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

****

**In this article, I have jotted down all the techniques in the form of sub-topics that I will be explaining one by one. And those pointers are as follows:**

* **Introduction**
* **Problem Statement (HR Analytics, Attrition in HR and attrition affecting companies)**
* **Dataset Link**
* **Data Analysis**
* **EDA**
* **Pre-Processing Data**
* **Building Machine Learning Models**
* **Concluding Remarks**
* **About the Author**

1. **Introduction**

In order to start with exercise, I have used IBM HR Analytics Employee Attrition & Performance Dataset, which was downloaded from GitHub. 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.

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 analyzing 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**

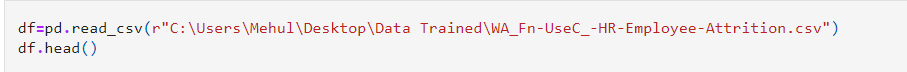
A major problem in high employee attrition is its cost to an organization. Job postings, hiring processes, paperwork, and new hire training are some of the common expenses of losing employees and replacing them. Additionally, regular employee turnover prohibits your organization from increasing its collective 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.

1. **Dataset link:**

https://github.com/dsrscientist/IBM\_HR\_Attrition\_Rate\_Analytics

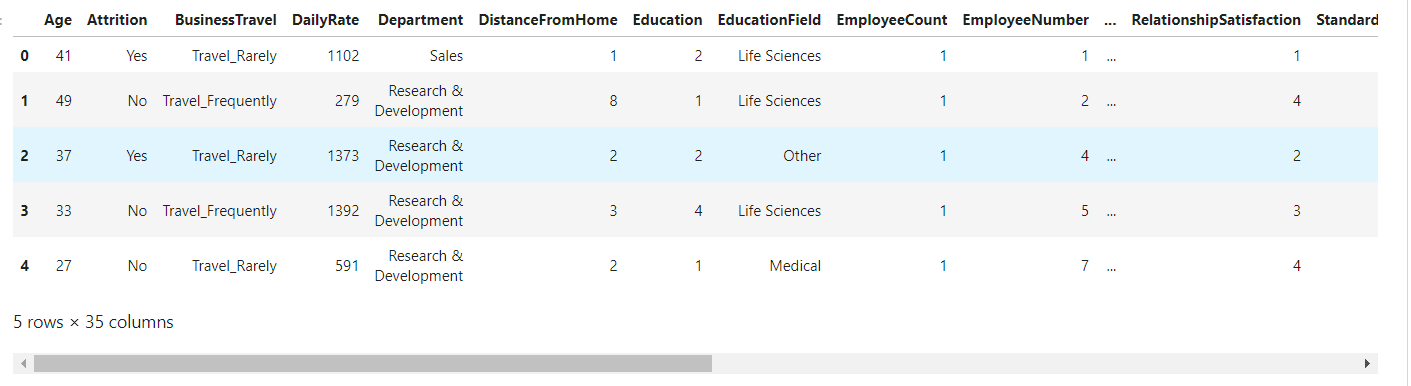
**Importing all the required libraries**

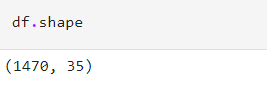
****

Before we begin with any process, we need to get the dataset in our Jupyter Notebook that can be achieved by a single step.  
  
****

This gives us our first five dataset stored in the variable name “df” for our dataframe.

1. **Data Analysis**



****

In the above line of code, we can see that the total number of rows present in our data is 1470 and the total number of columns are 35. Since, it is a dataset with reasonably higher number of rows and columns the visualization gets truncated.

1. **Exploratory Data Analysis (EDA)**

* **What is EDA**

**Exploratory Data Analysis, or EDA**, is an important step in any Data Analysis or Data Science project. EDA is the process of investigating the dataset to discover patterns, and anomalies (outliers), and form hypotheses based on our understanding of the dataset

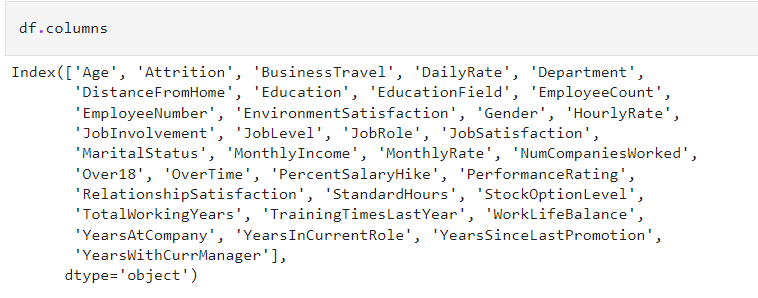
* **Purpose of EDA**

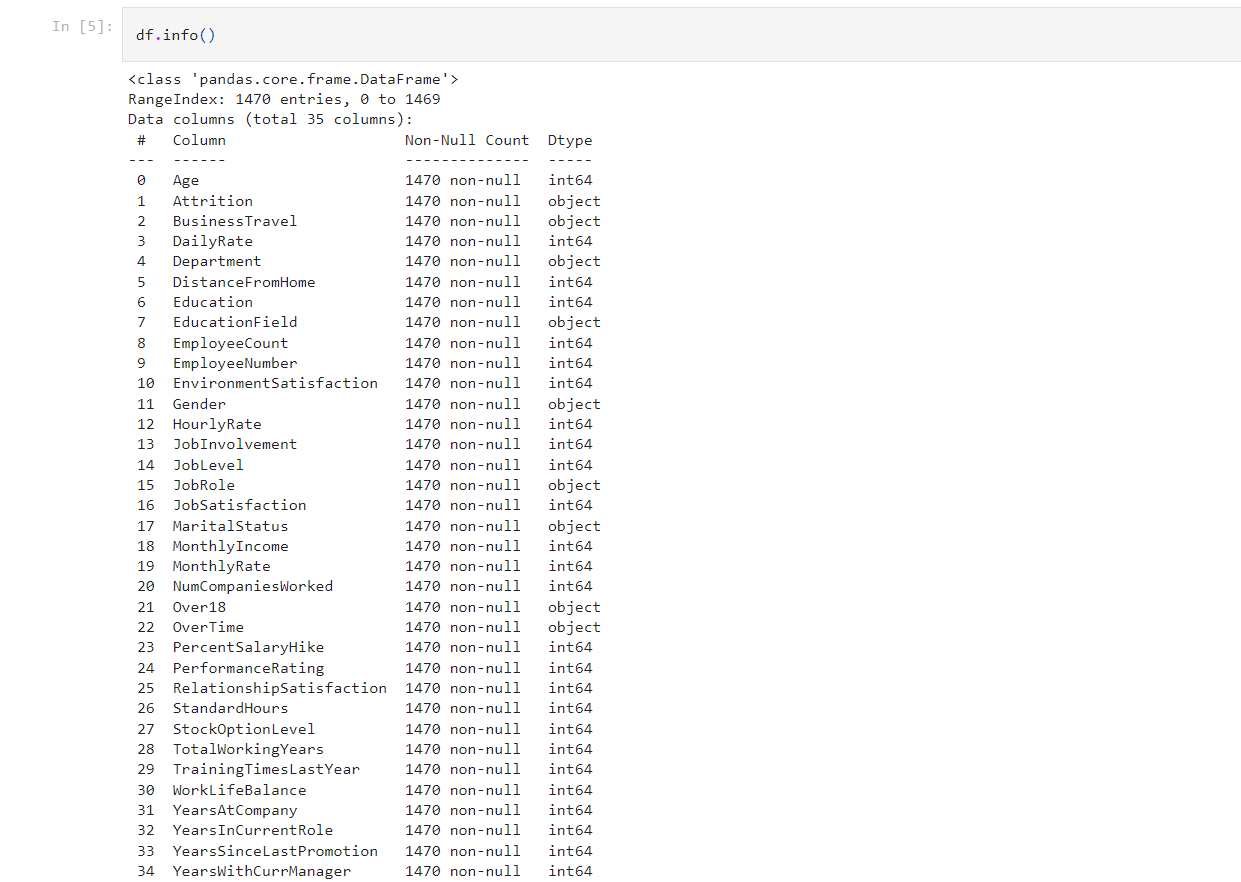
The primary goal of EDA is to maximize the analyst's insight into a data set and into the underlying structure of a data set, while providing all of the specific items that an analyst would want to extract from a data set, such as**: a good-fitting, parsimonious model.**

* **Types of EDA**

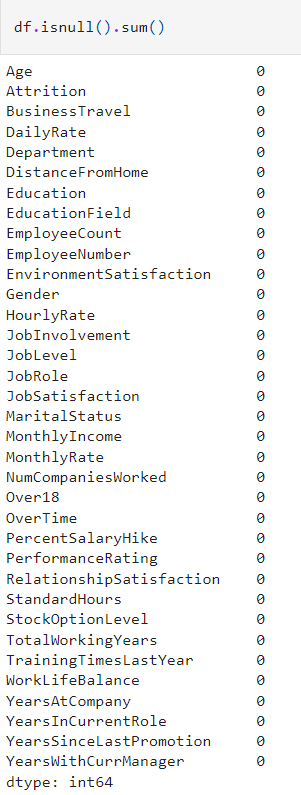
The four types of EDA are: **univariate non-graphical, multivariate non- graphical, univariate graphical, and multivariate graphical**.

