**Attrition Assignment**

**Launching**

**import pandas as pd**

**import numpy as np**

**import matplotlib.pyplot as plt**

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

**dataset.head()**

**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**

**[5 rows x 24 columns]**

**dataset.columns**

**Out[11]:**

**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')**

**Data Treatment**

**dataset.isnull()**

**Out[12]:**

**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.dropna()

**Out[13]:**

**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]**

**datset.duplicated()**

**Out[14]:**

**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[15]:**

**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**

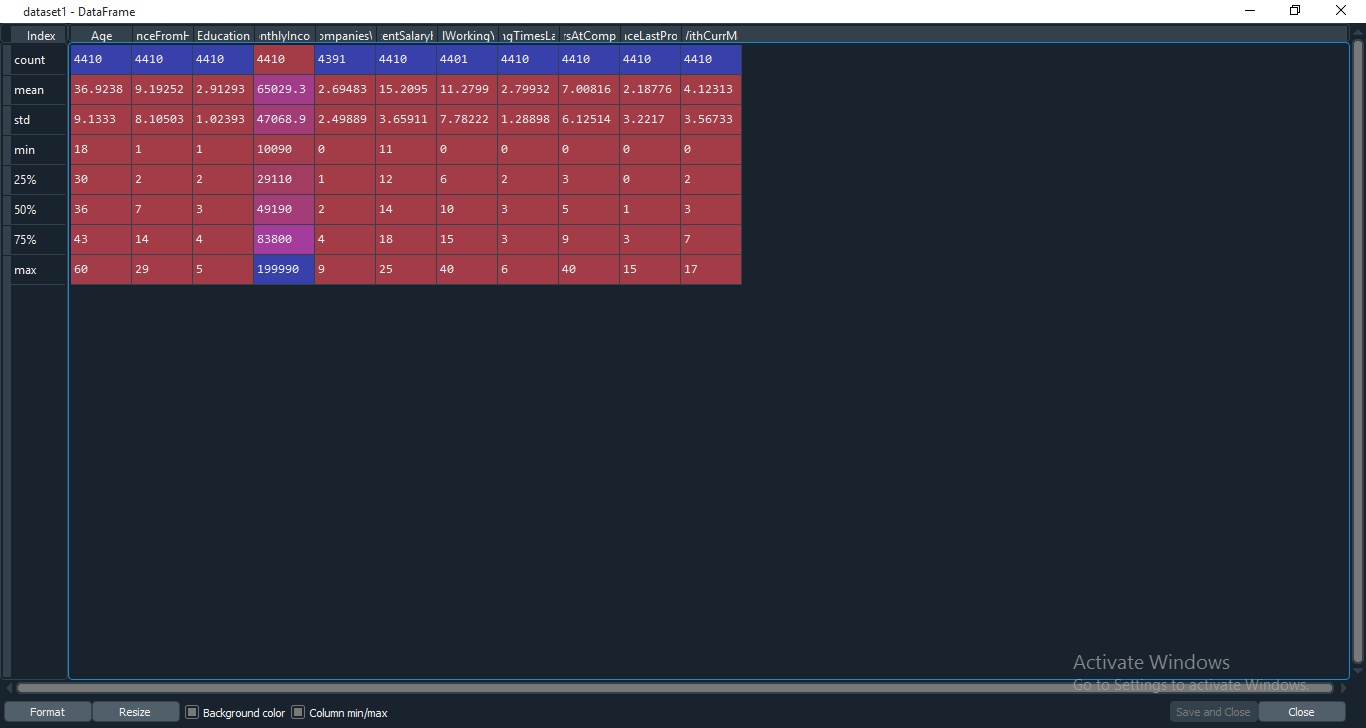
**4407 25 No ... 1 2**

**4408 42 No ... 7 8**

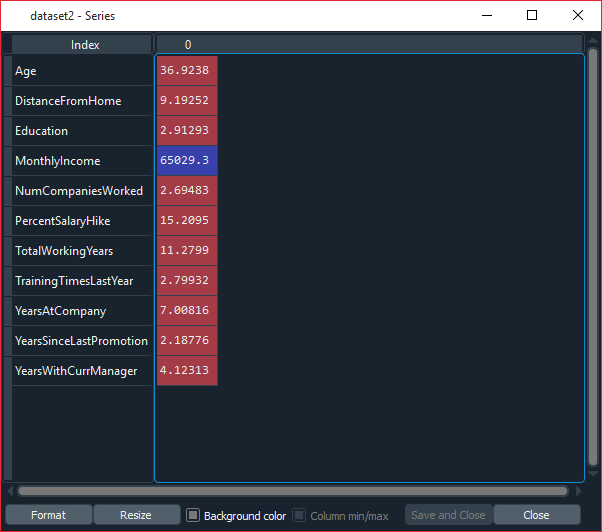
