

```
In [4]: import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

```
In [9]: df=pd.read_csv("C:/Users/DELL/Downloads/HR-Employee-Attrition.csv")
df
```

Out[9]:

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

1470 rows × 35 columns

```
In [10]: df.head()
```

Out[10]:

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeNumber
0	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	1	1
1	49	No	Travel_Frequently	279	Research & Development	8	1	Life Sciences	1	2
2	37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other	1	4
3	33	No	Travel_Frequently	1392	Research & Development	3	4	Life Sciences	1	5
4	27	No	Travel_Rarely	591	Research & Development	2	1	Medical	1	7

5 rows × 35 columns

```
In [11]: df.tail()
```

Out[11]:

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

5 rows × 35 columns

```
In [12]: df.describe() #for statistical
```

Out[12]:

	Age	DailyRate	DistanceFromHome	Education	EmployeeCount	EmployeeNumber	EnvironmentSatisfaction	HourlyRate
count	1470.000000	1470.000000	1470.000000	1470.000000	1470.0	1470.000000	1470.000000	1470.000000
mean	36.923810	802.485714	9.192517	2.912925	1.0	1024.865306	2.721769	65.891156
std	9.135373	403.509100	8.106864	1.024165	0.0	602.024335	1.093082	20.329428
min	18.000000	102.000000	1.000000	1.000000	1.0	1.000000	1.000000	30.000000
25%	30.000000	465.000000	2.000000	2.000000	1.0	491.250000	2.000000	48.000000
50%	36.000000	802.000000	7.000000	3.000000	1.0	1020.500000	3.000000	66.000000
75%	43.000000	1157.000000	14.000000	4.000000	1.0	1555.750000	4.000000	83.750000
max	60.000000	1499.000000	29.000000	5.000000	1.0	2068.000000	4.000000	100.000000

8 rows × 26 columns

In [18]:

df.describe(include="O")

Out[18]:

	Attrition	BusinessTravel	Department	EducationField	Gender	JobRole	MaritalStatus	Over18	OverTime
count	1470	1470	1470	1470	1470	1470	1470	1470	1470
unique	2	3	3	6	2	9	3	1	2
top	No	Travel_Rarely	Research & Development	Life Sciences	Male	Sales Executive	Married	Y	No
freq	1233	1043	961	606	882	326	673	1470	1054

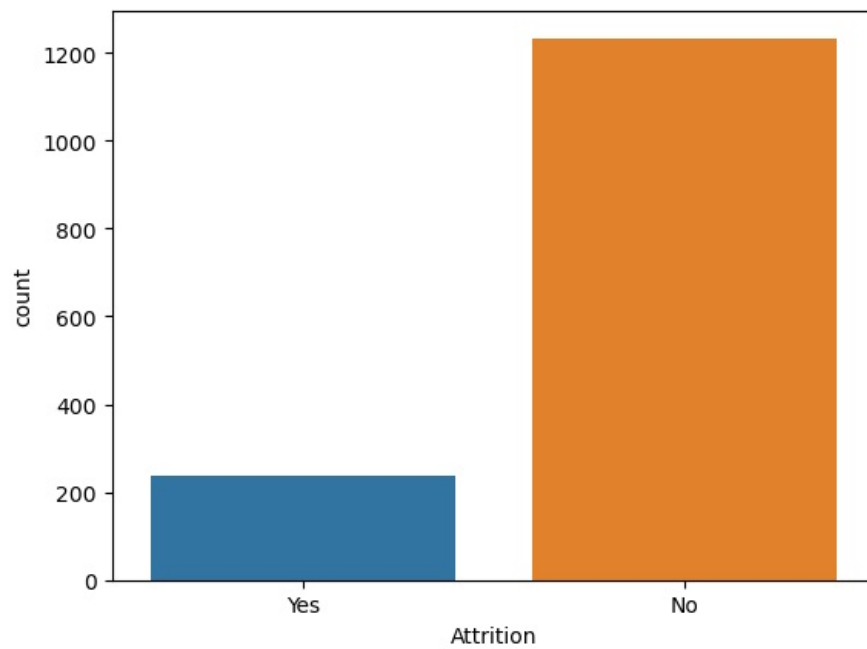
In [16]:

df.info()

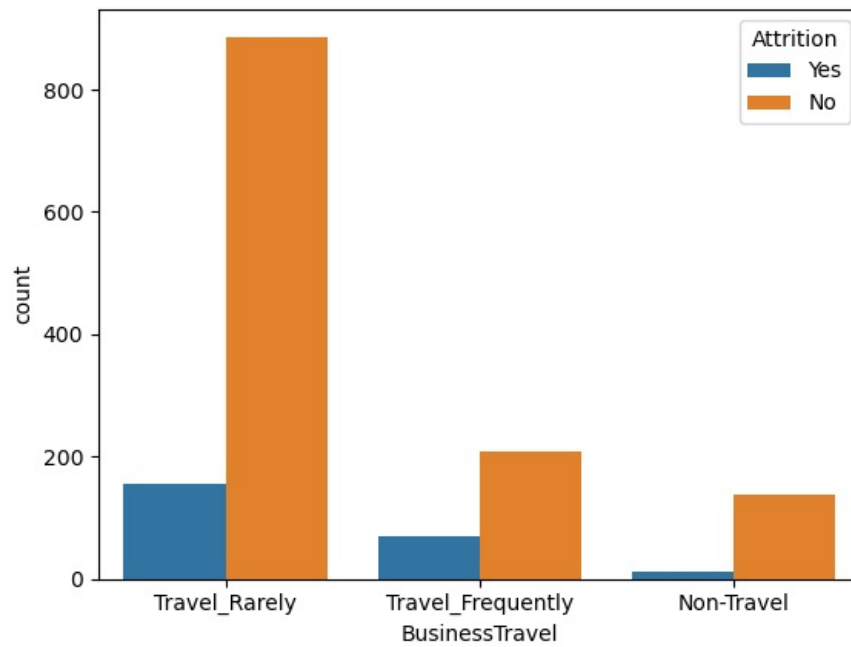
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1470 entries, 0 to 1469
Data columns (total 35 columns):
Column Non-Null Count Dtype
--- -
0 Age 1470 non-null int64
1 Attrition 1470 non-null object
2 BusinessTravel 1470 non-null object
3 DailyRate 1470 non-null int64
4 Department 1470 non-null object
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 [19]:

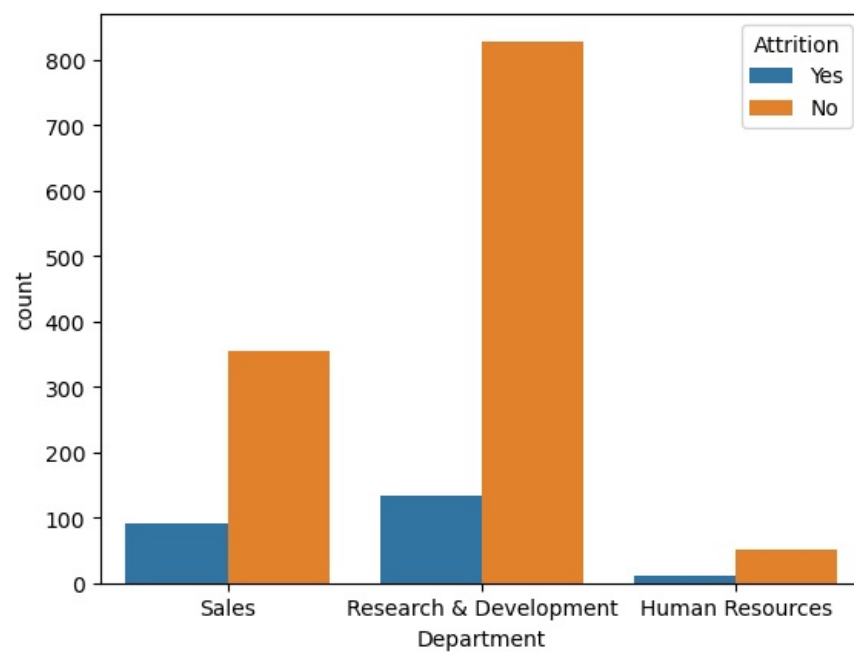
sns.countplot(x=df.Attrition)
plt.show()



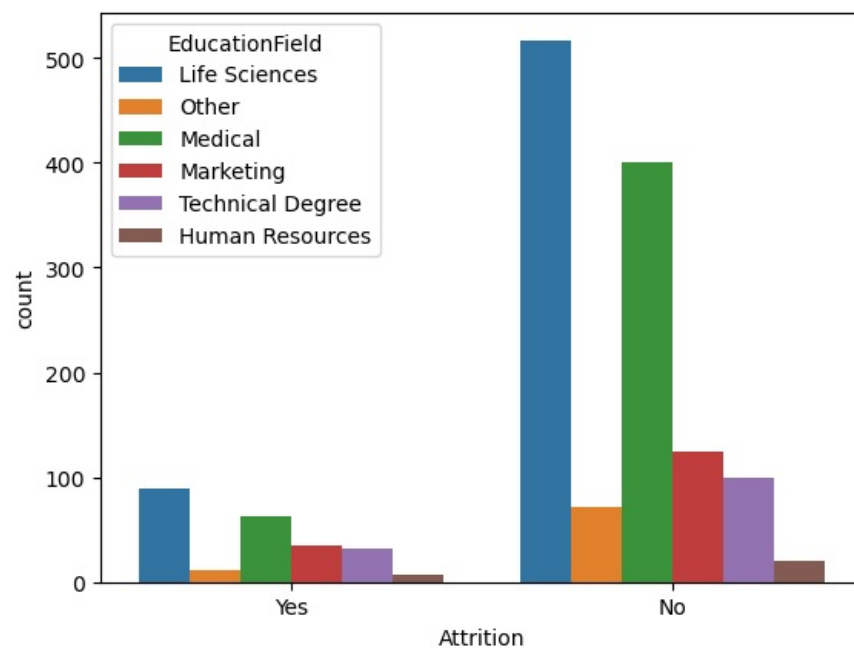
```
In [20]: #Impact Of Business travel On Attrition
sns.countplot(hue=df.Attrition,x=df.BusinessTravel)
plt.show()
```



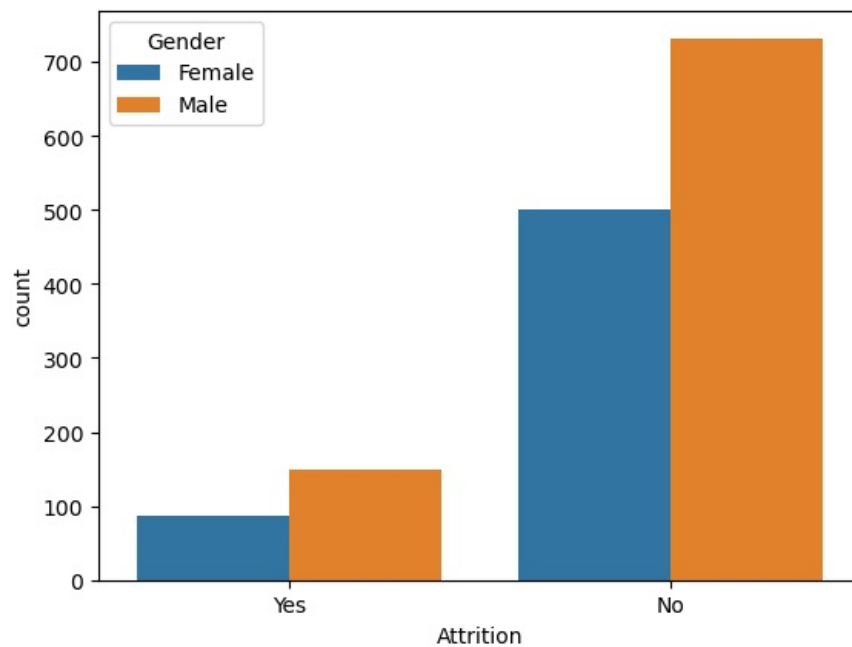
```
In [21]: #impact on department on Attrition
sns.countplot(hue=df.Attrition,x=df.Department)
plt.show()
```



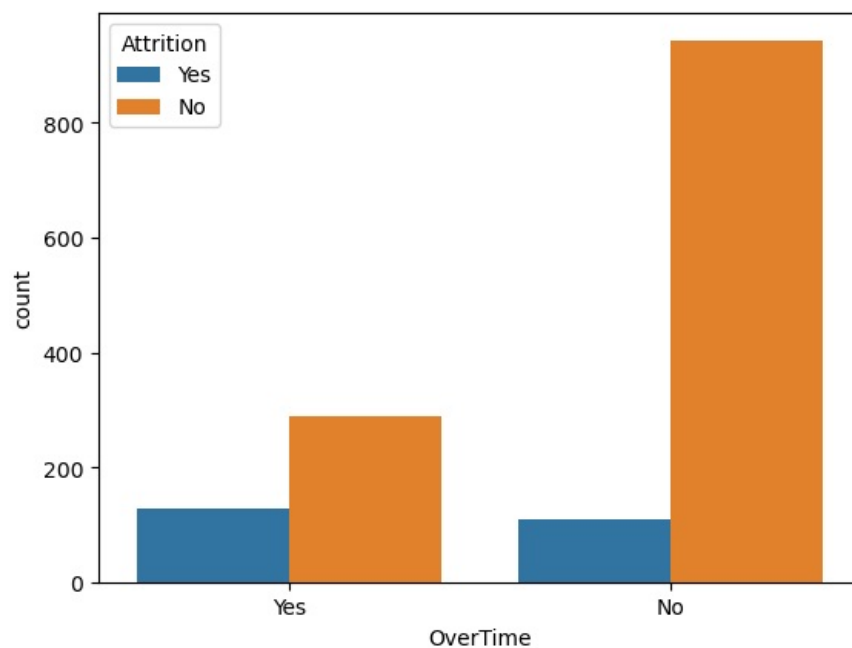
```
In [22]: #Impact of Education field on Attrition
sns.countplot(x=df.Attrition,hue=df.EducationField)
plt.show()
```



