**Flight Price Prediction**

**Contents**

1. [Problem Definition](https://towardsdatascience.com/building-an-employee-churn-model-in-python-to-develop-a-strategic-retention-plan-57d5bd882c2d#c099)
2. [Data Analysis](https://towardsdatascience.com/building-an-employee-churn-model-in-python-to-develop-a-strategic-retention-plan-57d5bd882c2d#6270)
3. [EDA Concluding Remarks](https://towardsdatascience.com/building-an-employee-churn-model-in-python-to-develop-a-strategic-retention-plan-57d5bd882c2d#030a)
4. [Pre-processing Pipeline](https://towardsdatascience.com/building-an-employee-churn-model-in-python-to-develop-a-strategic-retention-plan-57d5bd882c2d#dd40)
5. [Building Machine Learning Models](https://towardsdatascience.com/building-an-employee-churn-model-in-python-to-develop-a-strategic-retention-plan-57d5bd882c2d#8341)
6. [Concluding Remarks](https://towardsdatascience.com/building-an-employee-churn-model-in-python-to-develop-a-strategic-retention-plan-57d5bd882c2d#13e2)

**Problem Statement:**

A flight ticket's fare varies depending on a variety of factors such as length, number of stops, source, day, airline, and many others. Each airline uses its own algorithms and rules to determine the fare. Flight ticket prices can be difficult to anticipate; we may see a price today, but check the price of the same flight tomorrow; it will be a different storey. We've all heard travellers complain about how unpredictable flight ticket prices are.

Machine Learning allows us to estimate the price even more accurately than the real price. Flight prices are often difficult to predict because they fluctuate.

The goal of this blog is to predict the cost of a plane ticket using a dataset. The target value is a continuous value, which is the flight fare. Though here Regression algorithms will be coming in use with the help of Python. The comparison of various different algorithms will take place here solely to obtain the highest accuracy, which will aid in obtaining the most accurate flight price.

We have two sets of files,

A train data file with the target variable (i.e.) Flight price. Which will be used for EDA, pre-processing pipeline, building the model training the model, evaluating the model.

A test data file (i.e.) Without target variable, will be used to predict the prices.

**Data Analysis:**

The flight dataset contains 10683 rows and 11 columns.



These are the columns in the dataset.

Airline: The name of the carrier.

Date\_of\_Journey: Journey Date.

Source: The source from where the flight is taking off.

Destination: The destination points where the Flight lands.

Route: The route reserved by the flight to reach the last stop.

Dep\_Time: The time when the journey starts from the source.

Arrival\_Time: Time of arrival at the destination.

Duration: Total time taken (duration of the flight).

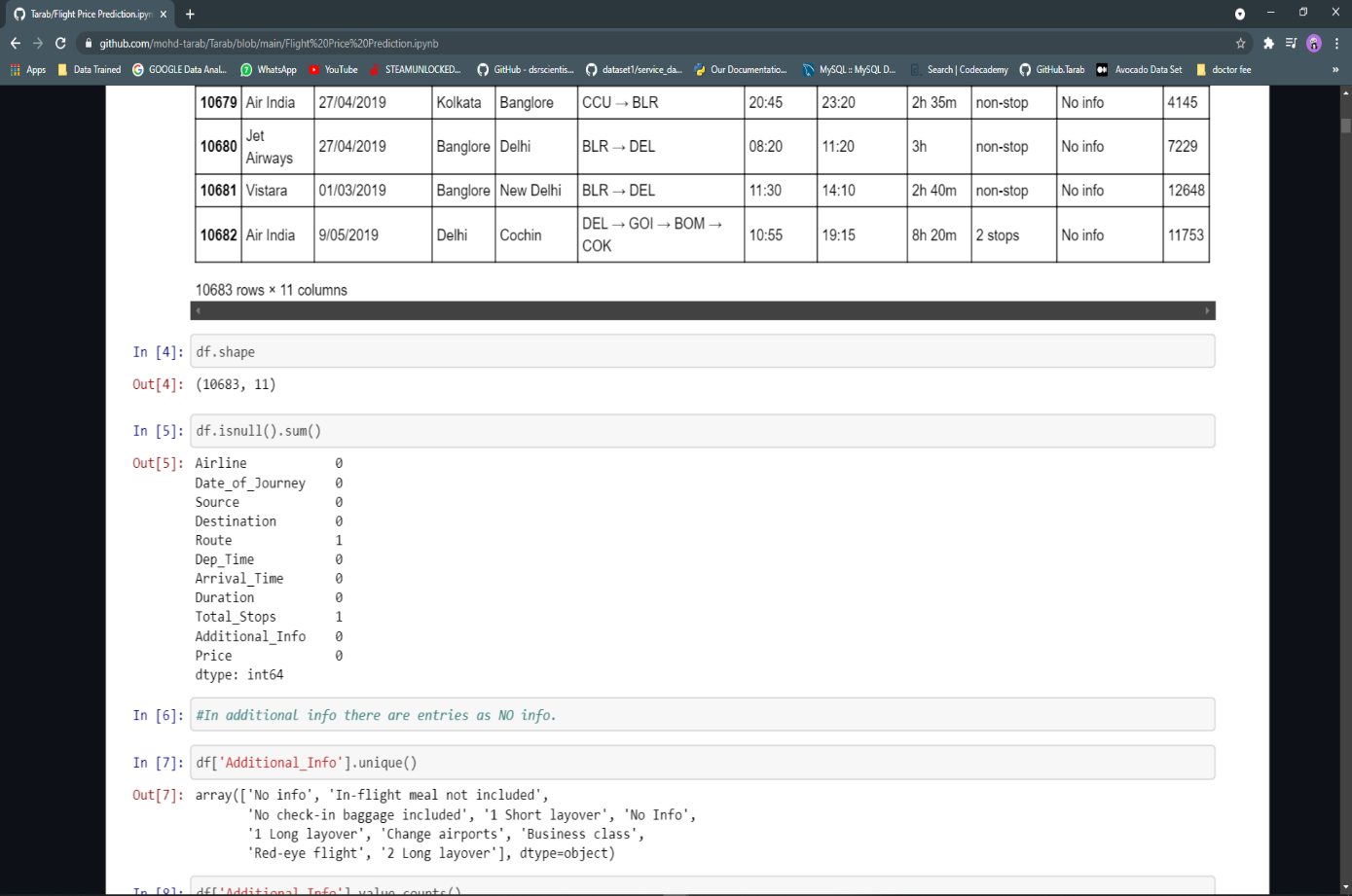
Total\_Stops: Total number of stops between the source and destination.

Additional\_Info: Additional information about the aircraft

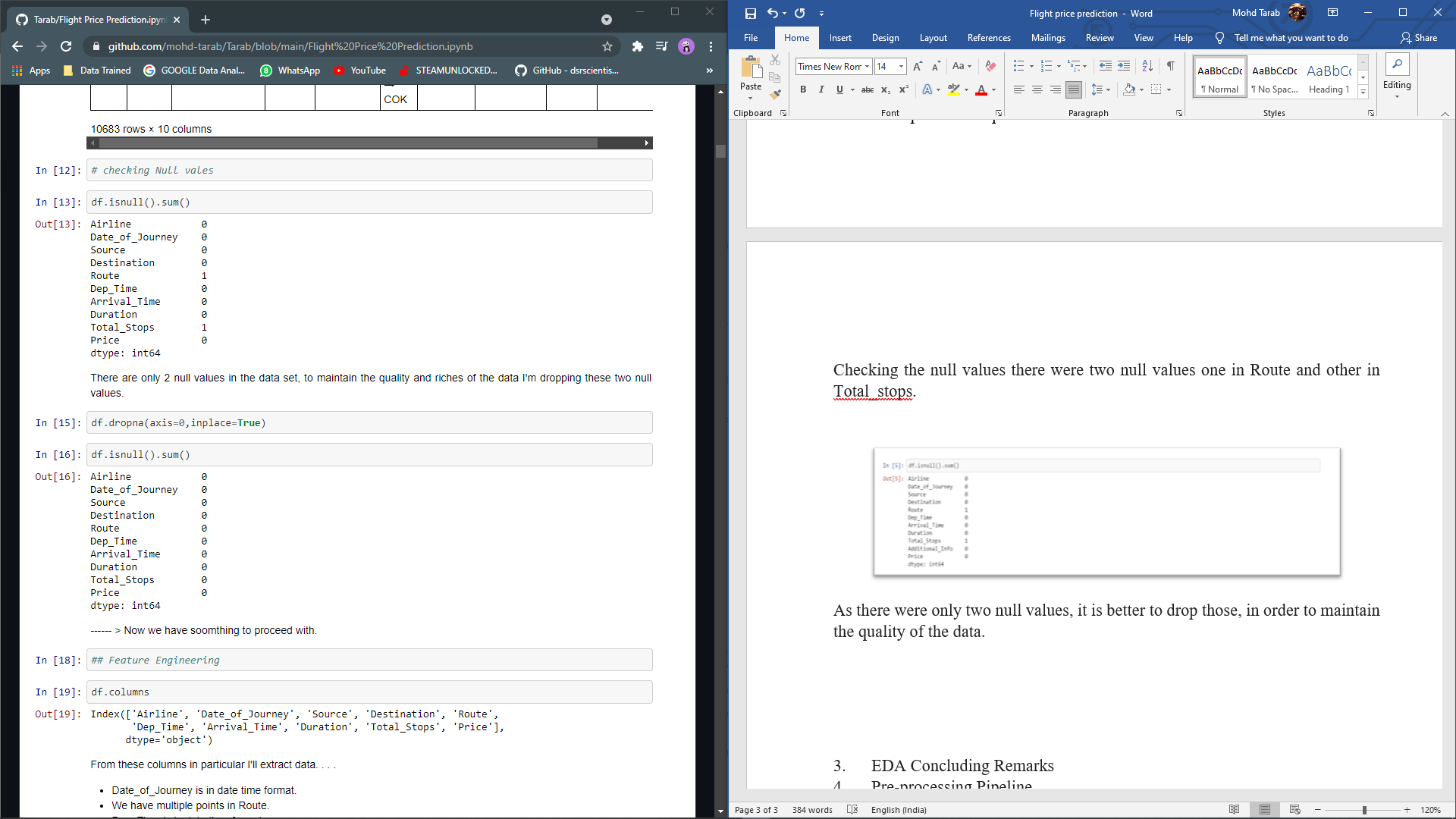
Price: The price of the plane ticket

Checking the null values

There were two null values one in Route and other in Total stops.

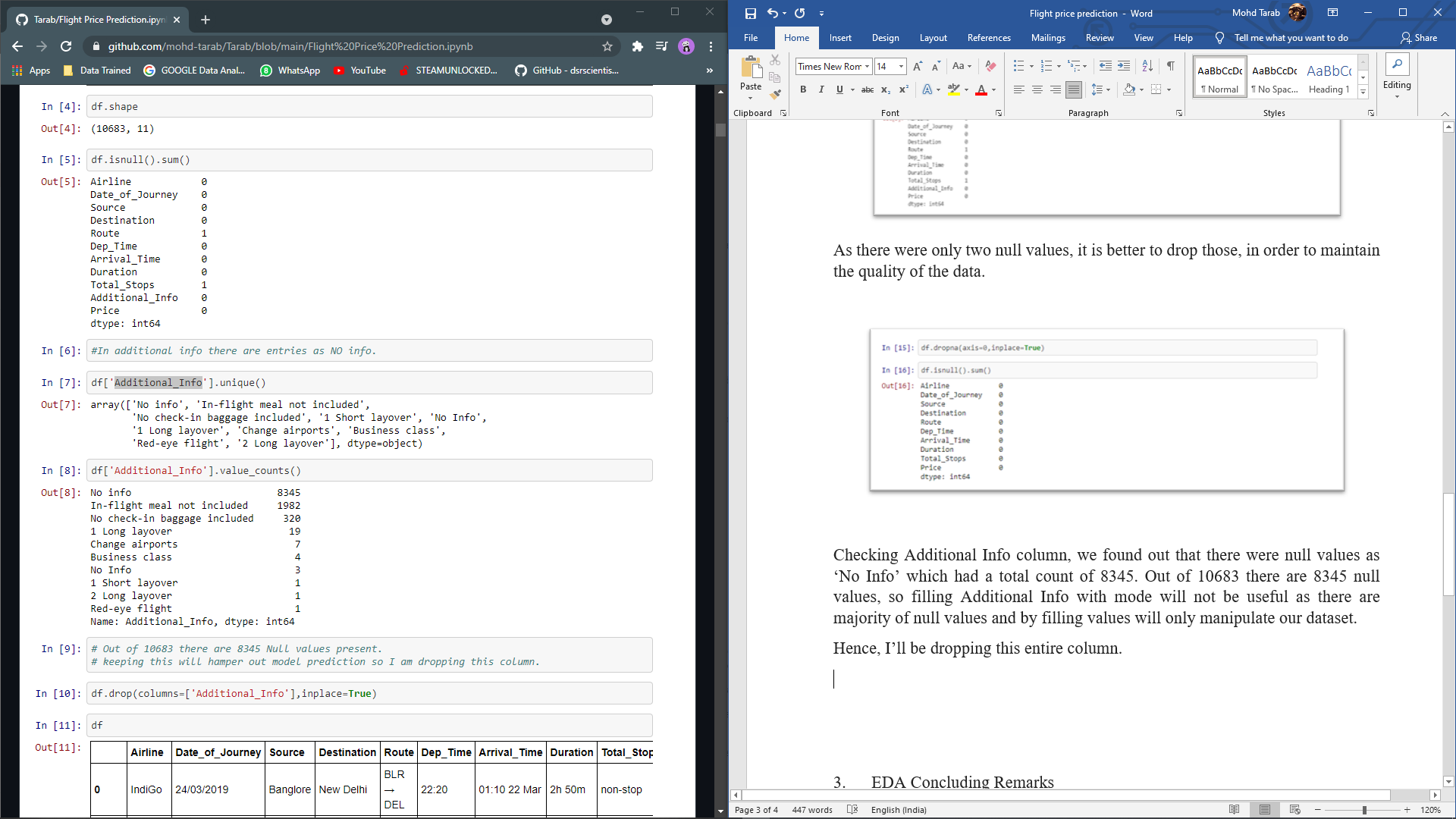


As there were only two null values, it is better to drop those, in order to maintain the quality of the data.



