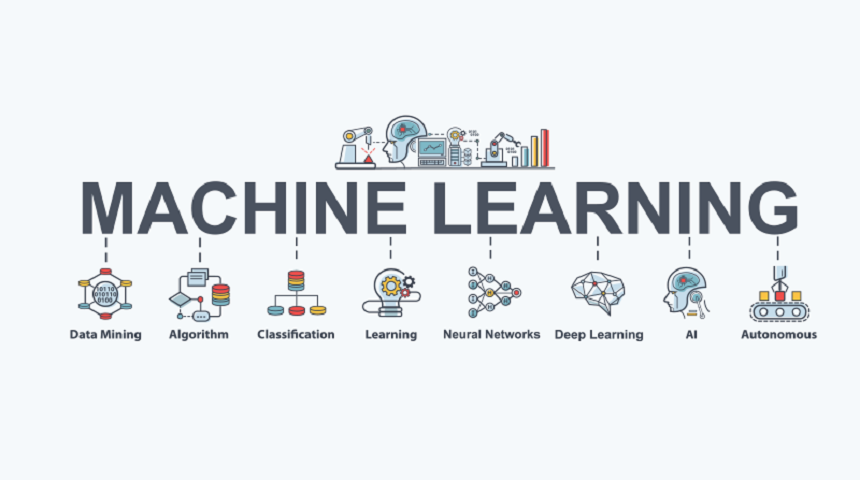
**HR-Employee-Attrition**

In this Python machine learning project, using the Python libraries scikit-learn, numpy, pandas, Seaborn, Matplotlib we will build a model to predict the approval of loan application. We’ll load the data, get the features and perform feature extractions. We explore few models and calculate the accuracy of the model and fine tuning for prediction.



1. **Problem Definition:**

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.

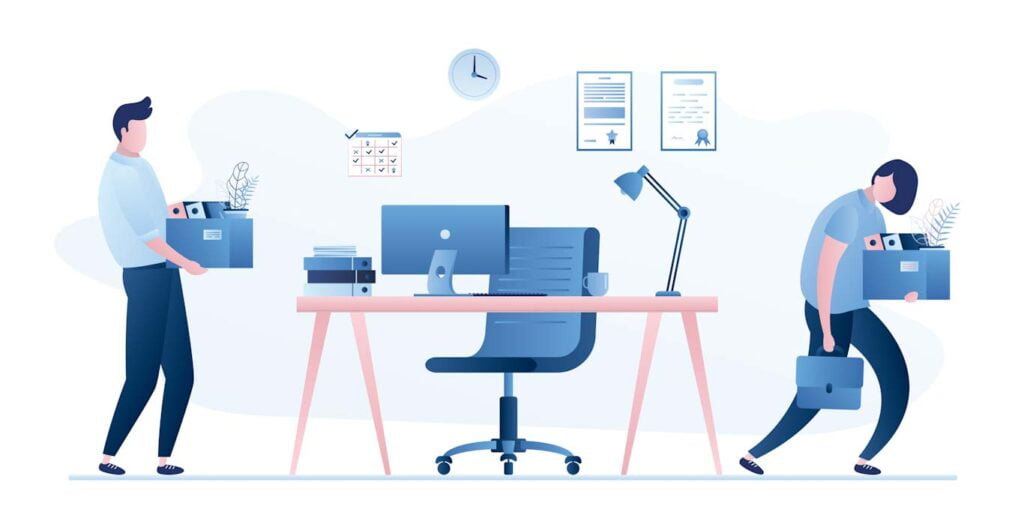
The aim of these programs is to increase the effectiveness of their employees.

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.

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.

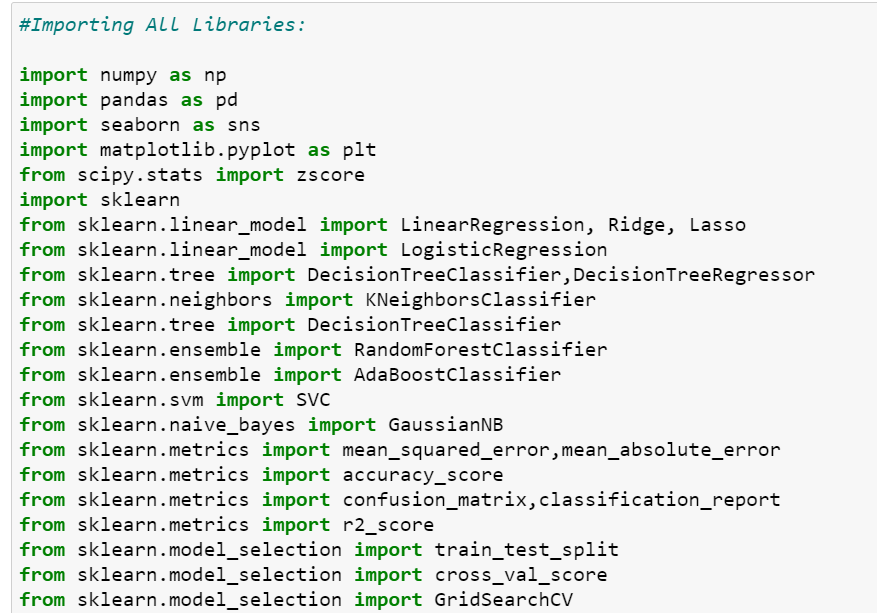
Errors and issues are more likely if you constantly have new workers.

**To Predict: Attrition**

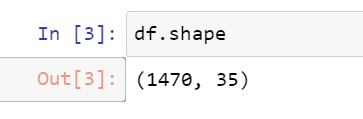
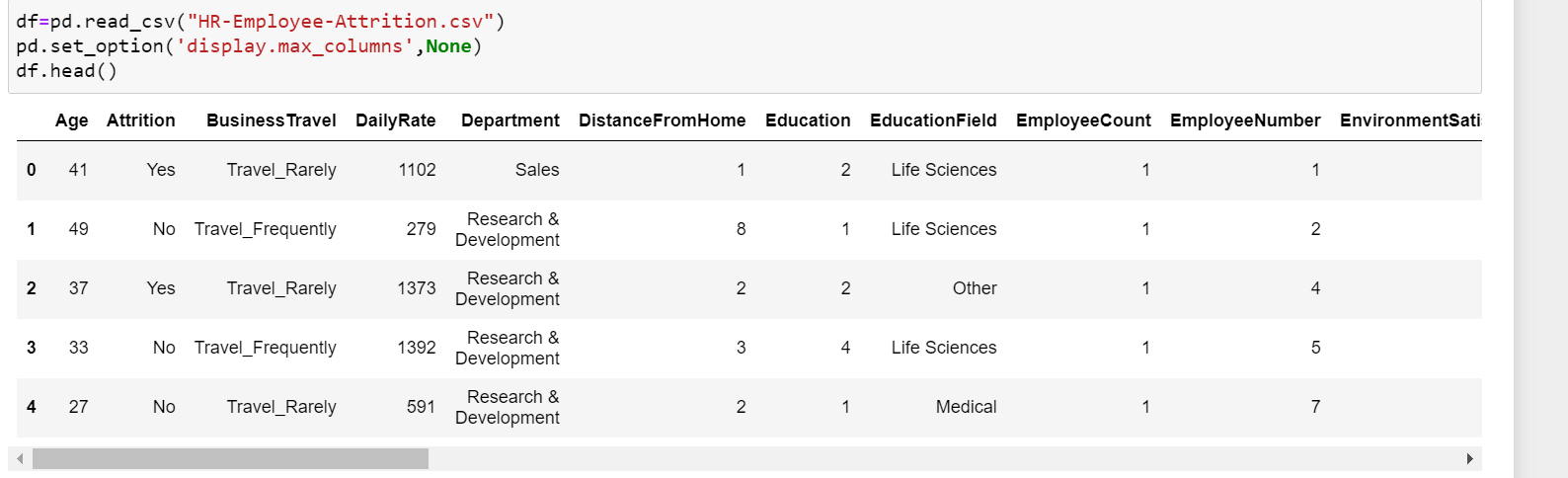
1. **Data Analysis:**

Data analysis helps us to know about the given dataset and what to predict. So before getting in to detail lets know about data columns before getting in to the actual problem. Now let us understand the data columns first so that we will get an idea about the dataset.

Initially we will imported the libraries necessary for further processes of EDA and model bulding.



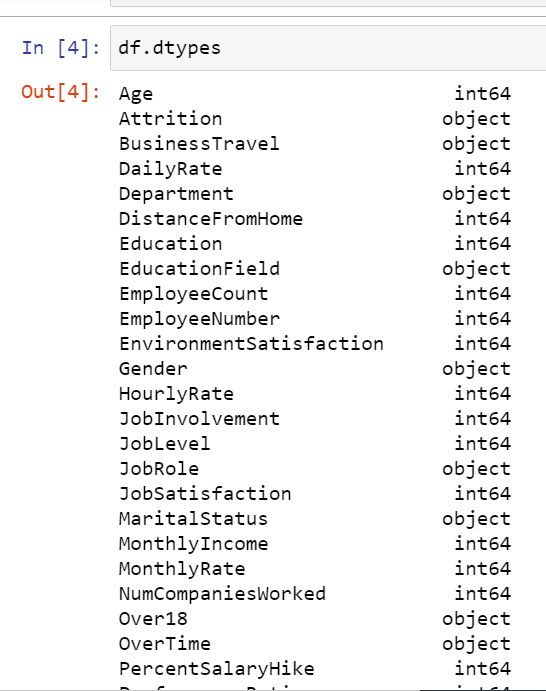
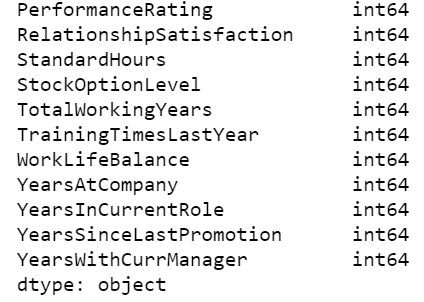
We are importing the dataset from csv format and storing it in Data Frame as below:

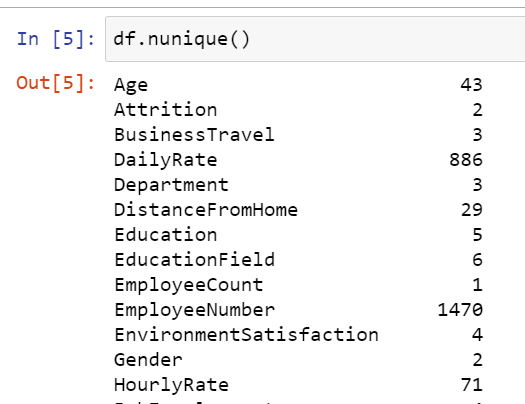
****

* There are 1470 rows and 35 columns. Hence it is small dataset.
* Also we can notice that we have Independent columns and dependent columns so it is Supervised learning.
* Target Variable is “Attrition” which has to be predicted.
* As we need to classify whether the “Attrition” is Yes or No, the above problem is a classification type problem.

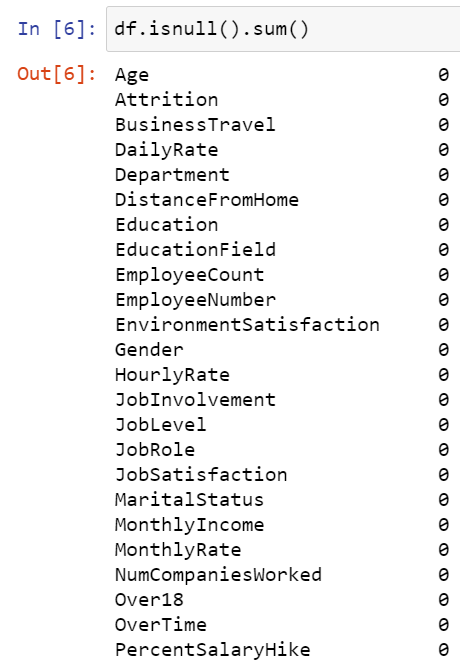
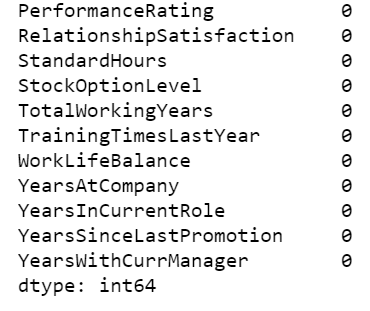
Now let us look in to the each variable and can make some understandings:

Get the data types of each column:

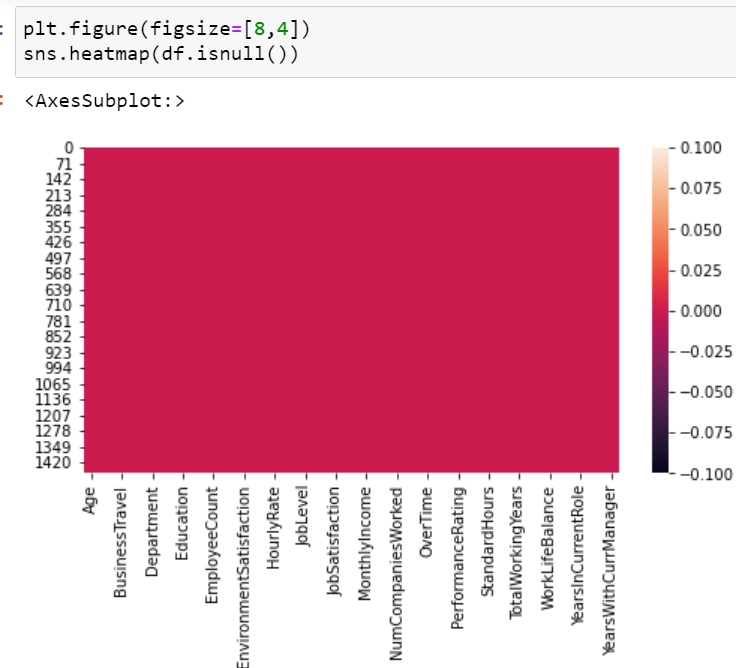
 



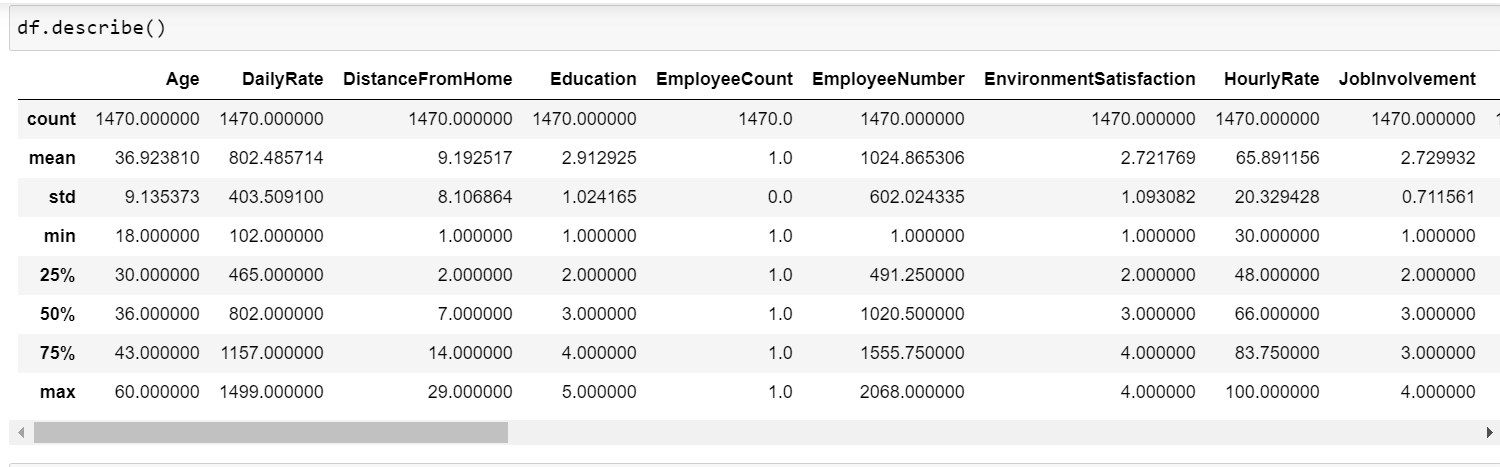
Since there are some categorical and continuous columns we need to check for nunique() method. 'Attrition' is Target variable. Other columns are Independent variables. Almost all columns have different value counts. Only 'EmployeeCount','StandardHours','Over18' have only one value. May be not required for model building.

From the above isnull() function we can see that there are no null values in dataset.



Above Heat map describes about the null values not present in it as color is full single plain color.

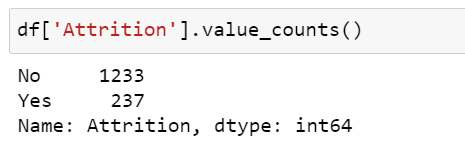


This describe method almost says maximum details of dataset.

Count shows all the rows are not null as , same counts.

Mean and Median are not having that much of big difference.

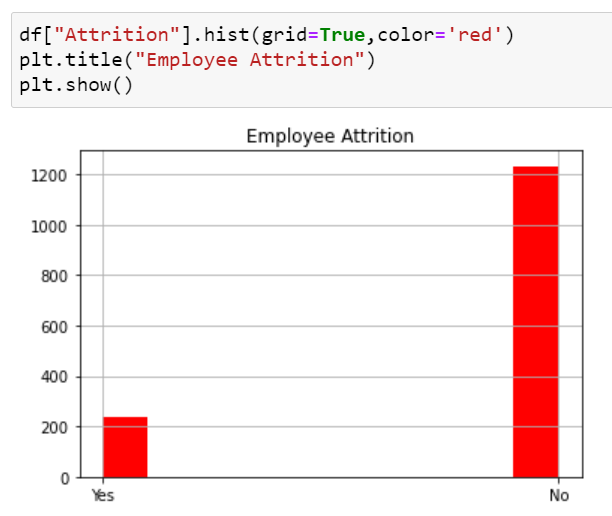
Min and Max have big difference for some of the columns.



Target variable has class imbalance as class “No” has 1233 records and class”Yes” has 237 records.

**EDA:**

Lets analyze using **Univariate Analysis:**



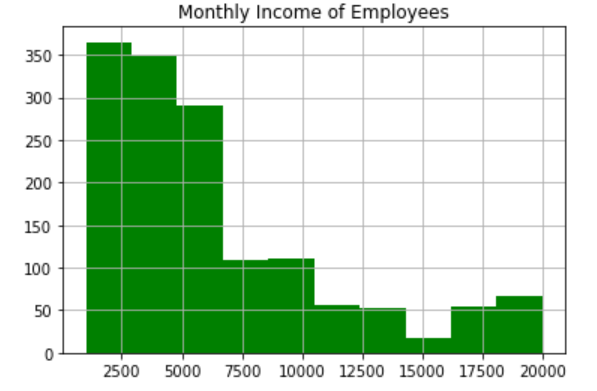
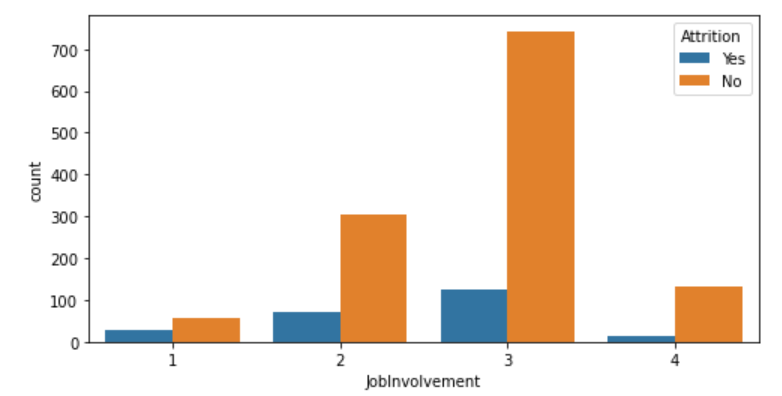
The Univariate analysis explains that there is an imbalance in the Target variable.

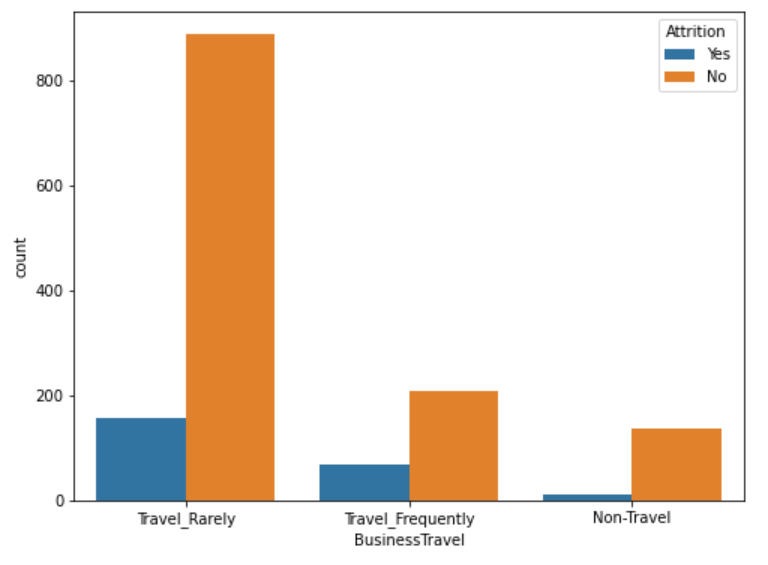
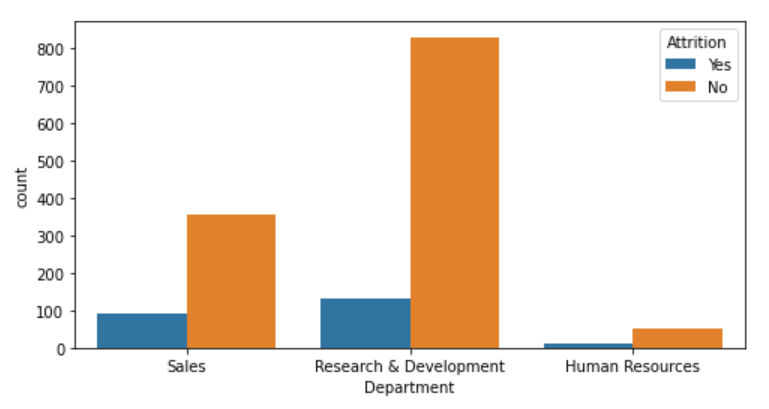
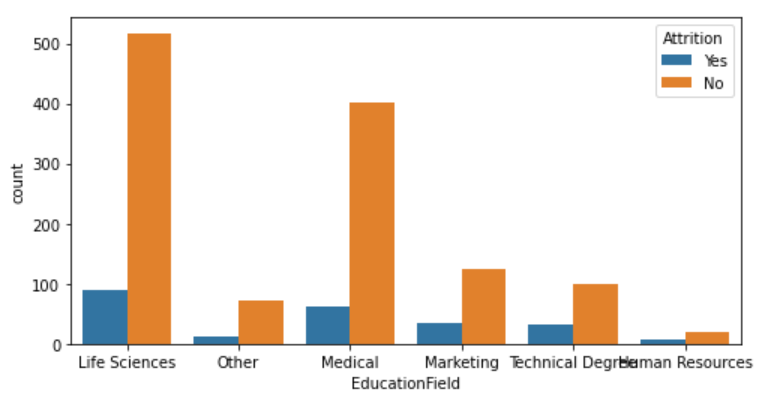
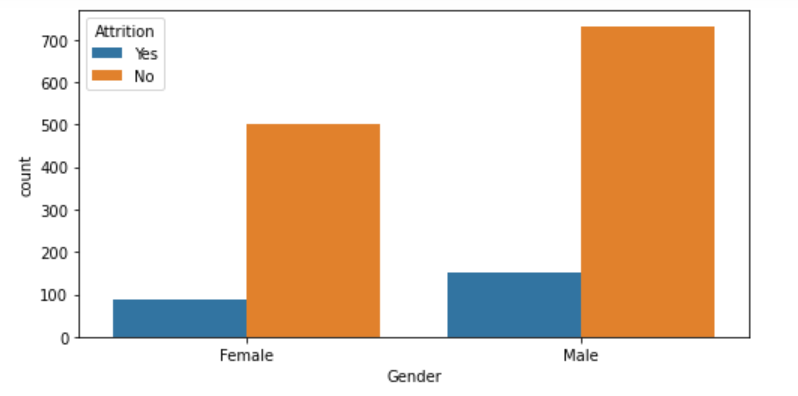
'No' type class value is very big in count as compared to 'Yes'.

