

Flight Price Prediction



Submitted by:

Manoj.I.V

**ACKNOWLEDGMENT**

This includes mentioning of all the references, research papers, data sources, professionals and other resources that helped you and guided you in completion of the project.

**INTRODUCTION**

* Business Problem Framing

Describe the business problem and how this problem can be related to the real world.

Answer: The travel from one country to another has become common these days. These algorithms are required so that it would be every cost effective if a price of the flight can be predicted.

* Conceptual Background of the Domain Problem

Describe the domain related concepts that you think will be useful for better understanding of the project.

Answer: Type of Airlines, time of flight, destination, source, price of the fuel, number of stops.

* Review of Literature

This is a comprehensive summary of the research done on the topic. The review should enumerate, describe, summarize, evaluate and clarify the research done.

Answer: Data was scraped at different websites like make my trip, yatra.com, skyscanner.com, official websites of airlines. So the different costs of the airlines were taken into consideration.

* Motivation for the Problem Undertaken

Describe your objective behind to make this project, this domain and what the motivation is behind.

Answer: The objective is to build economic model which can predict cost of the flight. This model is to save money during travel for far places.

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

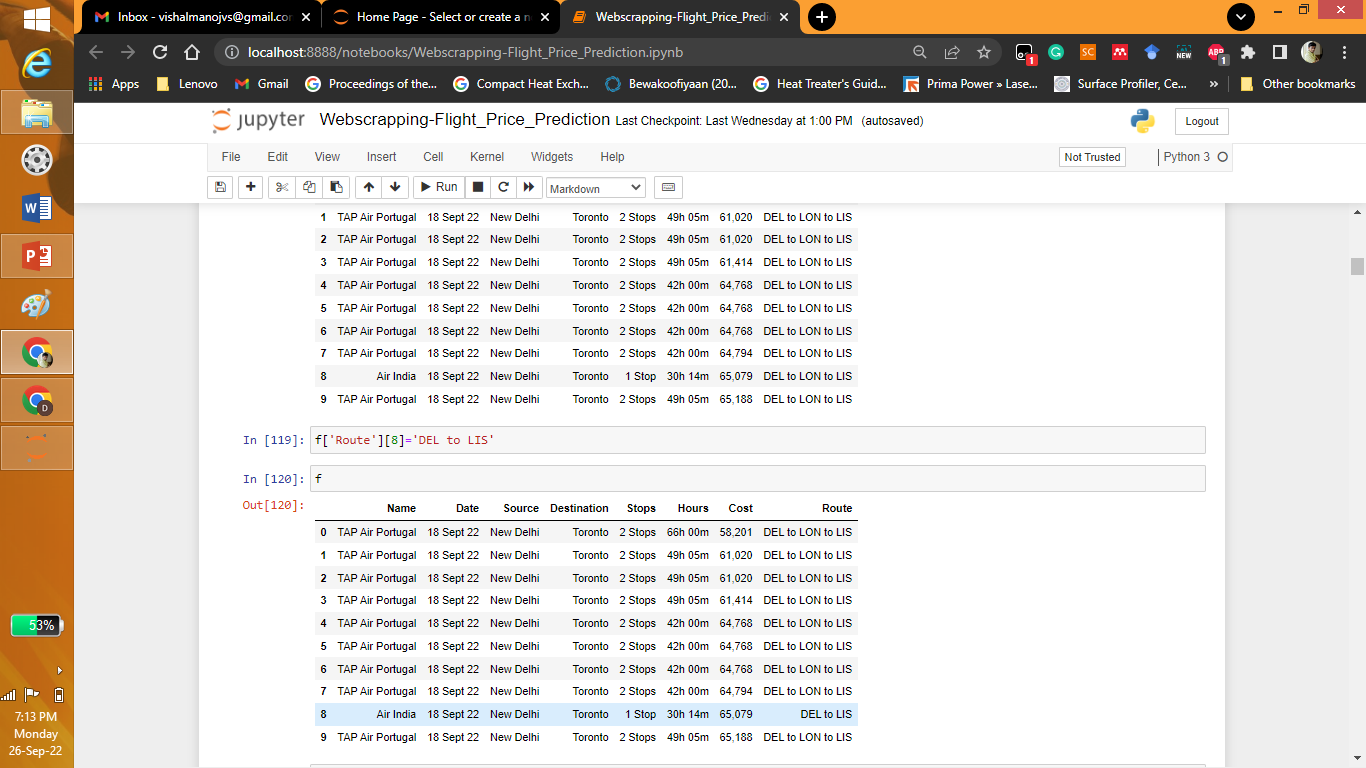
Describe the mathematical, statistical and analytics modelling done during this project along with the proper justification.

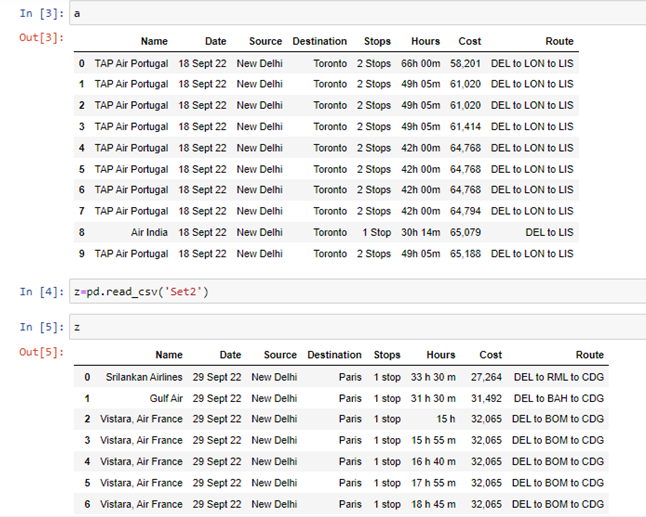
Answer: Different models used in the project are linear regression, logistic regression, Random Forest Regressor, XGBRegressor, AdaBoostRegressor, KNeighborsRegressor, SVR, Gradient Boosting Regressor.

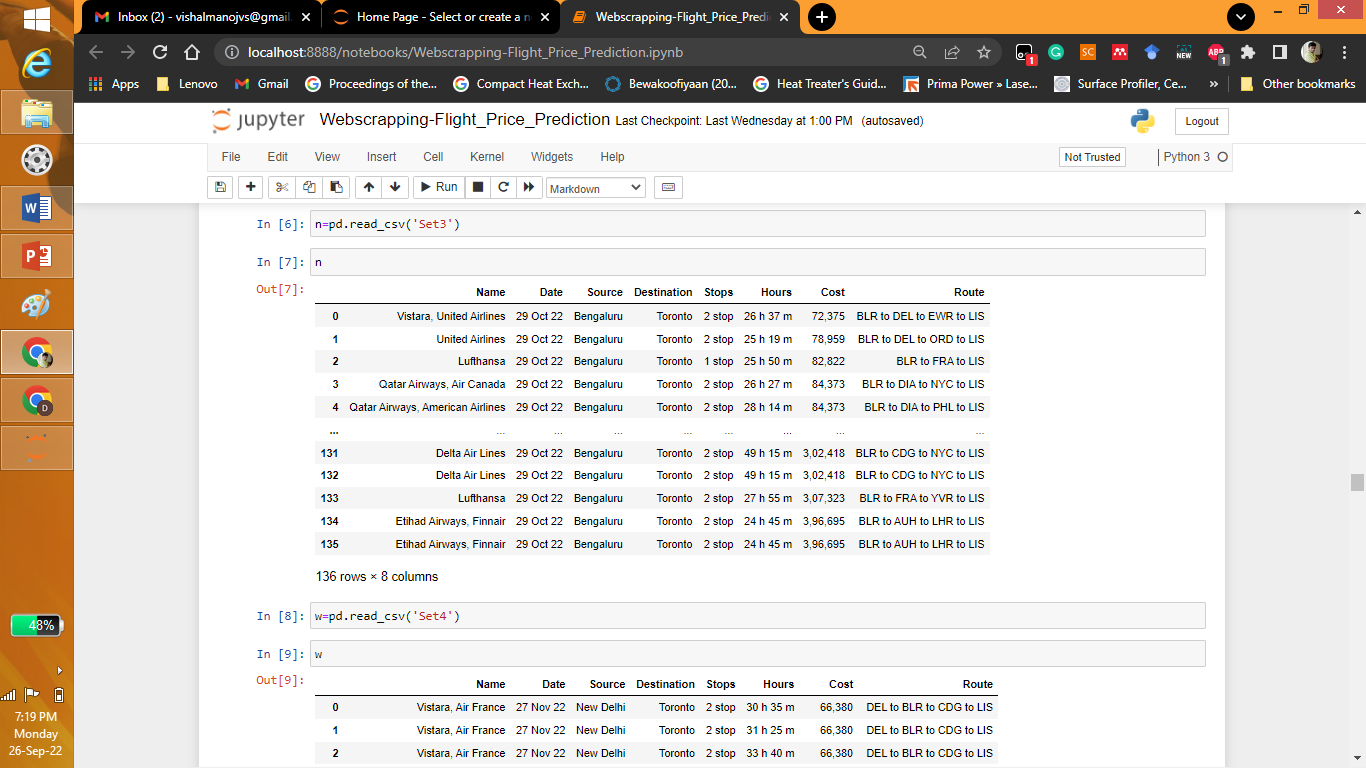
* Data Sources and their formats

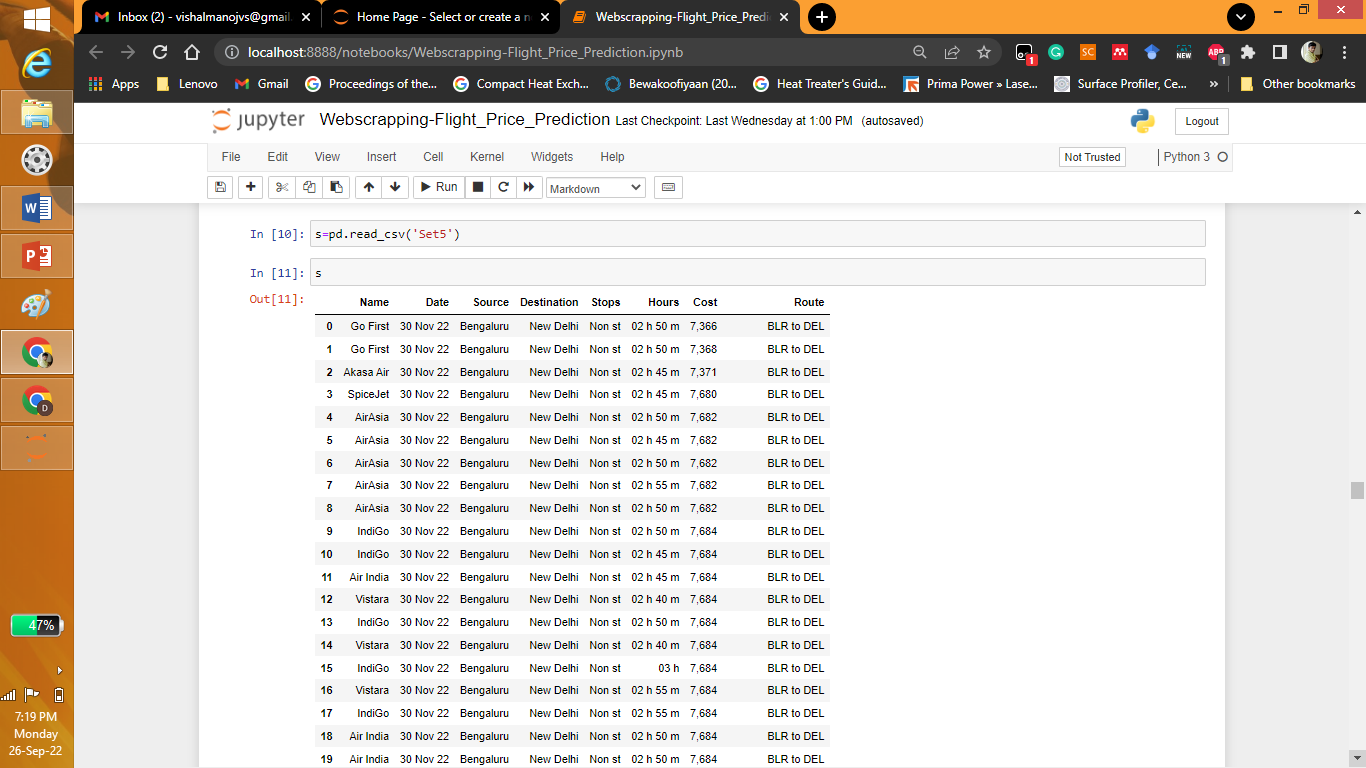
What are the data sources, their origins, their formats and other details that you find necessary? They can be described here. Provide a proper data description. You can also add a snapshot of the data.

