MeriSKILL_Task_03("HR Analytics") 🌐 🔍 📝

INTRODUCTION :

Given a HR Analytics data set we need to perform Exploratory data analysis in which we will perform Data Cleaning and Data Visualisation.

Data Description 🔍 📊 💡

This dataset is originally from the National Institute of Diabetes and Digestive and KidneyDiseases. The objective of the dataset is to diagnostically predict whether a patient has diabetesbased on certain diagnostic measurements included in the dataset. Several constraints were placedon the selection of these instances from a larger database. In particular, all patients here are femalesat least 21 years old of Pima Indian heritage.2From the data set in the (.csv) File We can find several variables, some of them are independent(several medical predictor variables) and only one target dependent variable (Outcome).

Import Libraries 📚

In [2]:
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

Loading data Set and Preprocessing Of Data(Data Cleaning)

In [6]:
```
HR_data=pd.read_csv("C:\\Users\\Ujjaval Raj\\Downloads\\HR-Employee-Attrition.csv")
HR_data_head=HR_data.head()
HR_data_tail=HR_data.tail()
print(HR_data_head)
print(HR_data_tail)
```

```
   Age Attrition     BusinessTravel  DailyRate              Department  \
0   41       Yes      Travel_Rarely       1102                   Sales
1   49        No  Travel_Frequently        279  Research & Development
2   37       Yes      Travel_Rarely       1373  Research & Development
3   33        No  Travel_Frequently       1392  Research & Development
4   27        No      Travel_Rarely        591  Research & Development

   DistanceFromHome  Education EducationField  EmployeeCount  EmployeeNumber  \
0                 1          2  Life Sciences              1               1
1                 8          1  Life Sciences              1               2
2                 2          2          Other              1               4
3                 3          4  Life Sciences              1               5
4                 2          1        Medical              1               7

   ...  RelationshipSatisfaction StandardHours  StockOptionLevel  \
0  ...                         1            80                 0
1  ...                         4            80                 1
2  ...                         2            80                 0
3  ...                         3            80                 0
4  ...                         4            80                 1

   TotalWorkingYears  TrainingTimesLastYear WorkLifeBalance  YearsAtCompany  \
0                  8                      0               1               6
1                 10                      3               3              10
2                  7                      3               3               0
3                  8                      3               3               8
4                  6                      3               3               2

  YearsInCurrentRole  YearsSinceLastPromotion  YearsWithCurrManager
0                  4                        0                     5
1                  7                        1                     7
2                  0                        0                     0
3                  7                        3                     0
4                  2                        2                     2

[5 rows x 35 columns]
      Age Attrition     BusinessTravel  DailyRate              Department  \
1465   36        No  Travel_Frequently        884  Research & Development
1466   39        No      Travel_Rarely        613  Research & Development
1467   27        No      Travel_Rarely        155  Research & Development
1468   49        No  Travel_Frequently       1023                   Sales
1469   34        No      Travel_Rarely        628  Research & Development

      DistanceFromHome  Education EducationField  EmployeeCount  \
1465                23          2        Medical              1
1466                 6          1        Medical              1
1467                 4          3  Life Sciences              1
1468                 2          3        Medical              1
1469                 8          3        Medical              1

      EmployeeNumber  ...  RelationshipSatisfaction StandardHours  \
1465            2061  ...                         3            80
1466            2062  ...                         1            80
1467            2064  ...                         2            80
1468            2065  ...                         4            80
1469            2068  ...                         1            80

      StockOptionLevel  TotalWorkingYears  TrainingTimesLastYear  \
1465                 1                 17                      3
```

```
1466                 1                  9                     5
1467                 1                  6                     0
1468                 0                 17                     3
1469                 0                  6                     3

     WorkLifeBalance  YearsAtCompany YearsInCurrentRole  \
1465               3               5                  2
1466               3               7                  7
1467               3               6                  2
1468               2               9                  6
1469               4               4                  3

      YearsSinceLastPromotion  YearsWithCurrManager
1465                        0                     3
1466                        1                     7
1467                        0                     3
1468                        0                     8
1469                        1                     2

[5 rows x 35 columns]
```

Checking for null values and duplicates

In [12]:
```
HR_data_null=HR_data.isnull().sum()
HR_data_duplicate=HR_data.duplicated().sum()
print(HR_data_null)
print(HR_data_duplicate)
```

```
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
0
```

Obtaining Datatypes and Describing data set with basic statistics

In [15]:
```
HR_data.info()
HR_data.describe()
```

```
<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
```

Out[15]:

	Age	DailyRate	DistanceFromHome	Education	EmployeeCount	EmployeeNumb
count	1470.000000	1470.000000	1470.000000	1470.000000	1470.0	1470.000(
mean	36.923810	802.485714	9.192517	2.912925	1.0	1024.865(
std	9.135373	403.509100	8.106864	1.024165	0.0	602.024(
min	18.000000	102.000000	1.000000	1.000000	1.0	1.000(
25%	30.000000	465.000000	2.000000	2.000000	1.0	491.250(
50%	36.000000	802.000000	7.000000	3.000000	1.0	1020.500(
75%	43.000000	1157.000000	14.000000	4.000000	1.0	1555.750(
max	60.000000	1499.000000	29.000000	5.000000	1.0	2068.000(

8 rows × 26 columns

Deleting column standardHours and Over18 as data is repeated for all rows(relundancy) .

In [18]:
```
HR_data = HR_data.drop(columns=['Over18','StandardHours'])
HR_data.head(1)
```

Out[18]:

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	Educatio
0	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sc

1 rows × 34 columns

Renaming the columns.

In [19]:
```
HR_data = HR_data.rename(columns={'NumCompaniesWorked': 'NumberOfWorkedCompanies'})
HR_data.head(1)
```

Out[19]:

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	Educatio
0	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sc

1 rows × 34 columns

Dropping Not available values (NAN).

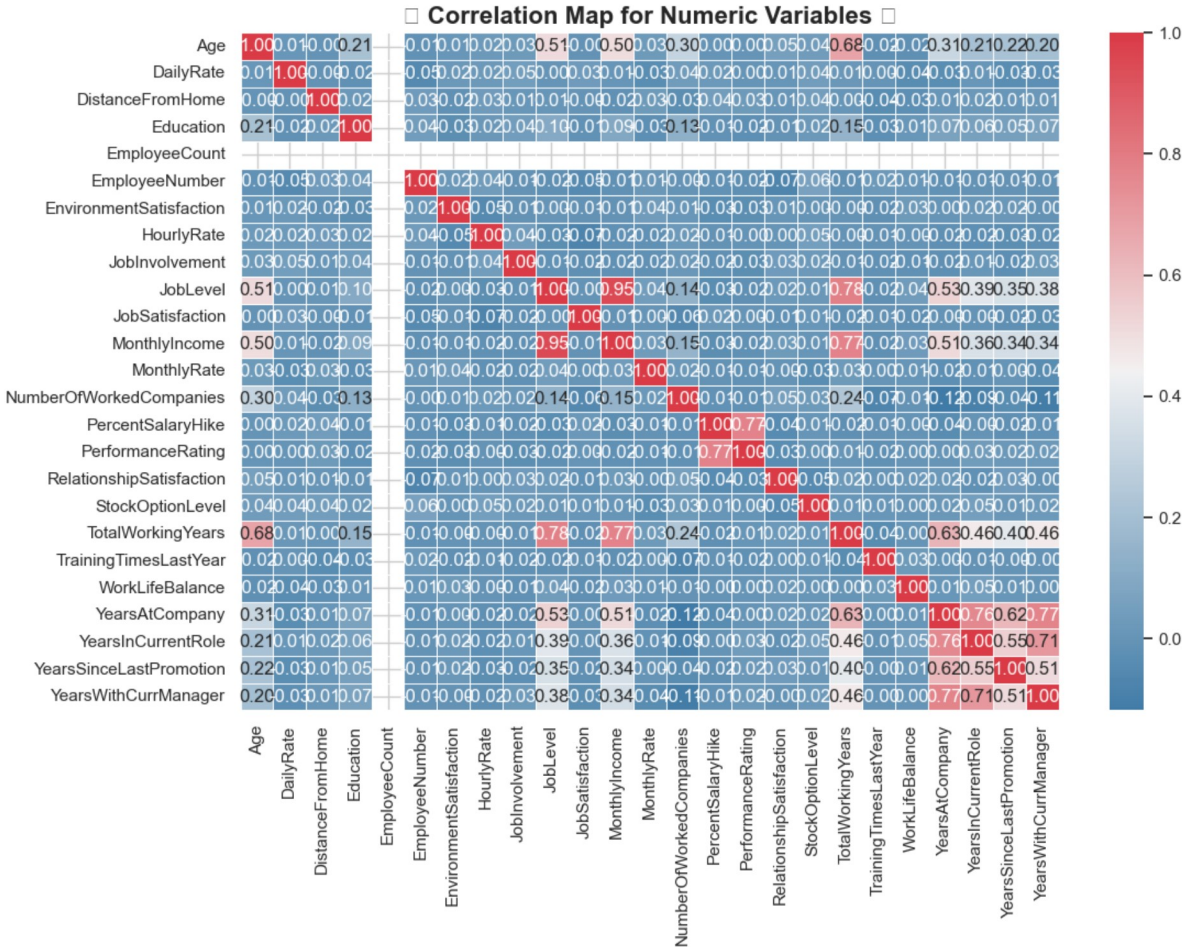
In [20]:
```
HR_data = HR_data.dropna()
```

Visualisation of the HR data

Plot a correlation map for all numeric variables

In [24]:
```
numeric_columns = HR_data.select_dtypes(include=['float64', 'int64']).columns

# Calculate the correlation matrix for numeric columns only
correlation_matrix = HR_data[numeric_columns].corr()

sns.set_theme(style="whitegrid")

plt.figure(figsize=(12, 8))
cmap = sns.diverging_palette(240, 10, as_cmap=True)
sns.heatmap(correlation_matrix, annot=True, cmap=cmap, fmt=".2f", linewidths=.5)

plt.title(' ⚡ Correlation Map for Numeric Variables ⚡ ', fontsize=16, fontweight='

plt.show()
```

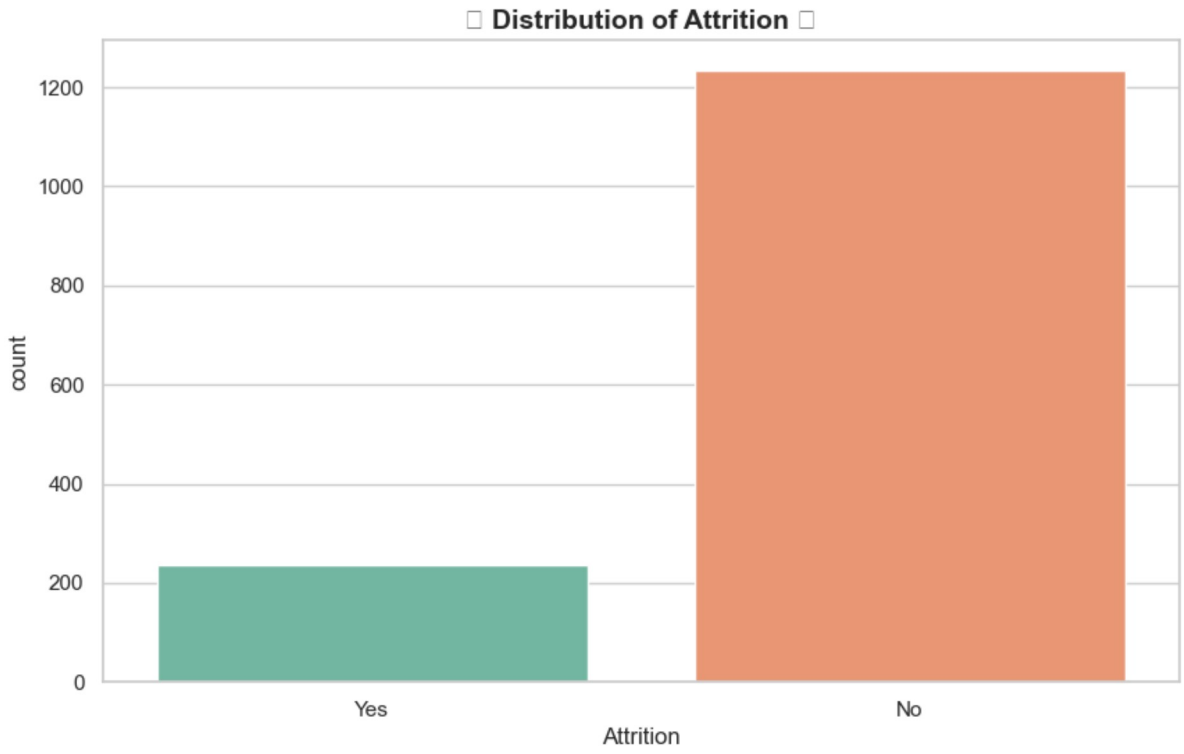
```
C:\Users\Ujjaval Raj\AppData\Roaming\Python\Python311\site-packages\IPython\core\py
labtools.py:152: UserWarning: Glyph 9889 (\N{HIGH VOLTAGE SIGN}) missing from curre
nt font.
  fig.canvas.print_figure(bytes_io, **kw)
```



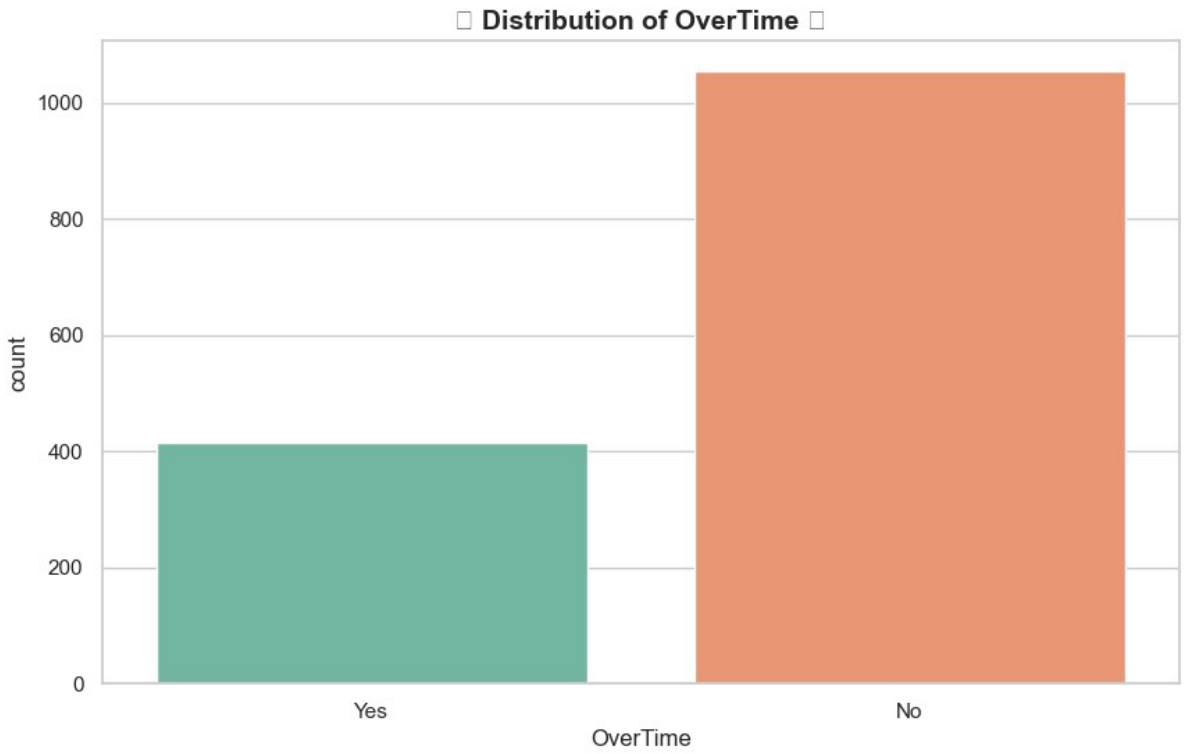
Plot a correlation map for all numeric variables

