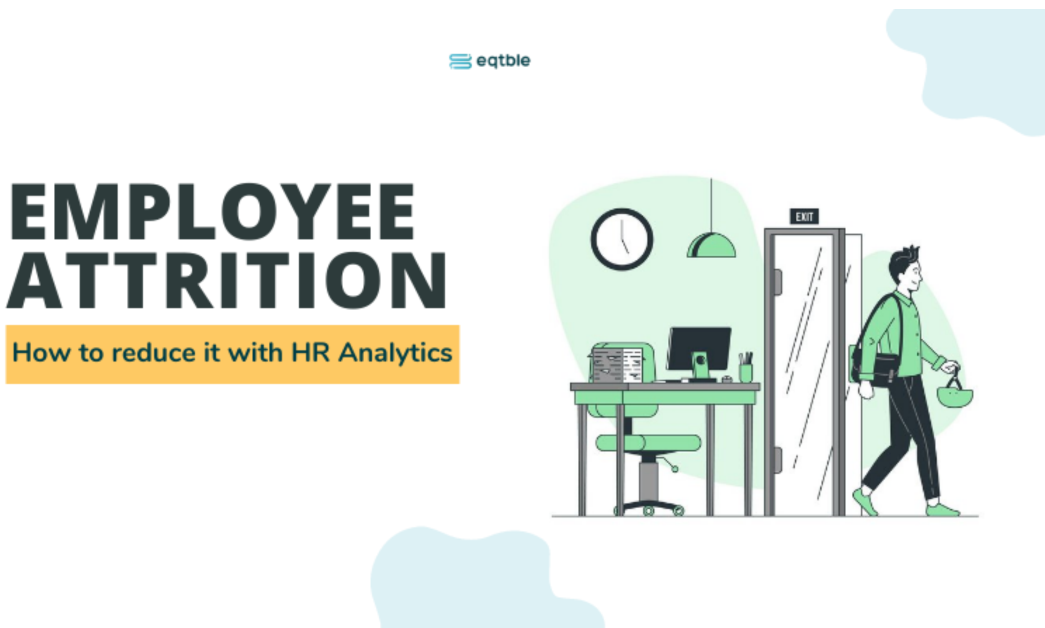
project-hr-employee-attrition-data

November 21, 2023

[1]:
```
pip install Pillow
```

```
Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: Pillow in c:\programdata\anaconda3\lib\site-
packages (9.4.0)
Note: you may need to restart the kernel to use updated packages.
```

[3]:
```
from IPython.display import Image, display

# Specify the path to your image file
image_path = r"C:\Users\jangi\Downloads\HR_Employee_Attrition_iamge.png"

# Display the image
display(Image(filename=image_path))
```



0.1 HR Employee Attrition Data Analysis

0.2 Objective

- The aim of this dataset is to build a model that can predict the attrition of the employees based on employee factors.

Import the Modules

0.2.1 EDA and Preprocessing

0.2.2 Machine Learning Algorithms

- Logistic Regression
- Random Forest
- Decision Tree

[71]:
```
import pandas as pd
import numpy as np


import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline

from sklearn.metrics import confusion_matrix, classification_report,␣
↪accuracy_score
from sklearn.model_selection import cross_val_score


import warnings
warnings.filterwarnings('ignore')
```

0.3 Read the CSV File

[4]:
```
df = pd.read_csv(r"C:\Users\jangi\Downloads\HR_Employee_Attrition_Data.csv")
```

[5]:
```
df.head()
```

[5]:
```
   Age Attrition     BusinessTravel  DailyRate              Department  \
0   41       Yes      Travel_Rarely       1102                   Sales
1   49        No  Travel_Frequently        279  Research & Development
2   37       Yes      Travel_Rarely       1373  Research & Development
3   33        No  Travel_Frequently       1392  Research & Development
4   27        No      Travel_Rarely        591  Research & Development

   DistanceFromHome  Education EducationField  EmployeeCount  EmployeeNumber  \
0                 1          2  Life Sciences              1               1
1                 8          1  Life Sciences              1               2
2                 2          2          Other              1               3
```

```
3                3          4  Life Sciences             1                4
4                2          1        Medical             1                5

   …  RelationshipSatisfaction StandardHours  StockOptionLevel  \
0  …                         1            80                 0
1  …                         4            80                 1
2  …                         2            80                 0
3  …                         3            80                 0
4  …                         4            80                 1

   TotalWorkingYears  TrainingTimesLastYear WorkLifeBalance  YearsAtCompany  \
0                  8                      0               1               6
1                 10                      3               3              10
2                  7                      3               3               0
3                  8                      3               3               8
4                  6                      3               3               2

  YearsInCurrentRole  YearsSinceLastPromotion  YearsWithCurrManager
0                  4                        0                     5
1                  7                        1                     7
2                  0                        0                     0
3                  7                        3                     0
4                  2                        2                     2

[5 rows x 35 columns]
```

0.4 Basic information

[4]:
```
df.shape
```

[4]:
```
(2940, 35)
```

[5]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2940 entries, 0 to 2939
Data columns (total 35 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Age                       2940 non-null   int64
 1   Attrition                 2940 non-null   object
 2   BusinessTravel            2940 non-null   object
 3   DailyRate                 2940 non-null   int64
 4   Department                2940 non-null   object
 5   DistanceFromHome          2940 non-null   int64
 6   Education                 2940 non-null   int64
 7   EducationField            2940 non-null   object
```

```
 8   EmployeeCount             2940 non-null   int64
 9   EmployeeNumber            2940 non-null   int64
 10  EnvironmentSatisfaction   2940 non-null   int64
 11  Gender                    2940 non-null   object
 12  HourlyRate                2940 non-null   int64
 13  JobInvolvement            2940 non-null   int64
 14  JobLevel                  2940 non-null   int64
 15  JobRole                   2940 non-null   object
 16  JobSatisfaction           2940 non-null   int64
 17  MaritalStatus             2940 non-null   object
 18  MonthlyIncome             2940 non-null   int64
 19  MonthlyRate               2940 non-null   int64
 20  NumCompaniesWorked        2940 non-null   int64
 21  Over18                    2940 non-null   object
 22  OverTime                  2940 non-null   object
 23  PercentSalaryHike         2940 non-null   int64
 24  PerformanceRating         2940 non-null   int64
 25  RelationshipSatisfaction  2940 non-null   int64
 26  StandardHours             2940 non-null   int64
 27  StockOptionLevel          2940 non-null   int64
 28  TotalWorkingYears         2940 non-null   int64
 29  TrainingTimesLastYear     2940 non-null   int64
 30  WorkLifeBalance           2940 non-null   int64
 31  YearsAtCompany            2940 non-null   int64
 32  YearsInCurrentRole        2940 non-null   int64
 33  YearsSinceLastPromotion   2940 non-null   int64
 34  YearsWithCurrManager      2940 non-null   int64
dtypes: int64(26), object(9)
memory usage: 804.0+ KB
```

[6]:
```
df.describe()
```

[6]:
```
               Age    DailyRate  DistanceFromHome    Education  EmployeeCount  \
count  2940.000000  2940.000000       2940.000000  2940.000000         2940.0
mean     36.923810   802.485714          9.192517     2.912925            1.0
std       9.133819   403.440447          8.105485     1.023991            0.0
min      18.000000   102.000000          1.000000     1.000000            1.0
25%      30.000000   465.000000          2.000000     2.000000            1.0
50%      36.000000   802.000000          7.000000     3.000000            1.0
75%      43.000000  1157.000000         14.000000     4.000000            1.0
max      60.000000  1499.000000         29.000000     5.000000            1.0

       EmployeeNumber  EnvironmentSatisfaction   HourlyRate  JobInvolvement  \
count     2940.000000              2940.000000  2940.000000     2940.000000
mean      1470.500000                 2.721769    65.891156        2.729932
std        848.849221                 1.092896    20.325969        0.711440
min          1.000000                 1.000000    30.000000        1.000000
```

```
25%        735.750000                  2.000000   48.000000        2.000000
50%       1470.500000                  3.000000   66.000000        3.000000
75%       2205.250000                  4.000000   84.000000        3.000000
max       2940.000000                  4.000000  100.000000        4.000000

          JobLevel  …  RelationshipSatisfaction  StandardHours  \
count  2940.000000  …               2940.000000         2940.0
mean      2.063946  …                  2.712245           80.0
std       1.106752  …                  1.081025            0.0
min       1.000000  …                  1.000000           80.0
25%       1.000000  …                  2.000000           80.0
50%       2.000000  …                  3.000000           80.0
75%       3.000000  …                  4.000000           80.0
max       5.000000  …                  4.000000           80.0

       StockOptionLevel  TotalWorkingYears  TrainingTimesLastYear  \
count       2940.000000        2940.000000            2940.000000
mean           0.793878          11.279592               2.799320
std            0.851932           7.779458               1.289051
min            0.000000           0.000000               0.000000
25%            0.000000           6.000000               2.000000
50%            1.000000          10.000000               3.000000
75%            1.000000          15.000000               3.000000
max            3.000000          40.000000               6.000000

       WorkLifeBalance  YearsAtCompany  YearsInCurrentRole  \
count      2940.000000     2940.000000         2940.000000
mean          2.761224        7.008163            4.229252
std           0.706356        6.125483            3.622521
min           1.000000        0.000000            0.000000
25%           2.000000        3.000000            2.000000
50%           3.000000        5.000000            3.000000
75%           3.000000        9.000000            7.000000
max           4.000000       40.000000           18.000000

       YearsSinceLastPromotion  YearsWithCurrManager
count              2940.000000           2940.000000
mean                  2.187755              4.123129
std                   3.221882              3.567529
min                   0.000000              0.000000
25%                   0.000000              2.000000
50%                   1.000000              3.000000
75%                   3.000000              7.000000
max                  15.000000             17.000000

[8 rows x 26 columns]
```

[7]:
```python
df.isna().sum()
```

[7]:
```
Age                         0
Attrition                   0
BusinessTravel              0
DailyRate                   0
Department                  0
DistanceFromHome            0
Education                   0
EducationField              0
EmployeeCount               0
EmployeeNumber              0
EnvironmentSatisfaction     0
Gender                      0
HourlyRate                  0
JobInvolvement              0
JobLevel                    0
JobRole                     0
JobSatisfaction             0
MaritalStatus               0
MonthlyIncome               0
MonthlyRate                 0
NumCompaniesWorked          0
Over18                      0
OverTime                    0
PercentSalaryHike           0
PerformanceRating           0
RelationshipSatisfaction    0
StandardHours               0
StockOptionLevel            0
TotalWorkingYears           0
TrainingTimesLastYear       0
WorkLifeBalance             0
YearsAtCompany              0
YearsInCurrentRole          0
YearsSinceLastPromotion     0
YearsWithCurrManager        0
dtype: int64
```

[8]:
```python
df.duplicated()
```

[8]:
```
0       False
1       False
2       False
3       False
4       False
        …
```

```
2935    False
2936    False
2937    False
2938    False
2939    False
Length: 2940, dtype: bool
```

0.5 EDA & PreProcessing

[263]:
```
df.describe().T.style.background_gradient(cmap = 'Reds')
```

[263]:
```
<pandas.io.formats.style.Styler at 0x24edc099ab0>
```

[10]:
```
pip install dataprep
```

```
Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: dataprep in
c:\users\jangi\appdata\roaming\python\python310\site-packages (0.4.5)
Requirement already satisfied: wordcloud<2.0,>=1.8 in
c:\users\jangi\appdata\roaming\python\python310\site-packages (from dataprep)
(1.9.2)
Requirement already satisfied: nltk<4.0.0,>=3.6.7 in
c:\programdata\anaconda3\lib\site-packages (from dataprep) (3.7)
Requirement already satisfied: rapidfuzz<3.0.0,>=2.1.2 in
c:\users\jangi\appdata\roaming\python\python310\site-packages (from dataprep)
(2.15.2)
Requirement already satisfied: numpy<2.0,>=1.21 in
c:\programdata\anaconda3\lib\site-packages (from dataprep) (1.23.5)
Requirement already satisfied: dask[array,dataframe,delayed]>=2022.3.0 in
c:\programdata\anaconda3\lib\site-packages (from dataprep) (2022.7.0)
Requirement already satisfied: jinja2<3.1,>=3.0 in
c:\users\jangi\appdata\roaming\python\python310\site-packages (from dataprep)
(3.0.3)
Requirement already satisfied: ipywidgets<8.0,>=7.5 in
c:\programdata\anaconda3\lib\site-packages (from dataprep) (7.6.5)
Requirement already satisfied: bokeh<3,>=2 in c:\programdata\anaconda3\lib\site-
packages (from dataprep) (2.4.3)
Requirement already satisfied: pydot<2.0.0,>=1.4.2 in
c:\users\jangi\appdata\roaming\python\python310\site-packages (from dataprep)
(1.4.2)
Requirement already satisfied: python-crfsuite==0.9.8 in
c:\users\jangi\appdata\roaming\python\python310\site-packages (from dataprep)
(0.9.8)
Requirement already satisfied: varname<0.9.0,>=0.8.1 in
c:\users\jangi\appdata\roaming\python\python310\site-packages (from dataprep)
(0.8.3)
Requirement already satisfied: tqdm<5.0,>=4.48 in
```

```
c:\programdata\anaconda3\lib\site-packages (from dataprep) (4.64.1)
Requirement already satisfied: sqlalchemy==1.3.24 in
c:\users\jangi\appdata\roaming\python\python310\site-packages (from dataprep)
(1.3.24)
Requirement already satisfied: python-stdnum<2.0,>=1.16 in
c:\users\jangi\appdata\roaming\python\python310\site-packages (from dataprep)
(1.19)
Requirement already satisfied: aiohttp<4.0,>=3.6 in
c:\users\jangi\appdata\roaming\python\python310\site-packages (from dataprep)
(3.8.6)
Requirement already satisfied: metaphone<0.7,>=0.6 in
c:\users\jangi\appdata\roaming\python\python310\site-packages (from dataprep)
(0.6)
Requirement already satisfied: pydantic<2.0,>=1.6 in
c:\users\jangi\appdata\roaming\python\python310\site-packages (from dataprep)
(1.10.13)
Requirement already satisfied: pandas<2.0,>=1.1 in
c:\programdata\anaconda3\lib\site-packages (from dataprep) (1.5.3)
Requirement already satisfied: jsonpath-ng<2.0,>=1.5 in
c:\users\jangi\appdata\roaming\python\python310\site-packages (from dataprep)
(1.6.0)
Requirement already satisfied: flask_cors<4.0.0,>=3.0.10 in
c:\users\jangi\appdata\roaming\python\python310\site-packages (from dataprep)
(3.0.10)
Requirement already satisfied: regex<2022.0.0,>=2021.8.3 in
c:\users\jangi\appdata\roaming\python\python310\site-packages (from dataprep)
(2021.11.10)
Requirement already satisfied: scipy<2.0,>=1.8 in
c:\programdata\anaconda3\lib\site-packages (from dataprep) (1.10.0)
Requirement already satisfied: flask<3,>=2 in c:\programdata\anaconda3\lib\site-
packages (from dataprep) (2.2.2)
Requirement already satisfied: async-timeout<5.0,>=4.0.0a3 in
c:\users\jangi\appdata\roaming\python\python310\site-packages (from
aiohttp<4.0,>=3.6->dataprep) (4.0.3)
Requirement already satisfied: multidict<7.0,>=4.5 in
c:\users\jangi\appdata\roaming\python\python310\site-packages (from
aiohttp<4.0,>=3.6->dataprep) (6.0.4)
Requirement already satisfied: aiosignal>=1.1.2 in
c:\users\jangi\appdata\roaming\python\python310\site-packages (from
aiohttp<4.0,>=3.6->dataprep) (1.3.1)
Requirement already satisfied: frozenlist>=1.1.1 in
c:\users\jangi\appdata\roaming\python\python310\site-packages (from
aiohttp<4.0,>=3.6->dataprep) (1.4.0)
Requirement already satisfied: attrs>=17.3.0 in
c:\programdata\anaconda3\lib\site-packages (from aiohttp<4.0,>=3.6->dataprep)
(22.1.0)
Requirement already satisfied: yarl<2.0,>=1.0 in
c:\users\jangi\appdata\roaming\python\python310\site-packages (from
```

```
aiohttp<4.0,>=3.6->dataprep) (1.9.2)
Requirement already satisfied: charset-normalizer<4.0,>=2.0 in
c:\programdata\anaconda3\lib\site-packages (from aiohttp<4.0,>=3.6->dataprep)
(2.0.4)
Requirement already satisfied: tornado>=5.1 in
c:\programdata\anaconda3\lib\site-packages (from bokeh<3,>=2->dataprep) (6.1)
Requirement already satisfied: pillow>=7.1.0 in
c:\programdata\anaconda3\lib\site-packages (from bokeh<3,>=2->dataprep) (9.4.0)
Requirement already satisfied: PyYAML>=3.10 in
c:\programdata\anaconda3\lib\site-packages (from bokeh<3,>=2->dataprep) (6.0)
Requirement already satisfied: typing-extensions>=3.10.0 in
c:\programdata\anaconda3\lib\site-packages (from bokeh<3,>=2->dataprep) (4.4.0)
Requirement already satisfied: packaging>=16.8 in
c:\programdata\anaconda3\lib\site-packages (from bokeh<3,>=2->dataprep) (22.0)
Requirement already satisfied: fsspec>=0.6.0 in
c:\programdata\anaconda3\lib\site-packages (from
dask[array,dataframe,delayed]>=2022.3.0->dataprep) (2022.11.0)
Requirement already satisfied: partd>=0.3.10 in
c:\programdata\anaconda3\lib\site-packages (from
dask[array,dataframe,delayed]>=2022.3.0->dataprep) (1.2.0)
Requirement already satisfied: cloudpickle>=1.1.1 in
c:\programdata\anaconda3\lib\site-packages (from
dask[array,dataframe,delayed]>=2022.3.0->dataprep) (2.0.0)
Requirement already satisfied: toolz>=0.8.2 in
c:\programdata\anaconda3\lib\site-packages (from
dask[array,dataframe,delayed]>=2022.3.0->dataprep) (0.12.0)
Requirement already satisfied: itsdangerous>=2.0 in
c:\programdata\anaconda3\lib\site-packages (from flask<3,>=2->dataprep) (2.0.1)
Requirement already satisfied: click>=8.0 in c:\programdata\anaconda3\lib\site-
packages (from flask<3,>=2->dataprep) (8.0.4)
Requirement already satisfied: Werkzeug>=2.2.2 in
c:\programdata\anaconda3\lib\site-packages (from flask<3,>=2->dataprep) (2.2.2)
Requirement already satisfied: Six in c:\programdata\anaconda3\lib\site-packages
(from flask_cors<4.0.0,>=3.0.10->dataprep) (1.16.0)
Requirement already satisfied: ipykernel>=4.5.1 in
c:\programdata\anaconda3\lib\site-packages (from ipywidgets<8.0,>=7.5->dataprep)
(6.19.2)
Requirement already satisfied: ipython-genutils~=0.2.0 in
c:\programdata\anaconda3\lib\site-packages (from ipywidgets<8.0,>=7.5->dataprep)
(0.2.0)
Requirement already satisfied: nbformat>=4.2.0 in
c:\programdata\anaconda3\lib\site-packages (from ipywidgets<8.0,>=7.5->dataprep)
(5.7.0)
Requirement already satisfied: ipython>=4.0.0 in
c:\programdata\anaconda3\lib\site-packages (from ipywidgets<8.0,>=7.5->dataprep)
(8.10.0)
Requirement already satisfied: widgetsnbextension~=3.5.0 in
c:\programdata\anaconda3\lib\site-packages (from ipywidgets<8.0,>=7.5->dataprep)
```

```
(3.5.2)
Requirement already satisfied: jupyterlab-widgets>=1.0.0 in
c:\programdata\anaconda3\lib\site-packages (from ipywidgets<8.0,>=7.5->dataprep)
(1.0.0)
Requirement already satisfied: traitlets>=4.3.1 in
c:\programdata\anaconda3\lib\site-packages (from ipywidgets<8.0,>=7.5->dataprep)
(5.7.1)
Requirement already satisfied: MarkupSafe>=2.0 in
c:\programdata\anaconda3\lib\site-packages (from jinja2<3.1,>=3.0->dataprep)
(2.1.1)
Requirement already satisfied: ply in c:\programdata\anaconda3\lib\site-packages
(from jsonpath-ng<2.0,>=1.5->dataprep) (3.11)
Requirement already satisfied: joblib in c:\programdata\anaconda3\lib\site-
packages (from nltk<4.0.0,>=3.6.7->dataprep) (1.1.1)
Requirement already satisfied: python-dateutil>=2.8.1 in
c:\programdata\anaconda3\lib\site-packages (from pandas<2.0,>=1.1->dataprep)
(2.8.2)
Requirement already satisfied: pytz>=2020.1 in
c:\programdata\anaconda3\lib\site-packages (from pandas<2.0,>=1.1->dataprep)
(2022.7)
Requirement already satisfied: pyparsing>=2.1.4 in
c:\programdata\anaconda3\lib\site-packages (from pydot<2.0.0,>=1.4.2->dataprep)
(3.0.9)
Requirement already satisfied: colorama in c:\programdata\anaconda3\lib\site-
packages (from tqdm<5.0,>=4.48->dataprep) (0.4.6)
Requirement already satisfied: pure_eval<1.0.0 in
c:\programdata\anaconda3\lib\site-packages (from
varname<0.9.0,>=0.8.1->dataprep) (0.2.2)
Requirement already satisfied: asttokens<3.0.0,>=2.0.0 in
c:\programdata\anaconda3\lib\site-packages (from
varname<0.9.0,>=0.8.1->dataprep) (2.0.5)
Requirement already satisfied: executing<0.9.0,>=0.8.3 in
c:\programdata\anaconda3\lib\site-packages (from
varname<0.9.0,>=0.8.1->dataprep) (0.8.3)
Requirement already satisfied: matplotlib in c:\programdata\anaconda3\lib\site-
packages (from wordcloud<2.0,>=1.8->dataprep) (3.7.0)
Requirement already satisfied: nest-asyncio in
c:\programdata\anaconda3\lib\site-packages (from
ipykernel>=4.5.1->ipywidgets<8.0,>=7.5->dataprep) (1.5.6)
Requirement already satisfied: pyzmq>=17 in c:\programdata\anaconda3\lib\site-
packages (from ipykernel>=4.5.1->ipywidgets<8.0,>=7.5->dataprep) (23.2.0)
Requirement already satisfied: comm>=0.1.1 in c:\programdata\anaconda3\lib\site-
packages (from ipykernel>=4.5.1->ipywidgets<8.0,>=7.5->dataprep) (0.1.2)
Requirement already satisfied: debugpy>=1.0 in
c:\programdata\anaconda3\lib\site-packages (from
ipykernel>=4.5.1->ipywidgets<8.0,>=7.5->dataprep) (1.5.1)
Requirement already satisfied: psutil in c:\programdata\anaconda3\lib\site-
packages (from ipykernel>=4.5.1->ipywidgets<8.0,>=7.5->dataprep) (5.9.0)
```

