Day 7 Notes

1. Launching

```
In [1]:
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
df= pd.read_csv("general_data.csv")
```

```
In [2]:
df.columns
```

```
Out[2]: 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')
```

```
In [27]:
df.describe()
```

Out[27]:

	Age	DistanceFromHome	Education	EmployeeCount	EmployeeID	JobLevel	MonthlyIncome	NumCompaniesWorked	PercentSalaryHike	S
count	4410.000000	4410.000000	4410.000000	4410.0	4410.000000	4410.000000	4410.000000	4391.000000	4410.000000	
mean	36.923810	9.192517	2.912925	1.0	2205.500000	2.063946	65029.312925	2.694830	15.209524	
std	9.133301	8.105026	1.023933	0.0	1273.201673	1.106689	47068.888559	2.498887	3.659108	
min	18.000000	1.000000	1.000000	1.0	1.000000	1.000000	10090.000000	0.000000	11.000000	
25%	30.000000	2.000000	2.000000	1.0	1103.250000	1.000000	29110.000000	1.000000	12.000000	
50%	36.000000	7.000000	3.000000	1.0	2205.500000	2.000000	49190.000000	2.000000	14.000000	
75%	43.000000	14.000000	4.000000	1.0	3307.750000	3.000000	83800.000000	4.000000	18.000000	
max	60.000000	29.000000	5.000000	1.0	4410.000000	5.000000	199990.000000	9.000000	25.000000	

2. Data Treatment:

```
In [4]:
df.isnull()
df.dropna()
```

Out[4]:

	Age	Attrition	BusinessTravel	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeID	Gender	...	NumCompaniesW
0	51	No	Travel_Rarely	Sales	6	2	Life Sciences	1	1	Female	...	
1	31	Yes	Travel_Frequently	Research & Development	10	1	Life Sciences	1	2	Female	...	
2	32	No	Travel_Frequently	Research & Development	17	4	Other	1	3	Male	...	
3	38	No	Non-Travel	Research & Development	2	5	Life Sciences	1	4	Male	...	
4	32	No	Travel_Rarely	Research & Development	10	1	Medical	1	5	Male	...	
...	...	...	...	...	...	...	...	...	...	...	...	
4404	29	No	Travel_Rarely	Sales	4	3	Other	1	4405	Female	...	
4405	42	No	Travel_Rarely	Research & Development	5	4	Medical	1	4406	Female	...	
4406	29	No	Travel_Rarely	Research & Development	2	4	Medical	1	4407	Male	...	
4407	25	No	Travel_Rarely	Research & Development	25	2	Life Sciences	1	4408	Male	...	
4408	42	No	Travel_Rarely	Sales	18	2	Medical	1	4409	Male	...	

4382 rows × 24 columns

```
In [5]:
df.drop_duplicates()
```

Out[5]:

	Age	Attrition	BusinessTravel	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeID	Gender	...	NumCompaniesW
0	51	No	Travel_Rarely	Sales	6	2	Life Sciences	1	1	Female	...	
1	31	Yes	Travel_Frequently	Research & Development	10	1	Life Sciences	1	2	Female	...	
2	32	No	Travel_Frequently	Research & Development	17	4	Other	1	3	Male	...	
3	38	No	Non-Travel	Research & Development	2	5	Life Sciences	1	4	Male	...	
4	32	No	Travel_Rarely	Research & Development	10	1	Medical	1	5	Male	...	
...	...	...	...	...	...	...	...	...	...	...	...	
4405	42	No	Travel_Rarely	Research & Development	5	4	Medical	1	4406	Female	...	
4406	29	No	Travel_Rarely	Research & Development	2	4	Medical	1	4407	Male	...	
4407	25	No	Travel_Rarely	Research & Development	25	2	Life Sciences	1	4408	Male	...	
4408	42	No	Travel_Rarely	Sales	18	2	Medical	1	4409	Male	...	
4409	40	No	Travel_Rarely	Research & Development	28	3	Medical	1	4410	Male	...	

4410 rows × 24 columns

3. Univariate Analysis:

a. describe()

```
In [6]: df.describe()
```

Out[6]:

ercentSalaryHike	StandardHours	StockOptionLevel	TotalWorkingYears	TrainingTimesLastYear	YearsAtCompany	YearsSinceLastPromotion	YearsWithCurrManager
4410.000000	4410.0	4410.000000	4401.000000	4410.000000	4410.000000	4410.000000	4410.000000
15.209524	8.0	0.793878	11.279936	2.799320	7.008163	2.187755	4.123129
3.659108	0.0	0.851883	7.782222	1.288978	6.125135	3.221699	3.567327
11.000000	8.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
12.000000	8.0	0.000000	6.000000	2.000000	3.000000	0.000000	2.000000
14.000000	8.0	1.000000	10.000000	3.000000	5.000000	1.000000	3.000000
18.000000	8.0	1.000000	15.000000	3.000000	9.000000	3.000000	7.000000
25.000000	8.0	3.000000	40.000000	6.000000	40.000000	15.000000	17.000000

b. mean()

```
In [8]: Hike","TotalWorkingYears","TrainingTimesLastYear","YearsAtCompany","YearsSinceLastPromotion","YearsWithCurrManager"]].mean()
```

Out[8]:

```
Age                          36.923810
DistanceFromHome              9.192517
Education                     2.912925
MonthlyIncome             65029.312925
NumCompaniesWorked            2.694830
PercentSalaryHike            15.209524
TotalWorkingYears            11.279936
TrainingTimesLastYear         2.799320
YearsAtCompany                7.008163
YearsSinceLastPromotion       2.187755
YearsWithCurrManager          4.123129
dtype: float64
```

c. median()

```
In [9]: ke","TotalWorkingYears","TrainingTimesLastYear","YearsAtCompany","YearsSinceLastPromotion","YearsWithCurrManager"]].median()
```

Out[9]:

```
Age                          36.0
DistanceFromHome              7.0
Education                     3.0
MonthlyIncome             49190.0
NumCompaniesWorked            2.0
PercentSalaryHike            14.0
TotalWorkingYears            10.0
TrainingTimesLastYear         3.0
YearsAtCompany                5.0
YearsSinceLastPromotion       1.0
YearsWithCurrManager          3.0
dtype: float64
```

d. mode()

```
In [11]: Hike","TotalWorkingYears","TrainingTimesLastYear","YearsAtCompany","YearsSinceLastPromotion","YearsWithCurrManager"]].mode()
```

Out[11]:

	Age	DistanceFromHome	Education	MonthlyIncome	NumCompaniesWorked	PercentSalaryHike	TotalWorkingYears	TrainingTimesLastYear	YearsAtCompany
0	35	2	3	23420	1.0	11	10.0	2	5

e. std()

```
In [12]: 1 yHike","TotalWorkingYears","TrainingTimesLastYear","YearsAtCompany","YearsSinceLastPromotion","YearsWithCurrManager"]].std()

Out[12]: Age                        9.133301
         DistanceFromHome           8.105026
         Education                  1.023933
         MonthlyIncome          47068.888559
         NumCompaniesWorked         2.498887
         PercentSalaryHike          3.659108
         TotalWorkingYears          7.782222
         TrainingTimesLastYear      1.288978
         YearsAtCompany             6.125135
         YearsSinceLastPromotion    3.221699
         YearsWithCurrManager       3.567327
         dtype: float64
```

f. var()

```
In [13]: 1 yHike","TotalWorkingYears","TrainingTimesLastYear","YearsAtCompany","YearsSinceLastPromotion","YearsWithCurrManager"]].var()

Out[13]: Age                      8.341719e+01
         DistanceFromHome         6.569144e+01
         Education                1.048438e+00
         MonthlyIncome            2.215480e+09
         NumCompaniesWorked       6.244436e+00
         PercentSalaryHike        1.338907e+01
         TotalWorkingYears        6.056298e+01
         TrainingTimesLastYear    1.661465e+00
         YearsAtCompany           3.751728e+01
         YearsSinceLastPromotion  1.037935e+01
         YearsWithCurrManager     1.272582e+01
         dtype: float64
```

g. kurt()

```
In [14]: 1 Hike","TotalWorkingYears","TrainingTimesLastYear","YearsAtCompany","YearsSinceLastPromotion","YearsWithCurrManager"]].kurt()

Out[14]: Age                      -0.405951
         DistanceFromHome         -0.227045
         Education                -0.560569
         MonthlyIncome             1.000232
         NumCompaniesWorked        0.007287
         PercentSalaryHike        -0.302638
         TotalWorkingYears         0.912936
         TrainingTimesLastYear     0.491149
         YearsAtCompany            3.923864
         YearsSinceLastPromotion   3.601761
         YearsWithCurrManager      0.167949
         dtype: float64
```

4. Inference of the analyst:

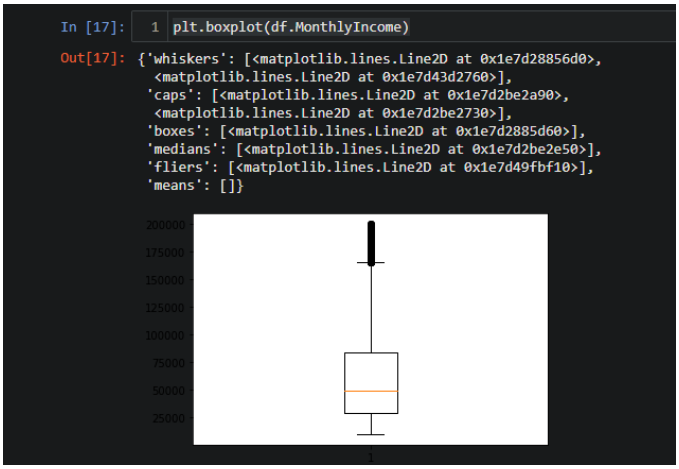
1. All the above variables show positive skewness except TrainingTimesLastYear which is a negatively skewed and Age is Normally distributed.

2. Age, DistanceFromHome, Education & PercentSalaryHike are Platykurtic. NumCompaniesWorked, TotalWorkingYears, TrainingTimesLastYear, YearsWithCurrManager are Mesokurtic.

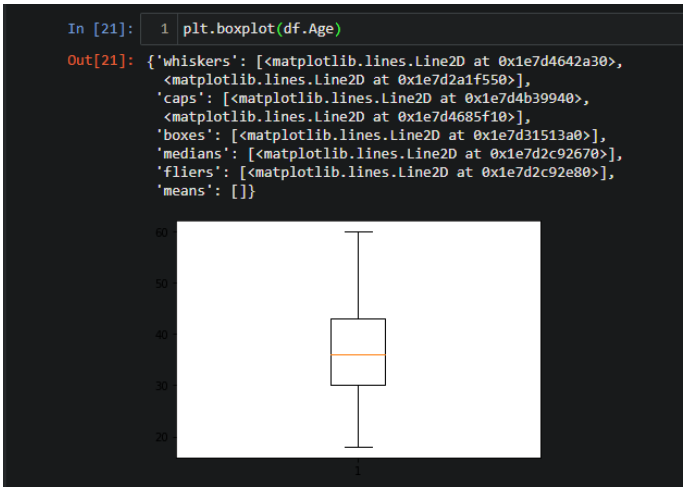
MonthlyIncome, YearsAtCompany, YearsSinceLastPromotion are are Leptokurtic.

5. Outliers

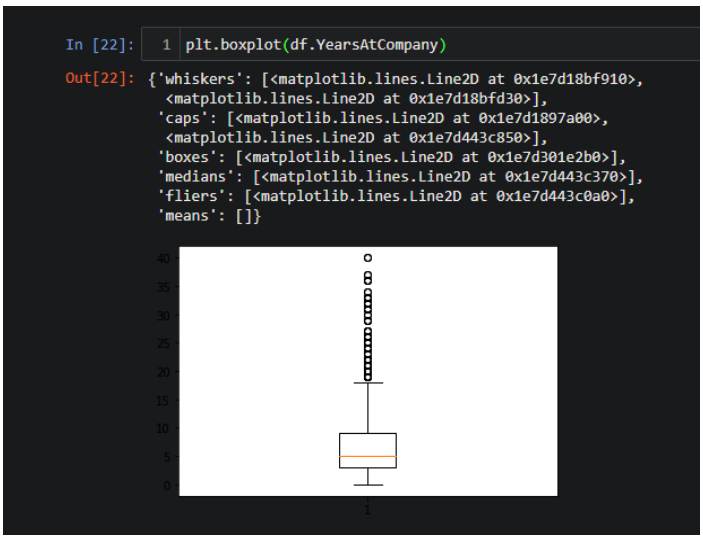
1. MonthlyIncome is Positively Skewed with many outliers.



2. Age is normally skewed with no outliers.



3. YearsAtCompany is negatively skewed with many outliers.



4. PercentSalaryHike is positively skewed with no outliers.