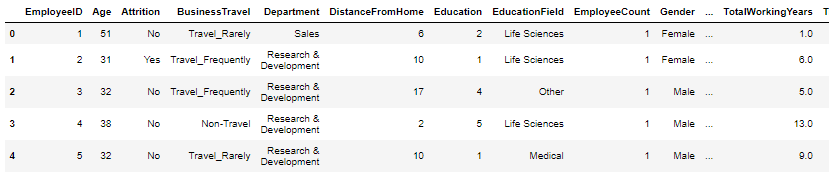
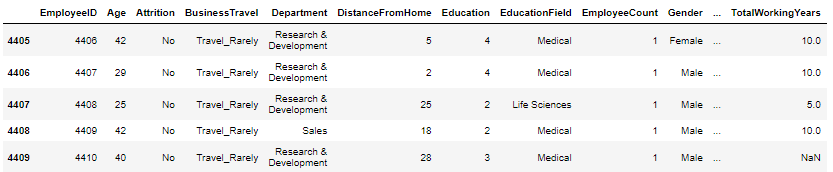
**DOCUMENTATION ON ATTRITION DATA SET**

The top five rows of the data set



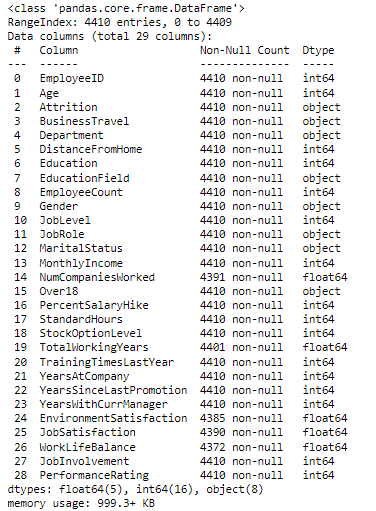
The bottom five rows of the data set



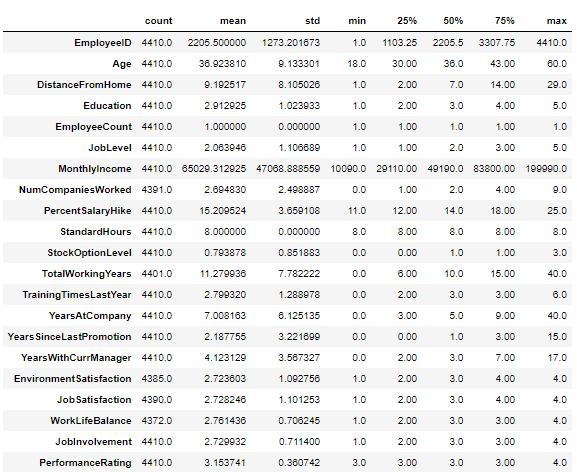
Shape of the data is as follows



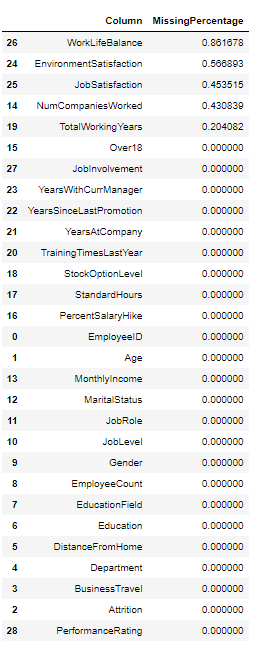
Info about the data set



Data summary of the data set

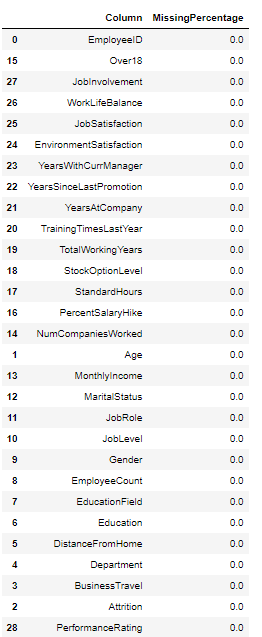


The following table shows the percentage of missing values in the different features present in the data



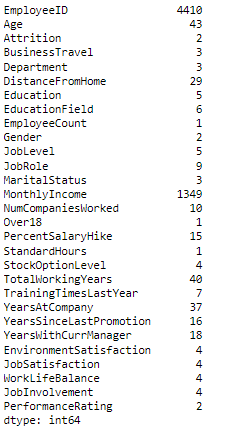
Now we impute the missing values with their median values

The missing percentages after imputing the missing the values with the median values



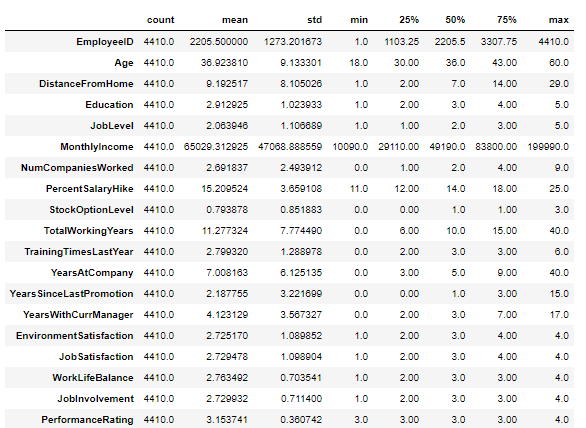
* There are no duplicate values present in the dataset

The unique categories present in each feature is as follows

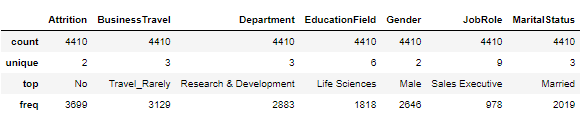


* From above, we can understand the number of value\_counts in each of the feature
* Emplyee Id shows that way because the ID is always unique to each of the row present in the data.
* Now we drop few features such as 'EmployeeCount', 'Over18', 'StandardHours' since they have got only 1 value counts and it doesn’t give any insight.