In [42]:
```
# Categorical Variables Visualizations
categorical_variables = ['Attrition','OverTime','BusinessTravel','Department','Educ

sns.set_theme(style="whitegrid")

# Iterate through categorical variables for visualizations
for variable in categorical_variables:

    plt.figure(figsize=(10, 6))
    sns.countplot(x=variable, data=HR_data, palette='Set2')
    plt.title(f' Distribution of {variable} ', fontsize=14, fontweight='bold')


    plt.show()
```

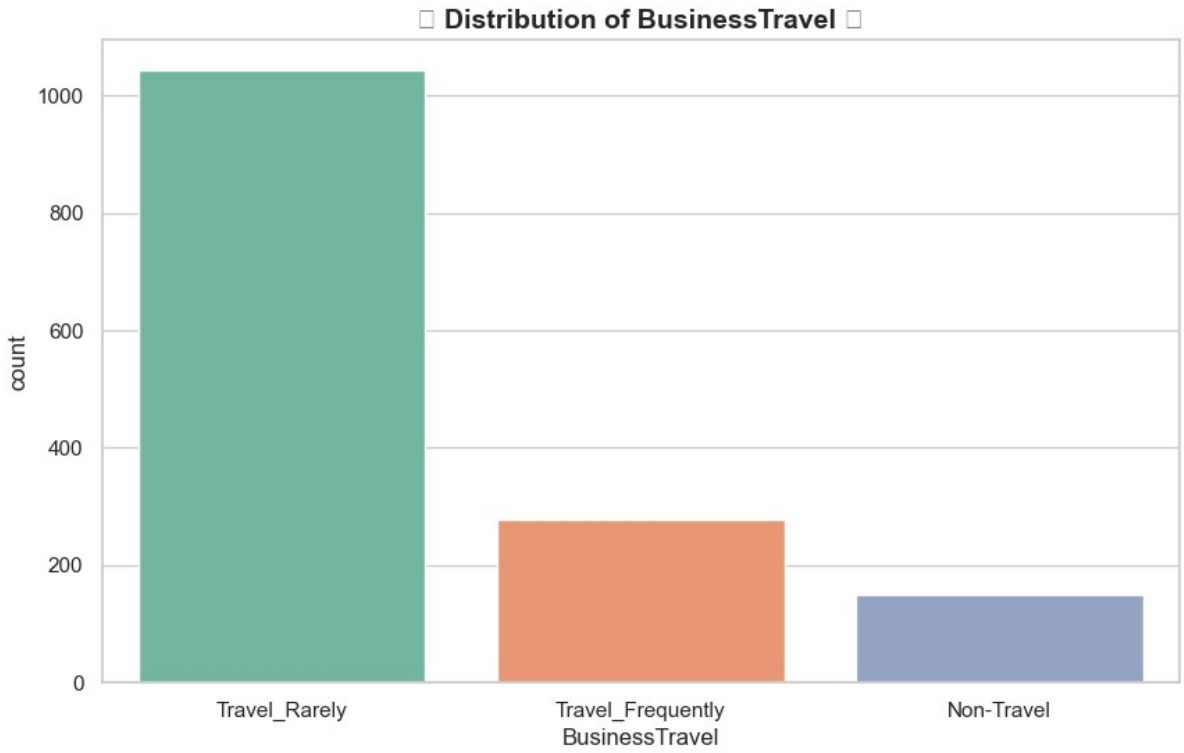
```
C:\Users\Ujjaval Raj\AppData\Roaming\Python\Python311\site-packages\IPython\core\py
labtools.py:152: UserWarning: Glyph 128202 (\N{BAR CHART}) missing from current fon
t.
  fig.canvas.print_figure(bytes_io, **kw)
```



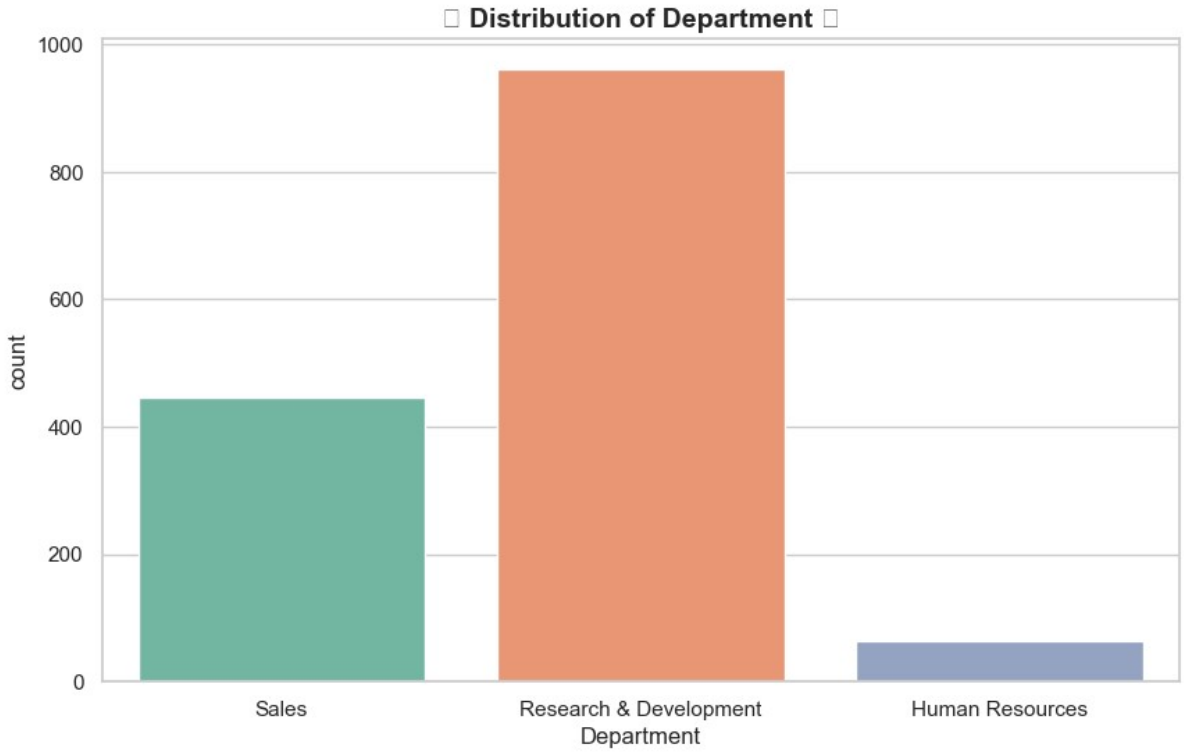

```
C:\Users\Ujjaval Raj\AppData\Roaming\Python\Python311\site-packages\IPython\core\py
labtools.py:152: UserWarning: Glyph 128202 (\N{BAR CHART}) missing from current fon
t.
  fig.canvas.print_figure(bytes_io, **kw)
```