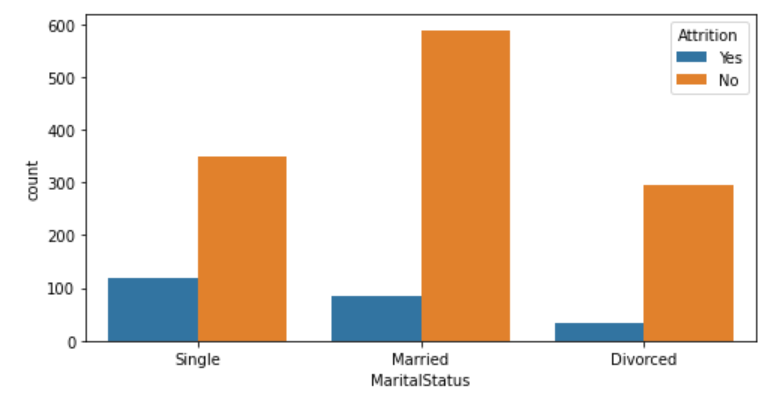
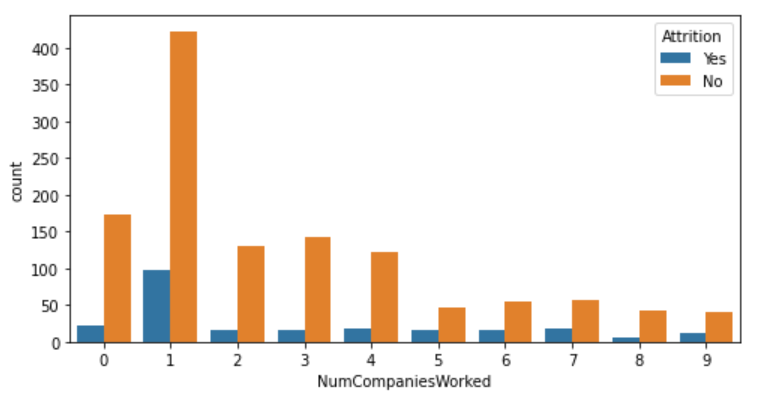
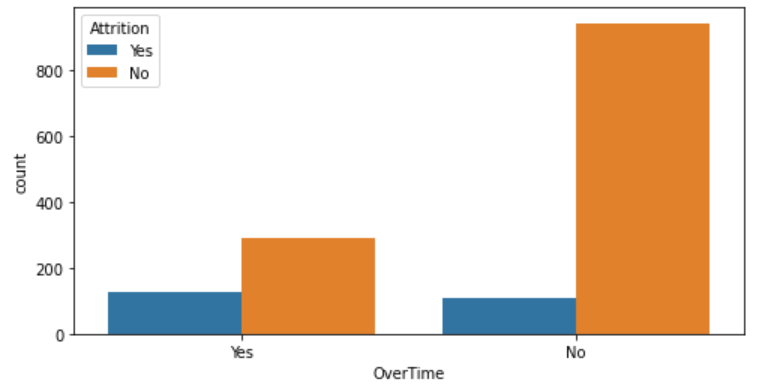
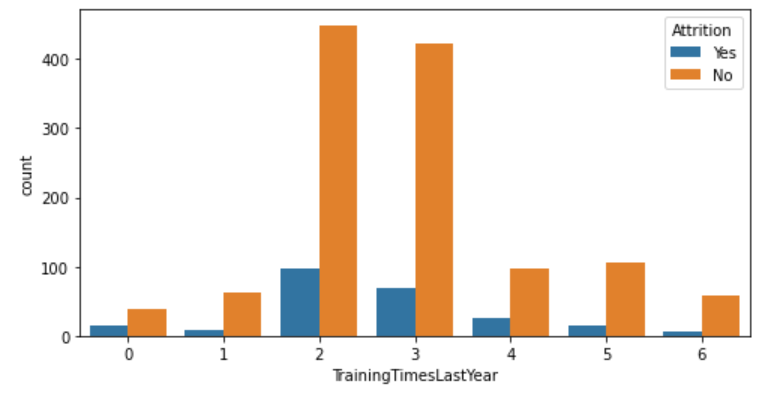
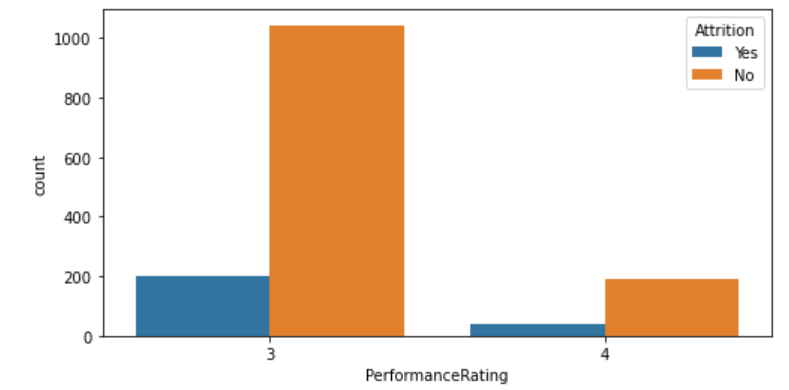
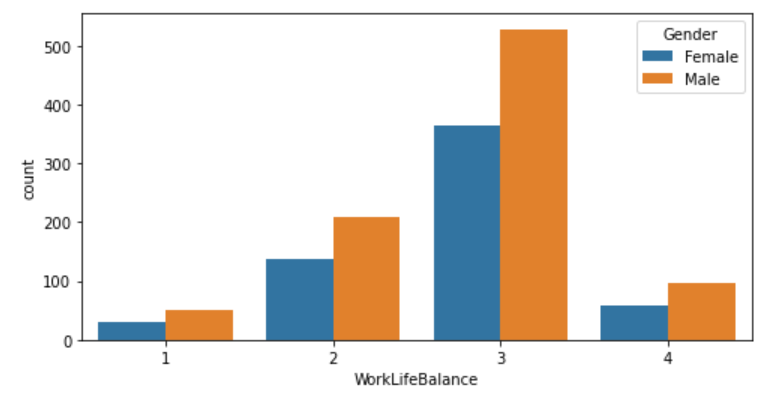
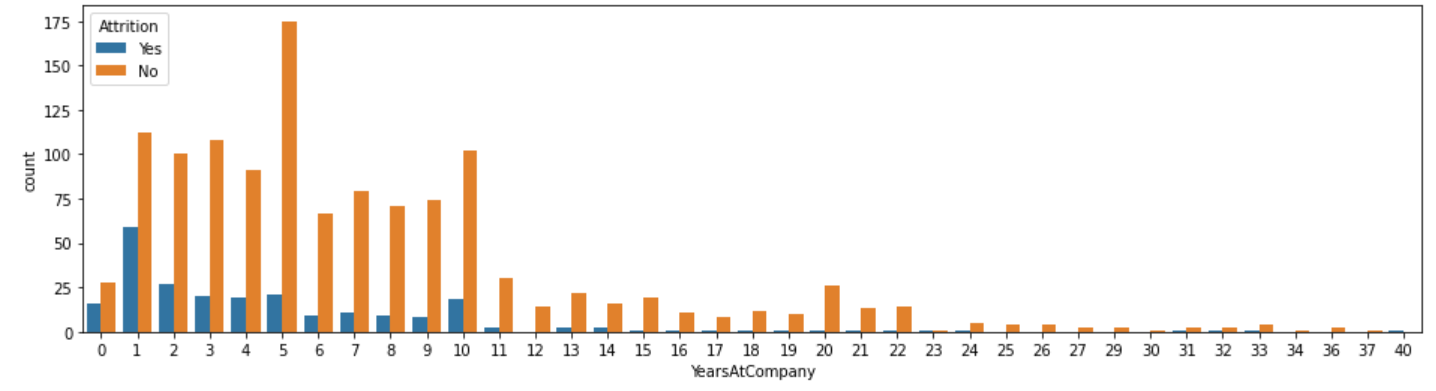
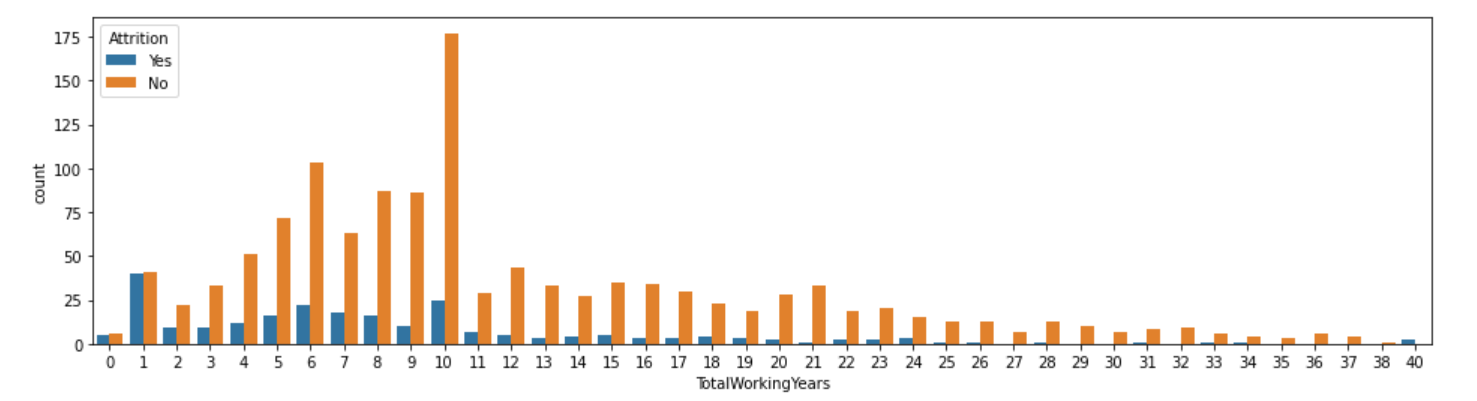
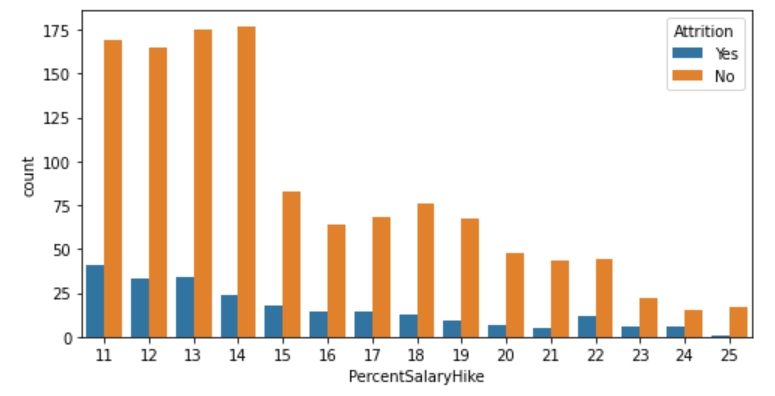
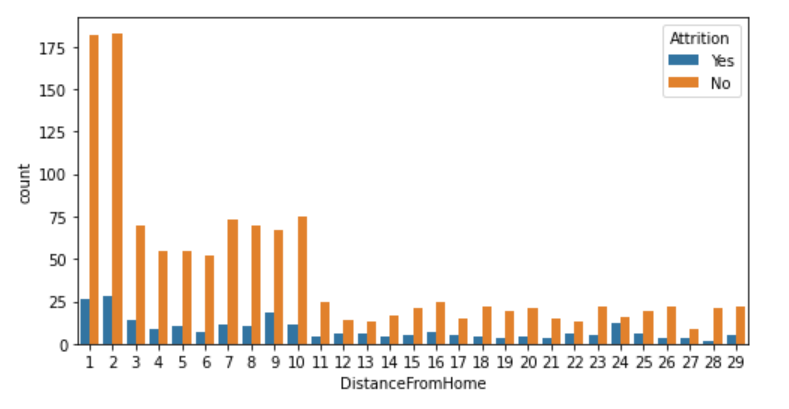
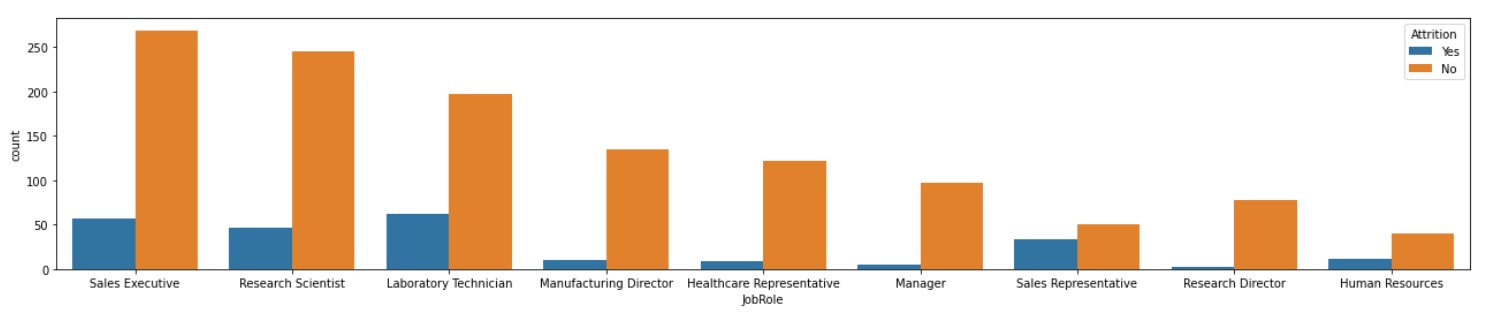
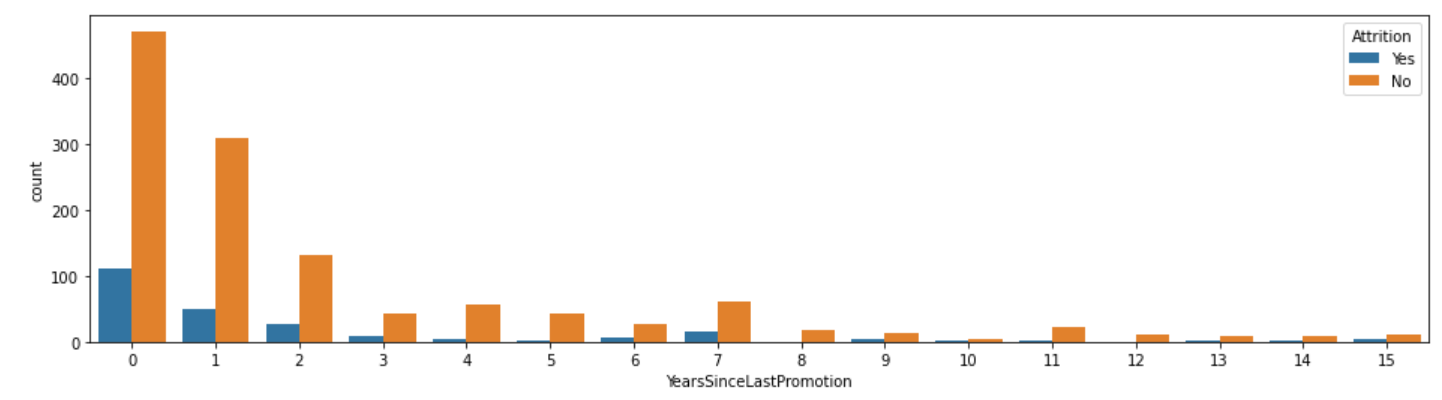
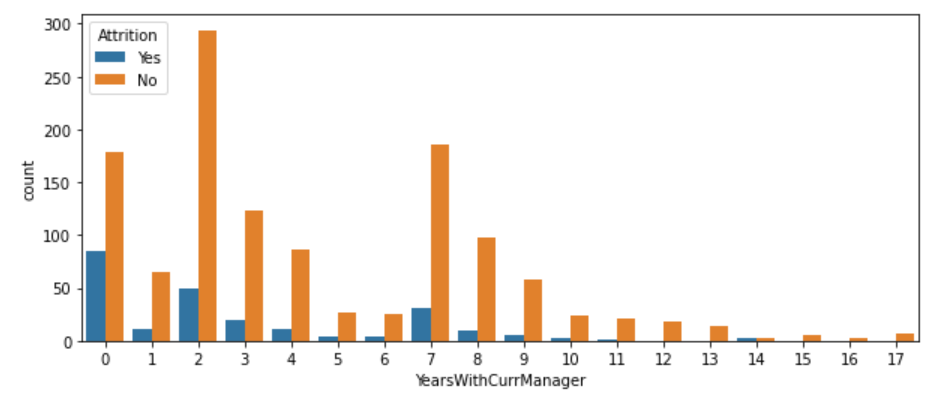
No 1233