**4409 40 No ... 3 9**

**[4410 rows x 24 columns]**

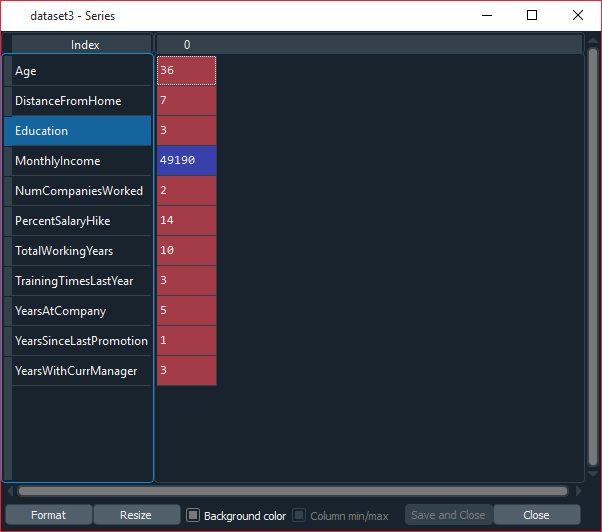
**dataset1=dataset[['Age','DistanceFromHome','Education','MonthlyIncome','NumCompaniesWorked','PercentSalaryHike','TotalWorkingYears','TrainingTimesLastYear','YearsAtCompany','YearsSinceLastPromotion','YearsWithCurrManager']].describe()**



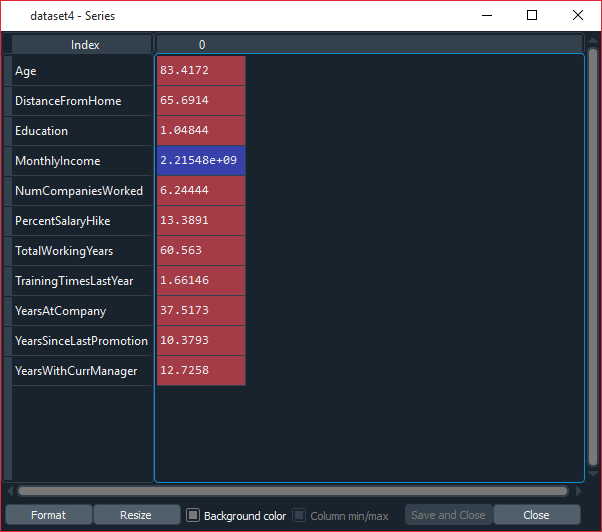
**dataset2=dataset[['Age','DistanceFromHome','Education','MonthlyIncome','NumCompaniesWorked','PercentSalaryHike','TotalWorkingYears','TrainingTimesLastYear','YearsAtCompany','YearsSinceLastPromotion','YearsWithCurrManager']].mean()**



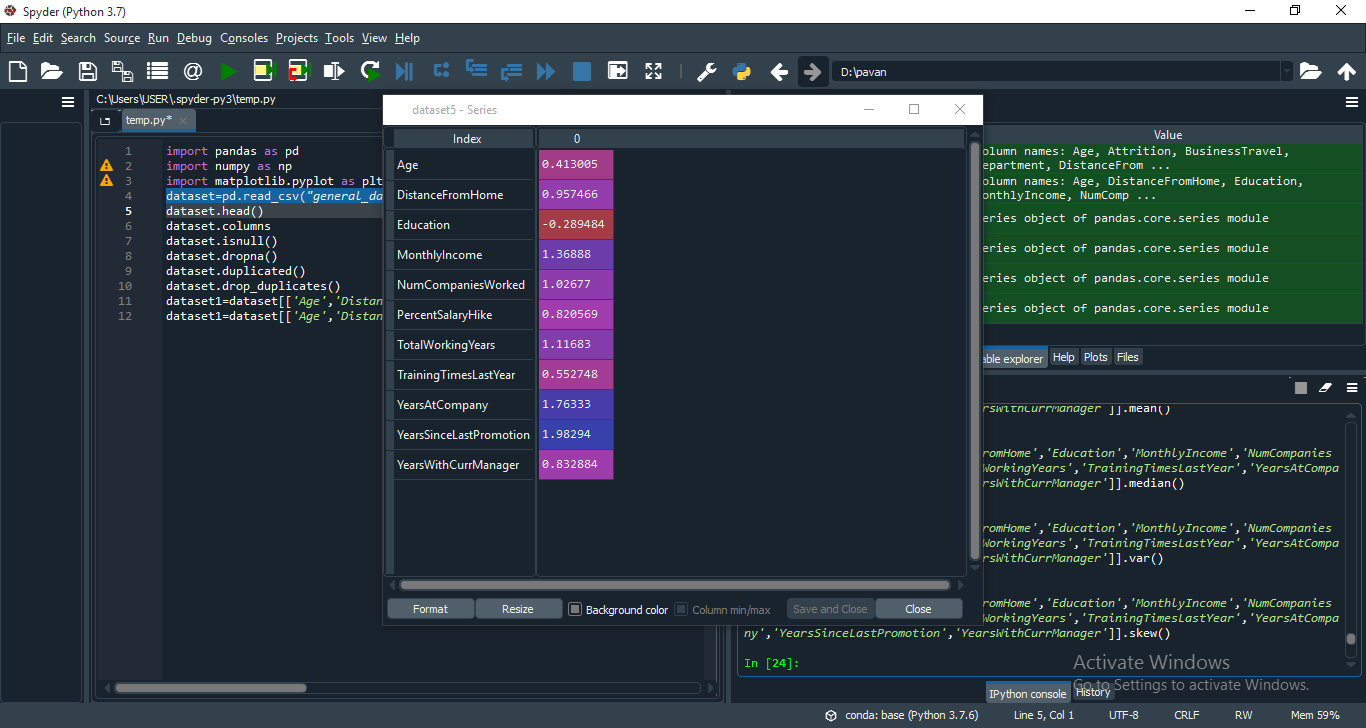
**dataset3=dataset[['Age','DistanceFromHome','Education','MonthlyIncome','NumCompaniesWorked','PercentSalaryHike','TotalWorkingYears','TrainingTimesLastYear','YearsAtCompany','YearsSinceLastPromotion','YearsWithCurrManager']].median()**



