import pandas as pd

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

dataset.head()

Out[3]:

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]

dataset.columns

Out[4]:

Index(['Age', 'Attrition', 'BusinessTravel', 'Department', 'DistanceFromHome',

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

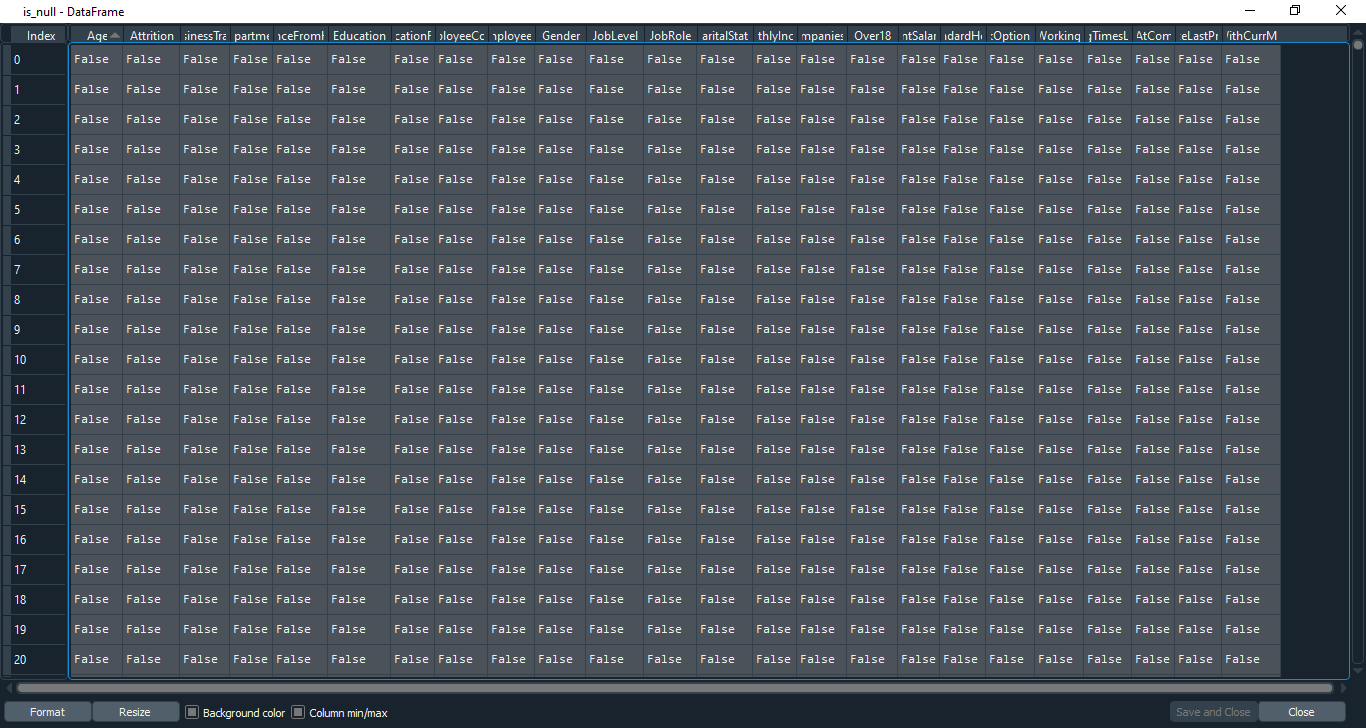
'JobLevel', 'JobRole', 'MaritalStatus', 'MonthlyIncome',

