In [ ]:

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

In [ ]:

```
import pandas as pd

# Load the dataset
file_path = '/content/Attrition data.csv'
data = pd.read_csv(file_path)

# Display the first few rows
print("First few rows of the dataset:")
print(data.head())

# Get a summary of the DataFrame
print("\nDataFrame info:")
print(data.info())

# Check for missing values
print("\nMissing values in each column:")
print(data.isnull().sum())

# Display the columns to identify the empty column
print("\nColumns in the dataset:")
print(data.columns)

print("\nSummary statistics after cleaning:")
print(data.describe())
```

```
First few rows of the dataset:
   EmployeeID  Age Attrition     BusinessTravel              Department  \
0           1   51        No      Travel_Rarely                   Sales
1           2   31       Yes  Travel_Frequently  Research & Development
2           3   32        No  Travel_Frequently  Research & Development
3           4   38        No         Non-Travel  Research & Development
4           5   32        No      Travel_Rarely  Research & Development

   DistanceFromHome  Education EducationField  EmployeeCount  Gender  ...  \
0                 6          2  Life Sciences              1  Female  ...
1                10          1  Life Sciences              1  Female  ...
2                17          4          Other              1    Male  ...
3                 2          5  Life Sciences              1    Male  ...
4                10          1        Medical              1    Male  ...

   TotalWorkingYears TrainingTimesLastYear YearsAtCompany  \
0                1.0                     6              1
1                6.0                     3              5
2                5.0                     2              5
3               13.0                     5              8
4                9.0                     2              6

   YearsSinceLastPromotion  YearsWithCurrManager EnvironmentSatisfaction  \
0                        0                     0                     3.0
1                        1                     4                     3.0
2                        0                     3                     2.0
3                        7                     5                     4.0
4                        0                     4                     4.0

   JobSatisfaction  WorkLifeBalance  JobInvolvement  PerformanceRating
0              4.0              2.0               3                  3
1              2.0              4.0               2                  4
2              2.0              1.0               3                  3
3              4.0              3.0               2                  3
4              1.0              3.0               3                  3
```

```
[5 rows x 29 columns]

DataFrame info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4410 entries, 0 to 4409
Data columns (total 29 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   EmployeeID               4410 non-null   int64
 1   Age                      4410 non-null   int64
 2   Attrition                4410 non-null   object
 3   BusinessTravel           4410 non-null   object
 4   Department               4410 non-null   object
 5   DistanceFromHome         4410 non-null   int64
 6   Education                4410 non-null   int64
 7   EducationField           4410 non-null   object
 8   EmployeeCount            4410 non-null   int64
 9   Gender                   4410 non-null   object
 10  JobLevel                 4410 non-null   int64
 11  JobRole                  4410 non-null   object
 12  MaritalStatus            4410 non-null   object
 13  MonthlyIncome            4410 non-null   int64
 14  NumCompaniesWorked       4391 non-null   float64
 15  Over18                   4410 non-null   object
 16  PercentSalaryHike        4410 non-null   int64
 17  StandardHours            4410 non-null   int64
 18  StockOptionLevel         4410 non-null   int64
 19  TotalWorkingYears        4401 non-null   float64
 20  TrainingTimesLastYear    4410 non-null   int64
 21  YearsAtCompany           4410 non-null   int64
 22  YearsSinceLastPromotion  4410 non-null   int64
 23  YearsWithCurrManager     4410 non-null   int64
 24  EnvironmentSatisfaction  4385 non-null   float64
 25  JobSatisfaction          4390 non-null   float64
 26  WorkLifeBalance          4372 non-null   float64
 27  JobInvolvement           4410 non-null   int64
 28  PerformanceRating        4410 non-null   int64
dtypes: float64(5), int64(16), object(8)
memory usage: 999.3+ KB
None

Missing values in each column:
EmployeeID                  0
Age                         0
Attrition                   0
BusinessTravel              0
Department                  0
DistanceFromHome            0
Education                   0
EducationField              0
EmployeeCount               0
Gender                      0
JobLevel                    0
JobRole                     0
MaritalStatus               0
MonthlyIncome               0
NumCompaniesWorked         19
Over18                      0
PercentSalaryHike           0
StandardHours               0
StockOptionLevel            0
TotalWorkingYears           9
TrainingTimesLastYear       0
YearsAtCompany              0
YearsSinceLastPromotion     0
YearsWithCurrManager        0
EnvironmentSatisfaction    25
JobSatisfaction            20
WorkLifeBalance            38
JobInvolvement              0
PerformanceRating           0
```

```
dtype: int64

Columns in the dataset:
Index(['EmployeeID', 'Age', 'Attrition', 'BusinessTravel', 'Department',
       'DistanceFromHome', 'Education', 'EducationField', 'EmployeeCount',
       'Gender', 'JobLevel', 'JobRole', 'MaritalStatus', 'MonthlyIncome',
       'NumCompaniesWorked', 'Over18', 'PercentSalaryHike', 'StandardHours',
       'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear',
       'YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManager',
       'EnvironmentSatisfaction', 'JobSatisfaction', 'WorkLifeBalance',
       'JobInvolvement', 'PerformanceRating'],
      dtype='object')

Summary statistics after cleaning:
        EmployeeID          Age  DistanceFromHome    Education  EmployeeCount  \
count  4410.000000  4410.000000       4410.000000  4410.000000         4410.0
mean   2205.500000    36.923810          9.192517     2.912925            1.0
std    1273.201673     9.133301          8.105026     1.023933            0.0
min       1.000000    18.000000          1.000000     1.000000            1.0
25%    1103.250000    30.000000          2.000000     2.000000            1.0
50%    2205.500000    36.000000          7.000000     3.000000            1.0
75%    3307.750000    43.000000         14.000000     4.000000            1.0
max    4410.000000    60.000000         29.000000     5.000000            1.0

          JobLevel  MonthlyIncome  NumCompaniesWorked  PercentSalaryHike  \
count  4410.000000    4410.000000         4391.000000        4410.000000
mean      2.063946   65029.312925            2.694830          15.209524
std       1.106689   47068.888559            2.498887           3.659108
min       1.000000   10090.000000            0.000000          11.000000
25%       1.000000   29110.000000            1.000000          12.000000
50%       2.000000   49190.000000            2.000000          14.000000
75%       3.000000   83800.000000            4.000000          18.000000
max       5.000000  199990.000000            9.000000          25.000000

       StandardHours  ...  TotalWorkingYears  TrainingTimesLastYear  \
count         4410.0  ...        4401.000000            4410.000000
mean             8.0  ...          11.279936               2.799320
std              0.0  ...           7.782222               1.288978
min              8.0  ...           0.000000               0.000000
25%              8.0  ...           6.000000               2.000000
50%              8.0  ...          10.000000               3.000000
75%              8.0  ...          15.000000               3.000000
max              8.0  ...          40.000000               6.000000

       YearsAtCompany  YearsSinceLastPromotion  YearsWithCurrManager  \
count     4410.000000              4410.000000           4410.000000
mean         7.008163                 2.187755              4.123129
std          6.125135                 3.221699              3.567327
min          0.000000                 0.000000              0.000000
25%          3.000000                 0.000000              2.000000
50%          5.000000                 1.000000              3.000000
75%          9.000000                 3.000000              7.000000
max         40.000000                15.000000             17.000000

       EnvironmentSatisfaction  JobSatisfaction  WorkLifeBalance  \
count              4385.000000      4390.000000      4372.000000
mean                  2.723603         2.728246         2.761436
std                   1.092756         1.101253         0.706245
min                   1.000000         1.000000         1.000000
25%                   2.000000         2.000000         2.000000
50%                   3.000000         3.000000         3.000000
75%                   4.000000         4.000000         3.000000
max                   4.000000         4.000000         4.000000

       JobInvolvement  PerformanceRating
count     4410.000000        4410.000000
mean         2.729932           3.153741
std          0.711400           0.360742
min          1.000000           3.000000
25%          2.000000           3.000000
50%          3.000000           3.000000
75%          3.000000           3.000000
```

```
max           4.000000            4.000000

[8 rows x 21 columns]
```

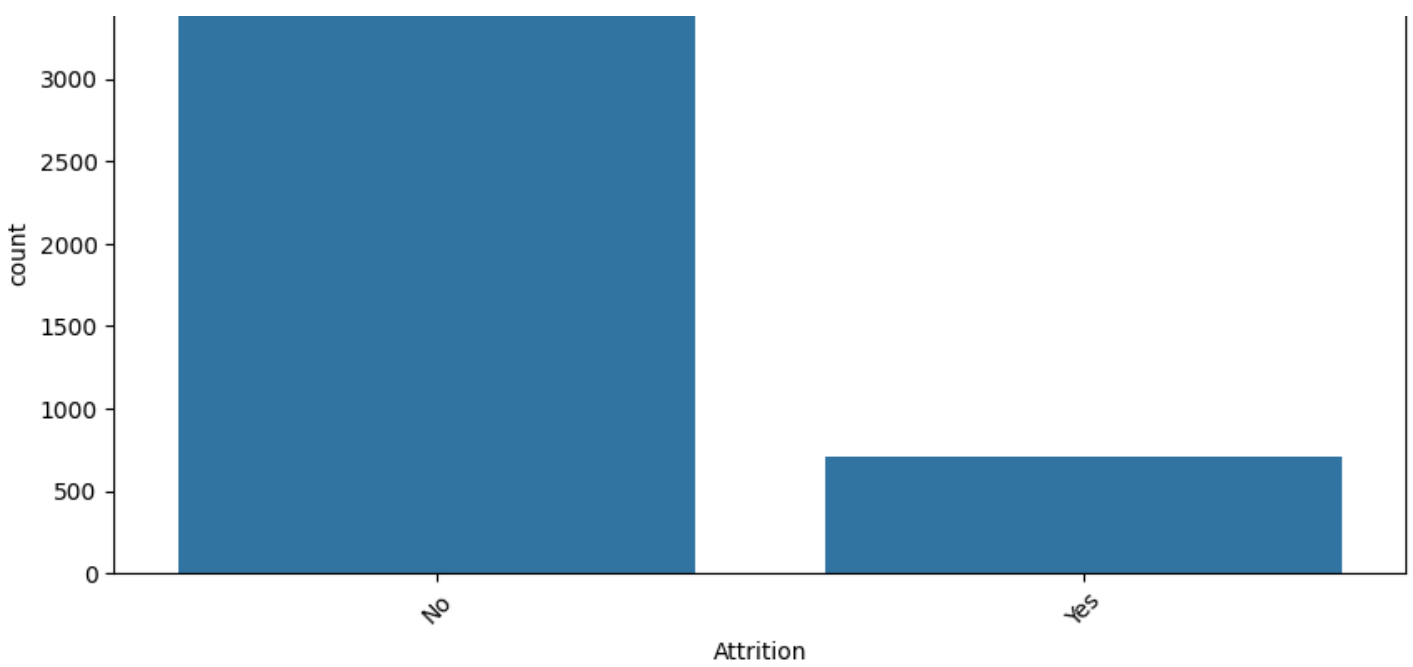
In [ ]:

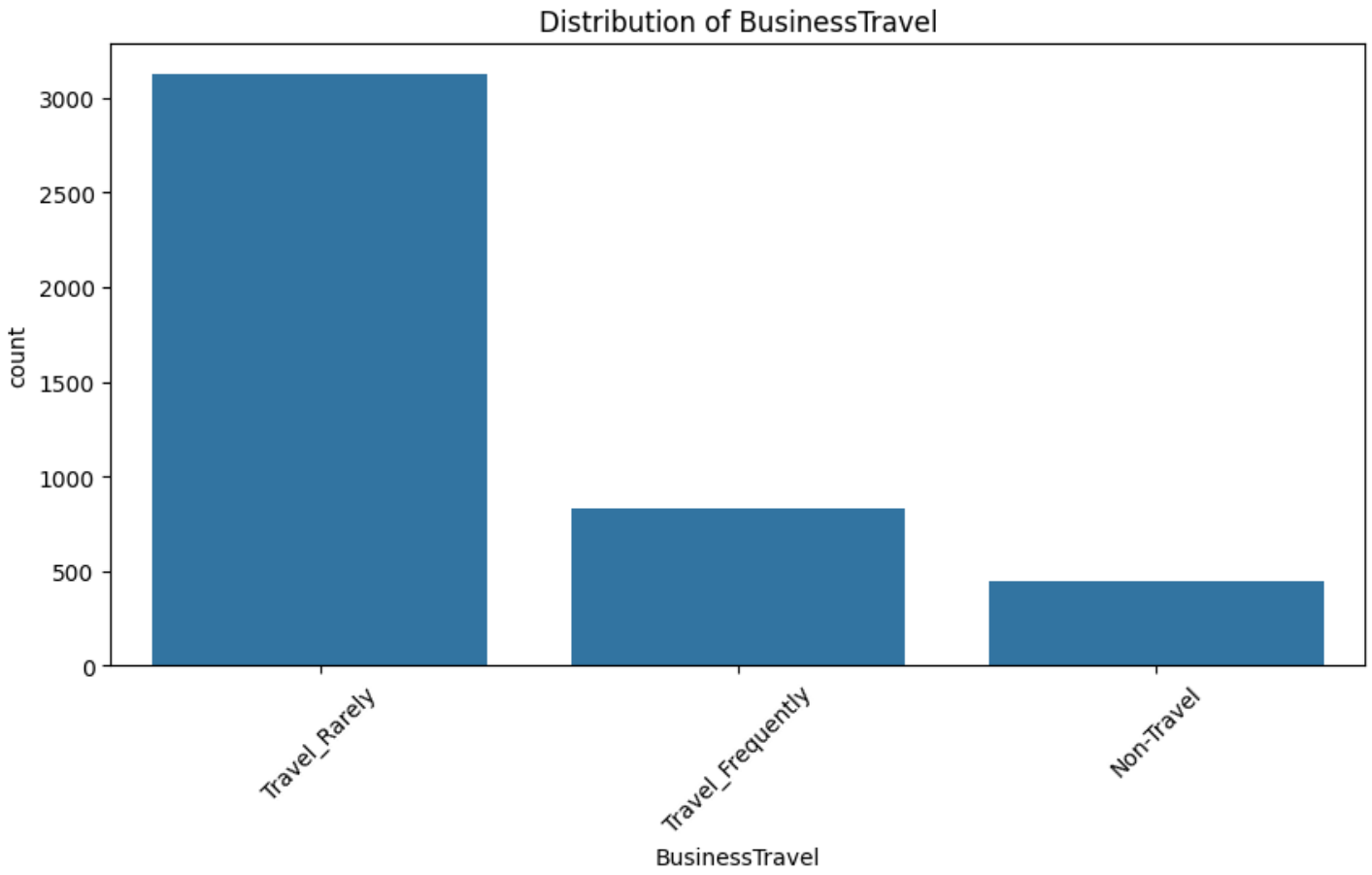
```
```

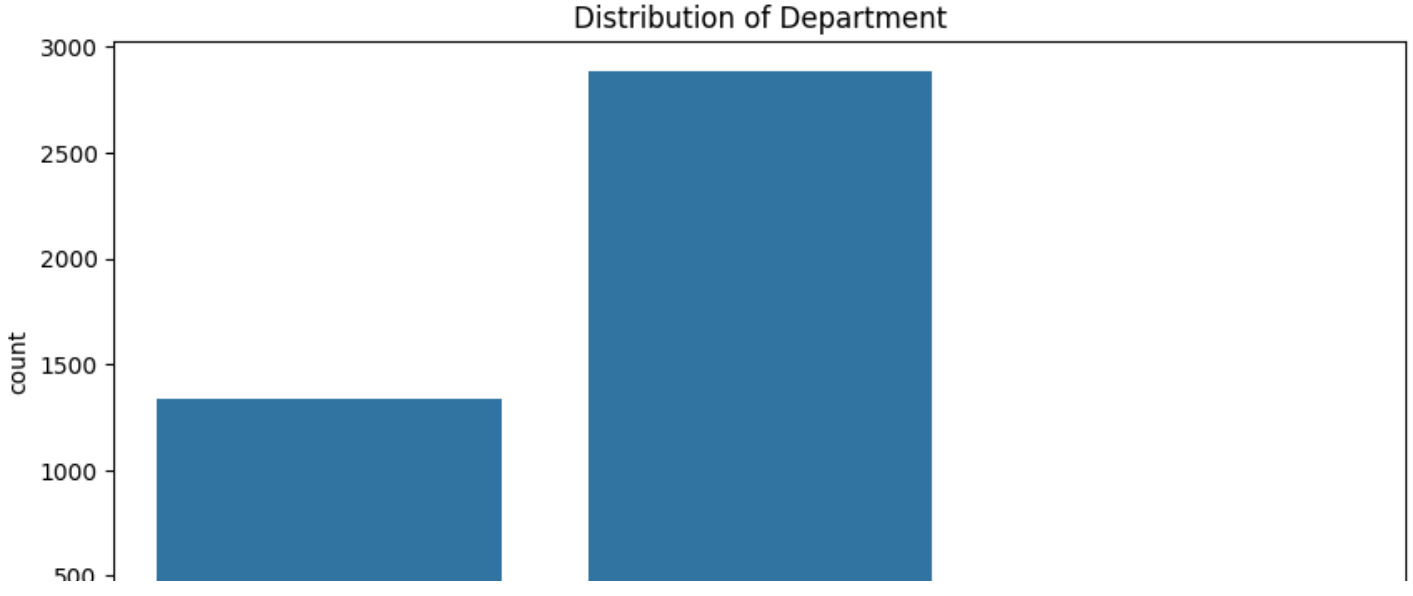
EDA

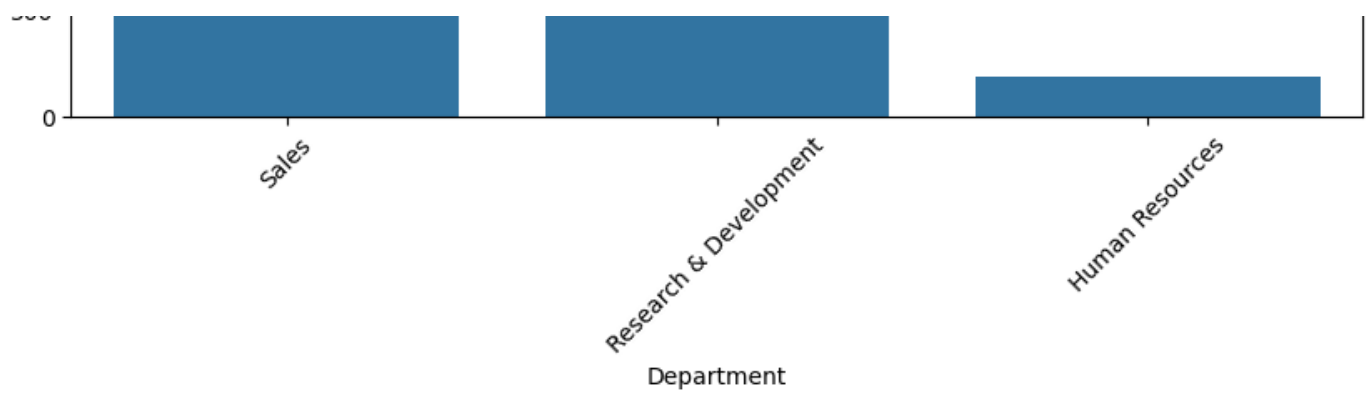
In [ ]:

