```python
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import streamlit as st

file_path = "Attrition data (1).csv"
data = pd.read_csv(file_path)

# Display the first few rows of the dataset
print(data.head(10))
```

```
   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
5           6   46        No        Travel_Rarely  Research &
Development
6           7   28       Yes        Travel_Rarely  Research &
Development
7           8   29        No        Travel_Rarely  Research &
Development
8           9   31        No        Travel_Rarely  Research &
Development
9          10   25        No         Non-Travel  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   ...
5                 8          3  Life Sciences              1
Female  ...
6                11          2        Medical              1
Male   ...
```

```
7                  18          3  Life Sciences            1
Male   ...
8                   1          3  Life Sciences            1
Male   ...
9                   7          4        Medical            1
Female  ...

   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
5               28.0                      5               7
6                5.0                      2               0
7               10.0                      2               0
8               10.0                      2               9
9                6.0                      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
5                        7                      7
3.0
6                        0                      0
1.0
7                        0                      0
1.0
8                        7                      8
2.0
9                        1                      5
2.0

   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	2.0	2.0	3	3
6	3.0	1.0	3	4
7	2.0	3.0	3	4
8	4.0	3.0	3	4
9	1.0	3.0	3	3

[10 rows x 29 columns]

```python
print(data.info())
print(data.describe())
```

```
<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
         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]

missing_values = data.isnull().sum()
print(missing_values[missing_values > 0])
```

```
NumCompaniesWorked            19
TotalWorkingYears              9
EnvironmentSatisfaction       25
JobSatisfaction               20
WorkLifeBalance               38
dtype: int64
```

```python
data['Age'] = data['Age'].fillna(data['Age'].median())
data['MonthlyIncome'] =
data['MonthlyIncome'].fillna(data['MonthlyIncome'].median())
data['Department'] =
data['Department'].fillna(data['Department'].mode()[0])

data = pd.get_dummies(data, drop_first=True)

data.dtypes
```

```
EmployeeID                              int64
Age                                     int64
DistanceFromHome                        int64
Education                               int64
EmployeeCount                           int64
JobLevel                                int64
MonthlyIncome                           int64
NumCompaniesWorked                    float64
PercentSalaryHike                       int64
StandardHours                           int64
StockOptionLevel                        int64
TotalWorkingYears                     float64
TrainingTimesLastYear                   int64
YearsAtCompany                          int64
YearsSinceLastPromotion                 int64
YearsWithCurrManager                    int64
EnvironmentSatisfaction               float64
JobSatisfaction                       float64
WorkLifeBalance                       float64
JobInvolvement                          int64
PerformanceRating                       int64
Attrition_Yes                            bool
BusinessTravel_Travel_Frequently         bool
BusinessTravel_Travel_Rarely             bool
Department_Research & Development        bool
Department_Sales                         bool
EducationField_Life Sciences             bool
EducationField_Marketing                 bool
EducationField_Medical                   bool
EducationField_Other                     bool
EducationField_Technical Degree          bool
Gender_Male                              bool
JobRole_Human Resources                  bool
```

```
JobRole_Laboratory Technician          bool
JobRole_Manager                        bool
JobRole_Manufacturing Director         bool
JobRole_Research Director              bool
JobRole_Research Scientist             bool
JobRole_Sales Executive                bool
JobRole_Sales Representative           bool
MaritalStatus_Married                  bool
MaritalStatus_Single                   bool
dtype: object
```

```python
# Distribution of attrition
plt.figure(figsize=(6, 4))
sns.countplot(x='Attrition_Yes', data=data)
plt.title('Attrition Count')
plt.show()
```


Attrition Count
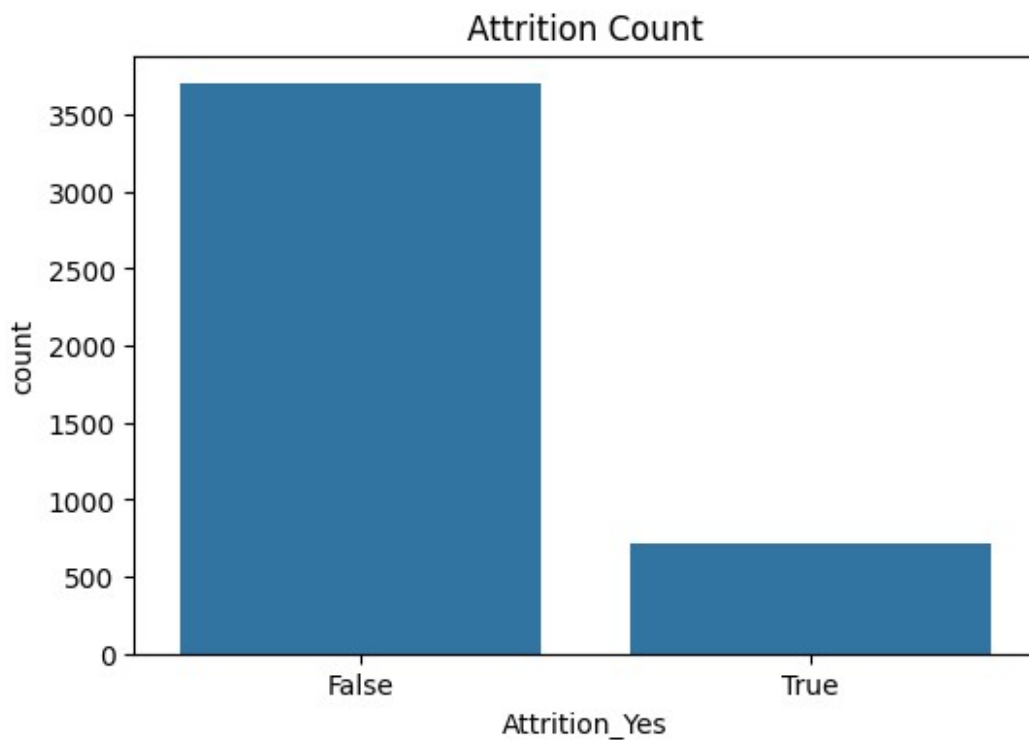
```python
# Correlation heatmap for numeric features
corr = data.corr()
plt.figure(figsize=(10, 6))
sns.heatmap(corr, annot=True, cmap='coolwarm', fmt='.2f')
plt.title('Correlation Heatmap')
plt.show()
```

Correlation Heatmap
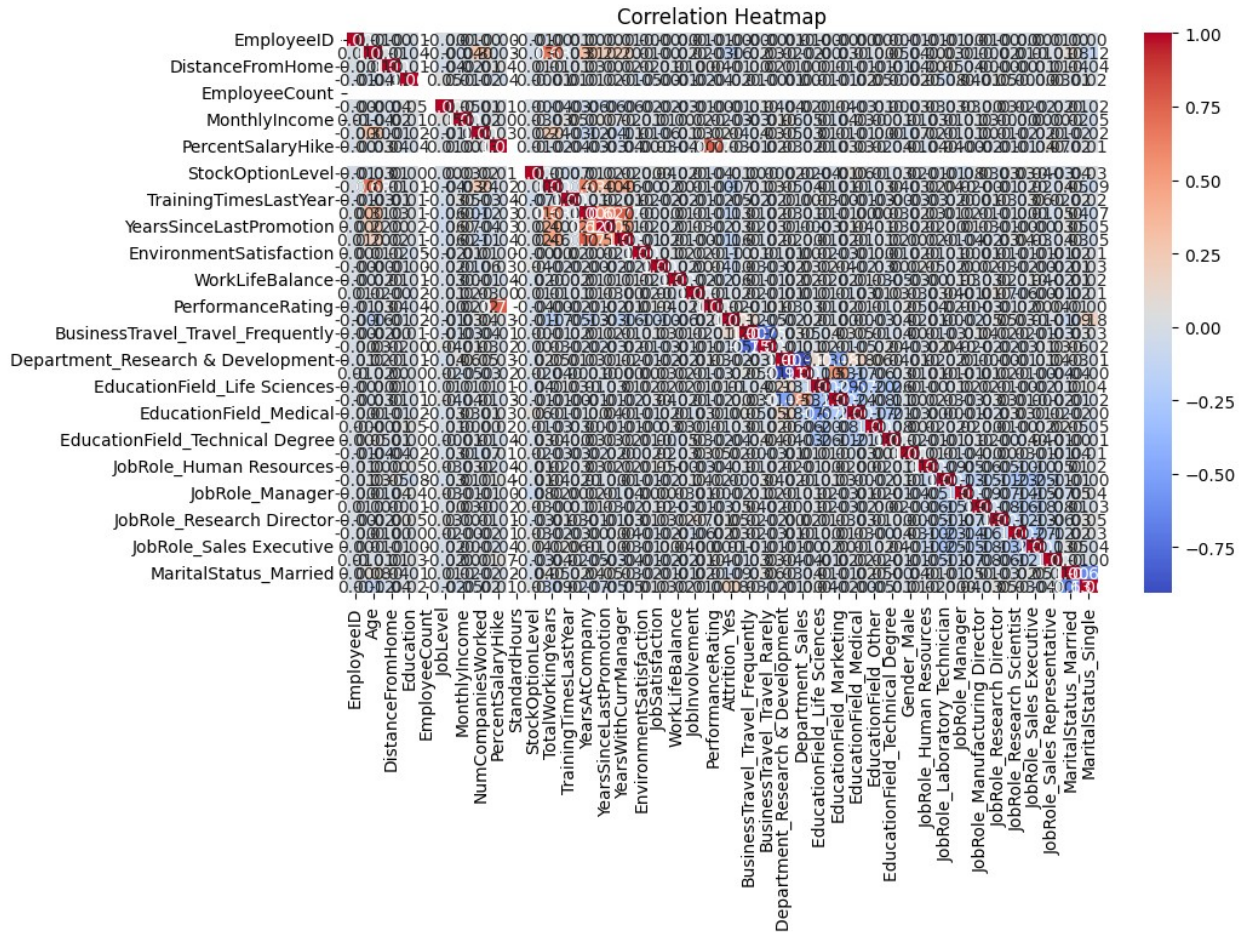

