Bharat Intern

Name: Sham johari

Task 3

Employee and attrition and performance:

In this project, you will need to evaluate each factor and its relationship with attrition, for example, the distance from home to office, the job role impact on attrition, etc

In [40]:

```
# important libraries
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

Exploratory Data Analysis

In [41]:

```
df = pd.read_csv('WA_Fn-UseC_-HR-Employee-Attrition.csv')
df.head(10)
```

Out[41]:

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	I
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	32	No	Travel_Frequently	1005	Research & Development	2	2	
6	59	No	Travel_Rarely	1324	Research & Development	3	3	
7	30	No	Travel_Rarely	1358	Research & Development	24	1	
8	38	No	Travel_Frequently	216	Research & Development	23	3	
9	36	No	Travel_Rarely	1299	Research & Development	27	3	

10 rows × 35 columns

In [42]:

```
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 [43]:

```
df.columns
```

Out[43]:

```
Index(['Age', 'Attrition', 'BusinessTravel', 'DailyRate', 'Department',
       'DistanceFromHome', 'Education', 'EducationField', 'EmployeeCoun
t',
       'EmployeeNumber', 'EnvironmentSatisfaction', 'Gender', 'HourlyRat
e',
       'JobInvolvement', 'JobLevel', 'JobRole', 'JobSatisfaction',
       'MaritalStatus', 'MonthlyIncome', 'MonthlyRate', 'NumCompaniesWork
ed',
       'Over18', 'OverTime', 'PercentSalaryHike', 'PerformanceRating',
       'RelationshipSatisfaction', 'StandardHours', 'StockOptionLevel',
       'TotalWorkingYears', 'TrainingTimesLastYear', 'WorkLifeBalance',
       'YearsAtCompany', 'YearsInCurrentRole', 'YearsSinceLastPromotion',
       'YearsWithCurrManager'],
      dtype='object')
```

In [44]:

```
df.describe().T
```

Out[44]:

	count	mean	std	min	25%	50%	
Age	1470.0	36.923810	9.135373	18.0	30.00	36.0	4
DailyRate	1470.0	802.485714	403.509100	102.0	465.00	802.0	115
DistanceFromHome	1470.0	9.192517	8.106864	1.0	2.00	7.0	1
Education	1470.0	2.912925	1.024165	1.0	2.00	3.0	
EmployeeCount	1470.0	1.000000	0.000000	1.0	1.00	1.0	
EmployeeNumber	1470.0	1024.865306	602.024335	1.0	491.25	1020.5	155
EnvironmentSatisfaction	1470.0	2.721769	1.093082	1.0	2.00	3.0	
HourlyRate	1470.0	65.891156	20.329428	30.0	48.00	66.0	8
JobInvolvement	1470.0	2.729932	0.711561	1.0	2.00	3.0	
JobLevel	1470.0	2.063946	1.106940	1.0	1.00	2.0	
JobSatisfaction	1470.0	2.728571	1.102846	1.0	2.00	3.0	
MonthlyIncome	1470.0	6502.931293	4707.956783	1009.0	2911.00	4919.0	837
MonthlyRate	1470.0	14313.103401	7117.786044	2094.0	8047.00	14235.5	2046
NumCompaniesWorked	1470.0	2.693197	2.498009	0.0	1.00	2.0	
PercentSalaryHike	1470.0	15.209524	3.659938	11.0	12.00	14.0	1
PerformanceRating	1470.0	3.153741	0.360824	3.0	3.00	3.0	
RelationshipSatisfaction	1470.0	2.712245	1.081209	1.0	2.00	3.0	
StandardHours	1470.0	80.000000	0.000000	80.0	80.00	80.0	8
StockOptionLevel	1470.0	0.793878	0.852077	0.0	0.00	1.0	
TotalWorkingYears	1470.0	11.279592	7.780782	0.0	6.00	10.0	1
TrainingTimesLastYear	1470.0	2.799320	1.289271	0.0	2.00	3.0	
WorkLifeBalance	1470.0	2.761224	0.706476	1.0	2.00	3.0	
YearsAtCompany	1470.0	7.008163	6.126525	0.0	3.00	5.0	
YearsInCurrentRole	1470.0	4.229252	3.623137	0.0	2.00	3.0	
YearsSinceLastPromotion	1470.0	2.187755	3.222430	0.0	0.00	1.0	
YearsWithCurrManager	1470.0	4.123129	3.568136	0.0	2.00	3.0	

In [45]:

```
df.shape
```

Out[45]:

```
(1470, 35)
```

In [46]:

```
df.nunique()
```

Out[46]:

```
Age                         43
Attrition                    2
BusinessTravel               3
DailyRate                  886
Department                   3
DistanceFromHome            29
Education                    5
EducationField               6
EmployeeCount                1
EmployeeNumber            1470
EnvironmentSatisfaction      4
Gender                       2
HourlyRate                  71
JobInvolvement               4
JobLevel                     5
JobRole                      9
JobSatisfaction              4
MaritalStatus                3
MonthlyIncome             1349
MonthlyRate               1427
NumCompaniesWorked          10
Over18                       1
OverTime                     2
PercentSalaryHike           15
PerformanceRating            2
RelationshipSatisfaction     4
StandardHours                1
StockOptionLevel             4
TotalWorkingYears           40
TrainingTimesLastYear        7
WorkLifeBalance              4
YearsAtCompany              37
YearsInCurrentRole          19
YearsSinceLastPromotion     16
YearsWithCurrManager        18
dtype: int64
```

In [47]:

```
df.drop(['EmployeeCount','StandardHours','Over18','EmployeeNumber'], axis=1, inplace=Tru
```

In [48]:

```
columns = list(df.columns)
categorical = [data for data in columns if df[data].dtype=='object']
categorical
```

Out[48]:

```
['Attrition',
 'BusinessTravel',
 'Department',
 'EducationField',
 'Gender',
 'JobRole',
 'MaritalStatus',
 'OverTime']
```

In [49]:

```
for data in categorical:
    print(pd.crosstab(df[data],df['Attrition'],margins=True))
    print('-------------------------------------------------')
```

Attribution	No	Yes	All
Attribution			
No	1233	0	1233
Yes	0	237	237
All	1233	237	1470

Attribution	No	Yes	All
BusinessTravel			
Non-Travel	138	12	150
Travel_Frequently	208	69	277
Travel_Rarely	887	156	1043
All	1233	237	1470

Attribution	No	Yes	All
Department			
Human Resources	51	12	63
Research & Development	828	133	961
Sales	354	92	446
All	1233	237	1470

Attribution	No	Yes	All
EducationField			
Human Resources	20	7	27
Life Sciences	517	89	606
Marketing	124	35	159
Medical	401	63	464
Other	71	11	82
Technical Degree	100	32	132
All	1233	237	1470

Attribution	No	Yes	All
Gender			
Female	501	87	588
Male	732	150	882
All	1233	237	1470

Attribution	No	Yes	All
JobRole			
Healthcare Representative	122	9	131
Human Resources	40	12	52
Laboratory Technician	197	62	259
Manager	97	5	102
Manufacturing Director	135	10	145
Research Director	78	2	80
Research Scientist	245	47	292
Sales Executive	269	57	326
Sales Representative	50	33	83
All	1233	237	1470

Attribution	No	Yes	All
MaritalStatus			
Divorced	294	33	327
Married	589	84	673
Single	350	120	470
All	1233	237	1470

Attribution	No	Yes	All
OverTime			
No	944	110	1054
Yes	289	127	416

```
All        1233  237  1470
-----------------------------------------------
```

Data Visualization

In [ ]:

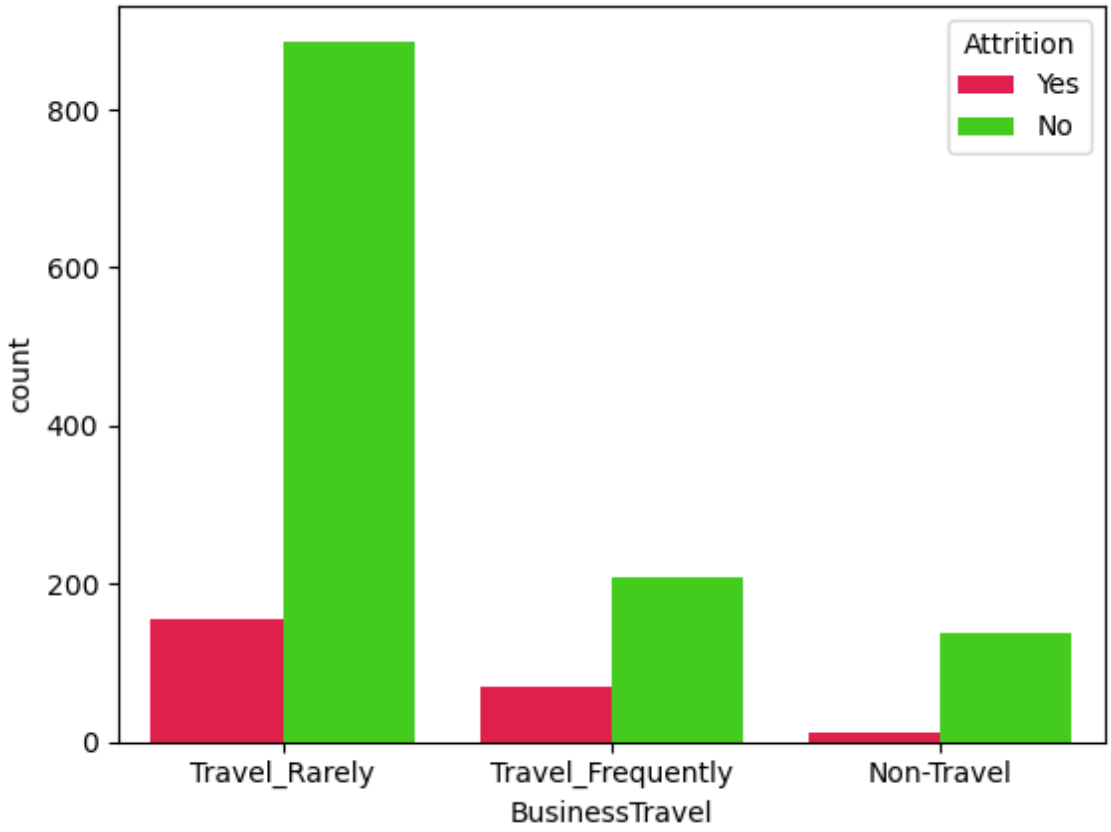
```
```

In [50]:

```
sns.countplot(x='Gender', hue='Attrition', data=df, palette='prism_r')
plt.show()
```

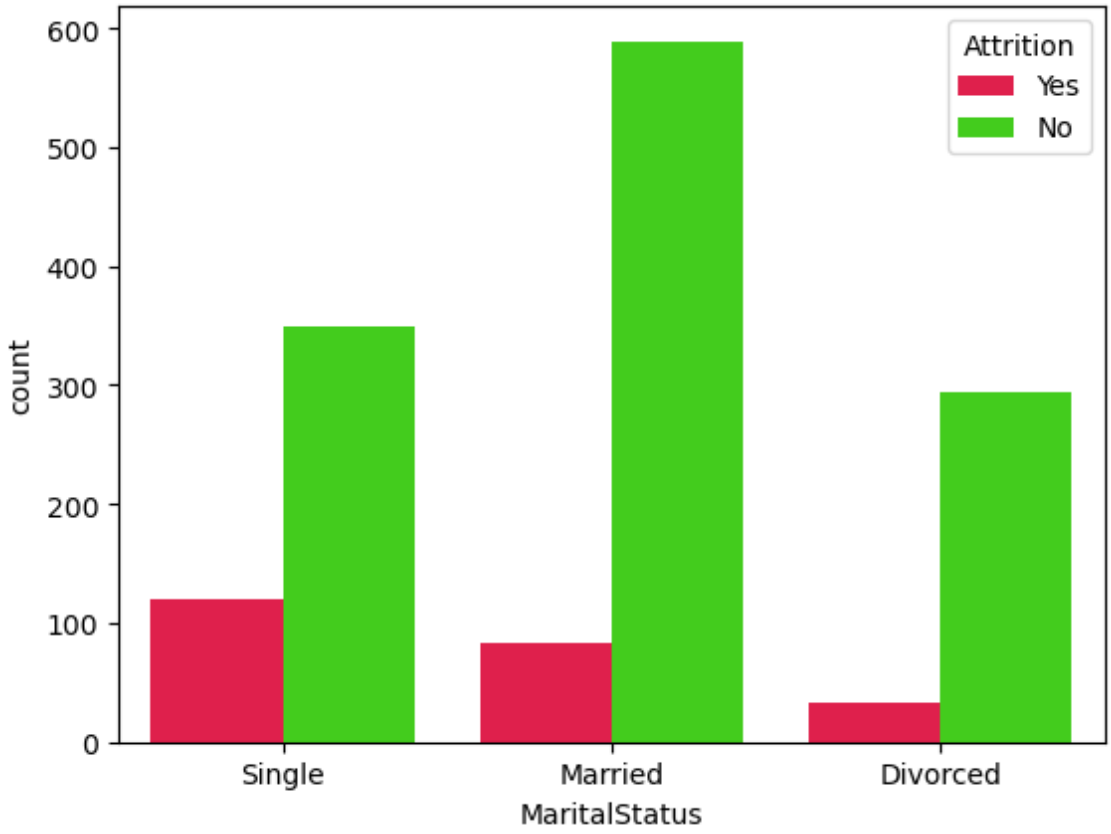
In [51]:

```
sns.countplot(x='BusinessTravel', hue='Attrition', data=df, palette='prism_r')
plt.show()
```



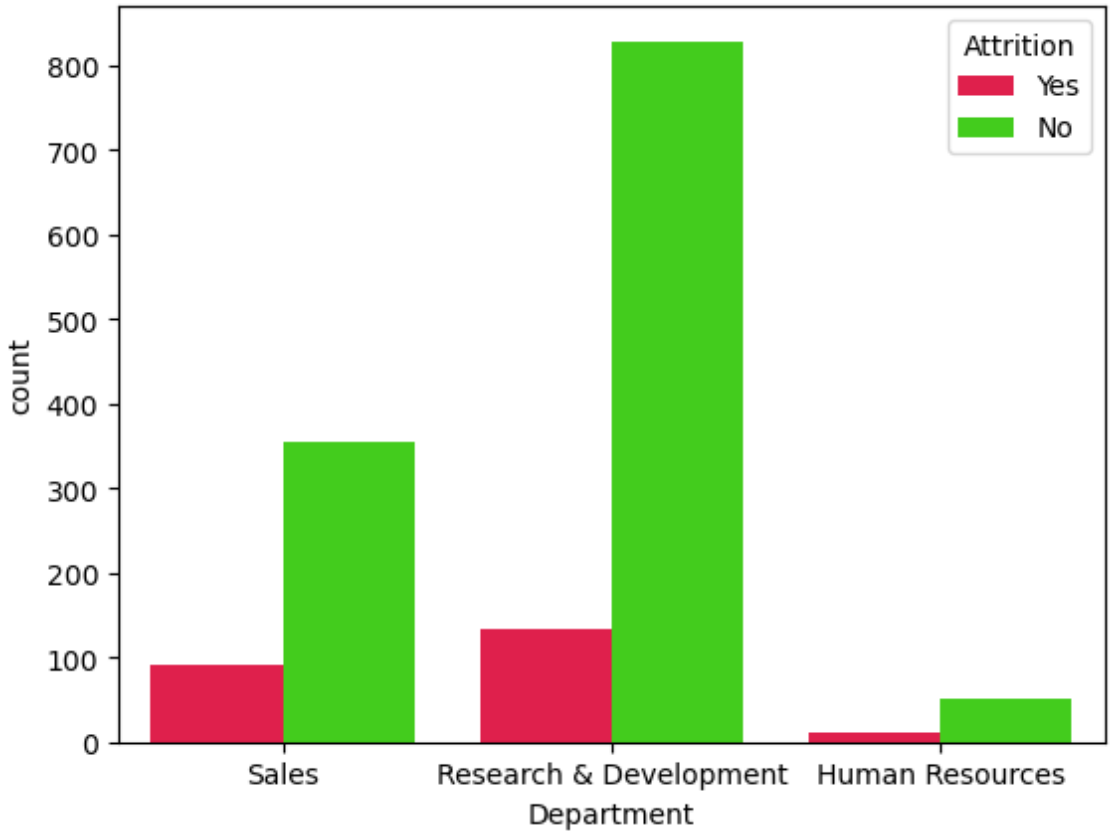
In [52]:

```
sns.countplot(x='MaritalStatus', hue='Attrition', data=df, palette='prism_r')
plt.show()
```



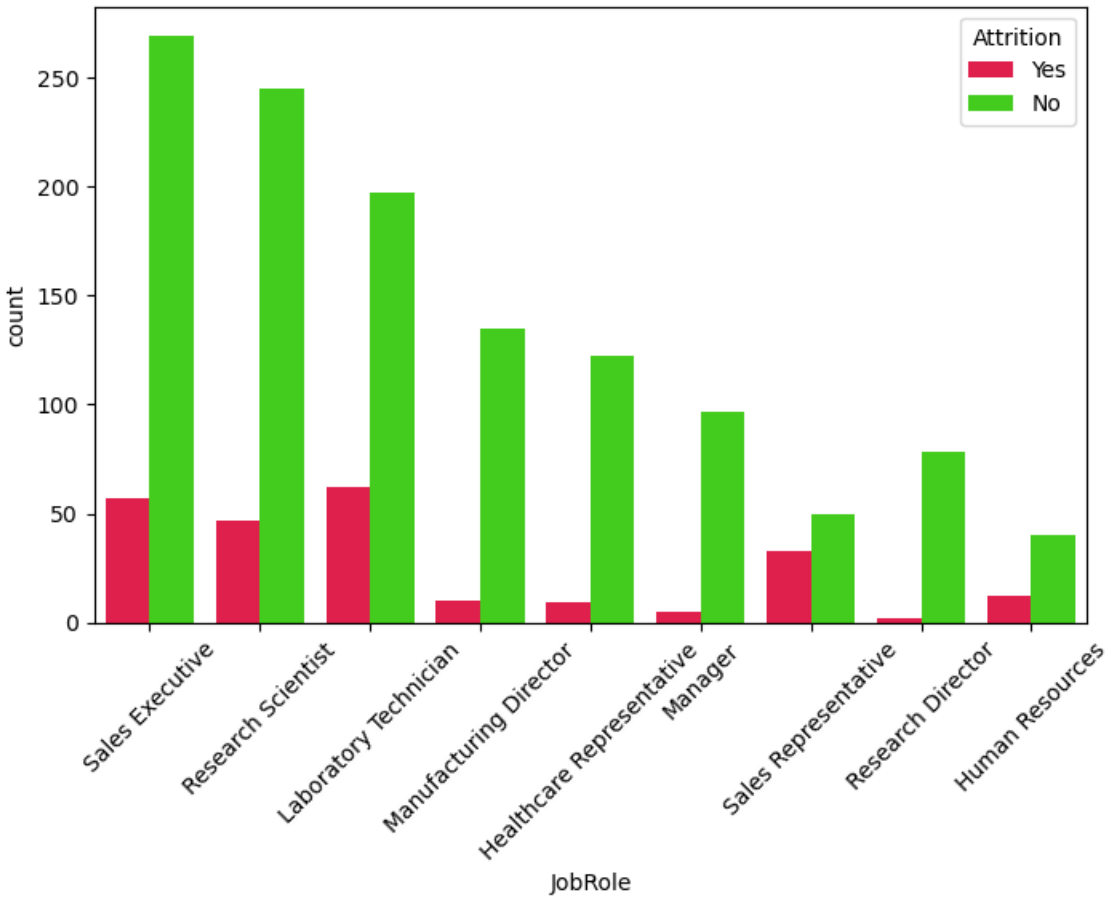
In [53]:

```
sns.countplot(x='Department', hue='Attrition', data=df, palette='prism_r')
plt.show()
```



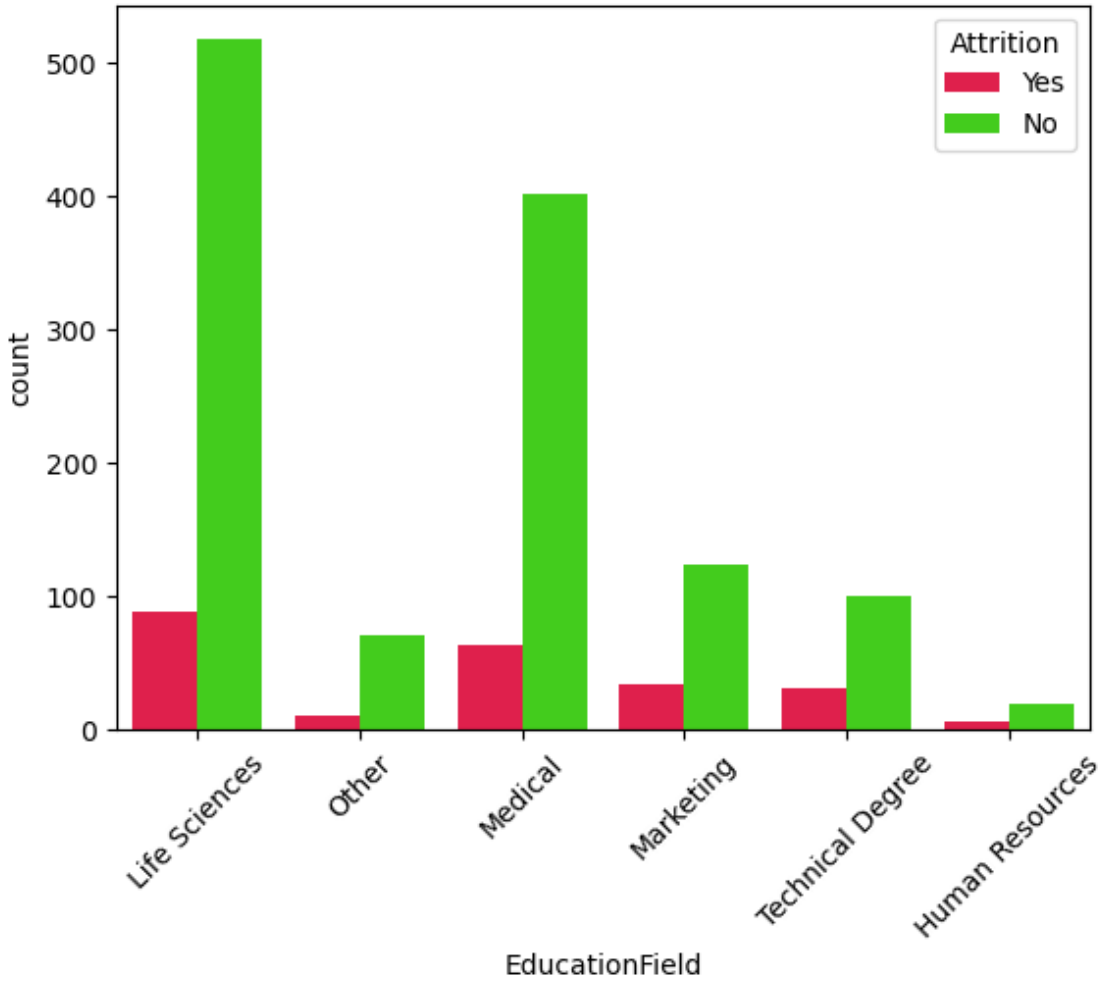
In [54]:

```
plt.figure(figsize=(8,5))
sns.countplot(x='JobRole', hue='Attrition', data=df, palette='prism_r')
plt.xticks(rotation=45)
plt.show()
```



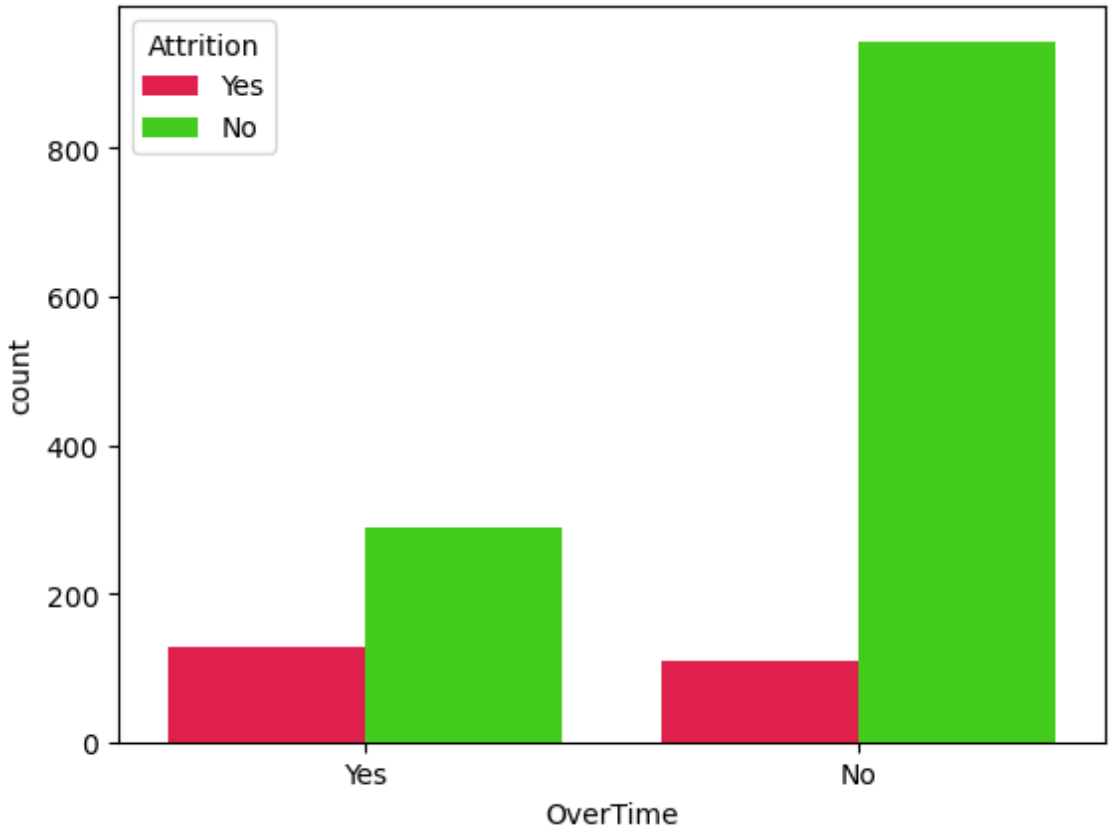
In [55]:

```
sns.countplot(x='EducationField', hue='Attrition', data=df, palette='prism_r')
plt.xticks(rotation=45)
plt.show()
```



In [56]:

```
sns.countplot(x='OverTime', hue='Attrition', data=df, palette='prism_r')
plt.show()
```

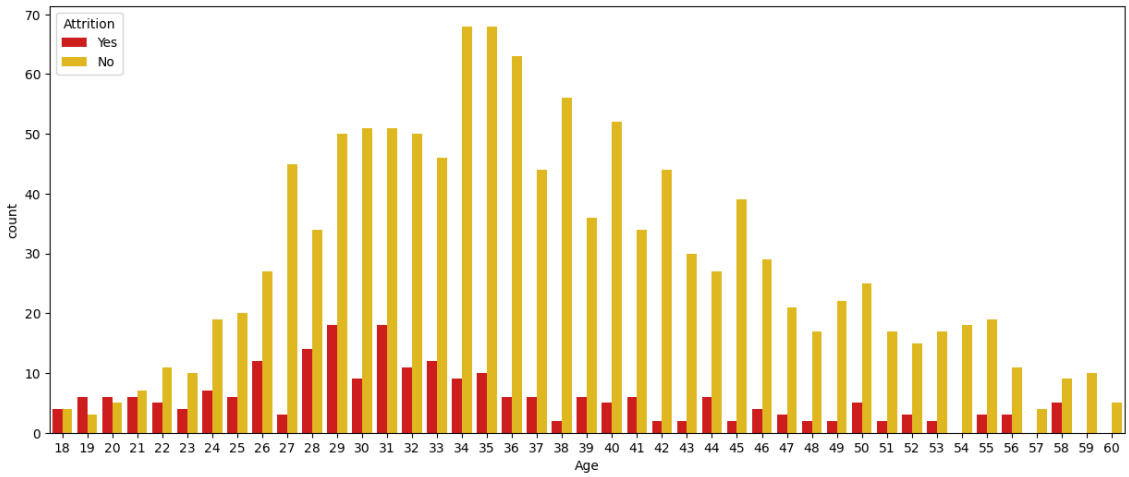


Some Observations:

1. Gender: Male employees quit more than female employees.
2. Business Travel: The employees who travel rarely are more likely to quit than other employees.
3. Marital Status: Employees who are single tend to quit their jobs more than the married or divorced.
4. Department: Research and Development employees don't quit their jobs as much as the other departments.
5. Job Role: Sales Executives, Laboratory Technicians and Research Scientists are more likely to quit than other employees.
6. Education Field: Employees from Life Sciences, Medical and Marketing educational background are more likely to stay than other employees of different educational background.
7. Over Time: Employees who do over time, quit more.

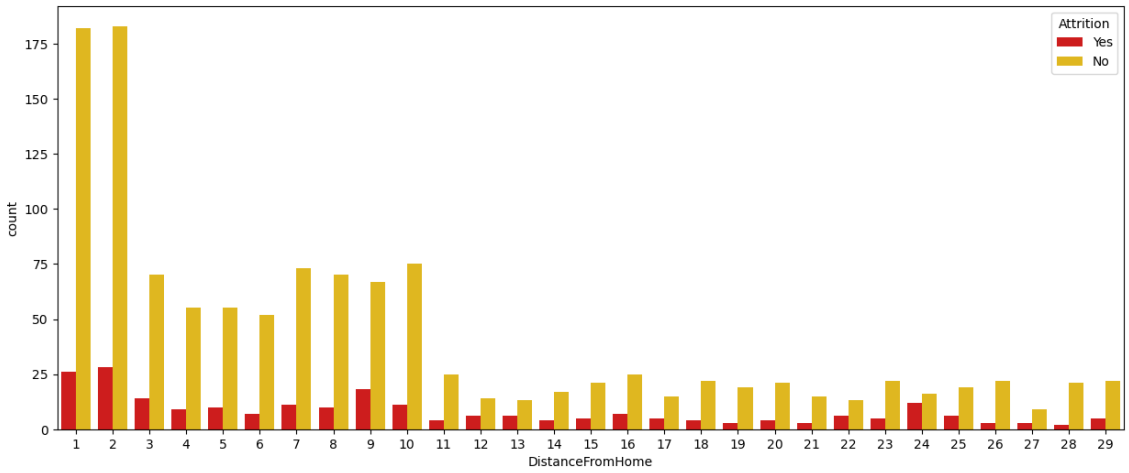
In [57]:

```
plt.figure(figsize=(15,6))
sns.countplot(x='Age', hue='Attrition', data=df, palette='hot')
plt.show()
```



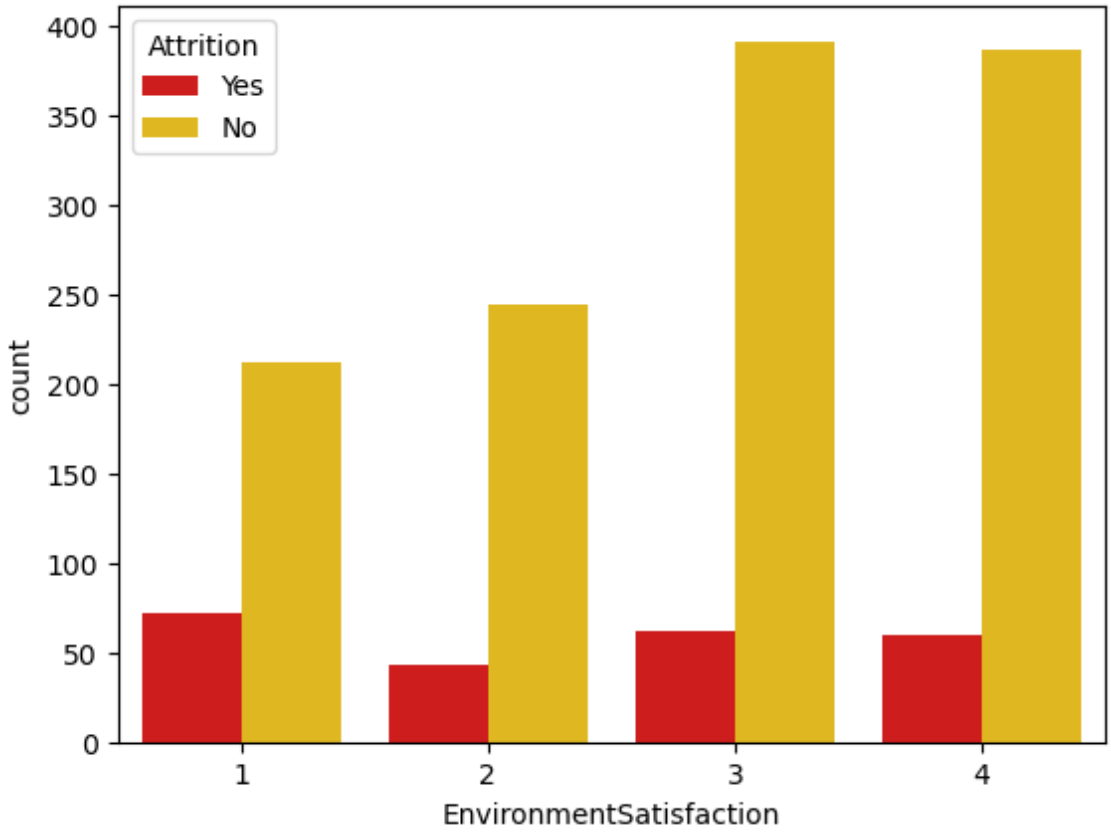
In [58]:

```
plt.figure(figsize=(15,6))
sns.countplot(x='DistanceFromHome', hue='Attrition', data=df, palette='hot')
plt.show()
```



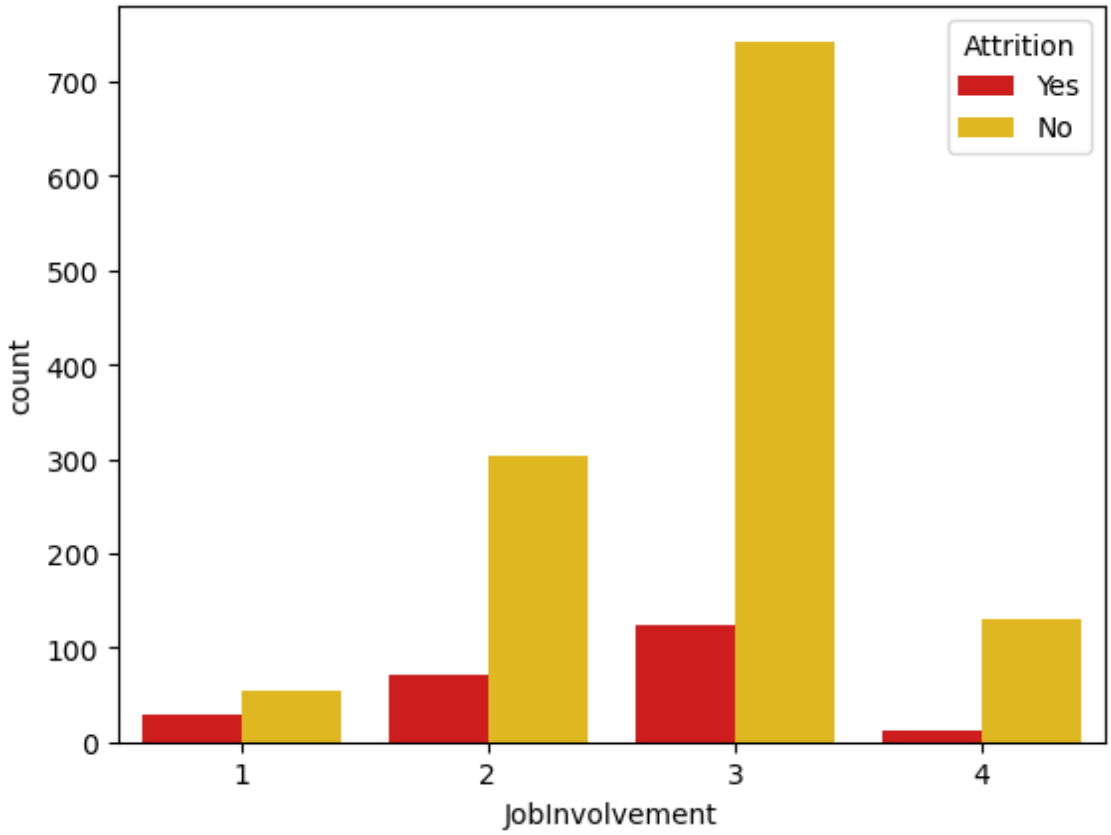
In [59]:

```
sns.countplot(x='EnvironmentSatisfaction', hue='Attrition', data=df, palette='hot')
plt.show()
```



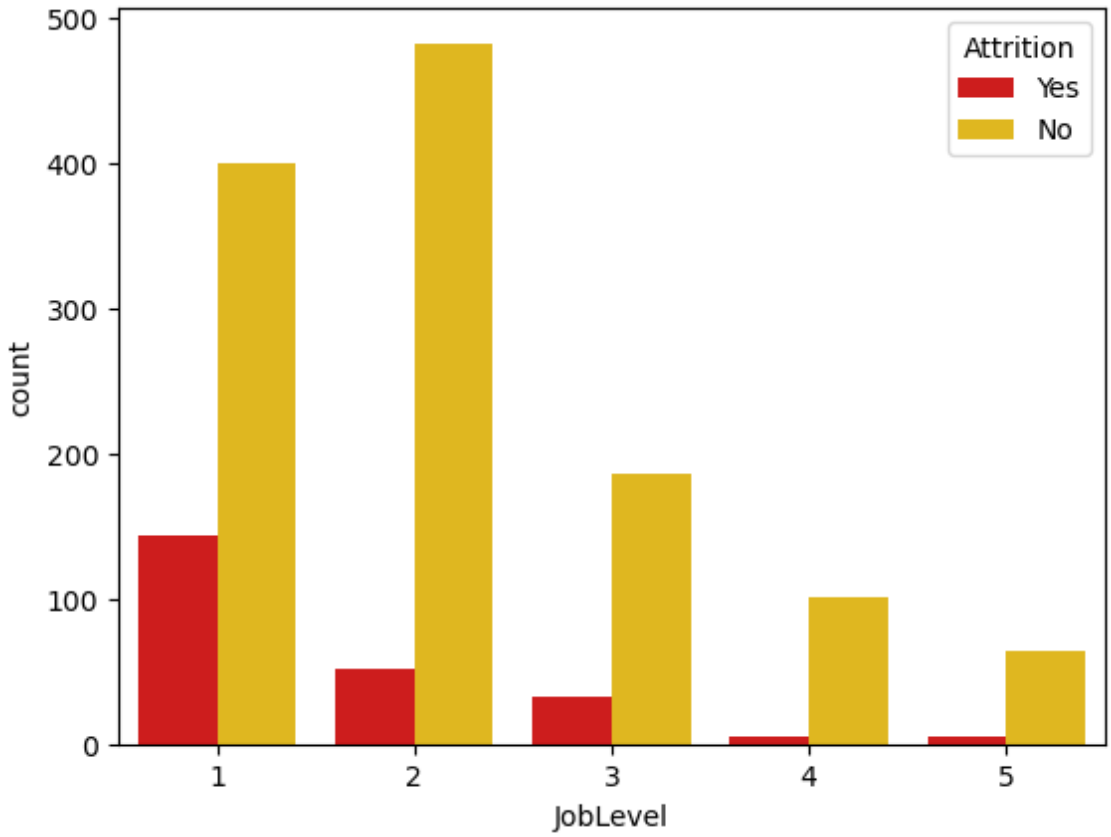
In [60]:

```
sns.countplot(x='JobInvolvement', hue='Attrition', data=df, palette='hot')
plt.show()
```



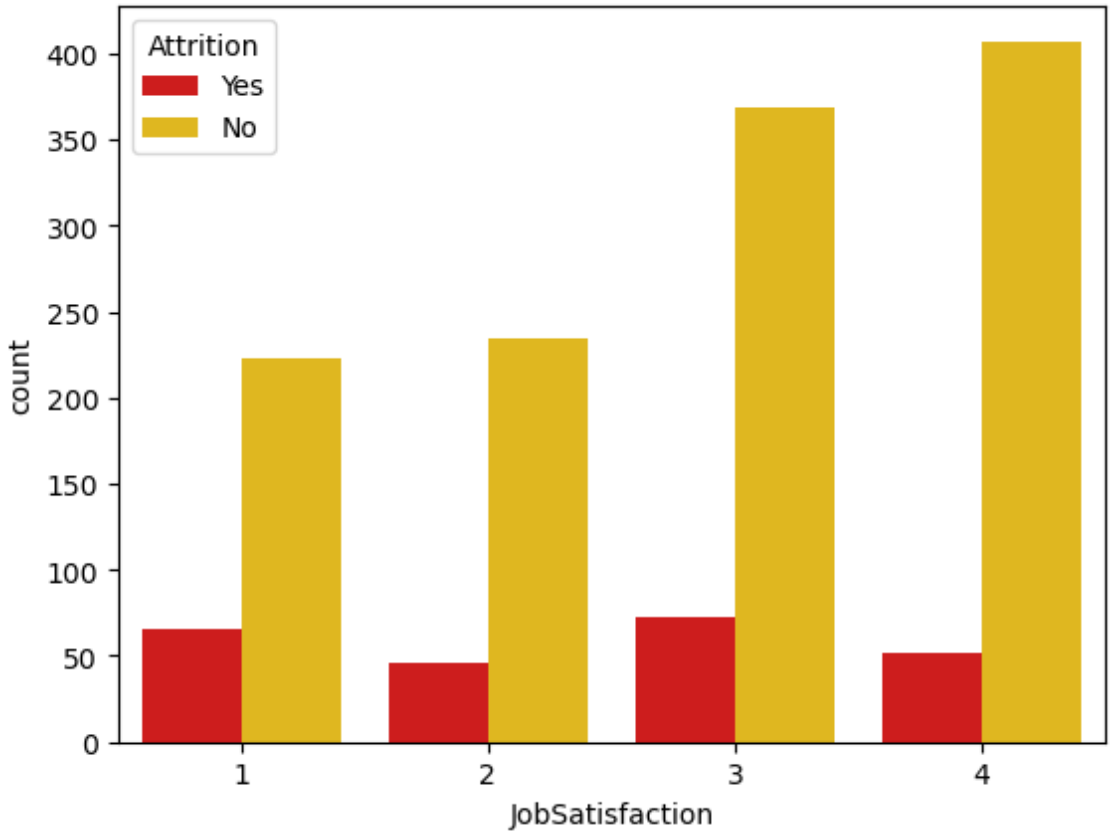
In [61]:

```
sns.countplot(x='JobLevel', hue='Attrition', data=df, palette='hot')
plt.show()
```



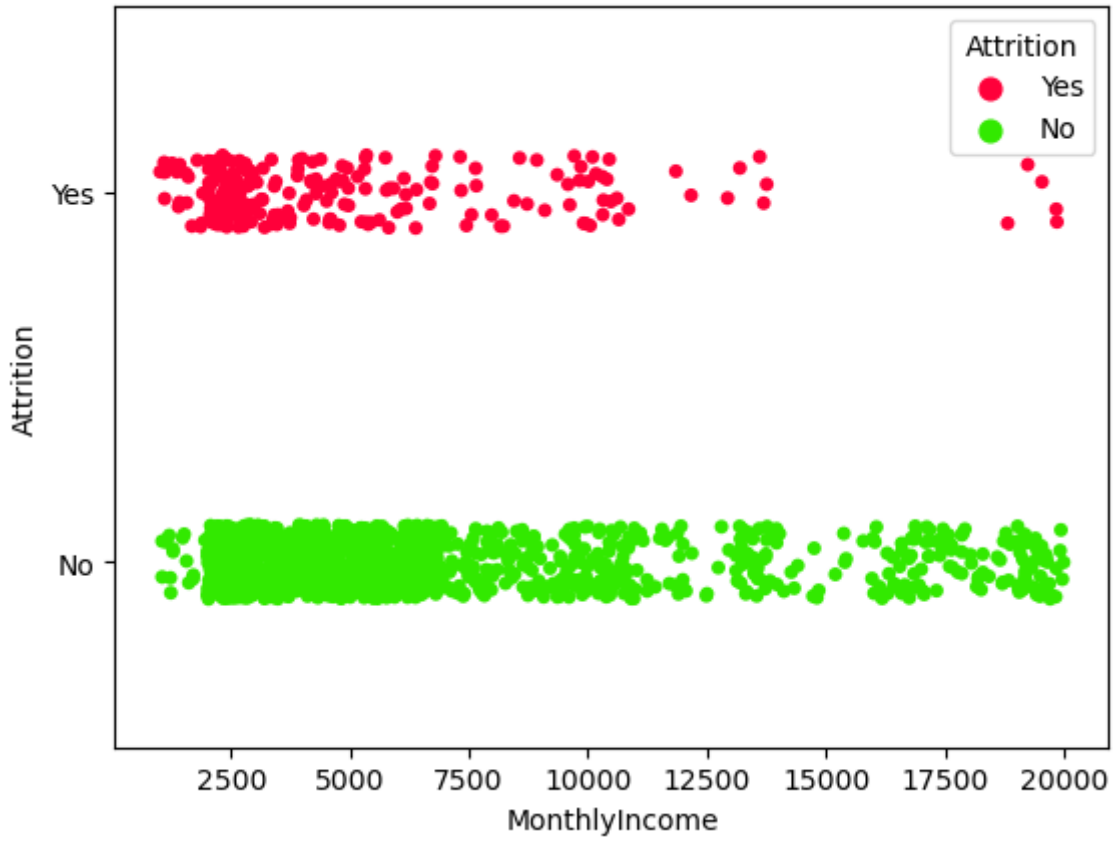
In [62]:

```
sns.countplot(x='JobSatisfaction', hue='Attrition', data=df, palette='hot')
plt.show()
```



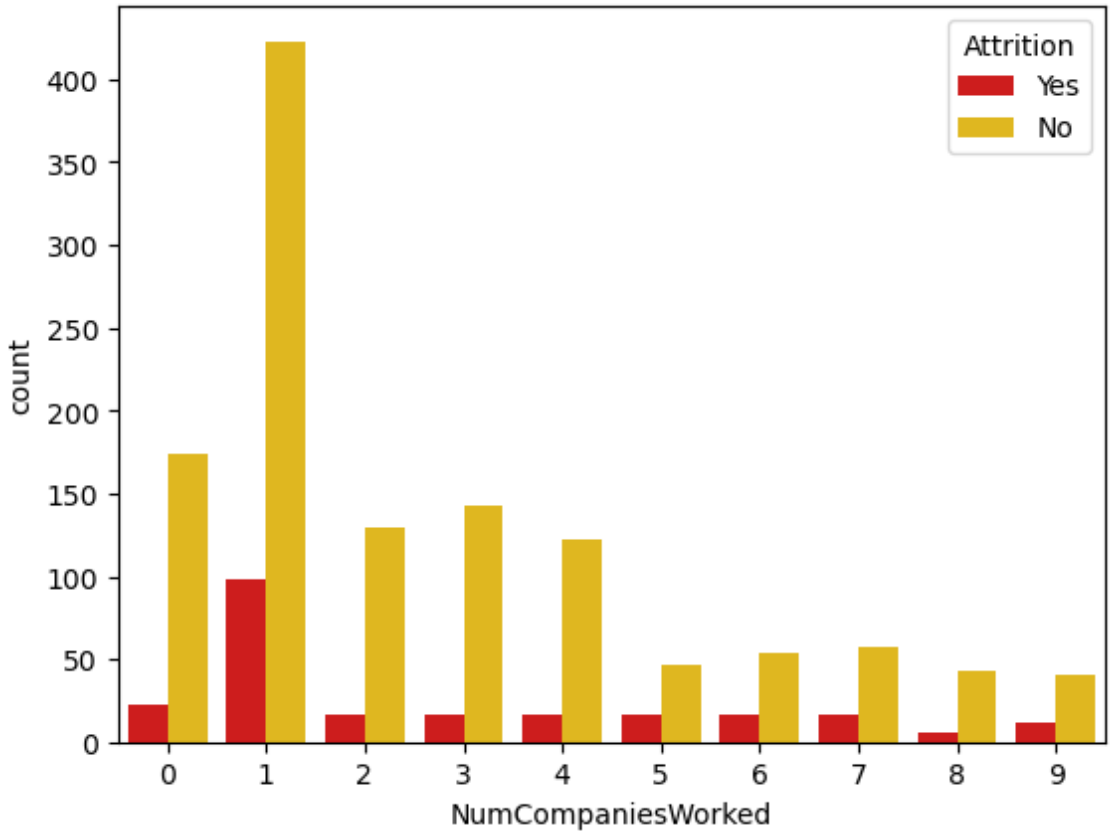
In [63]:

```
sns.stripplot(data=df, x='MonthlyIncome', y='Attrition', palette='prism_r', hue='Attriti
plt.show()
```



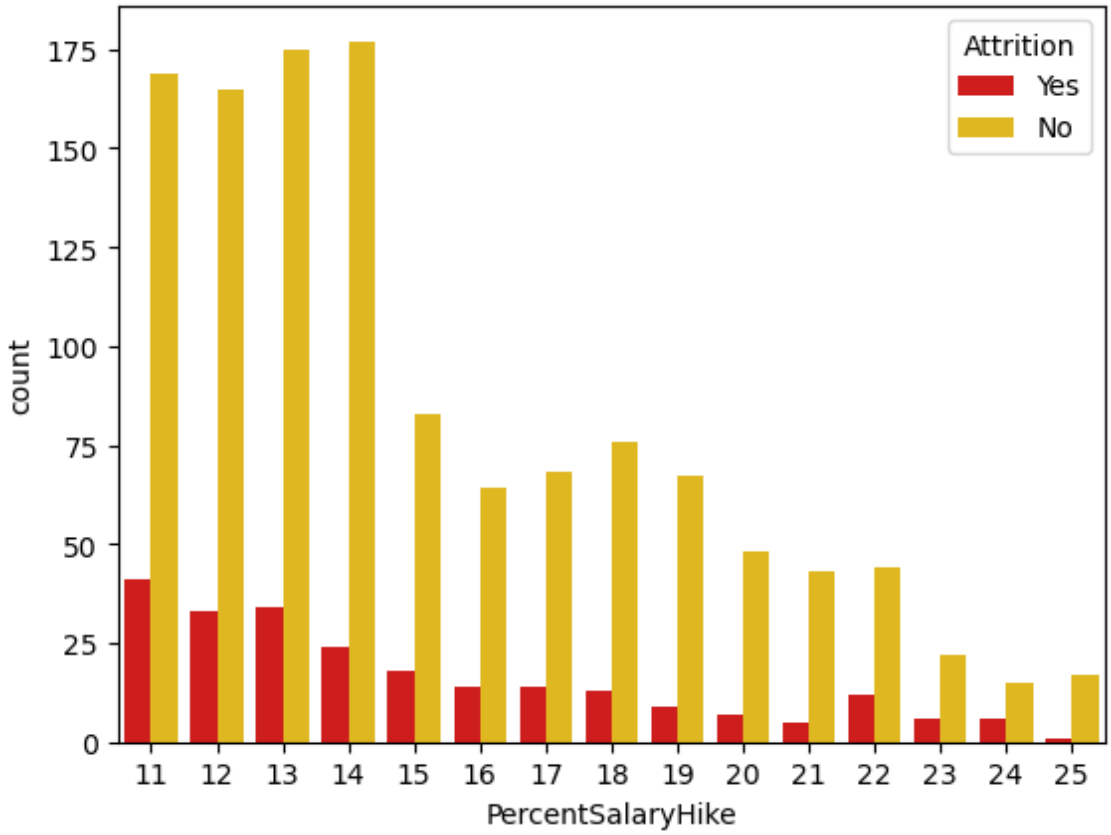
In [64]:

```
sns.countplot(x='NumCompaniesWorked', hue='Attrition', data=df, palette='hot')
plt.show()
```



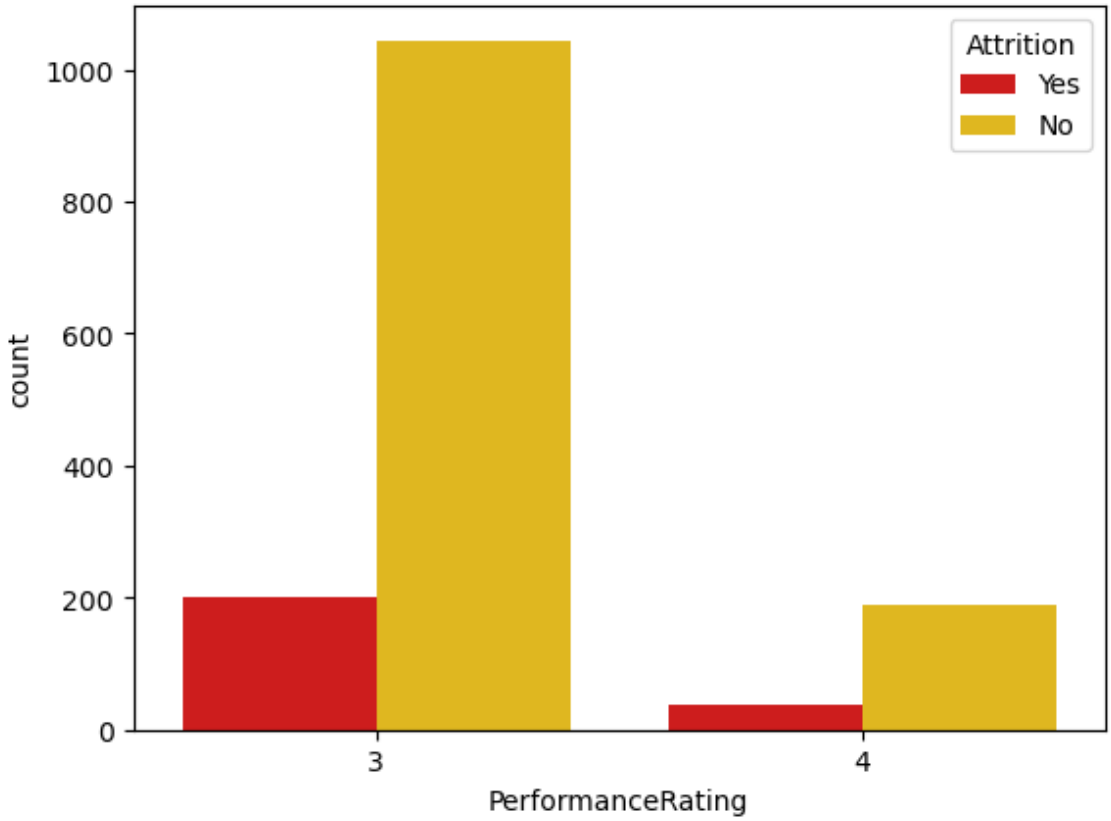
In [65]:

```
sns.countplot(x='PercentSalaryHike', hue='Attrition', data=df, palette='hot')
plt.show()
```



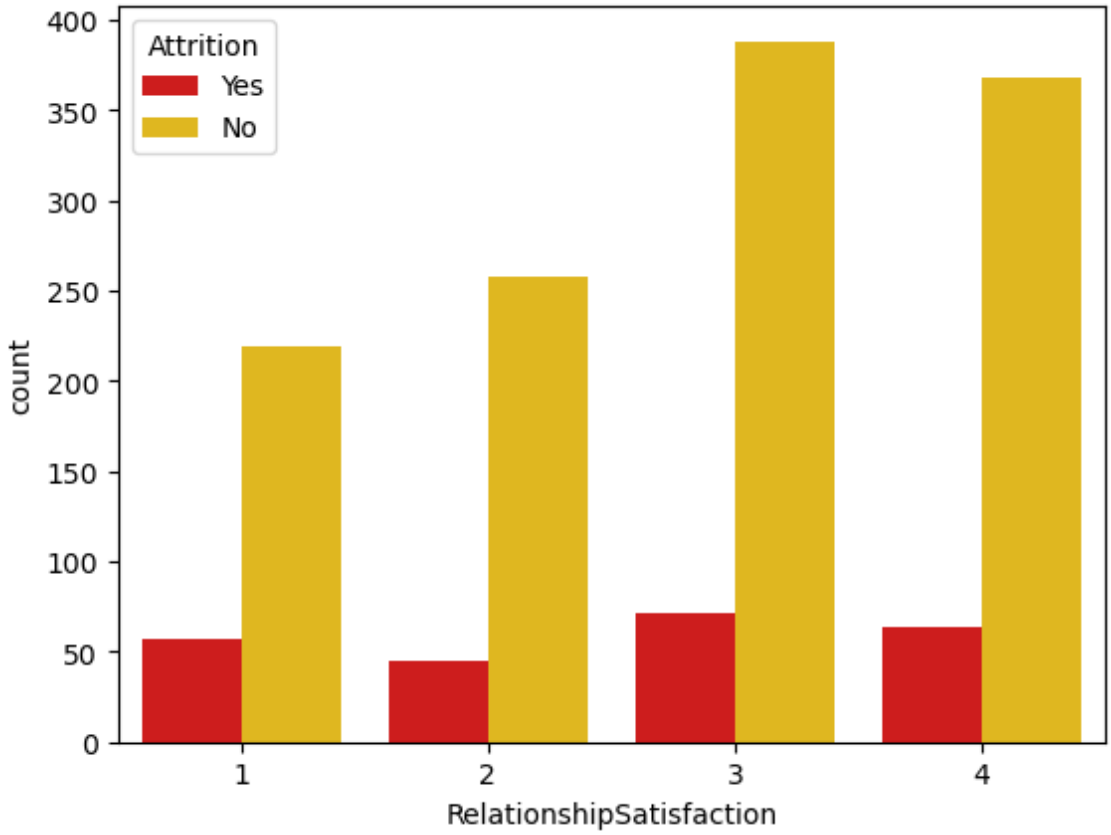
In [66]:

```
sns.countplot(x='PerformanceRating', hue='Attrition', data=df, palette='hot')
plt.show()
```



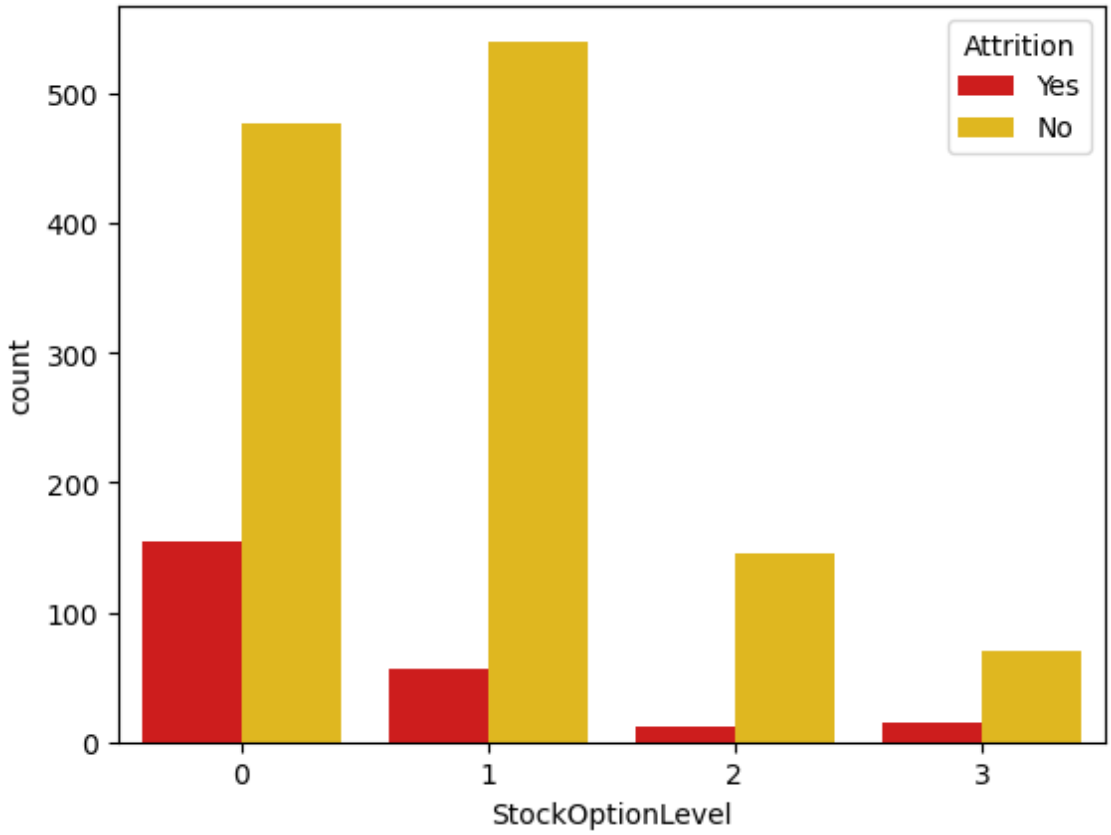
In [67]:

```
sns.countplot(x='RelationshipSatisfaction', hue='Attrition', data=df, palette='hot')
plt.show()
```



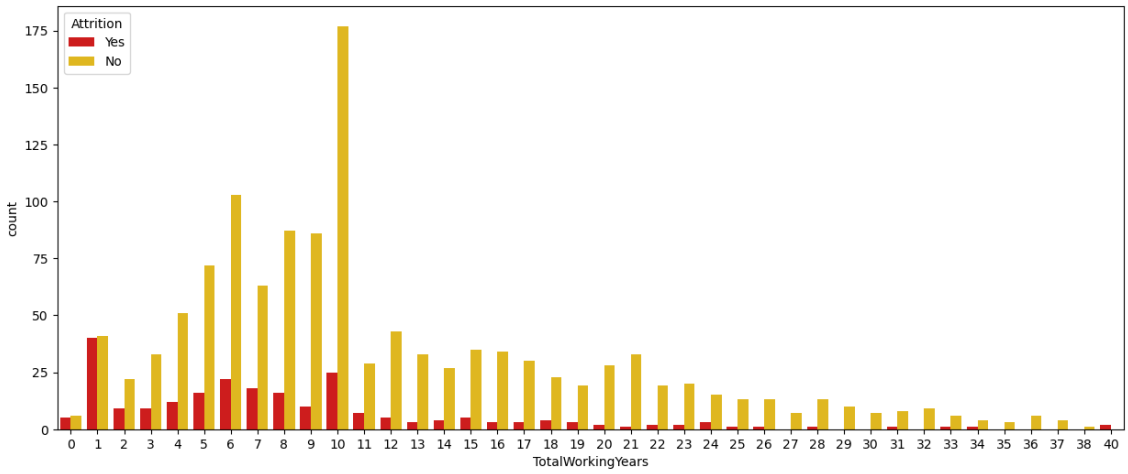
In [68]:

```
sns.countplot(x='StockOptionLevel', hue='Attrition', data=df, palette='hot')
plt.show()
```



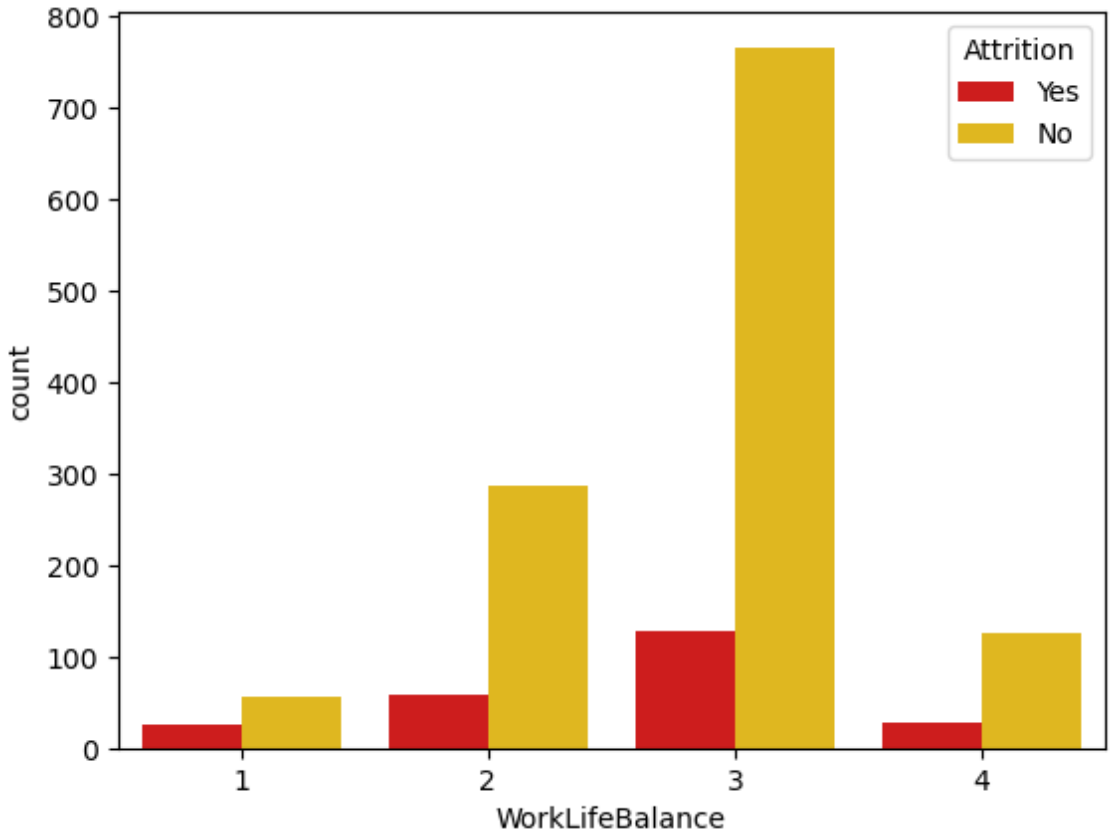
In [69]:

```
plt.figure(figsize=(15,6))
sns.countplot(x='TotalWorkingYears', hue='Attrition', data=df, palette='hot')
plt.show()
```



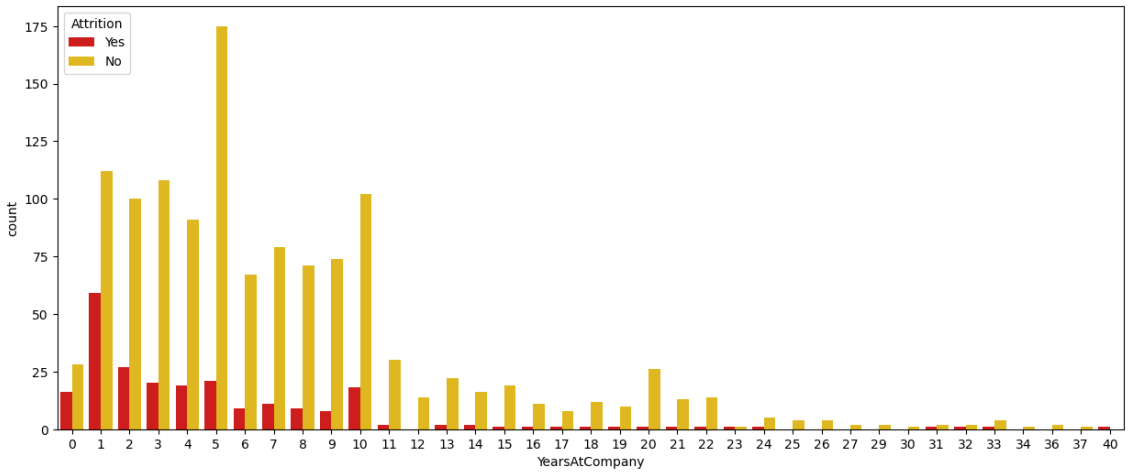
In [70]:

```
sns.countplot(x='WorkLifeBalance', hue='Attrition', data=df, palette='hot')
plt.show()
```



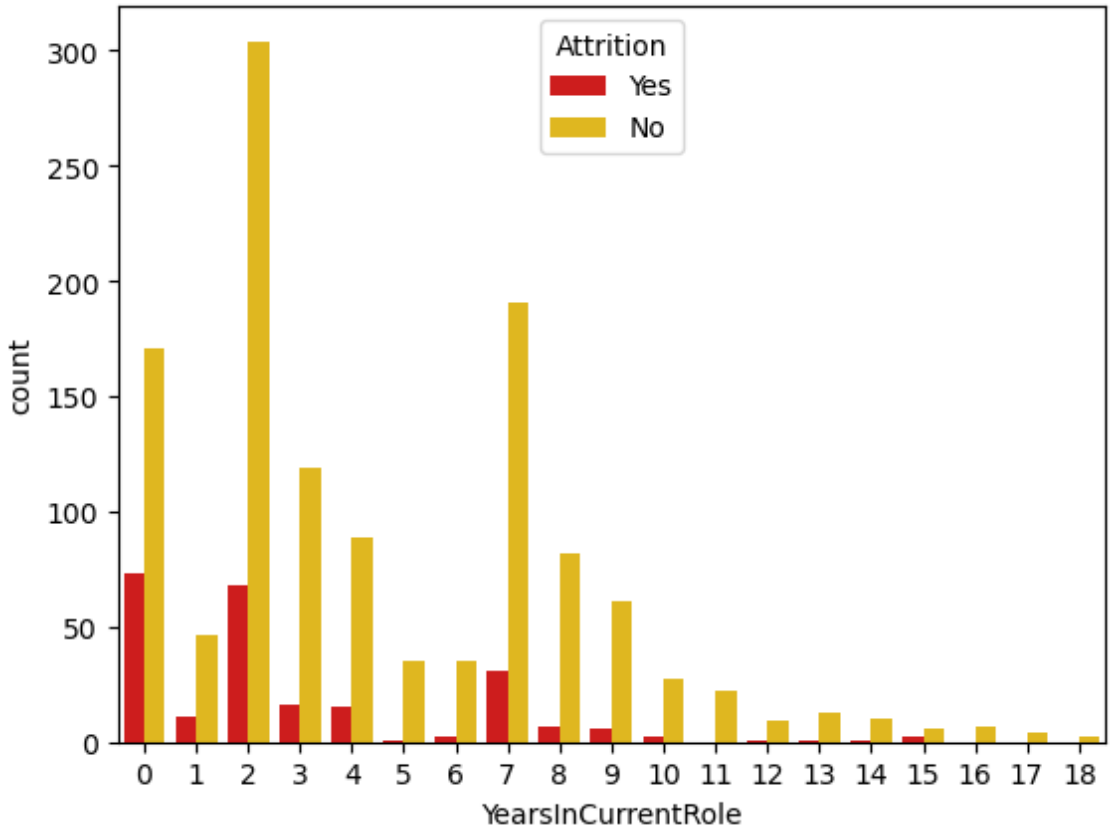
In [71]:

```
plt.figure(figsize=(15,6))
sns.countplot(x='YearsAtCompany', hue='Attrition', data=df, palette='hot')
plt.show()
```