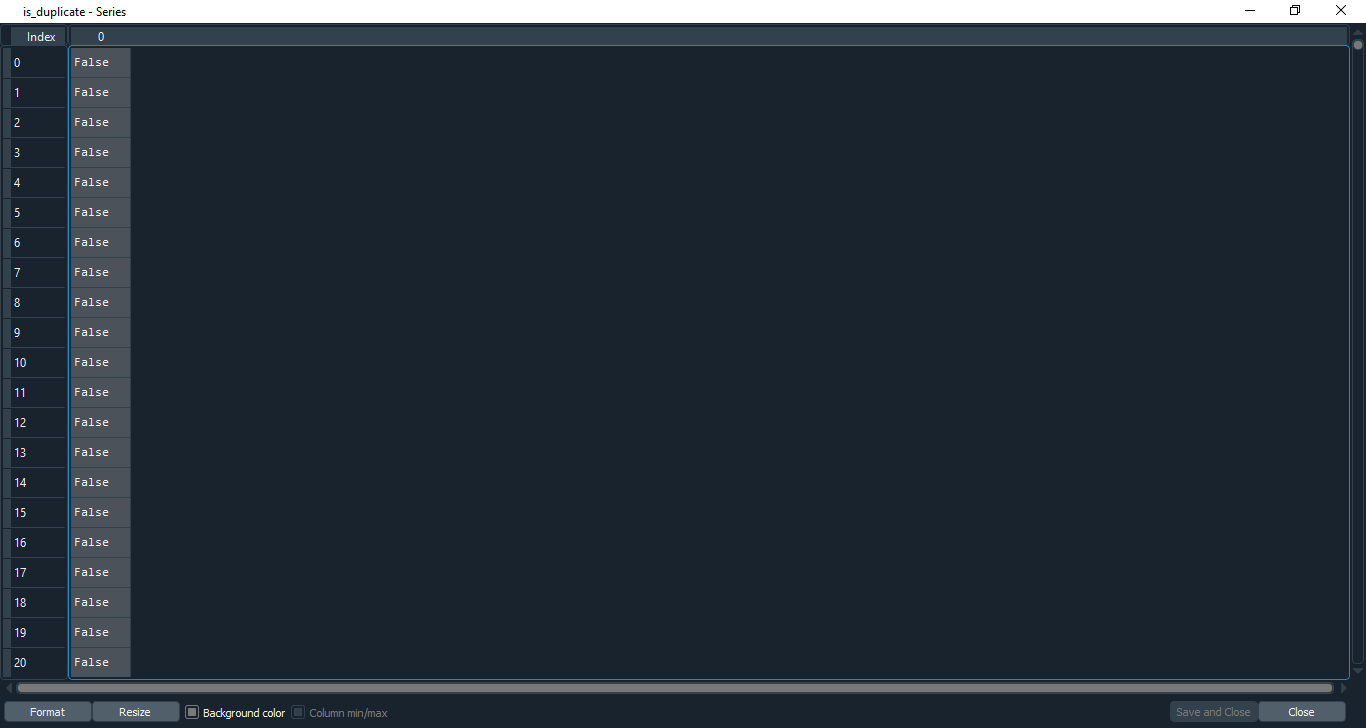
'NumCompaniesWorked', 'Over18', 'PercentSalaryHike', 'StandardHours',

