**HR ANALYTICS - UNDERSTANDING THE ATTRITION IN HR**

***We know that attrition is the departure of employees from the organization for different reasons like resignation, termination, death or retirement. Every year companies hire a particular number of employees & they invest lot of time and money in training those employees, not just this but they provide training programs also within the companies for their existing employees as well. The ultimate aim of doing these programs & giving training is to increase the effectiveness of the employees. But if we talk about HR Analytics, first question arises to most of us is, where HR Analytics fit in this? Is it just about improving the performance of employees? or else? or what? etc... etc...***

***Firstly, we need to understand is HR analytics which is Human resource analytics (HR analytics) also known as people analytics means the collection and application of talent data to improve critical talent and business outcomes. In other words, it 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.***

***Attrition is one of the problems that impacts all businesses, irrespective of geography, industry and size of the company. Employee attrition leads to significant costs for a business, including the cost of business disruption, hiring new staff and training new staff. 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 employees. A major problem in high employee attrition is its cost to an organization. Job postings, hiring processes, paperwork, new hire training is 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. For e.g. If your business is customer-facing or communicating with them than most of the customers prefer to interact with familiar people or voices they always communicate with. In that case chances of errors and issues are more likely if you constantly have new workers and for that HR Analytics help in analyzing the attrition.***

***To do the analysis we will use classification models to predict if an employee is likely to quit or not. It will be great method to increase the HR’s ability to intervene on time and remedy the situation to prevent attrition. While this model can be routinely run to identify employees who are most likely to quit, the key driver of success would be the human element of reaching out the employee, understanding the current situation of the employee and taking action to overcome with the controllable factors that can prevent attrition of the employee.***

***This data set represents an employee survey from IBM, indicating if there is attrition or not. The data set contains approximately 1500 entries. We will analyze the data set by providing modest improvement in identification of attrition & random allocation of probability of attrition.***

***While some level of attrition in a company is inevitable, minimizing it and being prepared for the cases that cannot be helped will significantly help improve the operations of most businesses. As a future development, with a sufficiently large data set, it would be used to run a segmentation on employees, to develop certain “at risk” categories of employees. This could generate new insights for the business on what drives attrition, insights that cannot be generated by merely informational interviews with employees.***

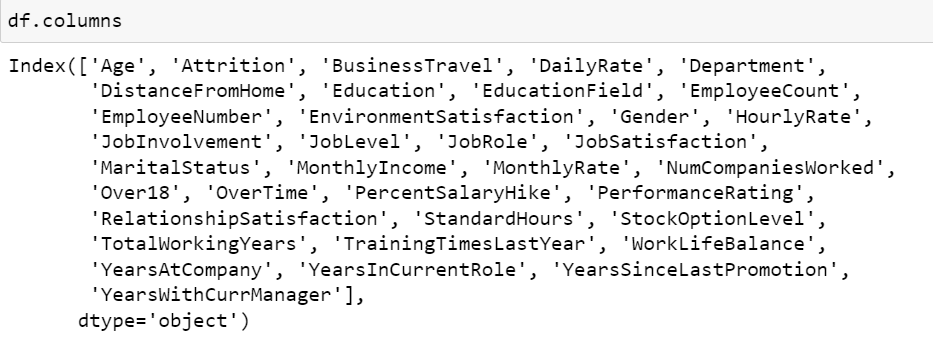
***IBM has gathered information on employee satisfaction, income, seniority and other different factors. It includes the data of total 1470 employees with 35 columns.***

***Before building any model, we need to check the dataset properly if it needs any type of changes or cleaning the data to remove unwanted data, missing values, rows or columns, duplicate values, data type conversion, etc.***

***I have tried here to clean some data which was not that necessary according to me and removed those columns from the dataset.***



***As the dataset has following variables which has the object data type with which we can’t run the model so, we changed the dataset to numeric data or Integer with Label Encoder where string data was present for processing the further methods & techniques. It will give us the exact results.***



So, like the ’Attrition’ which is output variable shown below I have converted following data: -

Attrition - [0 - 'No', 1 - 'Yes']

