In [1]:
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

In [3]:
```
df=pd.read_csv('WA_Fn-UseC_-HR-Employee-Attrition.csv')
df
```

Out[3]:

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	Em
0	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	1	
1	49	No	Travel_Frequently	279	Research & Development	8	1	Life Sciences	1	
2	37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other	1	
3	33	No	Travel_Frequently	1392	Research & Development	3	4	Life Sciences	1	
4	27	No	Travel_Rarely	591	Research & Development	2	1	Medical	1	
...	...	...	...	...	...	...	...	...	...	
1465	36	No	Travel_Frequently	884	Research & Development	23	2	Medical	1	
1466	39	No	Travel_Rarely	613	Research & Development	6	1	Medical	1	
1467	27	No	Travel_Rarely	155	Research & Development	4	3	Life Sciences	1	
1468	49	No	Travel_Frequently	1023	Sales	2	3	Medical	1	
1469	34	No	Travel_Rarely	628	Research & Development	8	3	Medical	1	

1470 rows × 35 columns

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

Out[5]:

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	Emplo
0	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	1	
1	49	No	Travel_Frequently	279	Research & Development	8	1	Life Sciences	1	
2	37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other	1	
3	33	No	Travel_Frequently	1392	Research & Development	3	4	Life Sciences	1	
4	27	No	Travel_Rarely	591	Research & Development	2	1	Medical	1	

5 rows × 35 columns

In [7]:
```
df.tail()
```

Out[7]:

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	Em
1465	36	No	Travel_Frequently	884	Research & Development	23	2	Medical	1	
1466	39	No	Travel_Rarely	613	Research & Development	6	1	Medical	1	
1467	27	No	Travel_Rarely	155	Research & Development	4	3	Life Sciences	1	
1468	49	No	Travel_Frequently	1023	Sales	2	3	Medical	1	
1469	34	No	Travel_Rarely	628	Research & Development	8	3	Medical	1	

5 rows × 35 columns

In [23]:
```
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
```

In [ ]:

In [11]:
```
df['BusinessTravel'].value_counts(normalize=True)*100
```

Out[11]:
```
BusinessTravel
Travel_Rarely        70.952381
Travel_Frequently    18.843537
Non-Travel           10.204082
Name: proportion, dtype: float64
```

In [13]:
```
df['EducationField'].value_counts(normalize=True)*100
```

Out[13]:
```
EducationField
Life Sciences       41.224490
Medical             31.564626
Marketing           10.816327
Technical Degree     8.979592
Other                5.578231
Human Resources      1.836735
Name: proportion, dtype: float64
```

In [15]:
```
df['Department'].value_counts(normalize=True)*100
```

Out[15]:
```
Department
Research & Development    65.374150
Sales                     30.340136
Human Resources            4.285714
Name: proportion, dtype: float64
```

In [17]:
```
df.isnull()
```

Out[17]:

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	Emp
0	False	False	False	False	False	False	False	False	False	
1	False	False	False	False	False	False	False	False	False	
2	False	False	False	False	False	False	False	False	False	
3	False	False	False	False	False	False	False	False	False	
4	False	False	False	False	False	False	False	False	False	
...	...	...	...	...	...	...	...	...	...	
1465	False	False	False	False	False	False	False	False	False	
1466	False	False	False	False	False	False	False	False	False	
1467	False	False	False	False	False	False	False	False	False	
1468	False	False	False	False	False	False	False	False	False	
1469	False	False	False	False	False	False	False	False	False	

1470 rows × 35 columns

In [21]:
```
df.drop_duplicates()
```

Out[21]:

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	Em
0	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	1	
1	49	No	Travel_Frequently	279	Research & Development	8	1	Life Sciences	1	
2	37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other	1	
3	33	No	Travel_Frequently	1392	Research & Development	3	4	Life Sciences	1	
4	27	No	Travel_Rarely	591	Research & Development	2	1	Medical	1	
...	...	...	...	...	...	...	...	...	...	
1465	36	No	Travel_Frequently	884	Research & Development	23	2	Medical	1	
1466	39	No	Travel_Rarely	613	Research & Development	6	1	Medical	1	
1467	27	No	Travel_Rarely	155	Research & Development	4	3	Life Sciences	1	
1468	49	No	Travel_Frequently	1023	Sales	2	3	Medical	1	
1469	34	No	Travel_Rarely	628	Research & Development	8	3	Medical	1	

