종합실습3 이변량분석(y-범주) : 직원 이직 분석

- 직원 이직 분석
  - 회사에서 최근 1~2년 사이 이직률이 상승하였습니다.
  - 여러분은, 직원들이 이직하는데 중요한 요인이 무엇인지 데이터를 기반으로 한 분석을 의뢰 받았습니다.

1.환경준비

- 라이브러리 불러오기

In [1]:
```
import pandas as pd
import numpy as np

import matplotlib.pyplot as plt
import seaborn as sns
from statsmodels.graphics.mosaicplot import mosaic      #mosaic plot!

import scipy.stats as spst
```

- 데이터 불러오기 : 다음의 예제 데이터를 사용합니다.

In [2]:
```
# 직원 이직 데이터
path = 'https://raw.githubusercontent.com/DA4BAM/dataset/master/Attrition_simple3.csv'
data = pd.read_csv(path)
data.head()
```

Out[2]:

	Attrition	Age	DistanceFromHome	EmployeeNumber	Gender	JobSatisfaction	MaritalStatus	Mor
0	0	27	2	1898	Female	3	Single	
1	0	27	9	1965	Male	4	Single	
2	0	44	2	1703	Female	4	Married	
3	0	42	2	1231	Male	1	Married	
4	0	32	1	2016	Female	4	Married	

- 변수설명
  - Attrition : 이직여부, Yes , No (Target)
  - Age : 나이
  - DistanceFromHome : 집-직장 거리(마일)
  - EmployeeNumber : 사번
  - Gender : 성별(Male, Female)
  - JobSatisfaction : 직무 만족도, 다음시트 참조
  - MaritalStatus : 결혼상태(Married, Single, Divorced)
  - MonthlyIncome : 월급(달러)
  - OverTime : 야근여부
  - PercentSalaryHike : 전년대비 급여인상율(%)
  - TotalWorkingYears : 총 근무 연수

2.범주-->범주

In [3]:
```
target = 'Attrition'
```

(1) Gender --> Attrition

In [4]:
```
feature = 'Gender'
```

- 교차표

In [5]:
```
# 두 범주별 빈도수를 교차표로 만들어 봅시다.
pd.crosstab(data[target], data[feature])
```

Out[5]:

Gender	Female	Male
Attrition		
0	157	248
1	66	129

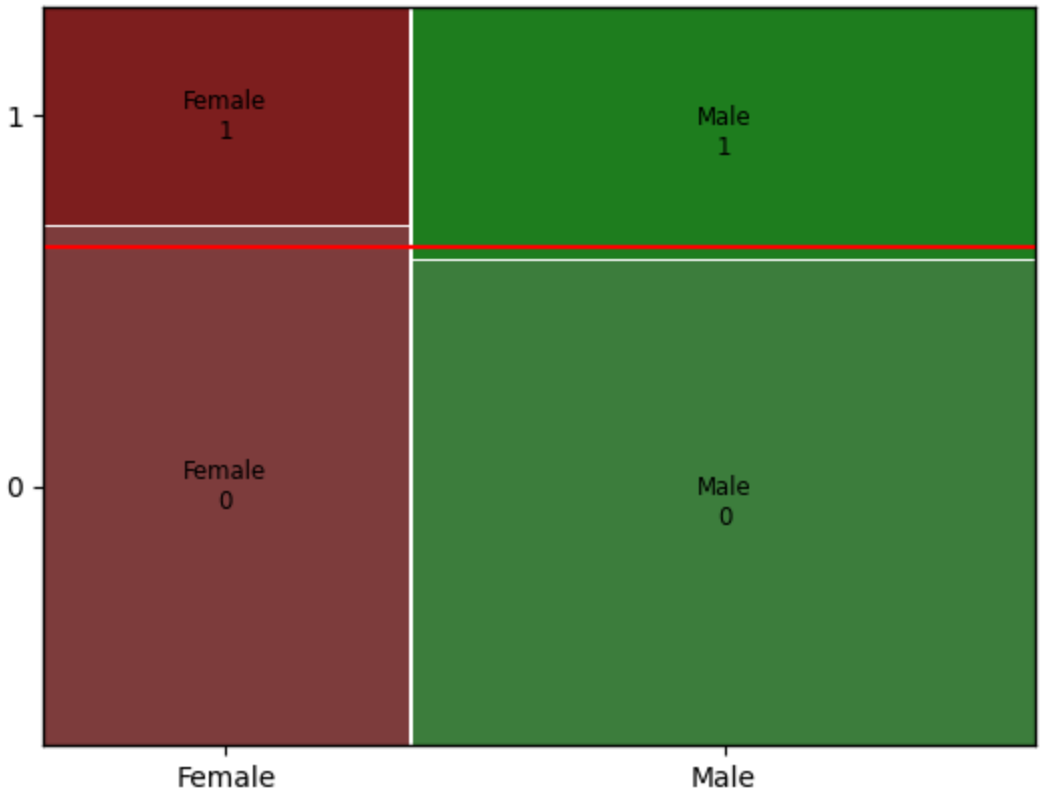
In [6]:
```
pd.crosstab(data[target], data[feature], normalize = 'columns')
```

Out[6]:

Gender	Female	Male
Attrition		
0	0.704036	0.657825
1	0.295964	0.342175

- 시각화

In [7]:
```
mosaic(data, [ feature,target])
plt.axhline(1- data[target].mean(), color = 'r')
plt.show()
```



- 수치화 : 카이제곱검정

In [8]:
```
# 먼저 집계
table = pd.crosstab(data[target], data[feature])
print('교차표\n', table)
print('-' * 100)

# 카이제곱검정
result = spst.chi2_contingency(table)
print('카이제곱통계량', result[0])
print('p-value', result[1])
print('자유도', result[2])
# print('기대빈도\n',result[3])
```

```
교차표
 Gender     Female  Male
Attrition
0             157   248
1              66   129
--------------------------------------------------------------------------------------------------
------
카이제곱통계량 1.1614318259891623
p-value 0.28116879016055174
자유도 1
```

- 파악된 내용을 기술해 봅시다.

