**Step1: Launching**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

dataset1=pd.read\_csv("general\_data.csv")

**dataset1.head()**

Out[3]:

Age Attrition ... YearsSinceLastPromotion YearsWithCurrManager

0 51 No ... 0 0

1 31 Yes ... 1 4

2 32 No ... 0 3

3 38 No ... 7 5

4 32 No ... 0 4

**dataset1.columns**

Out[5]:

Index(['Age', 'Attrition', 'BusinessTravel', 'Department', 'DistanceFromHome',

'Education', 'EducationField', 'EmployeeCount', 'EmployeeID', 'Gender',

'JobLevel', 'JobRole', 'MaritalStatus', 'MonthlyIncome',

'NumCompaniesWorked', 'Over18', 'PercentSalaryHike', 'StandardHours',

'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear',

'YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManager'],

dtype='object')

**Step 2 - Data Treatment:**

**dataset1.isnull()**

Out[6]:

Age Attrition ... YearsSinceLastPromotion YearsWithCurrManager

0 False False ... False False

1 False False ... False False

2 False False ... False False

3 False False ... False False

4 False False ... False False

... ... ... ... ...

4405 False False ... False False

4406 False False ... False False

4407 False False ... False False

4408 False False ... False False

4409 False False ... False False

**dataset1.duplicated()**

Out[7]:

0 False

1 False

2 False

3 False

4 False

4405 False

4406 False

4407 False

4408 False

4409 False

Length: 4410, dtype: bool

**dataset1.drop\_duplicates()**

Out[8]:

Age Attrition ... YearsSinceLastPromotion YearsWithCurrManager

0 51 No ... 0 0

1 31 Yes ... 1 4

2 32 No ... 0 3

3 38 No ... 7 5

4 32 No ... 0 4

... ... ... ... ...

4405 42 No ... 0 2

4406 29 No ... 0 2

4407 25 No ... 1 2

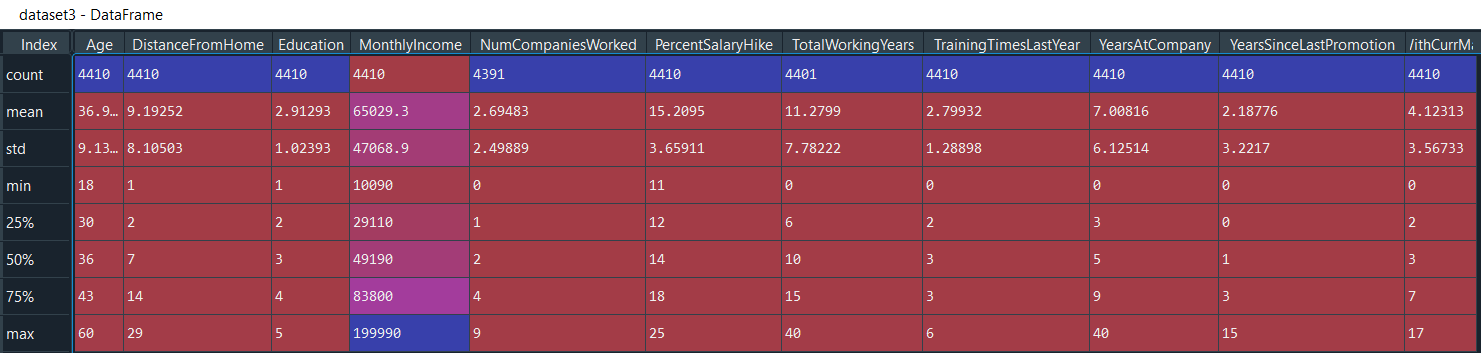
4408 42 No ... 7 8

4409 40 No ... 3 9

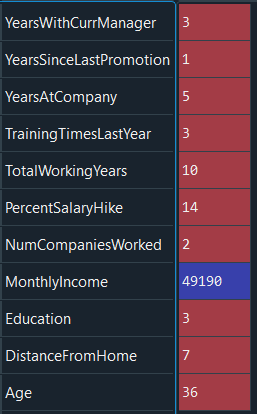
[4410 rows x 24 columns]

**Step 3 – Univariate**

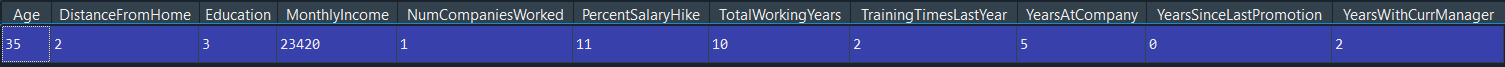
**dataset3=dataset1[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].describe()**



**dataset3=dataset1[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].median()**



**dataset3=dataset1[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].mode()**



**dataset3=dataset1[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].mean()**

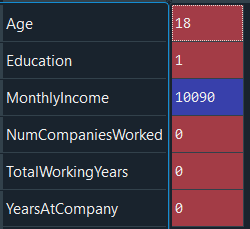


1. **The minimum age of employee in the company is 36.**

**dataset3=dataset1[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].var()**

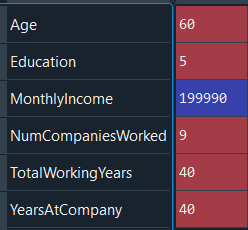


**dataset3=dataset1[['Age','Education','MonthlyIncome', 'NumCompaniesWorked','TotalWorkingYears', 'YearsAtCompany']].min()**



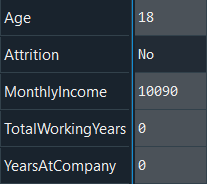
1. **The Minimum age of the employee in the company is 18 with salary of 10090.**

**dataset3=dataset1[['Age','Education','MonthlyIncome', 'NumCompaniesWorked','TotalWorkingYears', 'YearsAtCompany']].max()**

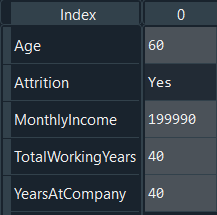


1. **The Maximum age of the employee in the company is 60 with the salary of 199990.**

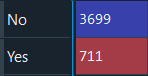
**dataset3=dataset1[['Age','Attrition','MonthlyIncome','TotalWorkingYears', 'YearsAtCompany']].min()**



**dataset3=dataset1[['Age','Attrition','MonthlyIncome','TotalWorkingYears', 'YearsAtCompany']].max()**

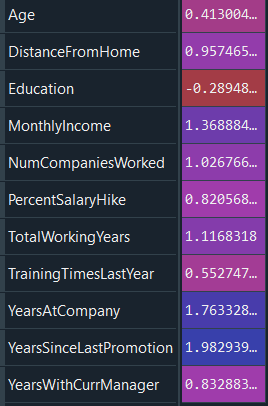


**dataset3=dataset1['Attrition'].value\_counts()**



1. **Count of the employee attrition, the number of employees that stayed (no) and the number that left (yes) the company**.

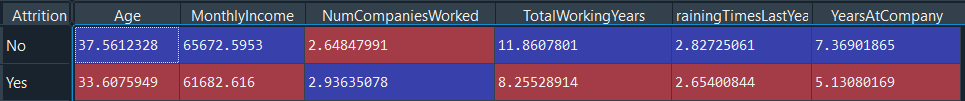
**dataset8=dataset1[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].skew()**



1. **All the above variables show positive skewness except education.**

**dataset4=dataset1[['Age','Attrition','MaritalStatus','MonthlyIncome', 'NumCompaniesWorked','TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany']]**

**hr=dataset4.groupby('Attrition').mean()**



1. **The average monthly work hours of employees who left the company is less than that of the employees who stayed.**
2. **The average monthly salary of employees who left the company is less than that of the employees who stayed.**

**dataset10=dataset1['MonthlyIncome']**

**plt.boxplot(dataset10)**

**Out[69]:**

**{'whiskers': [<matplotlib.lines.Line2D at 0x188a7165dc8>,**

**<matplotlib.lines.Line2D at 0x188a7165ec8>],**

**'caps': [<matplotlib.lines.Line2D at 0x188a7169d88>,**

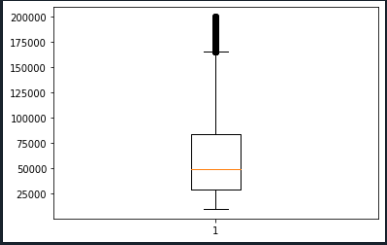
**<matplotlib.lines.Line2D at 0x188a7169e88>],**

**'boxes': [<matplotlib.lines.Line2D at 0x188a7165408>],**

**'medians': [<matplotlib.lines.Line2D at 0x188a716ed08>],**

**'fliers': [<matplotlib.lines.Line2D at 0x188a716ee08>],**

**'means': []}**



**8.Monthly income is rightly skewed has several outliers.**

**dataset10=dataset1['Age']**

**plt.boxplot(dataset10)**

**Out[67]:**

**{'whiskers': [<matplotlib.lines.Line2D at 0x188a7101f48>,**

**<matplotlib.lines.Line2D at 0x188a7104908>],**

**'caps': [<matplotlib.lines.Line2D at 0x188a7104f88>,**

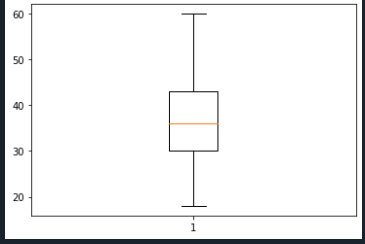
**<matplotlib.lines.Line2D at 0x188a7107908>],**

**'boxes': [<matplotlib.lines.Line2D at 0x188a7101748>],**

**'medians': [<matplotlib.lines.Line2D at 0x188a7107a08>],**

**'fliers': [<matplotlib.lines.Line2D at 0x188a710ca08>],**

**'means': []}**



**9.Age is normally distributed without outliers.**

**dataset10=dataset1['PercentSalaryHike']**

**plt.boxplot(dataset10)**

**Out[28]:**

**{'whiskers': [<matplotlib.lines.Line2D at 0x1d3833adf08>,**

**<matplotlib.lines.Line2D at 0x1d3833b18c8>],**

**'caps': [<matplotlib.lines.Line2D at 0x1d3833b1f48>,**

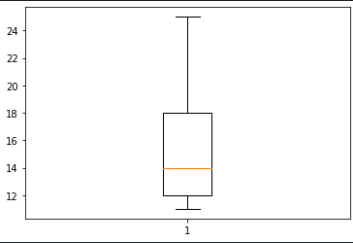
**<matplotlib.lines.Line2D at 0x1d3833b1fc8>],**

**'boxes': [<matplotlib.lines.Line2D at 0x1d3833ad608>],**

**'medians': [<matplotlib.lines.Line2D at 0x1d3833b4fc8>],**

**'fliers': [<matplotlib.lines.Line2D at 0x1d3833b4f88>],**

**'means': []}**



**10.Percent Salary Hike is normally distributed without outliers.**

**dataset10=dataset1['YearsAtCompany']**

**plt.boxplot(dataset10)**

**Out[30]:**

**{'whiskers': [<matplotlib.lines.Line2D at 0x1d383418fc8>,**

**<matplotlib.lines.Line2D at 0x1d383418f48>],**

**'caps': [<matplotlib.lines.Line2D at 0x1d38341af88>,**

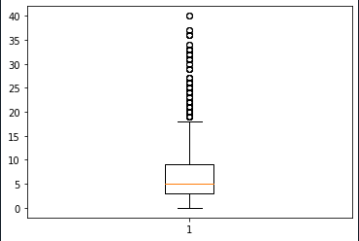
**<matplotlib.lines.Line2D at 0x1d38341af08>],**

**'boxes': [<matplotlib.lines.Line2D at 0x1d383418688>],**

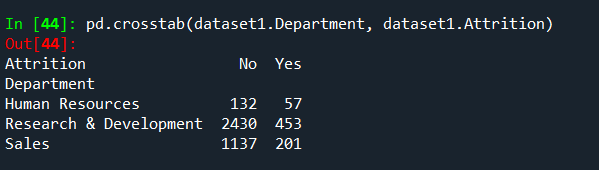
**'medians': [<matplotlib.lines.Line2D at 0x1d38341ef08>],**

**'fliers': [<matplotlib.lines.Line2D at 0x1d38341ee88>],**

**'means': []}**

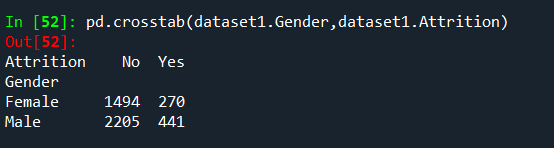


**11.Years At Company is rightly skewed has several outliers.**



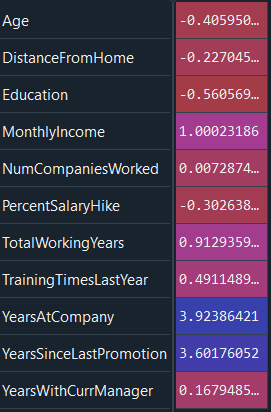
**12.The most attrition happened in Research and Development department.**

**13.The less attrition happened in Human Resource department.**



**14.There is more attrition of Male than Female.**

**dataset9=dataset1[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].kurt()**



**15.Age,DistanceFromHome,Education and PercentageSalaryHike are platykurtic while others are leptokurtic.**