New York City Taxi Trip Time Prediction

Samiksha Bandbuche
Prince Jain
Data science trainees,
Almabetter, Bangalore

Abstract:

New York City taxi rides form the core of the traffic in the city of New York. The many rides taken every day by New Yorkers in the busy city can give us a great idea of traffic times, road blockages, and so on. Predicting the duration of a taxi trip is very important since a user would always like to know precisely how much time it would require him to travel from one place to another.

Given the rising popularity of app-based taxi usage through common vendors like Ola and Uber, competitive pricing has to be offered to ensure users choose them. Prediction of duration and price of trips can help users to plan their trips properly, thus keeping potential margins for traffic con- gestions.

1. 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

2. Dataset Description:

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.

NYC Taxi Data.csv - the training set (contains 1458644 trip records)

Data fields

- 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
- trip_duration - duration of the trip in seconds

3. Research Methods:

In this paper, we propose a classification framework based on ensemble learning to New York City taxi duration prediction. The framework involved five steps, including data collection and understanding, data preprocessing, data cleaning, Exploratory Data Analysis (EDA) and various regression models.

3.1. Data collection and understanding: The data primarily contained the following attributes of information: id, vender id, date time, pickup datetime, passenger count, drop-off, datetime, pickup longitude, pickup latitude, drop-off longitude, drop-off latitude, store and fwd flag, trip duration and the action results. There were many fields under each type of information to enrich the data. There are 1458644 rows × 11 columns.

3.2. Data Preprocessing: In the dataset, Mean for Trip Duration is: 959.46.
Standard Deviation for Trip Duration is: 5237.07
While applying certain operations we drive that there are 4744 values that are false that means they are outliers. And 1451732 values are close to the mean value. They provide location in the form of longitude and latitude which is difficult to understand so combining them and converting them into location which provide pickup information which is useful for the determining distance and speed parameter.
There is certain distance formula are come across:

Euclidean Distance

Euclidean distance is calculated as the hypotenuse of a right triangle, just like in the Pythagorean theorem. This is simply a direct path from point A to point B. In the image below, this would be the black line. The Euclidean distance is roughly 1,417 miles. Although not perfect, this may be a good estimate for flight distance.

The Haversine (or great circle)

The distance is the angular distance between two points on the surface of a sphere. The first coordinate of each point is assumed to be the latitude, the second is the longitude, given in radians.

BEARING AND DISTANCES.

Bearing can be defined as the clockwise angular movement between two distant places

Manhattan Distance

The distance between two points measured along axes at right angles. In a plane with p1 at (x1, y1) and p2 at (x2, y2), it is |x1 - x2| + |y1 - y2|. This method has its problems but could be a good estimate in grid-based cities.

At this stage, the dataset was processed through data cleaning, feature engineering, and data normalization.

3.3. Data cleaning: Data cleaning aims to reduce the dimensions of the NYC dataset by detecting and deleting irrelevant or redundant attributes and case records. First, attribute fields that contained descriptive text or too many missing values were removed. We perform an operation in which passengers' counts are calculated if the count is more than six then consider it as an outlier because it is difficult for any taxi to carry more than six people. Second, missing values in specific attribute fields were filled and removed the trip duration which are far less or far bigger than the mean values.

- 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 canceled the trip due to some technical issue in software, etc.

3.4. Exploratory Data Analysis (EDA): Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods.

3.5. Regression Model testing:

Various ML regression models are testing and predicting the result for the best model and their differences are also compared.

4. Analysis:

This section consists of details regarding the visual results:

4.1 GPS Map of NYC taxi driven

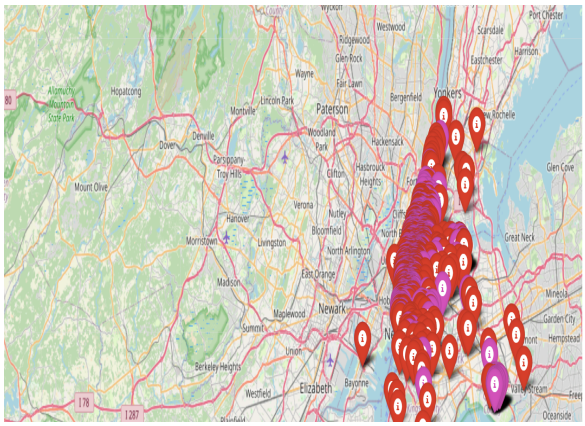


Figure 1: shown the GPS map of New York's city

4.2.1Trip duration bar chart:

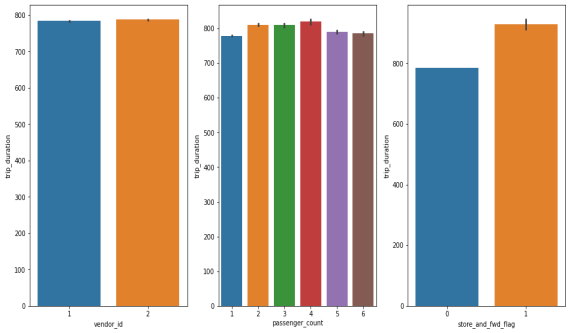


Figure2: Terrorist activities by region

Here we are plot a bar charts which is Vendor id', 'trip duration', 'passenger count, trip duration, 'store and fwd flag', trip duration respectively

4.2.2 Trip in 24 hours bar chart:

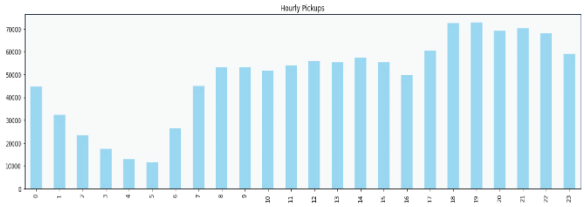
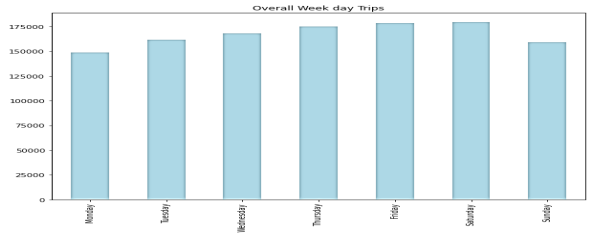


Figure 3: shows the variation per hours

4.2.3 Total trips per weekday:



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

4.2.3 Total trips per month:

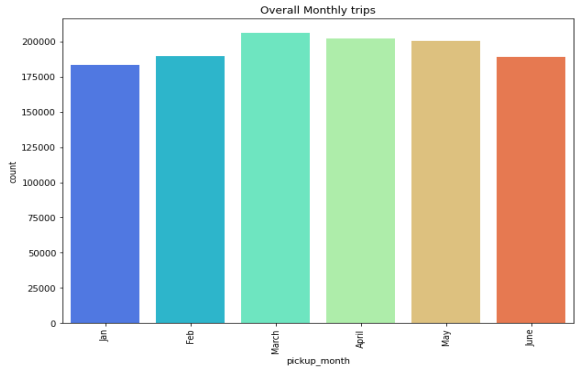
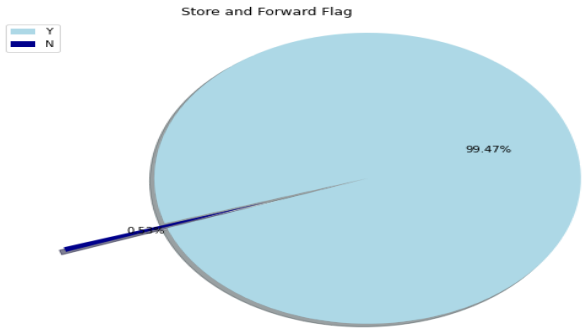


Figure 4: Total trips per month

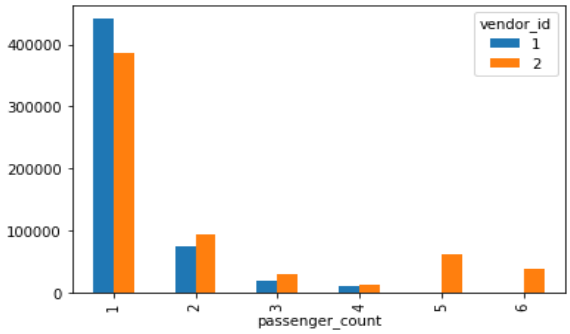
-Number of trips in a particular month March, April and May marking the highest.
-January being lowest probably due to extreme Snowfalls NYC.

4.3 Store and forward pie chart:



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

4.4 Passengers count and vendor:

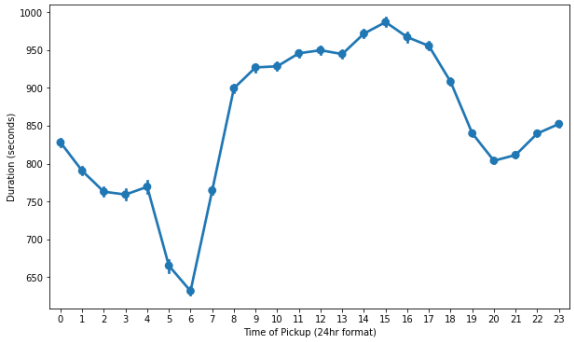


It seems that most of the big cars are served by the Vendor 2 including minivans because other than passenger1, vendor2 has majority in serving more than 1 passenger count and that explains it greater share of the market.

4.5 Trip Duration:

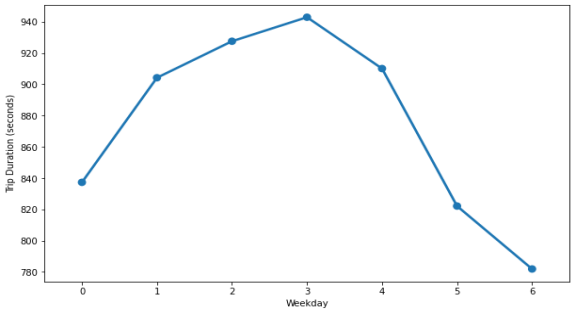
4.5.1 Trip Duration per hour:

We need to aggregate the total trip duration to plot it against the month. The aggregation measure can be anything like sum, mean, median or mode for the duration. Since we already did the outlier analysis, we can take the mean to visualize the pattern which should not result in the bias of the general trend.



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.

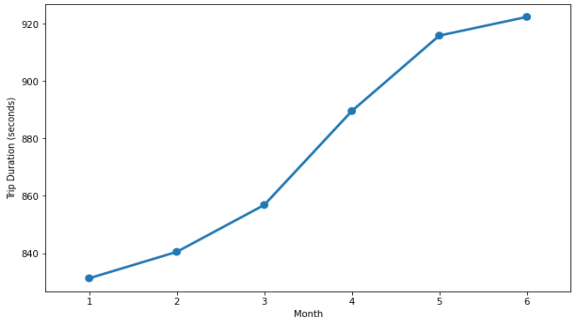
4.5.2 Trip duration per weekday:



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 the trip duration on Thursday is the longest among all days.

4.5.3 Trip duration per month:

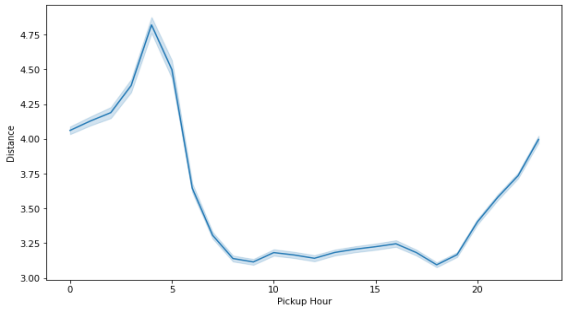


We can see an increasing trend in the average trip duration along with each subsequent month. The duration difference between each month is not much. It has increased gradually over a period of 6 months. It is lowest during February when winters start declining.
There might be some seasonal parameters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