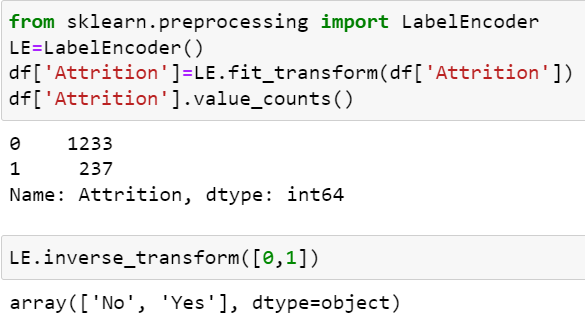
BusinessTravel - [0 - 'Non-Travel', 1 - 'Travel\_Frequently', 2 - 'Travel\_Rarely']

Department - [0 - 'Human Resources', 1 - 'Research & Development', 2 - 'Sales']

EducationField - [0 - 'Human Resources', 1 - 'Life Sciences', 2 - 'Marketing', 3 - 'Medical', 4 - 'Other', 5 - 'Technical Degree']

JobRole - [0 - 'Healthcare Representative', 1 - 'Human Resources', 2 - 'Laboratory Technician', 3 - 'Manager', 4 - 'Manufacturing Director', 5 - 'Research Director', 6 - 'Research Scientist', 7 - 'Sales Executive', 8 - 'Sales Representative']

OverTime - [0 - 'No', 1 - 'Yes']

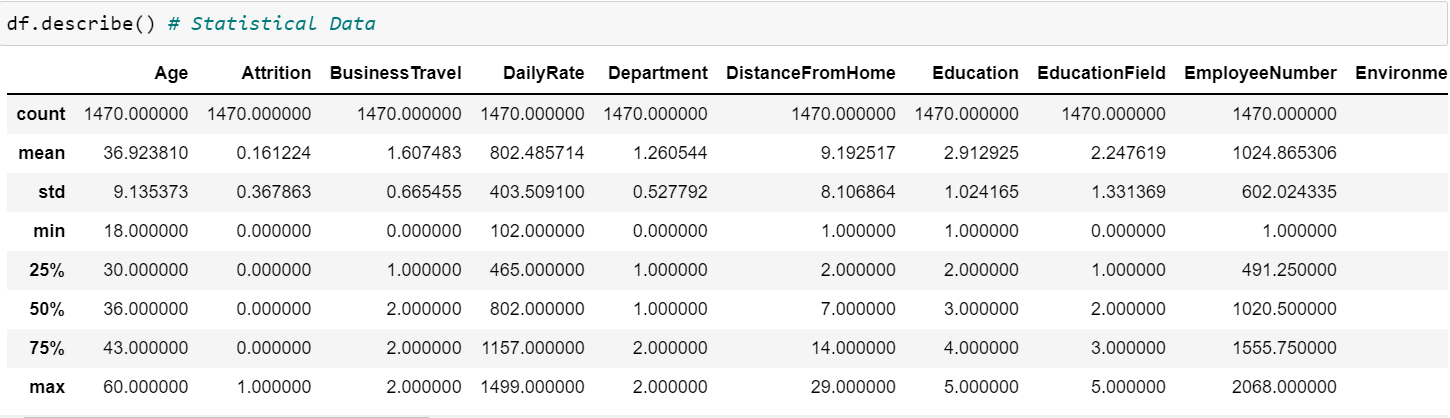


Checked if there are any null values are available but not found so it was good to go with other steps.

