Employee Attrition Prediction using Classification Tree

McCurr Healthcare Consultancy is an MNC that has thousands of employees spread out across the globe. The company believes in hiring the best talent available and retaining them for as long as possible. A huge amount of resources are spent on retaining existing employees through various initiatives. The Head of People Operations wants to bring down the cost of retaining employees. For this, he proposes limiting the incentives to only those employees who are at risk of attrition. As a recently hired Data Scientist in the People Operations Department, you have been asked to identify patterns in characteristics of employees who leave the organization. Also, you have to use this information to predict if an employee is at risk of attrition. This information will be used to target them with incentives.

Objective:

- To identify the different factors that drive attrition
- To make a model to predict if an employee will attrite or not

Dataset:

The data contains demographic details, work-related metrics and attrition flag.

- **EmployeeNumber** - Employee Identifier
- **Attrition** - Did the employee attrite?
- **Age** - Age of the employee
- **BusinessTravel** - Travel commitments for the job
- **DailyRate** - Data description not available**
- **Department** - Employee Department
- **DistanceFromHome** - Distance from work to home (in km)
- **Education** - 1-Below College, 2-College, 3-Bachelor, 4-Master,5-Doctor
- **EducationField** - Field of Education
- **EnvironmentSatisfaction** - 1-Low, 2-Medium, 3-High, 4-Very High
- **Gender** - Employee's gender
- **HourlyRate** - Data description not available**
- **JobInvolvement** - 1-Low, 2-Medium, 3-High, 4-Very High
- **JobLevel** - Level of job (1 to 5)
- **JobRole** - Job Roles
- **JobSatisfaction** - 1-Low, 2-Medium, 3-High, 4-Very High
- **MaritalStatus** - Marital Status
- **MonthlyIncome** - Monthly Salary
- **MonthlyRate** - Data description not available**

- **NumCompaniesWorked** - Number of companies worked at
- **Over18** - Over 18 years of age?
- **OverTime** - Overtime?
- **PercentSalaryHike** - The percentage increase in salary last year
- **PerformanceRating** - 1-Low, 2-Good, 3-Excellent, 4-Outstanding
- **RelationshipSatisfaction** - 1-Low, 2-Medium, 3-High, 4-Very High
- **StandardHours** - Standard Hours
- **StockOptionLevel** - Stock Option Level
- **TotalWorkingYears** - Total years worked
- **TrainingTimesLastYear** - Number of training attended last year
- **WorkLifeBalance** - 1-Low, 2-Good, 3-Excellent, 4-Outstanding
- **YearsAtCompany** - Years at Company
- **YearsInCurrentRole** - Years in the current role
- **YearsSinceLastPromotion** - Years since the last promotion
- **YearsWithCurrManager** - Years with the current manager

```
In [1]:  import pandas as pd
         import numpy as np
         import matplotlib.pyplot as plt
         import seaborn as sns

         from sklearn.preprocessing import StandardScaler # scale using z-score
         from sklearn.model_selection import train_test_split

         # Algorithms to use
         from sklearn.tree import DecisionTreeClassifier, plot_tree
         from sklearn.linear_model import LogisticRegression

         # Metrics for model evaluation
         from sklearn.metrics import confusion_matrix, classification_report, precisi

         # Tuning the model
         from sklearn.model_selection import GridSearchCV

         # # Ignore warnings
         # import warnings
         # warnings.filterwarnings("ignore")
```

```
In [2]:  # Read the CSV file from the local directory
         file_path = "../data/Data_HR_Employee_Attrition.csv"
         employees = pd.read_csv(file_path)
         employees
```

Out[2]:

	EmployeeNumber	Attrition	Age	BusinessTravel	DailyRate	Department	Dist
0	1	Yes	41	Travel_Rarely	1102	Sales	
1	2	No	49	Travel_Frequently	279	Research & Development	
2	3	Yes	37	Travel_Rarely	1373	Research & Development	
3	4	No	33	Travel_Frequently	1392	Research & Development	
4	5	No	27	Travel_Rarely	591	Research & Development	
...	...	...	...	...	...	...	
2935	2936	No	36	Travel_Frequently	884	Research & Development	
2936	2937	No	39	Travel_Rarely	613	Research & Development	
2937	2938	No	27	Travel_Rarely	155	Research & Development	
2938	2939	No	49	Travel_Frequently	1023	Sales	
2939	2940	No	34	Travel_Rarely	628	Research & Development	

2940 rows × 34 columns

In [3]:
```python
import requests
from io import StringIO

orig_url="https://drive.google.com/file/d/147Z67u4-bp_ZVlbc18dg6J3h9ORRlCcW/

file_id = orig_url.split('/')[-2]
dwn_url='https://drive.google.com/uc?export=download&id=' + file_id
url = requests.get(dwn_url).text
csv_raw = StringIO(url)

employees = pd.read_csv(csv_raw)
employees
```

Out[3]:

	EmployeeNumber	Attrition	Age	BusinessTravel	DailyRate	Department	Dis
0	1	Yes	41	Travel_Rarely	1102	Sales	
1	2	No	49	Travel_Frequently	279	Research & Development	
2	3	Yes	37	Travel_Rarely	1373	Research & Development	
3	4	No	33	Travel_Frequently	1392	Research & Development	
4	5	No	27	Travel_Rarely	591	Research & Development	
...	...	...	...	...	...	...	
2935	2936	No	36	Travel_Frequently	884	Research & Development	
2936	2937	No	39	Travel_Rarely	613	Research & Development	
2937	2938	No	27	Travel_Rarely	155	Research & Development	
2938	2939	No	49	Travel_Frequently	1023	Sales	
2939	2940	No	34	Travel_Rarely	628	Research & Development	

2940 rows × 34 columns

Exploratory Data Analysis (EDA)

In [4]:
```python
employees.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2940 entries, 0 to 2939
Data columns (total 34 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   EmployeeNumber            2940 non-null   int64
 1   Attrition                 2940 non-null   object
 2   Age                       2940 non-null   int64
 3   BusinessTravel            2940 non-null   object
 4   DailyRate                 2940 non-null   int64
 5   Department                2940 non-null   object
 6   DistanceFromHome          2940 non-null   int64
 7   Education                 2940 non-null   int64
 8   EducationField            2940 non-null   object
 9   EnvironmentSatisfaction   2940 non-null   int64
 10  Gender                    2940 non-null   object
 11  HourlyRate                2940 non-null   int64
 12  JobInvolvement            2940 non-null   int64
 13  JobLevel                  2940 non-null   int64
 14  JobRole                   2940 non-null   object
 15  JobSatisfaction           2940 non-null   int64
 16  MaritalStatus             2940 non-null   object
 17  MonthlyIncome             2940 non-null   int64
 18  MonthlyRate               2940 non-null   int64
 19  NumCompaniesWorked        2940 non-null   int64
 20  Over18                    2940 non-null   object
 21  OverTime                  2940 non-null   object
 22  PercentSalaryHike         2940 non-null   int64
 23  PerformanceRating         2940 non-null   int64
 24  RelationshipSatisfaction  2940 non-null   int64
 25  StandardHours             2940 non-null   int64
 26  StockOptionLevel          2940 non-null   int64
 27  TotalWorkingYears         2940 non-null   int64
 28  TrainingTimesLastYear     2940 non-null   int64
 29  WorkLifeBalance           2940 non-null   int64
 30  YearsAtCompany            2940 non-null   int64
 31  YearsInCurrentRole        2940 non-null   int64
 32  YearsSinceLastPromotion   2940 non-null   int64
 33  YearsWithCurrManager      2940 non-null   int64
dtypes: int64(25), object(9)
memory usage: 781.1+ KB
```

Observation:

- There are 2940 observations and 34 columns.
- All the column have 2940 non-null values i.e. there are no missing values in the data.

```python
In [5]:  # Checking unique values in each column
         employees.nunique()
```

```
Out[5]:  EmployeeNumber                2940
         Attrition                        2
         Age                             43
         BusinessTravel                   3
         DailyRate                      886
         Department                       3
         DistanceFromHome                29
         Education                        5
         EducationField                   6
         EnvironmentSatisfaction          4
         Gender                           2
         HourlyRate                      71
         JobInvolvement                   4
         JobLevel                         5
         JobRole                          9
         JobSatisfaction                  4
         MaritalStatus                    3
         MonthlyIncome                 1349
         MonthlyRate                   1427
         NumCompaniesWorked              10
         Over18                           1
         OverTime                         2
         PercentSalaryHike               15
         PerformanceRating                2
         RelationshipSatisfaction         4
         StandardHours                    1
         StockOptionLevel                 4
         TotalWorkingYears               40
         TrainingTimesLastYear            7
         WorkLifeBalance                  4
         YearsAtCompany                  37
         YearsInCurrentRole              19
         YearsSinceLastPromotion         16
         YearsWithCurrManager            18
         dtype: int64
```

Observation:

- `EmployeeNumber` is an identifier which is unique for each employee and we can drop this column as it would not add any value to our analysis.
- `Over18` and `StandardHours` have only 1 unique value. These column will not add any value to our model hence we can drop them.
- On the basis of number of unique values in each column and the data description, we can identify the continuous and categorical columns in the data.

Let's drop the columns mentioned above and define lists for numerical and categorical columns to apply explore them separately.

```
In [6]:  # Dropping the columns
         employees=employees.drop(['EmployeeNumber','Over18','StandardHours'],axis=1)
```

```
In [7]:  # Creating numerical columns
         num_cols=['DailyRate','Age','DistanceFromHome','MonthlyIncome','MonthlyRate'
```

```
            'YearsAtCompany','NumCompaniesWorked','HourlyRate',
            'YearsInCurrentRole','YearsSinceLastPromotion','YearsWithCurrManag

# Creating categorical variables
cat_cols= ['Attrition','OverTime','BusinessTravel', 'Department','Education'
           'StockOptionLevel','Gender', 'PerformanceRating', 'JobInvolvement
```

Univariate Analysis

In [8]:
```python
# Checking summary statistics of numerical columns
employees[num_cols].describe().T
```

Out[8]:

	count	mean	std	min	25%	50%
DailyRate	2940.0	802.485714	403.440447	102.0	465.0	802.0
Age	2940.0	36.923810	9.133819	18.0	30.0	36.0
DistanceFromHome	2940.0	9.192517	8.105485	1.0	2.0	7.0
MonthlyIncome	2940.0	6502.931293	4707.155770	1009.0	2911.0	4919.0
MonthlyRate	2940.0	14313.103401	7116.575021	2094.0	8045.0	14235.5
PercentSalaryHike	2940.0	15.209524	3.659315	11.0	12.0	14.0
TotalWorkingYears	2940.0	11.279592	7.779458	0.0	6.0	10.0
YearsAtCompany	2940.0	7.008163	6.125483	0.0	3.0	5.0
NumCompaniesWorked	2940.0	2.693197	2.497584	0.0	1.0	2.0
HourlyRate	2940.0	65.891156	20.325969	30.0	48.0	66.0
YearsInCurrentRole	2940.0	4.229252	3.622521	0.0	2.0	3.0
YearsSinceLastPromotion	2940.0	2.187755	3.221882	0.0	0.0	1.0
YearsWithCurrManager	2940.0	4.123129	3.567529	0.0	2.0	3.0
TrainingTimesLastYear	2940.0	2.799320	1.289051	0.0	2.0	3.0

- **Average employee age is around 37 years**. It has a high range, from 18 years to 60, indicating good age diversity in the organization.
- **At least 50% of the employees live within a 7 km radius** from the organization. However there are some extreme values, seeing as the maximum value is 29 km.
- **The average monthly income of an employee is USD 6500.** It has a high range of values from 1K-20K, which is to be expected for any organization's income distribution. There is a big difference between the 3rd quartile value (around USD 8400) and the maximum value (nearly USD 20000), showing that the **company's highest earners have a disproportionately large income** in comparison to the rest of the employees. Again, this is fairly common in most organizations.