In [ ]:
```
# 카이제곱검정으로는 관련이 없다고 나오나, 그래프로 볼때 약간 관련이 있다고 판단됨.
```

(2) JobSatisfaction --> Attrition

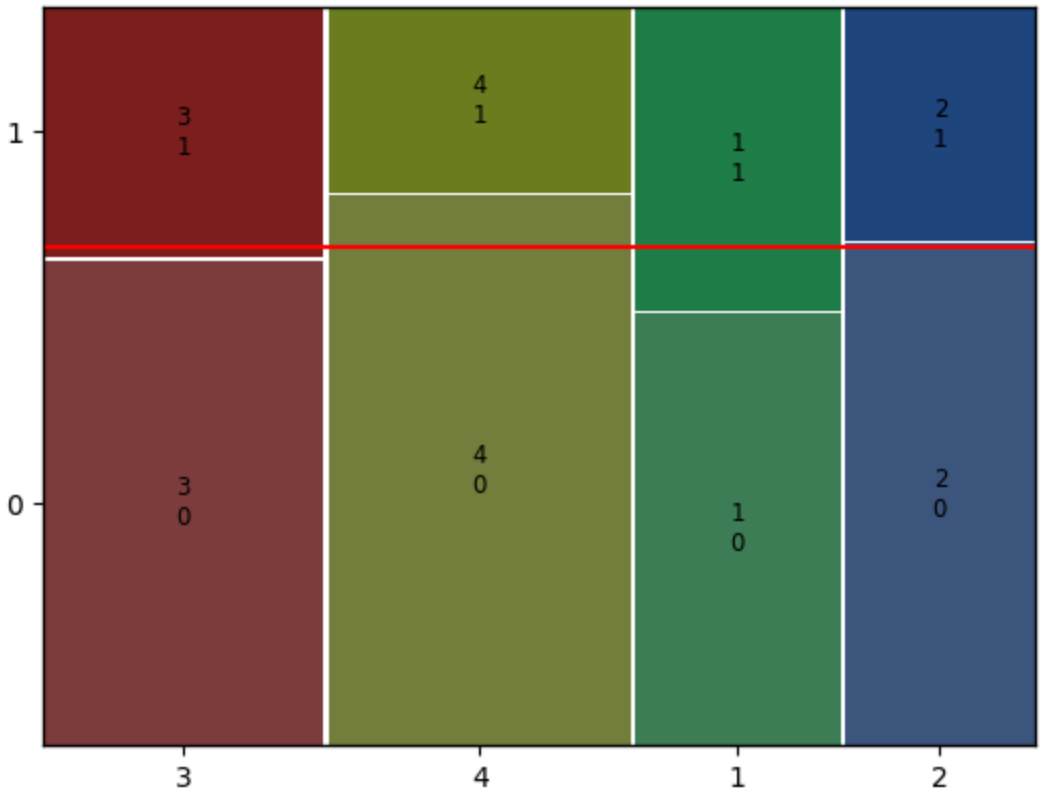
In [13]:
```
def edu_1(feature, target, data):
    # 두 범주별 빈도수를 교차표로 만들어 봅시다.
    print(pd.crosstab(data[target], data[feature]))
    print('-' * 100)
    print(pd.crosstab(data[target], data[feature], normalize = 'columns'))

    # 시각화
    mosaic(data, [ feature,target])
    plt.axhline(1- data[target].mean(), color = 'r')
    plt.show()

    # 먼저 집계
    table = pd.crosstab(data[target], data[feature])
    print('교차표\n', table)
    print('-' * 100)

    # 카이제곱검정
    result = spst.chi2_contingency(table)
    print('카이제곱통계량', result[0])
    print('p-value', result[1])
    print('자유도', result[2])
```

In [14]:
```
feature = 'JobSatisfaction' # 직무 만족도
```

- 교차표
- 시각화
- 수치화 : 카이제곱검정

In [15]:
```
edu_1(feature, target, data)
```

```
JobSatisfaction    1    2    3    4
Attrition
0                 74   79  114  138
1                 52   37   59   47
------------------------------------------------------------------------------------------------
------
JobSatisfaction         1         2         3         4
Attrition
0                0.587302  0.681034  0.65896  0.745946
1                0.412698  0.318966  0.34104  0.254054
```




```
교차표
 JobSatisfaction    1    2    3    4
Attrition
0                 74   79  114  138
1                 52   37   59   47
------------------------------------------------------------------------------------------------
------
카이제곱통계량 8.884191097554549
p-value 0.03087092125625072
자유도 3
```

- 파악된 내용을 기술해 봅시다.

