**Attrition Project**

**HR Analytics Case Study:**

**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 talentMore 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.

**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.dropna(inplace=True)**

**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

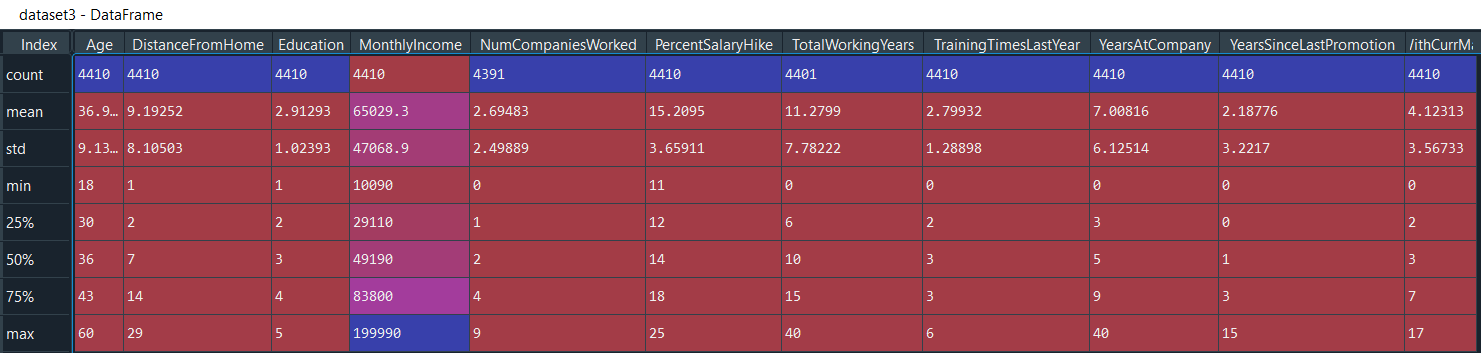
4408 42 No ... 7 8

4409 40 No ... 3 9

[4410 rows x 24 columns]

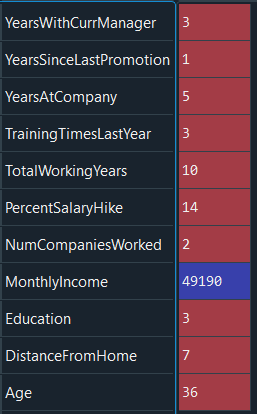
**Step 3 – Univariate**

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



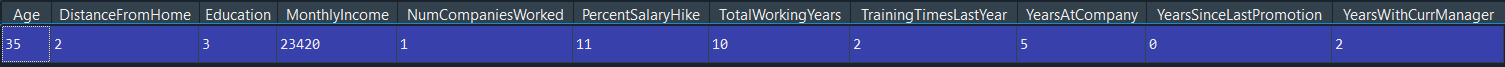
**Median:**

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



**Mode:**

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



**Mean:**

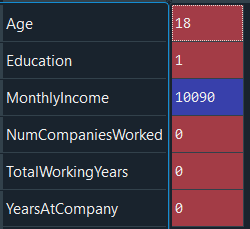
**dataset3=dataset1[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].mean()**



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



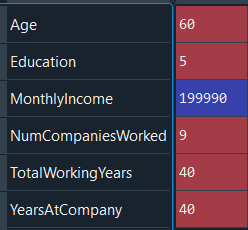
**dataset3=dataset1[['Age','Education','MonthlyIncome', 'NumCompaniesWorked','TotalWorkingYears', 'YearsAtCompany']].min()**



**Inference:**

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

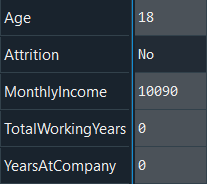
**dataset3=dataset1[['Age','Education','MonthlyIncome', 'NumCompaniesWorked','TotalWorkingYears', 'YearsAtCompany']].max()**



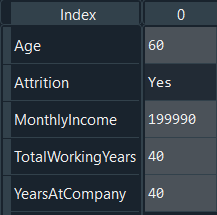
**Inference:**

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

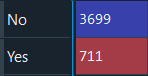
**dataset3=dataset1[['Age','Attrition','MonthlyIncome','TotalWorkingYears', 'YearsAtCompany']].min()**



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



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

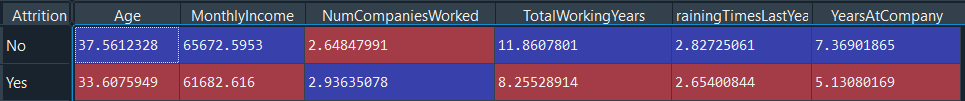


**Inference:**

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

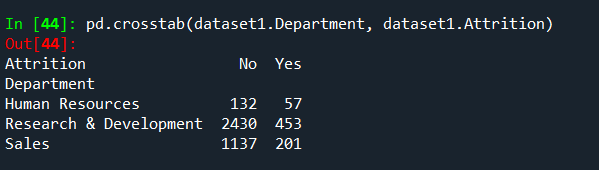
**dataset4=dataset1[['Age','Attrition','MaritalStatus','MonthlyIncome', 'NumCompaniesWorked','TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany']]**

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



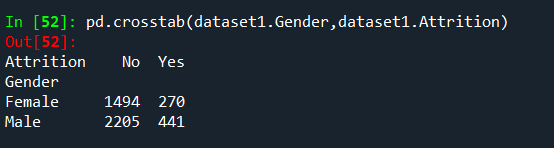
**Inferences:**

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.**



**Inferences:**

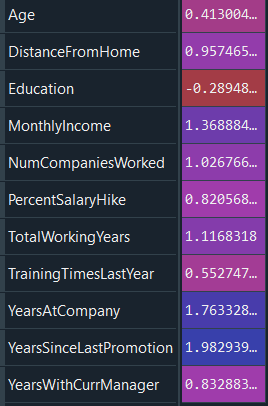
1. **The most attrition happened in Research and Development department.**
2. **The less attrition happened in Human Resource department.**



**Inference:**

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

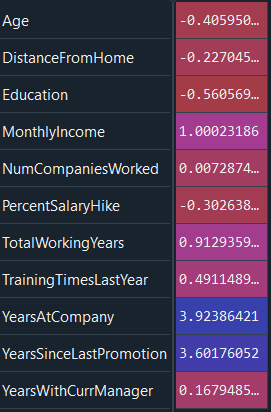
**dataset8=dataset1[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].skew()**



**Inference:**

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

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



**Inference:**

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

**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>,**

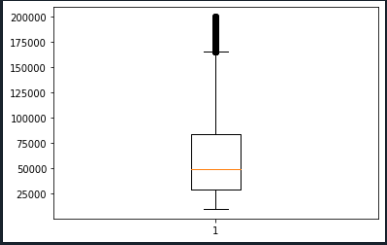
**<matplotlib.lines.Line2D at 0x188a7169e88>],**

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

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

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

**'means': []}**



**12.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>,**

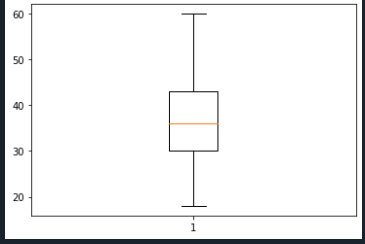
**<matplotlib.lines.Line2D at 0x188a7107908>],**

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

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

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

**'means': []}**



**13.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>,**

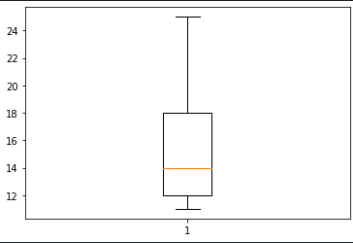
**<matplotlib.lines.Line2D at 0x1d3833b1fc8>],**

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

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

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

**'means': []}**



**14.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>,**

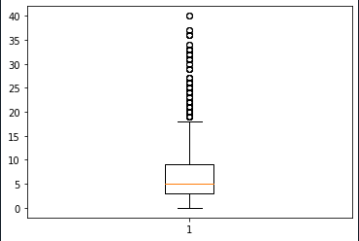
**<matplotlib.lines.Line2D at 0x1d38341af08>],**

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

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

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

**'means': []}**



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

Attr\_no= dataset1[dataset1['Attrition']== 0]

Attr\_no

Out[13]:

Age Attrition ... YearsSinceLastPromotion YearsWithCurrManager

0 51 0 ... 0 0

2 32 0 ... 0 3

3 38 0 ... 7 5

4 32 0 ... 0 4

5 46 0 ... 7 7

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

4404 29 0 ... 1 5

4405 42 0 ... 0 2

4406 29 0 ... 0 2

4407 25 0 ... 1 2

4408 42 0 ... 7 8

[3677 rows x 24 columns]

Attr\_yes= dataset1[dataset1['Attrition']== 1]

Attr\_yes

Out[15]:

Age Attrition ... YearsSinceLastPromotion YearsWithCurrManager

1 31 1 ... 1 4

6 28 1 ... 0 0

13 47 1 ... 9 9

28 44 1 ... 0 0

30 26 1 ... 0 2

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

4381 29 1 ... 0 1

4386 33 1 ... 0 4

4388 33 1 ... 1 7

4391 32 1 ... 1 2

4402 37 1 ... 0 0

[705 rows x 24 columns]

**Parametric and Nonparametric Test:**

**Mannwhitney test:**

**Attirition and Distance From Home:**

from scipy.stats import mannwhitneyu

stats,p=mannwhitneyu(Attr\_yes.DistanceFromHome,

Attr\_no.DistanceFromHome)

print(stats,p)

1295261.0 0.488538986087403

H0: There are no significant differences in the Distance from Home between those who left and those who stayed in the company.

Ha: There are significant differences in the Distance from Home between those who left and those who stayed in the company.

**Since p > 0.05, we reject the Alternate Hypothesis and null hypothesis is accepted.**

**Attirition and Monthly Income:**

stats, p = mannwhitneyu(Attr\_yes.MonthlyIncome,Attr\_no.MonthlyIncome)

print(stats,p)

1249573.5 0.06508807631576838

H0: There is no significant differences in the Monthly Income between those who left and those who stayed in the company.

Ha: There is significant differences in the Monthly Income between those who left and those who stayed in the company.

**Since p > 0.05, we reject the Alternate Hypothesis and null hypothesis is accepted.**

**Attirition and PercentSalaryHike:**

stats, p = mannwhitneyu(Attr\_yes.PercentSalaryHike, Attr\_no.PercentSalaryHike)

print(stats,p)

1231873.5 0.017810794960084964

H0: There is no significant differences in the percentage of salary hike between those who left and those who stayed in the company.

Ha: There is significant differences in the percentage of salary hike between those who left and those who stayed in the company.

**Since p < 0.05, alternative hypothesis is accepted and null hypothesis is rejected.**

**Chi Square Test:**

from scipy.stats import chi2\_contingency

MaritalStatus = {'Married': 0,'Single' : 1,'Divorced' : 2}

**Attrition and MaritalStatus:**

chitable = pd.crosstab (dataset.Attrition,dataset.MaritalStatus)

print (chitable)

MaritalStatus Divorced Married Single

Attrition

0 872 1756 1049

1 98 251 356

stats, p, dof, expeted = chi2\_contingency (chitable)

print(stats,p)

133.85785802925156 8.573051828219379e-30

H0: There is no dependency between Attrition and Marital Status.

Ha: There is dependency between Attrition and Marital Status.

**Since p < 0.05, alternative hypothesis is accepted and null hypothesis is rejected.**

**Attrition and Department:**

chitable = pd.crosstab(dataset.Attrition, dataset.Department)

stats, p, dof, expeted = chi2\_contingency (chitable)

print (chitable)

Department Human Resources Research & Development Sales

Attrition

0 132 2416 1129

1 55 449 201

print(stats,p)

25.89432541916022 2.382970570769315e-06

H0: There is no dependency between Attrition and Department.

Ha: There is dependency between Attrition and Department.

**Since p < 0.05, alternative hypothesis is accepted and null hypothesis is rejected.**

**Correlation:** dataset1["Attrition"]=dataset1["Attrition"].map({"Yes":1,"No":0})

dataset1["Gender"]=dataset1["Gender"].map({"Male":1,"Female":0})

**Correlation between Attrition and Age:**

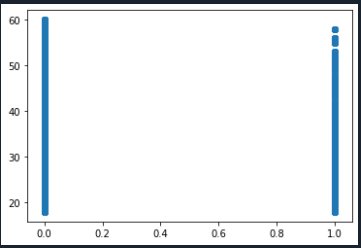
**stats,p=pearsonr(dataset.Attrition,dataset.Age)**

print(stats,p)

-0.159205006865775 1.996801615893625e-26

plt.scatter(dataset.Attrition,dataset.Age)

Out[16]: <matplotlib.collections.PathCollection at 0x2368c9fb808>



H0- Age doesn’t have significant impact on Attrition.

H1- Age has significant impact on Attrition.

**From r values we can conclude the age is negatively correlated with attrition.**

**As p <0.05, H0 i.e. null hypothesis is rejected.**

H1 Accepted i.e. Age has significant impact on Attrition.

**Correlation between Attrition and Gender:**

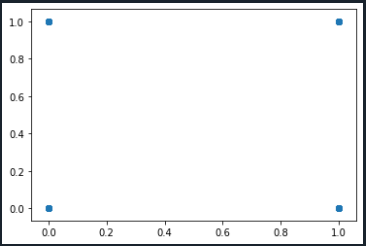
**stats,p=pearsonr(dataset.Attrition,dataset.Gender)**

print(stats,p)

0.018125078877010366 0.22881970951790567

plt.scatter(dataset.Attrition,dataset.Gender)

Out[19]: <matplotlib.collections.PathCollection at 0x2368caacdc8>



H0- Gender doesn’t have significant impact on Attrition.

H1- Gender has significant impact on Attrition.

**From r values we can conclude the gender is positively correlated with attrition.**

**As p >0.05, H0 i.e. null hypothesis is accepted.**

Gender doesn’t have significant impact on Attrition.

**Correlation between Attrition and MonthlyIncome:**

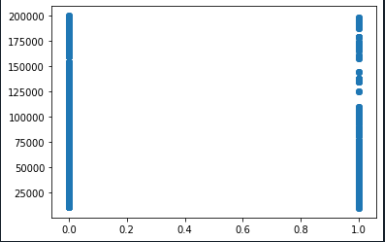
**stats,p=pearsonr(dataset.Attrition,dataset.MonthlyIncome)**

print(stats,p)

-0.031176281698114025 0.0384274849060192

plt.scatter(dataset.Attrition,dataset.MonthlyIncome)

Out[22]: <matplotlib.collections.PathCollection at 0x2368caacc48>



H0- Monthly Income doesn’t have significant impact on Attrition.

H1- Monthly Income has significant impact on Attrition.

**From r values we can conclude the Monthly Income is negatively correlated with attrition.**

**As p <0.05, H0 i.e. null hypothesis is rejected.**

H1 accepted Monthly Income has significant impact on Attrition.

**Correlation between Attrition and PercentSalaryHike:**

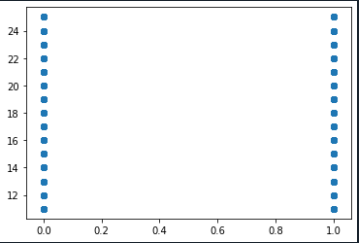
**stats,p=pearsonr(dataset.Attrition,dataset.PercentSalaryHike)**

print(stats,p)

0.03253259489105223 0.030743386433369824

plt.scatter(dataset.Attrition,dataset.PercentSalaryHike)

Out[25]: <matplotlib.collections.PathCollection at 0x2368cb89c88>



H0- Percent salary hike doesn’t have significant impact on Attrition.

H1- Percent salary hike has significant impact on Attrition.

**From r values we can conclude the percent salary hike is positively correlated with attrition.**

**As p <0.05, H0 i.e. null hypothesis is rejected.**

H1 accepetd- Percent salary hike has significant impact on Attrition.

**Correlation between Attrition and DistanceFromHome:**

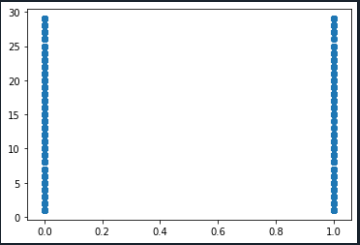
**stats,p=pearsonr(dataset.Attrition,dataset.DistanceFromHome)**

print(stats,p)

-0.009730141010179435 0.5182860428049617

plt.scatter(dataset.Attrition,dataset.DistanceFromHome)

Out[28]: <matplotlib.collections.PathCollection at 0x2368cbecd88>



H0- Distance from Home doesn’t have significant impact on Attrition.

H1- Distance from Home have significant impact on Attrition.

**From r values we can conclude the percent distance from home is negatively correlated with attrition.**

**As p >0.05, H0 i.e. null hypothesis is accepted.**

Distance from Home doesn’t have significant impact on Attrition.

**Correlation between Attrition and JobLevel:**

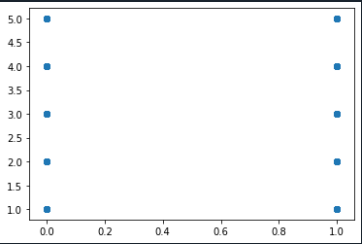
**stats,p=pearsonr(dataset.Attrition,dataset.JobLevel)**

print(stats,p)

-0.010289713287495079 0.49451717271828405

plt.scatter(dataset.Attrition,dataset.JobLevel)

Out[31]: <matplotlib.collections.PathCollection at 0x2368cbec148>



H0- Job Level doesn’t have significant impact on Attrition.

H1- Job Level have significant impact on Attrition.

**From r values we can conclude the job level is negatively correlated with attrition.**

**As p >0.05, H0 i.e. null hypothesis is accepted.**

Job Level doesn’t have significant impact on Attrition.

**Correlation between Attrition and YearsSinceLastPromotion:**

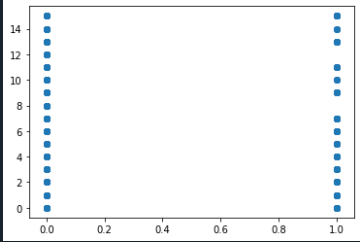
**stats,p=pearsonr(dataset.Attrition,dataset.YearsSinceLastPromotion)**

print(stats,p)

-0.03301877514258329 0.02833033618939086

plt.scatter(dataset.Attrition,dataset.YearsSinceLastPromotion)

Out[43]: <matplotlib.collections.PathCollection at 0x2368cc55248>



H0- Years since Last Promotion doesn’t have significant impact on Attrition.

H1- Years since Last Promotion have significant impact on Attrition.

**From r values we can conclude the job level is negatively correlated with attrition.**

**As p <0.05, H0 i.e. null hypothesis is rejected.**

H1 accepted -Years since Last Promotion have significant impact on Attrition.

**Correlation between Attrition and TotalWorkingYears:**

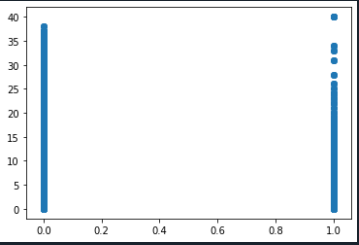
**stats,p=pearsonr(dataset.Attrition,dataset.TotalWorkingYears)**

print(stats,p)

-0.16966991684723265 1.1645434967153252e-29

plt.scatter(dataset.Attrition,dataset.TotalWorkingYears)

Out[10]: <matplotlib.collections.PathCollection at 0x201d8473e88>



H0- Total Working Years doesn’t have significant impact on Attrition.

H1- Total Working Years have significant impact on Attrition.

**From r values we can conclude the Total Working Years** **is negatively correlated with attrition.**

**As p <0.05, H0 i.e. null hypothesis is rejected.**

H1 accepted -Total Working Years have significant impact on Attrition.

**Correlation between Attrition and NumCompaniesWorked:**

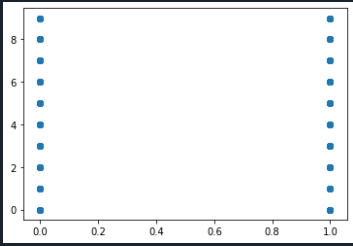
**stats,p=pearsonr(dataset.Attrition,dataset.NumCompaniesWorked)**

print(stats,p)

0.04283056724471892 0.004572057121624155

plt.scatter(dataset.Attrition,dataset.NumCompaniesWorked)

Out[14]: <matplotlib.collections.PathCollection at 0x201d85a7848>



H0- Number of companies worked doesn’t have significant impact on Attrition.

H1- Number of companies worked has significant impact on Attrition.

**From r values we can conclude the Number of companies worked** **is Positively correlated with attrition.**

**As p <0.05, H0 i.e. null hypothesis is rejected.**

**Correlation between Attrition and YearsAtCompany:**

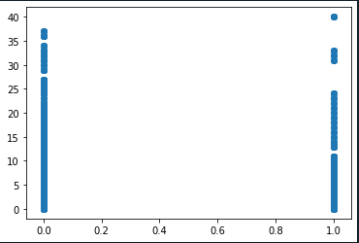
**stats,p=pearsonr(dataset.Attrition,dataset.YearsAtCompany)**

print(stats,p)

-0.1330026184252109 9.476118084864852e-19

plt.scatter(dataset.Attrition,dataset.YearsAtCompany)

Out[17]: <matplotlib.collections.PathCollection at 0x201d860a348>



H0- Years at Company worked doesn’t have significant impact on Attrition.

H1- Years at Company worked have significant impact on Attrition.

**From r values we can conclude the Number of companies worked** **is negatively correlated with attrition.**

**As p <0.05, H0 i.e. null hypothesis is rejected.**

H1 accepted - Years at Company worked have significant impact on Attrition.

**Conclusion:**

* As most attrition happened in Research and Development department, there is need to design good retention policies and strategies.
* Total working years and percent salary hike have significant impact on Attrition so there is need to review appraisal policies.
* Company should conduct exit interview to understand the expectations of employees at the workplace.
* Personality development programs should be conducted for boosting employee’s self-esteem.