In [33]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [35]:
```
df = pd.read_csv("C:/Users/noic3/OneDrive/Desktop/WA_Fn-UseC_-HR-Employee-Attrition.csv")
```

In [37]:
```
df.head()
```

Out[37]:

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeNur
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	

5 rows × 35 columns

In [41]:
```
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 [13]:

```
df.tail()
```

Out[13]:

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

5 rows × 35 columns

In [53]:

```
df.isnull().sum()
```

```
Out[53]: 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 [43]: df.duplicated().sum()
```

```
Out[43]: 0
```

In [45]:

```python
df.describe()
```

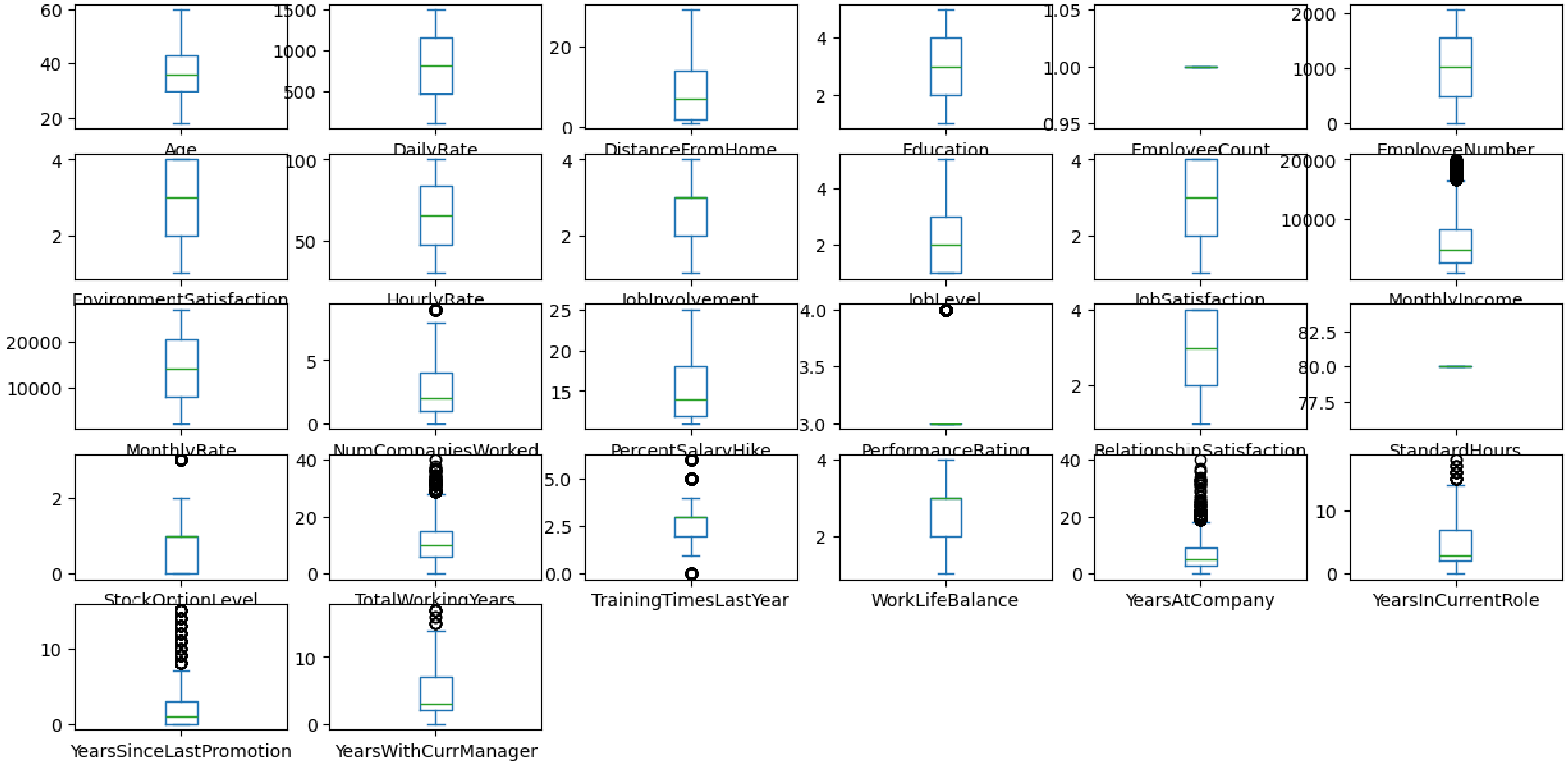
