Name: Ali Mohammed Taresh Ali

E-mail Id: [ali713626792@gmail.com](mailto:ali713626792@gmail.com)

------------------------------------------------------------------------------------------------------------------------------------------

**Day 7 Attration Assignment**

**Question 1:** ***Day 7 Attrition Assignment***

Attach the dataset and mentioned problem statement below,

● A sizable department has to be maintained, for the purposes of recruiting new talent

● More often than not, the new employees have to be trained for the job and/or given time to acclimatize themselves to the company

● Hence, the management has contracted an HR analytics firm to understand what factors they should focus on, in order to curb attrition.

● In other words, they want to know what changes they should make to their workplace, in order to get most of their employees to stay.

● Also, they want to know which of these variables is most important and needs to be addressed right away.

------------------------------------------------------------------------------------------------------------------------------------

**Step1: Launching:**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

dataset1=pd.read\_csv('general\_data.csv')

**dataset1.head()**

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.tail()**

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]

**dataset1.columns**

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:**

**dataset1.isnull()**

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()**

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()**

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.dropna()**

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

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

4404 29 No ... 1 5

4405 42 No ... 0 2

4406 29 No ... 0 2

4407 25 No ... 1 2

4408 42 No ... 7 8

[4382 rows x 24 columns]

**Step 3 - Univariate Analysis:**

**dataset2=dataset1.columns**

**dataset2**

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')

**dataset2=dataset1**[['Age','DistanceFromHome','Education','MonthlyIncome',

'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear',

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

**Dataset2**

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]

**dataset2=dataset1**[['Age','DistanceFromHome','Education','MonthlyIncome',

'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear',

'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].sample(25).describe()

**dataset2**

Age ... YearsWithCurrManager

count 25.000000 ... 25.000000

mean 37.240000 ... 5.040000

std 10.284616 ... 4.457952

min 21.000000 ... 0.000000

25% 29.000000 ... 2.000000

50% 36.000000 ... 4.000000

75% 43.000000 ... 8.000000

max 59.000000 ... 17.000000

[8 rows x 11 columns]

**dataset2=dataset1**[['Age','DistanceFromHome','Education','MonthlyIncome',

'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear',

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

**dataset2**

Age 36.923810

DistanceFromHome 9.192517

Education 2.912925

MonthlyIncome 65029.312925

NumCompaniesWorked 2.694830

PercentSalaryHike 15.209524

TotalWorkingYears 11.279936

TrainingTimesLastYear 2.799320

YearsAtCompany 7.008163

YearsSinceLastPromotion 2.187755

YearsWithCurrManager 4.123129

dtype: float64

**dataset2=dataset1**[['Age','DistanceFromHome','Education','MonthlyIncome',

'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear',

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

**Dataset2**

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

**dataset2=dataset1**[['Age','DistanceFromHome','Education','MonthlyIncome',

'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear',

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

**Dataset2**

Age DistanceFromHome ... YearsSinceLastPromotion YearsWithCurrManager

0 35 2 ... 0 2

[1 rows x 11 columns]

**dataset2=dataset1**[['Age','DistanceFromHome','Education','MonthlyIncome',

'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear',

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

**dataset2**

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

**dataset2=dataset1**[['Age','DistanceFromHome','Education','MonthlyIncome',

'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear',

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

**Dataset2**

Age 4410

DistanceFromHome 4410

Education 4410

MonthlyIncome 4410

NumCompaniesWorked 4391

PercentSalaryHike 4410

TotalWorkingYears 4401

TrainingTimesLastYear 4410

YearsAtCompany 4410

YearsSinceLastPromotion 4410

YearsWithCurrManager 4410

dtype: int64

**dataset2=dataset1**[['Age','DistanceFromHome','Education','MonthlyIncome',

'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear',

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

**Dataset2**

Age 162834.0

DistanceFromHome 40539.0

Education 12846.0

MonthlyIncome 286779270.0

NumCompaniesWorked 11833.0

PercentSalaryHike 67074.0

TotalWorkingYears 49643.0

TrainingTimesLastYear 12345.0

YearsAtCompany 30906.0

YearsSinceLastPromotion 9648.0

YearsWithCurrManager 18183.0

dtype: float64

**dataset2=dataset1**[['Age','DistanceFromHome','Education','MonthlyIncome',

'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear',

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

**Dataset2**

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

**dataset2=dataset1**[['Age','DistanceFromHome','Education','MonthlyIncome',

'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear',

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

**Dataset2**

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

**dataset2=dataset1**[['Age','DistanceFromHome','Education','MonthlyIncome',

'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear',

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

**Dataset2**

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

**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:**

**1-Scatter plot:**

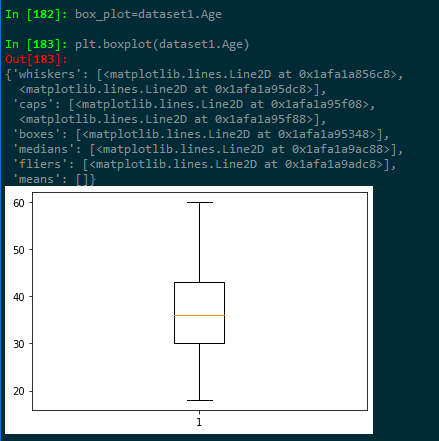
There’s no regression found while plotting Age, MonthlyIncome, TotalWorkingYears,

YearsAtCompany, etc., on a scatter plot.

**2-Box plot:**

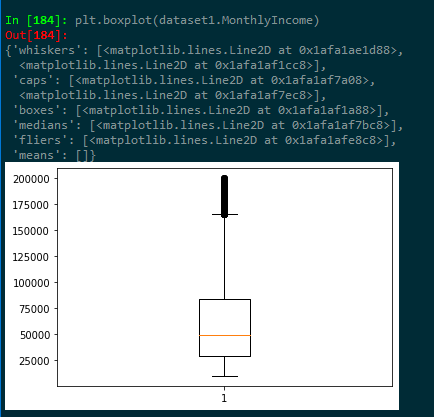
**box\_plot=dataset1.Age**

**plt.boxplot(dataset1.Age)**



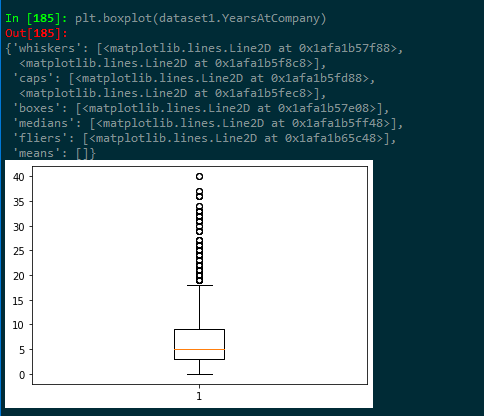
* **Age is normally distributed without any outliers**

**plt.boxplot(dataset1.MonthlyIncome)**



* **Monthly Income is Right skewed with several outliers**

**plt.boxplot(dataset1.YearsAtCompany)**



* **Years at company is also Right Skewed with several outliers observed.**