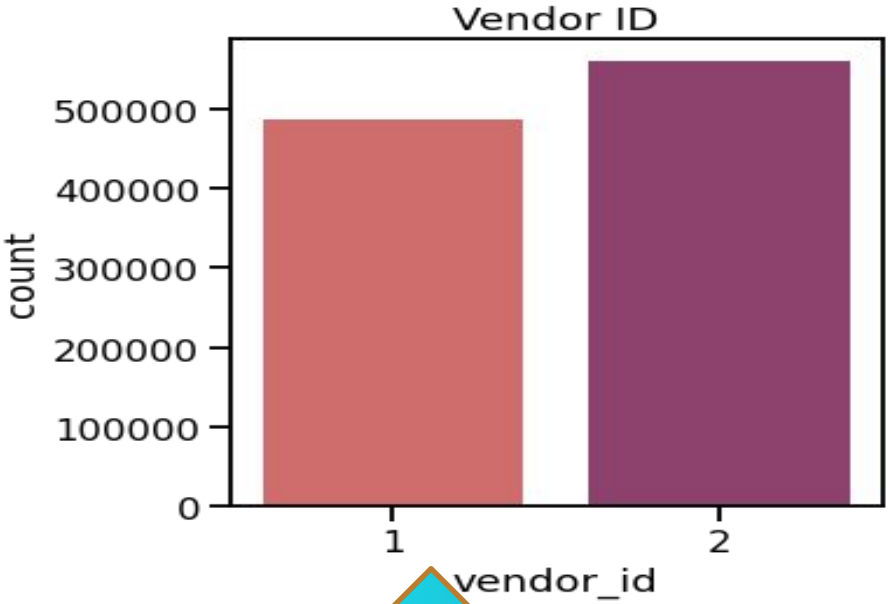
Descriptive Stats in Visual Form

- 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.

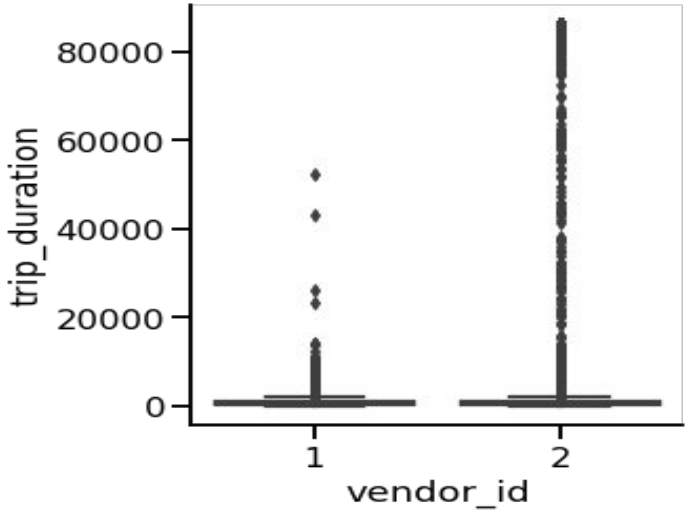


Analysis on : Vendor Id

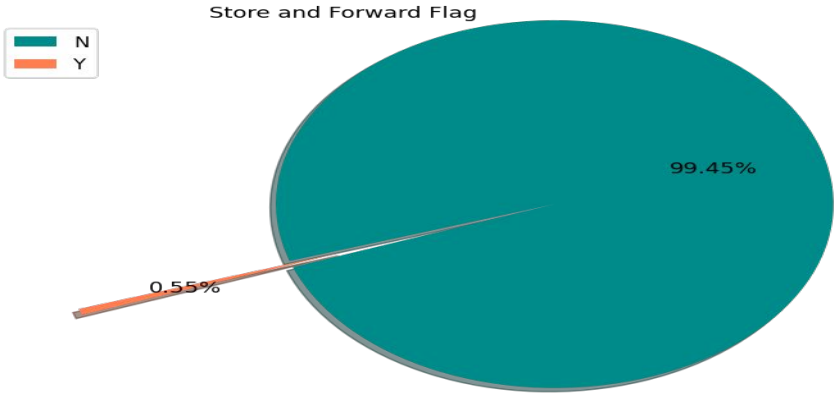


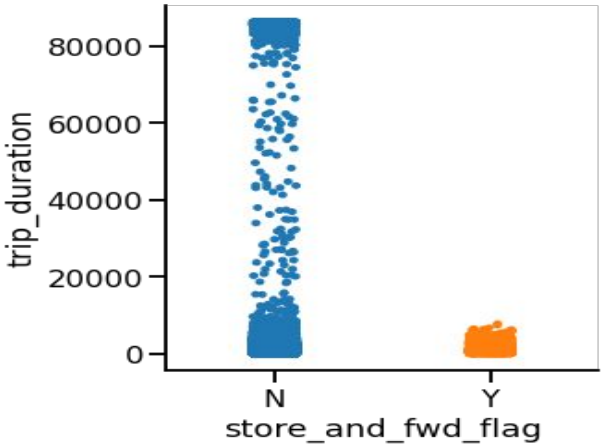
From Above Visualization, we can say that there are 2 vendors (Service Providers).
2nd Service provider is the most opted one by New Yorkers.

Vendor id 2 takes longer trips as compared to vendor 1



Analysis on : Store and Forward Flag

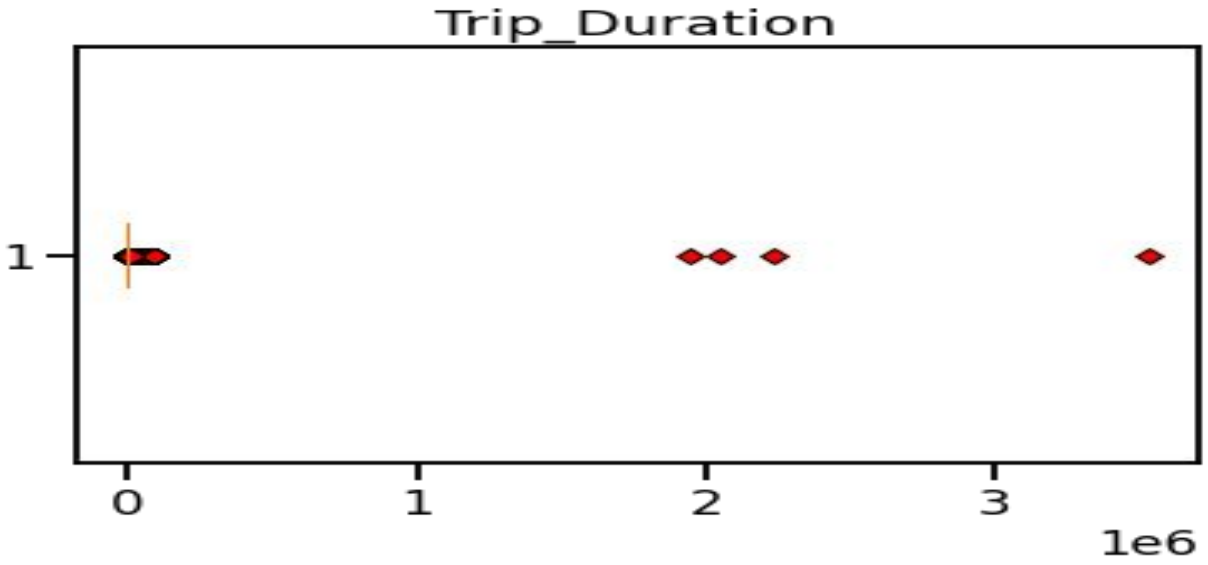




- We see there are less than 1% of trips that were stored before forwarding
- The number of N flag is much larger. We can see the relation with the trip duration

