Team Member's Name, Email and Contribution:

- Vinayak Marathe: vinmarathe100@live.com
- Riya Patel: riyapatelrp8308@gmail.com
- Muskan Kasere: muskankasere.mk@gmail.com

Contributor's Role :

Vinayak Marathe :

- Worked on Problem Statement with Business Objective
- Datasets Reading and Mapping the NYC datasets using Folium Library
- Worked on whole Feature Manipulation and Data Pre-Processing
- Prepared different type Model and implemented it (Model Implementation and Preparation)
- Performed Hypothesis Testing with 3 hypothetical statement
- Worked on Competitive Advantages and Business Goals of our Project
- Presentation Preparation

Riya Patel :

- Worked on Data Collection and Data Preview
- Performed Data Wrangling on dataset:
  Typecasting
  Finding top trips, Peak hours, longest trip, shortest trip
- Visualizing the Data using Maps, Scatter Plots, Count Plot, and Correlation Heatmap. And some Colours grading on Data visualisation.
- Handling data skewness, applying various transformation such as Logarithmic, boxcox on data
- Performed Outlier Detection and visualization and Outlier Treatment.
- Prepared some unseen data and done prediction on unseen data using saved model
- Prepared Evaluation Matrices and found insights

Muskan Kasere :

- Worked on Unique Values and Data Explanation
- Performed Data Wrangling on datasets:
  Assign new columns to the data frame
  Finding weekdays and weekends
- Data Visualization using Bar Plots, Point Plots, Route Map, Hist Plot, Pie Chart.
- Worked with latitude and longitude to visualize distribution and map.
- Performed Cross Validation and Hyper-Parameter Tuning
- Analysed Overall Results of the Project /Conclusion
- Summary and Technical Documentation.

GitHub Repo link.

1. Vinayak Marathe: https://github.com/v1git12
2. Riya Patel: https://github.com/riyapatelrp
3. Muskan Kasere: https://github.com/MuskanKasere

Project Link:
https://colab.research.google.com/drive/1pr7cLRwEfTGs85a2-nwCLmkbqfbq7ota?usp=sharing

Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

The NYC taxi trip time prediction project involves building a machine learning model to predict the time it takes for a taxi to travel from one location to another in New York City. The model takes into account various factors such as traffic conditions, weather, time of day, and origin-destination pairs to make its predictions.

In this dataset, there are 11 columns which have vendor_id, pickup_datetime, dropoff_datetime, passenger_count, longitude/latitude, atore_and_fwd_flag and trip duration.hosts and 1458644 rows which have all the information about NYC's Taxi. There are no null values in our dataset.

Our first task is to prepare dataset for our machine learning model. First we load dataset, we started with the **Know Your Data** it involves Null Values/Missing Values, Unique Values, Duplicate Values. Then the **Data Wrangling**, in this we make your dataset analysis ready and deep dive into the relationships of variables by knowing longest/shortest trip, trip duration, peak hours, number of trips in each hour, weekends/weekdays and many more.

In the **Data Visualization**, figuring out various aspects and relationships among the target and the independent variables. We will do certain steps like dropping unnecessary columns and outliers' removal and deeply analysis your data by using different Charts and Maps (using python libraries folium, seaborn, matlibplot) like Bar plot, Scatter Plot, Count Plot, Categorical Plot, Pie Chart, Point Plot, Route Map, Hist Plot, Correlation Heatmap etc.

Then the next step is **Hypothesis Testing**, in this we define three statements from dataset to obtain final conclusion about the statements through code and statistical testing. First is trip duration with a single passenger and trips with multiple passengers. Second is Trips with a higher distance have a longer duration? Third is Trips during weekdays have a longer duration compared to weekends? In all three we research hypothesis as a null hypothesis and alternate hypothesis then perform appropriate statistical test.

Next is **Feature Engineering & Data Pre-process**ing, in this we handle Outliers, Categorical Encoding, Textual Data Preprocessing, Feature Manipulation & Selection, Data Scaling, Data Splitting, and Handling Imbalanced Dataset.

Last is **ML Model Implementation**, the main part of the project where in train & test data, we divide it with 80:20 ratio. Firstly, we use Linear Regression which give us minimal result. Then, we used Ridge Regression and Lasso Regression and lastly we performed Random Forest which gave us the best accuracy of 72% (r2) among all of the algorithms. In these we explain the ML Model used and its performance using Evaluation metric Score Chart and Cross- Validation & Hyperparameter Tuning.

In conclusion, predicting taxi trip time accurately is an important task for optimizing transportation services in NYC. There have been many efforts to improve the accuracy of trip time predictions, including the use of advanced machine learning techniques, incorporating additional data sources, developing real-time prediction models, improving location accuracy, and incorporating user feedback.

Improving the accuracy of taxi trip time predictions has the potential to provide significant benefits:

1. **For transportation services in NYC**
2. **Including reducing wait times for passengers**
3. **Optimizing driver routes** and
4. **Improving overall transportation efficiency**

As such, it is an important area of research and development that will likely continue to receive attention and investment in the years to come.