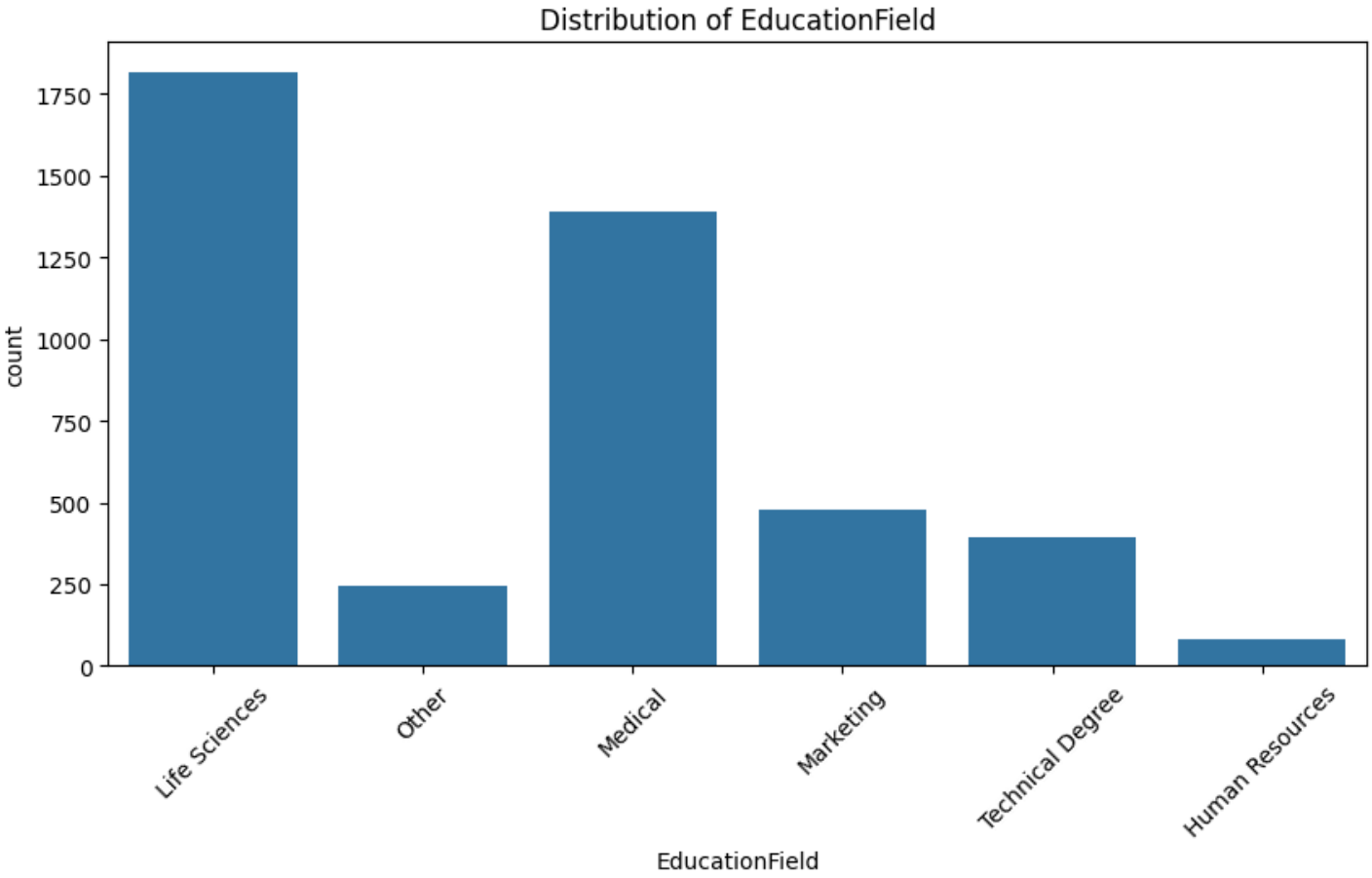
```
# Distribution of numerical features
data.hist(bins=20, figsize=(20, 15))
plt.show()

# Distribution of categorical features
categorical_columns = data.select_dtypes(include=['object']).columns

for col in categorical_columns:
    plt.figure(figsize=(10, 5))
    sns.countplot(x=col, data=data)
    plt.title(f'Distribution of {col}')
    plt.xticks(rotation=45)
    plt.show()

# Analyze the target variable 'Attrition'
sns.countplot(x='Attrition', data=data)
plt.title('Attrition Count')
plt.show()
```

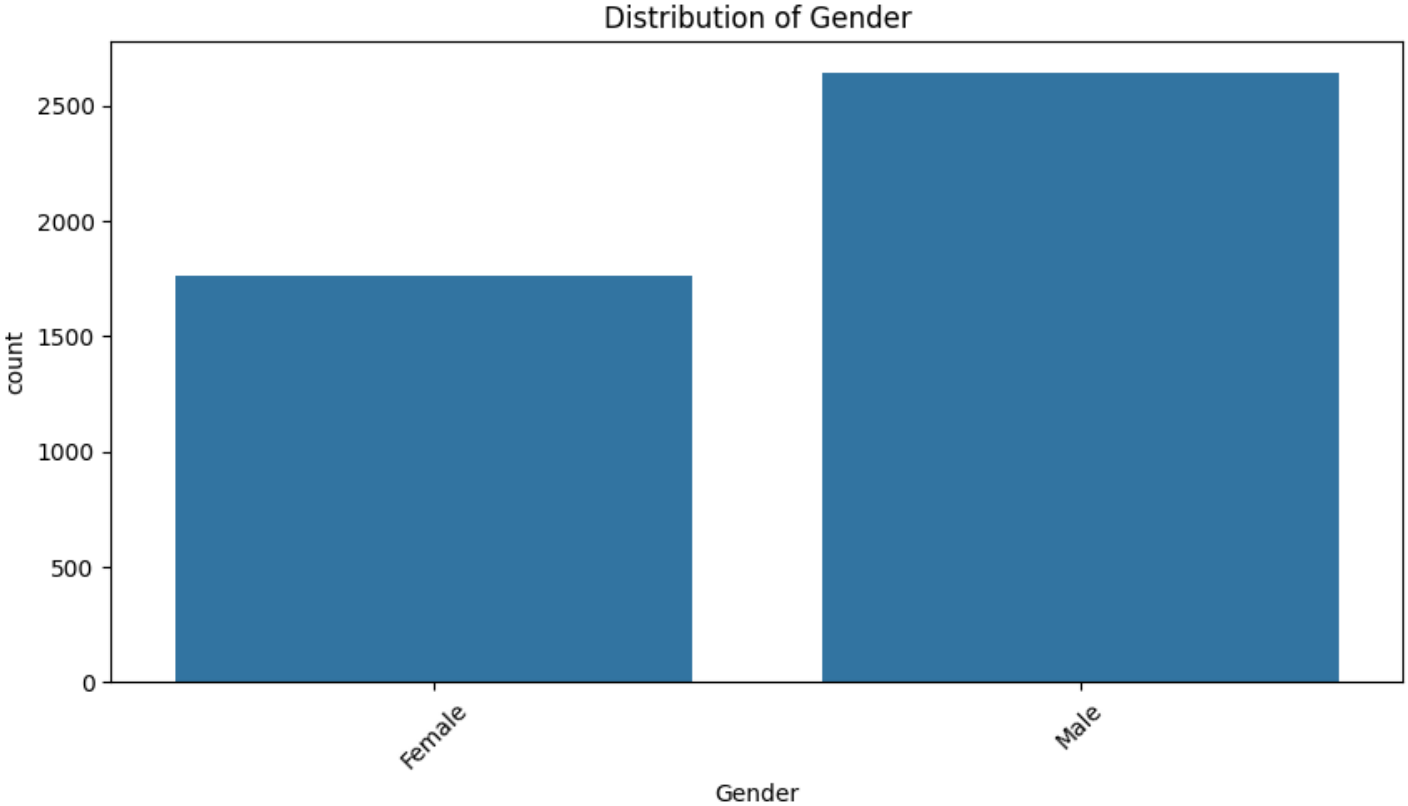

3000
2500
2000
1500
1000
500
0
count
No
Yes
Attrition

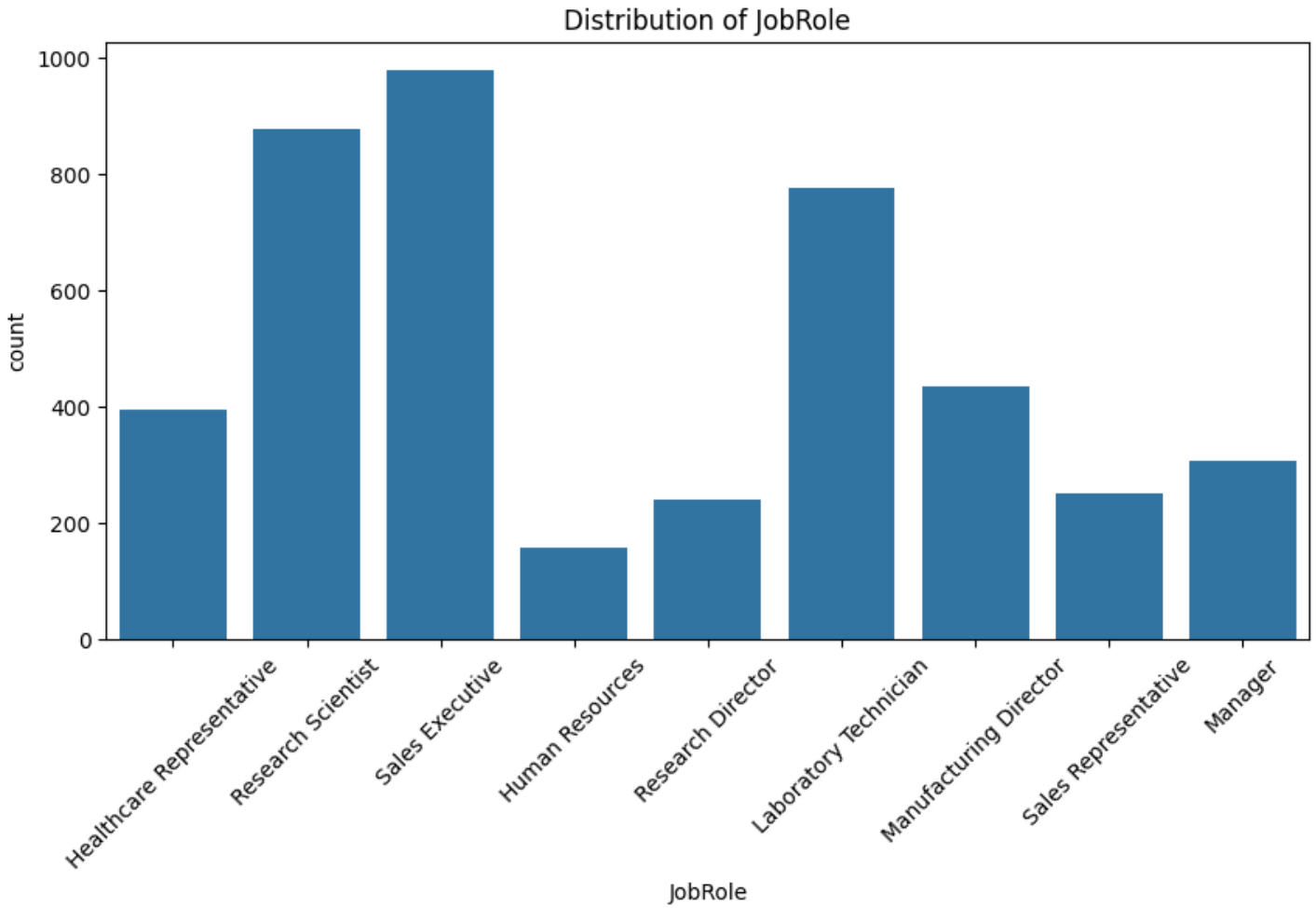

Distribution of BusinessTravel
3000
2500
2000
1500
1000
500
0
count
Travel_Rarely
Travel_Frequently
Non-Travel
BusinessTravel

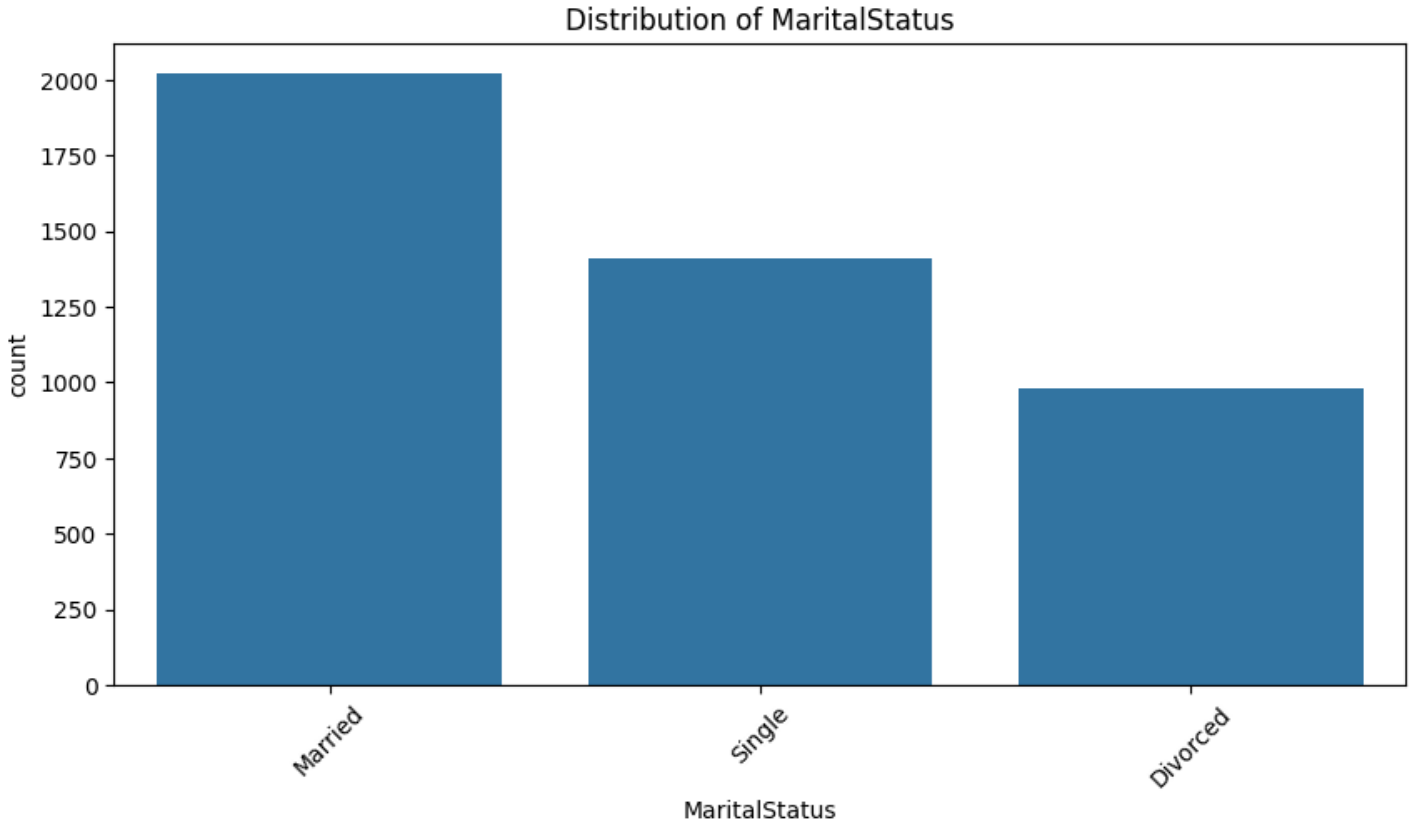

Distribution of Department
3000
2500
2000
1500
1000
500
count

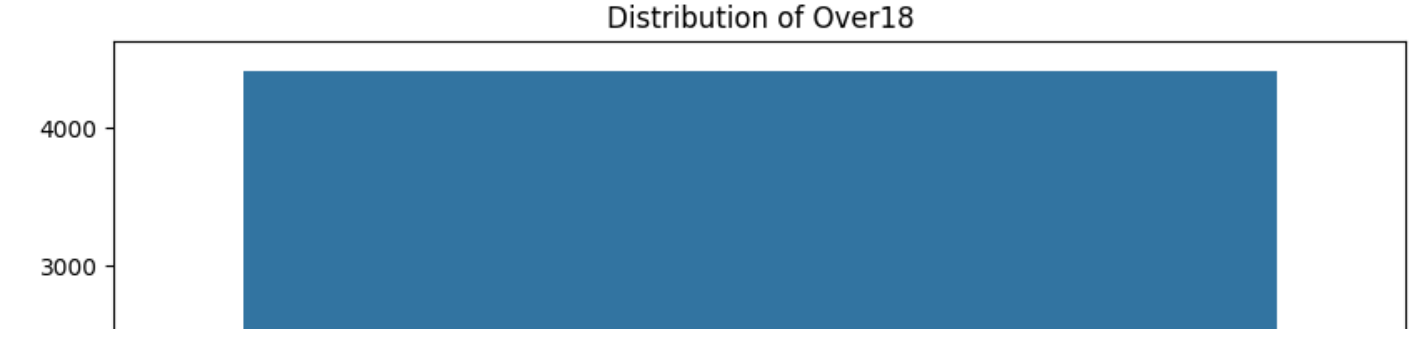

0
Sales
Research & Development
Human Resources
Department

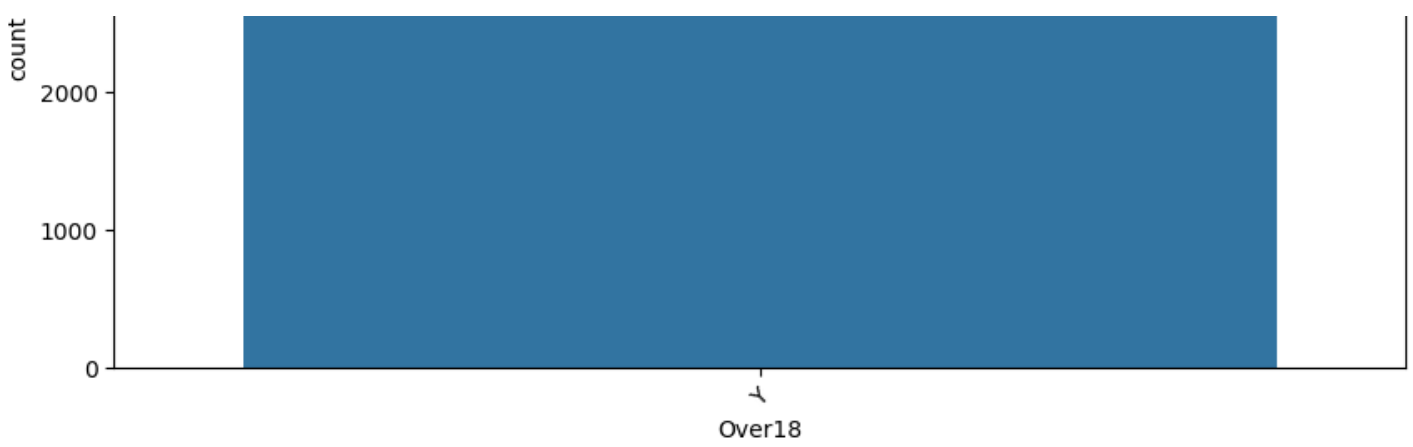

Distribution of EducationField
1750
1500
1250
1000
750
500
250
0
count
Life Sciences
Other
Medical
Marketing
Technical Degree
Human Resources
EducationField

