HR Analysis

Business Problem: Why employee left Our company ? OR why the rate of attriation become higher?

We have some steps before going to make a model to solve this Problem:

Importing Libraries

Data Collection

Data Cleaning

Explorate Data Analysis and try discovering some patterns

Solutions

Firstly we will import the libraries that i will use it to answer question.

Importing Libraries

In [150…
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import math as mt
import plotly.express as px
```

Data Collection

In [6]:
```
df=pd.read_csv("HR-Employee-Attrition.csv")
```

Data Cleaning

In [89]:
```
df.head(10)
```

Out[89]:

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Educatio
0	41	Yes	Travel_Rarely	1102	Sales	1	
1	49	No	Travel_Frequently	279	Research & Development	8	
2	37	Yes	Travel_Rarely	1373	Research & Development	2	
3	33	No	Travel_Frequently	1392	Research & Development	3	
4	27	No	Travel_Rarely	591	Research & Development	2	
5	32	No	Travel_Frequently	1005	Research & Development	2	
6	59	No	Travel_Rarely	1324	Research & Development	3	
7	30	No	Travel_Rarely	1358	Research & Development	24	
8	38	No	Travel_Frequently	216	Research & Development	23	
9	36	No	Travel_Rarely	1299	Research & Development	27	

10 rows × 32 columns

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

Out[8]:

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Educa
0	False	False	False	False	False	False	F
1	False	False	False	False	False	False	F
2	False	False	False	False	False	False	F
3	False	False	False	False	False	False	F
4	False	False	False	False	False	False	F
...	...	...	...	...	...	...	
1465	False	False	False	False	False	False	F
1466	False	False	False	False	False	False	F
1467	False	False	False	False	False	False	F
1468	False	False	False	False	False	False	F
1469	False	False	False	False	False	False	F

1470 rows × 35 columns

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

Out[9]:
```
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 [10]:
```
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 [23]:
```
df=df.drop(['Over18','EmployeeCount','StandardHours'],axis=1)
df.columns
```

Out[23]:
```
Index(['Age', 'Attrition', 'BusinessTravel', 'DailyRate', 'Department',
       'DistanceFromHome', 'Education', 'EducationField', 'EmployeeNumber',
       'EnvironmentSatisfaction', 'Gender', 'HourlyRate', 'JobInvolvement',
       'JobLevel', 'JobRole', 'JobSatisfaction', 'MaritalStatus',
       'MonthlyIncome', 'MonthlyRate', 'NumCompaniesWorked', 'OverTime',
       'PercentSalaryHike', 'PerformanceRating', 'RelationshipSatisfactio
n',
       'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear',
       'WorkLifeBalance', 'YearsAtCompany', 'YearsInCurrentRole',
       'YearsSinceLastPromotion', 'YearsWithCurrManager'],
      dtype='object')
```

In [38]:
```
df=df.drop_duplicates()
```

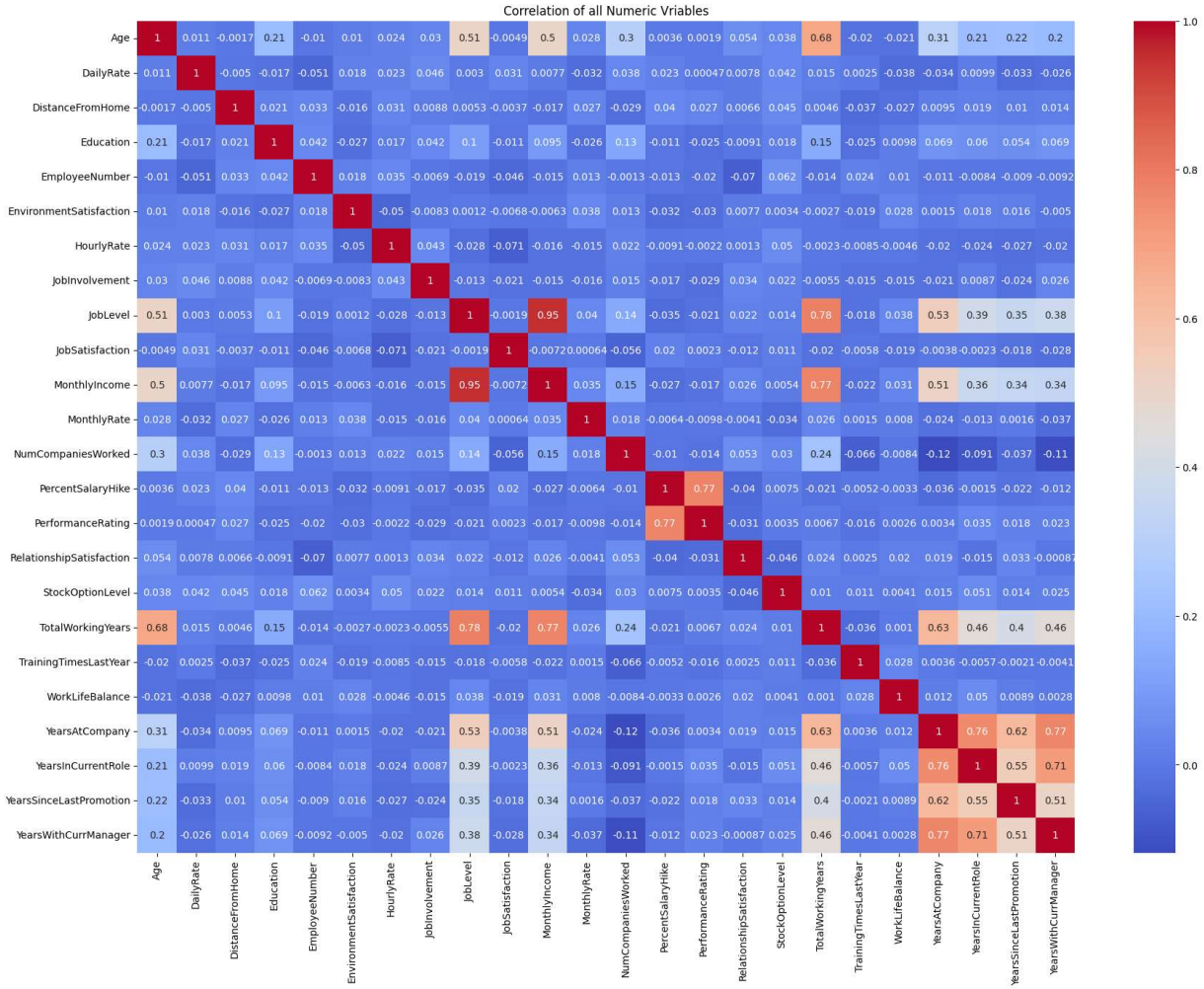
Exploratery Data Analysis

Now Let's Doing Explorate Data Analysis and try to finding solutions for Problems.

Q1) Find the correltion map between all numeric data?

In [186…
```
corr_df=df.corr()
plt.figure(figsize=(21,15))
sns.heatmap(corr_df,annot=True,cmap='coolwarm')
plt.title("Correlation of all Numeric Vriables")
plt.show();
```

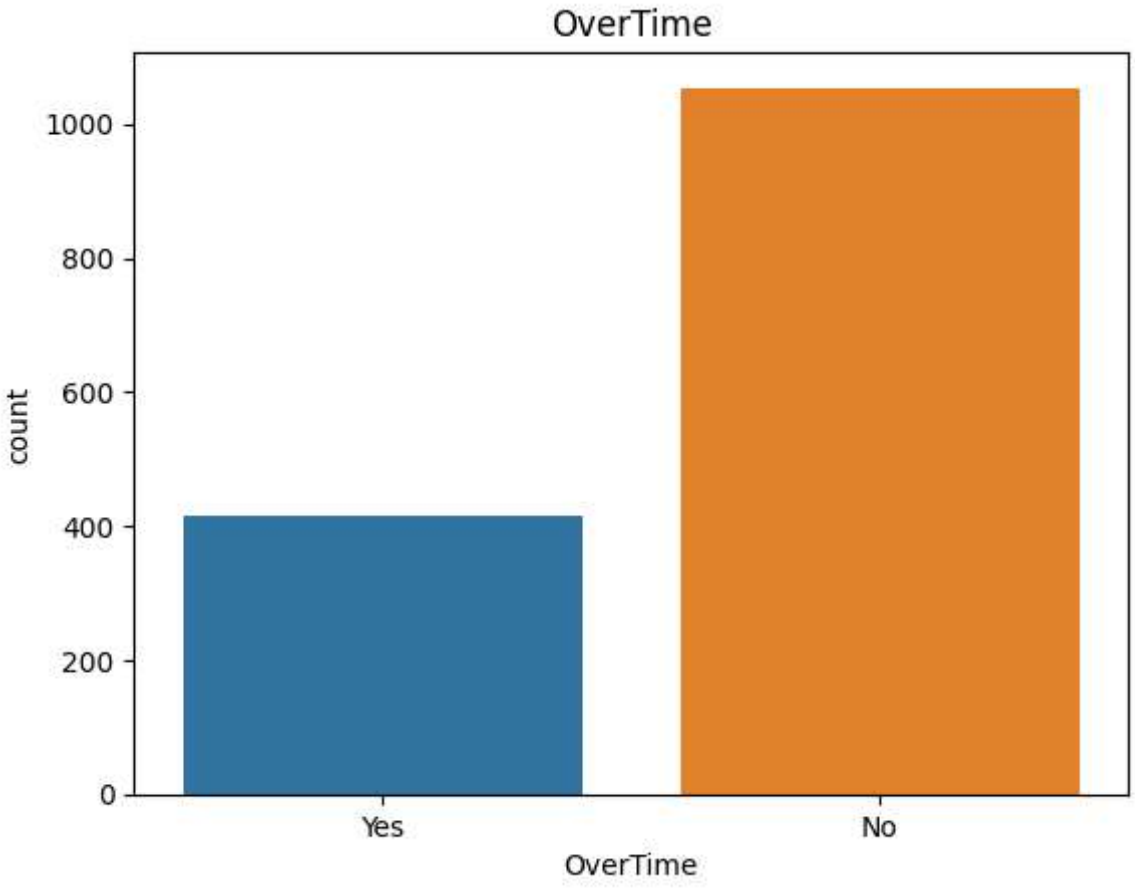
```
C:\Users\Saloni Singh\AppData\Local\Temp\ipykernel_6176\590578244.py:1: Futur
eWarning: The default value of numeric_only in DataFrame.corr is deprecated.
In a future version, it will default to False. Select only valid columns or s
pecify the value of numeric_only to silence this warning.
  corr_df=df.corr()
```



Q2) How many employee work overtime compared to those who do not?

In [69]:
```
sns.countplot(data=df, x='OverTime')
plt.title("OverTime")
```

Out[69]:
```
Text(0.5, 1.0, 'OverTime')
```



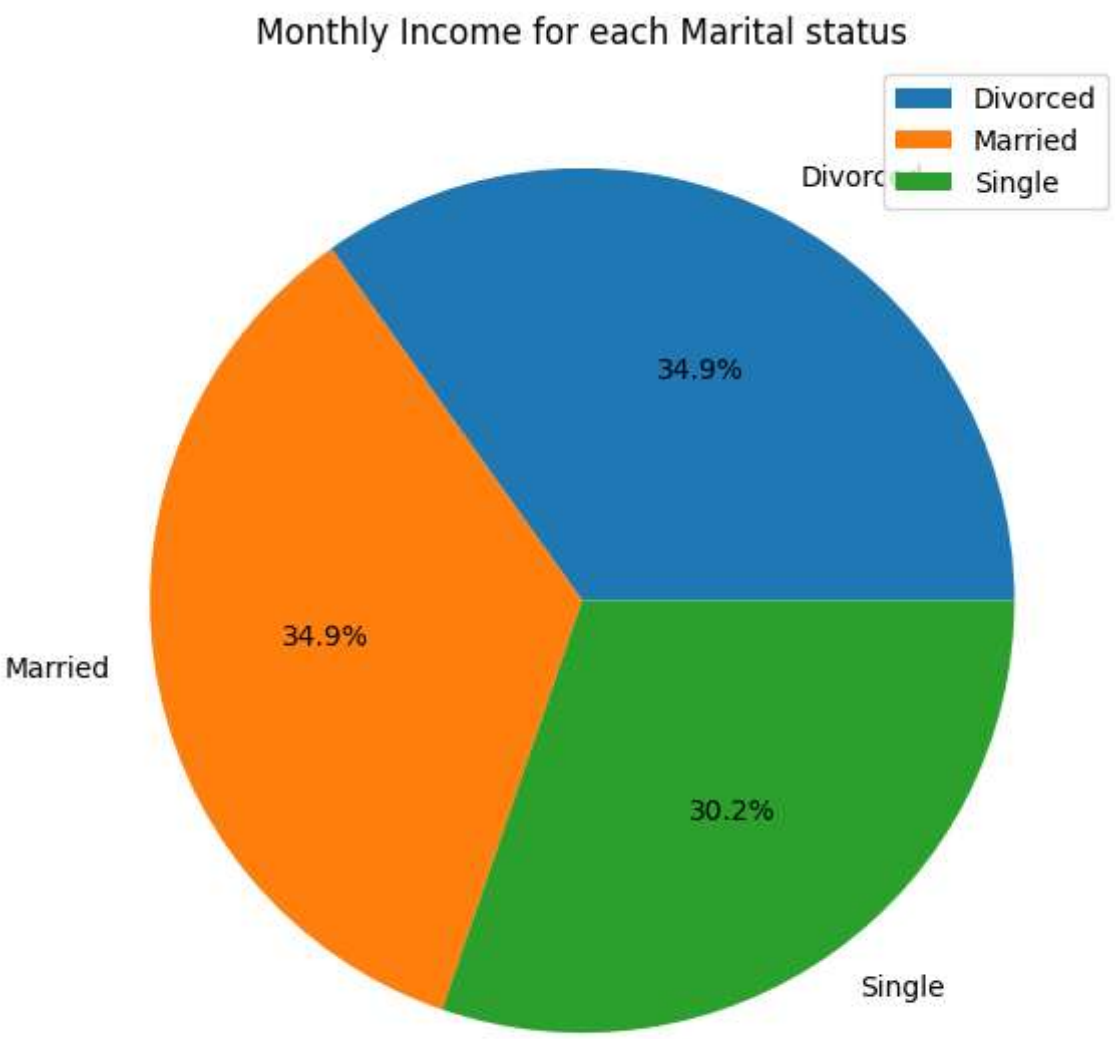
Q3) Do Married employee tend to have higher or lower salaries compared to their single or divorced countparts?

In [56]:
```
df.columns
```

Out[56]:
```
Index(['Age', 'Attrition', 'BusinessTravel', 'DailyRate', 'Department',
       'DistanceFromHome', 'Education', 'EducationField', 'EmployeeNumber',
       'EnvironmentSatisfaction', 'Gender', 'HourlyRate', 'JobInvolvement',
       'JobLevel', 'JobRole', 'JobSatisfaction', 'MaritalStatus',
       'MonthlyIncome', 'MonthlyRate', 'NumCompaniesWorked', 'OverTime',
       'PercentSalaryHike', 'PerformanceRating', 'RelationshipSatisfactio
n',
       'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear',
       'WorkLifeBalance', 'YearsAtCompany', 'YearsInCurrentRole',
       'YearsSinceLastPromotion', 'YearsWithCurrManager'],
      dtype='object')
```

In [98]:
```
q1=df.groupby('MaritalStatus')['MonthlyIncome'].mean()
MaritalStatus=q1.index.tolist()
MonthlyIncome=q1.tolist()
```

In [99]:
```
plt.figure(figsize=(10,7))
plt.pie(MonthlyIncome,labels=MaritalStatus,autopct='%1.1f%%')
plt.legend()
plt.title('Monthly Income for each Marital status');
```



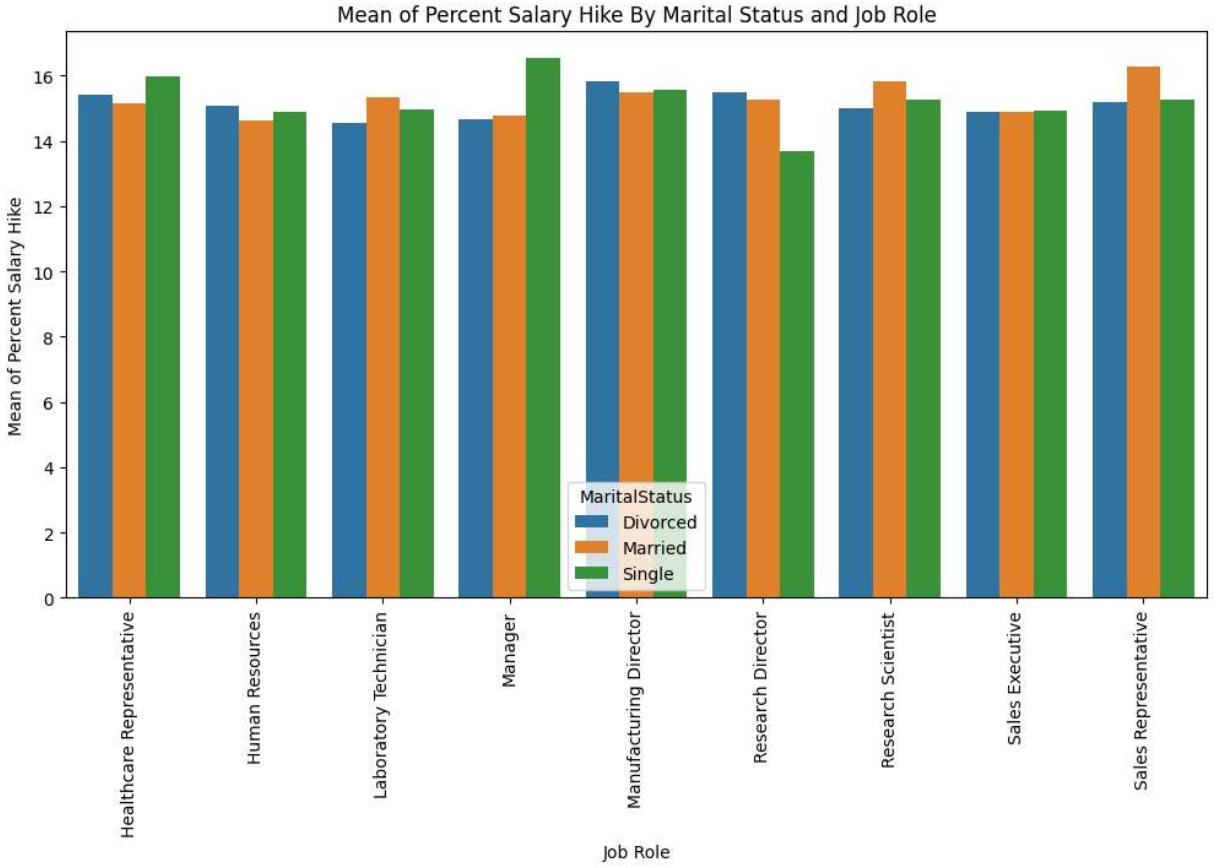
Q4) Does the percentage of salary hike differ for employee with job roles anf different marital statuses?