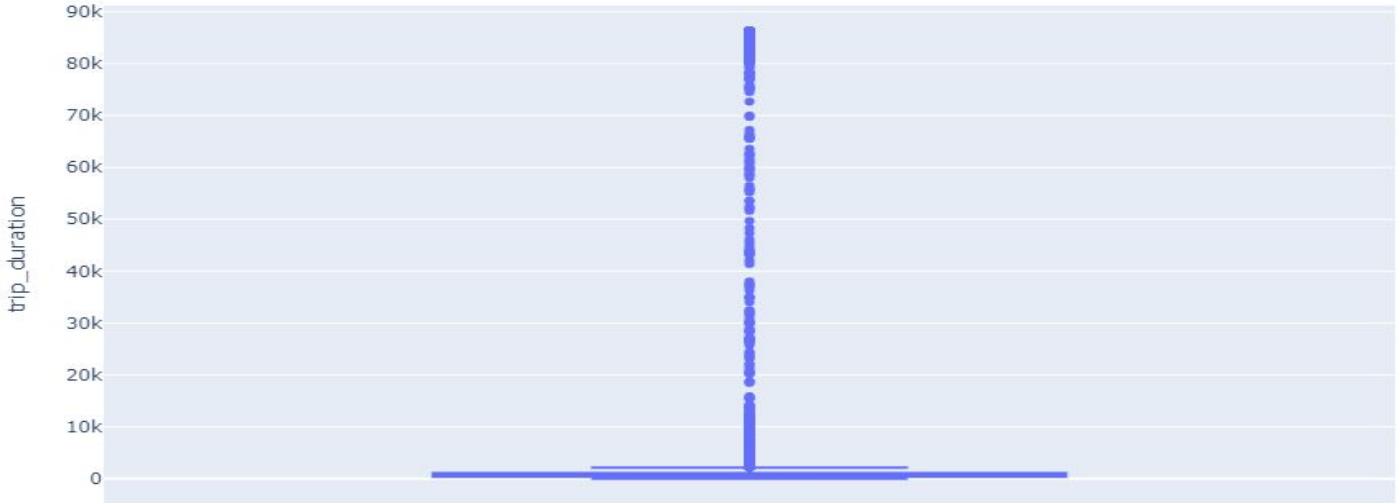
Trip duration is generally longer for trips whose flag was not stored.

Analysis on : Target Variable – Trip Duration



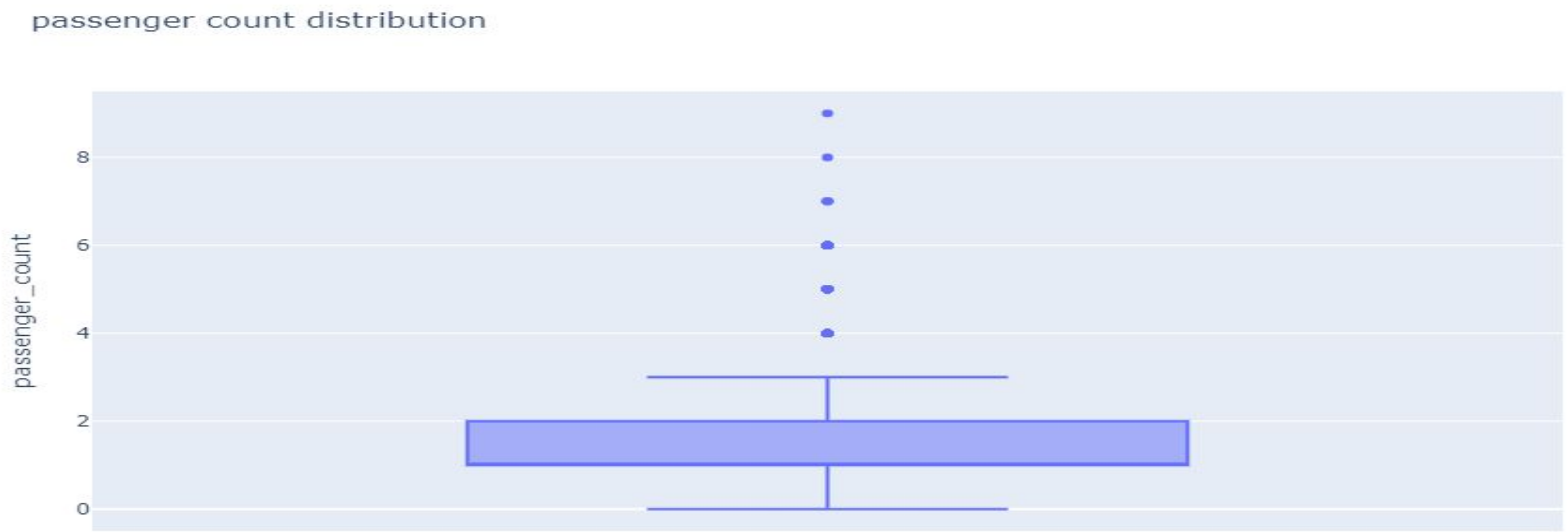
Probably in this visualization we can clearly see some outliers , their trips are lasting between 1900000 seconds (528 Hours) to somewhere around 3500000 seconds (972 hours) which is impossible in case of taxi trips , How can a taxi trip be that long ?It's Quite suspicious. We'll have to get rid of those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

Analysis on : Target Variable – Trip Duration



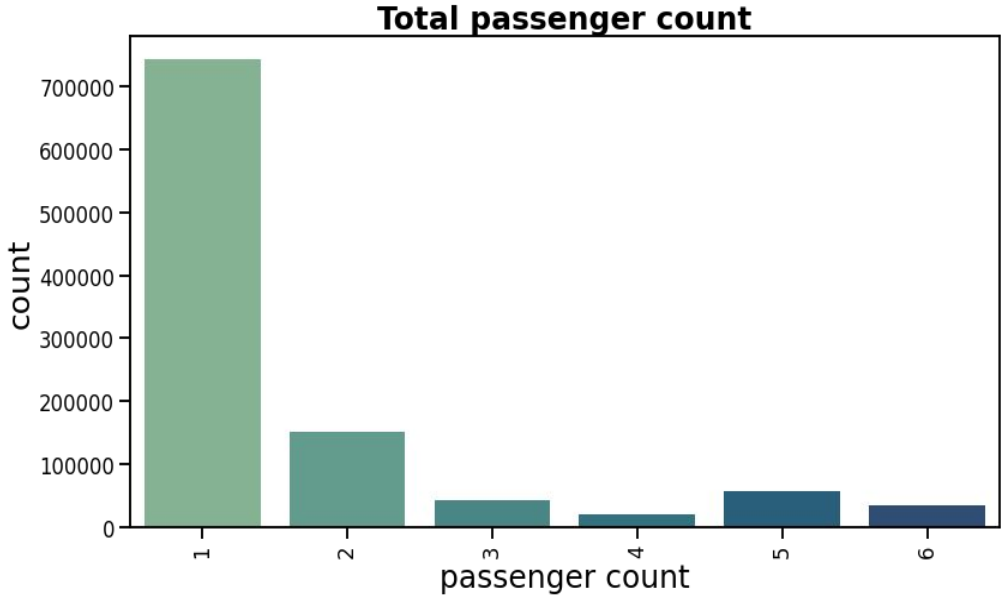
After dropping the outliers you can see now our target variable is fully free from the outliers.