- 그래프와 카에제곱을 보면 직무 만족도가 높을 수록 이직률이 낮은것으로 보임

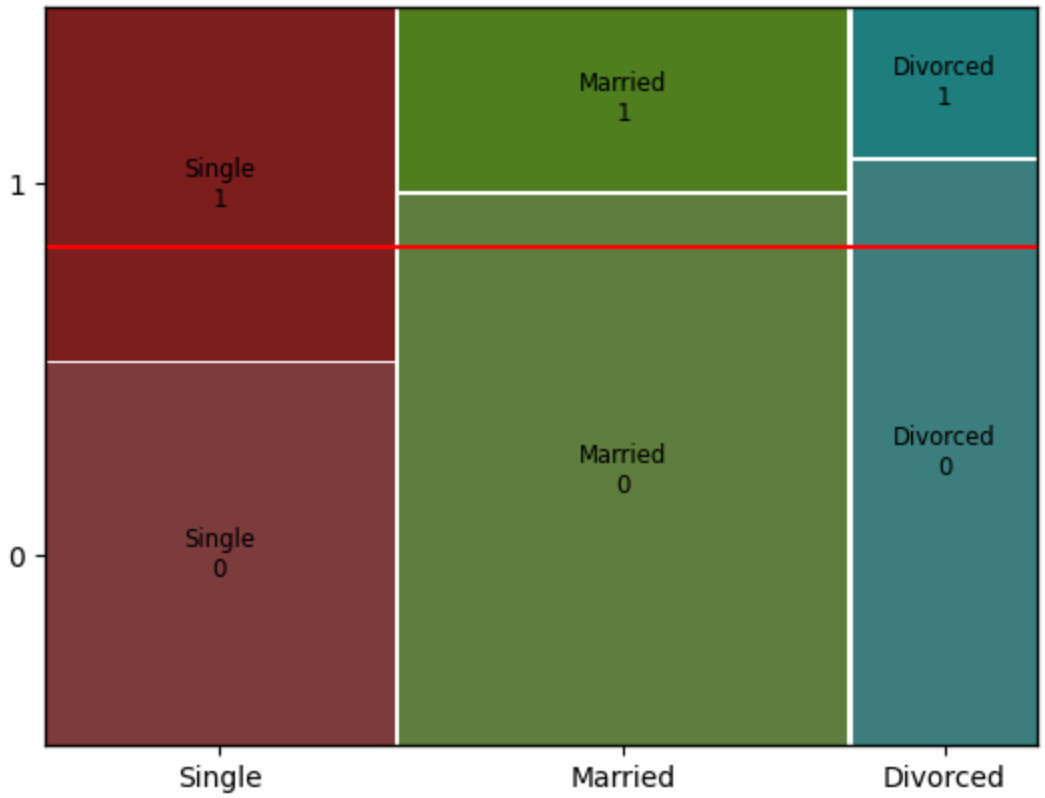
(3) MaritalStatus --> Attrition

In [19]:
```
feature = 'MaritalStatus' # 결혼상태
```

- 교차표
- 시각화
- 수치화 : 카이제곱검정

In [20]:
```
edu_1(feature, target, data)
```

```
MaritalStatus  Divorced  Married  Single
Attrition
0                    89      205     111
1                    23       69     103
------------------------------------------------------------------------------------------
------
MaritalStatus  Divorced   Married    Single
Attrition
0              0.794643  0.748175  0.518692
1              0.205357  0.251825  0.481308
```




```
교차표
 MaritalStatus  Divorced  Married  Single
Attrition
0                    89      205     111
1                    23       69     103
------------------------------------------------------------------------------------------
------
카이제곱통계량 37.829711907070525
p-value 6.100738829354226e-09
자유도 2
```

- 파악된 내용을 기술해 봅시다.

In [49]:
```
# 카이제곱, 그래프에 관련이 있어 보인다고 나옴
# 싱글 일수록 높고 이혼 한 사람 일수록 이직이 낮은 것으로 보임
```

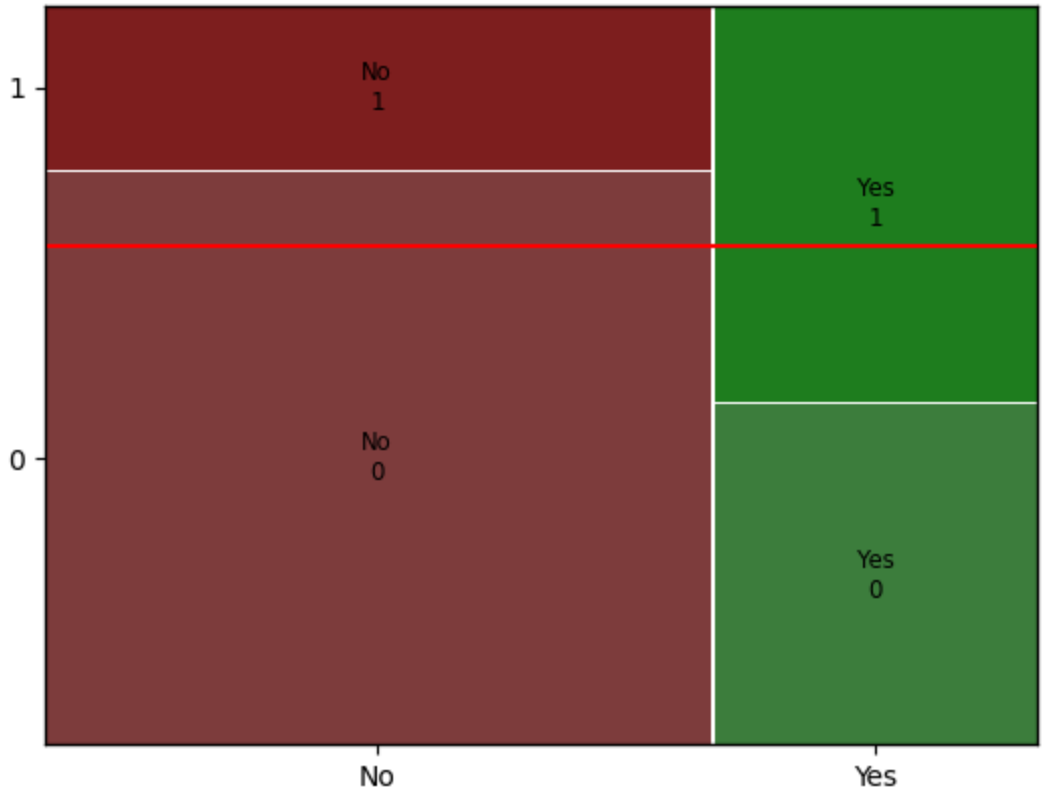
(4) OverTime --> Attrition

In [21]:
```
feature = 'OverTime' # 야근여부
```

- 교차표
- 시각화
- 수치화 : 카이제곱검정

In [22]:
```
edu_1(feature, target, data)
```

```
OverTime    No  Yes
Attrition
0          315   90
1           90  105
------------------------------------------------------------------------------------------------
------
OverTime         No       Yes
Attrition
0          0.777778  0.461538
1          0.222222  0.538462
```




```
교차표
 OverTime    No  Yes
Attrition
0          315   90
1           90  105
------------------------------------------------------------------------------------------------
------
카이제곱통계량 58.57149427899665
p-value 1.9603625783060702e-14
자유도 1
```

- 파악된 내용을 기술해 봅시다.

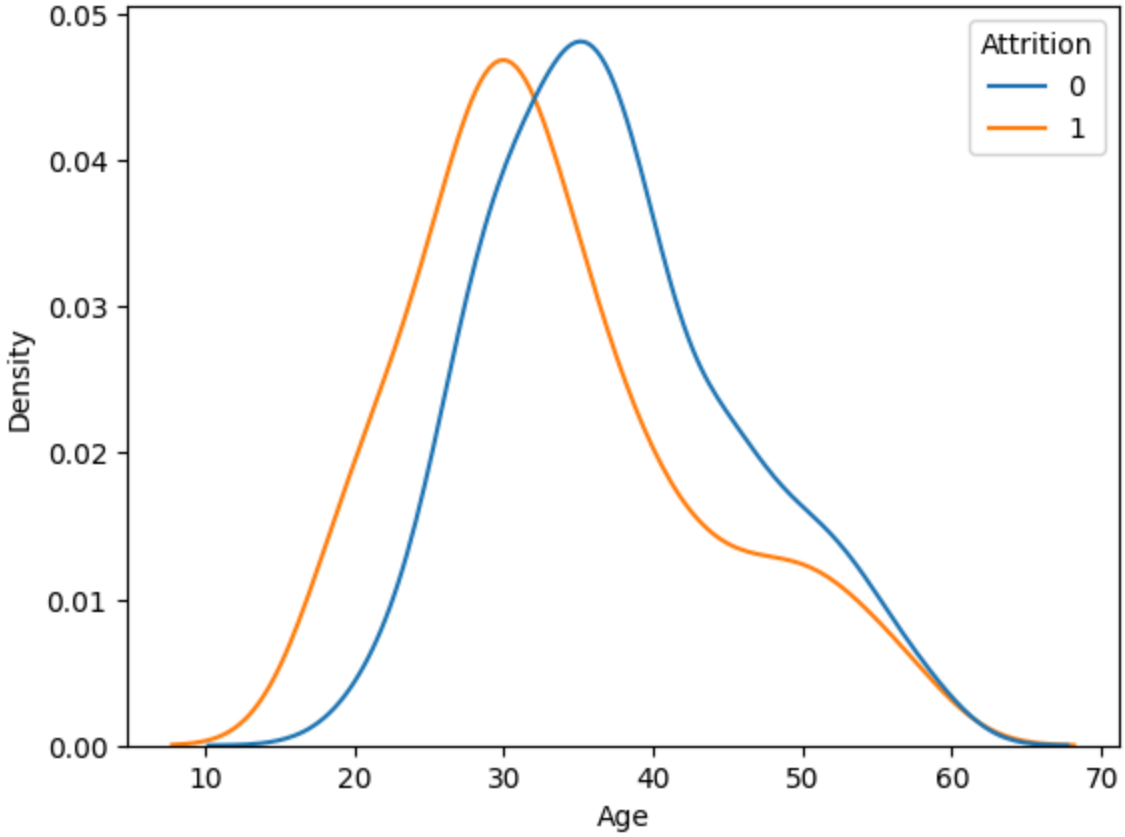
- 야근이 이직률에 관계가 있어 보임

3.숫자-->범주

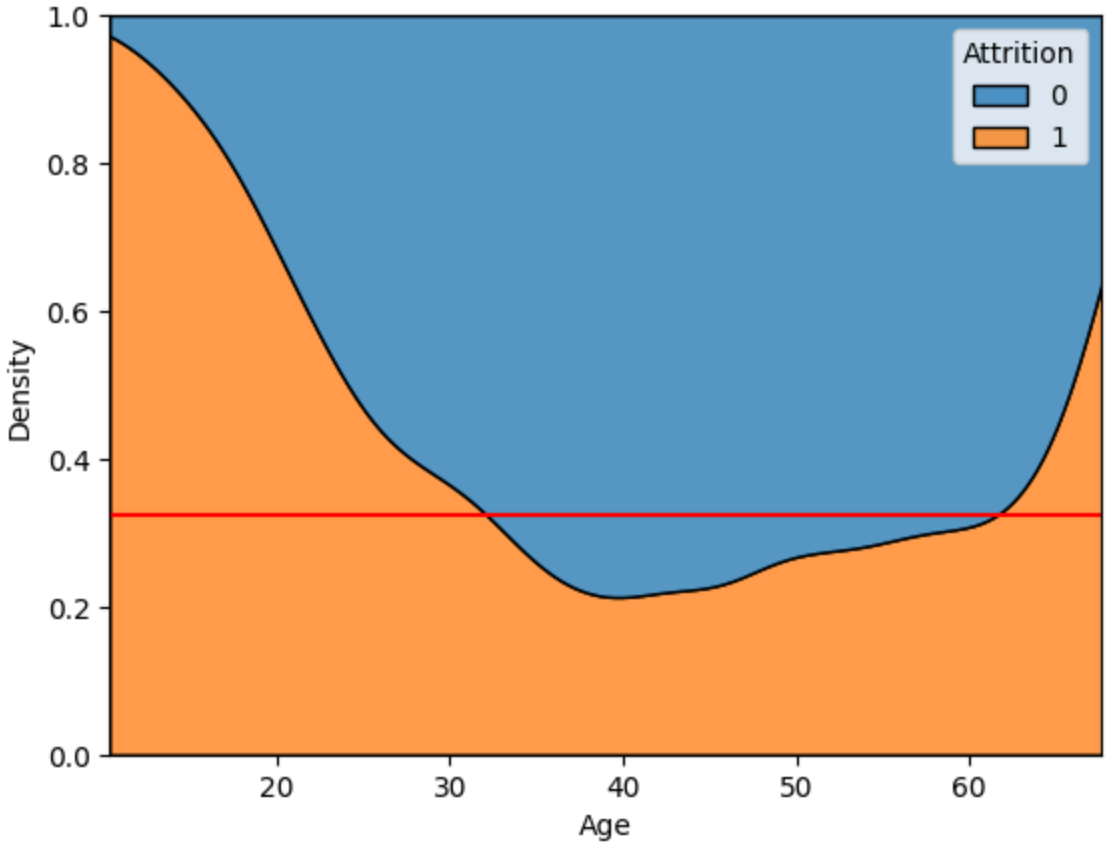
(1) Age --> Attrition

In [23]:
```
feature = 'Age'
```

In [24]:
```
sns.kdeplot(x= feature, data = data, hue = target,
            common_norm = False)
plt.show()
```



In [25]:
```
sns.kdeplot(x= feature, data = data, hue = target, multiple = 'fill')
plt.axhline(data[target].mean(), color = 'r')
plt.show()
```



- 파악된 내용을 기술해 봅시다.

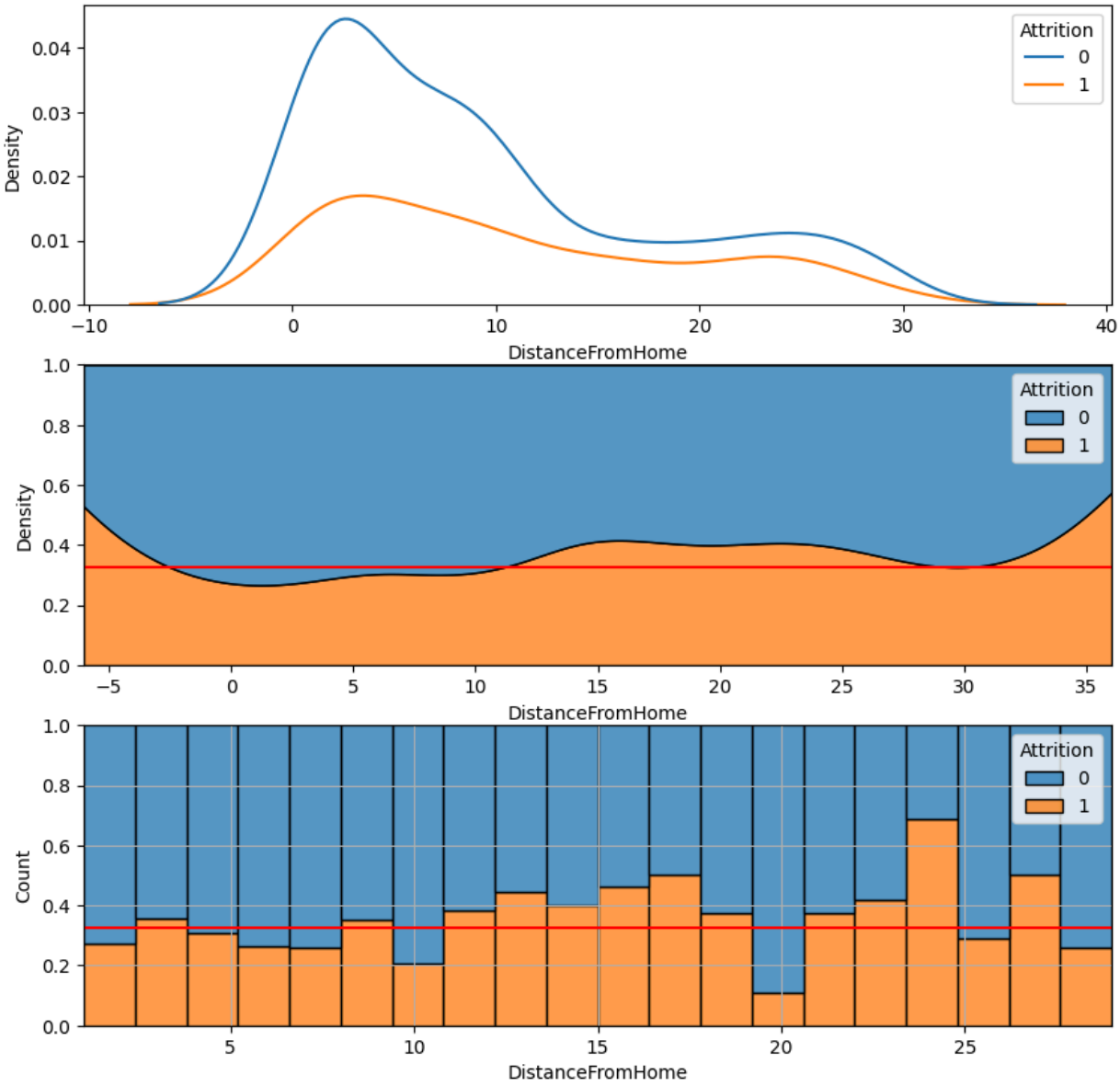
- 나이가 어릴 수록 이직률이 높고
- 30 ~ 50 대 사이가 낮다 아마 기혼자 일듯
- 정년을 앞둔 60대 이상 부터는 이직률이 낮다