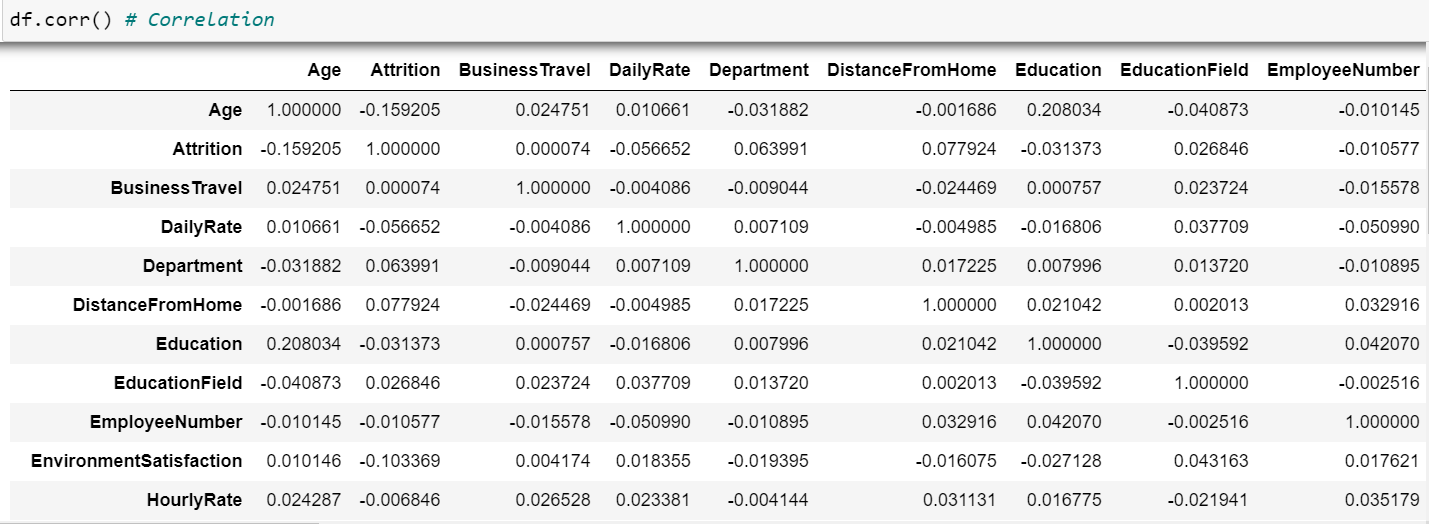
With the df.describe() method got the numerical/ statistical description of information for each column shown below: -

std - The standard deviation.

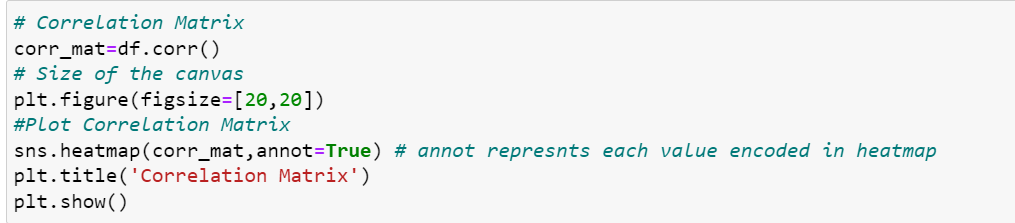
min - the minimum value.   
25% - The 25% percentile.   
50% - The 50% percentile.   
75% - The 75% percentile.   
max - the maximum value.

