Attrition Analysis

Step 1: Launching

```
import pandas as pd

dataset = pd.read_excel("Attrition Analysis Data.xlsx", sheet_name = 0)

dataset.head()
```

```
Out[3]:
   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]
```

```
dataset.tail()
```

```
Out[4]:
      Age Attrition  ... YearsSinceLastPromotion YearsWithCurrManager
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

[5 rows x 24 columns]
```

```
dataset.columns
```

```
Out[5]:
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.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
```

```
dataset1 =  dataset.drop_duplicates()
```

```
dataset1.isnull()
Out[8]:
       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]
```

```
working_dataset = dataset1.dropna()
```

```
working_dataset.info()
<class 'pandas.core.frame.DataFrame'>
Int64Index: 4382 entries, 0 to 4408
Data columns (total 24 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Age                      4382 non-null   int64
 1   Attrition                4382 non-null   object
 2   BusinessTravel           4382 non-null   object
 3   Department               4382 non-null   object
 4   DistanceFromHome         4382 non-null   int64
 5   Education                4382 non-null   int64
 6   EducationField           4382 non-null   object
 7   EmployeeCount            4382 non-null   int64
 8   EmployeeID               4382 non-null   int64
 9   Gender                   4382 non-null   object
 10  JobLevel                 4382 non-null   int64
 11  JobRole                 4382 non-null   object 
 12  MaritalStatus           4382 non-null   object 
 13  MonthlyIncome           4382 non-null   int64  
 14  NumCompaniesWorked      4382 non-null   float64
 15  Over18                  4382 non-null   object 
 16  PercentSalaryHike       4382 non-null   int64  
 17  StandardHours           4382 non-null   int64  
 18  StockOptionLevel        4382 non-null   int64  
 19  TotalWorkingYears       4382 non-null   float64
 20  TrainingTimesLastYear   4382 non-null   int64  
 21  YearsAtCompany          4382 non-null   int64  
 22  YearsSinceLastPromotion  4382 non-null   int64  
 23  YearsWithCurrManager    4382 non-null   int64  
dtypes: float64(2), int64(14), object(8)
memory usage: 855.9+ KB
```

Step 3: Univariate Analysis

```
dataset3 = working_dataset[['Age','DistanceFromHome','Education','EmployeeCount', 'EmployeeID',
     'JobLevel','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike', 'StandardHours',
      'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany',
      'YearsSinceLastPromotion', 'YearsWithCurrManager']]


dataset3.describe()
Out[12]: 
               Age  ...  YearsWithCurrManager
count  4382.000000  ...           4382.000000
mean     36.933364  ...              4.126198
std       9.137272  ...              3.569674
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 16 columns]
```

```
dataset3 = working_dataset[['Age','DistanceFromHome','Education','EmployeeCount', 'EmployeeID',
      'JobLevel','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike', 'StandardHours',
       'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany',
       'YearsSinceLastPromotion', 'YearsWithCurrManager']]
```

```
dataset3.median()
Out[16]:
Age                            36.0
DistanceFromHome                7.0
Education                       3.0
EmployeeCount                   1.0
EmployeeID                   2208.5
JobLevel                        2.0
MonthlyIncome               49190.0
NumCompaniesWorked              2.0
PercentSalaryHike              14.0
StandardHours                   8.0
StockOptionLevel                1.0
TotalWorkingYears              10.0
TrainingTimesLastYear           3.0
YearsAtCompany                  5.0
YearsSinceLastPromotion         1.0
YearsWithCurrManager            3.0
dtype: float64
```

```
dataset3.mode()

Out[17]:

    Age  DistanceFromHome  ...  YearsSinceLastPromotion  YearsWithCurrManager

0    35.0              2.0  ...                      0.0                   2.0
```

