LAUNCHING

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
dataset_1=pd.read_csv("general_data.csv")
dataset_1.head()
```

Out[5]:

	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]

DATA CLEANSING

```
dataset_1.isnull()
```

Out[7]:

	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_1.duplicated()

Out[8]:

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_1.drop_duplicates()

Out[11]:

	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]

Univariate Analysis

dataset3=dataset_1[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].describe()

Index	Age	istanceFromHom	Education	MonthlyIncome	mCompaniesWor	PercentSalaryHike	TotalWorkingYear	siningTimesLastY	YearsAtCompany	rsSinceLastPromo	rsWithCurrMana
count	4410	4410	4410	4410	4391	4410	4401	4410	4410	4410	4410
mean	36.9238	9.19252	2.91293	65029.3	2.69483	15.2095	11.2799	2.79932	7.00816	2.18776	4.12313
std	9.1333	8.10503	1.02393	47068.9	2.49889	3.65911	7.78222	1.28898	6.12514	3.2217	3.56733
min	18	1	1	10090	0	11	0	0	0	0	0
25%	30	2	2	29110	1	12	6	2	3	0	2
50%	36	7	3	49190	2	14	10	3	5	1	3
75%	43	14	4	83800	4	18	15	3	9	3	7
max	60	29	5	199990	9	25	40	6	40	15	17

dataset3=dataset_1[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].median()

Index	0
Age	36
DistanceFrom…	7
Education	3
MonthlyIncome	49190
NumCompanies…	2
PercentSalar…	14
TotalWorking…	10
TrainingTime…	3
YearsAtCompa…	5
YearsSinceLa…	1
YearsWithCur…	3

dataset3_mode=dataset_1[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].mode()

Index	Age	istanceFromHom	Education	MonthlyIncome	mCompaniesWor	PercentSalaryHike	TotalWorkingYear	siningTimesLastY	YearsAtCompany	rsSinceLastPromo	rsWithCurrMana
0	35	2	3	23420	1	11	10	2	5	0	2

```
dataset3_mean=dataset_1[['Age','DistanceFromHome','Education','MonthlyIncome',
'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear',
'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].mean()
```

Index	0
Age	36.9238
DistanceFrom…	9.19252
Education	2.91293
MonthlyIncome	65029.3
NumCompanies…	2.69483
PercentSalar…	15.2095
TotalWorking…	11.2799
TrainingTime…	2.79932
YearsAtCompa…	7.00816
YearsSinceLa…	2.18776
YearsWithCur…	4.12313

```
dataset3_var=dataset_1[['Age','DistanceFromHome','Education','MonthlyIncome',
'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear',
'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].var()
```

Index	0
Age	83.4172
DistanceFrom…	65.6914
Education	1.04844
MonthlyIncome	2.21548e+09
NumCcompanies…	6.24444
PercentSalar…	13.3891
TotalWorking…	60.563
TrainingTime…	1.66146
YearsAtCompa…	37.5173
YearsSinceLa…	10.3793
YearsWithCur…	12.7258

dataset3_skew=dataset_1[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].skew()

Index	0
Age	0.413005
DistanceFrom…	0.957466
Education	-0.289484
MonthlyIncome	1.36888
NumCompanies…	1.02677
PercentSalar…	0.820569
TotalWorking…	1.11683
TrainingTime…	0.552748
YearsAtCompa…	1.76333
YearsSinceLa…	1.98294
YearsWithCur…	0.832884

dataset3_kurt=dataset_1[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].kurt()

Index	0
Age	-0.405951
DistanceFrom…	-0.227045
Education	-0.560569
MonthlyIncome	1.00023
NumCompanies…	0.00728748
PercentSalar…	-0.302638
TotalWorking…	0.912936
TrainingTime…	0.491149
YearsAtCompa…	3.92386
YearsSinceLa…	3.60176
YearsWithCur…	0.167949

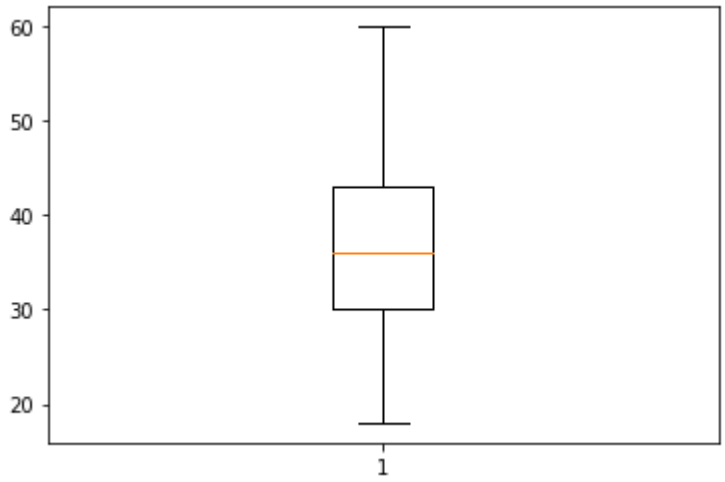
Inference from the analysis:

- 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

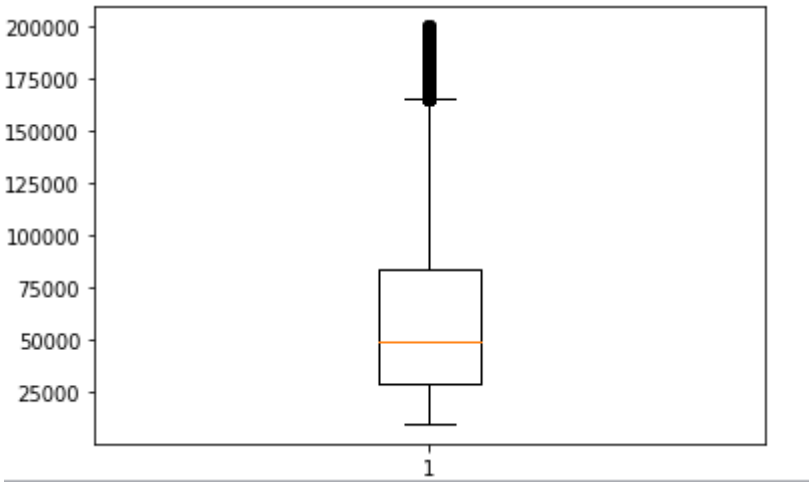
OUTLIERS

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

plt.boxplot(dataset_1.Age)



plt.boxplot(dataset_1.MonthlyIncome)



Monthly Income is Right skewed with several outliers

In [26]: plt.scatter(dataset_1.MonthlyIncome,dataset_1.YearsAtCompany)

```
Out[26]: <matplotlib.collections.PathCollection at 0xf0b64
```