ASSIGNMENT DAY 07

(STATISTICAL LEARNING PROJECT)

Step1 - Launching

```
import pandas as pd

dataset = pd.read_csv('general_data.csv')

dataset
```

o/p :

```
      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
```

```
dataset.head()
```

```
   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
```

```
dataset.columns
```

```
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')
```

Step 2 - Data Treatment

```
dataset.isnull()
```

```
      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
```

dataset.dropna()

```
      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
...   ...      ...  ...                    ...                  ...
4404   29       No  ...                      1                    5
4405   42       No  ...                      0                    2
4406   29       No  ...                      0                    2
4407   25       No  ...                      1                    2
4408   42       No  ...                      7                    8
```

dataset.duplicated()

```
0       False
1       False
2       False
3       False
4       False

4405    False
4406    False
4407    False
```

```
4408    False
4409    False
```

```
dataset.drop_duplicates()
```

```
      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
```

Step 3 – Univariate Analysis

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

dataset1
```

```
               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
```

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

dataset1
```

```
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
```

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

print(dataset1)
```

```
   Age  DistanceFromHome  ...  YearsSinceLastPromotion  YearsWithCurrManager
```

0   35   2 ...   0   2

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

dataset1
```

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

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

dataset1
```

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

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

dataset1
```

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

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

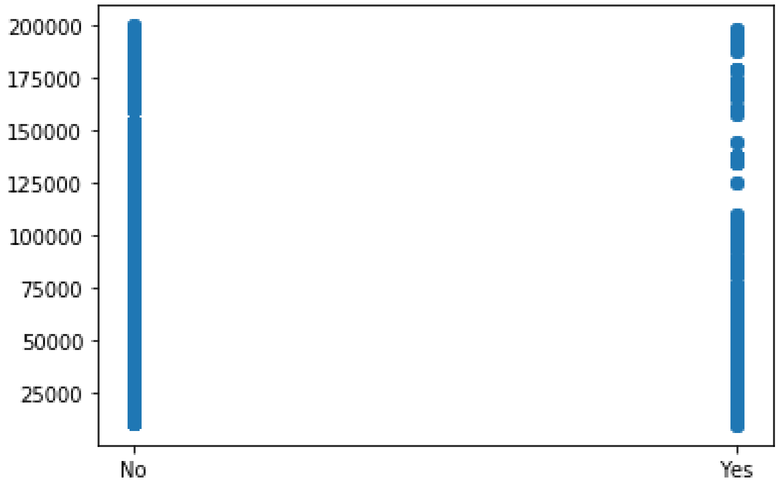
Outliers:

1. Scatter plot between Attrition and Age

```
import matplotlib.pyplot as plt

plt.scatter(dataset.Attrition,dataset.Age)
```

2. Scatter plot between Attrition and Monthlyincome

```
plt.scatter(dataset.Attrition,dataset.MonthlyIncome)
```



3. Scatter plot between Attrition and Education

```
plt.scatter(dataset.Attrition,dataset.Education)
```

4. Scatter plot between Attrition and DistanceFromHome

```
plt.scatter(dataset.Attrition,dataset.DistanceFromHome)
```

5. Scatter plot between Attrition and NumCompanies

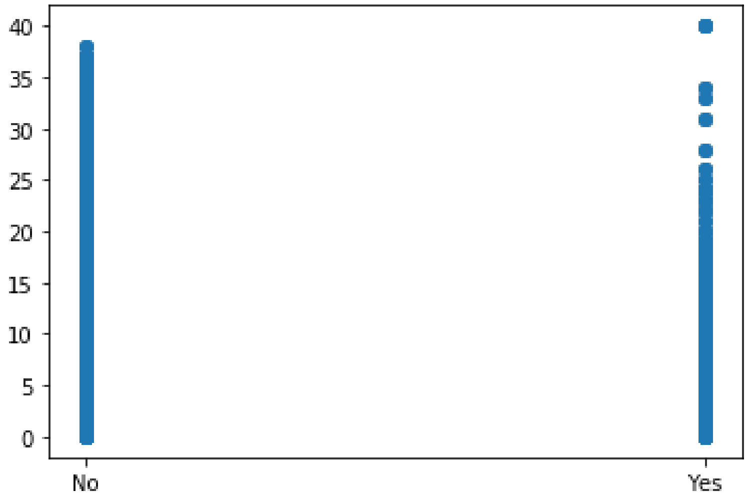
```
plt.scatter(dataset.Attrition,dataset.NumCompaniesWorked)
```