```
In [23]: #Gender And Attrition
sns.countplot(x=df.Attrition,hue=df.Gender)
plt.show()
```



```
In [24]: #Overtime And Attirtion
sns.countplot(hue=df.Attrition,x=df.OverTime)
plt.show()
```



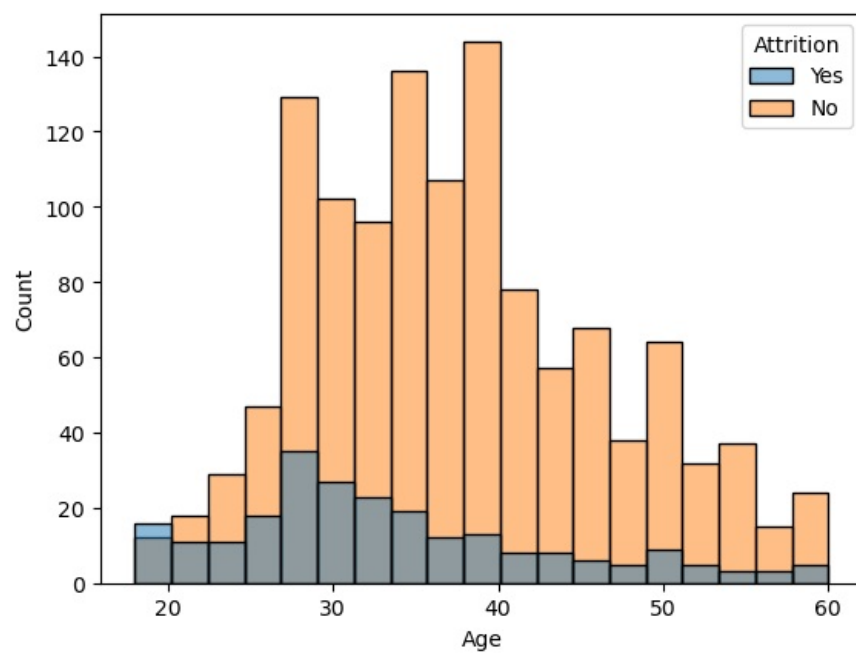
```
In [25]: #Analysis on Continuos data

numerical_col=[]
for column in df.columns:
    if df[column].dtype=="int64" and len(df[column].unique())>=10:
        numerical_col.append(column)
```

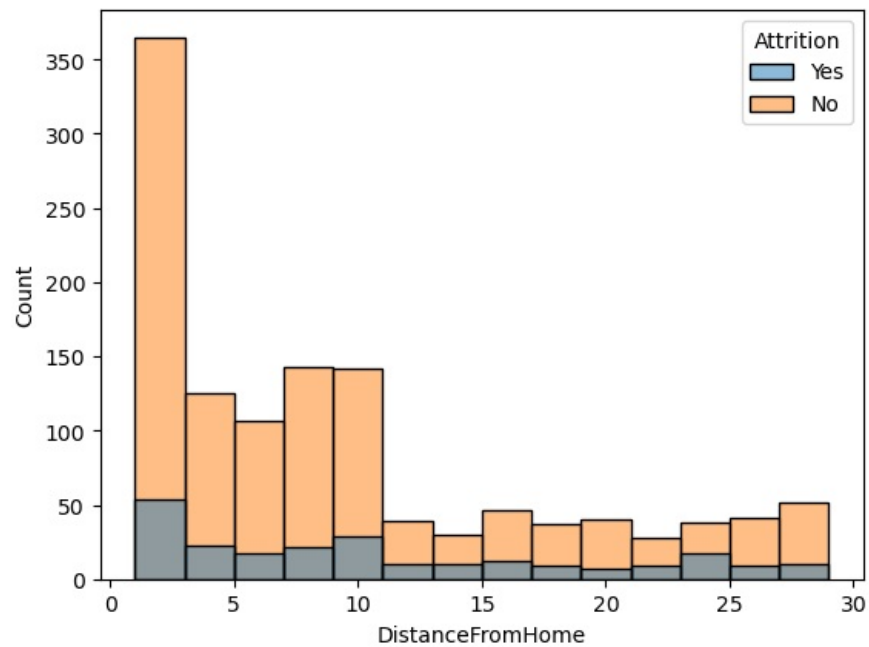
```
In [26]: numerical_col
```

