Assignment - Attrition

Step 1: Load the sheet/Data

import pandas as pd

import matplotlib.pyplot as mplt

dataset = pd.read\_csv("D:/AI\_ML\_Course/Day 7/general\_data.csv")

dataset.columns

Out[3]:

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

dataset.isnull()

Out[4]:

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[6]:

0 False

1 False

2 False

3 False

4 False

4405 False

4406 False

4407 False

4408 False

4409 False

Length: 4410, dtype: bool

dataset.drop\_duplicates()

Out[7]:

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

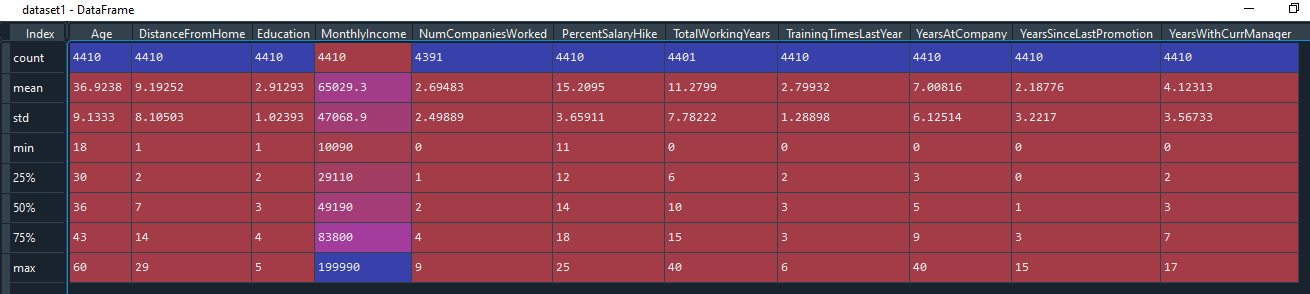
4409 40 No ... 3 9

[4410 rows x 24 columns]

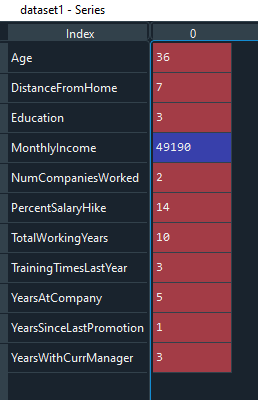
Step 3: Uni-Variate Analysis:

dataset1=dataset[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].describe()

Dataset1



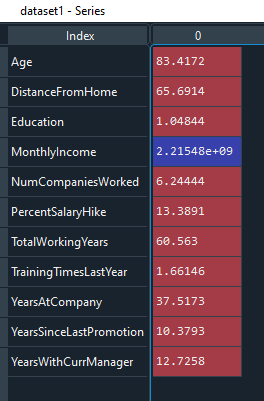
dataset1=dataset[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].median()



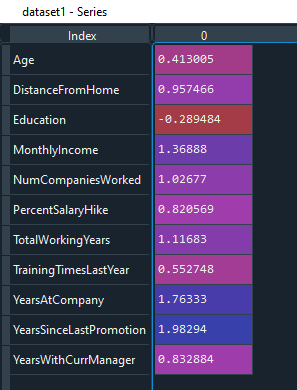
dataset1=dataset[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].mode()



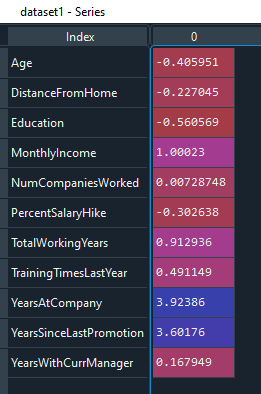
dataset1=dataset[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].var()



dataset1=dataset[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].skew()



dataset1=dataset[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].kurt()



|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Mean** | **Median** | **Mode** | **Variance** | **Std Deviation** | **IQR** | **Skewness** | **Kurtosis** |
| **Age (Yrs)** | 36.9 | 36 | 35 | 83.41 | 9.13 | 13 | 0.41 | -0.41 |
| **DistanceFromHome (Km)** | 9.19 | 7 | 2 | 65.69 | 8.1 | 12 | 0.96 | -0.23 |
| **Monthly Income (Rs)** | 65029 | 49190 | 23420 | 2215480270 | 47068 | 54690 | 1.37 | 1 |
| **PercentSalaryHike (%)** | 15 | 14 | 11 | 13.39 | 3.66 | 6 | 0.82 | -0.3 |
| **TotalWorkingYears (Yrs)** | 11.29 | 10 | 10 | 60.56 | 7.78 | 9 | 1.12 | 0.91 |
| **YearsAtCompany (Yrs)** | 7 | 5 | 5 | 37.52 | 6.12 | 6 | 1.76 | 3.92 |
| **YearsSinceLastPromotion (Yrs)** | 2 | 1 | 0 | 10.38 | 3.22 | 3 | 1.98 | 3.6 |
| **YearsWithCurrManager (Yrs)** | 4 | 3 | 2 | 12.73 | 3.57 | 5 | 0.83 | 0.17 |

Inference from the analysis:

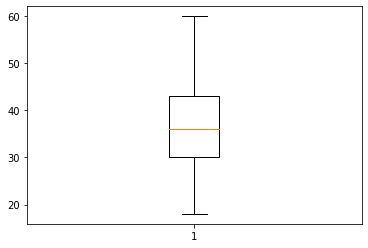
* All the above variables show positive skewness.
* Years\_At\_Company & Years\_Since\_LastPromotion are Leptokurtic i.e. more than 3 and all other variables are Platykurtic.
* The Mean\_Monthly\_Income’s IQR is at 54K suggesting companywide attrition across all income bands
* Mean age forms a near normal distribution with 13 years of IQR
* Mean Distance\_From\_Home is 12 Km which is higher.

Outliers:

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

box\_plot=dataset.Age

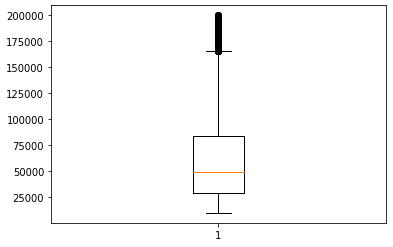
mplt. boxplot(box\_plot)



Age is normally distributed without any Outliers

box\_plot=dataset.MonthlyIncome

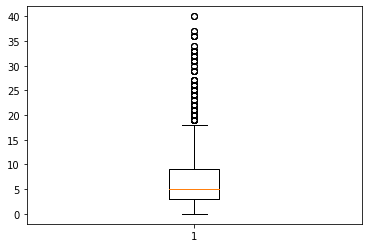
mplt.boxplot(box\_plot)



Monthly Income is right skewed with several Outliers

box\_plot=dataset.YearsAtCompany

mplt.boxplot(box\_plot)



Years at company is also Right skewed with several Outliers