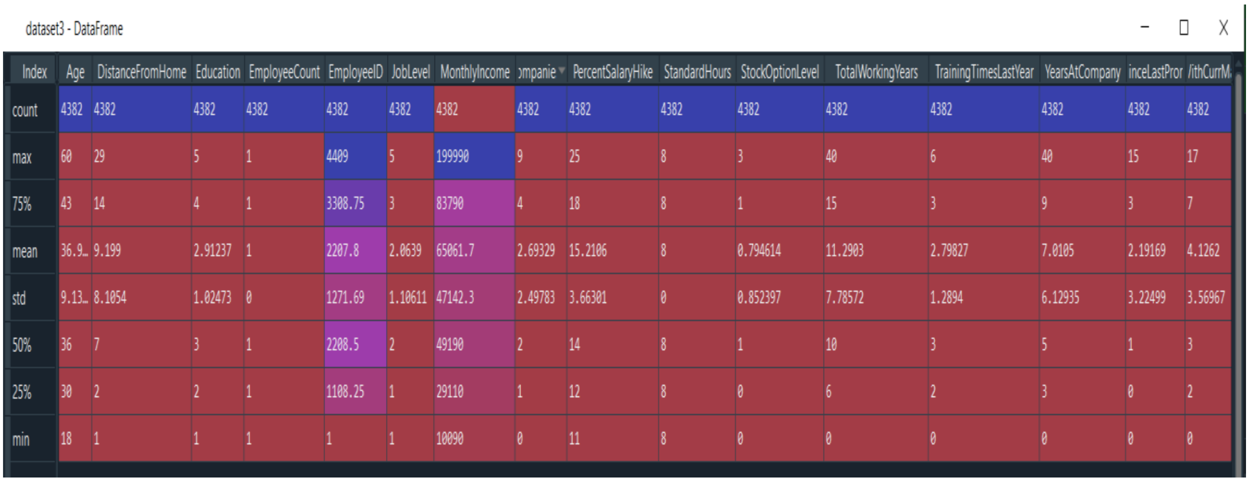
```
dataset3.var()

Out[23]:

Age                        8.348974e+01

DistanceFromHome           6.569744e+01

Education                  1.050068e+00

EmployeeCount              0.000000e+00

EmployeeID                 1.617192e+06

JobLevel                   1.223490e+00

MonthlyIncome              2.222397e+09

NumCompaniesWorked         6.239165e+00

PercentSalaryHike          1.341762e+01

StandardHours              0.000000e+00

StockOptionLevel           7.265814e-01

TotalWorkingYears          6.061739e+01

TrainingTimesLastYear      1.662558e+00

YearsAtCompany             3.756894e+01

YearsSinceLastPromotion    1.040059e+01

YearsWithCurrManager       1.274257e+01

dtype: float64
```

```
dataset3.skew()

Out[24]:

Age                        0.413048

DistanceFromHome           0.955517

Education                 -0.288977

EmployeeCount              0.000000
```

```
EmployeeID                -0.002335
JobLevel                   1.021797
MonthlyIncome              1.367457
NumCompaniesWorked         1.029174
PercentSalaryHike          0.819510
StandardHours              0.000000
StockOptionLevel           0.967263
TotalWorkingYears          1.115419
TrainingTimesLastYear      0.551818
YearsAtCompany             1.764619
YearsSinceLastPromotion    1.980992
YearsWithCurrManager       0.834277
dtype: float64

dataset3.kurt()
Out[25]:
Age                       -0.409517
DistanceFromHome          -0.230691
Education                 -0.565008
EmployeeCount              0.000000
EmployeeID                -1.198607
JobLevel                   0.388189
MonthlyIncome              0.990836
NumCompaniesWorked         0.014307
PercentSalaryHike         -0.306951
StandardHours              0.000000
StockOptionLevel           0.356755
TotalWorkingYears          0.909316
TrainingTimesLastYear      0.494215
YearsAtCompany             3.930726
YearsSinceLastPromotion    3.592162
```

```
YearsWithCurrManager      0.170703
dtype: float64
```

```
dataset3.std()
Out[27]:
Age                        9.137272
DistanceFromHome           8.105396
Education                  1.024728
EmployeeCount              0.000000
EmployeeID              1271.688783
JobLevel                   1.106115
MonthlyIncome          47142.310175
NumCompaniesWorked         2.497832
PercentSalaryHike          3.663007
StandardHours              0.000000
StockOptionLevel           0.852397
TotalWorkingYears          7.785717
TrainingTimesLastYear      1.289402
YearsAtCompany             6.129351
YearsSinceLastPromotion    3.224994
YearsWithCurrManager       3.569674
dtype: float64
```



dataset3 - DataFrame

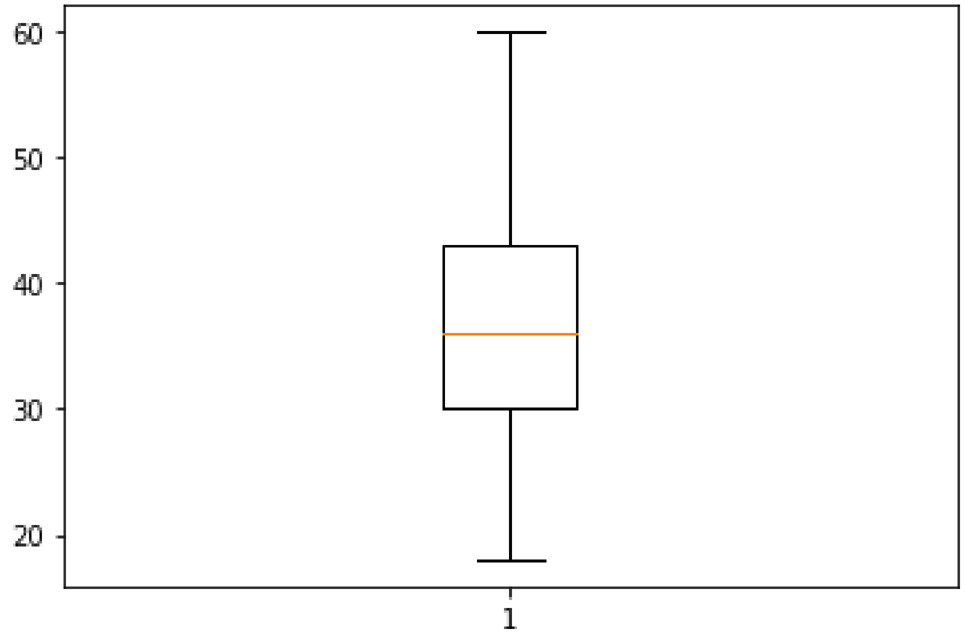
Index	Age	DistanceFromHome	Education	EmployeeCount	EmployeeID	JobLevel	MonthlyIncome	ɔmpanie	PercentSalaryHike	StandardHours	StockOptionLevel	TotalWorkingYears	TrainingTimesLastYear	YearsAtCompany	inceLastPror	/ithCurrM
count	4382	4382	4382	4382	4382	4382	4382	4382	4382	4382	4382	4382	4382	4382	4382	4382
max	60	29	5	1	4409	5	199990	9	25	8	3	40	6	40	15	17
75%	43	14	4	1	3308.75	3	83790	4	18	8	1	15	3	9	3	7
mean	36.9…	9.199	2.91237	1	2207.8	2.0639	65061.7	2.69329	15.2106	8	0.794614	11.2903	2.79827	7.0105	2.19169	4.1262
std	9.13…	8.1054	1.02473	0	1271.69	1.10611	47142.3	2.49783	3.66301	0	0.852397	7.78572	1.2894	6.12935	3.22499	3.56967
50%	36	7	3	1	2208.5	2	49190	2	14	8	1	10	3	5	1	3
25%	30	2	2	1	1108.25	1	29110	1	12	8	0	6	2	3	0	2
min	18	1	1	1	1	1	10090	0	11	8	0	0	0	0	0	0

