**HR Analytics Project- Understanding the attrition with HR analytic**

* **Introduction:-**

Companies hire new employees every year and spend lot of money and invest time on them to train. It is not only company provides training to their new employees but existing is also trained for effective delivery and to increase the productivity timely. “**Attrition is also an important parameter in a company and it is being reviewed rigorously by the top management on regular interval”.**

A question must be arising in the minds that how “HR analytics” helps in improve employees’ efficiency. HR analytic has significant role in process improvement. We will read further the relation and the contribution of HR analytics in our coming sections. Please read full article to get complete understanding.

This article is containing the following sub-topics

1. Problem Definition
2. How attrition impact the business
3. How to HR analysis help in understanding probable attrition case
4. Data Analysis

A. Understanding the data

3. EDA Concluding Remark

4. Pre-Processing Pipeline

5. Building Machine Learning Models

6. Concluding Remarks.

**Let’s get into details sub topic wise and understand:-**

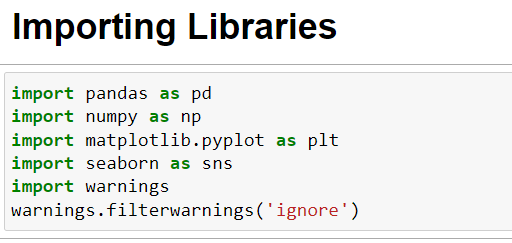
**Problem definition:-**

As we have already read in the introduction part that companies spend money and invest time to trained new hires and run many training programs internally for existing employees to enhance their work efficiency subsequently.

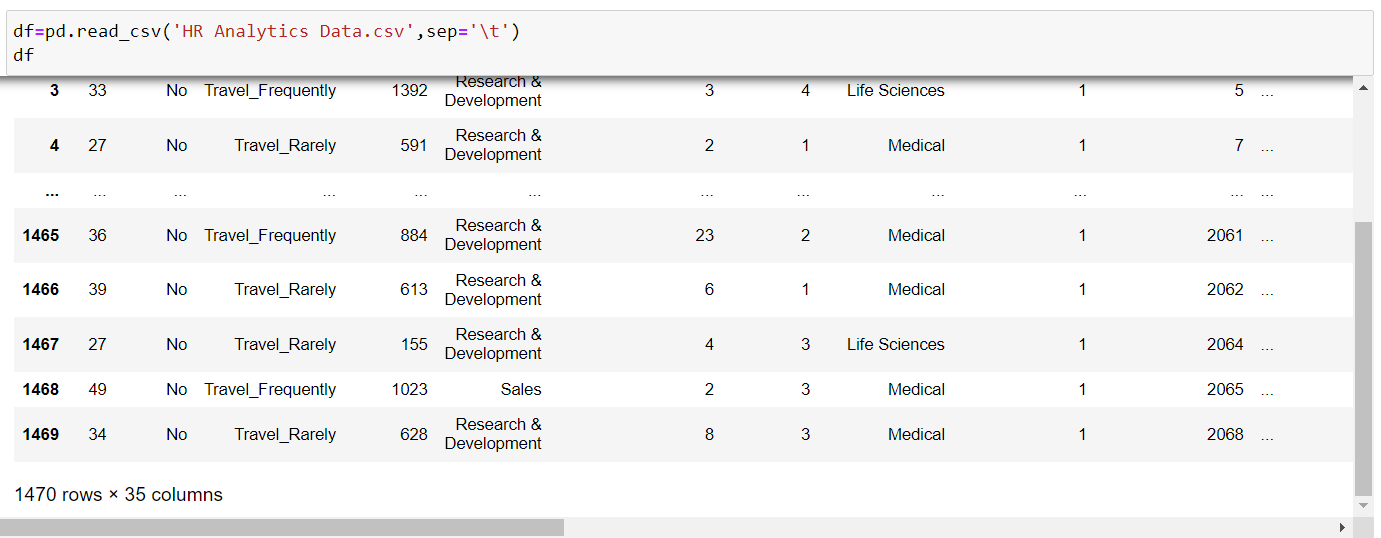
Attrition is considered a major & critical parameter in companies. It is said that old is gold and I believe this example fits perfectly in this case, experience brings lot of good work for the companies. HR conducts timely survey and basis on that gauge employee’s satisfaction and also they conduct many engagement activities to make employees feel good at work. Basis on survey data and historical attrition data, HR analytic brings many innovative idea and initiatives to control the attrition. Let’s get into deep and see how HR analytics provide insight

**Importing libraries:-**

For analyzing data, we would load dataset by using pandas’s read\_csv function , if you want to refresher your idea about pandas, please visit pandas official site and documents. we are also importing important libraries which will help in analyzing and model building.

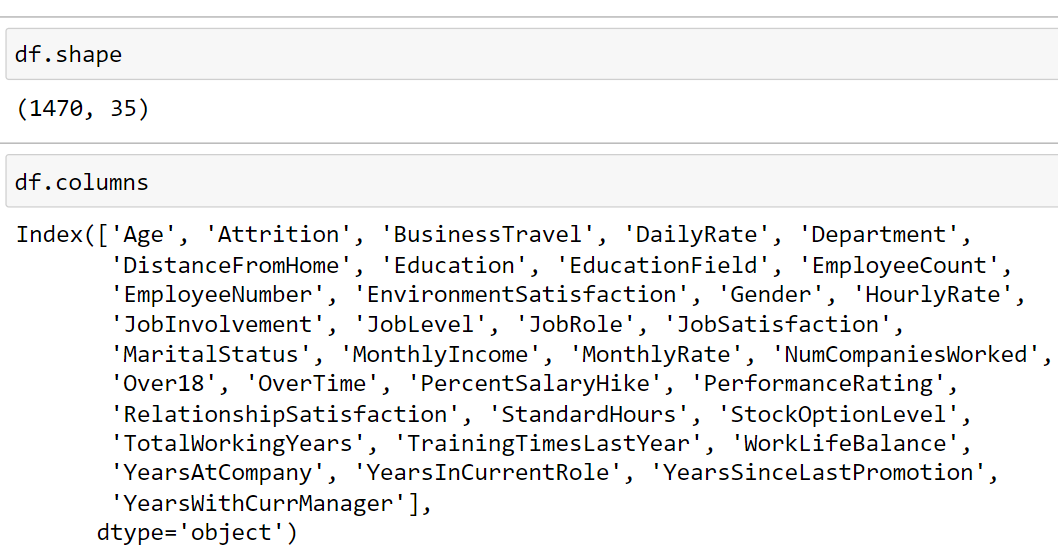


**Loading/gathering data:-**



Checking the first 5 and last 5 rows of our entire dataset. We can see that our dataset comprises of total 1470 rows and 35 columns. The column "Attrition" is our label that needs to be predicted in assisting the HR professionals to understand what causes attrition in an organization and help them with retention process. Apart from the "Attrition" column right now all the remaining columns are our features that we will be using to generate our prediction. Since our label column is based upon binary classes this becomes a Classification problem!

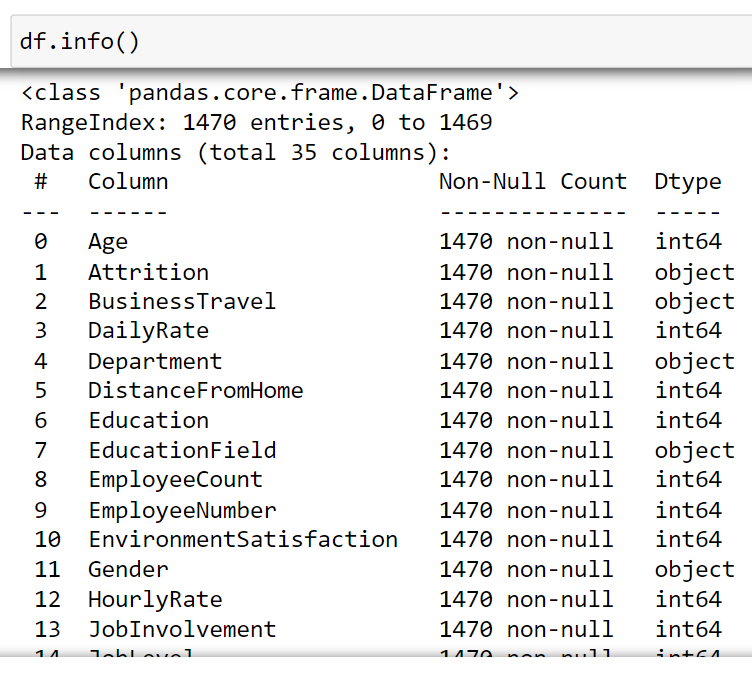
# Exploratory Data Analysis (EDA)



Dataset consists of 35 columns and 1470 rows. Let’s check the data type of each column

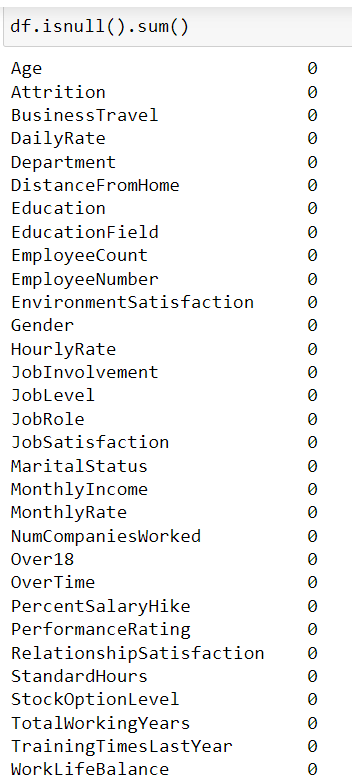
We can see data type of each feature. I have define manual function to check it but we can also use pandas dataframe.info() function to check it, this function provides information about dataset, number of observations, number of columns and missing values information.

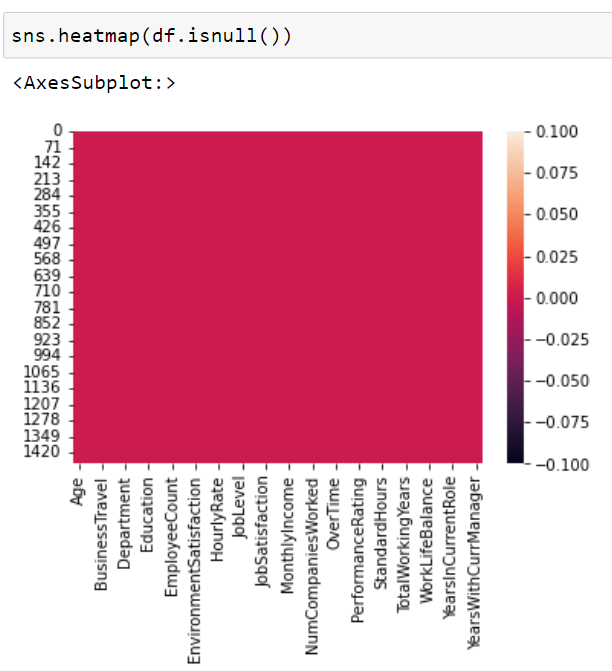
**Checking information:-**



Data consist of total 35 columns, 26 integer datatypes and 9 object datatypes.

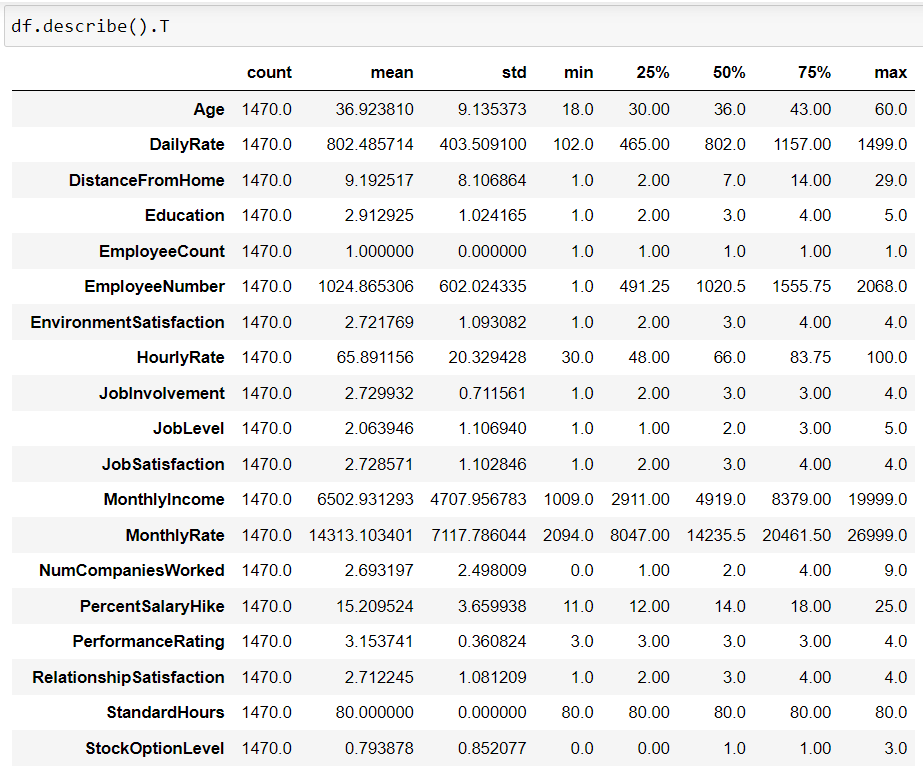
**Check missing values:-**

****

****

No missing value found in the data. Missing value can be checked by pandas’s .isnull() or isna() function

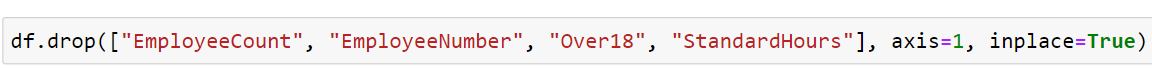
**Check statistical information by using describe function**



Understanding data’s stats is essentials for further analysis. Pandas describe() function provides insight about descriptive statistic information like mean, median, mode, standard deviation, min and max value and we can check the percentile of data.

