Assignment - Attrition

Step 1: Load the sheet/Data

```
import pandas as pd

import matplotlib.pyplot as mplt

dataset = pd.read_csv("D:/AI_ML_Course/Day 7/general_data.csv")
```

```
dataset.columns
```

```
Out[3]:

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()
```

```
Out[4]:

      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 24 columns]
```

```
dataset.duplicated()
```

```
Out[6]:

0       False
1       False
2       False
3       False
4       False


4405    False
4406    False
```

```
4407    False

4408    False

4409    False

Length: 4410, dtype: bool
```

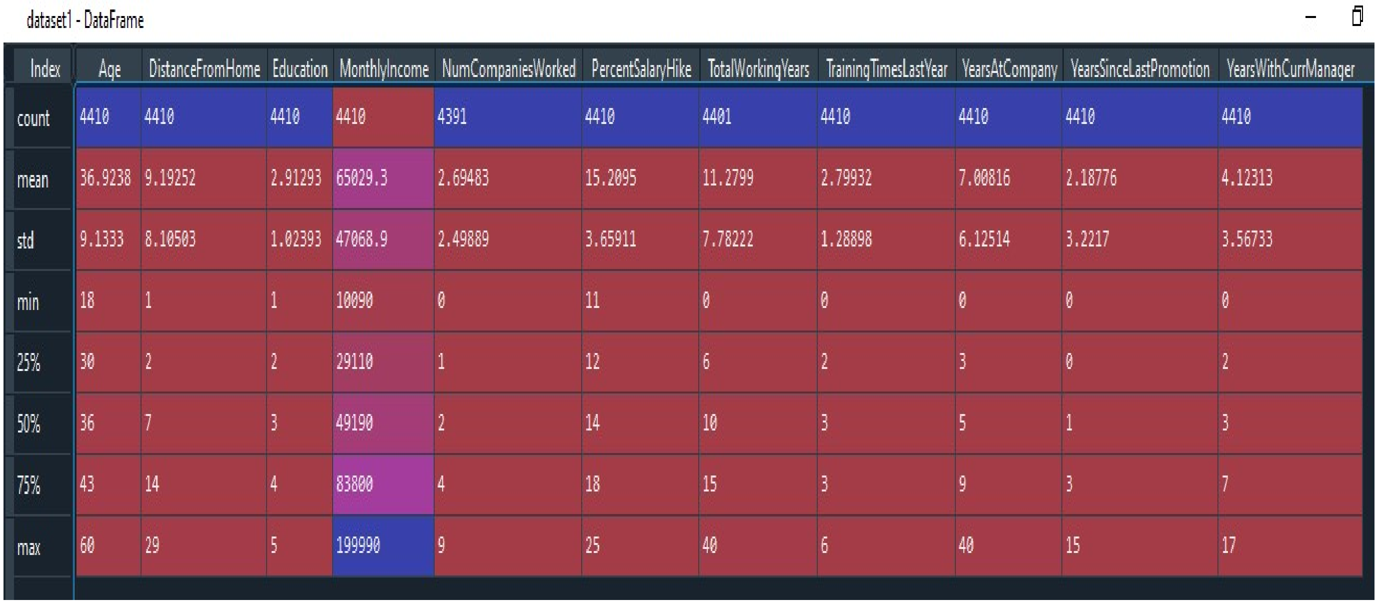
```
dataset.drop_duplicates()
```

```
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

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

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 3: Uni-Variate Analysis:

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

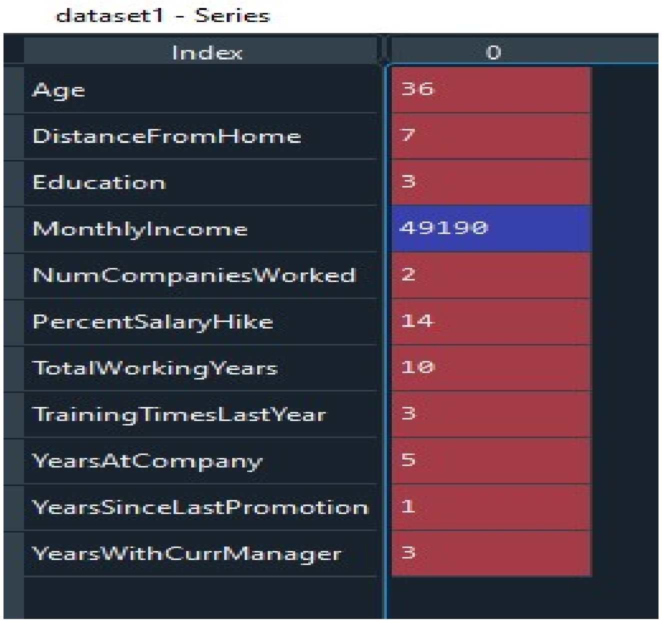
Dataset1



dataset1 - DataFrame

Index	Age	DistanceFromHome	Education	MonthlyIncome	NumCompaniesWorked	PercentSalaryHike	TotalWorkingYears	TrainingTimesLastYear	YearsAtCompany	YearsSinceLastPromotion	YearsWithCurrManager
count	4410	4410	4410	4410	4391	4410	4401	4410	4410	4410	4410
