Attrition Case Study

**Problem Statement**–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 acclimatise 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.

1. **Import the below library’s**

import pandas as pd

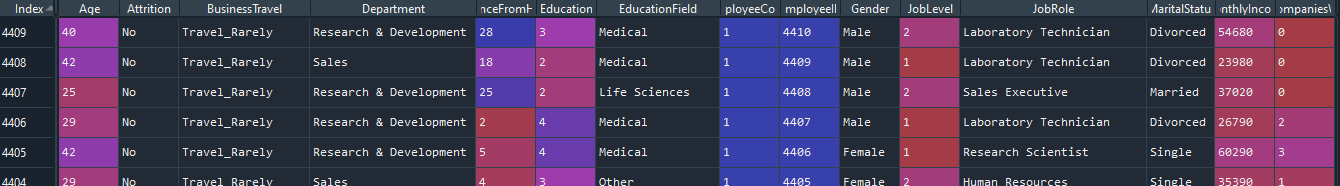
import numpy as np

import matplotlib.pyplot as plt

1. **Read the given file and display the data**

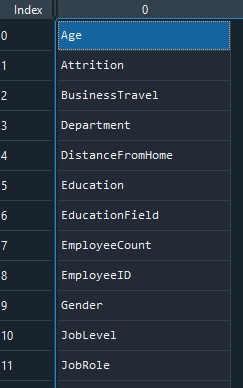
df = pd.read\_csv('general\_data.csv')

df.head()

****

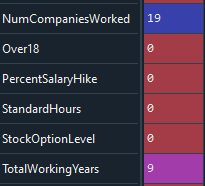
1. **To display the column name :**

df.columns

****

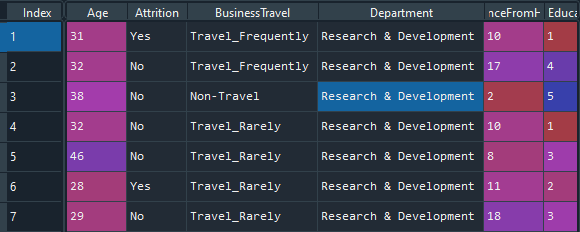
1. **To check the null values in columns:**

df\_null = df.isnull().sum()

****

1. **To drop the null values**

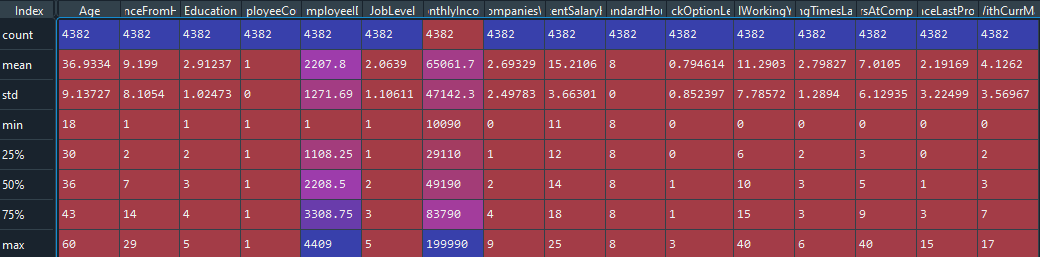
df1 = df.dropna()



**Univariate Analysis:**

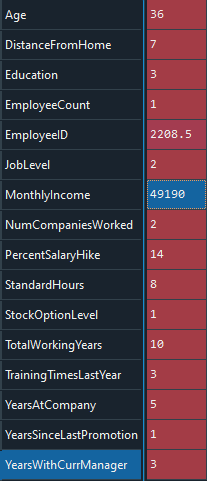
1. **To describe about the data**

df1\_des = df1.describe()



1. **To find the median values**

df\_median=df[['Age','DistanceFromHome','Education','MonthlyIncome','NumCompaniesWorked',PercentSalaryHike','TotalWorkingYears','TrainingTimesLastYear','YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].median()



1. **To find the mode**

df\_mode = df[['Age','DistanceFromHome','Education','MonthlyIncome',

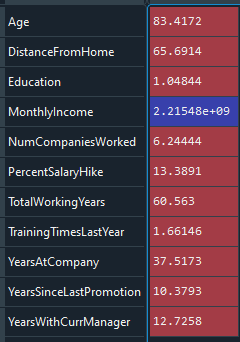
'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear',

'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].mode()

****

1. **To find the variant**

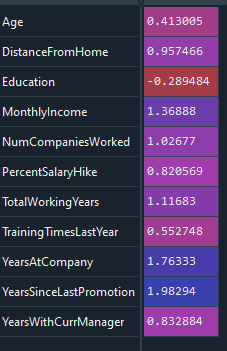
df\_var=df[['Age','DistanceFromHome','Education','MonthlyIncome','NumCompaniesWorked',’PercentSalaryHike','TotalWorkingYears',TrainingTimesLastYear','YearsAtCompany','YeasSinceLastPromotion', 'YearsWithCurrManager']].var()



1. **To find the skwenes**

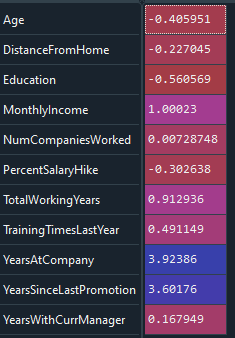
df\_skew=df[['Age','DistanceFromHome','Education','MonthlyIncome','NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear',

'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].skew()



1. **To find the kurtos**

df\_kurt=df[['Age','DistanceFromHome','Education','MonthlyIncome','NumCompaniesWord', 'PercentSalaryHike','TotalWorkingYears','TrainingTimesLastYear','YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].kurt()



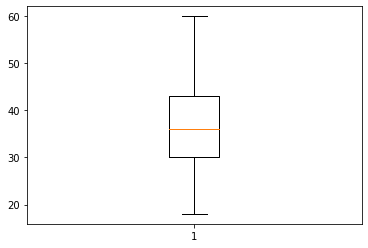
**Inference from the Analysis:**

* All the above variables are positively skewed (mean > median) except for Education which is negatively skewed.
* Age, DistanceFromHome, Education and PercentSalaryHike are platykurtic in nature while all the other variables are leptokurtic.
* Age forms a near normal distribution with 13 years of IQR.

1. **To Find outliers in age columns**

df\_box = df.Age

plt.boxplot(df\_box)

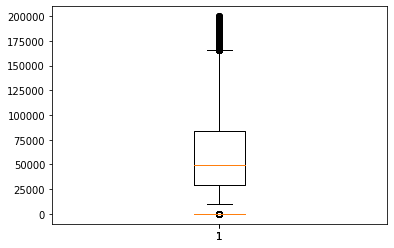


Age is no outerliers are found

1. **find outliers in monthly income columns**

df\_income = df.MonthlyIncome

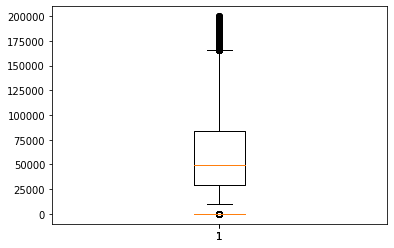
plt.boxplot(df\_income)



Montly income are Right skewed

1. **To find outliers in year at company columns**

plt.boxplot(df.YearsAtCompany)



Yearatcompany are Right skewed

1. **To find scatter plotts**

plt.scatter(df.Age,df.MonthlyIncome)

plt.xlable="Age"

plt.ylable="MonthlyIncome"

plt.show()