Checking Additional Info column, we found out that there were null values as ‘No Info’ which had a total count of 8345. Out of 10683 there are 8345 null values, so filling Additional Info with mode will not be useful as there are majority of null values and by filling values will only manipulate our dataset.

Hence, I’ll be dropping this entire column.

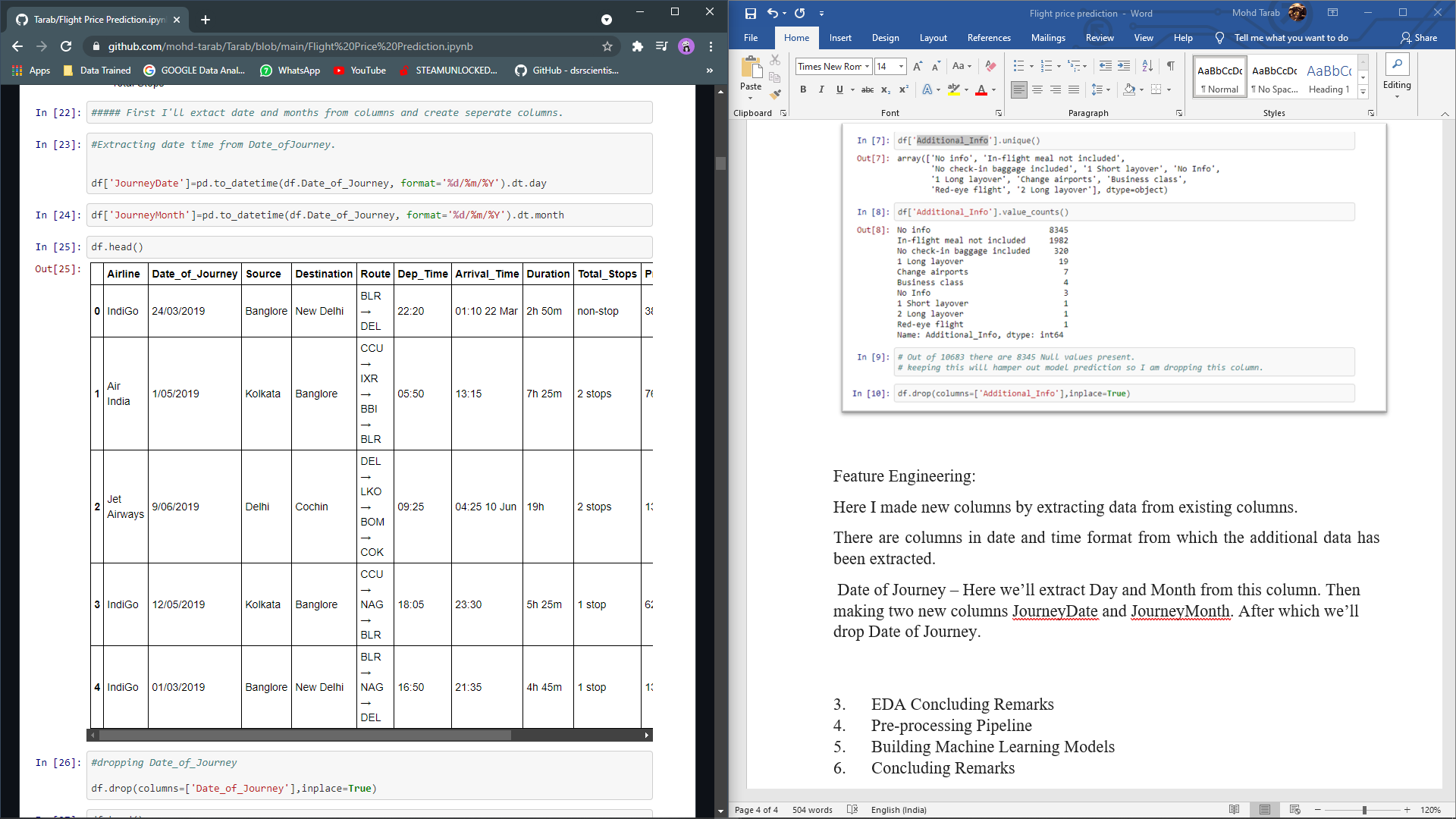


**Feature Engineering:**

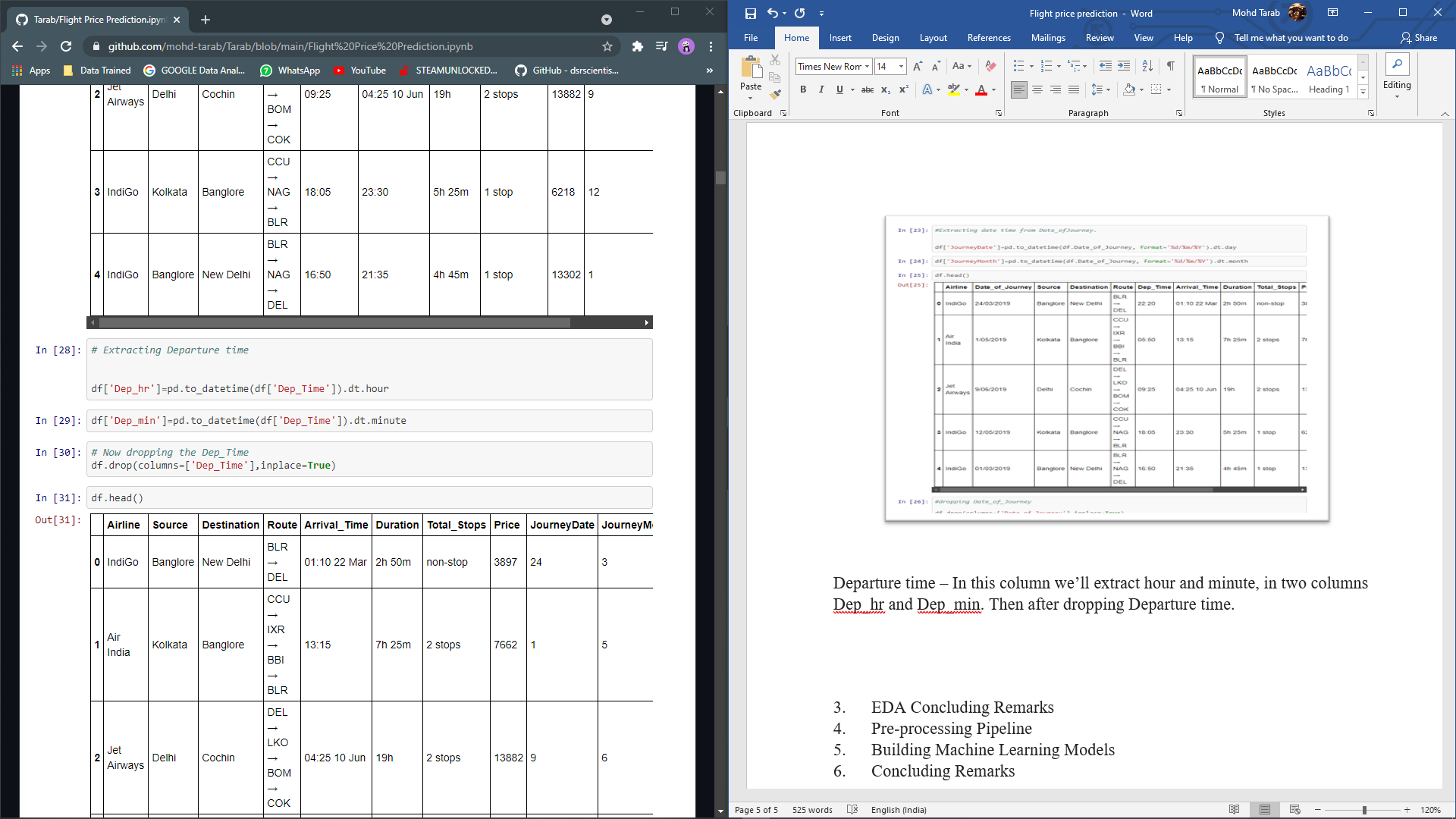
Here I made new columns by extracting data from existing columns.

There are columns in date and time format from which the additional data has been extracted.

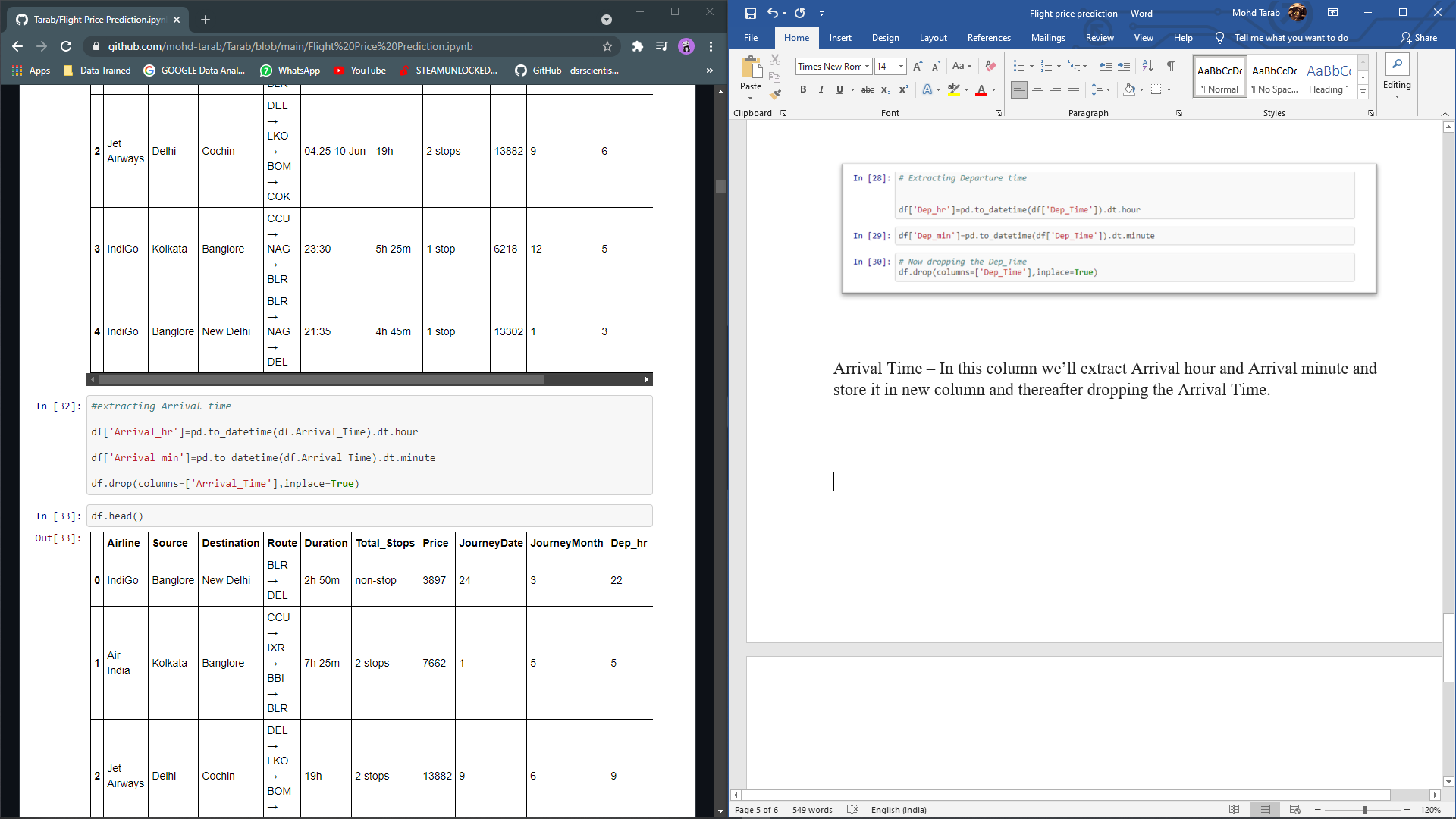
Date of Journey – Here we’ll extract Day and Month from this column. Then making two new columns JourneyDate and JourneyMonth. After which we’ll drop Date of Journey.



Departure time – In this column we’ll extract hour and minute, in two columns Dep\_hr and Dep\_min. Then after dropping Departure time.



