**HR Analytics Project**

**Introduction:**

Attrition is a problem 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. As such, there is great business interest in understanding the drivers of, and minimizing staff attrition.

Classification models to predict if an employee is likely to quit could greatly 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 remedy controllable factors that can prevent attrition of the employee.

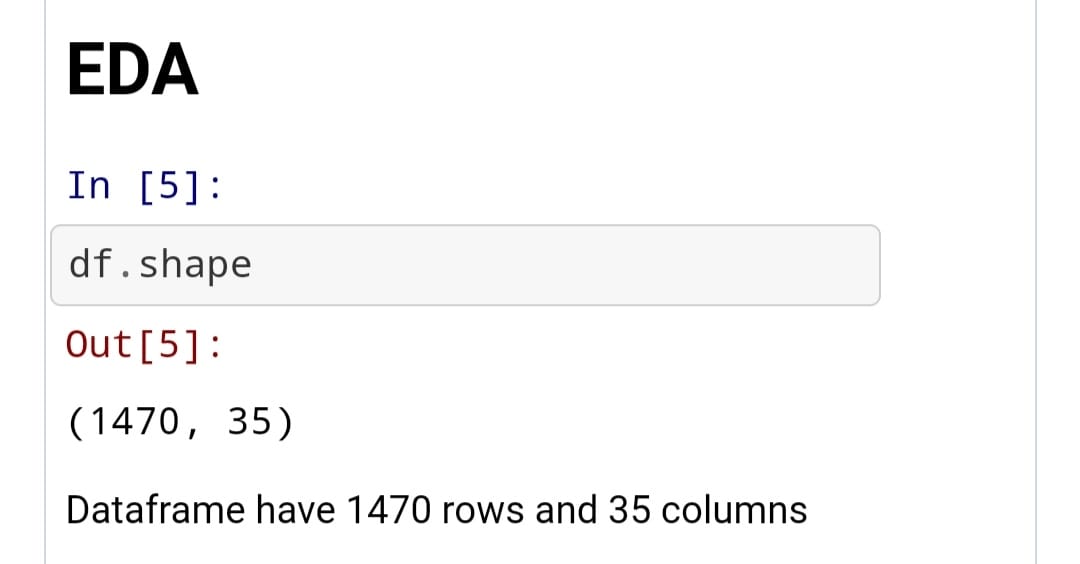
Attrition, from an overall perspective, is a progressive decrease or diminishing of a thing. From a business perspective, there are two different ways to characterize Attrition. Attrition for a business can be portrayed as either employee attrition or client attrition – both vital to comprehend as a business owner. Employee attrition is utilized to portray the decrease of employees. Employee attrition can occur for a large number of reasons. The reasons may include employees resigning, securing other position openings, or leaving because of misery.

Nonetheless, it is vital to note that for it to be characterized as employee attrition and not simply a piece of employee turnover, business owners or administrators should choose not to top off the particular position that is presently unfilled. Inside employee attrition, there is either voluntary or involuntary employee attrition. The distinctive factor for the two sorts of Attrition versus employee turnover is that with Attrition, the positions are not immediately topped off or topped off by any means. Voluntary employee attrition models incorporate employees leaving to seek after other open positions or resigning. Then again, involuntary employee attrition incorporates work position disposal because of business cutting back.

In this paper, we will make a model based on the information provided to predict the Attrition of an employee. Also, we will try to answer How does Attrition affects companies? Moreover, how does HR Analytics help in analyzing Attrition? We will discuss the first question, and for the second question, we will write the code and try to understand the process step by step.

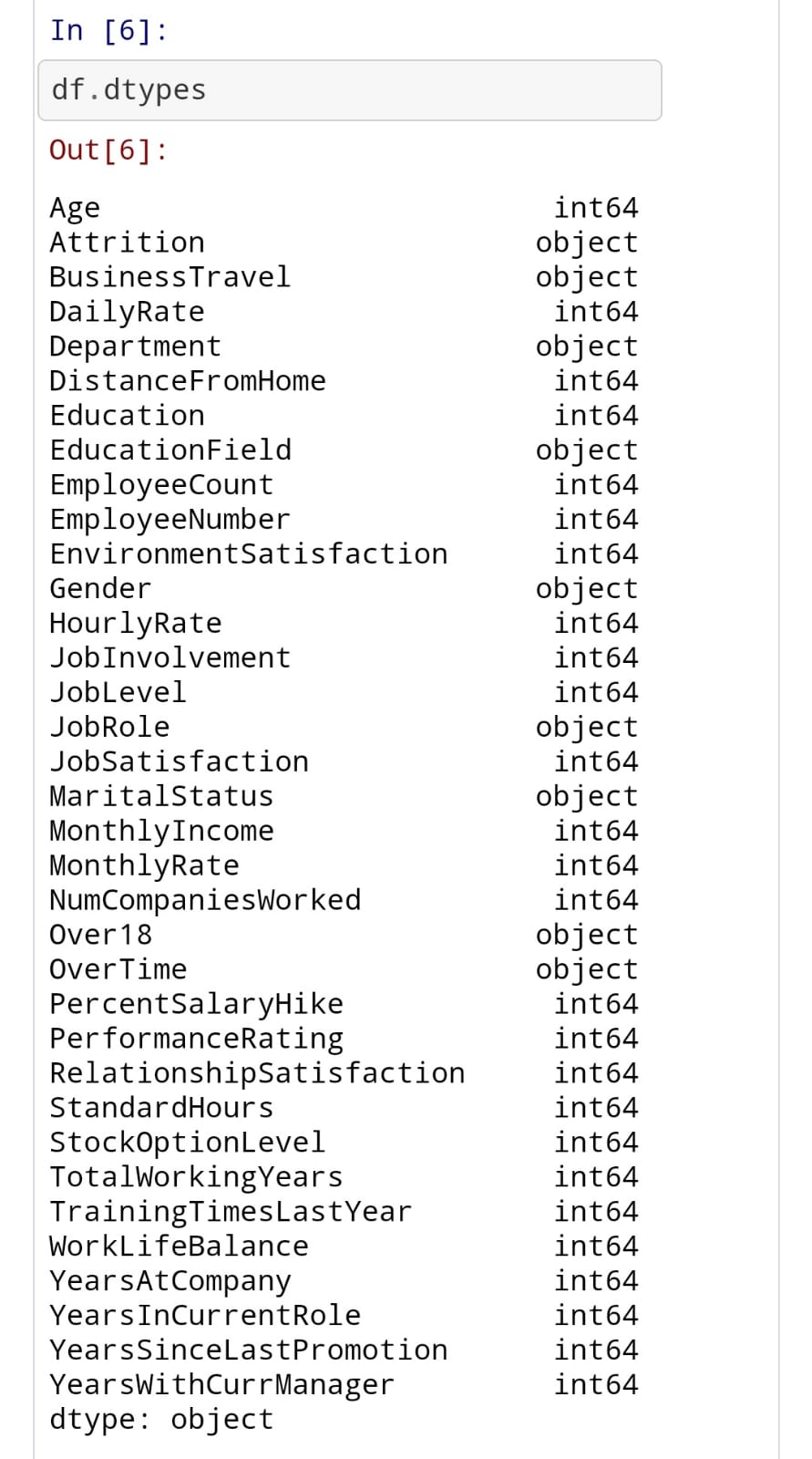
**Dataset:**

The dataset used in this project has been provided by Data trained for the assessment process. The dataset contains 1470 rows and 35 columns. In this dataset, our dependent, or we can say target variable is the Attrition column.

****

**EDA:**

In this attrition dataset, will perform the EDA to find out the most important metrics which are playing major role in employee attrition. It is idle to explore a data set with the various exploratory Method, especially when they can be done together for comparison. Every data scientist should compile a techniques in exploratory data analysis. Once we fully understand our data set, we may need to revisit one or more data munging tasks to refine or transform the data. The purpose of EDA is to obtain confidence in our data to a point where we will be ready to engage a machine learning algorithm. The very first thing to check in the data given is the type of the given variable.



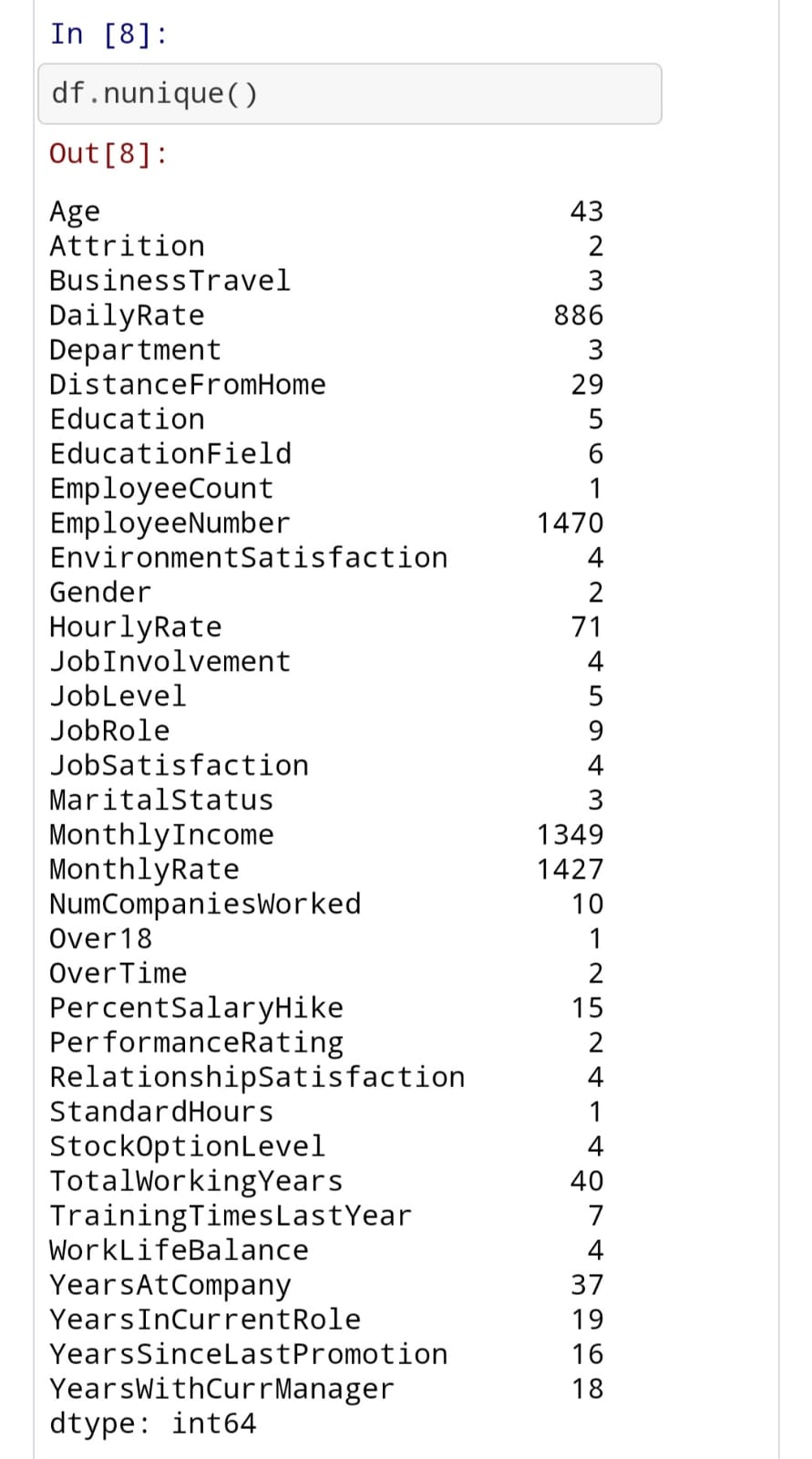
As shown in the image shown above, there are nine object type variables and 26 integer type variables.

**NULL VALUE:**



In the Null dataset, it is clear there is no null value in the dataset. As the total number of rows, this dataset contains 1470, and every variable has a 1470 non-null count.

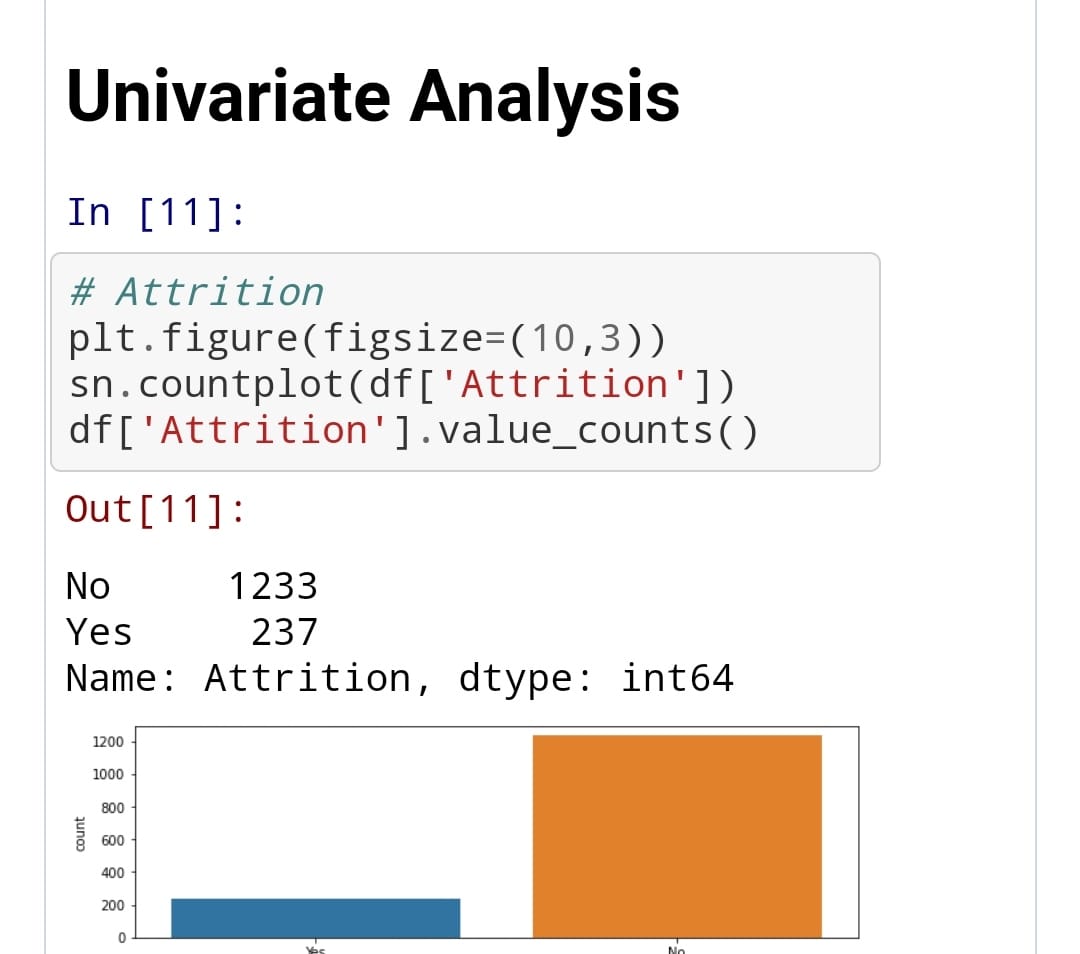
Further, we will try to examine the unique value of each variable. There are three columns with only one unique values.



**Univariate Analysis:**

Univariate Analysis is only one dependent variable. The objective of univariate analysis is to derive the data , define and summarize it ,and analyze the pattern present in it. In a dataset, it explores each variable separately.

* **Attrition**:

****

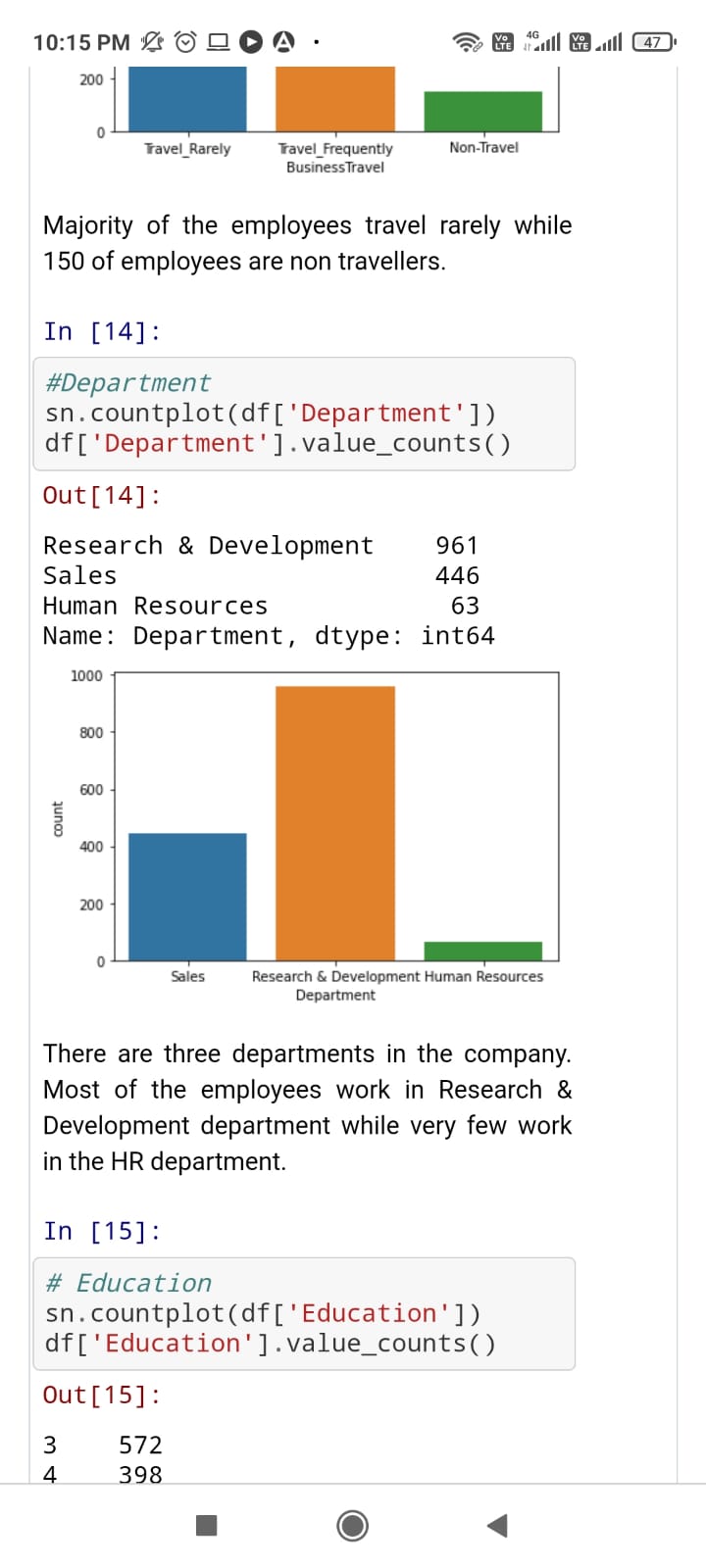
Dataset is highly imbalance as the label class has only 237 of yes and 1233 of no category. Attrition is our dependent variable. Also, from the image, we can see that the data is highly imbalanced. We need to do data balancing; otherwise, we will not make an effective model.

* **Business Travel**:

****

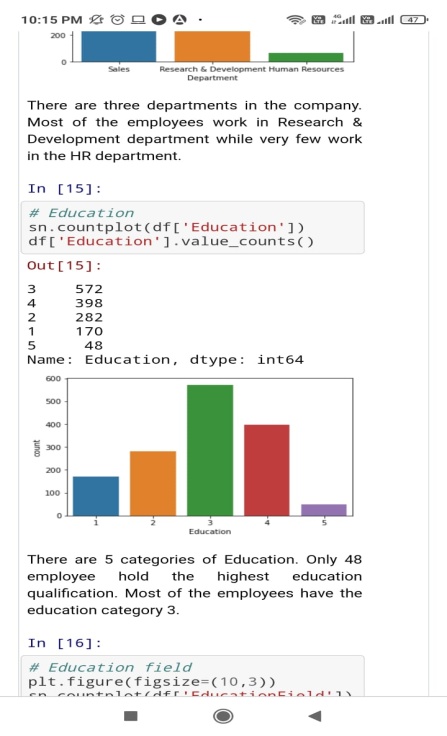
Majority of the employees travel rarely while 150 employees are non travellers.

* **Department:**



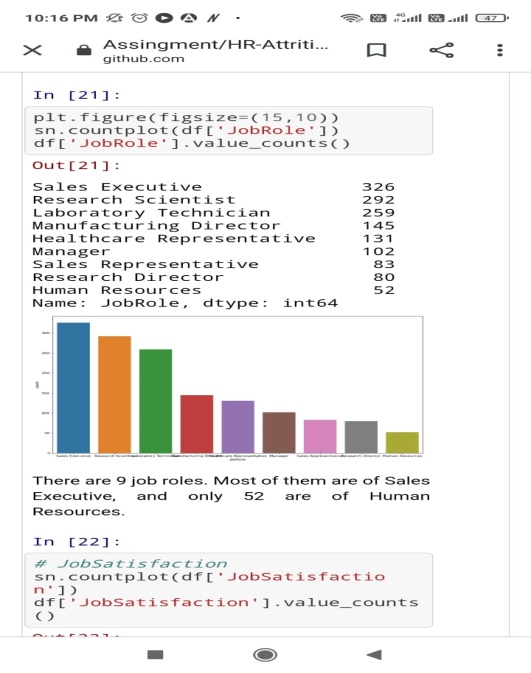
There are three departments in the company. Most of the employees work in Research& Development department while very few work in HR department. Majority of the employee works in Research & Development Department.

* **Education:**



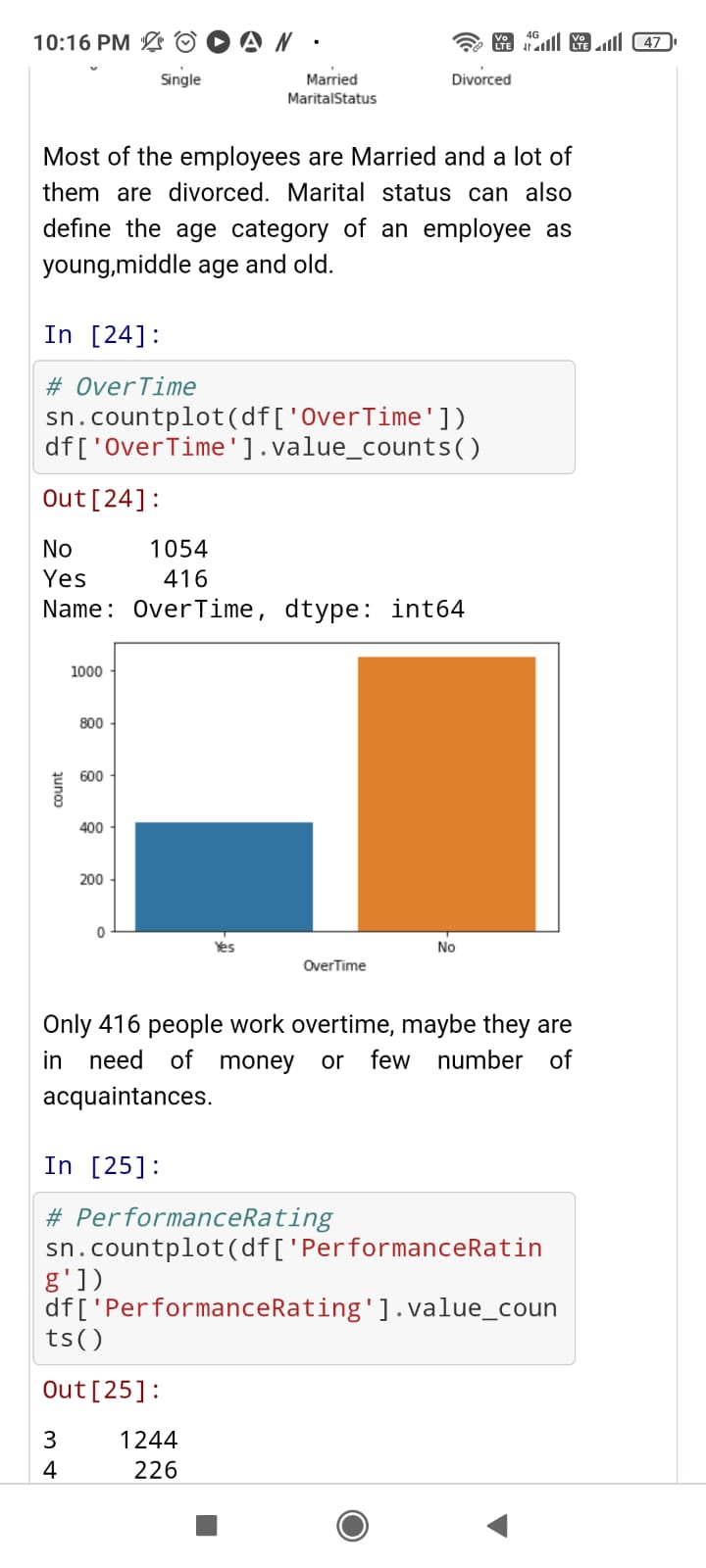
There are 5 categories of Education. Only 48 employee hold the highest education qualification. Most of the employees have the education category 3.

* **Job role:**

****

There are 9 job roles. Most of them are of Sales Executive, and only 52 are of Human Resources. Though Majority of the employee works in Research & Development department, the Majority of employees have job role of Sales Executive.

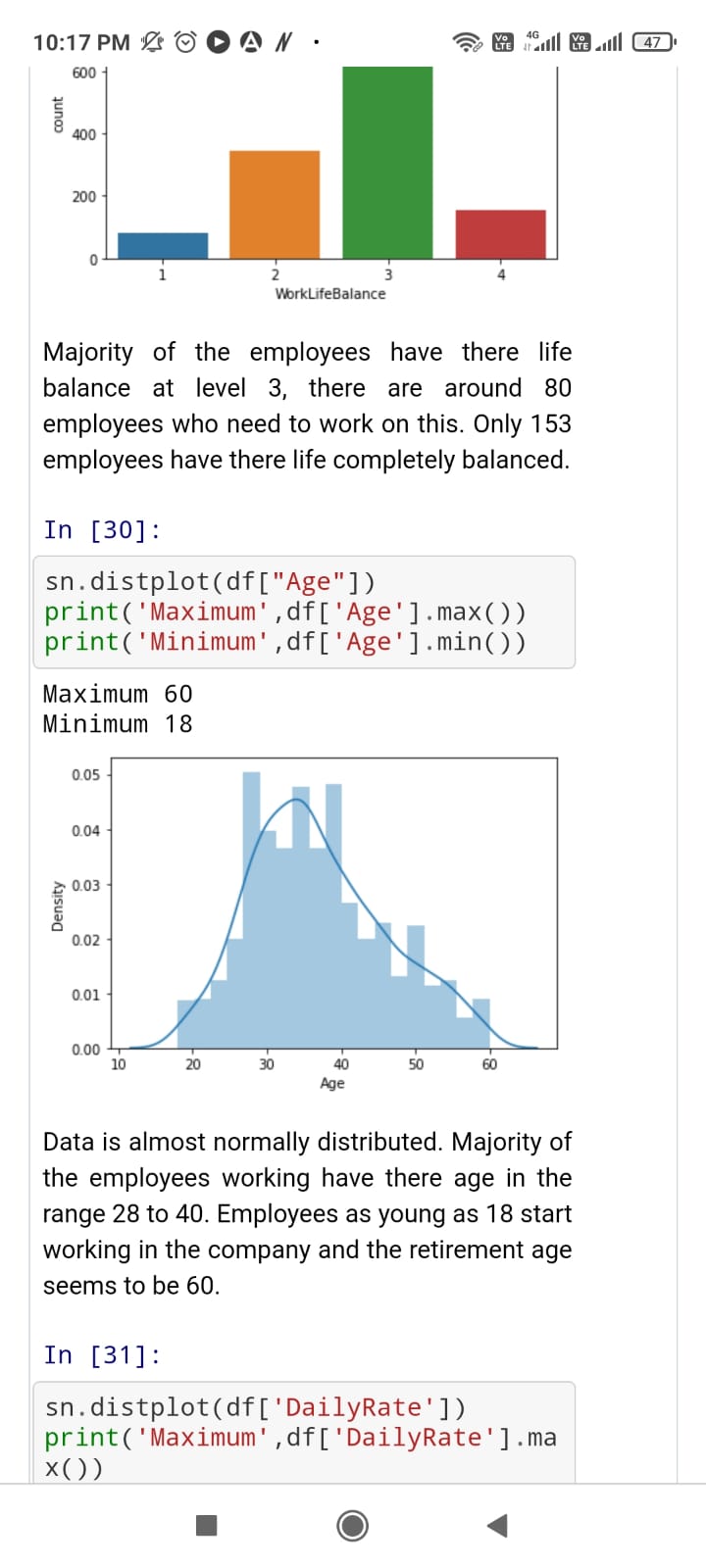
* **Over-Time:**



Only 416 people work overtime, maybe there are in need of money or few number of acquaintances. Though most employees are not doing overtime, however, there is a reasonable number of overtime employees.

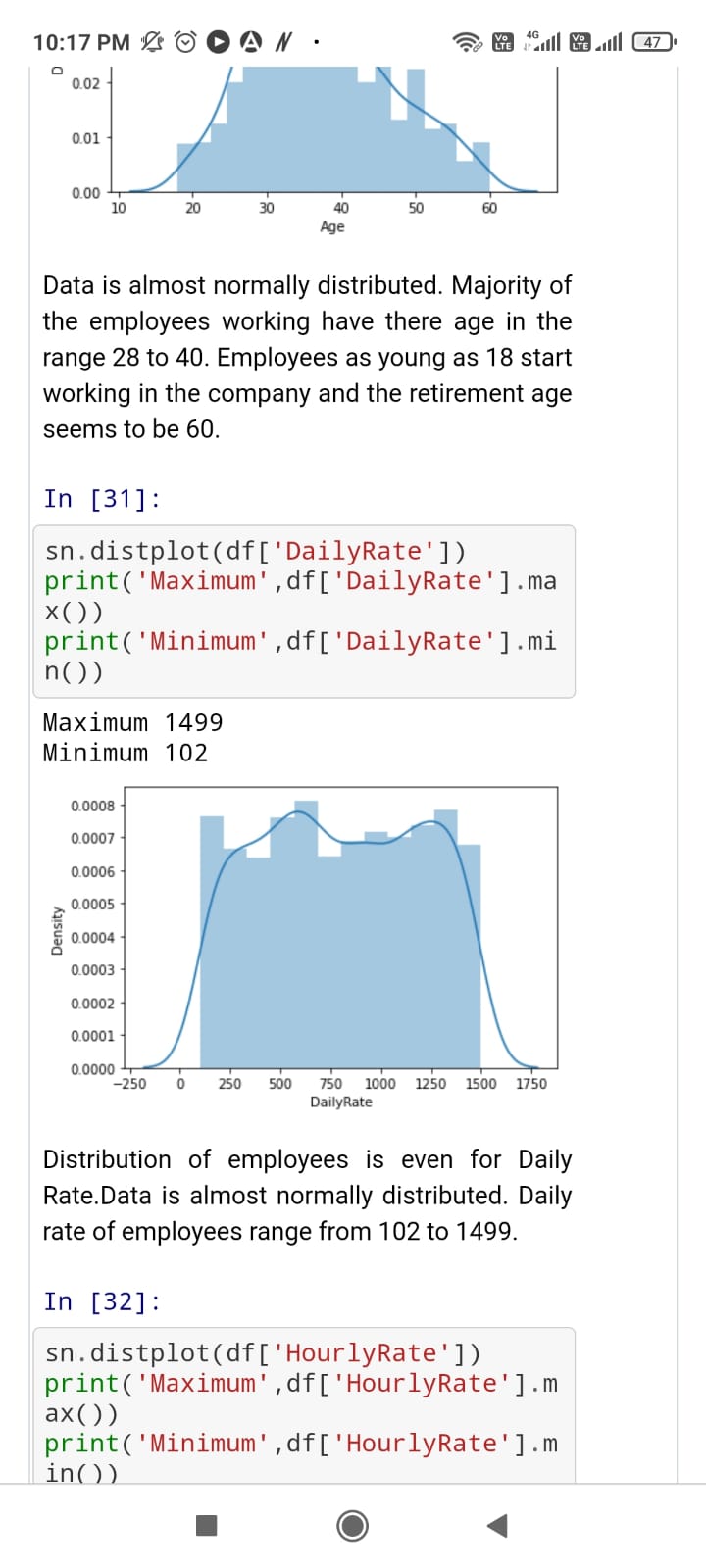
To analyze the integer variable, we are making a histogram function with the help of a distplot.

* Age:



Data is almost normally distributed. Majority of the employees working have there age in the range 28 to 40. Employees as young as 18 start working in the company and the retirement age seems to be 60.

* **Daily Rate:**

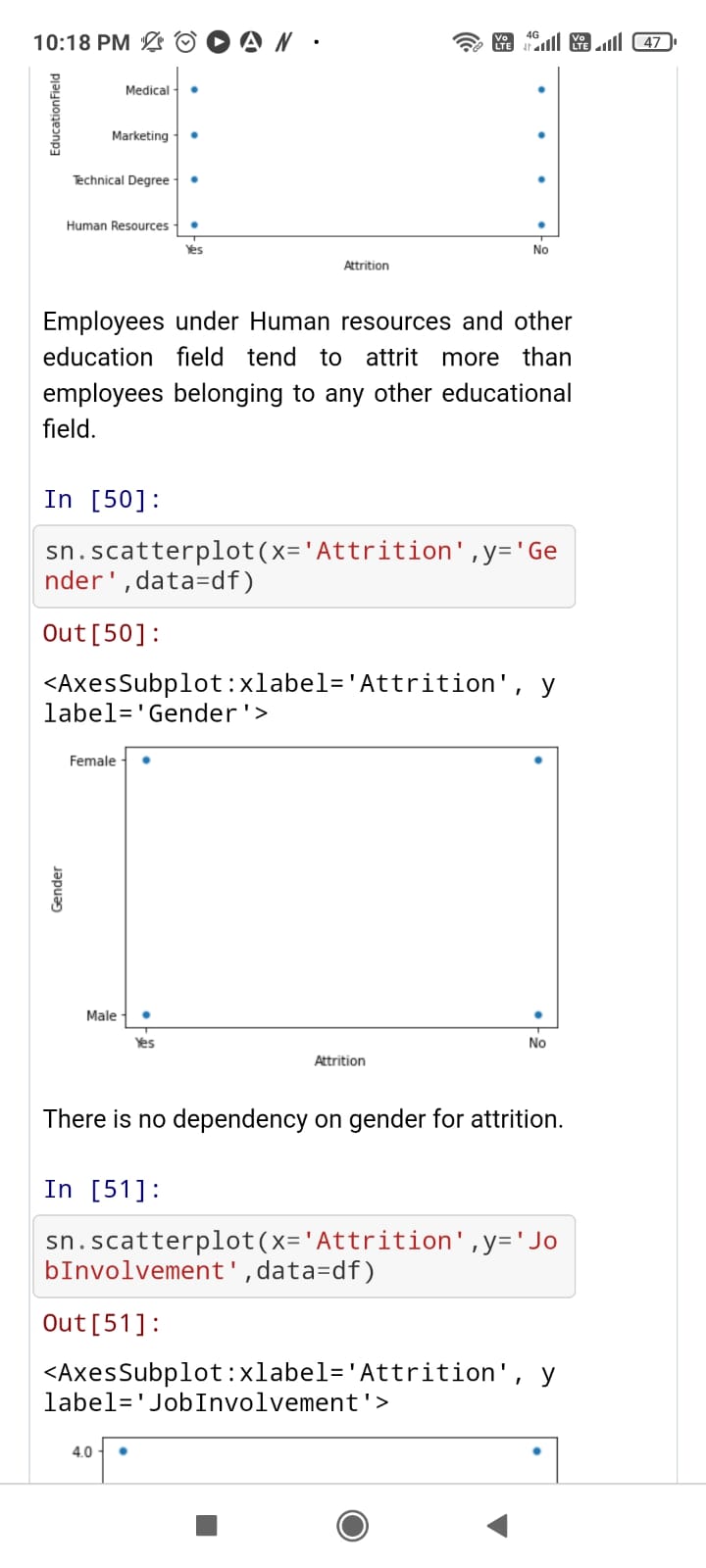
****

Distribution of employees is even for Daily Rate. Data is normally distributed. Daily rate of employees range from 102 to 1499.

