

FLIGHT TICKET PRICE PREDICTION

Submitted by:

NAMRATA BARASKAR

**ACKNOWLEDGMENT**

The internship opportunity I had with Flip Robo was a great chance for learning and professional development. Therefore, I consider myself as a very lucky individual as I was provided with an opportunity to be a part of it.

I would like to thank our SME for suggesting this project and for his whole hearted cooperation and constant encouragement throughout the project.

And I also like to thank the data trained mentors and Technical team members for helping me with the technical queries.

And these are the following website which I referred for the reference

1. <https://www.kaggle.com/>
2. <https://scikit-learn.org/>
3. [www.stackoverflow.com](http://www.stackoverflow.com)
4. [www.google.com](http://www.google.com)
5. [www.geeksforgeeks.org](http://www.geeksforgeeks.org)

**INTRODUCTION**

Business Problem Framing

* With respect to the season flight price also differs in the market, we have seen lot of changes in the Flight ticket price. Now some flight tickets are in demand hence making them costly and some are not in demand hence cheaper. With the change in market due to covid 19 impact, our customers are facing problems with their previous flight price valuation machine learning models. So, they are looking for new machine learning models from new data. We have to make flight ticket price valuation model.
* Motivation for the Problem Undertaken
* To understand real world problems where Machine Learning and Data Analysis can be applied to help to predict the prices in various domains to make better decisions with the help of which they can gain profit or can be escaped from any loss which otherwise could be possible without the study of data.

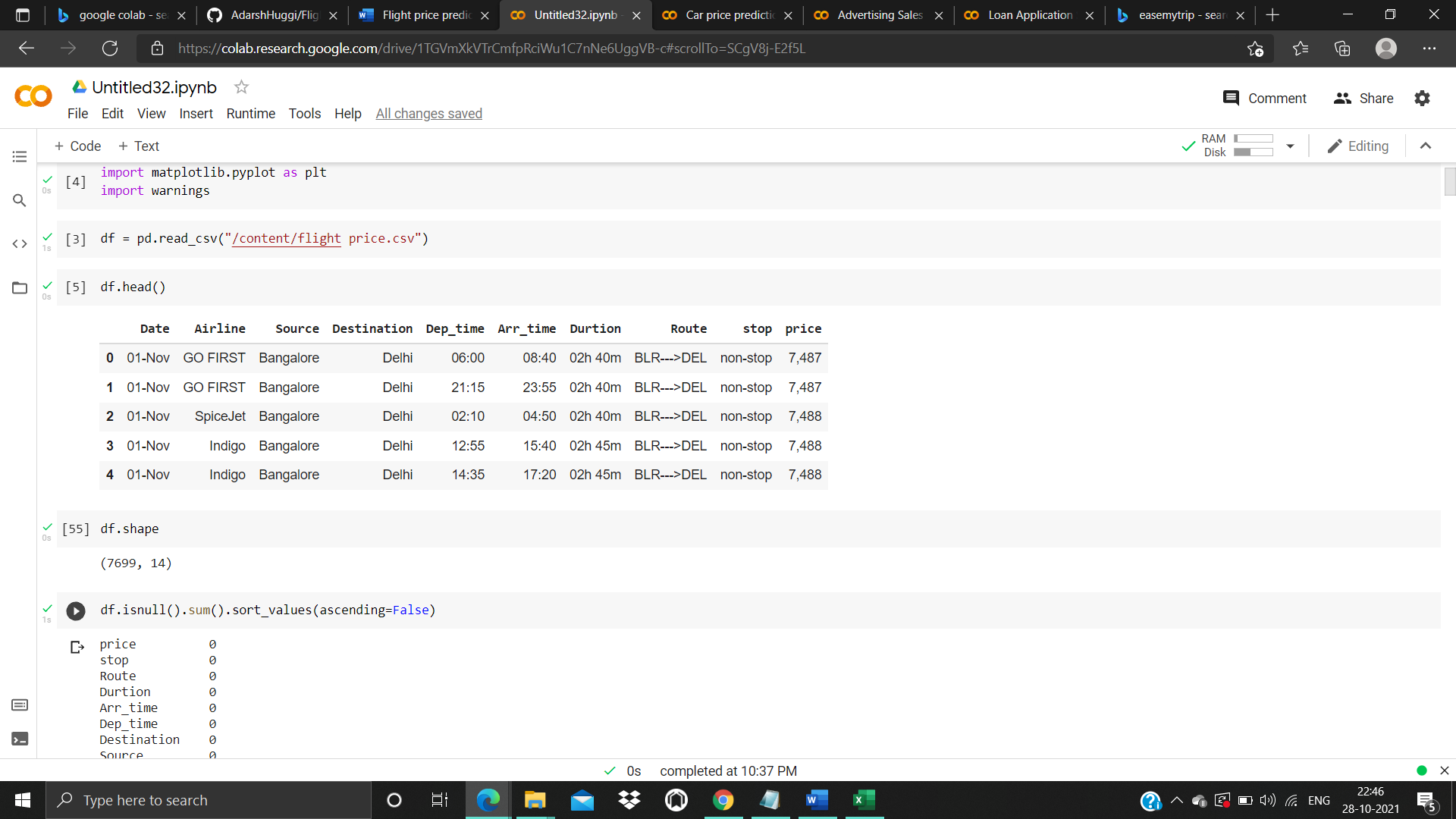
**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

In the whole research process various mathematical, statistical and analytics modelling has been done. There has been reduction of the columns because few of them was not necessary for the problem solving. And few of them was removed due to very less correlation with dependent variable. Since the dataset contains a lot of features hence feature selection has been also done.

* Data Sources and their formats

The data is scraped from the easemytrip website The data descriptions are as follow:- (7699, 14) rows and columns



* Data Pre-processing Done

The raw data is taken and performed various steps to reduce skewness, outlier, class imbalance and scaling. There were null value was present and removed the values from the dataset. Many outlier removal and skewness removal methods are tested and best method Is chosen in order to prevent data loss.

· The dataset contains 7699 rows and 14 columns

· Price is our dependent variable.

· We created new features from old ones.

· There are no null values in the dataset.

· Removed empty cells

* Data Inputs- Logic- Output Relationships

The input data contains 7699 rows and 14 columns.

Predictor variable are

* stop, Route, Durtion, Arr\_time, Dep\_time, Destination, Source, Airline, Date
* Target variable is Price of the flight ticket
* Hardware and Software Requirements and Tools Used

Hardware – Laptop

Software – google colab, jupyter notebook

Libraries- numpy, pandas, seaborn, matplotlib.pyplot, sklearn.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

Regression Model with following algorithms

* Linear Regression
* DecisionTreeRegressor
* Random forest regressor

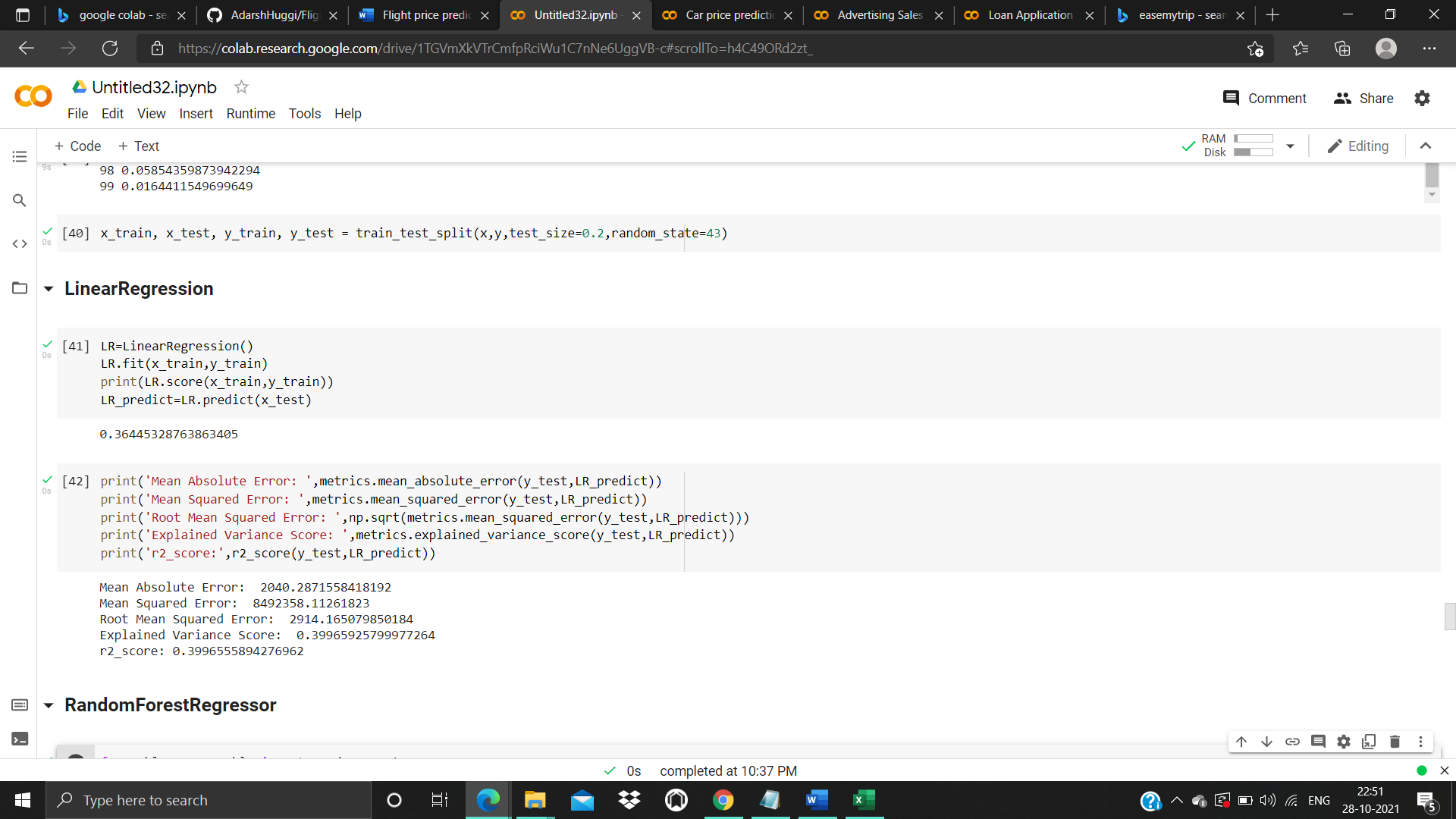
Evaluation metrics

* Mean square error
* Mean absolute error
* R2 score
* Root Mean Squared Error
* Testing of Identified Approaches (Algorithms)

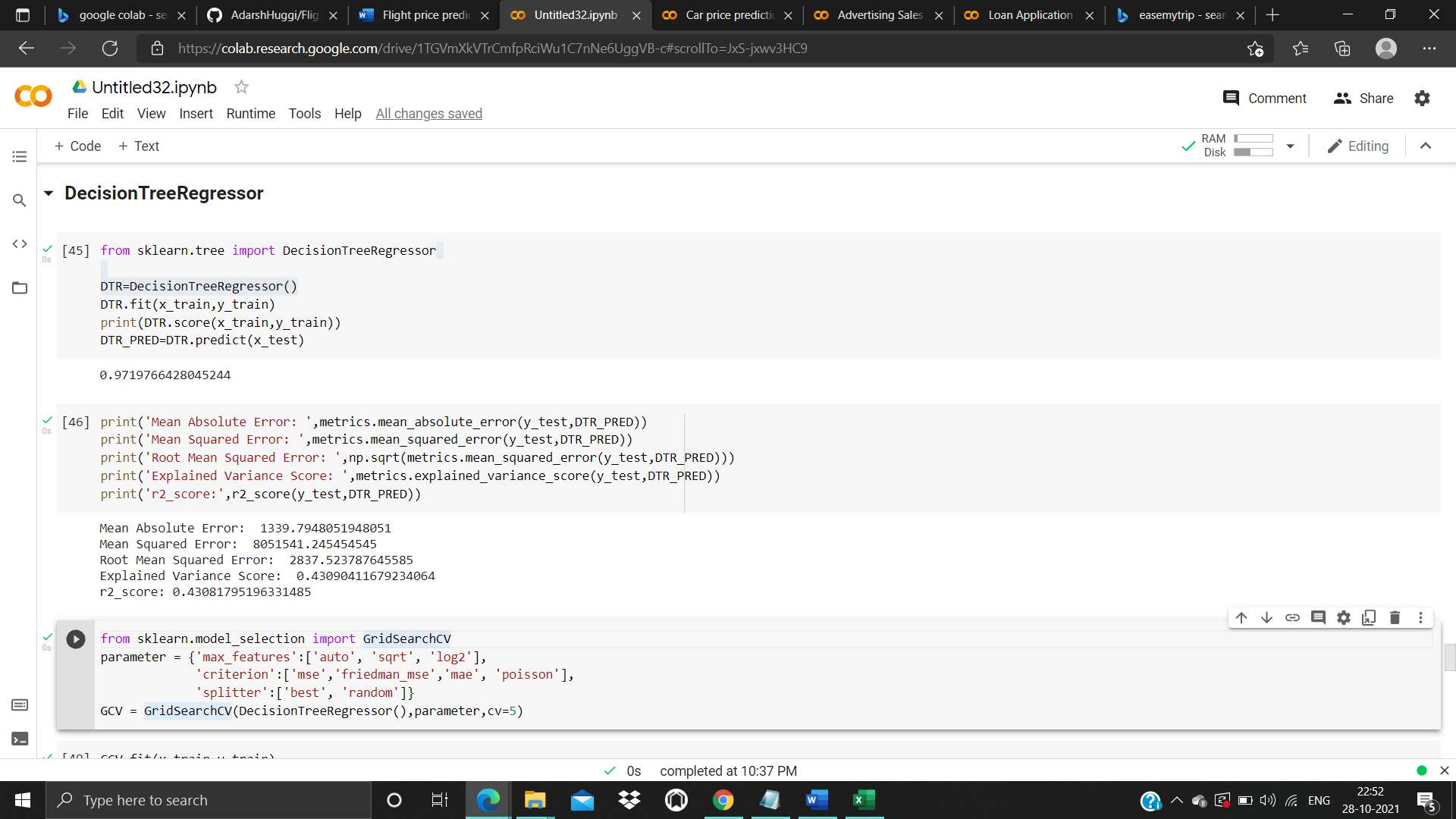
Listing down all the algorithms used for the training and testing.

* LR=LinearRegression()
* DT=DecisionTreeRegressor()
* rf=RandomForestRegressor()
* Run and Evaluate selected models

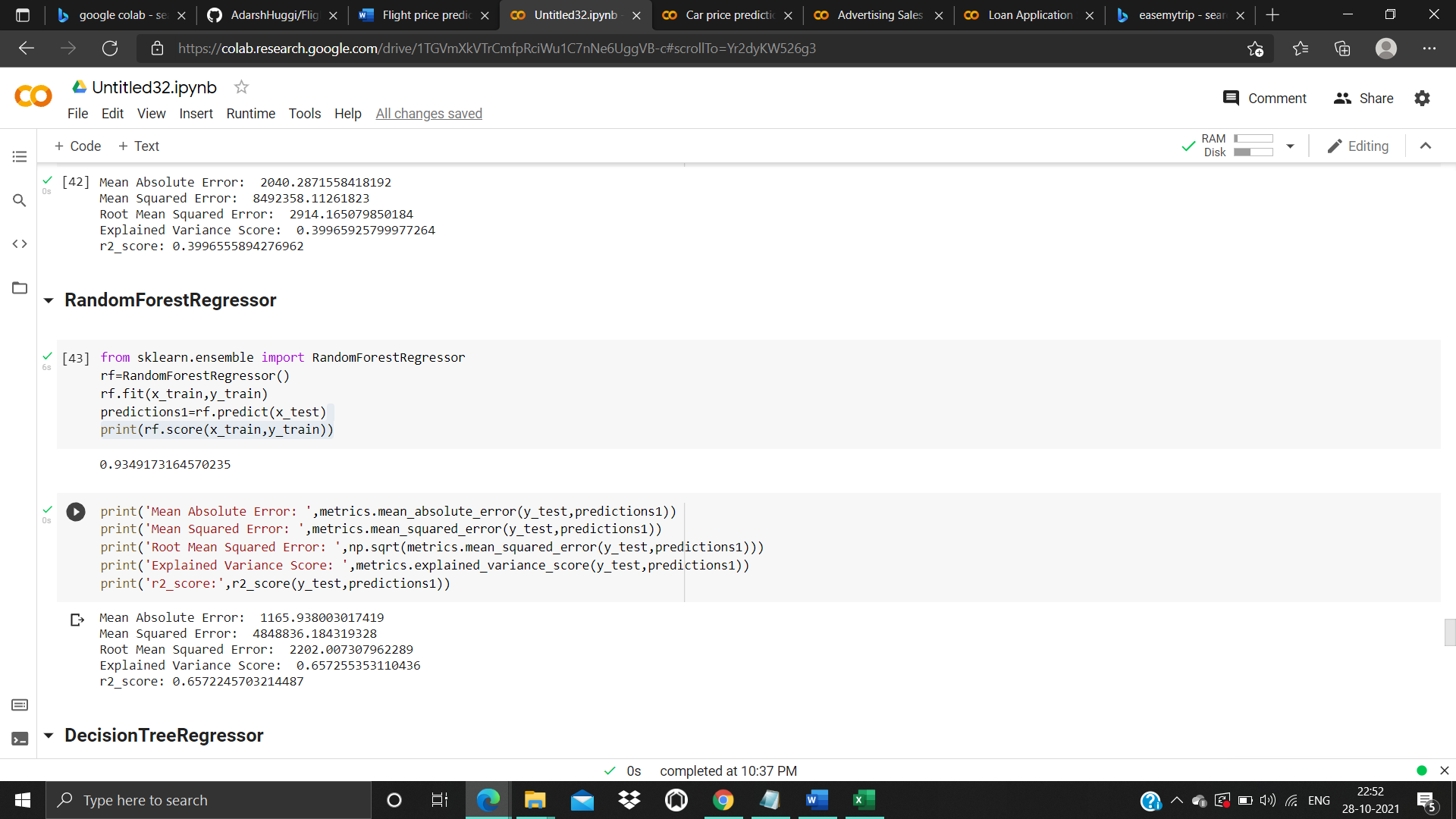
Describe all the algorithms used along with the snapshot of their code and what were the results observed over different evaluation metrics.