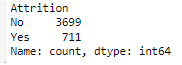
Data summary now

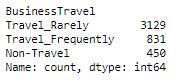


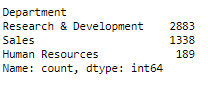
Data summary of only the ‘object’ datatype features

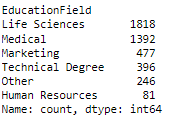


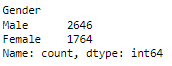
The value counts of different features is as follows :

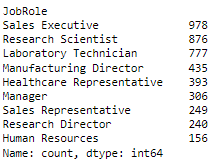




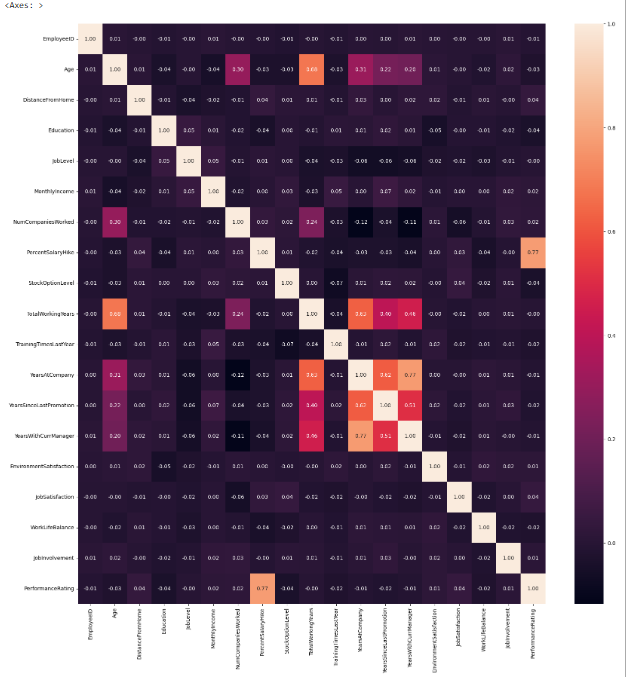








Correlation between the features

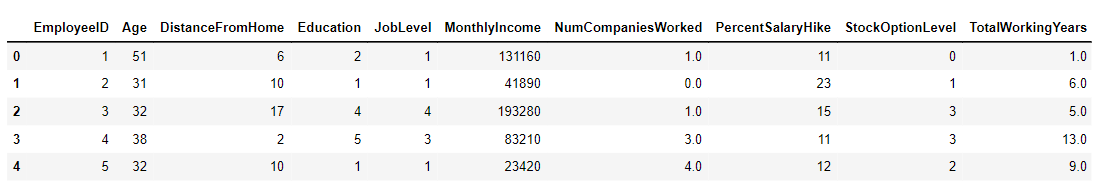


The above heat map is a bit messy because of the large number of features present in the data set.

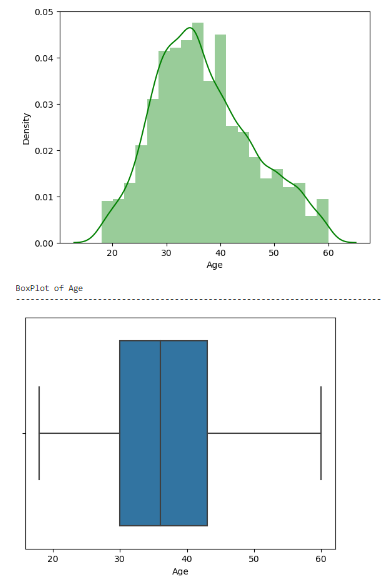
Seperating the dataframe having the numerical features:

