

# DAY 7 PROJECT

We need to analyse the reasons of attrition in the company and measures to stop that.

We have used Spyder-Ide for data-Analysis.

Firstly, we import numpy, pandas and matplotlib.

```
Console 1/A x
In [58]: import pandas as pd
In [59]: import numpy as np
In [60]: import matplotlib.pyplot as plt
In [61]: dataset=pd.read_csv("general_data(1).csv")
Traceback (most recent call last):
  File "<ipython-input-61-564fbc328df1>", line 1, in <module>
    dataset=pd.read_csv("general_data(1).csv")
```

Import the data file which is named as **general\_data** as variable “**dataset**”.

The excel sheet is distributed in two parts using filtering:

>>>> Attrition==Yes (Variable – “**dataset\_yes**”)

>>>>Attrition==No (Variable – “**dataset\_no**”)

```
Console 1/A x
File "pandas\_libs\parsers.pyx", line 674, in
pandas._libs.parsers.TextReader._setup_parser_source
FileNotFoundError: [Errno 2] File general_data(1).csv does not exist: 'general_data(1).csv'

In [62]: dataset=pd.read_csv("general_data (1).csv")
In [63]: dataset_no=pd.read_excel("general_data.xlsx",sheet_name=2)
In [64]: dataset_yes=pd.read_excel("general_data.xlsx",sheet_name=1)
```

| Name        | Type      | Size       | Value  |
|-------------|-----------|------------|--|
| dataset     | DataFrame | (4382, 24) | Column names: Age, Attrition, BusinessTravel, Department, DistanceFrom ... |
| dataset_no  | DataFrame | (3677, 24) | Column names: Age, Attrition, BusinessTravel, Department, DistanceFrom ... |
| dataset_yes | DataFrame | (705, 24)  | Column names: Age, Attrition, BusinessTravel, Department, DistanceFrom ... |

## Data Cleansing:

### 1. Checking for null value:

```
In [65]: dataset.isnull()
Out[65]:
```

|      | 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]
```

Result: No null values

### 2. Checking for duplicates :

```
Variable explorer  Help  Plots  Files

Console 1/A x
[4410 rows x 24 columns]

In [66]: dataset.duplicated()
Out[66]:
```

|      |       |
|------|-------|
| 0    | False |
| 1    | False |
| 2    | False |
| 3    | False |
| 4    | False |
| ...  | ...   |
| 4405 | False |
| 4406 | False |
| 4407 | False |
| 4408 | False |
| 4409 | False |

```
Length: 4410, dtype: bool

In [67]: dataset.dropna()
Out[67]:
```

|   | Age | Attrition | ... | YearsSinceLastPromotion | YearsWithCurrManager |
|---|-----|-----------|-----|-------------------------|----------------------|
| 0 | 51  | No        | ... | 0                       | 0                    |

Result: No duplicate values

### 3. Checking for NA and dropping it:

Dataset:

```
In [67]: dataset.dropna()
Out[67]:
```

|      | 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                    |
| ...  | ... | ...       | ... | ...                     | ...                  |
| 4404 | 29  | No        | ... | 1                       | 5                    |
| 4405 | 42  | No        | ... | 0                       | 2                    |
| 4406 | 29  | No        | ... | 0                       | 2                    |
| 4407 | 25  | No        | ... | 1                       | 2                    |
| 4408 | 42  | No        | ... | 7                       | 8                    |

```
[4382 rows x 24 columns]
In [68]: dataset=dataset.dropna()
```

Result: The rows have been reduced due to removal of NA entries.

### Analysis In General:

Getting important information about each column:

```
dtype='object')

In [86]: dataset1=dataset[['Age','DistanceFromHome',
...:                       'Education', 'JobLevel','MonthlyIncome',
...:                       'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears',
'TrainingTimesLastYear',
...:                       'YearsAtCompany', 'YearsSinceLastPromotion',
'YearsWithCurrManager']].describe()
```

| dataset1 - DataFrame |       |       |                  |           |          |               |                    |                   |                   |                       |                |                         |                      |
|----------------------|-------|-------|------------------|-----------|----------|---------------|--------------------|-------------------|-------------------|-----------------------|----------------|-------------------------|----------------------|
|                      | Index | Age   | DistanceFromHorr | Education | JobLevel | MonthlyIncome | NumCompaniesWorked | PercentSalaryHike | TotalWorkingYears | TrainingTimesLastYear | YearsAtCompany | YearsSinceLastPromotion | YearsWithCurrManager |
|                      | count | 18    | 1                | 1         | 1        | 10690         | 0                  | 11                | 0                 | 0                     | 0              | 0                       | 0                    |
|                      | mean  | 30    | 2                | 2         | 1        | 29110         | 1                  | 12                | 6                 | 2                     | 3              | 0                       | 2                    |
|                      | std   | 36    | 7                | 3         | 2        | 49190         | 2                  | 14                | 10                | 3                     | 5              | 1                       | 3                    |
|                      | min   | 9...  | 8.1054           | 1.02473   | 1.10611  | 47142.3       | 2.49783            | 3.66301           | 7.78572           | 1.2894                | 6.12935        | 3.22499                 | 3.56967              |
|                      | 25%   | 36... | 9.199            | 2.91237   | 2.0639   | 65061.7       | 2.69329            | 15.2106           | 11.2903           | 2.79827               | 7.0105         | 2.19169                 | 4.1262               |
|                      | 50%   | 43    | 14               | 4         | 3        | 83790         | 4                  | 18                | 15                | 3                     | 9              | 3                       | 7                    |
|                      | 75%   | 60    | 29               | 5         | 5        | 199990        | 9                  | 25                | 40                | 6                     | 40             | 15                      | 17                   |
|                      | max   | 43... | 4382             | 4382      | 4382     | 4382          | 4382               | 4382              | 4382              | 4382                  | 4382           | 4382                    | 4382                 |
|                      |       |       |                  |           |          |               |                    |                   |                   |                       |                |                         |                      |

Median ,Mode and Variance:

```
'YearsWithCurrManager']]).describe()
In [89]: dataset2=dataset[['Age','DistanceFromHome',
...:      'Education', 'JobLevel','MonthlyIncome',
...:      'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears',
'TrainingTimesLastYear',
...:      'YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManager']].median()
In [90]:
```

| dataset2 - Series       |       |
|-------------------------|-------|
| Index                   | 0     |
| Age                     | 36    |
| DistanceFromHome        | 7     |
| Education               | 3     |
| JobLevel                | 2     |
| MonthlyIncome           | 49190 |
| NumCompaniesWorked      | 2     |
| PercentSalaryHike       | 14    |
| TotalWorkingYears       | 10    |
| TrainingTimesLastYear   | 3     |
| YearsAtCompany          | 5     |
| YearsSinceLastPromotion | 1     |
| YearsWithCurrManager    | 3     |

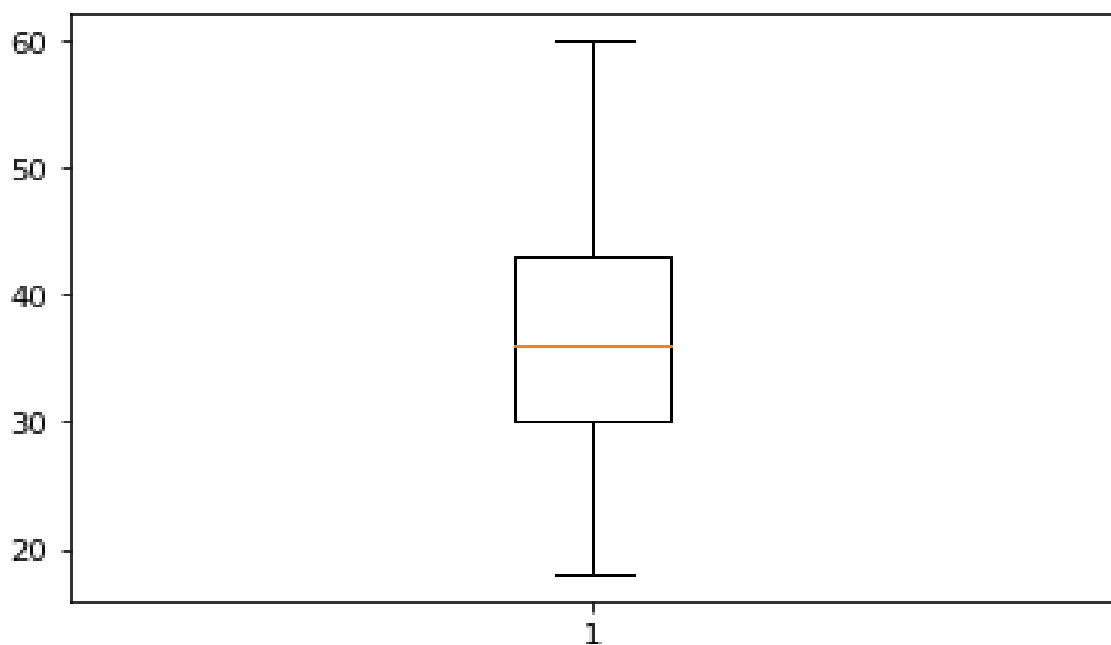
| Index                   | 0     |
|-------------------------|-------|
| Age                     | 35    |
| DistanceFromHome        | 2     |
| Education               | 3     |
| JobLevel                | 1     |
| MonthlyIncome           | 23420 |
| NumCompaniesWorked      | 1     |
| PercentSalaryHike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

| Index | 0          |
|-------|------------|
| 0     | 83.4897    |
| 1     | 65.6974    |
| 2     | 1.05007    |
| 3     | 1.22349    |
| 4     | 2.2224e+09 |
| 5     | 6.23917    |
| 6     | 13.4176    |
| 7     | 60.6174    |
| 8     | 1.66256    |