Inference:

- All the above variables in the dataset3 are positively skewed except for Education and EmployeeID which are negatively skewed.
- Age, Distance from home, Education and Percent Salary Hike are platyokurtic in nature whereas all other values are leptokurtic
- For age mean, median mode is nearly same and hence is normally distributed with an IQR (Q3 – Q1) of 13 years.

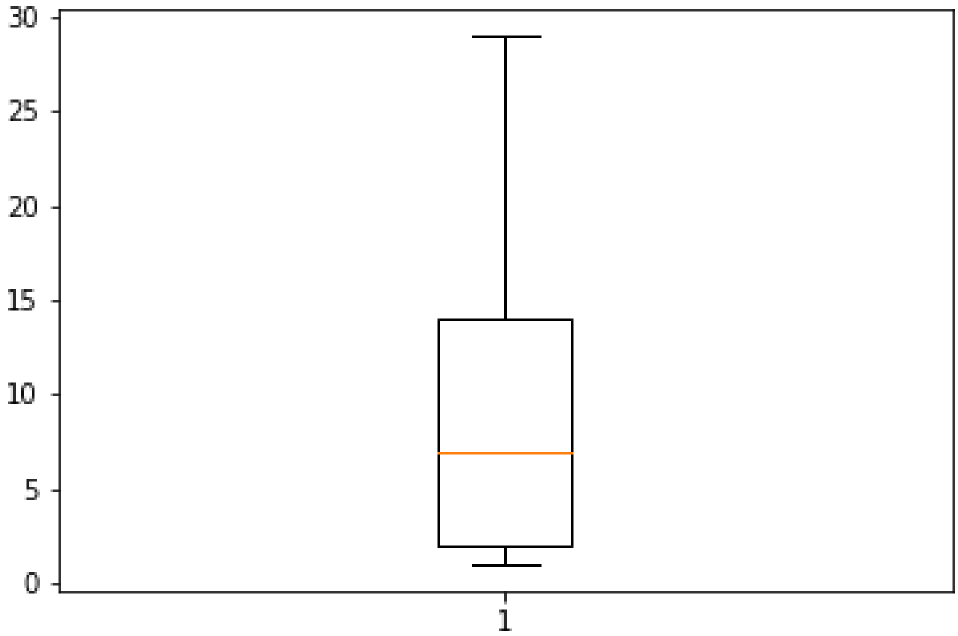
Outliers:

```
import matplotlib.pyplot as plt

plt.boxplot(dataset3.Age)

Out[30]:

{'whiskers': [<matplotlib.lines.Line2D at 0x2877e7b2a08>,

  <matplotlib.lines.Line2D at 0x2877e7987c8>],

 'caps': [<matplotlib.lines.Line2D at 0x2877efa8608>,

  <matplotlib.lines.Line2D at 0x2877efaab08>],

 'boxes': [<matplotlib.lines.Line2D at 0x2877e7b4d08>],

 'medians': [<matplotlib.lines.Line2D at 0x2877f087748>],

 'fliers': [<matplotlib.lines.Line2D at 0x2877efbdcc8>],

 'means': []}
```



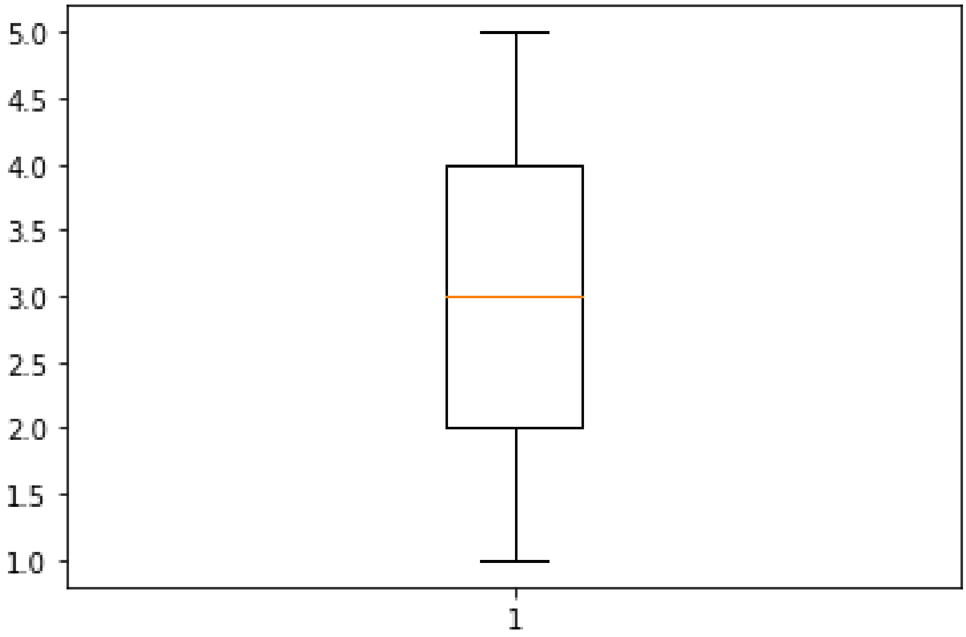
As mean, median and mode are equal, Age is normally distributed without any outliers.

```
plt.boxplot(dataset3.DistanceFromHome)

Out[31]:

{'whiskers': [<matplotlib.lines.Line2D at 0x2877e6f6788>,

  <matplotlib.lines.Line2D at 0x2877e6f69c8>],

 'caps': [<matplotlib.lines.Line2D at 0x2877e6f6448>,

  <matplotlib.lines.Line2D at 0x2877e703388>],

 'boxes': [<matplotlib.lines.Line2D at 0x2877e6e9648>],

 'medians': [<matplotlib.lines.Line2D at 0x2877d7c7148>],

 'fliers': [<matplotlib.lines.Line2D at 0x2877e6cbec8>],

 'means': []}
```



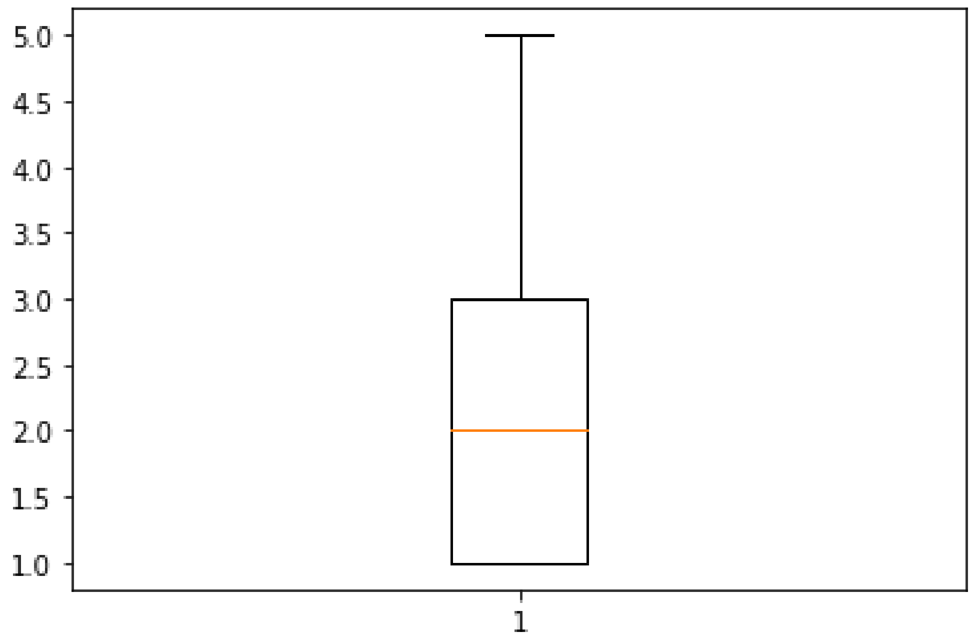
DistanceFromHome is Right Skewed without any outliers.

```
plt.boxplot(dataset3.Education)

Out[32]:

{'whiskers': [<matplotlib.lines.Line2D at 0x2877da59b48>,

  <matplotlib.lines.Line2D at 0x2877da59448>],

 'caps': [<matplotlib.lines.Line2D at 0x2877e6de988>,
 <matplotlib.lines.Line2D at 0x2877ea83fc8>],
 'boxes': [<matplotlib.lines.Line2D at 0x2877eda8588>],
 'medians': [<matplotlib.lines.Line2D at 0x2877ea83b88>],
 'fliers': [<matplotlib.lines.Line2D at 0x2877e6f0b88>],
 'means': []}
```



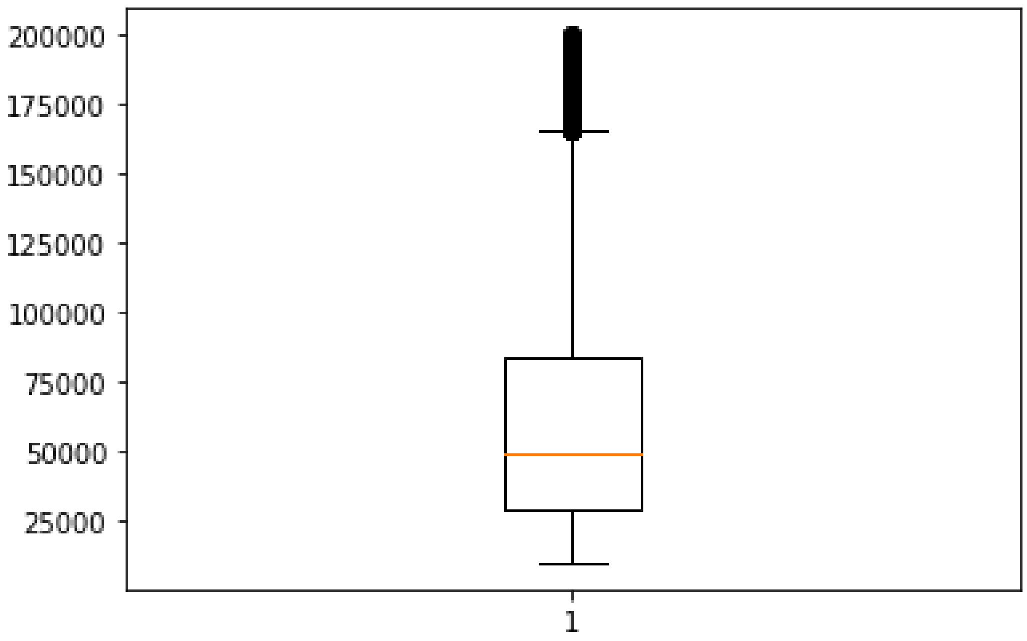
Education is normally distributed without any outliers.

```
plt.boxplot(dataset3.JobLevel)
Out[34]:
{'whiskers': [<matplotlib.lines.Line2D at 0x2877ec31cc8>,
 <matplotlib.lines.Line2D at 0x2877ec31108>],
 'caps': [<matplotlib.lines.Line2D at 0x2877edde3c8>,
 <matplotlib.lines.Line2D at 0x2877edde748>],
 'boxes': [<matplotlib.lines.Line2D at 0x2877ec315c8>],
 'medians': [<matplotlib.lines.Line2D at 0x2877ee67b88>],
 'fliers': [<matplotlib.lines.Line2D at 0x2877ee67648>],
 'means': []}
```



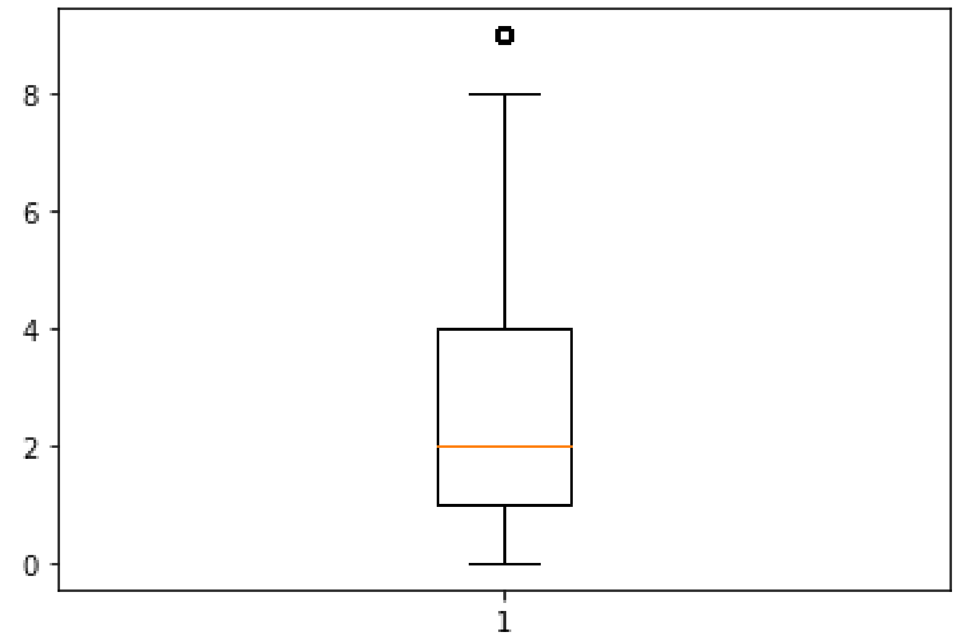
JobLevel is normally distributed without any outliers.

```
plt.boxplot(dataset3.MonthlyIncome)

Out[35]:

{'whiskers': [<matplotlib.lines.Line2D at 0x2877ef26548>,

  <matplotlib.lines.Line2D at 0x2877eef1d08>],

 'caps': [<matplotlib.lines.Line2D at 0x2877eef1888>,

  <matplotlib.lines.Line2D at 0x2877eeb7c48>],

 'boxes': [<matplotlib.lines.Line2D at 0x2877ef26208>],

 'medians': [<matplotlib.lines.Line2D at 0x2877eeb7688>],

 'fliers': [<matplotlib.lines.Line2D at 0x2877ee582c8>],

 'means': []}
```



MonthlyIncome is Right Skewed with several outliers. To remove outliers restrict the Monthly Income to 160000

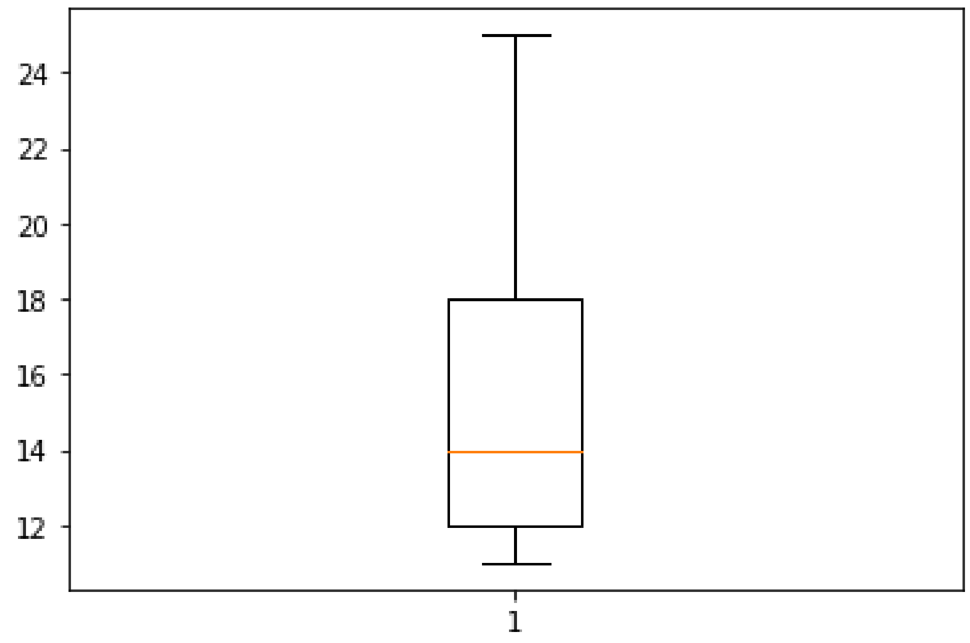
```
plt.boxplot(dataset3.NumCompaniesWorked)

Out[36]:

{'whiskers': [<matplotlib.lines.Line2D at 0x2877eec7348>,
  <matplotlib.lines.Line2D at 0x2877ec4ee08>],
 'caps': [<matplotlib.lines.Line2D at 0x2877ef60608>,
  <matplotlib.lines.Line2D at 0x2877ef60bc8>],
 'boxes': [<matplotlib.lines.Line2D at 0x2877ec4e888>],
 'medians': [<matplotlib.lines.Line2D at 0x2877ec48548>],
 'fliers': [<matplotlib.lines.Line2D at 0x2877ec48b48>],
 'means': []}
```



NumCompaniesWorked is Right Skewed with few outliers.

```
plt.boxplot(dataset3.PercentSalaryHike)
```

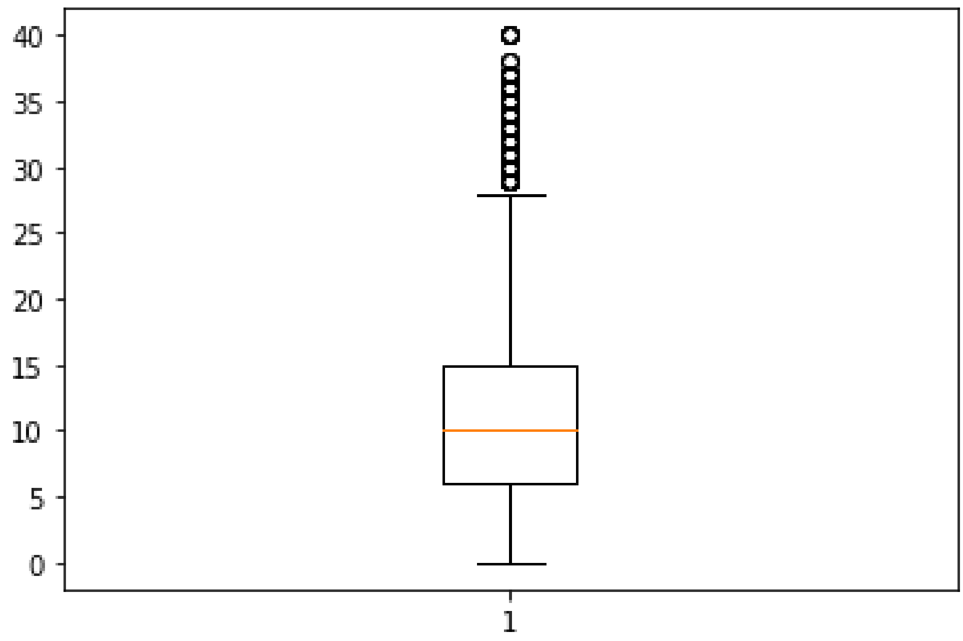
```
Out[37]:

{'whiskers': [<matplotlib.lines.Line2D at 0x2877ed683c8>,
  <matplotlib.lines.Line2D at 0x2877ed686c8>],
 'caps': [<matplotlib.lines.Line2D at 0x2877ed68cc8>,
  <matplotlib.lines.Line2D at 0x2877ec36e88>],
 'boxes': [<matplotlib.lines.Line2D at 0x2877ed5bac8>],
 'medians': [<matplotlib.lines.Line2D at 0x2877ec360c8>],
 'fliers': [<matplotlib.lines.Line2D at 0x2877ec28bc8>],
 'means': []}
```



PercentSalaryHike is Right Skewed without any outliers.

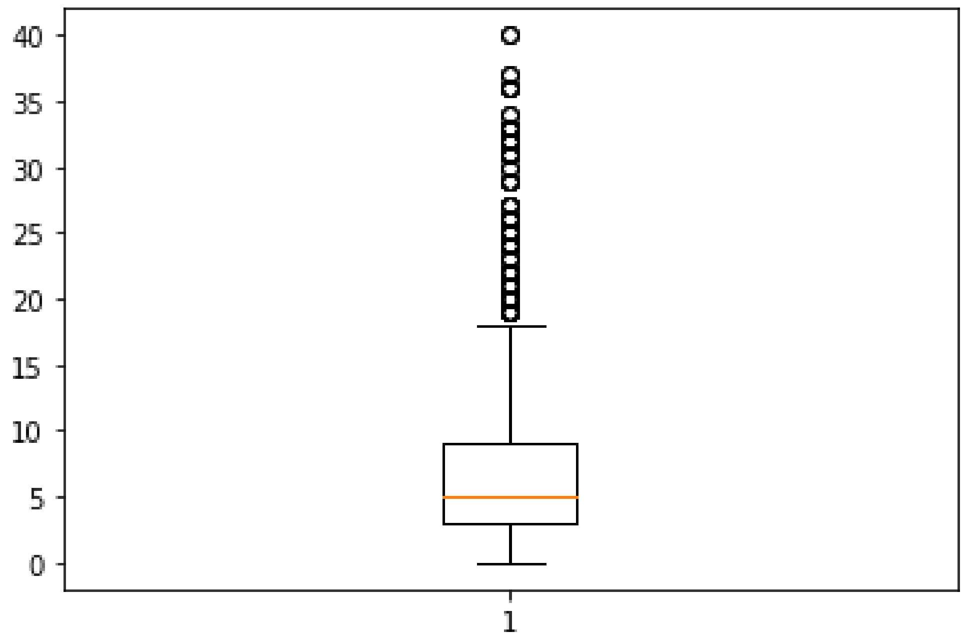
```
plt.boxplot(dataset3.TotalWorkingYears)
```

```
Out[38]:
{'whiskers': [<matplotlib.lines.Line2D at 0x2877ecacf88>,
  <matplotlib.lines.Line2D at 0x2877ecbd488>],
 'caps': [<matplotlib.lines.Line2D at 0x2877ecbda88>,
  <matplotlib.lines.Line2D at 0x2877ecc0bc8>],
 'boxes': [<matplotlib.lines.Line2D at 0x2877ecaca48>],
 'medians': [<matplotlib.lines.Line2D at 0x2877ecc0608>],
 'fliers': [<matplotlib.lines.Line2D at 0x2877ecc63c8>],
 'means': []}
```