**Bivariate Analysis**:

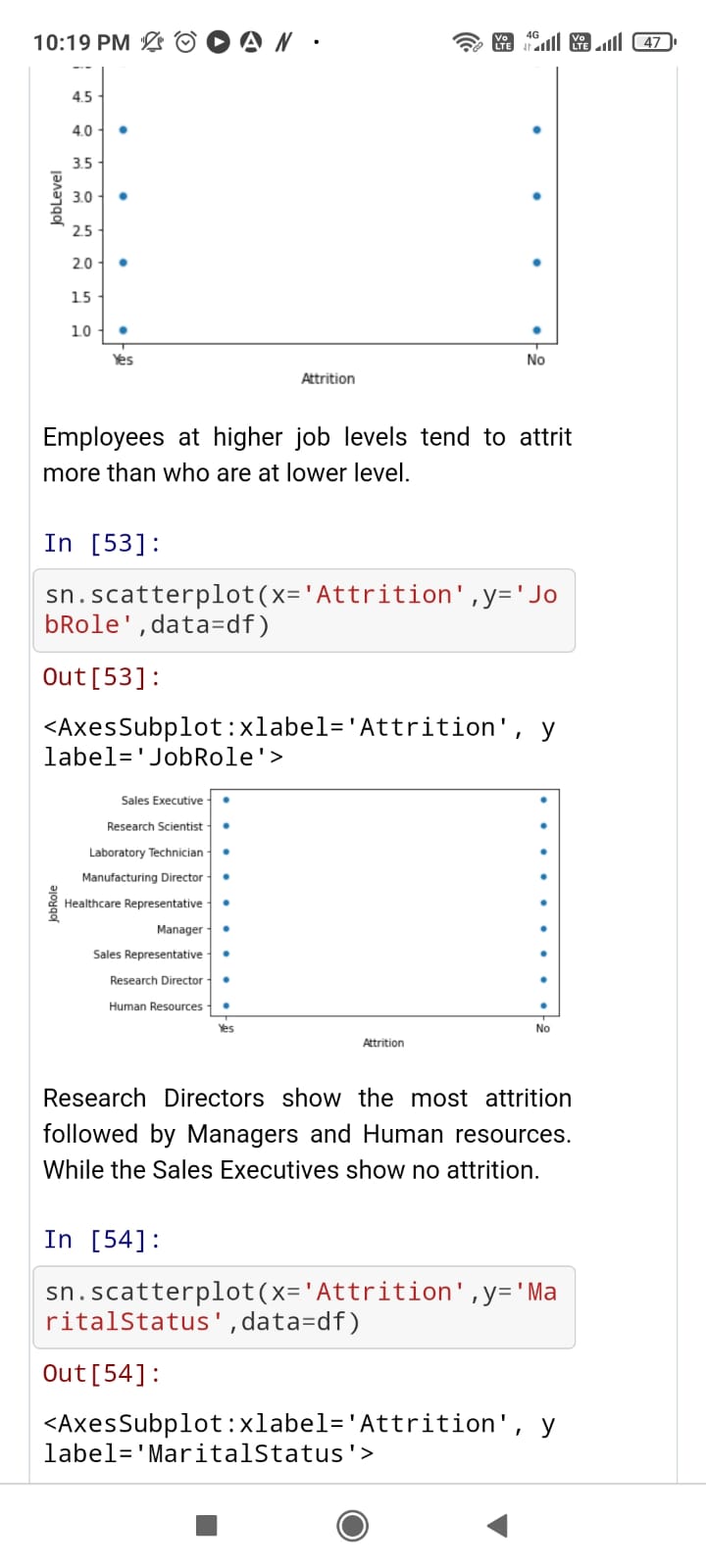
Bi means two and variate means variable, so here there are two variables. The analysis is related to cause and the relationship between the two variables. As the Attrition columns contain only two unique values, we already know it is a binary classification problem. However, some columns only contain one unique value; these columns are Standard Hours, Over18, and employee count. These columns will add no insight; it will also not need the model building; we will drop these values.

* Gender/ Attrition:



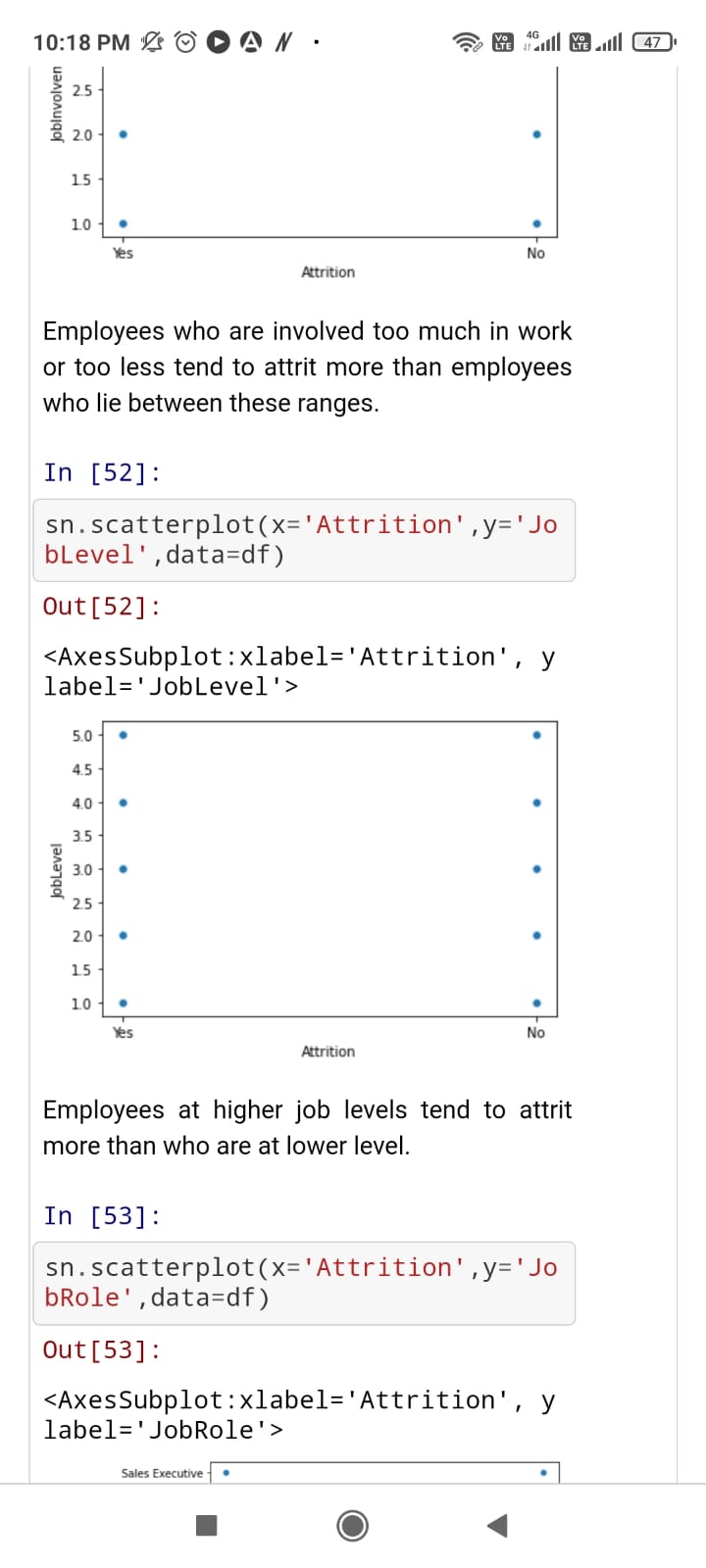
There is no dependency on gender for attrition. As indicated in the image, Male employees are more inclined towards Attrition than female.

* Job role/Attrition:



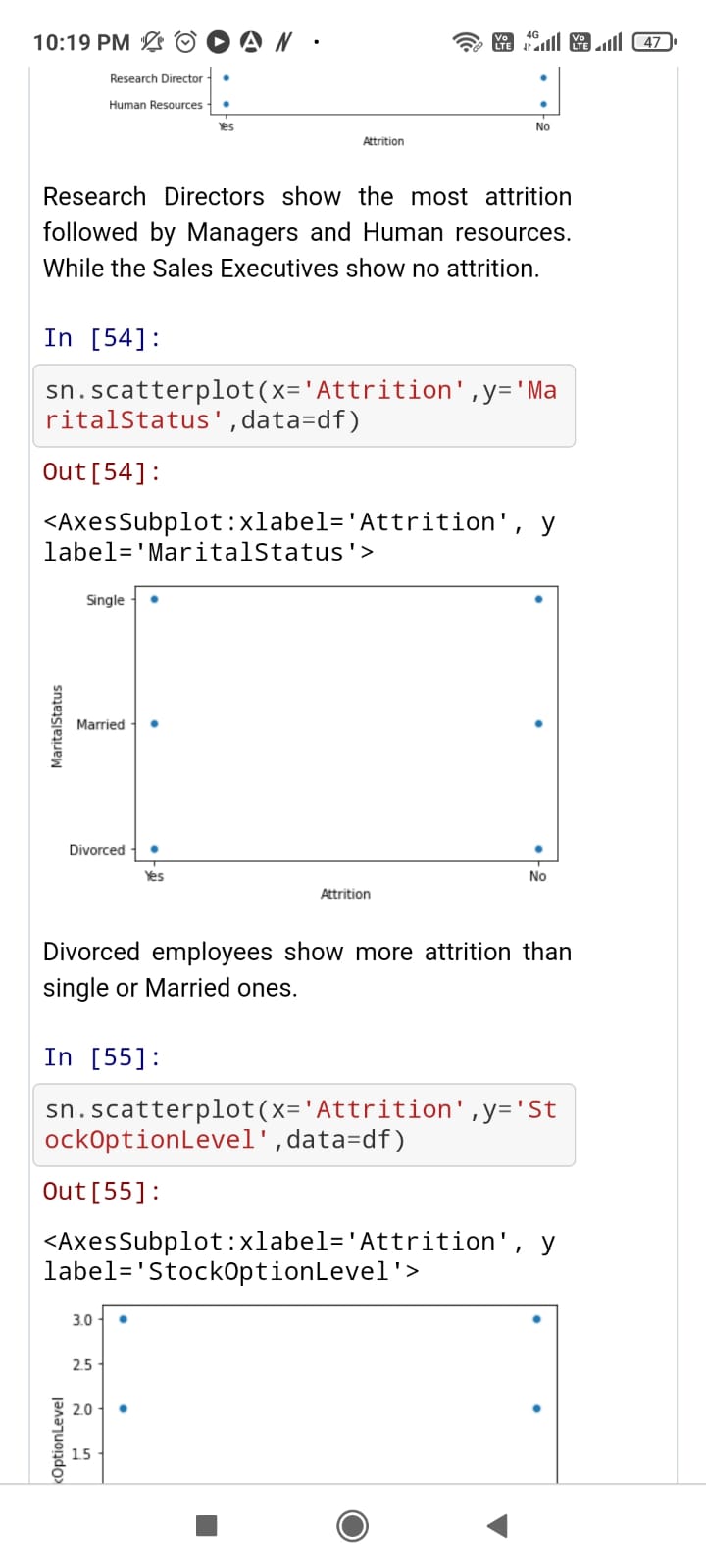
Research Directors show the most attrition followed by Managers and Human Resouces. While the Sales Executive show no attrition. Employees who are Sales Executive, Research Scientist, and Lab rotary Technician are more prone towards Attrition than other job role employs.