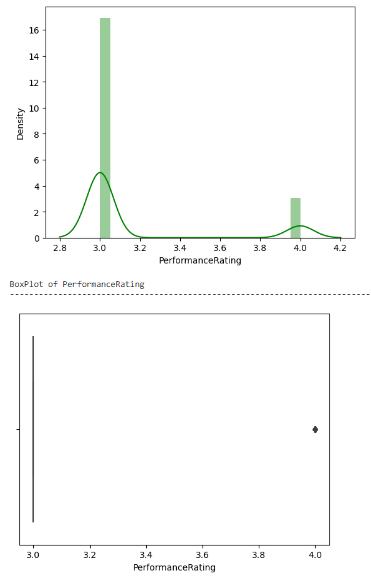
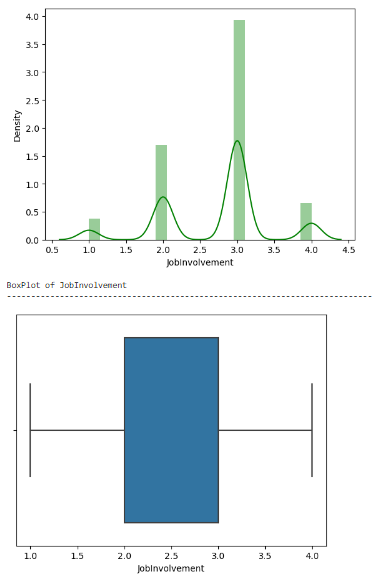
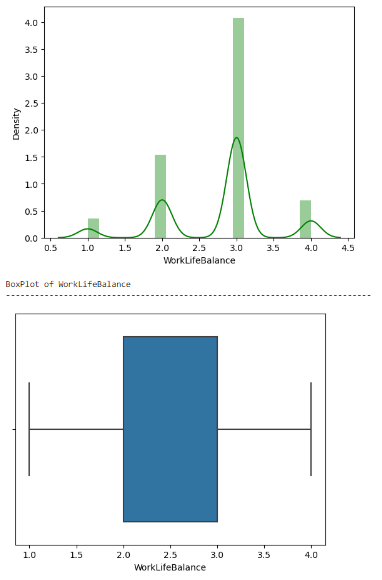
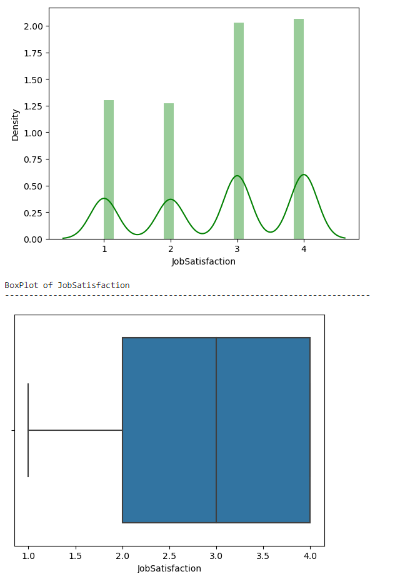
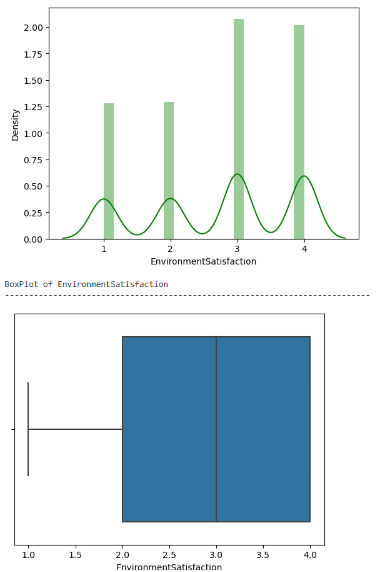
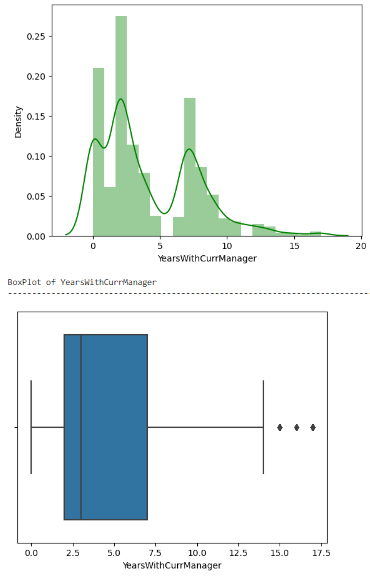
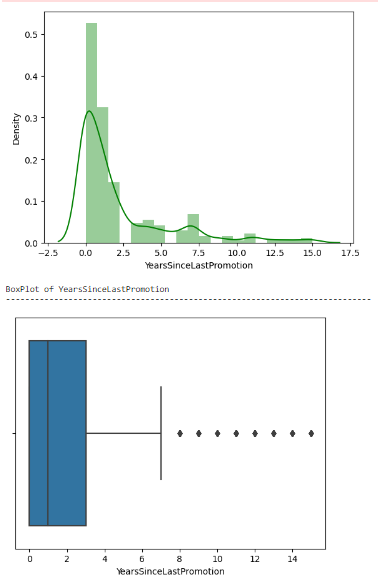
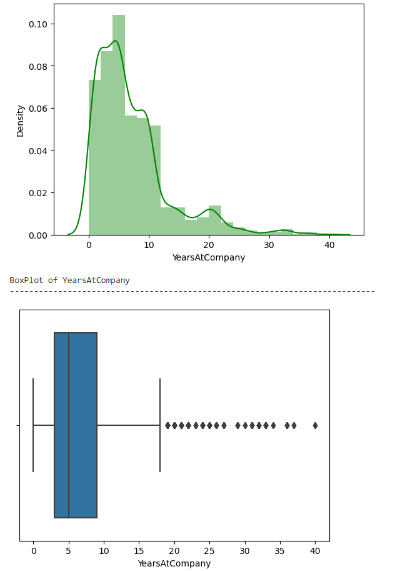
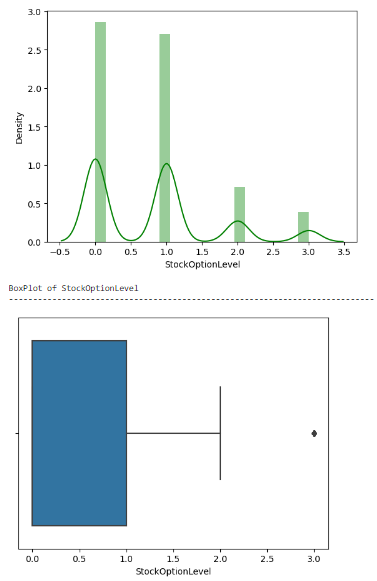
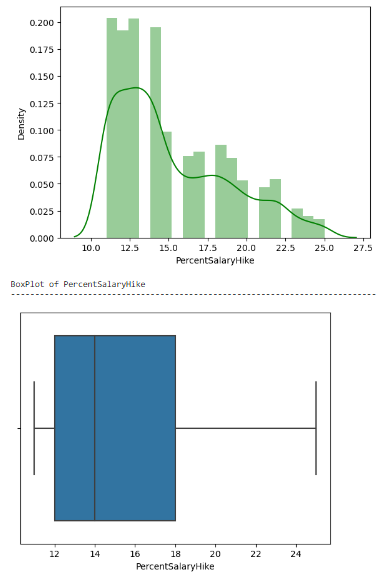
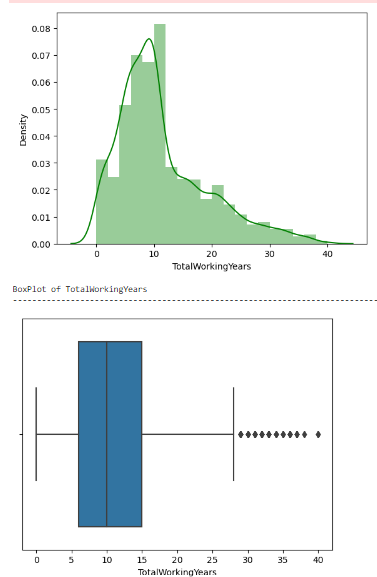
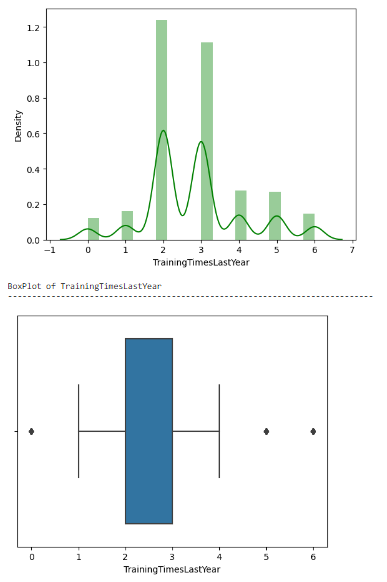
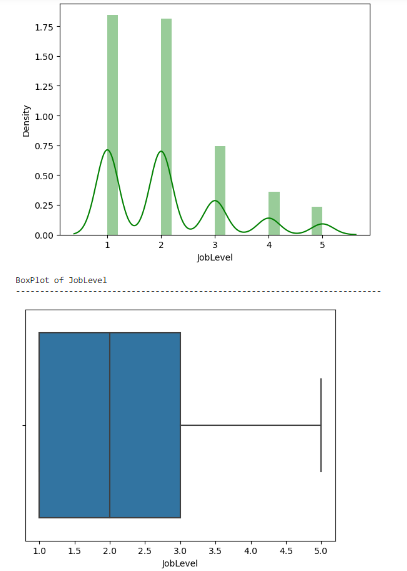
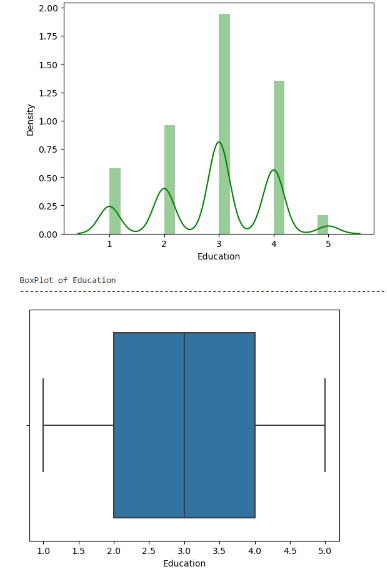
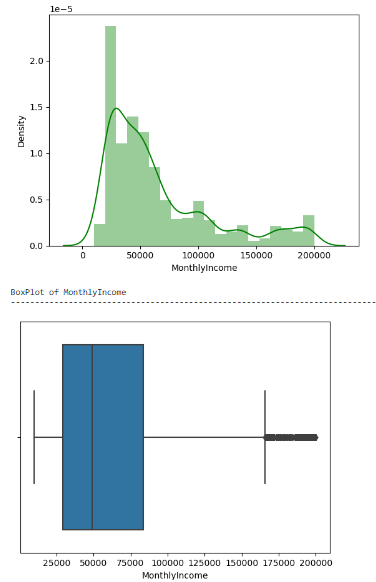
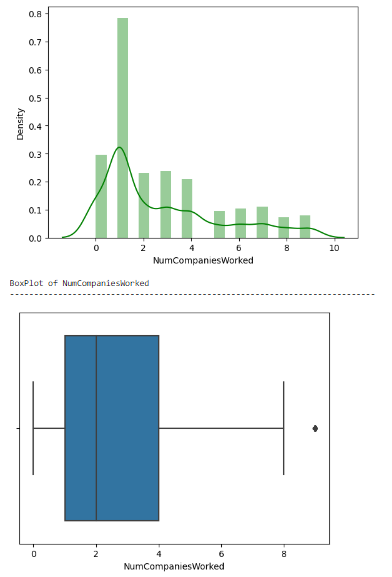
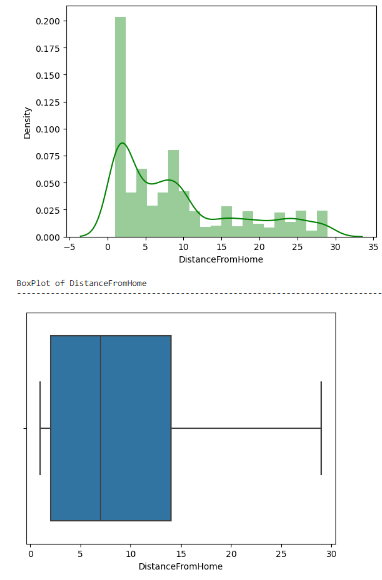
Now we separate only the features having the datatype ‘float’ and ‘integer’ into a new data frame