With the help of describe method in transpose format we are able to take a look at our column details clearly. The count column once again confirms that there are no missing data concern in our dataset. However if we check the min column we do notice columns with zero as their values. However considering the columns that have 0 value in them it is quite possible for freshers in an organisation to have them as 0 marked in their records.

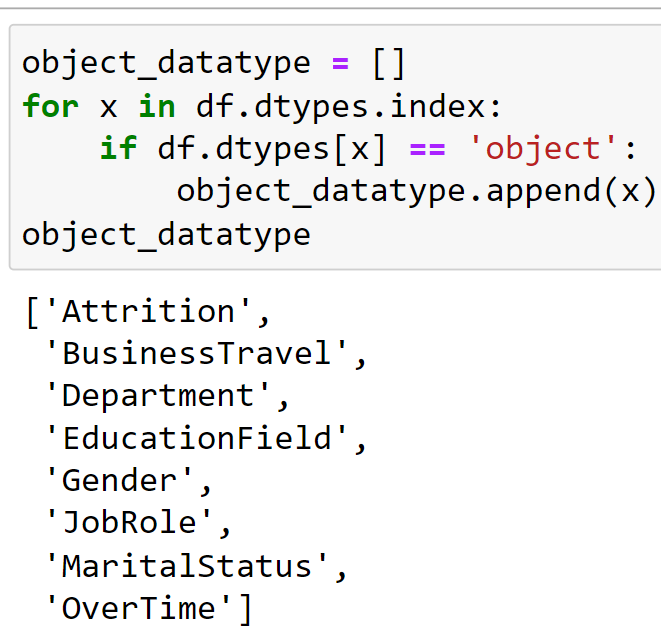
**Dropping unwanted column:-**



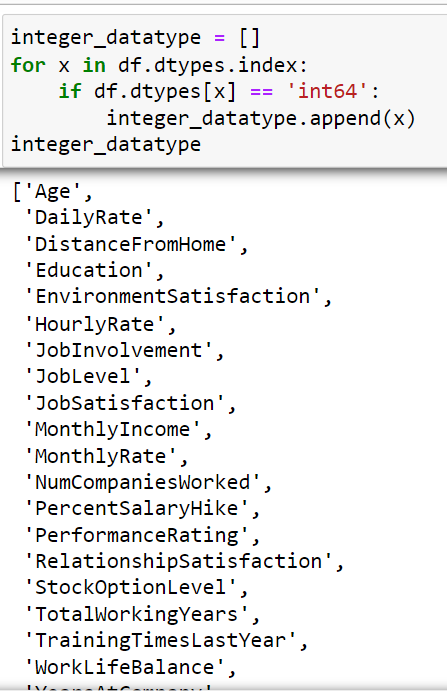
I am dropping all the unwanted columns after careful observations done above.

EmployeeCount - All the rows in this column are filled with just a single number "1" which cannot provide much information related to attrition of an employee EmployeeNumber - Since it is just a unique number provided to each employee that has nothing to do with attrition Over18 - As per child labour law any person below the age of 18 is not eligible for employement in India and also this particular column has 1 single value for all the rows therefore it does not add much value considering the attrition of the employees StandardHours - Again in this column we have single value for all the rows that is "80" hours as a standardized policy so does not make much difference with respect to attrition as there is no partiality in these terms on any employee

**Dropping unwanted column:-**

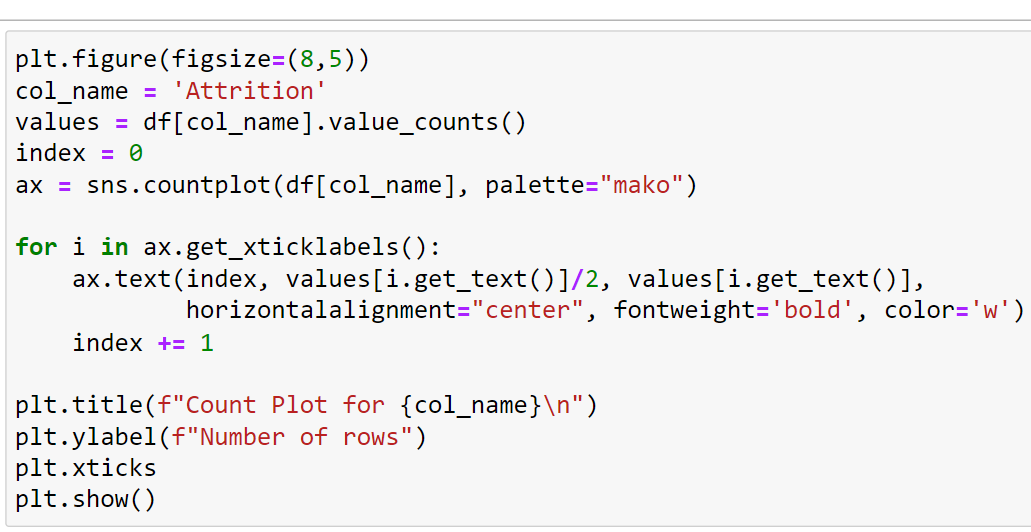


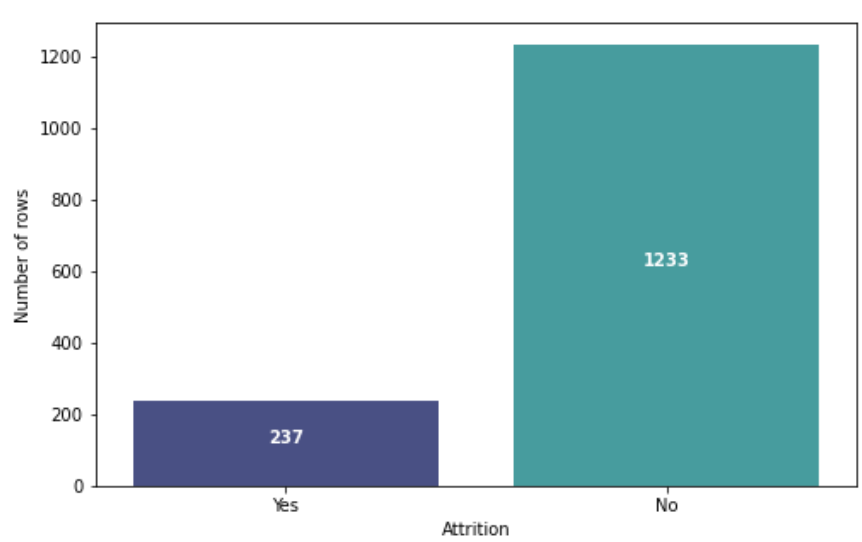
Here we have a list of all the 8 columns that hold object datatype that will need to be encoded into numerical format before creating our classification model.



Here we have a list of all the 23 columns that hold integer datatype that will comprise the numerical data part of our ML project

**Visualizations:-**



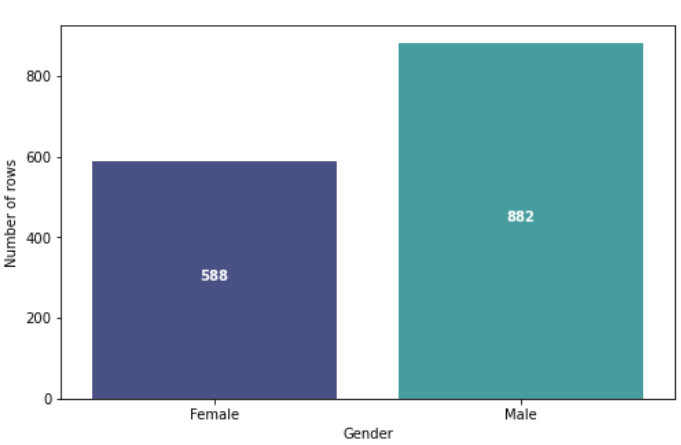


The attrition rate is = 237/1470 = 16.1 %.

This indicates that the data set is an imbalanced dataset where the number of observations belonging to class 1 (No) is significantly higher than those belonging to class 0 (True)

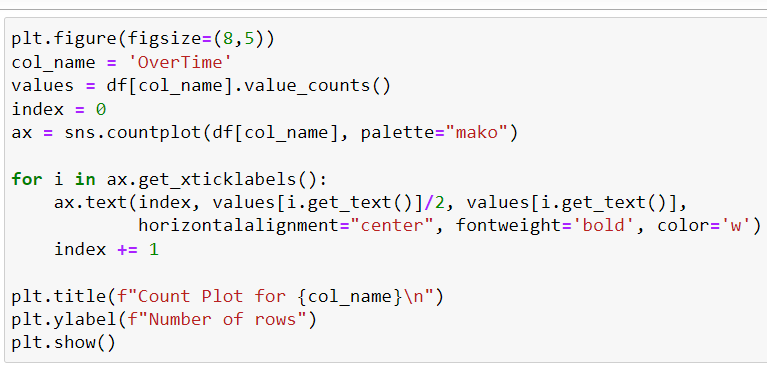
**Visualizations:-**

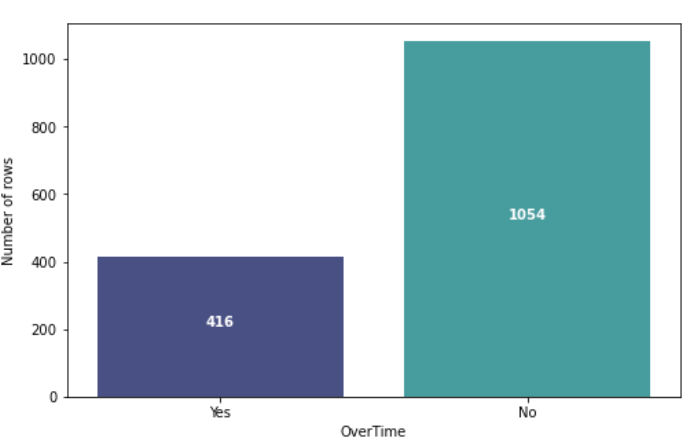




In the Gender column we can see that the number of male employees is higher than the female employees. This is one of the situation in mostly all the organizational workforce across India

**Visualizations:-**

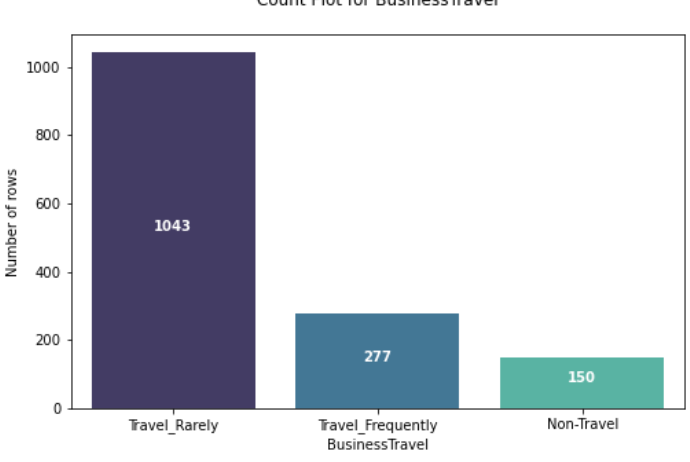




The OverTime column shows us the number of employees who do over time and the one's who do not. And it looks like from the above count plot that employees do not prefer doing over time in the company.