mean	36.9238	9.19252	2.91293	65029.3	2.69483	15.2095	11.2799	2.79932	7.00816	2.18776	4.12313
std	9.1333	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

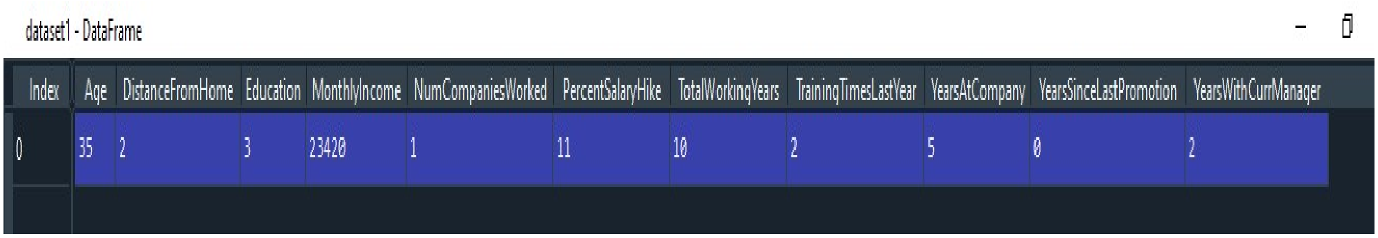
dataset1=dataset[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].median()



dataset1 - Series

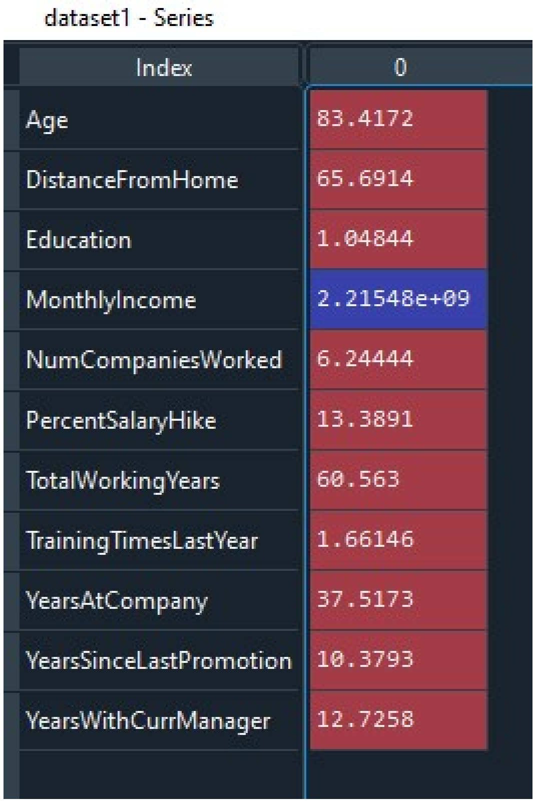
Index	0
Age	36
DistanceFromHome	7
Education	3
MonthlyIncome	49190
NumCompaniesWorked	2
PercentSalaryHike	14
TotalWorkingYears	10
TrainingTimesLastYear	3
YearsAtCompany	5
YearsSinceLastPromotion	1
YearsWithCurrManager	3

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



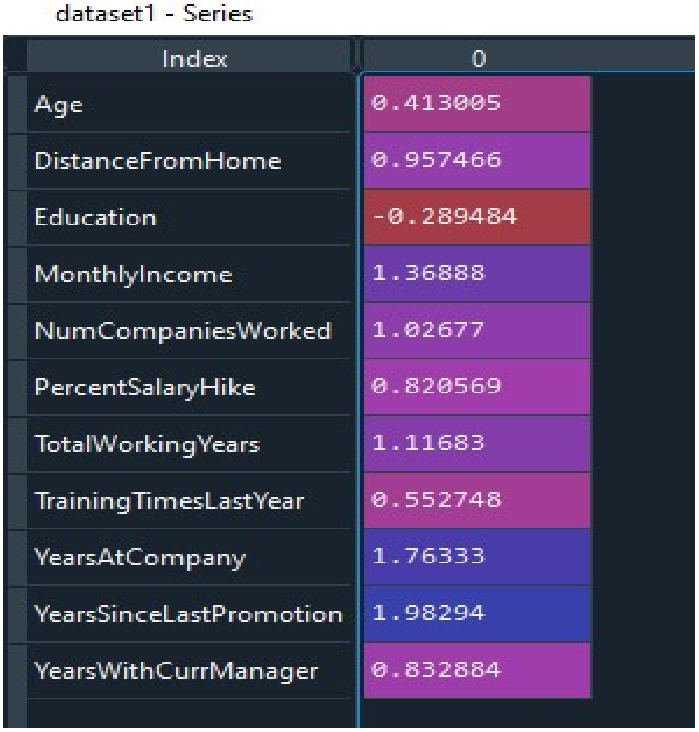
Index	Age	DistanceFromHome	Education	MonthlyIncome	NumCompaniesWorked	PercentSalaryHike	TotalWorkingYears	TrainingTimesLastYear	YearsAtCompany	YearsSinceLastPromotion	YearsWithCurrManager
0	35	2	3	23420	1	11	10	2	5	0	2

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



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

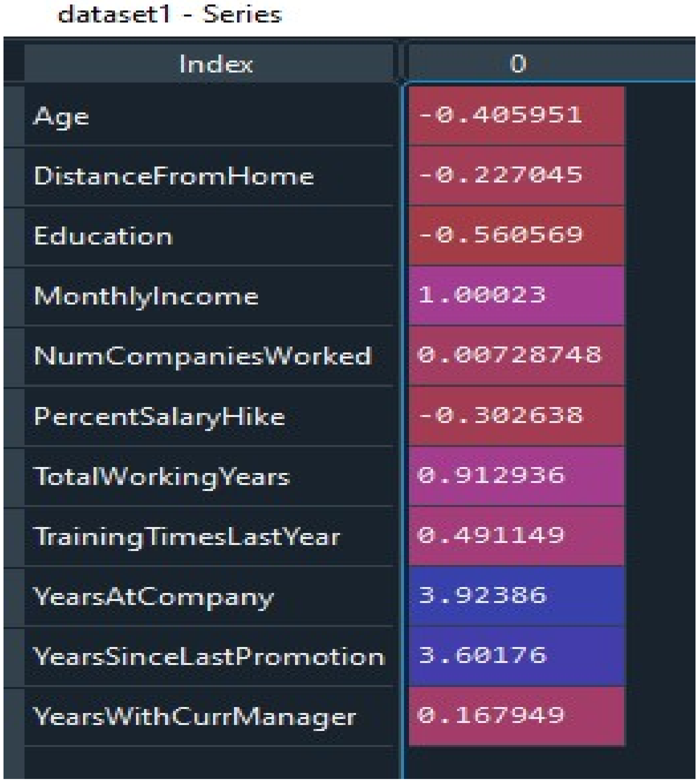
```
dataset1=dataset[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked',
'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear',
'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].skew()
```



dataset1 - Series