```python
# Compare Attrition by various features (e.g., Age, Tenure,
Department)
plt.figure(figsize=(8, 6))
sns.boxplot(x='Attrition_Yes', y='Age', data=data)
plt.title('Attrition by Age')
plt.show()
```

Attrition by Age
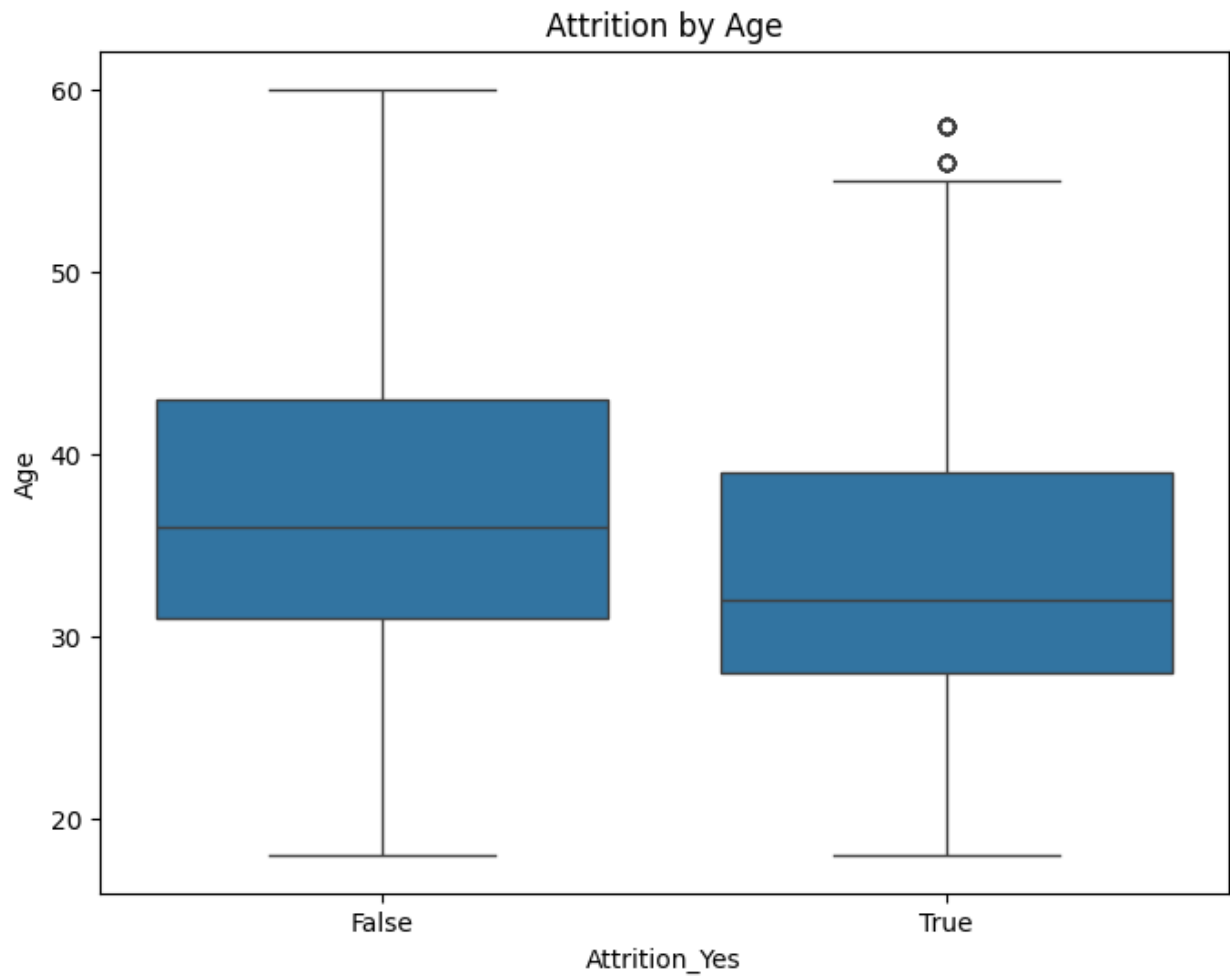