Out[45]:

	Age	DailyRate	DistanceFromHome	Education	EmployeeCount	EmployeeNumber	EnvironmentSatisfaction	HourlyRat
count	1470.000000	1470.000000	1470.000000	1470.000000	1470.0	1470.000000	1470.000000	1470.00000
mean	36.923810	802.485714	9.192517	2.912925	1.0	1024.865306	2.721769	65.89115
std	9.135373	403.509100	8.106864	1.024165	0.0	602.024335	1.093082	20.32942
min	18.000000	102.000000	1.000000	1.000000	1.0	1.000000	1.000000	30.00000
25%	30.000000	465.000000	2.000000	2.000000	1.0	491.250000	2.000000	48.00000
50%	36.000000	802.000000	7.000000	3.000000	1.0	1020.500000	3.000000	66.00000
75%	43.000000	1157.000000	14.000000	4.000000	1.0	1555.750000	4.000000	83.75000
max	60.000000	1499.000000	29.000000	5.000000	1.0	2068.000000	4.000000	100.00000

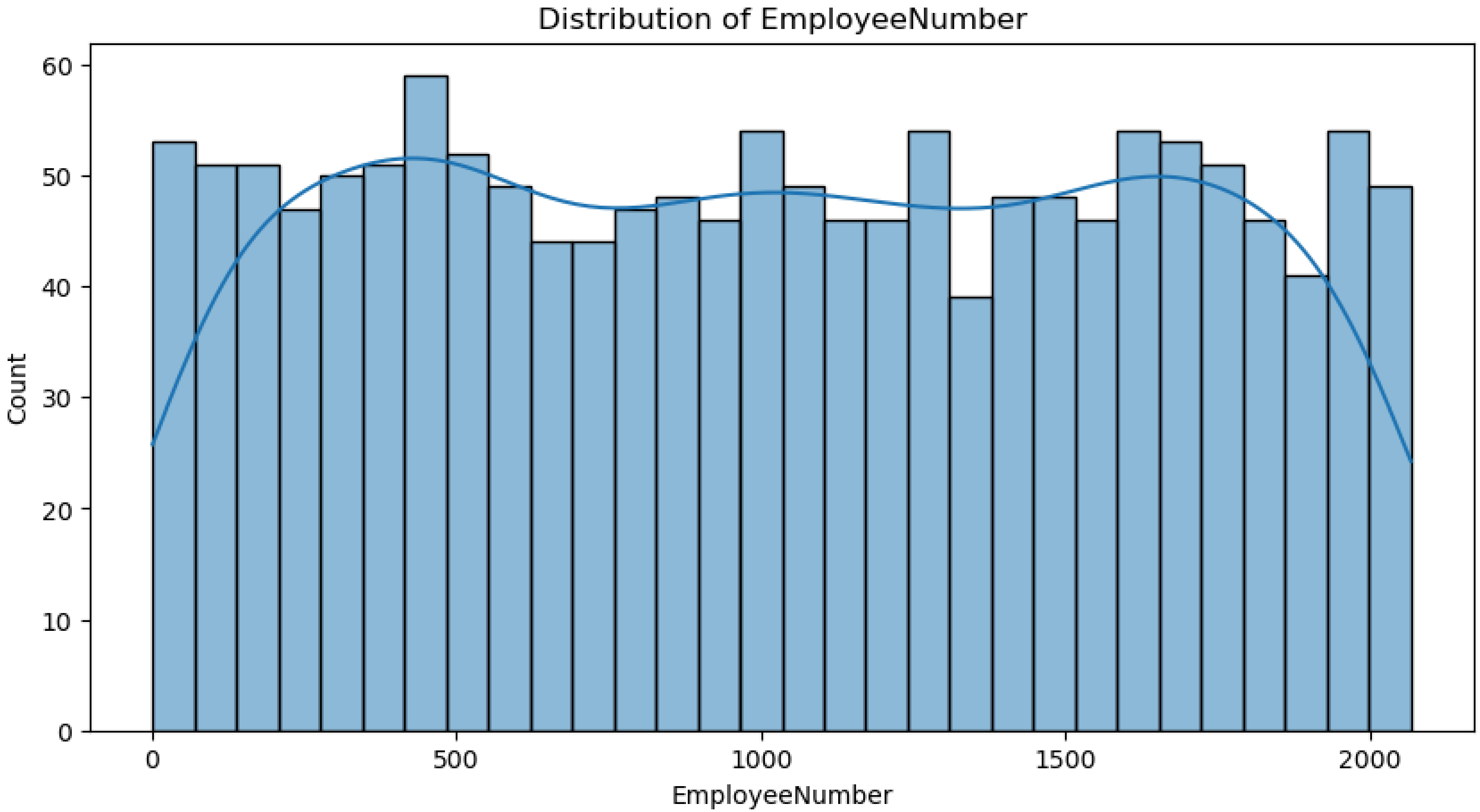
8 rows × 26 columns

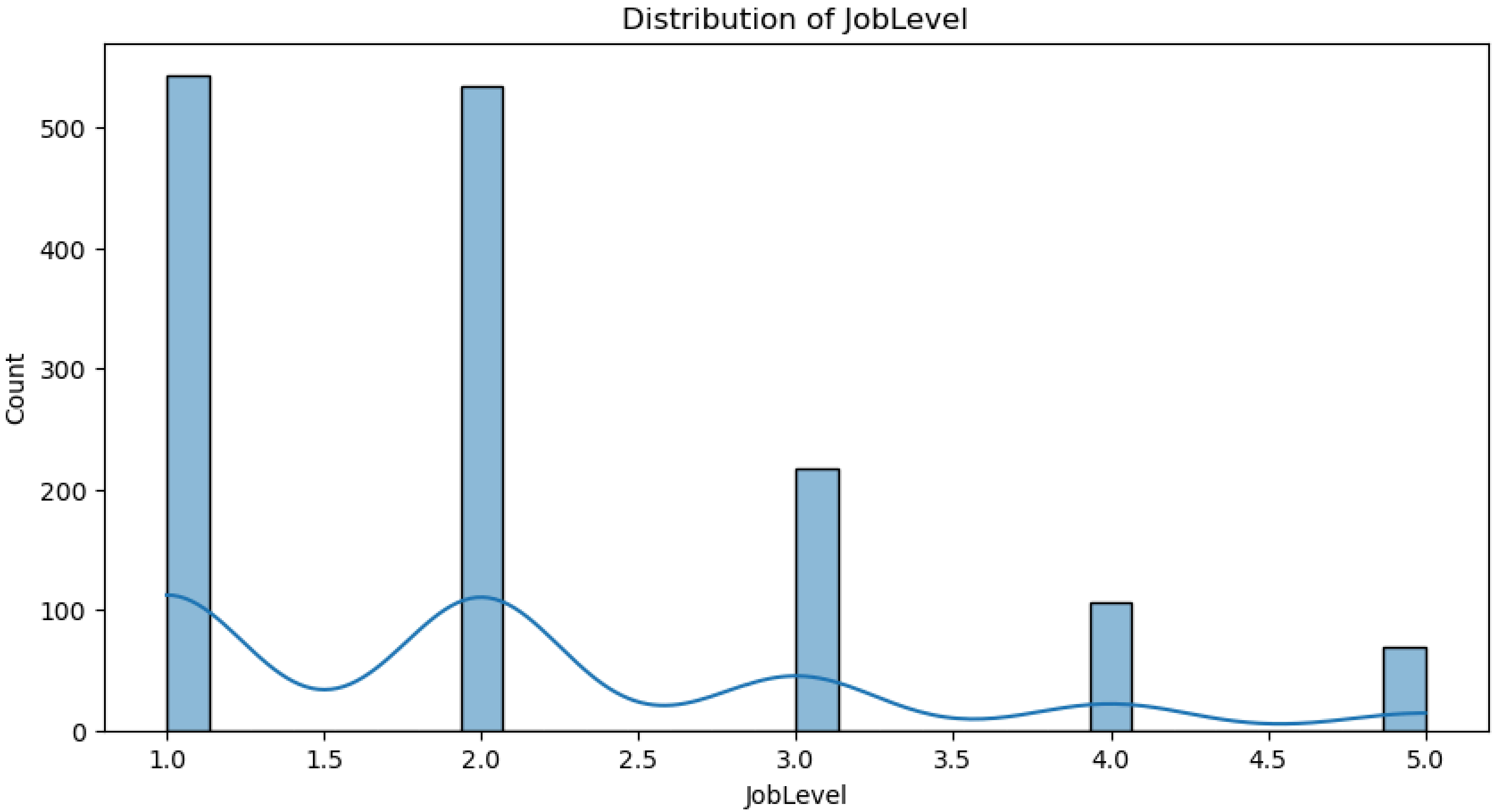
In [47]:

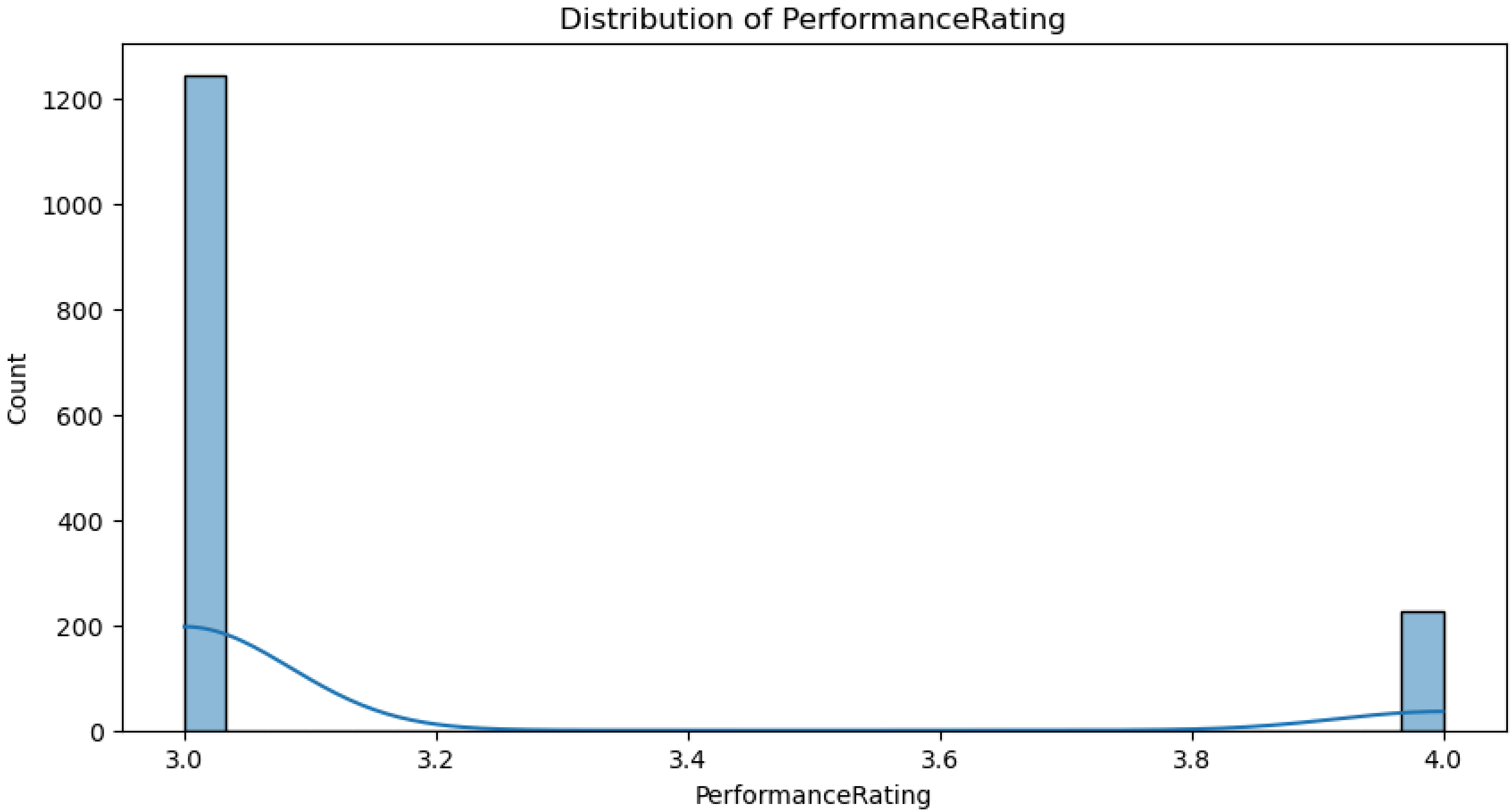
```python
# Boxplot to check for outliers
df.plot(kind='box', subplots=True, layout=(10,6), figsize=(15,15))
plt.show()
```

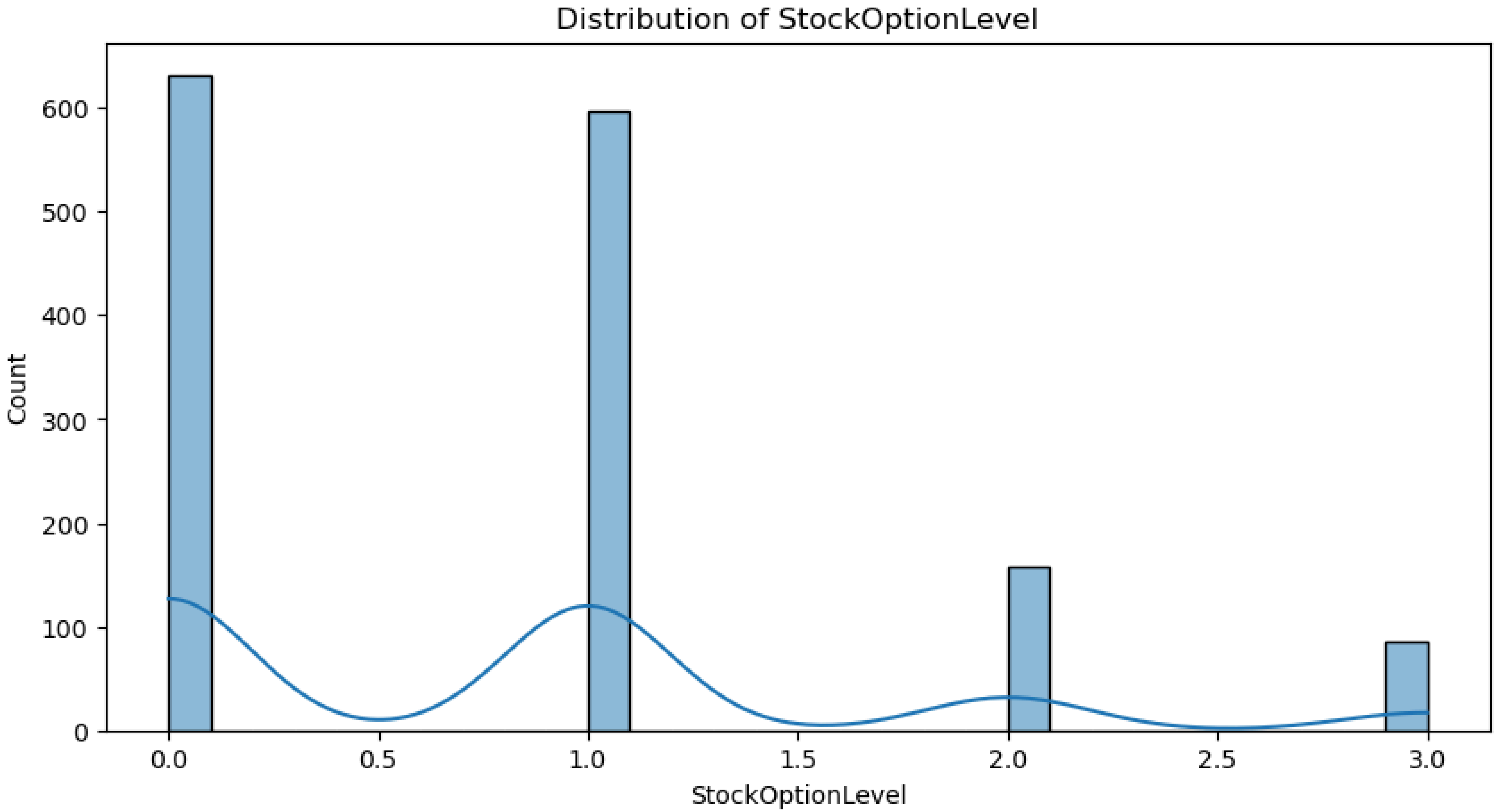


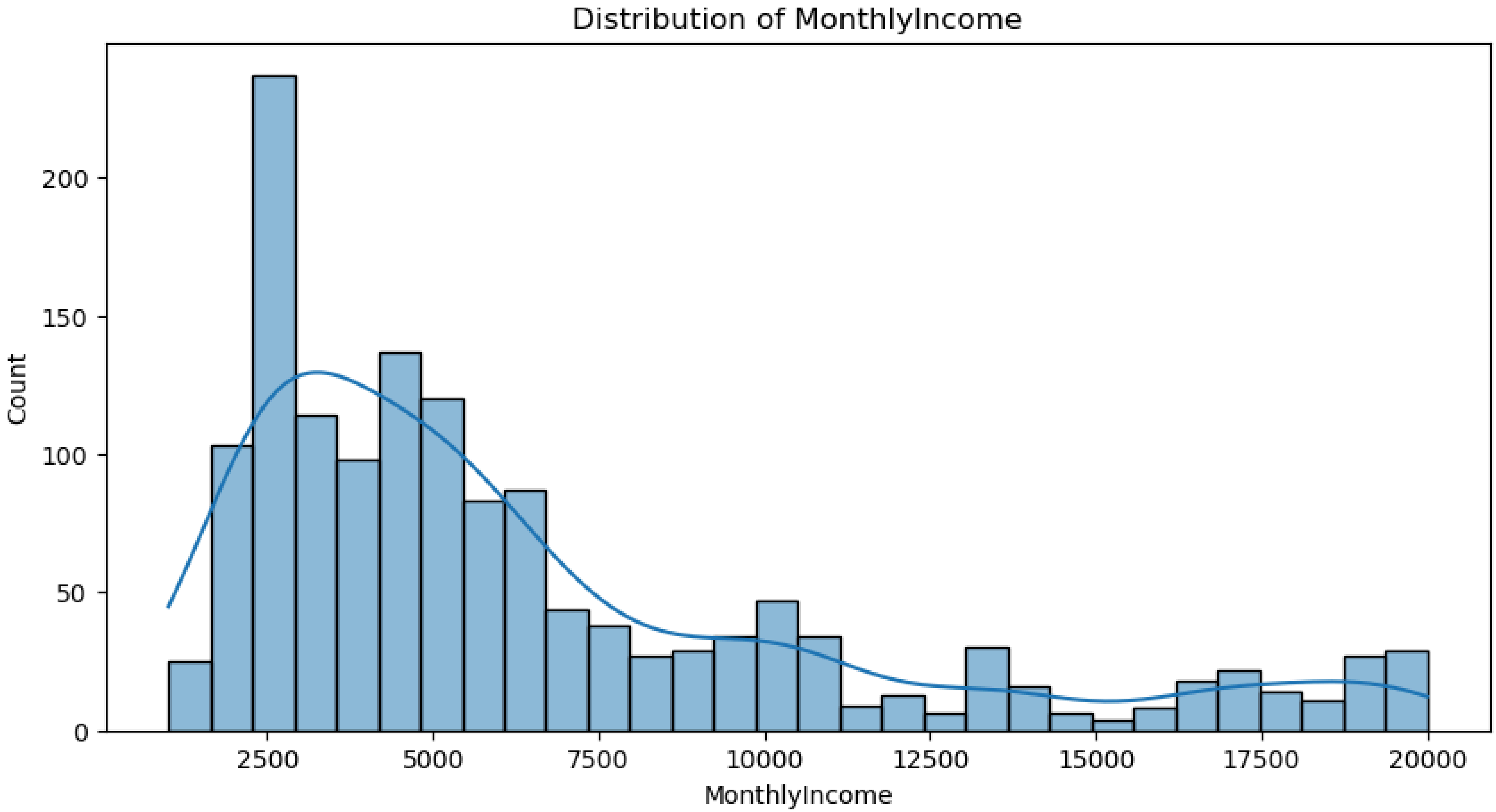
In [49]:
```
numerical_columns = list(set(df.describe().columns.to_list()))
for col in numerical_columns:
    plt.figure(figsize=(10, 5))
    sns.histplot(df[col], bins=30, kde=True)
    plt.title(f'Distribution of {col}')
    plt.show()
```

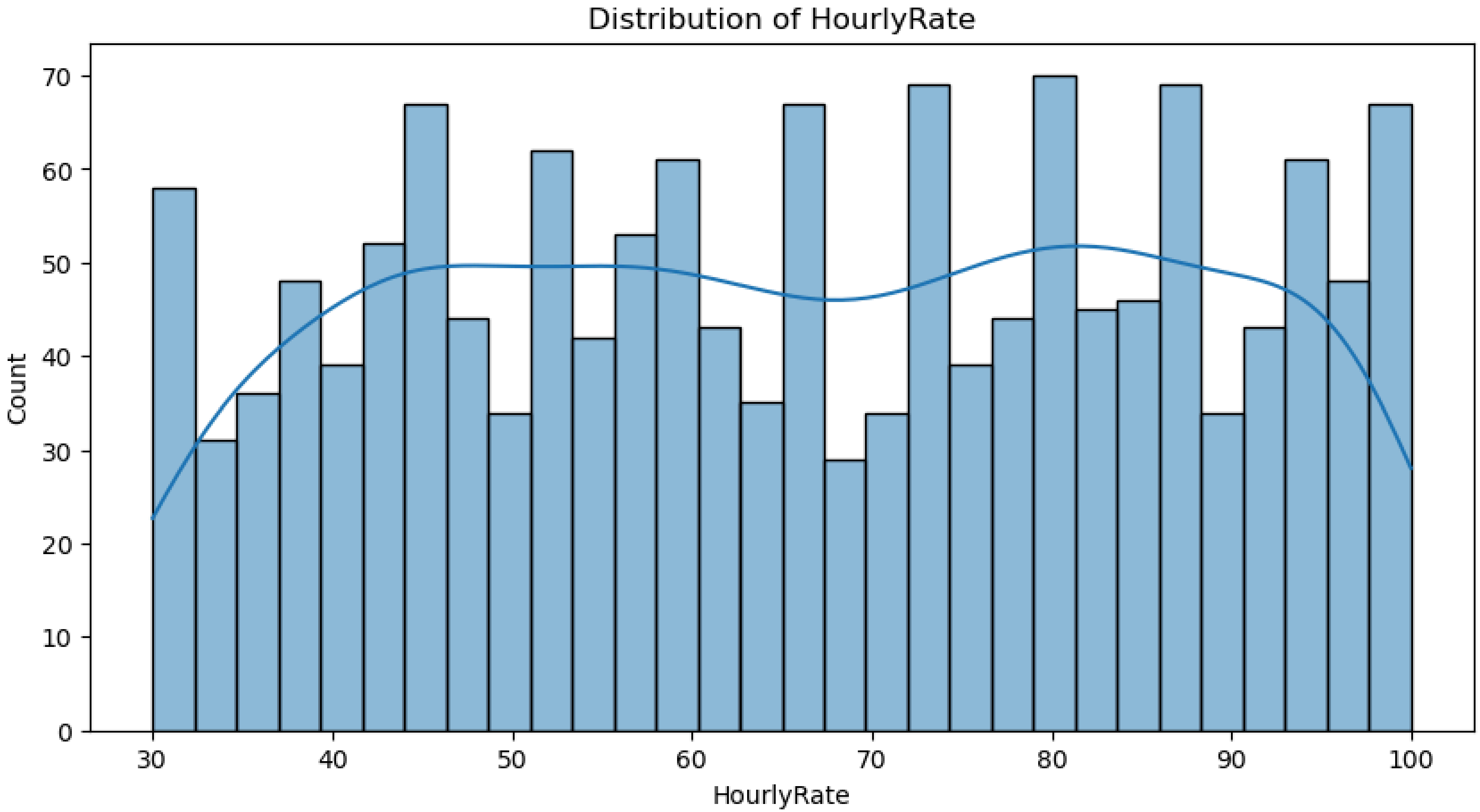

Distribution of EmployeeNumber
Count
0
10
20
30
40
50
60
0
500
1000
1500
2000
EmployeeNumber

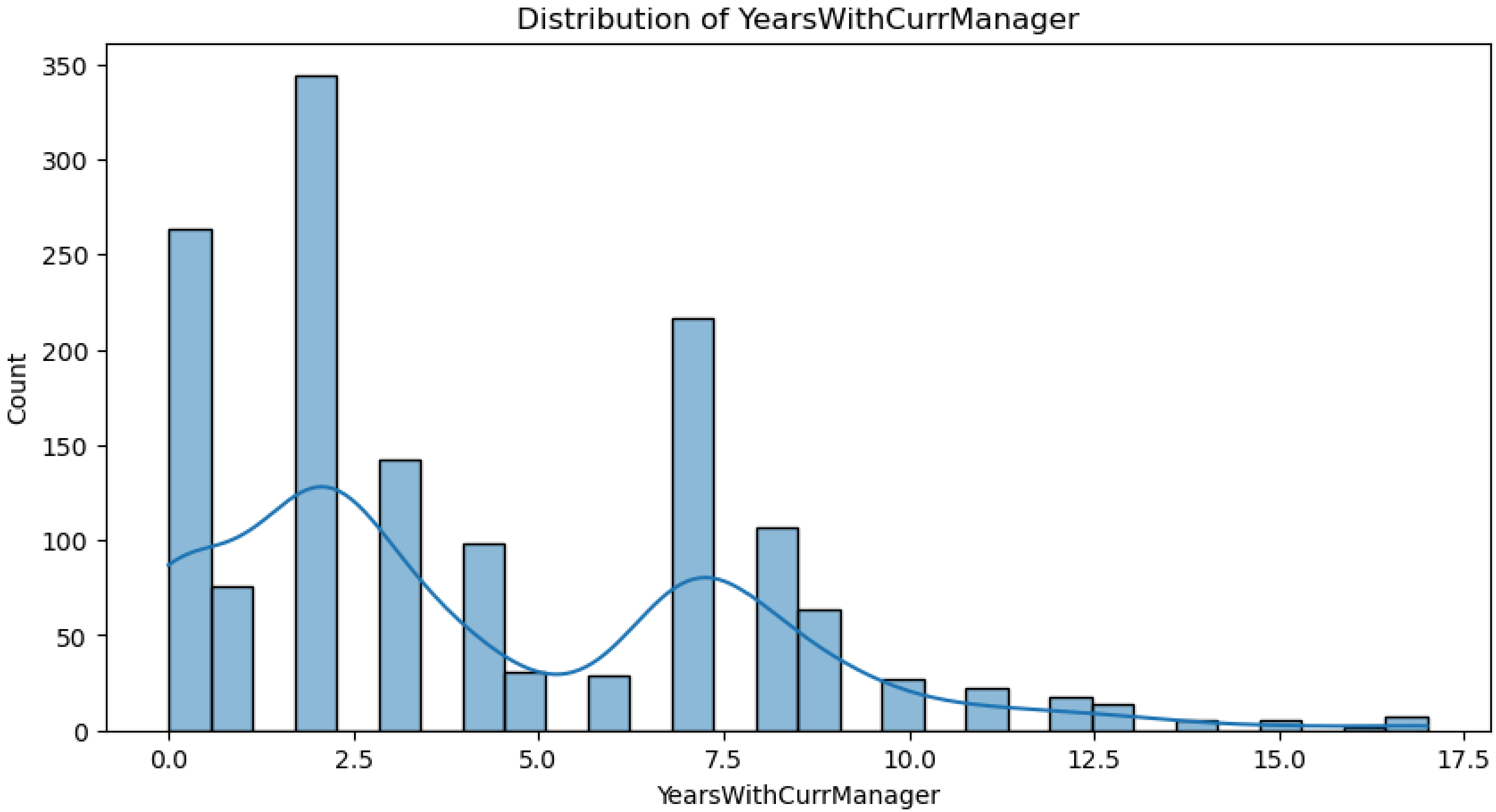

Distribution of JobLevel
Count
0
100
200
300
400
500
JobLevel
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0

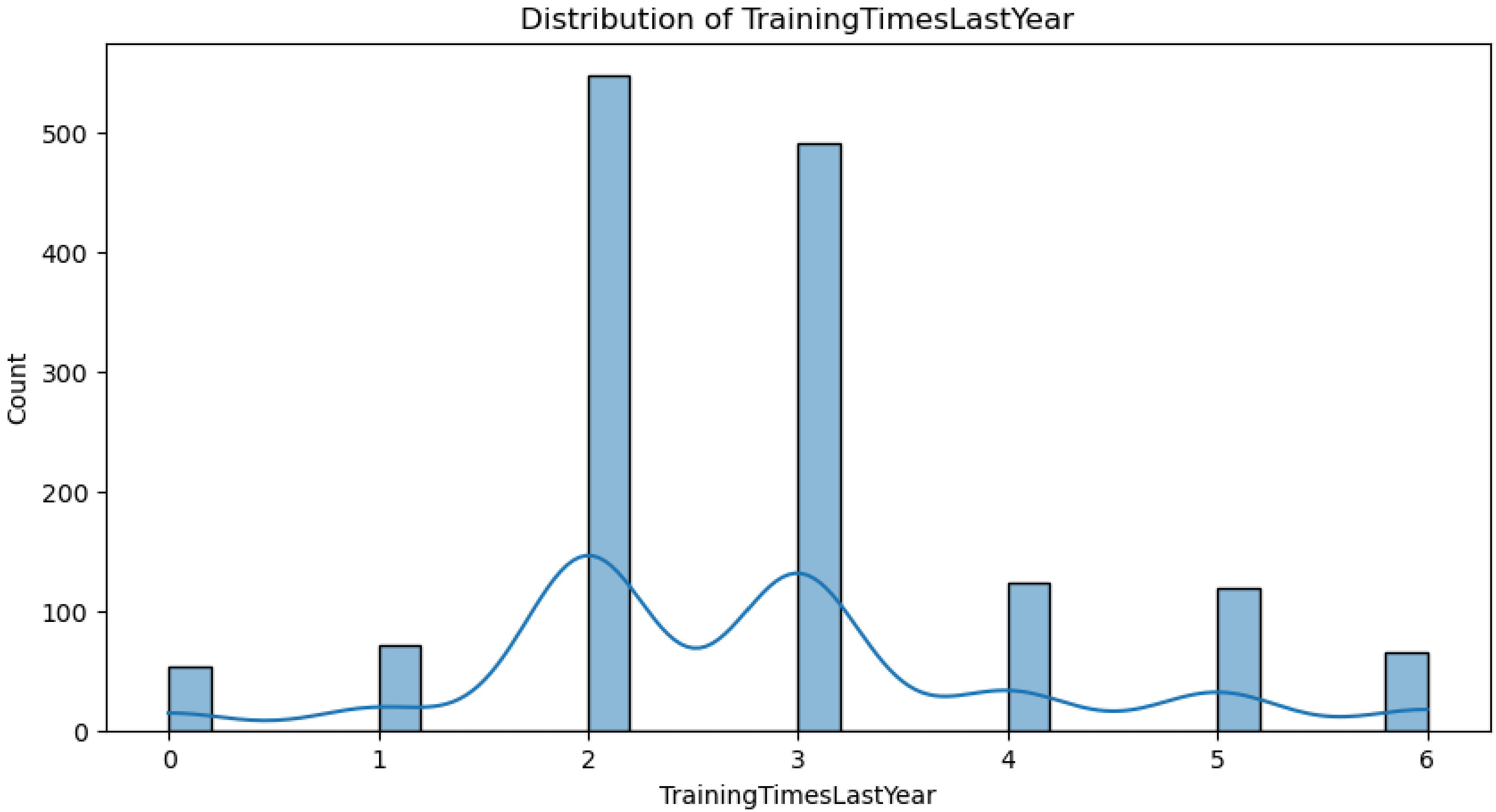

Distribution of PerformanceRating
Count
1200
1000
800
600
400
200
0
3.0
3.2
3.4
3.6
3.8
4.0
PerformanceRating

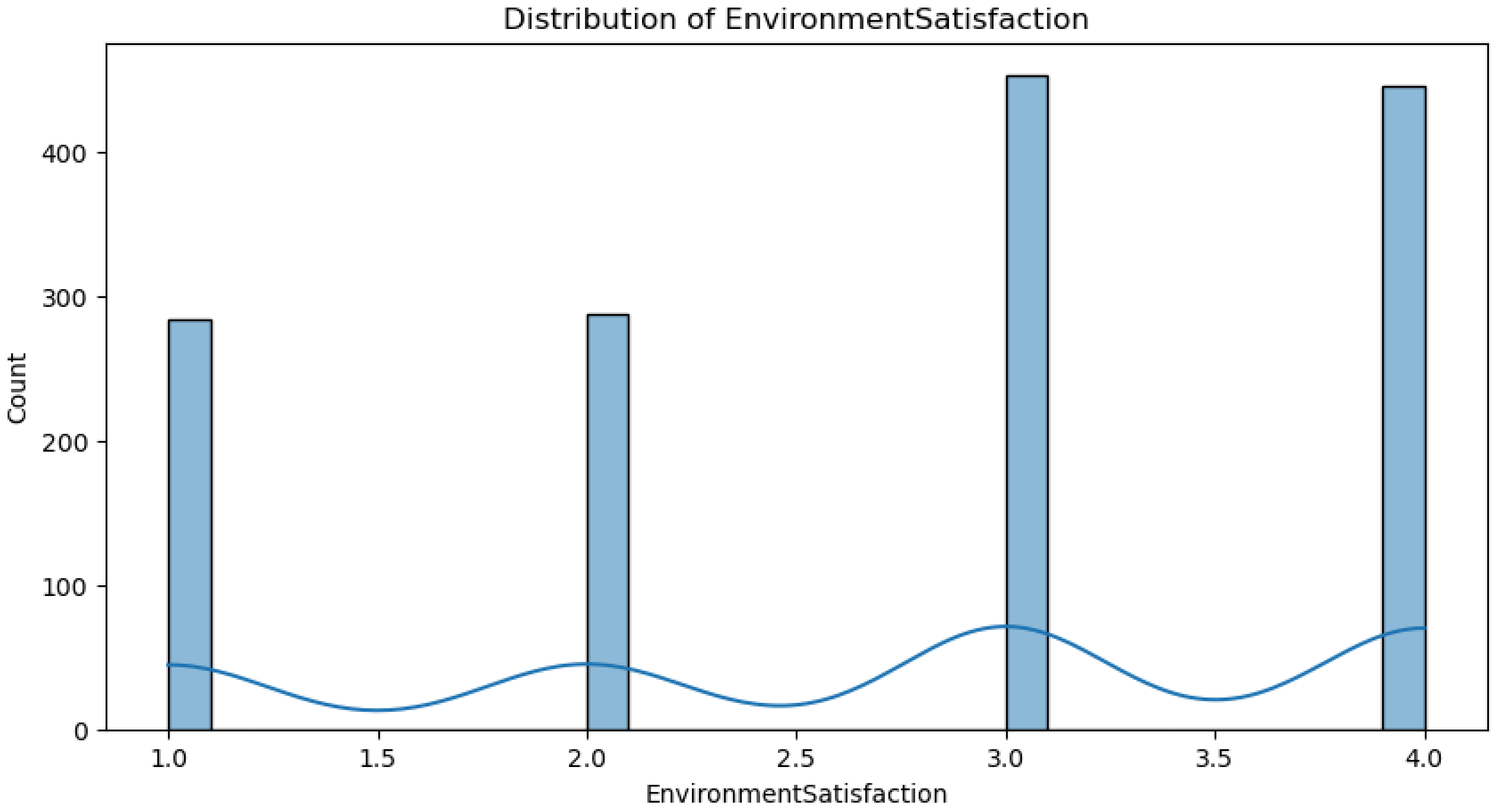

Distribution of StockOptionLevel
Count
0
100
200
300
400
500
600
0.0
0.5
1.0
1.5
2.0
2.5
3.0
StockOptionLevel

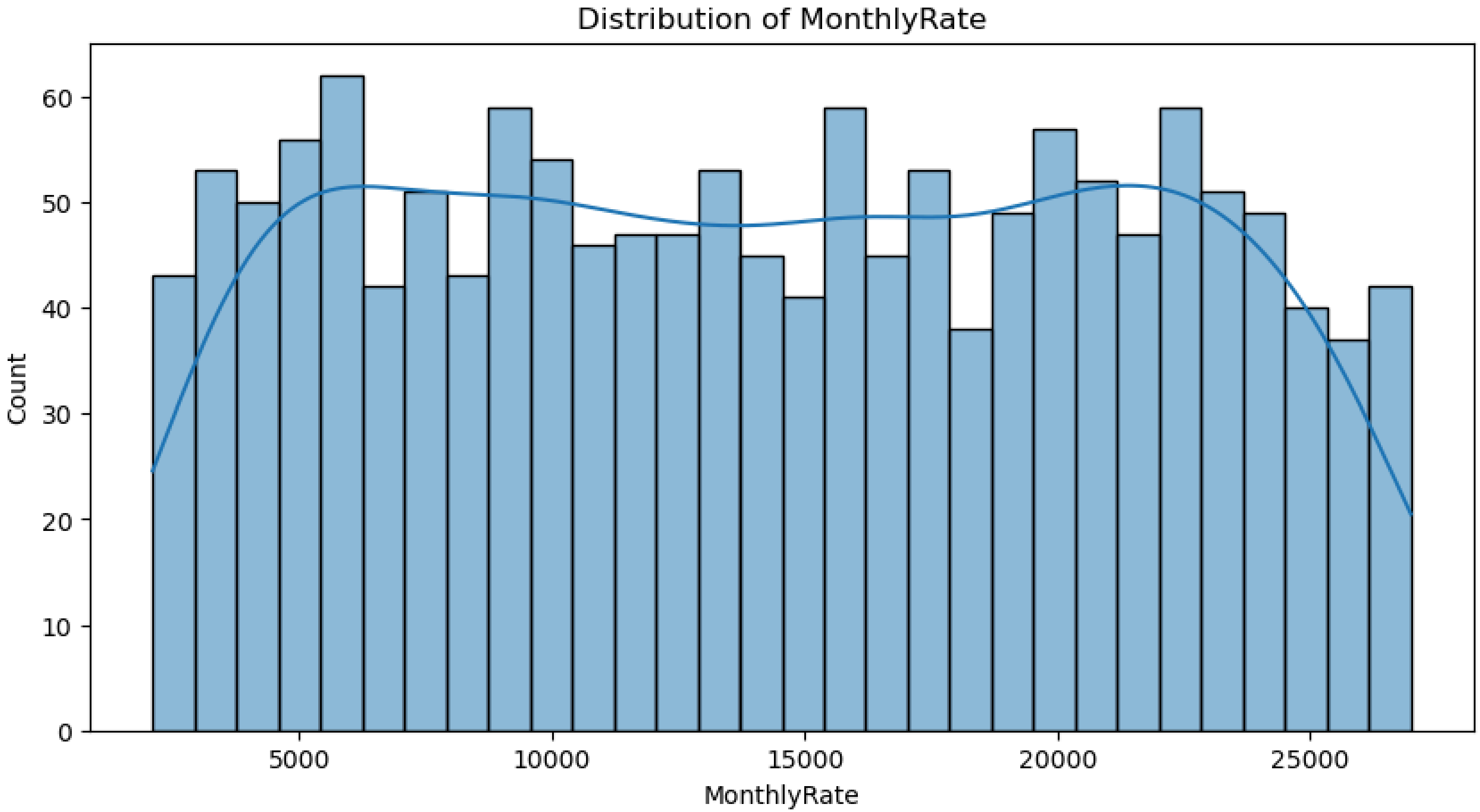

Distribution of MonthlyIncome
Count
0
50
100
150
200
2500
5000
7500
10000
12500
15000
17500
20000
MonthlyIncome

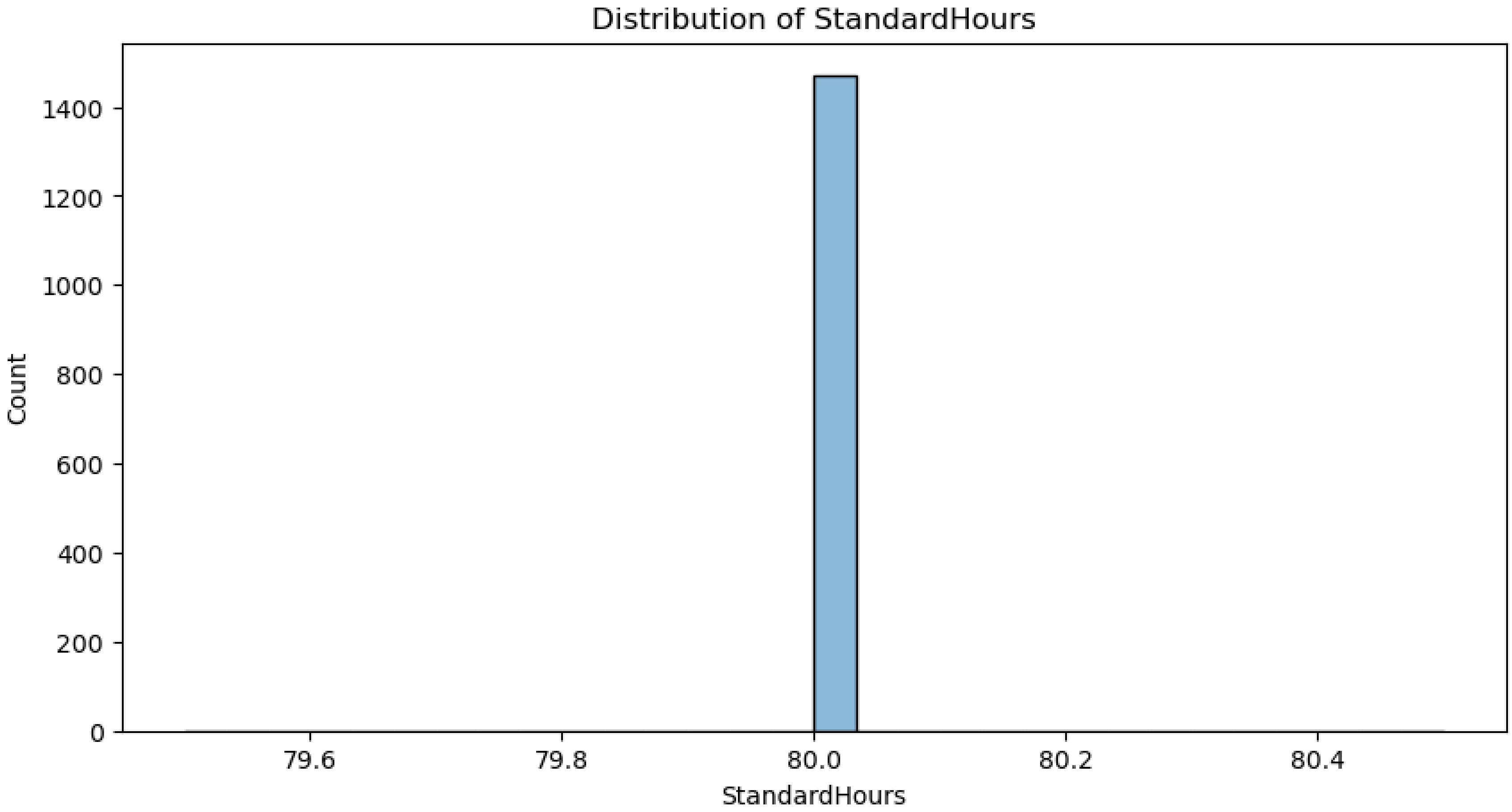

Distribution of HourlyRate
