```
# -*- coding: utf-8 -*-
"""
Spyder Editor

This is a temporary script file.
"""

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt


dataset1= pd.read_excel('data_dictionary (1).xlsx',sheet_name=0)

mycsv = pd.read_csv("general_data (1).csv")

mycsv.head()
"""
mycsv.head()
Out[10]:
   Age        ...         YearsWithCurrManager
0   51        ...                            0
1   31        ...                            4
2   32        ...                            3
3   38        ...                            5
4   32        ...                            4

"""



mycsv.coloumns

"""


mycsv.columns
Out[14]:
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')

"""

mycsv.isnull()

"""
mycsv.isnull()
Out[15]:
      Age      ...     YearsWithCurrManager
0     False     ...                 False
1     False     ...                 False
2     False     ...                 False
3     False     ...                 False
4     False     ...                 False
5     False     ...                 False
6     False     ...                 False

"""

mycsv.duplicate()



"""
   ...      ...                 ...
4380  False     ...                 False
4381  False     ...                 False
4382  False     ...                 False
4383  False     ...                 False
4384  False     ...                 False
4385  False     ...                 False
4386  False     ...                 False
4387  False     ...                 False

"""

mycsv.drop_duplicates()

"""

mycsv.drop_duplicates()
```

```
Out[17]:
    Age      ...       YearsWithCurrManager
0     51       ...                    0
1     31       ...                    4
2     32       ...                    3
3     38       ...                    5
4     32       ...                    4
5     46       ...                    7
6     28       ...                    0

"""
dataset2=mycsv[['Age','DistanceFromHome','Education','MonthlyIncome',
'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear',
'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].describe()


"""

Age     DistanceFromHome   Education       MonthlyIncome          NumCompaniesWorked
PercentSalaryHike      TotalWorkingYears    TrainingTimesLastYear          YearsAtCompany
YearsSinceLastPromotion      YearsWithCurrManager
count   4410.0 4410.0 4410.0 4410.0 4391.0 4410.0 4401.0 4410.0 4410.0 4410.0 4410.0
mean    36.923809523809524 9.19251700680272     2.912925170068027   65029.31292517007
2.6948303347756775 15.209523809523809 11.279936378095888 2.7993197278911564
7.0081632653061225 2.1877551020408164 4.12312925170068
std     9.133301271011184   8.10502551890526     1.0239326286269606 47068.88855947343
2.498886888807146   3.6591075162983544 7.782222140911688   1.288978169704257
6.125135444967677   3.2216993206893245 3.5673267440708067
min     18.0    1.0     1.0     10090.0         0.0     11.0    0.0     0.0     0.0     0.0     0.0
25%     30.0    2.0     2.0     29110.0         1.0     12.0    6.0     2.0     3.0     0.0     2.0
50%     36.0    7.0     3.0     49190.0         2.0     14.0    10.0    3.0     5.0     1.0     3.0
75%     43.0    14.0    4.0     83800.0         4.0     18.0    15.0    3.0     9.0     3.0     7.0
max     60.0    29.0    5.0     199990.0        9.0     25.0    40.0    6.0     40.0    15.0    17.0


"""

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

"""
        0
```

```
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

"""
```

```
dataset4=mycsv[['Age','DistanceFromHome','Education','MonthlyIncome',
'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear',
'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].mean()


"""

	0
Age	36.923809523809524
DistanceFromHome	9.19251700680272
Education	2.912925170068027
MonthlyIncome		65029.31292517007
NumCompaniesWorked		2.6948303347756775
PercentSalaryHike	15.209523809523809
TotalWorkingYears	11.279936378095888
TrainingTimesLastYear		2.7993197278911564
YearsAtCompany	7.0081632653061225
YearsSinceLastPromotion	2.1877551020408164
YearsWithCurrManager		4.12312925170068

"""
```

```
dataset4=mycsv[['Age','DistanceFromHome','Education','MonthlyIncome',
'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear',
'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].mode()
```

```
"""
         Age     DistanceFromHome    Education       MonthlyIncome
NumCompaniesWorked        PercentSalaryHike     TotalWorkingYears
TrainingTimesLastYear       YearsAtCompany       YearsSinceLastPromotion
YearsWithCurrManager
0         35       2         3         23420  1.0       11        10.0     2         5         0         2


"""
```

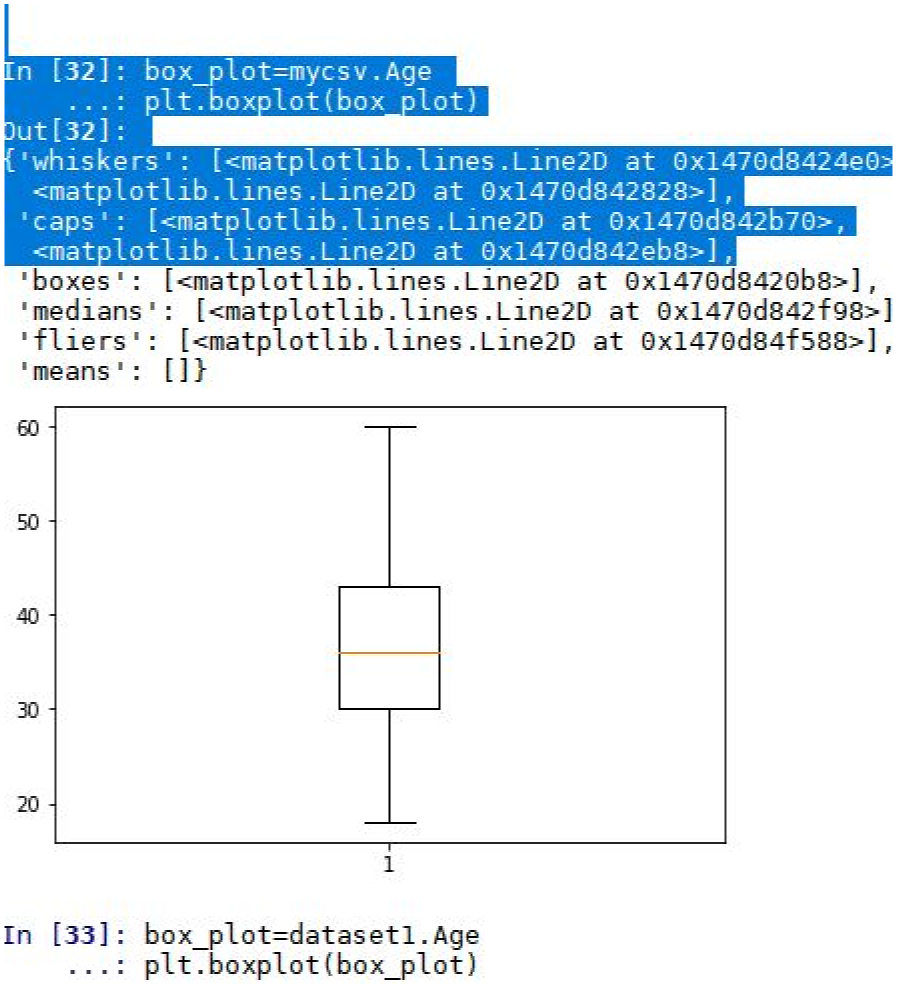
```
dataset6=mycsv[['Age','DistanceFromHome','Education','MonthlyIncome',
'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear',
'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].var()
```

```
"""
         0
Age      83.41719210705452
DistanceFromHome    65.69143866210547
Education       1.048438027966917
MonthlyIncome          2215480270.2241287
NumCompaniesWorked        6.244435683052258
PercentSalaryHike     13.389067815831112
TotalWorkingYears     60.56298145049609
TrainingTimesLastYear        1.6614647219741365
YearsAtCompany       37.51728421919939
YearsSinceLastPromotion     10.379346512930056
YearsWithCurrManager        12.725820098962824

"""
```

```
dataset7=mycsv[['Age','DistanceFromHome','Education','MonthlyIncome',
'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear',
'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].skew()
```

```
"""
         0
Age      0.41300495269768406
DistanceFromHome    0.9574657463788941
Education       -0.2894838784116763
```

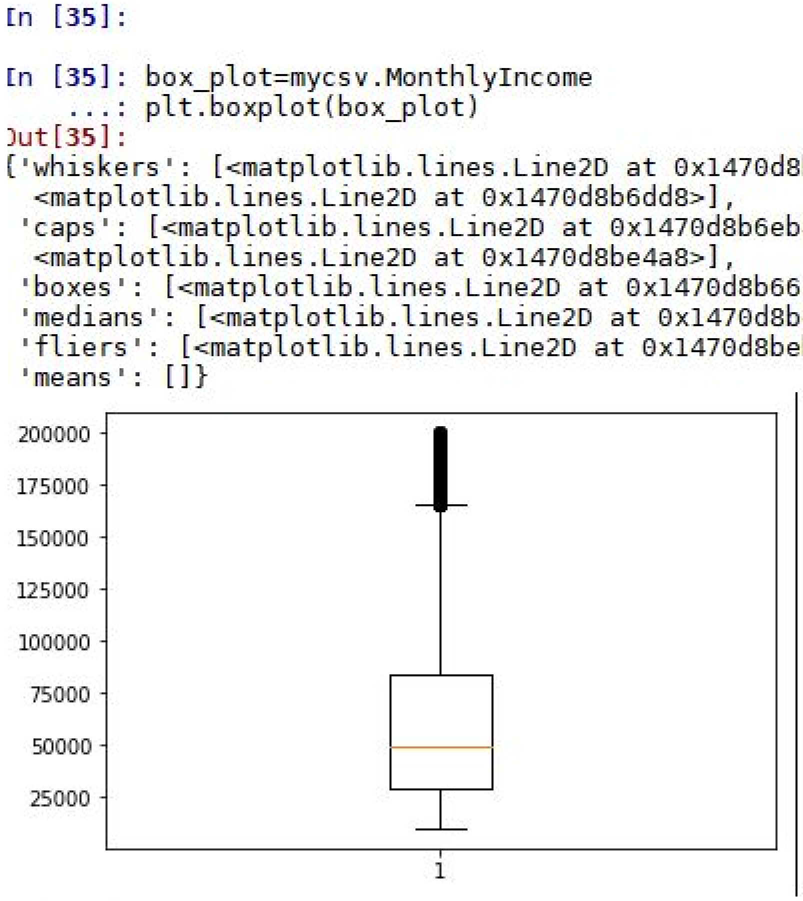
```
MonthlyIncome          1.3688841631898667
NumCompaniesWorked         1.0267666759708942
PercentSalaryHike      0.8205689837508037
TotalWorkingYears      1.1168317963678807
TrainingTimesLastYear          0.5527476257400273
YearsAtCompany       1.7633282316663832
YearsSinceLastPromotion      1.9829391562991707
YearsWithCurrManager         0.8328836111367132

"""



dataset8=mycsv[['Age','DistanceFromHome','Education','MonthlyIncome',
'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear',
'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].kurt()

"""
          0
Age     -0.4059505398497185
DistanceFromHome   -0.227045354876517
Education       -0.5605690113243802
MonthlyIncome          1.0002318550155116
NumCompaniesWorked         0.007287480878091834
PercentSalaryHike     -0.30263839310442986
TotalWorkingYears     0.9129359960798036
TrainingTimesLastYear          0.49114899850172034
YearsAtCompany       3.9238642054012636
YearsSinceLastPromotion      3.6017605183177106
YearsWithCurrManager         0.16794854278413895

"""


"""
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
"""
```



```
In [32]: box_plot=mycsv.Age
    ...: plt.boxplot(box_plot)
Out[32]:
{'whiskers': [<matplotlib.lines.Line2D at 0x1470d8424e0>
  <matplotlib.lines.Line2D at 0x1470d842828>],
 'caps': [<matplotlib.lines.Line2D at 0x1470d842b70>,
  <matplotlib.lines.Line2D at 0x1470d842eb8>],
 'boxes': [<matplotlib.lines.Line2D at 0x1470d8420b8>],
 'medians': [<matplotlib.lines.Line2D at 0x1470d842f98>]
 'fliers': [<matplotlib.lines.Line2D at 0x1470d84f588>],
 'means': []}
```

```
In [33]: box_plot=dataset1.Age
    ...: plt.boxplot(box_plot)
```

Age is mesocritic with skew is 0



```
[n [35]:

[n [35]: box_plot=mycsv.MonthlyIncome
    ...: plt.boxplot(box_plot)
)ut[35]:
{'whiskers': [<matplotlib.lines.Line2D at 0x1470d8b6a90>,
  <matplotlib.lines.Line2D at 0x1470d8b6dd8>],
 'caps': [<matplotlib.lines.Line2D at 0x1470d8b6eb8>,
  <matplotlib.lines.Line2D at 0x1470d8be4a8>],
 'boxes': [<matplotlib.lines.Line2D at 0x1470d8b66a0>],
 'medians': [<matplotlib.lines.Line2D at 0x1470d8be7f0>],
 'fliers': [<matplotlib.lines.Line2D at 0x1470d8beb38>],
 'means': []}
```

Monthly income is also right skewed with lot of outlayers

```
n [36]: box_plot=mycsv.YearsAtCompany
   ...: plt.boxplot(box_plot)
ut[36]:
'whiskers': [<matplotlib.lines.Line2D at 0x1470d915d30>,
 <matplotlib.lines.Line2D at 0x1470d915e10>],
'caps': [<matplotlib.lines.Line2D at 0x1470d91e400>,
 <matplotlib.lines.Line2D at 0x1470d91e748>],
'boxes': [<matplotlib.lines.Line2D at 0x1470d915940>],
'medians': [<matplotlib.lines.Line2D at 0x1470d91ea90>],
'fliers': [<matplotlib.lines.Line2D at 0x1470d91edd8>],
'means': []}
```

```
n [37]: |
```

Years is right skewed with lot of outlyers