HR DATA ANALYSIS PROJECT

In [1]:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
```

In [2]:

```
df =pd.read_csv("C:/Users/Asus/Downloads/HR Data.csv")
df
```

Out[2]:

	Attrition	Business Travel	CF_age band	CF_attrition label	Department	Education Field	emp no	Employee Number	Gender	Job Role	...	Performance Rating	F
0	Yes	Travel_Rarely	35 - 44	Ex-Employees	Sales	Life Sciences	STAFF-1	1	Female	Sales Executive	...	3	
1	No	Travel_Frequently	45 - 54	Current Employees	R&D	Life Sciences	STAFF-2	2	Male	Research Scientist	...	4	
2	Yes	Travel_Rarely	35 - 44	Ex-Employees	R&D	Other	STAFF-4	4	Male	Laboratory Technician	...	3	
3	No	Travel_Frequently	25 - 34	Current Employees	R&D	Life Sciences	STAFF-5	5	Female	Research Scientist	...	3	
4	No	Travel_Rarely	25 - 34	Current Employees	R&D	Medical	STAFF-7	7	Male	Laboratory Technician	...	3	
...	...	...	...	...	...	...	...	...	...	...	...	...	
1465	Yes	Non-Travel	25 - 34	Ex-Employees	R&D	Technical Degree	STAFF-1905	1905	Male	Research Scientist	...	4	
1466	Yes	Travel_Frequently	25 - 34	Ex-Employees	R&D	Life Sciences	STAFF-1868	1868	Male	Research Scientist	...	4	
1467	Yes	Travel_Frequently	35 - 44	Ex-Employees	Sales	Other	STAFF-1667	1667	Male	Sales Executive	...	4	
1468	Yes	Travel_Rarely	Under 25	Ex-Employees	R&D	Life Sciences	STAFF-1878	1878	Male	Research Scientist	...	4	
1469	Yes	Travel_Rarely	Under 25	Ex-Employees	Sales	Life Sciences	STAFF-1702	1702	Male	Sales Representative	...	4	

1470 rows × 41 columns

In [3]:

```
p = df.head(10)
p
```

Out[3]:

	Attrition	Business Travel	CF_age band	CF_attrition label	Department	Education Field	emp no	Employee Number	Gender	Job Role	...	Performance Rating	Rela Sati
0	Yes	Travel_Rarely	35 - 44	Ex-Employees	Sales	Life Sciences	STAFF-1	1	Female	Sales Executive	...	3	
1	No	Travel_Frequently	45 - 54	Current Employees	R&D	Life Sciences	STAFF-2	2	Male	Research Scientist	...	4	
2	Yes	Travel_Rarely	35 - 44	Ex-Employees	R&D	Other	STAFF-4	4	Male	Laboratory Technician	...	3	
3	No	Travel_Frequently	25 - 34	Current Employees	R&D	Life Sciences	STAFF-5	5	Female	Research Scientist	...	3	
4	No	Travel_Rarely	25 - 34	Current Employees	R&D	Medical	STAFF-7	7	Male	Laboratory Technician	...	3	
5	No	Travel_Frequently	25 - 34	Current Employees	R&D	Life Sciences	STAFF-8	8	Male	Laboratory Technician	...	3	
6	No	Travel_Rarely	Over 55	Current Employees	R&D	Medical	STAFF-10	10	Female	Laboratory Technician	...	4	
7	No	Travel_Rarely	25 - 34	Current Employees	R&D	Life Sciences	STAFF-11	11	Male	Laboratory Technician	...	4	
8	No	Travel_Frequently	35 - 44	Current Employees	R&D	Life Sciences	STAFF-12	12	Male	Manufacturing Director	...	4	
9	No	Travel_Rarely	35 - 44	Current Employees	R&D	Medical	STAFF-13	13	Male	Healthcare Representative	...	3	

10 rows × 41 columns

In [4]:

```
p1 = df.tail(10)
p1
```

Out[4]:

	Attrition	Business Travel	CF_age band	CF_attrition label	Department	Education Field	emp no	Employee Number	Gender	Job Role	...	Performance Rating
1460	Yes	Travel_Rarely	45 - 54	Ex-Employees	Sales	Life Sciences	STAFF-1869	1869	Female	Sales Executive	...	4
1461	Yes	Travel_Frequently	45 - 54	Ex-Employees	R&D	Life Sciences	STAFF-1420	1420	Male	Laboratory Technician	...	4
1462	Yes	Non-Travel	35 - 44	Ex-Employees	R&D	Life Sciences	STAFF-1458	1458	Female	Laboratory Technician	...	4
1463	Yes	Travel_Rarely	25 - 34	Ex-Employees	Sales	Medical	STAFF-1489	1489	Female	Sales Executive	...	4
1464	Yes	Travel_Rarely	25 - 34	Ex-Employees	Sales	Life Sciences	STAFF-1758	1758	Female	Sales Executive	...	4
1465	Yes	Non-Travel	25 - 34	Ex-Employees	R&D	Technical Degree	STAFF-1905	1905	Male	Research Scientist	...	4
1466	Yes	Travel_Frequently	25 - 34	Ex-Employees	R&D	Life Sciences	STAFF-1868	1868	Male	Research Scientist	...	4
1467	Yes	Travel_Frequently	35 - 44	Ex-Employees	Sales	Other	STAFF-1667	1667	Male	Sales Executive	...	4
1468	Yes	Travel_Rarely	Under 25	Ex-Employees	R&D	Life Sciences	STAFF-1878	1878	Male	Research Scientist	...	4
1469	Yes	Travel_Rarely	Under 25	Ex-Employees	Sales	Life Sciences	STAFF-1702	1702	Male	Sales Representative	...	4

10 rows × 41 columns

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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1470 entries, 0 to 1469
Data columns (total 41 columns):
 #   Column                      Non-Null Count  Dtype 
---  ------                      --------------  ----- 
 0   Attrition                   1470 non-null   object
 1   Business Travel             1470 non-null   object
 2   CF_age band                 1470 non-null   object
 3   CF_attrition label          1470 non-null   object
 4   Department                  1470 non-null   object
 5   Education Field             1470 non-null   object
 6   emp no                      1470 non-null   object
 7   Employee Number             1470 non-null   int64 
 8   Gender                      1470 non-null   object
 9   Job Role                    1470 non-null   object
 10  Marital Status              1470 non-null   object
 11  Over Time                   1470 non-null   object
 12  Over18                      1470 non-null   object
 13  Training Times Last Year    1470 non-null   int64 
 14  -2                          1470 non-null   int64 
 15  0                           1470 non-null   int64 
 16  Age                         1470 non-null   int64 
 17  CF_current Employee         1470 non-null   int64 
 18  Daily Rate                  1470 non-null   int64 
 19  Distance From Home          1470 non-null   int64 
 20  Education                   1470 non-null   object
 21  Employee Count              1470 non-null   int64 
 22  Environment Satisfaction    1470 non-null   int64 
 23  Hourly Rate                 1470 non-null   int64 
 24  Job Involvement             1470 non-null   int64 
 25  Job Level                   1470 non-null   int64 
 26  Job Satisfaction            1470 non-null   int64 
 27  Monthly Income              1470 non-null   int64 
 28  Monthly Rate                1470 non-null   int64 
 29  Num Companies Worked        1470 non-null   int64 
 30  Percent Salary Hike         1470 non-null   int64 
 31  Performance Rating          1470 non-null   int64 
 32  Relationship Satisfaction   1470 non-null   int64 
 33  Standard Hours              1470 non-null   int64 
 34  Stock Option Level          1470 non-null   int64 
 35  Total Working Years         1470 non-null   int64 
 36  Work Life Balance           1470 non-null   int64 
 37  Years At Company            1470 non-null   int64 
 38  Years In Current Role       1470 non-null   int64 
 39  Years Since Last Promotion  1470 non-null   int64 
 40  Years With Curr Manager     1470 non-null   int64 
dtypes: int64(28), object(13)
memory usage: 471.0+ KB
```

In [6]:
```
df.shape
```

Out[6]:
```
(1470, 41)
```

In [7]:
```
df.Attrition.value_counts()
```

Out[7]:
```
No     1233
Yes     237
Name: Attrition, dtype: int64
```

In [8]:
```
df.describe()
```

Out[8]:

	Employee Number	Training Times Last Year	-2	0	Age	CF_current Employee	Daily Rate	Distance From Home	Employee Count	Environment Satisfaction	...	Performan Rati
count	1470.000000	1470.000000	1470.0	1470.0	1470.000000	1470.000000	1470.000000	1470.000000	1470.0	1470.000000	...	1470.0000
mean	1024.865306	2.799320	-2.0	0.0	36.923810	0.838776	802.485714	9.192517	1.0	2.721769	...	3.1537
std	602.024335	1.289271	0.0	0.0	9.135373	0.367863	403.509100	8.106864	0.0	1.093082	...	0.3608
min	1.000000	0.000000	-2.0	0.0	18.000000	0.000000	102.000000	1.000000	1.0	1.000000	...	3.0000
25%	491.250000	2.000000	-2.0	0.0	30.000000	1.000000	465.000000	2.000000	1.0	2.000000	...	3.0000
50%	1020.500000	3.000000	-2.0	0.0	36.000000	1.000000	802.000000	7.000000	1.0	3.000000	...	3.0000
75%	1555.750000	3.000000	-2.0	0.0	43.000000	1.000000	1157.000000	14.000000	1.0	4.000000	...	3.0000
max	2068.000000	6.000000	-2.0	0.0	60.000000	1.000000	1499.000000	29.000000	1.0	4.000000	...	4.0000