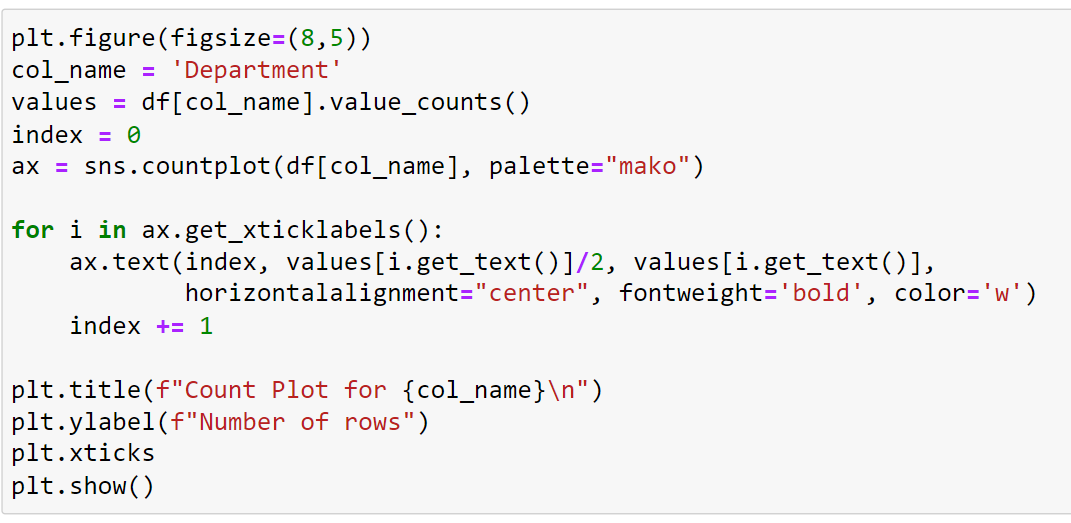
**Count plot for Business Travel:-**





In the Business Travel column we see a majority of number in the Travel\_Rarely value while Non\_Travel are the least of them

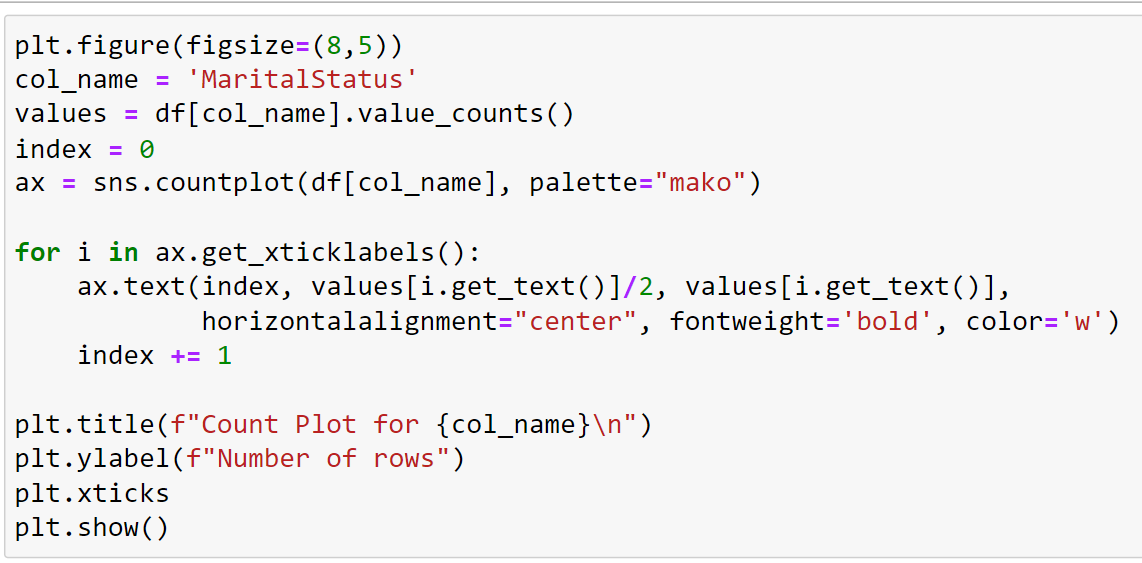
**Count plot for Department:-**

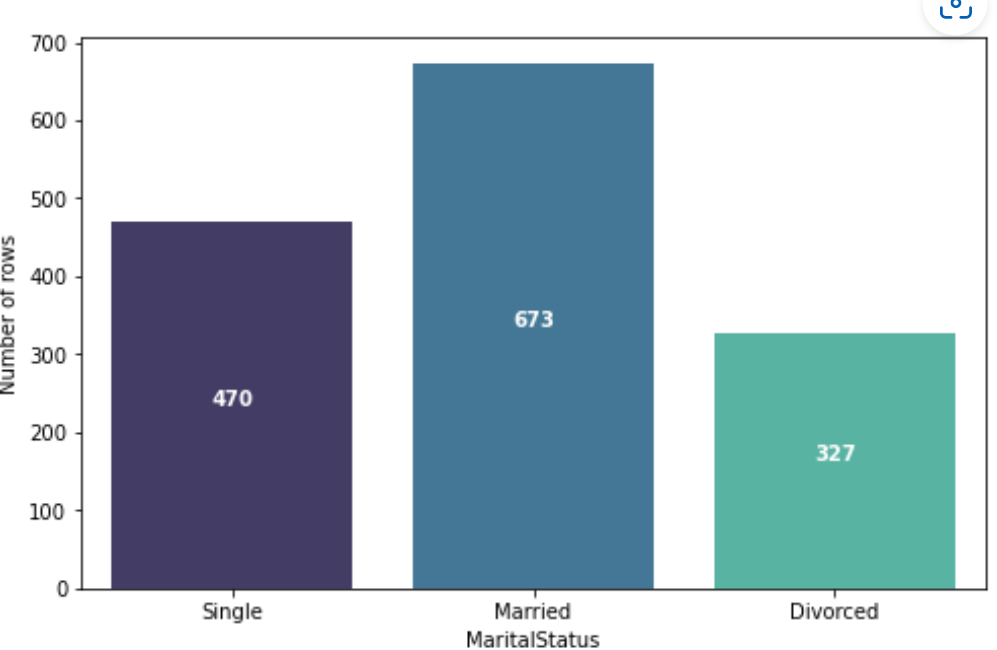
****

****

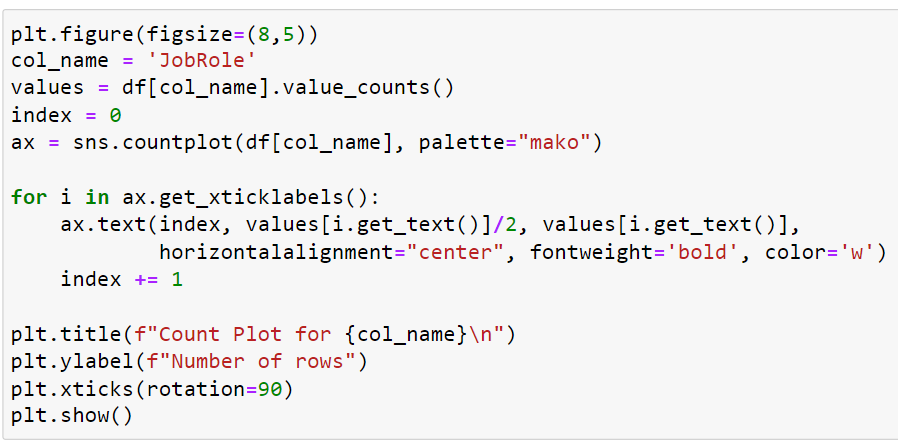
In the Department column we see lots of values for R&D department however the least number of employees are in HR since an organisation would need a limited number of Human Resources team it makes sense for it to have the least value.

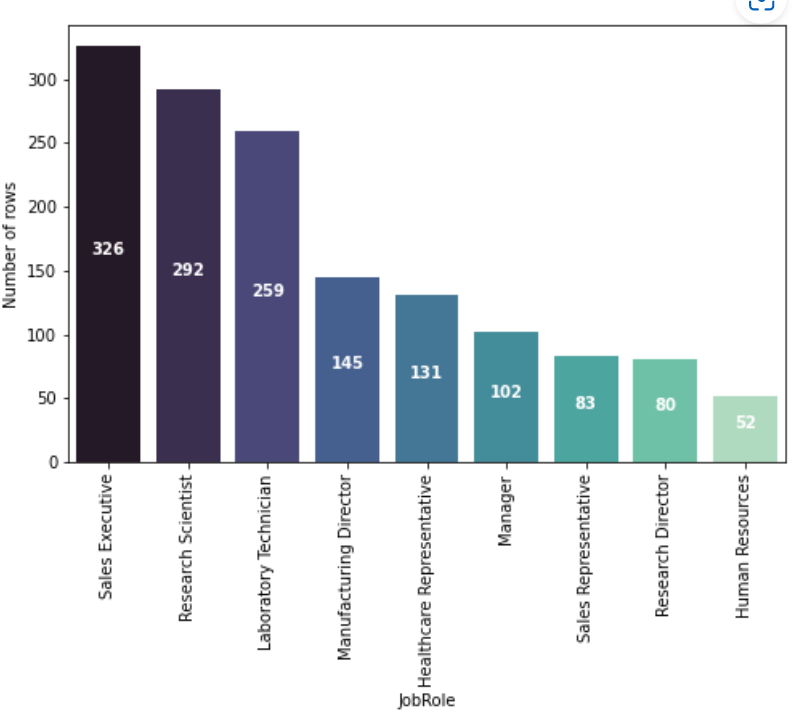
**Count plot for MaritalStatus: -**

****

****

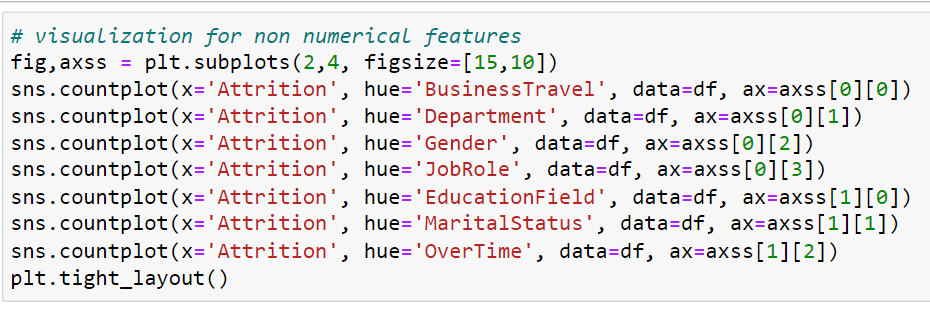
**Count plot for MaritalStatus: -**

****

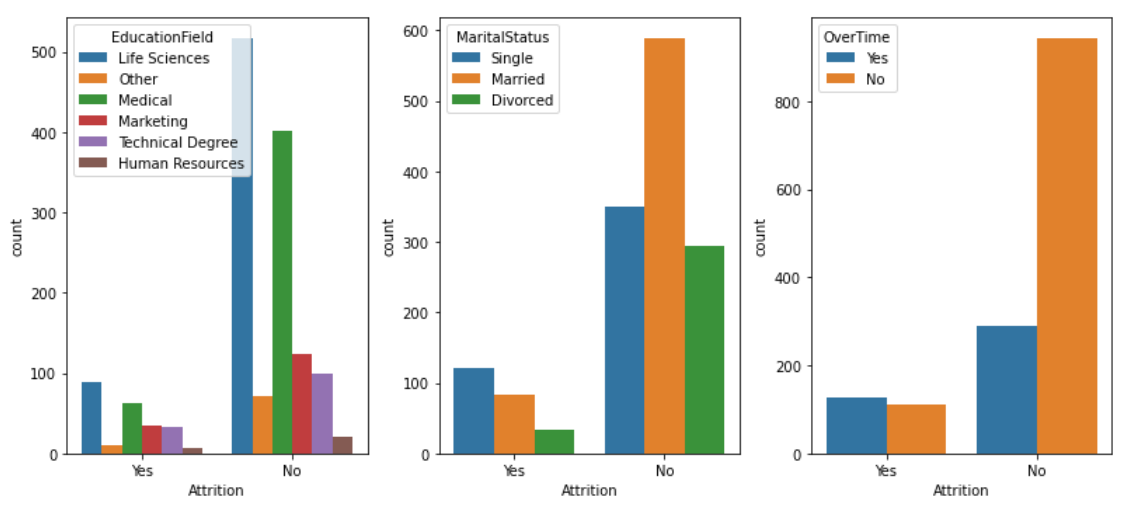
****

Pertaining to the JobRole column we can see that the highest number is found in sales executive designation since they are the grass root level employees who are the most productive and handle the business value directly. And as notice in other plots we have the lowest number for HR roles as the employees present in that department are the lowest as well.

**Feature wise attrition analysis:-**

****

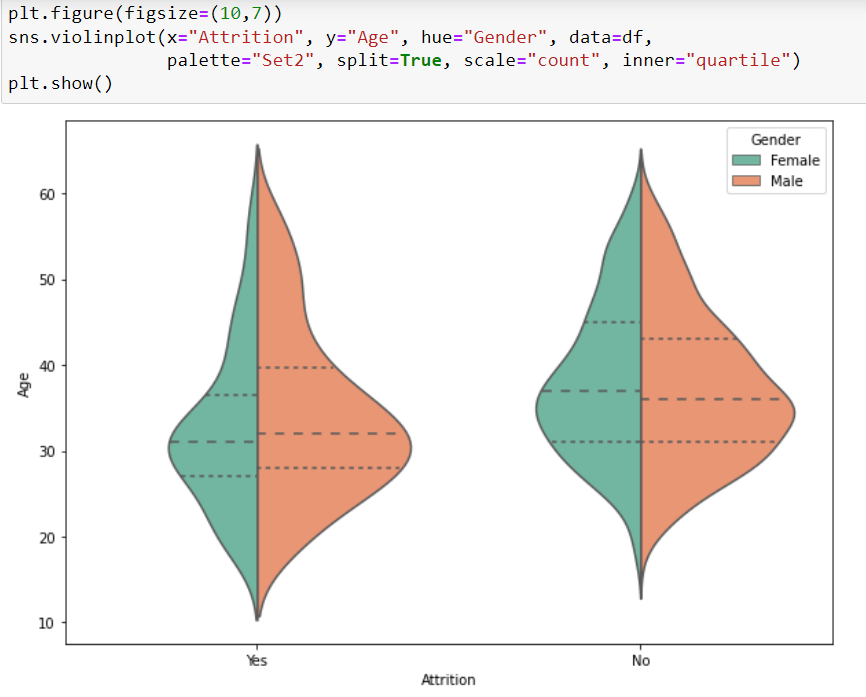
****

****

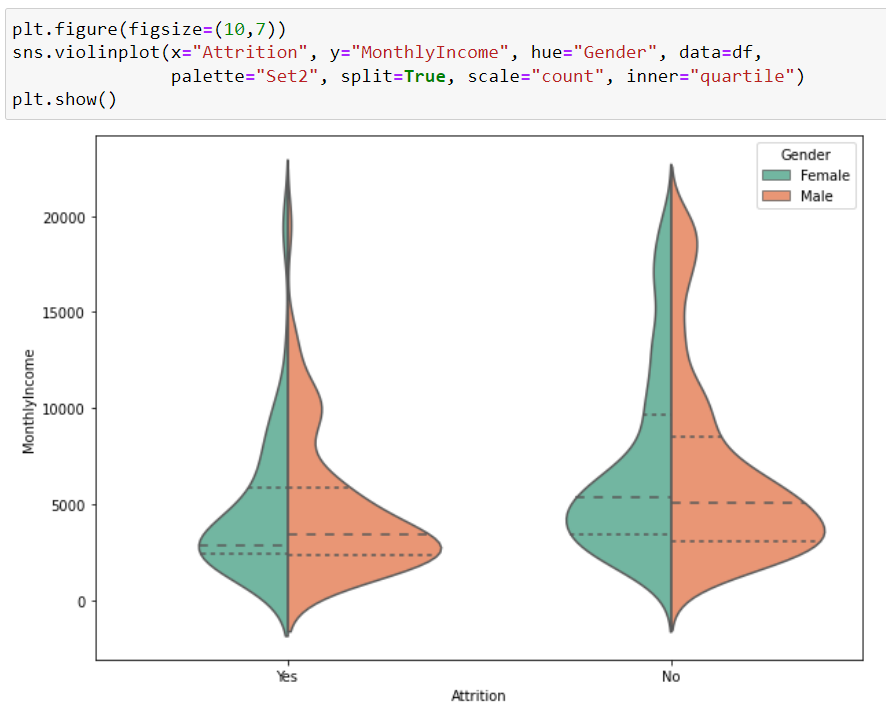
It seems that people who left the company they probably wanted to change or looking for other opportunity. Age is also one factor; they were younger than the other people who didn’t attrite. We can observe couple of more factors that people attrited like average daily rate, distance from home, monthly income.