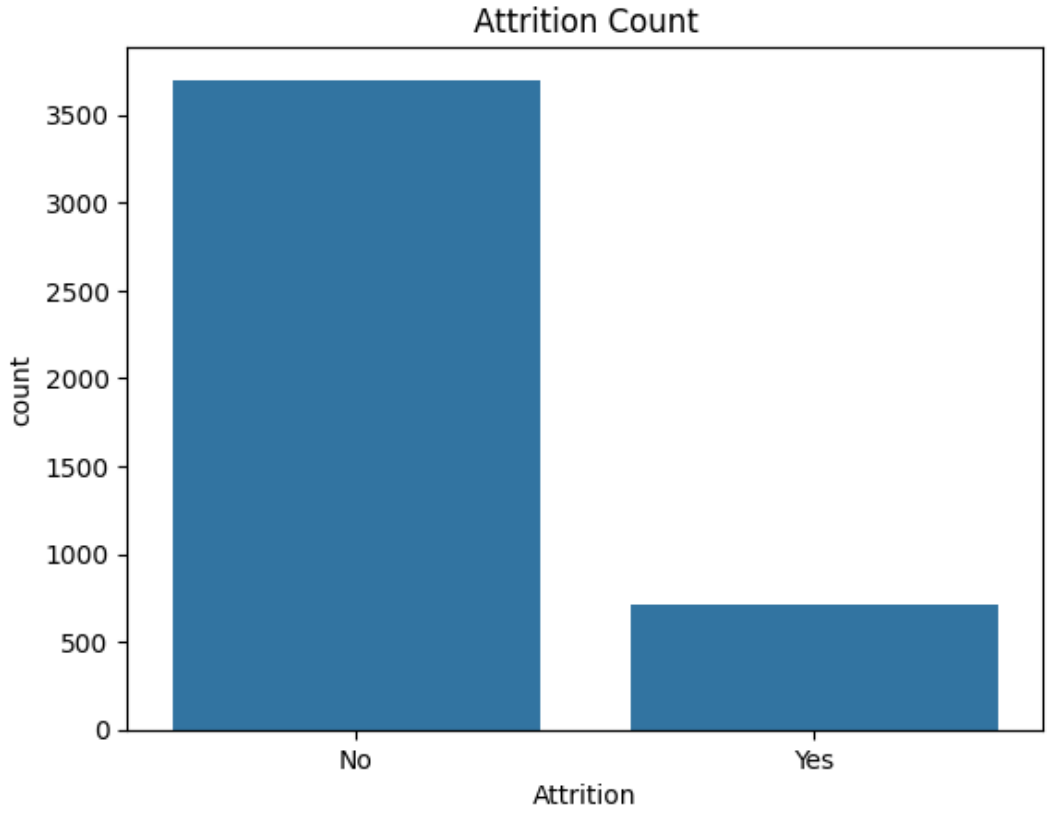

Distribution of Gender
2500
2000
1500
1000
500
0
count
Female
Male
Gender


Distribution of JobRole
count
JobRole
Healthcare Representative
Research Scientist
Sales Executive
Human Resources
Research Director
Laboratory Technician
Manufacturing Director
Sales Representative
Manager


Distribution of MaritalStatus
count
MaritalStatus
Married
Single
Divorced

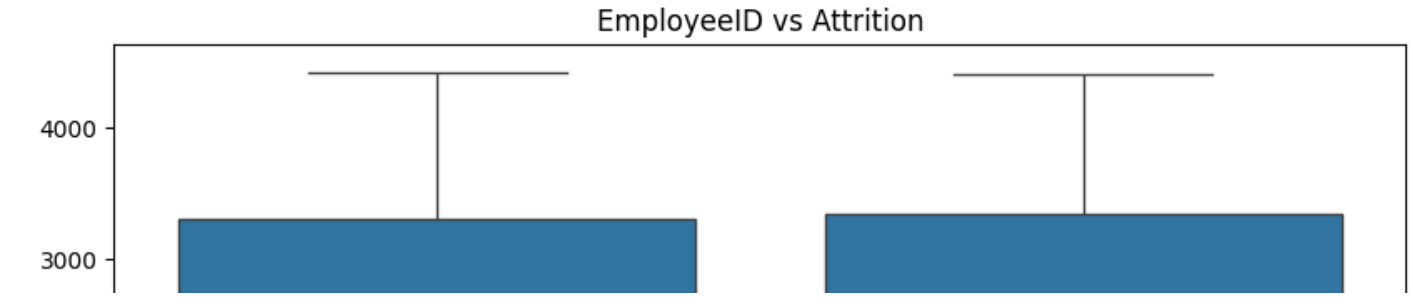

Distribution of Over18

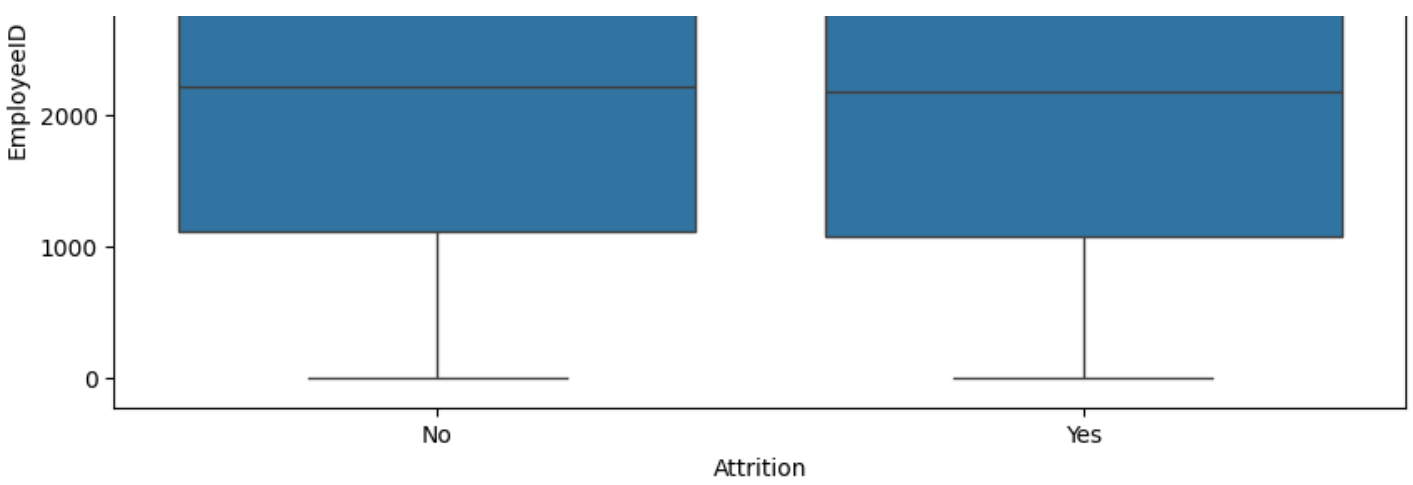


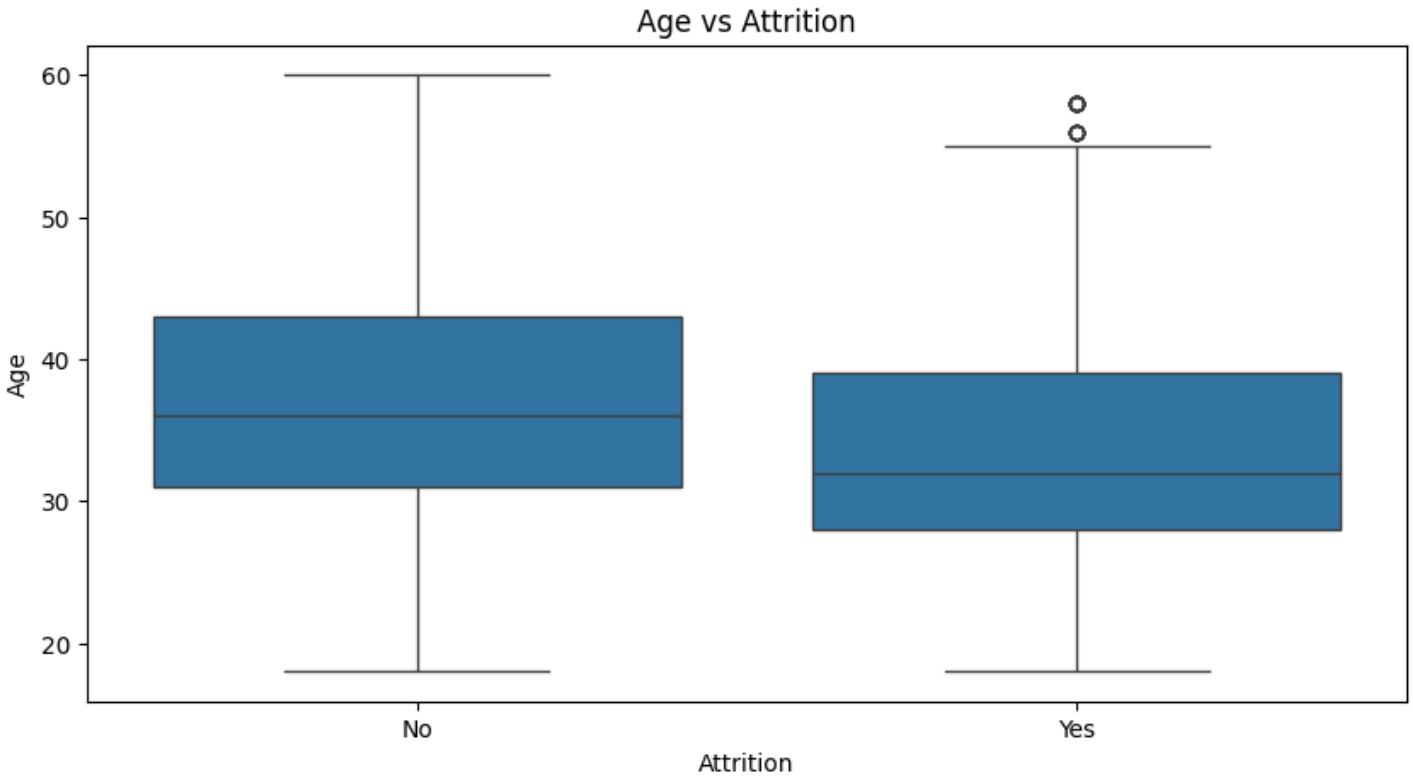


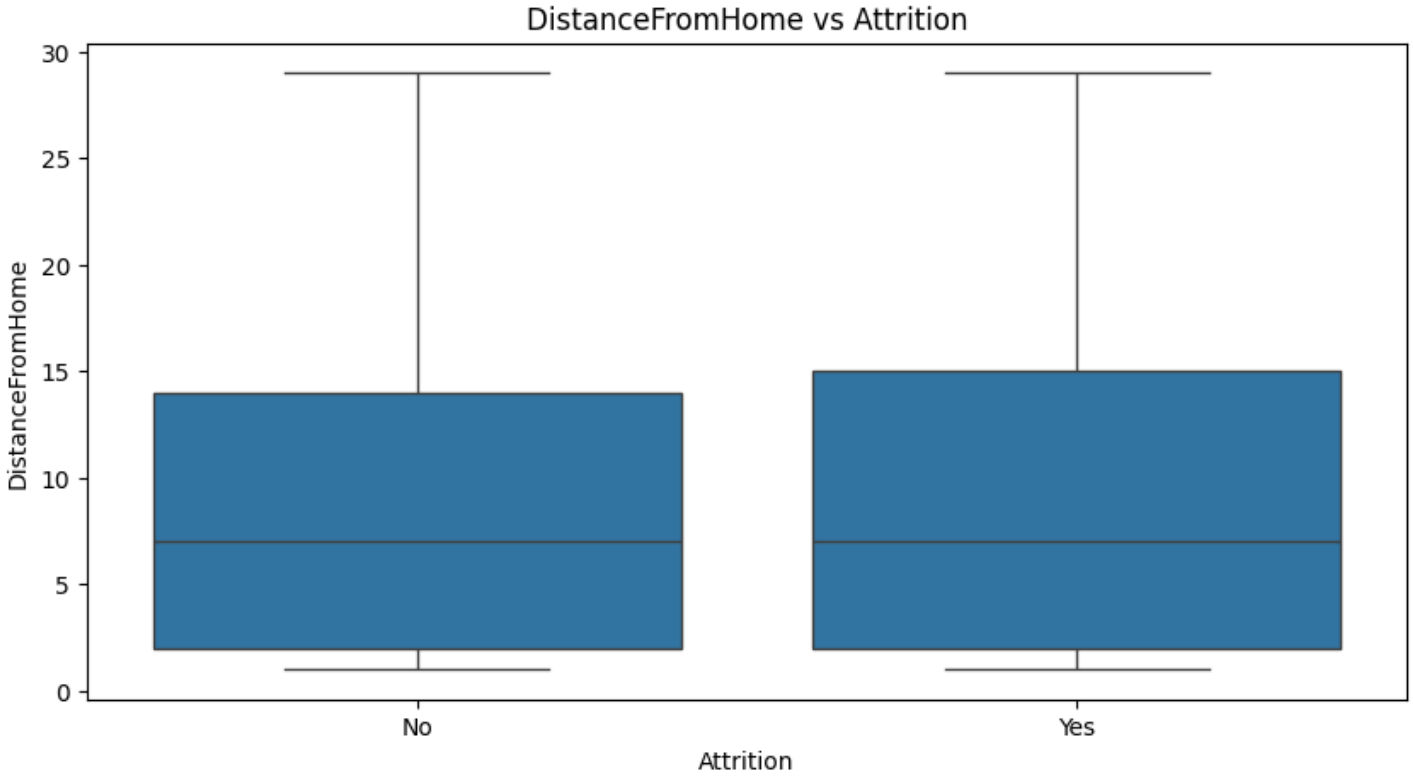
In [ ]:

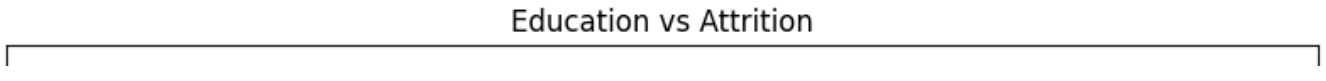
```
# Numerical features vs Attrition
numerical_columns = data.select_dtypes(include=['number']).columns

for col in numerical_columns:
    plt.figure(figsize=(10, 5))
    sns.boxplot(x='Attrition', y=col, data=data)
    plt.title(f'{col} vs Attrition')
    plt.show()

# Categorical features vs Attrition
for col in categorical_columns:
    plt.figure(figsize=(10, 5))
    sns.countplot(x=col, hue='Attrition', data=data)
    plt.title(f'{col} vs Attrition')
    plt.xticks(rotation=45)
    plt.show()
```

