ATTRITION ASSIGNMENT

STEP 1 : LAUNCHING

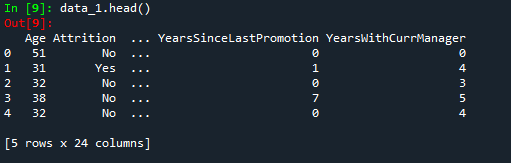
import pandas as pd

import numpy as np

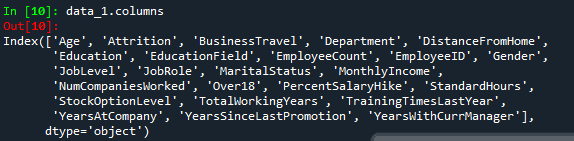
import matplotlib.pyplot as mp

data\_1=pd.read\_csv('general\_data.csv')

data\_1.head()

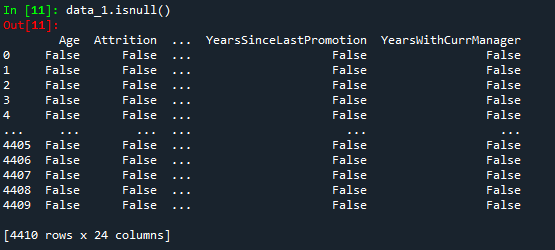


data\_1.columns

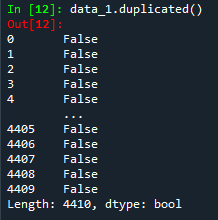


STEP 2:DATA TREATMENT

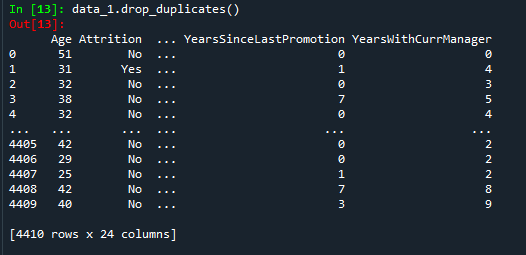
data\_1.isnull()



data\_1.duplicated()

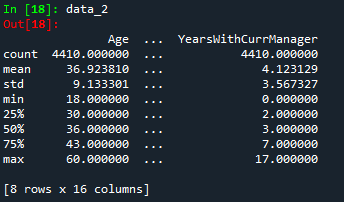


data\_1.drop\_duplicates()

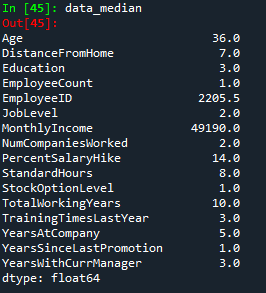


STEP 3: UNIVARIATE ANALYSIS

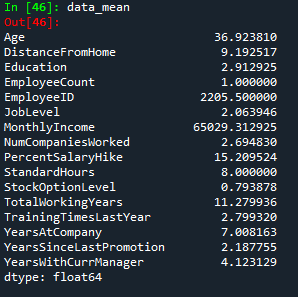
data\_2=data\_1[['Age', 'Attrition', 'BusinessTravel', 'Department', 'DistanceFromHome', 'Education', 'EducationField', 'EmployeeCount', 'EmployeeID', 'Gender', 'JobLevel', 'JobRole', 'MaritalStatus', 'MonthlyIncome', 'NumCompaniesWorked', 'Over18', 'PercentSalaryHike', 'StandardHours', 'StockOptionLevel', 'TotalWorkingYears', 'YearsAtCompany', 'TrainingTimesLastYear','YearsSinceLastPromotion', 'YearsWithCurrManager']].describe()



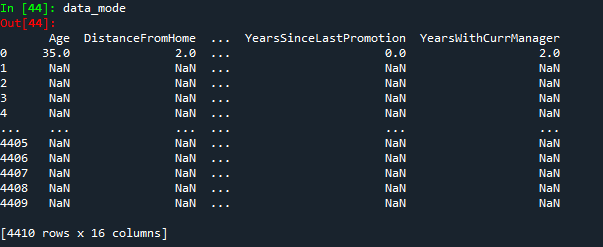
data\_median=data\_1[['Age', 'DistanceFromHome', 'Education', 'EmployeeCount', 'EmployeeID', 'JobLevel', 'MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike', 'StandardHours', 'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].median()