- **Average salary hike of an employee is around 15%.** At least 50% of employees got a salary hike 14% or less, with the maximum salary hike being 25%.
- Average number of years an employee is associated with the company is 7.
- **On average, the number of years since an employee got a promotion is 2.18**. The majority of employees have been promoted since the last year.

```
In [9]:  # Observing the distribution of numerical columns
         employees[num_cols].hist(figsize=(14,14))
         plt.show()
```
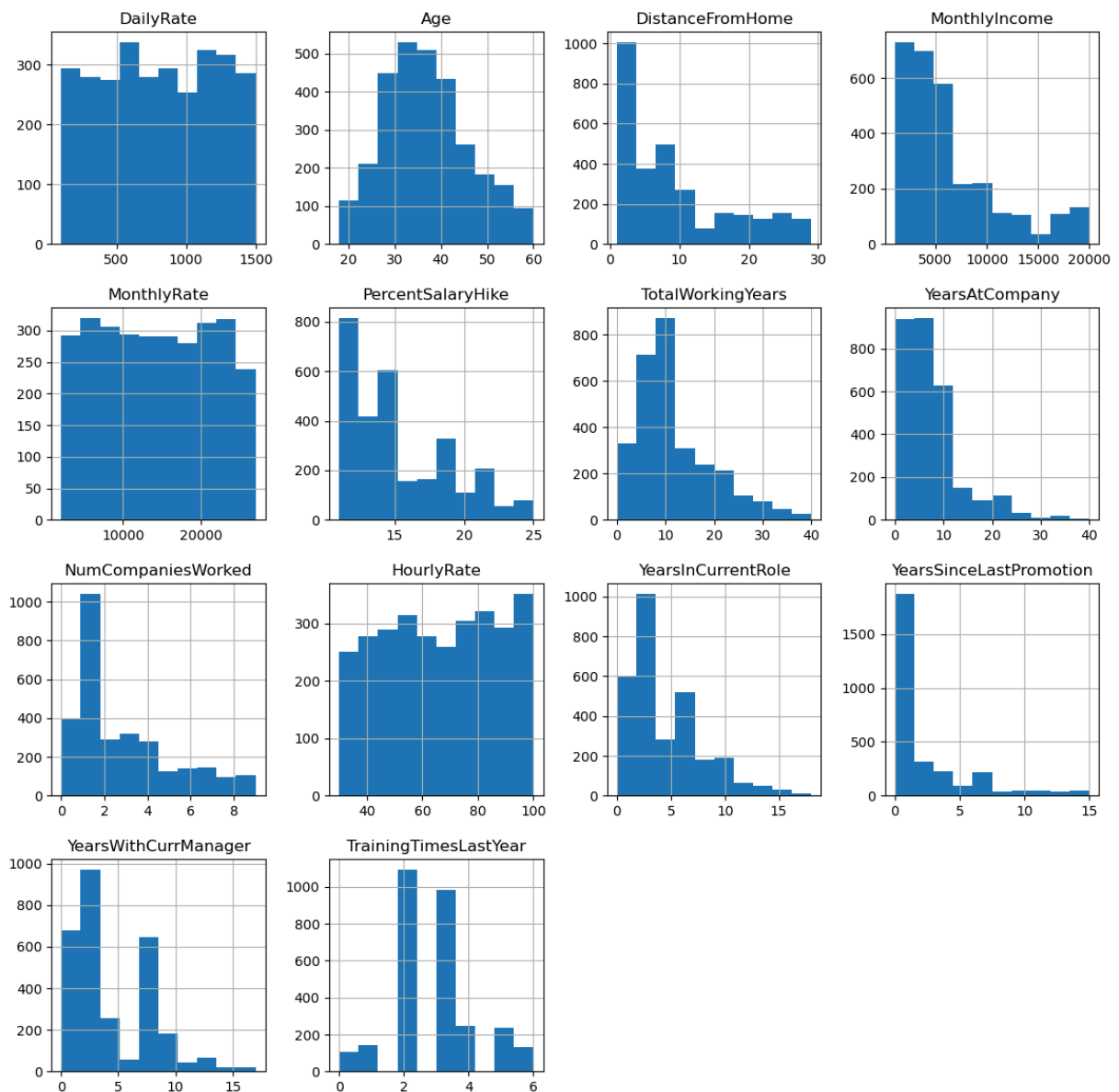

Observations:

- **The age distribution is close to a normal distribution** with the majority of employees between the ages of 25 and 50.

- **The percentage salary hike is skewed to the right**, which means employees are mostly getting lower percentage salary increases.

- **MonthlyIncome and TotalWorkingYears are skewed to the right**, indicating that the majority of workers are in entry / mid-level positions in the organization.

- **DistanceFromHome also has a right skewed distribution**, meaning most employees live close to work but there are a few that live further away.

- **On average, an employee has worked at 2.5 companies.** Most employees have worked at only 1 company.

- **The YearsAtCompany variable distribution shows a good proportion of workers with 10+ years**, indicating a significant number of loyal employees at the organization.

- **The YearsInCurrentRole distribution has three peaks at 0, 2, and 7.** There are a few employees that have even stayed in the same role for 15 years and more.

- **The YearsSinceLastPromotion variable distribution indicates that some employees have not received a promotion in 10-15 years and are still working in the organization.** These employees are assumed to be high work-experience employees in upper-management roles, such as co-founders, C-suite employees and the like.

- The distributions of DailyRate, HourlyRate and MonthlyRate appear to be uniform and do not provide much information. It could be that daily rate refers to the income earned per extra day worked while hourly rate could refer to the same concept applied for extra hours worked per day. Since these rates tend to be broadly similiar for multiple employees in the same department, that explains the uniform distribution they show.

In [10]:
```python
# Checking the proportion of sub-categories in each categorical column
for i in cat_cols:
    # This expresses the results as proportions rather than counts.
    print(employees[i].value_counts(normalize=True))
    print('*'*40)
```

```
Attrition
No     0.838776
Yes    0.161224
Name: proportion, dtype: float64
****************************************
OverTime
No     0.717007
Yes    0.282993
Name: proportion, dtype: float64
****************************************
BusinessTravel
Travel_Rarely        0.709524
Travel_Frequently    0.188435
Non-Travel           0.102041
Name: proportion, dtype: float64
****************************************
Department
Research & Development    0.653741
Sales                     0.303401
Human Resources           0.042857
Name: proportion, dtype: float64
****************************************
Education
3    0.389116
4    0.270748
2    0.191837
1    0.115646
5    0.032653
Name: proportion, dtype: float64
****************************************
EducationField
Life Sciences       0.412245
Medical             0.315646
Marketing           0.108163
Technical Degree    0.089796
Other               0.055782
Human Resources     0.018367
Name: proportion, dtype: float64
****************************************
JobSatisfaction
4    0.312245
3    0.300680
1    0.196599
2    0.190476
Name: proportion, dtype: float64
****************************************
EnvironmentSatisfaction
3    0.308163
4    0.303401
2    0.195238
1    0.193197
Name: proportion, dtype: float64
****************************************
WorkLifeBalance
3    0.607483
2    0.234014
```

```
4     0.104082
1     0.054422
Name: proportion, dtype: float64
**************************************
StockOptionLevel
0     0.429252
1     0.405442
2     0.107483
3     0.057823
Name: proportion, dtype: float64
**************************************
Gender
Male      0.6
Female    0.4
Name: proportion, dtype: float64
**************************************
PerformanceRating
3     0.846259
4     0.153741
Name: proportion, dtype: float64
**************************************
JobInvolvement
3     0.590476
2     0.255102
4     0.097959
1     0.056463
Name: proportion, dtype: float64
**************************************
JobLevel
1     0.369388
2     0.363265
3     0.148299
4     0.072109
5     0.046939
Name: proportion, dtype: float64
**************************************
JobRole
Sales Executive             0.221769
Research Scientist          0.198639
Laboratory Technician       0.176190
Manufacturing Director      0.098639
Healthcare Representative    0.089116
Manager                     0.069388
Sales Representative        0.056463
Research Director           0.054422
Human Resources             0.035374
Name: proportion, dtype: float64
**************************************
MaritalStatus
Married     0.457823
Single      0.319728
Divorced    0.222449
Name: proportion, dtype: float64
**************************************
RelationshipSatisfaction
3     0.312245
```

```
4    0.293878
2    0.206122
1    0.187755
Name: proportion, dtype: float64
****************************************
```

Observations:

- **The employee attrition rate is 16%.**
- **Around 28% of the employees are working overtime.** This number appears to be on the higher side, and might indicate a stressed employee work-life.
- 71% of the employees have traveled rarely, while around 19% have to travel frequently.
- Around 73% of the employees come from an educational background in the Life Sciences and Medical fields.
- Over 65% of employees work in the Research & Development department of the organization.
- **Nearly 40% of the employees have low (1) or medium-low (2) job satisfaction** and environment satisfaction in the organization, indicating that the morale of the company appears to be somewhat low.
- **Over 30% of the employees show low (1) to medium-low (2) job involvement.**
- Over 80% of the employees either have none or very less stock options.
- **In terms of performance ratings, none of the employees have rated lower than 3 (excellent).** About 85% of employees have a performance rating equal to 3 (excellent), while the remaining have a rating of 4 (outstanding). This could either mean that the majority of employees are top performers, or the more likely scenerio is that the organization could be highly lenient with its performance appraisal process.

Problem 1 - Data Summary Observations:

1.a. The employee attrition rate is **16%.**

- Around 28% of the employees are working overtime.

- **Implications: This suggests that there could be a problem with work-life balance in the company.**
  High rates of overtime may lead to stress.

- 71% of the employees have traveled rarely, while around 19% have to travel frequently.

- Around 73% of the employees come from an educational background in the Life Sciences and Medical fields.

1.b. Over **65%** of employees work in the Research & Development department of the organization.

1.c. Nearly **38.7% (19.7 + 19.0)** of the employees have low (1) or medium-low (2) job satisfaction in the organization.

- **Implications: A significant percentage of employees are dissatisfied with their jobs. This could lead to low productivity, disengagement, and higher attrition rates. The company should investigate key reasons for dissatisfaction and implement strategies to improve employee satisfaction**

1.d. Over **31.1% (5.6 + 25.5)** of the employees show low (1) to medium-low (2) job involvement.

1.e. Over **83.4% (40.5 + 42.9)** of the employees either have none or very less stock options.

- In terms of performance ratings, none of the employees have rated lower than 3 (excellent). About 85% of employees have a performance rating equal to 3 (excellent), while the remaining have a rating of 4 (outstanding). This could either mean that the majority of employees are top performers, or the more likely scenerio is that the organization could be highly lenient with its performance appraisal process.