```
Requirement already satisfied: jupyter-client>=6.1.12 in
c:\programdata\anaconda3\lib\site-packages (from
ipykernel>=4.5.1->ipywidgets<8.0,>=7.5->dataprep) (7.3.4)
Requirement already satisfied: matplotlib-inline>=0.1 in
c:\programdata\anaconda3\lib\site-packages (from
ipykernel>=4.5.1->ipywidgets<8.0,>=7.5->dataprep) (0.1.6)
Requirement already satisfied: pickleshare in c:\programdata\anaconda3\lib\site-
packages (from ipython>=4.0.0->ipywidgets<8.0,>=7.5->dataprep) (0.7.5)
Requirement already satisfied: prompt-toolkit<3.1.0,>=3.0.30 in
c:\programdata\anaconda3\lib\site-packages (from
ipython>=4.0.0->ipywidgets<8.0,>=7.5->dataprep) (3.0.36)
Requirement already satisfied: jedi>=0.16 in c:\programdata\anaconda3\lib\site-
packages (from ipython>=4.0.0->ipywidgets<8.0,>=7.5->dataprep) (0.18.1)
Requirement already satisfied: pygments>=2.4.0 in
c:\programdata\anaconda3\lib\site-packages (from
ipython>=4.0.0->ipywidgets<8.0,>=7.5->dataprep) (2.11.2)
Requirement already satisfied: backcall in c:\programdata\anaconda3\lib\site-
packages (from ipython>=4.0.0->ipywidgets<8.0,>=7.5->dataprep) (0.2.0)
Requirement already satisfied: stack-data in c:\programdata\anaconda3\lib\site-
packages (from ipython>=4.0.0->ipywidgets<8.0,>=7.5->dataprep) (0.2.0)
Requirement already satisfied: decorator in c:\programdata\anaconda3\lib\site-
packages (from ipython>=4.0.0->ipywidgets<8.0,>=7.5->dataprep) (5.1.1)
Requirement already satisfied: fastjsonschema in
c:\programdata\anaconda3\lib\site-packages (from
nbformat>=4.2.0->ipywidgets<8.0,>=7.5->dataprep) (2.16.2)
Requirement already satisfied: jupyter-core in
c:\programdata\anaconda3\lib\site-packages (from
nbformat>=4.2.0->ipywidgets<8.0,>=7.5->dataprep) (5.2.0)
Requirement already satisfied: jsonschema>=2.6 in
c:\programdata\anaconda3\lib\site-packages (from
nbformat>=4.2.0->ipywidgets<8.0,>=7.5->dataprep) (4.17.3)
Requirement already satisfied: locket in c:\programdata\anaconda3\lib\site-
packages (from partd>=0.3.10->dask[array,dataframe,delayed]>=2022.3.0->dataprep)
(1.0.0)
Requirement already satisfied: notebook>=4.4.1 in
c:\programdata\anaconda3\lib\site-packages (from
widgetsnbextension~=3.5.0->ipywidgets<8.0,>=7.5->dataprep) (6.5.2)
Requirement already satisfied: idna>=2.0 in c:\programdata\anaconda3\lib\site-
packages (from yarl<2.0,>=1.0->aiohttp<4.0,>=3.6->dataprep) (3.4)
Requirement already satisfied: cycler>=0.10 in
c:\programdata\anaconda3\lib\site-packages (from
matplotlib->wordcloud<2.0,>=1.8->dataprep) (0.11.0)
Requirement already satisfied: fonttools>=4.22.0 in
c:\programdata\anaconda3\lib\site-packages (from
matplotlib->wordcloud<2.0,>=1.8->dataprep) (4.25.0)
Requirement already satisfied: kiwisolver>=1.0.1 in
c:\programdata\anaconda3\lib\site-packages (from
matplotlib->wordcloud<2.0,>=1.8->dataprep) (1.4.4)
```

```
Requirement already satisfied: contourpy>=1.0.1 in
c:\programdata\anaconda3\lib\site-packages (from
matplotlib->wordcloud<2.0,>=1.8->dataprep) (1.0.5)
Requirement already satisfied: parso<0.9.0,>=0.8.0 in
c:\programdata\anaconda3\lib\site-packages (from
jedi>=0.16->ipython>=4.0.0->ipywidgets<8.0,>=7.5->dataprep) (0.8.3)
Requirement already satisfied: pyrsistent!=0.17.0,!=0.17.1,!=0.17.2,>=0.14.0 in
c:\programdata\anaconda3\lib\site-packages (from
jsonschema>=2.6->nbformat>=4.2.0->ipywidgets<8.0,>=7.5->dataprep) (0.18.0)
Requirement already satisfied: entrypoints in c:\programdata\anaconda3\lib\site-
packages (from jupyter-
client>=6.1.12->ipykernel>=4.5.1->ipywidgets<8.0,>=7.5->dataprep) (0.4)
Requirement already satisfied: platformdirs>=2.5 in
c:\programdata\anaconda3\lib\site-packages (from jupyter-
core->nbformat>=4.2.0->ipywidgets<8.0,>=7.5->dataprep) (2.5.2)
Requirement already satisfied: pywin32>=1.0 in
c:\programdata\anaconda3\lib\site-packages (from jupyter-
core->nbformat>=4.2.0->ipywidgets<8.0,>=7.5->dataprep) (305.1)
Requirement already satisfied: Send2Trash>=1.8.0 in
c:\programdata\anaconda3\lib\site-packages (from
notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets<8.0,>=7.5->dataprep)
(1.8.0)
Requirement already satisfied: nbclassic>=0.4.7 in
c:\programdata\anaconda3\lib\site-packages (from
notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets<8.0,>=7.5->dataprep)
(0.5.2)
Requirement already satisfied: terminado>=0.8.3 in
c:\programdata\anaconda3\lib\site-packages (from
notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets<8.0,>=7.5->dataprep)
(0.17.1)
Requirement already satisfied: nbconvert>=5 in
c:\programdata\anaconda3\lib\site-packages (from
notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets<8.0,>=7.5->dataprep)
(6.5.4)
Requirement already satisfied: argon2-cffi in c:\programdata\anaconda3\lib\site-
packages (from
notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets<8.0,>=7.5->dataprep)
(21.3.0)
Requirement already satisfied: prometheus-client in
c:\programdata\anaconda3\lib\site-packages (from
notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets<8.0,>=7.5->dataprep)
(0.14.1)
Requirement already satisfied: wcwidth in c:\programdata\anaconda3\lib\site-
packages (from prompt-
toolkit<3.1.0,>=3.0.30->ipython>=4.0.0->ipywidgets<8.0,>=7.5->dataprep) (0.2.5)
Requirement already satisfied: notebook-shim>=0.1.0 in
c:\programdata\anaconda3\lib\site-packages (from nbclassic>=0.4.7->notebook>=4.4
.1->widgetsnbextension~=3.5.0->ipywidgets<8.0,>=7.5->dataprep) (0.2.2)
```

```
Requirement already satisfied: jupyter-server>=1.8 in
c:\programdata\anaconda3\lib\site-packages (from nbclassic>=0.4.7->notebook>=4.4
.1->widgetsnbextension~=3.5.0->ipywidgets<8.0,>=7.5->dataprep) (1.23.4)
Requirement already satisfied: bleach in c:\programdata\anaconda3\lib\site-
packages (from nbconvert>=5->notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidg
ets<8.0,>=7.5->dataprep) (4.1.0)
Requirement already satisfied: lxml in c:\programdata\anaconda3\lib\site-
packages (from nbconvert>=5->notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidg
ets<8.0,>=7.5->dataprep) (4.9.1)
Requirement already satisfied: jupyterlab-pygments in
c:\programdata\anaconda3\lib\site-packages (from nbconvert>=5->notebook>=4.4.1->
widgetsnbextension~=3.5.0->ipywidgets<8.0,>=7.5->dataprep) (0.1.2)
Requirement already satisfied: defusedxml in c:\programdata\anaconda3\lib\site-
packages (from nbconvert>=5->notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidg
ets<8.0,>=7.5->dataprep) (0.7.1)
Requirement already satisfied: beautifulsoup4 in
c:\programdata\anaconda3\lib\site-packages (from nbconvert>=5->notebook>=4.4.1->
widgetsnbextension~=3.5.0->ipywidgets<8.0,>=7.5->dataprep) (4.11.1)
Requirement already satisfied: nbclient>=0.5.0 in
c:\programdata\anaconda3\lib\site-packages (from nbconvert>=5->notebook>=4.4.1->
widgetsnbextension~=3.5.0->ipywidgets<8.0,>=7.5->dataprep) (0.5.13)
Requirement already satisfied: tinycss2 in c:\programdata\anaconda3\lib\site-
packages (from nbconvert>=5->notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidg
ets<8.0,>=7.5->dataprep) (1.2.1)
Requirement already satisfied: mistune<2,>=0.8.1 in
c:\programdata\anaconda3\lib\site-packages (from nbconvert>=5->notebook>=4.4.1->
widgetsnbextension~=3.5.0->ipywidgets<8.0,>=7.5->dataprep) (0.8.4)
Requirement already satisfied: pandocfilters>=1.4.1 in
c:\programdata\anaconda3\lib\site-packages (from nbconvert>=5->notebook>=4.4.1->
widgetsnbextension~=3.5.0->ipywidgets<8.0,>=7.5->dataprep) (1.5.0)
Requirement already satisfied: pywinpty>=1.1.0 in
c:\programdata\anaconda3\lib\site-packages (from terminado>=0.8.3->notebook>=4.4
.1->widgetsnbextension~=3.5.0->ipywidgets<8.0,>=7.5->dataprep) (2.0.10)
Requirement already satisfied: argon2-cffi-bindings in
c:\programdata\anaconda3\lib\site-packages (from argon2-cffi->notebook>=4.4.1->w
idgetsnbextension~=3.5.0->ipywidgets<8.0,>=7.5->dataprep) (21.2.0)
Requirement already satisfied: websocket-client in
c:\programdata\anaconda3\lib\site-packages (from jupyter-server>=1.8->nbclassic>
=0.4.7->notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets<8.0,>=7.5->datapr
ep) (0.58.0)
Requirement already satisfied: anyio<4,>=3.1.0 in
c:\programdata\anaconda3\lib\site-packages (from jupyter-server>=1.8->nbclassic>
=0.4.7->notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets<8.0,>=7.5->datapr
ep) (3.5.0)
Requirement already satisfied: cffi>=1.0.1 in c:\programdata\anaconda3\lib\site-
packages (from argon2-cffi-bindings->argon2-cffi->notebook>=4.4.1->widgetsnbexte
nsion~=3.5.0->ipywidgets<8.0,>=7.5->dataprep) (1.15.1)
Requirement already satisfied: soupsieve>1.2 in
```

```
c:\programdata\anaconda3\lib\site-packages (from beautifulsoup4->nbconvert>=5->n
otebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets<8.0,>=7.5->dataprep)
(2.3.2.post1)
Requirement already satisfied: webencodings in
c:\programdata\anaconda3\lib\site-packages (from bleach->nbconvert>=5->notebook>
=4.4.1->widgetsnbextension~=3.5.0->ipywidgets<8.0,>=7.5->dataprep) (0.5.1)
Requirement already satisfied: sniffio>=1.1 in
c:\programdata\anaconda3\lib\site-packages (from anyio<4,>=3.1.0->jupyter-server
>=1.8->nbclassic>=0.4.7->notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets<
8.0,>=7.5->dataprep) (1.2.0)
Requirement already satisfied: pycparser in c:\programdata\anaconda3\lib\site-
packages (from cffi>=1.0.1->argon2-cffi-bindings->argon2-cffi->notebook>=4.4.1->
widgetsnbextension~=3.5.0->ipywidgets<8.0,>=7.5->dataprep) (2.21)
Note: you may need to restart the kernel to use updated packages.
```

[264]:
```
from dataprep.eda import create_report
report = create_report(df, title= 'Data Report')
report
```

```
  0%|          | 0/4747 [00:00<?, ?it/s]
```

[264]:

[76]:
```
df.columns
```

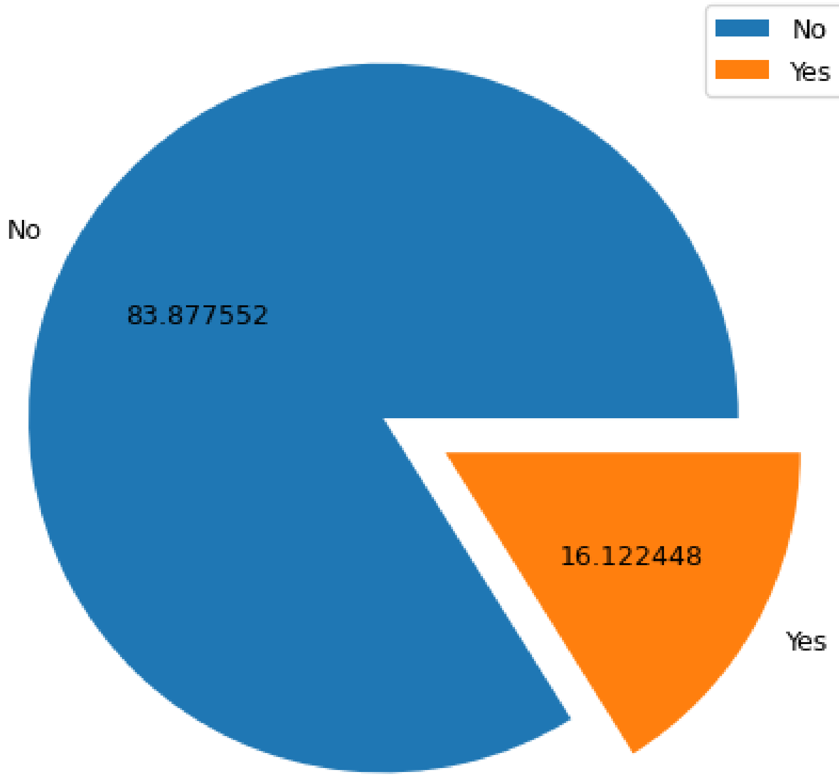
[76]:
```
Index(['Age', 'Attrition', 'BusinessTravel', 'DailyRate', 'Department',
       'DistanceFromHome', 'Education', 'EducationField', 'EmployeeCount',
       'EmployeeNumber', 'EnvironmentSatisfaction', 'Gender', 'HourlyRate',
       'JobInvolvement', 'JobLevel', 'JobRole', 'JobSatisfaction',
       'MaritalStatus', 'MonthlyIncome', 'MonthlyRate', 'NumCompaniesWorked',
       'Over18', 'OverTime', 'PercentSalaryHike', 'PerformanceRating',
       'RelationshipSatisfaction', 'StandardHours', 'StockOptionLevel',
       'TotalWorkingYears', 'TrainingTimesLastYear', 'WorkLifeBalance',
       'YearsAtCompany', 'YearsInCurrentRole', 'YearsSinceLastPromotion',
       'YearsWithCurrManager'],
      dtype='object')
```

[79]:
```
categorical_col = df.select_dtypes(include = ['object']).columns
numerical_col = df.select_dtypes(exclude = ['object']).columns
```

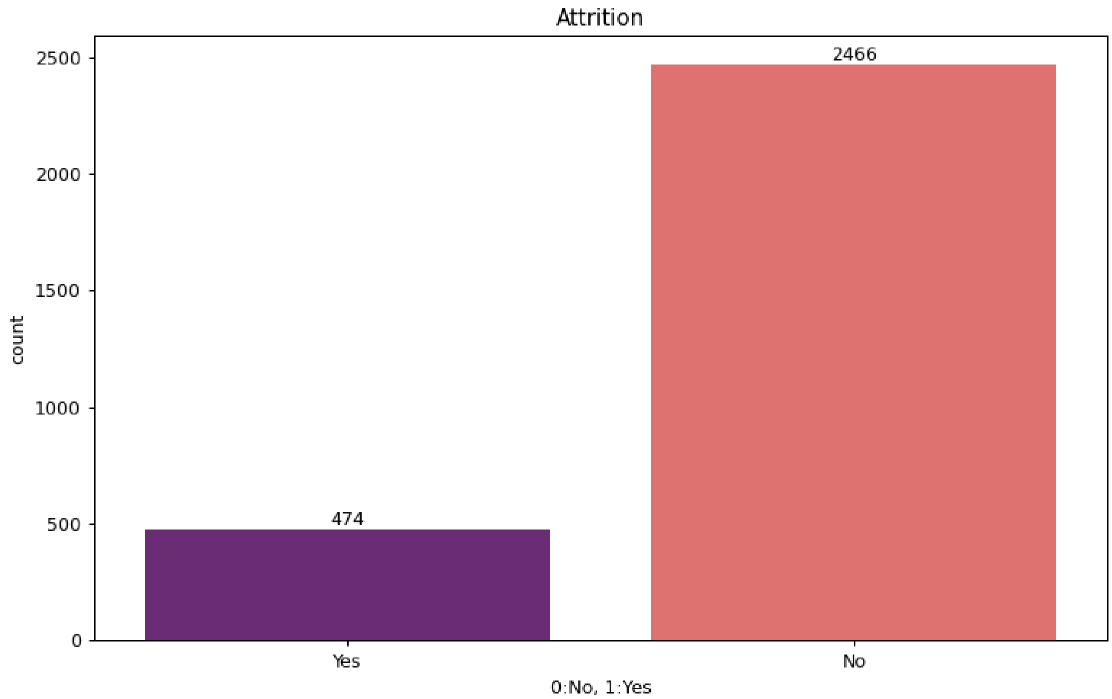
[80]:
```
df['Gender'].replace(['F'],'Female', inplace = True)
df['MaritalStatus'].replace(['M'],'Married', inplace = True)
```

[81]:
```
plt.figure(figsize = (10,6), dpi = 90)
plt.pie(df['Attrition'].value_counts(), labels = df['Attrition'].value_counts().
 ↪index,
        autopct = "%2f", explode = (0.1,0.1))
```

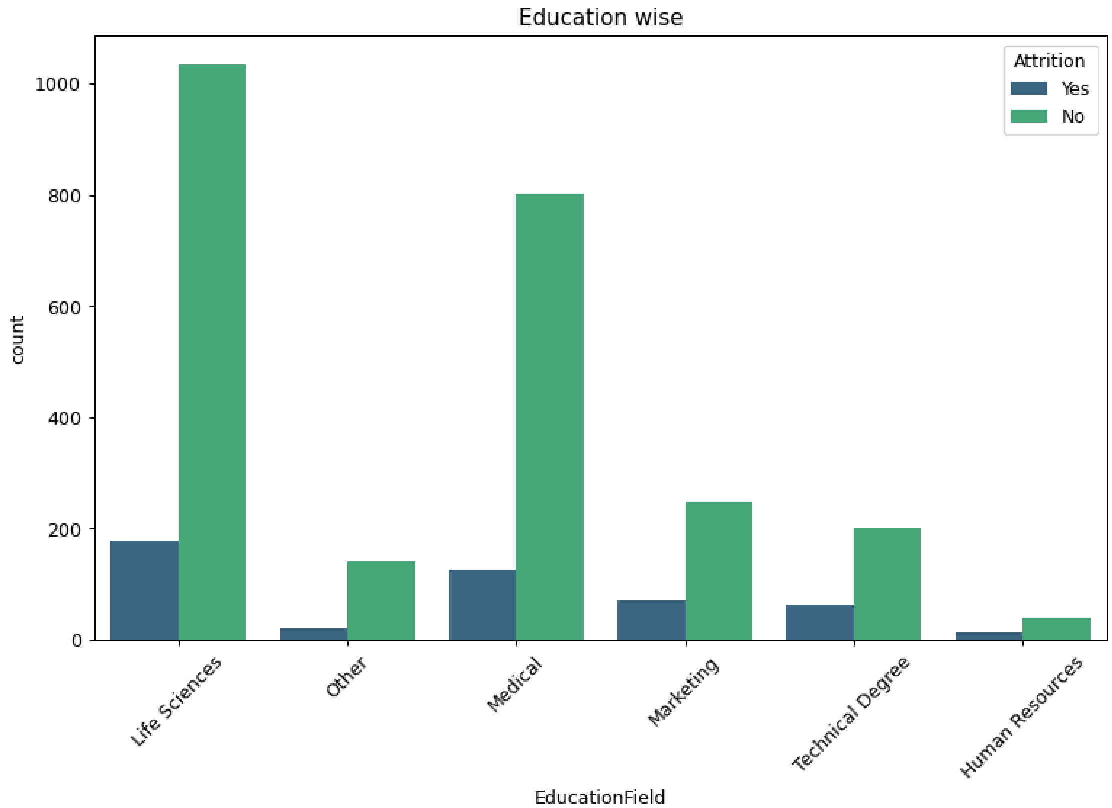
```
plt.legend()
plt.show()
```



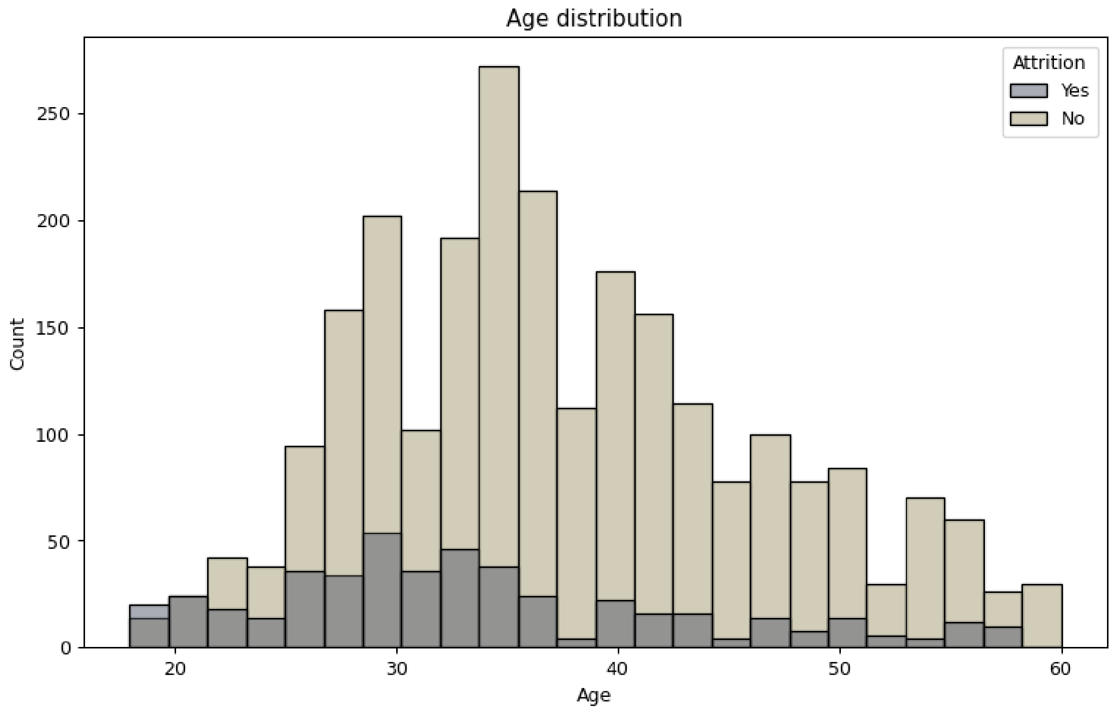
[14]:
```
plt.figure(figsize = (10,6), dpi = 90)
ax = sns.countplot(x = "Attrition",data = df , palette='magma')
plt.title('Attrition')
plt.xlabel('0:No, 1:Yes')
for i in ax.containers:
    ax.bar_label(i)
```



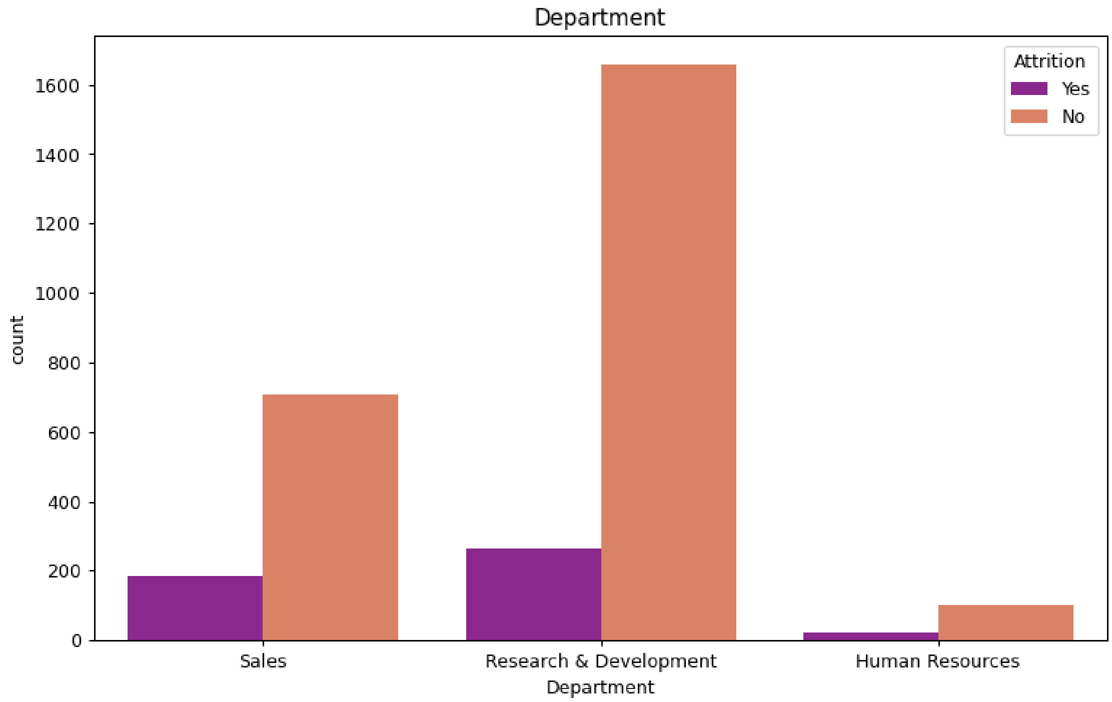
[15]:
```
plt.figure(figsize = (10,6), dpi = 90)
sns.countplot(x = "EducationField",hue='Attrition', data = df,␣
    ↪palette='viridis')
plt.title('Education wise')
plt.xticks(rotation=45)
plt.show()
```



[16]:
```
plt.figure(figsize = (10,6), dpi = 90)
sns.histplot(x = "Age", hue='Attrition', data = df, palette='cividis')
plt.title('Age distribution')
plt.show()
```

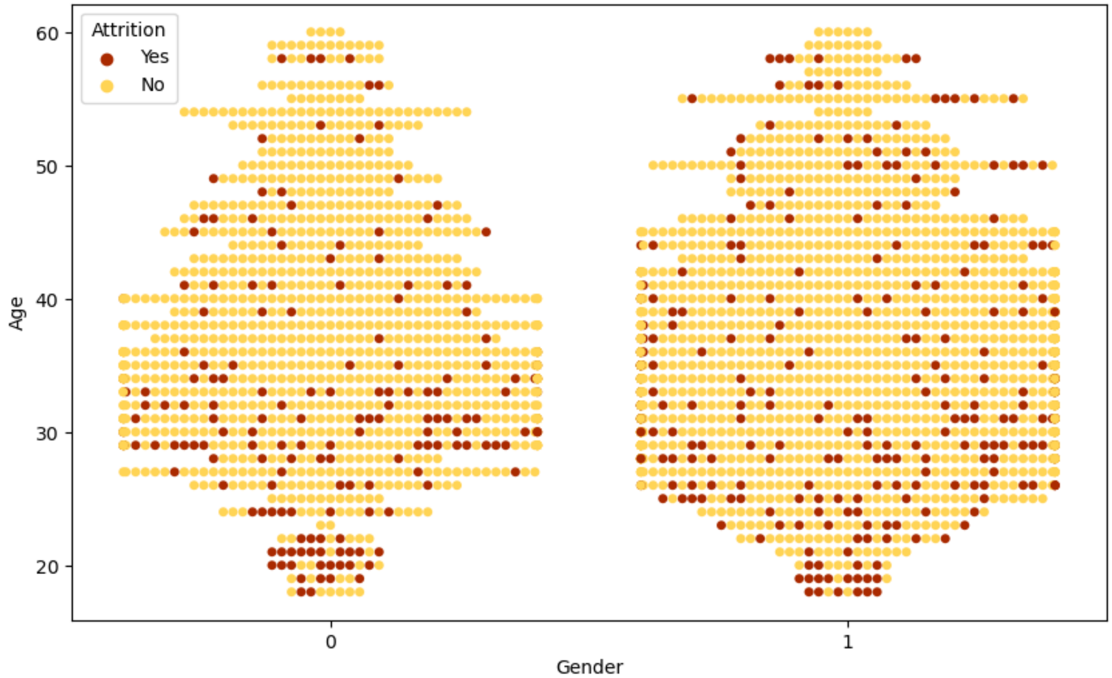