4.6 Distance:

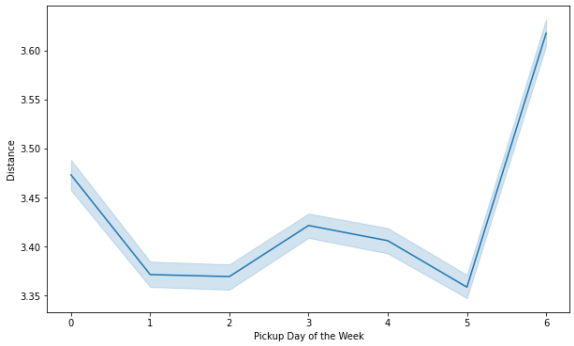
4.6.1 Distance per hour:

Now, let us check how the distance is distributed against different variables. We know that trip distance must be more or less proportional to the trip duration if we ignore general traffic and other stuff on the road. Let's visualize this for each-hour-now.



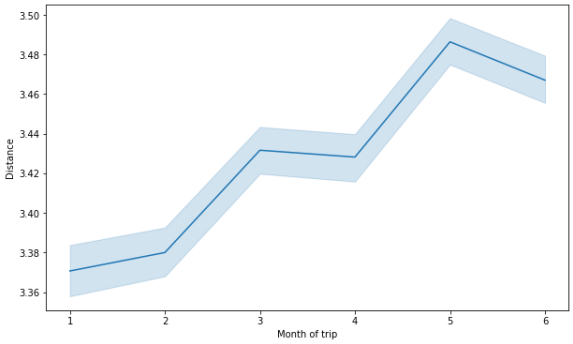
- Trip distance is highest during early morning hours which can account for some things like:
    1. Outstation trips taken during the weekends.
    2. Longer trips towards the city airport which is located in the outskirts of the city.
- Trip distance is fairly equal from morning till the evening varying around 3 - 3.5 kms.
- It starts increasing gradually towards the late-night hours starting from evening till-5-AM and decrease steeply towards morning.

4.6.2 Distance per Weekday:



So, it's a fairly equal distribution with average distance metric varying around 3.5 km/h with Sunday being at the top may be due to outstation trips or night trips towards the airport.

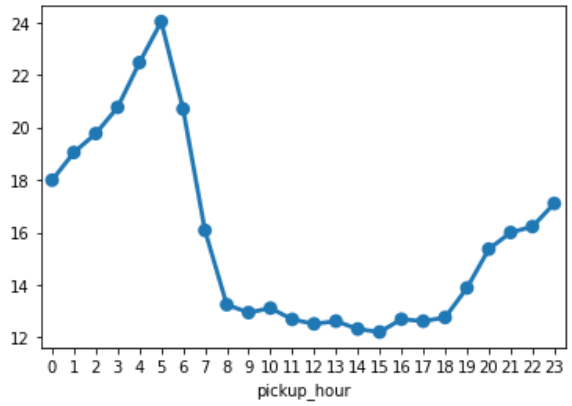
4.6.3 Distance per month:



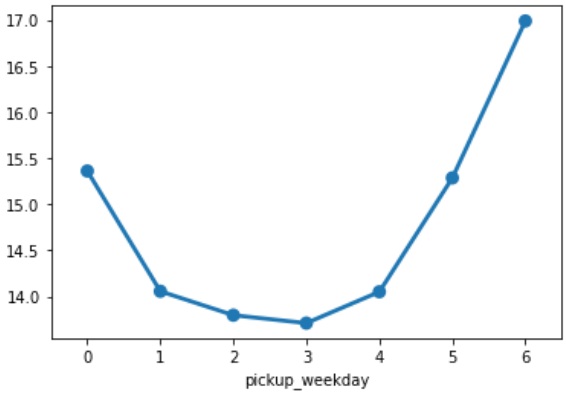
Here also the distribution is almost equivalent, varying mostly around 3.5km/h with 5th month being the highest in the average distance and 2nd month being the lowest.

4.7 Speed:

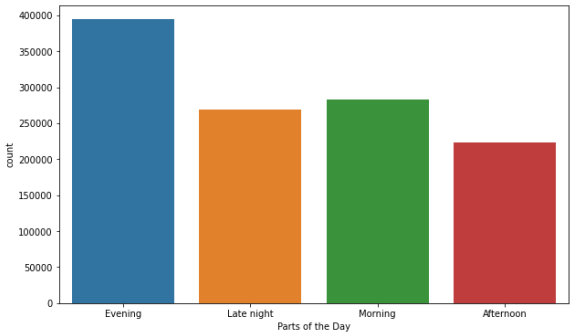
4.7.1 Average speed per hour:



- The average trend is totally in line with the normal circumstances.
- Average-speed tends to increase after late evening and continues to increase gradually till the late early morning hours.
- Average taxi speed is highest at 5 AM in the morning, then it declines steeply as the office hours approaches.
- Average taxi speed is more or less same during the office hours i.e., from 8 AM till 6PM in the evening.



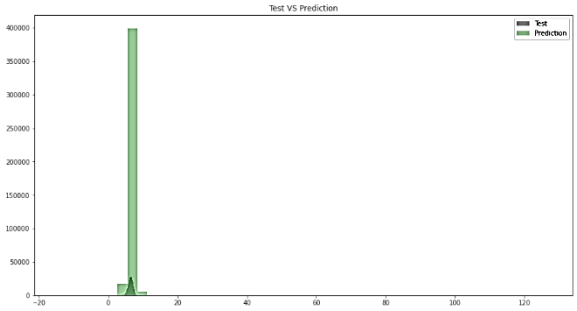
- Average taxi speed is higher on weekend as compared to the weekdays which is obvious when there is mostly rush of office goers and business owners.
- Even on Monday the average taxi speed is shown higher which is quite surprising when it is one of the busiest days after the weekend. There can be several possibilities for such behaviors
  1. Lot of customers who come back from outstation in the early hours of Monday before 6 AM to attend office on time.
  2. Early morning hours customers who come from the airports after vacation to attend office/business on time for the coming week.
- There could be some more reasons as well which only a local must be aware of.
- We also can't deny the anomalies in the dataset. which is quite cumbersome to spot in such a large dataset.



4.7.2Average speed per weekday:

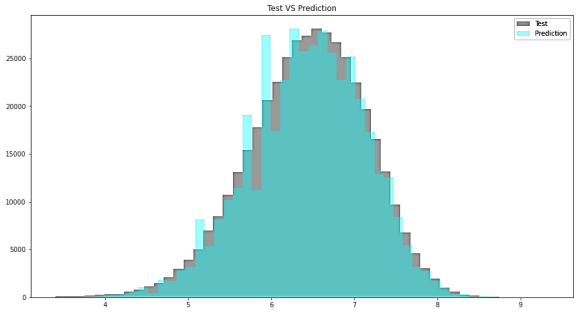
5.Regression Analysis:

5.1. Linear Regression:



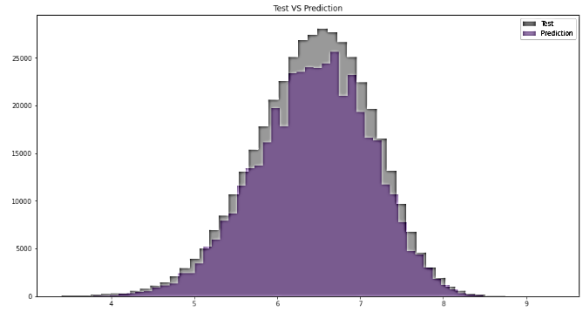
- From the above visualization, we can clearly identify that the Linear Regression isn't performing good.
- The Actual Data (in Grey) and Predicted values (in Green) are so much differing. We can conclude that Linear Regression doesn't seem like a right choice for Trip duration prediction.

5.2. Decision Tree:



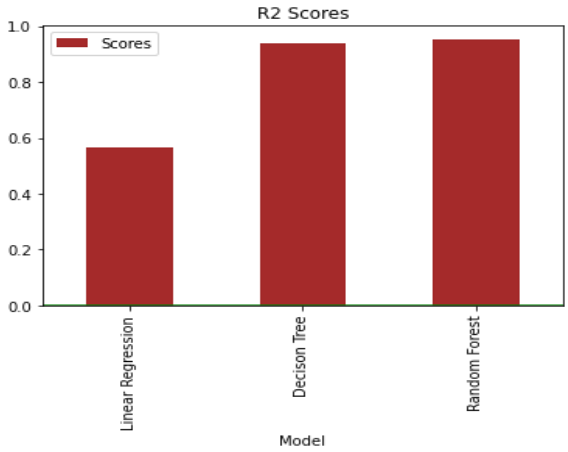
- From the above visualization, we can clearly identify that the Decision Tree Algorithm is performing good.
- The Actual Data (in Grey) and Predicted values (in blue) are as close as possible. We can conclude that Decision Tree could be a good choice for Trip duration prediction.

5.3 Random Forest:



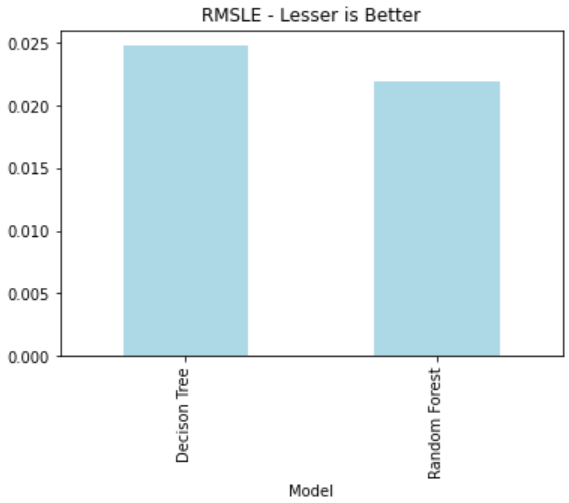
- From the above visualization, we can clearly identify that the Random Forest Algorithm is also performing good.
- The Actual Data (in Grey) and Predicted values (in violet) are as close as possible. We can conclude that Random Forest could be a good choice for Trip duration prediction.
- Similarly, we can Hyper tune Random Forest to get the most out of it.

5.4 R2 Scores Evaluation:



Although, our Evaluation Metric isn't R2 Score but I'm just plotting them to check the Good Fit. We're getting good fit score for Decision Tree and Random Forest, i.e., close to 1.0.

5.5 RMSLE Evaluation:



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.

6. Technologies used:

Python: Python is a high-level interpreted language that supports different platforms like Windows, Linux, Mac, Raspberry Pi, etc. Python can be used for creating web applications, database systems, handling big data, and performing complex mathematical calculations. Python can be treated in an object-oriented, functional or procedural way.

Google Colab: Colaboratory, or "Colab" for short, is a product from Google Research. Colab allows anybody to write and execute arbitrary python code through the browser, and is especially well suited to machine learning, data analysis and education. More technically, Colab is a hosted Jupyter notebook service that requires no setup to use, while providing access free of charge to computing resources including GPUs.

Python packages: Following are some of the python packages used in this project.

Matplotlib: Matplotlib is a visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualiz- ation library built on NumPy arrays and designed to work with the broader SciPy stack. It was introduced by John Hunter in the year 2002.One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.

Pandas: Pandas is an open-source library that is made mainly for working with relational or labeled data both easily and intuitively. It provides various data structures and operations for manipulating numerical data and time series. This library is built on top of the NumPy library. Pandas is fast and it has high performance & productivity for users.

NumPy: It provides structures for multiple dimensional array objects and tools for related operations. NumPy is usually used for high performance scientific computational tasks.

Seaborn: Seaborn is a visualization library for statistical graphics plotting in Python. It provides default styles and color palettes to make statistical plots more attractive. It is built on the top of matplotlib library and also closely integrated to the data structures from pandas. Seaborn aims to make visualization the central part of exploring and understanding data. It provides dataset-oriented APIs, so that we can switch between different visual representations for the same variables for better understanding of the dataset**.**

Linear regression: Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

Decision tree regression: observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output. Continuous output means that the output/result is not discrete, i.e., it is not represented just by a discrete, known set of numbers or values.

Random Forest Regression: is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines

predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.

6. Conclusion:

- It is to be noted that highest number 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.
- 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.

References:

1. Google .com
2. Kaggle.com
3. GeeksofGod.com