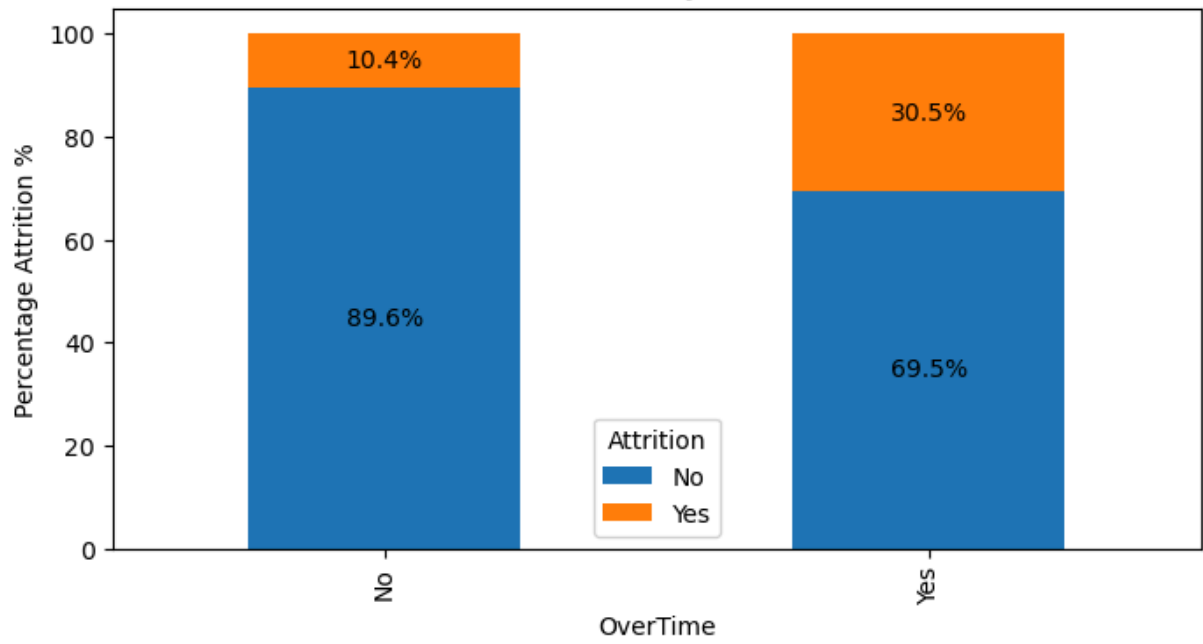## Bivariate and Multivariate analysis

In [11]:
```python
# for i in cat_cols:
#     if i!='Attrition':
#         (pd.crosstab(employees[i],employees['Attrition'],normalize='index'
#         plt.ylabel('Percentage Attrition %')

# Plot stacked bar charts with percentage labels
for i in cat_cols:
    if i != 'Attrition':  # Skip 'Attrition' itself
        attrition_rates = pd.crosstab(employees[i], employees['Attrition'],
        ax = attrition_rates.plot(kind='bar', figsize=(8, 4), stacked=True)

        # Add labels to each bar
        for p in ax.patches:
            width = p.get_width()
            height = p.get_height()
            x, y = p.get_xy()
            if height > 0:  # Only label non-zero bars
                ax.text(x + width / 2, y + height / 2, f'{height:.1f}%', ha=

        plt.ylabel('Percentage Attrition %')
        plt.title(f'Attrition Rate by {i}')
        plt.legend(title="Attrition", labels=['No', 'Yes'])
        plt.show()
```
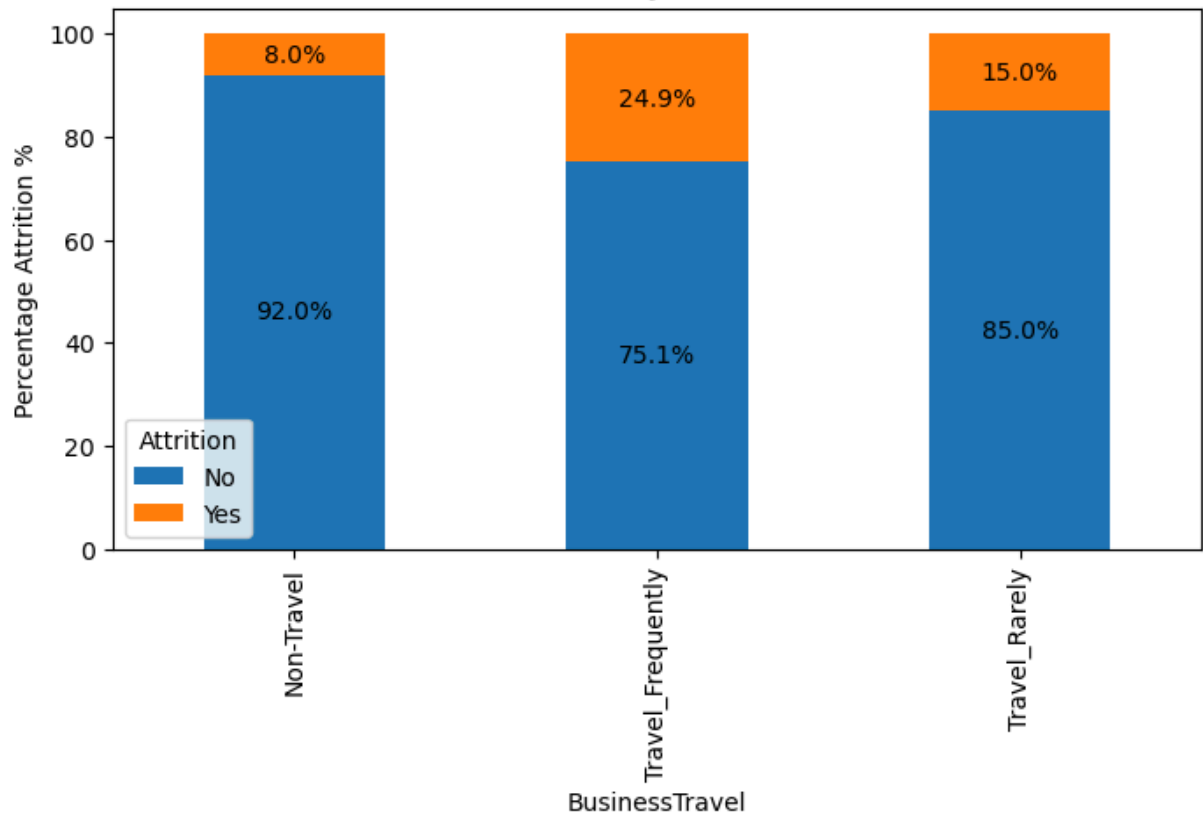
Attrition Rate by OverTime

Attrition Rate by BusinessTravel

Attrition Rate by Department

Percentage Attrition %

| Attrition |
| No |
| Yes |

- Human Resources: 19.0% / 81.0%
- Research & Development: 13.8% / 86.2%
- Sales: 20.6% / 79.4%

Attrition Rate by Education

Percentage Attrition %

| Attrition |
| No |
| Yes |

- 1: 18.2% / 81.8%
- 2: 15.6% / 84.4%
- 3: 17.3% / 82.7%
- 4: 14.6% / 85.4%
- 5: 10.4% / 89.6%

Attrition Rate by EducationField

Attrition Rate by JobSatisfaction
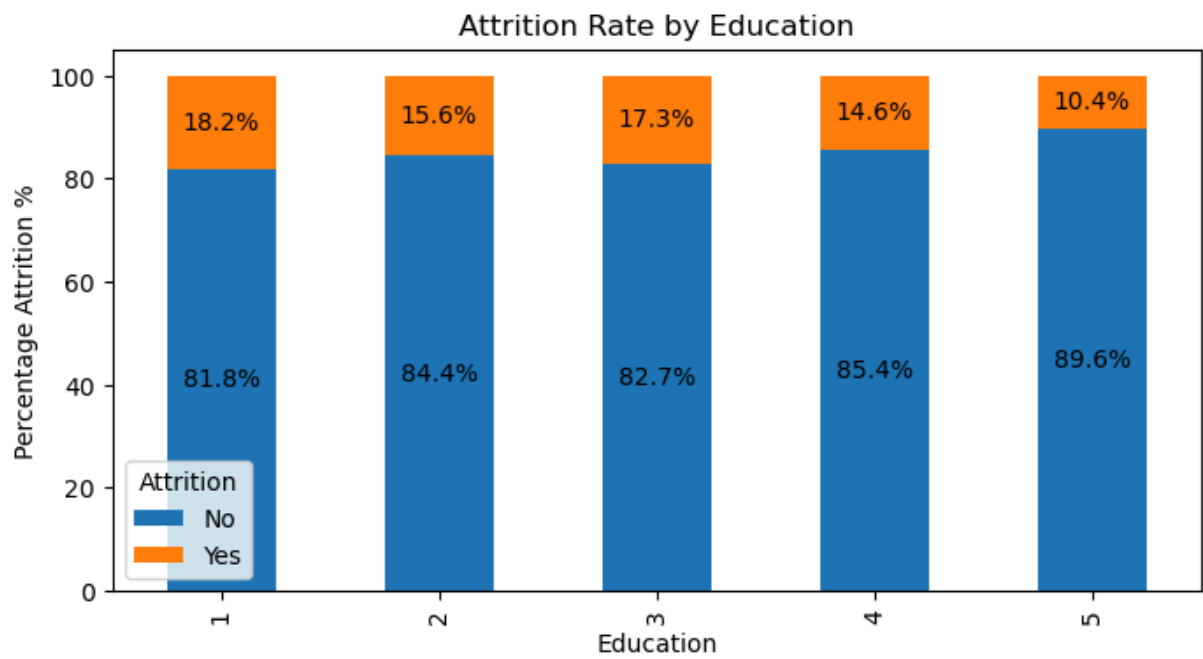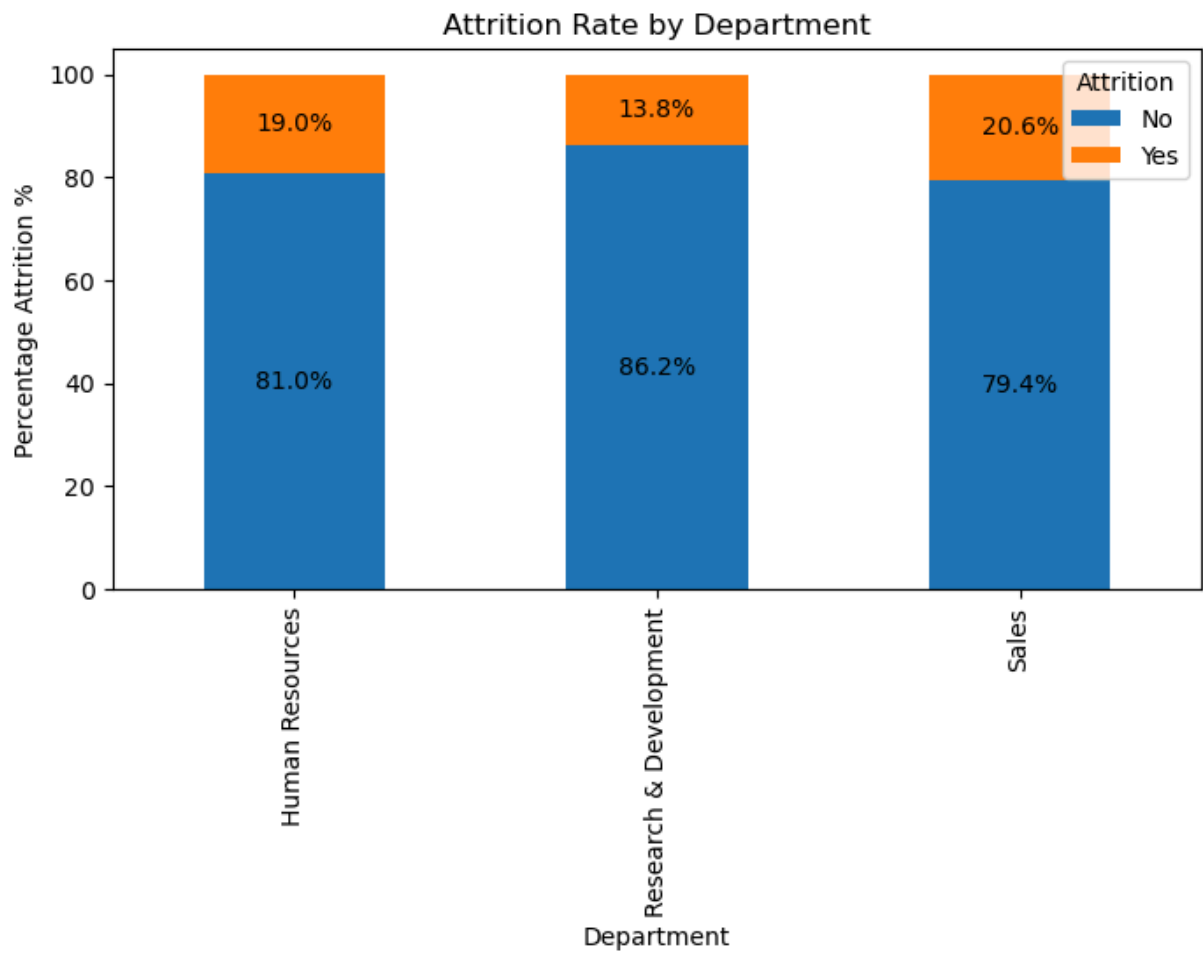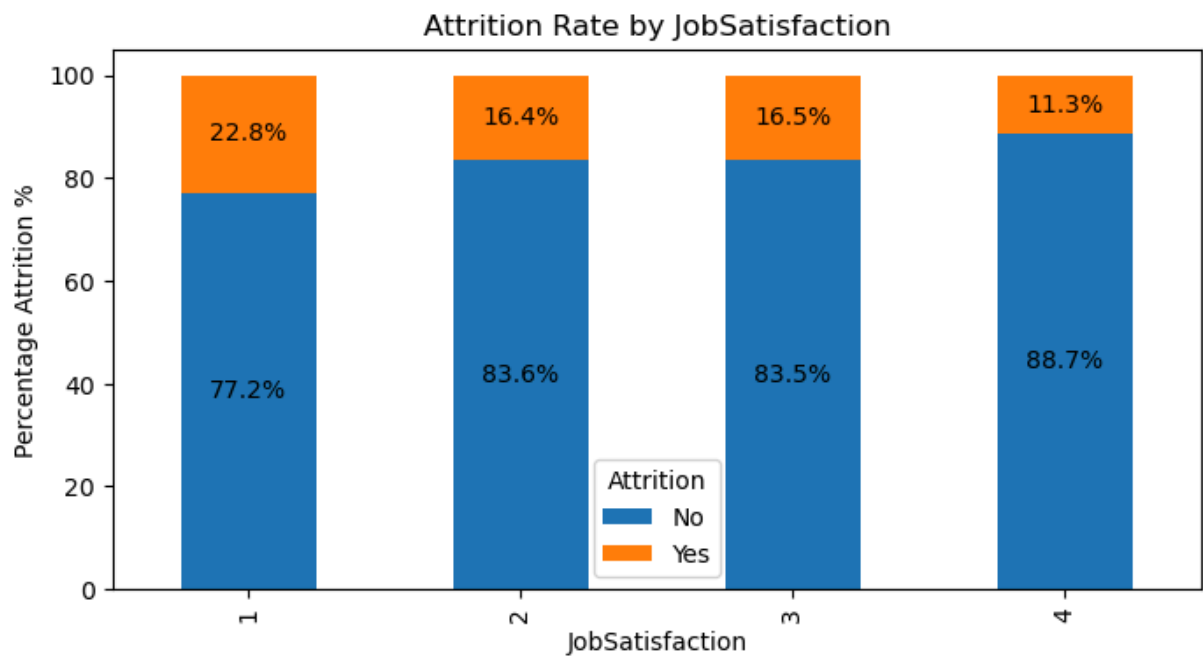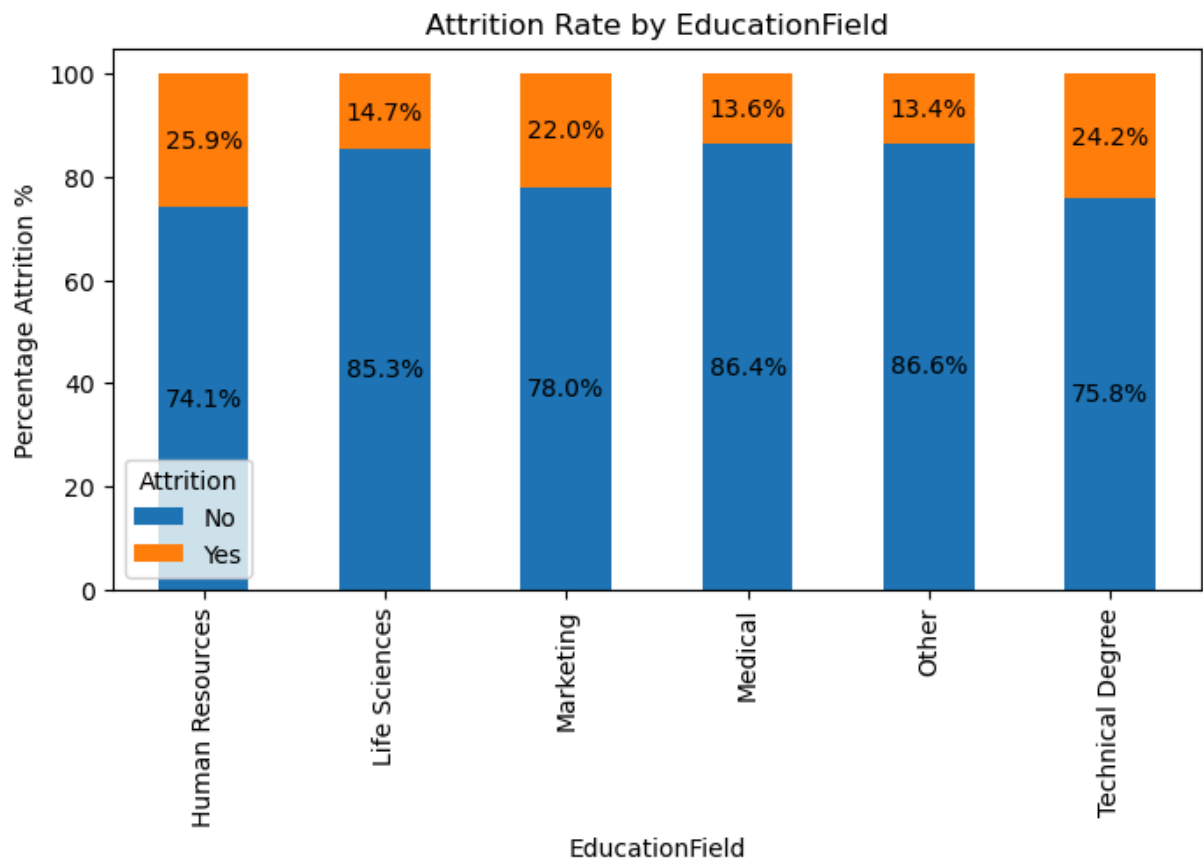
Attrition Rate by EnvironmentSatisfaction