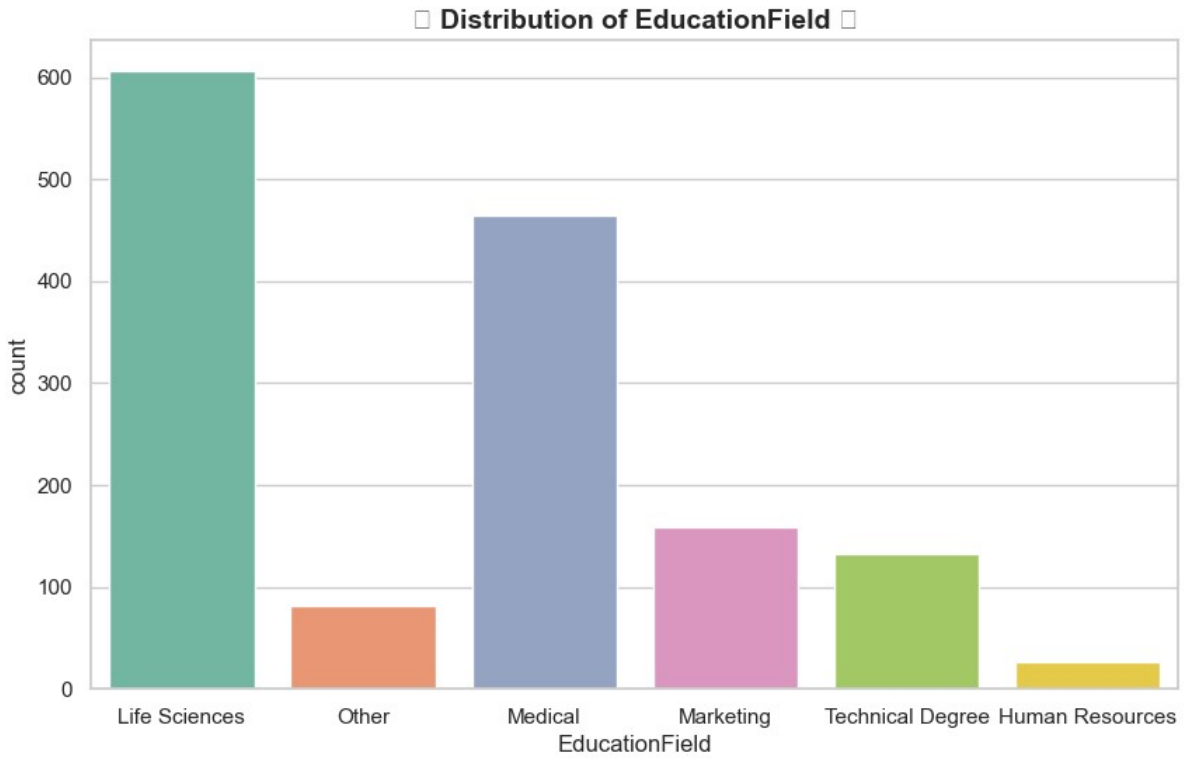

```
C:\Users\Ujjaval Raj\AppData\Roaming\Python\Python311\site-packages\IPython\core\py
labtools.py:152: UserWarning: Glyph 128202 (\N{BAR CHART}) missing from current fon
t.
  fig.canvas.print_figure(bytes_io, **kw)
```



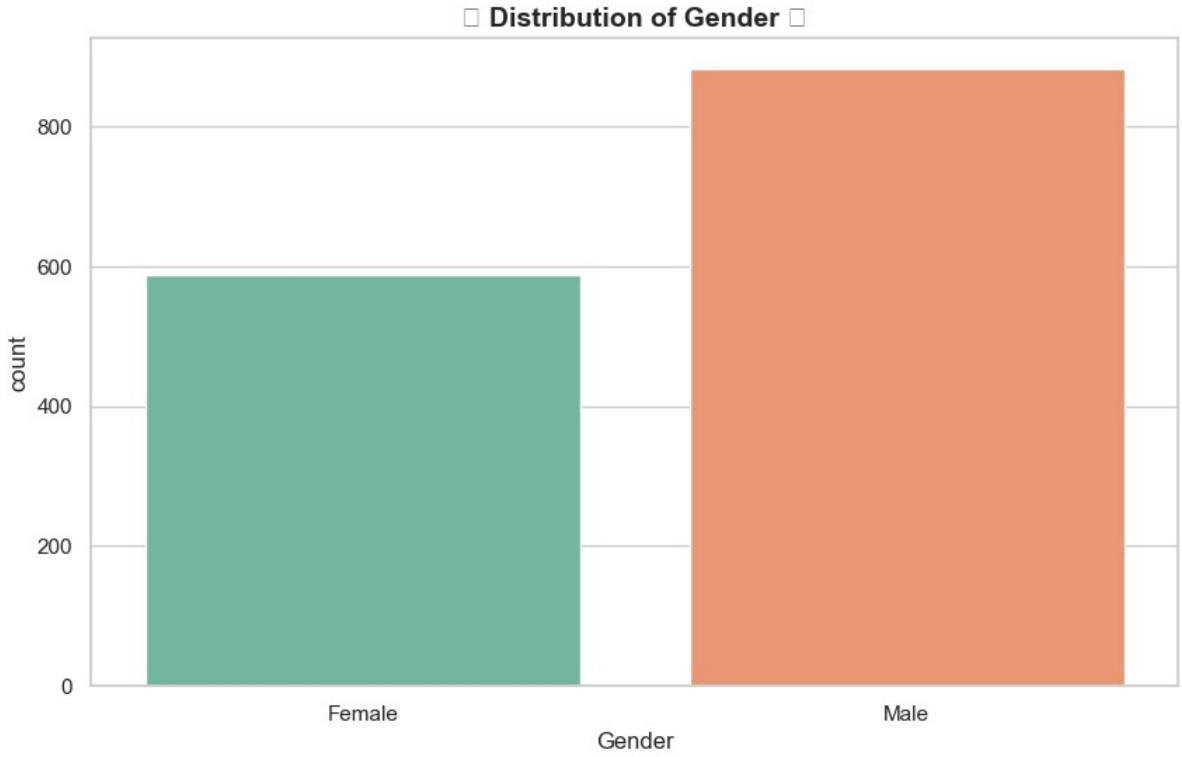

```
C:\Users\Ujjaval Raj\AppData\Roaming\Python\Python311\site-packages\IPython\core\py
labtools.py:152: UserWarning: Glyph 128202 (\N{BAR CHART}) missing from current fon
t.
  fig.canvas.print_figure(bytes_io, **kw)
```



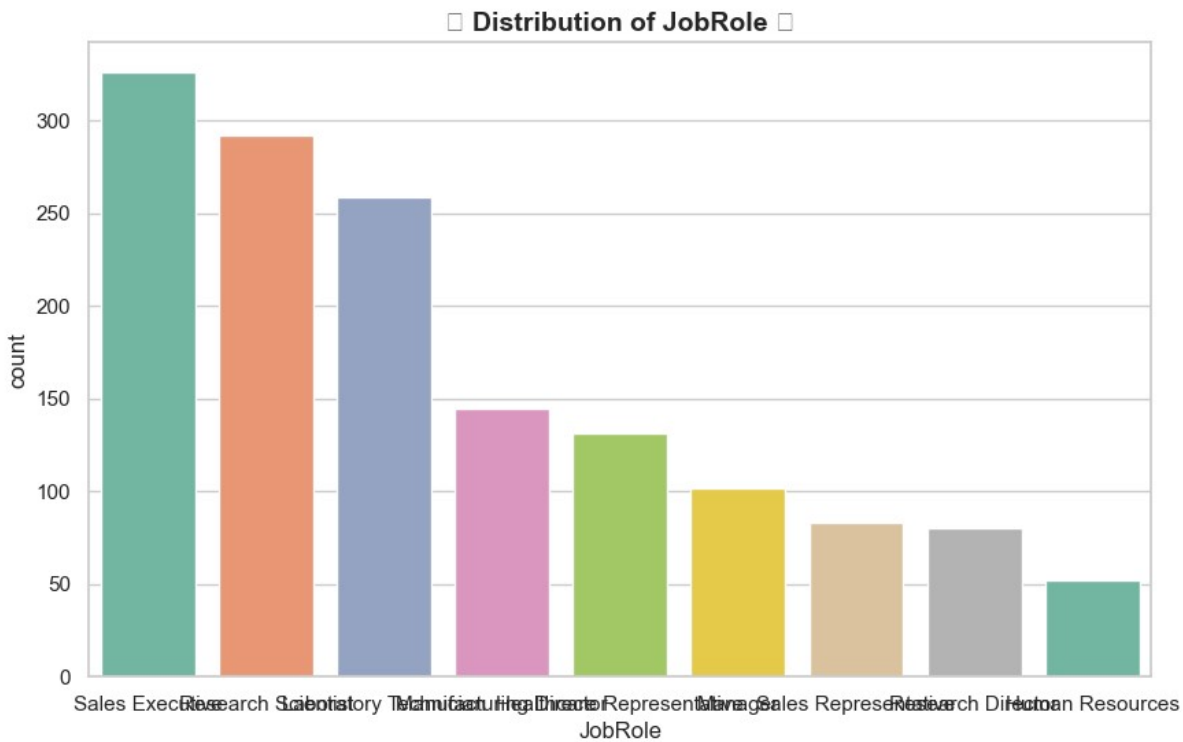

```
C:\Users\Ujjaval Raj\AppData\Roaming\Python\Python311\site-packages\IPython\core\py
labtools.py:152: UserWarning: Glyph 128202 (\N{BAR CHART}) missing from current fon
t.
  fig.canvas.print_figure(bytes_io, **kw)
```



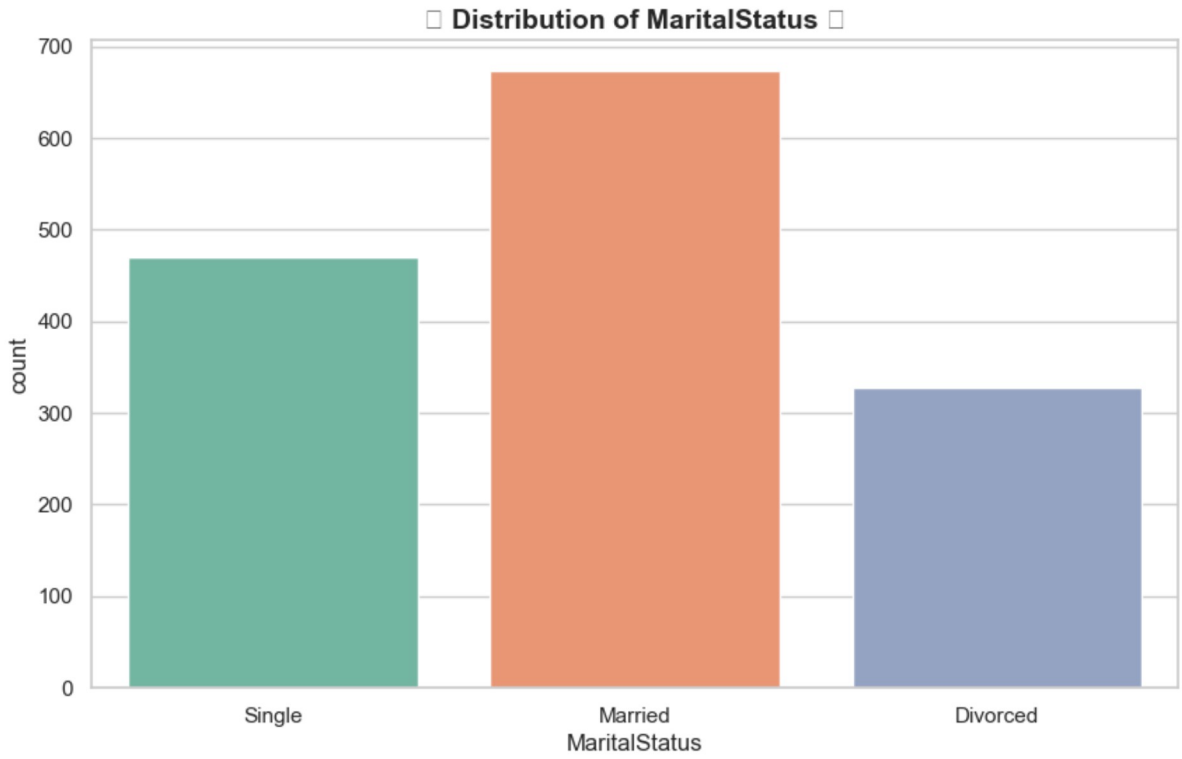

```
C:\Users\Ujjaval Raj\AppData\Roaming\Python\Python311\site-packages\IPython\core\py
labtools.py:152: UserWarning: Glyph 128202 (\N{BAR CHART}) missing from current fon
t.
  fig.canvas.print_figure(bytes_io, **kw)
```




```
C:\Users\Ujjaval Raj\AppData\Roaming\Python\Python311\site-packages\IPython\core\py
labtools.py:152: UserWarning: Glyph 128202 (\N{BAR CHART}) missing from current fon
t.
  fig.canvas.print_figure(bytes_io, **kw)
```

