IBM HR Analytics Employee Attrition and Performance

October 5, 2021

This notebook is a python based machine learning study of the given data to look into factors which lead the attrition of emplyees int the company

1 Problem Definition

The problem is to find whether or not an employee will leave the company

2 Evaluation Matrices

We are going to use accuracy, precision, recall and F1-score to check the validity of the model prediction.

3 Data

There are 35 attributes in the data and the data set is available at Kaggle competion

3.1 Data Dictionary: -

There are different 35 columns in the dataset which are self explanatory from their names

1)Age - Age of the person
2) Attirtion - Target variable
3) BusinessTravel - The employee travel rarely, frequently or there are employees who never travel
4) DailyRate -
5) Department - There are three departments-Sales, Human resource and Research and Development
6) DistanceFromHome - How far the employees living from the office
7) Education - There are four levels of education 1 to 4
8) EducationField - Field of education of the employee. There are six field of education in the data
9) EmployeeCount - 1 for all —— can be eliminated from the table
10) EmployeeNumber - Just an emplyee number —– can be eliminated from the table
11) EnvironmentSatisfaction - How much the employee satisfied with the environment range from 1 to 5
12) Gender - Male or Female
13) HourlyRate -
14) JobInvolvement - How much the employee involve in the job (1 to 4)
15) JobLevel - Job levels of the employee labelled from 1 to 5

16) JobRole - Position or Job title
17) JobSatisfaction - Level of emplyee with the job satisfaction from 1 to 4
18) MaritalStatus - married, single or divorced
19) MonthlyIncome - Monthly earning of the employee
20) MonthlyRate -
21) NumCompaniesWorked - The number of companies an employee worked
22) Over18 -Does the emplyees age above 18 or not. All employees are above 18 —can be eliminated
23) OverTime - Weather an empoyee is working overtime or not
24) PercentSalaryHike -Raise in the salary of emplyee
25) PerformanceRating -Performace of the worker range from 1 to 4 but the data contain only 3 and 4
26) RelationshipSatisfaction - How much an employee satisfied with colleagues realtions range from 1 to 4
27) StandardHours - it is 80 for all ——— can be eliminated
28) StockOptionLevel -Whether or not the the employee has stock option
29) TotalWorkingYears - Total working years of an employee
30) TrainingTimesLastYear - How many times the employee got training in the last year
31) WorkLifeBalance -The level of work life balance range from 1 to 4
32) YearsAtCompany - Years of the employee working in the company
33) YearsInCurrentRole - Totoal years of an employee at current role
34) YearsSinceLastPromotion -Years since last promotion of an employee
35) YearsWithCurrManager -How long the an employee working with the current manager

3.2 Data Processing- Before Data Analysis

3.2.1 Data Mapping - There is no need of data mapping becuase all the data are from one souce and in the form of an excel sheet

3.2.2 Data Cleaning -

[6]:
```
 # Import Libraries
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

[7]:
```
# load data from excel file
employee_df=pd.read_csv("WA_Fn-UseC_-HR-Employee-Attrition.csv")
```

[8]:
```
employee_df
```

[8]:
```
      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
…    …        …                 …        …                       …
1465   36         No  Travel_Frequently        884  Research & Development
1466   39         No      Travel_Rarely        613  Research & Development
1467   27         No      Travel_Rarely        155  Research & Development
1468   49         No  Travel_Frequently       1023                   Sales
1469   34         No      Travel_Rarely        628  Research & Development

      DistanceFromHome  Education EducationField  EmployeeCount  \
0                    1          2  Life Sciences              1
1                    8          1  Life Sciences              1
2                    2          2          Other              1
3                    3          4  Life Sciences              1
4                    2          1        Medical              1
…                  …          …              …             …
1465                23          2        Medical              1
1466                 6          1        Medical              1
1467                 4          3  Life Sciences              1
1468                 2          3        Medical              1
1469                 8          3        Medical              1

      EmployeeNumber  …  RelationshipSatisfaction StandardHours  \
0                  1  …                         1            80
1                  2  …                         4            80
2                  4  …                         2            80
3                  5  …                         3            80
4                  7  …                         4            80
…                …  …                       …           …
1465            2061  …                         3            80
1466            2062  …                         1            80
1467            2064  …                         2            80
1468            2065  …                         4            80
1469            2068  …                         1            80

      StockOptionLevel  TotalWorkingYears  TrainingTimesLastYear  \
0                    0                  8                      0
1                    1                 10                      3
2                    0                  7                      3
3                    0                  8                      3
4                    1                  6                      3
…                  …                …                    …
1465                 1                 17                      3
1466                 1                  9                      5
1467                 1                  6                      0
1468                 0                 17                      3
1469                 0                  6                      3
```

```
      WorkLifeBalance  YearsAtCompany YearsInCurrentRole  \
0                   1               6                  4
1                   3              10                  7
2                   3               0                  0
3                   3               8                  7
4                   3               2                  2
…                 …               …                …
1465                3               5                  2
1466                3               7                  7
1467                3               6                  2
1468                2               9                  6
1469                4               4                  3

      YearsSinceLastPromotion  YearsWithCurrManager
0                           0                     5
1                           1                     7
2                           0                     0
3                           3                     0
4                           2                     2
…                         …                   …
1465                        0                     3
1466                        1                     7
1467                        0                     3
1468                        0                     8
1469                        1                     2

[1470 rows x 35 columns]
```

[9]:
```python
employee_df.head()
```

[9]:
```
   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               4
3                 3          4  Life Sciences              1               5
4                 2          1        Medical              1               7

   …  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]
```

[10]:
```python
employee_df.describe()
```

[10]:
```
               Age    DailyRate  DistanceFromHome    Education  EmployeeCount  \
count  1470.000000  1470.000000       1470.000000  1470.000000         1470.0
mean     36.923810   802.485714          9.192517     2.912925            1.0
std       9.135373   403.509100          8.106864     1.024165            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     1470.000000              1470.000000  1470.000000     1470.000000
mean      1024.865306                 2.721769    65.891156        2.729932
std        602.024335                 1.093082    20.329428        0.711561
min          1.000000                 1.000000    30.000000        1.000000
25%        491.250000                 2.000000    48.000000        2.000000
50%       1020.500000                 3.000000    66.000000        3.000000
75%       1555.750000                 4.000000    83.750000        3.000000
max       2068.000000                 4.000000   100.000000        4.000000

          JobLevel  ...  RelationshipSatisfaction  StandardHours  \
count  1470.000000  ...               1470.000000         1470.0
mean      2.063946  ...                  2.712245           80.0
std       1.106940  ...                  1.081209            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       1470.000000        1470.000000            1470.000000
mean           0.793878          11.279592               2.799320
std            0.852077           7.780782               1.289271
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      1470.000000     1470.000000         1470.000000
mean          2.761224        7.008163            4.229252
std           0.706476        6.126525            3.623137
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              1470.000000           1470.000000
mean                  2.187755              4.123129
std                   3.222430              3.568136
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]
```

[11]:
```
employee_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1470 entries, 0 to 1469
Data columns (total 35 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Age                       1470 non-null   int64
 1   Attrition                 1470 non-null   object
 2   BusinessTravel            1470 non-null   object
 3   DailyRate                 1470 non-null   int64
```

```
 4   Department                1470 non-null   object
 5   DistanceFromHome          1470 non-null   int64
 6   Education                 1470 non-null   int64
 7   EducationField            1470 non-null   object
 8   EmployeeCount             1470 non-null   int64
 9   EmployeeNumber            1470 non-null   int64
 10  EnvironmentSatisfaction   1470 non-null   int64
 11  Gender                    1470 non-null   object
 12  HourlyRate                1470 non-null   int64
 13  JobInvolvement            1470 non-null   int64
 14  JobLevel                  1470 non-null   int64
 15  JobRole                   1470 non-null   object
 16  JobSatisfaction           1470 non-null   int64
 17  MaritalStatus             1470 non-null   object
 18  MonthlyIncome             1470 non-null   int64
 19  MonthlyRate               1470 non-null   int64
 20  NumCompaniesWorked        1470 non-null   int64
 21  Over18                    1470 non-null   object
 22  OverTime                  1470 non-null   object
 23  PercentSalaryHike         1470 non-null   int64
 24  PerformanceRating         1470 non-null   int64
 25  RelationshipSatisfaction  1470 non-null   int64
 26  StandardHours             1470 non-null   int64
 27  StockOptionLevel          1470 non-null   int64
 28  TotalWorkingYears         1470 non-null   int64
 29  TrainingTimesLastYear     1470 non-null   int64
 30  WorkLifeBalance           1470 non-null   int64
 31  YearsAtCompany            1470 non-null   int64
 32  YearsInCurrentRole        1470 non-null   int64
 33  YearsSinceLastPromotion   1470 non-null   int64
 34  YearsWithCurrManager      1470 non-null   int64
dtypes: int64(26), object(9)
memory usage: 402.1+ KB
```

[12]:
```
employee_df.isna().sum()
```

[12]:
```
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
```

[13]:

```
fig,ax=plt.subplots(figsize=(10,5))
sns.heatmap(employee_df.isnull(), yticklabels=False, cbar=False, cmap='Greens');
txt="Figure 1: Plot of 'na' counts for field"
plt.figtext(0.5, -0.4, txt, wrap=True, horizontalalignment='center',␣
 ↪fontsize=14);
```

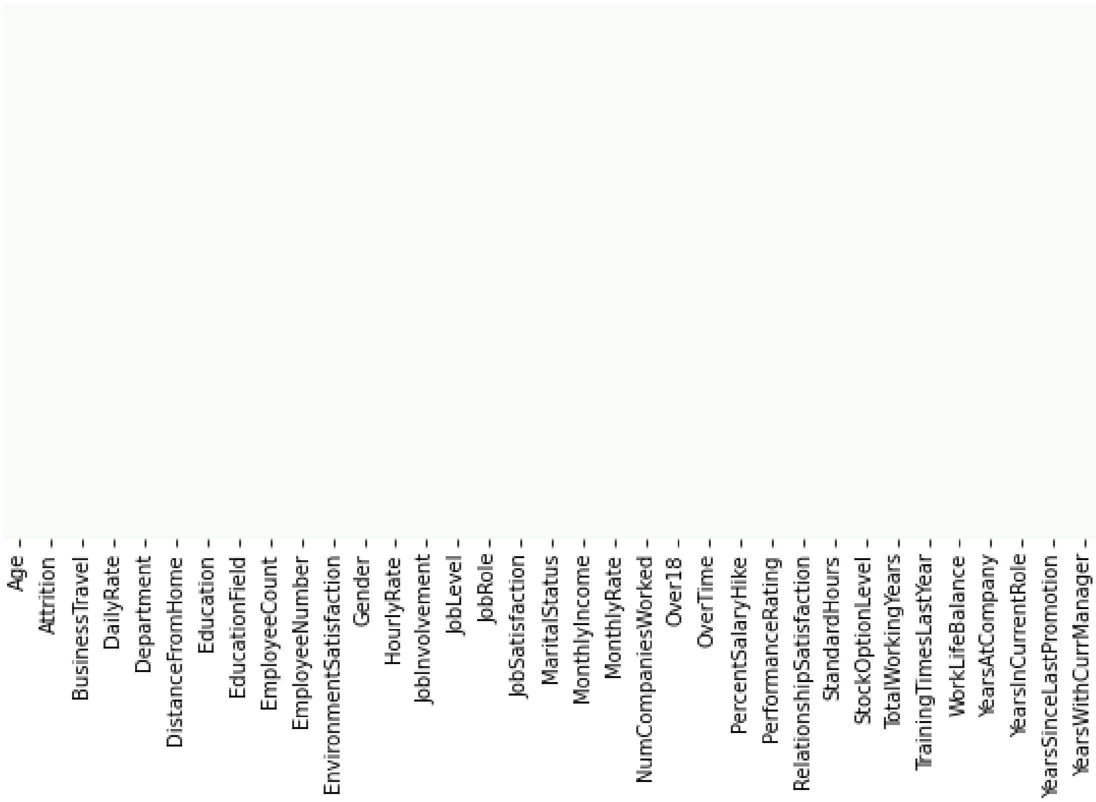


Figure 1: Plot of 'na' counts for field

There is no null value

[14]:
```
employee_df["EmployeeCount"]
```

[14]:
```
0       1
1       1
2       1
3       1
4       1
       ..
1465    1
1466    1
1467    1
1468    1
1469    1
Name: EmployeeCount, Length: 1470, dtype: int64
```

[15]:
```
employee_df["EmployeeCount"].value_counts()
```

[15]:
```
1    1470
Name: EmployeeCount, dtype: int64
```

[16]:
```
employee_df["Over18"].value_counts()
```

[16]:
```
Y    1470
Name: Over18, dtype: int64
```

[17]:
```
employee_df["StandardHours"].value_counts()
```

[17]:
```
80    1470
Name: StandardHours, dtype: int64
```

[18]:
```
employee_df1=employee_df.
 ↪drop(['EmployeeCount','Over18','StandardHours','EmployeeNumber'], axis=1)
employee_df1
```

[18]:
```
      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
…    …         …                …       …                     …
1465   36        No  Travel_Frequently        884  Research & Development
1466   39        No      Travel_Rarely        613  Research & Development
1467   27        No      Travel_Rarely        155  Research & Development
1468   49        No  Travel_Frequently       1023                   Sales
1469   34        No      Travel_Rarely        628  Research & Development

      DistanceFromHome  Education EducationField  EnvironmentSatisfaction  \
0                    1          2  Life Sciences                        2
1                    8          1  Life Sciences                        3
2                    2          2          Other                        4
3                    3          4  Life Sciences                        4
4                    2          1        Medical                        1
…                  …         …             …                       …
1465                23          2        Medical                        3
1466                 6          1        Medical                        4
1467                 4          3  Life Sciences                        2
1468                 2          3        Medical                        4
1469                 8          3        Medical                        2

      Gender  …  PerformanceRating  RelationshipSatisfaction  \
0     Female  …                  3                         1
1       Male  …                  4                         4
2       Male  …                  3                         2
```

```
3     Female  …                  3                         3
4       Male  …                  3                         4
…        …  …                 …                       …
1465    Male  …                  3                         3
1466    Male  …                  3                         1
1467    Male  …                  4                         2
1468    Male  …                  3                         4
1469    Male  …                  3                         1

      StockOptionLevel TotalWorkingYears  TrainingTimesLastYear  \
0                    0                 8                      0
1                    1                10                      3
2                    0                 7                      3
3                    0                 8                      3
4                    1                 6                      3
…                  …               …                    …
1465                 1                17                      3
1466                 1                 9                      5
1467                 1                 6                      0
1468                 0                17                      3
1469                 0                 6                      3

     WorkLifeBalance  YearsAtCompany  YearsInCurrentRole  \
0                  1               6                   4
1                  3              10                   7
2                  3               0                   0
3                  3               8                   7
4                  3               2                   2
…                …             …                 …
1465               3               5                   2
1466               3               7                   7
1467               3               6                   2
1468               2               9                   6
1469               4               4                   3

      YearsSinceLastPromotion YearsWithCurrManager
0                           0                    5
1                           1                    7
2                           0                    0
3                           3                    0
4                           2                    2
…                         …                  …
1465                        0                    3
1466                        1                    7
1467                        0                    3
1468                        0                    8
1469                        1                    2
```

```
[1470 rows x 31 columns]
```

We have dropped four columns and there is no null value in the table. So we can say that the data cleaning looks ok now and we can move to the next step of Data analysis.

3.3 Data Analsis -

```
[19]: employee_df1.hist(bins=10, figsize=(20,80),layout=(13,3), xlabelsize=15,␣
       ↪ylabelsize=15, color='g',legend=15);
      txt="Figure 2: Histogram for different entries"
      plt.figtext(0.5, 0.4, txt, wrap=True, horizontalalignment='center',␣
       ↪fontsize=14);
```

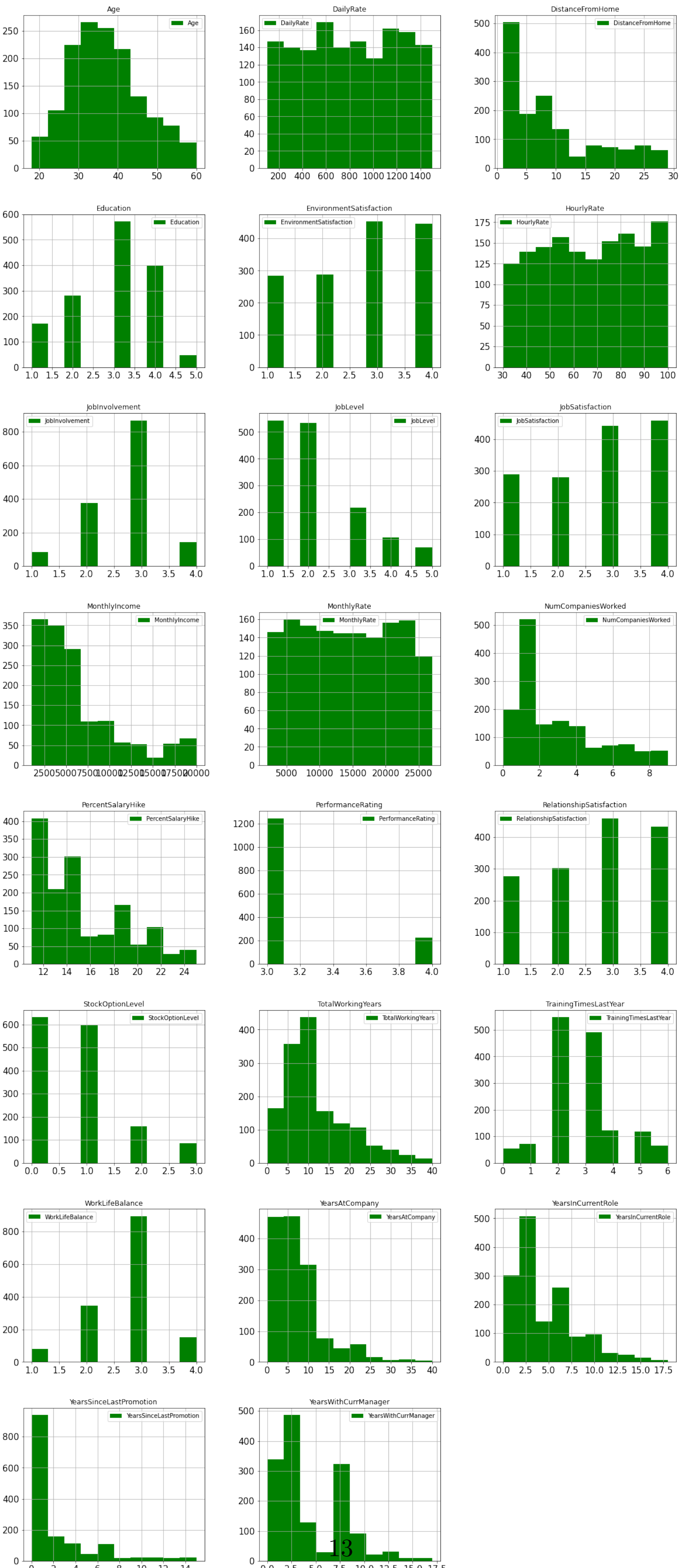


Figure 2: Histogram for different entries

??

[20]:
```
employee_df1["Attrition"]=employee_df1["Attrition"].apply(lambda x:1 if␣
↪x=="Yes" else 0)
```

[21]:
```
correlations=employee_df1.corr()
f,ax=plt.subplots(figsize=(20,20))
sns.heatmap(correlations,annot=True)
txt="Figure 3: Correlation Matrix"
plt.figtext(0.4, 0.01, txt, wrap=True, horizontalalignment='center',␣
↪fontsize=14);
```

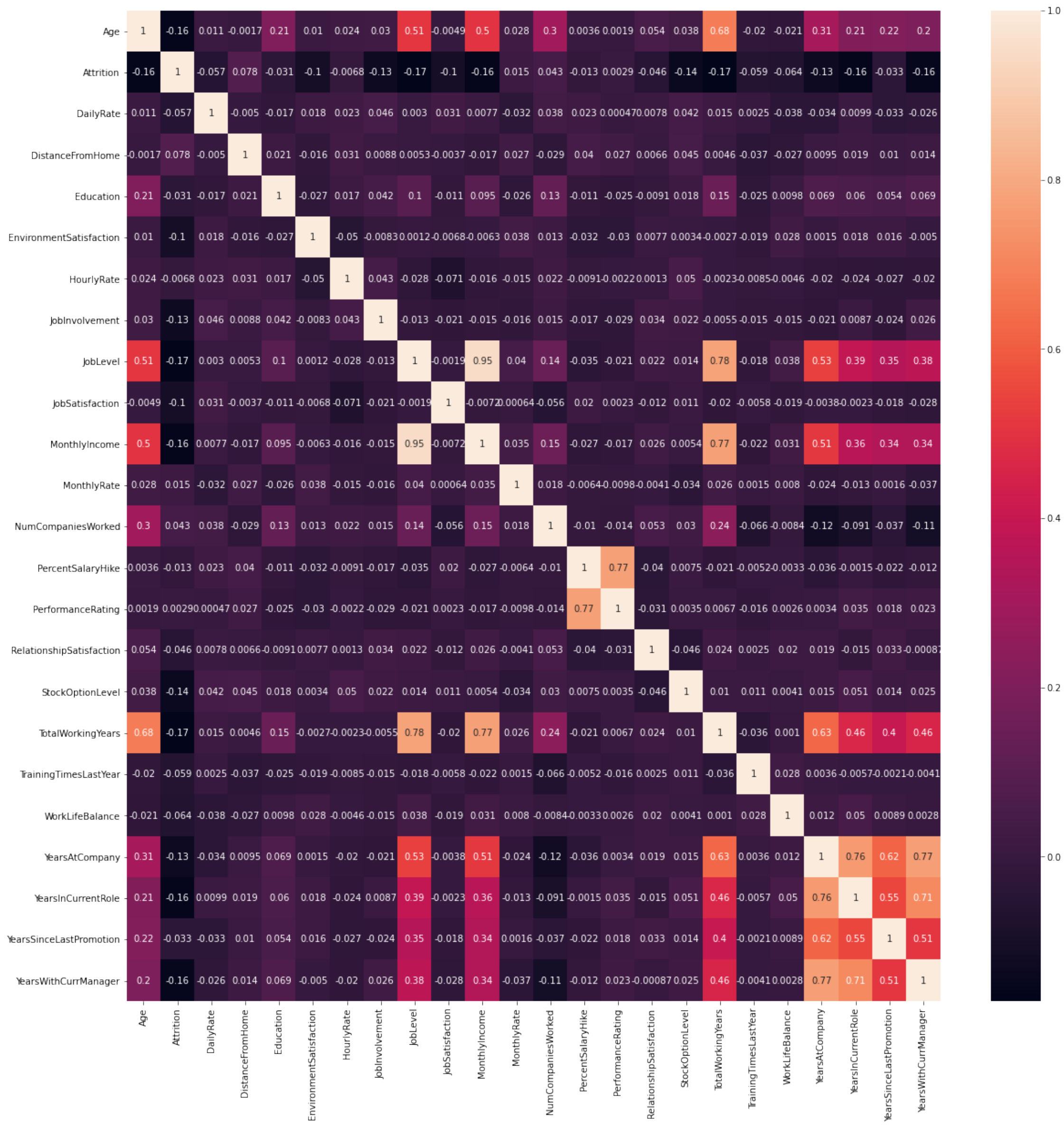


Figure 3: Correlation Matrix

The follwoing conclussions can be drawn from the correlation matix:

1. Overtime is very strongly correlated with attrition. This means people don't like to work overtime. Also "Distance from home" and "NumCompaniesWorked" are also positively correlated with attrition. Performace rating also have small correlation with attrition.

2. There different factors which are negatively correlated to the attrition but their values is very low:
   1) "Age" -people with more age like to stay in the company
   2) Job involvement-Employees' more involved in the work, they like to work in the comapny
   3) Job level- Means people at higher position like to stay in the company
   4) Monthly Income- High income emplyees like to stay in the company

5) Total Working years- Employees with more experinces like to stay in the company
6) year At comapny-Menas Emplyees working from long time in the company, they dont want to leave the comapny
7) Years in current role- It looks from the data that people don't like to change their role
8) Years with current manager- More emplyees like to work with the same manager

3. There is very strong correlation have been observed :
1) YearsAtCompany, Totalworkingyears, YearsIn CurentRole, YearsSinceLast Promotion and YearsWithCurrentManagers are all related field and function of time and increased with time. Similarly Age, Job level and MonthlyIncome are also raised with these five factors
2) There is 773) Also there is 95

We have observed some correlations between different factors and attrition. However, we have not observed any strong correlation between attrition and any other factors. Therefore to deep dive into the exploritory analysis in detals we further analyze all other factors and their behavious.

[22]:
```
Left_employees=employee_df1[employee_df1["Attrition"]==1]
```

[23]:
```
Left_employees
```

[23]:
```
      Age  Attrition     BusinessTravel  DailyRate              Department  \
0      41          1      Travel_Rarely       1102                   Sales
2      37          1      Travel_Rarely       1373  Research & Development
14     28          1      Travel_Rarely        103  Research & Development
21     36          1      Travel_Rarely       1218                   Sales
24     34          1      Travel_Rarely        699  Research & Development
…     …        …                …        …                     …
1438   23          1  Travel_Frequently        638                   Sales
1442   29          1      Travel_Rarely       1092  Research & Development
1444   56          1      Travel_Rarely        310  Research & Development
1452   50          1  Travel_Frequently        878                   Sales
1461   50          1      Travel_Rarely        410                   Sales

      DistanceFromHome  Education    EducationField  EnvironmentSatisfaction  \
0                    1          2     Life Sciences                        2
2                    2          2             Other                        4
14                  24          3     Life Sciences                        3
21                   9          4     Life Sciences                        3
24                   6          1           Medical                        2
…                  …        …               …                      …
1438                 9          3         Marketing                        4
1442                 1          4           Medical                        1
1444                 7          2  Technical Degree                        4
1452                 1          4     Life Sciences                        2
1461                28          3         Marketing                        4

      Gender  …  PerformanceRating  RelationshipSatisfaction  \
```

```
0     Female  …                 3                        1
2       Male  …                 3                        2
14      Male  …                 3                        2
21      Male  …                 4                        2
24      Male  …                 3                        3
…        …  …              …                        …
1438    Male  …                 3                        1
1442    Male  …                 3                        2
1444    Male  …                 3                        4
1452    Male  …                 3                        4
1461    Male  …                 3                        2

      StockOptionLevel TotalWorkingYears  TrainingTimesLastYear  \
0                    0                 8                      0
2                    0                 7                      3
14                   0                 6                      4
21                   0                10                      4
24                   0                 8                      2
…                  …                …                     …
1438                 1                 1                      3
1442                 3                 4                      3
1444                 1                14                      4
1452                 2                12                      3
1461                 1                20                      3

     WorkLifeBalance  YearsAtCompany  YearsInCurrentRole  \
0                  1               6                   4
2                  3               0                   0
14                 3               4                   2
21                 3               5                   3
24                 3               4                   2
…                …              …                  …
1438               2               1                   0
1442               4               2                   2
1444               1              10                   9
1452               3               6                   3
1461               3               3                   2

      YearsSinceLastPromotion YearsWithCurrManager
0                           0                    5
2                           0                    0
14                          0                    3
21                          0                    3
24                          1                    3
…                        …                   …
1438                        1                    0
1442                        2                    2
```

```
1444                        9                    8
1452                        0                    1
1461                        2                    0

[237 rows x 31 columns]
```

[24]:
```python
Stayed_employees=employee_df1[employee_df1["Attrition"]==0]
Stayed_employees
```

[24]:
```
      Age  Attrition     BusinessTravel  DailyRate              Department  \
1      49          0  Travel_Frequently        279  Research & Development
3      33          0  Travel_Frequently       1392  Research & Development
4      27          0      Travel_Rarely        591  Research & Development
5      32          0  Travel_Frequently       1005  Research & Development
6      59          0      Travel_Rarely       1324  Research & Development
…     …          …                  …        …                       …
1465   36          0  Travel_Frequently        884  Research & Development
1466   39          0      Travel_Rarely        613  Research & Development
1467   27          0      Travel_Rarely        155  Research & Development
1468   49          0  Travel_Frequently       1023                   Sales
1469   34          0      Travel_Rarely        628  Research & Development

      DistanceFromHome  Education EducationField  EnvironmentSatisfaction  \
1                    8          1  Life Sciences                        3
3                    3          4  Life Sciences                        4
4                    2          1        Medical                        1
5                    2          2  Life Sciences                        4
6                    3          3        Medical                        3
…                  …        …            …                      …
1465                23          2        Medical                        3
1466                 6          1        Medical                        4
1467                 4          3  Life Sciences                        2
1468                 2          3        Medical                        4
1469                 8          3        Medical                        2

      Gender  …  PerformanceRating  RelationshipSatisfaction  \
1       Male  …                  4                         4
3     Female  …                  3                         3
4       Male  …                  3                         4
5       Male  …                  3                         3
6     Female  …                  4                         1
…       …  …                …                       …
1465    Male  …                  3                         3
1466    Male  …                  3                         1
1467    Male  …                  4                         2
1468    Male  …                  3                         4
1469    Male  …                  3                         1
```

```
      StockOptionLevel  TotalWorkingYears  TrainingTimesLastYear  \
1                    1                 10                      3
3                    0                  8                      3
4                    1                  6                      3
5                    0                  8                      2
6                    3                 12                      3
…                  …                …                    …
1465                 1                 17                      3
1466                 1                  9                      5
1467                 1                  6                      0
1468                 0                 17                      3
1469                 0                  6                      3

     WorkLifeBalance  YearsAtCompany  YearsInCurrentRole  \
1                  3              10                   7
3                  3               8                   7
4                  3               2                   2
5                  2               7                   7
6                  2               1                   0
…               …             …                 …
1465               3               5                   2
1466               3               7                   7
1467               3               6                   2
1468               2               9                   6
1469               4               4                   3

      YearsSinceLastPromotion  YearsWithCurrManager
1                           1                     7
3                           3                     0
4                           2                     2
5                           3                     6
6                           0                     0
…                        …                   …
1465                        0                     3
1466                        1                     7
1467                        0                     3
1468                        0                     8
1469                        1                     2

[1233 rows x 31 columns]
```

[25]:
```
Stayed_employees.describe()
```

[25]:
```
               Age  Attrition    DailyRate  DistanceFromHome    Education  \
count  1233.000000     1233.0  1233.000000       1233.000000  1233.000000
mean     37.561233        0.0   812.504461          8.915653     2.927007
```

```
std       8.888360        0.0   403.208379          8.012633     1.027002
min      18.000000        0.0   102.000000          1.000000     1.000000
25%      31.000000        0.0   477.000000          2.000000     2.000000
50%      36.000000        0.0   817.000000          7.000000     3.000000
75%      43.000000        0.0  1176.000000         13.000000     4.000000
max      60.000000        0.0  1499.000000         29.000000     5.000000

       EnvironmentSatisfaction   HourlyRate  JobInvolvement     JobLevel  \
count              1233.000000  1233.000000     1233.000000  1233.000000
mean                  2.771290    65.952149        2.770479     2.145985
std                   1.071132    20.380754        0.692050     1.117933
min                   1.000000    30.000000        1.000000     1.000000
25%                   2.000000    48.000000        2.000000     1.000000
50%                   3.000000    66.000000        3.000000     2.000000
75%                   4.000000    83.000000        3.000000     3.000000
max                   4.000000   100.000000        4.000000     5.000000

       JobSatisfaction  …  PerformanceRating  RelationshipSatisfaction  \
count      1233.000000  …        1233.000000               1233.000000
mean          2.778589  …           3.153285                  2.733982
std           1.093277  …           0.360408                  1.071603
min           1.000000  …           3.000000                  1.000000
25%           2.000000  …           3.000000                  2.000000
50%           3.000000  …           3.000000                  3.000000
75%           4.000000  …           3.000000                  4.000000
max           4.000000  …           4.000000                  4.000000

       StockOptionLevel  TotalWorkingYears  TrainingTimesLastYear  \
count       1233.000000        1233.000000            1233.000000
mean           0.845093          11.862936               2.832928
std            0.841985           7.760719               1.293585
min            0.000000           0.000000               0.000000
25%            0.000000           6.000000               2.000000
50%            1.000000          10.000000               3.000000
75%            1.000000          16.000000               3.000000
max            3.000000          38.000000               6.000000

       WorkLifeBalance  YearsAtCompany  YearsInCurrentRole  \
count      1233.000000     1233.000000         1233.000000
mean          2.781022        7.369019            4.484185
std           0.681907        6.096298            3.649402
min           1.000000        0.000000            0.000000
25%           2.000000        3.000000            2.000000
50%           3.000000        6.000000            3.000000
75%           3.000000       10.000000            7.000000
max           4.000000       37.000000           18.000000
```

```
       YearsSinceLastPromotion  YearsWithCurrManager
count              1233.000000           1233.000000
mean                  2.234388              4.367397
std                   3.234762              3.594116
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 24 columns]
```

[26]:
```
Left_employees.describe()
```

[26]:
```
              Age  Attrition    DailyRate  DistanceFromHome   Education  \
count  237.000000      237.0   237.000000        237.000000  237.000000
mean    33.607595        1.0   750.362869         10.632911    2.839662
std      9.689350        0.0   401.899519          8.452525    1.008244
min     18.000000        1.0   103.000000          1.000000    1.000000
25%     28.000000        1.0   408.000000          3.000000    2.000000
50%     32.000000        1.0   699.000000          9.000000    3.000000
75%     39.000000        1.0  1092.000000         17.000000    4.000000
max     58.000000        1.0  1496.000000         29.000000    5.000000

       EnvironmentSatisfaction  HourlyRate  JobInvolvement    JobLevel  \
count               237.000000  237.000000      237.000000  237.000000
mean                  2.464135   65.573840        2.518987    1.637131
std                   1.169791   20.099958        0.773405    0.940594
min                   1.000000   31.000000        1.000000    1.000000
25%                   1.000000   50.000000        2.000000    1.000000
50%                   3.000000   66.000000        3.000000    1.000000
75%                   4.000000   84.000000        3.000000    2.000000
max                   4.000000  100.000000        4.000000    5.000000

       JobSatisfaction  …  PerformanceRating  RelationshipSatisfaction  \
count       237.000000  …         237.000000                237.000000
mean          2.468354  …           3.156118                  2.599156
std           1.118058  …           0.363735                  1.125437
min           1.000000  …           3.000000                  1.000000
25%           1.000000  …           3.000000                  2.000000
50%           3.000000  …           3.000000                  3.000000
75%           3.000000  …           3.000000                  4.000000
max           4.000000  …           4.000000                  4.000000

       StockOptionLevel  TotalWorkingYears  TrainingTimesLastYear  \
count        237.000000         237.000000             237.000000
mean           0.527426           8.244726               2.624473
```

```
std          0.856361         7.169204             1.254784
min          0.000000         0.000000             0.000000
25%          0.000000         3.000000             2.000000
50%          0.000000         7.000000             2.000000
75%          1.000000        10.000000             3.000000
max          3.000000        40.000000             6.000000

       WorkLifeBalance  YearsAtCompany  YearsInCurrentRole  \
count       237.000000      237.000000          237.000000
mean          2.658228        5.130802            2.902954
std           0.816453        5.949984            3.174827
min           1.000000        0.000000            0.000000
25%           2.000000        1.000000            0.000000
50%           3.000000        3.000000            2.000000
75%           3.000000        7.000000            4.000000
max           4.000000       40.000000           15.000000

       YearsSinceLastPromotion  YearsWithCurrManager
count               237.000000            237.000000
mean                  1.945148              2.852321
std                   3.153077              3.143349
min                   0.000000              0.000000
25%                   0.000000              0.000000
50%                   1.000000              2.000000
75%                   2.000000              5.000000
max                  15.000000             14.000000

[8 rows x 24 columns]
```

[27]:

```
print("Percentage of employees left ", len(Left_employees)/
 ↪len(employee_df1)*100)
```

```
Percentage of employees left  16.122448979591837
```

[28]:

```
employee_df1.hist(["Age"], bins=10)
txt="Figure 4: Histogram for Age"
plt.figtext(0.5, 0.03, txt, wrap=True, horizontalalignment='center',␣
 ↪fontsize=14);
```

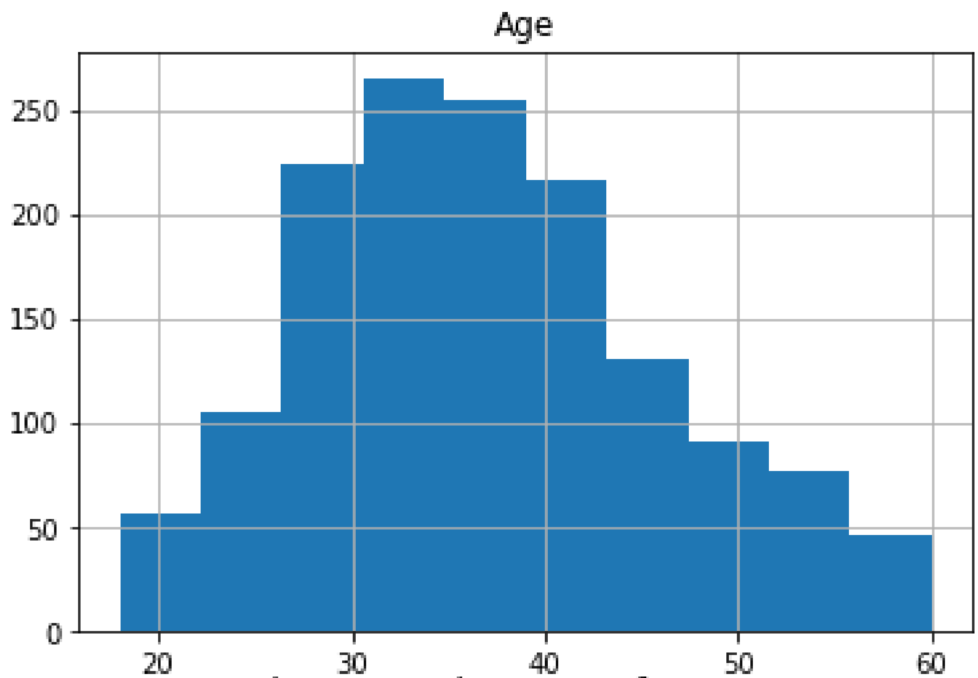


Figure 4: Histogram for Age

[29]:
```
employee_test=employee_df
```

[30]:
```
employee_test["Test"]=employee_df1["Age"].apply(lambda x: 1 if x>28 and x< 42␣
↪else 0)
```

[31]:
```
employee_test["Test"].value_counts()
```

[31]:
```
1    787
0    683
Name: Test, dtype: int64
```

[32]:
```
employee_df1["Age"].apply(lambda x: 1 if x>28 and x< 42 else 0).value_counts()
```

[32]:
```
1    787
0    683
Name: Age, dtype: int64
```

More than half of the employees are falling under the age from 29 to 41. It can also be seen from the histograph

[33]:
```
df=employee_df1[["Attrition","Age"]]
```

[34]:
```
fig, ax = plt.subplots(figsize=(10,4))
for key, grp in employee_df1.groupby(['Attrition']):
```

```
    ax.scatter(grp['Age'], grp['Attrition'], label=key)

ax.legend()
txt="Figure 5: Scatter plot of Age  for attrition (1) or non attrition (0) "
plt.figtext(0.5, 0.02, txt, wrap=True, horizontalalignment='center',␣
 ↪fontsize=14);
plt.show()
```

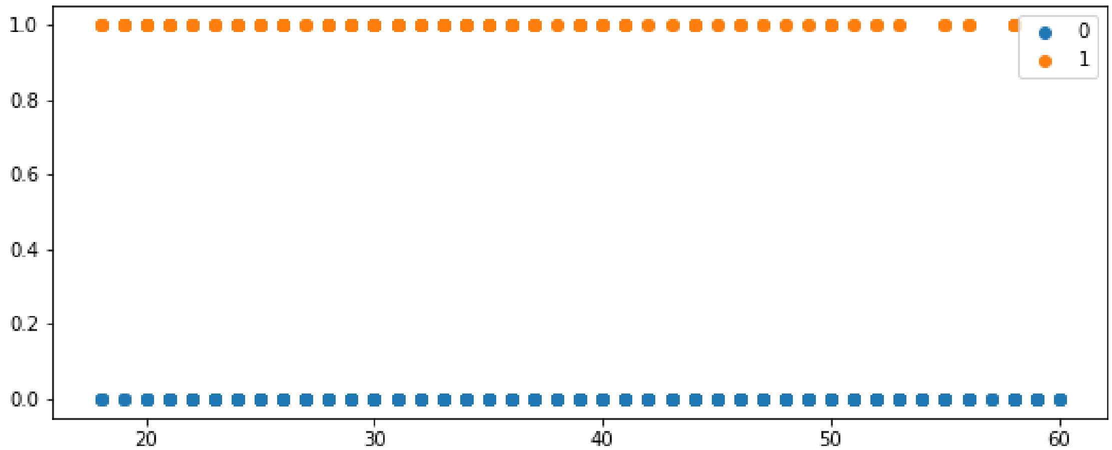


Figure 5: Scatter plot of Age for attrition (1) or non attrition (0)

[35]:
```
df1=df.groupby(['Attrition',"Age"]).count()
```

[36]:
```
import math
plt.figure(figsize=(20,10))
ax=sns.countplot(data=df, x="Age",hue='Attrition');
for p in ax.patches:
    ax.annotate(f' {p.get_height():.0f}', xy = (p.get_x()+p.get_width()/ 2, p.
 ↪get_height()+1),
                   ha='center',
                   va='center',
                   size=14,
                   xytext=(0, 8),
                   textcoords='offset points'
                )
txt="Figure 6: Comparision stayed and left employees for different ages"
plt.figtext(0.5, 0.02, txt, wrap=True, horizontalalignment='center',␣
 ↪fontsize=14);
```

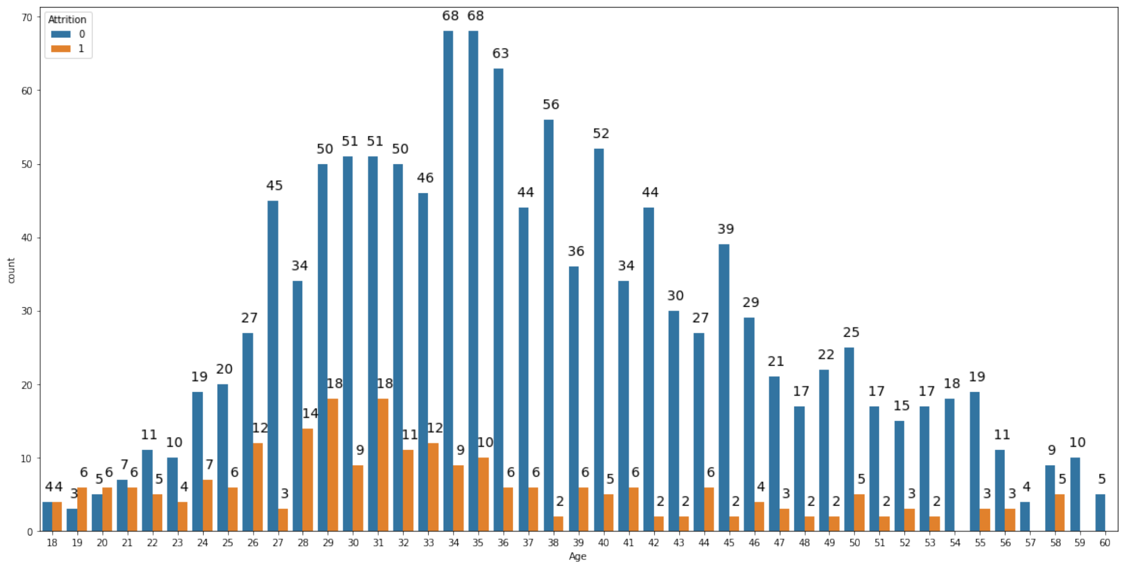


Figure 6: Comparision stayed and left employees for different ages

[37]:
```
plt.figure(figsize=(20,10))
w = 10
b = math.ceil((employee_df1["Age"].max() - employee_df1["Age"].min())/w)
ax=sns.histplot(data=employee_df1, x='Age', hue='Attrition',bins=b, );
y=[]
for i in ax.patches:
      x=i.get_height()
      y.append(x)
i=0
v=1
h=0
for p in ax.patches:
    if i>=b:
        k=y[i]/(y[i]+y[i-b])*100
        v=5
        h=1
    else:
        k=y[i]/(y[i]+y[i+b])*100
        t=1


    ax.annotate(f' {k:.2f} %', xy = (p.get_x()+p.get_width()/ 2+h, p.
 ↪get_height()+v),
                   ha='center',
                   va='center',
                   size=14,
                   xytext=(0, 8),
```

```
                    textcoords='offset points'
                )
    i=i+1
x = np.arange(0, 61)
plt.xticks(x);
txt="Figure 7: Comparision of different age groups for left (1) and stayed␣
↪employees (0)"
plt.figtext(0.5, 0.03, txt, wrap=True, horizontalalignment='center',␣
↪fontsize=14);
```

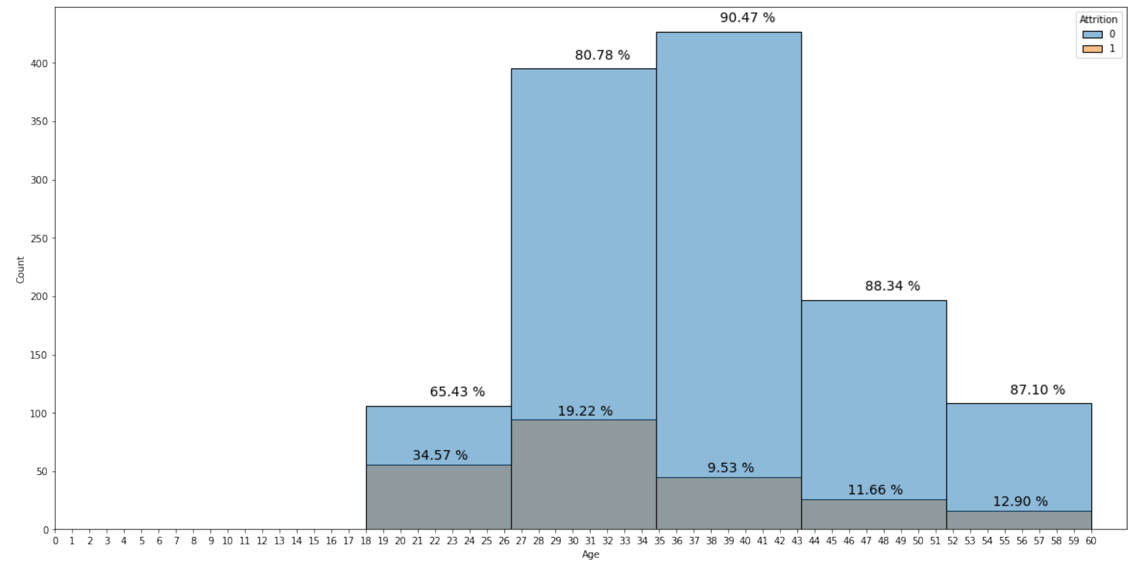


Figure 7: Comparision of different age groups for left (1) and stayed employees (0)

It can be obeserved from the above graph that employee with age group from 18-35 are more likely to attrition.

[38]:
```
employee_df["BusinessTravel"]
```

[38]:
```
0           Travel_Rarely
1       Travel_Frequently
2           Travel_Rarely
3       Travel_Frequently
4           Travel_Rarely
              …
1465    Travel_Frequently
1466        Travel_Rarely
1467        Travel_Rarely
1468    Travel_Frequently
1469        Travel_Rarely
Name: BusinessTravel, Length: 1470, dtype: object
```

[39]:
```
employee_df1["BusinessTravel"].value_counts()
```

[39]:
```
Travel_Rarely        1043
Travel_Frequently     277
Non-Travel            150
Name: BusinessTravel, dtype: int64
```

[40]:
```
plt.figure(figsize=(15,8))
ax=sns.countplot(data=employee_df1, x="BusinessTravel",hue='Attrition');
b=3
y=[]
for i in ax.patches:
      y.append(i.get_height())

i=0
for p in ax.patches:
    if i>=b:
        k=y[i]/(y[i]+y[i-b])*100
    else:
        k=y[i]/(y[i]+y[i+b])*100
        v=1
    ax.annotate(f' {k:.0f} %', xy = (p.get_x()+p.get_width()/ 2, p.
 ↪get_height()+v),
                   ha='center',
                   va='center',
                   size=14,
                   xytext=(0, 8),
                   textcoords='offset points'
               )
    i=i+1
txt="Figure 8: Comparision of rarely travelled, frequently travelled and␣
 ↪non-travelled employess for attrition"
plt.figtext(0.5, 0.01, txt, wrap=True, horizontalalignment='center',␣
 ↪fontsize=14);
```

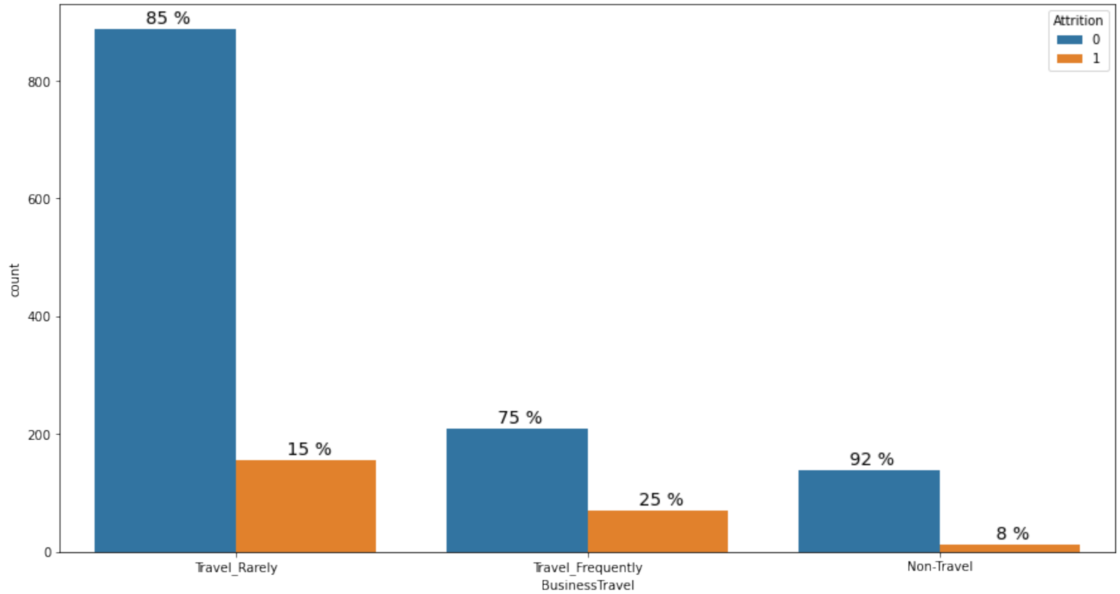


Figure 8: Comparision of rarely travelled, frequently travelled and non-travelled employess for attrition

It can be observed that attrition rate is higher fro employess who travel frequently and lowest for non traveler. It can be said that business travel is playing an important role in the attriton of employees

[41]:
```
plt.figure(figsize=(15,8))
ax=sns.countplot(data=employee_df1, x="MaritalStatus",hue='Attrition')
b=3
y=[]
for i in ax.patches:
      y.append(i.get_height())

i=0
for p in ax.patches:
    if i>=b:
        k=y[i]/(y[i]+y[i-b])*100
    else:
        k=y[i]/(y[i]+y[i+b])*100
        v=1
    ax.annotate(f' {k:.0f} %', xy = (p.get_x()+p.get_width()/ 2, p.
 ↪get_height()+v),
                   ha='center',
                   va='center',
                   size=14,
                   xytext=(0, 8),
                   textcoords='offset points'
               )
```

```
    i=i+1
txt="Figure 9: Percentage comaprision of employess who left as compared to␣
↪stayed emplyees fro their marital status"
plt.figtext(0.5, 0.01, txt, wrap=True, horizontalalignment='center',␣
↪fontsize=14);
```

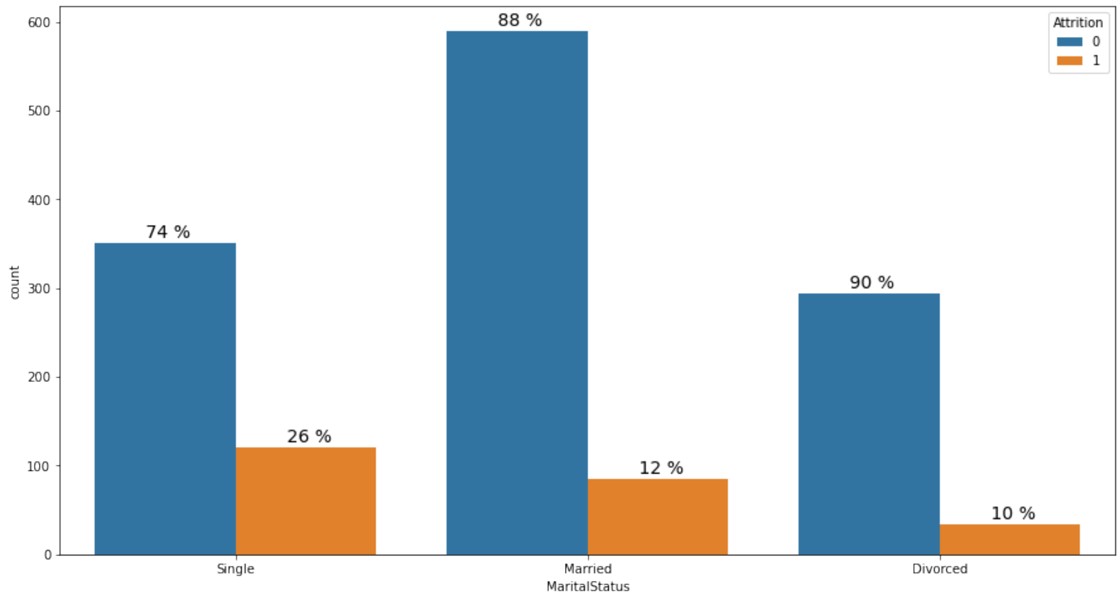


Figure 9: Percentage comaprision of employess who left as compared to stayed emplyees fro their marital status

It can be observed that attrition rate is higher for employess single employees. Marrited and divorced employees would not like to move to another place

[42]:
```
plt.figure(figsize=(15,15))
g=sns.catplot(data=employee_df1, x="BusinessTravel",hue='Attrition',␣
↪col="MaritalStatus", kind="count")
y=[]
for ax in g.axes.ravel():
    for i in ax.patches:
          y.append(i.get_height())

for ax in g.axes.ravel():
    for p in ax.patches:
        ax.annotate(f' {p.get_height():.0f}', xy = (p.get_x()+p.get_width()/ 2,␣
↪p.get_height()),
                       ha='center',
                       va='center',
                       size=14,
                       xytext=(0, 8),
                       textcoords='offset points'
```

```
                   )
txt="Figure 10: Comaprision of employess who left as compared to stayed␣
↪emplyees for bussiness travel based on theri marital status"
plt.figtext(0.5, -0.1, txt, wrap=True, horizontalalignment='center',␣
↪fontsize=14);
```

```
<Figure size 1080x1080 with 0 Axes>
```

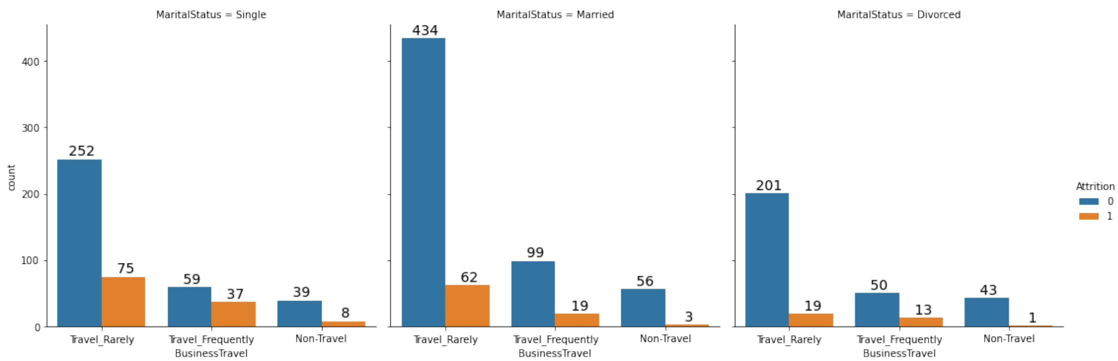


Figure 10: Comaprision of employess who left as compared to stayed emplyees for bussiness travel based on theri marital status

[43]:

```
plt.figure(figsize=(20,10))
g=sns.catplot(data=employee_df1, x="BusinessTravel",hue='Attrition',␣
↪col="MaritalStatus", kind="count", palette="viridis_r")
y=[]
for ax in g.axes.ravel():
    for i in ax.patches:
          y.append(i.get_height())
i=0
b=3
l=3
for ax in g.axes.ravel():
    for p in ax.patches:
        if i>=l:
            k=y[i]/(y[i]+y[i-b])*100
        else:
            k=y[i]/(y[i]+y[i+b])*100
            v=1
        ax.annotate(f' {k:.0f} %', xy = (p.get_x()+p.get_width()/ 2, p.
↪get_height()),
                       ha='center',
                       va='center',
                       size=14,
                       xytext=(0, 8),
                       textcoords='offset points'
                   )
```

```
        i=i+1
    l=l+6
txt="Figure 11: Comaprision of employess who left as compared to stayed␣
↪emplyees for bussiness travel based on theri marital status"
plt.figtext(0.5, -0.1, txt, wrap=True, horizontalalignment='center',␣
↪fontsize=14);
```

<Figure size 1440x720 with 0 Axes>

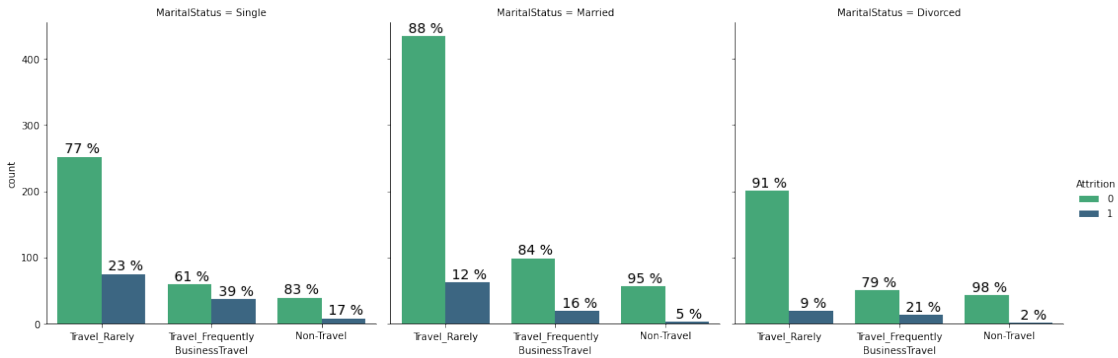


Figure 11: Comaprision of employess who left as compared to stayed emplyees for bussiness travel based on theri marital status

The follwoing are some observation drawn for marital status and busness travel:

1. It can observed that for married employees that who travel frquently have higher percentage of attrition than non traveler and rarely traveler. This percentage is more higher for the employees who are divorced. May be they are single parent and dont want to travel.
2. There is another factor about this that the company only give more responsibilities to single employees for travel.
3. So we can say that business travel and marital status both contribution to attriation of employees

[44]:
```
plt.figure(figsize=(10,8))
ax=sns.countplot( data=employee_df1, x="Department", palette='Greens')
for p in ax.patches:
    ax.annotate(f' {p.get_height():.0f}', xy = (p.get_x()+p.get_width()/ 2, p.
↪get_height()+v),
                   ha='center',
                   va='center',
                   size=14,
                   xytext=(0, 8),
                   textcoords='offset points'
               )
txt="Figure 12: Number of employees in different departments"
plt.figtext(0.5, 0.01, txt, wrap=True, horizontalalignment='center',␣
↪fontsize=14);
```

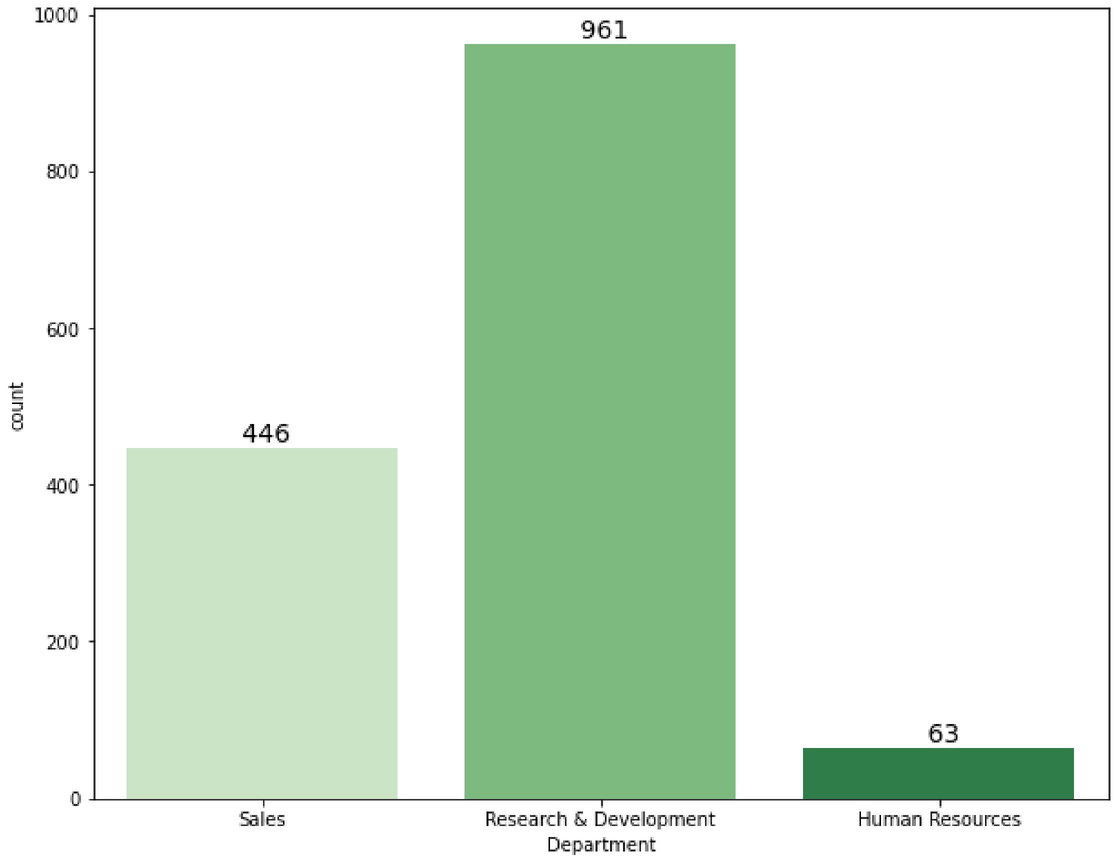


Figure 12: Number of employees in different departments

[45]:
```
plt.figure(figsize=(15,8))
ax=sns.countplot( data=employee_df1, x="Department",␣
↪hue="Attrition",palette='Greens')
for p in ax.patches:
    ax.annotate(f' {p.get_height():.0f}', xy = (p.get_x()+p.get_width()/ 2, p.
↪get_height()+v),
                   ha='center',
                   va='center',
                   size=14,
                   xytext=(0, 8),
                   textcoords='offset points'
               )
txt="Figure 13: Number of left and stayed employees in different departments"
plt.figtext(0.5, 0.01, txt, wrap=True, horizontalalignment='center',␣
↪fontsize=14);
```

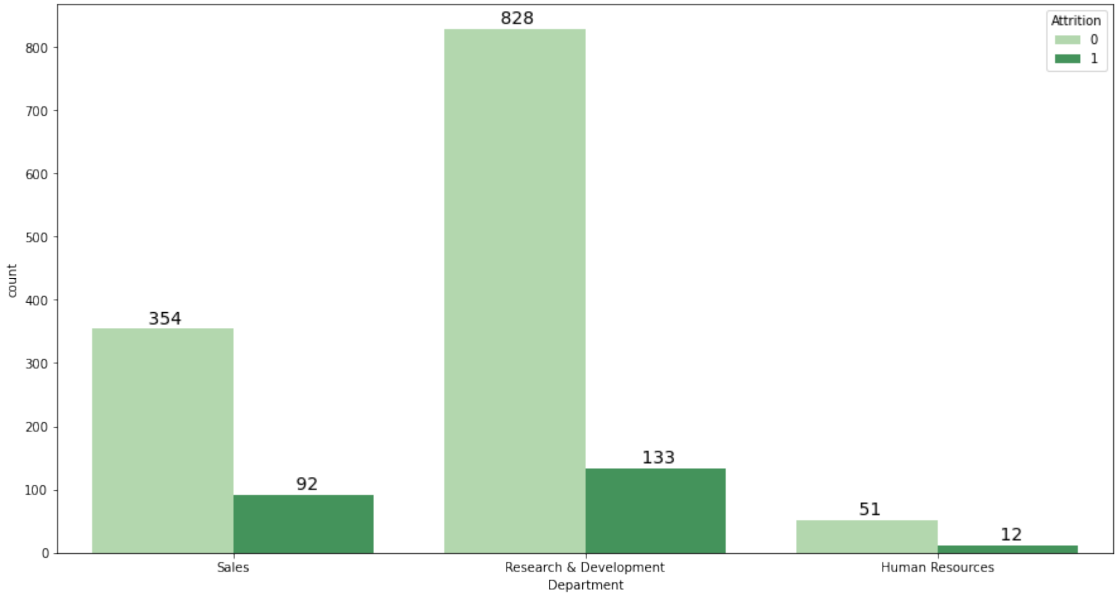


Figure 13: Number of left and stayed employees in different departments

[46]:
```
plt.figure(figsize=(15,8))
ax=sns.countplot(data=employee_df1,␣
↪x="Department",hue='Attrition',palette='Greens')
b=3
y=[]
for i in ax.patches:
      y.append(i.get_height())

i=0
for p in ax.patches:
    if i>=b:
        k=y[i]/(y[i]+y[i-b])*100
    else:
        k=y[i]/(y[i]+y[i+b])*100
        v=1
    ax.annotate(f' {k:.0f} %', xy = (p.get_x()+p.get_width()/ 2, p.
↪get_height()+v),
                    ha='center',
                    va='center',
                    size=14,
                    xytext=(0, 8),
                    textcoords='offset points'
                )
    i=i+1
```

```
txt="Figure 14: Percentage comparision of left and stayed employees in␣
↪different departments"
plt.figtext(0.5, 0.01, txt, wrap=True, horizontalalignment='center',␣
↪fontsize=14);
```

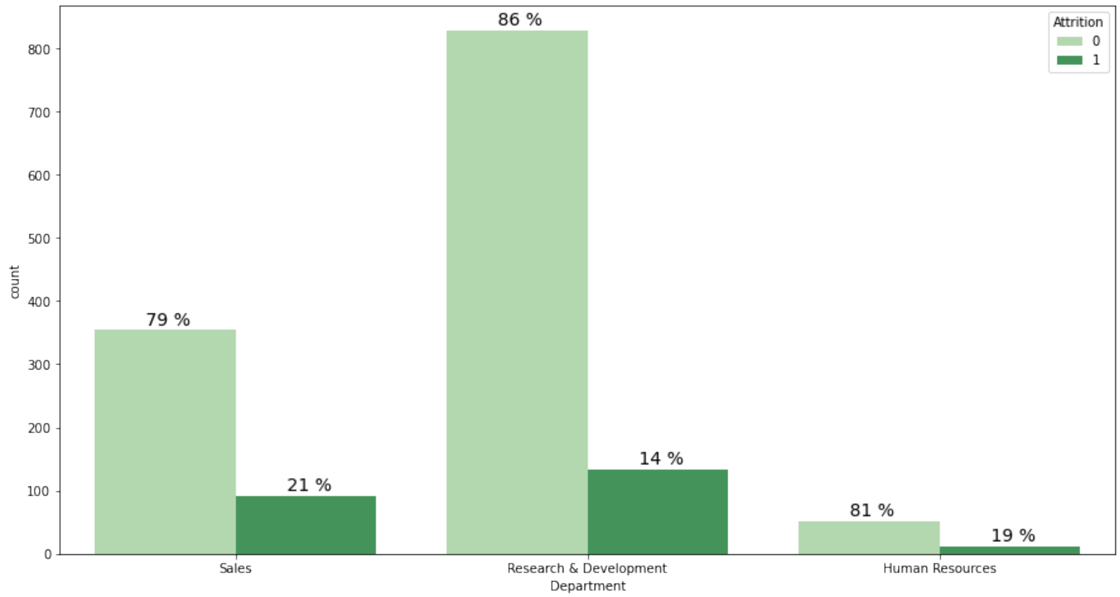


Figure 14: Percentage comparision of left and stayed employees in different departments

The sales and human resource employees have higher chances to leave the company.May be they have more exposer to new opportunities and have strong network connections.

[47]:
```
plt.figure(figsize=(20,10))
ax=sns.countplot(data=employee_df1, x="Department",hue='Age')
txt="Figure 15: Department wise Age distribution"
plt.figtext(0.5, 0.01, txt, wrap=True, horizontalalignment='center',␣
↪fontsize=14);
```

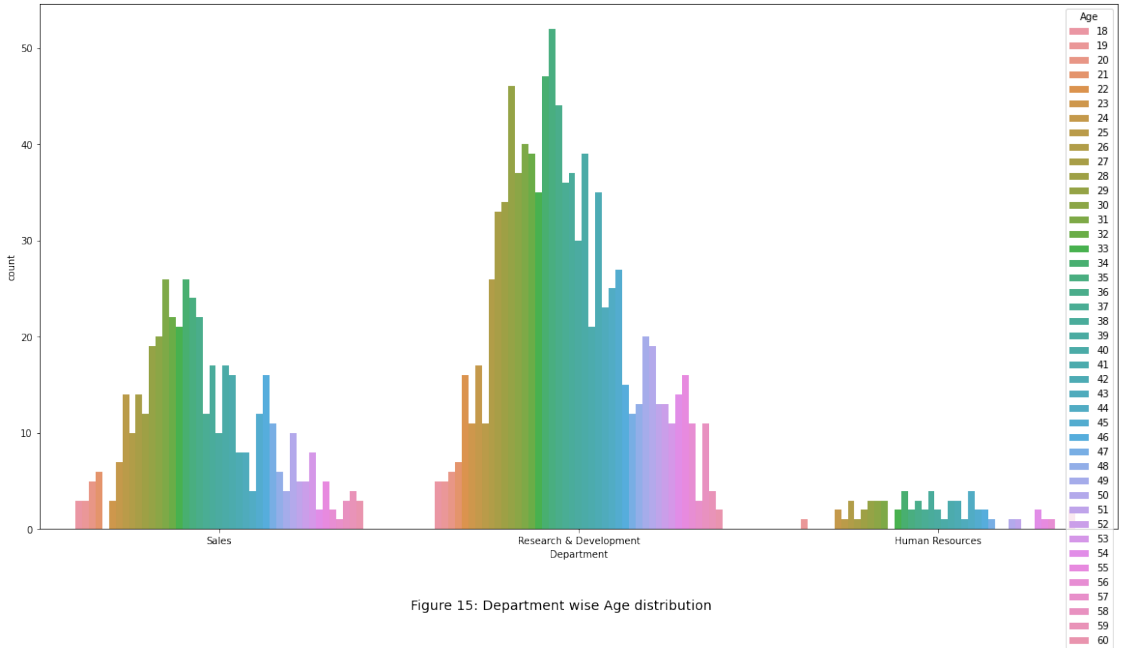


Figure 15: Department wise Age distribution

Every department have mixed employees of all age groups

[48]:
```
plt.figure(figsize=(20,10))
ax=sns.countplot(data=employee_df1, x="JobRole",hue='Attrition')
y=[]
for i in ax.patches:
      y.append(i.get_height())
i=0
b=9
for p in ax.patches:
    if i>=b:
        k=y[i]/(y[i]+y[i-b])*100
    else:
        k=y[i]/(y[i]+y[i+b])*100

    ax.annotate(f' {k:.0f}%', xy = (p.get_x()+p.get_width()/ 2, p.get_height()),
                   ha='center',
                   va='center',
                   size=13,
                   xytext=(0, 8),
                   textcoords='offset points'
               )
    i=i+1
txt="Figure 16: Attrition of employees based on job roles"
plt.figtext(0.5, 0.03, txt, wrap=True, horizontalalignment='center',␣
 ↪fontsize=14);
```

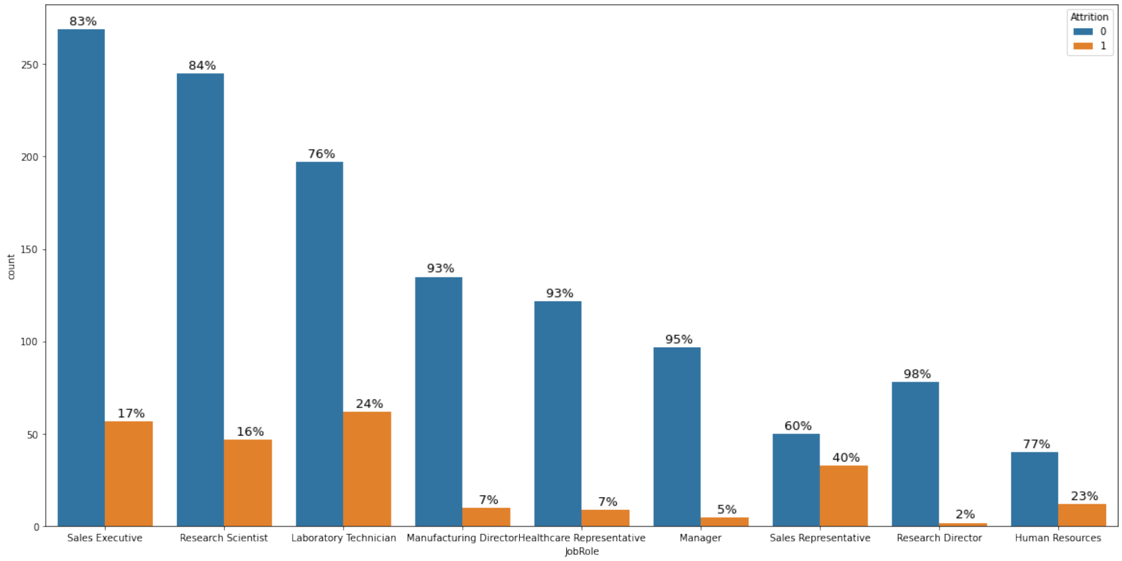


Figure 16: Attrition of employees based on job roles

Sales repersentative are among the maximum who left the comapny

[49]:
```
plt.figure(figsize=(25,20))

g=sns.catplot(data=employee_df1, col="Department",x='EducationField',␣
↪hue="Attrition", kind="count", palette="viridis_r")
g.set_xticklabels(rotation=30)

for ax in g.axes.ravel():
    for p in ax.patches:
        ax.annotate(f' {p.get_height():.0f}', xy = (p.get_x()+p.get_width()/␣
↪2, p.get_height()),
                    ha='center',
                    va='center',
                    size=13,
                    xytext=(0, 8),
                    textcoords='offset points'
                )
txt="Figure 17: Department wise number of employees educated in six different␣
↪areas"
plt.figtext(0.5, -0.23, txt, wrap=True, horizontalalignment='center',␣
↪fontsize=14);
```

```
<Figure size 1800x1440 with 0 Axes>
```

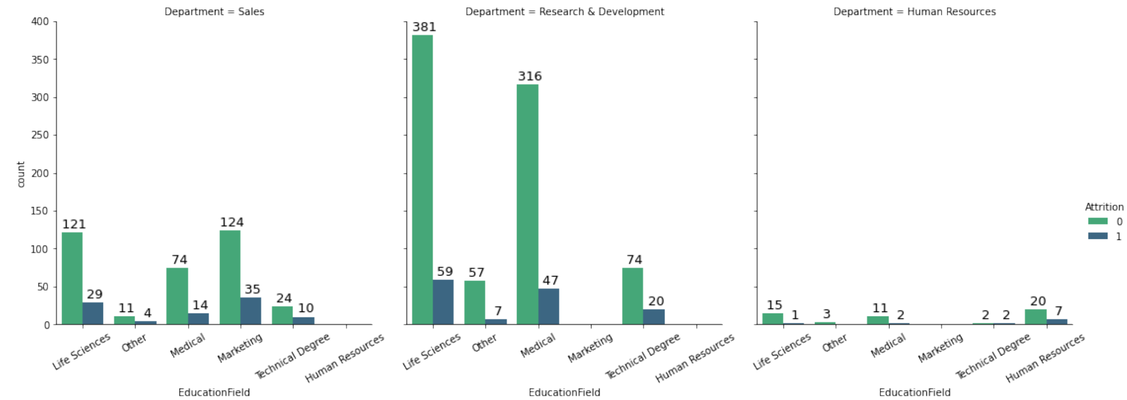


Figure 17: Department wise number of employees educated in six different areas

[50]:
```
g=sns.catplot(data=employee_df1, col="Department",x='EducationField',
    hue="Attrition", kind="count", palette="viridis_r")
g.set_xticklabels(rotation=30)
y=[]
for ax in g.axes.ravel():
    for i in ax.patches:
        y.append(i.get_height())
i=0
b=6
l=6

for ax in g.axes.ravel():
    for p in ax.patches:
        if i>=l:
            k=y[i]/(y[i]+y[i-b])*100
        else:
            k=y[i]/(y[i]+y[i+b])*100
            v=1
        ax.annotate(f' {k:.0f}%', xy = (p.get_x()+p.get_width()/ 2, p.
    get_height()),
                    ha='center',
                    va='center',
                    size=13,
                    xytext=(0, 8),
                    textcoords='offset points'
                )

        i=i+1
    l=l+12
txt="Figure 18: Department wise percentage of employees educated in six
    different areas"
```

```
plt.figtext(0.5, -0.23, txt, wrap=True, horizontalalignment='center',␣
↪fontsize=14);
```

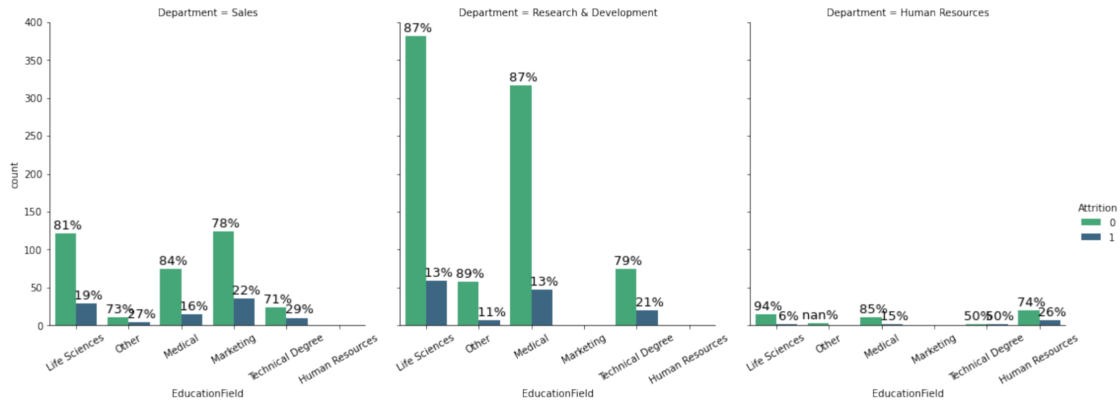


Figure 18: Department wise percentage of employees educated in six different areas

From the above graph there is no clear indication that the attrition is specific to field of education. Also we canont say that the attrition is happening becuse of the working in different area then the field of education.

[51]:
```
g=sns.catplot(data=employee_df1, col="JobRole", col_wrap=3,x='EducationField',␣
↪hue="Attrition", kind="count", palette="viridis_r")
g.set_xticklabels(rotation=30)

for ax in g.axes.ravel():
    for p in ax.patches:
        ax.annotate(f' {p.get_height():.0f}', xy = (p.get_x()+p.get_width()/␣
↪2, p.get_height()),
                        ha='center',
                        va='center',
                        size=13,
                        xytext=(0, 8),
                        textcoords='offset points'
                    )
txt="Figure 19: Employees count (left(1) and stayed (0)) for different job␣
↪roles based on theri field of education"
plt.figtext(0.5, -0.07, txt, wrap=True, horizontalalignment='center',␣
↪fontsize=14);
```

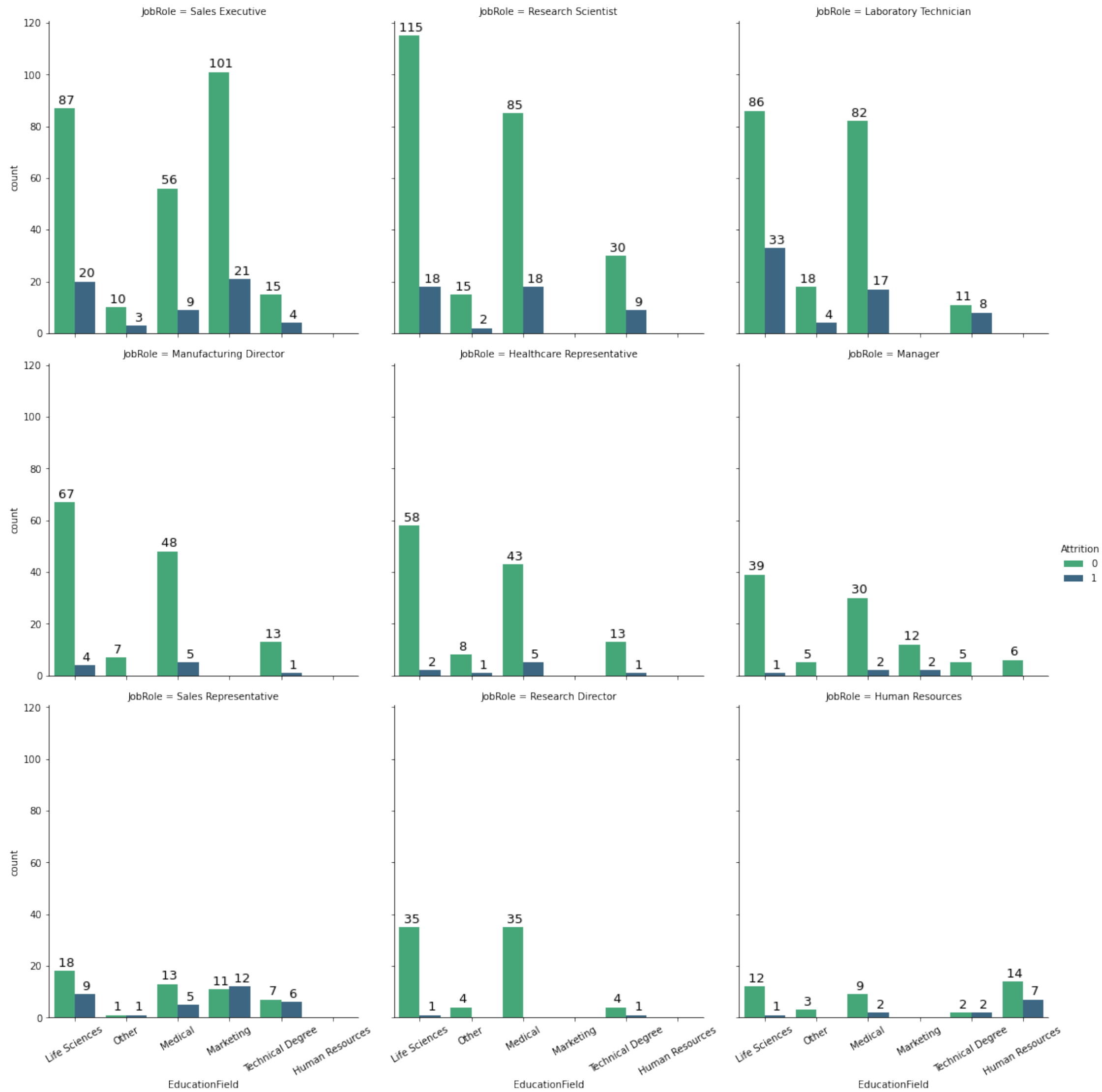


Figure 19: Employees count (left(1) and stayed (0)) for different job roles based on theri field of education

Again no conclusion can be drawn about attrition based on mismatch between education and job roles. Already most of the employees in the company are in their corresponding field of study. No marketing educated employee in the role of research directors, research scientist, laboratory technician, etc. and similar for other cases. However, It can be observed that almost 50% employee who have marketing degree left in the company for the role of sales representative. It is already observed that sales representatives among the maximum who left the comapany. May be the are not getting enough salery or there is an issue with the manager. It can also be seen that the 27 % Laboratory Assitants with life sciecne degree left the company. Also 33% human resource employees with human resource education left the comapany.

[52]:
```
plt.figure(figsize=(20,10))
g=sns.kdeplot(data=employee_df1,x='DistanceFromHome', hue="Attrition", 
 ↪shade=True, palette="OrRd")
txt="Figure 20: Desnsity distribution of left(1) and stayed (0) employees based␣
 ↪on distance from home"
plt.figtext(0.5, 0.02, txt, wrap=True, horizontalalignment='center',␣
 ↪fontsize=16);
```

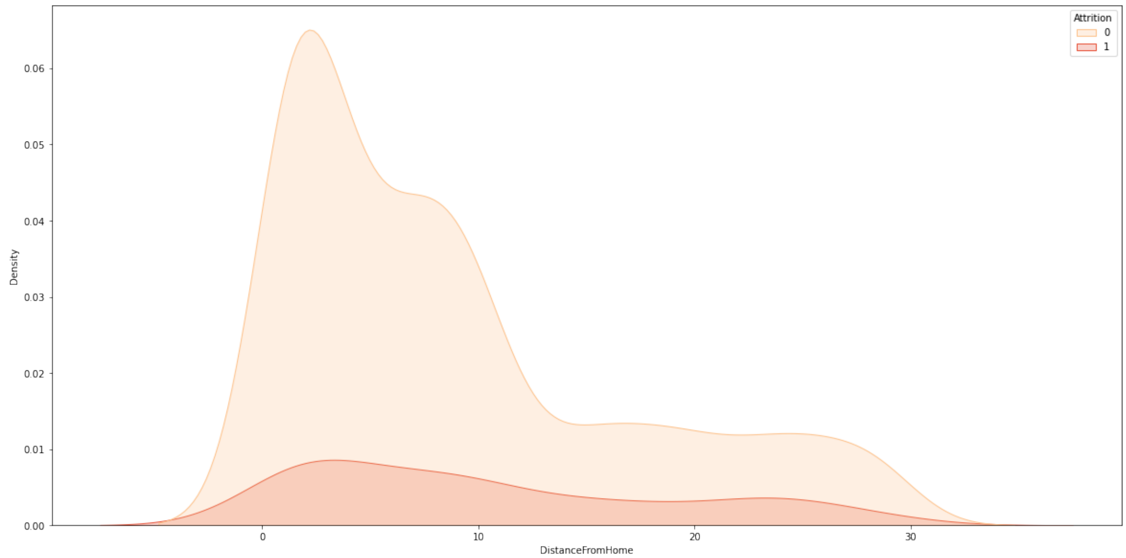


Figure 20: Desnsity distribution of left(1) and stayed (0) employees based on distance from home

It can be obereved that the percentage attrition is more around 25km. However, most of the employees are living less that 35 km from the office. This is the distance which maximum people can travel. But we even then we can see the effect of distace from home.

[53]:
```
plt.figure(figsize=(20,10))
ax=sns.countplot(data=employee_df1, x="Education",hue='Attrition')
y=[]
for i in ax.patches:
      y.append(i.get_height())
i=0
b=5
for p in ax.patches:
    if i>=b:
        k=y[i]/(y[i]+y[i-b])*100
    else:
        k=y[i]/(y[i]+y[i+b])*100

    ax.annotate(f' {k:.0f}%', xy = (p.get_x()+p.get_width()/ 2, p.get_height()),
```

```
                    ha='center',
                    va='center',
                    size=13,
                    xytext=(0, 8),
                    textcoords='offset points'
                )
    i=i+1;
txt="Figure 21: Percentage of left(1) and stayed (0) employees based on their␣
↪level of education"
plt.figtext(0.5, 0.02, txt, wrap=True, horizontalalignment='center',␣
↪fontsize=16);
```

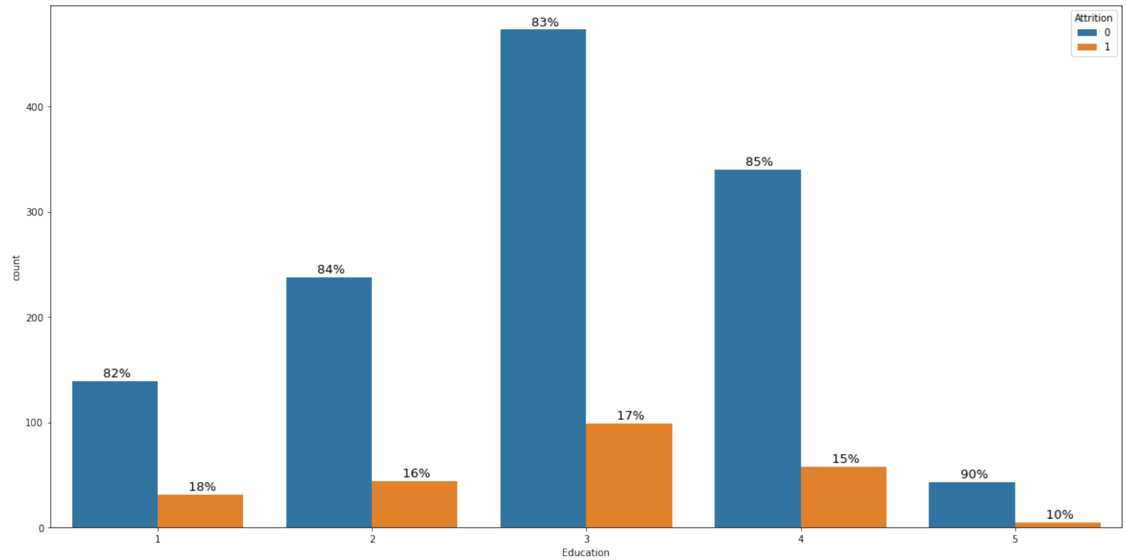


Figure 21: Percentage of left(1) and stayed (0) employees based on their level of education

There is not direct relationship between level of education and attrition.

[54]:
```
def text2(x,**kwargs):
    ax=plt.gca()
    i=0
    b=5
    y=[]
    for i in ax.patches:
        y.append(i.get_height())
    i=0
    b=int(len(y)/2) ## The step is dfferent than simple Facetgrid plots (dis,␣
↪rel,etc.)
    for p in ax.patches:
        if i>=b:
            k=y[i]/(y[i]+y[i-b])*100
```

```
        else:
            k=y[i]/(y[i]+y[i+b])*100
        if math.isnan(k):
            k=100
        ax.annotate(f' {k:.0f}%', xy = (p.get_x()+p.get_width()/ 2, p.
↪get_height()),
                       ha='center',
                       va='center',
                       size=18,
                       xytext=(0, 8),
                       textcoords='offset points'
                   )
        i=i+1
g=sns.FacetGrid(data=employee_df1, col="JobRole", col_wrap=3, height=8)
g.map_dataframe(sns.countplot, x='Education', hue="Attrition",palette="winter" )
g.map_dataframe(text2,'Education')
txt="Figure 22: Percentage comparision of attrition for different job roles for␣
↪different level of education"
plt.figtext(0.5, -0.03, txt, wrap=True, horizontalalignment='center',␣
↪fontsize=20);
```

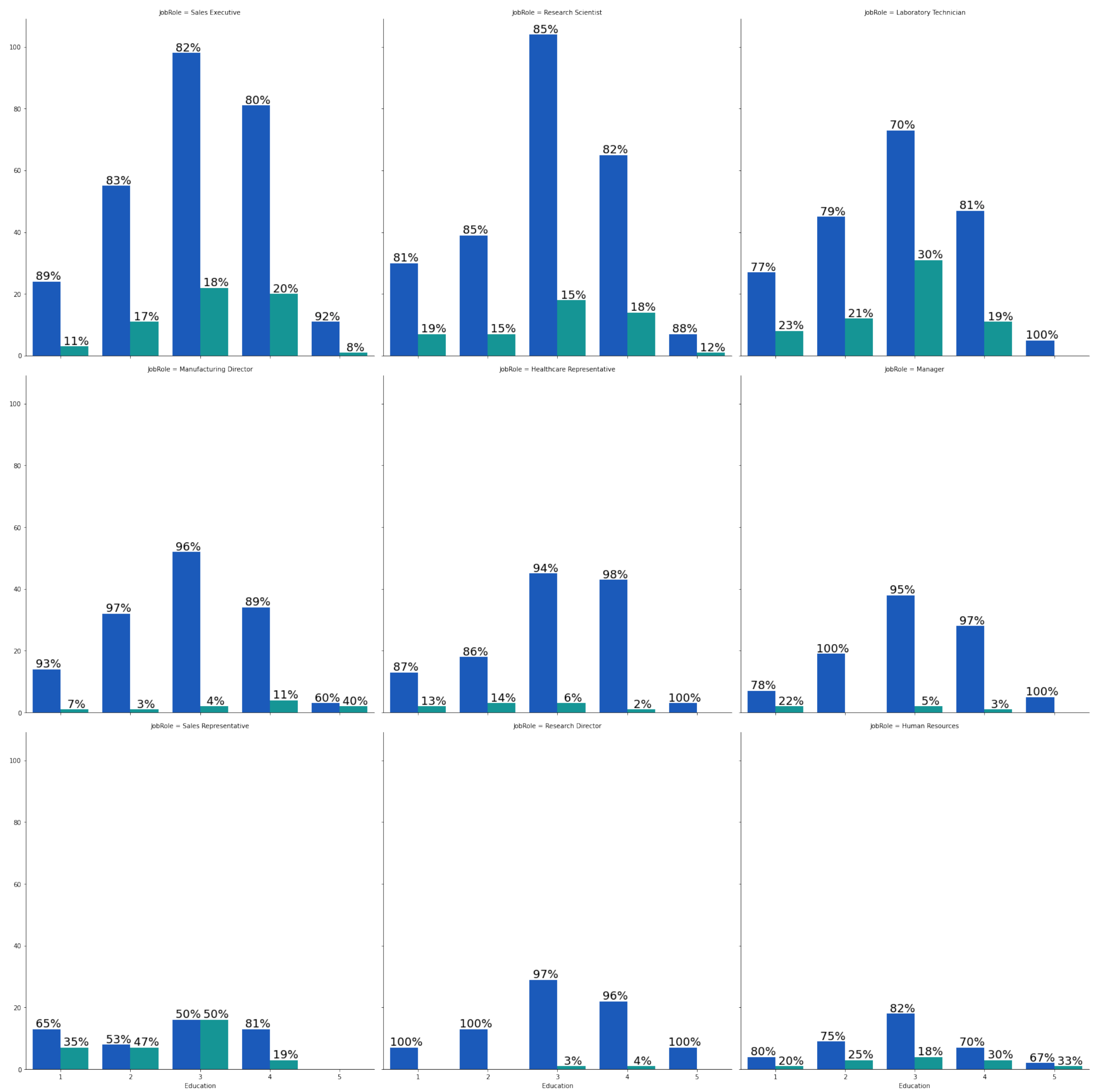


Figure 22: Percentage comparision of attrition for different job roles for different level of education

It can be observed that the in Human Resources the high level (4, 5) educated employees have higher percentage of attrition. Also the manufacturing directors with level 5 education have greater chances for attrition. For other roles the employees look satisfied with their postion regardless of their education. We can also observed that research directors have very low chances of attrition. This is may most of them are older and does not want to change the job.

[55]:
```
import scipy
import random

plt.figure(figsize=(20,10))
ax=sns.kdeplot(data=employee_df1,x='Age', hue="JobRole",lw=3, palette="Paired")
```

```
for i in range(len(ax.lines)):
#     r = lambda: random.randint(0,255)
#     c = '#%02X%02X%02X' % (r(),r(),r())
    kdeline = ax.lines[i]
    x_points = kdeline.get_xdata()
    y_points = kdeline.get_ydata()
    mean=np.sum(np.multiply(x_points,y_points))/np.sum(y_points)
    height = np.interp(mean, x_points, y_points)
    ax.vlines(mean, 0, height, color="gray",ls="--", lw=4)
    ax.fill_between(x_points, 0, y_points, facecolor='green', alpha=0.1)

txt="Figure 23: Kde plot for age for different job roles. The vertical line indicate the average age for each role."
plt.figtext(0.5, -0.03, txt, wrap=True, horizontalalignment='center', fontsize=17);
```

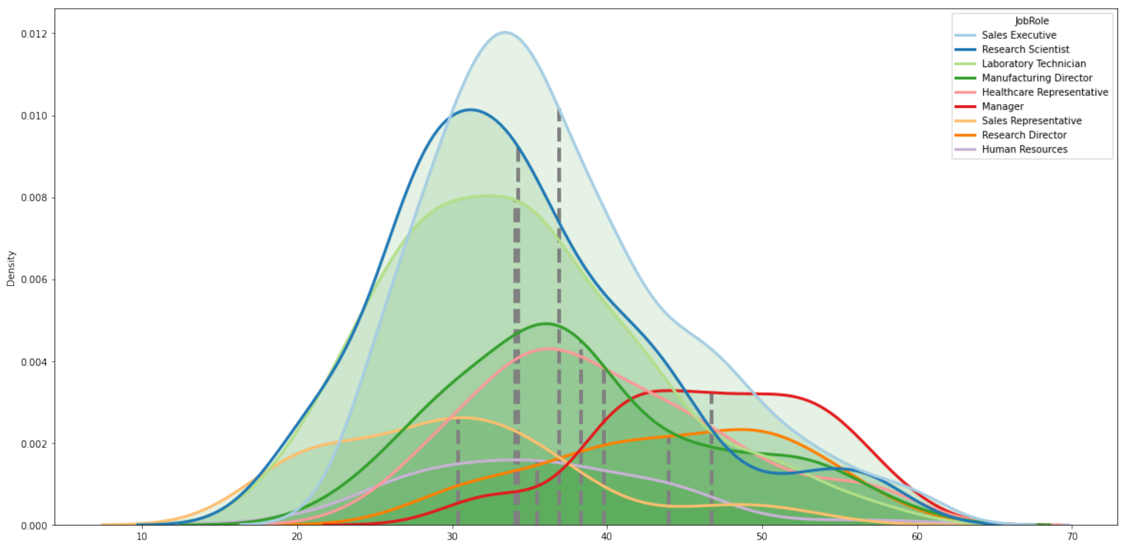


Figure 23: Kde plot for age for different job roles. The vertical line indicate the average age for each role.

Since the average age for Research directors and Managers are approximately 44 years and 47 years and they have the lowest chances of attrition. May be the age played an imporant role in the attrition. Lets check for managers and research directors

[56]:
```
df=employee_df1[(employee_df1['JobRole']=="Manager") | (employee_df1['JobRole']=="Research Director")]
df=df[['Attrition','Age','JobRole']]
df
```

[56]:
```
      Attrition  Age            JobRole
18            0   53            Manager
22            0   34  Research Director
25            0   53            Manager
29            0   46            Manager
45            1   41  Research Director
…           …   …                 …
1421          0   47  Research Director
1430          0   38  Research Director
1432          0   37  Research Director
1437          0   39            Manager
1443          0   42            Manager

[182 rows x 3 columns]
```

[57]:
```python
plt.figure(figsize=(20,10))
df1=df.groupby(['Attrition','Age']).count()
df1.plot(kind='bar', figsize=(20,10));
txt="Figure 24: Number of stayed and left employees for different ages"
plt.figtext(0.5, -0.03, txt, wrap=True, horizontalalignment='center',␣
 ↪fontsize=17);
```

```
<Figure size 1440x720 with 0 Axes>
```

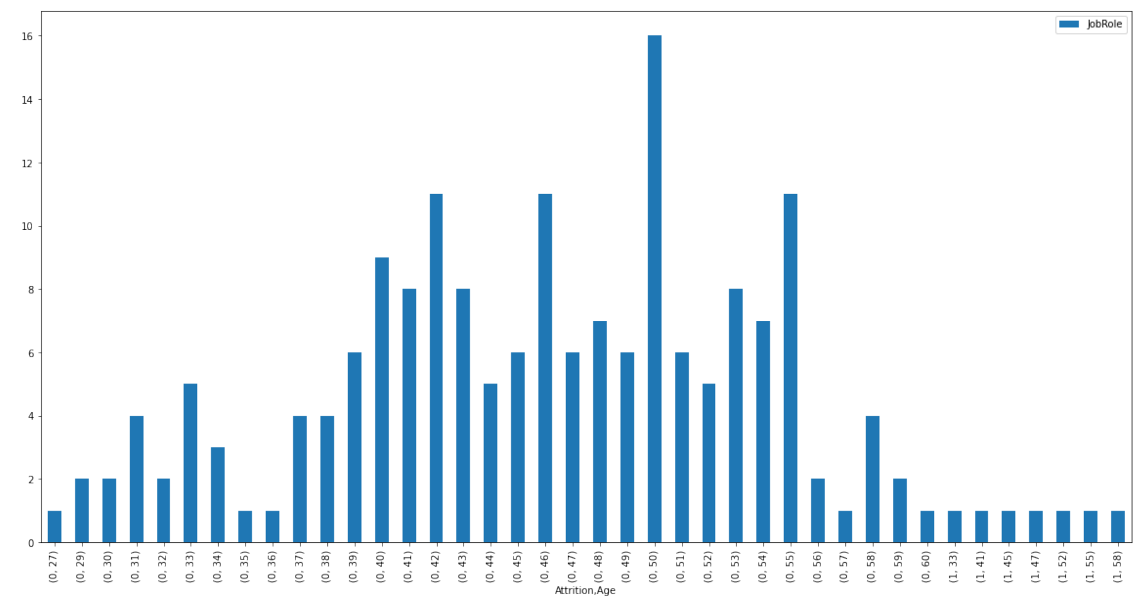


Figure 24: Number of stayed and left employees for different ages

So the number of left managers and research direcotrs is very low and we cannot say that the attrition is only because of the age. We have actually found less attrition in the lower age group in these Job Roles. The attrition in these Job roles may be less

because these are the higher positions and there are less opportunities.

[58]:
```
plt.figure(figsize=(20,10))
ax=sns.countplot(data=employee_df1, x="EnvironmentSatisfaction",␣
↪hue="Attrition")
y=[]
for i in ax.patches:
    y.append(i.get_height())
i=0
b=4
for p in ax.patches:
    if i>=b:
        k=y[i]/(y[i]+y[i-b])*100
    else:
        k=y[i]/(y[i]+y[i+b])*100

    ax.annotate(f' {k:.0f}%', xy = (p.get_x()+p.get_width()/ 2, p.get_height()),
                ha='center',
                va='center',
                size=13,
                xytext=(0, 8),
                textcoords='offset points'
            )
    i=i+1
txt="Figure 25: Percentage of stayed and left employees for different level of␣
↪environmental stisfaction"
plt.figtext(0.5, -0.03, txt, wrap=True, horizontalalignment='center',␣
↪fontsize=17);
```

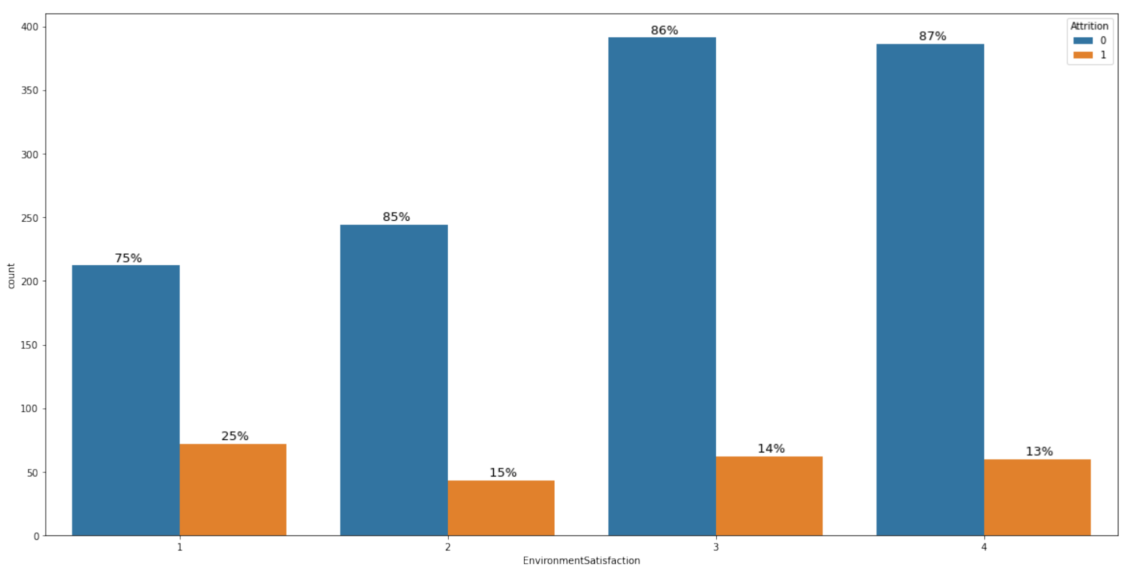


Figure 25: Percentage of stayed and left employees for different level of environmental stisfaction

It can be observed that the environmental satisfaction is also a factor responsible for attrition. We can observed the decreasing trend of attrition with increasing level of environmental satisfaction.

[59]:
```
plt.figure(figsize=(20,20))
plt.subplot(221)
sns.lineplot(data=employee_df1, x="JobLevel", y="EnvironmentSatisfaction")
plt.legend("A",loc="upper right", fontsize=17)
plt.subplot(222)
plt.xticks(rotation=45)
sns.lineplot(data=employee_df1, x="JobRole", y="EnvironmentSatisfaction")
plt.legend("B",loc="upper right",fontsize=17)
plt.subplot(223)
sns.lineplot(data=employee_df1, x="EducationField", y="EnvironmentSatisfaction")
plt.legend("C",loc="upper right", fontsize=17)
plt.xticks(rotation=45)
plt.subplot(224)
sns.lineplot(data=employee_df1, x="Education", y="EnvironmentSatisfaction")
plt.legend("D",loc="upper right", fontsize=17)
plt.tight_layout()

txt="Figure 26: Environmental satisfaction as fuction of A) Job Level, B)␣
↪JobRole, C) Education Field and D) Dducation"
plt.figtext(0.5, -0.03, txt, wrap=True, horizontalalignment='center',␣
↪fontsize=17);
```

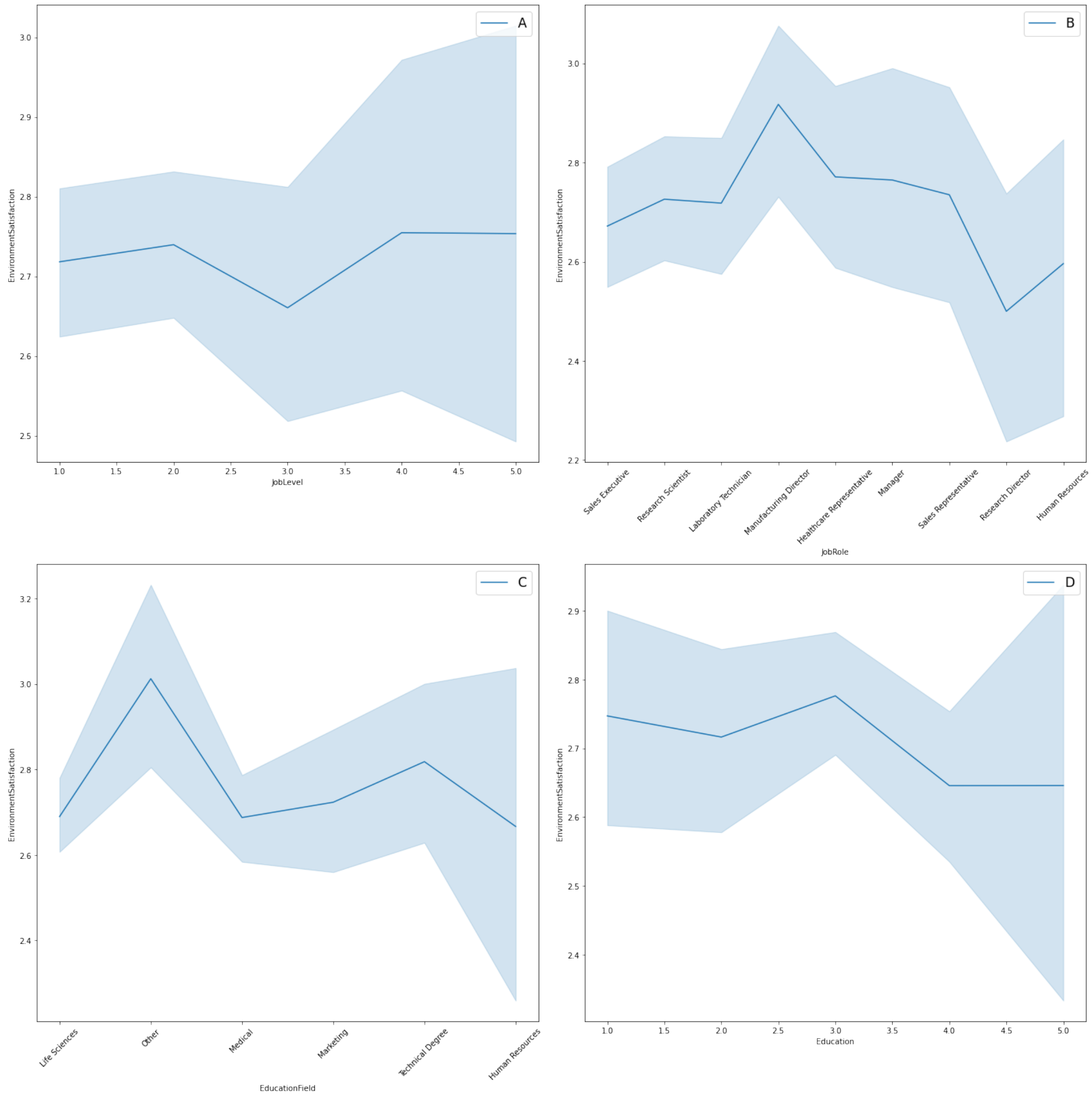


Figure 26: Environmental satisfaction as fuction of A) Job Level, B) JobRole, C) Education Field and D) Dducation

It can observed that the research directors are among those who have lowest envirmonmental satisafction and Manufactoring directors are highly satisfied by their working envirmonemnts. Simlarly highly educated employees are less satisfied by their environment as compared others. Also employees who are educated in the human resources are less satisfied with the environment. May be they have more human interactions and that make them think like that.

[60]:
```
plt.figure(figsize=(20,10))
ax=sns.countplot(data=employee_df1, x="Gender", hue="Attrition")
y=[]
for i in ax.patches:
    y.append(i.get_height())
```

```
i=0
b=2
for p in ax.patches:
    if i>=b:
        k=y[i]/(y[i]+y[i-b])*100
    else:
        k=y[i]/(y[i]+y[i+b])*100

    ax.annotate(f' {k:.0f}%', xy = (p.get_x()+p.get_width()/ 2, p.get_height()),
                    ha='center',
                    va='center',
                    size=13,
                    xytext=(0, 8),
                    textcoords='offset points'
                )
    i=i+1
txt="Figure 27: Attrition among different genders"
plt.figtext(0.5, -0.03, txt, wrap=True, horizontalalignment='center',␣
 ↪fontsize=17);
```

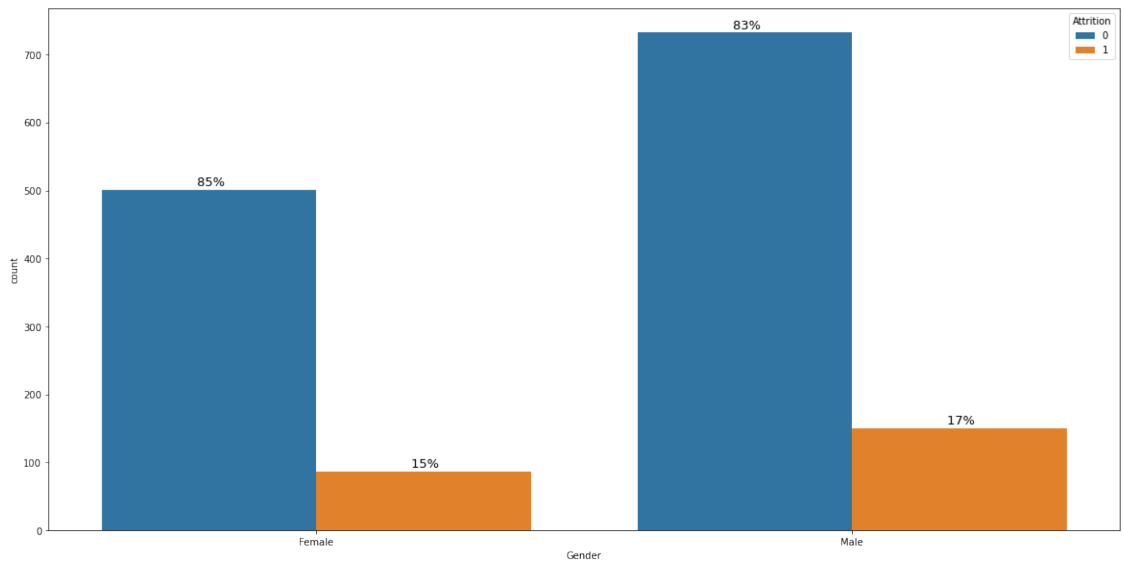


Figure 27: Attrition among different genders

It can be observed that the percentage of attrition is almost same for male and female, even the female attrition is less. It looks the company giving equal opportunity to males and females.

[61]:
```
plt.figure(figsize=(20,10))
ax=sns.countplot(data=employee_df1, x="JobInvolvement", hue="Attrition")
```

```
y=[]
for i in ax.patches:
    y.append(i.get_height())

i=0
b=4
for p in ax.patches:
    if i>=b:
        k=y[i]/(y[i]+y[i-b])*100
    else:
        k=y[i]/(y[i]+y[i+b])*100

    ax.annotate(f' {k:.0f}%', xy = (p.get_x()+p.get_width()/ 2, p.get_height()),
                   ha='center',
                   va='center',
                   size=13,
                   xytext=(0, 8),
                   textcoords='offset points'
               )
    i=i+1
txt="Figure 28: Attrition vs Job involvment"
plt.figtext(0.5, -0.03, txt, wrap=True, horizontalalignment='center',␣
 ↪fontsize=17);
```

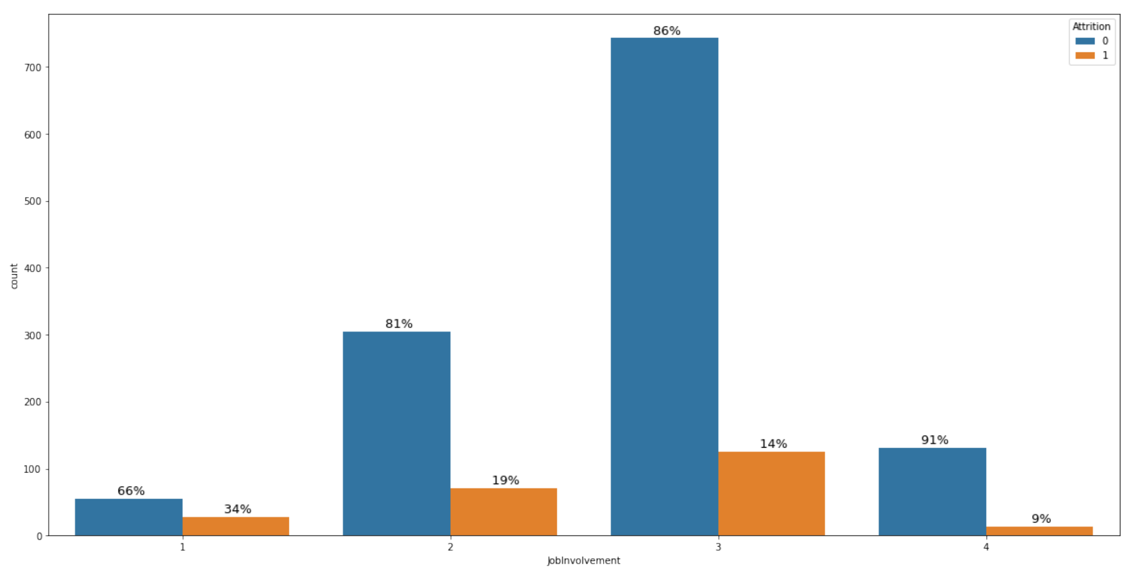


Figure 28: Attrition vs Job involvment

The bargraph shows the attrition rate decreasing with the increase in job involvement.The job involvement has some level of correlation with Education which we can

explore.

[62]:
```
g=sns.catplot(data=employee_df1, x="JobInvolvement", col_wrap=3,␣
↪col='Education', hue="Attrition", kind="count", palette="viridis_r")

y=[]
for ax in g.axes.ravel():
    for i in ax.patches:
          y.append(i.get_height())
i=0
b=4
l=4

for ax in g.axes.ravel():
    for p in ax.patches:
        if i>=l:
            k=y[i]/(y[i]+y[i-b])*100
        else:
            k=y[i]/(y[i]+y[i+b])*100
            v=1
        if math.isnan(k):
            k=100

        ax.annotate(f' {k:.0f}%', xy = (p.get_x()+p.get_width()/ 2, p.
↪get_height()),
                       ha='center',
                       va='center',
                       size=13,
                       xytext=(0, 8),
                       textcoords='offset points'
                   )

        i=i+1
    l=l+2*b
txt="Figure 29: percentage of stayed vs feft employees with different level of␣
↪Job involvment for different education levels "
plt.figtext(0.5, -0.04, txt, wrap=True, horizontalalignment='center',␣
↪fontsize=17);
```

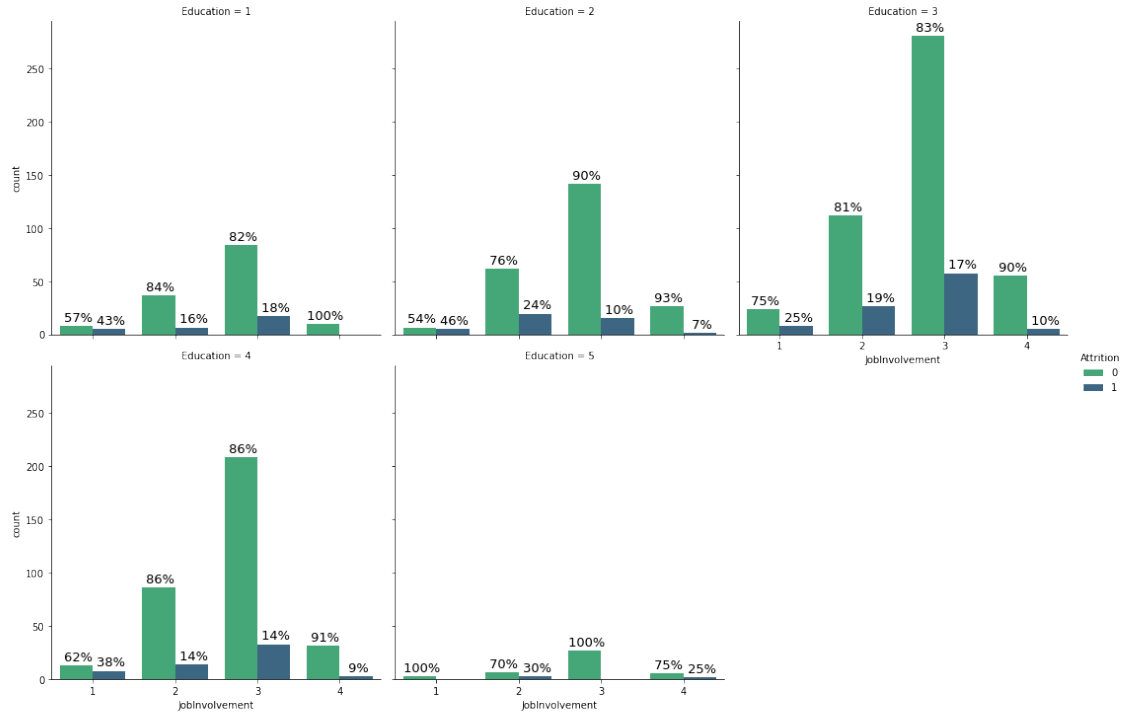


Figure 29: percentage of stayed vs feft employees with different level of Job involvment for different education levels

It can be observed that attrition is more for lower level of job involments for all level of education except level 5.

[63]:
```
plt.figure(figsize=(20,10))
ax=sns.countplot(data=employee_df1, x="OverTime", hue="Attrition")
y=[]
for i in ax.patches:
    y.append(i.get_height())
i=0
b=2
for p in ax.patches:
    if i>=b:
        k=y[i]/(y[i]+y[i-b])*100
    else:
        k=y[i]/(y[i]+y[i+b])*100

    ax.annotate(f' {k:.0f}%', xy = (p.get_x()+p.get_width()/ 2, p.get_height()),
                ha='center',
                va='center',
                size=13,
                xytext=(0, 8),
                textcoords='offset points'
            )
    i=i+1
```

```
txt="Figure 30: Ration of atrition among overtime vs non-overtime emplyess"
plt.figtext(0.5, -0.04, txt, wrap=True, horizontalalignment='center',
 ↪fontsize=17);
```

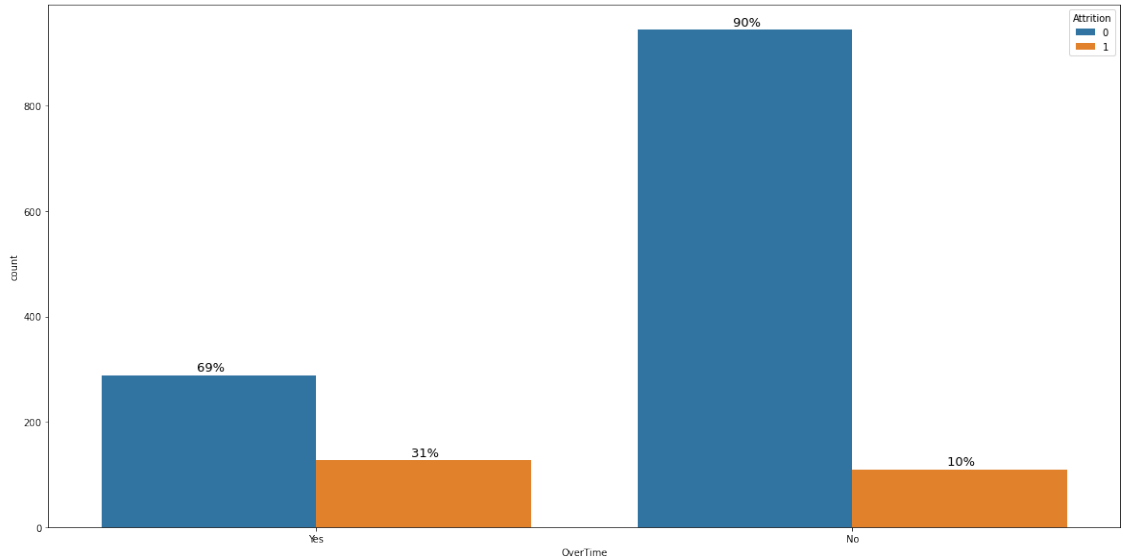


Figure 30: Ration of atrition among overtime vs non-overtime emplyess

It is clear that the employees dont like to work for overtime. The attrition percentage of emplyees who woking overtime are more as compared to those who dont work overtime. We can search who are the emplyees who are doing these overtime.

[64]:
```
plt.figure(figsize=(20,10))
w = 2000
b = math.ceil((employee_df1["MonthlyIncome"].max() -
 ↪employee_df1["MonthlyIncome"].min())/w)
ax=sns.histplot(data=employee_df1, x='MonthlyIncome', hue='OverTime',bins=b, );
y=[]
for i in ax.patches:
      x=i.get_height()
      y.append(x)
i=0
v=3
h=0
for p in ax.patches:
    if i>=b:
        k=y[i]/(y[i]+y[i-b])*100
        v=5
        h=1
    else:
```

```
        k=y[i]/(y[i]+y[i+b])*100
        t=1


    ax.annotate(f' {k:.2f} %', xy = (p.get_x()+p.get_width()/ 2+h, p.
 ↪get_height()+v),
                   ha='center',
                   va='center',
                   size=14,
                   xytext=(0, 8),
                   textcoords='offset points'
                )
    i=i+1
txt="Figure 31: Percentage comparision of Oovertime vs non-overtime employees␣
 ↪using histogram for different monthly income groups"
plt.figtext(0.5, -0.04, txt, wrap=True, horizontalalignment='center',␣
 ↪fontsize=17);
```

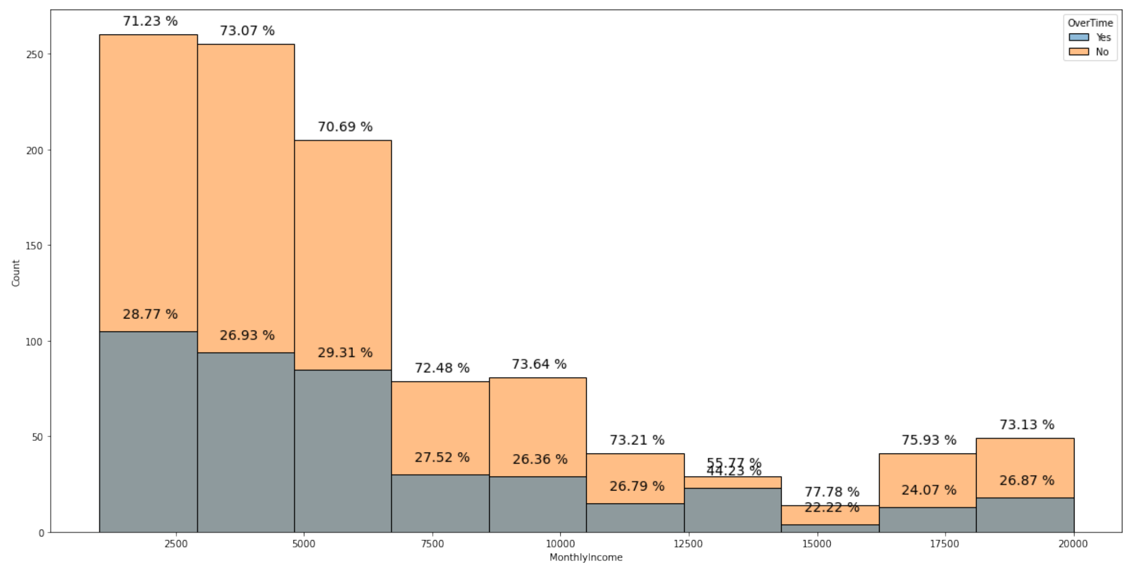


Figure 31: Percentage comparision of Oovertime vs non-overtime employees using histogram for different monthly income groups

The overtime percentage is almost same in all income regions. However, we can noticed that about half of the emmplyees who are have salary between 12000 to 14000 range are doing overtime.

[65]:
```
fig, ax=plt.subplots(figsize=(10,10))
sns.boxplot(x=employee_df1['MonthlyIncome'], y=employee_df1['JobRole'])

txt="Figure 32:Monthly income range for different job roles"
```

```
plt.figtext(0.5, 0.01, txt, wrap=True, horizontalalignment='center',␣
↪fontsize=17);
```

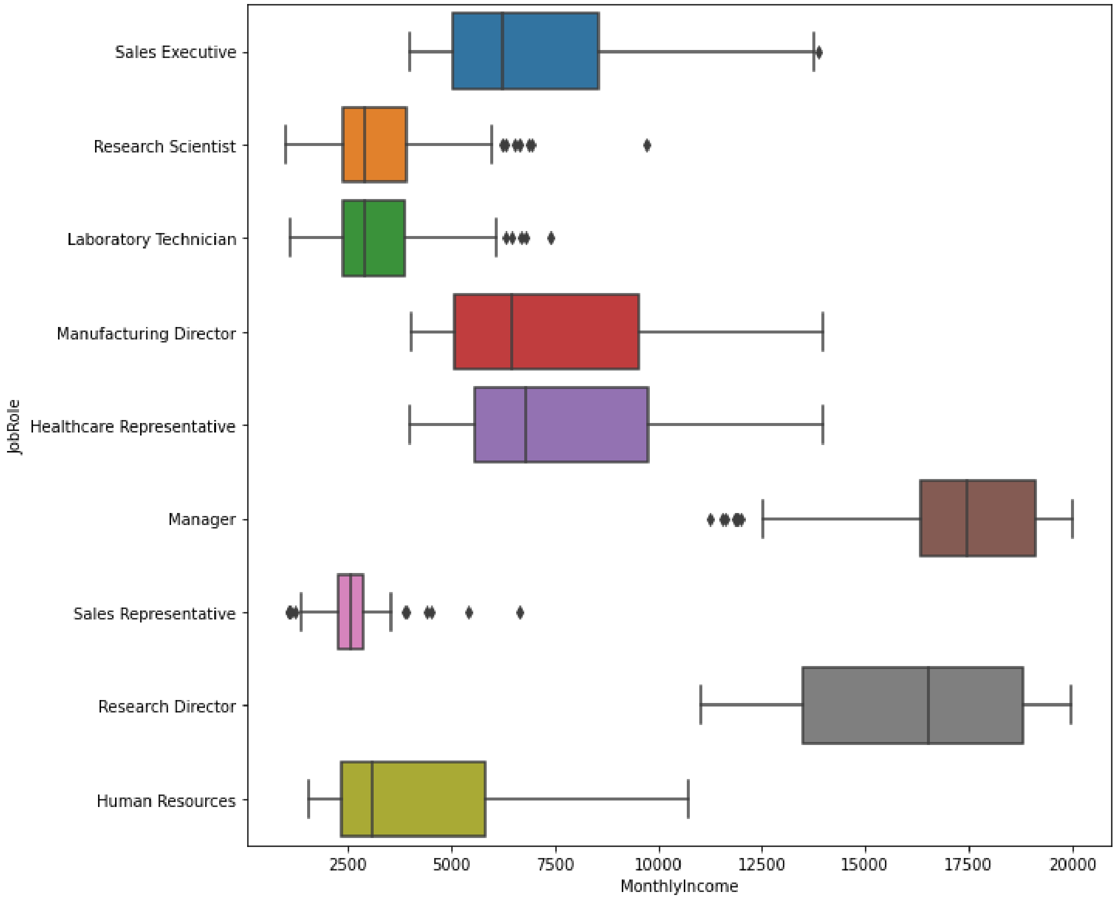


Figure 32:Monthly income range for different job roles

So we look in the salary we can say that the research Directors might need to do more the overtime as most of them are among 12000 to 14000. Also they have the lowest level of environmental satisfaction.*

[66]:
```
plt.figure(figsize=(20,10))
ax=sns.histplot(data=employee_df1, x="JobRole", hue="OverTime")
y=[]
for i in ax.patches:
      y.append(i.get_height())
i=0
b=9
for p in ax.patches:
    if i>=b:
        k=y[i]/(y[i]+y[i-b])*100
```

