***HR Analytics Project- Understanding the Attrition in HR***

In this article, I’ll be going through the process of building a machine learning models to predict the Employees turn over and it will take you through each and every step-in detail and helps you to understand the whole machine learning model building process.



***Introduction***

Employees are the backbone of any organization. Its performance is heavily based on the quality of the employees and retaining them. With employee attrition, organizations are faced with number of challenges:

* Expensive in terms of both money and time to train new employees
* Loss of experienced employees
* Impact on productivity
* Impact on profit

***Step 1: 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 analysing 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.

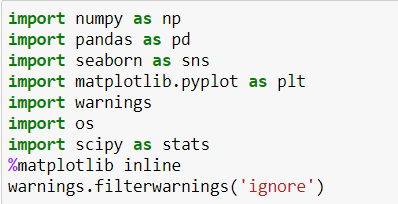
**Attrition affecting Companies**

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.

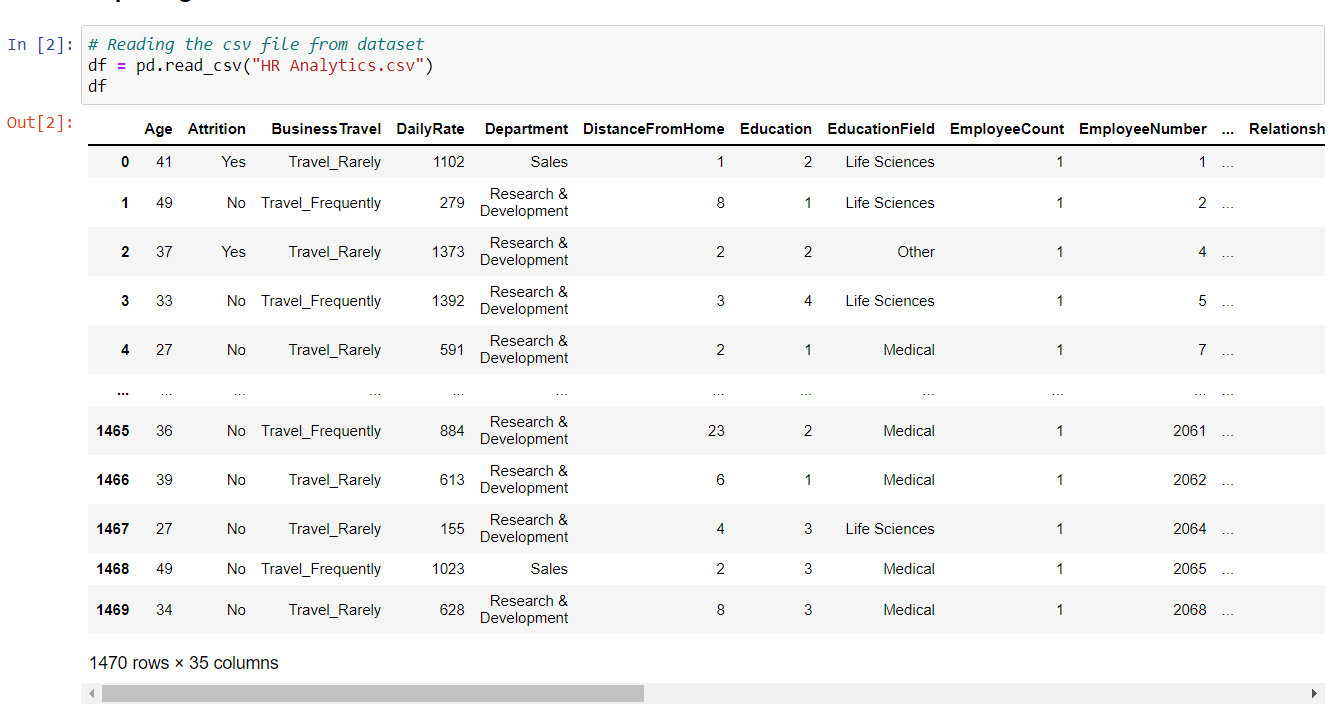
***Step 2: Data Analysis***

The process of cleaning, transforming and extracting data to discover the useful information for business decision making is called data analysis.

**Importing necessary libraries**

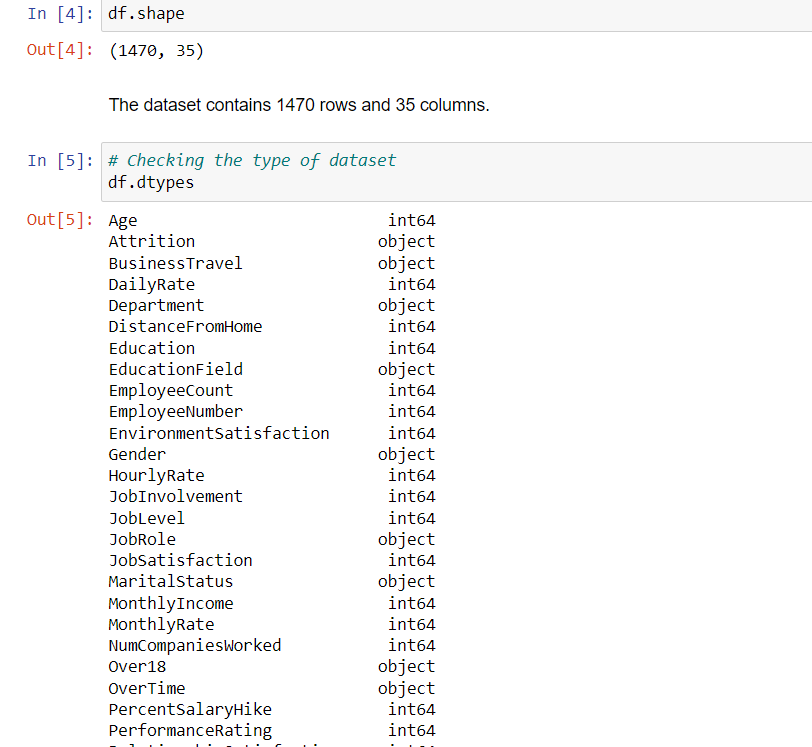


***Importing the Dataset***

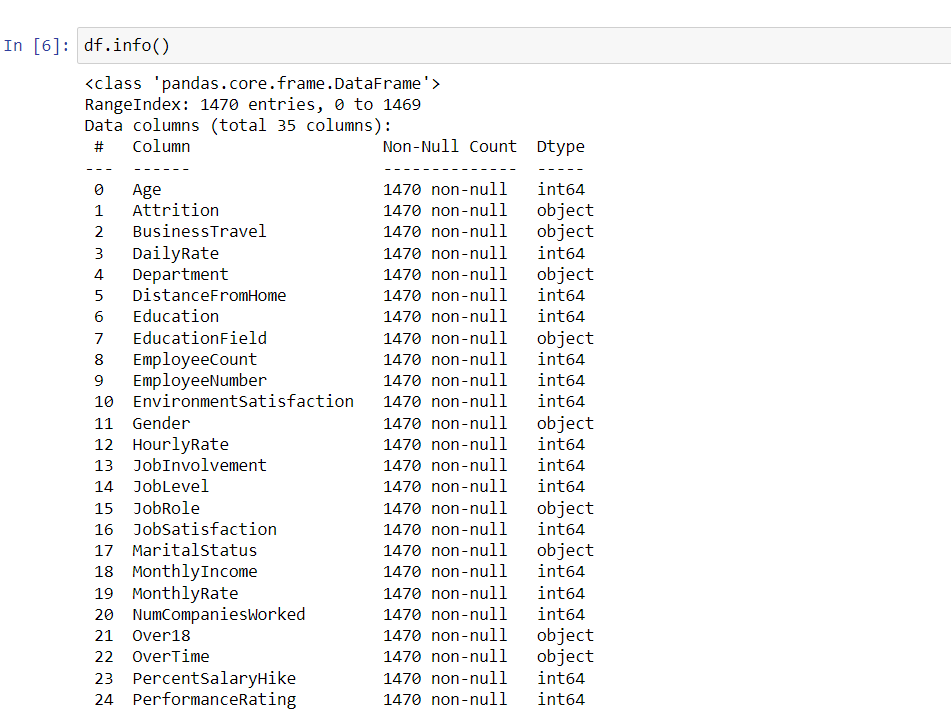
******

In this dataset **"Attrition""** is our target variable which has two classes. So this is a **"Classification type"** problem.

***Exploratory Data Analysis (EDA)***

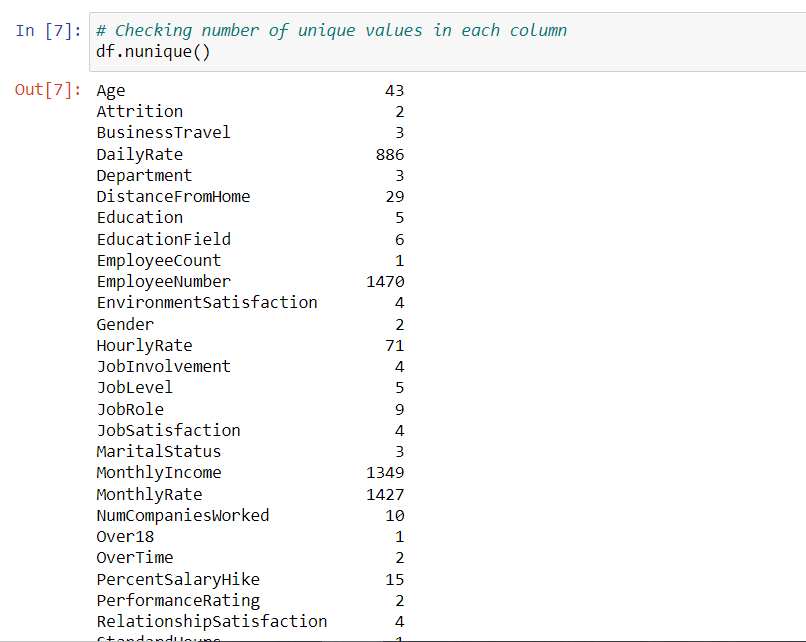


* I run the df.shape which gives the number of rows and columns present in the dataset and our dataset has 1470 rows and 35 columns.
* Then i run the df.dtypes which gives the types of dataset present in the dataset and in our dataset we have integer and objective data types.
* Checking the dataset Information by df.info() :



This gives the information about the dataset which includes indexing type, column type, no-null values and memory usage.

* Checking the unique values in the dataset- df.nunique()



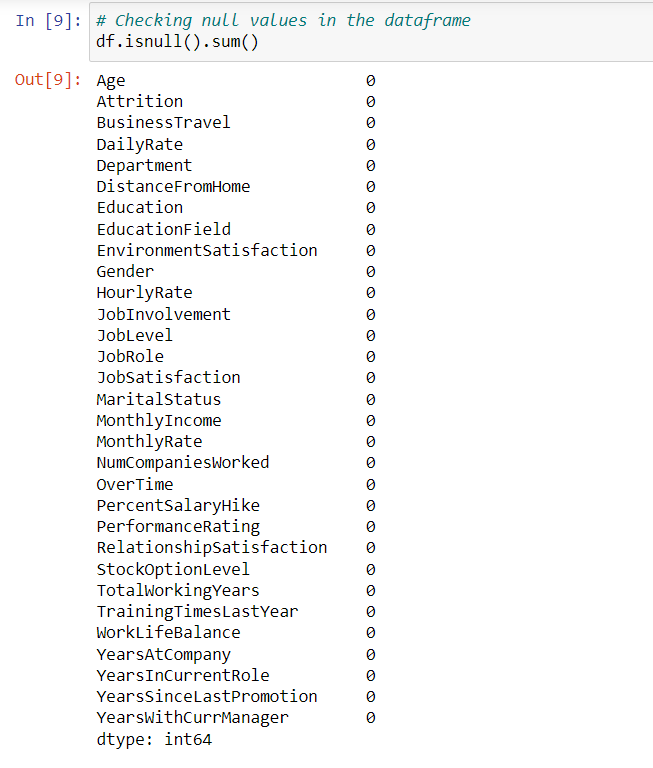
Which shows the number of unique values present in each columns.

Since the columns EmployeeCount, Over18 and StandardHours have only 1 count so they can be dropped since they won't affect our model. Also EmployeeNumber is taken on the basis of unique ID of the employees which does not helps so we can drop this too.

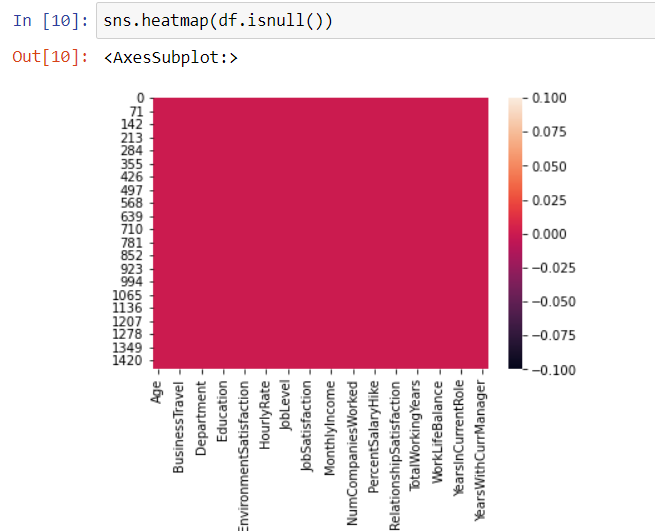


After dropping the few columns, dataset contains 1470 rows and 31 columns.

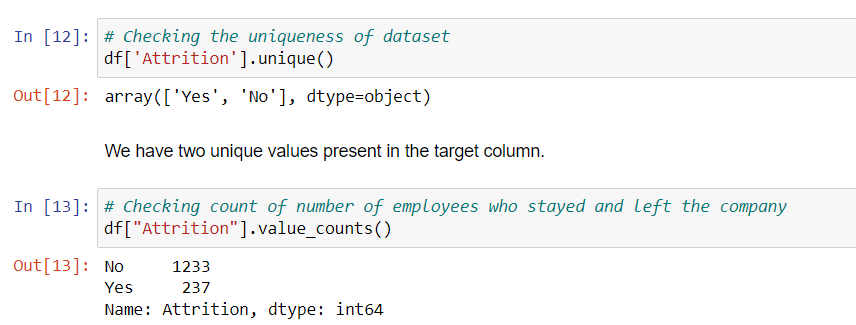
* Let’s check if there any null values- df.isnull.sum()



There are no null values in the dataset and we can observe it in the heat map visualization also.

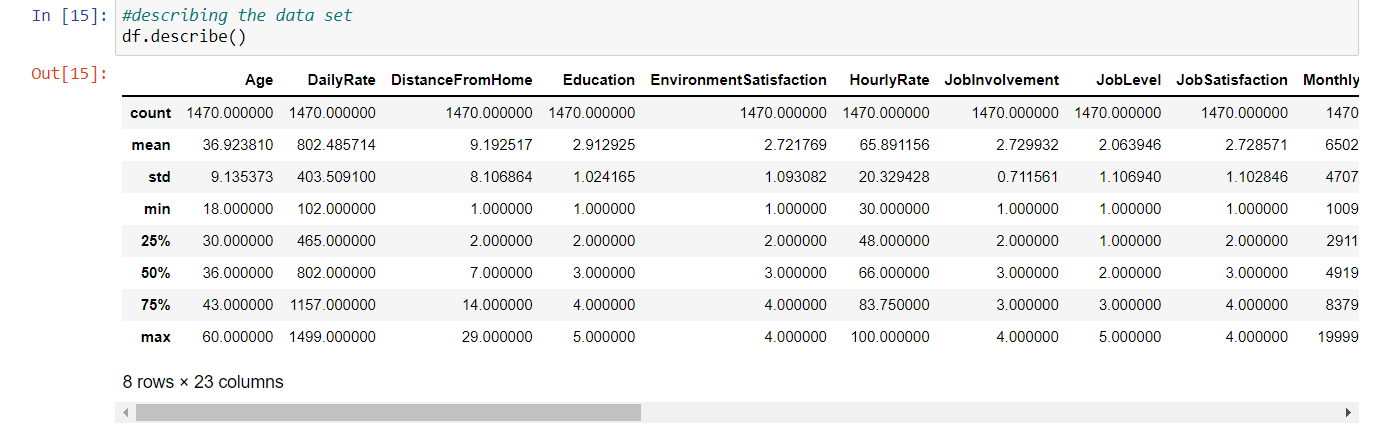


* Let’s check the uniqueness and value counts for target variable column.



We can see here , we have a data imbalance issue in the target variable column. Will balance this later with appropriate method.

* Describing the dataset – df.describe()

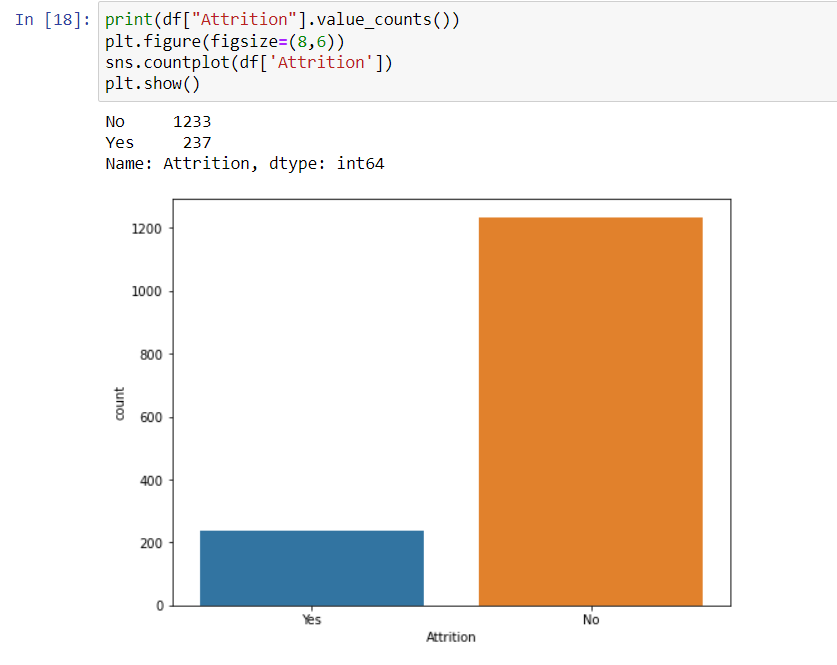


* The counts of all the columns are same which means there are no missing values present in the data.
* The mean is more than the median(50%) in most of the columns which means they are skewed to right. The min age of the employee is 18 and max is 60 and most of the employees are in between 36.
* In few columns the median(50%) is more than the mean which means they are skews to left.
* Some of the columns have huge difference in mean and the standard deviation.

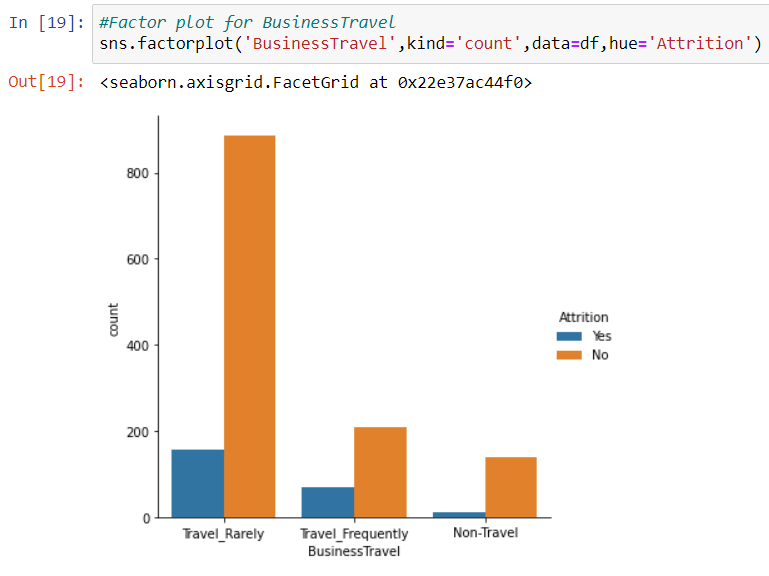
***Data Visualization***

Univariate Analysis

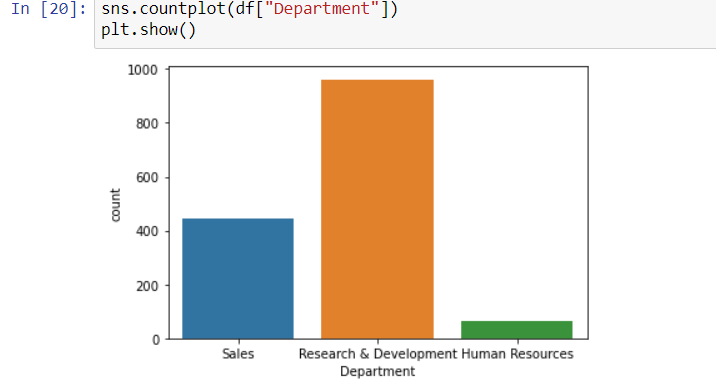
Plotting categorical columns



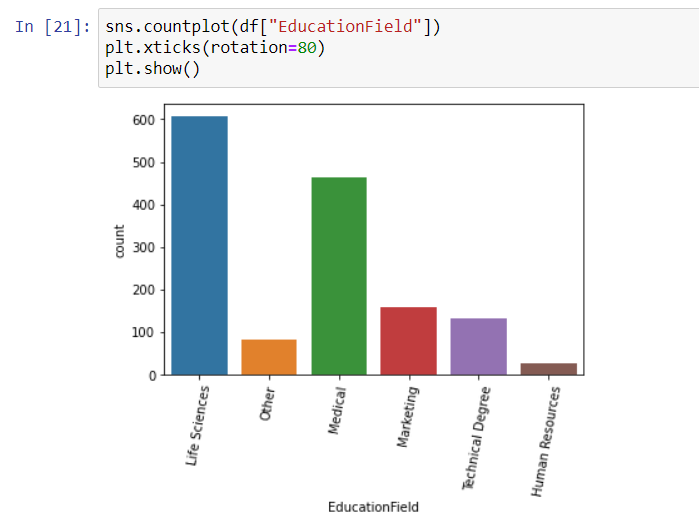
On studying the Attrition column, we can see that less employees facing the attrition and most of the people are not.



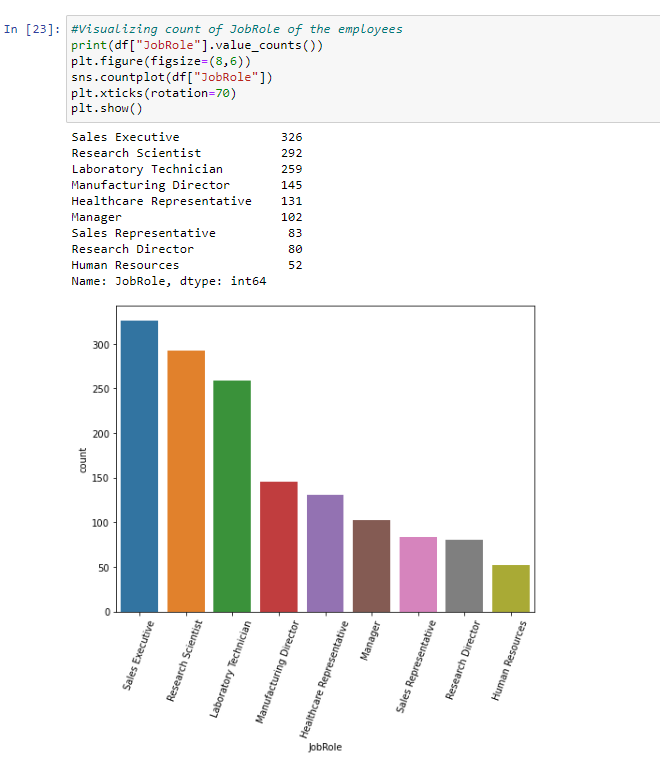
The above plot shows most of the employees travel rarely and few employees travel frequently.



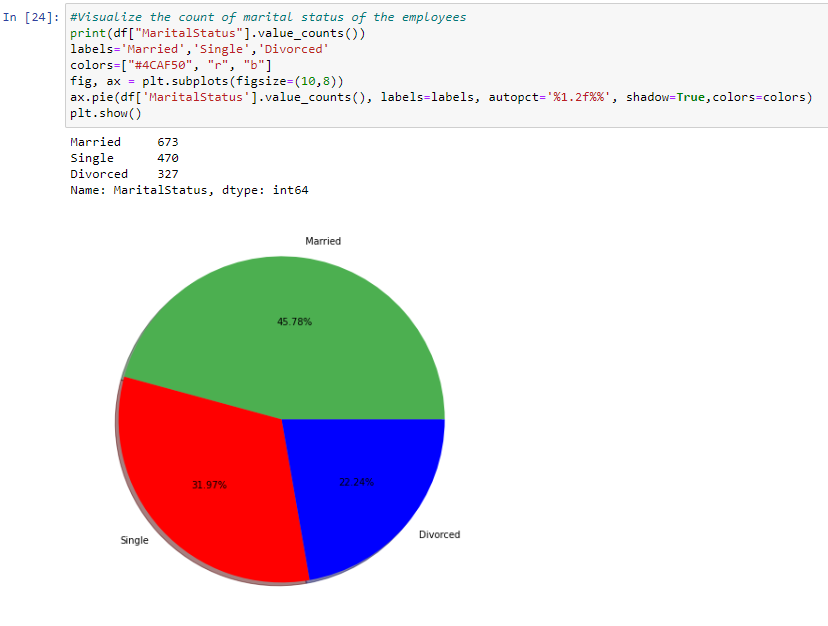
Shows the employees works in different department, more employees are in the Research & Development and followed by Sales and Human Resources.