```
[17]: plt.figure(figsize = (10,6), dpi = 90)
      ax = sns.countplot(x = "Department", hue='Attrition', data = df,␣
       ↪palette='plasma')
      plt.title('Department')
      plt.show()
```



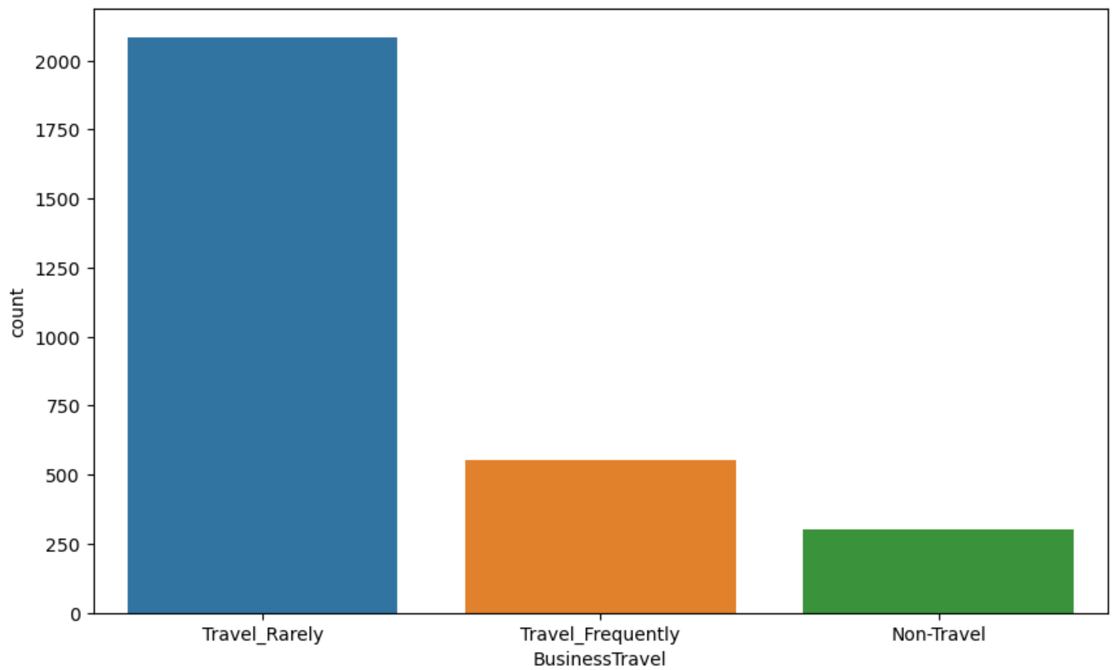
[265]:
```
plt.figure(figsize = (10,6), dpi = 100)
ax = sns.swarmplot(x = "Gender", y = "Age", hue = "Attrition" ,data = df,␣
 ↪palette = "afmhot")
plt.xlabel("Gender")
plt.ylabel("Age")

plt.show()
```

```
Using categorical units to plot a list of strings that are all parsable as
floats or dates. If these strings should be plotted as numbers, cast to the
appropriate data type before plotting.
Using categorical units to plot a list of strings that are all parsable as
floats or dates. If these strings should be plotted as numbers, cast to the
appropriate data type before plotting.
```

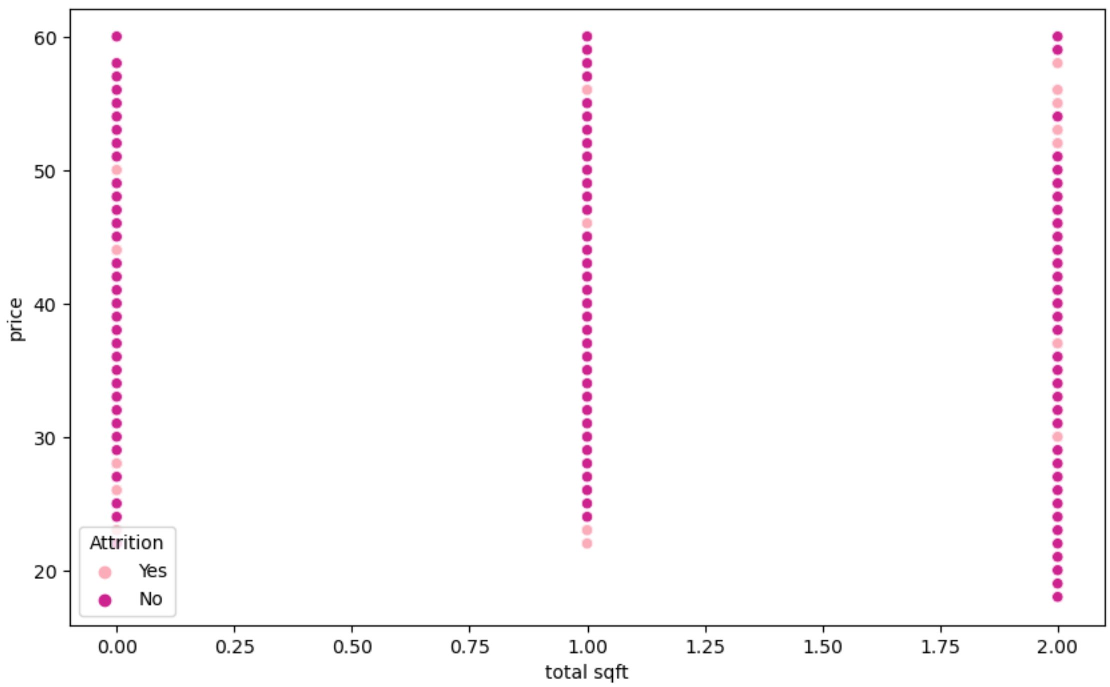


[266]:
```
plt.figure(figsize = (10,6))
print(df['BusinessTravel'].value_counts())
sns.countplot(x = "BusinessTravel", data = df)
plt.show()
```

```
Travel_Rarely        2086
Travel_Frequently     554
Non-Travel            300
Name: BusinessTravel, dtype: int64
```

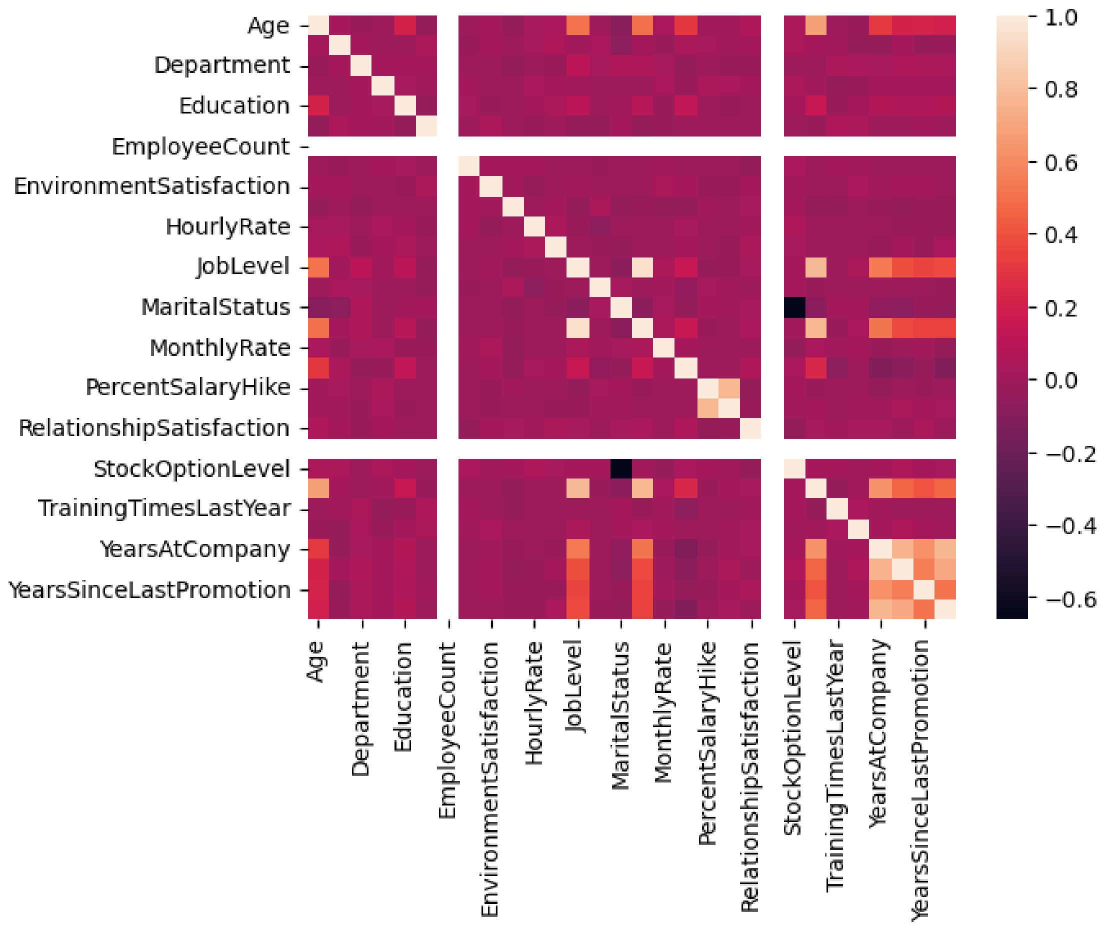


[267]:
```
plt.figure(figsize = (10,6), dpi = 100)
ax = sns.scatterplot(x = "MaritalStatus", y = "Age", hue = "Attrition" ,data =␣
 ↪df , palette='RdPu')
plt.xlabel("total sqft")
plt.ylabel("price")

plt.show()
```



[268]:
```
sns.heatmap(df.corr())
```

[268]: <Axes: >



0.6 Check the Outliers

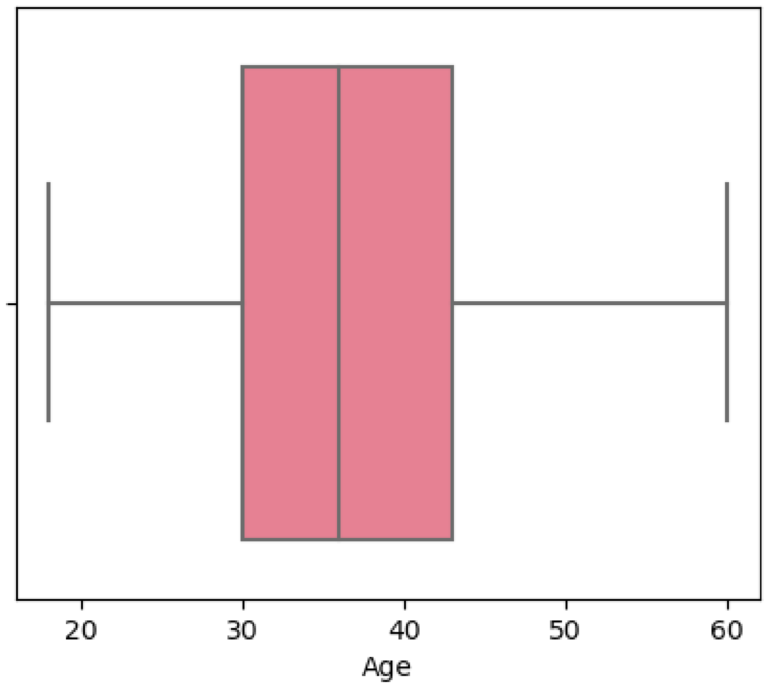
- A few data points that are significently different from the rest of the data points
- if any data points is far from the mean values can be treated as outliers
- outliers are only checked for numerical values
- to detect outliers we used boxplots

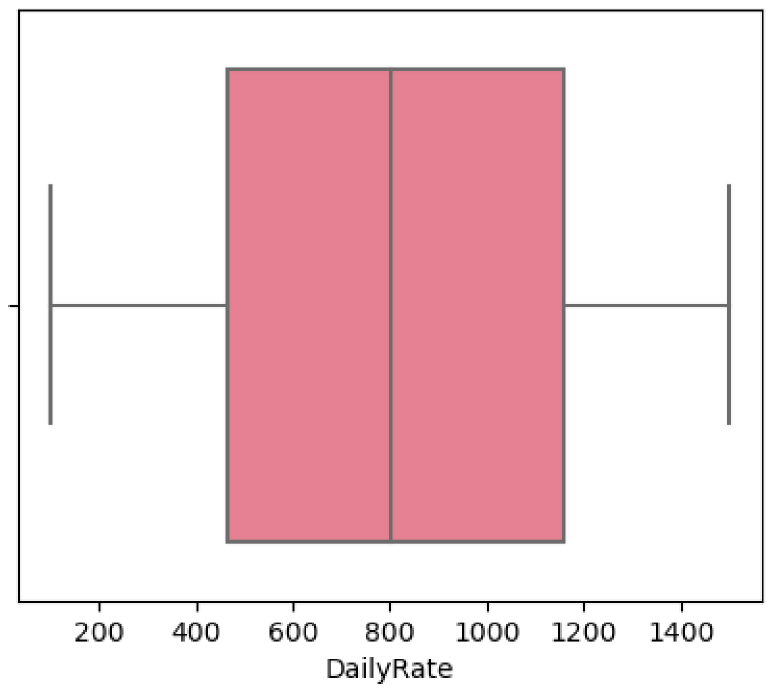
[72]:
```
df.columns
```

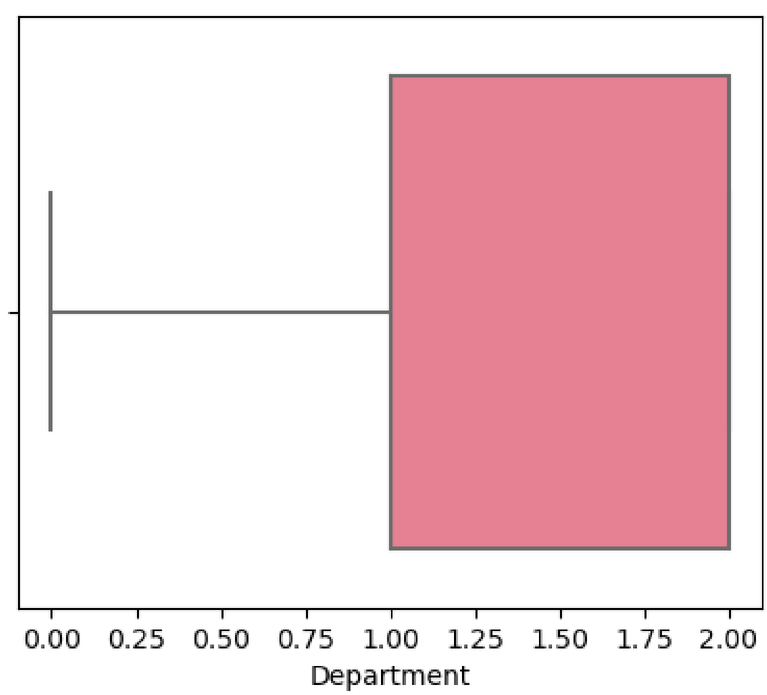
[72]:
```
Index(['Age', 'Attrition', 'BusinessTravel', 'DailyRate', 'Department',
       'DistanceFromHome', 'Education', 'EducationField', 'EmployeeCount',
       'EmployeeNumber', 'EnvironmentSatisfaction', 'Gender', 'HourlyRate',
       'JobInvolvement', 'JobLevel', 'JobRole', 'JobSatisfaction',
       'MaritalStatus', 'MonthlyIncome', 'MonthlyRate', 'NumCompaniesWorked',
       'Over18', 'OverTime', 'PercentSalaryHike', 'PerformanceRating',
       'RelationshipSatisfaction', 'StandardHours', 'StockOptionLevel',
       'TotalWorkingYears', 'TrainingTimesLastYear', 'WorkLifeBalance',
       'YearsAtCompany', 'YearsInCurrentRole', 'YearsSinceLastPromotion',
```

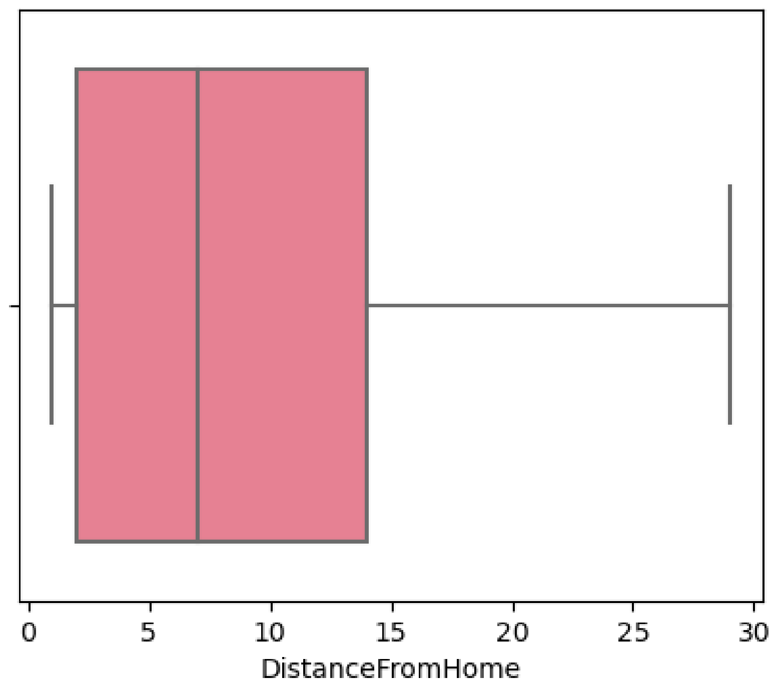
```
        'YearsWithCurrManager'],
       dtype='object')
```

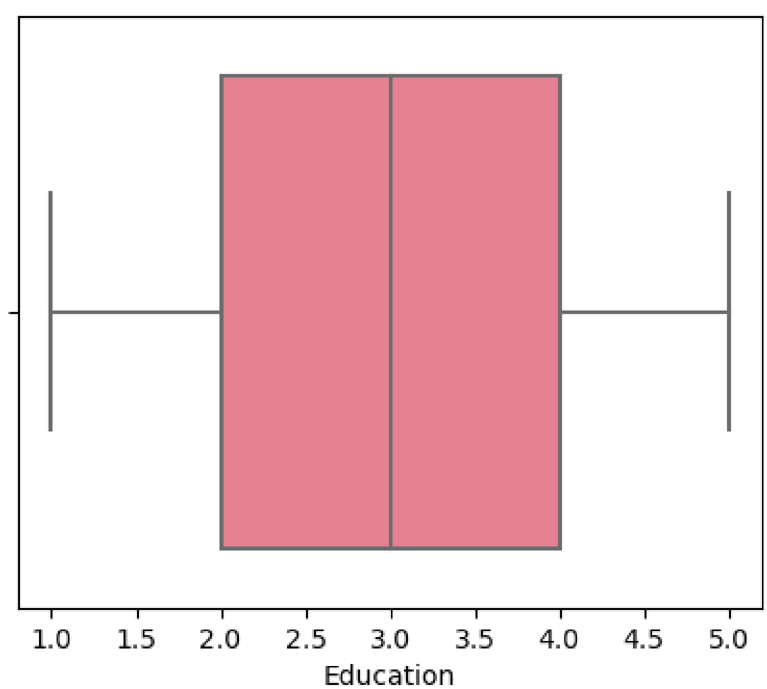
[269]:
```
def boxplots(col):
    plt.figure(figsize =(5,4))
    sns.boxplot(df, x=col, palette = 'husl')
    plt.show()

for i in list(df.select_dtypes(exclude=['object']).columns)[0:]:
    boxplots(i)
```

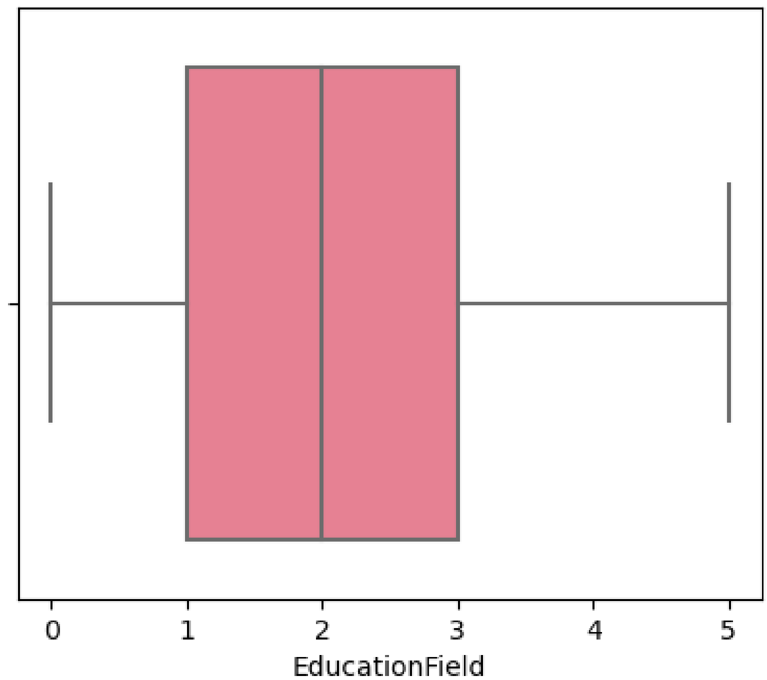


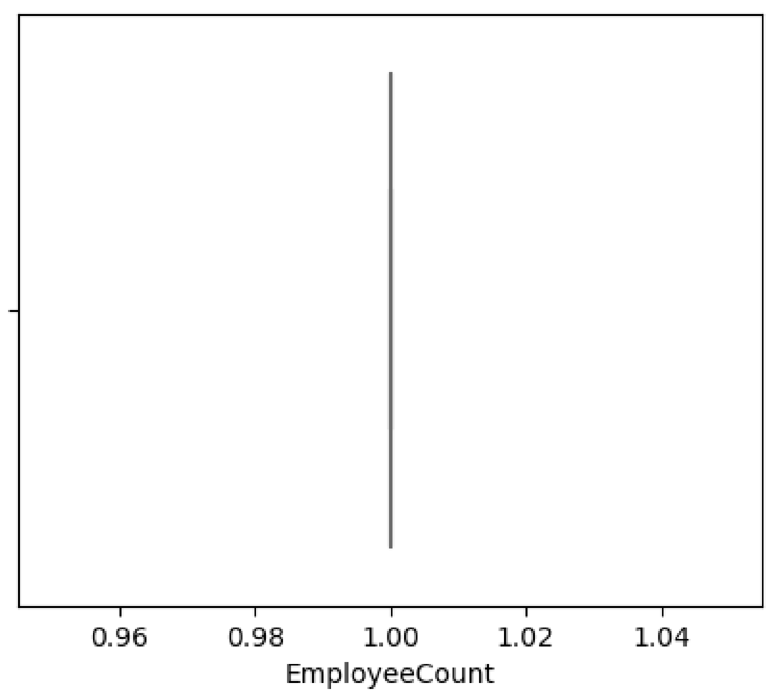

200
400
600
800
1000
1200
1400
DailyRate

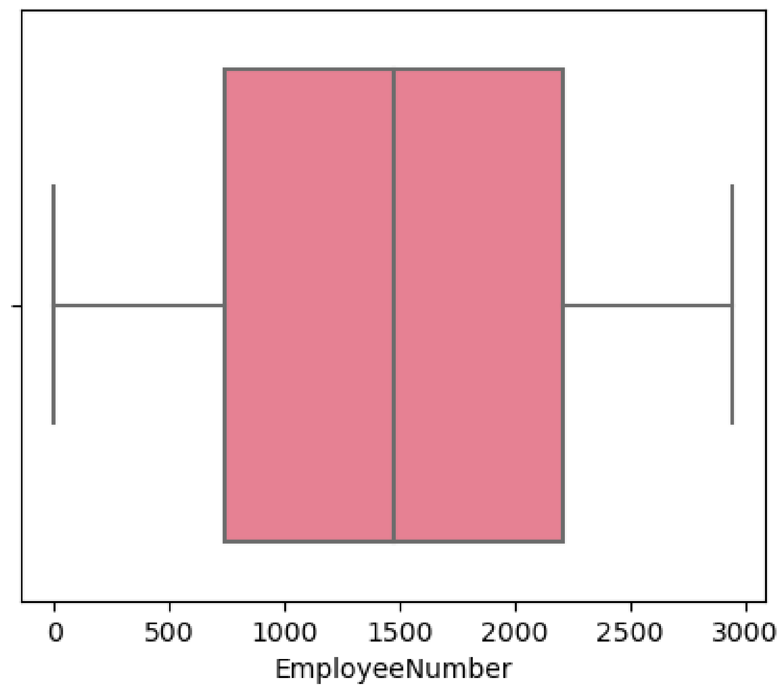

0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
Department

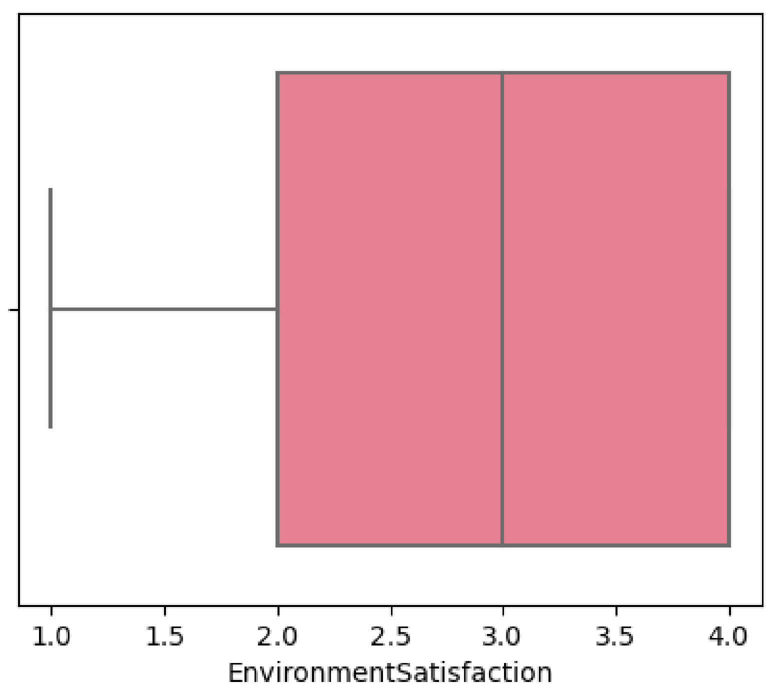

0
5
10
15
20
25
30
DistanceFromHome

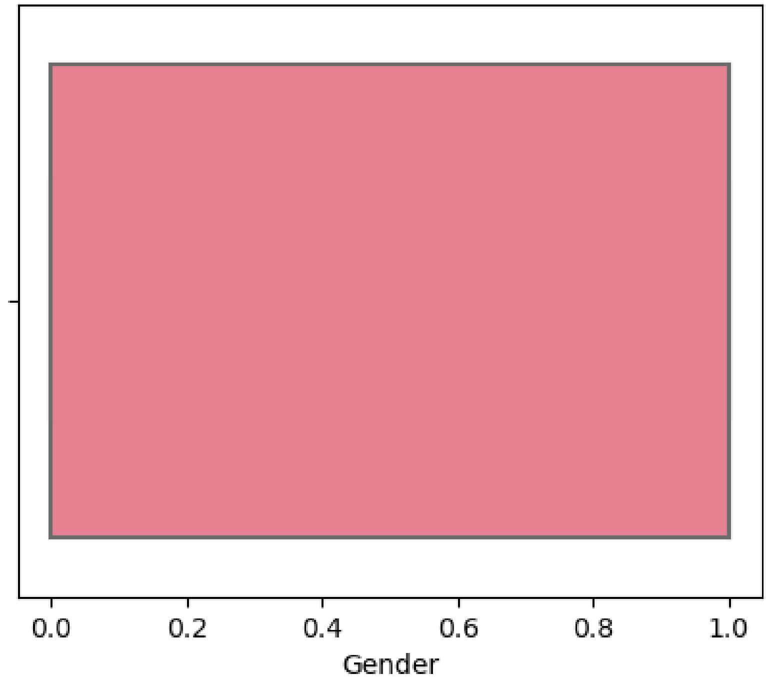

1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Education

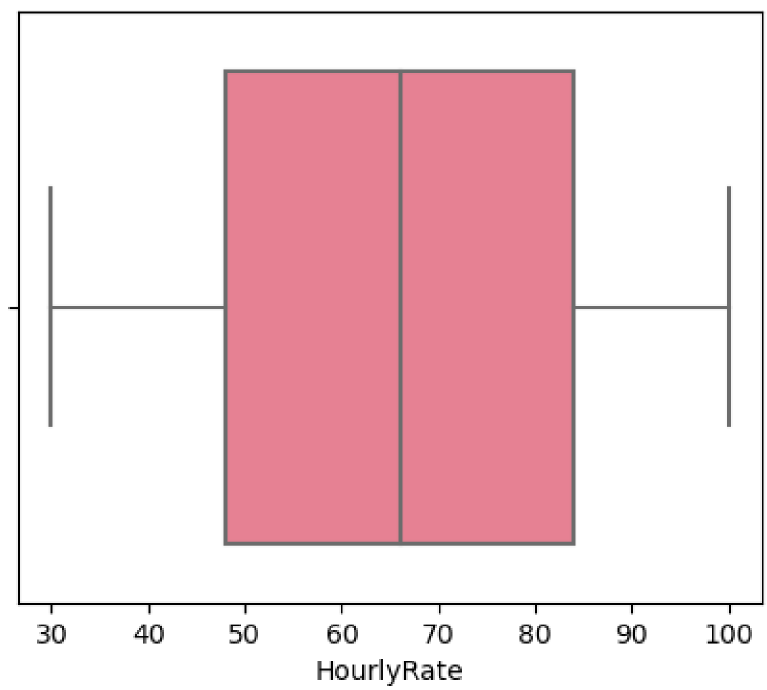

0
1
2
3
4
5
EducationField

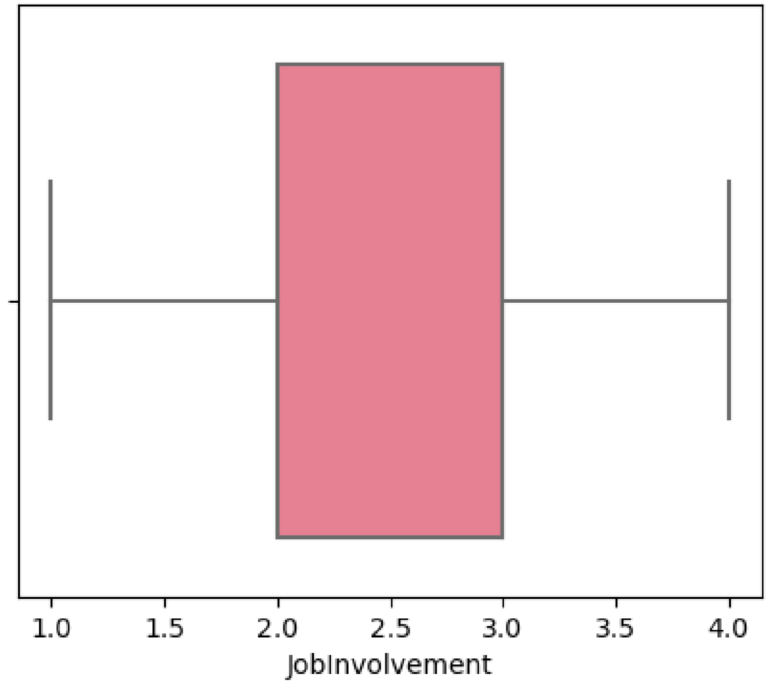

0.96
0.98
1.00
1.02
1.04
EmployeeCount

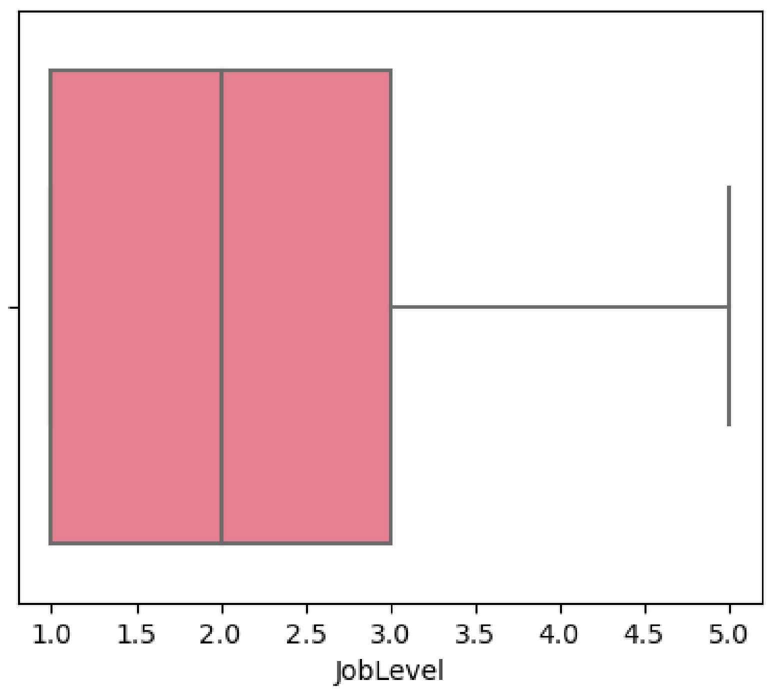

0
500
1000
1500
2000
2500
3000
EmployeeNumber

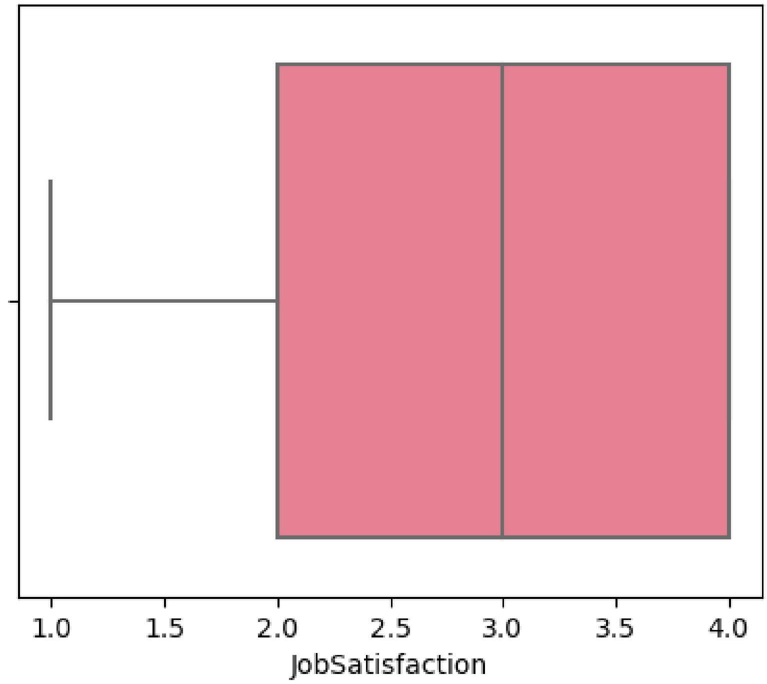

1.0
1.5
2.0
2.5
3.0
3.5
4.0
EnvironmentSatisfaction

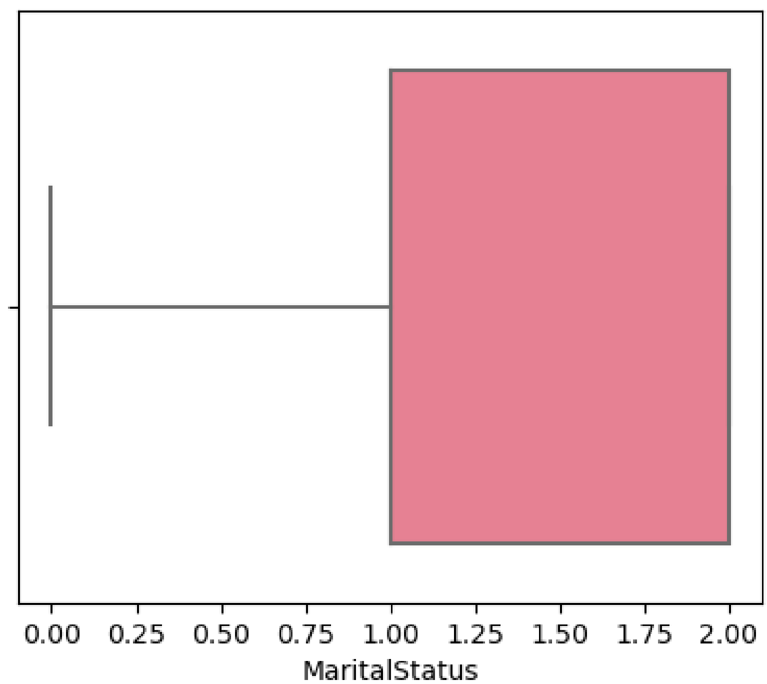

0.0
0.2
0.4
0.6
0.8
1.0
Gender

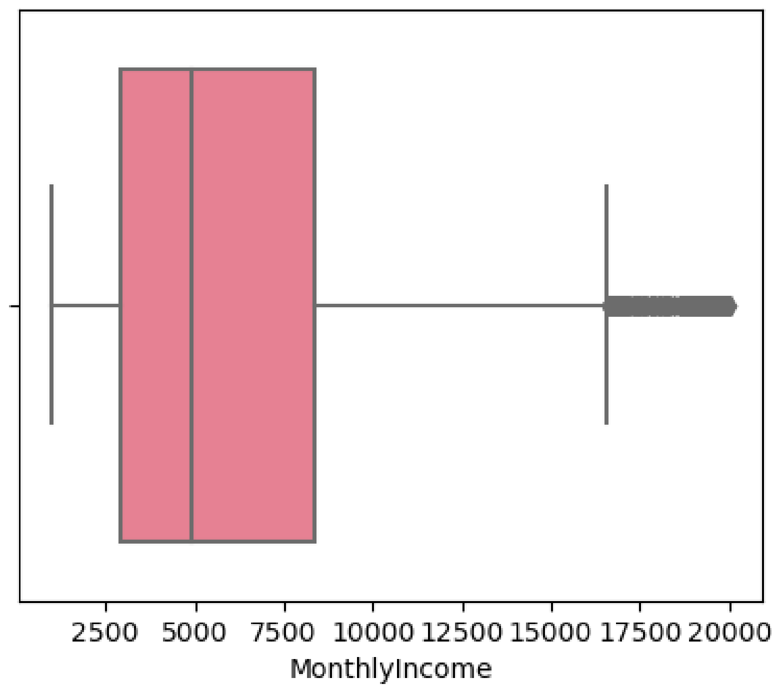

30
40
50
60
70
80
90
100
HourlyRate

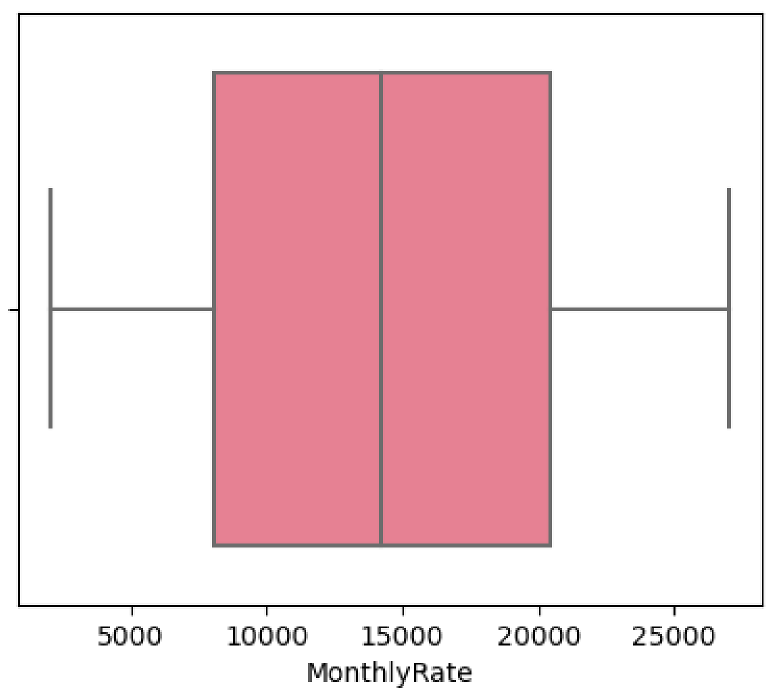

1.0
1.5
2.0
2.5
3.0
3.5
4.0
JobInvolvement

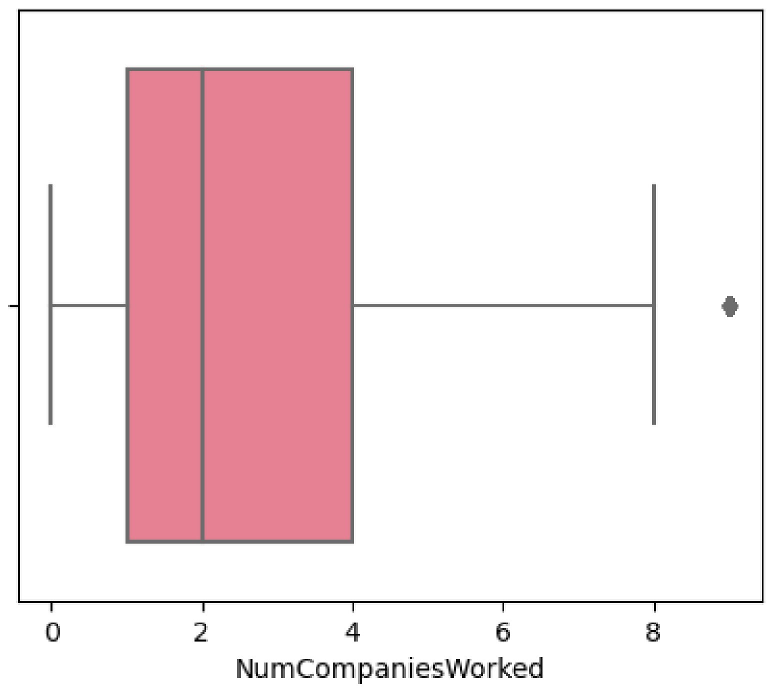

1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
JobLevel

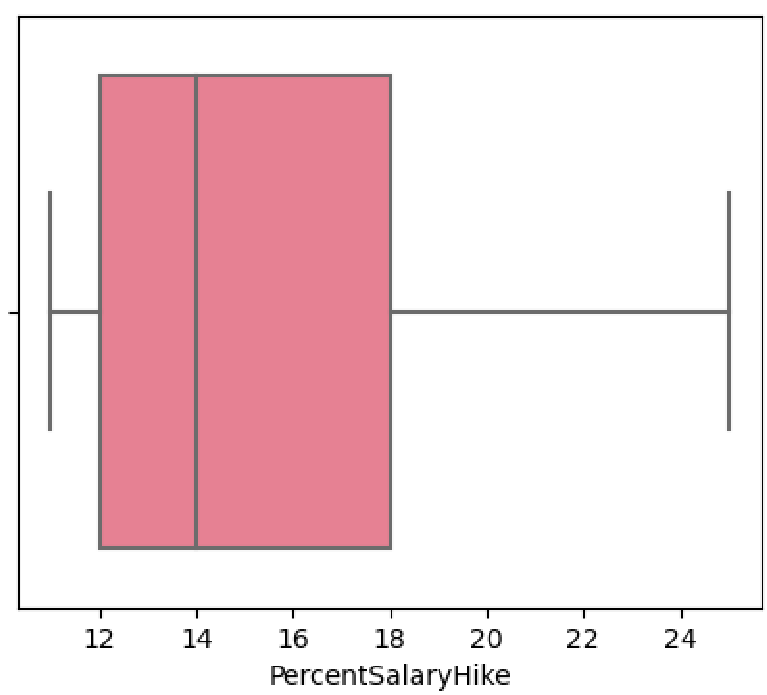

1.0
1.5
2.0
2.5
3.0
3.5
4.0
JobSatisfaction

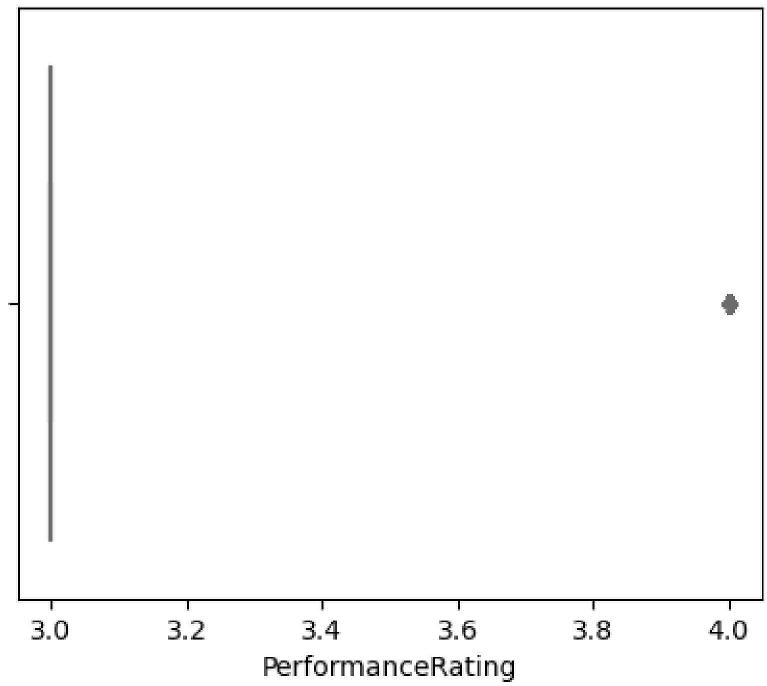

0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
MaritalStatus

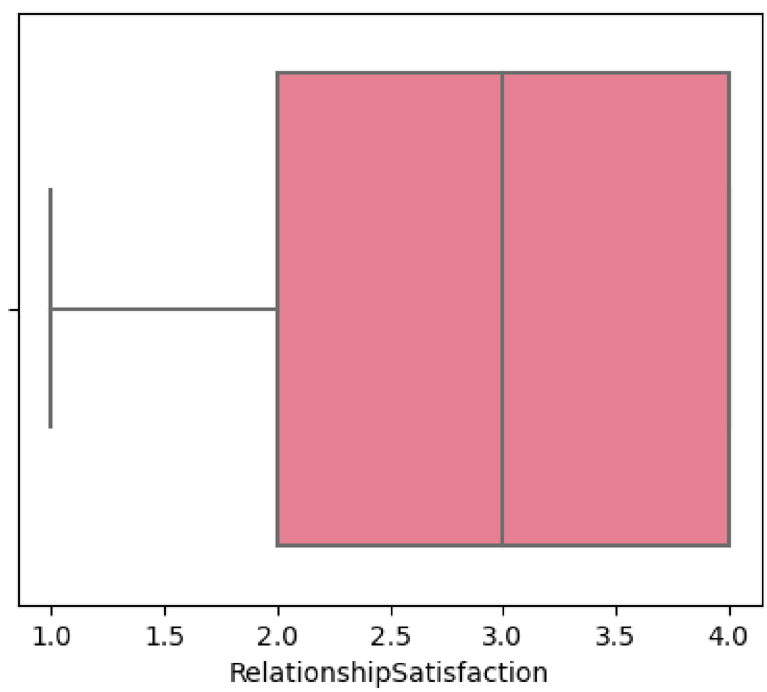

2500 5000 7500 10000 12500 15000 17500 20000
MonthlyIncome

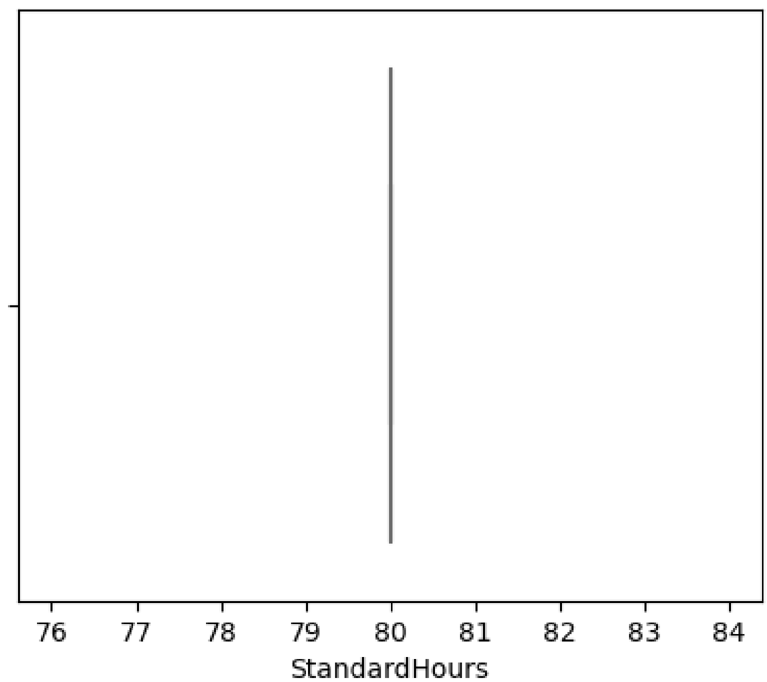

5000 10000 15000 20000 25000
MonthlyRate

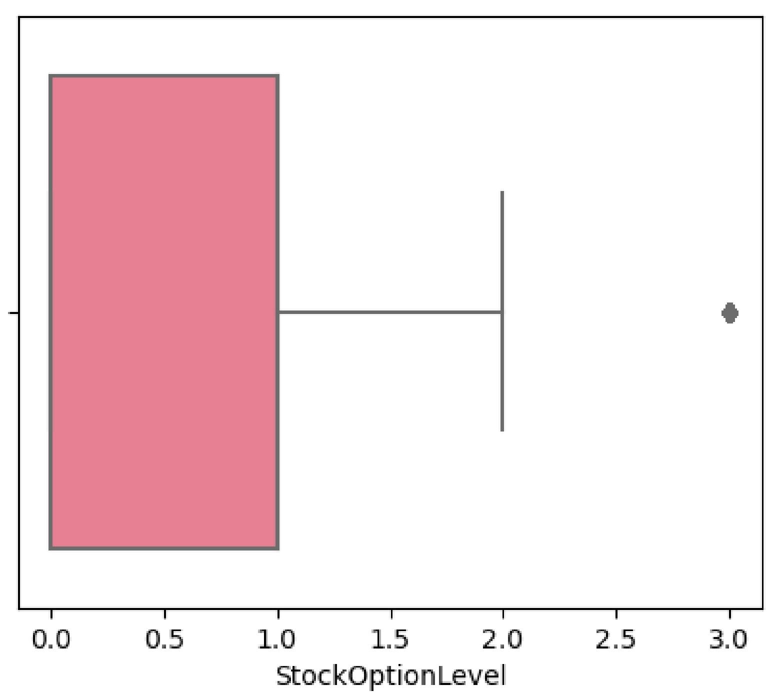

0
2
4
6
8
NumCompaniesWorked

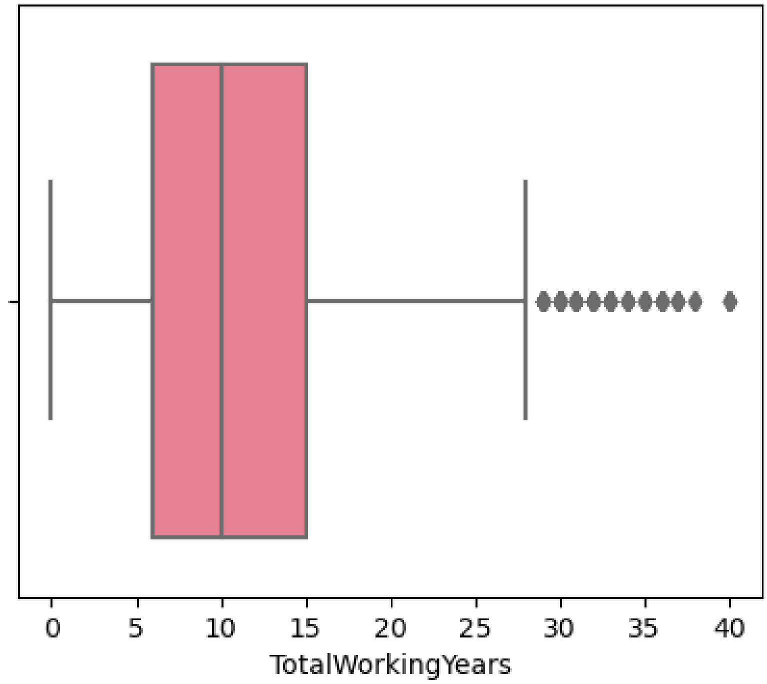

12
14
16
18
20
22
24
PercentSalaryHike

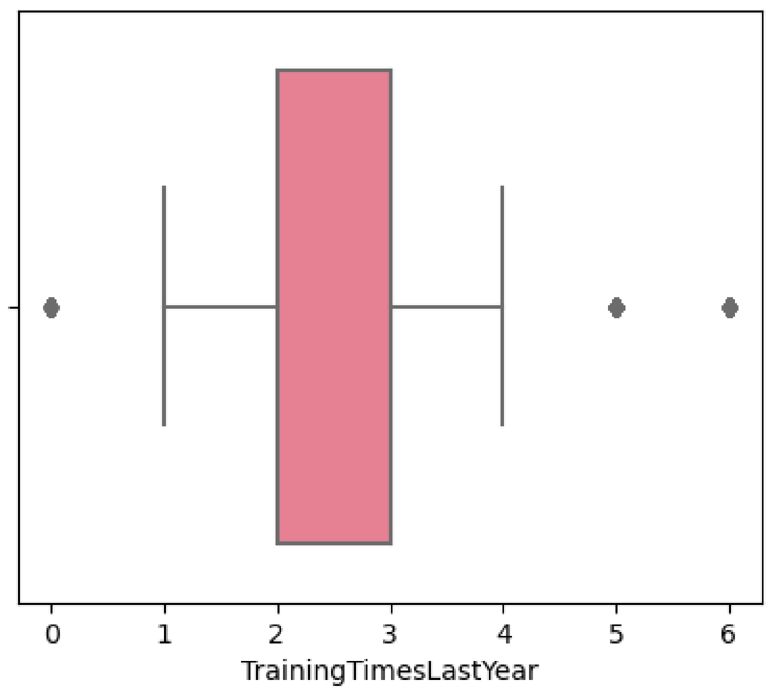

3.0
3.2
3.4
3.6
3.8
4.0
PerformanceRating

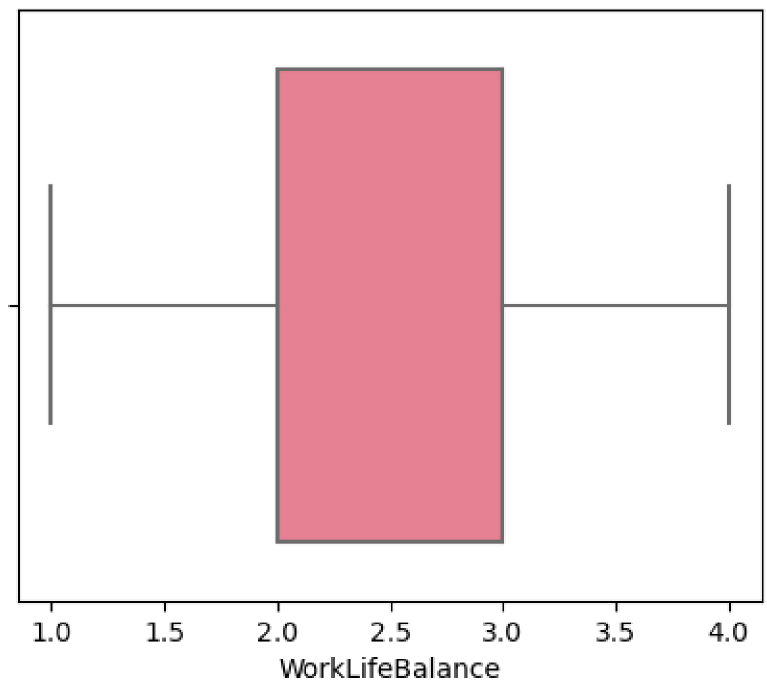

1.0
1.5
2.0
2.5
3.0
3.5
4.0
RelationshipSatisfaction

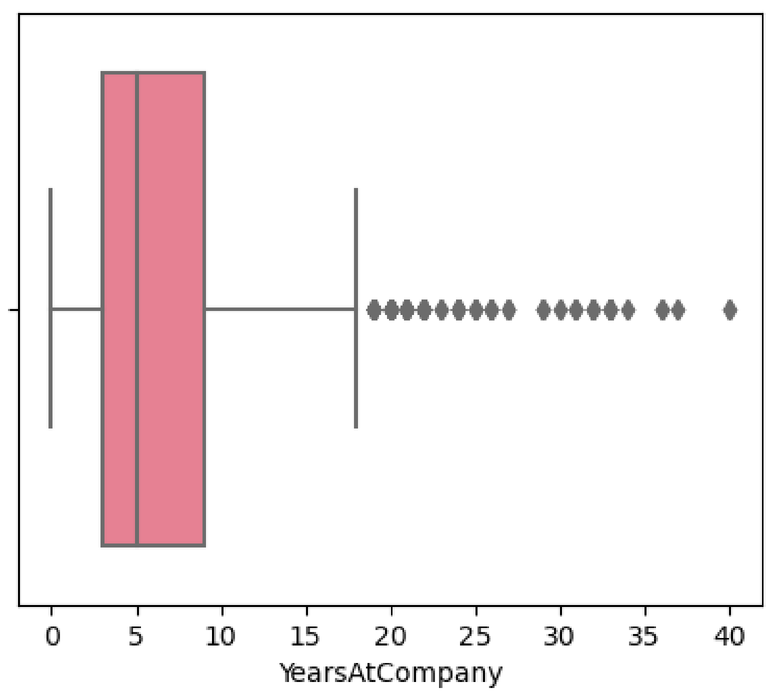

76
77
78
79
80
81
82
83
84
StandardHours

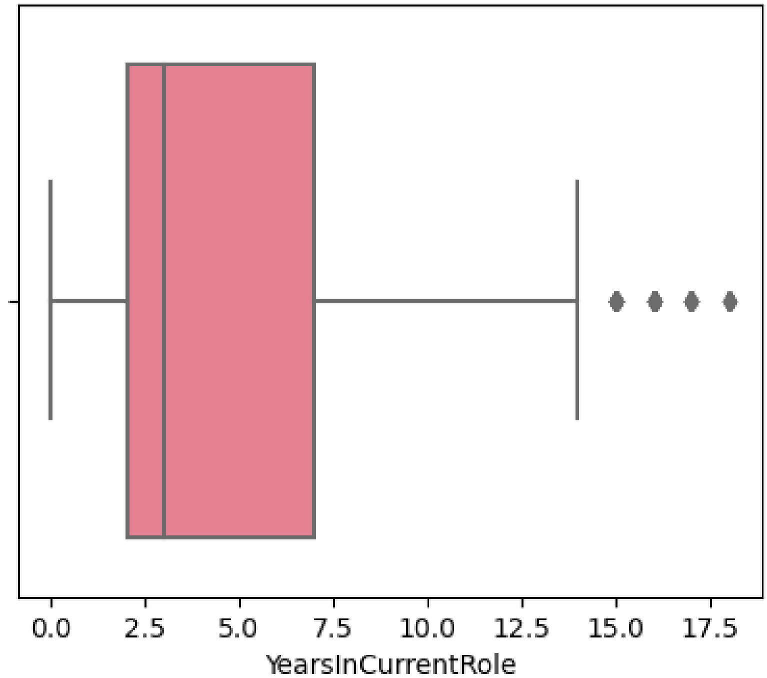

0.0
0.5
1.0
1.5
2.0
2.5
3.0
StockOptionLevel

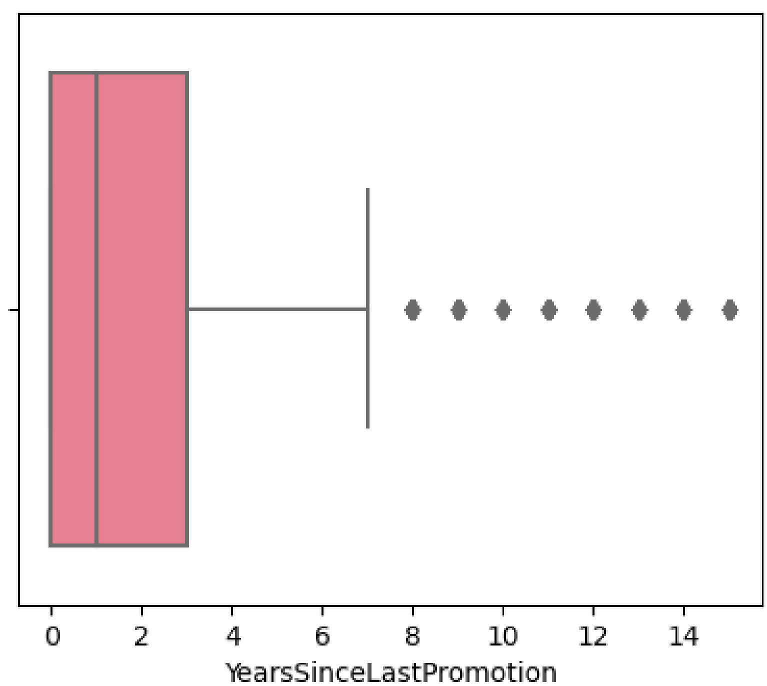

0
5
10
15
20
25
30
35
40
TotalWorkingYears

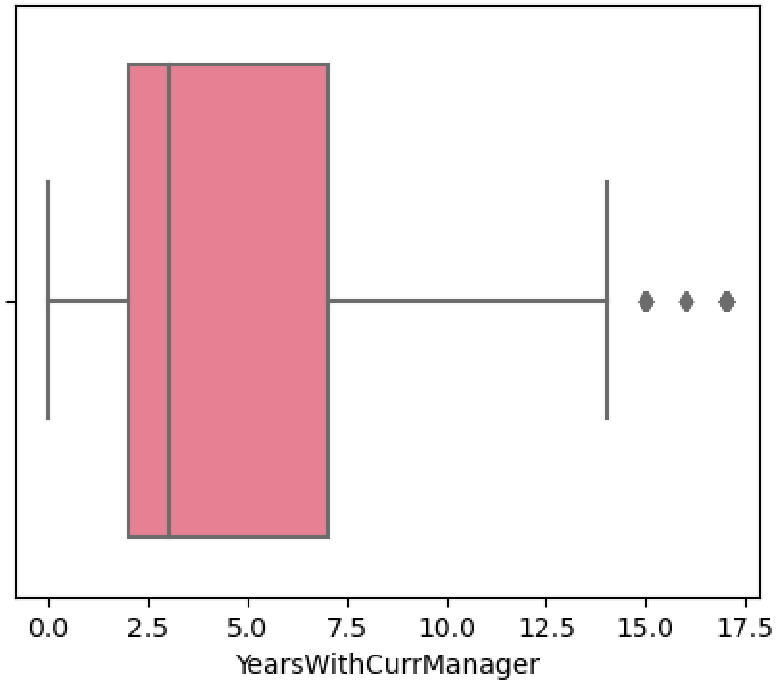

0
1
2
3
4
5
6
TrainingTimesLastYear


1.0
1.5
2.0
2.5
3.0
3.5
4.0
WorkLifeBalance


0
5
10
15
20
25
30
35
40
YearsAtCompany


0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
YearsInCurrentRole


0
2
4
6
8
10
12
14
YearsSinceLastPromotion



[270]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2940 entries, 0 to 2939
Data columns (total 35 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Age                       2940 non-null   int64
 1   Attrition                 2940 non-null   object
 2   BusinessTravel            2940 non-null   object
 3   DailyRate                 2940 non-null   int64
 4   Department                2940 non-null   int64
 5   DistanceFromHome          2940 non-null   int64
 6   Education                 2940 non-null   int64
 7   EducationField            2940 non-null   int64
 8   EmployeeCount             2940 non-null   int64
 9   EmployeeNumber            2940 non-null   int64
 10  EnvironmentSatisfaction   2940 non-null   int64
 11  Gender                    2940 non-null   int64
 12  HourlyRate                2940 non-null   int64
 13  JobInvolvement            2940 non-null   int64
 14  JobLevel                  2940 non-null   int64
 15  JobRole                   2940 non-null   object
 16  JobSatisfaction           2940 non-null   int64
 17  MaritalStatus             2940 non-null   int64
```

```
 18  MonthlyIncome             2940 non-null   int64
 19  MonthlyRate               2940 non-null   int64
 20  NumCompaniesWorked        2940 non-null   int64
 21  Over18                    2940 non-null   object
 22  OverTime                  2940 non-null   object
 23  PercentSalaryHike         2940 non-null   int64
 24  PerformanceRating         2940 non-null   int64
 25  RelationshipSatisfaction  2940 non-null   int64
 26  StandardHours             2940 non-null   int64
 27  StockOptionLevel          2940 non-null   int64
 28  TotalWorkingYears         2940 non-null   int64
 29  TrainingTimesLastYear     2940 non-null   int64
 30  WorkLifeBalance           2940 non-null   int64
 31  YearsAtCompany            2940 non-null   int64
 32  YearsInCurrentRole        2940 non-null   int64
 33  YearsSinceLastPromotion   2940 non-null   int64
 34  YearsWithCurrManager      2940 non-null   int64
dtypes: int64(30), object(5)
memory usage: 804.0+ KB
```

[271]:
```
df.describe(include='object').T
```

[271]:

	count	unique	top	freq
Attrition	2940	2	No	2466
BusinessTravel	2940	3	Travel_Rarely	2086
JobRole	2940	9	Sales Executive	652
Over18	2940	1	Y	2940
OverTime	2940	2	No	2108

[272]:
```
df.describe(include='int').T
```

[272]:

	count	mean	std	min	25% \
Age	2940.0	36.923810	9.133819	18.0	30.00
DailyRate	2940.0	802.485714	403.440447	102.0	465.00
Department	2940.0	1.260544	0.527703	0.0	1.00
DistanceFromHome	2940.0	9.192517	8.105485	1.0	2.00
Education	2940.0	2.912925	1.023991	1.0	2.00
EducationField	2940.0	2.247619	1.331143	0.0	1.00
EmployeeCount	2940.0	1.000000	0.000000	1.0	1.00
EmployeeNumber	2940.0	1470.500000	848.849221	1.0	735.75
EnvironmentSatisfaction	2940.0	2.721769	1.092896	1.0	2.00
Gender	2940.0	0.600000	0.489981	0.0	0.00
HourlyRate	2940.0	65.891156	20.325969	30.0	48.00
JobInvolvement	2940.0	2.729932	0.711440	1.0	2.00
JobLevel	2940.0	2.063946	1.106752	1.0	1.00
JobSatisfaction	2940.0	2.728571	1.102658	1.0	2.00
MaritalStatus	2940.0	1.097279	0.729997	0.0	1.00

MonthlyIncome	2940.0	6502.931293	4707.155770	1009.0	2911.00
MonthlyRate	2940.0	14313.103401	7116.575021	2094.0	8045.00
NumCompaniesWorked	2940.0	2.693197	2.497584	0.0	1.00
PercentSalaryHike	2940.0	15.209524	3.659315	11.0	12.00
PerformanceRating	2940.0	3.153741	0.360762	3.0	3.00
RelationshipSatisfaction	2940.0	2.712245	1.081025	1.0	2.00
StandardHours	2940.0	80.000000	0.000000	80.0	80.00
StockOptionLevel	2940.0	0.793878	0.851932	0.0	0.00
TotalWorkingYears	2940.0	11.279592	7.779458	0.0	6.00
TrainingTimesLastYear	2940.0	2.799320	1.289051	0.0	2.00
WorkLifeBalance	2940.0	2.761224	0.706356	1.0	2.00
YearsAtCompany	2940.0	7.008163	6.125483	0.0	3.00
YearsInCurrentRole	2940.0	4.229252	3.622521	0.0	2.00
YearsSinceLastPromotion	2940.0	2.187755	3.221882	0.0	0.00
YearsWithCurrManager	2940.0	4.123129	3.567529	0.0	2.00

	50%	75%	max
Age	36.0	43.00	60.0
DailyRate	802.0	1157.00	1499.0
Department	1.0	2.00	2.0
DistanceFromHome	7.0	14.00	29.0
Education	3.0	4.00	5.0
EducationField	2.0	3.00	5.0
EmployeeCount	1.0	1.00	1.0
EmployeeNumber	1470.5	2205.25	2940.0
EnvironmentSatisfaction	3.0	4.00	4.0
Gender	1.0	1.00	1.0
HourlyRate	66.0	84.00	100.0
JobInvolvement	3.0	3.00	4.0
JobLevel	2.0	3.00	5.0
JobSatisfaction	3.0	4.00	4.0
MaritalStatus	1.0	2.00	2.0
MonthlyIncome	4919.0	8380.00	19999.0
MonthlyRate	14235.5	20462.00	26999.0
NumCompaniesWorked	2.0	4.00	9.0
PercentSalaryHike	14.0	18.00	25.0
PerformanceRating	3.0	3.00	4.0
RelationshipSatisfaction	3.0	4.00	4.0
StandardHours	80.0	80.00	80.0
StockOptionLevel	1.0	1.00	3.0
TotalWorkingYears	10.0	15.00	40.0
TrainingTimesLastYear	3.0	3.00	6.0
WorkLifeBalance	3.0	3.00	4.0
YearsAtCompany	5.0	9.00	40.0
YearsInCurrentRole	3.0	7.00	18.0
YearsSinceLastPromotion	1.0	3.00	15.0
YearsWithCurrManager	3.0	7.00	17.0

0.7 Fix outlier or Remove outlier

0.8 IQR Method -

- we can cap the value between the upper bound and lower bound

[273]:
```python
def outlier(data):
    q1 = data.quantile(0.25)
    q3 = data.quantile(0.75)
    iqr = q3 - q1
    upper_bound = q3 + 1.5 * iqr
    lower_bound = q1 - 1.5 * iqr
    return data.clip(upper_bound,lower_bound)
```

[274]:
```python
df.head()
```

[274]:
```
   Age Attrition     BusinessTravel  DailyRate  Department  DistanceFromHome  \
0   41       Yes      Travel_Rarely       1102           2                 1
1   49        No  Travel_Frequently        279           1                 8
2   37       Yes      Travel_Rarely       1373           1                 2
3   33        No  Travel_Frequently       1392           1                 3
4   27        No      Travel_Rarely        591           1                 2

   Education  EducationField  EmployeeCount  EmployeeNumber  …  \
0          2               1              1               1  …
1          1               1              1               2  …
2          2               4              1               3  …
3          4               1              1               4  …
4          1               3              1               5  …

   RelationshipSatisfaction  StandardHours  StockOptionLevel  \
0                         1             80                 0
1                         4             80                 1
2                         2             80                 0
3                         3             80                 0
4                         4             80                 1

   TotalWorkingYears  TrainingTimesLastYear WorkLifeBalance  YearsAtCompany  \
0                  8                      0               1               6
1                 10                      3               3              10
2                  7                      3               3               0
3                  8                      3               3               8
4                  6                      3               3               2

   YearsInCurrentRole  YearsSinceLastPromotion  YearsWithCurrManager
0                   4                        0                     5
1                   7                        1                     7
2                   0                        0                     0
```

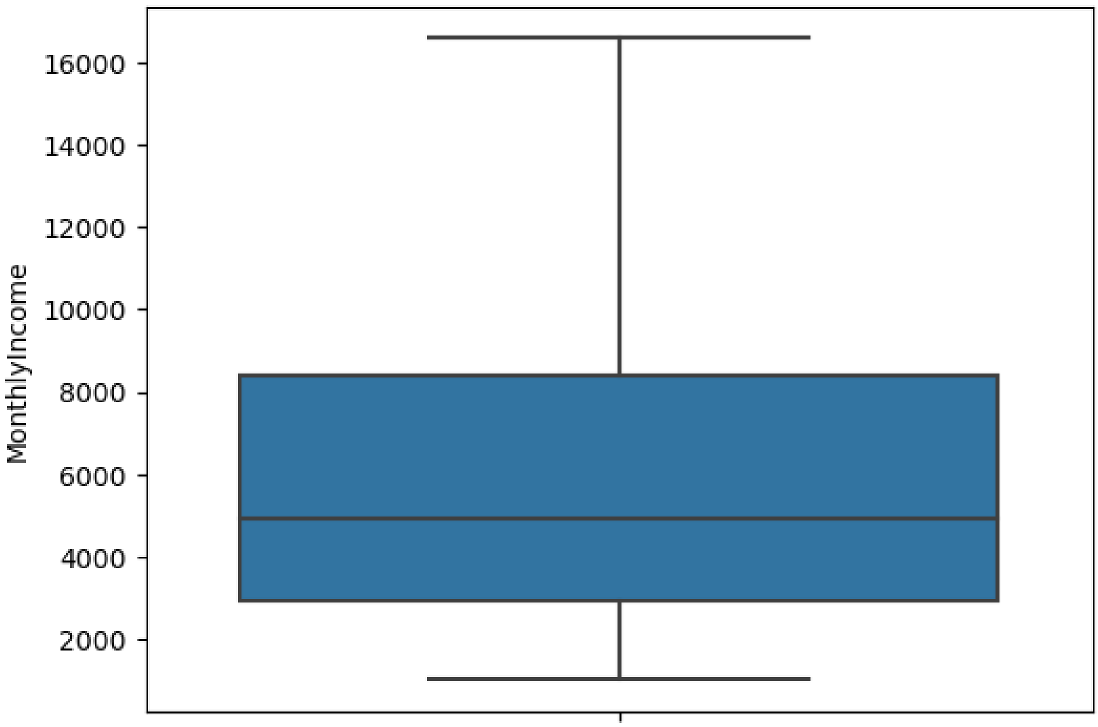
3	7	3	0
4	2	2	2

[5 rows x 35 columns]

[275]:
```
df["MonthlyIncome"] = outlier(df.MonthlyIncome)
```

[276]:
```
sns.boxplot(y="MonthlyIncome" , data=df )
```

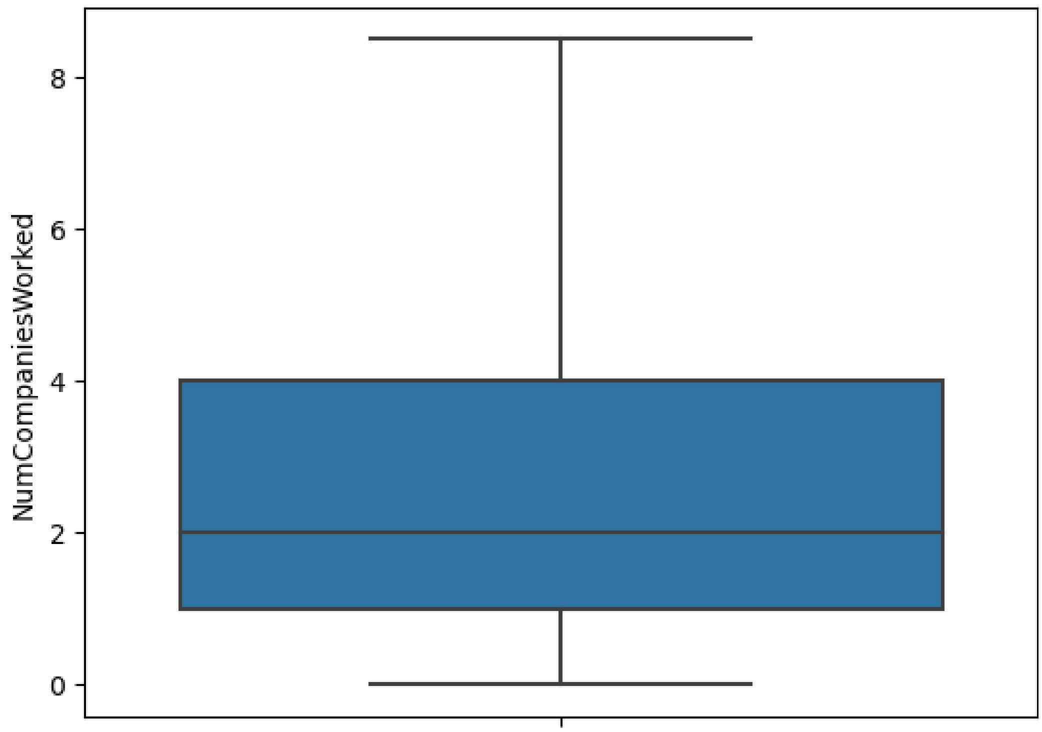
[276]: <Axes: ylabel='MonthlyIncome'>



[277]:
```
df["NumCompaniesWorked"] = outlier(df.NumCompaniesWorked)
```

[278]:
```
sns.boxplot(y="NumCompaniesWorked" , data=df )
```

[278]: <Axes: ylabel='NumCompaniesWorked'>



[279]:
```
df["PerformanceRating"] = outlier(df.PerformanceRating)
```

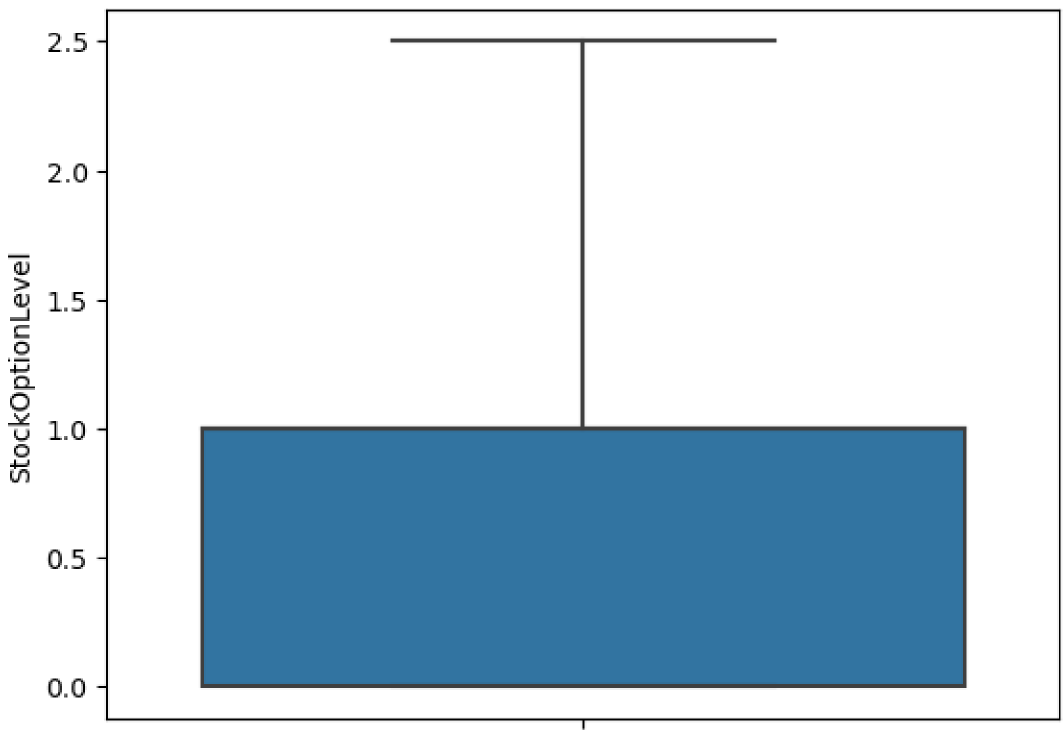
[280]:
```
sns.boxplot(y="PerformanceRating" , data=df )
```

[280]: <Axes: ylabel='PerformanceRating'>

[281]:
```
df["StockOptionLevel"] = outlier(df.StockOptionLevel)
```

[282]:
```
sns.boxplot(y="StockOptionLevel" , data=df )
```

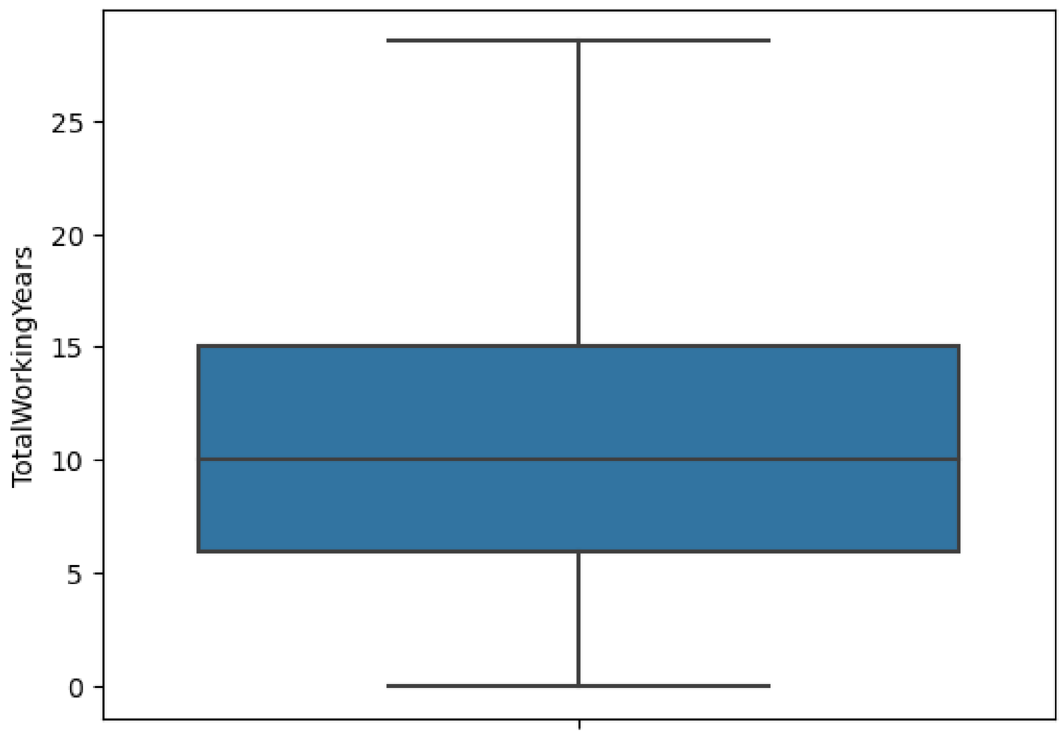
[282]: `<Axes: ylabel='StockOptionLevel'>`



[283]:
```
df["TotalWorkingYears"] = outlier(df.TotalWorkingYears)
```

[284]:
```
sns.boxplot(y="TotalWorkingYears" , data=df )
```

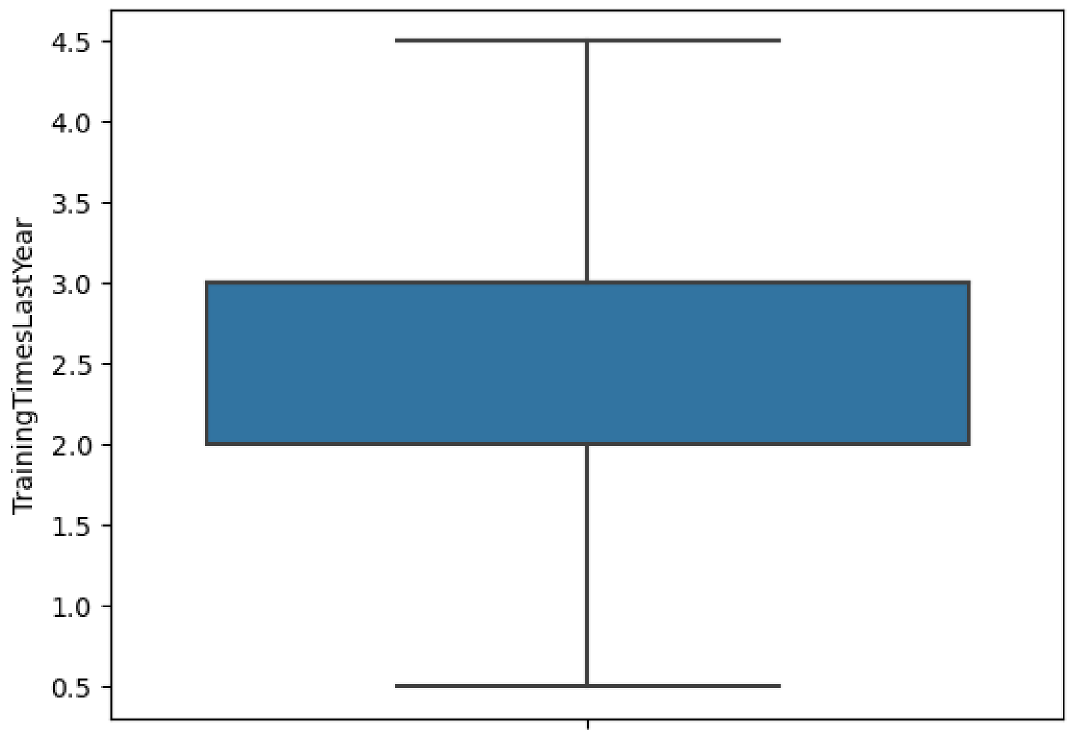
[284]: <Axes: ylabel='TotalWorkingYears'>



[285]:
```
df["TrainingTimesLastYear"] = outlier(df.TrainingTimesLastYear)
```

[286]:
```
sns.boxplot(y="TrainingTimesLastYear" , data=df )
```

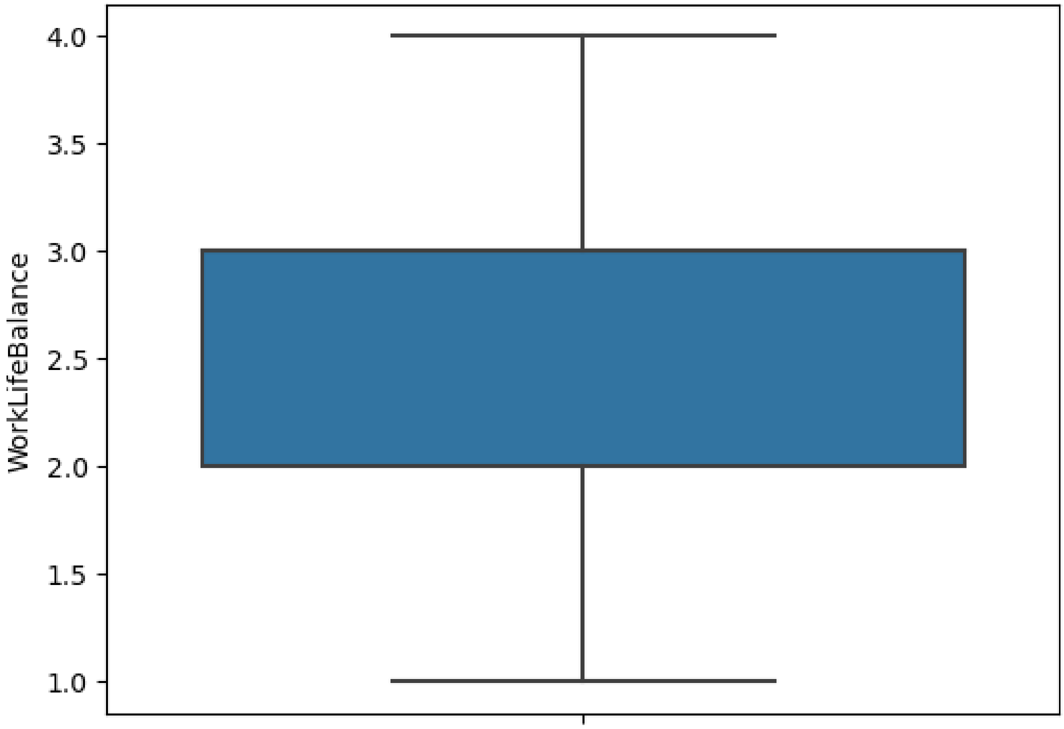
[286]: <Axes: ylabel='TrainingTimesLastYear'>



[287]:
```
df["WorkLifeBalance"] = outlier(df.WorkLifeBalance)
```

[288]:
```
sns.boxplot(y="WorkLifeBalance" , data=df )
```

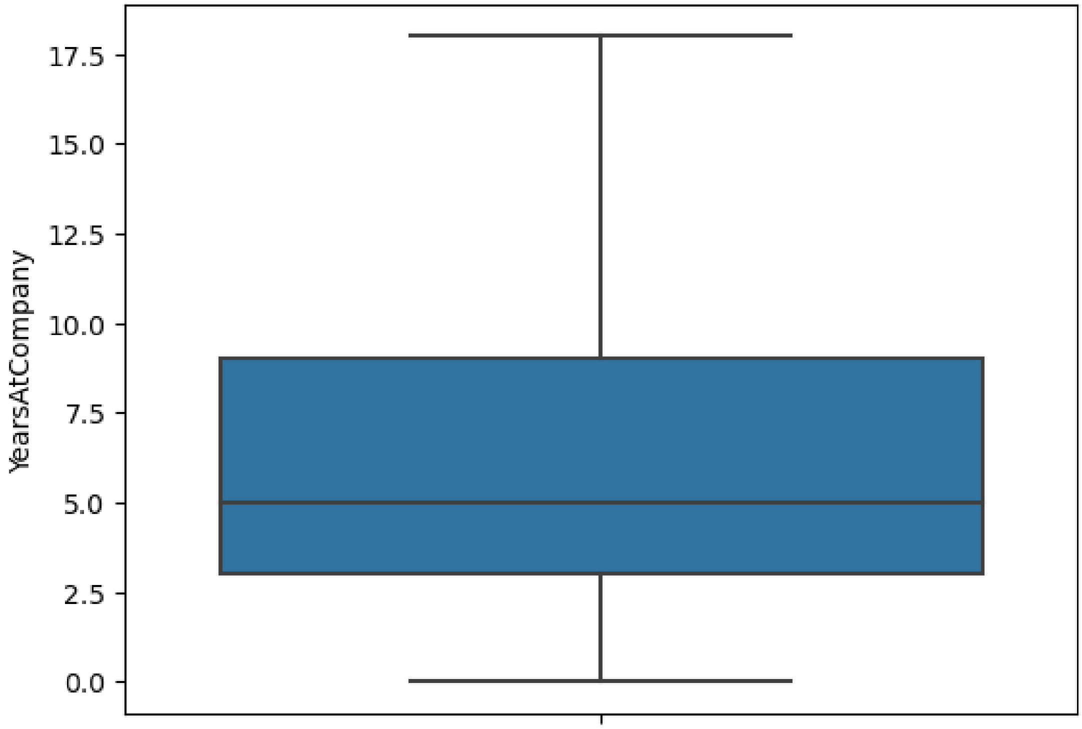
[288]: <Axes: ylabel='WorkLifeBalance'>



[289]:
```
df["YearsAtCompany"] = outlier(df.YearsAtCompany)
```

[290]:
```
sns.boxplot(y="YearsAtCompany" , data=df )
```

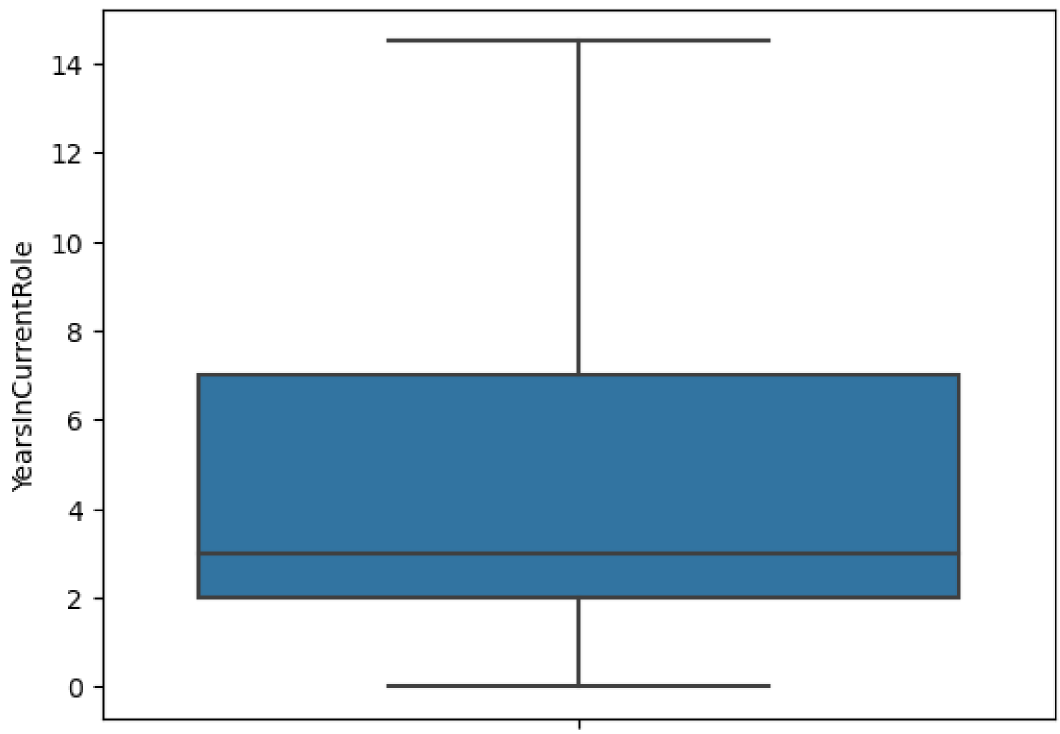
[290]: <Axes: ylabel='YearsAtCompany'>



[291]:
```
df["YearsInCurrentRole"] = outlier(df.YearsInCurrentRole)
```

[292]:
```
sns.boxplot(y="YearsInCurrentRole" , data=df )
```

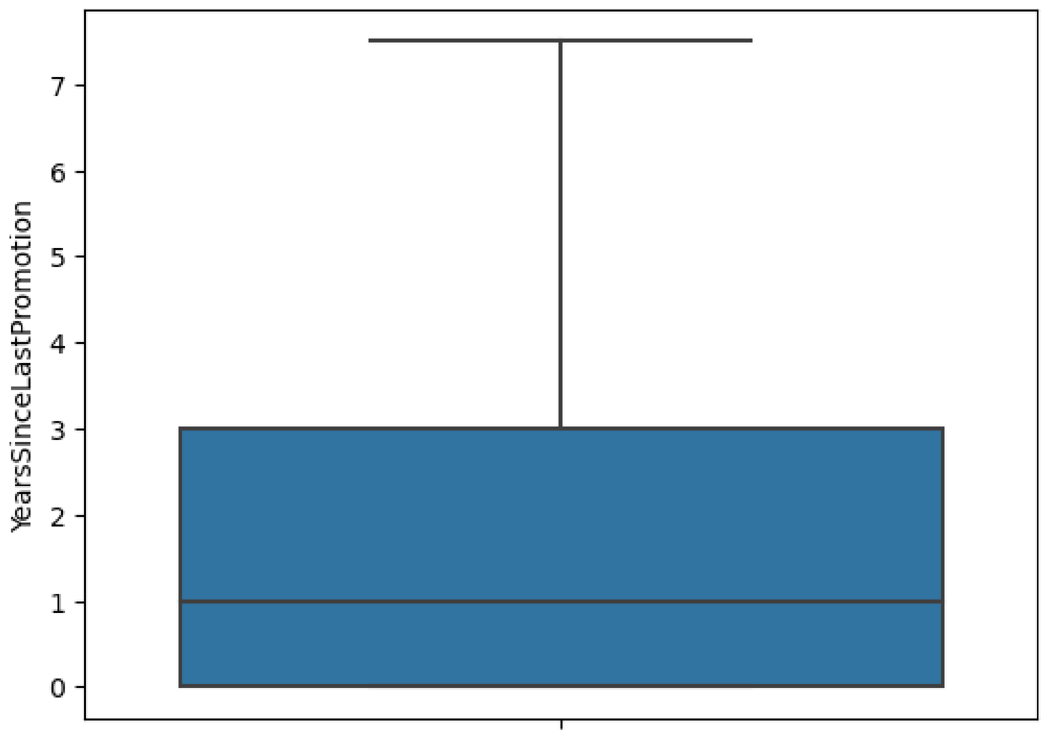
[292]: <Axes: ylabel='YearsInCurrentRole'>



[293]:
```
df["YearsSinceLastPromotion"] = outlier(df.YearsSinceLastPromotion)
```

[294]:
```
sns.boxplot(y="YearsSinceLastPromotion" , data=df )
```

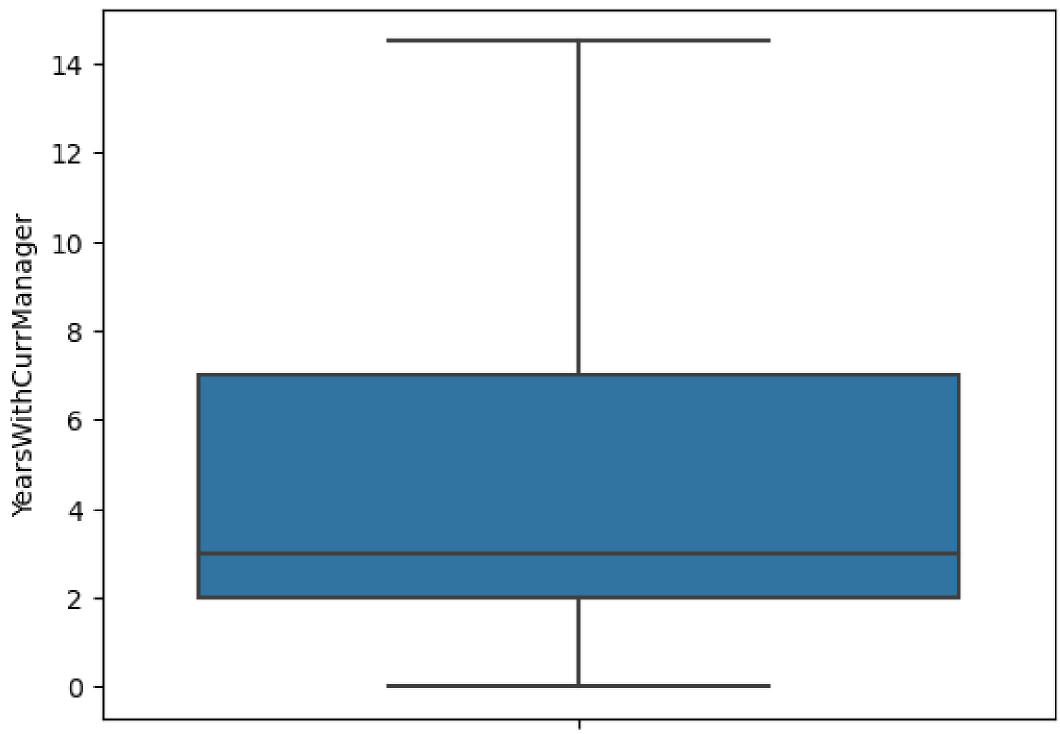
[294]:
```
<Axes: ylabel='YearsSinceLastPromotion'>
```



[295]:
```
df["YearsWithCurrManager"] = outlier(df.YearsWithCurrManager)
```

[296]:
```
sns.boxplot(y="YearsWithCurrManager" , data=df )
```

[296]: <Axes: ylabel='YearsWithCurrManager'>



0.9 Models Buildings :

- logistic Regression
- Random Forest
- Decision Tree

[11]:
```
correlations = df.corr()
correlations
```

[11]:

	Age	DailyRate	DistanceFromHome	Education	\
Age	1.000000	0.010661	-0.001686	0.208034	
DailyRate	0.010661	1.000000	-0.004985	-0.016806	
DistanceFromHome	-0.001686	-0.004985	1.000000	0.021042	
Education	0.208034	-0.016806	0.021042	1.000000	
EmployeeCount	NaN	NaN	NaN	NaN	
EmployeeNumber	-0.005175	-0.025742	0.016464	0.020950	
EnvironmentSatisfaction	0.010146	0.018355	-0.016075	-0.027128	
HourlyRate	0.024287	0.023381	0.031131	0.016775	
JobInvolvement	0.029820	0.046135	0.008783	0.042438	
JobLevel	0.509604	0.002966	0.005303	0.101589	
JobSatisfaction	-0.004892	0.030571	-0.003669	-0.011296	
MonthlyIncome	0.497855	0.007707	-0.017014	0.094961	
MonthlyRate	0.028051	-0.032182	0.027473	-0.026084	

```
NumCompaniesWorked        0.299635  0.038153          -0.029251  0.126317
PercentSalaryHike         0.003634  0.022704           0.040235 -0.011111
PerformanceRating         0.001904  0.000473           0.027110 -0.024539
RelationshipSatisfaction  0.053535  0.007846           0.006557 -0.009118
StandardHours                  NaN       NaN                NaN       NaN
StockOptionLevel          0.037510  0.042143           0.044872  0.018422
TotalWorkingYears         0.680381  0.014515           0.004628  0.148280
TrainingTimesLastYear    -0.019621  0.002453          -0.036942 -0.025100
WorkLifeBalance          -0.021490 -0.037848          -0.026556  0.009819
YearsAtCompany            0.311309 -0.034055           0.009508  0.069114
YearsInCurrentRole        0.212901  0.009932           0.018845  0.060236
YearsSinceLastPromotion   0.216513 -0.033229           0.010029  0.054254
YearsWithCurrManager      0.202089 -0.026363           0.014406  0.069065

                          EmployeeCount  EmployeeNumber  \
Age                                 NaN       -0.005175
DailyRate                           NaN       -0.025742
DistanceFromHome                    NaN        0.016464
Education                           NaN        0.020950
EmployeeCount                       NaN             NaN
EmployeeNumber                      NaN        1.000000
EnvironmentSatisfaction             NaN        0.008712
HourlyRate                          NaN        0.017377
JobInvolvement                      NaN       -0.003552
JobLevel                            NaN       -0.009020
JobSatisfaction                     NaN       -0.022970
MonthlyIncome                       NaN       -0.007188
MonthlyRate                         NaN        0.006177
NumCompaniesWorked                  NaN       -0.000345
PercentSalaryHike                   NaN       -0.006685
PerformanceRating                   NaN       -0.010338
RelationshipSatisfaction            NaN       -0.034827
StandardHours                       NaN             NaN
StockOptionLevel                    NaN        0.031226
TotalWorkingYears                   NaN       -0.007047
TrainingTimesLastYear               NaN        0.011953
WorkLifeBalance                     NaN        0.005370
YearsAtCompany                      NaN       -0.005779
YearsInCurrentRole                  NaN       -0.004427
YearsSinceLastPromotion             NaN       -0.004575
YearsWithCurrManager                NaN       -0.004716

                          EnvironmentSatisfaction  HourlyRate  JobInvolvement  \
Age                                      0.010146    0.024287        0.029820
DailyRate                                0.018355    0.023381        0.046135
DistanceFromHome                        -0.016075    0.031131        0.008783
Education                               -0.027128    0.016775        0.042438
```

```
EmployeeCount                        NaN        NaN            NaN
EmployeeNumber                  0.008712   0.017377      -0.003552
EnvironmentSatisfaction         1.000000  -0.049857      -0.008278
HourlyRate                     -0.049857   1.000000       0.042861
JobInvolvement                 -0.008278   0.042861       1.000000
JobLevel                        0.001212  -0.027853      -0.012630
JobSatisfaction                -0.006784  -0.071335      -0.021476
MonthlyIncome                  -0.006259  -0.015794      -0.015271
MonthlyRate                     0.037600  -0.015297      -0.016322
NumCompaniesWorked              0.012594   0.022157       0.015012
PercentSalaryHike              -0.031701  -0.009062      -0.017205
PerformanceRating              -0.029548  -0.002172      -0.029071
RelationshipSatisfaction        0.007665   0.001330       0.034297
StandardHours                        NaN        NaN            NaN
StockOptionLevel                0.003432   0.050263       0.021523
TotalWorkingYears              -0.002693  -0.002334      -0.005533
TrainingTimesLastYear          -0.019359  -0.008548      -0.015338
WorkLifeBalance                 0.027627  -0.004607      -0.014617
YearsAtCompany                  0.001458  -0.019582      -0.021355
YearsInCurrentRole              0.018007  -0.024106       0.008717
YearsSinceLastPromotion         0.016194  -0.026716      -0.024184
YearsWithCurrManager           -0.004999  -0.020123       0.025976

                          JobLevel  …  RelationshipSatisfaction  \
Age                       0.509604  …                  0.053535
DailyRate                 0.002966  …                  0.007846
DistanceFromHome          0.005303  …                  0.006557
Education                 0.101589  …                 -0.009118
EmployeeCount                  NaN  …                       NaN
EmployeeNumber           -0.009020  …                 -0.034827
EnvironmentSatisfaction   0.001212  …                  0.007665
HourlyRate               -0.027853  …                  0.001330
JobInvolvement           -0.012630  …                  0.034297
JobLevel                  1.000000  …                  0.021642
JobSatisfaction          -0.001944  …                 -0.012454
MonthlyIncome             0.950300  …                  0.025873
MonthlyRate               0.039563  …                 -0.004085
NumCompaniesWorked        0.142501  …                  0.052733
PercentSalaryHike        -0.034730  …                 -0.040490
PerformanceRating        -0.021222  …                 -0.031351
RelationshipSatisfaction  0.021642  …                  1.000000
StandardHours                  NaN  …                       NaN
StockOptionLevel          0.013984  …                 -0.045952
TotalWorkingYears         0.782208  …                  0.024054
TrainingTimesLastYear    -0.018191  …                  0.002497
WorkLifeBalance           0.037818  …                  0.019604
YearsAtCompany            0.534739  …                  0.019367
```

```
YearsInCurrentRole        0.389447  ...                 -0.015123
YearsSinceLastPromotion   0.353885  ...                  0.033493
YearsWithCurrManager      0.375281  ...                 -0.000867

                          StandardHours  StockOptionLevel  TotalWorkingYears  \
Age                                 NaN          0.037510           0.680381
DailyRate                           NaN          0.042143           0.014515
DistanceFromHome                    NaN          0.044872           0.004628
Education                           NaN          0.018422           0.148280
EmployeeCount                       NaN               NaN                NaN
EmployeeNumber                      NaN          0.031226          -0.007047
EnvironmentSatisfaction             NaN          0.003432          -0.002693
HourlyRate                          NaN          0.050263          -0.002334
JobInvolvement                      NaN          0.021523          -0.005533
JobLevel                            NaN          0.013984           0.782208
JobSatisfaction                     NaN          0.010690          -0.020185
MonthlyIncome                       NaN          0.005408           0.772893
MonthlyRate                         NaN         -0.034323           0.026442
NumCompaniesWorked                  NaN          0.030075           0.237639
PercentSalaryHike                   NaN          0.007528          -0.020608
PerformanceRating                   NaN          0.003506           0.006744
RelationshipSatisfaction            NaN         -0.045952           0.024054
StandardHours                       NaN               NaN                NaN
StockOptionLevel                    NaN          1.000000           0.010136
TotalWorkingYears                   NaN          0.010136           1.000000
TrainingTimesLastYear               NaN          0.011274          -0.035662
WorkLifeBalance                     NaN          0.004129           0.001008
YearsAtCompany                      NaN          0.015058           0.628133
YearsInCurrentRole                  NaN          0.050818           0.460365
YearsSinceLastPromotion             NaN          0.014352           0.404858
YearsWithCurrManager                NaN          0.024698           0.459188

                          TrainingTimesLastYear  WorkLifeBalance  \
Age                                   -0.019621        -0.021490
DailyRate                              0.002453        -0.037848
DistanceFromHome                      -0.036942        -0.026556
Education                             -0.025100         0.009819
EmployeeCount                               NaN              NaN
EmployeeNumber                         0.011953         0.005370
EnvironmentSatisfaction               -0.019359         0.027627
HourlyRate                            -0.008548        -0.004607
JobInvolvement                        -0.015338        -0.014617
JobLevel                              -0.018191         0.037818
JobSatisfaction                       -0.005779        -0.019459
MonthlyIncome                         -0.021736         0.030683
MonthlyRate                            0.001467         0.007963
NumCompaniesWorked                    -0.066054        -0.008366
```

```
PercentSalaryHike                    -0.005221         -0.003280
PerformanceRating                    -0.015579          0.002572
RelationshipSatisfaction              0.002497          0.019604
StandardHours                              NaN               NaN
StockOptionLevel                      0.011274          0.004129
TotalWorkingYears                    -0.035662          0.001008
TrainingTimesLastYear                 1.000000          0.028072
WorkLifeBalance                       0.028072          1.000000
YearsAtCompany                        0.003569          0.012089
YearsInCurrentRole                   -0.005738          0.049856
YearsSinceLastPromotion              -0.002067          0.008941
YearsWithCurrManager                 -0.004096          0.002759

                          YearsAtCompany  YearsInCurrentRole  \
Age                             0.311309            0.212901
DailyRate                      -0.034055            0.009932
DistanceFromHome                0.009508            0.018845
Education                       0.069114            0.060236
EmployeeCount                        NaN                 NaN
EmployeeNumber                 -0.005779           -0.004427
EnvironmentSatisfaction         0.001458            0.018007
HourlyRate                     -0.019582           -0.024106
JobInvolvement                 -0.021355            0.008717
JobLevel                        0.534739            0.389447
JobSatisfaction                -0.003803           -0.002305
MonthlyIncome                   0.514285            0.363818
MonthlyRate                    -0.023655           -0.012815
NumCompaniesWorked             -0.118421           -0.090754
PercentSalaryHike              -0.035991           -0.001520
PerformanceRating               0.003435            0.034986
RelationshipSatisfaction        0.019367           -0.015123
StandardHours                        NaN                 NaN
StockOptionLevel                0.015058            0.050818
TotalWorkingYears               0.628133            0.460365
TrainingTimesLastYear           0.003569           -0.005738
WorkLifeBalance                 0.012089            0.049856
YearsAtCompany                  1.000000            0.758754
YearsInCurrentRole              0.758754            1.000000
YearsSinceLastPromotion         0.618409            0.548056
YearsWithCurrManager            0.769212            0.714365

                          YearsSinceLastPromotion  YearsWithCurrManager
Age                                      0.216513              0.202089
DailyRate                               -0.033229             -0.026363
DistanceFromHome                         0.010029              0.014406
Education                                0.054254              0.069065
EmployeeCount                                 NaN                   NaN
```

```
EmployeeNumber                                 -0.004575             -0.004716
EnvironmentSatisfaction                         0.016194             -0.004999
HourlyRate                                     -0.026716             -0.020123
JobInvolvement                                 -0.024184              0.025976
JobLevel                                        0.353885              0.375281
JobSatisfaction                                -0.018214             -0.027656
MonthlyIncome                                   0.344978              0.344079
MonthlyRate                                     0.001567             -0.036746
NumCompaniesWorked                             -0.036814             -0.110319
PercentSalaryHike                              -0.022154             -0.011985
PerformanceRating                               0.017896              0.022827
RelationshipSatisfaction                        0.033493             -0.000867
StandardHours                                        NaN                   NaN
StockOptionLevel                                0.014352              0.024698
TotalWorkingYears                               0.404858              0.459188
TrainingTimesLastYear                          -0.002067             -0.004096
WorkLifeBalance                                 0.008941              0.002759
YearsAtCompany                                  0.618409              0.769212
YearsInCurrentRole                              0.548056              0.714365
YearsSinceLastPromotion                         1.000000              0.510224
YearsWithCurrManager                            0.510224              1.000000

[26 rows x 26 columns]
```

0.10 Model-1 : Logistic Regression Algorthim

[12]:
```
df.head()
```

[12]:
```
   Age Attrition     BusinessTravel  DailyRate              Department  \
0   41       Yes      Travel_Rarely       1102                   Sales
1   49        No  Travel_Frequently        279  Research & Development
2   37       Yes      Travel_Rarely       1373  Research & Development
3   33        No  Travel_Frequently       1392  Research & Development
4   27        No      Travel_Rarely        591  Research & Development

   DistanceFromHome  Education EducationField  EmployeeCount  EmployeeNumber  \
0                 1          2  Life Sciences              1               1
1                 8          1  Life Sciences              1               2
2                 2          2          Other              1               3
3                 3          4  Life Sciences              1               4
4                 2          1        Medical              1               5

   …  RelationshipSatisfaction StandardHours  StockOptionLevel  \
0  …                         1            80                 0
1  …                         4            80                 1
2  …                         2            80                 0
3  …                         3            80                 0
```

```
4  …                          4          80                  1

   TotalWorkingYears  TrainingTimesLastYear WorkLifeBalance  YearsAtCompany  \
0                  8                      0               1               6
1                 10                      3               3              10
2                  7                      3               3               0
3                  8                      3               3               8
4                  6                      3               3               2

  YearsInCurrentRole  YearsSinceLastPromotion  YearsWithCurrManager
0                  4                        0                     5
1                  7                        1                     7
2                  0                        0                     0
3                  7                        3                     0
4                  2                        2                     2

[5 rows x 35 columns]
```

[13]:
```
df['Gender'].replace(['F'],'Female', inplace = True)
df['MaritalStatus'].replace(['M'],'Married', inplace = True)
```

[14]:
```
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()

df['Department'] = le.fit_transform(df['Department'])
df['EducationField'] = le.fit_transform(df['EducationField'])
df['Gender'] = le.fit_transform(df['Gender'])
df['MaritalStatus'] = le.fit_transform(df['MaritalStatus'])
```

0.11 Data(Train-Test) Split

[15]:
```
x=df.drop(['Gender','Attrition','JobRole','Over18',␣
 ↪'OverTime','BusinessTravel'], saxis=1)
y=df[['Gender']]
```

[16]:
```
x.head(2)
```

[16]:
```
   Age  DailyRate  Department  DistanceFromHome  Education  EducationField  \
0   41       1102           2                 1          2               1
1   49        279           1                 8          1               1

   EmployeeCount  EmployeeNumber  EnvironmentSatisfaction  HourlyRate  …  \
0              1               1                        2          94  …
1              1               2                        3          61  …

   RelationshipSatisfaction  StandardHours  StockOptionLevel  \
```

```
0                          1             80                  0
1                          4             80                  1

   TotalWorkingYears  TrainingTimesLastYear  WorkLifeBalance  YearsAtCompany  \
0                  8                      0                1               6
1                 10                      3                3              10

   YearsInCurrentRole  YearsSinceLastPromotion  YearsWithCurrManager
0                   4                        0                     5
1                   7                        1                     7

[2 rows x 29 columns]
```

[17]:
```
x.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2940 entries, 0 to 2939
Data columns (total 29 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Age                       2940 non-null   int64
 1   DailyRate                 2940 non-null   int64
 2   Department                2940 non-null   int32
 3   DistanceFromHome          2940 non-null   int64
 4   Education                 2940 non-null   int64
 5   EducationField            2940 non-null   int32
 6   EmployeeCount             2940 non-null   int64
 7   EmployeeNumber            2940 non-null   int64
 8   EnvironmentSatisfaction   2940 non-null   int64
 9   HourlyRate                2940 non-null   int64
 10  JobInvolvement            2940 non-null   int64
 11  JobLevel                  2940 non-null   int64
 12  JobSatisfaction           2940 non-null   int64
 13  MaritalStatus             2940 non-null   int32
 14  MonthlyIncome             2940 non-null   int64
 15  MonthlyRate               2940 non-null   int64
 16  NumCompaniesWorked        2940 non-null   int64
 17  PercentSalaryHike         2940 non-null   int64
 18  PerformanceRating         2940 non-null   int64
 19  RelationshipSatisfaction  2940 non-null   int64
 20  StandardHours             2940 non-null   int64
 21  StockOptionLevel          2940 non-null   int64
 22  TotalWorkingYears         2940 non-null   int64
 23  TrainingTimesLastYear     2940 non-null   int64
 24  WorkLifeBalance           2940 non-null   int64
 25  YearsAtCompany            2940 non-null   int64
 26  YearsInCurrentRole        2940 non-null   int64
```

```
 27  YearsSinceLastPromotion   2940 non-null   int64
 28  YearsWithCurrManager      2940 non-null   int64
dtypes: int32(3), int64(26)
memory usage: 631.8 KB
```

[18]:
```
y.head(2)
```

[18]:

	Gender
0	0
1	1

[19]:
```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2,␣
 ↪random_state=0)
```

[20]:
```
x_train.shape , x_test.shape
```

[20]: ((2352, 29), (588, 29))

[21]:
```
y_train.shape , y_test.shape
```

[21]: ((2352, 1), (588, 1))

0.12 Logistic Regression Method

[22]:
```
from sklearn.linear_model import LogisticRegression
logit = LogisticRegression(random_state= 100)
logit.fit(x_train, y_train)
```

[22]: LogisticRegression(random_state=100)

0.13 Prediction

[23]:
```
y_pred_train_log = logit.predict(x_train)
y_pred_test_log = logit.predict(x_test)
```

0.14 Evaluate test data Accuracy

[24]:
```
from sklearn.metrics import confusion_matrix,classification_report,␣
 ↪accuracy_score
accuracy_log_test=accuracy_score(y_test,y_pred_test_log)
print('Logistic regression Test accuracy:', accuracy_score(y_test,␣
 ↪y_pred_test_log))
```

```
Logistic regression Test accuracy: 0.6071428571428571
```

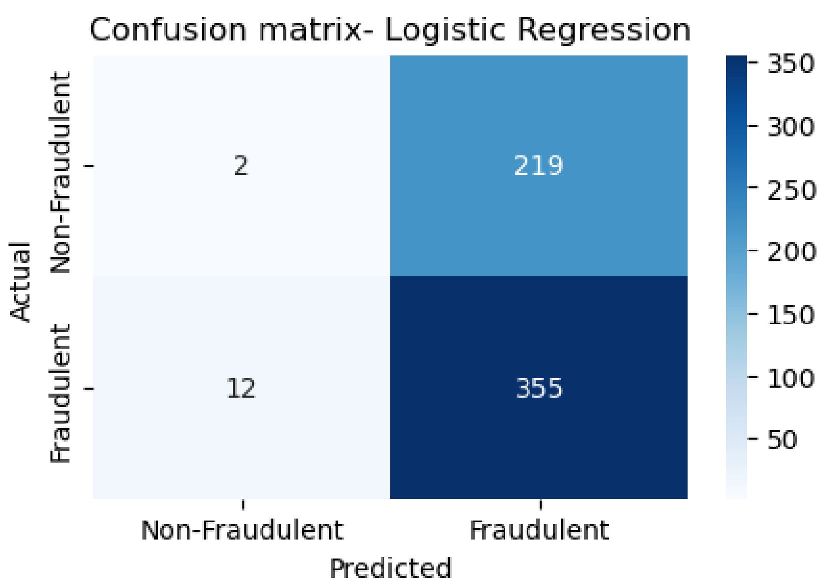
0.15 Evaluate train data Accuracy

[25]:
```
from sklearn.metrics import confusion_matrix,classification_report,␣
↪accuracy_score
accuracy_train = accuracy_score(y_train,y_pred_train_log)

print('Logistic regression Train accuracy:', accuracy_score(y_train,␣
↪y_pred_train_log))
```

```
Logistic regression Train accuracy: 0.594812925170068
```

0.16 Confusion Martrix - Logistic Regression

[256]:
```
Labels = ['Non-Fraudulent', 'Fraudulent']
plt.figure(figsize=(5,3))
sns.heatmap(confusion_matrix(y_test,y_pred_test_log),xticklabels=Labels,
            yticklabels=Labels,cmap='Blues',annot=True, fmt='g')
plt.title("Confusion matrix- Logistic Regression")
plt.ylabel('Actual')
plt.xlabel('Predicted')
plt.show()
```



0.17 AUC (Area under the curve) & ROC (Receiver operating characteristics)

- It is one of the most important evaluation metrics for checking classification model's performance.
- It is also written as AUROC (Area Under the Receiver Operating Characteristics)
- ROC is a probability curve and AUC represents the degree or measure of separability.

- It tells how much the model is capable of distinguishing between classes.

[26]:
```
from sklearn.metrics import roc_auc_score
logit_roc_auc = roc_auc_score(y_test, y_pred_test_log)
print(logit_roc_auc)
```

```
0.48817611303586617
```

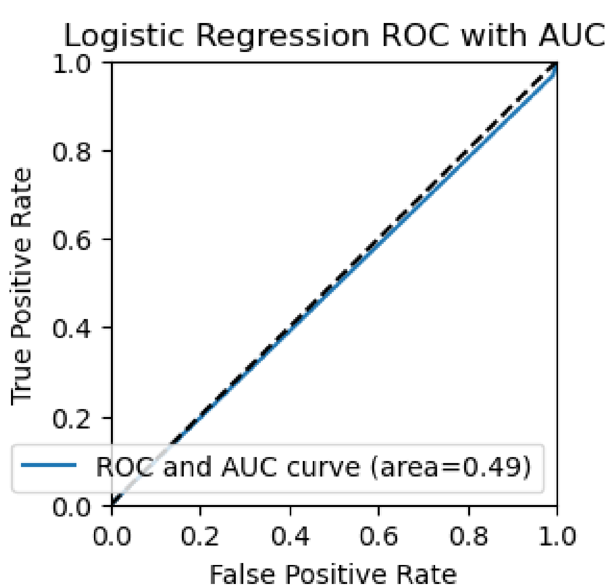
[27]:
```
from sklearn.metrics import roc_curve

fpr, tpr, thresholds = roc_curve(y_test, y_pred_test_log)
display(fpr[:10])
display(tpr[:10])
display(thresholds[:10])
```

```
array([0.        , 0.99095023, 1.        ])
```

```
array([0.        , 0.96730245, 1.        ])
```

```
array([2, 1, 0])
```

[28]:
```
plt.figure(figsize=(3,3))
plt.plot(fpr, tpr, label="ROC and AUC curve (area=%0.2f)" % logit_roc_auc)
plt.plot([0,1],[0,1], 'k--')
plt.xlim([0.0,1.0])
plt.ylim([0.0,1.0])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title("Logistic Regression ROC with AUC")
plt.legend(loc='lower right')
plt.show()
```



0.18 Model-2 : Random Forest Algorithm

0.18.1 Feature Scaling

[29]:
```
x.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2940 entries, 0 to 2939
Data columns (total 29 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Age                       2940 non-null   int64
 1   DailyRate                 2940 non-null   int64
 2   Department                2940 non-null   int32
 3   DistanceFromHome          2940 non-null   int64
 4   Education                 2940 non-null   int64
 5   EducationField            2940 non-null   int32
 6   EmployeeCount             2940 non-null   int64
 7   EmployeeNumber            2940 non-null   int64
 8   EnvironmentSatisfaction   2940 non-null   int64
 9   HourlyRate                2940 non-null   int64
 10  JobInvolvement            2940 non-null   int64
 11  JobLevel                  2940 non-null   int64
 12  JobSatisfaction           2940 non-null   int64
 13  MaritalStatus             2940 non-null   int32
 14  MonthlyIncome             2940 non-null   int64
 15  MonthlyRate               2940 non-null   int64
 16  NumCompaniesWorked        2940 non-null   int64
 17  PercentSalaryHike         2940 non-null   int64
 18  PerformanceRating         2940 non-null   int64
 19  RelationshipSatisfaction  2940 non-null   int64
 20  StandardHours             2940 non-null   int64
 21  StockOptionLevel          2940 non-null   int64
 22  TotalWorkingYears         2940 non-null   int64
 23  TrainingTimesLastYear     2940 non-null   int64
 24  WorkLifeBalance           2940 non-null   int64
 25  YearsAtCompany            2940 non-null   int64
 26  YearsInCurrentRole        2940 non-null   int64
 27  YearsSinceLastPromotion   2940 non-null   int64
 28  YearsWithCurrManager      2940 non-null   int64
dtypes: int32(3), int64(26)
memory usage: 631.8 KB
```

[30]:
```
from sklearn.preprocessing import StandardScaler
sc=StandardScaler()
```

```
x1=sc.fit_transform(x)
pd.DataFrame(x1).head(2)
```

[30]:
```
          0         1         2         3         4         5    6         7  \
0  0.446350  0.742527  1.401512 -1.010909 -0.891688 -0.937414  0.0 -1.731462
1  1.322365 -1.297775 -0.493817 -0.147150 -1.868426 -0.937414  0.0 -1.730284

          8         9  …        19   20        21        22        23  \
0 -0.660531  1.383138  … -1.584178  0.0 -0.932014 -0.421642 -2.171982
1  0.254625 -0.240677  …  1.191438  0.0  0.241988 -0.164511  0.155707

         24        25        26        27        28
0 -2.493820 -0.164613 -0.063296 -0.679146  0.245834
1  0.338096  0.488508  0.764998 -0.368715  0.806541

[2 rows x 29 columns]
```

0.18.2 Check Balance Data

[31]:
```
y.value_counts()
```

[31]:
```
Gender
1         1764
0         1176
dtype: int64
```

0.18.3 Conculsion - Data is Imbalaced

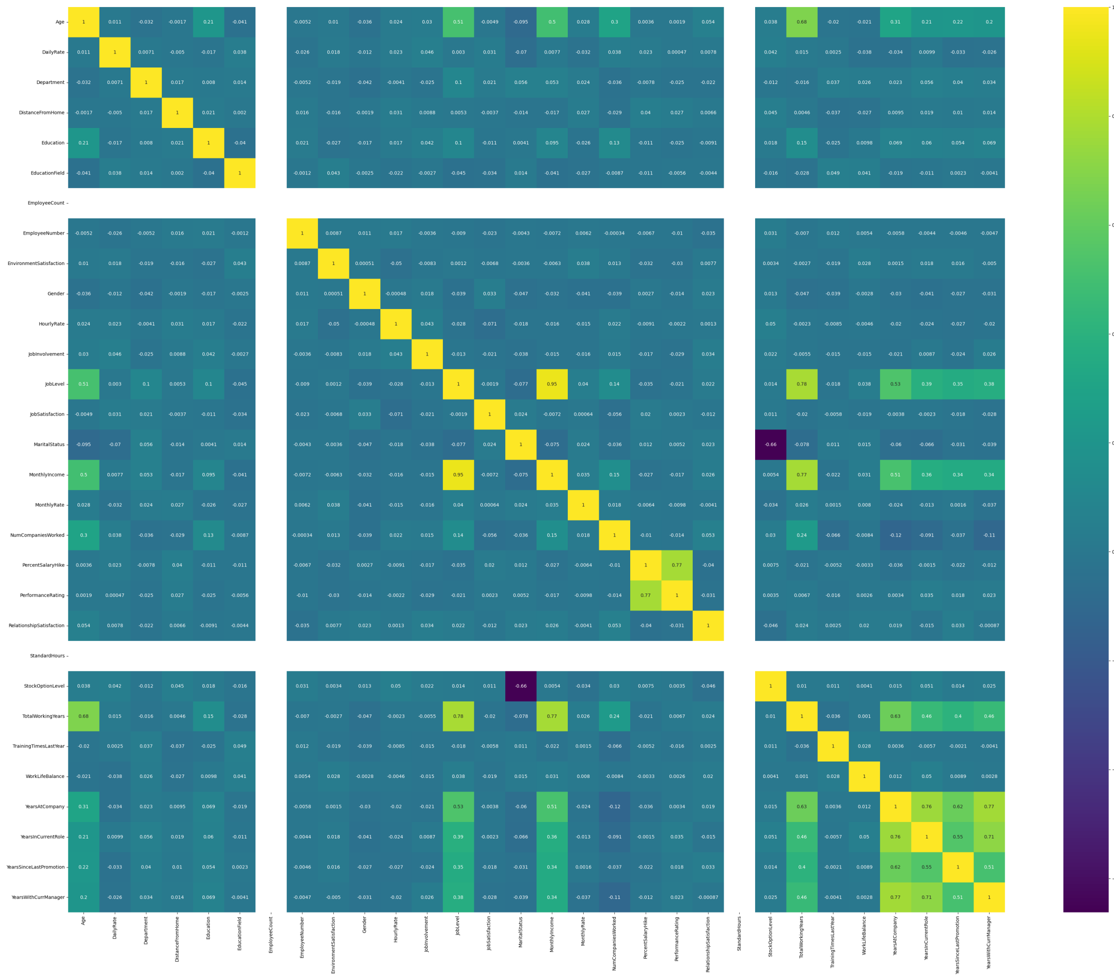
[33]:
```
# Done Under Sampling to balaced the data
import imblearn
from imblearn.under_sampling import RandomUnderSampler
ros = RandomUnderSampler()
x_un,y_un = ros.fit_resample(x1,y)
print(x_un.shape,y_un.shape,y.shape)
```

```
(2352, 29) (2352, 1) (2940, 1)
```

[34]:
```
y_un.value_counts()
```

[34]:
```
Gender
0         1176
1         1176
dtype: int64
```

[41]:
```
plt.figure(figsize = (45, 35))
sns.heatmap(df.corr(), annot = True, cmap = 'viridis')
plt.show()
```



0.19 Model Building

[43]:
```
from sklearn.ensemble import RandomForestClassifier
rf = RandomForestClassifier(n_estimators = 200, oob_score = False)
rf.fit(x_train,y_train)
```

[43]:
```
RandomForestClassifier(n_estimators=200)
```

0.20 Prediction

[44]:
```
y_pred_train_rf = rf.predict(x_train)
y_pred_test_rf = rf.predict(x_test)
```

0.21 Evaluate

[45]:
```
accuracy_rf_test = accuracy_score(y_test,y_pred_test_rf)
accuracy_rf_train = accuracy_score(y_train,y_pred_train_rf)
print('Random Forest - Train accuracy:', accuracy_score(y_train,␣
 ↪y_pred_train_rf))
print('-------'*10)
print('Random Forest - Test accuracy:', accuracy_score(y_test, y_pred_test_rf))
```

```
Random Forest - Train accuracy: 1.0
----------------------------------------------------------------------
Random Forest - Test accuracy: 0.9149659863945578
```

0.22 Confusion Matrix

[46]:
```
print(confusion_matrix(y_test,y_pred_test_rf))
```

```
[[185  36]
 [ 14 353]]
```

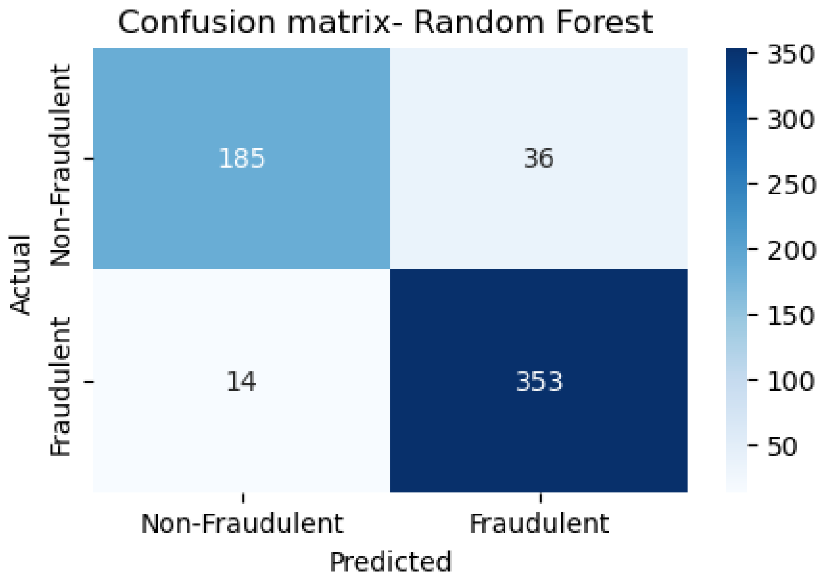
[47]:
```
print(confusion_matrix(y_train,y_pred_train_rf))
```

```
[[ 955    0]
 [   0 1397]]
```

[53]:
```
print('Random Forest Train data accuracy')
acc = accuracy_score (y_train, y_pred_train_rf)
print('Accuracy score is', acc)
```

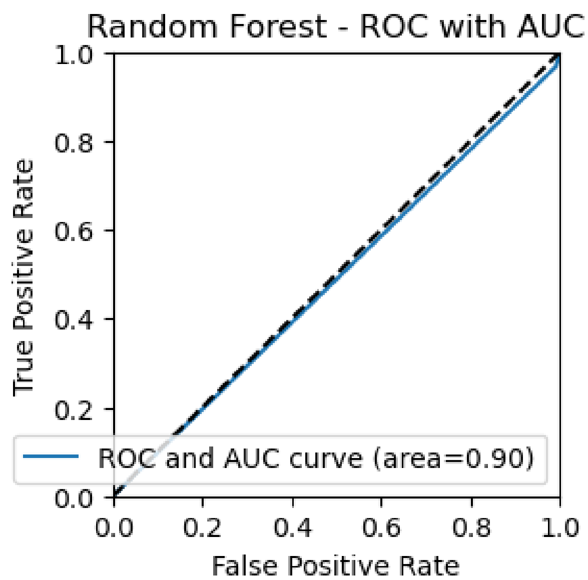
```
Random Forest Train data accuracy
Accuracy score is 1.0
```

[55]:
```
Labels = ['Non-Fraudulent', 'Fraudulent']
plt.figure(figsize = (5,3))
sns.heatmap(confusion_matrix(y_test,y_pred_test_rf), xticklabels = Labels,
            yticklabels = Labels, cmap = 'Blues', annot = True, fmt = 'g')
plt.title("Confusion matrix- Random Forest ")
plt.ylabel('Actual')
plt.xlabel('Predicted')
plt.show()
```



[56]:
```
rf_roc_auc = roc_auc_score(y_test, y_pred_test_rf)
print(rf_roc_auc)
plt.figure(figsize = (3,3))
plt.plot(fpr, tpr, label = "ROC and AUC curve (area=%0.2f)" % rf_roc_auc)
plt.plot([0,1],[0,1], 'k--')
plt.xlim([0.0,1.0])
plt.ylim([0.0,1.0])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title("Random Forest - ROC with AUC")
plt.legend(loc = 'lower right')
plt.show()
```

```
0.8994784667168062
```



0.22.1 Cross validation because of underfitting issue

[57]:
```
from sklearn.model_selection import cross_val_score
train_accuracy_rf = cross_val_score(rf, x_train, y_train, cv = 10)
crossval_train_rf = train_accuracy_rf.mean()
test_accuracy_rf = cross_val_score(rf, x_test, y_test, cv = 10)
crossval_test_rf = test_accuracy_rf.mean()

print('Random forest after Cross validation Train accuracy:', crossval_train_rf)
print('-------'*10)
print('Random forest  after Cross validation Test accuracy:', crossval_test_rf)
```

```
Random forest after Cross validation Train accuracy: 0.8877623512441399
----------------------------------------------------------------------
Random forest  after Cross validation Test accuracy: 0.6545587375803623
```

0.23 Model-3 : Decision Tree

- A decision tree uses the tree representation to solve the problem in which each leaf node corresponds to a class label and attributes are represented on the internal node of the tree.

0.24 Model building

[58]:
```
from sklearn.tree import DecisionTreeClassifier,plot_tree
dtree = DecisionTreeClassifier()
dtree.fit(x_train,y_train)
```

[58]: DecisionTreeClassifier()

0.25 Prediction

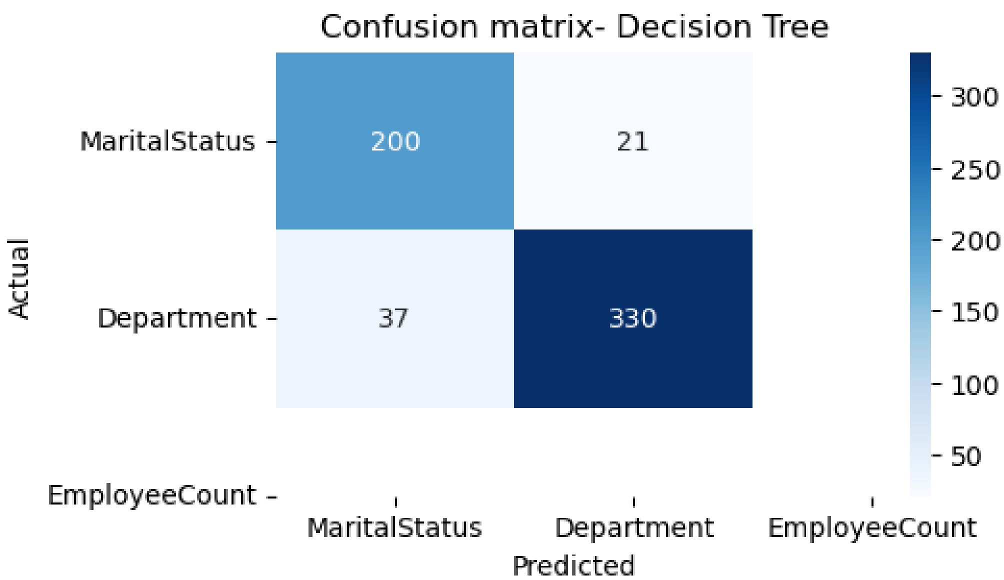
[60]:
```
y_pred_train_dtree = dtree.predict(x_train)
y_pred_test_dtree = dtree.predict(x_test)
```

0.26 Evaluate

[61]:
```
accuracy_dtree_test = accuracy_score(y_test,y_pred_test_dtree)
accuracy_dtree_train = accuracy_score(y_train,y_pred_train_dtree)

print('Decision Tree - Train accuracy:', accuracy_score(y_train,␣
 ↪y_pred_train_dtree))
print('-------'*10)
print('Decision Tree - Test accuracy:', accuracy_score(y_test,␣
 ↪y_pred_test_dtree))
```

```
Decision Tree - Train accuracy: 1.0
----------------------------------------------------------------------
Decision Tree - Test accuracy: 0.9013605442176871
```

[63]:
```
Labels = ['MaritalStatus','Department','EmployeeCount']
plt.figure(figsize = (5,3))
sns.heatmap(confusion_matrix(y_test,y_pred_test_dtree),xticklabels = Labels,
            yticklabels = Labels, cmap = 'Blues', annot = True, fmt = 'g')
plt.title("Confusion matrix- Decision Tree")
plt.ylabel('Actual')
plt.xlabel('Predicted')
plt.show()
```



0.26.1 Using Post prunning method to handle overfitting problem

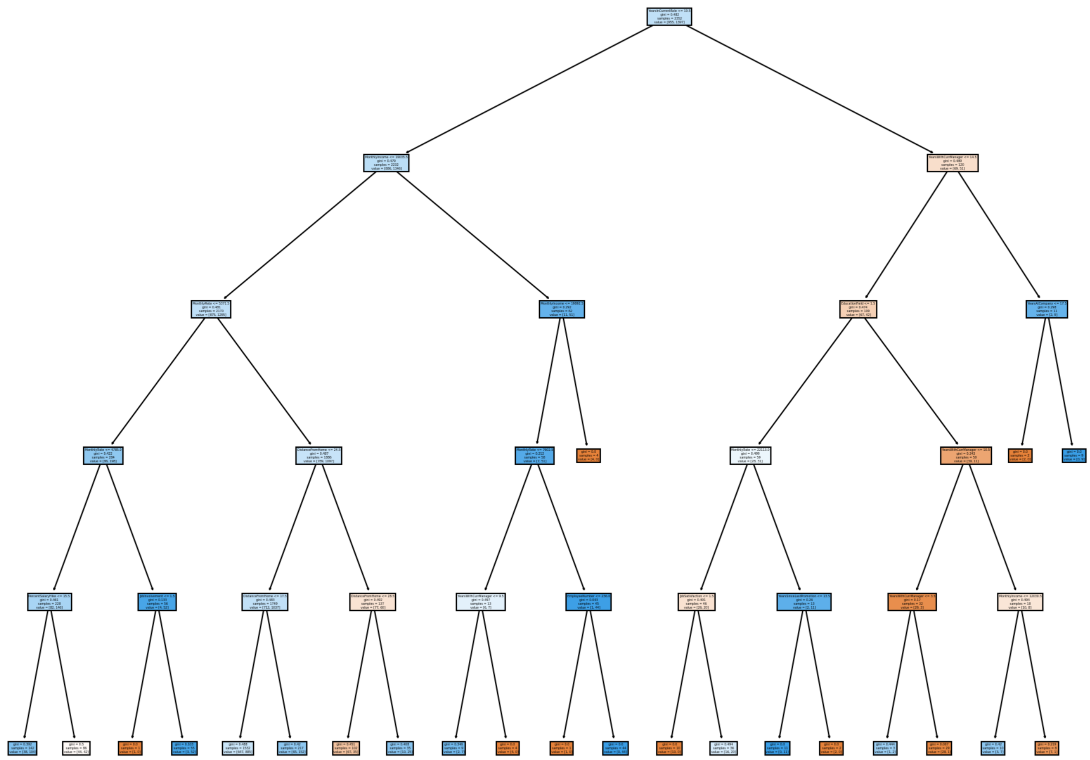
[66]:
```python
def dtree_model(model):
    model_preds = model.predict(x_test)
    print(classification_report(y_test,model_preds))
    print('\n')
    plt.figure(figsize = (15,12), dpi = 150)
    plot_tree(model, filled = True, feature_names = x.columns)
plt.show()
```

[67]:
```python
# max depth at 5
prunned_dtree = DecisionTreeClassifier(max_depth = 5)
prunned_dtree.fit(x_train,y_train)
```

[67]: DecisionTreeClassifier(max_depth=5)

[68]:
```python
dtree_model(prunned_dtree)
```

```
              precision    recall  f1-score   support

           0       0.55      0.16      0.25       221
           1       0.65      0.92      0.76       367

    accuracy                           0.63       588
   macro avg       0.60      0.54      0.50       588
weighted avg       0.61      0.63      0.57       588
```



0.26.2 Prediction

[73]:
```
y_pred_prunned_train = prunned_dtree.predict(x_train)
y_pred_prunned_test = prunned_dtree.predict(x_test)
```

0.26.3 Evaluate

[70]:
```
print('Decision Tree post prunning- Train accuracy:
↪',accuracy_score(y_train,y_pred_prunned_train))
print('-------'*10)
print('Decision Tree post prunning- Test accuracy:',␣
↪accuracy_score(y_test,y_pred_prunned_test))
```

```
Decision Tree post prunning- Train accuracy: 0.6326530612244898
----------------------------------------------------------------------
Decision Tree post prunning- Test accuracy: 0.6343537414965986
```

[ ]:

```

```