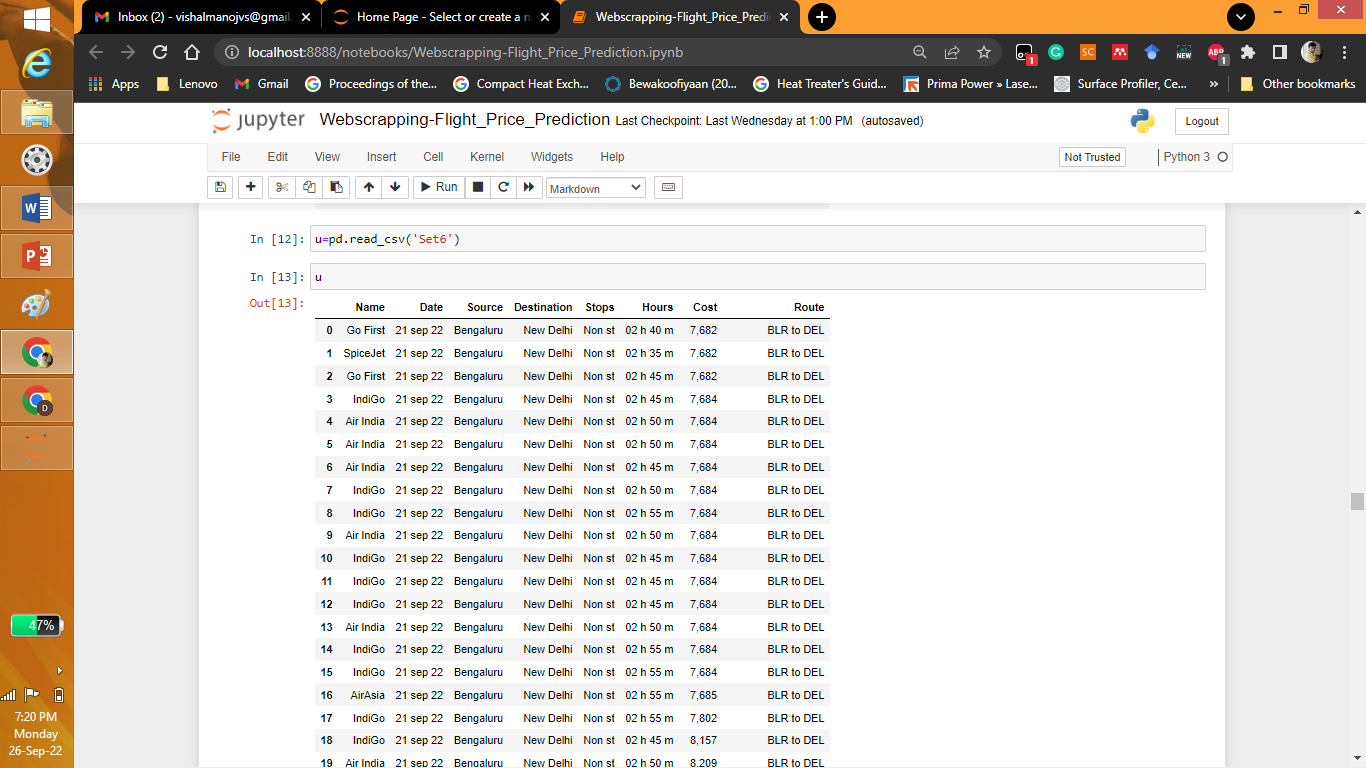
Answer: Data was scraped at different websites like make my trip, yatra.com, skyscanner.com, official websites of airlines.



