**HR Analytics Project**

Let’s understand what is the problem statement and what steps we took to complete the prediction.

**Introduction:-**

Companies hire new employees every year and spend lot of money and invest time on them to train. It is not only company provides training to their new employees but existing is also trained for effective delivery and to increase the productivity timely. “**Attrition is also an important parameter in a company and it is being reviewed rigorously by the top management on regular interval”.**

A question must be arising in the minds that how “HR analytics” helps in improve employees’ efficiency. HR analytic has significant role in process improvement. We will read further the relation and the contribution of HR analytics in our coming sections. Please read full article to get complete understanding.

This article is containing the following sub-topics

1. Problem Definition
2. How attrition impact the business
3. How to HR analysis help in understanding probable attrition case
4. Data Analysis

A. Understanding the data

3. EDA Concluding Remark

4. Pre-Processing Pipeline

5. Building Machine Learning Models

6. Concluding Remarks.

**Let’s get into details sub topic wise and understand:-**

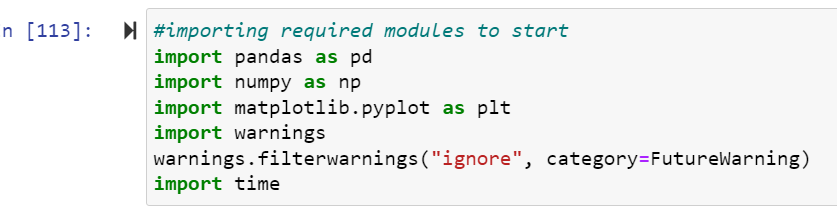
**Problem definition:-**

As we have already read in the introduction part that companies spend money and invest time to trained new hires and run many training programs internally for existing employees to enhance their work efficiency subsequently.

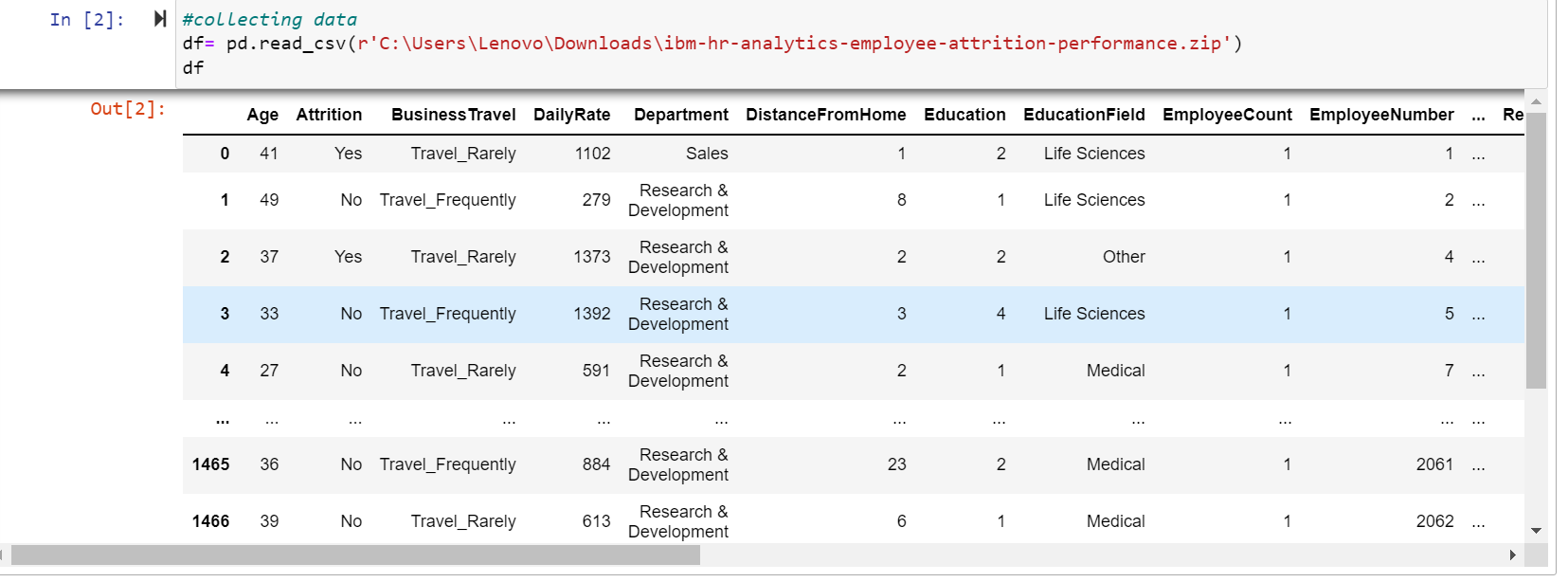
Attrition is considered a major & critical parameter in companies. It is said that old is gold and I believe this example fits perfectly in this case, experience brings lot of good work for the companies. HR conducts timely survey and basis on that gauge employee’s satisfaction and also they conduct many engagement activities to make employees feel good at work. Basis on survey data and historical attrition data, HR analytic brings many innovative idea and initiatives to control the attrition. Let’s get into deep and see how HR analytics provide insight

**Importing libraries:-**

For analysing data, we would load dataset by using pandas’ read\_csv function , if you want to refresher your idea about pandas, please visit pandas official site and documents. we are also importing important libraries which will help in analysing and model building.

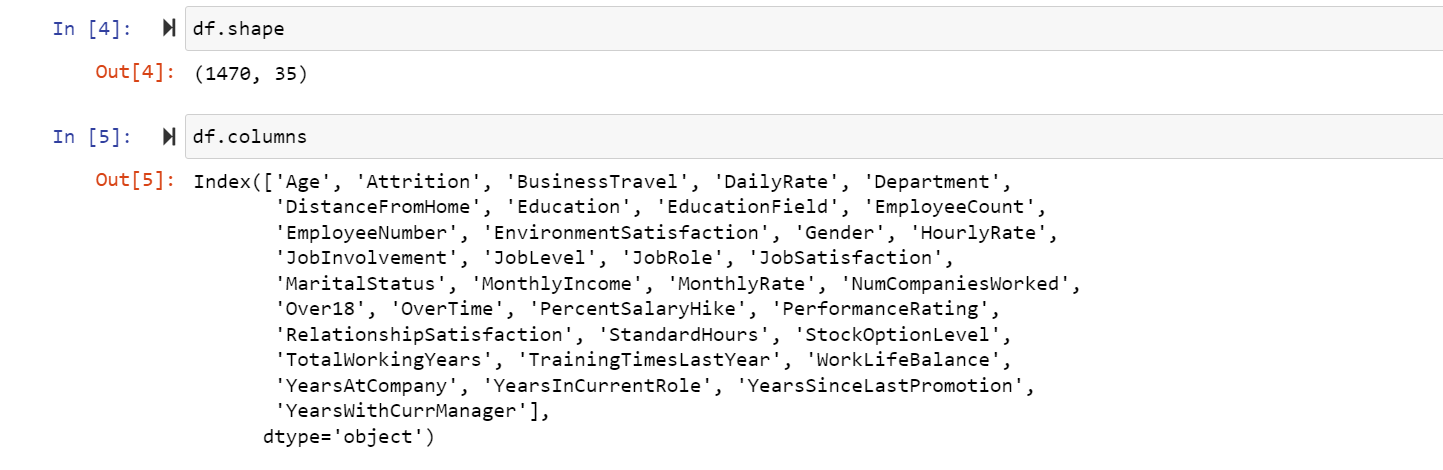


Once the basic libraries got imported to start the EDA process. Dataset called in the jupyter notebook by read\_csv.



Now, we will check the details of dataset like how many rows and column it has, and is there any null value in the data. Also what data type the data set have.

We will do the prediction after checking and correcting the data for prediction process.



