# Capstone Project – 2
# NYC Taxi Trip Time Prediction

**Project By-**
Samiksha Bandbuche
Prince Jain

# Point to be discussed:

- Introduction
- Data Summary
- What is Our Goal?
- Exploratory Data Analysis(EDA)
- Correlation Heat Map
- Machine Learning Model-Regression
- ML Model Prediction
- Model Evaluation Result
- 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.

# Data Summary:

The dataset is based on the 2016 NYC Yellow Cab trip record data made available in Big Query on Google Cloud Platform. The data was originally published by the NYC Taxi and Limousine Commission (TLC). The data was sampled and cleaned for the purposes of this project. Based on individual trip attributes, you should predict the duration of each trip in the test set.

**Data Set Name --** NYC Taxi Data.csv

**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’.

# What is our goal?


- Predict the demand for taxi services in geographic areas in the greater New-York city metropolitan area.

- Provide data that can allow taxi and ride-sharing companies to optimally allocate driver to specific location.

- Reduce unnecessary driving and shorten waiting time for customers.

# METHODOLOGY:

## Approch:

| Step 1: | Step 2: | Step 3: |
| --- | --- | --- |
| Data Preparation and Exploratory Data Analysis | Building Predictive Model using Multiple Techniques/ Algorithms | Optimal Model Identified through testing and evaluation |

# Machine Learning Algorithm:

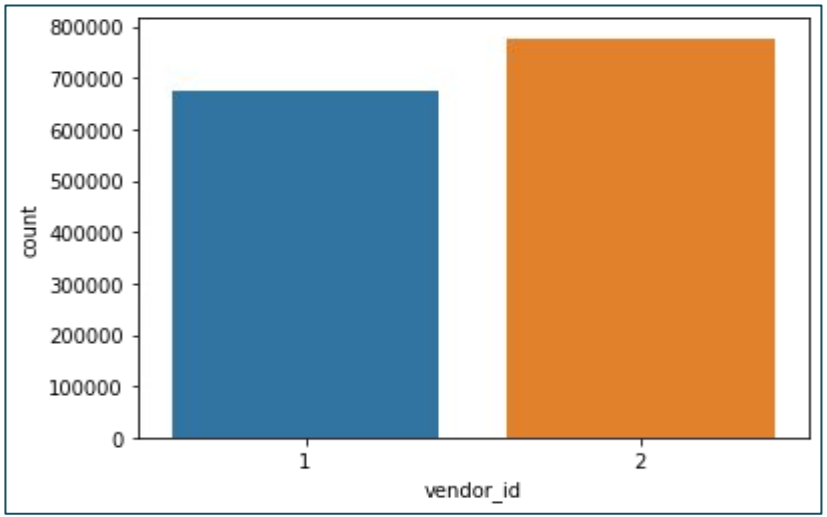
- Decomposition: PCA
- Linear Regression
- Decision Tree
- Random Forest

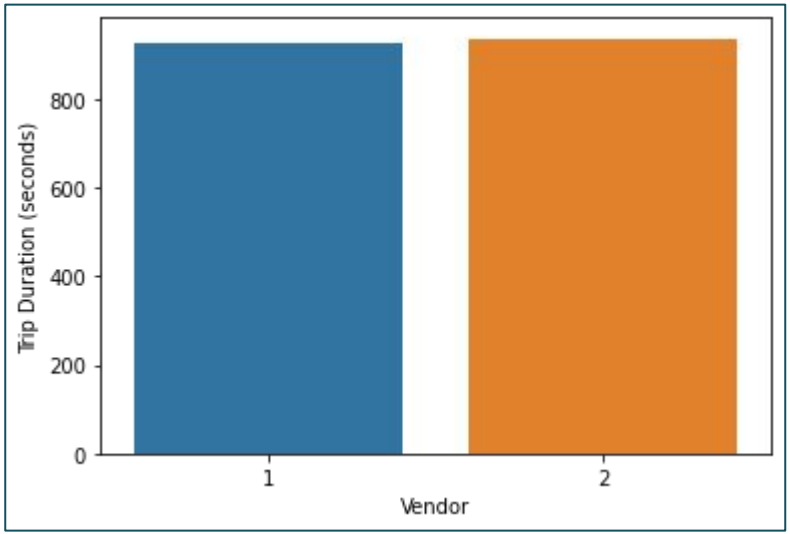
# Tools Used:

- Google Colab Research

# EXPLORATORY DATA ANALYSIS

# Analysis on : Vendor Id

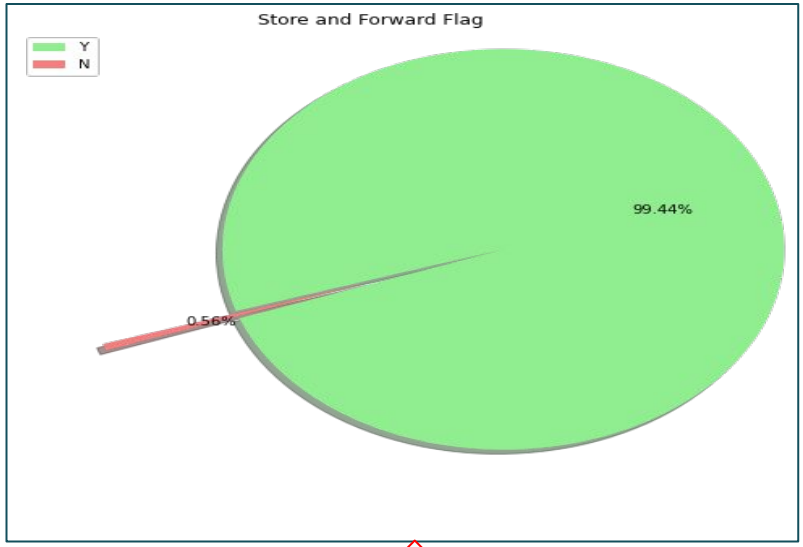



Though both the vendors seems to have almost equal market share.**Vendor 2** Service provider is the **most opted one by New Yorkers**.