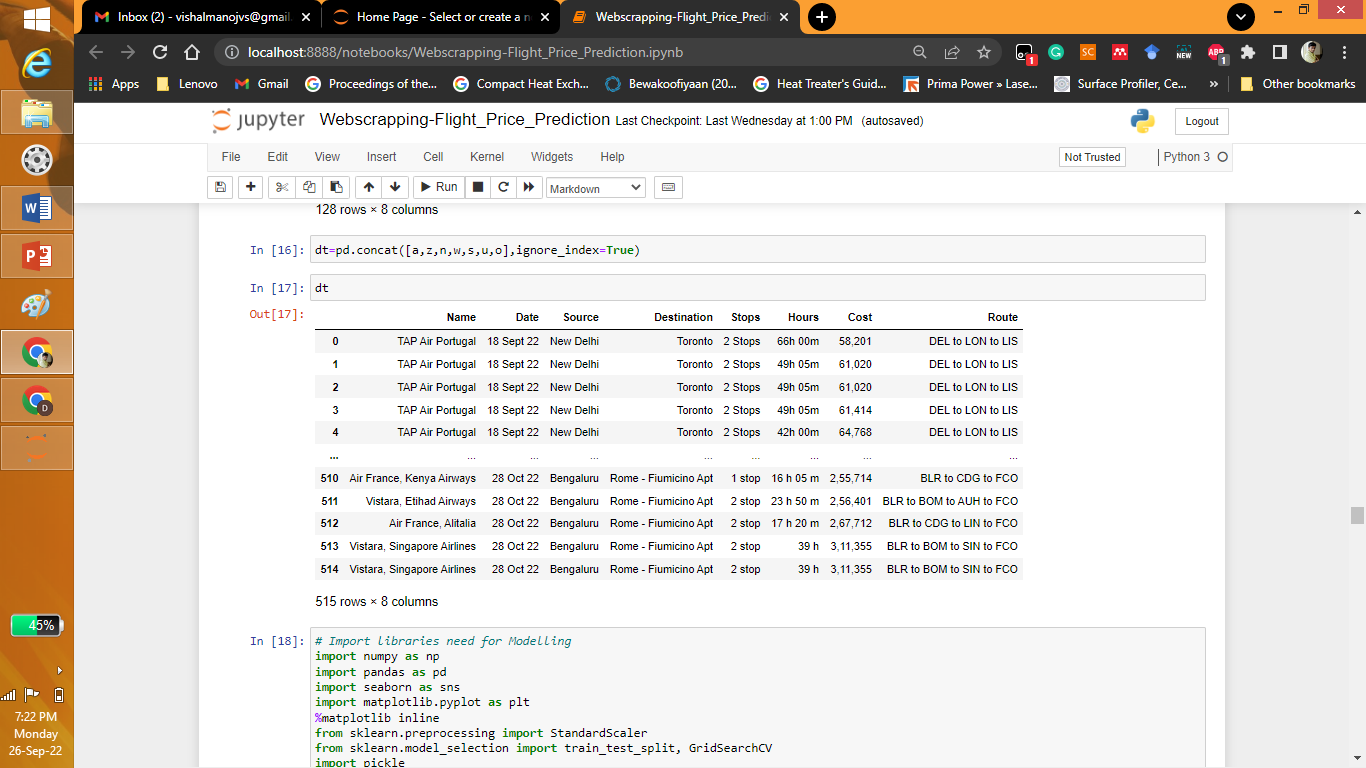


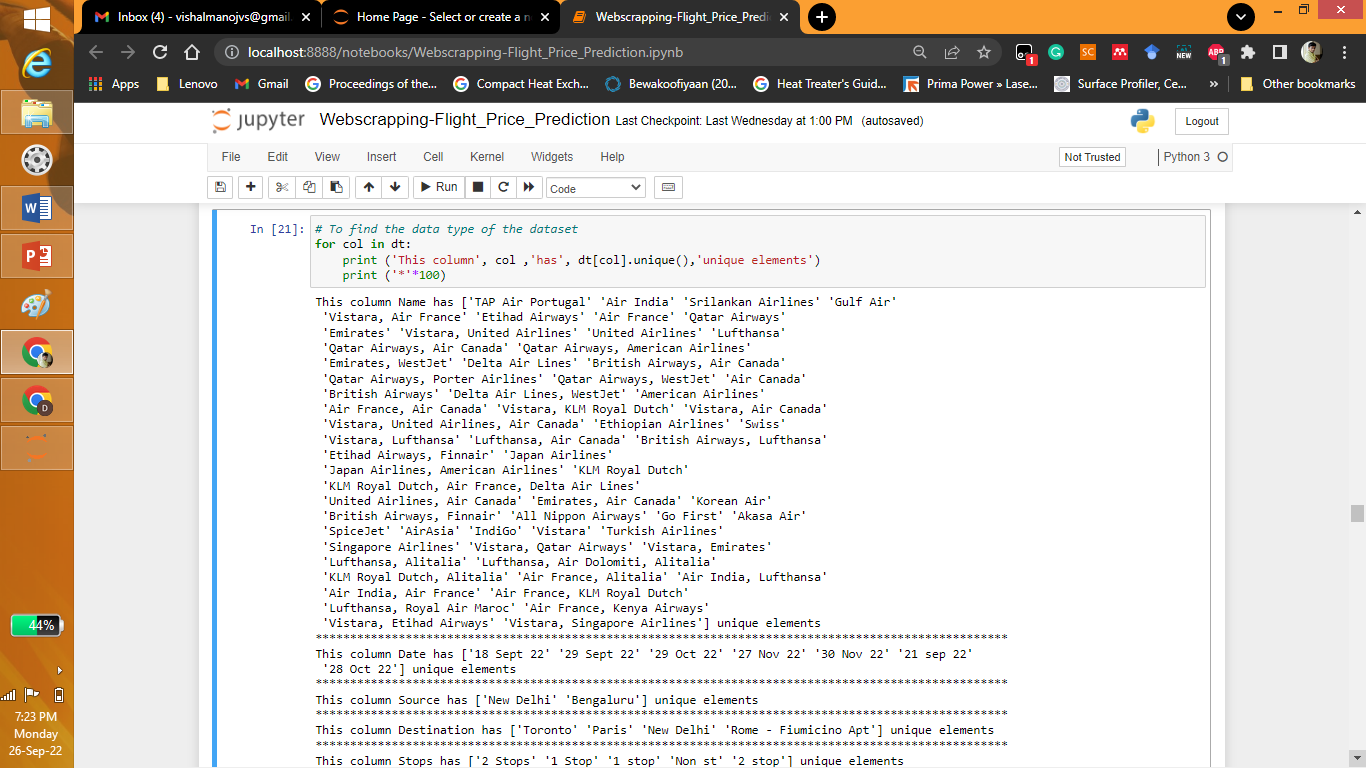


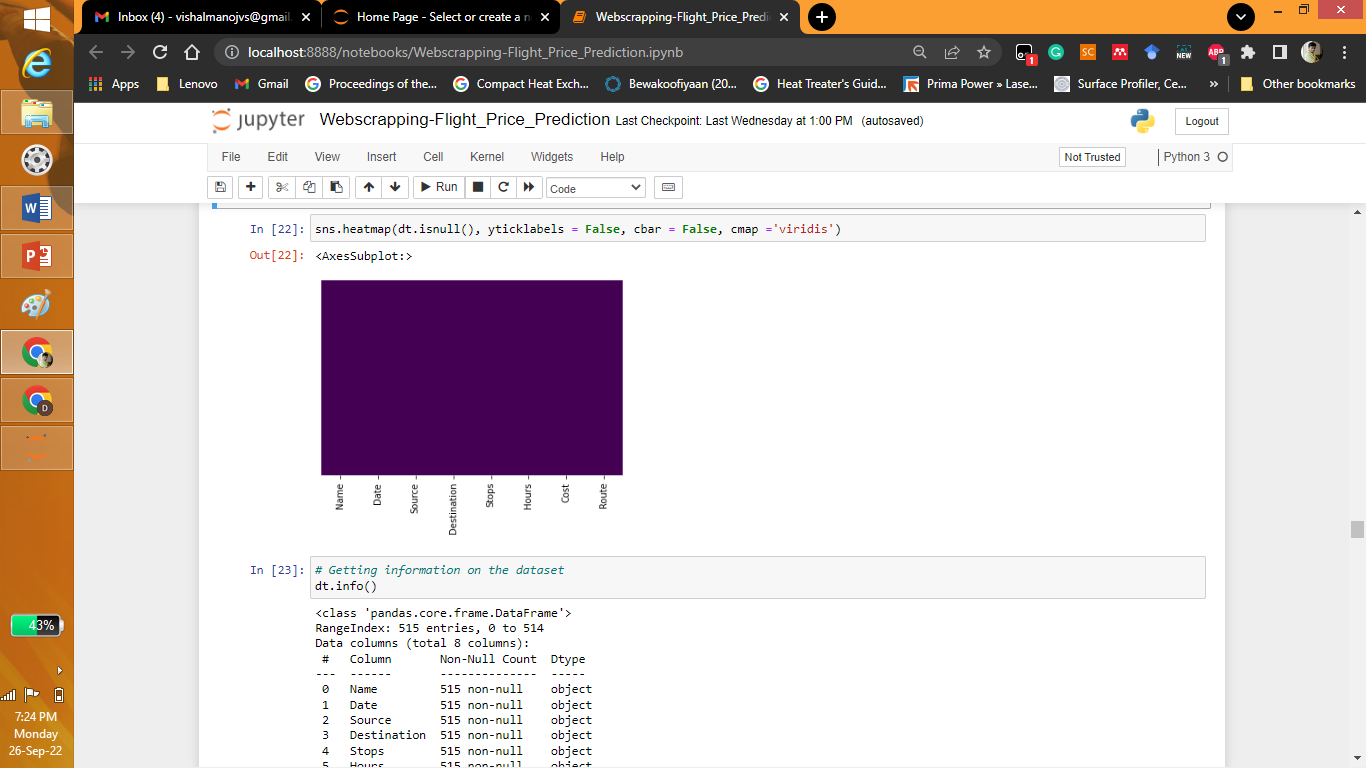








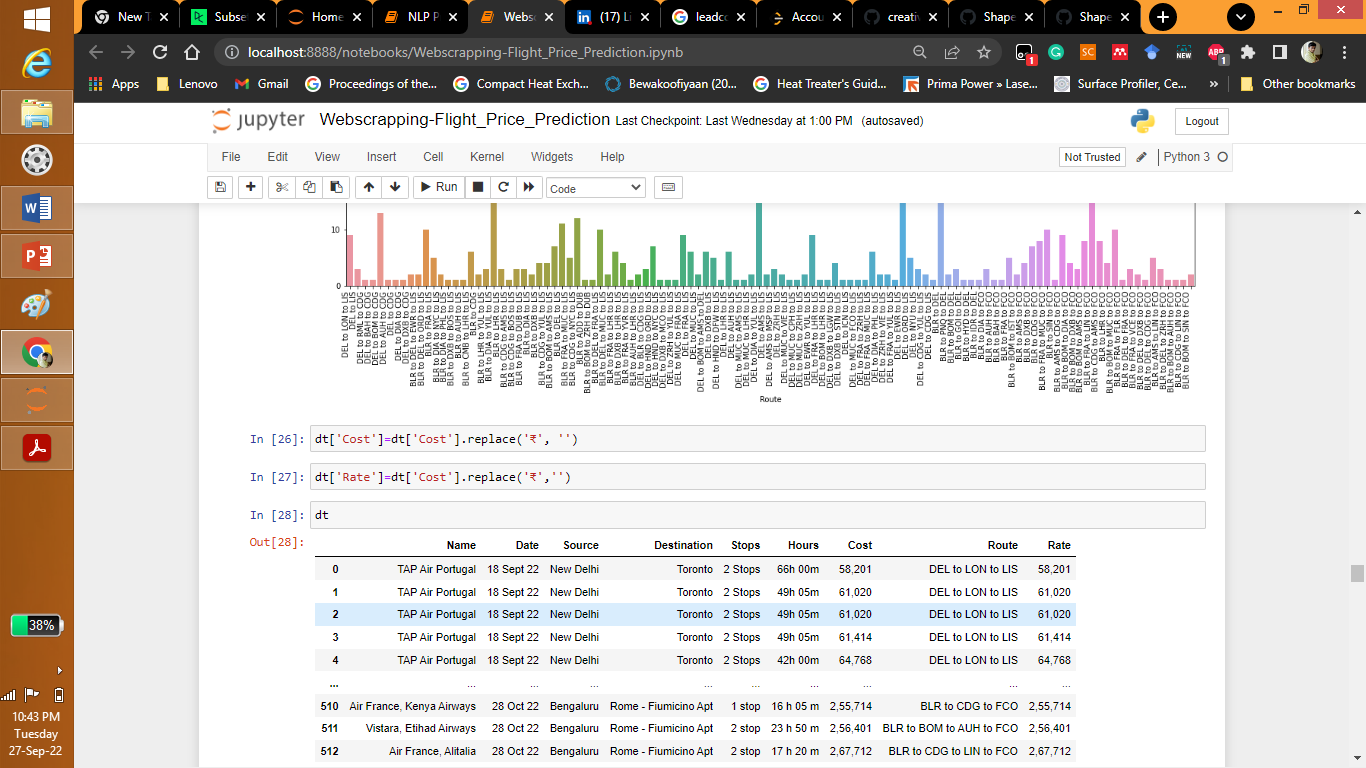


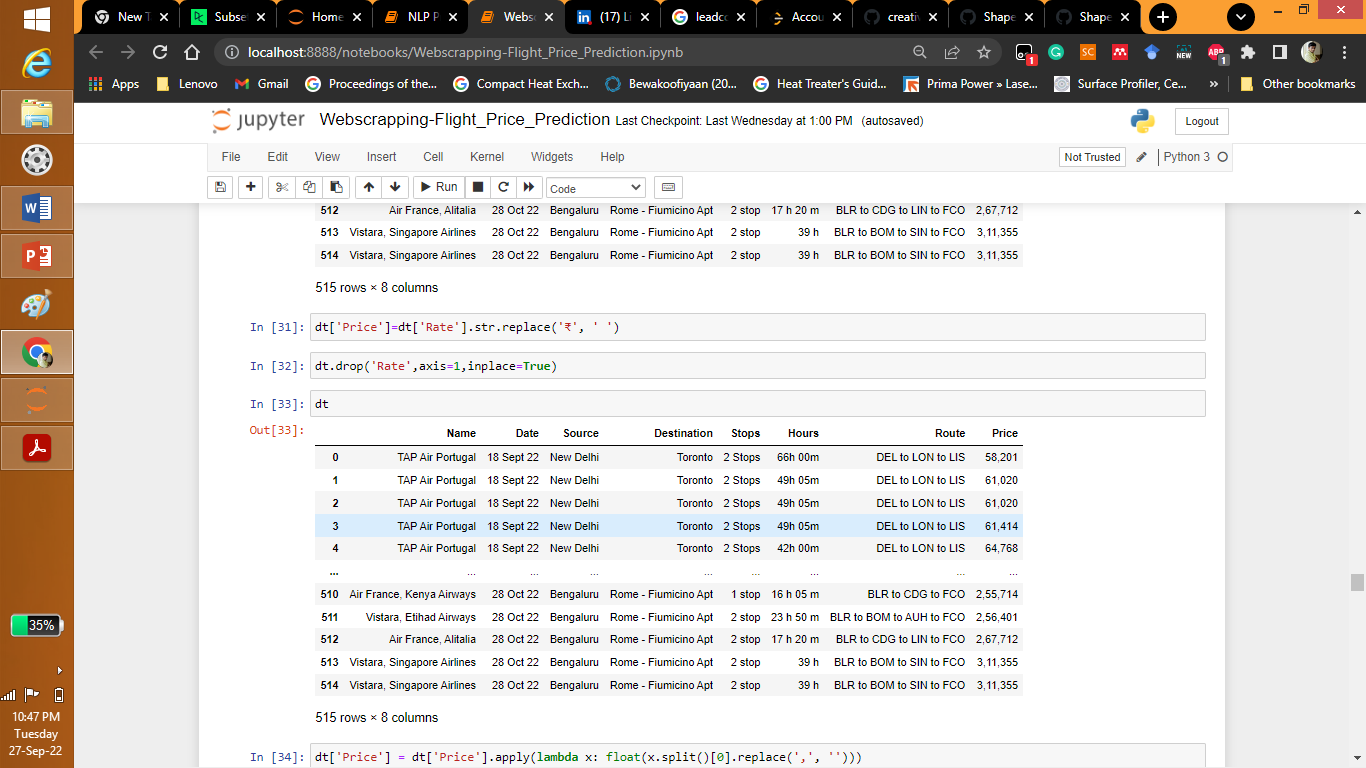


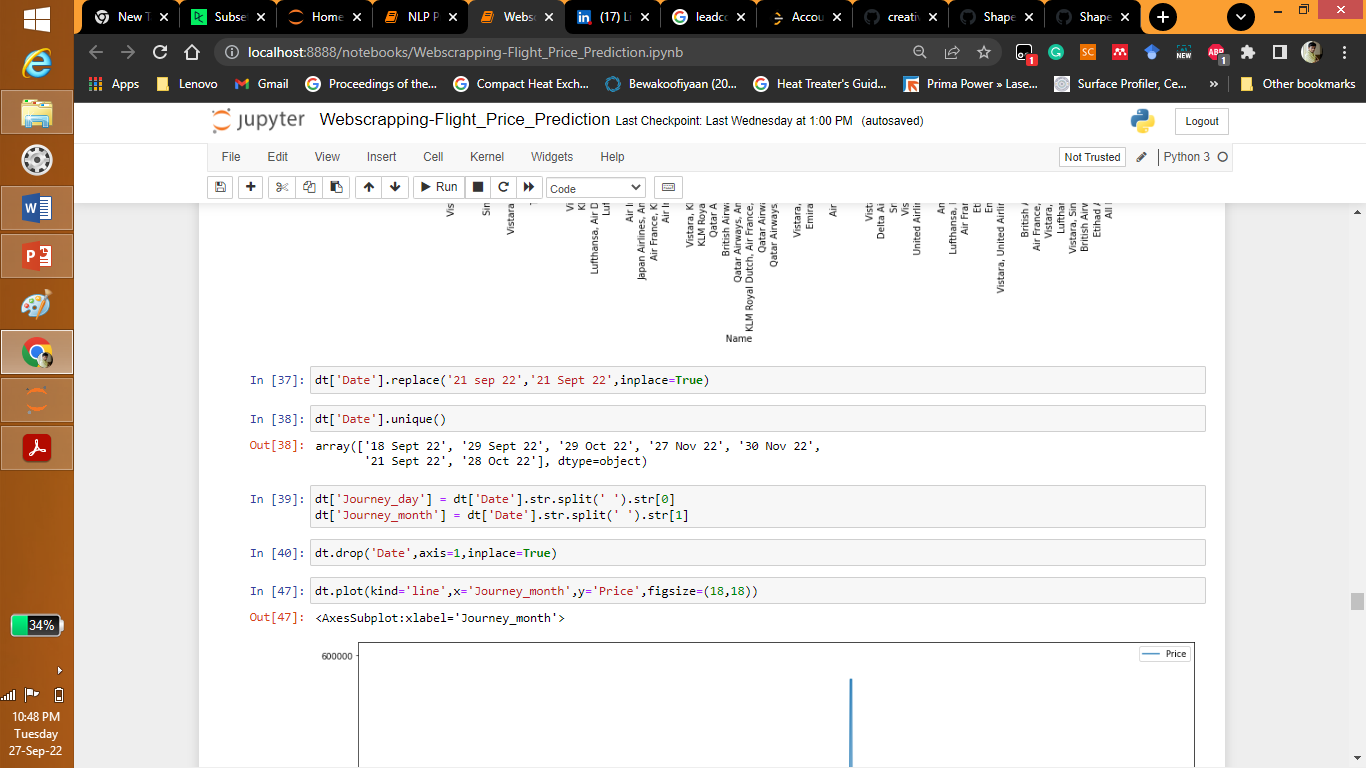
* Data Preprocessing Done

What were the steps followed for the cleaning of the data? What were the assumptions done and what were the next actions steps over that?

Answer: Converting data formats from object to float etc. Removing junk data.







* Data Inputs- Logic- Output Relationships

Describe the relationship behind the data input, its format, the logic in between and the output. Describe how the input affects the output.

Answer: Normally we use corr(), describe() to get better relation between input and output data. Basically every website considers the following Type of Airlines, time of flight, destination, source, price of the fuel, number of stops for the decision of the price.

* State the set of assumptions (if any) related to the problem under consideration

Here, you can describe any presumptions taken by you.

Answer: As Type of Airlines, time of flight, destination, source, price of the fuel, number of stops for the decision of the price in almost all the websites it is also taken to be input variable for the prediction.

* Hardware and Software Requirements and Tools Used

Listing down the hardware and software requirements along with the tools, libraries and packages used. Describe all the software tools used along with a detailed description of tasks done with those tools.

Answer: To analyse the present project we have used python libraries like numpy, pandas, seaborn, matplotlib, sklearn etc.

Numpy and pandas for converting the data to data frame, datacleaning etc.

Seaborn and matplotlib for explorative data analysis.

Sklearn for models, trainsplit and training etc.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

Describe the approaches you followed, both statistical and analytical, for solving of this problem.

Answer: We have used different models from data provided from the website. Data was cleaned and trained by the model.

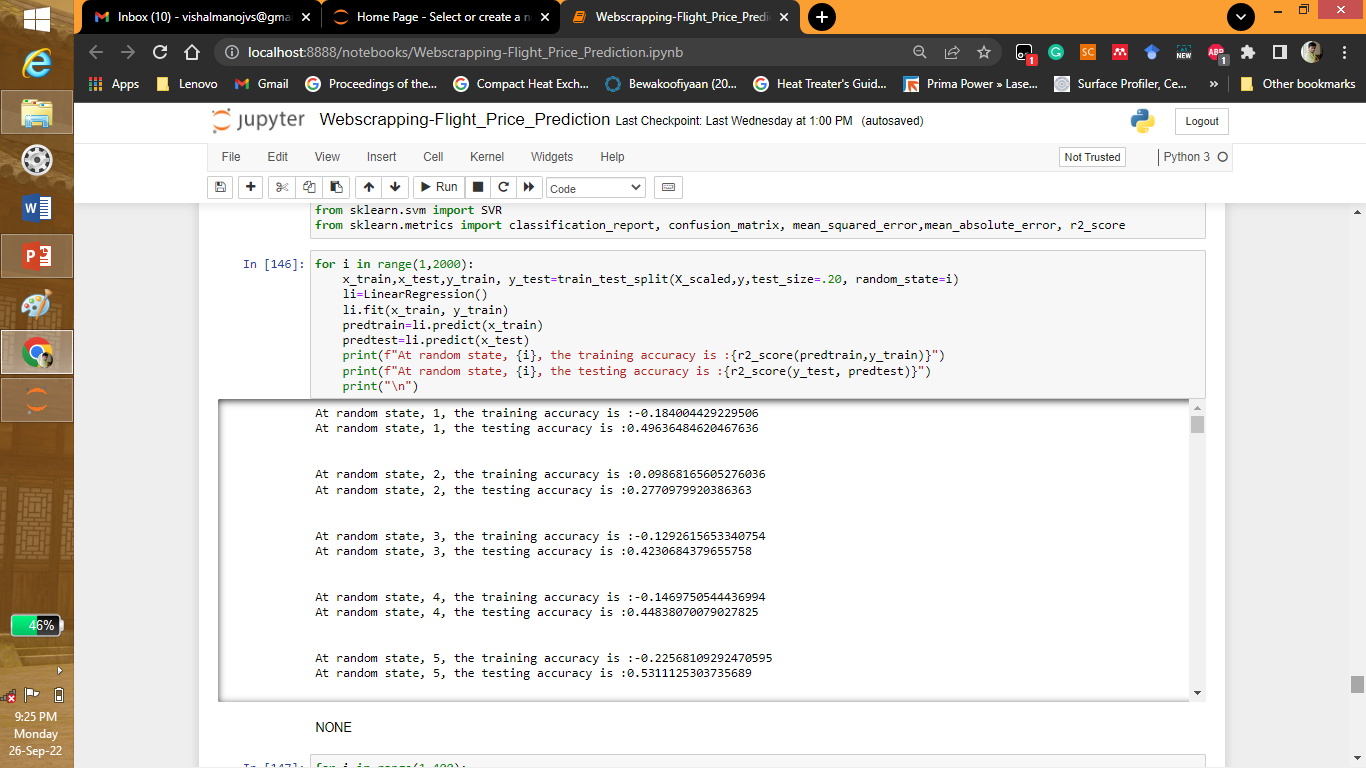
* Testing of Identified Approaches (Algorithms)

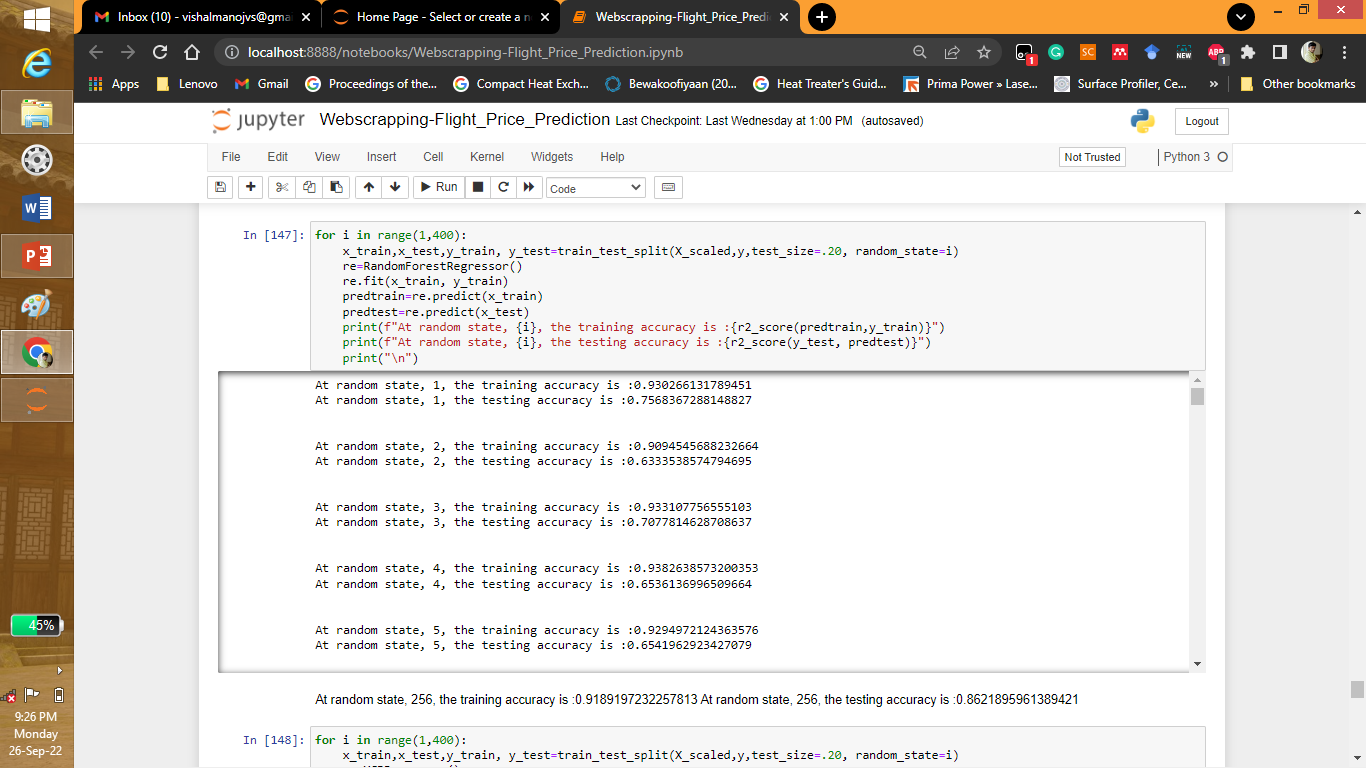
Listing down all the algorithms used for the training and testing.

