ATTRITION :

The rate at which employees leave an organization over a particular period of time, when the company does not seek to fill the position.

EMPLOYEE ATTRITION RATE :

Employee Attrition Rate is calculated as the percentage of employees who left the company in a given period to the total average number of employees within that period.

AIM OF PROJECT:

To help a company understand the attrition in their organization to minimize the attrition rate, so that turnover costs are reduced and employees can be retained for a longer period.To explore the dataset and it's different features Understand the relationship between the different varaibles Find the variables that tend to have a larger number of employees with positive attrition

Overview of project:

Exploratory data analysis of an attrition survey conducted. Built a model to calculate the attrition rate in an organisation.

Exploratory Data Analysis :

Exploratory Data Analysis is an initial process of analysis, in which we can summarize characteristics of data such as pattern, trends, outliers, and hypothesis testing using descriptive statistics and visualization.

Import Dataset and Libraries

In [1]:

```
import matplotlib.pyplot as plt
import numpy as np
import warnings
warnings.filterwarnings("ignore")
import pandas as pd
import seaborn as sns
```

In [2]:

```
attr = pd.read_csv("attrition data.csv")
```

In [3]:
```
attr
```

Out[3]:

	Age	Attrition	Department	DistanceFromHome	Education	EducationField	EnvironmentS
0	41	Yes	Sales	1	2	Life Sciences	
1	49	No	Research & Development	8	1	Life Sciences	
2	37	Yes	Research & Development	2	2	Other	
3	33	No	Research & Development	3	4	Life Sciences	
4	27	No	Research & Development	2	1	Medical	
...	...	...	...	...	...	...	
1465	36	No	Research & Development	23	2	Medical	
1466	39	No	Research & Development	6	1	Medical	
1467	27	No	Research & Development	4	3	Life Sciences	
1468	49	No	Sales	2	3	Medical	
1469	34	No	Research & Development	8	3	Medical	

1470 rows × 13 columns

EDA on the Datasets

In [4]:
```
attr.shape
```

Out[4]:
```
(1470, 13)
```

In [5]:
```
attr.isnull()
```

Out[5]:

	Age	Attrition	Department	DistanceFromHome	Education	EducationField	EnvironmentS
0	False	False	False	False	False	False	
1	False	False	False	False	False	False	
2	False	False	False	False	False	False	
3	False	False	False	False	False	False	
4	False	False	False	False	False	False	
...	...	...	...	...	...	...	
1465	False	False	False	False	False	False	
1466	False	False	False	False	False	False	
1467	False	False	False	False	False	False	
1468	False	False	False	False	False	False	
1469	False	False	False	False	False	False	

1470 rows × 13 columns

In [6]:
```
attr.isnull().sum()
```

Out[6]:
```
Age                        0
Attrition                  0
Department                 0
DistanceFromHome           0
Education                  0
EducationField             0
EnvironmentSatisfaction    0
JobSatisfaction            0
MaritalStatus              0
MonthlyIncome              0
NumCompaniesWorked         0
WorkLifeBalance            0
YearsAtCompany             0
dtype: int64
```

There are no null entries in existing employee dataset.

In [7]:

```
attr.describe()
```

Out[7]:

	Age	DistanceFromHome	Education	EnvironmentSatisfaction	JobSatisfaction
count	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000
mean	36.923810	9.192517	2.912925	2.721769	2.728571
std	9.135373	8.106864	1.024165	1.093082	1.102846
min	18.000000	1.000000	1.000000	1.000000	1.000000
25%	30.000000	2.000000	2.000000	2.000000	2.000000
50%	36.000000	7.000000	3.000000	3.000000	3.000000
75%	43.000000	14.000000	4.000000	4.000000	4.000000
max	60.000000	29.000000	5.000000	4.000000	4.000000

In [8]:

```
num = attr.select_dtypes(include=np.number).columns
len(num)
```

Out[8]:

```
9
```

In [9]:

```
for i in num:
    print("The no.of unique values in",i,"are: ",attr.loc[:,i].nunique())
```

```
The no.of unique values in Age are:  43
The no.of unique values in DistanceFromHome are:  29
The no.of unique values in Education are:  5
The no.of unique values in EnvironmentSatisfaction are:  4
The no.of unique values in JobSatisfaction are:  4
The no.of unique values in MonthlyIncome are:  1349
The no.of unique values in NumCompaniesWorked are:  10
The no.of unique values in WorkLifeBalance are:  4
The no.of unique values in YearsAtCompany are:  37
```

Column Description

Independent Variables were:

```
Age: Age of employees,
Department: Department of work,
Distance from home,
Education: 1-Below College; 2-College; 3-Bachelor; 4-Master; 5-Docto
r;
Education Field
Environment Satisfaction: 1-Low; 2-Medium; 3-High; 4-Very High;
Job Satisfaction: 1-Low; 2-Medium; 3-High; 4-Very High;
```

Top 5 values

In [10]:
```
attr.head()
```

Out[10]:

	Age	Attrition	Department	DistanceFromHome	Education	EducationField	EnvironmentSatis
0	41	Yes	Sales	1	2	Life Sciences	
1	49	No	Research & Development	8	1	Life Sciences	
2	37	Yes	Research & Development	2	2	Other	
3	33	No	Research & Development	3	4	Life Sciences	
4	27	No	Research & Development	2	1	Medical	

last 5 values

In [11]:
```
attr.tail()
```

Out[11]:

	Age	Attrition	Department	DistanceFromHome	Education	EducationField	EnvironmentS
1465	36	No	Research & Development	23	2	Medical	
1466	39	No	Research & Development	6	1	Medical	
1467	27	No	Research & Development	4	3	Life Sciences	
1468	49	No	Sales	2	3	Medical	
1469	34	No	Research & Development	8	3	Medical	

If there is any duplicate values

In [12]:

```
attr.duplicated()
```

Out[12]:

```
0       False
1       False
2       False
3       False
4       False
        ...
1465    False
1466    False
1467    False
1468    False
1469    False
Length: 1470, dtype: bool
```

Value counts of duplicate values

In [13]:

```
attr.duplicated().value_counts()
```

Out[13]:

```
False    1470
dtype: int64
```

There are No Duplicate Columns.

Unique values of every columns

In [14]:

```
attr.nunique()
```

Out[14]:

```
Age                           43
Attrition                      2
Department                     3
DistanceFromHome              29
Education                      5
EducationField                 6
EnvironmentSatisfaction        4
JobSatisfaction                4
MaritalStatus                  3
MonthlyIncome               1349
NumCompaniesWorked            10
WorkLifeBalance                4
YearsAtCompany                37
dtype: int64
```

Data Insights

In the given dataset, we have two types of employee one who stayed and another who left the company.

Attrition

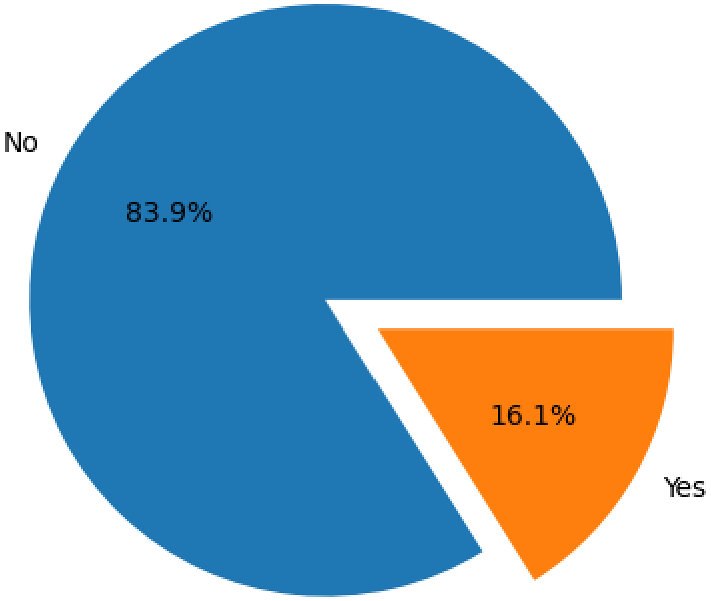
In [15]:
```
att_count=pd.DataFrame(attr["Attrition"].value_counts())
att_count
```

Out[15]:

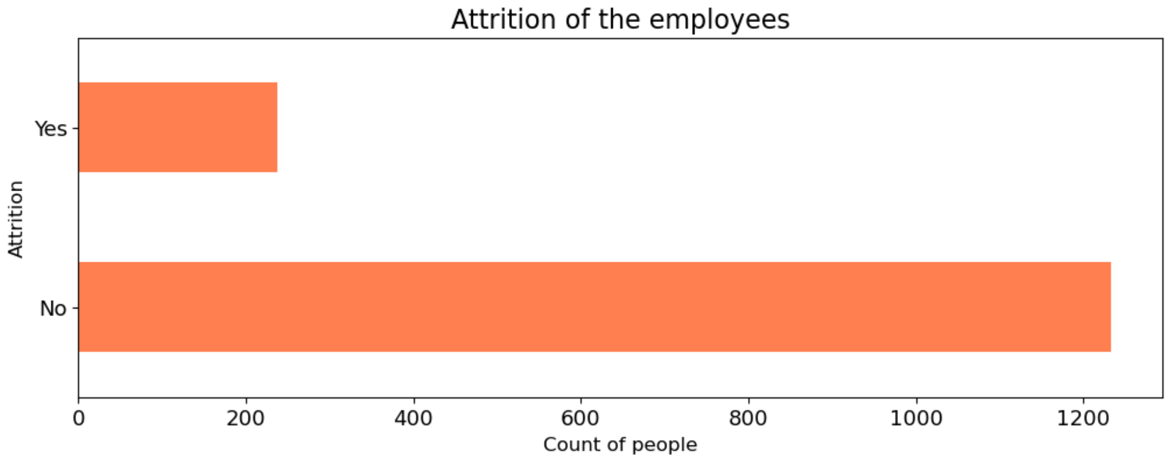
	Attrition
No	1233
Yes	237

Distribution of Attrition

In [16]:
```
plt.pie(att_count["Attrition"],labels=['No','Yes'],autopct='%.1f%%', explode=
plt.show()
```



In [17]:
```
attr['Attrition'].value_counts().plot(kind='barh', figsize=(12,4), color="cor

plt.xlabel("Count of people", fontsize=12)
plt.ylabel("Attrition", fontsize=12)
plt.title("Attrition of the employees", fontsize=16)
plt.show()
```



In [ ]:
```
```

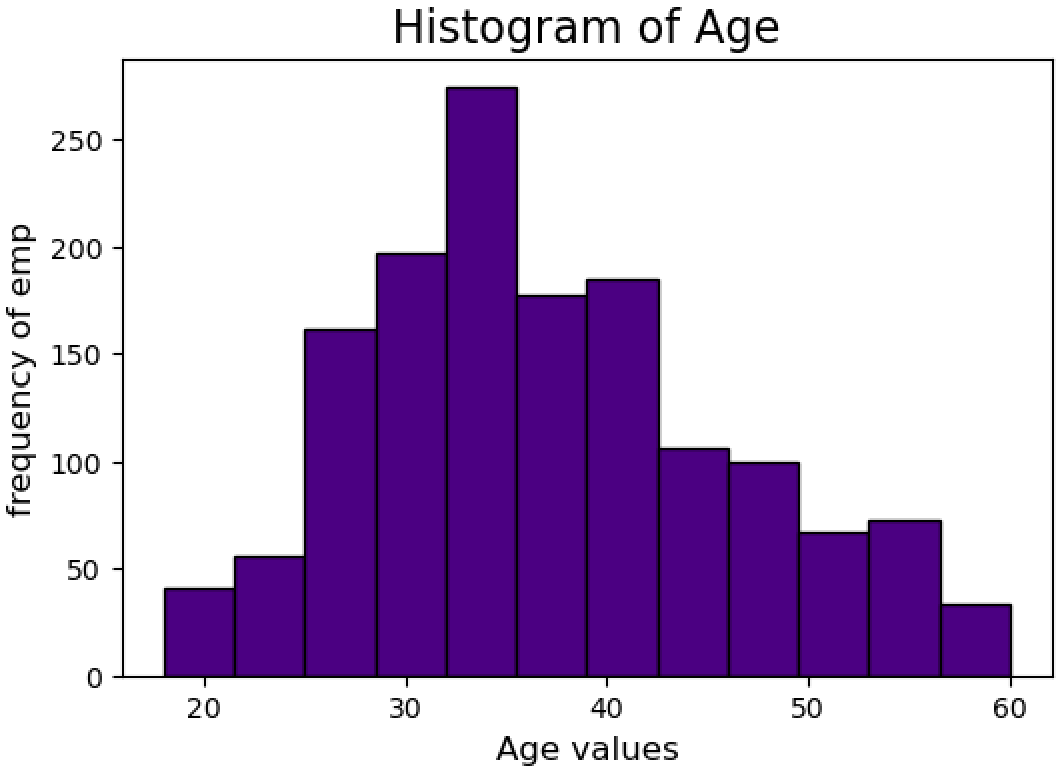
Age and Attrition

In [18]:
```
#Number of Unique values

attr["Age"].nunique()
```

Out[18]:
```
43
```

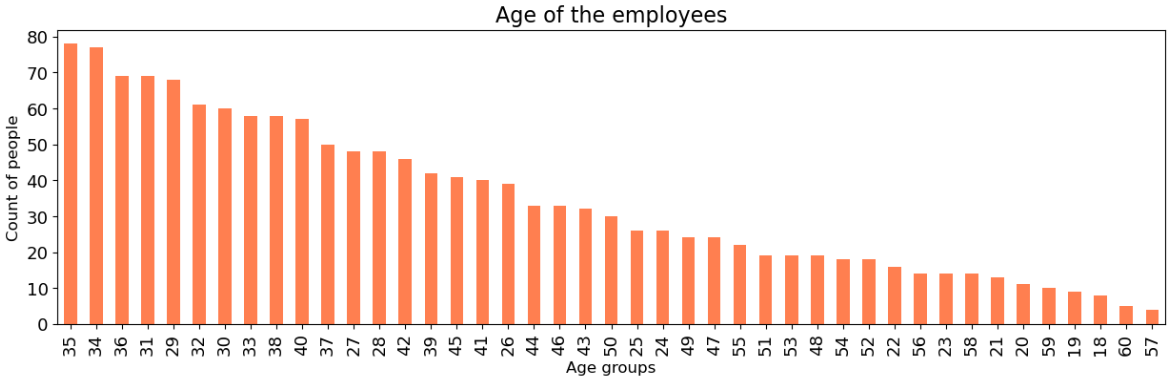
In [19]:

```
attr.hist(column="Age",
         grid=False,
         figsize=(6,4),
         color="indigo",
         edgecolor="black",
         bins=12)

plt.xlabel("Age values",fontsize=12)
plt.ylabel("frequency of emp",fontsize=12)
plt.title("Histogram of Age",fontsize=16)

plt.show()
```



Most of the employees, who have been a part of the company, tend to fall in the age range from 25 years to 45 years. (68–95–99 rule)

In [20]:
```
attr['Age'].value_counts().plot(kind='bar', figsize=(15,4), color="coral", fo

plt.xlabel("Age groups", fontsize=12)
plt.ylabel("Count of people", fontsize=12)
plt.title("Age of the employees", fontsize=16)
plt.show()
```



Maximum employees age lies in the range of 30 to 40

In [21]:

```
t = pd.crosstab(attr.Age,columns = attr.Attrition)
t
```

Out[21]:

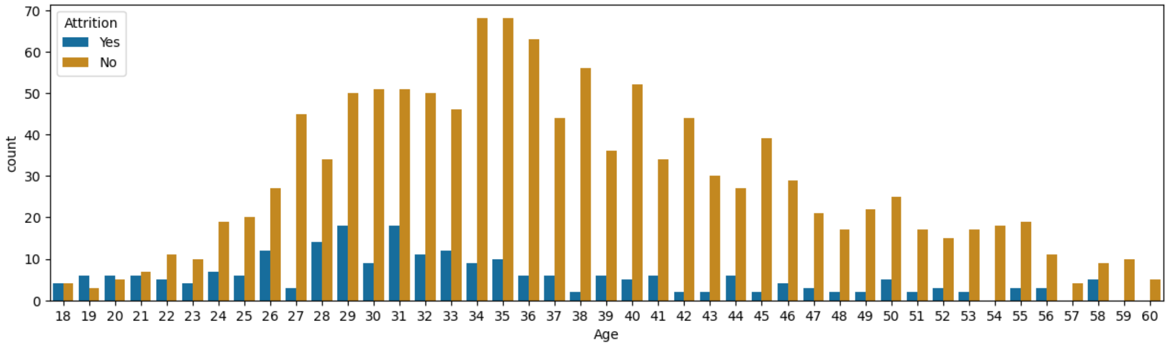
Attrition	No	Yes
Age		
18	4	4
19	3	6
20	5	6
21	7	6
22	11	5
23	10	4
24	19	7
25	20	6
26	27	12
27	45	3
28	34	14
29	50	18
30	51	9
31	51	18
32	50	11
33	46	12
34	68	9
35	68	10
36	63	6
37	44	6
38	56	2
39	36	6
40	52	5
41	34	6
42	44	2
43	30	2
44	27	6
45	39	2
46	29	4
47	21	3
48	17	2
49	22	2
50	25	5
51	17	2

Attrition	No	Yes
Age		
52	15	3
53	17	2
54	18	0
55	19	3
56	11	3
57	4	0
58	9	5
59	10	0
60	5	0

What age group is more likely to leave?

In [22]:

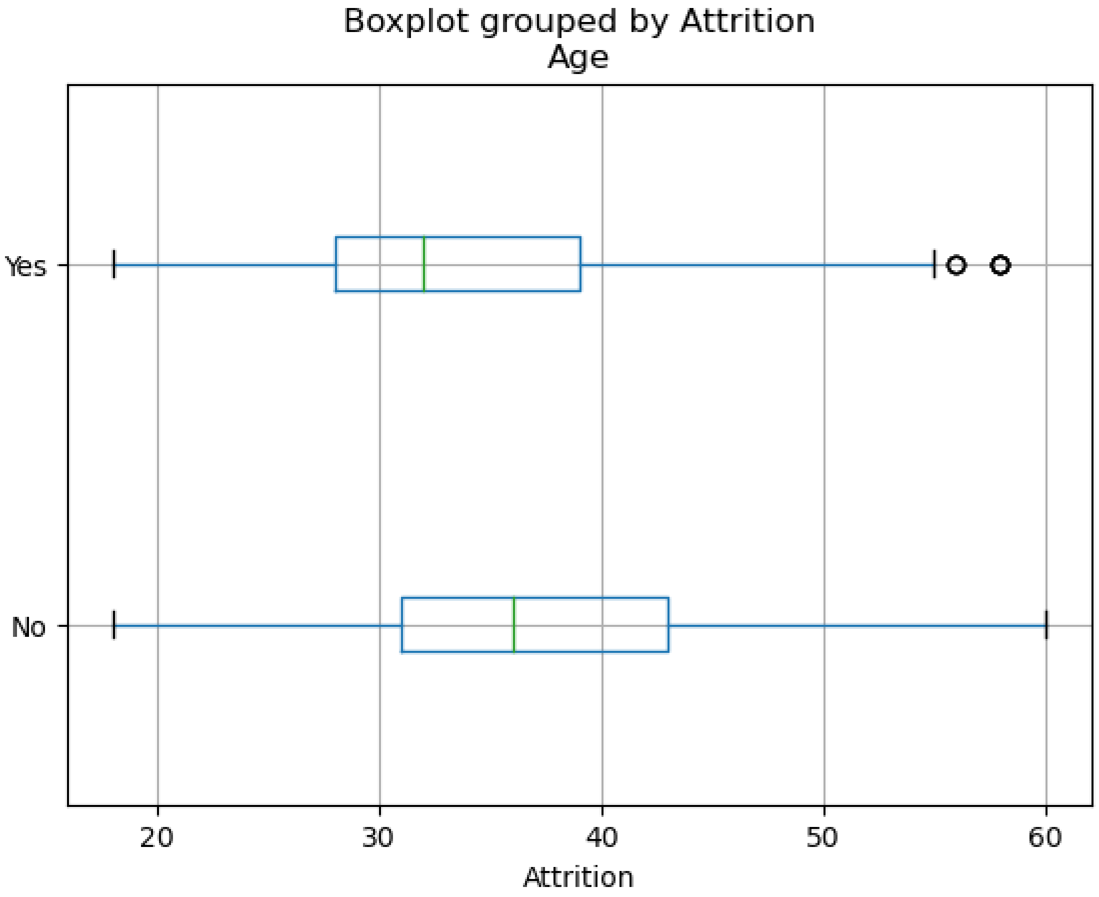
```
plt.subplots (figsize=(15,4))
sns.countplot (x='Age',  hue='Attrition', data= attr, palette = 'colorblind')
plt.show()
```



The proportion of employees who left was comparatively less among the senior level employees. Maybe these employees were at a higher authority and were less likely to leave that position.

Major Attrition occured among the mid-career level employees. Career change and career growth could be one of the possibilities for such an attrition levels.

In [23]:
```
attr.boxplot(column = 'Age', by = 'Attrition', vert= False)
plt.show()
```



we can express that Age does not play a powerful role in influencing attrition but In the above plot we can see that affirmative attrition count is higher in late 20 and early 30s.

MonthlyIncome

There are 1470 employees under our analysis for monthly income (in our sample)

• Ordinal data(numerical)

• The employees in the organization are earning from 1009 To 19,999.

• 25%(1st quartile) of the employees under our analysis in our sample are earning monthly salary less than or equal to 2911.

• 50%(median-2nd quartile) of the employees under our analysis in our sample are earning monthly salary less than or equal to 4919.

• 75%(3rd quartile) of the employees under our analysis in our sample are earning monthly salary less than or equal to 8379.

- The mean monthly income received by the employees are 6502.93.

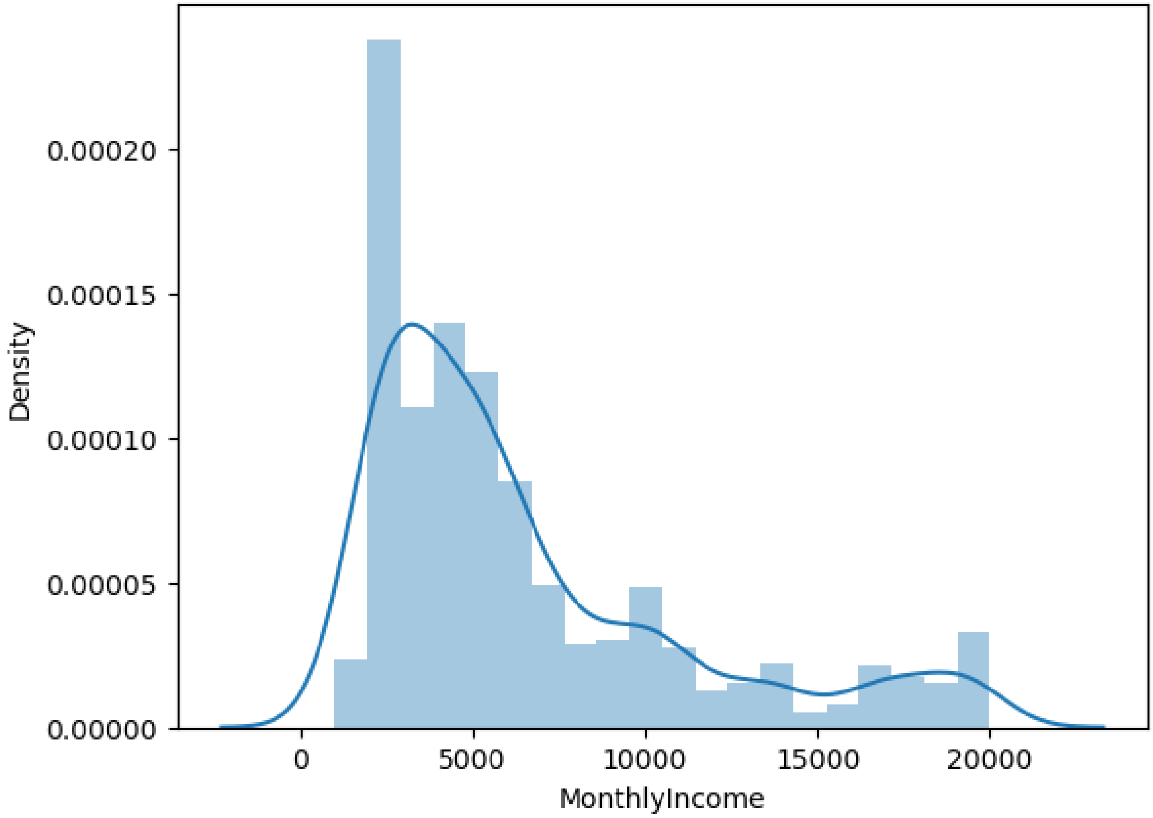
```
In [24]: attr.MonthlyIncome.describe()
```

```
Out[24]: count     1470.000000
         mean      6502.931293
         std       4707.956783
         min       1009.000000
         25%       2911.000000
         50%       4919.000000
         75%       8379.000000
         max      19999.000000
         Name: MonthlyIncome, dtype: float64
```

```
In [25]: attr["MonthlyIncome"].nunique()
```

```
Out[25]: 1349
```

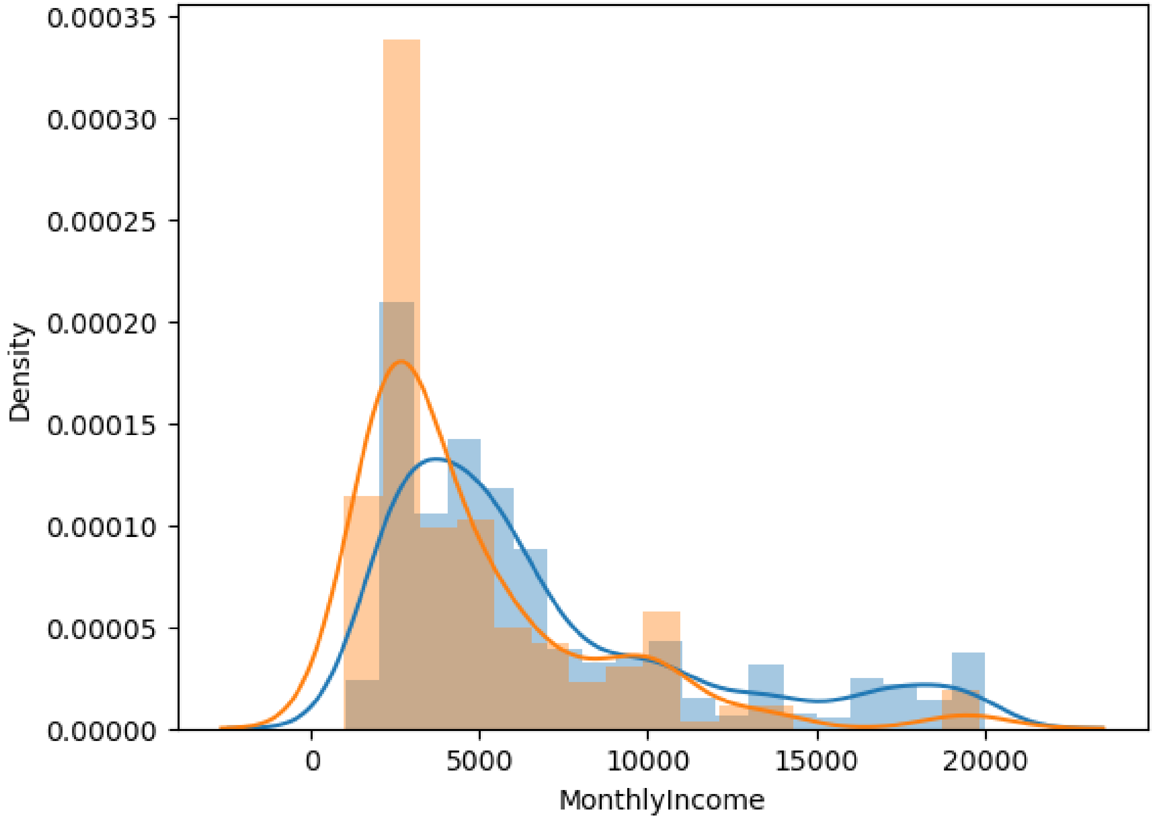
```
In [26]: sns.distplot(attr['MonthlyIncome'])
         plt.show()
```



This variable is left skewed distribution,meaning most of its values are located at the lower end(2000-6000)

In [27]:
```
p_attr = attr.loc[attr['Attrition'] == "Yes"]
n_attr = attr.loc[attr['Attrition'] == "No"]
```

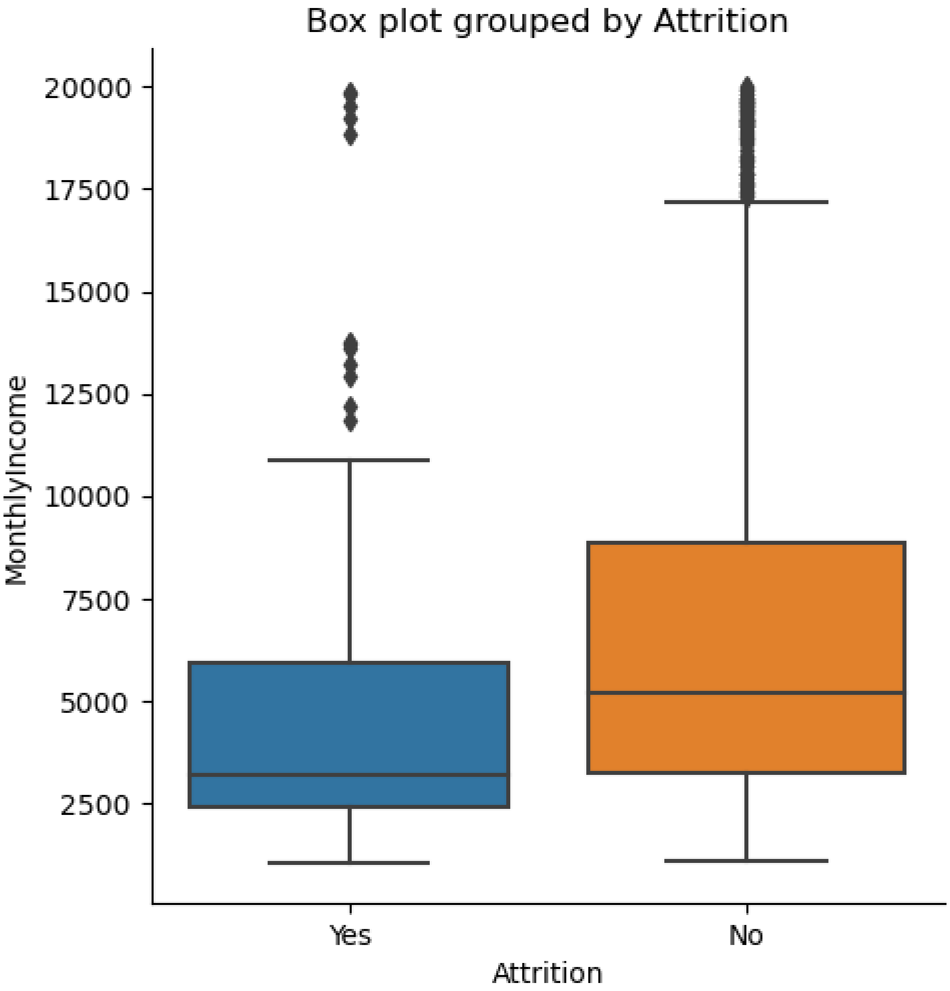
In [28]:
```
sns.distplot(n_attr['MonthlyIncome'], label='Negative attrition')
sns.distplot(p_attr['MonthlyIncome'], label='positive attrition')
plt.show()
```



There is high attrition rate among employees of relatively lower salary range, ie. less than 5000.

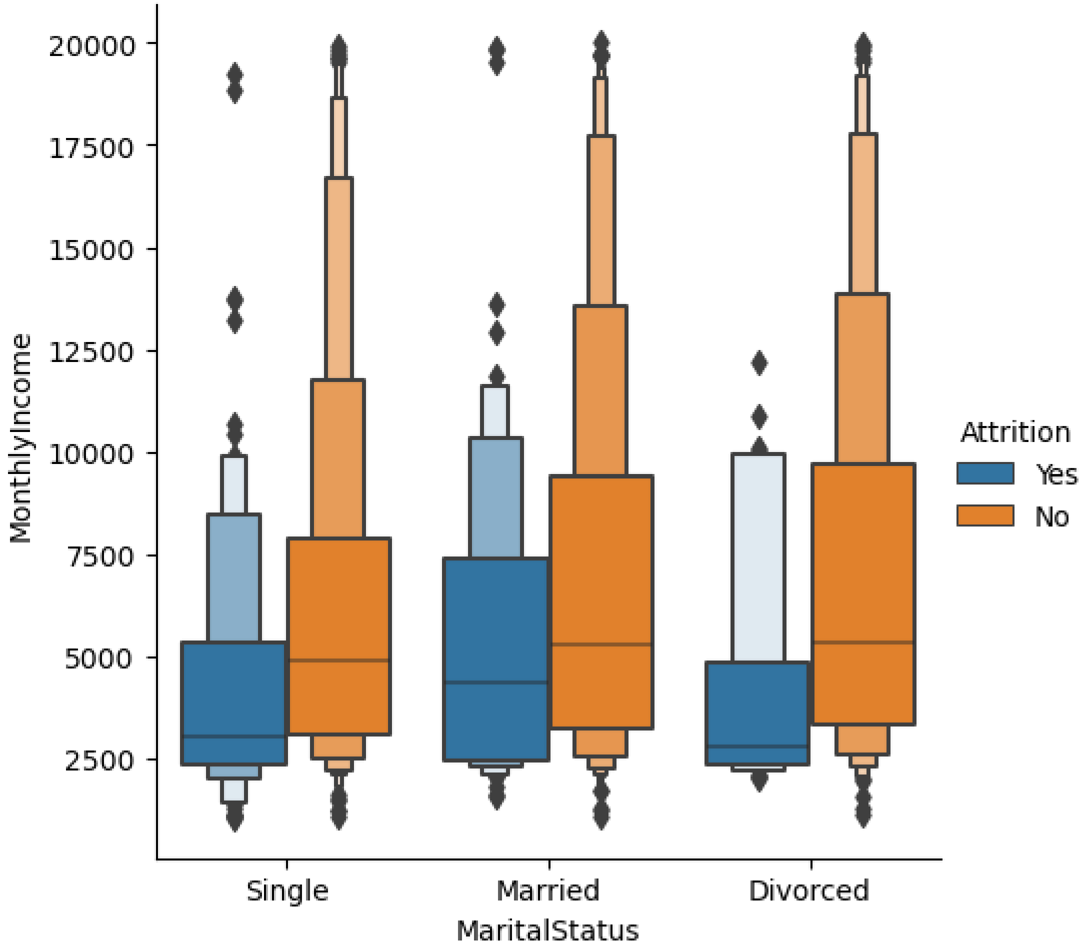
In [29]:

```
sns.catplot(x="Attrition",y="MonthlyIncome",kind="box",data=attr)
plt.title("Box plot grouped by Attrition")
plt.show()
```



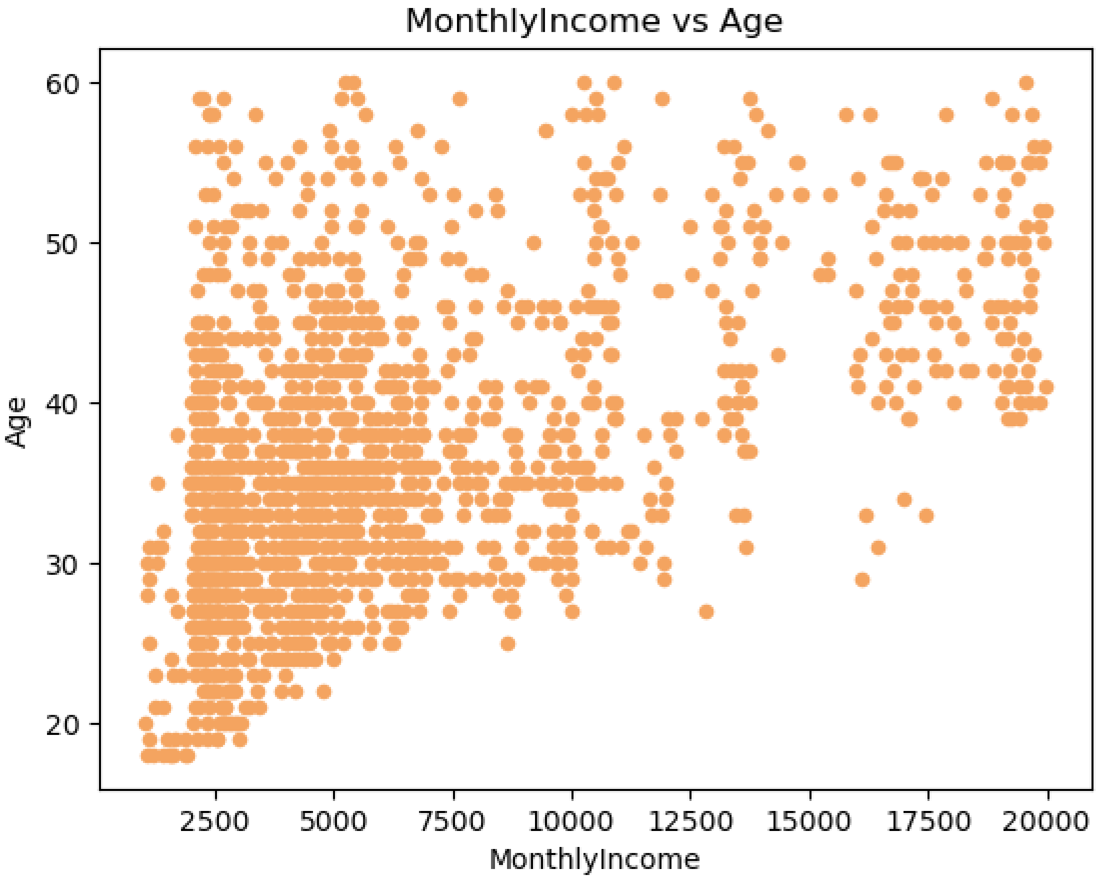
There is high attrition rate among employees of relatively lower salary range, ie. less than 5000.

```
In [30]: sns.catplot(x="MaritalStatus",y="MonthlyIncome",hue="Attrition" ,kind="boxen"
         plt.show()
```