Count
HourlyRate
0
10
20
30
40
50
60
70
30
40
50
60
70
80
90
100

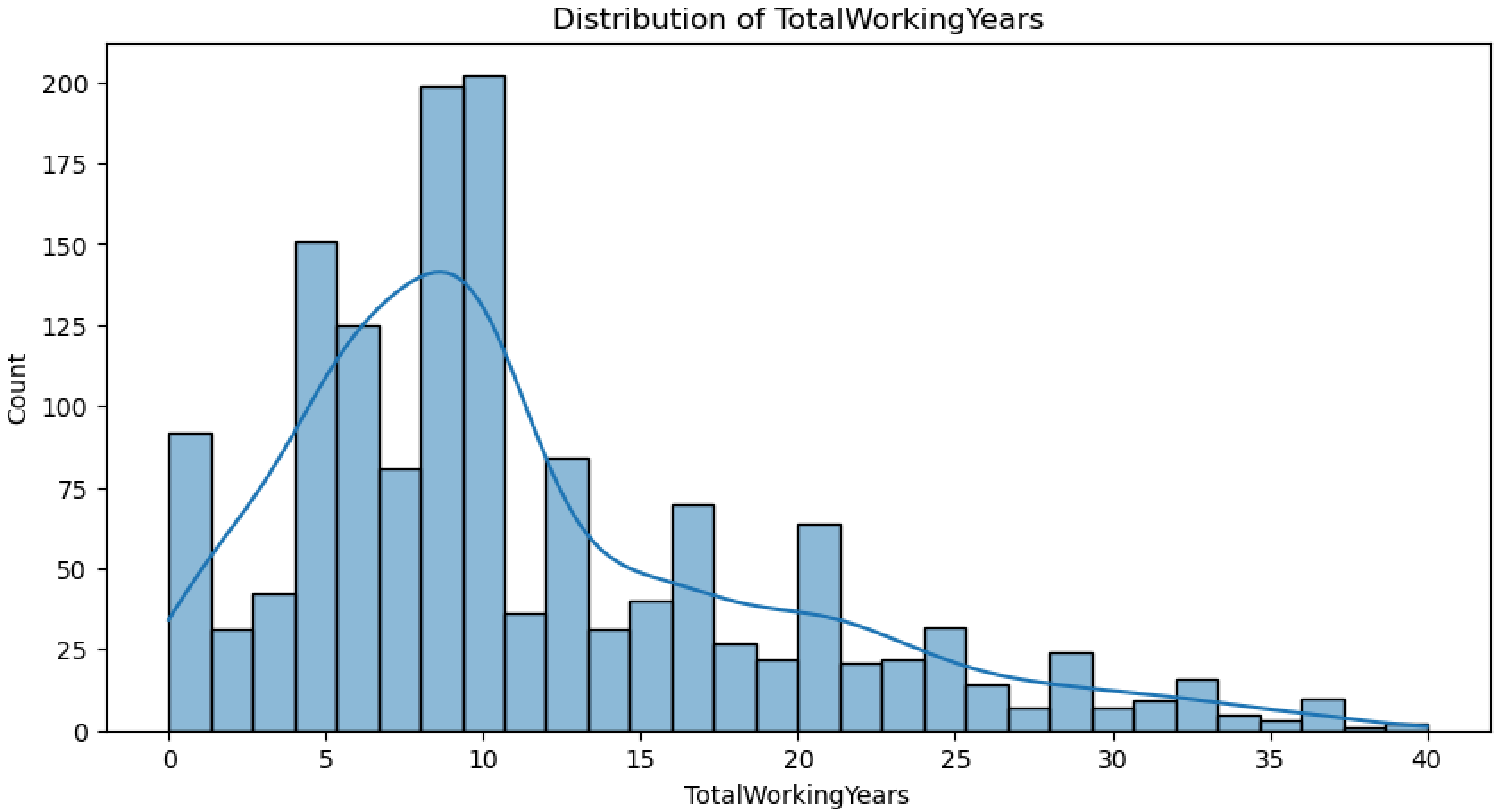

Distribution of YearsWithCurrManager
Count
0
50
100
150
200
250
300
350
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
YearsWithCurrManager

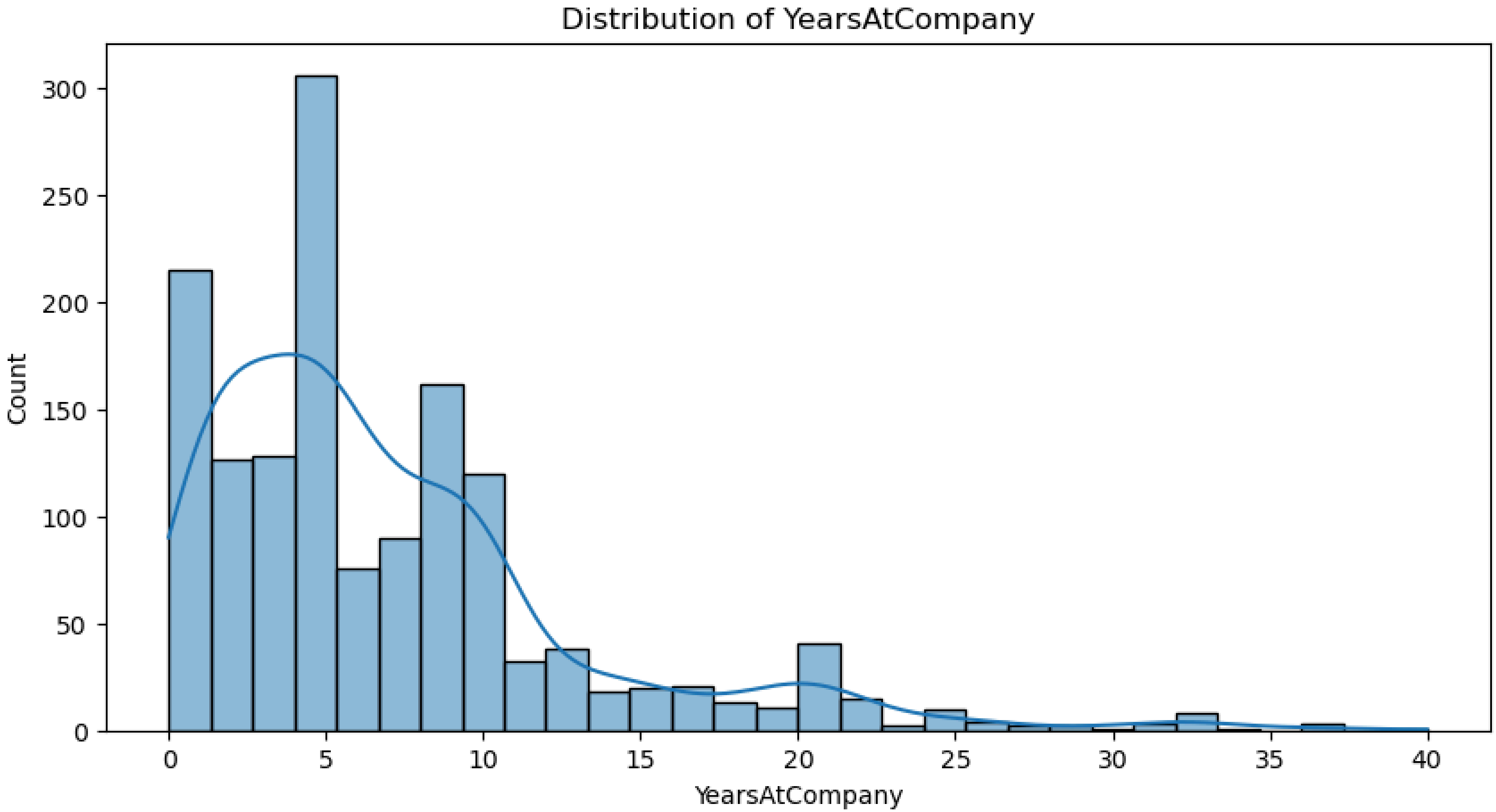

Distribution of TrainingTimesLastYear
Count
0
100
200
300
400
500
0
1
2
3
4
5
6
TrainingTimesLastYear

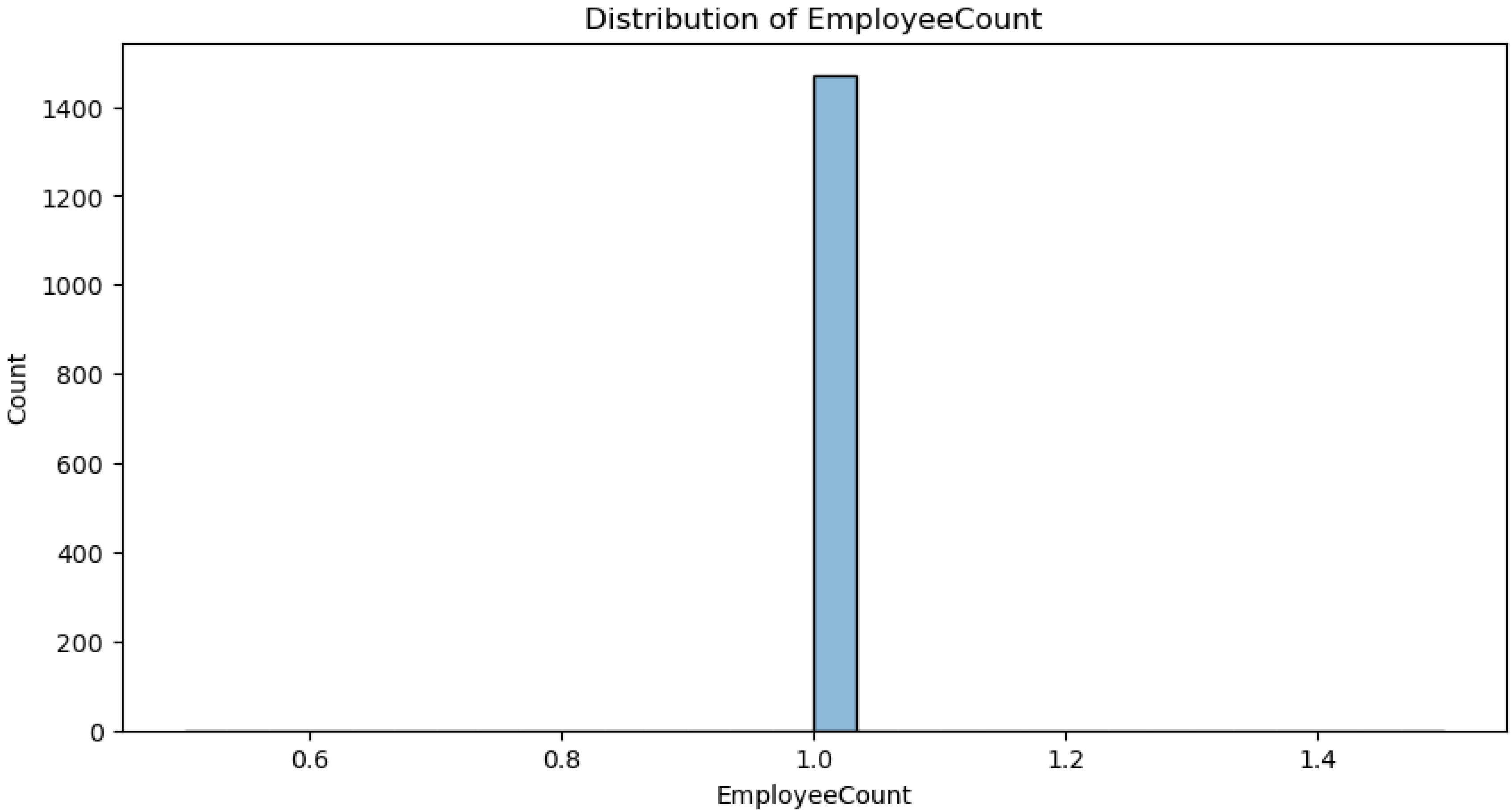

Distribution of EnvironmentSatisfaction
Count
0
100
200
300
400
1.0
1.5
2.0
2.5
3.0
3.5
4.0
EnvironmentSatisfaction

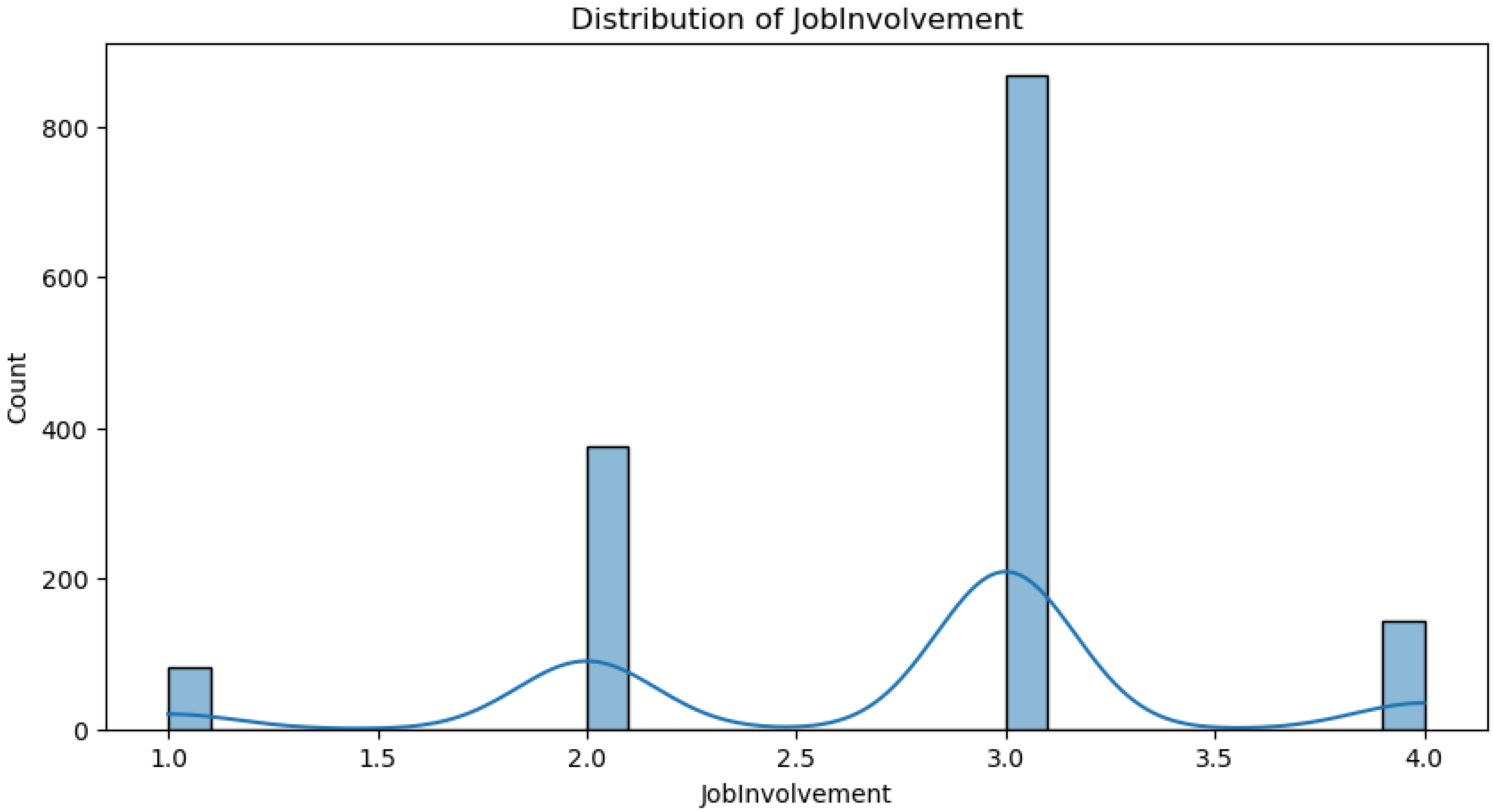

Distribution of MonthlyRate
Count
MonthlyRate
0
10
20
30
40
50
60
5000
10000
15000
20000
25000

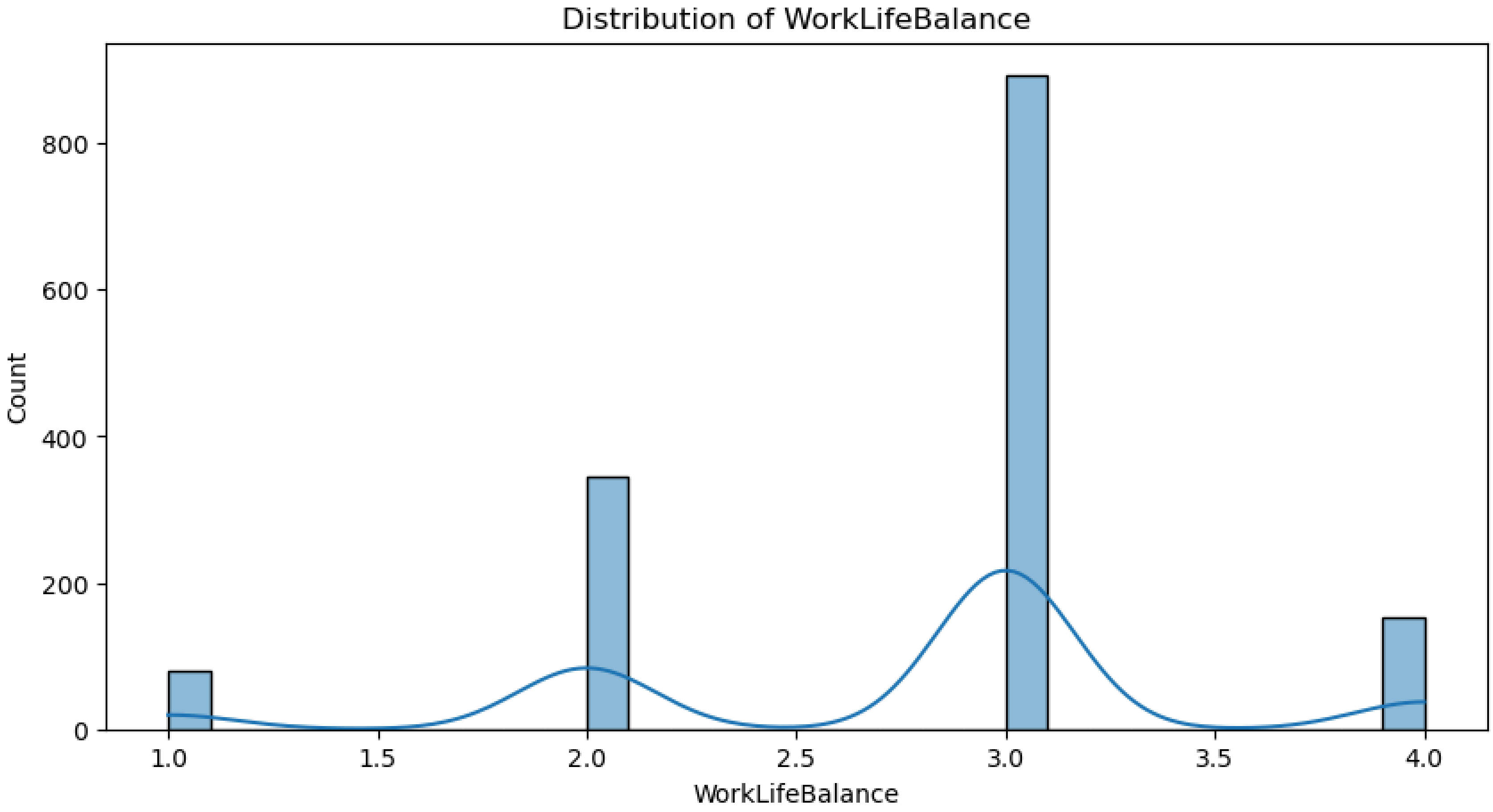

Distribution of StandardHours
Count
0
200
400
600
800
1000
1200
1400
79.6
79.8
80.0
80.2
80.4
StandardHours

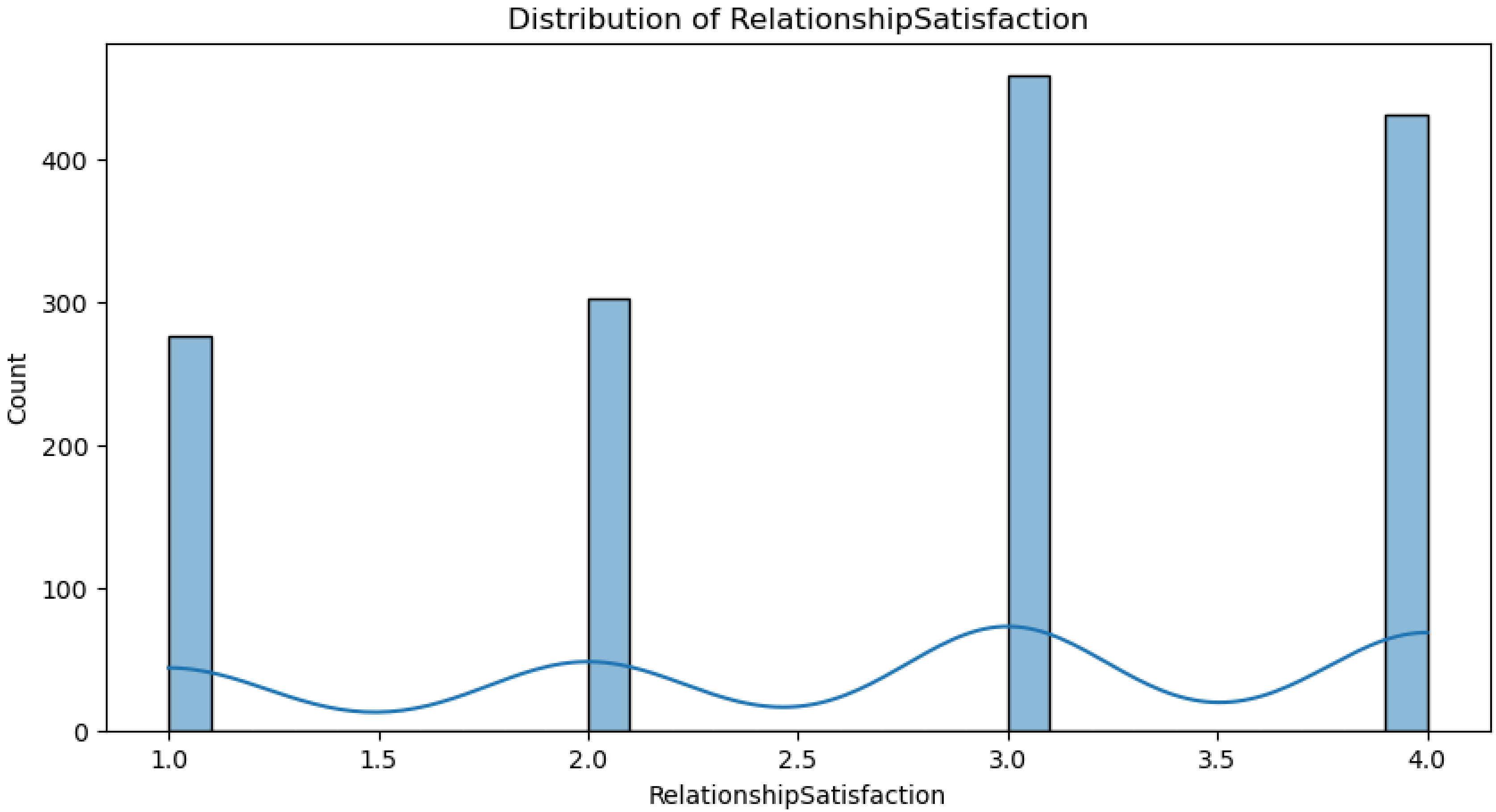

Distribution of TotalWorkingYears
Count
0
25
50
75
100
125
150
175
200
0
5
10
15
20
25
30
35
40
TotalWorkingYears

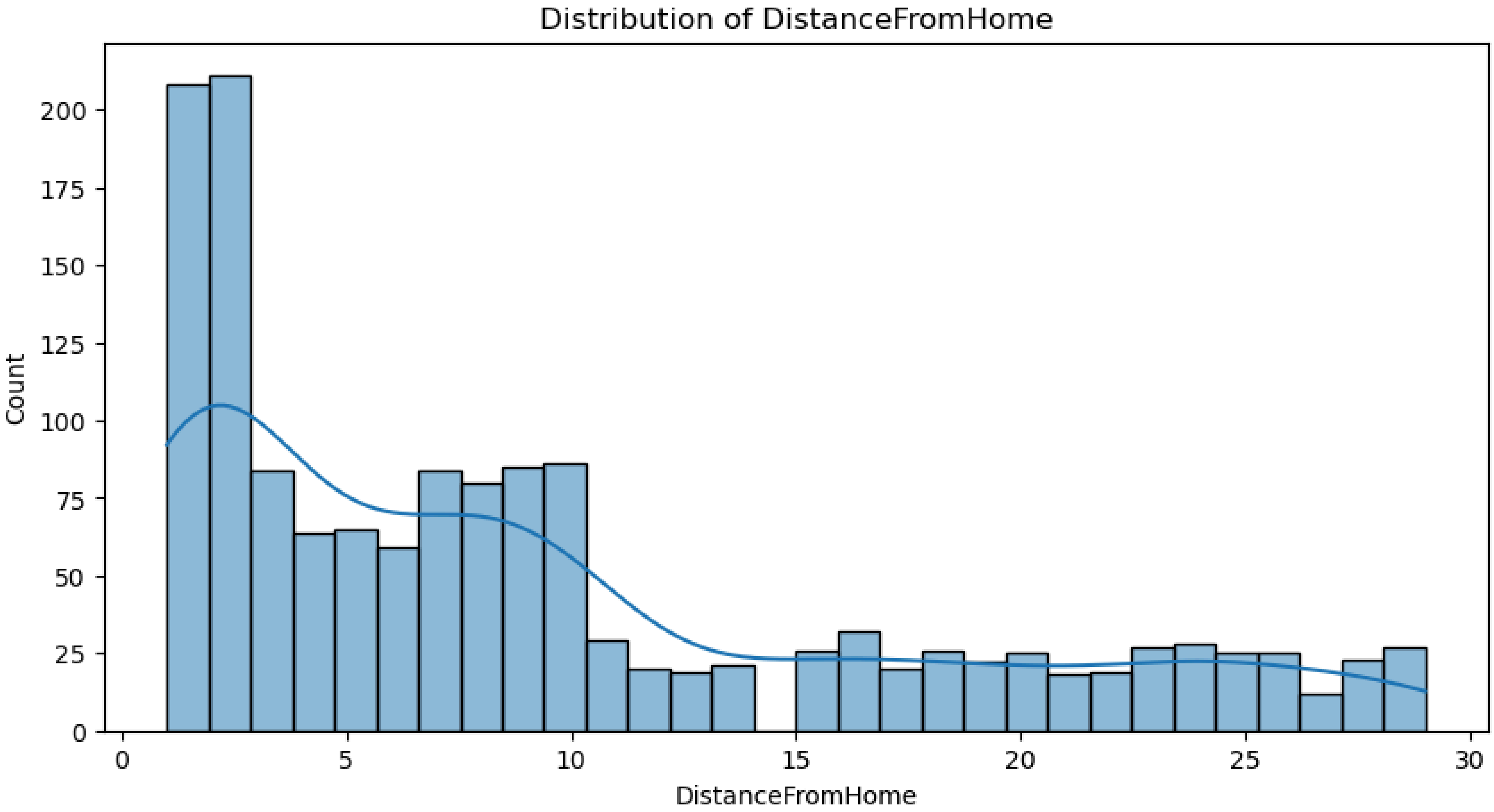

Distribution of YearsAtCompany
Count
0
50
100
150
200
250
300
0
5
10
15
20
25
30
35
40
YearsAtCompany

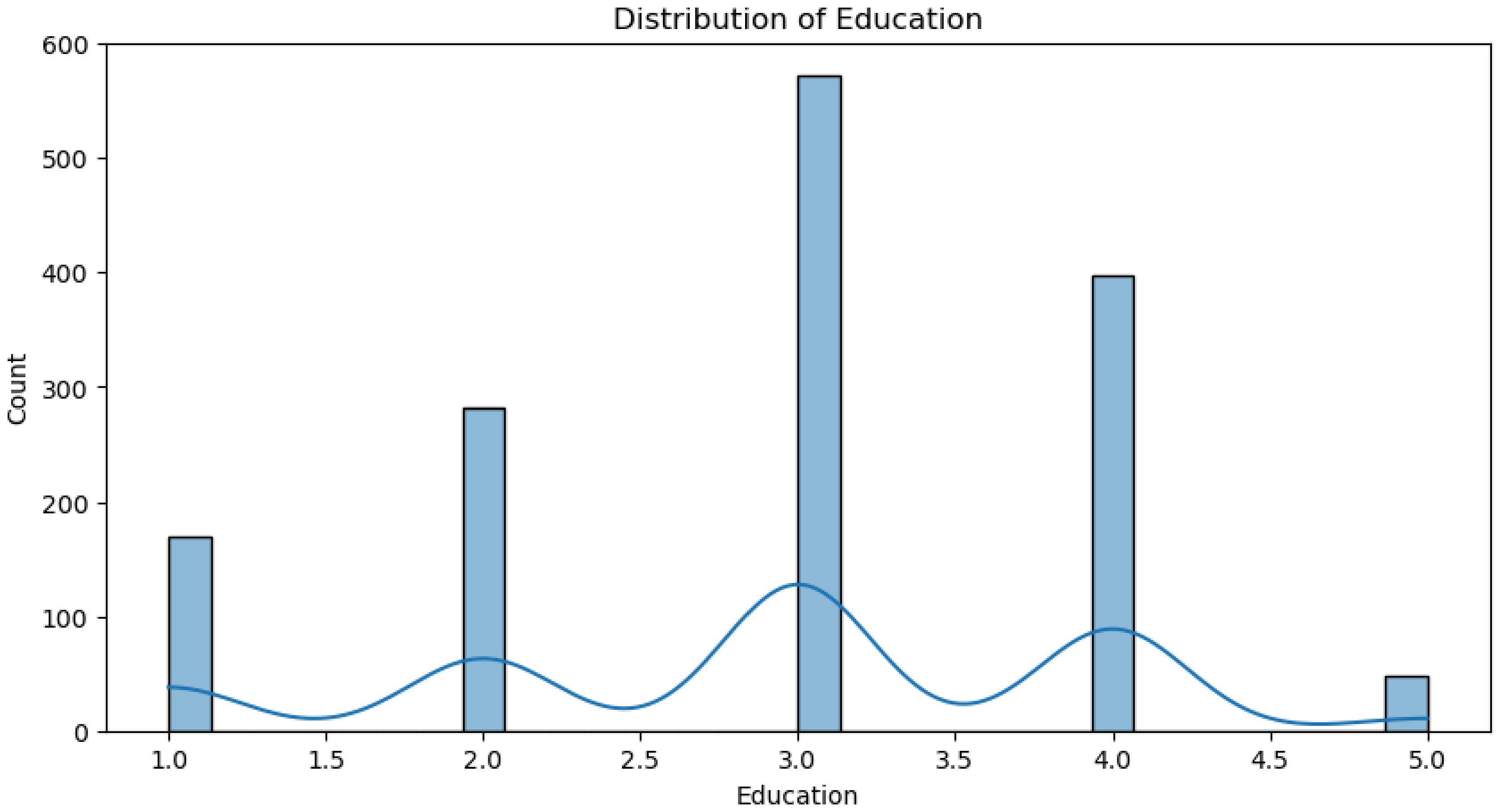

Distribution of EmployeeCount
Count
0
200
400
600
800
1000
1200
1400
0.6
0.8
1.0
1.2
1.4
EmployeeCount

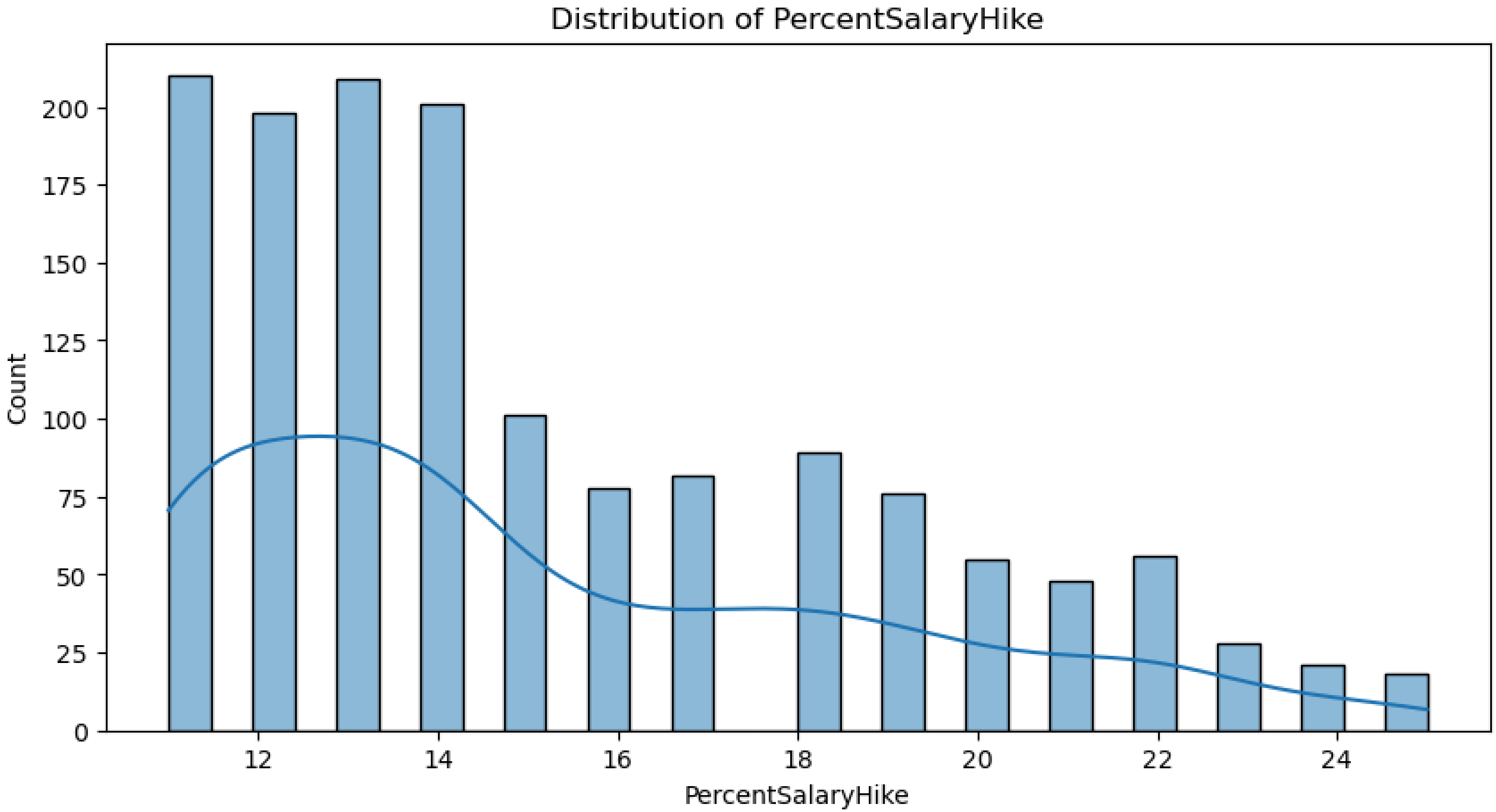

Distribution of JobInvolvement
Count
0
200
400
600
800
1.0
1.5
2.0
2.5
3.0
3.5
4.0
JobInvolvement

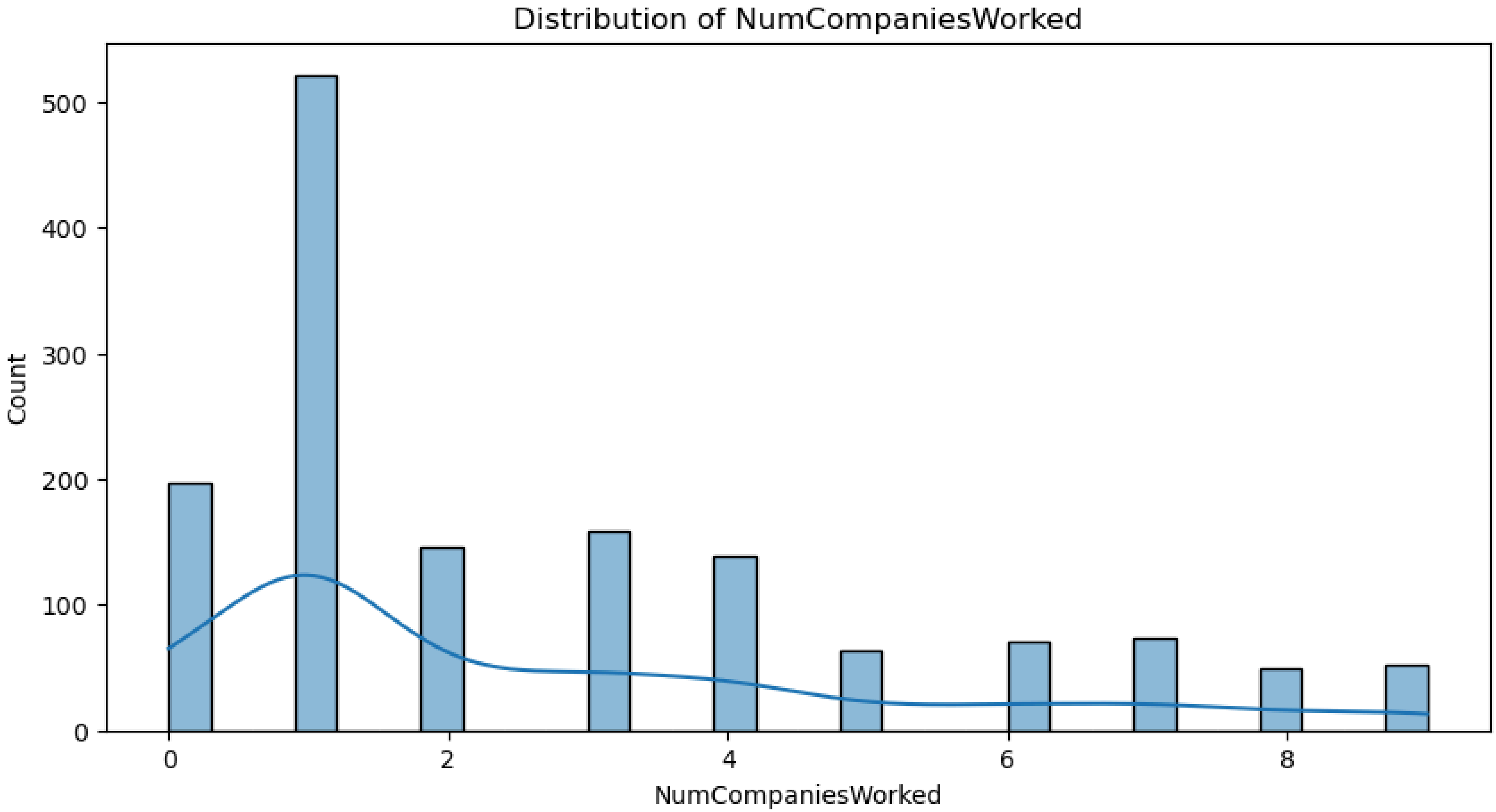