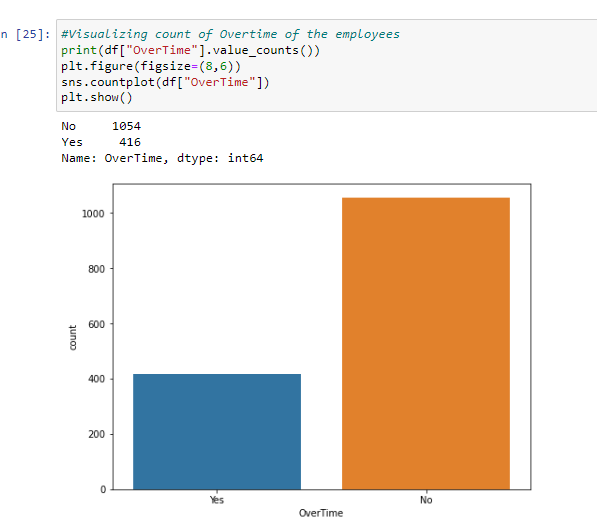
In the above count plot it shows that the employees educational background and most of the employees are from Life Science educational background.



In the organization there are several job roles exist, count of employees working in Sales Executive are more.



The above graph shows the marital status of the employees in the organization. Here percentage of married employees are more (45.78%)



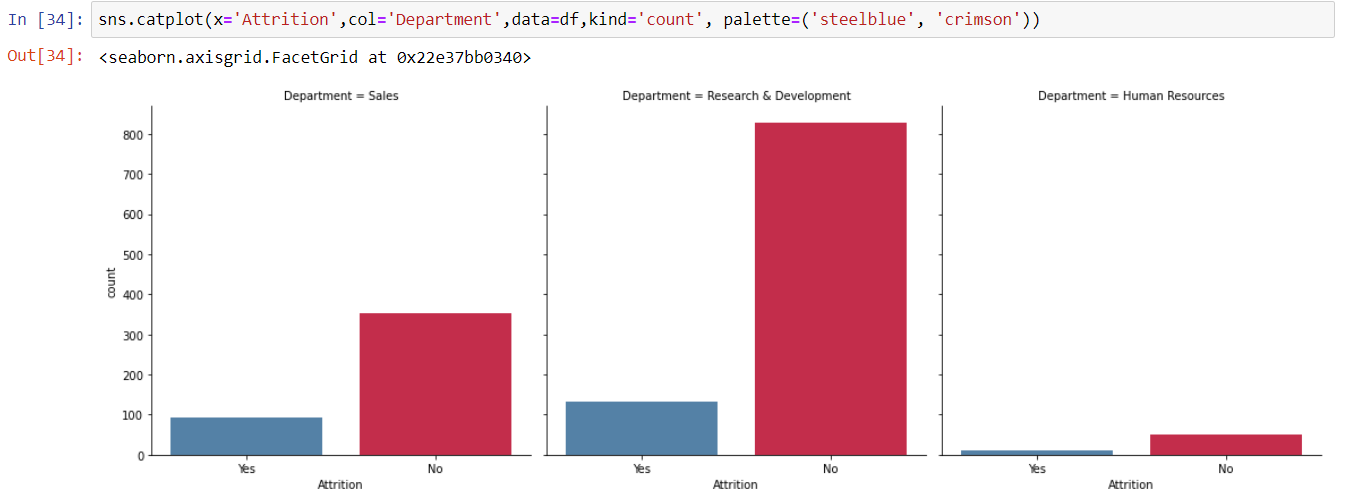
There are some employees working extra hours than the assigned hours.

* Plotting numerical columns.

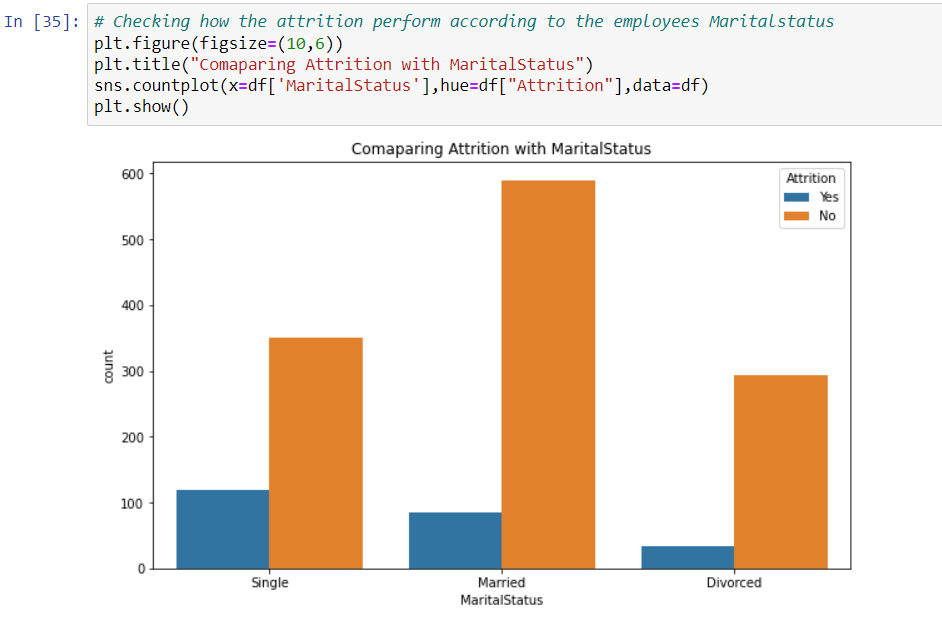


By plotting the numerical column, we can observe that except Age, DailyRate, HourlyRate and MonthlyRate all the other columns are skewed.

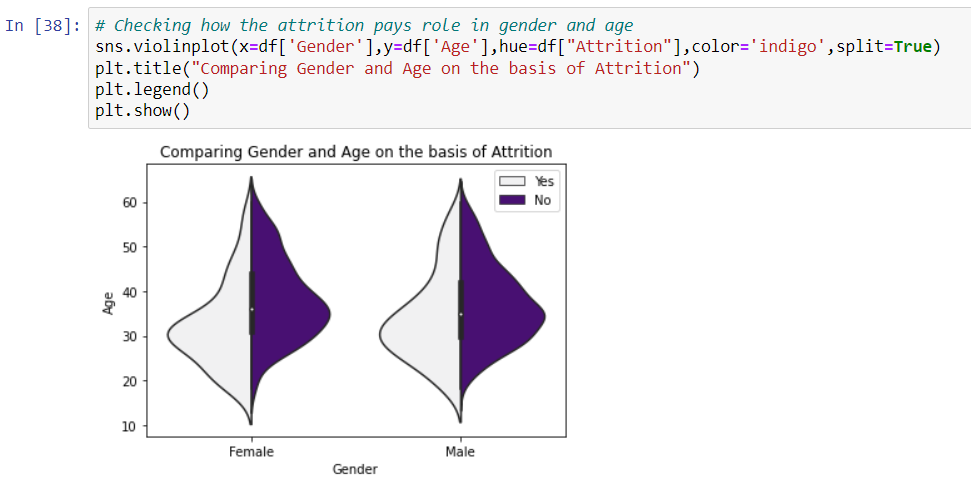
***Bivariate Analysis***



As department wise, in HR department employees are not really interested to leave the organization but the more employees from the R&D followed by the Sales department are ready to quit their job in the organization.



More number of employees, who are unmarried are facing the attrition than the married and divorced employees.

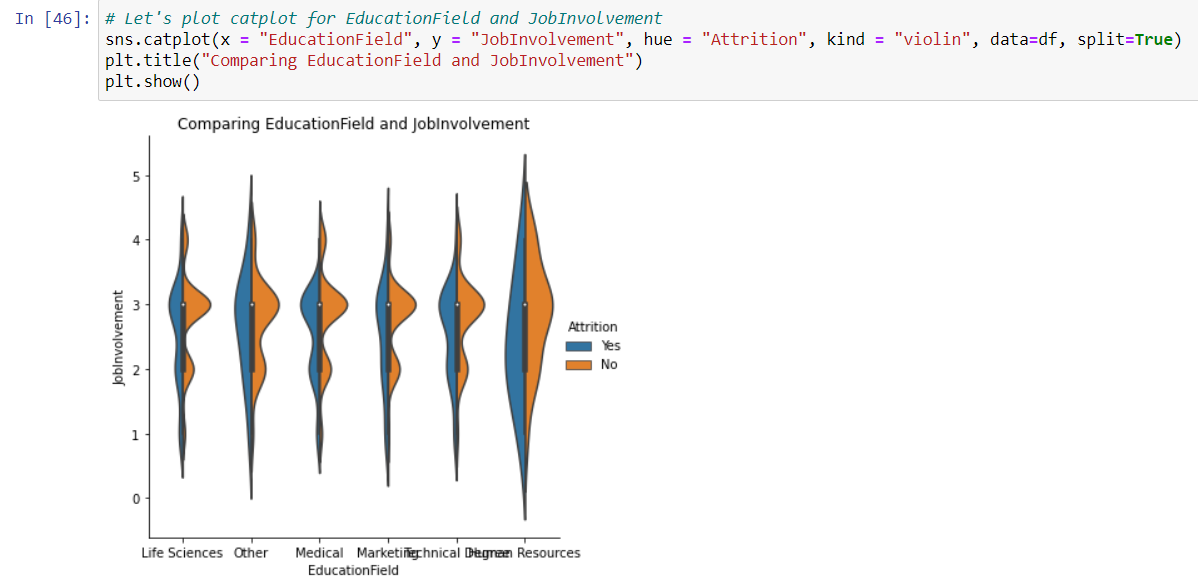


By running the above code, it shows that female and male employees with respect to their age, who are all ready to leave the organization.