1. Linear Regression

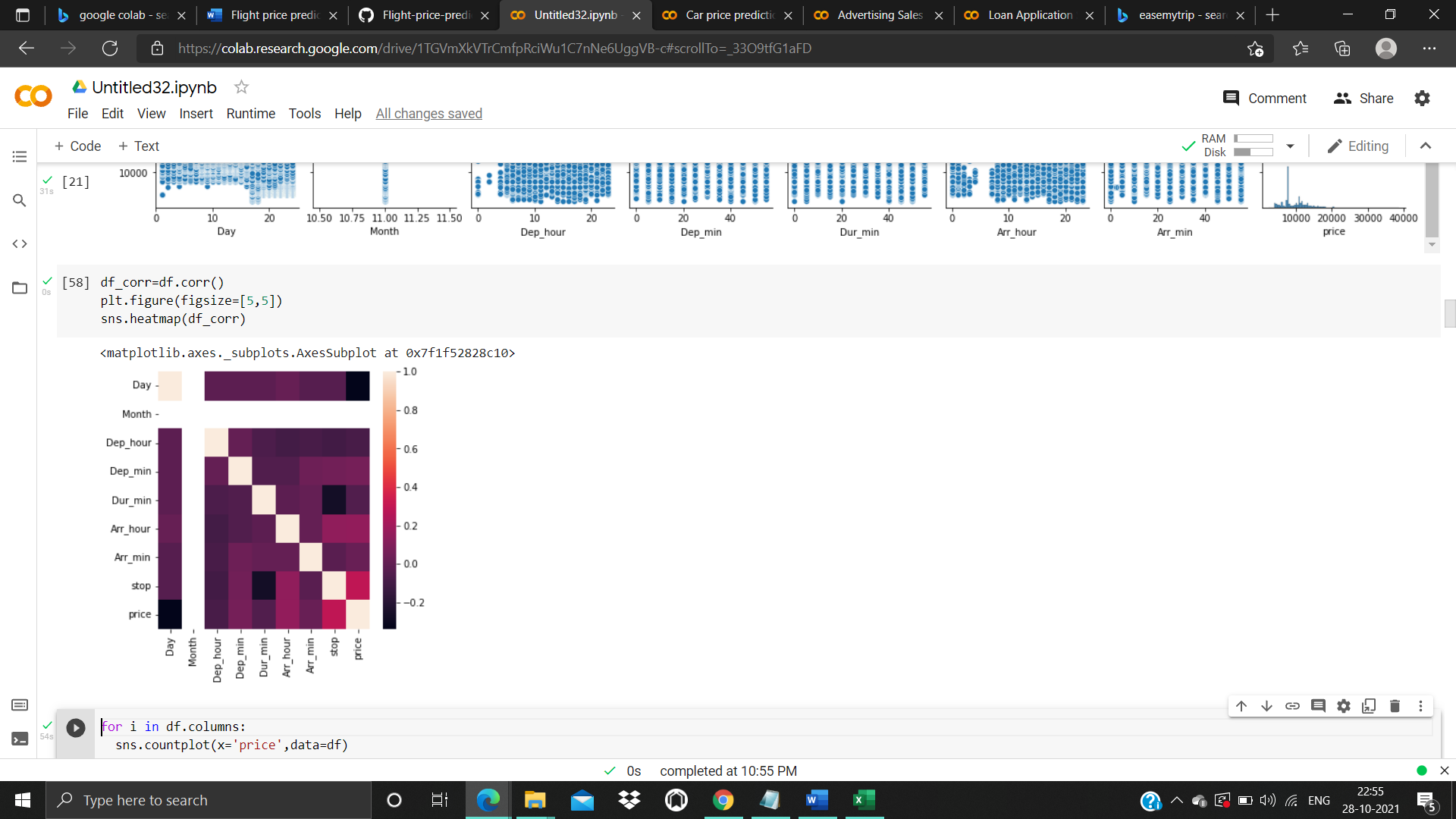


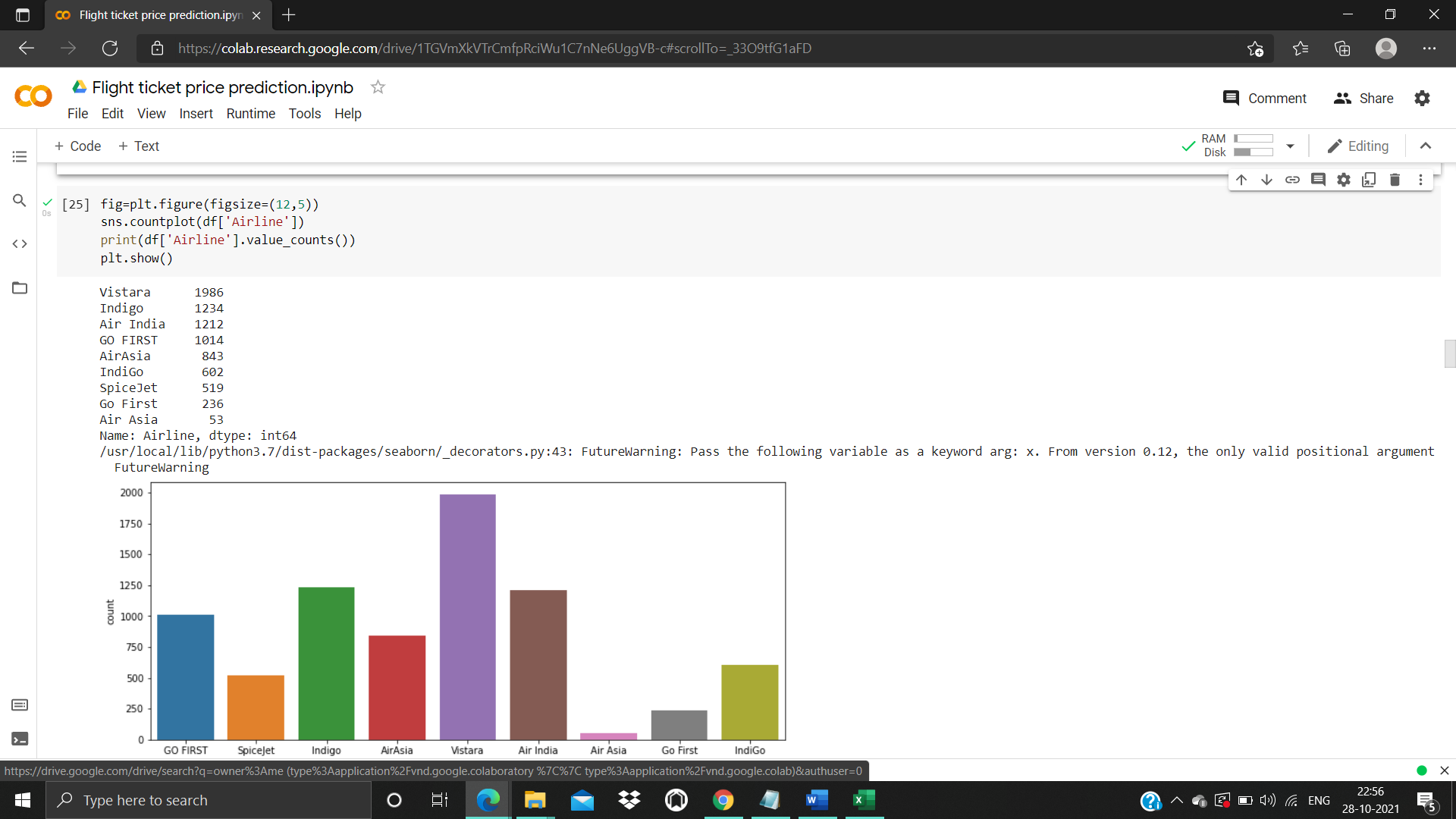
2. DecisionTreeRegressor

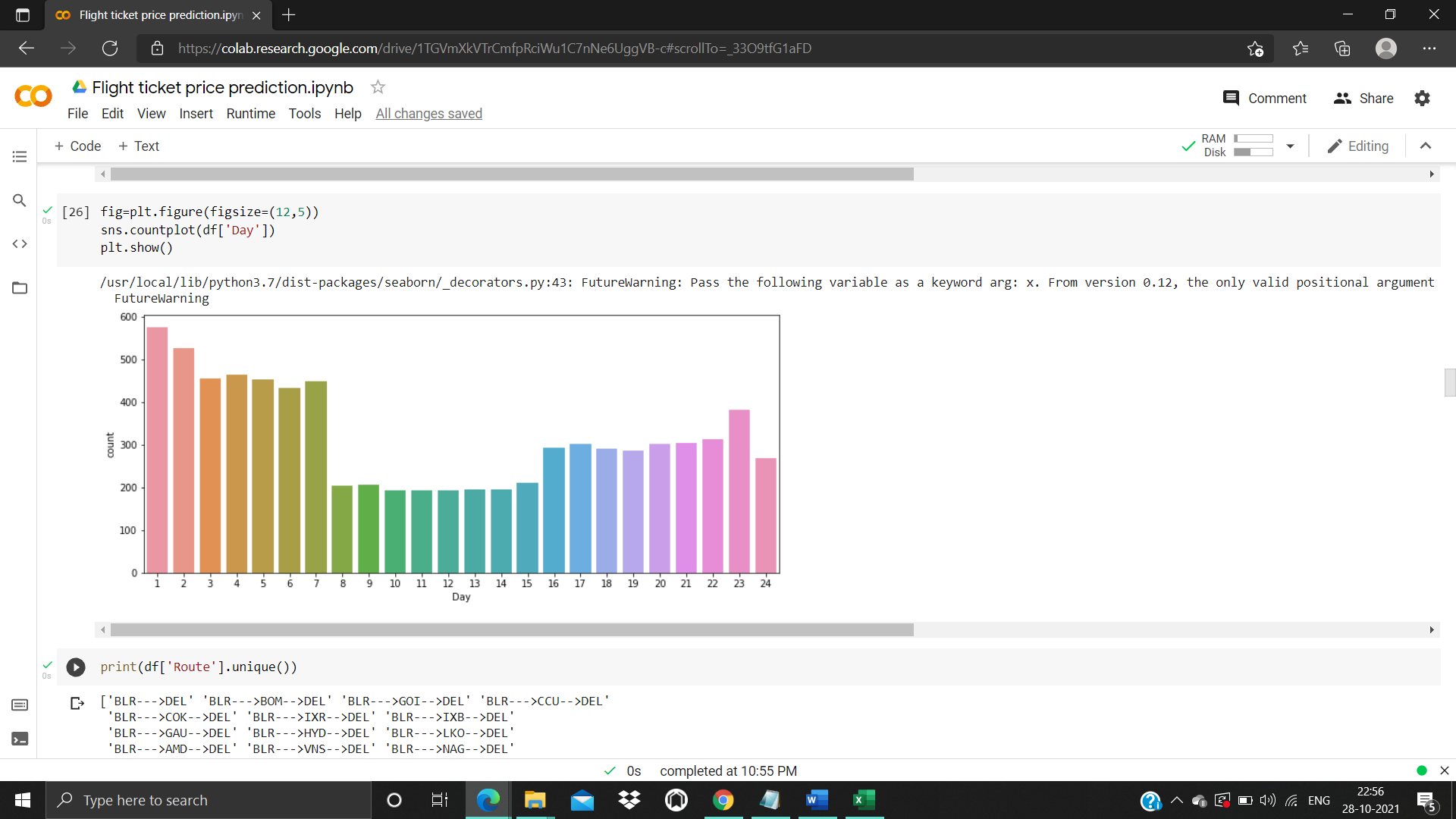


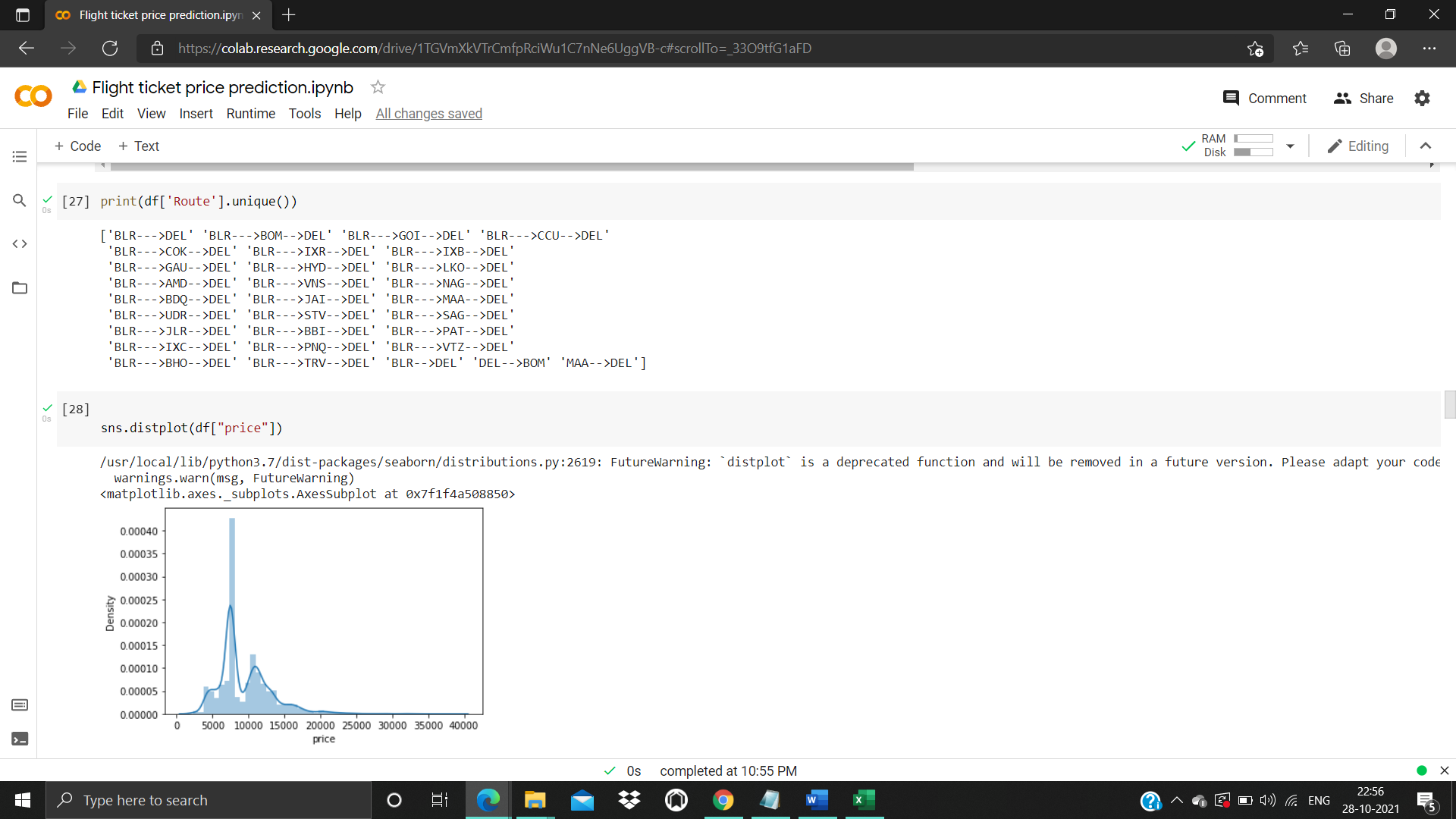
3.Random forest

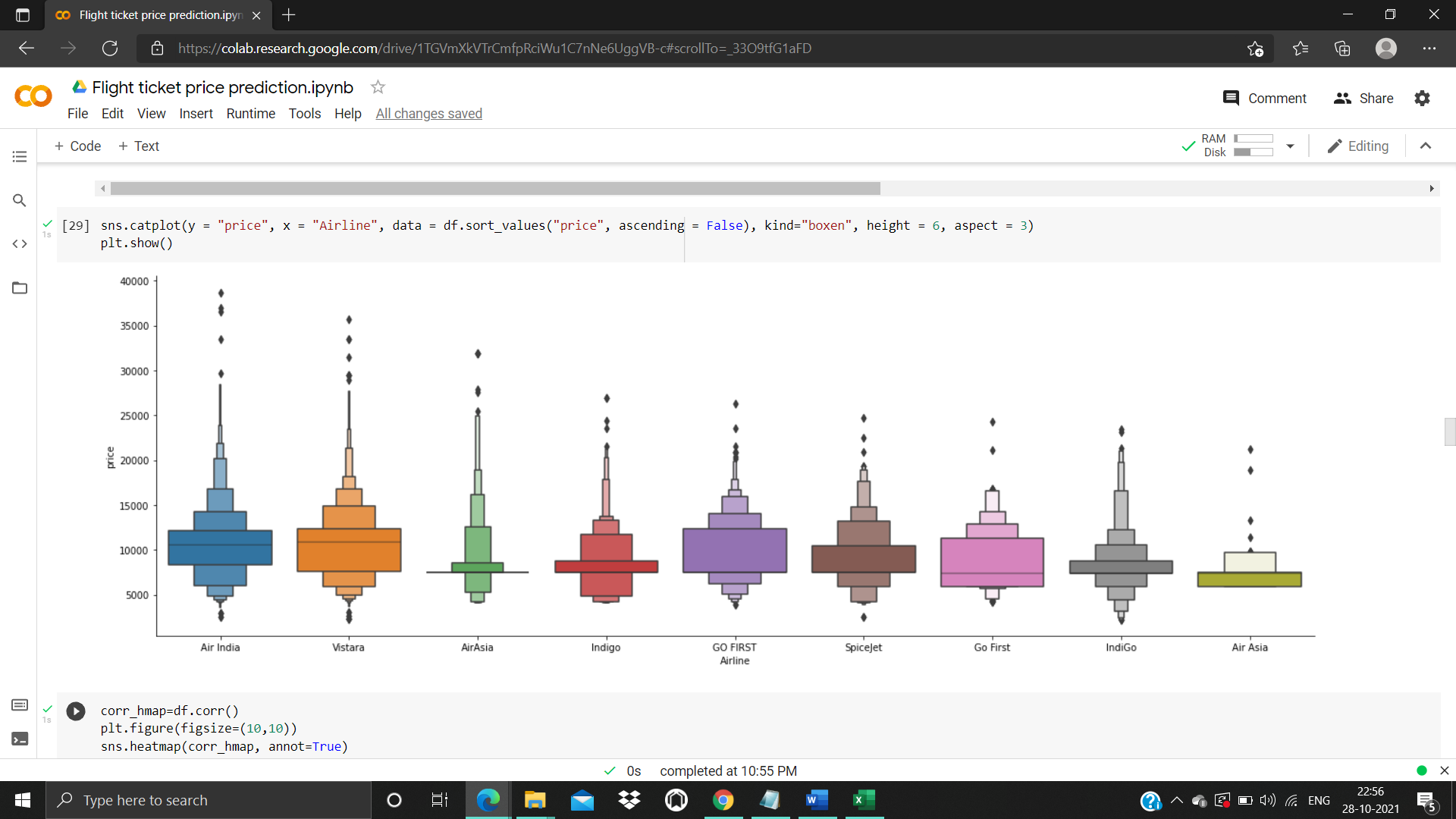
* Visualizations

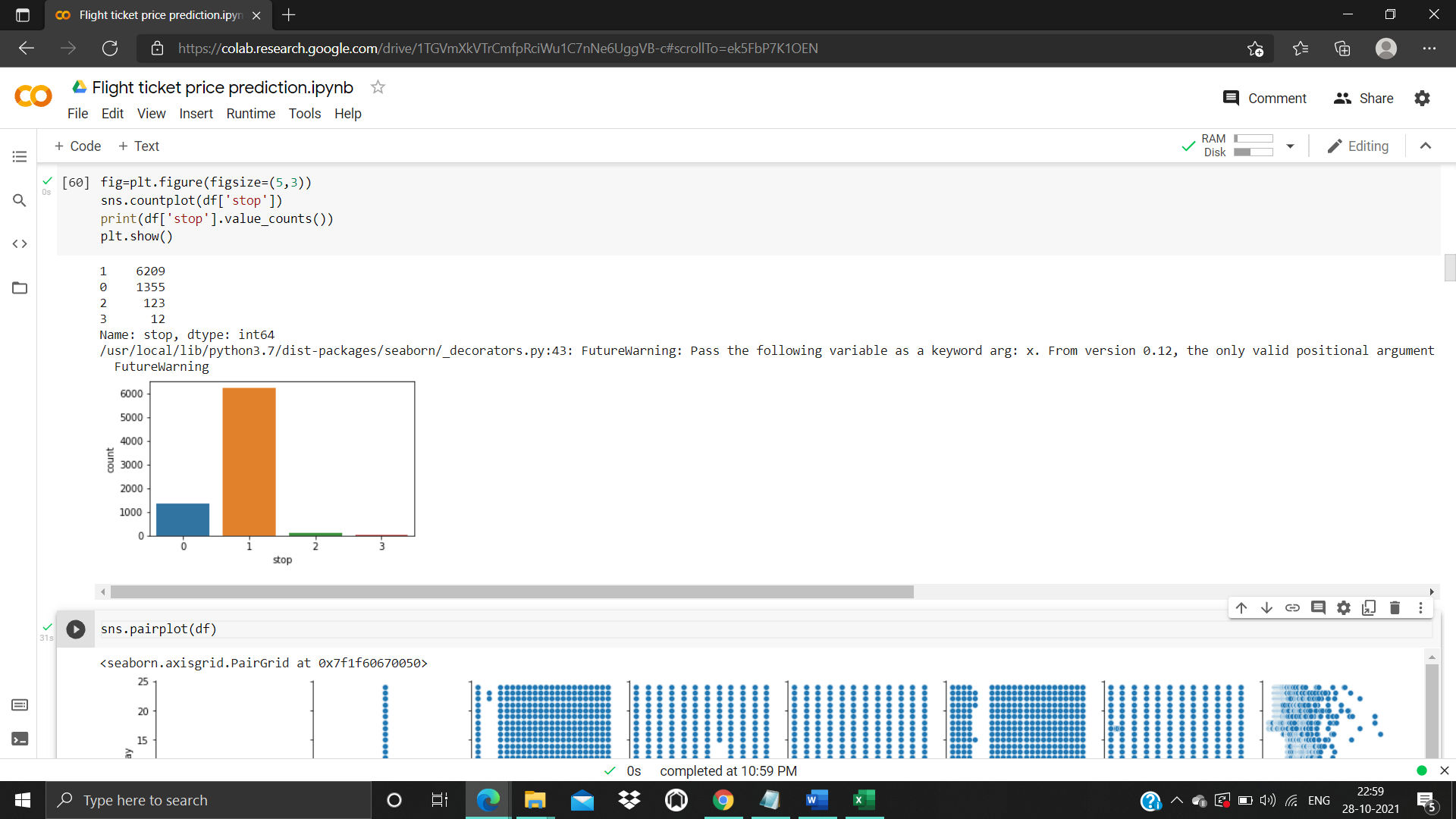












ANALYSIS ON THE DATA.

1. Do airfares change frequently?

Yes, Based on Season and Demand of the flight Tickets Price changes regularly.

2. Do they move in small increments or in large jumps?

 Ans. large jumps, If you buy a ticket before few of date of journey you will get arround 44% high price

3. Do they tend to go up or down over time?

Ans. Price of the flight goes up only down over time but its not reduces the price over time

4. What is the best time to buy so that the consumer can save the most by taking the least risk?

Ans. Consumer should buy the ticket before 30 days of date of joureney so they can save up to 45% to 50%.

**CONCLUSION**

* Key Findings and Conclusions of the Study

From this dataset I get to know that each feature play a very import role to understand the data. Data format plays a very important role in the visualization and Appling the models and algorithms. Importance of removing the skewness and outlier is important. Finding the best parameters for the algorithm also plays a important role in performance and accuracy of the model.

* Learning Outcomes of the Study in respect of Data Science

Learnt how to process the large number of data. Tried and learnt more about distribution of the data. The power of visualization is helpful for the understanding of data into the graphical representation its help me to understand that what data is trying to say, Data cleaning is one of the most important step to remove missing value or null value fill it by mean median or by mode or by 0.Setting a good parameters is more important for the model accuracy. Finding a best random state played a vital roll in finding a better model.

* Limitations of this work and Scope for Future Work

The techniques to increase the speed of the model need to be constructed. The future model can be constructed with the most co related data with the target variable in order to increase the speed of the model.