Distribution of WorkLifeBalance
Count
0
200
400
600
800
1.0
1.5
2.0
2.5
3.0
3.5
4.0
WorkLifeBalance

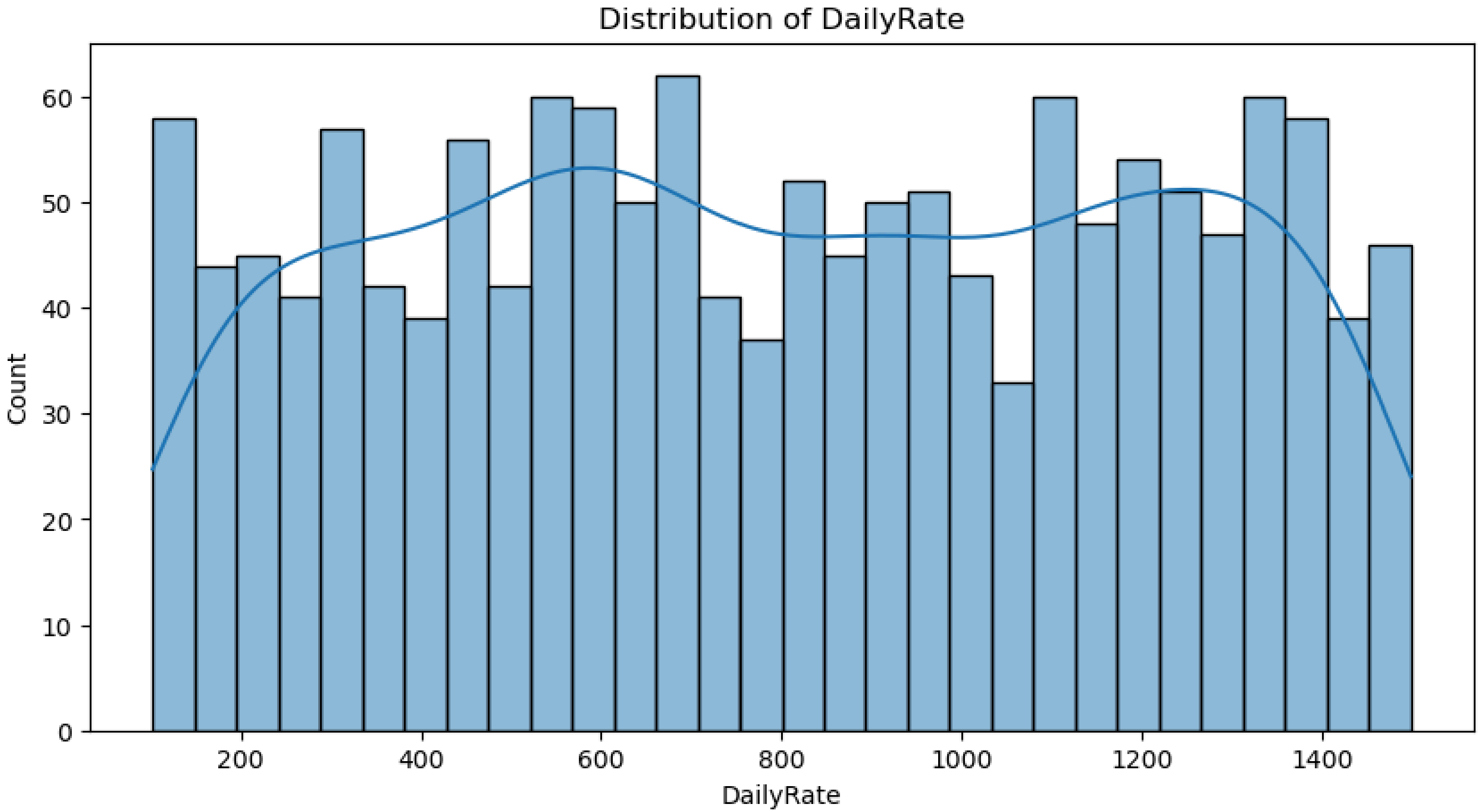

Distribution of RelationshipSatisfaction
Count
0
100
200
300
400
1.0
1.5
2.0
2.5
3.0
3.5
4.0
RelationshipSatisfaction

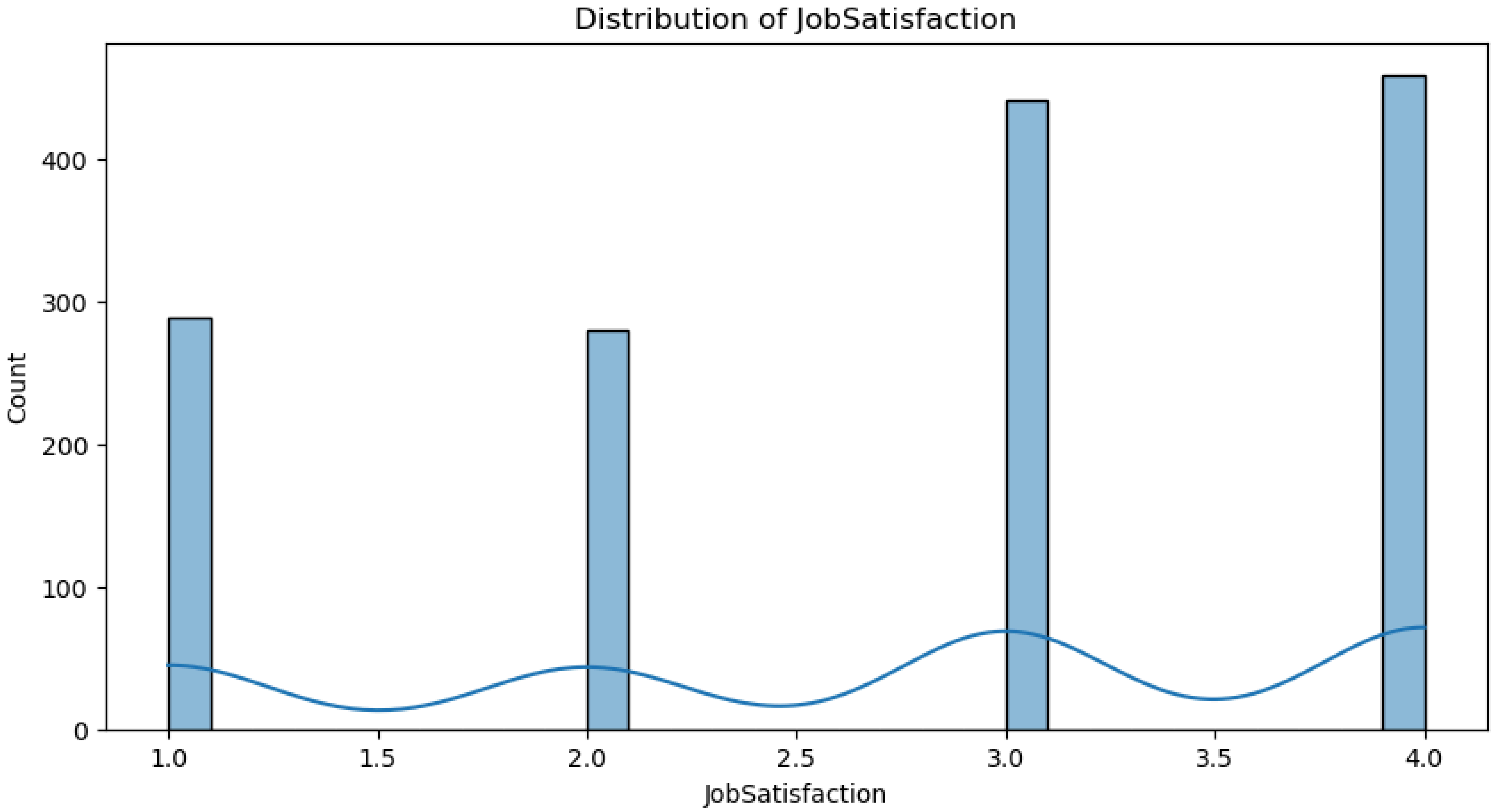

Distribution of DistanceFromHome
Count
0
25
50
75
100
125
150
175
200
0
5
10
15
20
25
30
DistanceFromHome

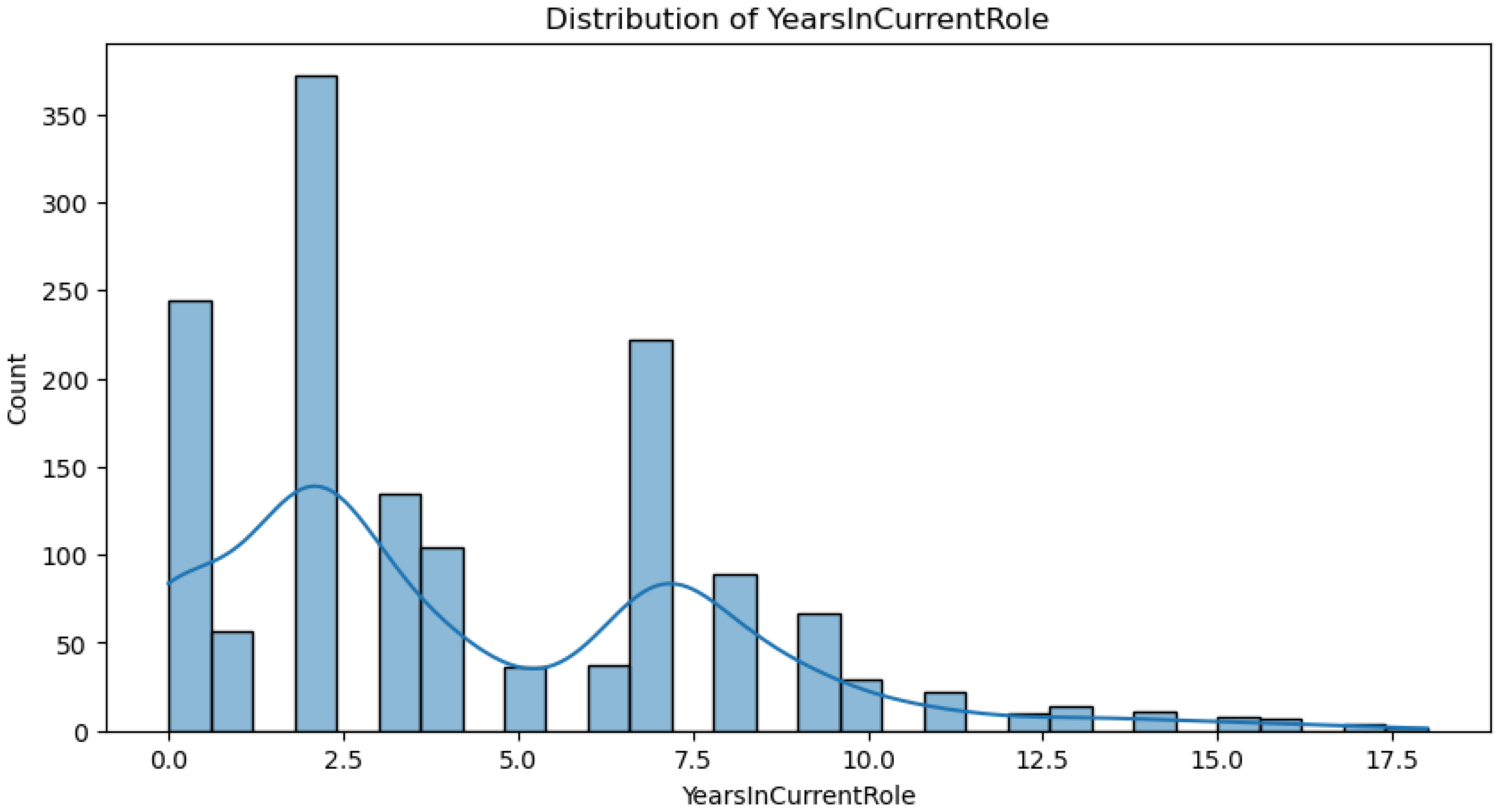

Distribution of Education
Count
0
100
200
300
400
500
600
Education
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0

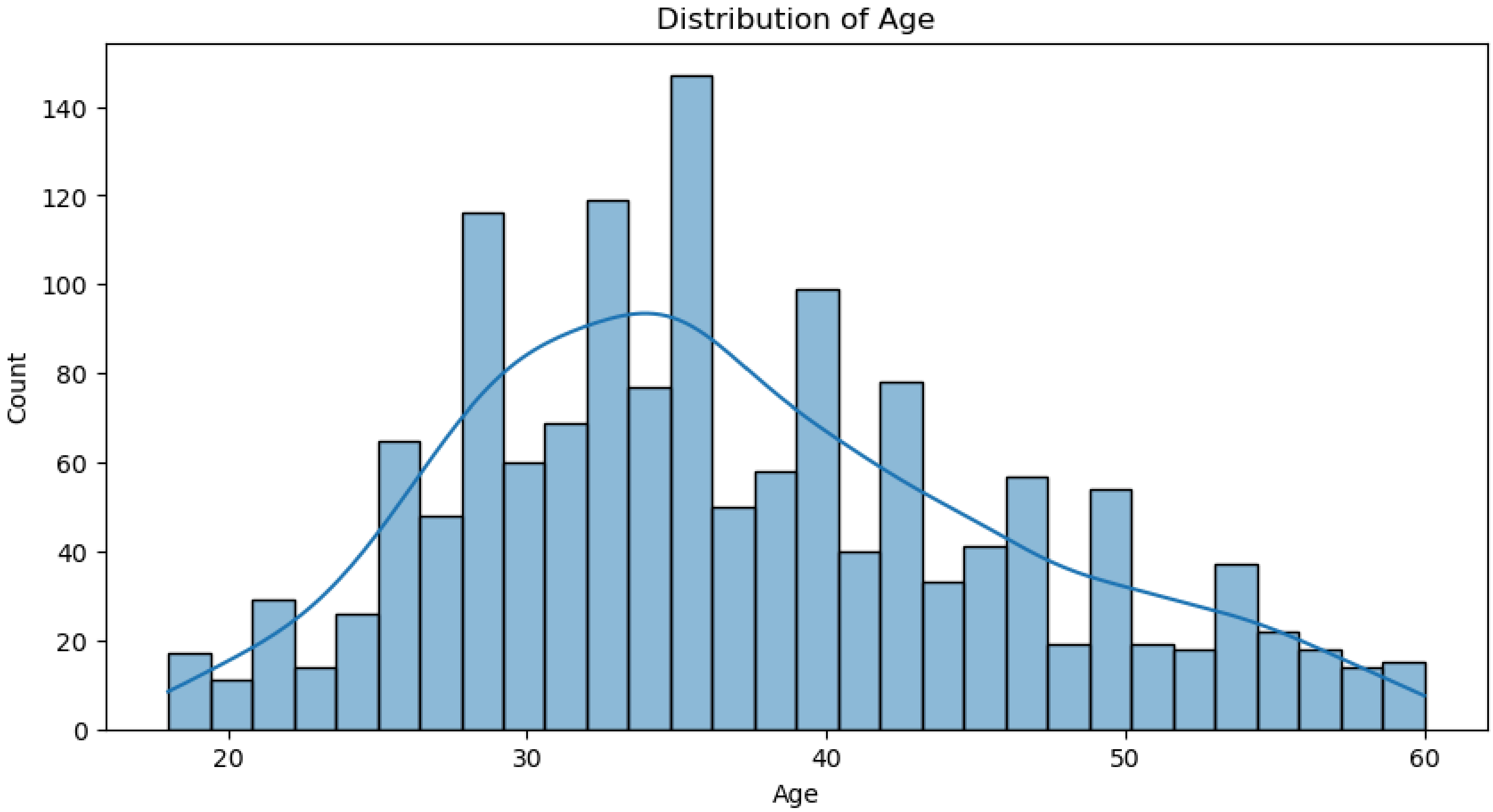

Distribution of PercentSalaryHike
Count
0
25
50
75
100
125
150
175
200
12
14
16
18
20
22
24
PercentSalaryHike