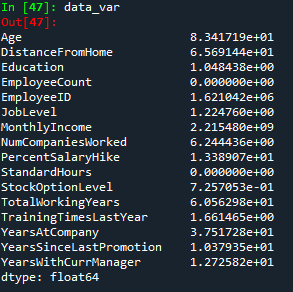
data\_mean=data\_1[['Age', 'DistanceFromHome', 'Education', 'EmployeeCount', 'EmployeeID','JobLevel', 'MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike', 'StandardHours', 'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManager']].mean()



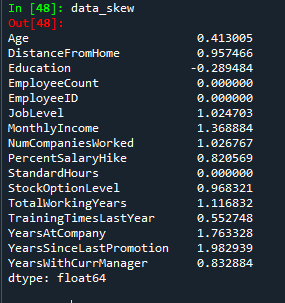
data\_mode=data\_1[['Age', 'DistanceFromHome', 'Education', 'EmployeeCount', 'EmployeeID','JobLevel', 'MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','StandardHours', 'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManager']].mode()



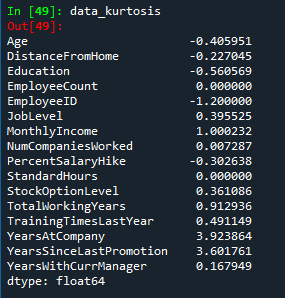
data\_var=data\_1[['Age', 'DistanceFromHome', 'Education', 'EmployeeCount', 'EmployeeID', 'JobLevel', 'MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','StandardHours', 'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManager']].var()



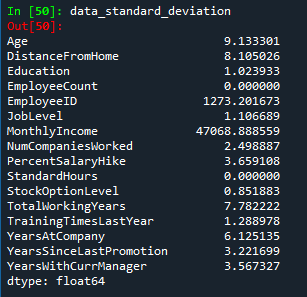
data\_skew=data\_1[['Age', 'DistanceFromHome', 'Education', 'EmployeeCount', 'EmployeeID', 'JobLevel', 'MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike', 'StandardHours', 'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManager']].skew()



data\_kurtosis=data\_1[['Age', 'DistanceFromHome', 'Education', 'EmployeeCount', 'EmployeeID', 'JobLevel', 'MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike', 'StandardHours', 'StockOptionLevel', 'TotalWorkingYears','TrainingTimesLastYear', 'YearsAtCompany', 'YearsSinceLastPromotion','YearsWithCurrManager']].kurt()



data\_standard\_deviation=data\_1[['Age', 'DistanceFromHome', 'Education', 'EmployeeCount', 'EmployeeID','JobLevel', 'MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike', 'StandardHours', 'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManager']].std()



**Inference from the analysis:**

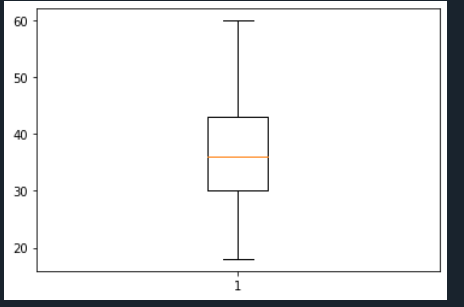
* All the above variables show positive skewness; while Age and Mean\_distance\_from\_home are leptokurtic and all other variables are platykurtic.
* The Mean\_Monthly\_Income’s IQR is at 54K suggesting company wide attrition across all income bands
* Mean age forms a near normal distribution with 13 years of IQR

**OUTLIERS:**

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

box\_plot=data\_1.Age

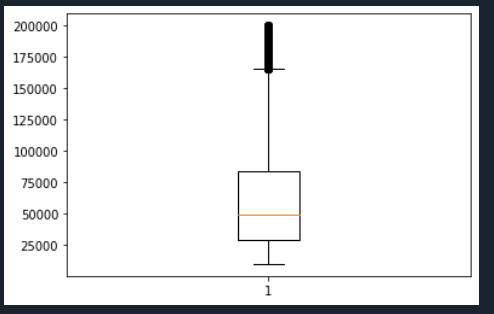
mp.boxplot(box\_plot)



