HR Analytics Project- Understanding the Attrition in HR

Table of Contents

[**1.** **Introduction** 2](#_Toc110790138)

[**2.** **Problem Definition** 2](#_Toc110790139)

[**3.** **Data Analysis** 2](#_Toc110790140)

[ **Checking for Null Values** 3](#_Toc110790141)

[**4.** **Data Visualization** 4](#_Toc110790142)

[ **Distribution of Target Variable - Attrition** 4](#_Toc110790143)

[ **Uni-variate Analysis of Nominal Columns** 5](#_Toc110790144)

[ **Distribution Plots for Continuous Columns** 7](#_Toc110790145)

[ **Bivariate Analysis** 9](#_Toc110790146)

[ **Relationship with the Target variable** 10](#_Toc110790147)

[ **Analysing Statical Summary of the Data** 12](#_Toc110790148)

[**5.** **Data Preparation & Cleaning** 12](#_Toc110790149)

[ **Correlation** 13](#_Toc110790150)

[ **Label Encoding** 13](#_Toc110790151)

[ **Checking Correlation with the Target Variable** 14](#_Toc110790152)

[ **Data Balancing** 15](#_Toc110790153)

[ **Treating Outliers** 15](#_Toc110790154)

[ **Treating Skewness** 17](#_Toc110790155)

[ **Treating Multicollinearity** 18](#_Toc110790156)

[ **Data Scaling** 19](#_Toc110790157)

[**6.** **Splitting the Data** 20](#_Toc110790158)

[**7.** **Testing with Different Models** 21](#_Toc110790159)

[**8.** **Ensemble Techniques** 22](#_Toc110790160)

[**9.** **Cross-Validation of Algorithms** 23](#_Toc110790161)

[**10.** **Hyperparameter tuning** 24](#_Toc110790162)

[**11.** **Cross-Validation of Best Model** 24](#_Toc110790163)

[**12.** **AUC ROC Curve** 25](#_Toc110790164)

[**13.** **Saving the best Model** 26](#_Toc110790165)

[**14.** **Conclusion** 26](#_Toc110790166)

HR Analytics Project- Understanding the Attrition in HR

1. **Introduction**

Attrition in human resources refers to the gradual loss of employees’ overtime. In general, relatively high attrition is problematic for companies. HR professionals keep on designing company compensation programs, work culture, and motivation systems that will help the organization retain top employees.

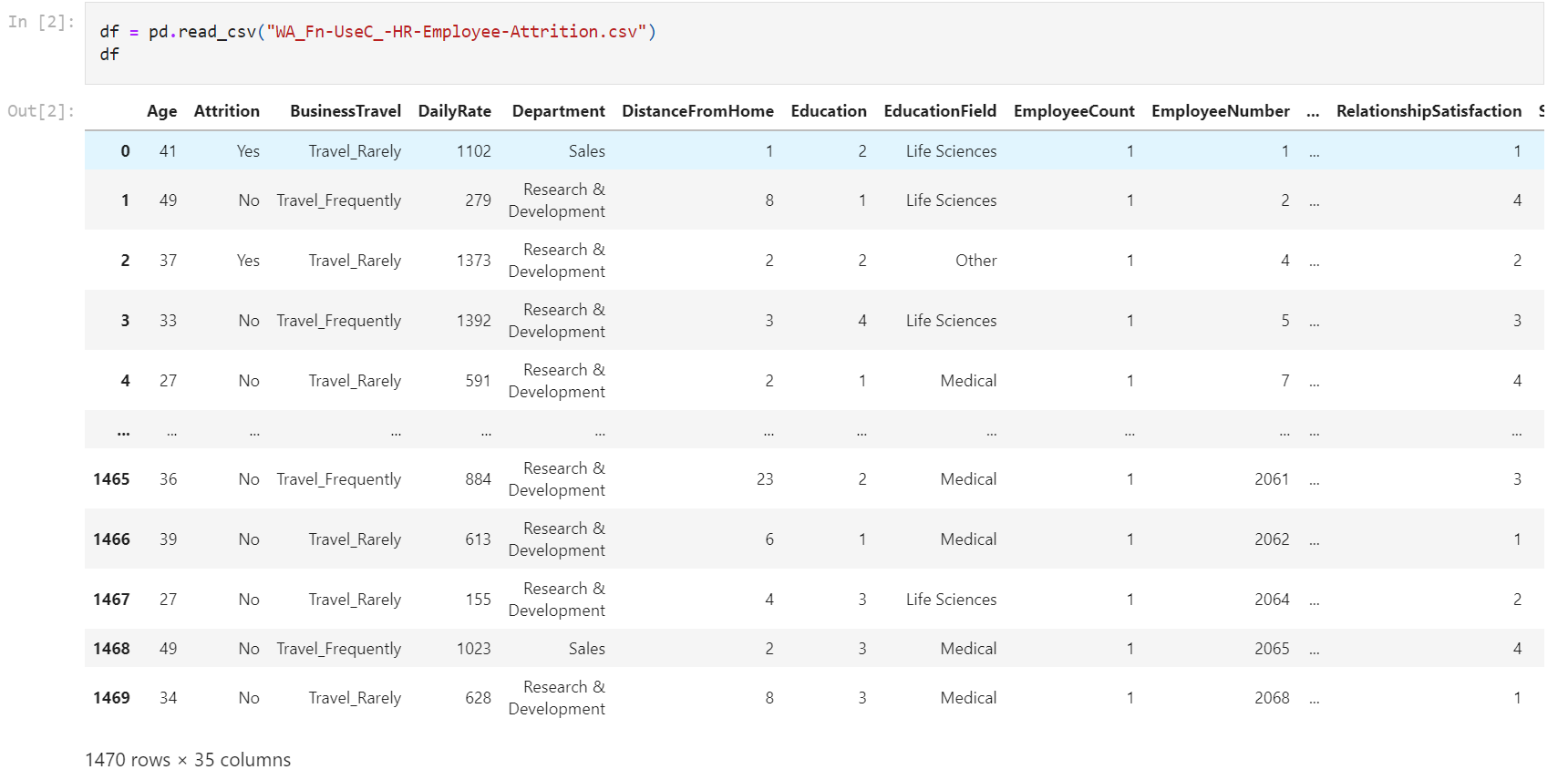
Attrition is critical in the business these days. It’s the main problem which is highlighted in all the organizations. Knowing that the organization loses key skills and knowledge. Managers and HR administrators are greatly interested in reducing the attrition in the organization in such a way that it will contribute to the maximum effectiveness and progress of the organization because employees are the most treasured assets of an organization.

1. **Problem Definition**

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 to happen if you constantly have new workers.

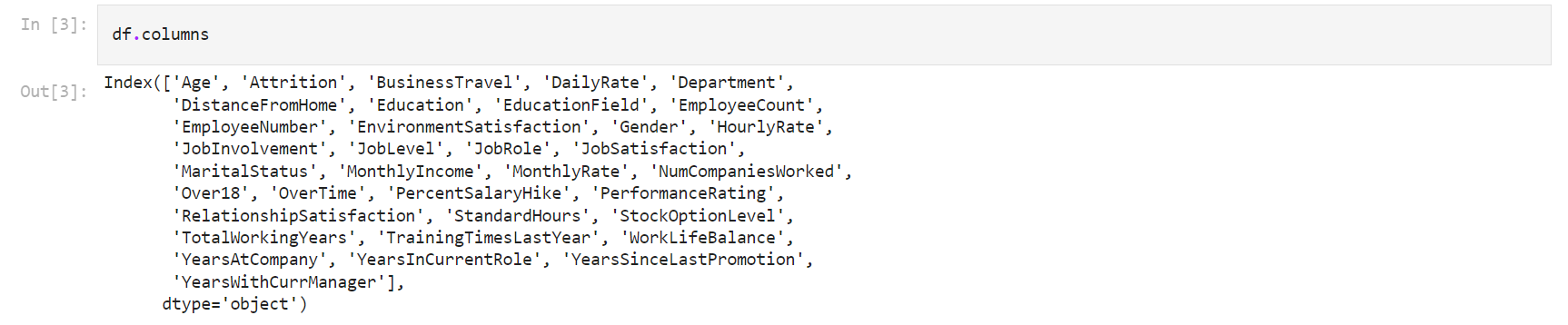
Therefore, the purpose of this exercise is to identify factors that lead to attrition and then predict employee attrition using multiple machine learning algorithms. The dataset has thirty-four independent variables. We will perform various EDA techniques to identify the patterns in the Data through data visualisation & use this Exploratory Data Analysis for building Machine Learning Algorithms.

1. **Data Analysis**

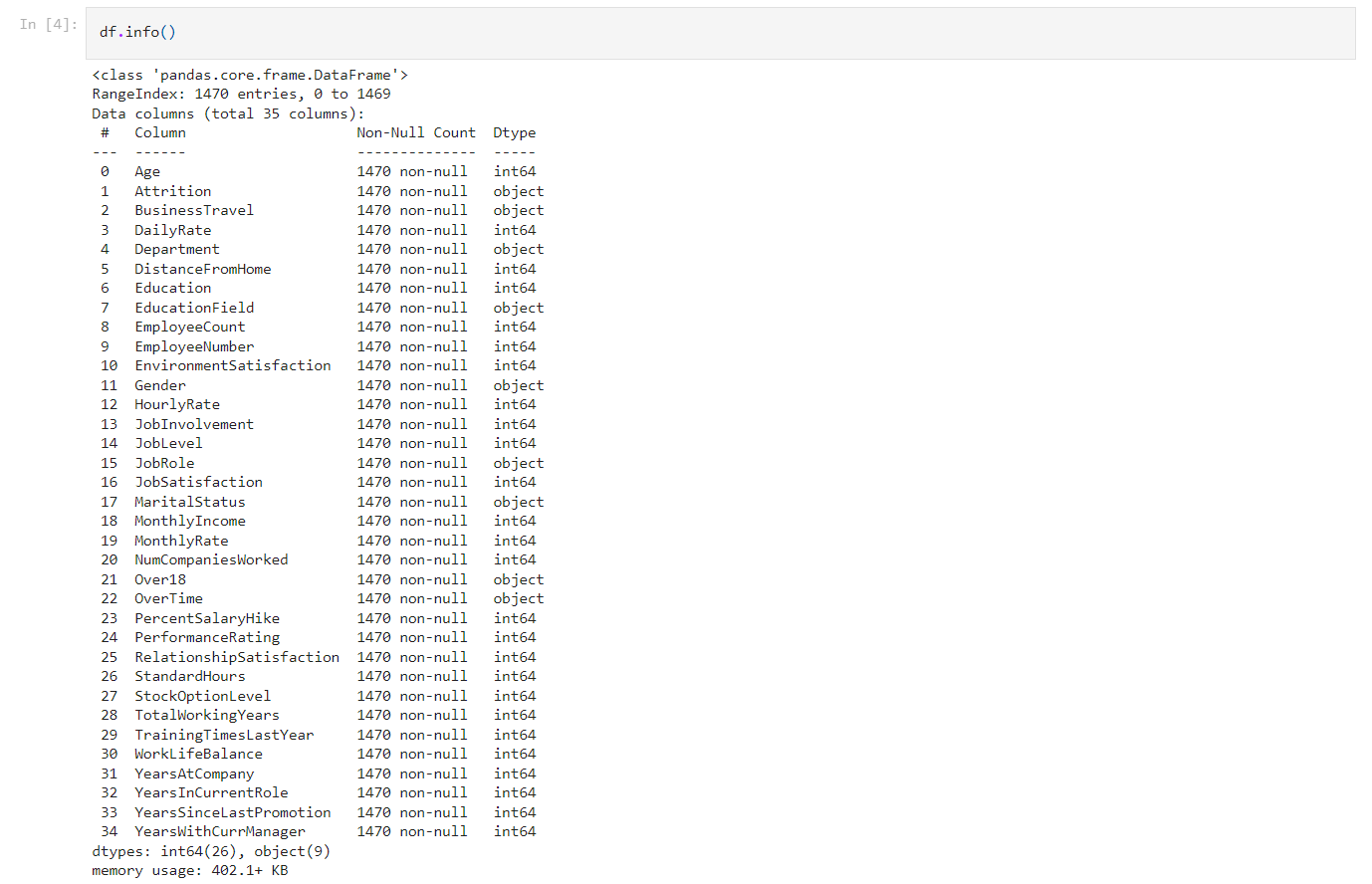


After importing all the necessary libraries. We will load the data into the python notebook.

We can see that the Dataset contains 1470 Rows & 35 Columns out of which ‘Attrition’ is our target column and all the rest 34 columns are Independent Variables.

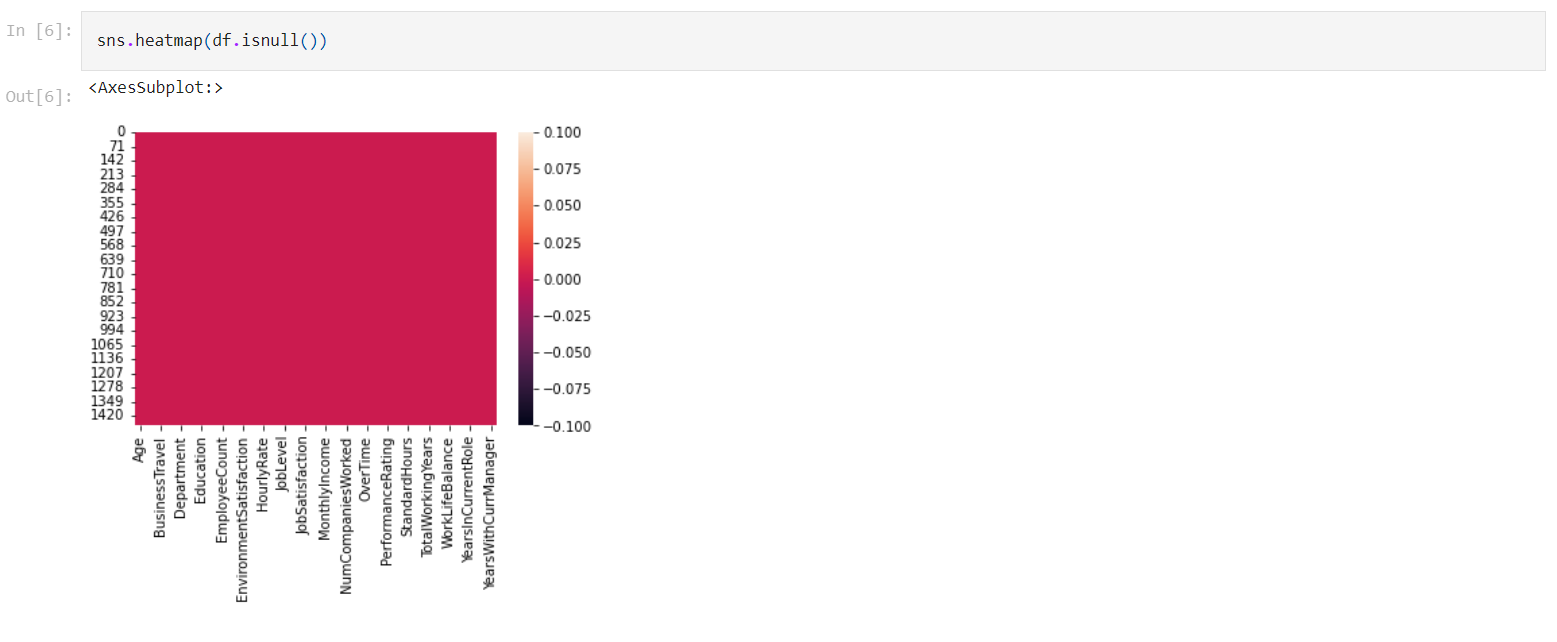


We can see that names of all the columns present in the Dataset.



We can see that ‘Attrition’ , ‘BusinessTravel’ , ‘Department’ , ’ EducationField’ , ’ Gender’ , ’ JobRole’ , ’MaritalStatus’ , ‘Over18’ & ‘OverTime’ columns are OBJECT type Data type. Therefore, we will change these values into integers using encoding techniques to use these features in the M.L. Model as Machine Learning only accepts numeric data.

* **Checking for Null Values**





We can see that that the Dataset is free from Null Values. So we can proceed further.

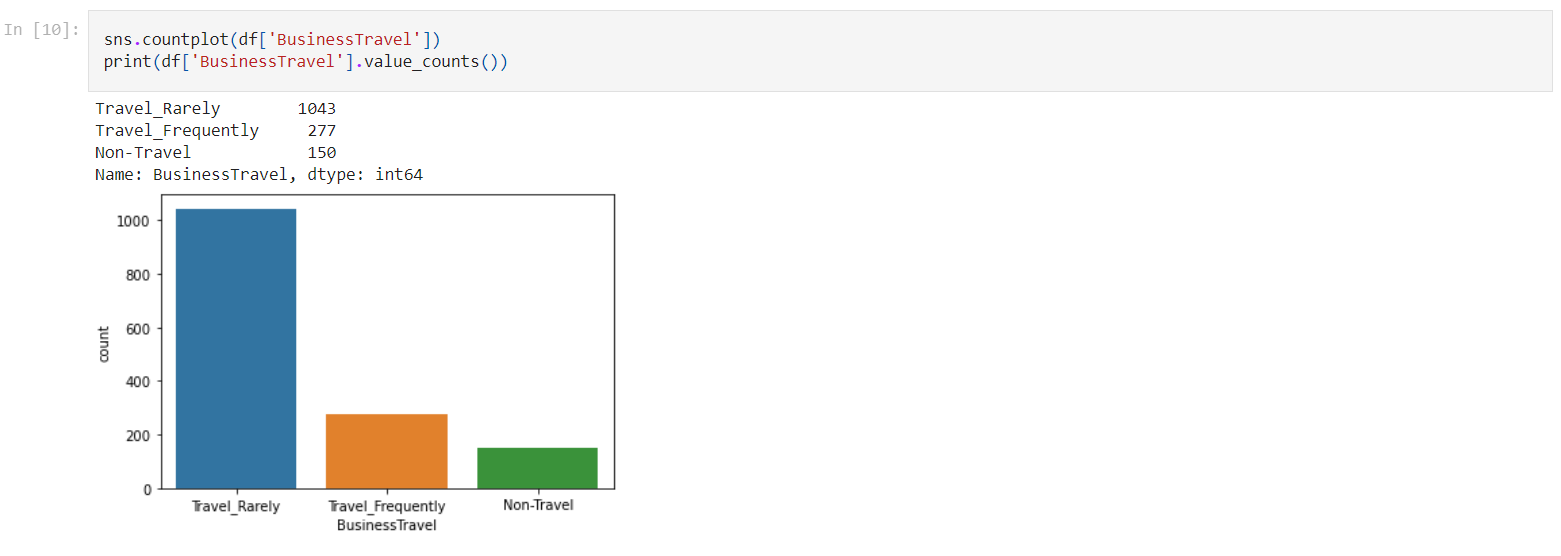
1. **Data Visualization**

* **Distribution of Target Variable - Attrition**



We can see from above that the organization has retained around 1233(84%) people while around 237(16%) people have left the organization.

