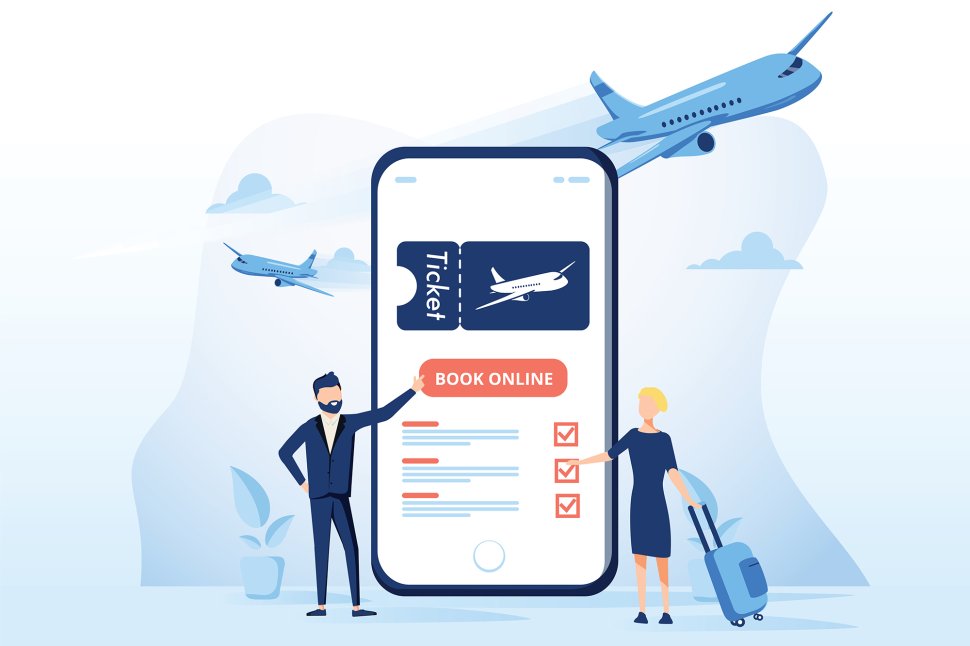
**TECHNOCOLABS DATA ANALYSIS INTERNSHIP**

**PROJECT REPORT**

**TITLE: Flight Prices Prediction Using Machine Learning**



**AIM:**

The principle focus of our project is to perform data analysis and train a model using the most popular Machine Learning algorithm –Random forest in order to analyse the historical data that is present regarding price prediction of flight tickets by analysing the optimal features.

**ABSTRACT:**

Someone who purchase flight tickets frequently would be able to predict the right time to procure a ticket to obtain the best deal. Many airlines change ticket prices for their revenue management. The airline may increase the prices when the demand is to be expected to increase the capacity. To estimate the minimum airfare, data for a specific air route has been collected including the features like departure time, arrival time and airways over a specific period. Features are extracted from the collected data to apply Data Modelling technique – Random forest Algorithm.

**INTRODUCTION:**

The flight ticket buying system is to purchase a ticket many days prior to flight takeoff so as to stay away from the effect of the most extreme charge. Mostly, aviation routes don’t agree this procedure. Plane organizations may diminish the cost at the time, they need to build the market and at the time when the tickets are less accessible. They may maximize the costs. So, the cost may rely upon different factors. To foresee the costs this venture uses AI to exhibit the ways of flight tickets after some time. All organizations have the privilege and opportunity to change its ticket costs at any time. Explorer can set aside cash by booking a ticket at the least costs. People who had travelled by flight frequently are aware of price fluctuations. The airlines use complex policies of Revenue Management for execution of distinctive evaluating systems. The evaluating system as a result changes the charge depending on time, season, and festive days to change the header or footer on successive pages. The ultimate aim of the airways is to earn profit whereas the customer searches for the minimum rate. Customers usually try to buy the ticket well in advance of departure date so as to avoid hike in airfare as date comes closer. But actually, this is not the fact. The customer may wind up by giving more than they ought to for the same seat.