The first five rows of the data frame looks like

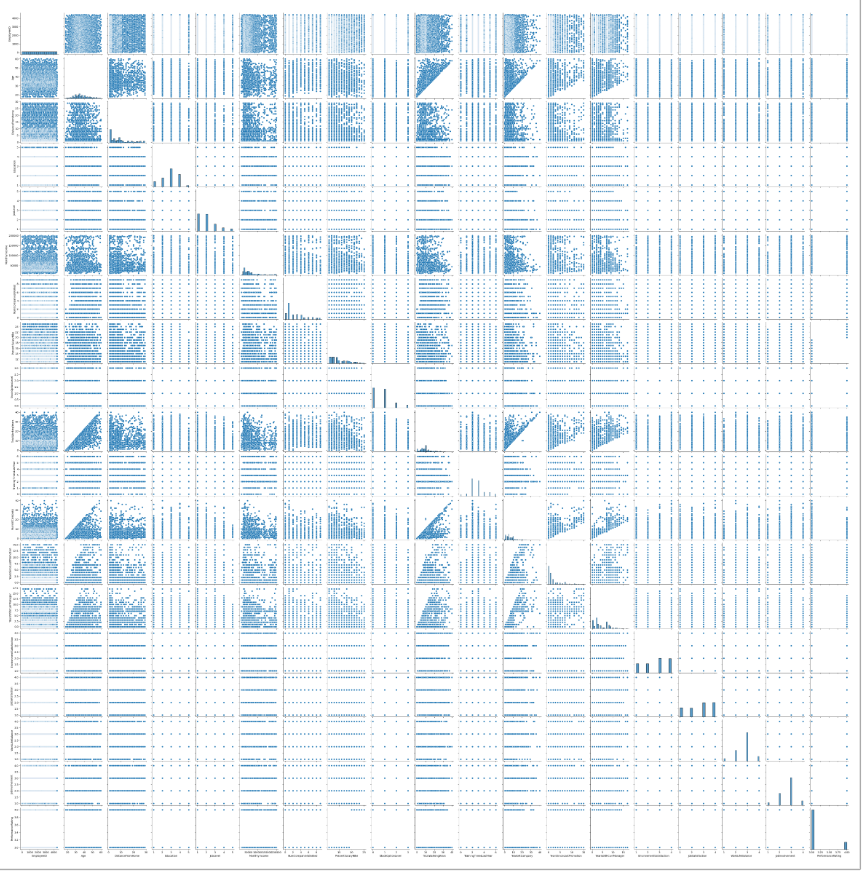


UNIVARIATE ANALYSIS :

The boxplots and histograms of various features is as follows:



Pairplot :

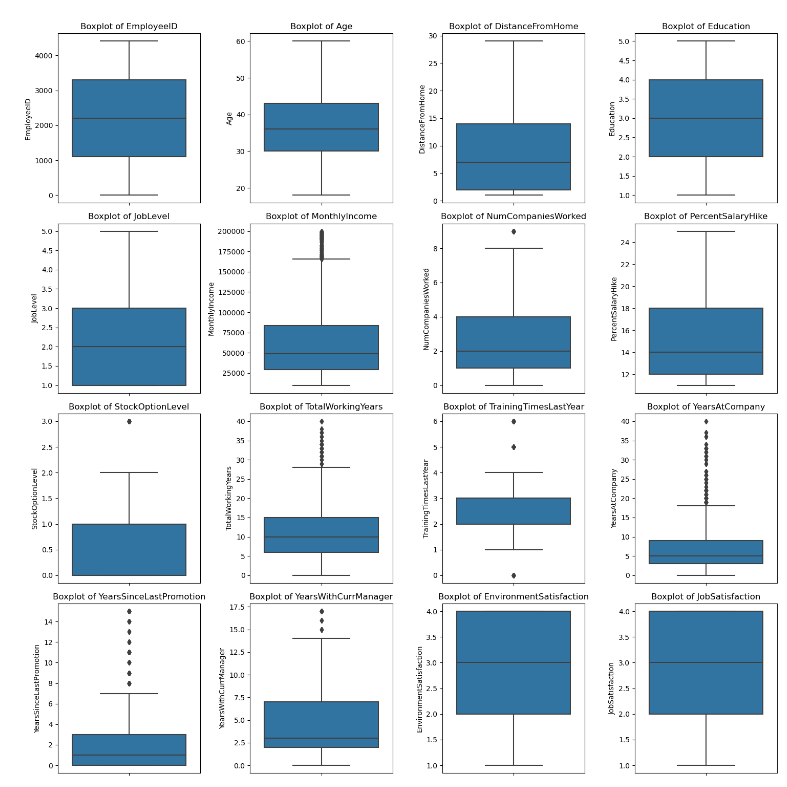


The above pairplot shows the trend of each and every feature with the other feature

The pairplot looks a bit messy due to large number of features present

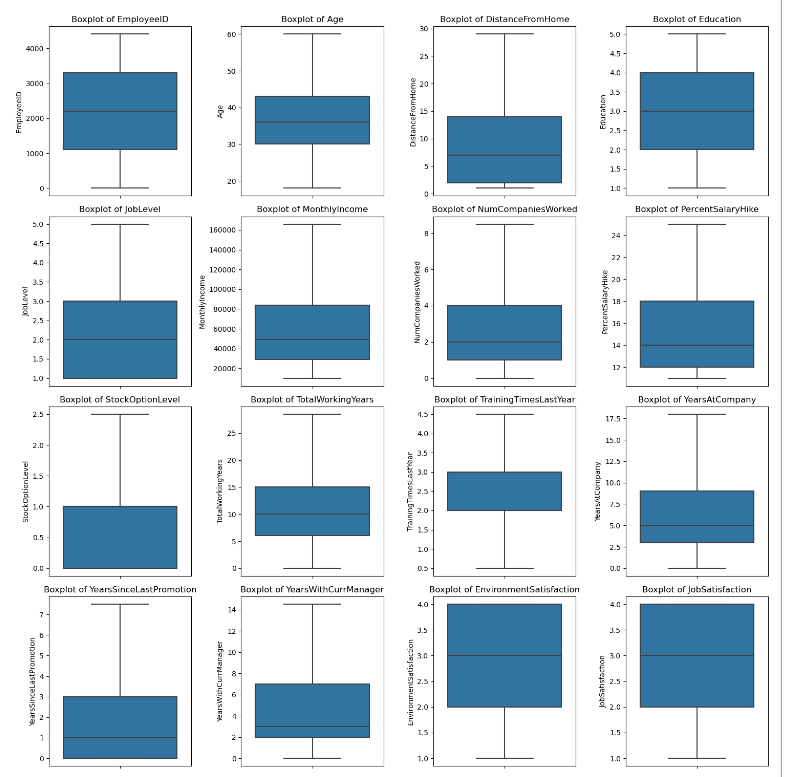
From the above pairplot, we may note that there is positive correlation, no correlation in some cases and sfew with shattered data values.

The box plots before outlier treatment :

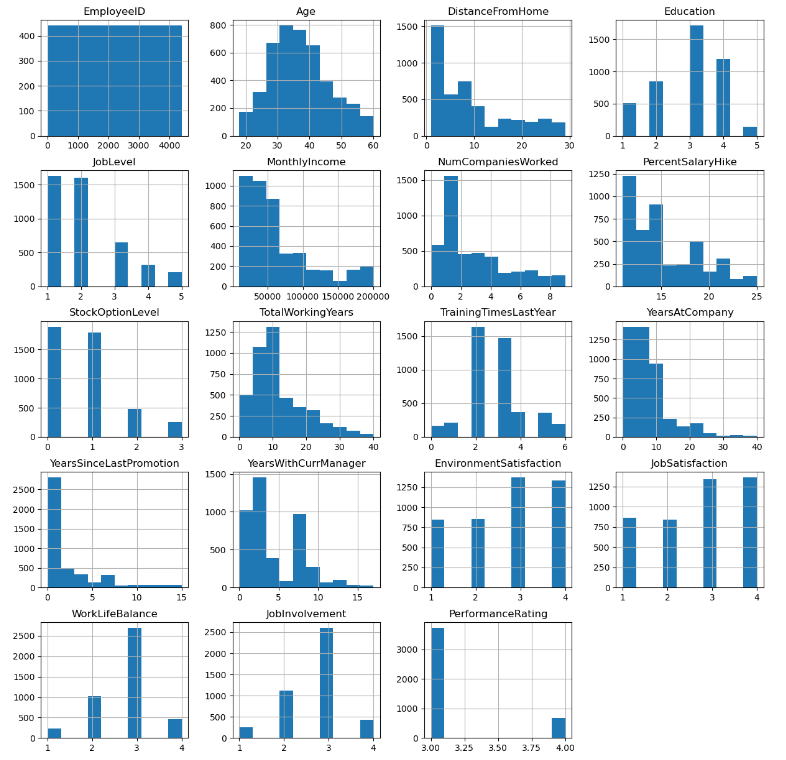


* There are outliers present in the data, these needs to be treated
* There are many methods in which outliers can be treated
* We choose IQR method to treat them
* So, we treat them using IQR method. In this method, any observation that is less than Q1- 1.5 IQR or more than Q3 + 1.5 IQR is considered an outlier.

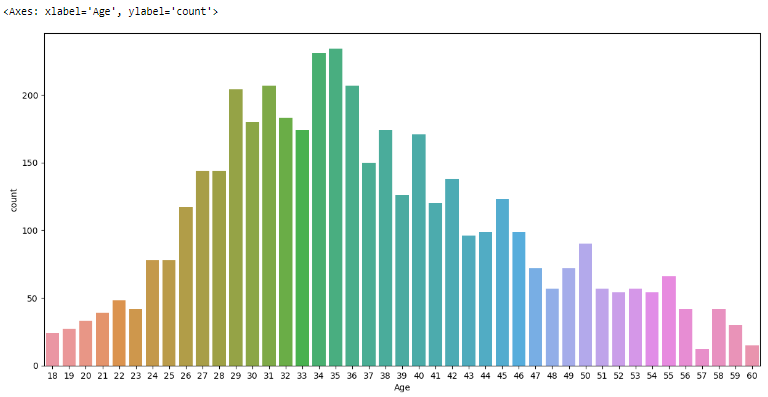
After outlier treatment :



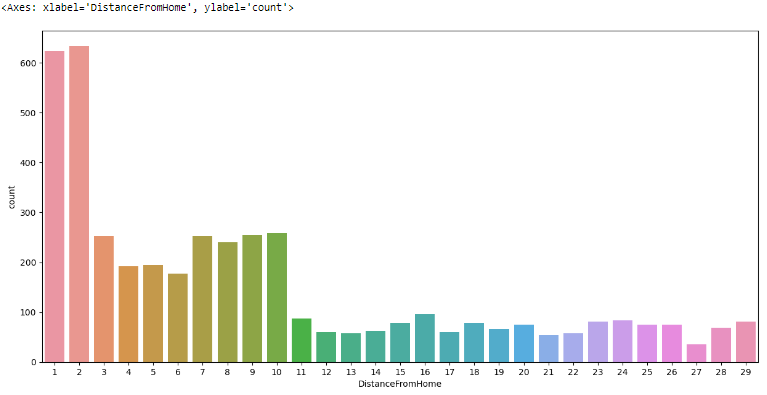
The histograms is as follows :

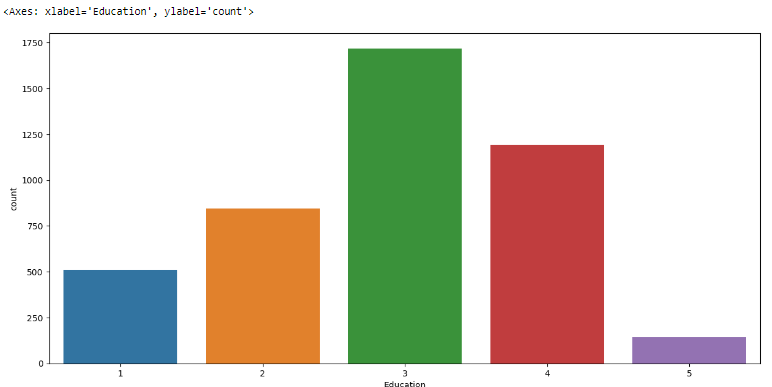


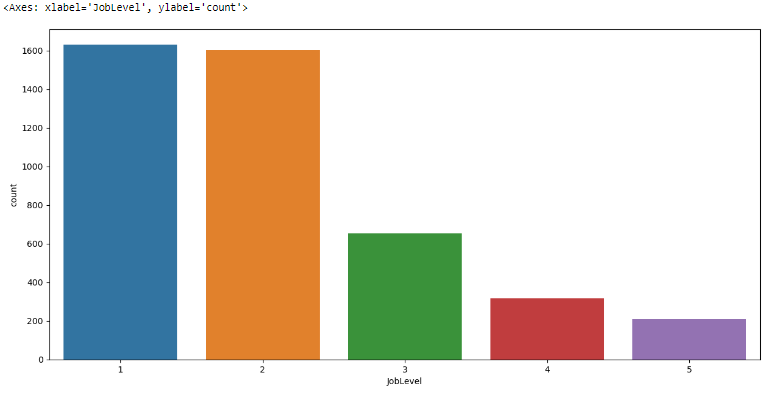
The countplot for different features present :

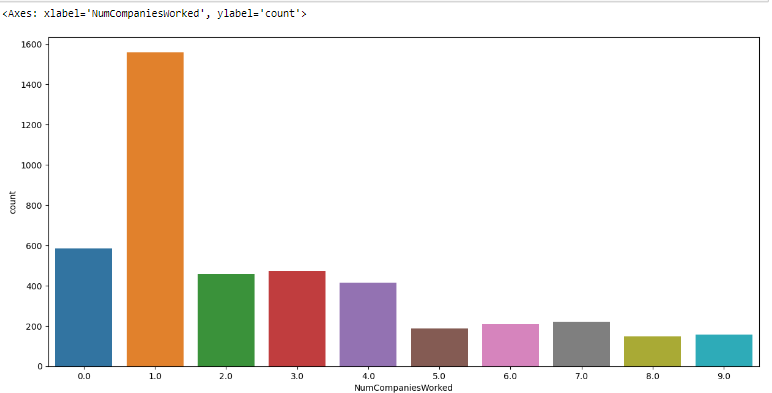


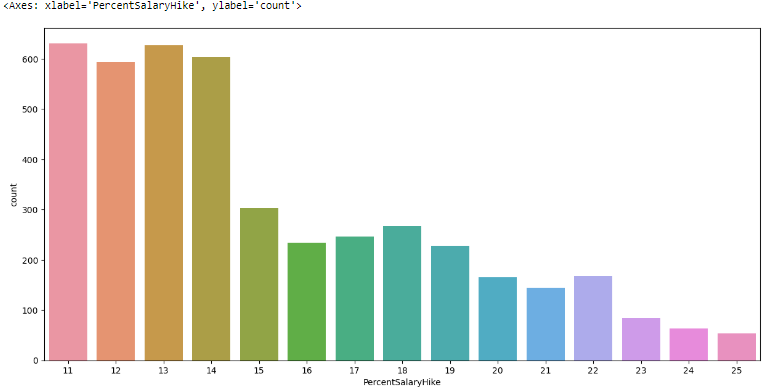
Most of the attritions happening mostly between the age groups 30-40

