In [15]:

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

# Load the dataset
file_path = '/users/Jobin/Desktop/WA_Fn-UseC_-HR-Employee-Attrition.csv'
employee_data = pd.read_csv(file_path)
employee_data.head(100)
```

Out[15]:

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	Emp
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	
...	...	...	...	...	...	...	...	...	...	
95	54	No	Travel_Rarely	1217	Research & Development	2	4	Technical Degree	1	
96	24	No	Travel_Rarely	1353	Sales	3	2	Other	1	
97	28	No	Non-Travel	120	Sales	4	3	Medical	1	
98	58	No	Travel_Rarely	682	Sales	10	4	Medical	1	
99	44	No	Non-Travel	489	Research & Development	23	3	Medical	1	

100 rows × 35 columns

In [18]:

```python
# Data cleaning
# Check for missing values
missing_values = employee_data.isnull().sum()
print("Missing Values:\n", missing_values)

# Check for duplicate records
duplicates = employee_data.duplicated().sum()
print("Duplicate Records:", duplicates)

# Handle missing values (replace NaN values or drop rows/columns)
# Example: Drop rows with missing values
employee_data_cleaned = employee_data.dropna()

# Handle duplicate records (drop duplicates if necessary)
# Example: Drop duplicate records
employee_data_cleaned = employee_data_cleaned.drop_duplicates()
```

```
Missing Values:
 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
Duplicate Records: 0
```

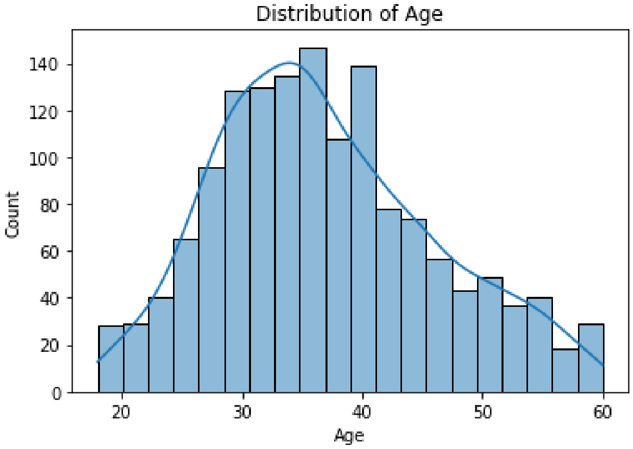
In [19]:

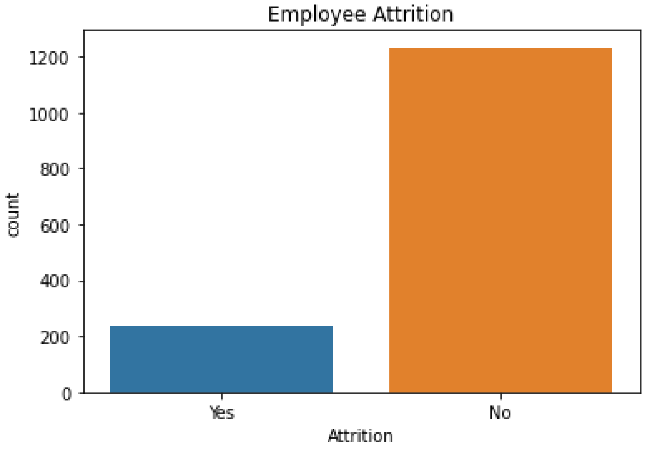
```python
# Handle outliers (use appropriate method)
# Example: Identify and remove outliers in the 'Age' column using IQR
Q1 = employee_data_cleaned['Age'].quantile(0.25)
Q3 = employee_data_cleaned['Age'].quantile(0.75)
IQR = Q3 - Q1
employee_data_cleaned = employee_data_cleaned[
    (employee_data_cleaned['Age'] >= Q1 - 1.5 * IQR) & (employee_data_cleaned['Age'] <= Q3 + 1.5 * IQR)
]

# Ensure data consistency
# Example: Check unique values in the 'Department' column
unique_departments = employee_data_cleaned['Department'].unique()
print("Unique Departments:", unique_departments)
```

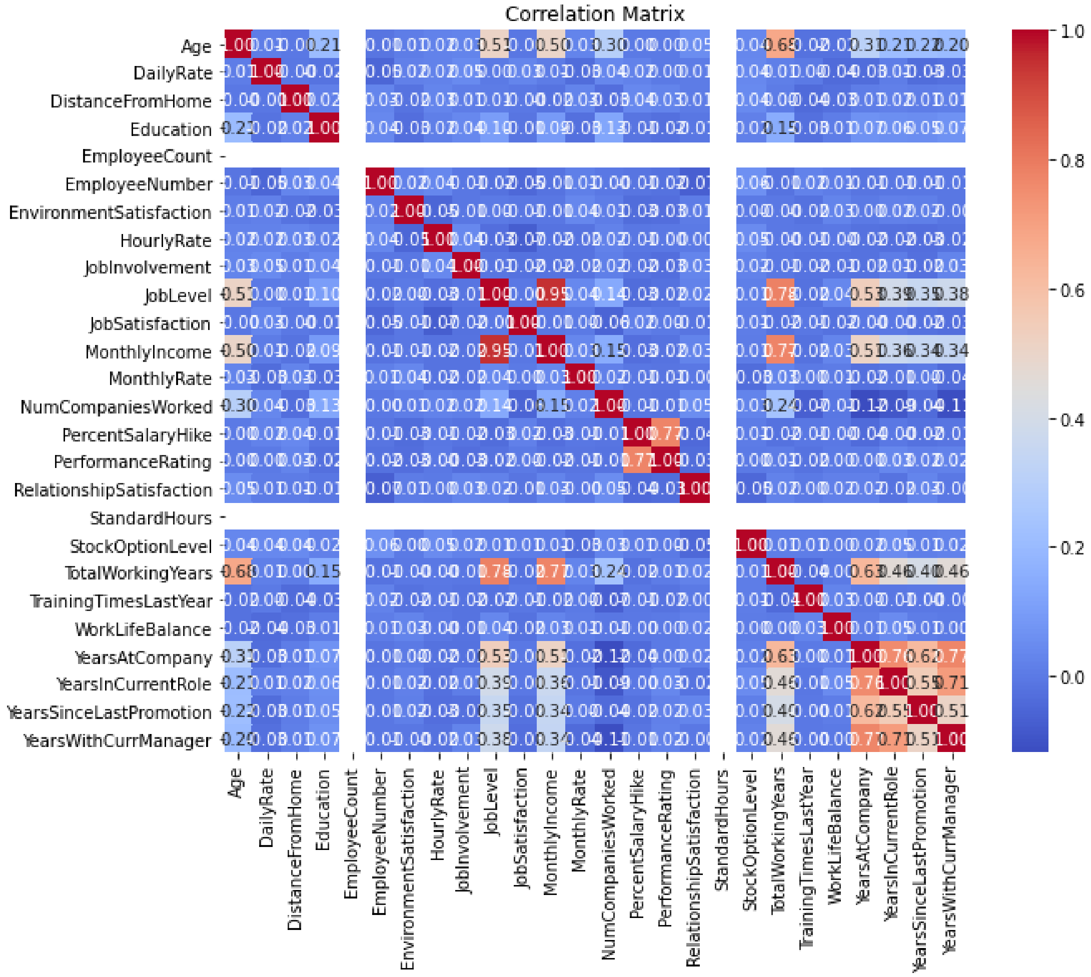
```
Unique Departments: ['Sales' 'Research & Development' 'Human Resources']
```

In [28]:

```python
# Explore the distribution of features
sns.histplot(employee_data_cleaned['Age'], bins=20, kde=True)
plt.title('Distribution of Age')
plt.show()

# Visualize employee attrition
sns.countplot(x='Attrition', data=employee_data_cleaned)
plt.title('Employee Attrition')
plt.show()

# Correlation matrix
correlation_matrix = employee_data_cleaned.corr()
plt.figure(figsize = (10,8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt='.2f')
plt.title('Correlation Matrix')
plt.show()

# Descriptive statistics
descriptive_stats = employee_data_cleaned.describe()
print("Descriptive Statistics:\n", descriptive_stats)
```


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

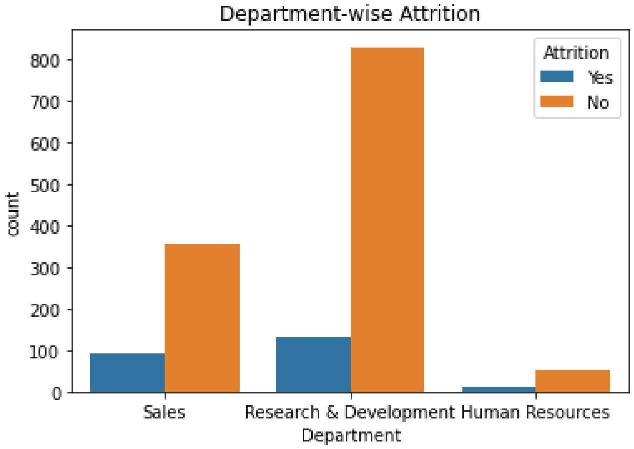

Employee Attrition
count
0
200
400
600
800
1000
1200
Yes
No
Attrition

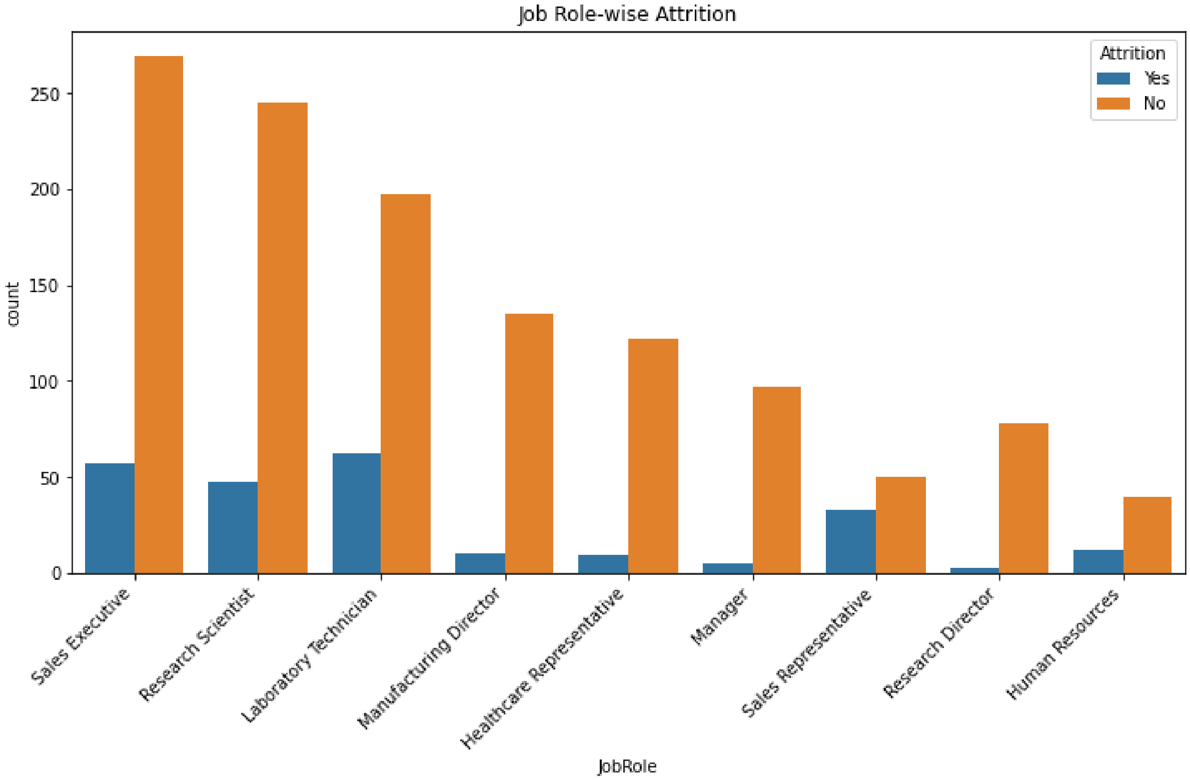

Correlation Matrix
Age
DailyRate
DistanceFromHome
Education
EmployeeCount
EmployeeNumber
EnvironmentSatisfaction
HourlyRate
JobInvolvement
JobLevel
JobSatisfaction
MonthlyIncome
MonthlyRate
NumCompaniesWorked
PercentSalaryHike
PerformanceRating
RelationshipSatisfaction
StandardHours
StockOptionLevel
TotalWorkingYears
TrainingTimesLastYear
WorkLifeBalance
YearsAtCompany
YearsInCurrentRole
YearsSinceLastPromotion
YearsWithCurrManager
1.0
0.8
0.6
0.4
0.2
0.0

```
Descriptive Statistics:
               Age    DailyRate  DistanceFromHome    Education  EmployeeCount  \
count  1470.000000  1470.000000       1470.000000  1470.000000         1470.0   
mean     36.923810   802.485714          9.192517     2.912925            1.0   
std       9.135373   403.509100          8.106864     1.024165            0.0   
min      18.000000   102.000000          1.000000     1.000000            1.0   
25%      30.000000   465.000000          2.000000     2.000000            1.0   
50%      36.000000   802.000000          7.000000     3.000000            1.0   
75%      43.000000  1157.000000         14.000000     4.000000            1.0   
max      60.000000  1499.000000         29.000000     5.000000            1.0   

       EmployeeNumber  EnvironmentSatisfaction   HourlyRate  JobInvolvement  \
count     1470.000000              1470.000000  1470.000000     1470.000000   
mean      1024.865306                 2.721769    65.891156        2.729932   
std        602.024335                 1.093082    20.329428        0.711561   
min          1.000000                 1.000000    30.000000        1.000000   
25%        491.250000                 2.000000    48.000000        2.000000   
50%       1020.500000                 3.000000    66.000000        3.000000   
75%       1555.750000                 4.000000    83.750000        3.000000   
max       2068.000000                 4.000000   100.000000        4.000000   

          JobLevel  ...  RelationshipSatisfaction  StandardHours  \
count  1470.000000  ...               1470.000000         1470.0   
mean      2.063946  ...                  2.712245           80.0   
std       1.106940  ...                  1.081209            0.0   
min       1.000000  ...                  1.000000           80.0   
25%       1.000000  ...                  2.000000           80.0   
50%       2.000000  ...                  3.000000           80.0   
75%       3.000000  ...                  4.000000           80.0   
max       5.000000  ...                  4.000000           80.0   

       StockOptionLevel  TotalWorkingYears  TrainingTimesLastYear  \
count       1470.000000        1470.000000            1470.000000   
mean           0.793878          11.279592               2.799320   
std            0.852077           7.780782               1.289271   
min            0.000000           0.000000               0.000000   
25%            0.000000           6.000000               2.000000   
50%            1.000000          10.000000               3.000000   
75%            1.000000          15.000000               3.000000   
max            3.000000          40.000000               6.000000   

       WorkLifeBalance  YearsAtCompany  YearsInCurrentRole  \
count      1470.000000     1470.000000         1470.000000   
mean          2.761224        7.008163            4.229252   
std           0.706476        6.126525            3.623137   
min           1.000000        0.000000            0.000000   
25%           2.000000        3.000000            2.000000   
50%           3.000000        5.000000            3.000000   
75%           3.000000        9.000000            7.000000   
max           4.000000       40.000000           18.000000   

       YearsSinceLastPromotion  YearsWithCurrManager  
count              1470.000000           1470.000000  
mean                  2.187755              4.123129  
std                   3.222430              3.568136  
min                   0.000000              0.000000  
25%                   0.000000              2.000000  
50%                   1.000000              3.000000  
75%                   3.000000              7.000000  
max                  15.000000             17.000000  

[8 rows x 26 columns]
```

In [25]:

```
# Department-wise Attrition
sns.countplot(x='Department', hue='Attrition', data=employee_data_cleaned)
plt.title('Department-wise Attrition')
plt.show()

# Job Role-wise Attrition
plt.figure(figsize=(12, 6))
sns.countplot(x='JobRole', hue='Attrition', data=employee_data_cleaned)
plt.title('Job Role-wise Attrition')
plt.xticks(rotation=45, ha='right')
plt.show()
```


Department-wise Attrition
Attrition
Yes
No
count
0
100
200
300
400
500
600
700
800
Sales
Research & Development
Human Resources
Department


Job Role-wise Attrition
Attrition
Yes
No
count
0
50
100
150
200
250
Sales Executive
Research Scientist
Laboratory Technician
Manufacturing Director
Healthcare Representative
Manager
Sales Representative
Research Director
Human Resources
JobRole