In [76]:
```
df.columns
```

Out[76]:
```
Index(['Age', 'Attrition', 'BusinessTravel', 'DailyRate', 'Department',
       'DistanceFromHome', 'Education', 'EducationField', 'EmployeeNumber',
       'EnvironmentSatisfaction', 'Gender', 'HourlyRate', 'JobInvolvement',
       'JobLevel', 'JobRole', 'JobSatisfaction', 'MaritalStatus',
       'MonthlyIncome', 'MonthlyRate', 'NumCompaniesWorked', 'OverTime',
       'PercentSalaryHike', 'PerformanceRating', 'RelationshipSatisfactio
n',
       'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear',
       'WorkLifeBalance', 'YearsAtCompany', 'YearsInCurrentRole',
       'YearsSinceLastPromotion', 'YearsWithCurrManager'],
      dtype='object')
```

In [127…
```
#tSalaryGroup the data by MaritalStatus and Jobrole and calculate the sum of
q2=df.groupby(['MaritalStatus','JobRole'])['PercentSalaryHike'].mean().reset
```

```
#Create a bar plot using Seaborn
plt.figure(figsize=(12,6))
sns.barplot(data=q2,x='JobRole',y='PercentSalaryHike',hue='MaritalStatus')
plt.xticks(rotation=90)
plt.xlabel('Job Role')
plt.ylabel('Mean of Percent Salary Hike')
plt.title("Mean of Percent Salary Hike By Marital Status and Job Role")
plt.show();
```



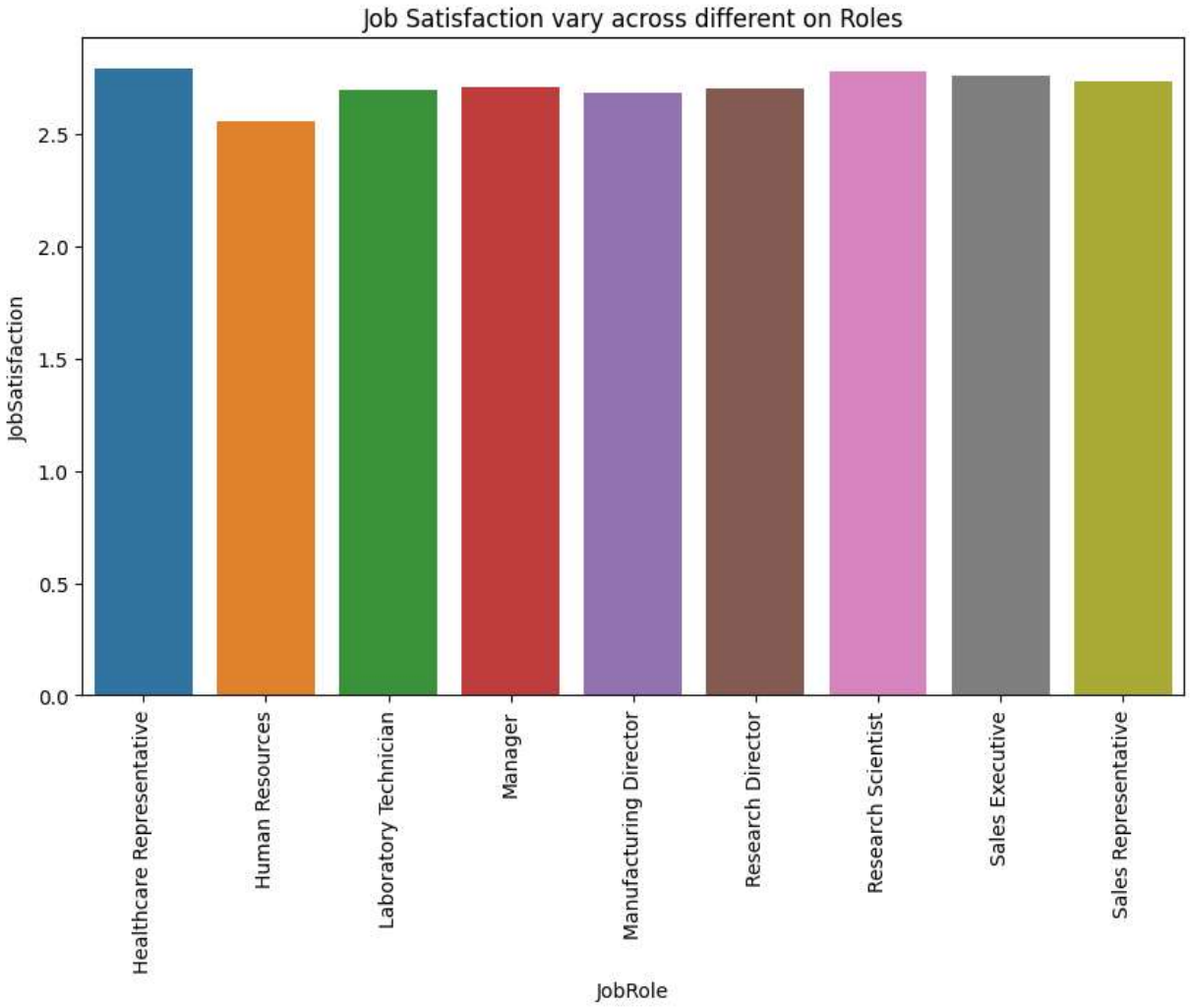
Q5)How does job statisfaction vary across different job roles?