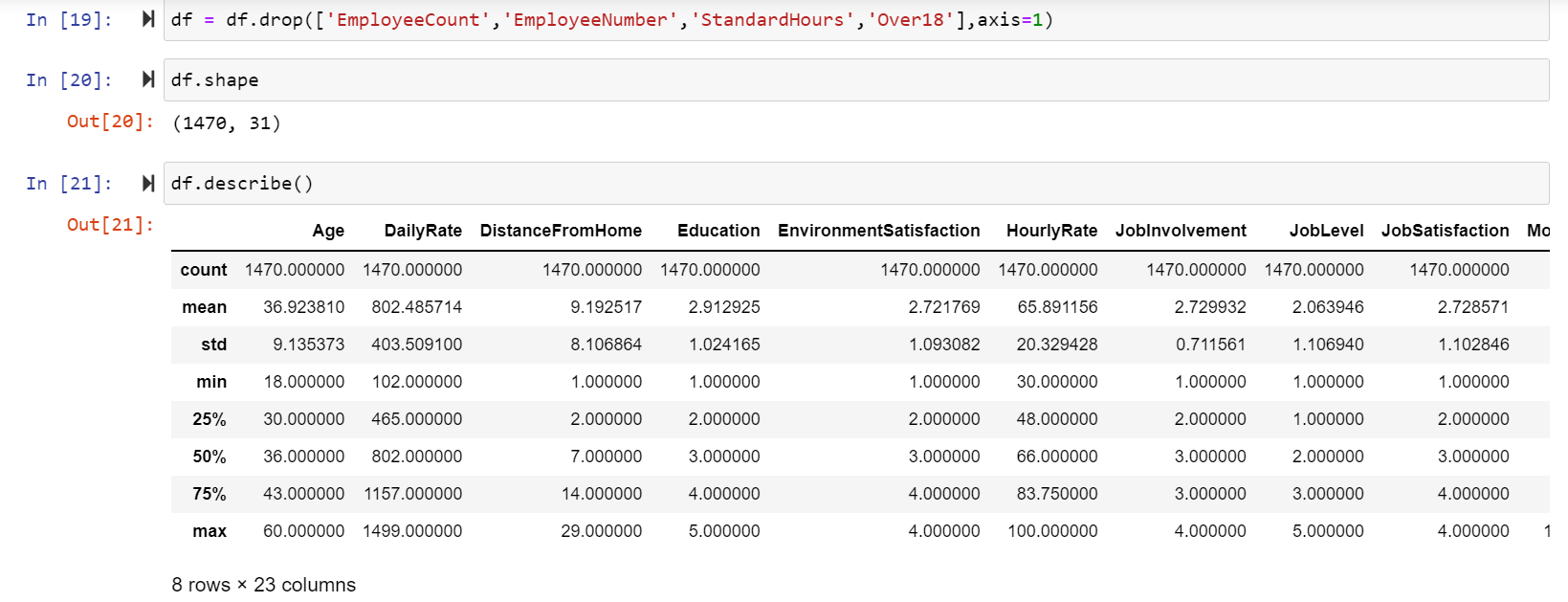
It shows that the data have 35 columns with the names mentioned and 1470 columns.

We can check and remove those columns which do not have any role in predicting the data.



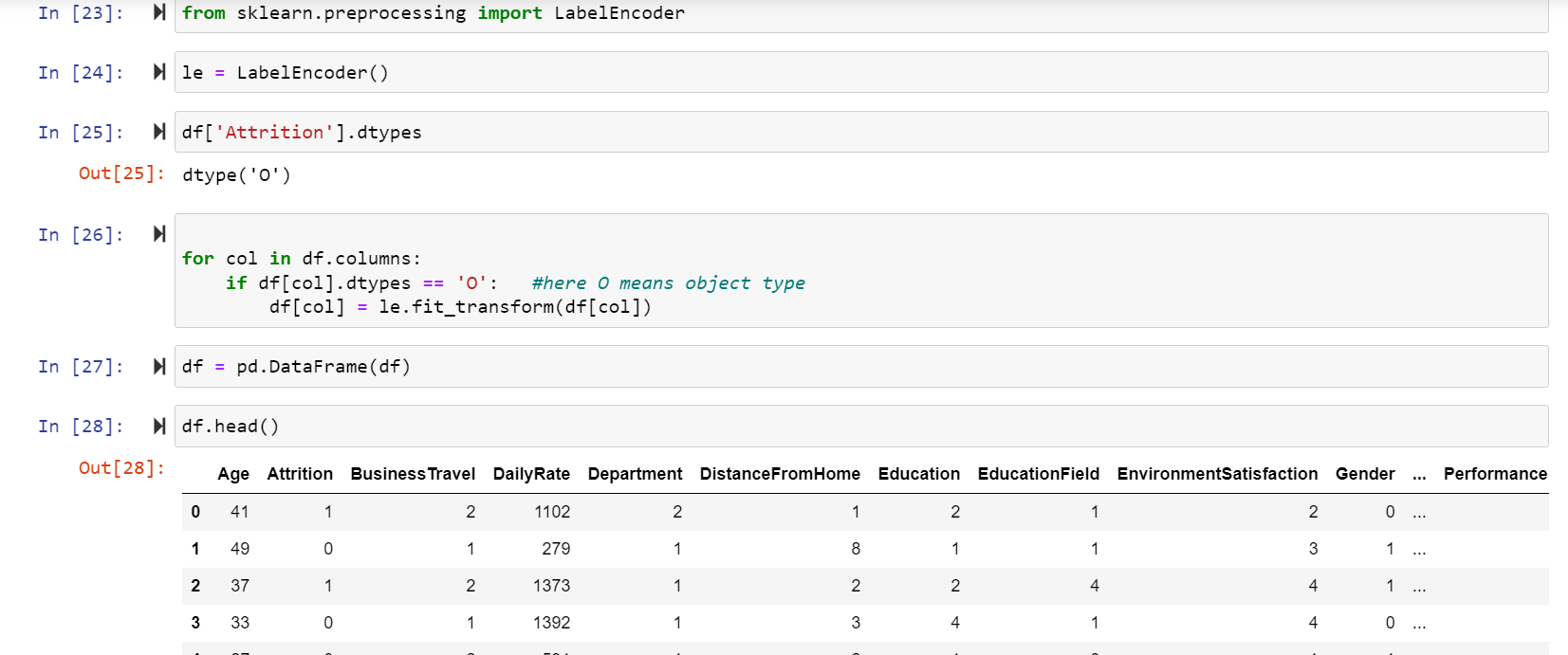
Here I will be removing 2 columns EmployeeCount and EmployeeNumber, as they are not affecting the change in the number of Attrition, they are just used to refer an employee uniquely and also removing Standard hours, Over18, because there is no variance in these columns, 80 hrs are fixed value for all the employees, and all employees are Over18 so it will have null effect on our model.

**Now the new data on which we will work have different shape. Let’s check that.**



We have some object values as well which we need to change into int to perform EDA.

**We will use LabelEncoder for that.**

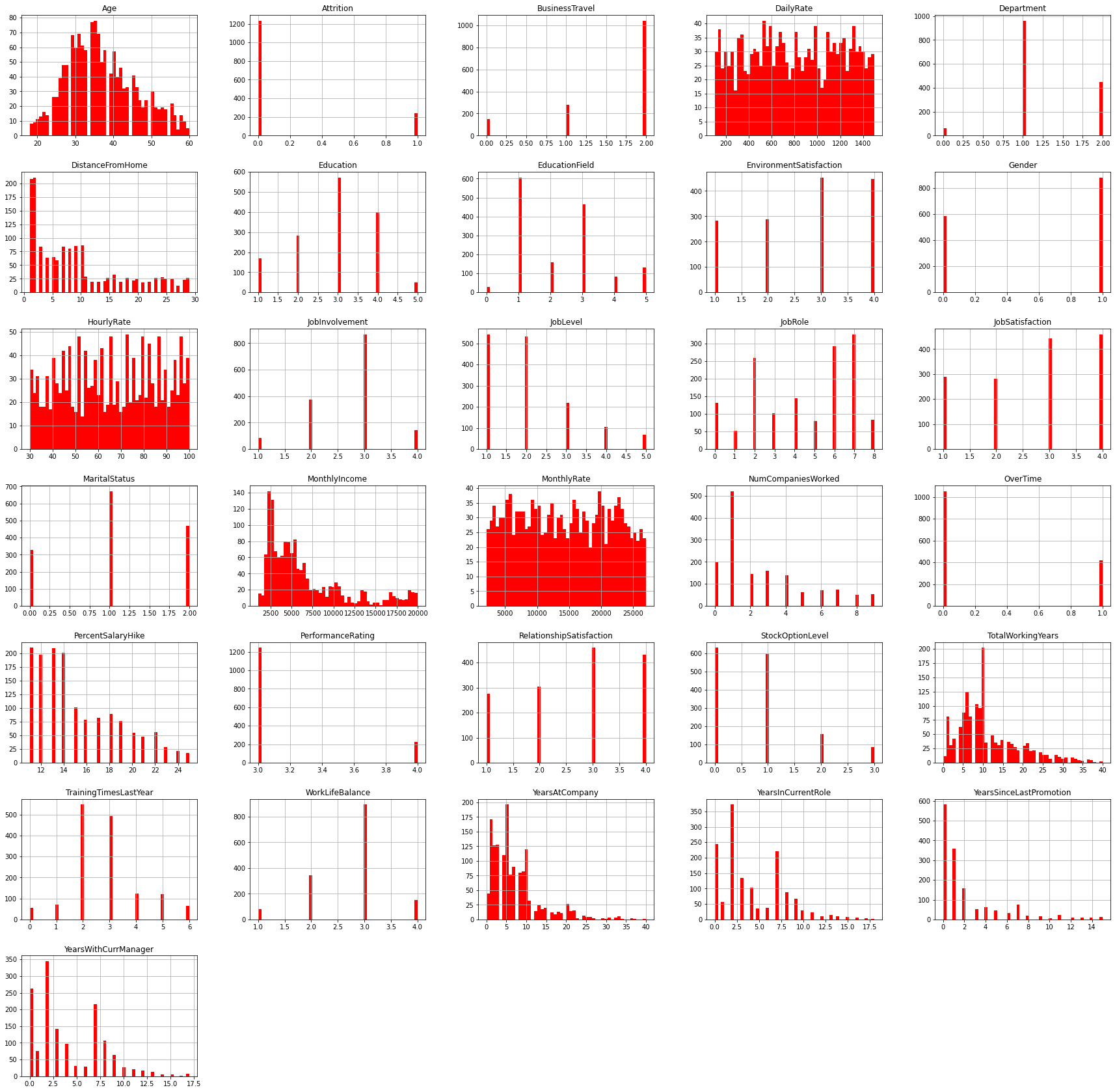


Here we can see, all columns have int values now. We can perform the data visualization.

**Data Visualization**

Now we will check the data with single and double variant to check how they impact with the output.

**Univariate Analysis**

Let’s check the relation of some columns here as an example:

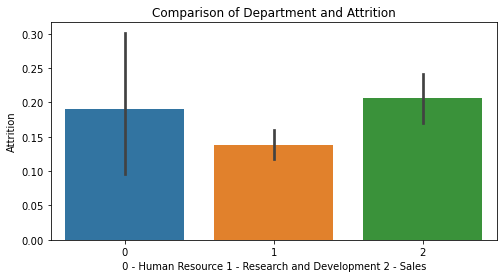
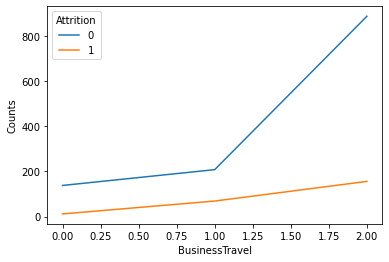
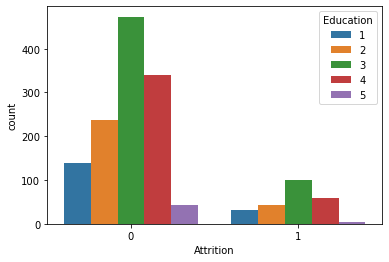
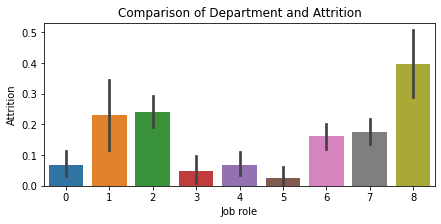
1 - AGE: As the age of an employee increases after a limit the workforce diminishes, there could be many reason for that like Retirement, Health Issues etc.

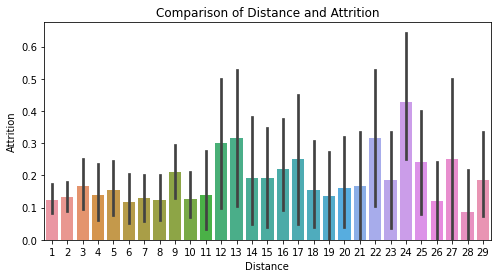
2 - Attrition: Here we can see that the number of employees leaving the company is lower. We will be analysing later the reason of Attrition. Attrotion 0 = No , Attrition 1 = Yes