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

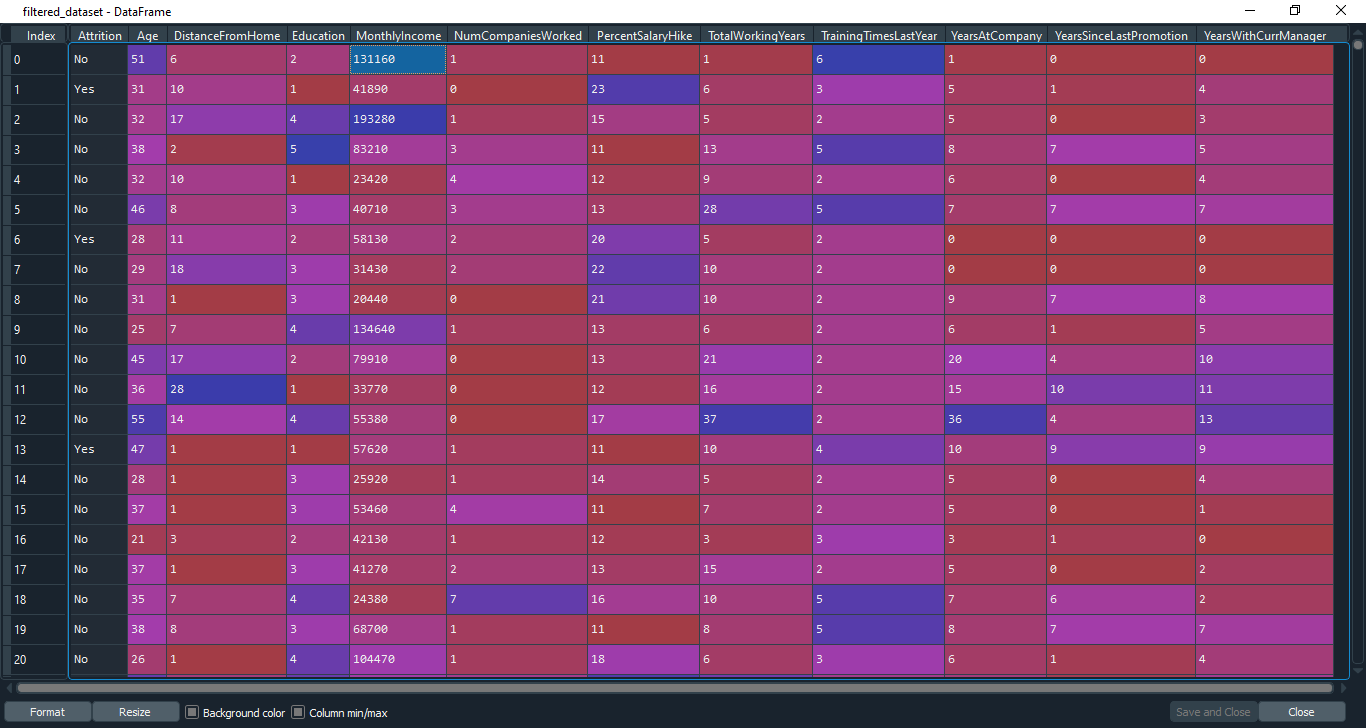
'YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManager'],

dtype='object')

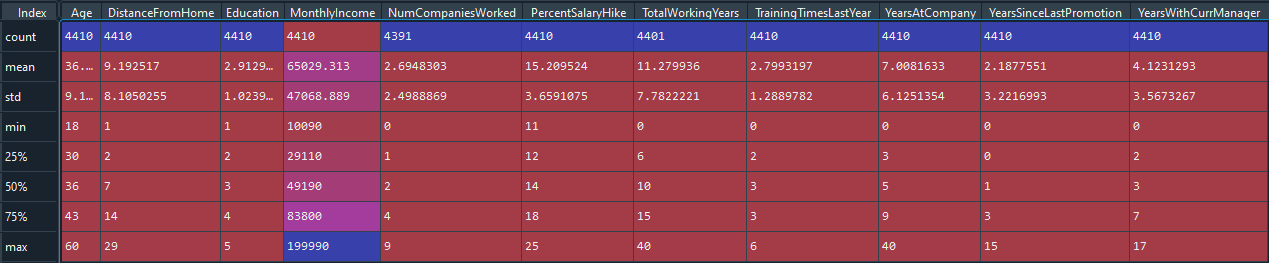
is\_null = dataset.isnull()

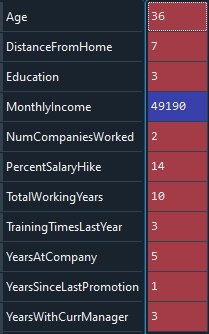
is\_duplicate = dataset.duplicated()