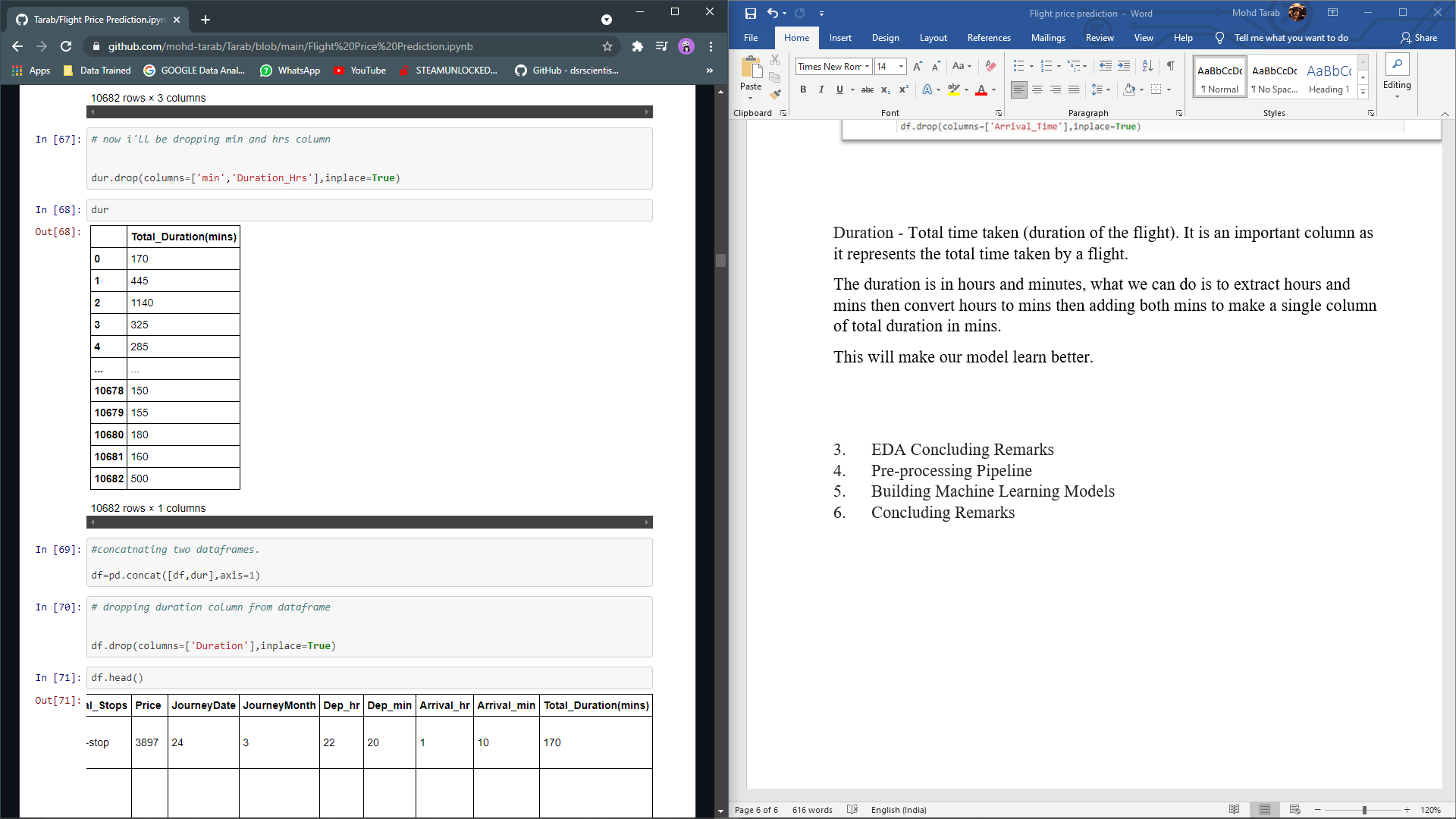
Arrival Time – In this column we’ll extract Arrival hour and Arrival minute and store it in new column and thereafter dropping the Arrival Time.



Duration - Total time taken (duration of the flight). It is an important column as it represents the total time taken by a flight.

The duration is in hours and minutes, what we can do is to extract hours and mins then convert hours to mins then adding both mins to make a single column of total duration in mins.

This will make our model learn better.



There is a catch in the data frame, Route and Total stops are the same thing, and both of them show the same result for example if a flight takes a Route from DEL → LKO → BOM → COK(Cochin) it will have 2 stops LKO and BOM. Therefore, in this way total number of stops and Routes are working in same manner. Hence dropping Route column will be a better thing to do.

**Univariate Analysis**

Airline

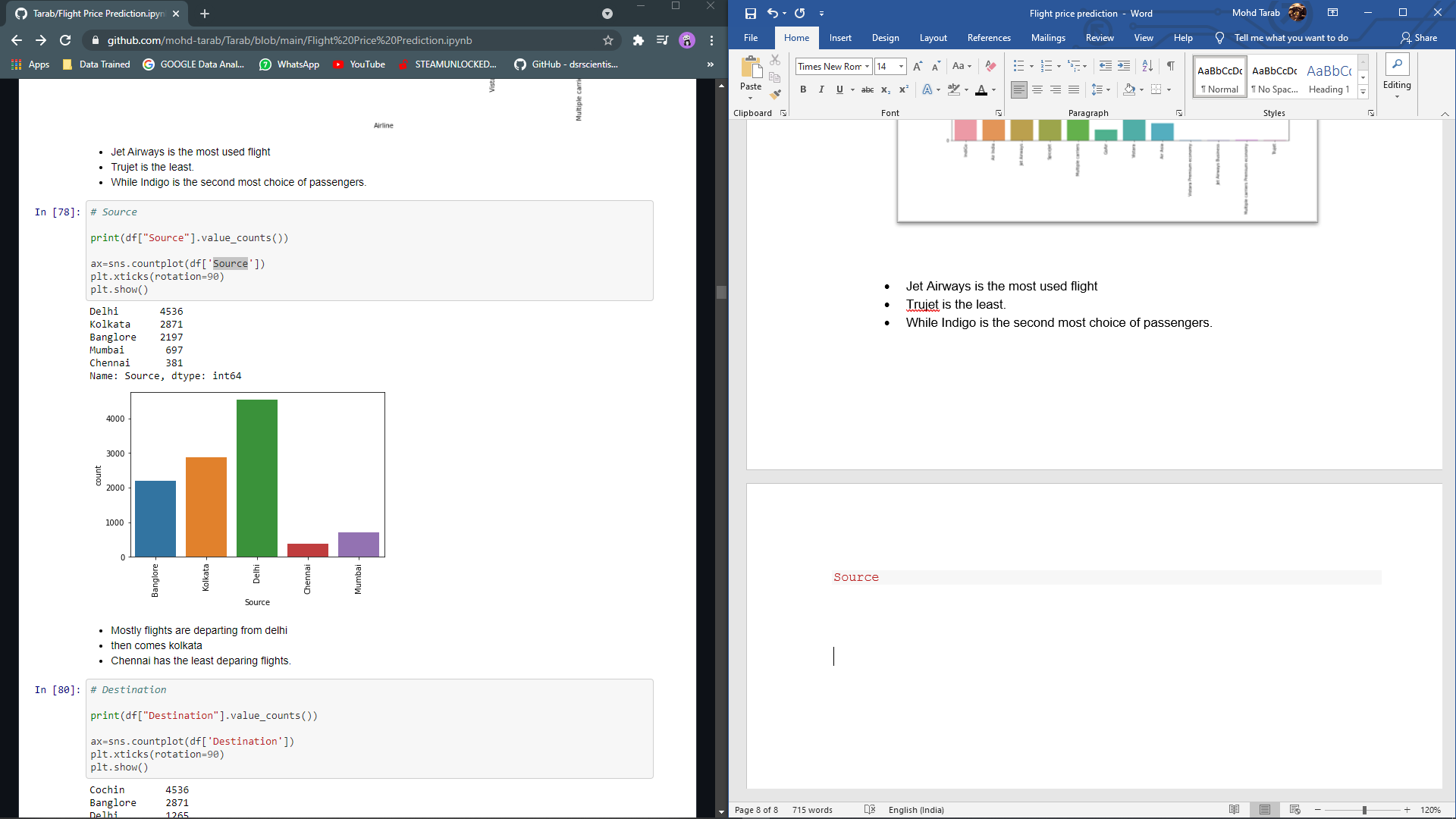


Jet Airways is the most used flight

Trujet is the least.

While Indigo is the second most choice of passengers.

Source

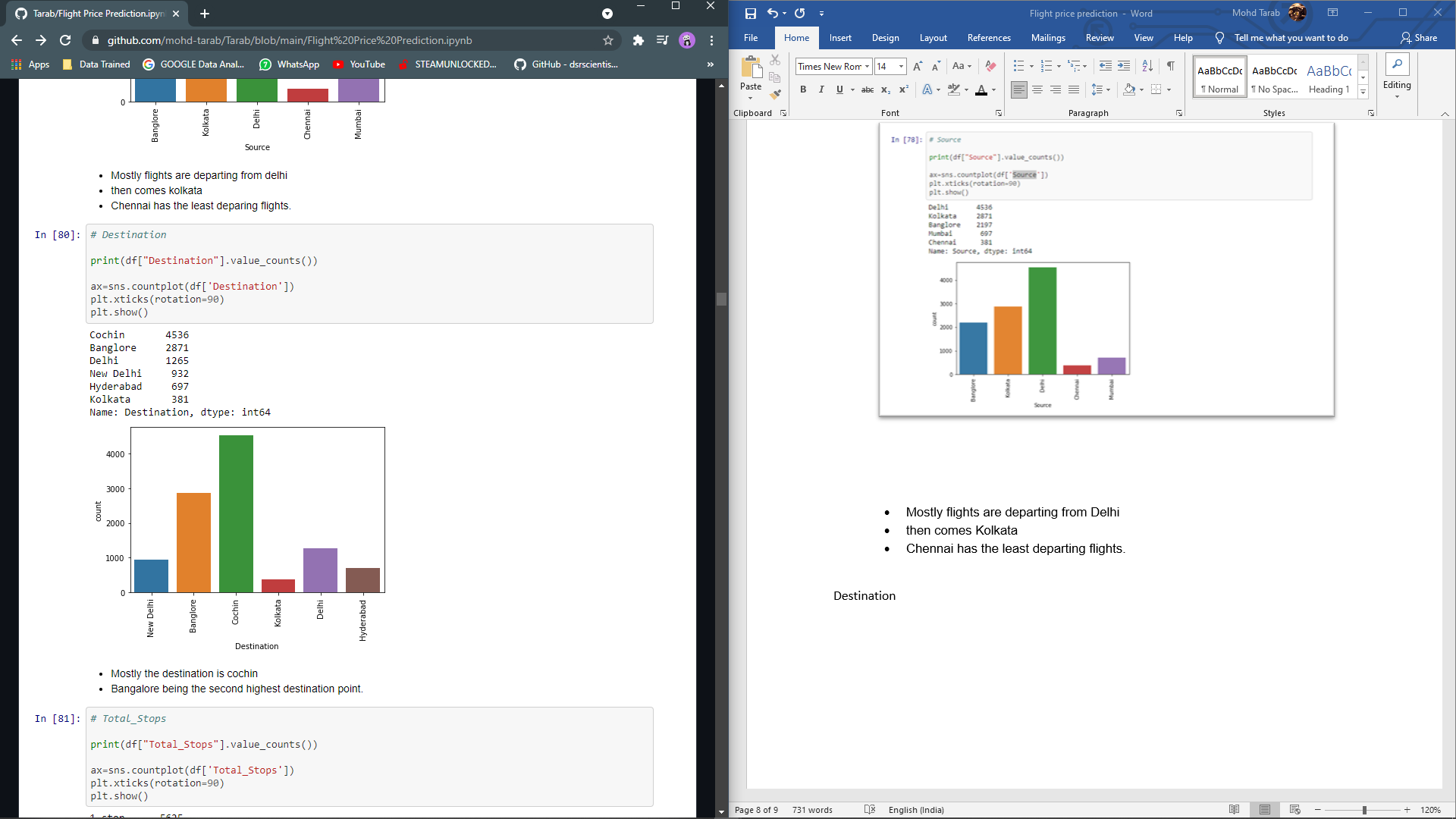


Mostly flights are departing from Delhi

then comes Kolkata

Chennai has the least departing flights.

Destination

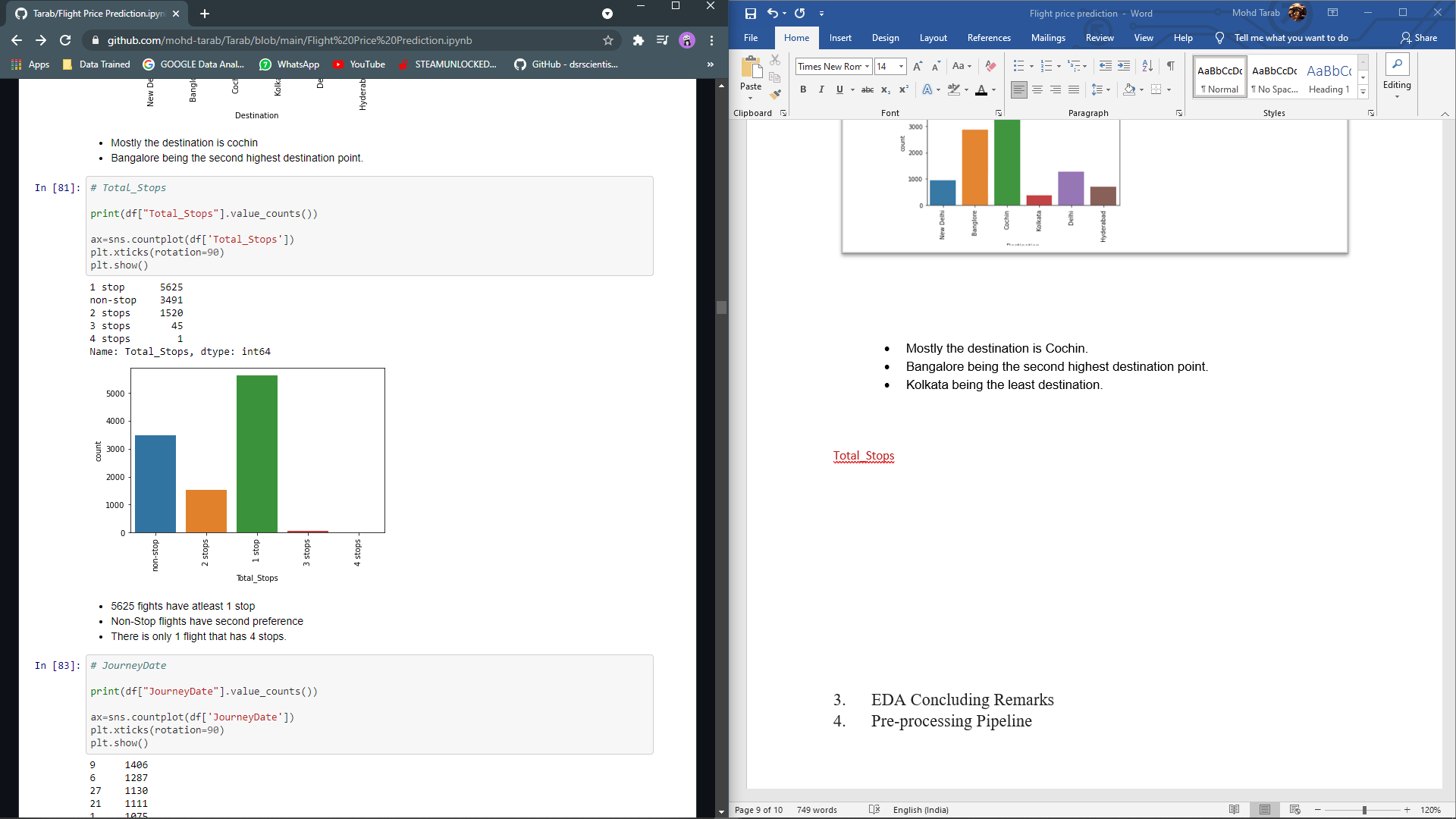


Mostly the destination is Cochin.