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

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



dataset1 - 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
Age (Yrs)	36.9	36	35	83.41	9.13	13	0.41	-0.41
DistanceFromHome (Km)	9.19	7	2	65.69	8.1	12	0.96	-0.23
Monthly Income (Rs)	65029	49190	23420	2215480270	47068	54690	1.37	1
PercentSalaryHike (%)	15	14	11	13.39	3.66	6	0.82	-0.3
TotalWorkingYears (Yrs)	11.29	10	10	60.56	7.78	9	1.12	0.91
YearsAtCompany (Yrs)	7	5	5	37.52	6.12	6	1.76	3.92
YearsSinceLastPromotion (Yrs)	2	1	0	10.38	3.22	3	1.98	3.6
YearsWithCurrManager (Yrs)	4	3	2	12.73	3.57	5	0.83	0.17

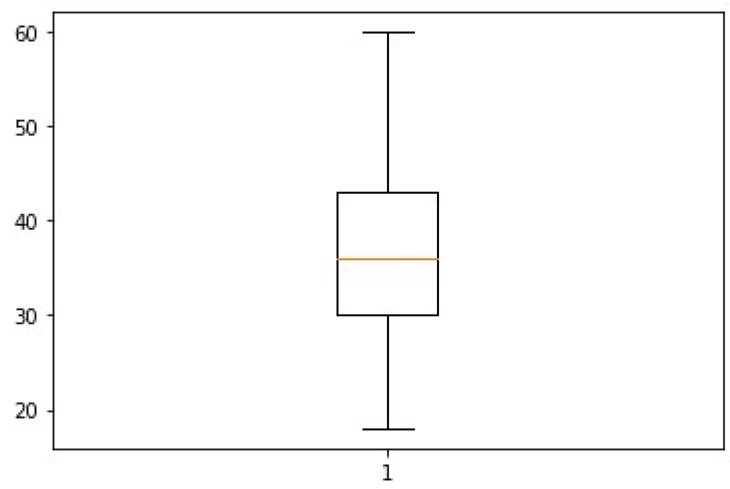
Inference from the analysis:

- All the above variables show positive skewness.
- Years_At_Company & Years_Since_LastPromotion are Leptokurtic i.e. more than 3 and all other variables are Platykurtic.
- The Mean_Monthly_Income's IQR is at 54K suggesting companywide attrition across all income bands
- Mean age forms a near normal distribution with 13 years of IQR
- Mean Distance_From_Home is 12 Km which is higher.

Outliers:

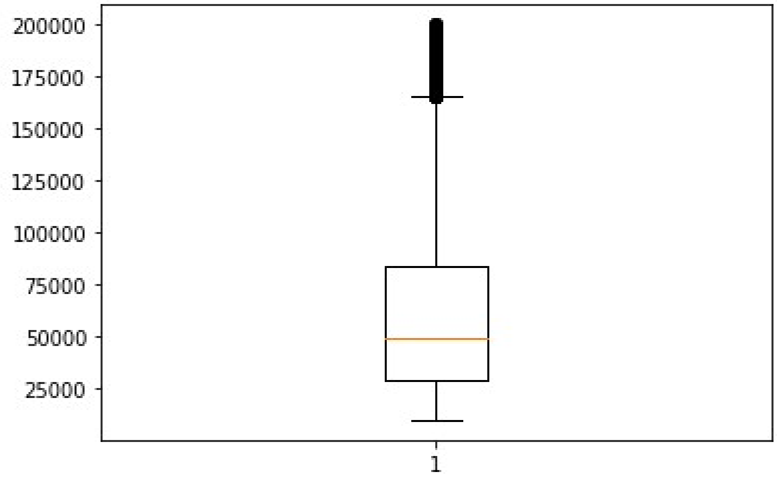
There's no regression found while plotting Age, MonthlyIncome, TotalWorkingYears, YearsAtCompany, etc., on a scatter plot

