PREDICTING ARRIVAL DELAY OF TRAINS

- First I import all the necessary libraries
  Pandas
  Numpy
  Matplotlib
  Seaborn
  Scikit-learn

- Then read the data in jupyter notebook

- Since the data given was having lots of missing rows i handled those missing values by dropping that missing rows. I also used seaborn to visualize the missing data.

- Our data was also having some categorical variable (“stations”) I used **LabelEncoder** to change this categorical variable into numerical

- Also ML models can’t train dates timestamp so I split the timestamps columns (“ ScheduleArrival”, ”ScheduleDeparture” etc..) into two different columns of respective date and times and then convert this into integers so that our model can easily train the data.

- Then I have done some **Exploratory data analysis** by plotting some scatter plot using Matplotlib and seaborn

- I separately predict the Accual_date and Acctual_time of Model.

- Now since we have lots of columns some columns are not much relevant for the target variable(“ AcctualArrival”). So, I have done some feature selection to remove those irrelevant columns in order to increase the accuracy of the model. I used the **chi-squared** method and **Heatmap** method to do the feature selection of the model.

- After the feature selection I split the data 70% for the training and rest 30% for testing purpose by using **train_test_split** then I trained my model with different Algorithms such as **LinearRegression, DescisionTree** and **XGBOOST** out of these LinearRegression gave us highest accuracy of about 99.99 on test set for the date and about 97.33 for the time on test dataset.
- Finally I combined those two predicted columns.