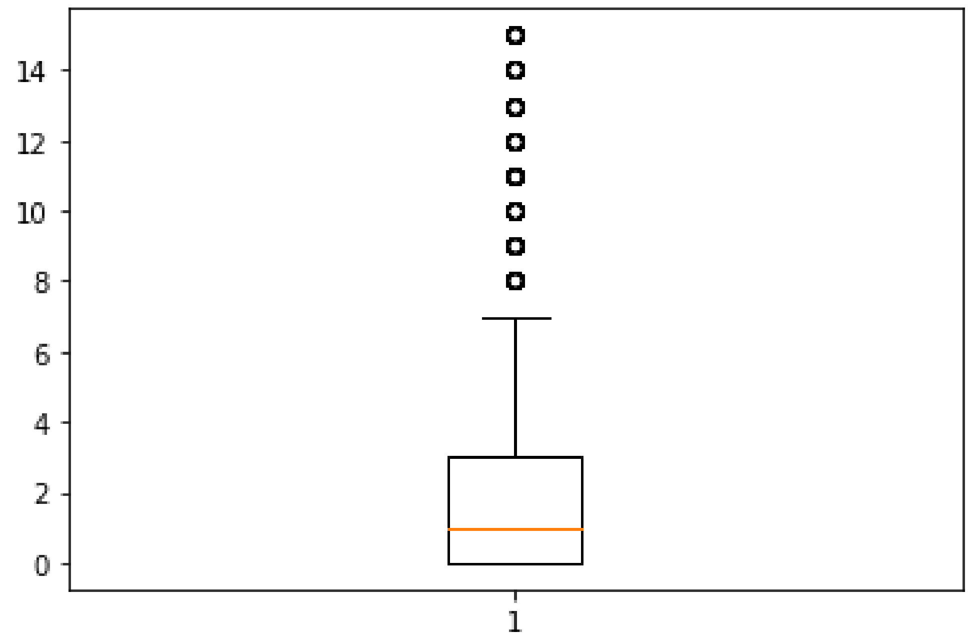
TotalWorkingYears is normally distributed with several outliers.

```
plt.boxplot(dataset3.YearsAtCompany)

Out[39]:

{'whiskers': [<matplotlib.lines.Line2D at 0x2877ed3a6c8>,

  <matplotlib.lines.Line2D at 0x2877ed3a9c8>],

 'caps': [<matplotlib.lines.Line2D at 0x2877ebdd288>,

  <matplotlib.lines.Line2D at 0x2877ebdd948>],

 'boxes': [<matplotlib.lines.Line2D at 0x2877ed33e48>],

 'medians': [<matplotlib.lines.Line2D at 0x2877ebdddc8>],

 'fliers': [<matplotlib.lines.Line2D at 0x2877ebc08c8>],

 'means': []}
```



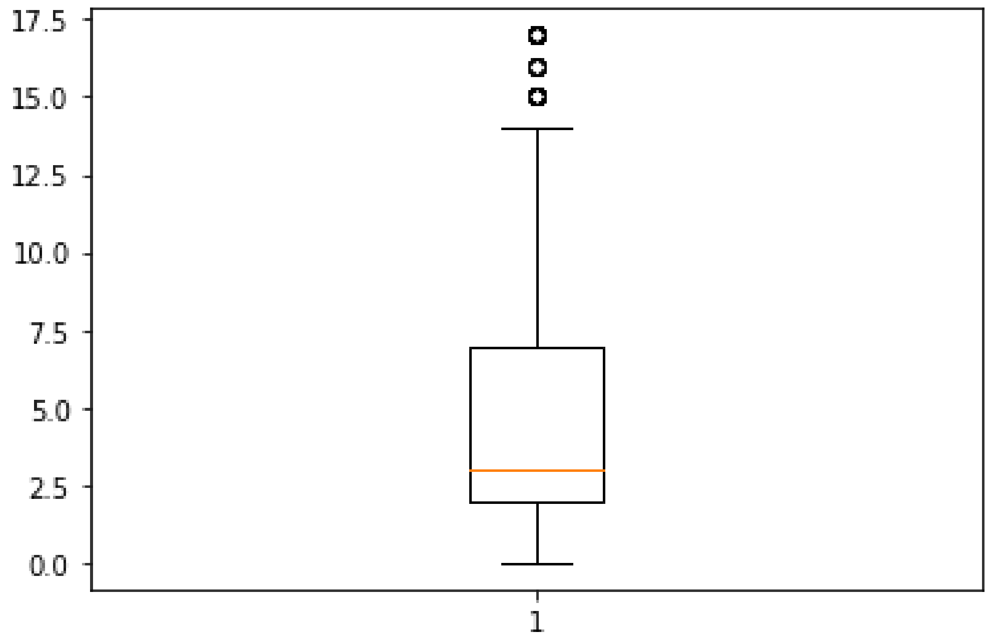
YearsAtCompany is Right Skewed with several outliers.

```
plt.boxplot(dataset3.YearsSinceLastPromotion)

Out[40]:

{'whiskers': [<matplotlib.lines.Line2D at 0x2877ebb6988>,

  <matplotlib.lines.Line2D at 0x2877ebb6ac8>],

 'caps': [<matplotlib.lines.Line2D at 0x2877eb9e448>,

  <matplotlib.lines.Line2D at 0x2877eb9ea48>],

 'boxes': [<matplotlib.lines.Line2D at 0x2877ebb6188>],

 'medians': [<matplotlib.lines.Line2D at 0x2877eb7e648>],

 'fliers': [<matplotlib.lines.Line2D at 0x2877eb7e208>],

 'means': []}
```



YearsSinceLastPromotion is Right Skewed with several outliers.

```
plt.boxplot(dataset3.YearsWithCurrManager)

Out[41]:

{'whiskers': [<matplotlib.lines.Line2D at 0x2877eb12208>,

  <matplotlib.lines.Line2D at 0x2877eb12508>],

 'caps': [<matplotlib.lines.Line2D at 0x2877eb12ac8>,

  <matplotlib.lines.Line2D at 0x2877eb3a488>],

 'boxes': [<matplotlib.lines.Line2D at 0x2877eb38a08>],

 'medians': [<matplotlib.lines.Line2D at 0x2877eb3aa48>],

 'fliers': [<matplotlib.lines.Line2D at 0x2877eb03408>],

 'means': []}
```



YearsWithCurrManager is Right Skewed with few outliers.