ATTRITION ASSIGNMENT

STEP 1 = LAUNCHING :

import pandas as p

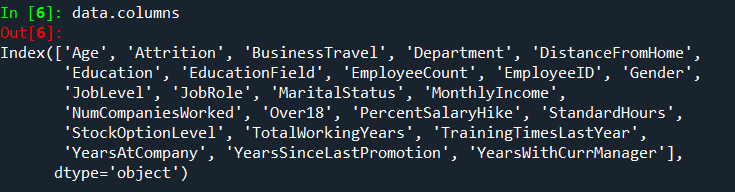
import numpy as n

import matpolib.pyplot as pl

data=pd.read\_csv(“general\_data.csv”)

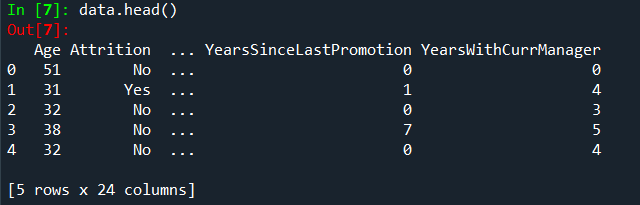
->To find column names .

data.columns



->To find the data of first 5 rows.

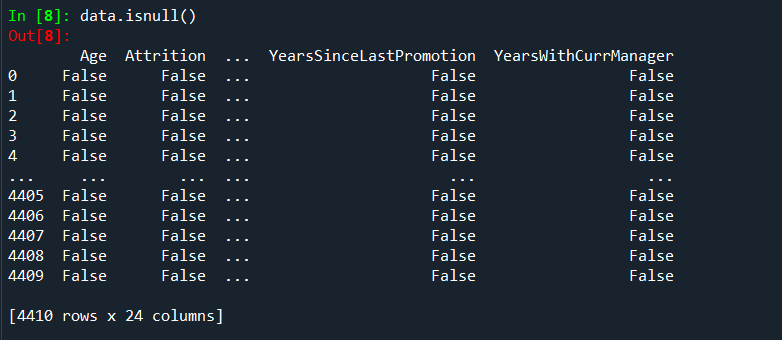
data.head( )



STEP 2 = DATA TREATMENT :

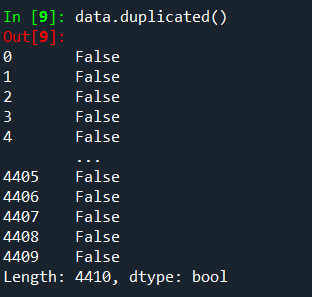
->To find out null values in the table.

data.isnull( )



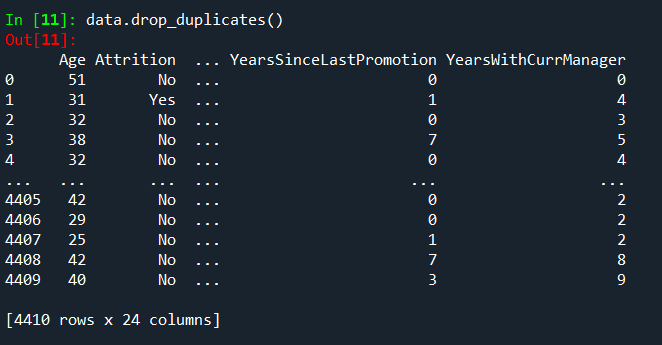
->To find out duplicated values of table.

data.duplicated( )



->To drop all duplicated values of the table.

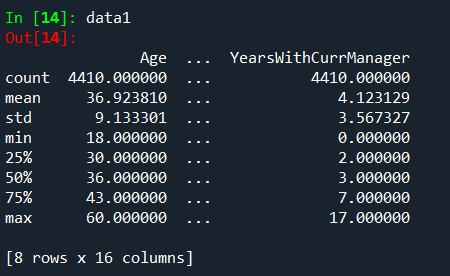
data.drop\_duplicates( )



STEP 3 = UNIVARIATE ANALYSIS :

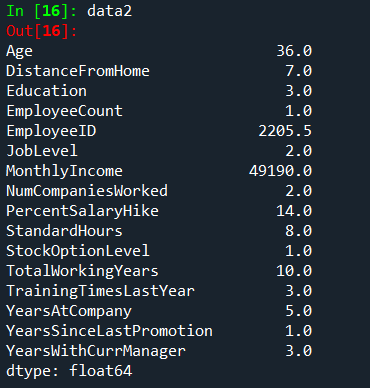
->To describe the whole table.

data1=data[['Age', 'Attrition', 'BusinessTravel', 'Department', 'DistanceFromHome','Education', 'EducationField', 'EmployeeCount', 'EmployeeID', 'Gender','JobLevel', 'JobRole', 'MaritalStatus', 'MonthlyIncome', 'NumCompaniesWorked', 'Over18', 'PercentSalaryHike', 'StandardHours','StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear','YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManager']].describe( )



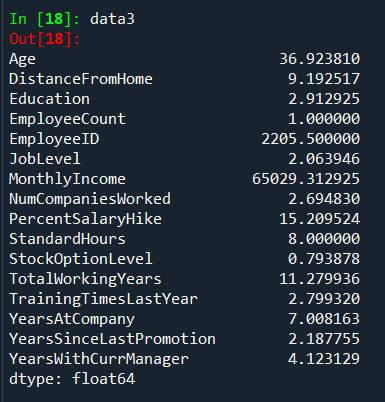
->To find out median of each column .

data2=data[['Age', 'Attrition', 'BusinessTravel', 'Department', 'DistanceFromHome','Education', 'EducationField', 'EmployeeCount', 'EmployeeID', 'Gender','JobLevel', 'JobRole', 'MaritalStatus', 'MonthlyIncome', 'NumCompaniesWorked', 'Over18', 'PercentSalaryHike', 'StandardHours','StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear','YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManager']].median( )



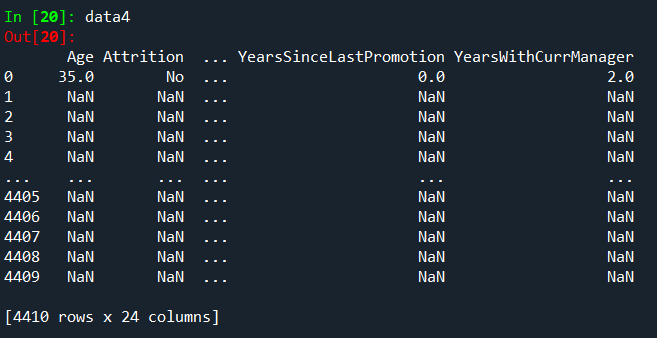
->To find out mean of each column.

data3=data[['Age', 'Attrition', 'BusinessTravel', 'Department', 'DistanceFromHome','Education', 'EducationField', 'EmployeeCount', 'EmployeeID', 'Gender','JobLevel', 'JobRole', 'MaritalStatus', 'MonthlyIncome', 'NumCompaniesWorked', 'Over18', 'PercentSalaryHike', 'StandardHours','StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear','YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManager']].mean( )



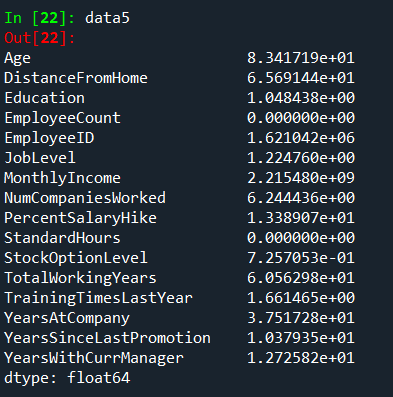
->To find out mode.

data4=data[['Age', 'Attrition', 'BusinessTravel', 'Department', 'DistanceFromHome','Education', 'EducationField', 'EmployeeCount', 'EmployeeID', 'Gender','JobLevel', 'JobRole', 'MaritalStatus', 'MonthlyIncome', 'NumCompaniesWorked', 'Over18', 'PercentSalaryHike', 'StandardHours','StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear','YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManager']].mode( )



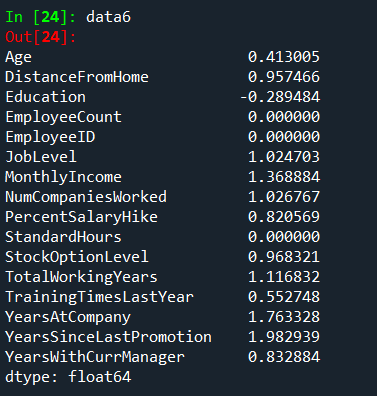
->To find variance of each column.

data5=data[['Age', 'Attrition', 'BusinessTravel', 'Department', 'DistanceFromHome','Education', 'EducationField', 'EmployeeCount', 'EmployeeID', 'Gender','JobLevel', 'JobRole', 'MaritalStatus', 'MonthlyIncome', 'NumCompaniesWorked', 'Over18', 'PercentSalaryHike', 'StandardHours','StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear','YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManager']].var( )



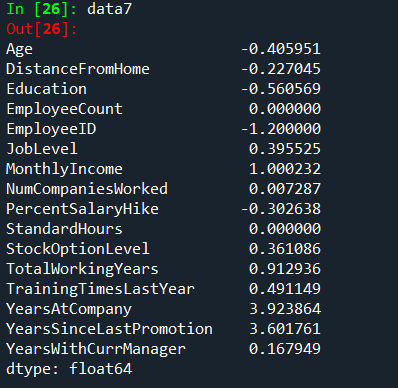
->To find skewness.

data6=data[['Age', 'Attrition', 'BusinessTravel', 'Department', 'DistanceFromHome','Education', 'EducationField', 'EmployeeCount', 'EmployeeID', 'Gender','JobLevel', 'JobRole', 'MaritalStatus', 'MonthlyIncome', 'NumCompaniesWorked', 'Over18', 'PercentSalaryHike', 'StandardHours','StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear','YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManager']].skew( )



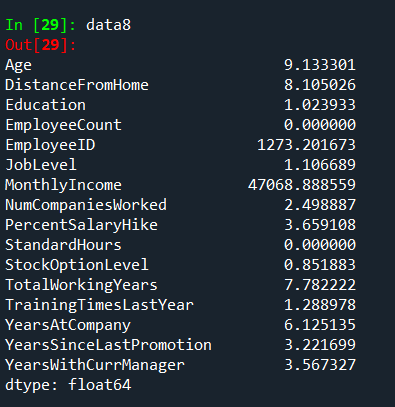
->To find out kurtosis.

data7=data[['Age', 'Attrition', 'BusinessTravel', 'Department', 'DistanceFromHome','Education', 'EducationField', 'EmployeeCount', 'EmployeeID', 'Gender','JobLevel', 'JobRole', 'MaritalStatus', 'MonthlyIncome', 'NumCompaniesWorked', 'Over18', 'PercentSalaryHike', 'StandardHours','StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear','YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManager']].kurt( )

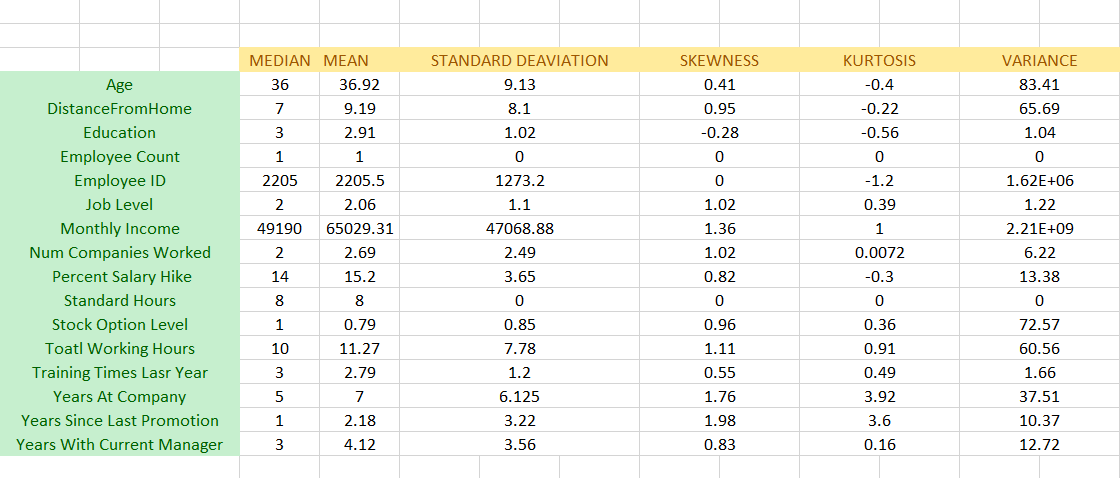


->To find standard deviation .

data8=data[['Age', 'Attrition', 'BusinessTravel', 'Department', 'DistanceFromHome','Education', 'EducationField', 'EmployeeCount', 'EmployeeID', 'Gender','JobLevel', 'JobRole', 'MaritalStatus', 'MonthlyIncome', 'NumCompaniesWorked', 'Over18', 'PercentSalaryHike', 'StandardHours','StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear','YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManager']].std( )



INFERENCE :



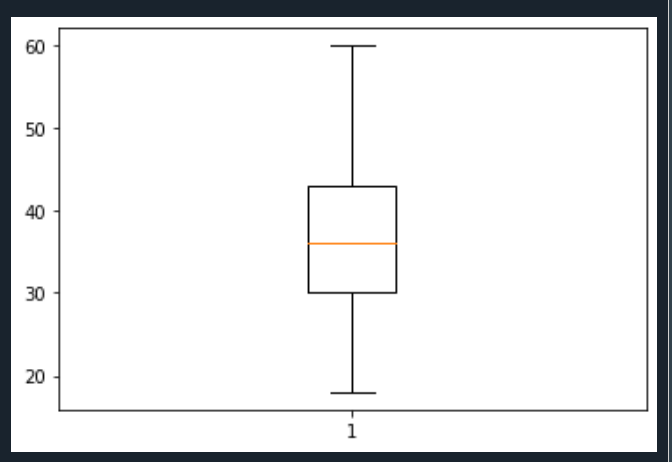
->All the above variables show positive skewness; while Age &Mean\_distance\_from\_home are leptokurtic and all other variables are platykurtic.

OUTLIERS :

There’s no regression found while plotting Age, MonthlyIncome, TotalWorkingYears , YearsAtCompany, etc., on a scatter plot.

box\_plot=data.Age

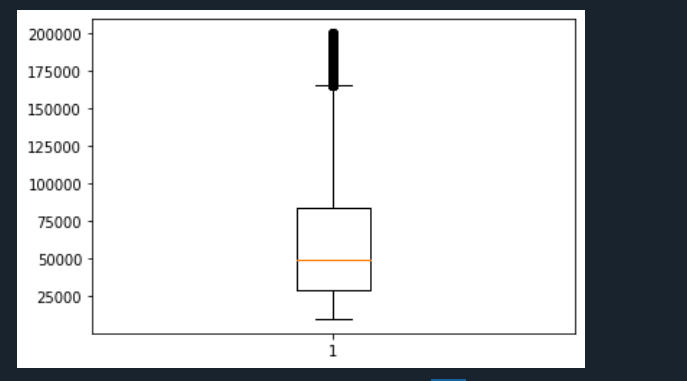
pl.boxplot(box\_plot)



Age is normally distributed without any outliers

box\_plot=data.MonthlyIncome

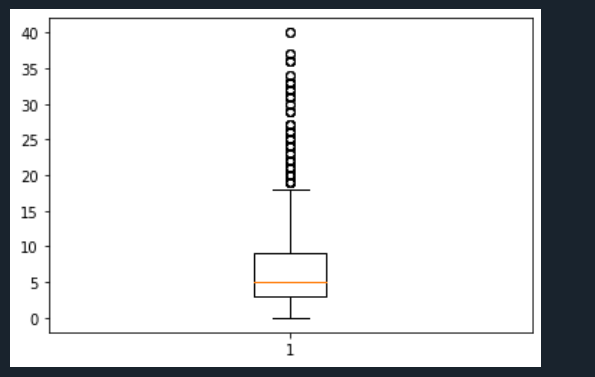
pl.boxplot(box\_plot)



Monthly Income is Right skewed with several outliers

box\_plot=data.YearsAtCompany

pl.boxplot(box\_plot)



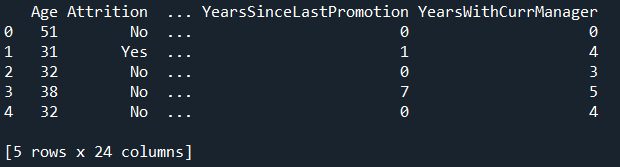
Years at company is also Right Skewed with several outliers observed.

**4>STATISTICAL TEST ( MANN-WHITNEY)**

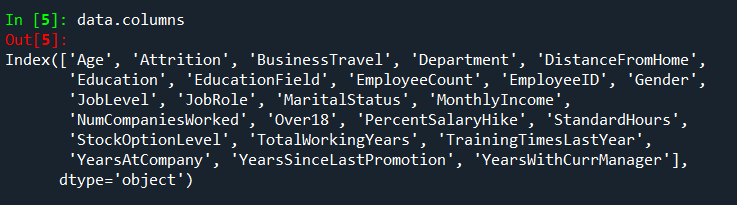
import pandas as pd

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

data.head()



data.columns



**MANN WHTNEY TEST**

import pandas as pd

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

dummy=pd.get\_dummies(df['Attrition'])

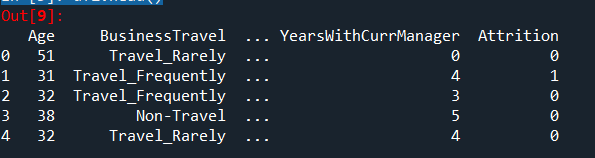
df2=pd.concat((df,dummy),axis=1)

df2=df2.drop(['Attrition'],axis=1)

df2=df2.drop(['No'],axis=1)

df2=df2.rename(columns={"Yes":"Attrition"})

df2.head()



**ATTRITION VS DISTANCE FROM HOME.**

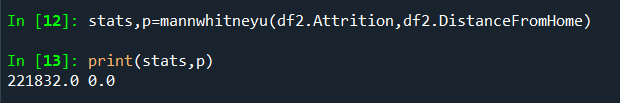
H0 = There is no significant difference between attrition yes and no for distance from home

HA= There is significant difference between attrition yes and no for distance from home

from scipy.stats import mannwhitneyu

stats,p=mannwhitneyu(df2.Attrition,df2.DistanceFromHome)

print(stats,p)



As the P value of 0.0 is < 0.05, the H0 is rejected and HA is accepted.

So there is difference in attrition and distance from home.

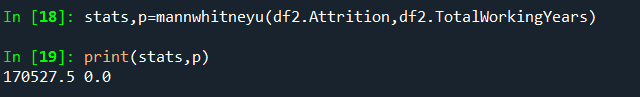
**ATTRITION VS TOTAL WORKING YEARS**

H0 = There is no significant difference between attrition yes and no for total working years.

HA= There is significant difference between attrition yes and no for total working years.

stats,p=mannwhitneyu(df2.Attrition,df2.TotalWorkingYears)

print(stats,p)



As the P value of 0.0 is < 0.05, the H0 is rejected and Ha is accepted.

So there is difference in attrition and total working years.

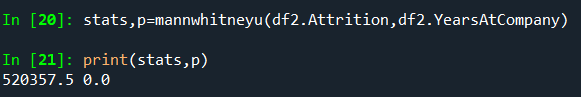
**ATTRITION VS YEARS AT COMPANY**

H0 = There is no significant difference between attrition yes and no for years at company.

HA= There is significant difference between attrition yes and no for years at company.

stats,p=mannwhitneyu(df2.Attrition,df2.YearsAtCompany)

print(stats,p)



As the P value of 0.0 is < 0.05, the H0 is rejected and Ha is accepted.

So there is difference in attrition and years at company.