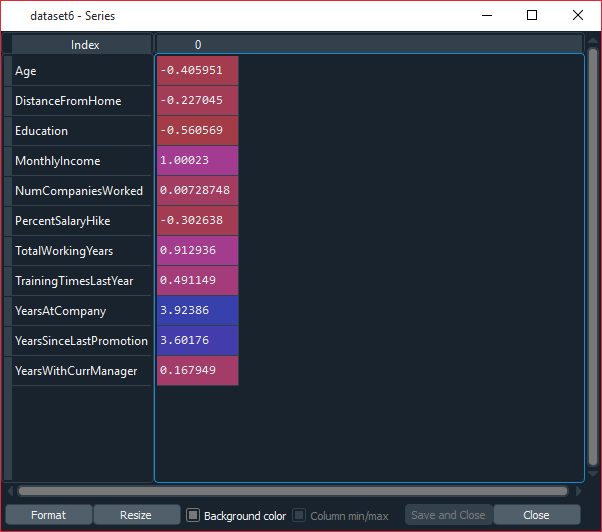
**dataset4=dataset[['Age','DistanceFromHome','Education','MonthlyIncome','NumCompaniesWorked','PercentSalaryHike','TotalWorkingYears','TrainingTimesLastYear','YearsAtCompany','YearsSinceLastPromotion','YearsWithCurrManager']].var()**



**dataset5=dataset[['Age','DistanceFromHome','Education','MonthlyIncome','NumCompaniesWorked','PercentSalaryHike','TotalWorkingYears','TrainingTimesLastYear','YearsAtCompany','YearsSinceLastPromotion','YearsWithCurrManager']].skew()**

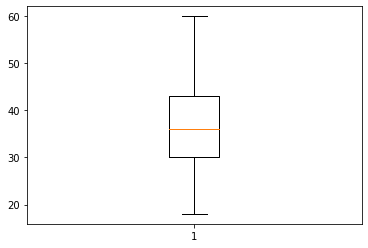


**dataset6=dataset[['Age','DistanceFromHome','Education','MonthlyIncome','NumCompaniesWorked','PercentSalaryHike','TotalWorkingYears','TrainingTimesLastYear','YearsAtCompany','YearsSinceLastPromotion','YearsWithCurrManager']].kurt()**



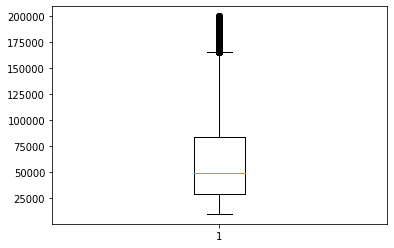
**All the variables above shows positive skewness.**

**plt.boxplot(dataset.Age)**

****

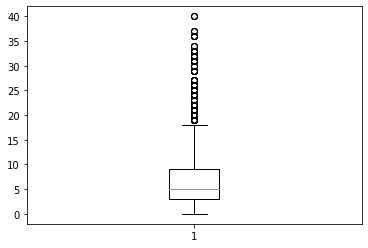
**Normaly distributed age groups without any ouliers**

**plt.boxplot(dataset.MonthlyIncome)**

****

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

**plt.boxplot(dataset.YearsAtCompany)**

****

**Right skewed with several outliers**