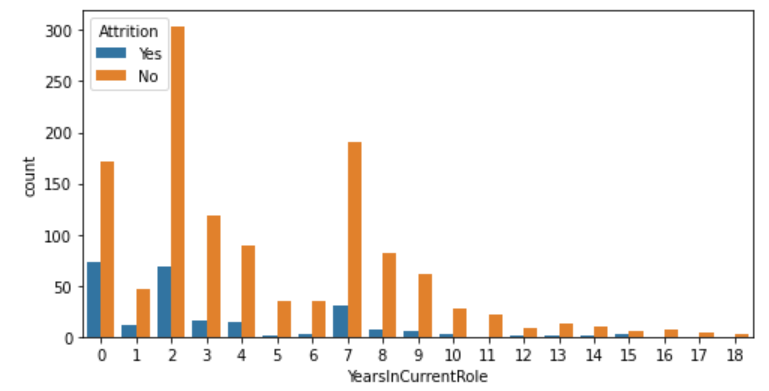
Yes 237

So we need to do Sampling. Either Oversampling or Under sampling.

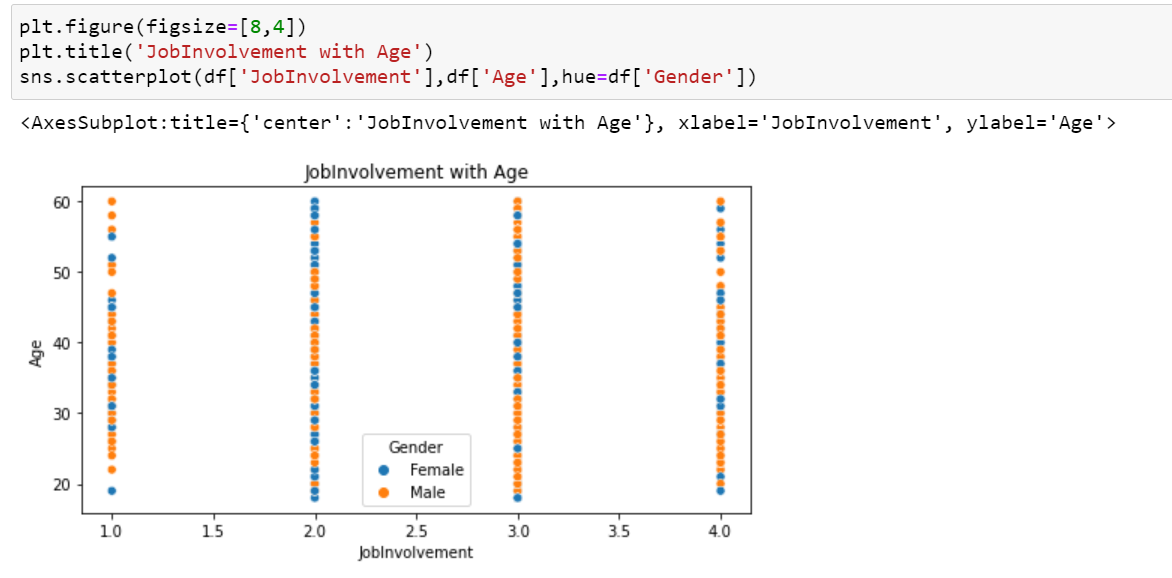
   

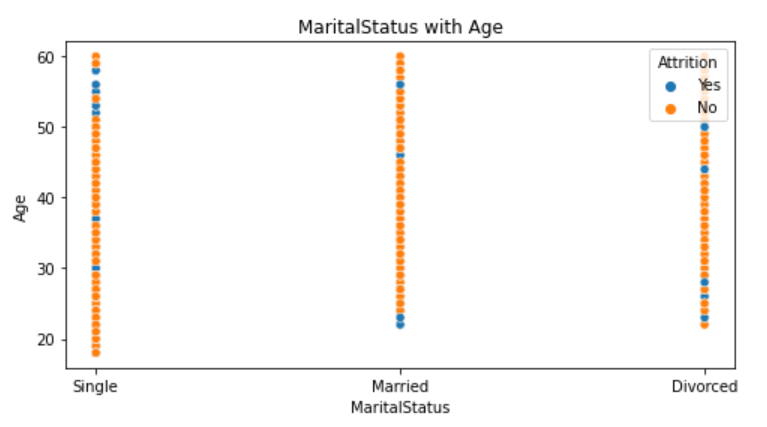


**EDA Remarks:**

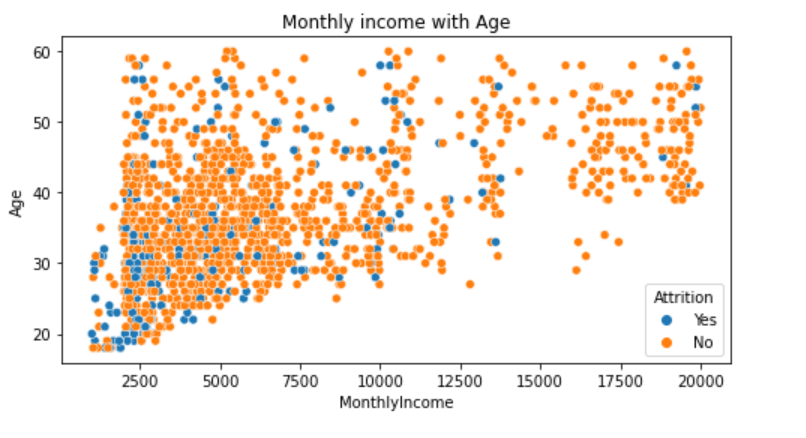
* Employees who travel rarely is big in count as compare with Frequent Business travel.
* Also Attrition is high on the employees who travel rarely compared to that of others.
* In Research and Development department employee count is high with high count in Attrition when compared with Sales and Human Resource department.
* Short distance people are higher in count. And higher counts in ‘yes’ attrition with the short distance from Home compared with higher distance.
* More number of employees are from 'Life Sciences' field. Next is from Technical Degree. Attrition is in high counts in Life Science dept as compared to other dept.
* Number of Male employees are higher than Female and the male count for Attrition is slightly higher than females.
* Job Involvement we can find its moderate as more counts in 2nd and 3rd level. Number of counts with 2nd and 3rd level seems to have high without Attrition but when compared with level 1 and 4 , it has higher attrition.
* The count is higher without attrition in 'Sales Executive' and 'Research scientist' as compared with other roles of job. But the attrition is higher with Laboratory technician and Sales executive. Research Director and Manager are the two roles who has the less attrition comparatively on other roles of jobs.
* Higher number of counts in Married peoples compared to Singles and Divorced. Attrition count is higher in Singles.
* Large number of employees in small scale salary between 2500 to 5000$.
* We can see that 9 is the maximum number of companies worked by an employee. And number of employees with Less years worked is high.ie. mostly with 1 year.
* Number of count is high with the people not having overtime.
* Maximum counts of no attrition are with 11,12,13,14 percentage . i.e with minimum salary hike percent.
* Number of counts in performance rating is less with 4rating and higher counts in 3 rating.
* We can see that there is a wide range of working experience with employees (from 0 to 40). More number of counts are there in 10 years of experiences. And Attrition is high with more counts in 1 year experience.
* We can see that male has higher work life balancing as compared with female employees.
* Number of years working at present company is large in the range of 5 years and 1 year.
* Higher counts of people in current role is 2 years and 7 years.
* Promotion are with higher counts of people with less than 2 years.
* There is a big count in people working with current managers for 2 year and 7 years. Very minimum number of years for long term working with same manager.



Job involvement is missing in the ages between 48-60.Also Males have higher job involvemet than Females in point 4.



We can see that singles are highly populated and can see 'Yes' Attrition in higher age factors of Single kind of employees.



its densely populate with people with small scale salary in the age groups of 25-45.

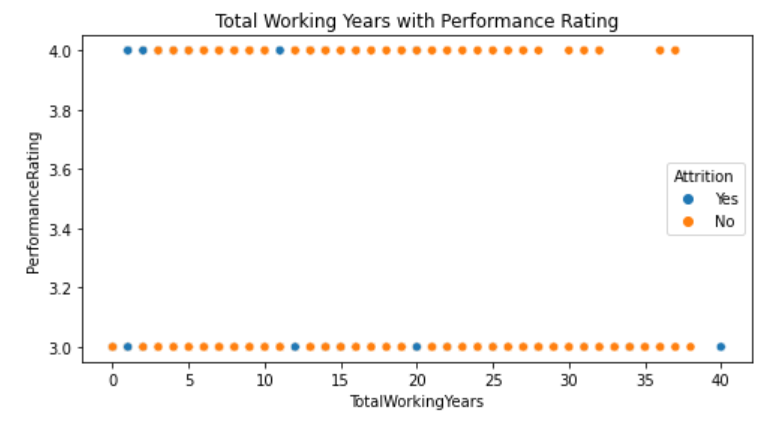
Also higher scale salaries are with higher age employees.

And we could find that there is attrition higher in the salary ranges of 2000 to 3000$ and of ages between 20 to 30.



Number of companies worked is higher with more job involvements.

Higher the job involvement, lower the attrition.

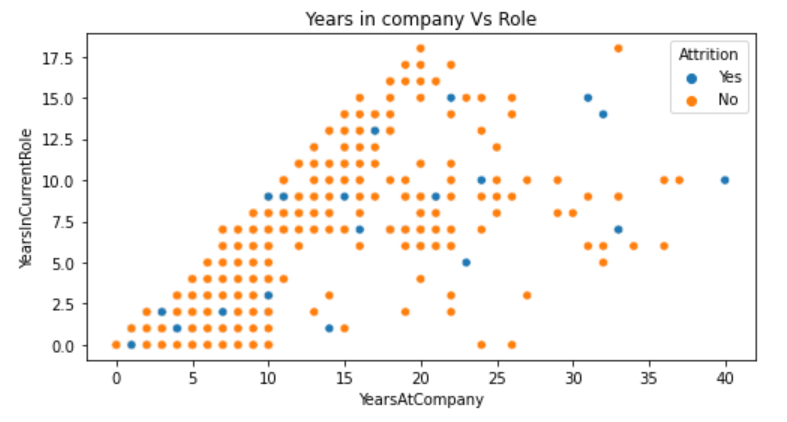


We can find that there is lacking Performance rating with higher number of working years between 33 to 40.