8 rows × 28 columns

In [9]:
```
df.isnull().sum()
```

Out[9]:
```
Attrition                     0
Business Travel               0
CF_age band                   0
CF_attrition label            0
Department                    0
Education Field               0
emp no                        0
Employee Number               0
Gender                        0
Job Role                      0
Marital Status                0
Over Time                     0
Over18                        0
Training Times Last Year      0
-2                            0
0                             0
Age                           0
CF_current Employee           0
Daily Rate                    0
Distance From Home            0
Education                     0
Employee Count                0
Environment Satisfaction      0
Hourly Rate                   0
Job Involvement               0
Job Level                     0
Job Satisfaction              0
Monthly Income                0
Monthly Rate                  0
Num Companies Worked          0
Percent Salary Hike           0
Performance Rating            0
Relationship Satisfaction     0
Standard Hours                0
Stock Option Level            0
Total Working Years           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
```

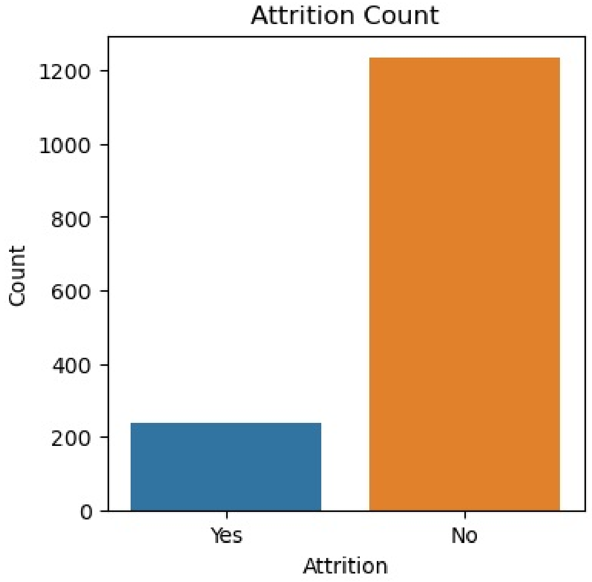
In [10]:
```
print(df.columns)
```

```
Index(['Attrition', 'Business Travel', 'CF_age band', 'CF_attrition label',
       'Department', 'Education Field', 'emp no', 'Employee Number', 'Gender',
       'Job Role', 'Marital Status', 'Over Time', 'Over18',
       'Training Times Last Year', '-2', '0', 'Age', 'CF_current Employee',
       'Daily Rate', 'Distance From Home', 'Education', 'Employee Count',
       'Environment Satisfaction', 'Hourly Rate', 'Job Involvement',
       'Job Level', 'Job Satisfaction', 'Monthly Income', 'Monthly Rate',
       'Num Companies Worked', 'Percent Salary Hike', 'Performance Rating',
       'Relationship Satisfaction', 'Standard Hours', 'Stock Option Level',
       'Total Working Years', 'Work Life Balance', 'Years At Company',
       'Years In Current Role', 'Years Since Last Promotion',
       'Years With Curr Manager'],
      dtype='object')
```

Count plot showing total no of attrition count

In [11]:

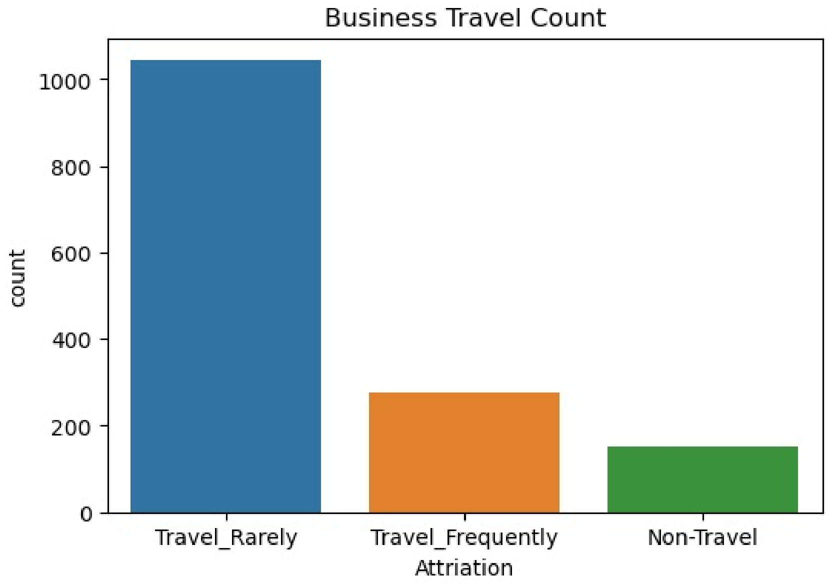
```
plt.figure(figsize=(4, 4))
sns.countplot(x='Attrition', data=df)
plt.title('Attrition Count')
plt.xlabel('Attrition')
plt.ylabel('Count')
plt.show()
plt.show()
```



count plot showing Business Travel Count

In [12]:

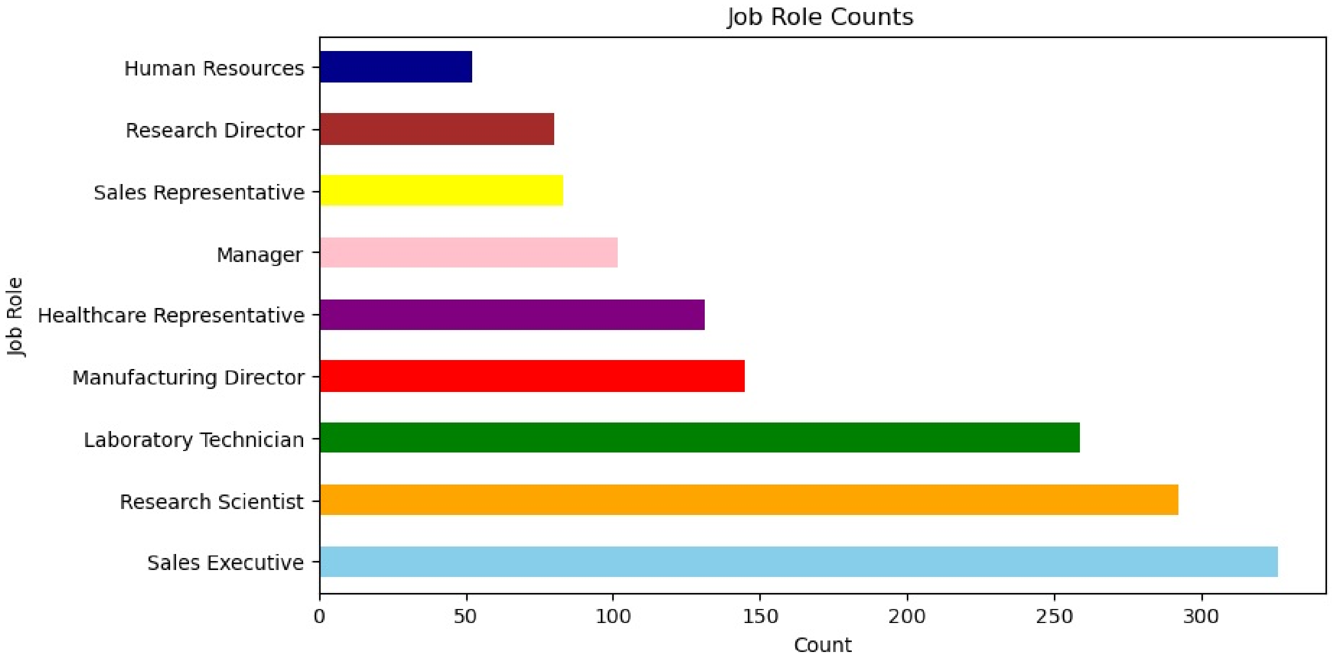
```
plt.figure(figsize=(6, 4))
sns.countplot(x='Business Travel', data=df)
plt.title('Business Travel Count')
plt.xlabel('Attriation')
plt.ylabel('count')
plt.show()
```



Total no of job role count

In [13]:

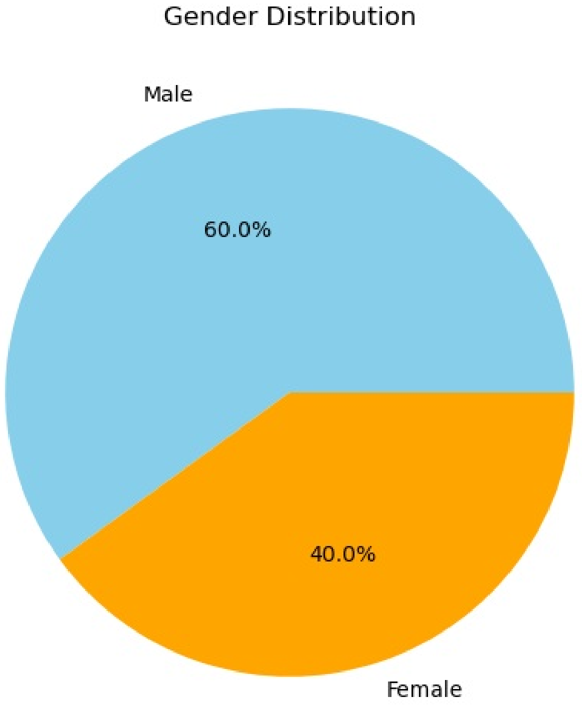
```
job_role_counts = df['Job Role'].value_counts()
colors = ['skyblue', 'orange', 'green', 'red', 'purple', 'pink', 'yellow', 'brown', 'darkblue']
plt.figure(figsize=(9, 5))
job_role_counts.plot(kind='barh', color=colors)
plt.title('Job Role Counts')
plt.xlabel('Count')
plt.ylabel('Job Role')
plt.show()
```



Total no of Gender Distribution

In [14]:

```
gender_counts = df['Gender'].value_counts()
plt.figure(figsize=(6, 6))
plt.pie(gender_counts, labels=gender_counts.index, autopct='%1.1f%%', colors=colors)
plt.title('Gender Distribution')
plt.show()
```

