**AIRLINE TICKET FARE PREDICTION**

**A PROJECT REPORT**

***Submitted by TEAM 8***

SUSHMA SINDHE – TEAM LEAD

HARSHITHA VEJELLA – STORYTELLING AND DATA VISUALIZATION

PHANIDEEPAK GADDAM – DATA COLLECTION AND PREPROCESSING

LAKSHMI SAI KISHORE SAVARAPU – DATA SCIENTIST

***Under the guidance of***

**Professor Ms. Ardiana Sula**

**For the Course DSCI-6007-02**

**DISTRIBUTED AND SCALABLE DATA ENGINEERING**

Logo

Description automatically generated

**UNIVERSITY OF NEW HAVEN**

**WEST HAVEN, CONNECTICUT**

**FALL 2023**

**TABLE OF CONTENTS**

**TITLE PAGE NO**

**ABSTRACT 03**

**CHALLENGE 04**

**SOLUTION 04**

**USER STORY 05**

**GRAPHICAL VIEW OF THE MODEL 06**

**CRISP DM 06**

BUSINESS UNDERSTANDING 07

BUSINESS QUESTIONS 07

DATASET 08

DATA UNDERSTANDING 08

DATA PREPARATION 10

MODELLING A SOLUTION 15

EVALUATION OF THE MODELS 17

DEPLOYMENT OF THE MODEL 18

**INFERENCE 20**

**FUTURE SCOPE 21**

**REFERENCES 21**

**ABSTRACT**

The Airline industry is a significant component of the global travel and transportation sector. It encompasses the buying and selling of airline tickets for passenger travel, both domestic and international. Travelers look for best deals and the Airlines try to make profit by selling their tickets. As the industry evolves, building a flight ticket fare prediction model will be a valuable tool for travelers and the airline industry. These models can benefit both consumers looking for the best deals and airlines striving to optimize their pricing strategies.

Airline ticket prices are influenced by a multitude of factors, making them a complex and dynamic pricing challenge. However, with the power of data and machine learning, we can create models that provide accurate fare predictions. As the industry evolves, these models can adapt to changing market dynamics, ultimately benefiting both consumers and airlines.

Data science and machine learning have a wide range of applications across various domains and industries. These technologies have the potential to extract valuable insights, make predictions, and automate tasks, leading to improved decision-making and efficiency. Predicting airline ticket fares is a complex task that involves multiple variables and factors. Data Science and Machine learning techniques can be used to build an effective solution for predicting airline ticket fares.

Dataset has been collected for the above requirement and preprocessing has been performed on it using various python libraries. Feature extraction has been performed to get valuable insights from the dataset. Machine Learning models like Linear Regressor and Random Forest regressor have been build and the models performance has been evaluated and it is observed that the model is able to perform well with a good prediction rate and can be deployed for real time usage.

**CHALLENGE**

How to predict the price of an Airline ticket based on certain features?

Buying an airline ticket depends on various factors like source, destination, number of stops between source and destination, price range, flight duration, airline company etc. Travelers look for the best deal at a low price with certain specifications while buying a flight ticket. Building a model considering all these factors is a challenging task. The model has to include various requirements like:

What is the source and destination?

What is the price range?

Number of tickets needed?

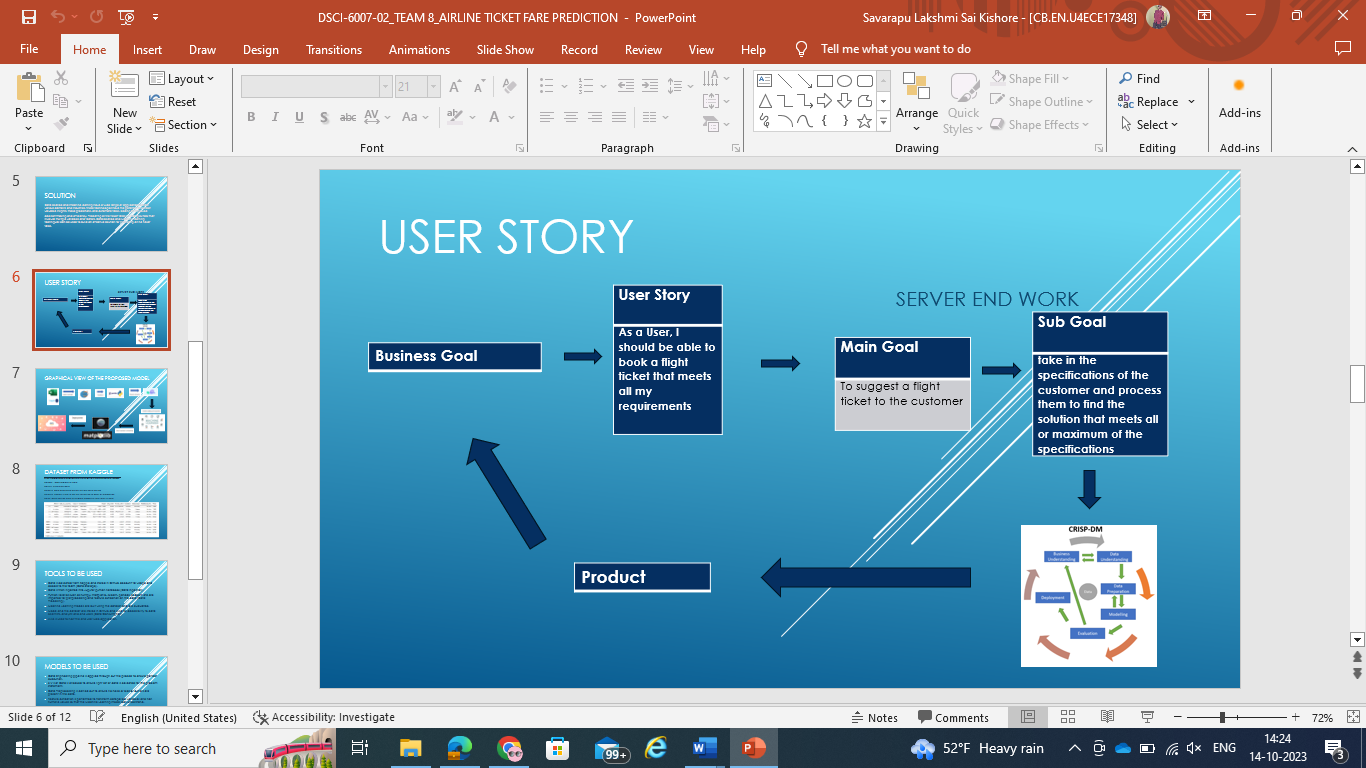
Number of stops in between source and destination?

What is the flight duration?

**SOLUTION**

Considering all sorts of requirements like the above, a Machine Learning model can be built that can predict the price of an airline ticket. Data Science and Machine Learning techniques have been used to solve these kinds of problems in various domains and industries. Many industries have implemented Data Science and Machine Learning techniques in real time cases and have achieved great yield. Fraud detection, Recommendation systems, E-commerce pricing, Healthcare diagnostics are some examples of real time applications of Data Science and Machine Learning techniques. With boundless number of applications of Data Science and Machine Learning Techniques, it is possible to build an efficient model that can be used in real time for predicting Airline ticket price.

**USER STORY**



Title: Flight ticket Fare Prediction

User story: As a user, I should be able to get a predicted price of a flight ticket depending on my requirements so that I can make a decision on my travel plans.

Acceptance Criteria:

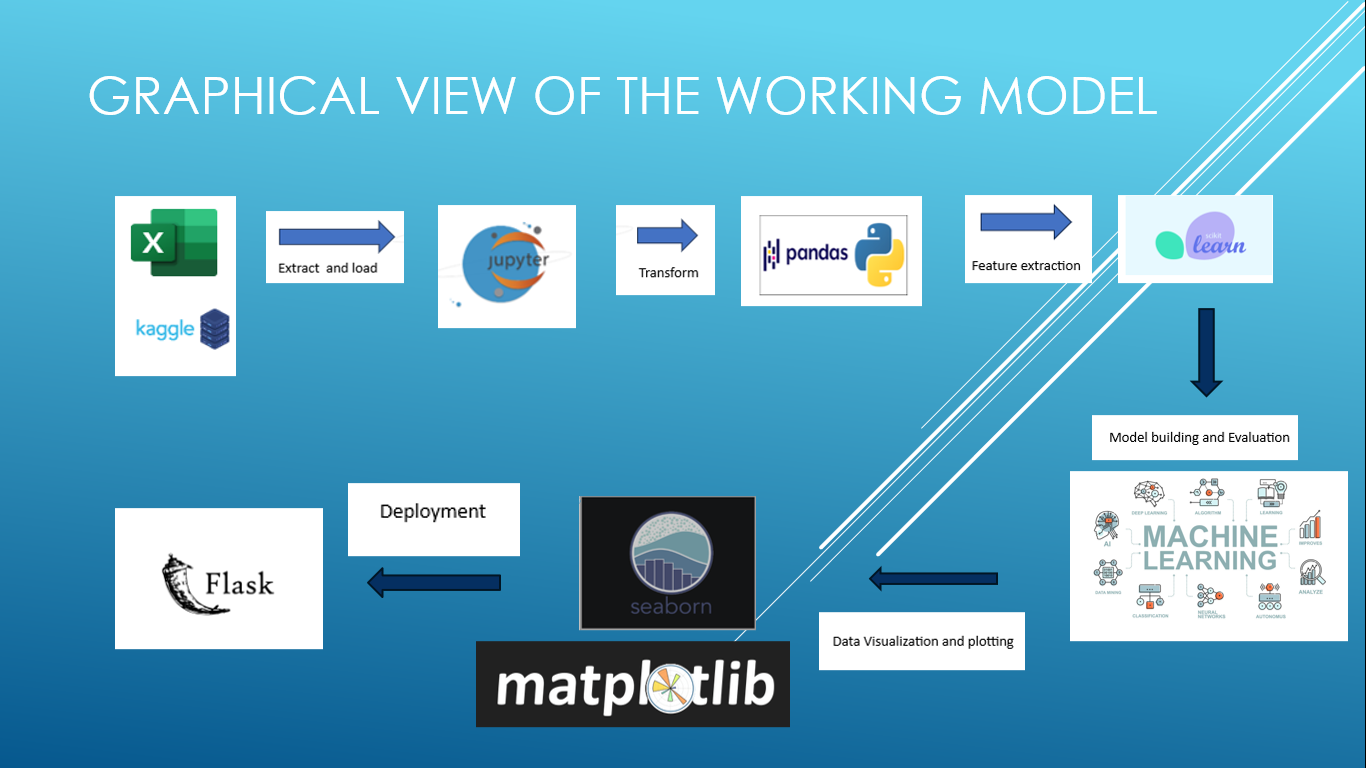
Input Parameters:

Source and destination, Number of stops in between, Flight duration, Airline Company, etc.

Historical data is required for the model building so that the input parameters can be considered while predicting the ticket price and suggesting it to the user.

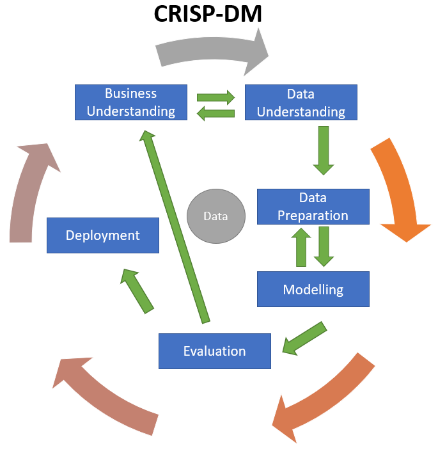
Output: Machine Learning model is built that takes input of all the parameters and predicts the ticket price to the customer.

**GRAPHICAL VIEW OF THE WORKING MODEL**



**CRISP-DM IMPLEMENTATION**

CRISP-DM stands for Cross Industry Standard Process.



The steps involved in CRISP-DM Methodology are:

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modelling a solution
5. Evaluation of the model
6. Deployment of the model

We are going to implement CRISP-DM Methodology to the project in order to build a successful working Model.

**Business Understanding:**

Primary goal:

To predict the price of an Airline ticket based on given inputs such as source, destination, arrival time, departure time, number of stops in between, duration of journey, date of journey etc.

Stakeholder Identification:

This model helps Airline companies as well as Passengers looking for flight tickets. Travel Agencies can also be taken into consideration here.

Scope and Constraints:

The data is generated before the pandemic i.e., 2019. And the sources and destinations are confined to cities in the country India.

**Business Questions:**

1. Which Airline company flight tickets are expensive?
2. Which Airline company flight tickets are economical?
3. Which location is the most frequent origin and which is the most frequent destination?
4. Which is the busiest route?

**DATASET**

<https://www.kaggle.com/datasets/nikhilmittal/flight-fare-prediction-mh>

5 V’s of Data is checked to ensure right set of data is collected for the problem statement.

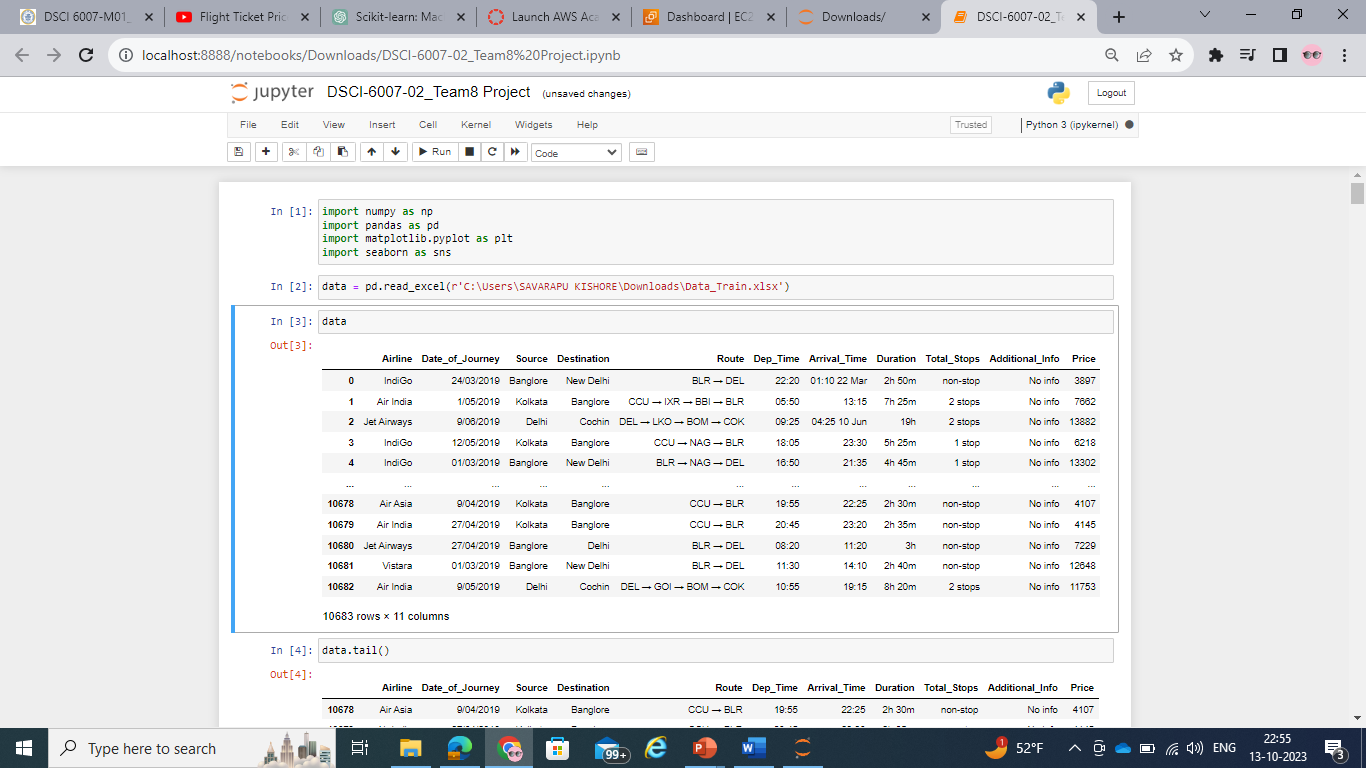
Volume: 10683 samples of data.

Variety: Structured Data.

Velocity: Data generated during the pre-covid period.

Veracity: Kaggle is said to be the trusted go to point of information.

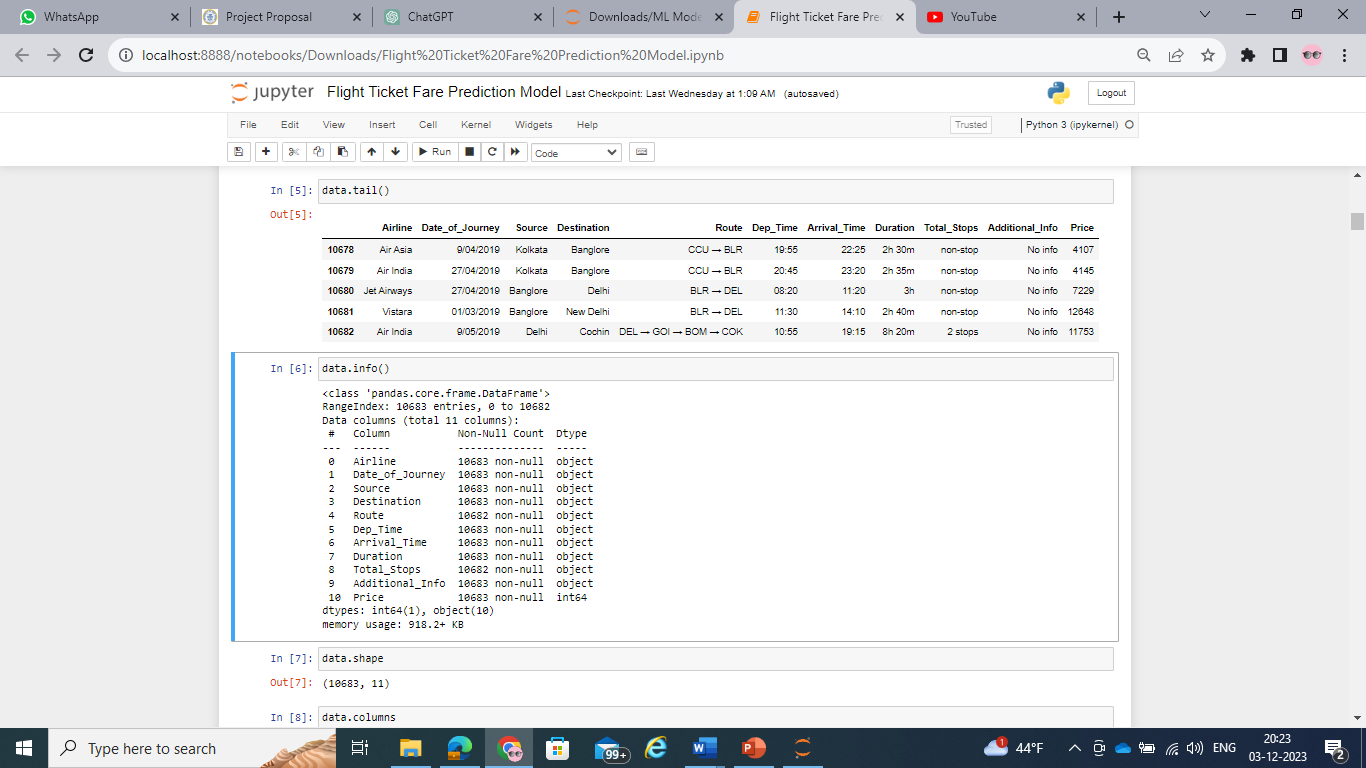
Value: Data can be used to estimate domestic flight fares in India.



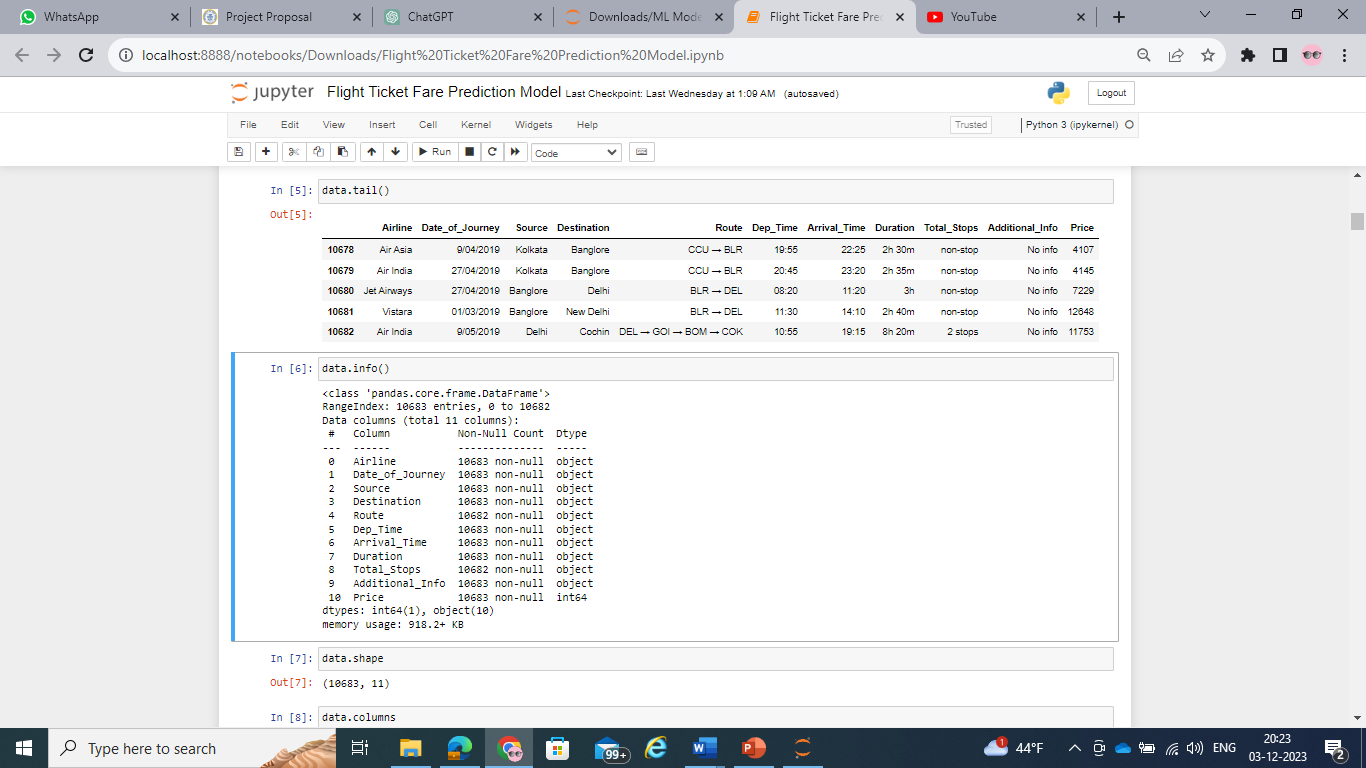
**Data Understanding:**

This is the phase in a data science project where the focus lies on comprehensively exploring and familiarizing oneself with the dataset that will be used for analysis and modelling. It is a crucial step in data mining or machine learning process.

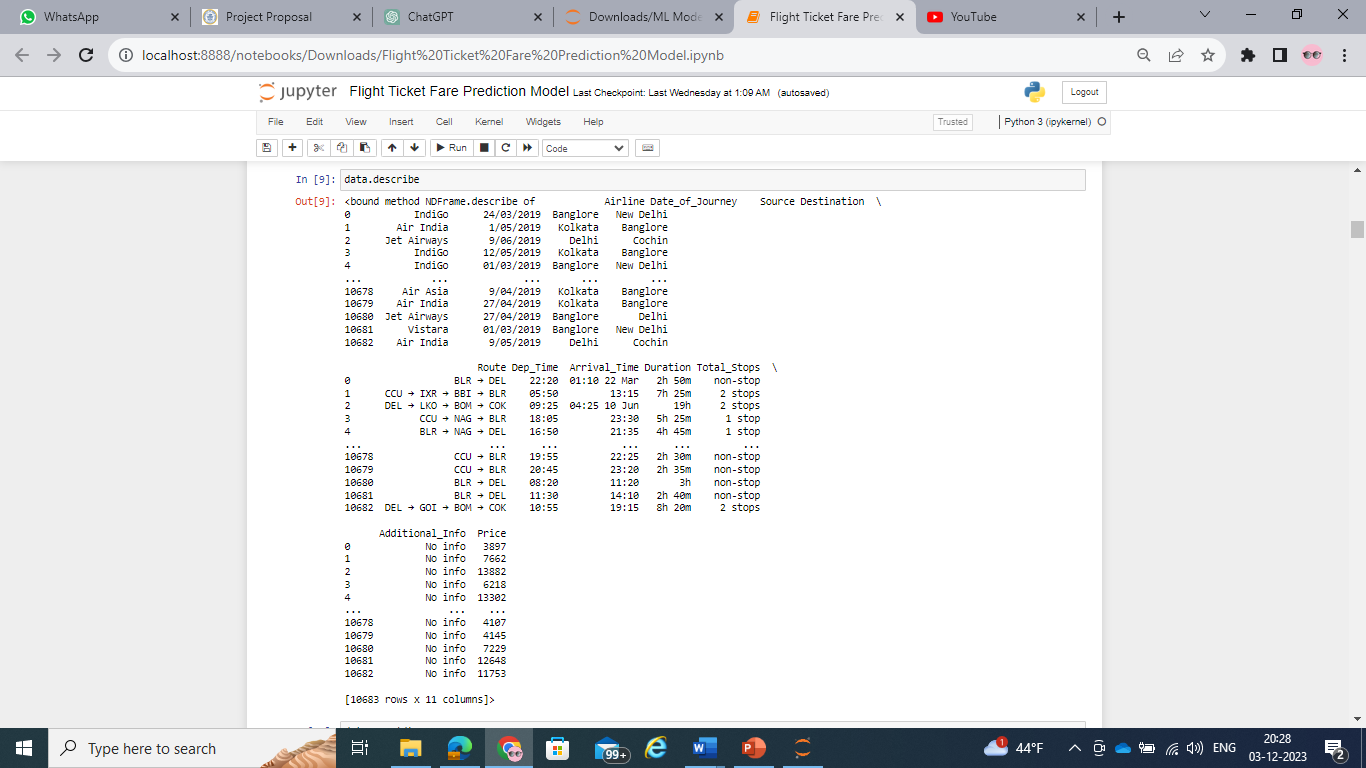
The data.info() function in python is typically used with the pandas library to retrieve information about the data.



The data.shape gives the number of columns and rows in the dataset. Columns are the features of the dataset and rows are the samples of it.



The data.describe function in python is used to generate descriptive statistics of a dataframe. When it is applied to a dataframe, it provides the summary statistics for the numerical columns in the dataset.



**Data Preparation:**

Data Preparation, also known as Data Preprocessing, is a crucial phase in any data science or machine learning project. It involves transforming raw data into a clean, organized format that is suitable for analysis or modeling. Here are key steps involved in data preparation:

Handling missing values, dealing with outliers, encoding categorical variables, feature engineering, train test splitting of data, data transformation, handling skewed data etc.

Identify Missing Data: Determine where data is missing or null in the dataset.

Imputation or Removal: Fill missing values using techniques like mean, median, mode imputation or remove rows/columns with excessive missing data.

Detect Outliers: Identify and handle outliers that might skew analysis or modeling.

Transformation, or Removal: Apply techniques to address outliers based on the context of the data.

Convert Categorical Data: Encode categorical variables into numerical format using techniques like one-hot encoding or label encoding.

Scale Numerical Features: Normalize or scale numerical features to a similar range to prevent dominance of certain features in the model.

Create New Features: Generate new features from existing ones that might improve model performance.

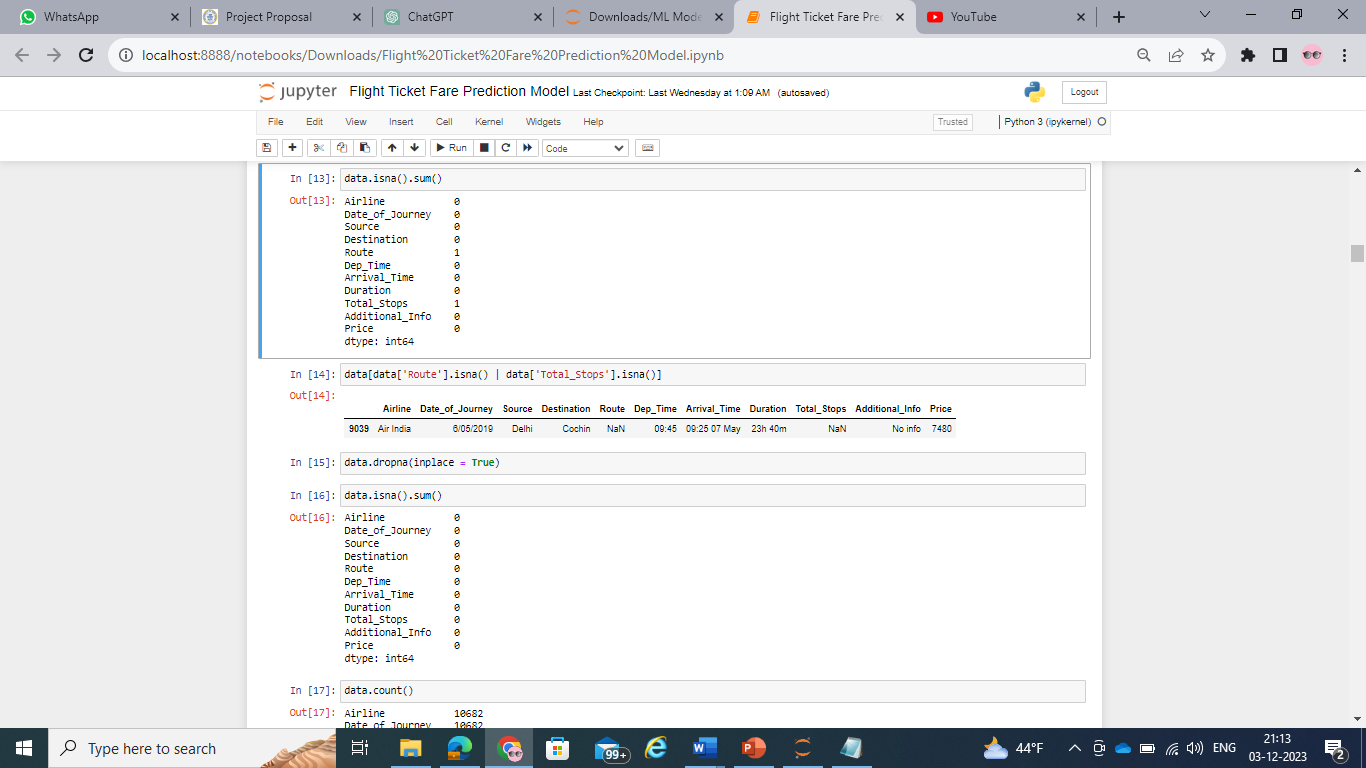
Dimensionality Reduction: Use techniques like PCA or feature selection to reduce the number of features while retaining relevant information.

Split Data: Divide the dataset into training and testing sets to train the model on one portion and validate its performance on another.

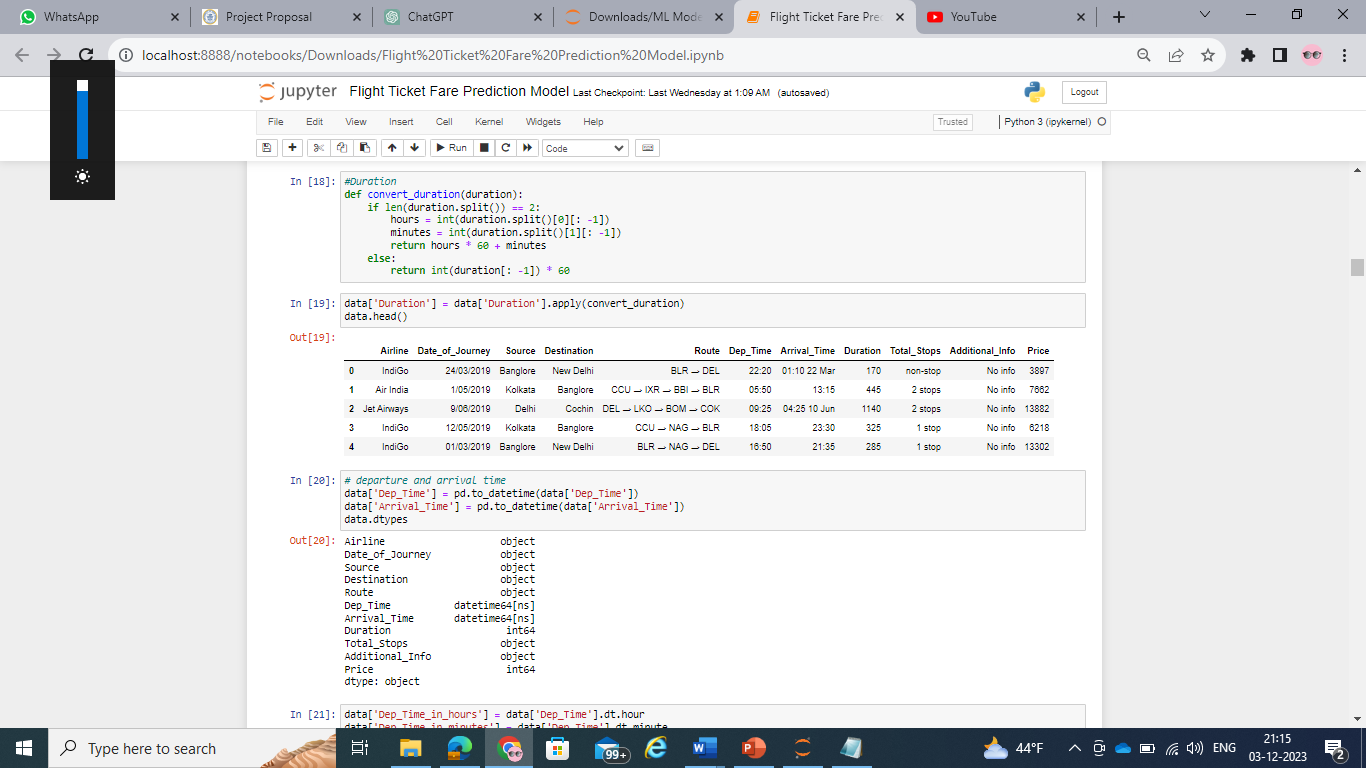
Prepare Data for Modeling: Reshape, reformat, or prepare data specifically tailored for the chosen machine learning algorithms.

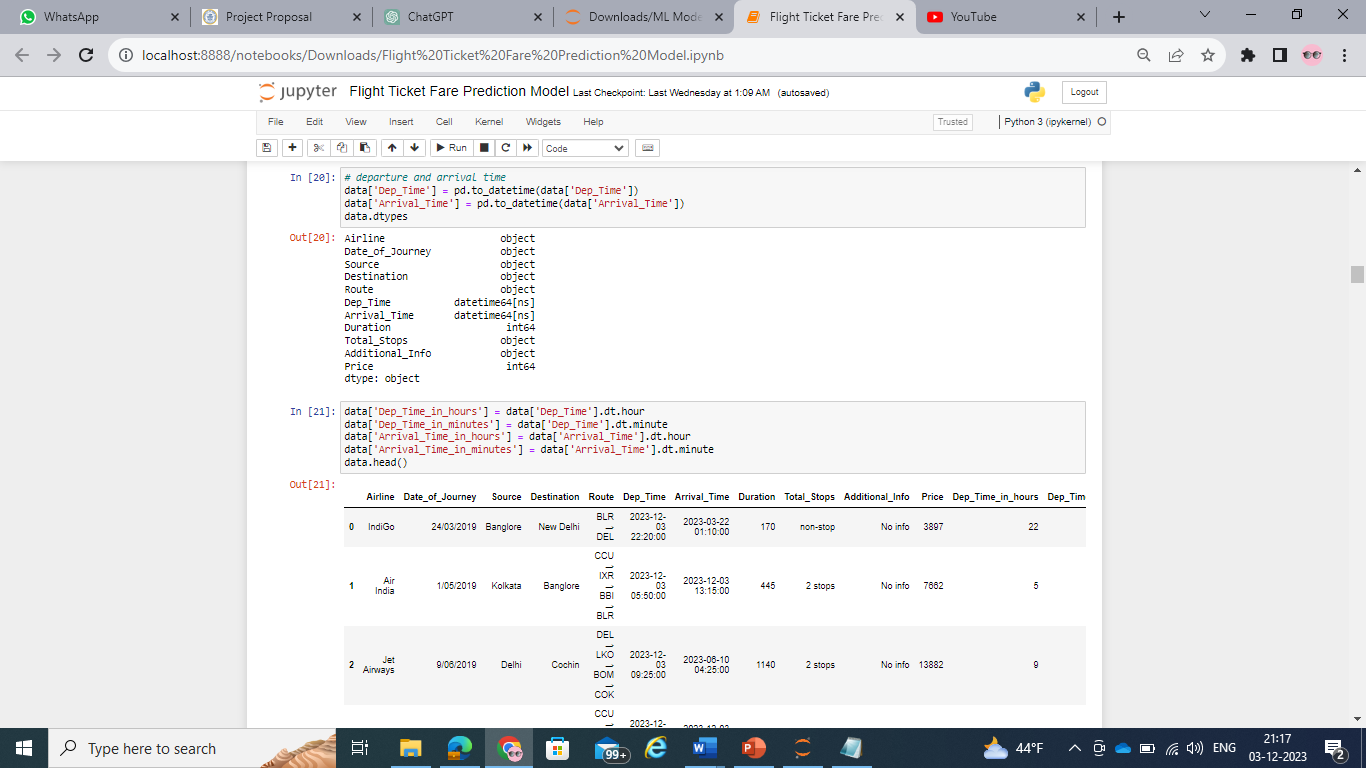
Normalize Skewed Data: Address skewed distributions in certain columns through techniques like log transformation.

Handling of null values:

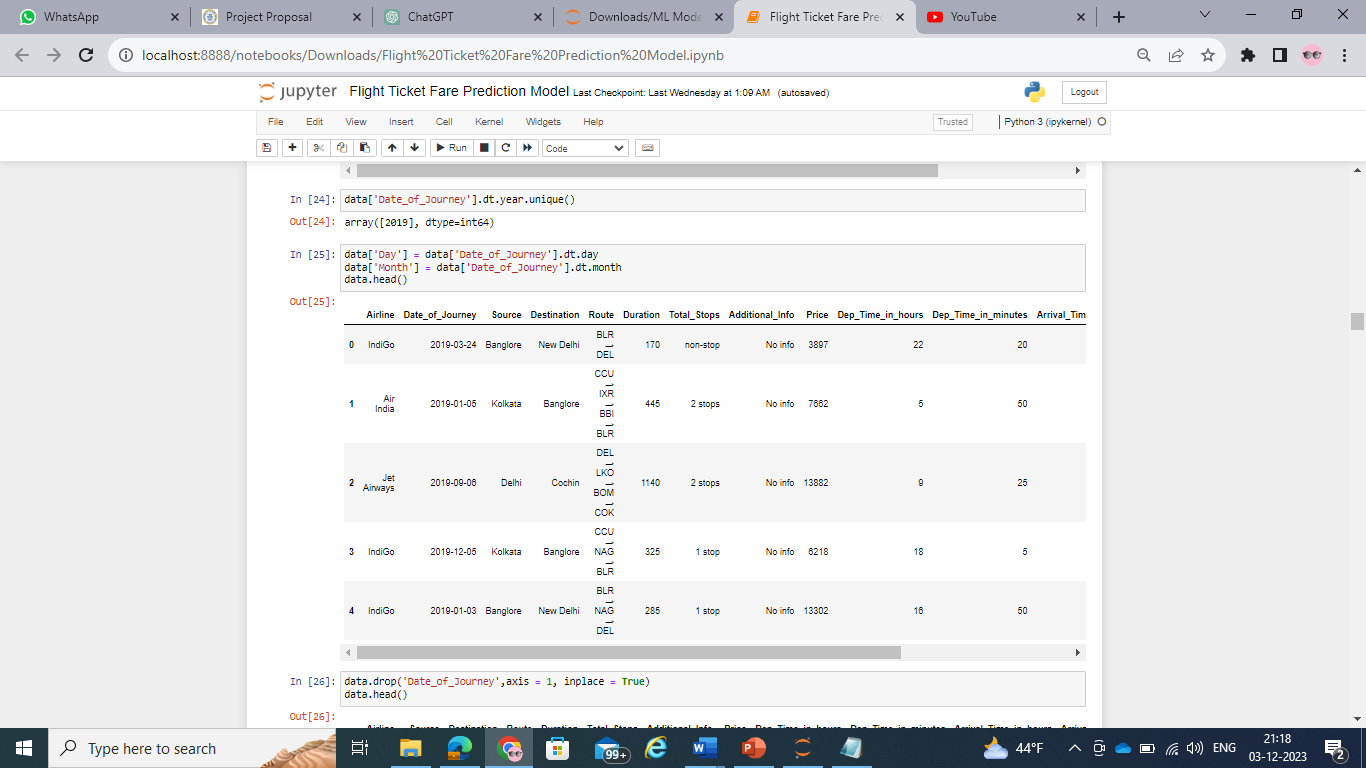


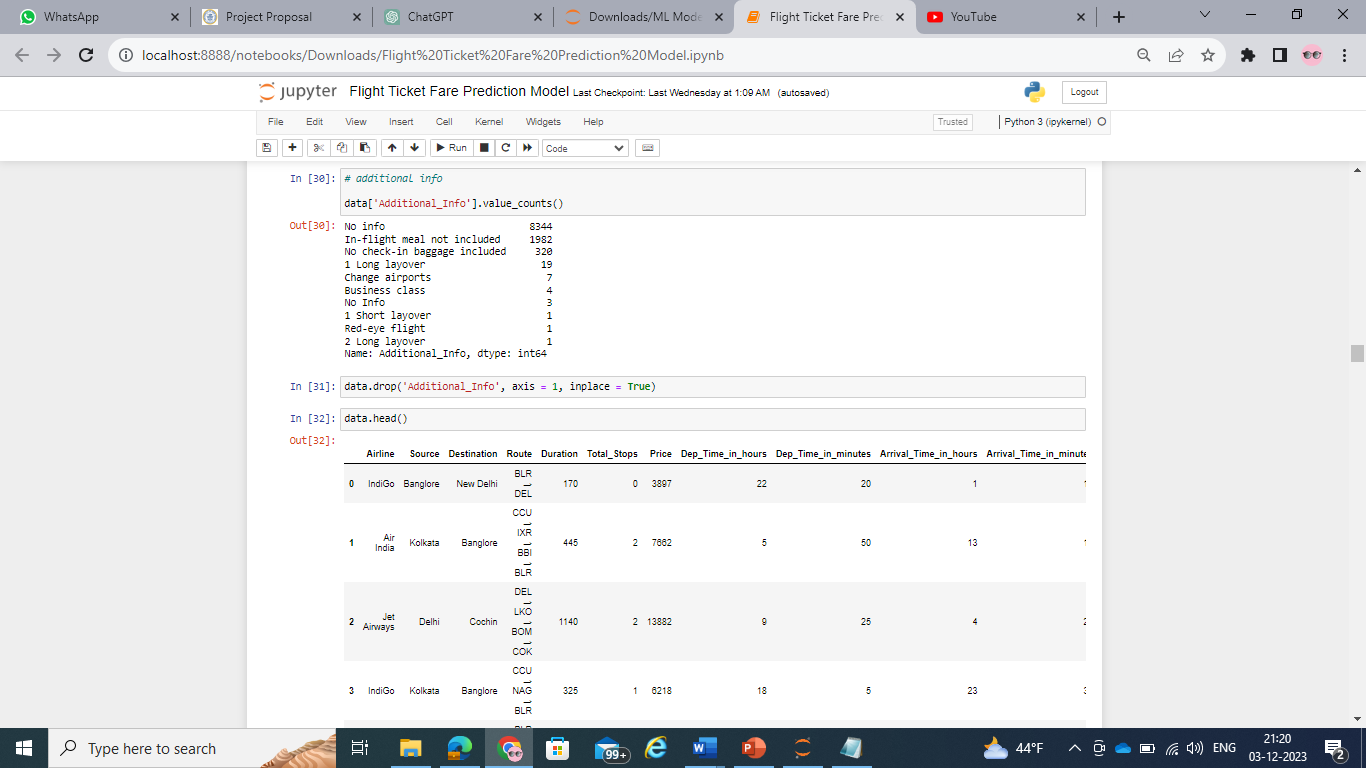
Type conversion of features:

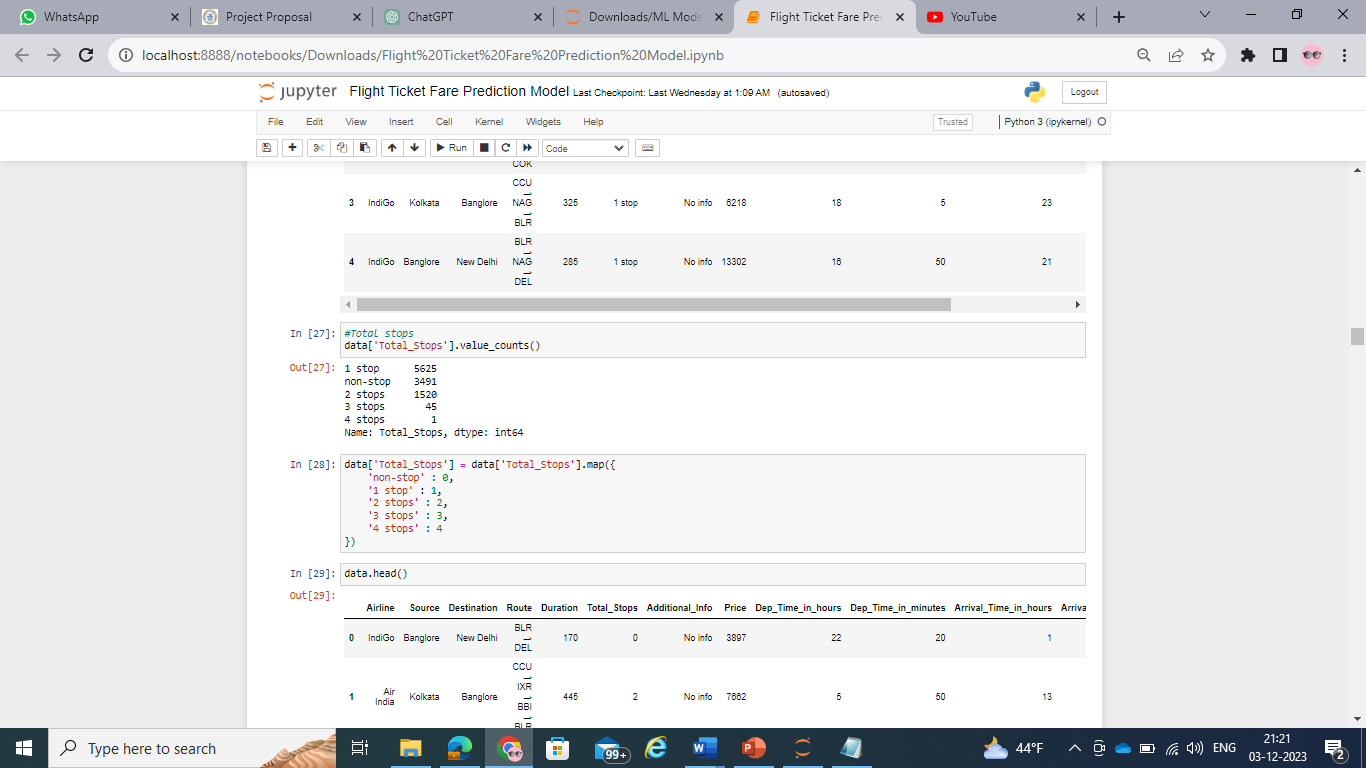




Deleting unnecessary features:

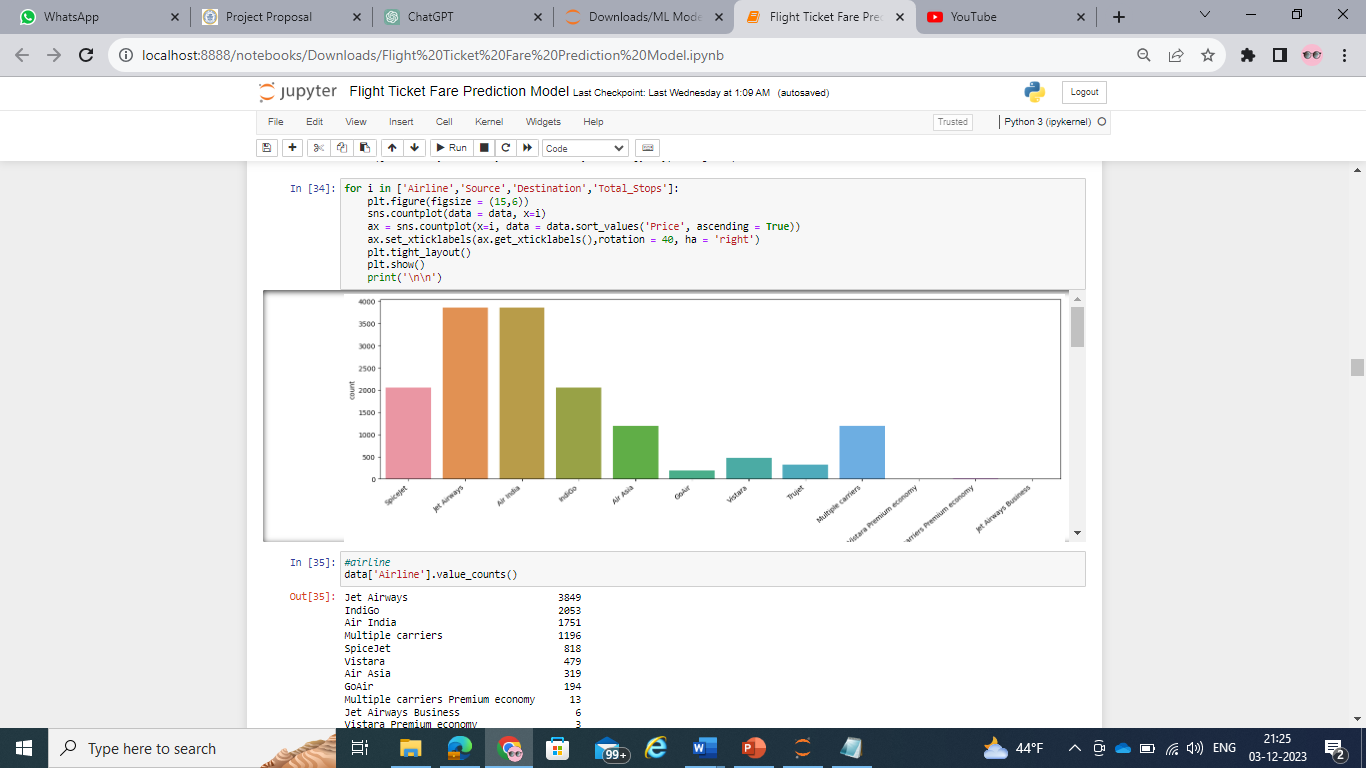


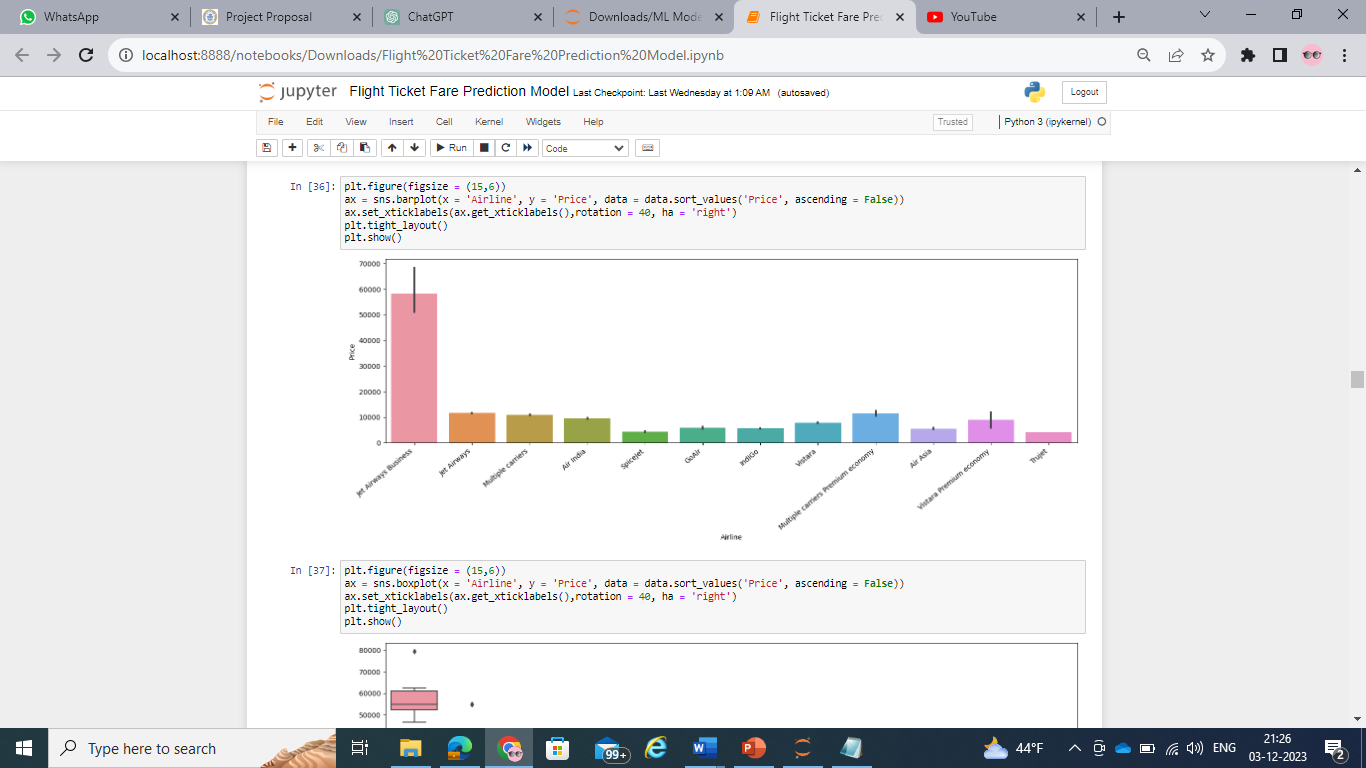


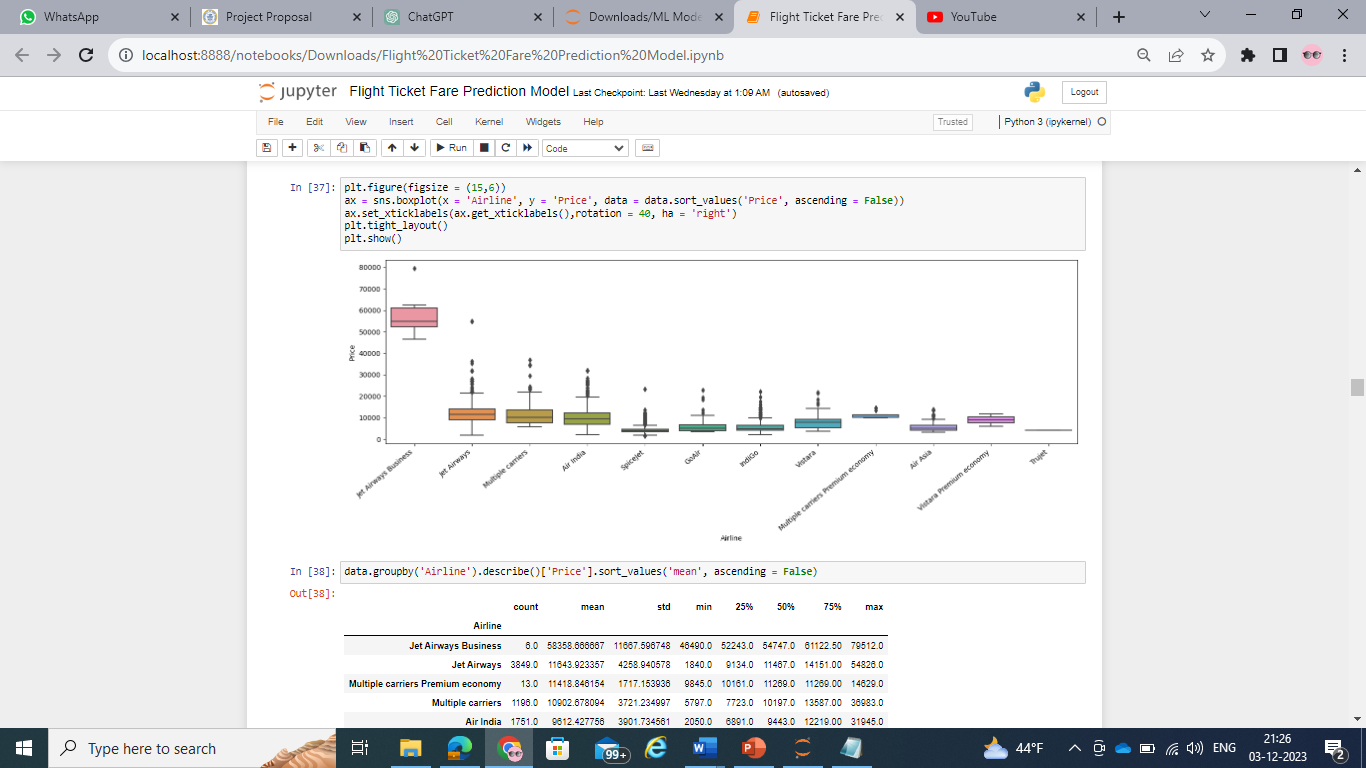


Data Visualization:

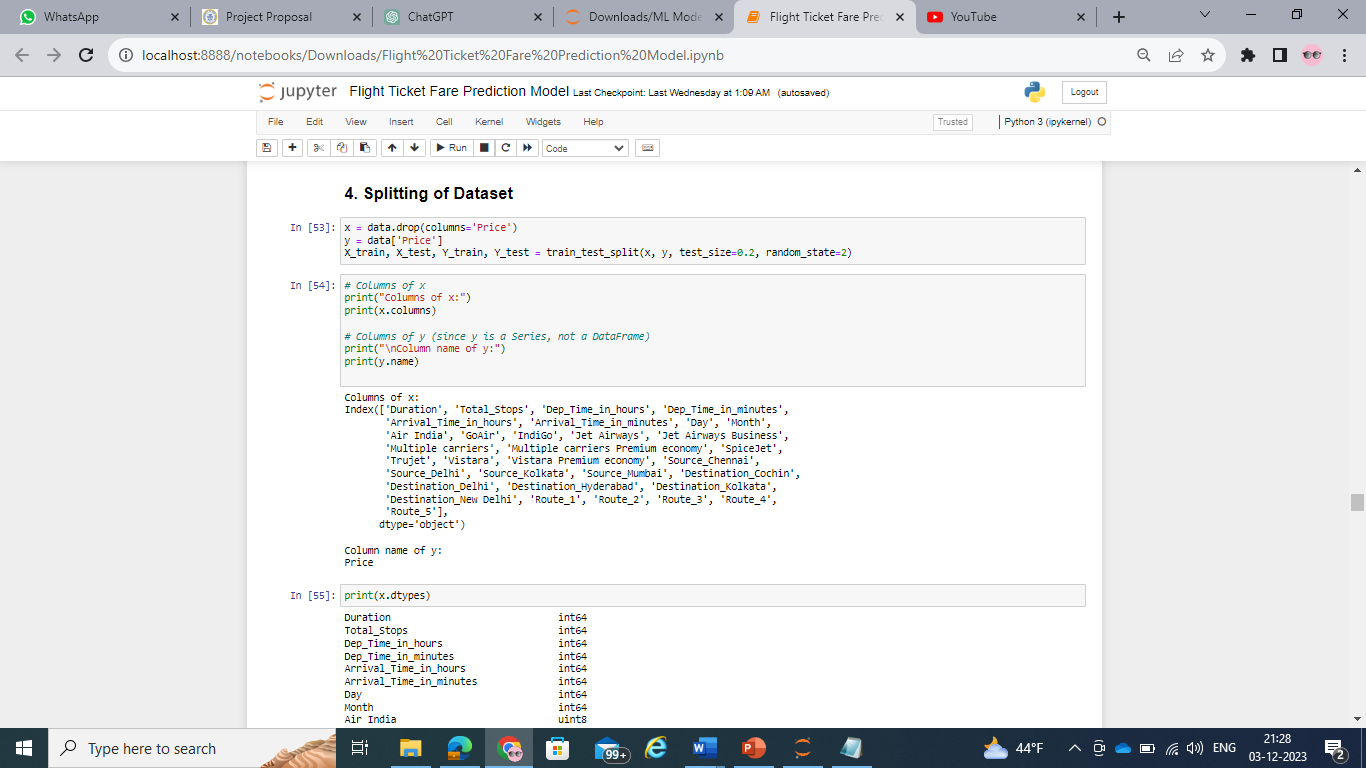
It is the process of making graphical plots on the data. It is performed to draw hidden insights about the data.



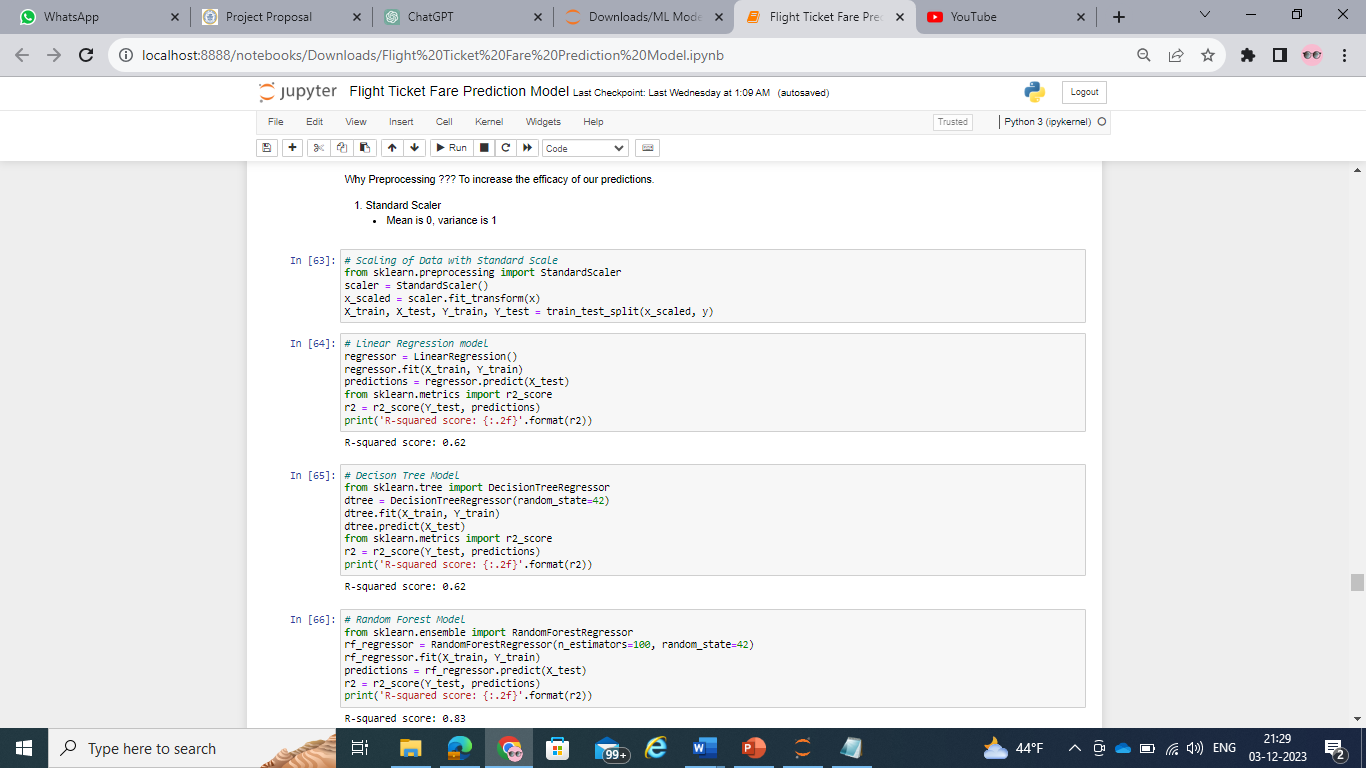




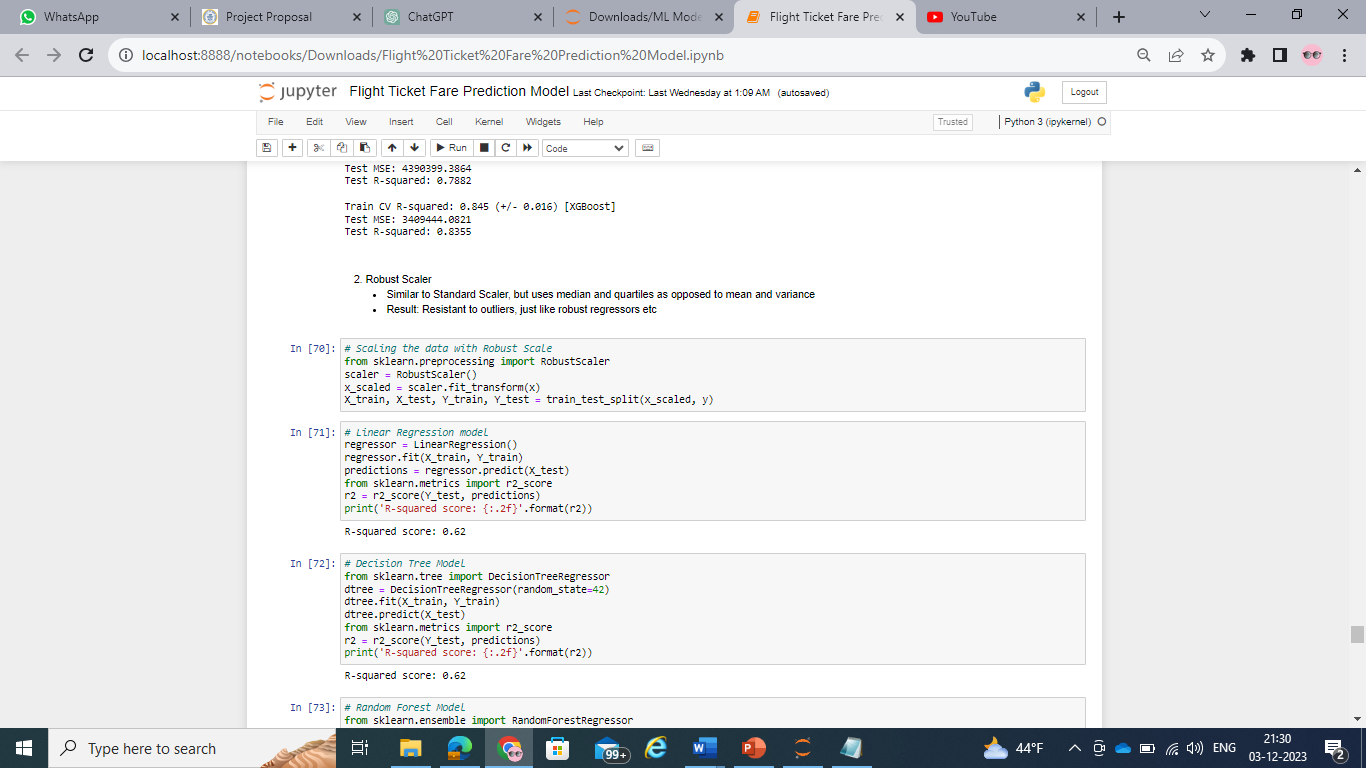
Splitting of data:



Filtering the dataset with Standard scale:



Filtering the dataset with Robust scale:



**Modelling a solution:**

As the target variable is a continuous variable, Regression models listed below are built using the dataset for the problem statement.

• Linear Regression model

• Decision Tree model

• Random Forest model

• Ordinary Least Squared Model

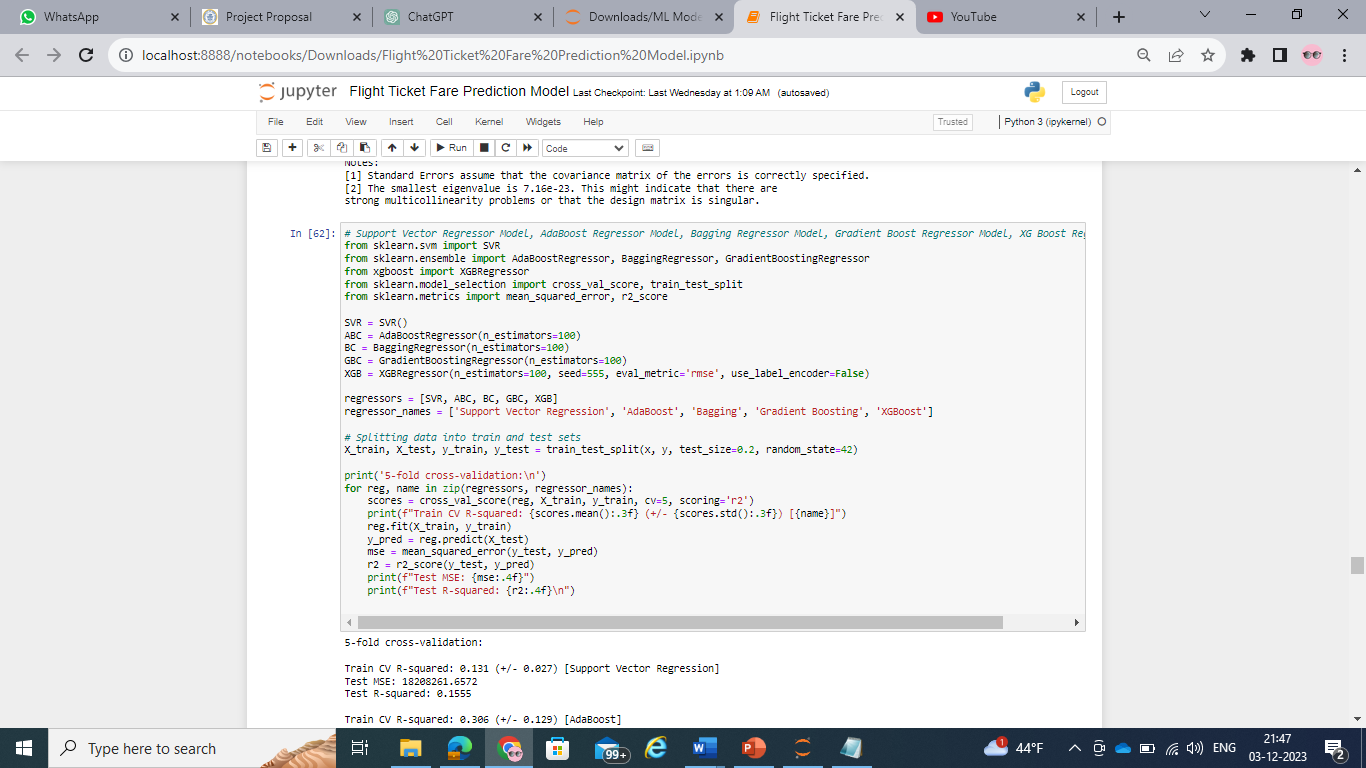
• Support Vector Regressor Model

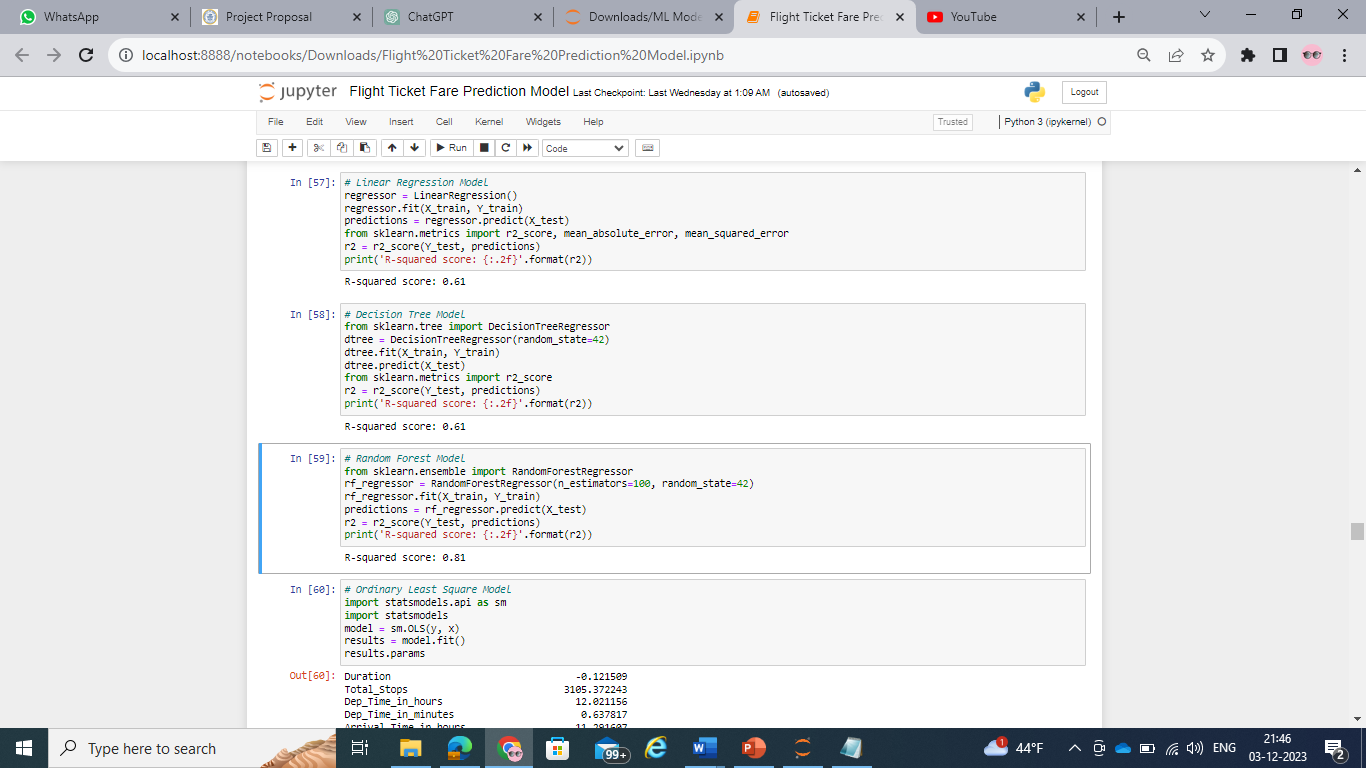
• AdaBoost Regressor Model

• Bagging Regressor Model

• Gradient Boost Regressor Model

• XG Boost Regressor Model





**Evaluation of the models:**

The built models are evaluated by calculating the R squared value and Mean Absolute Error.

The coefficient of determination, often denoted as R-squared (R²), is a statistical measure that represents the proportion of variance in the dependent variable (target) that is explained by the independent variables (features) in a regression model.

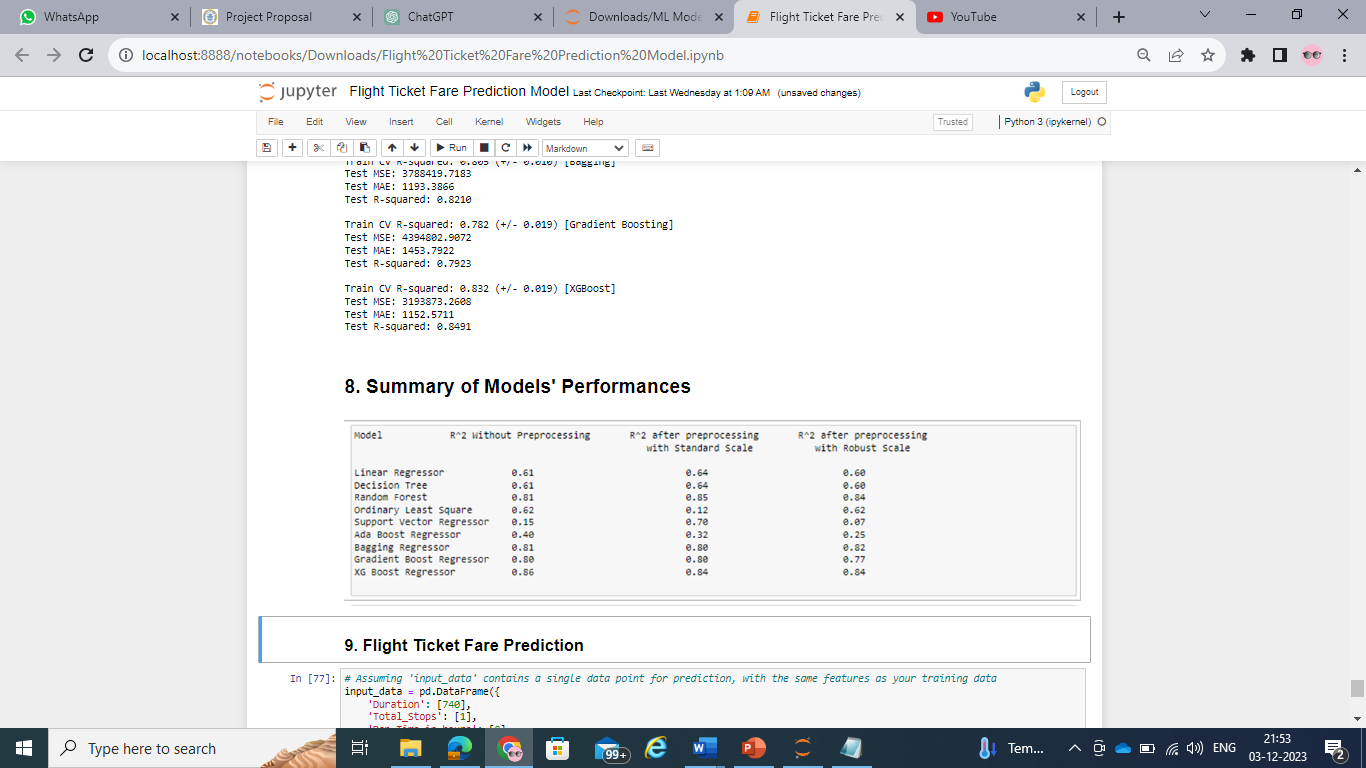
In simple terms, R-squared measures how well the independent variables explain the variability of the dependent variable. It ranges between 0 and 1, with:

0 indicating that the model does not explain any variance in the target variable.

1 indicating that the model perfectly explains the variance in the target variable.

In the context of regression analysis, when you fit a regression model to a dataset, the R-squared value tells you how well the model fits the observed data.

The R-squared values for the models are shown below:



Mean Absolute Error was also calculated for the models.

Mean Absolute Error (MAE) is a metric used to evaluate the accuracy of a predictive model, often in the context of regression analysis. It measures the average absolute difference between the actual and predicted values. The formula for calculating MAE is:

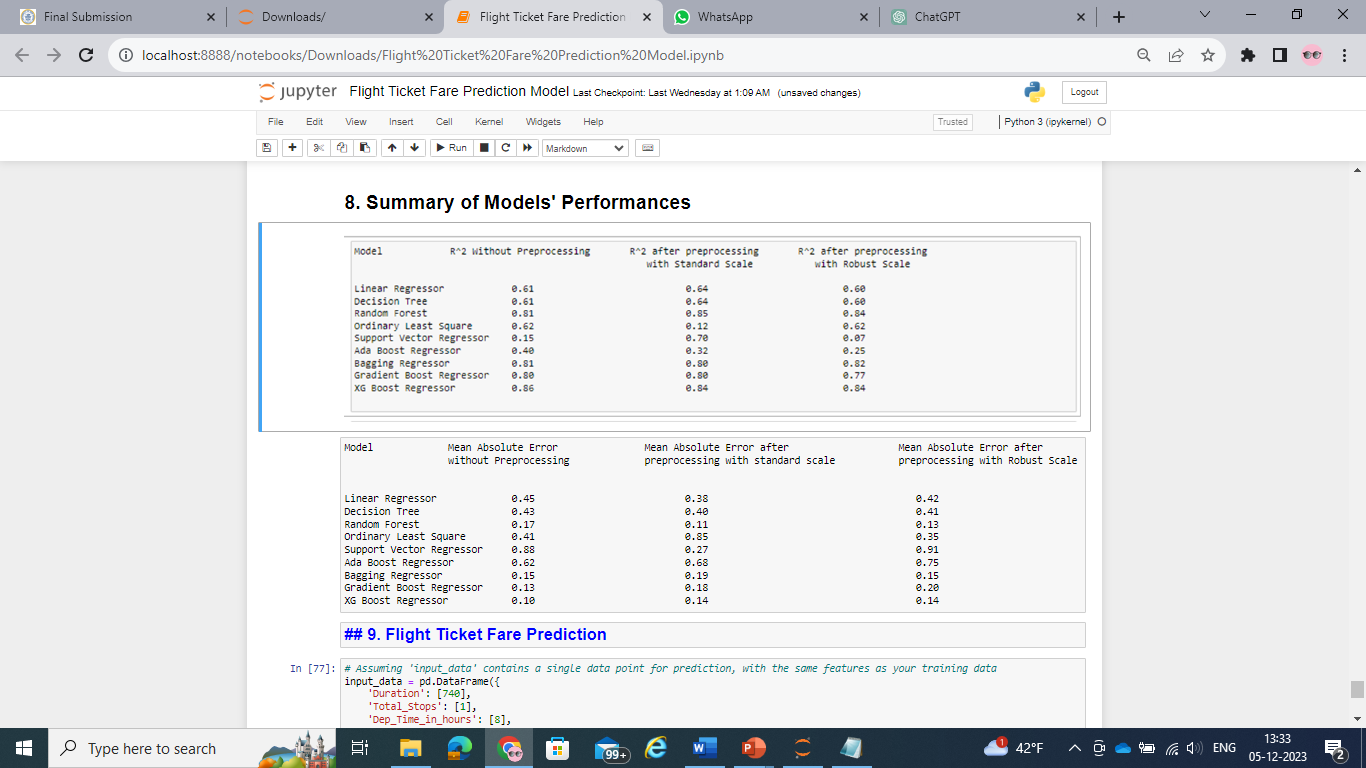
MAE = (1/n) Σ(i=1 to n) |y\_i – ŷ\_i|

y\_i = prediction, ŷ\_i = True value, n = total number of data points

In simpler terms, MAE is the average of the absolute differences between the actual and predicted values. It provides a straightforward and easy-to-interpret measure of the average magnitude of errors in the predictions.

A lower MAE indicates better accuracy, as it means that, on average, the model's predictions are closer to the actual values. MAE is less sensitive to outliers compared to some other metrics like Mean Squared Error (MSE), making it a good choice when dealing with datasets that may contain extreme values.

The Mean Absolute Error values for the models are shown below:

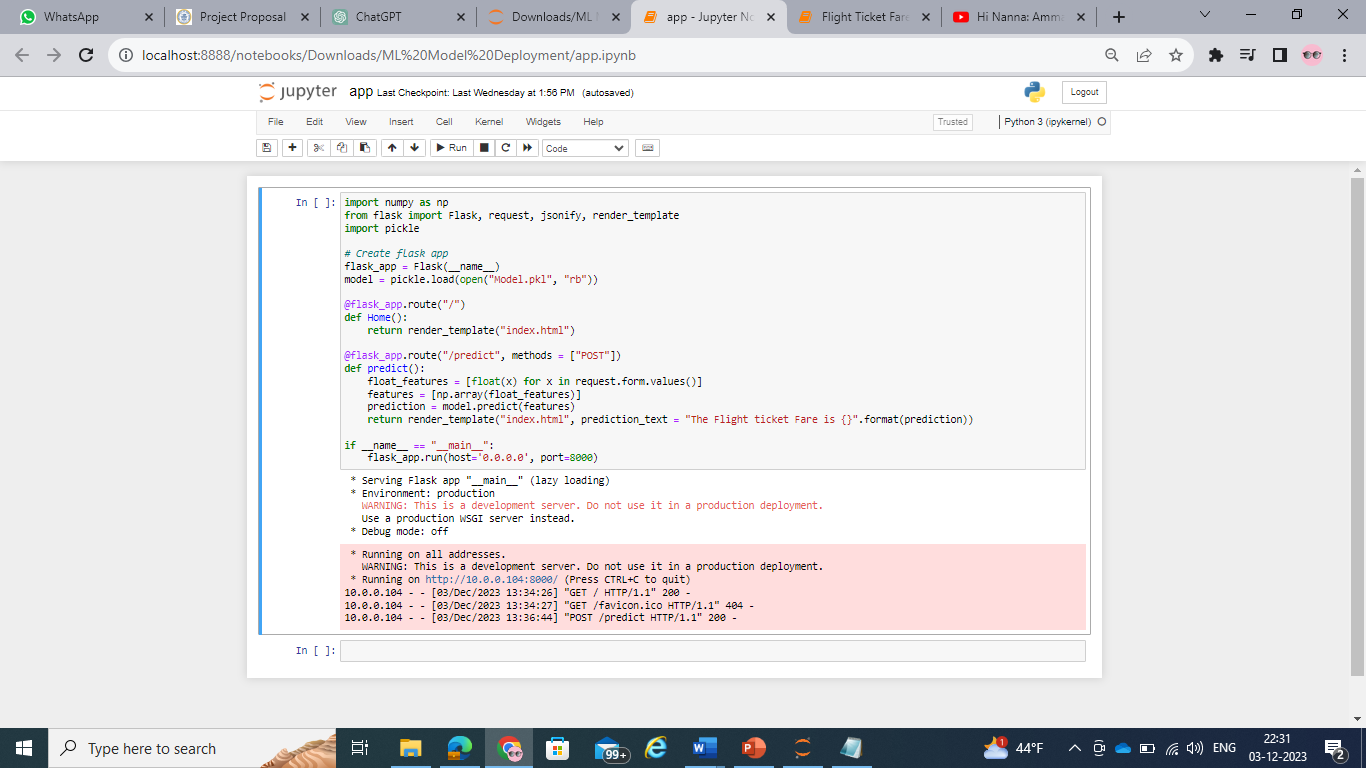


**Deployment of the model:**

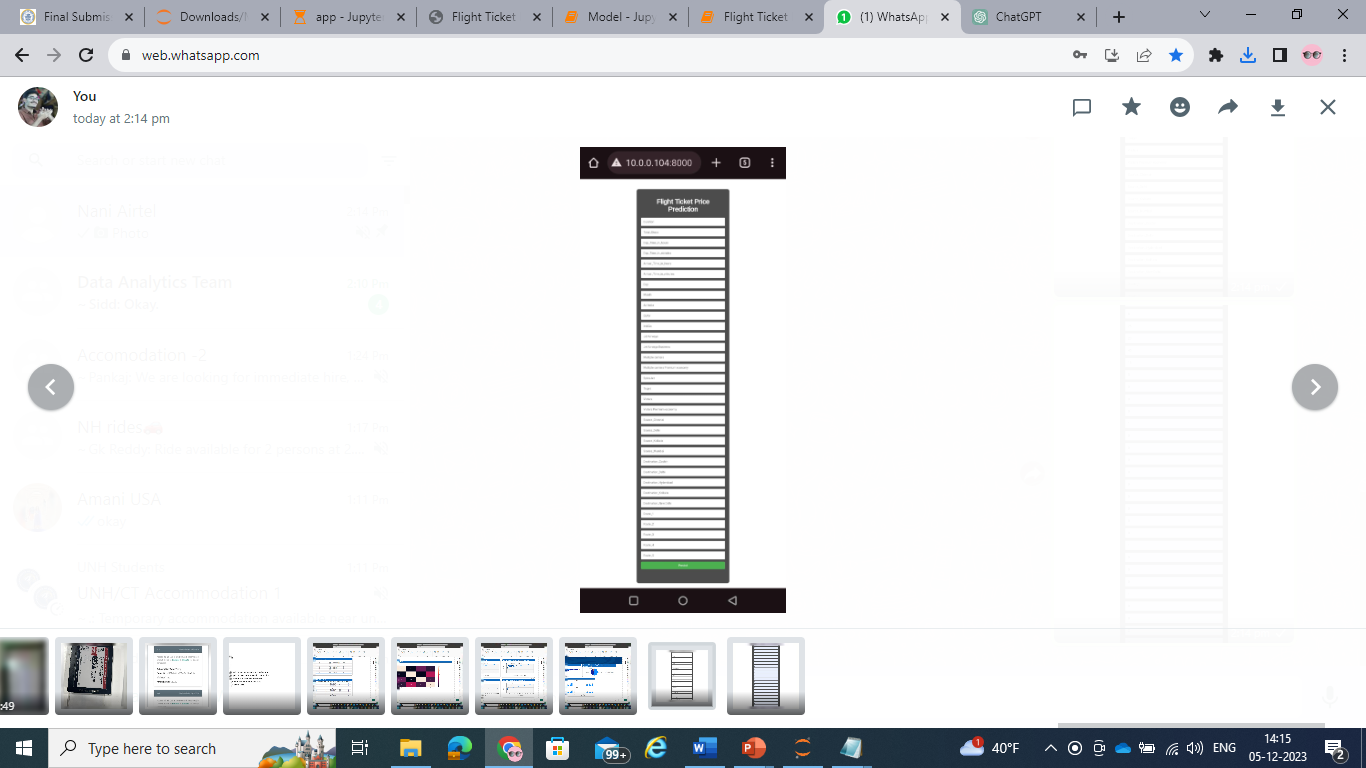
Deploying a model involves making it available for use in a production environment where it can generate predictions or perform the intended task.

Of all the models built, from evaluation it is concluded that XG Boost Regressor model performed better than all the other models. So, deploying that model will give better predictions and more business value to the end users.

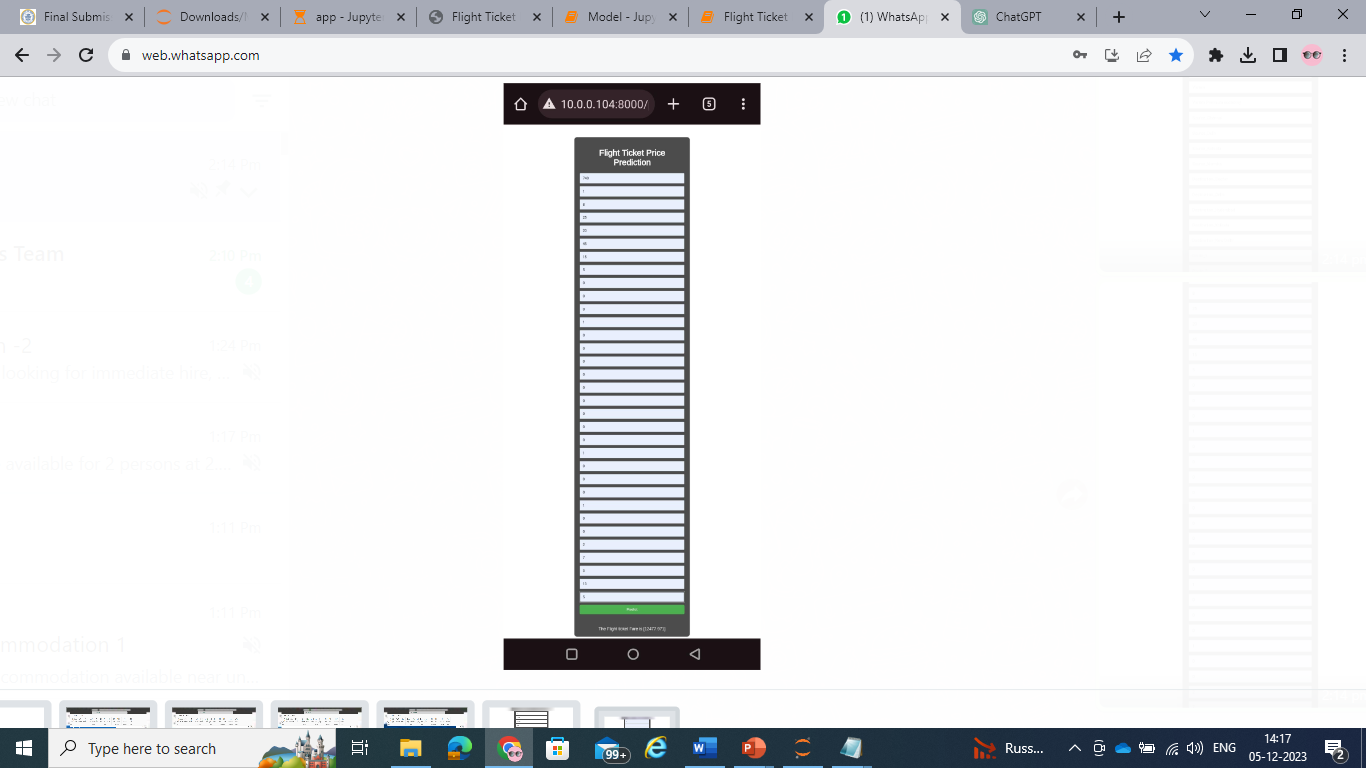
Flask is a popular web framework in Python used for building web applications, including deploying machine learning models as web services or APIs.



The http address: <http://10.0.0.104:8000/> given by the flask environment is where the end users can give the input to get the output.



Feeding inputs on the page will give the predicted output. In this case, the output is the flight ticket price in Indian rupees.



**INFERENCE**

The XG Boost Machine Learning model gave better performance than the other regressor models. After deploying the model using Flask web server, the model gave required outputs and met all the business requirements. This model can be used to get the flight ticket price by providing the inputs to it through the flask web application.

**FUTURE SCOPE**

The regressor model performance can be evaluated with other performance metrices.

A Custom ensemble model (Super Learner) model can be built using multiple models to improve the performance of the model and get better results from it.

The model is confined to only domestic flights in India. Models can be built to predict ticket prices for international flights so that when deployed, the model can be used by international airlines and passengers.

**REFERENCES**

* <https://youtu.be/WjLou_0YVMo?si=O_MsRoI5q4ooTXeC>
* <https://youtu.be/46bRwRvW0PI?si=KIchjcJvDOZkfBMq>
* <https://youtu.be/KFPX0FLhscs?si=VSFVpvoKYssh3epv>
* <https://www.kaggle.com/datasets/nikhilmittal/flight-fare-prediction-mh/code>