| 9     | 37.5689    |
| 10    | 10.4006    |
| 11    | 12.7426    |

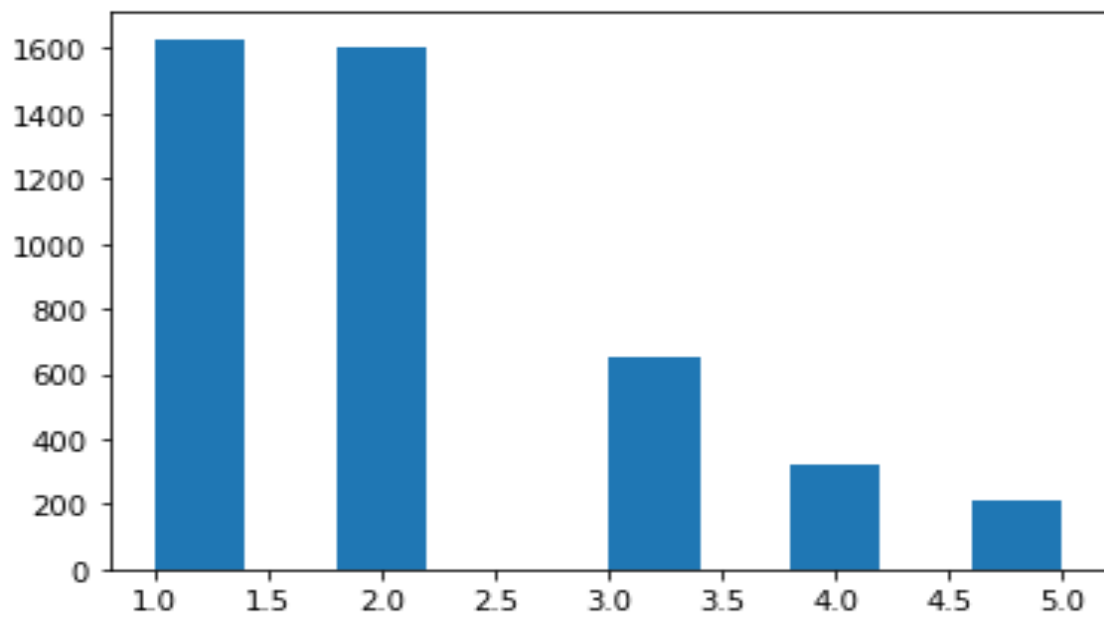
**Attrition:** About 700 employees left which is about 15.96%.



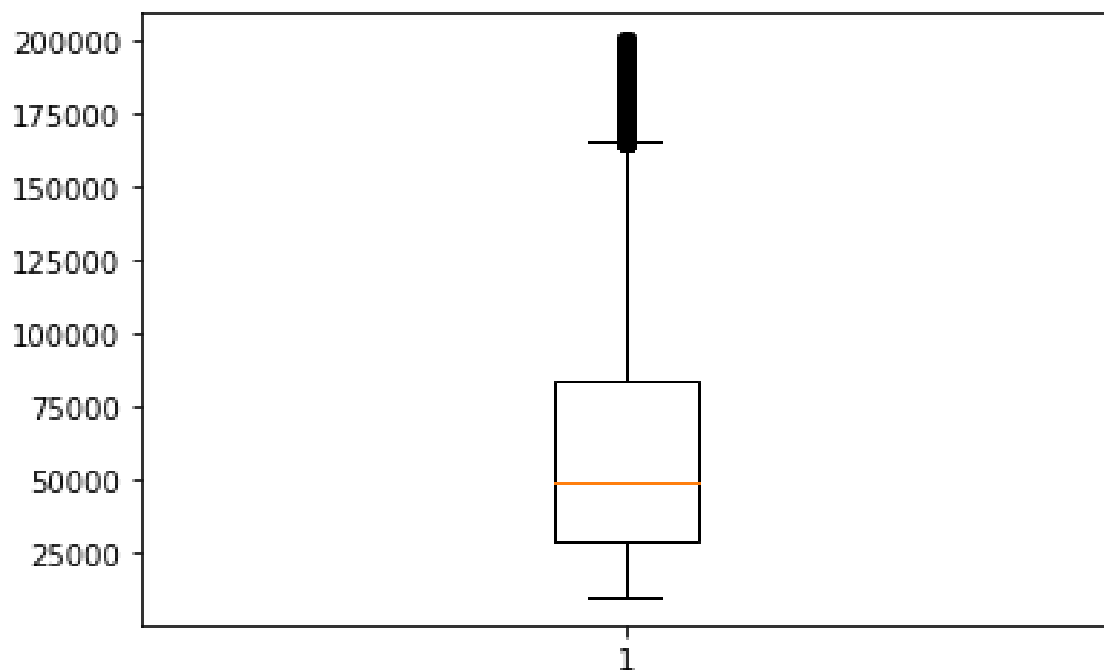
**Age:** Average age is about 36 (from the table). Moreover, the age is equally distributed.



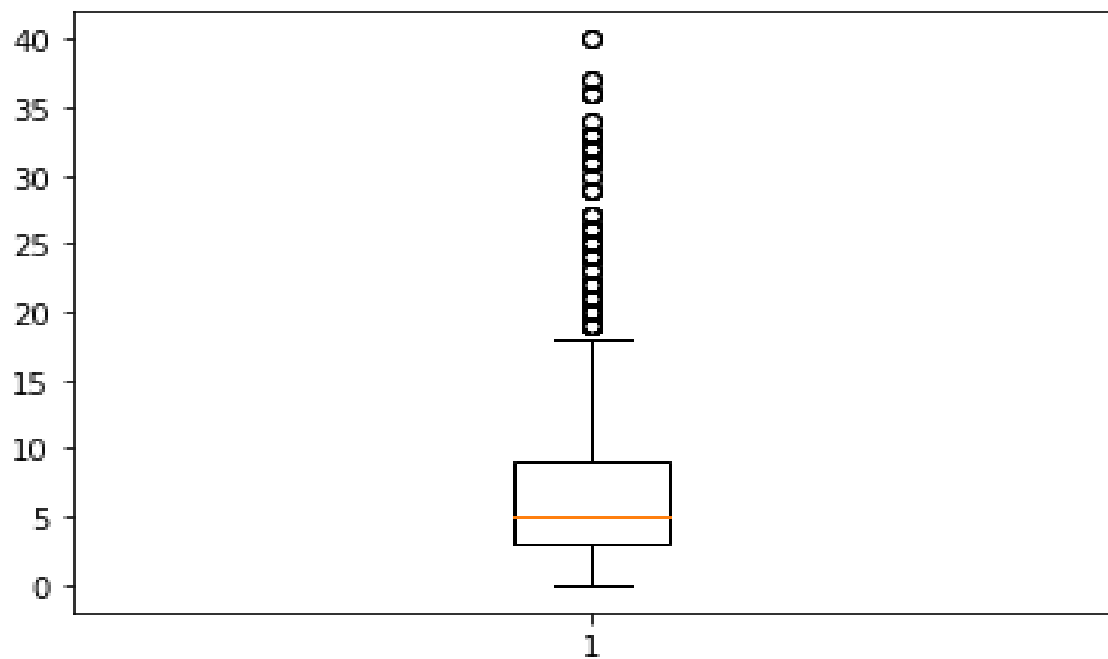
**Job-level:** The starting levels of job are highly densed.



**Monthly-Income:** We can see few outliers in this section.



**Years In Company:** The average time period is about 3 but there are few experienced employees too.





## Analysis for employee who left the company:

As seen the data frame “**dataset\_yes**” contains information of employees who left the job

The screenshot shows the Spyder Python IDE interface. In the center, a DataFrame viewer window titled "dataset\_yes - DataFrame" displays the following data:

| Index | Age | Attrition | BusinessTravel    | Department             | DistanceFromHome | Education | EducationField  | EmployeeCount |
|-------|-----|-----------|-------------------|------------------------|------------------|-----------|-----------------|---------------|
| 0     | 31  | Yes       | Travel_Frequently | Research & Development | 10               | 1         | Life Sciences   | 1             |
| 1     | 28  | Yes       | Travel_Rarely     | Research & Development | 11               | 2         | Medical         | 1             |
| 2     | 47  | Yes       | Non-Travel        | Research & Development | 1                | 1         | Medical         | 1             |
| 3     | 44  | Yes       | Travel_Frequently | Research & Development | 1                | 2         | Medical         | 1             |
| 4     | 26  | Yes       | Travel_Rarely     | Research & Development | 4                | 3         | Medical         | 1             |
| 5     | 26  | Yes       | Travel_Rarely     | Research & Development | 8                | 3         | Medical         | 1             |
| 6     | 18  | Yes       | Travel_Rarely     | Research & Development | 1                | 4         | Life Sciences   | 1             |
| 7     | 52  | Yes       | Travel_Rarely     | Research & Development | 7                | 1         | Life Sciences   | 1             |
| 8     | 28  | Yes       | Travel_Rarely     | Research & Development | 9                | 4         | Medical         | 1             |
| 9     | 39  | Yes       | Travel_Rarely     | Research & Development | 1                | 1         | Medical         | 1             |
| 10    | 29  | Yes       | Travel_Rarely     | Research & Development | 29               | 3         | Medical         | 1             |
| 11    | 21  | Yes       | Travel_Frequently | Research & Development | 9                | 3         | Medical         | 1             |
| 12    | 33  | Yes       | Travel_Rarely     | Human Resources        | 28               | 2         | Human Resources | 1             |

The general information :

|       | Index | Age     | DistanceFromHome | Education | JobLevel | MonthlyIncome | OverseasAssignments | PercentSalaryHike | WorkingTime | NumberOfTimesPromoted | PercentageLastYearSalaryIncrease | WithCurrentCompany |
|-------|-------|---------|------------------|-----------|----------|---------------|---------------------|-------------------|-------------|-----------------------|----------------------------------|--------------------|
| count |       | 705     | 705              | 705       | 705      | 705           | 705                 | 705               | 705         | 705                   | 705                              | 705                |
| mean  |       | 33.6284 | 9.02411          | 2.87234   | 2.03262  | 61815         | 2.93759             | 15.4879           | 8.27376     | 2.65816               | 5.14894                          | 1.96028            |
| std   |       | 9.67884 | 7.75518          | 1.01446   | 1.04871  | 44890.5       | 2.68128             | 3.78584           | 7.17676     | 1.1559                | 5.96097                          | 3.15753            |
| min   |       | 18      | 1                | 1         | 1        | 10090         | 0                   | 11                | 0           | 0                     | 0                                | 0                  |
| 25%   |       | 28      | 2                | 2         | 1        | 28440         | 1                   | 12                | 3           | 2                     | 1                                | 0                  |
| 50%   |       | 32      | 7                | 3         | 2        | 49080         | 1                   | 14                | 7           | 3                     | 3                                | 2                  |
| 75%   |       | 39      | 15               | 4         | 2        | 71040         | 5                   | 18                | 10          | 3                     | 7                                | 5                  |
| max   |       | 58      | 29               | 5         | 5        | 198590        | 9                   | 25                | 40          | 6                     | 40                               | 14                 |

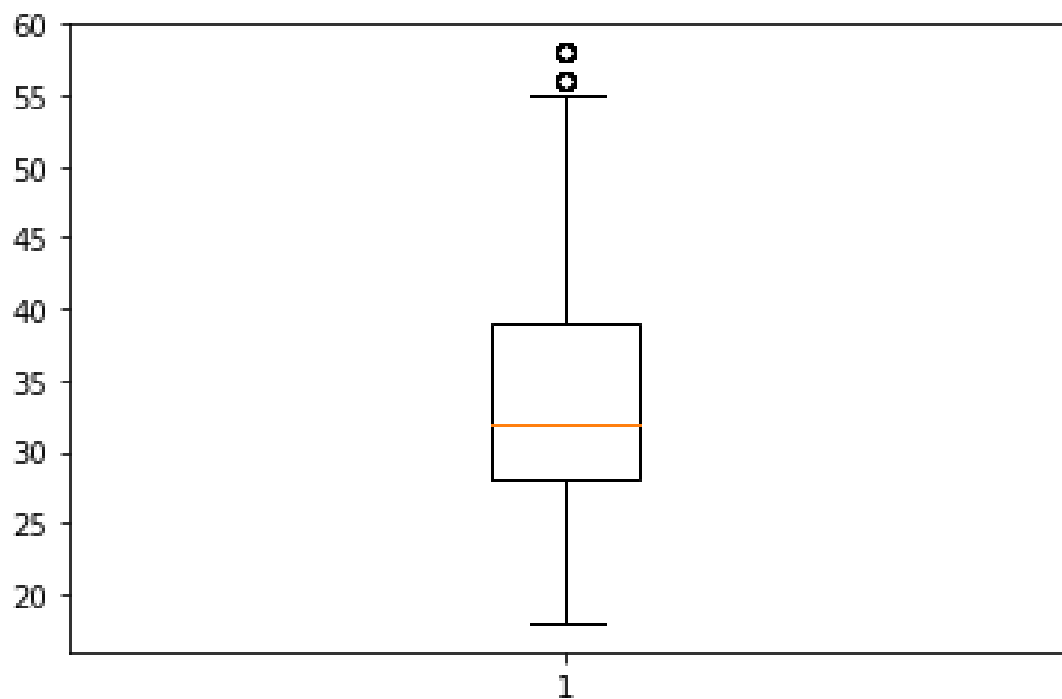
Mode:

| Index | Age | DistanceFromHome | Education | JobLevel | MonthlyIncome | NumCompaniesWorked | PercentSalaryHike | TotalWorkingYears | TrainingTimesLastYear | YearsAtCompany | YearsSinceLastPromotion | YearsWithCurrManager |
|-------|-----|------------------|-----------|----------|---------------|--------------------|-------------------|-------------------|-----------------------|----------------|-------------------------|----------------------|
| 0     | 31  | 2                | 3         | 2        | 25590         | 1                  | 13                | 1                 | 2                     | 1              | 0                       | 0                    |
| 1     | nan | nan              | nan       | nan      | 27410         | nan                | nan               | nan               | nan                   | nan            | nan                     | nan                  |
| 2     | nan | nan              | nan       | nan      | 27430         | nan                | nan               | nan               | nan                   | nan            | nan                     | nan                  |
| 3     | nan | nan              | nan       | nan      | 28860         | nan                | nan               | nan               | nan                   | nan            | nan                     | nan                  |
| 4     | nan | nan              | nan       | nan      | 55620         | nan                | nan               | nan               | nan                   | nan            | nan                     | nan                  |

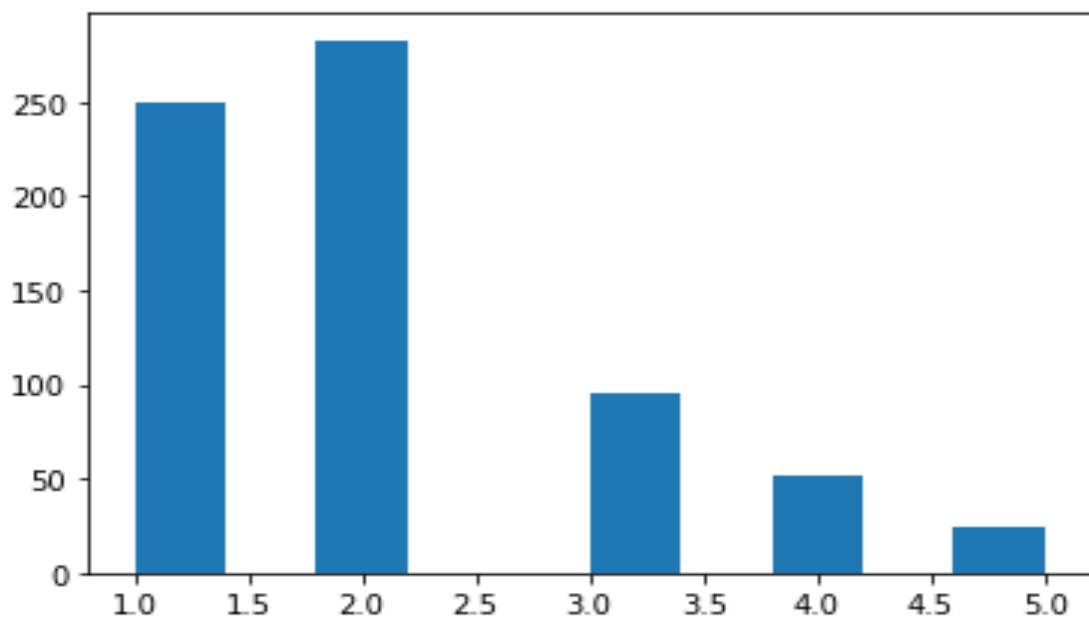
| dataset2_yes - Series   |       |
|-------------------------|-------|
| Index                   | 0     |
| Age                     | 32    |
| DistanceFromHome        | 7     |
| Education               | 3     |
| JobLevel                | 2     |
| MonthlyIncome           | 49080 |
| NumCompaniesWorked      | 1     |
| PercentSalaryHike       | 14    |
| TotalWorkingYears       | 7     |
| TrainingTimesLastYear   | 3     |
| YearsAtCompany          | 3     |
| YearsSinceLastPromotion | 1     |
| YearsWithCurrManager    | 2     |

Median: -----

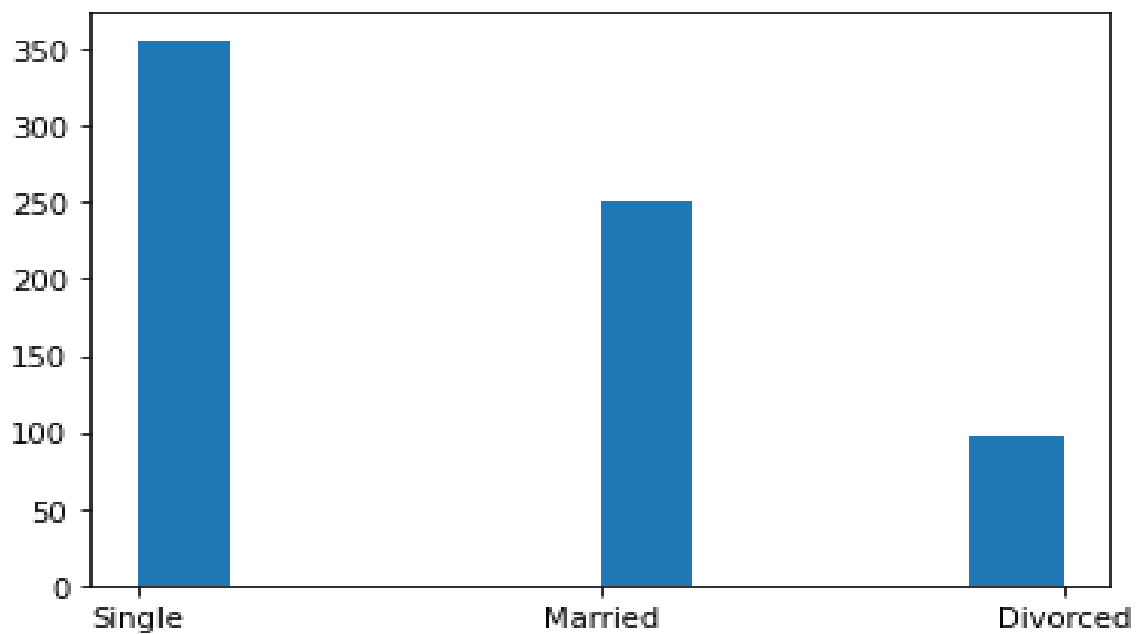
**Age:** Mainly people of 32-33 years leave to get new job and their are some outliers who probably leave because the are exhausted by their work.



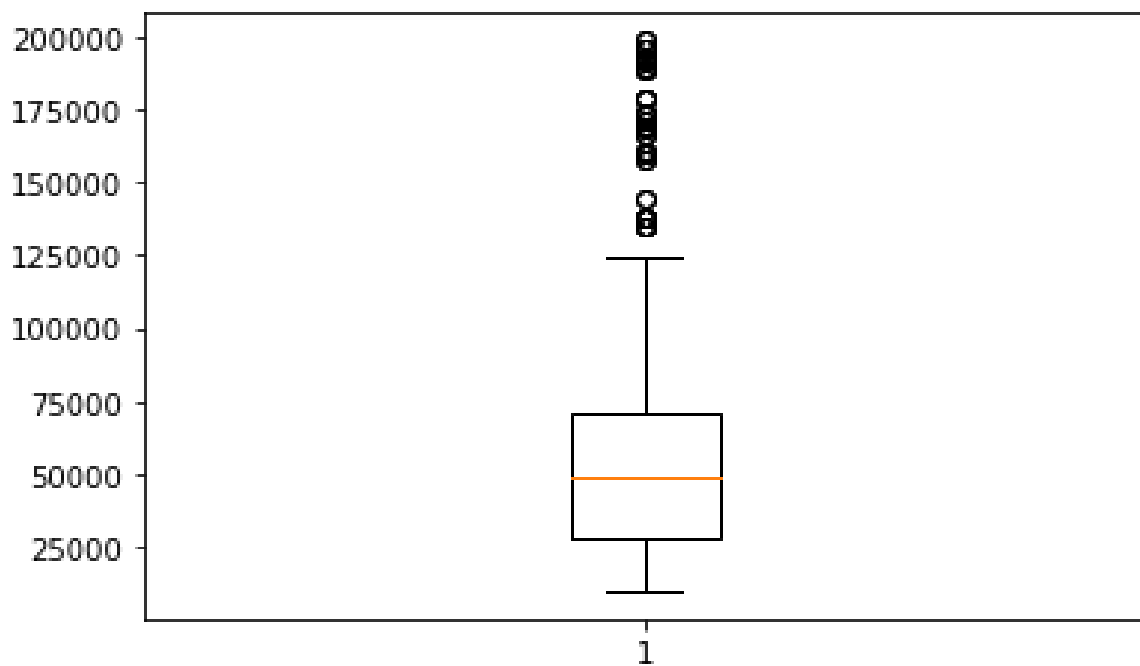
**Job Level-** Employees in starting of their carrier rush from on company to another to find a perfect Job profile (as they are not satisfied with the work they are doing)



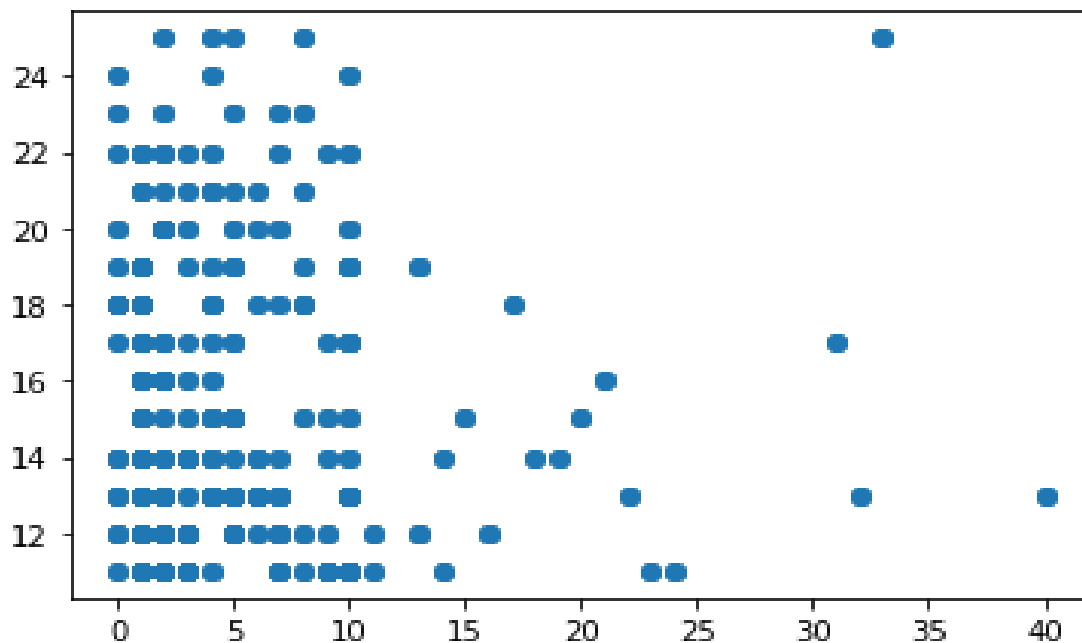
**Marital status:** Single employees are more ready to take risk and eager to transit.



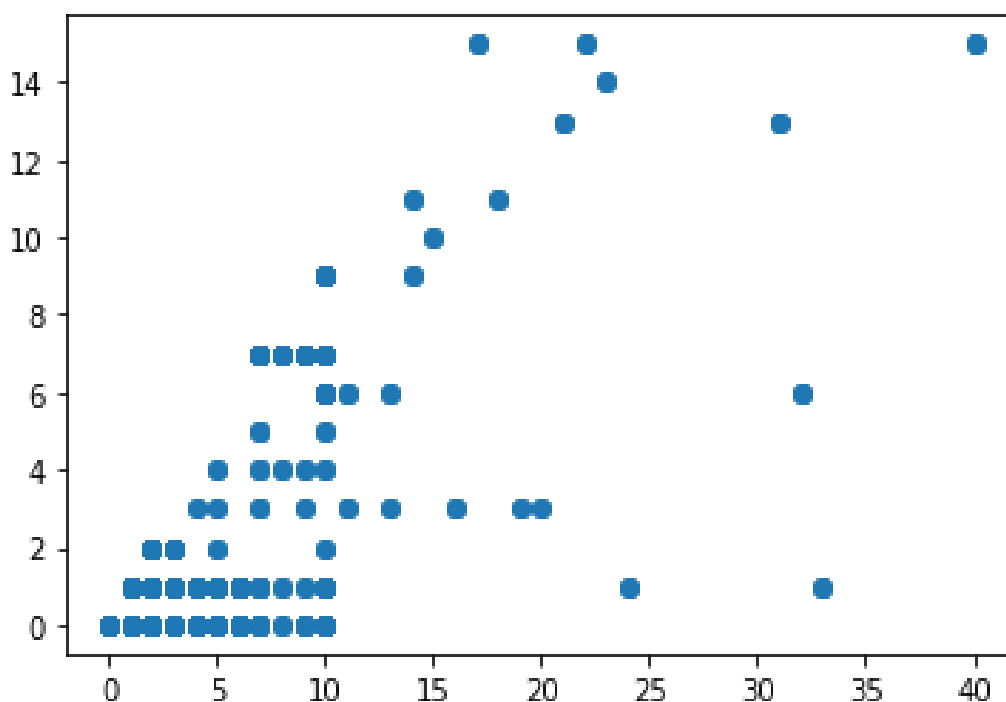
**Monthly income:** The attrition rate is evenly distributed between the employees regardless of their monthly income. And we can also observe that few high-paid employees have also left.



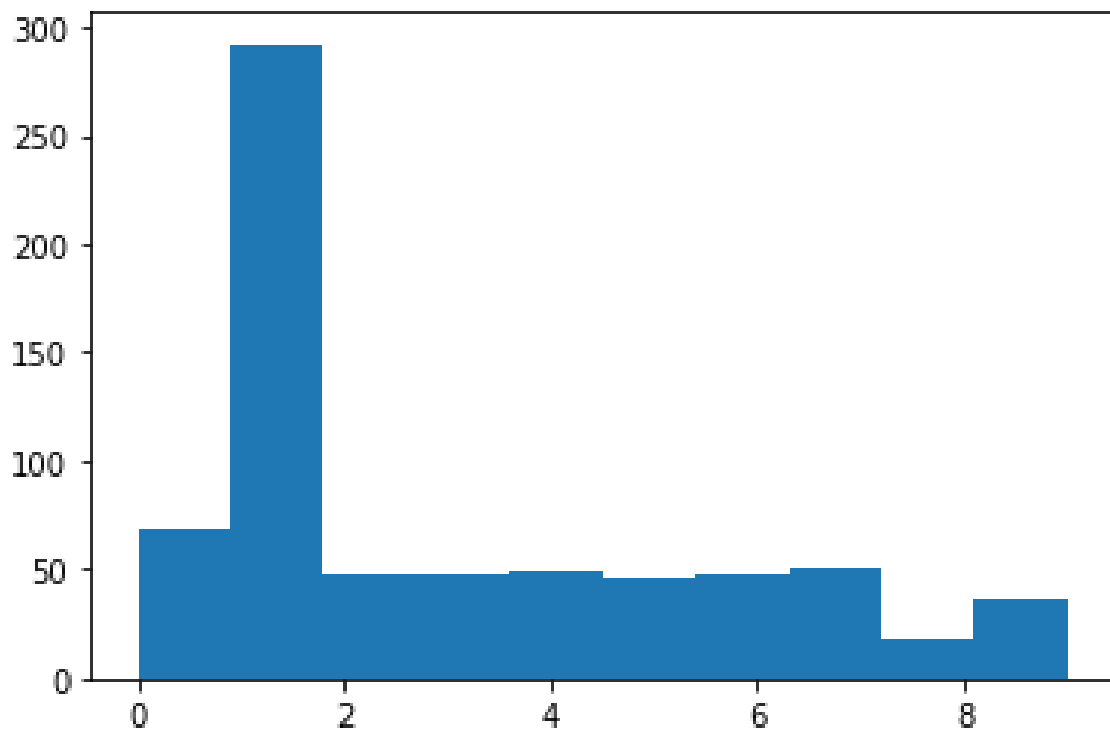
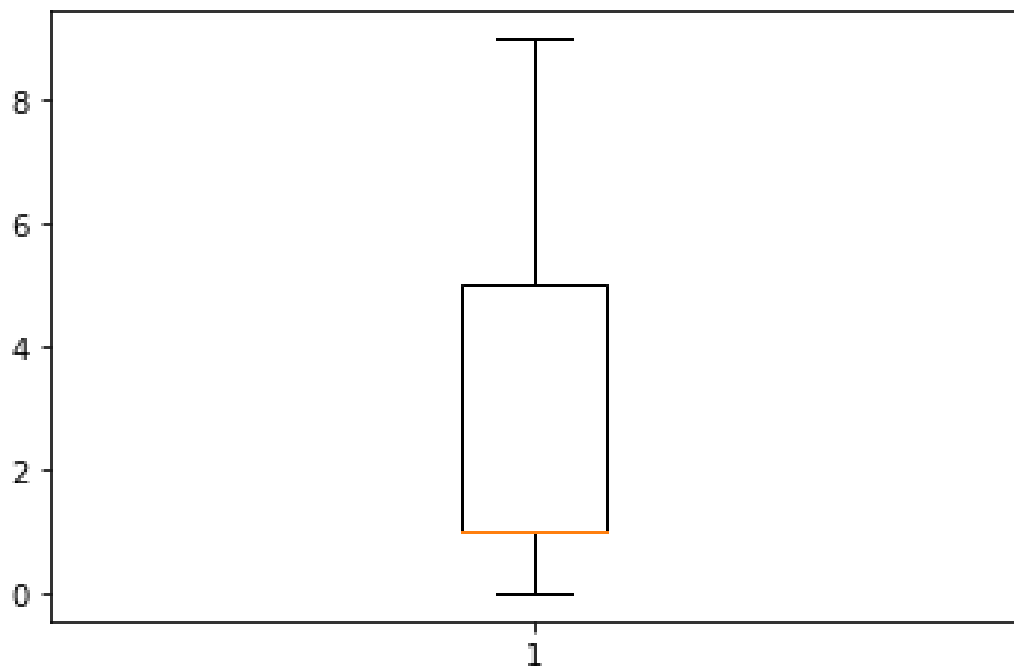
**Monthly-Hike vs Time at Company:** This scatter plot is significant in-order to know if one of the main reasons of attrition may be unhappiness of employees towards their salary hike. From this plot we observe that majorly employees who had spend about 0-10 years are more likely to leave .



**Time from last Promotion vs Time at Company:** The plot shows that recently promoted employees have been leaving the company.



**Companies worked:** By this plot we may observe that 1<sup>st</sup> quartile and median line intersect . It means that the employees who have worked for 1 company are most likely to leave in order to experiment or to get better opportunities.



## **Conclusion:**

We conclude that single employees who are in the age group of 32-22 having a mediocre job level(profile) and who have not worked in more companies are most likely to leave the company in order to gain opportunity and experience.