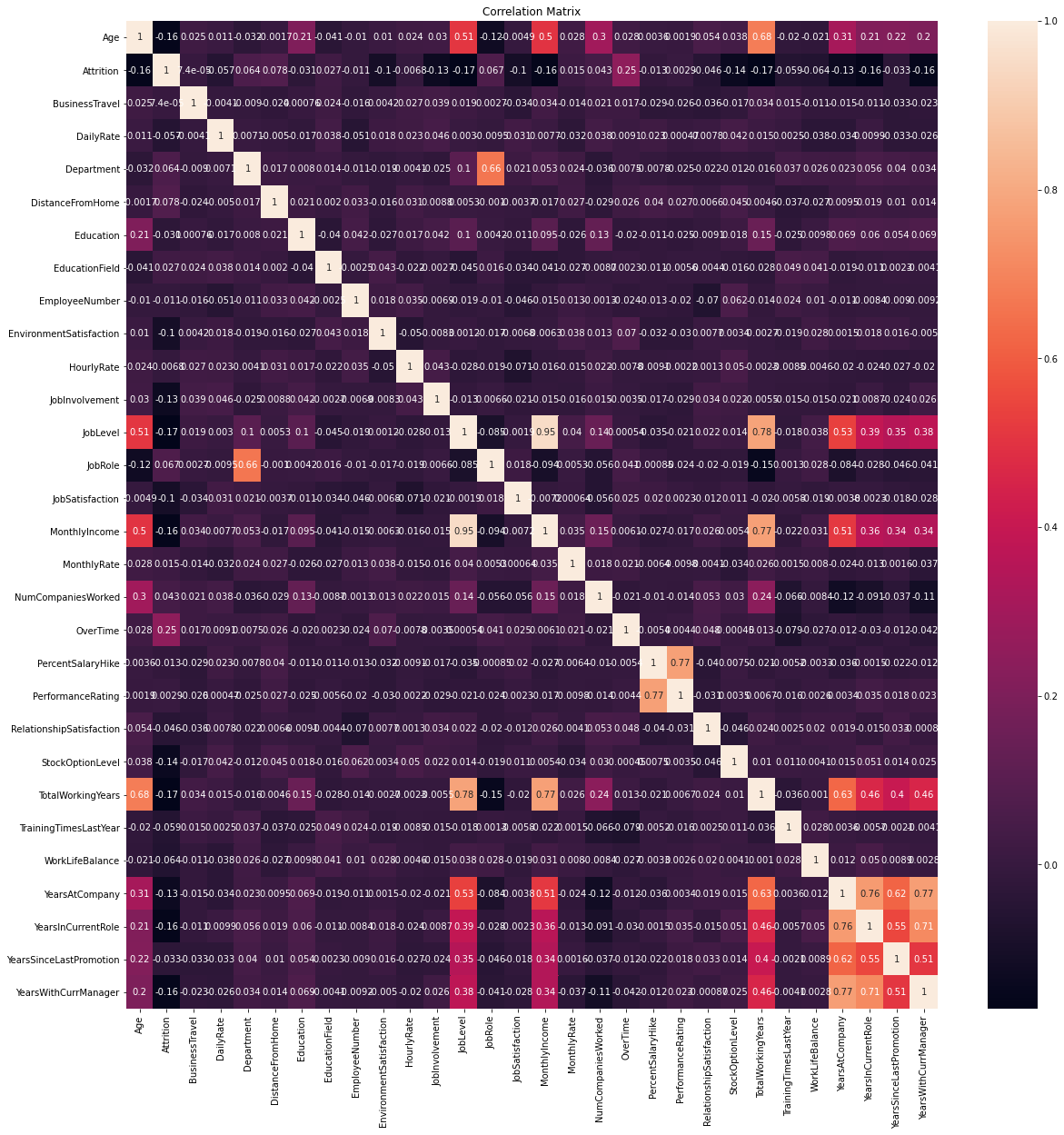


Then, I checked the correlation among all the variables with each other.

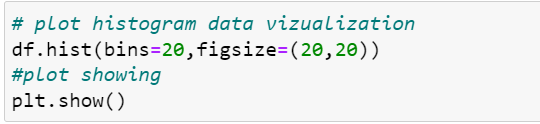


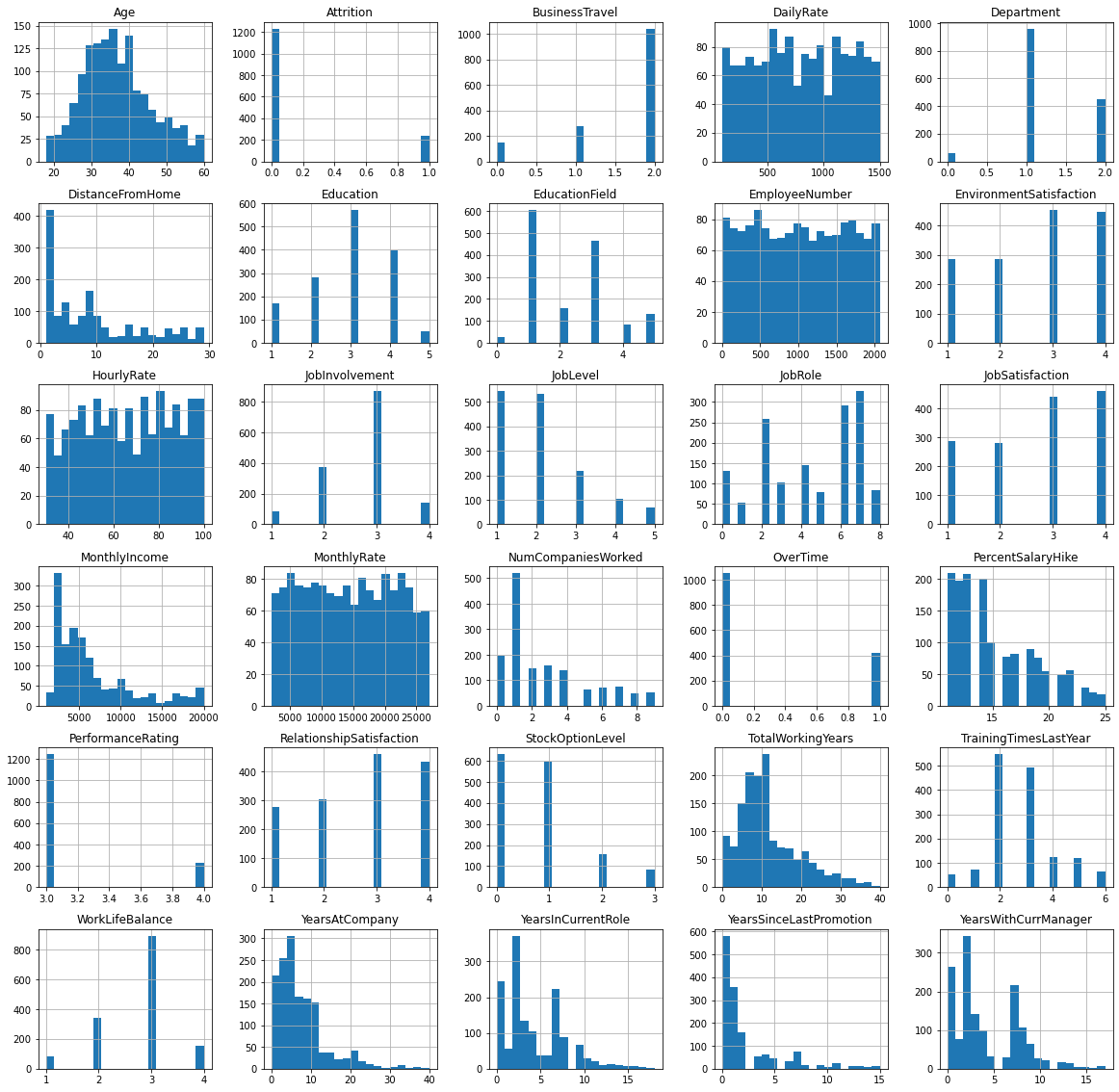
***And shown the graph of “Correlation Matrix”. It shows a correlation between two discrete dimensions, using colored cells to represent data from usually a monochromatic scale.***

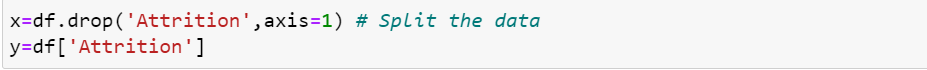




I used Histogram to visualize the data and according to me it was the best graph as there are many input variables which has categorical data.

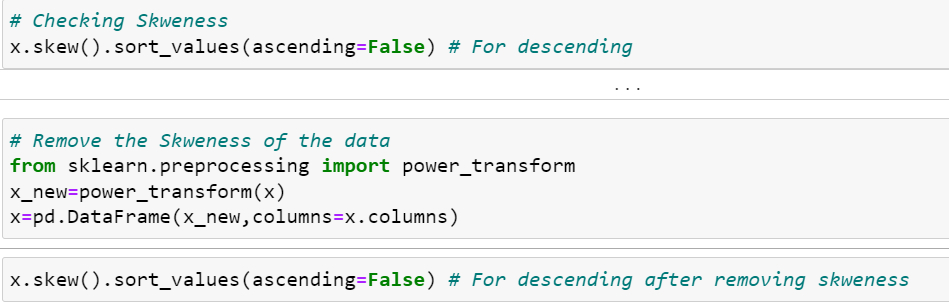




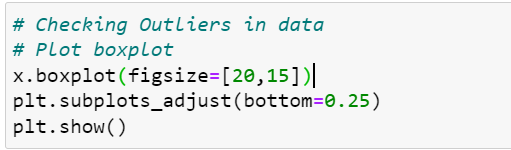


I split the data before checking the skewness and outliers because the output is categorical data.

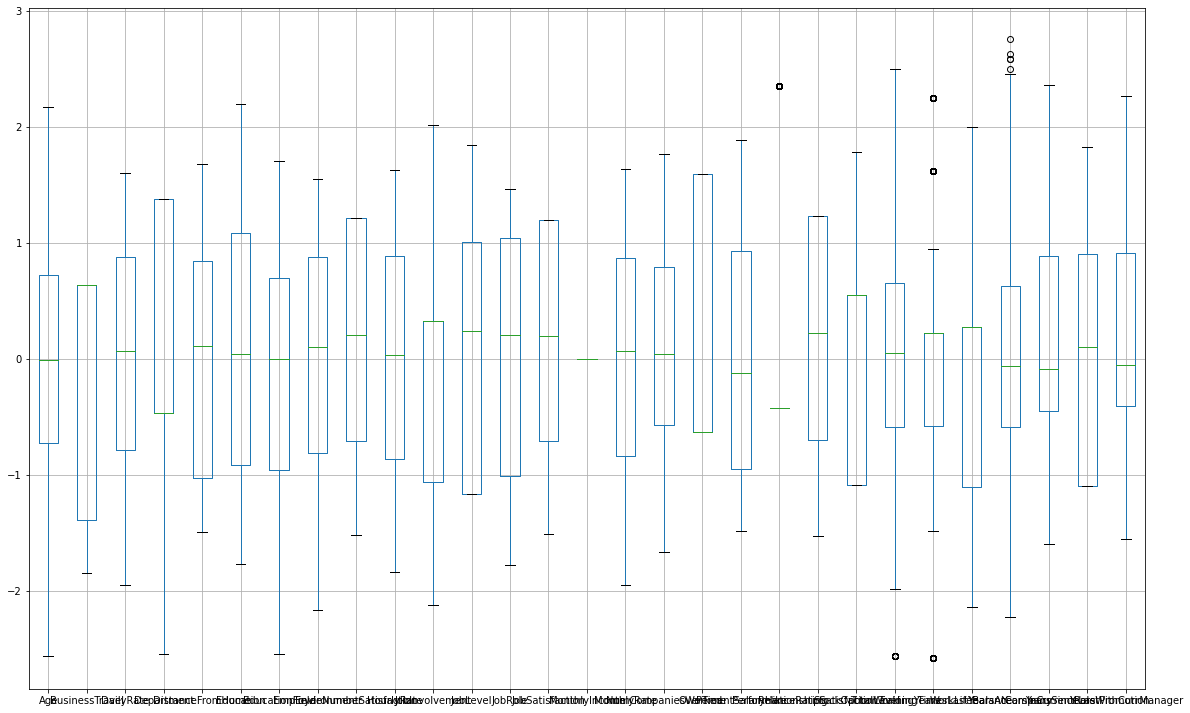
After splitting the data removed the skewness & Outliers from the dataset.



For removing the outliers, I used boxplot.



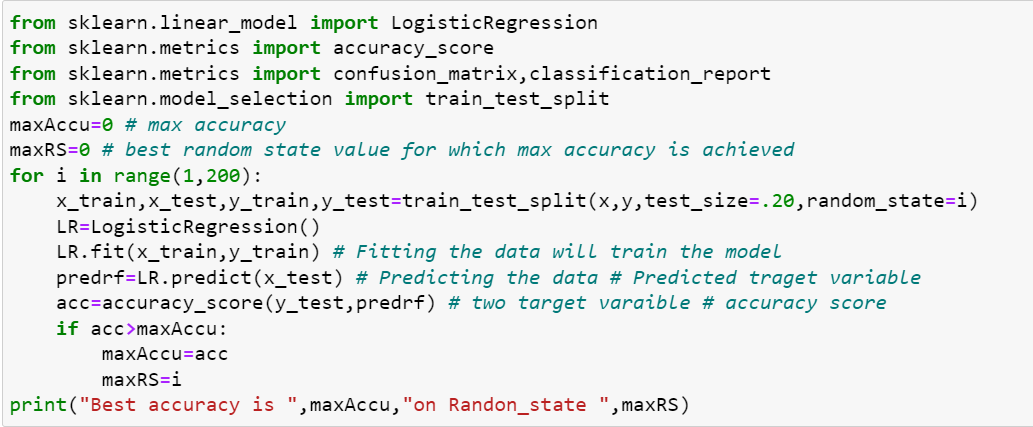
When I checked the outliers, it was very less as shown and for only 1500 data I didn’t remove the outliers and avoid unnecessary reduce the amount of data.