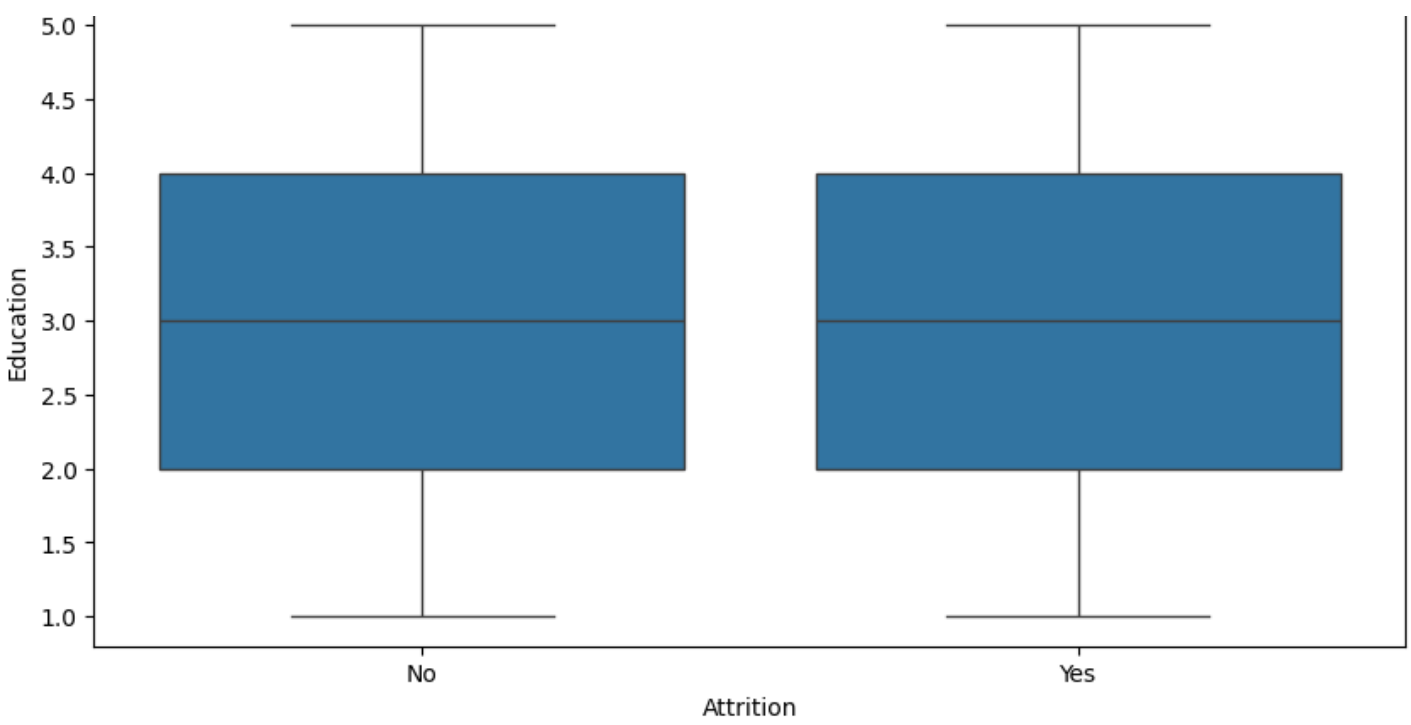


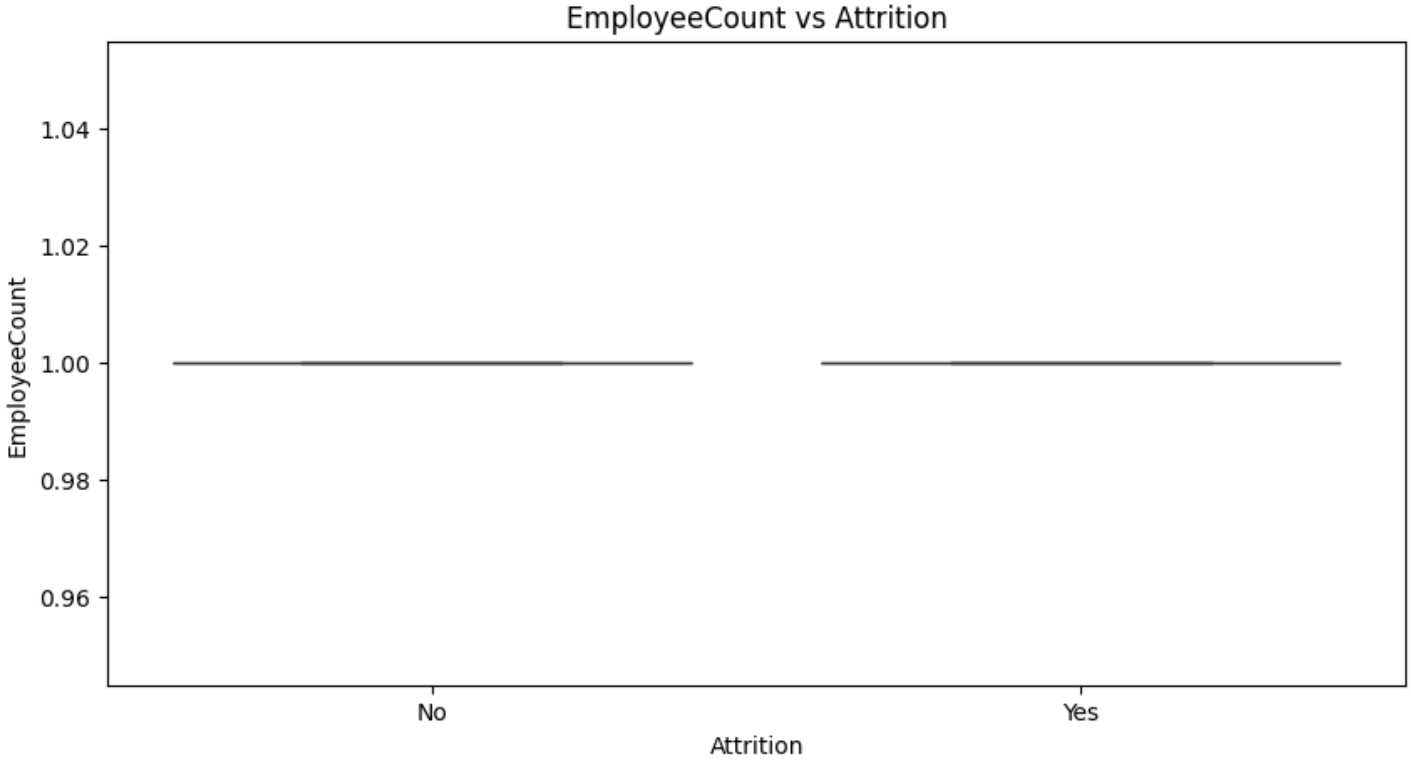

EmployeeID
2000
1000
0
No
Yes
Attrition

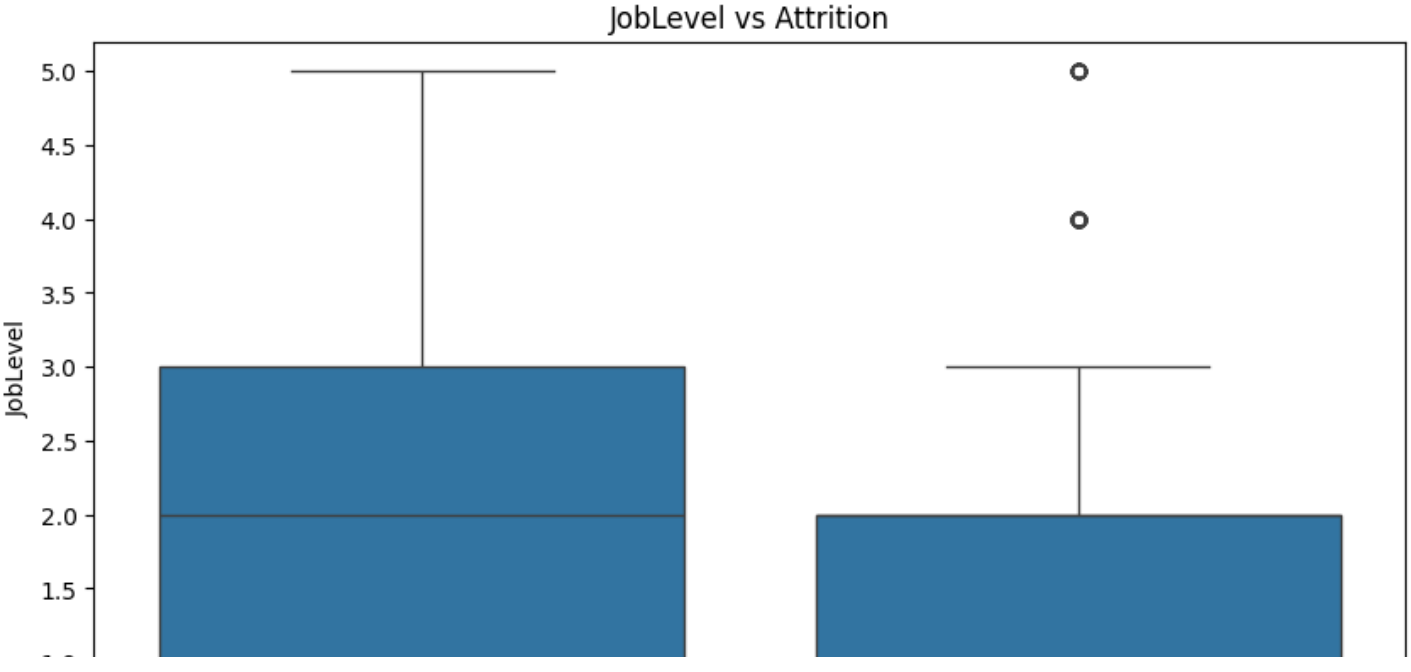

Age vs Attrition
Age
60
50
40
30
20
No
Yes
Attrition


DistanceFromHome vs Attrition
DistanceFromHome
30
25
20
15
10
5
0
No
Yes
Attrition

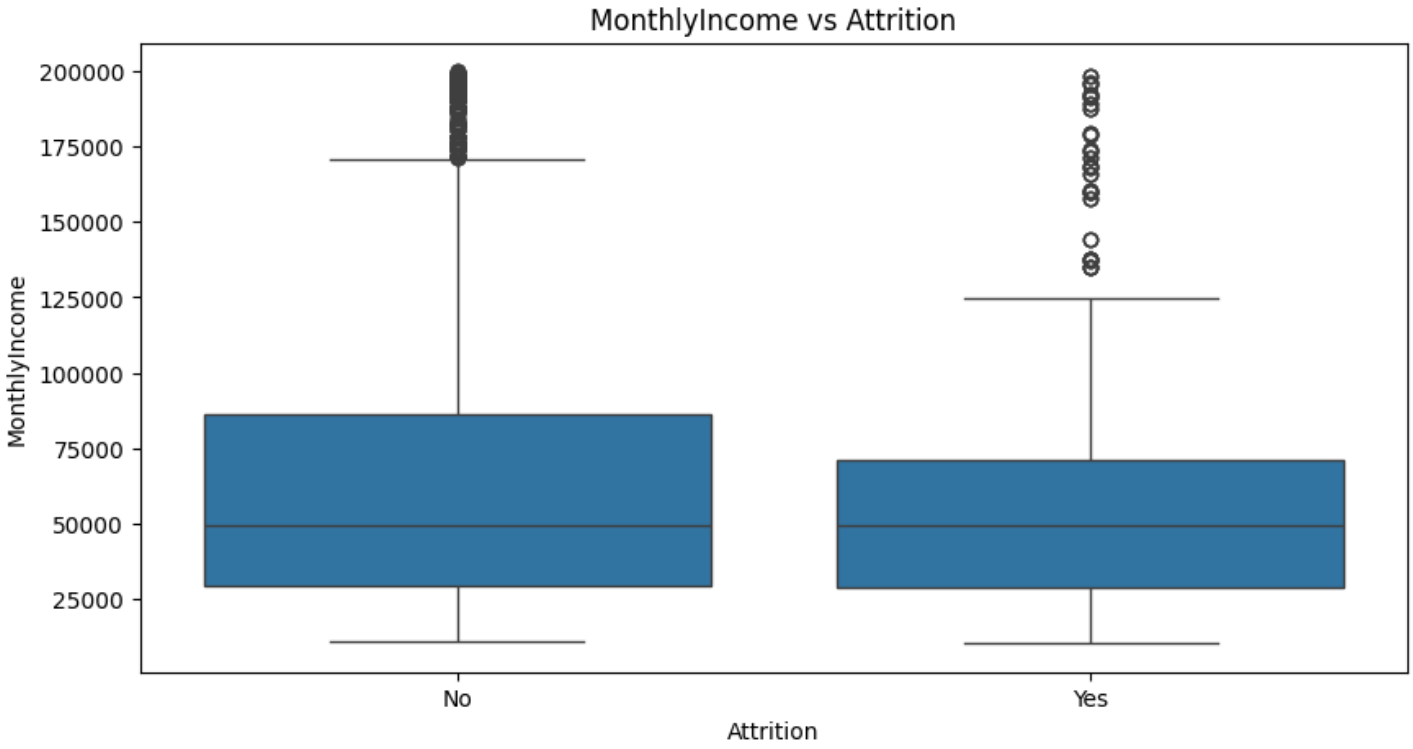

Education vs Attrition

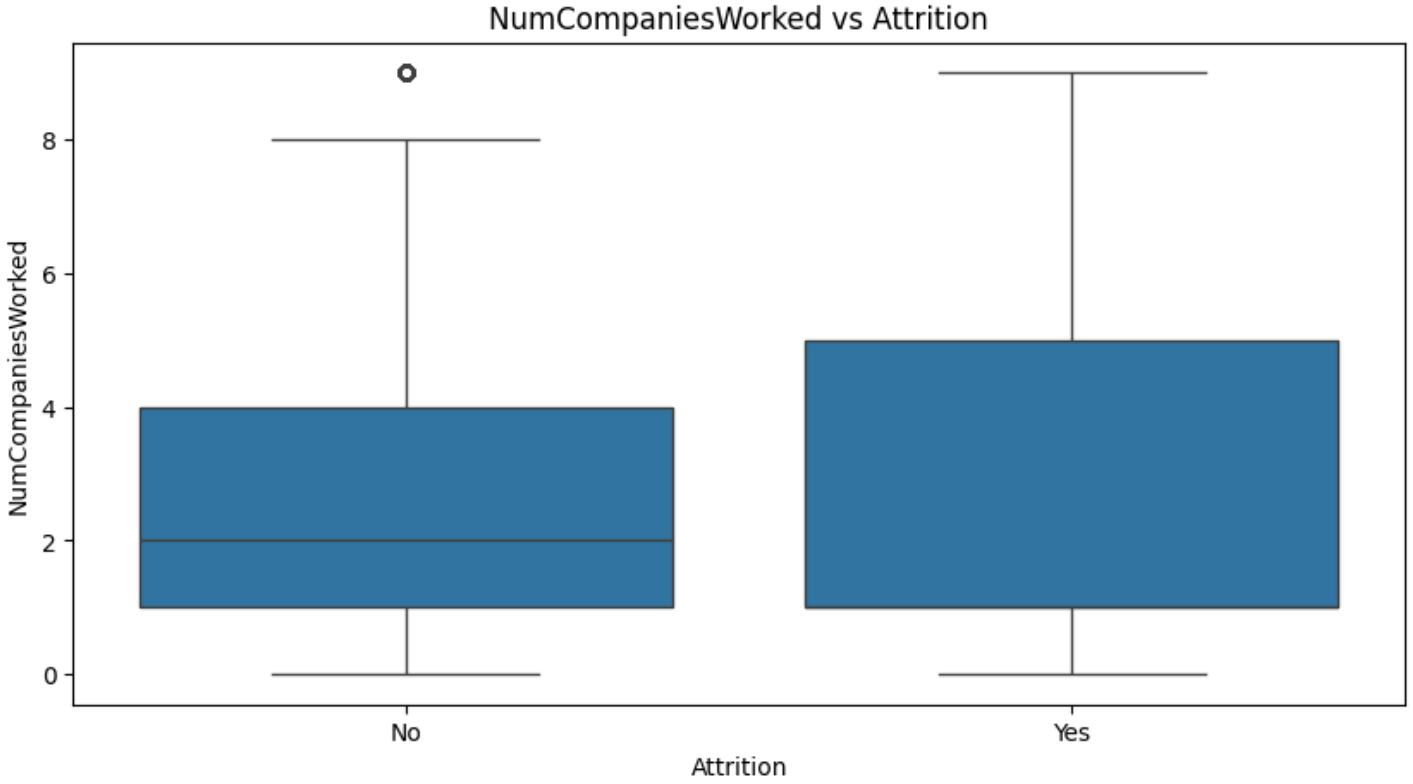

5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
Education
No
Yes
Attrition

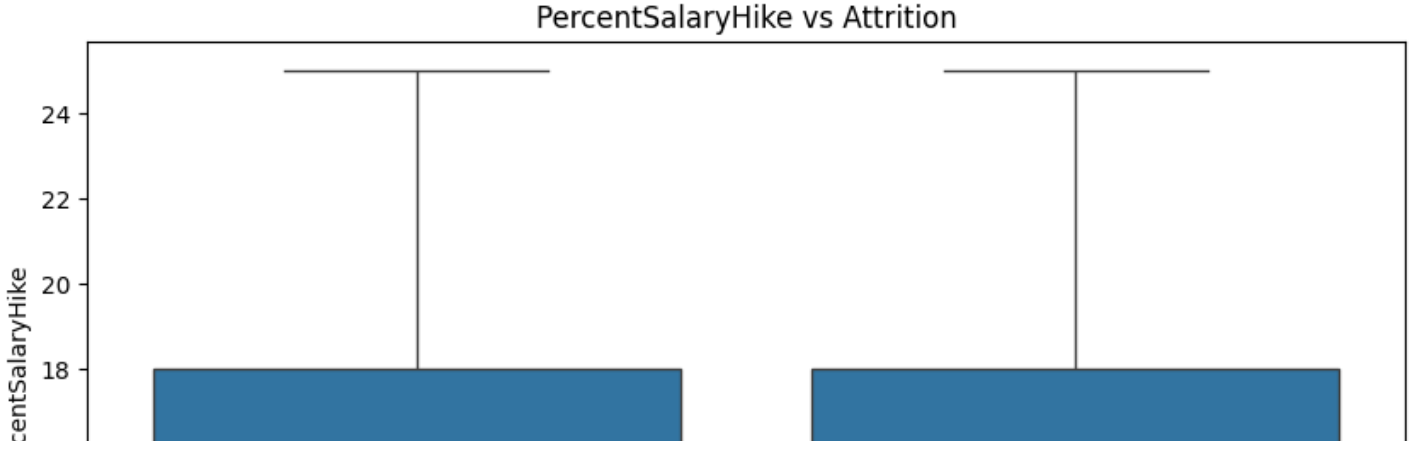

EmployeeCount vs Attrition
1.04
1.02
1.00
0.98
0.96
EmployeeCount
No
Yes
Attrition

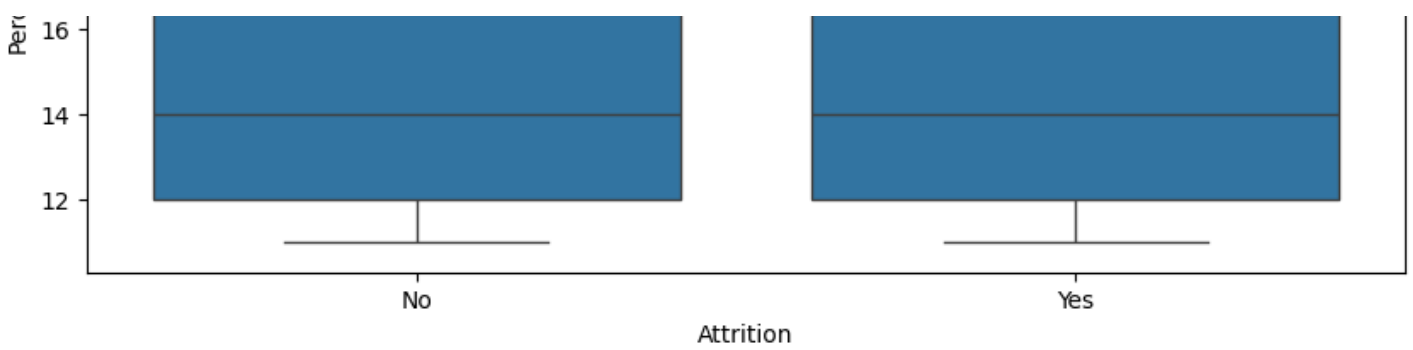

JobLevel vs Attrition
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
JobLevel

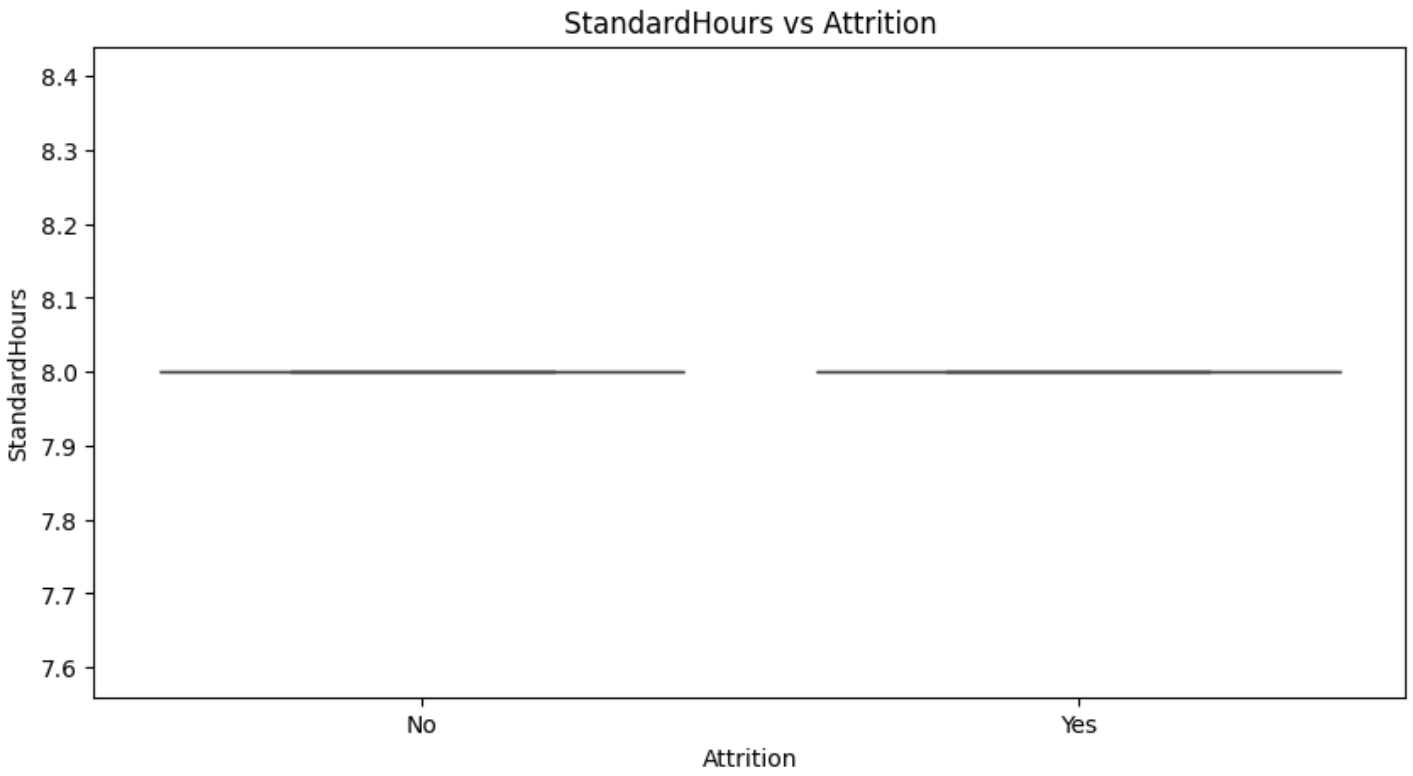

1.0
No
Yes
Attrition

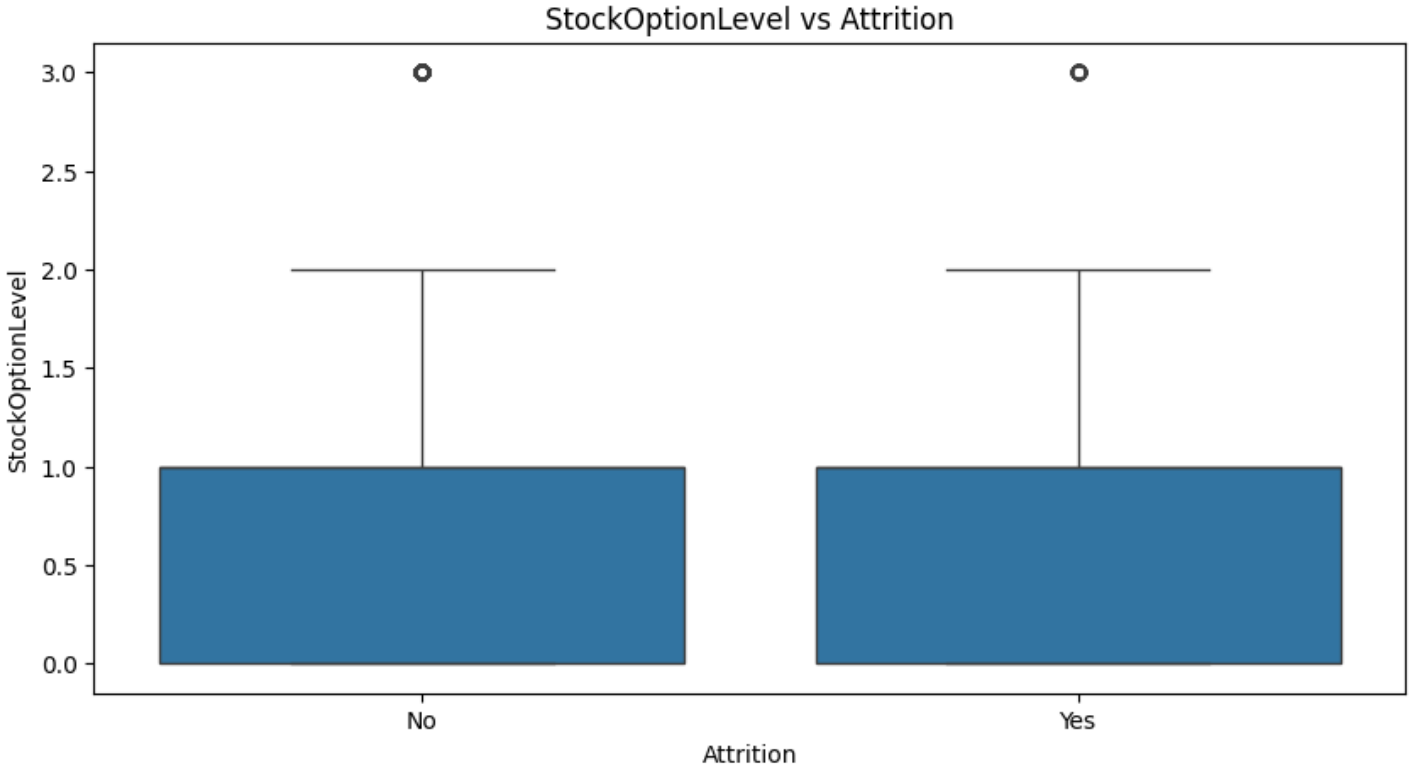

MonthlyIncome vs Attrition
MonthlyIncome
200000
175000
150000
125000
100000
75000
50000
25000
No
Yes
Attrition

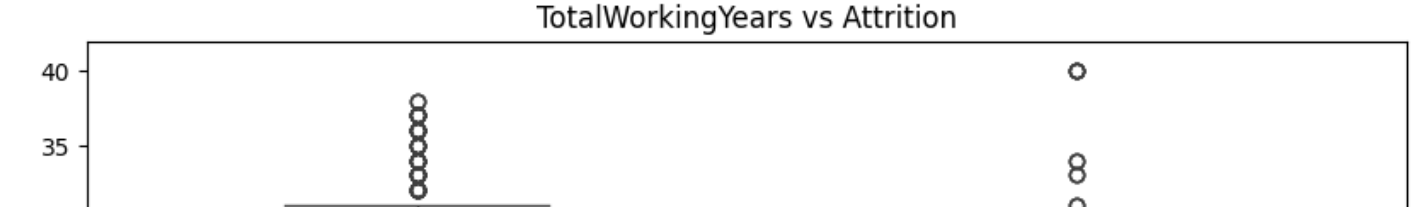

NumCompaniesWorked vs Attrition
NumCompaniesWorked
8
6
4
2
0
No
Yes
Attrition

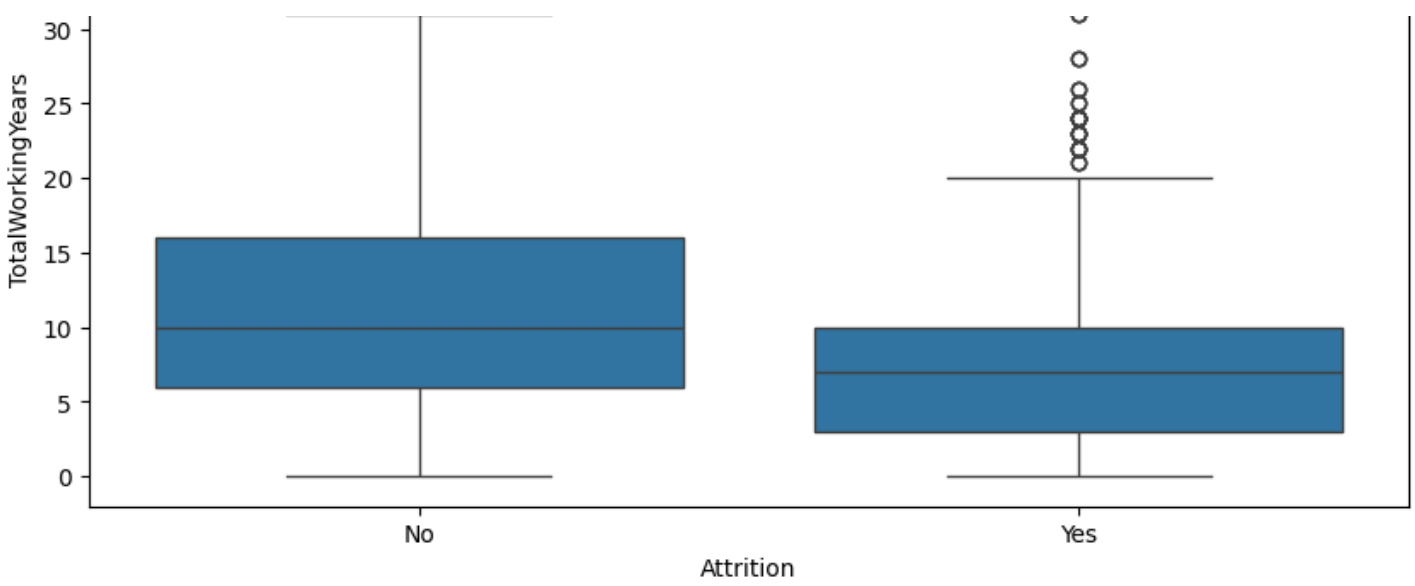

PercentSalaryHike vs Attrition
24
22
20
18

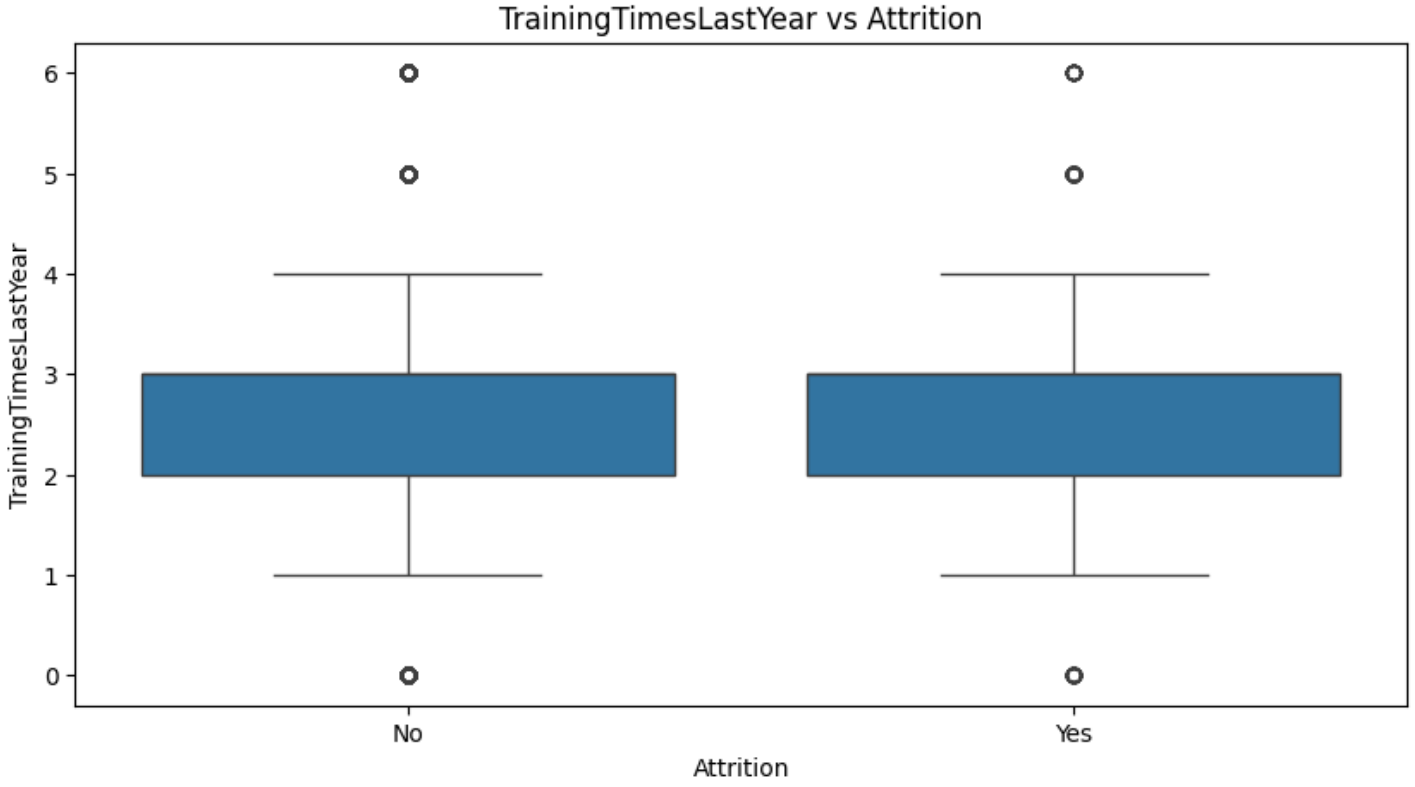

16
14
12
No
Yes
Attrition

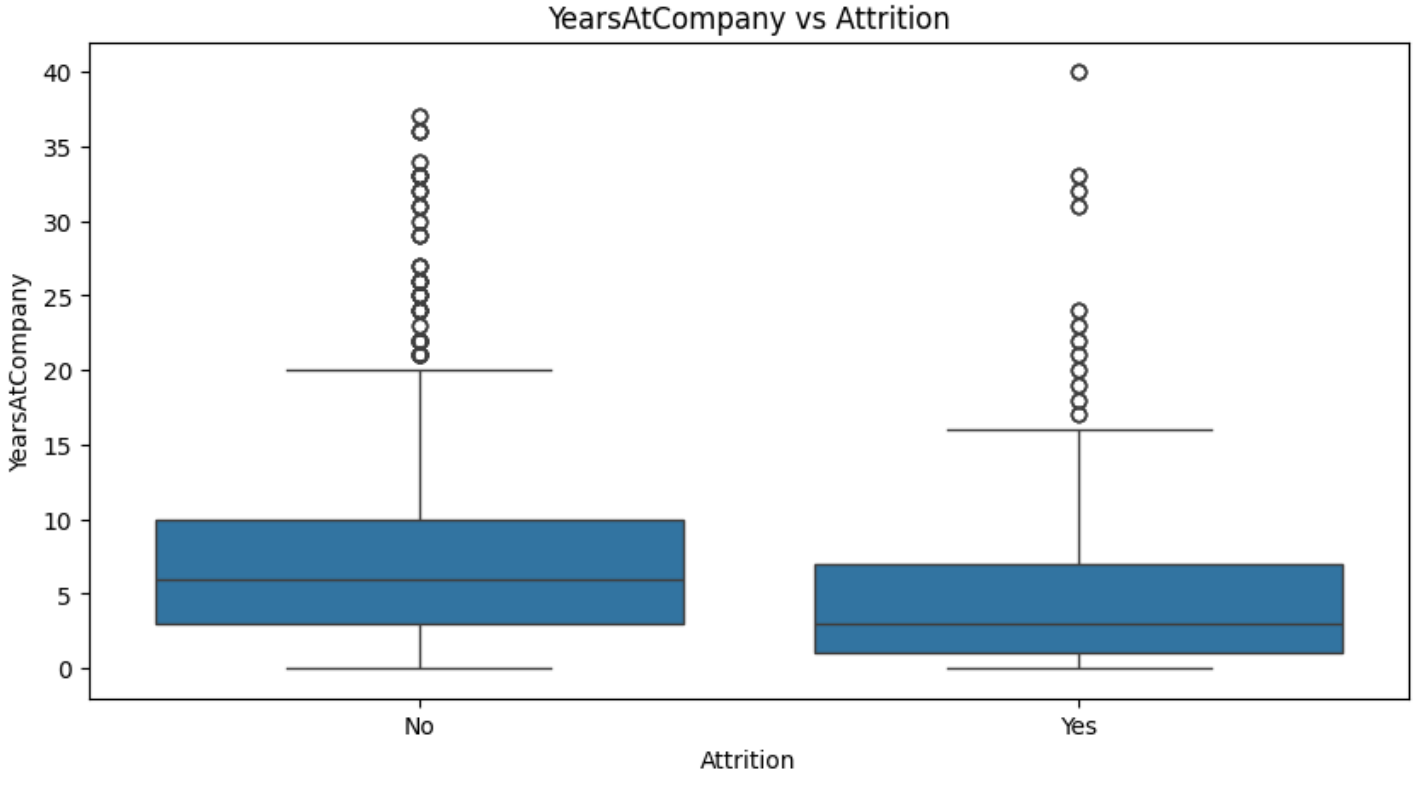

StandardHours vs Attrition
8.4
8.3
8.2
8.1
8.0
7.9
7.8
7.7
7.6
StandardHours
No
Yes
Attrition

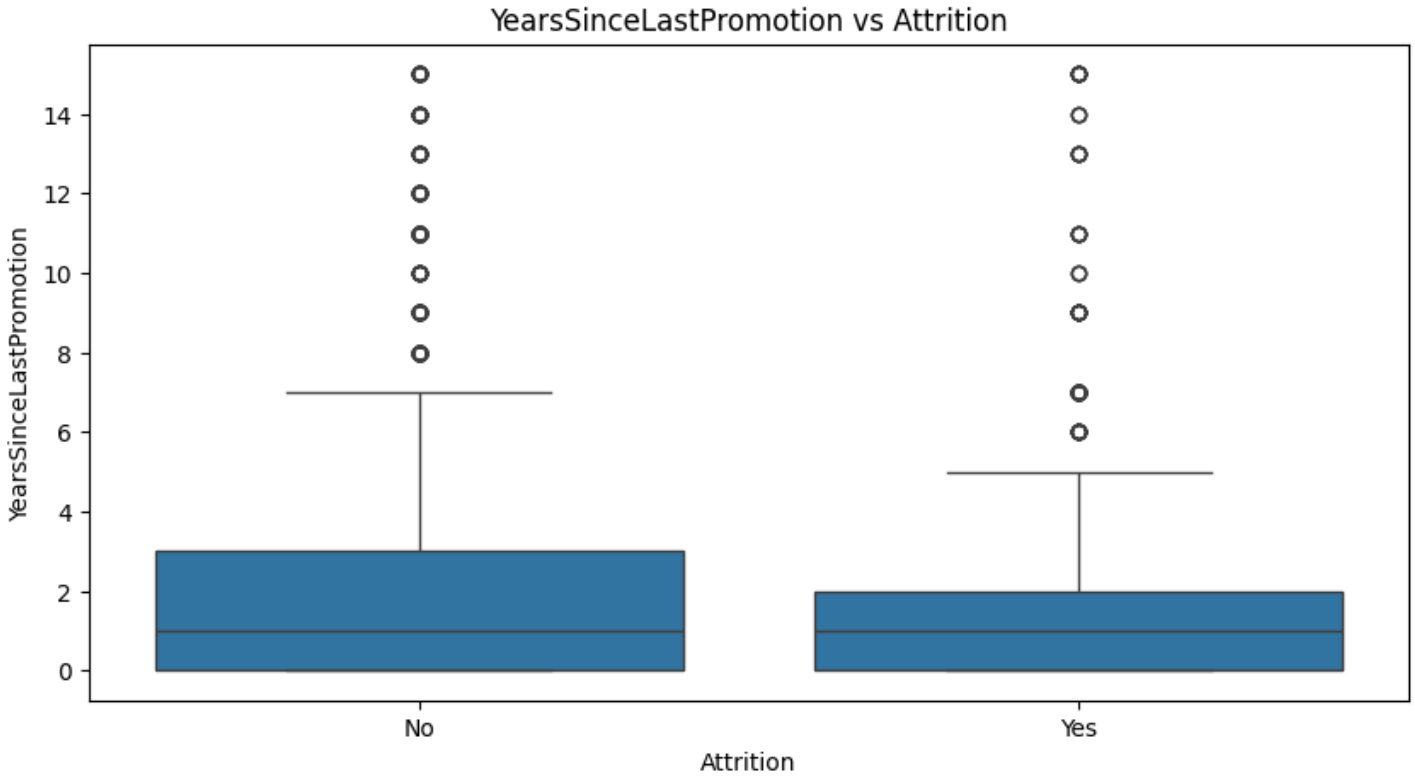

StockOptionLevel vs Attrition
3.0
2.5
2.0
1.5
1.0
0.5
0.0
StockOptionLevel
No
Yes
Attrition

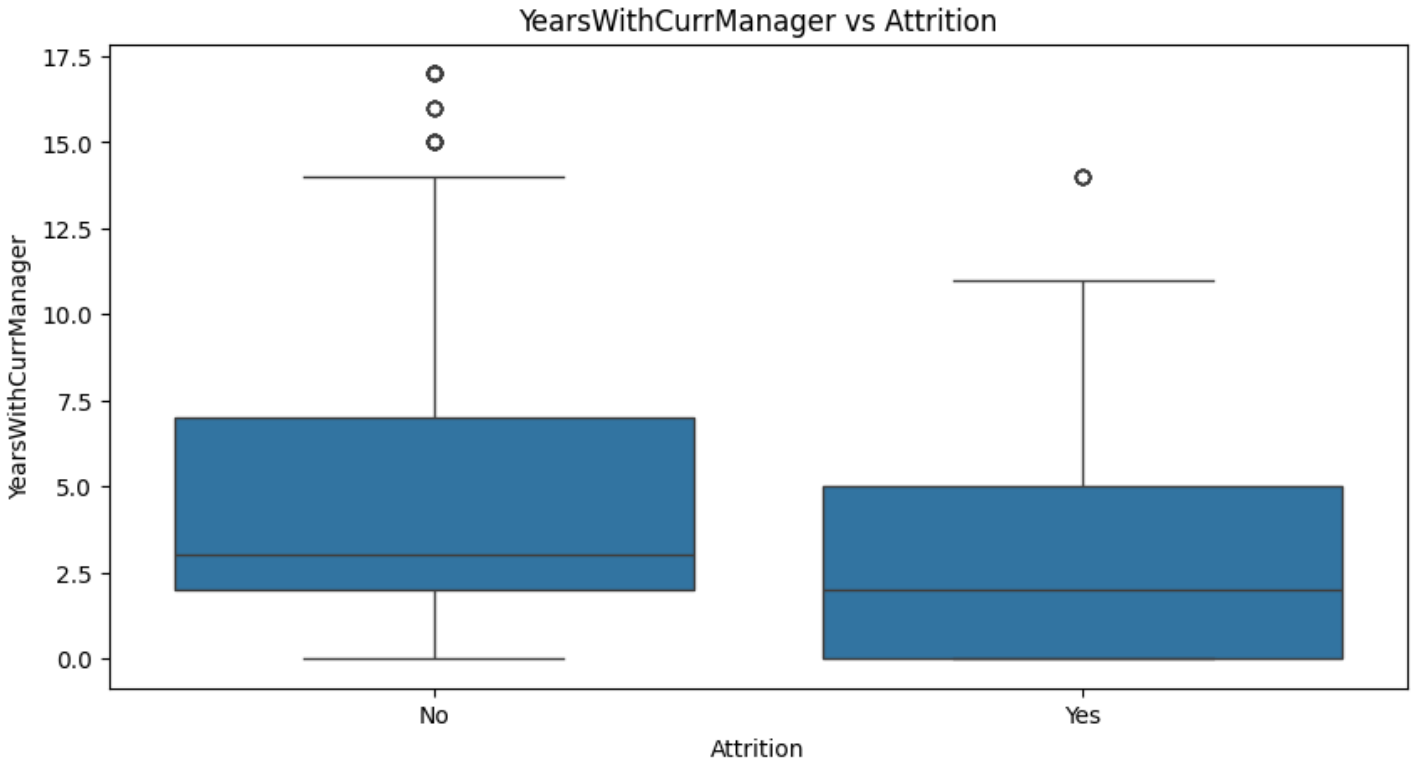

TotalWorkingYears vs Attrition
40
35

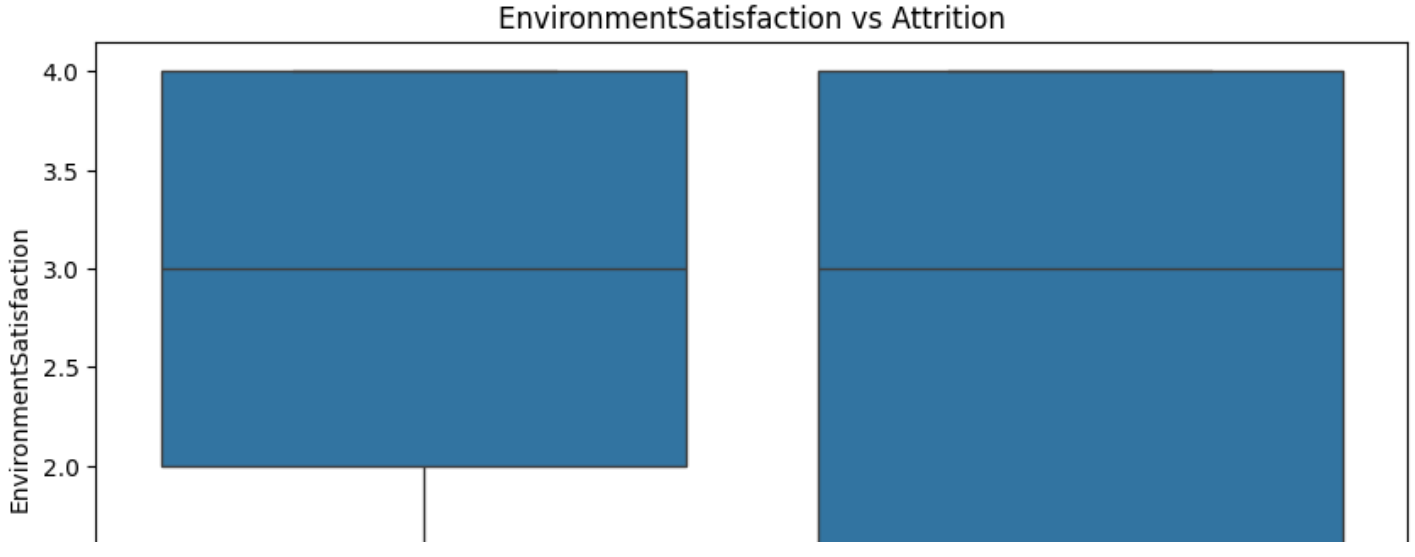

TotalWorkingYears
30
25
20
15
10
5
0
No
Yes
Attrition

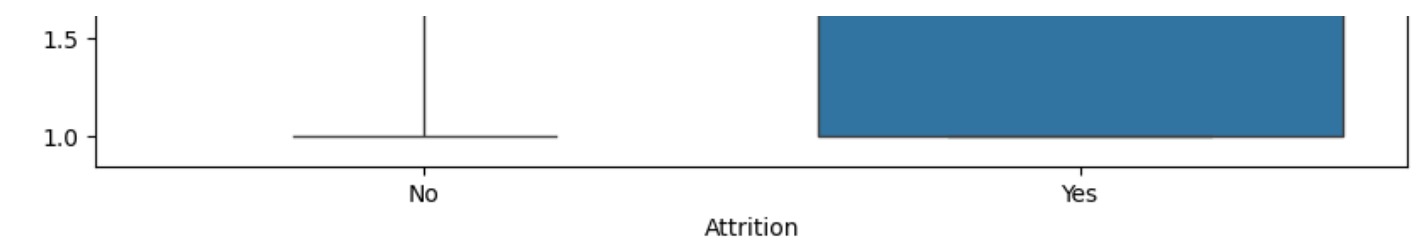

TrainingTimesLastYear vs Attrition
TrainingTimesLastYear
6
5
4
3
2
1
0
No
Yes
Attrition

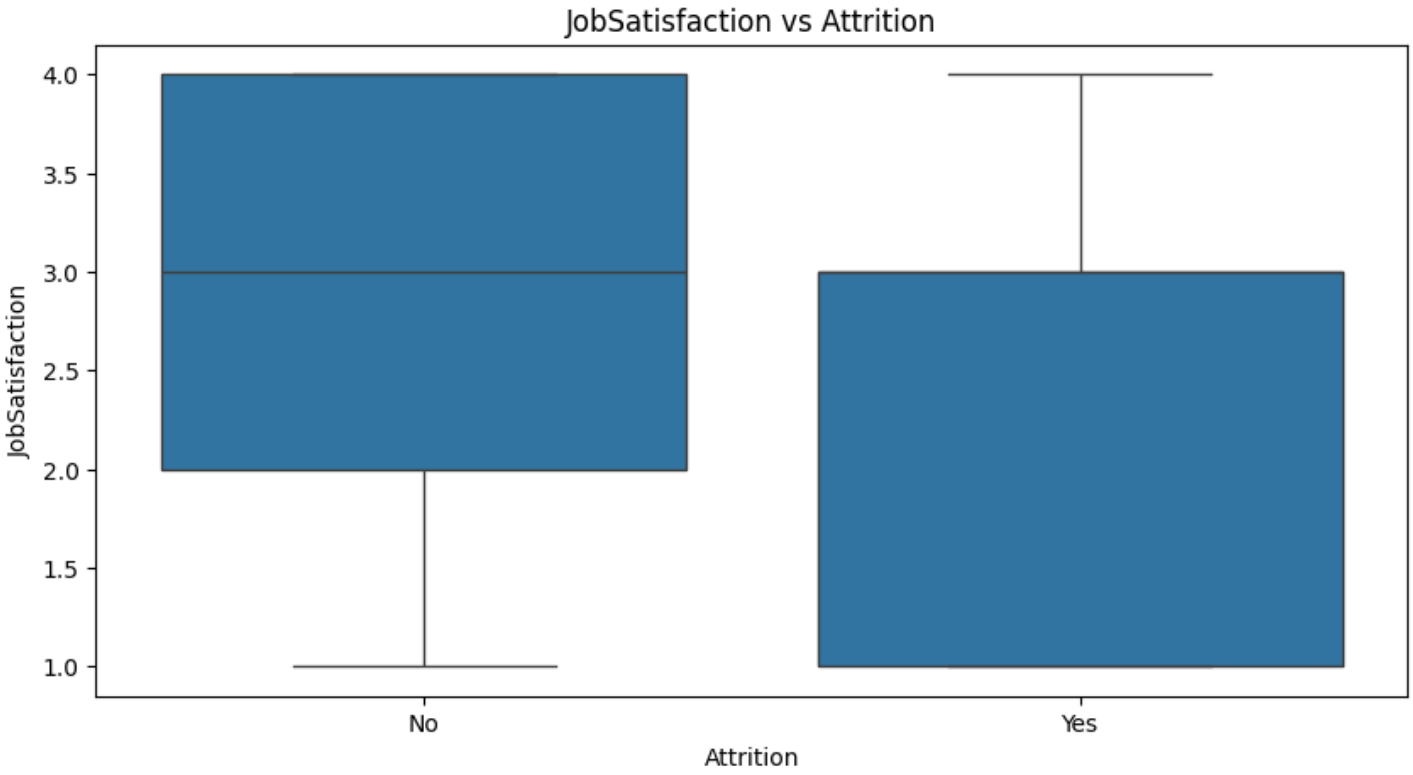

YearsAtCompany vs Attrition
YearsAtCompany
40
35
30
25
20
15
10
5
0
No
Yes
Attrition

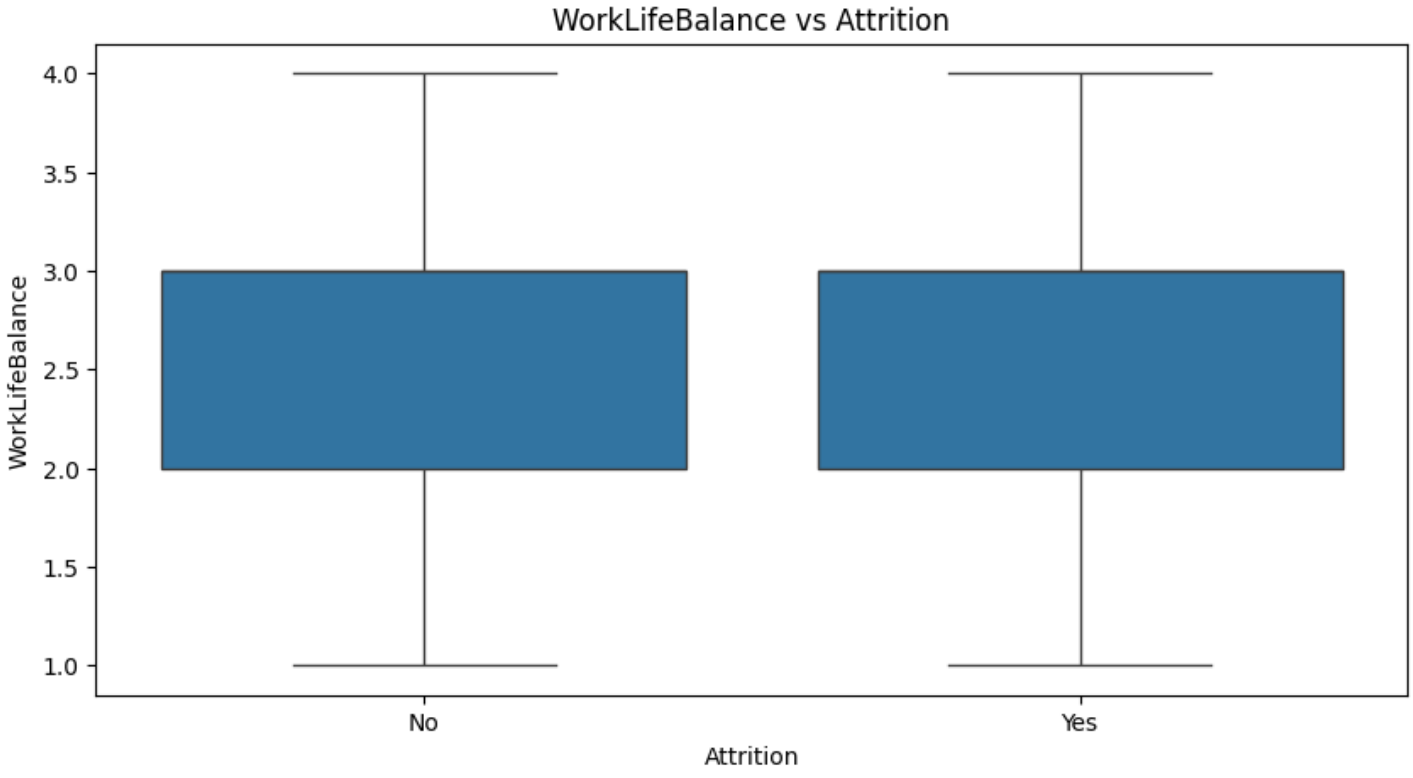

YearsSinceLastPromotion vs Attrition
YearsSinceLastPromotion
0
2
4
6
8
10
12
14
No
Yes
Attrition

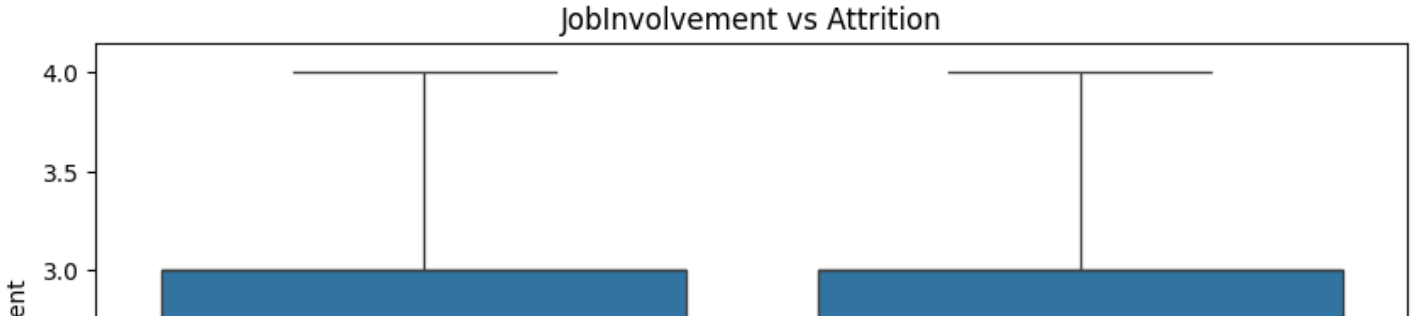

YearsWithCurrManager vs Attrition
YearsWithCurrManager
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
No
Yes
Attrition