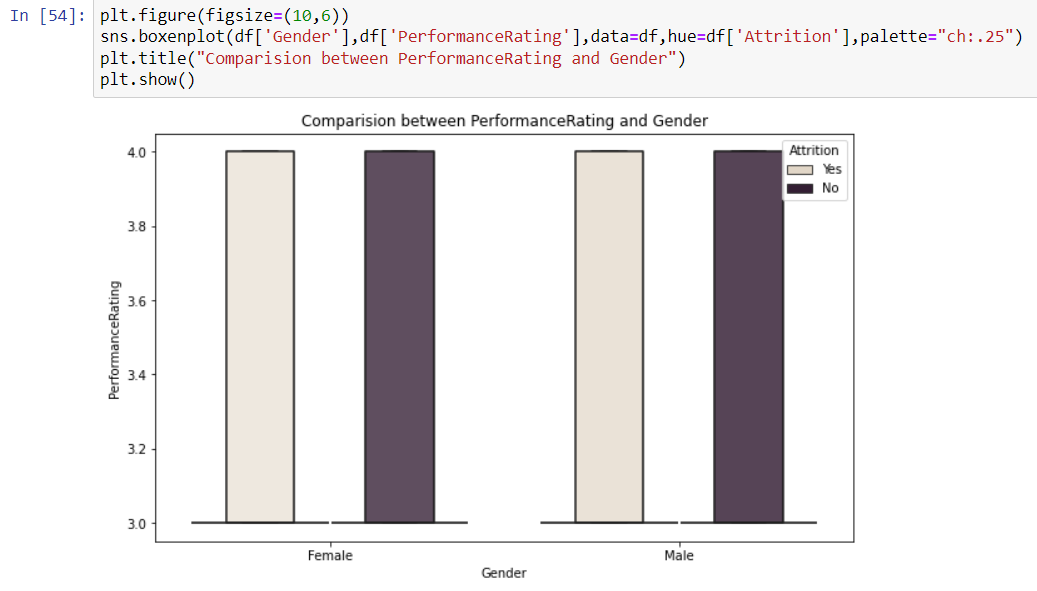
Here the female employees age between 25-35yrs and male employees age between 25-40yrs are interested in attrition.



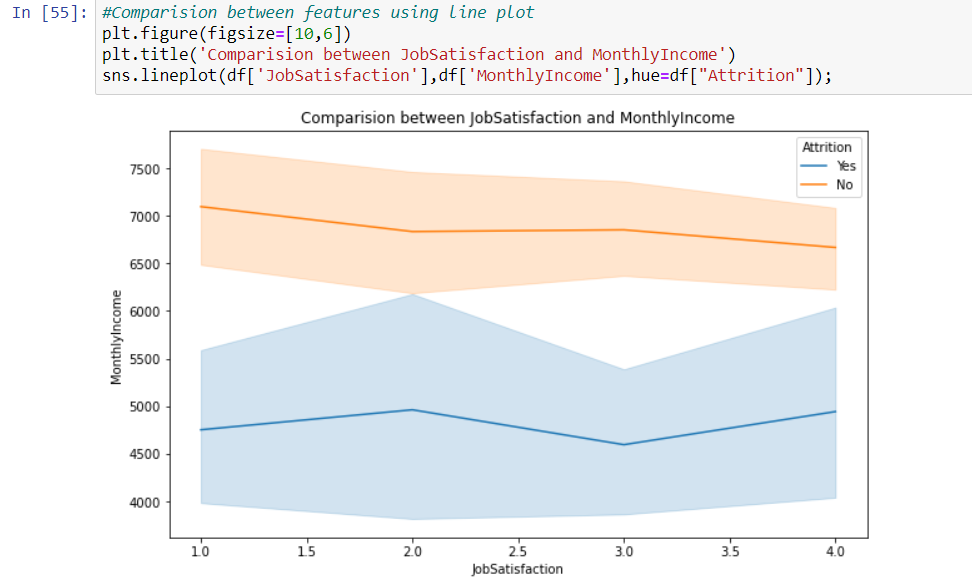
he employees who’s reside is near to the organization is high in number and the attrition levels are lower for distance less than 15. As the distance from the home increases, attrition level also increases.



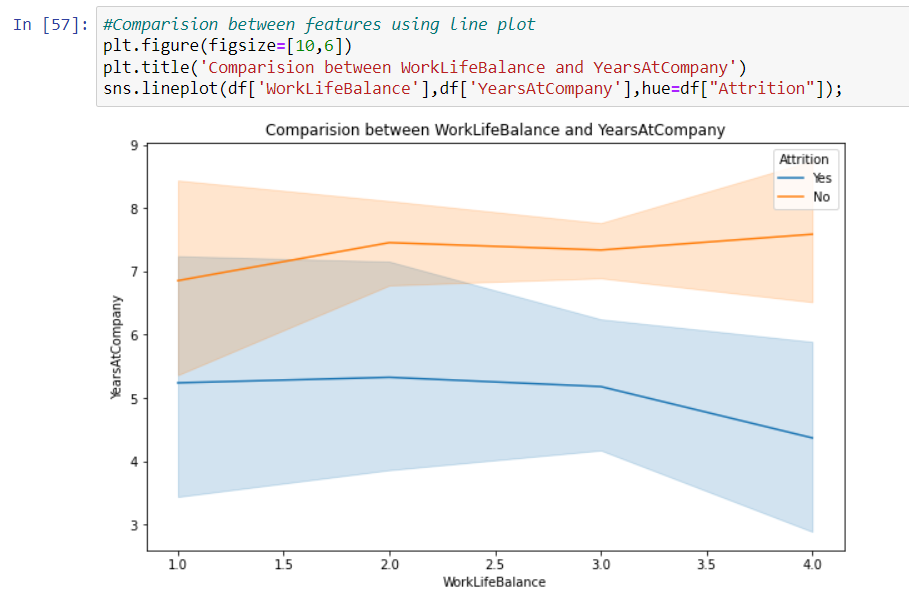
From the above graph it clears that irrespective of education field, all the employees facing the Attrition.



As we can see there is no significant difference in Performance Rating between male and female and the attrition also affecting them equally.



For monthly income greater than 6500 has no attrition whereas for less than 6500 has attrition irrespective of job satisfaction.



Attrition is high for new employees compare to the experienced one.

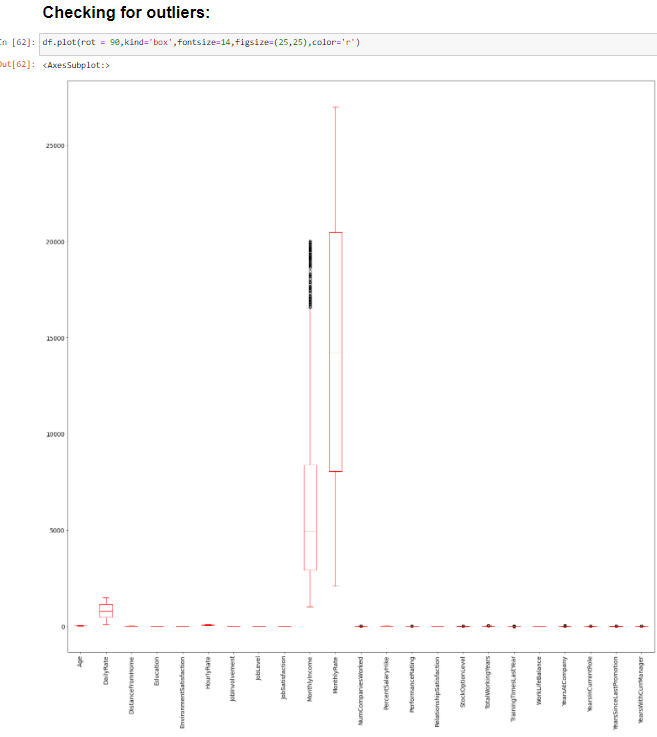
In a similar manner we can plot the graphs for other columns too and we can make a good visualization by them.