```
box_plot=dataset.Age

mplt. boxplot(box_plot)
```



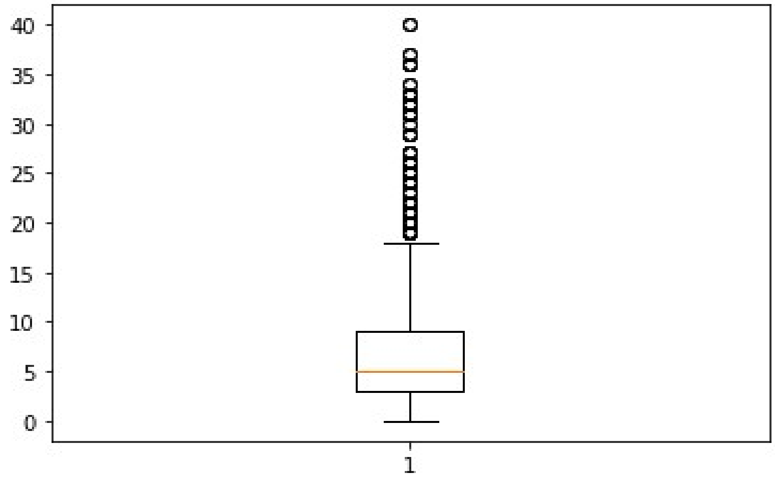
Age is normally distributed without any Outliers

```
box_plot=dataset.MonthlyIncome

mplt.boxplot(box_plot)
```



Monthly Income is right skewed with several Outliers

```
box_plot=dataset.YearsAtCompany

mplt.boxplot(box_plot)
```



Years at company is also Right skewed with several Outliers