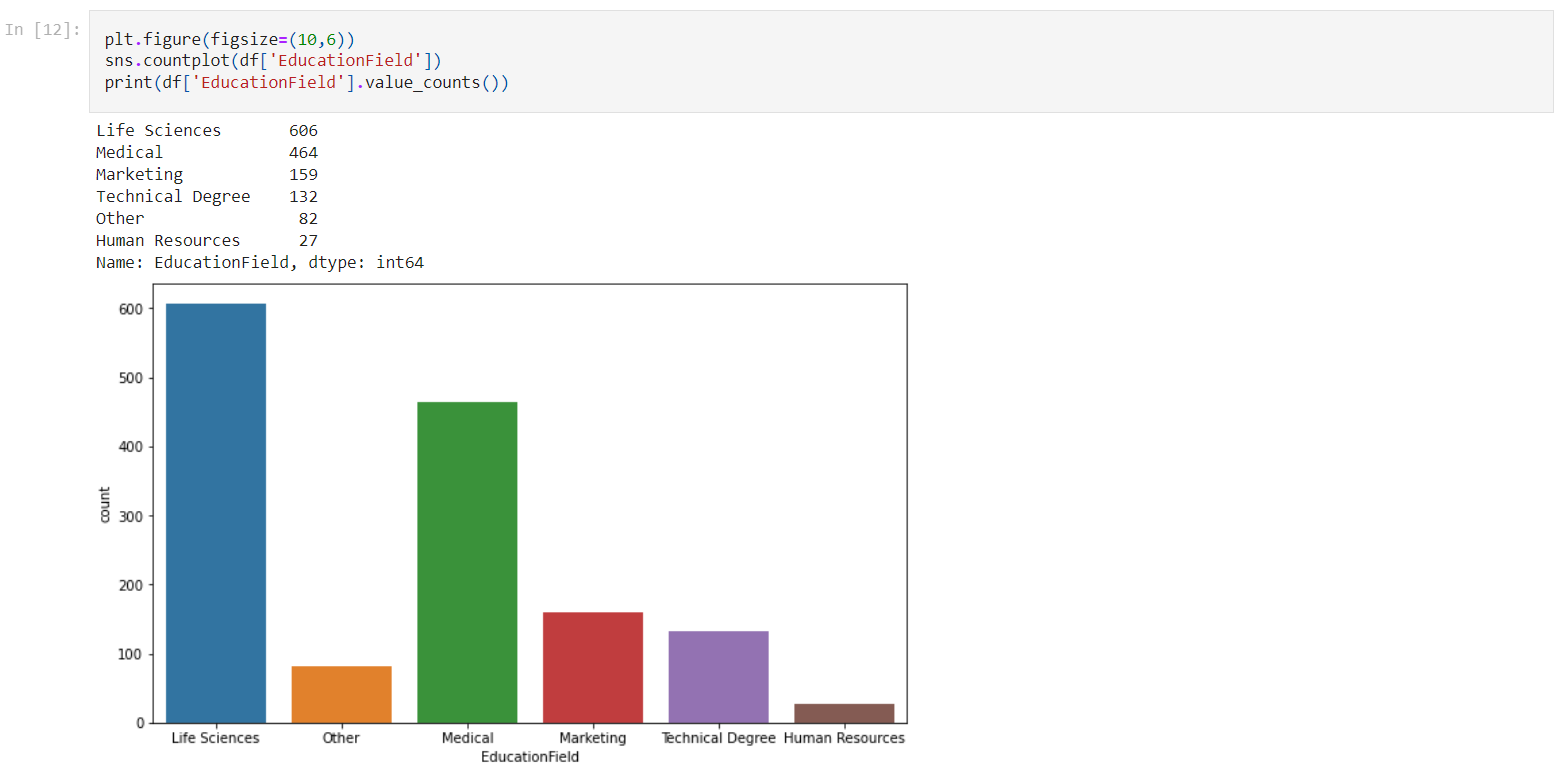
* **Uni-variate Analysis of Nominal Columns**



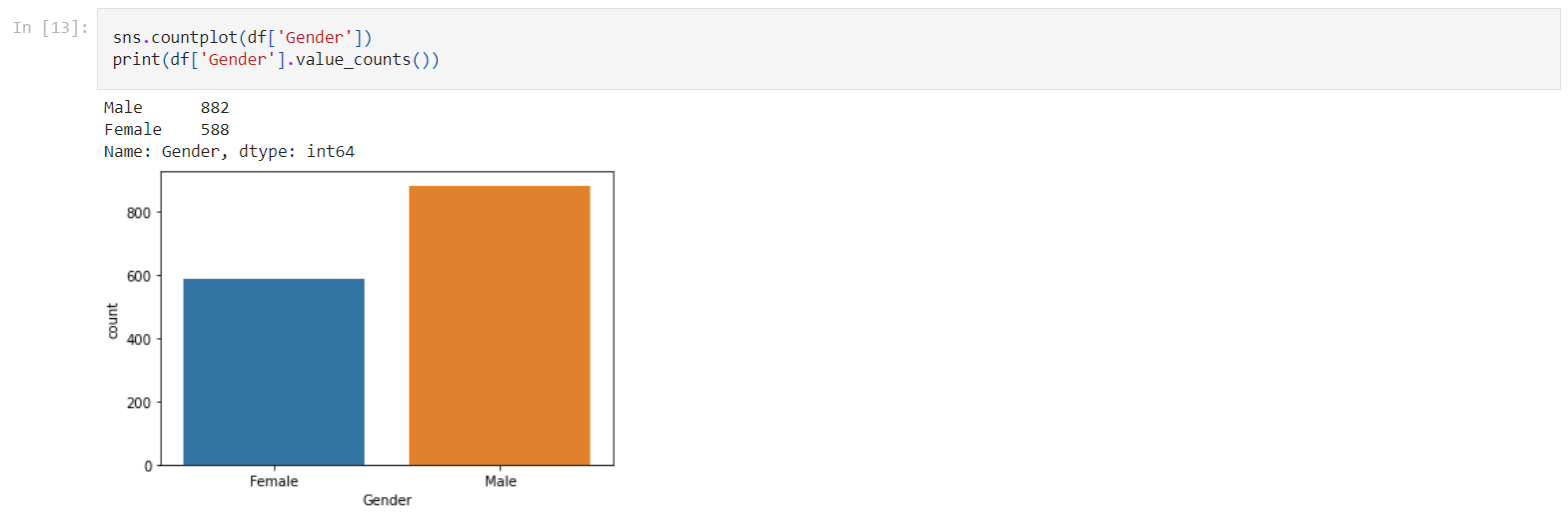
We can see that Most of the employees have to Travel Rarely for the organization.



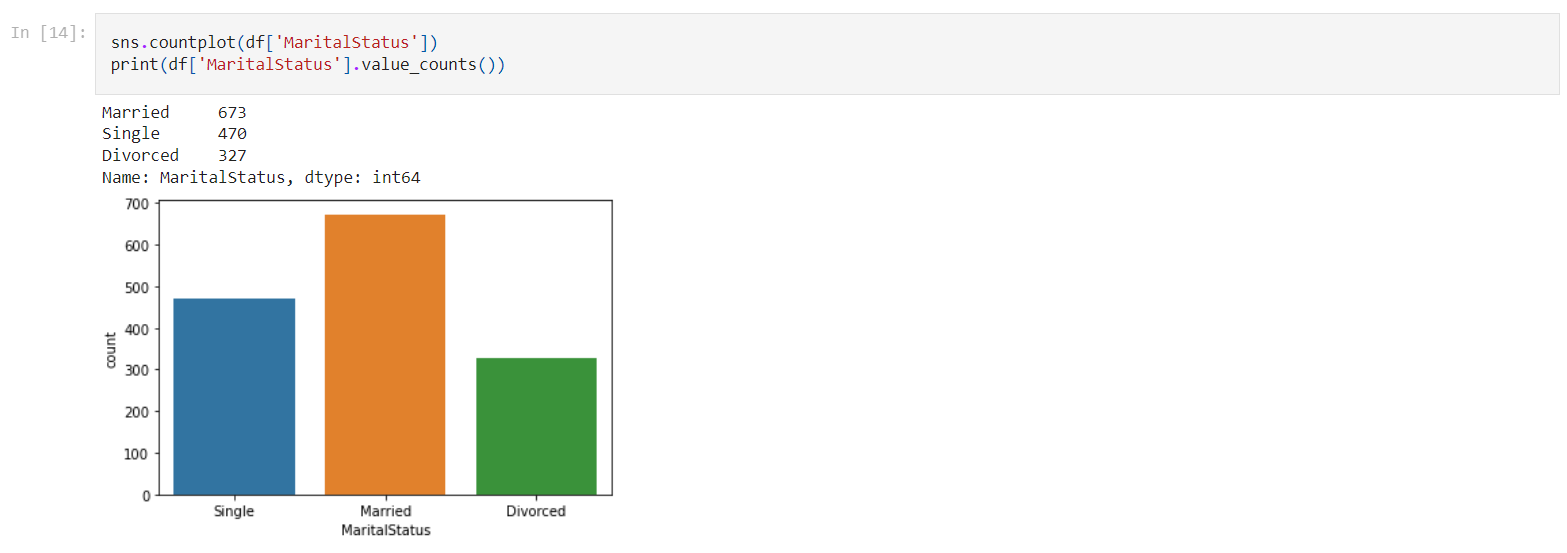
Most of the Employees belong to R&D department while the Human resource department contains least number of Employees.

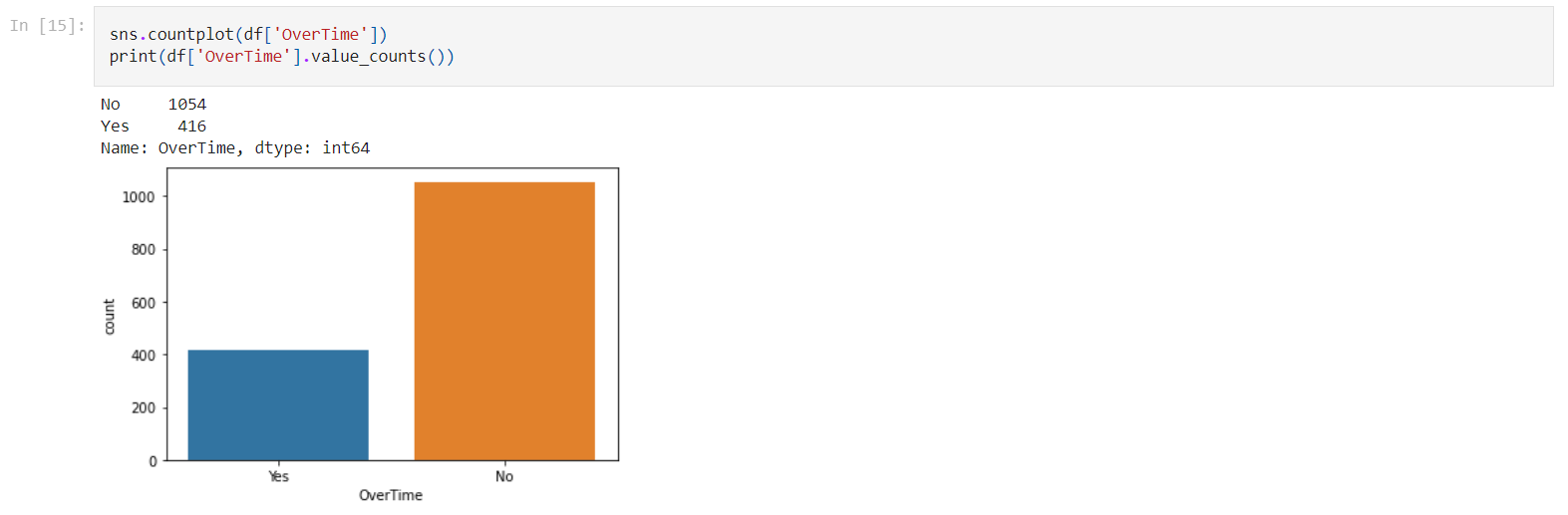


We can see that most of the employees in the organization belong 'Life Sciences' or 'Medical' Education Background.

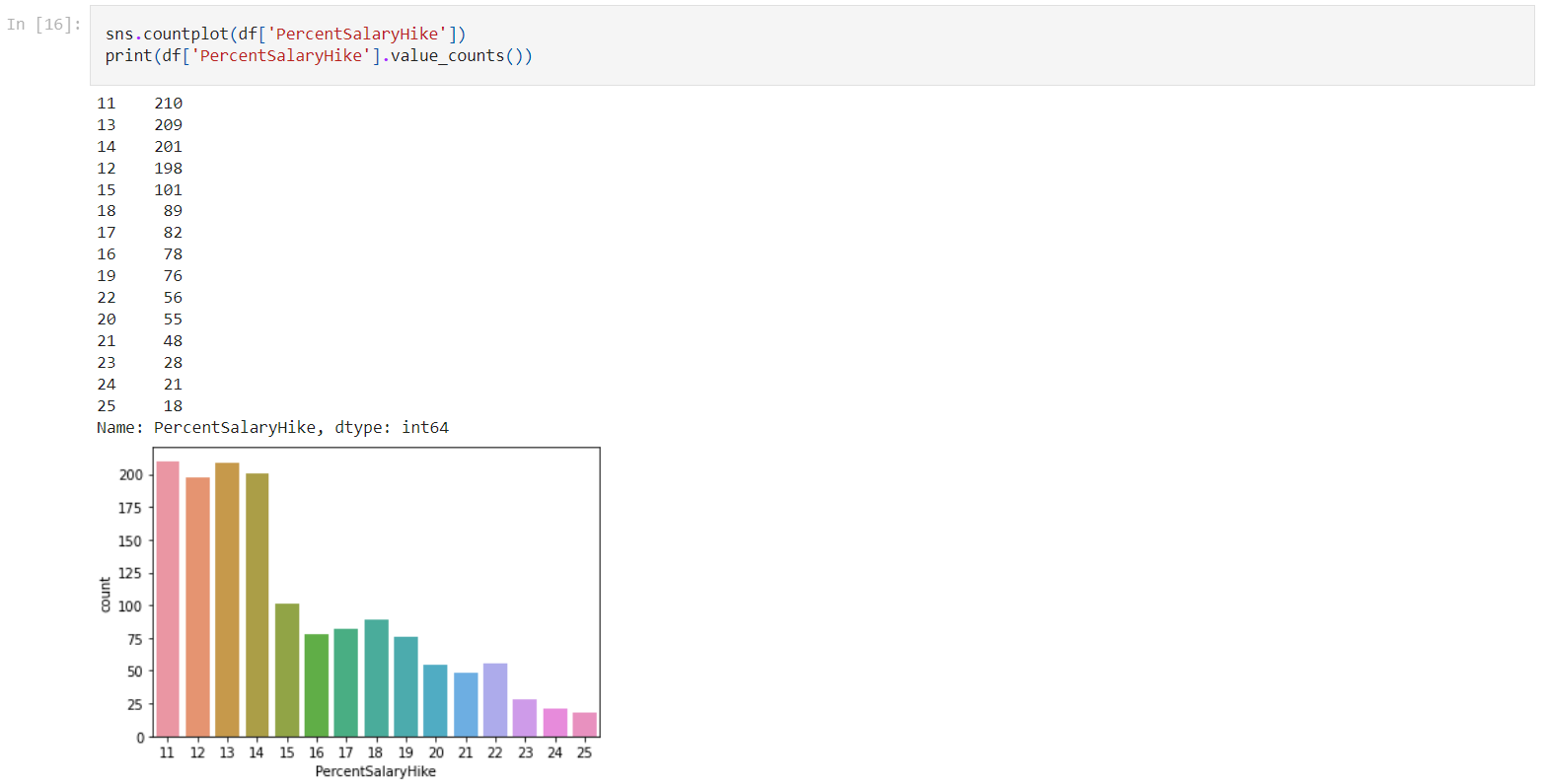


The organizations have very less Female employees as compared to Male employees.

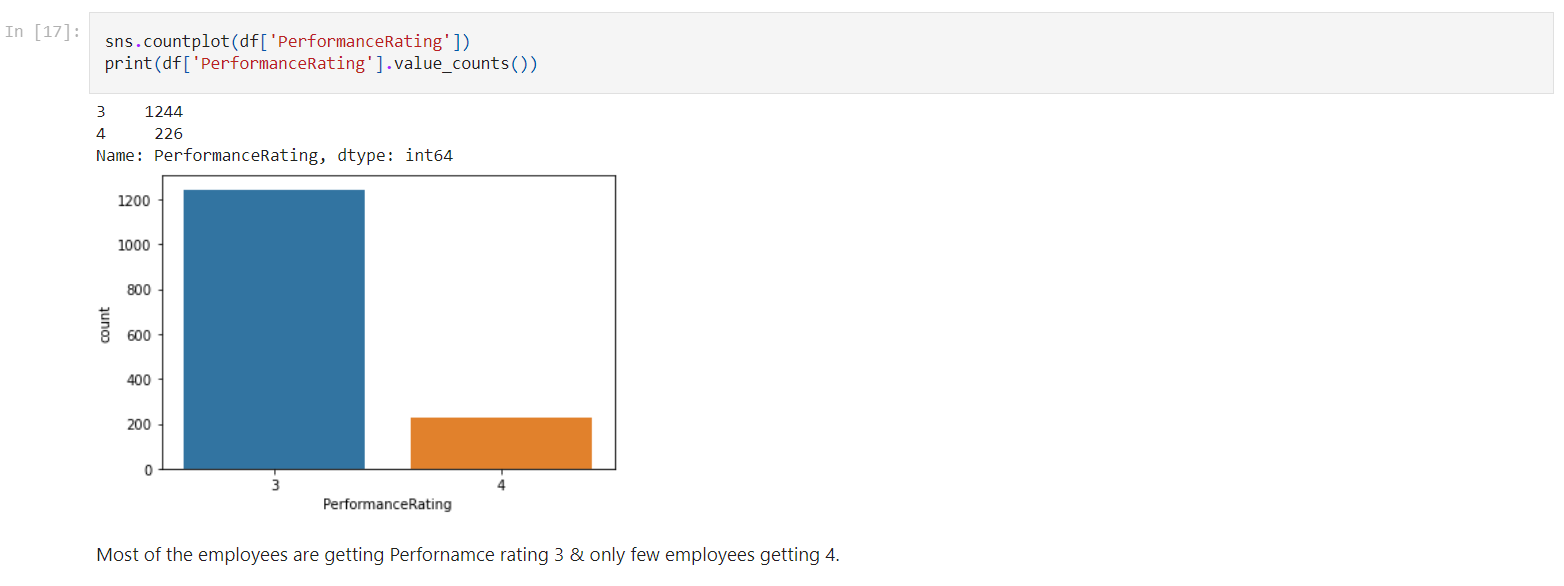
 Quantity of Employees who are Married are the Highest.



Most of the employees i.e. 1054 do not do overtime while only a small quantity of employees i.e. 417 work Overtime as well for the organization.

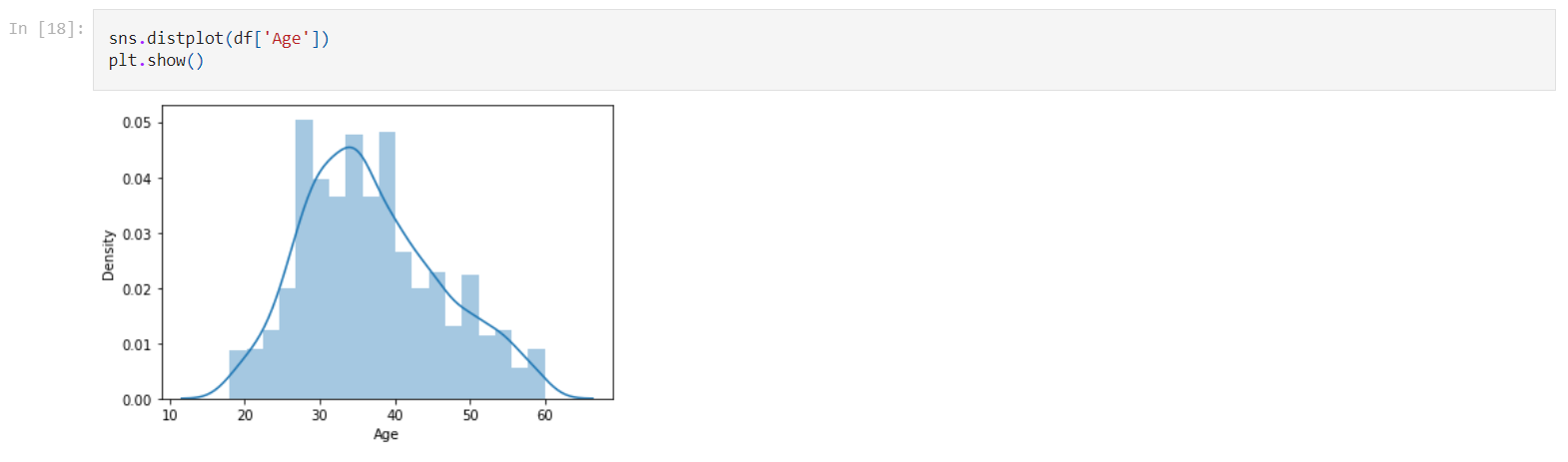


Most of the employees are getting only around 11%-15% Salary Hike while the employees getting salary hike b/w 20%-25% are least in number.

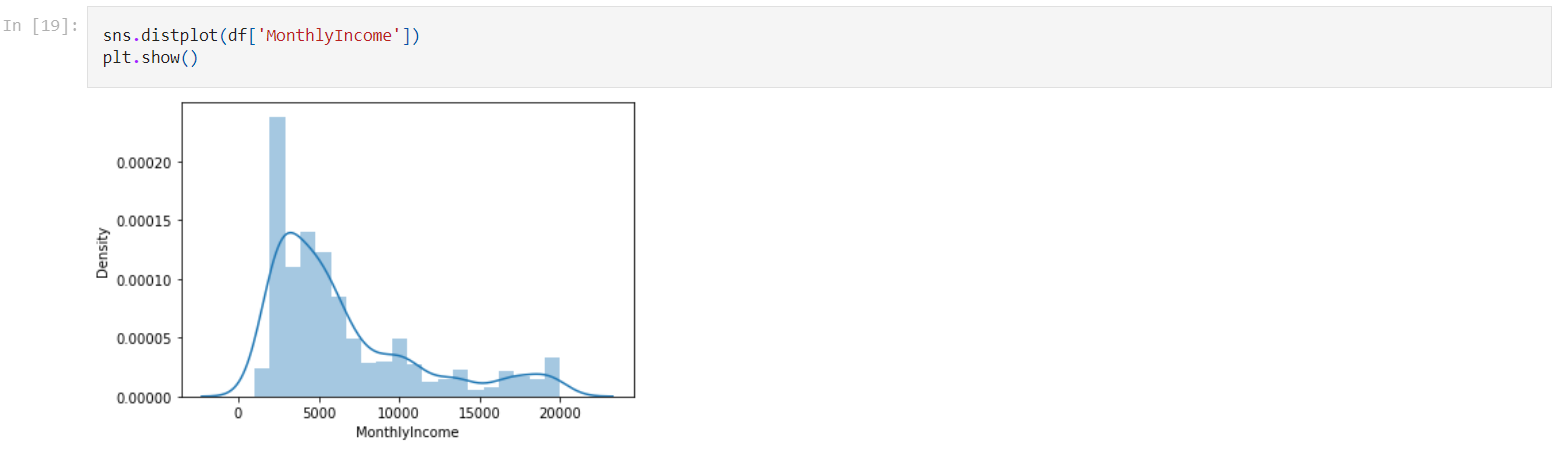


Majority of the employees are getting Performance rating 3 & only few of the employees getting 4.

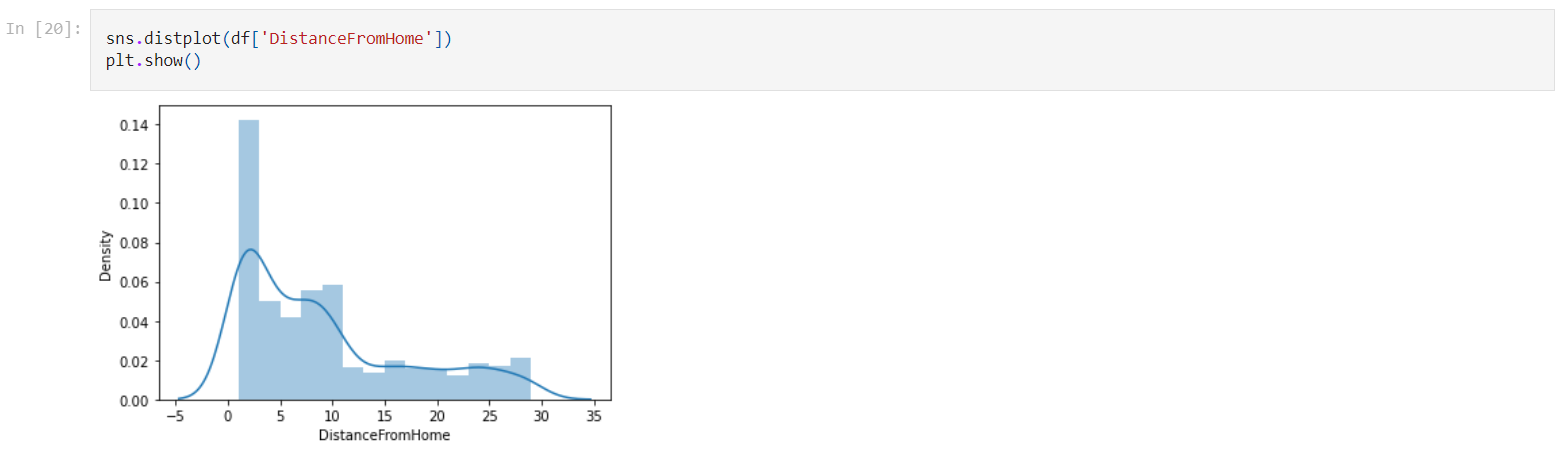
* **Distribution Plots for Continuous Columns**



We can see that most of the employees fall between 30-40 years of Age Range.



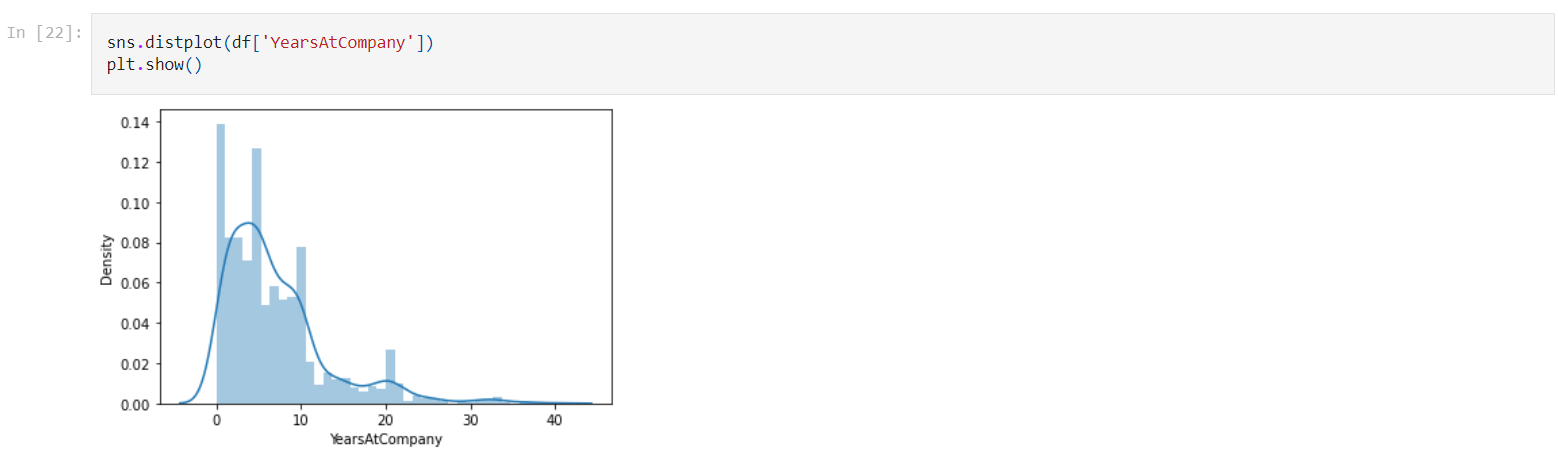
Majority of Employees are getting Monthly Income b/w 2000-5000 Units.



Large group of employees have the distance from home to the organization b/w 0-5 Units. And the density of employees decreases as we move towards 30 units.



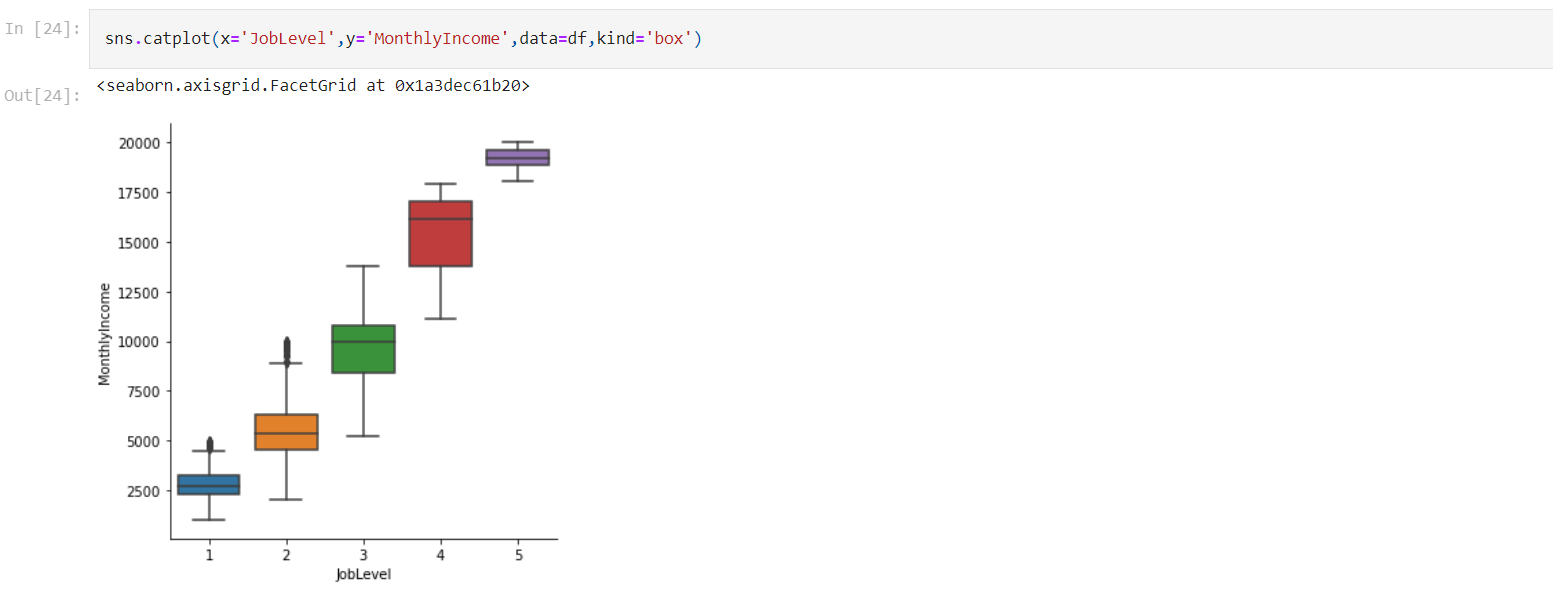
Majority of employees are having a Total Work Experience of 7-10 Years.



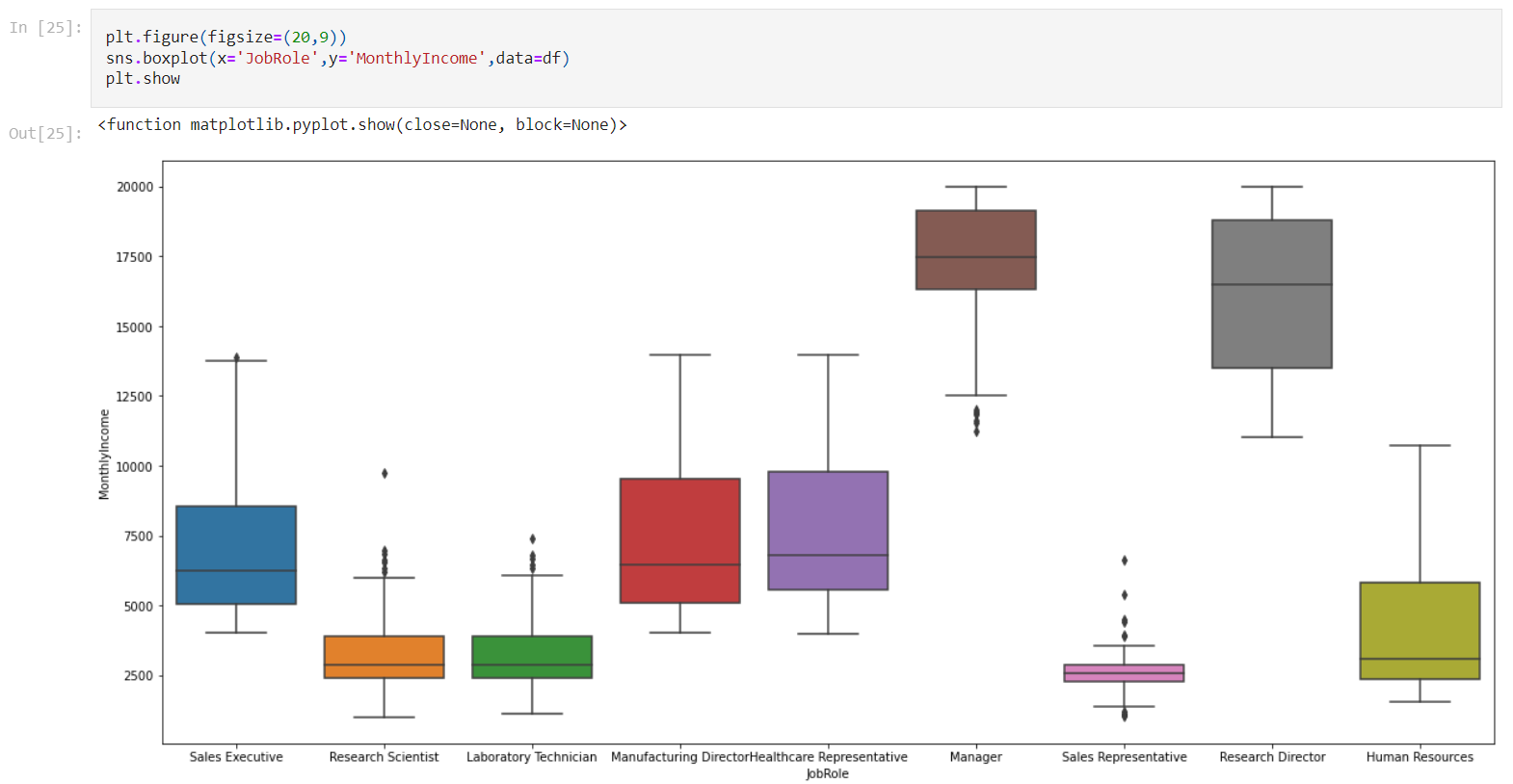
Majority of employees are working only b/w 0-5 Years for the company.

Also, We can see from above Distributions Plots that Skewness is present in MonthlyIncome, DistanceFromHome, TotalWorkingYears, YearsAtCompany & YearsInCurrentRole columns.

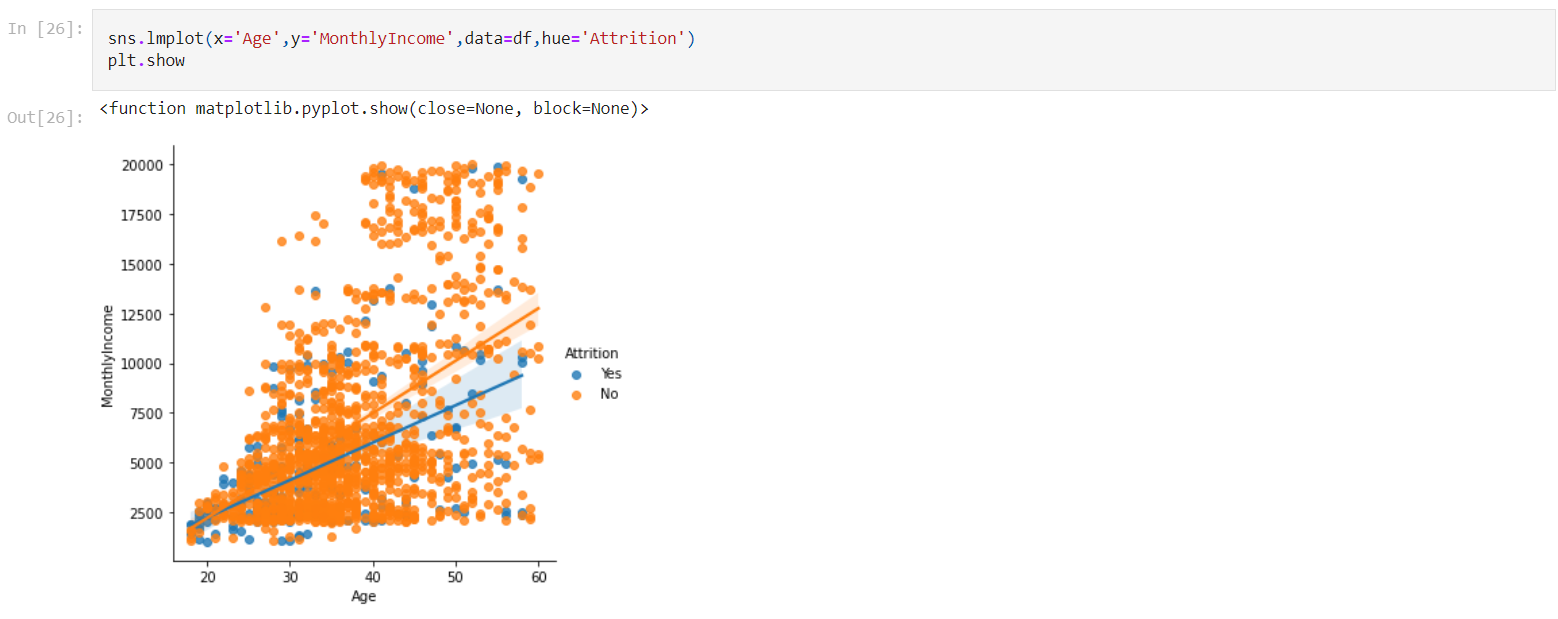
* **Bivariate Analysis**



From above plot we can clearly see big Difference in Monthly Income across different job levels.



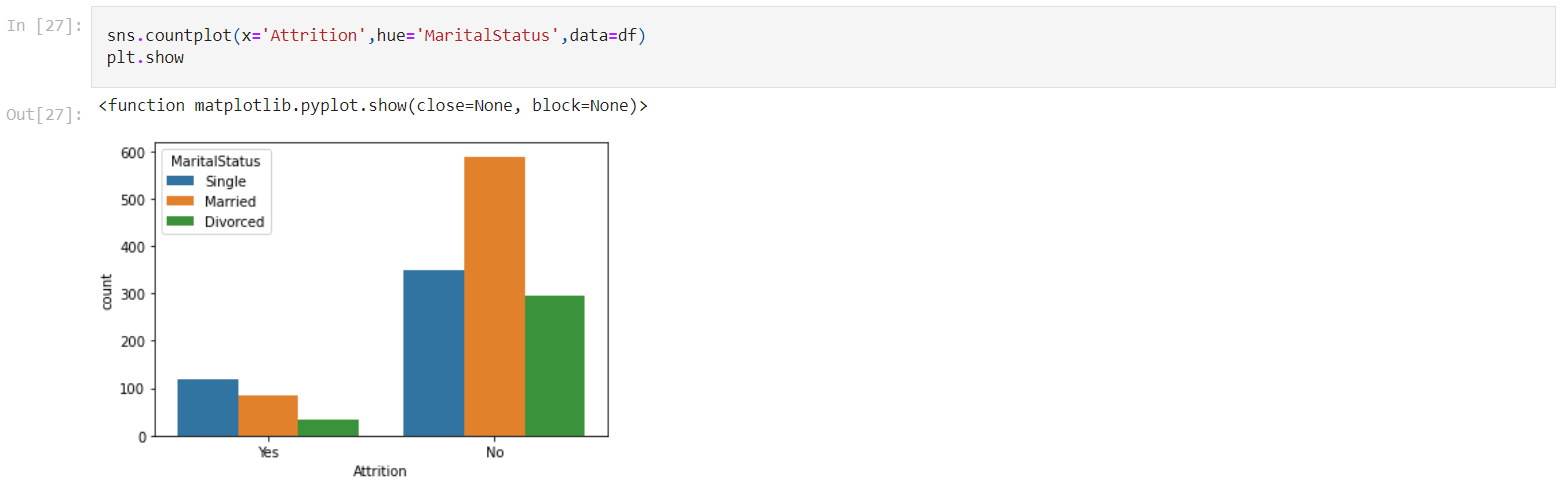
The 'Manager' & 'Research Director' are the highest paying jobs in the company while 'Sales Representative’, 'Research Scientist' & 'Laboratory Technician' employees are earning least.



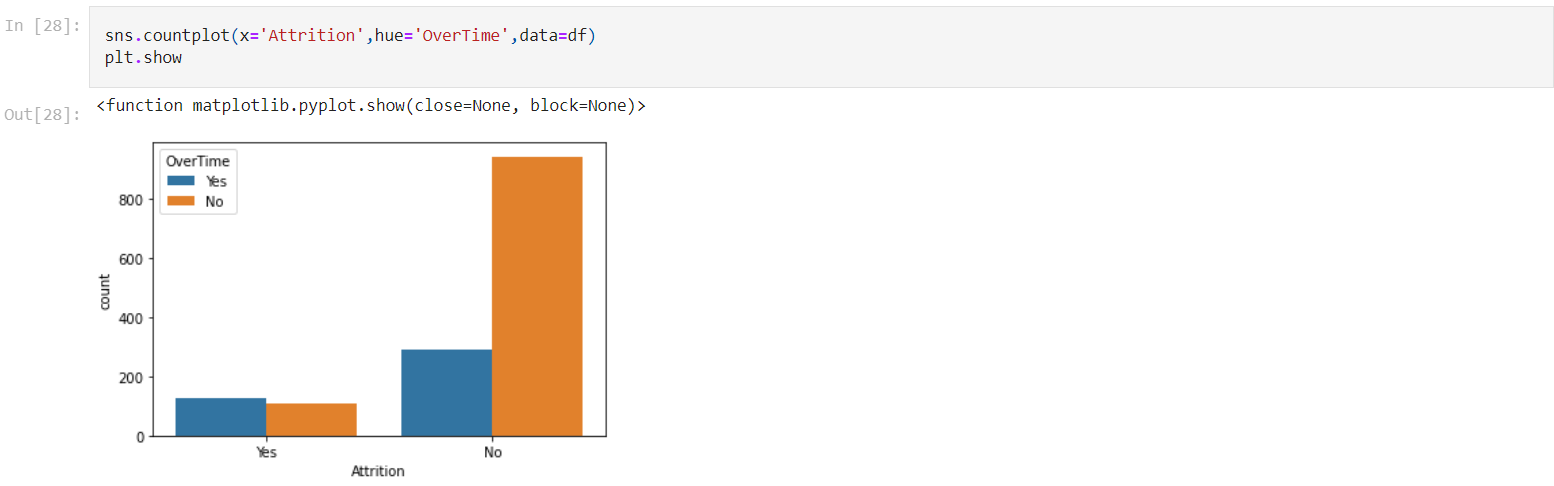
We can see that Monthly Income increases with increase in age. However the increase is not same for both Attrition groups. And the employees that leave the organization tend to have higher Income with age.

* **Relationship with the Target variable**

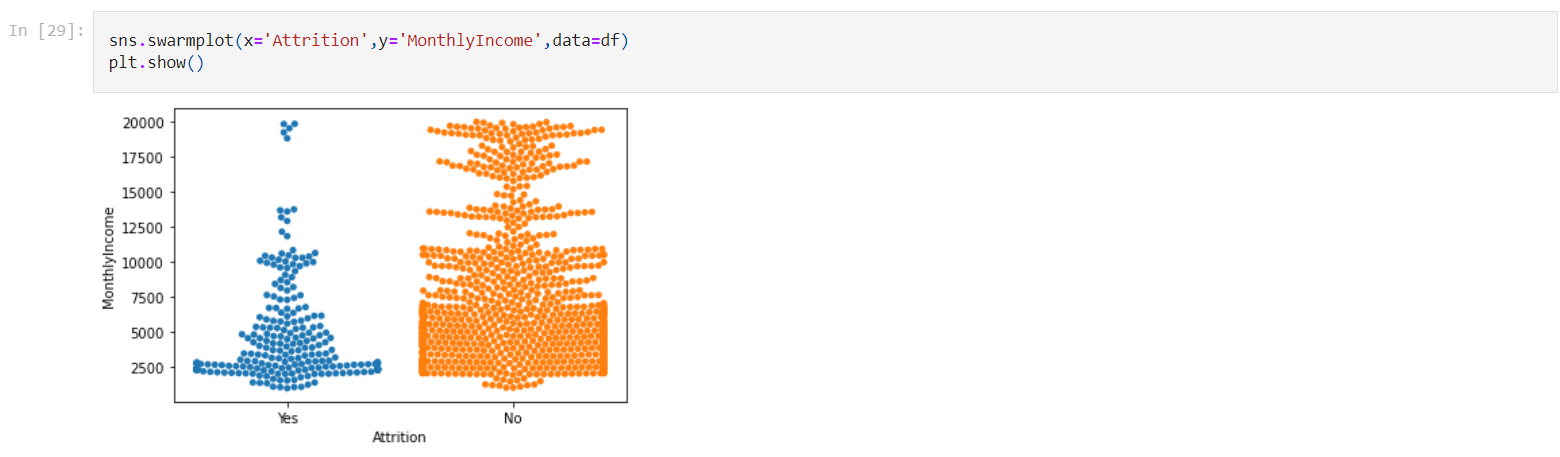
Now let’s explore how the dependent variables are related to the Target i.e. Attrition Column.



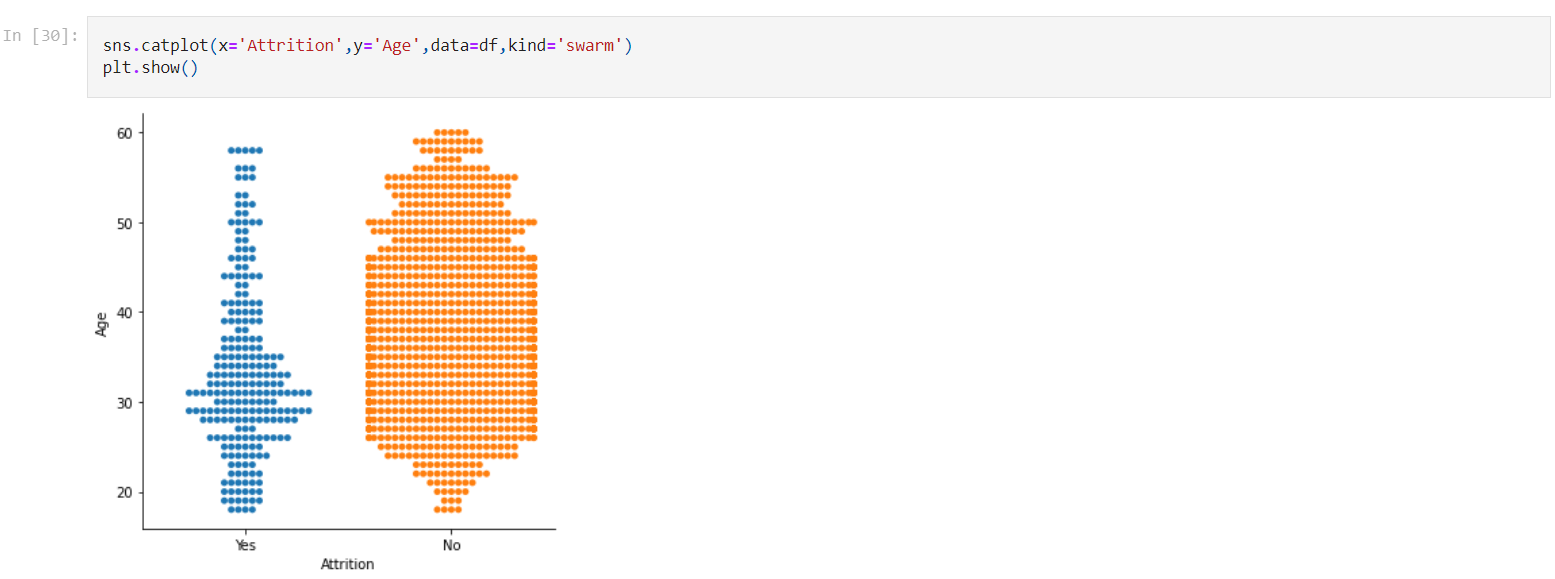
From above plot we can see that the 'Single' people are more likely to leave the organization than the ones that are married or divorced.



We can see that the employees working overtime for the company are more likely to leave the organization than the ones who are not doing overtime.



From the above Plot we can see that the People with Low Monthly Income i.e. b/w 2000 - 3000 Units are more likely to leave the organization than the ones earning more.

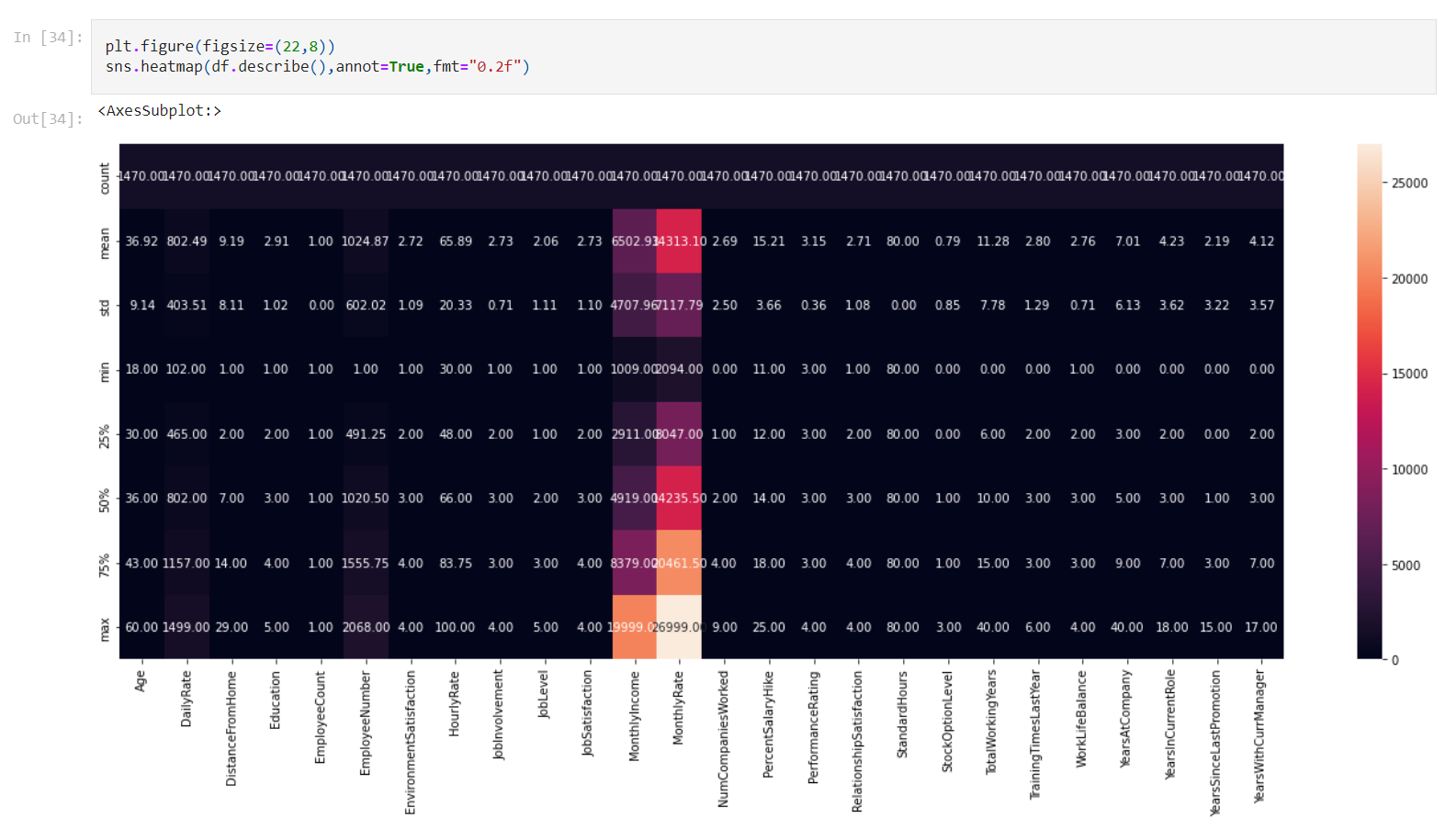


From above plot we can see that Attrition is highest in people b/w age group 28-32.



We can see that Attrition is highest for employees getting Salary Hike of only around 12.5% & it gradually decreases as the Salary Hike increases.

* **Analysing Statical Summary of the Data**

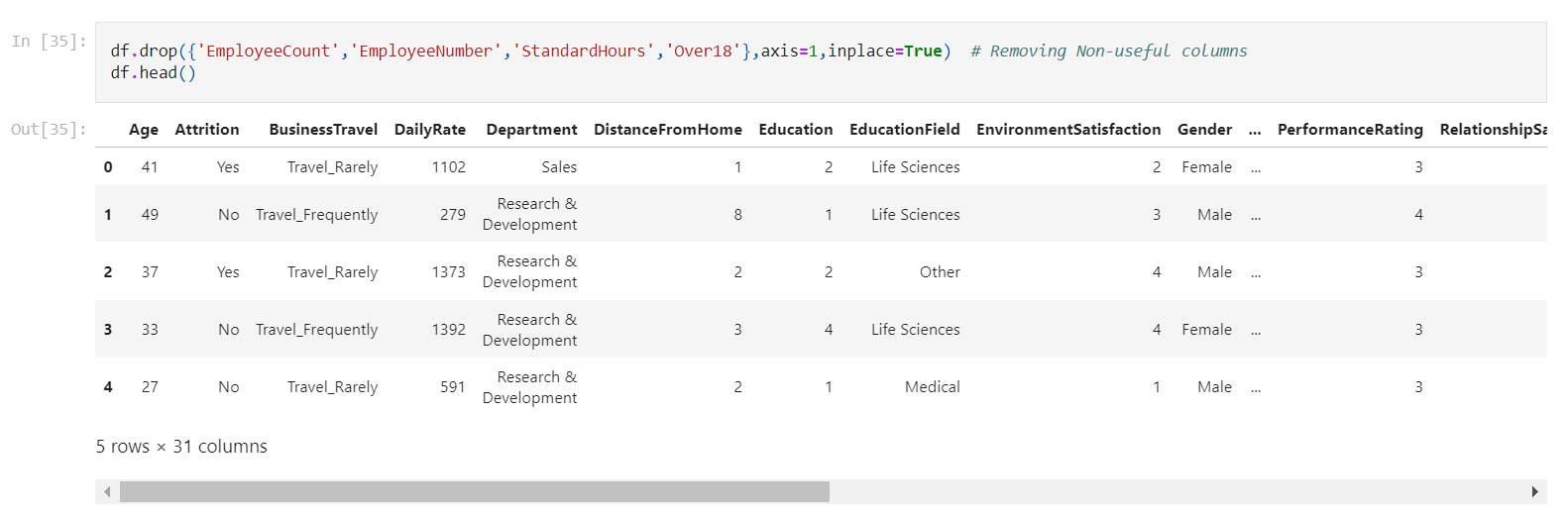


Observations : -

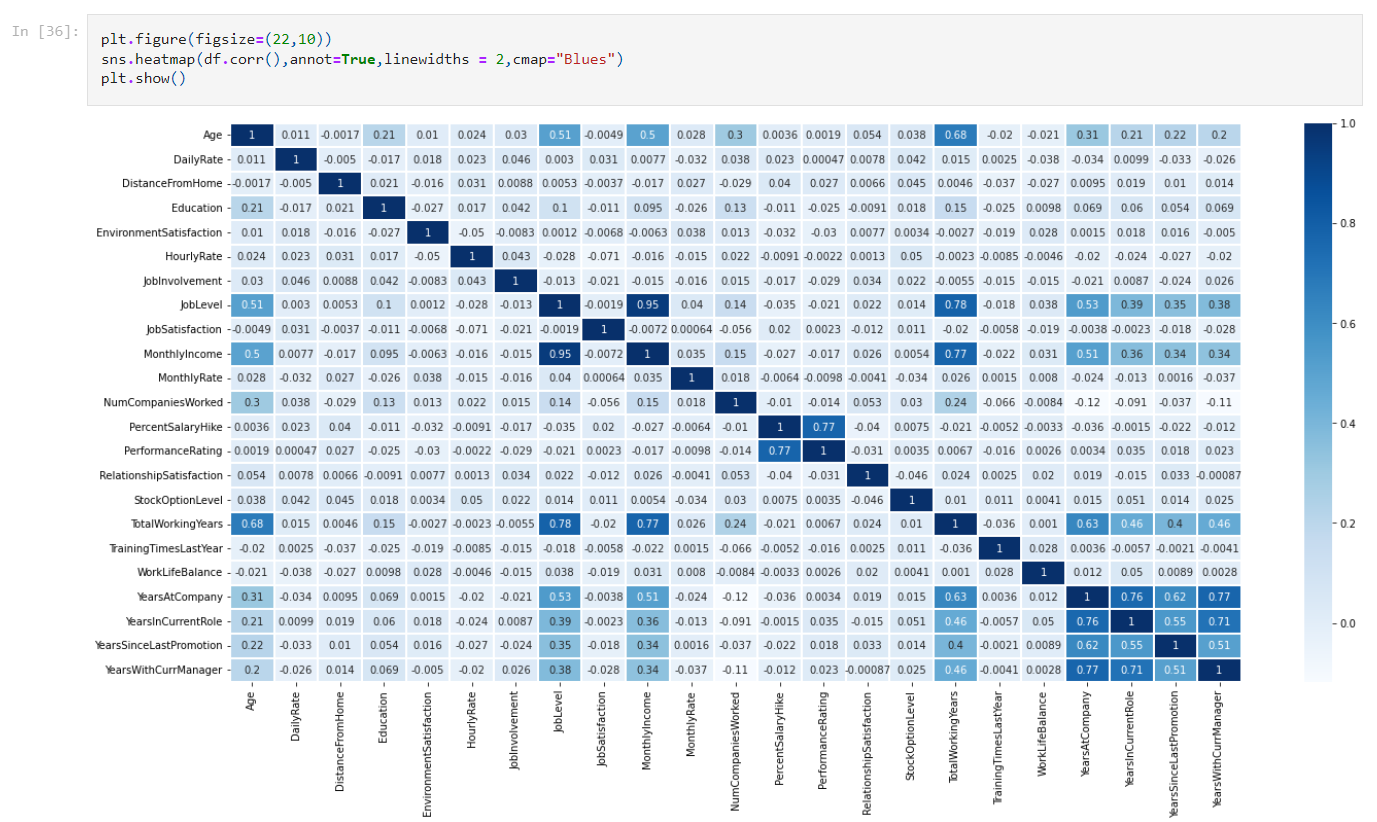
1. Mean > Median, Hence the data is right skewed for columns :- 'DistanceFromHome' & 'years At company'. We will treat this skewness before building a model.
2. Also big difference is present b/w 75th percentile & max value of the column for ‘MonthlyIncome’ , ‘TotalWorkingYears’ , ‘YearsAtCompany’ , ‘YearsInCurrentRole’ , ‘YearsSinceLastPromotion’ & ‘YearsWithCurrManager’ which means Outliers are present in the Data & we have to remove these outliers for building a better Model.
3. **Data Preparation & Cleaning**

EmployeeCount','EmployeeNumber','StandardHours' & 'Over18' Feature Columns are not useful for the model. As they are not provide any useful information in determining the Attrition Rate.

Therefore, we will drop these columns.



* + **Correlation**



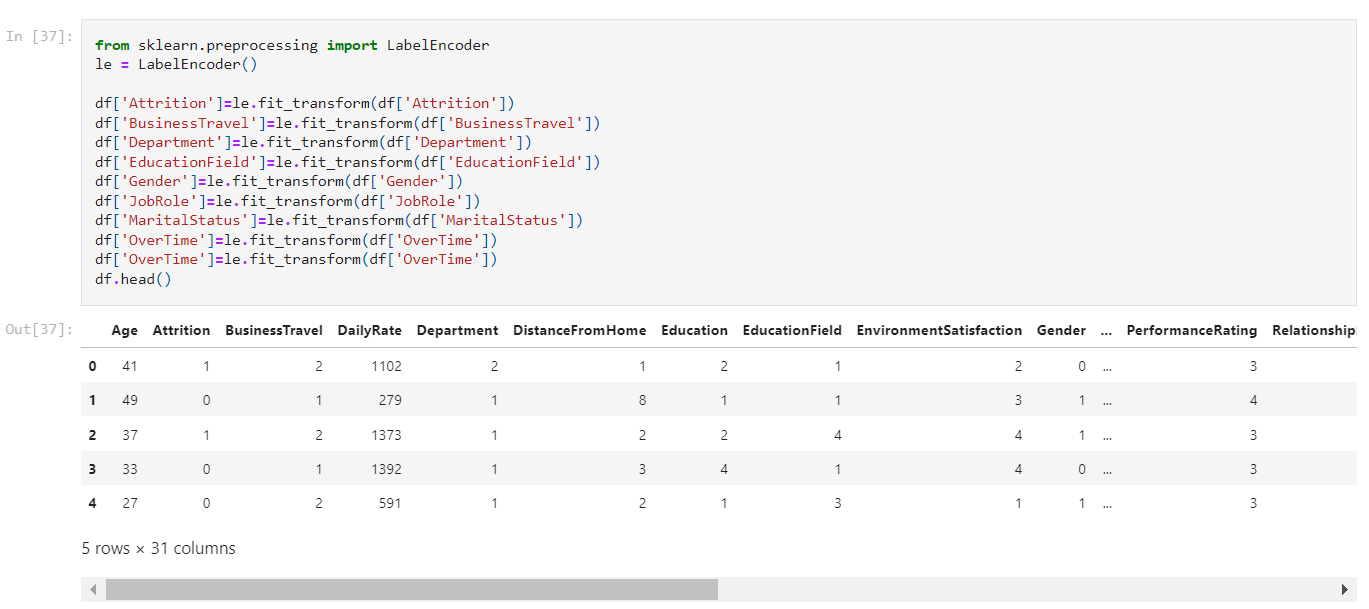
From the correlation chart we can see that the some columns are having high correlation(>0.7) b/w them. These columns are:-

1. Years At Company & Years In current role
2. Years At Company & Years With current manager
3. Years In current role & Years With current manager
4. Monthly Income & Total Working Years
5. Monthly Income & Job Level
6. Job Level & Total Working Years
7. Percent Salary Hike & Performance Rating
8. Age & Total Working Years

It means there is Multicollinearity present in the Data. Therefore, we will use VIF Values to determine this Multicollinearity & remove it from the dataset.

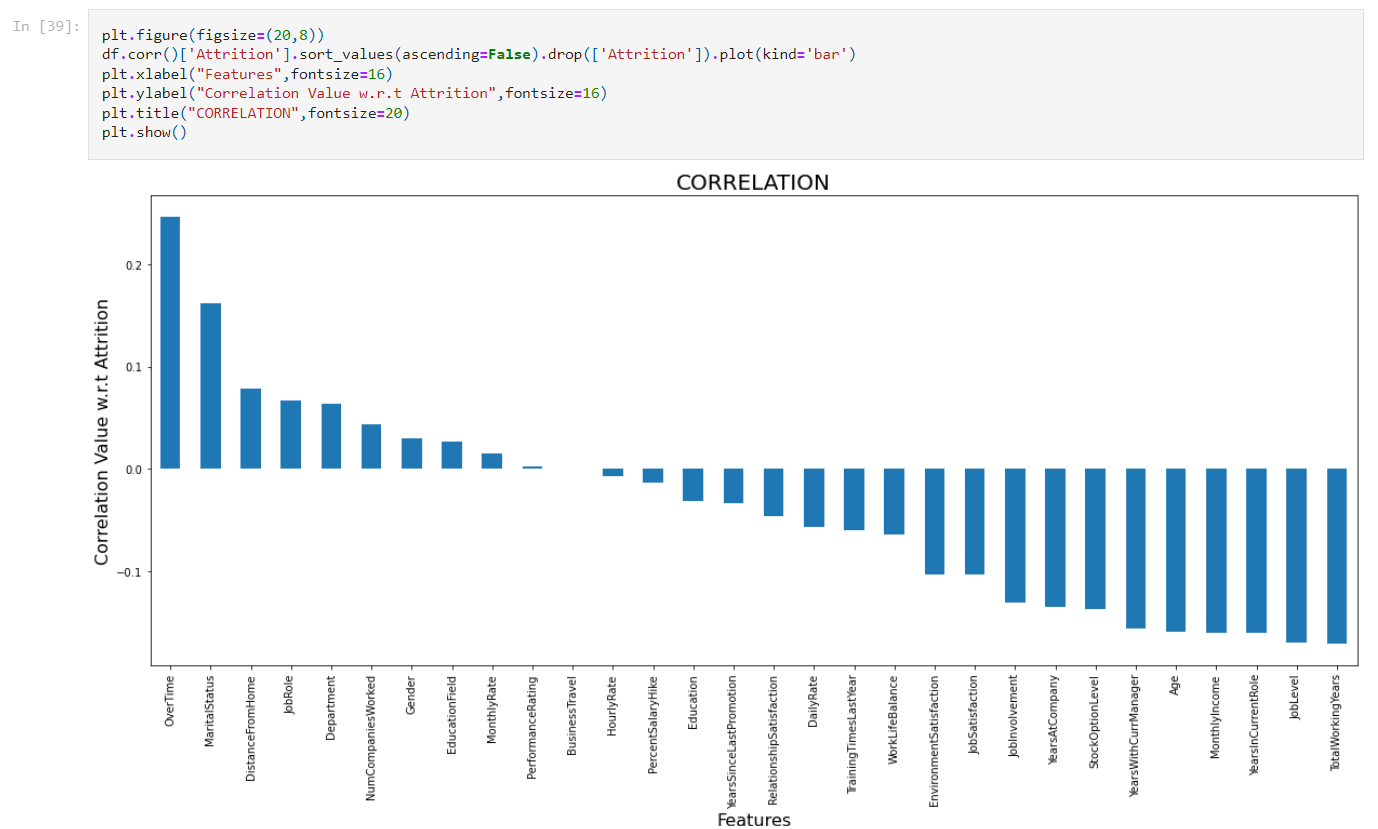
* + **Label Encoding**

We will convert the Columns with String values into integer values to use those features into M.L. model.



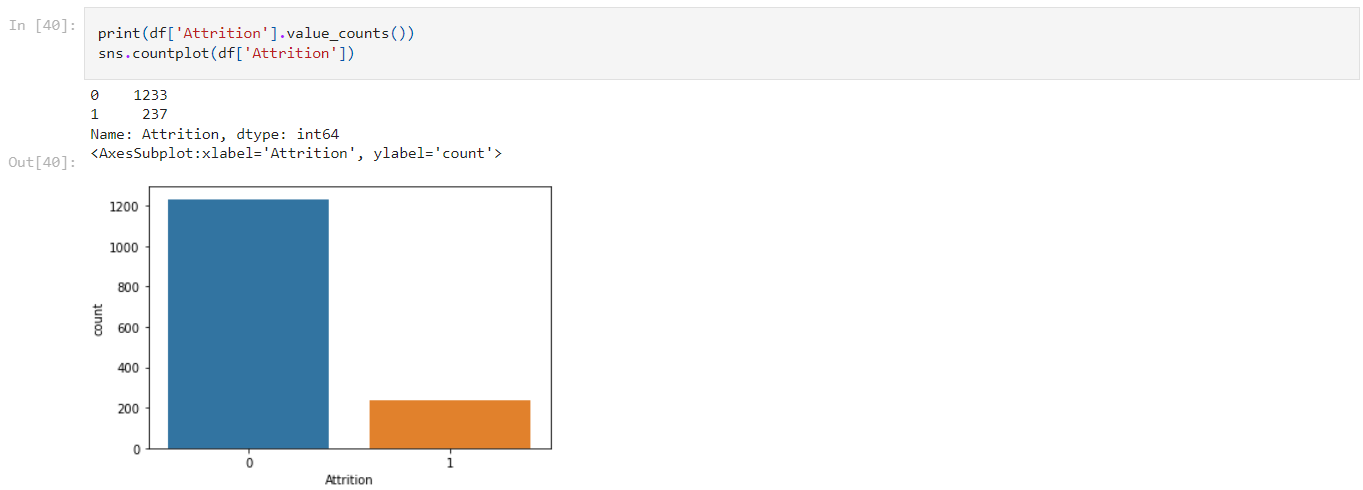
We have converted the ‘Attrition’ , ‘BusinessTravel’ , ‘Department’ , ‘EducationField’ , ‘Gender’ , ‘JobRole’ , ‘MaritalStatus’ & ‘OverTime’ columns into Numeric Values.

* + **Checking Correlation with the Target Variable**



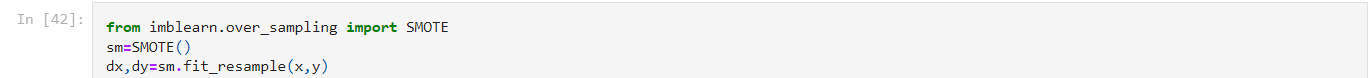
Observations : -

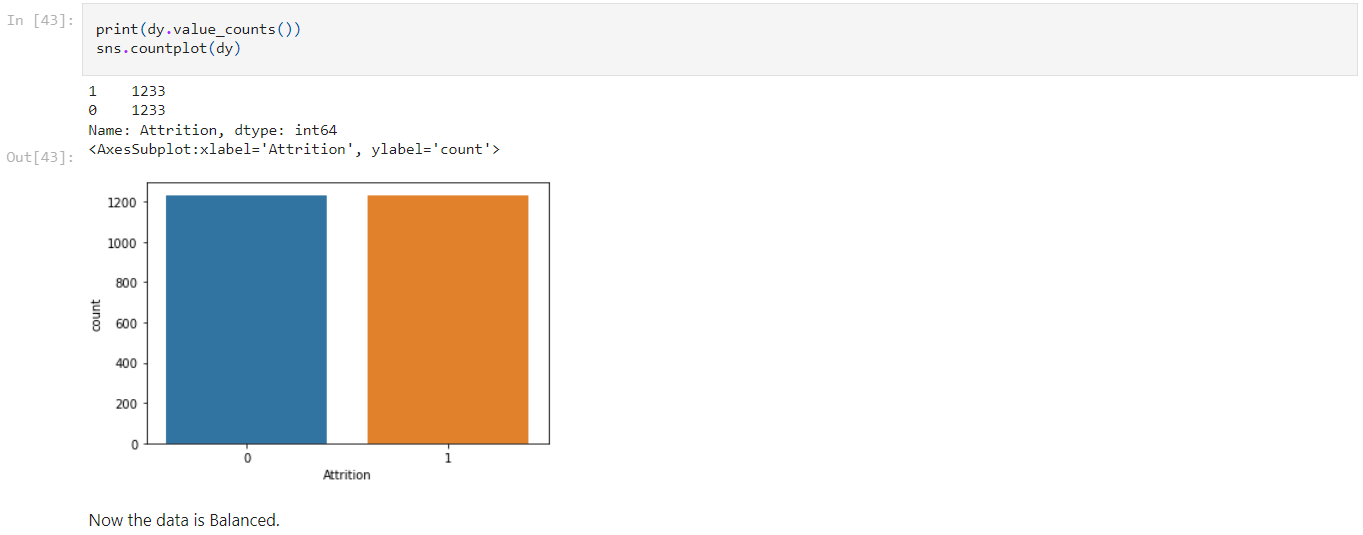
1. 'Overtime' is showing highest +ve correlation with Attrition whereas Total Working Years has highest -ve correlation with the target column.
2. 'Business Travel' has very least or no correlation with Atrrition also ‘HourlyRate’ & ‘PerformanceRating’ show very less correlation with the Target Varible i.e. Attrition.
   * **Data Balancing**



As we can see from above that It is an Imbalanced Dataset because the No. of Entries for Attrition 'YES' is way more than for 'NO' values.

Therefore we will use SMOTE to balance the Data.

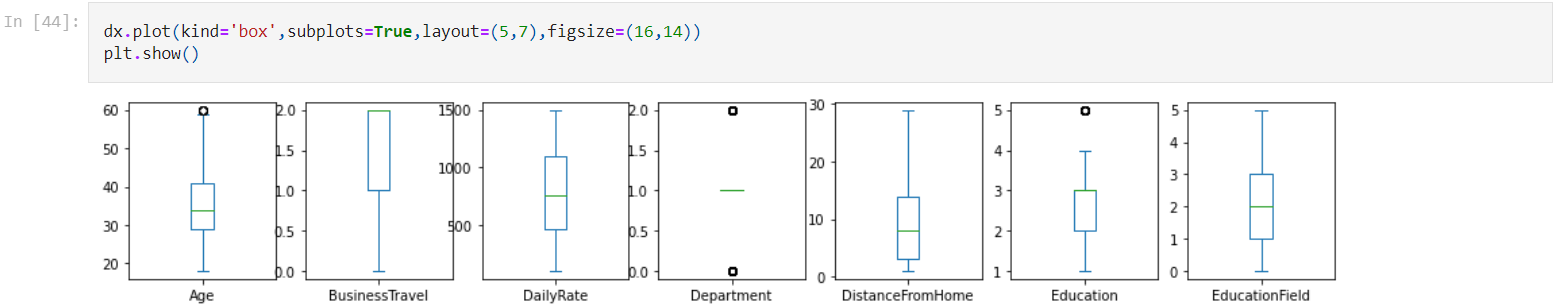


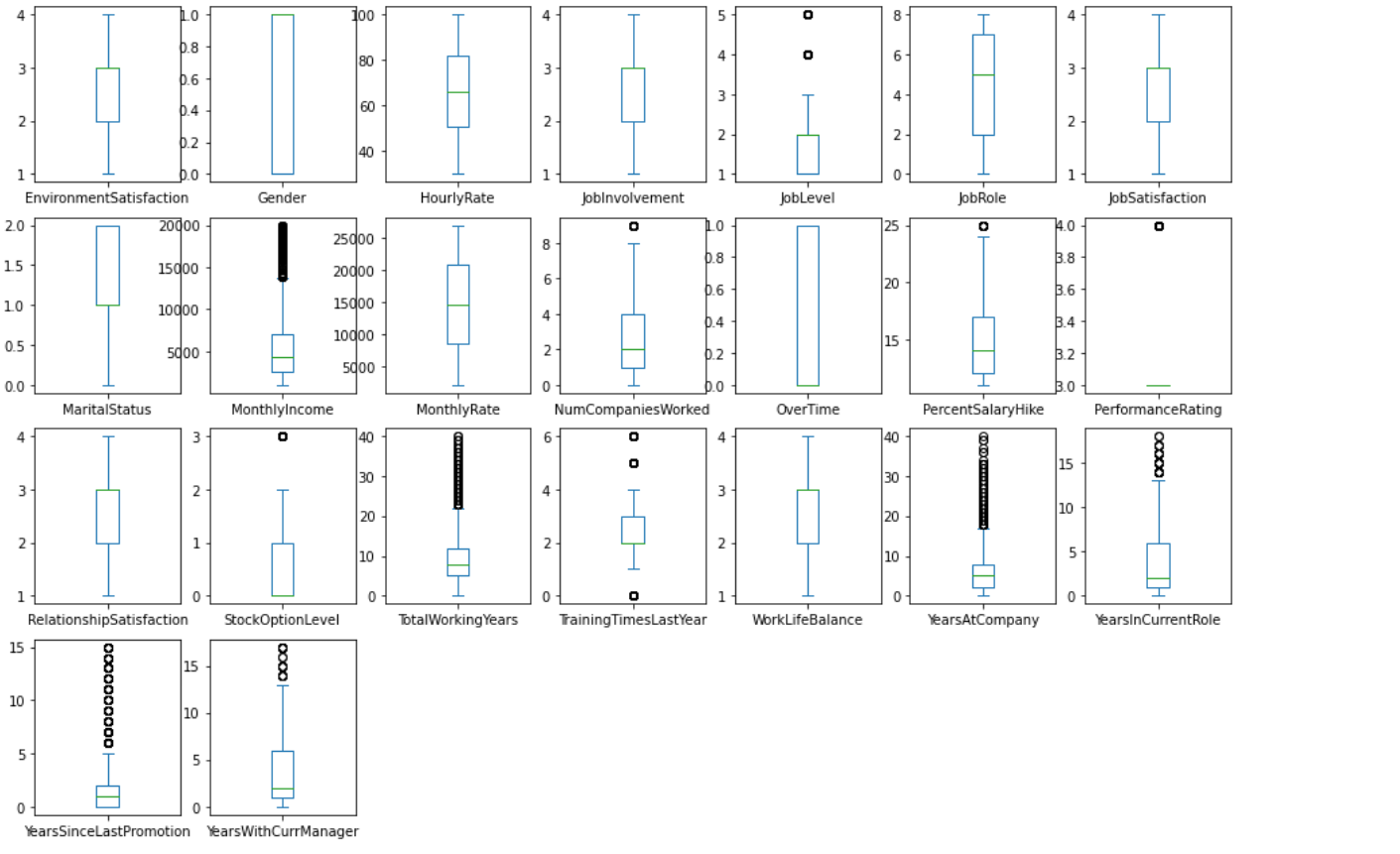


After using SMOTE the Dataset has increased to 2466 rows. As SMOTE creates dummy rows from the information present in the Dataset to balance the Data.

* + **Treating Outliers**

We will use boxplots to determine whether the Outliers are present in the Dataset.





While checking outliers we will only consider columns with continuous values.