***Step 3: Exploratory Data Analysis (EDA)***

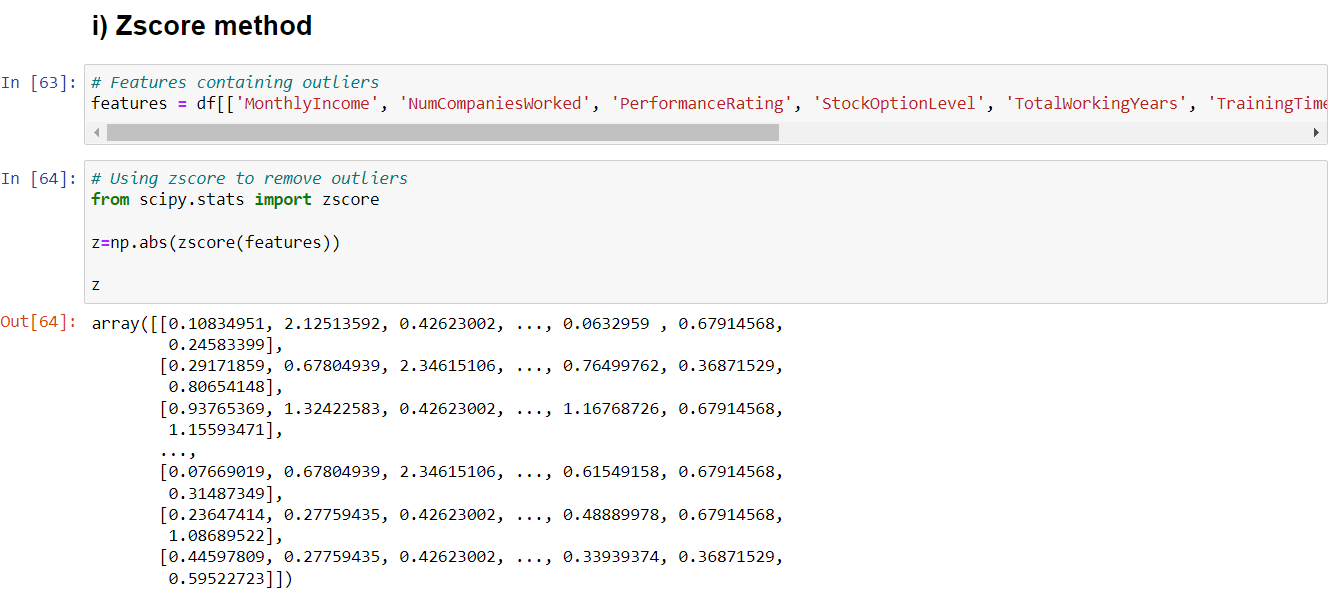
There are some outliers and skewness present In some columns, I used the boxplot to identify the outliers and found the outliers in MonthlyIncome, NumCompaniesWorked, performanceRating, StockOptionLevel, TotalWorkingYears, TrainingTimesLastYear, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion and YearsWithCurrManager columns.

And also skewness present in DistanceFromHome, JobInvolvement, JobLevel, MonthlyIncome, NumCompaniesWorked, PercentSalaryHike, PerformanceRating, StockOptionLevel, TotalWorkingYears, TrainingTimesLastYear, WorkLifeBalance, YearsAtCompany, YearsInCurrentRole, YearsWithCurrManager.

* Checking for Outliers:

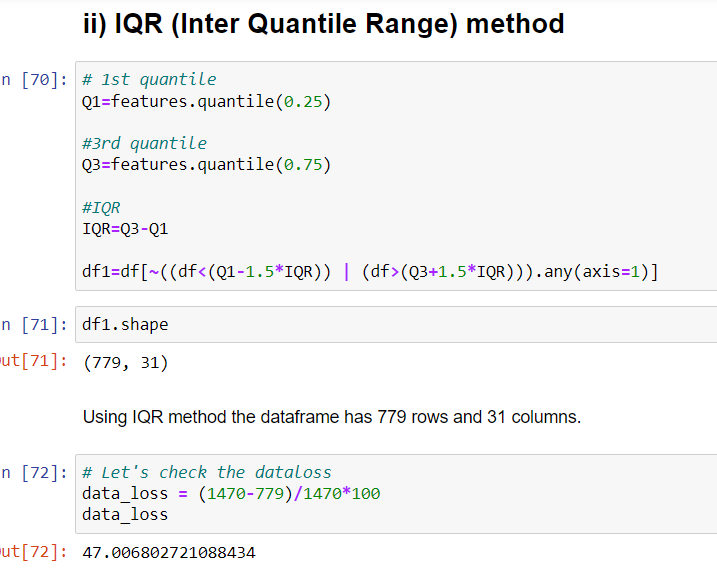


* Removing the Outliers by Zscore method:



By using zscore method outliers has been removed and the data loss from this method is 5.64% so let me check with IQR method then I will decide which method should I consider.

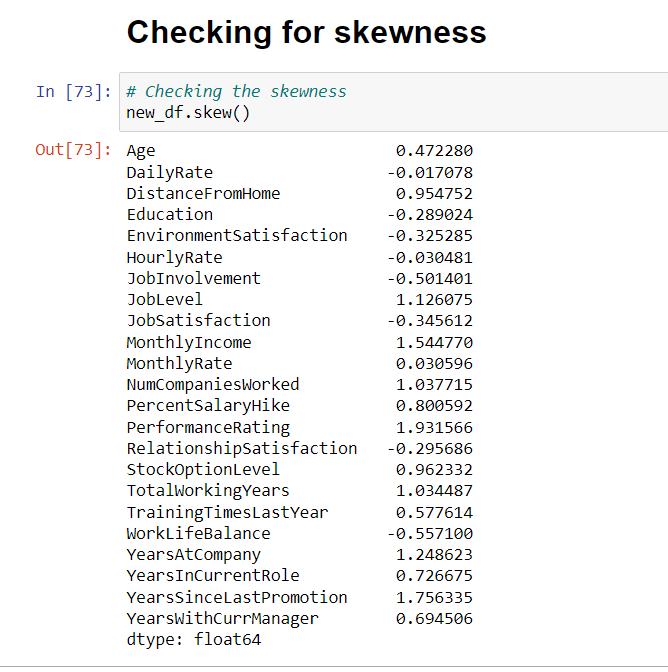
* Removing the Outliers by IQR method:



By using IQR method outliers has been removed and the data loss is 47% which is very high, so I’m considering the zscore method.

* Checking for skewness:

If the data is not distributed normally on the central value is known as skewness.



# Skewness present in many columns, Presence of skewness more than +0.5 and -0.5 is not acceptable as it will impact on our accuracy and I’m removing the skewness by using yeo-Johnson method.

# 

# Label encoding: Use to convert Labels into numerical form so machine learning algorithms can decide in a better way how those labels must be operated.

# 

# Here I have converted categorical column into numerical form using Label encoder.

# *Correlation between the target variable and independent variables using HEAT map*

# Correlation is use to find the relation between two or more variables.

# 

# The heatmap contains both positive and negative correlation.

# OverTime is positively correlated with the label.

# YearsAtCompany and YearsInCurrentRole are highly correlated each other.

# Job level and Monthly income are highly correlated each other.

# *Step 4: Pre-Processing Pipeline*

# Separating the feature and label into x and y :

# 

# I have separated feature and label into x and y. And there is a data imbalance issue in target variable. Let’s use SMOTE to overcome from this.

# 

## Since the skewness of the data is in the acceptable range and the data is also normally distributed in the columns, in such case we can make use of Standard Scaler method else we can make use of Min Max scaler method.

# *Feature Scaling using Standard Scalarization*

# 

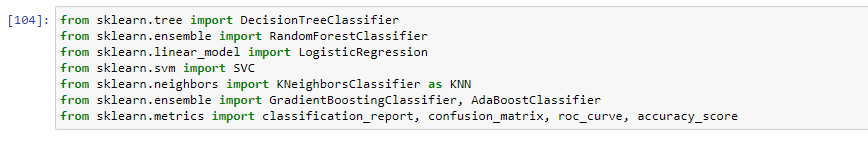
## Standard Scaler helps to get standardized distribution, which makes mean = 0 and scales the data to unit variance. It helps in improving our model accuracy also solve the issue of data biasness.

***Step 5: Building Machine Learning Models***

Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed.