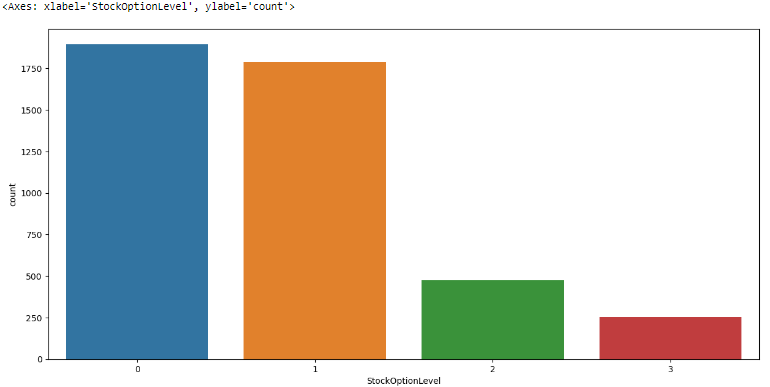


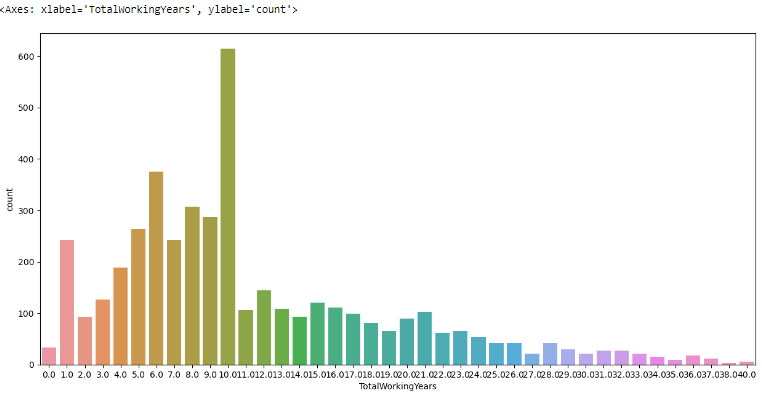


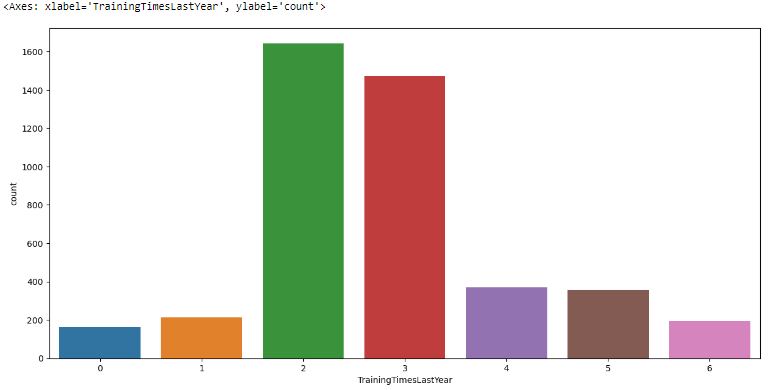


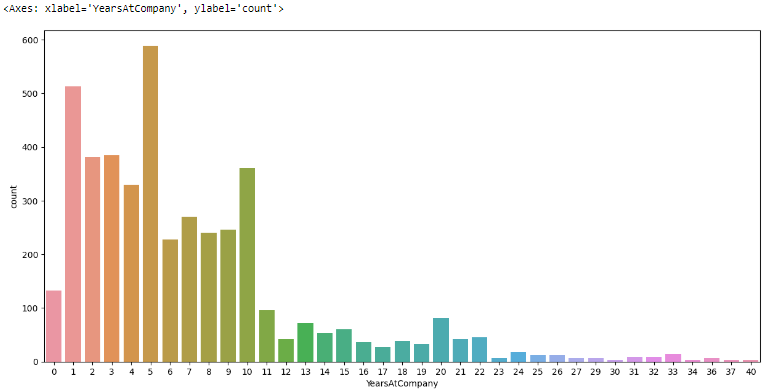


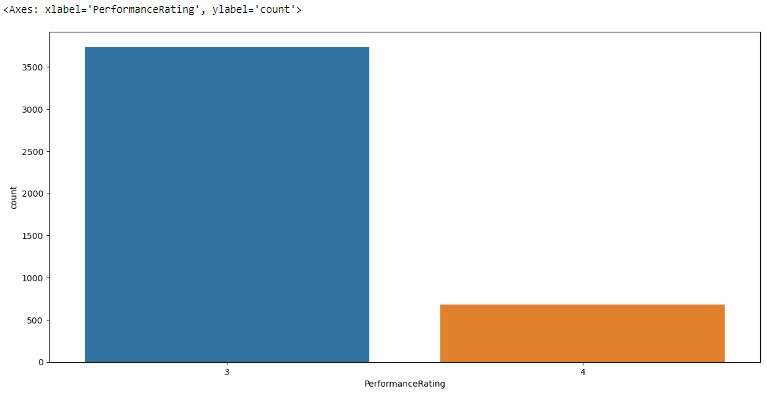
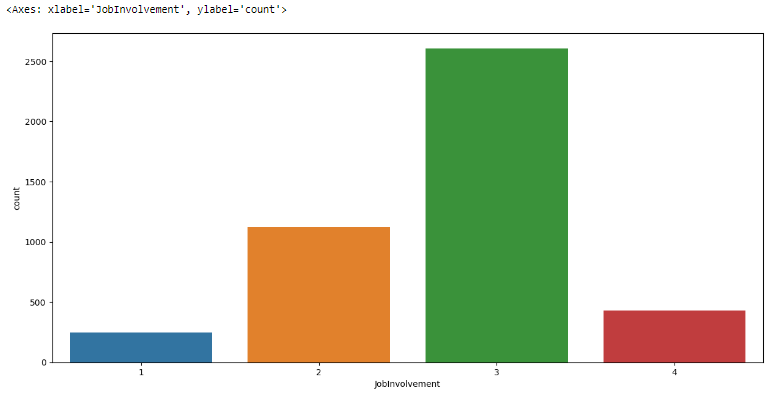
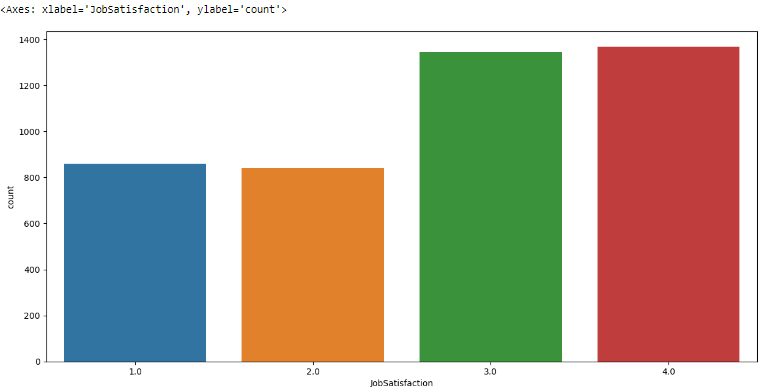
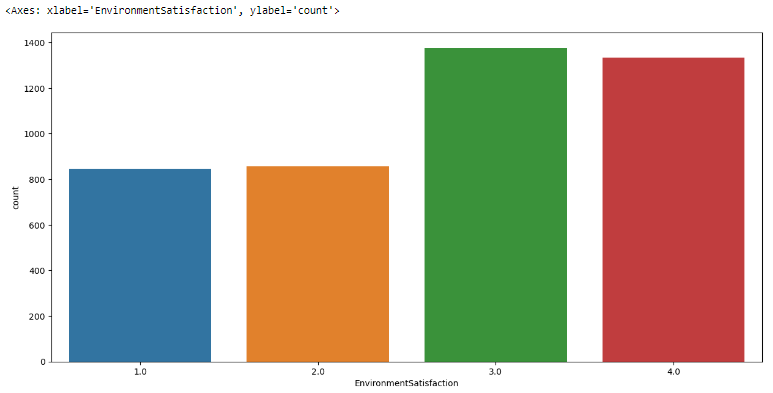
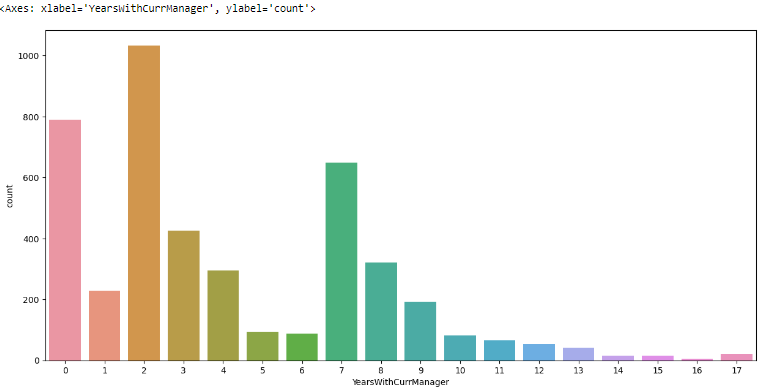
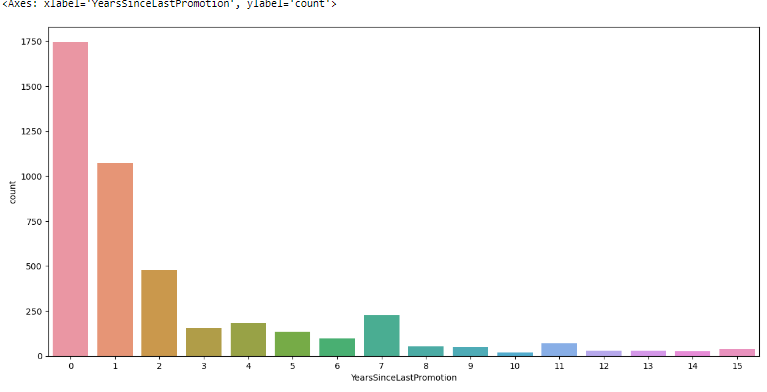


