A Comprehensive Guide-HR Analytics Project



In this article, we’re going to discuss employee attrition prediction i.e. predicting that employee will leave the current company (or will resign from the current company) and we will do this using several machine learning algorithms (basically 9 ML algorithms) but this article is gonna be completely step by step explanation. So let’s get started.

## **Need of Employee Attrition prediction**

1. **Managing workforce:**If the supervisors or HR came to know about some employees that they will be planning to leave the company then they could get in touch with those employees which can help them to stay back or they can manage the workforce by hiring the new alternative of those employees.
2. **Smooth pipeline:** If all the employees in the current project are working continuously on a project then the pipeline of that project will be smooth but if suppose one efficient asset of the project(employee) suddenly leave that company then the workflow will be not so smooth
3. **Hiring Management:** If HR of one particular project came to know about the employee who is willing to leave the company then he/she can manage the number of hiring and they can get the valuable asset whenever they need so for the efficient flow of work.

**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 over time. 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.*

Hope the basics made sense. Let’s move on to problem statement and through coding we will try finding out how HR Analytics help in understanding attrition.

**Table of content**

1. Problem Definition
2. Dataset link
3. Data Analysis
4. EDA
5. Pre-processing Data
6. Building Machine Learning Models
   * + - Logistic Regression
       - Decision Tree Classifier
       - Random Forest Classifier
       - K Neighbors Classifier
       - Support Vector Classifier
       - XGB Classifier
       - Gaussian Naïve B
       - Ada Boost Classifier
       - Gradient Boosting Classifier
7. Concluding Remarks

**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?

**Dataset Link**

<https://github.com/Pratyush-Ghosh/DataScience-Projects/blob/main/hr%20analytics/WA_Fn-UseC_-HR-Employee-Attrition.csv>

**Data Analysis**

In order to start with exercise, I have used IBM HR Analytics Employee Attrition & Performance Dataset. The dataset includes features like Age, Employee Role, Daily Rate, Job Satisfaction, Years At Company, Years in Current Role etc. For this exercise, we will try to study the factors that lead to employee attrition. This is a fictional data set created by IBM data scientists.

Let’s get start with the work.

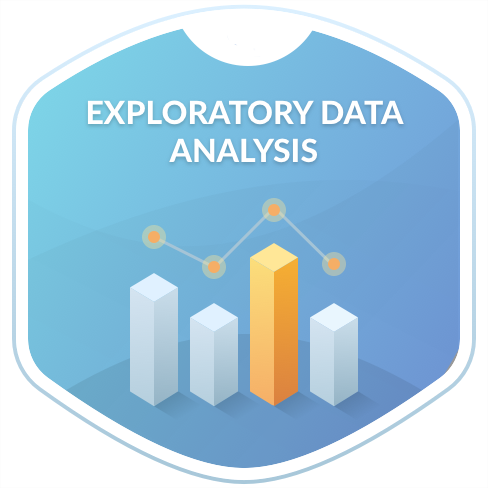
# Importing all required libraries:

# 

# Data Preparation:

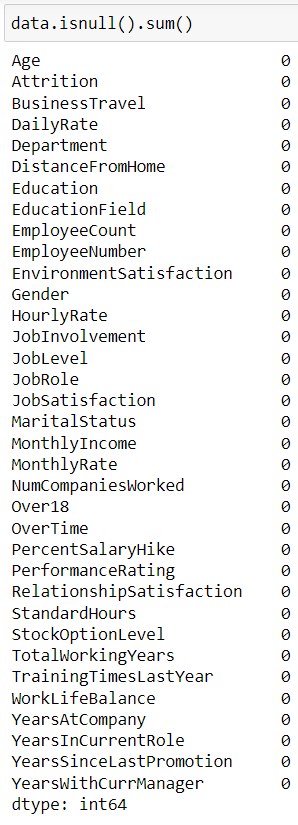
# Here, I download the csv file in my local storage and called it.

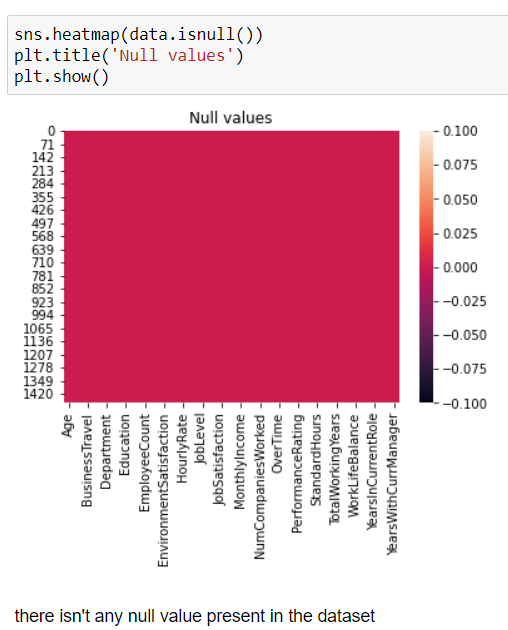
# EDA: Exploratory Data Analysis



Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

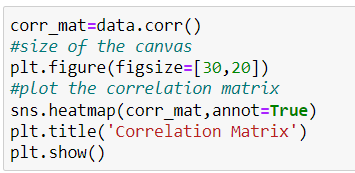
It is a good practice to understand the data first and try to gather as many insights from it. EDA is all about making sense of data in hand

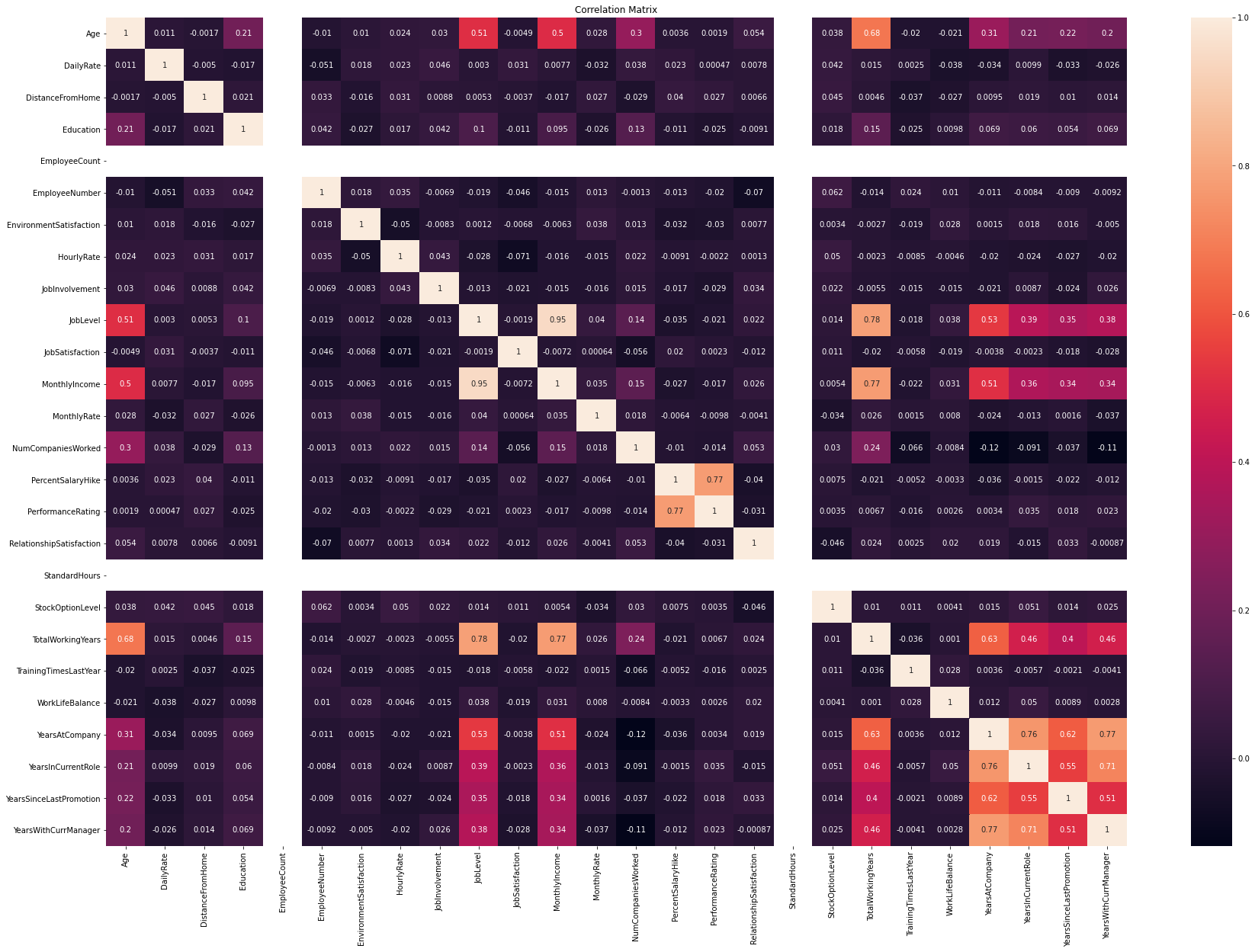
****

****

Luckily, we don’t have any missing values from the provided dataset

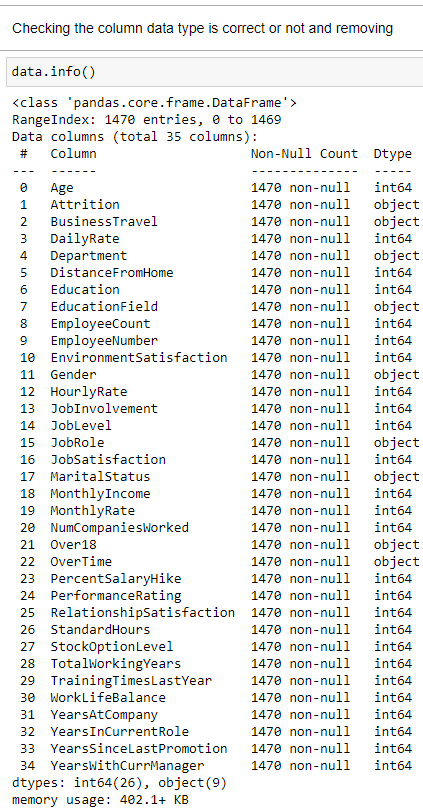
Now, let’s check the correlation using heat map



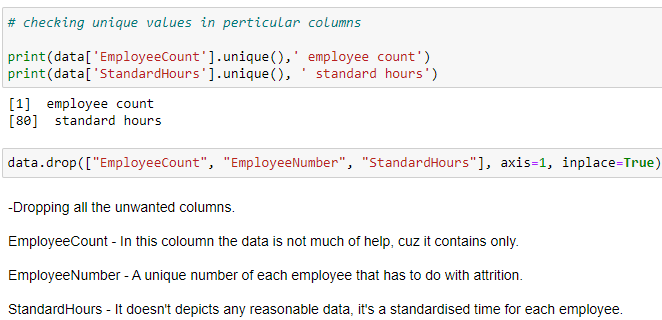


* From heat map we can say that the column MonthlyIncome is totally arround 95% depends on JobLevelOfEmployee.
* Columns of Age, JobLevel, MonthlyIncome, TotalWorkingYears, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, and YearsWithCurrentManager are showing some variation to each other.
* Other features are not showing any strong relation between them.

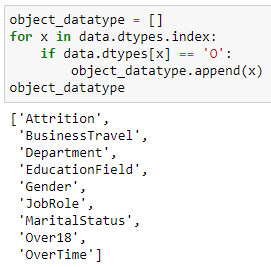
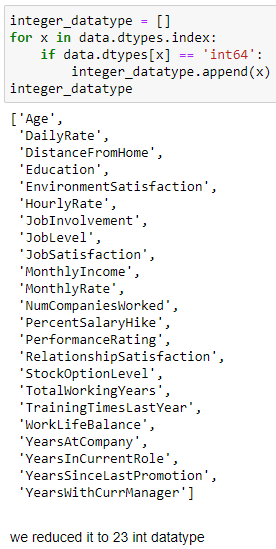
Now, there might be some unnecessary columns which we don’t want in our final models. So we have to get rid of them, for that we have to get information of each column.

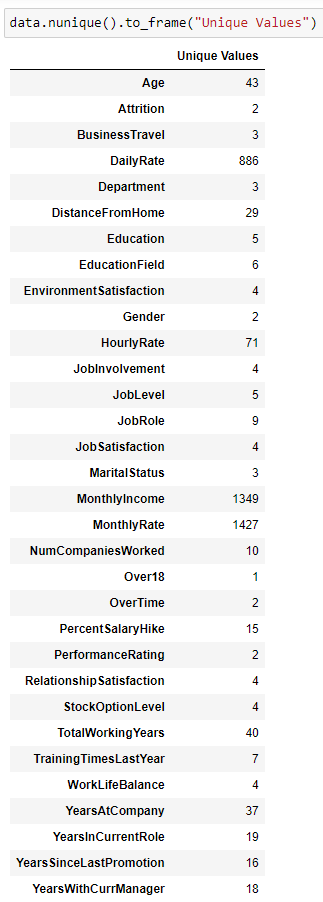


So from this we came up droping…….



One of the things that I like to do is separate the object datatype and numeric datatype values that allows for easier processing in further steps. The code to do that is a simple for loop usage.

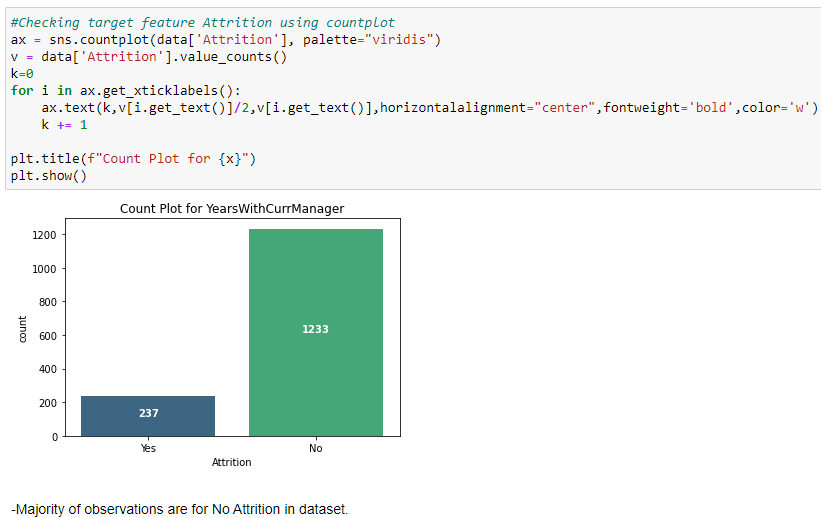


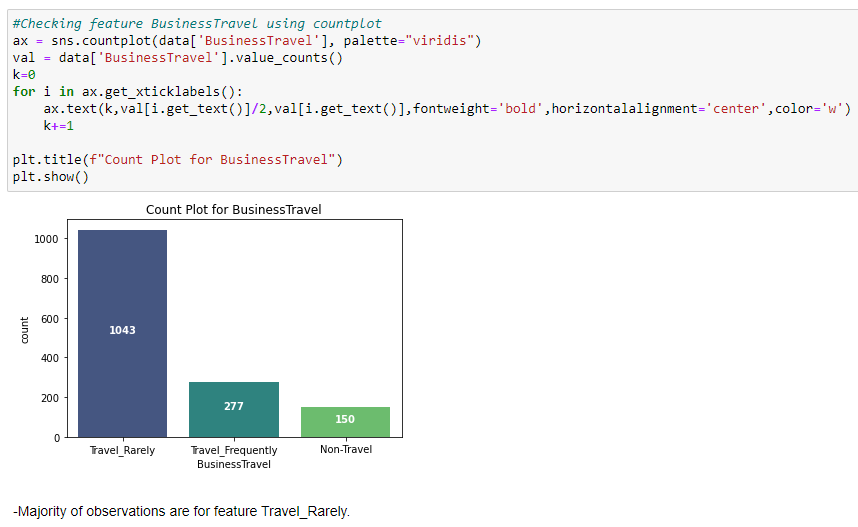
NumPy library is used in python to create one or more dimensional arrays, and it has many functions to work with the array. The unique() function is one of this library’s useful functions to find out the unique values of an array and return the sorted unique values. This function can also return a tuple of array values, the array of the associative indices, and the number of times each unique value appears in the main array.

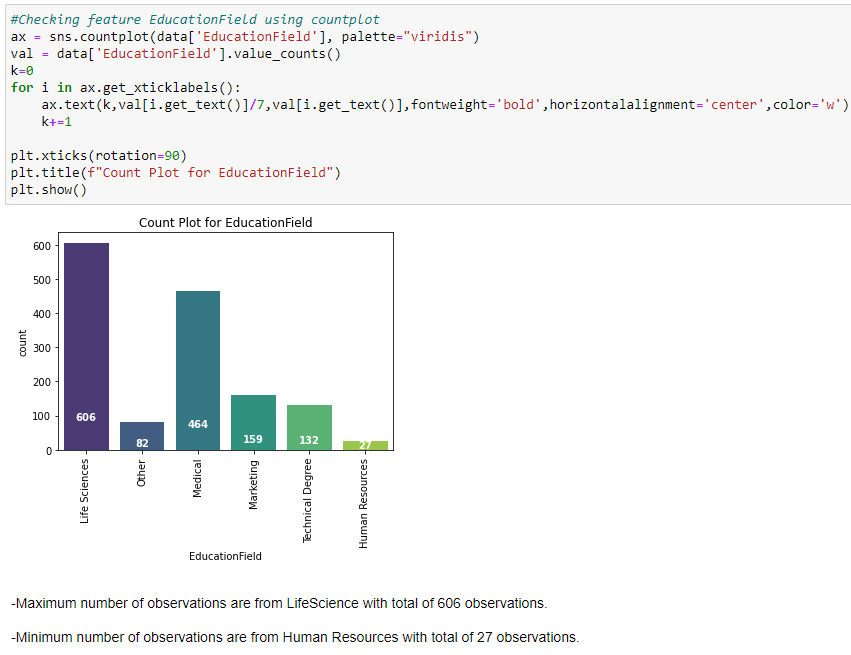
These line of codes provide us the output where we get an entire list of column names with unique data covered in the dataset rows providing a numerical data and then a description of those values for categorical object datatype columns.

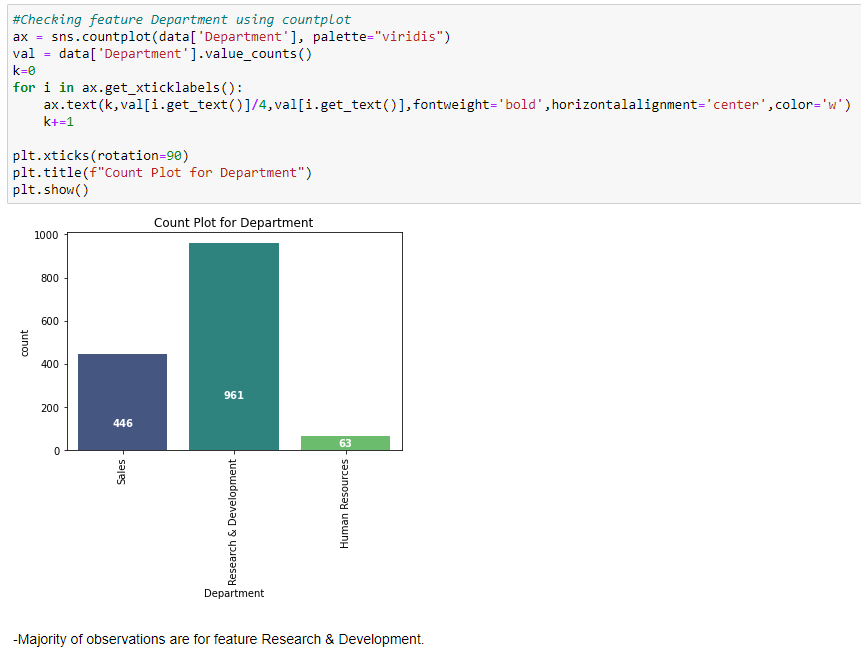
we will, take care of those columns more in near future since one or two unique value will be liability to our prediction but before that we have to get some visualization.

Visualization:



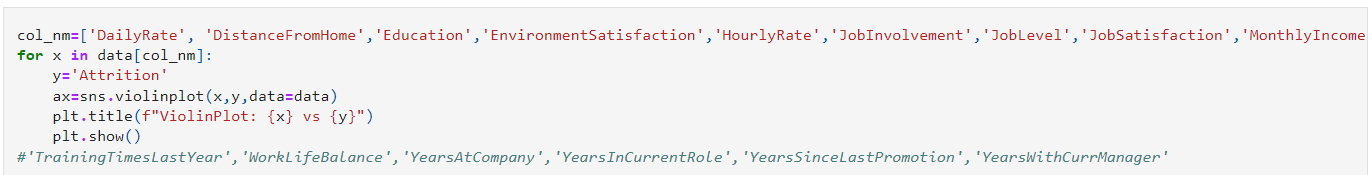




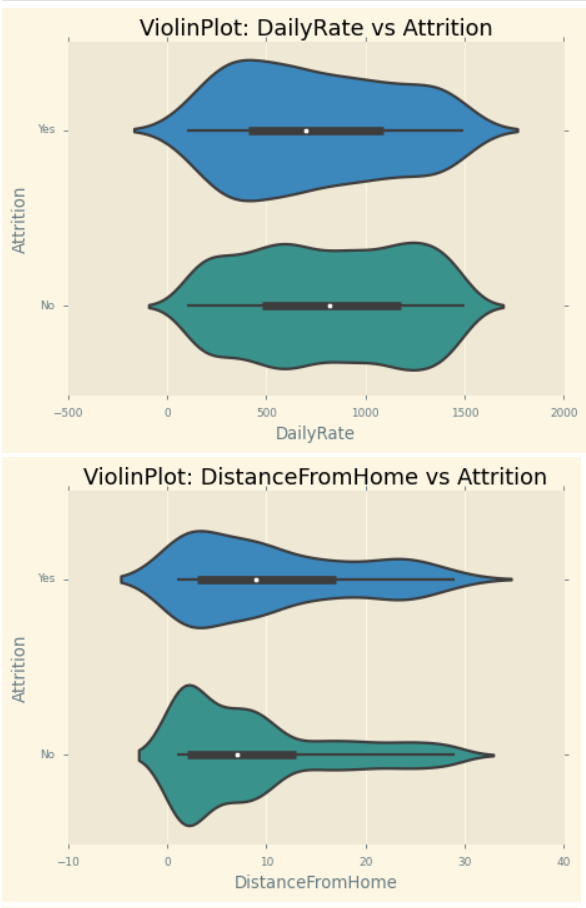


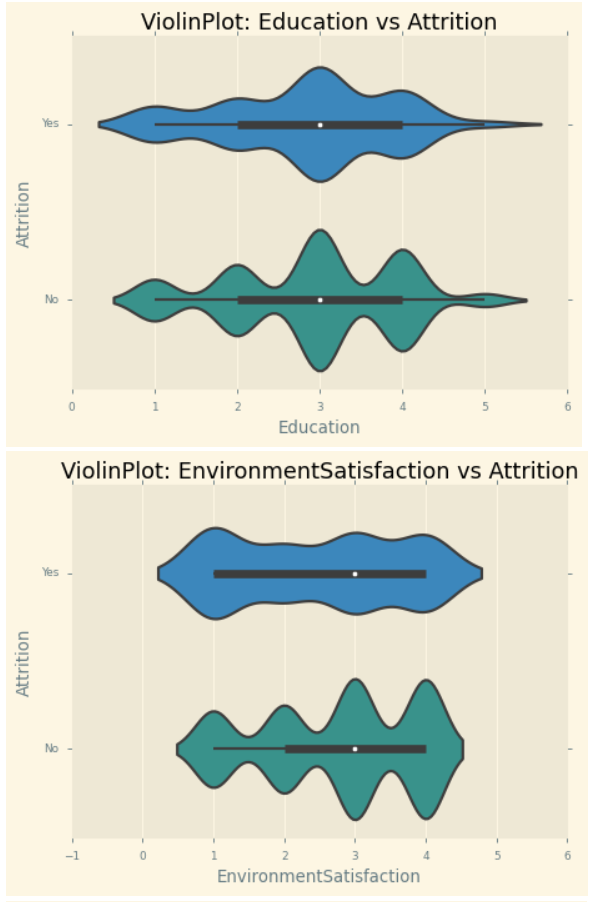
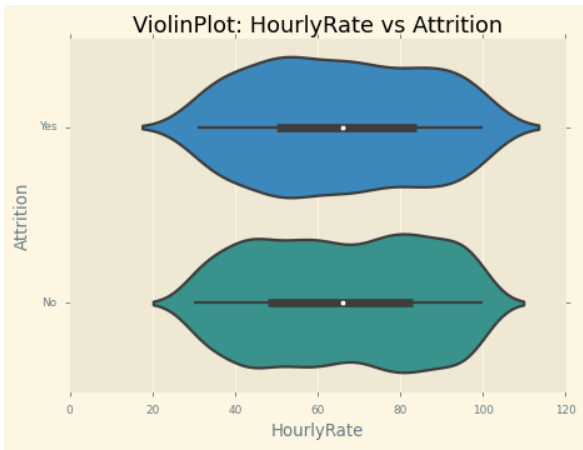


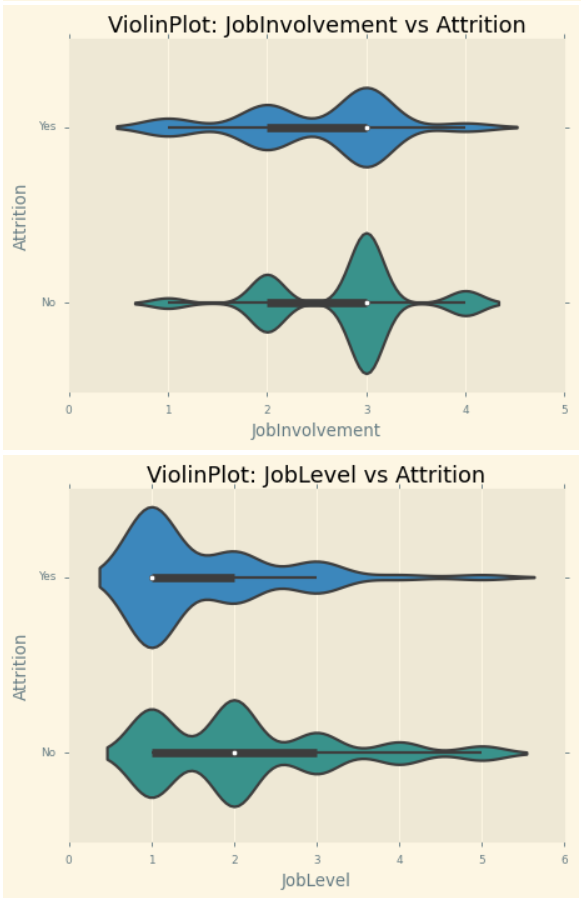
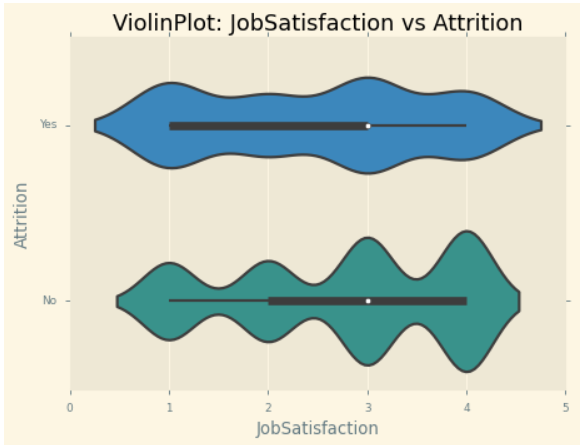


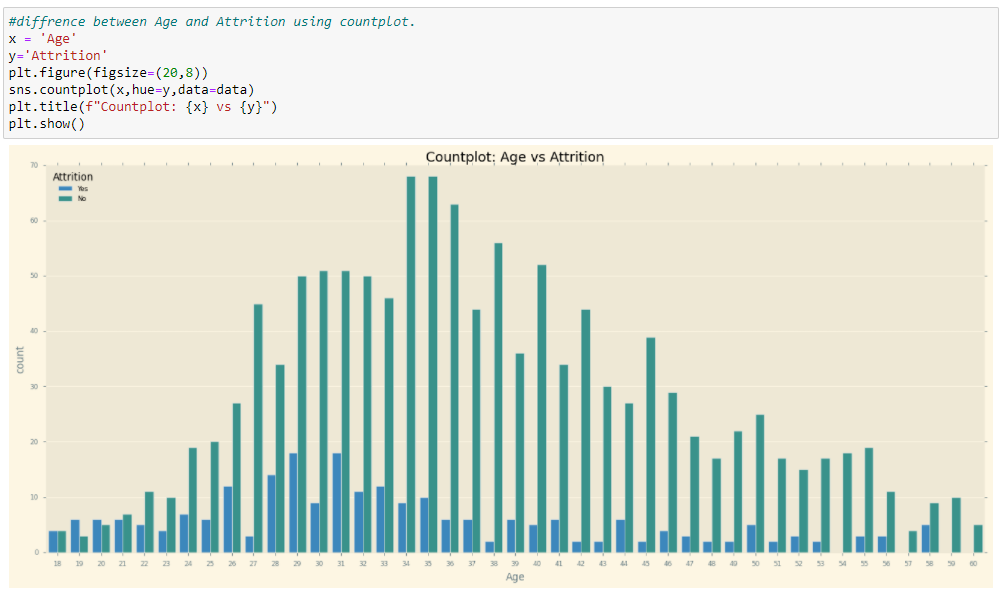
Bi-Variant:

The accurate visualization are been plotted in loop with violetplot.



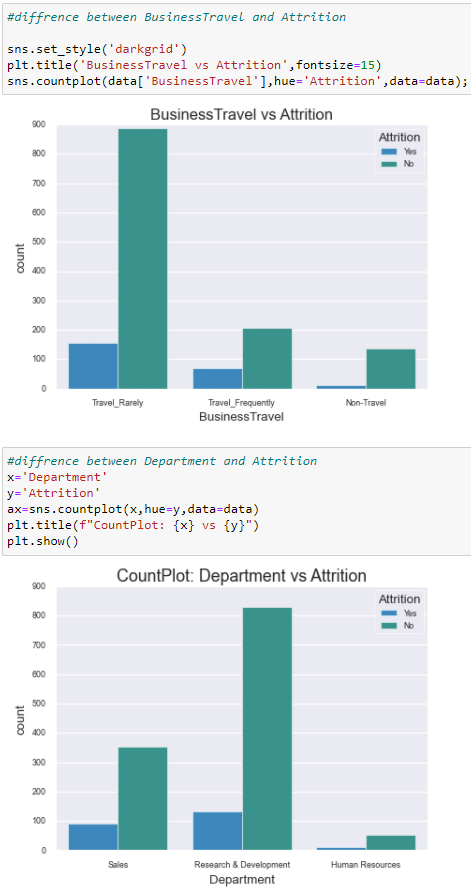
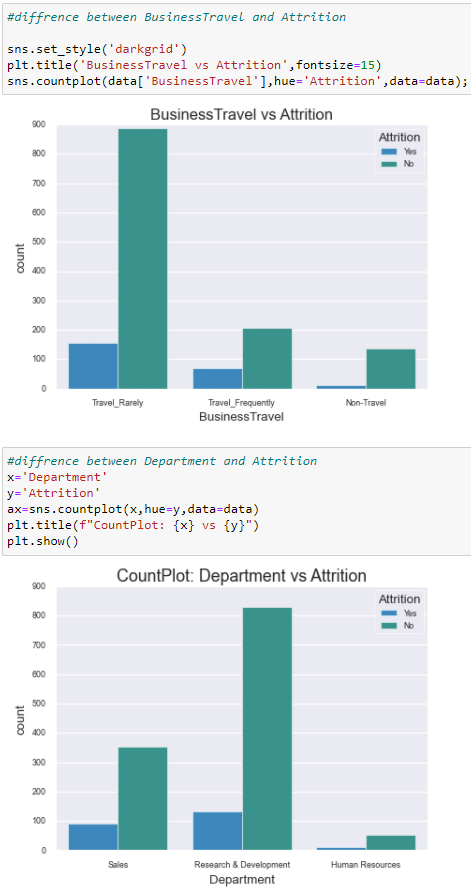


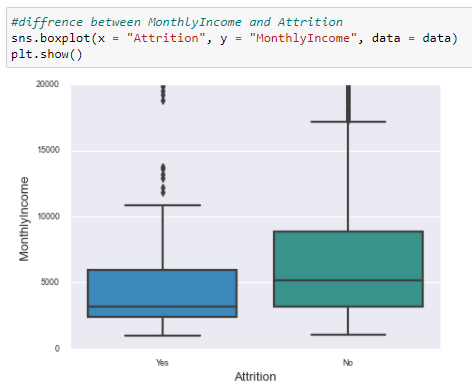




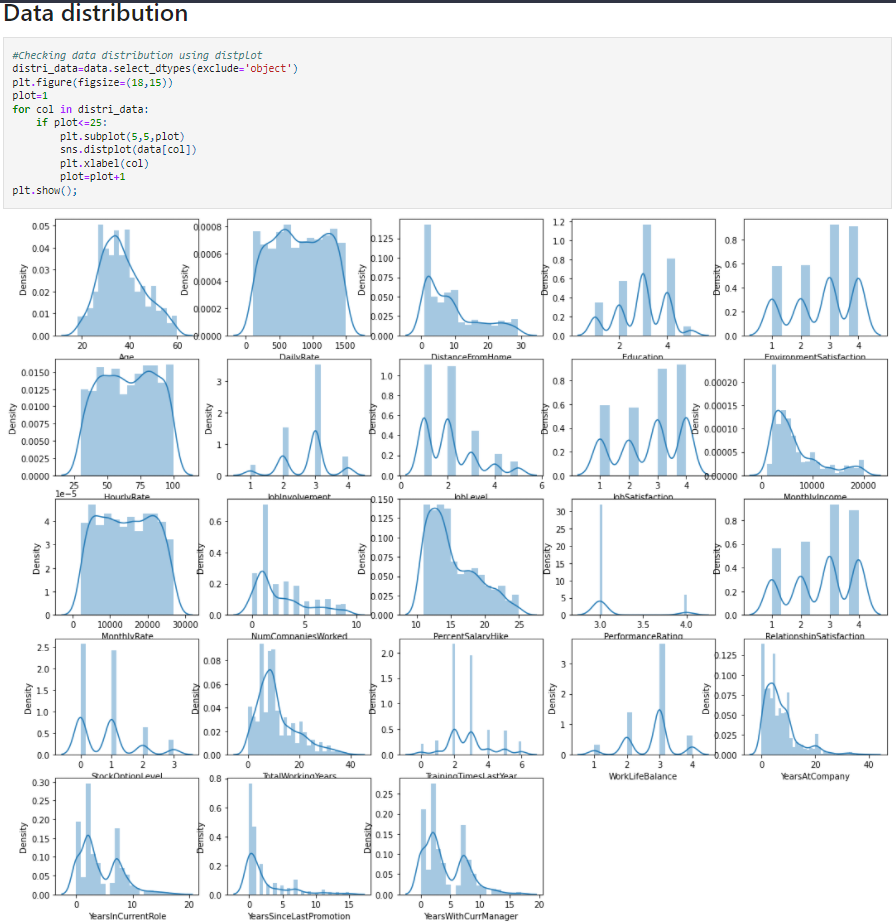
]]]]]

employees whose age group is between 27 & 42 has highest tendency to job change



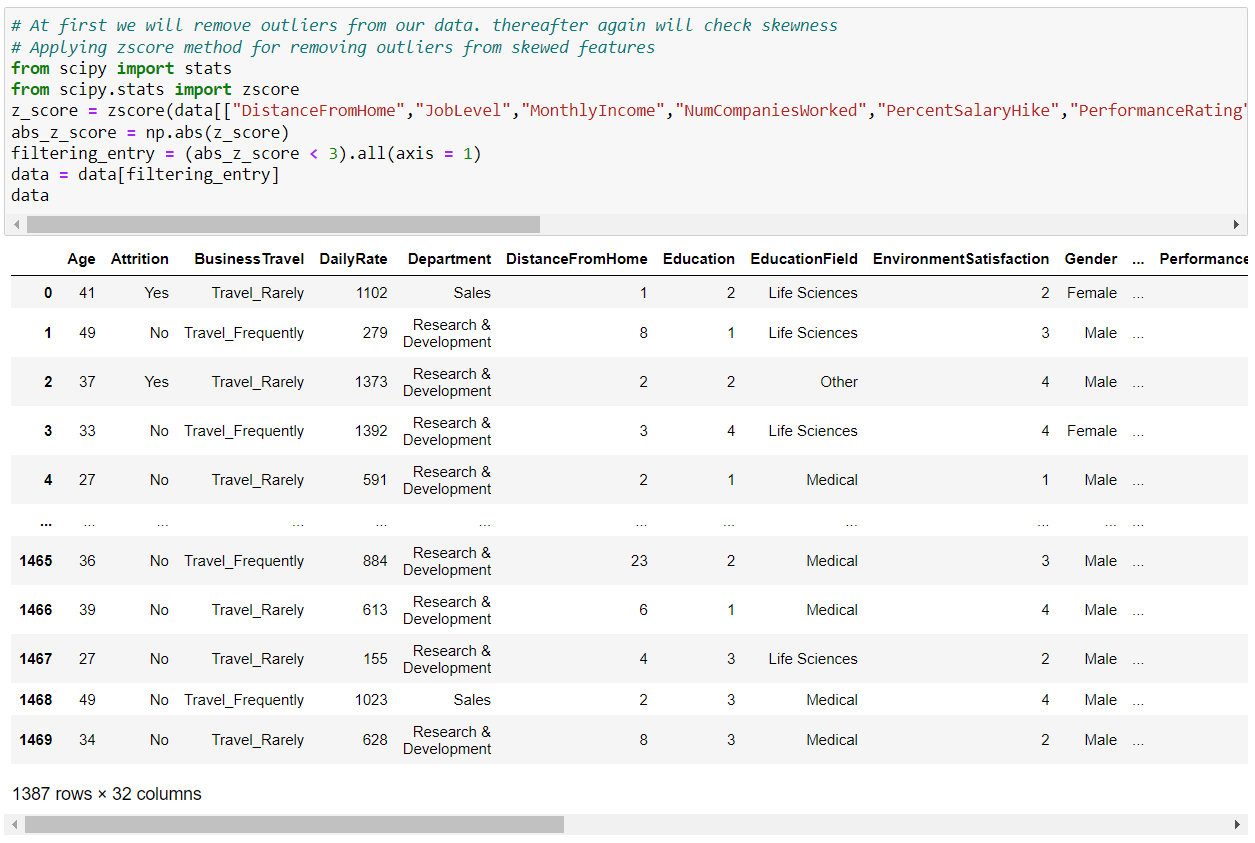


We saw so many graphs representing each other solely to the purpose of our better understanding. The more we create such graph the more we will get the whole purpose behind getting the prediction.

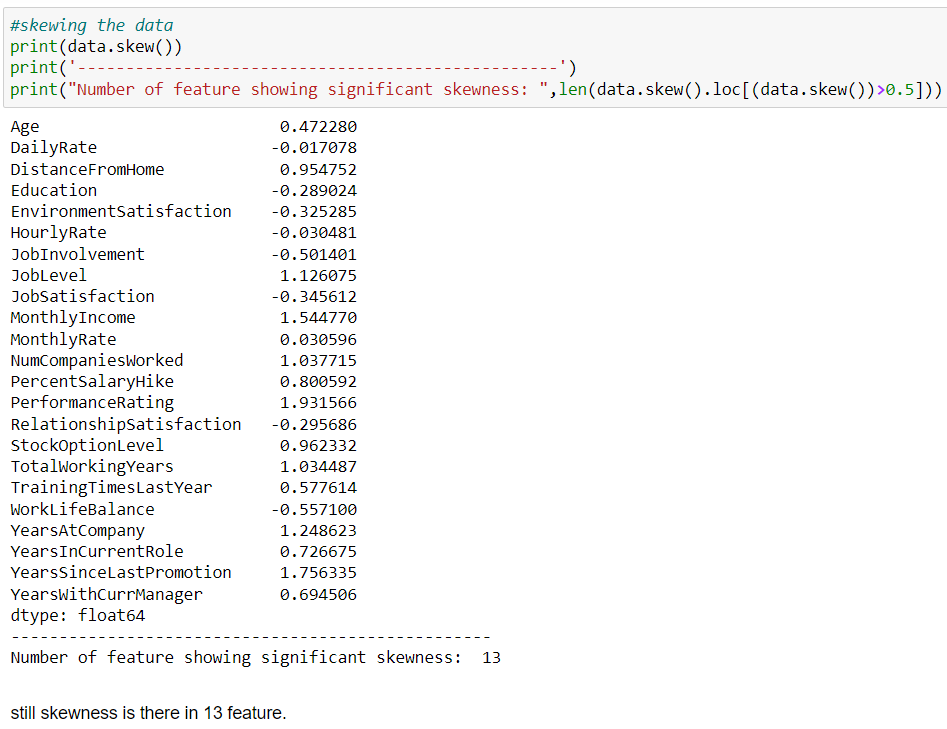


Except Feature Age which is somewhat distributed normally but not in well curve, all other features are not distributed normally which indicates data is heavily skewed and needs to be treated accordingly before sending for model training.

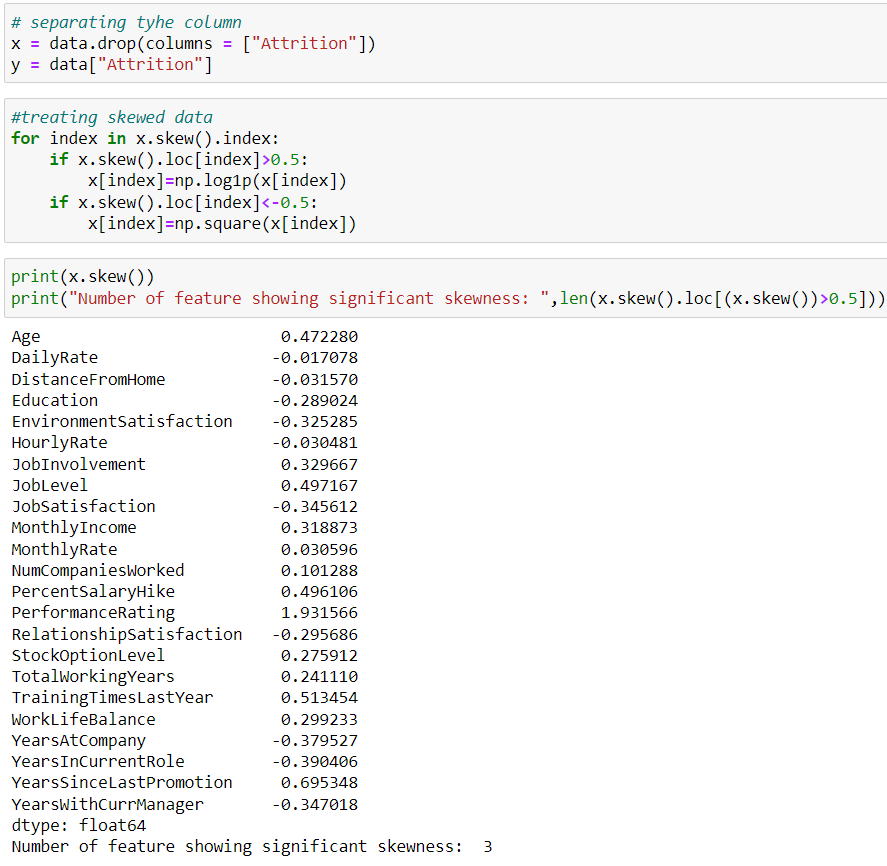
**Pre-processing Data:**

****In the pre-processing step I am going to tackle all the miss fits and fix them one by one starting with the problem that out dataset has object datatype values where as our Machine Learning models can only understand numeric values

After removing outliers from the data we are losing more that 5% data, we can't consider it is safe to remove outliers, we can't take this risk to lose the data. Then after we will check the skewness later. And in above figure we applied zscore method for removing outliers from skewed features.

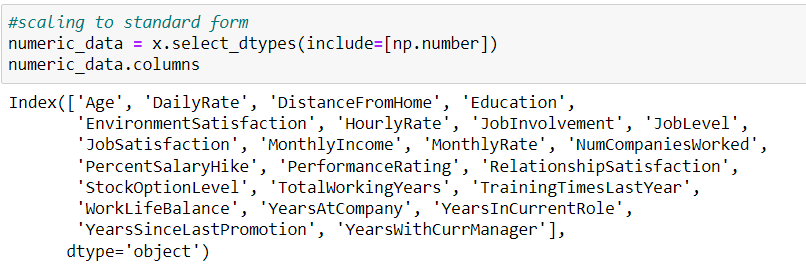
\

We skewed the dataframe and still there are 13 features with skewness.

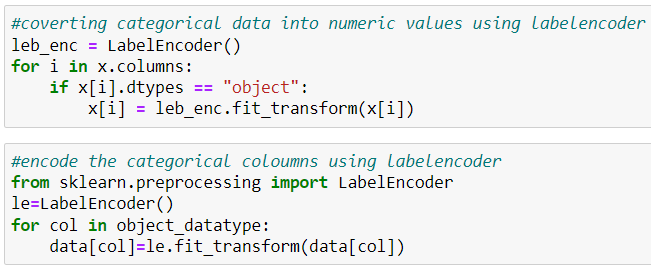
We have to reduce the skewness features as soon as possible and for that we have separate the columns.

After dealing with the data concerns I will then split our columns into feature and label. I am storing the feature columns in x and the target label column in the y variable and we skewed them again and checked a better result out of it.

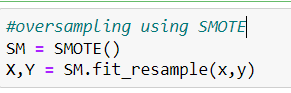
We can apply the StandardScaler to the attrition dataset directly to standardize the input variables.



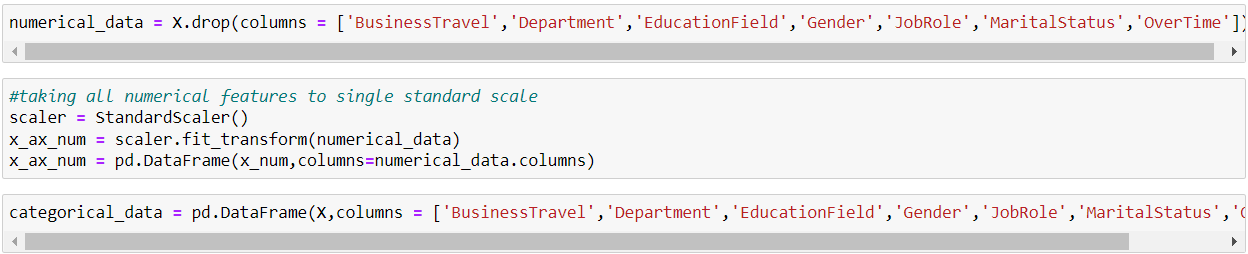
Then I will also scale the feature columns that is stored in the X variable to avoid any kind of biasness over column values. Some integers cover thousands place and some cover hundreds or tens place then it can make the machine learning model assume the column with thousands place has a higher importance when in real that won’t be true due to difference in unit range.



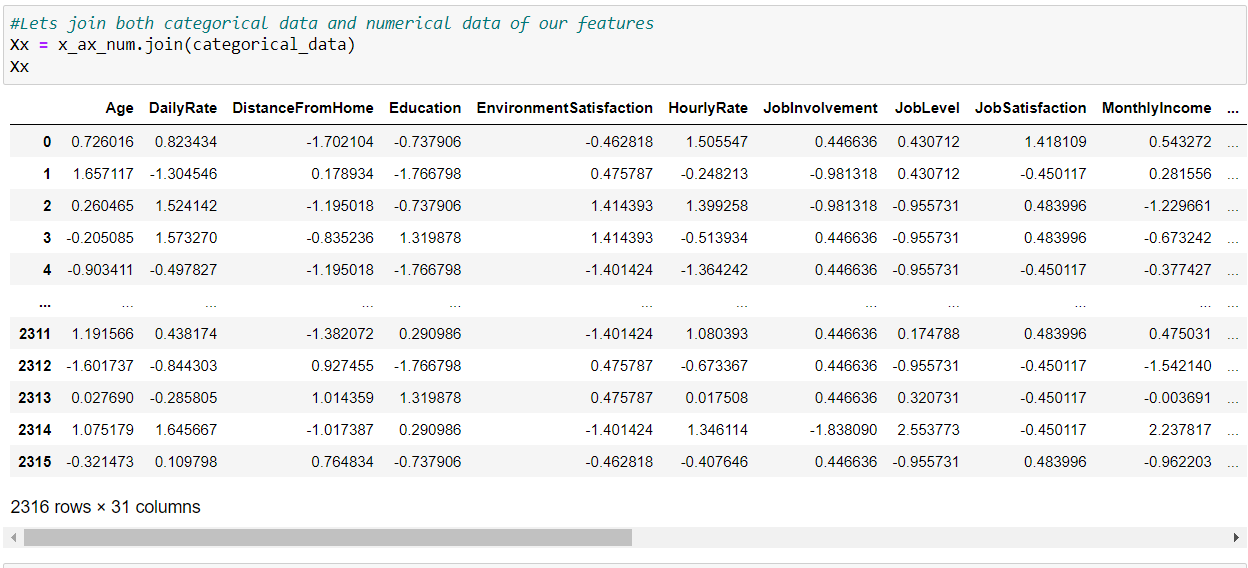
Now the most important thing is converting a categorical data into numerical values since we have to work on numbers for prediction because the object datatype will not be any of our use in prediction, prediction works on numbers not alphabets, so we will convert it numerical if possible using the LabelEncoder. Which is the best bet we can have right now in this situation. Then we encoded the categorical columns using LabelEncoder.



But there was an imbalance between the label classes. If you would notice the value displayed in the count plot earlier, there was a huge difference between the “Yes” and “No” data. Therefore, I will have to resolve it as the imbalance can make our machine learning model biased towards the “No” value.

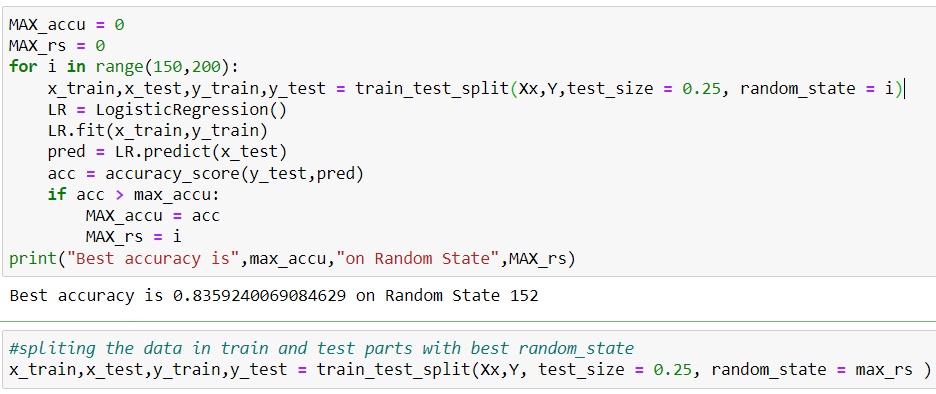


We dropped the object datatype, from the partion ‘x’, and named it in a variable as numerical\_data. And then again we took all the remaining numerical feature to single standard scale. And then merging the remaining feature into one

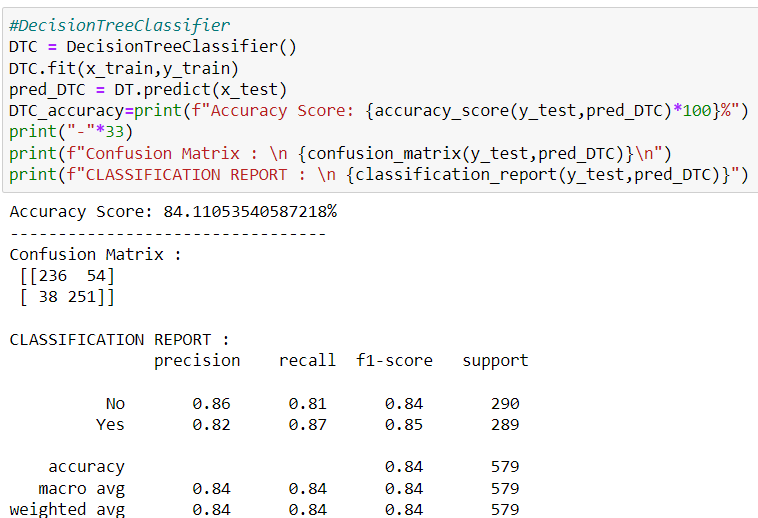


We joined them into one and displayed the dataframe after data-processing.

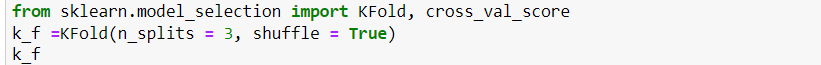
**Building Machine Learning Models**

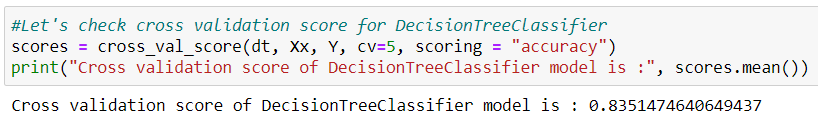
We have our final dataset. We now have to start modelling- Predicting the Attrition. Wait? Are you also confused like me? We already have the Attrition data then what is it here to predict? Well most of the time in Regression and classification problem, you run your model with the available values and check the metrics like accuracy of the model by comparing observed values with true values. If you won’t have the true values how you would know that the predictions are correct. Now you will realize that, how important the training data phase is. We train the model in a way that it can predict (almost) correct results.

It is always advisable to build more than 5 machine learning models so that you can choose from the best performing model and then apply hyper parameter tuning to make it perform even better. I am going to use the Decision Tree Classifier as my choice of classification model as I see it is doing better than the other models I used.



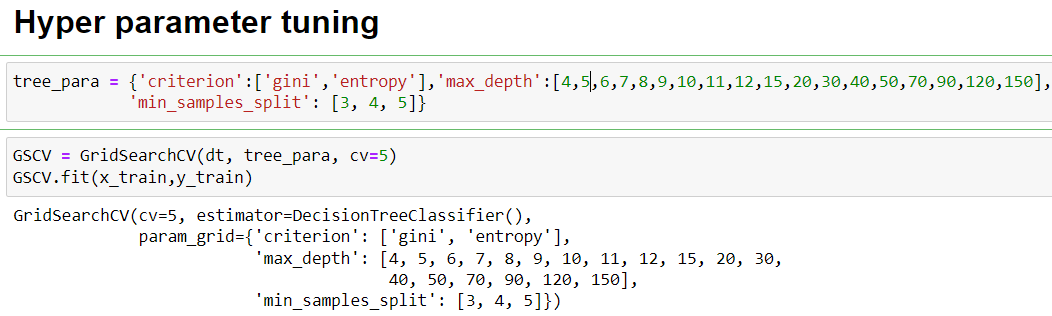
We have to check cross validation to figure out with model is the best.



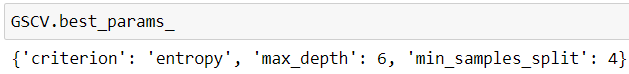




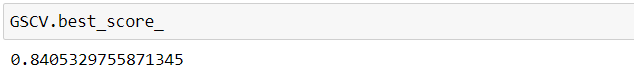


Lowest difference was with Decision and it was the best model we have.

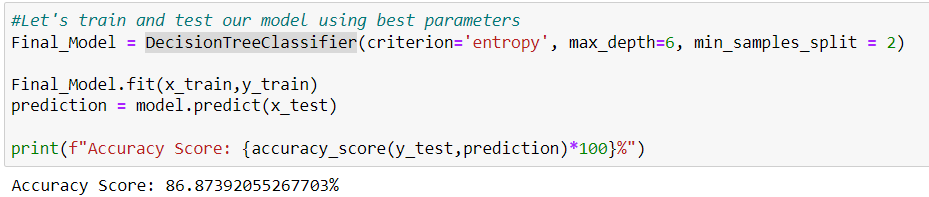
We gave some parameters to our hyper parameter tuning exclusively for Decision tree classifier.



Using the best parameter let’s see the output.

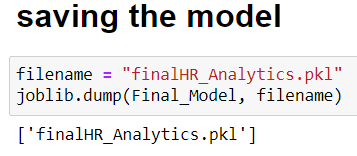


Let’s train the model and test it all up.



We came up with better result, it’s 86.9%

Now , we have to save our model for better use.



**Conclusion**

Throughout this post, we saw Data is important in Human Resource department (actually in most of places it is important). We saw how we can avoid using correlated values and why it is important not to use those while modelling. We used Decision Tree and learned how it can be very advantageous over other available machine learning algorithm. Most of all we found factors which are most important to employees and if are not fulfilled might lead to Attrition.

If you got stuck somewhere you can connect with me on LinkedIn, refer to this [Link](https://www.linkedin.com/in/pratyush-ghosh-a778621b6/).

#### About me

Hello everyone, currently I’m an under graduate student pursuing B.Tech Degree, specialisation on Electronics and Communication Engineering. I might be new to this data science world, since I came to about this way later but I’m trying n’ giving my best to quench my thirst towards data science.