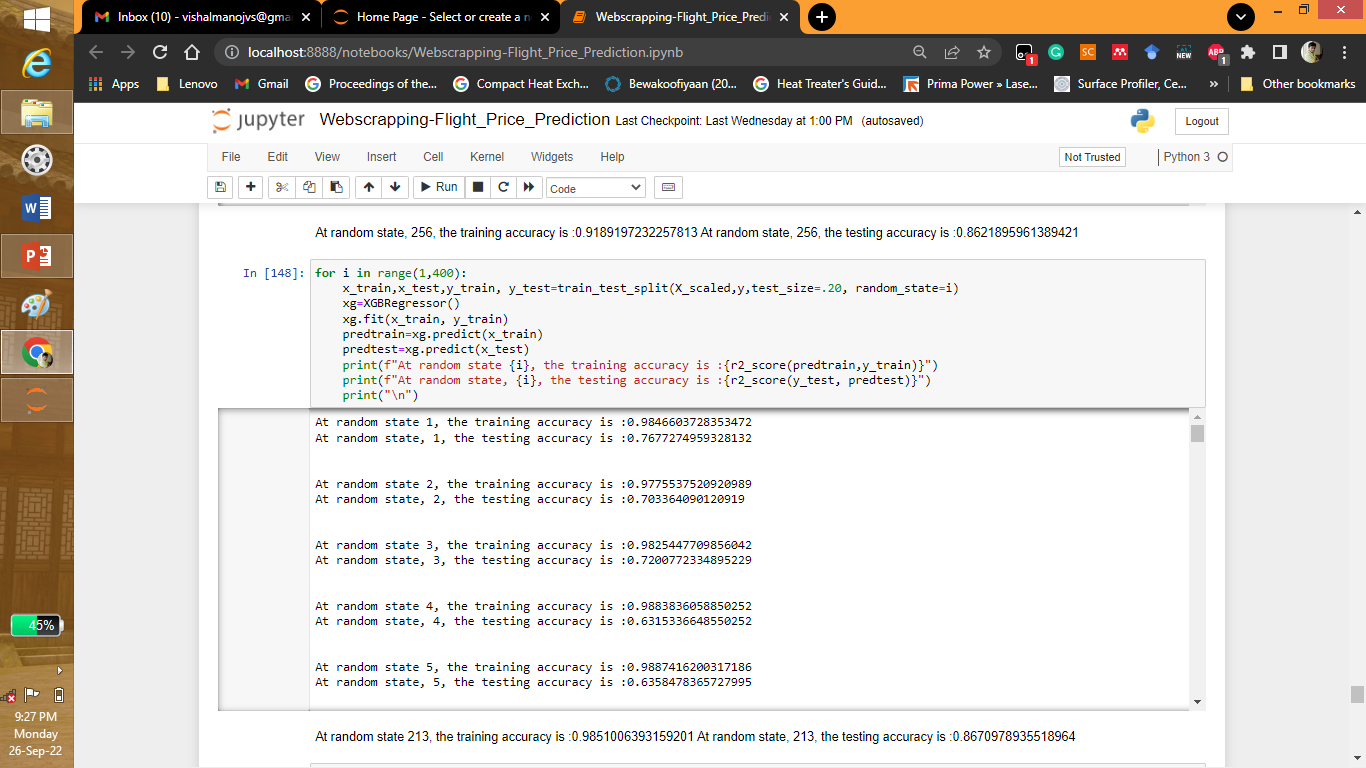
Answer: : Different models used in the project are linear regression, logistic regression, Random Forest Regressor, XGBRegressor, AdaBoostRegressor, KNeighborsRegressor, SVR, Gradient Boosting Regressor.

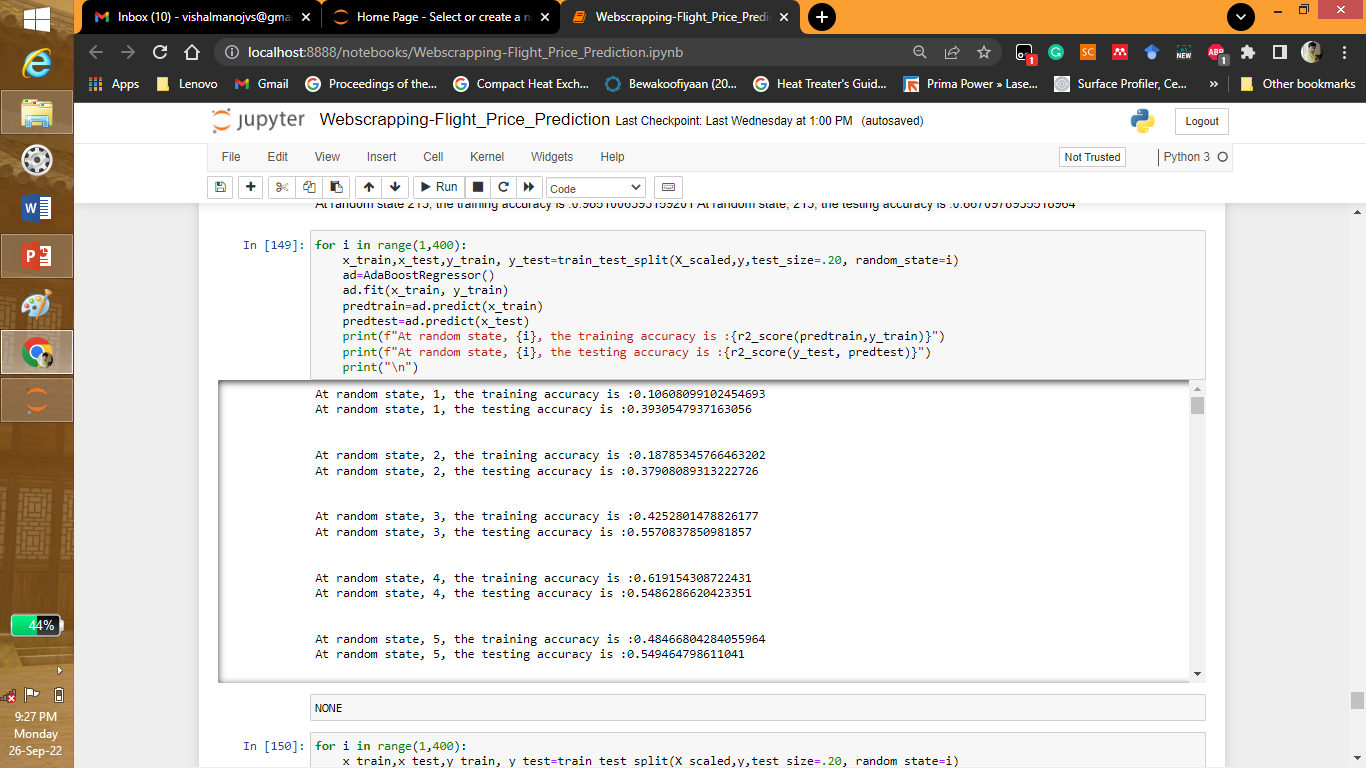
* Run and Evaluate selected models

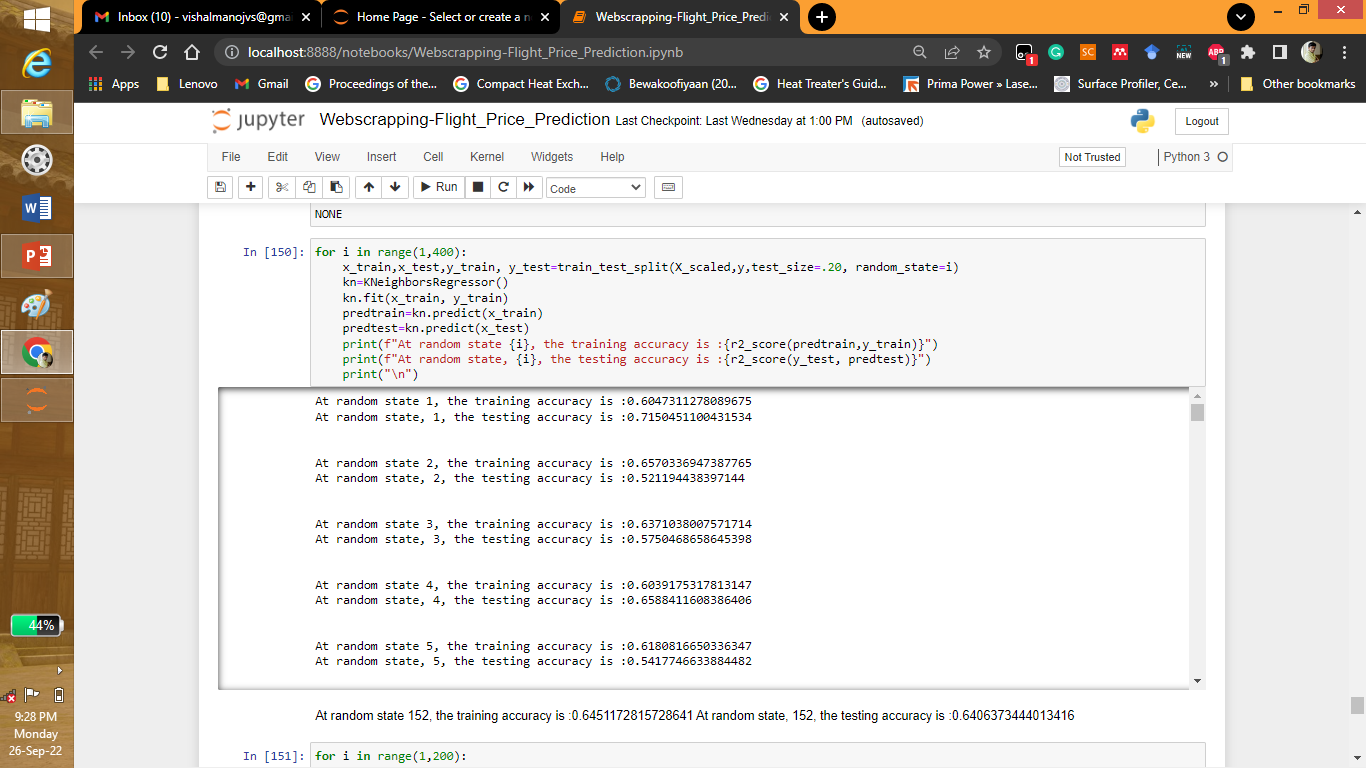
Describe all the algorithms used along with the snapshot of their code and what were the results observed over different evaluation metrics.

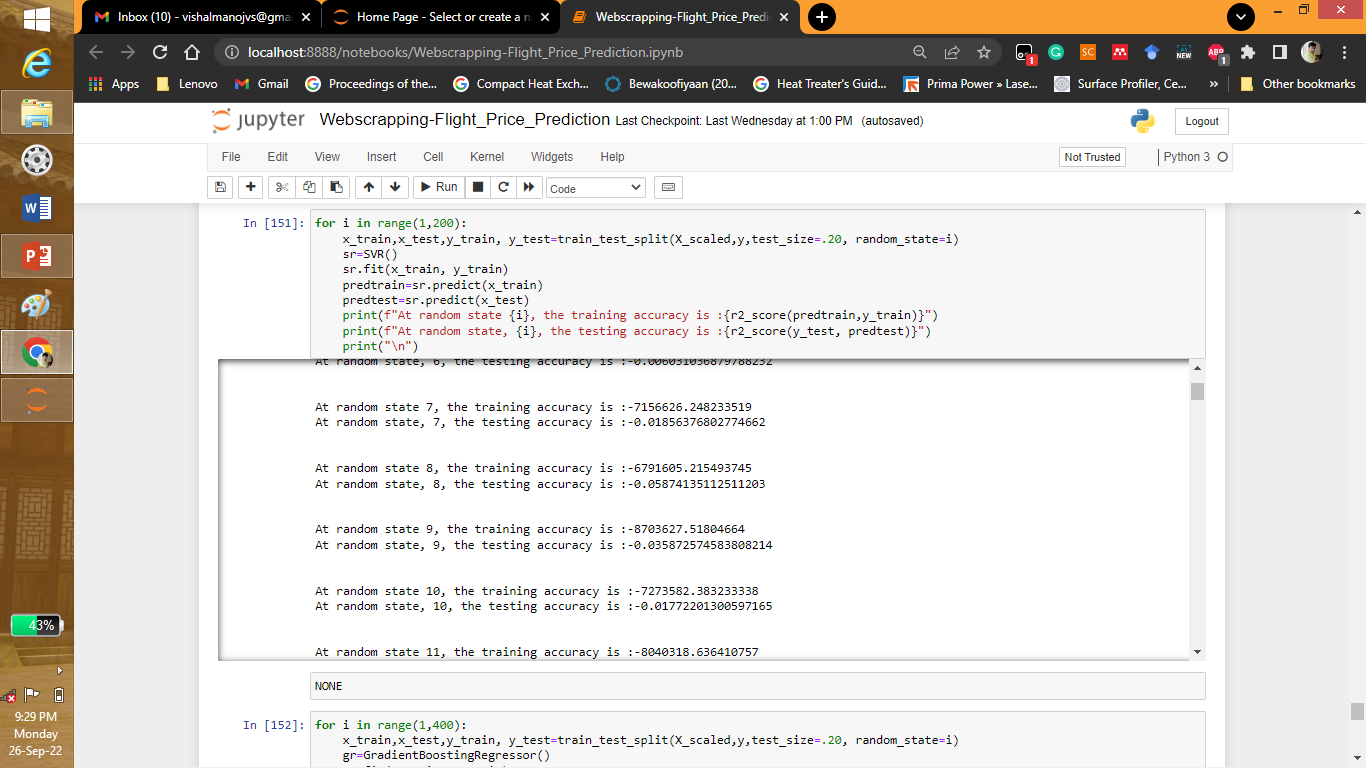


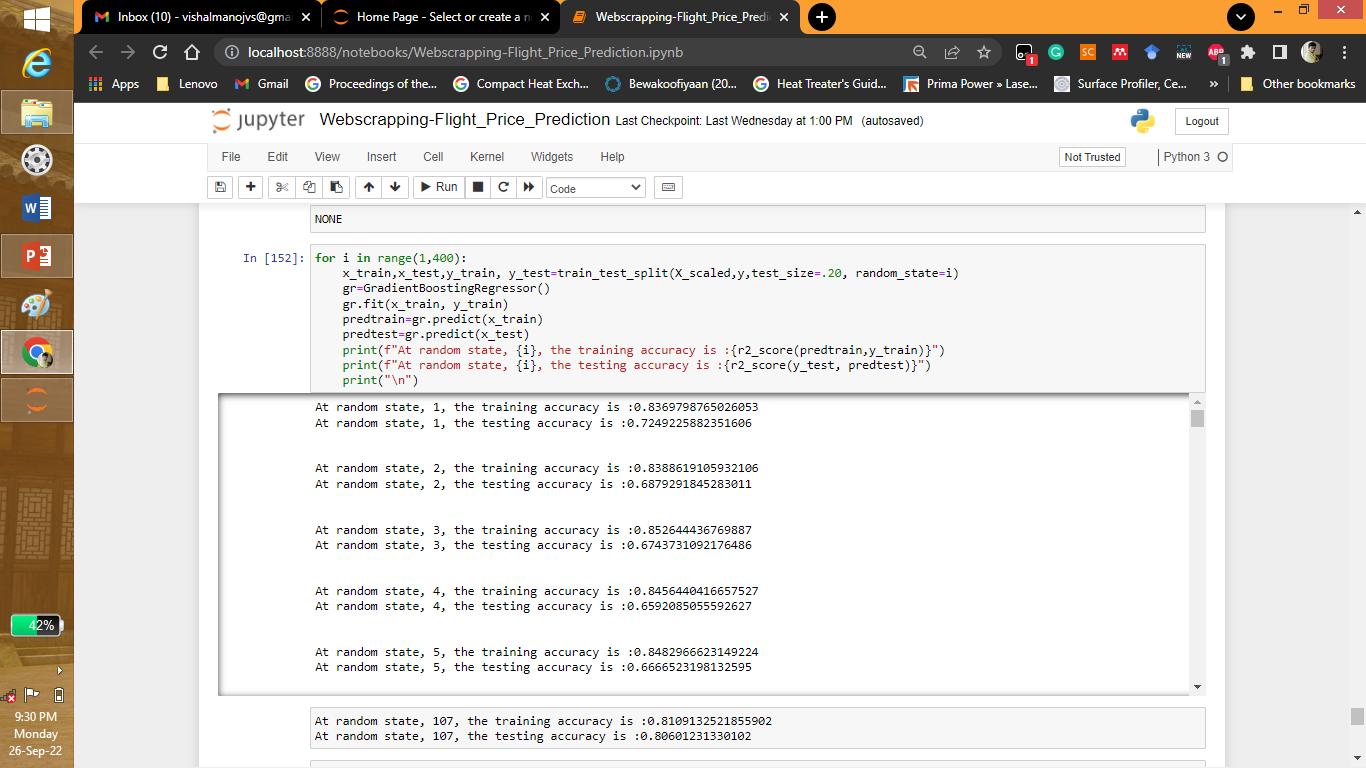


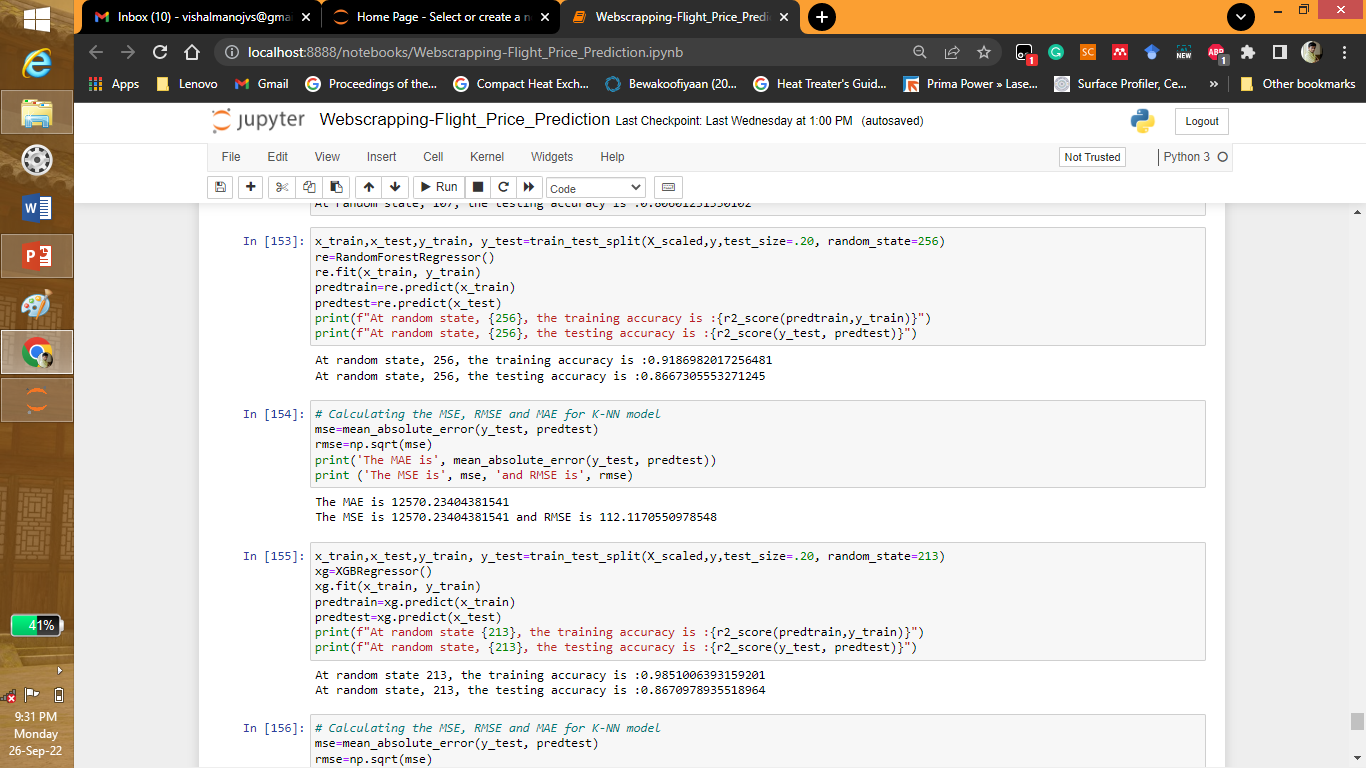


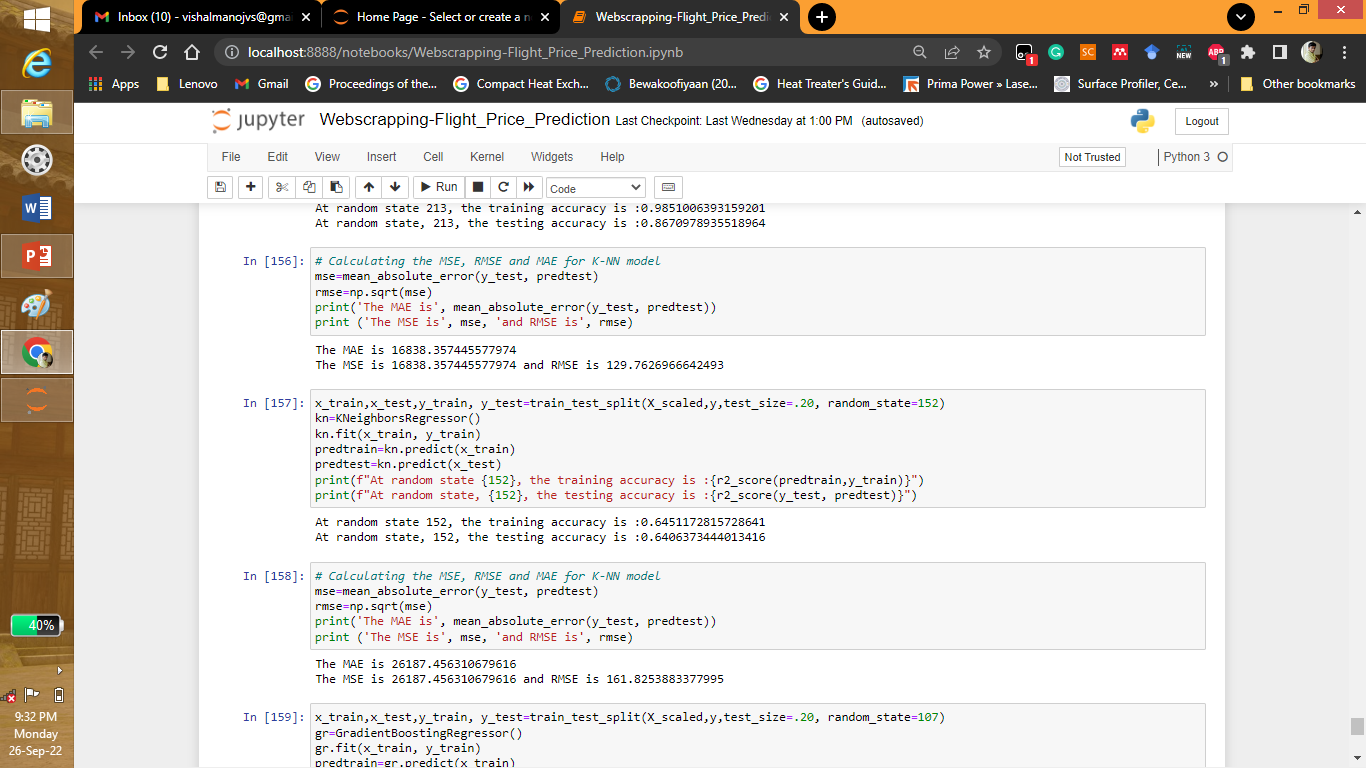


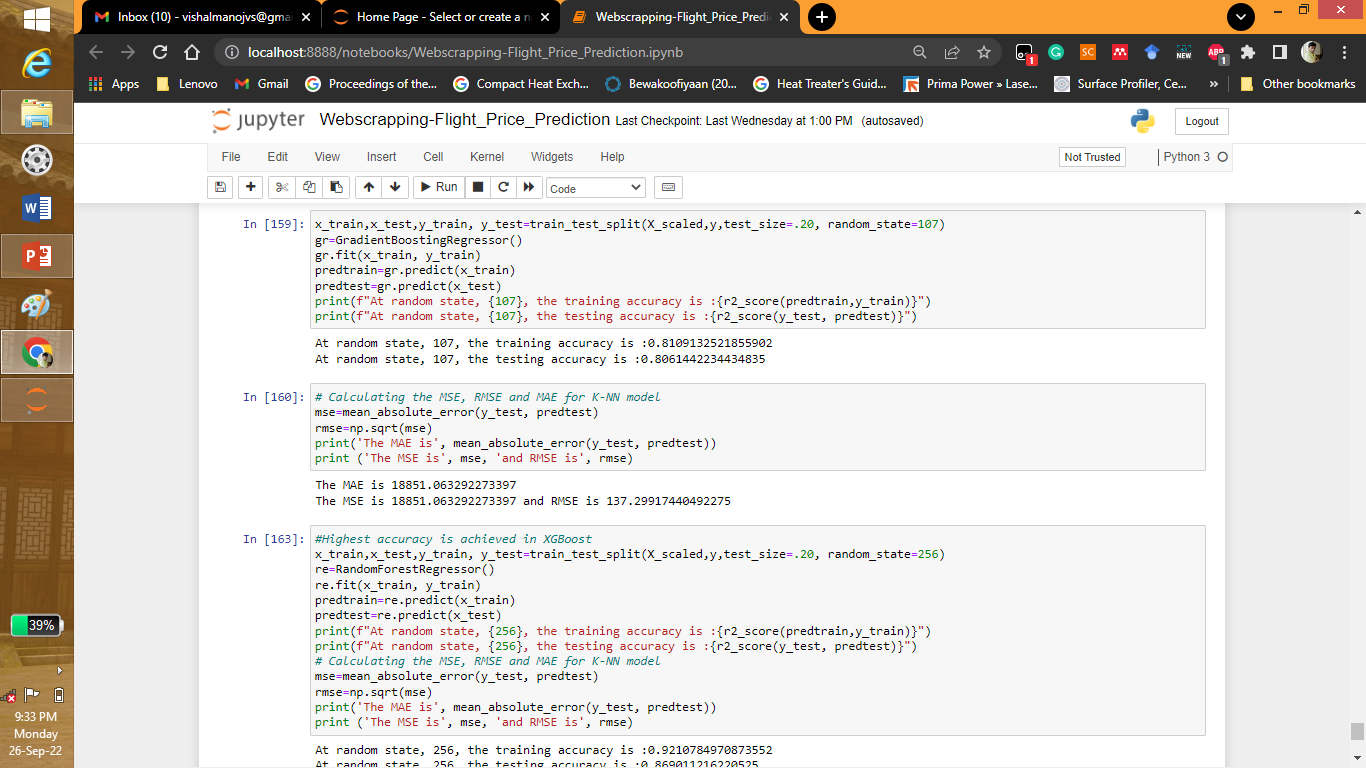


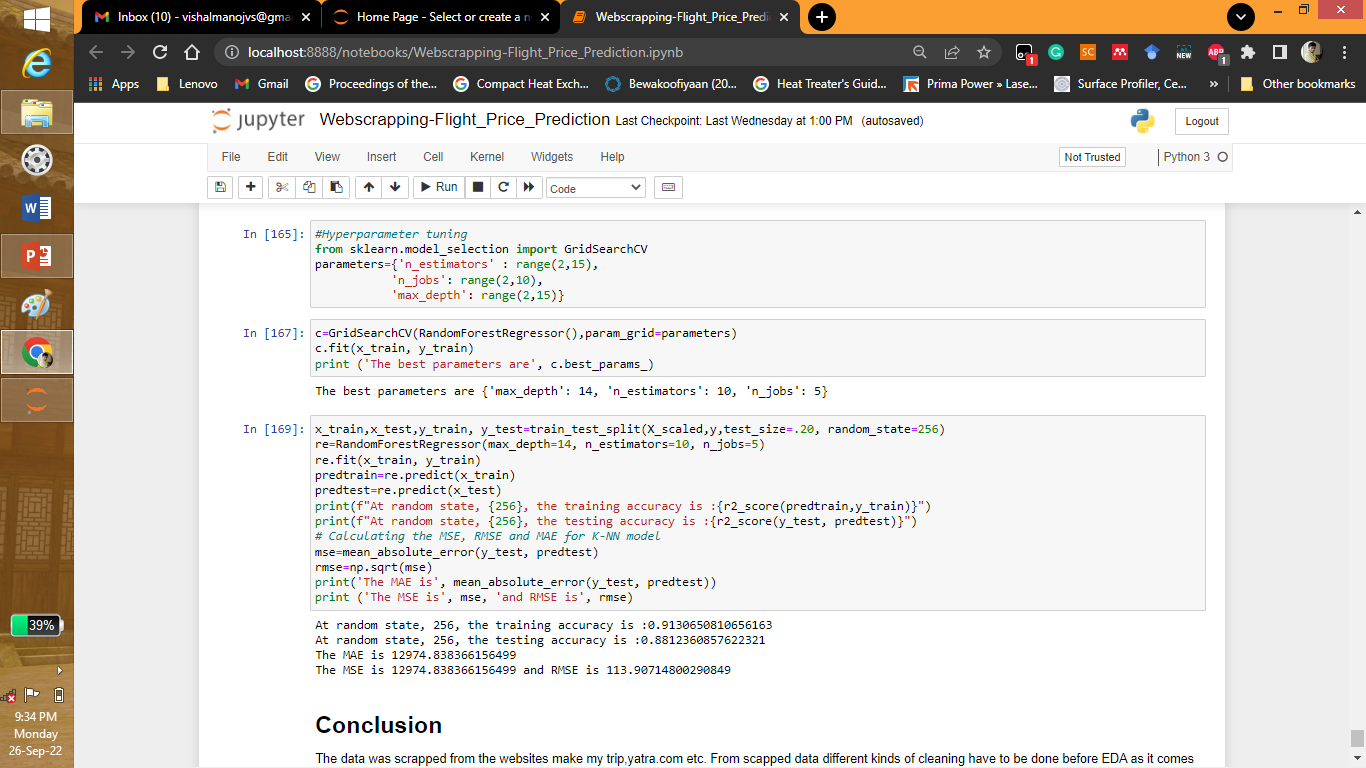












* Key Metrics for success in solving problem under consideration

What were the key metrics used along with justification for using it? You may also include statistical metrics used if any.