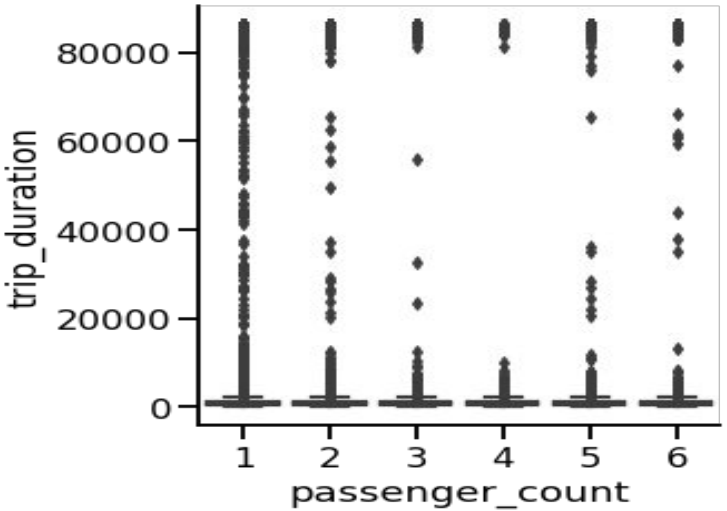
Analysis on : Passenger Count



- 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.)

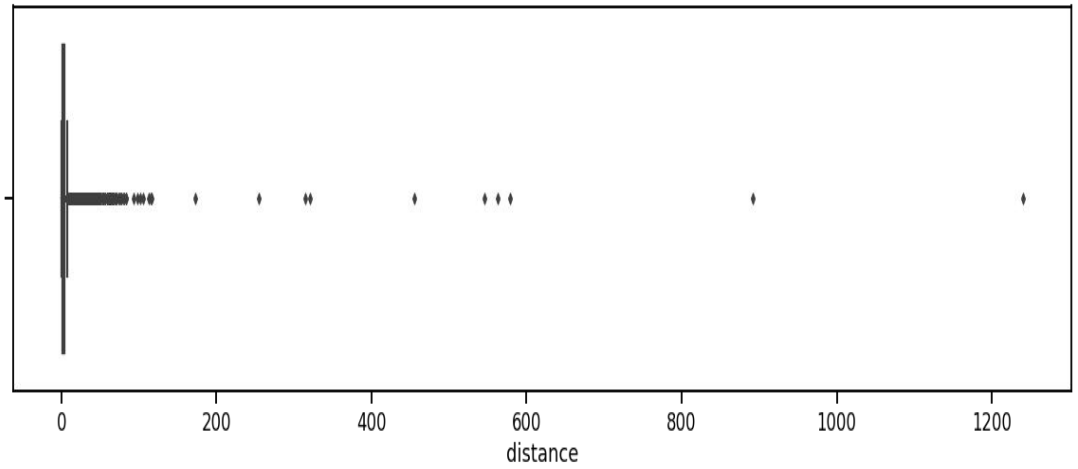


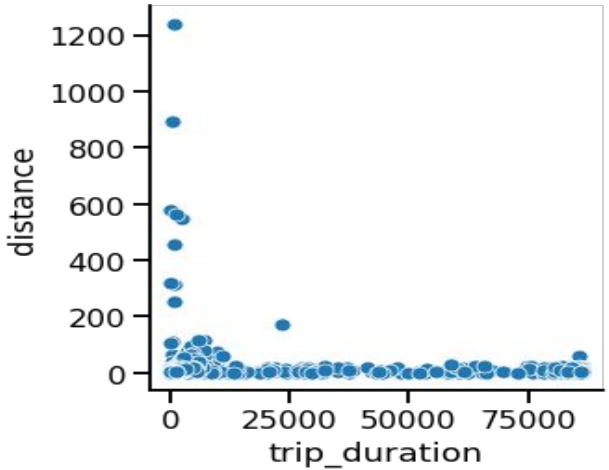


Now, that seems like a fair distribution. We see the highest amount of trips are with 1 passenger.

There is no visible relation between trip duration and passenger count

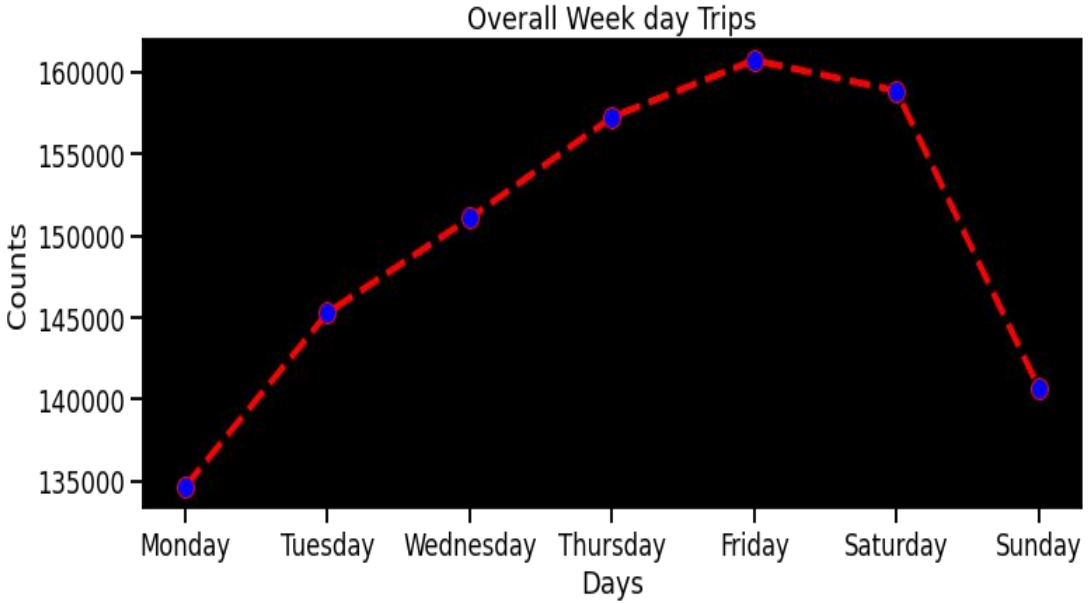
Analysis on : Distance

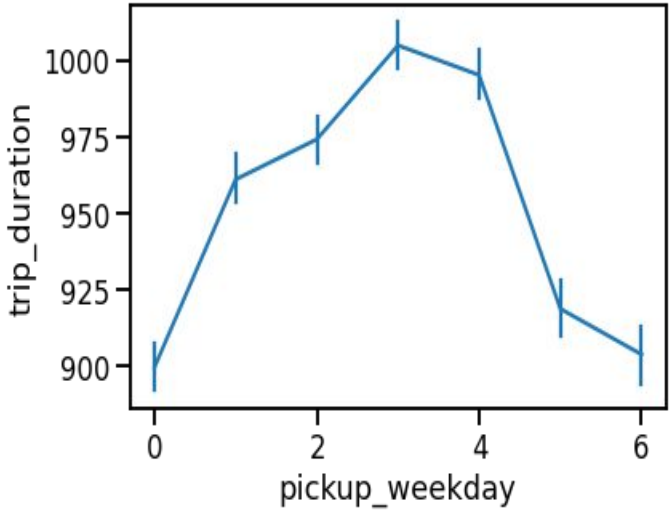




- We can see there are trips which trip duration is as short as 0 seconds and yet covering a large distance. And, trips with 0 km distance has long trip durations.
- The reasons for 0 km distance can be:
  - i. The drop off location couldn't be tracked.
  - ii. The driver deliberately took this ride to complete a target ride number.
  - iii. The passengers canceled the trip.

Analysis on : Trip Duration on a weekday

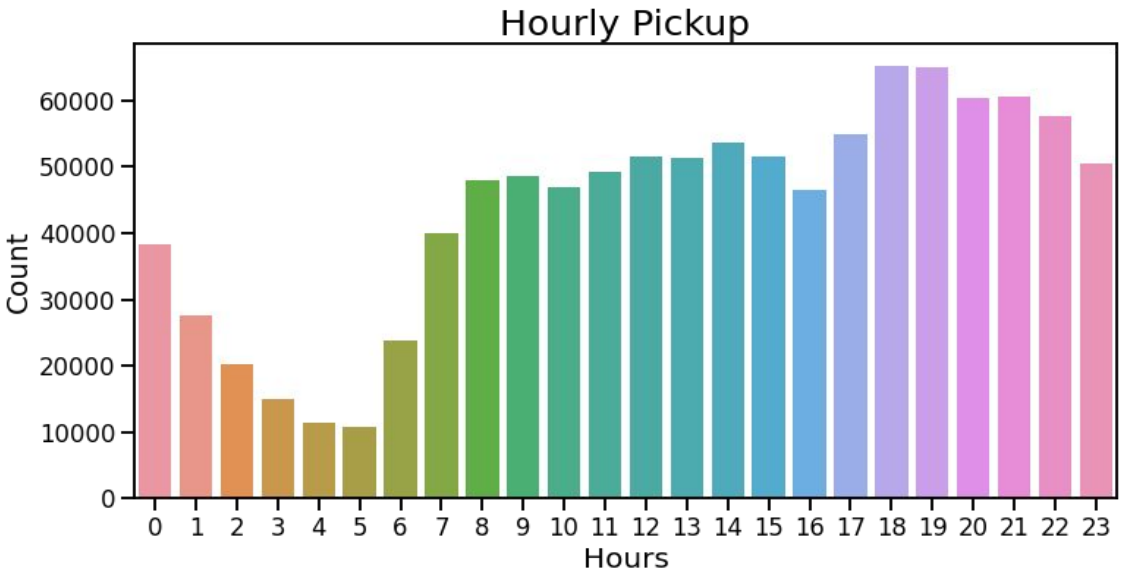


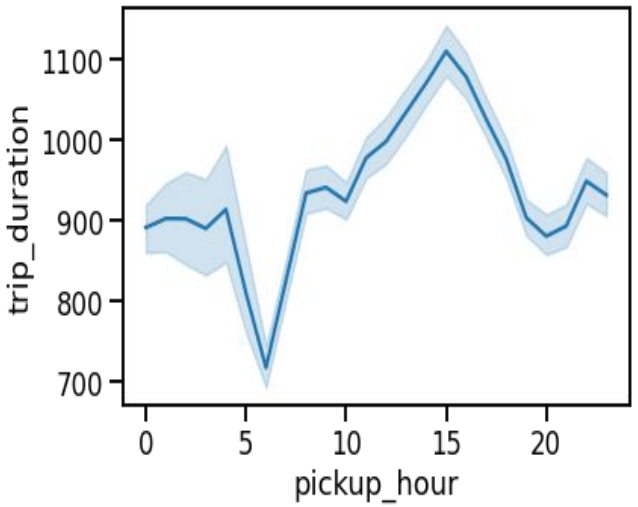


Observations tells us that Fridays and Saturdays are those days in a week when New Yorkers prefer to roam in the city.

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

Analysis on : Trip Duration per hour

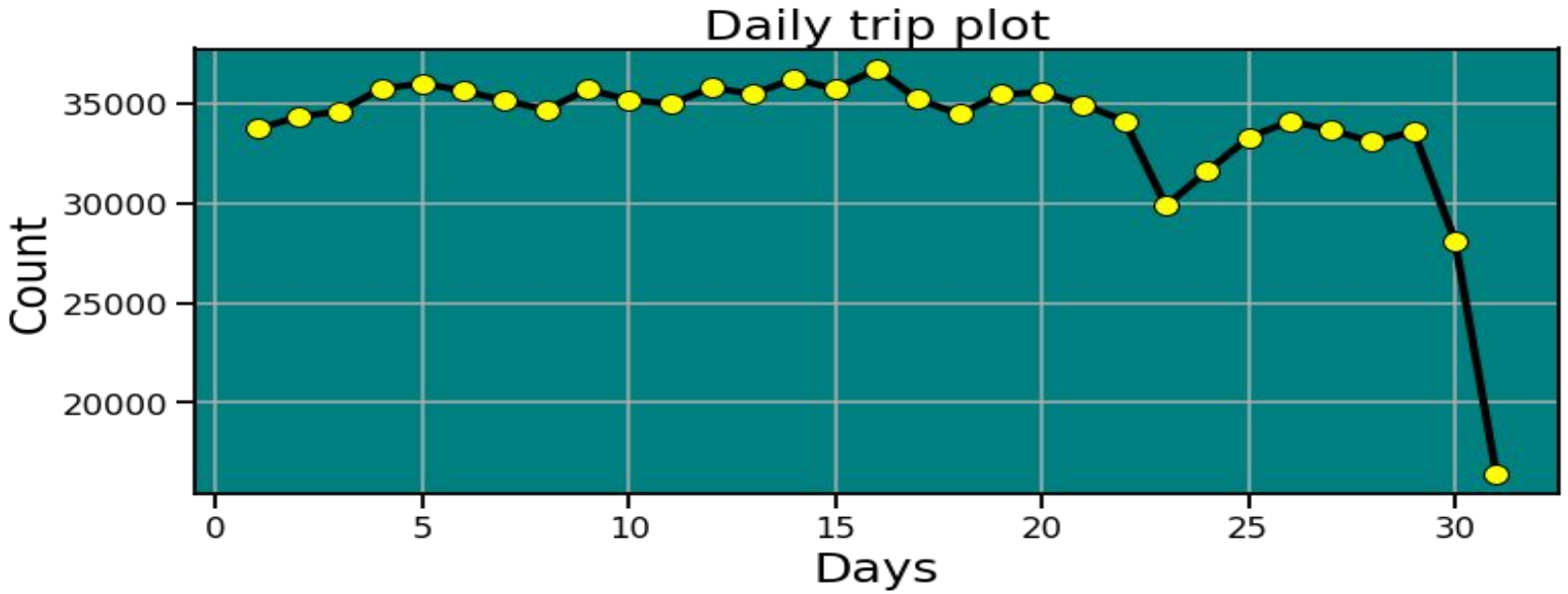




we observe that most pickups and drops occur in the evening (5 pm to 10 pm) probably office leaving time. While the least drops and pickups occur during between 1am to 6am.

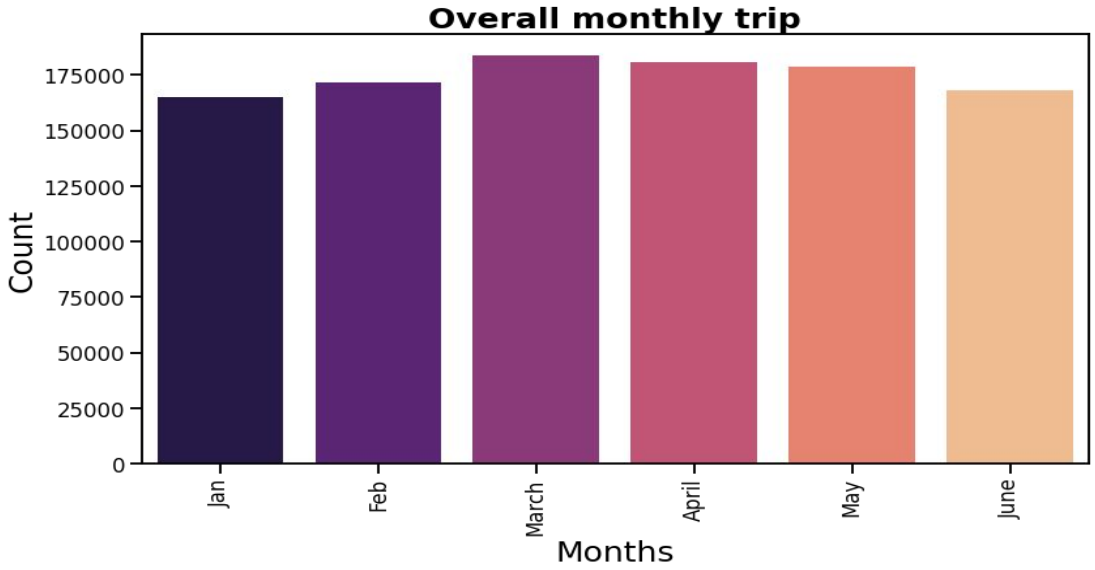
We see the trip duration is the maximum around 3 pm which may be because of traffic on the roads. Trip duration is the lowest around 6 am as streets may not be busy.

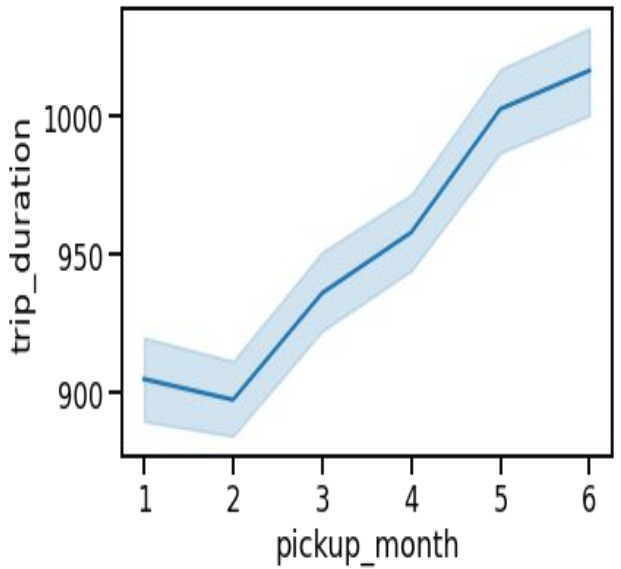
Analysis on : Trip Duration in a month



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

Analysis on : Trip Duration in 6 months

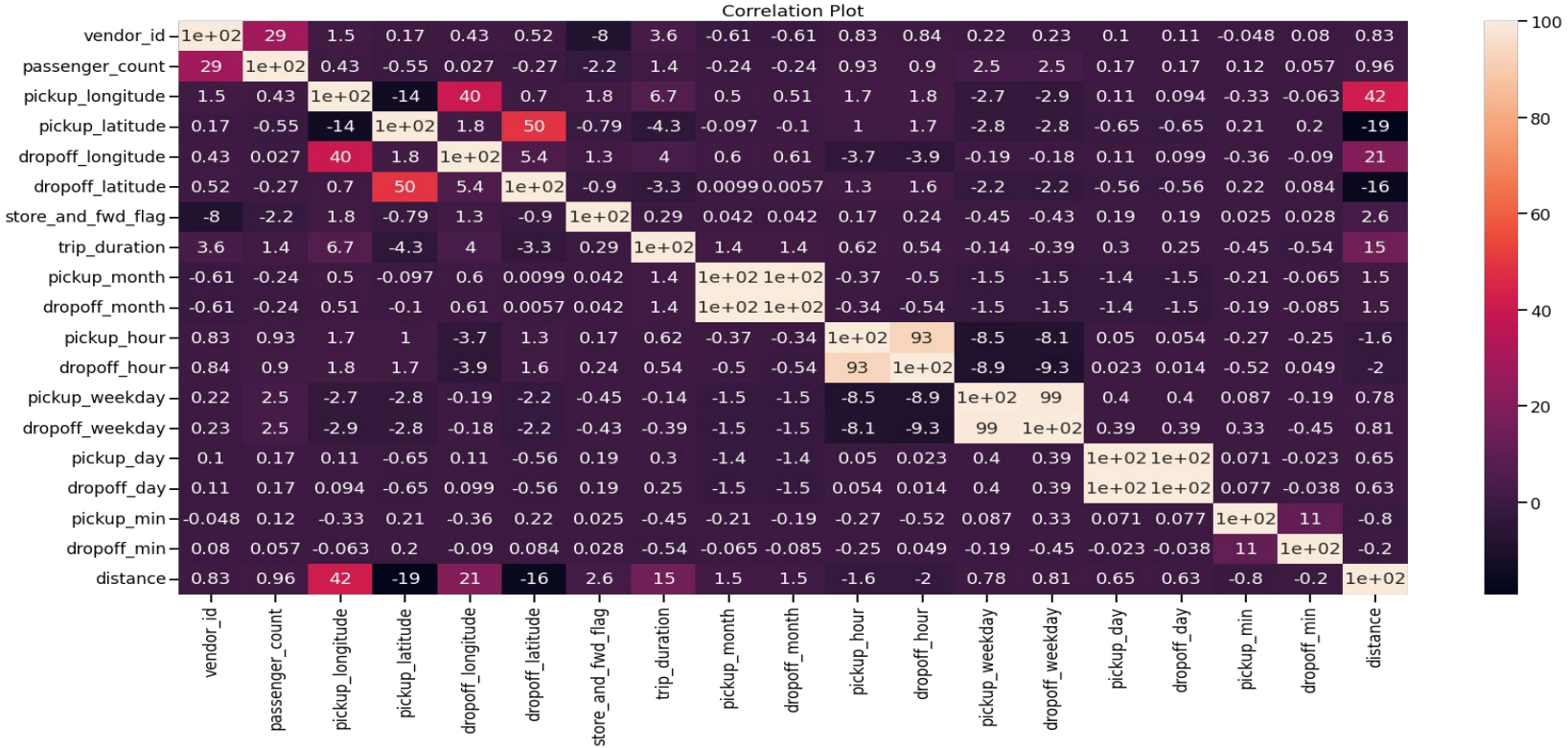




- We've data of 6 months.
- Number of trips in a particular month - March and April marking the highest.
- January being lowest probably due to extreme SnowFall NYC.