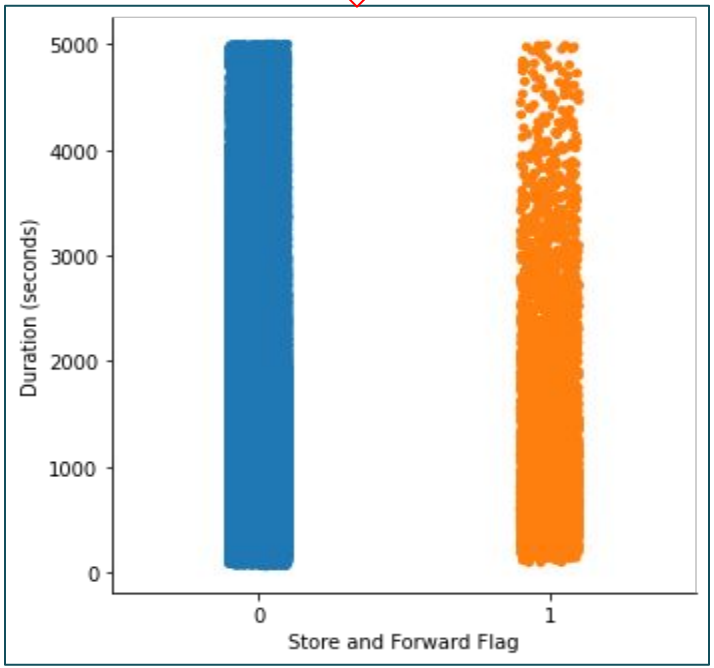
**Trip Duration Per Vendor:** **Average trip duration** for **vendor 2** is **higher** than **vendor 1** by approx **200 seconds** i.e. atleast **3 minutes per trip**.

# Analysis on : Store and Forward Flag

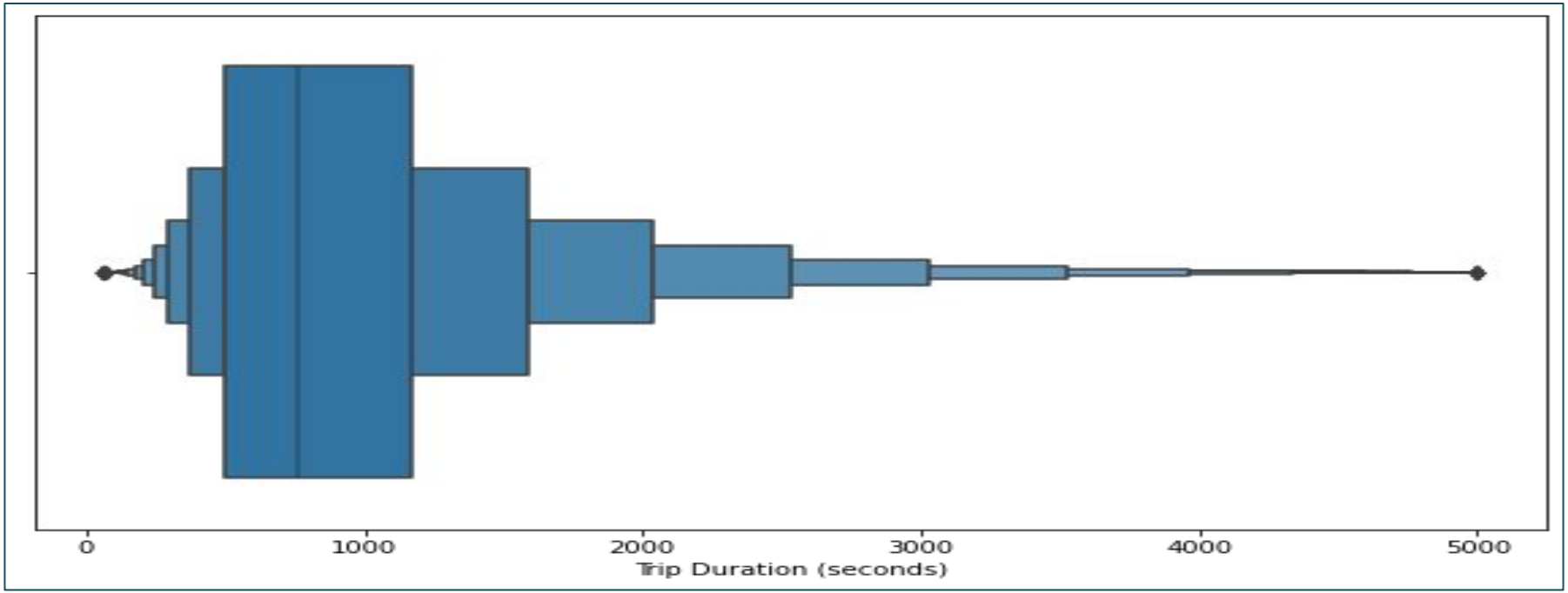



Visualization tells us that there were very **few trips** of which the records were stored in memory due to no connection to the server.

Trip Duration v/s Flag:
There is **not much difference between N and Y**.