Age is normally distributed without any outliers

box\_plot=data\_1.MonthlyIncome

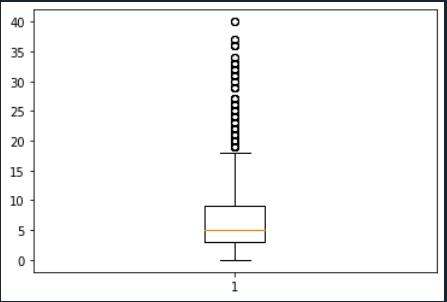
mp.boxplot(box\_plot)



Monthly Income is Right skewed with several outliers

box\_plot=data\_1.YearsAtCompany

mp.boxplot(box\_plot)



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

**Statistical tests**

**Mann Whitney test**

**Attrition Vs Monthly Income**

import pandas as pd

data=pd.read\_csv('file:///C:/Users/Admin/general\_data.csv')

data.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

[5 rows x 24 columns]

data.columns

Out[4]:

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')

data.Attrition=data.Attrition.map({"Yes":1,"No":0})

data.Attrition

Out[6]:

0 0

1 1

2 0

3 0

4 0

..

4405 0

4406 0

4407 0

4408 0

4409 0

Name: Attrition, Length: 4410, dtype: int64

from scipy.stats import mannwhitneyu

stats,p=mannwhitneyu(data.Attrition,data.MonthlyIncome)

print(stats,p)

0.0 0.0

H0: There is no significant differences in the Monthly income between attrition (Y) and attirition (N)

Ha: There is significant differences in the Monthly income between attrition (Y) and attirition (N)

The p valve is 0.0 which is less than 0.05 we reject the H0 hypothesis and accept the alternative hypothesis.

**Attrition Vs Distance From Home**

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

print(stats,p)

221832.0 0.0

H0: There is no significant differences in the Distance From Home between attrition (Y) and attirition (N)

Ha: There is significant differences in the Distance From Home between attrition (Y) and attirition (N)

The p valve is 0.0 which is less than 0.05 we reject the H0 hypothesis and accept the alternative hypothesis.

**Attrition Vs Total Working Years**

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

print(stats,p)

170527.5 0.0

H0: There is no significant differences in the Total Working Years between attrition (Y) and attirition (N)

Ha: There is significant differences in the Total Working Years between attrition (Y) and attirition (N)

The p valve is 0.0 which is less than 0.05 we reject the H0 hypothesis and accept the alternative hypothesis.

**Attrition Vs Years at company**

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

print(stats,p)

520357.5 0.0

H0: There is no significant differences in the Years At Company between attrition (Y) and attirition (N)

Ha: There is significant differences in the Years At Company between attrition (Y) and attirition (N)

The p valve is 0.0 which is less than 0.05 we reject the H0 hypothesis and accept the alternative hypothesis.

**Attrition Vs YearsWithCurrentManager**

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

print(stats,p)

2101288.5 0.0

H0: There is no significant differences in the Years With Current Manager between attrition (Y) and attirition (N)

Ha: There is significant differences in the Years With Current Manager between attrition (Y) and attirition (N)

The p valve is 0.0 which is less than 0.05 we reject the H0 hypothesis and accept the alternative hypothesis.

**T Test**

**Attrition Vs Distance From Home**

from scipy.stats import ttest\_ind

stats,p=ttest\_ind(data.Attrition,data.DistanceFromHome)

print(stats,p)

-73.92105563691779 0.0

H0: There is no significant differences in the Distance From Home between attrition (Y) and attirition (N)

Ha: There is significant differences in the Distance From Home between attrition (Y) and attirition (N)

The p valve is 0.0 which is less than 0.05 we reject the H0 hypothesis and accept the alternative hypothesis.

**Attrition Vs Monthly Income**

stats,p=ttest\_ind(data.Attrition,data.MonthlyIncome)

print(stats,p)

-91.74733118564392 0.0

H0: There is no significant differences in the Monthly Income between attrition (Y) and attirition (N)

Ha: There is significant differences in the Monthly Income between attrition (Y) and attirition (N)

The p valve is 0.0 which is less than 0.05 we reject the H0 hypothesis and accept the alternative hypothesis.

**Attrition Vs Years at company**

stats,p=ttest\_ind(data.Attrition,data.YearsAtCompany)

print(stats,p)

-74.10006092710509 0.0

H0: There is no significant differences in the Years At Company between attrition (Y) and attirition (N)

Ha: There is significant differences in the Years At Company between attrition (Y) and attirition (N)

The p valve is 0.0 which is less than 0.05 we reject the H0 hypothesis and accept the alternative hypothesis.

**Attrition Vs YearsWithCurrentManager**

stats,p=ttest\_ind(data.Attrition,data.YearsWithCurrManager)

print(stats,p)

-73.36426551326637 0.0

H0: There is no significant differences in the Years With Current Manager between attrition (Y) and attirition (N)

Ha: There is significant differences in the Years With Current Manager between attrition (Y) and attirition (N)

The p valve is 0.0 which is less than 0.05 we reject the H0 hypothesis and accept the alternative hypothesis.

**CORRELATION**

import pandas as pd

import numpy as np

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

df.head()

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

dummy.head()

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

df2.head()

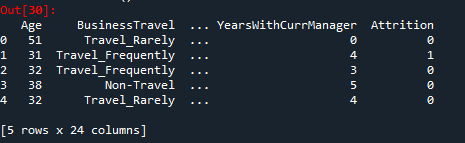
df2=df2.drop(['Attrition'],axis=1)

df2.head()

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

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

df2.head()



#correlation between Attrition and distance from home

from scipy.stats import pearsonr

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")

-0.009730141010179438 0.5182860428049617

NEGETIVE CORRELATION

#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")

-0.13439221398997386 3.163883122493571e-19

NEGETIVE CORRELATION

#correlation between Attrition and Total working years

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

print(stats, p)

if (stats==0):

print("NO CORRELATION")

elif(stats<0):

print("NEGETIVE CORRELATION")

else:

print("POSITIVE CORRELATION")

-0.17011136355964646 5.4731597518148054e-30

NEGETIVE CORRELATION

#correlation between Attrition and MonthlyIncome

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

print(stats, p)

if (stats==0):

print("NO CORRELATION")

elif(stats<0):

print("NEGETIVE CORRELATION")

else:

print("POSITIVE CORRELATION")

-0.031176281698114025 0.0384274849060192

NEGETIVE CORRELATION

#correlation between Attrition and Yeras with currmanager

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")

-0.15619931590162422 1.7339322652951965e-25

NEGETIVE CORRELATION

**INFERENCE :**

Attrition and DistanceFromHome:

As r = -0.009, there’s low negative correlation between Attrition and DistanceFromHome

As the P value of 0.518 is > 0.05, we are accepting H0 and hence there’s no significant correlation between Attrition and DistanceFromHome

Attrition and MonthlyIncome:

As r = -0.031, there’s low negative correlation between Attrition and MonthlyIncome

As the P value of 0.038 is < 0.05, we are accepting Ha and hence there’s significant correlation between Attrition and MonthlyIncome

Attrition and TotalWorkingYears:

As r = -0.17, there’s low negative correlation between Attrition and TotalWorkingYears

As the P value is < 0.05, we are accepting Ha and hence there’s significant correlation between Attrition and TotalWorkingYears

Attrition & YearsAtCompany:

As r = -0.1343, there’s low negative correlation between Attrition and YearsAtCompany

As the P value is < 0.05, we are accepting Ha and hence there’s significant correlation between Attrition and YearsAtCompany

Attrition & YearsWithCurrManager:

As r = -0.1561, there’s low negative correlation between Attrition and YearsWithCurrManager

As the P value is < 0.05, we are accepting Ha and hence there’s significant correlation between Attrition and YearsWithCurrManager