These two codes below give us the name and list of all the columns marking them to be either integer, float or object datatype depending on the values present inside the columns.



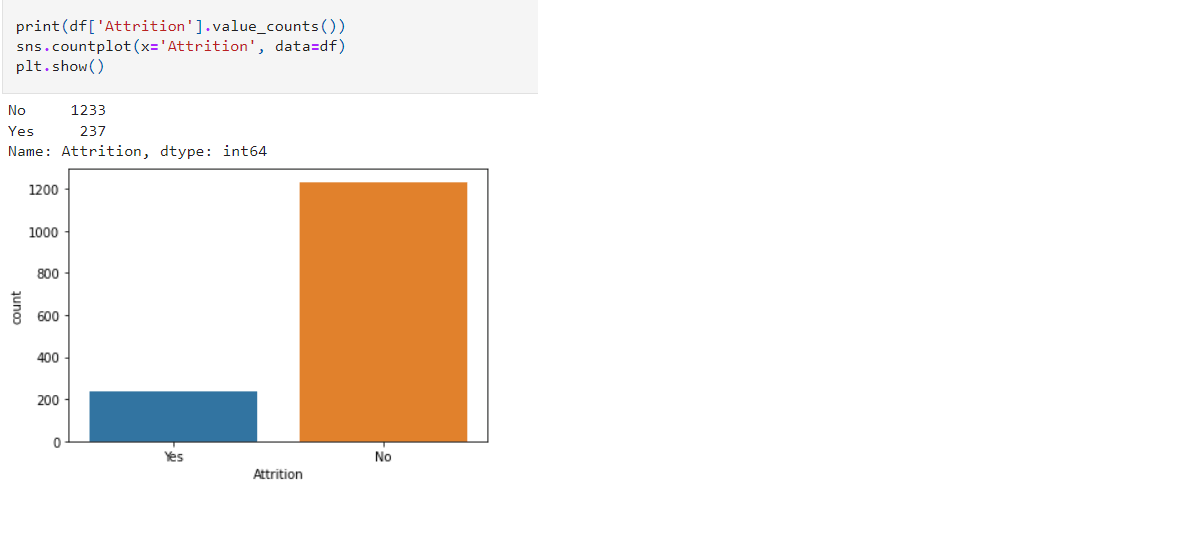


This code gives us the missing values information in a tabular format that looks something like this.

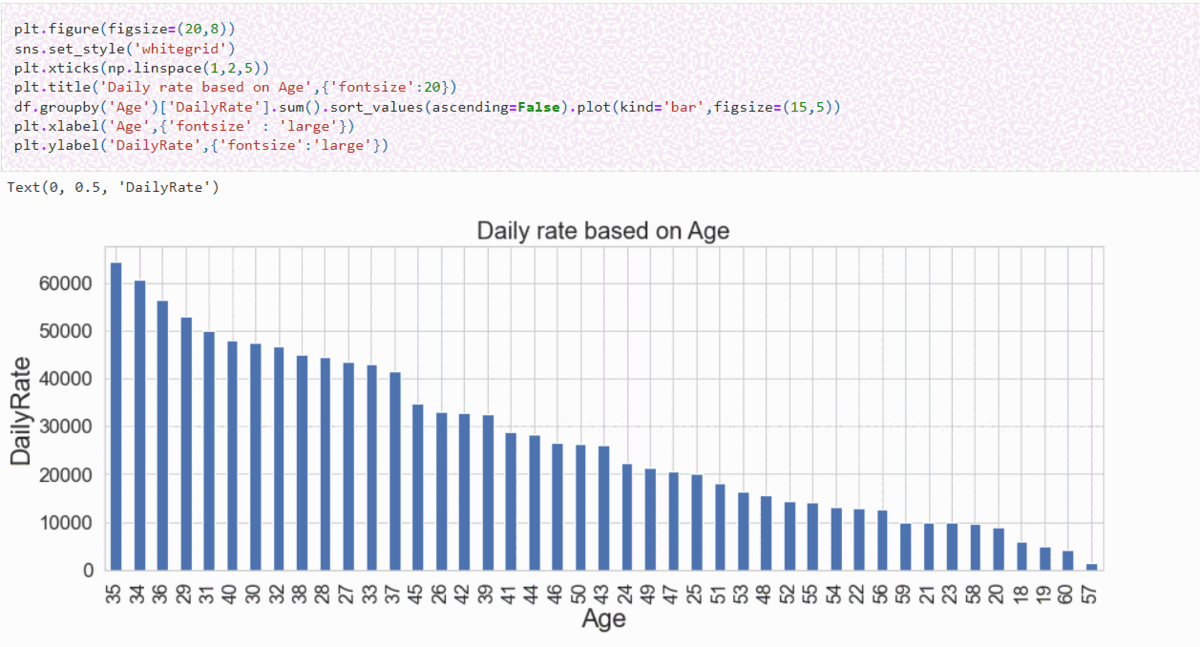


Now that we were able to confirm our dataset being free of any missing data, we will drop any duplicates that might be present using the code below.

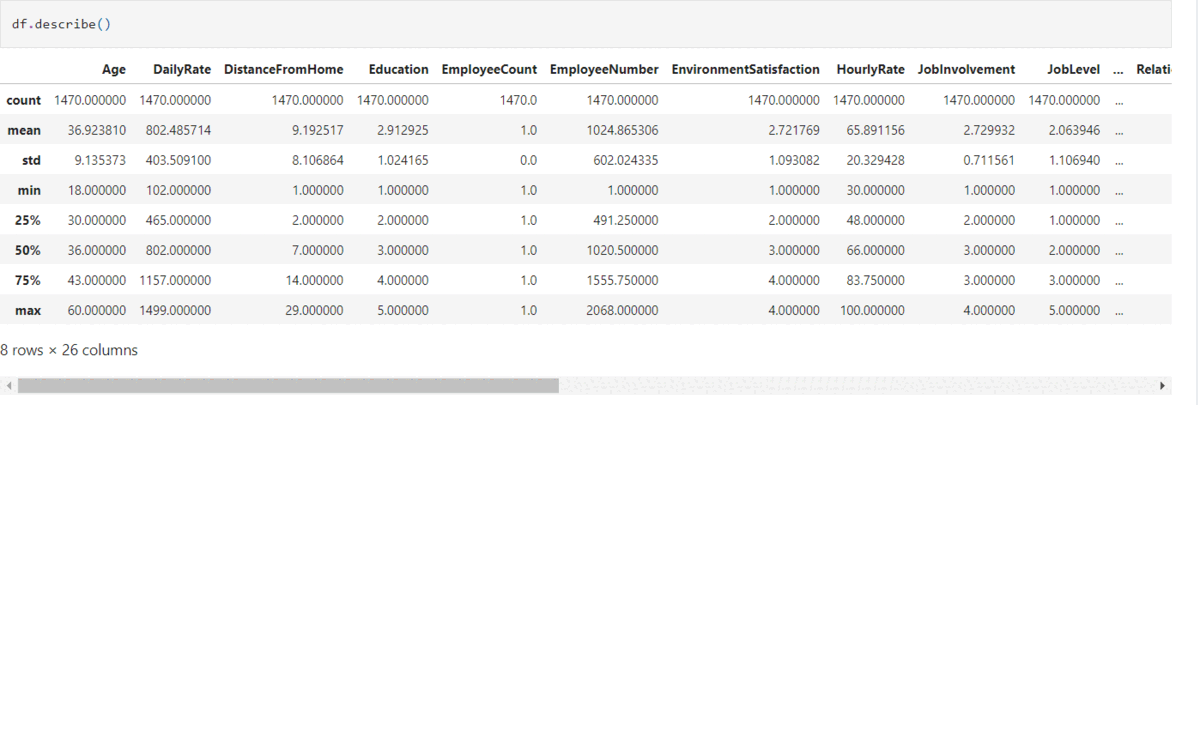
* **Univariate Analysis**



* **Bivariate**



Next, we move on to using the describe method to take a look at the count value, mean data, standard deviation information and the minimum, maximum, 25% quartile, ignored. Take a look at the below code and you will get an idea on how to use it.50% quartile and 75% quartile details. As the describe method works best for numeric data all the object (text) type data gets.



Once we have used the code the output provided are format to accommodate all the columns from our dataset in tabular as well as visual format.

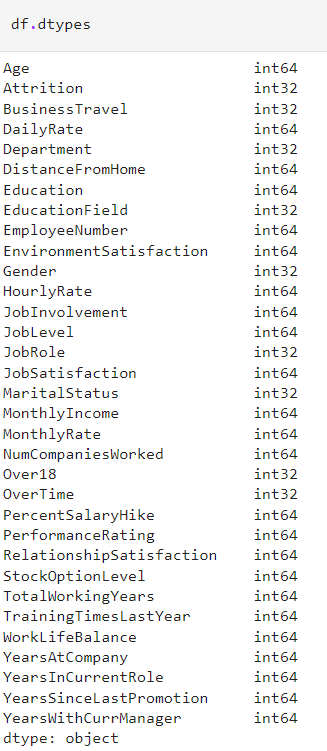
1. **Pre-Processing Data**

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. I am making use of the encoding methods to convert all the object datatype values. For our label I am using Label Encoder.

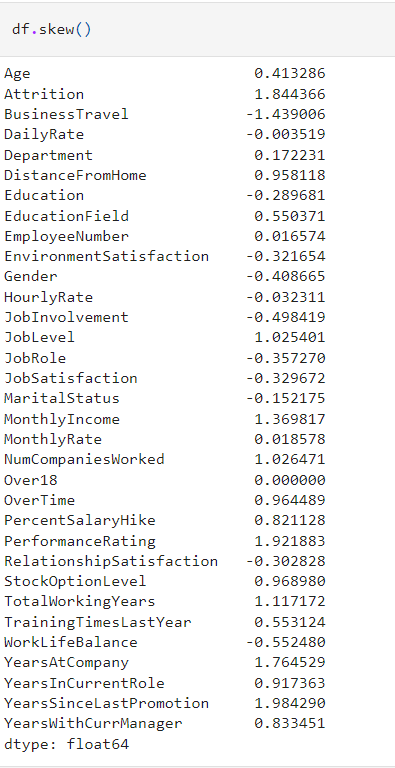
* **Converting Categorical data into Numerical data through Label Encoder**

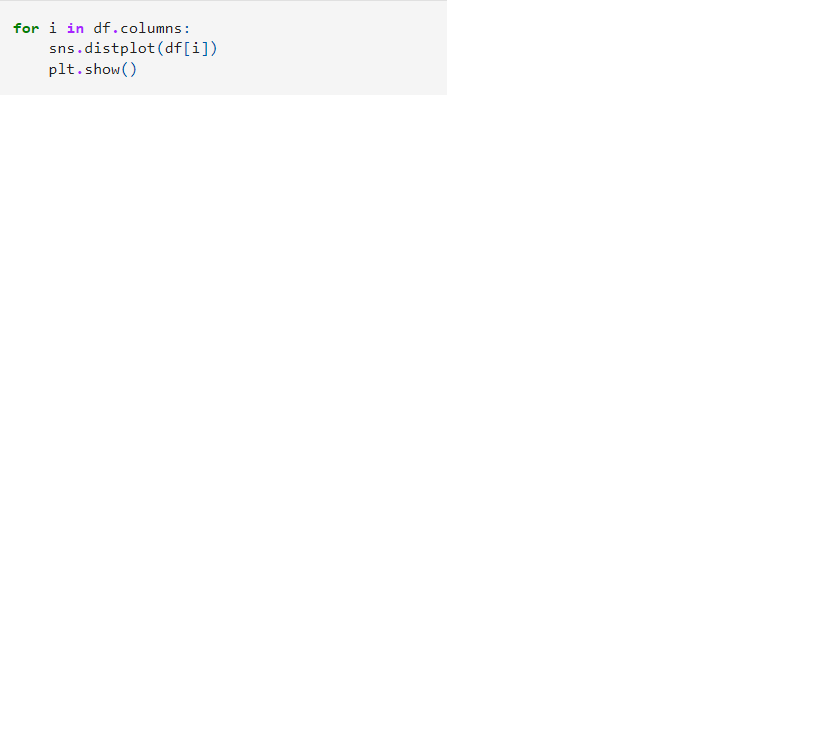


* **Checking the Datatype**

****

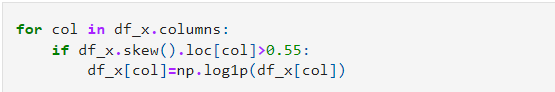
* **Checking Skewness**

****

****

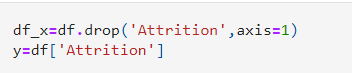
* **Treating Skewness using Log Transform**

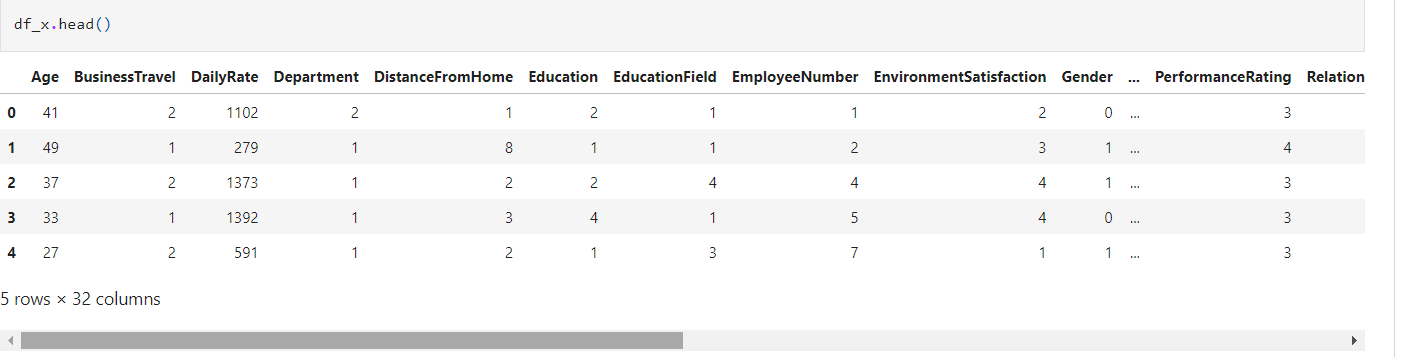
I am using the Log transformation to deal with the skewness since the acceptable range lies between +/-0.5 value for each column.

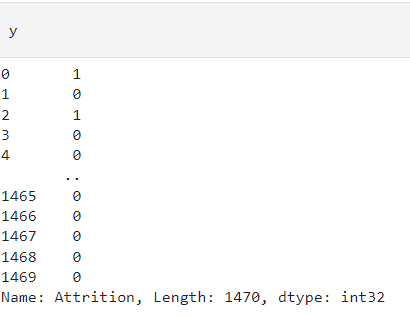
****

* **Splitting X and y**

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.

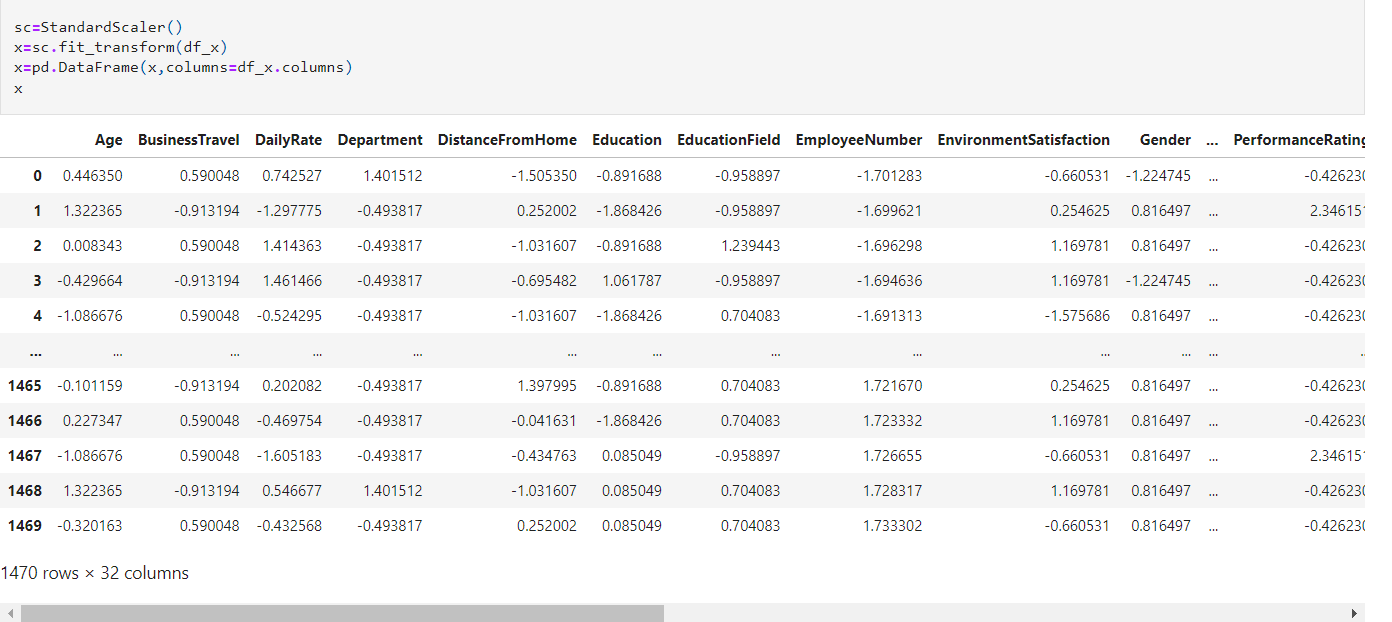




****

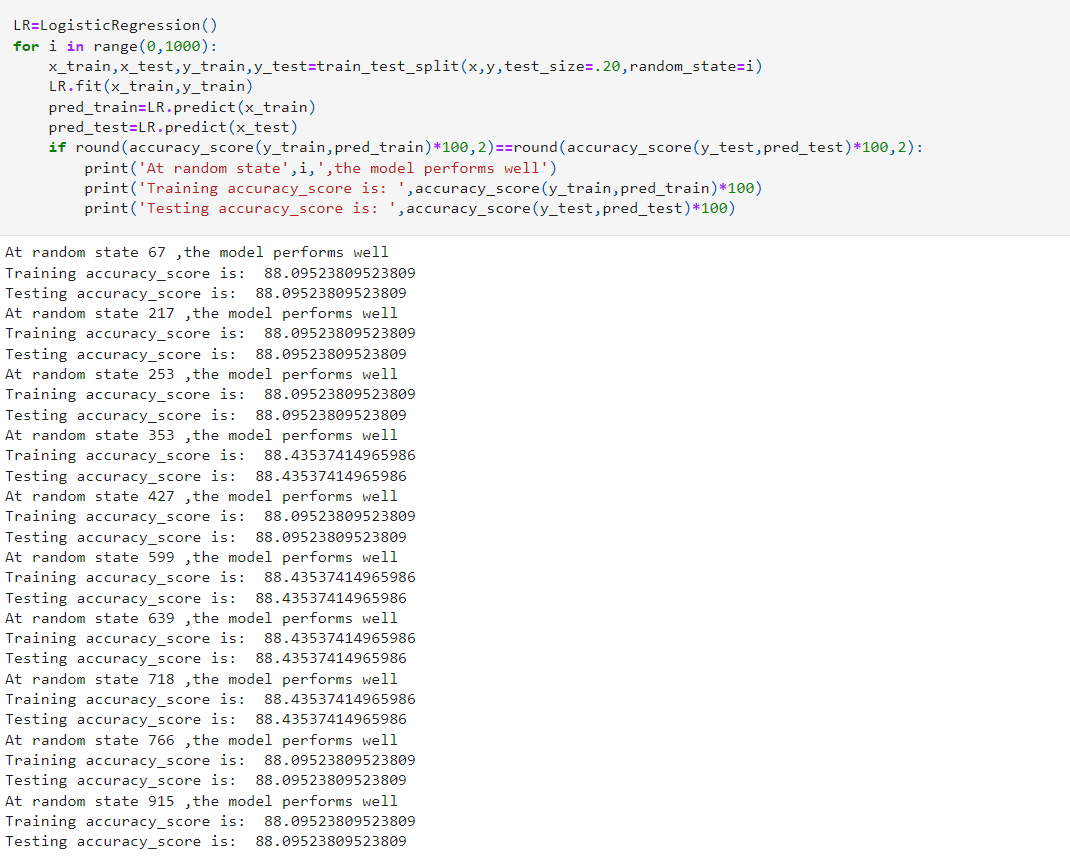
* **StandardScaler**

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.