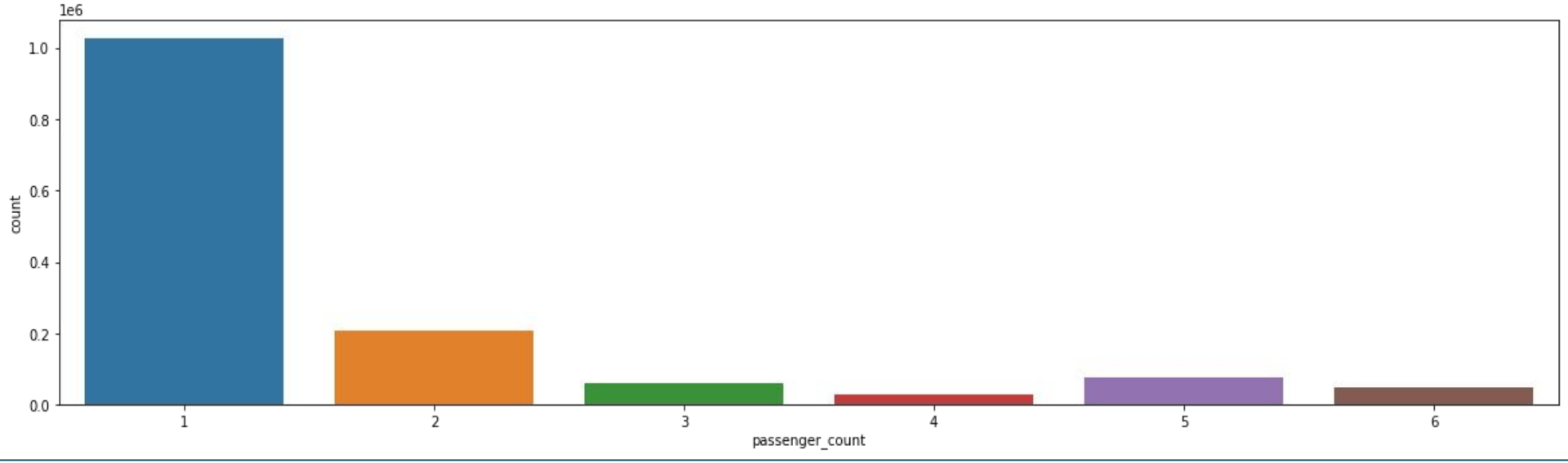


# Analysis on : Target Variable – Trip Duration



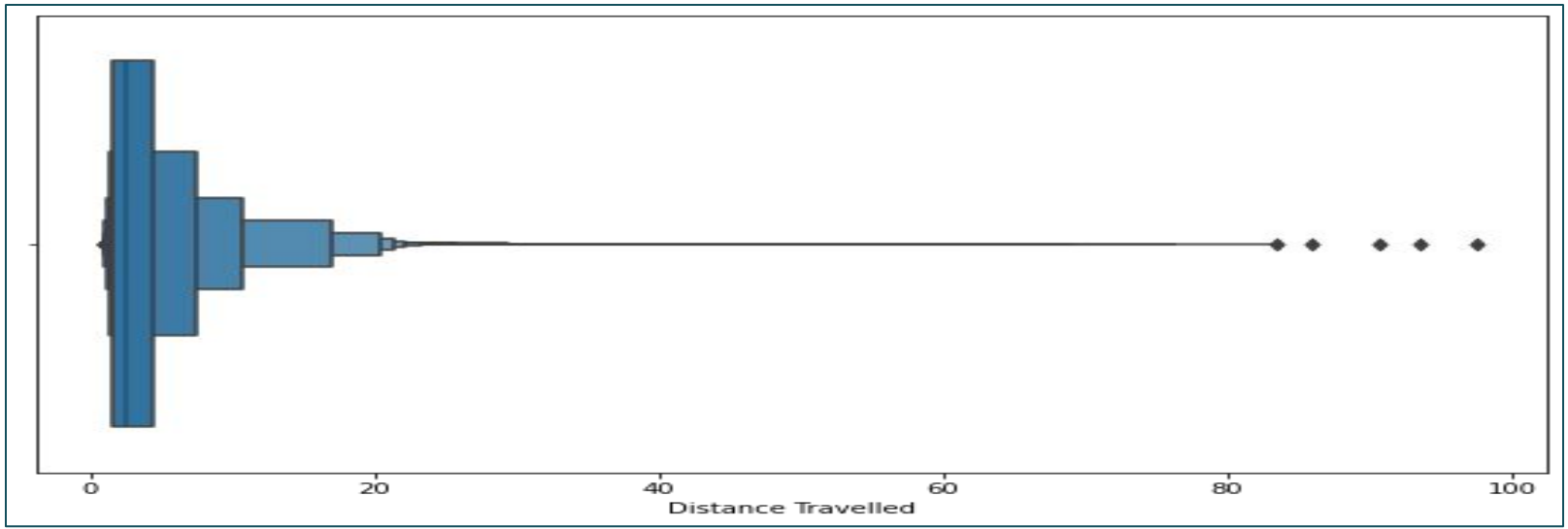

Many of the trips durations took between **10-20 mins** . As observed most of the trips took **0 - 30 mins** to complete

# Analysis on : Passenger Count




As per above observations, it is to be noted that **highest amount** of trips were taken by a **single passenger** and large group of people travelling together is rare compared to single passenger.