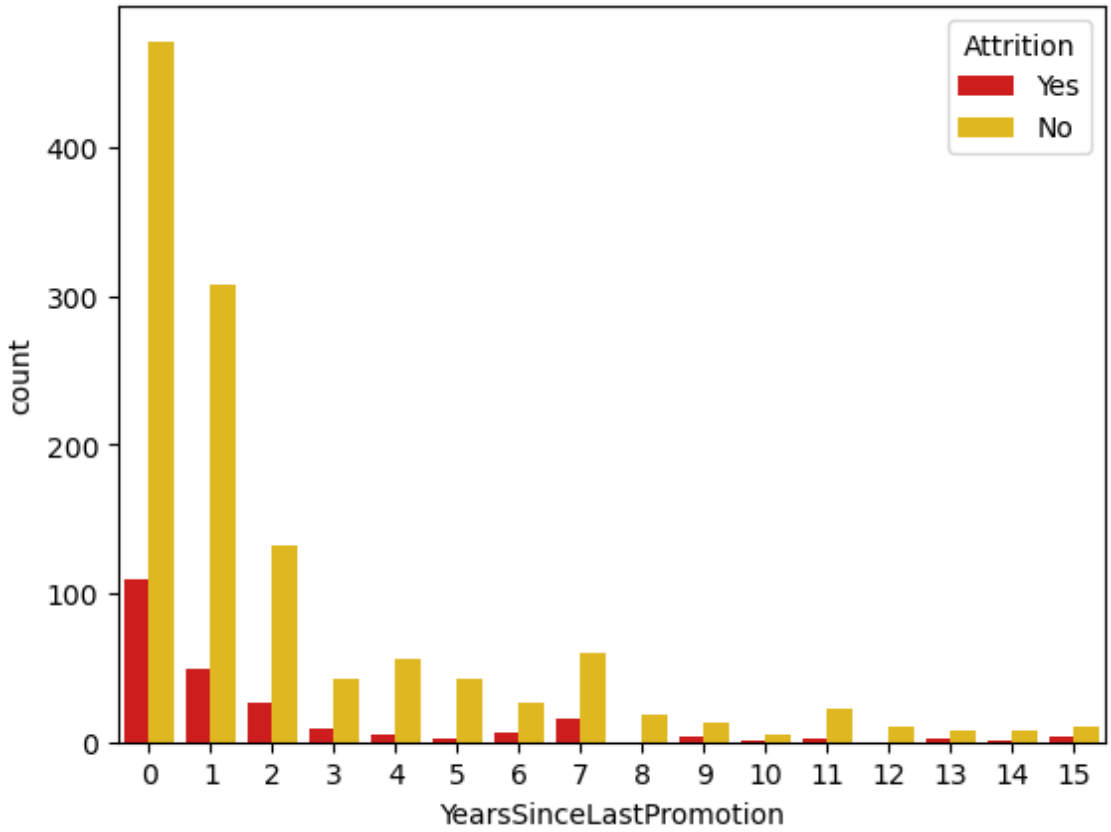
In [72]:

```
sns.countplot(x='YearsInCurrentRole', hue='Attrition', data=df, palette='hot')
plt.show()
```



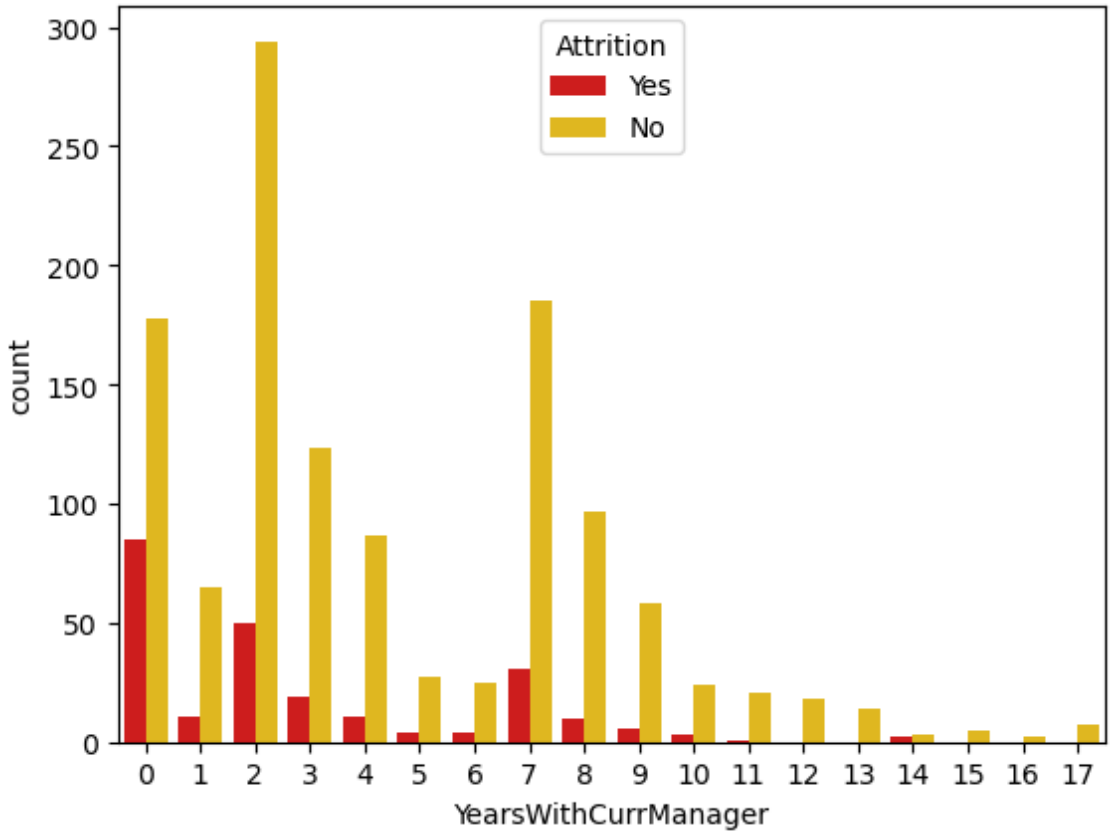
In [73]:

```
sns.countplot(x='YearsSinceLastPromotion', hue='Attrition', data=df, palette='hot')
plt.show()
```



In [74]:

```
sns.countplot(x='YearsWithCurrManager', hue='Attrition', data=df, palette='hot')
plt.show()
```



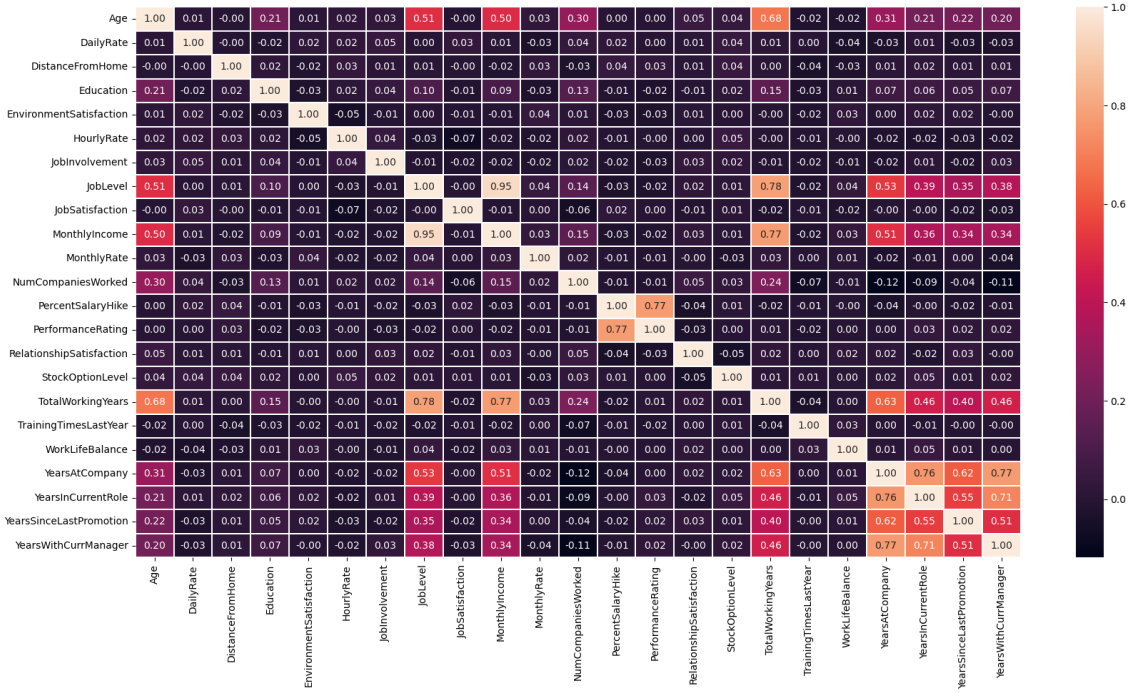
Some Observations:

1. Young employees aged below 22 yrs, quit their jobs more than the rest.
2. Employees who travel more than 10 kms to reach office, are more likely to quit.
3. Environment Satisfaction, Job Satisfaction, Relationship Satisfaction, Job Involvement, Performance Rating, Stock Option Level, Work Life Balance: these features don't really help us in understanding the employees' attrition.
4. Employees with low Job Level, Monthly Income, Percent Salary Hike, Total Working Years, Years At Company are prone to quitting their jobs.
5. Employees who have worked in less than 2 comapanies, are more likely to stay.
6. Employees who have received promotion recently within 2 years, will stay than employees who haven't received any promotion for a long time.
7. Employees who have spent more than 2 years with their current manager, are more likely to stay.

Correlation Matrix

In [75]:

```
plt.figure(figsize=(20,10))
sns.heatmap(df.corr(),annot=True,fmt='.2f',linewidth='0.2')
plt.show()
```



Some Observations:

1. Job Level and Monthly Income are highly correlated.
2. Monthly Income is highly correlated with Total Working Hours.
3. Job Level and Total Working Hours are highly correlated.
4. Performance Rating is highly correlated with Percent Salary Hike.
5. Years in Current Role and Years with Current Manager has high correlation with Years at Company.

Thank you...

In [ ]:

```
```