1. **Building Machine Learning Model**

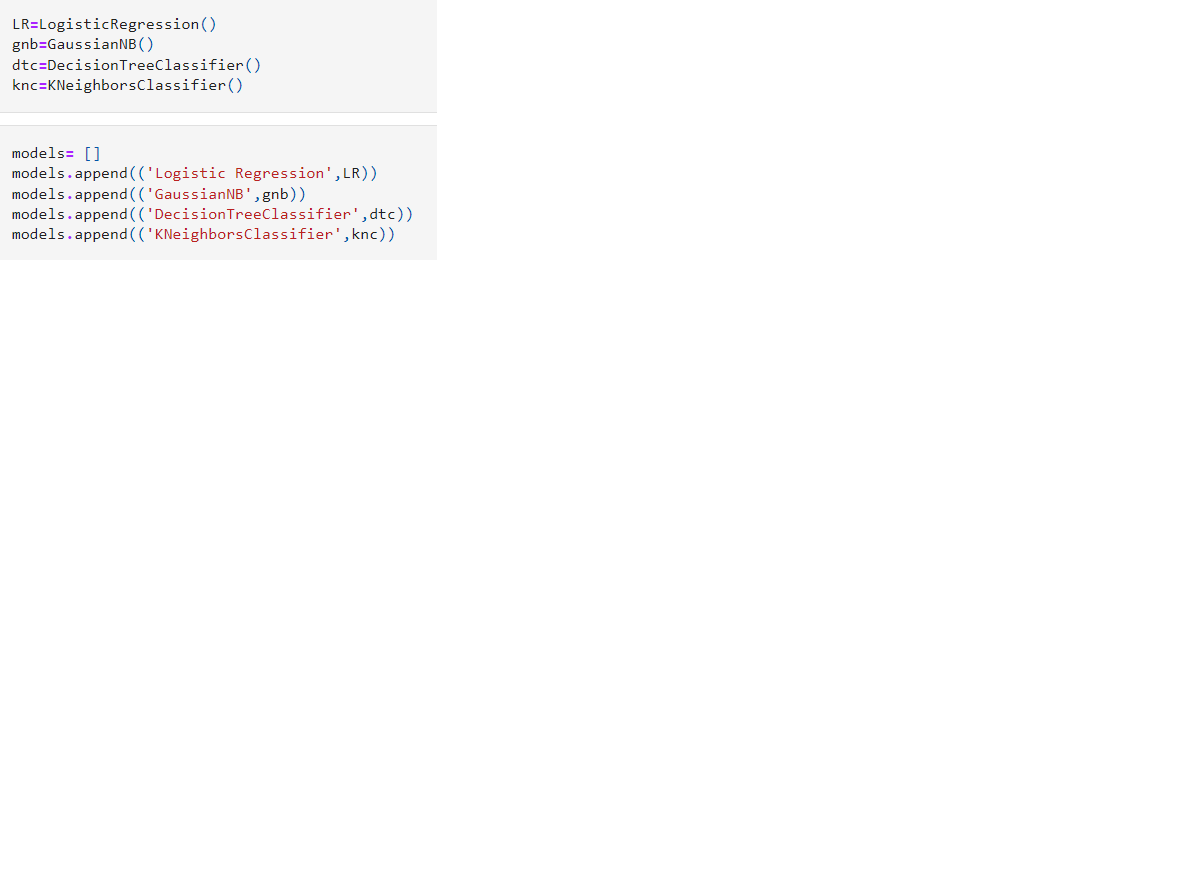
I would like to share a simple piece of code that allows us to choose a fitting random state for the machine learning models.



Then I will use the train test split to bifurcate our entire data set into training data and testing data. Here I am using 75% data for training purpose and 25% data for testing purpose. Some people provide training and test data separately as well and hence it completely depends on you how you want to use this step.



* **Finding out the best model**



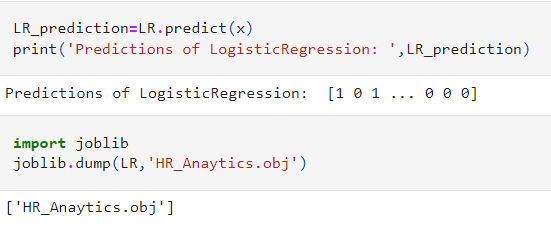
* **Hyperparameter Tuning**

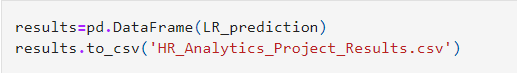
It is always advisable to build more than 5 machine learning models so that you can choose from the best performing model and then apply hyper parameter tuning to make it perform even better. I am going to use the **Logistic Regression** as my choice of classification model as I see it is doing better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

****

* **Ensemble technique for boosting the score**

****

* **Finalizing the Model** 
* **Saving the Model**

****

1. **Concluding Remark**

Kindly allow me to provide a quick recap on all the steps that we went through starting from understanding the Problem Definition then going through the Data Analysis and EDA processes. We went through the necessary Pre-processing Data steps before the final Building Machine Learning Models step came into picture.  
  
What I do is code my entire project on my own and then take a peek at the internet to look through other’s coding style for inspiration and understand if I can incorporate anything to improvise further on accuracy or beautify the visuals. However, I have seen many people doing the complete opposite whereupon they don’t practise or create their own unique coding style first and rather copy paste lines from the web and perform some sort of messy patch work and when asked to explain might not be capable of conveying functioning or usage of those code blocks.

**Throughout this project, 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 Logistic Regression 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**.

Created a self-made unique data story telling commandment list and follow it along with the standard project life cycle. Hope this at length article helps you in gaining the initial knowledge on building your first project from scratch

1. **About the author**

Pursuing Post Graduate Diploma in Data Science, Machine Learning and Neural Network. I am an aspiring Data Scientist whose purpose is to learn in detail all the concepts needed for Data Science. I am passionate about Data Science and have skills that help me derive valuable insights from data, such as Data Manipulation, Data Visualization, Data Analysis, EDA, and Machine Learning. I am also open to get some feedback from anyone that will help me in improving too! The content that I have written is solely my view of the project but it’s definitely inspired by others over the internet who have worked on similar projects before me.