# Assignment-3
# Kaggle Challenge - Taxi Fare Prediction

**Top 2-scores :**
**(i) Method A : 4.375**
**(ii) Method B : 4.543**

## Data Set Description -

Contains following columns

- key
- pickup_datetime
- pickup_longitude
- pickup_latitude
- dropoff_longitude
- dropoff_latitude
- passenger_count

## Data Set Exploration -

- key object
- pickup_datetime object
- pickup_longitude float64
- pickup_latitude float64
- dropoff_longitude float64
- dropoff_latitude float64
- passenger_count int64

## Preprocessing of data-

- Firstly we remove all the NaN values in the dataset using
  **train = train.dropna(how = 'any', axis = 'rows')**
  we can remove all of them as they are less in number

- In second step we added a new column in the dataframe named '**Distance**', which is calculated using the haversine formula
- Splitted the key column in " Year, Month, Date ,Day of Week, Hour, Minute" and then used for training
- Dropped rows that had pickup coordinates outside the range or dropoff coordinates outside the range from test data
- As now we have distance in km, we can drop the columns like pickup_latitute, pickup_longitute, dropoff_latitute, dropoff_longitute
- We have removed all the rows in which distance is less than 0
- Also we have removed the rows in which fare_amount is less than 0
- After exploring the dataset I find passenger count is greater than 6 in some of the rows but, in real life taxi cannot accommodate more than 6 so, i removed that rows also
- Trained the model only on rows in which latitude and longitude values are in the range of test data latitude and longitude

# Description of both the methods :

# Method for score 1-

### Random Forest classifier-

- As random forest is immune to overfitting, We were getting **score** of 4.34 which is less than the scores i get with other classifiers
- Random forest is a ensemble classifier
- We fix the parameter as follows
    - ❖ n_estimators=350
    - ❖ max_depth=5
    - ❖ max_leaves=10
    - ❖ other parameter has there default values

# Method for score 2-

### XGBoost classifier-

- Boosting is an ensemble technique where new models are added to correct the errors made by existing models

- Sparse Aware implementation with automatic handling of missing data values so, in this it handles NaN by its own
- XGBoost is an algorithm that has recently been dominating applied machine learning  for structured or tabular data
- **We fix the parameter as follows-**
    - n_estimators=300
    - max_depth=3
    - max_leaves=9
    - other parameter has there default values

**Analysis of all the methods used :**

- I tried for logistic regression classifier also but i am getting a score of 5.45 using that, it may be because logistic regression is  vulnerable to overfitting
- In Linear regression i got a score of 5.5, may be because the fare amount i.e the output is not linearly dependent on the input array we are giving
- As the xgboost and the random forest are the ensemble classifiers so they are giving better score than the other classifiers and also both the classifiers are well suited for tabular form data  and frequently used in kaggle challenges problem, as the data is generally in tabular form