```python
# Attrition rate by Department
plt.figure(figsize=(8, 6))
sns.barplot(x='Department_Sales', y='Attrition_Yes', data=data)
plt.title('Attrition by IT Department')
plt.show()
```
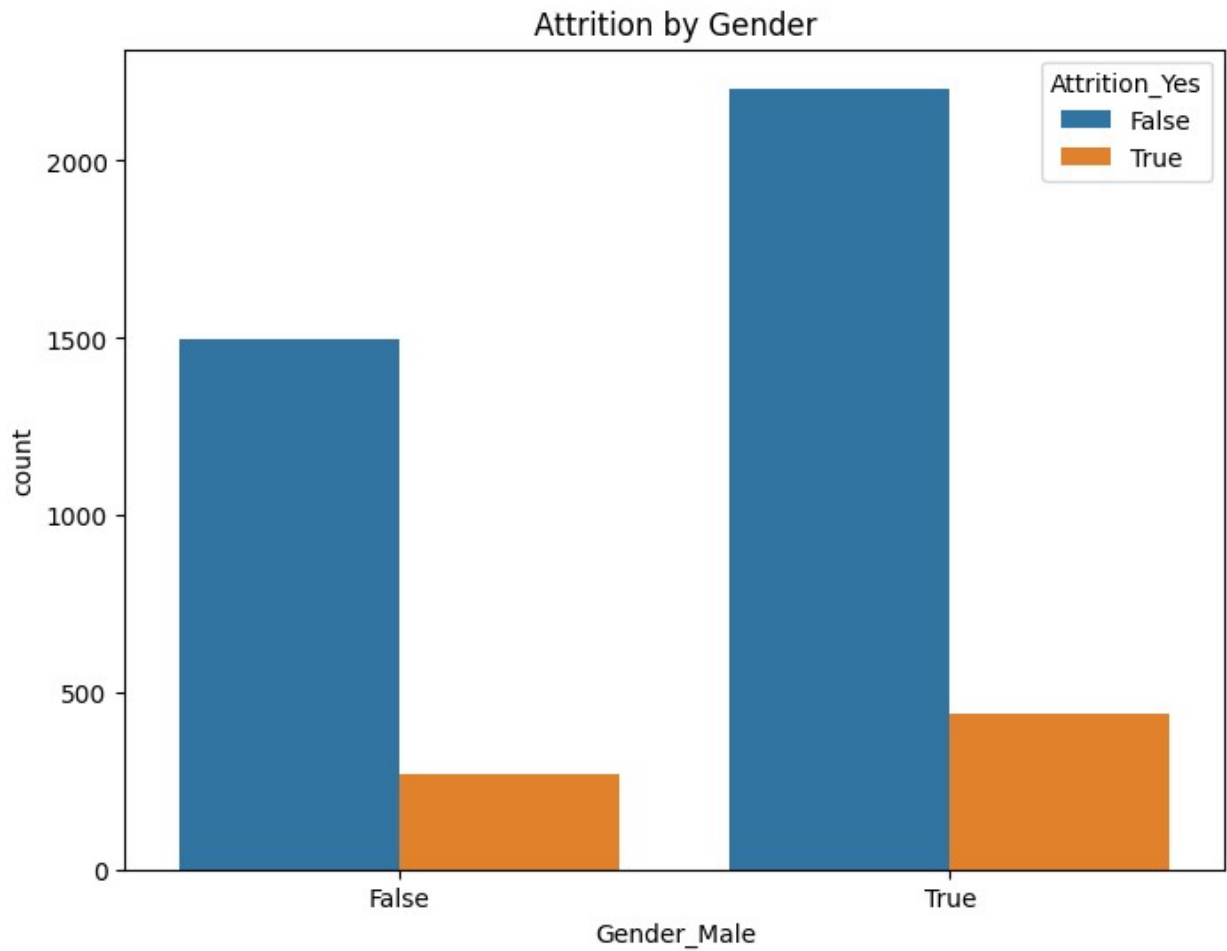
Attrition by IT Department
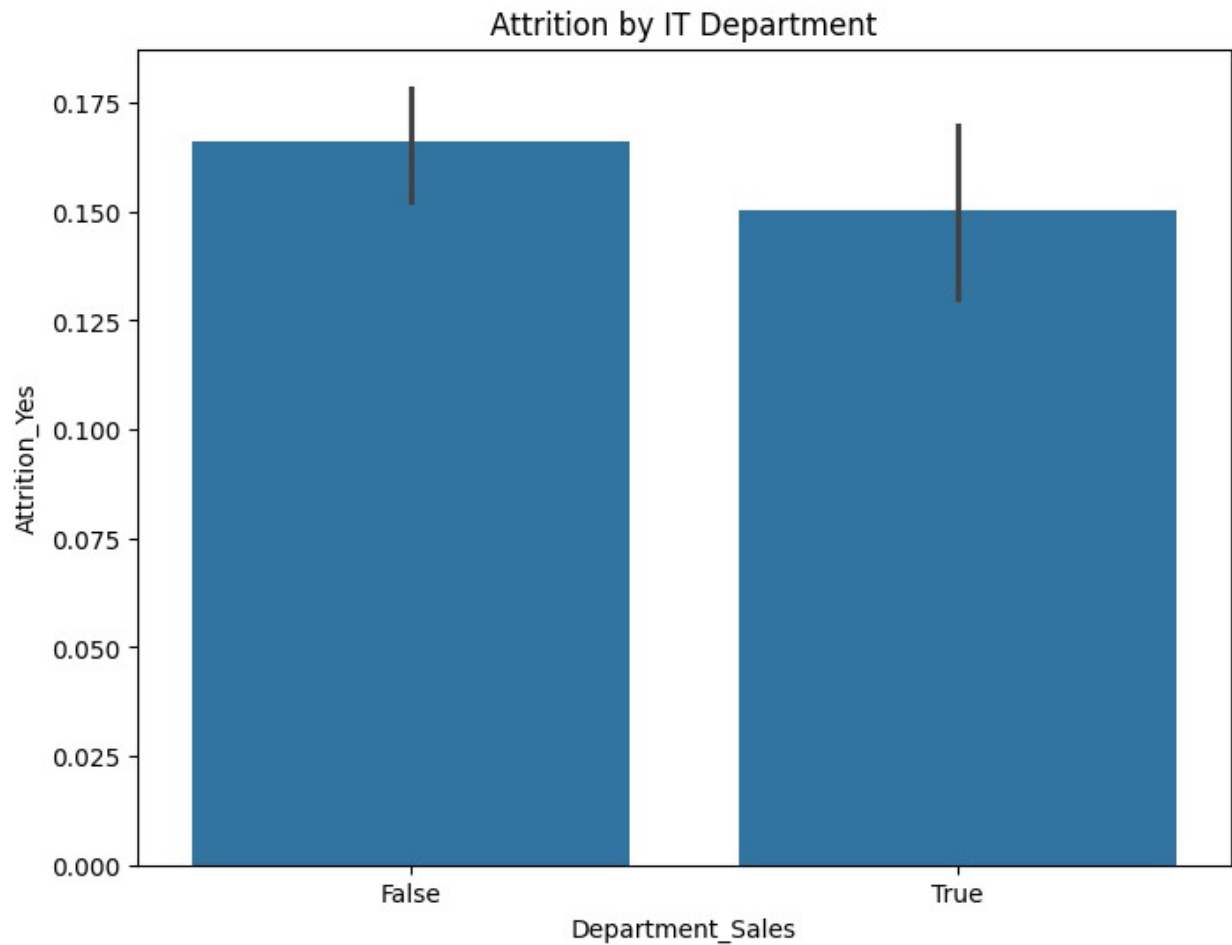

```
plt.figure(figsize=(8, 6))
sns.countplot(x='Gender_Male', hue='Attrition_Yes', data=data)
plt.title('Attrition by Gender')
plt.show()
```

```
<Figure size 1000x600 with 0 Axes>
```

```
<Figure size 800x600 with 0 Axes>
```

```
<Figure size 800x600 with 0 Axes>
```

```
<Figure size 800x600 with 0 Axes>
```

Attrition by Gender

```
plt.figure(figsize=(8, 6))
sns.boxplot(x='Gender_Male', y='MonthlyIncome', hue='Attrition_Yes',
data=data)
plt.title('Monthly Income by Gender and Attrition')
plt.show()

<Figure size 800x600 with 0 Axes>
```

Monthly Income by Gender and Attrition