1470 rows × 35 columns

In [38]:
```
df1 = df.loc[:, df.nunique() > 1]
df1
```

Out[38]:

	Age	DailyRate	DistanceFromHome	Education	EnvironmentSatisfaction	HourlyRate	JobInvolvement	JobLevel	JobSatisfact
0	41	1102	1	2	2	94	3	2	
1	49	279	8	1	3	61	2	2	
2	37	1373	2	2	4	92	2	1	
3	33	1392	3	4	4	56	3	1	
4	27	591	2	1	1	40	3	1	
...	...	...	...	...	...	...	...	...	
1465	36	884	23	2	3	41	4	2	
1466	39	613	6	1	4	42	2	3	
1467	27	155	4	3	2	87	4	2	
1468	49	1023	2	3	4	63	2	2	
1469	34	628	8	3	2	82	4	2	

1470 rows × 45 columns

In [40]:
```
df2 = pd.get_dummies(df, drop_first=True)
df2
```

Out[40]:

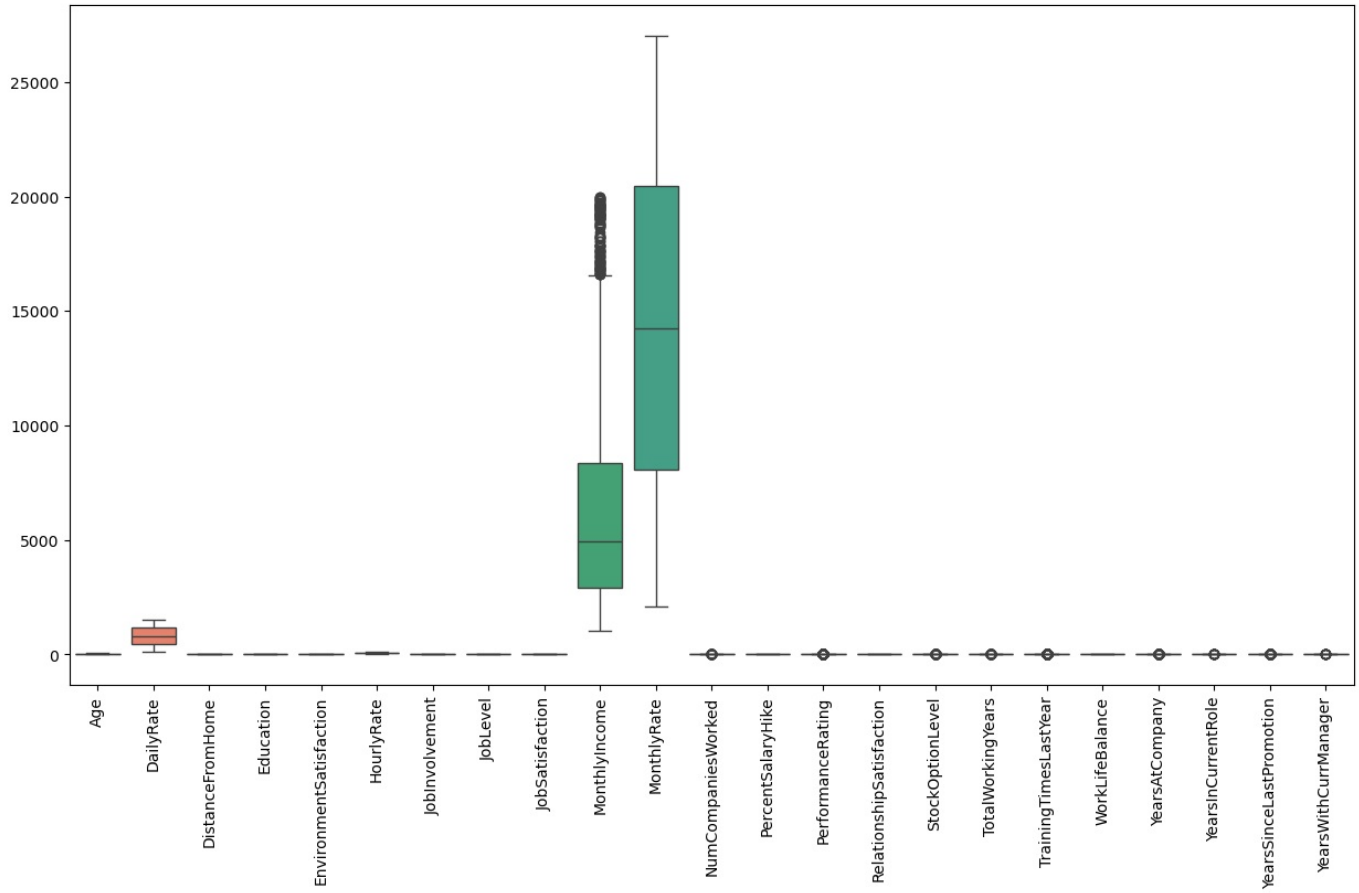
	Age	DailyRate	DistanceFromHome	Education	EnvironmentSatisfaction	HourlyRate	JobInvolvement	JobLevel	JobSatisfact
0	41	1102	1	2	2	94	3	2	
1	49	279	8	1	3	61	2	2	
2	37	1373	2	2	4	92	2	1	
3	33	1392	3	4	4	56	3	1	
4	27	591	2	1	1	40	3	1	
...	...	...	...	...	...	...	...	...	
1465	36	884	23	2	3	41	4	2	
1466	39	613	6	1	4	42	2	3	
1467	27	155	4	3	2	87	4	2	
1468	49	1023	2	3	4	63	2	2	
1469	34	628	8	3	2	82	4	2	

1470 rows × 45 columns

In [42]:
```
df.dtypes
```

Out[42]:
```
Age                                int64
DailyRate                          int64
DistanceFromHome                   int64
Education                          int64
EnvironmentSatisfaction            int64
HourlyRate                         int64
JobInvolvement                     int64
JobLevel                           int64
JobSatisfaction                    int64
MonthlyIncome                      int64
MonthlyRate                        int64
NumCompaniesWorked                 int64
PercentSalaryHike                  int64
PerformanceRating                  int64
RelationshipSatisfaction           int64
StockOptionLevel                   int64
TotalWorkingYears                  int64
TrainingTimesLastYear              int64
WorkLifeBalance                    int64
YearsAtCompany                     int64
YearsInCurrentRole                 int64
YearsSinceLastPromotion            int64
YearsWithCurrManager               int64
Attrition_Yes                       bool
BusinessTravel_Travel_Frequently    bool
BusinessTravel_Travel_Rarely        bool
Department_Research & Development   bool
Department_Sales                    bool
EducationField_Life Sciences        bool
EducationField_Marketing            bool
EducationField_Medical              bool
EducationField_Other                bool
EducationField_Technical Degree     bool
Gender_Male                         bool
JobRole_Human Resources             bool
JobRole_Laboratory Technician       bool
JobRole_Manager                     bool
JobRole_Manufacturing Director      bool
JobRole_Research Director           bool
JobRole_Research Scientist          bool
JobRole_Sales Executive             bool
JobRole_Sales Representative        bool
MaritalStatus_Married               bool
MaritalStatus_Single                bool
OverTime_Yes                        bool
dtype: object
```

In [44]:
```
plt.figure(figsize=(15, 8))
sns.boxplot(data=df.select_dtypes(include=['int64', 'float64']))
plt.xticks(rotation=90)
plt.show()
```



In [46]:
```
df.describe()
print("Final dataset shape:", df.shape)
```

```
Final dataset shape: (1470, 45)
```

In [48]:
```
df.to_csv('HR_analysis')
```

In [ ]: