**Day 10 Assignment**

**Step 1: Launching**

Python 3.7.6 (default, Jan 8 2020, 20:23:39) [MSC v.1916 64 bit (AMD64)]

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

IPython 7.12.0 -- An enhanced Interactive Python.

In [**1**]: import pandas as pd

In [**2**]: import numpy as np

In [**3**]: import matplotlib.pyplot as plt

In [**5**]: data = pd.read\_excel('general\_data.xlsx',sheet\_name = 0)

In [**6**]: data.head()

Out[**6**]:

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]

In [**7**]: data.columns

Out[**7**]:

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**

In [**8**]: data.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]

In [**9**]: data.duplicated()

Out[**9**]:

0 False

1 False

2 False

3 False

4 False

...

4405 False

4406 False

4407 False

4408 False

4409 False

Length: 4410, dtype: bool

In [**10**]: data.drop\_duplicates()

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]

**Step 3: Univariate**

* **Analysis:**

In [**11**]: data2 = data[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].describe()

In [**12**]: data2

Out[**12**]:

Age ... YearsWithCurrManager

count 4410.000000 ... 4410.000000

mean 36.923810 ... 4.123129

std 9.133301 ... 3.567327

min 18.000000 ... 0.000000

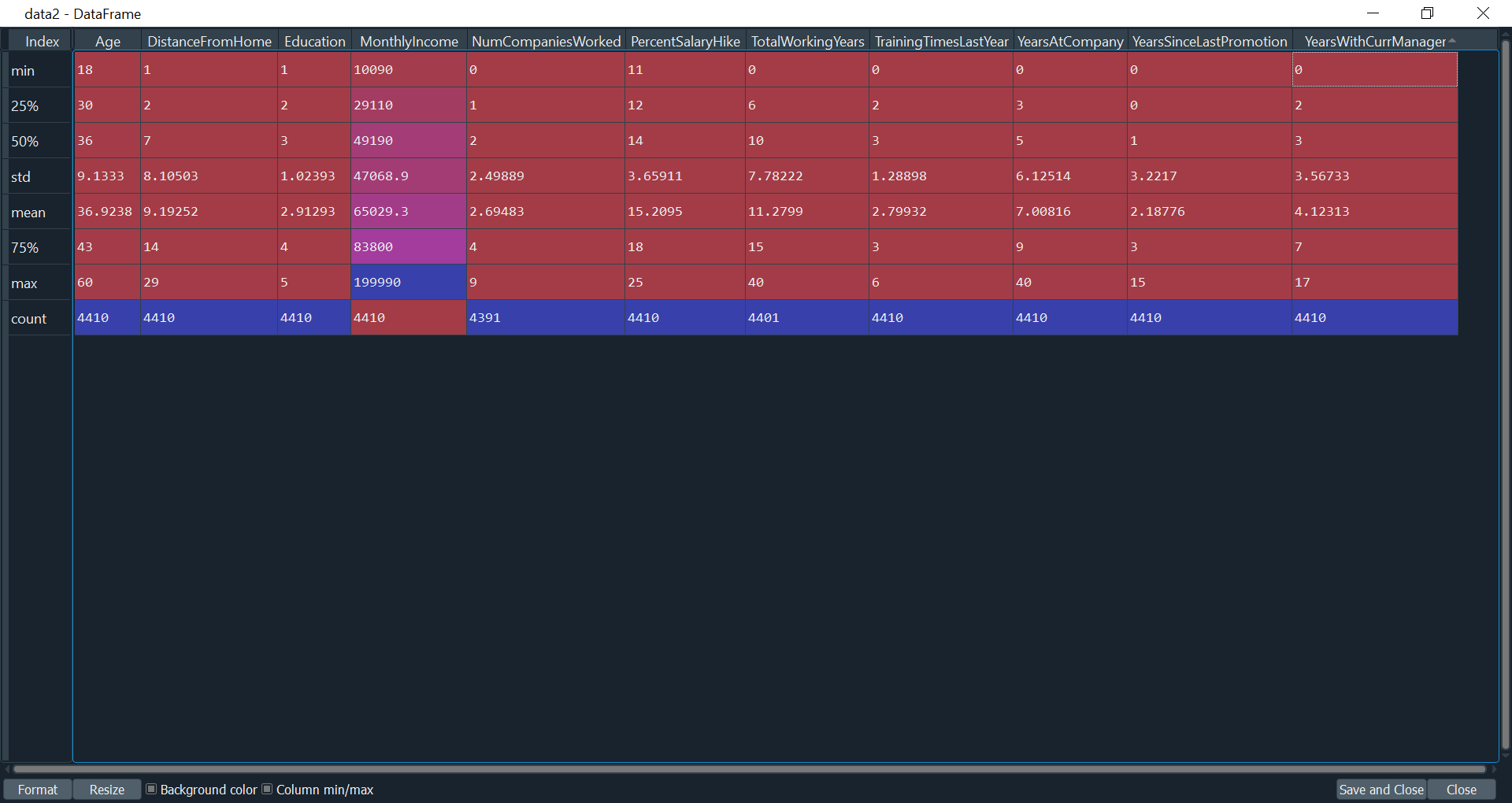
25% 30.000000 ... 2.000000

50% 36.000000 ... 3.000000

75% 43.000000 ... 7.000000

max 60.000000 ... 17.000000

[8 rows x 11 columns]



In [**13**]: data3 = data[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].median()

In [**14**]: data3

Out[**14**]:

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

dtype: float64

In [**15**]: data4 = data[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].mode()

In [**16**]: data4

Out[**16**]:

Age DistanceFromHome ... YearsSinceLastPromotion YearsWithCurrManager

0 35 2 ... 0 2

[1 rows x 11 columns]

In [**19**]: data5 = data[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].var()

In [**20**]: data5

Out[**20**]:

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

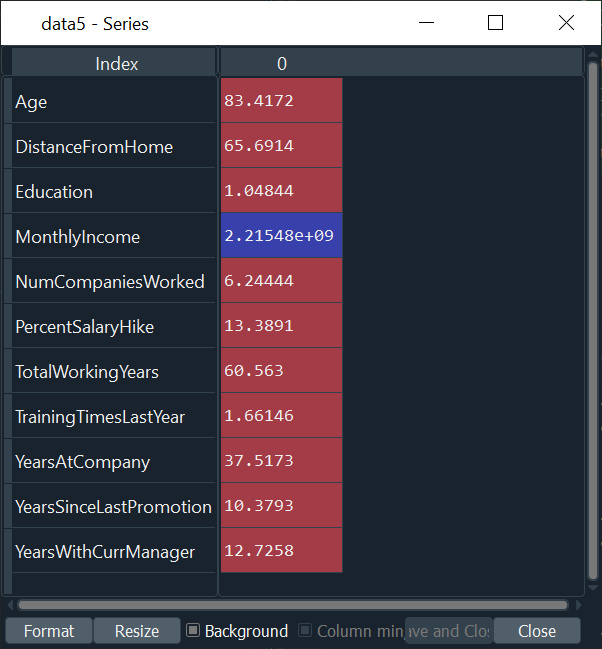
TrainingTimesLastYear 1.661465e+00

YearsAtCompany 3.751728e+01

YearsSinceLastPromotion 1.037935e+01

YearsWithCurrManager 1.272582e+01

dtype: float64



In [**21**]: data6 = data[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].skew()

In [**22**]: data6

Out[**22**]:

Age 0.413005

DistanceFromHome 0.957466

Education -0.289484

MonthlyIncome 1.368884

NumCompaniesWorked 1.026767

PercentSalaryHike 0.820569

TotalWorkingYears 1.116832

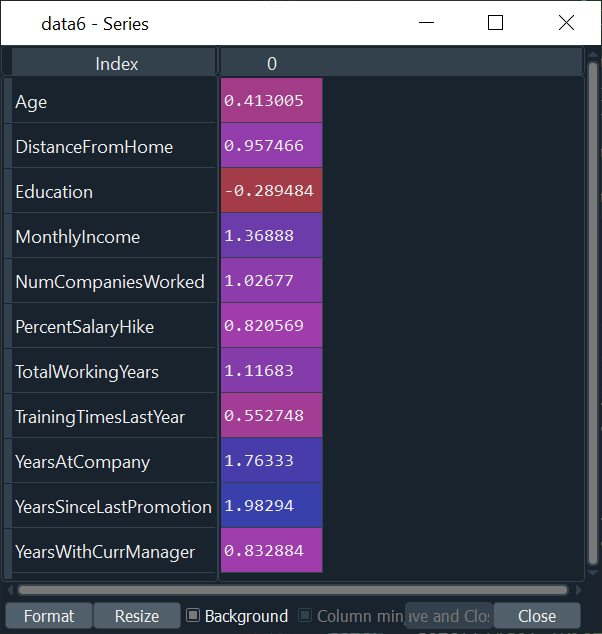
TrainingTimesLastYear 0.552748

YearsAtCompany 1.763328

YearsSinceLastPromotion 1.982939

YearsWithCurrManager 0.832884

dtype: float64



In [**23**]: data7 = data[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].kurt()

In [**24**]: data7

Out[**24**]:

Age -0.405951

DistanceFromHome -0.227045

Education -0.560569

MonthlyIncome 1.000232

NumCompaniesWorked 0.007287

PercentSalaryHike -0.302638

TotalWorkingYears 0.912936

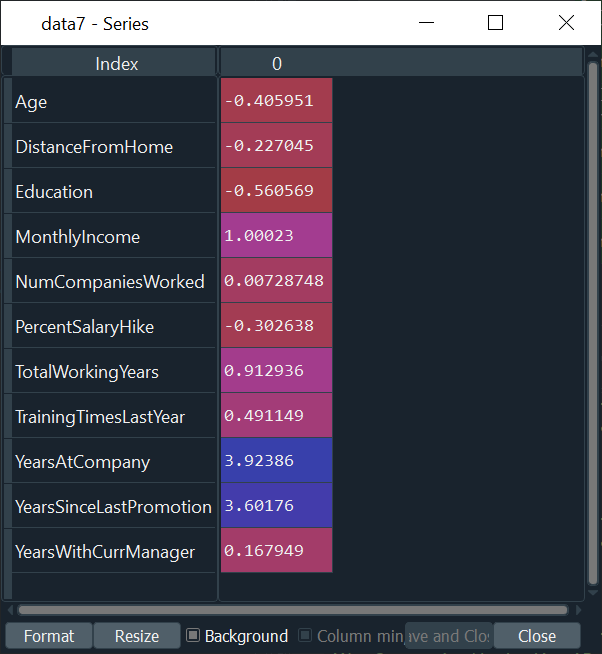
TrainingTimesLastYear 0.491149

YearsAtCompany 3.923864

YearsSinceLastPromotion 3.601761

YearsWithCurrManager 0.167949

dtype: float64



* **Outlies:**

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

In [**25**]: box\_plot1 = data.Age

In [**26**]: plt.boxplot(box\_plot1)

Out[**26**]:

{'whiskers': [<matplotlib.lines.Line2D at 0x1d7385c7d08>,

<matplotlib.lines.Line2D at 0x1d73810ce88>],

'caps': [<matplotlib.lines.Line2D at 0x1d73810c288>,

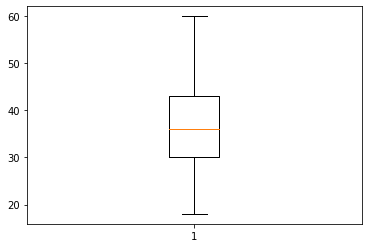
<matplotlib.lines.Line2D at 0x1d73857d308>],

'boxes': [<matplotlib.lines.Line2D at 0x1d7385a4cc8>],

'medians': [<matplotlib.lines.Line2D at 0x1d738489d08>],

'fliers': [<matplotlib.lines.Line2D at 0x1d738607dc8>],

'means': []}



In [**28**]: box\_plot2 = data.MonthlyIncome

In [**29**]: plt.boxplot(box\_plot2)

Out[**29**]:

{'whiskers': [<matplotlib.lines.Line2D at 0x1d738749788>,

<matplotlib.lines.Line2D at 0x1d738749d08>],

'caps': [<matplotlib.lines.Line2D at 0x1d738749e08>,

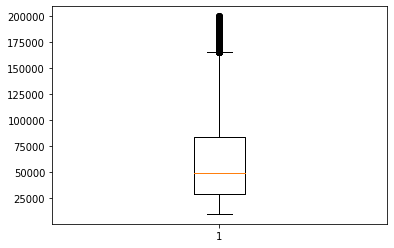
<matplotlib.lines.Line2D at 0x1d738749e88>],

'boxes': [<matplotlib.lines.Line2D at 0x1d738740fc8>],

'medians': [<matplotlib.lines.Line2D at 0x1d73874ed88>],

'fliers': [<matplotlib.lines.Line2D at 0x1d73874ee88>],

'means': []}



In [**32**]: box\_plot3 = data.YearsAtCompany

In [**33**]: plt.boxplot(box\_plot3)

Out[**33**]:

{'whiskers': [<matplotlib.lines.Line2D at 0x1d73a9fbfc8>,

<matplotlib.lines.Line2D at 0x1d73aa04a48>],

'caps': [<matplotlib.lines.Line2D at 0x1d73aa04fc8>,

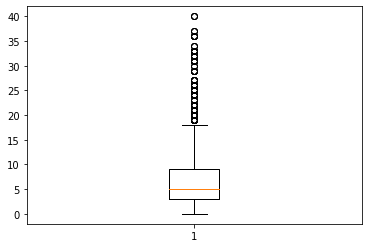
<matplotlib.lines.Line2D at 0x1d73aa04bc8>],

'boxes': [<matplotlib.lines.Line2D at 0x1d73a9fbe88>],

'medians': [<matplotlib.lines.Line2D at 0x1d739381ac8>],

'fliers': [<matplotlib.lines.Line2D at 0x1d739381fc8>],

'means': []}



In [**34**]: