**Project Name: Flight Price Prediction**

***-Analysis by Shravani Natakala***

**About Project:**

Flight tickets prices for various airlines from the months of March to June of 2019 also from various cities were collected. A machine learning model to predict the price of the flight ticket based on features such as Airline, Date of Journey, Dep Time, Arrival Time, Total Stops etc was built with the below mentioned process.

**Problem Definition:**

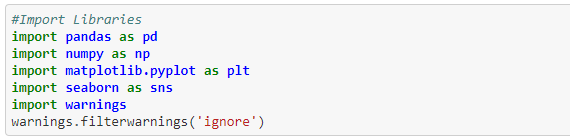
Flight ticket prices can be something hard to guess, today we might see a price, check out the price of the same flight tomorrow, it will be a different story. We might have often heard travellers saying that flight ticket prices are so unpredictable. Here you will be provided with prices of flight tickets for various airlines between the months of March and June of 2019 and between various cities.

Size of training set: **10683** records

Size of test set: **2671** records

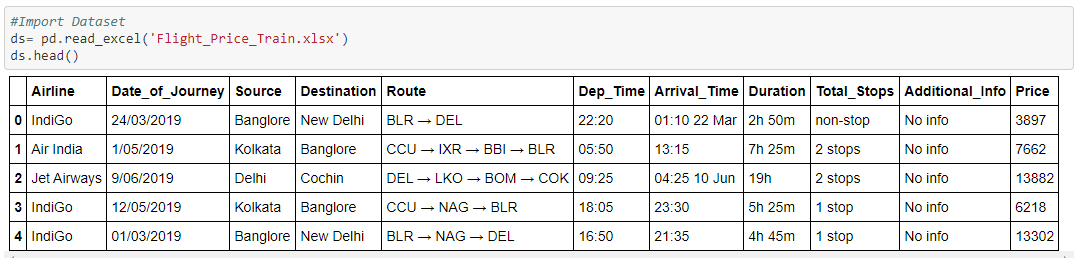
**Data Analysis**

* **Importing Libraries**



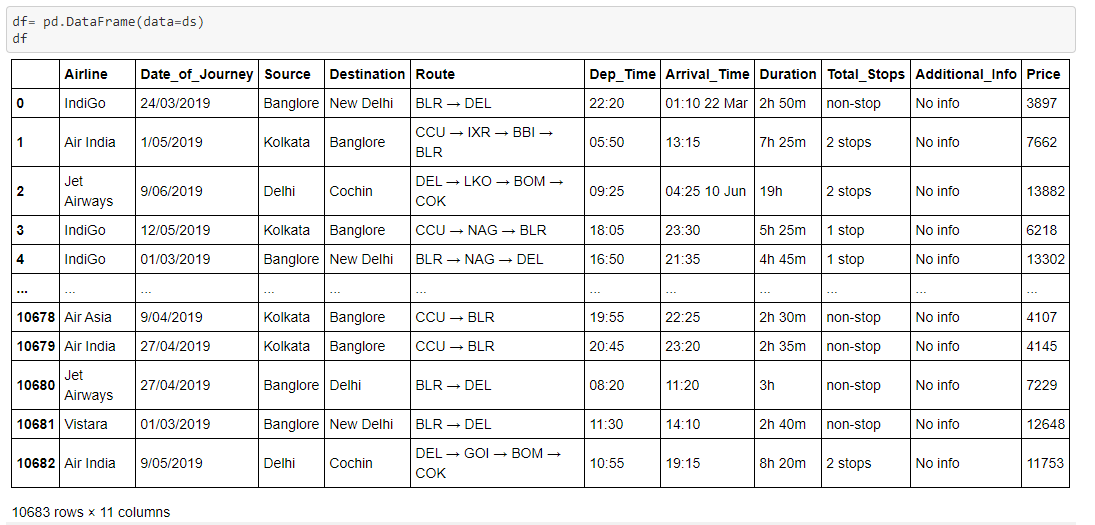
For proper EDA analysis and to import the required Data Libraries are imported.

* **Importing Dataset**



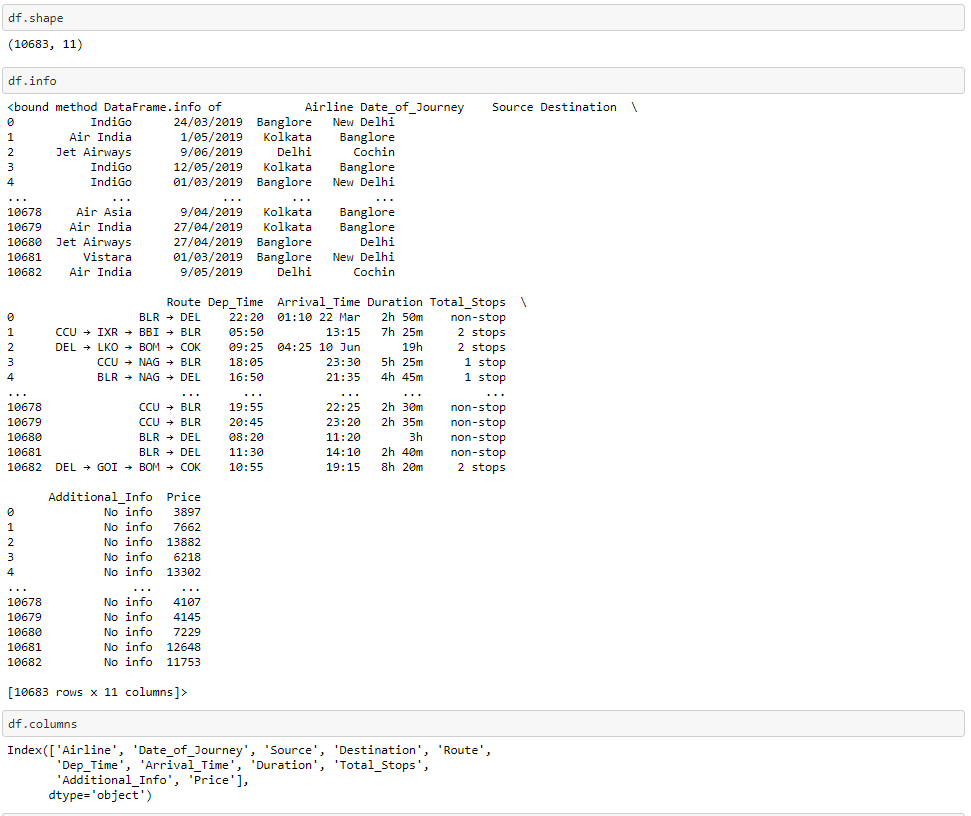
Particular Head() command is used to see the first 5 rows of the dataset.

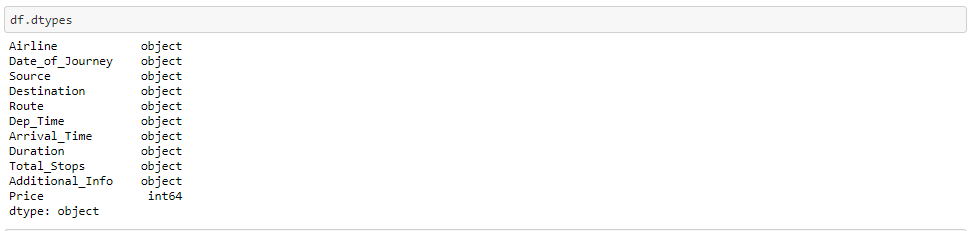
* **Reshaping data into DataFrame using Pandas library by organizing and providing index/column values appropriately.**



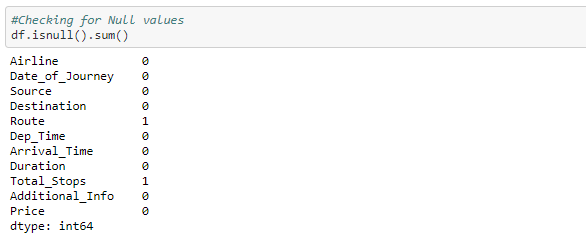
The dataset has been reshaped and organized as DataFrame which is mutable and could be used as required according to the project needs.

* **Appling shape attribute to find the shape and size of the DataFrame**



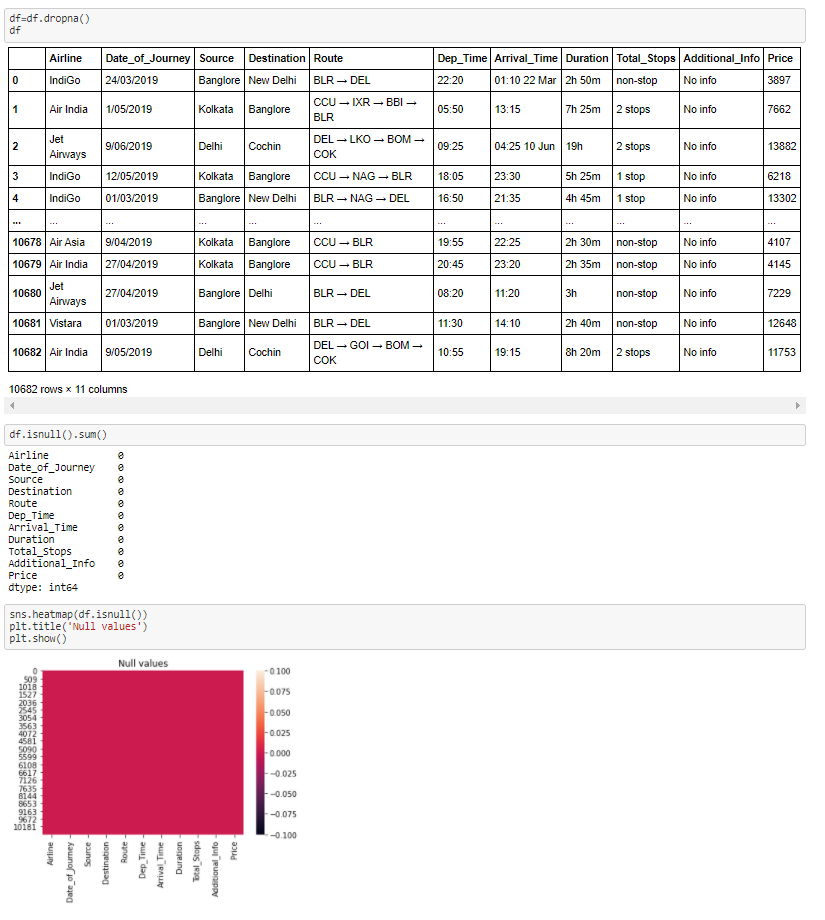


* **Detecting for Null values to find the missing values in the DataFrame**



Observations: Null values found in the Route and Total\_Stops columns where the values are shown as 1 which indicates that data is missing in those particular columns.

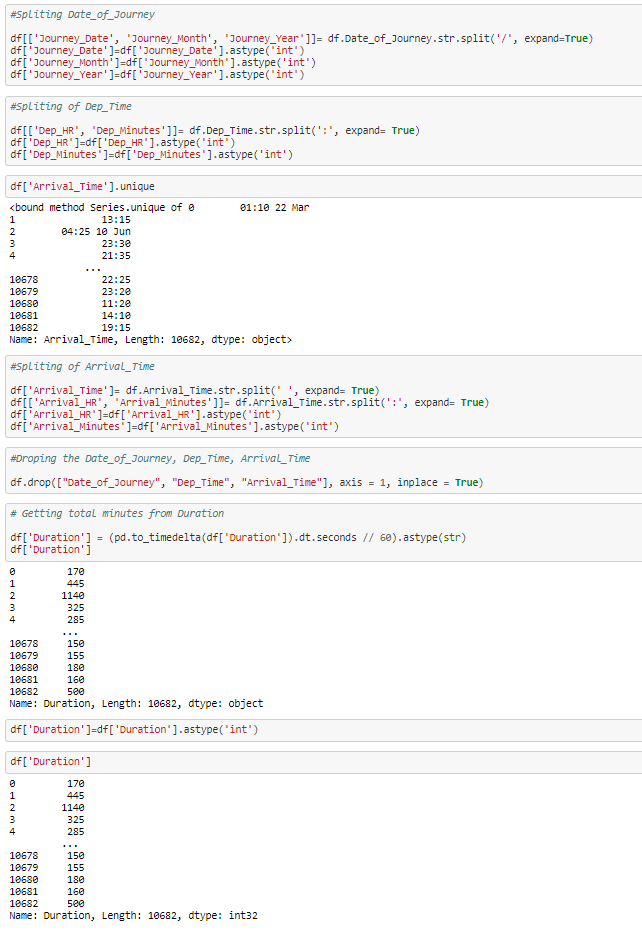
* **Dropping the columns which contains null/missing values**

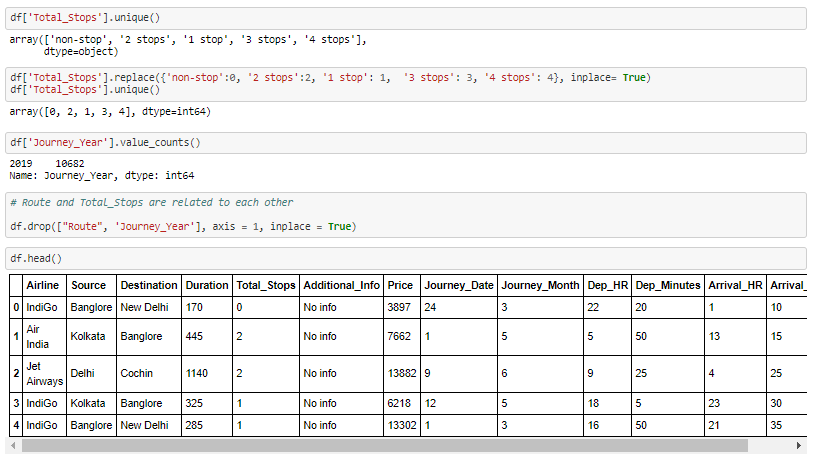
****

Observations: Dropped the columns that had missing data in the DataFrame.

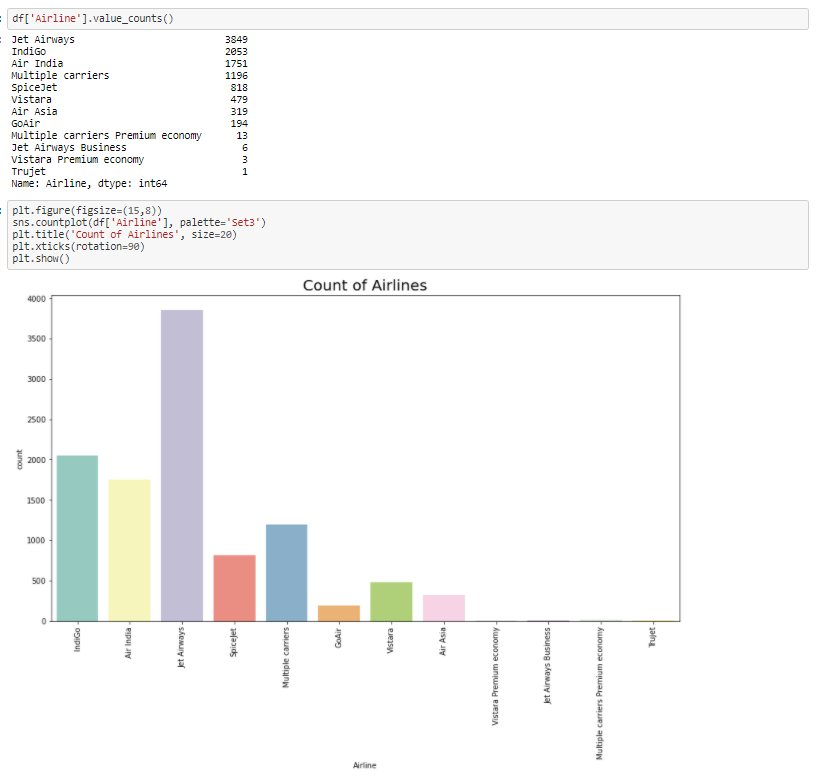
* **Using Feature Engineering technique to create new features from the existing data which helps to gain more insight into the data.**

Adding more features to the existing DataFrame and eliminating unwanted features.

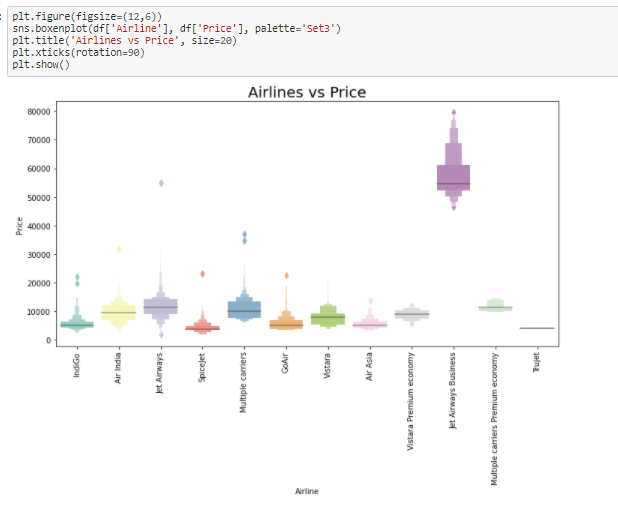




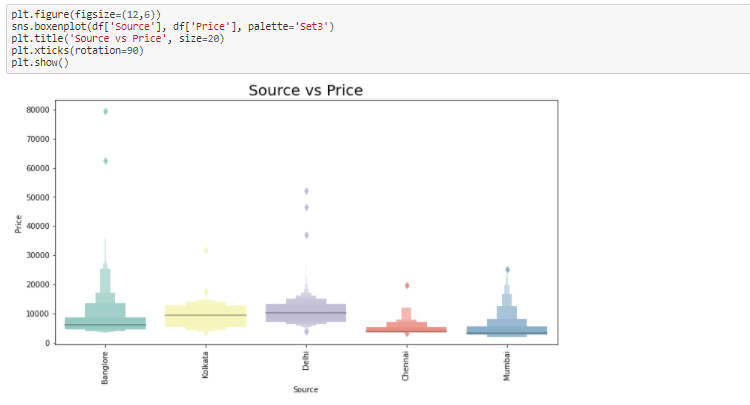
* **Exploratory Data Analysis (EDA)**



Observations: Jet-Airways is having huge count of 3849.



Observations: We can see that for Jet Airways Business having maximum price rate which is distributed from min. 50000 to 80000, whereas spice Jet having less than all minimum price rate which is distributed from min. 2500 to 10000.



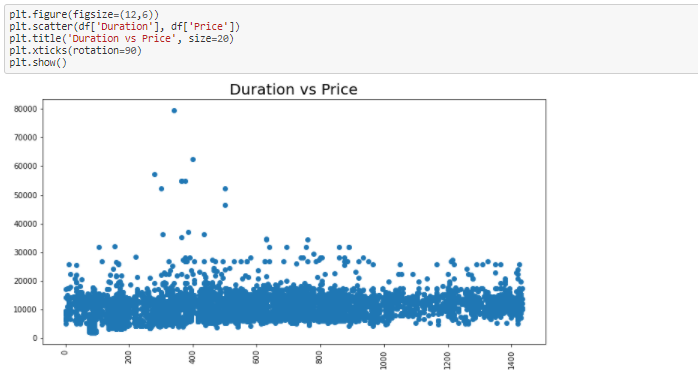
Observations: Flight which are arriving from Kolkata and Delhi has more price distributed between 5000 to 15000.



Observations: First of every month is pricey than other days comparatively.



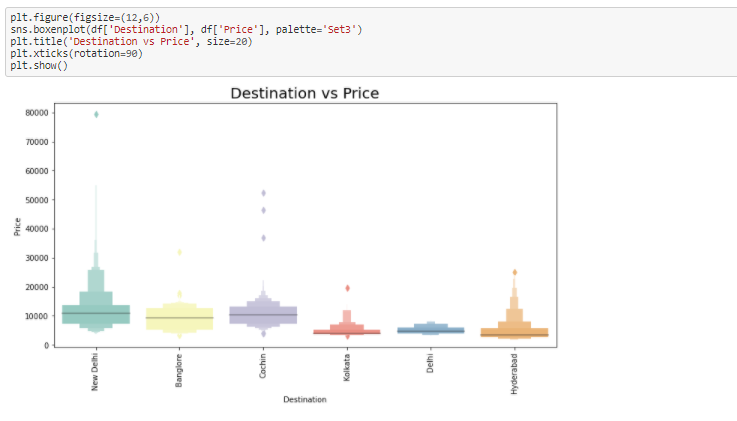
Observations: March month flight prices are costlier when compared.



Observations: Almost for all duration flight price is distributed in between 5000 to 20000.



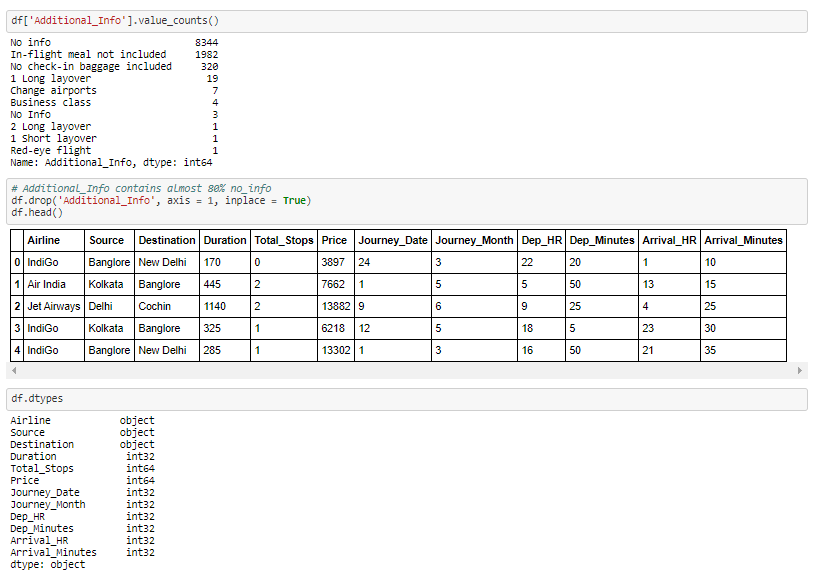
Observations: As number of total\_stops increases, price of flight tickets also increases.



Observations: Flights those are going to New Delhi are having more price distribution.

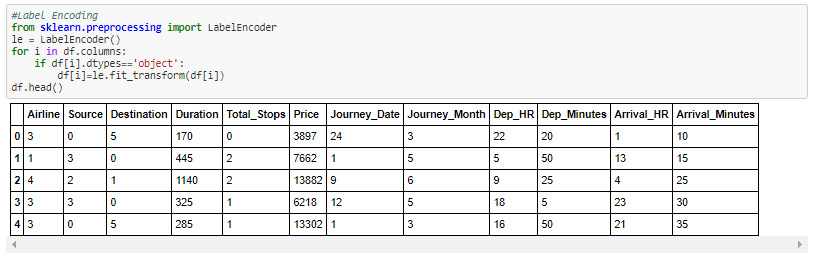


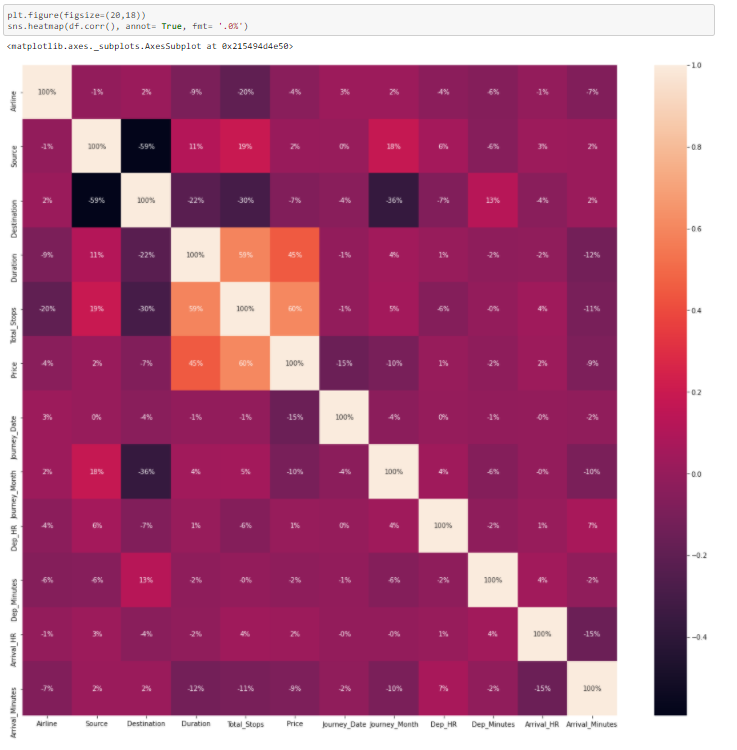
Observations: Business class bear the highest price.



* **Encoding to keep the data safe**

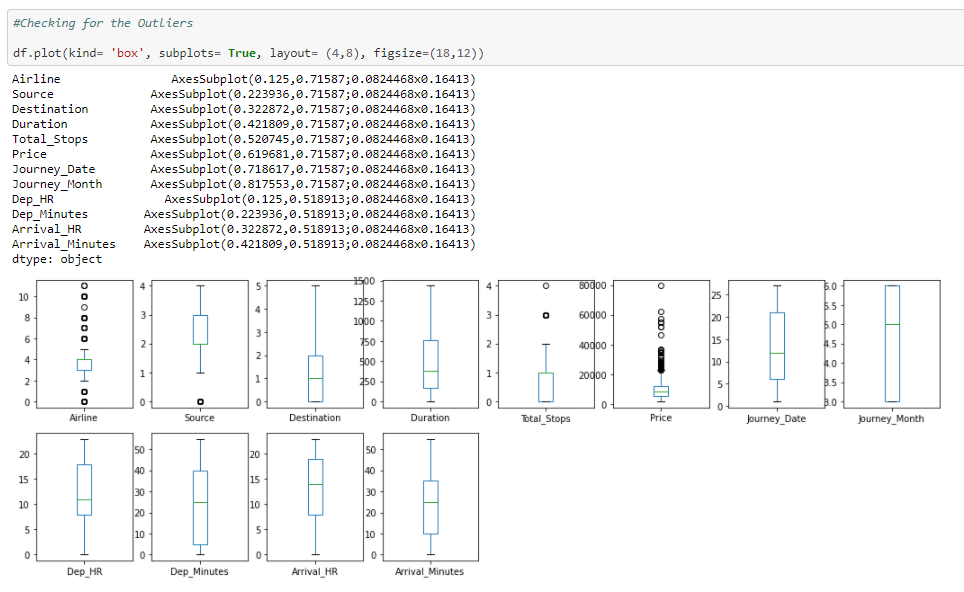
Here we are going to convert all object data type into numerical data type by using Label Encoding.





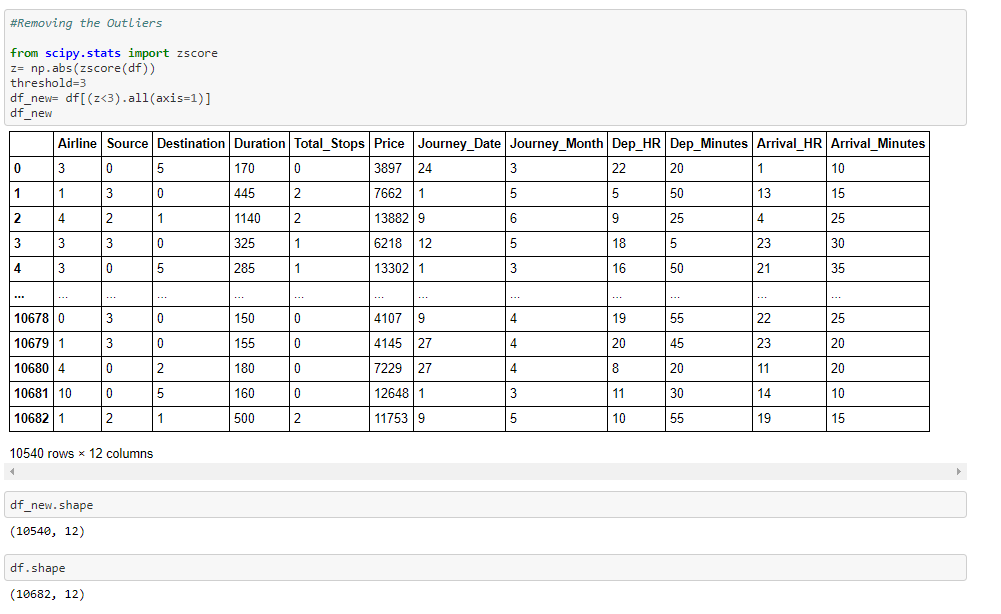
Observations: Flight price is highly correlated with total stops and Duration of flight.

* **Checking For Outliers**

****

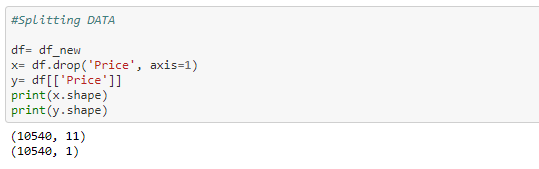
Observations: There are maximum outliers are present in Price. That one we need to remove.

* **Removing Outliers by trimming the data from the DataFrame**

****

Observations: We can observe that there are almost 142 rows removed due to outliers.

* **Splitting Data**

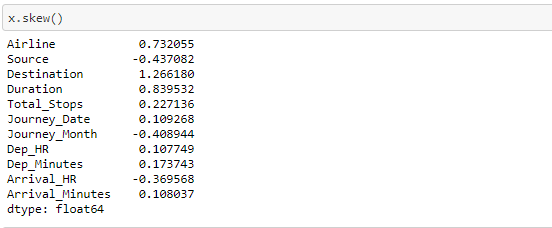
****

**Pre-processing Pipeline**

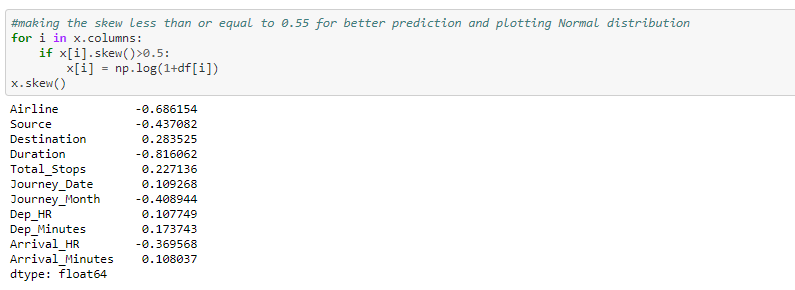
* **Checking for Skewness**

****

Observations: it is observed that Duration, Dep\_Hr are having skewness.

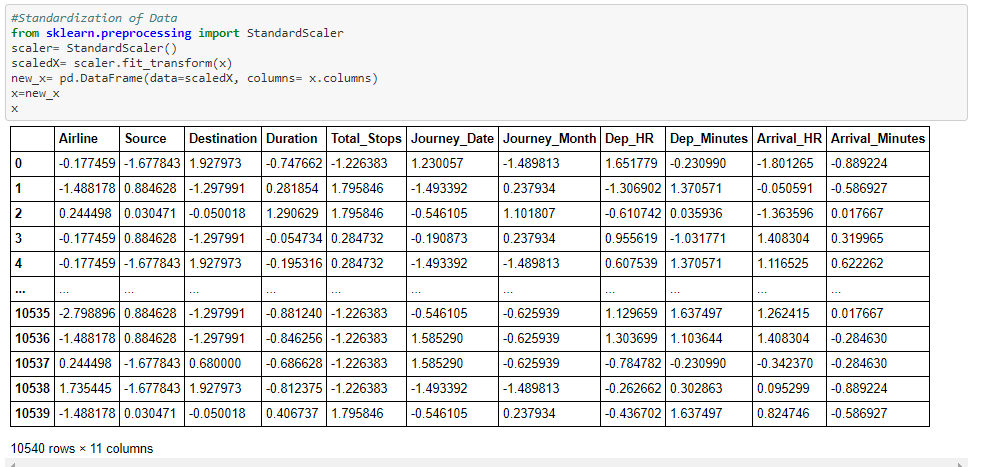
****

* **Removing Skewness**

****

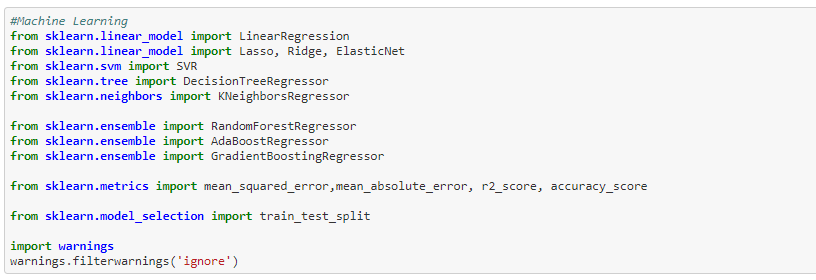
* **Standardization of Data**

Here we are doing Standardization i.e. scaling of data by using StandardScaler.

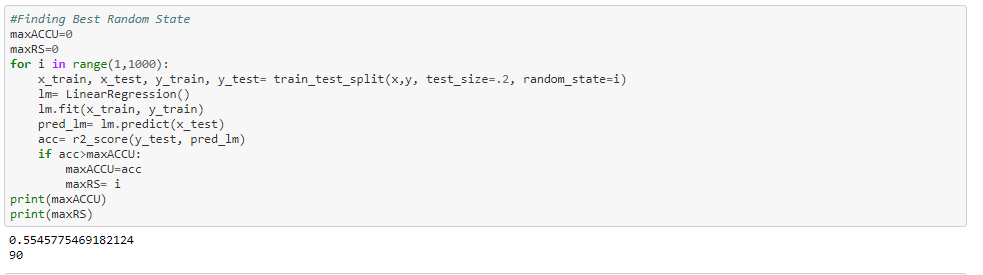


* **Machine Learning**

1. Importing Libraries for Machine Learning.

****

1. Finding the best random state.

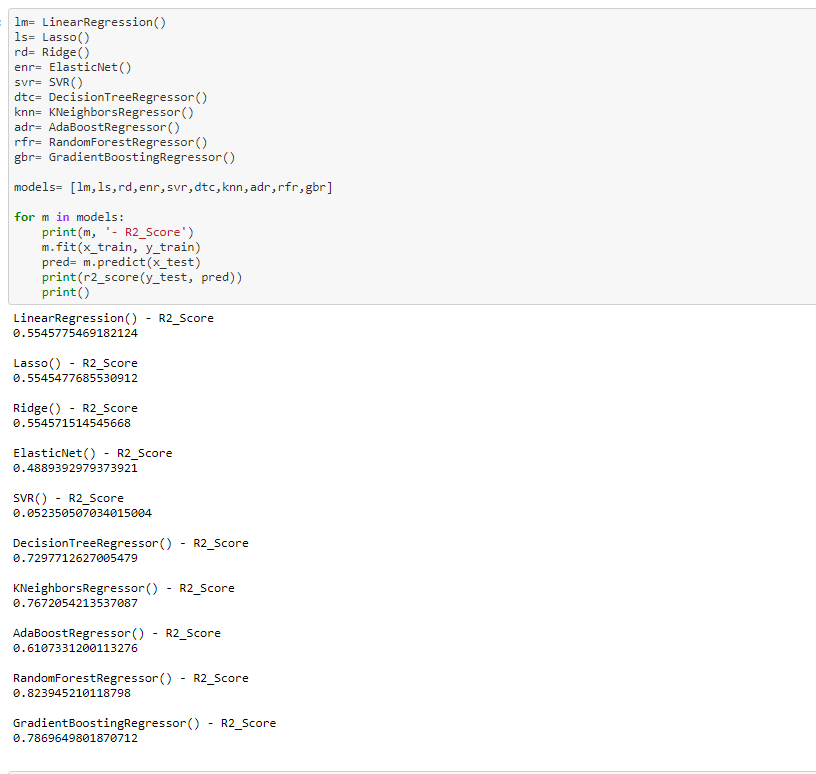


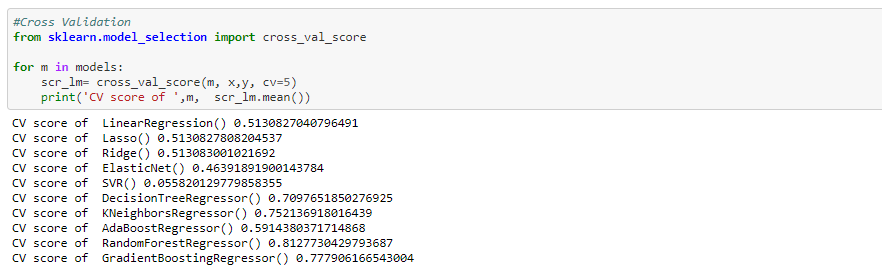
Here we can see that best random state is 90.

1. Splitting the data for training and testing purpose.



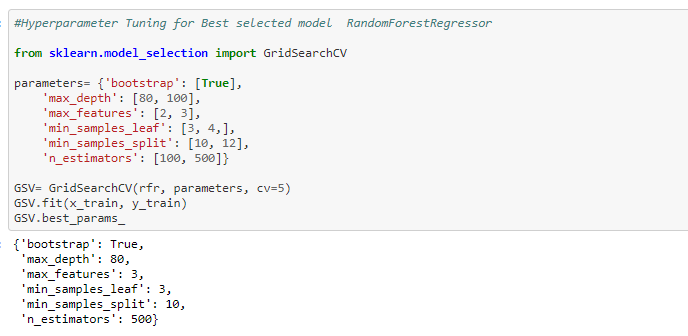
1. Selecting the best model.



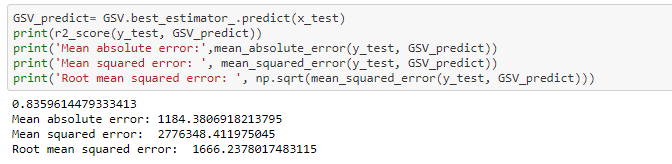


Observations: RandomForestRegressor model is having effectiveness of machine learning model, so we will select RandomForestRegressor model for final predictions.

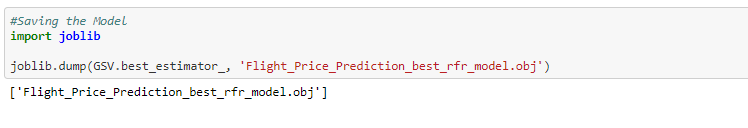
1. Hyperparameter Tuning for Best selected model RandomForestRegressor



1. Getting best of RandomForestRegressor model from best Parameters.



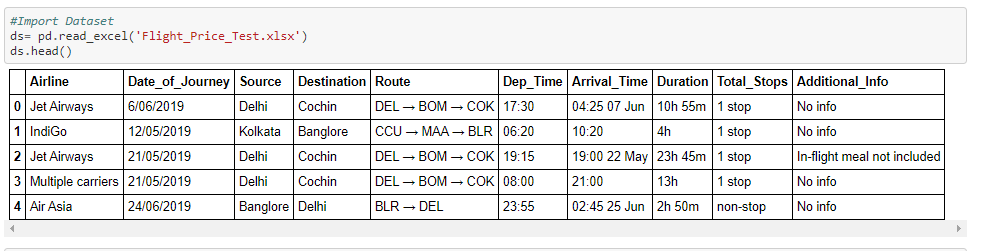
* **Saving the best model**

****

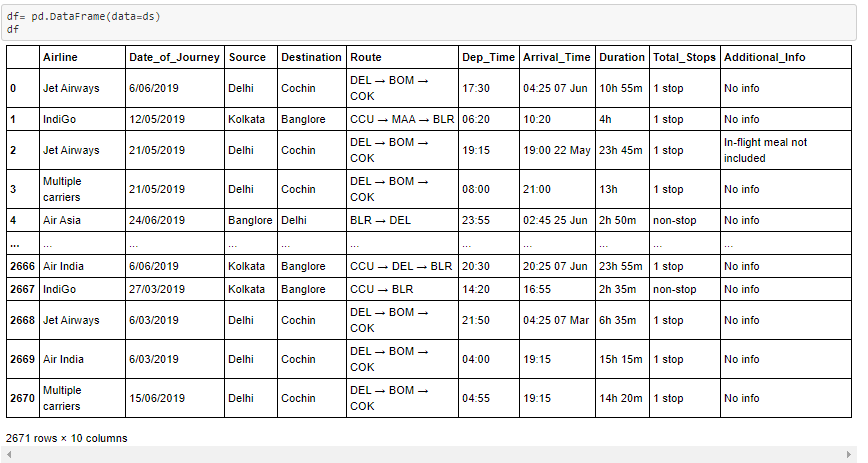
* **Working on Test Data**

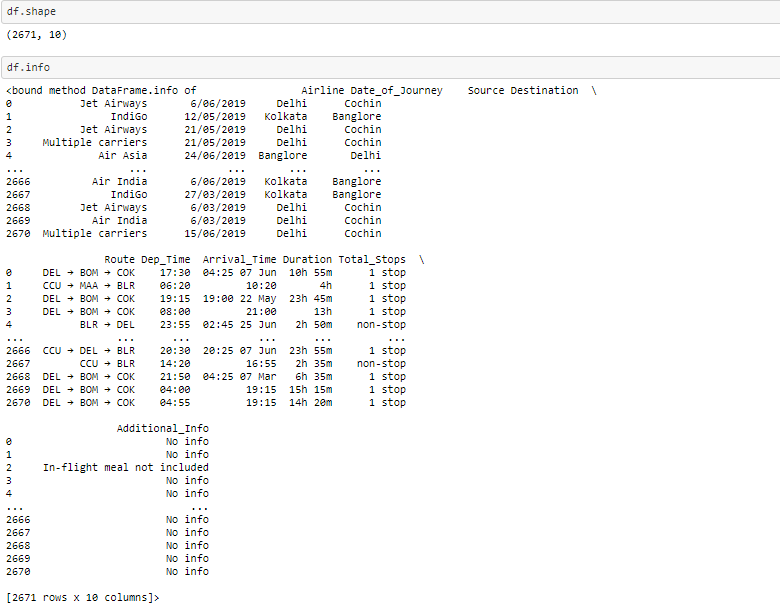
Data cleaning is carried out, Feature Engineering and EDA on test data, then pass this data into the best selected model.

1. Import Test data

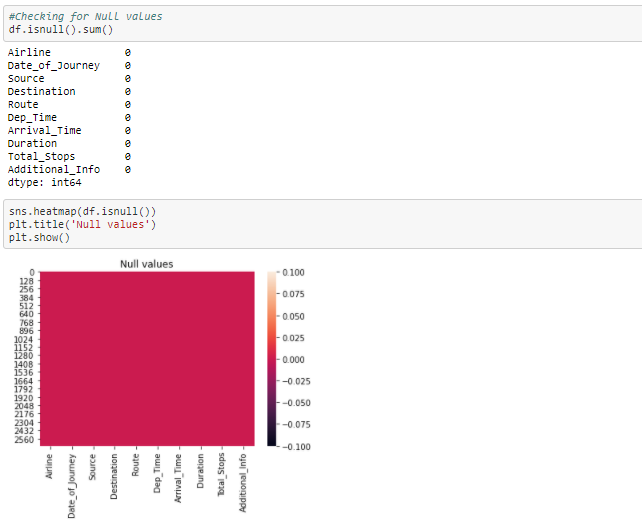


1. Fitting test data into DataFrame and use of pandas DataFrame methods.



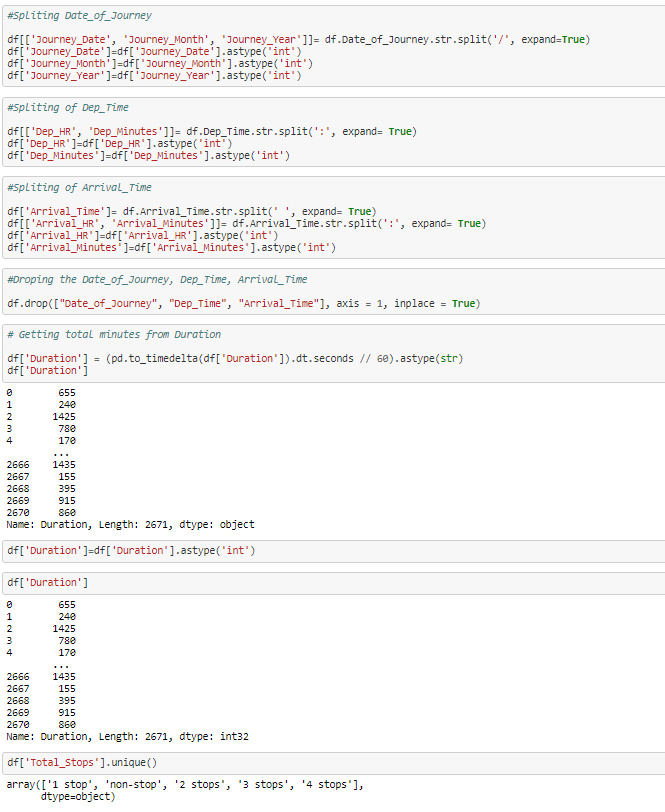


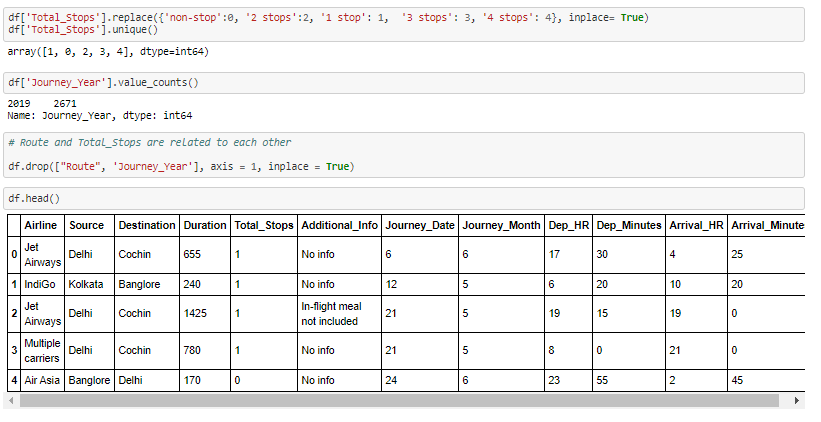
1. Checking of Null Values and removing if Null values present.



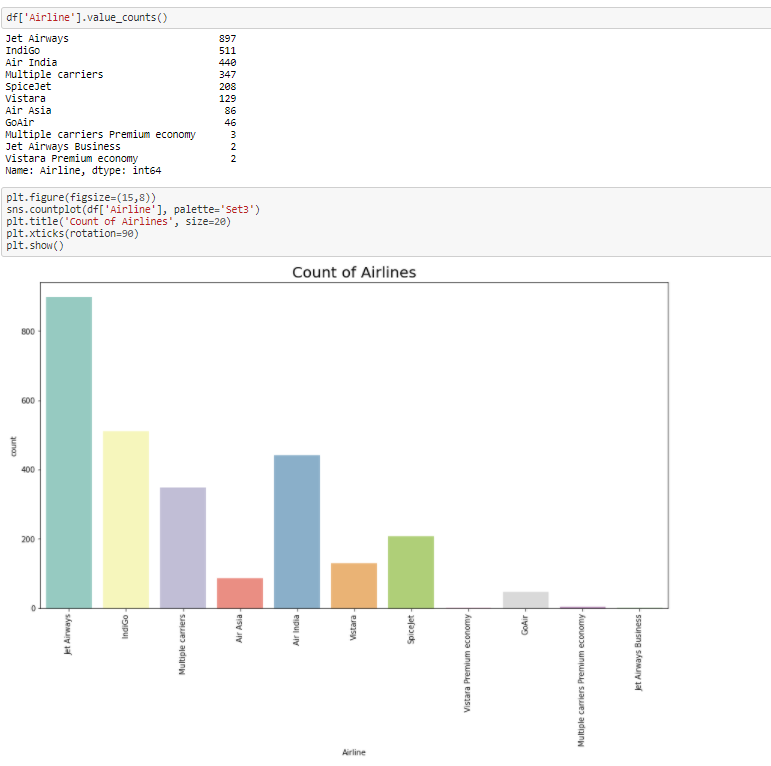
Observations: No null values are present in dataset.

1. Feature engineering for Test Data

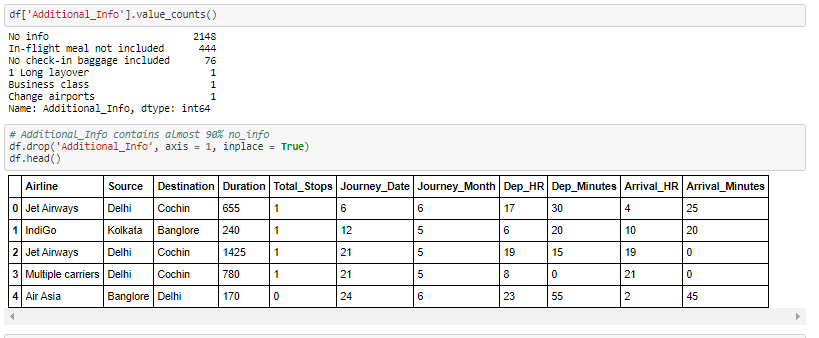


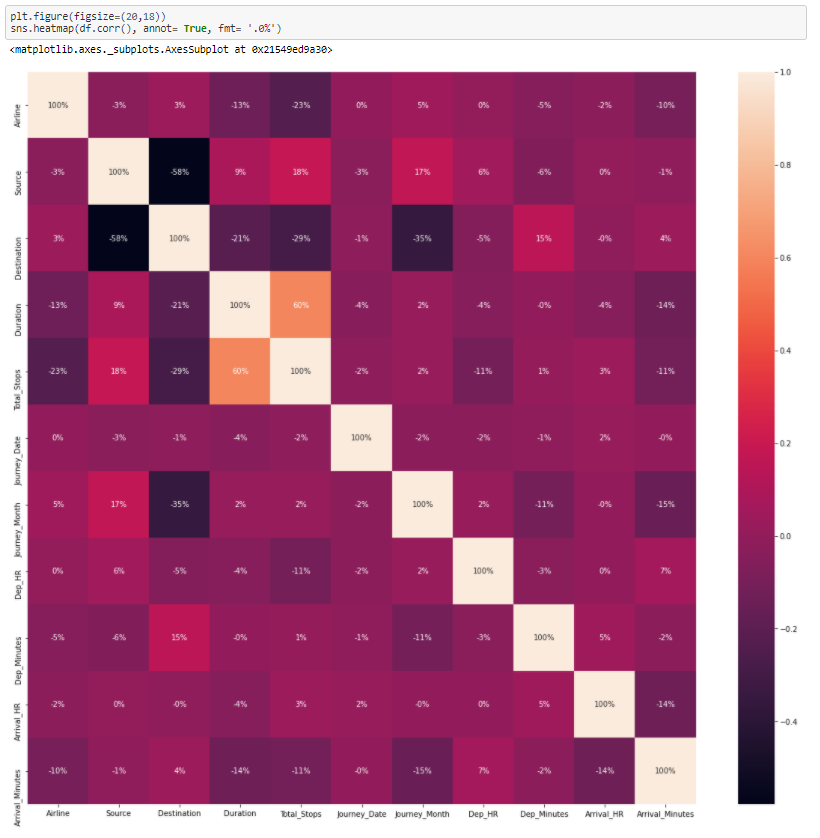


1. EDA for Test data



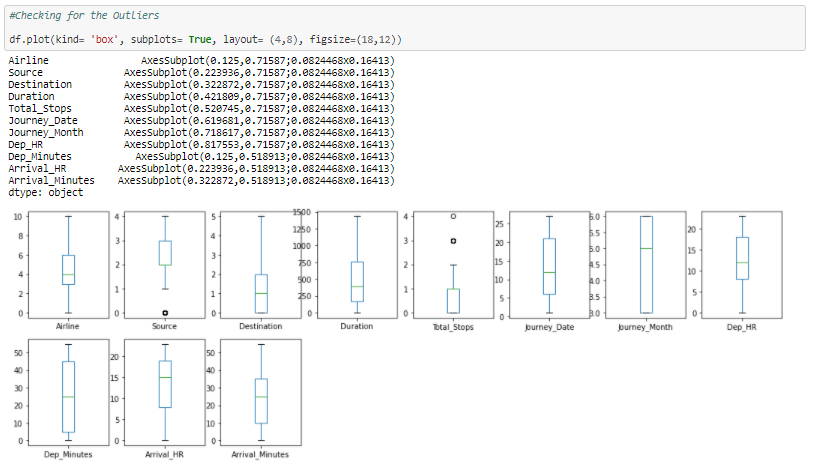
Observations: There are most of Jet-Airways having almost 900 counts.





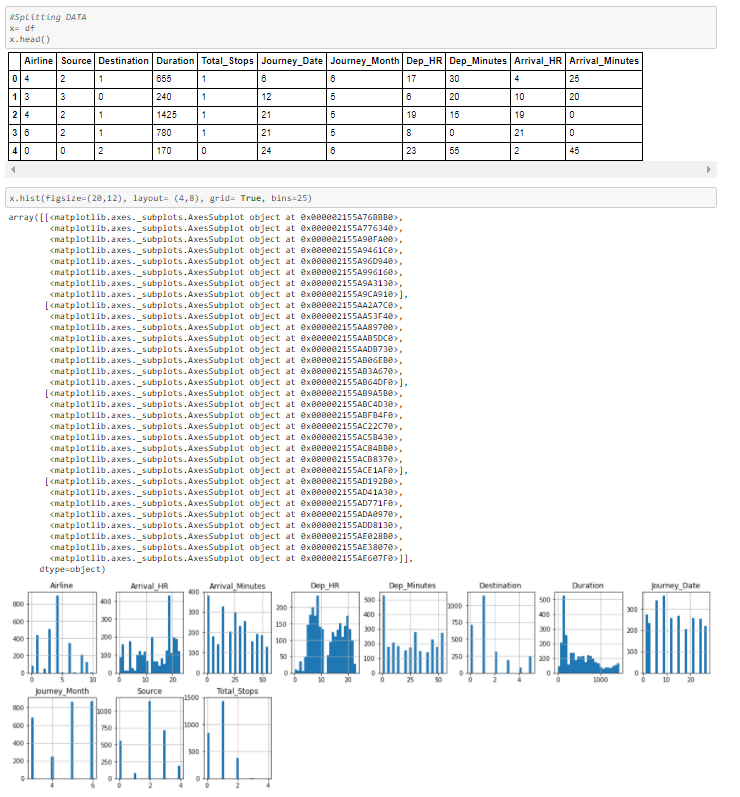
Observations: We can see that Duration is highly correlated with total stops.

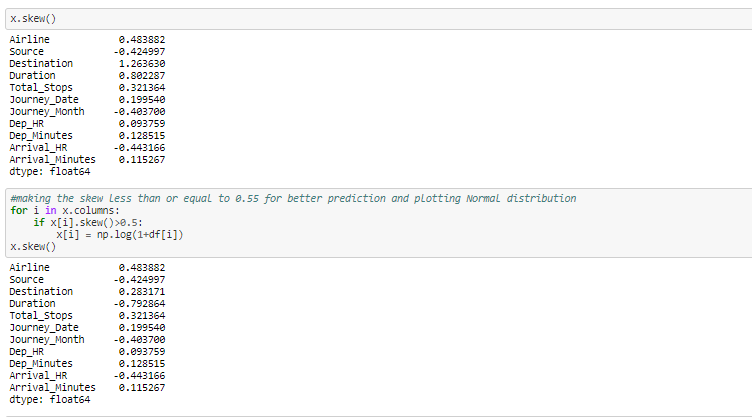
1. Checking for Outliers and removing if present.



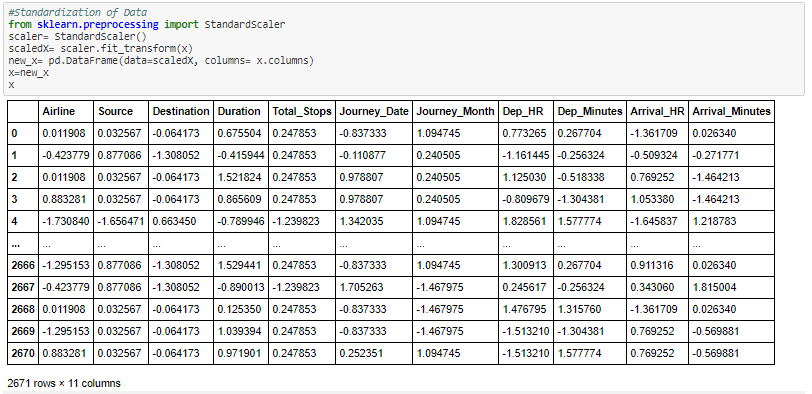
Observations: We can see that there no outliers present.

1. Checking for Skewness and removing if present.

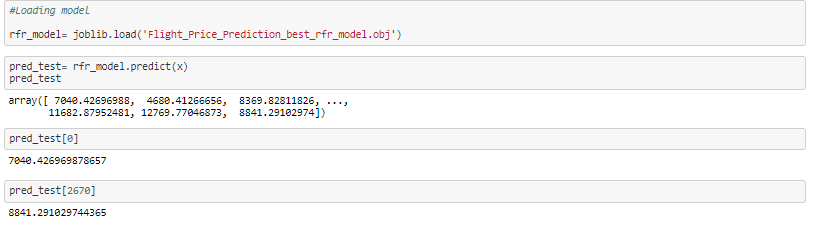




1. Standardization of Test Data



* **Loading the best model and fitting Test data**

****

* **Conclusion**

For Analysis of Flight\_Price\_Prediction project we have selected RandomForestRegressor model which having accuracy of 84%.

# Project Name: HR Analytics Project – Understanding the Attrition in HR

***-Analysis by Shravani Natakala***

**About Project:**

This HR Analytics Project takes HR data and uses machine learning models to predict which employees will be more likely to leave company by given some attributes. Such model would help an organization predict employee attrition and define a strategy to reduce costly problem. For each employee, in addition to whether the employee left or not i.e. attrition, there are attributes / features such as age, employee role, daily rate, job satisfaction, years at the company, years in current role, etc.

**Problem Statement:**

Every year a lot of companies hire a number of employees. The companies invest time and money in training those employees, not just this but there are training programs within the companies for their existing employees as well. The aim of these programs is to increase the effectiveness of their employees. But where HR Analytics fit in this? and is it just about improving the performance of employees?

**HR Analytics**

Human resource analytics (HR analytics) is an area in the field of analytics that refers to applying analytic processes to the human resource department of an organization in the hope of improving employee performance and therefore getting a better return on investment. HR analytics does not just deal with gathering data on employee efficiency. Instead, **it aims to provide insight into each process by gathering data and then using it to make relevant decisions about how to improve these processes.**

**Attrition in HR**

Attrition in human resources refers to the gradual loss of employees overtime. In general, relatively high attrition is problematic for companies. HR professionals often assume a leadership role in designing company compensation programs, work culture, and motivation systems that help the organization retain top employees.

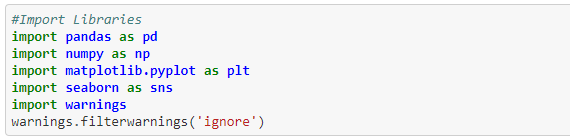
How does Attrition affect companies? and how does HR Analytics help in analyzing attrition? We will discuss the first question here and for the second question, we will write the code and try to understand the process step by step.

**Attrition affecting Companies**

A major problem in high employee attrition is its cost to an organization. Job postings, hiring processes, paperwork, and new hire training are some of the common expenses of losing employees and replacing them. Additionally, regular employee turnover prohibits your organization from increasing its collective knowledge base and experience over time. This is especially concerning if your business is customer-facing, as customers often prefer to interact with familiar people. Errors and issues are more likely if you constantly have new workers.

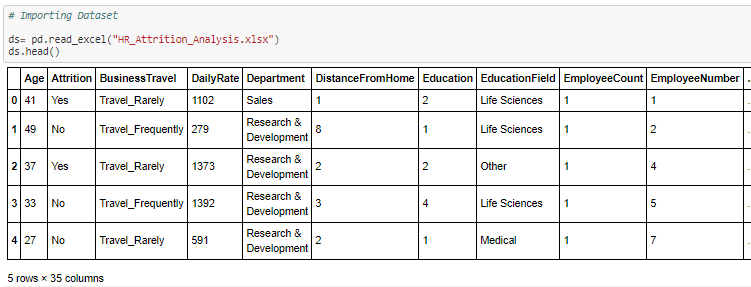
**Data Analysis**

* **Import Libraries**



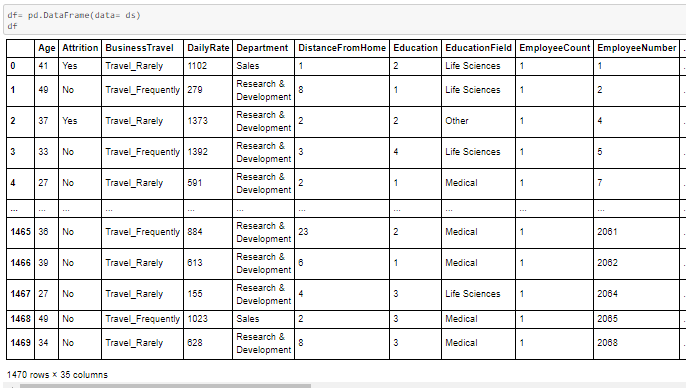
Here we import libraries for getting data and for making EDA Analysis.

* **Import Dataset**



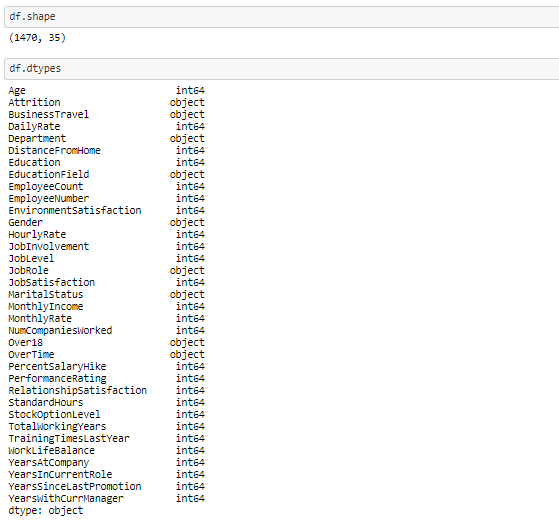
We can see now top 5 rows of Dataset.

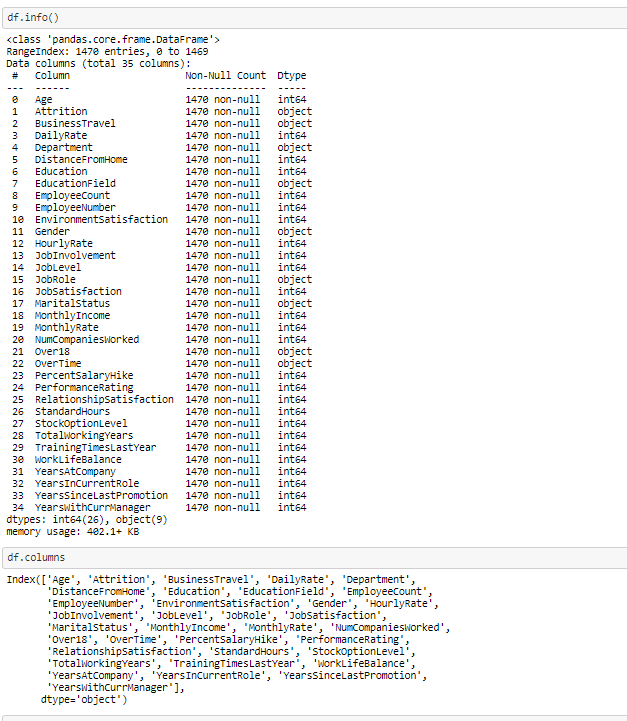
* **Fitting data into Pandas Dataframe**

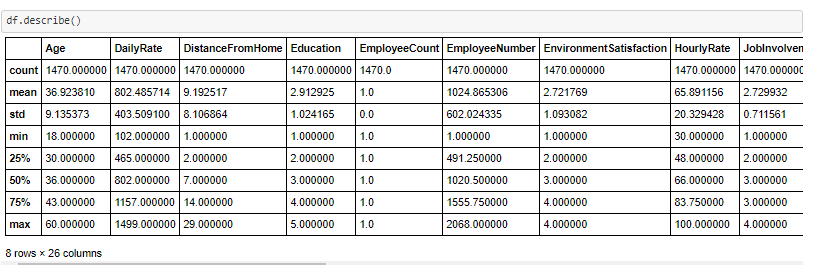


Here we fitted our dataset into Pandas Dataframe, so that we can do all Dataframe methods.

* **Appling Pandas DataFrame methods**

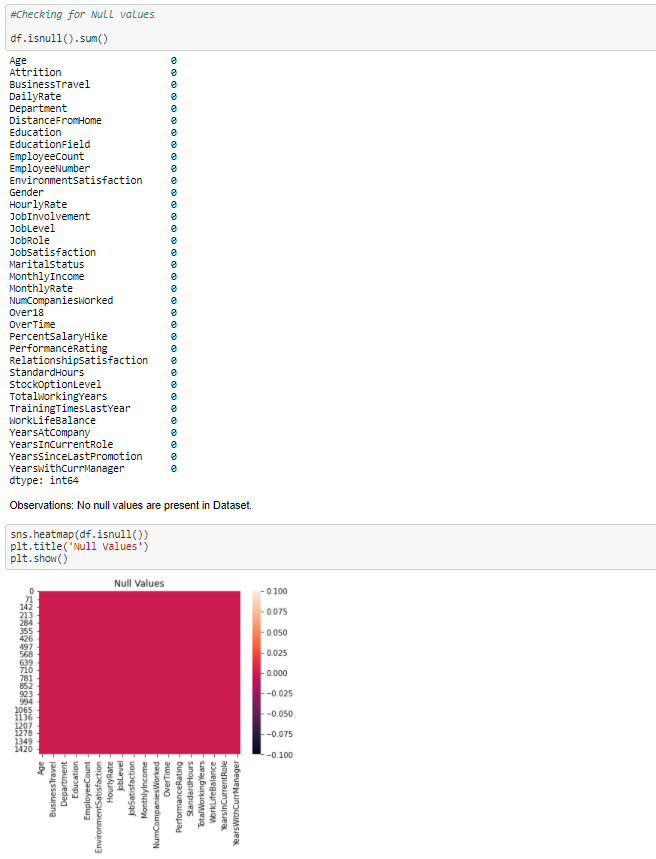






Observations: Age of the employee is in between 18 and 60. For all columns count is 1470, that means no null values present in dataset. EmployeeCount and StandardHours are single value column. Mean and 50% values are same for almost all columns. Also difference in 75% and max value is less that means chance of presence of outliers are less.

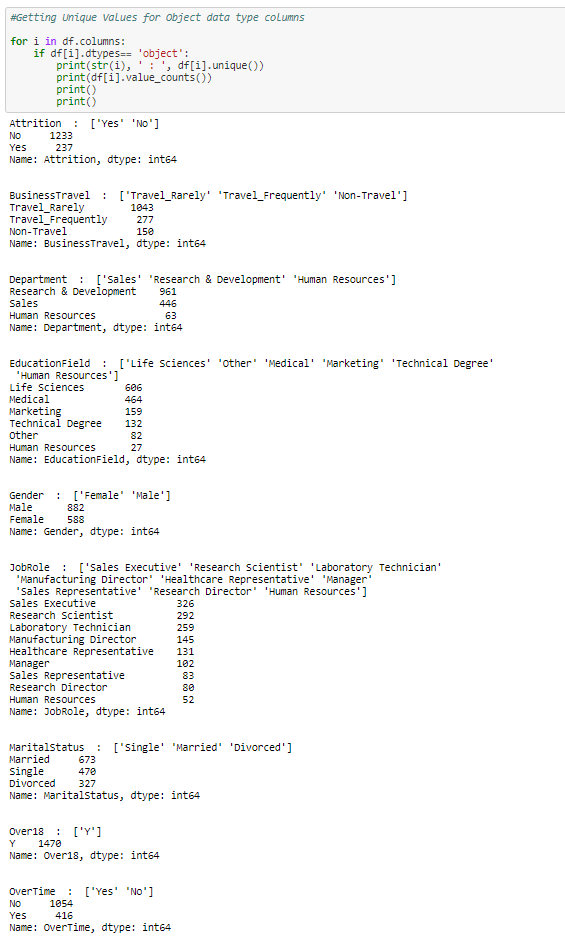
* **Checking for Null values**



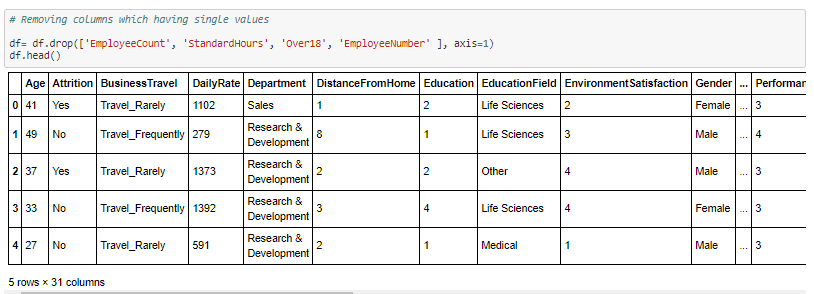
Observations: No null values are present in Dataset.

* **Feature Engineering**

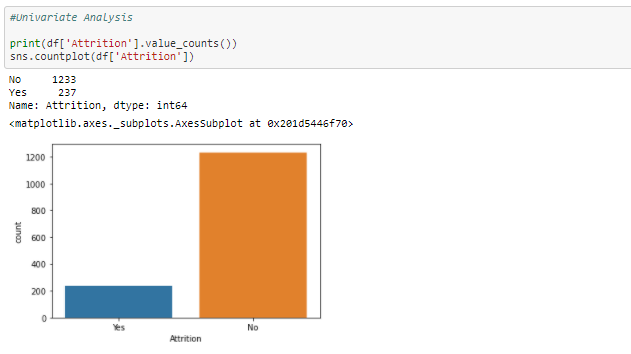
Here we are going to add more features in our dataset from existing one and deleting unwanted features.



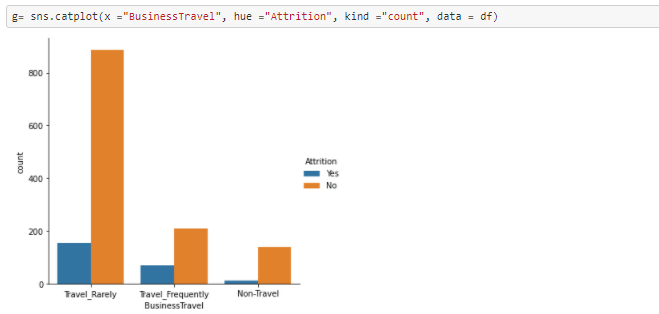
Observations: Here we can see that some features having single value, those we need to remove as shown in below



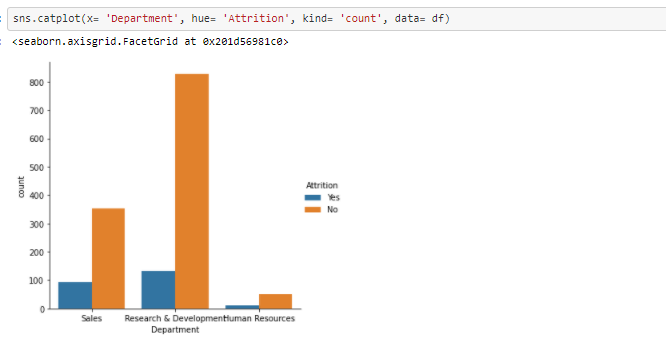
* **EDA**



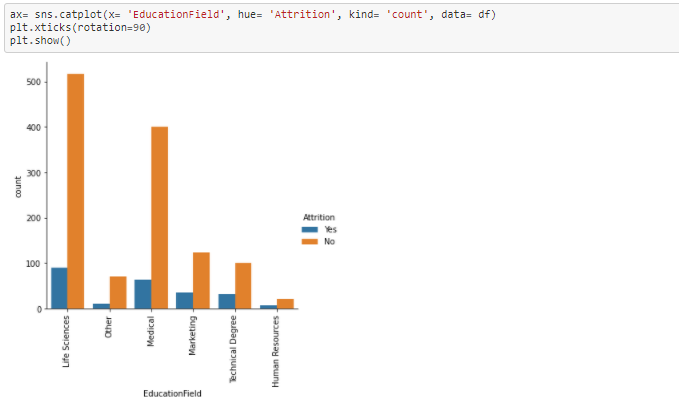
Observations: There are 237 employees left the company and 1233 employees are currently working in company.



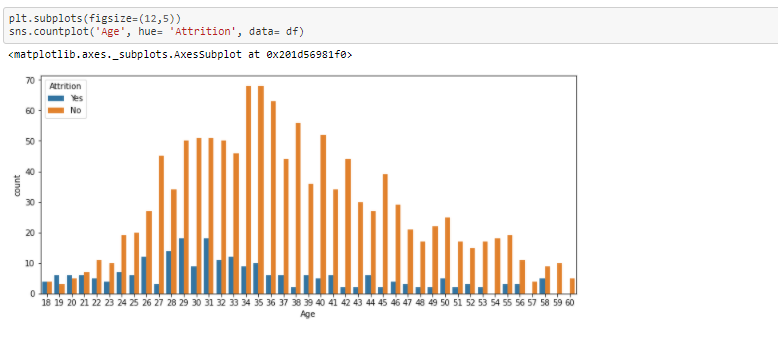
Observations: We can see that, rate of employees leaving is more in Frequently Travel than Rarely and Non\_Travler.



Observations: We can see that, rate of employees leaving is more in sales department than Research & Development and Human Resources.



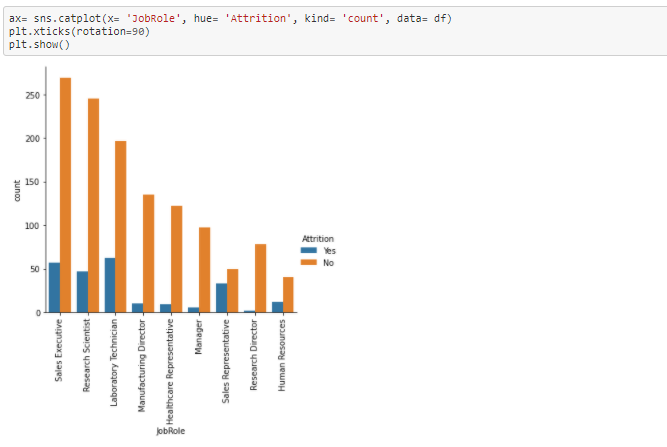
Observations: We can see that, rate of employees leaving is more in Life Science and Medical.



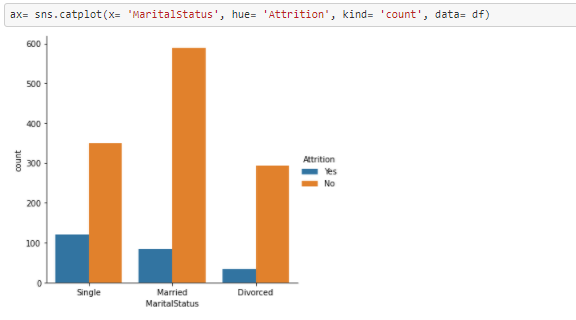
Observations: Maximum Employees at Age of 34 to 35 are currrently working in company where as maximum employees with Age of 29 and 31 are leaving the company.



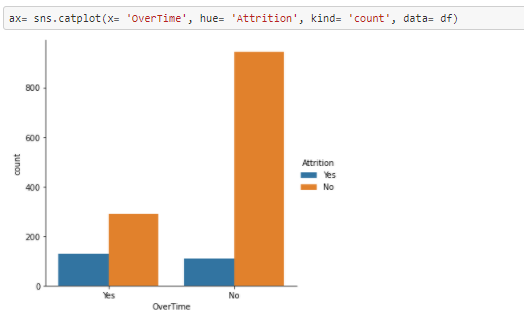
Observations: There is more Male employees are currently working in company. Also Male employees have left the company.



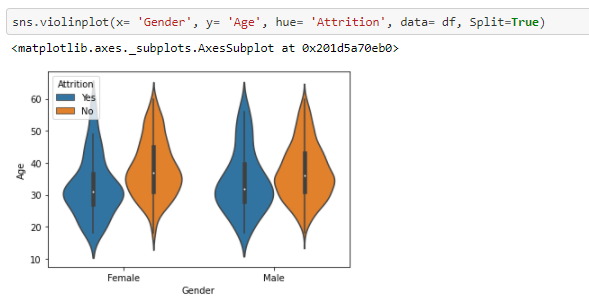
Observations: Currently employees with job role of Sales Executive and Research scientist are in company.



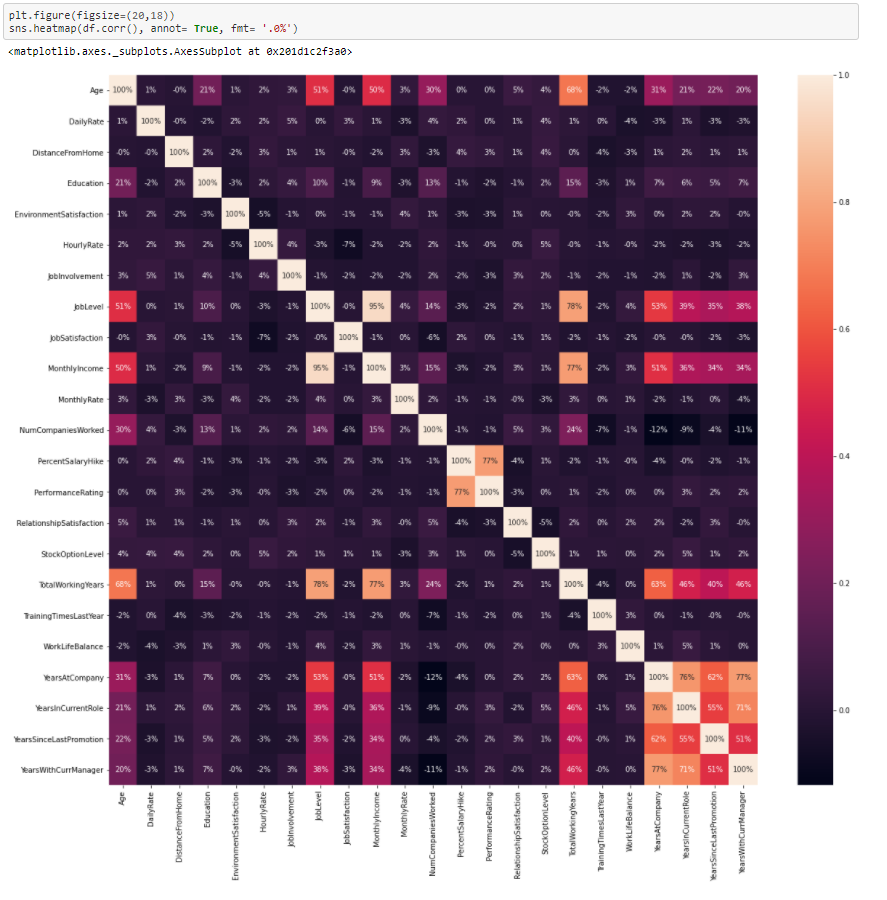
Observations: Currently Married employees are more in company. And where as Single employees have left.



Observations: Currently maximum employees are not doing overtime.



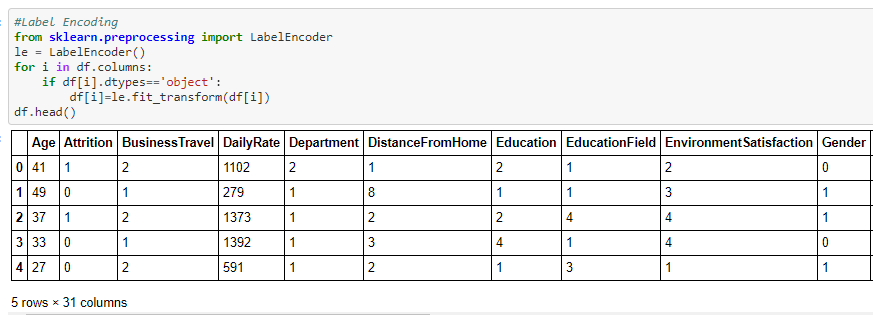
Observations: Female and male employees with age of 29 to 32 have left the company.



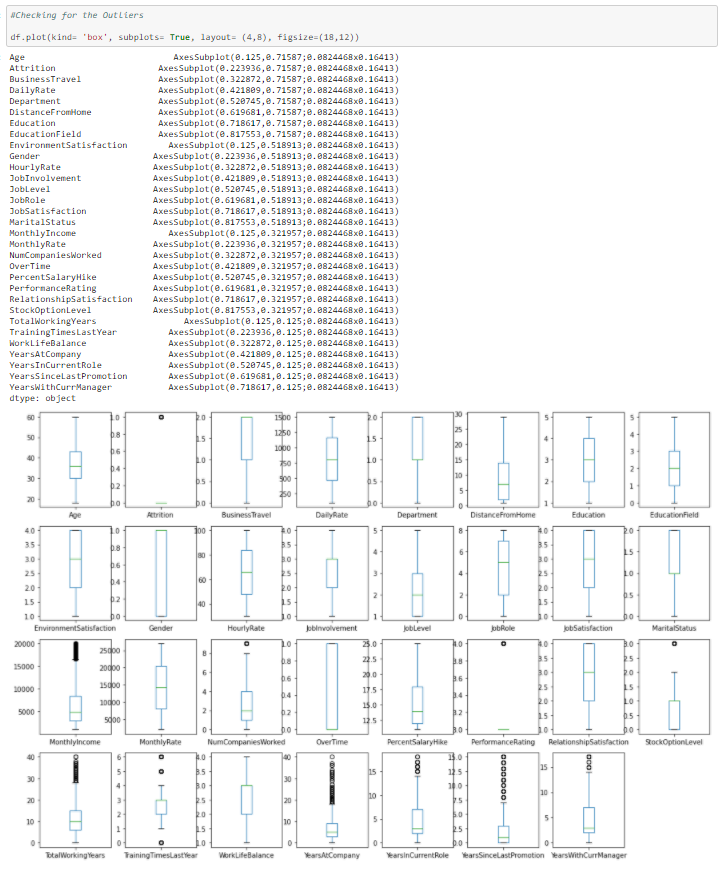
Observations: Total working years has higher positive correlation with age as longer working will get older employees. Also Total working years has higher correlation with jo level and monthly income. Higher coorelation of monthly income with job level.

* **Encoding**

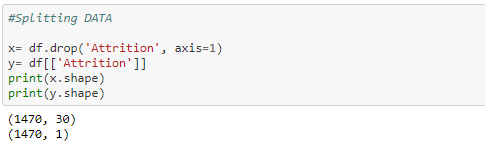
Here we are going to convert all object data type into numerical data type by using Label Encoding.



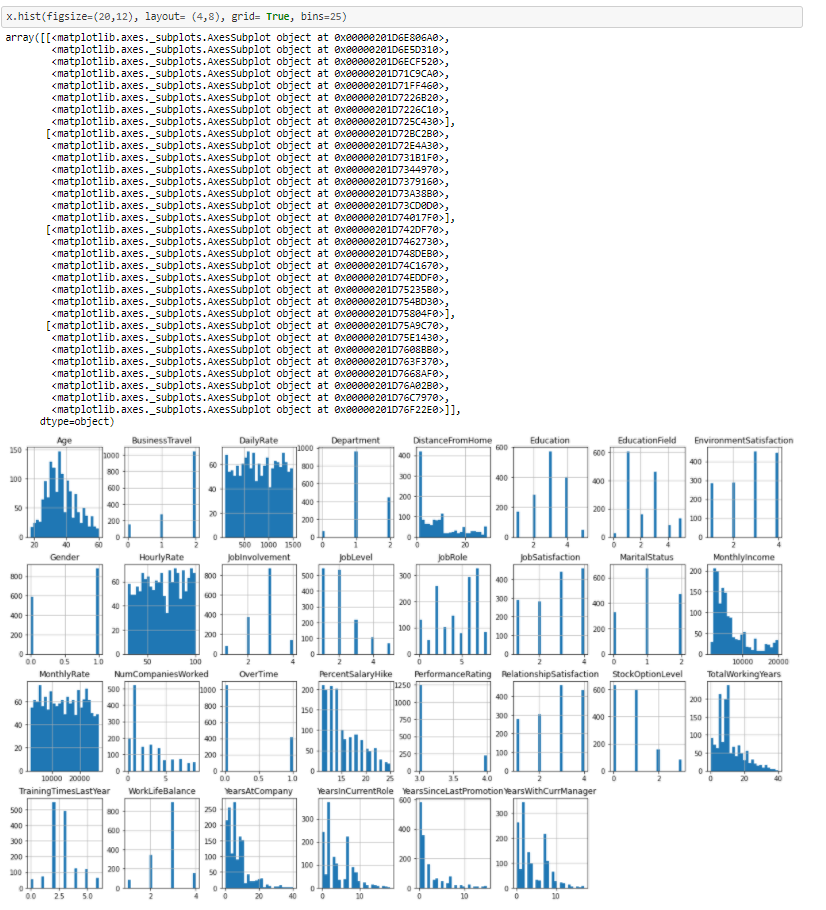
* **Checking For Outliers**

  
Observations: Almost for all columns outliers are no present, except some columns but that one also acceptable.

* **Splitting Data**

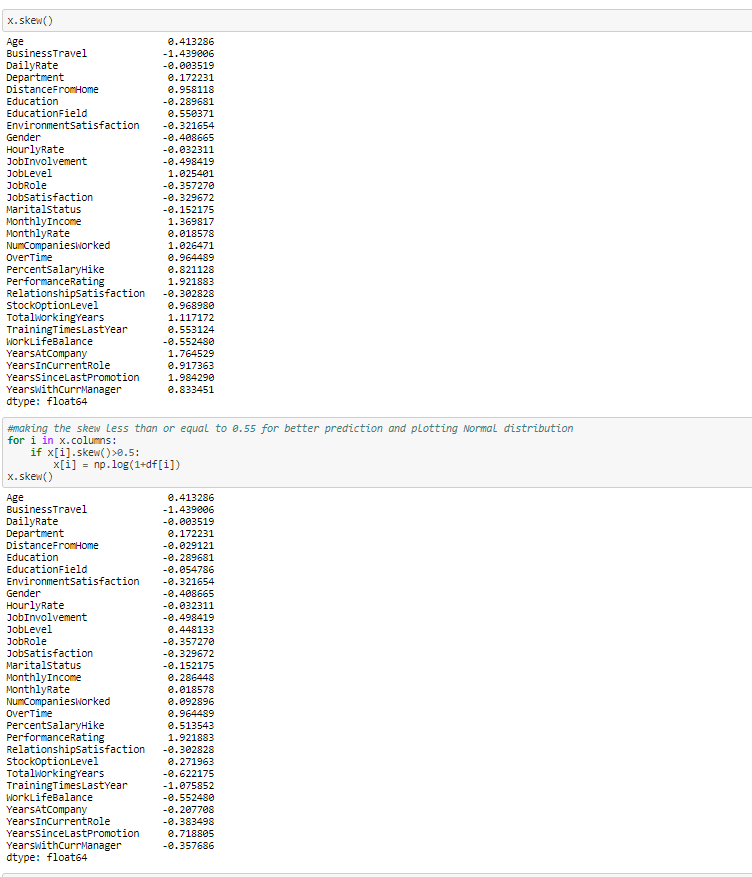
****

* **Checking for Skewness**

****

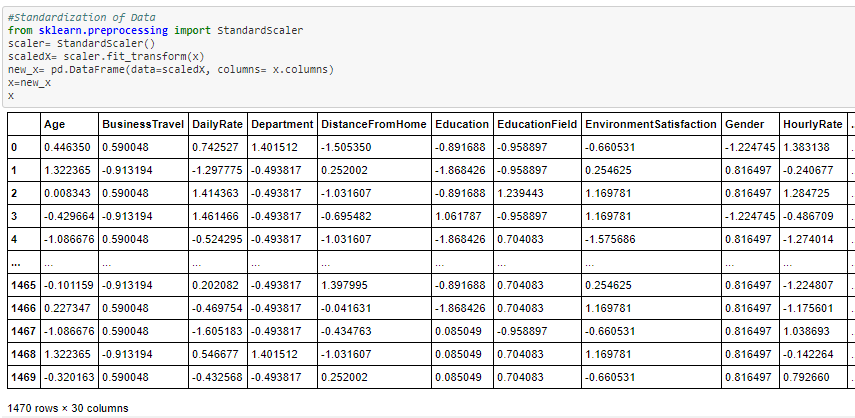
Observations: Age, Daily Rate are looks normally distributed. Where as rest all columns are some what left skewed as data values are not varies in terms of numbers.

* **Skewness removal**

****

* **Standardization of Data**

Here we are doing Standardization i.e. scaling of data by using StandardScaler.

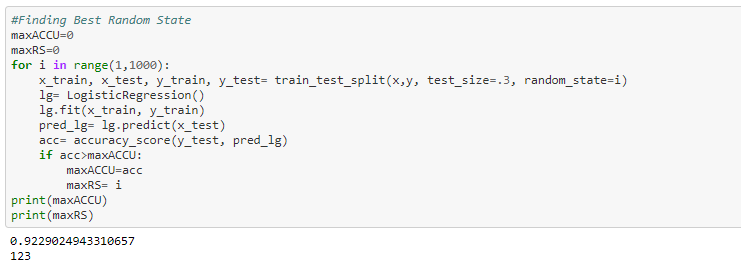


* **Machine Learning**

1. First we are going to import Libraries for Machine Learning.

****

1. Finding best random state.

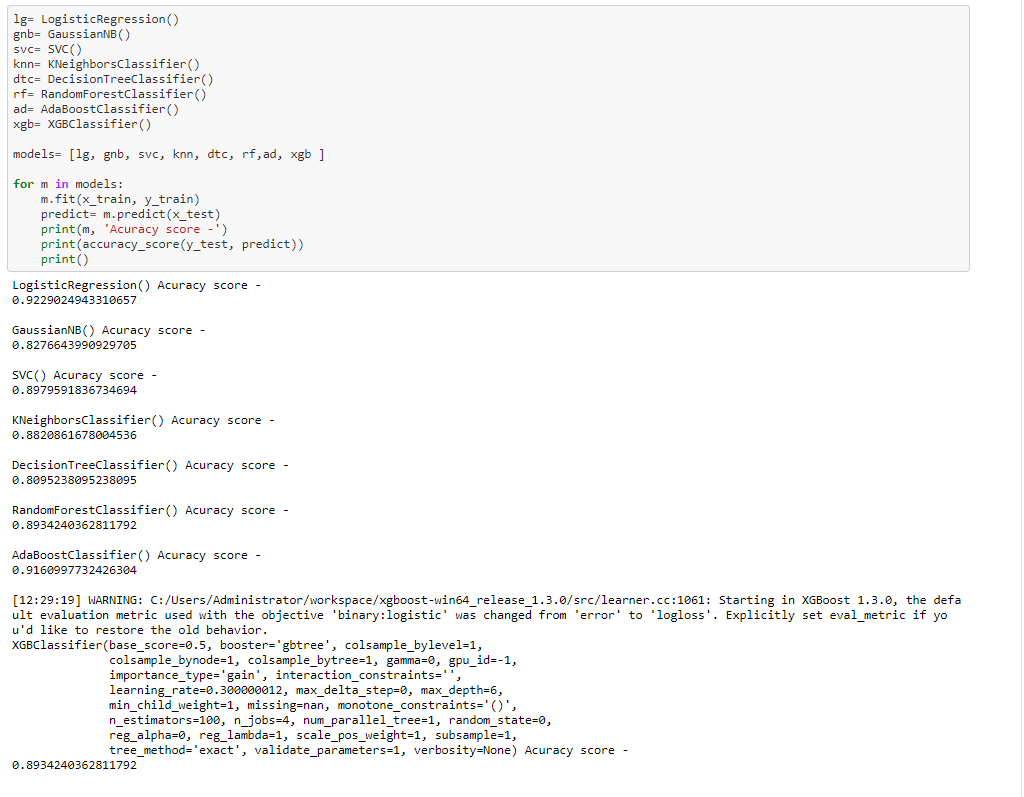


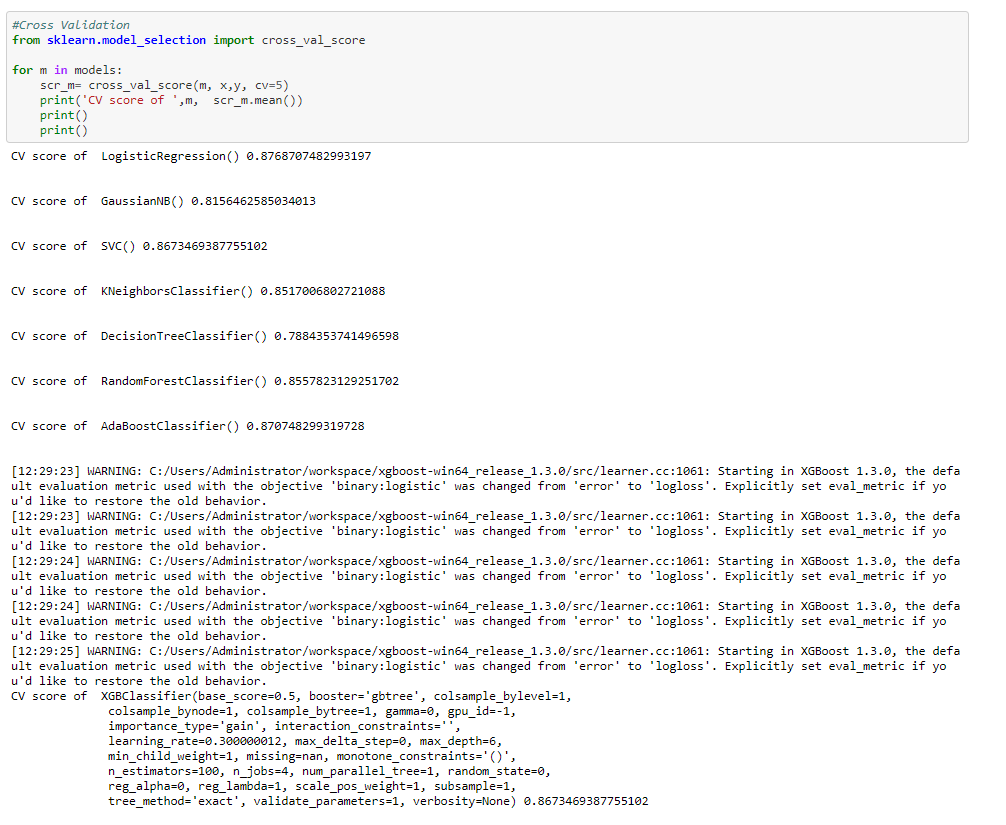
Here we can see that best random state is 123 which gives Accuracy of 92%..

1. Splitting the data for training and testing purpose.



1. Selection of best model.



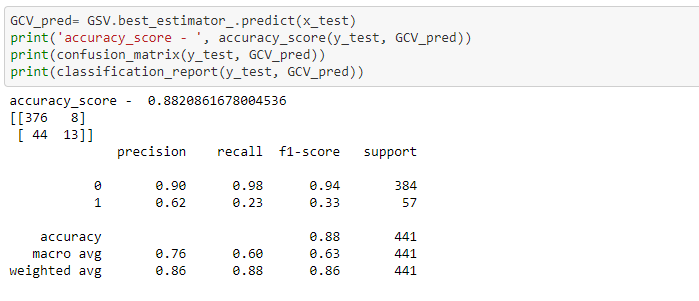


Observations: From above we can observe that for DecisionTreeClassifier model there is less difference in accuracy\_score and cross validation accuracy\_score so we will select DecisionTreeClassifier model for final predictions.

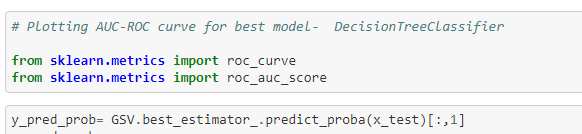
1. Hyperparameter Tuning for Best selected model DecisionTreeClassifier

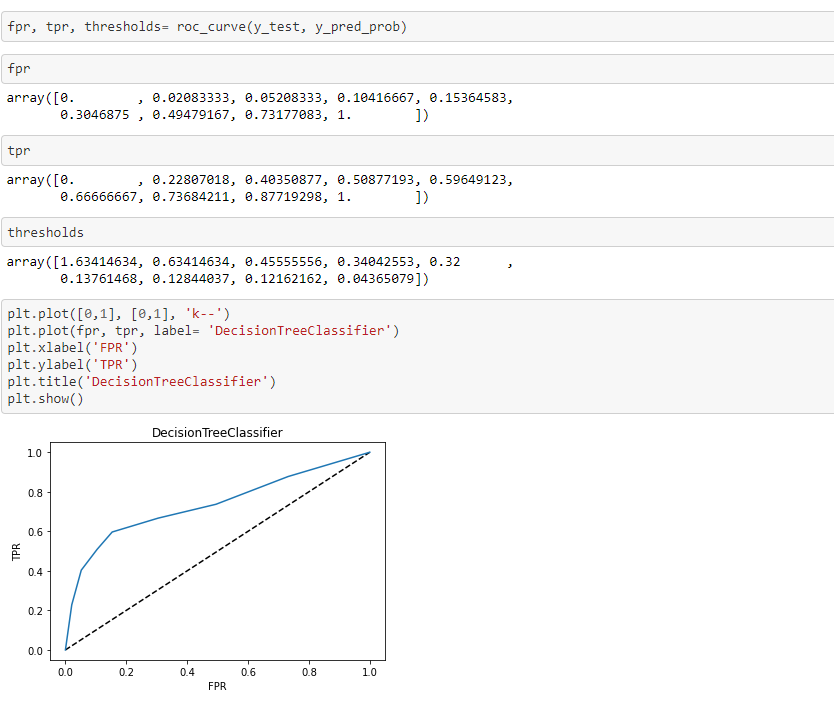


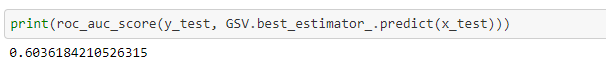
1. Getting best of DecisionTreeClassifier model from best Parameters.



* **Plotting AUC-ROC Curve**

****

****

****

Observations: We can observe that A plot of the ROC Curve confirms the AUC interpretation of a best model for most probability thresholds.

* **Saving the best model**

****

Conclusion: For Analysis of HR\_Attrition\_Analysis project we have selected DecisionTreeClassifier model which having accuracy 88.21%.