

# Assignment – Attrition Analysis

## Step 1: Load the sheet/Data

```
import pandas as pd
```

```
import matplotlib.pyplot as plt
```

```
dataset = pd.read_csv("D:/AI_ML_Course/Day 7/general_data.csv")
```

```
dataset.columns
```

```
Out[3]:
```

```
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')
```

## Step 2: Data Treatment

`dataset.isnull()`

Out[4]:

	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.duplicated()`

Out[6]:

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.drop\_duplicates()

Out[7]:

	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]

## Step 3: Univariate Analysis:

### A) Analysis with Complete Dataset

```
dataset1=dataset[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked',  
'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear',  
'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].describe()
```

#### Dataset1

dataset1 - DataFrame

Index	Age	DistanceFromHome	Education	MonthlyIncome	NumCompaniesWorked	PercentSalaryHike	TotalWorkingYears	TrainingTimesLastYear	YearsAtCompany	YearsSinceLastPromotion	YearsWithCurrManager
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

```
dataset1=dataset[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked',  
'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear',  
'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].median()
```

dataset1 - Series

Index	0
Age	36
DistanceFromHome	7
Education	3
MonthlyIncome	49190
NumCompaniesWorked	2
PercentSalaryHike	14
TotalWorkingYears	10
TrainingTimesLastYear	3
YearsAtCompany	5
YearsSinceLastPromotion	1
YearsWithCurrManager	3

```
dataset1=dataset[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked',
'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear',
'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].mode()
```

dataset1 - DataFrame

— 📄

Index	Age	DistanceFromHome	Education	MonthlyIncome	NumCompaniesWorked	PercentSalaryHike	TotalWorkingYears	TrainingTimesLastYear	YearsAtCompany	YearsSinceLastPromotion	YearsWithCurrManager
0	35	2	3	23420	1	11	10	2	5	0	2

```
dataset1=dataset[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked',
'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear',
'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].var()
```

dataset1 - Series

Index	0
Age	83.4172
DistanceFromHome	65.6914
Education	1.04844
MonthlyIncome	2.21548e+09
NumCompaniesWorked	6.24444
PercentSalaryHike	13.3891
TotalWorkingYears	60.563
TrainingTimesLastYear	1.66146
YearsAtCompany	37.5173
YearsSinceLastPromotion	10.3793
YearsWithCurrManager	12.7258

```
dataset1=dataset[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked',
'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear',
'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].skew()
```

dataset1 - Series

Index	0
Age	0.413005
DistanceFromHome	0.957466
Education	-0.289484
MonthlyIncome	1.36888
NumCompaniesWorked	1.02677
PercentSalaryHike	0.820569
TotalWorkingYears	1.11683
TrainingTimesLastYear	0.552748
YearsAtCompany	1.76333
YearsSinceLastPromotion	1.98294
YearsWithCurrManager	0.832884

```
dataset1=dataset[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked',
'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear',
'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].kurt()
```

dataset1 - Series

Index	0
Age	-0.405951
DistanceFromHome	-0.227045
Education	-0.560569
MonthlyIncome	1.00023
NumCompaniesWorked	0.00728748
PercentSalaryHike	-0.302638
TotalWorkingYears	0.912936
TrainingTimesLastYear	0.491149
YearsAtCompany	3.92386
YearsSinceLastPromotion	3.60176
YearsWithCurrManager	0.167949

	Mean	Median	Mode	Variance	Std Deviation	IQR	Skewness	Kurtosis
Age (Yrs)	36.9	36	35	83.41	9.13	13	0.41	-0.41
DistanceFromHome (Km)	9.19	7	2	65.69	8.1	12	0.96	-0.23
Monthly Income (Rs)	65029	49190	23420	2215480270	47068	54690	1.37	1
PercentSalaryHike (%)	15	14	11	13.39	3.66	6	0.82	-0.3
TotalWorkingYears (Yrs)	11.29	10	10	60.56	7.78	9	1.12	0.91
YearsAtCompany (Yrs)	7	5	5	37.52	6.12	6	1.76	3.92
YearsSinceLastPromotion (Yrs)	2	1	0	10.38	3.22	3	1.98	3.6
YearsWithCurrManager (Yrs)	4	3	2	12.73	3.57	5	0.83	0.17

### Inference from the analysis:

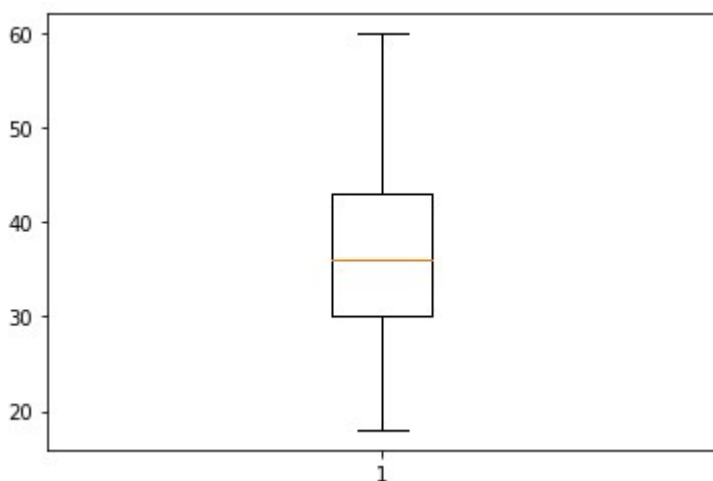
- All the above variables show positive skewness.
- Years\_At\_Company & Years\_Since\_LastPromotion are Leptokurtic i.e. more than 3 and all other variables are Platykurtic.
- The Mean\_Monthly\_Income's IQR is at 54K suggesting companywide attrition across all income bands
- Mean age forms a near normal distribution with 13 years of IQR
- Mean Distance\_From\_Home is 12 Km of IQR which is higher.

### Outliers:

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

```
box_plot=dataset.Age
```

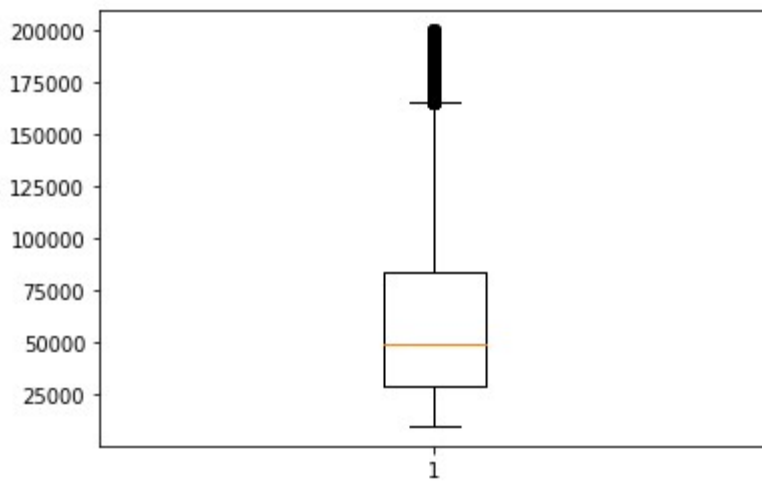
```
mplt. boxplot(box_plot)
```



Age is normally distributed without any Outliers

```
box_plot=dataset.MonthlyIncome
```

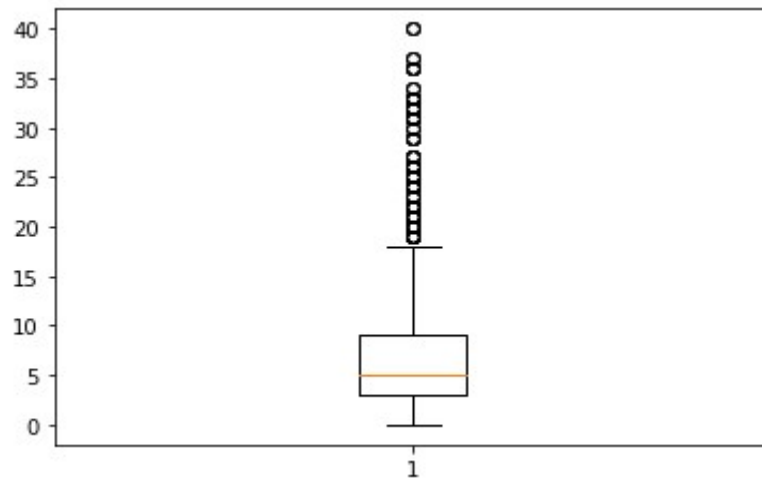
```
mplt.boxplot(box_plot)
```



Monthly Income is right skewed with several Outliers

```
box_plot=dataset.YearsAtCompany
```

```
mplt.boxplot(box_plot)
```



Years at company is also Right skewed with several Outliers



## B) Analysis with dataset having Attrition as Yes

```
dataset1=dataset[dataset['Attrition']=='Yes']
```

```
dataset2=dataset1[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked',  
'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear',  
'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].describe()
```

dataset2 - DataFrame

Index	Age	DistanceFromHome	Education	MonthlyIncome	NumCompaniesWorked	PercentSalaryHike	TotalWorkingYears	TrainingTimesLastYear	YearsAtCompany	YearsSinceLastPromotion	YearsWithCurrManager
count	711	711	711	711	707	711	709	711	711	711	711
mean	33.0066	9.01266	2.87764	61682.6	2.93635	15.481	8.25529	2.65401	5.1308	1.94515	2.85232
std	9.6076	7.77237	1.01423	44792.1	2.67877	3.77529	7.16402	1.15483	5.9416	3.14863	3.13892
min	18	1	1	10090	0	11	0	0	0	0	0
25%	28	2	2	28440	1	12	3	2	1	0	0
50%	32	7	3	49080	1	14	7	3	3	1	2
75%	39	15	4	71040	5	18	10	3	7	2	5
max	58	29	5	198590	9	25	40	6	40	15	14

```
dataset2=dataset1[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked',  
'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear',  
'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].mean()
```

dataset2 - Series

Index	0
Age	33.6076
DistanceFromHome	9.01266
Education	2.87764
MonthlyIncome	61682.6
NumCompaniesWorked	2.93635
PercentSalaryHike	15.481
TotalWorkingYears	8.25529
TrainingTimesLastYear	2.65401
YearsAtCompany	5.1308
YearsSinceLastPromotion	1.94515
YearsWithCurrManager	2.85232

```
dataset2=dataset1[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked',
'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear',
'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].median()
```

dataset2 - Series

Index	0
Age	32
DistanceFromHome	7
Education	3
MonthlyIncome	49080
NumCompaniesWorked	1
PercentSalaryHike	14
TotalWorkingYears	7
TrainingTimesLastYear	3
YearsAtCompany	3
YearsSinceLastPromotion	1
YearsWithCurrManager	2

```
dataset2=dataset1[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked',
'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear',
'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].mode()
```

dataset2 - DataFrame

Index	Age	DistanceFromHome	Education	MonthlyIncome	NumCompaniesWorked	PercentSalaryHike	TotalWorkingYears	TrainingTimesLastYear	YearsAtCompany	YearsSinceLastPromotion	YearsWithCurrManager
0	29	2	3	25590	1	13	1	2	1	0	0

```
dataset2=dataset1[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked',
'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear',
'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].var()
```

dataset2 - Series

Index	0
Age	93.619
DistanceFromHome	60.4097
Education	1.02867
MonthlyIncome	2.00633e+09
NumCompaniesWorked	7.17583
PercentSalaryHike	14.2528
TotalWorkingYears	51.3232
TrainingTimesLastYear	1.33364
YearsAtCompany	35.3026
YearsSinceLastPromotion	9.91389
YearsWithCurrManager	9.85281

```
dataset2=dataset1[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked',
'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear',
'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].skew()
```

dataset2 - Series

Index	0
Age	0.712699
DistanceFromHome	0.96442
Education	-0.216055
MonthlyIncome	1.54071
NumCompaniesWorked	0.865799
PercentSalaryHike	0.763693
TotalWorkingYears	1.67932
TrainingTimesLastYear	0.422124
YearsAtCompany	2.67088
YearsSinceLastPromotion	2.20816
YearsWithCurrManager	1.02549

```
dataset2=dataset1[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked',
'PercentSalaryHike','TotalWorkingYears', 'TrainingTimesLastYear',
'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].kurt()
```

dataset2 - Series

Index	0
Age	-0.0731438
DistanceFromHome	-0.0417711
Education	-0.616494
MonthlyIncome	1.65818
NumCompaniesWorked	-0.548699
PercentSalaryHike	-0.418411
TotalWorkingYears	3.70317
TrainingTimesLastYear	0.960134
YearsAtCompany	9.45611
YearsSinceLastPromotion	4.77593
YearsWithCurrManager	0.242923

	Mean	Median	Mode	Variance	Std Deviation	IQR	Skewness	Kurtosis
Age (Yrs)	33	32	29	93	9.67	11	0.71	-0.073
DistanceFromHome (Km)	9	7	2	60	7.77	13	0.96	-0.04
MonthlyIncome (Rs)	61682	49080	25590		45000	42600	1.54	1.65
PercentSalaryHike (%)	15.4	14	13	14.25	3.77	6	0.76	-0.41
TotalWorkingYears (Yrs)	8	7	1	51	7.16	7	1.67	3.7
YearsAtCompany (Yrs)	5.13	3	1	35	5.94	6	2.67	9.45
YearsSinceLastPromotion (Yrs)	1.94	1	0	9.91	3.14	2	2.2	4.77
YearsWithCurrManager (Yrs)	2.85	2	0	9.85	3.13	5	1.02	0.24

### Inference from the analysis:

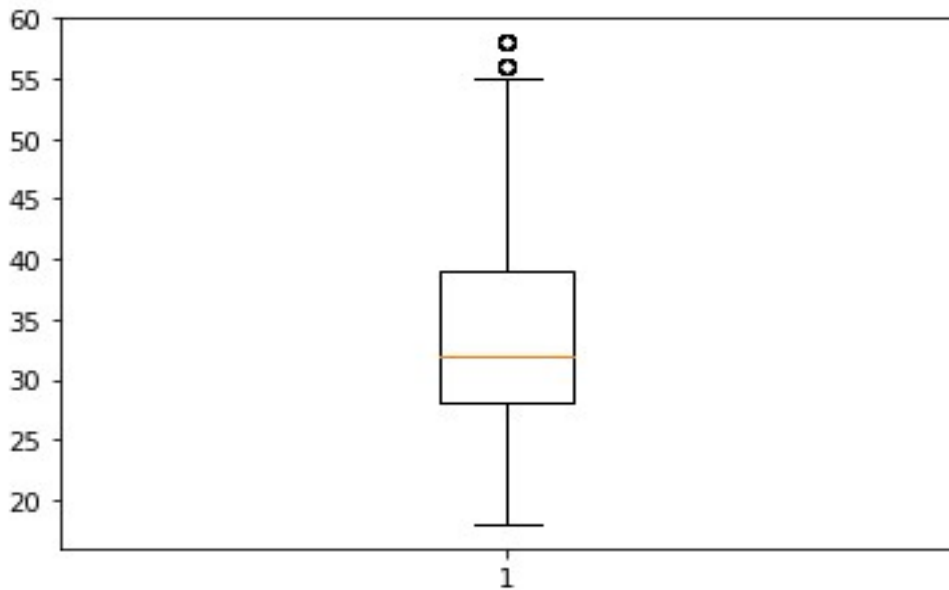
- All the above variables show positive skewness.
- Years\_At\_Company, Years\_Since\_LastPromotion & Total\_Working\_Years are Leptokurtic i.e. more than 3 and all other variables are Platykurtic.
- The Mean\_Monthly\_Income's IQR is at 42K suggesting companywide attrition across all income bands
- Mean age forms a near normal distribution with 11 years of IQR
- Mean Distance\_From\_Home is 13 Km of IQR which is higher.

## Outliers:

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

```
box_plot=dataset1.Age
```

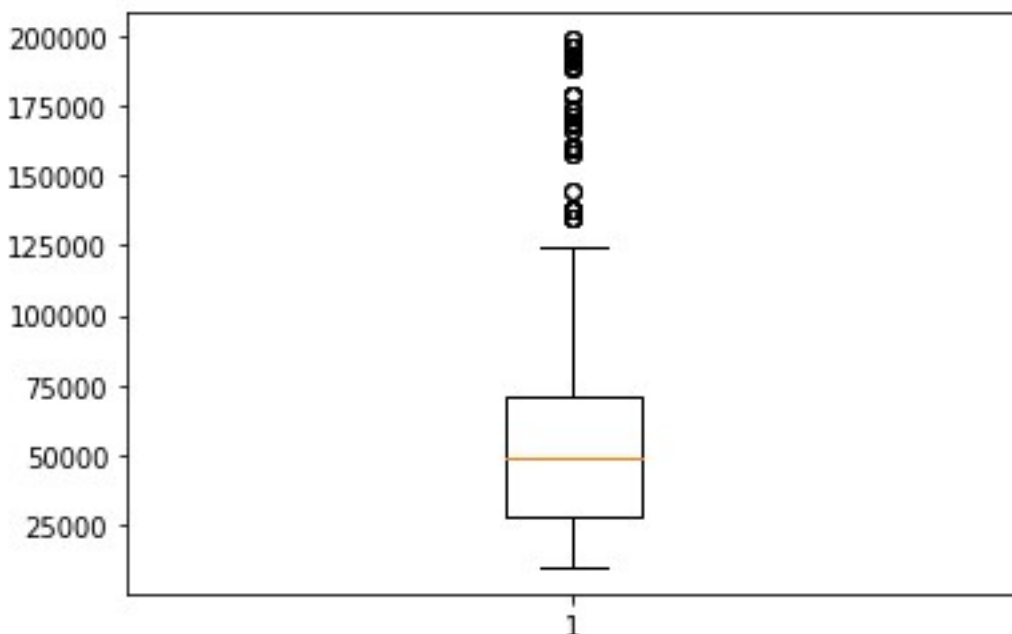
```
mplt.boxplot(box_plot)
```



Age is normally distributed but there are few Outliers. This might be retired/voluntary retired employees.

```
box_plot=dataset1.MonthlyIncome
```

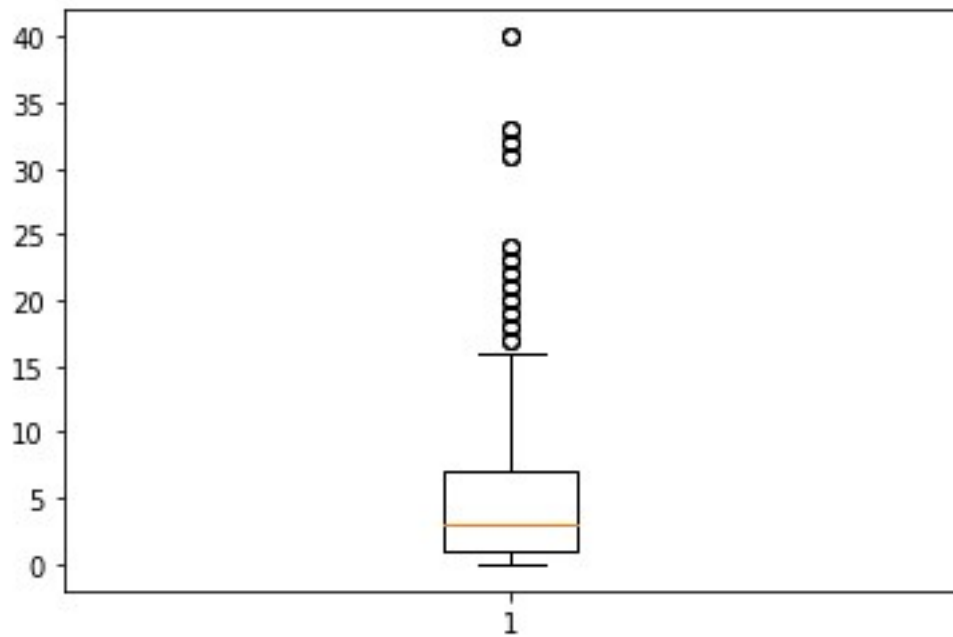
```
mplt.boxplot(box_plot)
```



Monthly Income is Right Skewed with several Outliers

```
box_plot=dataset1.YearsAtCompany
```

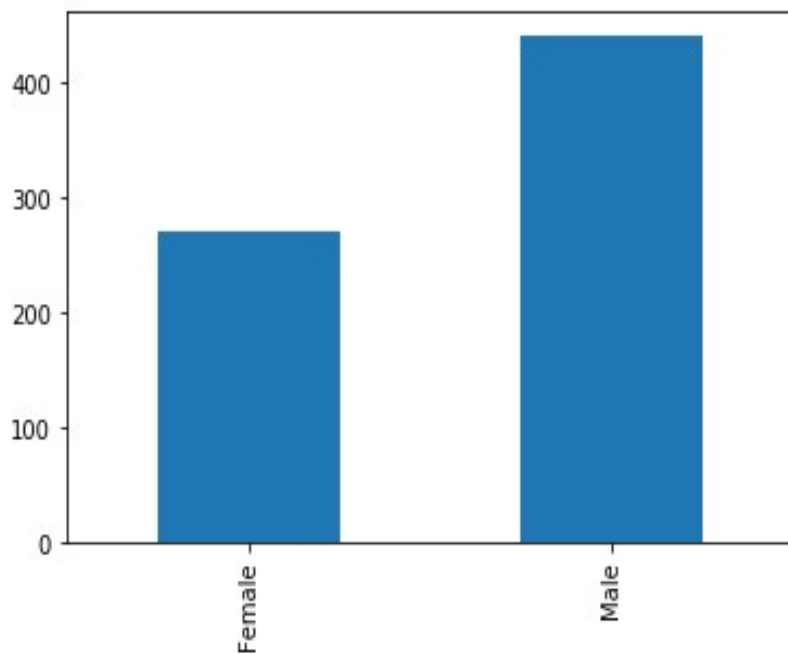
```
mplt.boxplot(box_plot)
```



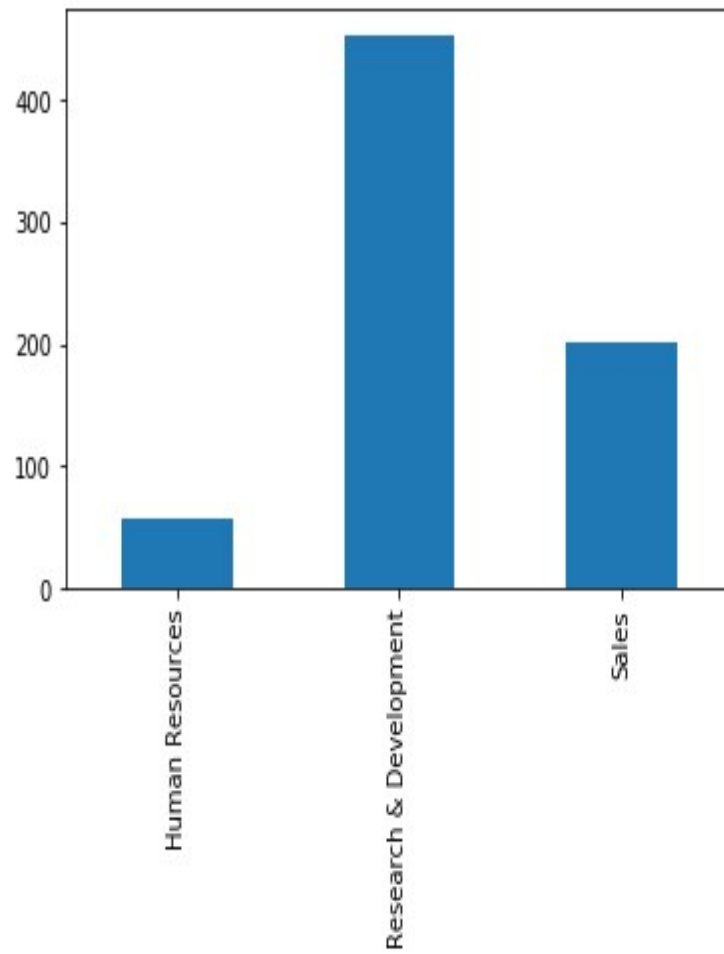
Years at company is also Right skewed with several Outliers

## Step 4: Visualization

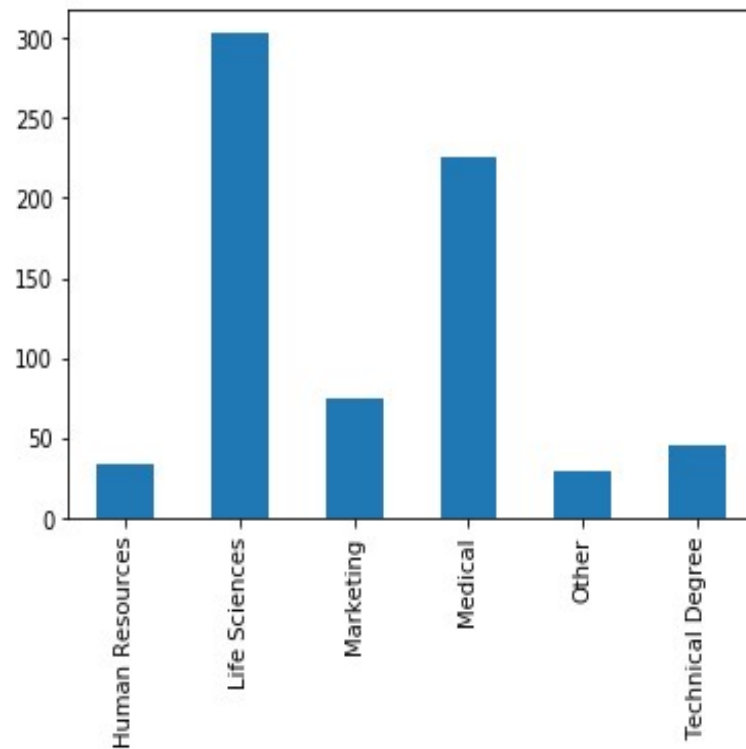
### A) Attrition Vs Gender



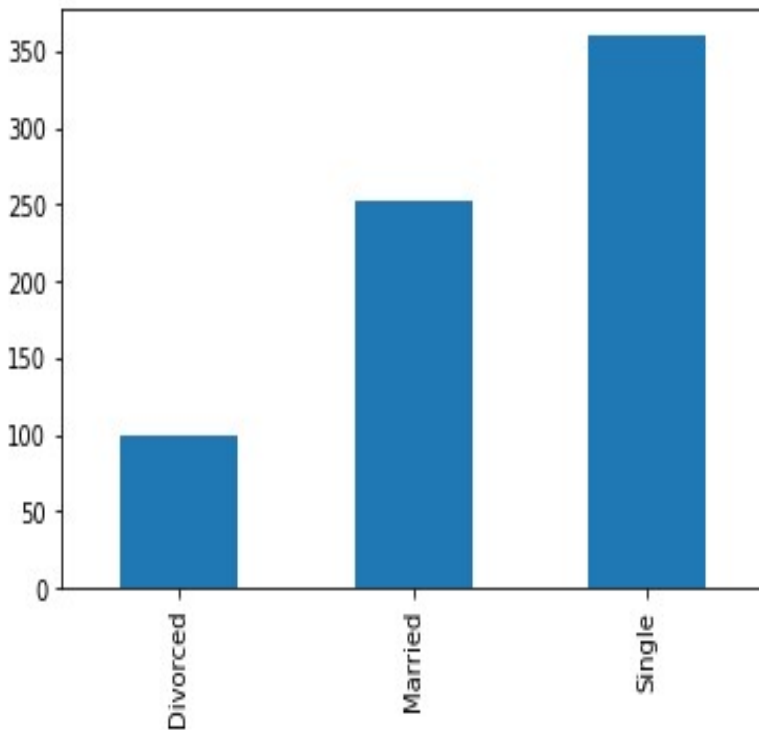
## B) Attrition Vs Department



## C) Attrition Vs Education Fields



#### D) Attrition Vs Marital Status



### Step 5 – Statistical Tests (Mann-Whitney)

#### Attrition Vs Distance from Home

```
from scipy.stats import mannwhitneyu  
  
dataset1=dataset[dataset['Attrition']=='Yes']  
  
a1 = dataset1.DistanceFromHome  
  
dataset2 = dataset[dataset['Attrition']=='No']  
  
a2 = dataset2.DistanceFromHome  
  
stat, p = mannwhitneyu(a1,a2)  
  
print(stat, p)  
  
1312110.0 0.4629185205822659
```

As the P value of 0.46 is  $> 0.05$ , the  $H_0$  is accepted and  $H_a$  is rejected.