Bangalore being the second highest destination point.

Kolkata being the least destination.

Total Stops



5625 fights have at least 1 stop.

Non-Stop flights have second preference.

There is only 1 flight that has 4 stops.

Journey Date

Most preferred date is 9th.

Least preferred date is 18th

Journey Month

May and June are the most travelled months,

while April being the least.

Dep\_hr

Most (915) fights are taking off at 9AM.

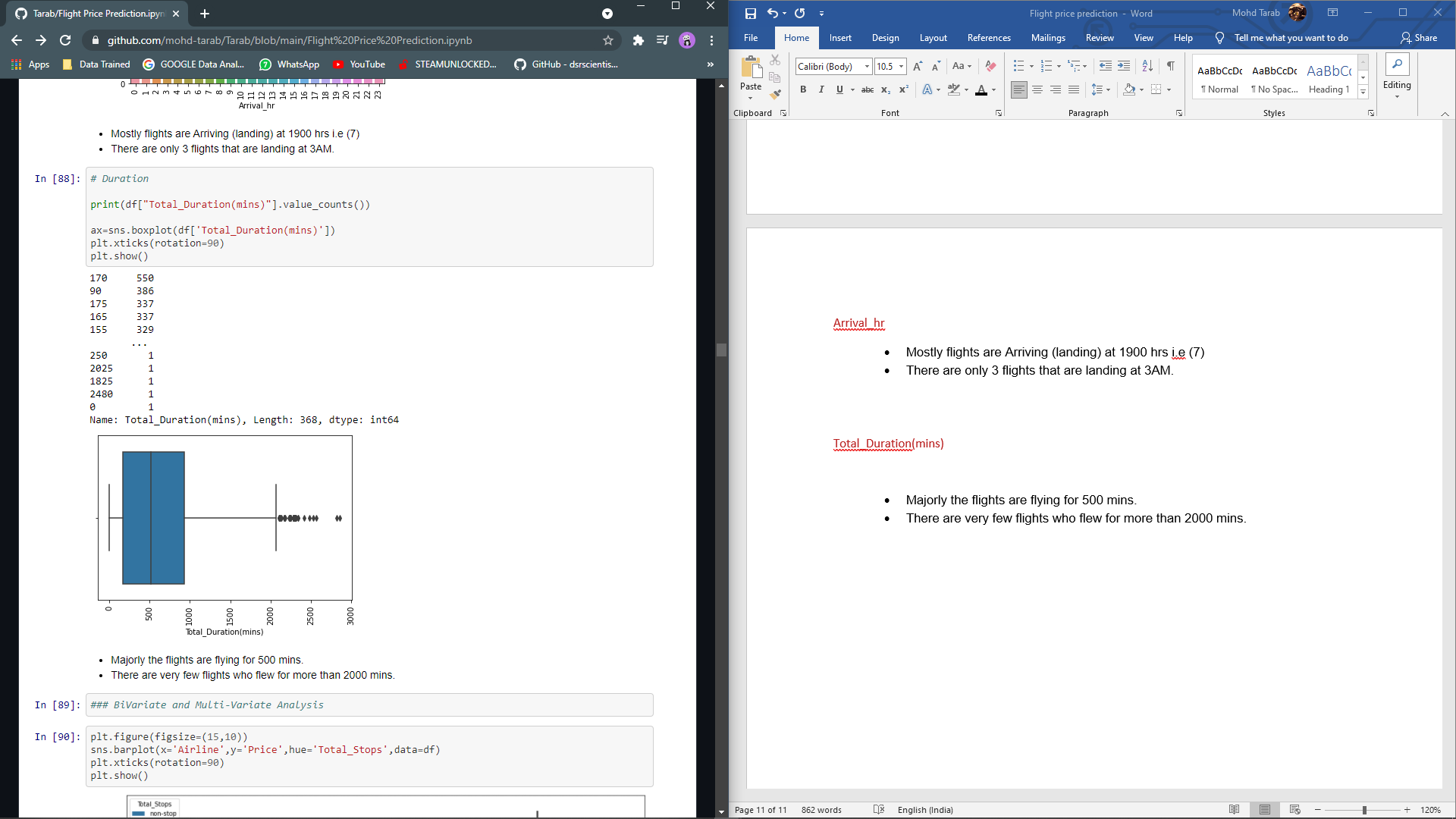
Only 24 flights are taking off at 3AM, making it the least dep time.

Arrival\_hr

Mostly flights are Arriving (landing) at 1900 hrs i.e (7)

There are only 3 flights that are landing at 3AM.

Total\_Duration(mins)

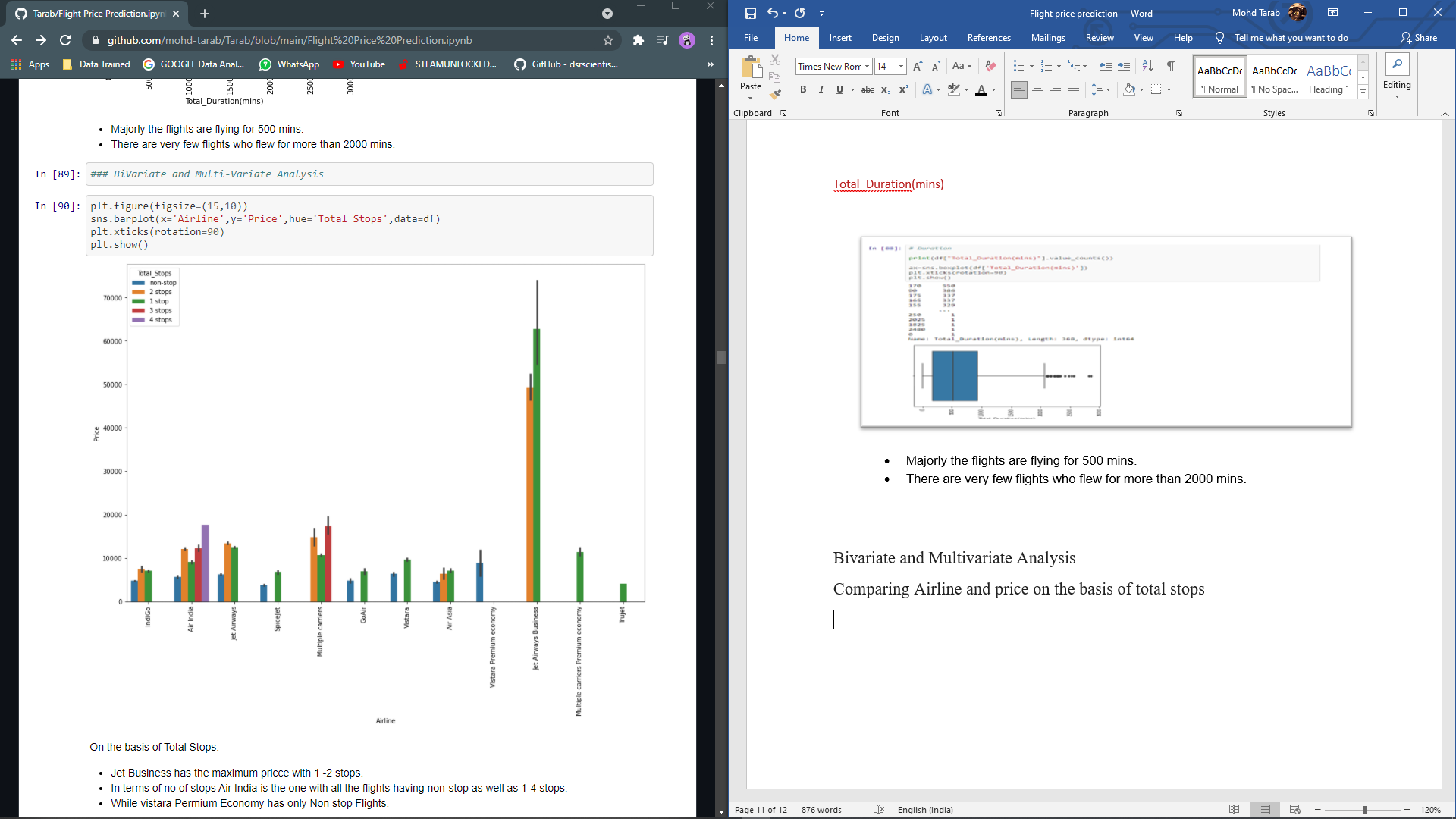


Majorly the flights are flying for 500 mins.

There are very few flights who flew for more than 2000 mins.

**Bivariate and Multivariate Analysis**

Comparing Airline and price on the basis of total stops

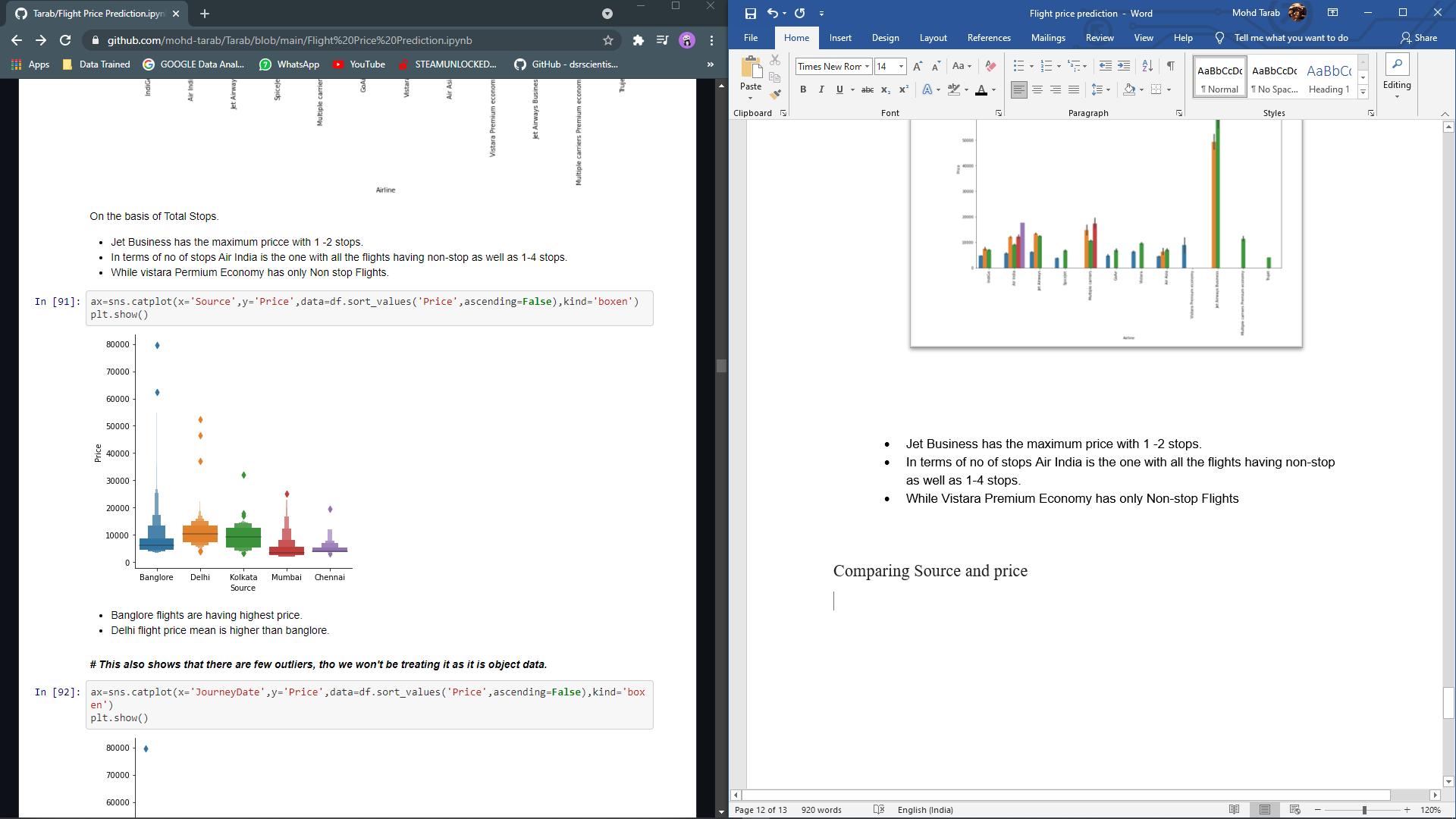


Jet Business has the maximum price with 1 -2 stops.

In terms of no of stops Air India is the one with all the flights having non-stop as well as 1-4 stops.

While Vistara Premium Economy has only Non-stop Flights

Comparing Source and price

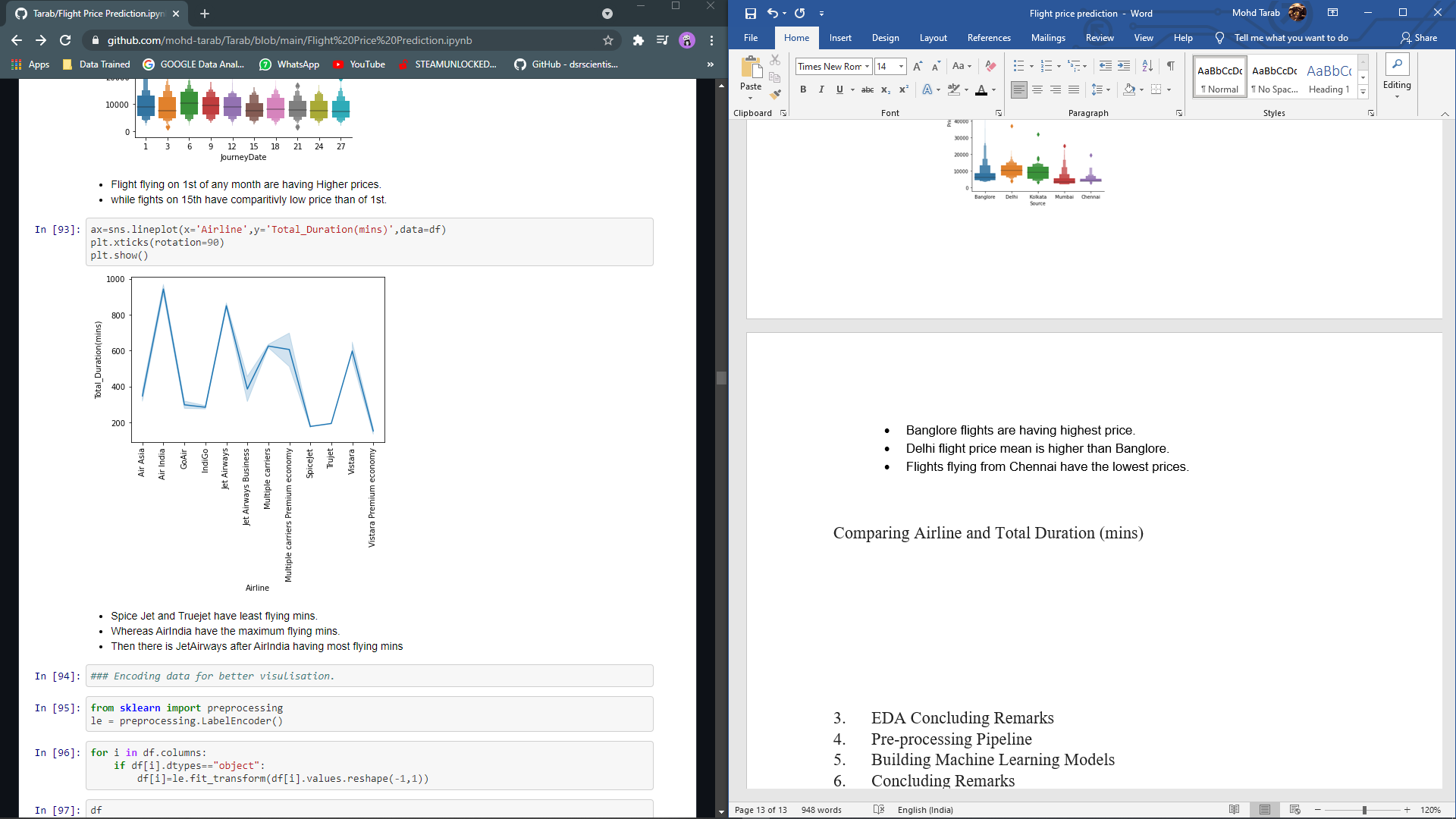


Banglore flights are having highest price.

Delhi flight price mean is higher than Banglore.

Flights flying from Chennai have the lowest prices.

Comparing Airline and Total Duration (mins)



Spice Jet and True jet have least flying mins.

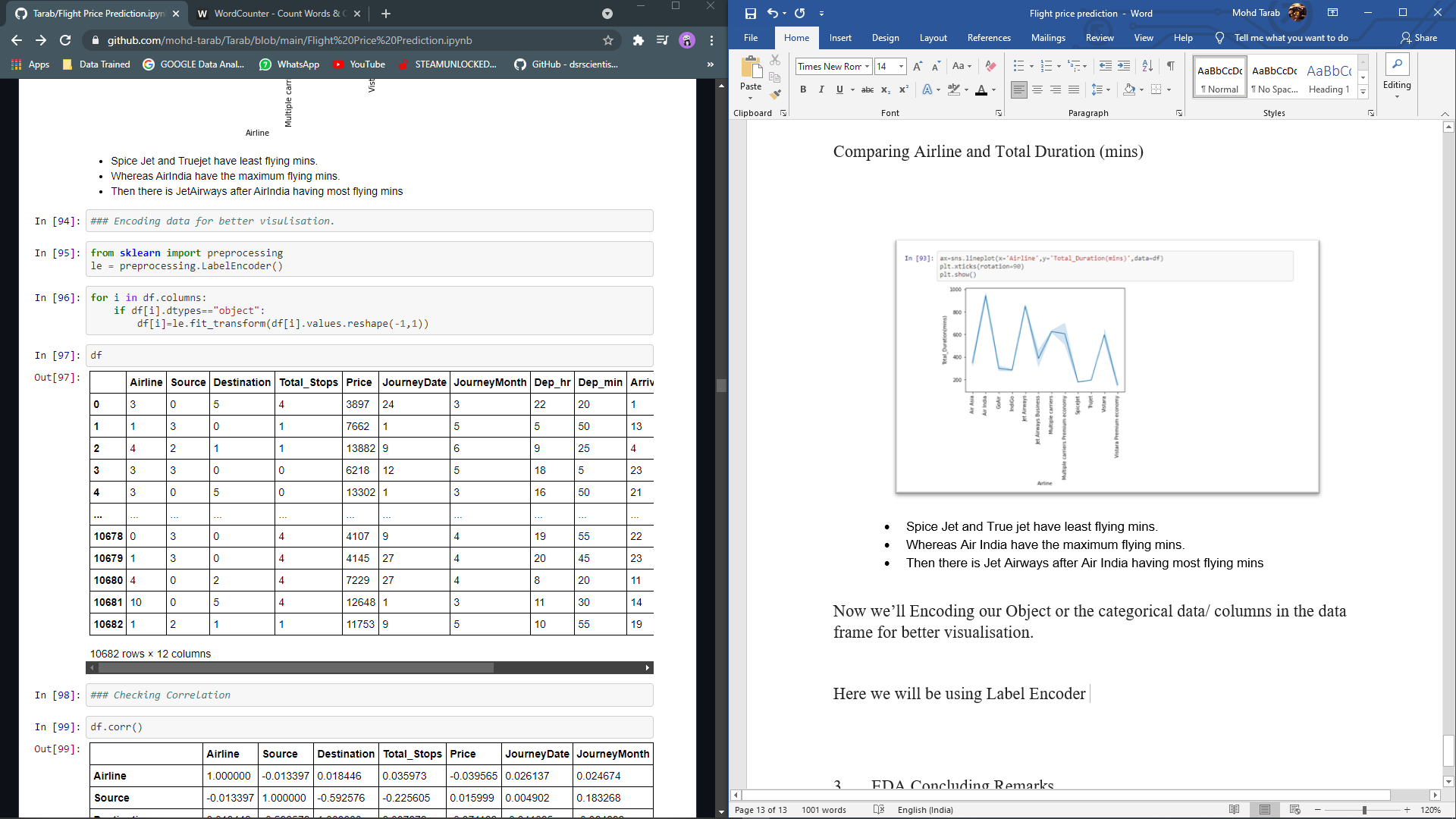
Whereas Air India have the maximum flying mins.

Then there is Jet Airways after Air India having most flying mins

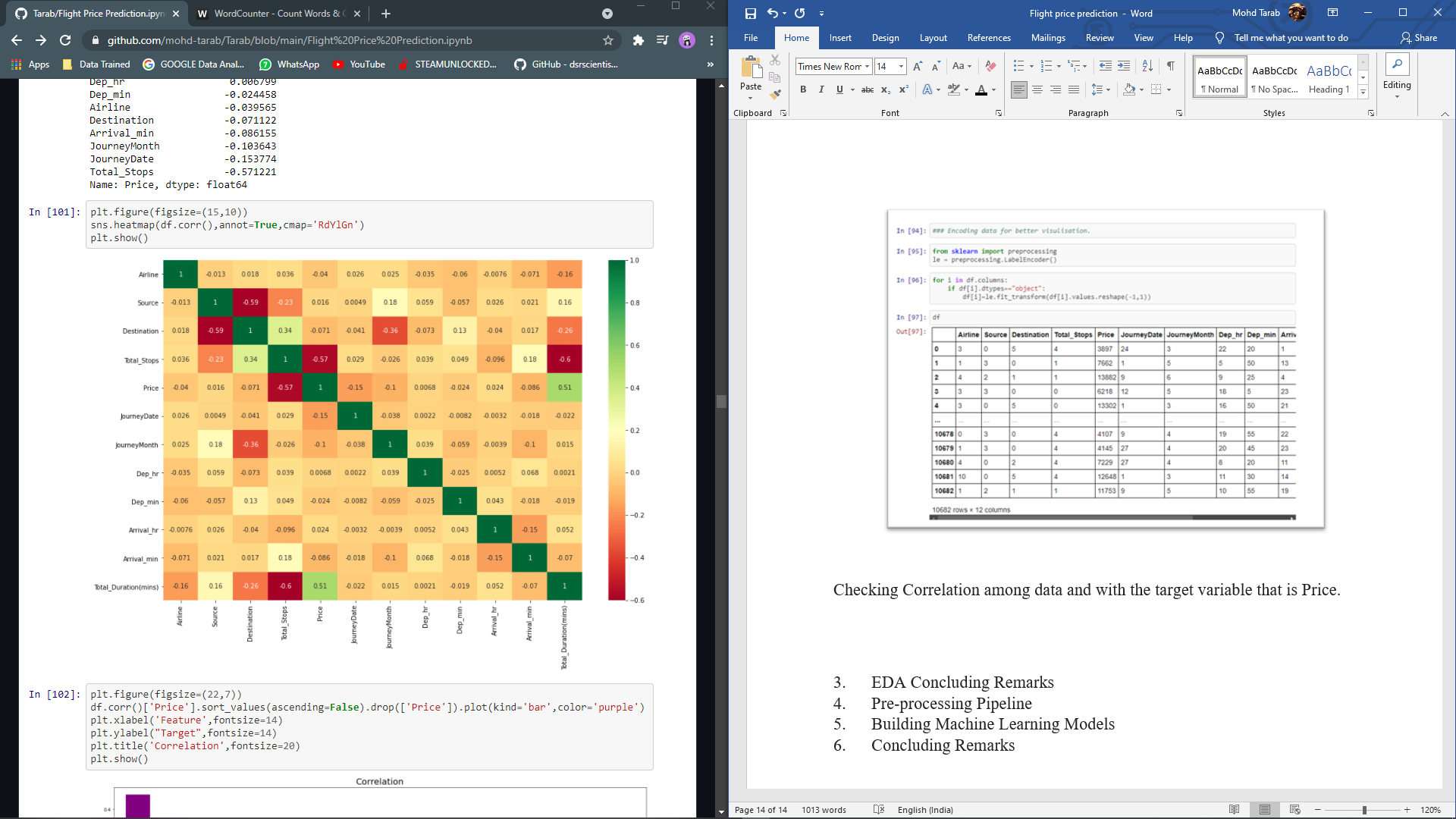
Now we’ll Encoding our Object or the categorical data/ columns in the data frame for better visualisation.

Here we will be using Label Encoder

The Label Encoder class is used to convert categorical text data into model-understandable numerical data. To label encode a column, we simply import the LabelEncoder class from the sklearn library, fit and transform the data column, and then update the original text data with the new encoded data.



Checking Correlation among data and with the target variable that is Price.



Total Duration is most correlated with the price, which means the more the flying hours the more the price will be.

while Total Stops is negatively corelated, which means the more the stops the more the price.

Journey Date is also a factor which is inversely correlated with price.

**EDA Concluding Remarks**

Here in EDA, we did univariate analysis, bivariate analysis and multi variate analysis.

We found out few key insights, like

* Jet Airways is the most popular airline, While Indigo is the second most popular option for passengers.
* Trujet is the cheapest option.
* The majority of flights departing from Delhi. Then there's Kolkata, which has the fewest departing flights.
* The most popular destination is Cochin, with Bangalore ranking second.
* There has been at least one stoppage in 5625 fights. Non-stop flights are given second priority. There is only one flight with four stops.
* The 9th is the most preferred date. The 18th is the least preferred date.
* May and June are the busiest months for travel, while April is the least.
* The majority of (915) fights begin at 9 a.m.
* At 3 a.m., only 24 flights take off, making it the shortest dep time.
* The majority of flights arrive (land) at 1900 hrs, i.e. (7). There are only three flights that arrive at 3 a.m.
* The majority of the flights are 500 minutes long. There are very few flights that lasted over than 2000 minutes.
* With one or two stops, Jet Business is the most expensive option.
* In terms of the number of stops, Air India is the one, with all flights having both non-stop and 1-4 stops.
* Vistara Premium Economy offers only direct flights.
* Flights to Bangalore are the most expensive. The average flight price from Delhi to Bangalore is higher.
* Flights departing on the first of any month have higher prices. While fights on the 15th are less expensive than those on the 1st.
* Spice Jet and Truejet have the shortest flight times. While Air India has the most flying minutes.
* Then there's Jet Airways, which comes in second to Air India in terms of flying minutes.
* Total Duration is the most closely related to price.
* While Total Stops is negatively related, which means that the more stops there are, the higher the price.

**Pre-processing Pipeline**

In pre-processing pipeline, we are going to check the skewness and outliers then splitting our dataset into two features and target. Later applying a scaling method to scale down data.

Checking Outliers.

In this, we are checking for any value in respective columns that is an outlier and is affecting our data set.

We found out there are outliers in these columns...

Airline

Price

Total Duration

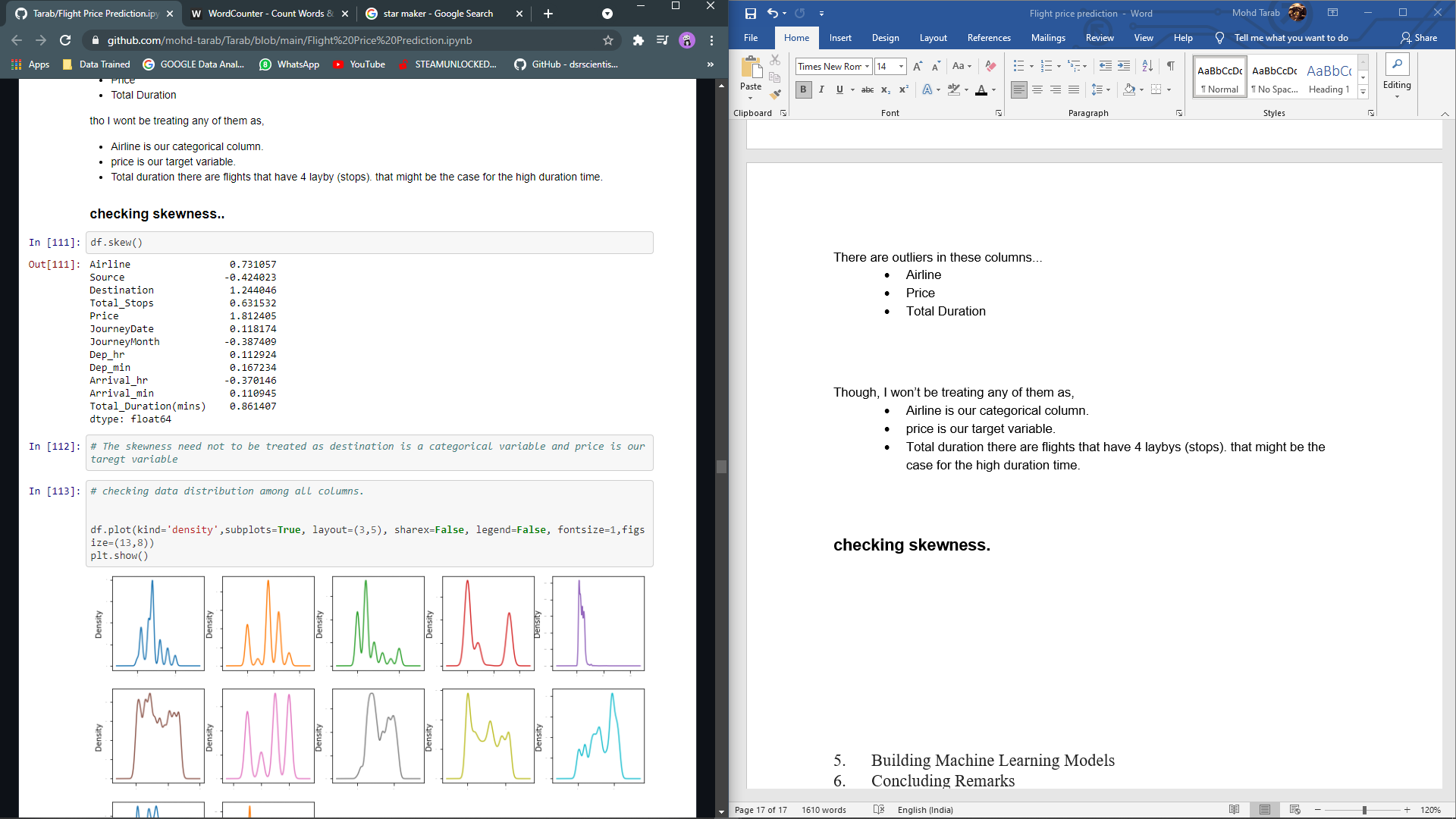
Though, I won’t be treating any of them as,

Airline is our categorical column.

price is our target variable.

Total duration there are flights that have 4 laybys (stops). that might be the case for the high duration time.

Checking skewness.



Keeping the threshold value of (+-0.5) we have skewness in following columns

Airline

Destination

Total stops

Price

Total Duration

Though we are not going to treat skewness as

Airline is a categorical column

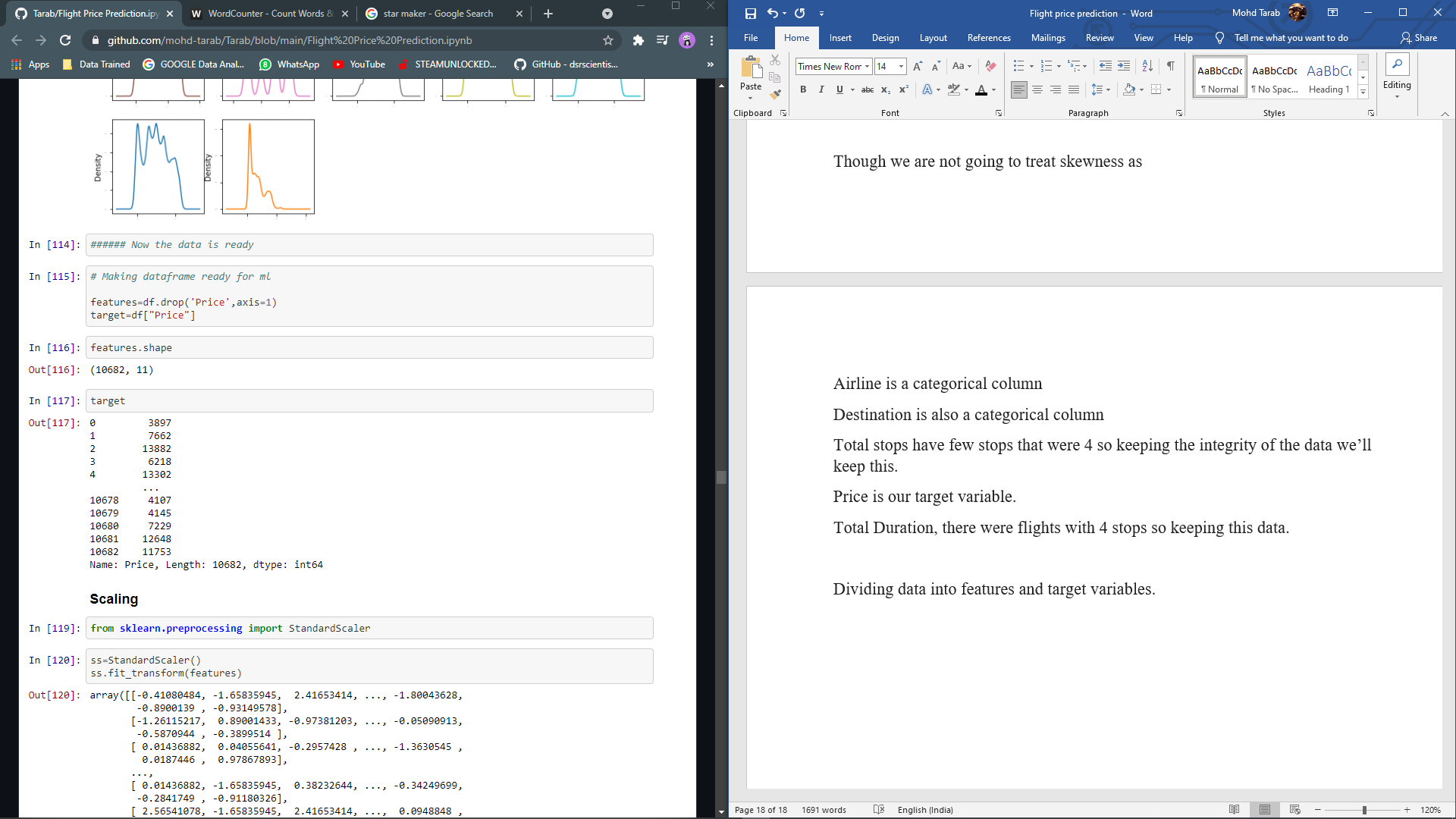
Destination is also a categorical column

Total stops have few stops that were 4 so keeping the integrity of the data we’ll keep this.

Price is our target variable.

Total Duration, there were flights with 4 stops so keeping this data.

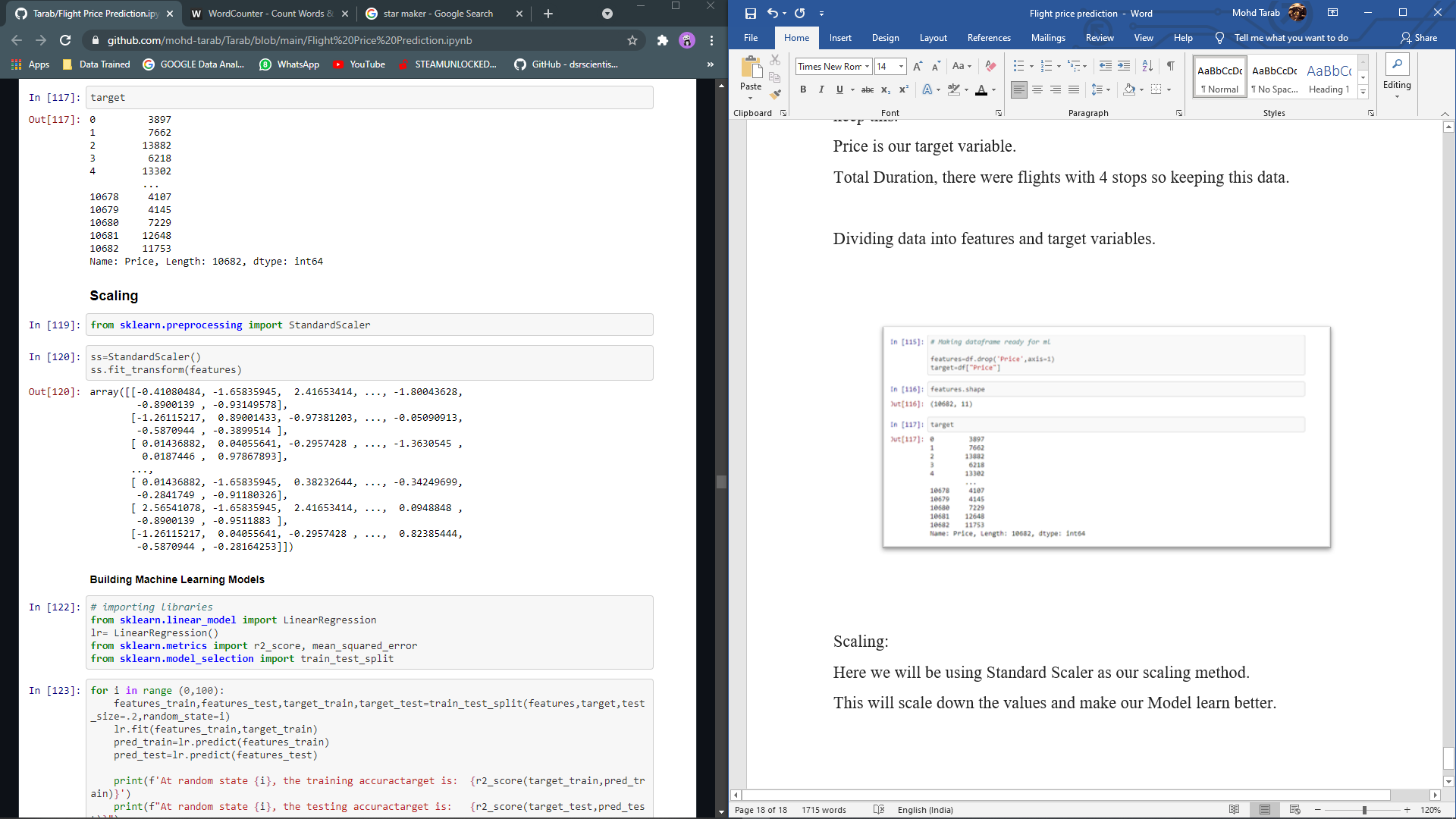
Dividing data into features and target variables.



Scaling:

Here we will be using Standard Scaler as our scaling method. Scaling is needed Because the values in each column vary greatly, scaling is required.

This will scale down the values and make our Model learn better.



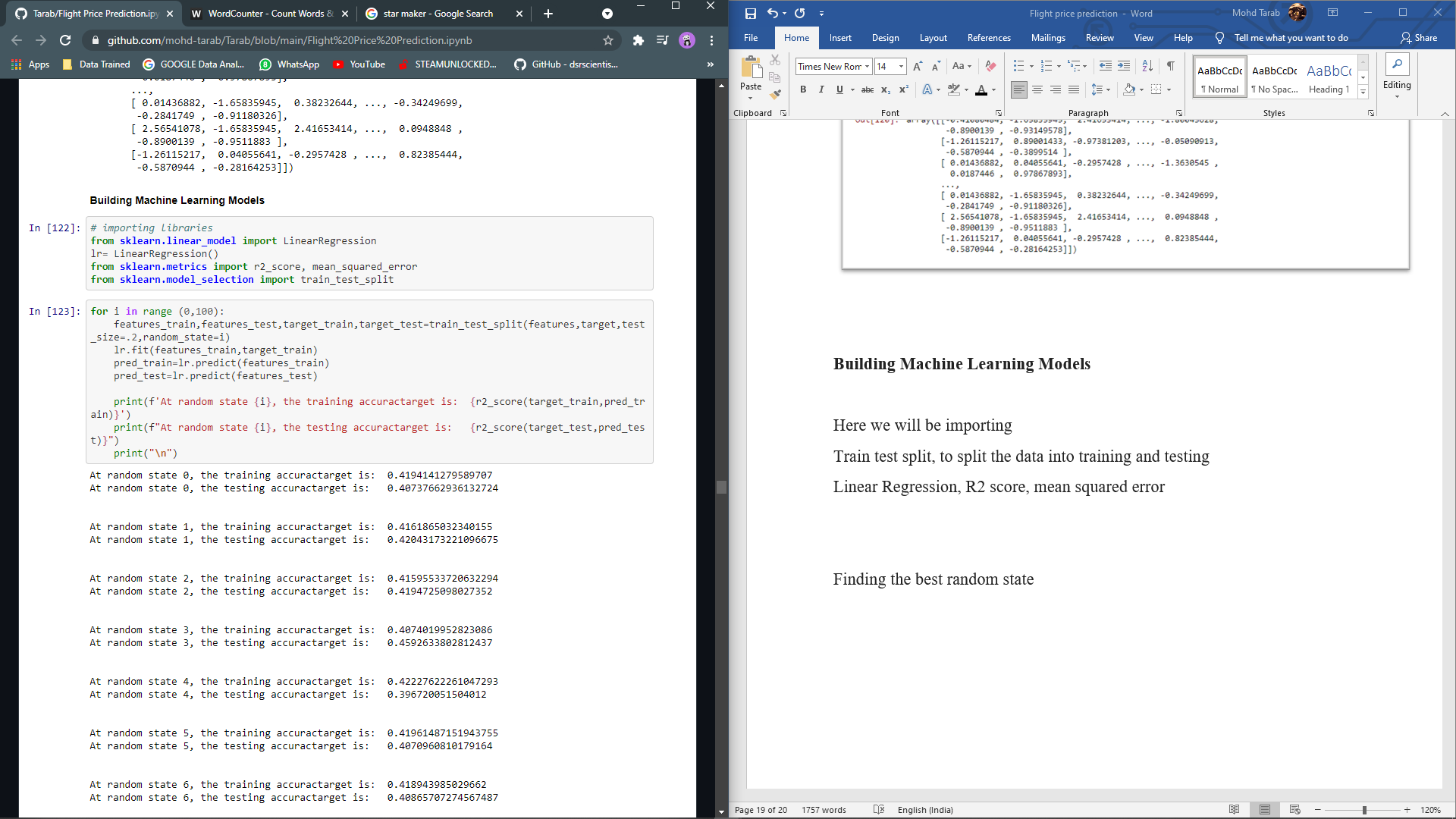
**Building Machine Learning Models**

Here we will be importing

Train test split, to split the data into training and testing

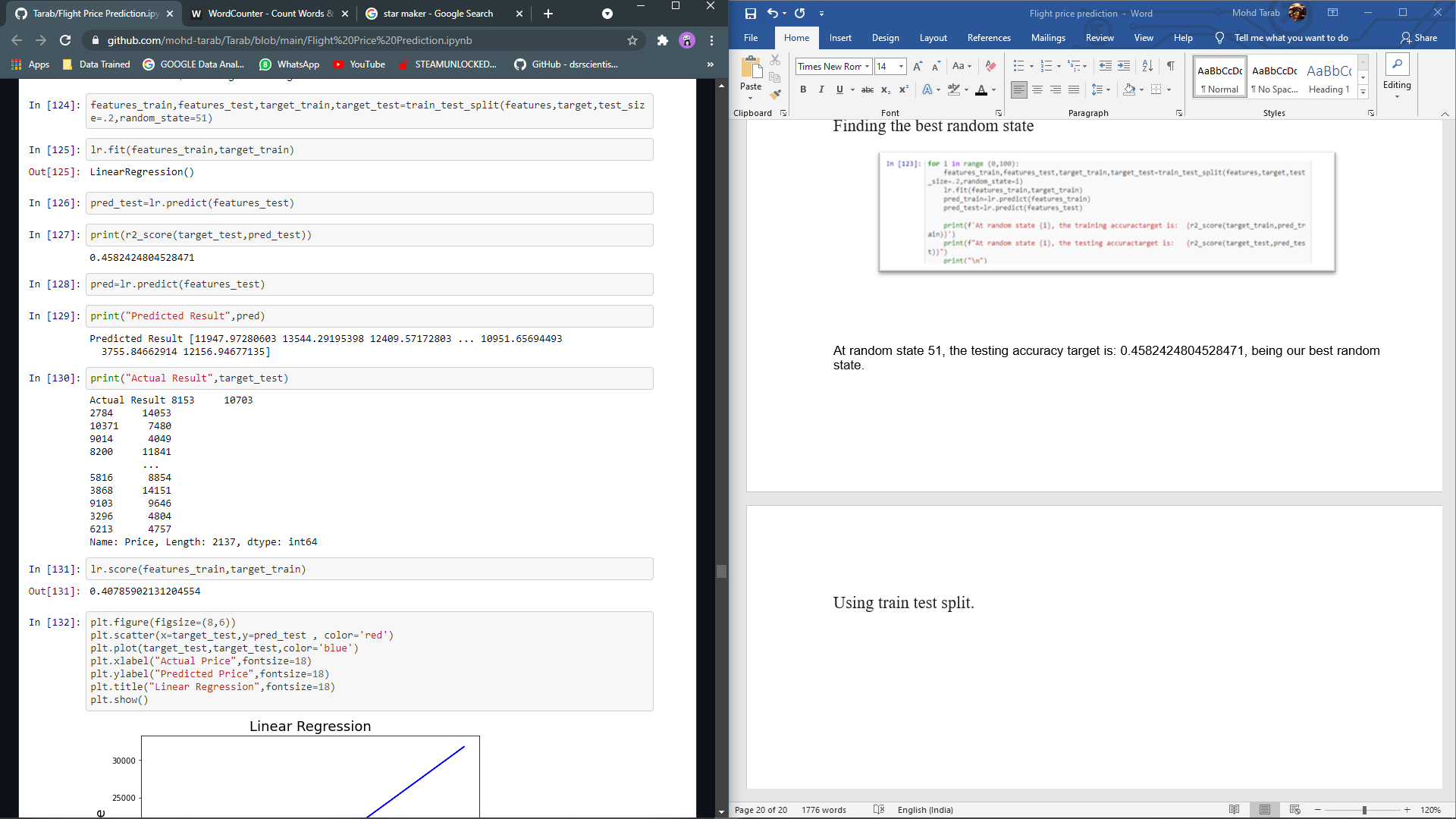
Linear Regression, R2 score, mean squared error

Finding the best random state

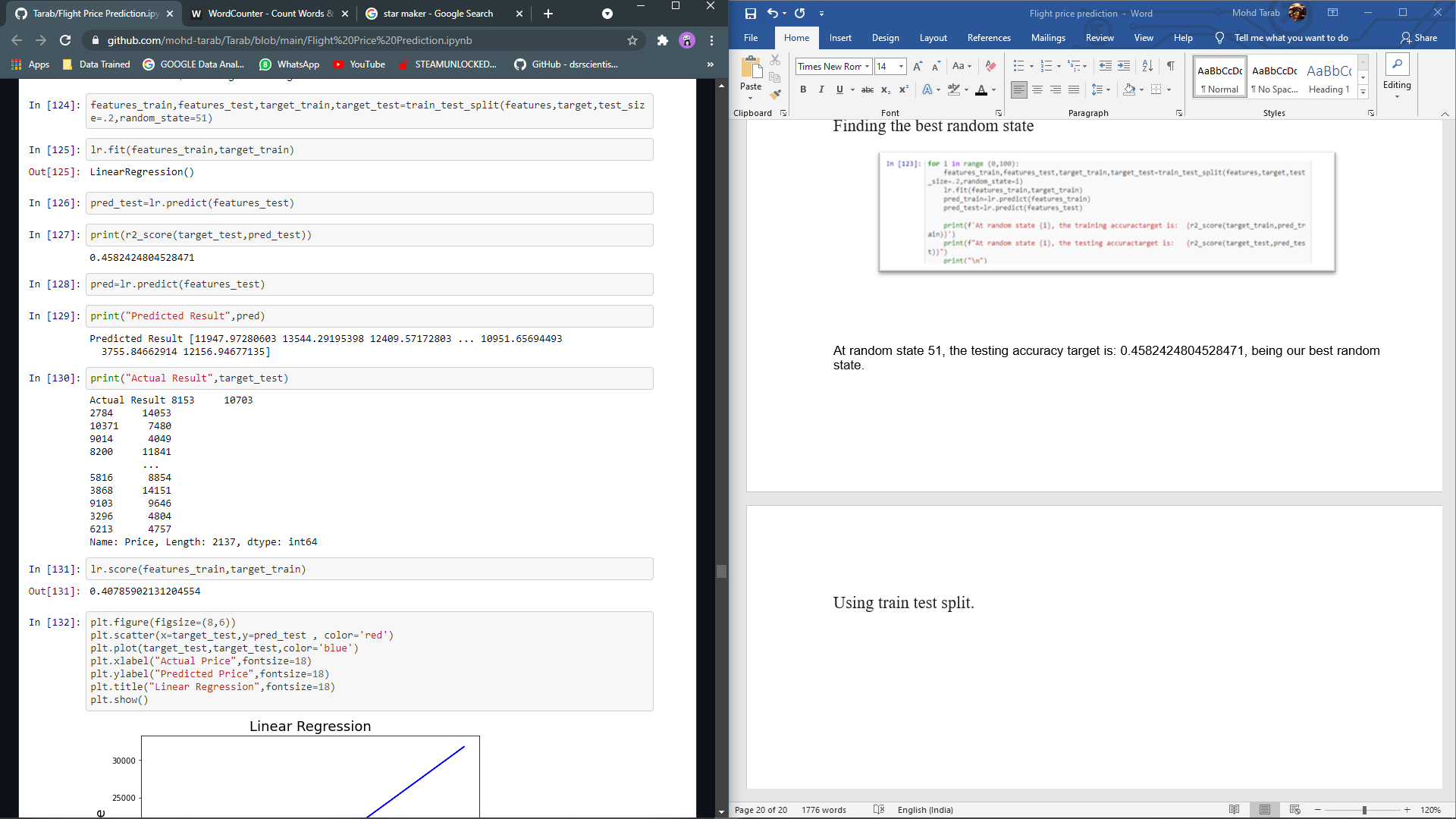


At random state 51, the testing accuracy target is: 0.4582424804528471, being our best random state.

Using train test split.

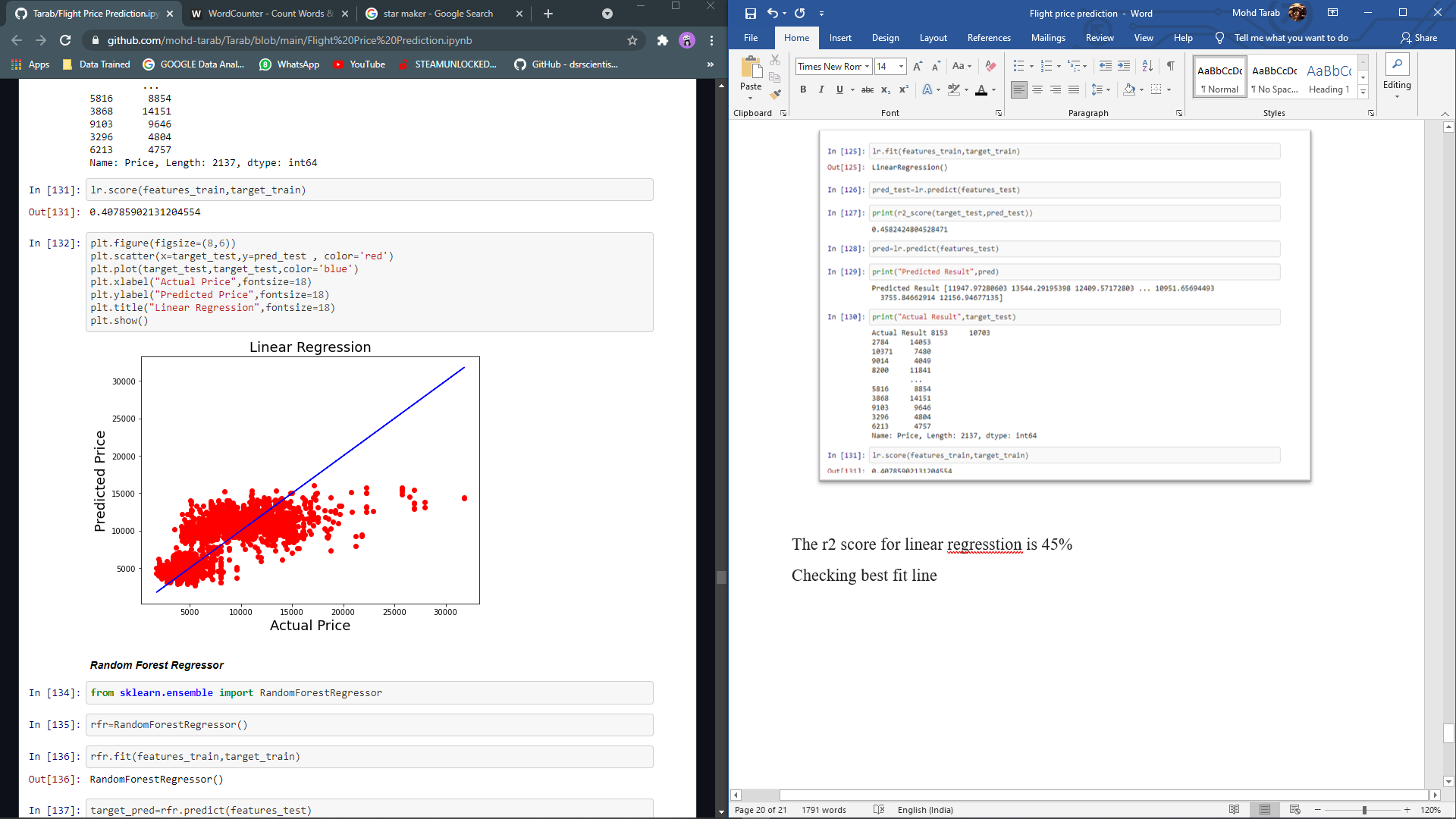


Using Linear Regression

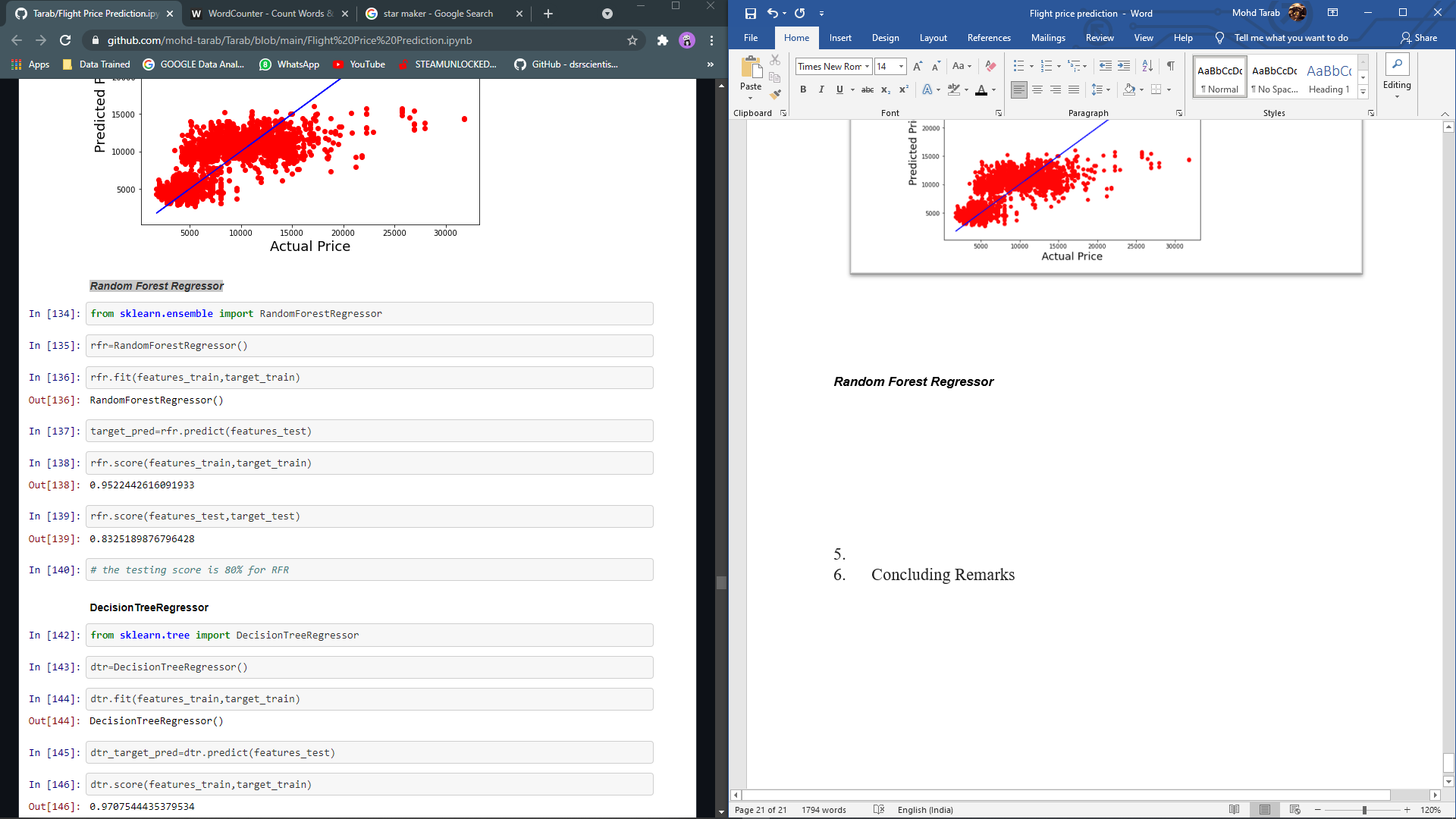


The r2 score for linear regresstion is 45%

Checking best fit line

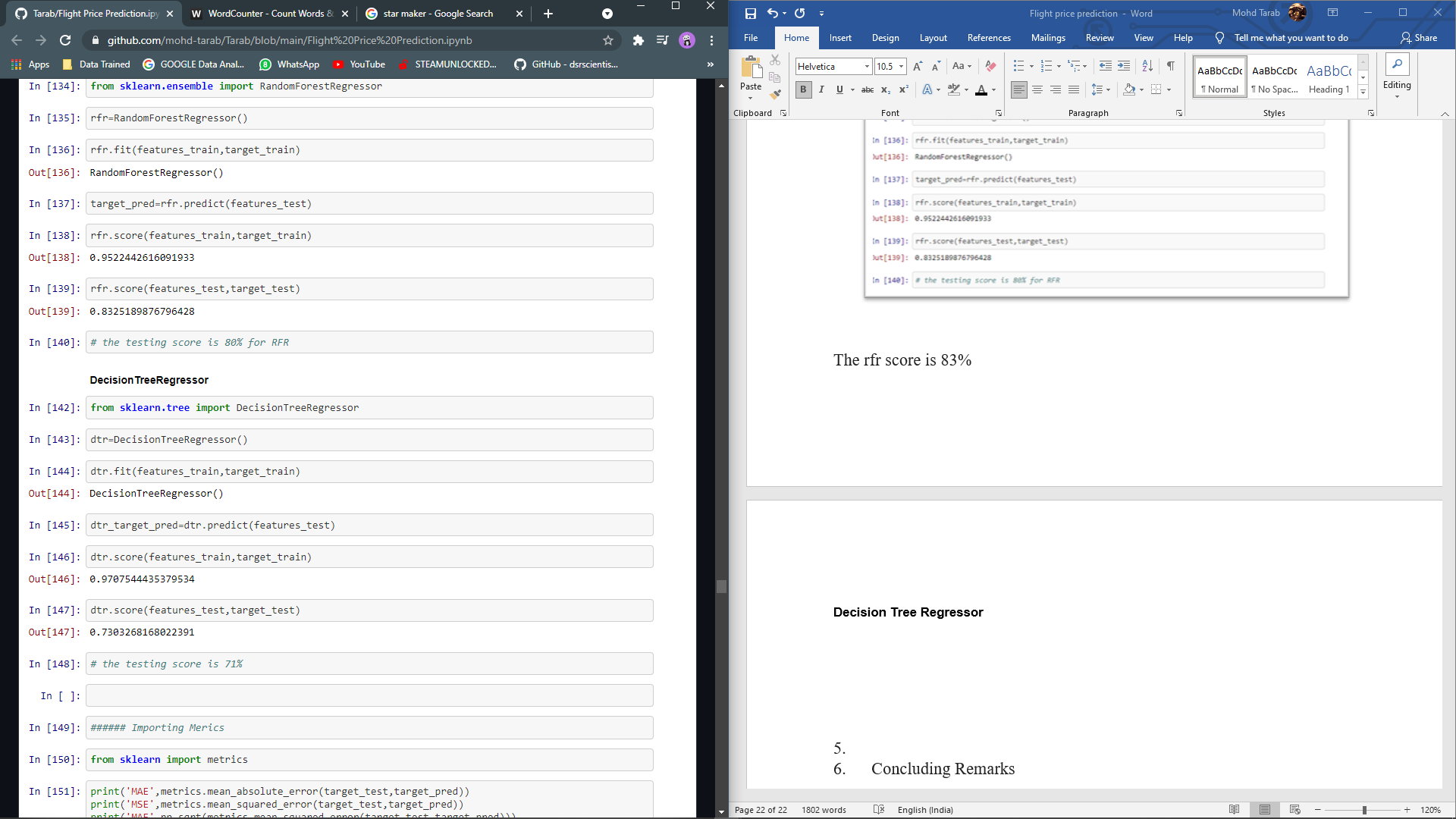


Random Forest Regressor



The rfr score is 83%

Decision Tree Regressor

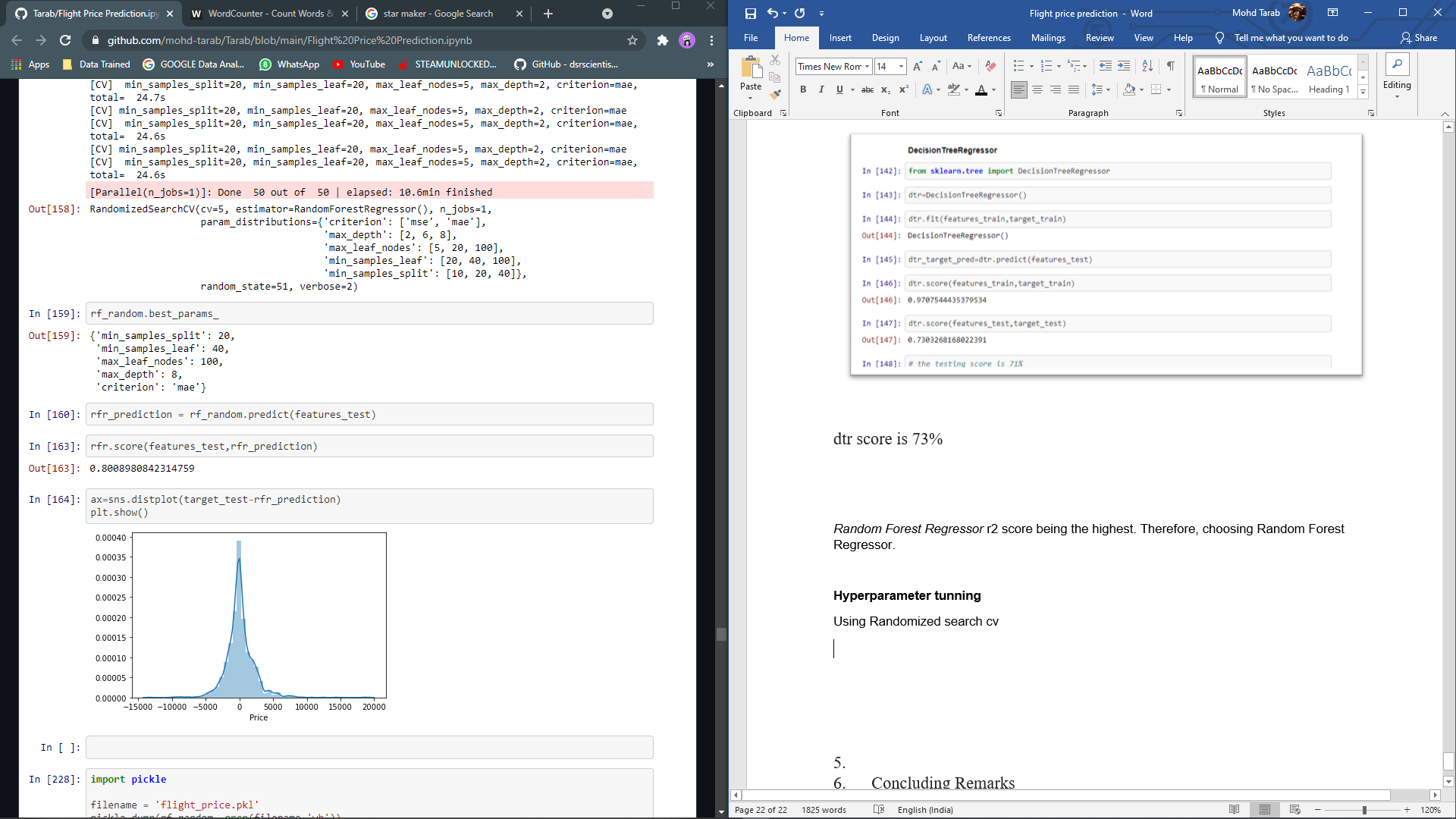


dtr score is 73%

Random Forest Regressor r2 score being the highest. Therefore, choosing Random Forest Regressor.

Hyperparameter tunning

Using Randomized search cv



The prediction score of randomized search is 80% and the by plotting the graph we see a gaussian like structure, this shows that the model is not bias.

**Saving the model, it is always a smart option to save the learned model because in the real world, only saved object files will be needed and called from websites, etc. to predict new data.**

Later we will apply the same steps for test data set.

**Concluding Remarks**

In this case study, a Machine Learning model is built to forecast airline fares. Several features were extracted from the dataset and combined with the help of Machine Learning to predict flight prices. With the assistance of the aforementioned techniques, the proposed model can predict the flight fare with an adjusted R squared score of 80 percent.

The most important thought in this type of problem is Feature Engineering. You can see how we handled categorical and numerical data, as well as how we built various ML models on the same dataset. We also examine each model's RMSE score to determine how it should perform in our test dataset.

However, there are still opportunities for improvement in this model.

If we get some information like seat location, when the ticket was purchased, special occasion on the departure date, and so on, our model will be able to predict the flight fare more accurately in the future.