**Let us explore more things:-**

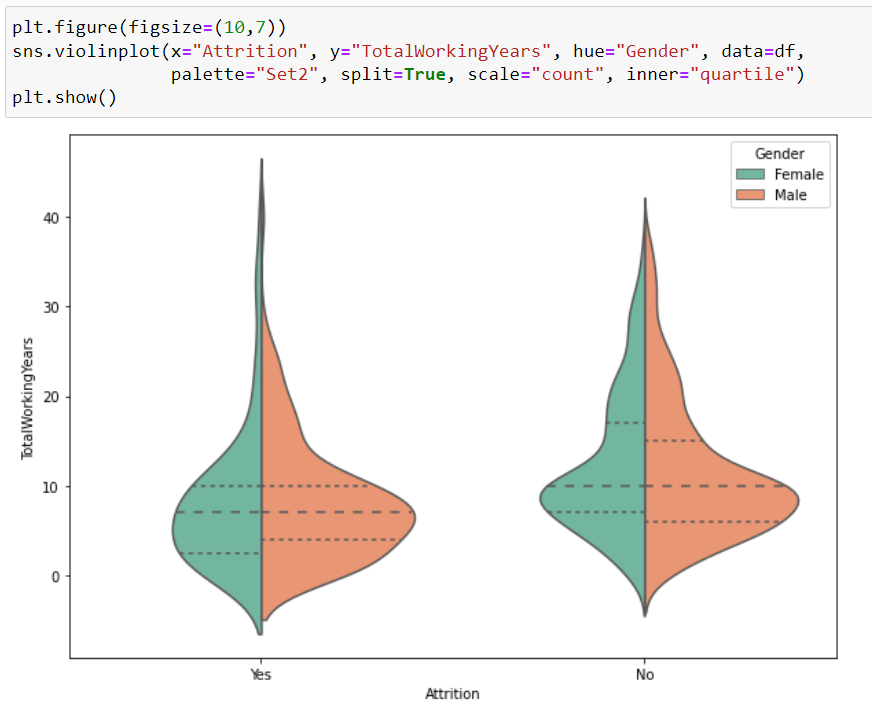
I tried checking the relationship between travel roles and non travel roles, have seen no relation between attrition and travel profile, I thought of may be travelling would be high as people didn’t like to travel frequently



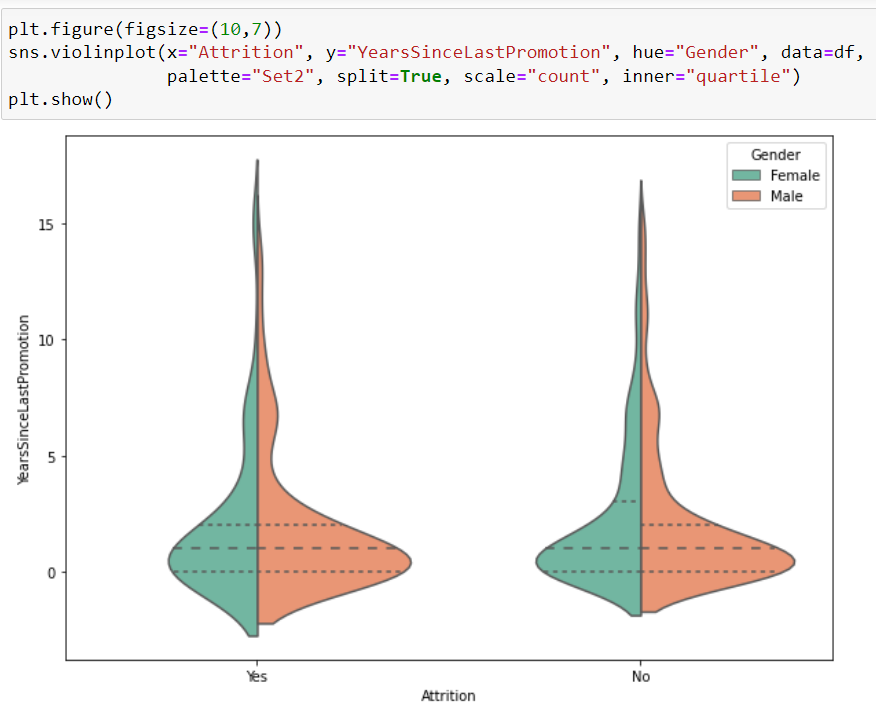
In the above violin plot we see that the Attrition counts both in male and female are high when the are in their mid 20's and 30's of age.



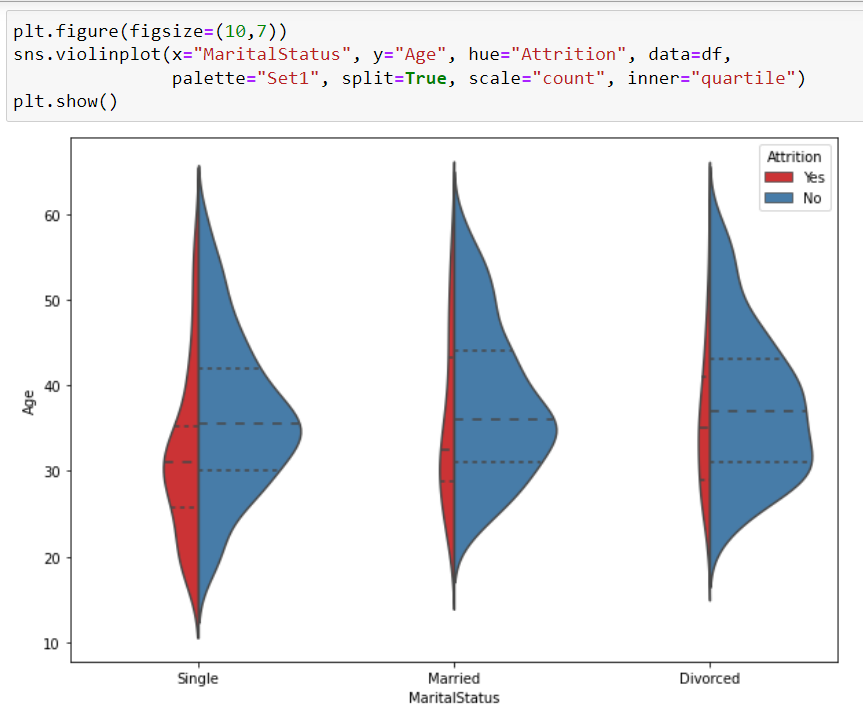
In the above plot we can see that the Attrition peaks for both male and female employees when the monthly income is less than 5000.



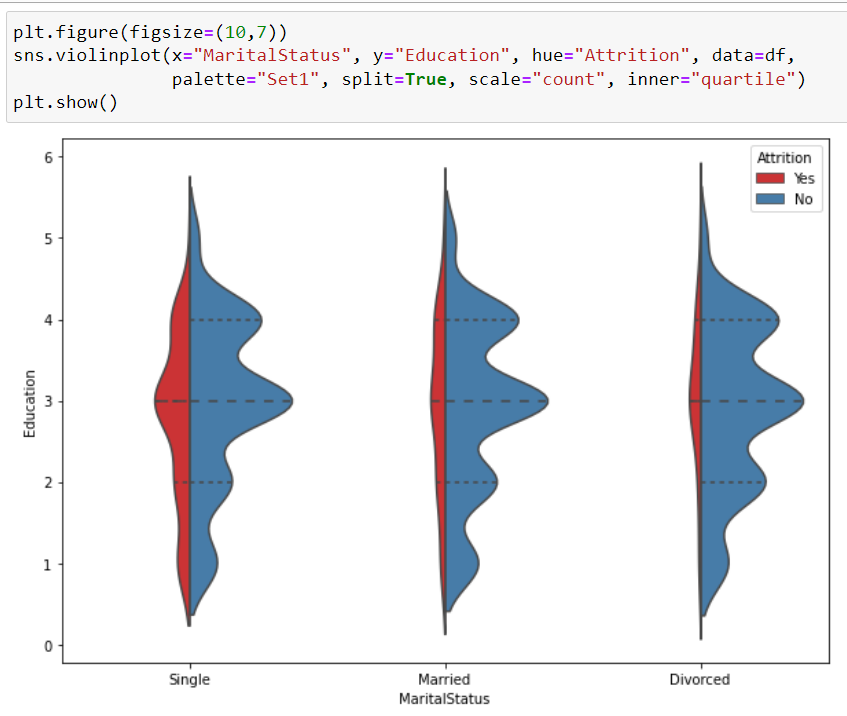
In the above violin plot we notice that the Attrition for both male ad female employees occur when they are in their experience range of 1-10 years.



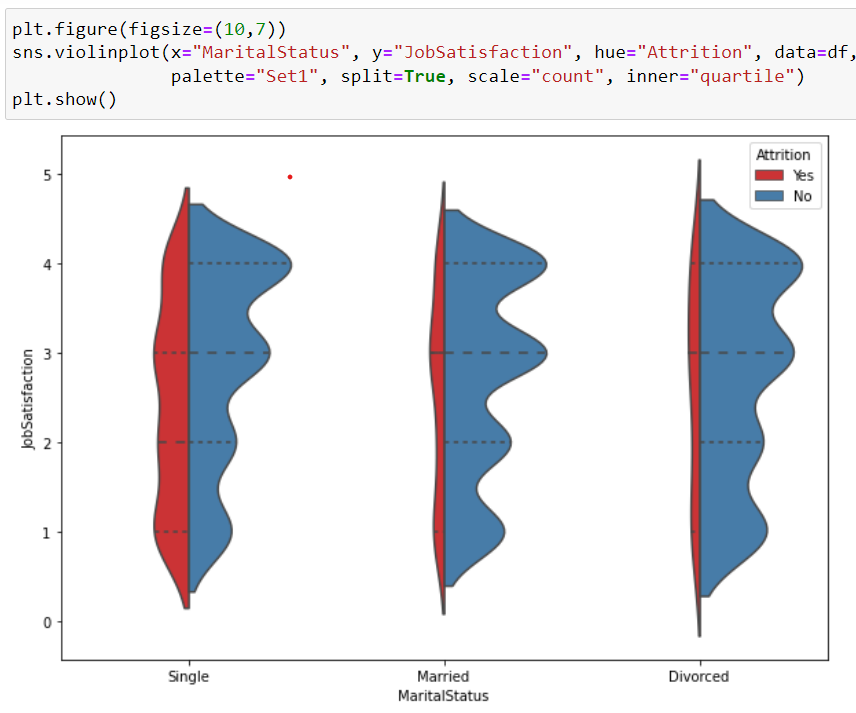
In the above plot we see that the Attrition for both make and female employees happen when they do not see prootions happening after years of gaining experience.



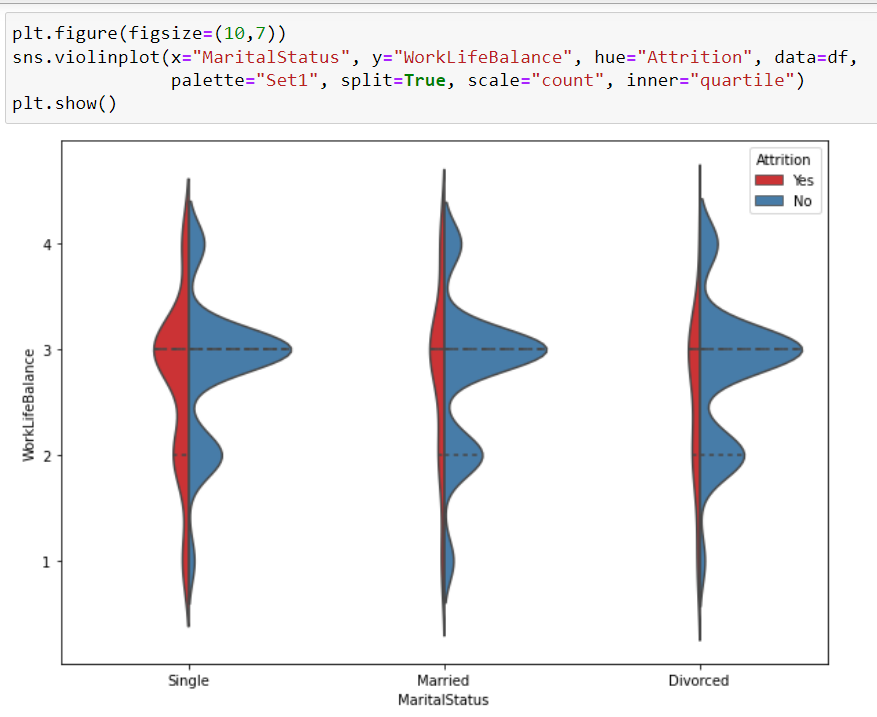
In the above violin plot we can see that the Attrition rate is quite less in employees when they are married or divorced as compared to when they are single and have lesser responsibilities to deal with at their age

****

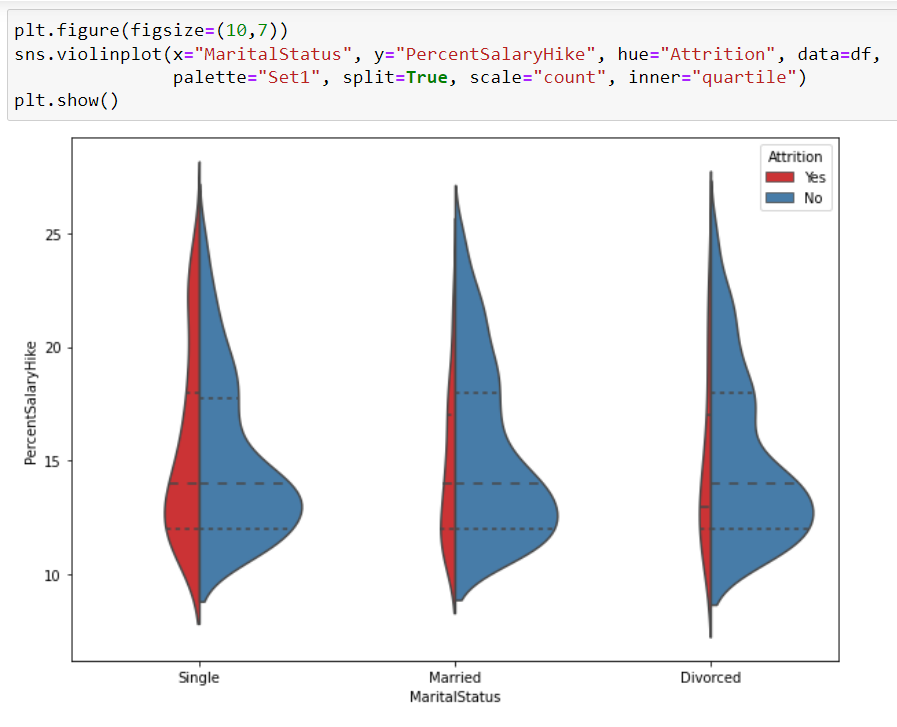
In the above plot we notice that once again employees who are married or divorced and with good education choose stability in life rather than the one's who are single and are okay to take risks and oppotunities in life

****

In the above violin part we can see that the job satisfaction part for singles is not that great compared to employees who are married or divorced may be due to the year of experience difference that makes a huge gap in pay scale. But we do notice stability and lesser attrition rate amongst employees who are married or divorced



In the above plot we can see that Work Life balance maintained by singles are quite less therefore there are attritions observed as they have to achieved lots of skills to get better in their career.

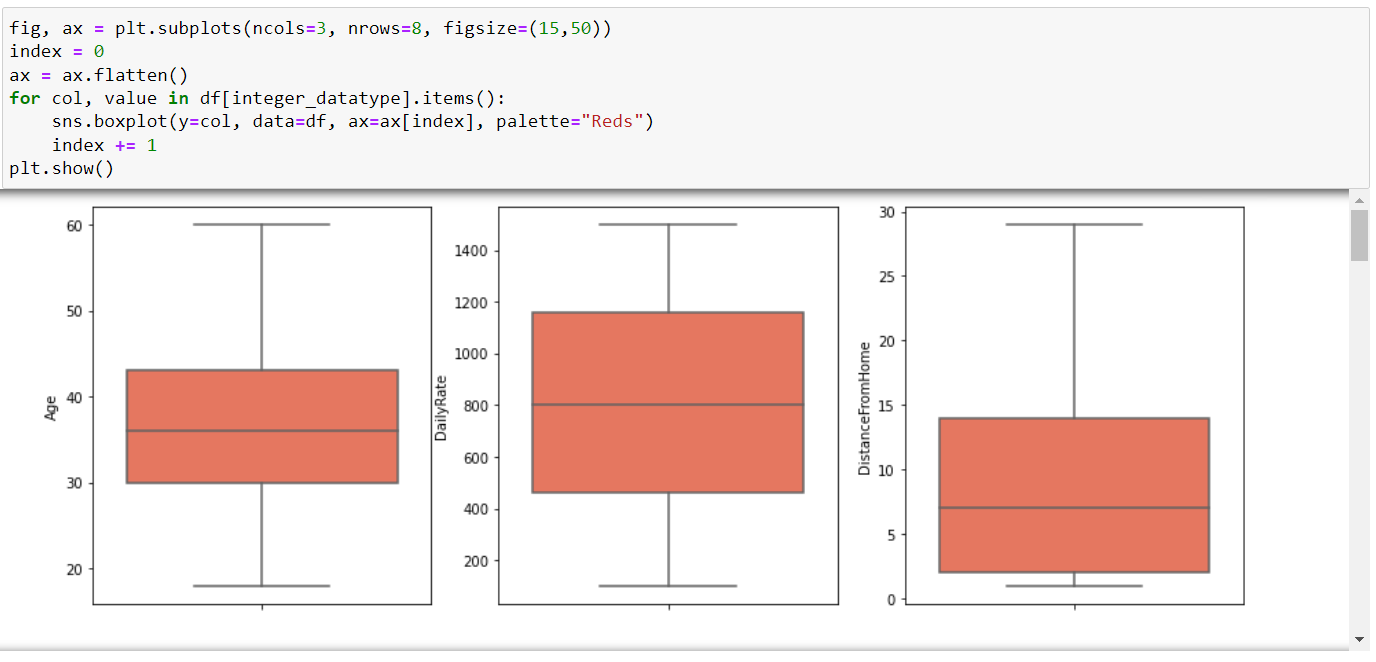
****

In the above violin plot we can see that the Percent Salary Hike plays a major role when it come to Attrition amongst the Singles as compared to their married or divorced counterparts.



If monthly income is less than 5000 then irresptive of maritial status attrition is high in each case.

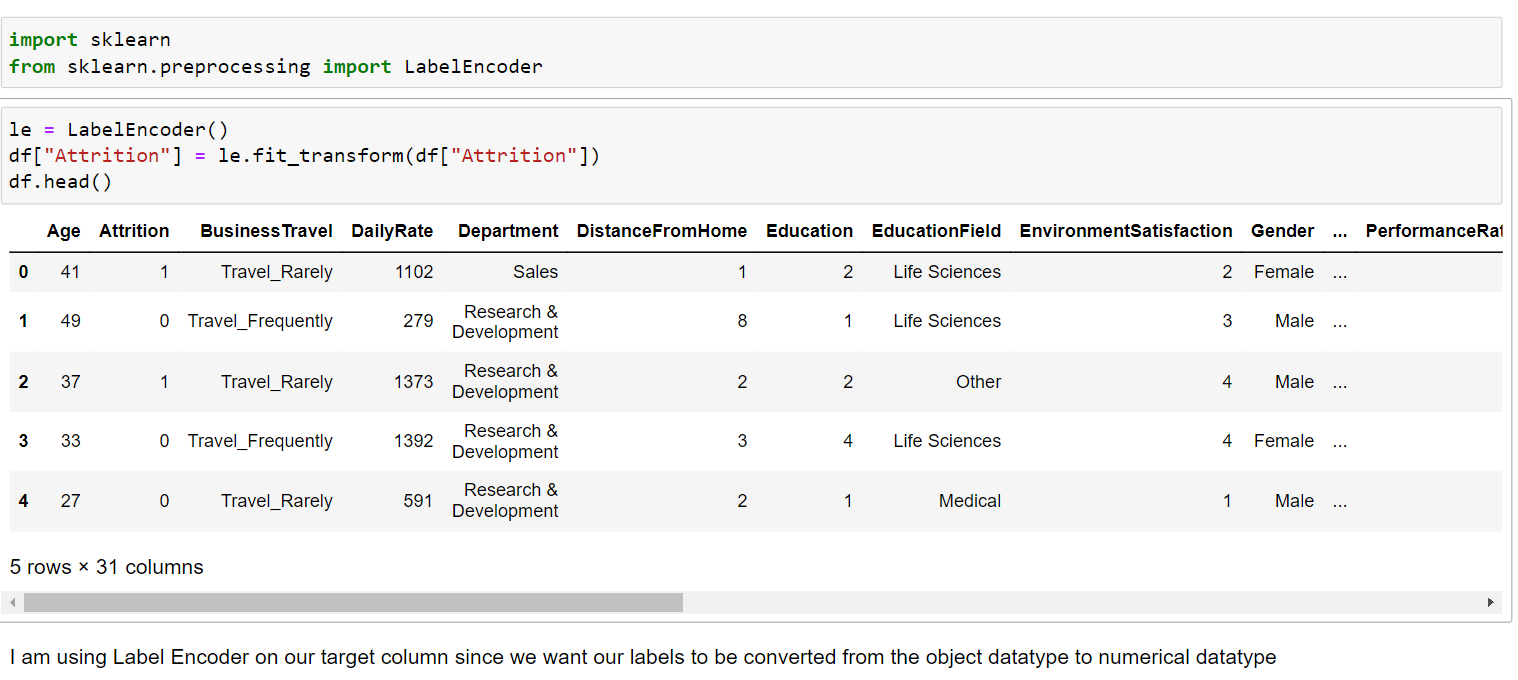
**Outlier detection using boxplot**

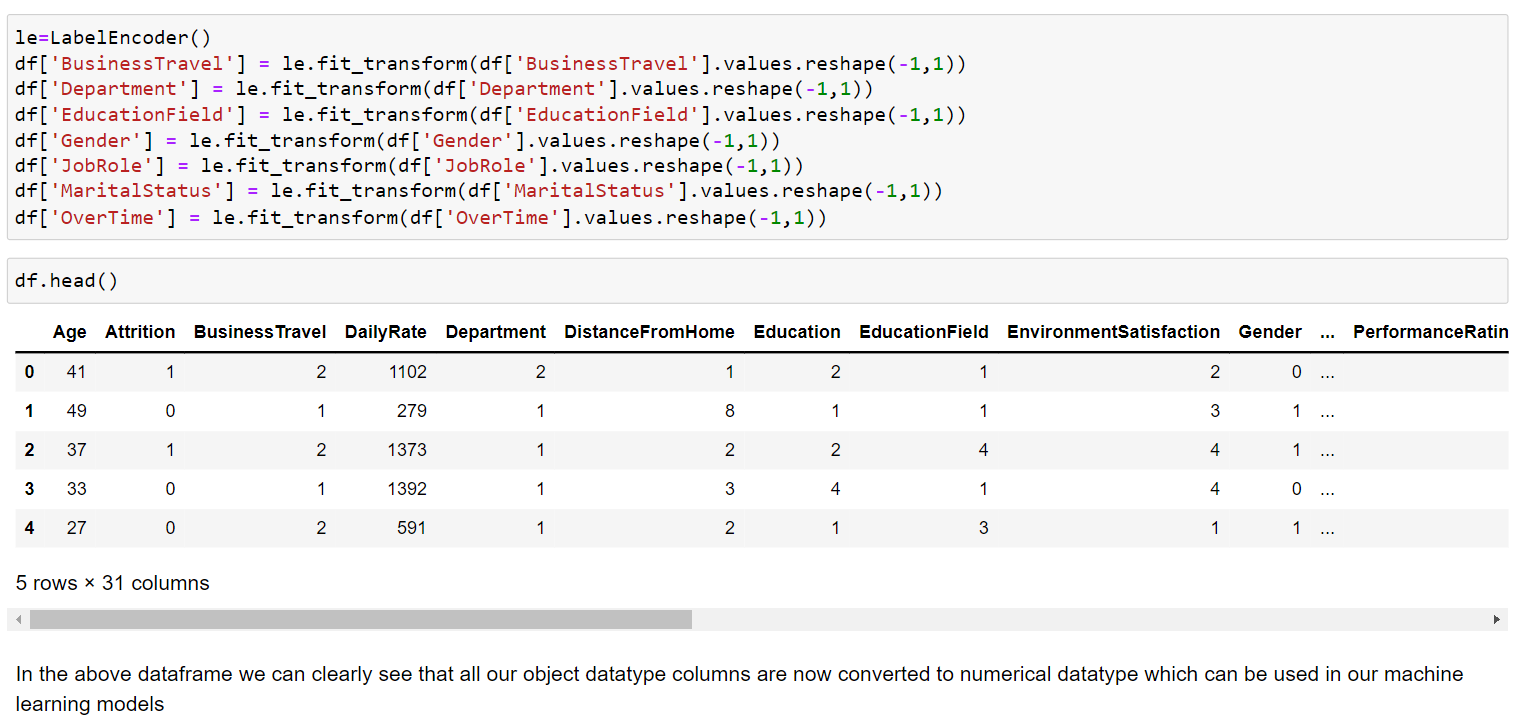


We have created a box plot visual for all our integer datatype columns to check for outliers. We do see some of the columns where there are presence of outliers and we will need to treat it accordingly.

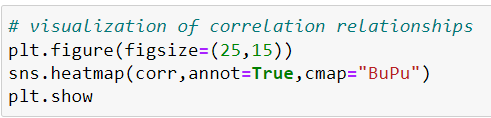
* MonthlyIncome
* NumCompaniesWorked
* PerformanceRating
* StockOptionLevel
* TotalWorkingYears
* TrainingTimesLastYear
* YearsAtCompany
* YearsInCurrentRole
* YearsSinceLastPromotion
* YearsWithCurrManager

**Encoding the categorical object datatype columns**

****

****

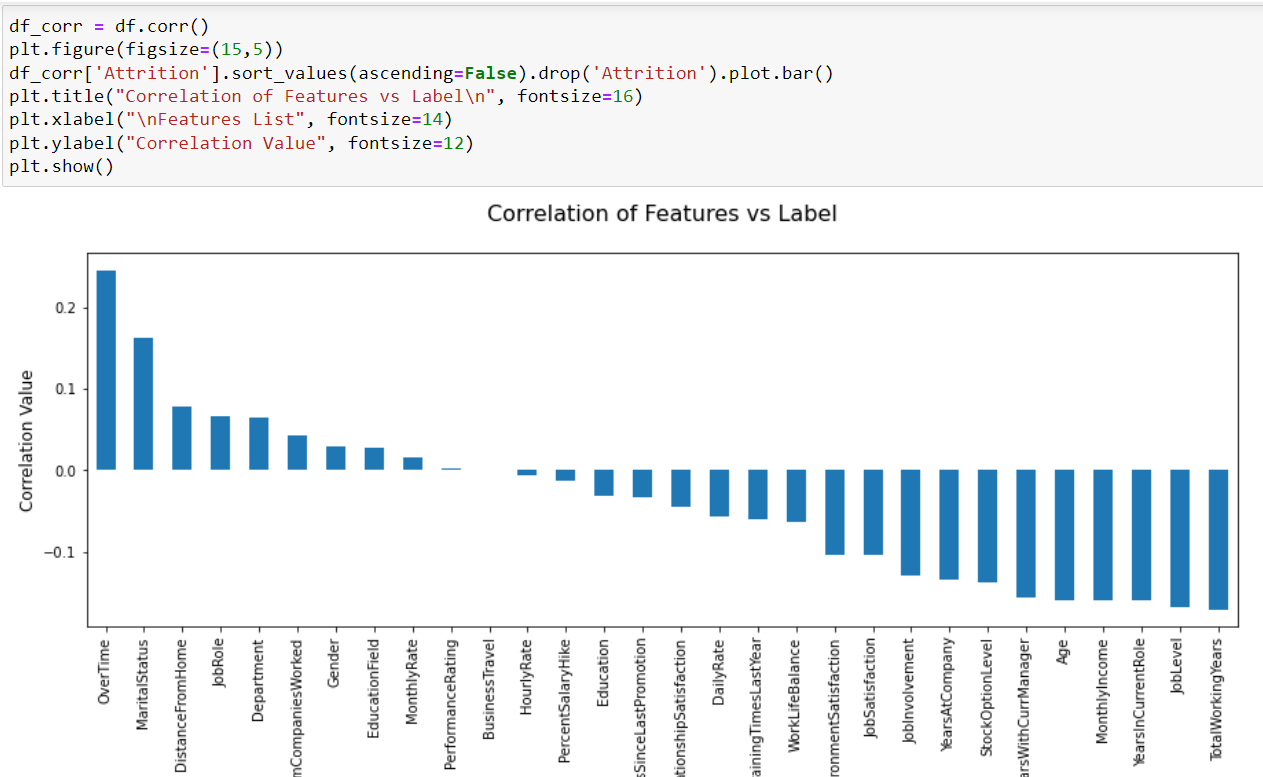
**Visualization of correlation relationships**

****

****

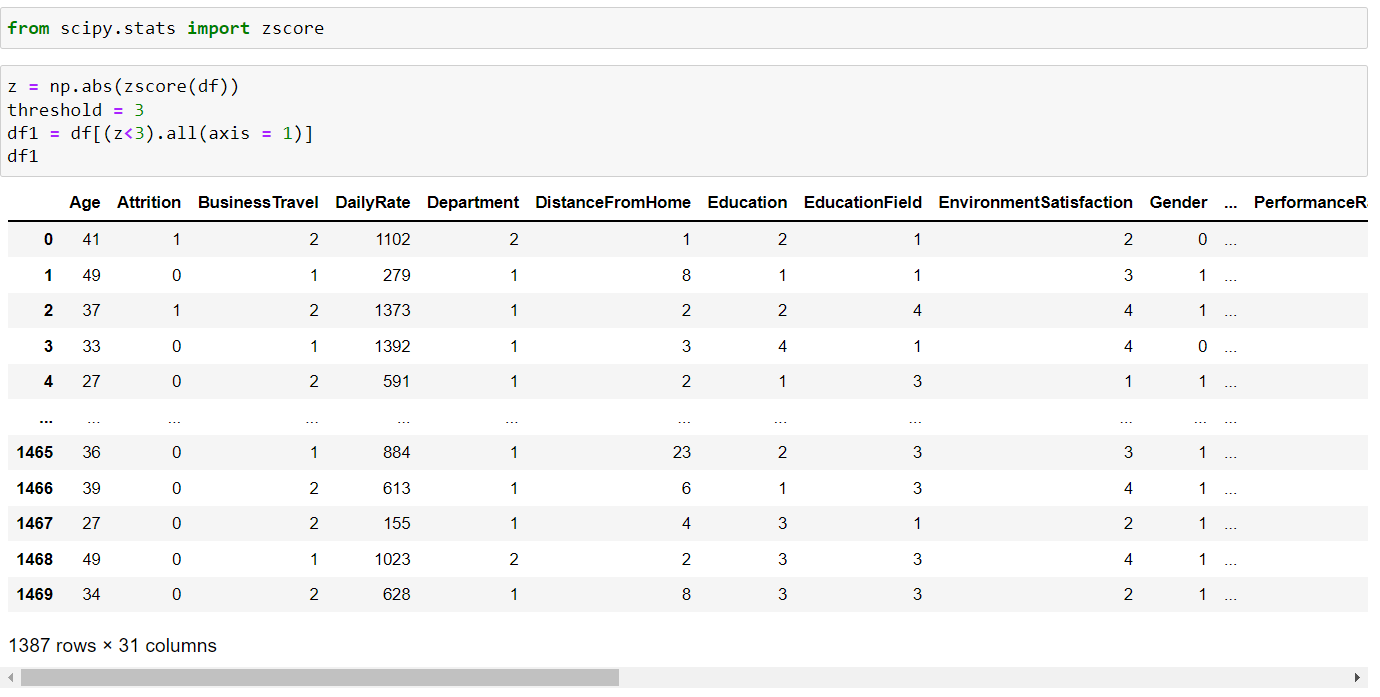
In the above heatmap we can see that our target label "Attrition" has both positive and negative correlations with the feature columns. Also we see very less or negligible amount of multi colinearity so we will not have to worry about it. Since the one's which are reflecting the value are inter dependent on those feature columns and I intend to retain and keep them

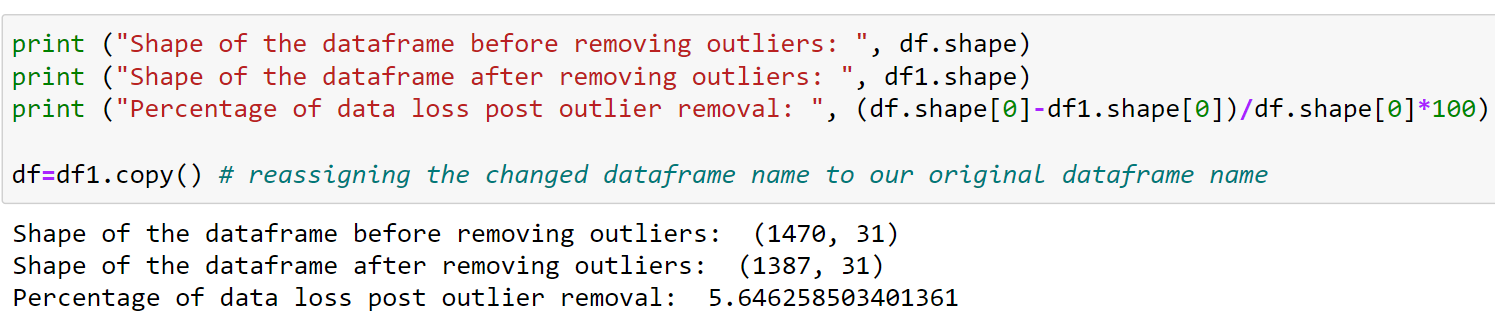
**Correlation of Features vs Label**

****

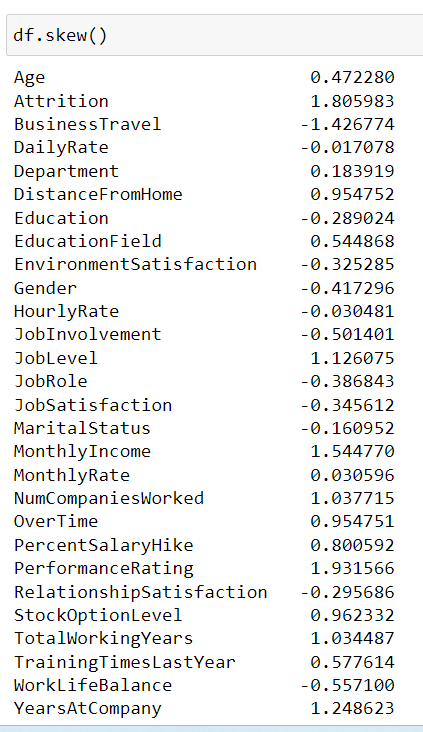
In the above Bar Plot we are able to clearly define the feature columns that are positively correlated with our label and the feature columns that are negatively correlated with our label

**Using Z Score to remove outliers**

****

****

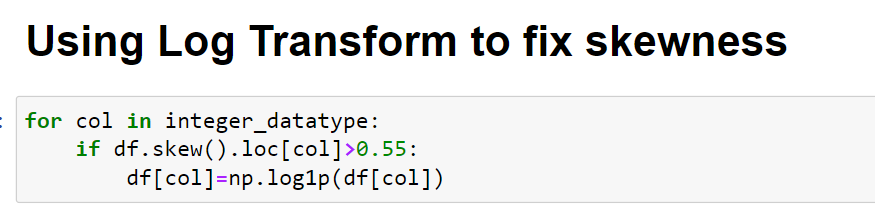
**Checking Skewness:-**

****

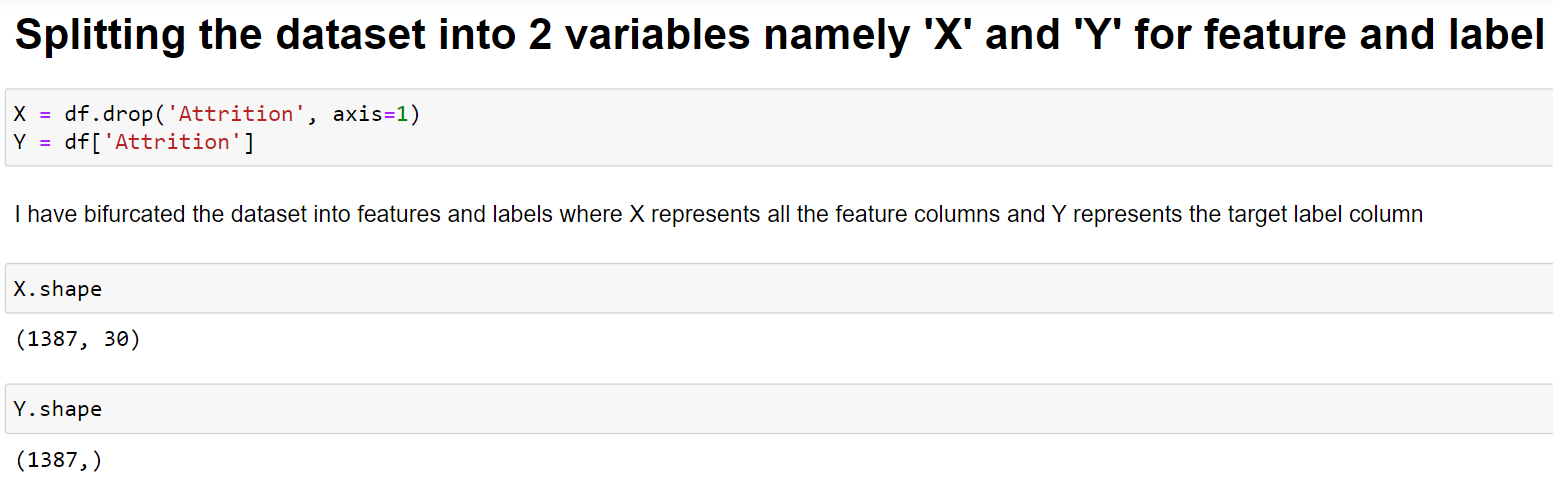
With the skew method we see that there are columns present in our dataset that are above the acceptable range of +/-0.5 value