$H_0$ : There is no significant difference in the Distance\_From\_Home between attrition (Y) and attrition (N)

$H_a$ : There is a significant difference in the Distance\_From\_Home between attrition (Y) and attrition (N)



### Attrition Vs Monthly Income

```
dataset1=dataset[dataset['Attrition']=='Yes']
```

```
a1 = dataset1.MonthlyIncome
```

```
dataset2 = dataset[dataset['Attrition']=='No']
```

```
a2 = dataset2.MonthlyIncome
```

```
stat, p = mannwhitneyu(a1,a2)
```

```
print(stat, p)
```

```
1264900.5 0.053577283839938566
```

As the P value of 0.053 is  $> 0.05$ , the  $H_0$  is accepted and  $H_a$  is rejected.

$H_0$ : There is no significant difference in the Monthly\_Income between attrition (Y) and attrition (N)

$H_a$ : There is a significant difference in the Monthly\_Income between attrition (Y) and attrition (N)

### Attrition Vs Years at Company

```
dataset1=dataset[dataset['Attrition']=='Yes']
```

```
a1 = dataset1.YearsAtCompany
```

```
dataset2 = dataset[dataset['Attrition']=='No']
```

```
a2 = dataset2.YearsAtCompany
```

```
stat, p = mannwhitneyu(a1,a2)
```

```
print(stat, p)
```

```
923238.0 6.047598261692858e-37
```

As the P value of 0.0 is  $< 0.05$ , the  $H_0$  is rejected and  $H_a$  is accepted.

$H_0$ : There is no significant difference in the Years\_At\_Company between attrition (Y) and attrition (N)

$H_a$ : There is a significant difference in the Years\_At\_Company between attrition (Y) and attrition (N)

### Attrition Vs YearsWithCurrManager

```
dataset1=dataset[dataset['Attrition']=='Yes']
```

```
a1 = dataset1.YearsWithCurrManager
```

```
dataset2 = dataset[dataset['Attrition']=='No']
```

```
a2 = dataset2.YearsWithCurrManager
```

```
stat, p = mannwhitneyu(a1,a2)
```

```
print(stat, p)
```

957253.5 1.2365483142169853e-31

As the P value of 0.0 is  $< 0.05$ , the  $H_0$  is rejected and  $H_a$  is accepted.

$H_0$ : There is no significant difference in the Years\_With\_Curr\_Manager between attrition (Y) and attrition (N)

$H_a$ : There is a significant difference in the Years\_With\_Curr\_Manager between attrition (Y) and attrition (N)

### Attrition Vs YearsSinceLastPromotion

```
dataset1=dataset[dataset['Attrition']=='Yes']
```

```
a1 = dataset1. YearsSinceLastPromotion
```

```
dataset2 = dataset[dataset['Attrition']=='No']
```

```
a2 = dataset2. YearsSinceLastPromotion
```

```
stat, p = mannwhitneyu(a1,a2)
```

```
print(stat, p)
```

1209366.0 0.0002021180346719736

As the P value of 0.0002 is  $< 0.05$ , the  $H_0$  is rejected and  $H_a$  is accepted.

$H_0$ : There is no significant difference in the Years\_Since\_Last\_Promotion between attrition (Y) and attrition (N)

$H_a$ : There is a significant difference in the Years\_Since\_Last\_Promotion between attrition (Y) and attrition (N)

## Step 6: Statistical Test (Separate T Test)

### Attrition Vs Distance from Home

```
from scipy.stats import ttest_ind
```

```
dataset1=dataset[dataset['Attrition']=='Yes']
```

```
a1 = dataset1.DistanceFromHome
```

```
dataset2 = dataset[dataset['Attrition']=='No']
```

```
a2 = dataset2.DistanceFromHome
```

```
stat, p = ttest_ind(a1,a2)
```

```
print(stat, p)
```

-0.6460416038042738 0.518286042805572

As the P value of 0.51 is  $> 0.05$ , the  $H_0$  is accepted and  $H_a$  is rejected.

$H_0$ : There is no significant difference in the Distance\_From\_Home between attrition (Y) and attrition (N)

$H_a$ : There is a significant difference in the Distance\_From\_Home between attrition (Y) and attrition (N)

### Attrition Vs Monthly Income

```
dataset1=dataset[dataset['Attrition']=='Yes']  
  
a1 = dataset1.MonthlyIncome  
  
dataset2 = dataset[dataset['Attrition']=='No']  
  
a2 = dataset2.MonthlyIncome  
  
stat, p = ttest_ind(a1,a2)  
  
print(stat, p)  
  
-2.0708863763619316 0.03842748490605113
```

As the P value of 0.038 is  $< 0.05$ , the  $H_0$  is rejected and  $H_a$  is accepted.

$H_0$ : There is no significant difference in the Monthly\_Income between attrition (Y) and attrition (N)

$H_a$ : There is a significant difference in the Monthly\_Income between attrition (Y) and attrition (N)

### Attrition Vs Years at Company

```
dataset1=dataset[dataset['Attrition']=='Yes']  
  
a1 = dataset1.YearsAtCompany  
  
dataset2 = dataset[dataset['Attrition']=='No']  
  
a2 = dataset2.YearsAtCompany  
  
stat, p = ttest_ind(a1,a2)  
  
print(stat, p)  
  
-9.004357011787226 3.163883122491456e-19
```

As the P value of 0.0 is  $< 0.05$ , the  $H_0$  is rejected and  $H_a$  is accepted.

$H_0$ : There is no significant difference in the Years\_At\_Company between attrition (Y) and attrition (N)

$H_a$ : There is a significant difference in the Years\_At\_Company between attrition (Y) and attrition (N)

### Attrition Vs YearsWithCurrManager

```
dataset1=dataset[dataset['Attrition']=='Yes']  
  
a1 = dataset1.YearsWithCurrManager  
  
dataset2 = dataset[dataset['Attrition']=='No']  
  
a2 = dataset2.YearsWithCurrManager  
  
stat, p = ttest_ind(a1,a2)  
  
print(stat, p)
```

-10.499379408703438 1.7339322652918153e-25

As the P value of 0.0 is  $< 0.05$ , the  $H_0$  is rejected and  $H_a$  is accepted.

$H_0$ : There is no significant difference in the Years\_With\_Curr\_Manager between attrition (Y) and attrition (N)

$H_a$ : There is a significant difference in the Years\_With\_Curr\_Manager between attrition (Y) and attrition (N)

### Attrition Vs YearsSinceLastPromotion

```
dataset1=dataset[dataset['Attrition']=='Yes']
```

```
a1 = dataset1. YearsSinceLastPromotion
```

```
dataset2 = dataset[dataset['Attrition']=='No']
```

```
a2 = dataset2. YearsSinceLastPromotion
```

```
stat, p = ttest_ind(a1,a2)
```

```
print(stat, p)
```

-2.1934039604328843 0.028330336189428353

As the P value of 0.028 is  $< 0.05$ , the  $H_0$  is rejected and  $H_a$  is accepted.

$H_0$ : There is no significant difference in the Years\_Since\_Last\_Promotion between attrition (Y) and attrition (N)

$H_a$ : There is a significant difference in the Years\_Since\_Last\_Promotion between attrition (Y) and attrition (N)

## Step 7: Correlation Analysis

In order to find the interdependency of the variables DistanceFromHome, MonthlyIncome, TotalWorkingYears, YearsAtCompany, YearsWithCurrManager, YearsSinceLastPromotion from that of Attrition, we executed the Correlation Analysis as follows.

```
import pandas as pd
```

```
dataset1 = pd.read_csv("D:/AI_ML_Course/Day 7/general_data.csv")
```

```
from scipy.stats import pearsonr
```

```
stats, pdfh=pearsonr(dataset1.Attrition,dataset1.DistanceFromHome)
```

```
print(stats, pdfh)
```

-0.009730141010179674 0.5182860428050771

```
stats, pmi=pearsonr(dataset1.Attrition,dataset1.MonthlyIncome)
```

```
print(stats, pmi)
```

-0.031176281698115007 0.03842748490600132

```
stats, pyac=pearsonr(dataset1.Attrition,dataset1.YearsAtCompany)
```

```
print(stats, pyac)
```

```
-0.1343922139899772 3.1638831224877484e-19
```

```
stats, pywcm=pearsonr(dataset1.Attrition,dataset1.YearsWithCurrManager)
```

```
print(stats, pywcm)
```

```
-0.15619931590162847 1.7339322652896276e-25
```

```
stats, pyslp=pearsonr(dataset1.Attrition,dataset1.YearsSinceLastPromotion)
```

```
print(stats, pyslp)
```

```
-0.03301877514258434 0.028330336189396753
```

The Inference for the above Analysis is as follows:

Attrition & Distance from Home:

As  $r = -0.009$ , there's low negative correlation between Attrition and DistanceFromHome

As the P value of 0.518 is  $> 0.05$ , we are accepting  $H_0$  and hence there's no significant correlation between Attrition & DistanceFromHome

Attrition & MonthlyIncome:

As  $r = -0.031$ , there's low negative correlation between Attrition and MonthlyIncome

As the P value of 0.038 is  $< 0.05$ , we are accepting  $H_a$  and hence there's significant correlation between Attrition & MonthlyIncome

Attrition & YearsAtCompany:

As  $r = -0.1343$ , there's low negative correlation between Attrition and YearsAtCompany

As the P value is  $0.0 < 0.05$ , we are accepting  $H_a$  and hence there's significant correlation between Attrition & YearsAtCompany

Attrition & YearsWithCurrManager:

As  $r = -0.1561$ , there's low negative correlation between Attrition and YearsWithCurrManager

As the P value is  $0.0 > 0.05$ , we are accepting  $H_a$  and hence there's significant correlation between Attrition & YearsWithCurrManager

Attrition & YearsSinceLastPromotion:

As  $r = -0.033$ , there's low negative correlation between Attrition and YearsSinceLastPromotion

As the P value 0.028 is  $< 0.05$ , we are accepting  $H_a$  and hence there's significant correlation between Attrition & YearsSinceLastPromotion