**Attrition Assignment**

import pandas as pd import numpy as np import matplotlib.pyplot as plt

dataset1== pd.read\_excel('E:\Rutuja\LockDown\AI and ML\Day 7-Notes and assignments\Assignment\general\_data.xlsx', sheet\_name=0)

dataset1.head()

Out[7]:

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]

dataset1.columns

Out[8]:

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')

dataset1

Out[9]:

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 2: Data Treatment**

dataset1.isnull

Out[10]:

<bound method DataFrame.isnull of 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]

dataset1.duplicated()

Out[11]:

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[12]:

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]

**Step3- Univariate Analysis**

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

dataset3

Out[15]:

Age ... YearsWithCurrManager

count 4410.000000 ... 4410.000000

mean 36.923810 ... 4.123129

std 9.133301 ... 3.567327

min 18.000000 ... 0.000000

25% 30.000000 ... 2.000000

50% 36.000000 ... 3.000000

75% 43.000000 ... 7.000000

max 60.000000 ... 17.000000

[8 rows x 11 columns]

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

dataset3

Out[17]:

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[19]:

Age DistanceFromHome ... YearsSinceLastPromotion YearsWithCurrManager

0 35 2 ... 0 2

[1 rows x 11 columns]

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

dataset3

Out[21]:

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[23]:

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

dataset3

Out[25]:

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 variable show positive skewness, while Age and Mean of distance\_from\_home are leptokurtic and all other variables are platykurtic.

The Mean of Monthly\_Income’s IQR is at 54K suggesting company wide attrition across all the 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

plot.boxplot(box\_plot)

Out[30]:

{'whiskers': [<matplotlib.lines.Line2D at 0x16522c62188>,

<matplotlib.lines.Line2D at 0x16522c6ce48>],

'caps': [<matplotlib.lines.Line2D at 0x16522c74e48>,

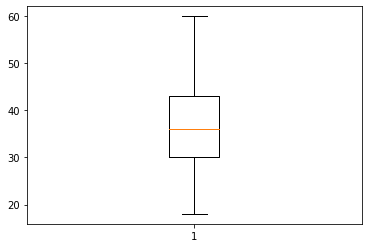
<matplotlib.lines.Line2D at 0x16522c74f48>],

'boxes': [<matplotlib.lines.Line2D at 0x16522c6cc88>],

'medians': [<matplotlib.lines.Line2D at 0x16522c74fc8>],

'fliers': [<matplotlib.lines.Line2D at 0x16522c79ec8>],

'means': []}



Age is normally distributed without any outliers

box\_plot = dataset1.MonthlyIncome

plot.boxplot(box\_plot)

Out[32]:

{'whiskers': [<matplotlib.lines.Line2D at 0x165211147c8>,

<matplotlib.lines.Line2D at 0x16521114d48>],

'caps': [<matplotlib.lines.Line2D at 0x16521146108>,

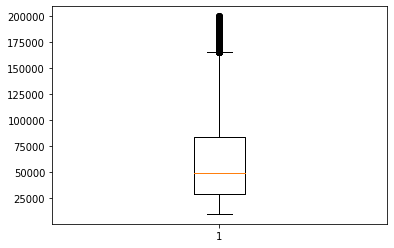
<matplotlib.lines.Line2D at 0x16521407188>],

'boxes': [<matplotlib.lines.Line2D at 0x16521128088>],

'medians': [<matplotlib.lines.Line2D at 0x16521255588>],

'fliers': [<matplotlib.lines.Line2D at 0x1652112e988>],

'means': []}



Monthly Income is Right skewed with several outliers.

plot.boxplot(box\_plot)

Out[35]:

{'whiskers': [<matplotlib.lines.Line2D at 0x1652113dc88>,

<matplotlib.lines.Line2D at 0x1652113d048>],

'caps': [<matplotlib.lines.Line2D at 0x16520d482c8>,

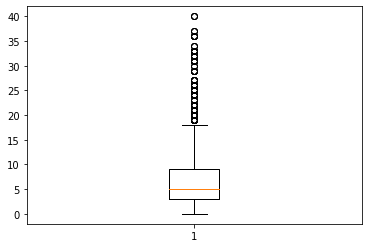
<matplotlib.lines.Line2D at 0x16520d48988>],

'boxes': [<matplotlib.lines.Line2D at 0x1652133ed08>],

'medians': [<matplotlib.lines.Line2D at 0x1652127bd88>],

'fliers': [<matplotlib.lines.Line2D at 0x1652110cc08>],

'means': []}



Years at Company is also Right skewed with several outliers observed.