* Job level/Attrition:



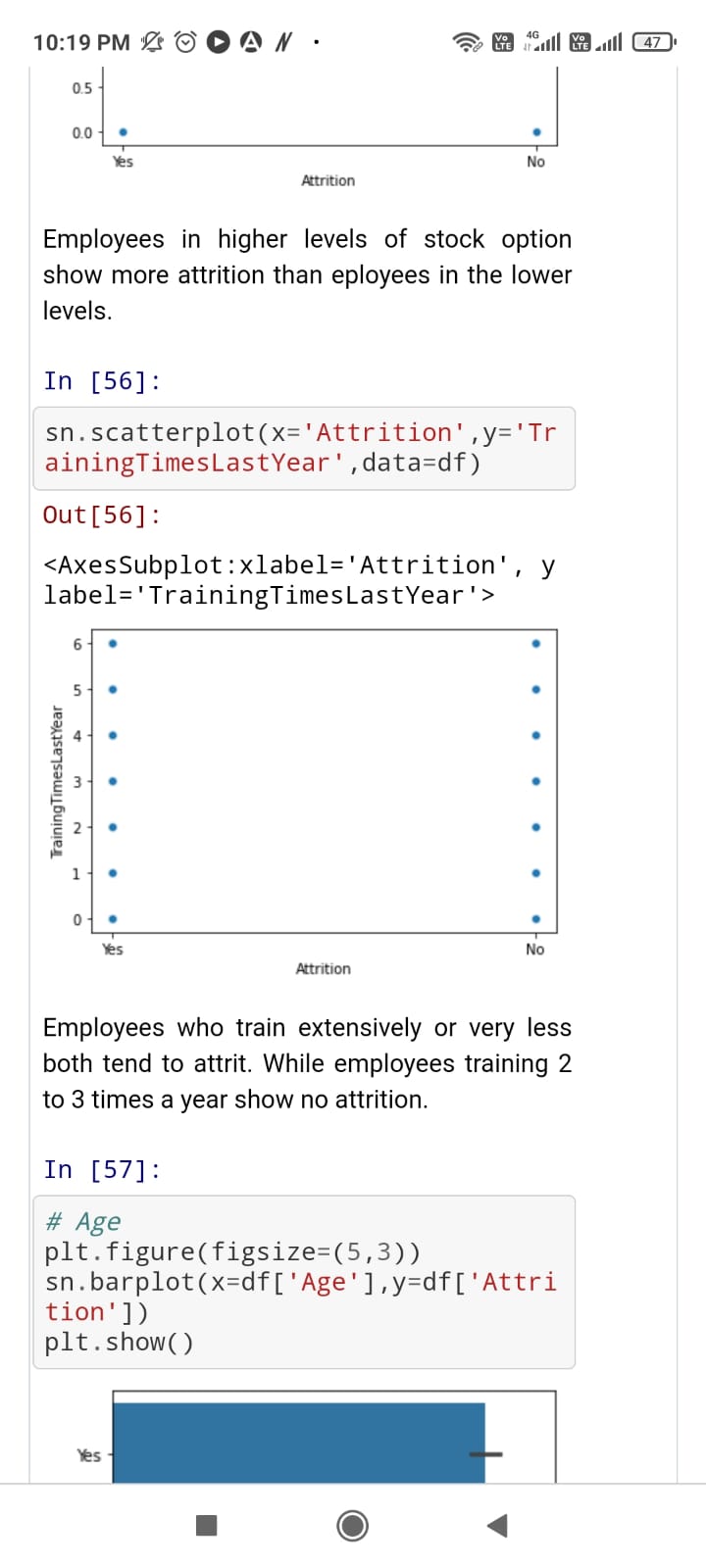
Employees at higher job levels tend to attrit more than who are at lower level. As the image indicates, employees on a lower job level are more prone to Attrition, while employees on the higher post are significantly less inclined towards Attrition.

* Marital Status/Attrition:



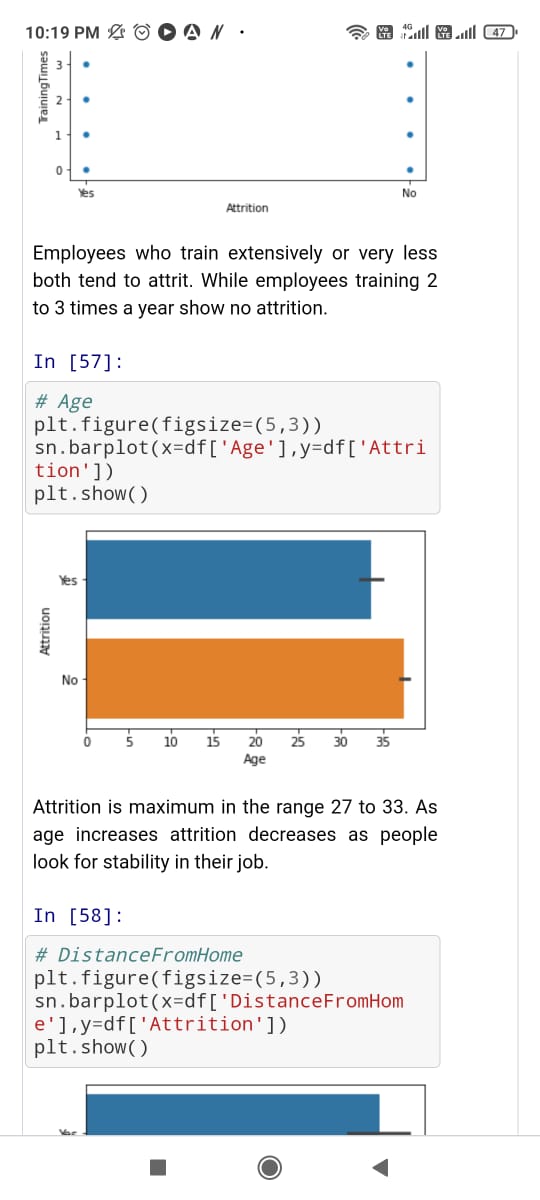
Divorced employees show more attrition than single or married ones.

* TrainingTimeLastYear/Attrition:



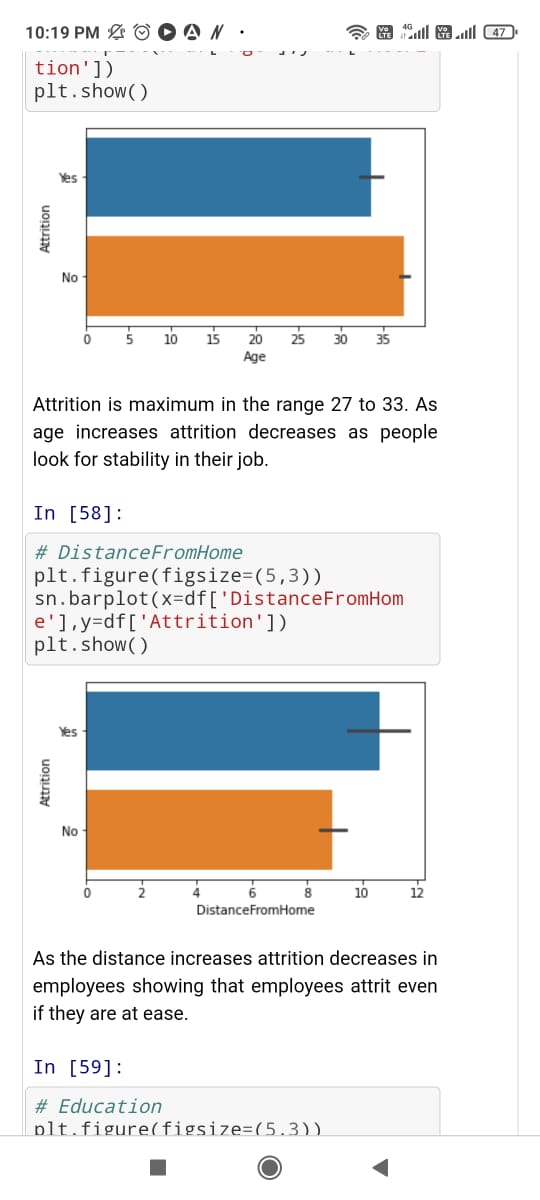
Employees who train extensively or very less both tend to attrit. While employees training 2 to 3 times a year show no attrition.

* Age/Attrition:



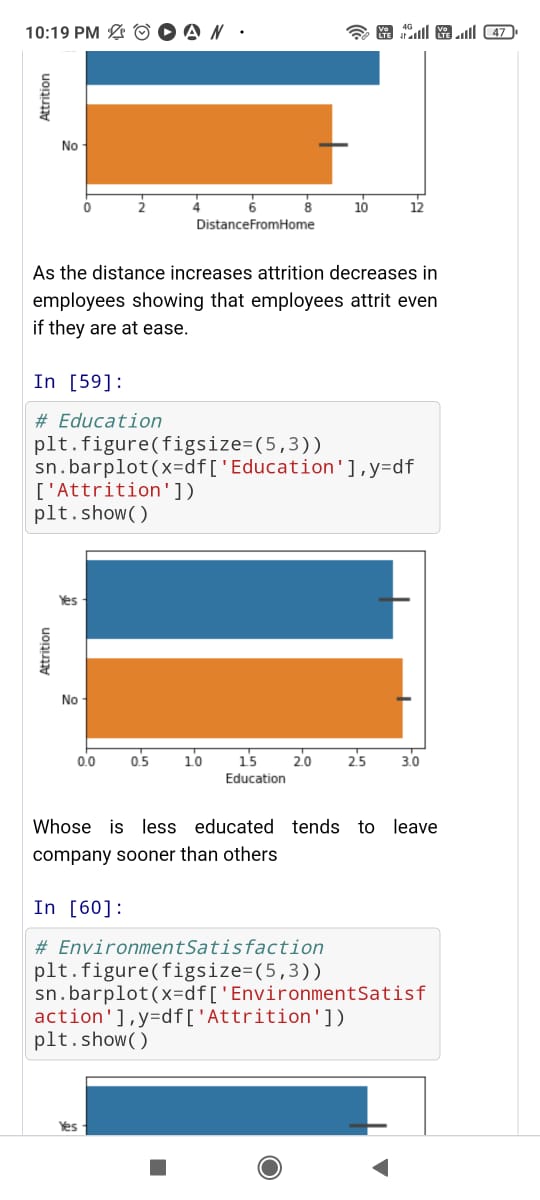
Attrition is maximum in the range27 to 33. As age increases attrition decreases as people look for stability in their job.

* DistanceFromHome/Attrition:



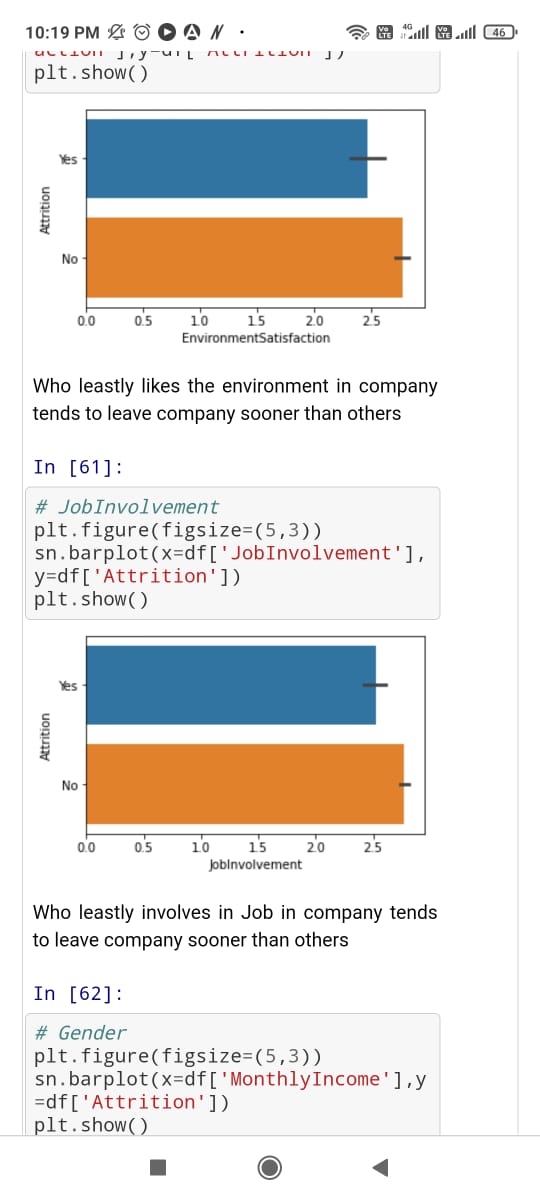
As the distance increases attrition decreases in employees showing that employees attrit even if they are at ease.

* Education/Attrition:



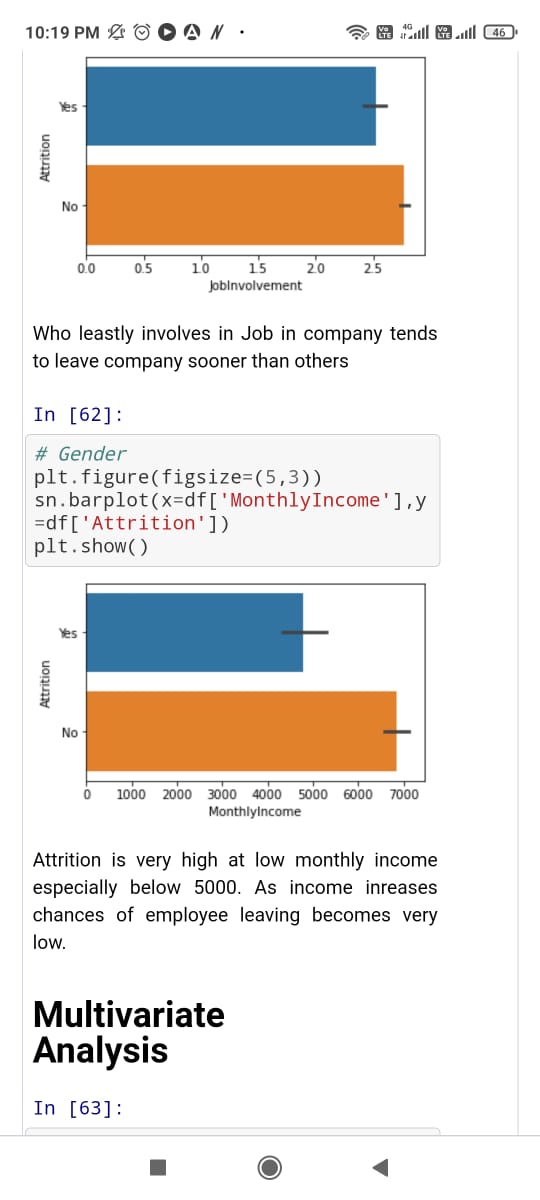
Less educated tend to leave company sooner than others.

* JobInvolvement/Attrition:



Least involves in job in company tends to leave company sooner than others.

* MonthlyIncome/Attrition:



Attrition is very high at low monthly income especially below 5000. As income increases chances of employee leaving becomes very low.

**Multivariate Analysis:**

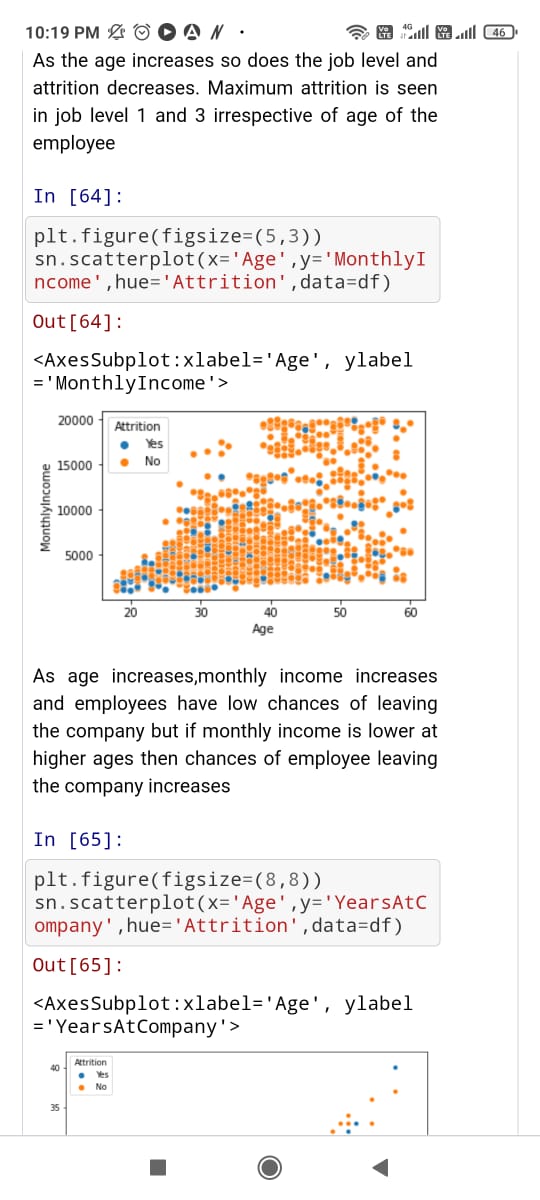
Multi variate analysis is required when more than two variables have to be analyzed simultaneously.

* Joblevel/Age/Attrition:



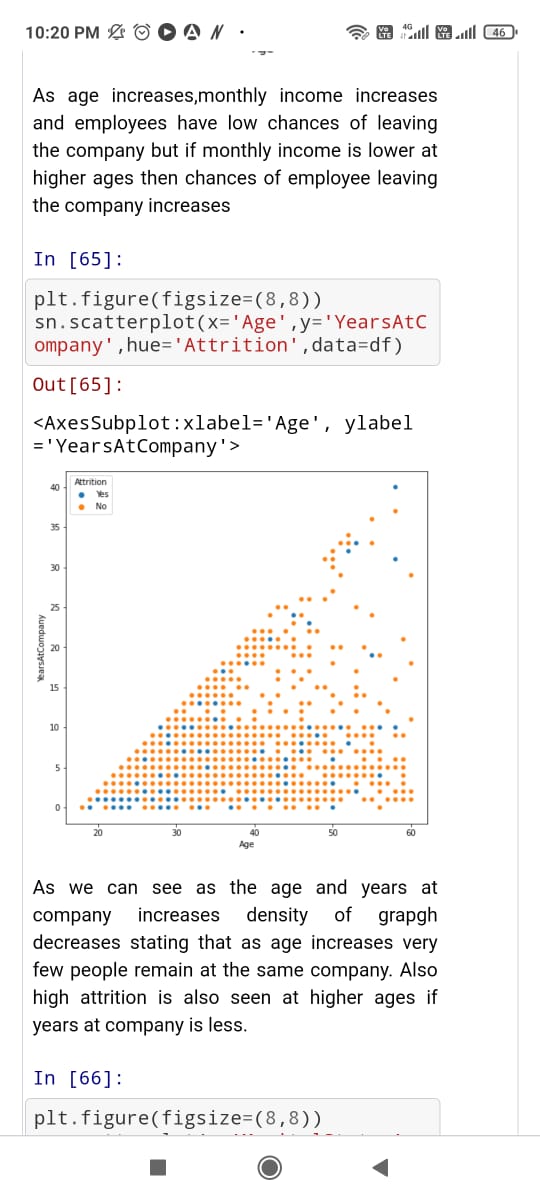
As the age increases so does the job level and attrition decreases. Maximum attrition is seen in job level 1 and 3 irrespective of age of the employee.

* MonthlyIncome/Age/Attrition:



As age increases , monthly income e increases and employees have low chances of leaving the company but if monthly income is lower at higher ages then chances of employee leaving the company increases.

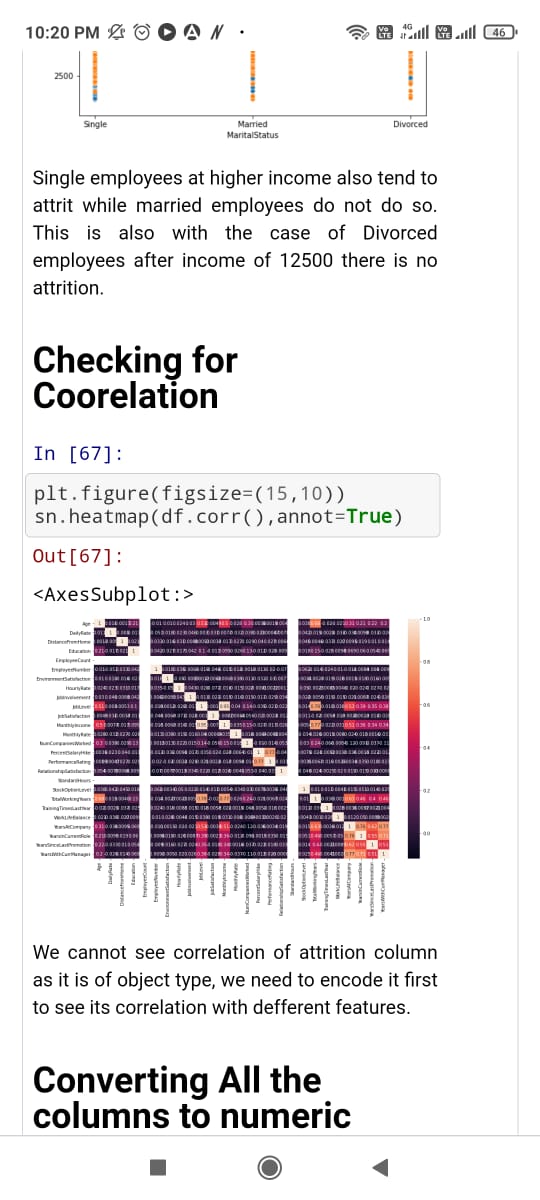
* YearsAtCompany/Age/Attrition:



As we can see as the age and years at company increases density of graph decreases. Starting that as the age increases very few people remain at the same company. Also high attrition is also seen at higher ages if years at company is less. Employees working with the company for a very long time are significantly less inclined towards Attrition. In contrast, employees who have been working for the last one or two are more inclined towards Attrition.

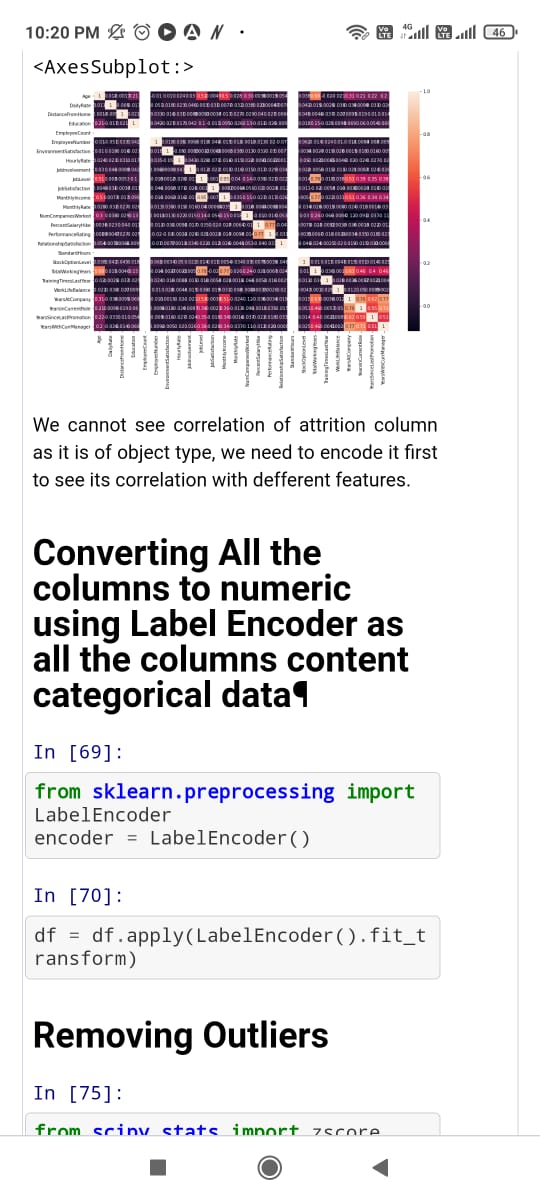
As we completed the analysis part, it is time to go for feature selection. We will use the correlation method to get the importance of the feature concerning the dependent variable. So, before proceeding, we need to encode all of the categorical data.

**Checking for Coorelation:**

****

As we can see in the image, over time and marital status are the most positively correlated variable with Attrition, while entire working years and job level are the most negatively correlated variables. Also, we can see in Performance rating, Business travel and Hourly rate are close to zero and significantly less correlated to Attrition we will drop these columns. Further, we can see that Over 18, Employee count, and Standards hours show nan because this variable contained only a single.

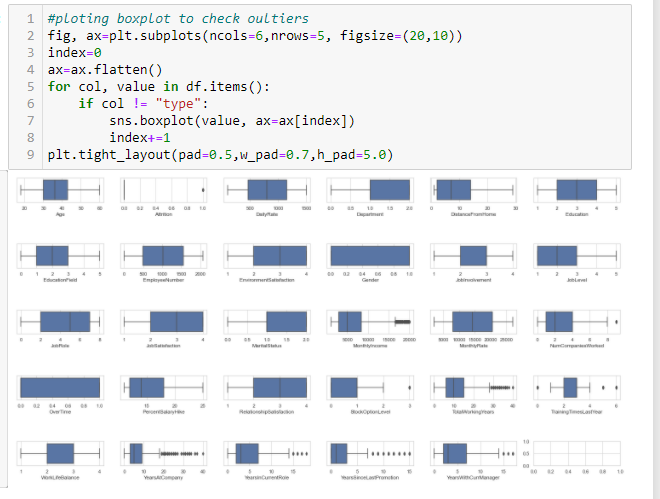
**Label Encoder:**

****

Here. we have used Label Encoder to transform the value of the given variables.

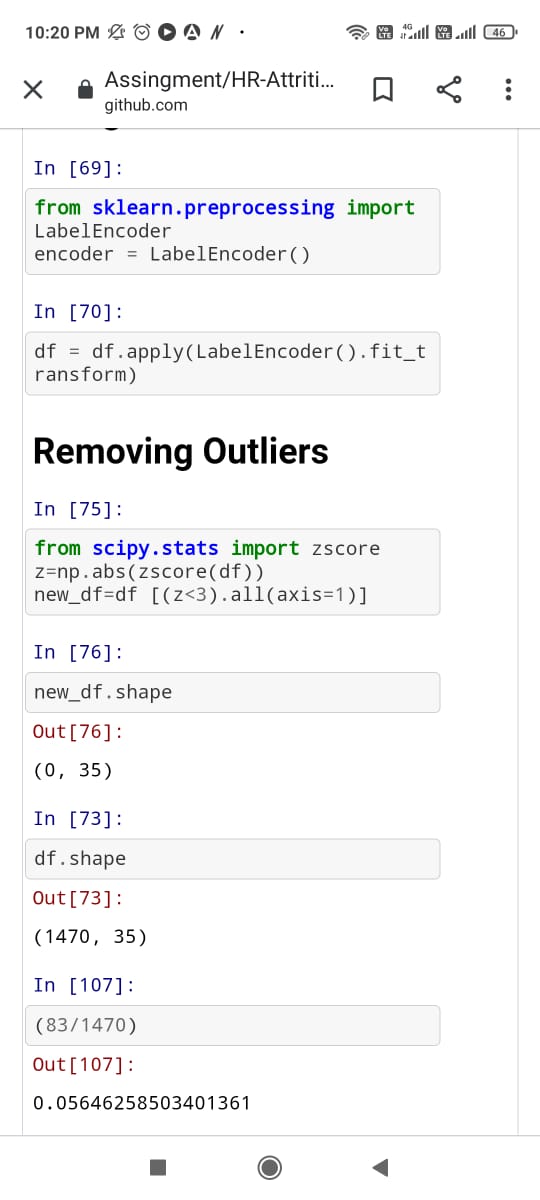
**Checking outliers:**

The next step is to see whether our data contains outliers or not.



There are outliers in some of the columns, so we will use the Zscore method to replace the outliers.

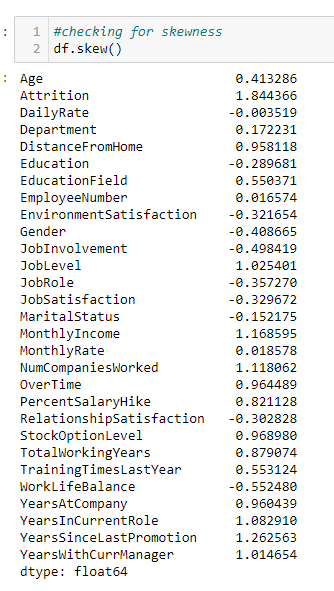
**Removing Outliers:**

****

The value of z-score tells how many standard deviations are away from the mean. If a z-score is equal to 0, it is on the mean. A positive z-score indicates the raw score is higher than the mean average. A negative z-score reveals the raw score is below the mean average. To remove the outliers in this dataset using z-score.

**Checking for Skewness**:

Further, before feeding the data for the prediction, we need to check the skewness of the data and treat it.

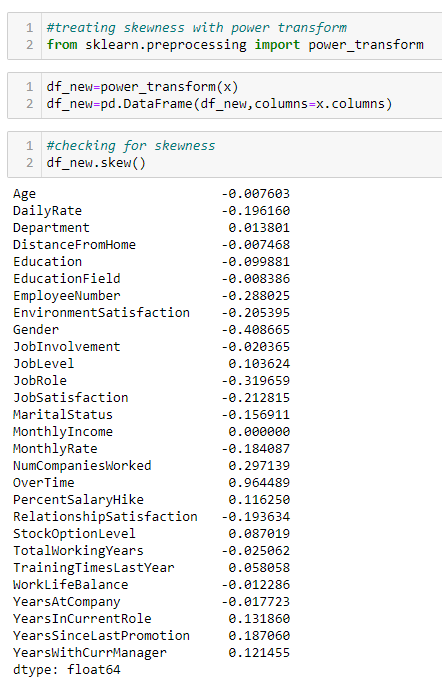


We will consider any value as skewness that is more than 0.5. As we can see, there is skewness in some columns. Price is our dependent variable, so we will not be changing anything in that column.

**Find the Best Model:**

So before proceeding further, we need to data in independent(x) and dependent(y) datasets. Except for Attrition, we have saved all the data in x. We have saved Attrition in y



Now we will treat skewness with the help of the Power Transform method.

Except for overtime, all the variable's skewness less than 0.5, implying the skewness problem has been treated. Overtime skewness will not be affecting our model as it is a categorical variable.

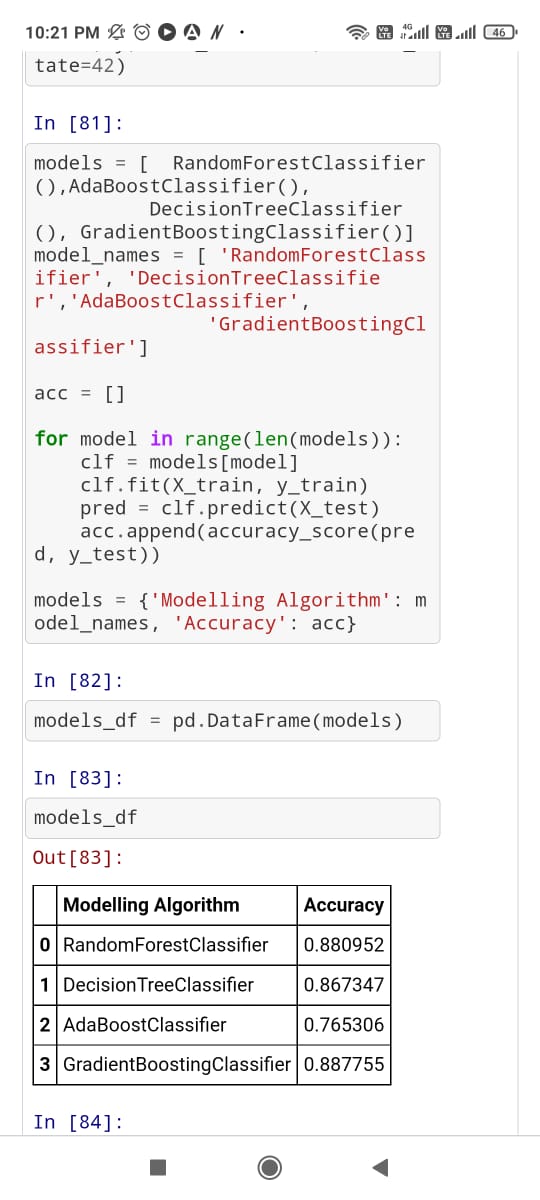
**Importing Libraries**:

Further, we will import all the necessary libraries for the model building process.

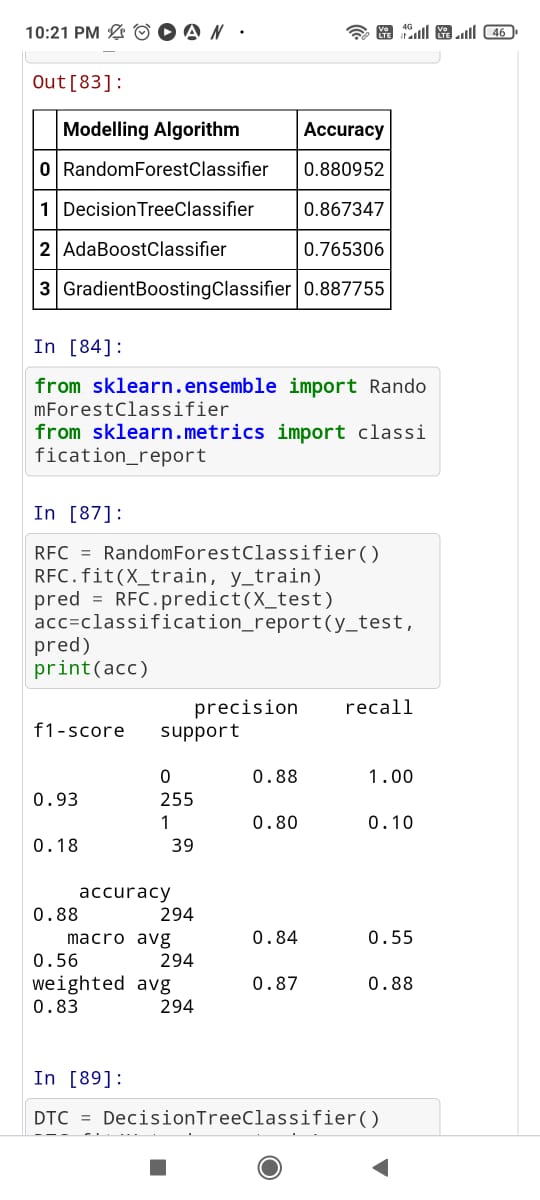


We have used an accuracy score, classification report and cross validation score to evaluate the model. Now we will split the data into train and test data to train the model and determine its performance.

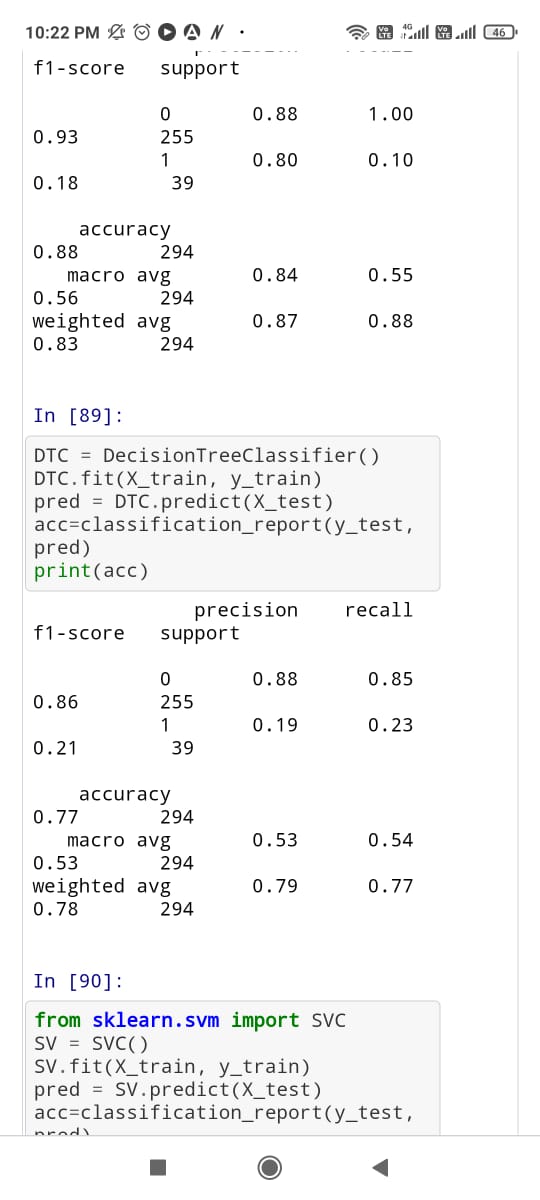




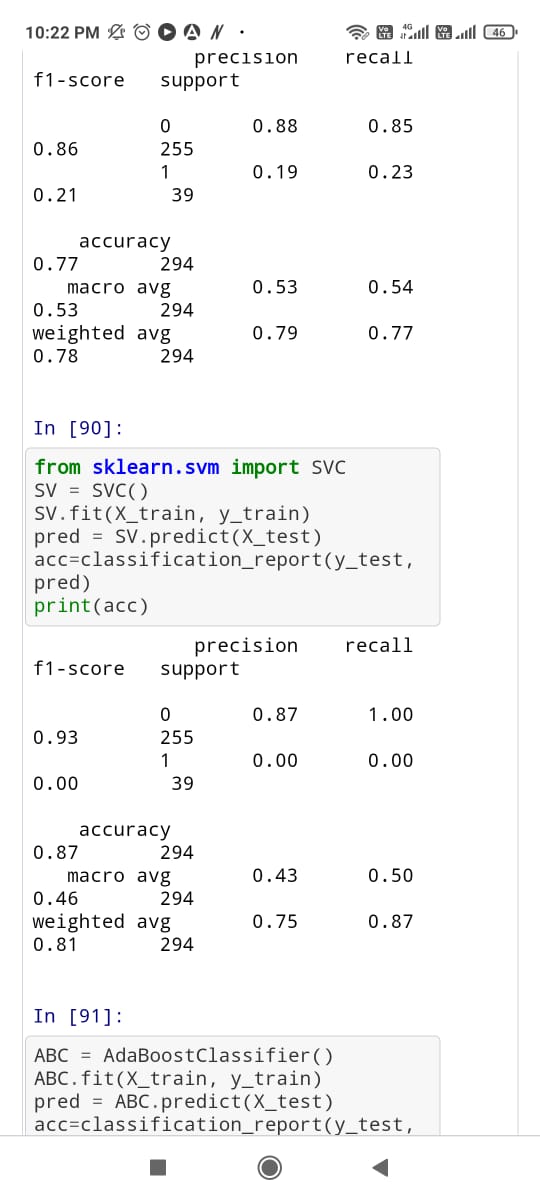
Here we have created a function model that will take the classification model's name, apply all the processes, and give information about the model's efficiency.



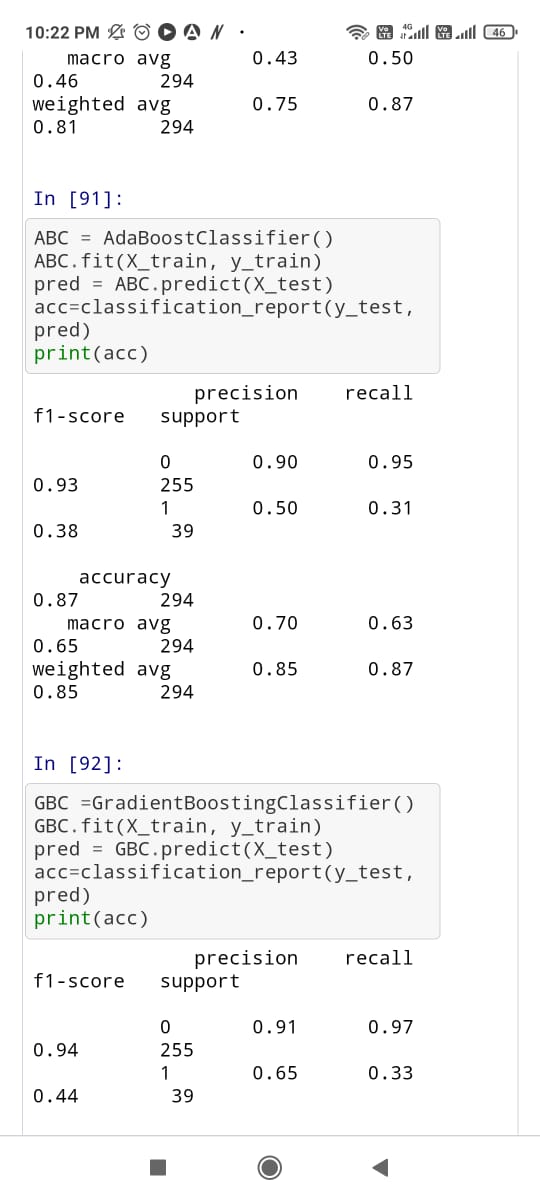
RandomForestClassifier accuracy score is 88 percent.



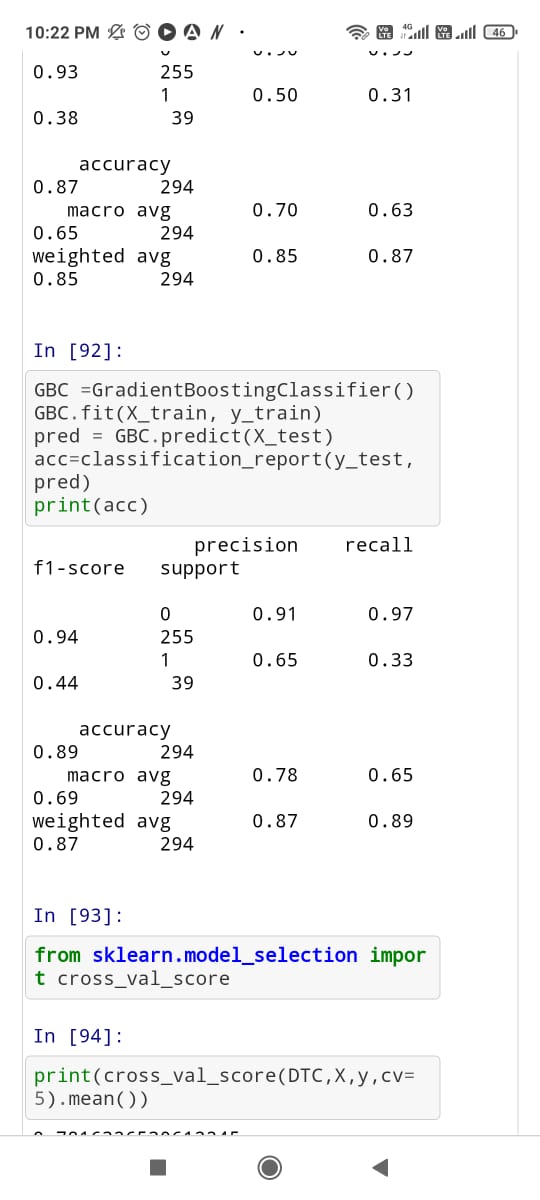
DecisionTreeClassifier accuracy score is 86 percent.



SVC accuracy score is 87percent.



AdaBoostClassifier accuracy score is 87 percent.



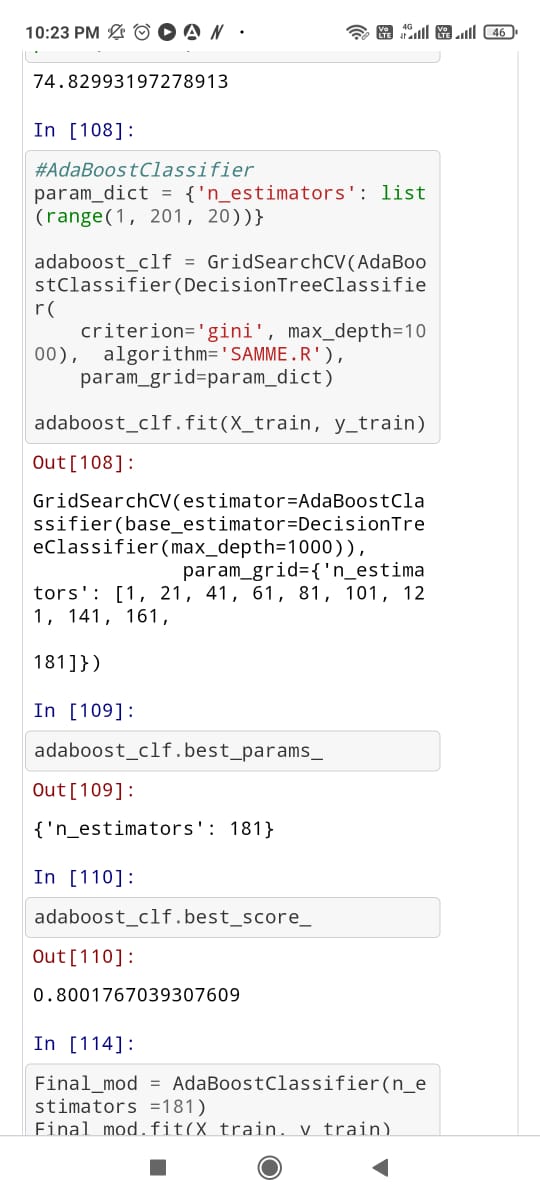
GradientBoostingClassifier accuracy score is 87 percent.

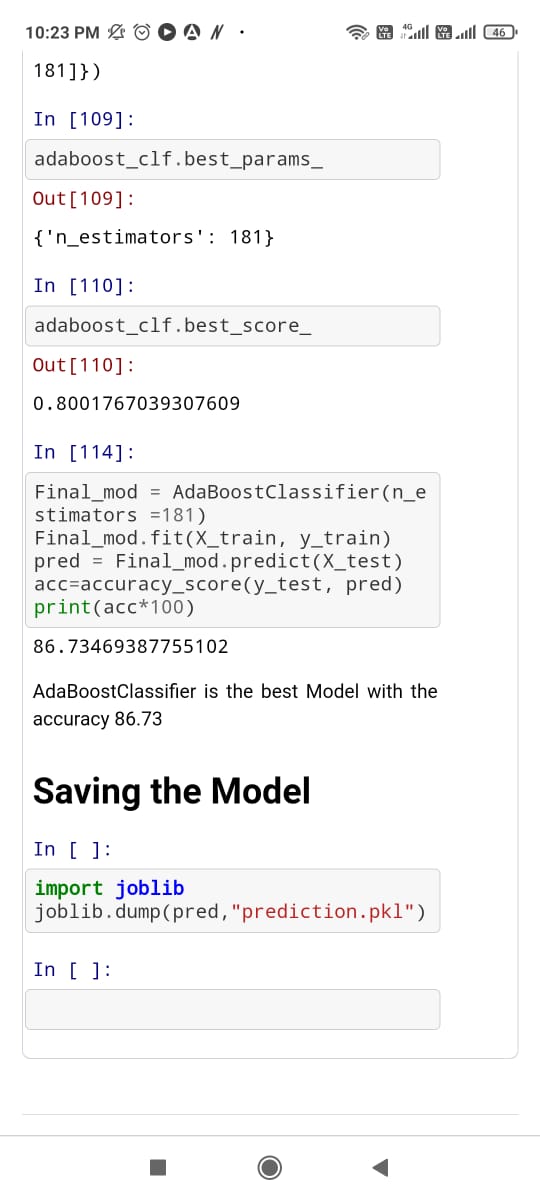
**Cross-validation-Score:**

****

There are two models with an accuracy score of 87 percent; however, AdaBoostModel is our best model as the difference between cross val score and accuracy score is minimum in AdaBoost. Now we will perform hyper parameter tunning to enhance our model.

**Hyper Parameter Tuning:**





As we can see, we have set several different parameters to do hyperparameter tunning; with the help of GridSearchCV. We trained our model with each parameter, then with the best parameters,we trained our model, and the accuracy score has resulted in 86.73 percent.

Conclusion:

In this paper, we have gone through the process of prediction model building. The paper showed how to clean data with various techniques and what will happen if we will not clean the data. We have also analysed the data graphically to find out insights from the given data. It was one of the main objectives. Also, we learned how to deal with imbalanced dependent data set as if we will not fix it, our model will not be good and will contain biases. This paper showed how to create a different type of models and how to evaluate them. Also, it showed how to approach with hyperparameter tuning of the selected model.