**Importing modules and loading dataset**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

dataset = pd.read\_csv('general\_data.csv')

dataset.head()

Out[5]:

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]

**Data Pre-processing**

dataset.isnull()

Out[6]:

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[7]:

0 False

1 False

2 False

3 False

4 False

4405 False

4406 False

4407 False

4408 False

4409 False

Length: 4410, dtype: bool

**Univariate Analysis**

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

ds

Out[9]:

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

[8 rows x 11 columns]

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

ds

Out[11]:

Age DistanceFromHome ... YearsSinceLastPromotion YearsWithCurrManager

0 35 2 ... 0 2

[1 rows x 11 columns]

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

ds

Out[13]:

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

dtype: float64

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

ds

Out[15]:

Age 9.133301

DistanceFromHome 8.105026

Education 1.023933

MonthlyIncome 47068.888559

NumCompaniesWorked 2.498887

PercentSalaryHike 3.659108

TotalWorkingYears 7.782222

TrainingTimesLastYear 1.288978

YearsAtCompany 6.125135

YearsSinceLastPromotion 3.221699

YearsWithCurrManager 3.567327

dtype: float64

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

ds

Out[17]:

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

dtype: float64

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

ds

Out[19]:

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

dtype: float64

**Outliers**

plt.boxplot(dataset.Age)

Out[20]:

{'whiskers': [<matplotlib.lines.Line2D at 0x2290d630948>,

<matplotlib.lines.Line2D at 0x2290d630f08>],

'caps': [<matplotlib.lines.Line2D at 0x2290d630b08>,

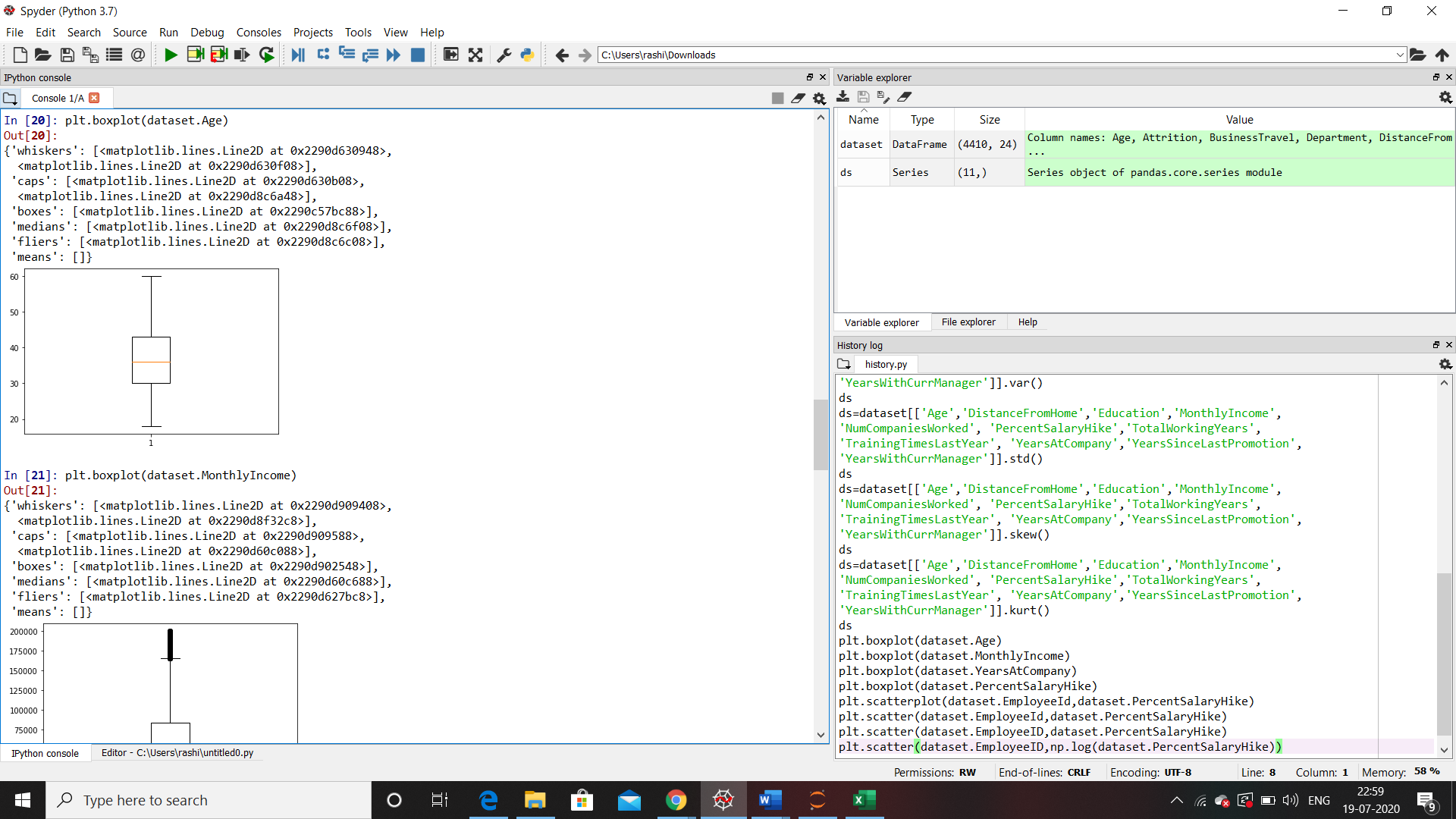
<matplotlib.lines.Line2D at 0x2290d8c6a48>],

