Python 3.7.6 (default, Jan 8 2020, 16:21:45) [MSC v.1916 32 bit (Intel)]

Type "copyright", "credits" or "license" for more information.

IPython 7.12.0 -- An enhanced Interactive Python.

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

dataset1=pd.read\_excel("general\_data.csv",sheet\_name=0)

Traceback (most recent call last):

File "<ipython-input-4-fbfafd74a502>", line 1, in <module>

dataset1=pd.read\_excel("general\_data.csv",sheet\_name=0)

File "C:\ProgramData\Anaconda3\lib\site-packages\pandas\io\excel\\_base.py", line 304, in read\_excel

io = ExcelFile(io, engine=engine)

File "C:\ProgramData\Anaconda3\lib\site-packages\pandas\io\excel\\_base.py", line 821, in \_\_init\_\_

self.\_reader = self.\_engines[engine](self.\_io)

File "C:\ProgramData\Anaconda3\lib\site-packages\pandas\io\excel\\_xlrd.py", line 21, in \_\_init\_\_

super().\_\_init\_\_(filepath\_or\_buffer)

File "C:\ProgramData\Anaconda3\lib\site-packages\pandas\io\excel\\_base.py", line 353, in \_\_init\_\_

self.book = self.load\_workbook(filepath\_or\_buffer)

File "C:\ProgramData\Anaconda3\lib\site-packages\pandas\io\excel\\_xlrd.py", line 36, in load\_workbook

return open\_workbook(filepath\_or\_buffer)

File "C:\ProgramData\Anaconda3\lib\site-packages\xlrd\\_\_init\_\_.py", line 157, in open\_workbook

ragged\_rows=ragged\_rows,

File "C:\ProgramData\Anaconda3\lib\site-packages\xlrd\book.py", line 92, in open\_workbook\_xls

biff\_version = bk.getbof(XL\_WORKBOOK\_GLOBALS)

File "C:\ProgramData\Anaconda3\lib\site-packages\xlrd\book.py", line 1278, in getbof

bof\_error('Expected BOF record; found %r' % self.mem[savpos:savpos+8])

File "C:\ProgramData\Anaconda3\lib\site-packages\xlrd\book.py", line 1272, in bof\_error

raise XLRDError('Unsupported format, or corrupt file: ' + msg)

XLRDError: Unsupported format, or corrupt file: Expected BOF record; found b'Age,Attr'

dataset1=pd.read\_excel("general\_data.csv",sheet\_name=0)

Traceback (most recent call last):

File "<ipython-input-5-fbfafd74a502>", line 1, in <module>

dataset1=pd.read\_excel("general\_data.csv",sheet\_name=0)

File "C:\ProgramData\Anaconda3\lib\site-packages\pandas\io\excel\\_base.py", line 304, in read\_excel

io = ExcelFile(io, engine=engine)

File "C:\ProgramData\Anaconda3\lib\site-packages\pandas\io\excel\\_base.py", line 821, in \_\_init\_\_

self.\_reader = self.\_engines[engine](self.\_io)

File "C:\ProgramData\Anaconda3\lib\site-packages\pandas\io\excel\\_xlrd.py", line 21, in \_\_init\_\_

super().\_\_init\_\_(filepath\_or\_buffer)

File "C:\ProgramData\Anaconda3\lib\site-packages\pandas\io\excel\\_base.py", line 353, in \_\_init\_\_

self.book = self.load\_workbook(filepath\_or\_buffer)

File "C:\ProgramData\Anaconda3\lib\site-packages\pandas\io\excel\\_xlrd.py", line 36, in load\_workbook

return open\_workbook(filepath\_or\_buffer)

File "C:\ProgramData\Anaconda3\lib\site-packages\xlrd\\_\_init\_\_.py", line 157, in open\_workbook

ragged\_rows=ragged\_rows,

File "C:\ProgramData\Anaconda3\lib\site-packages\xlrd\book.py", line 92, in open\_workbook\_xls

biff\_version = bk.getbof(XL\_WORKBOOK\_GLOBALS)

File "C:\ProgramData\Anaconda3\lib\site-packages\xlrd\book.py", line 1278, in getbof

bof\_error('Expected BOF record; found %r' % self.mem[savpos:savpos+8])

File "C:\ProgramData\Anaconda3\lib\site-packages\xlrd\book.py", line 1272, in bof\_error

raise XLRDError('Unsupported format, or corrupt file: ' + msg)

XLRDError: Unsupported format, or corrupt file: Expected BOF record; found b'Age,Attr'

dataset1=pd.read\_excel("general\_data",sheet\_name=0)

Traceback (most recent call last):

File "<ipython-input-6-6b0475f8189b>", line 1, in <module>

dataset1=pd.read\_excel("general\_data",sheet\_name=0)

File "C:\ProgramData\Anaconda3\lib\site-packages\pandas\io\excel\\_base.py", line 304, in read\_excel

io = ExcelFile(io, engine=engine)

File "C:\ProgramData\Anaconda3\lib\site-packages\pandas\io\excel\\_base.py", line 821, in \_\_init\_\_

self.\_reader = self.\_engines[engine](self.\_io)

File "C:\ProgramData\Anaconda3\lib\site-packages\pandas\io\excel\\_xlrd.py", line 21, in \_\_init\_\_

super().\_\_init\_\_(filepath\_or\_buffer)

File "C:\ProgramData\Anaconda3\lib\site-packages\pandas\io\excel\\_base.py", line 353, in \_\_init\_\_

self.book = self.load\_workbook(filepath\_or\_buffer)

File "C:\ProgramData\Anaconda3\lib\site-packages\pandas\io\excel\\_xlrd.py", line 36, in load\_workbook

return open\_workbook(filepath\_or\_buffer)

File "C:\ProgramData\Anaconda3\lib\site-packages\xlrd\\_\_init\_\_.py", line 111, in open\_workbook

with open(filename, "rb") as f:

FileNotFoundError: [Errno 2] No such file or directory: 'general\_data'

dataset1=pd.read\_excel("general\_data.csv",sheet\_name=1)

Traceback (most recent call last):

File "<ipython-input-7-1a984a59ea5c>", line 1, in <module>

dataset1=pd.read\_excel("general\_data.csv",sheet\_name=1)

File "C:\ProgramData\Anaconda3\lib\site-packages\pandas\io\excel\\_base.py", line 304, in read\_excel

io = ExcelFile(io, engine=engine)

File "C:\ProgramData\Anaconda3\lib\site-packages\pandas\io\excel\\_base.py", line 821, in \_\_init\_\_

self.\_reader = self.\_engines[engine](self.\_io)

File "C:\ProgramData\Anaconda3\lib\site-packages\pandas\io\excel\\_xlrd.py", line 21, in \_\_init\_\_

super().\_\_init\_\_(filepath\_or\_buffer)

File "C:\ProgramData\Anaconda3\lib\site-packages\pandas\io\excel\\_base.py", line 353, in \_\_init\_\_

self.book = self.load\_workbook(filepath\_or\_buffer)

File "C:\ProgramData\Anaconda3\lib\site-packages\pandas\io\excel\\_xlrd.py", line 36, in load\_workbook

return open\_workbook(filepath\_or\_buffer)

File "C:\ProgramData\Anaconda3\lib\site-packages\xlrd\\_\_init\_\_.py", line 157, in open\_workbook

ragged\_rows=ragged\_rows,

File "C:\ProgramData\Anaconda3\lib\site-packages\xlrd\book.py", line 92, in open\_workbook\_xls

biff\_version = bk.getbof(XL\_WORKBOOK\_GLOBALS)

File "C:\ProgramData\Anaconda3\lib\site-packages\xlrd\book.py", line 1278, in getbof

bof\_error('Expected BOF record; found %r' % self.mem[savpos:savpos+8])

File "C:\ProgramData\Anaconda3\lib\site-packages\xlrd\book.py", line 1272, in bof\_error

raise XLRDError('Unsupported format, or corrupt file: ' + msg)

XLRDError: Unsupported format, or corrupt file: Expected BOF record; found b'Age,Attr'

dataset1=pd.read\_excel("general\_data (1).csv",sheet\_name=1)

Traceback (most recent call last):

File "<ipython-input-8-3e63fbab9e7e>", line 1, in <module>

dataset1=pd.read\_excel("general\_data (1).csv",sheet\_name=1)

File "C:\ProgramData\Anaconda3\lib\site-packages\pandas\io\excel\\_base.py", line 304, in read\_excel

io = ExcelFile(io, engine=engine)

File "C:\ProgramData\Anaconda3\lib\site-packages\pandas\io\excel\\_base.py", line 821, in \_\_init\_\_

self.\_reader = self.\_engines[engine](self.\_io)

File "C:\ProgramData\Anaconda3\lib\site-packages\pandas\io\excel\\_xlrd.py", line 21, in \_\_init\_\_

super().\_\_init\_\_(filepath\_or\_buffer)

File "C:\ProgramData\Anaconda3\lib\site-packages\pandas\io\excel\\_base.py", line 353, in \_\_init\_\_

self.book = self.load\_workbook(filepath\_or\_buffer)

File "C:\ProgramData\Anaconda3\lib\site-packages\pandas\io\excel\\_xlrd.py", line 36, in load\_workbook

return open\_workbook(filepath\_or\_buffer)

File "C:\ProgramData\Anaconda3\lib\site-packages\xlrd\\_\_init\_\_.py", line 157, in open\_workbook

ragged\_rows=ragged\_rows,

File "C:\ProgramData\Anaconda3\lib\site-packages\xlrd\book.py", line 92, in open\_workbook\_xls

biff\_version = bk.getbof(XL\_WORKBOOK\_GLOBALS)

File "C:\ProgramData\Anaconda3\lib\site-packages\xlrd\book.py", line 1278, in getbof

bof\_error('Expected BOF record; found %r' % self.mem[savpos:savpos+8])

File "C:\ProgramData\Anaconda3\lib\site-packages\xlrd\book.py", line 1272, in bof\_error

raise XLRDError('Unsupported format, or corrupt file: ' + msg)

XLRDError: Unsupported format, or corrupt file: Expected BOF record; found b'Age,Attr'

dataset1=pd.read\_excel("machine assign.xlsx",sheet\_name=0)

dataset1

Out[10]:

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]

dataset1.head()

Out[11]:

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]

dataset1.columns

Out[12]:

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')

Index(['Variable', 'Meaning', 'Levels'], dtype='object')

Traceback (most recent call last):

File "<ipython-input-13-c0535564bb97>", line 1, in <module>

Index(['Variable', 'Meaning', 'Levels'], dtype='object')

NameError: name 'Index' is not defined

dataset1.isnull()

Out[14]:

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]

dataset1.duplicated()

Out[15]:

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.drop\_duplicates()

Out[16]:

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]

dataset3=dataset1[['Age', 'Attrition', 'BusinessTravel', 'Department', 'DistanceFromHome',

'Education', 'EducationField', 'EmployeeCount', 'EmployeeID', 'Gender',

'JobLevel', 'JobRole', 'MaritalStatus', 'MonthlyIncome',

'NumCompaniesWorked', 'Over18', 'PercentSalaryHike', 'StandardHours',

'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear',

'YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManager']].describe()

dataset3

dataset3

File "<ipython-input-17-36d22151df95>", line 9

dataset3

^

IndentationError: unexpected indent

dataset3=dataset1[['Age', 'Attrition', 'BusinessTravel', 'Department', 'DistanceFromHome',

'Education', 'EducationField', 'EmployeeCount', 'EmployeeID', 'Gender',

'JobLevel', 'JobRole', 'MaritalStatus', 'MonthlyIncome',

'NumCompaniesWorked', 'Over18', 'PercentSalaryHike', 'StandardHours',

'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear',

'YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManager']].describe()

dataset3

Out[18]:

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 16 columns]

dataset3=dataset1[['Age', 'Attrition', 'BusinessTravel', 'Department', 'DistanceFromHome',

'Education', 'EducationField', 'EmployeeCount', 'EmployeeID', 'Gender',

'JobLevel', 'JobRole', 'MaritalStatus', 'MonthlyIncome',

'NumCompaniesWorked', 'Over18', 'PercentSalaryHike', 'StandardHours',

'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear',

'YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManager']].median()

dataset3

Out[20]:

Age 36.0

DistanceFromHome 7.0

Education 3.0

EmployeeCount 1.0

EmployeeID 2205.5

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=dataset1[['Age', 'Attrition', 'BusinessTravel', 'Department', 'DistanceFromHome',

'Education', 'EducationField', 'EmployeeCount', 'EmployeeID', 'Gender',

'JobLevel', 'JobRole', 'MaritalStatus', 'MonthlyIncome',

'NumCompaniesWorked', 'Over18', 'PercentSalaryHike', 'StandardHours',

'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear',

'YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManager']].mode()

dataset3

Out[22]:

Age Attrition ... YearsSinceLastPromotion YearsWithCurrManager

0 35.0 No ... 0.0 2.0

1 NaN NaN ... NaN NaN

2 NaN NaN ... NaN NaN

3 NaN NaN ... NaN NaN

4 NaN NaN ... NaN NaN

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

4405 NaN NaN ... NaN NaN

4406 NaN NaN ... NaN NaN

4407 NaN NaN ... NaN NaN

4408 NaN NaN ... NaN NaN

4409 NaN NaN ... NaN NaN

[4410 rows x 24 columns]

dataset3=dataset1[['Age', 'Attrition', 'BusinessTravel', 'Department', 'DistanceFromHome',

'Education', 'EducationField', 'EmployeeCount', 'EmployeeID', 'Gender',

'JobLevel', 'JobRole', 'MaritalStatus', 'MonthlyIncome',

'NumCompaniesWorked', 'Over18', 'PercentSalaryHike', 'StandardHours',

'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear',

'YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManager']].var()

dataset3

Out[24]:

Age 8.341719e+01

DistanceFromHome 6.569144e+01

Education 1.048438e+00

EmployeeCount 0.000000e+00

EmployeeID 1.621042e+06

JobLevel 1.224760e+00

MonthlyIncome 2.215480e+09

NumCompaniesWorked 6.244436e+00

PercentSalaryHike 1.338907e+01

StandardHours 0.000000e+00

StockOptionLevel 7.257053e-01

TotalWorkingYears 6.056298e+01

TrainingTimesLastYear 1.661465e+00

YearsAtCompany 3.751728e+01

YearsSinceLastPromotion 1.037935e+01

YearsWithCurrManager 1.272582e+01

dtype: float64

dataset3=dataset1[['Age', 'Attrition', 'BusinessTravel', 'Department', 'DistanceFromHome',

'Education', 'EducationField', 'EmployeeCount', 'EmployeeID', 'Gender',

'JobLevel', 'JobRole', 'MaritalStatus', 'MonthlyIncome',

'NumCompaniesWorked', 'Over18', 'PercentSalaryHike', 'StandardHours',

'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear',

'YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManager']].skew()

dataset3

Out[26]:

Age 0.413005

DistanceFromHome 0.957466

Education -0.289484

EmployeeCount 0.000000

EmployeeID 0.000000

JobLevel 1.024703

MonthlyIncome 1.368884

NumCompaniesWorked 1.026767

PercentSalaryHike 0.820569

StandardHours 0.000000

StockOptionLevel 0.968321

TotalWorkingYears 1.116832

TrainingTimesLastYear 0.552748

YearsAtCompany 1.763328

YearsSinceLastPromotion 1.982939

YearsWithCurrManager 0.832884

dtype: float64

dataset3=dataset1[['Age', 'Attrition', 'BusinessTravel', 'Department', 'DistanceFromHome',

'Education', 'EducationField', 'EmployeeCount', 'EmployeeID', 'Gender',

'JobLevel', 'JobRole', 'MaritalStatus', 'MonthlyIncome',

'NumCompaniesWorked', 'Over18', 'PercentSalaryHike', 'StandardHours',

'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear',

'YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManager']].kurt()

dataset3

Out[28]:

Age -0.405951

DistanceFromHome -0.227045

Education -0.560569

EmployeeCount 0.000000

EmployeeID -1.200000

JobLevel 0.395525

MonthlyIncome 1.000232

NumCompaniesWorked 0.007287

PercentSalaryHike -0.302638

StandardHours 0.000000

StockOptionLevel 0.361086

TotalWorkingYears 0.912936

TrainingTimesLastYear 0.491149

YearsAtCompany 3.923864

YearsSinceLastPromotion 3.601761

YearsWithCurrManager 0.167949

dtype: float64

box\_plot=dataset1.Age

plt.boxplot(box\_plot)

Out[30]:

{'whiskers': [<matplotlib.lines.Line2D at 0xd59b70>,

<matplotlib.lines.Line2D at 0xd59f30>],

'caps': [<matplotlib.lines.Line2D at 0xd591b0>,

<matplotlib.lines.Line2D at 0xda5910>],

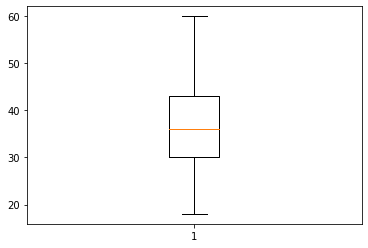
'boxes': [<matplotlib.lines.Line2D at 0xda5a10>],

'medians': [<matplotlib.lines.Line2D at 0xd76430>],

'fliers': [<matplotlib.lines.Line2D at 0xd762b0>],

'means': []}

Figures now render in the Plots pane by default. To make them also appear inline in the Console, uncheck "Mute Inline Plotting" under the Plots pane options menu.



box\_plot=dataset1.MonthlyIncome

plt.boxplot(box\_plot)

Out[32]:

{'whiskers': [<matplotlib.lines.Line2D at 0xde1050>,

<matplotlib.lines.Line2D at 0xde1d50>],

'caps': [<matplotlib.lines.Line2D at 0xe69250>,

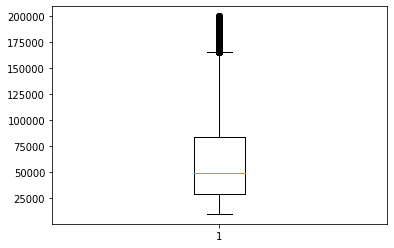
<matplotlib.lines.Line2D at 0xe694b0>],

'boxes': [<matplotlib.lines.Line2D at 0xde1030>],

'medians': [<matplotlib.lines.Line2D at 0xe69730>],

'fliers': [<matplotlib.lines.Line2D at 0xe69990>],

'means': []}



box\_plot=dataset1.YearsAtCompany

plt.boxplot(box\_plot)

Out[34]:

{'whiskers': [<matplotlib.lines.Line2D at 0x11159b0>,

<matplotlib.lines.Line2D at 0x1115d30>],

'caps': [<matplotlib.lines.Line2D at 0x111f210>,

<matplotlib.lines.Line2D at 0x111f470>],

'boxes': [<matplotlib.lines.Line2D at 0x1115990>],

'medians': [<matplotlib.lines.Line2D at 0x111f6f0>],

'fliers': [<matplotlib.lines.Line2D at 0x111f950>],

'means': []}