**Employee Attrition Analysis with Python**

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.

**Problem description:**

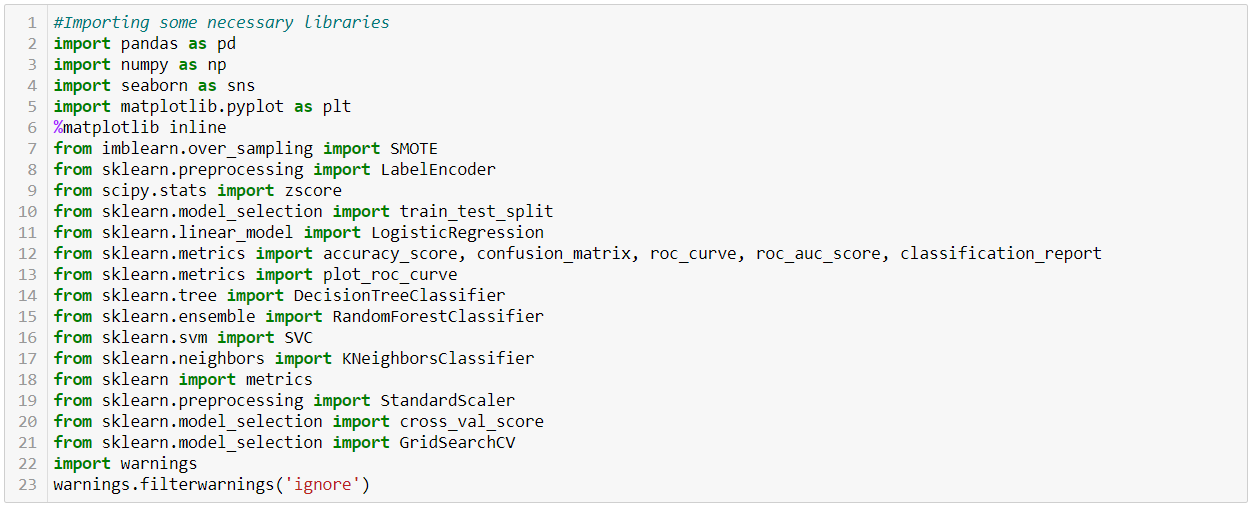
The goal is to predict the attrition which have classes in nature

**Methodology:**

Here, I am going to use 5 simple steps to analyze Employee Attrition

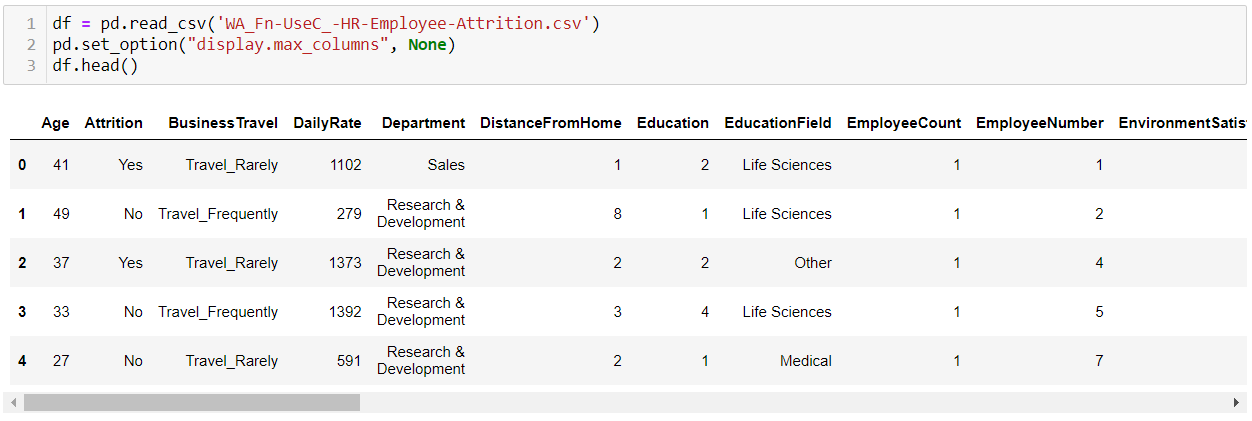
1. **DATA COLLECTION**
2. **DATA PRE PROCESSING**
3. **DIVIDING THE DATA into TWO PARTS “TRAINING” AND “TESTING”**
4. **BUILD UP THE MODEL USING “TRAINING DATA SET”**
5. **DO THE ACCURACY TEST USING “TESTING DATA SET”**

**Importing Library:**

****

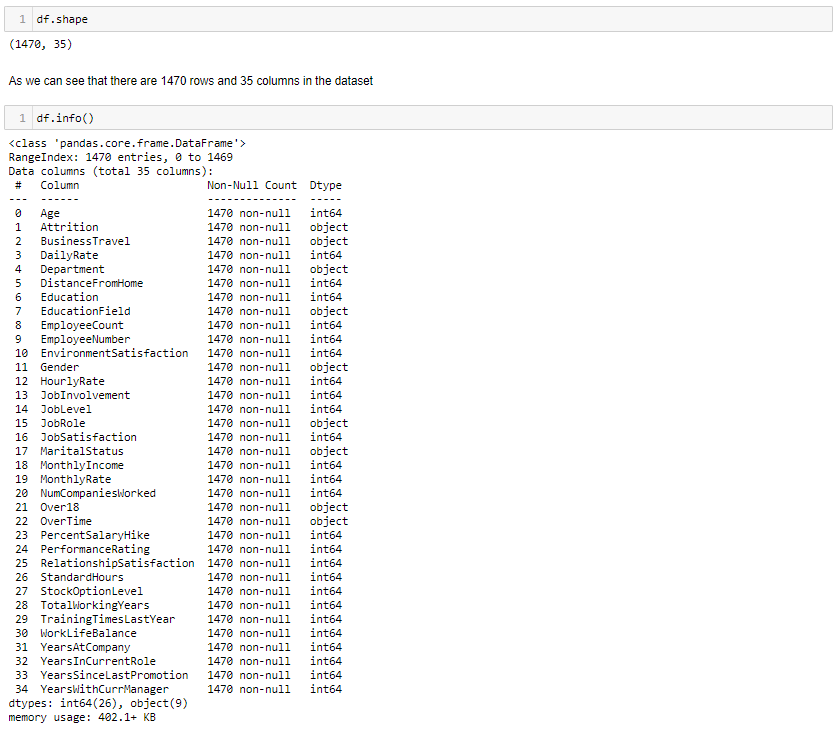
I am importing the all libraries which I required for EDA, Data visualization , Prediction and finding all Matrices . The reason of doing this is that it become easier to use all the import statement at one go and we don’t require to import statement at each point

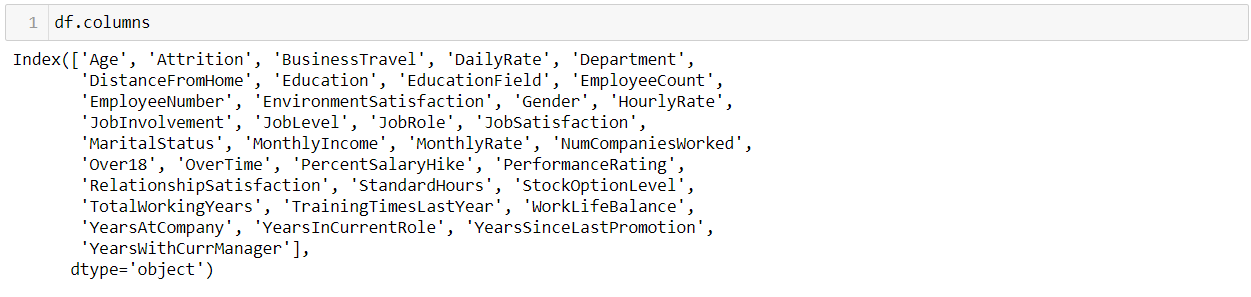
**Loading Dataset:**

****

Here I’m loading the dataset into variable i.e ‘ df ’ and processing the first 5 rows and all the data columns with pd.set\_option.

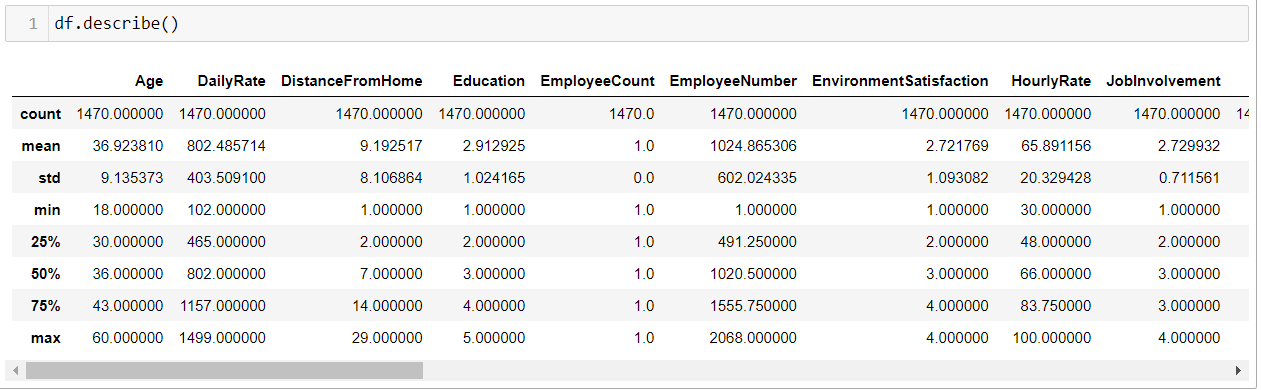
**Exploratory Data Analysis:**





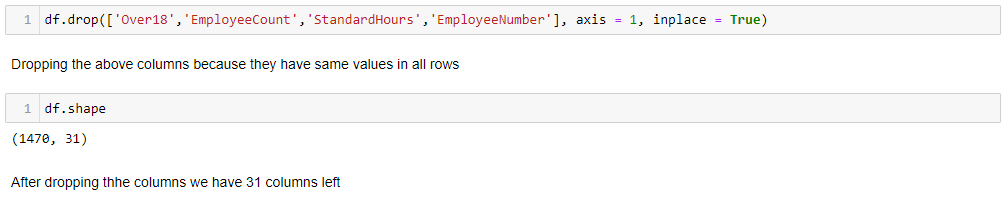
As seen in the above scenario :

* I checked the shape of the dataset where we get to know there are 1470 rows and 35 columns in this dataset.
* And we can see some of the columns have ‘Int64’ and some of them have ‘object’ dtype , There is no float dtype column in this dataset.
* And there are no nan’s present in this dataset.



**Detect the Missing values:**





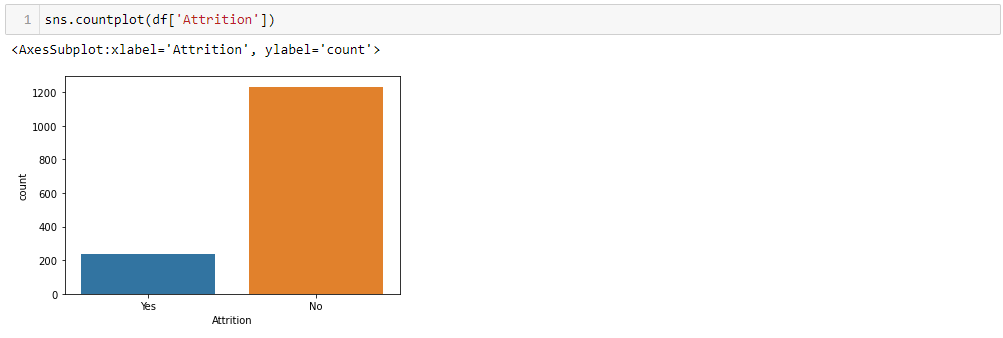
* From the above observation we can see there are no null values in the dataset and in ‘.describe’ we can see the mean and std and from min value to max everything is in a proper way.
* And I dropped some of the columns because they all have same values, and after droping the columns we can see the change in the shape of the dataset, The columns reduced to the 31.

**Data Visualization:**

In this portion we plot different graphs and try to visualize the data

We use different graph include:

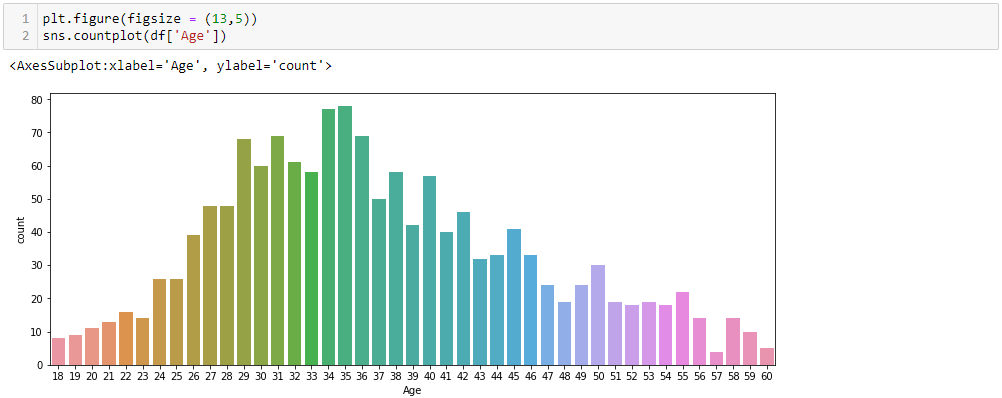
* Bar plot
* Count plot
* Box plot
* Dist plot
* Violin plot
* Cat plot
* Pairplot



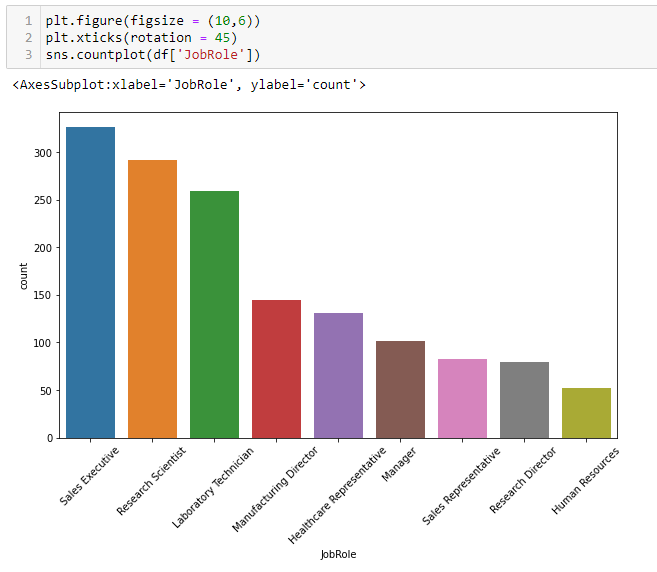
As we can see:

* Attrition: Yes , Around 18 to 20 % of Total Employee count along the table.
* Attrition: No, Around 80 % of Total Employee count along the table.

Observing the situation we can say that there is class imbalancing problem in the dataset.



From the above observation we can say that there are more people between the 30 to 40.



From the above observation we can say that most of the employees are in Sales Executive, Research Scientist and Lab Technician field. And less employees in HR job role.

**Now we will check outliers and skewness using boxplot and distplot.\**



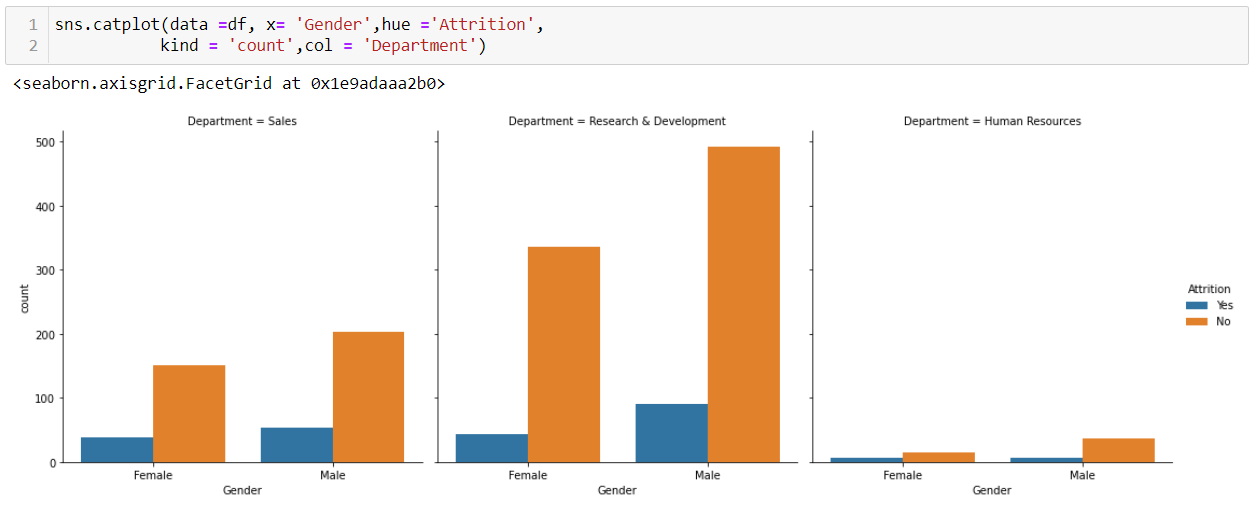
* From the above observation we can see that the ‘Age’ and ‘DailyRate’ both columns are clear, Both the columns are normally distributed and don’t have outliers
* But, We can see that below two columns means ‘TotalWorkingYears ’ & ‘YearsAtCompany’ both have outliers and data is positively skewed.
* ‘YearsAtCompany’ column have more outliers and data also more skewed that the other column. The people above 20 years at company are in outliers.



We will removed them in further process.

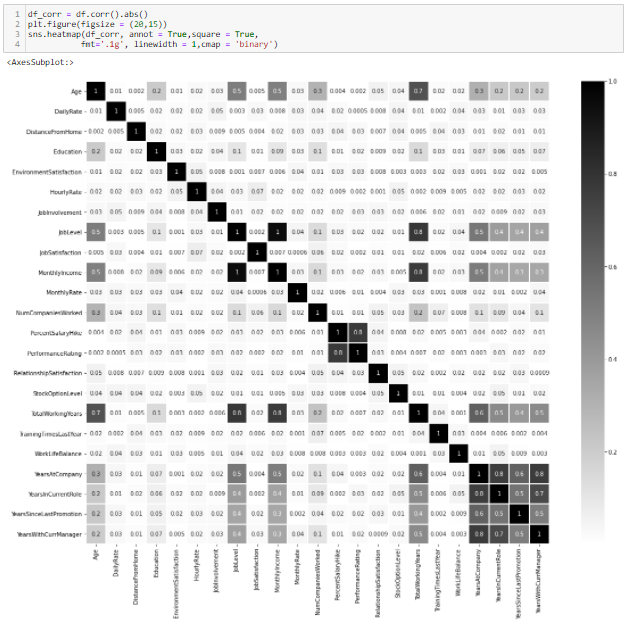


This is bar plot and violin plot and here we can see not much correlation between the Attrition and Gender column and have equal rate of attrition in age also.



* From the above observation we can see that Research& Development department have less attrition.
* And in Human Resources Department, as we saw earlier the HR department have less no. of employees, So the due to that attrition is also less in this department.

**HeatMap:**

****

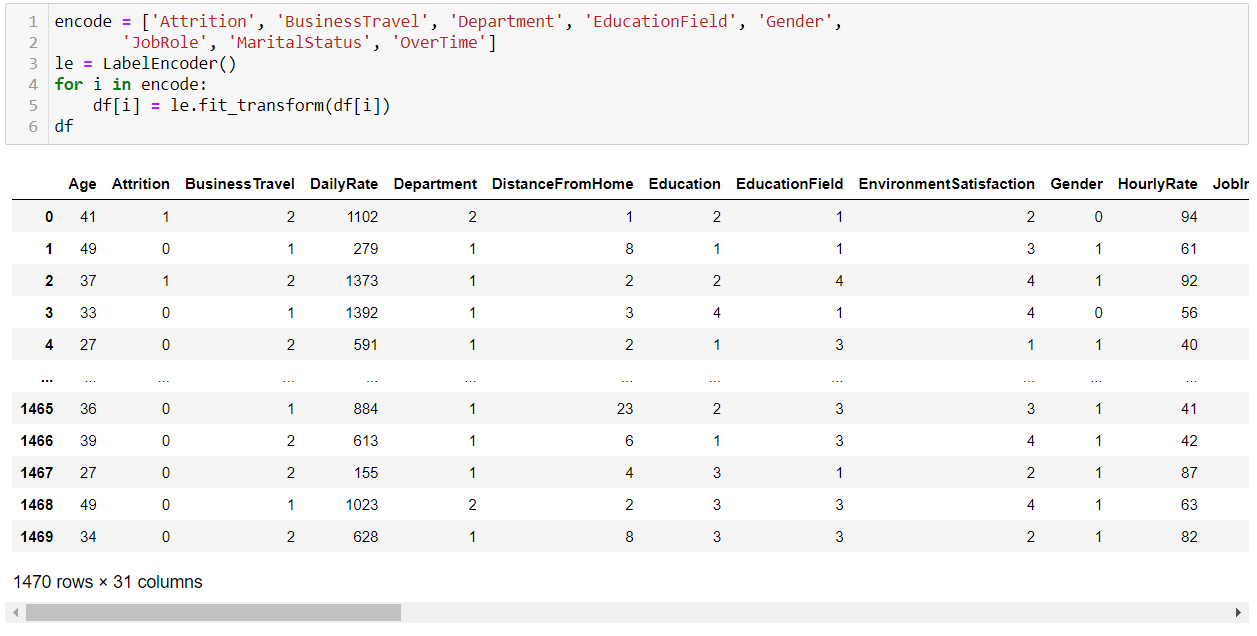
* The heatmap shows us the correlation between the columns and we can see that very less no. of columns has correlation with each other.
* The last 4 columns (YearsatCompany,YearsInCurretnRole,YearswithcurrManager) have major correlation with each other.

**Data Preprocessing:**

First we encode all the categorical columns using the LabelEncoder.

Sklearn provides very efficient tool for encoding the levels of categorical features into numeric values. Label Encoder encode labels with a value between 0 and n\_classes 1

Converting all categorical columns into numeric



As we can see all categorical values converted into numerical values.

Now we will proceed with the encode data.

**Seprating the Features and Target:**

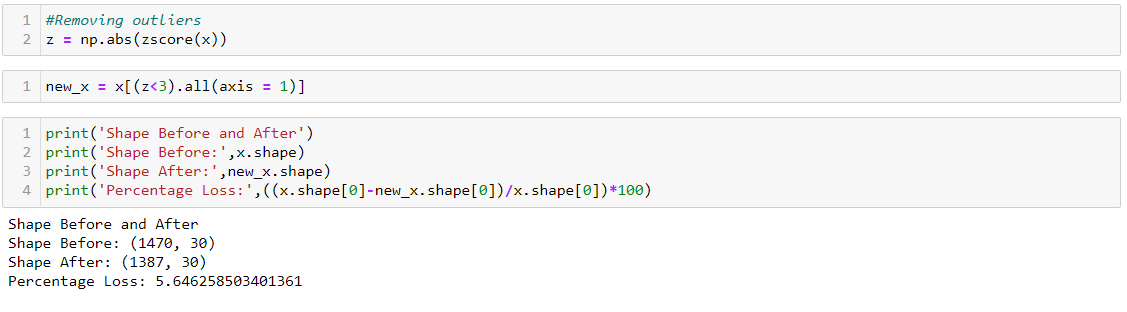
Now we have seprate the features and target column for the preprocessing

ab.png

X have all the features and y have target variable.

**Outliers:**

An outlier is a data point that is distance from all other observations. A data point that lies outside the the overall distribution of the data set.

As we saw earlier some of the columns have outliers and we have to remove them 

From the above image we can see that after removing outlier by using zscore method in result we loss 5.64 % of the data.

The rows reduced from 1470 to 1387

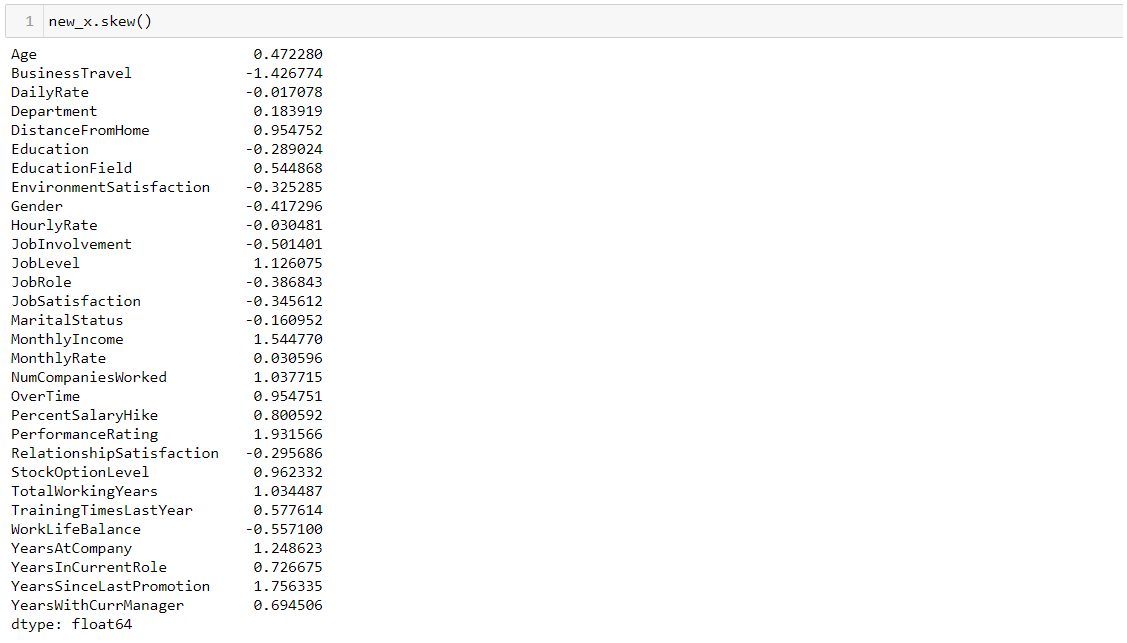
And that much data loss is acceptable, So we will proceed further with this new data.

**Skewness:**

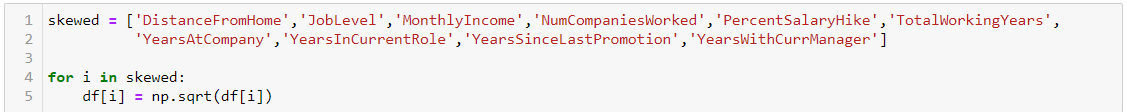
Skewness refers to a distortion or asymmetry that deviates from the symmetrical bell curve, or normal distribution, in a set of data. If the curve is shifted to the left or to the right, it is said to be skewed.

As we saw earlier some of the columns are positively skewed so we have to remove skewness of the columns

First we have check skewness



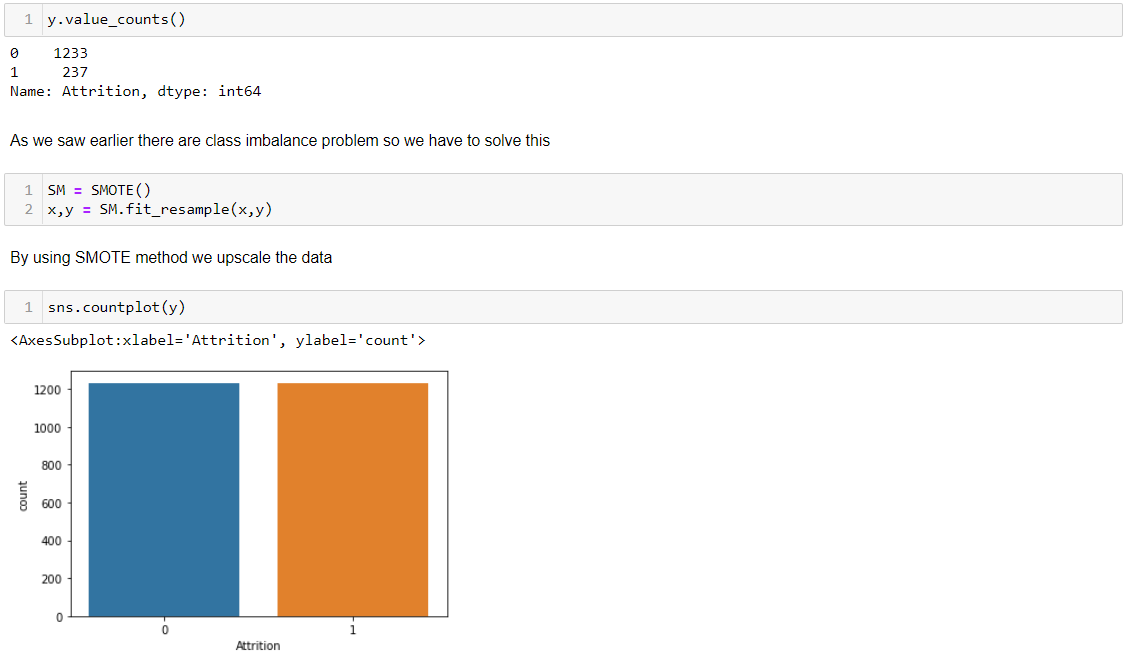
* From the above image we can see the columns with the skewness.
* The skewness value between 0.5 to -0.5 is acceptable. The data above this value can harm model
* We can only remove skewness of the numerical columns, If there is skewness in categorical ones, not to worry about them.
* So we have to remove skewness of numerical columns.



So by the above code we removed the skewness of the data

**Resampling the target variable:**

* As we check above, There were class imbalancing problem so we need to cure that
* If we proceed as it is then it will be problematic for the models and will not give good score.
* We do upscaling of the data using the SMOTE method.



As we can see both 0 (yes) and 1(No) have equal data.

**Standard Scaler:**

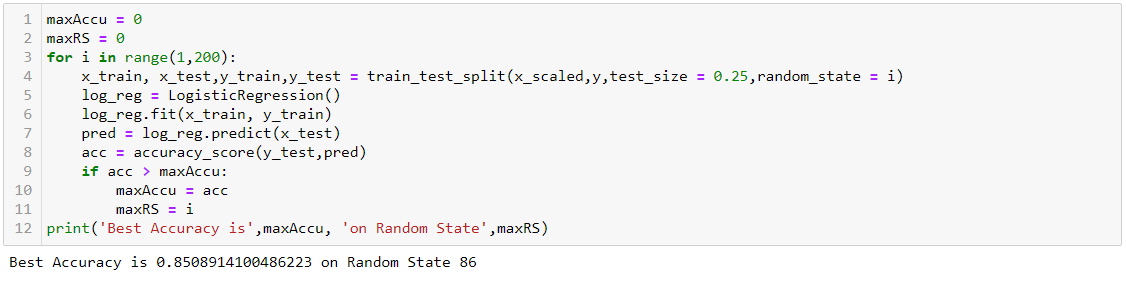
The idea behind StandardScaler is that it will transform your data such that its distribution will have a mean value 0 and standard deviation of 1. In case of multivariate data, this is done feature-wise (in other words independently for each column of the data).

We use standard scaler for scaling the data

**bc.png**

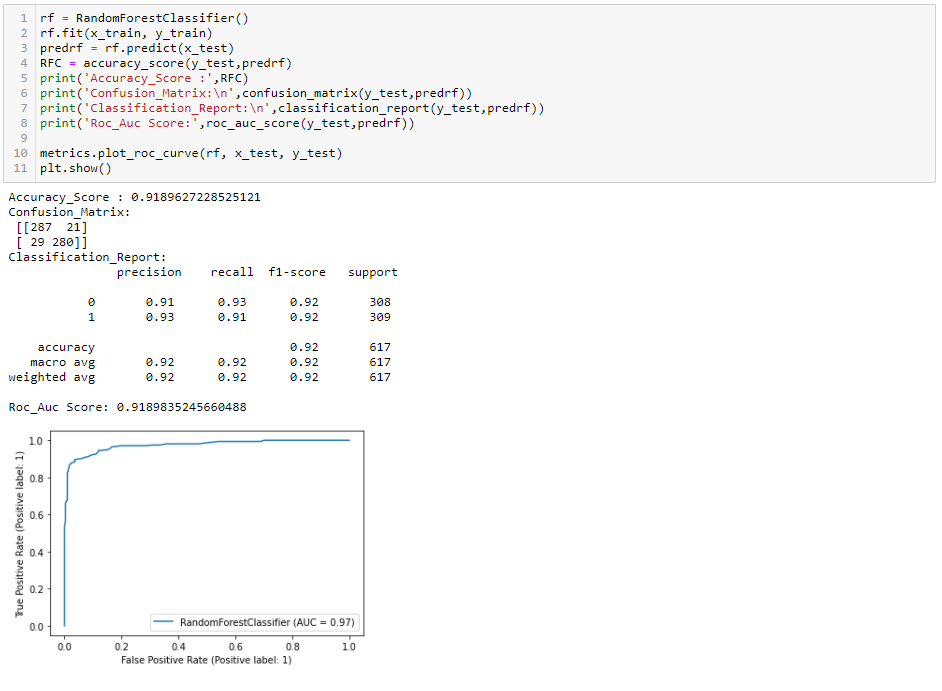
**Prediction:**

Above I am using the for loop which help me to find accuracy score at each random state and for the best state where acc score is maximum is come as output value.