```
Out[26]: ['Age',
'DailyRate',
'DistanceFromHome',
'EmployeeNumber',
'HourlyRate',
'MonthlyIncome',
'MonthlyRate',
'NumCompaniesWorked',
'PercentSalaryHike',
'TotalWorkingYears',
'YearsAtCompany',
'YearsInCurrentRole',
'YearsSinceLastPromotion',
'YearsWithCurrManager']
```

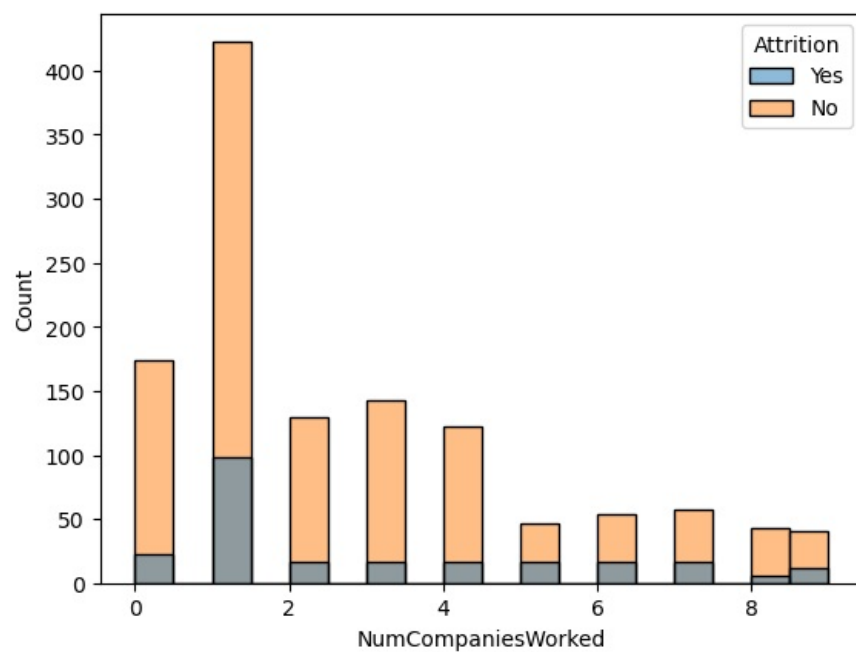
```
In [31]: #Impact on Age on Attrition
sns.histplot(hue=df.Attrition,x=df.Age)
plt.show()
```



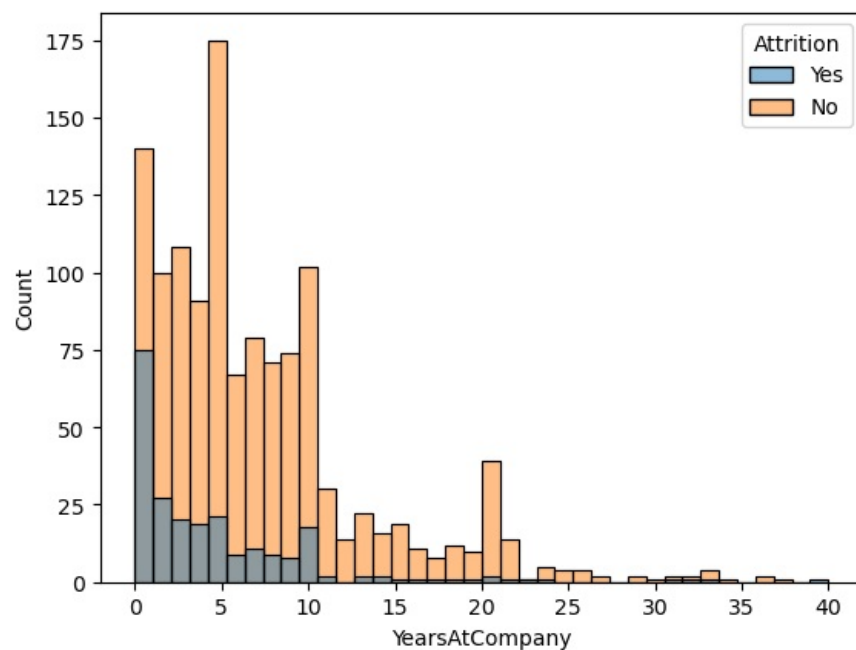
```
In [32]: #Distance from Home And Attrition
sns.histplot(hue=df.Attrition,x=df.DistanceFromHome)
plt.show()
```



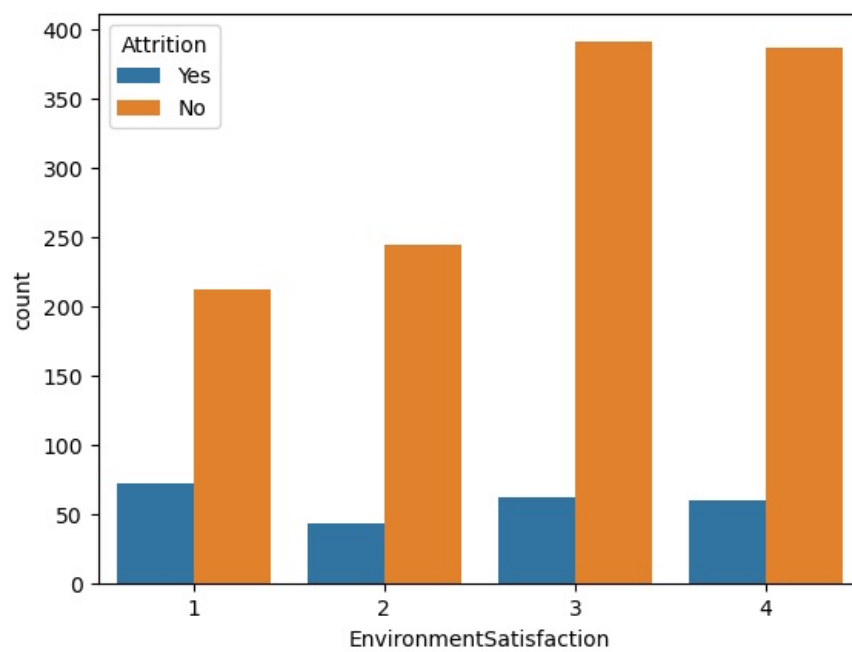
```
In [33]: #Impact of No. of Companies Worked
sns.histplot(hue=df.Attrition,x=df.NumCompaniesWorked)
plt.show()
```



```
In [35]: #years at the Company
sns.histplot(x=df.YearsAtCompany, hue=df.Attrition)
plt.show()
```



```
In [36]: #Impact on Environment And Job Satisfaction on Attrition
sns.countplot(hue=df.Attrition, x=df.EnvironmentSatisfaction)
plt.show()
```



```
In [37]: df.isnull().sum()
```

```
Out[37]: 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
```

```
In [39]: #Dropping unique ones
df.drop(['EmployeeCount', 'EmployeeNumber', 'Over18', 'StandardHours'], axis=1, inplace=True)
```

```
In [40]: df
```


Out[40]:

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

1470 rows × 31 columns

In [41]: df.Attrition.value_counts()

Out[41]:

No1233
Yes237
Name: Attrition, dtype: int64

In []:

#Insights/Impacts

1. There are more employees which travels rarely and are not satisfied with their job.
2. Research and Development department have more number of Attrition as compared to other two departments.
3. As for Attrition Yes ,there is minor difference between the Employees who are doing overtime and who are not
4. Employees in Age of 25 to 35 are more likely to leave the job.
5. After 40 age,the distibution tells us that Higher The Age Lesser will be Attrition.
6. Employees who has distance range of 0-10km, are more likely to leave the job.
7. Increase in the rate in environment as well as job satisfaction give rise to increase in iteration no.