6. Scatter plot between Attrition and PercentSalaryHike

```
plt.scatter(dataset.Attrition,dataset.PercentSalaryHike)
```

7. Scatter plot between Attrition and TotalWorkingYears

```
plt.scatter(dataset.Attrition,dataset.TotalWorkingYears)
```



8. Scatter plot between Attrition and TrainingTimesLastYear

```
plt.scatter(dataset.Attrition,dataset.TrainingTimesLastYear)
```

9. Scatter plot between Attrition and YearsAtCompany

```
plt.scatter(dataset.Attrition,dataset.YearsAtCompany)
```

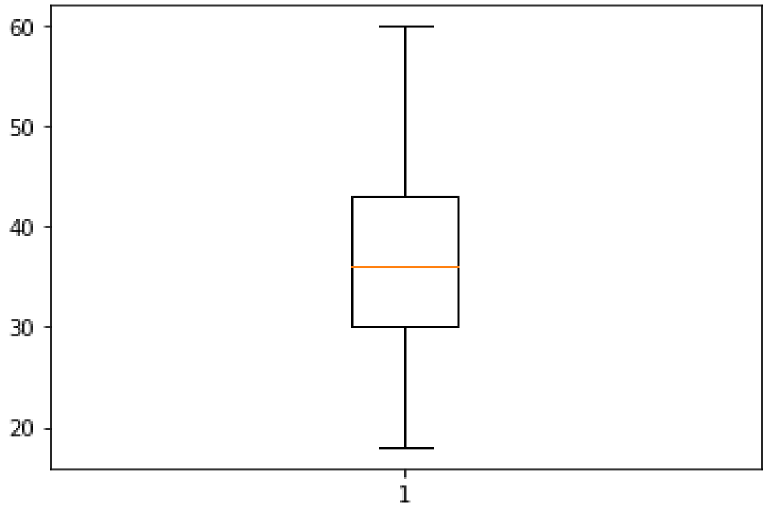
10. Scatter plot between Attrition and YearsSinceLastPromotion

```
plt.scatter(dataset.Attrition,dataset.YearsSinceLastPromotion)
```

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

- **Age**

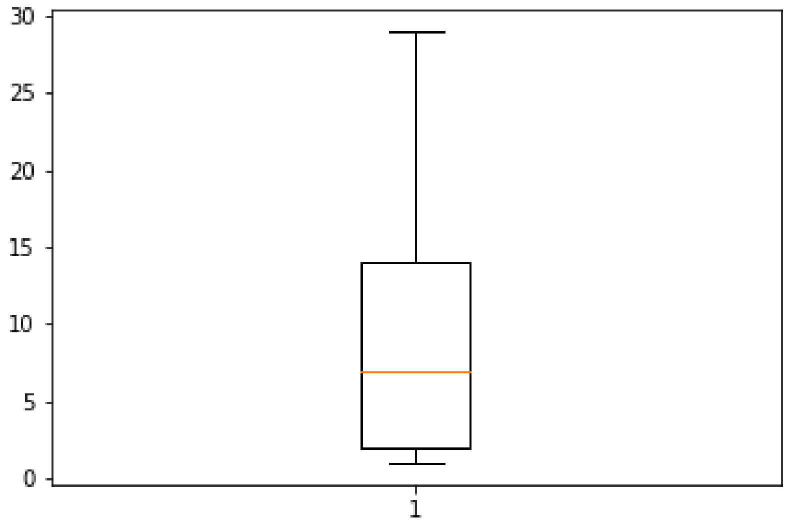
```
box_plot = dataset1.Age
plt.boxplot(box_plot)
```



Age is normally distributed without any outliers

- **DistanceFromHome**

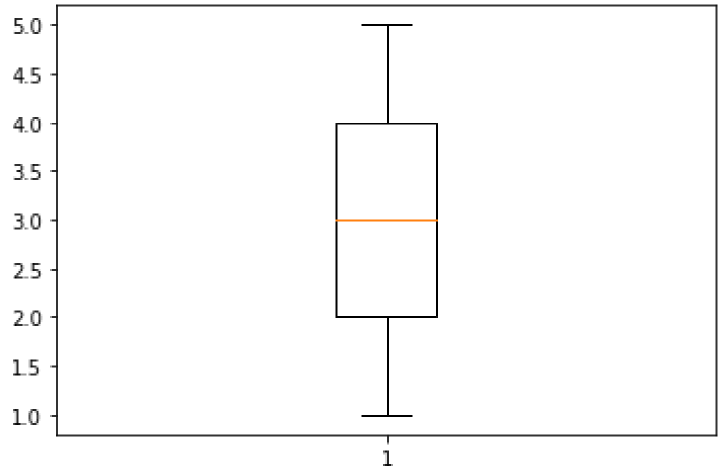
```
box_plot = dataset.DistanceFromHome
plt.boxplot(box_plot)
```



DistanceFromHome is Right skewed

- **Education**

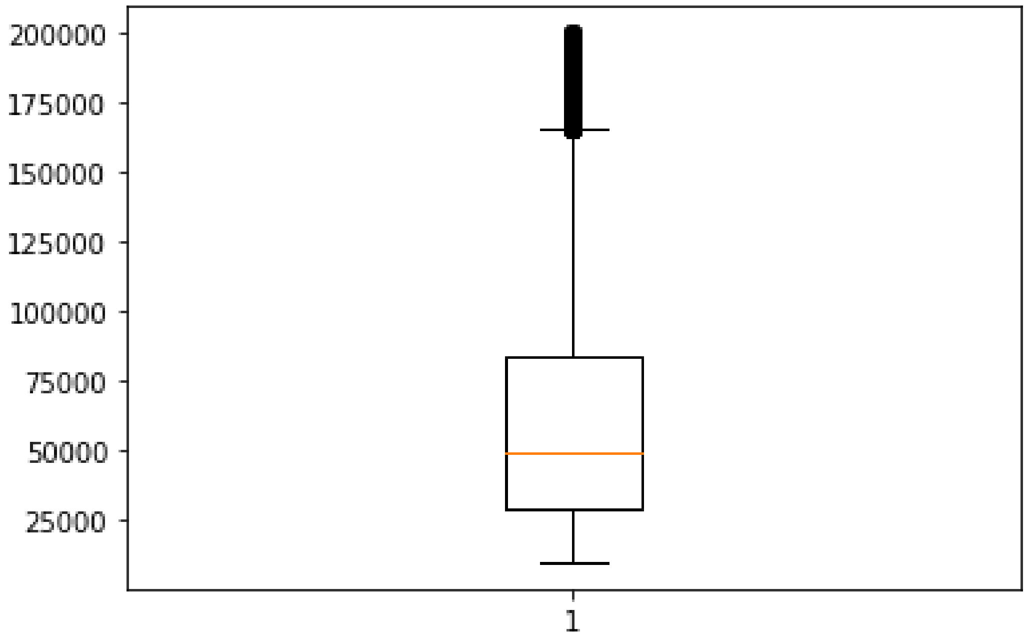
```
box_plot = dataset.Education

plt.boxplot(box_plot
```



Education is normally distributed without any outliers

- **MonthlyIncome**

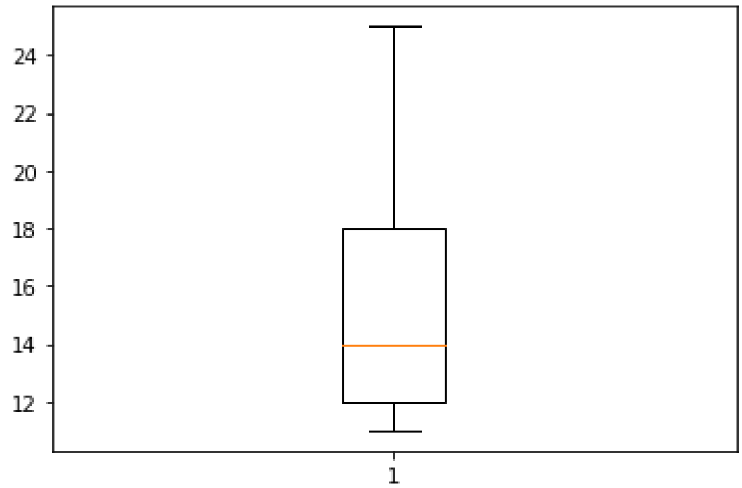
```
box_plot = dataset.MonthlyIncome

plt.boxplot(box_plot)
```



Monthly Income is Right skewed with several outliers

- **PercentSalaryHike**

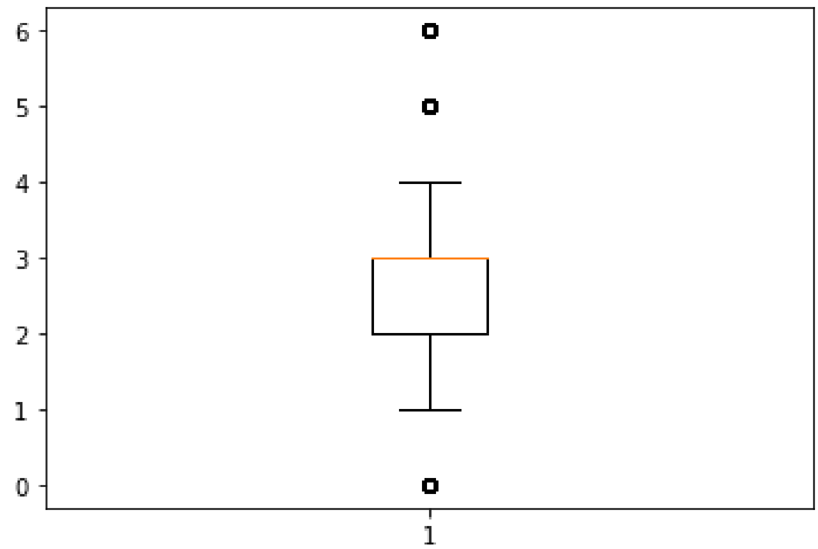
```
box_plot = dataset.PercentSalaryHike
plt.boxplot(box_plot)
```



PercentSalaryHike is Right skewed

- **TrainingTimesLastYear**

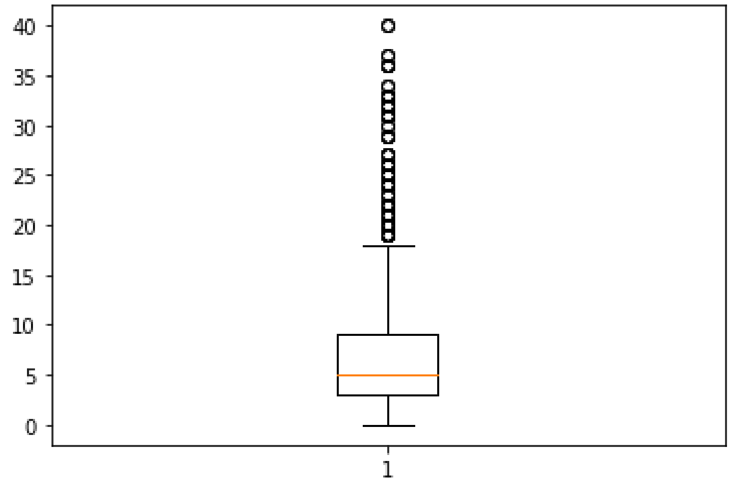
```
box_plot = dataset.TrainingTimesLastYear
plt.boxplot(box_plot)
```



TrainingTimesLastYear is Left skewed with several outlier

- **YearsAtCompany**

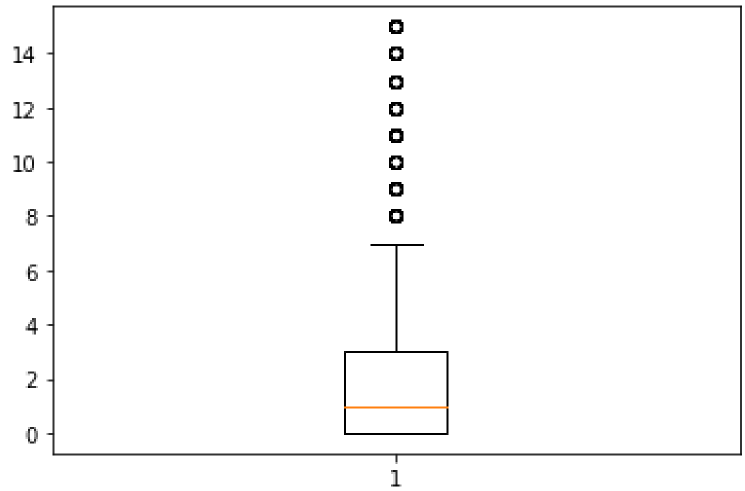
```
box_plot = dataset.YearsAtCompany

plt.boxplot(box_plot)
```



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

- **YearsSinceLastPromotion**

```
box_plot = dataset.YearsSinceLastPromotion

plt.boxplot(box_plot)
```



YearsSinceLastPromotion is also Right Skewed with several outliers observed.

Step 5 – Statistical Tests (Mann-Whitney)

```
from sklearn import preprocessing

Label_encode = preprocessing.LabelBinarizer()

dataset['Attrition'] = Label_encode.fit_transform(dataset['Attrition'])
```

- Attrition Vs Distance from Home

```
from scipy.stats import mannwhitneyu

a1=dataset.Attrition

a2=dataset.DistanceFromHome

stat, p=mannwhitneyu(a1,a2)

print(stat, p)

print('p-value',p)
```

221832.0 0.0

p-value 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
Ha: There is significant differences in the Distance From Home between attrition

- Attrition Vs Income

```
b1=dataset.Attrition

b2=dataset.MonthlyIncome

stat, p=mannwhitneyu(b1,b2)

print(stat, p)

print('p-value',p)
```

0.0 0.0

p-value 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
Ha: There is significant differences in the income between attrition

- <u>Attrition Vs Total Working Years</u>

```
b1=dataset.Attrition

b2=dataset.TotalWorkingYears

stat, p=mannwhitneyu(b1,b2)

print(stat, p)

print('p-value',p)
```

170527.5 0.0

p-value 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
Ha: There is significant differences in the Total Working Years between attrition

- <u>Attrition Vs Years at company</u>

```
a1=dataset.Attrition

a2=dataset.YearsAtCompany

stat, p=mannwhitneyu(a1,a2)

print(stat, p)

print('p-value',p)
```

520357.5 0.0

p-value 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
Ha: There is significant differences in the Years At Company between attrition

- Attrition Vs YearsWithCurrentManager

```
a1=dataset.Attrition

a2=dataset.YearsWithCurrManager

stat, p=mannwhitneyu(a1,a2)

print(stat, p)

print('p-value',p)
```

2101288.5 0.0

p-value 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
Ha: There is significant differences in the Years With Current Manager between attrition

---

Step 6 – Statistical Tests (Separate T Test)

- Attrition Vs Distance From Home

```
from scipy.stats import ttest_ind

z1=dataset.Attrition

z2=dataset.DistanceFromHome

stat, p=ttest_ind(z2,z1)

print(stat, p)

print('p-value',p)
```

73.92105563691779 0.0

p-value 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
Ha: There is significant differences in the Distance From Home between attrition

- Attrition Vs Yeats At Company

```
z1=dataset.Attrition

z2=dataset.YearsAtCompany

stat, p=ttest_ind(z2,z1)

print(stat, p)

print('p-value',p)
```

74.10006092710509 0.0

p-value 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
Ha: There is significant differences in the Years At Company between attrition

- Attrition Vs Income

```
z1=dataset.Attrition

z2=dataset.MonthlyIncome

stat, p=ttest_ind(z2,z1)

print(stat, p)

print('p-value',p)
```

91.74733118564392 0.0

p-value 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
Ha: There is significant differences in the Monthly Income between attrition

- Attrition Vs Years With Current Manager

```
z1=dataset.Attrition

z2=dataset.YearsWithCurrManager

stat, p=ttest_ind(z2,z1)

print(stat, p)

print('p-value',p)
```

73.36426551326637 0.0

p-value 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
Ha: There is significant differences in the Years With Current Manager between attrition

---

Step 8 – 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.

- Correlation between Attrition and DistanceFromHome

```
from scipy.stats import pearsonr

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

print(stats,p)

print('\nCorrelation :',stats,'\np-value :',p)



if stats == 0 :

   print('No correlation \n')

elif stats < 0 :

   print('\nNegative correlation \n')
```

```
else :

   print('Positive corrlation \n')



if p >= 0.5 :

   print("Accept null hypothesis \n")

elif p < 0.5 :

   print("Reject null hypothesis \n")
```

```
-0.009730141010179674 0.5182860428050771

Correlation : -0.009730141010179674
p-value : 0.5182860428050771

Negative correlation

Accept null hypothesis
```

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

- <u>Correlation between Attrition and MonthlyIncome</u>

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

print(stats,p)

print('\nCorrelation :',stats,'\np-value :',p)



if stats == 0 :

   print('No correlation \n')

elif stats < 0 :

   print('\nNegative correlation \n')

else :
```

```
    print('Positive corrlation \n')


if p >= 0.5 :
    print("Accept null hypothesis \n")
elif p < 0.5 :
    print("Reject null hypothesis \n")
```

```
-0.031176281698115007 0.03842748490600132

Correlation : -0.031176281698115007
p-value : 0.03842748490600132

Negative correlation

Reject null hypothesis
```

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

- Correlation between Attrition and TotalWorkingYears

```
stats,p = pearsonr(dataset.Attrition,dataset.TotalWorkingYears)
print(stats,p)
print('\nCorrelation :',stats,'\np-value :',p)


if stats == 0 :
    print('\nNo correlation \n')
elif stats < 0 :
    print('\nNegative correlation \n')
else :
    print('\nPositive correlation \n')
```

```
if p >= 0.5 :

   print("Accept null hypothesis \n")

elif p < 0.5 :

   print("Reject null hypothesis \n")
```

```
-0.17011136355964646 5.4731597518148054e-30

Correlation : -0.17011136355964646
p-value : 5.4731597518148054e-30

Negative correlation

Reject null hypothesis
```

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

- <u>Attrition & YearsAtCompany</u>

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

print(stats,p)

print('\nCorrelation :',stats,'\np-value :',p)



if stats == 0 :

   print('\nNo correlation \n')

elif stats < 0 :

   print('\nNegative correlation \n')

else :

   print('\nPositive correlation \n')
```

```
if p >= 0.5 :
   print("Accept null hypothesis \n")
elif p < 0.5 :
   print("Reject null hypothesis \n")
```

```
-0.1343922139899772 3.1638831224877484e-19

Correlation : -0.1343922139899772
p-value : 3.1638831224877484e-19

Negative correlation

Reject null hypothesis
```

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

```
stats,p = pearsonr(dataset.Attrition,dataset.YearsWithCurrManager)
print(stats,p)
print('\nCorrelation :',stats,'\np-value :',p)


if stats == 0 :
   print('\nNo correlation \n')
elif stats < 0 :
   print('\nNegative correlation \n')
else :
   print('\nPositive correlation \n')


if p >= 0.5 :
```

```
    print("Accept null hypothesis \n")
elif p < 0.5 :
    print("Reject null hypothesis \n")
```

```
-0.15619931590162847 1.7339322652896276e-25

Correlation : -0.15619931590162847

p-value : 1.7339322652896276e-25

Negative correlation

Reject null hypothesis
```

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