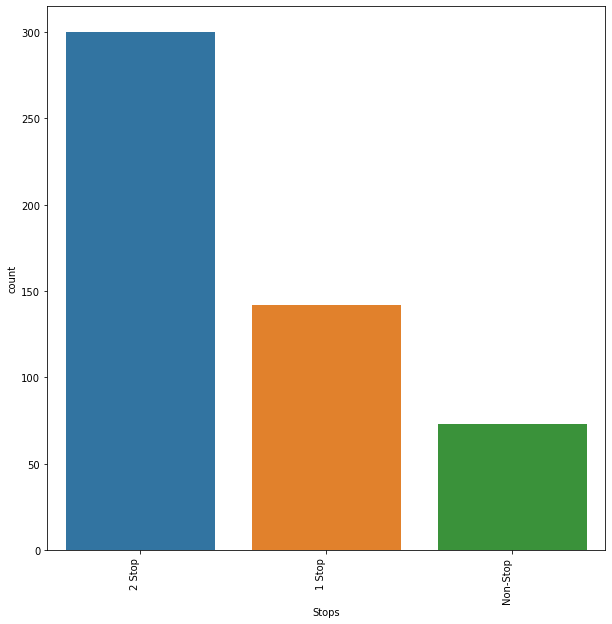
Answer: The metrics used were The R2score, MAE, MSE and RMSE.

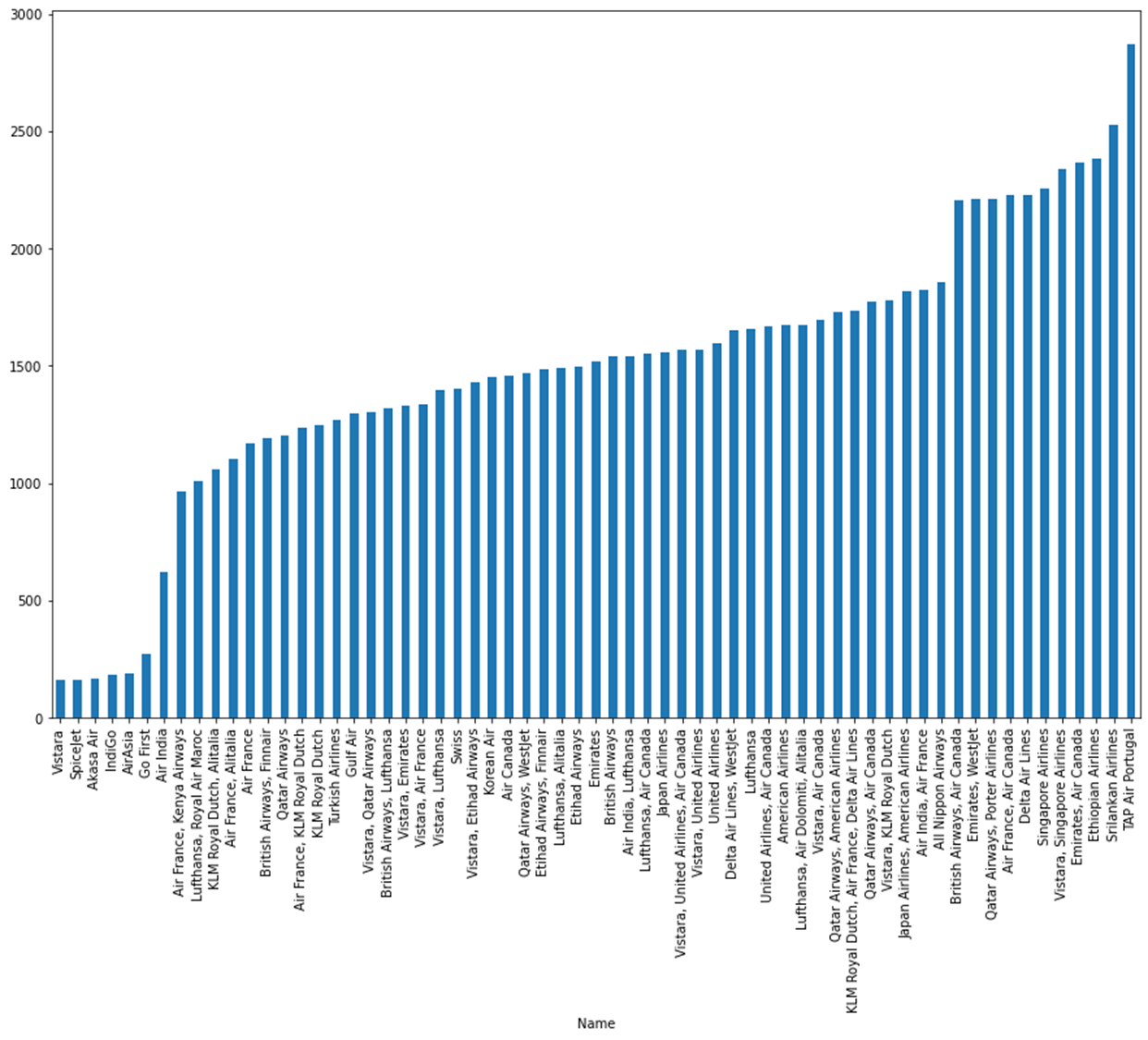
* Visualizations

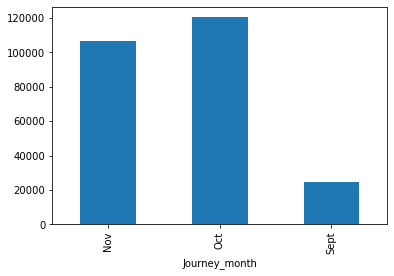
Mention all the plots made along with their pictures and what were the inferences and observations obtained from those. Describe them in detail.