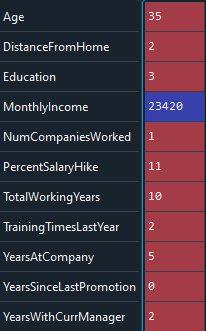
filtered\_dataset = dataset[['Attrition','Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']]

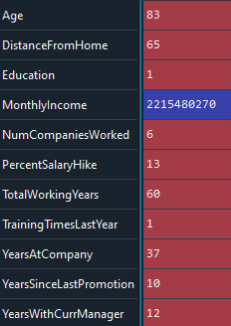


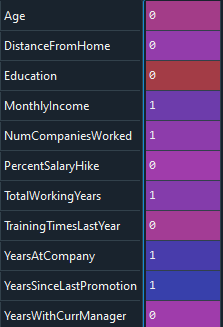
With these filtered data, we will analyse the data

description = filtered\_dataset.describe()

median = filtered\_dataset.median()

mode = filtered\_dataset.mode()

variance = filtered\_dataset.var()

skew = filtered\_dataset.skew()

kurt = filtered\_dataset.kurt()

Data Summary:

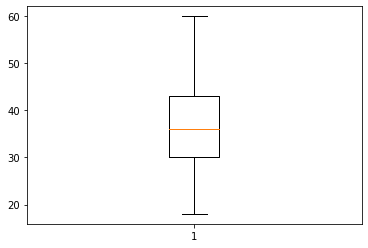
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Mean | Median | Mode | Variance | Standard  Deviation | IQR | Skewness  (Rounded) | Kurtosis  (Rounded) |
| Age(yrs) | 36.9 | 36 | 35 | 83 | 9 | 13 | 0.413 | -0.406 |
| Distance  From  Home(km) | 9.19 | 7 | 2 | 65 | 8 | 12 | 0.957 | -0.227 |
| Monthly  Income(Rs.) | 65029.313 | 49190 | 23420 | 2215480270 | 47068 | 54690 | 1.369 | 1.000 |
| TotalWorking  Years | 11.28 | 10 | 10 | 60 | 7 | 9 | 1.117 | 0.913 |
| YearsAt  Company | 7.01 | 3 | 5 | 37 | 6 | 6 | 1.763 | 3.923 |
| YearsSince  Last  Promotion | 2.19 | 1 | 0 | 10 | 3 | 3 | 1.983 | 3.602 |
| YearsWith  CurrManager | 4.12 | 3 | 2 | 12 | 3 | 5 | 0.833 | 0.168 |

Inferences from Above data:

1. All above variable show close to normal skewness, but all are slightly positive
2. YearsAtCompany and YearsSinceLastPromotion are leptokurtic, and others are mesokurtic to platykurtic
3. The MonthlyIncome IQR suggests income-bands wide Attrition, and thus that’s not the reason of Attrition
4. Mean age forms a near normal distribution with 13 years of IQR(close to 0 in perfect normality)

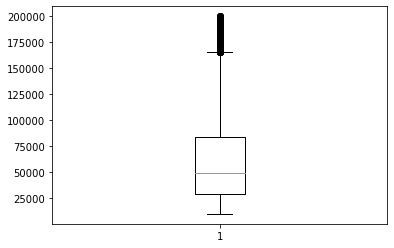
Outliers:

Age:

plt.boxplot(dataset.Age)

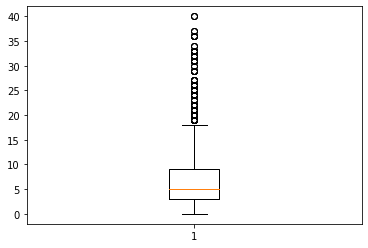
Age is normally distributed without any Outliers.

Monthly Income:

plt.boxplot(dataset.MonthlyIncome)

Monthly income is Positively(Right) skewed with several outliers

Years at company:

plt.boxplot(dataset.YearsAtCompany)

Years at company is also Positively(Right) Skewed with several outliers observed.