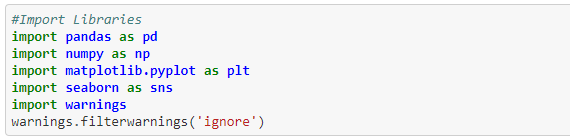
**Project Name: Flight Price Prediction**

*-Analysis by Vaibhav Banke*

**About Project:**

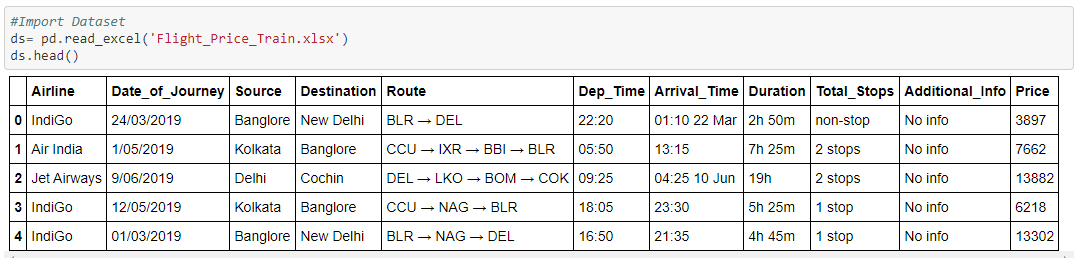
We have prices of flight tickets for various airlines between the months of March and June of 2019 and between various cities. So we are going to build a machine learning model to predict the price of the flight ticket based on features such as Airline, Date\_of\_Journey, Dep\_Time, Arrival\_TimeTotal\_Stops etc.

* **Import Libraries**



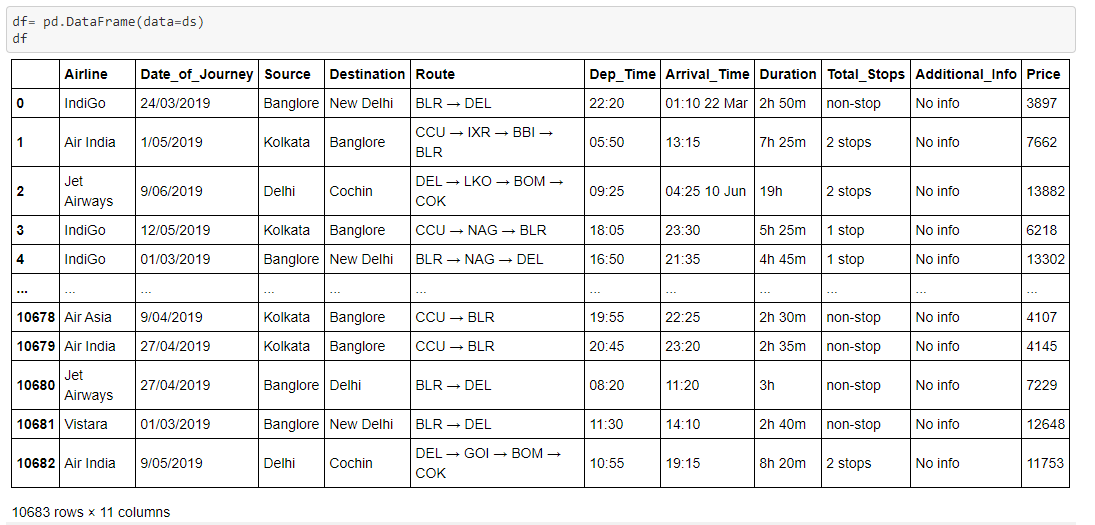
Here we import libraries for getting data and for making EDA Analysis.

* **Import Dataset**



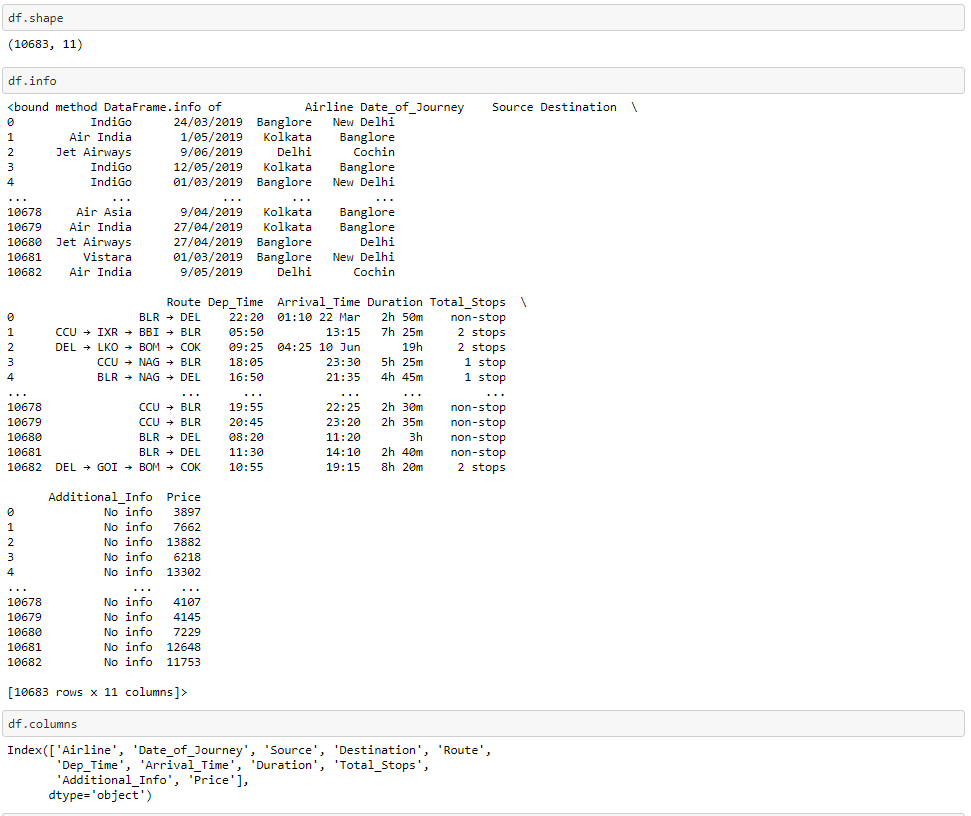
We can see now top 5 rows of Dataset.

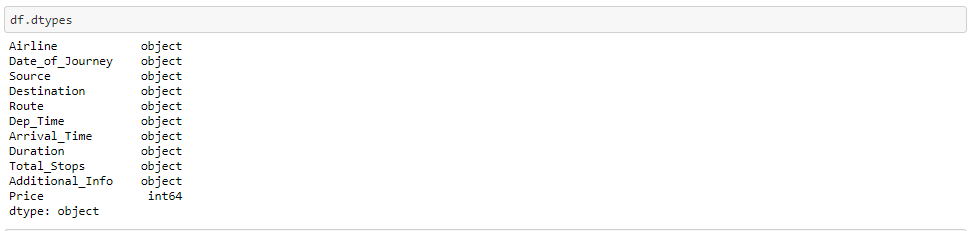
* **Fitting data into Pandas Dataframe**



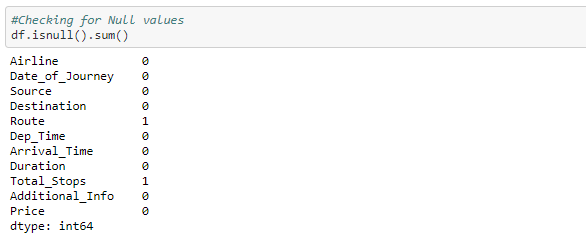
Here we fitted our dataset into Pandas Dataframe, so that we can do all Dataframe methods.

* **Appling Pandas DataFrame methods**



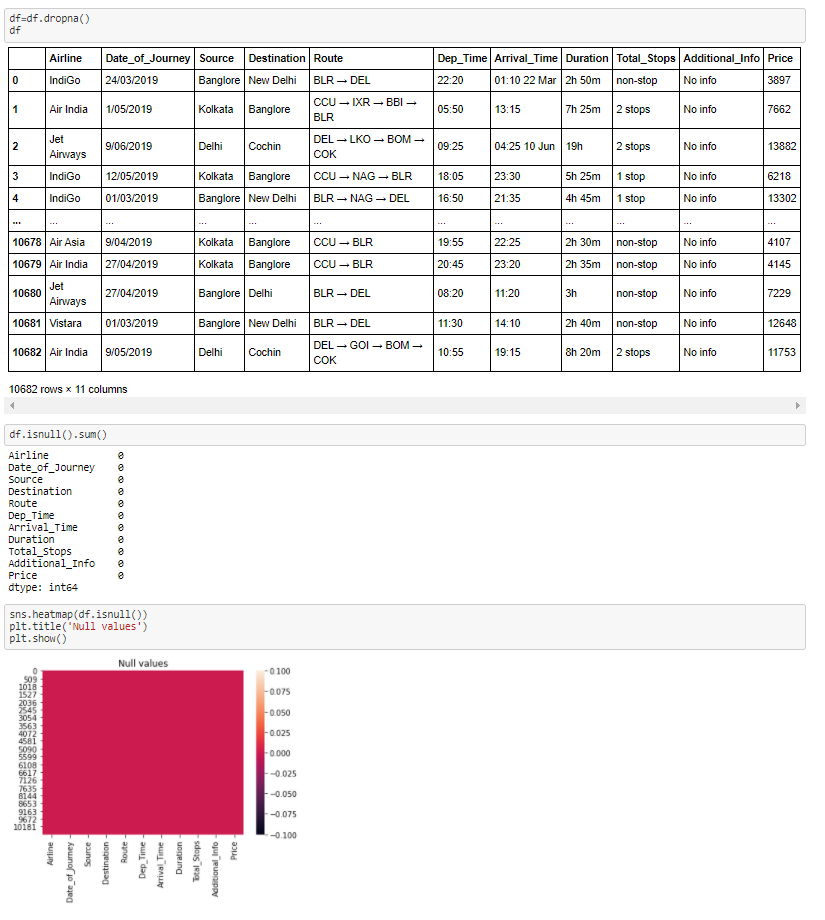


* **Checking for Null values**



Observations: There are Null Values and 1 in each in column Route and Total Stops, so we can drop these.

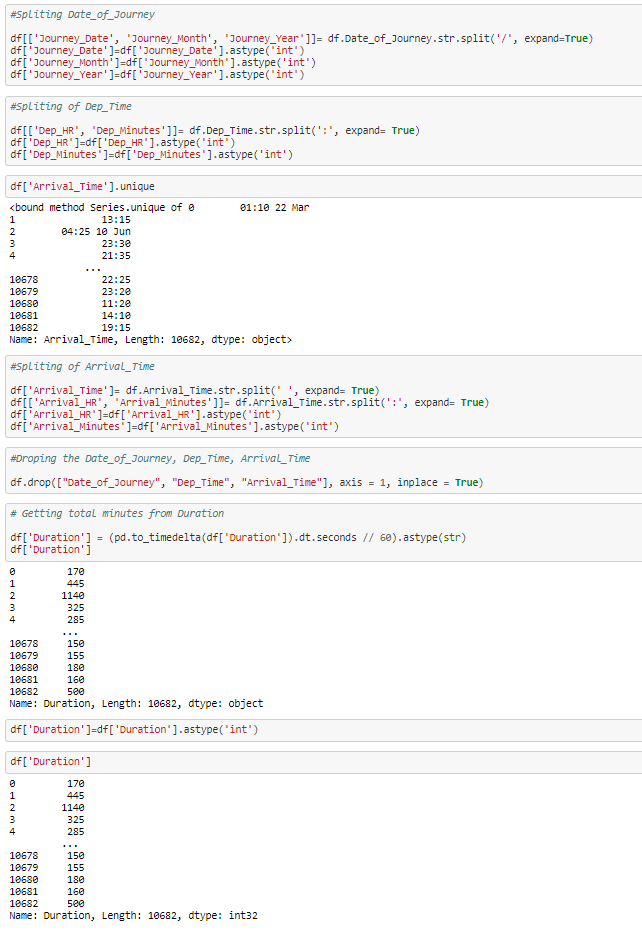
* **Removing Null values**

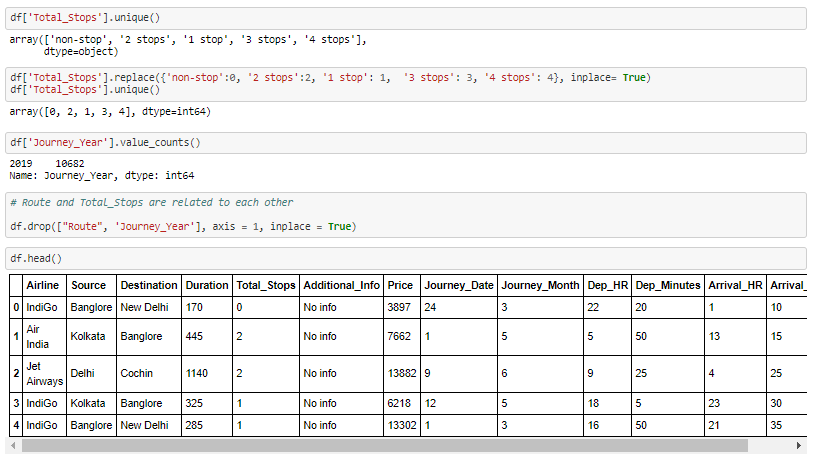
****

Observations: Now no null values present in dataset.