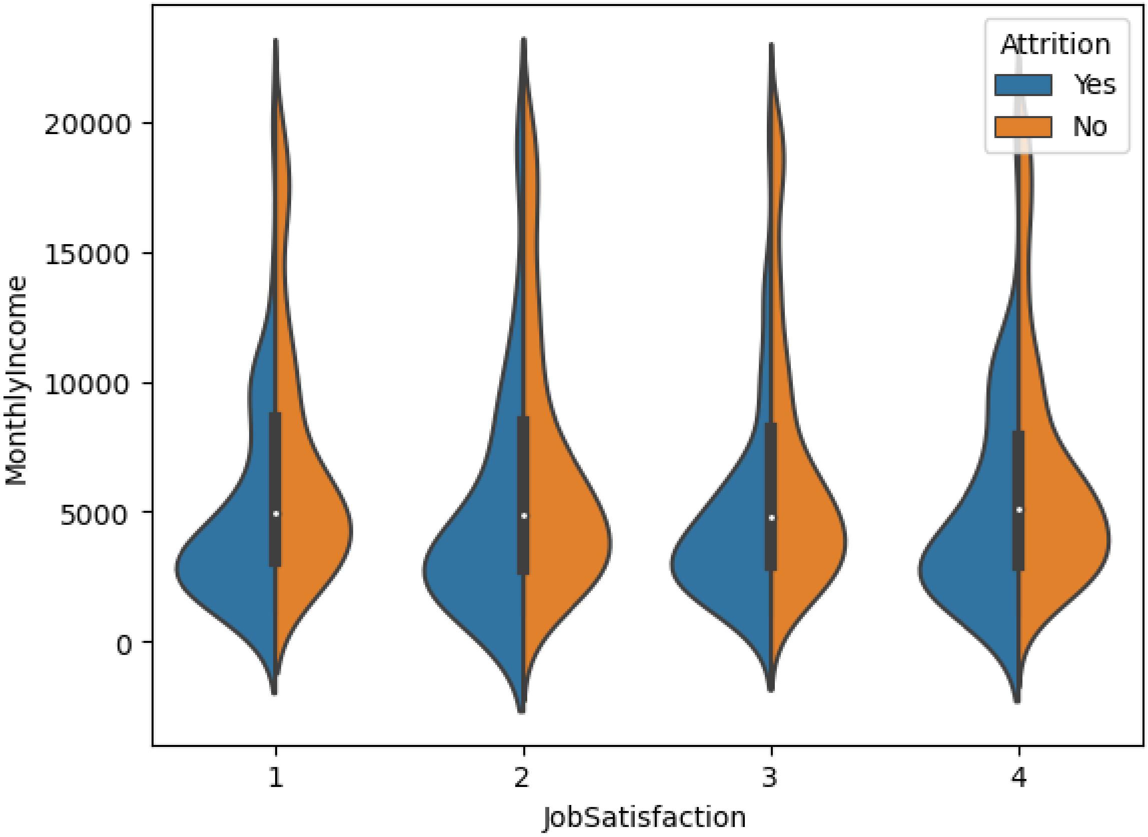


This will show monthly income and attrition using Marital Status category

In [31]:
```
attr.plot.scatter(x = 'MonthlyIncome', y = 'Age',
                  color = 'sandybrown', title = 'MonthlyIncome vs Age')
plt.show()
```



In [32]:
```
sns.violinplot(x ="JobSatisfaction", y ="MonthlyIncome", hue ="Attrition",
data = attr, split = True)
plt.show()
```



JobSatisfaction

There are 1470 employees under our analysis for Job Satisfaction (in our sample)

□ Nominal data(categorical)

□ There are 4 categories in this data sample: 1,2,3 and 4

□ Here, 1 corresponds to least Job Satisfaction and 4 corresponds to the highest Job Satisfaction.

□ The count analysis of the Job Satisfaction category: □ 4 459 □ 3 442 □ 1 289 □ 2 280

□ Highest rating is with ‘Job Satisfaction 4’ with a count of 459. □ Least rating is with ‘Job Satisfaction 1’ with a count of 280.

In [33]:
```
#Number of Unique values
attr["JobSatisfaction"].unique().size
```

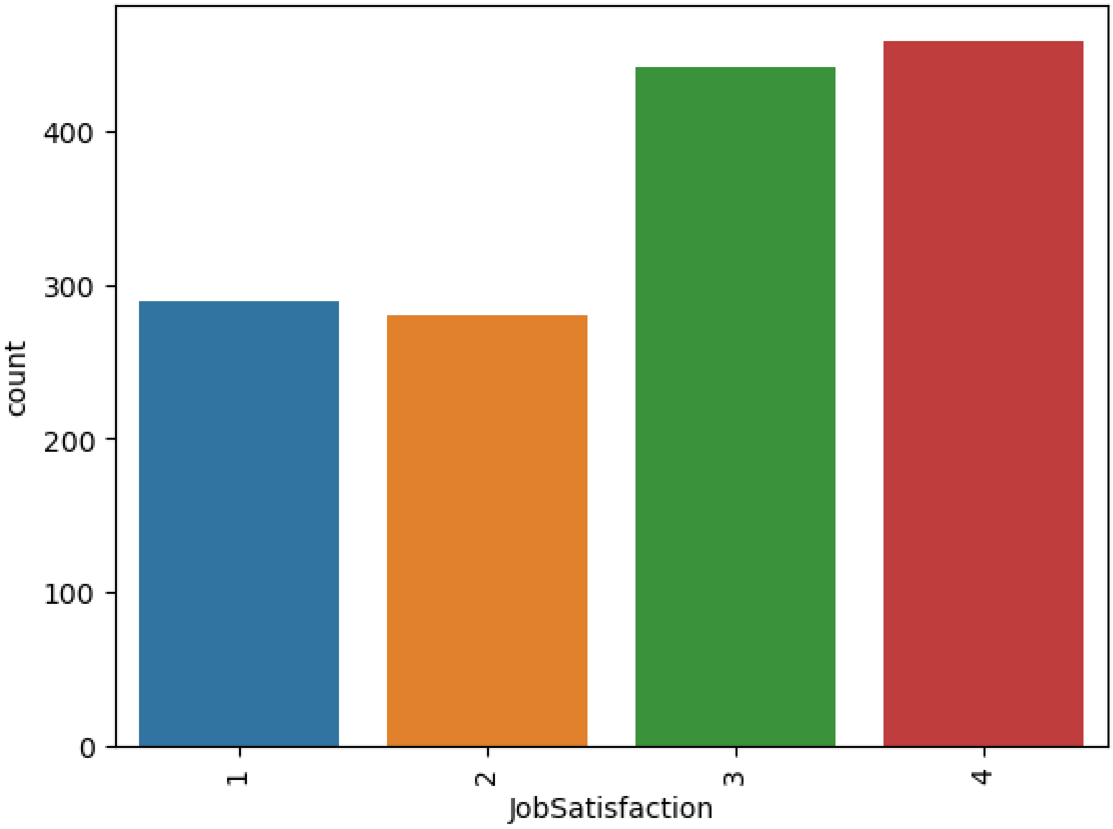
Out[33]:
```
4
```

In [34]:
```
attr["JobSatisfaction"].value_counts()
```

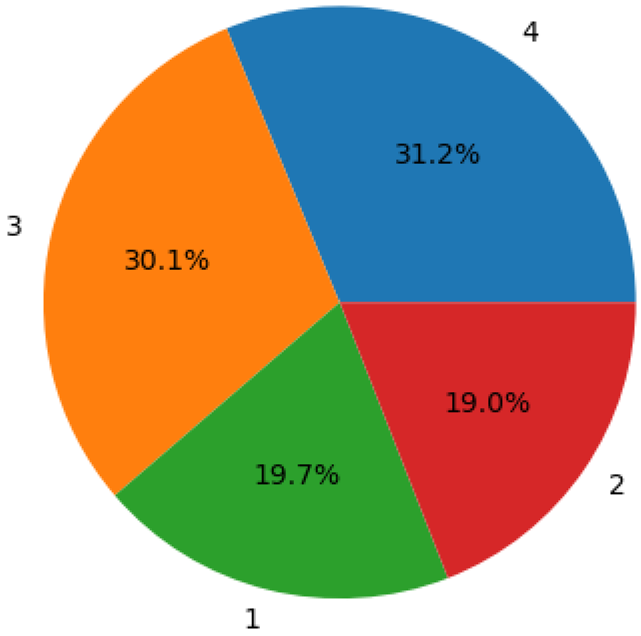
Out[34]:
```
4    459
3    442
1    289
2    280
Name: JobSatisfaction, dtype: int64
```

Here JobSatisfaction Counts are shown by Graph

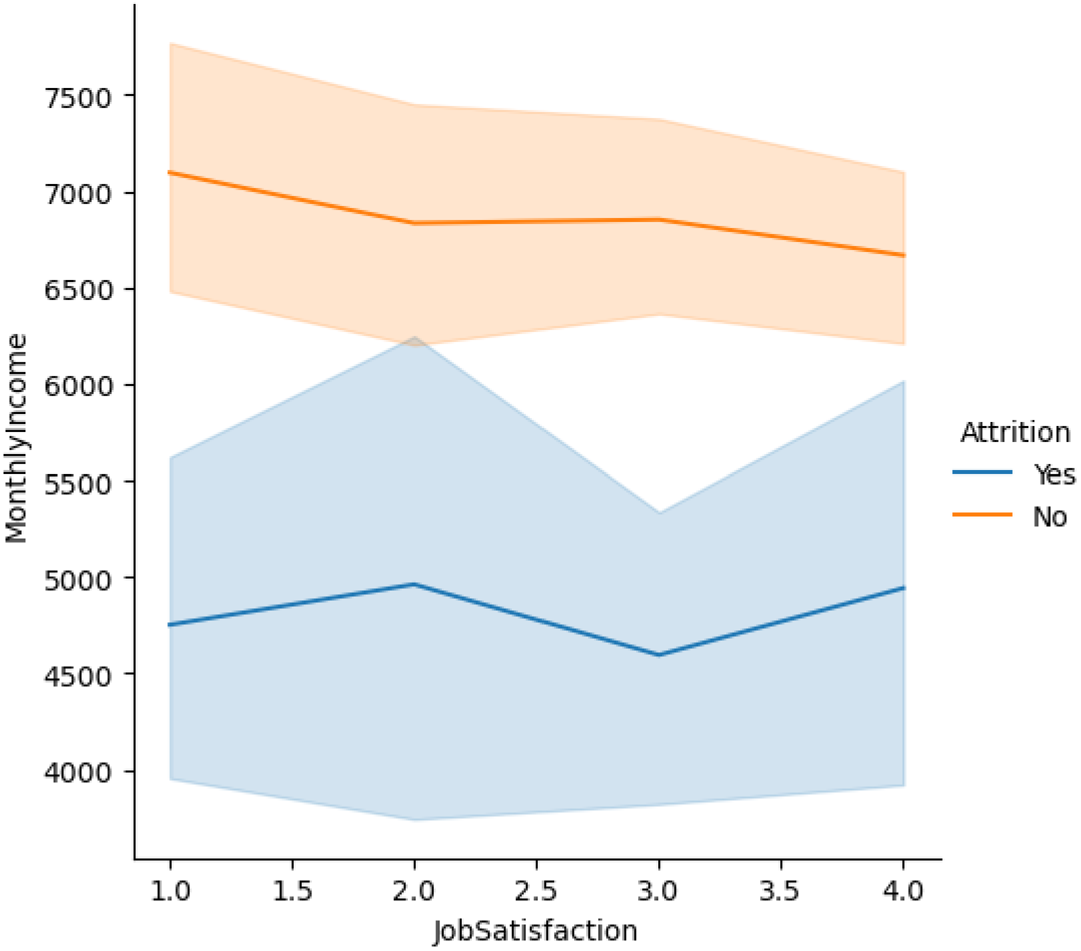
In [35]:
```
sns.countplot(attr["JobSatisfaction"])
plt.xticks(rotation=90)
plt.show()
```



In [36]:
```
labels = ['4', '3','1','2']
plt.pie(attr.JobSatisfaction.value_counts(), labels=labels, autopct='%.1f%%')
plt.show()
```



In [37]:
```
sns.relplot(x="JobSatisfaction",y="MonthlyIncome",hue="Attrition",kind="line"
plt.show()
```



Department

• Nominal data(categorical)

• There are 3 categories in this data sample: Human Resources, Sales and Research & Development.

• The count of employees in these categories are: Human Resources = 63 Sales = 446 Research and Development = 961

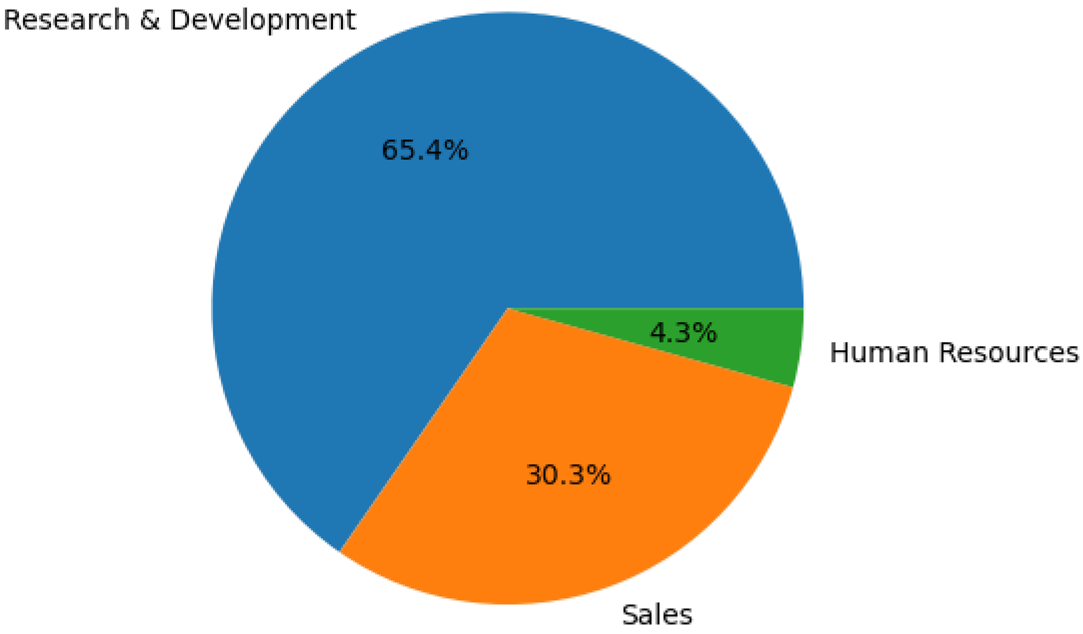
• Under our analysis in the sample, 65.37% of the employees work in the Research and Development department, which is the maximum, 30.34% of the employees work in the Sales department and 4.28% of the employees work in the Human Resources department which is the least in the organization.

In [38]:
```
attr["Department"].value_counts()
```

Out[38]:
```
Research & Development    961
Sales                     446
Human Resources            63
Name: Department, dtype: int64
```

Plot of different department

In [39]:
```
labels = ['Research & Development', 'Sales', 'Human Resources']
plt.pie(attr.Department.value_counts(), labels=labels, autopct='%.1f%%')
plt.show()
```



In [40]:
```
t3 = pd.crosstab(attr.Department,columns = attr.Attrition)
t3
```

Out[40]:

Attrition	No	Yes
Department		
Human Resources	51	12
Research & Development	828	133
Sales	354	92

Replace Attrition to numeric values

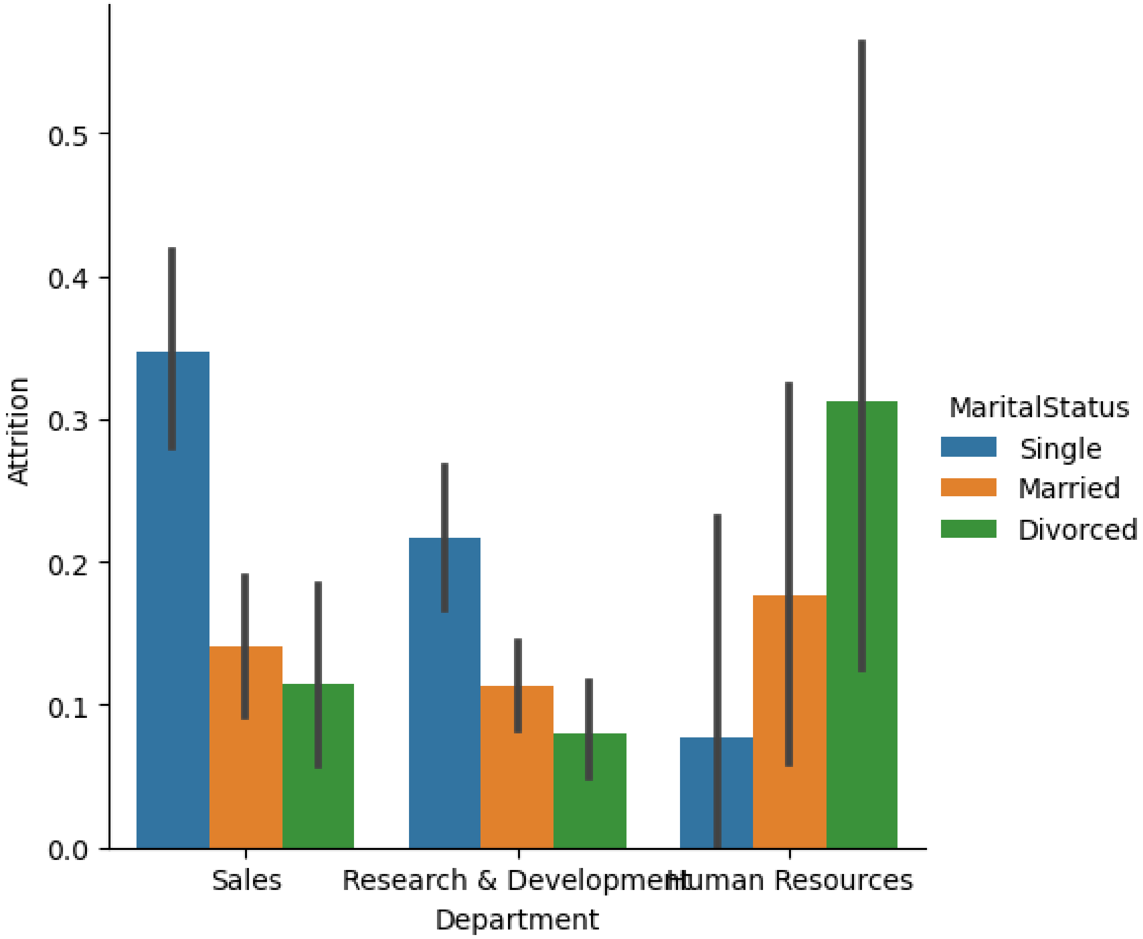
In [41]:
```
attr = attr.replace("Yes", 1)
```

In [42]:
```
attr = attr.replace("No", 0)
```

In [43]:
```
attr.groupby('Department')['Attrition'].mean()
```

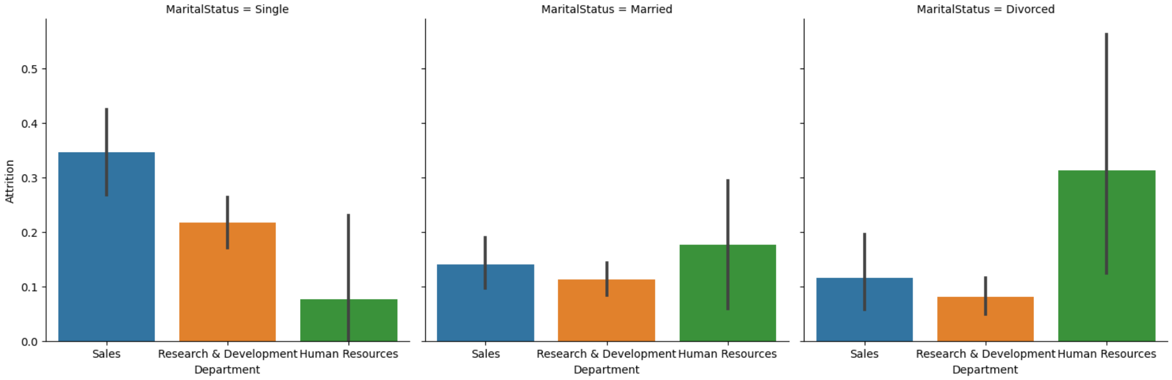
Out[43]:
```
Department
Human Resources           0.190476
Research & Development    0.138398
Sales                     0.206278
Name: Attrition, dtype: float64
```

In [44]:
```
sns.catplot(x="Department",y="Attrition" ,hue="MaritalStatus",kind="bar",data
plt.show()
```



Employees in R&D department seem to have less attrition rate among other department employees, with higher attrition in both Human Resources and Sales Department.

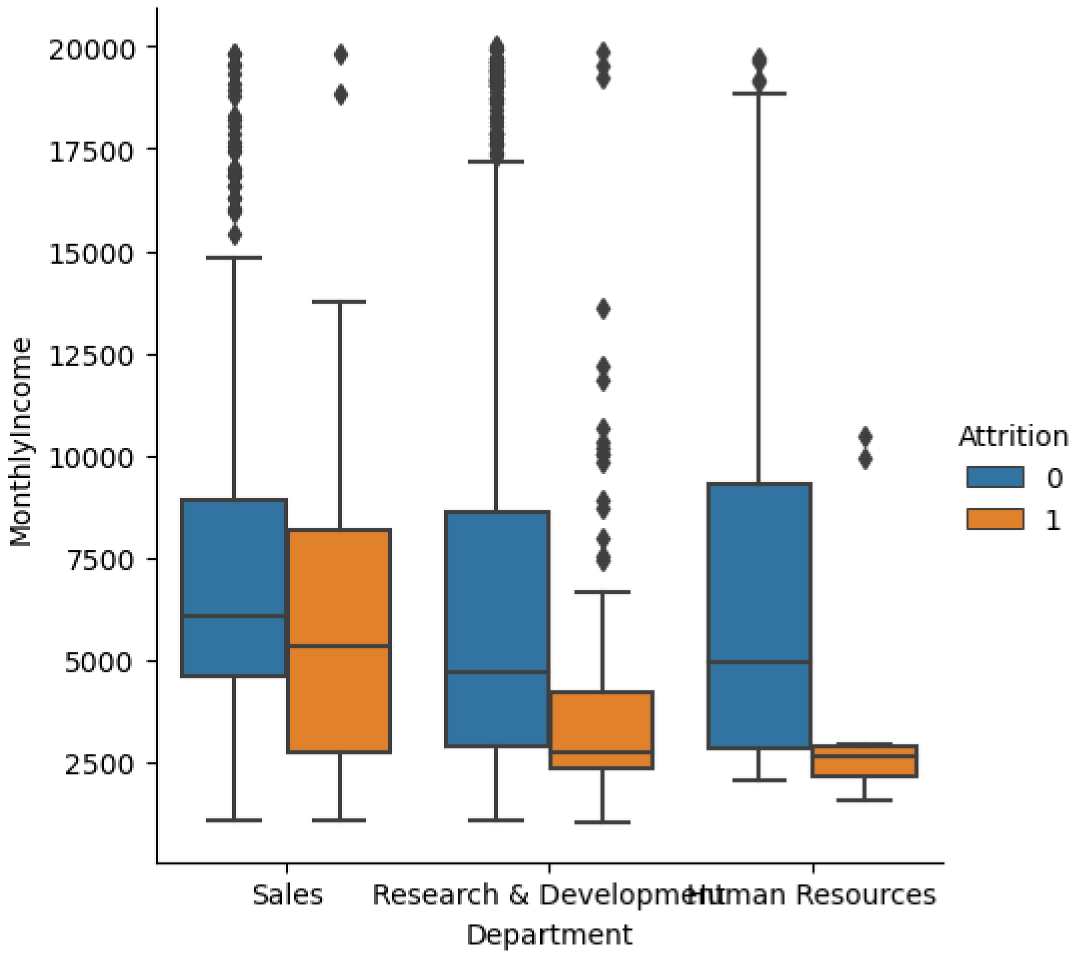
```
In [45]: sns.catplot(x="Department",y="Attrition" ,col="MaritalStatus",kind="bar",data
         plt.show()
```



when drilled down by Marital status, sales singles seem to have almost twice the attrition rate of sales married employees.

Categorical plot between MonthlyIncome and Department and Attrition

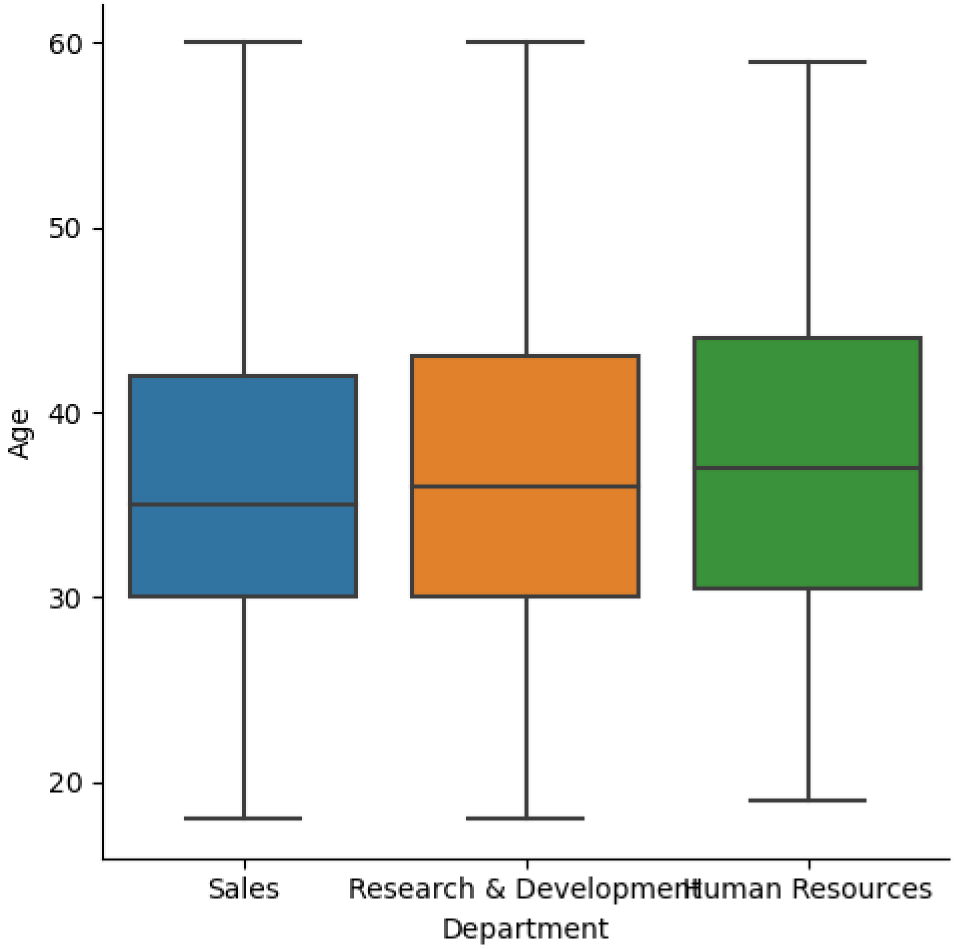
In [46]:
```
sns.catplot(x="Department",y="MonthlyIncome" ,hue="Attrition",kind="box",data
plt.show()
```



Here shows less income employee from human resorce has more Attrition

In [47]:

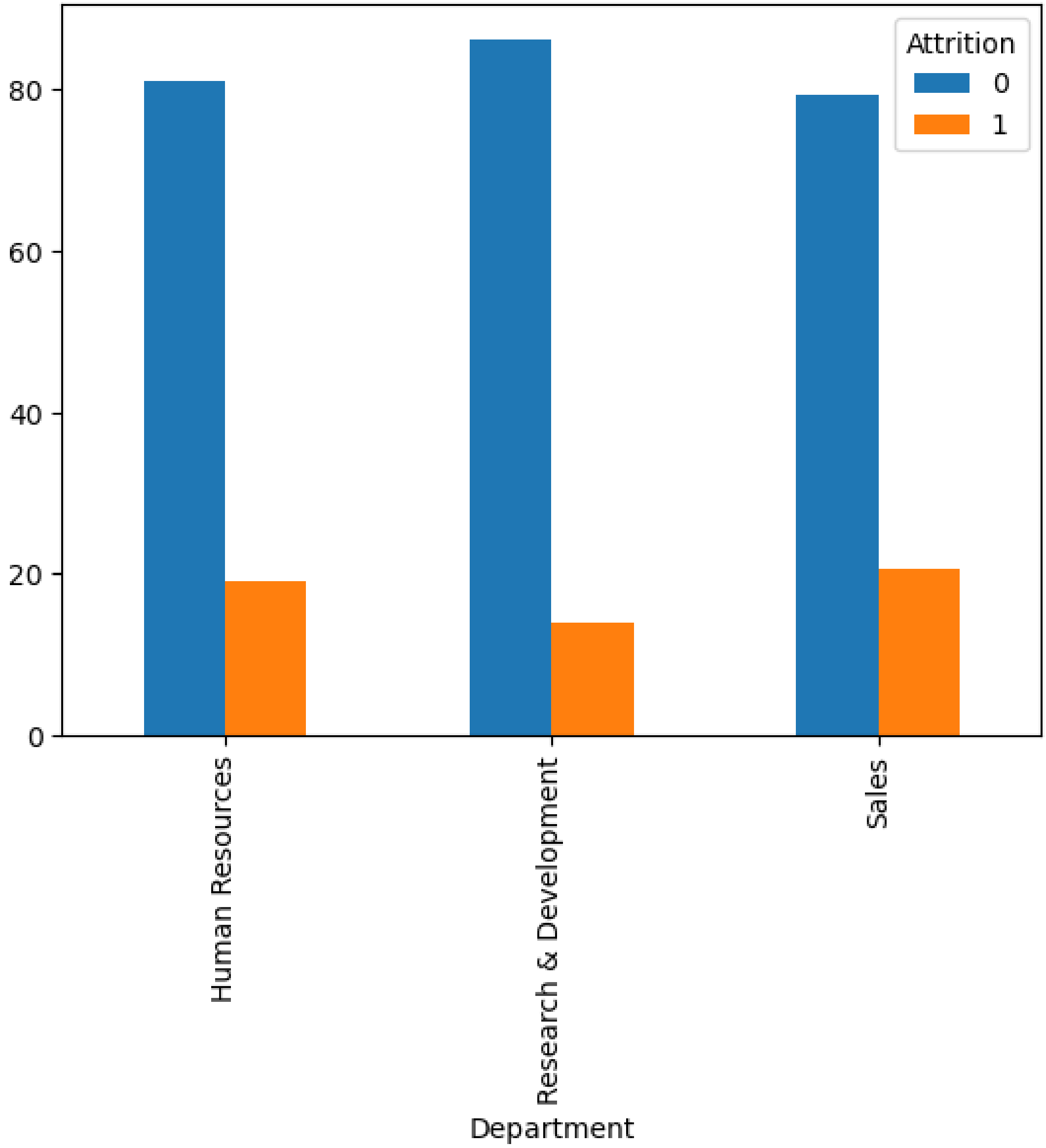
```
sns.catplot(x="Department",y="Age" ,kind="box",data=attr)
plt.show()
```



It shows relation between departments and age which normally distributed

In [48]:

```
tab=pd.crosstab(attr["Department"],attr["Attrition"],normalize='index')*100

tab.plot(kind="bar")
plt.show()


print(tab)
```



```
Attrition                       0          1
Department
Human Resources         80.952381  19.047619
Research & Development  86.160250  13.839750
Sales                   79.372197  20.627803
```

Education Field

- Nominal data(categorical)

- There are 6 categories in this data sample: Human Resources, Life sciences, Marketing, Medical, Other and Technical degree

- The count of employees from these educational backgrounds are: Human Resources = 27 Life Sciences = 606 Marketing = 159 Medical = 464 Technical Degree = 132 Other = 82

- Under our analysis in the sample, 41.22% of the employees are from life sciences educational background, which is the maximum, 31.56% of the employees are from medical educational background, 10.81% of the employees are from marketing educational background, 8.97% of the employees are from technical degree educational background, 5.57% of the employees are from other educational background, 1.83% of the employees are from Human Resources background, which is the least in the organization.

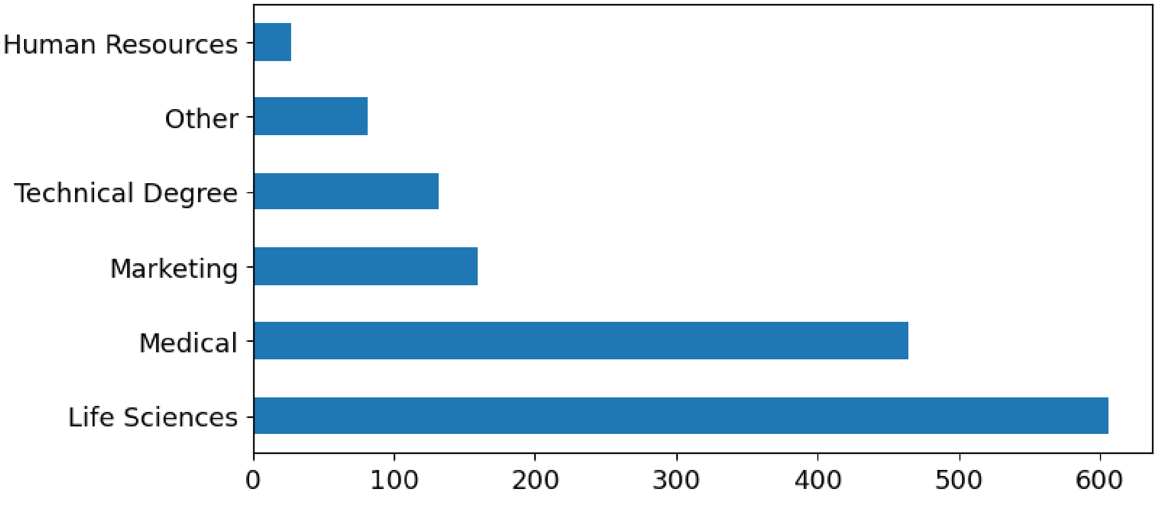
In [49]:
```
#Number of Unique values
attr["EducationField"].nunique()
```

Out[49]:
```
6
```

In [50]:
```
attr["EducationField"].value_counts()
```

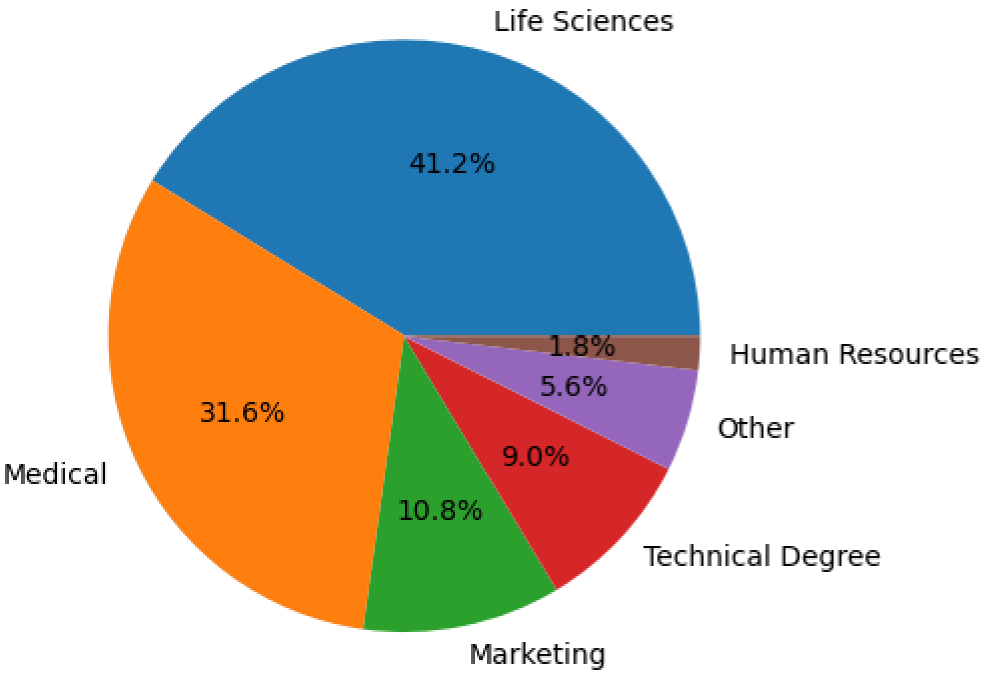
Out[50]:
```
Life Sciences       606
Medical             464
Marketing           159
Technical Degree    132
Other                82
Human Resources      27
Name: EducationField, dtype: int64
```

In [51]:
```
attr["EducationField"].value_counts().plot(kind="barh",figsize=(8,4),fontsize
plt.show()
```



In [52]:

```
labels = ['Life Sciences', 'Medical', 'Marketing','Technical Degree','Other',
plt.pie(attr.EducationField.value_counts(), labels=labels, autopct='%.1f%%')
plt.show()
```



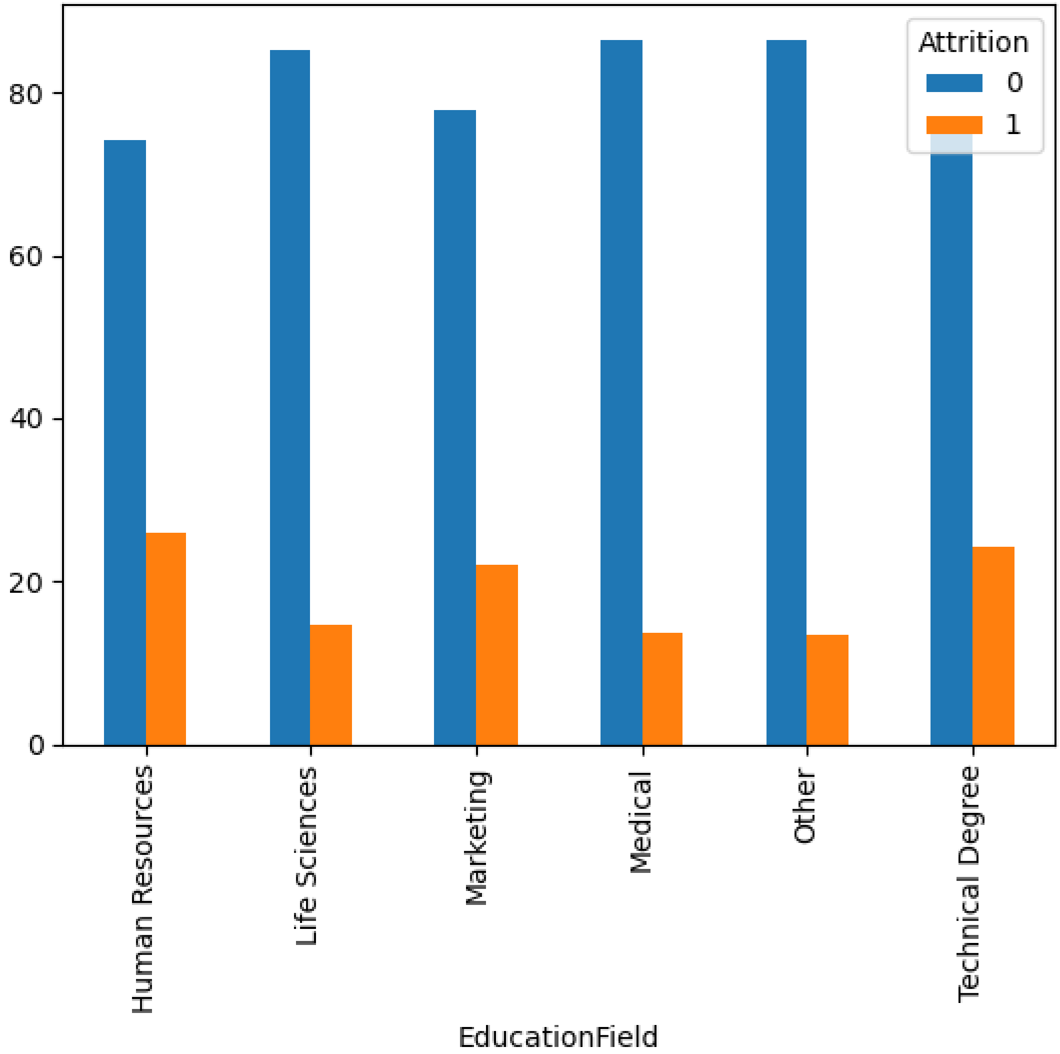
In [53]:

```
t4 = pd.crosstab(attr.EducationField,columns = attr.Attrition)
t4
```

Out[53]:

Attrition	0	1
EducationField		
Human Resources	20	7
Life Sciences	517	89
Marketing	124	35
Medical	401	63
Other	71	11
Technical Degree	100	32

```
In [54]: tab=pd.crosstab(attr["EducationField"],attr["Attrition"],normalize='index')*1

         tab.plot(kind="bar")
         plt.show()


         print(tab)
```



```
Attrition                0          1
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
```

We can see that HR(25%),Marketing(22%),Technical(24%) education have attrition rates than the average of 16%. Hence we can say that these 3 education result in higher attrition

Can low salary be one of the reasons?