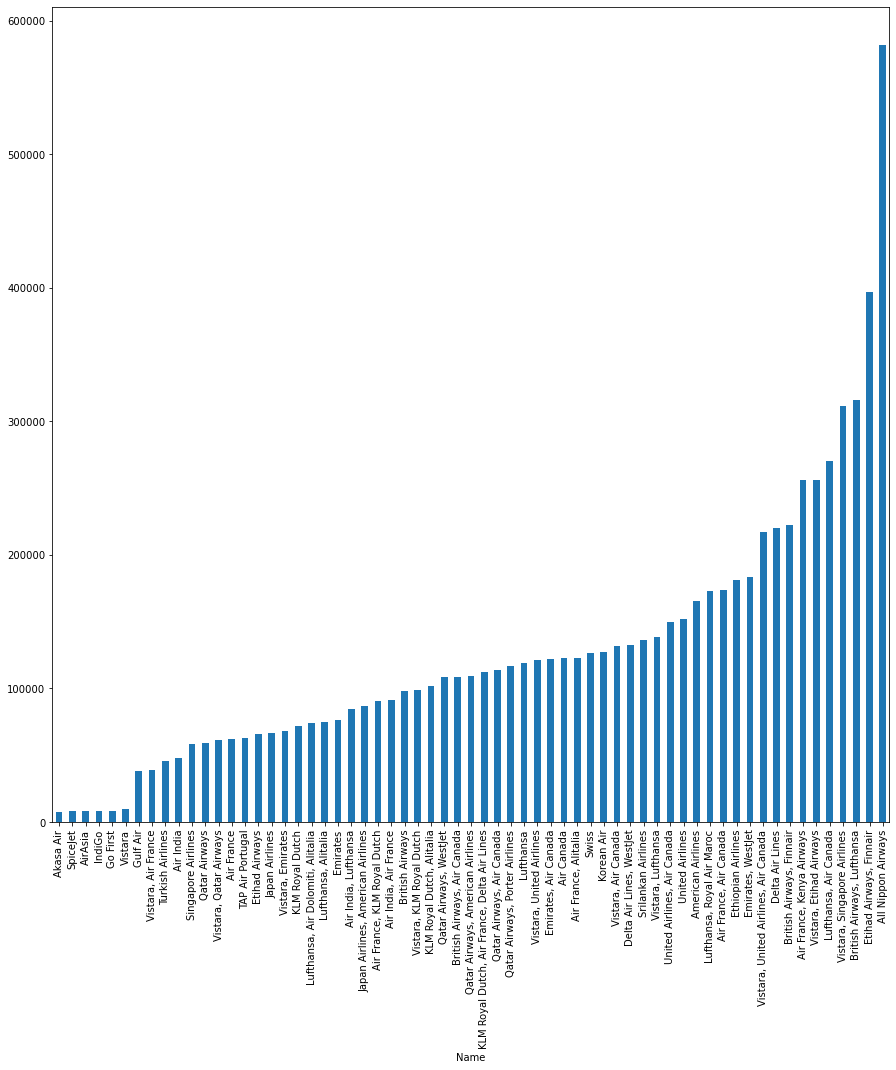
If different platforms were used, mention that as well.

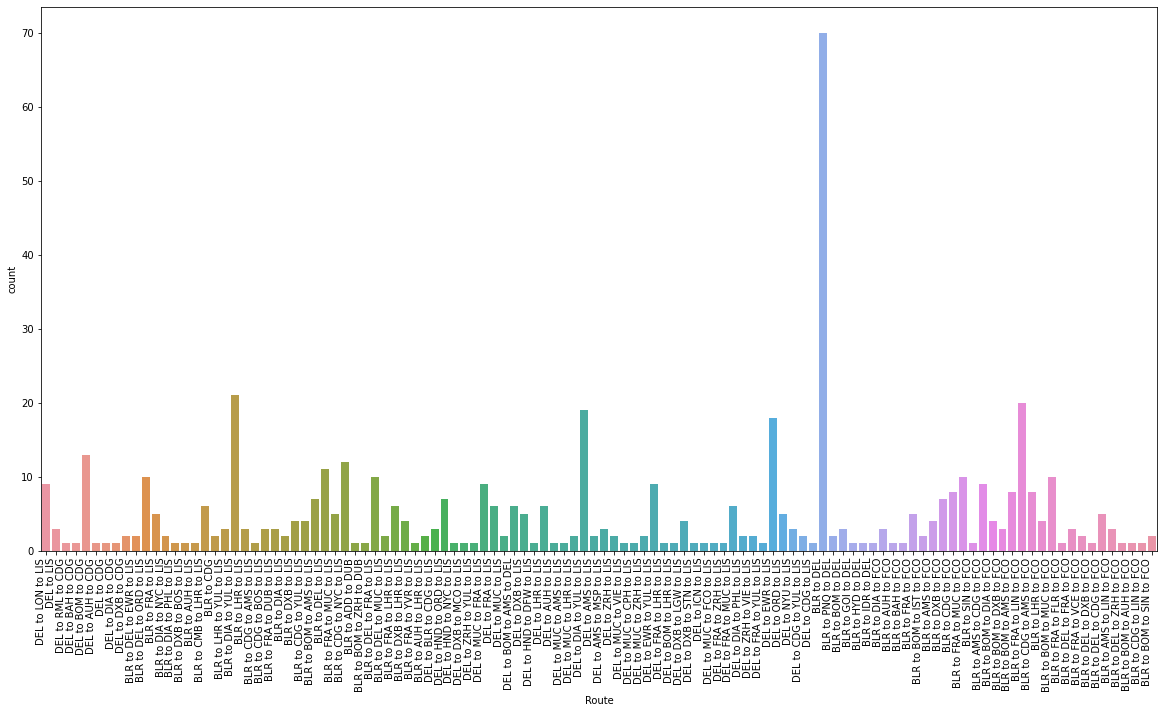
Answer: The different plots were made to predict the cost, number of stops, destination, source etc. only python was used for explorative data analysis.

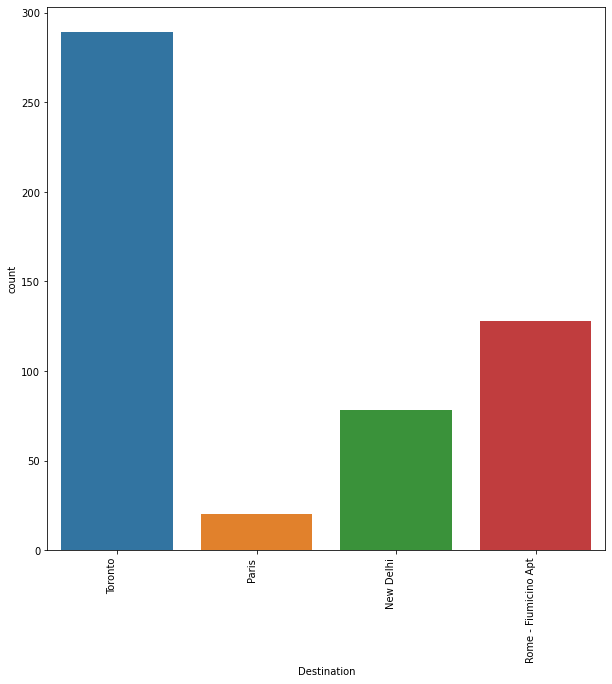


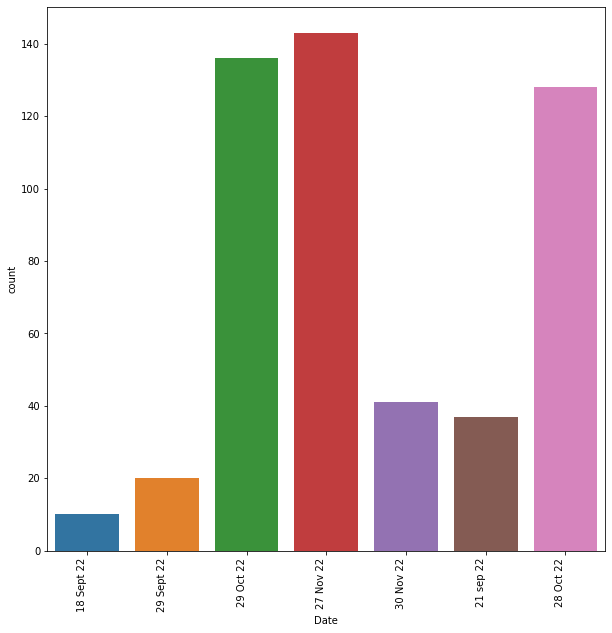














* Interpretation of the Results

Give a summary of what results were interpreted from the visualizations, preprocessing and modelling.

Answer: It was seen that TAP AIR RORTUGAL was the airways having highest amount of travelling time. The month of October most travelling occurred according to the dataset. Most of them have 2 stops at least to commute. ALL NIPPON AIRWAYS is the most costliest. Most travelling occurred on 27 NOV 2022. From the dataset it was seen that TORONTO was the most destination place.

**CONCLUSION**

* Key Findings and Conclusions of the Study

Describe the key findings, inferences, observations from the whole problem.

Answer: It was seen that TAP AIR RORTUGAL was the airways having highest amount of travelling time. The month of October most travelling occurred according to the dataset. Most of them have 2 stops at least to commute. ALL NIPPON AIRWAYS is the most costliest. Most travelling occurred on 27 NOV 2022. From the dataset it was seen that TORONTO was the most destination place.

* Learning Outcomes of the Study in respect of Data Science

List down your learnings obtained about the power of visualization, data cleaning and various algorithms used. You can describe which algorithm works best in which situation and what challenges you faced while working on this project and how did you overcome that.

Answer: From the graphs obtained by the python different conclusions were drawn. As the data was supposed to be taken from the different websites there were challenges in scraping them. Joining and Cleaning of data was also the next challenge as there were many column and junk data.

Limitations of this work and Scope for Future Work

What are the limitations of this solution provided, the future scope? What all steps/techniques can be followed to further extend this study and improve the results.

Answer: Different models were tested but it was seen that after hyperparameter tuning it was found to have at random state, 256, the training accuracy is :0.9130650810656163. At random state, 256, the testing accuracy is :0.8812360857622321. The MAE is 12974.838366156499. The MSE is 12974.838366156499 and RMSE is 113.90714800290849. in order to increase this accuracy different dataset have to be still collected for different destinations. Different deep learning methods can also be employed.