**ATTRITION ASSIGNMENT**

STEP 1=LAUNCHING:

import pandas as pd

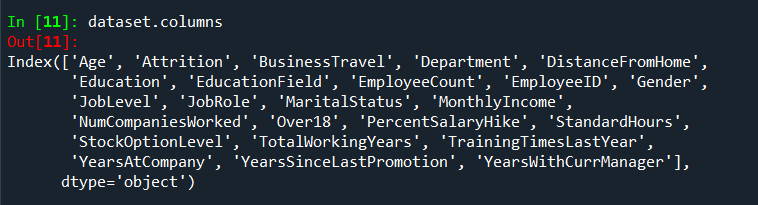
import numpy as n

import matplotlib.pyplot as pl

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

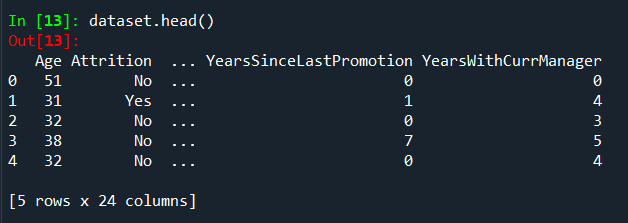
-->To find columns names

data.columns



-->To find the data of first 5 rows

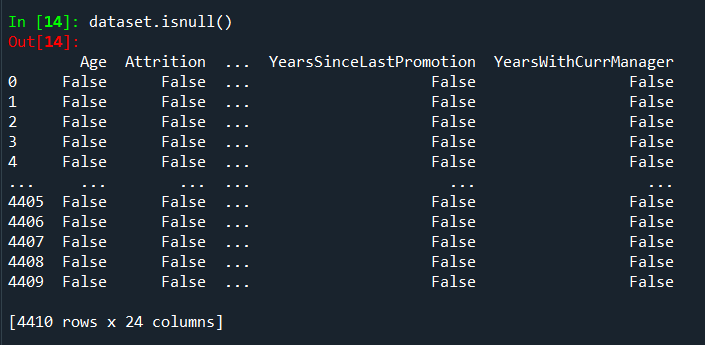
data.head()



STEP2=DATA TREATMENT:

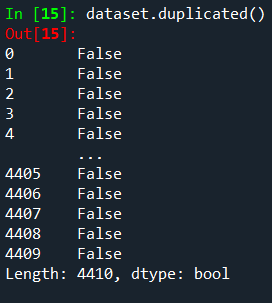
-->To find out null values in the table.

data.isnull()



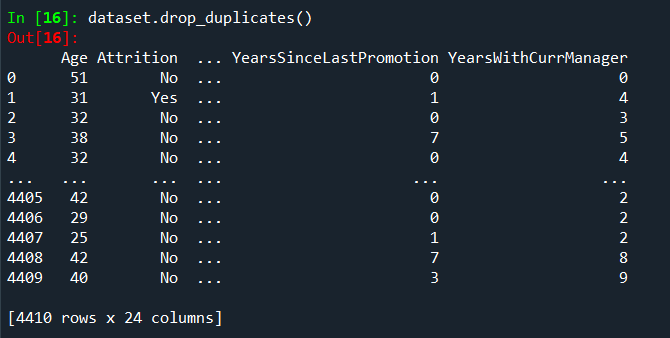
-->To find out duplicated values of table.

data.duplicated()



-->To drop all duplicated values of the table.

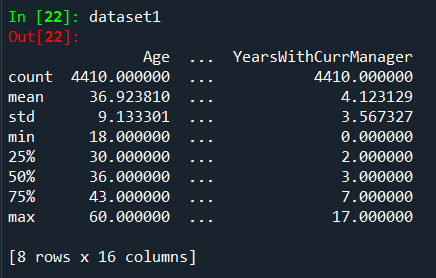
data.drop\_duplicates()



STEP3=UNIVARIATE ANALYSIS:

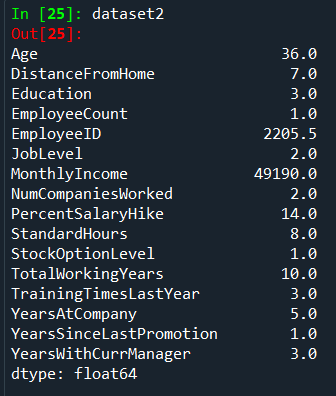
-->To describe the whole table.

dataset1=dataset[['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.

dataset2=datasetdataset[['Age', 'DistanceFromHome', 'Education', 'EmployeeCount', 'EmployeeID', 'JobLevel', 'MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike', 'StandardHours', 'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManager']].median()



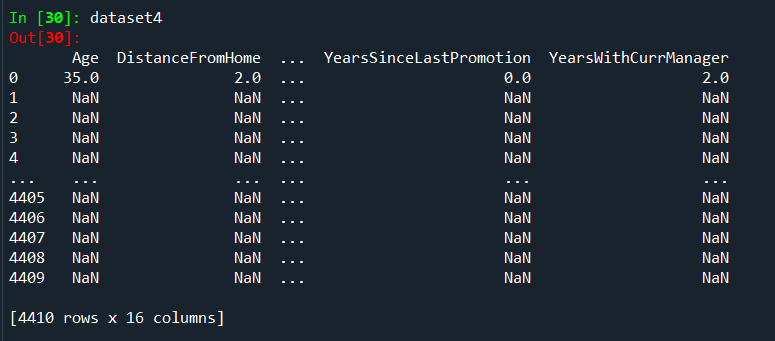
-->To find out mean of each cloumn

dataset3==datasetdataset[['Age', 'DistanceFromHome', 'Education', 'EmployeeCount', 'EmployeeID', 'JobLevel', 'MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike', 'StandardHours', 'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManager']].mean()



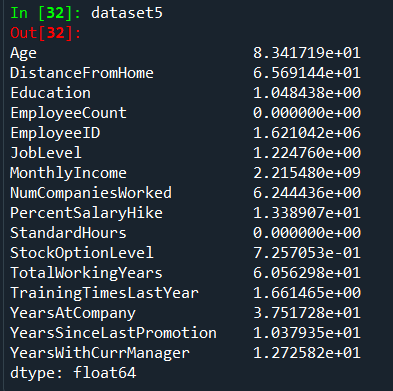
-->To find out mode.

dataset4=dataset[['Age', 'DistanceFromHome', 'Education', 'EmployeeCount', 'EmployeeID', 'JobLevel', 'MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike', 'StandardHours', 'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManager']].mode()



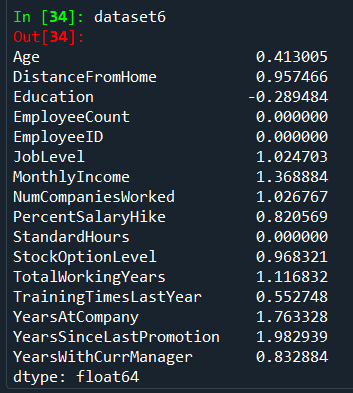
-->To find variance of each column.

dataset5=dataset[['Age', 'DistanceFromHome', 'Education', 'EmployeeCount', 'EmployeeID', 'JobLevel', 'MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike', 'StandardHours', 'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManager']].var()



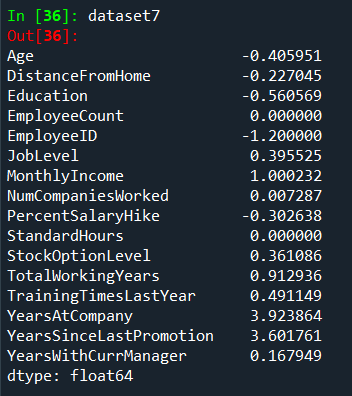
-->To find skewness.