In [59]:
```
def edu_2(feaure, target, data):
    plt.figure(figsize=(10, 10))
    plt.subplot(3, 1, 1)
    sns.kdeplot(x=feature, data=data, hue=target, common_norm=True)

    plt.subplot(3, 1, 2)
    sns.kdeplot(x=feature, data=data, hue=target, multiple='fill')
    plt.axhline(data[target].mean(), color='r')

    plt.subplot(3,1,3)
    sns.histplot(x = feature, data = data, bins = 20, hue = target, multiple = 'fill')
    plt.axhline(data[target].mean(), color = 'r')
    plt.xlim(data[feature].min(), data[feature].max())
    plt.grid()

    plt.show()
```

(2) DistanceFromHome --> Attrition

In [60]:
```
feature = 'DistanceFromHome' # 집-직장 거리(마일)
```

In [61]:
```
edu_2(feature, target, data)
```



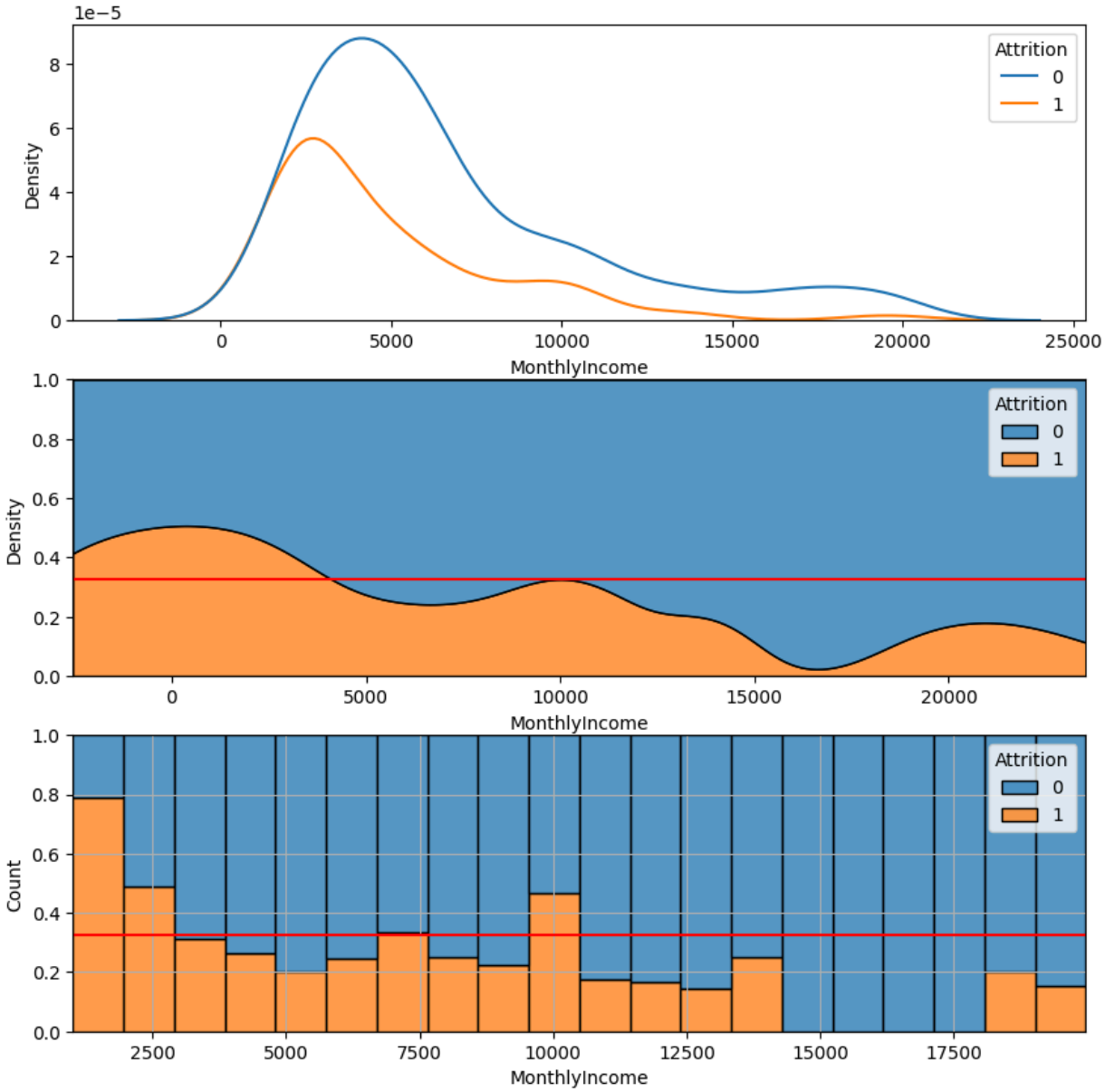
- 파악된 내용을 기술해 봅시다.

- 집과 직장과의 거리와 이직률은 약간 있어 보임

(3) MonthlyIncome --> Attrition

In [62]:
```
feature = 'MonthlyIncome' # 월급(달러)
```

In [63]:
```
edu_2(feature, target, data)
```



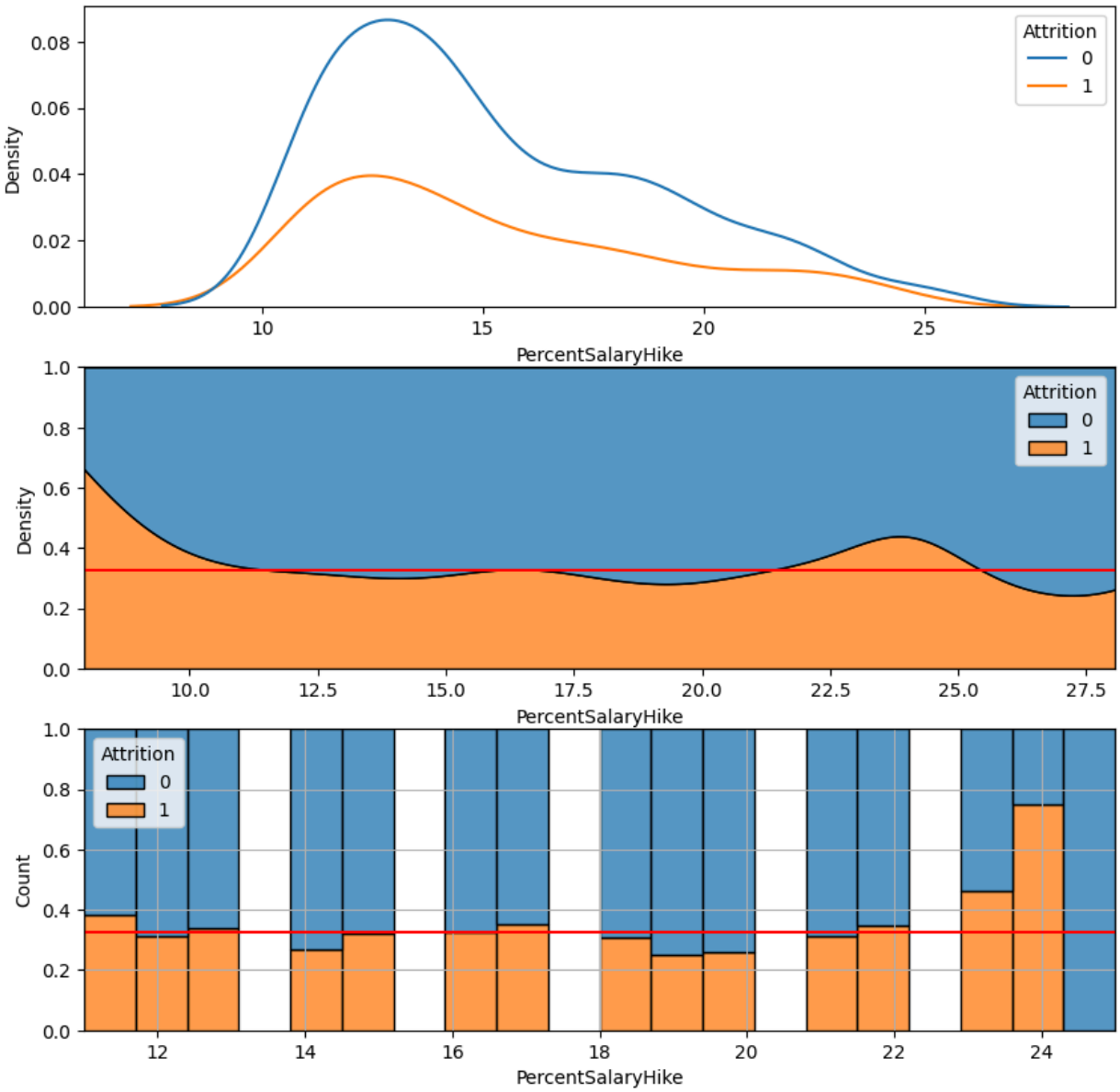
- 파악된 내용을 기술해 봅시다.

In [44]:
```
# 월급이 높아질수록  이직률이 낮아 지는 것이 보임
```

(4) PercentSalaryHike --> Attrition

In [64]:
```
feature = 'PercentSalaryHike' # 전년대비 급여인상율(%)
```

In [65]:
```
edu_2(feature, target, data)
```



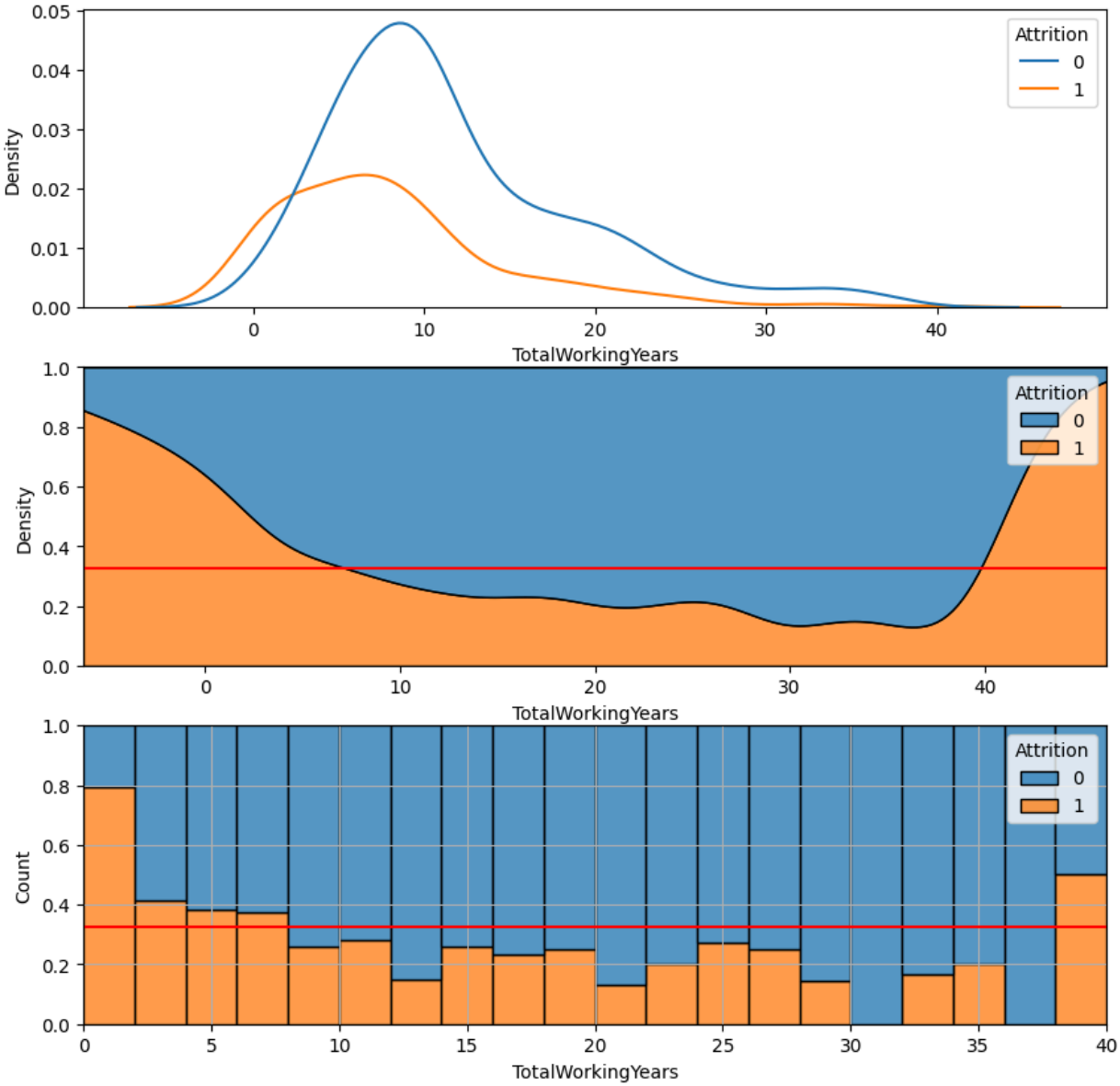
- 파악된 내용을 기술해 봅시다.

In [66]:
```
# 인상률이 낮을 때 이직류을 보이고
# 또 인상 폭이 커질 때 이직률이 올라가는 것을 보임 (책임감?, 월급이 원래 낮았나?)
```

(5) TotalWorkingYears --> Attrition

In [67]:
```
feature = 'TotalWorkingYears' # 총 근무 연수
```

In [68]:
```
edu_2(feature, target, data)
```



- 파악된 내용을 기술해 봅시다.

In [53]:
```
# 총 근무 연수가 적을 수록 이직률이 높고 연수가 많아질 수록 이직률이 높지만
# 36년 정도 부터는 이직률이 높아지는 것이 보임 (아마 정년이 다되어 퇴직이 많은 것인 가능성이 있
```

4.관계 정리하기

① 강한관계

In [55]:
```
# MaritalStatus
# OverTime
# Age
# MonthlyIncome
# TotalWorkingYears
```

② 중간관계

In [ ]:
```
#JobSatisfaction
```

③ 관계없음(약한 관계)

In [ ]:
```
# Gender
# DistanceFromHome
# PercentSalaryHike
```