# Attrition assignment day 7

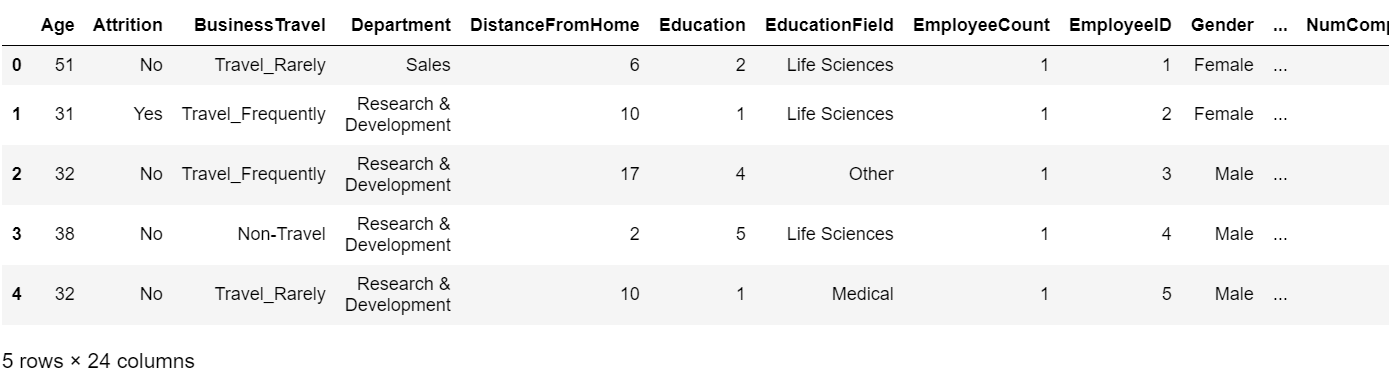
# ## Importing the library and dataset

import pandas as pd

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

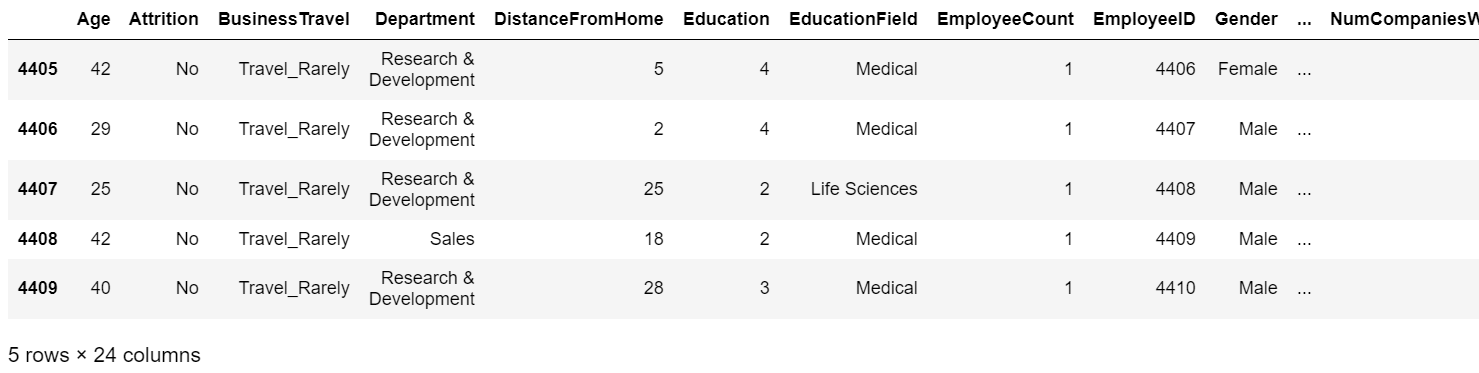
## ### reading the first 5 entries

dataset.head()



## ### reading the last 5 entries

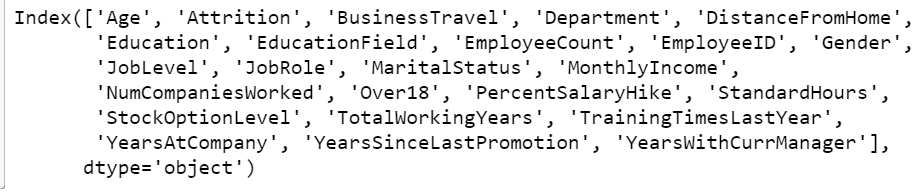
dataset.tail()