In [101…
```
df['JobRole'].value_counts()
```

Out[101]:
```
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: JobRole, dtype: int64
```

In [124…
```
q3=df.groupby('JobRole')['JobSatisfaction'].mean().reset_index()
plt.figure(figsize=(10,6))
sns.barplot(data=q3,x='JobRole',y='JobSatisfaction')
plt.xticks(rotation=90)
```

```
plt.title("Job Satisfaction vary across different on Roles")
plt.show()
```



Q6)Are there are any relation ship between age and monthly income?

In [129…
```
df.columns
```

Out[129]:
```
Index(['Age', 'Attrition', 'BusinessTravel', 'DailyRate', 'Department',
       'DistanceFromHome', 'Education', 'EducationField', 'EmployeeNumbe
r',
       'EnvironmentSatisfaction', 'Gender', 'HourlyRate', 'JobInvolvemen
t',
       'JobLevel', 'JobRole', 'JobSatisfaction', 'MaritalStatus',
       'MonthlyIncome', 'MonthlyRate', 'NumCompaniesWorked', 'OverTime',
       'PercentSalaryHike', 'PerformanceRating', 'RelationshipSatisfactio
n',
       'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear',
       'WorkLifeBalance', 'YearsAtCompany', 'YearsInCurrentRole',
       'YearsSinceLastPromotion', 'YearsWithCurrManager'],
      dtype='object')
```

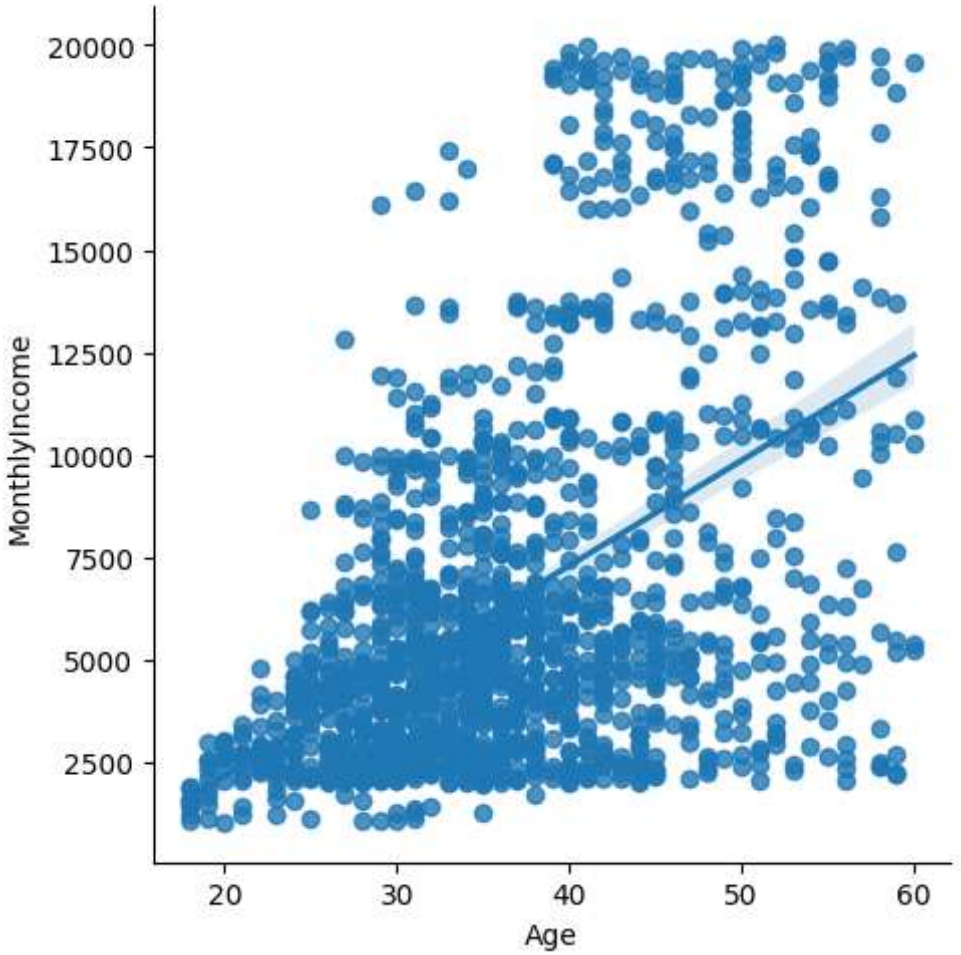
In [135…
```
q4_corr=df[['Age','MonthlyIncome']].corr()
q4_corr
```

Out[135]:

	Age	MonthlyIncome
Age	1.000000	0.497855
MonthlyIncome	0.497855	1.000000

In [140…
```
plt.figure(figsize=(12,8))
sns.lmplot(data=df,x='Age',y='MonthlyIncome')
plt.show();
```

```
C:\python 3.11\Lib\site-packages\seaborn\axisgrid.py:118: UserWarning: The fi
gure layout has changed to tight
  self._figure.tight_layout(*args, **kwargs)
<Figure size 1200x800 with 0 Axes>
```

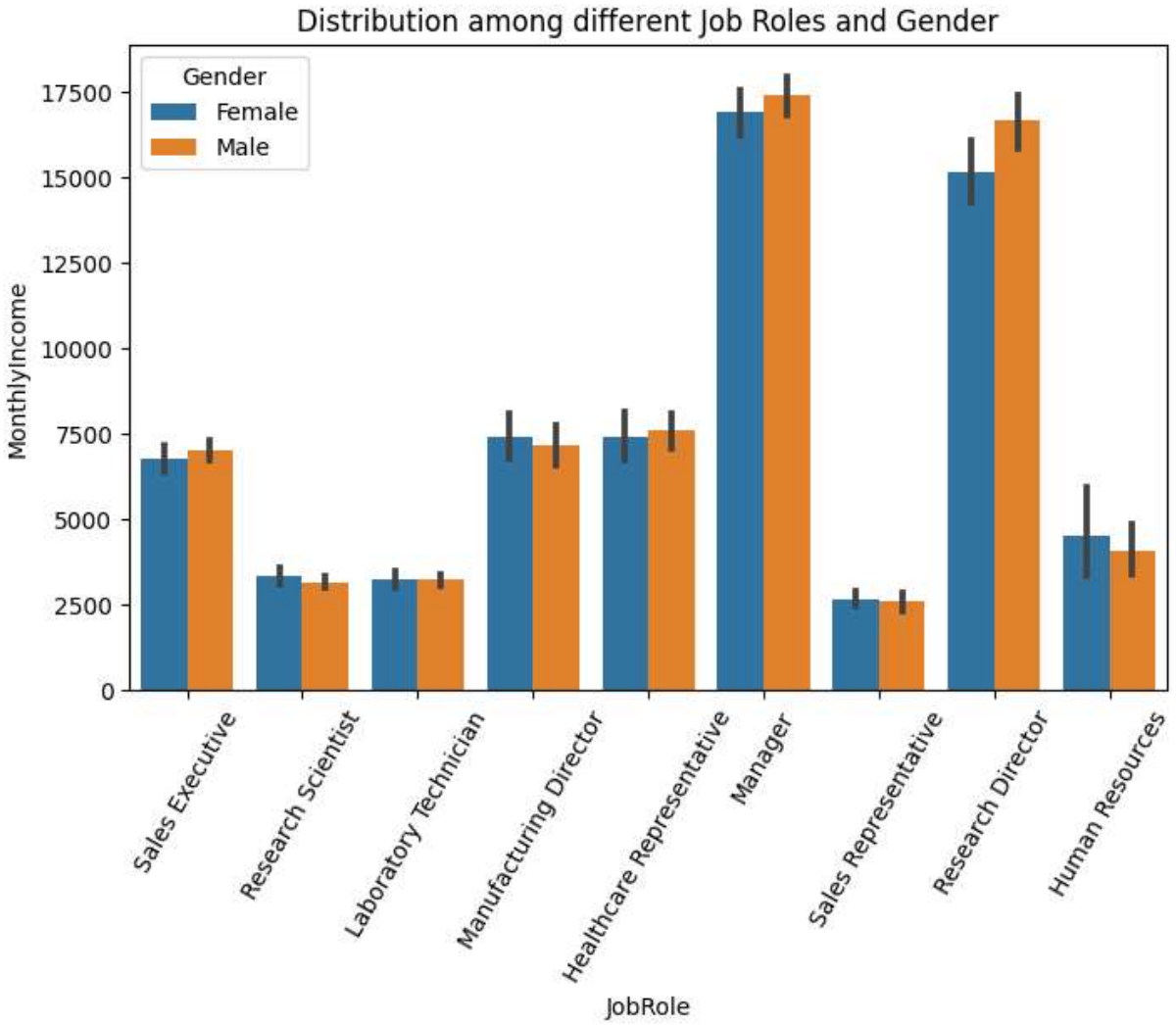


Q7) What is the Salary distribution among different job roles and diffrent Gender?

In [141…
```
df.columns
```

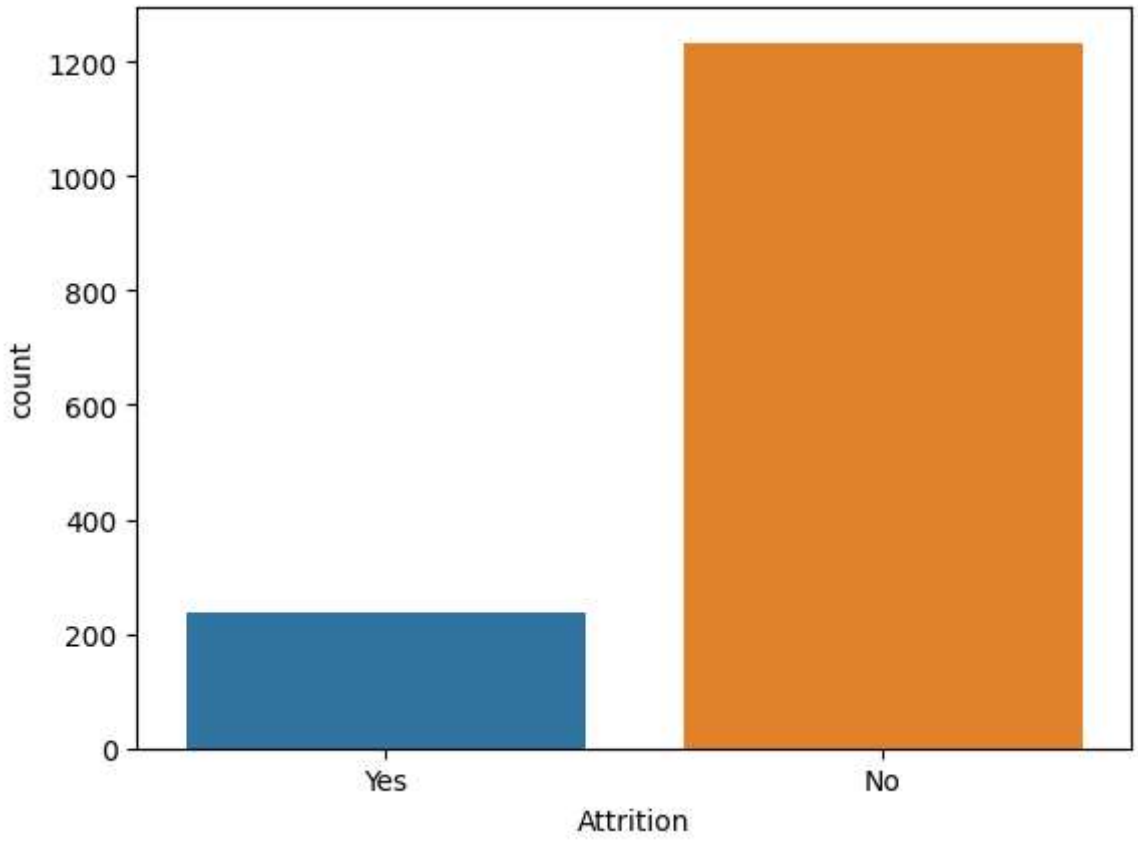
Out[141]:
```
Index(['Age', 'Attrition', 'BusinessTravel', 'DailyRate', 'Department',
       'DistanceFromHome', 'Education', 'EducationField', 'EmployeeNumbe
r',
       'EnvironmentSatisfaction', 'Gender', 'HourlyRate', 'JobInvolvemen
t',
       'JobLevel', 'JobRole', 'JobSatisfaction', 'MaritalStatus',
       'MonthlyIncome', 'MonthlyRate', 'NumCompaniesWorked', 'OverTime',
       'PercentSalaryHike', 'PerformanceRating', 'RelationshipSatisfactio
n',
       'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear',
       'WorkLifeBalance', 'YearsAtCompany', 'YearsInCurrentRole',
       'YearsSinceLastPromotion', 'YearsWithCurrManager'],
      dtype='object')
```

In [146…
```
plt.figure(figsize=(8,5))
sns.barplot(data=df,x='JobRole',y='MonthlyIncome',hue='Gender')