```
    else:
        k=y[i]/(y[i]+y[i+b])*100
    ax.annotate(f' {k:.2f} %', xy = (p.get_x()+p.get_width()/ 2, p.
 ↪get_height()),
                    ha='center',
                    va='center',
                    size=14,
                    xytext=(0, 8),
                    textcoords='offset points'
                )
    i=i+1
txt="Figure 33: Overtime vs non overtime percentage of employees on the basis␣
 ↪of their job roles"
plt.figtext(0.5, 0.01, txt, wrap=True, horizontalalignment='center',␣
 ↪fontsize=17);
```

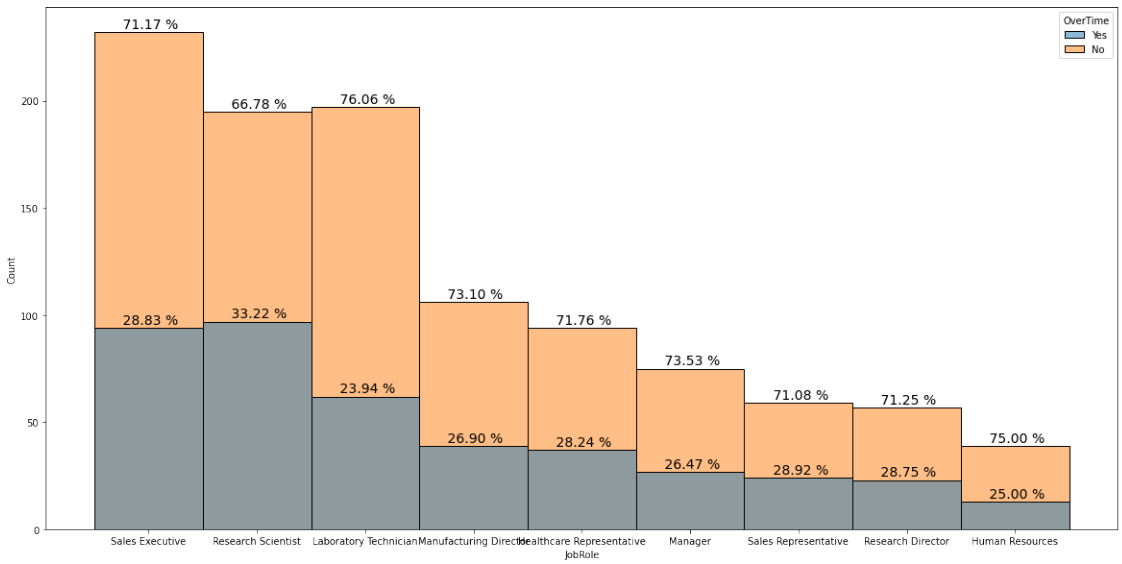


Figure 33: Overtime vs non overtime percentage of employees on the basis of their job roles

We can identifiy the employees who are more involved in overtime from their monthly income range but not on the bases of the Job Roles. From job role it can be seen that Research Scientist have the higher percentage of employees who are doing overtime. However, overtime is almost same for each role within the range of 10% differences.

[67]:
```
plt.figure(figsize=(20,10))
ax=sns.countplot(data=employee_df1, x="JobSatisfaction", hue="Attrition")
y=[]
for i in ax.patches:
      y.append(i.get_height())
i=0
```

```
b=4
for p in ax.patches:
    if i>=b:
        k=y[i]/(y[i]+y[i-b])*100
    else:
        k=y[i]/(y[i]+y[i+b])*100
    ax.annotate(f' {k:.2f} %', xy = (p.get_x()+p.get_width()/ 2, p.
 ↪get_height()),
                    ha='center',
                    va='center',
                    size=14,
                    xytext=(0, 8),
                    textcoords='offset points'
                )
    i=i+1

txt="Figure 34: Percentage attrition of employees for different Job␣
 ↪Satisfaction Levels"
plt.figtext(0.5, 0.01, txt, wrap=True, horizontalalignment='center',␣
 ↪fontsize=17);
```

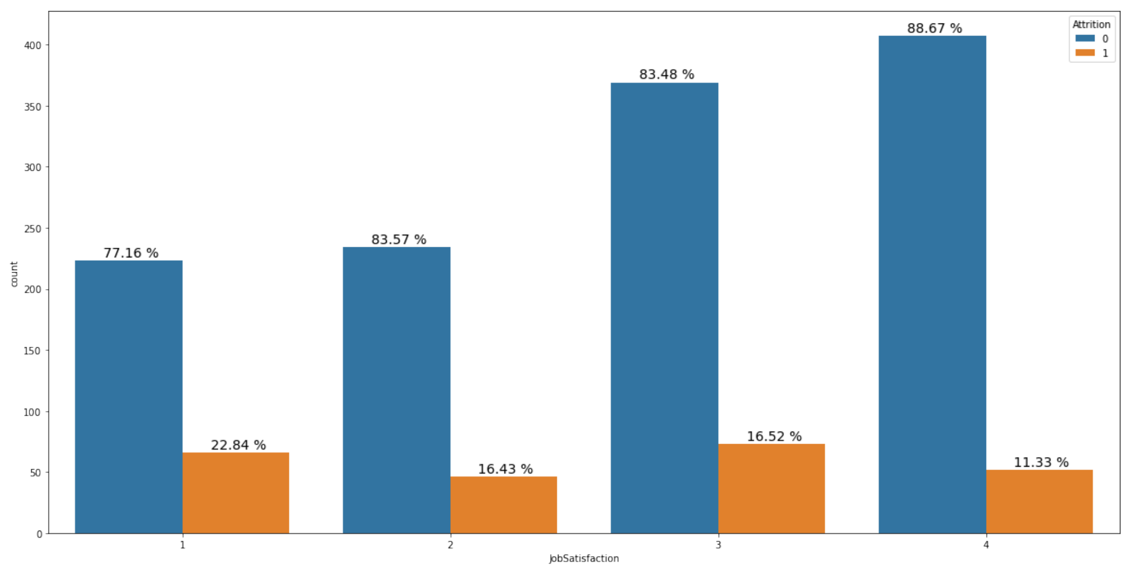


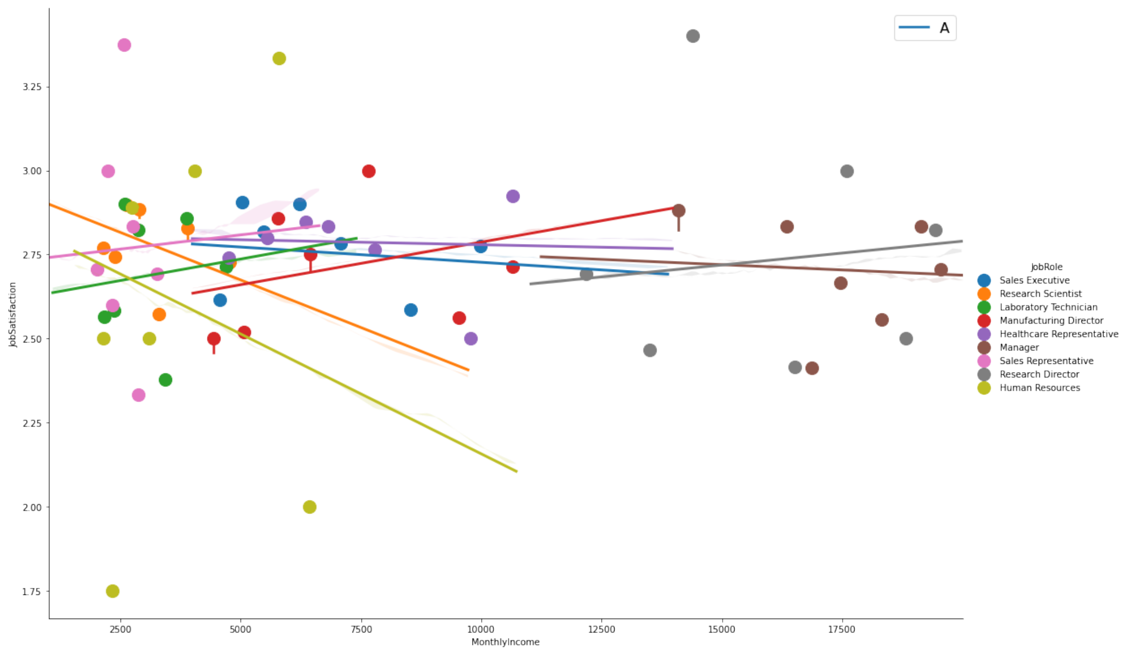
Figure 34: Percentage attrition of employees for different Job Satisfaction Levels

The job satisafaction also contribute for attrition. The employees with Jobsatiasaction level "1", have higher chances to left hte company. The corellation matrix also shows that the Job Satisfaction is negatively correlated with Attrition. From correlation matrix, we have not observed any correlation between JobSatisfaction and any other parameter.

[68]:

```
sns.lmplot(x='MonthlyIncome', y='JobSatisfaction',
           hue='JobRole', data=employee_df,
           x_bins=7, ci=1,robust=True,n_boot=100, height=10,
           aspect=1.5,
           scatter_kws={"s": 200},
           line_kws={'lw': 3})
plt.legend("A",loc="upper right", fontsize=16)
sns.lmplot(x='MonthlyIncome', y='JobSatisfaction',
           hue='JobLevel', data=employee_df, height=8,
           aspect=1.5,
            x_bins=7, n_boot=100,
           scatter_kws={"s": 300},
           line_kws={'lw': 2})
plt.legend("B",loc="upper right", fontsize=16)

txt="Figure 35: LM plots between monthly income and job satisfaction levels for␣
↪different A) job roles B) Job Levels"
plt.figtext(0.5, -0.04, txt, wrap=True, horizontalalignment='center',␣
↪fontsize=17);
```



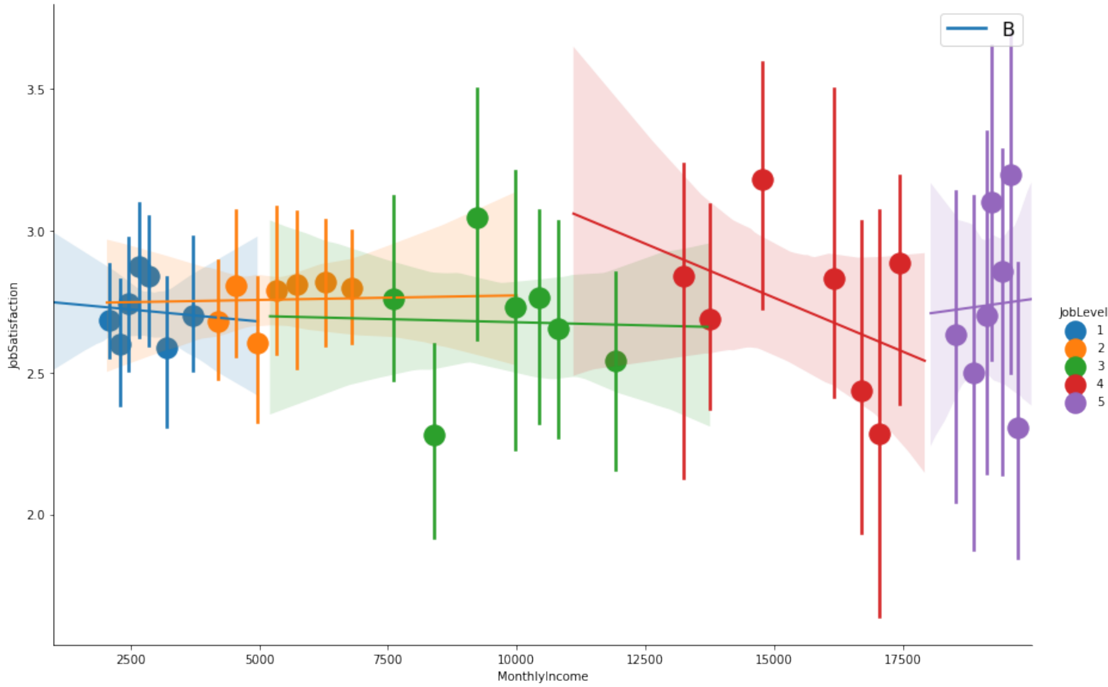


Figure 35: LM plots between monthly income and job satisfaction levels for different A) job roles B) Job Levels

It can be observed that the job satisfaction decresing for the Human resource and research scientist with the increase in monthly income. Also employees at joblevel 4 are getting less satisfied with the increase in monthly income.

[69]:
```
plt.figure(figsize=(20,10))
w = 2500
b = math.ceil((employee_df1["MonthlyIncome"].max() -␣
↪employee_df1["MonthlyIncome"].min())/w)
ax=sns.histplot(data=employee_df1, x='MonthlyIncome', hue='Attrition',bins=b, );
y=[]
for i in ax.patches:
      y.append(i.get_height())
i=0
v=3
h=0
for p in ax.patches:
    if i>=b:
        k=y[i]/(y[i]+y[i-b])*100
        v=5
        h=1
    else:
        k=y[i]/(y[i]+y[i+b])*100
        t=1
```

```
    ax.annotate(f' {k:.2f} %', xy = (p.get_x()+p.get_width()/ 2+h, p.
 ↪get_height()+v),
                    ha='center',
                    va='center',
                    size=14,
                    xytext=(0, 8),
                    textcoords='offset points'
                )
    i=i+1

txt="Figure 36: percentage attrition for different monthly  income ranges"
plt.figtext(0.5, 0.02, txt, wrap=True, horizontalalignment='center',␣
 ↪fontsize=17);
```

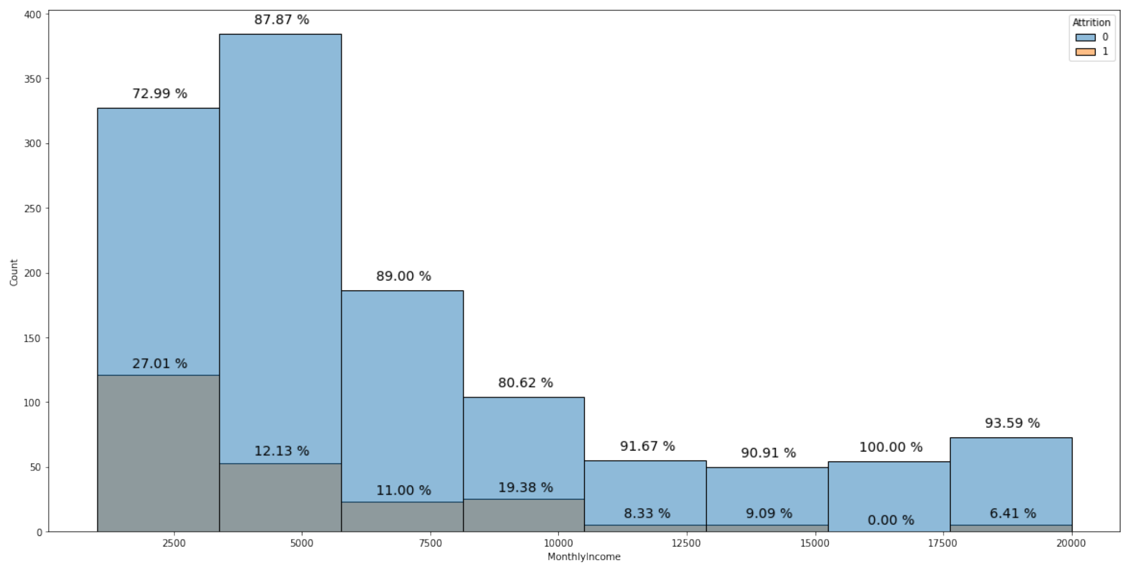


Figure 36: percentage attrition for different monthly income ranges

The above plot shows a downword trend of atrition with the increase of monthlyIncome with some exeptions. Employees with monthly income lower 3300 have higher chances to left the company. Mainly human resource, sales executive, research scintist and laboratory scientist fall under this weges group.

[70]:
```
fig, ax=plt.subplots(figsize=(10,10))
sns.lineplot(x=employee_df1['NumCompaniesWorked'], y=employee_df1['Attrition'])
txt="Figure 37: Plot between number of companies worked and attriton"
plt.figtext(0.5, 0.02, txt, wrap=True, horizontalalignment='center',␣
 ↪fontsize=17);
```

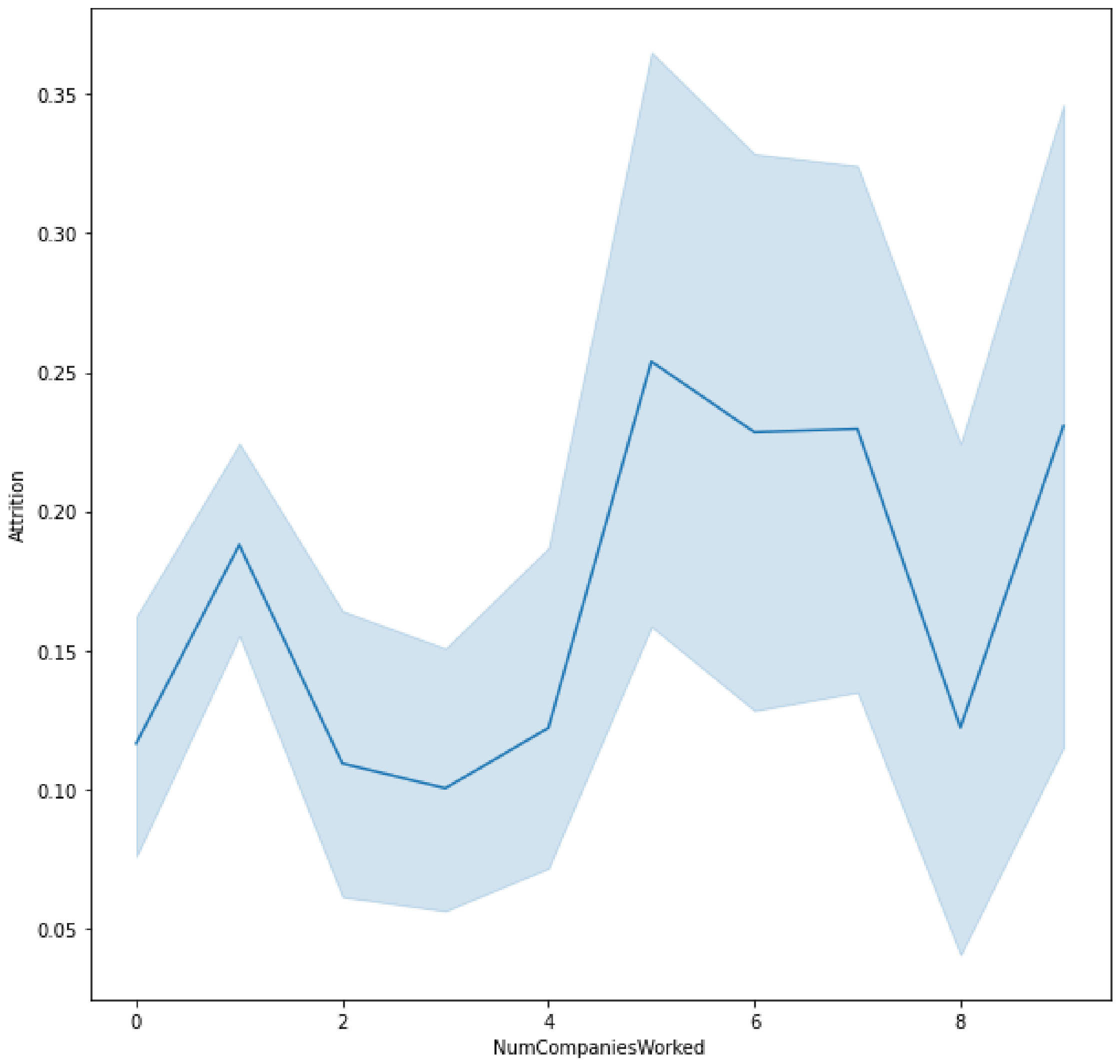


Figure 37: Plot between number of companies worked and attriton

The attrition is increasing with the number of companies workeed but it is not a straight line.

[71]:
```
plt.figure(figsize=(20,10))
w = 1
b = math.ceil((employee_df1["PercentSalaryHike"].max() -
 ↪employee_df1["PercentSalaryHike"].min())/w)
ax=sns.histplot(data=employee_df1, x='PercentSalaryHike',
 ↪hue='Attrition',bins=b, );
y=[]
for i in ax.patches:
      y.append(i.get_height())
```

```
i=0
v=3
h=0
for p in ax.patches:
    if i>=b:
        k=y[i]/(y[i]+y[i-b])*100
        v=5
        h=1
    else:
        k=y[i]/(y[i]+y[i+b])*100
        t=1


    ax.annotate(f' {k:.2f} %', xy = (p.get_x()+p.get_width()/ 2+h, p.
 ↪get_height()+v),
                    ha='center',
                    va='center',
                    size=14,
                    xytext=(0, 8),
                    textcoords='offset points'
                )
    i=i+1

txt="Figure 38: Plot between number of companies worked and attriton"
plt.figtext(0.5, 0.02, txt, wrap=True, horizontalalignment='center',␣
 ↪fontsize=17);
```

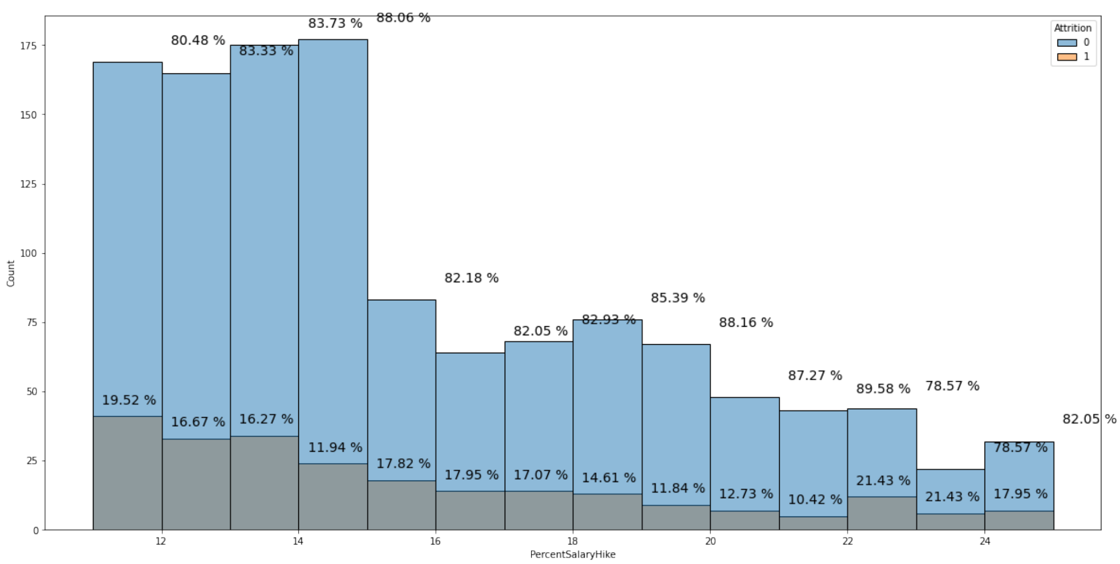


Figure 38: Plot between number of companies worked and attriton

No relation has been observed between PercentSalaryHike and attrition.So we can say that the emplyees who are leaving the company does not left because they want the salary hike. The comapny might providing the proper hike to the employees based on their performance and YearsOfWorking.

[72]:
```
plt.figure(figsize=(20,10))
ax=sns.displot(data=employee_df1, kind='hist', x='PercentSalaryHike',␣
↪hue="PerformanceRating", col="Attrition");
txt="Figure 39: Salary hike for diffent performance ratings for Stayed and left␣
↪employees"
plt.figtext(0.5, -0.04, txt, wrap=True, horizontalalignment='center',␣
↪fontsize=15);
```

```
<Figure size 1440x720 with 0 Axes>
```

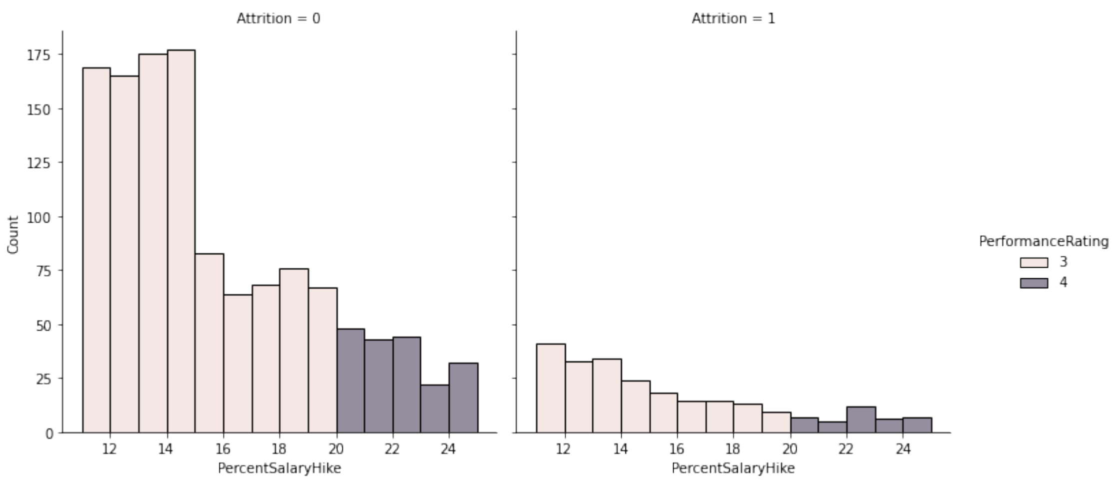


Figure 39: Salary hike for diffent performance ratings for Stayed and left employees

[73]:
```
plt.figure(figsize=(20,10))
ax=sns.relplot(data=employee_df1, kind='scatter', x='YearsAtCompany',␣
↪y="PercentSalaryHike", col="PerformanceRating");
txt="Figure 40: Percentage Salary hike for years in the comapny for diffent␣
↪performance ratings"
plt.figtext(0.5, -0.08 , txt, wrap=True, horizontalalignment='center',␣
↪fontsize=15);
```

```
<Figure size 1440x720 with 0 Axes>
```

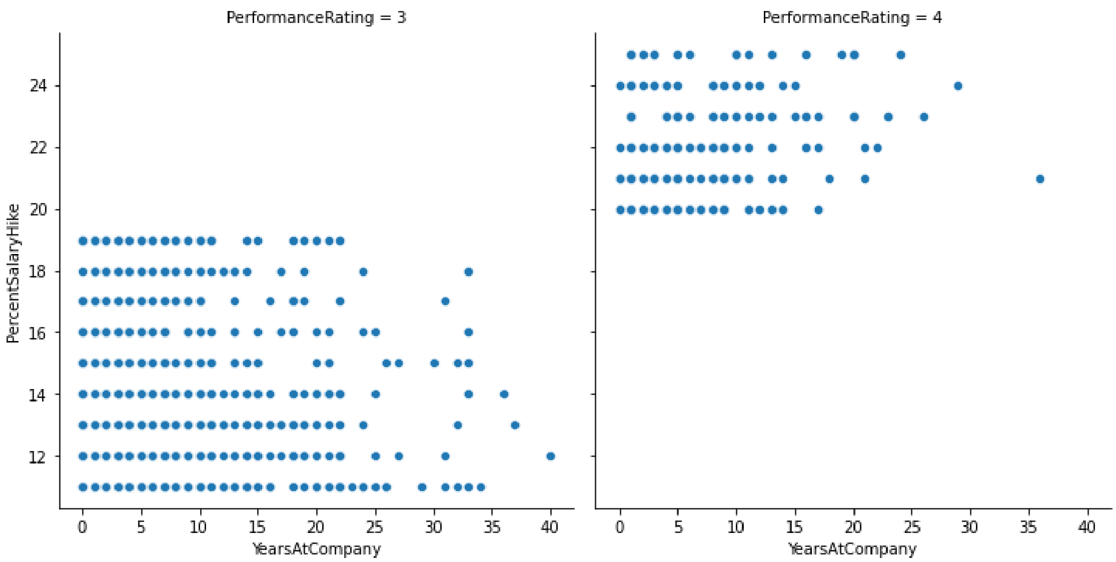


Figure 40: Percentage Salary hike for years in the comapny for diffent performance ratings

We can notice that the salary hike does not have any relation with the experince (YearsAtCompany). It is unform for all employees but have strong relation with performance rating.

[74]:
```
plt.figure(figsize=(15,20))
plt.subplot(211)
sns.boxplot(x=employee_df1['PercentSalaryHike'], y=employee_df1['JobRole'])
plt.legend("A",loc="upper right", fontsize=16)

plt.subplot(212)
sns.boxplot(x=employee_df1['PercentSalaryHike'], y=employee_df1['Department'])
plt.legend("B", fontsize=16)
txt="Figure 41: Percentage Salary hike for differetn A) Job Roles and B)␣
 ↪Departments"

plt.figtext(0.5,0.08 , txt, wrap=True, horizontalalignment='center',␣
 ↪fontsize=15);
```

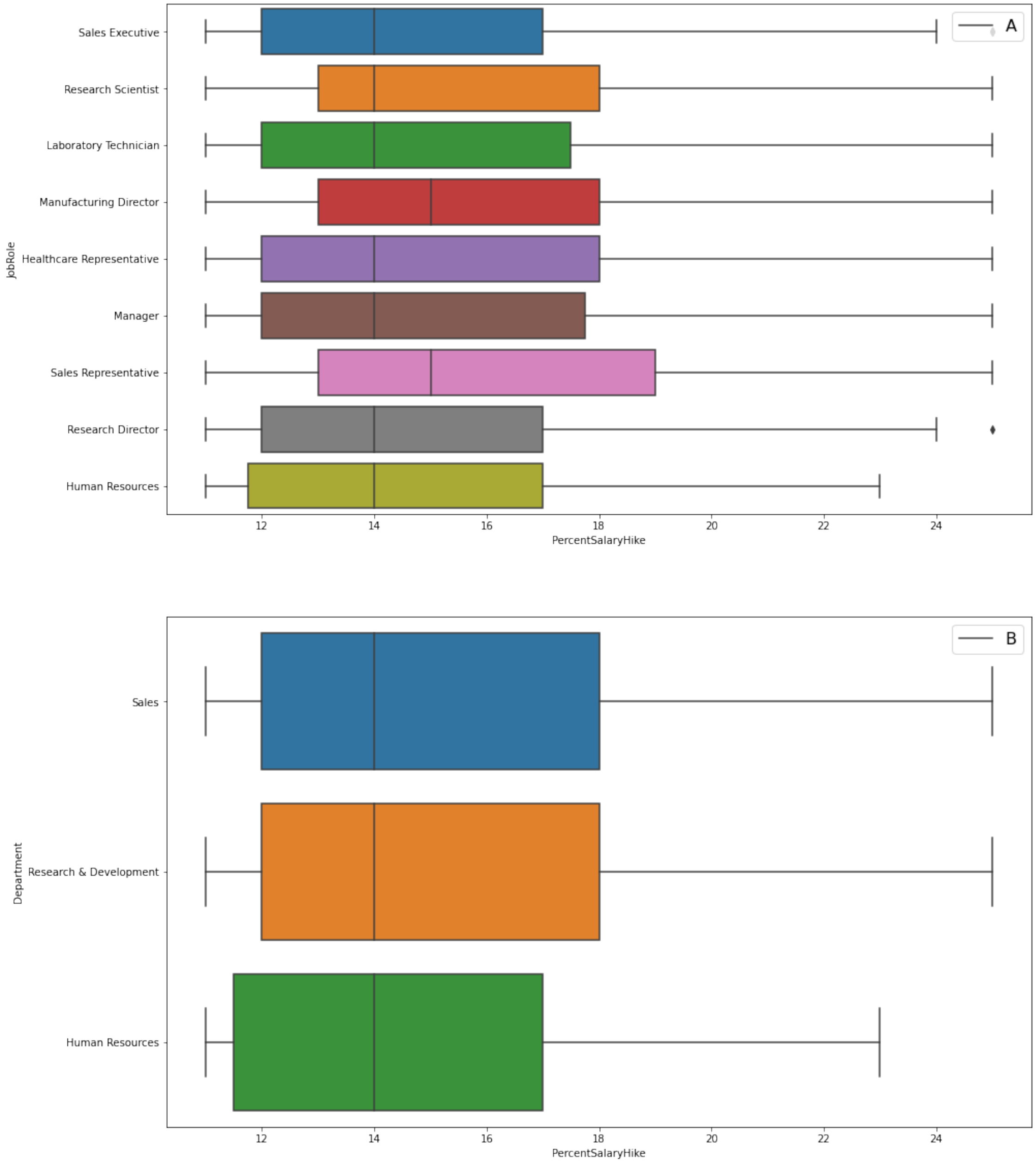


Figure 41: Percentage Salary hike for differetn A) Job Roles and B) Departments

The sales and research development has higer hike than human resource. Also sales representative hike is more than any other role. May be the hike is given to overcome high attrition among sales representative.

[75]:
```
plt.figure(figsize=(20,10))
ax=sns.countplot(data=employee_df1, x="RelationshipSatisfaction", hue="Attrition")
y=[]
```

```python
for i in ax.patches:
      y.append(i.get_height())
i=0
b=4
for p in ax.patches:
    if i>=b:
        k=y[i]/(y[i]+y[i-b])*100
    else:
        k=y[i]/(y[i]+y[i+b])*100
    ax.annotate(f' {k:.2f} %', xy = (p.get_x()+p.get_width()/ 2, p.
 ↪get_height()),
                   ha='center',
                   va='center',
                   size=14,
                   xytext=(0, 8),
                   textcoords='offset points'
                )
    i=i+1
txt="Figure 42: Attrition percentage on the among different levels of␣
 ↪relationship satisfaction"

plt.figtext(0.5,0.05 , txt, wrap=True, horizontalalignment='center',␣
 ↪fontsize=15);
```

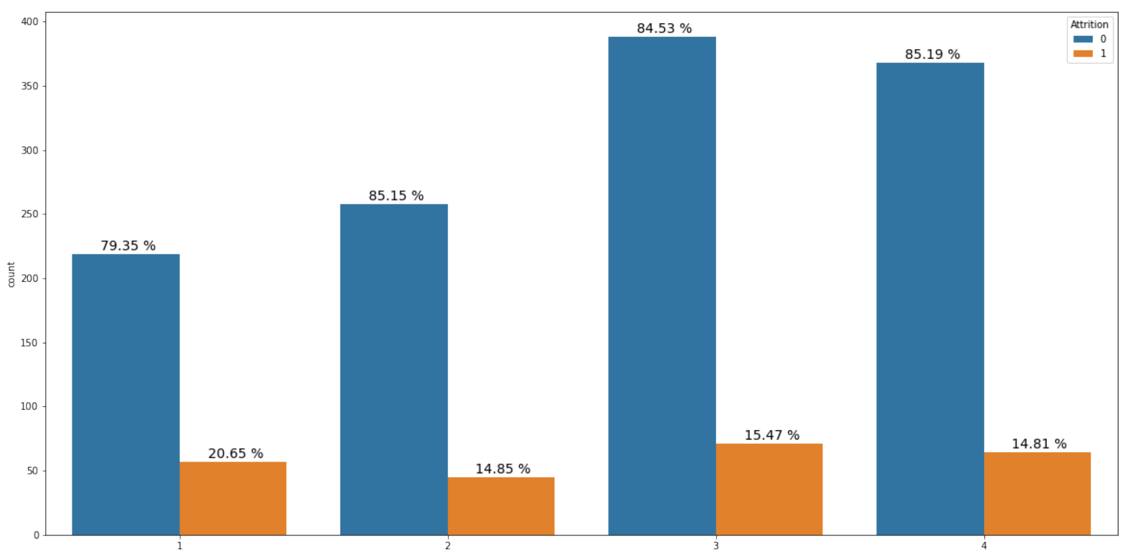


Figure 42: Attrition percentage on the among different levels of relationship satisfaction

Relationship satisfaction with level 1 has higher attrition rate but the difference is not much to be consider as a big factor for attrition.

[76]:

```
plt.figure(figsize=(20,10))
ax=sns.countplot(data=employee_df1, x="StockOptionLevel", hue="Attrition")
y=[]
for i in ax.patches:
      y.append(i.get_height())
i=0
b=4
for p in ax.patches:
    if i>=b:
        k=y[i]/(y[i]+y[i-b])*100
    else:
        k=y[i]/(y[i]+y[i+b])*100
    ax.annotate(f' {k:.2f} %', xy = (p.get_x()+p.get_width()/ 2, p.
 ↪get_height()),
                   ha='center',
                   va='center',
                   size=14,
                   xytext=(0, 8),
                   textcoords='offset points'
                )
    i=i+1

txt="Figure 43: Attrition percentage on the among different stock option levels"
plt.figtext(0.5,0.05 , txt, wrap=True, horizontalalignment='center',␣
 ↪fontsize=15);
```

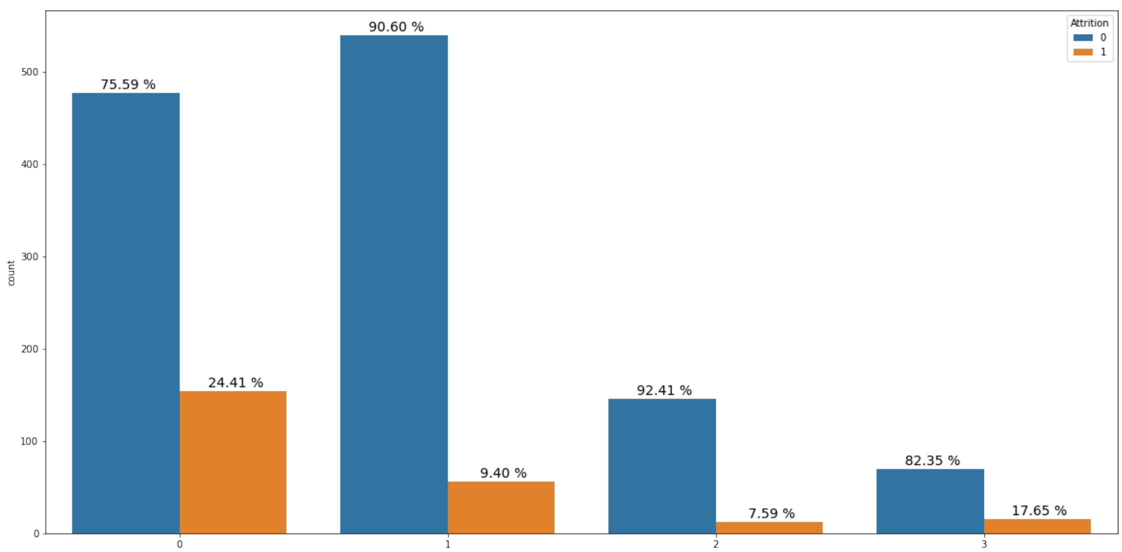


Figure 43: Attrition percentage on the among different stock option levels

Stock option level is playing a role in the attrition. The emplyees with level 0 have higher rate of attrition. The trend is not dicreasing for level 1 and 2 but for level 3

it again raised. But it can be said that providing stoock option will have effect on decrease the attrition.

[77]:
```
plt.figure(figsize=(20,20))
plt.subplot(221)
ax1=sns.countplot(data=employee_df1, x='Gender', hue='Gender')
y1=[]
for i in ax1.patches:
      y1.append(i.get_height())

for p in ax1.patches:
    ax1.annotate(f' {p.get_height():.2f}', xy = (p.get_x()+p.get_width()/ 2, p.
 ↪get_height()+17),
                   ha='center',
                   va='center',
                   size=14,
                   xytext=(0, 8),
                   textcoords='offset points'
                )
#plt.legend("A", fontsize=15)
ax1.legend(title="A", loc="upper left", labels=['Female', 'Male'])

plt.subplot(222)
ax2=sns.countplot(data=employee_df1, x='StockOptionLevel', hue='Gender', ␣
 ↪linewidth=2, edgecolor=(0,0,0))

y=[]
for i in ax.patches:
      y.append(i.get_height())
i=0
b=4
for p in ax.patches:
    if i>=b:
        k=y[i]/y1[1]*100
    else:
        k=y[i]/(y1[0])*100
    ax.annotate(f' {k:.2f} %', xy = (p.get_x()+p.get_width()/ 2, p.
 ↪get_height()),
                   ha='center',
                   va='center',
                   size=14,
                   xytext=(0, 8),
                   textcoords='offset points'
                )
    i=i+1
#plt.legend("B", fontsize=15)
ax2.legend(title="B", loc="upper right", labels=['Female', 'Male'])
```

```
txt="Figure 44: Male and females counts for A) total employees and B) different␣
↪stock level options"
plt.figtext(0.5, 0.45 , txt, wrap=True, horizontalalignment='center',␣
↪fontsize=15);
```

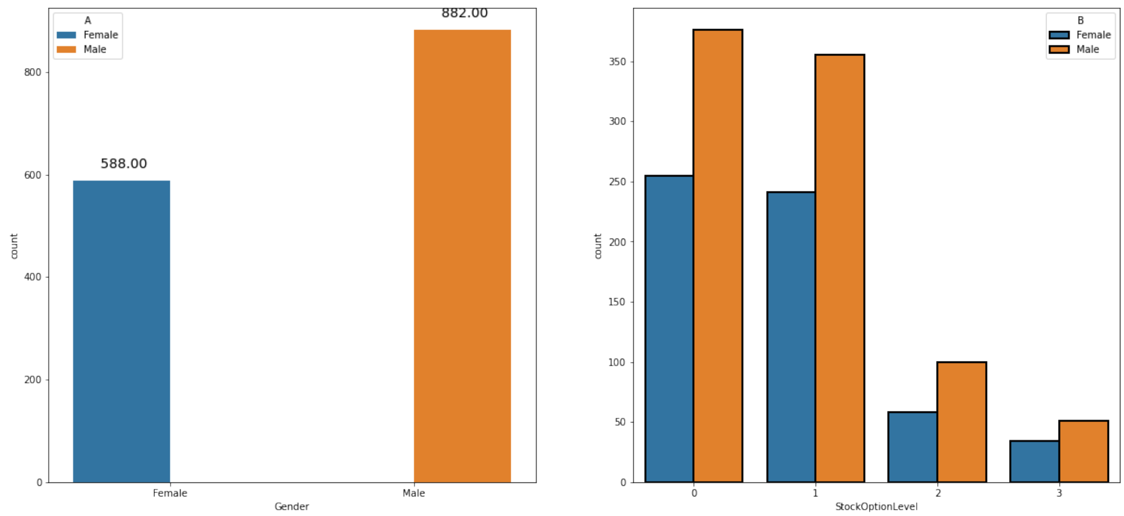


Figure 44: Male and females counts for A) total employees and B) different stock level options

The male/female percentage of taking different levels stock are almost same.

[78]:
```
plt.figure(figsize=(20,10))
ax=sns.countplot(data=employee_df1, x="TrainingTimesLastYear", hue="Attrition")
y=[]
for i in ax.patches:
      y.append(i.get_height())
i=0
b=7
for p in ax.patches:
    if i>=b:
        k=y[i]/(y[i]+y[i-b])*100
    else:
        k=y[i]/(y[i]+y[i+b])*100
    ax.annotate(f' {k:.2f} %', xy = (p.get_x()+p.get_width()/ 2, p.
↪get_height()),
                   ha='center',
                   va='center',
                   size=14,
                   xytext=(0, 8),
                   textcoords='offset points'
```

```
                )
    i=i+1
txt="Figure 45: Percentage attrition for different training times in a year."
plt.figtext(0.5, 0.05 , txt, wrap=True, horizontalalignment='center',␣
 ↪fontsize=15);
```

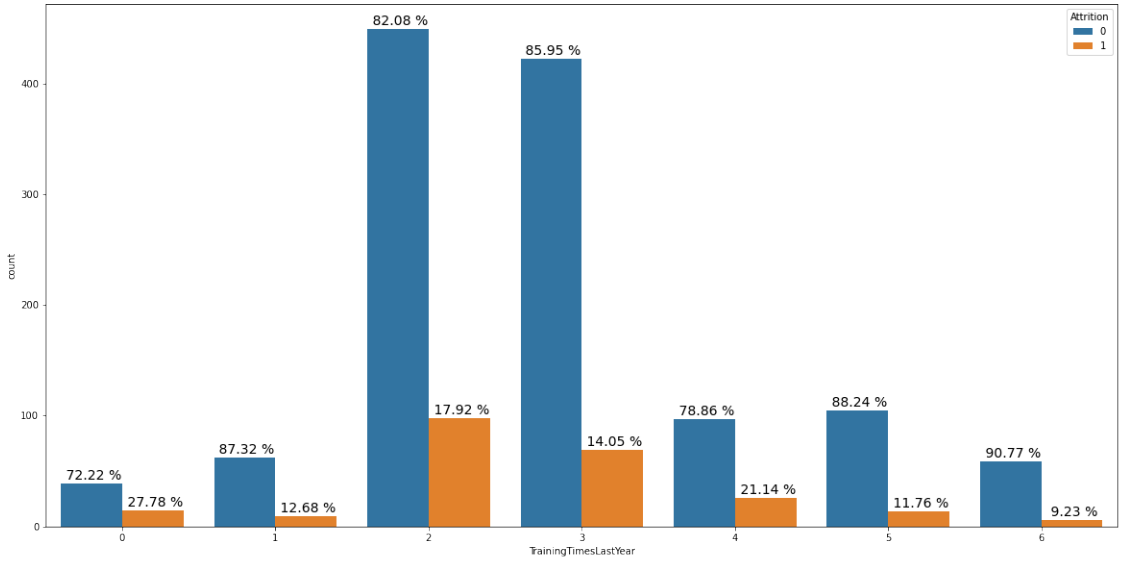


Figure 45: Percentage attrition for different training times in a year.

No conclusion can be drwan from the training time last year but more attrition is happening who did not get any training. We can also look who are the employees are getting more training.

[138]:
```
g=sns.catplot(data=employee_df1, kind="count", col_wrap=2, x='JobRole',␣
 ↪col='TrainingTimesLastYear', height=6)
g.set_xticklabels(rotation=90)

txt="Figure 46: Count plot of employees for different roles for different␣
 ↪training times."
plt.figtext(0.5, -0.09 , txt, wrap=True, horizontalalignment='center',␣
 ↪fontsize=15);
```

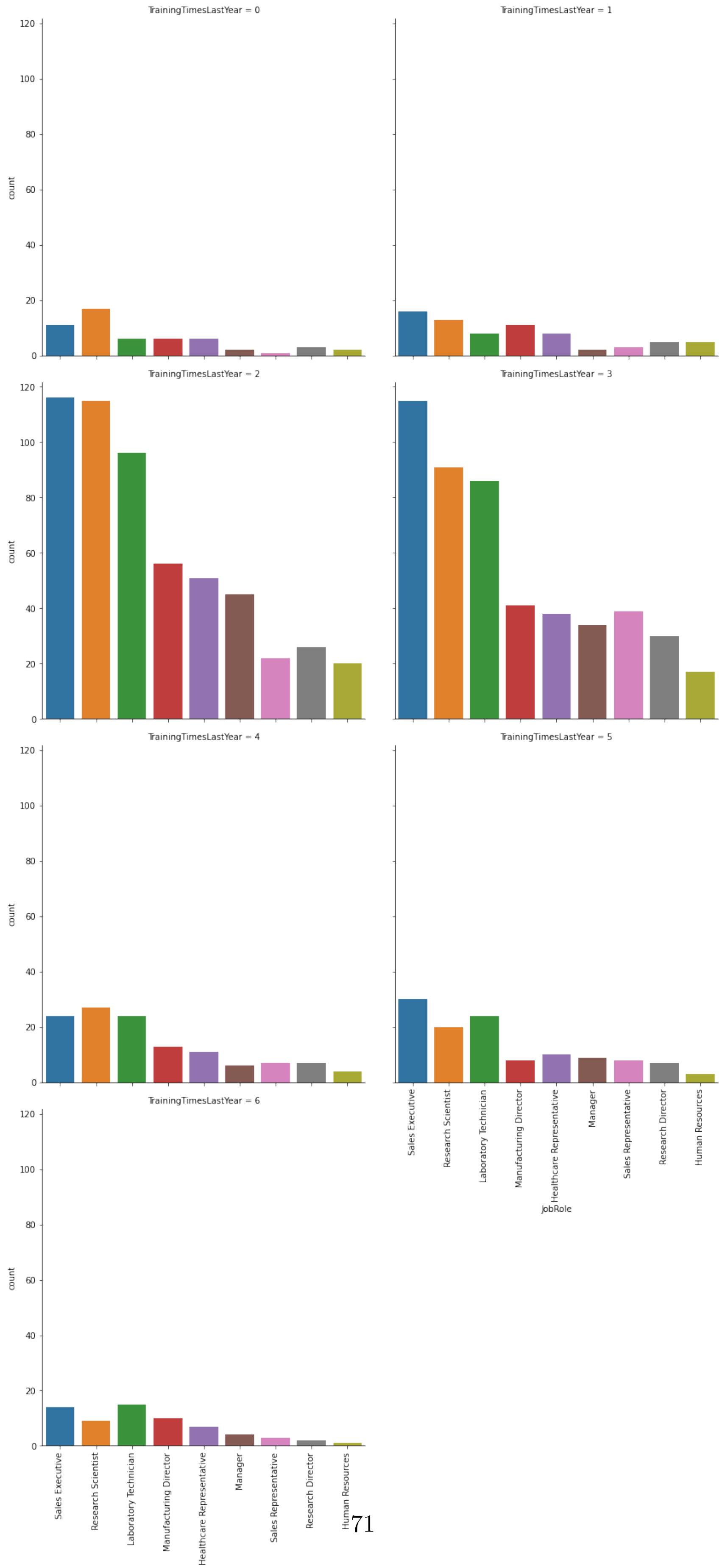


Figure 46: Count plot of employees for different roles for different training times.

Maximum employees in the comapny for all roles are getting 2 and 3 times training in the company.

[ ]:

```

```

[80]:

```
plt.figure(figsize=(20,10))
ax=sns.lineplot(data=employee_df1, x="TrainingTimesLastYear",␣
↪y="PerformanceRating")
txt="Figure 47: Training times vs performance rating"
plt.figtext(0.5, 0.05 , txt, wrap=True, horizontalalignment='center',␣
↪fontsize=15);
```

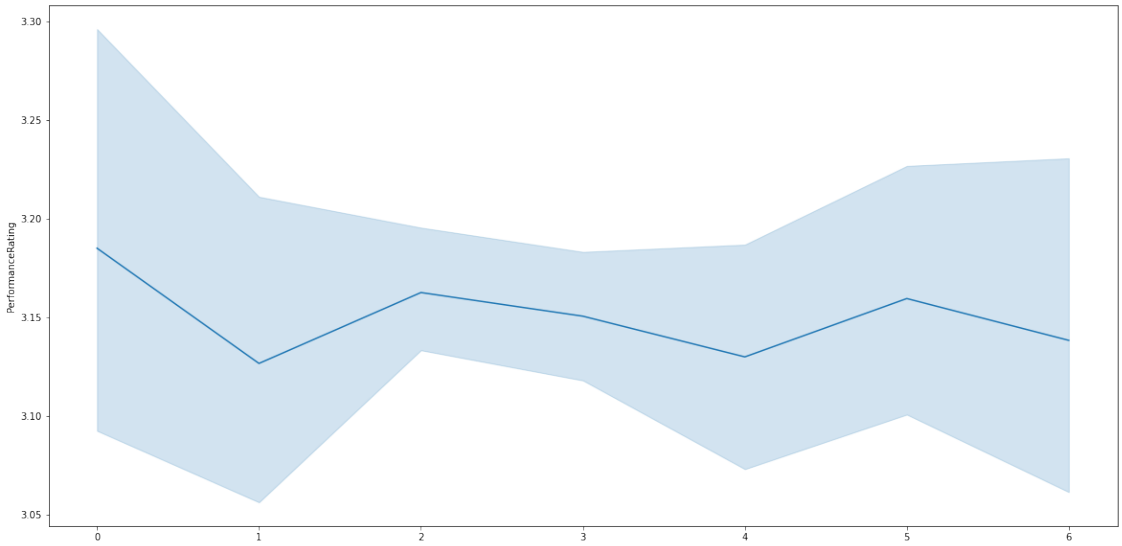


Figure 47: Training times vs performance rating

The training does not increase the performance rating.

[81]:

```
plt.figure(figsize=(20,10))
markers = {"+","a", "s","X"}
ax=sns.lineplot(data=employee_df1, hue="JobRole",
                y="PerformanceRating",  x="TrainingTimesLastYear", lw=4, ci=40 )
txt="Figure 48: Training times vs performance rating for different roles"
plt.figtext(0.5, 0.05 , txt, wrap=True, horizontalalignment='center',␣
↪fontsize=15);
```

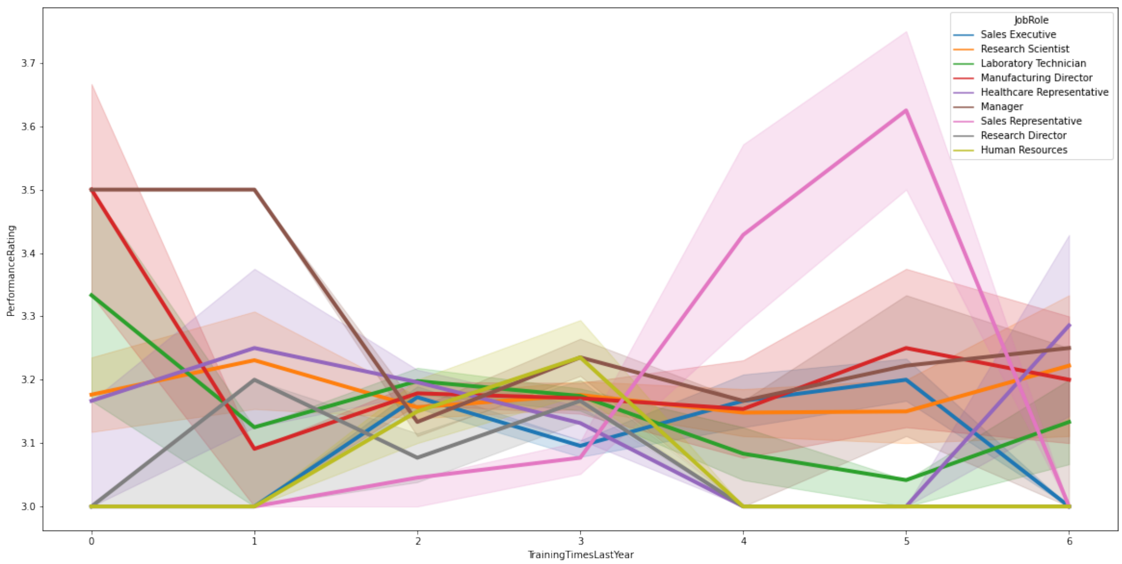


Figure 48: Training times vs performance rating for different roles

The performace rating actually increasing for sales representative. So it is better to provide more training to sales representatives than other roles. But provoding training more that 5 times is actually droping their performace

[82]:
```
plt.figure(figsize=(20,10))
ax=sns.countplot(data=employee_df1, x="WorkLifeBalance", hue="Attrition")
y=[]
for i in ax.patches:
      y.append(i.get_height())
i=0
b=4
for p in ax.patches:
    if i>=b:
        k=y[i]/(y[i]+y[i-b])*100
    else:
        k=y[i]/(y[i]+y[i+b])*100
    ax.annotate(f' {k:.2f} %', xy = (p.get_x()+p.get_width()/ 2, p.
 ↪get_height()),
                   ha='center',
                   va='center',
                   size=14,
                   xytext=(0, 8),
                   textcoords='offset points'
                )
    i=i+1
txt="Figure 48: Count plot for differetn levels of work life balance."
plt.figtext(0.5, 0.05 , txt, wrap=True, horizontalalignment='center',␣
 ↪fontsize=15);
```

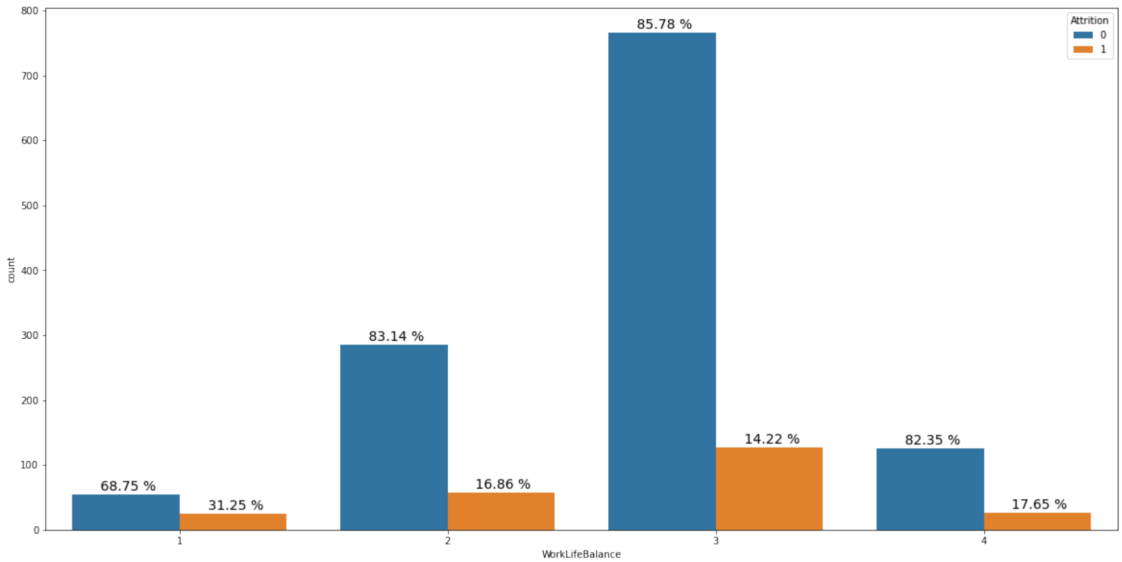


Figure 48: Count plot for differetn levels of work life balance.

The employees with WorkLife balance with level 1 have more chances to left the company.

[83]:
```
plt.figure(figsize=(20,10))
plt.subplot(211)
ax=sns.lineplot(data=employee_df1, y="WorkLifeBalance", x="JobLevel")
ax.legend(title="A", loc="upper right")
plt.subplot(212)
ax=sns.lineplot(data=employee_df1, y="WorkLifeBalance", x="JobLevel",␣
 ↪hue="JobRole", ci=10)
ax.legend(title="B", loc="upper right")
txt="Figure 49: Work life balance for A) Job level B) job level for different␣
 ↪job roles."
plt.figtext(0.5, 0.05 , txt, wrap=True, horizontalalignment='center',␣
 ↪fontsize=15);
```

```
No handles with labels found to put in legend.
```

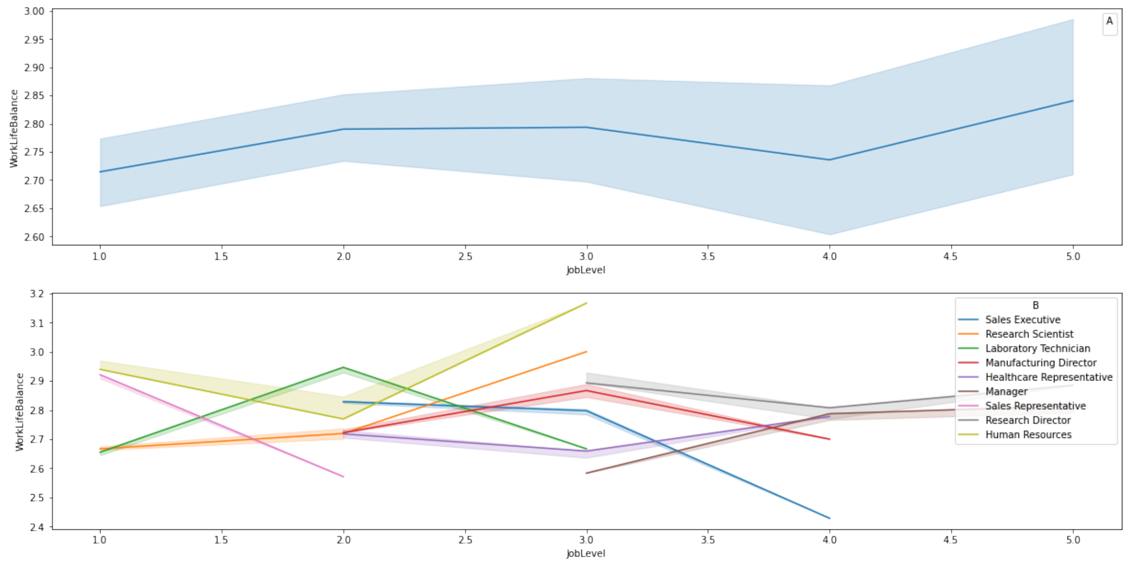


Figure 49: Work life balance for A) Job level B) job level for different job roles.

The workLifeBalance decreasing for Sales Representatives and Sales Executive with the increase of job Level

[84]:
```
sns.lmplot(data=employee_df,x='MonthlyIncome', y='WorkLifeBalance',
           hue='JobRole', height=10,
           aspect=1.5, x_bins=7, ci=6,
           scatter_kws={"s": 200},
           line_kws={'lw': 3})

txt="Figure 50: Work life balance for monthly income for different job roles."
plt.figtext(0.5, -0.05 , txt, wrap=True, horizontalalignment='center',␣
 ↪fontsize=15);
```

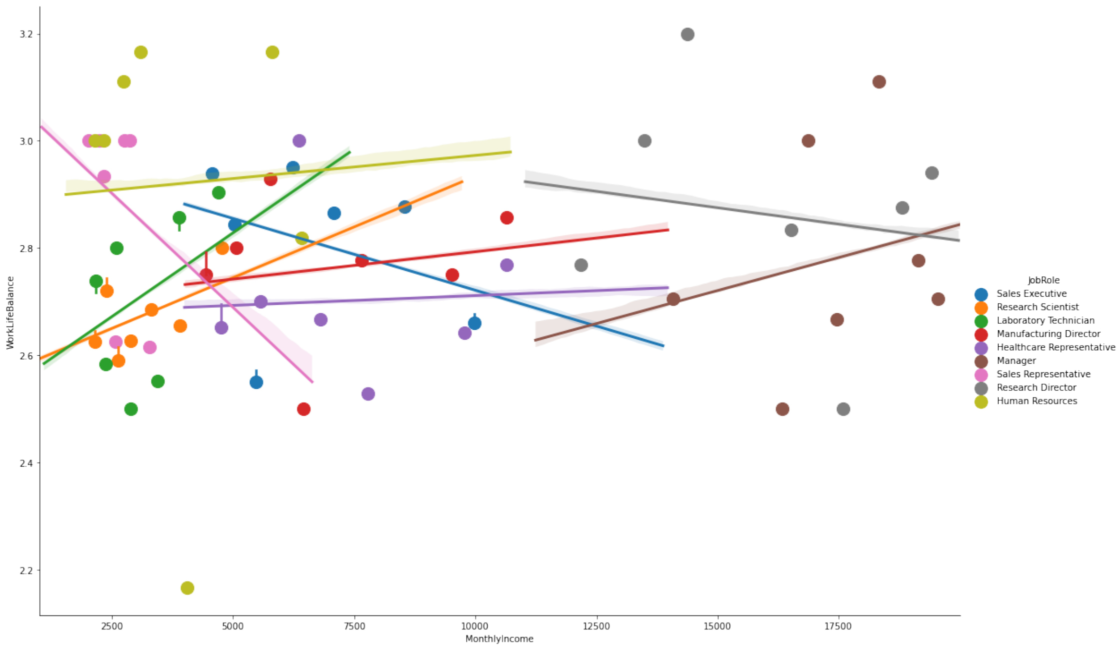


Figure 50: Work life balance for monthly income for different job roles.

The workLifeBalance actually decreasing for Sales Representatives and Sales Executive with the increase of Monthly Income. Thus the worklife balance and job satisafaction for sales representaitves and sales eecutives decreasing with the increase of monthly income and jobe levels. This may be the one of the main cause for more attrition in Sales Representatives roles.

[85]:
```
plt.figure(figsize=(20,20))
plt.subplot(321)
ax1=sns.kdeplot(x=employee_df1["TotalWorkingYears"], hue=employee_df1["Attrition"]);
ax1.legend(title="A", loc="upper right", labels=employee_df["Attrition"])
plt.subplot(322)
ax2=sns.kdeplot(x=employee_df1["YearsAtCompany"], hue=employee_df1["Attrition"]);
ax2.legend(title="B", loc="upper right", labels=employee_df["Attrition"])
plt.subplot(323)
ax3=sns.kdeplot(x=employee_df1["YearsInCurrentRole"], hue=employee_df1["Attrition"]);
ax3.legend(title="C", loc="upper right", labels=employee_df["Attrition"])
plt.subplot(324)
ax4=sns.kdeplot(x=employee_df1["YearsSinceLastPromotion"], hue=employee_df1["Attrition"])
ax4.legend(title="D", loc="upper right", labels=employee_df["Attrition"])
plt.subplot(325)
```

```
ax5=sns.kdeplot(x=employee_df1["YearsWithCurrManager"],␣
↪hue=employee_df1["Attrition"])
ax5.legend(title="E", loc="upper right", labels=employee_df["Attrition"])
txt="Figure 51: Kde plot A) TotalWorkingYears, B) YearsAtCompany, C)␣
↪YearsInCurrentRole, D) YearsSinceLastPromotion and E) YearsWithCurrManager␣
↪along with attrition."
plt.figtext(0.5, 0.05 , txt, wrap=True, horizontalalignment='center',␣
↪fontsize=15);
```

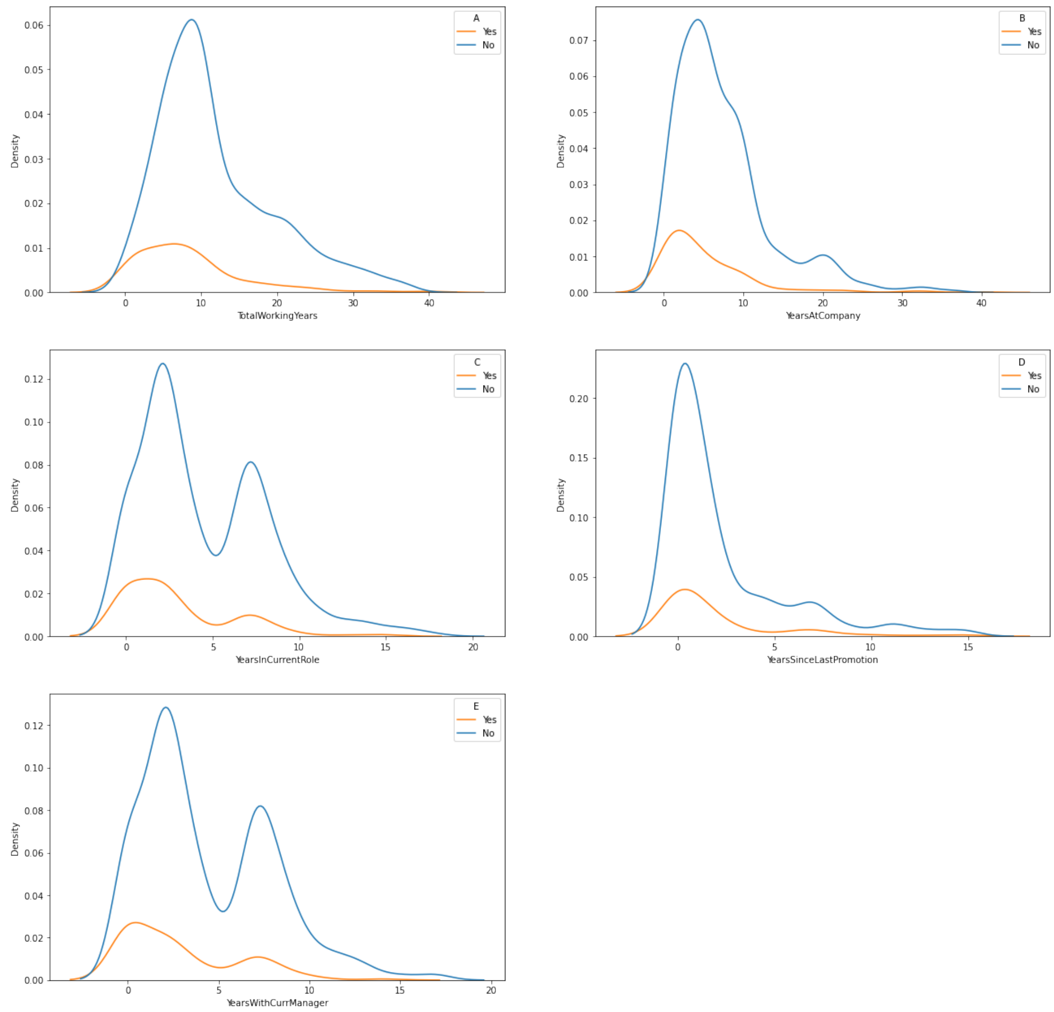


Figure 51: Kde plot A) TotalWorkingYears, B) YearsAtCompany, C) YearsInCurrentRole, D) YearsSinceLastPromotion and E) YearsWithCurrManager along with attrition.

[86]:
```
plt.figure(figsize=(20,10))
ax=sns.countplot(data=employee_df1[employee_df1["YearsAtCompany"]<20],␣
↪x="YearsAtCompany", hue="Attrition");
y=[]
```

```
for i in ax.patches:
      y.append(i.get_height())
i=0
b=20
for p in ax.patches:
    if i>=b:
        k=y[i]/(y[i]+y[i-b])*100
    else:
        k=y[i]/(y[i]+y[i+b])*100
    ax.annotate(f' {k:.2f} %', xy = (p.get_x()+p.get_width()/ 2, p.
 ↪get_height()),
                   ha='center',
                   va='center',
                   size=14,
                   xytext=(0, 8),
                   textcoords='offset points'
                )
    i=i+1

txt="Figure 52: Pervcentage atrition for Years at company"
plt.figtext(0.5, 0.05 , txt, wrap=True, horizontalalignment='center',␣
 ↪fontsize=15);
```

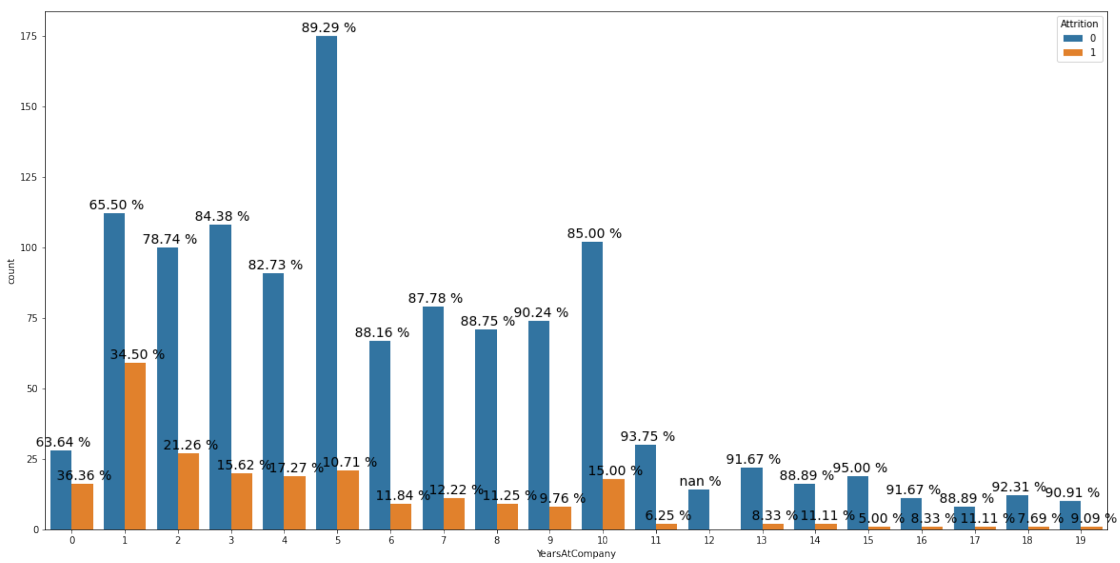


Figure 52: Pervcentage atrition for Years at company

Actually we can see that the attrition is more during the first three years after that its decreasing

[87]:
```
plt.figure(figsize=(20,10))
```

```
sns.scatterplot(data=employee_df1, y="TotalWorkingYears", x="MonthlyIncome",␣
↪hue="JobRole", s=200)
txt="Figure 53: Total working year vs monthly income plot for different job␣
↪roles."
plt.figtext(0.5, 0.05 , txt, wrap=True, horizontalalignment='center',␣
↪fontsize=15);
```

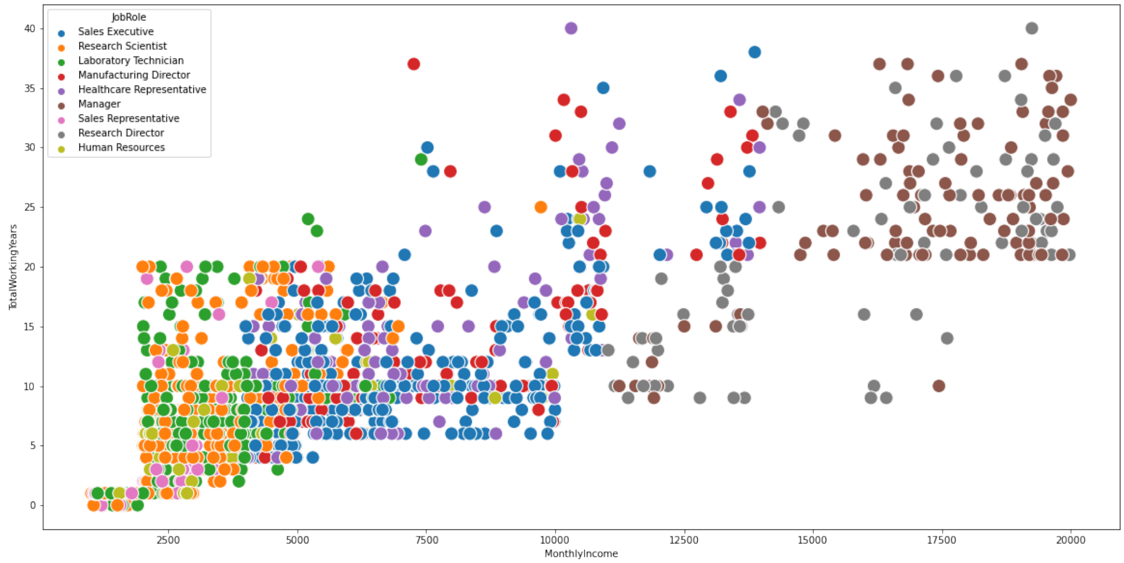


Figure 53: Total working year vs monthly income plot for different job roles.

Research scientists, laboratory Technician and Sales respresntative are among the lowest salery group but their experince is below 20 years, Most of the Research Directors and Managers have experience of more than 20 years and they are in higher Income range. All other fall in between. The similar trend has been observed for Years in the company as follow.

[88]:
```
plt.figure(figsize=(15,7))
plt.subplot(311)
ax1=sns.countplot(data=employee_df1, x="YearsInCurrentRole")
for p in ax1.patches:
    ax1.annotate(f' {p.get_height():.2f}', xy = (p.get_x()+p.get_width()/ 2, p.
↪get_height()),
                   ha='center',
                   va='center',
                   size=14,
                   xytext=(0, 8),
                   textcoords='offset points'
                )
ax1.legend(title="A")
```

```
plt.subplot(312)
ax2=sns.countplot(data=employee_df1, x="YearsInCurrentRole", hue="Attrition")
for p in ax2.patches:
    ax2.annotate(f' {p.get_height():.2f}', xy = (p.get_x()+p.get_width()/ 2, p.
 ↪get_height()),
                   ha='center',
                   va='center',
                   size=14,
                   xytext=(0, 8),
                   textcoords='offset points'
                )

ax2.legend(title="B", loc="upper right", labels=[])

plt.subplot(313)
ax3=sns.countplot(data=employee_df1, x="YearsInCurrentRole", hue="Attrition")
y=[]
for i in ax3.patches:
      y.append(i.get_height())
i=0
b=19
for p in ax3.patches:
    if i>=b:
        k=y[i]/(y[i]+y[i-b])*100
    else:
        k=y[i]/(y[i]+y[i+b])*100
    ax3.annotate(f' {k:.2f} %', xy = (p.get_x()+p.get_width()/ 2, p.
 ↪get_height()),
                   ha='center',
                   va='center',
                   size=14,
                   xytext=(0, 8),
                   textcoords='offset points'
                )
    i=i+1
ax3.legend(title="C", loc="upper right", labels=[])

txt="Figure 54: Total working year vs monthly income plot for different job␣
 ↪roles."
plt.figtext(0.5, 0.03 , txt, wrap=True, horizontalalignment='center',␣
 ↪fontsize=15);
```

```
No handles with labels found to put in legend.
```

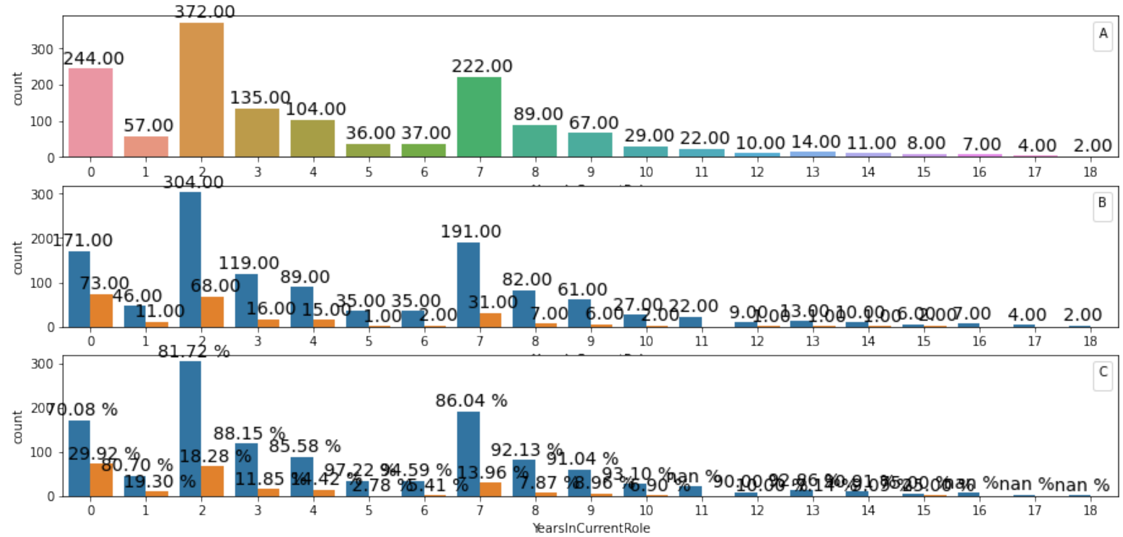


Figure 54: Total working year vs monthly income plot for different job roles.

The count plot shows that maximum employees stayes at the same role for their 1st, 3rd and 7th year. More percentage of employees left if after 1st year (almost 30%). We can also observe that the attrition is more for employees who are in their 1st, 2nd and 3d year at the same role. May they are waiting for pemotions. After third year till 7th year, the number of emplyees decreasing long with the huge decrease in attrition which indicates that most of the employees get permoted continously or their roles get changed. Also more employees are staying at the same role for their 7th year and then their is continous decrease which shows that the company give continous promotions as attrition rate is very less.

[89]:
```
plt.figure(figsize=(20,10))
g=sns.displot(data=employee_df1, kind="kde", x="YearsWithCurrManager",␣
↪col_wrap=3, col="JobLevel");

axes = g.axes.flatten()

for ax in axes:
#     t = ax.get_title().split(' = ')[1]
    kdeline = ax.lines[0]
    x_points = kdeline.get_xdata()
    y_points = kdeline.get_ydata()
    max_y = np.max(y_points)  # Find the maximum y value
    max_i = np.where(y_points==max_y)
    max_x=x_points[max_i]
    mode=max_y
    height= np.interp(max_x, x_points,y_points)
    ax.vlines(max_x, 0, height, color="gray",ls="--", lw=4)
    ax.fill_between(x_points, 0, y_points, facecolor='green', alpha=0.1)
```

```
txt="Figure 55: Kde plot for time with current manager for different job levels.
↪ The dotted line represents the mode value."
plt.figtext(0.5, -0.06 , txt, wrap=True, horizontalalignment='center',␣
↪fontsize=15);
```

```
<Figure size 1440x720 with 0 Axes>
```

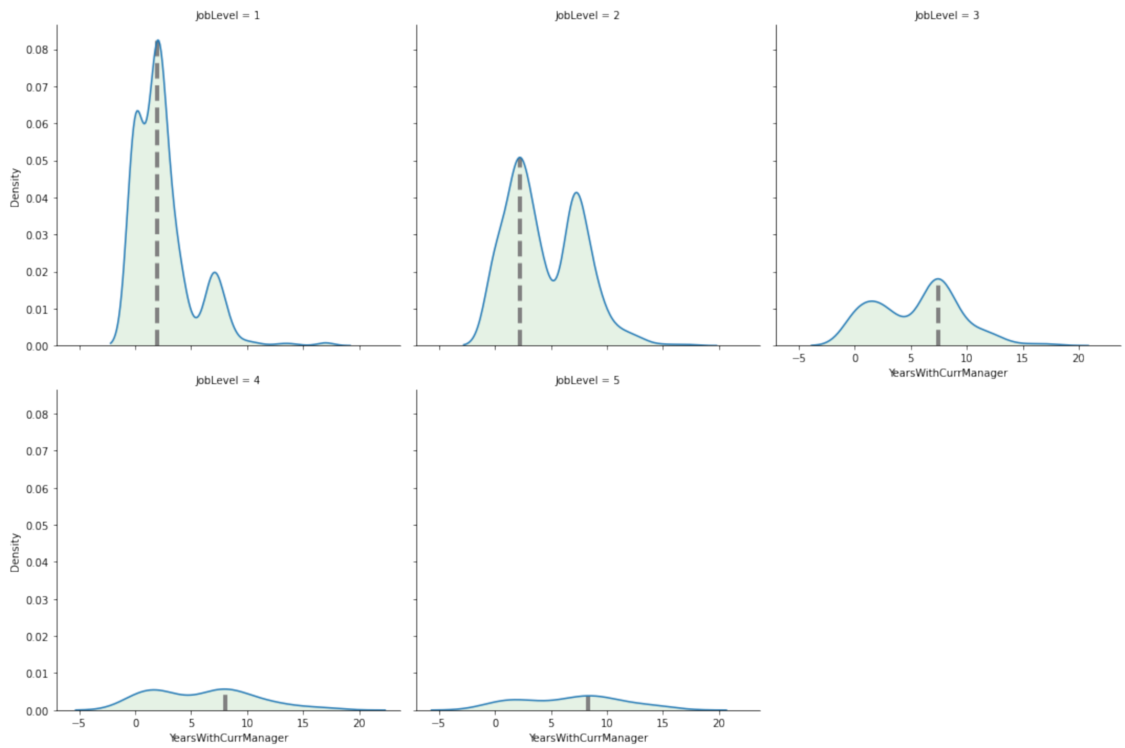


Figure 55: Kde plot for time with current manager for different job levels. The dotted line represents the mode value.

It can be observed that the number of employees at Job level 1 and 2 are is maximum with current manager for 1st year howver at Job level 3,4 and 5 the employees number is maximum at 7th year with the same manager.

3.4 Data Processing- After Data Analysis

After data cleaning we have left 31 numerical and six categorical attributes. We would like to convert these categorical variables into numerical variables and we can perform the data transformation

3.4.1 Data Transformation-

[90]:
```
employee_df1
```

[90]:

```
      Age  Attrition     BusinessTravel  DailyRate              Department  \
0      41          1      Travel_Rarely       1102                   Sales
1      49          0  Travel_Frequently        279  Research & Development
2      37          1      Travel_Rarely       1373  Research & Development
3      33          0  Travel_Frequently       1392  Research & Development
4      27          0      Travel_Rarely        591  Research & Development
…     …          …                …        …                     …
1465   36          0  Travel_Frequently        884  Research & Development
1466   39          0      Travel_Rarely        613  Research & Development
1467   27          0      Travel_Rarely        155  Research & Development
1468   49          0  Travel_Frequently       1023                   Sales
1469   34          0      Travel_Rarely        628  Research & Development

      DistanceFromHome  Education EducationField  EnvironmentSatisfaction  \
0                    1          2  Life Sciences                        2
1                    8          1  Life Sciences                        3
2                    2          2          Other                        4
3                    3          4  Life Sciences                        4
4                    2          1        Medical                        1
…                  …        …            …                      …
1465                23          2        Medical                        3
1466                 6          1        Medical                        4
1467                 4          3  Life Sciences                        2
1468                 2          3        Medical                        4
1469                 8          3        Medical                        2

      Gender  …  PerformanceRating  RelationshipSatisfaction  \
0     Female  …                  3                         1
1       Male  …                  4                         4
2       Male  …                  3                         2
3     Female  …                  3                         3
4       Male  …                  3                         4
…       …  …                …                       …
1465    Male  …                  3                         3
1466    Male  …                  3                         1
1467    Male  …                  4                         2
1468    Male  …                  3                         4
1469    Male  …                  3                         1

      StockOptionLevel TotalWorkingYears  TrainingTimesLastYear  \
0                    0                 8                      0
1                    1                10                      3
2                    0                 7                      3
3                    0                 8                      3
4                    1                 6                      3
…                  …               …                    …
1465                 1                17                      3
```

```
1466                 1                 9                       5
1467                 1                 6                       0
1468                 0                17                       3
1469                 0                 6                       3

     WorkLifeBalance  YearsAtCompany  YearsInCurrentRole  \
0                  1               6                   4
1                  3              10                   7
2                  3               0                   0
3                  3               8                   7
4                  3               2                   2
…                …              …                  …
1465               3               5                   2
1466               3               7                   7
1467               3               6                   2
1468               2               9                   6
1469               4               4                   3

      YearsSinceLastPromotion YearsWithCurrManager
0                           0                    5
1                           1                    7
2                           0                    0
3                           3                    0
4                           2                    2
…                         …                   …
1465                        0                    3
1466                        1                    7
1467                        0                    3
1468                        0                    8
1469                        1                    2

[1470 rows x 31 columns]
```

[91]:
```python
## Lets also transfer Over time
employee_df1["OverTime"]=employee_df1["OverTime"].apply(lambda x:1 if x=="Yes"␣
↪else 0 )
```

[92]:
```python
employee_df1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1470 entries, 0 to 1469
Data columns (total 31 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Age                       1470 non-null   int64
 1   Attrition                 1470 non-null   int64
 2   BusinessTravel            1470 non-null   object
 3   DailyRate                 1470 non-null   int64
```

```
 4   Department                1470 non-null   object
 5   DistanceFromHome          1470 non-null   int64
 6   Education                 1470 non-null   int64
 7   EducationField            1470 non-null   object
 8   EnvironmentSatisfaction   1470 non-null   int64
 9   Gender                    1470 non-null   object
 10  HourlyRate                1470 non-null   int64
 11  JobInvolvement            1470 non-null   int64
 12  JobLevel                  1470 non-null   int64
 13  JobRole                   1470 non-null   object
 14  JobSatisfaction           1470 non-null   int64
 15  MaritalStatus             1470 non-null   object
 16  MonthlyIncome             1470 non-null   int64
 17  MonthlyRate               1470 non-null   int64
 18  NumCompaniesWorked        1470 non-null   int64
 19  OverTime                  1470 non-null   int64
 20  PercentSalaryHike         1470 non-null   int64
 21  PerformanceRating         1470 non-null   int64
 22  RelationshipSatisfaction  1470 non-null   int64
 23  StockOptionLevel          1470 non-null   int64
 24  TotalWorkingYears         1470 non-null   int64
 25  TrainingTimesLastYear     1470 non-null   int64
 26  WorkLifeBalance           1470 non-null   int64
 27  YearsAtCompany            1470 non-null   int64
 28  YearsInCurrentRole        1470 non-null   int64
 29  YearsSinceLastPromotion   1470 non-null   int64
 30  YearsWithCurrManager      1470 non-null   int64
dtypes: int64(25), object(6)
memory usage: 356.1+ KB
```

[93]:
```
categorical_df=employee_df1[["BusinessTravel","Department","EducationField","Gender","MaritalS
```

[94]:
```
categorical_df
```

[94]:
```
         BusinessTravel              Department EducationField  Gender  \
0         Travel_Rarely                   Sales  Life Sciences  Female
1     Travel_Frequently  Research & Development  Life Sciences    Male
2         Travel_Rarely  Research & Development          Other    Male
3     Travel_Frequently  Research & Development  Life Sciences  Female
4         Travel_Rarely  Research & Development        Medical    Male
…                   …                       …            …      …
1465  Travel_Frequently  Research & Development        Medical    Male
1466      Travel_Rarely  Research & Development        Medical    Male
1467      Travel_Rarely  Research & Development  Life Sciences    Male
1468  Travel_Frequently                   Sales        Medical    Male
1469      Travel_Rarely  Research & Development        Medical    Male
```

```
    MaritalStatus                     JobRole
0          Single             Sales Executive
1         Married          Research Scientist
2          Single       Laboratory Technician
3         Married          Research Scientist
4         Married       Laboratory Technician
…             …                           …
1465      Married       Laboratory Technician
1466      Married   Healthcare Representative
1467      Married      Manufacturing Director
1468      Married             Sales Executive
1469      Married       Laboratory Technician

[1470 rows x 6 columns]
```

[95]:
```
from sklearn.preprocessing import OneHotEncoder
onehotencoder=OneHotEncoder()
categorical_df=onehotencoder.fit_transform(categorical_df).toarray()
```

[96]:
```
categorical_df
```

[96]:
```
array([[0., 0., 1., …, 0., 1., 0.],
       [0., 1., 0., …, 1., 0., 0.],
       [0., 0., 1., …, 0., 0., 0.],
       …,
       [0., 0., 1., …, 0., 0., 0.],
       [0., 1., 0., …, 0., 1., 0.],
       [0., 0., 1., …, 0., 0., 0.]])
```

[97]:
```
categorical_df=pd.DataFrame(categorical_df)
```

[98]:
```
categorical_df
```

[98]:
```
       0    1    2    3    4    5    6    7    8    9  …   16   17   18  \
0     0.0  0.0  1.0  0.0  0.0  1.0  0.0  1.0  0.0  0.0  …  1.0  0.0  0.0
1     0.0  1.0  0.0  0.0  1.0  0.0  0.0  1.0  0.0  0.0  …  0.0  0.0  0.0
2     0.0  0.0  1.0  0.0  1.0  0.0  0.0  0.0  0.0  0.0  …  1.0  0.0  0.0
3     0.0  1.0  0.0  0.0  1.0  0.0  0.0  1.0  0.0  0.0  …  0.0  0.0  0.0
4     0.0  0.0  1.0  0.0  1.0  0.0  0.0  0.0  0.0  1.0  …  0.0  0.0  0.0
…     …   …   …   …   …   …   …   …   …   …   …   …   …   …
1465  0.0  1.0  0.0  0.0  1.0  0.0  0.0  0.0  0.0  1.0  …  0.0  0.0  0.0
1466  0.0  0.0  1.0  0.0  1.0  0.0  0.0  0.0  0.0  1.0  …  0.0  1.0  0.0
1467  0.0  0.0  1.0  0.0  1.0  0.0  0.0  1.0  0.0  0.0  …  0.0  0.0  0.0
1468  0.0  1.0  0.0  0.0  0.0  1.0  0.0  0.0  0.0  1.0  …  0.0  0.0  0.0
1469  0.0  0.0  1.0  0.0  1.0  0.0  0.0  0.0  0.0  1.0  …  0.0  0.0  0.0

       19   20   21   22   23   24   25
```

```
0     0.0  0.0  0.0  0.0  0.0  1.0  0.0
1     0.0  0.0  0.0  0.0  1.0  0.0  0.0
2     1.0  0.0  0.0  0.0  0.0  0.0  0.0
3     0.0  0.0  0.0  0.0  1.0  0.0  0.0
4     1.0  0.0  0.0  0.0  0.0  0.0  0.0
…   …   …   …   …   …   …   …
1465  1.0  0.0  0.0  0.0  0.0  0.0  0.0
1466  0.0  0.0  0.0  0.0  0.0  0.0  0.0
1467  0.0  0.0  1.0  0.0  0.0  0.0  0.0
1468  0.0  0.0  0.0  0.0  0.0  1.0  0.0
1469  1.0  0.0  0.0  0.0  0.0  0.0  0.0

[1470 rows x 26 columns]
```

[99]:

```
categorical_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1470 entries, 0 to 1469
Data columns (total 26 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   0       1470 non-null   float64
 1   1       1470 non-null   float64
 2   2       1470 non-null   float64
 3   3       1470 non-null   float64
 4   4       1470 non-null   float64
 5   5       1470 non-null   float64
 6   6       1470 non-null   float64
 7   7       1470 non-null   float64
 8   8       1470 non-null   float64
 9   9       1470 non-null   float64
 10  10      1470 non-null   float64
 11  11      1470 non-null   float64
 12  12      1470 non-null   float64
 13  13      1470 non-null   float64
 14  14      1470 non-null   float64
 15  15      1470 non-null   float64
 16  16      1470 non-null   float64
 17  17      1470 non-null   float64
 18  18      1470 non-null   float64
 19  19      1470 non-null   float64
 20  20      1470 non-null   float64
 21  21      1470 non-null   float64
 22  22      1470 non-null   float64
 23  23      1470 non-null   float64
 24  24      1470 non-null   float64
 25  25      1470 non-null   float64
dtypes: float64(26)
```

```
memory usage: 298.7 KB
```

[100]:
```python
## let collect the numerical varaible without target variable
numerical_df=employee_df1.
↪drop(["Attrition","BusinessTravel","Department","EducationField","Gender","MaritalStatus","
↪axis=1)
```

[101]:
```python
numerical_df
```

[101]:
```
      Age  DailyRate  DistanceFromHome  Education  EnvironmentSatisfaction  \
0      41       1102                 1          2                        2
1      49        279                 8          1                        3
2      37       1373                 2          2                        4
3      33       1392                 3          4                        4
4      27        591                 2          1                        1
…     …          …                 …          …                      …
1465   36        884                23          2                        3
1466   39        613                 6          1                        4
1467   27        155                 4          3                        2
1468   49       1023                 2          3                        4
1469   34        628                 8          3                        2

      HourlyRate  JobInvolvement  JobLevel  JobSatisfaction  MonthlyIncome  \
0             94               3         2                4           5993
1             61               2         2                2           5130
2             92               2         1                3           2090
3             56               3         1                3           2909
4             40               3         1                2           3468
…            …               …        …               …             …
1465          41               4         2                4           2571
1466          42               2         3                1           9991
1467          87               4         2                2           6142
1468          63               2         2                2           5390
1469          82               4         2                3           4404

      …  PerformanceRating  RelationshipSatisfaction  StockOptionLevel  \
0     …                  3                         1                 0
1     …                  4                         4                 1
2     …                  3                         2                 0
3     …                  3                         3                 0
4     …                  3                         4                 1
…    …                …                        …                …
1465  …                  3                         3                 1
1466  …                  3                         1                 1
1467  …                  4                         2                 1
1468  …                  3                         4                 0
1469  …                  3                         1                 0
```

```
      TotalWorkingYears  TrainingTimesLastYear  WorkLifeBalance  \
0                     8                      0                1
1                    10                      3                3
2                     7                      3                3
3                     8                      3                3
4                     6                      3                3
…                   …                    …                …
1465                 17                      3                3
1466                  9                      5                3
1467                  6                      0                3
1468                 17                      3                2
1469                  6                      3                4

      YearsAtCompany  YearsInCurrentRole  YearsSinceLastPromotion  \
0                  6                   4                        0
1                 10                   7                        1
2                  0                   0                        0
3                  8                   7                        3
4                  2                   2                        2
…                …                   …                      …
1465               5                   2                        0
1466               7                   7                        1
1467               6                   2                        0
1468               9                   6                        0
1469               4                   3                        1

      YearsWithCurrManager
0                        5
1                        7
2                        0
3                        0
4                        2
…                      …
1465                     3
1466                     7
1467                     3
1468                     8
1469                     2

[1470 rows x 24 columns]
```

[102]:
```
numerical_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1470 entries, 0 to 1469
Data columns (total 24 columns):
 #   Column                    Non-Null Count  Dtype
```

```
---  ------                    --------------  -----
 0   Age                       1470 non-null   int64
 1   DailyRate                 1470 non-null   int64
 2   DistanceFromHome          1470 non-null   int64
 3   Education                 1470 non-null   int64
 4   EnvironmentSatisfaction   1470 non-null   int64
 5   HourlyRate                1470 non-null   int64
 6   JobInvolvement            1470 non-null   int64
 7   JobLevel                  1470 non-null   int64
 8   JobSatisfaction           1470 non-null   int64
 9   MonthlyIncome             1470 non-null   int64
 10  MonthlyRate               1470 non-null   int64
 11  NumCompaniesWorked        1470 non-null   int64
 12  OverTime                  1470 non-null   int64
 13  PercentSalaryHike         1470 non-null   int64
 14  PerformanceRating         1470 non-null   int64
 15  RelationshipSatisfaction  1470 non-null   int64
 16  StockOptionLevel          1470 non-null   int64
 17  TotalWorkingYears         1470 non-null   int64
 18  TrainingTimesLastYear     1470 non-null   int64
 19  WorkLifeBalance           1470 non-null   int64
 20  YearsAtCompany            1470 non-null   int64
 21  YearsInCurrentRole        1470 non-null   int64
 22  YearsSinceLastPromotion   1470 non-null   int64
 23  YearsWithCurrManager      1470 non-null   int64
dtypes: int64(24)
memory usage: 275.8 KB
```

[103]:
```
employee_df2=pd.concat([numerical_df,categorical_df], axis=1)
```

[104]:
```
employee_df2
```

[104]:
```
      Age  DailyRate  DistanceFromHome  Education  EnvironmentSatisfaction  \
0      41       1102                 1          2                        2
1      49        279                 8          1                        3
2      37       1373                 2          2                        4
3      33       1392                 3          4                        4
4      27        591                 2          1                        1
…     …         …                …          …                        …
1465   36        884                23          2                        3
1466   39        613                 6          1                        4
1467   27        155                 4          3                        2
1468   49       1023                 2          3                        4
1469   34        628                 8          3                        2

      HourlyRate  JobInvolvement  JobLevel  JobSatisfaction  MonthlyIncome  \
0             94               3         2                4           5993
```

```
1            61            2       2            2        5130
2            92            2       1            3        2090
3            56            3       1            3        2909
4            40            3       1            2        3468
…          …           …      …           …        …
1465         41            4       2            4        2571
1466         42            2       3            1        9991
1467         87            4       2            2        6142
1468         63            2       2            2        5390
1469         82            4       2            3        4404

      …   16   17   18   19   20   21   22   23   24   25
0     …  1.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  1.0  0.0
1     …  0.0  0.0  0.0  0.0  0.0  0.0  0.0  1.0  0.0  0.0
2     …  1.0  0.0  0.0  1.0  0.0  0.0  0.0  0.0  0.0  0.0
3     …  0.0  0.0  0.0  0.0  0.0  0.0  0.0  1.0  0.0  0.0
4     …  0.0  0.0  0.0  1.0  0.0  0.0  0.0  0.0  0.0  0.0
…   …  …  …  …  …  …  …  …  …  …  …
1465  …  0.0  0.0  0.0  1.0  0.0  0.0  0.0  0.0  0.0  0.0
1466  …  0.0  1.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0
1467  …  0.0  0.0  0.0  0.0  0.0  1.0  0.0  0.0  0.0  0.0
1468  …  0.0  0.0  0.0  0.0  0.0  0.0  0.0  0.0  1.0  0.0
1469  …  0.0  0.0  0.0  1.0  0.0  0.0  0.0  0.0  0.0  0.0

[1470 rows x 50 columns]
```

[105]:
```
employee_df2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1470 entries, 0 to 1469
Data columns (total 50 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Age                       1470 non-null   int64
 1   DailyRate                 1470 non-null   int64
 2   DistanceFromHome          1470 non-null   int64
 3   Education                 1470 non-null   int64
 4   EnvironmentSatisfaction   1470 non-null   int64
 5   HourlyRate                1470 non-null   int64
 6   JobInvolvement            1470 non-null   int64
 7   JobLevel                  1470 non-null   int64
 8   JobSatisfaction           1470 non-null   int64
 9   MonthlyIncome             1470 non-null   int64
 10  MonthlyRate               1470 non-null   int64
 11  NumCompaniesWorked        1470 non-null   int64
 12  OverTime                  1470 non-null   int64
 13  PercentSalaryHike         1470 non-null   int64
 14  PerformanceRating         1470 non-null   int64
```

```
 15  RelationshipSatisfaction  1470 non-null   int64
 16  StockOptionLevel          1470 non-null   int64
 17  TotalWorkingYears         1470 non-null   int64
 18  TrainingTimesLastYear     1470 non-null   int64
 19  WorkLifeBalance           1470 non-null   int64
 20  YearsAtCompany            1470 non-null   int64
 21  YearsInCurrentRole        1470 non-null   int64
 22  YearsSinceLastPromotion   1470 non-null   int64
 23  YearsWithCurrManager      1470 non-null   int64
 24  0                         1470 non-null   float64
 25  1                         1470 non-null   float64
 26  2                         1470 non-null   float64
 27  3                         1470 non-null   float64
 28  4                         1470 non-null   float64
 29  5                         1470 non-null   float64
 30  6                         1470 non-null   float64
 31  7                         1470 non-null   float64
 32  8                         1470 non-null   float64
 33  9                         1470 non-null   float64
 34  10                        1470 non-null   float64
 35  11                        1470 non-null   float64
 36  12                        1470 non-null   float64
 37  13                        1470 non-null   float64
 38  14                        1470 non-null   float64
 39  15                        1470 non-null   float64
 40  16                        1470 non-null   float64
 41  17                        1470 non-null   float64
 42  18                        1470 non-null   float64
 43  19                        1470 non-null   float64
 44  20                        1470 non-null   float64
 45  21                        1470 non-null   float64
 46  22                        1470 non-null   float64
 47  23                        1470 non-null   float64
 48  24                        1470 non-null   float64
 49  25                        1470 non-null   float64
dtypes: float64(26), int64(24)
memory usage: 574.3 KB
```

3.4.2 Data Normalization-

[106]:
```python
from sklearn.preprocessing import MinMaxScaler
scaler=MinMaxScaler()
x=scaler.fit_transform(employee_df2)
```

[107]:
```python
x
```

[107]:
```
array([[0.54761905, 0.71581961, 0.        , …, 0.        , 1.        ,
        0.        ],
       [0.73809524, 0.12670007, 0.25      , …, 1.        , 0.        ,
        0.        ],
       [0.45238095, 0.90980673, 0.03571429, …, 0.        , 0.        ,
        0.        ],
       …,
       [0.21428571, 0.03793844, 0.10714286, …, 0.        , 0.        ,
        0.        ],
       [0.73809524, 0.65926986, 0.03571429, …, 0.        , 1.        ,
        0.        ],
       [0.38095238, 0.37652112, 0.25      , …, 0.        , 0.        ,
        0.        ]])
```

[108]:
```
y=employee_df1["Attrition"]
```

[109]:
```
y
```

[109]:
```
0       1
1       0
2       1
3       0
4       0
       ..
1465    0
1466    0
1467    0
1468    0
1469    0
Name: Attrition, Length: 1470, dtype: int64
```

3.5 Data Sampling-

[110]:
```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test= train_test_split(x,y, test_size=0.25)
```

[111]:
```
x_train.shape
```

[111]:
```
(1102, 50)
```

[112]:
```
x_test.shape
```

[112]:
```
(368, 50)
```

4 Building the Model

[113]:
```
y.value_counts()
```

[113]:
```
0    1233
1     237
Name: Attrition, dtype: int64
```

This is a unbalanced binary problem

[114]:
```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, confusion_matrix,␣
 ↪classification_report

model=LogisticRegression()
model.fit(x_train,y_train)
```

[114]:
```
LogisticRegression()
```

[115]:
```
y_train
```

[115]:
```
833     0
68      0
1030    0
1188    0
647     0
       ..
1145    0
499     0
24      1
424     0
1353    1
Name: Attrition, Length: 1102, dtype: int64
```

[116]:
```
y_pred=model.predict(x_test)
```

[117]:
```
y_pred
```

[117]:
```
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1,
       0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,
       0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1,
       1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1,
       0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
```

```
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,
0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0,
0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0], dtype=int64)
```

[118]:
```
accuracy=accuracy_score(y_pred, y_test)
```

[119]:
```
accuracy
```

[119]:
```
0.8967391304347826
```

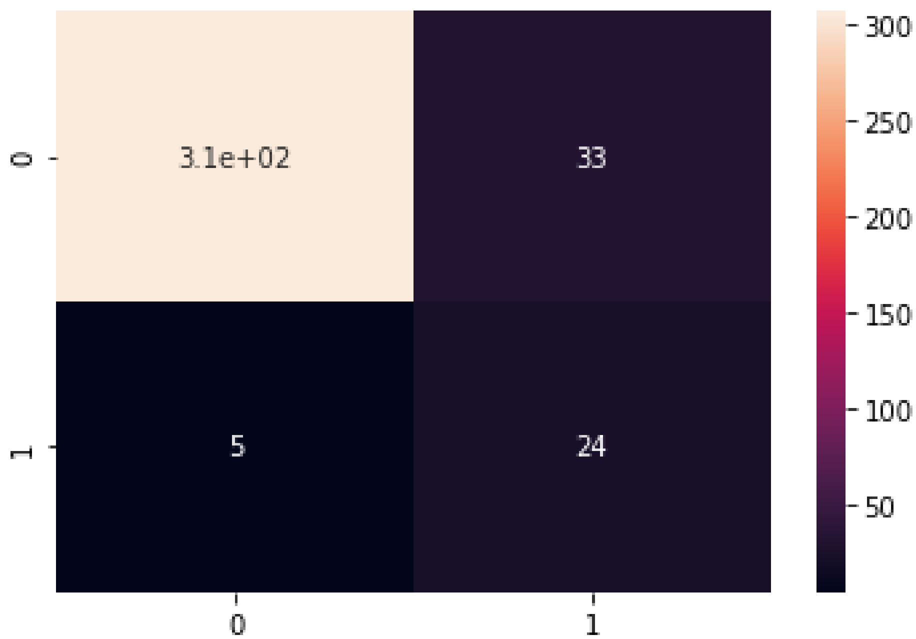
[120]:
```
conf_matrix=confusion_matrix(y_pred, y_test)
```

[121]:
```
conf_matrix
```

[121]:
```
array([[306,  33],
       [  5,  24]], dtype=int64)
```

[122]:
```
sns.heatmap(conf_matrix,annot=True)
```

[122]:
```
<AxesSubplot:>
```



[123]:
```
print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.90      0.98      0.94       311
           1       0.83      0.42      0.56        57

    accuracy                           0.90       368
   macro avg       0.87      0.70      0.75       368
weighted avg       0.89      0.90      0.88       368
```

Lets try Random Forest

[124]:
```
from sklearn.ensemble import RandomForestClassifier
model=RandomForestClassifier()
model.fit(x_train, y_train)
y_pred=model.predict(x_test)
y_pred
```

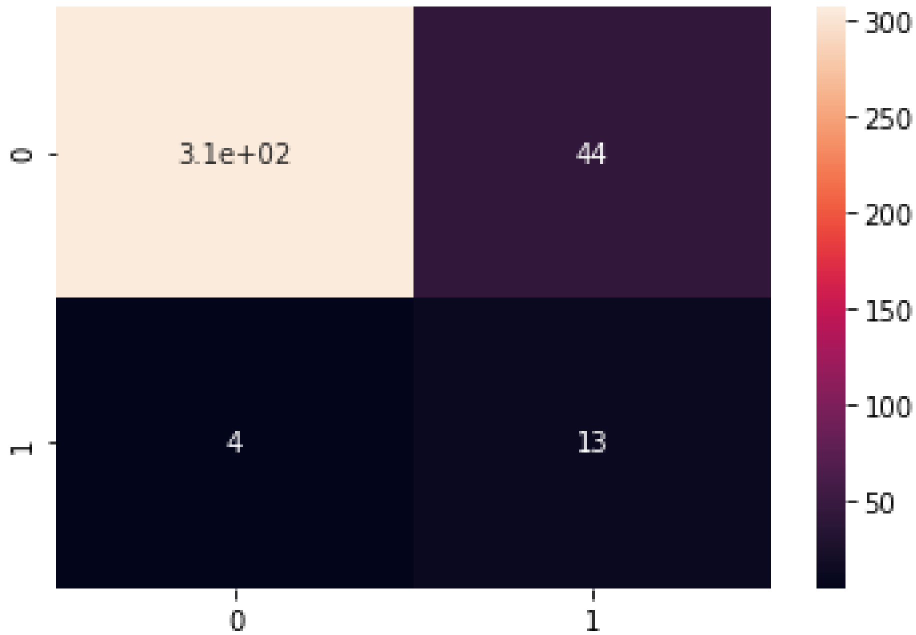
[124]:
```
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,
       0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0], dtype=int64)
```

[125]:
```
accuracy=accuracy_score(y_pred, y_test)
accuracy
```

[125]:
```
0.8695652173913043
```

[126]:
```
conf_matrix=confusion_matrix(y_pred, y_test)
sns.heatmap(conf_matrix,annot=True)
```

[126]:
```
<AxesSubplot:>
```



[127]:
```
print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.87      0.99      0.93       311
           1       0.76      0.23      0.35        57

    accuracy                           0.87       368
   macro avg       0.82      0.61      0.64       368
weighted avg       0.86      0.87      0.84       368
```

Lets try tenserflow sequential model

[128]:
```
import tensorflow as tf
model =tf.keras.models.Sequential()
model.add(tf.keras.layers.Dense(units=500, activation='relu', input_shape=(50,␣
 ↪)))
model.add(tf.keras.layers.Dense(units=500, activation='relu'))
model.add(tf.keras.layers.Dense(units=500, activation='relu'))
model.add(tf.keras.layers.Dense(units=1, activation='sigmoid'))
```

[129]:
```
model.summary()
```

```
Model: "sequential"

_________________________________________________________________
Layer (type)                 Output Shape              Param #
```

```
=================================================================
dense (Dense)                (None, 500)               25500

_________________________________________________________________
dense_1 (Dense)              (None, 500)               250500

_________________________________________________________________
dense_2 (Dense)              (None, 500)               250500

_________________________________________________________________
dense_3 (Dense)              (None, 1)                 501
=================================================================
Total params: 527,001
Trainable params: 527,001
Non-trainable params: 0

_________________________________________________________________
```

[130]:
```
model.compile(optimizer='Adam', loss='binary_crossentropy',␣
↪metrics=['accuracy'])
```

[131]:
```
model.fit(x_train, y_train,epochs=100,batch_size=50)
```

```
Epoch 1/100
23/23 [==============================] - 1s 6ms/step - loss: 0.4335 - accuracy:
0.8140
Epoch 2/100
23/23 [==============================] - 0s 6ms/step - loss: 0.3542 - accuracy:
0.8512
Epoch 3/100
23/23 [==============================] - 0s 5ms/step - loss: 0.3179 - accuracy:
0.8775
Epoch 4/100
23/23 [==============================] - 0s 5ms/step - loss: 0.2855 - accuracy:
0.8947
Epoch 5/100
23/23 [==============================] - 0s 5ms/step - loss: 0.2488 - accuracy:
0.9102
Epoch 6/100
23/23 [==============================] - 0s 6ms/step - loss: 0.2171 - accuracy:
0.9129
Epoch 7/100
23/23 [==============================] - 0s 5ms/step - loss: 0.1867 - accuracy:
0.9292
Epoch 8/100
23/23 [==============================] - 0s 6ms/step - loss: 0.1707 - accuracy:
0.9347
Epoch 9/100
23/23 [==============================] - 0s 5ms/step - loss: 0.1780 - accuracy:
0.9410
Epoch 10/100
23/23 [==============================] - 0s 6ms/step - loss: 0.1223 - accuracy:
```

```
0.9537
Epoch 11/100
23/23 [==============================] - 0s 5ms/step - loss: 0.0909 - accuracy:
0.9655
Epoch 12/100
23/23 [==============================] - 0s 5ms/step - loss: 0.0657 - accuracy:
0.9764
Epoch 13/100
23/23 [==============================] - 0s 5ms/step - loss: 0.0357 - accuracy:
0.9918
Epoch 14/100
23/23 [==============================] - 0s 5ms/step - loss: 0.0379 - accuracy:
0.9855
Epoch 15/100
23/23 [==============================] - 0s 5ms/step - loss: 0.0258 - accuracy:
0.9936
Epoch 16/100
23/23 [==============================] - 0s 5ms/step - loss: 0.0135 - accuracy:
0.9973
Epoch 17/100
23/23 [==============================] - 0s 6ms/step - loss: 0.0158 - accuracy:
0.9955
Epoch 18/100
23/23 [==============================] - 0s 6ms/step - loss: 0.0110 - accuracy:
0.9991
Epoch 19/100
23/23 [==============================] - 0s 5ms/step - loss: 0.0059 - accuracy:
1.0000
Epoch 20/100
23/23 [==============================] - 0s 5ms/step - loss: 0.0022 - accuracy:
1.0000
Epoch 21/100
23/23 [==============================] - 0s 5ms/step - loss: 8.0908e-04 -
accuracy: 1.0000
Epoch 22/100
23/23 [==============================] - 0s 5ms/step - loss: 4.9947e-04 -
accuracy: 1.0000
Epoch 23/100
23/23 [==============================] - 0s 5ms/step - loss: 3.6981e-04 -
accuracy: 1.0000
Epoch 24/100
23/23 [==============================] - 0s 5ms/step - loss: 2.9676e-04 -
accuracy: 1.0000: 0s - loss: 2.8511e-04 - accuracy: 1.00
Epoch 25/100
23/23 [==============================] - 0s 5ms/step - loss: 2.6548e-04 -
accuracy: 1.0000
Epoch 26/100
23/23 [==============================] - 0s 5ms/step - loss: 2.1515e-04 -
```

```
accuracy: 1.0000
Epoch 27/100
23/23 [==============================] - 0s 5ms/step - loss: 1.8791e-04 -
accuracy: 1.0000
Epoch 28/100
23/23 [==============================] - 0s 5ms/step - loss: 1.6768e-04 -
accuracy: 1.0000
Epoch 29/100
23/23 [==============================] - 0s 5ms/step - loss: 1.4954e-04 -
accuracy: 1.0000
Epoch 30/100
23/23 [==============================] - 0s 5ms/step - loss: 1.3616e-04 -
accuracy: 1.0000
Epoch 31/100
23/23 [==============================] - 0s 5ms/step - loss: 1.2175e-04 -
accuracy: 1.0000
Epoch 32/100
23/23 [==============================] - 0s 5ms/step - loss: 1.1180e-04 -
accuracy: 1.0000
Epoch 33/100
23/23 [==============================] - 0s 5ms/step - loss: 1.0226e-04 -
accuracy: 1.0000
Epoch 34/100
23/23 [==============================] - 0s 5ms/step - loss: 9.3191e-05 -
accuracy: 1.0000
Epoch 35/100
23/23 [==============================] - 0s 5ms/step - loss: 8.5914e-05 -
accuracy: 1.0000
Epoch 36/100
23/23 [==============================] - 0s 5ms/step - loss: 7.9084e-05 -
accuracy: 1.0000
Epoch 37/100
23/23 [==============================] - 0s 5ms/step - loss: 7.2785e-05 -
accuracy: 1.0000
Epoch 38/100
23/23 [==============================] - 0s 5ms/step - loss: 6.7521e-05 -
accuracy: 1.0000
Epoch 39/100
23/23 [==============================] - 0s 5ms/step - loss: 6.2227e-05 -
accuracy: 1.0000
Epoch 40/100
23/23 [==============================] - 0s 5ms/step - loss: 5.8260e-05 -
accuracy: 1.0000
Epoch 41/100
23/23 [==============================] - 0s 5ms/step - loss: 5.3937e-05 -
accuracy: 1.0000
Epoch 42/100
23/23 [==============================] - 0s 5ms/step - loss: 4.9688e-05 -
```

```
accuracy: 1.0000
Epoch 43/100
23/23 [==============================] - 0s 4ms/step - loss: 4.6774e-05 -
accuracy: 1.0000
Epoch 44/100
23/23 [==============================] - 0s 5ms/step - loss: 4.3776e-05 -
accuracy: 1.0000
Epoch 45/100
23/23 [==============================] - 0s 5ms/step - loss: 4.0810e-05 -
accuracy: 1.0000
Epoch 46/100
23/23 [==============================] - 0s 5ms/step - loss: 3.8442e-05 -
accuracy: 1.0000
Epoch 47/100
23/23 [==============================] - 0s 6ms/step - loss: 3.6186e-05 -
accuracy: 1.0000
Epoch 48/100
23/23 [==============================] - 0s 5ms/step - loss: 3.4703e-05 -
accuracy: 1.0000
Epoch 49/100
23/23 [==============================] - 0s 5ms/step - loss: 3.3447e-05 -
accuracy: 1.0000
Epoch 50/100
23/23 [==============================] - 0s 5ms/step - loss: 3.0297e-05 -
accuracy: 1.0000
Epoch 51/100
23/23 [==============================] - 0s 5ms/step - loss: 2.8436e-05 -
accuracy: 1.0000
Epoch 52/100
23/23 [==============================] - 0s 5ms/step - loss: 2.6865e-05 -
accuracy: 1.0000
Epoch 53/100
23/23 [==============================] - 0s 5ms/step - loss: 2.5490e-05 -
accuracy: 1.0000
Epoch 54/100
23/23 [==============================] - 0s 5ms/step - loss: 2.4240e-05 -
accuracy: 1.0000
Epoch 55/100
23/23 [==============================] - 0s 5ms/step - loss: 2.2822e-05 -
accuracy: 1.0000
Epoch 56/100
23/23 [==============================] - 0s 5ms/step - loss: 2.1809e-05 -
accuracy: 1.0000
Epoch 57/100
23/23 [==============================] - 0s 5ms/step - loss: 2.0682e-05 -
accuracy: 1.0000
Epoch 58/100
23/23 [==============================] - 0s 5ms/step - loss: 1.9730e-05 -
```

```
accuracy: 1.0000
Epoch 59/100
23/23 [==============================] - 0s 5ms/step - loss: 1.8820e-05 -
accuracy: 1.0000
Epoch 60/100
23/23 [==============================] - 0s 5ms/step - loss: 1.8021e-05 -
accuracy: 1.0000
Epoch 61/100
23/23 [==============================] - 0s 5ms/step - loss: 1.7144e-05 -
accuracy: 1.0000
Epoch 62/100
23/23 [==============================] - 0s 5ms/step - loss: 1.6420e-05 -
accuracy: 1.0000
Epoch 63/100
23/23 [==============================] - 0s 5ms/step - loss: 1.5721e-05 -
accuracy: 1.0000
Epoch 64/100
23/23 [==============================] - 0s 5ms/step - loss: 1.5049e-05 -
accuracy: 1.0000
Epoch 65/100
23/23 [==============================] - 0s 5ms/step - loss: 1.4407e-05 -
accuracy: 1.0000
Epoch 66/100
23/23 [==============================] - 0s 6ms/step - loss: 1.3820e-05 -
accuracy: 1.0000
Epoch 67/100
23/23 [==============================] - 0s 5ms/step - loss: 1.3261e-05 -
accuracy: 1.0000
Epoch 68/100
23/23 [==============================] - 0s 5ms/step - loss: 1.2744e-05 -
accuracy: 1.0000
Epoch 69/100
23/23 [==============================] - 0s 5ms/step - loss: 1.2288e-05 -
accuracy: 1.0000
Epoch 70/100
23/23 [==============================] - 0s 5ms/step - loss: 1.1762e-05 -
accuracy: 1.0000
Epoch 71/100
23/23 [==============================] - 0s 5ms/step - loss: 1.1280e-05 -
accuracy: 1.0000
Epoch 72/100
23/23 [==============================] - 0s 5ms/step - loss: 1.0913e-05 -
accuracy: 1.0000
Epoch 73/100
23/23 [==============================] - 0s 5ms/step - loss: 1.0488e-05 -
accuracy: 1.0000
Epoch 74/100
23/23 [==============================] - 0s 5ms/step - loss: 1.0110e-05 -
```

```
accuracy: 1.0000
Epoch 75/100
23/23 [==============================] - 0s 5ms/step - loss: 9.7530e-06 -
accuracy: 1.0000
Epoch 76/100
23/23 [==============================] - 0s 5ms/step - loss: 9.3964e-06 -
accuracy: 1.0000
Epoch 77/100
23/23 [==============================] - 0s 5ms/step - loss: 9.0769e-06 -
accuracy: 1.0000
Epoch 78/100
23/23 [==============================] - 0s 5ms/step - loss: 8.7834e-06 -
accuracy: 1.0000
Epoch 79/100
23/23 [==============================] - 0s 5ms/step - loss: 8.4797e-06 -
accuracy: 1.0000
Epoch 80/100
23/23 [==============================] - 0s 5ms/step - loss: 8.2046e-06 -
accuracy: 1.0000
Epoch 81/100
23/23 [==============================] - 0s 6ms/step - loss: 7.9997e-06 -
accuracy: 1.0000
Epoch 82/100
23/23 [==============================] - 0s 6ms/step - loss: 7.6418e-06 -
accuracy: 1.0000
Epoch 83/100
23/23 [==============================] - 0s 6ms/step - loss: 7.3748e-06 -
accuracy: 1.0000
Epoch 84/100
23/23 [==============================] - 0s 5ms/step - loss: 7.1586e-06 -
accuracy: 1.0000
Epoch 85/100
23/23 [==============================] - 0s 5ms/step - loss: 6.9302e-06 -
accuracy: 1.0000
Epoch 86/100
23/23 [==============================] - 0s 5ms/step - loss: 6.6984e-06 -
accuracy: 1.0000
Epoch 87/100
23/23 [==============================] - 0s 5ms/step - loss: 6.4703e-06 -
accuracy: 1.0000
Epoch 88/100
23/23 [==============================] - 0s 5ms/step - loss: 6.2849e-06 -
accuracy: 1.0000
Epoch 89/100
23/23 [==============================] - 0s 5ms/step - loss: 6.0760e-06 -
accuracy: 1.0000
Epoch 90/100
23/23 [==============================] - 0s 5ms/step - loss: 5.8997e-06 -
```

```
accuracy: 1.0000
Epoch 91/100
23/23 [==============================] - 0s 5ms/step - loss: 5.7603e-06 -
accuracy: 1.0000
Epoch 92/100
23/23 [==============================] - 0s 5ms/step - loss: 5.5731e-06 -
accuracy: 1.0000
Epoch 93/100
23/23 [==============================] - 0s 5ms/step - loss: 5.3998e-06 -
accuracy: 1.0000
Epoch 94/100
23/23 [==============================] - 0s 5ms/step - loss: 5.2463e-06 -
accuracy: 1.0000
Epoch 95/100
23/23 [==============================] - 0s 5ms/step - loss: 5.0922e-06 -
accuracy: 1.0000
Epoch 96/100
23/23 [==============================] - 0s 5ms/step - loss: 4.9627e-06 -
accuracy: 1.0000
Epoch 97/100
23/23 [==============================] - 0s 5ms/step - loss: 4.8138e-06 -
accuracy: 1.0000
Epoch 98/100
23/23 [==============================] - 0s 5ms/step - loss: 4.6756e-06 -
accuracy: 1.0000
Epoch 99/100
23/23 [==============================] - 0s 5ms/step - loss: 4.5456e-06 -
accuracy: 1.0000
Epoch 100/100
23/23 [==============================] - 0s 5ms/step - loss: 4.4270e-06 -
accuracy: 1.0000
```

[131]: <keras.callbacks.History at 0x22778251640>

[132]:
```
y_pred=model.predict(x_test)
y_pred=(y_pred>0.5)
```

[133]:
```
print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.91      0.95      0.93       311
           1       0.63      0.47      0.54        57

    accuracy                           0.88       368
   macro avg       0.77      0.71      0.73       368
weighted avg       0.86      0.88      0.87       368
```

5 Conclusions/Results

(A) The follwoing conclussions can be drawn from the correlation matix (Figure 3):

1) Overtime is very strongly correlated with attrition. This means people don't like to work overtime. Also "Distance from home" and "NumCompaniesWorked" are also positively correlated with attrition. Performace rating also have small correlation with attrition.

2) There are different factors which are negatively correlated to the attrition but their values is very low:

   a) "Age" -people with more age like to stay in the company.

   b) Job involvement-Employees' more involved in the work, attrition is low.

   c) Job level- Means people at higher position like to stay in the company.

   d) Monthly Income- High income emplyees like to stay in the company.

   e) Total Working years- Employees with more experinces like to stay in the company.

   f) year At comapny-Menas Emplyees working from long time in the company, they dont want to leave the comapny.

   g) Years in current role- It looks from the data that people don't like to change their role.

   h) Years with current manager- More emplyees like to work with the same manager.

3) There is very strong correlation have been observed :

   a) YearsAtCompany, Totalworkingyears, YearsIn CurentRole, YearsSinceLast Promotion and YearsWithCurrentManagers are all related field and function of time and increased with time. Similarly Age, Job level and MonthlyIncome are also raised with these five factors.

   b) There is 77% correlation has been found between performace rating and PercentSalary-Hike, which is understood that the company raise the salary based on the performace of the employeees.

   c) Also there is 95% correlation haas been observed between monthly income and Job level.

(B) The main obervation from the exploritory analysis are as discussed below:-

1) We can observer from Figure 7 that the attrition rate is more for the age group fom 18-35 years. Every department have mixed employees of all age groups(Figure 15).

2) Attrition rate is higher for employess who travel frequently and lowest for non traveler. It can be said that business travel is playing an important role in the attriton of employees (Figure 8).

3) Attrition is higher for single employees and marrited and divorced employees would not like to move to another place (Figure 9).

4) Figure 11 indicates that married employees who travel frquently have higher percentage of attrition than non traveler and rarely traveler. This percentage is more higher for the em-

ployees who are divorced. May be they are single parent and dont want to travel. It is also possible that the company give more travel responsibilities to single employees.

5) The sales and human resource employees have higher chances to leave the company. May be they have more exposer to new opportunities and have strong network connections (Figure 14).

6) Figure 16 shows that the sales repersentative have the highest percentage attrition.

7) It is clear from Figure 17 that there the attrition is not specific to any/some field(s) of education. Also we canont say that the attrition is happening becuse of the working in different area from their field of education (Figure 18).

8) Again no conclusion can be drawn (Figure 19) about attrition based on mismatch between education and job roles. Already most of the employees in the company are in their corresponding field of Education. No marketing educated employee in the role of research directors, research scientist, laboratory technician, etc. and similar for other cases. However, it can be observed that almost 50% sales representative who have marketing degree left the company. It is already observed that sales representatives among the maximum who left the comapany. May be the are not getting enough salery or there is an issue with the manager. It can also be seen that the 27% Laboratory Assitants with life sciecne degree left the company and 33% human resource employees with human resource education left the comapany.

9) It can be obereved from Figure 20 that the percentage attrition is more around 25km. However, most of the employees are living closer than 35 km fromthe office. This is the distance which maximum people can travel. But we can see the effect of distace from home.

10) There is not direct relationship between level of education and attrition.

11) It can be observed from Figure 22 that highly educated level (4, 5) human resource employees have higher percentage of attrition. Also the manufacturing directors with level 5 education have greater chances for attrition. For other roles the employees look satisfied with their postion regardless of their education. We can also observed that research directors have very low chances of attrition. This is may most of them are older and does not want to change the job since the average age for research directors and managers are approximately 44 years and 47 years (Figure 23). However, Figure 24 indicates that there is less attrition the lower age group for these Job Roles.

12) It can be observed that the environmental satisfaction can be one of the factor responsible for attrition. We can observed the decreasing trend of attrition with increasing level of environmental satisfaction (Figure 25). The research directors are among those who have lowest envirmonmental satisafction but manufactoring directors are highly satisfied by their working envirmonemnts. Simlarly highly educated employees are less satisfied by their environment as compared others. Also employees who are educated in the human resources are less satisfied with the environment (Figure 26).

13) It can be observed that the percentage of attrition is almost same for male and female, actually the female attrition is less (Figure 27).

14) The bargraph (Figure 28) shows that the attrition rate decreasing with the increase in job involvement which is also indicated from the correlation matrix. It can also be observed that attrition is more for lower levels of job involments for all level of education except level 5 (Figure 29).

15) It is clear from correlation matrix (Figure 3) that the employees dont like to work for overtime.The same can be confirmed from Figure 30. Figure 31 shows that the overtime percentage is almost same in all income regions. However, we can noticed that about half of the emmplyees who are have salary between 12000 to 14000 range are doing overtime. If we look in the salary range, we can say that the research directors might need to do more the overtime as most of them are among 12000 to 14000 salary range. Also they have the lowest level of environmental satisfaction (Figure 32). However, if we look into the job roles verses overtime plot (Figure 33), we cannot say only research directors are among who are doing more overitme. Actually, research scientist have the higher percentage of employees who are doing overtime. Overtime is almost same for each role within the range of 10% differences.

16) The job satisafaction also contribute for attrition. The employees with job satiasaction level "1", have higher chances to left the company. The corellation matrix also shows negative correlation between job satisfaction and attrition. From correlation matrix, we have not observed any correlation between job satisfaction and any other parameter. The LM plot in Figure 35 (A) shows that the job satisfaction decresing for the human resource and research scientist with the increase in monthly income. Also employees at joblevel 4 are getting less satisfied with the increase in monthly income (Figure 35 (B)).

17) We can see a downword trend of atrition with the increase of monthlyIncome (from Figure 36) with some exeptions. Employees with monthly income lower 3300 have higher chances to left the company. Mainly human resource, sales executive, research scintist and laboratory scientist fall under this weges group.

18) Figure 37 shows that the attrition is increasing with the number of companies workeed but it is not a straight line. No relation has been observed between percentage salary hike and attrition. So we can say that the emplyees who left the company did not leave because they want salary hike (Figure 38). The comapny might providing the proper hike to the employees based on their performance. We can also notice that the salary hike does not have any relation with the experince/years at company and uniform for all employees. But have strongly related to performance rating (Figure 40).

19) The sales and research development has higer hike than human resource. Also sales representative hike is more than any other role. May be the hike is given to overcome high attrition among sales representative (Figure 41).

20) Relationship satisfaction with level 1 has higher attrition rate but the difference is not much to be consider as a big factor responsible for attrition.

21) Stock option level looks playing responsible for attrition. The emplyees with level 0 have higher rate of attrition. The trend is dicreasing for level 1 and 2 but for level 3 it again raised. However we can say that providing higher level of stoock option decreased the attrition (Figure 43).

22) If we look into the training times in a year, we can observed that most of the employees of all roles are getting 2 and 3 times training in the company (Figure 46). Providing more training should increase the performace but we have not found any correlation. However, if we search performce of differnt roles we can observed that the performace rating increasing only for sales representative. So it is better to provide more training to sales representatives than other roles. But providing training more that 5 times is actually start droping their performace (Figure 47).

23) The employees with WorkLife balance of level 1 have more chances to left the company. The work life balance decreasing for sales representatives and sales executive with the increase of job Level. Another way to look into that the work life balance decreasing for sales representatives and sales executive with the increase of monthly income (Figure 50). This may be the one of the main cause for more attrition among sales sepresentatives.

24) Research scientists, laboratory Technician and Sales respresntative are among the lowest salery group but their experince is below 20 years. Most of the research directors and managers have experience of more than 20 years and they are in higher income range. All other fall in between. The similar trend has been observed for years in the company (Figure 53).

25) The count plot in Figure 54 shows that maximum employees stayes at the same role for their 1st, 3rd and 7th year. More percentage of employees left after 1st year (almost 30%). We can also observe that the attrition is more for employees who are in their 1st, 2nd and 3rd year at the same role. May be they are waiting for pemotions and left. From 3rd to 7th year, the number of emplyees decreasing long with the huge decrease in attrition rate which indicates that most of the employees get permoted continously or their roles get changed. Also more employees are staying at the same role for their 7th year and then their is continous decrease which shows that the company give continous promotions as attrition rate is very low for these years.

[ ]:

```

```