Here, we have 4410 entries in total

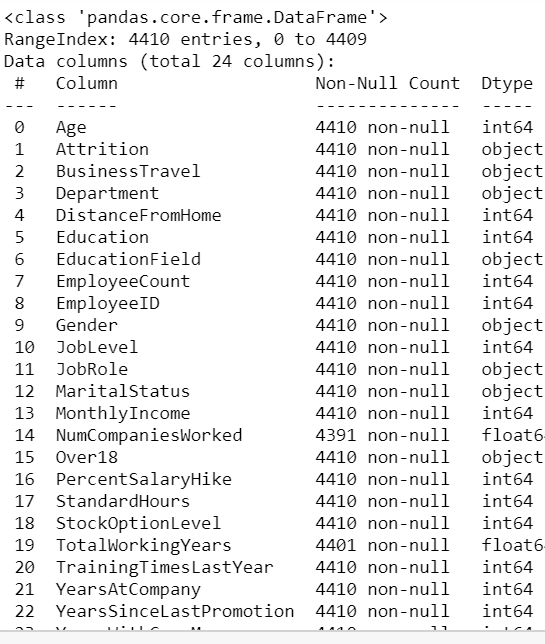
## ### printing the columns/variables

dataset.columns



## ### checking the description

dataset.info()



### #### we have 24 variables in the attrition dataset available

## ### checking for null values and duplicates

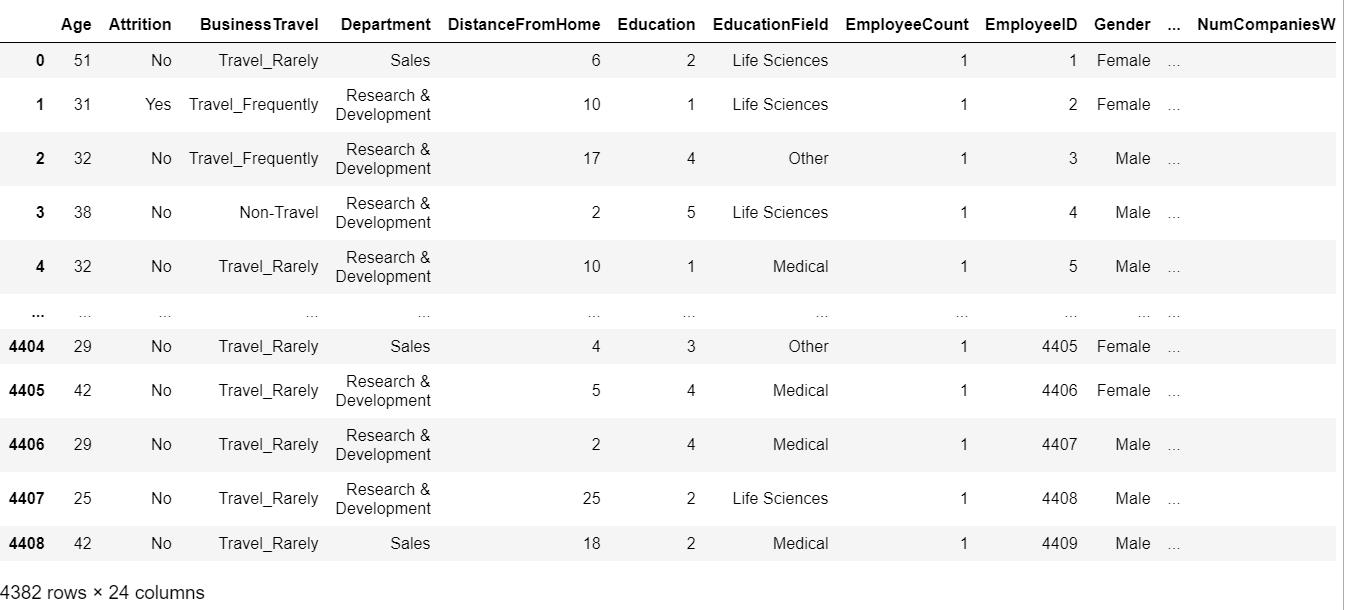
Here, we will be checking for null values and duplicates and dropping them if present

dataset.isnull()