Attrition Rate by WorkLifeBalance

Attrition Rate by StockOptionLevel

Percentage Attrition %

- 0: 75.6% No, 24.4% Yes
- 1: 90.6% No, 9.4% Yes
- 2: 92.4% No, 7.6% Yes
- 3: 82.4% No, 17.6% Yes

StockOptionLevel

Attrition
- No
- Yes

Attrition Rate by Gender

Percentage Attrition %

- Female: 85.2% No, 14.8% Yes
- Male: 83.0% No, 17.0% Yes

Gender

Attrition
- No
- Yes

Attrition Rate by PerformanceRating

Attrition Rate by JobInvolvement

Attrition Rate by JobLevel

Attrition Rate by JobRole

Attrition Rate by MaritalStatus

Attrition Rate by RelationshipSatisfaction


```
In [12]:  # Calculate attrition percentages for each category
          for i in cat_cols:
              if i != 'Attrition':  # Skip 'Attrition' itself
                  attrition_rates = pd.crosstab(employees[i], employees['Attrition'],
                  print(f"Attrition rates for {i}:\n{attrition_rates}\n")
```

```
Attrition rates for OverTime:
Attrition          No        Yes
OverTime
No            89.563567   10.436433
Yes           69.471154   30.528846

Attrition rates for BusinessTravel:
Attrition                  No        Yes
BusinessTravel
Non-Travel            92.000000    8.000000
Travel_Frequently    75.090253   24.909747
Travel_Rarely        85.043145   14.956855

Attrition rates for Department:
Attrition                      No        Yes
Department
Human Resources           80.952381   19.047619
Research & Development    86.160250   13.839750
Sales                     79.372197   20.627803

Attrition rates for Education:
Attrition          No        Yes
Education
1             81.764706   18.235294
2             84.397163   15.602837
3             82.692308   17.307692
4             85.427136   14.572864
5             89.583333   10.416667

Attrition rates for EducationField:
Attrition                No        Yes
EducationField
Human Resources     74.074074   25.925926
Life Sciences       85.313531   14.686469
Marketing           77.987421   22.012579
Medical             86.422414   13.577586
Other               86.585366   13.414634
Technical Degree    75.757576   24.242424

Attrition rates for JobSatisfaction:
Attrition              No        Yes
JobSatisfaction
1               77.162630   22.837370
2               83.571429   16.428571
3               83.484163   16.515837
4               88.671024   11.328976

Attrition rates for EnvironmentSatisfaction:
Attrition                      No        Yes
EnvironmentSatisfaction
1                       74.647887   25.352113
2                       85.017422   14.982578
3                       86.313466   13.686534
4                       86.547085   13.452915

Attrition rates for WorkLifeBalance:
```