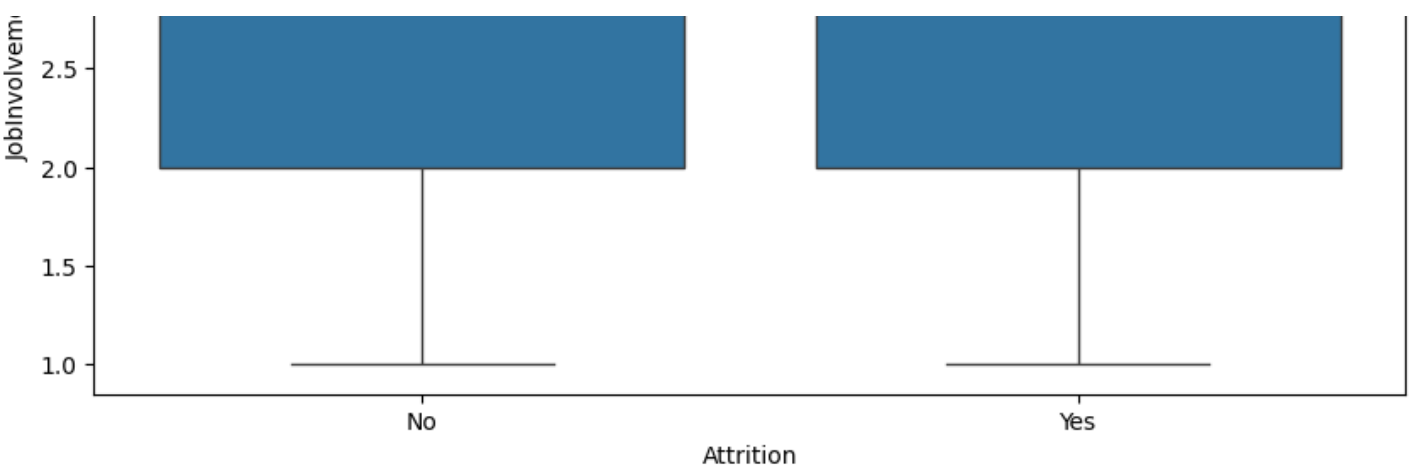

EnvironmentSatisfaction vs Attrition
EnvironmentSatisfaction
2.0
2.5
3.0
3.5
4.0

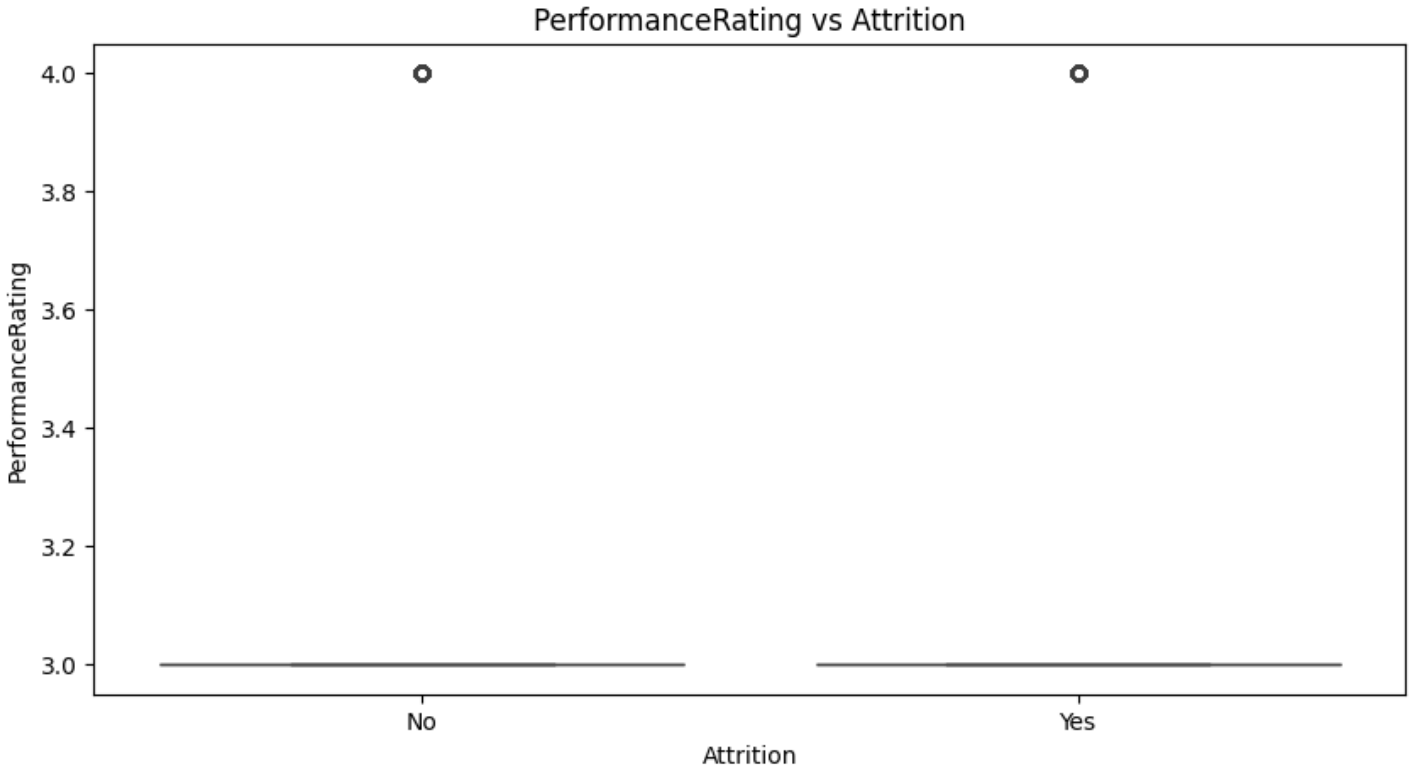

1.5
1.0
No
Yes
Attrition

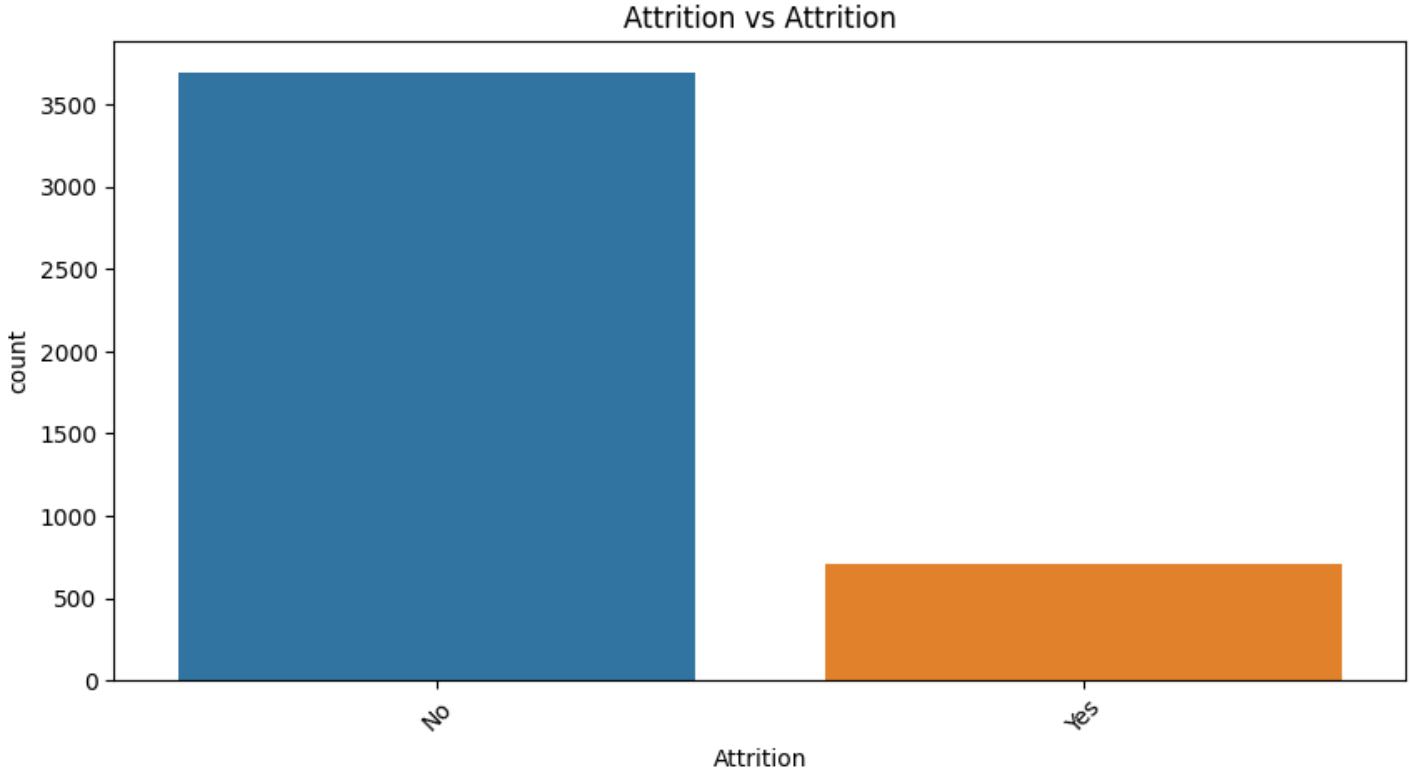

JobSatisfaction vs Attrition
JobSatisfaction
4.0
3.5
3.0
2.5
2.0
1.5
1.0
No
Yes
Attrition

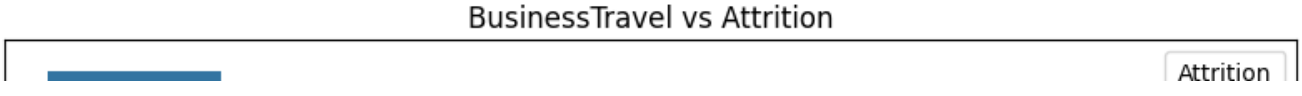

WorkLifeBalance vs Attrition
WorkLifeBalance
4.0
3.5
3.0
2.5
2.0
1.5
1.0
No
Yes
Attrition

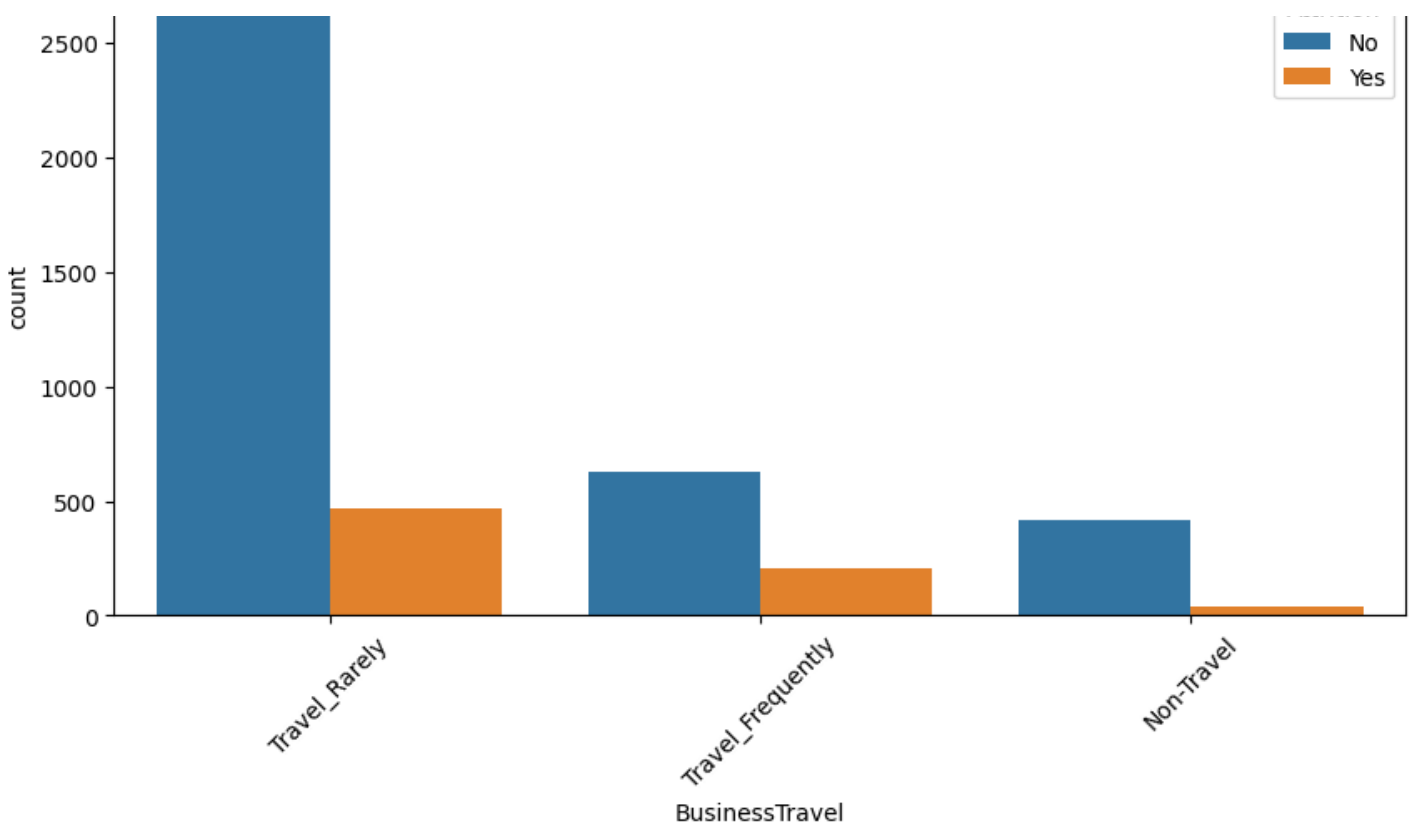

JobInvolvement vs Attrition
4.0
3.5
3.0

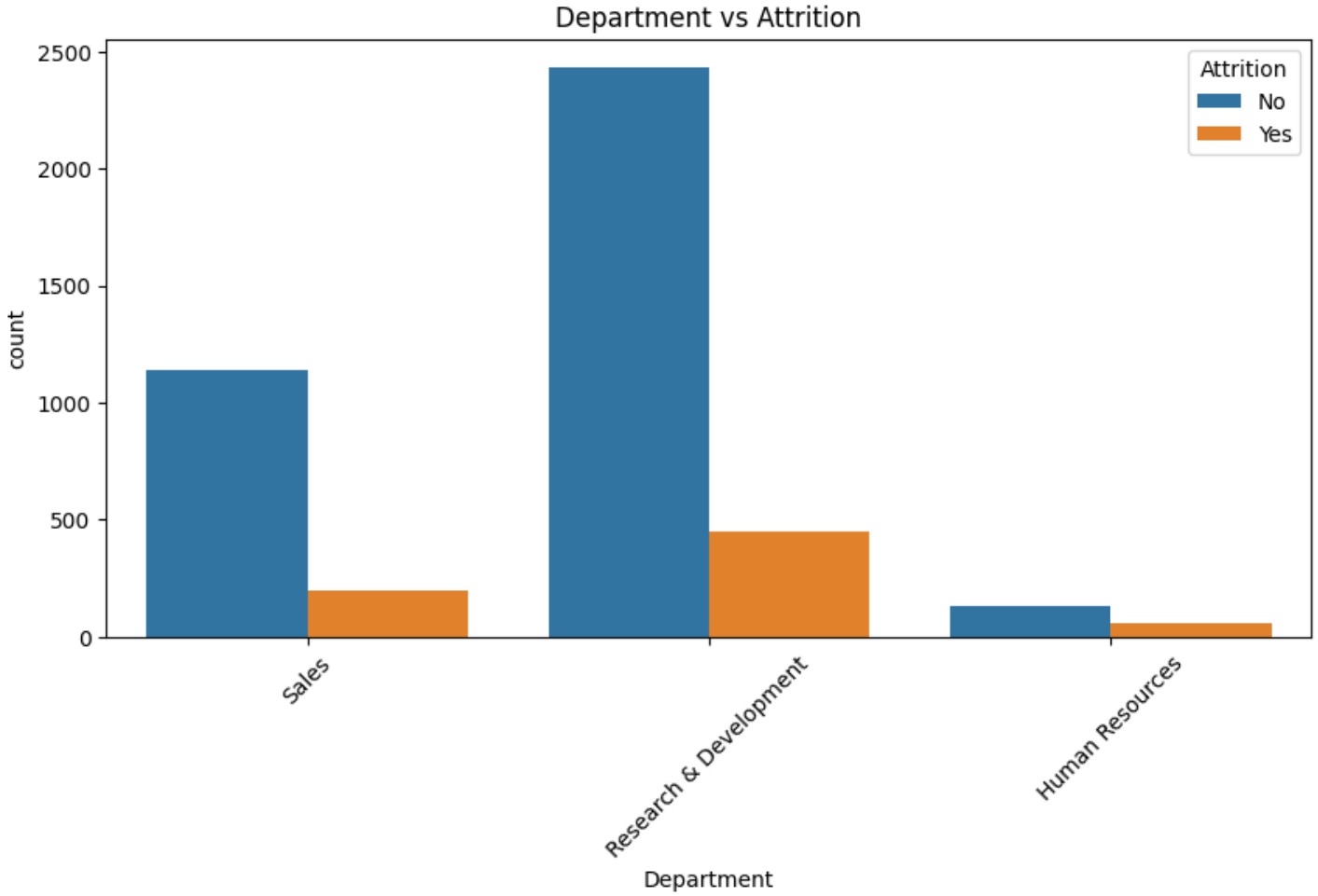

JobInvolvem
2.5
2.0
1.5
1.0
No
Yes
Attrition

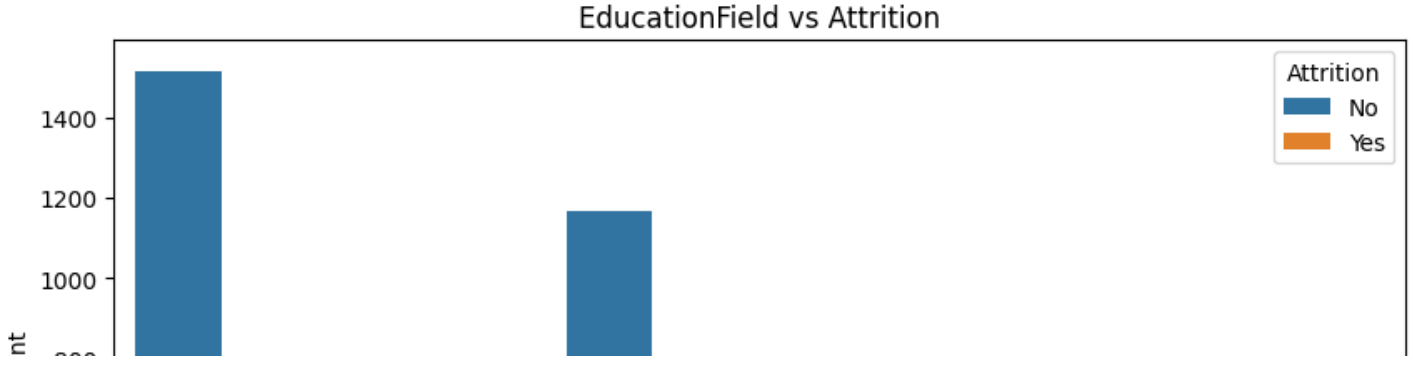

PerformanceRating vs Attrition
4.0
3.8
3.6
3.4
3.2
3.0
PerformanceRating
No
Yes
Attrition

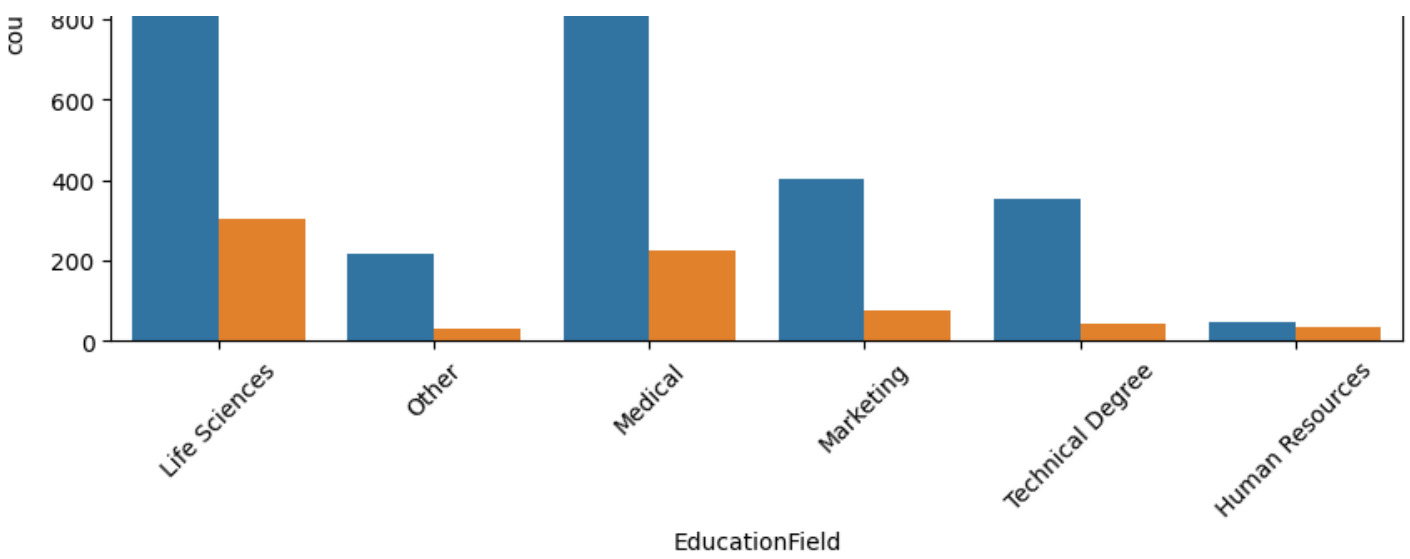

Attrition vs Attrition
3500
3000
2500
2000
1500
1000
500
0
count
No
Yes
Attrition

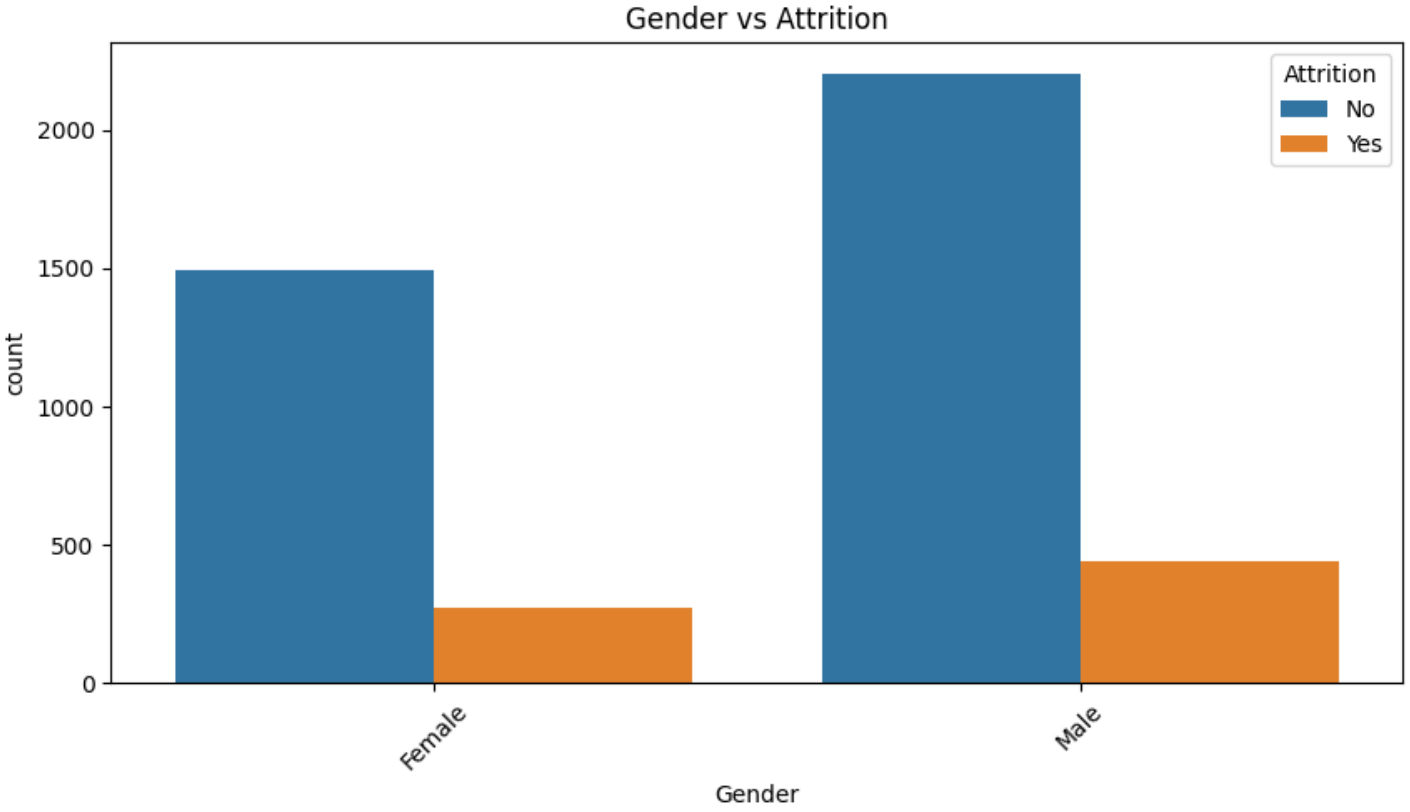

BusinessTravel vs Attrition
Attrition

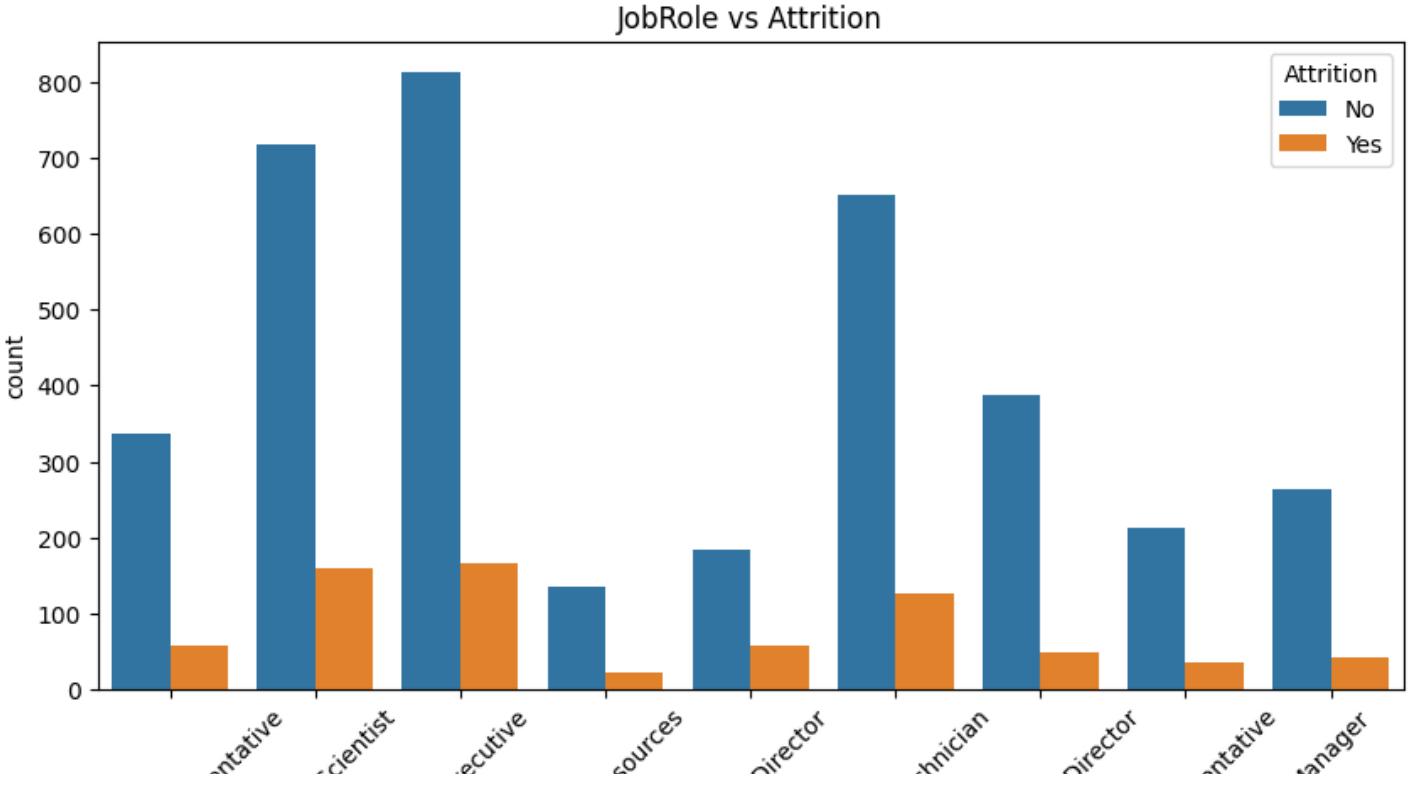

2500
2000
1500
1000
500
0
count
No
Yes
Travel_Rarely
Travel_Frequently
Non-Travel
BusinessTravel

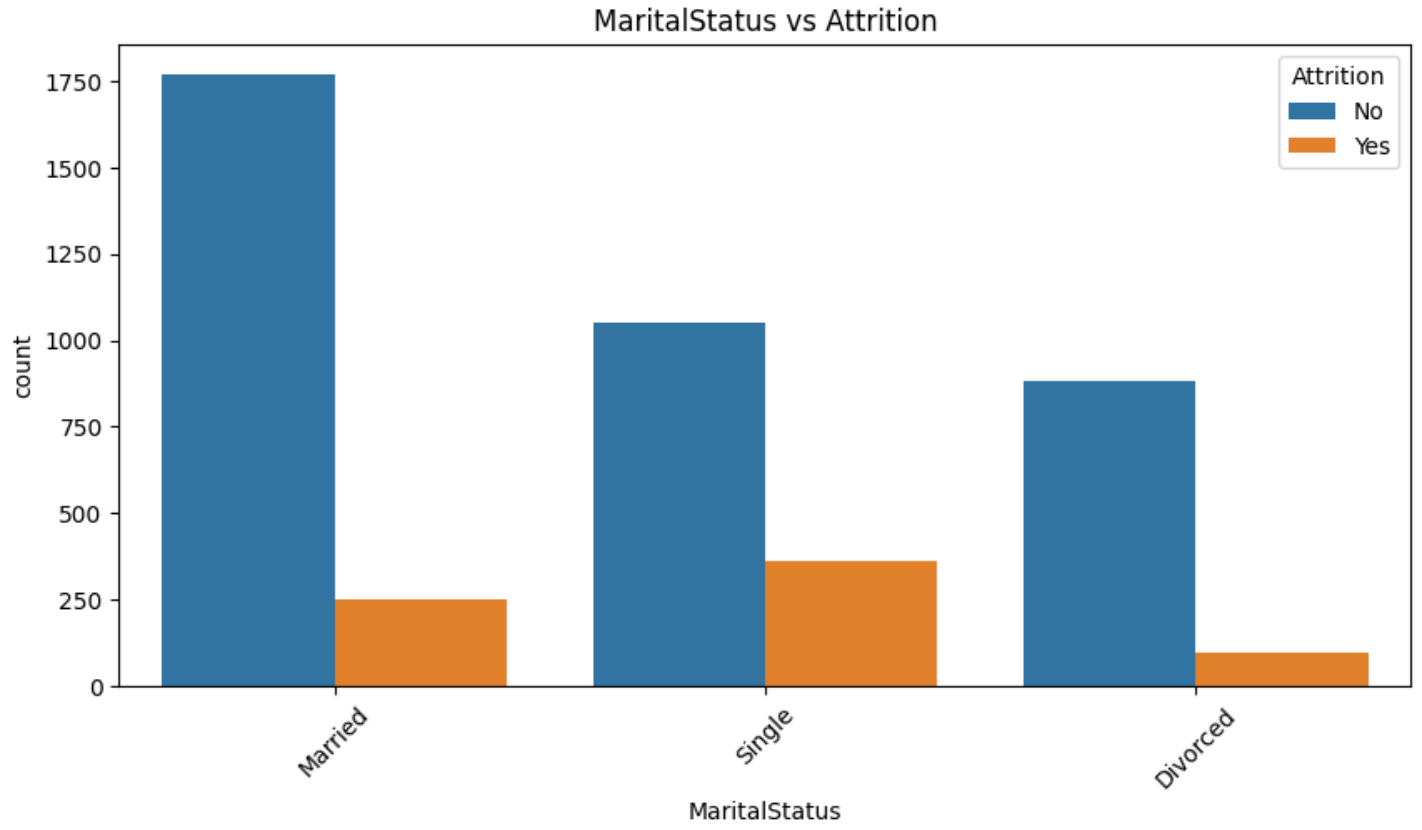

Department vs Attrition
2500
2000
1500
1000
500
0
count
Attrition
No
Yes
Sales
Research & Development
Human Resources
Department

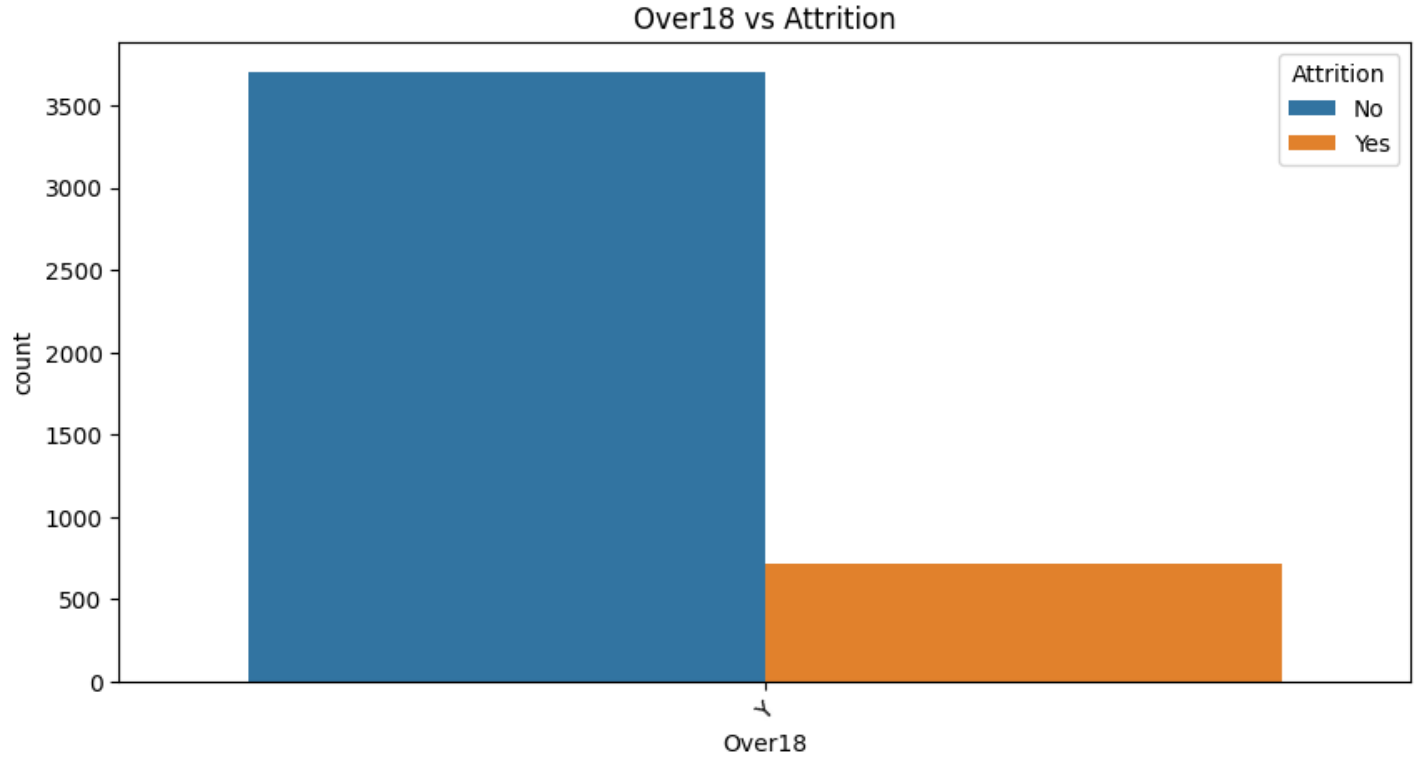

EducationField vs Attrition
1400
1200
1000
Attrition
No
Yes


800
600
400
200
0
Life Sciences
Other
Medical
Marketing
Technical Degree
Human Resources
EducationField


Gender vs Attrition
Attrition
No
Yes
2000
1500
1000
500
0
count
Female
Male
Gender


JobRole vs Attrition
Attrition
No
Yes
800
700
600
500
400
300
200
100
0
count





OBSERVATION

The high attrition rate in the company depends on low-income, bad employment experience and limited work-life balance conditions with poor perspective growth. For instance high attrition could exist in front end teams like sales due to higher pressure work or variable pay structures and Research and Development may witness exits because of no scope on innovation opportunity for the employees growth. Stress can likewise impact the HR division dealing with representative relations and associations. The issues can be dealt with the help of targeted strategies like uplifting job satisfaction, offering competitive salaries and improving career progression paths to

slash attrition rates particularly in these worst-hit departments.