**Assignment Day 11 | July 20**

**Question:**

Find the correlation of Attrition variable with all other variables in the dataset

**Problem Statement:**

A large company named XYZ employs, at any given point of time, around 4000 employees. However, every year, around 15% of its employees leave the company and need to be replaced with the talent pool available in the job market. The management believes that this level of attrition (employees leaving, either on their own or because they got fired) is bad for the company, because of the following reasons -

The former employees projects get delayed, which makes it difficult to meet timelines, resulting in a reputation loss among consumers and partners A sizeable department has to be maintained, for the purposes of recruiting new talent More often than not, the new employees have to be trained for the job and/or given time to acclimatize themselves to the company

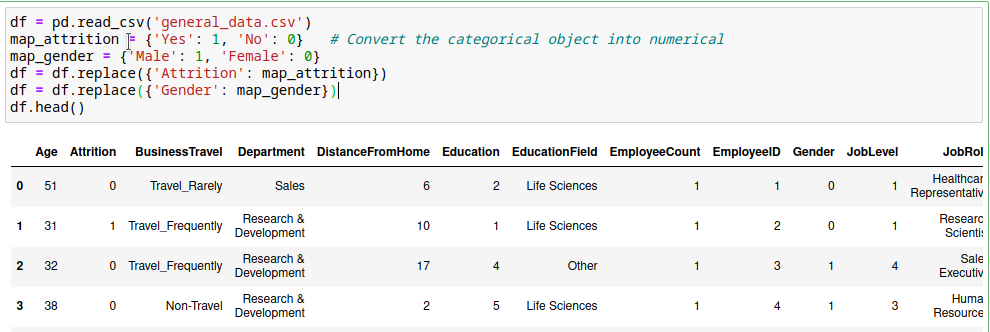
Hence, the management has contracted an HR analytics firm to understand what factors they should focus on, in order to curb attrition. In other words, they want to know what changes they should make to their workplace, in order to get most of their employees to stay. Also, they want to know which of these variables is most important and needs to be addressed right away.

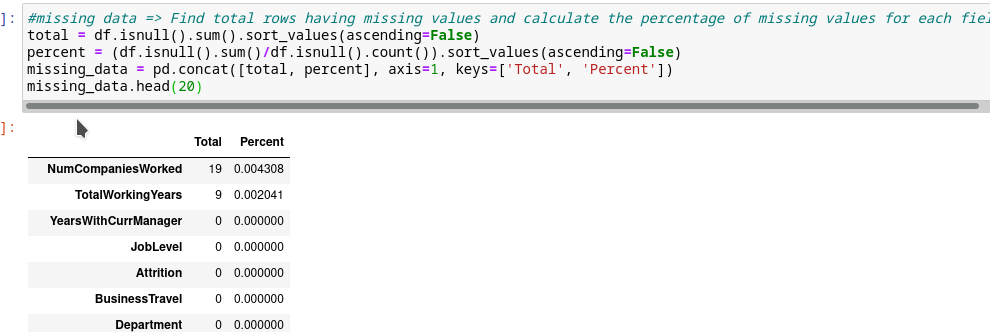
Since you are one of the star analysts at the firm, this project has been given to you.

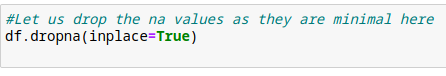
Goal of the case study you are required to model the probability of attrition. The results thus obtained will be used by the management to understand what changes they should make to their workplace, in order to get most of their employees to stay.

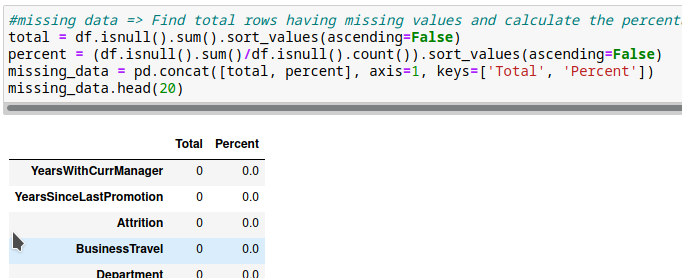
**Solutions:**

We can find out the correlation of Attrition variable with other variables and state how the independent variables are related to or dependent variable (Attrition).

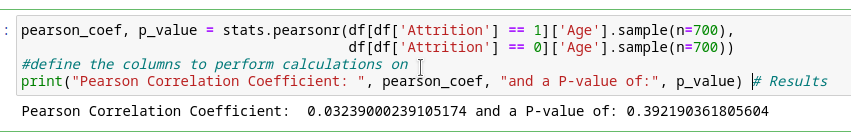
* To find the correlation of Attrition variable with other numerical variables, we converted Attrition variable into numerical variable by mapping the yes and no value with 1 and 0.
* We checked for any missing or null values and removed it from our dataset as they were very minimal.

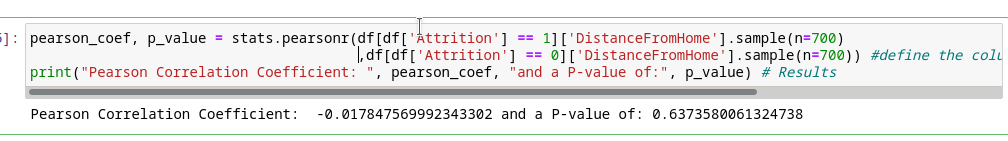


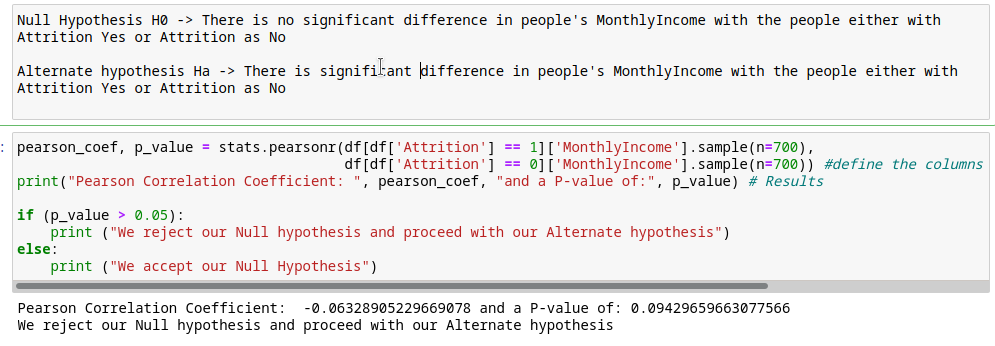




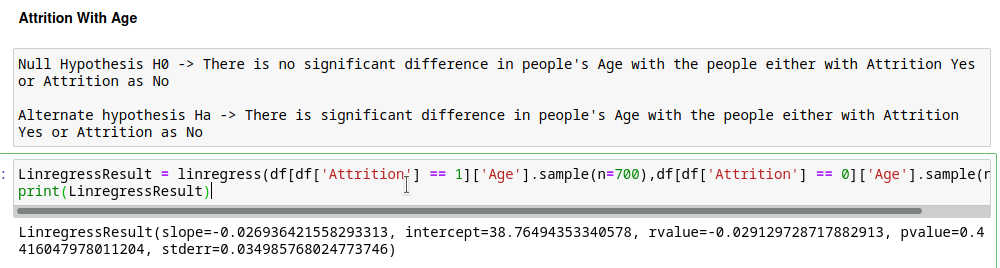
* We have found the Pearsonr correlation in python for different variable using the scipy.stats package.



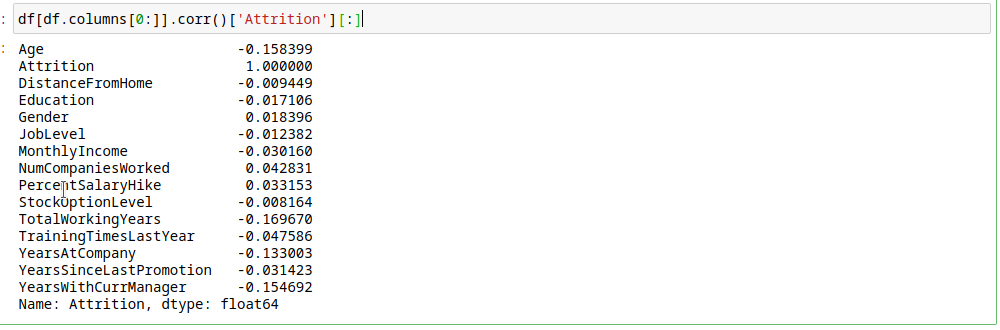


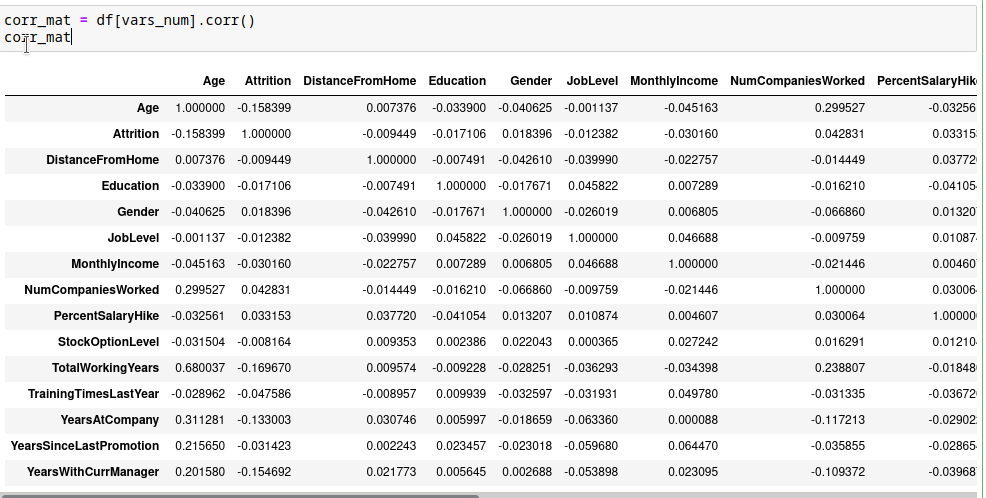


Also we found the correlation using linregress package

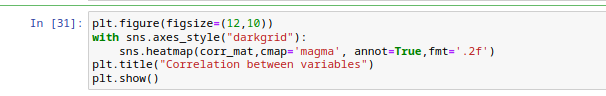


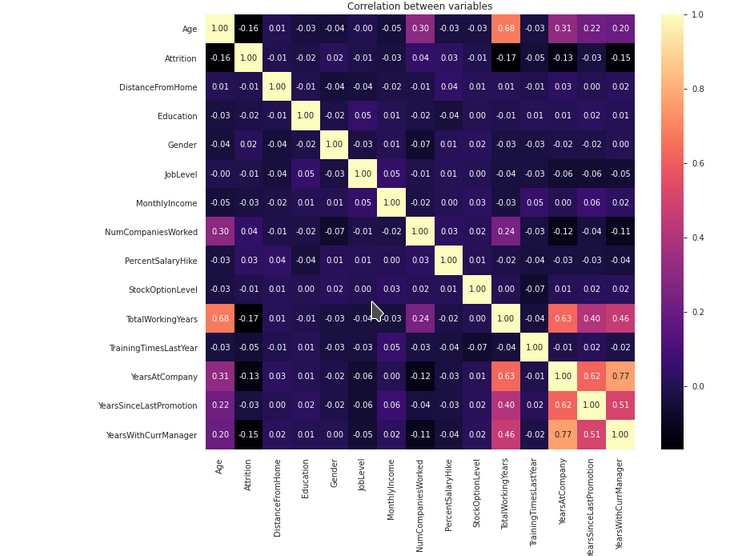
* Also we found the correlation of Attrition variable with all other numerical variables using pandas corr() function.





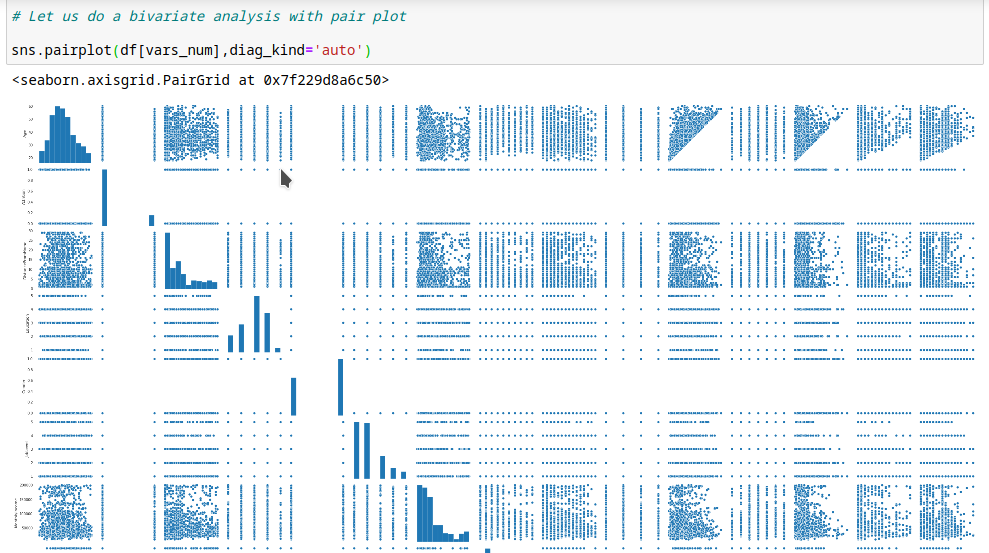
* We visualized the correlation in a heatmap in one matrix form.



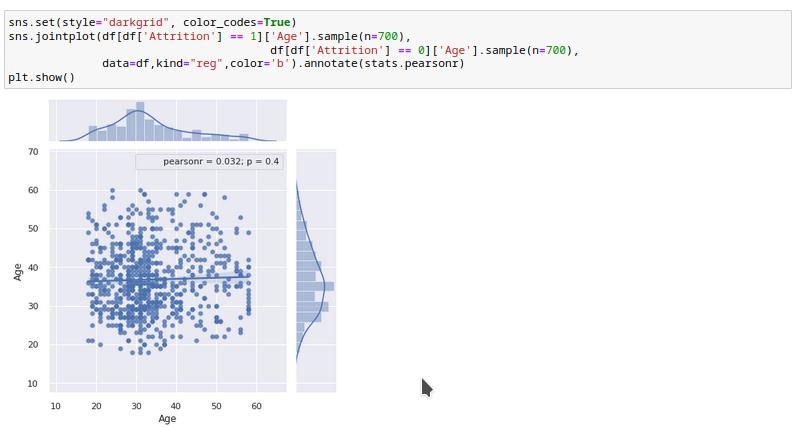


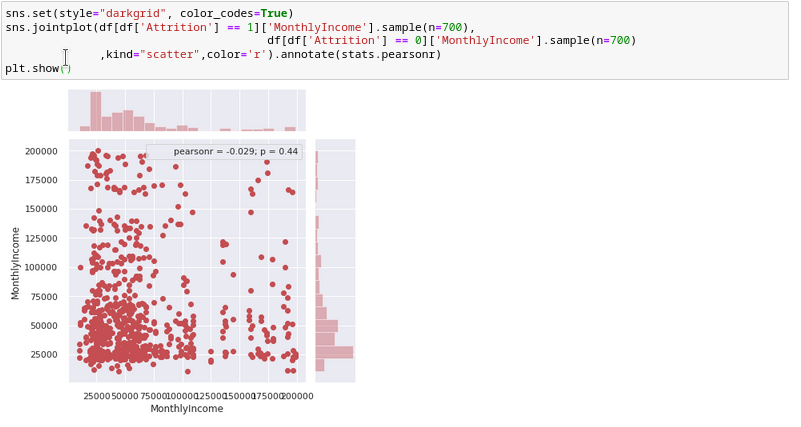
We can see here we get all the correlations of Attrition variable with all other as well as the correlations between the independent variables among themselves

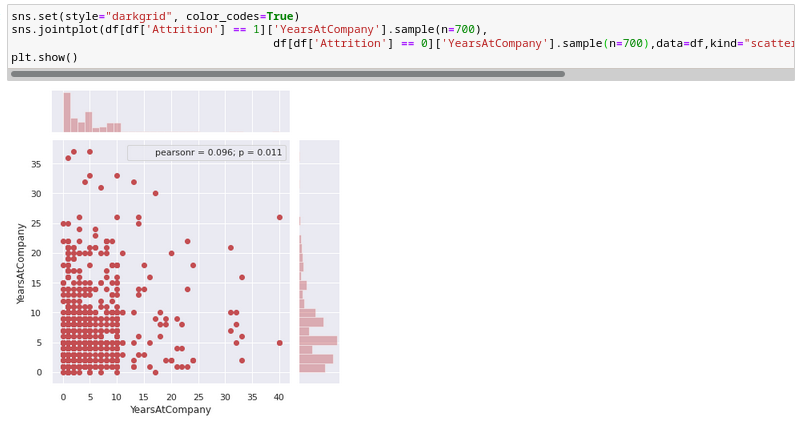
* Also we plotted scatter plot to find the relationship and co-linearity of independent and dependent variable and independent variables among themselves.

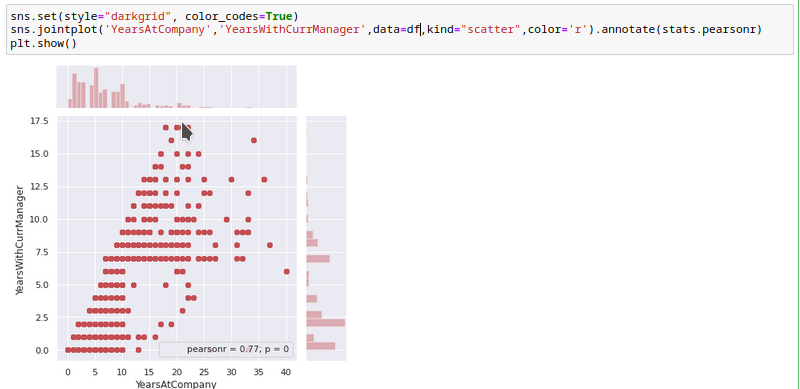


We also plot few joint plots stating the pearsonr coefficient in it.









So from above all visualizations and calculations we can draw some insights about correlation:

* There are almost weak correlation for all the independent variables with our dependent variable i.e Attrition variable
* Though the correlation is less but we have p-values which are greater than 0.05 so the variables like Age,DistanceFromHome,MonthlyIncome,YearsAtCompany,YearsSinceLastPromotion all have a significant relation with the dependent variable i.e Attrition
* There is a little stronger correlation between YearsAtCompany and YearsWithCurrentManager.