plt.title("Distribution among different Job Roles and Gender")
plt.xticks(rotation=60)
plt.show();
```



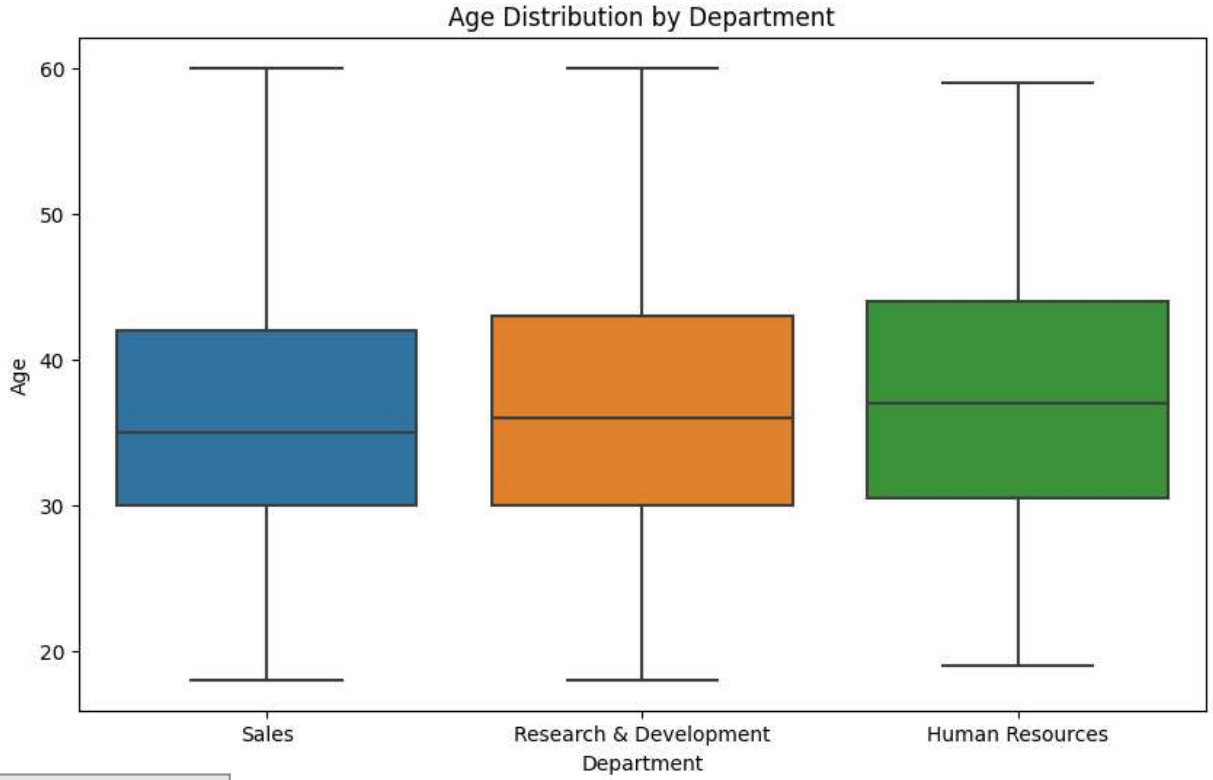
In [147…
```
sns.countplot(x=df['Attrition'])
```

Out[147]:
```
<Axes: xlabel='Attrition', ylabel='count'>
```

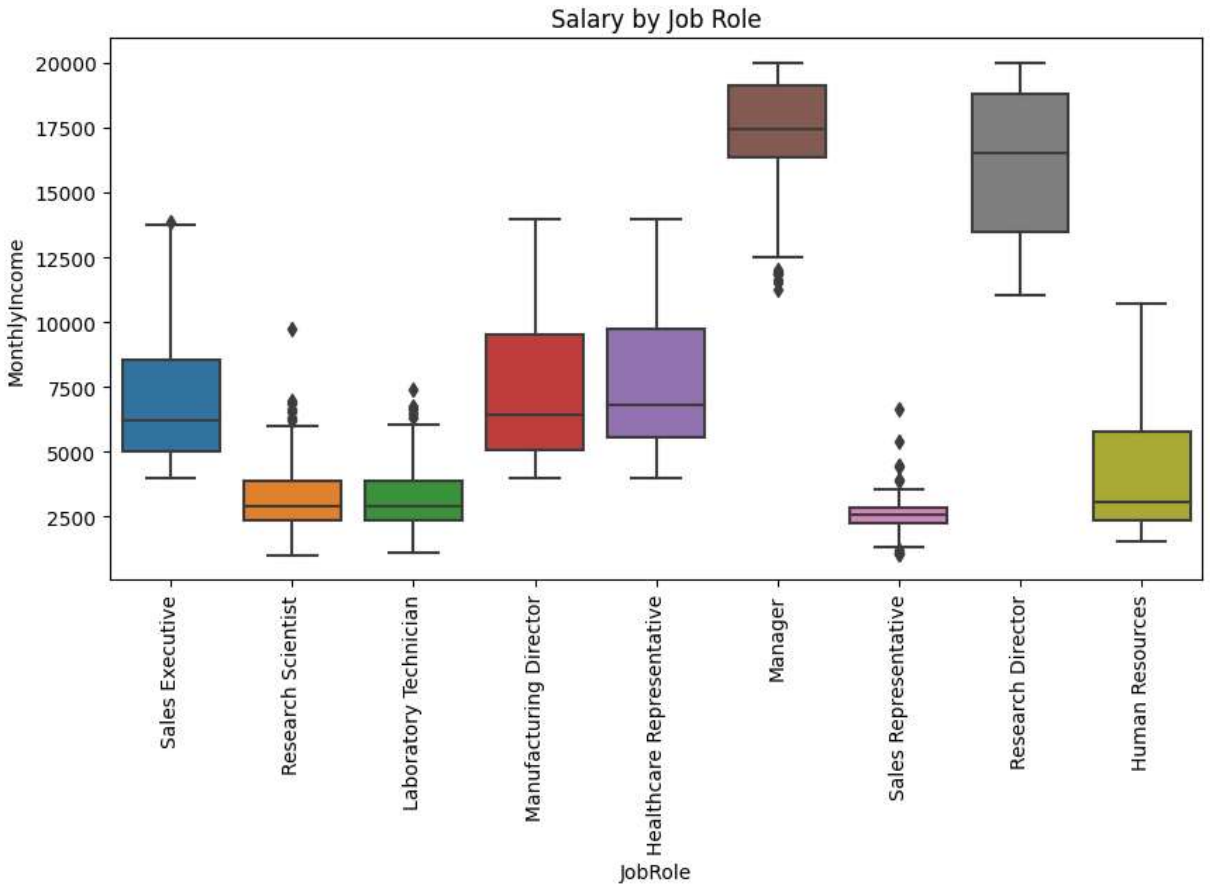


In [158…

```
#Age Distribution by Department
plt.figure(figsize=(10,6))
sns.boxplot(data=df,x='Department',y='Age')
plt.title("Age Distribution by Department")
plt.show();
```

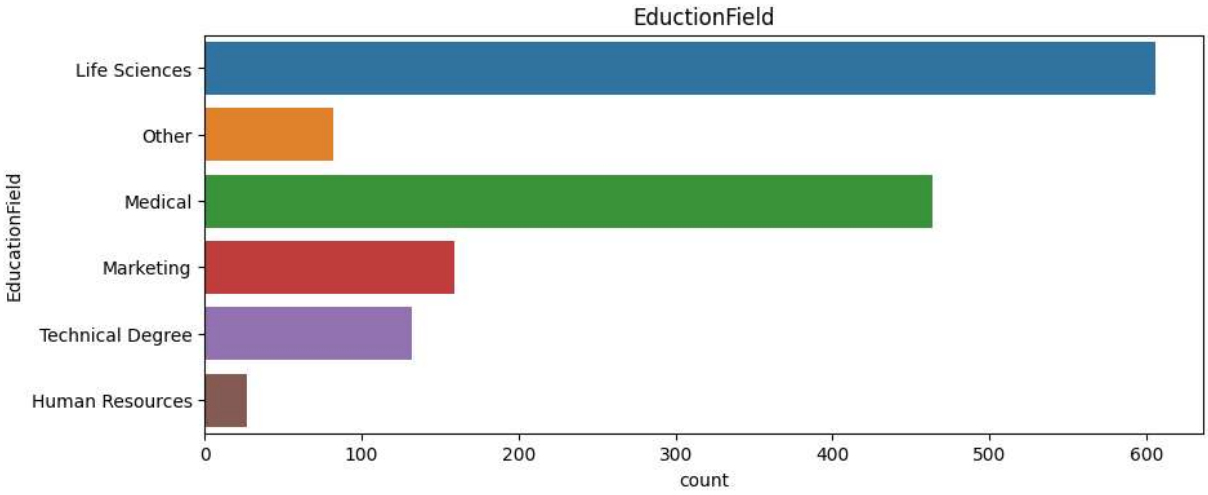


In [171…
```
#Salary By Job Role
plt.figure(figsize=(10,5))
sns.boxplot(data=df,x='JobRole',y='MonthlyIncome')
plt.xticks(rotation=90)
plt.title("Salary by Job Role")
plt.show();
```



In [188…
```
#Education Field
plt.figure(figsize=(10,4))
sns.countplot(data=df,y='EducationField')
plt.title("EductionField")
```

Out[188]: Text(0.5, 1.0, 'EductionField')



Thankyou