dataset6=dataset[['Age', 'DistanceFromHome', 'Education', 'EmployeeCount', 'EmployeeID', 'JobLevel', 'MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike', 'StandardHours', 'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManager']].skew()



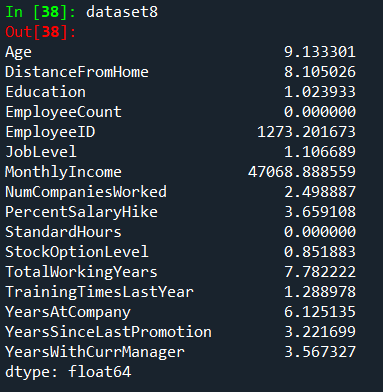
-->To find out kurtosis.

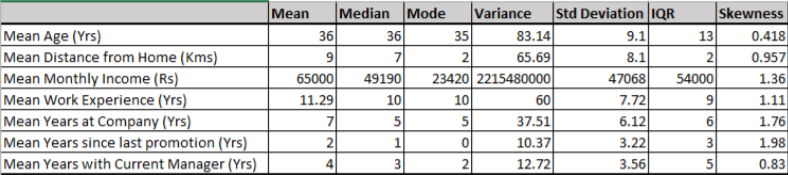
dataset7=dataset[['Age', 'DistanceFromHome', 'Education', 'EmployeeCount', 'EmployeeID', 'JobLevel', 'MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike', 'StandardHours', 'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManager']].kurt()



-->To find standard deviation.

dataset8=dataset[['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 & 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

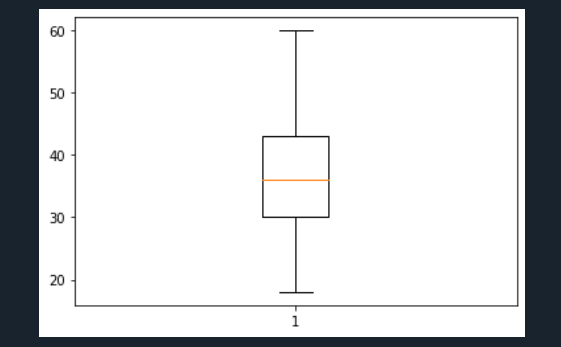
• Mean age forms a near normal distribution with 13 years of IQR

OUTLIERS:

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

box\_plot=data.Age

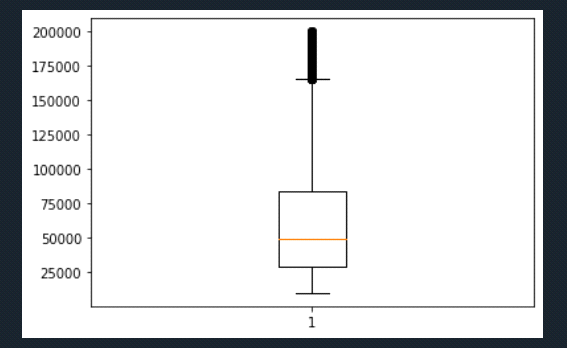
pl.boxplot(box\_plot)



Age is normally distributed without any outliers.

box\_plot=data.MonthlyIncome

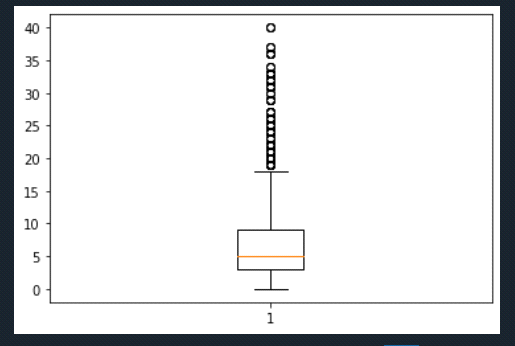
pl.boxplot(box\_plot)



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.

– ->**Statistical Tests (Mann-Whitney)**

import pandas as pd

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

dataset.columns

Out[3]:

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')

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

dataset.Attrition

Out[5]:

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

**Attrition Vs MonthlyIncome**

from scipy.stats import mannwhitneyu

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

print(stats,p)

1. 0.0

As the P value is a 0.0, which is < than 0.05, the H0 is rejected and Ha is accepted. H0: There is no significant differences in the monthlyincome between attrition (Y) and attirition (N) Ha: There is significant differences in the monthlyincome between attrition (Y) and attirition (N)

**Attrition Vs DistanceFromHome**

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

print(stats,p)

221832.0 0.0

As the P value of 0.0 is < 0.05, the H0 is rejected and Ha is accepted. H0: There is no significant differences in the DistanceFromHome between attrition (Y) and attirition (N) Ha: There is significant differences in the DistanceFromHome between attrition (Y) and attirition (N)

**Attrition Vs TotalWorkingYears**

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

print(stats,p)

170527.5 0.0

As the P value is again 0.0, which is < than 0.05, the H0 is rejected and ha is accepted. H0: There is no significant differences in the TotalWorkingYears between attrition (Y) and attirition (N) Ha: There is significant differences in the TotalWorkingYears between attrition (Y) and attirition (N)

**Attrition Vs YearsAtCompany**

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

print(stats,p)

520357.5 0.0

As the P value is again 0.0, which is < than 0.05, the H0 is rejected and ha is accepted. H0: There is no significant differences in the YearsAtCompany between attrition (Y) and attirition (N) Ha: There is significant differences in the YearsAtCompany between attrition (Y) and attirition (N)

**Attrition Vs YearsWithCurrManager**

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

print(stats,p)

2101288.5 0.0

As the P value is again 0.0, which is < than 0.05, the H0 is rejected and ha is accepted. H0: There is no significant differences in the YearsWithCurrManager between attrition (Y) and attirition (N) Ha: There is significant differences in the YearsWithCurrentManager between attrition (Y) and attirition (N)

**Statistical Tests (Separate T Test)**

**Attrition Vs DistanceFromHome**

from scipy.stats import ttest\_ind

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

print(stats,p)

-73.92105563691779 0.0

As the P value of 0.0 is < 0.05, the H0 is rejected and Ha is accepted. H0: There is no significant differences in the DistanceFromHome between attrition (Y) and attirition (N) Ha: There is significant differences in the DistanceFromHome between attrition (Y) and attirition (N)

**Attrition Vs MonthlyIncome**

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

print(stats,p)

-91.74733118564392 0.0

As the P value is a 0.0, which is < than 0.05, the H0 is rejected and Ha is accepted. H0: There is no significant differences in the monthlyincome between attrition (Y) and attirition (N) Ha: There is significant differences in the monthlyincome between attrition (Y) and attirition (N)

**Attrition Vs YearsAtCompany**

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

print(stats,p)

-74.10006092710509 0.0

As the P value is again 0.0, which is < than 0.05, the H0 is rejected and ha is accepted. 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)

**Attrition Vs YearsWithCurrentManager**

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

print(stats,p)

-73.36426551326637 0.0

As the P value is again 0.0, which is < than 0.05, the H0 is rejected and ha is accepted. 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)

**Unsupervised Learning –**

Correlation Analysis In order to find the interdependency of the variables DistanceFromHome, MonthlyIncome, TotalWorkingYears, YearsAtCompany, YearsWithCurrManager from that of Attrition, we executed the Correlation Analysis as follows.

**from** **scipy.stats** **import** pearsonr

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

print(stats,p)

-0.031176281698115017 0.03842748490600132

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

print(stats,p)

-0.009730141010179667 0.5182860428050771

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

print(stats,p)

-0.1343922139899772 3.1638831224877484e-19

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

print(stats,p)

-0.1561993159016286 1.7339322652900218e-25

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

print(stats, p)

-0.17011136355964646 5.4731597518148054e-30

**The inference of the above analysis are as follows:**

Attrition & 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 & DistanceFromHome Attrition & 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 & MonthlyIncome Attrition & 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 & 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 & 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 & YearsWithCurrManager.