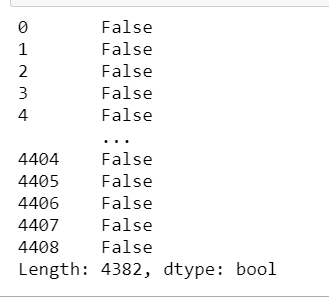
dataset1 = dataset.dropna()

dataset1



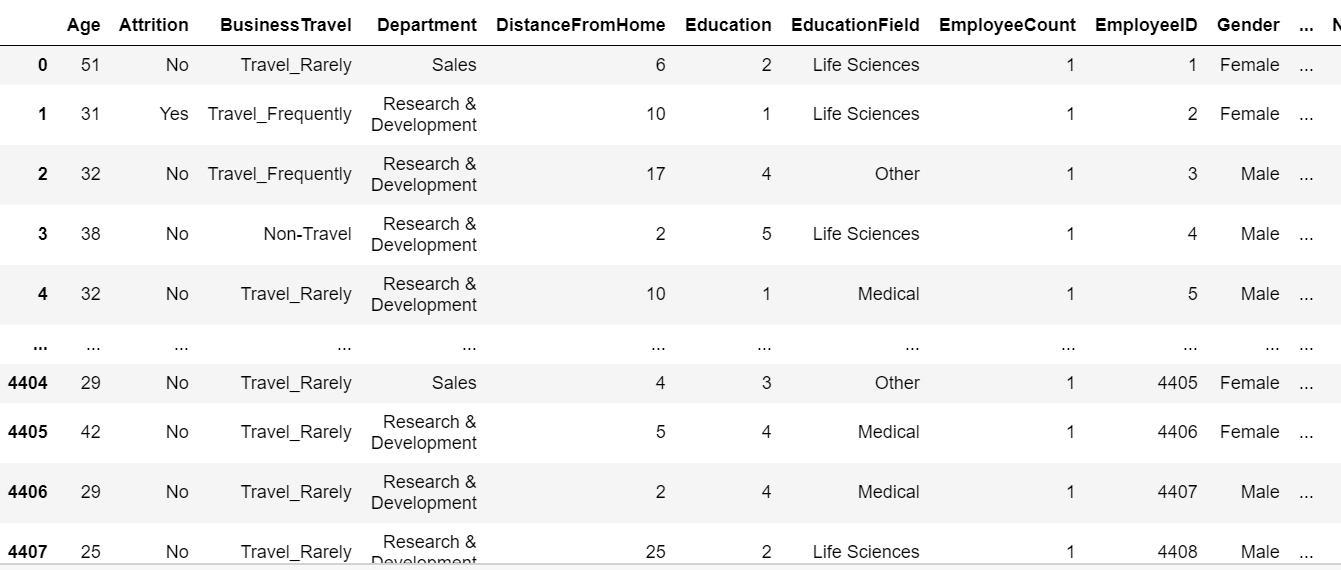
### #### from 4410 entries to 4382 entries, we can say there were, 28 null values

dataset1.duplicated()



dataset2 = dataset1.drop\_duplicates()

dataset2



### #### no duplicates present

# ## Univariate analysis

df0 = dataset2.describe()

df0



## ### finding the median

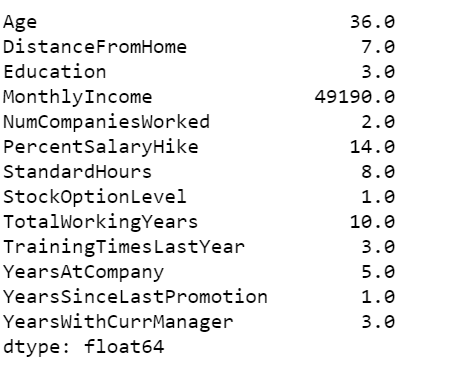
df1 = dataset2[['Age','DistanceFromHome','Education','MonthlyIncome',

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

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

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

df1



## ### finding mode

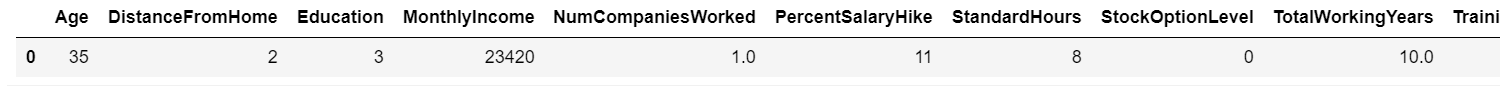
df2 = dataset2[['Age','DistanceFromHome','Education','MonthlyIncome',

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

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

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

df2



## ### finding variance

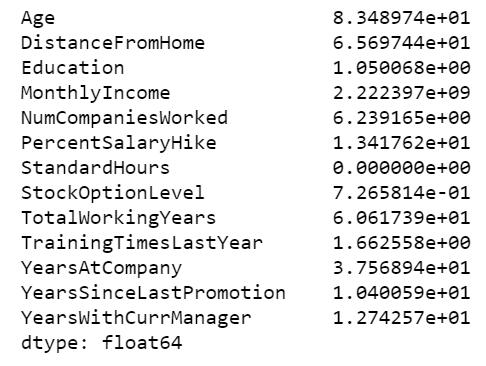
df3 = dataset2[['Age','DistanceFromHome','Education','MonthlyIncome',

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

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

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

df3



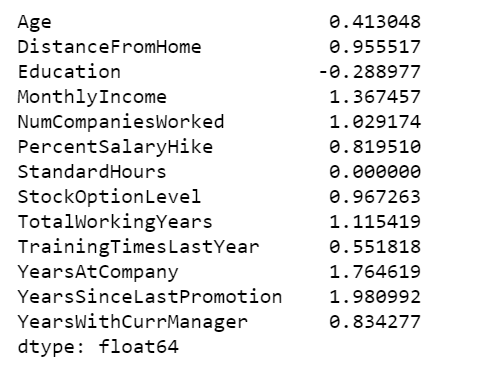
df4 = dataset2[['Age','DistanceFromHome','Education','MonthlyIncome',

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

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

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

df4



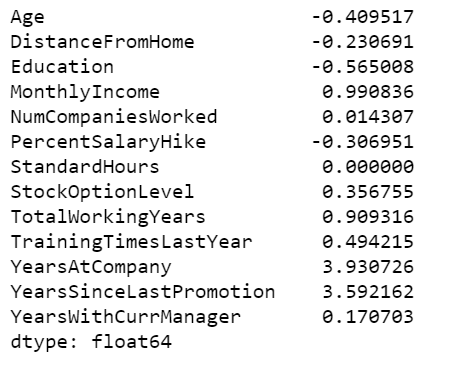
df5 = dataset2[['Age','DistanceFromHome','Education','MonthlyIncome',

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

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

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

df5



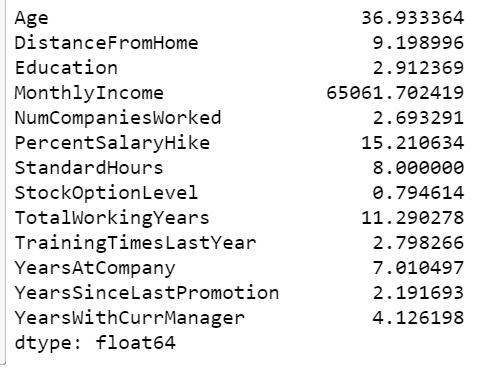
df6 = dataset2[['Age','DistanceFromHome','Education','MonthlyIncome',

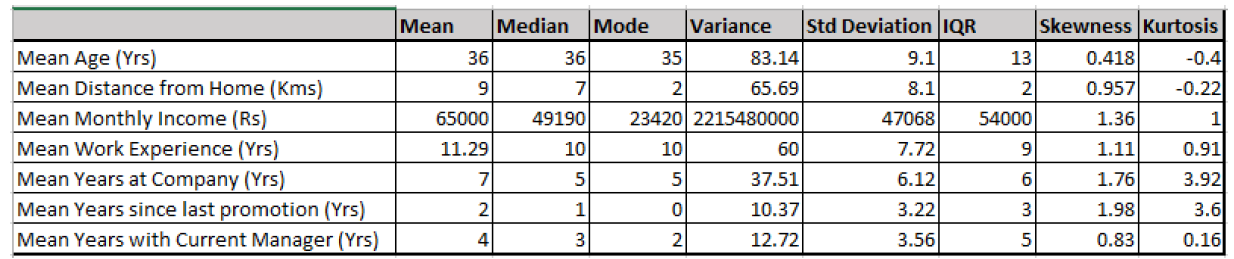
'NumCompaniesWorked', 'PercentSalaryHike', 'StandardHours',