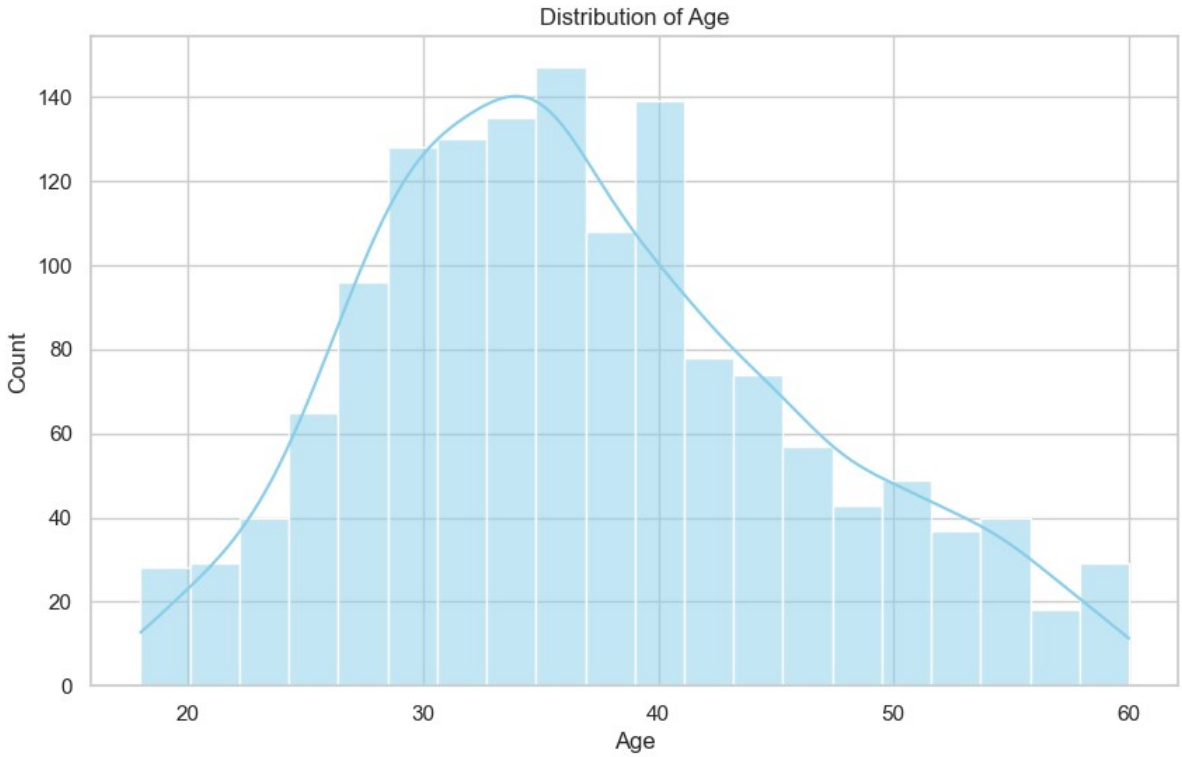


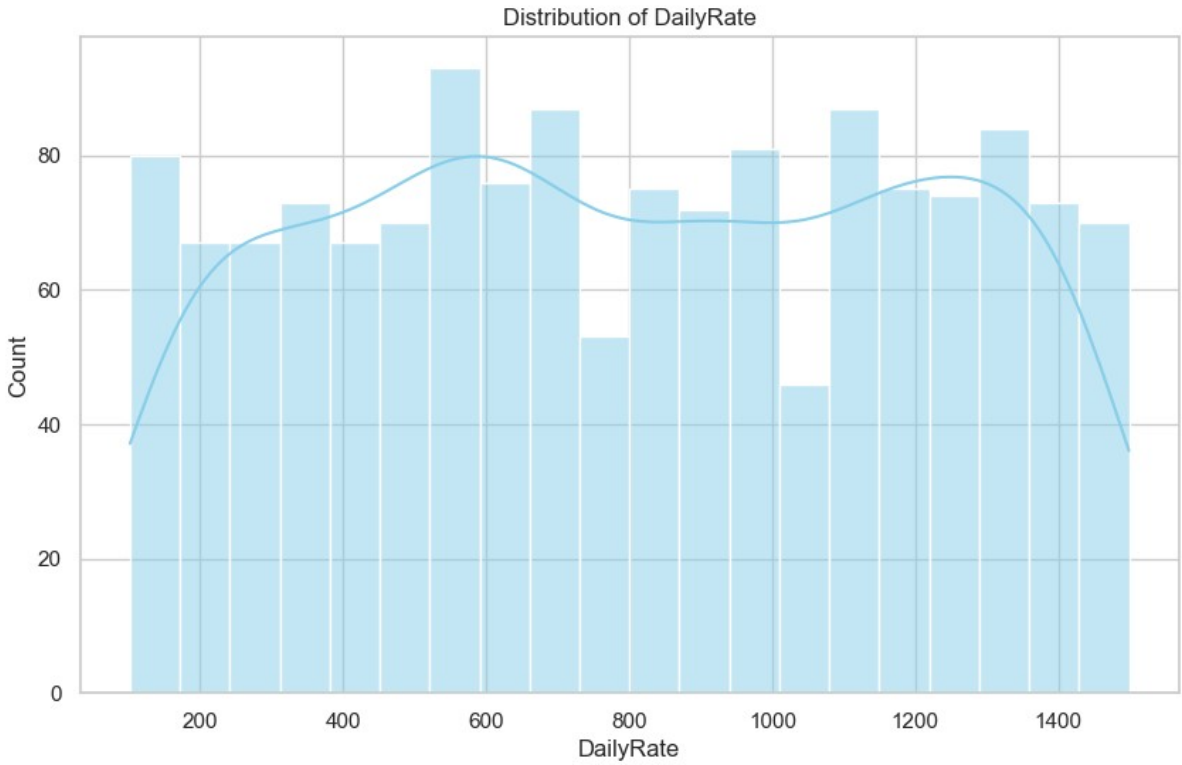

```
C:\Users\Ujjaval Raj\AppData\Roaming\Python\Python311\site-packages\IPython\core\py
labtools.py:152: UserWarning: Glyph 128202 (\N{BAR CHART}) missing from current fon
t.
  fig.canvas.print_figure(bytes_io, **kw)
```

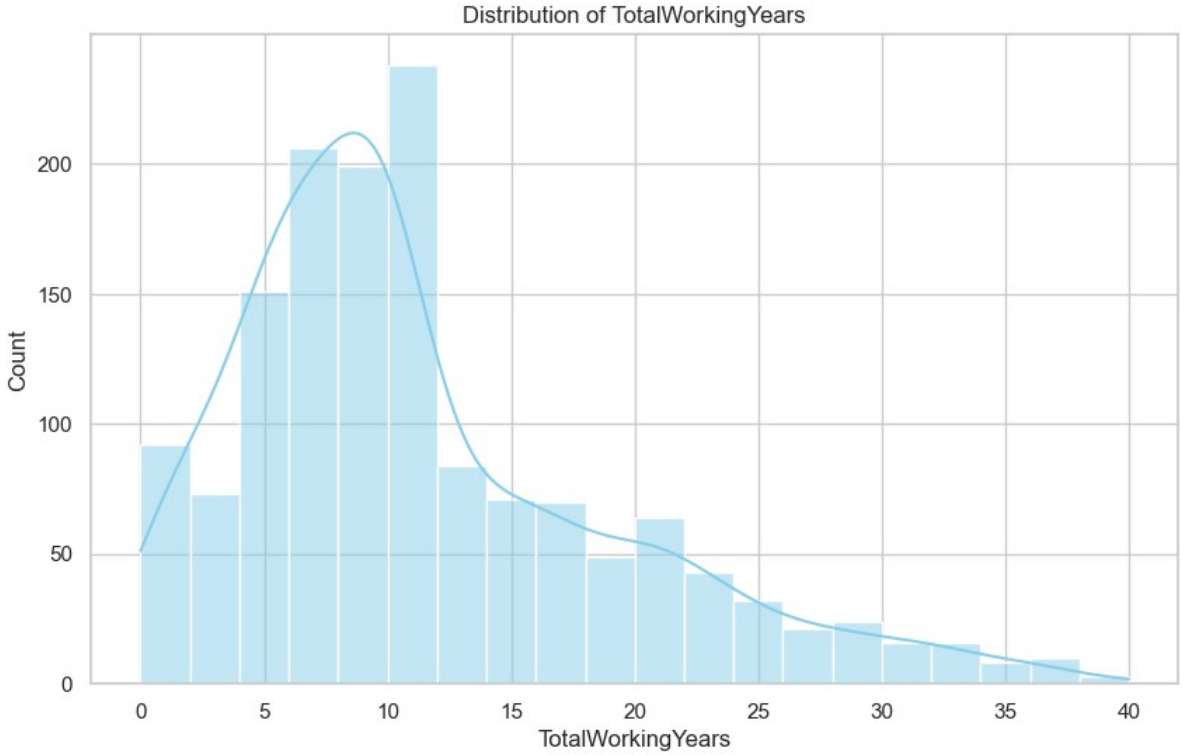


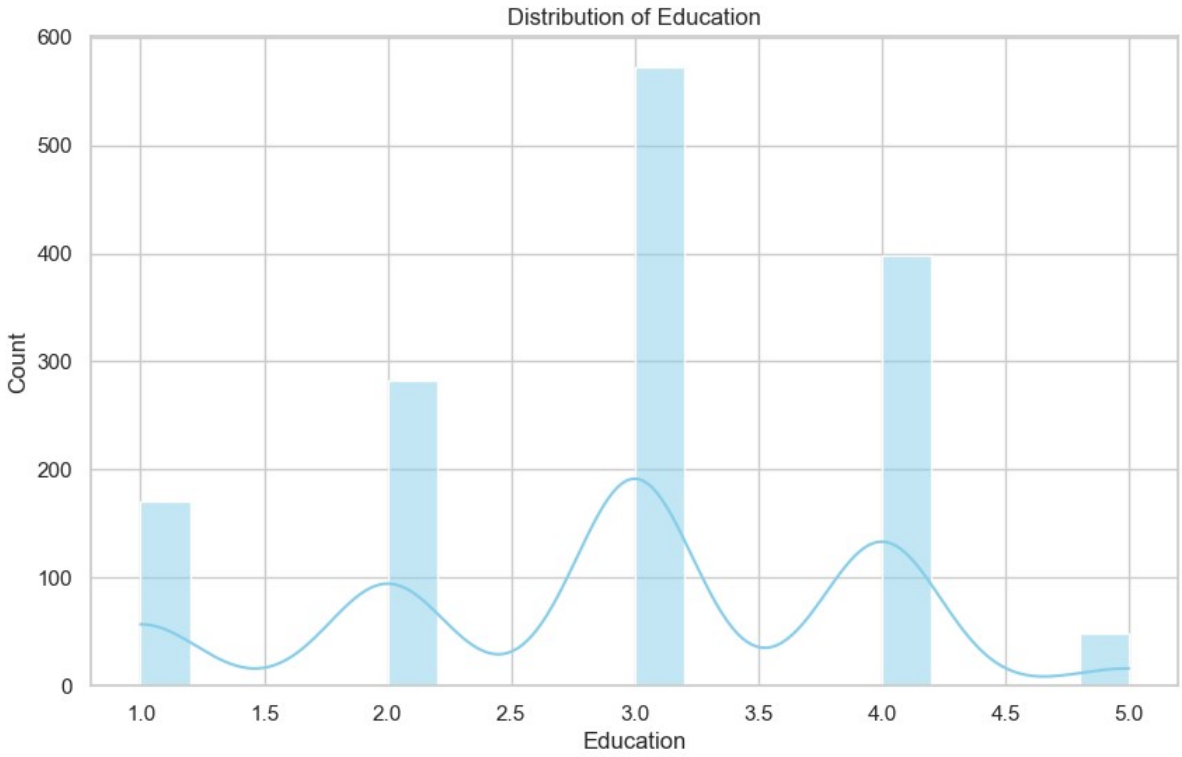
Plot a correlation map for all numeric variables

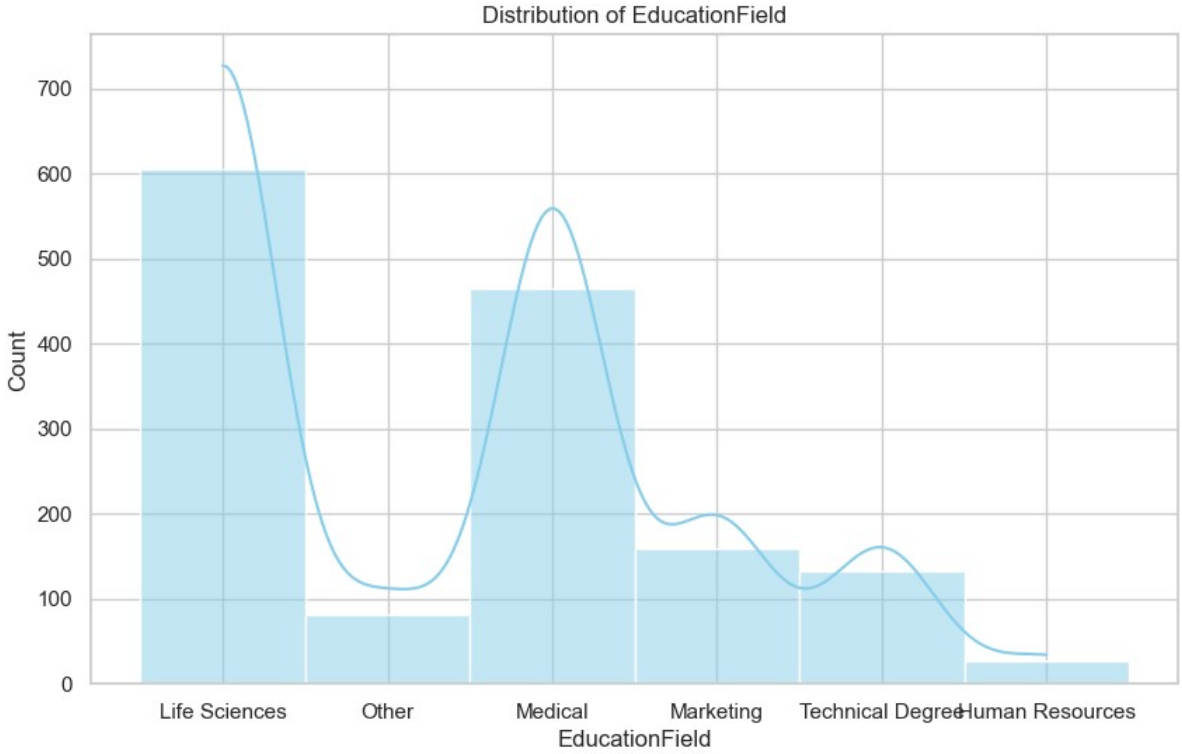
In [41]:
```
# Visualizations for numeric variables
numeric_variables = ['Age', 'DailyRate','TotalWorkingYears', 'Education', 'Educatio

for variable in numeric_variables:
    plt.figure(figsize=(10, 6))
    sns.histplot(HR_data[variable], bins=20, kde=True, color='skyblue')
    plt.title(f'Distribution of {variable}')
    plt.show()
```

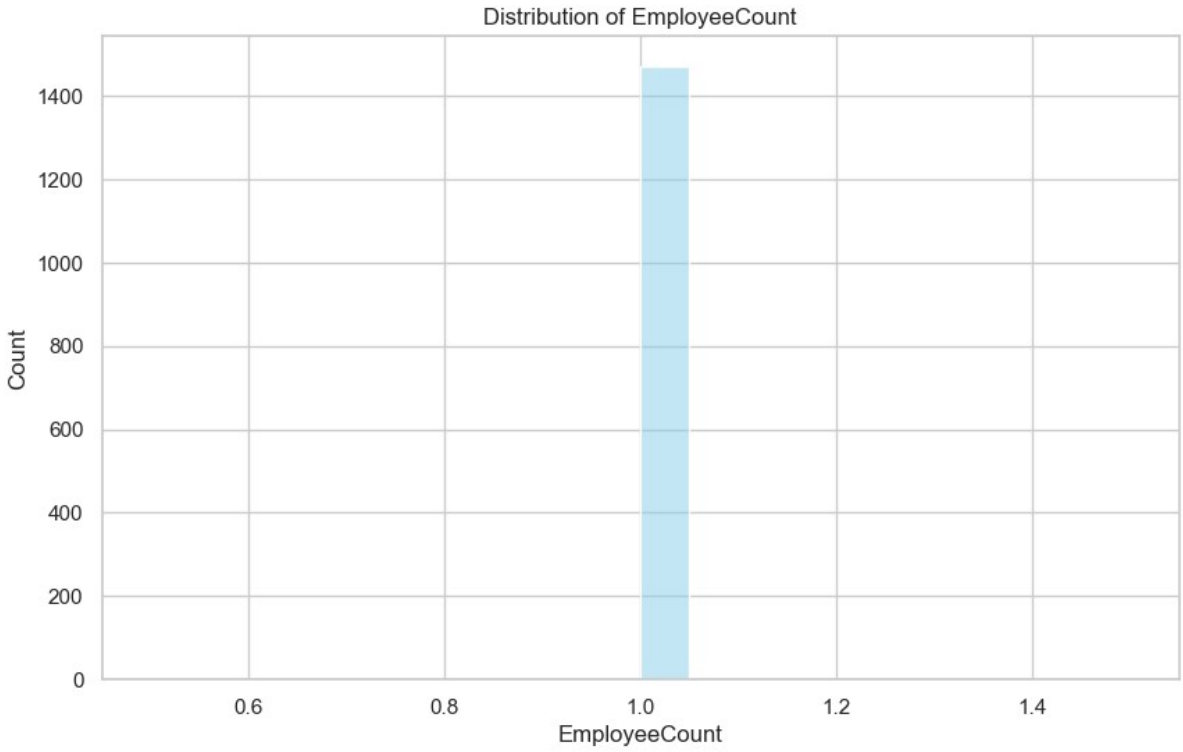

Distribution of Age
Count
Age
0
20
40
60
80
100
120
140
20
30
40
50
60

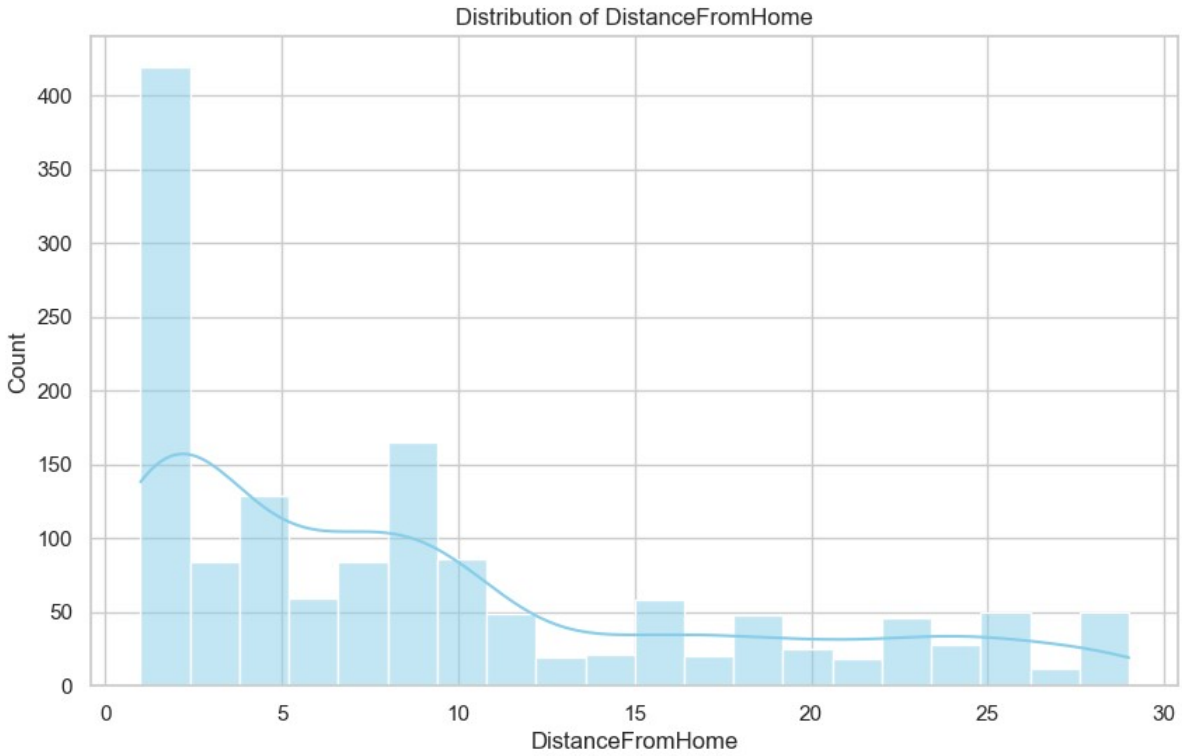

Distribution of DailyRate
Count
DailyRate
0
20
40
60
80
200
400
600
800
1000
1200
1400


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


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

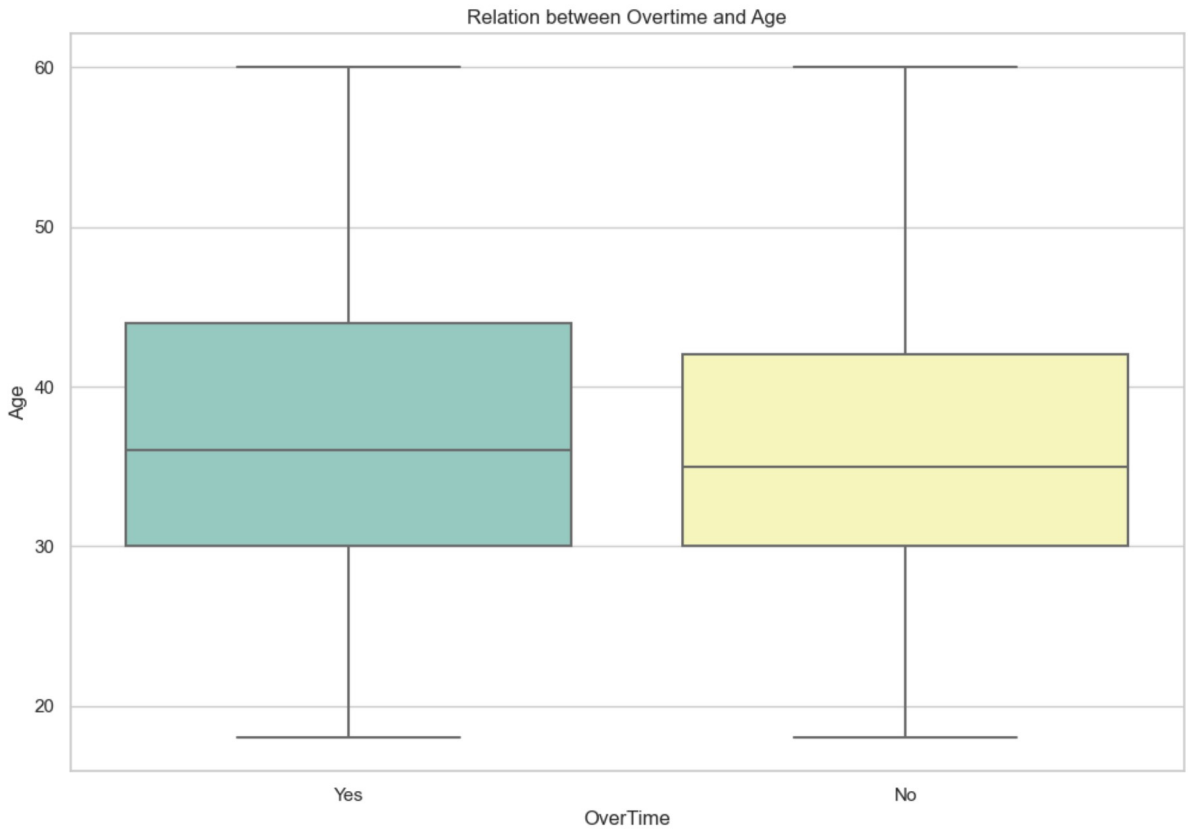

Distribution of EducationField
Count
0
100
200
300
400
500
600
700
Life Sciences
Other
Medical
Marketing
Technical Degree
Human Resources
EducationField


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



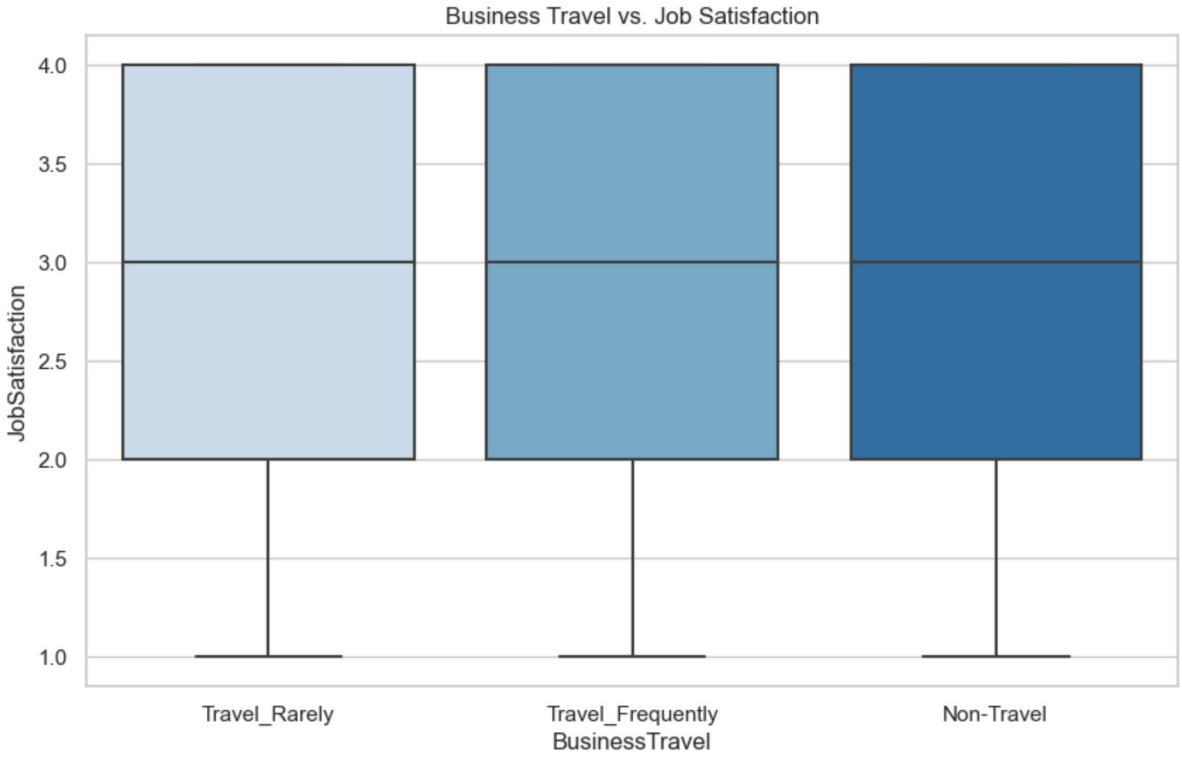
Relation between Overtime and Age

In [46]:
```
plt.figure(figsize=(12, 8))
sns.boxplot(x='OverTime', y='Age', data=HR_data, palette='Set3')
plt.title('Relation between Overtime and Age')
plt.show()
```



Relation between BuisenessTravel and JObSatisfaction

In [47]:
```
plt.figure(figsize=(10, 6))
sns.boxplot(x='BusinessTravel', y='JobSatisfaction', data=HR_data, palette='Blues')
plt.title('Business Travel vs. Job Satisfaction')
plt.show()
```



Relation between MaritalStatus and TotalWorkingYears

In [48]:
```
plt.figure(figsize=(10, 6))
sns.boxplot(x='MaritalStatus', y='TotalWorkingYears', data=HR_data, palette='husl')
plt.title('Marital Status vs. Total Working Years')
plt.show()
```