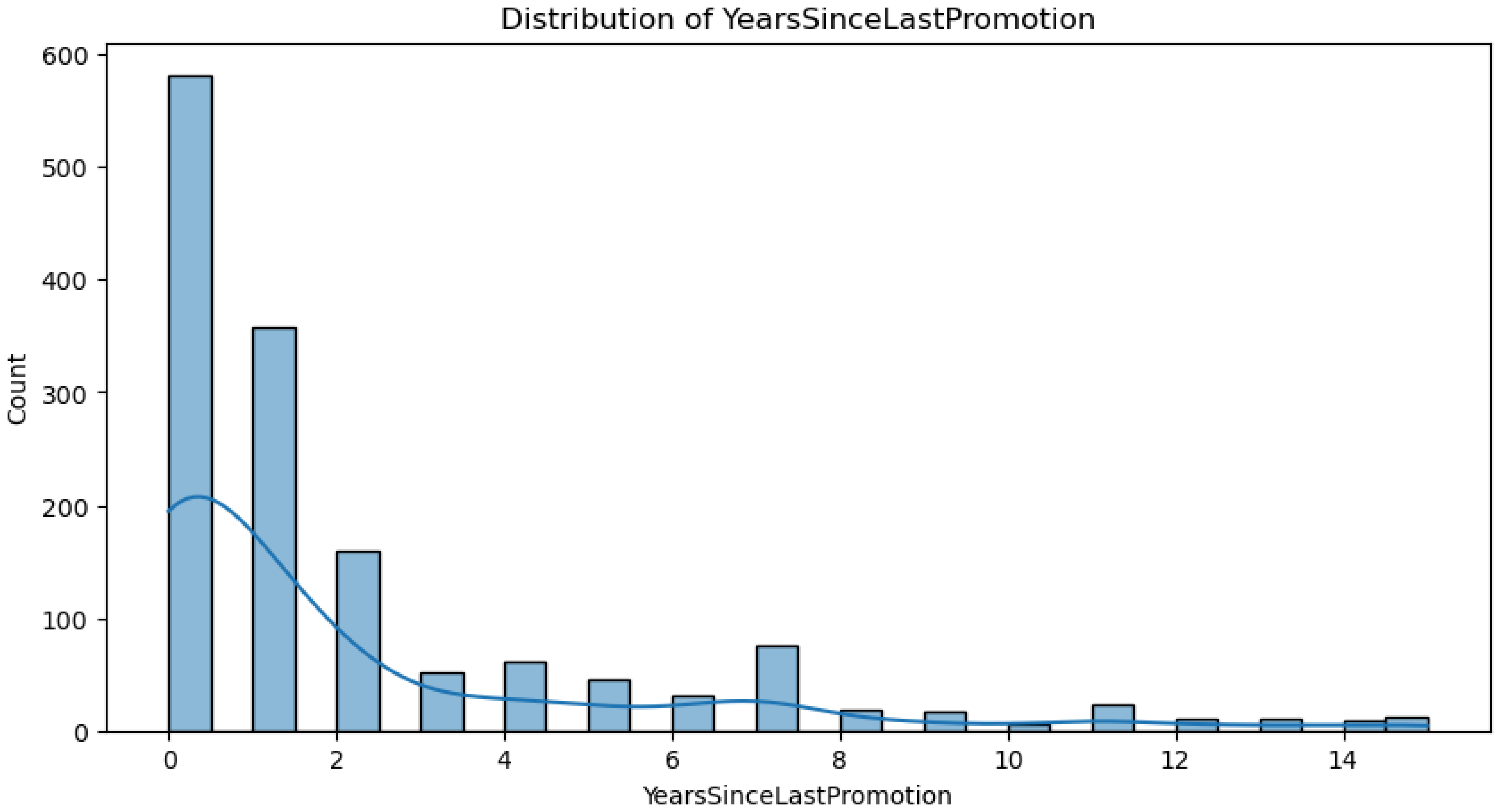

Distribution of NumCompaniesWorked
Count
0
100
200
300
400
500
0
2
4
6
8
NumCompaniesWorked


Distribution of DailyRate
Count
0
10
20
30
40
50
60
DailyRate
200
400
600
800
1000
1200
1400


Distribution of JobSatisfaction
Count
0
100
200
300
400
1.0
1.5
2.0
2.5
3.0
3.5
4.0
JobSatisfaction


Distribution of YearsInCurrentRole
Count
0
50
100
150
200
250
300
350
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
YearsInCurrentRole


Distribution of Age
Count
Age
0
20
40
60
80
100
120
140
20
30
40
50
60


Distribution of YearsSinceLastPromotion
Count
0
100
200
300
400
500
600
YearsSinceLastPromotion
0
2
4
6
8
10
12
14

```
In [ ]:
```

```
In [ ]:
```

```
In [ ]:
```

```
In [ ]:
```