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[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

[5 rows x 18 columns]

>>>dataset1.columns

Out[3]:

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[47]:

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[23]:

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[1]:

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

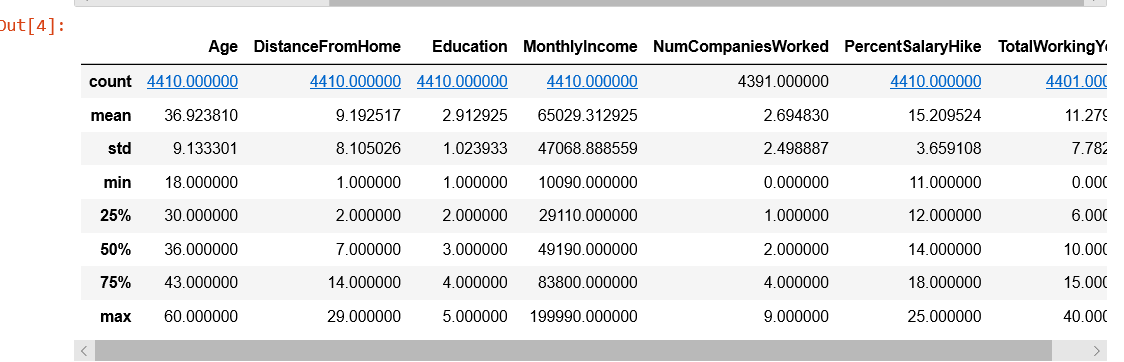
4409 40 No ... 3 9

[1498 rows x 18 columns]

**Step 3 – Univariate Analysis:**

>>>dataset3=dataset1[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears']].describe()

>>>dataset3



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

>>>dataset3

Out[2]:

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[69]:

Age 35

DistanceFromHome 2

Education 3

MonthlyIncome 23420

NumCompaniesWorked 1

PercentSalaryHike 11

TotalWorkingYears 10

TrainingTimesLastYear 2

YearsAtCompany 5.0

YearsSinceLastPromotion 0

YearsWithCurrManager 2

dtype: float64

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

>>>dataset3

output-1

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[5]: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

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:**

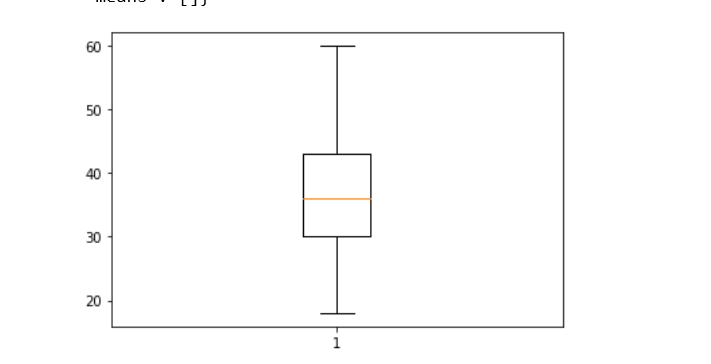
There’s no regression found while plotting Age, MonthlyIncome

, TotalWorkingYears, YearsAtCompany, etc., on a scatter plot

box\_plot=dataset1.Age plt.boxplot(box\_plot)

Out[23]:

**{'whiskers': [<matplotlib.lines.Line2D at 0x1ed72394608>, <matplotlib.lines.Line2D at 0x1ed723ae708>], 'caps': [<matplotlib.lines.Line2D at 0x1ed723aedc8>, <matplotlib.lines.Line2D at 0x1ed723b2f48>], 'boxes': [<matplotlib.lines.Line2D at 0x1ed723a3288>], 'medians': [<matplotlib.lines.Line2D at 0x1ed723b6ec8>], 'fliers': [<matplotlib.lines.Line2D at 0x1ed723b9c48>], 'means': []}**

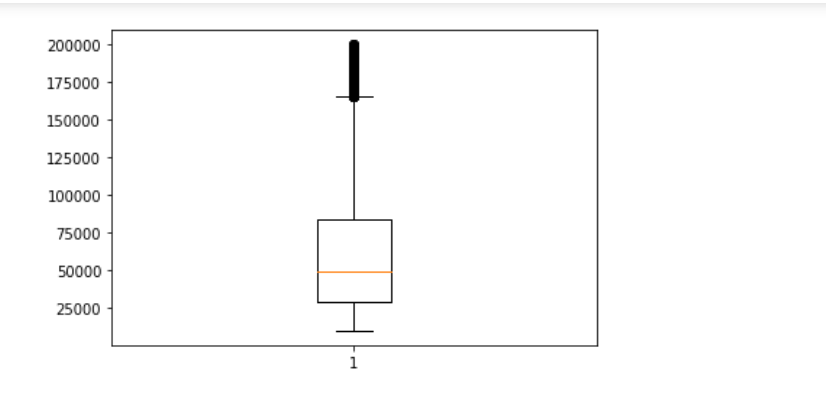
******

\*\*\*Age is normally distributed without any outliers

>>>box\_plot=dataset1.MonthlyIncome

>>>plt.boxplot(box\_plot)

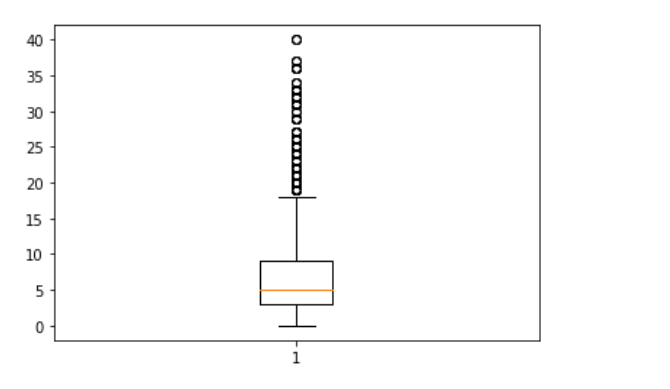
{'whiskers': [<matplotlib.lines.Line2D at 0x1ed7243c388>, <matplotlib.lines.Line2D at 0x1ed7243cb88>], 'caps': [<matplotlib.lines.Line2D at 0x1ed7243acc8>, <matplotlib.lines.Line2D at 0x1ed724365c8>], 'boxes': [<matplotlib.lines.Line2D at 0x1ed7243cec8>], 'medians': [<matplotlib.lines.Line2D at 0x1ed7242e4c8>], 'fliers': [<matplotlib.lines.Line2D at 0x1ed7242e5c8>], 'means': []}



\*\*Monthly Income is Right skewed with several outliers

>>>box\_plot=dataset1.YearsAtCompany >>>plt.boxplot(box\_plot)

Output:  
{'whiskers': [<matplotlib.lines.Line2D at 0x1ed6c21da08>, <matplotlib.lines.Line2D at 0x1ed6c21df88>], 'caps': [<matplotlib.lines.Line2D at 0x1ed6c21db88>, <matplotlib.lines.Line2D at 0x1ed6c21fa88>], 'boxes': [<matplotlib.lines.Line2D at 0x1ed6c21d308>], 'medians': [<matplotlib.lines.Line2D at 0x1ed6c21ff88>], 'fliers': [<matplotlib.lines.Line2D at 0x1ed6c21fc08>], 'means': []}



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

**Step 5 – Statistical Tests (Mann-Whitney)**

**Attrition Vs Distance from Home**

import pandas as pd

dataset=pd.read\_excel('EmployeeAttrition.xlsx', sheet\_name=1)

dataset.head()

Out[3]:

DistanceFromHome\_Yes ... YearsWithCurrManager\_No

0 0 ... 0

1 10 ... 0

2 0 ... 3

3 0 ... 5

4 0 ... 4

[5 rows x 10 columns]

dataset.columns

Out[4]:

Index(['DistanceFromHome\_Yes', 'DistanceFromHome\_No', 'MonthlyIncome\_Yes',

'MonthlyIncome\_No', 'TotalWorkingYears\_Yes', 'TotalWorkingYears\_No',

'YearsAtCompany\_Yes', 'YearsAtCompany\_No', 'YearsWithCurrManager\_Yes',

'YearsWithCurrManager\_No'],

dtype='object')

from scipy.stats import mannwhitneyu

a1=dataset.DistanceFromHome\_Yes

a2=dataset.DistanceFromHome\_No

stat, p=mannwhitneyu(a1,a2)

print(stat, p)

3132625.5 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 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)

**Attrition Vs Income**

a1=dataset.MonthlyIncome\_Yes

a2=dataset.MonthlyIncome\_No

stat, p=mannwhitneyu(a1,a2)

print(stat, p)

3085416.0 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 income between attrition (Y) and attirition (N)

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

**Attrition Vs Total Working Years**

a1=dataset.TotalWorkingYears\_Yes

a2=dataset.TotalWorkingYears\_No

stat, p=mannwhitneyu(a1,a2)

print(stat, p)

2760982.0 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 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)

**Attrition Vs Years at company**

a1=dataset.YearsAtCompany\_Yes

a2=dataset.YearsAtCompany\_No

stat, p=mannwhitneyu(a1,a2)

print(stat, p)

2882047.5 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 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**

a1=dataset.YearsWithCurrManager\_Yes

a2=dataset.YearsWithCurrManager\_No

stat, p=mannwhitneyu(a1,a2)

print(stat, p)

3674749.5 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 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)

**Step 6 – Statistical Tests (Separate T Test)**

**Attrition Vs Distance From Home**

from scipy.stats import ttest\_ind

dataset.columns

Out[49]:

Index(['DistanceFromHome\_Yes', 'DistanceFromHome\_No', 'MonthlyIncome\_Yes',

'MonthlyIncome\_No', 'TotalWorkingYears\_Yes', 'TotalWorkingYears\_No',

'YearsAtCompany\_Yes', 'YearsAtCompany\_No', 'YearsWithCurrManager\_Yes',

'YearsWithCurrManager\_No'],

dtype='object')

z1=dataset.DistanceFromHome\_Yes

z2=dataset.DistanceFromHome\_No

stat, p=ttest\_ind(z2,z1)

print(stat, p)

44.45445917636664 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 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)

**Attrition Vs Income**

z1=dataset.MonthlyIncome\_Yes

z2=dataset.MonthlyIncome\_No

stat, p=ttest\_ind(z2, z1)

print(stat, p)

52.09279408504947 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 Monthly Income between attrition (Y) and attirition (N)

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

**Attrition Vs Yeats At Company**

z1=dataset.YearsAtCompany\_Yes

z2=dataset.YearsAtCompany\_No

stat, p=ttest\_ind(z2, z1)

print(stat, p)

51.45296941515692 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 Years With Current Manager**

z1=dataset.YearsWithCurrManager\_Yes

z2=dataset.YearsWithCurrManager\_No

stat, p=ttest\_ind(z2, z1)

print(stat, p)

53.02424349024521 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)

**Step 8 – Unsupervised Learning - Correlation Analysis**

dataset=pd.read\_excel('EmployeeAtrrition.xlsx’, sheet\_name=0)

from scipy.stats import pearsonr

dataset['TotalWorkingYears']=dataset['TotalWorkingYears'].fillna(11.28)

dataset.columns

Out[258]:

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

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

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

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

'YearsSinceLastPromotion', 'YearsWithCurrManager'],

dtype='object')

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

print(stats, p)

-0.009730141010179438 0.5182860428049617

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

print(stats, p)

-0.031176281698114025 0.0384274849060192

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

print(stats, p)

-0.17011136355964646 5.4731597518148054e-30

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

print(stats, p)

-0.13439221398997386 3.163883122493571e-19

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

print(stats, p)

-0.15619931590162422 1.7339322652951965e-25

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