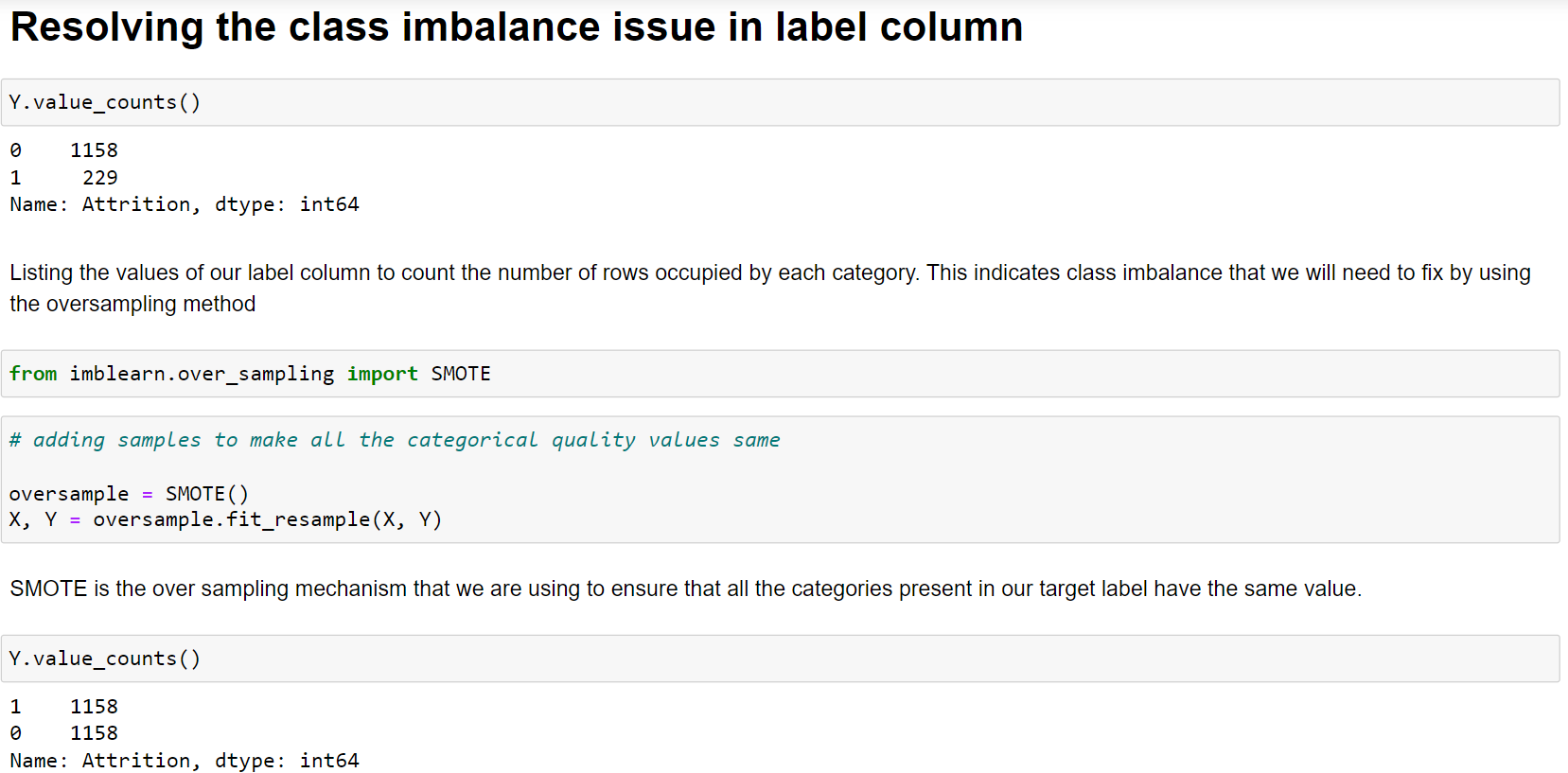
Skewness is a measure of the symmetry in a distribution. ... It measures the amount of probability in the tails.

Standard threshold is 0.5 which is considered to be used to modelling, if it goes up we need it to be corrected. There are several ways to correct it, a few operations which is applied to correct them, these are square root transformation, cube root transformation, log transformation (for non zero values), logp1 transformation (if particular column have zero values in it), boxcox transformation for non zero and positive data point, power transformer ( in scikit learn power transformer, we have two method 1. yeo-johnson' which is default one and 2. Boxcox. So apply transformation as per the column nature.

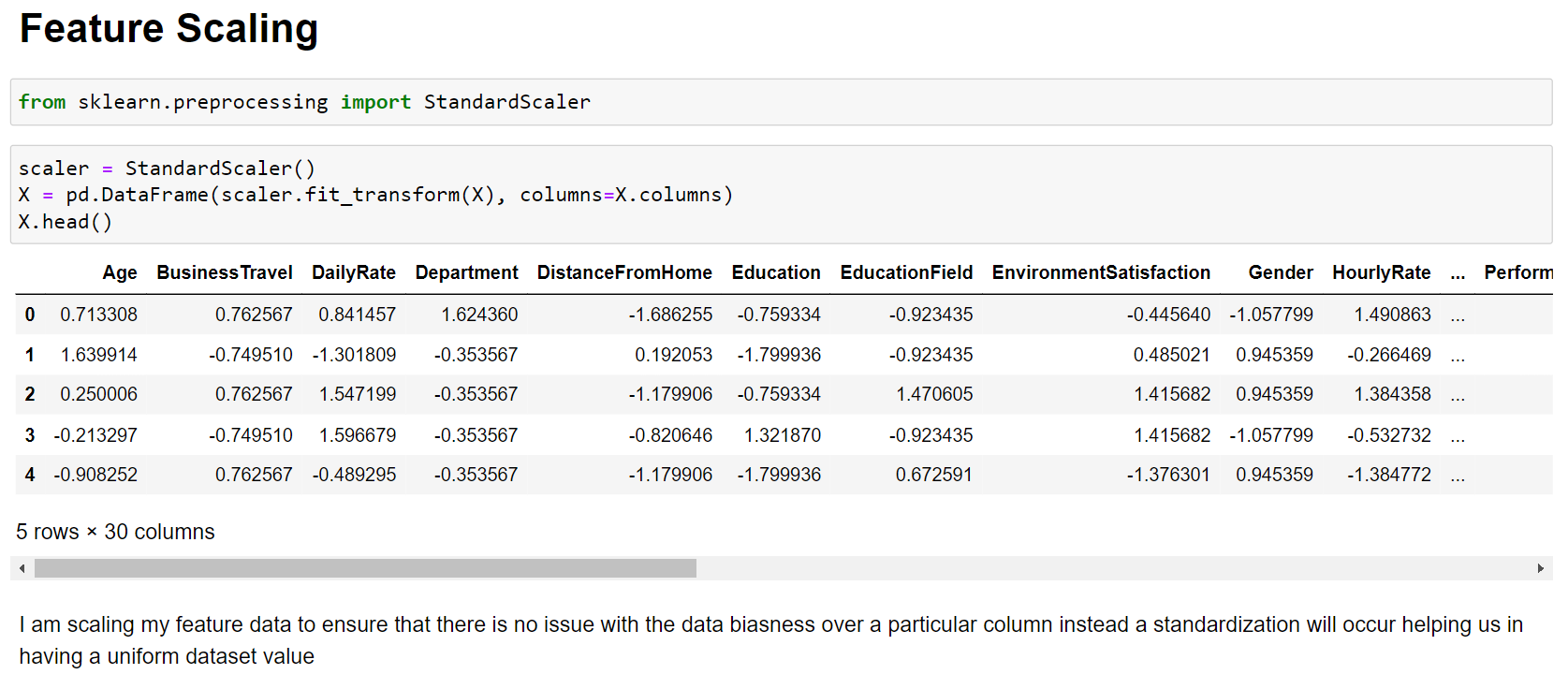


I have applied Log Transformation on our numerical integer datatype columns to ensure that we do not have skewness in our dataset.





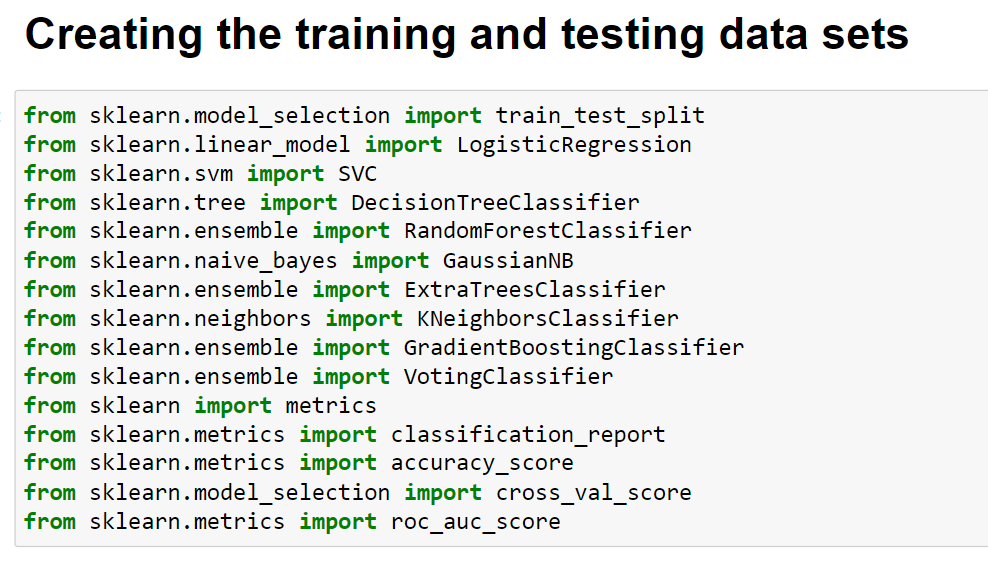
After applying over sampling, we are once again listing the values of our label column to cross verify the updated information. Here we see that we have successfully resolved the class imbalance problem and now all the categories have same data ensuring that the machine learning model does not get biased towards one category.



* **Building Machine Learning Models**

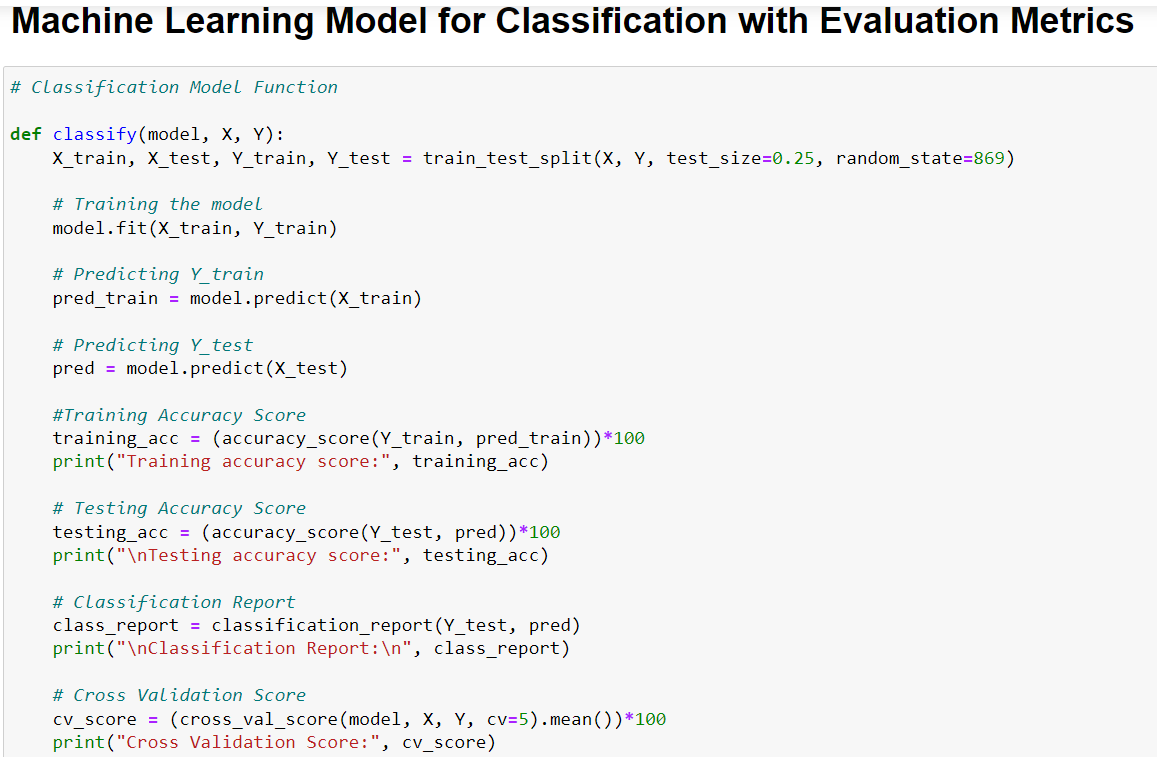
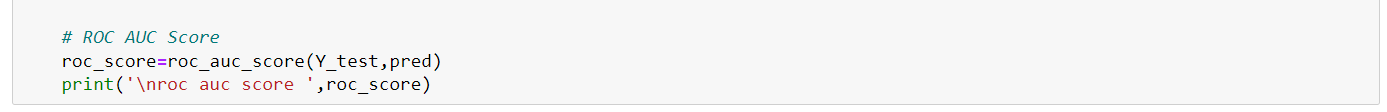
**Applying ML algorithm**

We will use multiple machine learning algorithms to train model and pick the best one.

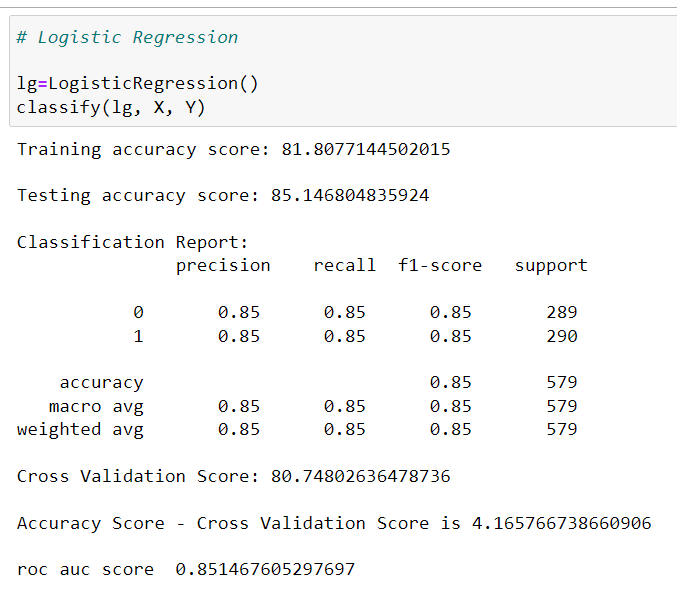


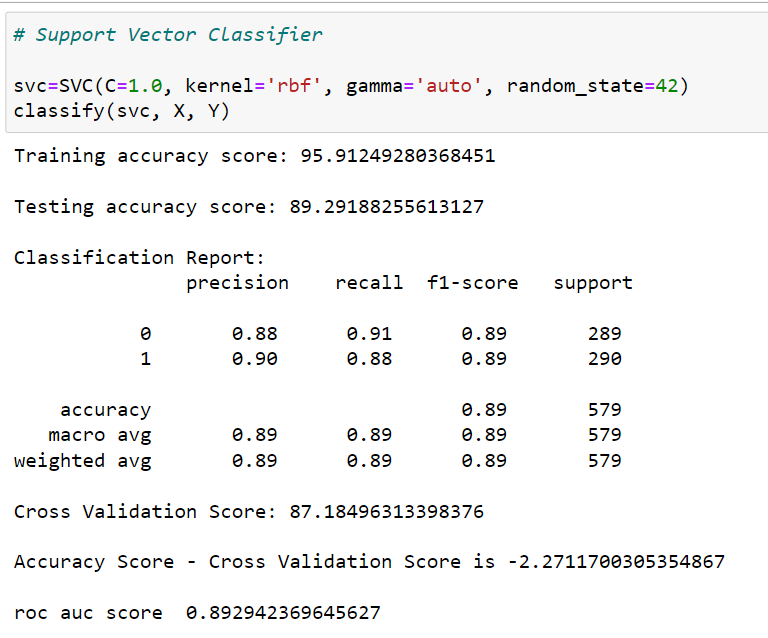


**Applying Multiple ML algorithm**

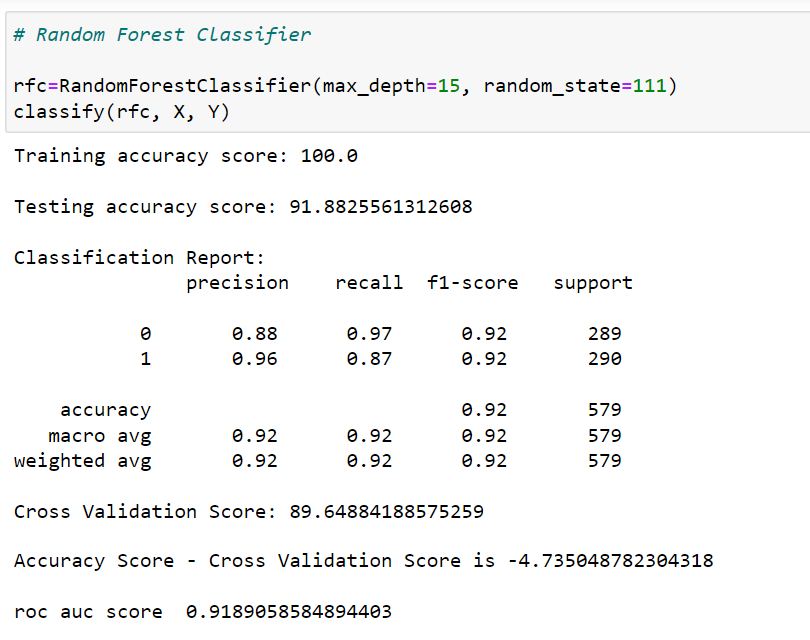
 

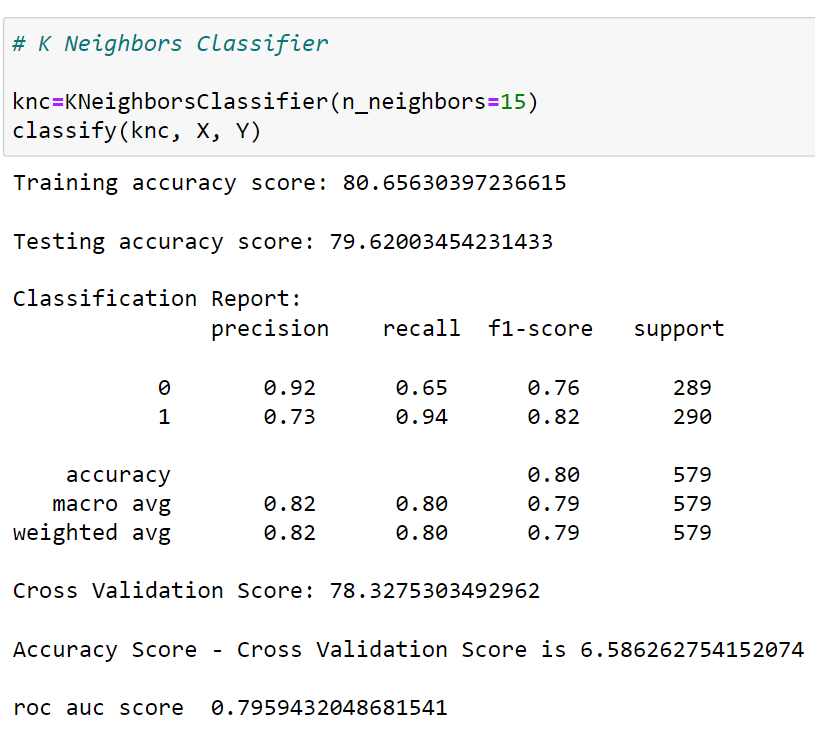
I have defined a class that will perform the train-test split, training of machine learning model, predicting the label value, getting the accuracy score, generating the classification report, getting the cross validation score and the result of difference between the accuracy score and cross validation score for any machine learning model that calls for this function



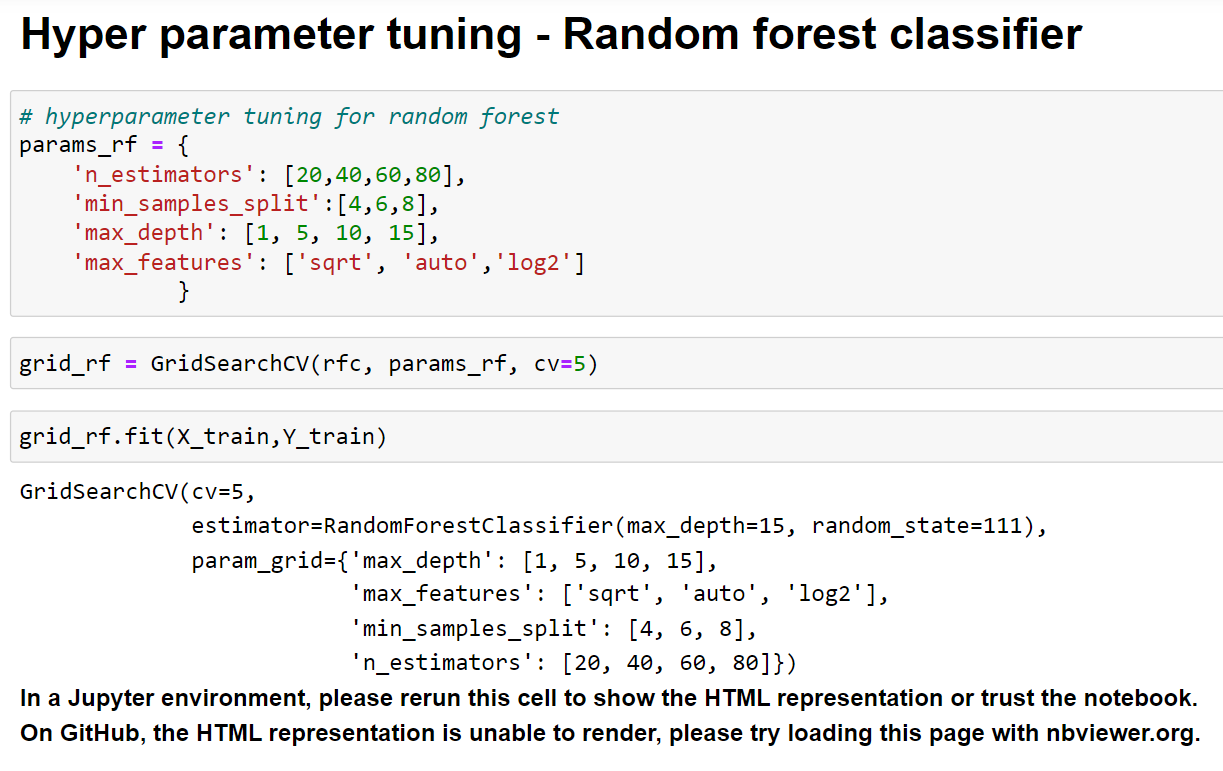






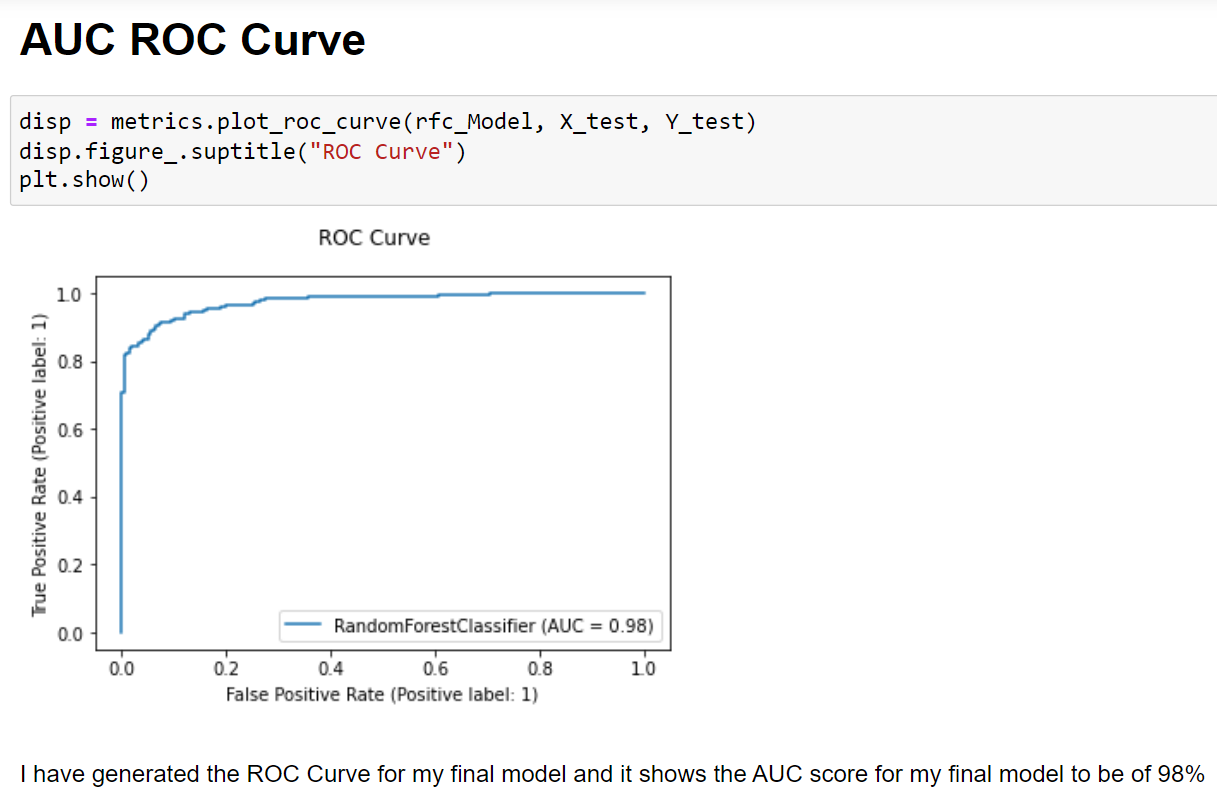


Random Forest is giving better result as compare to other algorithms, so i have decided it to select as best and process further operation like hyper parameter tuning and cross validation



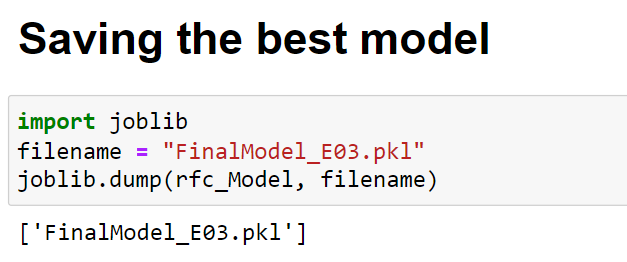


I have used Gridsearch CV to find best parameter of random forest.





**Model saving**



**Conclusion:-**

Dataset was quite clear, there was no missing value. It was mix of categorical and numerical features. We have performed multiple analyses to check that which factor plays important role in attrition. We checked outlier and found that few columns have some extreme value but it is very close to upper whisker and we didn’t try treating them because the ensemble methods will deal with them. I have checked correlated of each features and found that couple of features were correlated so have deleted them.  
As we saw at the initial phase of analysis that data was imbalance, we have corrected that by applying oversampling technique and then Model was trained. Random forest has given best F1 score and has taken it for final model.