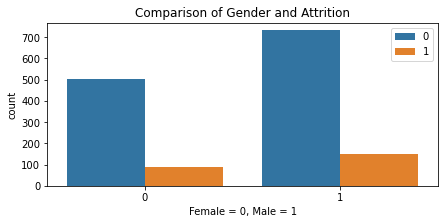
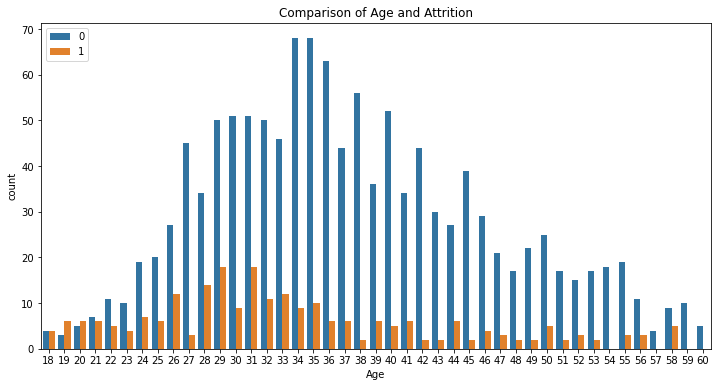
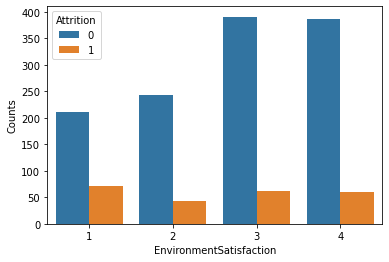
3 - Business Travel: We can confirm that the employee who are traveling Rarely are more likely to stay in a company, which is obvious as if one has to travel a lot for work, it will make any indivisual frustrated, unless traveling is not related work but for holiday.

4 - Daily Rates: A Daily rate is the billing cost for an individual's services for a single day. It is sometimes called a per diem. Daily rate is almost uniformaly distributed.understood as if there are more work load exceeding a limit then it could be possible that an indivisual try to quit.

# Bivariate Analysis





From the graphs above, we use different figures bar and data visualization.

Eg- Here we can see that the attrition is maximum between the age groups 26-33. At a very younger age, i.e. from 18-20, the chances of an employee leaving the organization is far more- since they are exploring at that point of time. It reaches a break even point at the age of 21.

Eg-From the above graph it can be seen that at the initial phase the employment there is more rate of Attrition,at 1 level the attrition rate is 25.35% highest from all other levels. A new employee takes time to keep up with the office Environment, From 2 to 3 we see that there is an increase at attrition rate, the reason could be if an employee is shifted to a new process or work place.

# OUTLIER DETECTION/REMOVAL:

# 

# 

We see that couple of high values in the monthly income column above the top whisker but rest of the columns don’t have outlier.

# CORRELATION MATRIX

# Lets check the correlation between the data and check with the data visualization.

# 

# 

I see that Joblevel and Monthly Income are very Highly correlated at 0.94, I will remove monthly income for my model building data set from input variable, as With Job level there is more accurate and negative linear relation with Arrition.

some of the independent features (basically which includes long duration(yearly)) are positively correlated to each other.

# Splitting into inputs and Target

# 

We saw in the above section while checking the proposition of target features, it wasn’t equal, it was data imbalance problem and we discussed that would see the solution so here is the solution:-

# 

# CROSS VALIDATION TEST

# 

# From the above Cross Validation test , we got a score of 85.72 with random forest classifier. I will be buliding my model using that only

# 

# here we have got an accuracy score of 87.55%, lets also use Gridsearch CV for parameter Tuning

# Logistic Regression

# 

# 

# AUC ROC curve

# 

# 

# 

**Using multiple algorithms:-**

I have used logistic regression, decision tree Random forest is giving better result as compare to other algorithms, so i have decided it to select as best and process further operation like hyper parameter tuning and cross validation

I have used random forest as first algorithm, it is givive 100 accuarcy on training test and approx 91 test set.i have used for loop to find the best random state. We can see that random\_state 275 is giving best accuracy with good recall and precision score out of other random state.

The 100 accuracy on training set will definitely raise question in the mind. To answer this question, i have checked the test score as well as roc\_auc\_score and it has been found that roc\_auc\_score is 92.02 which means that model is able to identify 92.02 data correctly.

# 

I have used Gridsearch CV to find best parameter of random forest.

Now, we are going to perform cross validation, cross validation train model on all the data points so that model learn every pattern and provide average score of all the models. We get score by using mean function and we check the standard deviation as well of the score. Which ensure us that given score doesn’t have risk factor.

# 

**Conclusion:-**

Dataset was quite clear, there was no missing value. It was mix of categorical and numerical features. We have performed multiple analyses to check that which factor plays important role in attrition. We checked outlier and found that few columns have some extreme value but it is very close to upper whisker and we didn’t try treating them because the ensemble methods will deal with them. I have checked correlated of each features and found that couple of features were correlated so have deleted them.  
As we saw at the initial phase of analysis that data was imbalance, we have corrected that by applying oversampling technique and then Model was trained. Random forest has given best F1 score and has taken it for final model.