**OVERVIEW:**

* Data Segmentation and Data Cleaning
* Exploratory Data Analysis using python’s data visualisation libraries
* Training the model based on the historical data available.

**DATASET:**

The dataset we have considered, has information of different airlines with informative features of a flight that determine the prices of the air tickets effectively. This phase is very important since it defines the problem under solving.

Original format of the dataset: XLSX

A brief explanation of every column in the dataset is as follows:

➢ **Out Date –** The date of the journey

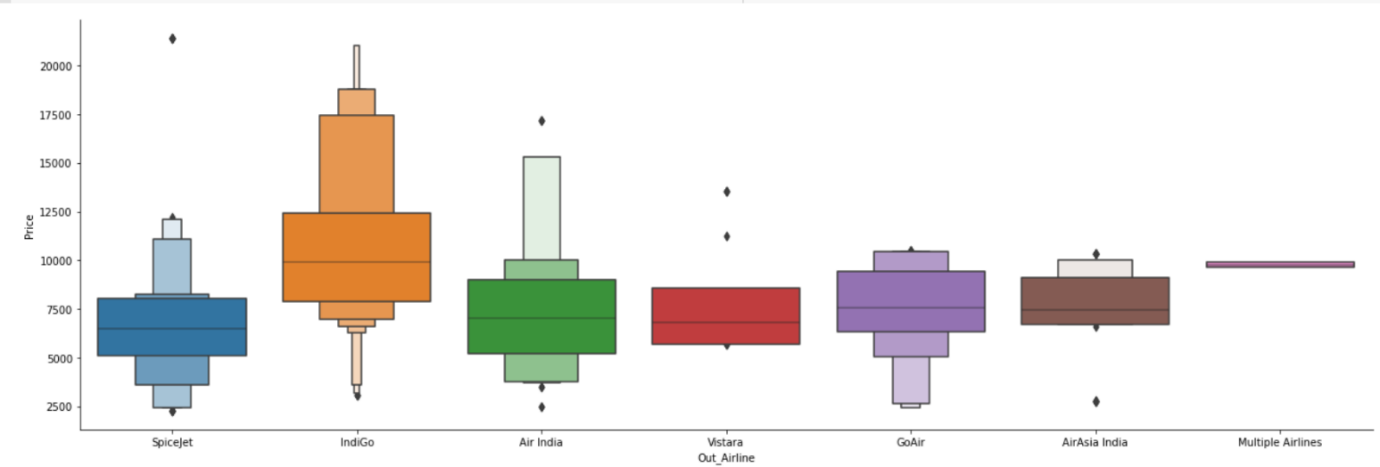
* **Out-Time –** The timestamp at which journey commences from the source
* **Out-Cities –** The source from which the service begins.
* **Out-Airline –** The type of airline chosen for the journey
* **Return-Date–** Date of return journey
* **Return Time-**The timestamp at which journey starts for return journey type.
* **Return Cities –** The destination from which the service/journey begins
* **Return Airline-** The type of airline chosen for return journey.
* **Out travel time-** Time taken to reach from source to destination.
* **Return travel time-** Time taken to reach from destination to source
* **Out Journey Type-** One stop or non- stop.
* **Return Journey Type-** One stop or non-stop
* **Out Stop Cities-** intermediate cities for one-stop type journey
* **Return Stop Cities-** intermediate cities for return one-stop type journey
* **Timestamp-**time stamp of the particular journey
* **Sort­**- best or cheap
* **Price** –The price of the ticket.

**DATA SEGMENTATION AND DATA CLEANING:**

* In this project, we have prepared a processed dataset by and collected the clear-cut data available online.
* Using pandas data frame, we have calculated the mean of every column.
* By using the dropna we have dropped unnecessary columns.
* While doing so we have dropped the columns which have unique values.
* Performed encoding for ordinal and nominal data

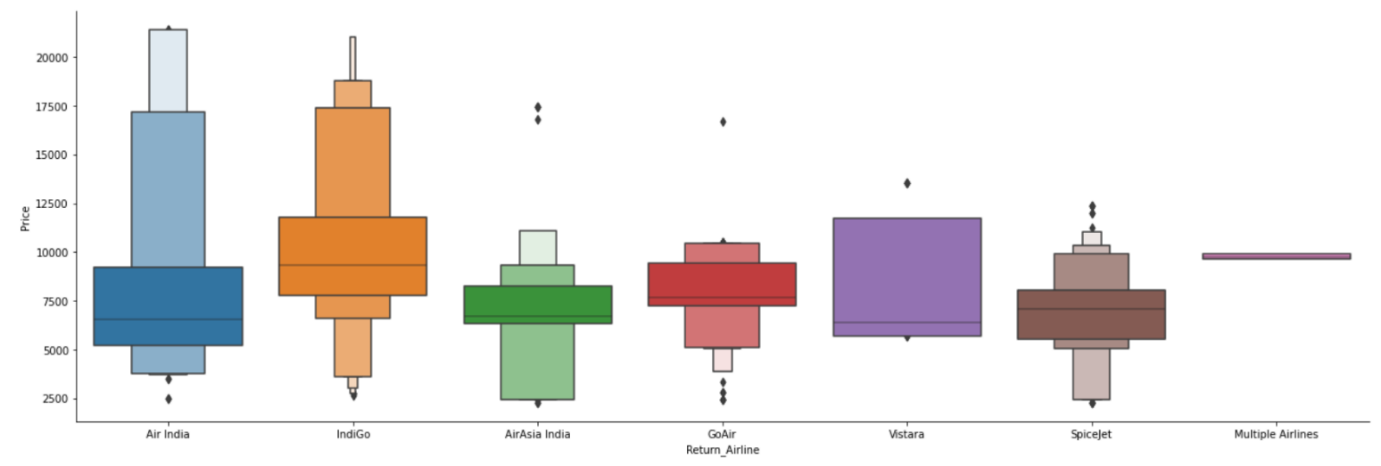
**EXPLORATORY DATA ANALYSIS:**

# ➢ PLOT 1



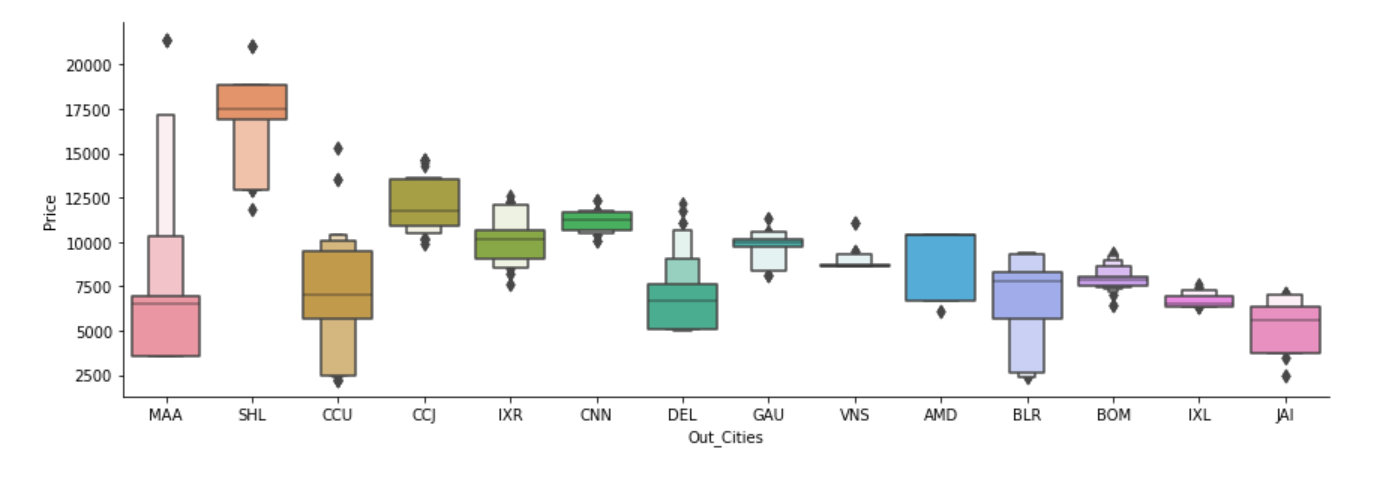
The above plot represents out Airline vs price

# ➢ PLOT 2



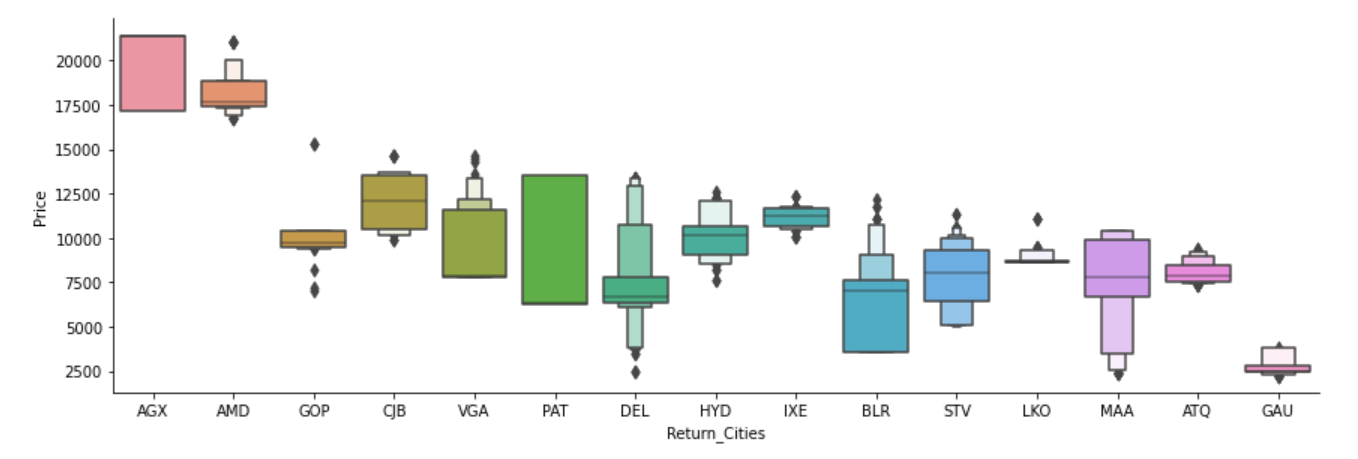
The above plot represents return Airline vs price

# ➢ PLOT 3



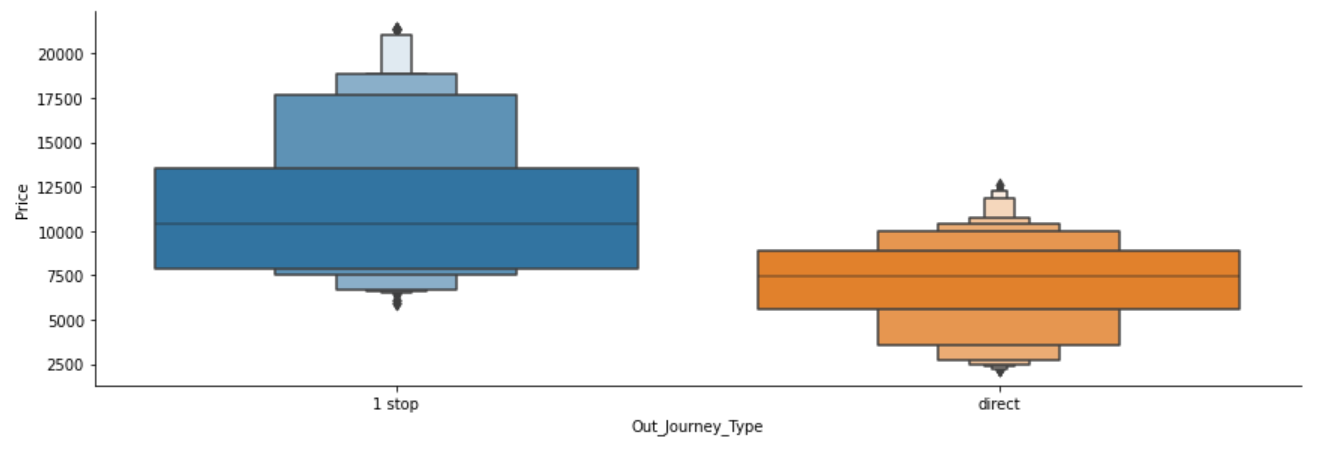
The above plot out cities vs price

# ➢ PLOT 4



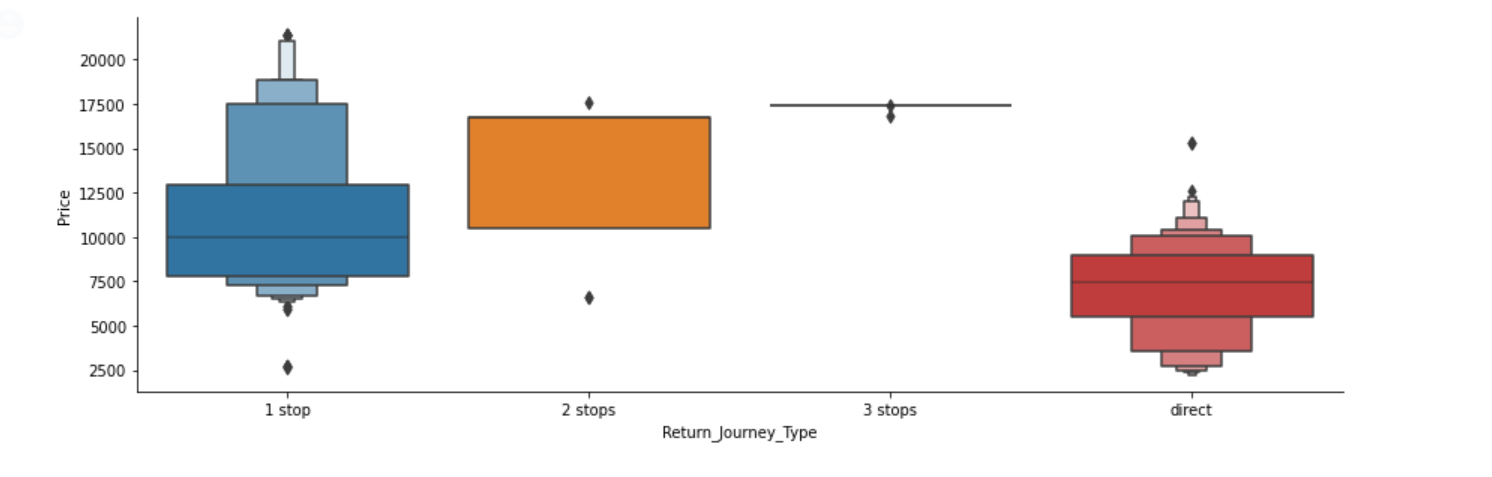
The above plot represents return cities vs price.

# ➢ PLOT - 5



The above interactive plot represents out journey type vs price

# ➢ PLOT 6



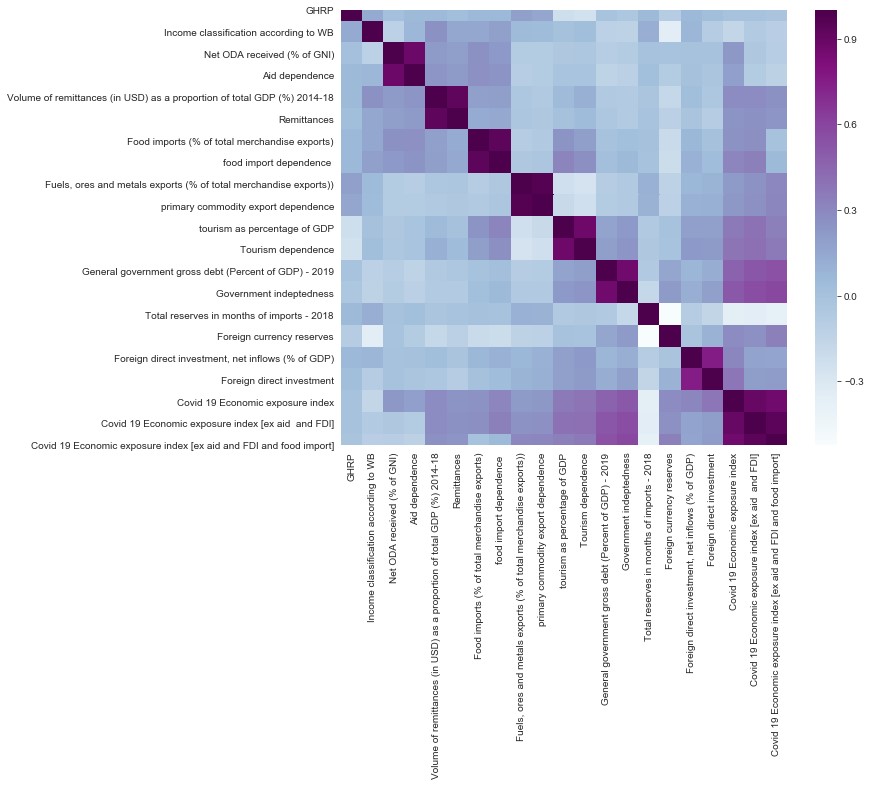
The above plot represents return journey type vs price

**TRAINING THE MODEL:**

Globalization and the rapid integration of markets due the COVID-19 pandemic have a prominent effect on the EEI which aims to reflect the performance of the companies with significant exposure to specific regions or countries, regardless of their domicile.

Our project aims to train 3 models for -

1. COVID-19 EEI
2. COVID-19 EEI [ex aid and FDI]
3. COVID-19 [ex aid, FDI and Food imports]



We have used used correlation in order to select the columns to train the model and columns which had a correlation of 0.2 and above were used.

# ➢ MODEL FOR COVID-19 EEI

The correlated columns considered for this model are as follows:

ODA received (% of GNI)','Volume of remittances (in USD) as a proportion of total GDP (%) 2014-18',

'Remittances',

'Food imports (% of total merchandise exports)',

'food import dependence ',

'Fuels, ores and metals exports (% of total merchandise exports))',

'primary commodity export dependence', 'tourism as percentage of GDP',

'Tourism dependence',

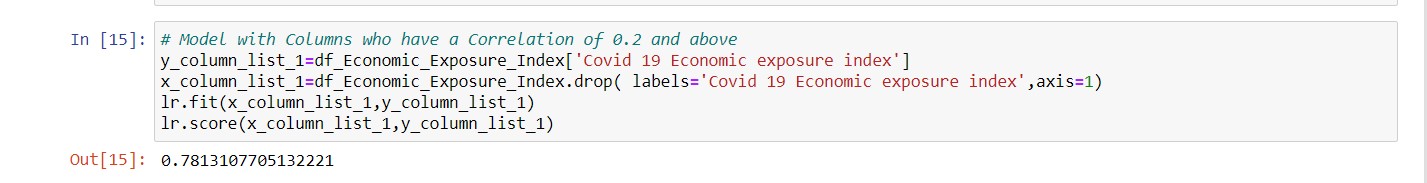
'General government gross debt (Percent of GDP) - 2019',

'Government indebtedness', 'Total reserves in months of imports - 2018',

'Foreign currency reserves',

'Foreign direct investment, net inflows (% of GDP)','Foreign direct investment’

Accuracy of the model:



# ➢ MODEL FOR COVID-19 EEI [EX AID & FDI]

The correlated columns considered for this model are as follows:

'Volume of remittances (in USD) as a proportion of total GDP (%) 2014-18',

'Remittances', 'Food imports (% of total merchandise exports)',

'food import dependence ',

'Fuels, ores and metals exports (% of total merchandise exports))',

'primary commodity export dependence', 'tourism as percentage of GDP',

'Tourism dependence',

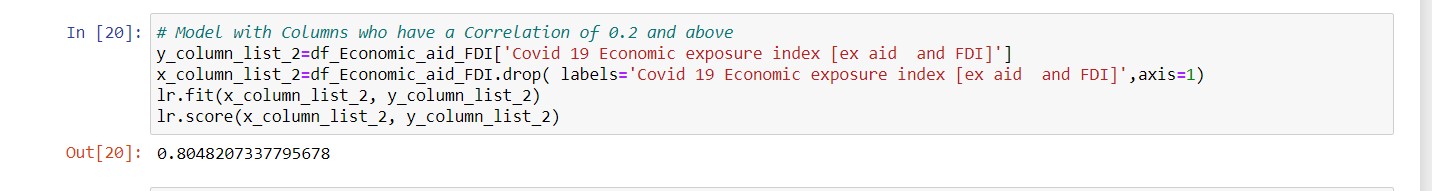
'General government gross debt (Percent of GDP) - 2019',

'Government indebtedness', 'Total reserves in months of imports - 2018',

'Foreign currency reserves',

'Foreign direct investment',' ‘Covid 19 Economic exposure index [ex aid and FDI]’

Accuracy of the model:



# MODEL FOR COVID-19 [EX AID, FDI & FOOD IMPORTS]

The correlated columns considered for this model are as follows:

'Volume of remittances (in USD) as a proportion of total GDP (%) 2014-18',

'Remittances', 'Food imports (% of total merchandise exports)',

'food import dependence ',

'Fuels, ores and metals exports (% of total merchandise exports))',

'primary commodity export dependence', 'tourism as percentage of GDP',

'Tourism dependence',

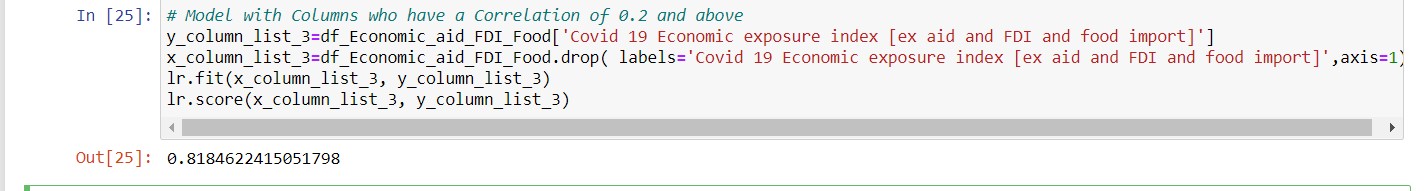
'General government gross debt (Percent of GDP) - 2019',

'Government indebtedness', 'Total reserves in months of imports - 2018',

'Foreign currency reserves',

'Foreign direct investment’, ‘Covid 19 Economic exposure index [ex aid and FDI and food import’

Accuracy of the model is as follows:



**Random Forest Regression**

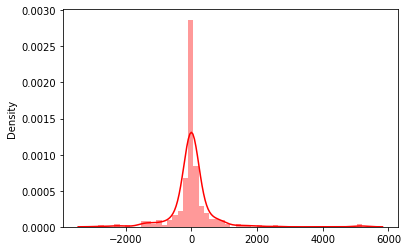
X = Clean\_data.drop(['Price'],axis=1).values

y = Clean\_data['Price'].values

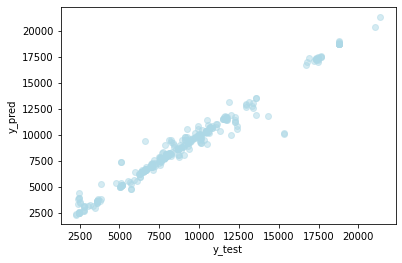
X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.3, random\_state=42)

X\_train.shape, X\_test.shape, y\_train.shape, y\_test.shape

**plot 1**



**Plot 2**



MAE: 303.2761510131435

MSE: 431690.134995819

RMSE: 657.0313044260669

metrics.r2\_score(y\_test, y\_pred)

0.9722960694759464

**Accuracy of our model is 0.972296**

**TEAM MEMBERS:**

# Snehal Kshirsagar,