HR EMPLOYEE ATTRITION

Problem statement

Addressing and reducing employee attrition in the HR department of a Corporation by analyzing the underlying causes, exploring the impact on organizational performance, and developing effective retention strategies.

IMPORTING THE PYTHON LIBRARIES

In [2]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt # visualizing data %matplotlib inline
import seaborn as sns
```

IMPORTING DATASET

In [3]:
```
df=pd.read_csv(r'C:\Users\Dell\Desktop\meriskill\python code\HR-Employee-Attrition.
#to avoid error use encoding
```

In [24]:
```
df.shape
```

Out[24]:
```
(1470, 35)
```

DataFrame is made up of 1470 rows and 35 columns of data.

In [25]:
```
df.head()
```

Out[25]:

	ï»¿Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	Edu
0	41	Yes	Travel_Rarely	1102	Sales	1	2	
1	49	No	Travel_Frequently	279	Research & Development	8	1	
2	37	Yes	Travel_Rarely	1373	Research & Development	2	2	
3	33	No	Travel_Frequently	1392	Research & Development	3	4	
4	27	No	Travel_Rarely	591	Research & Development	2	1	

5 rows × 35 columns

The head() function is primarily used to view the first few rows of a dataset

In [26]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1470 entries, 0 to 1469
Data columns (total 35 columns):
 #   Column                    Non-Null Count  Dtype 
---  ------                    --------------  ----- 
 0   ï»¿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 [27]:
```
pd.isnull(df)
```

Out[27]:

	ï»¿Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	Ec
0	False	False	False	False	False	False	False	
1	False	False	False	False	False	False	False	
2	False	False	False	False	False	False	False	
3	False	False	False	False	False	False	False	
4	False	False	False	False	False	False	False	
...	...	...	...	...	...	...	...	
1465	False	False	False	False	False	False	False	
1466	False	False	False	False	False	False	False	
1467	False	False	False	False	False	False	False	
1468	False	False	False	False	False	False	False	
1469	False	False	False	False	False	False	False	

1470 rows × 35 columns

The isnull() method returns a DataFrame object where all the values are replaced with a Boolean value True for NULL values, and otherwise False.

In [28]:
```
pd.isnull(df).sum()
```

Out[28]:
```
ï»¿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
```

There no null values in any column

In [29]:
```
df.columns
```

Out[29]:
```
Index(['ï»¿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')
```

df.columns gives the heading of all the columns.We can clearly identify that the Age column needs proper renaming

In [31]:
```
#rename column
df.rename(columns= {'ï»¿Age':'Age'}, inplace=True)
```

In [31]:
```
df.describe()
```

Out[31]:

	Age	DailyRate	DistanceFromHome	Education	EmployeeCount	EmployeeNuml
count	1470.000000	1470.000000	1470.000000	1470.000000	1470.0	1470.000(
mean	36.923810	802.485714	9.192517	2.912925	1.0	1024.865:
std	9.135373	403.509100	8.106864	1.024165	0.0	602.024:
min	18.000000	102.000000	1.000000	1.000000	1.0	1.000(
25%	30.000000	465.000000	2.000000	2.000000	1.0	491.250(
50%	36.000000	802.000000	7.000000	3.000000	1.0	1020.500(
75%	43.000000	1157.000000	14.000000	4.000000	1.0	1555.750(
max	60.000000	1499.000000	29.000000	5.000000	1.0	2068.000(

8 rows × 26 columns

INTRODUCTION TO THE PROJECT

Target column is attrition
ATTRITION-Attrition is departure of employees from an organization for any reason
Attrition 'Yes' means employees want to leave the company
Attrition 'No' means employees dont want to leave the company
In this project we are going to identify the variable factors affecting ATTRITION

Data information

Age: Represents the age of the employees in the dataset.

Attrition: Likely a binary variable indicating whether an employee has left the company (attrition) or is still employed.

BusinessTravel: Indicates the frequency or type of business travel undertaken by the employees (e.g., "Travel Frequently," "Travel Rarely," "Non-Travel")

. **DailyRate**: The daily rate of pay or salary for the employees.

Department: Specifies the department or division within the company where the employees work (e.g., "Sales," "HR," "Research & Development").

DistanceFromHome: Shows the distance in miles between an employee's home and their workplace.

Education: Typically represents the highest level of education attained by the employees (e.g., High School, Bachelor's, Master's, etc.).

EducationField: Indicates the field or area of study in which the employee's education is focused (e.g., "Marketing," "Engineering," "Life Sciences").

EmployeeCount: The number of employees included in this particular dataset.

EmployeeNumber: A unique identifier or code for each employee.

EnvironmentSatisfaction: A measure of how satisfied employees are with their work environment.

Gender: Specifies the gender of the employees (e.g., "Male" or "Female").

HourlyRate: The hourly rate of pay or salary for the employees.

JobInvolvement: Reflects the level of involvement or engagement of employees in their jobs.

JobLevel: Represents the hierarchical level or position of employees within the company (e.g., "Entry-Level," "Mid-Level," "Senior-Level").

JobRole: Specifies the specific role or job title of the employees (e.g., "Sales Representative," "Research Scientist," "Manager").

JobSatisfaction: Indicates the satisfaction level of employees with their jobs.

MaritalStatus: Describes the marital status of employees (e.g., "Single," "Married," "Divorced").

MonthlyIncome: The total monthly income or salary of the employees.

MonthlyRate: A rate or amount associated with the monthly compensation of employees.

NumCompaniesWorked: Represents the number of different companies the employees have worked for.

Over18: This column seems unusual and might represent a binary variable, possibly indicating whether the employees are over 18 years old.

OverTime: Indicates whether employees work overtime or not.

PercentSalaryHike: Shows the percentage increase in salary or wage for the employees.

PerformanceRating: Reflects the performance rating or evaluation of employees.

RelationshipSatisfaction: A measure of how satisfied employees are with their relationships at work.

StandardHours: The standard number of working hours in a day or week.

StockOptionLevel: Reflects the level of stock options granted to employees.

TotalWorkingYears: Represents the total number of years of working experience for each employee.

TrainingTimesLastYear: Shows the number of training times or courses attended by employees in the last year.

WorkLifeBalance: Reflects the perceived balance between work and personal life by employees.

YearsAtCompany: The number of years an employee has been with the current company.

YearsInCurrentRole: Indicates the number of years an employee has spent in their current job role.

YearsSinceLastPromotion: The number of years since an employee's last promotion.

YearsWithCurrManager: Represents the number of years an employee has been under their current manager.

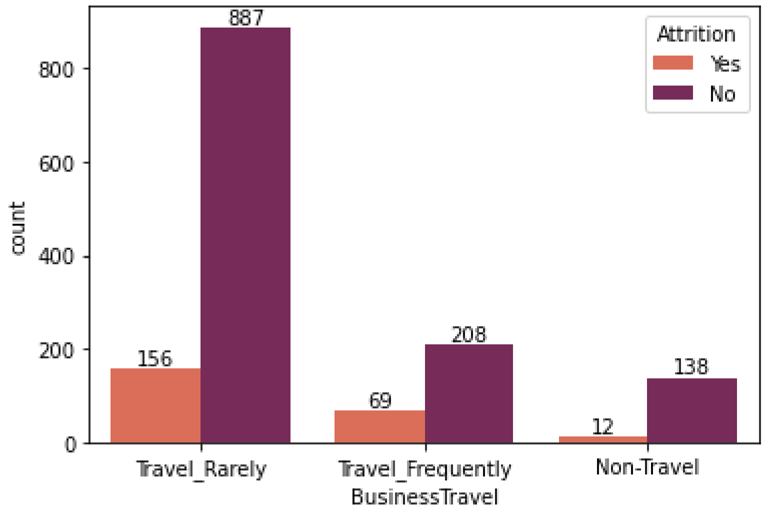
EXPLORATORY DATA ANALYSIS

1) Impact Of Business Travel On Attrition

In [4]:
```
ax=sns.countplot(data=df,x="BusinessTravel",hue="Attrition",palette='rocket_r')

for container in ax.containers:
        ax.bar_label(container, fmt="%d", label_type="edge")

plt.show()
```



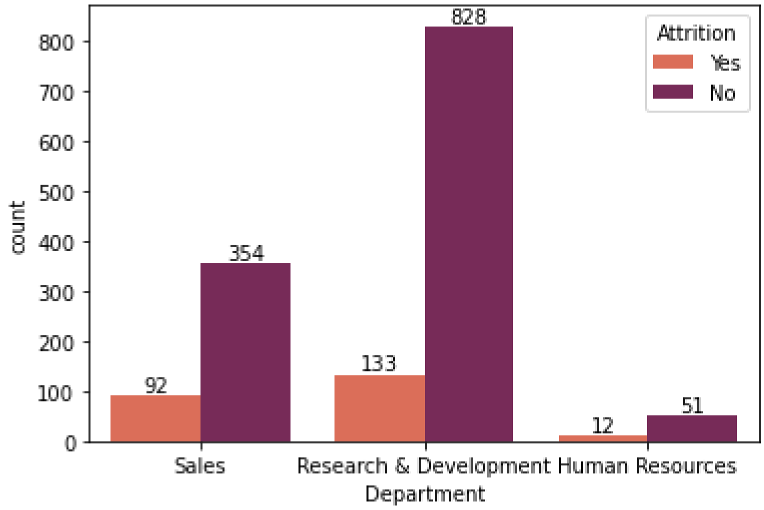
The graph tells us that a company has more count of employee who travel rarely.
Travel rate of employee is less.
There are 887 employees who travel rarely are satisfied by there job although 156 employees who travel rarely are not satisfied.that could be any other region.
14.95% of employees who travel rarely are not satisfied with the job

2)Impact of Department on Attrition

In [6]:
```
ax=sns.countplot(data=df,x="Department",hue="Attrition",palette='rocket_r')

for container in ax.containers:
        ax.bar_label(container, fmt="%d", label_type="edge")


plt.show()
```



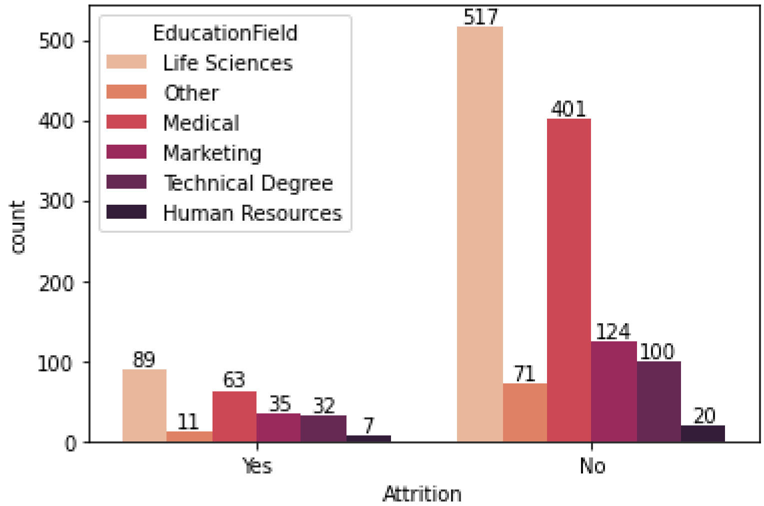
There are mainly three Departement-1>Sales 2>Research and Development 3>Human Resource.
Research and Developement has 133 attrition"yes" than other departments.
The number of employee in research and developement is also high.
20% of sales executive are leaving the company followed by 13% in Research and Developement.

3)Impact of Education Field on Attrition

In [9]:
```
ax=sns.countplot(data=df,x="Attrition",hue="EducationField",palette='rocket_r')

for container in ax.containers:
        ax.bar_label(container, fmt="%d", label_type="edge")

plt.show()
```



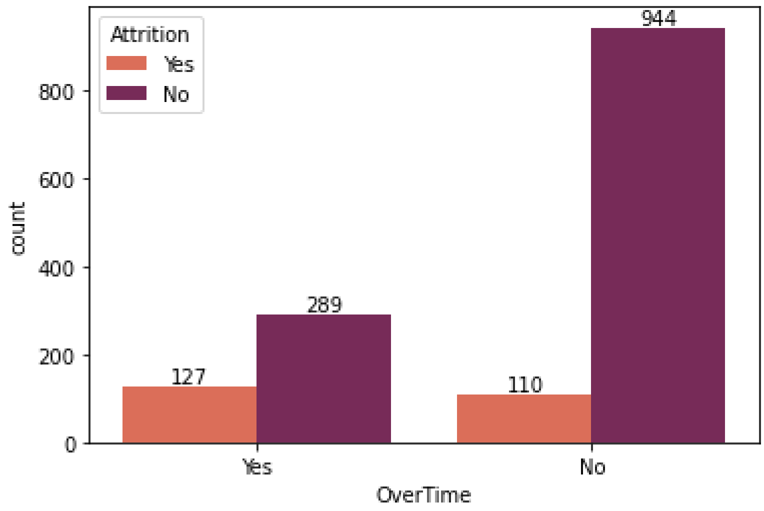
Employees who are from Life Science and medical background are highest as comparision to others.
89 Life Science background employees are leaving company followed by 63 Medical Employees.
Attrition rate of Human Resource Departement is low.

4)Overtime and Attrition

In [10]:
```
ax=sns.countplot(data=df,x="OverTime",hue="Attrition",palette='rocket_r')

for container in ax.containers:
        ax.bar_label(container, fmt="%d", label_type="edge")

plt.show()
```

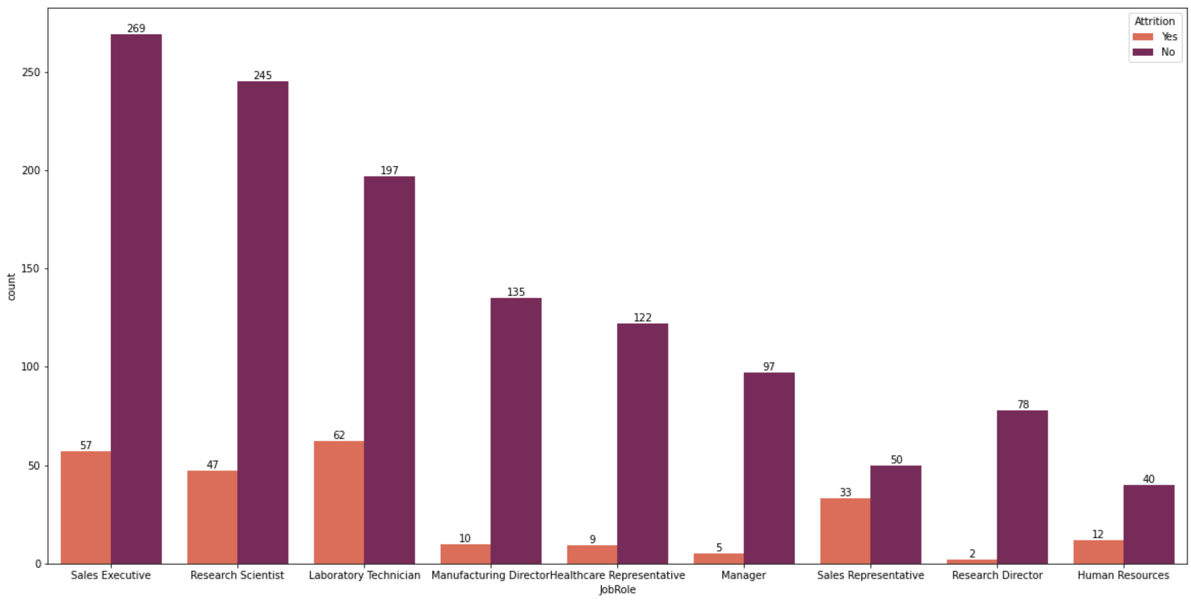


Attrition count of a employee who is doing overtime and who is not doing overtime is somehow similar.
From the graph we can say that Overtime is not much effecting the Attrition at all.

5)Impact of job role on Attrition

In [11]:
```
plt.figure(figsize=(20,10),facecolor='white')
ax=sns.countplot(data=df,x="JobRole",hue="Attrition",palette='rocket_r')