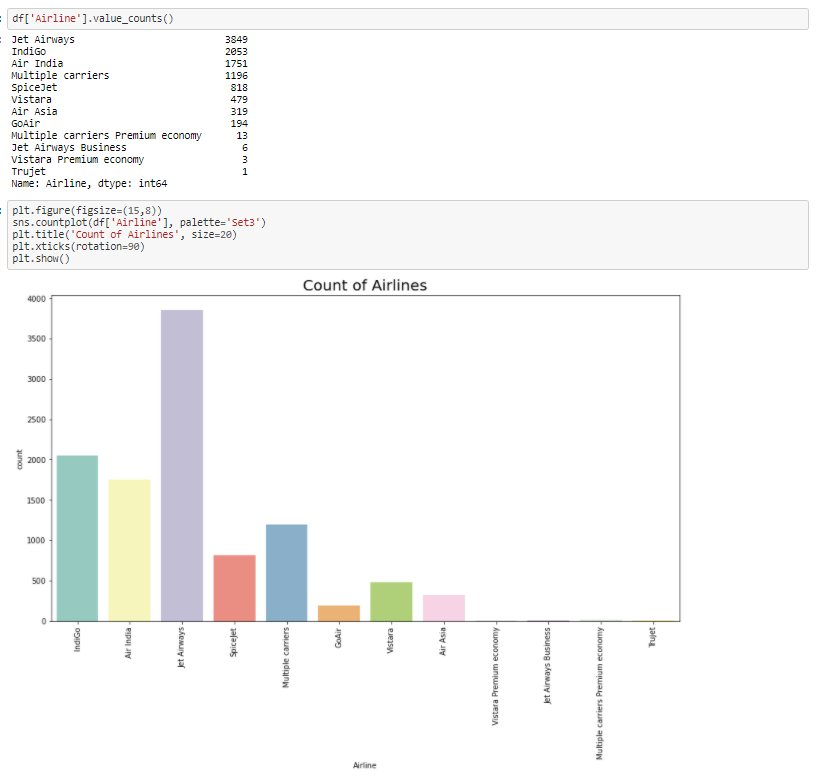
* **Feature Engineering**

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

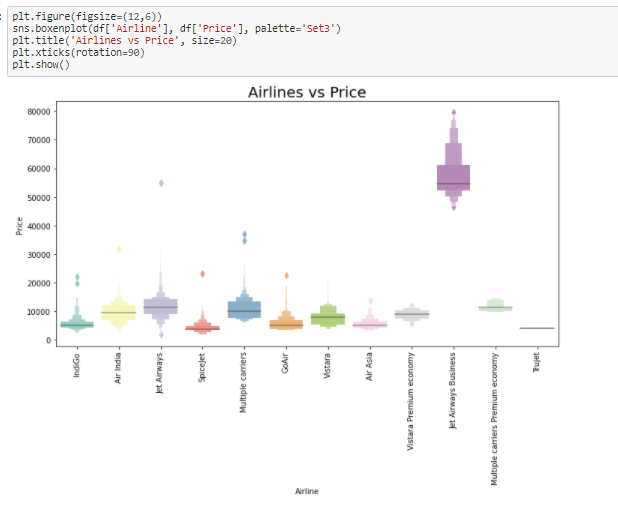




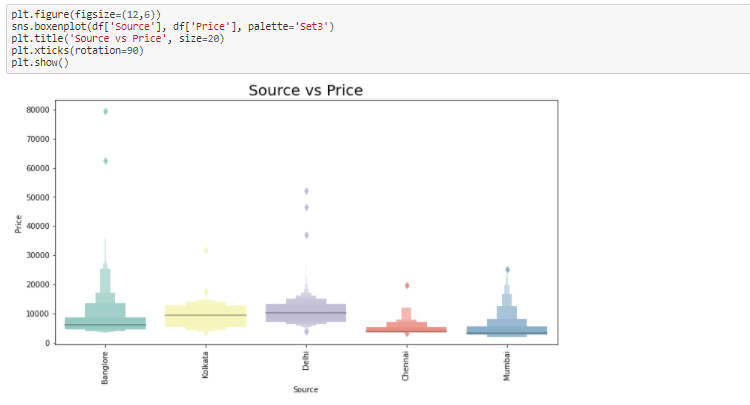
* **EDA**



Observations: There are most of Jet-Airways having almost 3800 counts.



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.



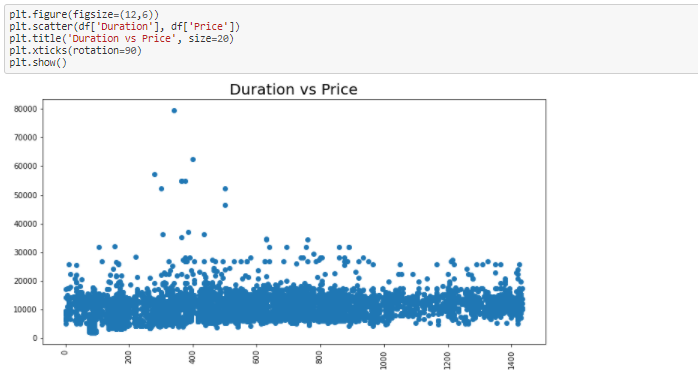
Observations: Flight which are started from Lokata and Delhi has more price distributed in 5000 to 15000.



Observations: For month starting i.e. on date 1, there is maximum price value.



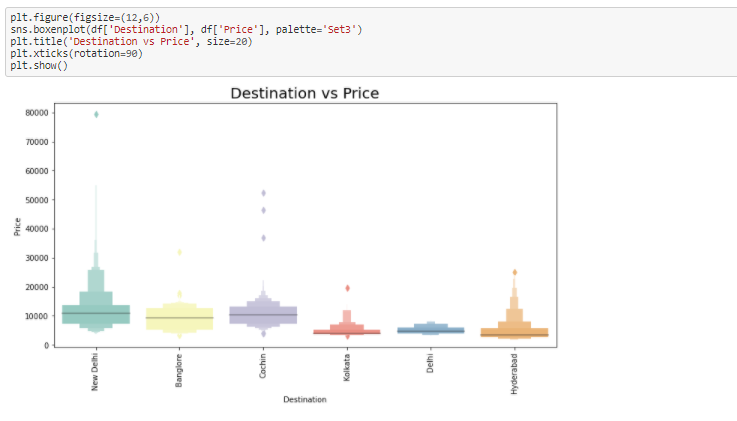
Observations: There is maximum flight price in 3rd i.t. in March month.



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



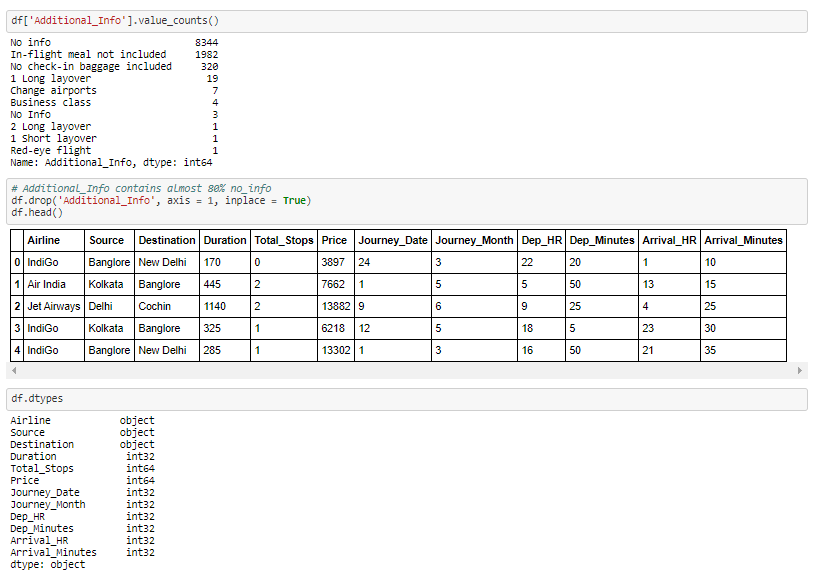
Observations: As number of total\_stops increases, price of flight also increases.



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

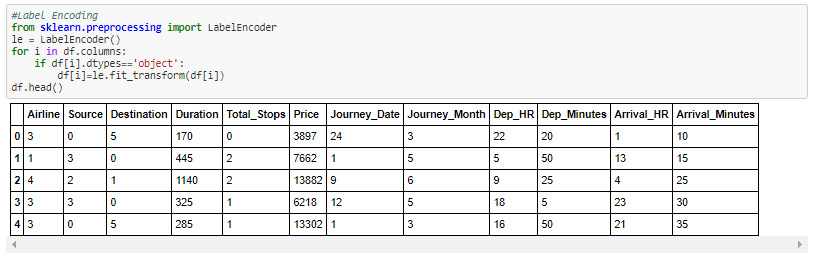


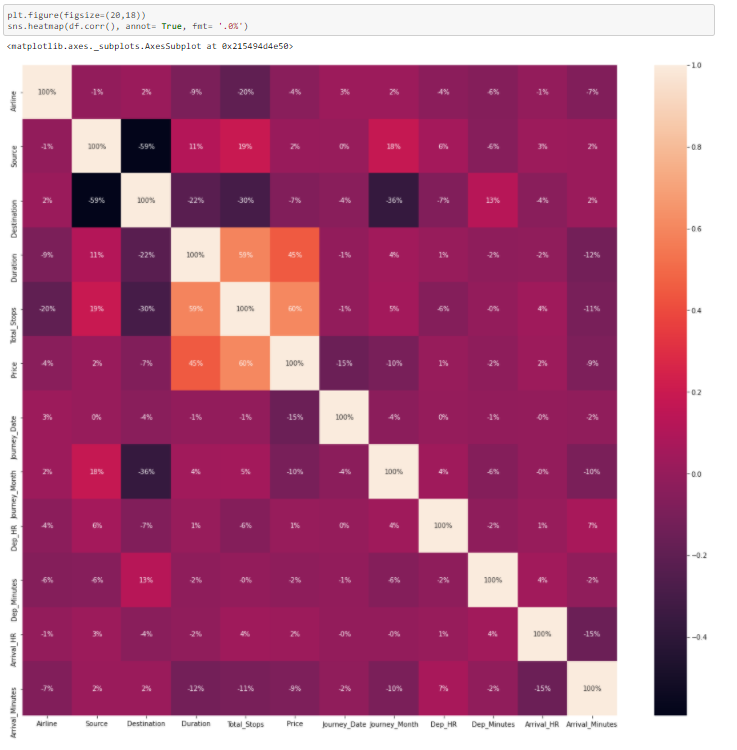
Observations: For Business class there is higher price.



* **Encoding**

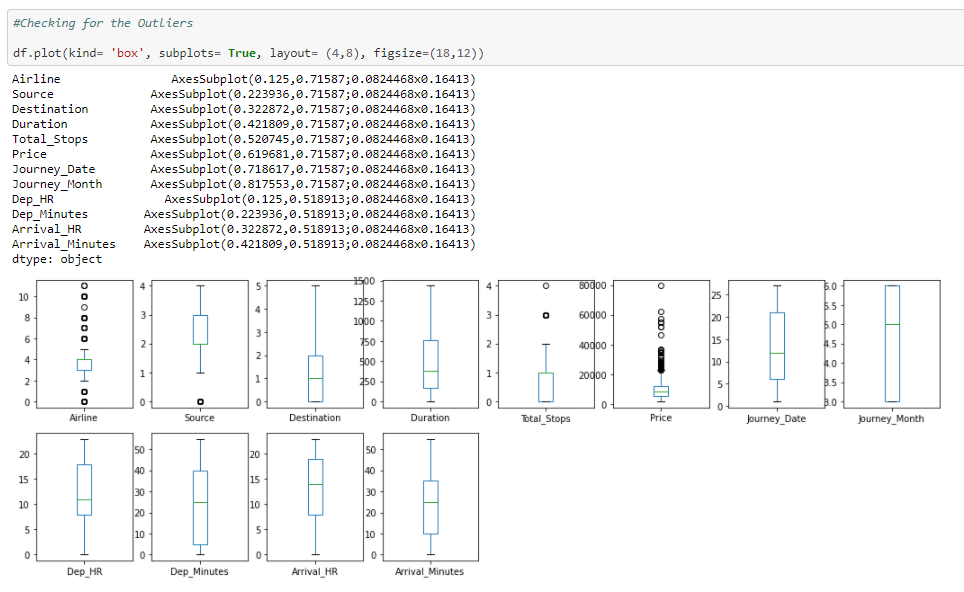
Here we are going to convert all object data type into numerical data type by using Label Encoding.





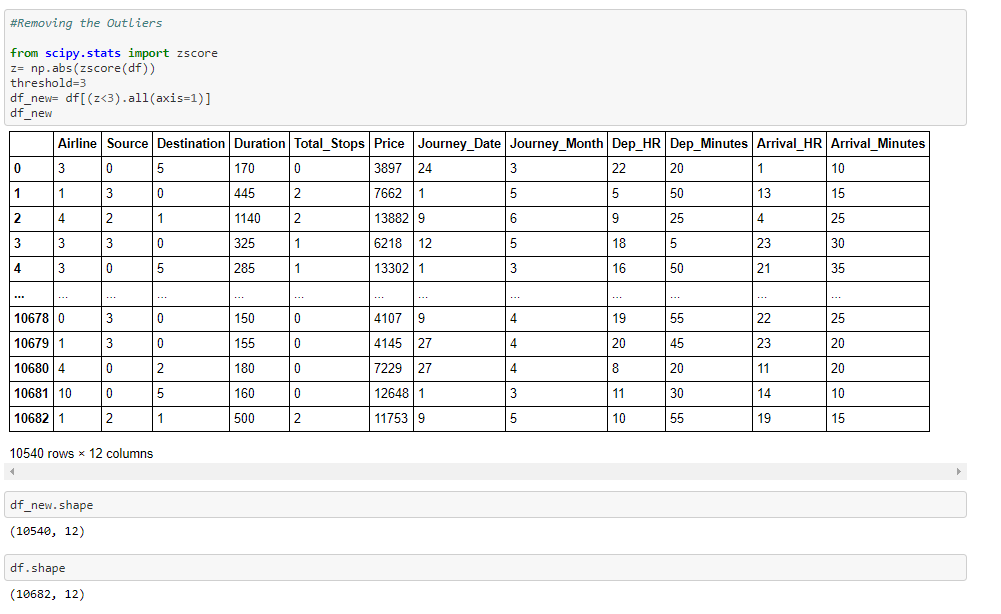
Observations: Flight price is highly correlated with total stops and Duration of flight.

* **Checking For Outliers**

****

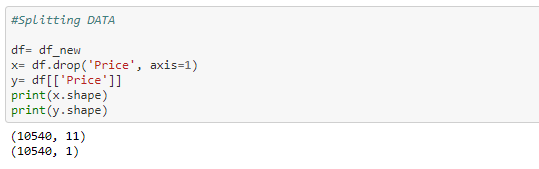
Observations: There are maximum outliers are present in Price. That one we need to remove.

* **Outliers Removing**

****

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

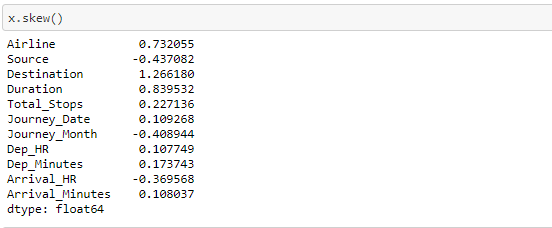
* **Splitting Data**

****

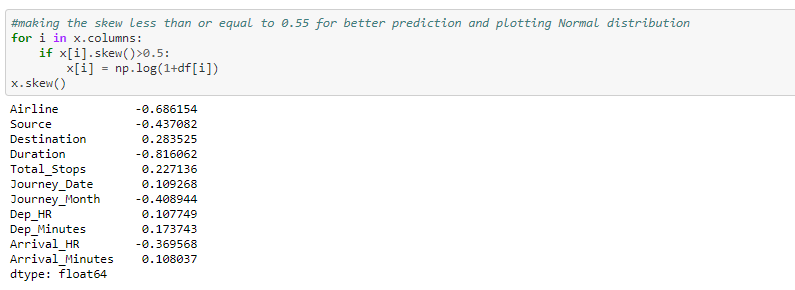
* **Checking for Skewness**

****

Observations: We can see that duration, Dep\_Hr are somewhat left skewed.

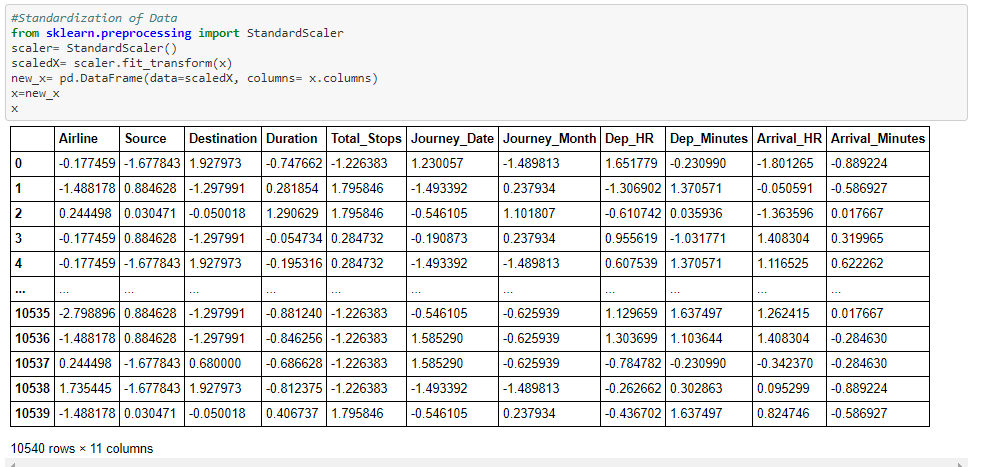
****

* **Skewness removal**

****

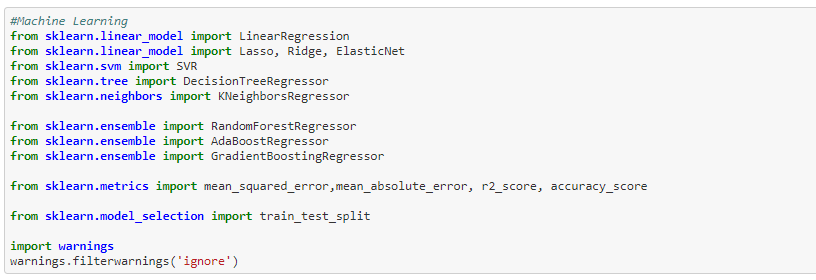
* **Standardization of Data**

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

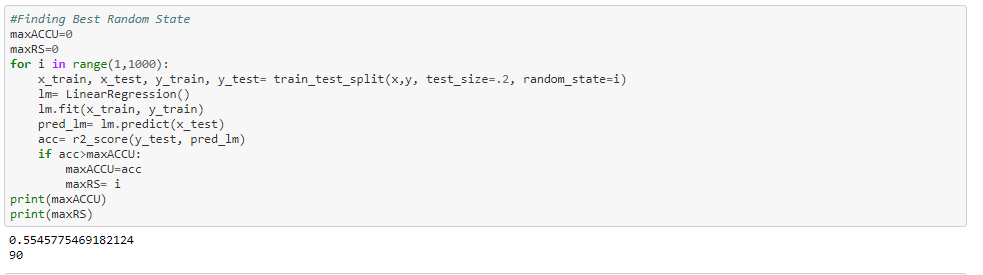


* **Machine Learning**

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

****

1. Finding best random state.

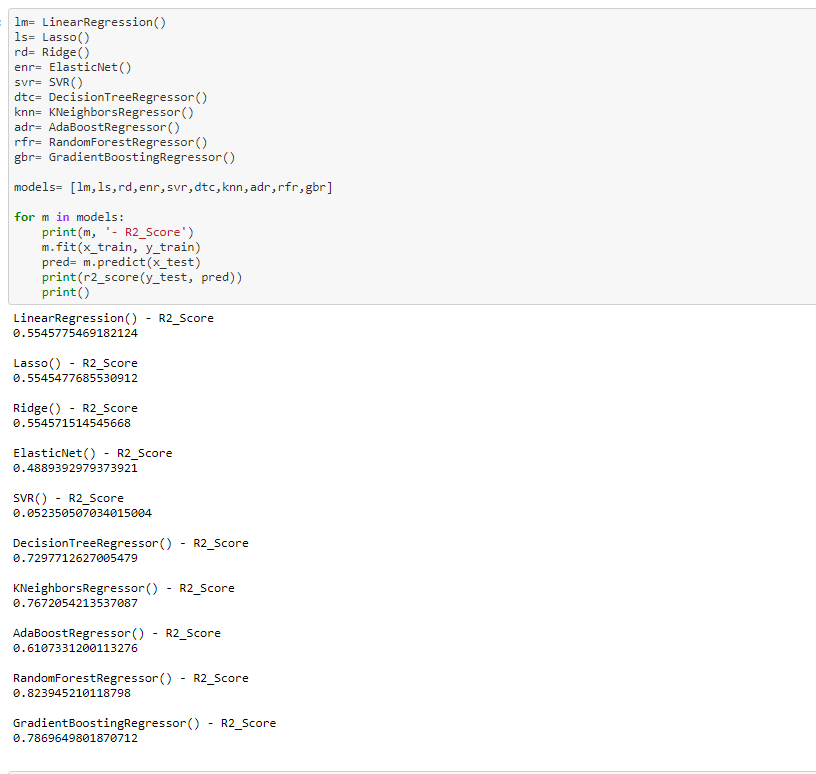


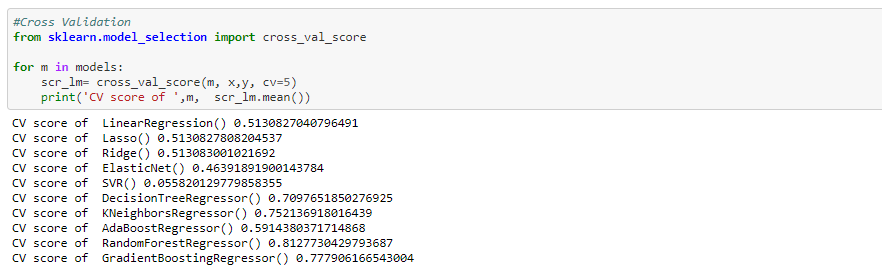
Here we can see that best random state is 90.

1. Splitting the data for training and testing purpose.



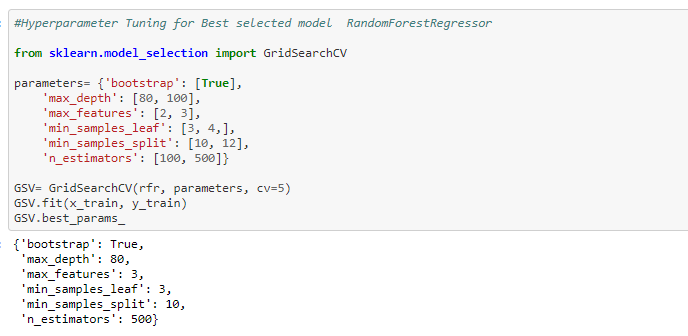
1. Selection of best model.



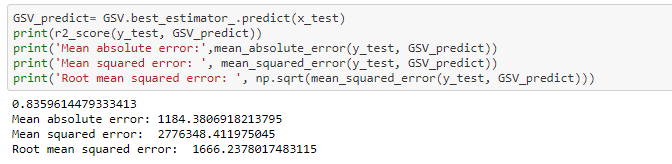


Observations: From above we can observe that for RandomForestRegressor model there is less difference in r2\_score and cross validation r2\_score so we will select RandomForestRegressor model for final predictions.

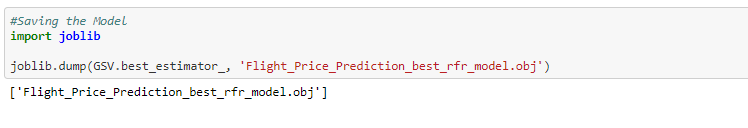
1. Hyperparameter Tuning for Best selected model RandomForestRegressor



1. Getting best of RandomForestRegressor model from best Parameters.



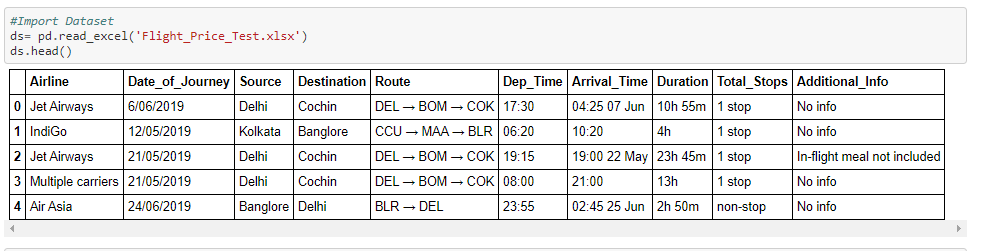
* **Saving the best model**

****

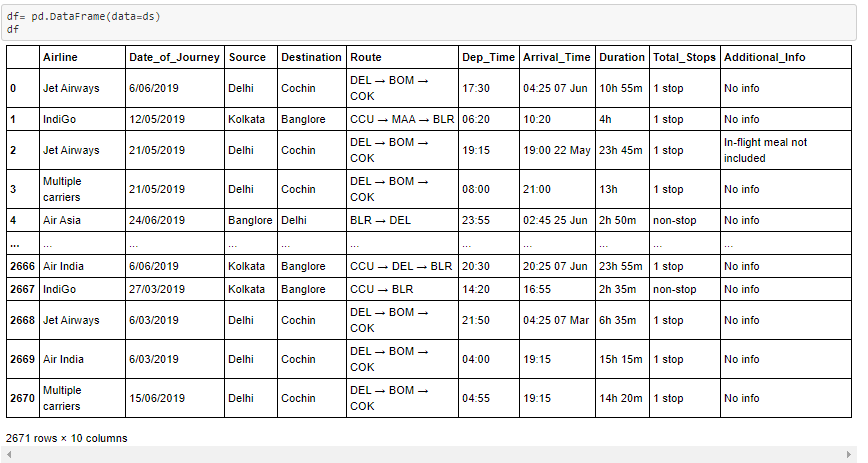
* **Working on Test Data**

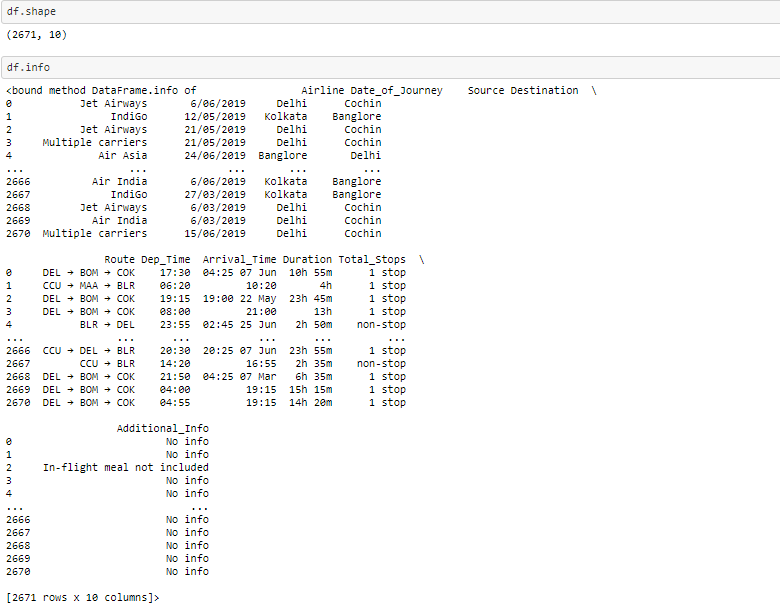
Now we will do all data cleaning, Feature Engineering and EDA on test data, then we will pass this data into our best selected model.

1. Import Test data

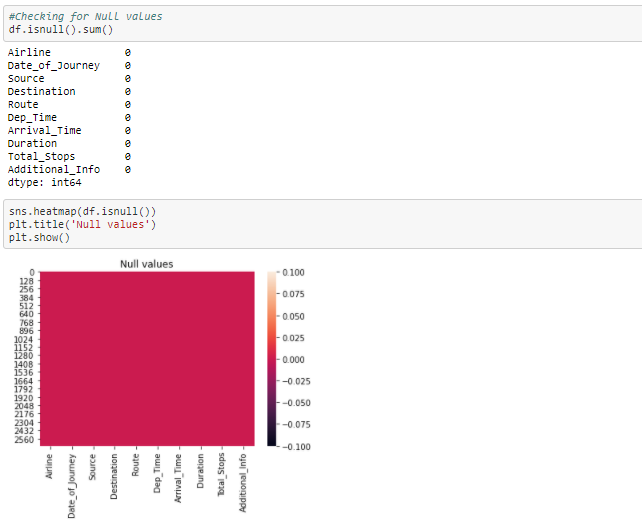


1. Fitting test data into dataframe and use of pandas dataframe methods.



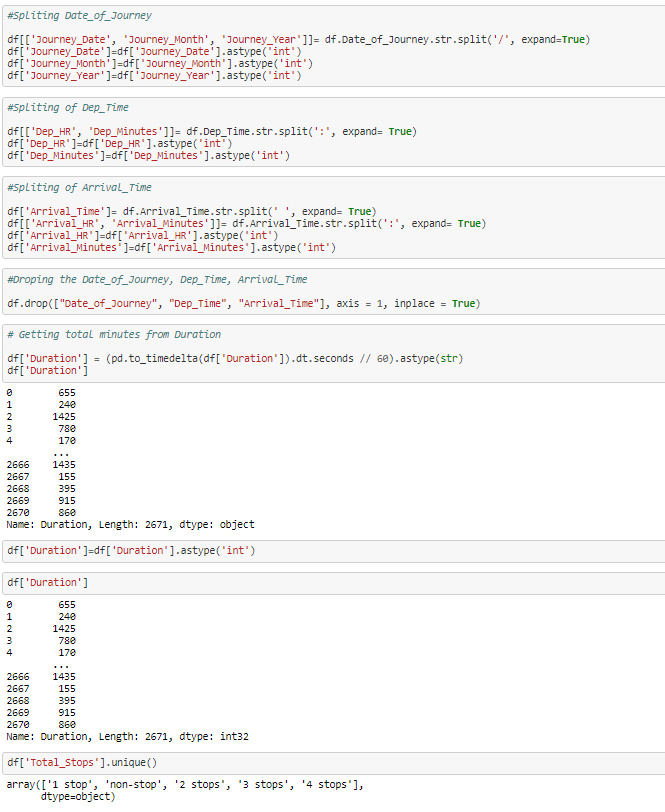


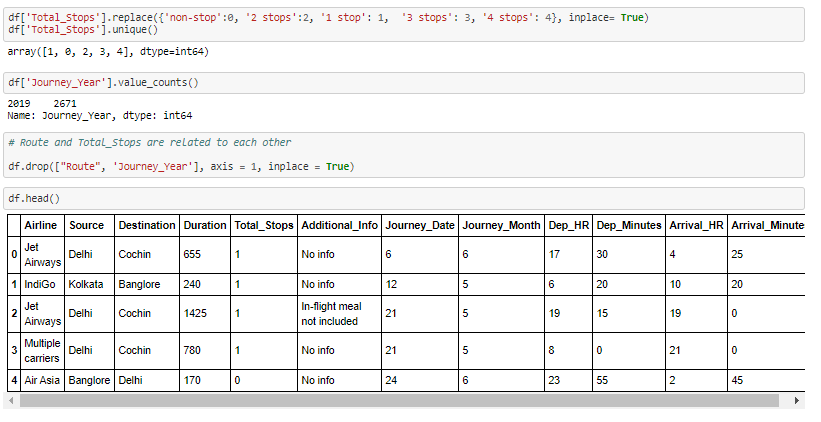
1. Checking of Null Values and removing if Null values present.



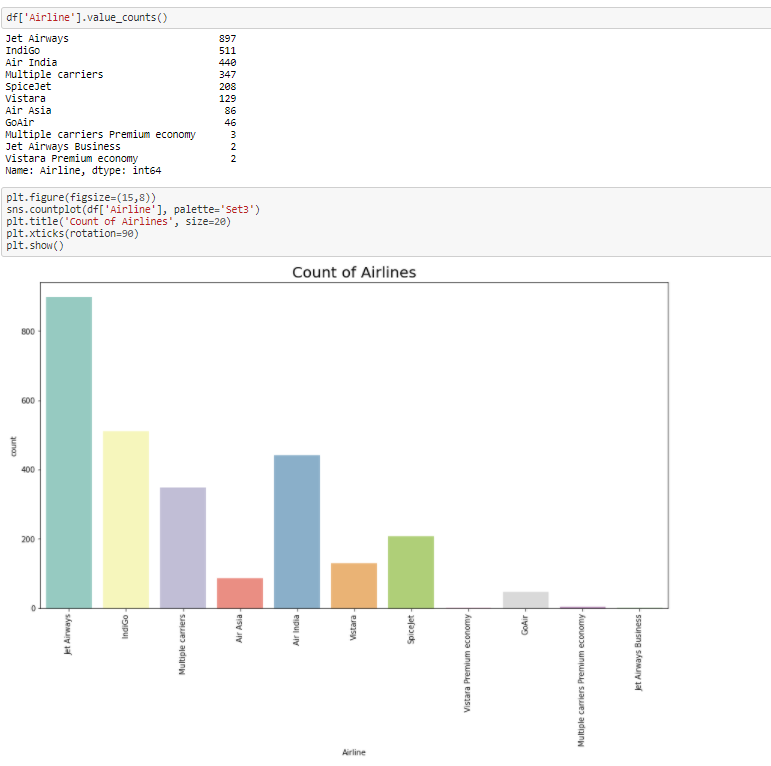
Observations: No null values are present in dataset.

1. Feature engineering for Test Data

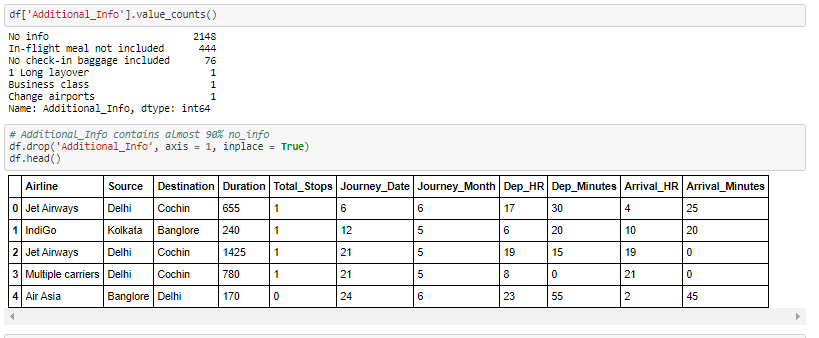


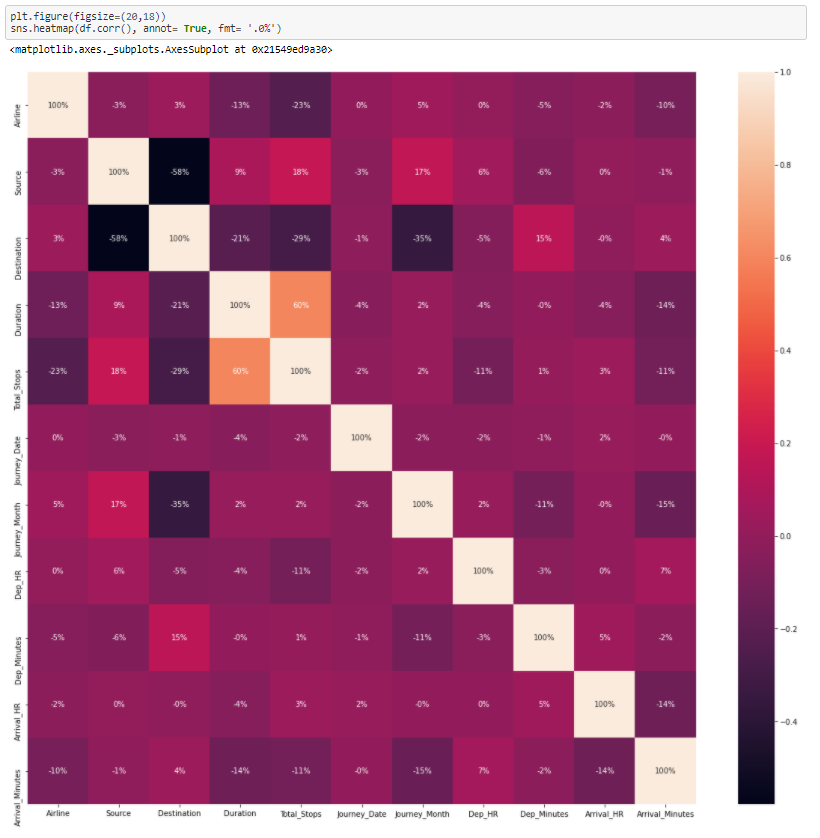


1. EDA for Test data



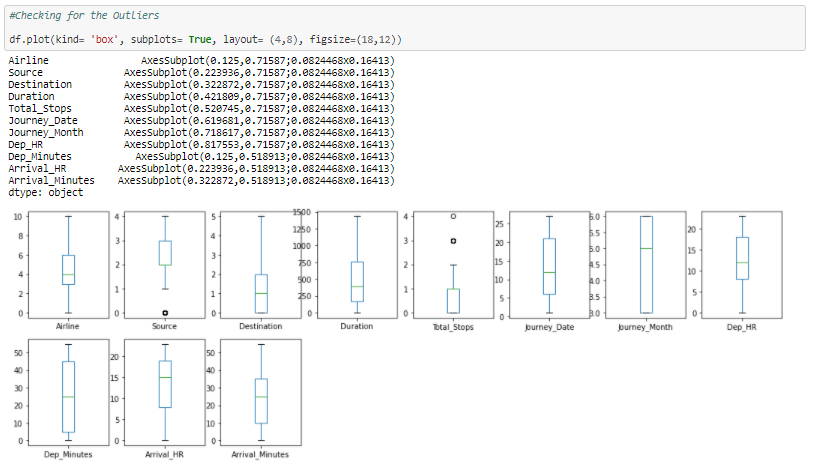
Observations: There are most of Jet-Airways having almost 900 counts.





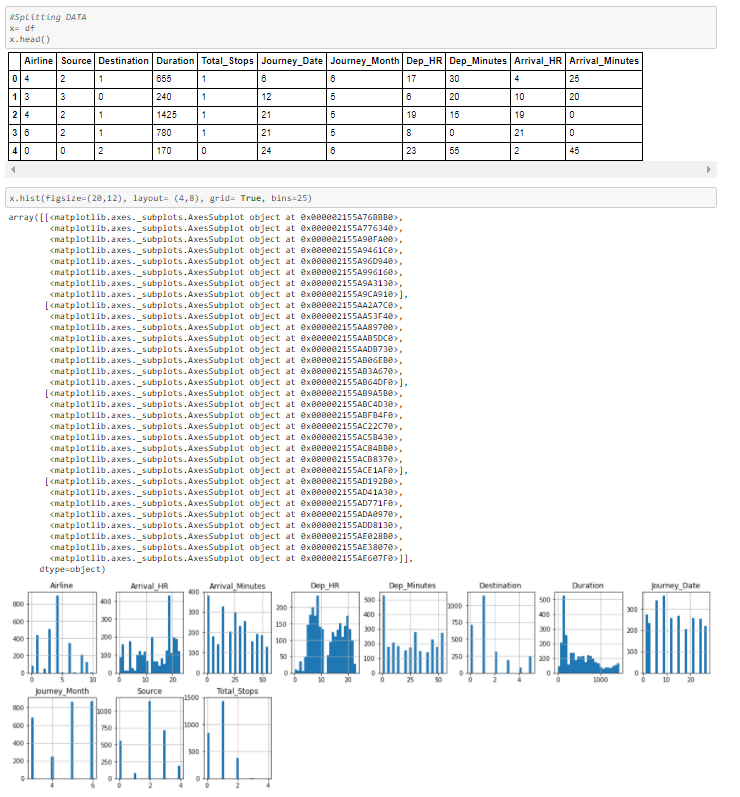
Observations: We can see that Duration is highly correlated with total stops.

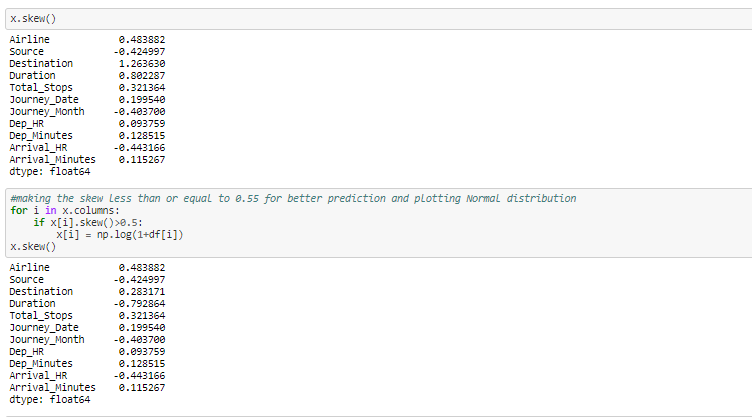
1. Checking for Outliers and removing if present.



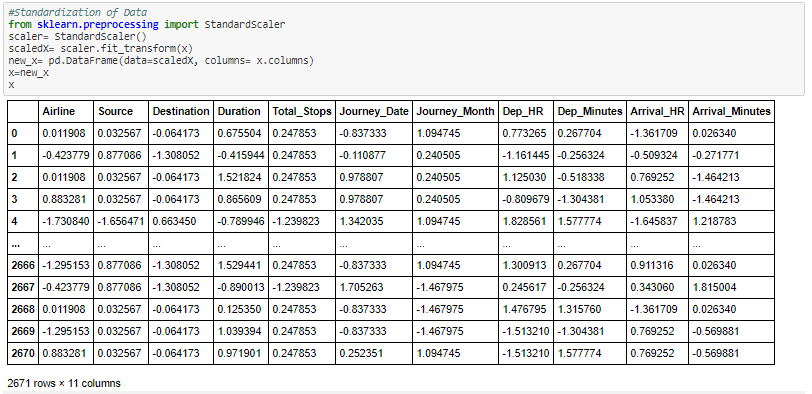
Observations: We can see that there no outliers present.

1. Checking for Skewness and removing if present.

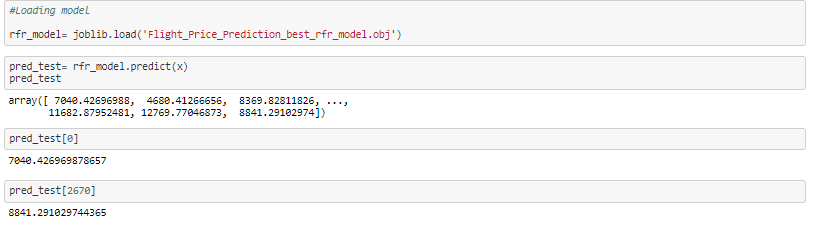




1. Standardization of Test Data



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

****

* **Conclusion**

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

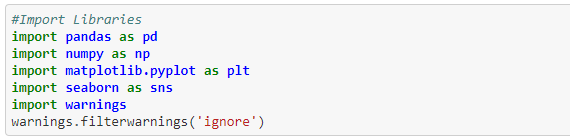
# Project Name: HR Analytics Project

*-Analysis by Vaibhav Banke*

**About Project:**

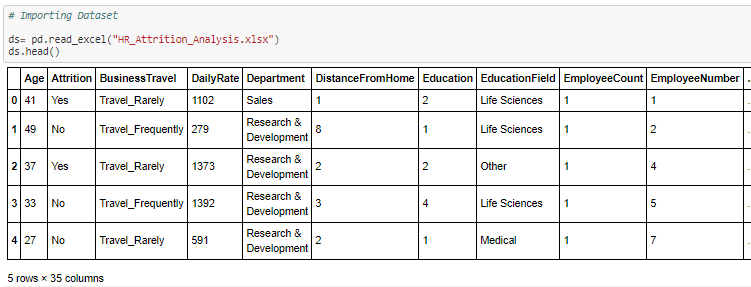
This HR Analytics Project takes HR data and uses machine learning models to predict which employees will be more likely to leave company by given some attributes. Such model would help an organization predict employee attrition and define a strategy to reduce costly problem. For each employee, in addition 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.

* **Import Libraries**



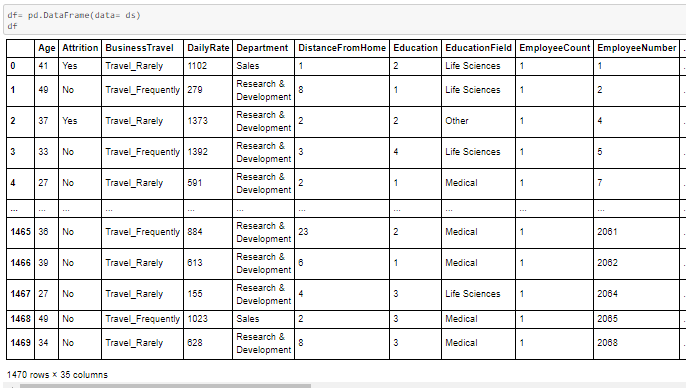
Here we import libraries for getting data and for making EDA Analysis.

* **Import Dataset**



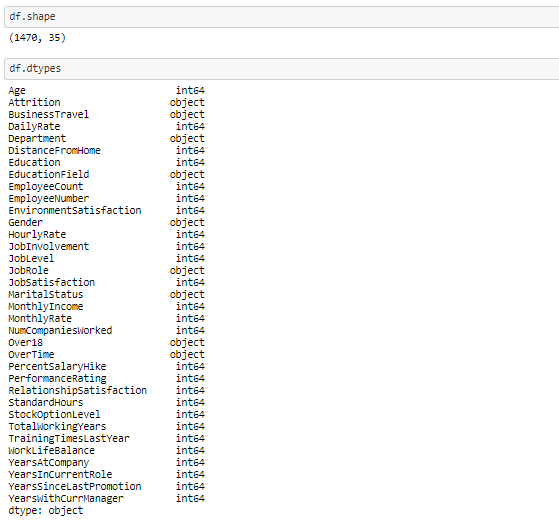
We can see now top 5 rows of Dataset.

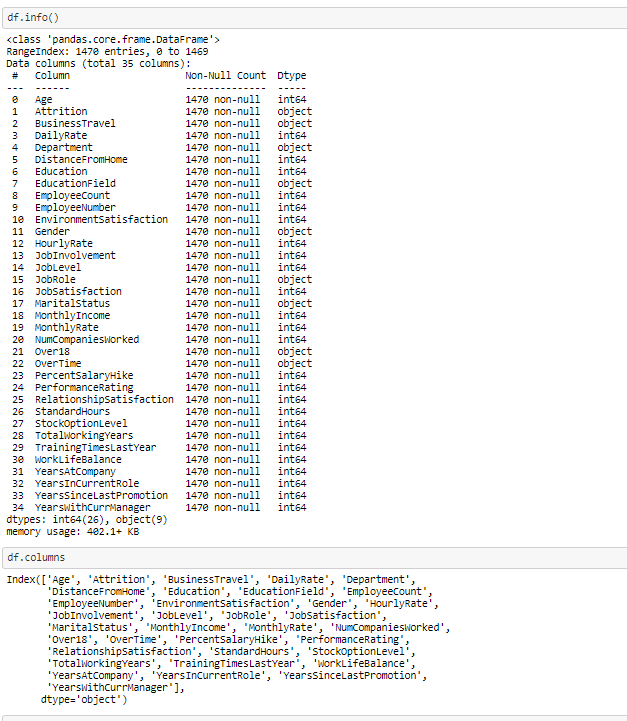
* **Fitting data into Pandas Dataframe**

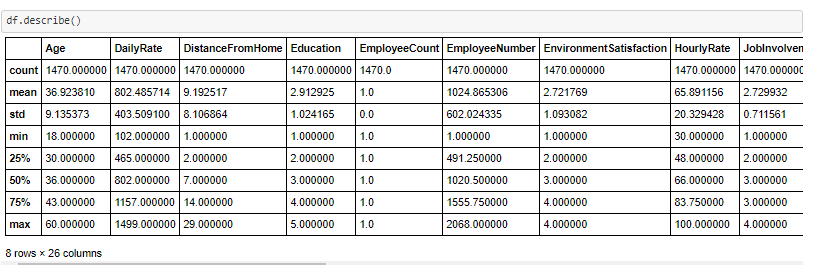


Here we fitted our dataset into Pandas Dataframe, so that we can do all Dataframe methods.

* **Appling Pandas DataFrame methods**

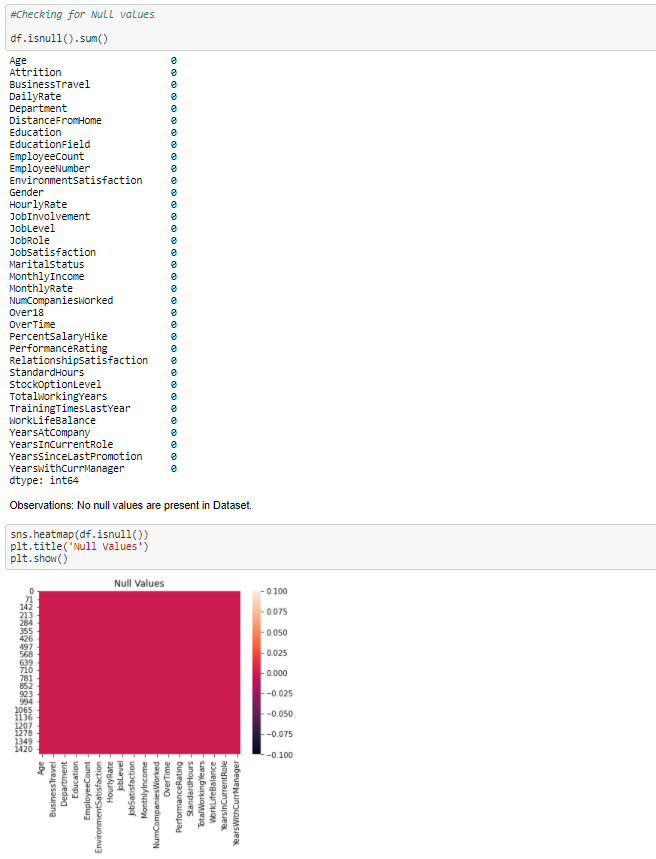






Observations: Age of the employee is in between 18 and 60. For all columns count is 1470, that means no null values present in dataset. EmployeeCount and StandardHours are single value column. Mean and 50% values are same for almost all columns. Also difference in 75% and max value is less that means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

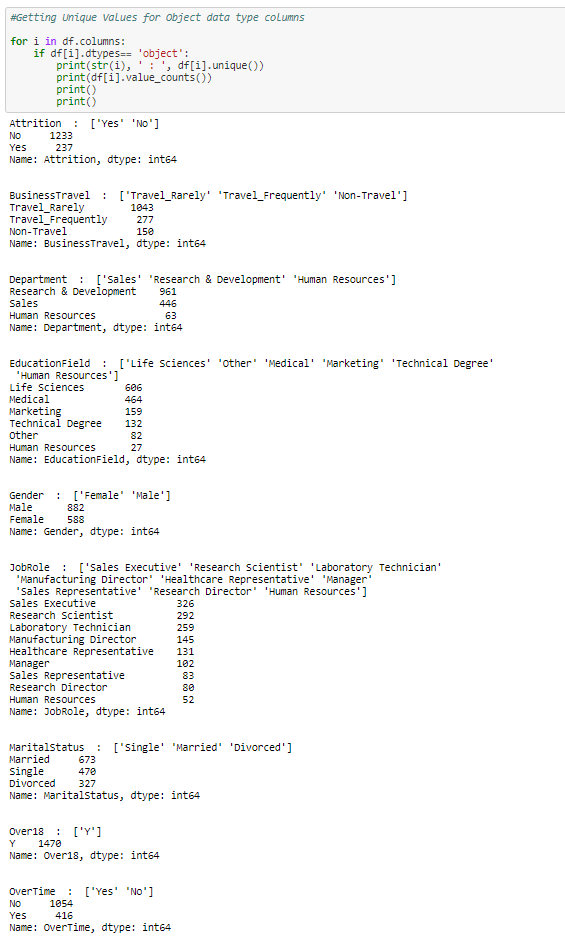
* **Checking for Null values**



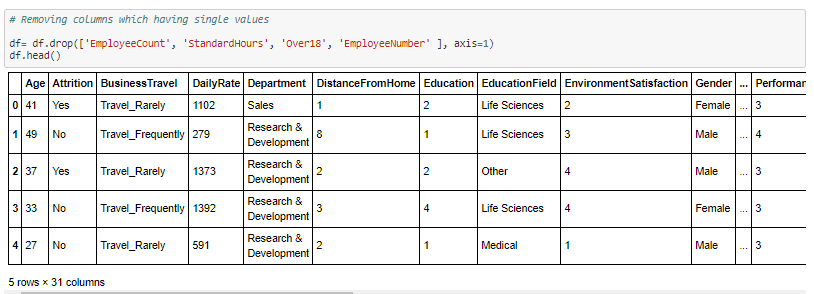
Observations: No null values are present in Dataset.

* **Feature Engineering**

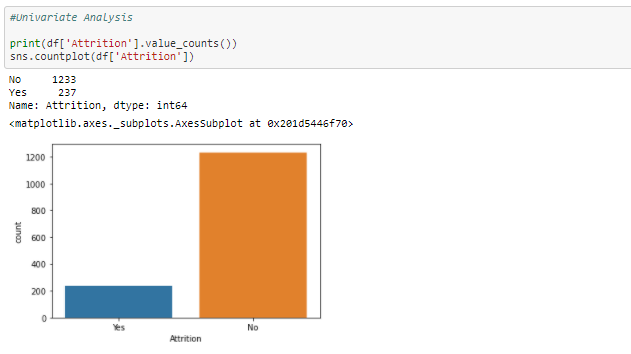
Here we are going to add more features in our dataset from existing one and deleting unwanted features.



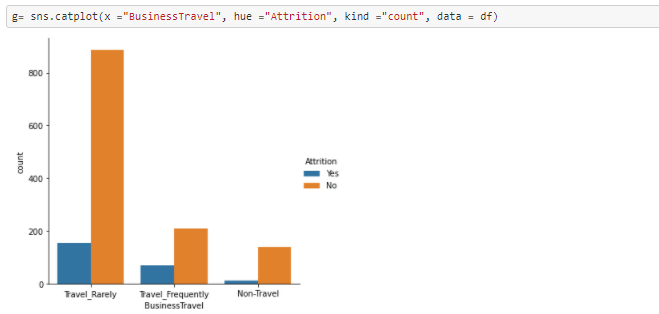
Observations: Here we can see that some features having single value, those we need to remove as shown in below



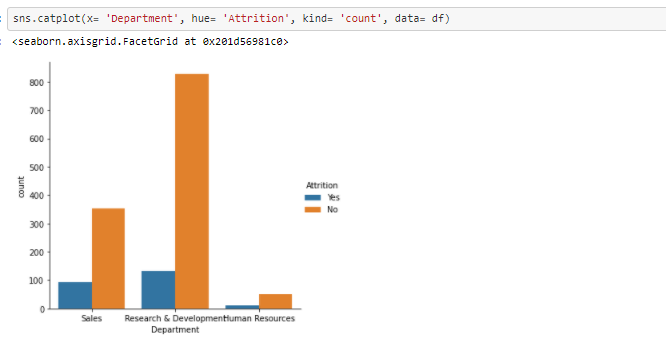
* **EDA**



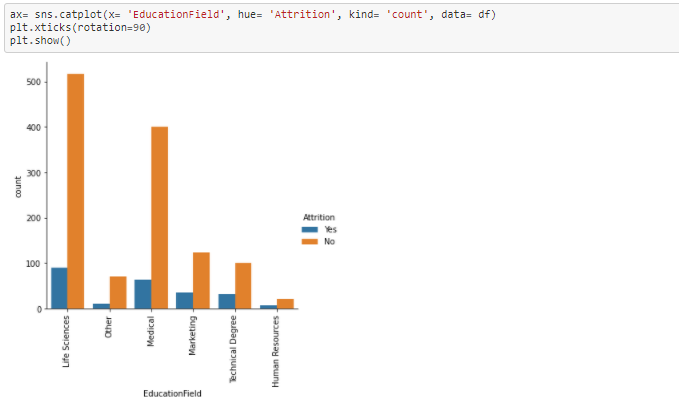
Observations: There are 237 employees left the company and 1233 employees are currently working in company.



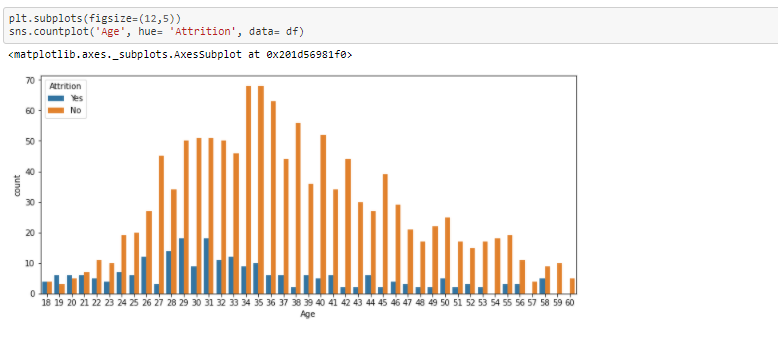
Observations: We can see that, rate of employees leaving is more in Frequently Travel than Rarely and Non\_Travler.



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



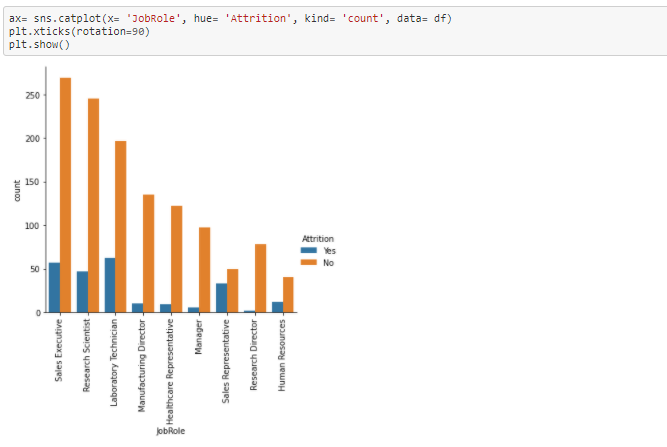
Observations: We can see that, rate of employees leaving is more in Life Science and Medical.



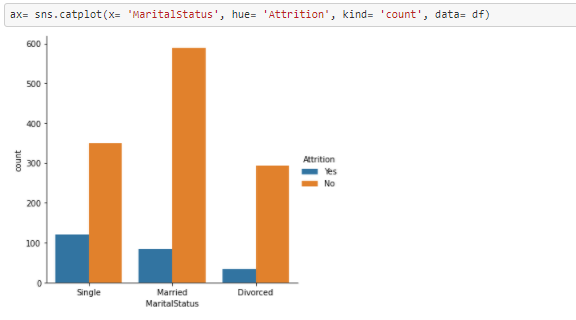
Observations: Maximum Employees at Age of 34 to 35 are currrently working in company where as maximum employees with Age of 29 and 31 are leaving the company.



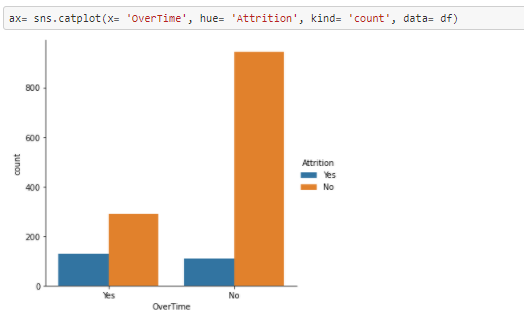
Observations: There is more Male employees are currently working in company. Also Male employees have left the company.



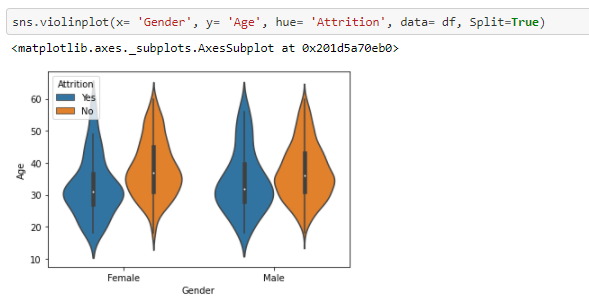
Observations: Currently employees with job role of Sales Executive and Research scientist are in company.



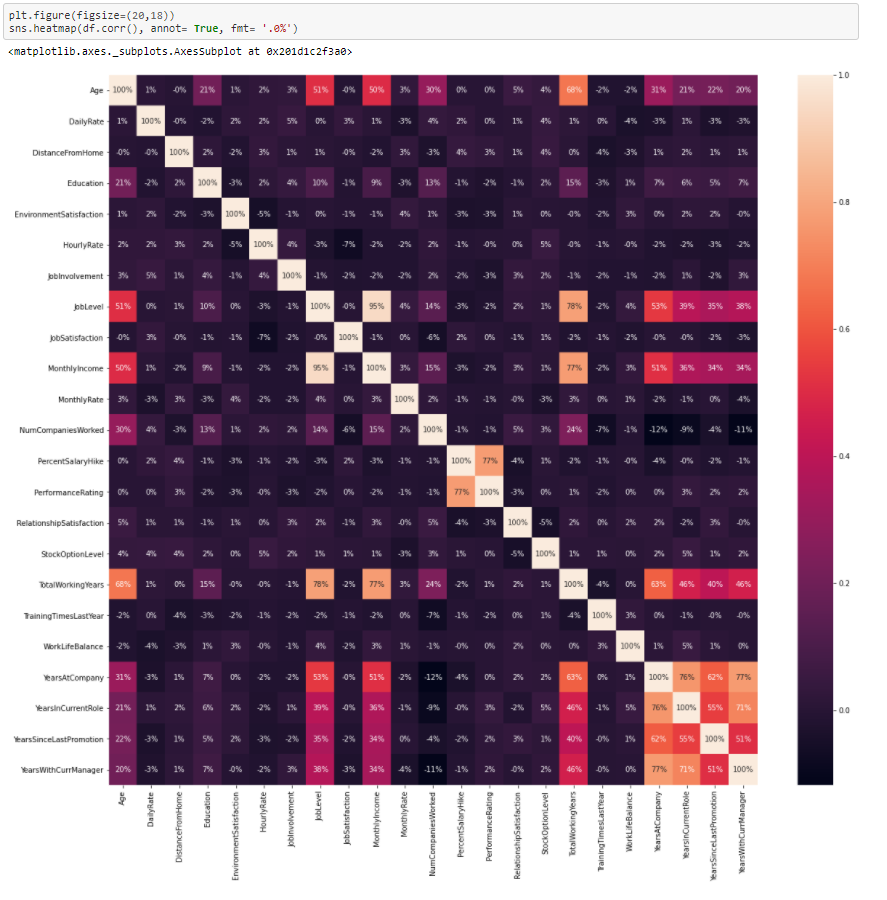
Observations: Currently Married employees are more in company. And where as Single employees have left.



Observations: Currently maximum employees are not doing overtime.



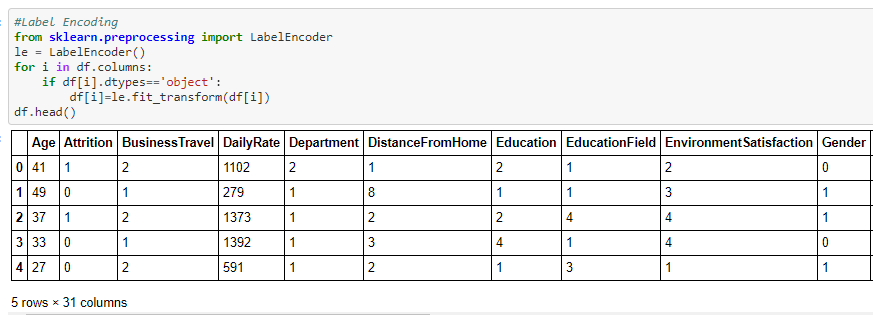
Observations: Female and male employees with age of 29 to 32 have left the company.



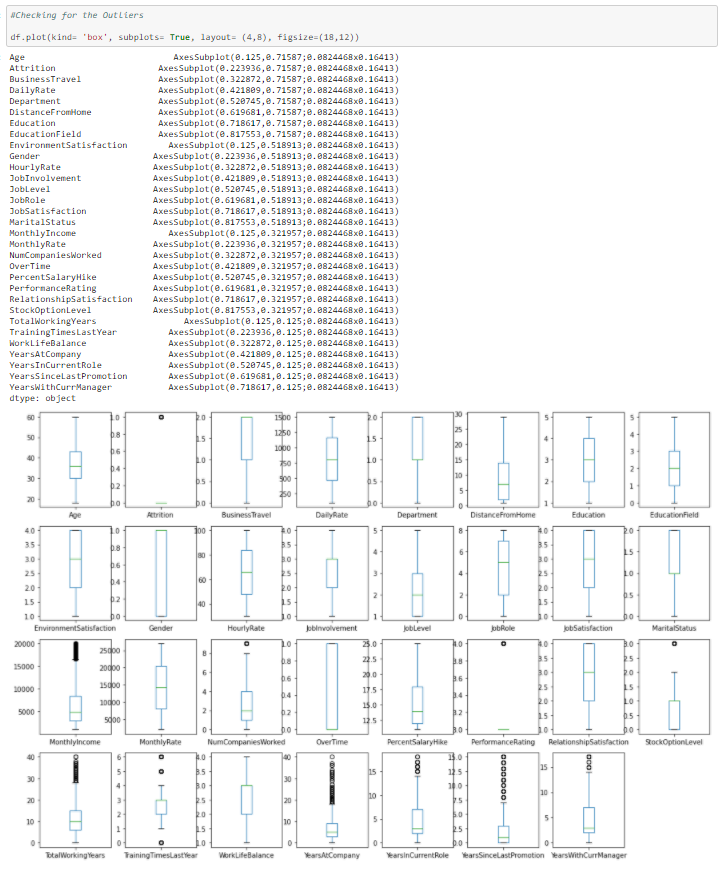
Observations: Total working years has higher positive correlation with age as longer working will get older employees. Also Total working years has higher correlation with jo level and monthly income. Higher coorelation of monthly income with job level.

* **Encoding**

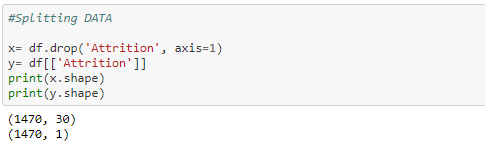
Here we are going to convert all object data type into numerical data type by using Label Encoding.



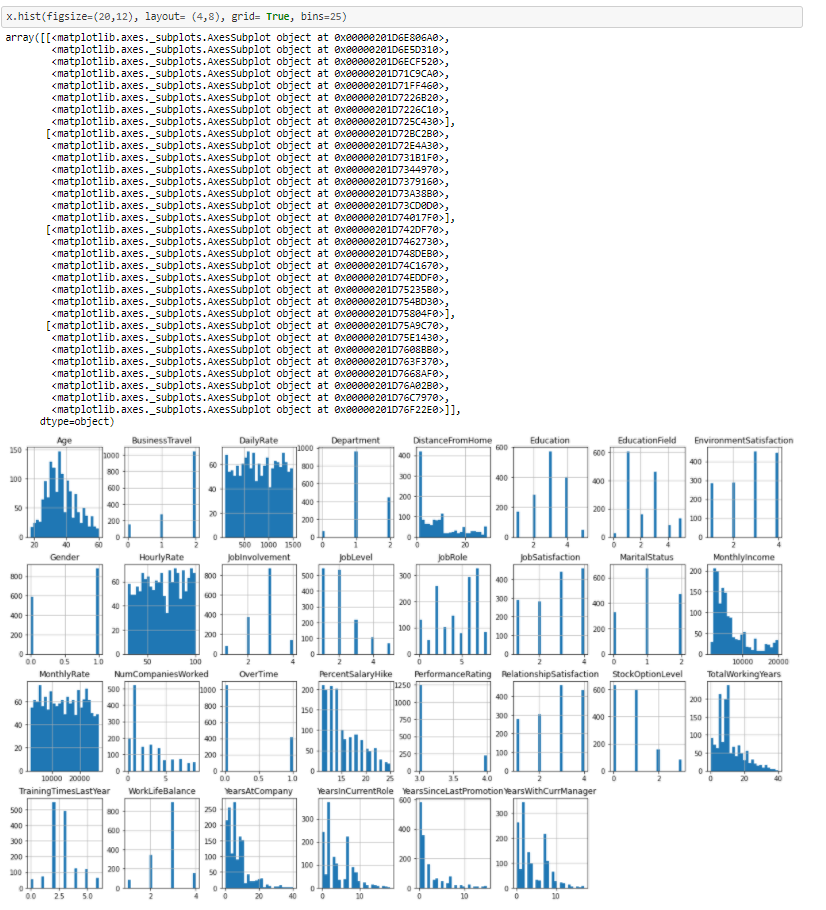
* **Checking For Outliers**

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

* **Splitting Data**

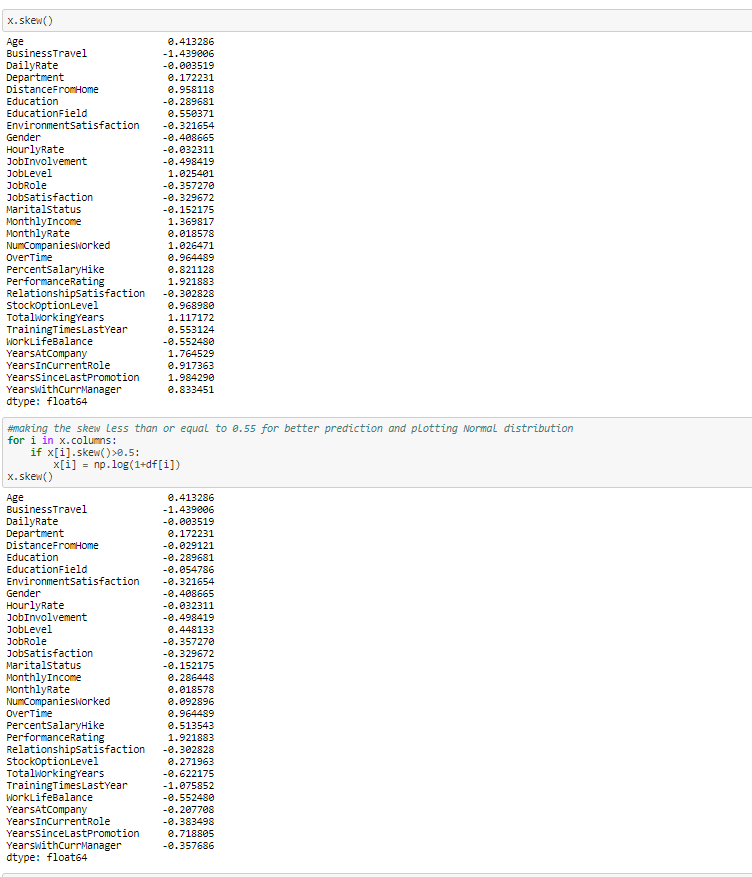
****

* **Checking for Skewness**

****

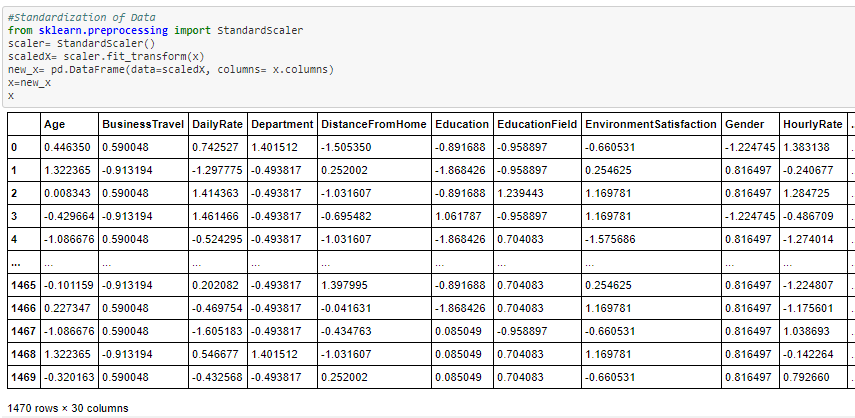
Observations: Age, Daily Rate are looks normally distributed. Where as rest all columns are some what left skewed as data values are not varies in terms of numbers.

* **Skewness removal**

****

* **Standardization of Data**

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

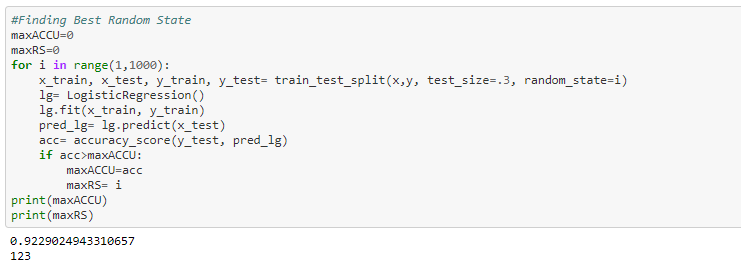


* **Machine Learning**

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

****

1. Finding best random state.

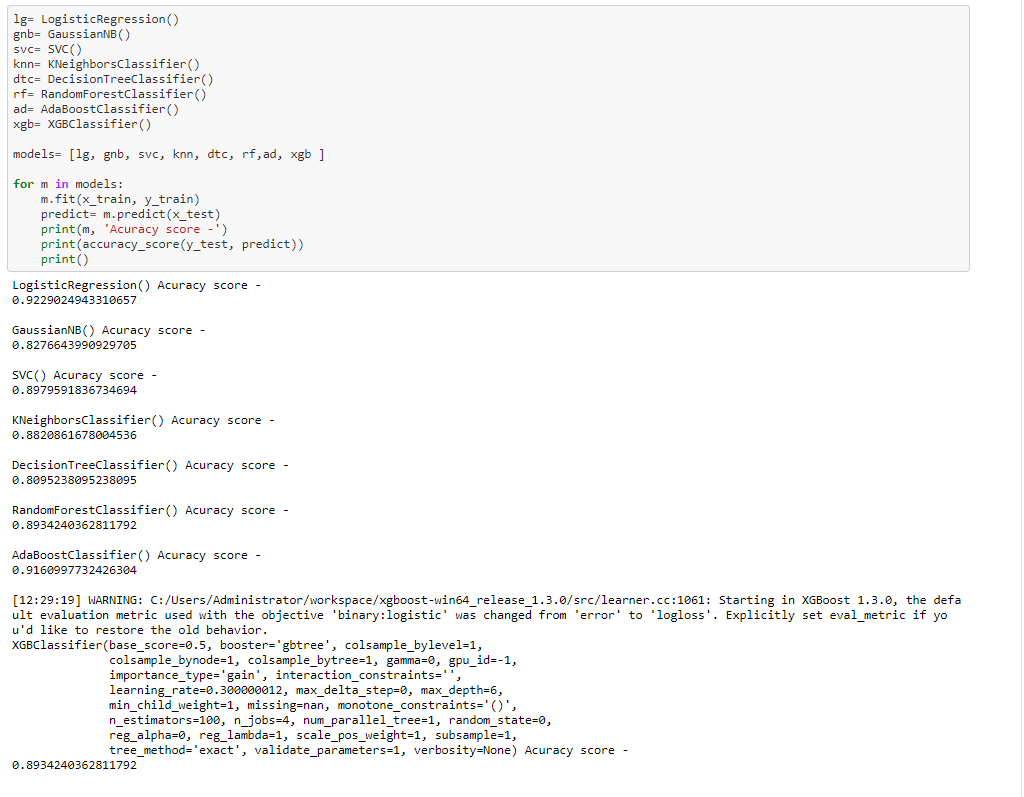


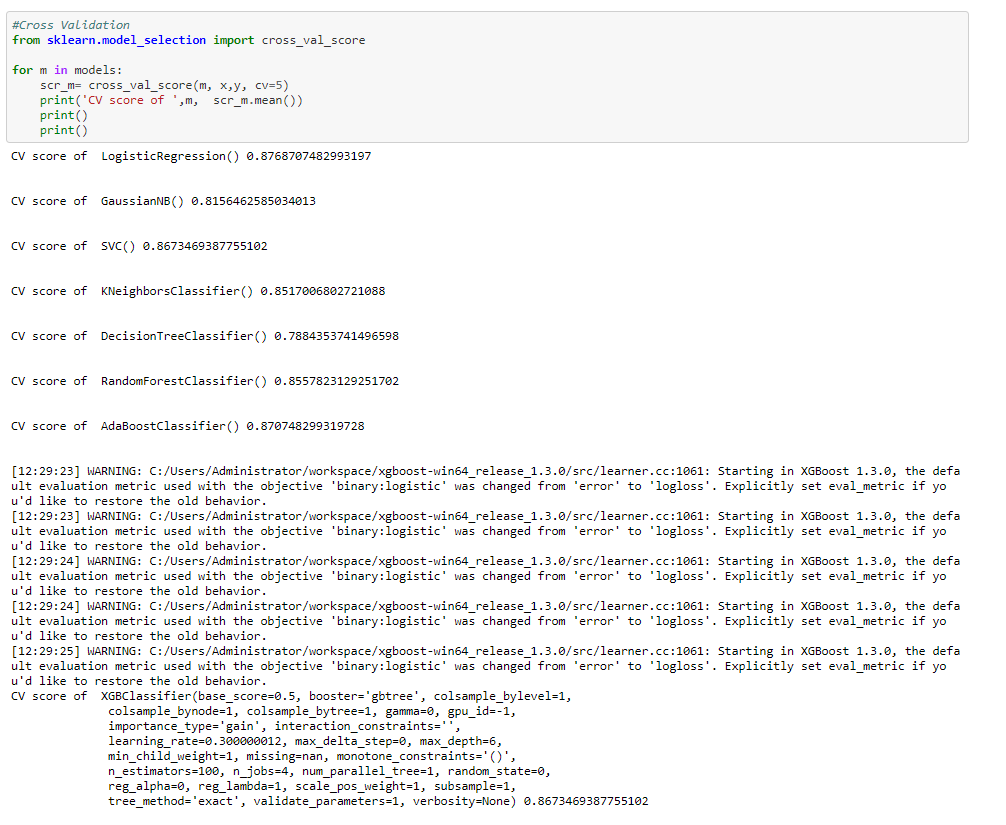
Here we can see that best random state is 123 which gives Accuracy of 92%..

1. Splitting the data for training and testing purpose.



1. Selection of best model.



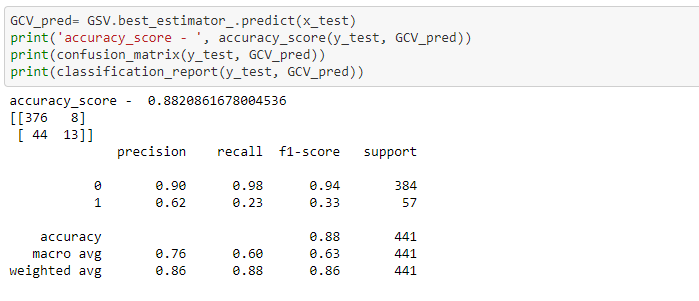


Observations: From above we can observe that for DecisionTreeClassifier model there is less difference in accuracy\_score and cross validation accuracy\_score so we will select DecisionTreeClassifier model for final predictions.

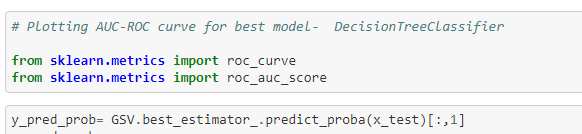
1. Hyperparameter Tuning for Best selected model DecisionTreeClassifier

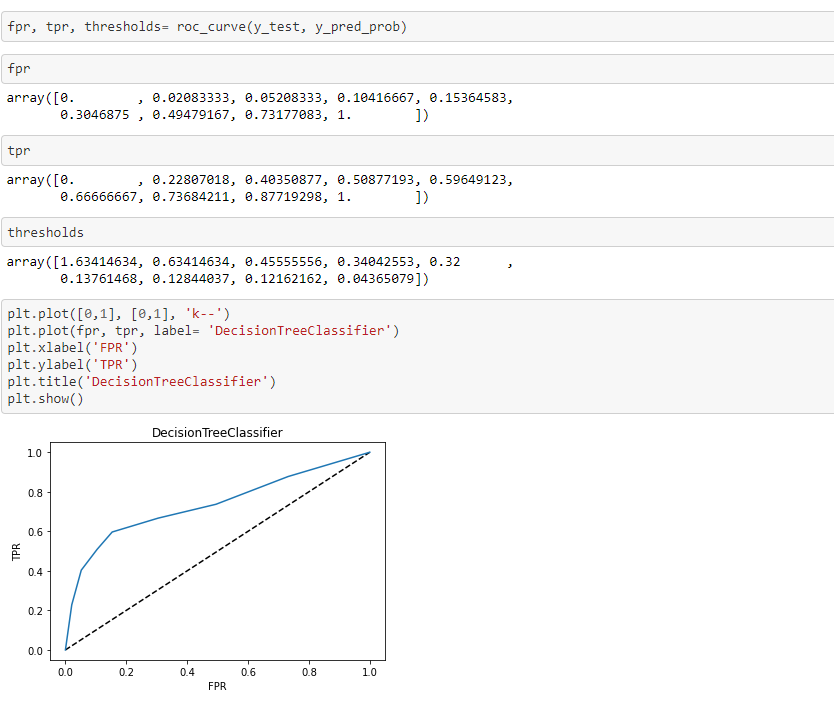


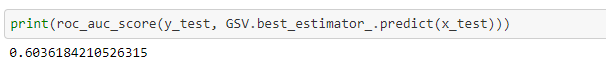
1. Getting best of DecisionTreeClassifier model from best Parameters.



* **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%.