***ASSIGNMENT***

***Step1 - Launching***

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

dataset1=pd.read\_excel(‘general\_data.xlsx', sheet\_name=0)

dataset1.head()

Out[41]:

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 18 columns]

dataset1.columns

Out[42]:

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

'Education', 'EducationField', 'Gender', 'JobRole', 'MaritalStatus',

'MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike',

'TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany',

'YearsSinceLastPromotion', 'YearsWithCurrManager'],

dtype='object')

***Step 2 - Data Treatment:***

dataset1.isnull()

Out[17]:

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

[4410 rows x 18 columns]

dataset1.duplicated()

Out[21]:

0 False

1 False

2 False

3 False

4 False

4405 True

4406 True

4407 True

4408 True

4409 False

Length: 4410, dtype: bool

dataset1.drop\_duplicates()

Out[23]:

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

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

3818 28 Yes ... 0 0

3910 41 No ... 1 2

4226 36 No ... 0 0

4395 40 No ... 4 7

4409 40 No ... 3 9

[1498 rows x 18 columns]

***Step 3 – Univariate Analysis:***

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

dataset3

Out[12]:

Age DistanceFromHome Education MonthlyIncome \

count 4410.000000 4410.000000 4410.000000 4410.000000

mean 36.923810 9.192517 2.912925 65029.312925

std 9.133301 8.105026 1.023933 47068.888559

min 18.000000 1.000000 1.000000 10090.000000

25% 30.000000 2.000000 2.000000 29110.000000

50% 36.000000 7.000000 3.000000 49190.000000

75% 43.000000 14.000000 4.000000 83800.000000

max 60.000000 29.000000 5.000000 199990.000000

NumCompaniesWorked PercentSalaryHike TotalWorkingYears \

count 4391.000000 4410.000000 4401.000000

mean 2.694830 15.209524 11.279936

std 2.498887 3.659108 7.782222

min 0.000000 11.000000 0.000000

25% 1.000000 12.000000 6.000000

50% 2.000000 14.000000 10.000000

75% 4.000000 18.000000 15.000000

max 9.000000 25.000000 40.000000

TrainingTimesLastYear YearsAtCompany YearsSinceLastPromotion \

count 4410.000000 4410.000000 4410.000000

mean 2.799320 7.008163 2.187755

std 1.288978 6.125135 3.221699

min 0.000000 0.000000 0.000000

25% 2.000000 3.000000 0.000000

50% 3.000000 5.000000 1.000000

75% 3.000000 9.000000 3.000000

max 6.000000 40.000000 15.000000

YearsWithCurrManager

count 4410.000000

mean 4.123129

std 3.567327

min 0.000000

25% 2.000000

50% 3.000000

75% 7.000000

max 17.000000

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

dataset3

Out[14]:

Age 36.0

DistanceFromHome 7.0

Education 3.0

MonthlyIncome 49190.0

NumCompaniesWorked 2.0

PercentSalaryHike 14.0

TotalWorkingYears 10.0

TrainingTimesLastYear 3.0

YearsAtCompany 5.0

YearsSinceLastPromotion 1.0

YearsWithCurrManager 3.0

dtype: float64

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

dataset3

Out[16]:

Age DistanceFromHome Education MonthlyIncome NumCompaniesWorked \

0 35 2 3 23420 1.0

PercentSalaryHike TotalWorkingYears TrainingTimesLastYear \

0 11 10.0 2

YearsAtCompany YearsSinceLastPromotion YearsWithCurrManager

0 5 0 2

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

dataset3

Out[18]:

Age 8.341719e+01

DistanceFromHome 6.569144e+01

Education 1.048438e+00

MonthlyIncome 2.215480e+09

NumCompaniesWorked 6.244436e+00

PercentSalaryHike 1.338907e+01

TotalWorkingYears 6.056298e+01

TrainingTimesLastYear 1.661465e+00

YearsAtCompany 3.751728e+01

YearsSinceLastPromotion 1.037935e+01

YearsWithCurrManager 1.272582e+01

dtype: float64

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

dataset3

Out[20]:

Age 0.413005

DistanceFromHome 0.957466

Education -0.289484

MonthlyIncome 1.368884

NumCompaniesWorked 1.026767

PercentSalaryHike 0.820569

TotalWorkingYears 1.116832

TrainingTimesLastYear 0.552748

YearsAtCompany 1.763328

YearsSinceLastPromotion 1.982939

YearsWithCurrManager 0.832884

dtype: float64

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

dataset3

Out[22]:

Age -0.405951

DistanceFromHome -0.227045

Education -0.560569

MonthlyIncome 1.000232

NumCompaniesWorked 0.007287

PercentSalaryHike -0.302638

TotalWorkingYears 0.912936

TrainingTimesLastYear 0.491149

YearsAtCompany 3.923864

YearsSinceLastPromotion 3.601761

YearsWithCurrManager 0.167949

dtype: float64

**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:**

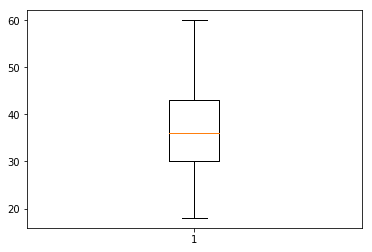
There’s no regression found while plotting Age, MonthlyIncome, TotalWorkingYears, YearsAtCompany, etc., on a scatter plot

box\_plot=dataset1.Age

plt.boxplot(box\_plot)

Out[24]:

<matplotlib.lines.Line2D at 0x20382d59160>]}

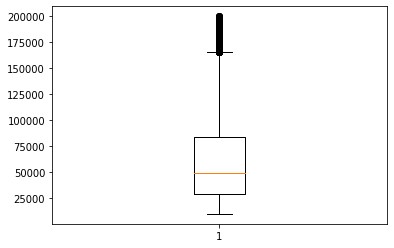


Age is normally distributed without any outliers

box\_plot=dataset1.MonthlyIncome

plt.boxplot(box\_plot)

Out[26]:

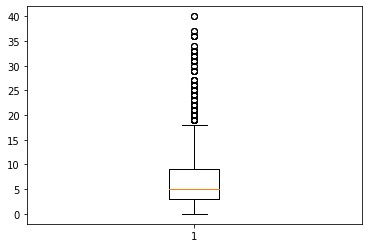


Monthly Income is Right skewed with several outliers

box\_plot=dataset1.YearsAtCompany

plt.boxplot(box\_plot)

Out[28]:



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