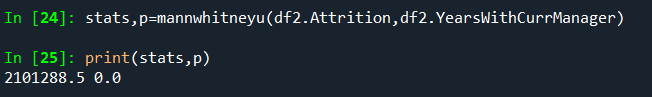
**ATTRITION VS YEARS WITH CURRENT MANAGER**

H0 = There is no significant difference between attrition yes and no for years with current manager.

HA= There is significant difference between attrition yes and no for years with current manager.

stats,p=mannwhitneyu(df2.Attrition,df2.YearsWithCurrManager)

print(stats,p)



As the P value of 0.0 is < 0.05, the H0 is rejected and HA is accepted.

So there is difference in attrition and years with current manager.

**CORRELATION BETWEEN 2 VARIABLES**

import pandas as pd

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

CORRELATION BETWEEN ATTRITION AND AGE

from scipy.stats import pearsonr

stats,p=pearsonr(df2.Attrition,df2.Age)

print(stats,p)

if (stats==0):

print("NO CORRELATION")

elif(stats<0):

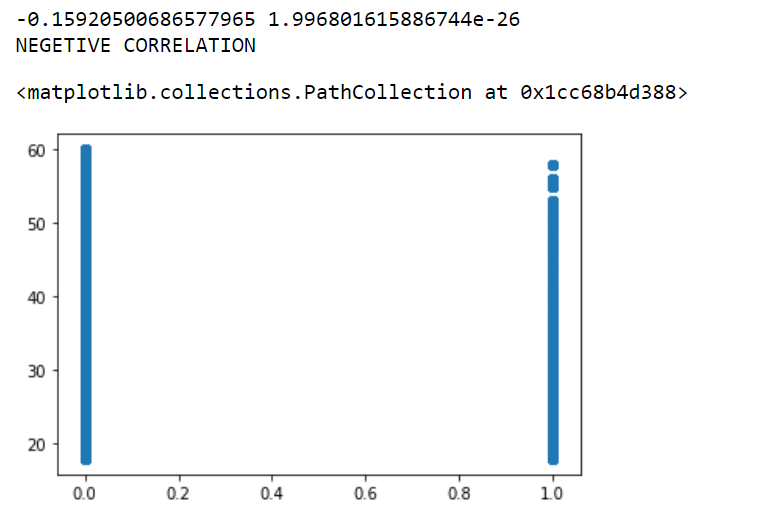
print("NEGETIVE CORRELATION")

else:

print("POSITIVE CORRELATION")

import matplotlib.pyplot as plt

plt.scatter(df2.Attrition,df2.Age)



It is low negative correlation.

H0 = There is no significant difference between attrition and age.

HA = There is significant difference between attrition and age.

As p value less than .05 so null hypothesis is rejected , so there is significant difference between attrition and age.

CORRELATION BETWEEN ATTRITION AND YEARS AT COMPANY

stats,p=pearsonr(df2.Attrition,df2.YearsAtCompany)

print(stats,p)

if (stats==0):

print("NO CORRELATION")

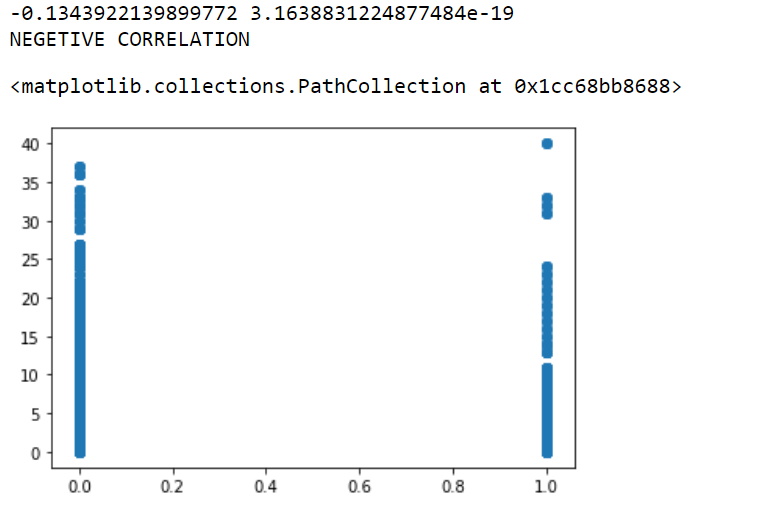
elif(stats<0):

print("NEGETIVE CORRELATION")

else:

print("POSITIVE CORRELATION")

plt.scatter(df2.Attrition,df2.YearsAtCompany)



It is low negative correlation.

H0 = There is no significant difference between attrition and years at company.

HA = There is significant difference between attrition and years at company.

As p value less than .05 so null hypothesis is rejected . so there is significant difference between attrition and years at company.

CORRELATION BETWEEN ATTRITION AND YEARS SINCE LAST PROMOTION

stats,p=pearsonr(df2.Attrition,df2.YearsSinceLastPromotion)

print(stats,p)

if (stats==0):

print("NO CORRELATION")

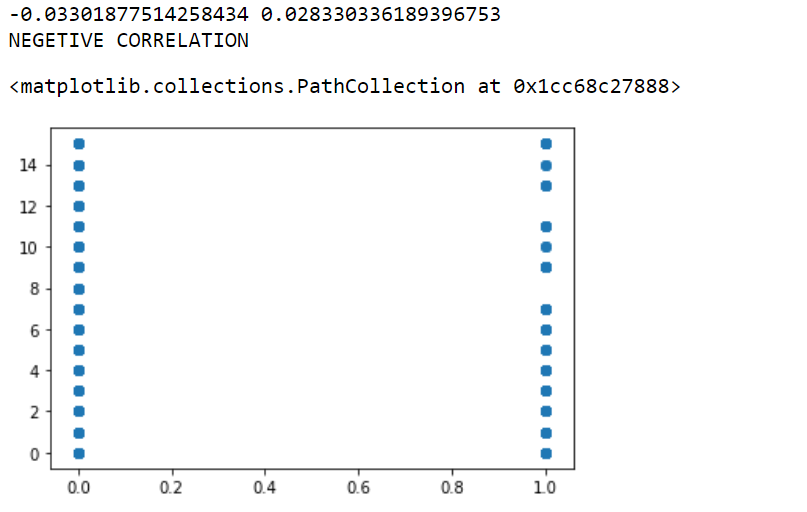
elif(stats<0):

print("NEGETIVE CORRELATION")

else:

print("POSITIVE CORRELATION")

plt.scatter(df2.Attrition,df2.YearsSinceLastPromotion)



It is low negative correlation.

H0 = There is no significant difference between attrition and years since last promotion.

HA = There is significant difference between attrition and years since last promotion.

As p value less than .05 so null hypothesis is rejected , so there is significant difference between attrition and years since last promotion.

CORRELATION BETWEEN ATTRITION AND TRAINING TIMES LAST YEAR

stats,p=pearsonr(df2.Attrition,df2.TrainingTimesLastYear)

print(stats,p)

if (stats==0):

print("NO CORRELATION")

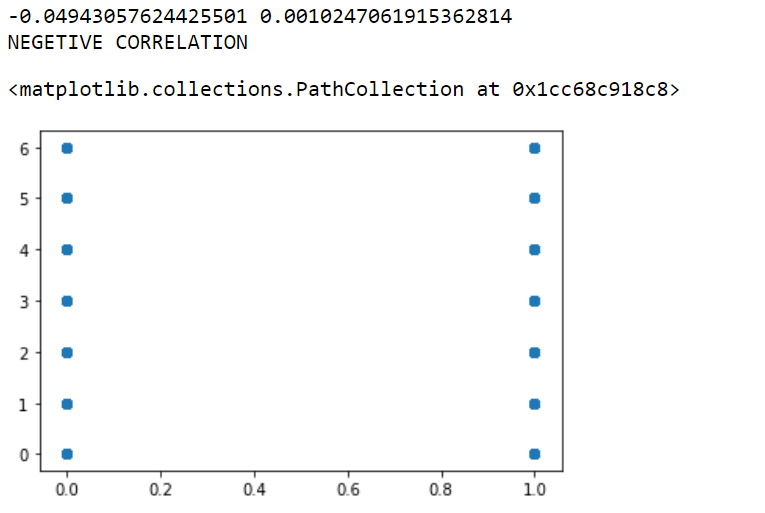
elif(stats<0):

print("NEGETIVE CORRELATION")

else:

print("POSITIVE CORRELATION")

plt.scatter(df2.Attrition,df2.TrainingTimesLastYear)

It is low negative correlation.

H0 = There is no significant difference between attrition and training times last year.

HA = There is significant difference between attrition and training times last year.

As p value less than .05 so null hypothesis is rejected , so there is significant difference between attrition and training times last year .

CORRELATION BETWEEN ATTRITION AND STOCK OPTION LEVEL

stats,p=pearsonr(df2.Attrition,df2.StockOptionLevel)

print(stats,p)

if (stats==0):

print("NO CORRELATION")

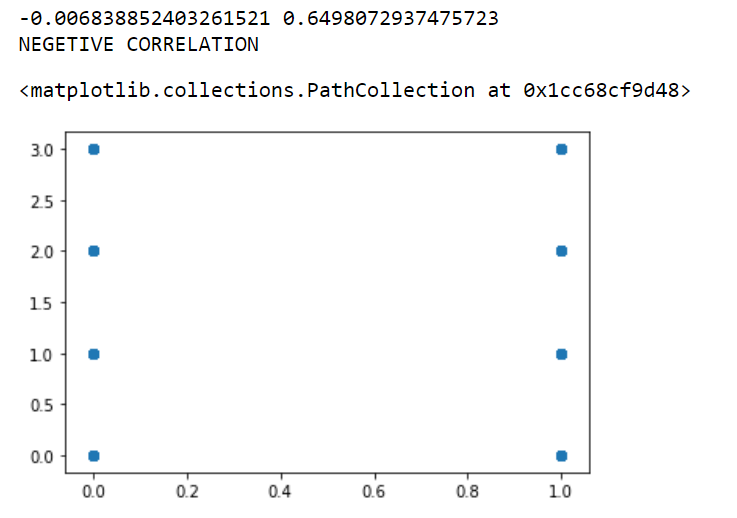
elif(stats<0):

print("NEGETIVE CORRELATION")

else:

print("POSITIVE CORRELATION")

plt.scatter(df2.Attrition,df2.StockOptionLevel)



It is low negative correlation.

H0 = There is no significant difference between attrition and stock option level.

HA = There is significant difference between attrition and stock option level.

As p value more than .05 so null hypothesis is accepted , so there is no significant difference between attrition and stock option level.

CORRELATION BETWEEN ATTRITION AND PERCENTAGE SALARY HIKE

stats,p=pearsonr(df2.Attrition,df2.PercentSalaryHike)

print(stats,p)

if (stats==0):

print("NO CORRELATION")

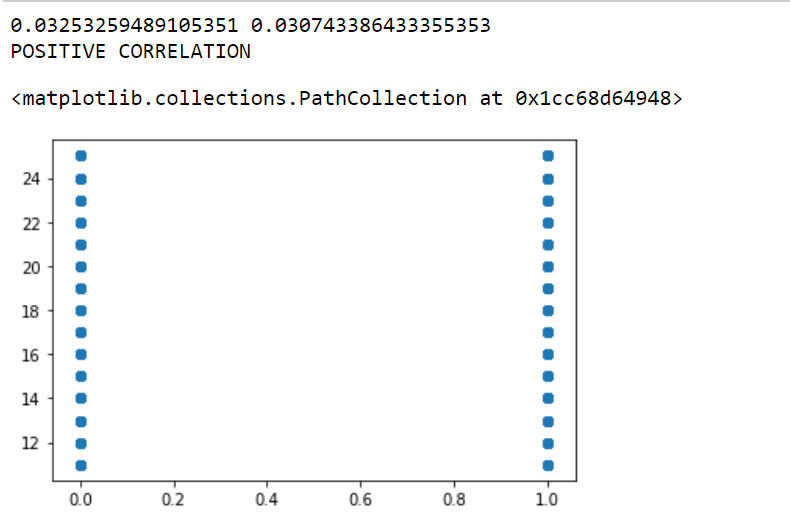
elif(stats<0):

print("NEGETIVE CORRELATION")

else:

print("POSITIVE CORRELATION")

plt.scatter(df2.Attrition,df2.PercentSalaryHike)



It is low positive correlation.

H0 = There is no significant difference between attrition and percent salary hike.

HA = There is significant difference between attrition and percent salary hike.

As p value less than .05 so null hypothesis is rejected , so there there is significant difference between attrition and percent salary hike.

CORRELATION BETWEEN ATTRITION AND JOB LEVEL

stats,p=pearsonr(df2.Attrition,df2.JobLevel)

print(stats,p)

if (stats==0):

print("NO CORRELATION")

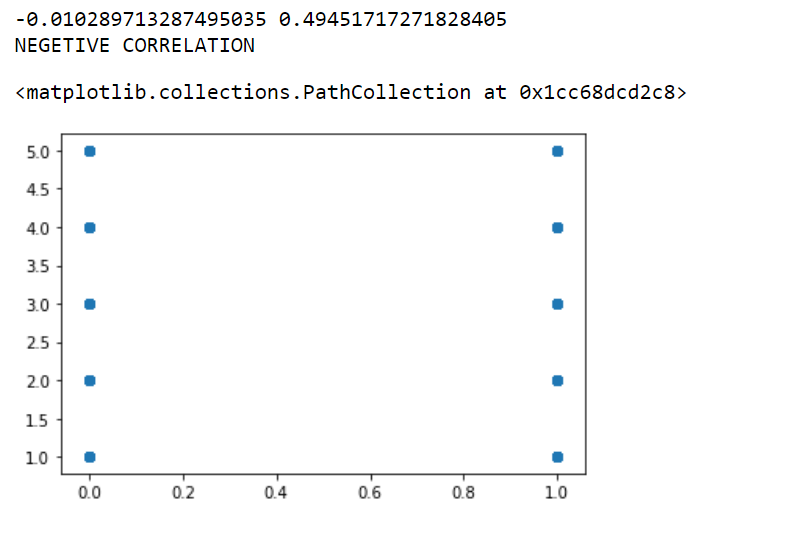
elif(stats<0):

print("NEGETIVE CORRELATION")

else:

print("POSITIVE CORRELATION")

plt.scatter(df2.Attrition,df2.JobLevel)



It is low negative correlation.

H0 = There is no significant difference between attrition and job level.

HA = There is significant difference between attrition and job level.

As p value more than .05 so null hypothesis is accepted , so there there is no significant difference between attrition and job level.

CORRELATION BETWEEN ATTRITIONA ND DISTANCE FROM HOME

stats,p=pearsonr(df2.Attrition,df2.DistanceFromHome)

print(stats,p)

if (stats==0):

print("NO CORRELATION")

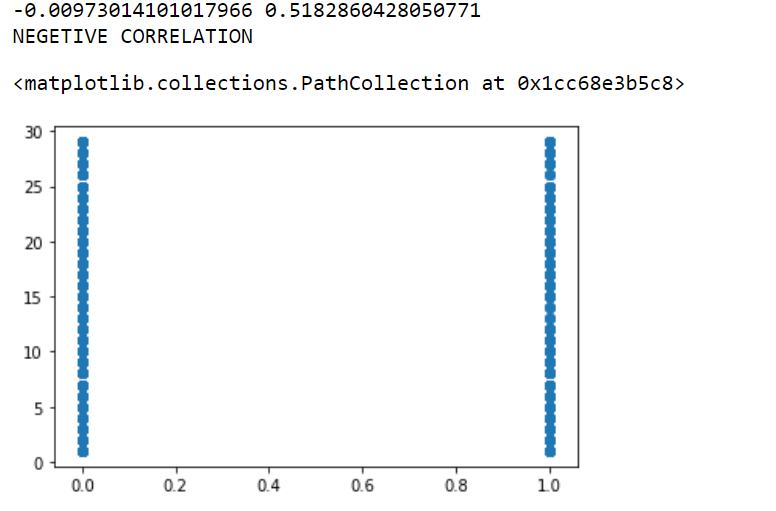
elif(stats<0):

print("NEGETIVE CORRELATION")

else:

print("POSITIVE CORRELATION")

plt.scatter(df2.Attrition,df2.DistanceFromHome)



It is low negative correlation.

H0 = There is no significant difference between attrition and distance from home.

HA = There is significant difference between attrition and distance from home.

As p value more than .05 so null hypothesis is accepted , so there is no significant difference between attrition and distance from home.

CORRELATION BETWEEN ATTRITION AND EDUCATION

stats,p=pearsonr(df2.Attrition,df2.Education)

print(stats,p)

if (stats==0):

print("NO CORRELATION")

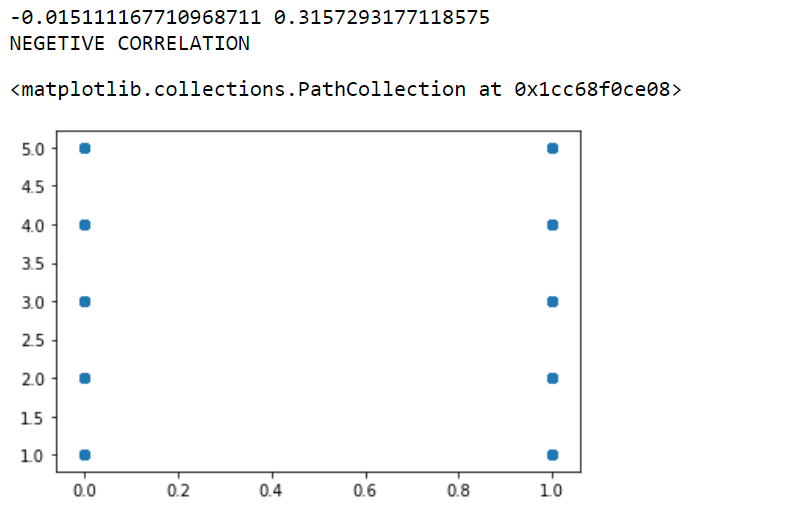
elif(stats<0):

print("NEGETIVE CORRELATION")

else:

print("POSITIVE CORRELATION")

plt.scatter(df2.Attrition,df2.Education)



It is low negative correlation.

H0 = There is no significant difference between attrition and education.

HA = There is significant difference between attrition and education.

As p value more than .05 so null hypothesis is accepted , so there is no significant difference between attrition and education.

CORRELATION BETWEEN ATTRTION AND YEARS WITH CURRENT MANAGER

stats,p=pearsonr(df2.Attrition,df2.YearsWithCurrManager)

print(stats,p)

if (stats==0):

print("NO CORRELATION")

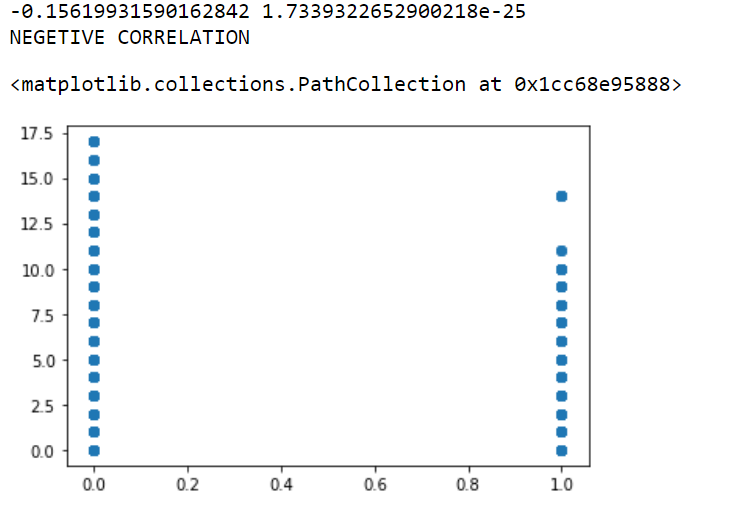
elif(stats<0):

print("NEGETIVE CORRELATION")

else:

print("POSITIVE CORRELATION")

plt.scatter(df2.Attrition,df2.YearsWithCurrManager)



It is low negative correlation.

H0 = There is no significant difference between attrition and years with current manager.

HA = There is significant difference between attrition and years with current manager.

As p value less than .05 so null hypothesis is rejected , so there is significant difference between attrition and years with current manager.