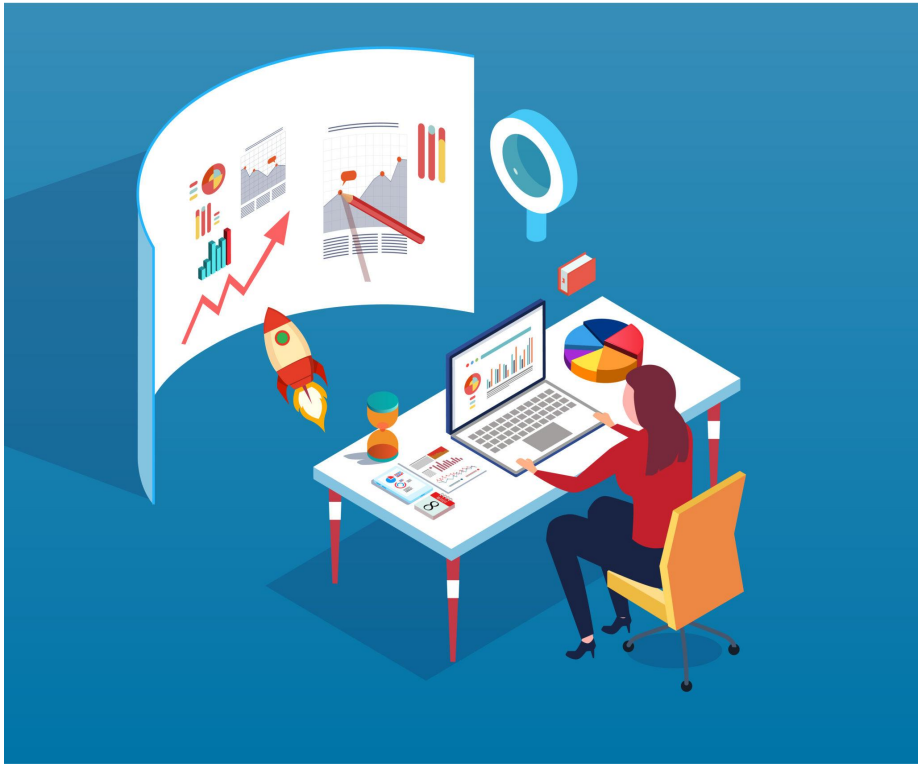
From February, we can see trip duration rising every month.

Analysis on : Correlation Heat map

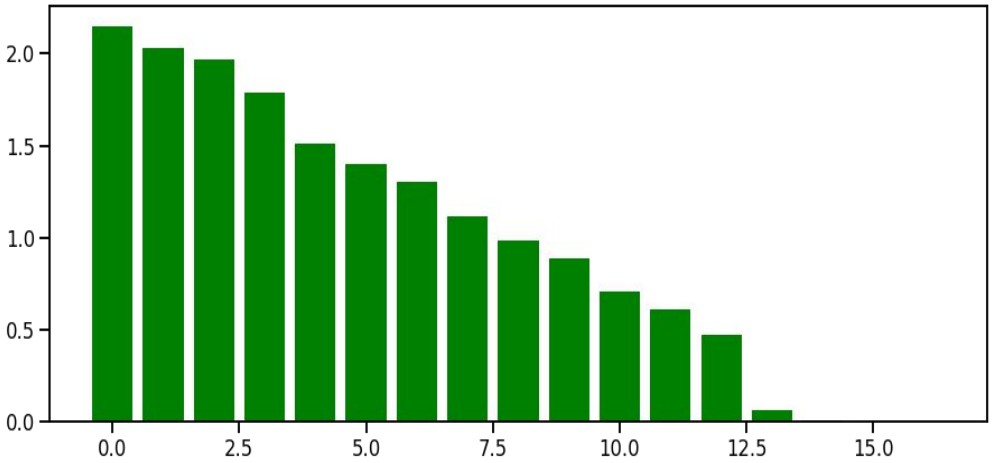


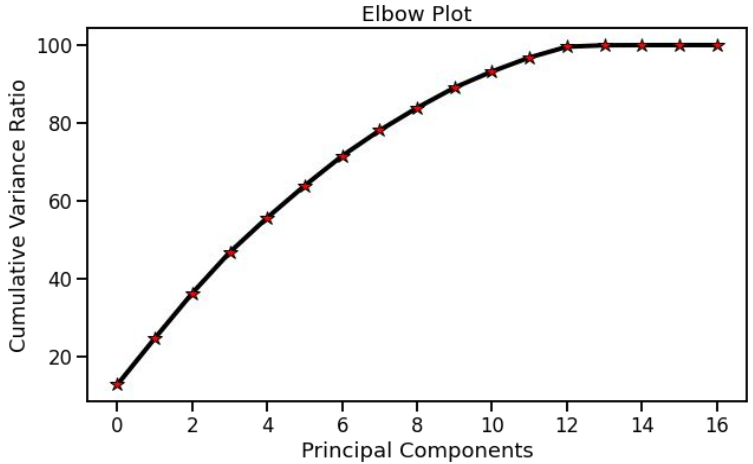
Decomposition of Data: PCA

Principal component analysis (PCA) simpfies Principal component analysis (PCA) simplifies the complexity in high-dimensional data while retaining trends and patterns. It does this by transforming the data into fewer dimensions, which act as summaries of features.



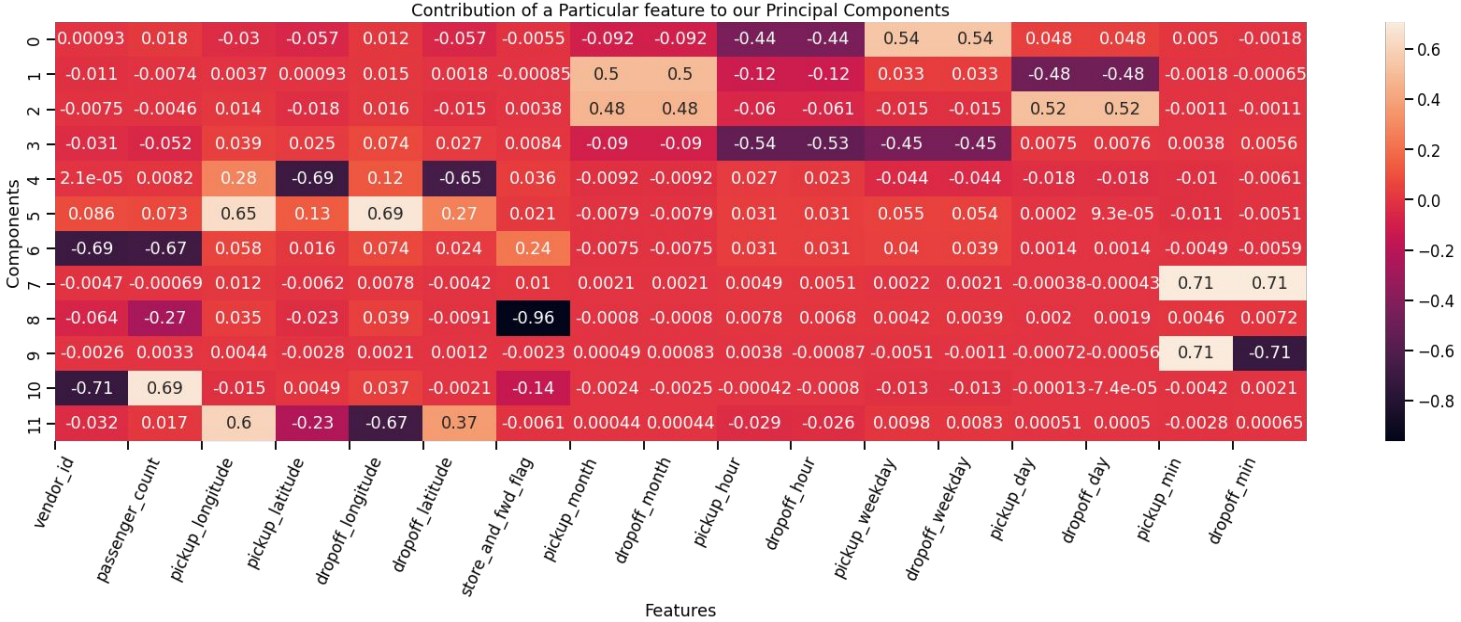
Analysis on : Principal Component Analysis(PCA)





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.

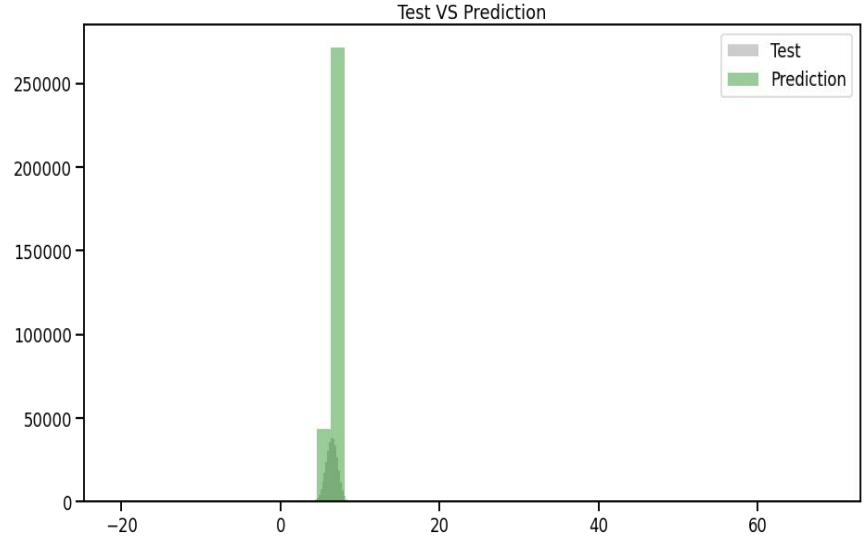
Analysis on: Feature Contribution

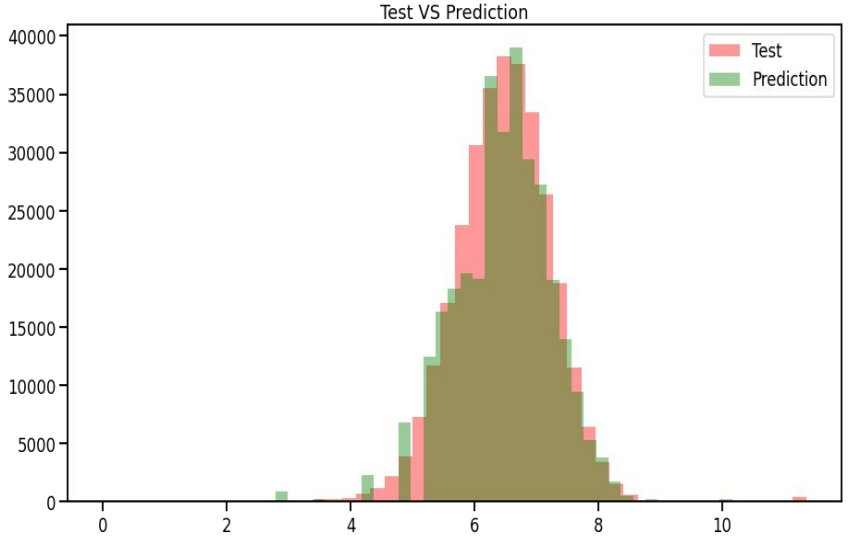


- Above plot gives us detailed ideology of which feature has contributed more or less to our each Principal Component.
- Principal Components are our new features which consists of Information from every other original Feature we have.
- We reduce the Dimensions using PCA by retaining as much as Information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

Machine Learning Model – Regression

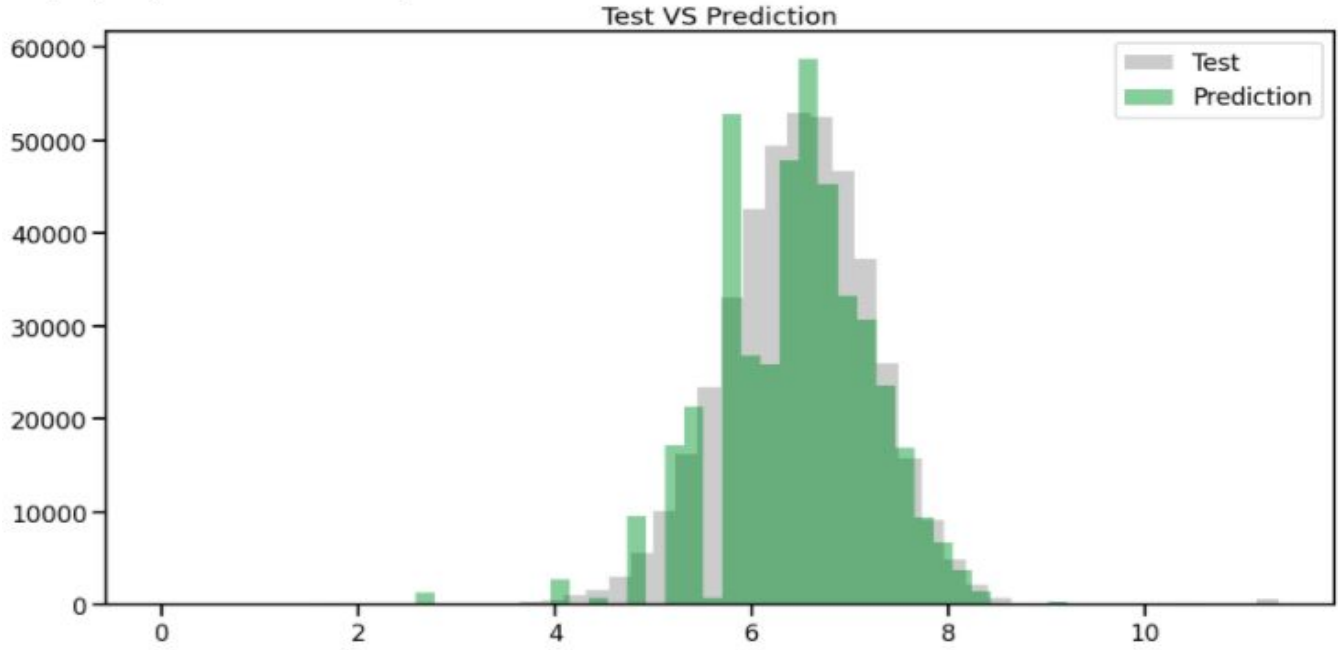
Analysis on: ML Model Prediction with PCA





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.

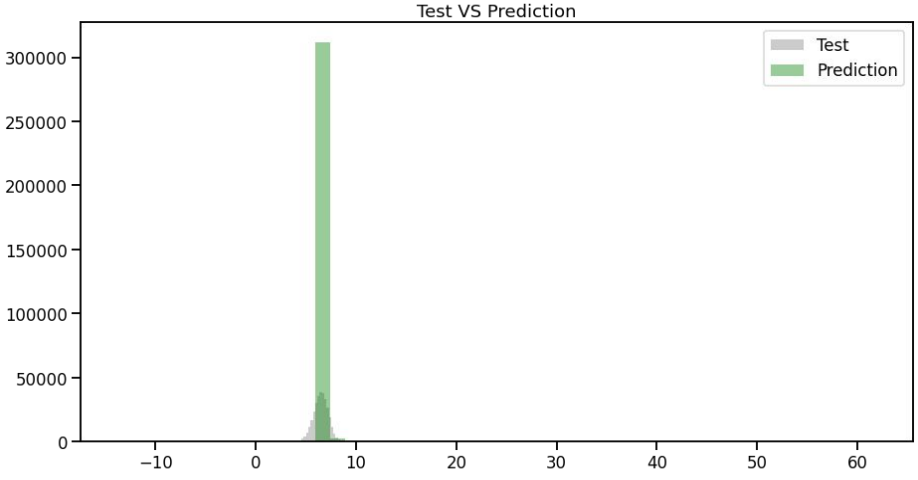
Analysis on : Model Evaluation Result with PCA

- 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

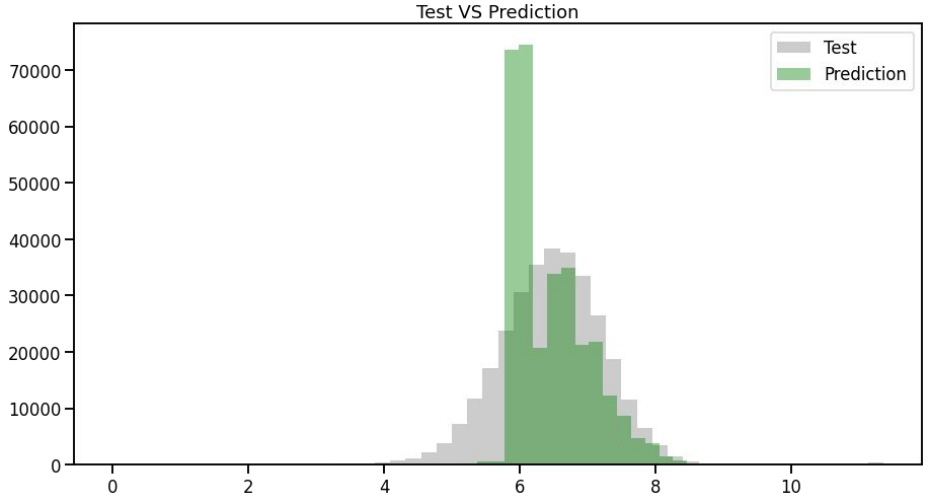
Algorithms	Training Score	Validation Score	Cross Validation Score	R2-Score	RMSE
Linear Regression	0.0661	0.04564	0.06082	-10.221	--
Decision Tree	0.9506	0.9418	0.9406	0.9389	0.0306
Random Forest	0.9522	0.9454	0.9451	0.9424	0.0300

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

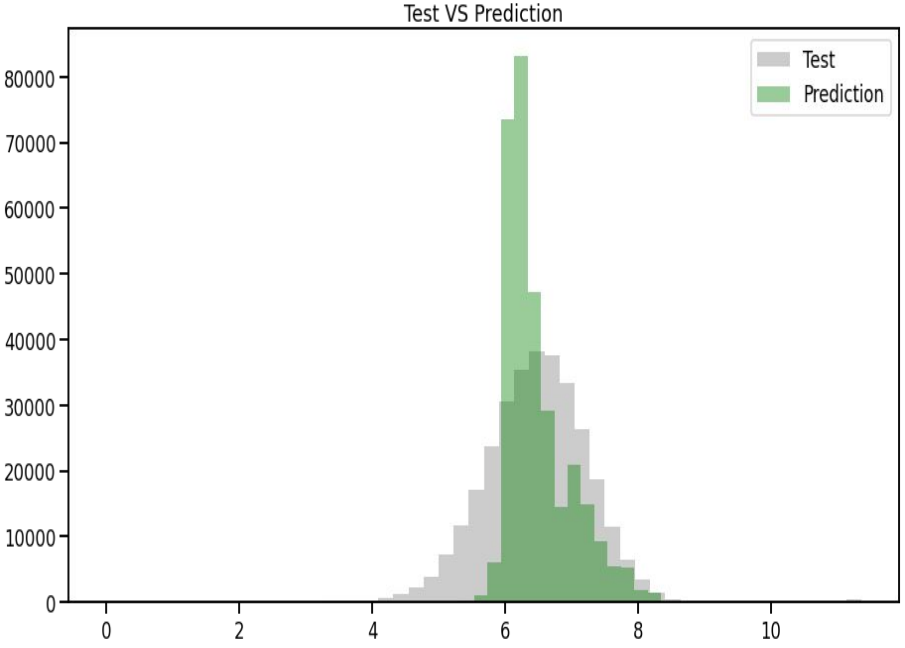
Analysis on: ML Model Prediction without PCA



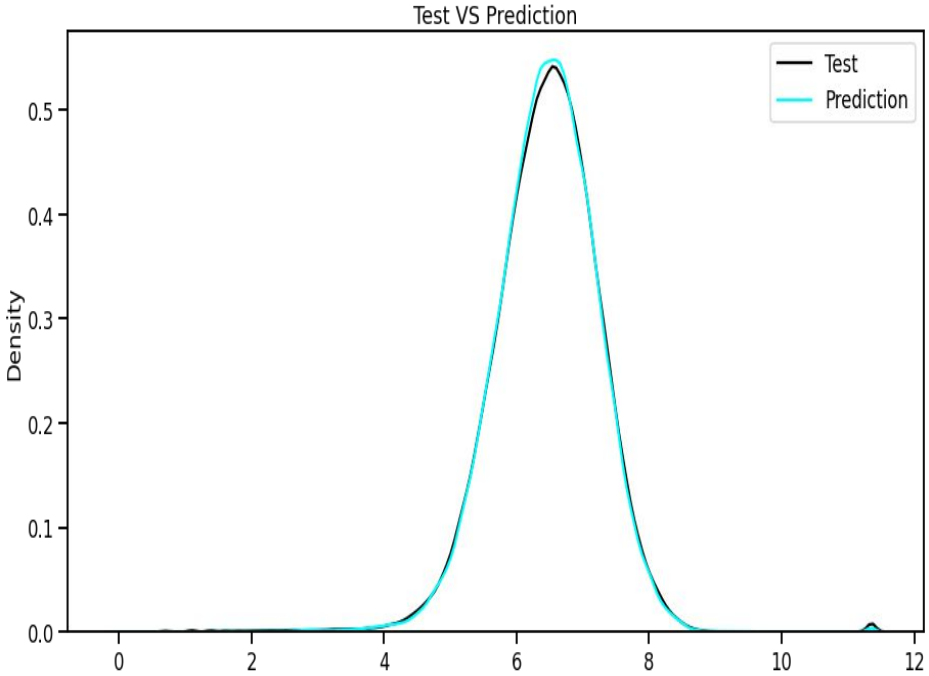
Linear Regression



Decision Tree



Random Forest



Decision Tree with GridsearchCV

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	0.192	0.1766	0.1881	0.1766	--
Decision Tree	0.4672	0.4576	0.4469	0.4576	0.0884
Decision Tree with Gridsearch cv	0.9091	0.5520	--	0.5130	0.08314
Random Forest	0.4805	0.4727	0.4702	0.4727	0.08767

- 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(vendor 2) 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.

- Apply Standard Scaling on the Dataset to Normalize the values.
- Further, Apply PCA to reduce dimensions, as you'll extract features from our primary Date Time Feature. Those additional features might lead our model to suffer from "Curse of dimensionality" and could drastically affect performance.
- Pass the PCA Transformed data in our ML Regression Algorithms and Evaluate results.
- We can perform hyper tuning on our Algorithm to get the most out of it but Hyper Tuning consume lot of time and resources of the system depending upon the how big the Data we have and what algorithm we're using. It will go through number of Iterations and try to come up with the best possible value for us.
- we also applied some other type of algorithms like xgboost & elastic net but result are not acceptable so that's why i think that there is no need to show these type of algorithms .

THANK YOU

Q & A