'boxes': [<matplotlib.lines.Line2D at 0x2290c57bc88>],

'medians': [<matplotlib.lines.Line2D at 0x2290d8c6f08>],

'fliers': [<matplotlib.lines.Line2D at 0x2290d8c6c08>],

'means': []}



plt.boxplot(dataset.MonthlyIncome)

Out[21]:

{'whiskers': [<matplotlib.lines.Line2D at 0x2290d909408>,

<matplotlib.lines.Line2D at 0x2290d8f32c8>],

'caps': [<matplotlib.lines.Line2D at 0x2290d909588>,

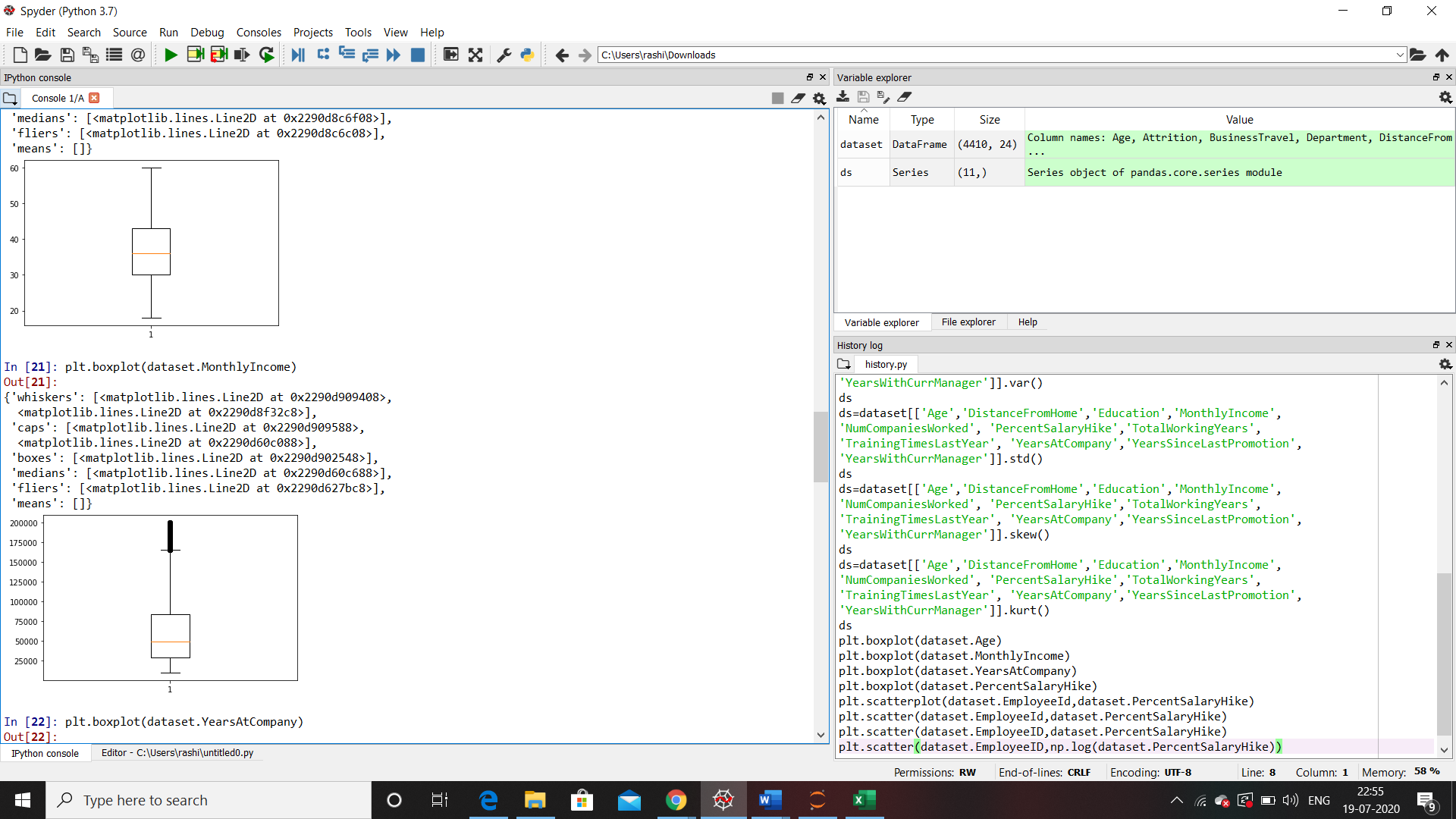
<matplotlib.lines.Line2D at 0x2290d60c088>],

'boxes': [<matplotlib.lines.Line2D at 0x2290d902548>],

'medians': [<matplotlib.lines.Line2D at 0x2290d60c688>],

'fliers': [<matplotlib.lines.Line2D at 0x2290d627bc8>],

'means': []}



plt.boxplot(dataset.YearsAtCompany)

Out[22]:

{'whiskers': [<matplotlib.lines.Line2D at 0x2290d9aff88>,

<matplotlib.lines.Line2D at 0x2290d9afc08>],

'caps': [<matplotlib.lines.Line2D at 0x2290d9b2f48>,

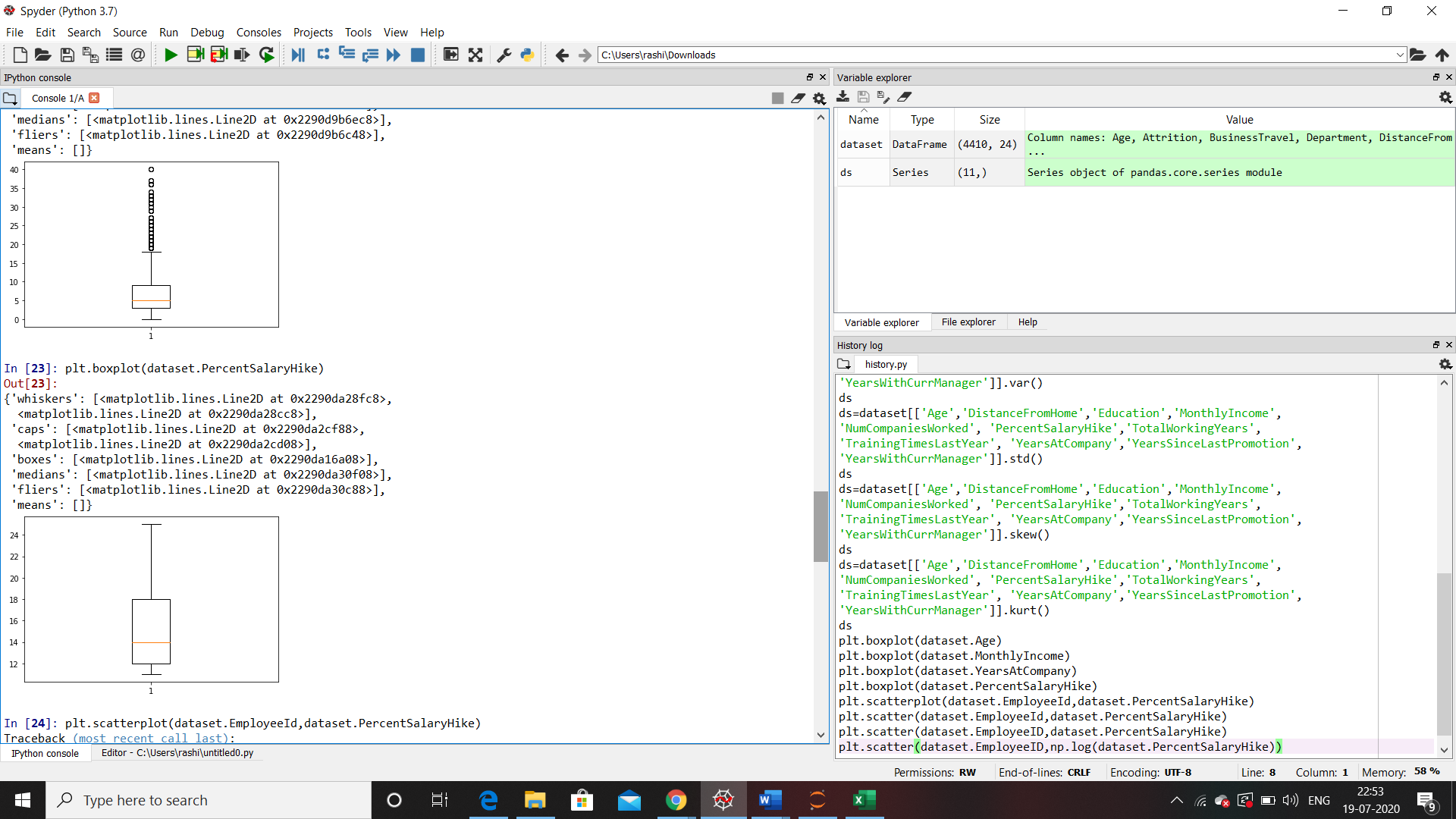
<matplotlib.lines.Line2D at 0x2290d9b2cc8>],

'boxes': [<matplotlib.lines.Line2D at 0x2290d99d948>],

'medians': [<matplotlib.lines.Line2D at 0x2290d9b6ec8>],

'fliers': [<matplotlib.lines.Line2D at 0x2290d9b6c48>],

'means': []}



plt.boxplot(dataset.PercentSalaryHike)

Out[23]:

{'whiskers': [<matplotlib.lines.Line2D at 0x2290da28fc8>,

<matplotlib.lines.Line2D at 0x2290da28cc8>],

'caps': [<matplotlib.lines.Line2D at 0x2290da2cf88>,

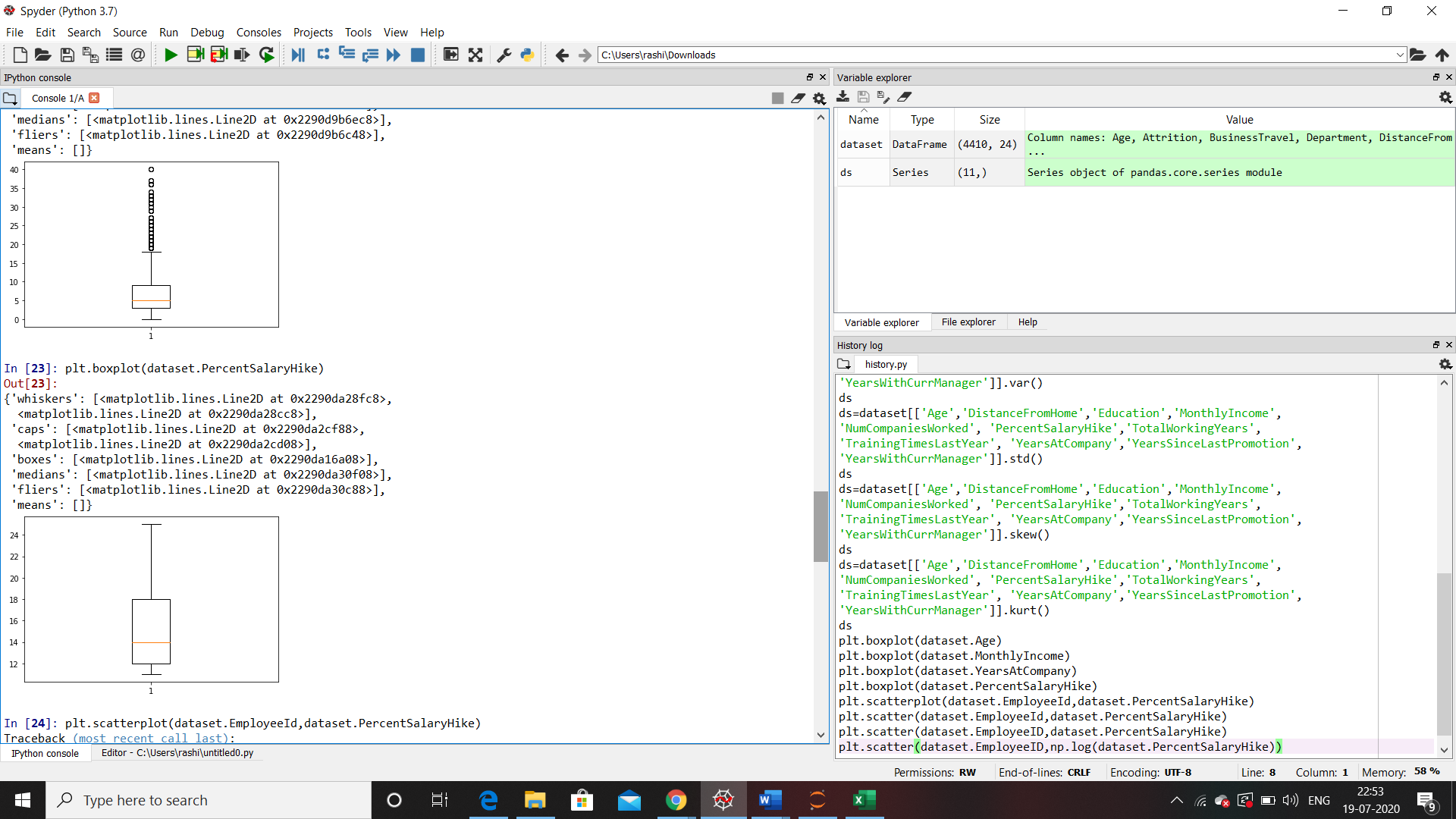
<matplotlib.lines.Line2D at 0x2290da2cd08>],

'boxes': [<matplotlib.lines.Line2D at 0x2290da16a08>],

'medians': [<matplotlib.lines.Line2D at 0x2290da30f08>],

'fliers': [<matplotlib.lines.Line2D at 0x2290da30c88>],

'means': []}



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 .