for container in ax.containers:
        ax.bar_label(container, fmt="%d", label_type="edge")
plt.show()
```

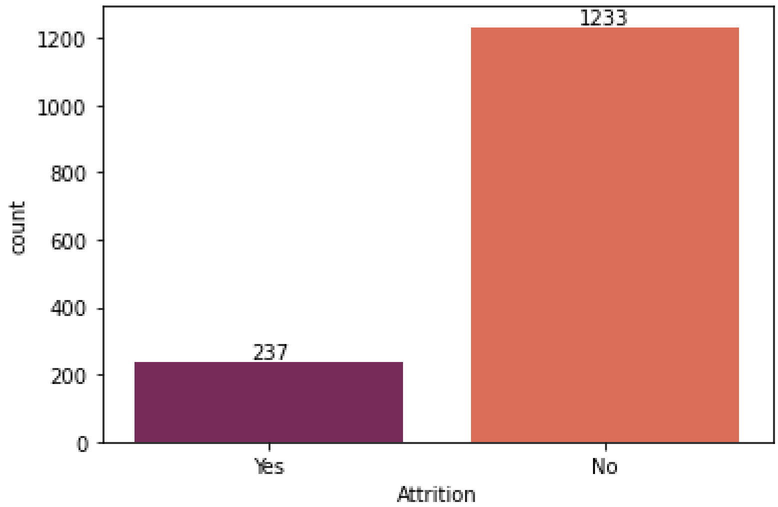


Laboratory Technician,Sales Executive and Research Scientist are the top three jobroles in which employees are leaving and also the number employees are higher in these roles.
62 Laboratory Technician are more willing to leave which is higher than any other jobroles
Around 23% of laboratory technician are leaving which higher than the sales (17%)and research scientist(16%)

6)Total attrition among the employees

In [5]:
```
ax=sns.countplot(data=df,x="Attrition",palette='rocket')

for container in ax.containers:
        ax.bar_label(container, fmt="%d", label_type="edge")
plt.show()
```

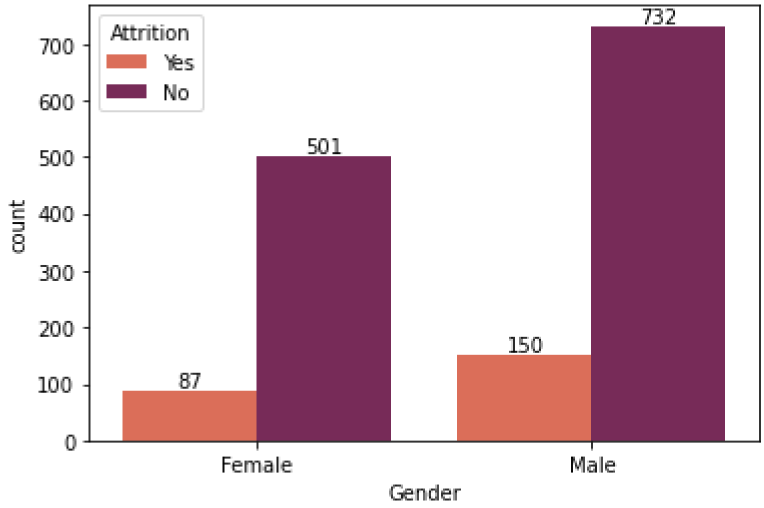


There are 1470 employees in the company.
Aprox 16.12 % (237) of employees are affected by attrition.

7)Impact of Gender on Attrition

In [13]:
```
ax=sns.countplot(data=df,x="Gender",hue="Attrition",palette='rocket_r')

for container in ax.containers:
        ax.bar_label(container, fmt="%d", label_type="edge")
plt.show()
```

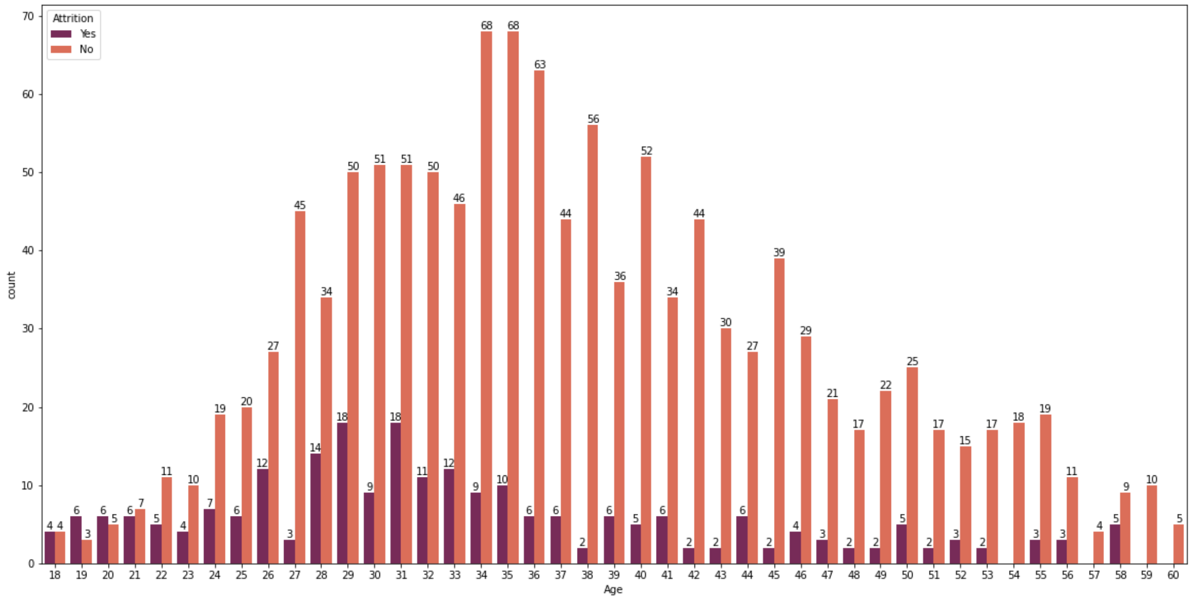


Women comprise 40% of the workforce and men comprise 60%.
Around 17% of male are leaving which is higher than the 15% of female.

8)Impact of Age on Attrition

In [38]:
```
plt.figure(figsize=(20,10),facecolor='white')
ax=sns.countplot(data=df,x="Age",hue="Attrition",palette='rocket')

for container in ax.containers:
        ax.bar_label(container, fmt="%d", label_type="edge")
plt.show()
```

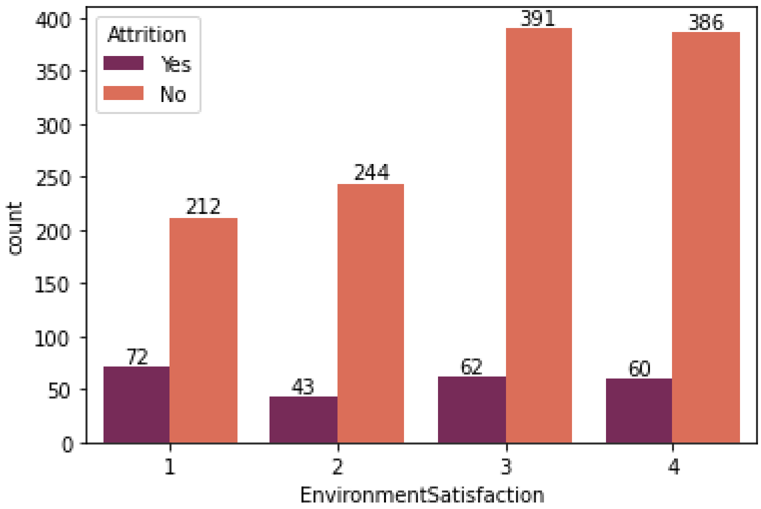


Employees aged between 25 and 37 are more likely to leave the job.
After the age of 40, there are fewer chances of leaving the company.
Employees aged 34 and 35 have the highest attrition rate, with 68 employees in this age group wanting to leave the company, more than any other age group.
Age is the most prominent factor affecting Attrition"yes".

9)Impact of Environment And Job Satisfaction

In [51]:
```
ax=sns.countplot(data=df,x="EnvironmentSatisfaction",hue="Attrition",palette='rocke

for container in ax.containers:
        ax.bar_label(container, fmt="%d", label_type="edge")
plt.show()
```



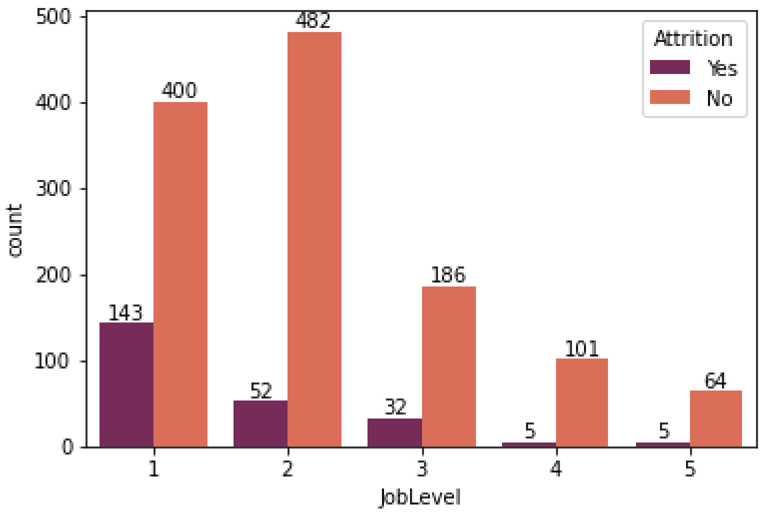
Employee who are satisfied by there job and environment are more likely to stay.
72 employees are more willing to leave their jobs because of poor job satisfaction. This is a significant number, and it shows that job satisfaction is a major factor in employee retention.

10)Job level Impacting Attrition

In [55]:
```
ax=sns.countplot(data=df,x="JobLevel",hue="Attrition", palette='rocket')

for container in ax.containers:
        ax.bar_label(container, fmt="%d", label_type="edge")

plt.show()
```



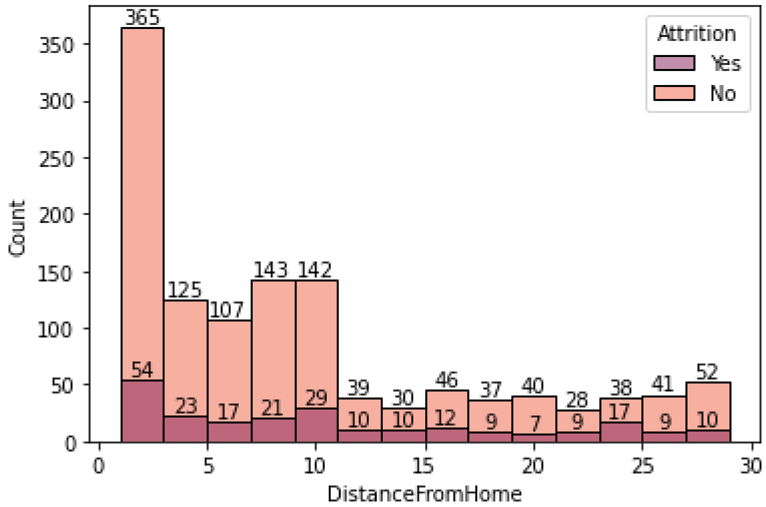
26.33% (143)entry-level employees are more likely to leave than senior-level employees.
This is because employees at higher levels typically have more responsibility, autonomy, compensation, and opportunities for career advancement.

ANALYSIS WITH THE HELP OF HISTOGRAPH

1)Distance from home and Attrition

In [39]:
```
ax=sns.histplot(data=df,x="DistanceFromHome",hue="Attrition",palette='rocket')

for container in ax.containers:
        ax.bar_label(container, fmt="%d", label_type="edge")
plt.show()
```



Employee who has distance range of 0-10 miles are more likely to leave the job.Attrition is not depend on just one factor.

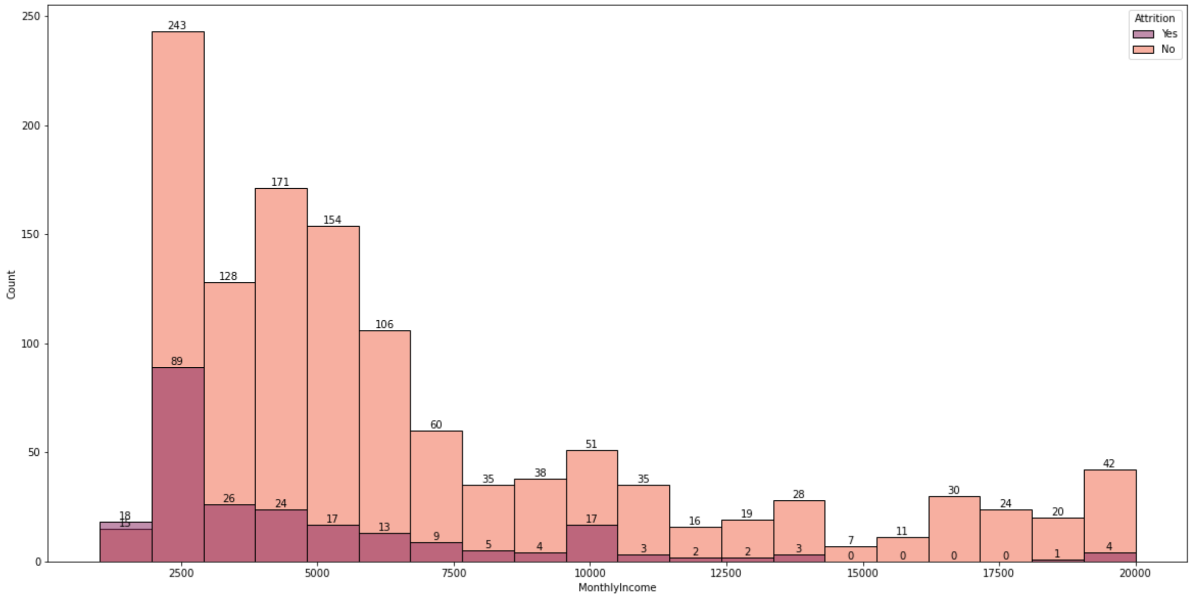
We can also conclude that the lesser the distance more employees are working.

There are 79 employee who has distance range of 0-5 miles are willing leave the company.

2)Monthly Income Trend with respect to Attrition

In [42]:
```
plt.figure(figsize=(20,10),facecolor='white')
ax=sns.histplot(data=df,x="MonthlyIncome",hue="Attrition",palette='rocket')

for container in ax.containers:
        ax.bar_label(container, fmt="%d", label_type="edge")
plt.show()
```



By the plot we can conclude that there are 115 empolyee who has monthly income between 2400 to 2600 doller are more willing to leave

Yes, we can conclude that attrition is indirectly proportional to monthly income. This means that as monthly income increases, attrition"yes" decreases. This is because employees with higher incomes are more likely to be satisfied with their jobs and less likely to leave for other opportunities.

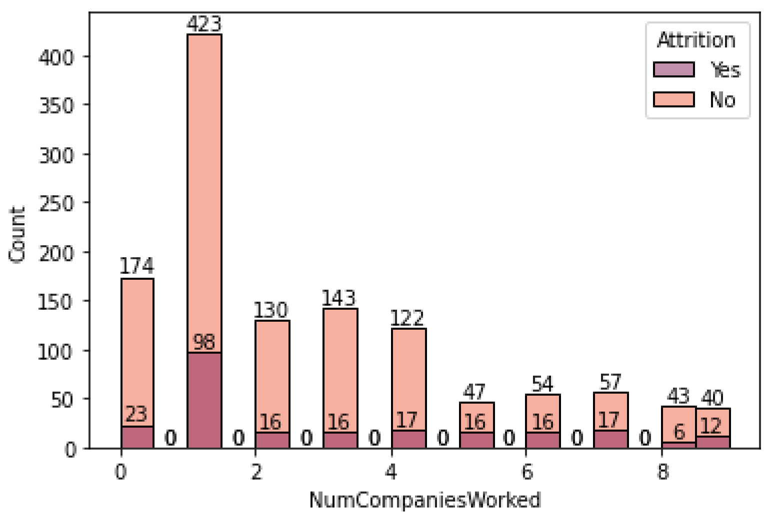
The fact that employees having a monthly income between 15,000 and 18,000 doller have zero attrition"yes" supports this conclusion. It suggests that this income range is a threshold where employees are most likely to be satisfied with their jobs and stay with the company.

3)Impact of Number of company worked

In [44]:

```
ax=sns.histplot(data=df,x="NumCompaniesWorked",hue="Attrition",palette='rocket')

for container in ax.containers:
        ax.bar_label(container, fmt="%d", label_type="edge")
plt.show()
```

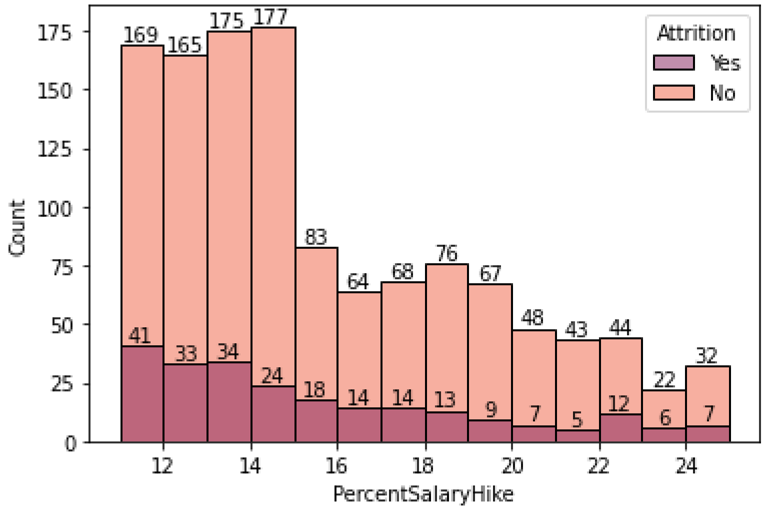


Employee who previously worked in only 1 company are more willing to leave the company.
There are 98 employees who worked in only one company previously are more willing to leave.
Employees who worked in multiple company are more willing to stay.

4)Impact of Salary hike on Attrition

In [43]:
```
ax=sns.histplot(data=df,x="PercentSalaryHike",hue="Attrition",palette='rocket')

for container in ax.containers:
        ax.bar_label(container, fmt="%d", label_type="edge")
plt.show()
```

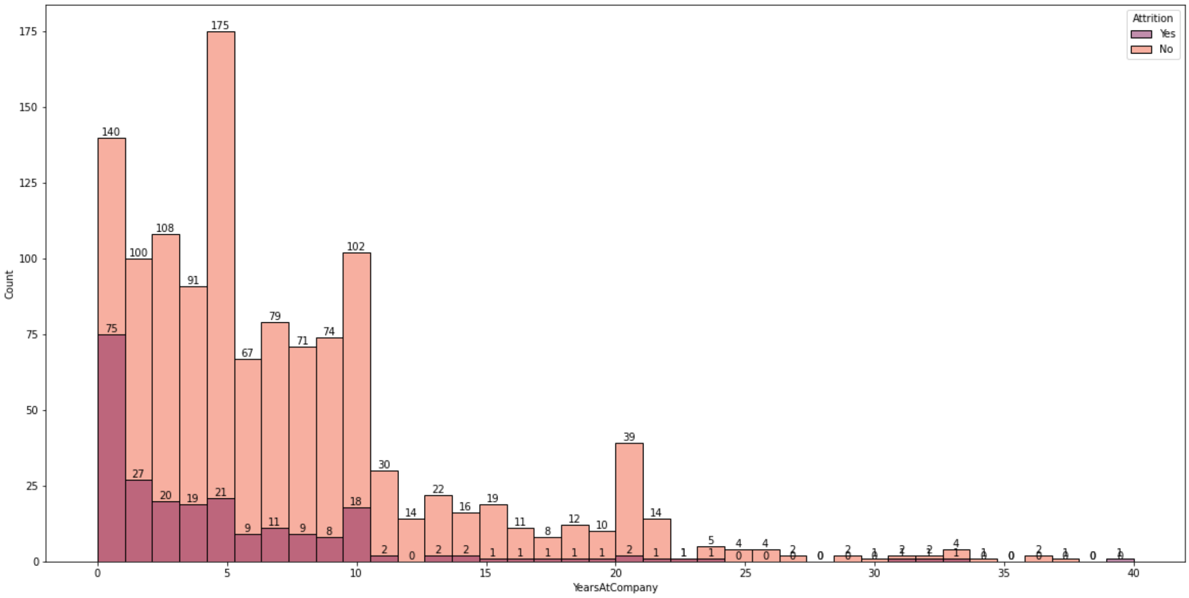


The graph suggests that employees are more likely to leave if they receive a salary hike of between 11% and 14%.
Empolyees are less likely to leave if they receive a salary hike of between 20% and 25%.
Percentage salary hike is inversely proportional to Attrition "yes"
108 employees having percentage hike in between 10 to 14 % are willing to leave
Employee having percentage hike 19 to 22% are less willing to leave.

5)Impact of Years at company

In [47]:
```
plt.figure(figsize=(20,10),facecolor='white')
ax=sns.histplot(data=df,x="YearsAtCompany",hue="Attrition",palette='rocket')

for container in ax.containers:
        ax.bar_label(container, fmt="%d", label_type="edge")
plt.show()
```



75 Fresher are leaving the company.
Employee having experience of 1 to 10 years are more likely to leave than employees of 25 year of experience.

CORELATION MATRIX IN HEATMAP

In [57]:
```
corr=df.corr()
corr
```

Out[57]:

	Age	DailyRate	DistanceFromHome	Education	EmployeeCount	Em
Age	1.000000	0.010661	-0.001686	0.208034	NaN	
DailyRate	0.010661	1.000000	-0.004985	-0.016806	NaN	
DistanceFromHome	-0.001686	-0.004985	1.000000	0.021042	NaN	
Education	0.208034	-0.016806	0.021042	1.000000	NaN	
EmployeeCount	NaN	NaN	NaN	NaN	NaN	
EmployeeNumber	-0.010145	-0.050990	0.032916	0.042070	NaN	
EnvironmentSatisfaction	0.010146	0.018355	-0.016075	-0.027128	NaN	
HourlyRate	0.024287	0.023381	0.031131	0.016775	NaN	
JobInvolvement	0.029820	0.046135	0.008783	0.042438	NaN	
JobLevel	0.509604	0.002966	0.005303	0.101589	NaN	
JobSatisfaction	-0.004892	0.030571	-0.003669	-0.011296	NaN	
MonthlyIncome	0.497855	0.007707	-0.017014	0.094961	NaN	
MonthlyRate	0.028051	-0.032182	0.027473	-0.026084	NaN	
NumCompaniesWorked	0.299635	0.038153	-0.029251	0.126317	NaN	
PercentSalaryHike	0.003634	0.022704	0.040235	-0.011111	NaN	
PerformanceRating	0.001904	0.000473	0.027110	-0.024539	NaN	
RelationshipSatisfaction	0.053535	0.007846	0.006557	-0.009118	NaN	
StandardHours	NaN	NaN	NaN	NaN	NaN	
StockOptionLevel	0.037510	0.042143	0.044872	0.018422	NaN	
TotalWorkingYears	0.680381	0.014515	0.004628	0.148280	NaN	
TrainingTimesLastYear	-0.019621	0.002453	-0.036942	-0.025100	NaN	
WorkLifeBalance	-0.021490	-0.037848	-0.026556	0.009819	NaN	
YearsAtCompany	0.311309	-0.034055	0.009508	0.069114	NaN	
YearsInCurrentRole	0.212901	0.009932	0.018845	0.060236	NaN	
YearsSinceLastPromotion	0.216513	-0.033229	0.010029	0.054254	NaN	
YearsWithCurrManager	0.202089	-0.026363	0.014406	0.069065	NaN	

26 rows × 26 columns

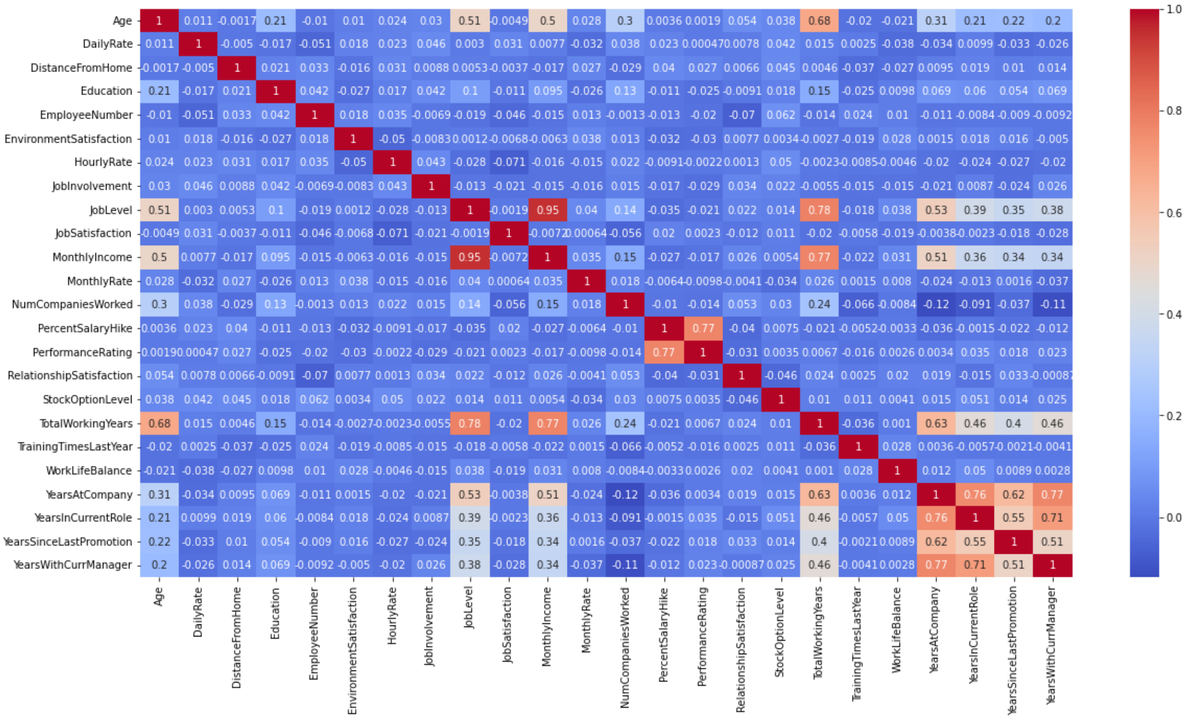
In [60]:

```
df.drop(['EmployeeCount','StandardHours'],axis=1,inplace=True)
```

In [61]:

```
corr=df.corr()
plt.figure(figsize=(20,10))
sns.heatmap(corr,annot=True,cmap='coolwarm')
```

Out[61]: <AxesSubplot:>



A correlation coefficient of 0.95 between job level and monthly income indicates a very strong positive correlation. This means that there is a very strong relationship between the two variables, such that as job level increases, monthly income also tends to increase.

In other words, employees with higher job levels tend to have higher monthly incomes. This is because higher job levels typically require more skills and experience, which are rewarded with higher compensation.

A correlation coefficient of 0.77 between total working year and monthly income indicates a strong positive correlation. This means that there is a strong relationship between the two variables, such that as total working year increases, monthly income also tends to increase.

In other words, employees with more total working years tend to have higher monthly incomes. This is because employees with more total working years typically have more skills and experience, which are rewarded with higher compensation.