# Analysis on : Distance

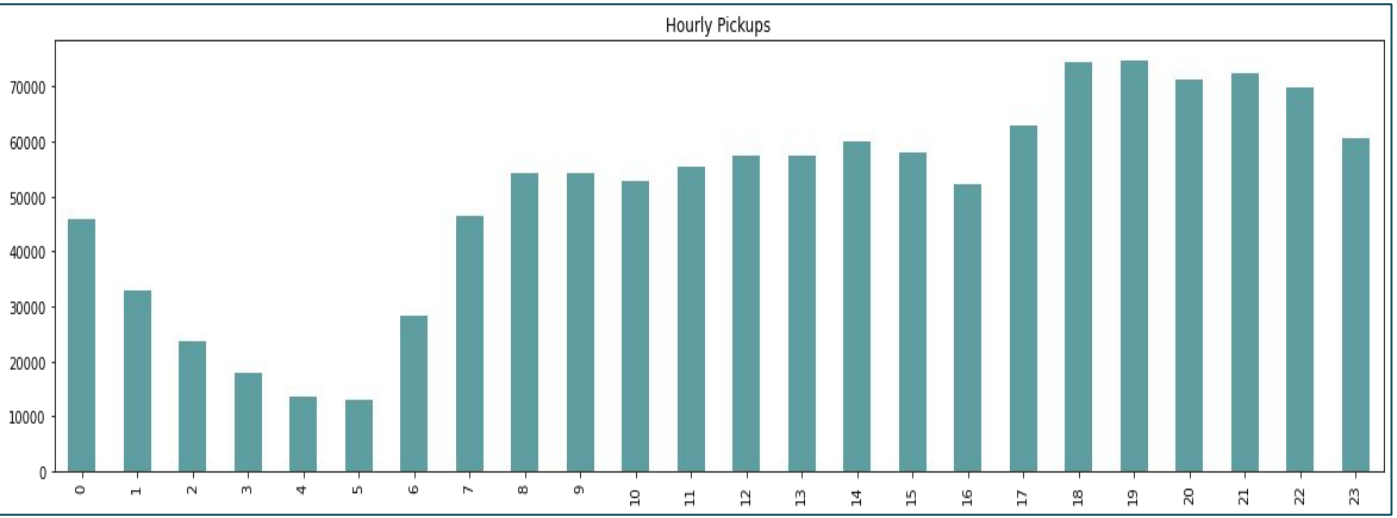



There are some trips with over **100 km** distance and some trips with **0 km** distance.

The possible reasons for zero km trips can be:

- The drop off location couldn't be tracked.
- The passengers or driver cancelled the trip due to some or issue technical issue in software, etc.

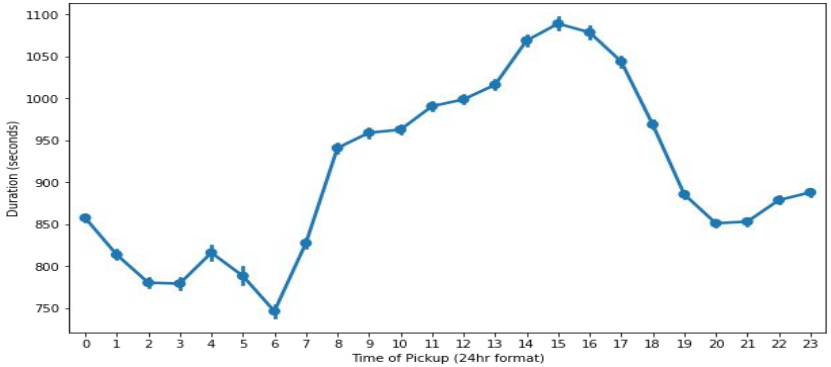
# Analysis on : Trip Duration per Hour



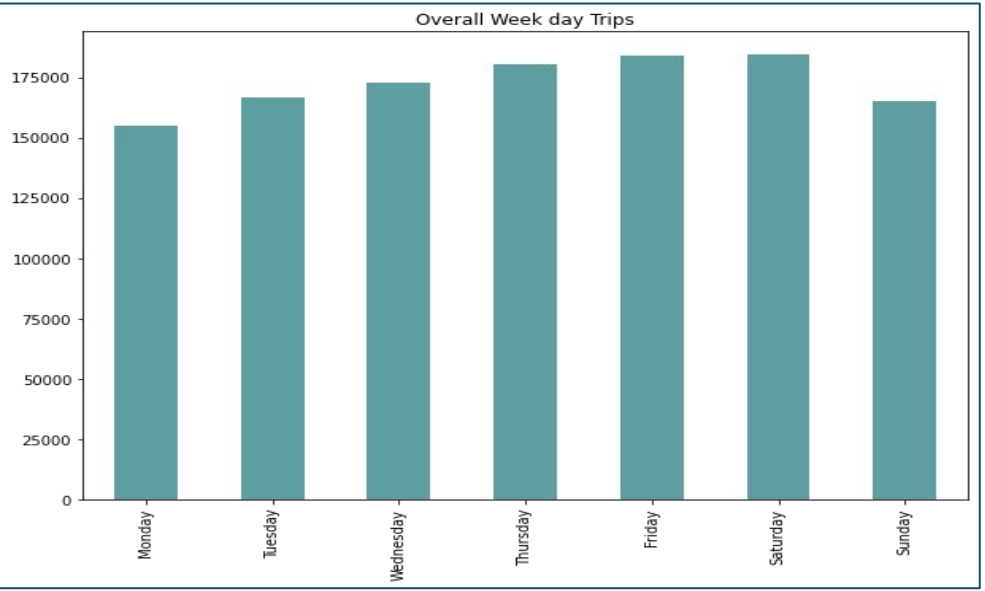

**In which hour we get to see maximum pickups ?**

Rush hours (5 pm to 10 pm), probably office leaving time.

- 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.

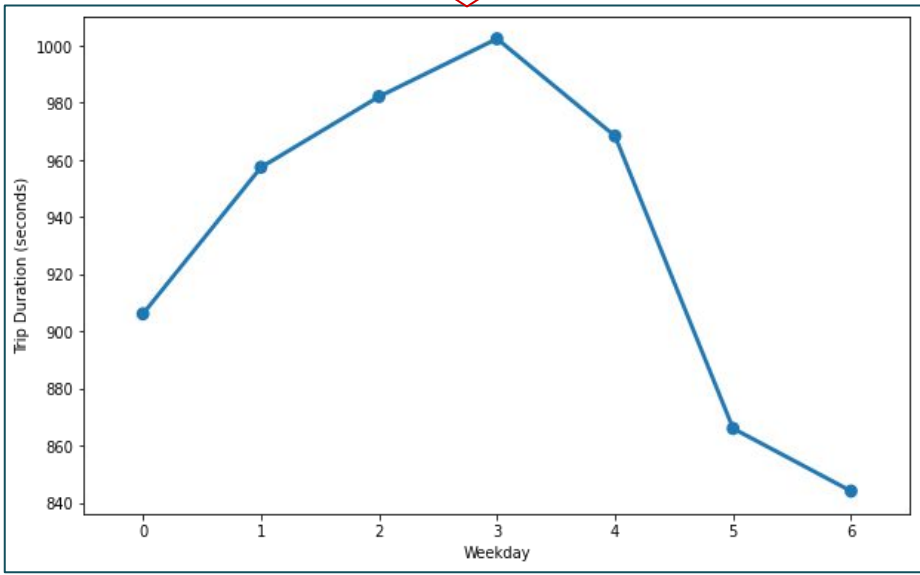


# Analysis on : Trip Duration on a weekday




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.

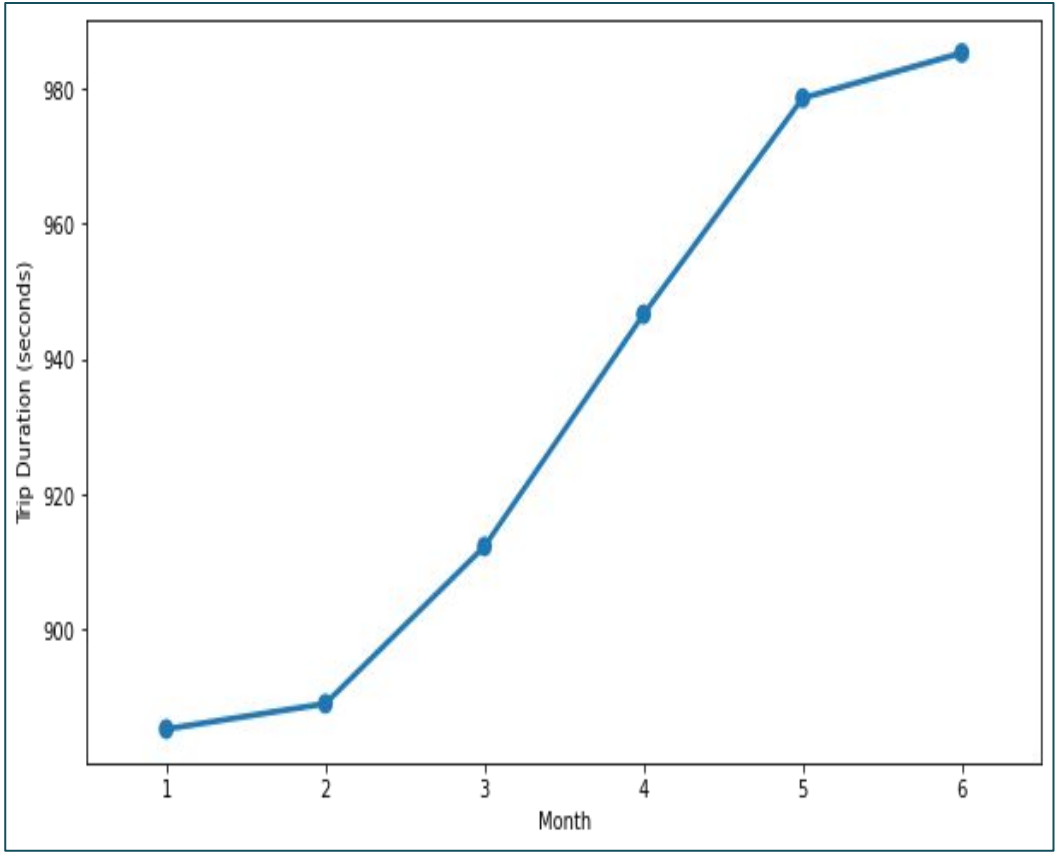
Also, it is observed that trip duration on thursday is longest among all days.



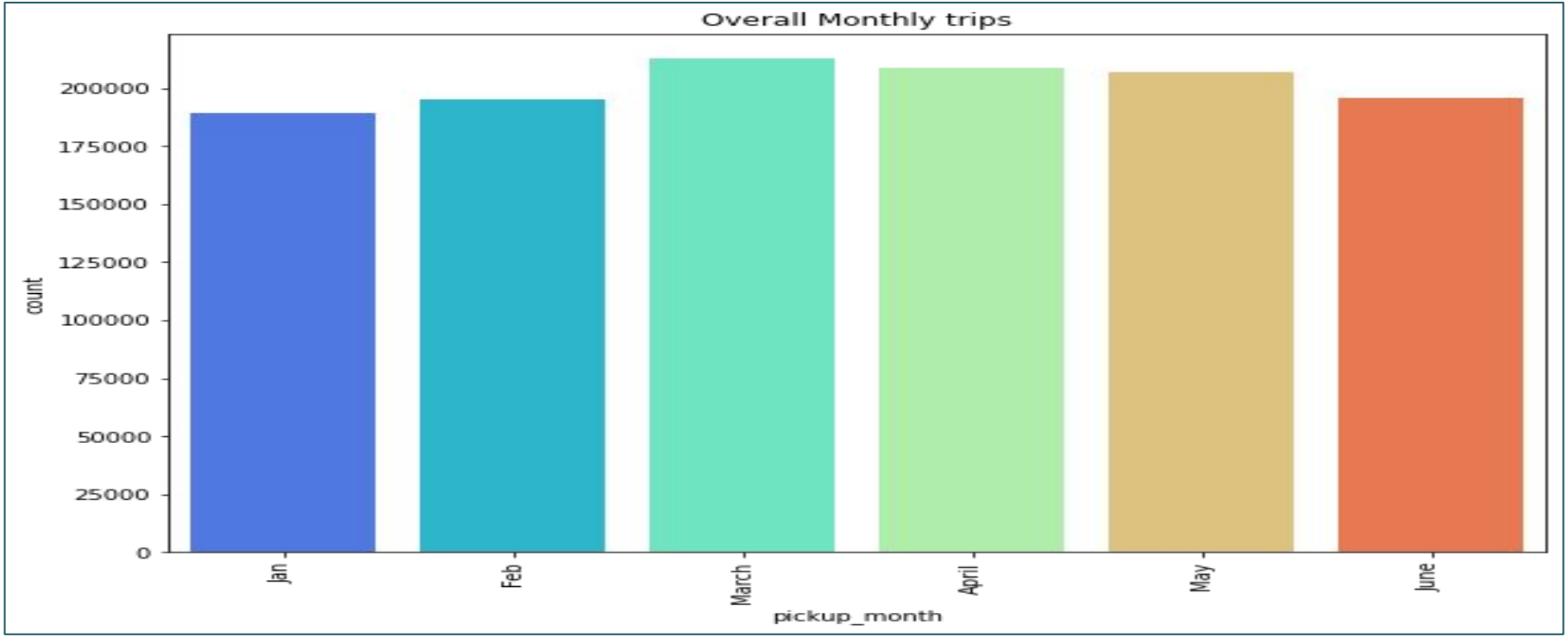

Observations tells us that **Fridays and Saturdays** are those days in a week when New Yorkers prefer to rome in the city. GREAT !!

# Analysis on : Trip Duration in a month

- We can see an **increasing trend** in the average trip duration along with each subsequent month.
- The duration difference between each month is not much. It has **increased** gradually over a **period of 6 months**.
- It is **lowest** during **february** when winters starts declining.
- There might be some seasonal parameters like wind/rain which can be a factor of this gradual increase in trip duration over a period. Like May is generally the considered as the wettest month in NYC and which is inline with our visualization. As it generally takes longer on the roads due to traffic jams during rainy season. So naturally the trip duration would increase towards April May and June.

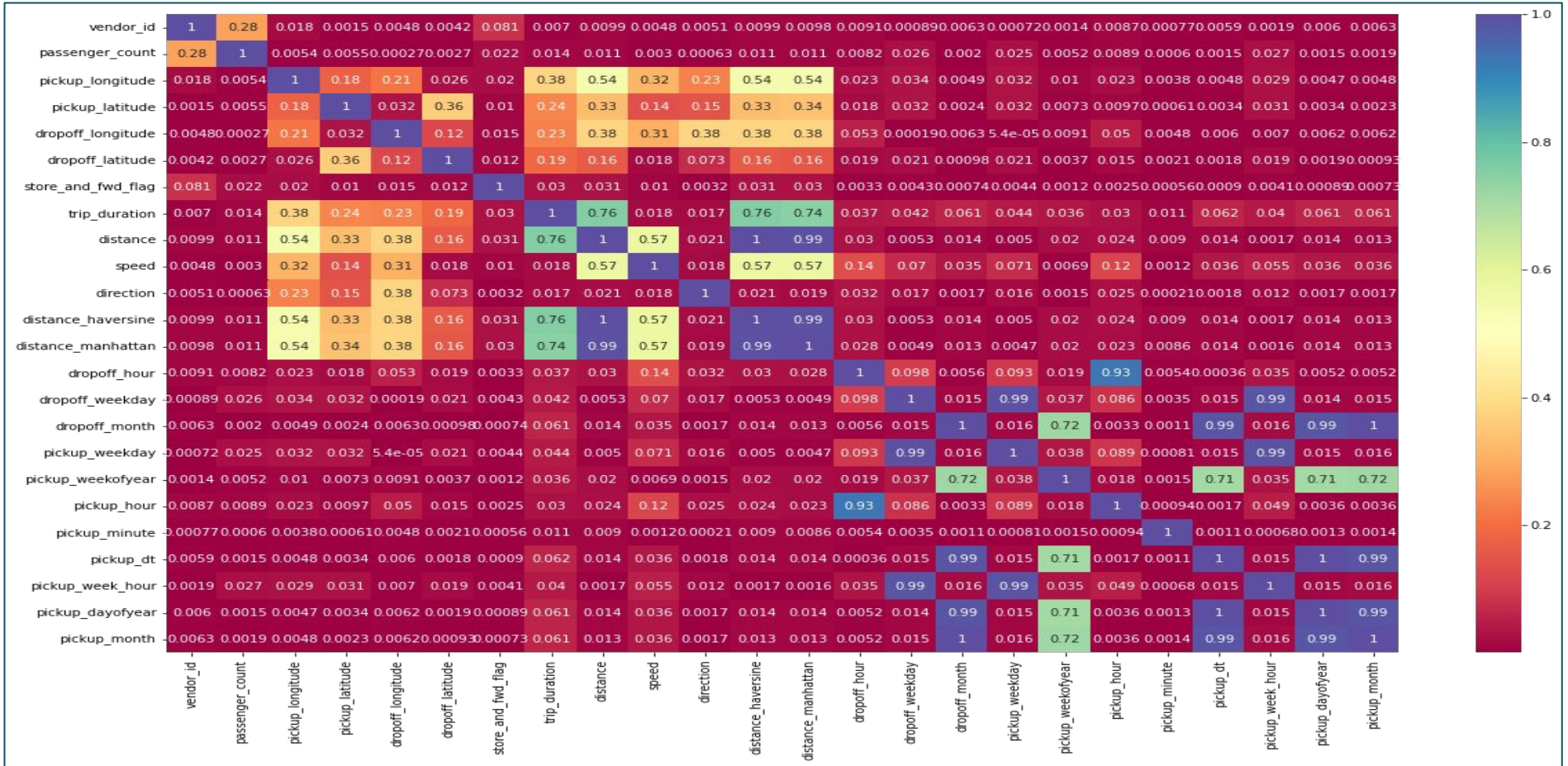


# Analysis on : Trips in 6 months



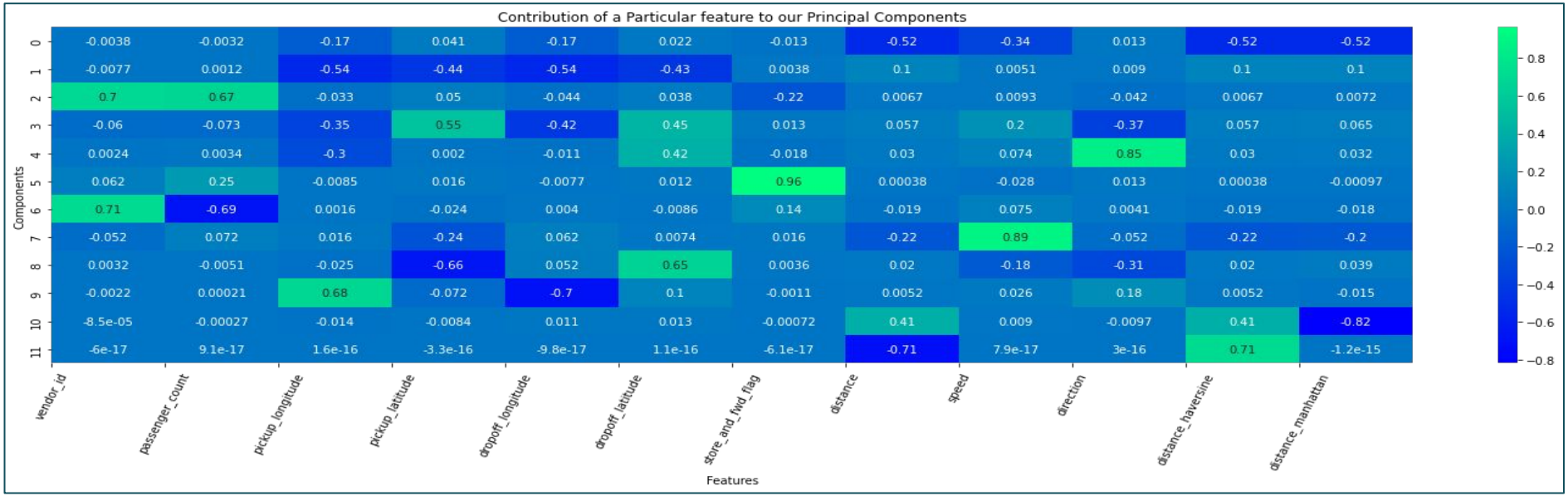

- Number of trips in a particular month **March,April and May** marking the **highest**.
- **January** being **lowest** probably due to extreme SnowFall NYC.

# Analysis on : Correlation Heat map



# Machine Learning Model – Regression

## Principal Component Analysis (PCA):




- 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.

# ML Model Prediction:

## Why Linear Regression , Decision Tree and Random Forest ?

**Linear regression:**

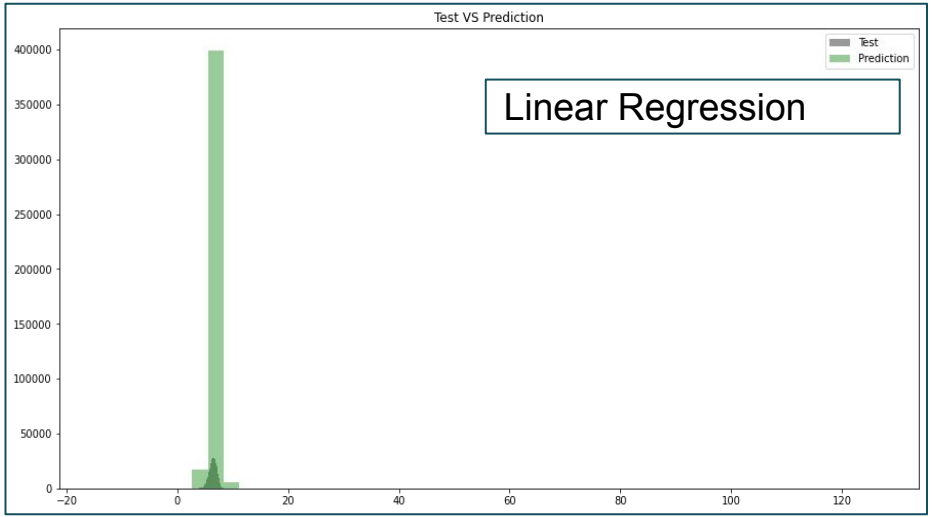
- Simple to explain.
- Model training and prediction are fast.
- No tuning is required except regularization.

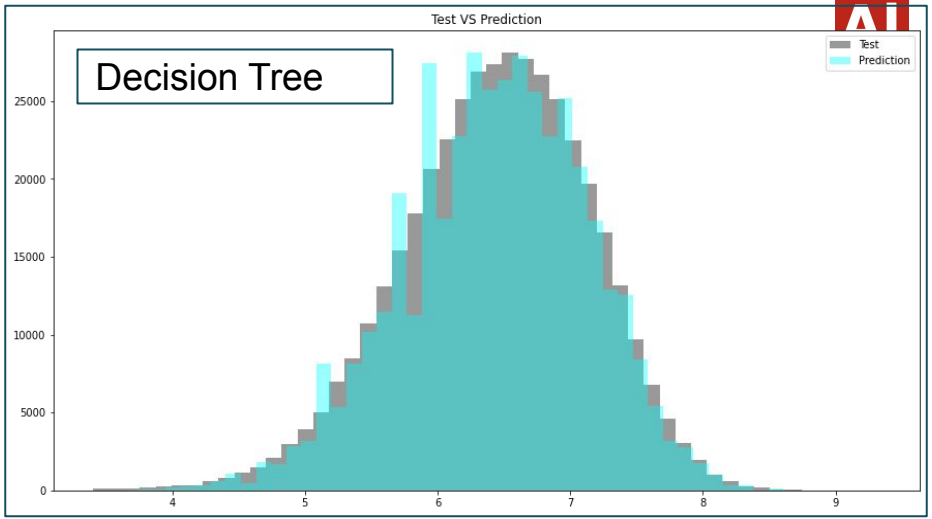
**Decision Tree:**

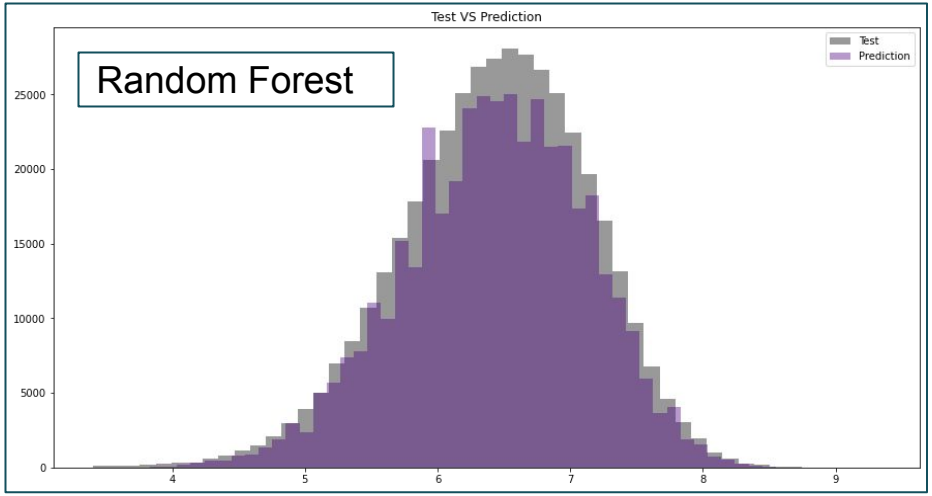
- Decision trees are very intuitive and easy to explain.
- They follow the same pattern of thinking that humans use when making decisions.
- Decision trees are a common-sense technique to find the best solutions to problems with uncertainty.

**Random Forest:**

- It is one of the most accurate learning algorithms available.
- Random Forest consists of multiple Decision Trees - Results from multiple trees are then merged to give best possible final outcome.
- Random forests overcome several problems with decision trees like Reduction in overfitting.




We can observe from above visualization, that our **Decision Tree model and Random Forest model are good performers**. As, **Random Forest is providing us reduced RMSLE**, we can say that it's a model to Opt for.

# Model Evaluation Result:

| Algorithms | Training Score | Validation Score | Cross Validation Score | R2-Score | RMSLE |
|---|---|---|---|---|---|
| Linear Regression | 0.7523 | 0.7523 | 0.7522 | 0.5675 | — |
| Decision Tree | 0.9483 | 0.9441 | 0.9447 | 0.9409 | 0.0248 |
| Random Forest | 0.9577 | 0.9546 | 0.9566 | 0.9515 | 0.0224 |

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

# Conclusion:

- It is to be noted that **highest amount** of trips were taken by a **single passenger** and **large group of people** travelling together is rare compared to single passenger.
- Observed that **Vendor 2** taxi service provider is **most Frequently used** by **New Yorkers**.
- As observed most of the trips took **0 - 30 mins to complete** (1800 seconds).
- There were very **few trips** of which the records were **not stored** in memory due to no connection to the server.
- The **rush hours** are **5 pm to 10 pm**, probably because they are office leaving time.

# Conclusion:(Contd…..)

- Number of trips in a particular month **March,April and May** marking the **highest**.
- **Average trip duration** is **lowest at 6 AM** when there is minimal traffic on the roads and 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.**
- 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.
- **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.

# So,which algorithm is better - Decision Tree or Random Forest ?

- One problem that might occur with **Decision Tree is that it can overfit**.
- Difference is - **A random forest is a collection of decision trees**.
- A **decision tree model** considers **all the features which makes it memorize everything**, it gets **overfitted on training data** which **couldn't predict well on unseen data**.
- A **random forest chooses few number of rows at random** and **interprets results** from all the Tress and **combines** it to get more **accurate and stable final result**.
- To predict the trip duration for a particular taxi, we can conclude that **Random Forest is the best models as compare to the other models**.

# Thank You!!