AFAME TECHNOLOGIES: HR DATA ANALYSIS

Importing Libraries

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
import warnings
warnings.simplefilter('ignore')
```

Dataset: HR DATA ANALYSIS

```
data=pd.read_csv("HR Data.csv")
data
```

	Age int64 18 - 60	Attrition object No 83.9% Yes 16.1%	BusinessTravel obj... Travel_Rarely 71% Travel_Freq... 18.8% Non-Travel 10.2%	DailyRate int64 102 - 1499	Department object Research &... 65.4% Sales 30.3% Human Reso... 4.3%	DistanceFromHome i 1 - 29	Education int64 1 - 5	EducationField obj... Life Sciences 41.2% Medical 31.6% 4 others 27.2%
0	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences
1	49	No	Travel_Frequently	279	Research & Devel...	8	1	Life Sciences
2	37	Yes	Travel_Rarely	1373	Research & Devel...	2	2	Other
3	33	No	Travel_Frequently	1392	Research & Devel...	3	4	Life Sciences
4	27	No	Travel_Rarely	591	Research & Devel...	2	1	Medical
5	32	No	Travel_Frequently	1005	Research & Devel...	2	2	Life Sciences
6	59	No	Travel_Rarely	1324	Research & Devel...	3	3	Medical
7	30	No	Travel_Rarely	1358	Research & Devel...	24	1	Life Sciences
8	38	No	Travel_Frequently	216	Research & Devel...	23	3	Life Sciences
9	36	No	Travel_Rarely	1299	Research & Devel...	27	3	Medical

1470 rows, 35 cols, showing 10 rows/page Page 1 of 147

```
data.describe()
```

	Age float64	DailyRate float64	DistanceFromHome f	Education float64	EmployeeCount flo...	EmployeeNumber f...	EnvironmentSatisf...	HourlyRate float64
cou...	1470	1470	1470	1470	1470	1470	1470	1470
me...	36.92380952	802.4857143	9.192517007	2.91292517	1	1024.865306	2.721768707	65.89115646
std	9.135373489	403.5090999	8.106864436	1.024164945	0	602.0243348	1.093082215	20.32942759
min	18	102	1	1	1	1	1	30
25%	30	465	2	2	1	491.25	2	48
50%	36	802	7	3	1	1020.5	3	66
75%	43	1157	14	4	1	1555.75	4	83.75
max	60	1499	29	5	1	2068	4	100

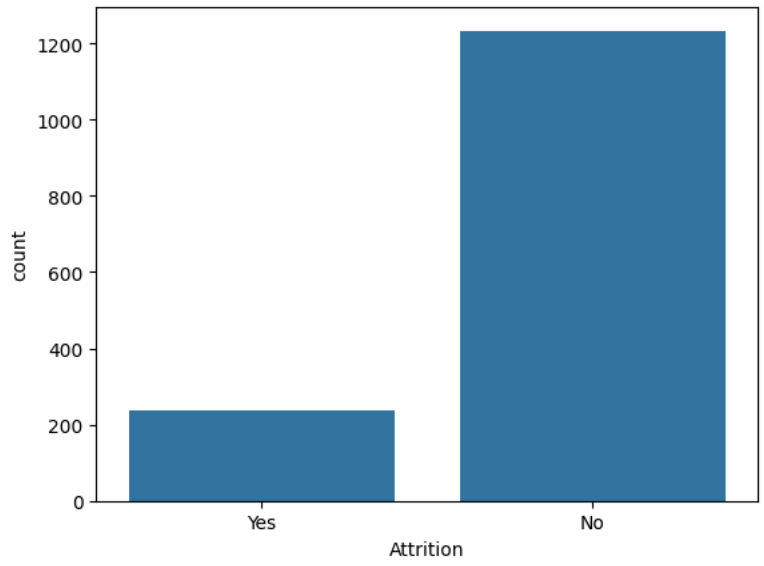
8 rows, 26 cols, showing 10 rows/page Page 1 of 1

```
data.info()
```

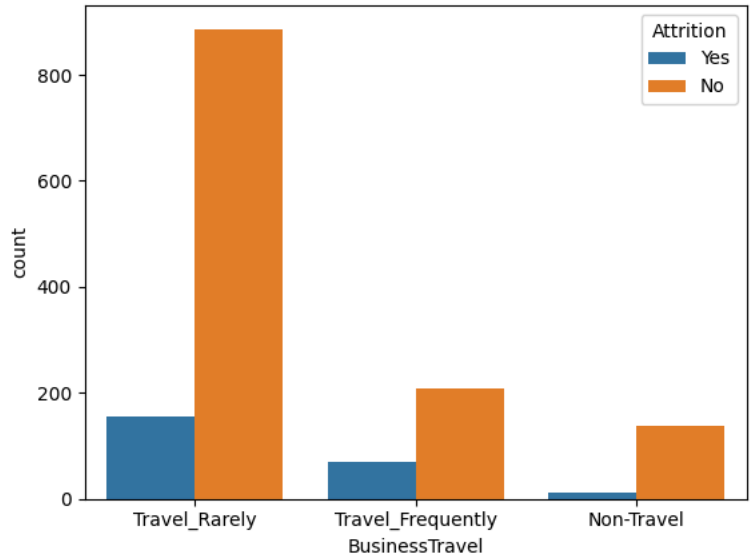
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1470 entries, 0 to 1469
Data columns (total 35 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Age                       1470 non-null   int64
 1   Attrition                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

Data Visualization

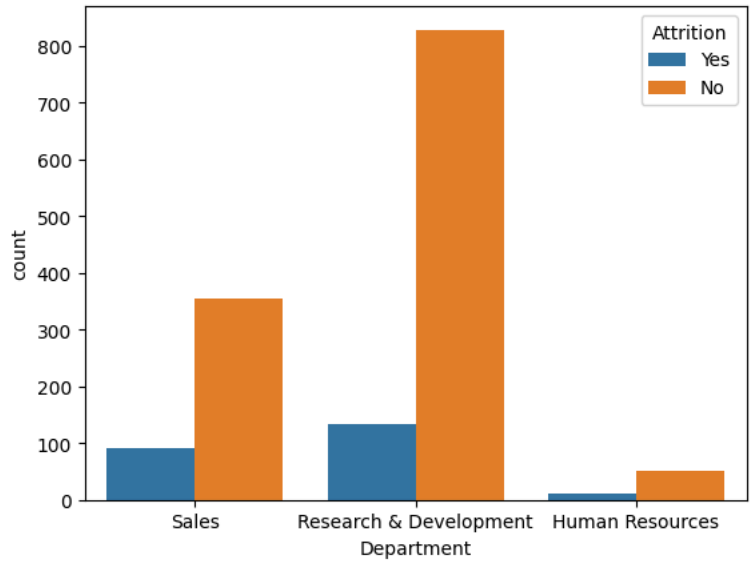
```
sns.countplot(x=data.Attrition)
plt.show()
```



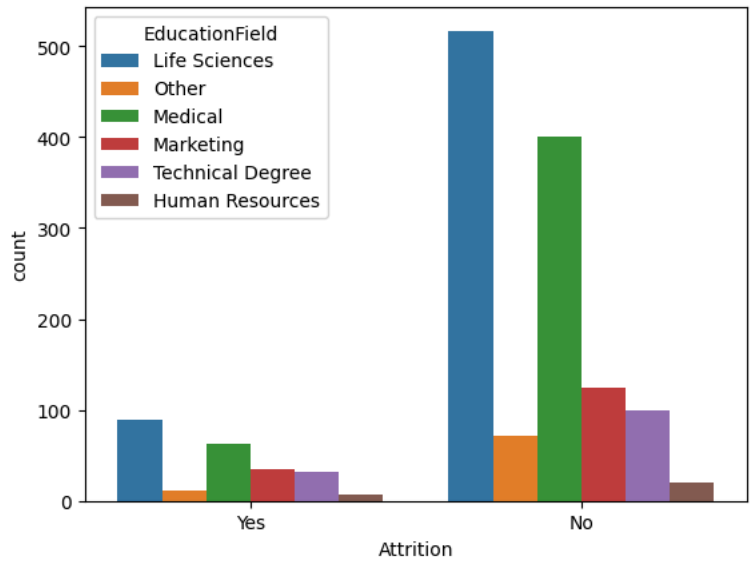
```
sns.countplot(hue=data.Attrition, x = data.BusinessTravel)
plt.show()
```



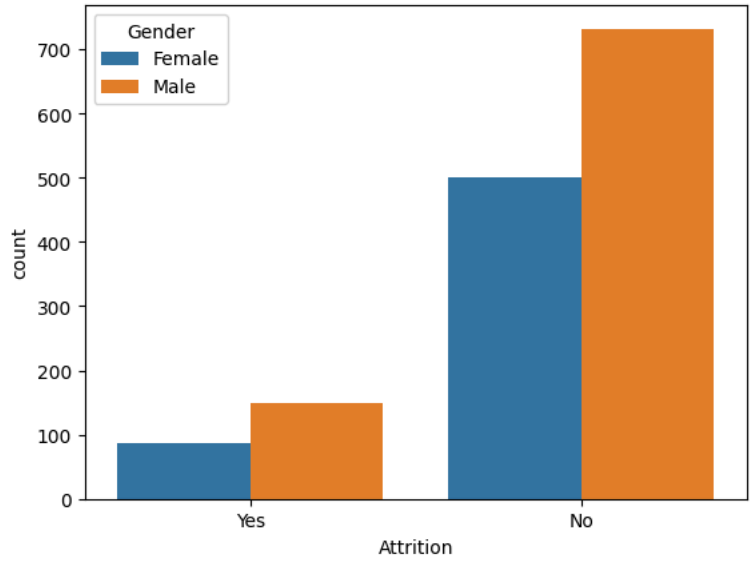

```
sns.countplot(hue=data.Attrition, x=data.Department)
plt.show()
```



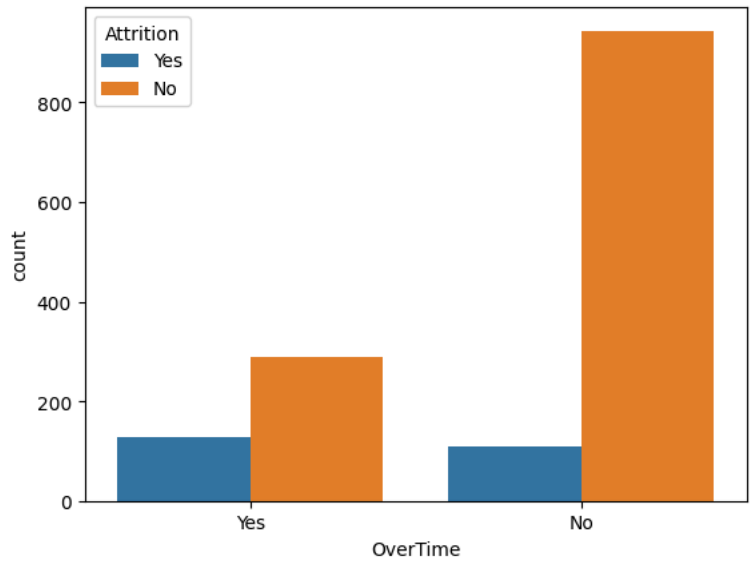
```
sns.countplot(x=data.Attrition, hue=data.EducationField)
plt.show()
```




```
sns.countplot(x=data.Attrition, hue=data.Gender)
plt.show()
```

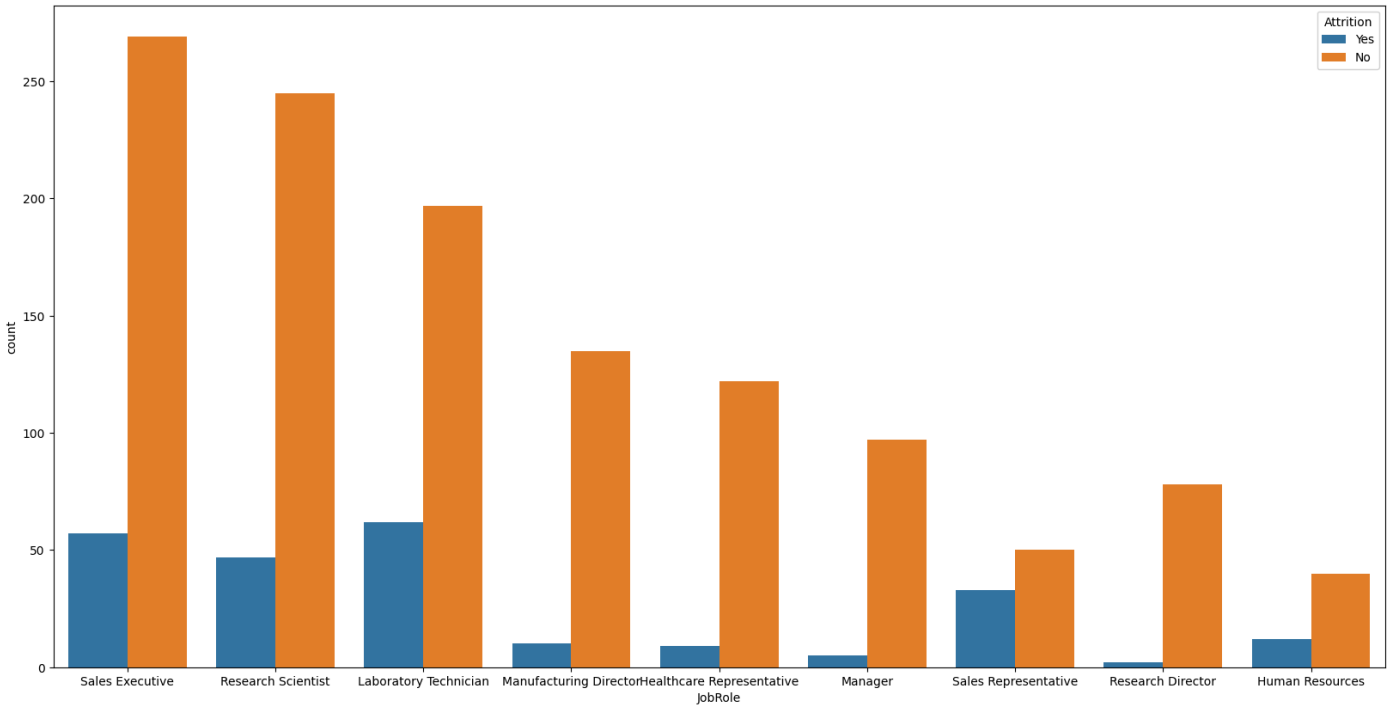


```
sns.countplot(hue=data.Attrition, x=data.OverTime)
plt.show()
```




```
plt.figure(figsize=(20,10), facecolor='white')
sns.countplot(x='JobRole', hue='Attrition',data=data)
plt.xlabel('JobRole',fontsize=10)
```

```
Text(0.5, 0, 'JobRole')
```



```
data.isnull().sum()
```

```
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
```

```
data.shape
```

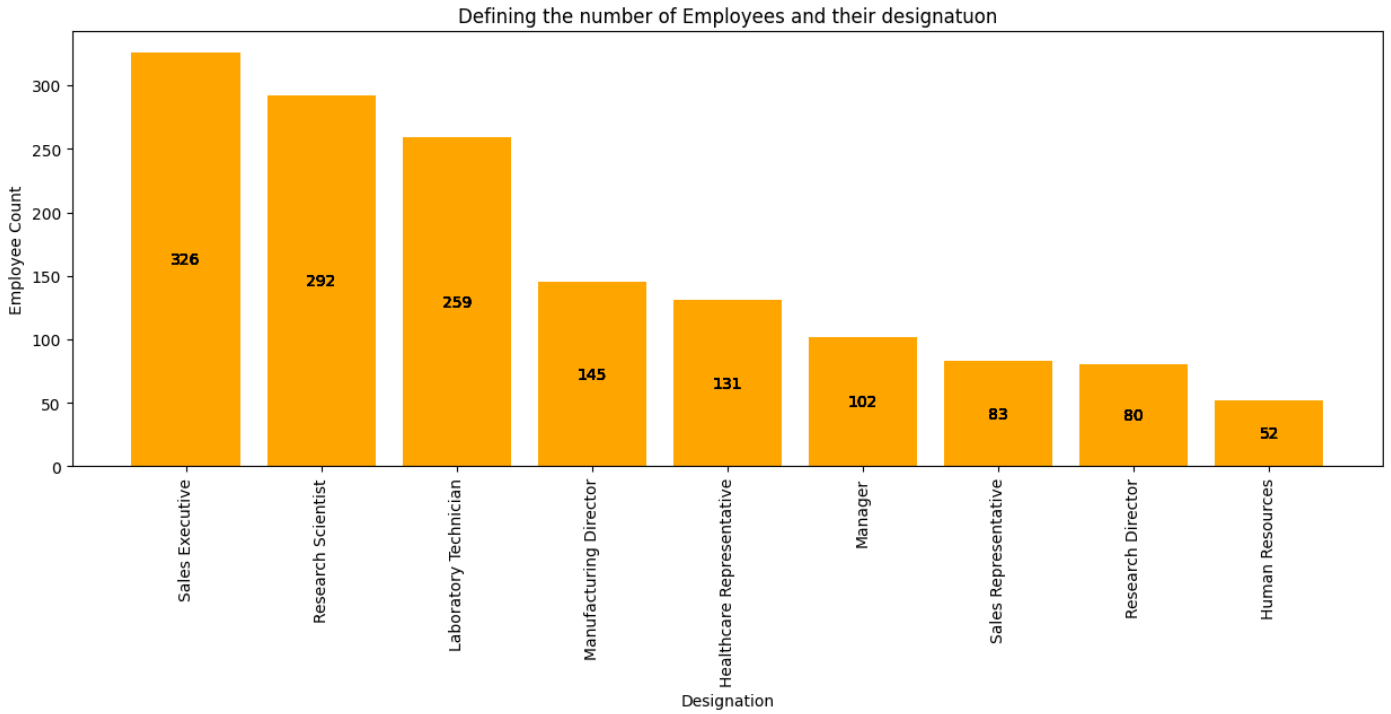
```
(1470, 35)
```

```
job_role_num=data['JobRole'].value_counts()
```

```
job_role_num
```

```
JobRole
Sales Executive              326
Research Scientist           292
Laboratory Technician        259
Manufacturing Director       145
Healthcare Representative    131
Manager                      102
Sales Representative          83
Research Director             80
Human Resources               52
Name: count, dtype: int64
```

```
plt.figure(figsize=(15,5))
for i in job_role_num.items():
    p = plt.bar(job_role_num.index, job_role_num.values,color='orange')
    plt.bar_label(p, label_type='center')
plt.xlabel('Designation')
plt.ylabel('Employee Count')
plt.title('Defining the number of Employees and their designatuon')
plt.xticks(rotation='vertical')
plt.show()
```



Data Preprocessing

```
data.columns
```

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

```
data = data.drop(columns=['EmployeeCount', 'StandardHours', 'Over18', 'EmployeeNumber'])
```

```
data.columns
```

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

```
data = data.rename(columns={
    'Attrition': 'Turnover',
    'BusinessTravel': 'Business_Travel',
    'DailyRate': 'Daily_Rate',
    'DistanceFromHome': 'Distance_From_Home',
    'EducationField': 'Education_Field',
    'EnvironmentSatisfaction': 'Environment_Satisfaction',
    'HourlyRate': 'Hourly_Rate',
    'JobInvolvement': 'Job_Involvement',
    'JobLevel': 'Job_Level',
    'JobRole': 'Job_Role',
    'JobSatisfaction': 'Job_Satisfaction',
    'MaritalStatus': 'Marital_Status',
    'MonthlyIncome': 'Monthly_Income',
    'MonthlyRate': 'Monthly_Rate',
    'NumCompaniesWorked': 'Num_Companies_Worked',
    'OverTime': 'Over_Time',
    'PercentSalaryHike': 'Percent_Salary_Hike',
    'PerformanceRating': 'Performance_Rating',
    'RelationshipSatisfaction': 'Relationship_Satisfaction',
    'StockOptionLevel': 'Stock_Option_Level',
    'TotalWorkingYears': 'Total_Working_Years',
    'TrainingTimesLastYear': 'Training_Times_Last_Year',
    'WorkLifeBalance': 'Work_Life_Balance',
    'YearsAtCompany': 'Years_At_Company',
    'YearsInCurrentRole': 'Years_In_Current_Role',
    'YearsSinceLastPromotion': 'Years_Since_Last_Promotion',
    'YearsWithCurrManager': 'Years_With_Curr_Manager'
})
```

```
data['Turnover'] = data['Turnover'].astype('category')
data['Business_Travel'] = data['Business_Travel'].astype('category')
data['Department'] = data['Department'].astype('category')
data['Education_Field'] = data['Education_Field'].astype('category')
data['Gender'] = data['Gender'].astype('category')
data['Job_Role'] = data['Job_Role'].astype('category')
data['Marital_Status'] = data['Marital_Status'].astype('category')
data['Over_Time'] = data['Over_Time'].astype('category')

#Eliminate NaN values
data = data.dropna()

data = data.drop_duplicates()
```

```
print(data.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   Turnover                  1470 non-null   category
 2   Business_Travel           1470 non-null   category
 3   Daily_Rate                1470 non-null   int64
 4   Department                1470 non-null   category
 5   Distance_From_Home        1470 non-null   int64
 6   Education                 1470 non-null   int64
 7   Education_Field           1470 non-null   category
 8   Environment_Satisfaction  1470 non-null   int64
 9   Gender                    1470 non-null   category
 10  Hourly_Rate               1470 non-null   int64
 11  Job_Involvement           1470 non-null   int64
 12  Job_Level                 1470 non-null   int64
 13  Job_Role                  1470 non-null   category
 14  Job_Satisfaction          1470 non-null   int64
 15  Marital_Status            1470 non-null   category
 16  Monthly_Income            1470 non-null   int64
 17  Monthly_Rate              1470 non-null   int64
 18  Num_Companies_Worked      1470 non-null   int64
 19  Over_Time                 1470 non-null   category
 20  Percent_Salary_Hike       1470 non-null   int64
 21  Performance_Rating        1470 non-null   int64
 22  Relationship_Satisfaction 1470 non-null   int64
 23  Stock_Option_Level        1470 non-null   int64
 24  Total_Working_Years       1470 non-null   int64
```

```
data
```

	Age int64 18 - 60	Turnover category No 83.9% Yes 16.1%	Business_Travel ca... Travel_Rarely 71% Travel_Freq... 18.8% Non-Travel 10.2%	Daily_Rate int64 102 - 1499	Department categ... Research &... 65.4% Sales 30.3% Human Reso... 4.3%	Distance_From_H... 1 - 29	Education int64 1 - 5	Education_Field ca... Life Sciences 41.2% Medical 31.6% 4 others 27.2%
0	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences
1	49	No	Travel_Frequently	279	Research & Devel...	8	1	Life Sciences
2	37	Yes	Travel_Rarely	1373	Research & Devel...	2	2	Other
3	33	No	Travel_Frequently	1392	Research & Devel...	3	4	Life Sciences
4	27	No	Travel_Rarely	591	Research & Devel...	2	1	Medical
5	32	No	Travel_Frequently	1005	Research & Devel...	2	2	Life Sciences
6	59	No	Travel_Rarely	1324	Research & Devel...	3	3	Medical
7	30	No	Travel_Rarely	1358	Research & Devel...	24	1	Life Sciences
8	38	No	Travel_Frequently	216	Research & Devel...	23	3	Life Sciences
9	36	No	Travel_Rarely	1299	Research & Devel...	27	3	Medical

1470 rows, 31 cols, showing 10 rows/page Page 1 of 147

Checking whether NaN value is present or not.

```
data.isnull().sum()
```

```
Age                         0
Turnover                    0
Business_Travel             0
Daily_Rate                  0
Department                  0
Distance_From_Home          0
Education                   0
Education_Field             0
Environment_Satisfaction    0
Gender                      0
Hourly_Rate                 0
Job_Involvement             0
Job_Level                   0
Job_Role                    0
Job_Satisfaction            0
Marital_Status              0
Monthly_Income              0
Monthly_Rate                0
Num_Companies_Worked        0
Over_Time                   0
Percent_Salary_Hike         0
Performance_Rating          0
Relationship_Satisfaction   0
Stock_Option_Level          0
Total_Working_Years         0
Training_Times_Last_Year    0
Work_Life_Balance           0
Years_At_Company            0
Years_In_Current_Role       0
Years_Since_Last_Promotion  0
Years_With_Curr_Manager     0
dtype: int64
```

Cleaned Dataset

```
# cleaned dataset
data.to_csv('cleaned_hr_data.csv', index=False)
```