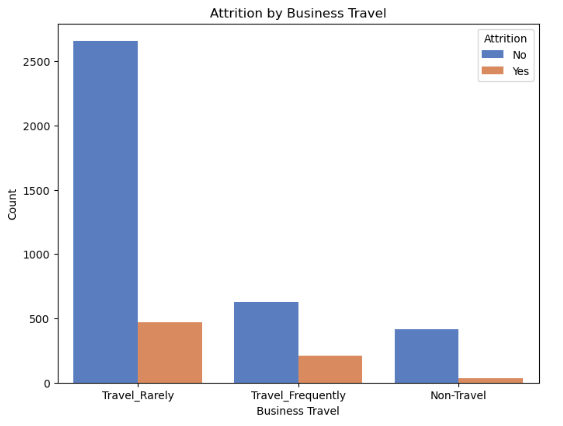


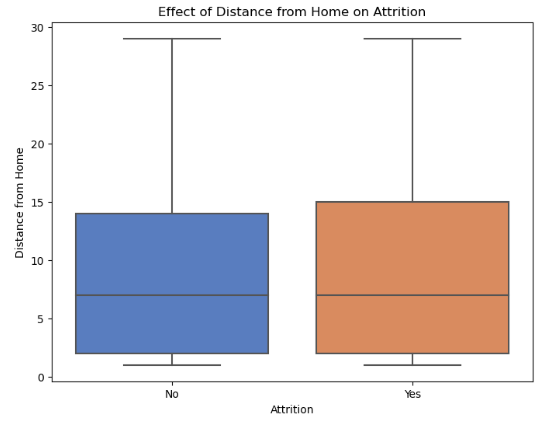


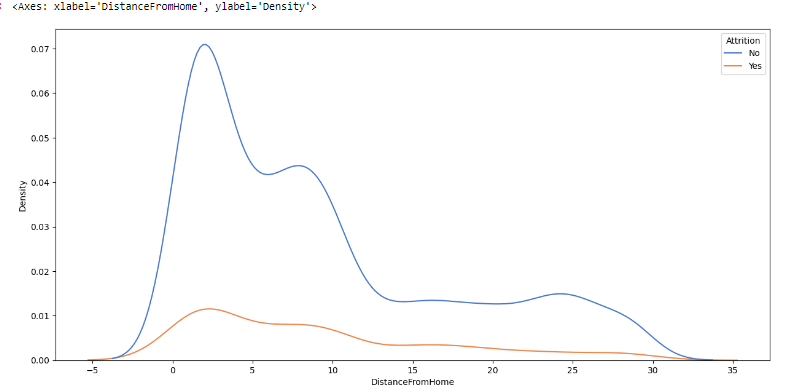


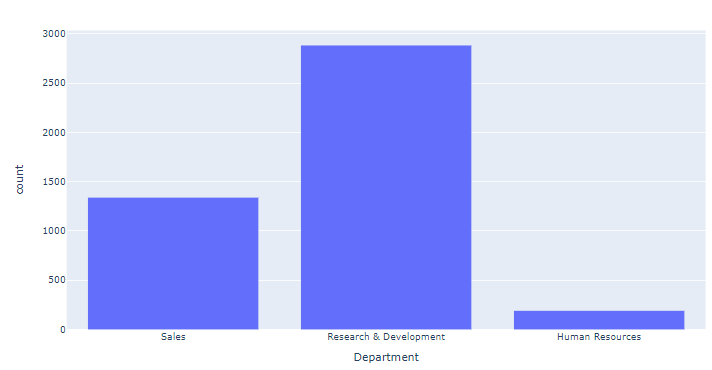


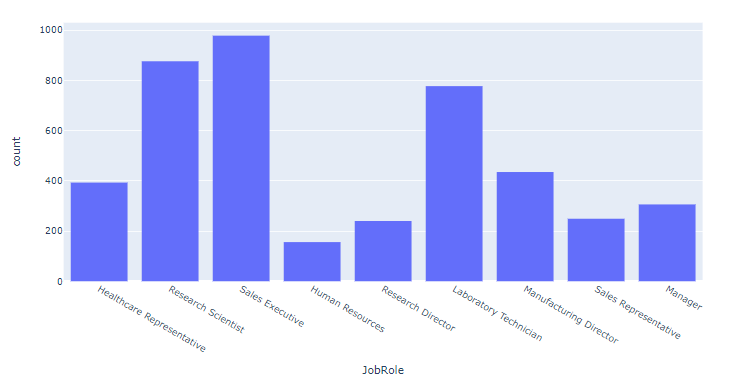


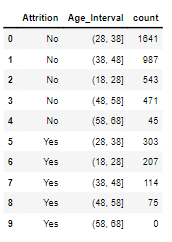
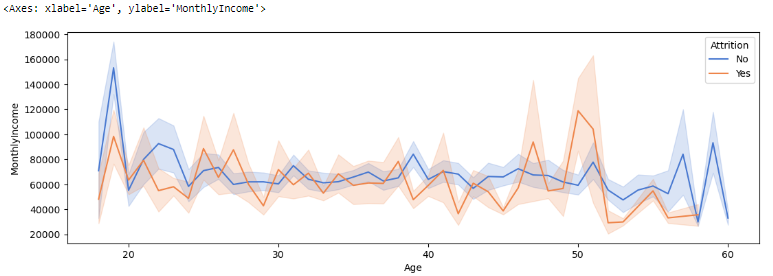


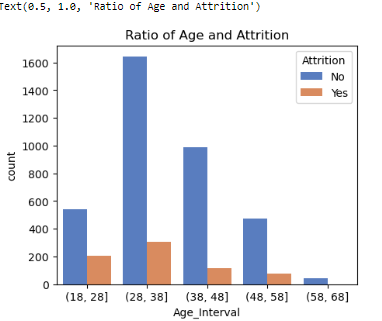
The attrition mostly happened with people who travel rarely.



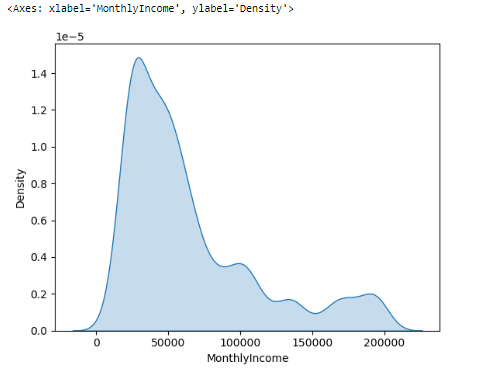


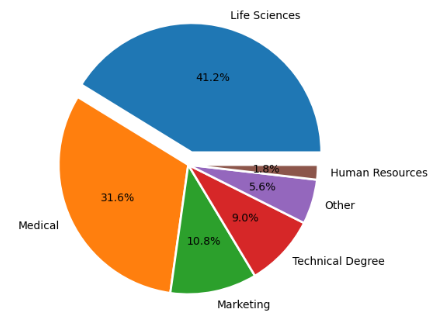
The most attrition happende in ‘ research and develpoment ‘ department.





The ratio of age and attrition plot shows that the age interval between 28 to 48





The most attrition happening in medical life science