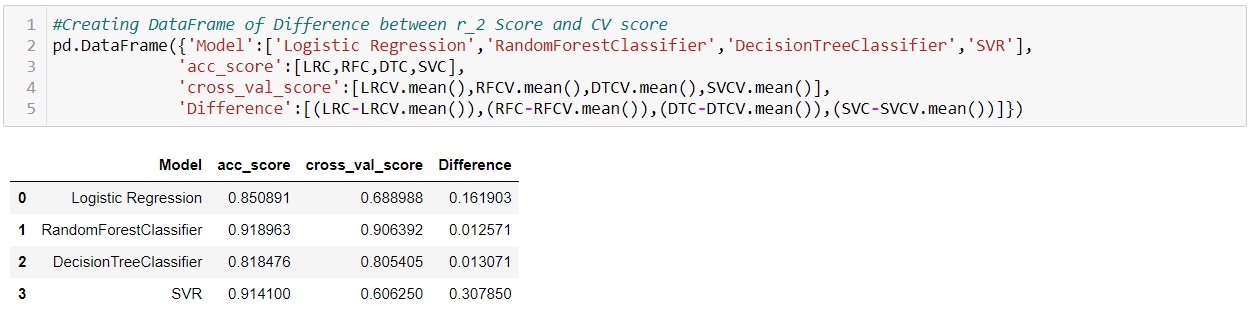
****

So we get 85.08 accuracy score at 86 Random State

Now we use this random state to train the model

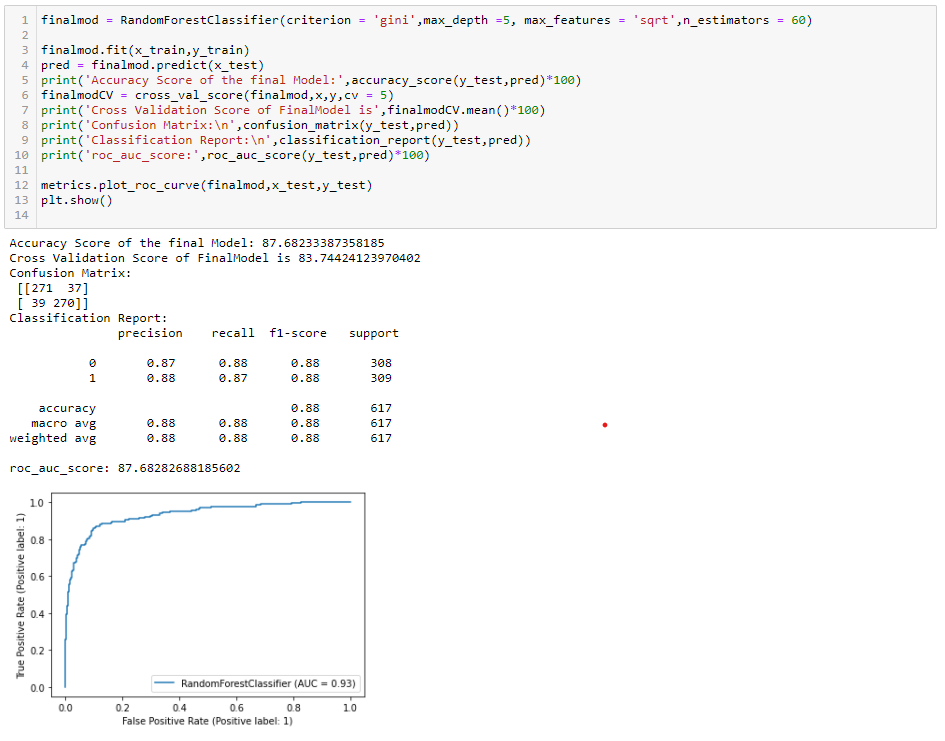


Using RandomForestClassifier we get 85.08 Accuracy score.



* As RandomForest we used 3 more algorithms for model building anf from that we get best acuuracy score from random forest.
* We plot the roc\_curve also
* But we can’t take it as a final score, We try to get best parameters for RandomForestClassifier by using GridSearchCV





* We get best params using the GridSearchCV
* As we can see by putting the best params to the model in result we get 87.68 Accuracy Score and 83.77 CV score.
* By using GridSearchCV the score is reduced, We get best score by the random parameters and it is not mandatory to get best score using the best parameters.
* We will save the random parameters model for the further work.



By using joblib we saved the model.

**Conclusion:**

We have successfully build a model and learned how to analyze employee attrition usingpython. Only with a couple of codes and a proper data set, a company can easily understand which areas needed to look after to make the workplace more comfortable for their employees and restore their human resource power for a longer period.