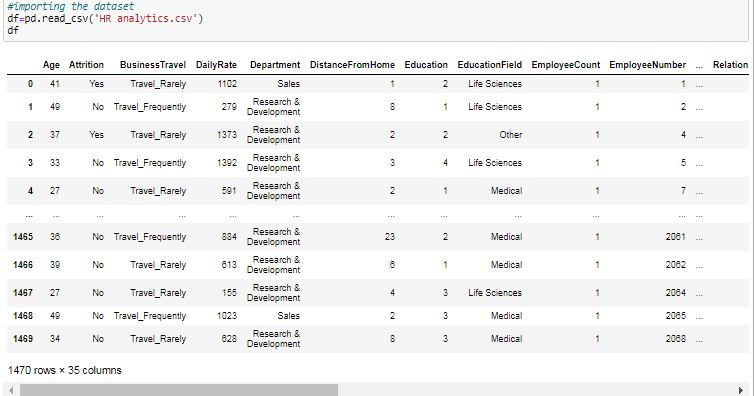
**HR Analytics Project- Understanding the Attrition in HR**

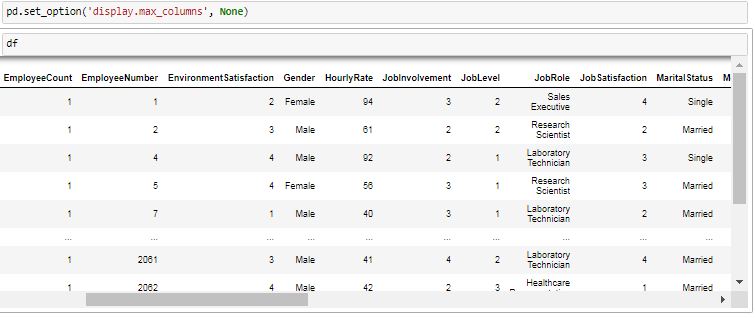
* Human resource analytics (HR analytics) is an area in the field of analytics that refers to applying analytic processes to the human resource department of an organization in the hope of improving employee performance and therefore getting a better return on investment. HR analytics does not just deal 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 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.
* We have to undrstand the data annd find out whether the employee is attrited or not.
* First we have to import every library which we are going to use.



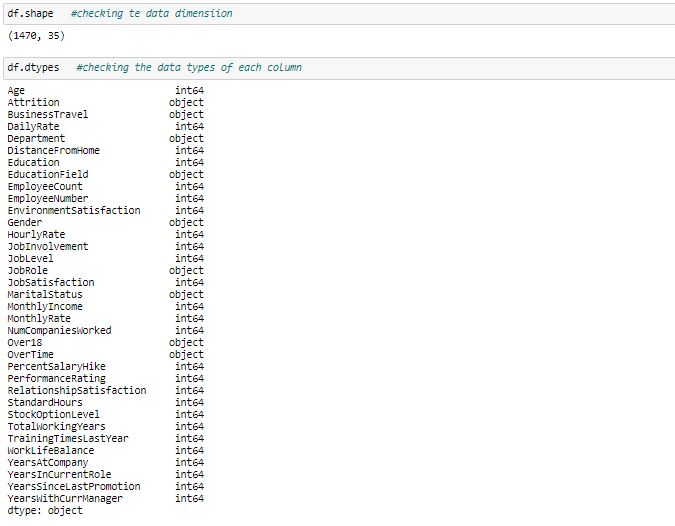
* Now we have to import the dataset.



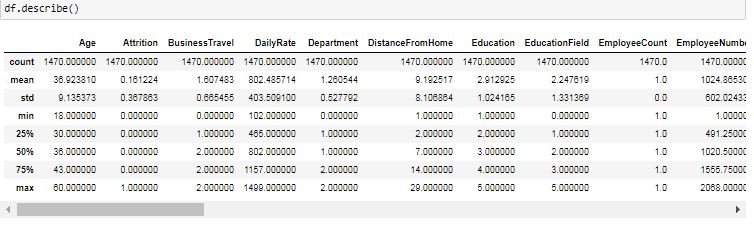
* Now we have to make a function to see all the columns of the dataset. There is a function in pandas ‘set\_option’ to display every column of the dataset.



* Now are checking the dimensions of the dataset and the datatypes of each column of the datatypes.

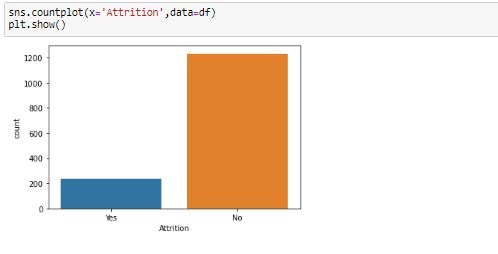


* Now we are checking the statistical description of the dataset by using ‘describe()’ function.



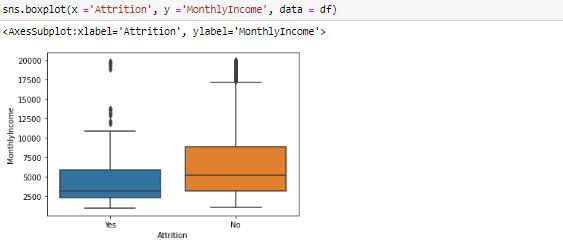
**Visualization of the data:-**

1. In this line of code we are using the number of attrited employee in the company by using countplot in the seaborne library.



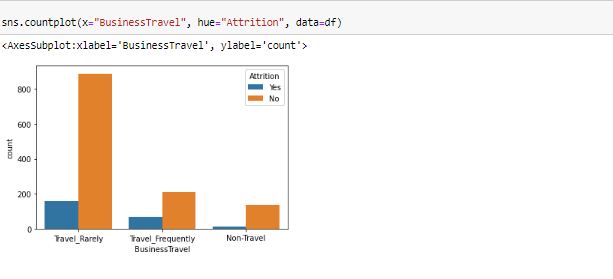
* As we see from above graph that number of not attrited people is more than no of attrited people.

1. In this line of code we are plotting a boxplot of ‘Attrition’ on x-axis and ‘Monthly Income ‘ on y-axis.



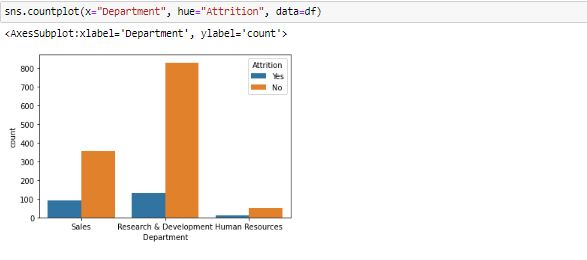
* As we see from the above graph that with monthly income more prople are not attrited as compare to people that are attrited.

1. In this line code we are plotting a countplot of ‘Business Travel’ on x-axis and taking the ‘Attrition’ in consideration.



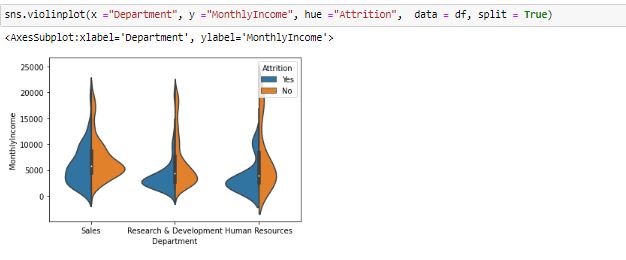
* As we see from the above graph that more no of people are not attrited if they are travel rarely, travel frequently or not travelling.

1. In this line of code we are plotting a countplot of ‘Department’ on x-axis and taking the ‘Attrition in consideration.



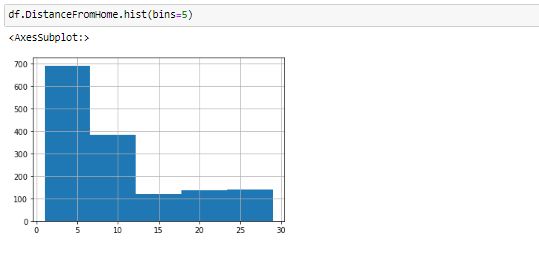
* We can observe from the above graph that in research and development department majority of the people are not attrited.
* In sales department also majority of the people are not attrited.
* In Human resource department also more no of people are not attrited.

1. In this line of code we are plotting violinplot of ‘Department’ on x-axis and ‘Monthly Income’ on y-axis and taking ‘Attrition’ in consideration.



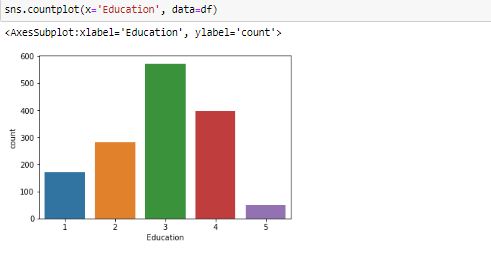
* As we see from the above graph that in Research and development department where salary is less than 5000 more number of people attrited but as the salary increases the attrited people are less as compare to not attrited people.
* In sales department we can see that when the salary is in range of 0 to 10000, the attrited people are less as compare to not attrited people.
* In HR department when the salary range is 0 to 5000, people are more attrited.

1. In this line of code we are making a histogram using matplotlib of ‘Distance from home’.



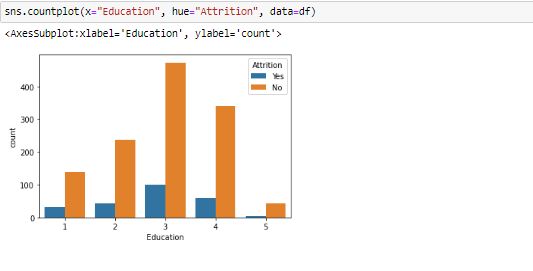
* As we see from the above graph that majority of the people live within 10km from the office.

1. In this line of code we are making a countplot for ‘Education’ of the employees in the company.



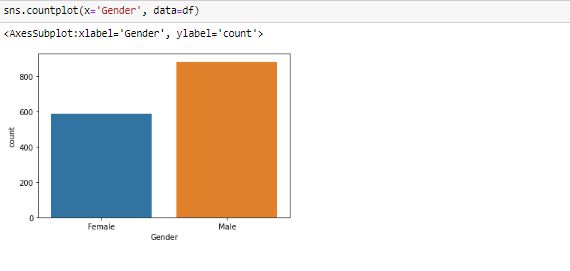
* As we see from the above graph that majority of the people have education level of 3.

1. Now we are plotting a countplot for the ‘Education’ of the employee and taking ‘Attrition’ in consideration. It will help us to see in what type of education the attrition level is high.



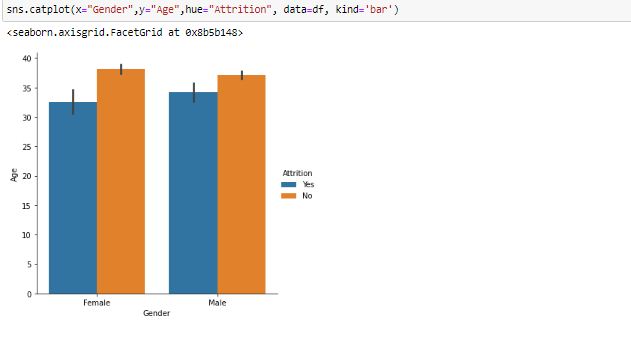
* As we see from the above graph that most of the peple are not attrited from every level of education.

1. Now we are see how many male and female employees working in the company.



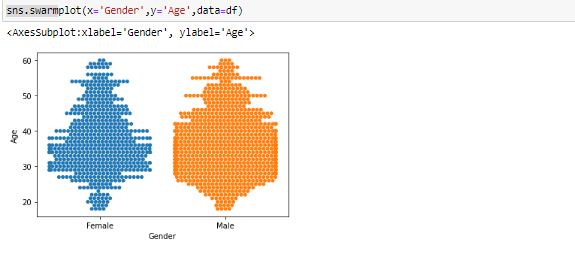
* We can observe from the above graph that the company has more male employees as compare to female.

1. Now are plotting a catplot by using matplotlib library of ‘Gender’ on x-axis and ‘Age’ on y-axis and taking ‘Attrition’ in consideration.



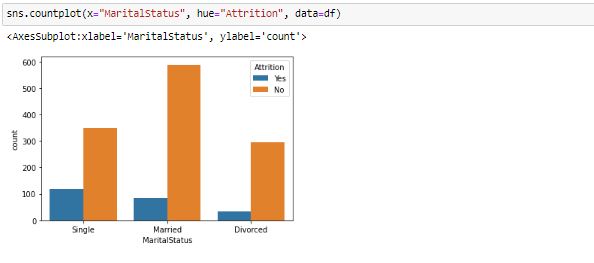
* As we see from the above graph that both male and female have less number of people who are Attrited.

1. Now in this line of code we are plotting swarmplot of ‘Gender’ on x-axis and ‘Age’ on y-axis.



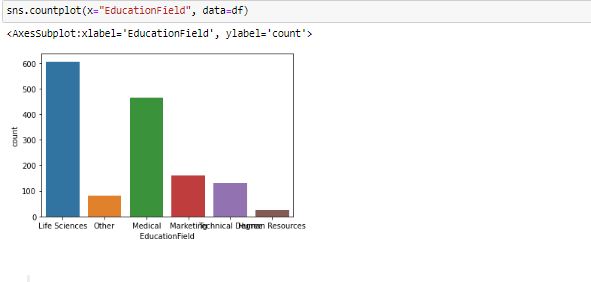
* We can observe from the above plot that the company has majority of the employees are in between age of 25 to 45.

1. In this line of code we are plotting a countplot of ‘Maritial Status’ on x-axis and taking ‘Attrition’ in consideration.



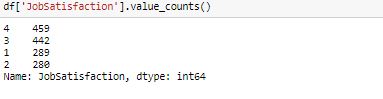
* We can observe from the above graph that people are less attrited with the company whether they are single, married or divorced.

1. Now we are checking how many education field in the dataset by using countplot. It help us to find how many employees are from which field.



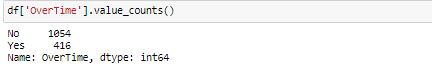
* We can observe from the above plot that the majority of the employees in the company are from life sciences field.

1. Now we are checking the satisfaction level of the job of the employee that they are doing by using a function called ‘value\_counts()’.



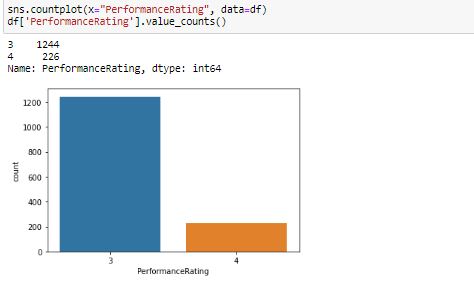
* From the above stats we can see that highest number of people are highly satisfied from the job.

1. Now we are checking how many people have done overtime by using a function called ‘value\_counts()’.



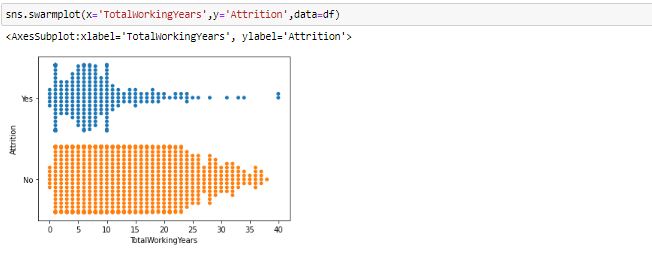
* We can see from the above stats that less number of people are doing overtime in the company.

1. Now we are doing making a countplot and value coun t of the ‘Performance Rating’ in a same line of code.



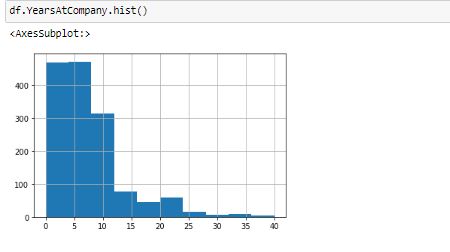
* From the above graph we can see that only 226 people have high performance rating.

1. In this line of code we are plotting a swarmplot of ‘Total Working Years’ on x-axis and ‘Attrition’ on y-axis.



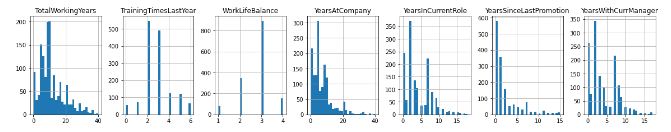
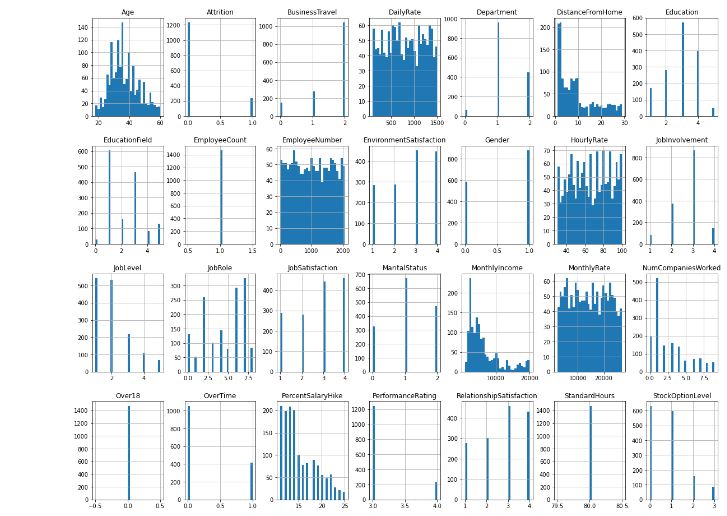
* From the above graph we can observe that employee in 1th year and 10th year number of people that are attrited are same that are not attrited.

1. Now we are plotting a histogram to depict the ‘Years at company’ of the employee. Through this we can check from how many years the employee are with the company.



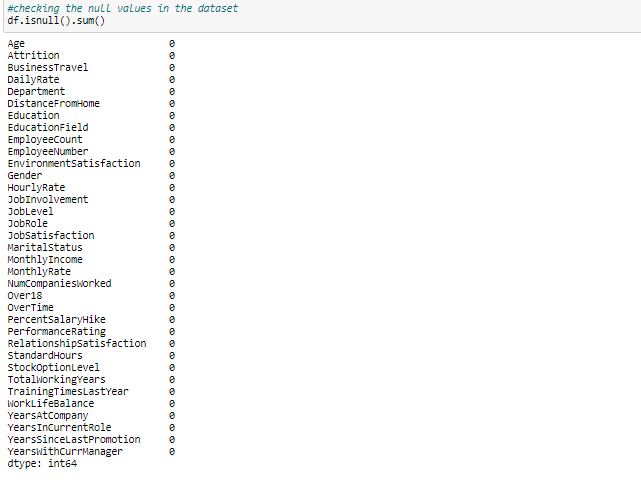
* From the above graph we can see that majority of the employees are with the company from past 7-8 years.

1. Now we are plotting histogram for every column in the dataset. Through we we can see what is the range of the data in every column of the dataset.

ad 24(1).JPG

* We can see that in some columns the range of the data is high.

1. Now we check that is their any null values in the dataset or not by using ‘isnull().sum()’ function.

we checking if it is 

* From the above stats we can see that their is no null values in the dataset.

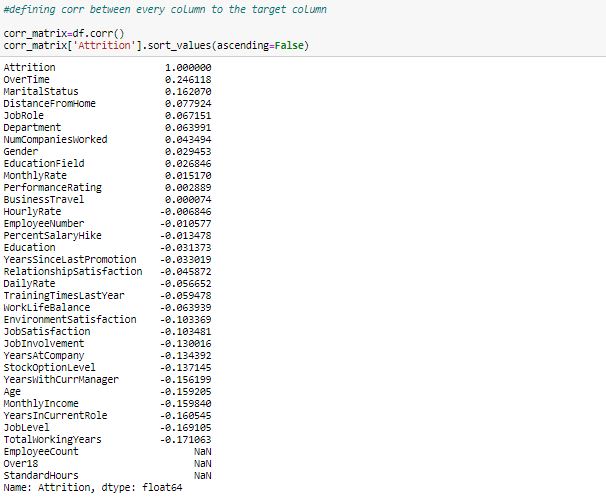
**Data Pre-Processing:-**

1. Now we are using a line of code to change the datatypes of the column from object to float. We are using for loop for this.



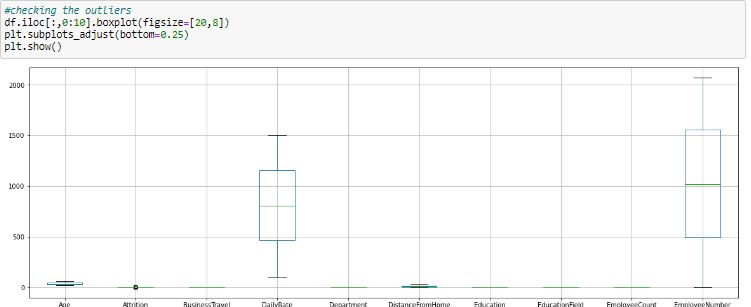
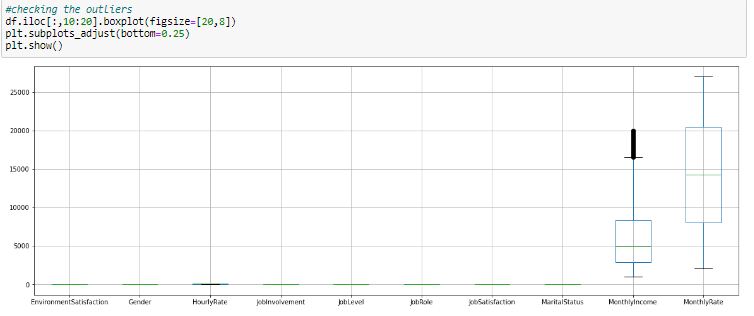
* First we give our ordinal encoder a instance enc then we use the loop for taking every column of the dataset on-by-one. Then this code will check if that column is object or not. If it is object then it will fit it in encoder and transform it and reshape it for making it in DataFrame again.
* With this line of code we can also use other encoding techniques like ‘Label Encoder’ etc.

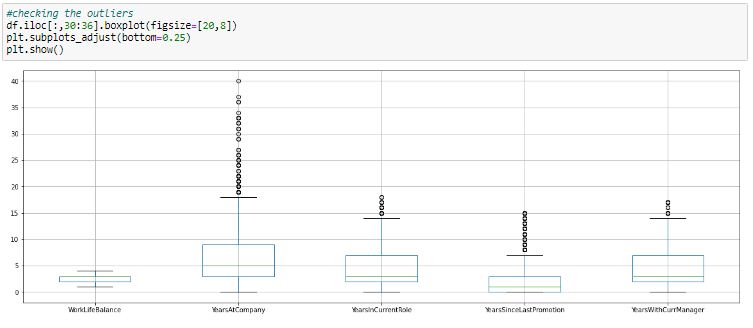
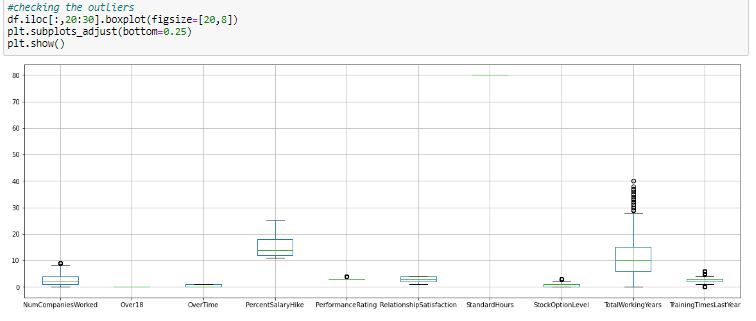
1. Now we are checking the correlation of every column with the target column that is ‘Attrition’.



* We can observe from the above stats that 'Overtime' column is most positively correlated column with the 'Attrition' column.
* We can also see that 'TotalWorkingYears' column is the most negatively correlated column with the 'Attrition' column.

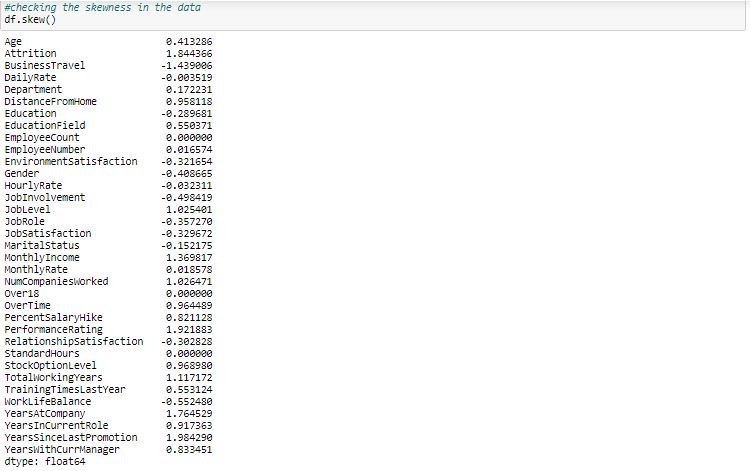
1. Now we have to check whether in our dataset there are outliers or not. We are checking the outliers in 10 columns each at one time because we have 35 columns and if we check outliers from every column inn one go, it will be very difficult for us to check.





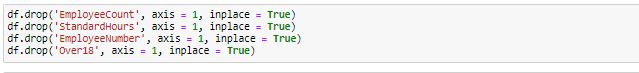
* We can observe that their are outliers in 'Monthly Income' column, But Some senior employees or higher position staff have more salaiers than the range of that column.
* We can see that their are considerable outliers in 'TotalWorkingYears', but senior staff from age above 50 have been working in the company so that outliers are those employees.
* We can see that their are considerable outliers in 'YearsAtCompany', But it might be possible that some senior employees have been working with the company from their starting of career.
* We can see that their are considerable outliers in 'YearsSinceLastPromotion', but promotion is based on performance, so might be possible that these employees are not performing well or these employees are in a position from where they can't be promoted further.

1. Now we are checking the skewness present in the dataset by using ‘skew()’ function present in the pandas.



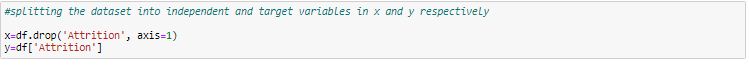
* As their is no considerable skewness present in the continous data column so we don't have to remove skewness from the data.

1. There are 4 columns in the dataset which we can drop from our dataset because these columns will not help us in prediction whether the employee is attrited or not. We are using ‘drop()’ function to drop these columns.



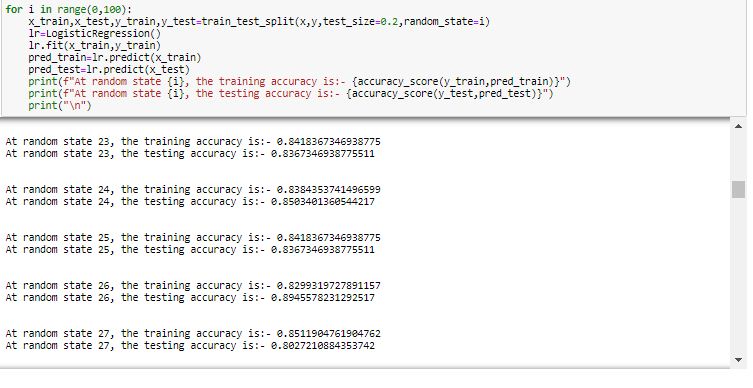
**Model Building:-**

1. Now we have to build a Machine Learning model to predict wheather the employee is attrited or not. For that we have to first split our dataset in ‘x’ and ‘y’ variable.



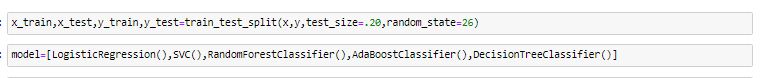
* In ‘x’ variable we have all the independent column of the dataset and in ‘y’ variable we have the target column that is ‘Attrition’.

1. Now we are checking the best random state for the dataset by taking a range of 0-100 and using logistic regression.

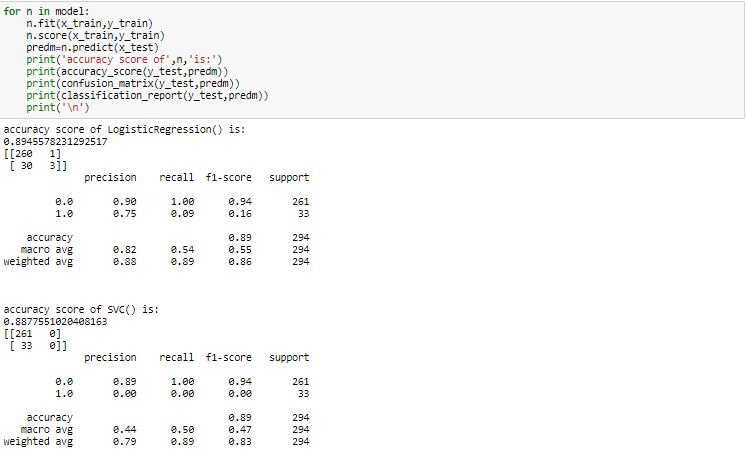


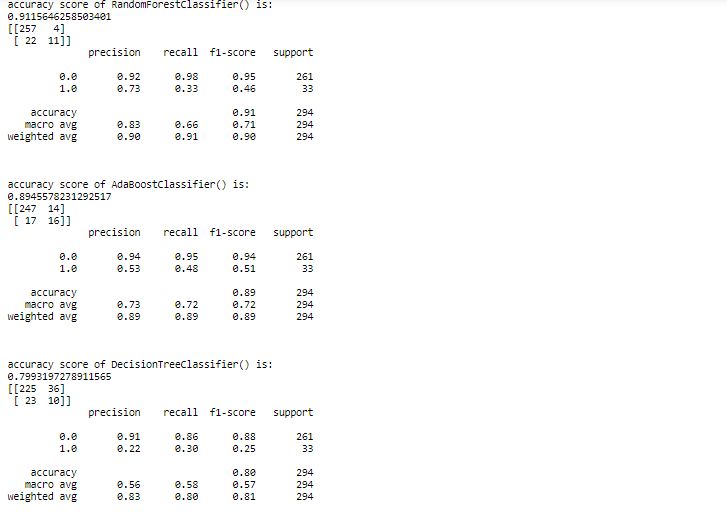
* From the above stats, at random state 26 the testing accuracy and training accuracy is hightest.

1. Now in this line of code we have to spilt our data into training data and testing data in 80:20 ratio and making a list of all the models we are using in this case study.



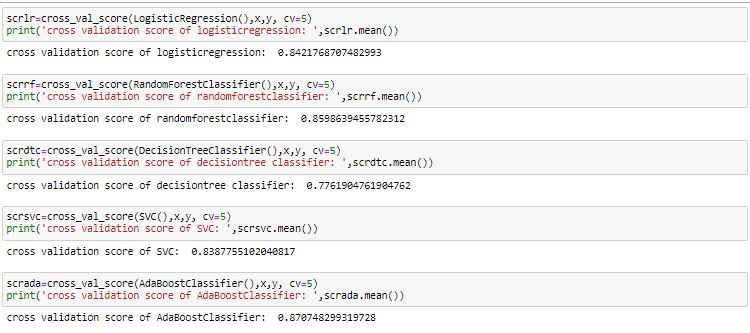
1. Now we are giving the data to our different models through a loop.





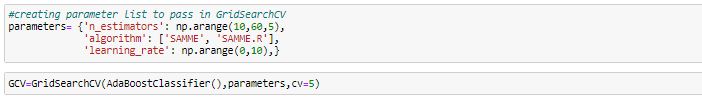
* As we can see that Random forest classifier has the best accuracy score.

1. Now checking the cross validation score to check whether our models are underfitted or overfitted.



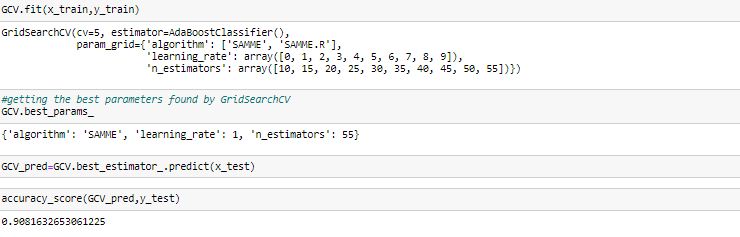
* As we can see the Minimum difference in accuracy score and cross validation score is for DecisionTreeClassifier and AdaBoostClassifier (2) so they are our best model.
* We are using AdaBoostClassifier.

1. Now we have to do Hyper parameter tunning to increase our model score. We are using GridSearchCV for this. First we make a list of all the parameters of the AdaBoostClassifier. Then we have to insert our model into the GridSearchCV with the parameters list.



* We have done the above mentioned things by taking the k-fold value 5.

1. Now we have to fit the parameters to our train dataset and get the best parameters by using ‘best\_params\_’ function of the GridSearchCV. After getting the best parameters we have to fit these parameters into our test data and find the upgraded accuracy score of our AdaBoostClassifier model.



* As we can see that after fitting the best parameters to the training and testing data, our accuracy score has increased to approximately 91%.

1. Now we will save this model by using the ‘joblib’ library and in that library we are using ‘dump’ function.

ad 43.JPG

* We save our model with best parameters as ‘HR Analytics’. We can call this model anytime and predict the ‘Attrition’ of any data of this type.

**Case Study:**

* 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 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.
* We have to understand the data annd find out whether the employee is attrited or not.

**Findings:**

* Number of not attrited people is more than no of attrited people.
* With monthly income more prople are not attrited as compare to people that are attrited.
* More no of people are not attrited if they are travel rarely, travel frequently or not travelling.
* In research and development department majority of the people are not attrited. In sales department also majority of the people are not attrited. In Human resource department also more no of people are not attrited.
* In Research and development department where salary is less than 5000 more number of people attrited but as the salary increases the attrited people are less as compare to not attrited people. In sales department we can see that when the salary is in range of 0 to 10000, the attrited people are less as compare to not attrited people. In HR department when the salary range is 0 to 5000, people are more attrited.
* Majority of the people live within 10km from the office.
* Majority of the people have education level of 3.
* Most of the peple are not attrited from every level of education.
* The company has more male employees as compare to female.
* Both male and female have less number of people who are Attrited.
* The company has majority of the employees are in between age of 25 to 45.
* People are less attrited with the company whether they are single, married or divorced.
* The majority of the employees in the company are from life sciences field.
* Highest number of people are highly satisfied from the job.
* Less number of people are doing overtime in the company.
* Only 226 people have high performance rating.
* Employee in 1th year and 10th year number of people that are attrited are same that are not attrited.
* Majority of the employees are with the company from past 7-8 years.
* Their is no null values in the dataset.
* 'Overtime' column is most positively correlated column with the 'Attrition' column. We can also see that 'TotalWorkingYears' column is the most negatively correlated column with the 'Attrition' column.

**Predictions:**

* At random state 26, the testing accuracy and the training accuracy is highest.
* The best accuracy\_score is for RandomForest Classifier.
* The Minimum difference in accuracy score and cross validation score is for DecisionTreeClassifier and AdaBoostClassifier (2) so they are our best model.
* After doing hyperparameter tuning. we increase the score of our model from 87% to 91%.