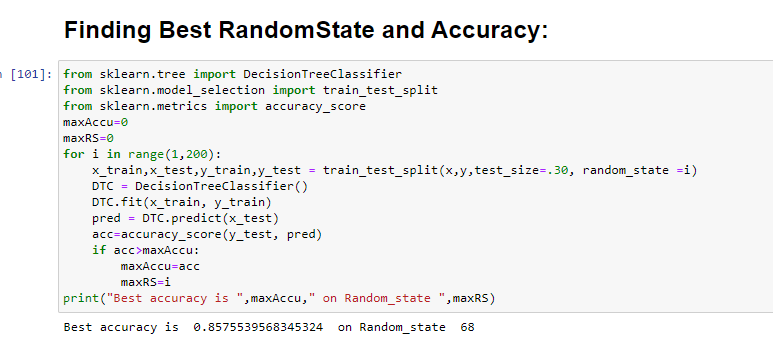
-Arthur Samuel

The main goal in this step is to develop a benchmark model serves us as a baseline, upon which we will measure the performance of a better and more tuned algorithm. We are using different Classification Technique and comparing them to see which algorithm is giving better performance.

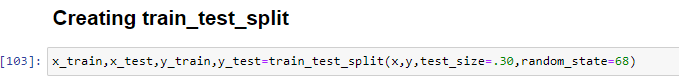


All the required machine learning algorithms are imported from sklearn library.

In this project I have used 7 different algorithms. The model which gives the best performance amongst them, we will be using that as best model for prediction. Before creating the model, let us check the best random state and accuracy using any of the DecisionTreeClassifier algorithm.

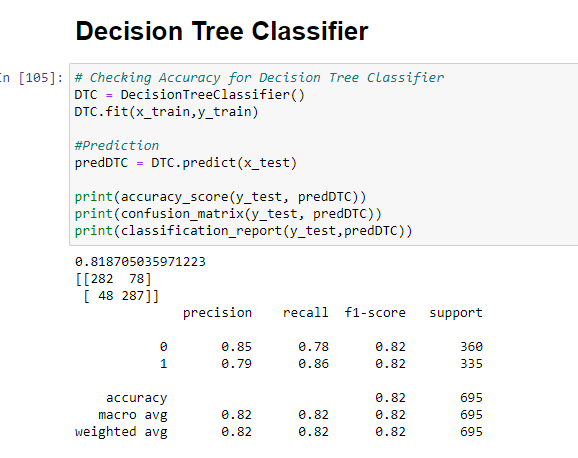


After we have found the value for best random state, we proceeded with the train test split function to create new training and testing datasets and fit them into the models to find our ideal models.



**DecisionTreeClassifier**

Decision Tree are versatile Machine learning algorithms that can perform both classification and regression tasks. They are powerful algorithms, capable of fitting complex datasets.

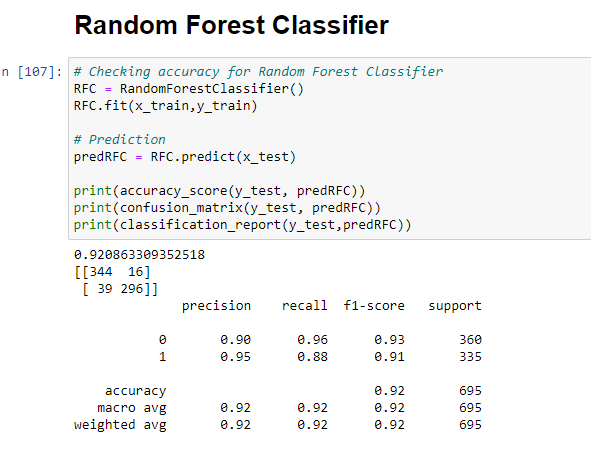




Decision Tree Classifier giving 81.87% accuracy and also we can observe the confusion matrix also known as an error matrix, is a specific table layout that allows visualization of the performance of an algorithm. This gives the values like True Positive, False Positive, True Negative and False Negatives.

**Random Forest Classifier**

Random Forest is an ensemble of Decision Tree and capable of performing both classification and regression task using multiple decision tree and a statical technique called “Bagging”.

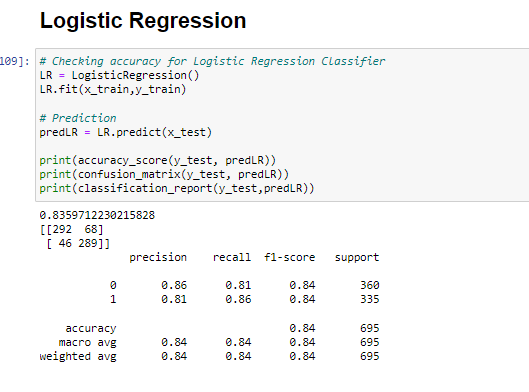


Random Forest Classifier giving the 92.08% of accuracy.



**Logistic Regression**

Logistic regression commonly used to estimate the probability that an instance belongs to a particular class.

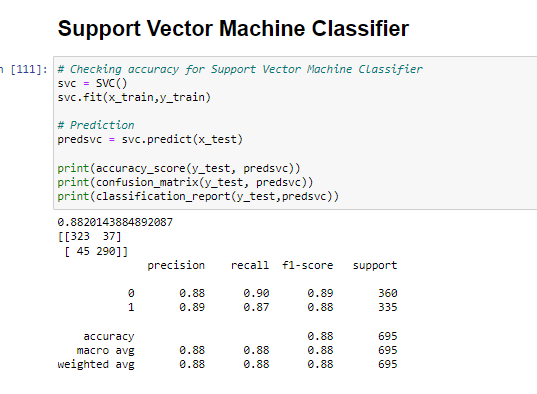


Logistic Regression giving the 83.59% accuracy.



**Support Vector Machine Classifier**

Support vector machine learning is capable of performing linear or non linear classification, regression and even outlier detection. It is one of the most popular models in machine learning and are particularly well suited for classification of complex small or medium sized datasets.

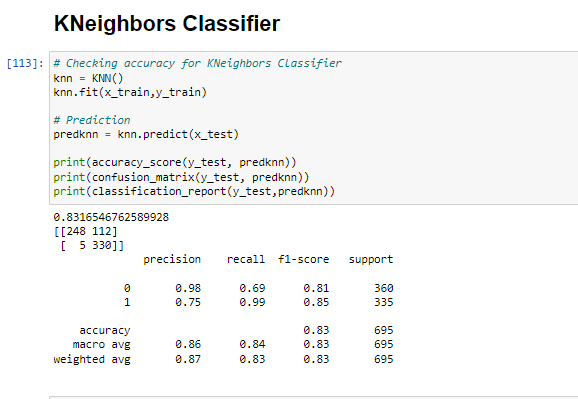


SVM giving 88.20% accuracy and the confusion matrix for SVM is



**KNeighbours Classifier**

K Nearest Neighbor algorithm falls under supervised learning and it is used most commonly in classification.

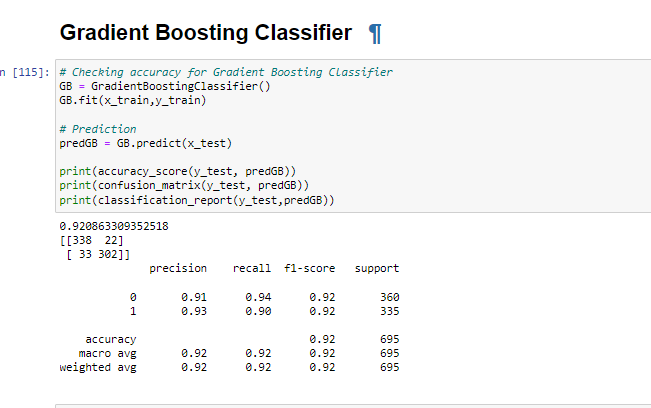


KNN is giving 83.16% of accuracy.



**Gradient Boosting Classifier**

Gradient boosting works by sequentially adding predictors to an ensemble, each one correcting its predecessor.

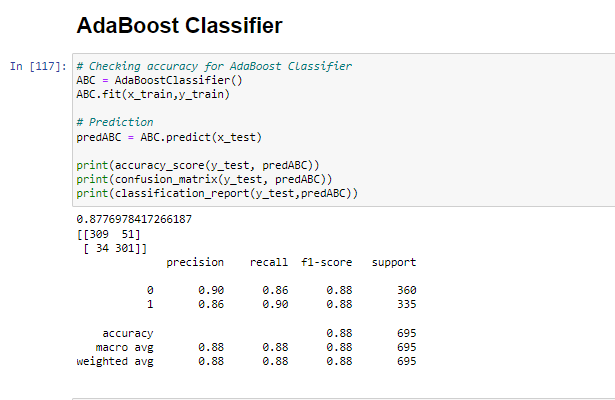


It is giving 92.08% accuracy.



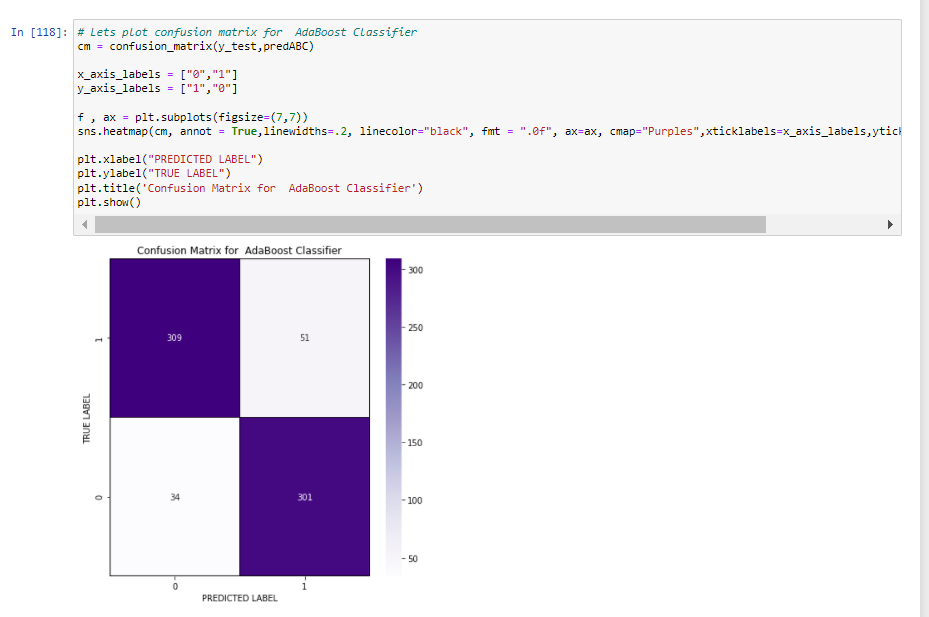
**AdaBoost Classifier**

One way for a new predictor to correct its predecessor is to pay a bit more attention to the training instances that the predecessor underfitted. This results in new predictors focusing more and more on the hard cases. This is the technique used by AdaBoost.



AdaBoosting classifier giving 87.76% accuracy.

And the confusion matrix AdaBoost classifier is



***Checking the Cross Validation Score***



**Difference between Accuracy and Cross validation Score**

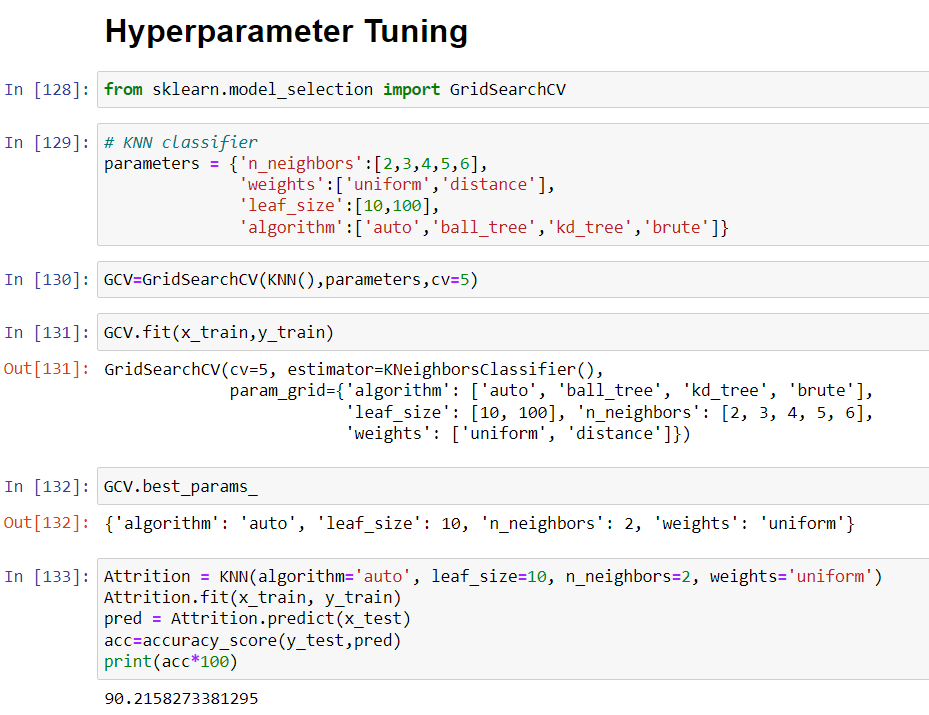
|  |  |  |  |
| --- | --- | --- | --- |
| **Models** | **Accuracy** | **CV Score** | **Difference** |
| DecisionTreeClasssifier | 81.87 | 80.66 | 1.21 |
| Random Forest Classifier | 92.08 | 89.64 | 2.44 |
| Logistic Regression | 83.59 | 81.22 | 2.37 |
| SVM Classifier | 88.20 | 85.97 | 2.23 |
| KNN Classifier | 83.16 | 83.03 | 0.13 |
| Gradient Boosting Classifier | 92.08 | 87.44 | 4.64 |
| AdaBoost Classifier | 87.76 | 83.90 | 3.86 |

The model KNN Classifier giving the very less difference compared to other models.

Since KNN Classifier is giving best Accuracy and CV score difference, I choose KNN Classifier as best fitting model. Let’s check whether we can increase the Accuracy score by using Hyper parameter tuning.

***Hyper Parameter Tuning (Using GridSearchCV)***

In GridSearchCV approach, machine learning is evaluated for a range of hyperparameter values. This approach is called “GridSearchCV” , because it searches for best set of hyperparameters from a grid of hyperparameters values.

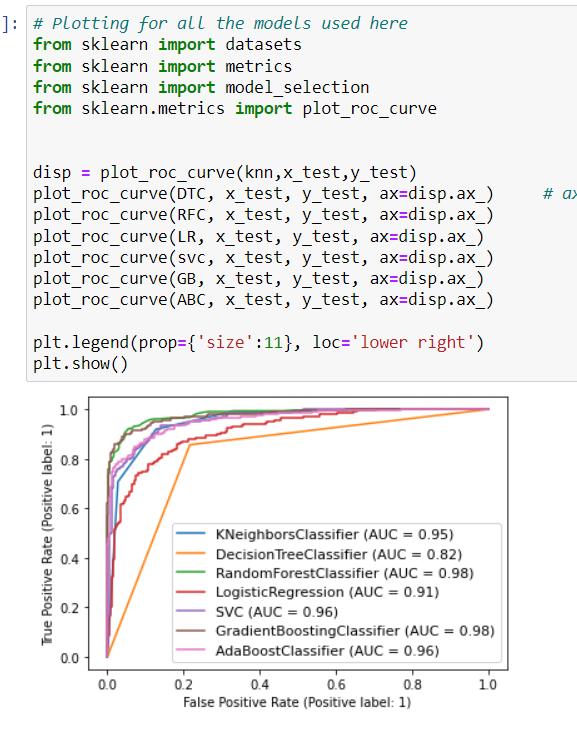


After tuning the best fitting model (KNN Classifier) using GridSearchCV, the accuracy of the model increased to 90.21%**.**

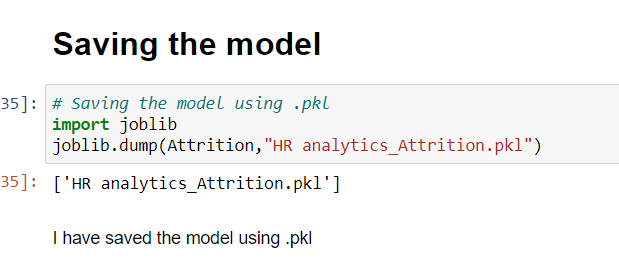
**AUC-ROC Curve:**

It is a graph that shows the performance of a classification model at all possible thresholds. The curve is plotted between two parameters

* True Positive Rate
* False Positive Rate



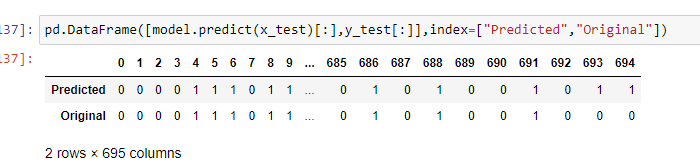
We have built the model and performed the hyper parameter tuning, now we will save the model and load the model to compare the actual and predicted values.





We have loaded the saved model to get predicted values.

Let’s compare the predicted and actual values.



We can see from the above observation, Predicted and the Original values are matching each other, which means model performance is good.

***Step 6: Conclusion Remarks***

In this project we have gone through the feature engineering which is the most important thing to get the better performance models, we have removed the outliers, skewness and also handled the categorical columns by encoding the data, scaled the data, handled the data imbalance and at last, we built the different classification models to predict the attrition and perform the hyper parameter tuning to improve the model accuracy by using different parameters.

With the help of above techniques, our model is able to give the good performance and it can help in the organization to understand the attrition and overcome from the issue.