

# ATTRITION PROJECT

Jupyter Project Assignment Day-7 Last Checkpoint: Last Monday at 10:58 PM (autosaved)

Logout

File Edit View Insert Cell Kernel Navigate Widgets Help

Trusted Python 3

Run Code

```
In [24]: import pandas as pd
```

```
In [25]: dataset = pd.read_csv("/home/jatin/ML & AI/general_data.csv")
dataset1 = pd.read_excel("/home/jatin/ML & AI/data_dictionary.xlsx")
```

Csv file is read

```
In [26]: dataset
```

```
Out[26]:
```

	Age	Attrition	BusinessTravel	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeID	Gender	...	NumCompa
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	...	

```
In [27]: dataset2 = dataset.drop_duplicates()
```

Duplicates values r dropped and saved in dataset2

```
In [28]: dataset2
```

```
Out[28]:
```

	Age	Attrition	BusinessTravel	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeID	Gender	...	NumCompa
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	...	

File Edit View Insert Cell Kernel Navigate Widgets Help

Trusted Python 3

Code

```
14.000000 8.0 1.000000 7.000000 3.000000 3.000000 1.000000 2.000000
18.000000 8.0 1.000000 10.000000 3.000000 7.000000 2.000000 5.000000
25.000000 8.0 3.000000 40.000000 6.000000 40.000000 15.000000 14.000000
```

In [37]: dataset1.head()

Out[37]:

	Variable	Meaning	Levels
0	Age	Age of the employee	NaN
1	Attrition	Whether the employee left in the previous year...	NaN
2	BusinessTravel	How frequently the employees travelled for bus...	NaN
3	Department	Department in company	NaN
4	DistanceFromHome	Distance from home in kms	NaN

In [38]: dataset2['PercentSalaryHike'].mean()

Out[38]: 15.481012658227849

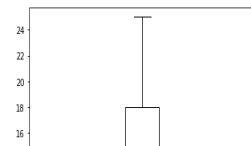
In [50]: dataset2['PercentSalaryHike'].median()

Out[50]: 14.0

In [40]: import matplotlib.pyplot as plt

In [42]: plt.boxplot(dataset2['PercentSalaryHike'])

Out[42]: {'whiskers': [<matplotlib.lines.Line2D at 0x77eff60bc90>, <matplotlib.lines.Line2D at 0x77eff572890>], 'caps': [<matplotlib.lines.Line2D at 0x77eff4ebfd0>, <matplotlib.lines.Line2D at 0x77eff4e4ed0>], 'boxes': [<matplotlib.lines.Line2D at 0x77eff27f210>], 'medians': [<matplotlib.lines.Line2D at 0x77eff305e50>], 'fliers': [<matplotlib.lines.Line2D at 0x77eff60be10>], 'means': []}



Here the mean > median for the percent salary hike thus data is positively skewed

File Edit View Insert Cell Kernel Navigate Widgets Help

Trusted Python 3

Code

In [43]: dataset2['PercentSalaryHike'].mean()

Out[43]: 15.481012658227849

In [44]: dataset2['PercentSalaryHike'].median()

Out[44]: 14.0

In [45]: dataset2['PercentSalaryHike'].mode()

Out[45]: 0 13  
dtype: int64

In [46]: dataset2['PercentSalaryHike'].skew()

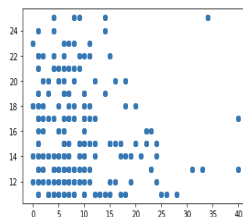
Out[46]: 0.7636927125846142

In [47]: dataset2['PercentSalaryHike'].kurt()

Out[47]: -0.418410993987842

In [51]: plt.scatter(dataset2['TotalWorkingYears'], dataset2['PercentSalaryHike'])

Out[51]: <matplotlib.collections.PathCollection at 0x77f018ee590>



In [20]: a['YearsSinceLastPromotion']

Out[20]: 1 1  
6 0  
13 9  
28 0  
30 0  
..

It is platykurtic kurtosis with flat shaped curve.

Jupyter Project Assignment Day-7 Last Checkpoint: Last Monday at 10:58 PM (autosaved)

File Edit View Insert Cell Kernel Navigate Widgets Help Trusted Python 3.1

```
In [29]: dataset2.dropna()
```

Out[29]:

	Age	Attrition	BusinessTravel	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeID	Gender	...	NumCompanies
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 x 24 columns

```
In [30]: dataset2.Attrition = dataset2.Attrition.replace('No','0')
dataset2.Attrition = dataset2.Attrition.replace('Yes','1')
```

Out[30]:

```
In [31]: dataset2.head()
```

Out[31]:

	BusinessTravel	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeID	Gender	...	NumCompaniesWorked	Over18
0	Travel_Rarely	Sales	6	2	Life Sciences	1	1	Female	...	1.0	Y
1	Travel_Frequently	Research & Development	10	1	Life Sciences	1	2	Female	...	0.0	Y
0	Travel_Frequently	Research & Development	17	4	Other	1	3	Male	...	1.0	Y
0	Non-Travel	Research & Development	2	5	Life Sciences	1	4	Male	...	3.0	Y
0	Travel_Rarely	Research & Development	10	1	Medical	1	5	Male	...	4.0	Y

Drop null values from the dataset2

Set Attrition value yes to binary 1 and no to binary 0

Jupyter Project Assignment Day-7 Last Checkpoint: Last Monday at 10:58 PM (autosaved)

File Edit View Insert Cell Kernel Navigate Widgets Help Trusted Python 3.1

```
Out[38]: 15.481812658227849
```

```
In [50]: dataset2['PercentSalaryHike'].median()
```

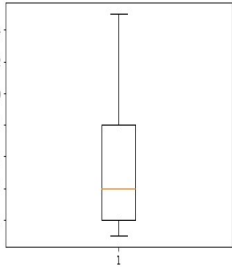
Out[50]: 14.0

```
In [40]: import matplotlib.pyplot as plt
```

```
In [42]: plt.boxplot(dataset2['PercentSalaryHike'])
```

Out[42]:

```
{'whiskers': [matplotlib.lines.Line2D at 0x7f7eff60bc90],
'caps': [matplotlib.lines.Line2D at 0x7f7eff572890],
'boxes': [matplotlib.lines.Line2D at 0x7f7eff4ebfd0],
'medians': [matplotlib.lines.Line2D at 0x7f7eff2f7210],
'fliers': [matplotlib.lines.Line2D at 0x7f7eff305e50],
'means': []}
```

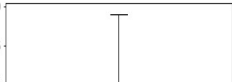


Here the median is lying at 14.0 and mean is lying slightly above it around 15.48

```
In [48]: plt.boxplot(dataset2['DistanceFromHome'])
```

Out[48]:

```
{'whiskers': [matplotlib.lines.Line2D at 0x7f7efaa605d0],
'caps': [matplotlib.lines.Line2D at 0x7f7efaa0f890],
'boxes': [matplotlib.lines.Line2D at 0x7f7efaa768d0],
'medians': [matplotlib.lines.Line2D at 0x7f7efdc7bd0],
'fliers': [matplotlib.lines.Line2D at 0x7f7efaa76250],
'means': []}
```



Jupyter Project Assignment Day-7 Last Checkpoint: Last Monday at 10:58 PM (autosaved)

File Edit View Insert Cell Kernel Navigate Widgets Help Trusted Python 3

min 35.0 2.0 3.0 23420.0 1.0 11.0 10.0 2.0

25% 35.0 2.0 3.0 23420.0 1.0 11.0 10.0 2.0

50% 35.0 2.0 3.0 23420.0 1.0 11.0 10.0 2.0

75% 35.0 2.0 3.0 23420.0 1.0 11.0 10.0 2.0

max 35.0 2.0 3.0 23420.0 1.0 11.0 10.0 2.0

```
In [147]: a[WorkingYears, 'TrainingTimesLastYear', 'YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManager']].skew()
```

```
In [148]: df0
```

```
Out[148]: Age                0.413805
DistanceFromHome          0.957466
Education                 -0.289484
MonthlyIncome             1.368884
NumCompaniesWorked        1.026767
PercentSalaryHike          0.820569
TotalWorkingYears         1.116832
TrainingTimesLastYear      0.552748
YearsAtCompany             1.763328
YearsSinceLastPromotion    1.982939
YearsWithCurrManager       0.832884
dtype: float64
```

Skewness of the given data

```
In [143]: df1 = dataset2[['Age', 'DistanceFromHome', 'Education', 'MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike', 'TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManager']].skew()
```

```
In [145]: df1
```

```
Out[145]: 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
```

Kurtosis of the given data

```
In [106]: dataset2.Attrition = dataset2.Attrition.replace('No', '0')
dataset2.Attrition = dataset2.Attrition.replace('Yes', '1')
```

## INFERENCE:

- > All the data is positively skewed except while Age and Mean\_distance from \_home are leptokurtic and all other variables are platykurtic
- > Mean age forms a near normal distribution with 13 years of IQR.