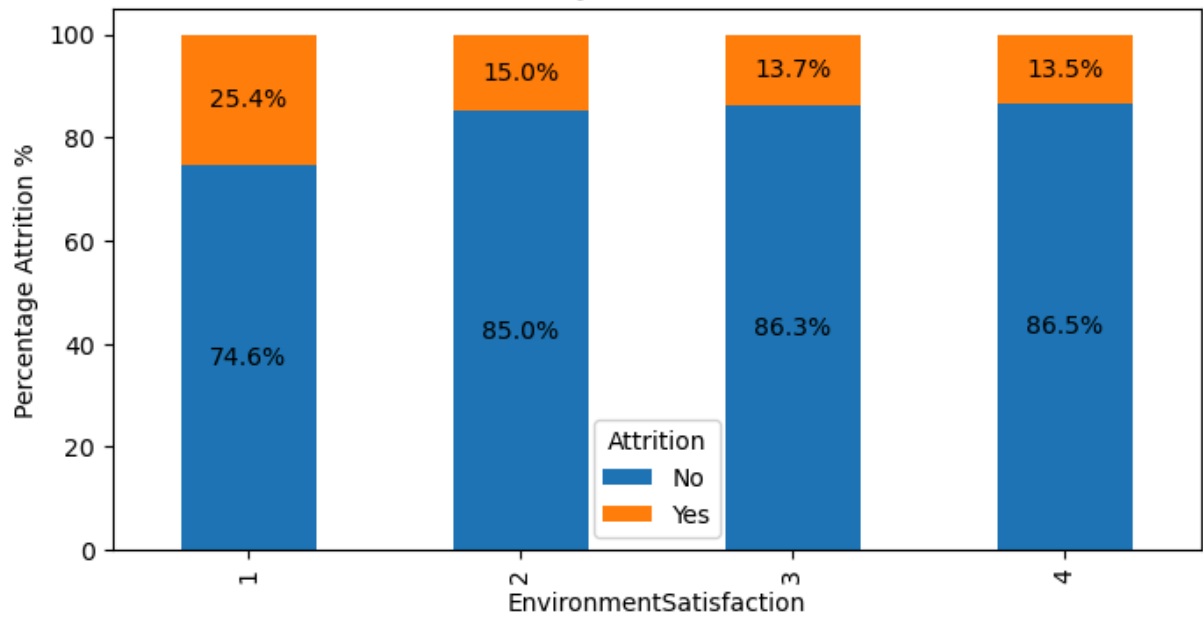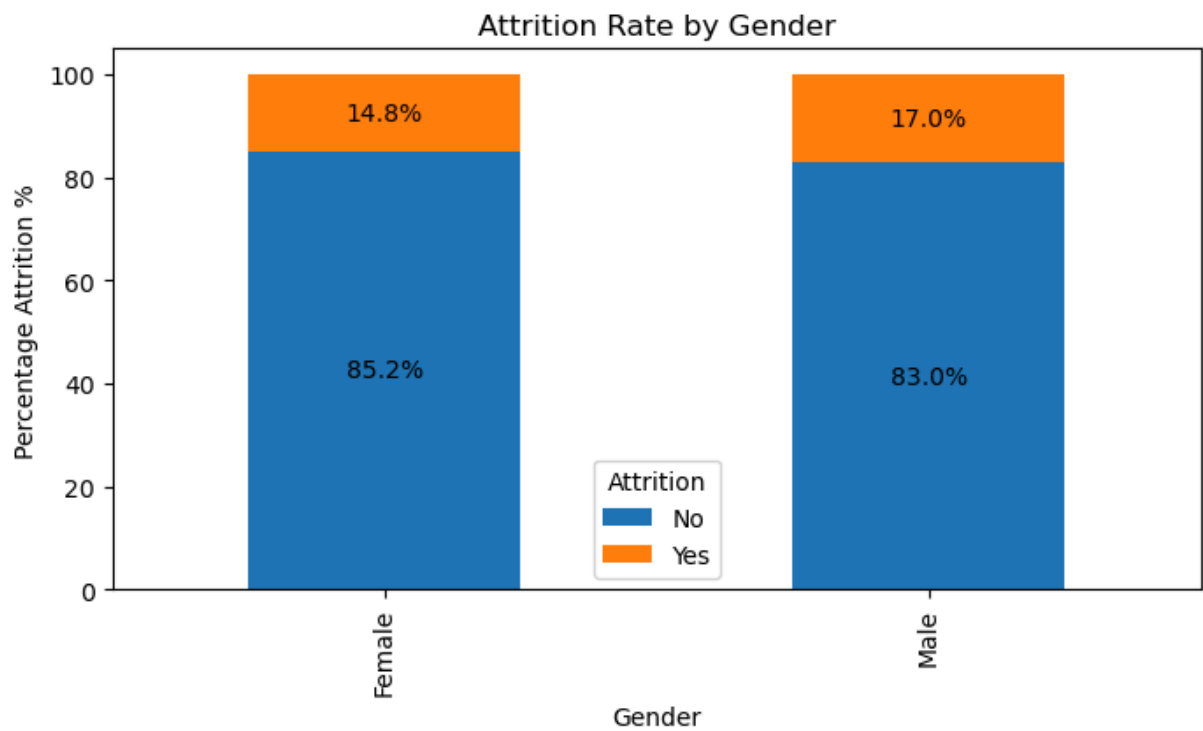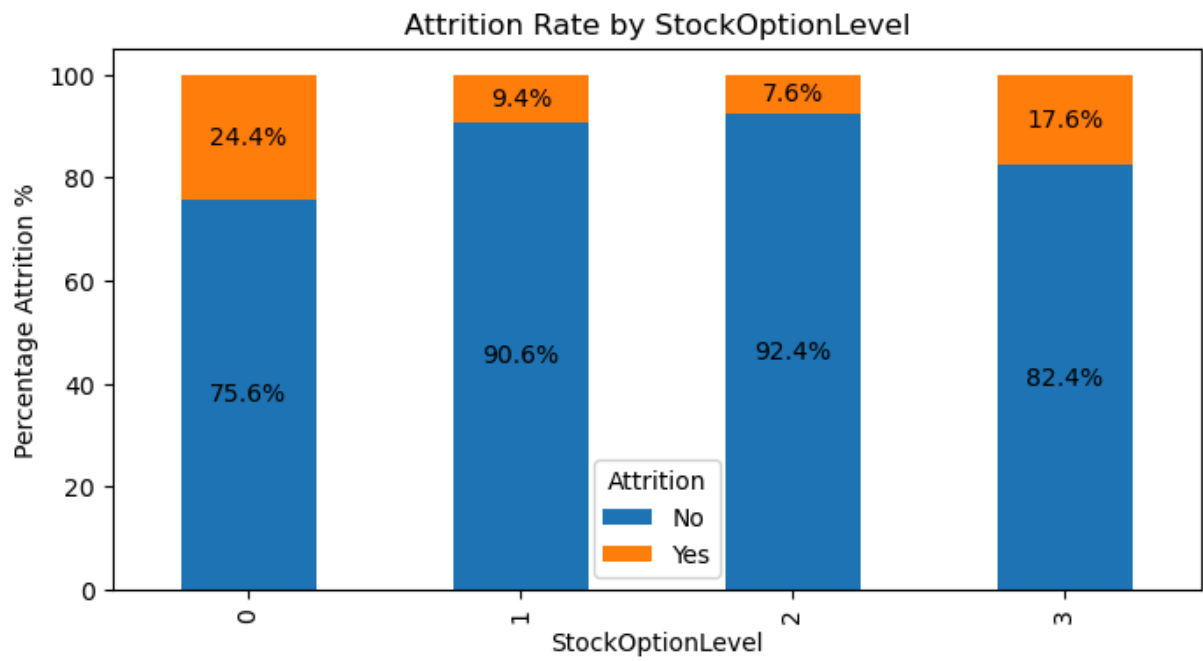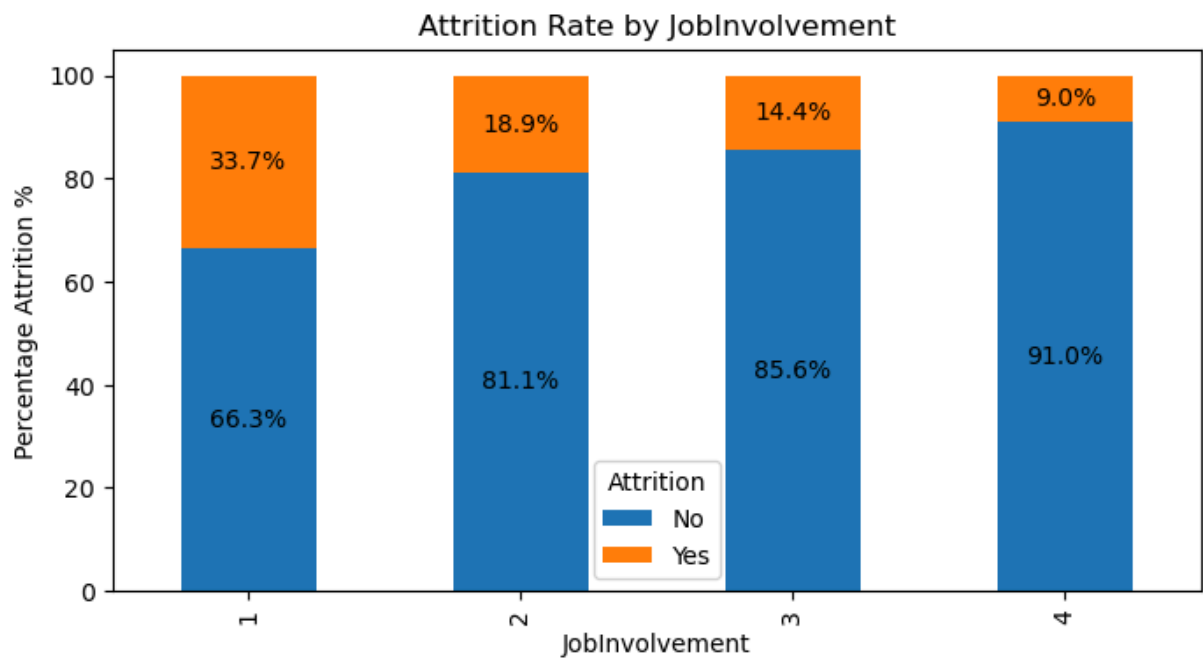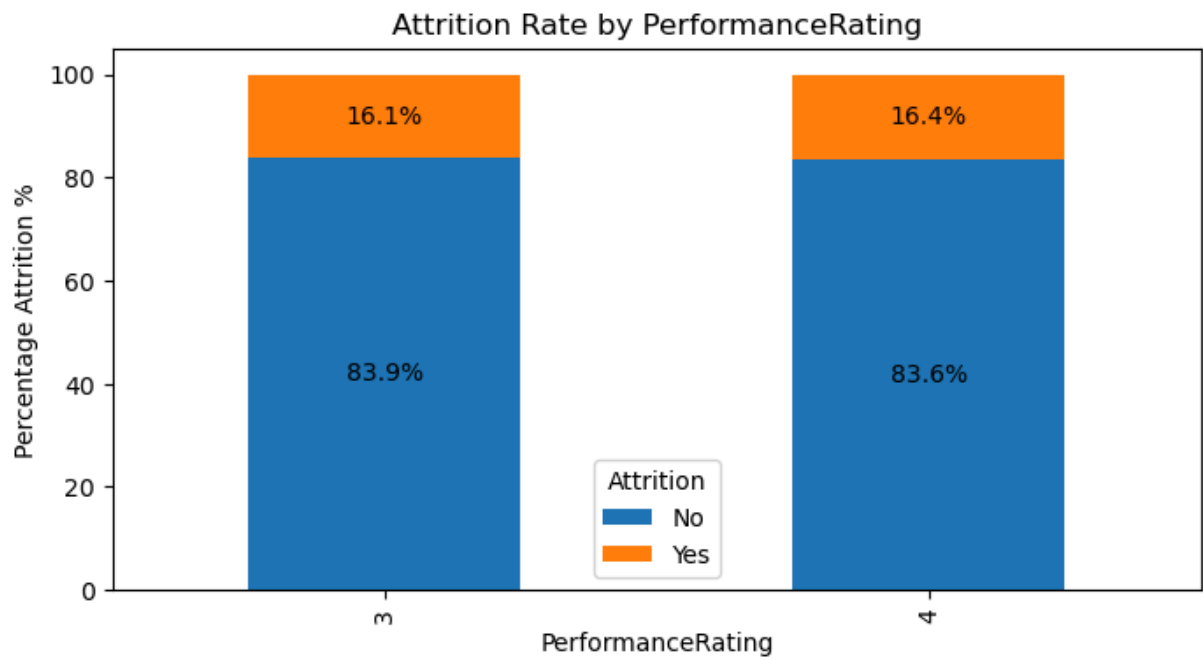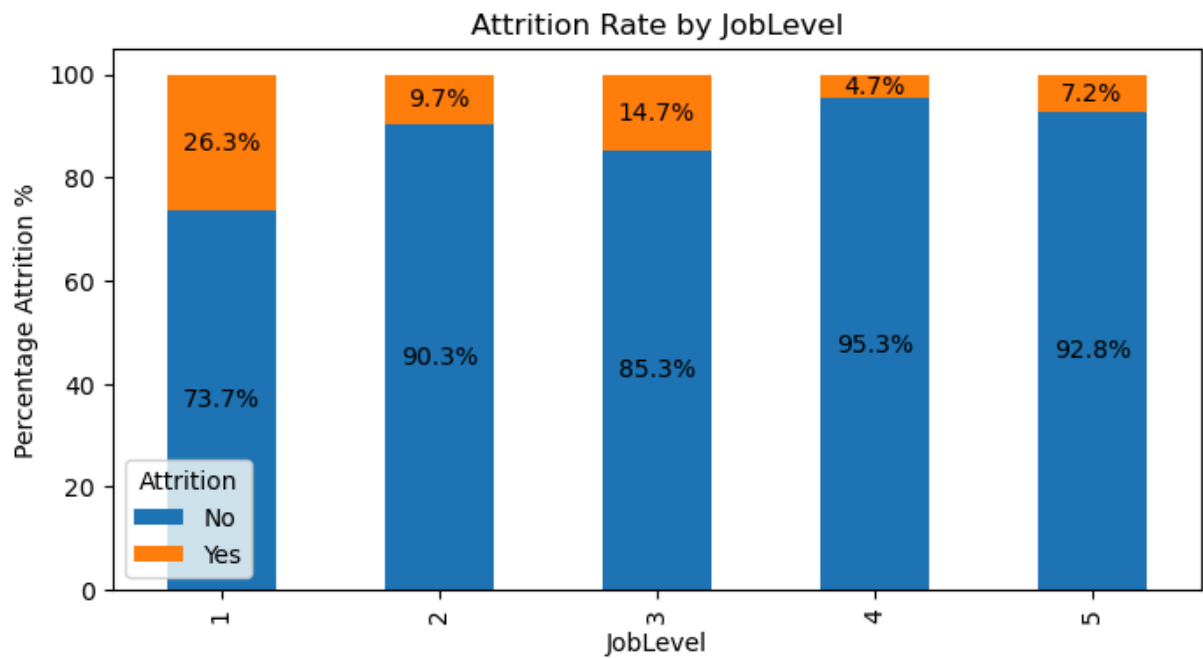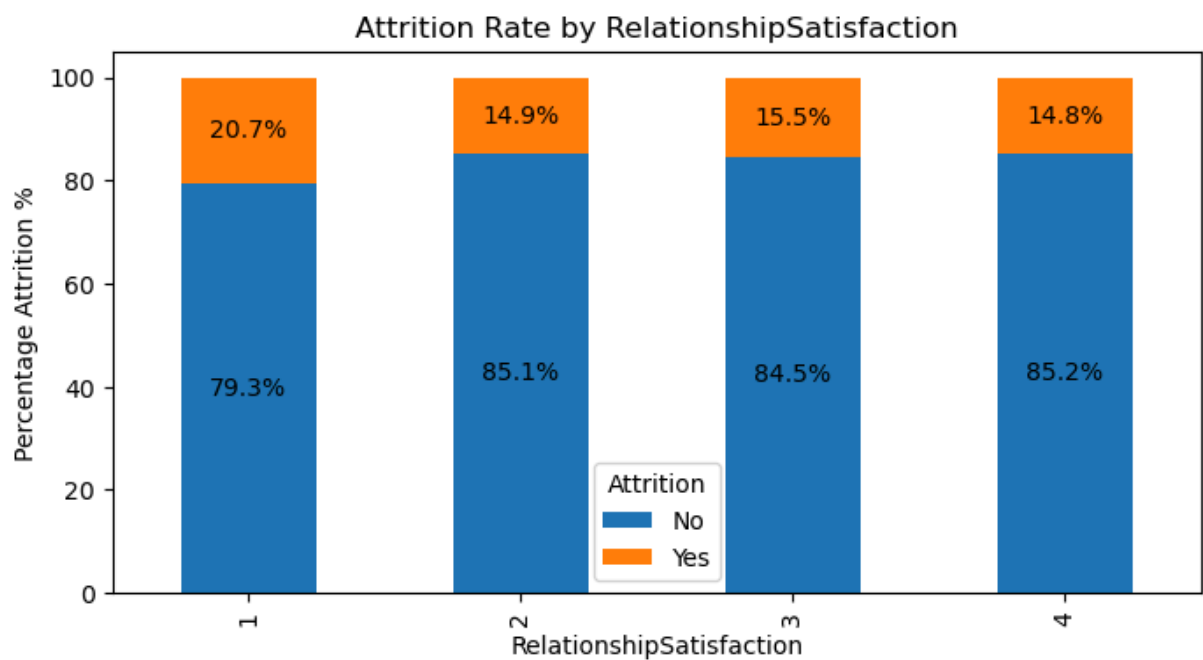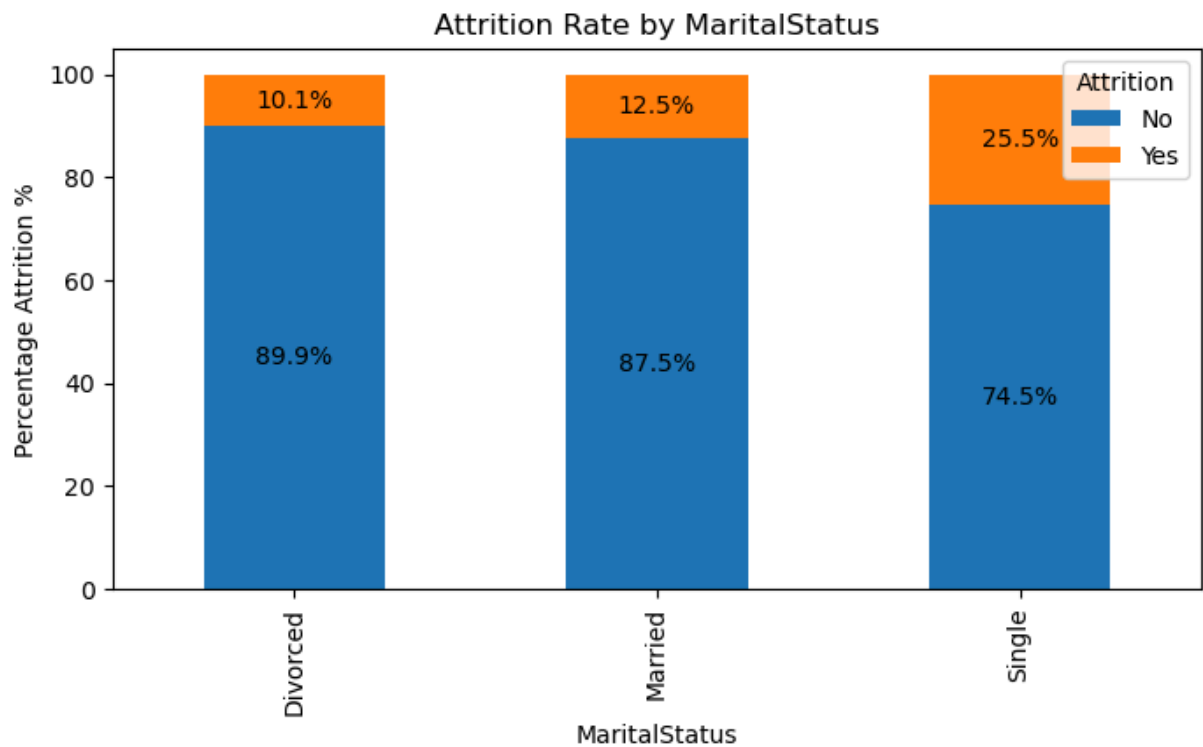
```
Attrition                          No          Yes
WorkLifeBalance
1                           68.750000   31.250000
2                           83.139535   16.860465
3                           85.778275   14.221725
4                           82.352941   17.647059


Attrition rates for StockOptionLevel:
Attrition                          No          Yes
StockOptionLevel
0                           75.594295   24.405705
1                           90.604027    9.395973
2                           92.405063    7.594937
3                           82.352941   17.647059


Attrition rates for Gender:
Attrition           No          Yes
Gender
Female       85.204082   14.795918
Male         82.993197   17.006803


Attrition rates for PerformanceRating:
Attrition                      No          Yes
PerformanceRating
3                       83.922830   16.077170
4                       83.628319   16.371681


Attrition rates for JobInvolvement:
Attrition               No          Yes
JobInvolvement
1                66.265060   33.734940
2                81.066667   18.933333
3                85.599078   14.400922
4                90.972222    9.027778


Attrition rates for JobLevel:
Attrition           No          Yes
JobLevel
1            73.664825   26.335175
2            90.262172    9.737828
3            85.321101   14.678899
4            95.283019    4.716981
5            92.753623    7.246377


Attrition rates for JobRole:
Attrition                              No          Yes
JobRole
Healthcare Representative       93.129771    6.870229
Human Resources                 76.923077   23.076923
Laboratory Technician           76.061776   23.938224
Manager                         95.098039    4.901961
Manufacturing Director          93.103448    6.896552
Research Director               97.500000    2.500000
Research Scientist              83.904110   16.095890
Sales Executive                 82.515337   17.484663
Sales Representative            60.240964   39.759036
```

```
Attrition rates for MaritalStatus:
Attrition               No        Yes
MaritalStatus
Divorced          89.908257  10.091743
Married           87.518574  12.481426
Single            74.468085  25.531915

Attrition rates for RelationshipSatisfaction:
Attrition                      No        Yes
RelationshipSatisfaction
1                         79.347826  20.652174
2                         85.148515  14.851485
3                         84.531590  15.468410
4                         85.185185  14.814815
```

Observations:

- **Employees working overtime have more than a 30% chance of attrition**,

which is very high compared to the 10% chance of attrition for employees who do not work extra hours.

- As seen earlier, the majority of employees work for the R&D department. The chance of attrition there is ~15%
- **Employees working as sales representatives have an attrition rate of around 40%** while HRs and Technicians have an attrition rate of around 25%. The sales and HR departments have higher attrition rates in comparison to an academic department like Research & Development, an observation that makes intuitive sense keeping in mind the differences in those job profiles. The high-pressure and incentive-based nature of Sales and Marketing roles may be contributing to their higher attrition rates.
- **The lower the employee's job involvement, the higher their attrition chances appear to be, with 1-rated JobInvolvement employees attriting at 35%.** The reason for this could be that employees with lower job involvement might feel left out or less valued and have already started to explore new options, leading to a higher attrition rate.
- **Employees at a lower job level also attrite more,** with 1-rated JobLevel employees showing a nearly 25% chance of attrition. These may be young employees who tend to explore more options in the initial stages of their careers.
- **A low work-life balance rating clearly leads employees to attrite**; 30% of those in the 1-rated category show attrition.

Problem 2 - Categorical Variable Observations:

2.a. Employees working overtime have more than a **30.5%** chance of attrition, which is very high compared to the **10.4%** chance of attrition for employees who do not work extra hours.

- As seen earlier, the majority of employees work for the R&D department. The chance of attrition there is ~13.8%

2.b. Employees working as sales representatives have an attrition rate of around **39.8%** while HRs and Technicians have an attrition rate of around **23%**.

- **Implications: Sales representatives show the highest attrition rate, possibly due to high-pressure sales targets, commission-based salaries, or lack of career stability. HR and Laboratory Technicians also experience higher attrition, which may be due to limited career advancement opportunities or job dissatisfaction.**

- The lower the employee's job involvement, the higher their attrition chances appear to be, with 1-rated JobInvolvement employees attriting at **33.7%**. The reason for this could be that employees with lower job involvement might feel left out or less valued and have already started to explore new options, leading to a higher attrition rate.

2.c. Employees at a lower job level also attrite more, with 1-rated JobLevel employees showing a nearly 25% chance of attrition.

- **Implications: Lower-level employees may leave due to low salary, limited career growth, and lack of job security. Providing career progression opportunities and fair compensation could help reduce attrition**

2.d. A low work-life balance rating clearly leads employees to attrite; **31.3%** of those in the 1-rated category show attrition.

```
In [13]:  # Mean of numerical varibles grouped by attrition
          employees.groupby(['Attrition'])[num_cols].mean()
```

Out[13]:

Attrition	DailyRate	Age	DistanceFromHome	MonthlyIncome	MonthlyRate
No	812.504461	37.561233	8.915653	6832.739659	14265.779400
Yes	750.362869	33.607595	10.632911	4787.092827	14559.308017

Observations:

- **Employees leaving the company have a nearly 30% lower average income and 30% lesser work experience than those who are not.** These could be the employees looking to explore new options and/or increase their salary with a company switch.
- **Employees showing attrition also tend to live 16% further from the office than those who are not**. The longer commute to and from work could mean they have to

spend more time/money every day, amd this could be leading to job dissatisfaction and wanting to leave the organization.

Problem 3 - Quantitative Data Observations:

3.a. Employees leaving the company have an over 30% lower average income and 30% less work experience (TotalWorkingYears) than those who are not.

Average Monthly Income of Employees Who Left:
$4,787 Average Monthly Income of Employees Who Stayed$:6,832
Average Total Working Years of Employees Who Left: 7.8 years
Average Total Working Years of Employees Who Stayed: 11.3 years

- **Implications:** Lower-income employees are at a higher risk of leaving. The company should ensure competitive salaries aligned with industry standards. Employees with fewer years of experience might feel less engaged or lack career growth opportunities, leading to higher attrition.

3.b. Employees showing attrition also tend to live 19% further from the office than those who are not.

Average Distance from Home of Employees Who Left: 10.6 km
Average Distance from Home of Employees Who Stayed: 8.9 km

- **Implications:** Employees who live farther from work might experience longer commutes, higher transportation costs, or work–life balance issues, leading to higher attrition. The company could consider remote work options or flexible work arrangements to retain employees who live farther away

Correlation Analysis

```
In [14]:  # Correation between numerical variables
          plt.figure(figsize=(15,8))
          sns.heatmap(employees[num_cols].corr(),annot=True, fmt='0.2f', cmap='YlGnBu'
```

Out[14]: `<Axes: >`

	DailyRate	Age	DistanceFromHome	MonthlyIncome	MonthlyRate	PercentSalaryHike	TotalWorkingYears	YearsAtCompany	NumCompaniesWorked	HourlyRate	YearsInCurrentRole	YearsSinceLastPromotion	YearsWithCurrManager	TrainingTimesLastYear
DailyRate	1.00	0.01	-0.00	0.01	-0.03	0.02	0.01	-0.03	0.04	0.02	0.01	-0.03	-0.03	0.00
Age	0.01	1.00	-0.00	0.50	0.03	0.00	0.68	0.31	0.30	0.02	0.21	0.22	0.20	-0.02
DistanceFromHome	-0.00	-0.00	1.00	-0.02	0.03	0.04	0.00	0.01	-0.03	0.03	0.02	0.01	0.01	-0.04
MonthlyIncome	0.01	0.50	-0.02	1.00	0.03	-0.03	0.77	0.51	0.15	-0.02	0.36	0.34	0.34	-0.02
MonthlyRate	-0.03	0.03	0.03	0.03	1.00	-0.01	0.03	-0.02	0.02	-0.02	-0.01	0.00	-0.04	0.00
PercentSalaryHike	0.02	0.00	0.04	-0.03	-0.01	1.00	-0.02	-0.04	-0.01	-0.01	-0.00	-0.02	-0.01	-0.01
TotalWorkingYears	0.01	0.68	0.00	0.77	0.03	-0.02	1.00	0.63	0.24	-0.00	0.46	0.40	0.46	-0.04
YearsAtCompany	-0.03	0.31	0.01	0.51	-0.02	-0.04	0.63	1.00	-0.12	-0.02	0.76	0.62	0.77	0.00
NumCompaniesWorked	0.04	0.30	-0.03	0.15	0.02	-0.01	0.24	-0.12	1.00	0.02	-0.09	-0.04	-0.11	-0.07
HourlyRate	0.02	0.02	0.03	-0.02	-0.02	-0.01	-0.00	-0.02	0.02	1.00	-0.02	-0.03	-0.02	-0.01
YearsInCurrentRole	0.01	0.21	0.02	0.36	-0.01	-0.00	0.46	0.76	-0.09	-0.02	1.00	0.55	0.71	-0.01
YearsSinceLastPromotion	-0.03	0.22	0.01	0.34	0.00	-0.02	0.40	0.62	-0.04	-0.03	0.55	1.00	0.51	-0.00
YearsWithCurrManager	-0.03	0.20	0.01	0.34	-0.04	-0.01	0.46	0.77	-0.11	-0.02	0.71	0.51	1.00	-0.00
TrainingTimesLastYear	0.00	-0.02	-0.04	-0.02	0.00	-0.01	-0.04	0.00	-0.07	-0.01	-0.01	-0.00	-0.00	1.00

In [15]:
```python
# Compute correlation matrix
corr = employees[num_cols].corr()

# Display the correlation values sorted by strength
corr_unstacked = corr.unstack().reset_index()
corr_unstacked.columns = ['Variable 1', 'Variable 2', 'Correlation']
corr_unstacked = corr_unstacked[corr_unstacked['Variable 1'] != corr_unstack

# Sort correlations in descending order
corr_unstacked = corr_unstacked.sort_values(by="Correlation", ascending=Fals
corr_unstacked.head(15)
```

	Variable 1	Variable 2	Correlation
48	MonthlyIncome	TotalWorkingYears	0.772893
87	TotalWorkingYears	MonthlyIncome	0.772893
175	YearsWithCurrManager	YearsAtCompany	0.769212
110	YearsAtCompany	YearsWithCurrManager	0.769212
108	YearsAtCompany	YearsInCurrentRole	0.758754
147	YearsInCurrentRole	YearsAtCompany	0.758754
152	YearsInCurrentRole	YearsWithCurrManager	0.714365
178	YearsWithCurrManager	YearsInCurrentRole	0.714365
20	Age	TotalWorkingYears	0.680381
85	TotalWorkingYears	Age	0.680381
104	YearsAtCompany	TotalWorkingYears	0.628133
91	TotalWorkingYears	YearsAtCompany	0.628133
161	YearsSinceLastPromotion	YearsAtCompany	0.618409
109	YearsAtCompany	YearsSinceLastPromotion	0.618409
151	YearsInCurrentRole	YearsSinceLastPromotion	0.548056

Observations:

- **Total work experience, monthly income, years at company and years with current manager are highly correlated with each other and with employee age** which is easy to understand as these variables show an increase with age for most employees.
- Years at company and years in current role are correlated with years since last promotion which means that the company is not giving promotions at the right time.

Problem 4 - Correlation Analysis:

4.a. YearsWithCurrManager & YearsAtCompany (0.77), MonthlyIncome & TotalWorkingYears (0.77), YearsAtCompany & YearsInCurrentRole (0.76) and YearsInCurrentRole & YearsWithCurrManager (0.71) are highly correlated with each other.

4.b. YearsAtCompany and **YearsInCurrent Role** are correlated with **YearsSinceLastPromotion** which means that the company is not giving promotions at the right time.

Employee Attrition (Classification)

1. Prepare data for modeling
2. Partition the data into train and test set.
3. Build model on the train data.
4. Tune the model if required.
5. Test the data on test set.

Prepare data for modeling

- **Drop columns:** EmployeeNumber, Over18, StandardHours
- **Convert categorical columns to numerical:** Use one-hot encoding to convert categorical columns to numerical.
- **Split data into X and Y:** Split the data into independent and dependent variables.
- **Split data into train and test set:** Split the data into train and test set.

```
In [16]:  # Creating list of dummy columns
          to_get_dummies_for = ['BusinessTravel', 'Department','Education', 'Education

          # Creating dummy variables
          employees = pd.get_dummies(data = employees, columns= to_get_dummies_for, dr

          # Mapping overtime and attrition
          dict_OverTime = {'Yes': 1, 'No':0}
          dict_attrition = {'Yes': 1, 'No': 0}

          employees['OverTime'] = employees.OverTime.map(dict_OverTime)
          employees['Attrition'] = employees.Attrition.map(dict_attrition)
```

```
In [17]:  # Separating the independent variables (X) and the dependent variable (Y)
          Y= employees.Attrition
          X= employees.drop(columns = ['Attrition'])
```

```
In [18]:  # Splitting the data into 70% train and 30% test set
          x_train,x_test,y_train,y_test=train_test_split(X,Y,test_size=0.3,random_stat
```

Model evaluation criterion

The model can make two types of wrong predictions:

1. Predicting an employee will attrite when the employee doesn't attrite
2. Predicting an employee will not attrite and the employee actually attrites

Which case is more important?

- **Predicting that the employee will not attrite but the employee attrites** i.e. losing out on a valuable employee or asset. This would be considered a major miss for any employee attrition predictor, and is hence the more important case of wrong predictions.

How to reduce this loss i.e the need to reduce False Negatives?

- **The company would want the Recall to be maximized**, the greater the Recall, the higher the chances of minimizing false negatives. Hence, the focus should be on increasing Recall (minimizing the false negatives) or in other words identifying the true positives (i.e. Class 1) very well, so that the company can provide incentives to control attrition rate especially for top-performers. This would help in optimizing the overall project cost towards retaining the best talent.

In [19]:
```python
# Create a function to calculate and print the classification report and cor

def metrics_score(actual, predicted):
    print(classification_report(actual, predicted))
    cm = confusion_matrix(actual, predicted)
    plt.figure(figsize=(8,5))
    sns.heatmap(cm, annot=True,  fmt='.2f', xticklabels=['Not Attrite', 'Att
    plt.ylabel('Actual')
    plt.xlabel('Predicted')
    plt.show()
```

Building the model

We will be building a single model: **Classification Tree**

- Tree induction is a supervised segmentation algorithm which is used for **binary classification problems** i.e. where the dependent variable is categorical and has only two possible values. The data is recursively partitioned to achieve maximum homogeneity of partitions (minimal entropy) with respect to the target variable.

In [20]:
```python
# Instantiate a DecisionTreeClassifier 'dt' with a maximum depth of 8
dt = DecisionTreeClassifier(max_depth=8, random_state=42)
```

In [21]:
```python
# Fit dt to the training set
dt.fit(x_train, y_train)
```

Out[21]:
```
         ▼          DecisionTreeClassifier          ⓘ ⍰

DecisionTreeClassifier(max_depth=8, random_state=42)
```

In [22]:
```python
y_pred = dt.predict(x_test)
y_pred[:20]
```

Out[22]: array([0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0])

In [23]:
```python
# Checking model performace of dt on training set
y_pred_train_dt = dt.predict(x_train)
metrics_score(y_train, y_pred_train_dt)

# Checking model performace of dt on test set
```
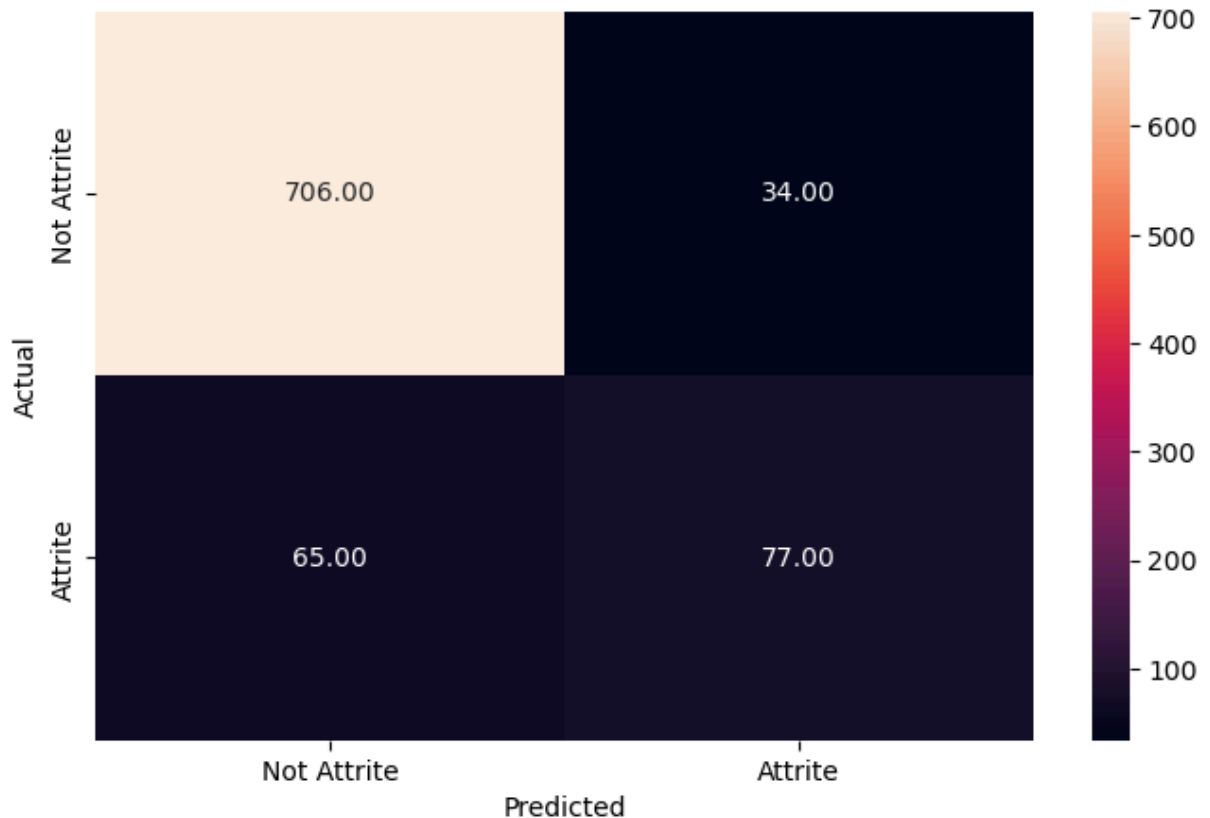
```
y_pred_test_dt = dt.predict(x_test)
metrics_score(y_test, y_pred_test_dt)
```

	precision	recall	f1-score	support
0	0.95	1.00	0.97	1726
1	0.97	0.71	0.82	332
accuracy			0.95	2058
macro avg	0.96	0.85	0.90	2058
weighted avg	0.95	0.95	0.95	2058
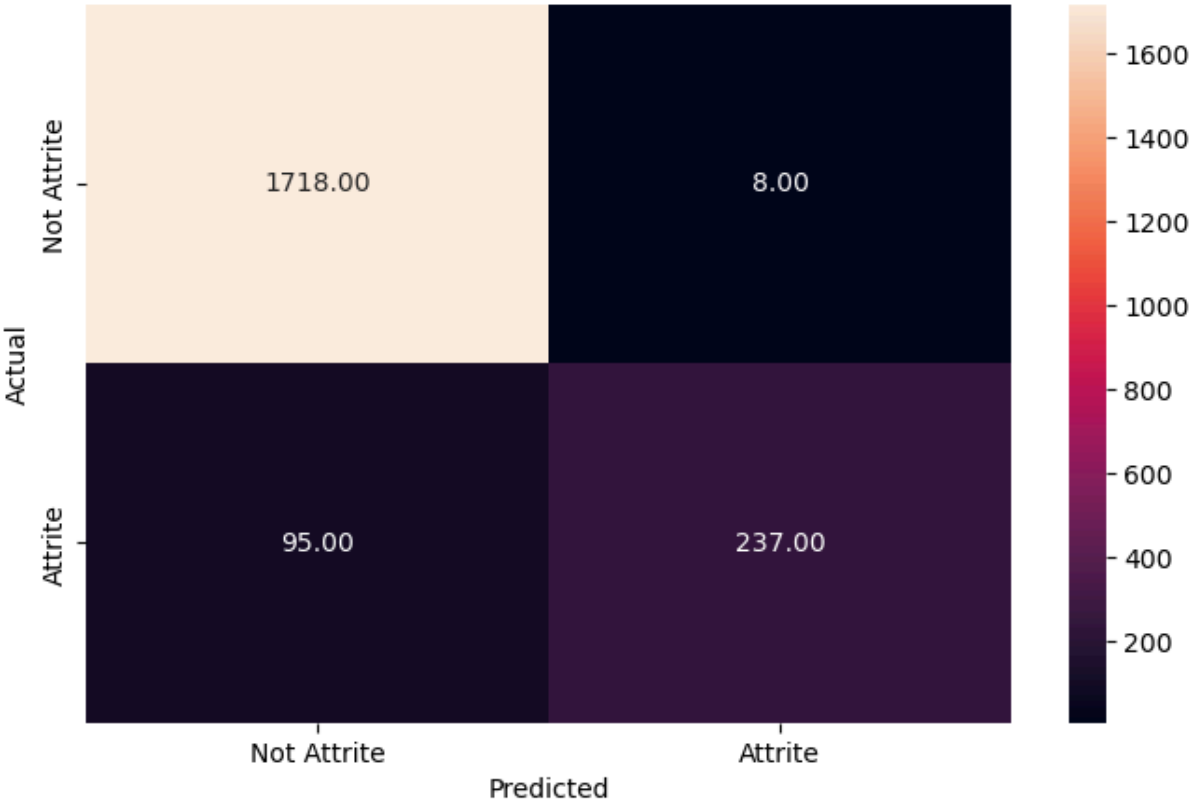
	precision	recall	f1-score	support
0	0.92	0.95	0.93	740
1	0.69	0.54	0.61	142
accuracy			0.89	882
macro avg	0.80	0.75	0.77	882
weighted avg	0.88	0.89	0.88	882

- The reported average includes the macro average which averages the unweighted mean per label, and the weighted average i.e. averaging the support-weighted mean per label.
- In classification, the class of interest is considered the positive class. Here, the class of interest is 1 i.e. identifying the employees at risk of attrition.

Reading the confusion matrix (clockwise):

- True Negative (Actual=0, Predicted=0): Model predicts that an employee would not attrite and the employee does not attrite

- False Positive (Actual=0, Predicted=1): Model predicts that an employee would attrite but the employee does not attrite

- False Negative (Actual=1, Predicted=0): Model predicts that an employee would not attrite but the employee attrites

- True Positive (Actual=1, Predicted=1): Model predicts that an employee would attrite and the employee actually attrites

Observations:

- The model is performing well in terms of accuracy.

- The recall for class 1 is quite low (`54%`), which implies that this model will not perform well in differentiating out the employees who have a high chance of leaving the company, and hence this model would not help reduce the attrition rate.
- The model also has a somewhat low precision (`69%`). A precision of ~0.70 suggests that there is a 30% (1 - 0.70) chance that the model will predict that a person is going to leave even though he/she would not, and the company would waste their time and energy on these employees who are not at risk of attrition.

Feature Importance

```python
In [24]:   importances = pd.DataFrame(data={
               'Attribute': x_train.columns,
               'Importance': dt.feature_importances_
           })

           importances = importances.sort_values(by='Importance', ascending=False)
           importances
```

	Attribute	Importance
5	MonthlyIncome	0.170471
8	OverTime	0.100399
3	HourlyRate	0.069340
13	TotalWorkingYears	0.061483
1	DailyRate	0.060270
15	WorkLifeBalance	0.055674
17	YearsInCurrentRole	0.052963
0	Age	0.043628
12	StockOptionLevel	0.039736
18	YearsSinceLastPromotion	0.039416
6	MonthlyRate	0.034077
14	TrainingTimesLastYear	0.031185
49	JobRole_Research Scientist	0.025486
4	JobSatisfaction	0.025416
9	PercentSalaryHike	0.024467
16	YearsAtCompany	0.021054
19	YearsWithCurrManager	0.019010
50	JobRole_Sales Executive	0.018728
29	EducationField_Marketing	0.014835
44	JobRole_Human Resources	0.012965
7	NumCompaniesWorked	0.012059
2	DistanceFromHome	0.011970
51	JobRole_Sales Representative	0.010790
24	Education_2	0.010681
34	EnvironmentSatisfaction_3	0.008856
20	BusinessTravel_Travel_Frequently	0.006280
40	JobLevel_2	0.004955
31	EducationField_Other	0.004844
27	Education_5	0.004422
11	RelationshipSatisfaction	0.003925
32	EducationField_Technical Degree	0.000615
28	EducationField_Life Sciences	0.000000

	Attribute	Importance
42	JobLevel_4	0.000000
52	MaritalStatus_Married	0.000000
10	PerformanceRating	0.000000
21	BusinessTravel_Travel_Rarely	0.000000
48	JobRole_Research Director	0.000000
47	JobRole_Manufacturing Director	0.000000
46	JobRole_Manager	0.000000
45	JobRole_Laboratory Technician	0.000000
22	Department_Research & Development	0.000000
43	JobLevel_5	0.000000
41	JobLevel_3	0.000000
30	EducationField_Medical	0.000000
23	Department_Sales	0.000000
39	JobInvolvement_4	0.000000
38	JobInvolvement_3	0.000000
37	JobInvolvement_2	0.000000
36	Gender_Male	0.000000
35	EnvironmentSatisfaction_4	0.000000
25	Education_3	0.000000
33	EnvironmentSatisfaction_2	0.000000
26	Education_4	0.000000
53	MaritalStatus_Single	0.000000

In [25]:
```python
# Define figure size in (width, height) for a single plot
plt.figure(figsize=(15,5))

plt.bar(x=importances['Attribute'], height=importances['Importance'], color=
plt.title('Feature importances obtained from coefficients', size=10)
plt.xticks(rotation='vertical')
plt.show()
```

Feature importances obtained from coefficients
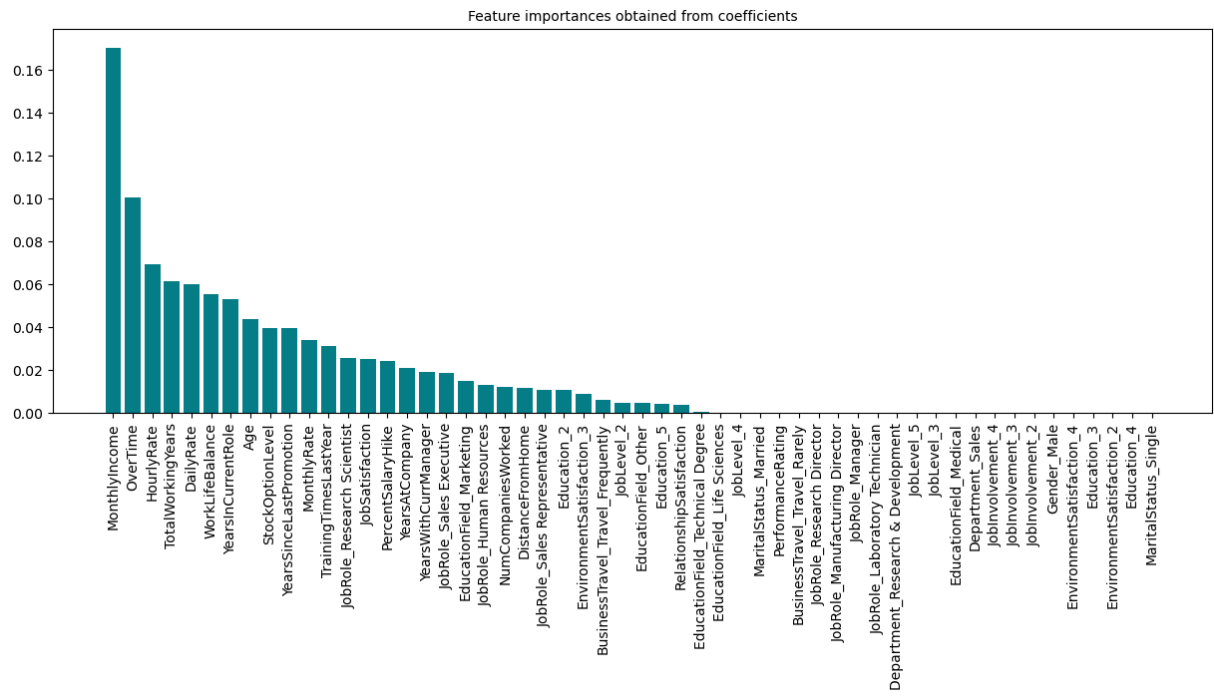
Conclusions:

- The classification tree was **able to identify some key factors involved with high attrition** in the organization. In a real-world setting, additional models would be attempted (Logistic Regression, Linear Discriminant, Support Vector Machines, etc.). Models would be systematically compared in order to select the one with the best performance. These techniques will be developed later in the course.

Problem 5 - Recommendations:

- **Increase Compensation for High-Risk Employees** A higher monthly income might lower the chance of an employee attriting. The company should make sure that all of its employees are compensated at least based on industry standards.

- **Address Work-Life Balance Issues**: 31.3% of employees with the lowest work-life balance rating attrite. Introduce flexible work arrangements (remote work, hybrid model) and workload management strategies to reduce stress and burnout.

- **Improve Career Growth & Promotion Opportunities**: Employees who haven't been promoted in a long time are at a higher risk of attrition. Implement a structured promotion plan, increase opportunities for internal mobility, and introduce mentorship programs to encourage professional development

- **Reduce Overtime Dependency**: Employees who work overtime have a 30.5% chance of attrition, compared to 10.4% for those who don't. Enforce reasonable overtime limits, monitor burnout indicators, and ensure employees are compensated adequately for extra work.

- **Targeted Retention Incentives for At-Risk Employees**: Sales Representatives (39.8% attrition rate) and HRs & Technicians (~23%) need targeted retention strategies. Offer equity options, bonuses, and career progression incentives to high-risk job roles to improve retention.

- **Minimize Commute Challenges**: Employees who live farther from work (10.6 km vs. 8.9 km) have a higher attrition rate. Consider relocation assistance, remote work options, or flexible start times to ease commute burdens.