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

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

df6





# Inferences:

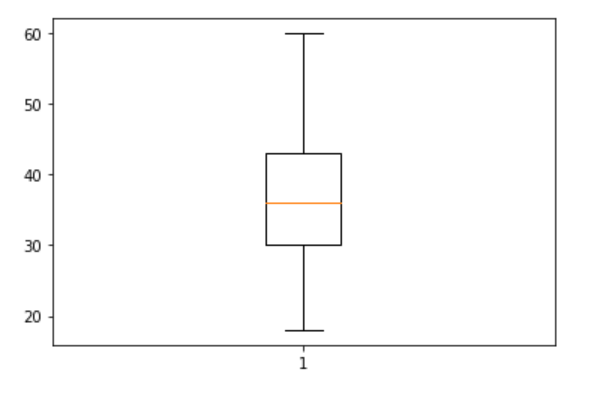
* Kurtosis of age and distance from home is negative. Rest all are positive. Hence, they’re leptokurtic and others are platykurtic
* The mean income is 65000 and its standard deviation is 47068 which is quite high. This means the company has people with huge differences in salaries.
* All the attributes show positive skewness meaning, mean > median

# ## Outliers

Outliers will be checked using box plot

import matplotlib.pyplot as plt

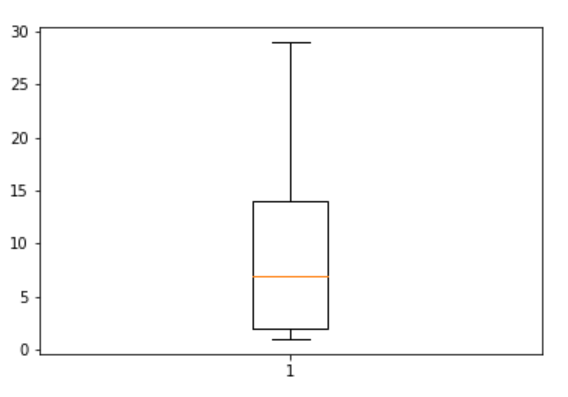
plt.boxplot(dataset2['Age'])



### AGE

conclusion: No outliers for the attribute age

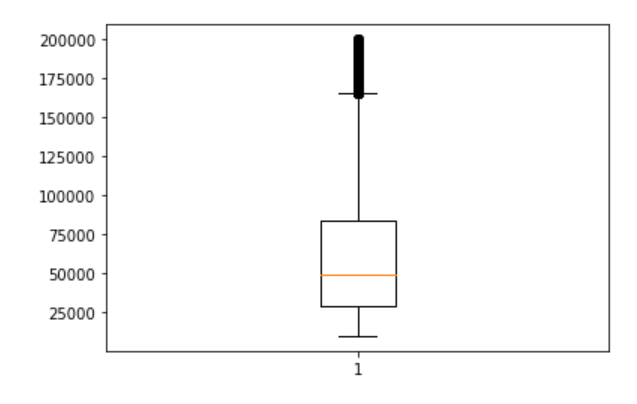
plt.boxplot(dataset2['DistanceFromHome'])



### DistanceFromHome

conclusion: No outliers for the attribute DistanceFromHome

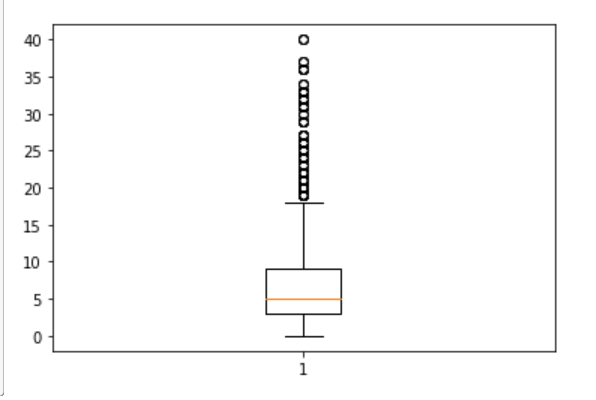
plt.boxplot(dataset2['MonthlyIncome'])



### MonthlyIncome

conclusion: Outliers exist for the attribute MonthlyIncome. Some employees have very large monthly income as compared to most others

plt.boxplot(dataset['YearsAtCompany'])



### Conclusion:

outliers exist for attribute years at company meaning some emplotees have been working here for very long according to the plot