Attrition Assignment

Step1 - Launching

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

dataset1=pd.read_excel('C:/group_Folder/TheDataScience/Dinesh/Group 1- HR Analytics - Employee Attrition rate analysis/Working_sheet.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')
```

```
dataset1
```

```
Out[45]:

     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 18 columns]
```

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[50]:
0       False
1       False
2       False
3       False
4       False
```

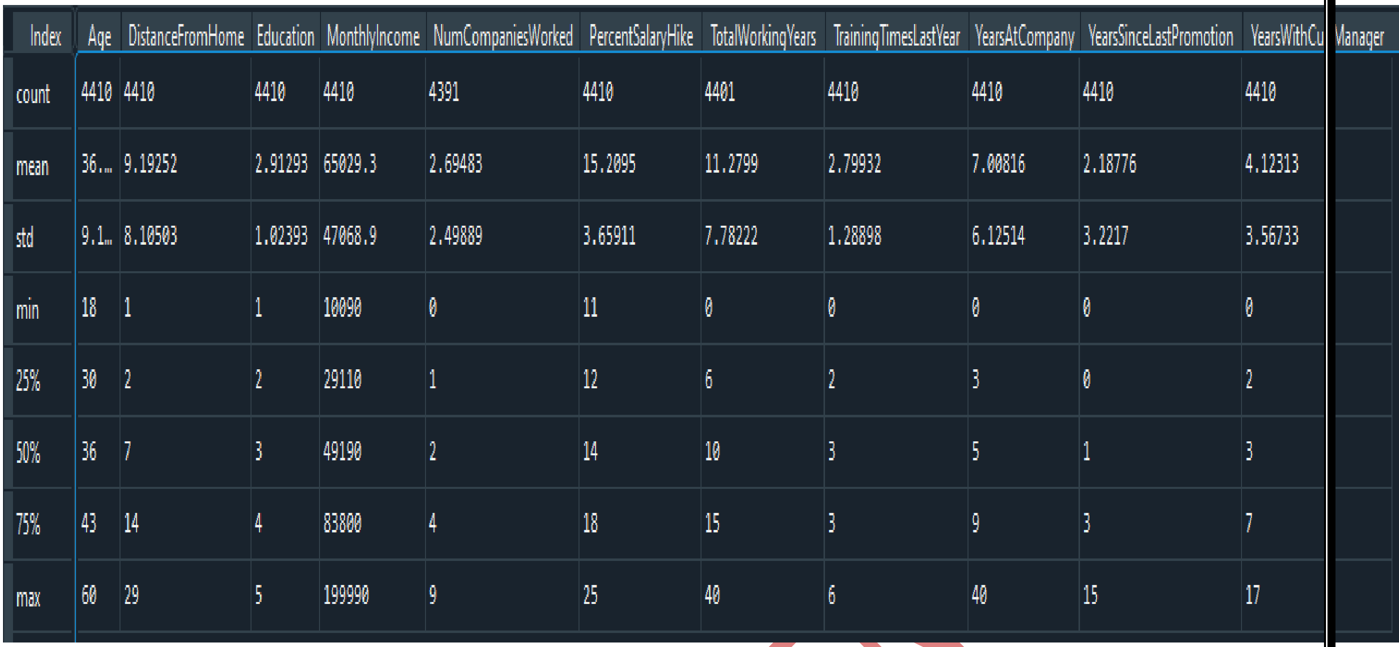
```
4405    True
4406    True
4407    True
4408    True
4409   False
Length: 4410, dtype: bool
```

```
dataset1.drop_duplicates()
```

```
Out[53]:
     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
```



Index	Age	DistanceFromHome	Education	MonthlyIncome	NumCompaniesWorked	PercentSalaryHike	TotalWorkingYears	TrainingTimesLastYear	YearsAtCompany	YearsSinceLastPromotion	YearsWithCu Manager
count	4410	4410	4410	4410	4391	4410	4401	4410	4410	4410	4410
mean	36…	9.19252	2.91293	65029.3	2.69483	15.2095	11.2799	2.79932	7.00816	2.18776	4.12313
std	9.1…	8.10503	1.02393	47068.9	2.49889	3.65911	7.78222	1.28898	6.12514	3.2217	3.56733
min	18	1	1	10090	0	11	0	0	0	0	0
25%	30	2	2	29110	1	12	6	2	3	0	2
50%	36	7	3	49190	2	14	10	3	5	1	3
75%	43	14	4	83800	4	18	15	3	9	3	7
max	60	29	5	199990	9	25	40	6	40	15	17