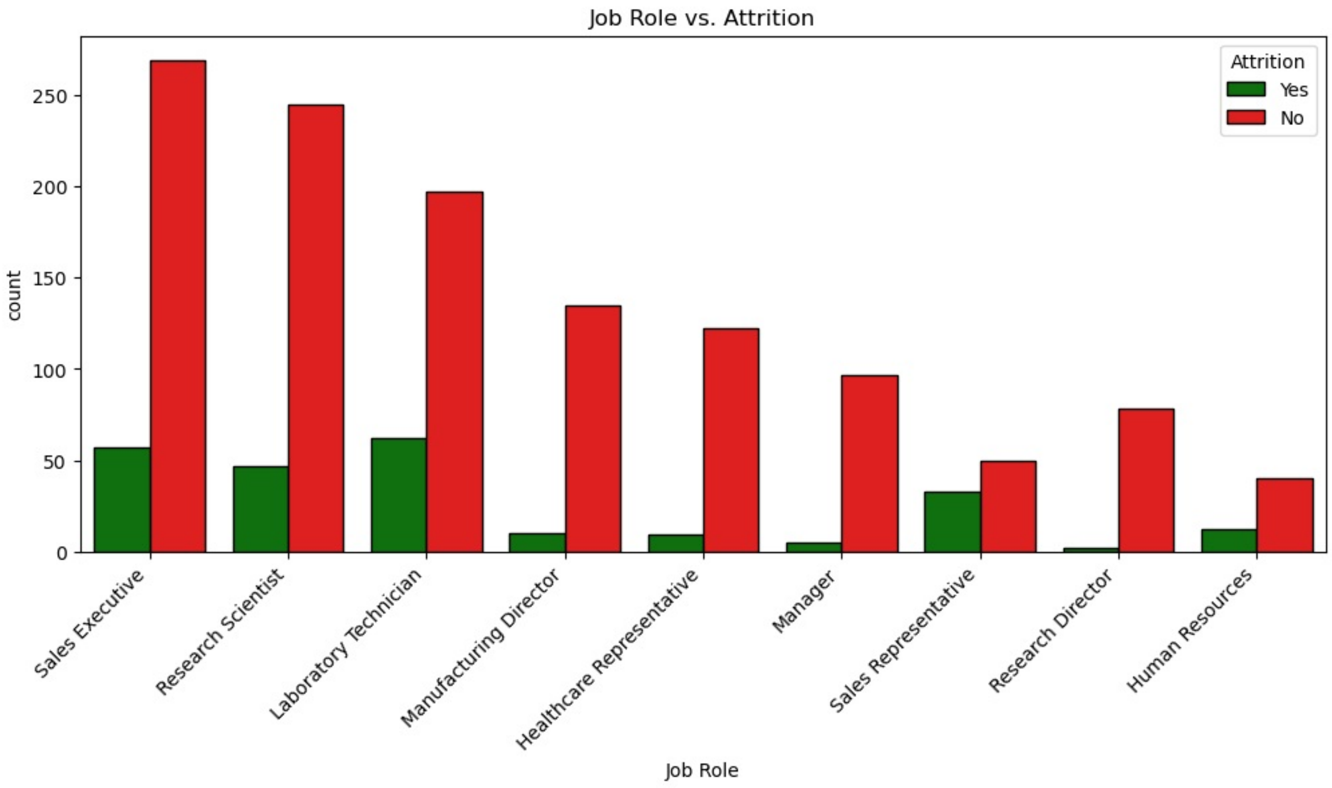


Count plot showing Job role vs Attrition count

In [15]:

```
plt.figure(figsize=(12, 5))
```

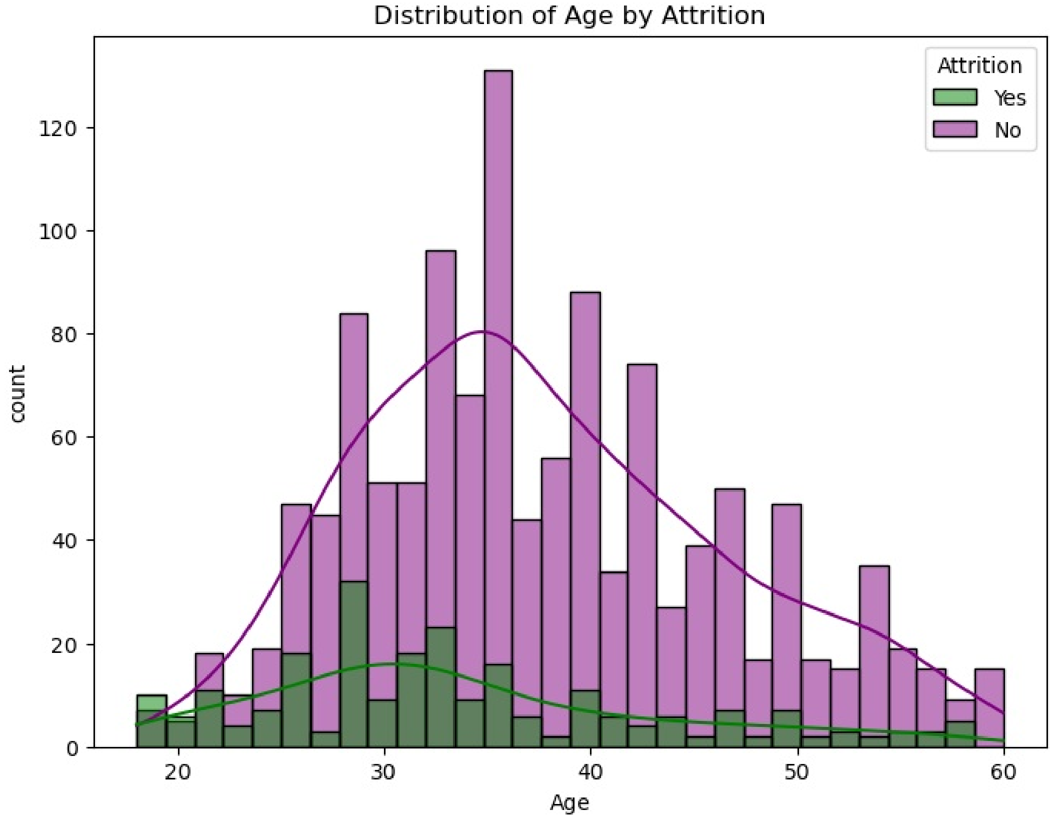
```
sns.countplot(x='Job Role', data=df, hue='Attrition', edgecolor='k', palette={'Yes': 'green', 'No': 'red'})
plt.title('Job Role vs. Attrition')
plt.xticks(rotation=45, ha='right')
plt.show()
```



Hist plot showing Distribution of Age by Attrition

In [16]:

```
plt.figure(figsize=(8, 6))
sns.histplot(data=df, x='Age', bins=30, kde=True, hue='Attrition', edgecolor='k', palette={'Yes': 'green', 'No'
plt.title('Distribution of Age by Attrition')
plt.xlabel('Age')
plt.ylabel('count')
plt.show()
```

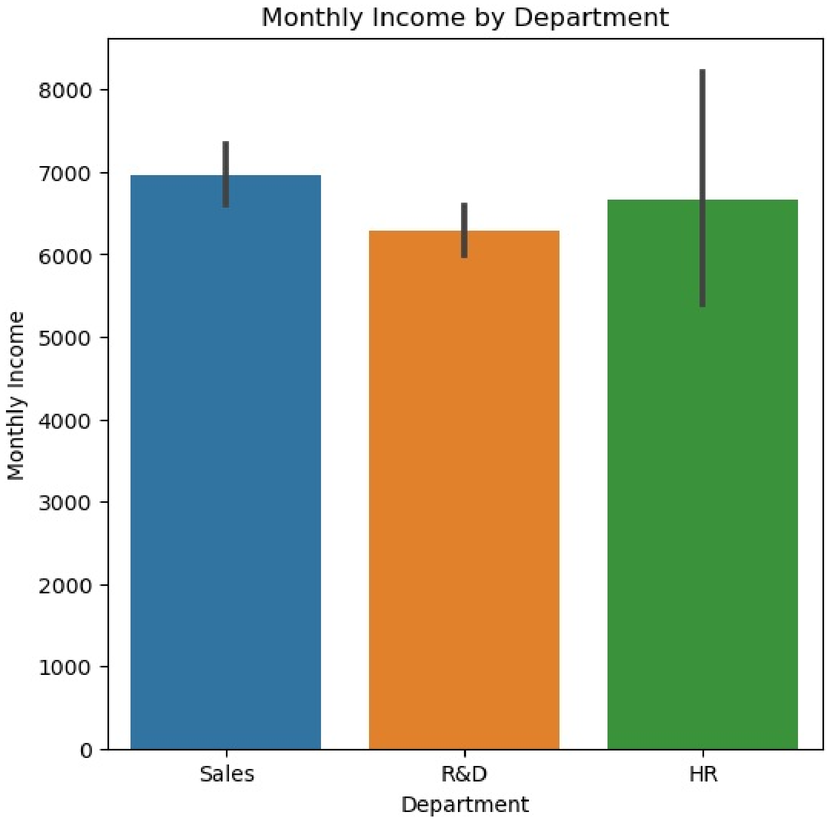


Bar Plot showing total no of monthly income by department

```
plt.figure(figsize=(8, 6))
```

In [17]:

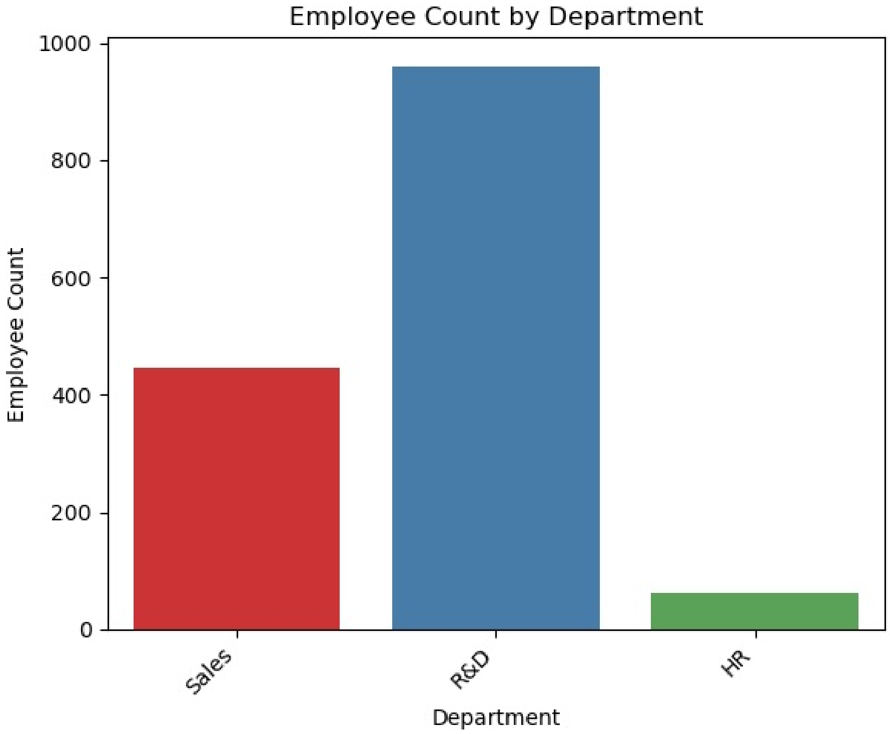
```python
plt.figure(figsize=(6, 6))
sns.barplot(x='Department', y='Monthly Income', data=df)
plt.title('Monthly Income by Department')
plt.show()
```



Count plot showing total no of employee count by department

In [18]:

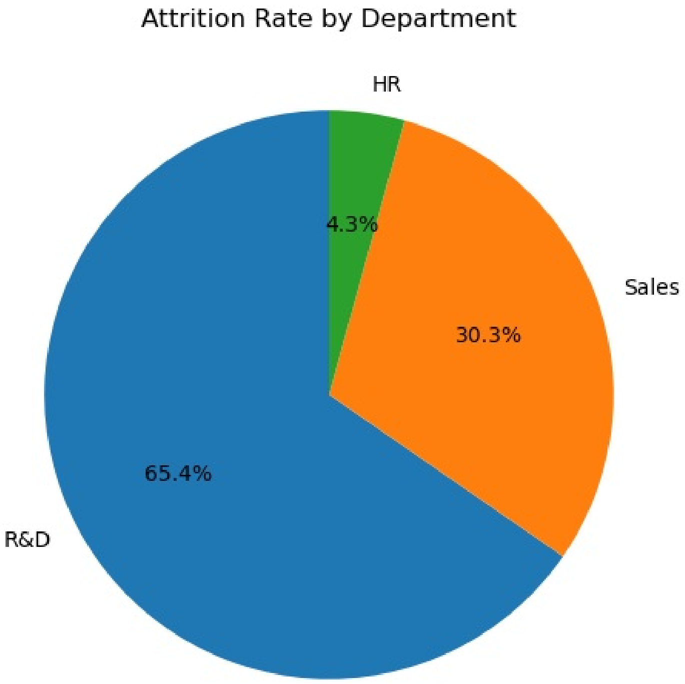
```python
plt.figure(figsize=(6, 5))
sns.countplot(data=df, x='Department', palette='Set1')
plt.xlabel('Department')
plt.ylabel('Employee Count')
plt.title('Employee Count by Department')
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
plt.show()
```



Pie Chart showing Attrition rate by Department

In [19]:

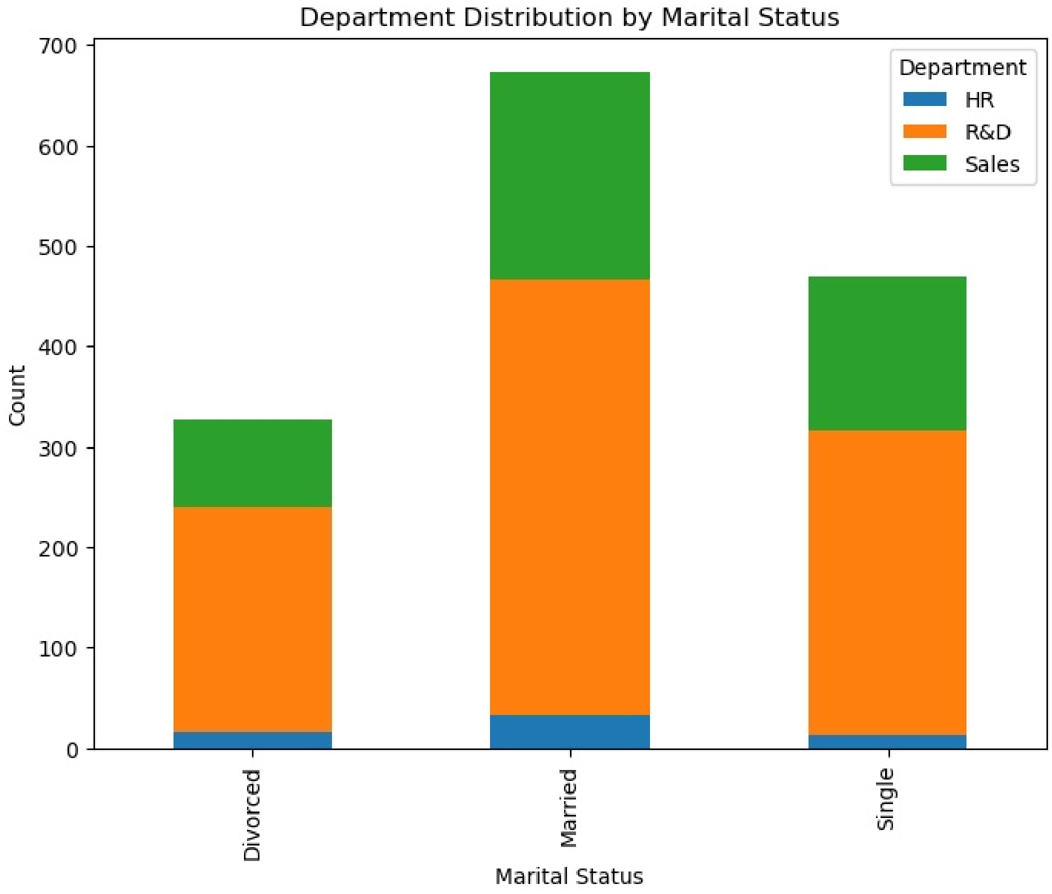
```python
attrition_by_department = df['Department'].value_counts()
plt.figure(figsize=(6, 6))
plt.pie(attrition_by_department, labels=attrition_by_department.index, autopct='%1.1f%%', startangle=90)
plt.title('Attrition Rate by Department')
plt.show()
```



Bar plot showing Distribution of department by marital status

In [20]:

```python
department_by_marital_status = df.groupby('Marital Status')['Department'].value_counts().unstack().fillna(0)
department_by_marital_status.plot(kind='bar', stacked=True, figsize=(8, 6))
plt.xlabel('Marital Status')
plt.ylabel('Count')
plt.title('Department Distribution by Marital Status')
plt.show()
```

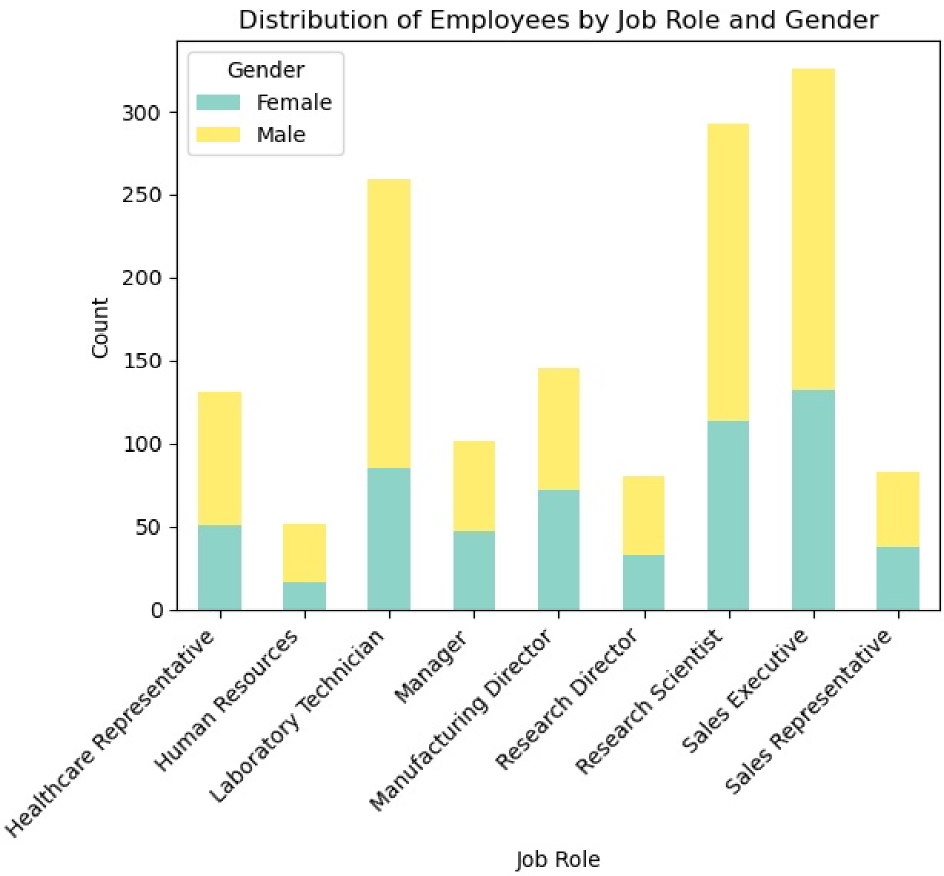


Bar plot showing Distribution of Employees by job role and gender

Bar plot showing Distribution of Employees by job role and gender

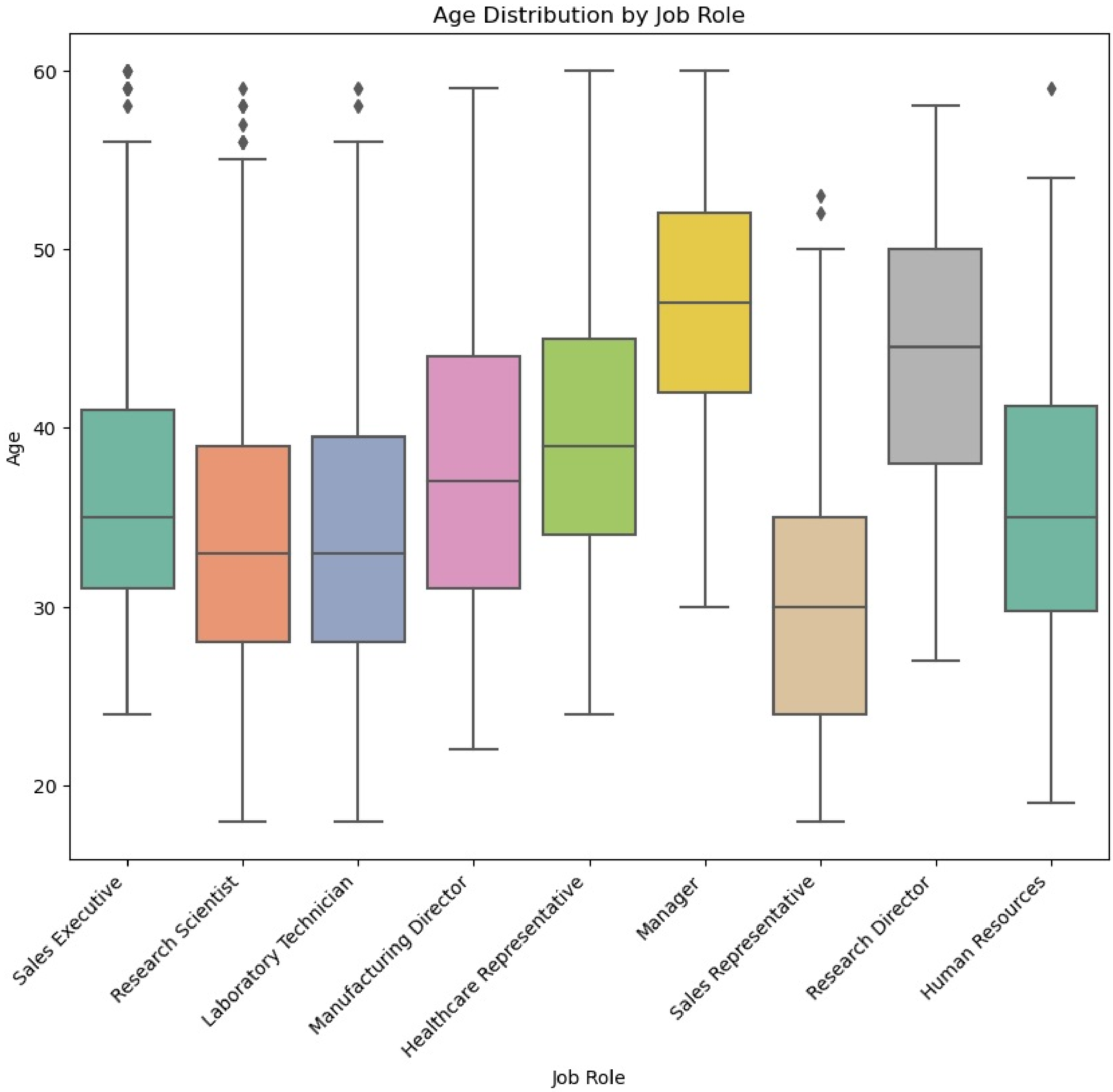
```
In [21]: pivot_df = df.groupby(['Job Role', 'Gender']).size().unstack()
         plt.figure(figsize=(15, 8))
         pivot_df.plot(kind='bar', stacked=True, colormap='Set3')
         plt.xlabel('Job Role')
         plt.ylabel('Count')
         plt.title('Distribution of Employees by Job Role and Gender')
         plt.xticks(rotation=45, ha='right')
         plt.show()
```

```
<Figure size 1500x800 with 0 Axes>
```



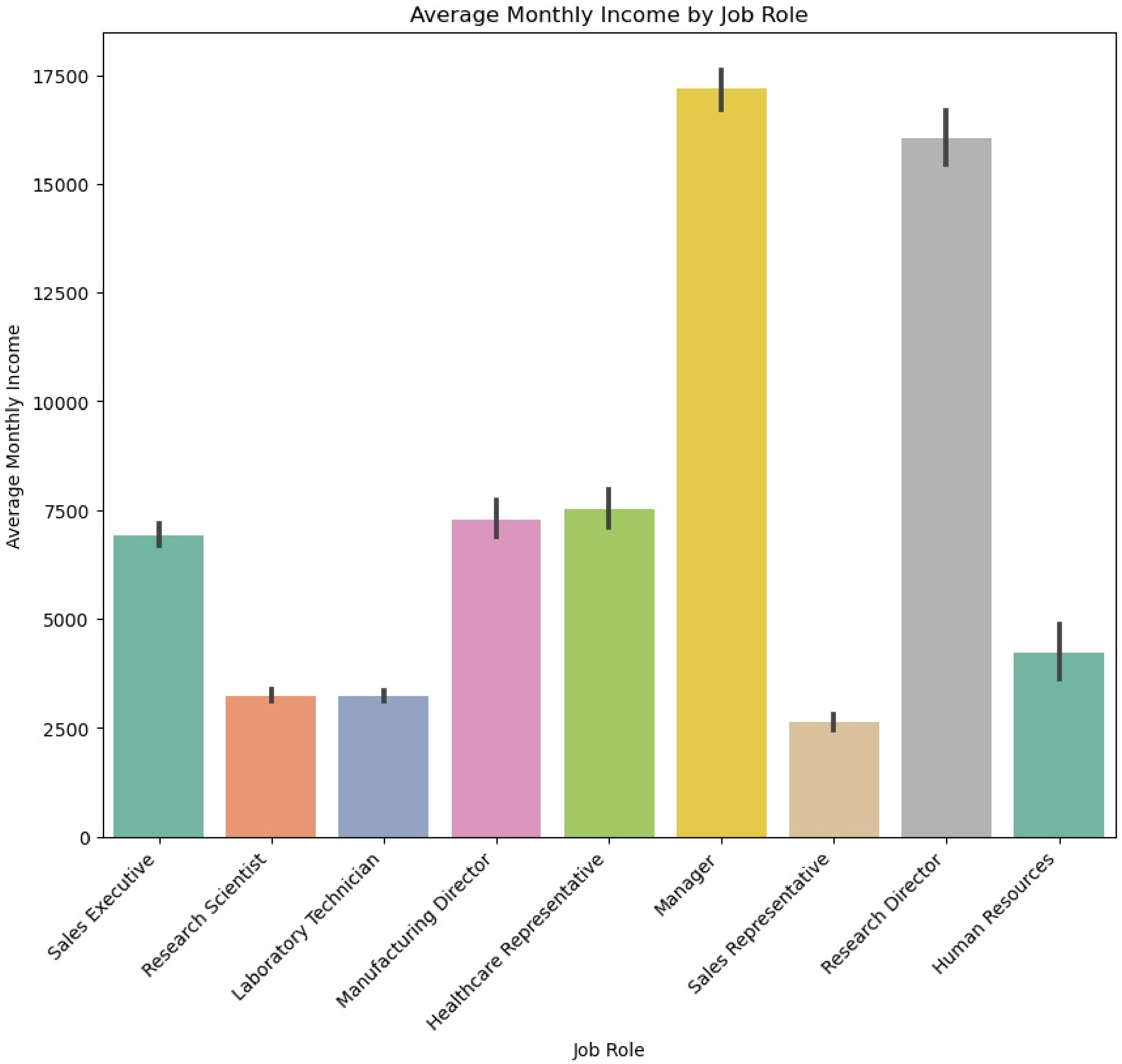
Box plot showing Distribution of age by job role

```
In [22]: plt.figure(figsize=(10, 8))
         sns.boxplot(data=df, x='Job Role', y='Age', palette='Set2')
         plt.xlabel('Job Role')
         plt.ylabel('Age')
         plt.title('Age Distribution by Job Role')
         plt.xticks(rotation=45, ha='right')
         plt.show()
```



In [23]:

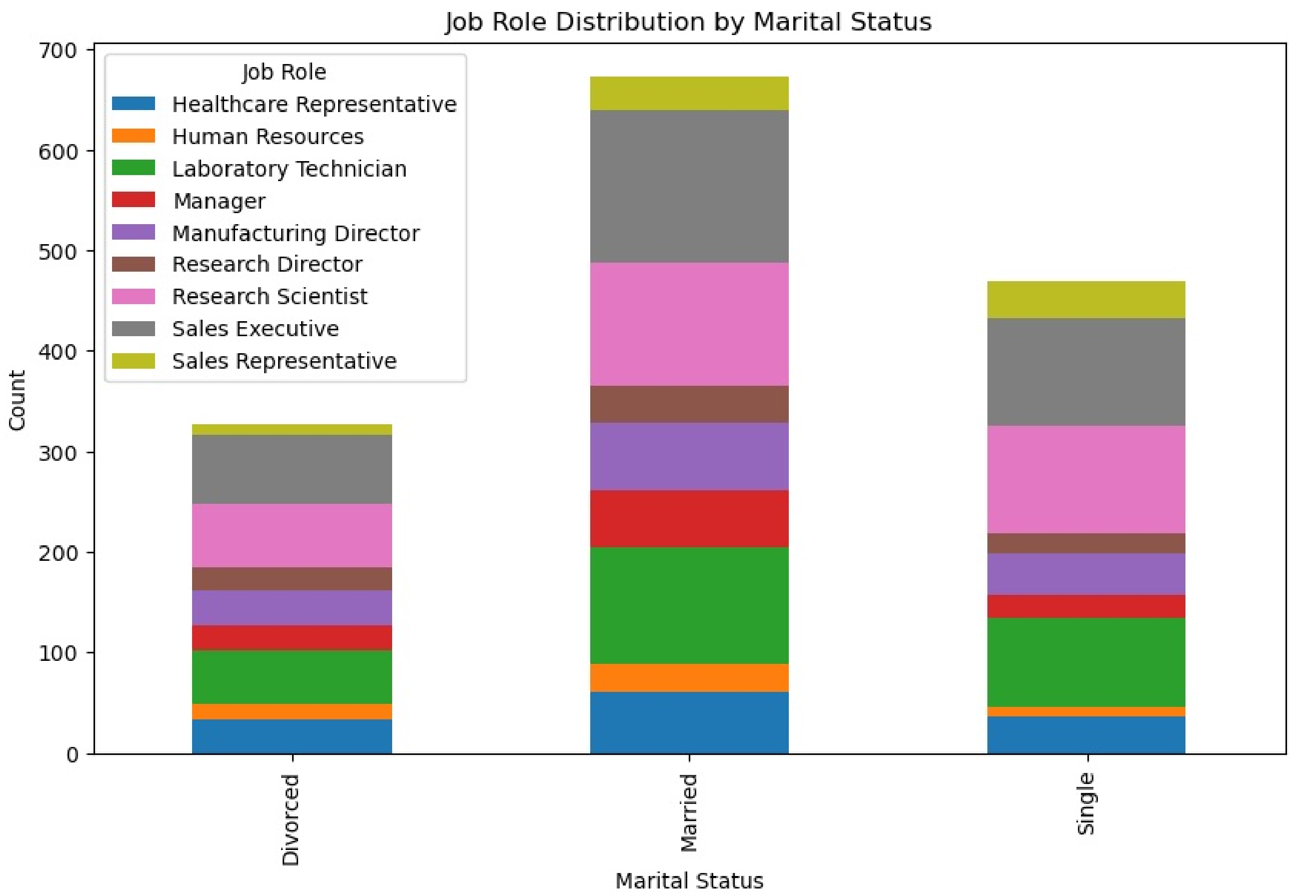
```
plt.figure(figsize=(10, 8))
sns.barplot(data=df, x='Job Role', y='Monthly Income', palette='Set2',)
plt.xlabel('Job Role')
plt.ylabel('Average Monthly Income')
plt.title('Average Monthly Income by Job Role')
plt.xticks(rotation=45, ha='right')
plt.show()
```



bar plot showing Distribution of job role by marital status

In [24]:

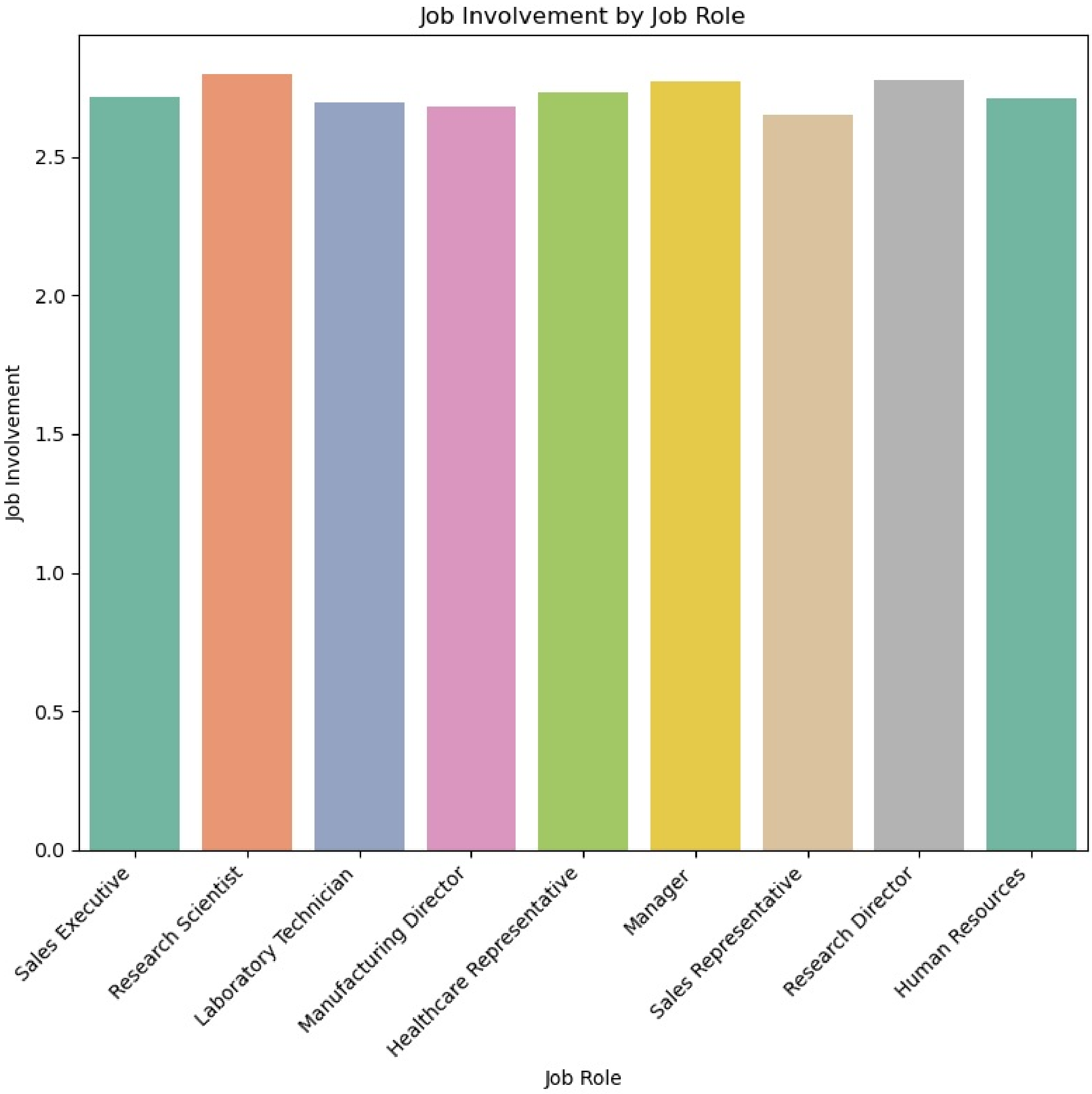
```
job_role_by_marital_status = df.groupby('Marital Status')['Job Role'].value_counts().unstack().fillna(0)
job_role_by_marital_status.plot(kind='bar', stacked=True, figsize=(10, 6))
plt.xlabel('Marital Status')
plt.ylabel('Count')
plt.title('Job Role Distribution by Marital Status')
plt.show()
```



Bar plot showing Job Involvement by Job role

In [25]:

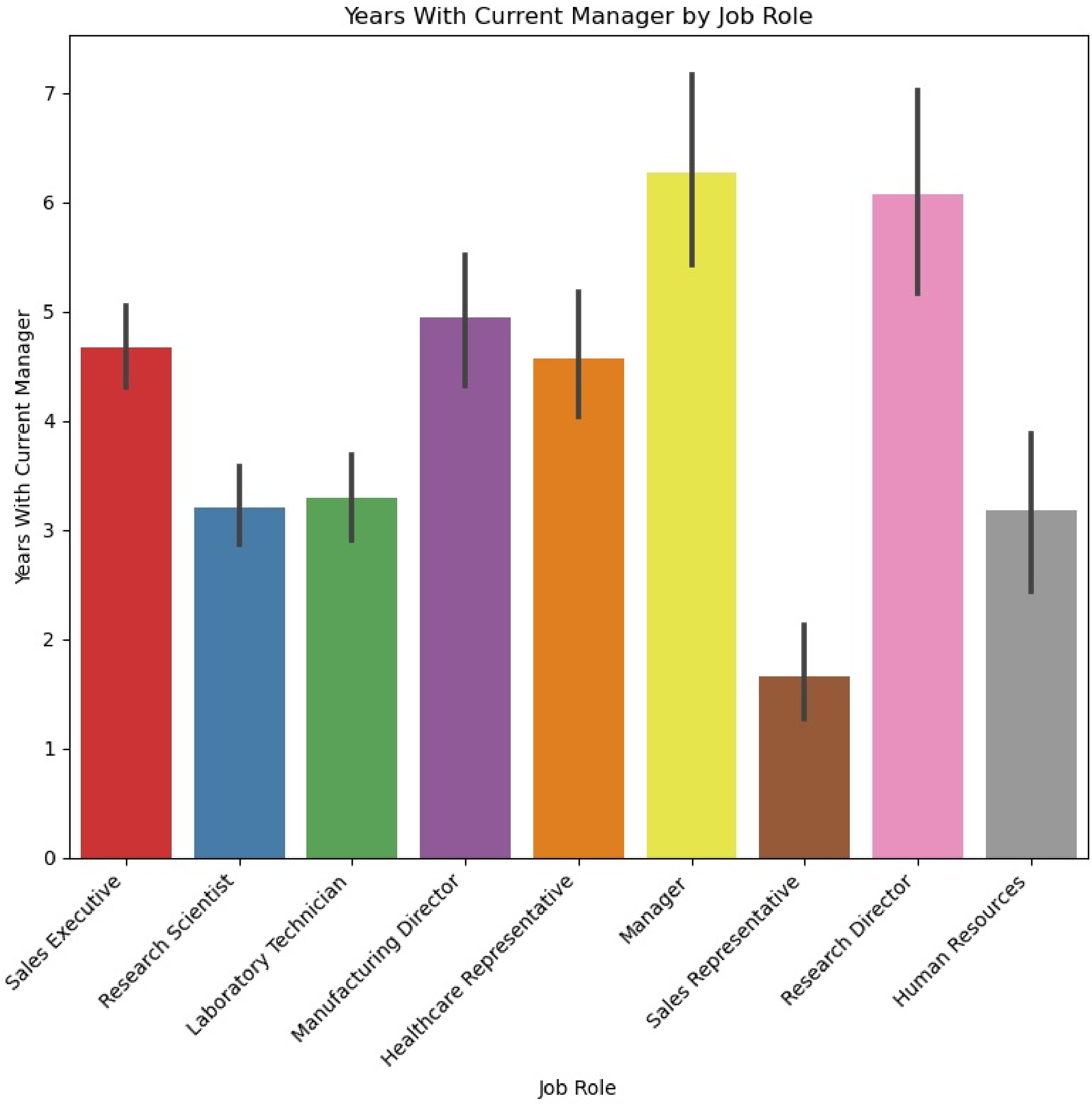
```
plt.figure(figsize=(8, 8))
sns.barplot(data=df, x='Job Role', y='Job Involvement', ci=None, palette='Set2')
plt.xlabel('Job Role')
plt.ylabel('Job Involvement')
plt.title('Job Involvement by Job Role')
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
plt.show()
```



Bar Plot showing Years with current manager by job role

In [26]:

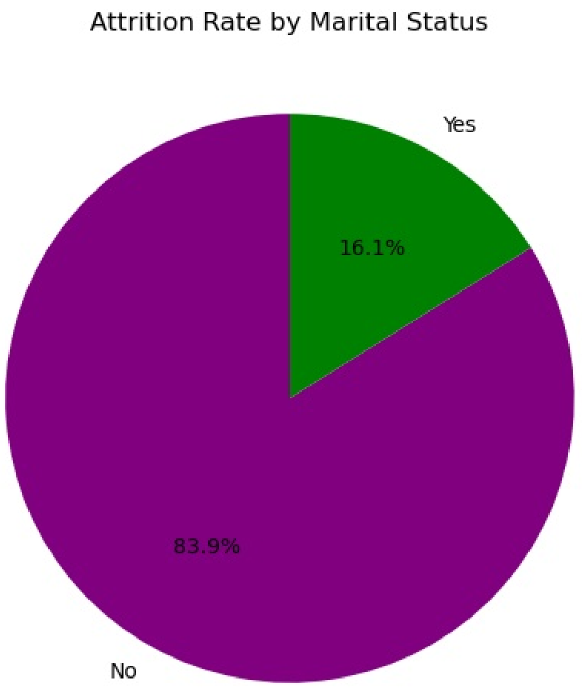
```
plt.figure(figsize=(8, 8))
sns.barplot(data=df, x='Job Role', y='Years With Curr Manager', palette='Set1')
plt.xlabel('Job Role')
plt.ylabel('Years With Current Manager')
plt.title('Years With Current Manager by Job Role')
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
plt.show()
```



Pie chart showing Attrition rate by marital Status

In [27]:

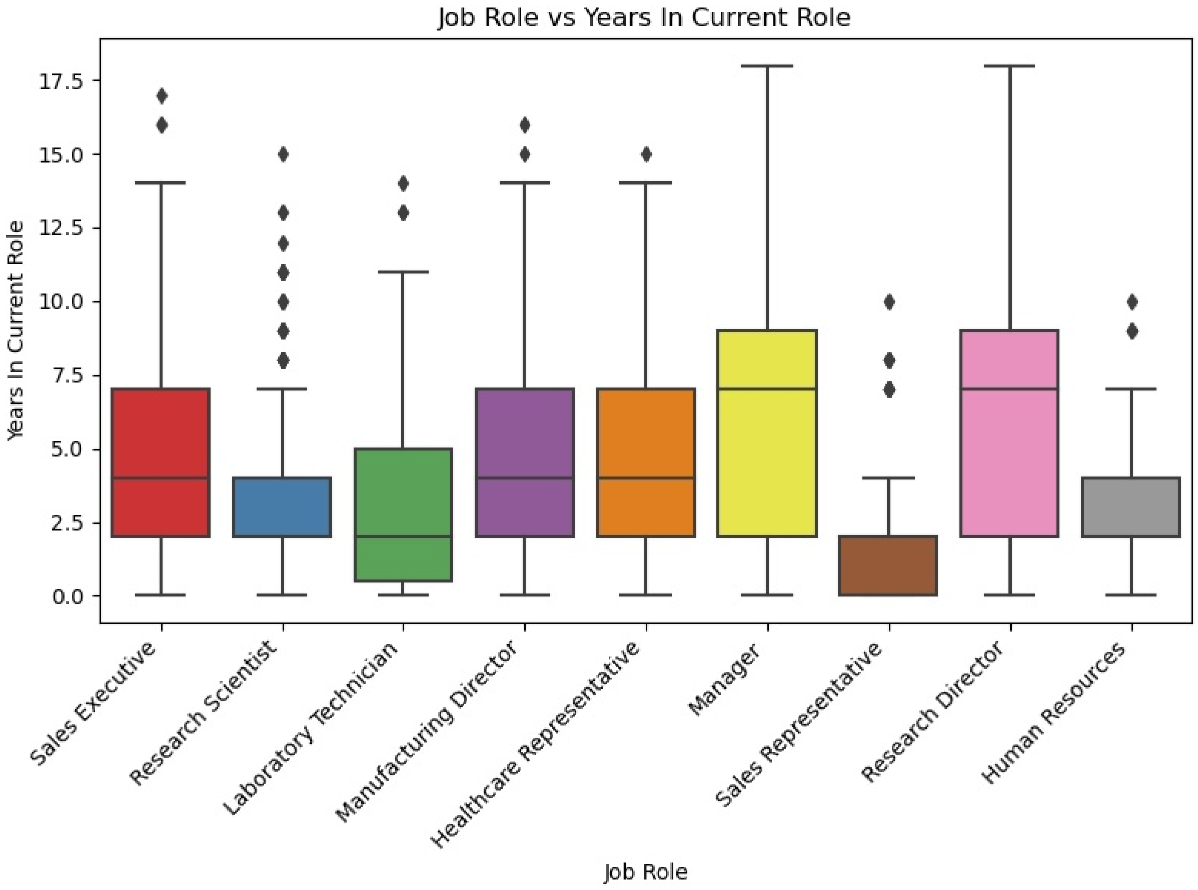
```
attrition_by_marital_status = df['Attrition'].value_counts()
plt.figure(figsize=(6, 6))
plt.pie(attrition_by_marital_status, labels=attrition_by_marital_status.index, autopct='%1.1f%%', startangle=90
plt.title('Attrition Rate by Marital Status')
plt.show()
```



Box plot showing Job role vs years in current role

In [28]:

```
plt.figure(figsize=(8, 6))
sns.boxplot(data=df, x='Job Role', y='Years In Current Role', palette='Set1')
plt.xlabel('Job Role')
plt.ylabel('Years In Current Role')
plt.title('Job Role vs Years In Current Role')
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
plt.show()
```

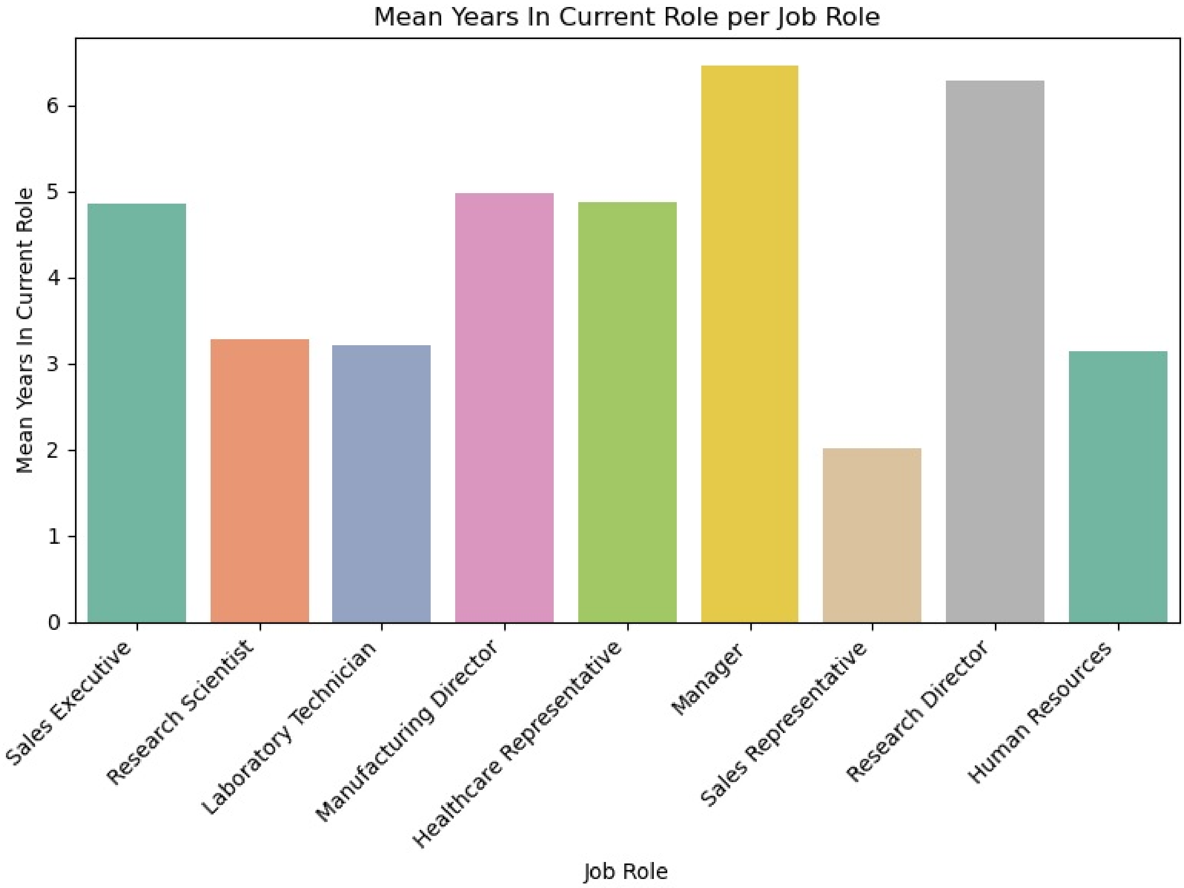


Bar plot showing mean years in currrent role per job role

In [29]:

```
plt.figure(figsize=(8, 6))
```

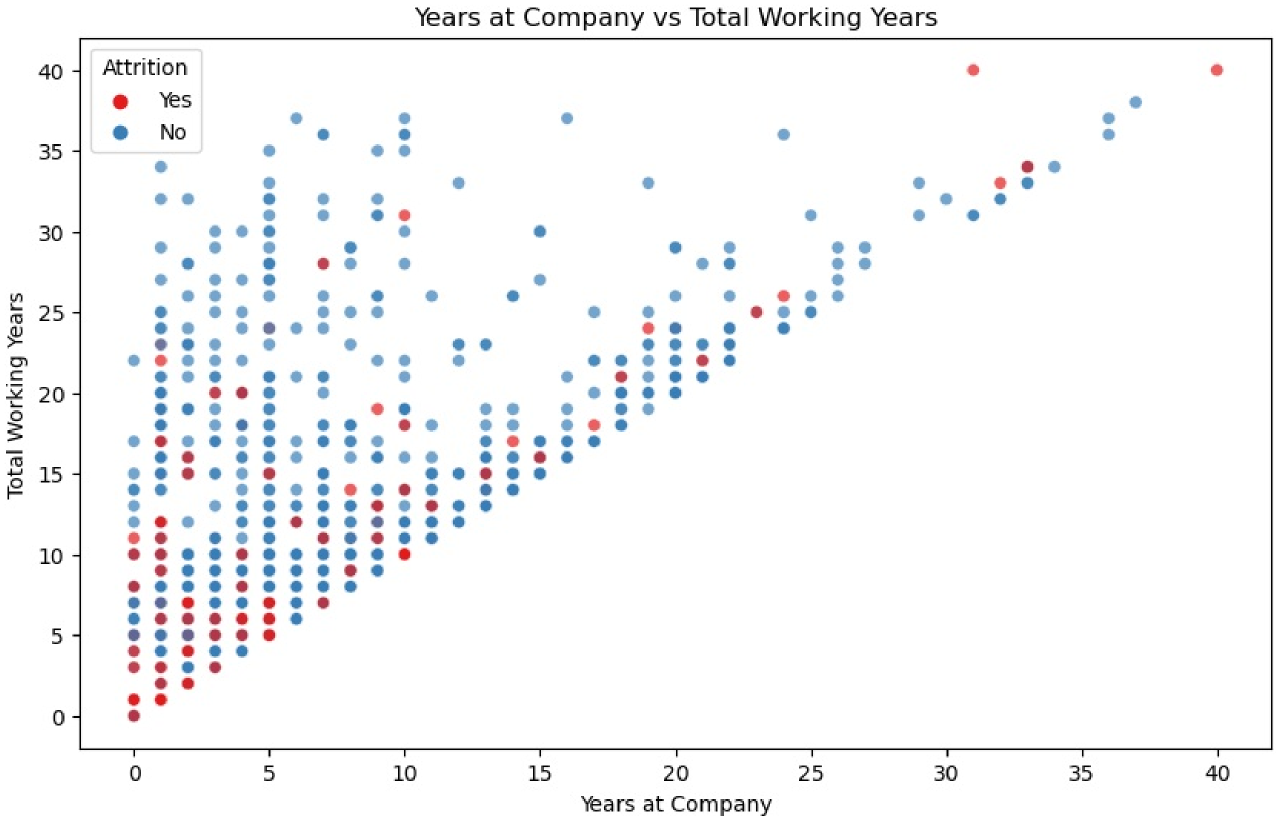
```
sns.barplot(data=df, x='Job Role', y='Years In Current Role', palette='Set2', ci=None)
plt.xlabel('Job Role')
plt.ylabel('Mean Years In Current Role')
plt.title('Mean Years In Current Role per Job Role')
plt.xticks(rotation=45, ha='right')
plt.tight_layout()
plt.show()
```



Scatter plot showing Total working years vs years at company

In [30]:

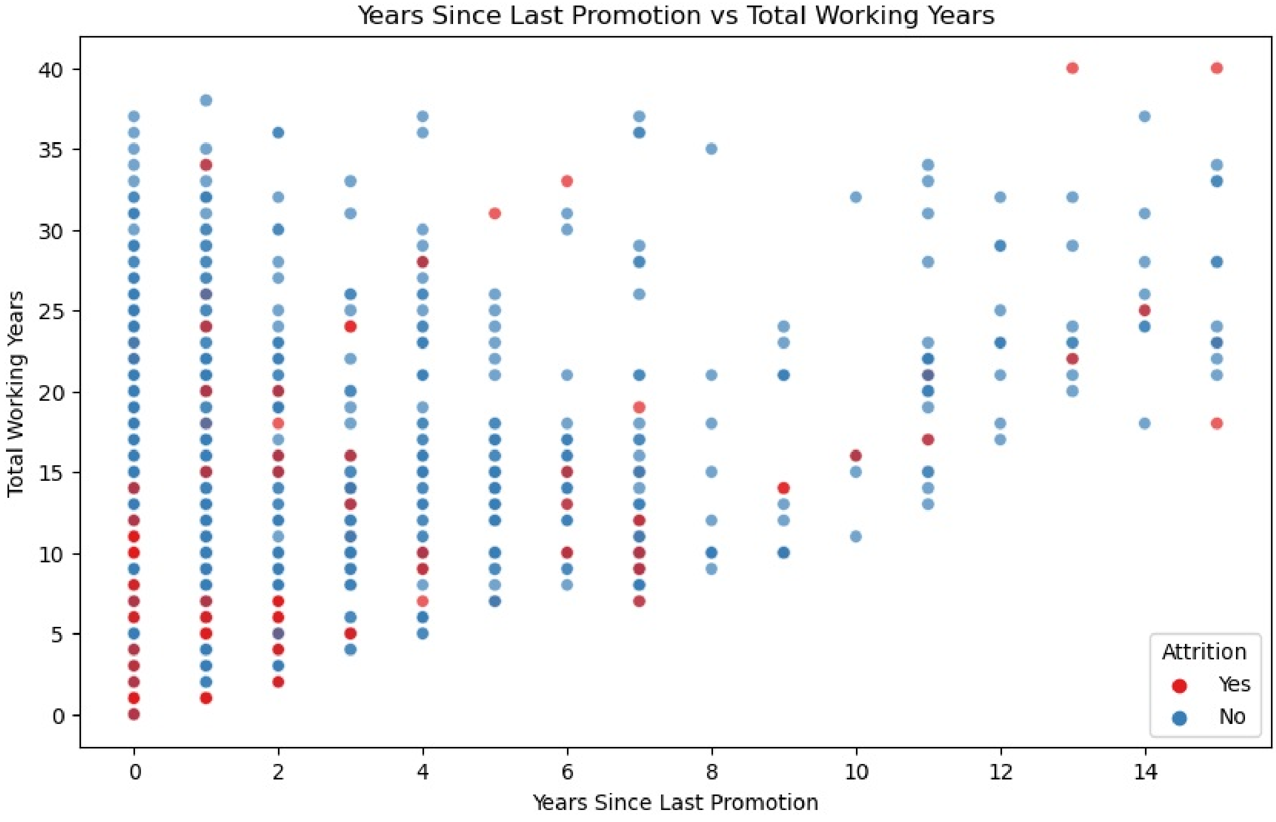
```
plt.figure(figsize=(10, 6))
sns.scatterplot(data=df, x='Years At Company', y='Total Working Years', hue='Attrition', palette='Set1', alpha=
plt.xlabel('Years at Company')
plt.ylabel('Total Working Years')
plt.title('Years at Company vs Total Working Years')
plt.show()
```



Scatter plot showing years since last promotion vs Total Working Years

In [31]:

```
plt.figure(figsize=(10, 6))
sns.scatterplot(data=df, x='Years Since Last Promotion', y='Total Working Years', hue='Attrition', palette='Set
plt.xlabel('Years Since Last Promotion')
plt.ylabel('Total Working Years')
plt.title('Years Since Last Promotion vs Total Working Years')
plt.show()
```



In [ ]:

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js