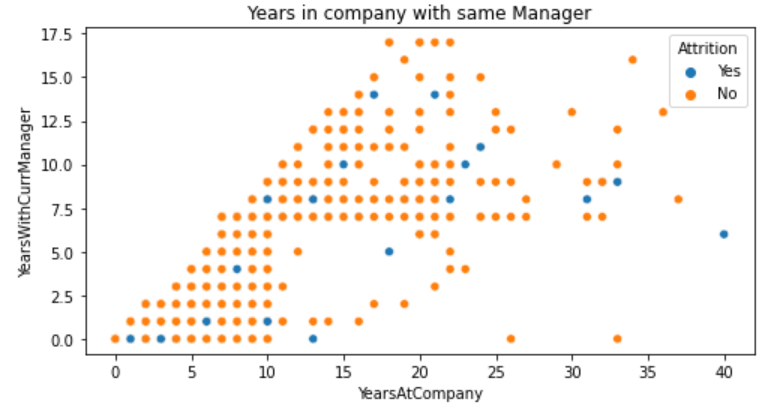
Males seem to be more with many years at company.

And we can notice that Attrition is with females when it come for long years.

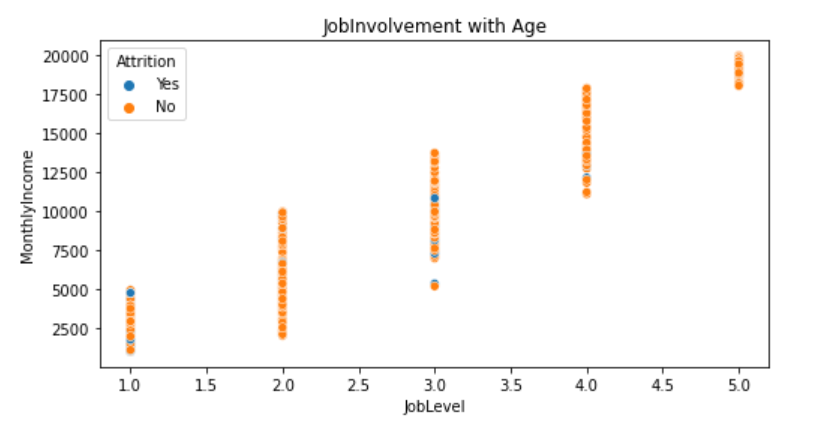


We can find some people with 18 years at same role same company as maximum.

Roles keep changing as there is increase in years at company.

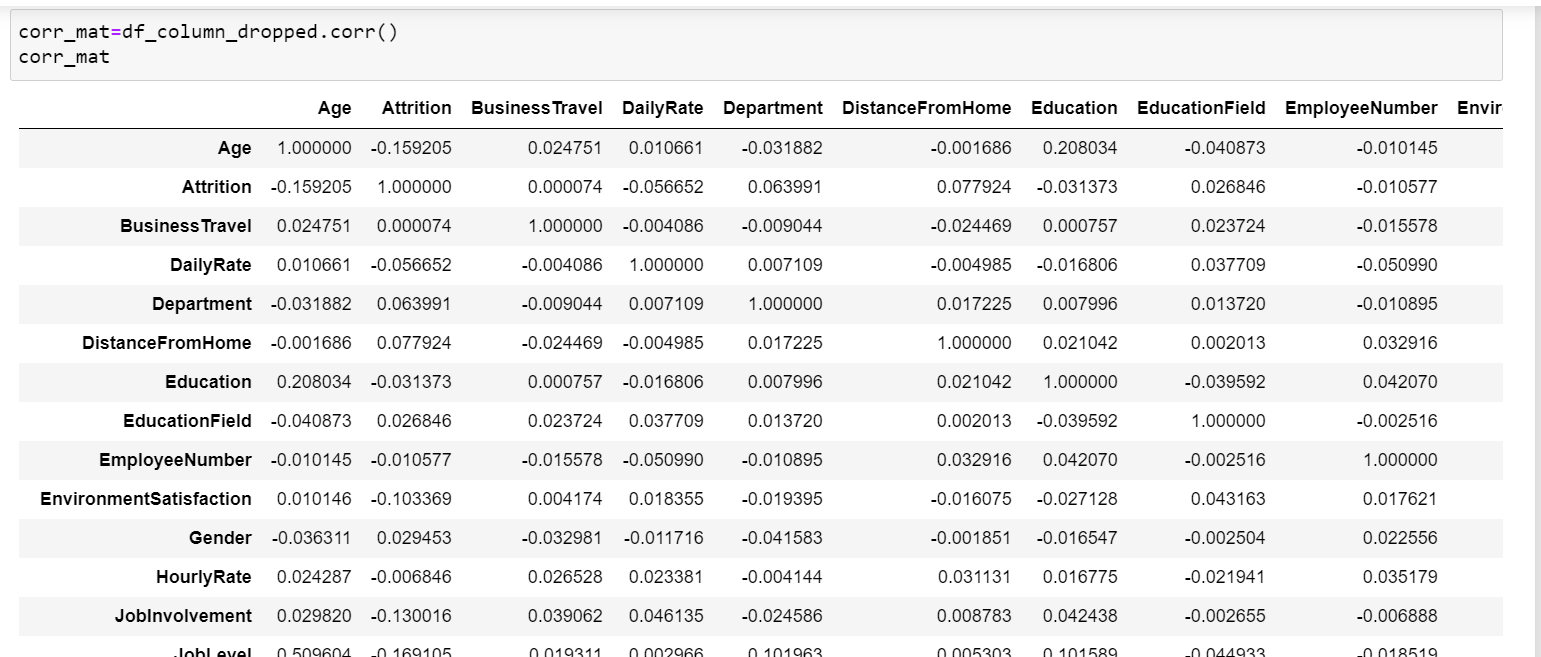


We can find the slope with increase in years at company there is increase of years with manager.



As the job level increases the Monthly income also increases.

We can find the attrition in low levels of job.



**3. EDA Concluding Remark:**

We can see that 'Attrition' is negatively correlated with

'YearsWithCurrManager', 'YeraInCurrentRole', 'YearsAtCompany', 'TotalWorkingYears', 'MnthlyIncome'.And positively correlated with 'OverTime', and 'MaritialStatus'.

Job role and monthly income is correlated with Age factor.

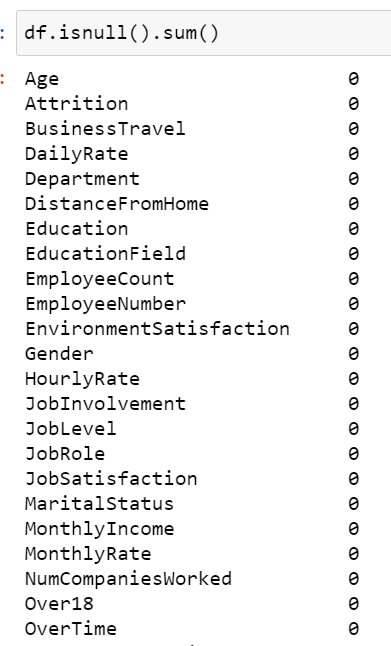
Also Total working years highly correlated with age.

We can find the columns like 'YearsAtCompany','YearsAtCurrRole','YearsSinceLastPromotion','YearsWithCurrManager' are highly inter correlated.

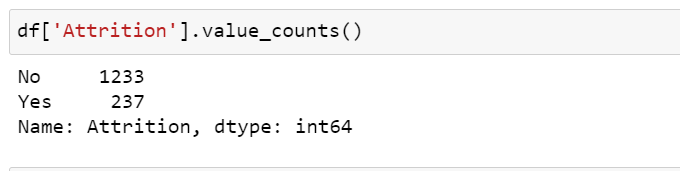
Marital Status and Stock option level is high in negative correlation.

'Precentagesalaryhike' is correlated with 'Performance rating'.

**4. Pre-Processing Pipeline:**

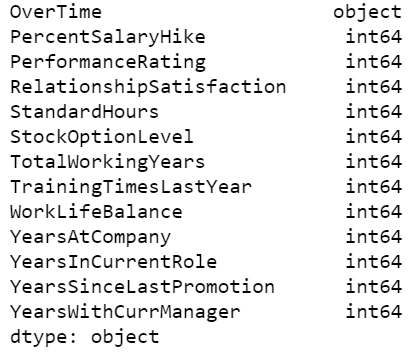
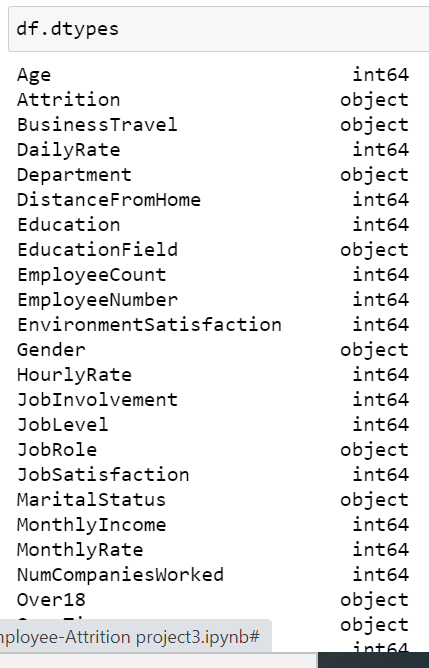


From the above isnull() function we can see that there are no null values in dataset.



Since this 'Attrition' column is Target variable column, we need to find the value count of it. and hence determined that

Yes with 237 and No with 1233.



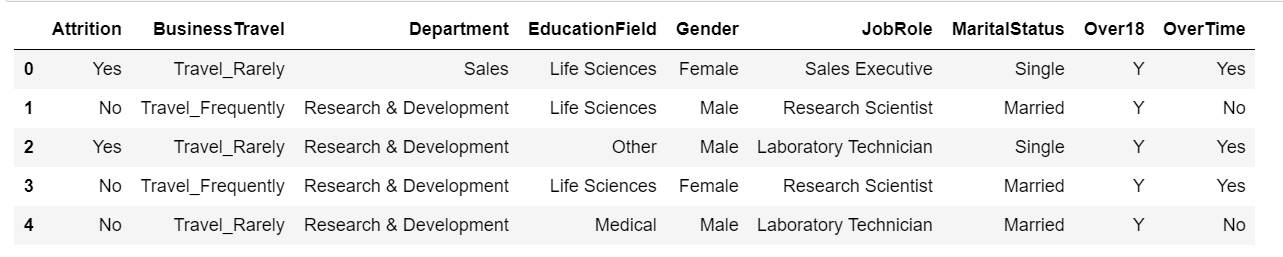
we have some of the Object variables we need to analyze and to transform those for model building.

Label Encoding:

obj\_df = df.select\_dtypes(include=['object']).copy()

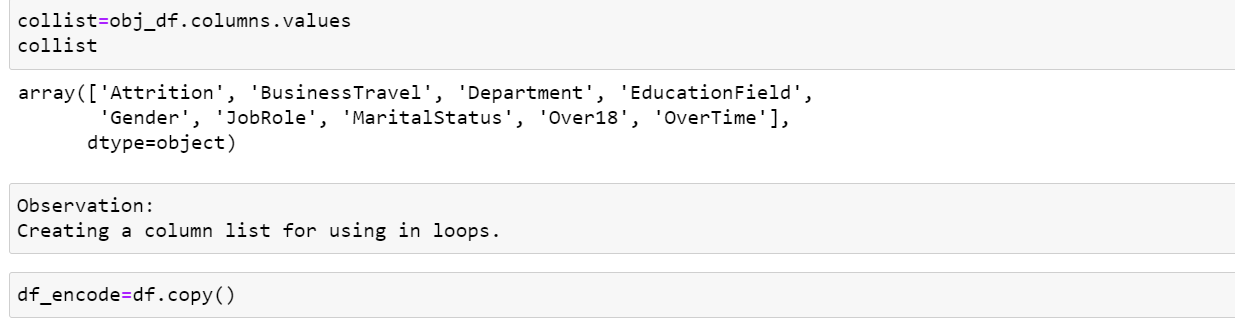
obj\_df.head()

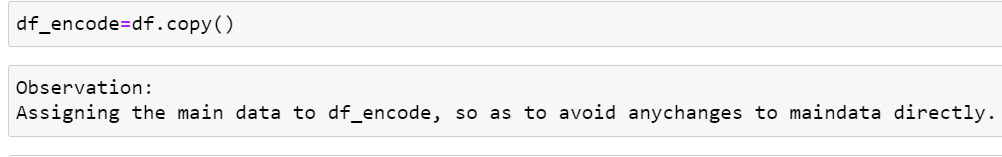
Filter of above code gives us the below data with object string columns alone filtered from actual dataset.



We cannot apply this column directly in model building. So we change it to Numeric kind of data.

Creating a column list for object columns using in loops :

****

****

**Code for transformation of object columns to numeric type for model building:**

code is used for transformation of the string columns by 'LabelEncoder()' method:

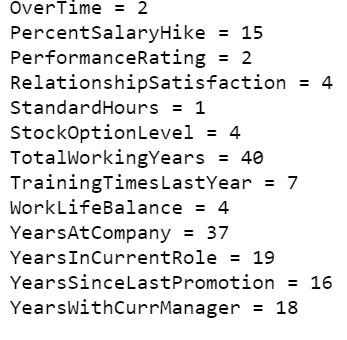
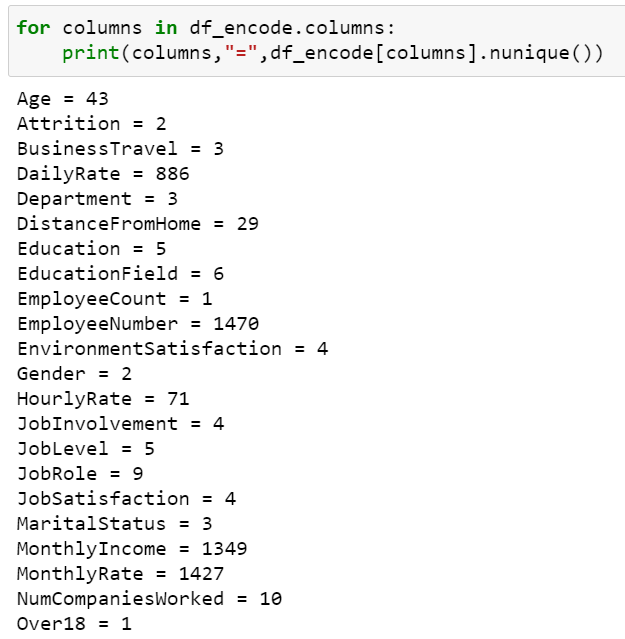
from sklearn.preprocessing import LabelEncoder

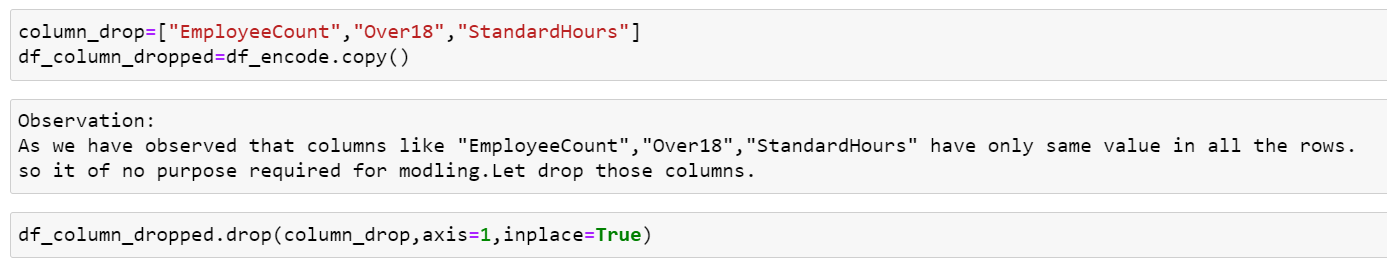
LE=LabelEncoder()

for column in collist:

df\_encode.loc[:,column] = LE.fit\_transform(df\_encode.loc[:,column]).

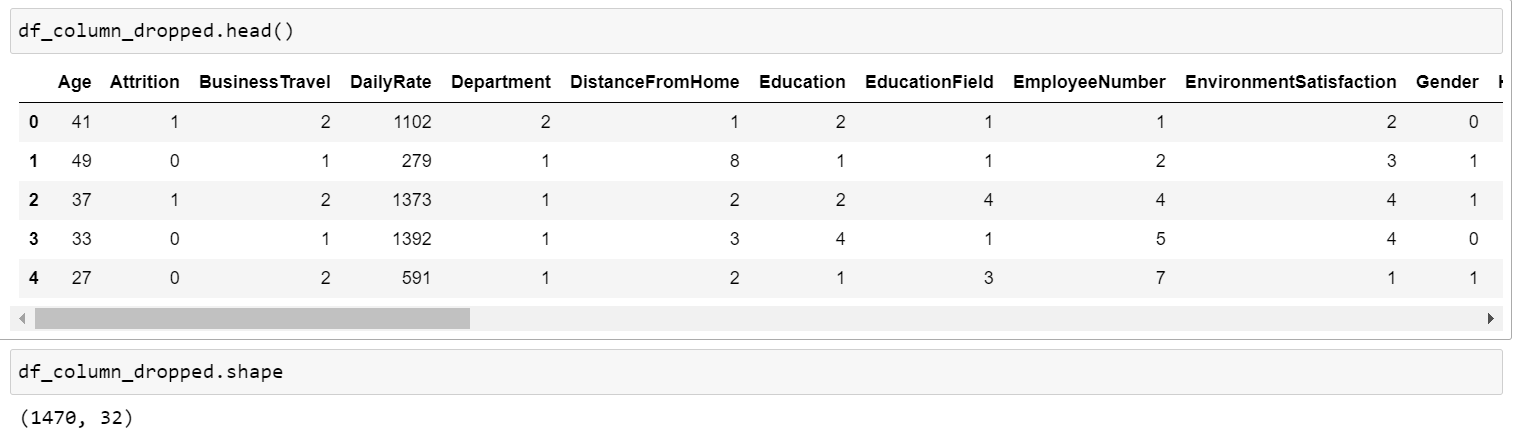
Now lets identify the value counts in an organized manner. so that we can identify and analyze for unwanted columns.





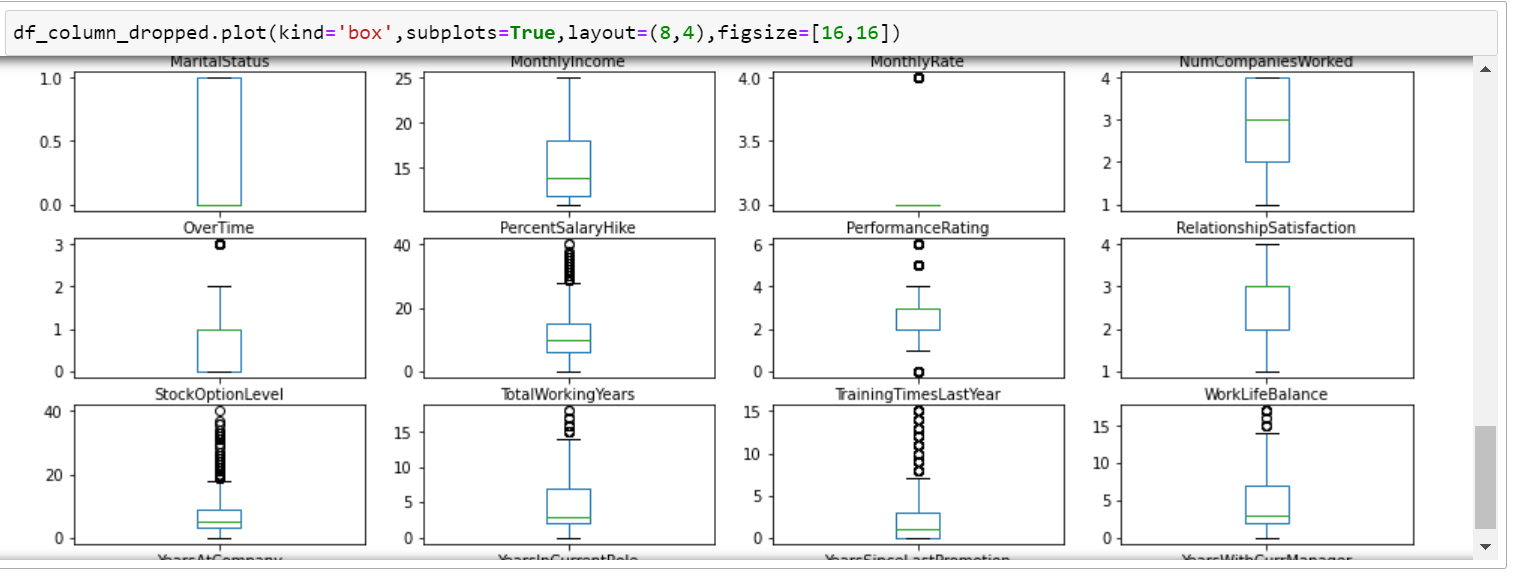
Lets drop "EmployeeCount","Over18","StandardHours" columns.

As these columns have only same value in all the rows. so it of no purpose required for modeling. Let drop those columns.



After dropping we can notice the change in number of columns count as 32.

**Outliers identification:**

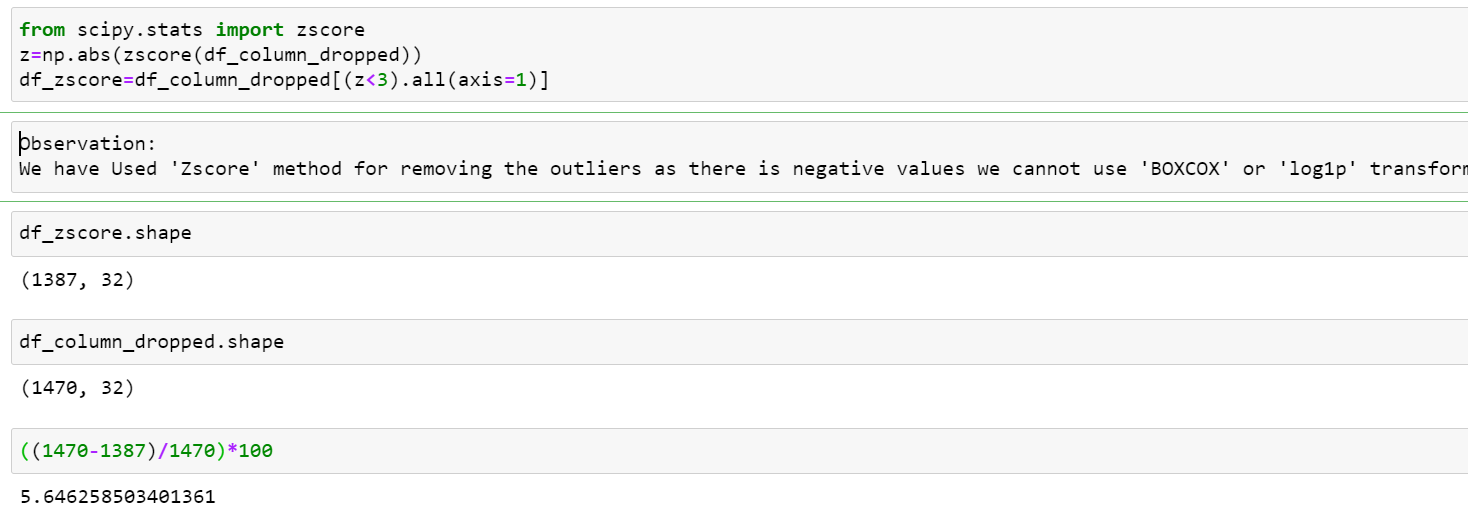


With the Box plot we can identify that there are some outliers beyond whiskers.

outliers in some columns like:

YearAtCompany,YearsSinceLastPromotion,YearsInCurrentRole,TotalWorkingYears.

We try to remove the **outliers** using the zscore method:



We can notice there is change in number of rows. Columns still remains the same.

And the data loss is up to 5%.not an issue ,So let proceed with the new set of data.

We have Used 'Zscore' method for removing the outliers as its effective over all columns at a time .

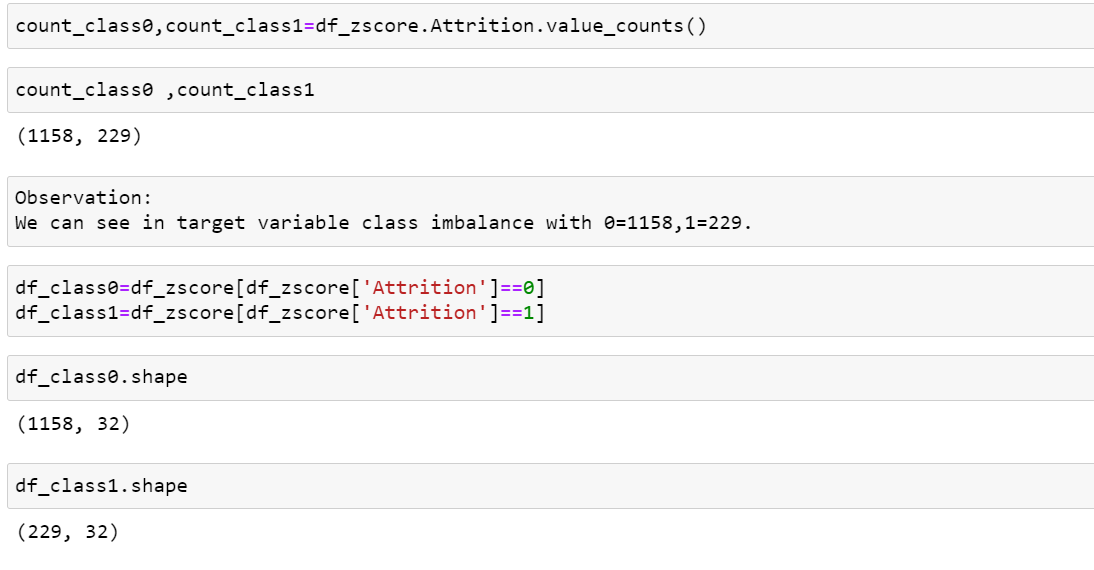
**Imbalanced Data Handling:**

As we mentioned in start we have imbalanced data. So we nee to balance it before applying in model building.

Or else we will not get good accuracy.

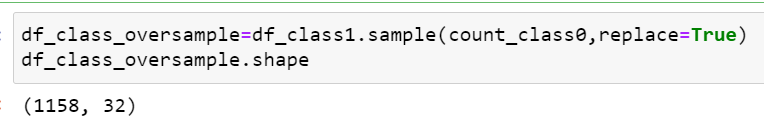
So lets proceed with Over-Sampling as it will not cut down any data or data loss will not be there.

So let do sampling with Over-Sampling for the lower count class in Target variable.



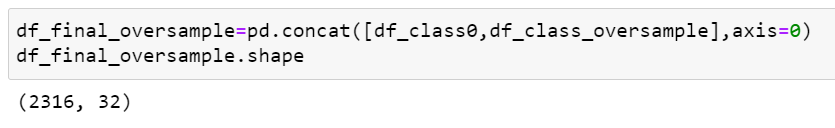
We have split the Target variable value counts into 2 different classes:

count\_class0, count\_class1 and 2 data frames.



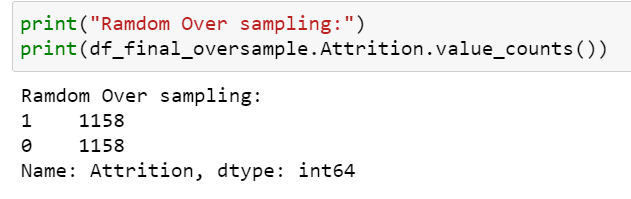
We are going to Add or oversample the 'df\_class1' as it has only 229 rows.

we will add the 'count\_class0' rows as required to equalize both class values.

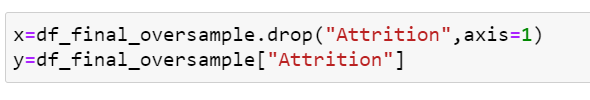


Hence we are concatenating the oversampled('df\_final\_oversample') class + with ('df\_class0') original class.

And hence row count is doubled. making the classes balanced.



**Now let’s do feature selection:**

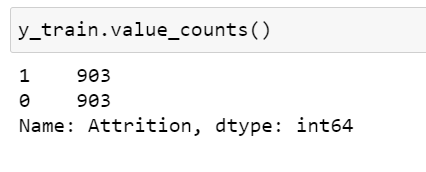


We apply train test split as regular step for processing model building.

x\_train,x\_test,y\_train,y\_test = train\_test\_split(x,y,test\_size=0.22,random\_state=6,stratify=y)

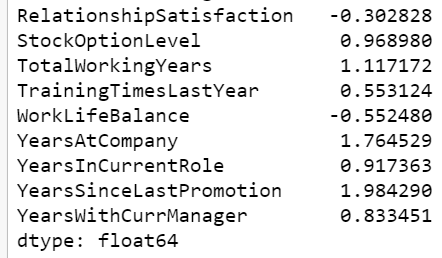
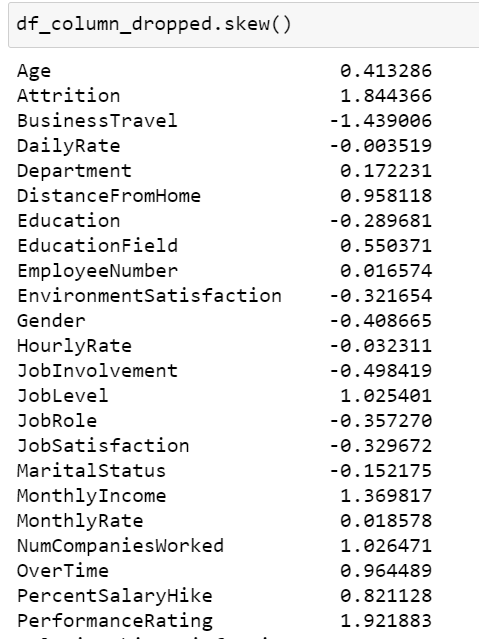
we use stratify for making the train\_test\_split that y is equalized from above sampling.

Below step clarifies that y\_train is perfect split after sampling. so its perfect to proceed further



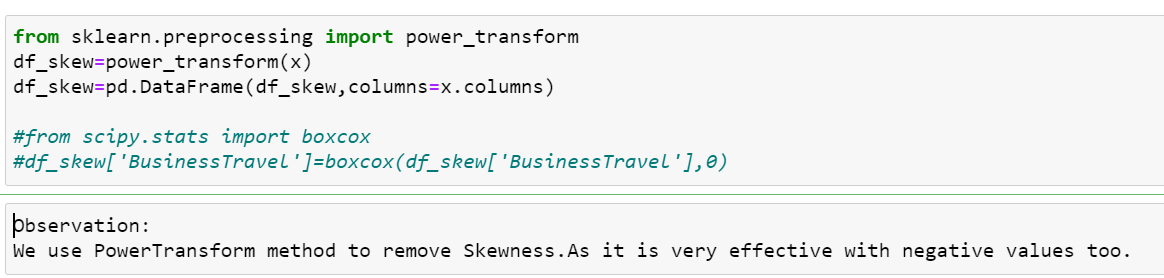
**Skewness checking:**

Before skewness removal the data are with huge skewness. We can notice that there are some columns with above 0.5 to -0.5 values.

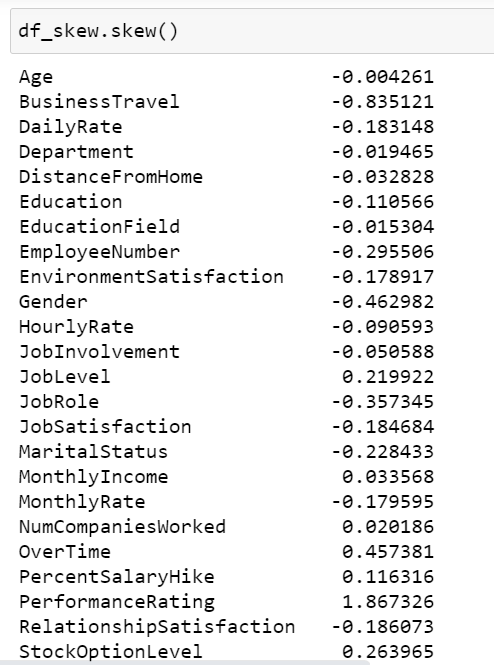
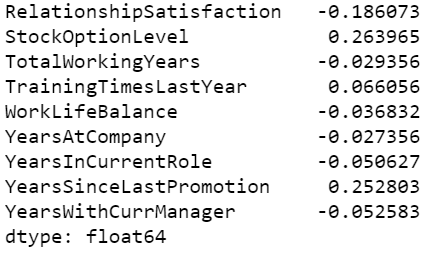


There is negative values we cannot use 'BOXCOX' or 'log1p' transform.

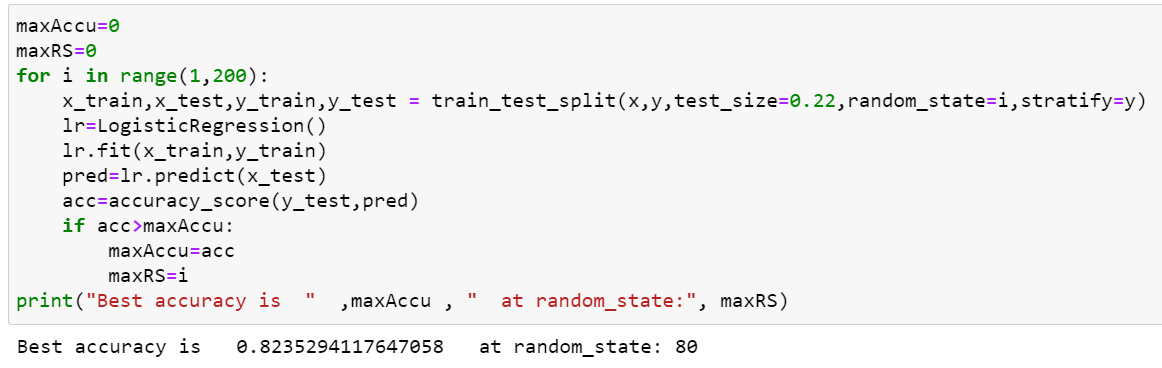
Let’s use power Transform method for removal of skewness:



After removal of skewness using transformation we could notice the difference of reduced skewnesss in most of the columns.

1. **Building Machine Learning Models:**
   1. **Best Random state checking**

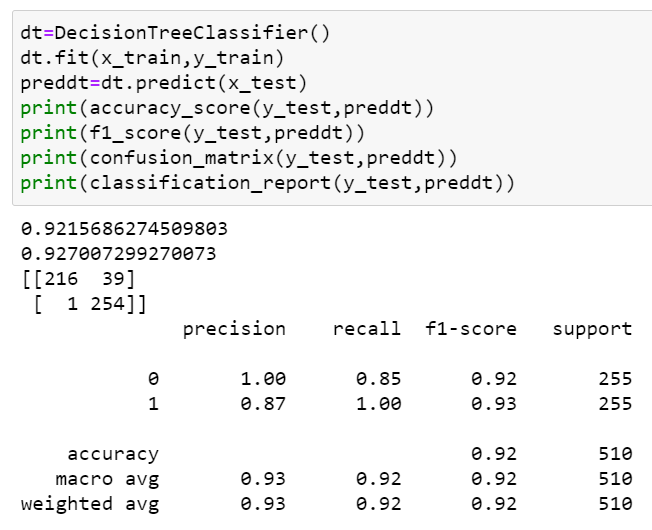
****

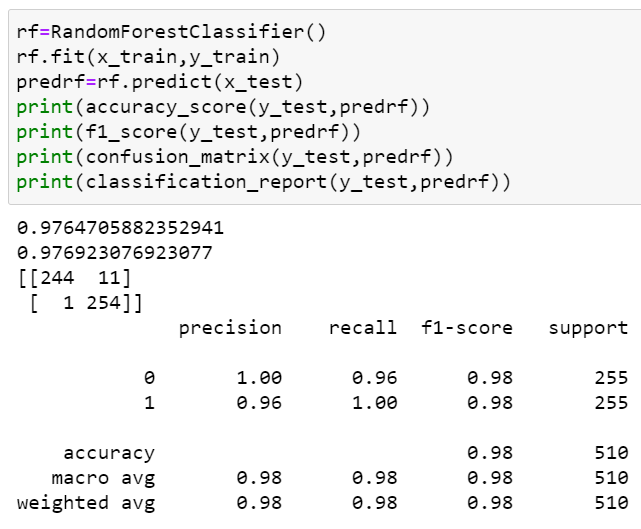
**We have used above code to get the best random state value with best accuracy percentage. So 80 is best with 82%**

* 1. **Applying models:**

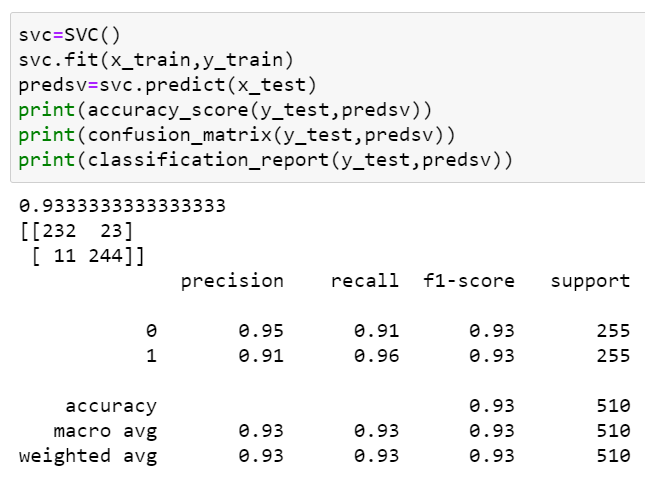
Lets check with 3 to 4 models for best performer

**1.Decision Tree Classifier()**

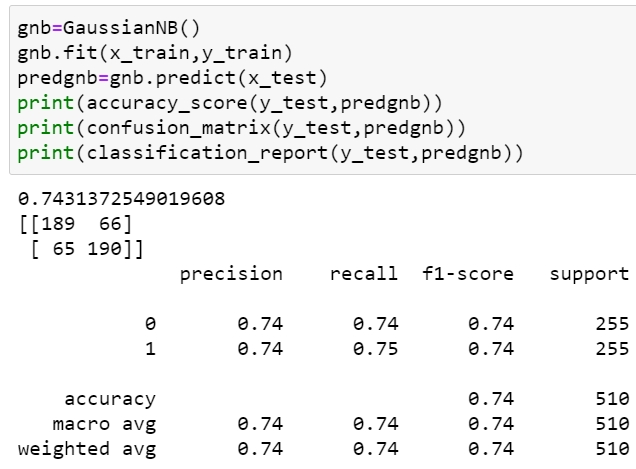
****

**2.Random Forest Classifier()**

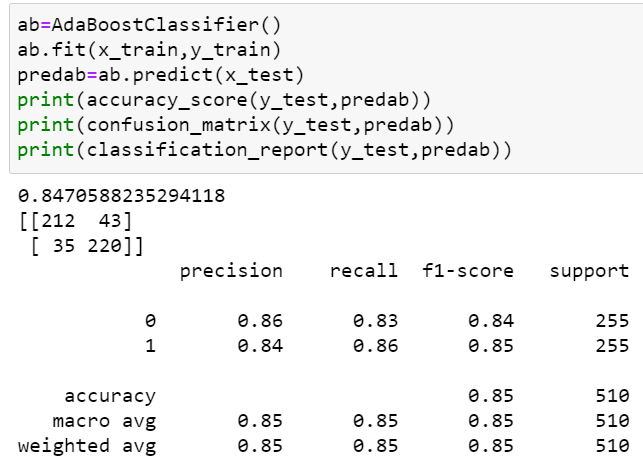
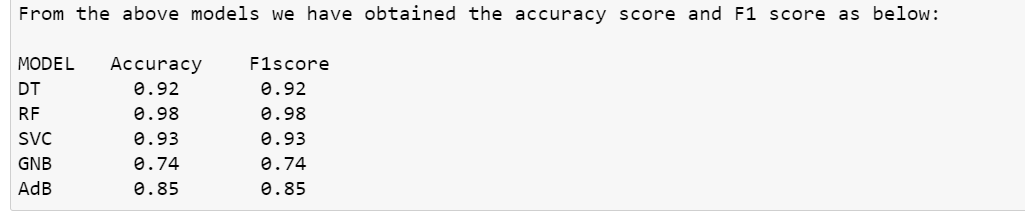
**3.SVC()**

****

**4.GaussianNB()**

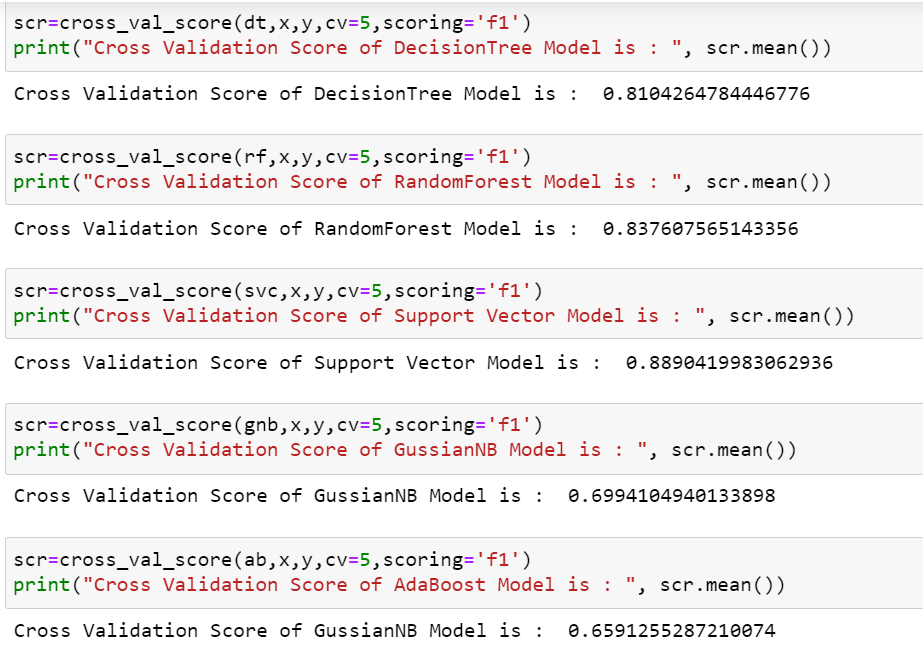
****

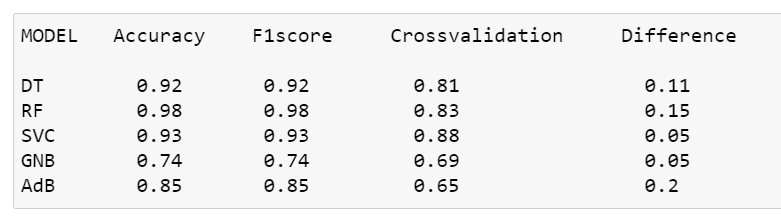
**5.AdaBoostClassifier()**

** **

**There may me over-fitting or under fitting with the accuracy scores above.**

**So let’s cross check with the Cross\_Validation methods.**

****

****

**So we have got difference with each model with cross Validations.**

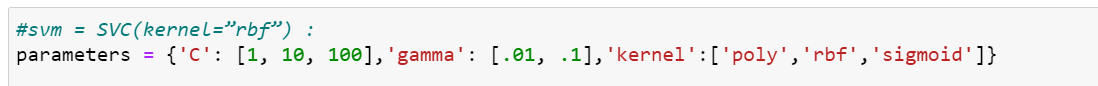
**So we will check difference of F1 score and Cross Validation.**

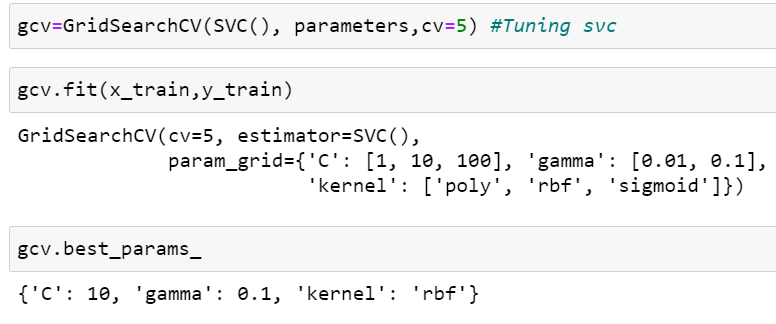
**finalize with minimum difference to proceed further for Parameter Tuning.**

**As per above, It is observed that SVC and GaussianNB both have same minimum difference. So we can will take these both models for tuning further.(SVC and GaussianNB is same)**

**Lets check for both of then in tuning.**

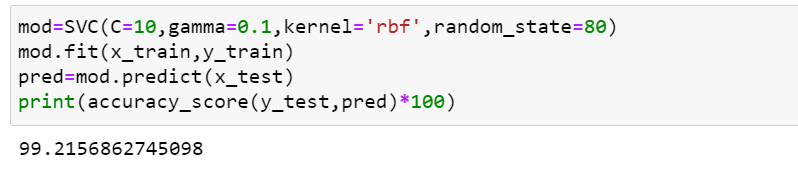
* 1. **Hyper Parameter Tuning:**
     1. **SVC**

****

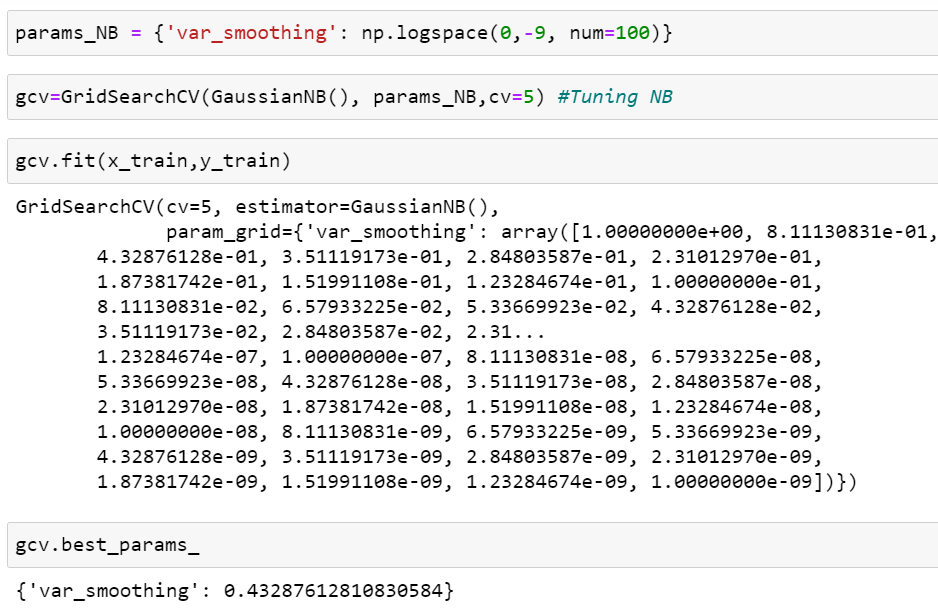
****

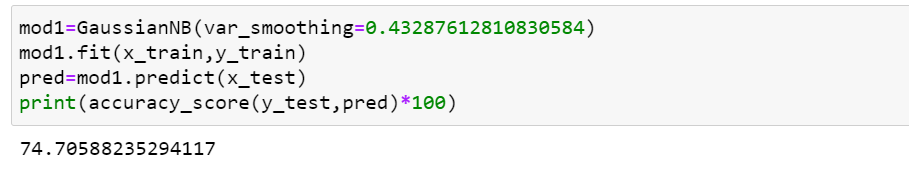
**We have obtained the best Parameters to apply in model from above.**

**Let’s pass it in selected SVC model and check the Accuracy % .**

****

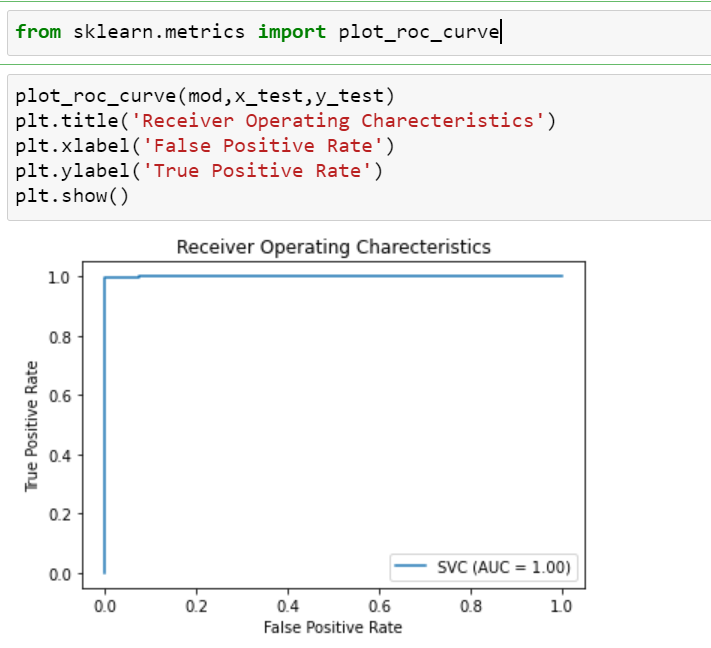
**5.3.2 GaussianNB**

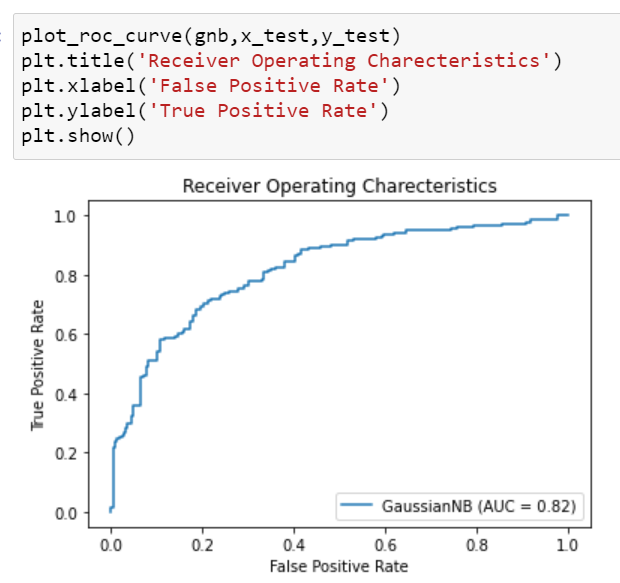
****

****

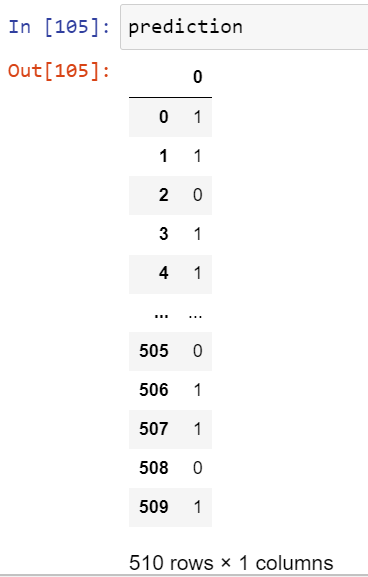
# We have got 99.21% accuracy from SVC model. Hence it is best model to proceed for Prediction.

### Lets Check with AUC-ROC curve % :





### Seems to be almost 100% with SVC model in AUC-ROC plot which is best percentage score.

****

1. **Concluding Remarks:**

After Hyper Parameter Tuning, we have got 99.21% accuracy from SVC model and good output from AUC-ROC graph.

Hence we have Saved the model and Checked by loading whether its working fine and checked with predicting.

In this Machine learning Project,we have learned to identify the Employee Attrition using various factors. We used SVC Classifier for this and made use of sklearn libraries to prepare the dataset.