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

dataset3
```

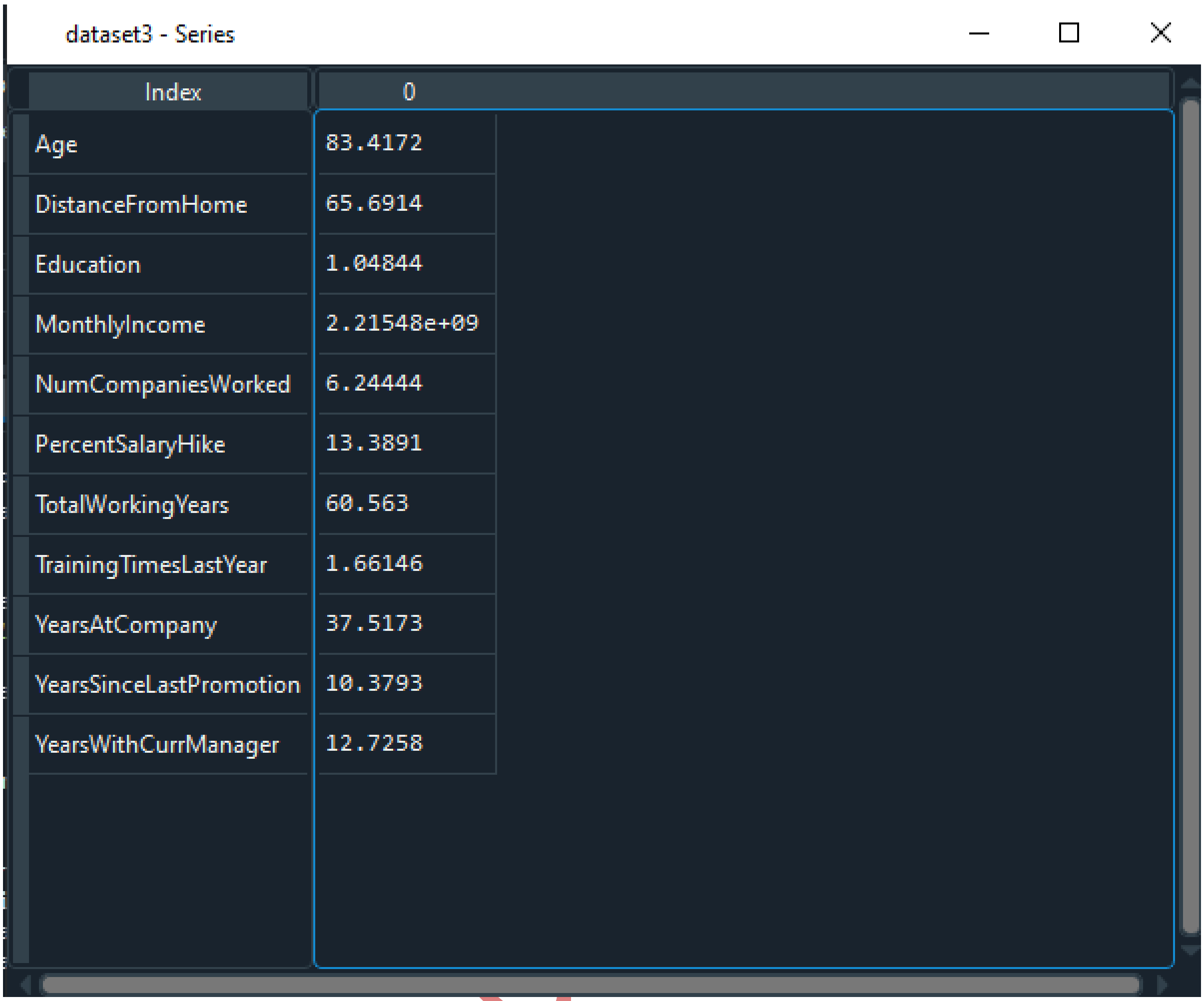
```
Out[67]:

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
```

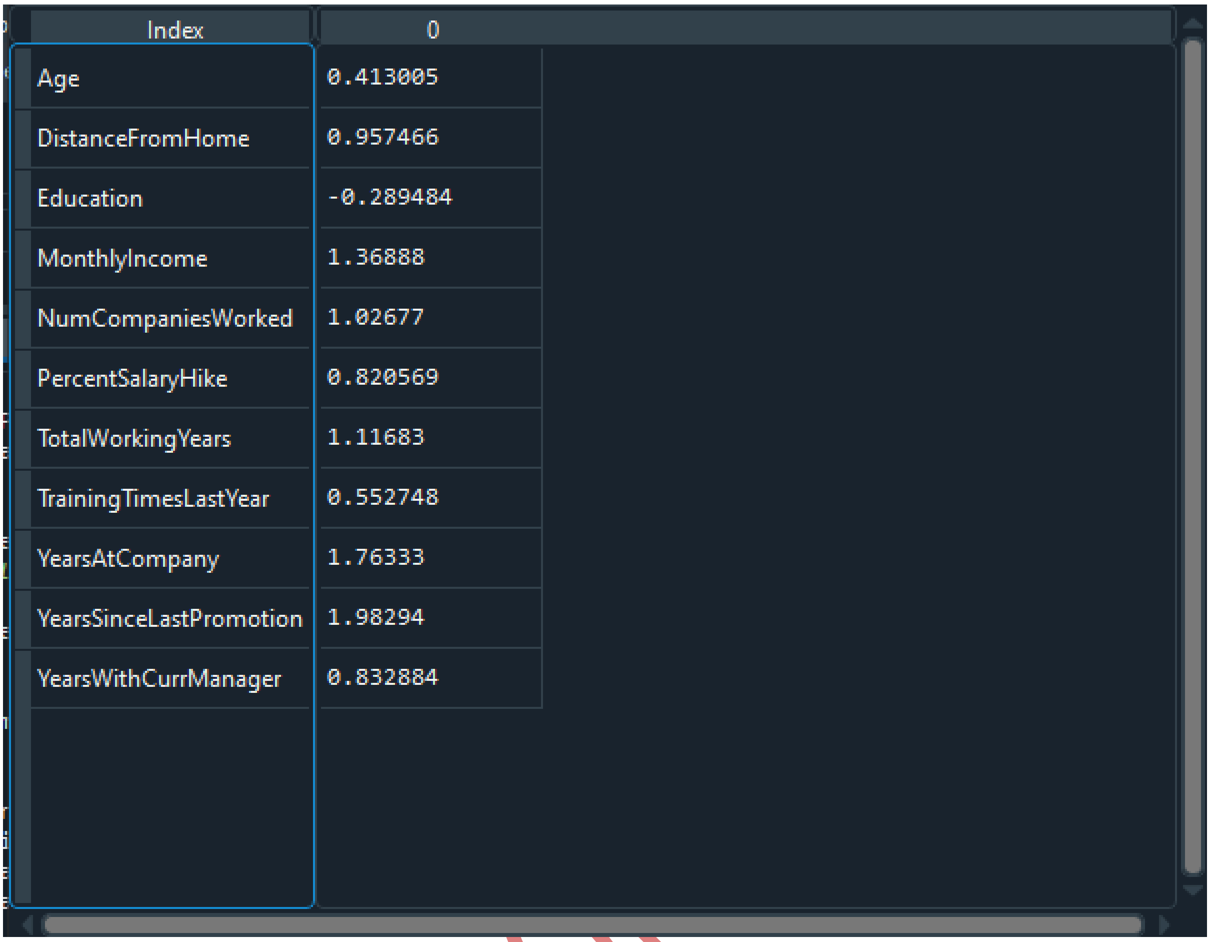
```
dataset3=dataset1[['Age','DistanceFromHome','Education','MonthlyIncome',
'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear',
'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].var()

dataset3
```

1



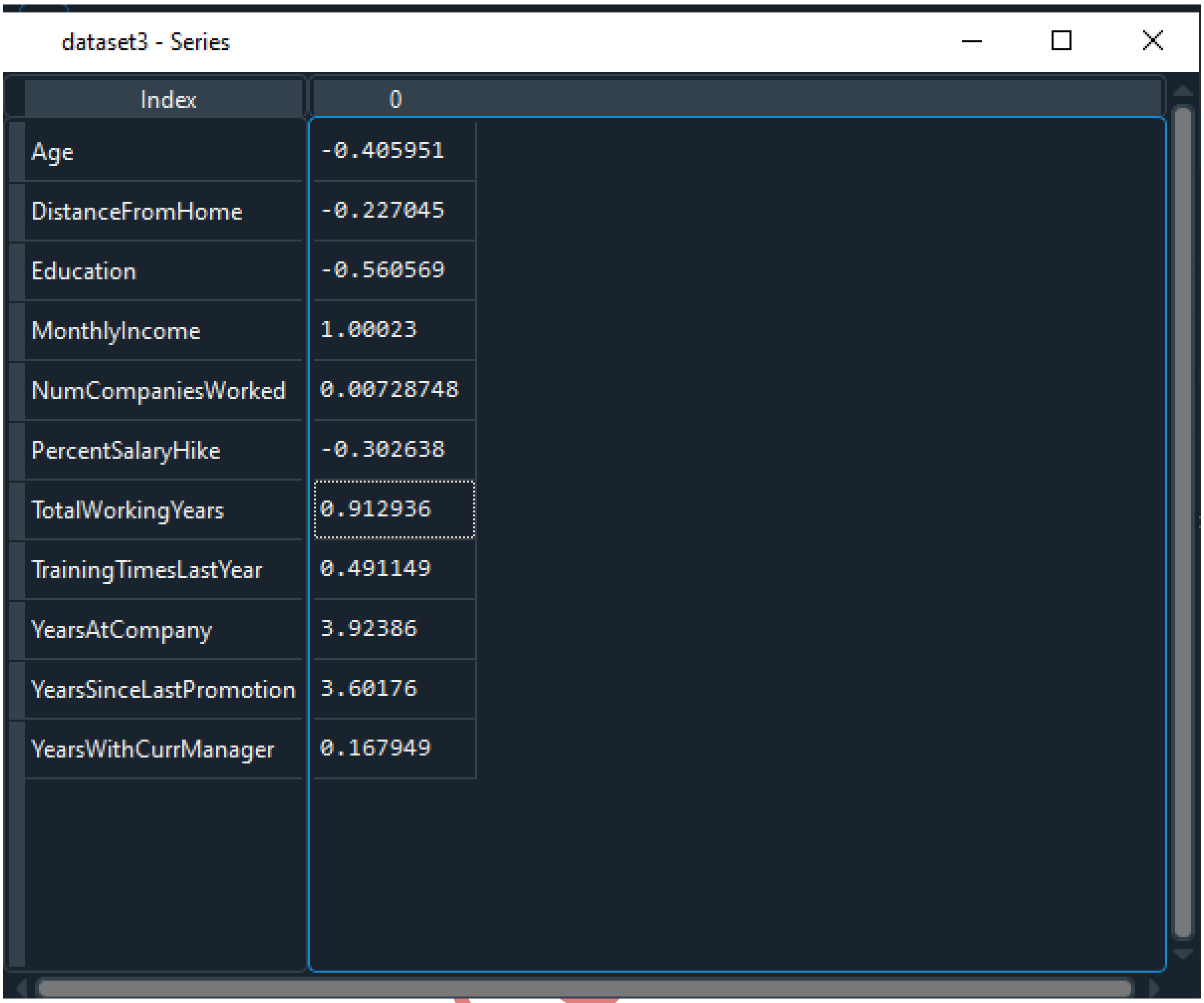
Index	0
Age	83.4172
DistanceFromHome	65.6914
Education	1.04844
MonthlyIncome	2.21548e+09
NumCompaniesWorked	6.24444
PercentSalaryHike	13.3891
TotalWorkingYears	60.563
TrainingTimesLastYear	1.66146
YearsAtCompany	37.5173
YearsSinceLastPromotion	10.3793
YearsWithCurrManager	12.7258

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

dataset3
```



Index	0
Age	0.413005
DistanceFromHome	0.957466
Education	-0.289484
MonthlyIncome	1.36888
NumCompaniesWorked	1.02677
PercentSalaryHike	0.820569
TotalWorkingYears	1.11683
TrainingTimesLastYear	0.552748
YearsAtCompany	1.76333
YearsSinceLastPromotion	1.98294
YearsWithCurrManager	0.832884

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

dataset3
```



dataset3 - Series

Index	0
Age	-0.405951
DistanceFromHome	-0.227045
Education	-0.560569
MonthlyIncome	1.00023
NumCompaniesWorked	0.00728748
PercentSalaryHike	-0.302638
TotalWorkingYears	0.912936
TrainingTimesLastYear	0.491149
YearsAtCompany	3.92386
YearsSinceLastPromotion	3.60176
YearsWithCurrManager	0.167949

	Mean	Median	Mode	Variance	Std Deviation	IQR	Skewness	Kurtosis
Mean Age (Yrs)	36	36	35	83.14	9.1	13	0.418	-0.4
Mean Distance from Home (Kms)	9	7	2	65.69	8.1	2	0.957	-0.22
Mean Monthly Income (Rs)	65000	49190	23420	2215480000	47068	54000	1.36	1
Mean Work Experience (Yrs)	11.29	10	10	60	7.72	9	1.11	0.91
Mean Years at Company (Yrs)	7	5	5	37.51	6.12	6	1.76	3.92
Mean Years since last promotion (Yrs)	2	1	0	10.37	3.22	3	1.98	.6
Mean Years with Current Manager (Yrs)	4	3	2	12.72	3.56	5	0.83	0.16

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

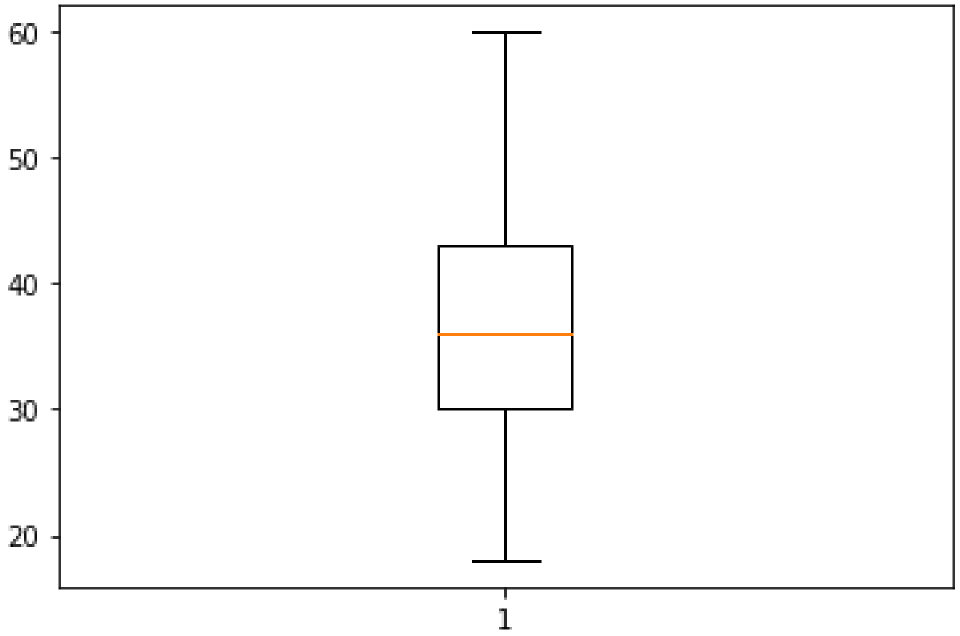
Outliers:

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

box_plot=dataset1.Age

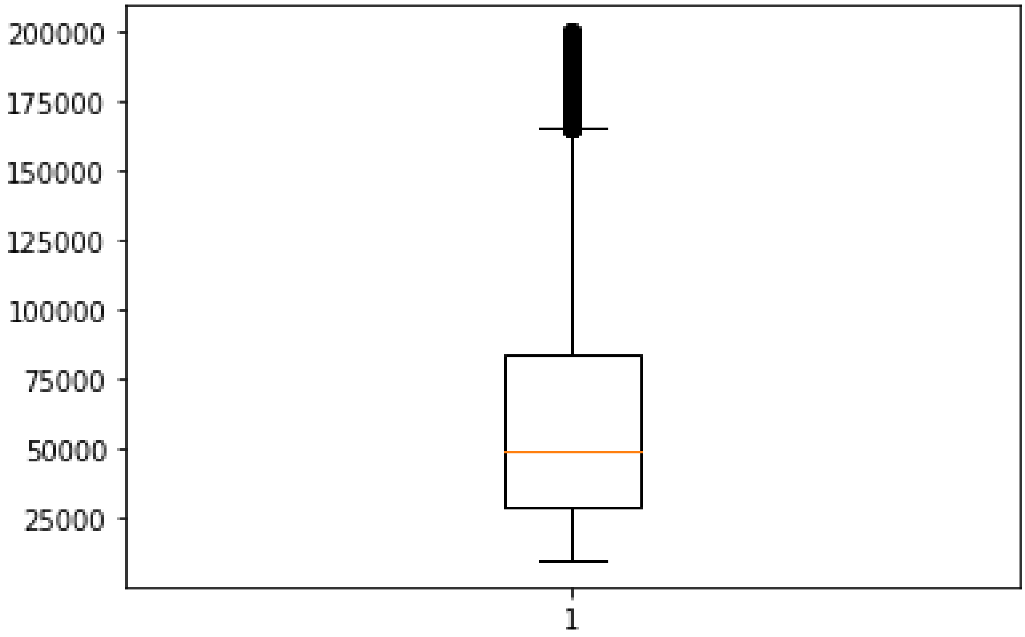
```
plt.boxplot(box_plot)
```

```
Out[23]:
```



Age is normally distributed without any outliers

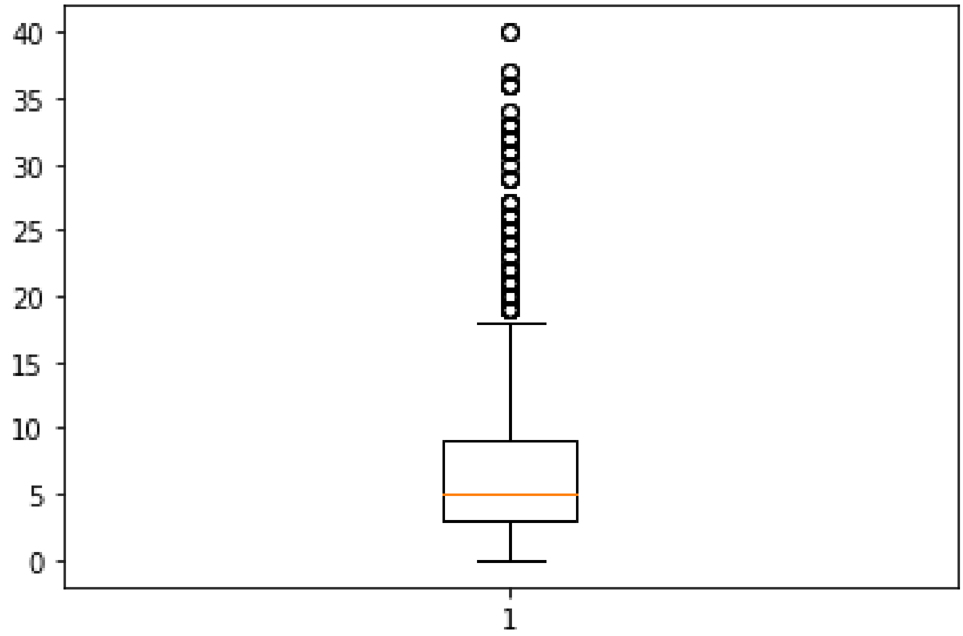
```
box_plot=dataset1.MonthlyIncome
plt.boxplot(box_plot)
```



Monthly Income is Right skewed with several outliers

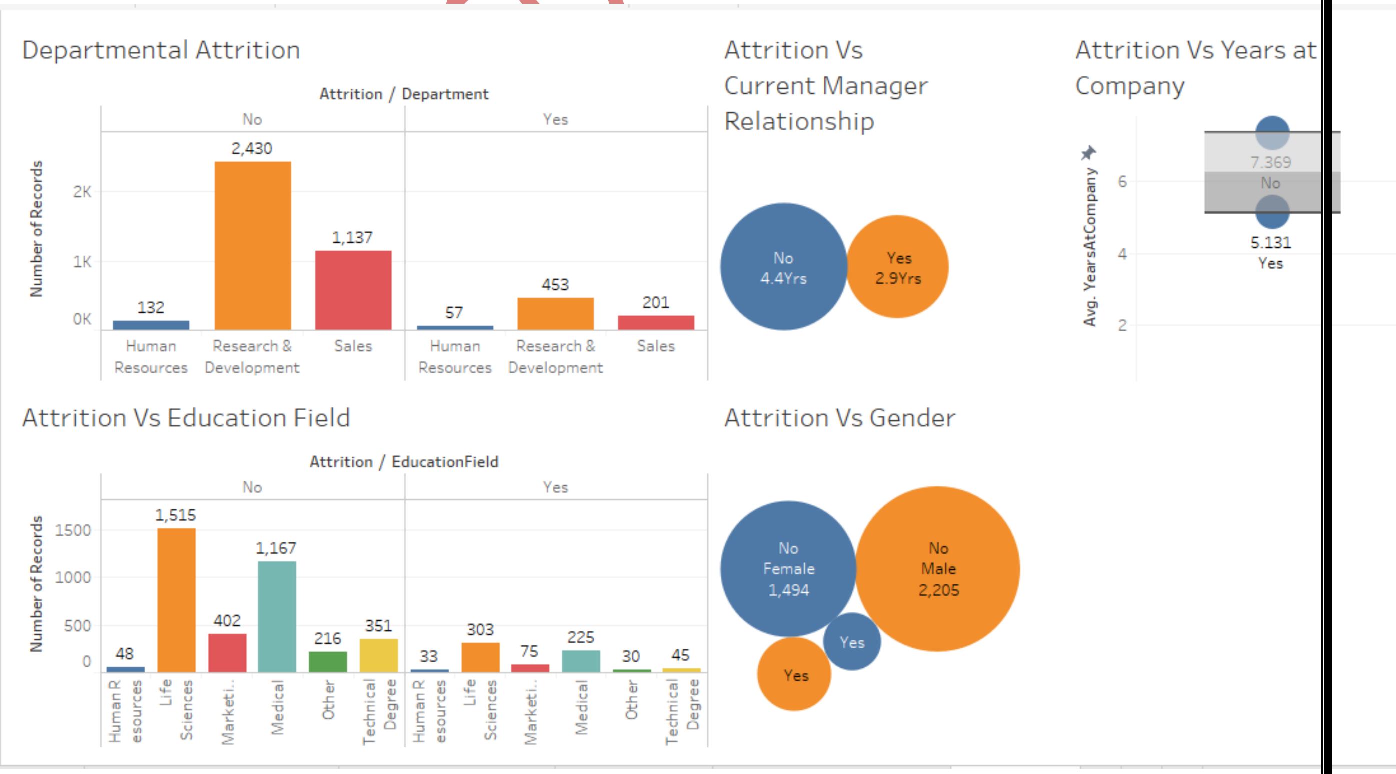
```
box_plot=dataset1.YearsAtCompany
```

```
plt.boxplot(box_plot)
```



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

Step 4 – Visualisation:



Step 5 – Statistical Tests (Mann-Whitney)

Attrition Vs Distance from Home

```
import pandas as pd

dataset=pd.read_excel('C:/Group_Folder/TheDataScience/Dinesh/Group 1- HR Analytics - Employee Attrition rate analysis/Working_sheet.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

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.

```
dataset=pd.read_excel('C:/Group_Folder/TheDataScience/Dinesh/Group 1- HR Analytics - Employee Attrition rate analysis/Working_sheet.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