Employees with Life Sciences, Medical and other educational background are leaving less as compared to HR,Marketing and Tech

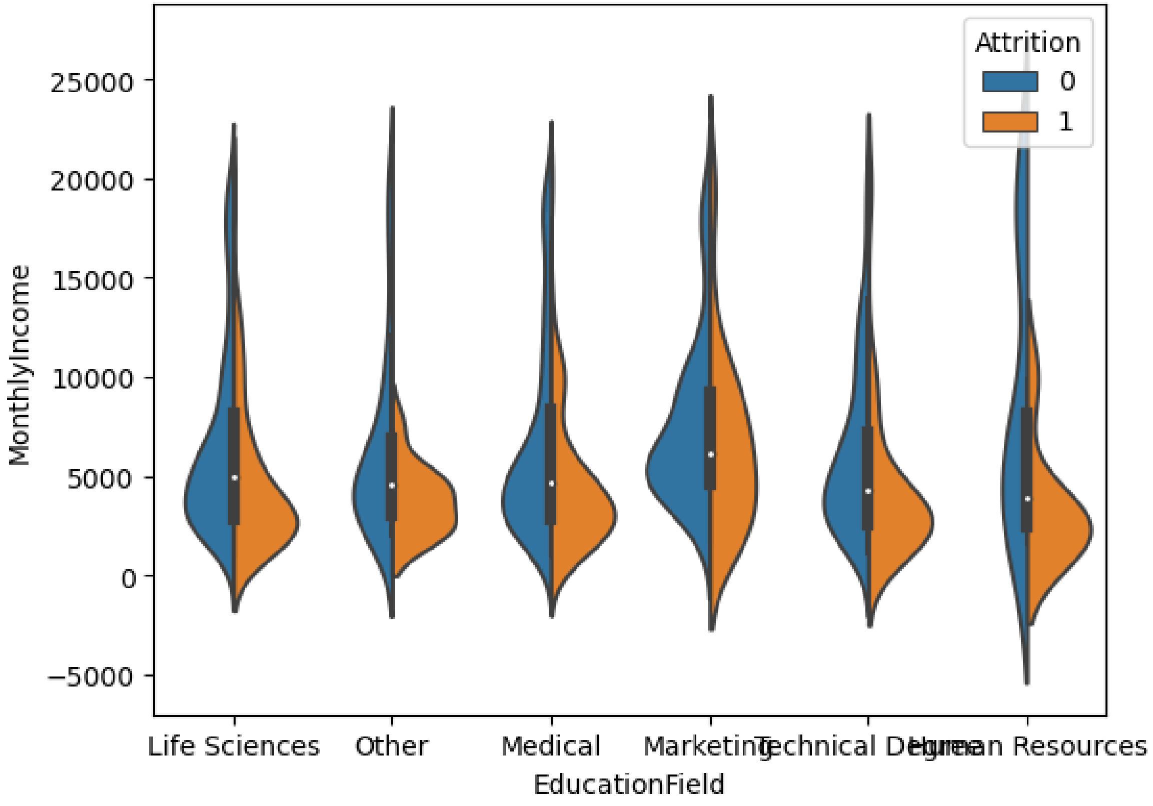
Can low salary be a reason? let's find out

```
In [55]: attr.groupby('EducationField').MonthlyIncome.mean()
```

```
Out[55]: EducationField
         Human Resources     7241.148148
         Life Sciences       6463.288779
         Marketing           7348.584906
         Medical             6510.036638
         Other               6071.548780
         Technical Degree    5758.303030
         Name: MonthlyIncome, dtype: float64
```

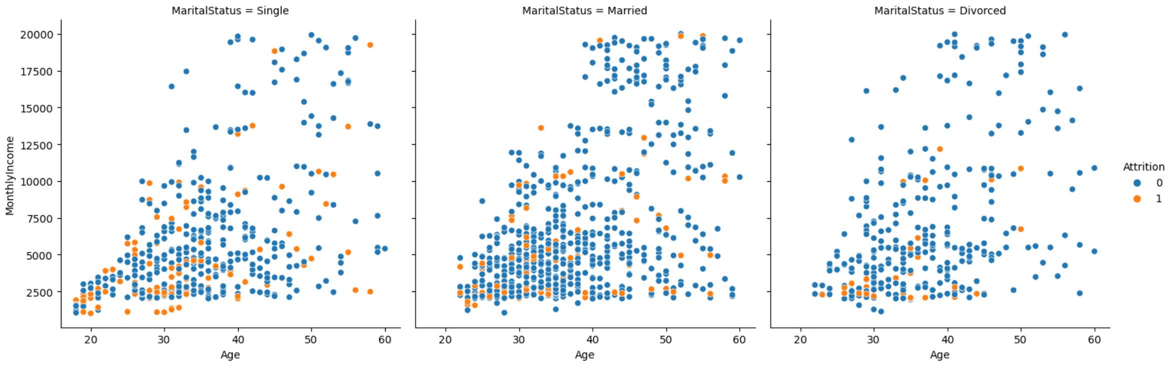
No,because even though the salary of HR and marketing employees is higher, their leaving rate is highest.

```
In [56]: sns.violinplot(x ="EducationField", y ="MonthlyIncome", hue ="Attrition",
         data = attr, split = True)
         plt.show()
```



Subplot

In [57]:
```
sns.relplot(x="Age",y="MonthlyIncome",hue="Attrition",col="MaritalStatus",dat
plt.show()
```



Environmental Satisfaction

There are 1470 employees under our analysis for Environment Satisfaction (in our sample)

• Nominal data(categorical)

• There are 4 categories in this data sample: 1,2,3 and 4.

• Here, 1 corresponds to least environment satisfaction and 4 corresponds to the highest environment satisfaction.

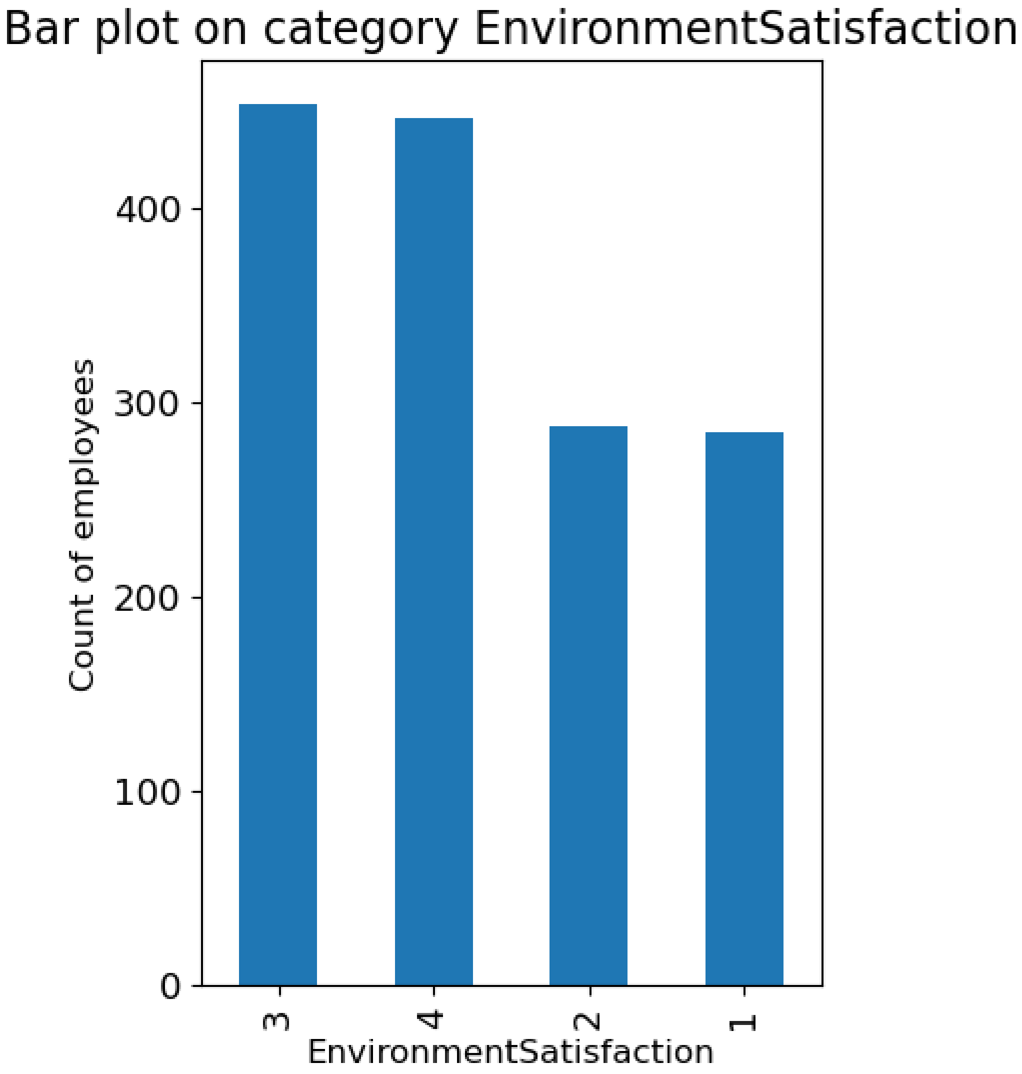
• The count analysis of the Environment satisfaction category: • The employees who gave Environment satisfaction rating 1 is 284(19.34%) • The employees who gave Environment satisfaction rating 2 is 287 (19.52%) • The employees who gave Environment satisfaction rating 3 is 453 (30.8%) • The employees who gave Environment satisfaction rating 4 is 446 (30.34%)

□ Most rating is with Environment satisfaction 3 with a count of 453 □ Least rating is with Environment satisfaction 1 with a count of 284

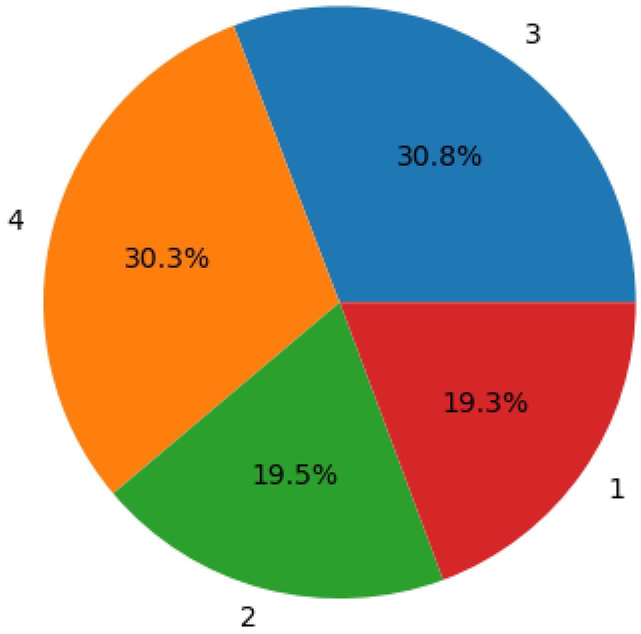
In [58]:
```
#Number of Unique values
attr["EnvironmentSatisfaction"].value_counts()
```

Out[58]:
```
3    453
4    446
2    287
1    284
Name: EnvironmentSatisfaction, dtype: int64
```

In [59]:
```
attr["EnvironmentSatisfaction"].value_counts().plot(kind="bar",figsize=(4,6),
plt.xlabel("EnvironmentSatisfaction", fontsize=12)
plt.ylabel("Count of employees", fontsize=12)
plt.title("Bar plot on category EnvironmentSatisfaction", fontsize=16)
plt.show()
```



In [60]:
```
labels = ['3', '4', '2','1']
plt.pie(attr.EnvironmentSatisfaction.value_counts(), labels=labels, autopct='
plt.show()
```



In [61]:
```
t5 = pd.crosstab(attr.EnvironmentSatisfaction,columns = attr.Attrition)
t5
```

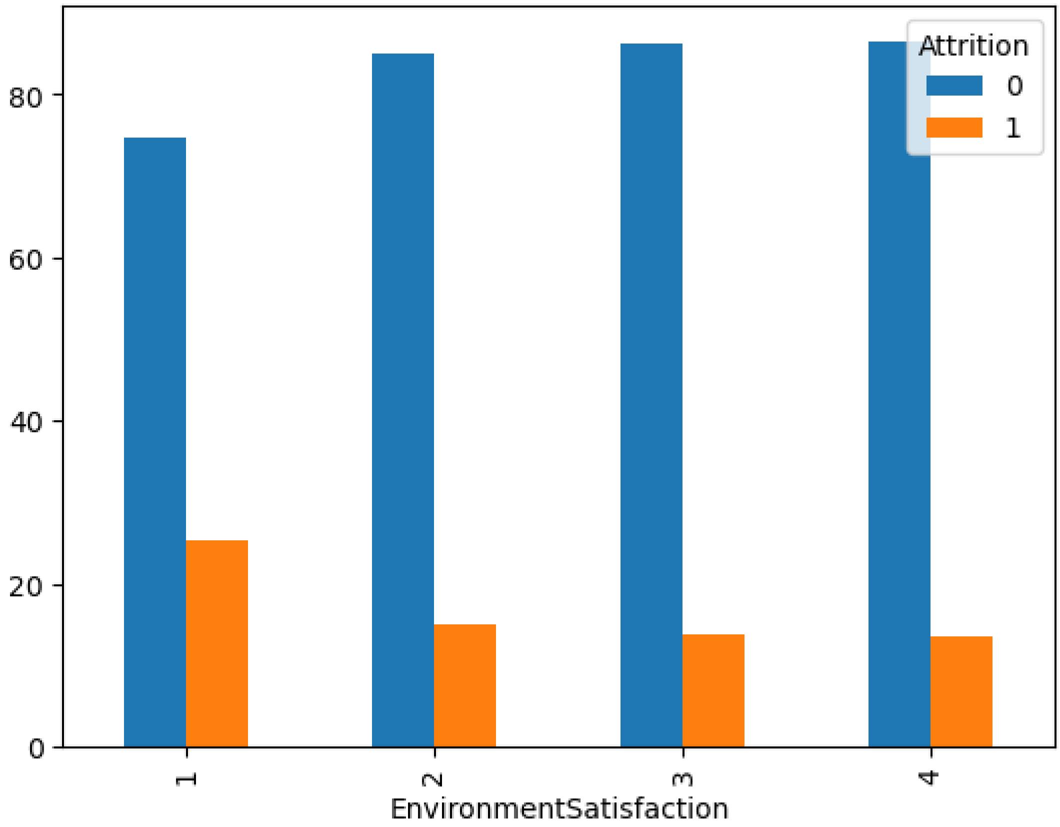
Out[61]:

Attrition	0	1
EnvironmentSatisfaction		
1	212	72
2	244	43
3	391	62
4	386	60

percentage of EnvironmentSatisfaction numbers using attrition

In [62]:
```
tab = pd.crosstab(attr["EnvironmentSatisfaction"],attr["Attrition"],normalize
print(tab)
tab.plot(kind='bar')
plt.show()
```

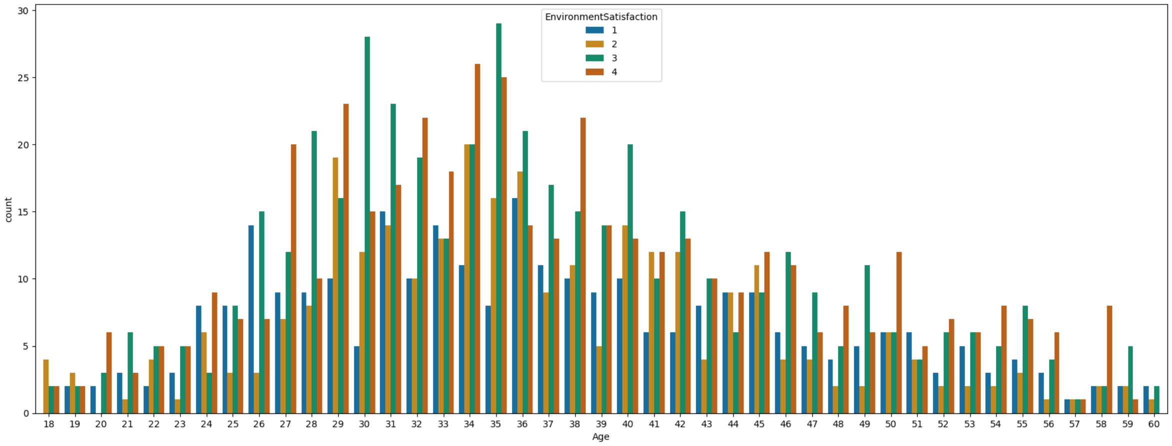
```
Attrition                        0          1
EnvironmentSatisfaction
1                        74.647887  25.352113
2                        85.017422  14.982578
3                        86.313466  13.686534
4                        86.547085  13.452915
```



Here, we can see that people with low environmental score of 1 have higher attrition rate(25%) than the average attrition rate of 16%

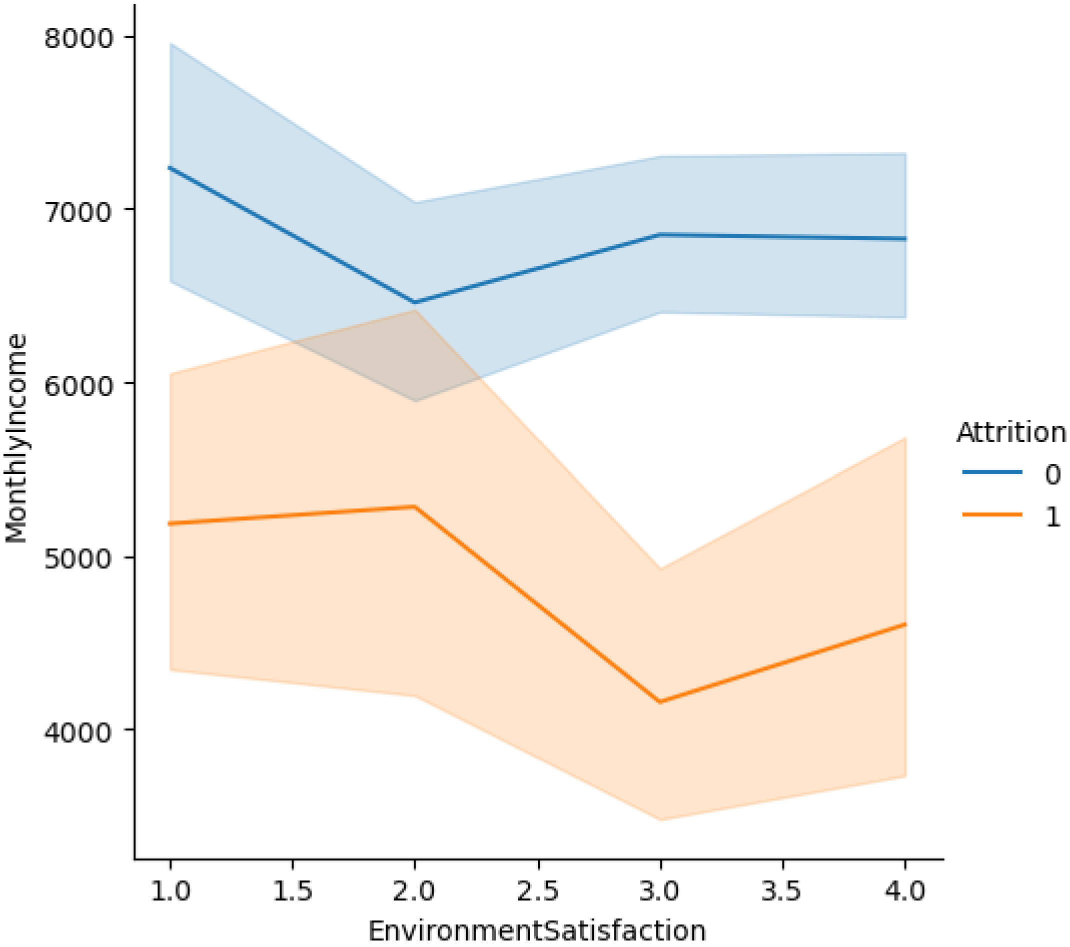
In [63]:

```
plt.subplots (figsize=(22,8))
sns.countplot (x='Age',  hue='EnvironmentSatisfaction', data= attr, palette =
plt.show()
```



In [64]:

```
sns.relplot(x="EnvironmentSatisfaction",y="MonthlyIncome",hue="Attrition",kin
plt.show()
```



MaritalStatus

There are 1470 employees under our analysis for Marital Status (in our sample)

□ Nominal data(categorical)

□ There are 3 categories in this data sample: Single, Married and Divorced.

□ The count analysis of the Marital Status category: □ Married 673 □ Single 470 □ Divorced 327

• Employees with the “Married” status have maximum count in the given data set. • Employees with the “Divorced” status have minimum count in the given data set.

In [65]:

```
attr["MaritalStatus"].value_counts()
```

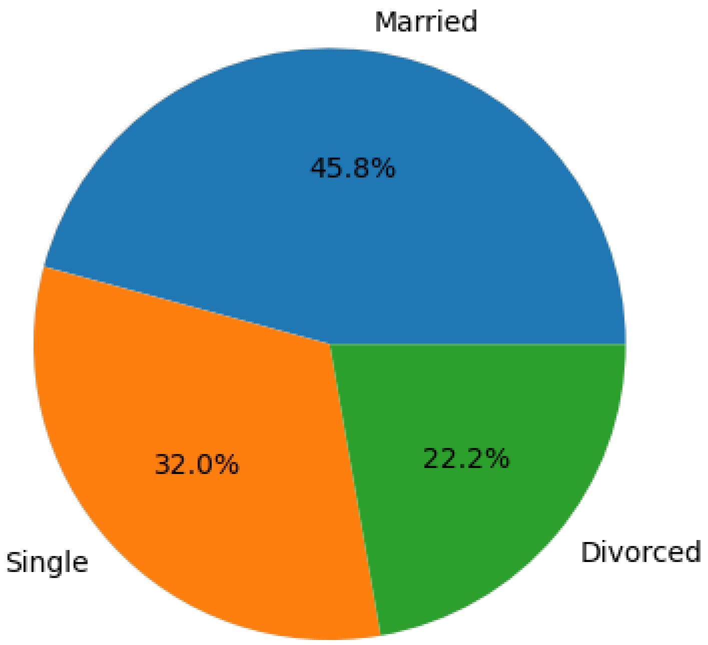
Out[65]:

```
Married     673
Single      470
Divorced    327
Name: MaritalStatus, dtype: int64
```

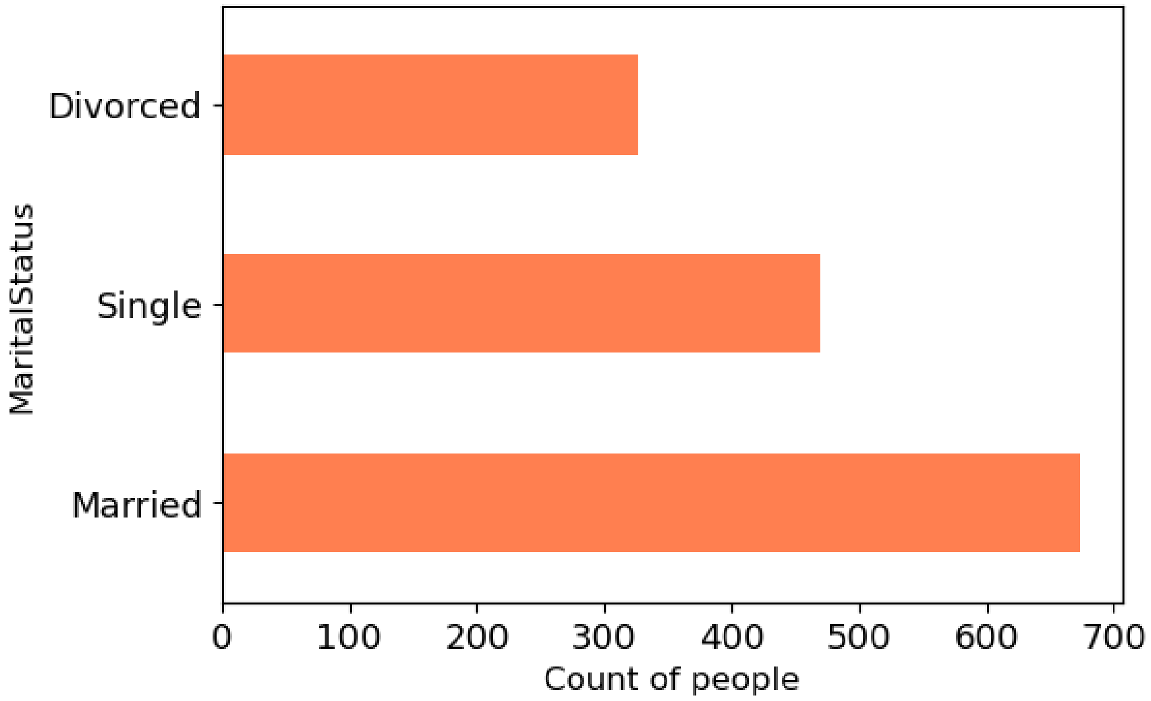
Who is leaving more? married ,un-married or divorced employees?

In [66]:

```
labels = ['Married', 'Single','Divorced']
plt.pie(attr.MaritalStatus.value_counts(), labels=labels, autopct='%.1f%%')
plt.show()
```



In [67]:
```
attr['MaritalStatus'].value_counts().plot(kind='barh', figsize=(6,4), color="

plt.xlabel("Count of people", fontsize=12)
plt.ylabel("MaritalStatus", fontsize=12)

plt.show()
```

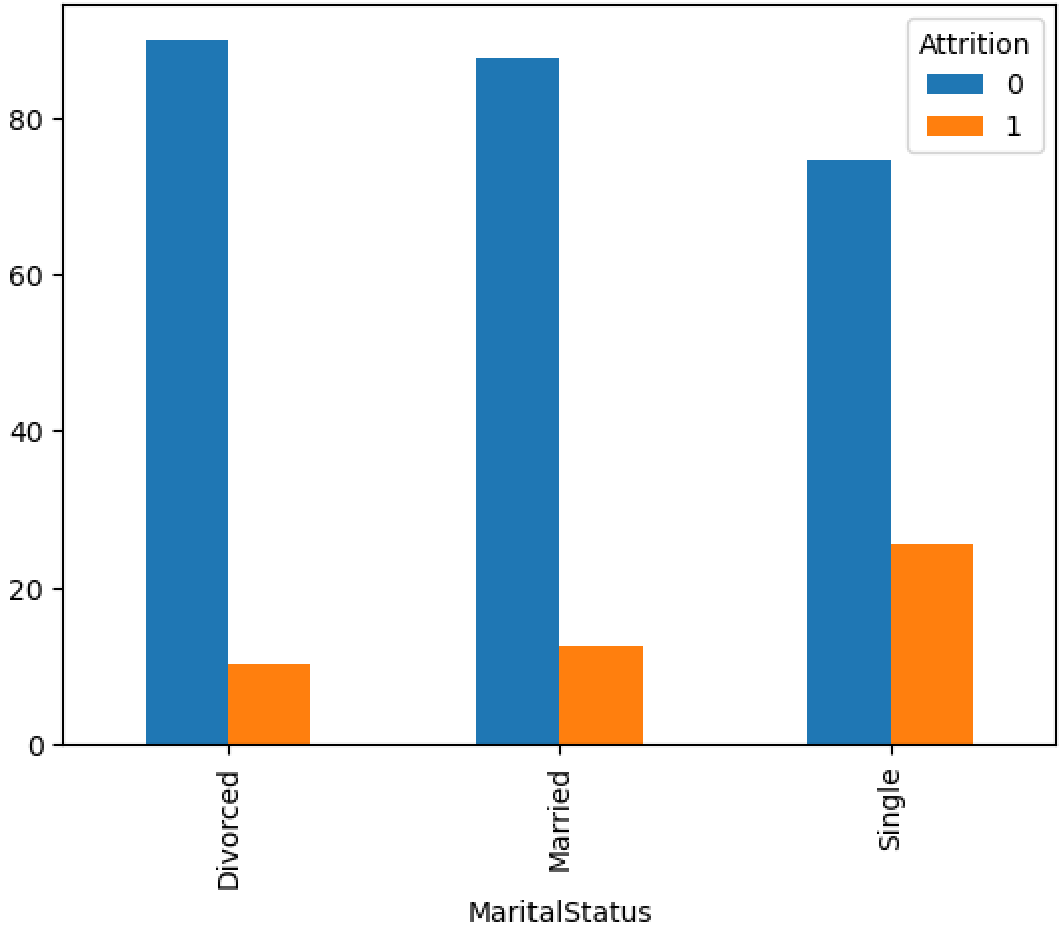


In [68]:
```
attr.groupby('MaritalStatus').Attrition.mean()
```

Out[68]:
```
MaritalStatus
Divorced    0.100917
Married     0.124814
Single      0.255319
Name: Attrition, dtype: float64
```

Singles tend to have a little higher attrition rate than others.

```
In [69]: tab = pd.crosstab(attr["MaritalStatus"],attr["Attrition"],normalize='index')*

         tab.plot(kind='bar')
         plt.show()
```



Single employee has high attrition rate(25%) compared to average attrition rate of 16%.Similarly, divorced employees have low attrition rate(10%) compared to the average of 16%.

WorkLifeBalance

There are 1470 employees under our analysis for Work Life Balance(in our sample) • Nominal data(categorical)

• There are 4 categories in this data sample: 1,2,3 and 4.

• Here, 1 corresponds to least Work Life Balance and 4 corresponds to the highest Work Life Balance.

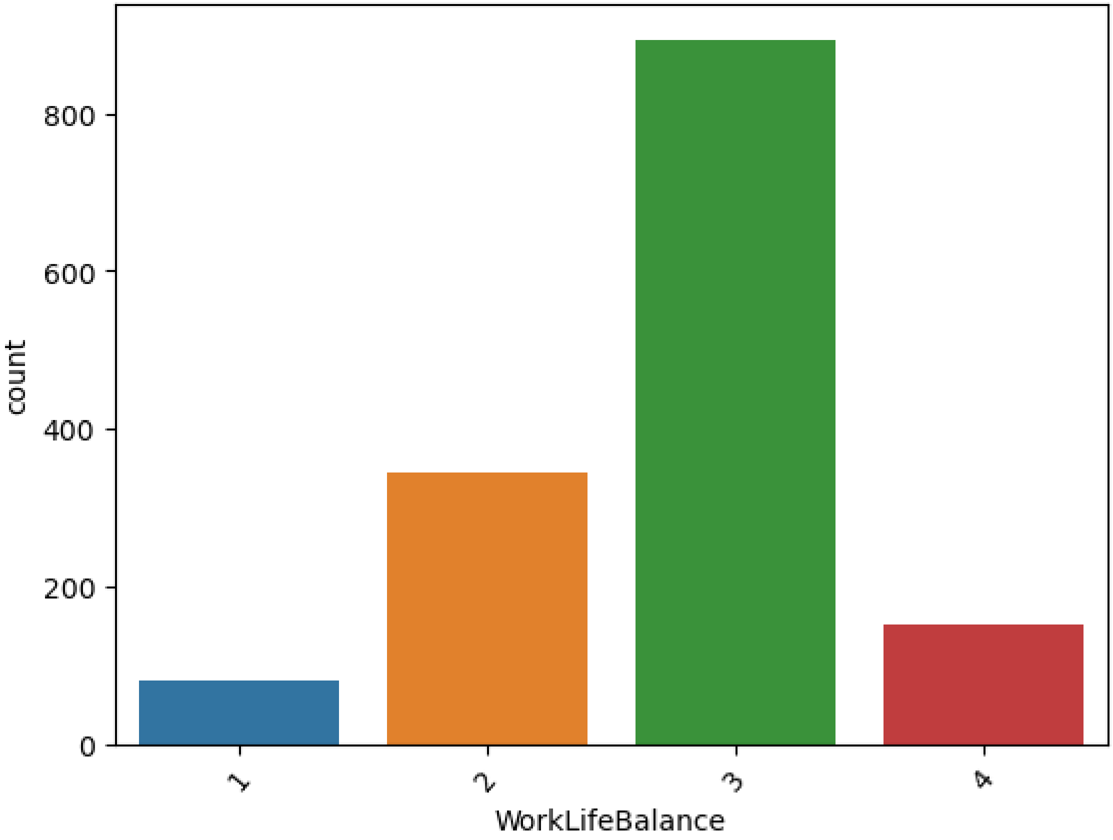
• The count analysis of the Work Life Balance category: • The employees who gave Work Life Balance rating 1 is 80 (5.44%) • The employees who gave Work Life Balance rating 2 is344 (23.40%) • The employees who gave Work Life Balance rating 3 is 893 (60.07%) • The employees who gave Work Life Balance rating 4 is 153 (10.40%)

□ Most rating is with Work Life Balance is 3 with a count of 893 □ Least rating is with Work Life Balance 1 with a count of 80

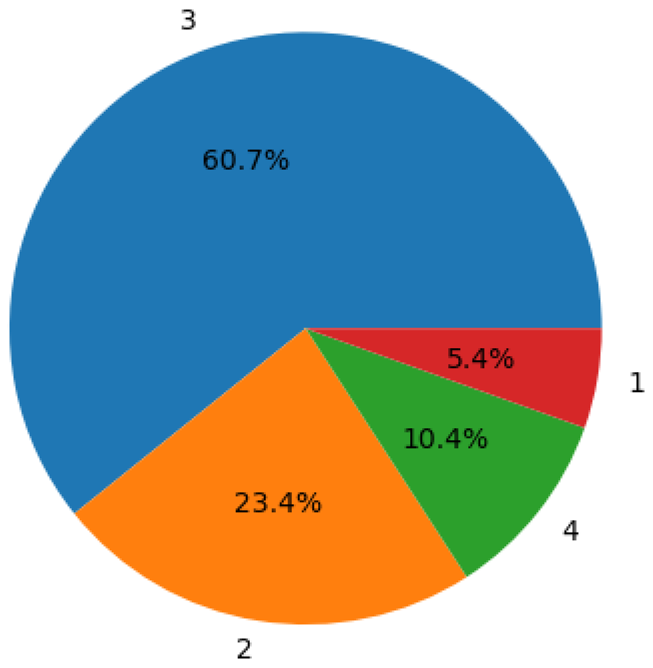
```
In [70]: attr["WorkLifeBalance"].value_counts()
```

```
Out[70]: 3    893
         2    344
         4    153
         1     80
         Name: WorkLifeBalance, dtype: int64
```

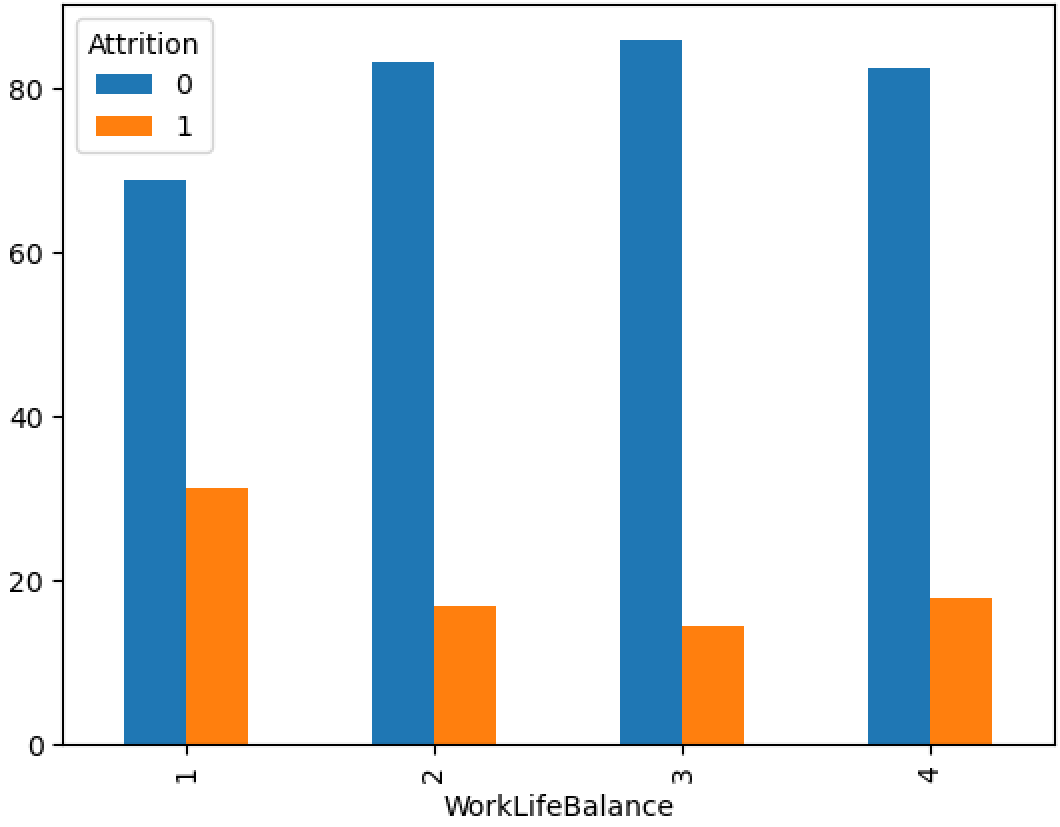
```
In [71]: sns.countplot(attr["WorkLifeBalance"])
         plt.xticks(rotation=50)
         plt.show()
```



```
In [72]: labels = ['3', '2','4','1']
         plt.pie(attr.WorkLifeBalance.value_counts(), labels=labels,autopct='%.1f%%')
         plt.show()
```



In [73]:
```
tab = pd.crosstab(attr["WorkLifeBalance"],attr["Attrition"],normalize='index'

tab.plot(kind='bar')
plt.show()
```



People with work life balance of 1 have high attrition rate(31%) compared to the population attrition rate of 16%.

Hence poor work life balance increases attrition.

Years At Company

There are 1470 employees under our analysis for Years At Company (in our sample)

• Ordinal data(numerical)

• Range of Years At Company is between 0 years to 40 years: This company has employees across the spectrum from zero years experience till employees with 40 years of experience.

• Mean years in company of employees in the data sample is 7.008

• 25%(1st quartile) of the employees under our analysis in our sample are of years less than or equal to 3 years. • 50%(median-2nd quartile) of the employees under our analysis in our sample are of years less than or equal to 5. It provides the years 'midpoint' of employees of organization; there are the same number of people who are have more years than the median years. • 75%(3rd quartile) of the employees under our analysis in our sample have Years At Company less than or equal to 9.

• At least 75% of the employees under our analysis in our sample have Years At Company between 5.24(mean-2sd) and 19.24(mean+2sd)

• The maximum number(mode) of employees with years in the company under our analysis in our sample is 5 years.

• The minimum number of employees of the years in the company under our analysis in our sample are of age 1. (4 in number)

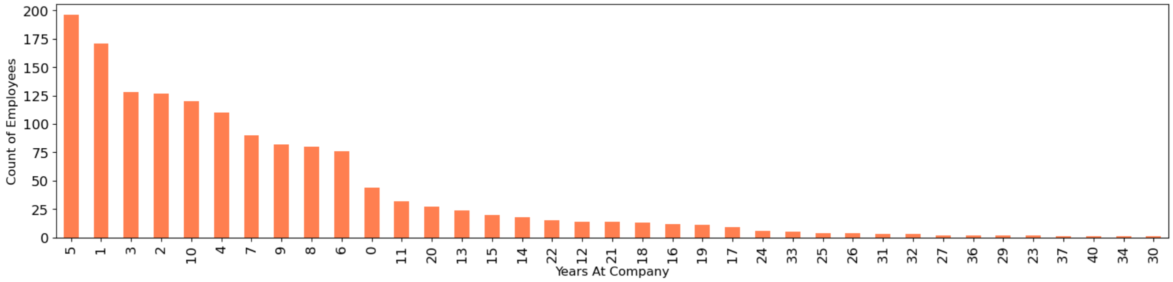
• The mean is greater than the median, so the distribution of variable mean will have its tail

```
In [74]: attr.YearsAtCompany.describe()
```

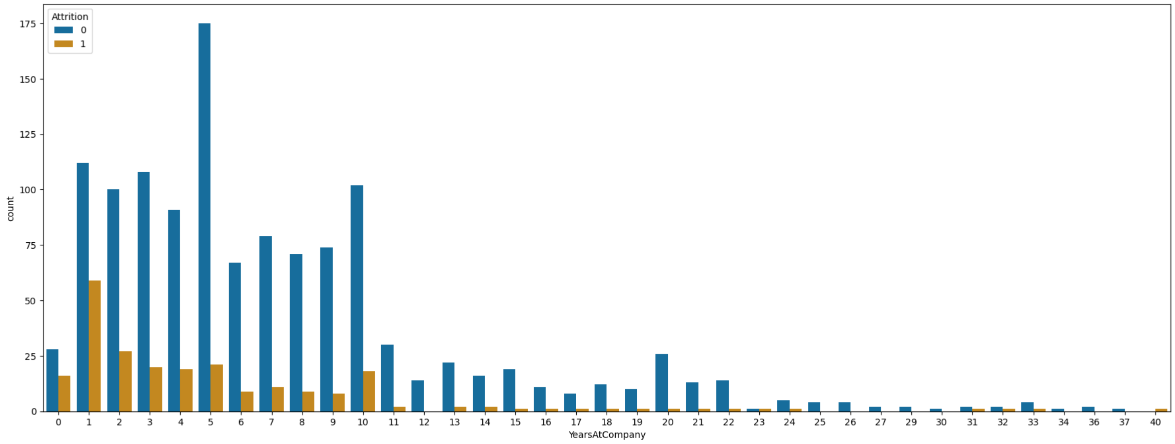
```
Out[74]: count    1470.000000
         mean        7.008163
         std         6.126525
         min         0.000000
         25%         3.000000
         50%         5.000000
         75%         9.000000
         max        40.000000
         Name: YearsAtCompany, dtype: float64
```

Are Employees leaving after working for certain number years?

```
In [75]: attr['YearsAtCompany'].value_counts().plot(kind='bar', figsize=(19,4), color=

         plt.xlabel("Years At Company", fontsize=12)
         plt.ylabel("Count of Employees", fontsize=12)
         plt.show()
```



In [76]:
```
plt.subplots (figsize=(22,8))
sns.countplot (x='YearsAtCompany',  hue='Attrition', data= attr, palette = 'c
plt.show()
```



Employees who have worked for less number of years tend to leave more.

Distance From Home

There are 1470 employees under our analysis for ‘Distance from Home’ (in our sample).
• Ordinal data(numerical)

• Range of ‘Distance from Home’ is between 1 years to 29 years.

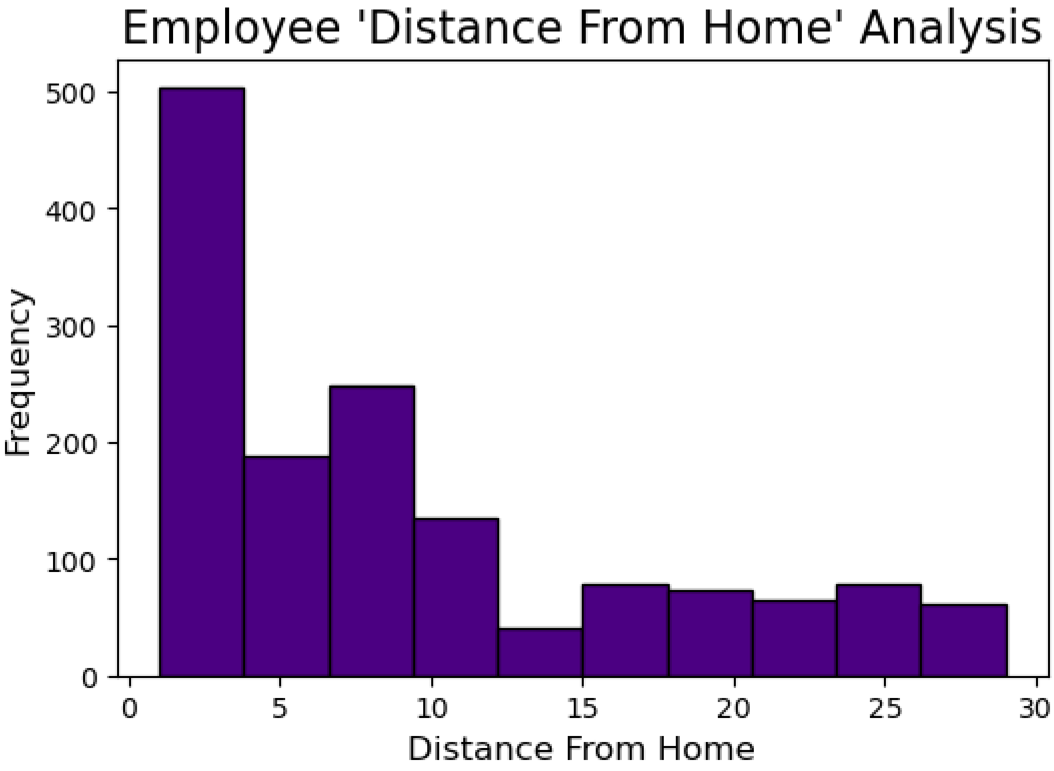
• Mean ‘Distance from Home’ of employees in the data sample is 9.19.

In [77]:
```
attr.DistanceFromHome.describe()
```

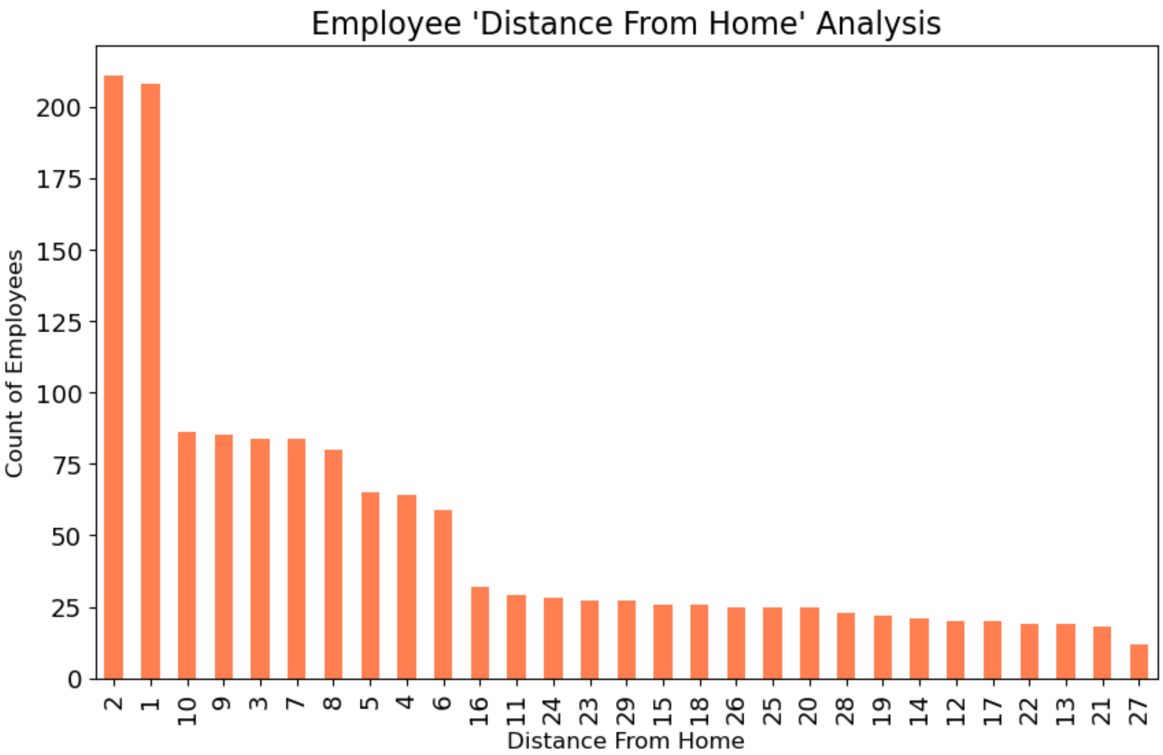
Out[77]:
```
count    1470.000000
mean        9.192517
std         8.106864
min         1.000000
25%         2.000000
50%         7.000000
75%        14.000000
max        29.000000
Name: DistanceFromHome, dtype: float64
```

In [78]:

```
attr.hist(column="DistanceFromHome",
          grid=False,
          figsize=(6,4),
          color="indigo",
          edgecolor="black",
          bins=10)

plt.xlabel("Distance From Home",fontsize=12)
plt.ylabel("Frequency",fontsize=12)
plt.title(" Employee 'Distance From Home' Analysis ",fontsize=16)

plt.show()
```

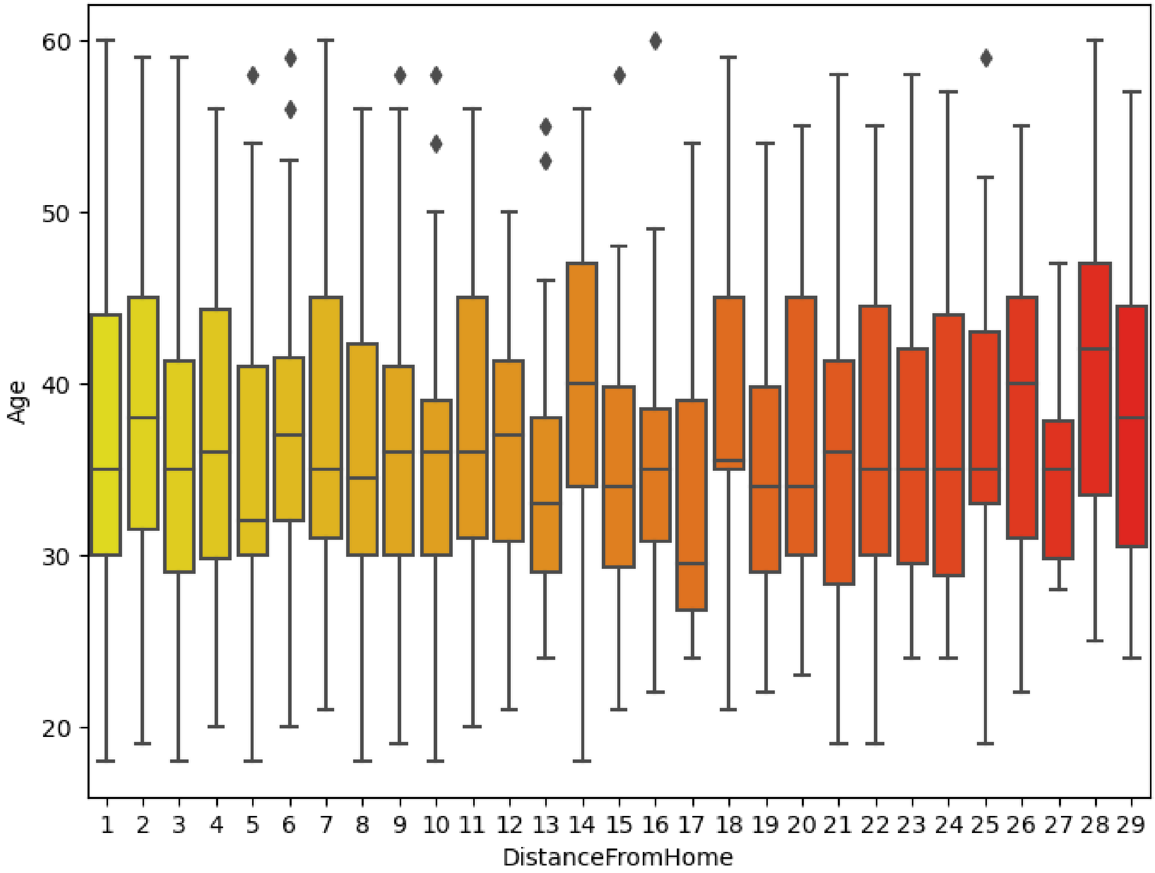


In [79]:
```
attr['DistanceFromHome'].value_counts().plot(kind='bar', figsize=(10,6), colo

plt.xlabel("Distance From Home", fontsize=12)
plt.ylabel("Count of Employees", fontsize=12)
plt.title(" Employee 'Distance From Home' Analysis ", fontsize=16)
plt.show()
```



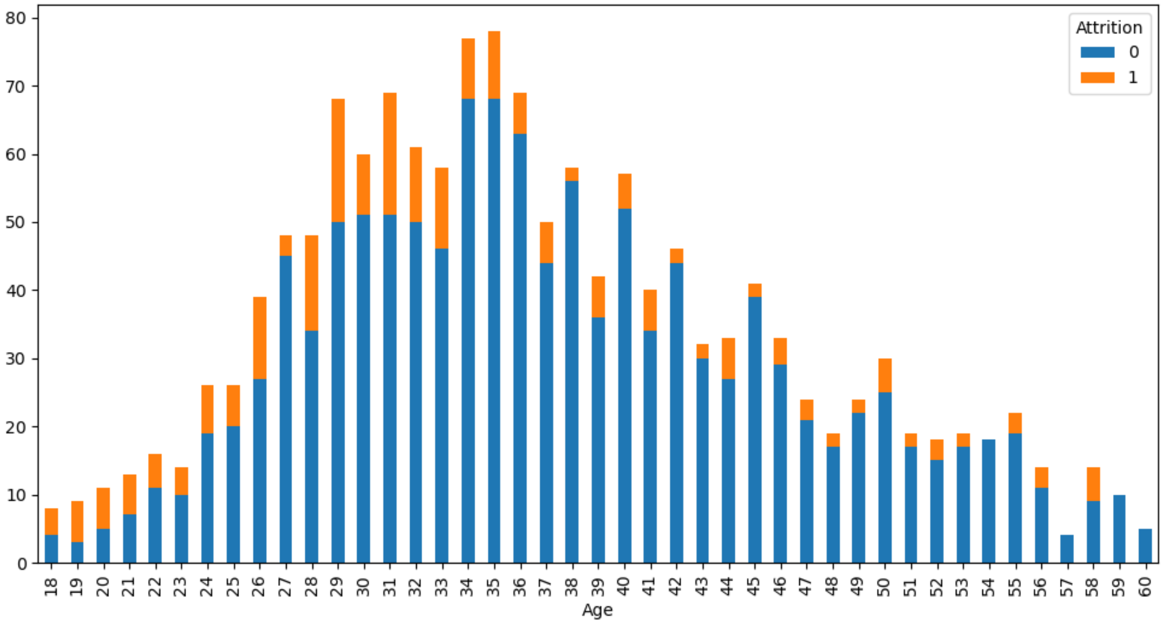
This graph shows most of the employees prefer to live colser to their home.

In [80]:
```
sns.boxplot(x = "DistanceFromHome",y= "Age",data = attr, palette = "autumn_r"
plt.gcf().set_size_inches(8, 6)
```



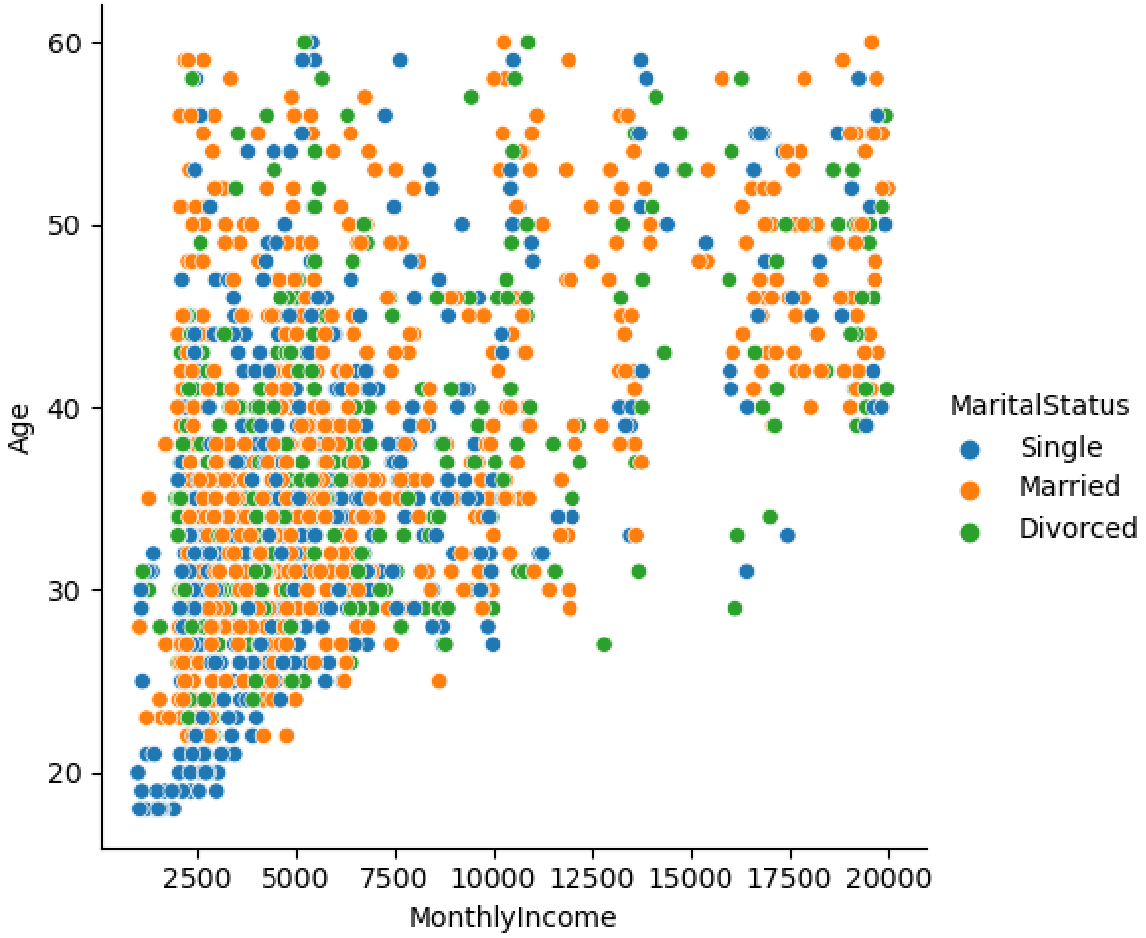
This box plot shows DistanceFromHome with respect to age.It does not show any difference so it is normally distributed.

```
In [81]: attr.groupby(['Age','Attrition'])['DistanceFromHome'].size().unstack().plot(k
         plt.show()
```



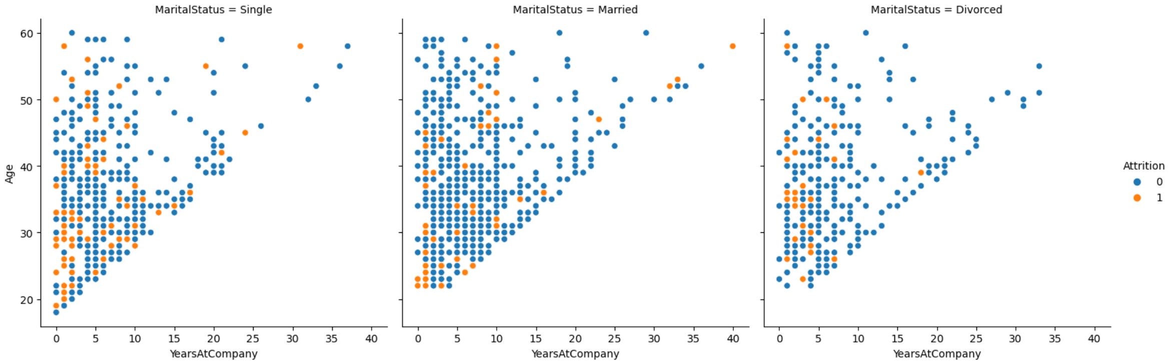
Relation plot between Monthly Income and Age by Marital status

In [82]:
```
sns.relplot(x="MonthlyIncome",y="Age",hue="MaritalStatus",data=attr)
plt.show()                                               ##
```

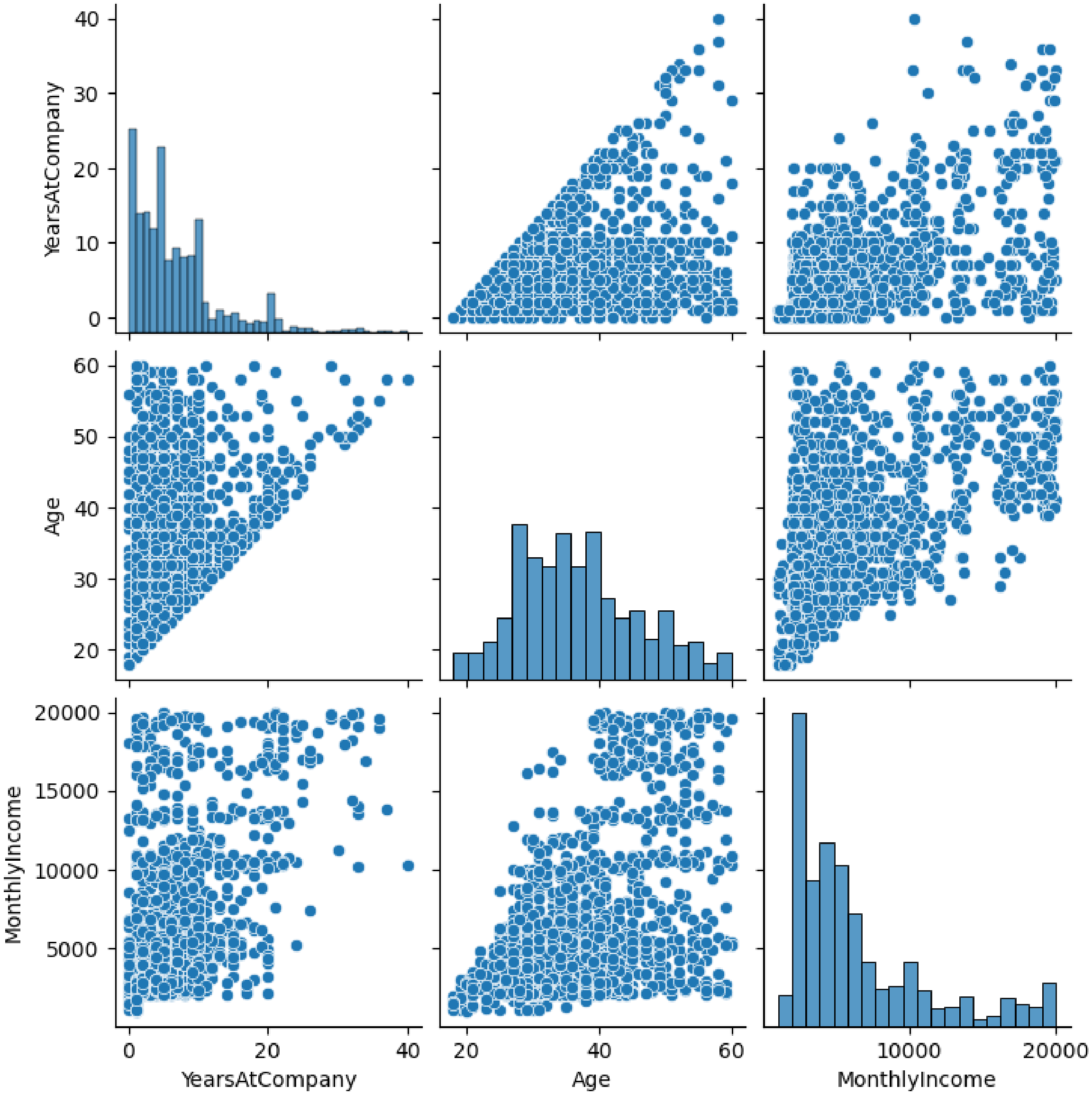


Relation between employee's YearsAtCompany and their age with attrition

In [83]:
```
sns.relplot(x="YearsAtCompany",y="Age",hue="Attrition",col="MaritalStatus",da
plt.show()
```



In [84]:
```
sns.pairplot(attr, vars=["YearsAtCompany","Age","MonthlyIncome"])
plt.show()
```



Here pair plot pairwise relationships in attrition dataset so that here all variables in attr will be plotted against each other variable in the dataset.

Corelation

In [85]:

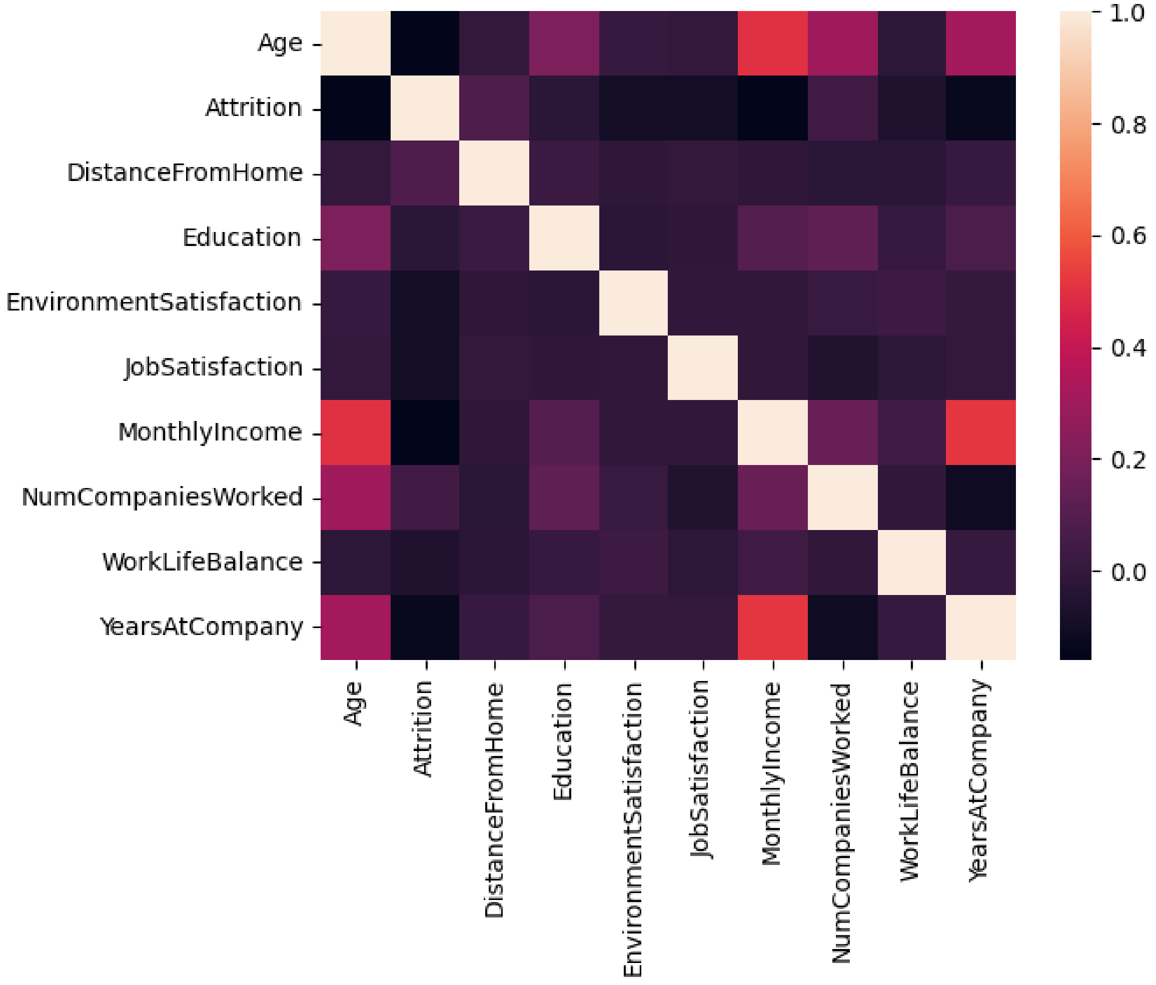
```
attr.corr()
```

Out[85]:

	Age	Attrition	DistanceFromHome	Education	EnvironmentSatisf
Age	1.000000	-0.159205	-0.001686	0.208034	0.0
Attrition	-0.159205	1.000000	0.077924	-0.031373	-0.1
DistanceFromHome	-0.001686	0.077924	1.000000	0.021042	-0.0
Education	0.208034	-0.031373	0.021042	1.000000	-0.0
EnvironmentSatisfaction	0.010146	-0.103369	-0.016075	-0.027128	1.0
JobSatisfaction	-0.004892	-0.103481	-0.003669	-0.011296	-0.0
MonthlyIncome	0.497855	-0.159840	-0.017014	0.094961	-0.0
NumCompaniesWorked	0.299635	0.043494	-0.029251	0.126317	0.0
WorkLifeBalance	-0.021490	-0.063939	-0.026556	0.009819	0.0
YearsAtCompany	0.311309	-0.134392	0.009508	0.069114	0.0

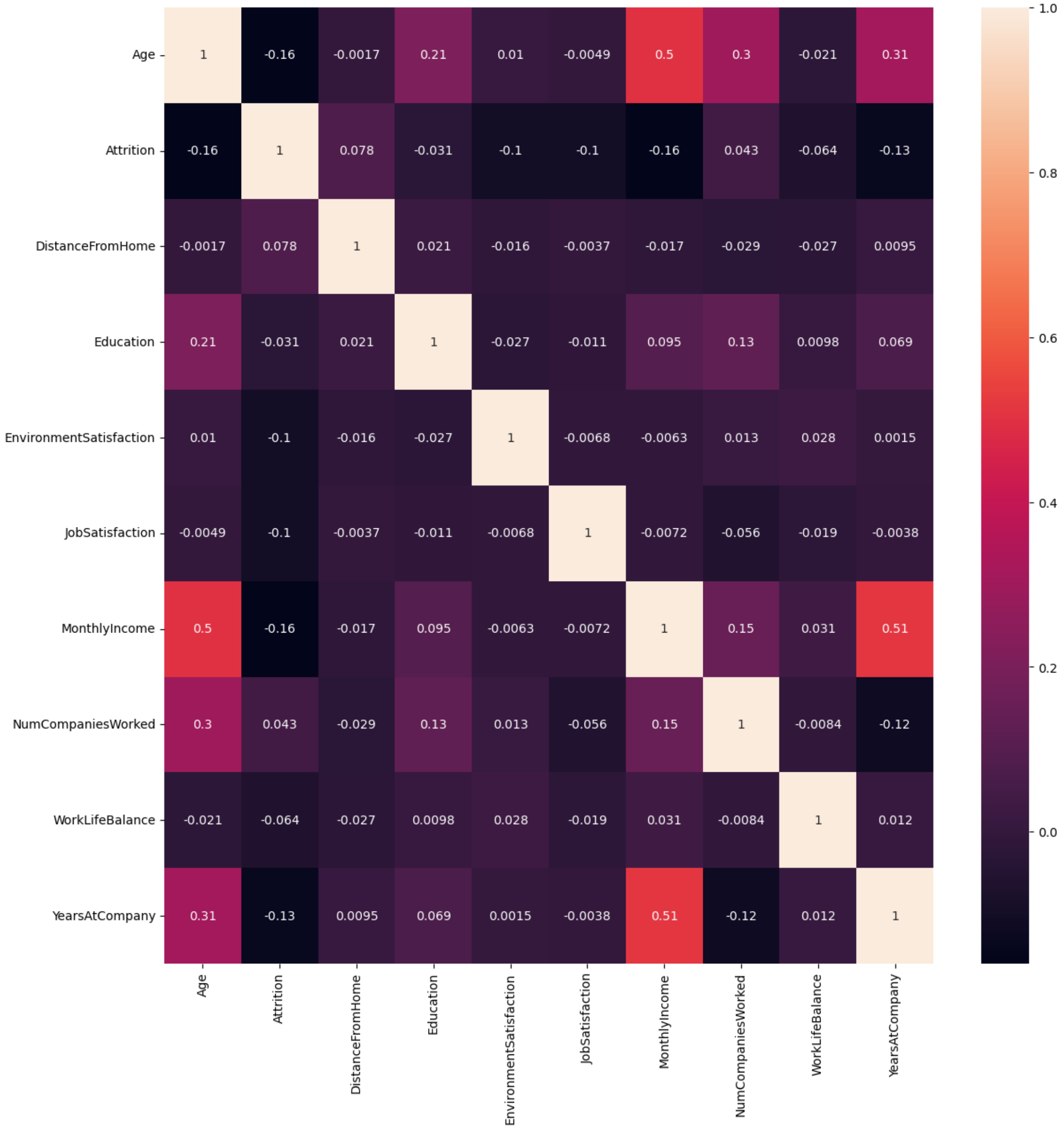
In [86]:

```
sns.heatmap(attr.corr())               ###heatmap
plt.show()
```



In [87]:

```
plt.figure(figsize = (14,14))
sns.heatmap(attr.corr(), annot=True)                ###heatmap
plt.show()
```



Variables such as TotalWorkingYears, YearsAtCompany, YearsInCurrentRoleare highly corelated to each other.

This graph shows Pearson’s correlation values, and there is a presence of high correlations values among different sets of variables such as Monthly Income, Total Working Hours and many more.

'Years at Company' and 'Monthly Imcome': 0.51 -------> More experiece more salary

'Age' and 'Monthly Income': 0.49 --------> old employees have more experience

'Age' and 'Years at Company' : 0.31

'Age' and 'Number of Companies Worked': 0.29

'Work Life Balance', 'Job Satisfaction' and 'Environment Satisfaction' do not have strong correlation with other features

pandas_profiling

```
In [88]: from pandas_profiling import ProfileReport
```

```
In [89]: profiling = ProfileReport(attr)
In [90]: profiling
```

```
Summarize dataset:   0%|          | 0/5 [00:00<?, ?it/s]

Generate report structure:   0%|          | 0/1 [00:00<?, ?it/s]

Render HTML:   0%|          | 0/1 [00:00<?, ?it/s]
```

Overview

Dataset statistics

Number of variables	13
Number of observations	1470
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	149.4 KiB
Average record size in memory	104.1 B

Variable types

Numeric	5
Categorical	8

Alerts

`Department` is highly overall correlated with `EducationField`	**High correlation**
`EducationField` is highly overall correlated with `Department`	**High correlation**

```
Out[90]:
```

```
In [91]: profiling.to_file("attr.html")
```

```
Export report to file:   0%|          | 0/1 [00:00<?, ?it/s]
```

It will save profiling data in another file named attr.html

Findings after performing Analysis :-

Based on Age:

Range of age is between 18 years to 60 years: This company has employees across the spectrum from recent graduates to the retirement age. Since the maximum age bracket with highest number of employees are between 30 -36, we can safely assume that the company prefers employees who have considerable work experience.

Based on Department:

Since 65% of the employees are in Research and development, we can assume that the company activities include innovation and introducing new products and services. Ostensibly, the goal of the company would be to take new products and services to market and add to the company's bottom line.

Based on Education Field:

Since 41% and 31% of the employees are from the Life sciences and medical backgrounds respectively, we can say that the company is majorly based out of bio-technology and R&D. This further strengthens our earlier propositions.

Based on Environment Satisfaction:

Since more than 60% of the employees have voted for the top 2 categories in environment satisfaction, we can say that the majority of employees feel comfortable working in this organization. However, 40% of the employees seem to need improvement in their working environment.

Based on Job Satisfaction:

Only 30% of employees in the organization have given the highest rating for job satisfaction. This shows that there is still a lot of room for improvement.

Based on Marital Status:

As 45% of the employees in the organization are married, the company appears to give a stable and well-paying job profile to the employees. Marital status is also an important indicator of organisational commitment. Generally, married people are more committed to their organisation than unmarried people.

Based on Monthly Income:

It is negatively correlated to attrition. Lesser the Monthly Income, higher the attrition rate. Employees with salary around 2500 are more prone to leave the company.

Based on WorkLifeBalance :

60% of employees have given second best rating. This could again hint at a 'neutral response bias' and it involves a certain subjectivity of interpreting a correct work-life balance.

Based on Years At Company:

The dataset reveals that a majority of the employees (75%) have a work experience of less than 9 years at the company. There appears to be a tapering down of career progression above 10 years. If the company can positively tackle this, it can greatly benefit the organisation.

Based on Distance From Home:

A large number of employees are residing close to the company (50% of the employees are at a distance of 7km). There is a possibility that accomodation facilities are provided by the company, since the maximum count of employees are at a distance of 2 km.

To reduce the attrition rate I would recommend:

Offer support: Provide work-life balance programs and flexible work arrangements that help employees manage their workload and reduce overtime. Offer mentorship and coaching programs that support employees in their current roles and help them develop the skills required for future roles.

Encourage career growth: Provide career advancement opportunities, training programs, and mentorship to support employee progression to higher job levels.

Offer competitive compensation: Offer competitive salaries and benefits that align with the market standards and recognize and reward long-serving employees for their commitment to the organization.

Foster a positive work environment: Provide a positive and inclusive work environment that encourages employee engagement and job satisfaction.

Gather employee feedback: Conduct regular employee engagement surveys to understand the underlying reasons for employee turnover and take corrective actions accordingly.

Project By Shalinee Kumari

In [ ]: