1.Luanching

In [1]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

# Dataset

df = pd.read_csv("general_data.csv")
```

In [2]:

```
df.head()
```

Out[2]:

	Age	Attrition	BusinessTravel	Department	DistanceFromHome	Education	Educa
0	51	No	Travel_Rarely	Sales	6	2	Life
1	31	Yes	Travel_Frequently	Research & Development	10	1	Life
2	32	No	Travel_Frequently	Research & Development	17	4	
3	38	No	Non-Travel	Research & Development	2	5	Life
4	32	No	Travel_Rarely	Research & Development	10	1	

In [5]:

```
# Columns in our dataset

df.columns
```

Out[5]:

```
Index(['Age', 'Attrition', 'BusinessTravel', 'Department', 'DistanceFromHo
me',
       'Education', 'EducationField', 'EmployeeCount', 'EmployeeID', 'Gend
er',
       'JobLevel', 'JobRole', 'MaritalStatus', 'MonthlyIncome',
       'NumCompaniesWorked', 'Over18', 'PercentSalaryHike', 'StandardHour
s',
       'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear',
       'YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManage
r'],
      dtype='object')
```

2.Data Treatment

In [8]:

```
df.isnull().sum()
```

Out[8]:

```
Age                         0
Attrition                   0
BusinessTravel              0
Department                  0
DistanceFromHome            0
Education                   0
EducationField              0
EmployeeCount               0
EmployeeID                  0
Gender                      0
JobLevel                    0
JobRole                     0
MaritalStatus               0
MonthlyIncome               0
NumCompaniesWorked         19
Over18                      0
PercentSalaryHike           0
StandardHours               0
StockOptionLevel            0
TotalWorkingYears           9
TrainingTimesLastYear       0
YearsAtCompany              0
YearsSinceLastPromotion     0
YearsWithCurrManager        0
dtype: int64
```

In [136]:

```
df.dropna().head()
```

Out[136]:

	Age	Attrition	BusinessTravel	Department	DistanceFromHome	Education	Educa
0	51	No	Travel_Rarely	Sales	6	2	Life
1	31	Yes	Travel_Frequently	Research & Development	10	1	Life
2	32	No	Travel_Frequently	Research & Development	17	4	
3	38	No	Non-Travel	Research & Development	2	5	Life
4	32	No	Travel_Rarely	Research & Development	10	1	

In [13]:

```
print("Duplicates In dataset :", df.duplicated().sum())
```

```
Duplicates In dataset : 0
```

3.Univariate Analysis

In [18]:

```
df1 = df[df.columns].describe()
df1
```

Out[18]:

	Age	DistanceFromHome	Education	EmployeeCount	EmployeeID	
count	4410.000000	4410.000000	4410.000000	4410.0	4410.000000	44
mean	36.923810	9.192517	2.912925	1.0	2205.500000	
std	9.133301	8.105026	1.023933	0.0	1273.201673	
min	18.000000	1.000000	1.000000	1.0	1.000000	
25%	30.000000	2.000000	2.000000	1.0	1103.250000	
50%	36.000000	7.000000	3.000000	1.0	2205.500000	
75%	43.000000	14.000000	4.000000	1.0	3307.750000	
max	60.000000	29.000000	5.000000	1.0	4410.000000	

In [105]:

```
mean=df[['Age', 'DistanceFromHome','MonthlyIncome','TotalWorkingYears',
    'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].mean()
mean
```

Out[105]:

```
Age                          36.923810
DistanceFromHome              9.192517
MonthlyIncome             65029.312925
TotalWorkingYears            11.279936
YearsAtCompany                7.008163
YearsSinceLastPromotion       2.187755
YearsWithCurrManager          4.123129
dtype: float64
```

In [104]:

```
med = df[['Age', 'DistanceFromHome','MonthlyIncome','TotalWorkingYears',
    'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].median()
med
```

Out[104]:

```
Age                          36.0
DistanceFromHome              7.0
MonthlyIncome             49190.0
TotalWorkingYears            10.0
YearsAtCompany                5.0
YearsSinceLastPromotion       1.0
YearsWithCurrManager          3.0
dtype: float64
```

In [124]:

```
mode = df[['Age', 'DistanceFromHome','MonthlyIncome','TotalWorkingYears',
    'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].mode().T
mode
```

Out[124]:

	0
Age	35.0
DistanceFromHome	2.0
MonthlyIncome	23420.0
TotalWorkingYears	10.0
YearsAtCompany	5.0
YearsSinceLastPromotion	0.0
YearsWithCurrManager	2.0

In [111]:

```
var = df[['Age', 'DistanceFromHome','MonthlyIncome','TotalWorkingYears',
    'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].var()
var
```

Out[111]:

```
Age                       8.341719e+01
DistanceFromHome          6.569144e+01
MonthlyIncome             2.215480e+09
TotalWorkingYears         6.056298e+01
YearsAtCompany            3.751728e+01
YearsSinceLastPromotion   1.037935e+01
YearsWithCurrManager      1.272582e+01
dtype: float64
```

In [112]:

```
std = df[['Age', 'DistanceFromHome','MonthlyIncome','TotalWorkingYears',
    'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].std()
std
```

Out[112]:

```
Age                           9.133301
DistanceFromHome              8.105026
MonthlyIncome             47068.888559
TotalWorkingYears             7.782222
YearsAtCompany                6.125135
YearsSinceLastPromotion       3.221699
YearsWithCurrManager          3.567327
dtype: float64
```

In [113]:

```
sk = df[['Age', 'DistanceFromHome','MonthlyIncome','TotalWorkingYears',
    'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].skew()
sk
```

Out[113]:

```
Age                        0.413005
DistanceFromHome           0.957466
MonthlyIncome              1.368884
TotalWorkingYears          1.116832
YearsAtCompany             1.763328
YearsSinceLastPromotion    1.982939
YearsWithCurrManager       0.832884
dtype: float64
```

In [114]:

```
kt = df[['Age', 'DistanceFromHome','MonthlyIncome','TotalWorkingYears',
    'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].kurt()
kt
```

Out[114]:

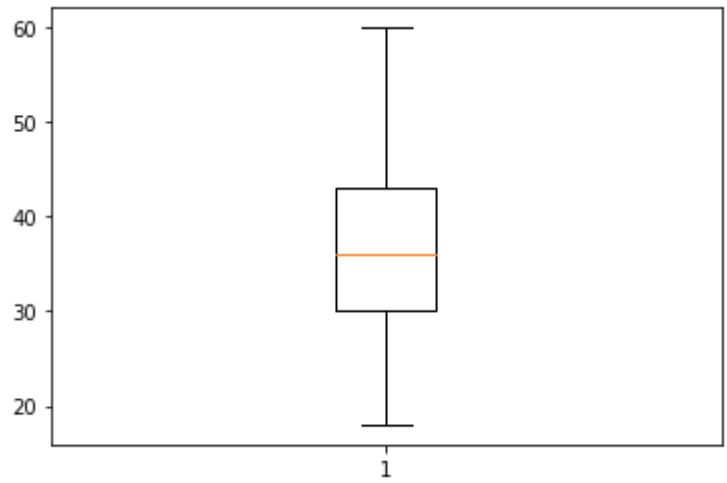
```
Age                       -0.405951
DistanceFromHome          -0.227045
MonthlyIncome              1.000232
TotalWorkingYears          0.912936
YearsAtCompany             3.923864
YearsSinceLastPromotion    3.601761
YearsWithCurrManager       0.167949
dtype: float64
```

Inference from the Above analysis :

- Mean Age Forms a near Normal Distribution with 13 Years of IQR
- All the above Variables show +Ve Skewness : Age & Distance from Home (Lipokurtic)

In [127]:

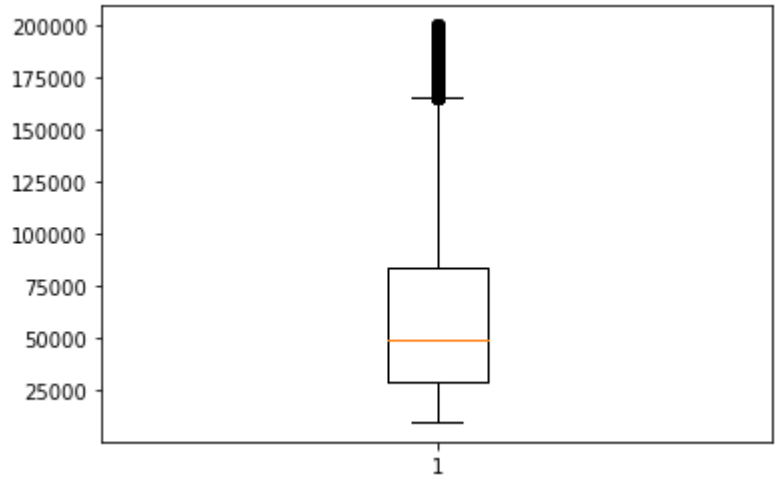
```
plt.boxplot(df.Age);
```



Age is Normally Distributed without any Outliers

In [129]:

```
plt.boxplot(df.MonthlyIncome);
```



MontlyIncome is Right Skewed With Several Outliers

In [ ]:

```

```