I have done Splitting of the cleaned data into two sets - a training set and a testing set. The training set is the set your model learns from. A testing set is used to check the accuracy of your model after training.

With the following code we find the best accuracy of the data and best random state.

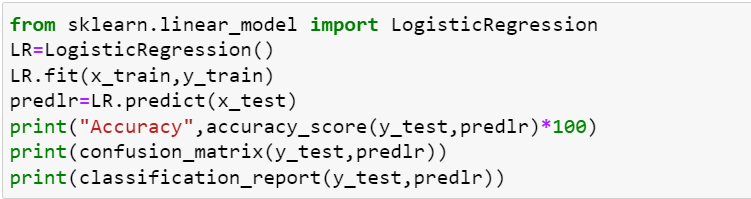
random state is basically used for reproducing your problem the same every time it is run. If you do not use a random\_state in train\_test\_split, every time you make the split you might get a different set of train and test data points and will not help you in debugging in case you get an issue.



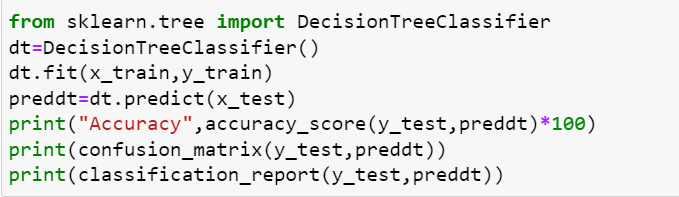


By splitting the data next, we build the model with different classification model building techniques to know which one is best fitting for the dataset. I used,

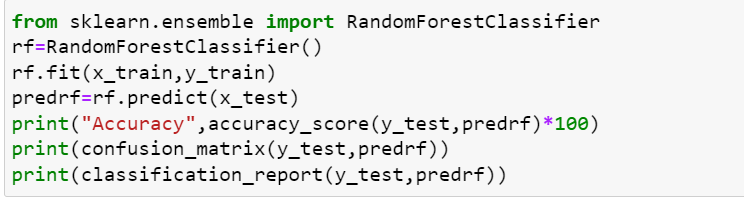
Logistic Regression



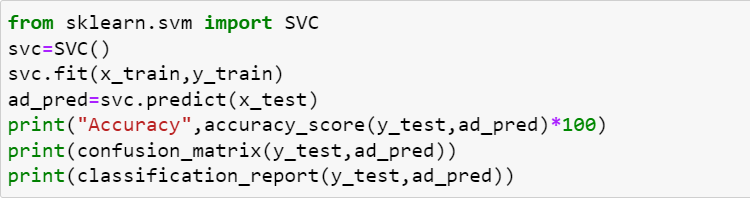
Decision Tree Classifier



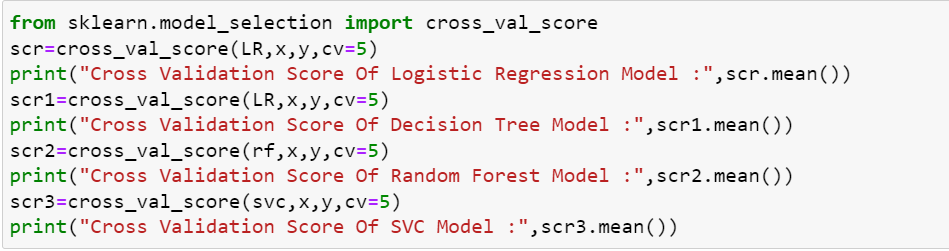
Random Forest Classifier



SVC - Support Vector Classifier



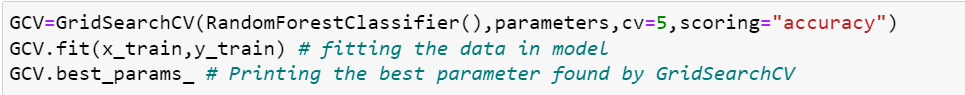
To check overfitting, I used Cross Validation technique to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model.

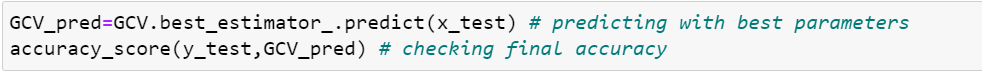


Hyperparameter tuning is the problem of choosing a set of optimal [hyperparameters](https://en.wikipedia.org/wiki/Hyperparameter_(machine_learning)) for a learning algorithm. Different machine learning models can require different constraints, weights or learning rates to generalize different data patterns. These measures are called hyperparameters, and have to be tuned so that the model can optimally solve the machine learning problem

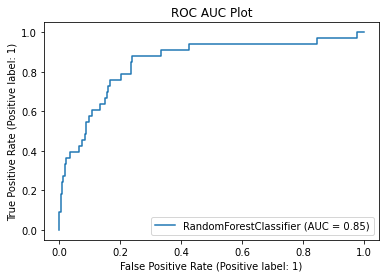
I used GridSearchCV for performing the hyperparameter tuning.





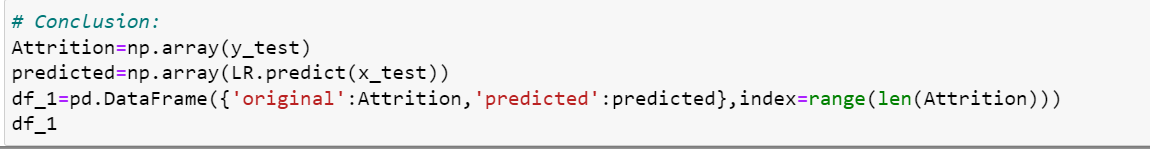


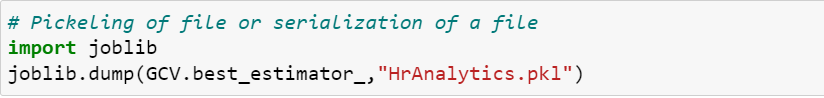
Next, I plot ROC AUC curve which shows the AUC Score & the distribution curve.



In the end I have tried my best to get good results and got the final accuracy is 90% & AUC Score is 85%.

With this I have saved my model & conclusion for the same on jupyter notebook.





The analysis shows that,

Correlation Heatmap gives the idea about the features related to each other in a negative or positive way. When I checked I found most of the features are positively related to each other & less numbers are of negatively corelated.

About the visualization of Histogram is quit sorted for many variables as if we see the plot above the data is categorical and it itself show the bifurcation of all the categories separately.

By removal of skewness in data there were no such major number of outliers exist in data shown in boxplot above so kept as it is.

All the models used are showing almost same score and we can choose any of them or there is no particular model we can say which suits best.

Cross Validation Score Of Logistic Regression Model : 0.8734693877551021  
Cross Validation Score Of Decision Tree Model : 0.8734693877551021  
Cross Validation Score Of Random Forest Model : 0.8551020408163266  
Cross Validation Score Of SVC Model : 0.8659863945578232

The data is showing 90% accuracy which is good. With this we can say that the data is correct and we can consider the data for further findings or processes.

Finally, I would like to conclude that no doubt Analytics helps to reduce the attrition rate or give a way to handle the issues related to employees, ultimately to understand what exactly employees' needs are or what they really want that is a big task which is not as easy as we think.

But if when we reach out to employees and find out how it will work well for them with their comfortability, their likings, their way to handle things, their way of doing things, their opinions on anything & try to work on it matters a lot according to me and will definitely help to reduce the attrition rate.

You can find the solution & dataset on below GitHub Links to refer,

Solution - <https://github.com/komalghatvilkar/Practice_Projects_ML/blob/main/HR%20Analytics%20Project-%20Understanding%20the%20Attrition%20in%20HR.ipynb>

Dataset - <https://github.com/dsrscientist/IBM_HR_Attrition_Rate_Analytics>