Assignment Day 7
By Akash Nauhwar

Case Study : Problem Statement A large company named XYZ, employs, at any given point of time, around 4000 employees. However, every year, around 15% of its employees leave the company and need to be replaced with the talent pool available in the job market. The management believes that this level of attrition (employees leaving, either on their own or because they got fired) is bad for the company, because of the following reasons -

The former employees' projects get delayed, which makes it difficult to meet timelines, resulting in a reputation loss among consumers and partners
A sizeable department has to be maintained, for the purposes of recruiting new talent. More often than not, the new employees have to be trained for the job and/or given time to acclimatize themselves to the company
Hence, the management has contracted an HR analytics firm to understand what factors they should focus on, in order to curb attrition. In other words, they want to know what changes they should make to their workplace, in order to get most of their employees to stay. Also, they want to know which of these variables is most important and needs to be addressed right away.

Since you are one of the star analysts at the firm, this project has been given to you.

Goal of the case study You are required to model the probability of attrition. The results thus obtained will be used by the management to understand what changes they should make to their workplace, in order to get most of their employees to stay.

Columns
EmployeeIDEmployee number/id
EnvironmentSatisfactionWork Environment Satisfaction Level
JobSatisfactionJob Involvement Level Job Involvement Level Job Involvement Level
WorkLifeBalanceWork life balance level

Step1 – Launching

```
In [1]: import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
```

```
In [2]: database = pd.read_csv("C:/Users/vishalshivhare/Desktop/general_data.csv")
```

```
In [3]: database.head()
```

Out[3]:

	Age	Attrition	BusinessTravel	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeID	Gender	...	NumCompaniesWork
0	51	No	Travel_Rarely	Sales	6	2	Life Sciences	1	1	Female	...	
1	37	Yes	Travel_Rarely	Human Resources	9	3	Human Resources	1	1313	Male	...	N
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	...	

5 rows × 24 columns

Step 2 - Data Treatment:

```
In [5]: databasea = database.drop_duplicates()
```

```
In [6]: database = database.dropna()
```

```
In [7]: database.head()
```

Out[7]:

	Age	Attrition	BusinessTravel	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeID	Gender	...	NumCompaniesWork
0	51	0	Travel_Rarely	Sales	6	2	Life Sciences	1	1	Female	...	
2	32	0	Travel_Frequently	Research & Development	17	4	Other	1	3	Male	...	
3	38	0	Non-Travel	Research & Development	2	5	Life Sciences	1	4	Male	...	
4	32	0	Travel_Rarely	Research & Development	10	1	Medical	1	5	Male	...	
5	46	0	Travel_Rarely	Research & Development	8	3	Life Sciences	1	6	Female	...	

5 rows × 24 columns

Step 3 – Univariate Analysis:

```
In [9]: databasea.mean()
```

```
Out[9]: Age                          36.923810
        Attrition                     0.161224
        DistanceFromHome              9.192517
        Education                     2.912925
        EmployeeCount                 1.000000
        EmployeeID                 2205.500000
        JobLevel                      2.063946
        MonthlyIncome             65029.312925
        NumCompaniesWorked            2.694830
        PercentSalaryHike            15.209524
        StandardHours                 8.000000
        StockOptionLevel              0.793878
        TotalWorkingYears            11.279936
        TrainingTimesLastYear         2.799320
        YearsAtCompany                7.008163
        YearsSinceLastPromotion       2.187755
        YearsWithCurrManager          4.123129
        dtype: float64
```

```
In [10]: database[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears', 'T
```

Out[10]:

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

```
In [11]: database.median()
```

```
Out[11]: Age                          36.0
         Attrition                     0.0
         DistanceFromHome              7.0
         Education                     3.0
         EmployeeCount                 1.0
         EmployeeID                 2208.5
         JobLevel                      2.0
         MonthlyIncome             49190.0
         NumCompaniesWorked            2.0
         PercentSalaryHike            14.0
         StandardHours                 8.0
         StockOptionLevel              1.0
         TotalWorkingYears            10.0
         TrainingTimesLastYear         3.0
         YearsAtCompany                5.0
         YearsSinceLastPromotion       1.0
         YearsWithCurrManager          3.0
         dtype: float64
```

```
In [12]: # Varience and Standard Deviation

In [13]: database.var()

Out[13]: Age                        8.348974e+01
         Attrition                  1.350321e-01
         DistanceFromHome           6.569744e+01
         Education                  1.050068e+00
         EmployeeCount              0.000000e+00
         EmployeeID                 1.617192e+06
         JobLevel                   1.223490e+00
         MonthlyIncome              2.222397e+09
         NumCompaniesWorked         6.239165e+00
         PercentSalaryHike          1.341762e+01
         StandardHours              0.000000e+00
         StockOptionLevel           7.265814e-01
         TotalWorkingYears          6.061739e+01
         TrainingTimesLastYear      1.662558e+00
         YearsAtCompany             3.756894e+01
         YearsSinceLastPromotion    1.040059e+01
         YearsWithCurrManager       1.274257e+01
         dtype: float64

In [14]: databasea.std()

Out[14]: Age                         9.133301
         Attrition                   0.367780
         DistanceFromHome            8.105026
         Education                   1.023933
         EmployeeCount               0.000000
         EmployeeID               1273.201673
         JobLevel                    1.106689
         MonthlyIncome           47068.888559
         NumCompaniesWorked          2.498887
         PercentSalaryHike           3.659108
         StandardHours               0.000000
         StockOptionLevel            0.851883
         TotalWorkingYears           7.782222
         TrainingTimesLastYear       1.288978
         YearsAtCompany              6.125135
         YearsSinceLastPromotion     3.221699
         YearsWithCurrManager        3.567327
         dtype: float64
```

In [15]:
```
database.skew()
```

Out[15]:
```
Age                        0.413048
Attrition                  1.846529
DistanceFromHome           0.955517
Education                 -0.288977
EmployeeCount              0.000000
EmployeeID                -0.002335
JobLevel                   1.021797
MonthlyIncome              1.367457
NumCompaniesWorked         1.029174
PercentSalaryHike          0.819510
StandardHours              0.000000
StockOptionLevel           0.967263
TotalWorkingYears          1.115419
TrainingTimesLastYear      0.551818
YearsAtCompany             1.764619
YearsSinceLastPromotion    1.980992
YearsWithCurrManager       0.834277
dtype: float64
```

In [16]:
```
database.kurt()
```

Out[16]:
```
Age                       -0.409517
Attrition                  1.410313
DistanceFromHome          -0.230691
Education                 -0.565008
EmployeeCount              0.000000
EmployeeID                -1.198607
JobLevel                   0.388189
MonthlyIncome              0.990836
NumCompaniesWorked         0.014307
PercentSalaryHike         -0.306951
StandardHours              0.000000
StockOptionLevel           0.356755
TotalWorkingYears          0.909316
TrainingTimesLastYear      0.494215
YearsAtCompany             3.930726
YearsSinceLastPromotion    3.592162
YearsWithCurrManager       0.170703
dtype: float64
```

Inference from the analysis:

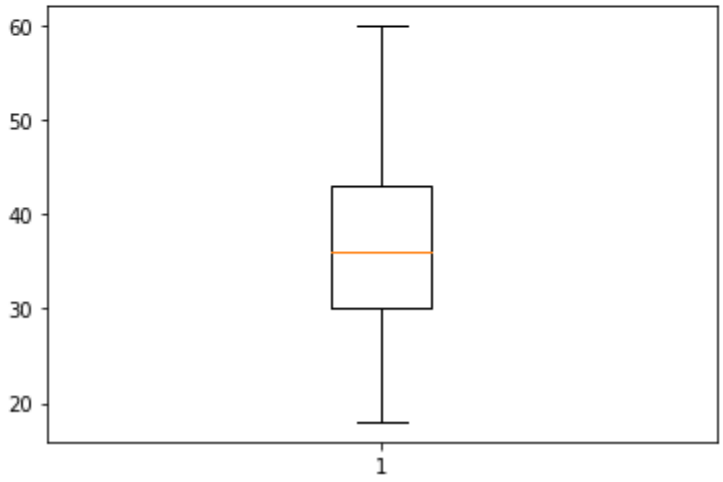
- All the above variables show positive skewness; while Age & Mean_distance_from_home are leptokurtic and all other variables are platykurtic.
- The Mean_Monthly_Income's IQR is at 54K suggesting company wide attrition across all income bands
- Mean age forms a near normal distribution with 13 years of IQR

Outliers:

Age is normally distributed without any outliers

```
In [18]: plt.boxplot(database.Age)
```

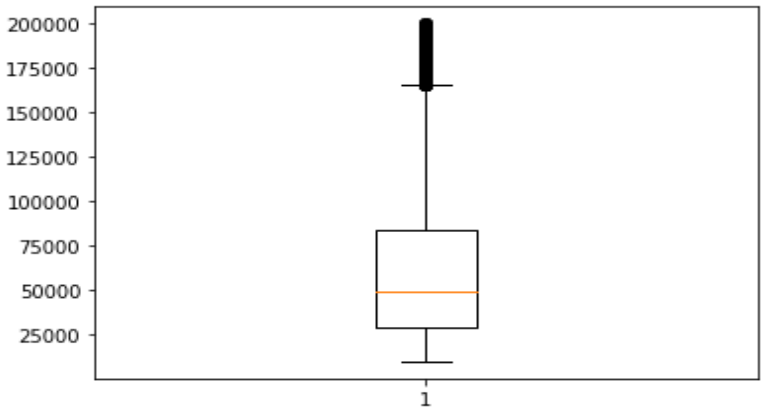
```
Out[18]: {'whiskers': [<matplotlib.lines.Line2D at 0x223bbbac188>,
                       <matplotlib.lines.Line2D at 0x223bbb98d08>],
          'caps': [<matplotlib.lines.Line2D at 0x223bbb98d88>,
                   <matplotlib.lines.Line2D at 0x223bbb50e08>],
          'boxes': [<matplotlib.lines.Line2D at 0x223bbba3f08>],
          'medians': [<matplotlib.lines.Line2D at 0x223bbb50f88>],
          'fliers': [<matplotlib.lines.Line2D at 0x223bbb4af08>],
          'means': []}
```



Monthly Income is Right skewed with several outliers

```
In [19]: plt.boxplot(database.MonthlyIncome)
```

```
Out[19]: {'whiskers': [<matplotlib.lines.Line2D at 0x223bc2aa788>,
                       <matplotlib.lines.Line2D at 0x223bc2aaf48>],
          'caps': [<matplotlib.lines.Line2D at 0x223bc2ae748>,
                   <matplotlib.lines.Line2D at 0x223bc2aeec8>],
          'boxes': [<matplotlib.lines.Line2D at 0x223bc2a6e08>],
          'medians': [<matplotlib.lines.Line2D at 0x223bc2b46c8>],
          'fliers': [<matplotlib.lines.Line2D at 0x223bc2b4e48>],
          'means': []}
```



Years at company is also Right Skewed with several outliers observed

```
In [20]: plt.boxplot(database.YearsAtCompany)
```

```
Out[20]: {'whiskers': [<matplotlib.lines.Line2D at 0x223bc320d08>,
                       <matplotlib.lines.Line2D at 0x223bc320e88>],
          'caps': [<matplotlib.lines.Line2D at 0x223bc2d0d88>,
                   <matplotlib.lines.Line2D at 0x223bc2d0f08>],
          'boxes': [<matplotlib.lines.Line2D at 0x223bc3204c8>],
          'medians': [<matplotlib.lines.Line2D at 0x223bc32ae88>],
          'fliers': [<matplotlib.lines.Line2D at 0x223bc32e708>],
          'means': []}
```