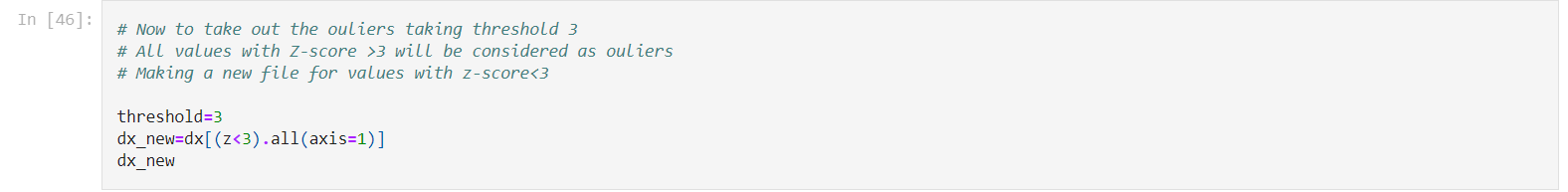
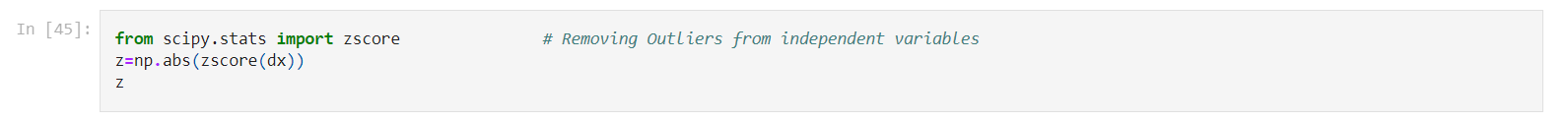
From the boxplots we can see that outliers are present in below mentioned columns:-

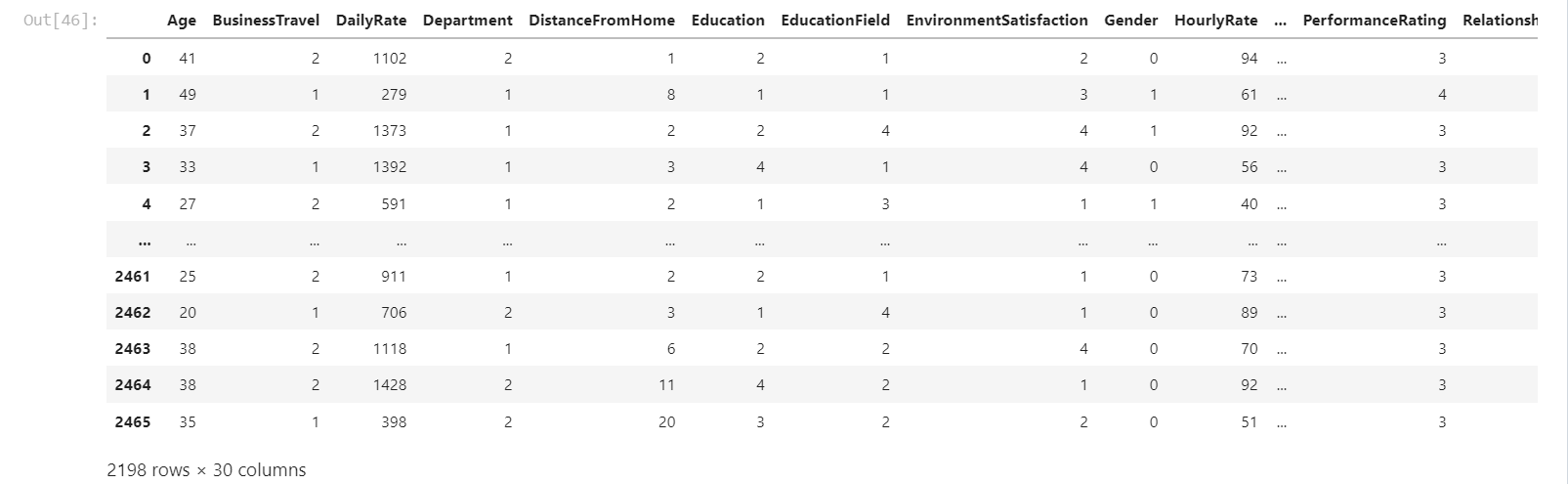
1. MonthlyIncome
2. TotalWorkingYears
3. YearsAtCompany
4. YearsInCurrentRole
5. YearsSinceLastPromotion
6. YearsWithCurrManager

We will use Z-score values to determine & remove the outliers.

For all the Z-Score values above 3 the data point will be considered as an outlier. We will make a new Dataframe removing all these values whose Z-score is above 3.

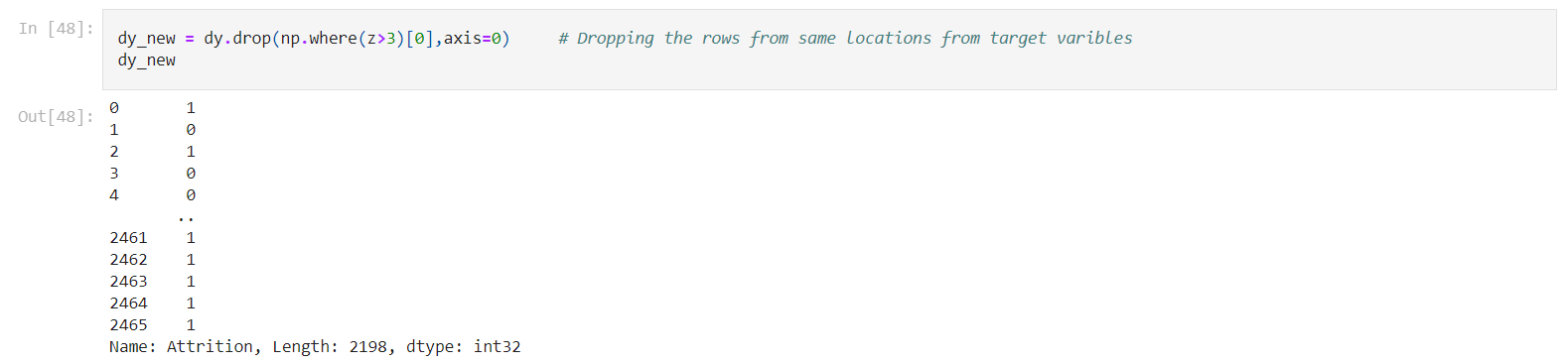
Also, outliers will not be removed from the Target column. Only the corresponding rows will be removed where the outliers are present in the Independent columns.





The rows now have been reduced from 2466 to 2198.

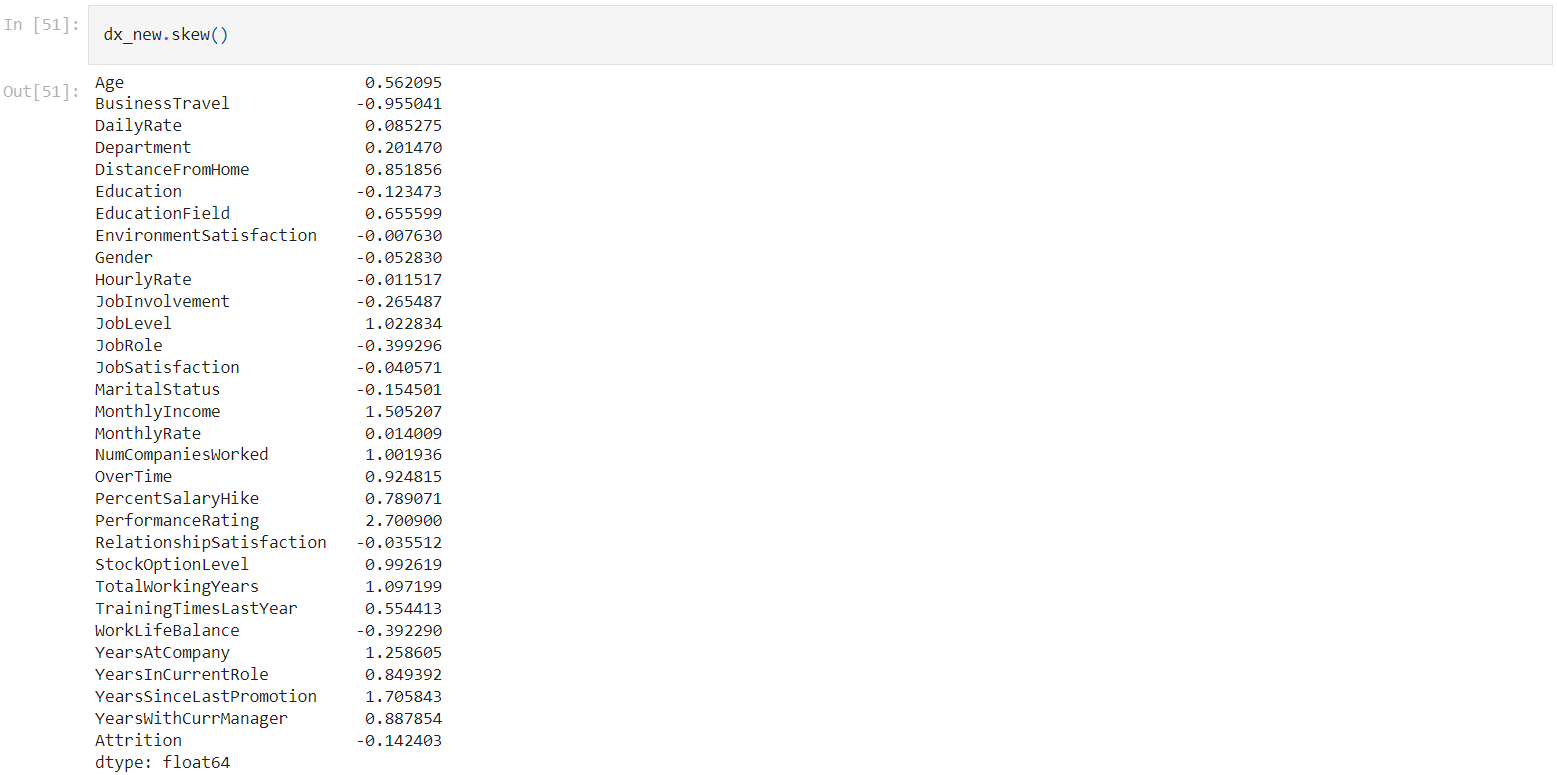
Now we will remove same rows from the Target Column.



Now the Dataset is free from Outliers.

* + **Treating Skewness**

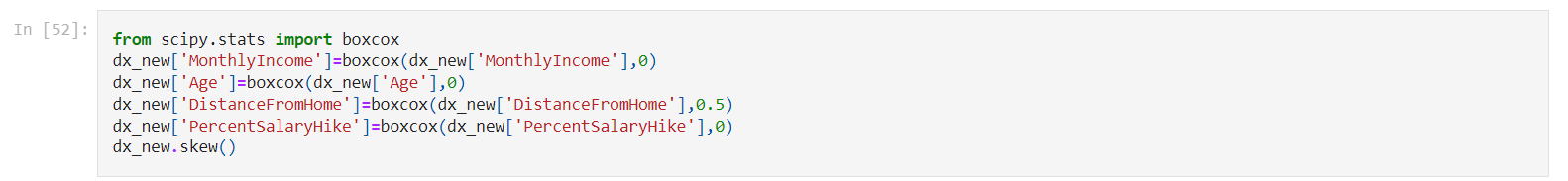
While checking skewness we will only consider Continuous columns.



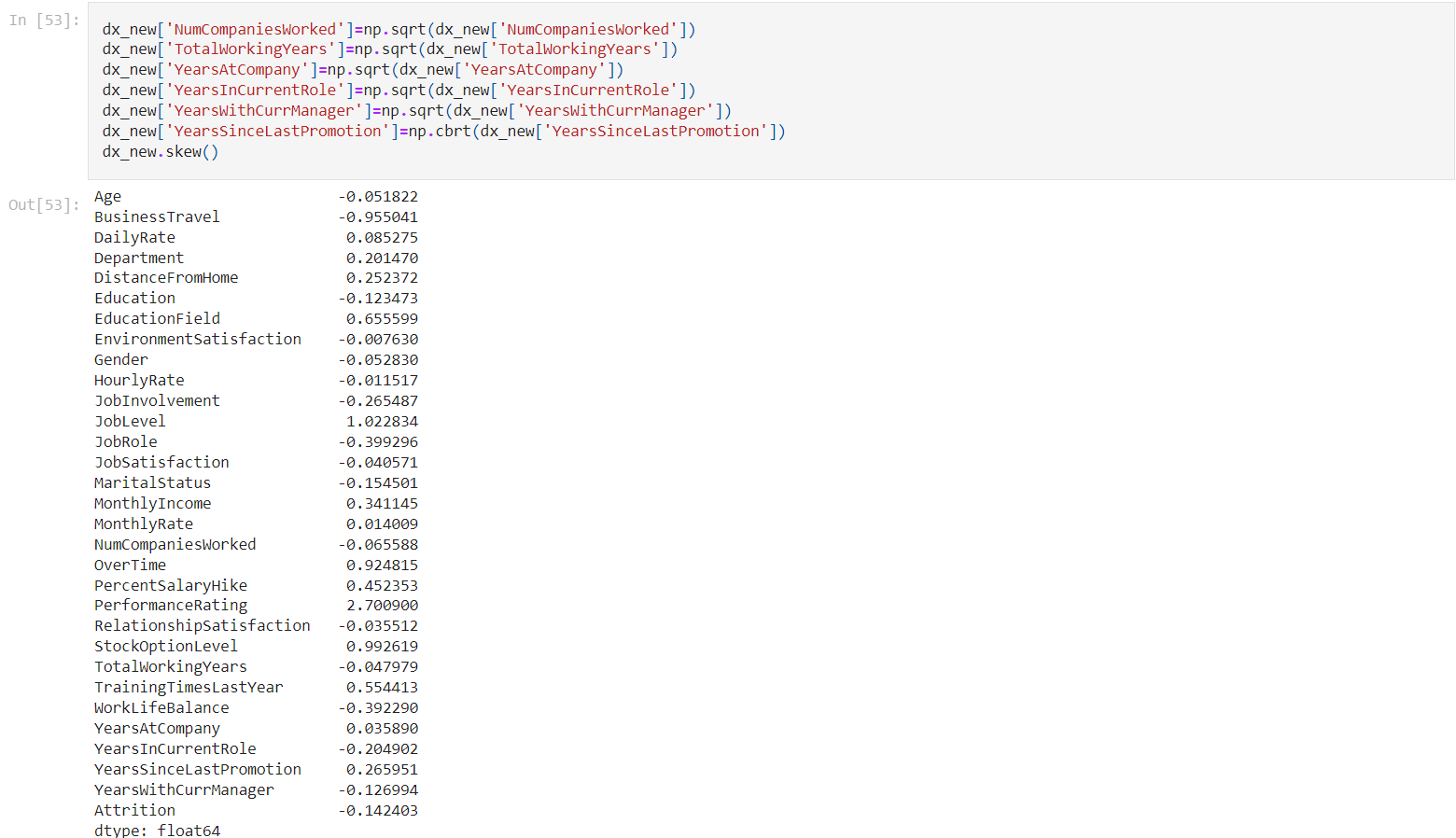
Taking Skewness threshold as +/- 0.5. We can see that the following columns are having skewnees that need to be treated.

1. Age
2. DistanceFromHome
3. MonthlyIncome
4. NumCompaniesWorked
5. PercentSalaryHike
6. TotalWorkingYears
7. YearsAtCompany
8. YearsInCurrentRole
9. YearsSinceLastPromotion
10. YearsWithCurrManager

Using **Boxcox Transformation** for Columns not containing 0 or -ve Values.



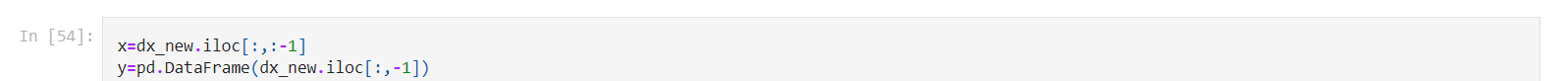
Using **Square Root & Cuberoot transformation** for rest columns.

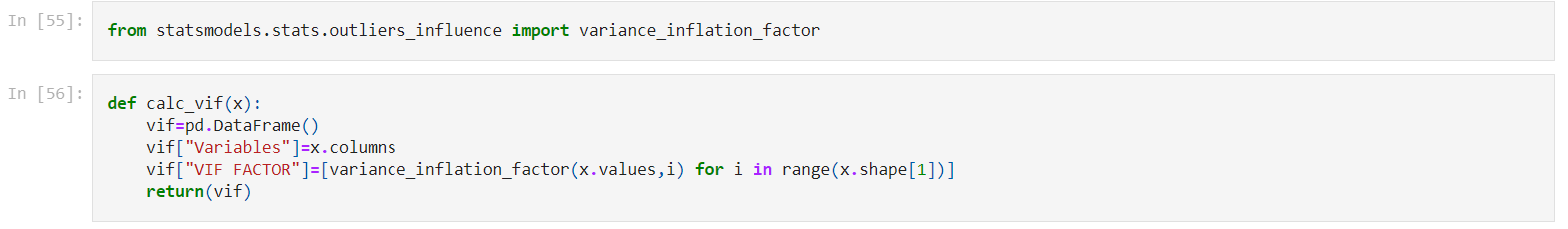


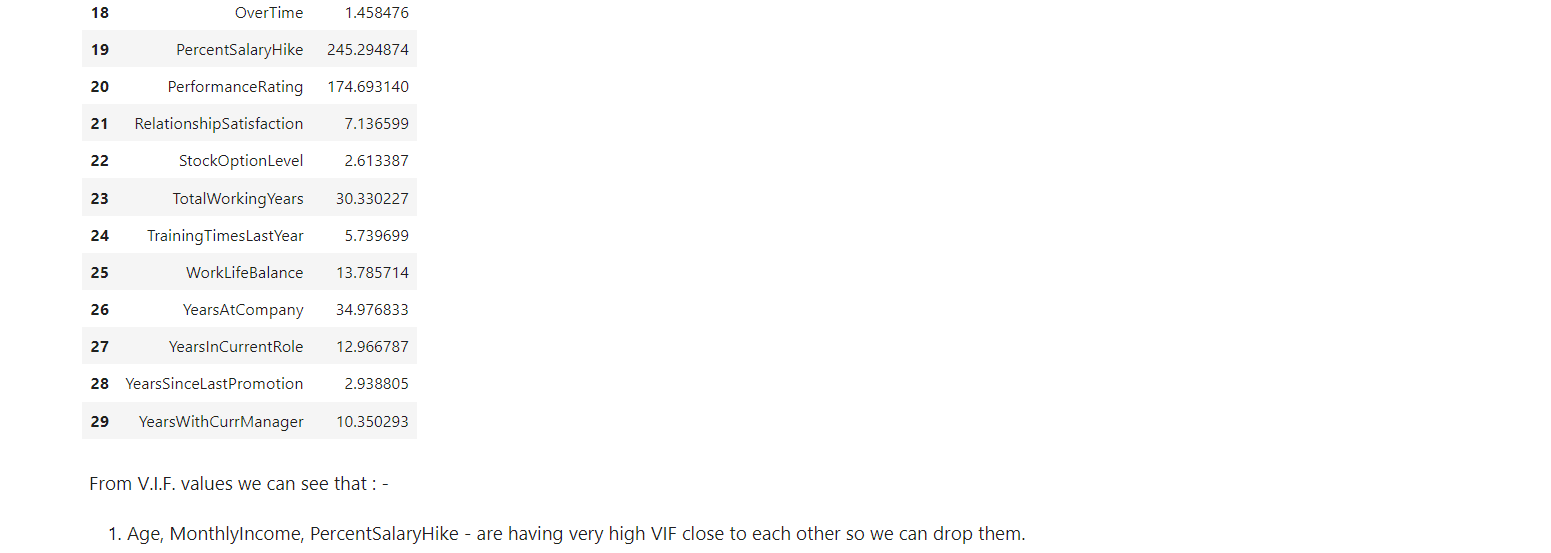
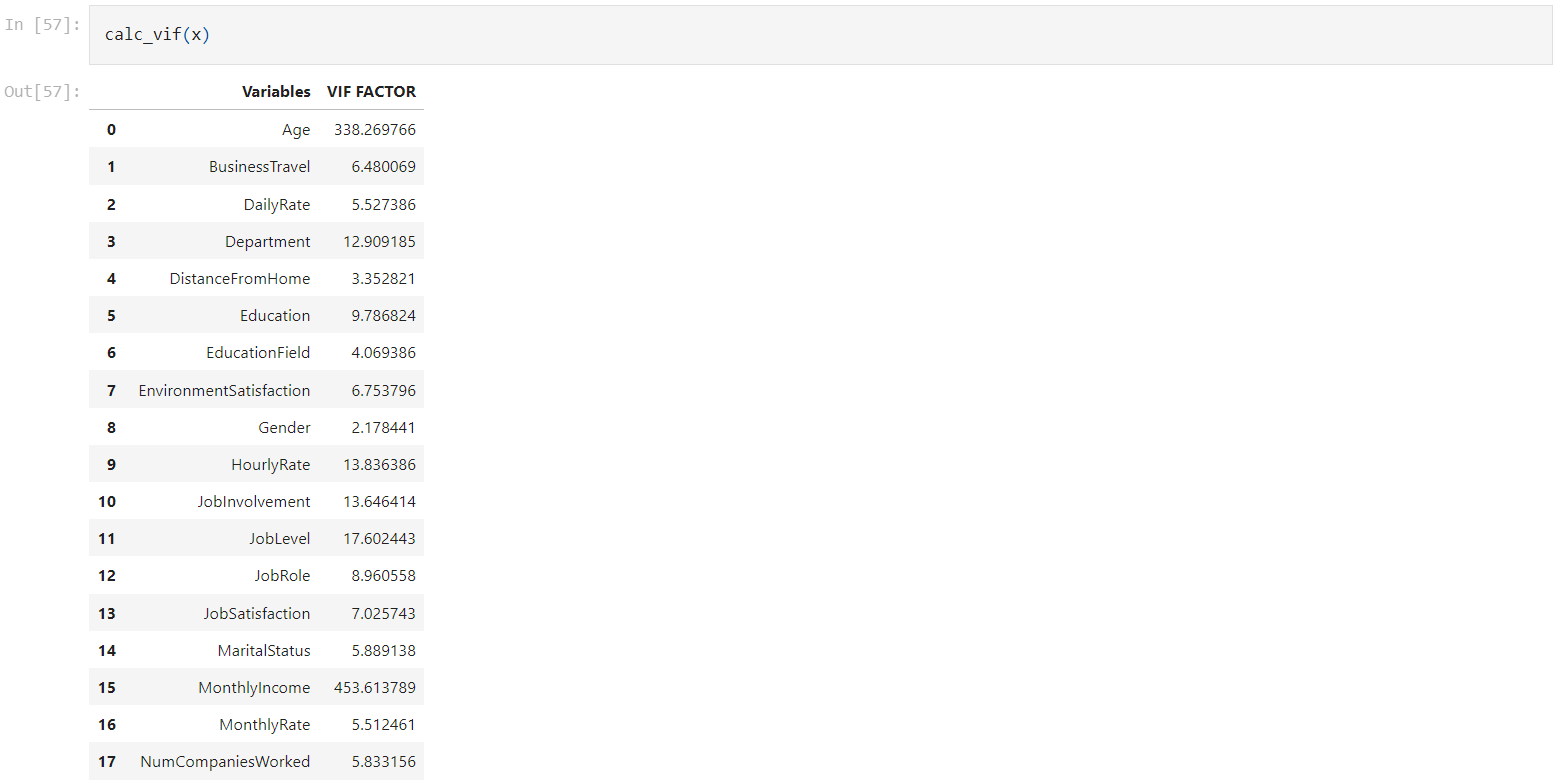
Now we can see that Skewness Values for all the columns with continuos values have been reduced to b/w +/-0.5.

* + **Treating Multicollinearity**

We will check V.I.F values for the Independent columns. And will remove those columns one by one that are having highest V.I.F. Values. Until the V.I.F. Values are in acceptable range.







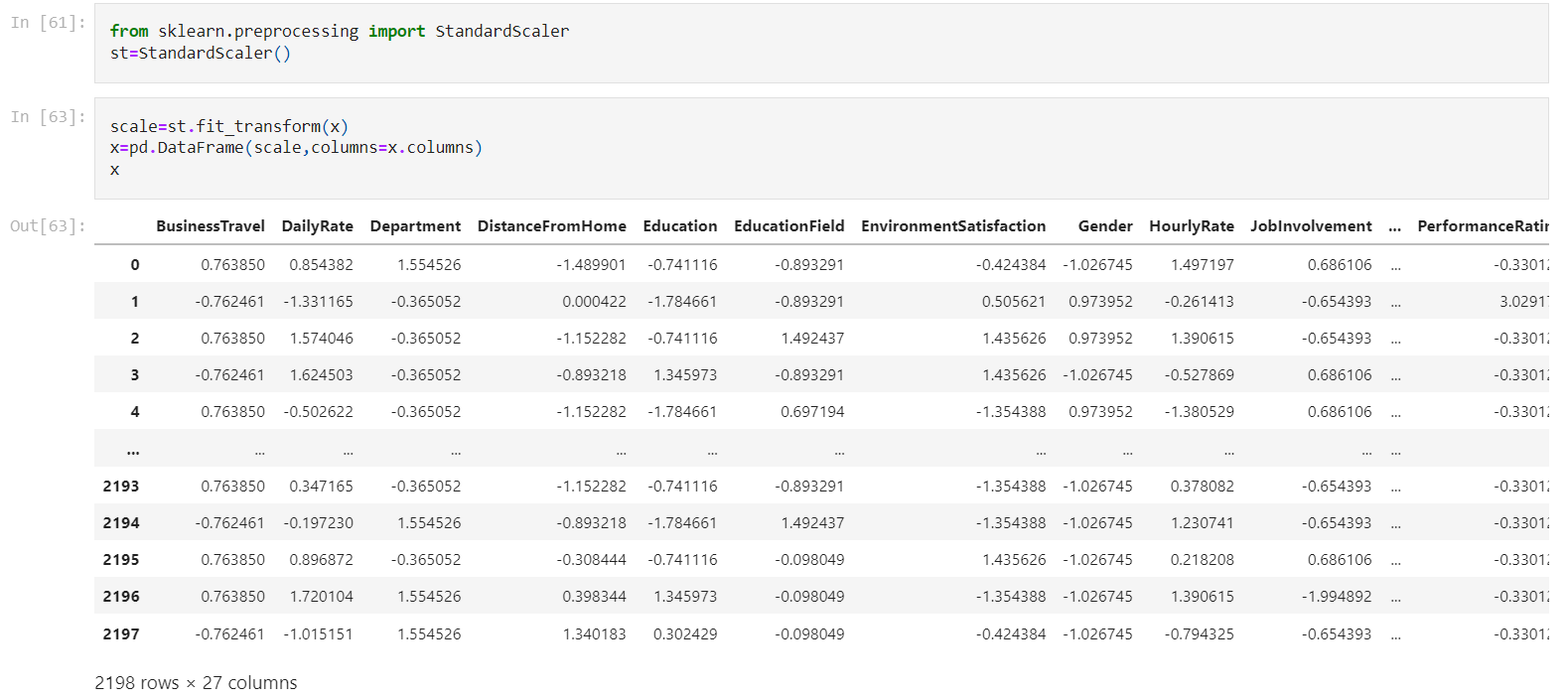
From checking the V.I.F. values we will drop ‘Age’ , ‘MonthlyIncome’ , ‘PercentSalaryHike’ columns as they are having very high VIF close to each other.

After removing these columns all the V.I.F. Values have reduced significantly & most of the values are below 10 which is acceptable.

* + **Data Scaling**

We will use encoding techniques to transform the Data into same scale as the independent columns values are in different scales.

From this data I am using StandardScaler. We can also use use Min-Max Scaler.



After using StandardScaler we can see that all the column values are now in same scale.

**NOTE : -** Now our Data is ready for Machine Learning algorithms. We have cleaned the data form noise values. The steps performed on the Data are as follows:-

1. Label Encoding – To convert the Columns with String values into numeric values.
2. Data Balancing – To balance the Data using Smote w.r.t. both Attrition type.
3. Outliers Removal – To remove the Outliers with Z-score.
4. Treating Skewness – Removing skewness with boxcox, square root & cube root transformation methods.
5. Treating Multicollinearity – Removing Multicollinearity through V.I.F.
6. Data Scaling - To scale the data with Standard Scaler.

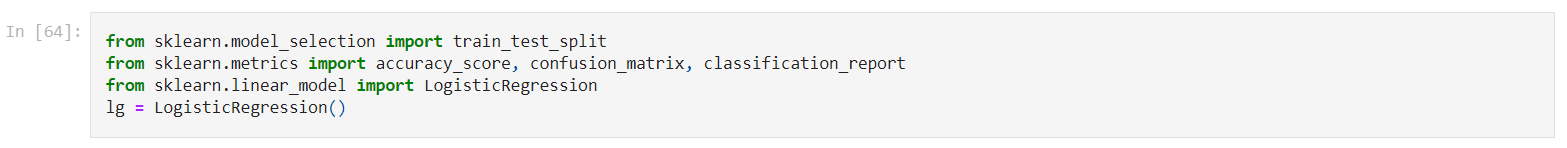
Now we will proceed with building various Machine Learning algorithms for this prepared Data.

1. **Splitting the Data**

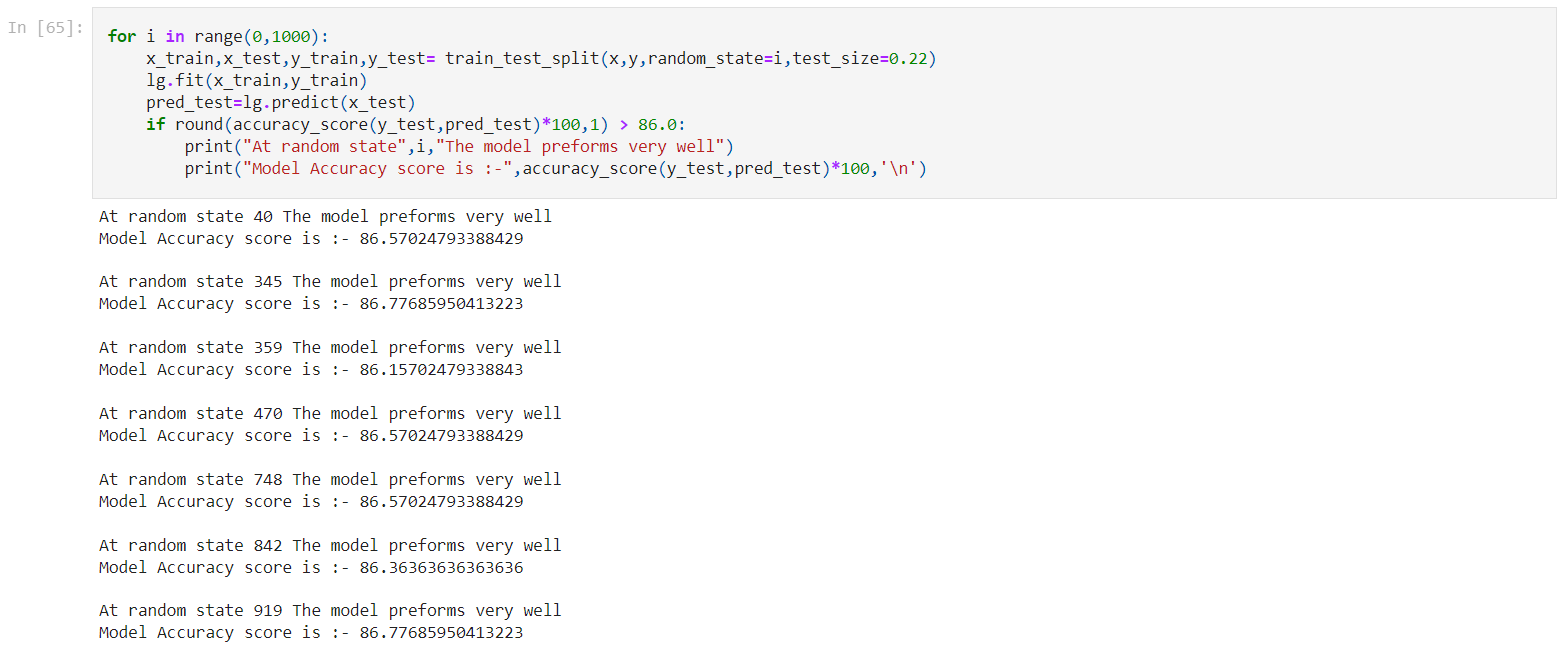
Now we will find the best Random State value to split the Data into Test & Train.

First we will import the necessary libraries.

Let’s import LogisticRegression to find out the best random state value.



Now we will use for loop to find out the best Random State value b/w 1-1000.



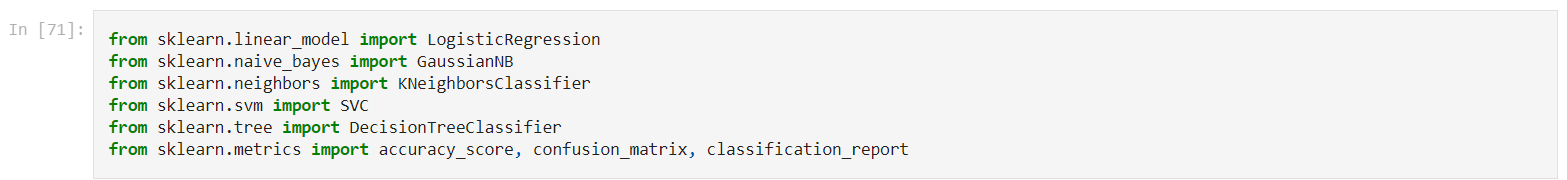
We can see that Model is working best at Random State 345 as the model accuracy score is highest at 345.

So, using random state = 345 for further working & splitting the Data.



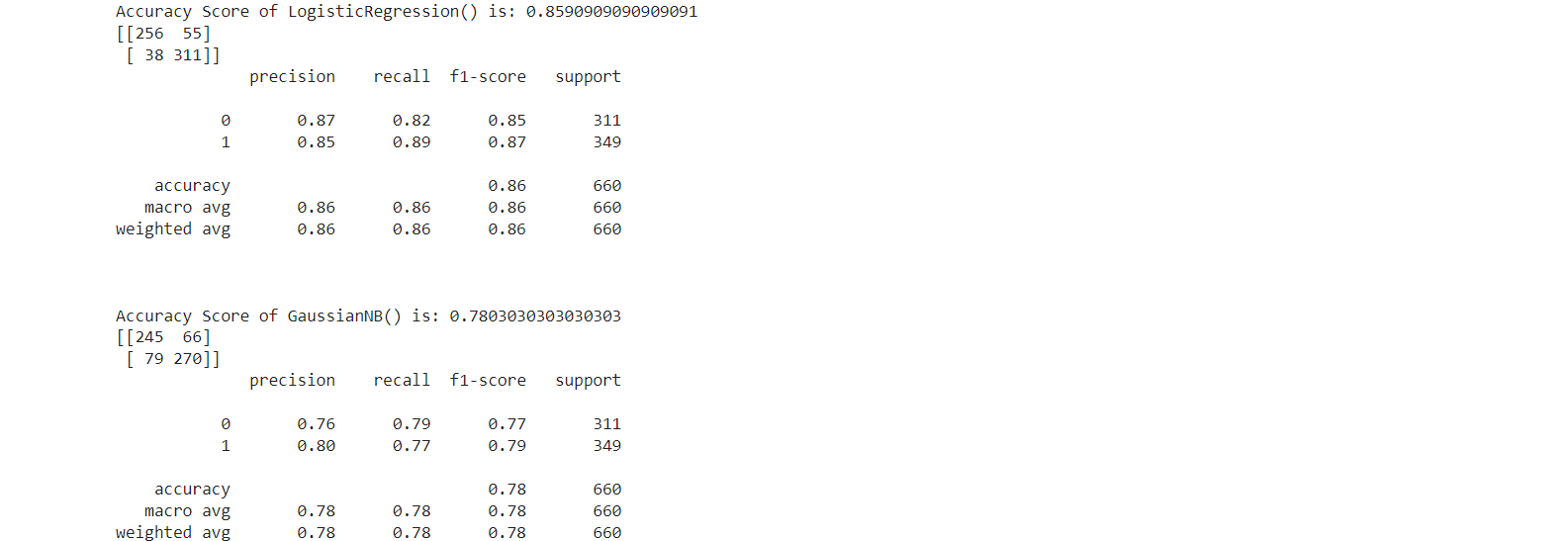
1. **Testing with Different Models**

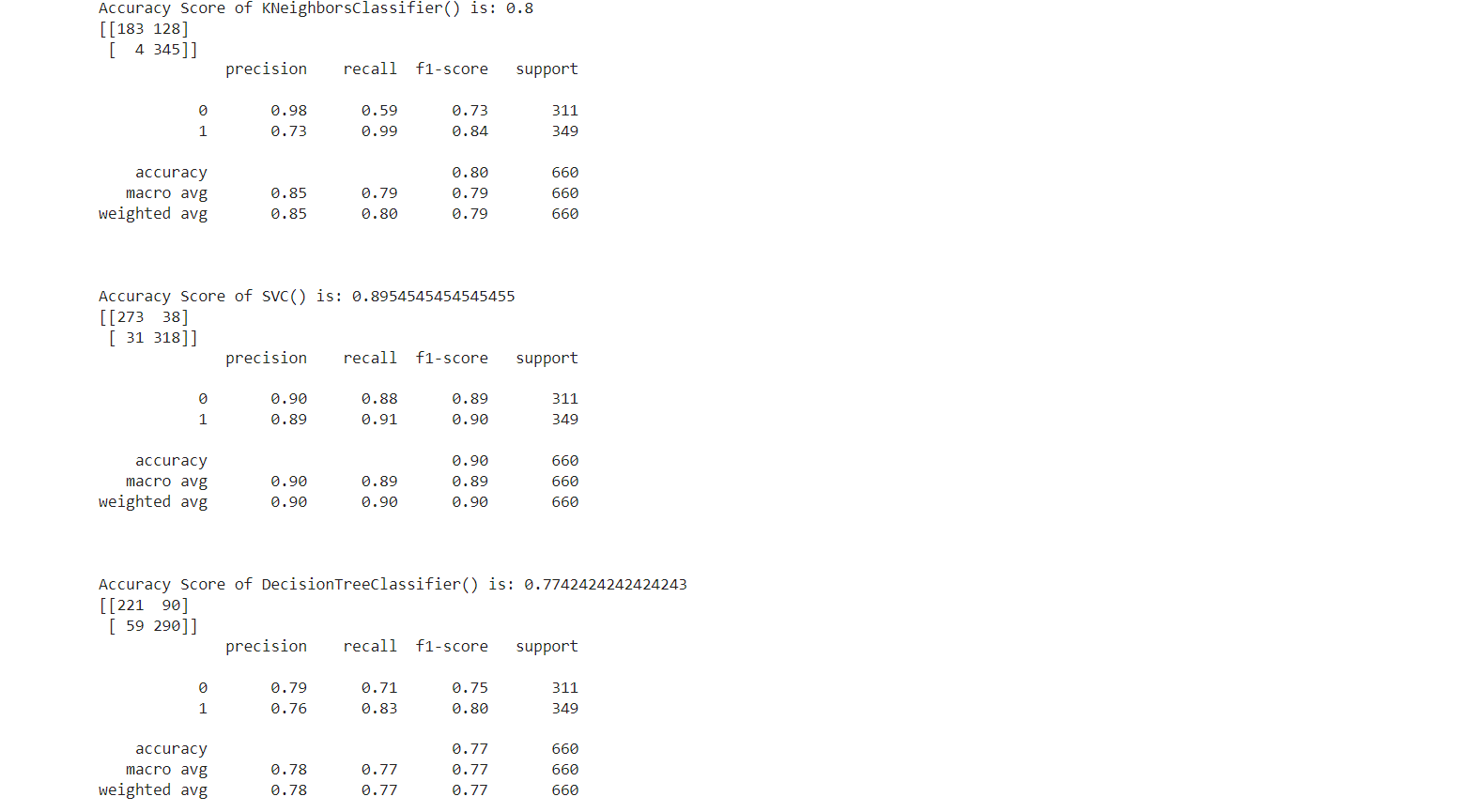
Now let’s import various models to test which algorithm is performing best will the Data.



Let’s check the accuracy scores for all algorithms.





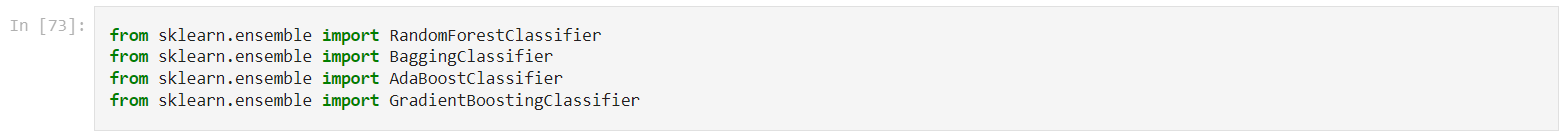


From above we can see that LogisticRegression() & Support Vector Classfier() are working best for the Data as they are obtaining highest Accuracy. Whereas, SVC is getting highest accuracy among all other algorithms.

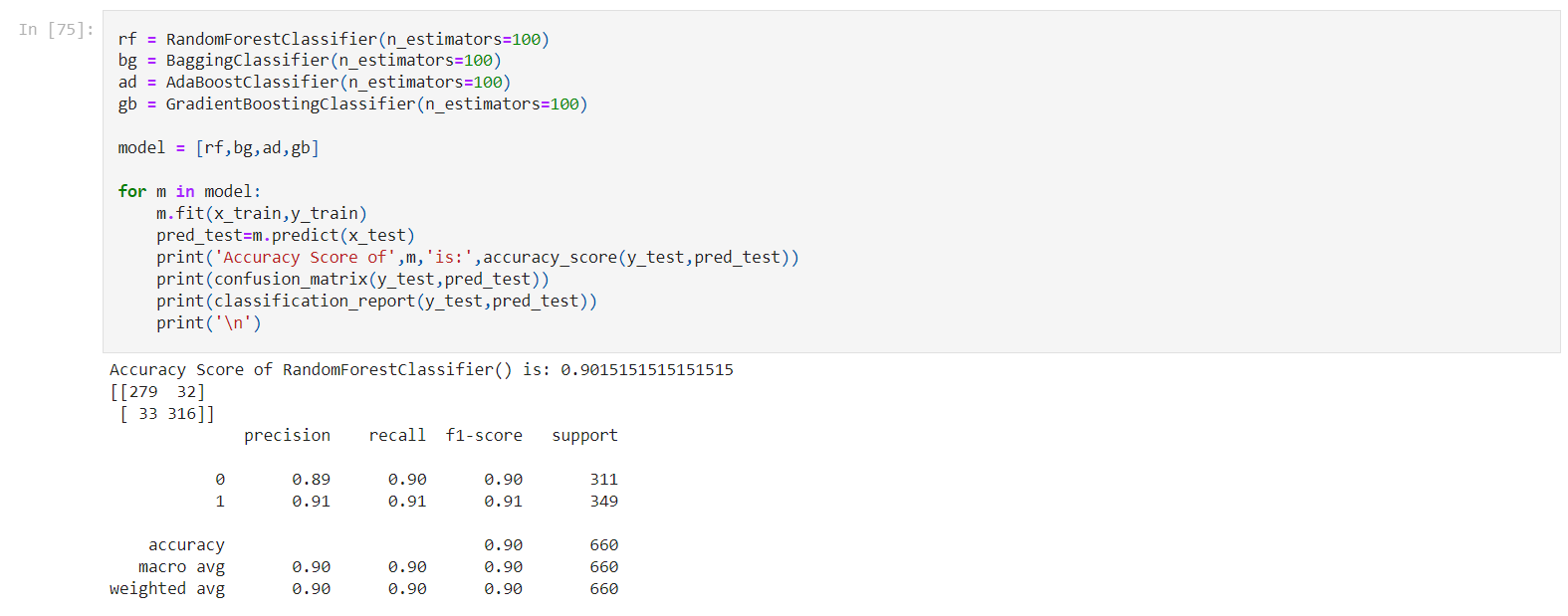
1. **Ensemble Techniques**

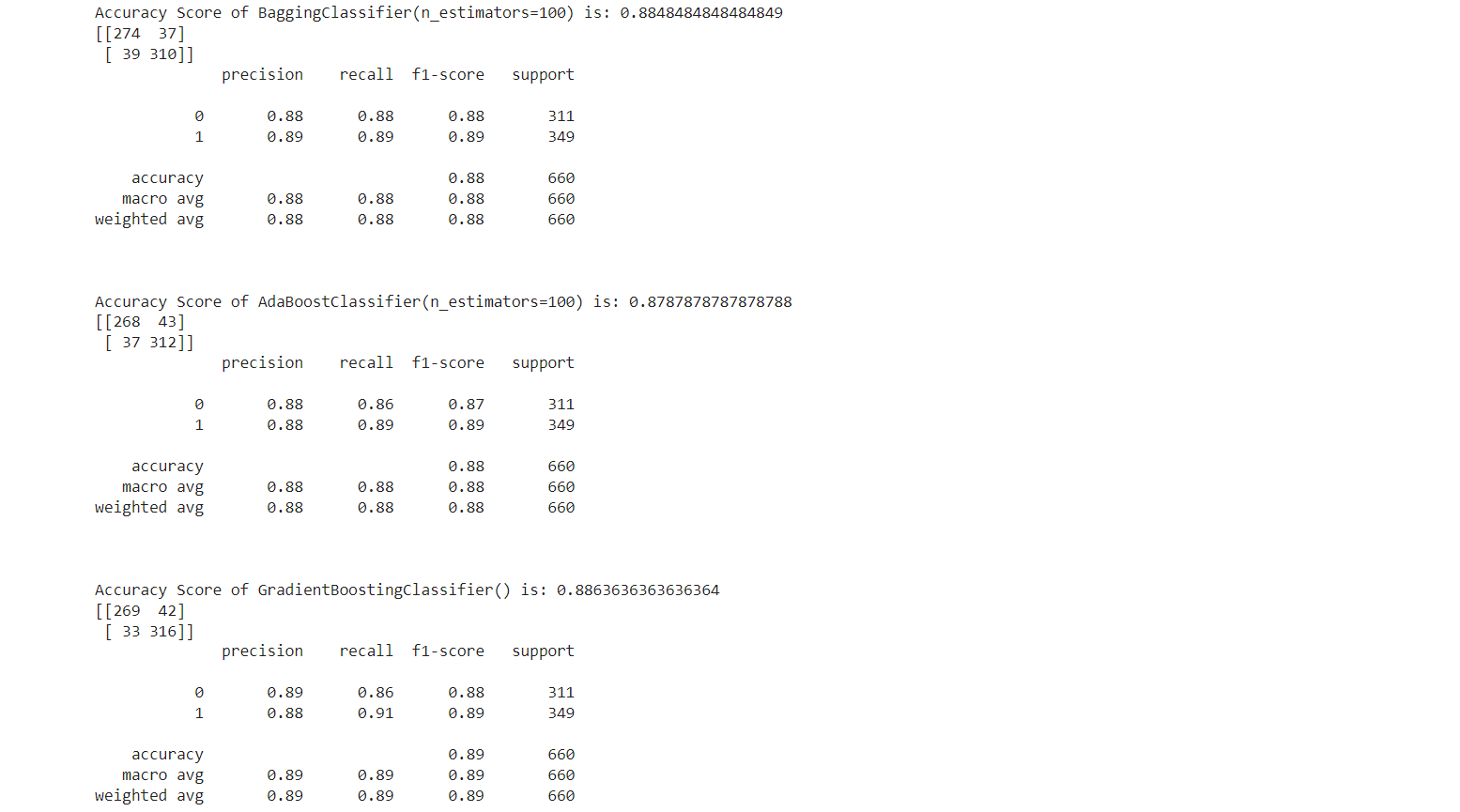
Let’s use various Bagging & Boosting Ensemble methods for training the data to increase the model performance.

We will import all the Ensemble algorithms.



Now checking the accuracy scores for the above methods.

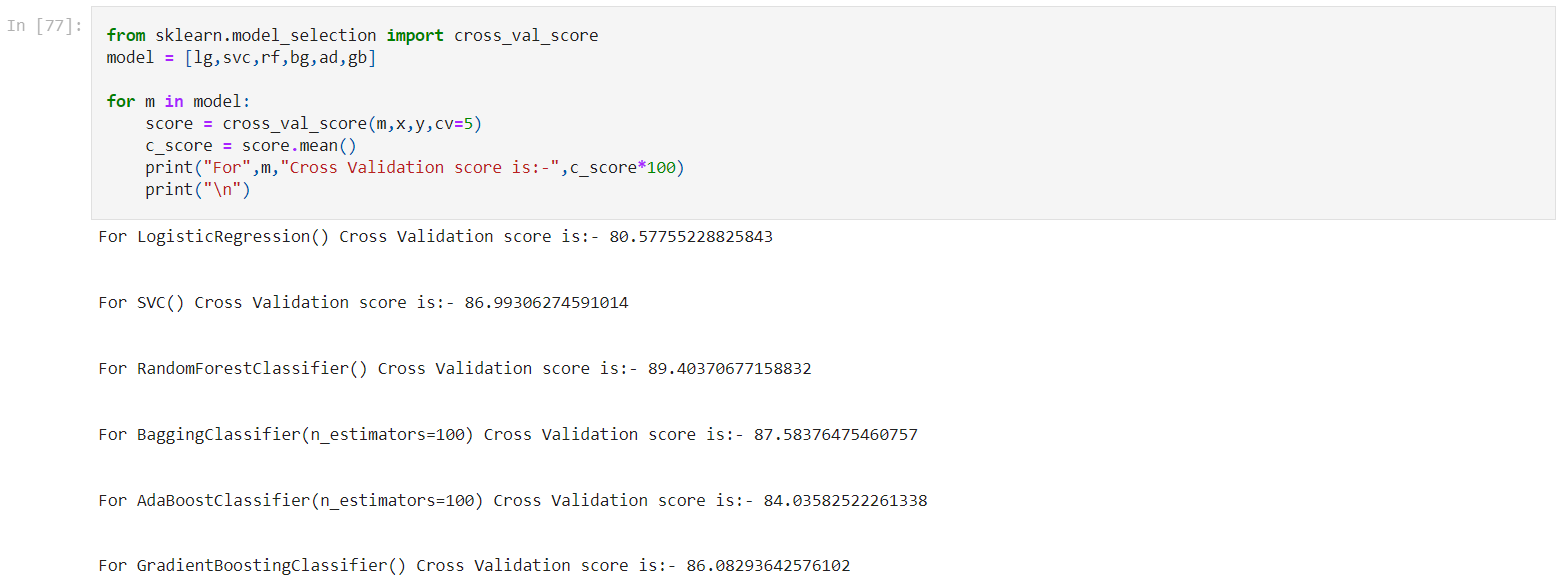




From above ensemble methods we can see that RandomForestClassifier() is working best for the Model as it is obtaining highest accuracy of 90%. Whereas other algorithms are also working well with the data as they are getting around 88% accuracy.

1. **Cross-Validation of Algorithms**

Let’s check the Cross-Val Scores of all the algorithms that are obtaining accuracy score more than 80%. To see which model is actually performing best with data.



Now from above we can confirm that RandomForestClassifier() is working Best for the data as its getting the highest Cross Val Score.

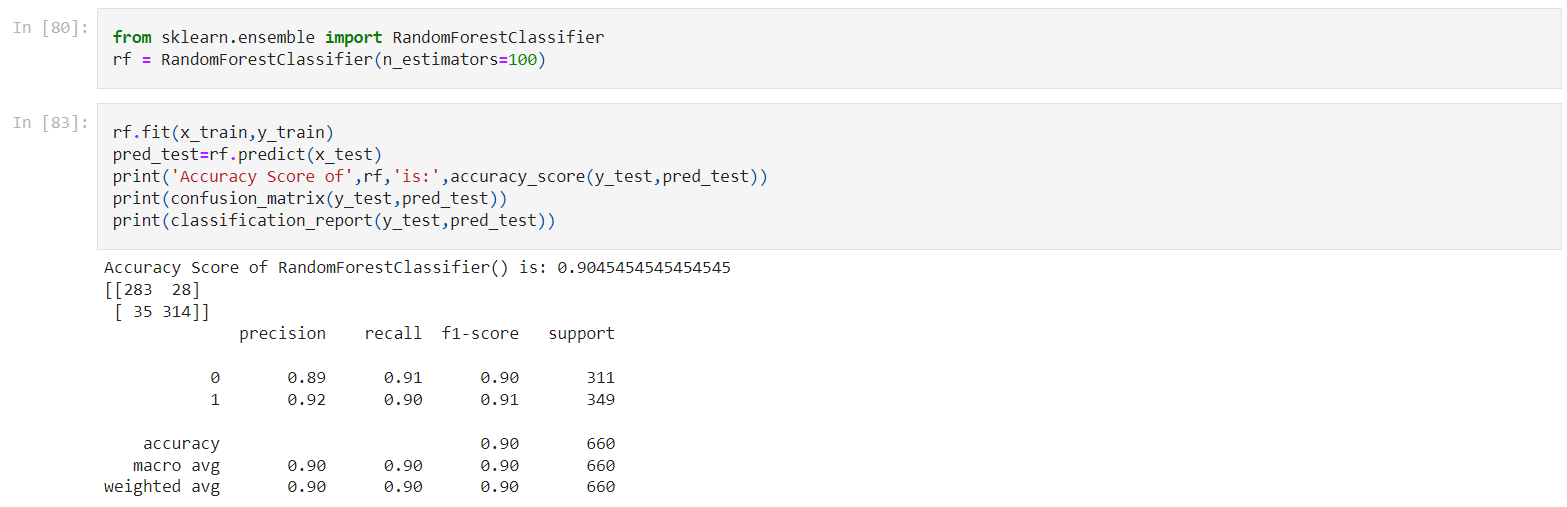
Therefore, using RandomForestClassifier as the Final Algorithm.

1. **Hyperparameter tuning**

Using GridSearchCV for further optimization of RandomForestClassifier(), to check which parameters are working best for RandomForestClassifier.



As we can see from above that the Model Performance(accuracy) has reduced from the default values of RandomForestClassifier(). Hence proceeding with the default parameter values of RandomForestClassifier() only.



1. **Cross-Validation of Best Model**

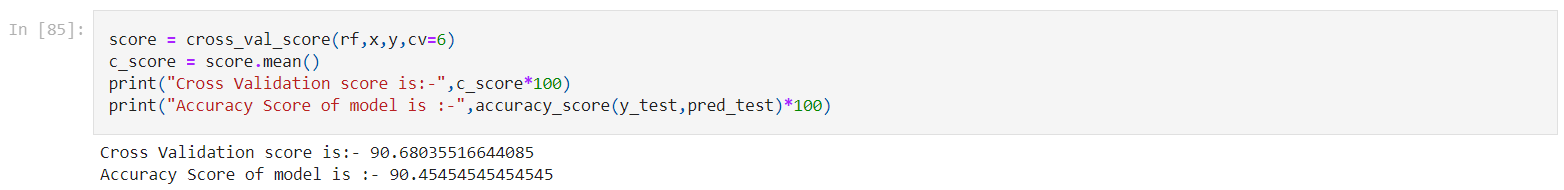
Let’s find out the cv value for which the Cross Val score is closest to the Accuracy Score of the Model.





Since Cross Validation Score is almost equal to Accuracy Score. Hence we are heading towards a good approach and there is no overfitting/underfitting.

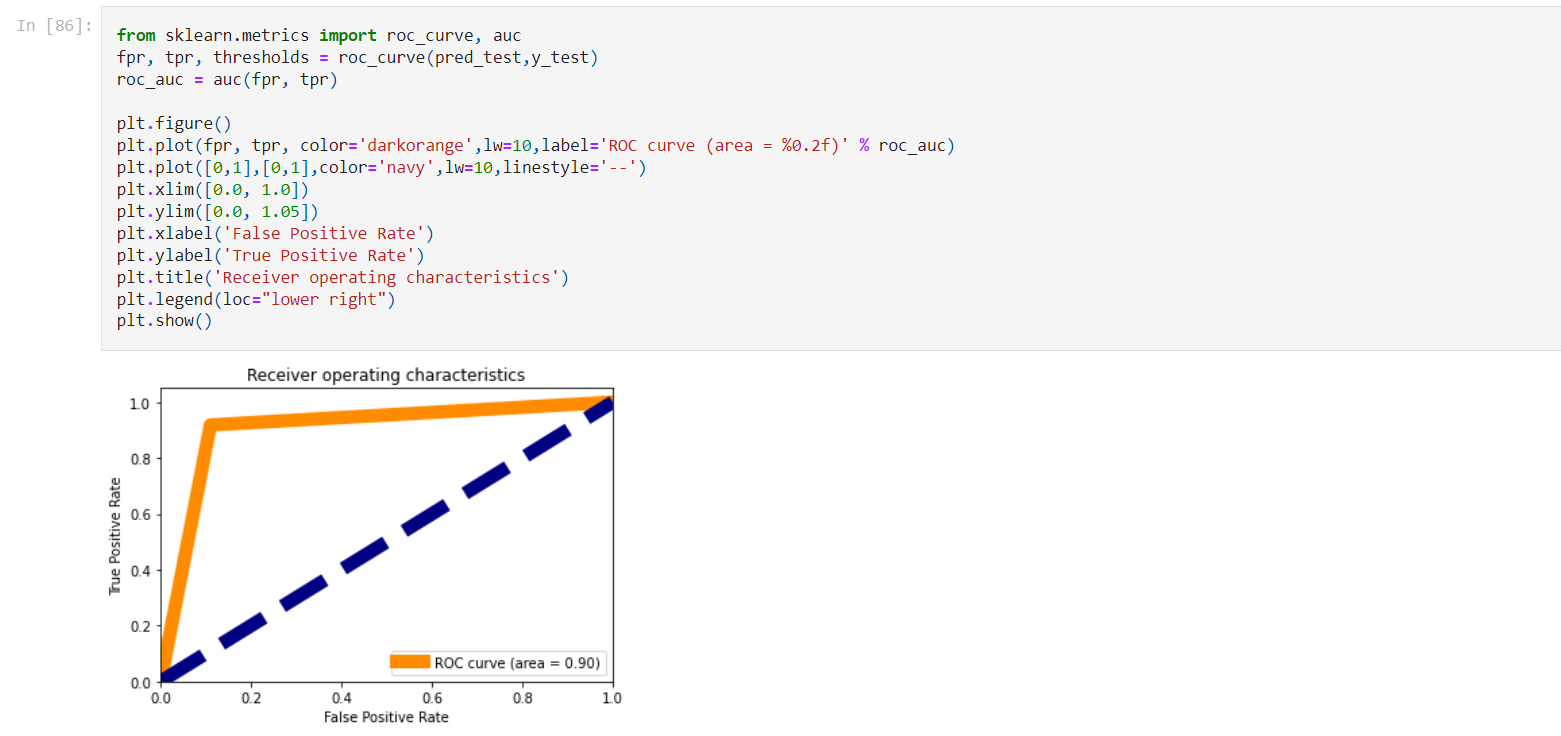
Accuracy score is closet to Cross Val Score at cv = 6. Therefore using cv = 6 for final model validation.



We can see that the Cross- Val Score & accuracy score are equal & both are obtaining more than 90% accuracy. Hence our model is performing good & there is no problem of overfitting or underfitting.

1. **AUC ROC Curve**

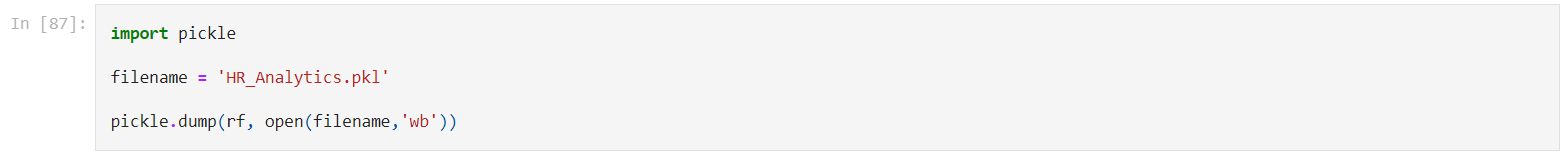
Let’s plot the ROC curve to validate the performance of the model.



We can see that 90% area is under the curve hence our model is performing good.

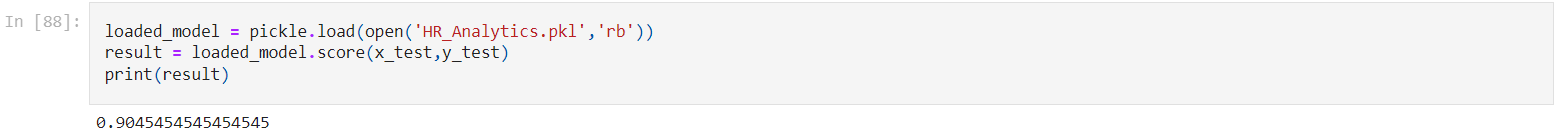
1. **Saving the best Model**

Let’s save the model using pickle library.



1. **Conclusion**

Now let’s check the saved model performance.



Now making a Data Frame for predicted & original values to see the results.



From above we can see that almost all the predicted & original values are matching. Hence our model is performing good.