이변량_범주 vs 범주

1.환경준비

(1) 라이브러리

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

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

import scipy.stats as spst
```

(2) 데이터 불러오기

- 다음의 예제 데이터를 사용합니다.

> 타이타닉 생존자

In [2]:
```
# 타이타닉 데이터
titanic = pd.read_csv('https://raw.githubusercontent.com/DA4BAM/dataset/master/titanic.1.csv')
titanic.head()
```

Out[2]:

	PassengerId	Survived	Pclass	Title	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarl
0	1	0	3	Mr	male	22.0	1	0	A/5 21171	7.2500	NaN	
1	2	1	1	Mrs	female	38.0	1	0	PC 17599	71.2833	C85	
2	3	1	3	Miss	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	
3	4	1	1	Mrs	female	35.0	1	0	113803	53.1000	C123	
4	5	0	3	Mr	male	35.0	0	0	373450	8.0500	NaN	

2.범주 --> 범주

(1) 교차표(pd.crosstab)

범주 vs 범주 를 비교하고 분석하기 위해서는 먼저 **교차표**를 만들어야 합니다.

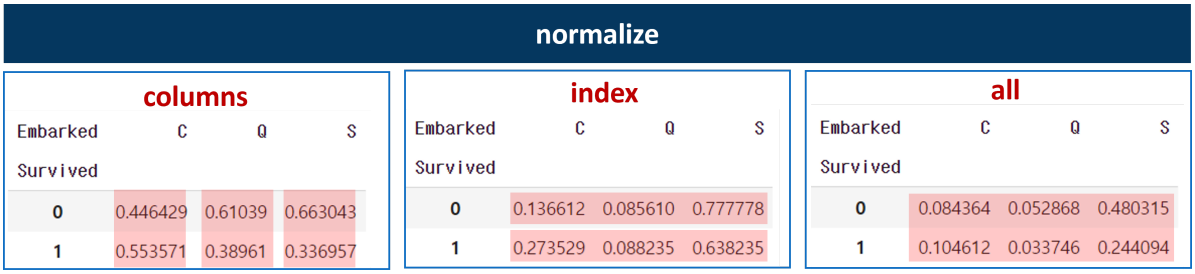
- pd.crosstab(행, 열)

In [3]:
```
# 두 범주별 빈도수를 교차표로 만들어 봅시다.
pd.crosstab(titanic['Survived'], titanic['Sex'])
```

Out[3]:

Sex	female	male
Survived		
0	81	468
1	233	109

- pd.crosstab(행, 열, normalize =)



In [4]:
```
pd.crosstab(titanic['Survived'], titanic['Sex'], normalize = 'columns')
```

Out[4]:

Sex	female	male
Survived		
0	0.257962	0.811092
1	0.742038	0.188908

In [5]:
```
pd.crosstab(titanic['Survived'], titanic['Sex'], normalize = 'index')
```

Out[5]:

Sex	female	male
Survived		
0	0.147541	0.852459
1	0.681287	0.318713

In [6]:
```
pd.crosstab(titanic['Survived'], titanic['Embarked'], normalize = 'all')
```

Out[6]:

Embarked	C	Q	S
Survived			
0	0.084175	0.05275	0.479237
1	0.104377	0.03367	0.245791

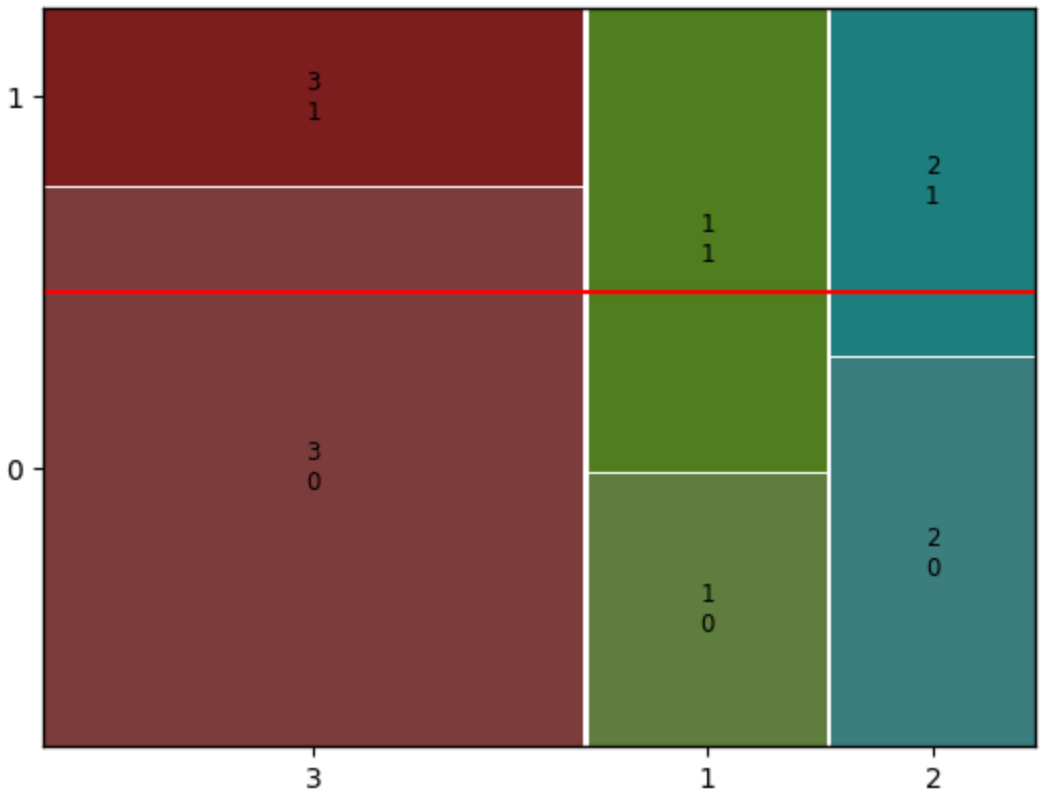
(2) 시각화

- mosaic
- (참조) 100% Stacked Bar

1) Pclass --> Survived

- mosaic plot
  - mosaic(dataframe, [feature, target])

In [7]:
```
# Pclass별 생존여부를 mosaic plot으로 그려 봅시다.
mosaic(titanic, [ 'Pclass','Survived'])
plt.axhline(1- titanic['Survived'].mean(), color = 'r')
plt.show()
```

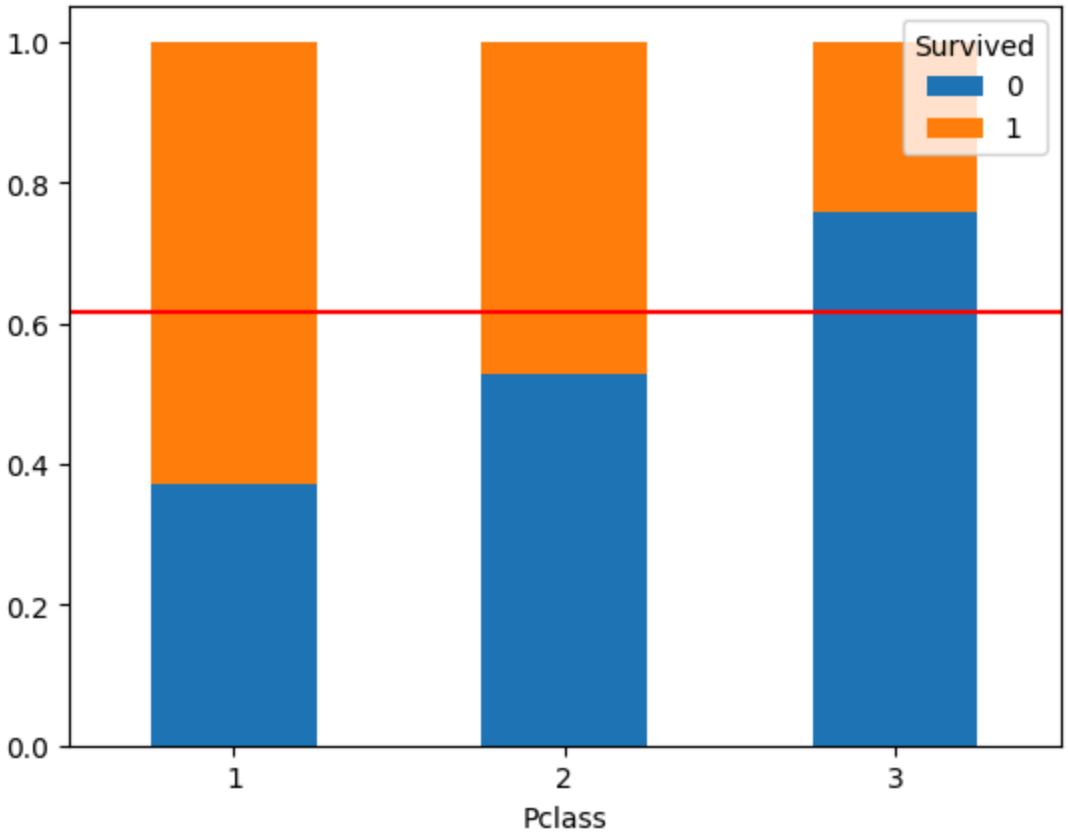


① X축 길이는 각 객실등급별 승객비율을 나타냅니다.
② 그 중 3등급 객실에 대해서 보면, y축의 길이는, 3등급 객실 승객 중에서 사망, 생존 비율을 의미합니다.

- 100% Stacked Bar
  - 먼저 crosstab으로 집계 : pd.crosstab(**feature**, **target**, normalize = **'index'**)
  - .plot.bar(stacked = true)
  - 전체 평균선 : plt.axhline()

In [8]:
```
# 비율만 보이고 이 순서대로 코드를 작성해야 함 ********************
temp = pd.crosstab(titanic['Pclass'], titanic['Survived'], normalize = 'index')
print(temp)
temp.plot.bar(stacked=True)
plt.axhline(1-titanic['Survived'].mean(), color = 'r')
plt.xticks(rotation=0)
plt.show()
```

```
Survived         0         1
Pclass
1         0.370370  0.629630
2         0.527174  0.472826
3         0.757637  0.242363
```



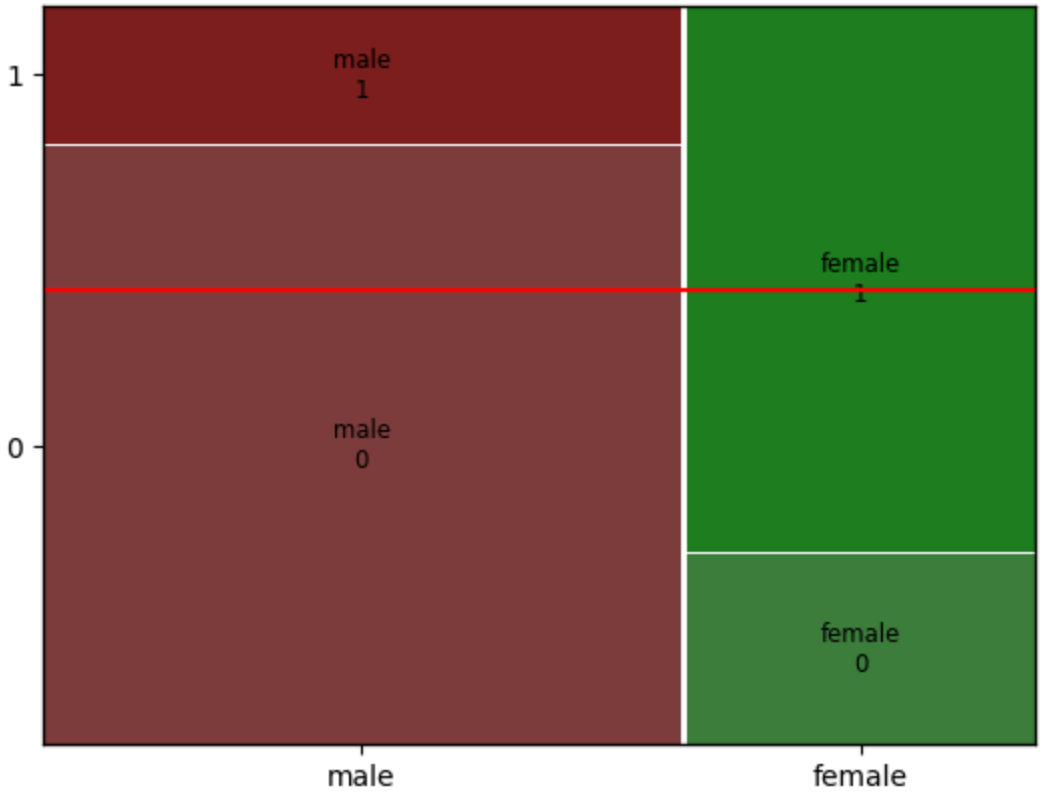
비율만 비교하므로 양에 대한 비교는 할 수 없다!

-연습문제-

아래 관계에 대해서 교차표와 시각화(mosaic)를 수행하고, feature와 target 간에 관계가 있는지 분석해 봅시다.

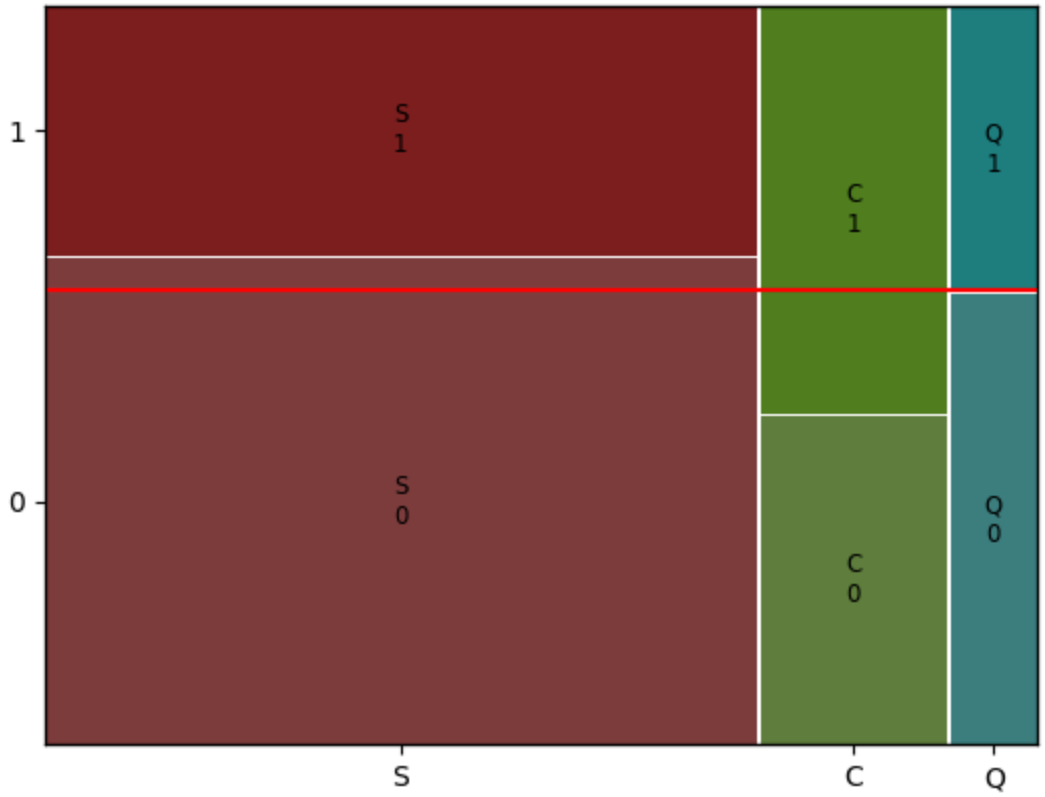
- [문1] Sex --> Survived

In [9]:
```
mosaic(titanic, ['Sex', 'Survived'])
plt.axhline(1- titanic['Survived'].mean(), color = 'r')
plt.show()
```



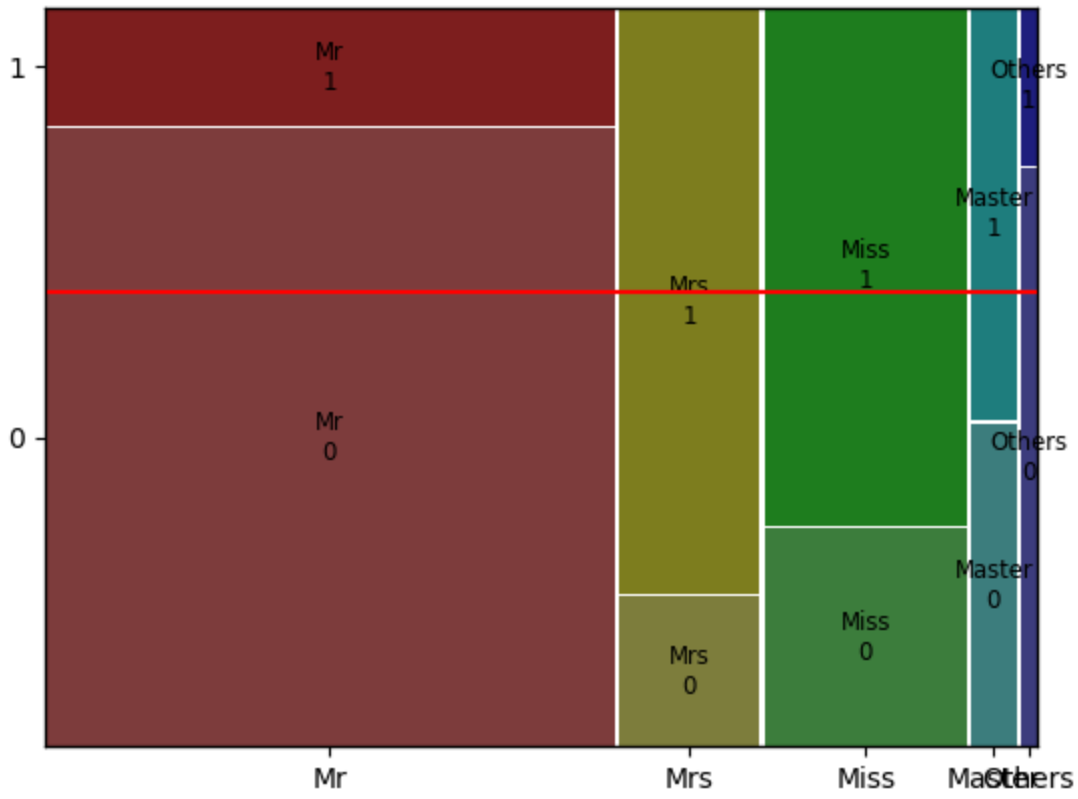
- [문2] Embarked --> Survived

In [10]:
```
mosaic(titanic, ['Embarked', 'Survived'])
plt.axhline(1- titanic['Survived'].mean(), color = 'r')
plt.show()
```



- [문3] Title --> Survived

In [11]:
```
mosaic(titanic, ['Title', 'Survived'])
plt.axhline(1- titanic['Survived'].mean(), color = 'r')
plt.show()
```



(3) 수치화 : 카이제곱검정

- 카이제곱검정 : 범주형 변수들 사이에 어떤 관계가 있는지, 수치화 하는 방법

$$x^2 = \sum \frac{(\text{관측빈도} - \text{기대빈도})^2}{\text{기대빈도}} = \frac{(+5)^2}{5} + \frac{(-5)^2}{20} + \frac{(-5)^2}{15} + \frac{(+5)^2}{60} = 8.33$$

- 카이 제곱 통계량은
  - 클수록 기대빈도로부터 실제 값에 차이가 크다는 의미.
  - 계산식으로 볼 때, 범주의 수가 늘어날 수록 값은 커지게 되어 있음.
  - 보통, 자유도의 2~3배 보다 크면, 차이가 있다고 본다.
- 범주형 변수의 자유도 : 범주의 수 - 1
- 카이제곱검정에서는
  - x 변수의 자유도 × y 변수의 자유도
  - 예 : Pclass --> Survived
    - Pclass : 범주가 3개, Survived : 2개
    - (3-1) * (2-1) = 2
    - 그러므로, 2의 2 ~ 3배인 4 ~ 6 보다 카이제곱 통계량이 크면, 차이가 있다고 볼수 있음.
- 타이타닉 데이터에서 객실등급과 생존여부 간의 카이제곱 검정을 수행해 봅시다.

In [12]:
```
pd.crosstab(titanic['Survived'], titanic['Pclass'])
```

Out[12]:

Pclass	1	2	3
Survived			
0	80	97	372
1	136	87	119

In [13]:
```
# 1) 먼저 교차표 집계- normalize 하면 안 됨
table = pd.crosstab(titanic['Survived'], titanic['Pclass'])
print(table)
print('-' * 50)

# 2) 카이제곱검정
spst.chi2_contingency(table)
```

```
Pclass      1   2    3
Survived
0          80  97  372
1         136  87  119
--------------------------------------------------
```

Out[13]:
```
Chi2ContingencyResult(statistic=102.88898875696056, pvalue=4.549251711298793e-23, dof=2, expec
ted_freq=array([[133.09090909, 113.37373737, 302.53535354],
       [ 82.90909091,  70.62626263, 188.46464646]]))
```

다음의 관계에 대해 수치화 해 봅시다.

- [문1] Sex --> Survived

In [19]:
```
# 1) 먼저 교차표 집계- normalize 하면 안 됨
table = pd.crosstab(titanic['Survived'], titanic['Sex'])
print(table)
print('-' * 50)

# 2) 카이제곱검정
spst.chi2_contingency(table)
```

```
Sex       female  male
Survived
0             81   468
1            233   109
--------------------------------------------------
```

Out[19]:
```
Chi2ContingencyResult(statistic=260.71702016732104, pvalue=1.1973570627755645e-58, dof=1, expe
cted_freq=array([[193.47474747, 355.52525253],
       [120.52525253, 221.47474747]]))
```

- [문2] Embarked --> Survived

In [23]:
```
# 1) 먼저 교차표 집계- normalize 하면 안 됨
table = pd.crosstab(titanic['Embarked'], titanic['Survived'])
print(table)
print('-' * 50)

# 2) 카이제곱검정
k_statistic, pvalue, dof, expected_freq = spst.chi2_continLgency(table)
print(f'카이제곱 통계량 : {k_statistic}')
print(f'P_value : {pvalue}')
print(f'자유도도 : {dof}')        # 적절한 자유도를 가진 모델을 선택하는 것이 중요
print(f'기대빈도 : {expected_freq}') # 기대 빈도가 높을수록, 관측된 데이터와 기대되는 데이터가
```

```
Survived    0    1
Embarked
C          75   93
Q          47   30
S         427  219
--------------------------------------------------
카이제곱 통계량 : 25.964452881874784
P_value : 2.3008626481449577e-06
자유도도 : 2
기대빈도 : [[103.51515152  64.48484848]
 [ 47.44444444  29.55555556]
 [398.04040404 247.95959596]]
```

- [문3] Title --> Survived

In [25]:
```
# 1) 먼저 교차표 집계- normalize 하면 안 됨
table = pd.crosstab(titanic['Title'], titanic['Survived'])
print(table)
print('-' * 50)

# 2) 카이제곱검정
k_statistic, pvalue, dof, expected_freq = spst.chi2_contingency(table)
print(f'카이제곱 통계량 : {k_statistic}')
print(f'P_value : {pvalue}')
print(f'자유도도 : {dof}')
print(f'기대빈도 : {expected_freq}')
```

```
Survived    0    1
Title
Master     18   23
Miss       55  130
Mr        439   84
Mrs        26  102
Others     11    3
--------------------------------------------------
카이제곱 통계량 : 289.1953165452417
P_value : 2.318405007221846e-61
자유도도 : 4
기대빈도 : [[ 25.26262626  15.73737374]
 [113.98989899  71.01010101]
 [322.25252525 200.74747475]
 [ 78.86868687  49.13131313]
 [  8.62626263   5.37373737]]
```

3.복습문제

- 항공기 탑승객의 만족도와 관련 있는 요인을 분석해 봅시다.
- 약 5천명의 탑승객에 대해서 탑승 경험을 바탕으로 데이터셋이 구성되어 있습니다.
  - Target
    - 탑승 만족도(satisfaction) : 만족 = 1, 불만 = 0
  - Feature
    - 성별, 나이, 여행타입, 객실등급, 비행거리, 객실등급, 비행거리, 식음료 만족도, 출발 지연시간

In [26]:
```
path = 'https://raw.githubusercontent.com/DA4BAM/dataset/master/Air_Satisfaction.csv'
cols = ['Gender', 'Age', 'Type of Travel', 'Class', 'Flight Distance', 'Food and drink',
        'Departure Delay in Minutes', 'satisfaction']
data = pd.read_csv(path, usecols = cols)
data['satisfaction'] = np.where(data['satisfaction'] =='satisfied', 1 ,0)
data.head()
```

Out[26]:

	Gender	Age	Type of Travel	Class	Flight Distance	Food and drink	Departure Delay in Minutes	satisfaction
0	Male	13	Personal Travel	Eco Plus	460	5	25	0
1	Male	25	Business travel	Business	235	1	1	0
2	Female	26	Business travel	Business	1142	5	0	1
3	Female	25	Business travel	Business	562	2	11	0
4	Male	61	Business travel	Business	214	4	0	1

다음의 변수 관계에 대해 그래프와 가설검정으로 분석하시오.

In [28]:
```
target = 'satisfaction'
```

In [38]:
```
def ed_n(var, target, data):
    mosaic(data, [var, target])
    plt.axhline(1, data[target].mean(), color='r')
    plt.show()
```

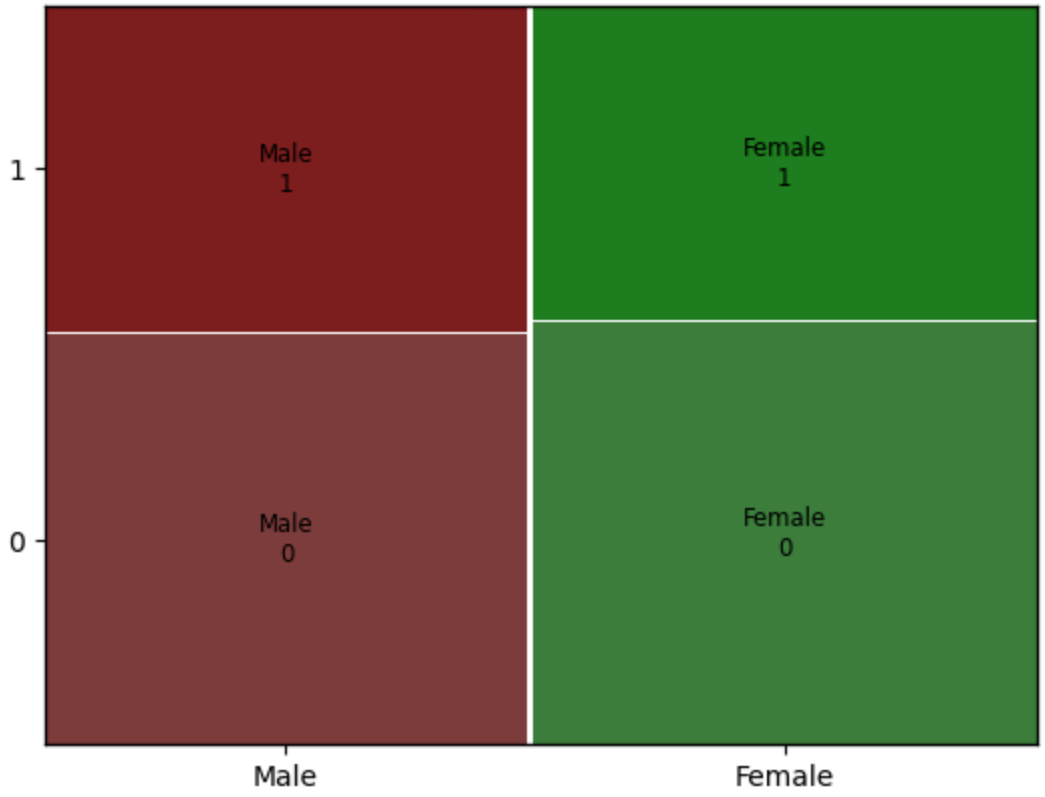
In [42]:
```
def chi2(var, target, data):
    table = pd.crosstab(data[var], data[target])
# 1) 먼저 교차표 집계- normalize 하면 안 됨
    print(table)
    print('-' * 50)

    # 2) 카이제곱검정
    k_statistic, pvalue, dof, expected_freq = spst.chi2_contingency(table)
    print(f'카이제곱 통계량 : {k_statistic}')
    print(f'P_value : {pvalue}')
    print(f'자유도도 : {dof}')
    print(f'기대빈도 : {expected_freq}')
```

(1) Gender --> Satisfaction

In [36]:
```
var = 'Gender'
```

- 시각화

In [39]:
```
ed_n(var, target, data)
```



- 수치화 : 카이제곱검정

In [43]:
```
chi2(var, target, data)
```

```
satisfaction     0     1
Gender
Female        1463  1088
Male          1362  1087
--------------------------------------------------
카이제곱 통계량 : 1.461470294787199
P_value : 0.2266963263128574
자유도도 : 1
기대빈도 : [[1441.315 1109.685]
 [1383.685 1065.315]]
```

- 파악된 내용을 기술해 봅시다.

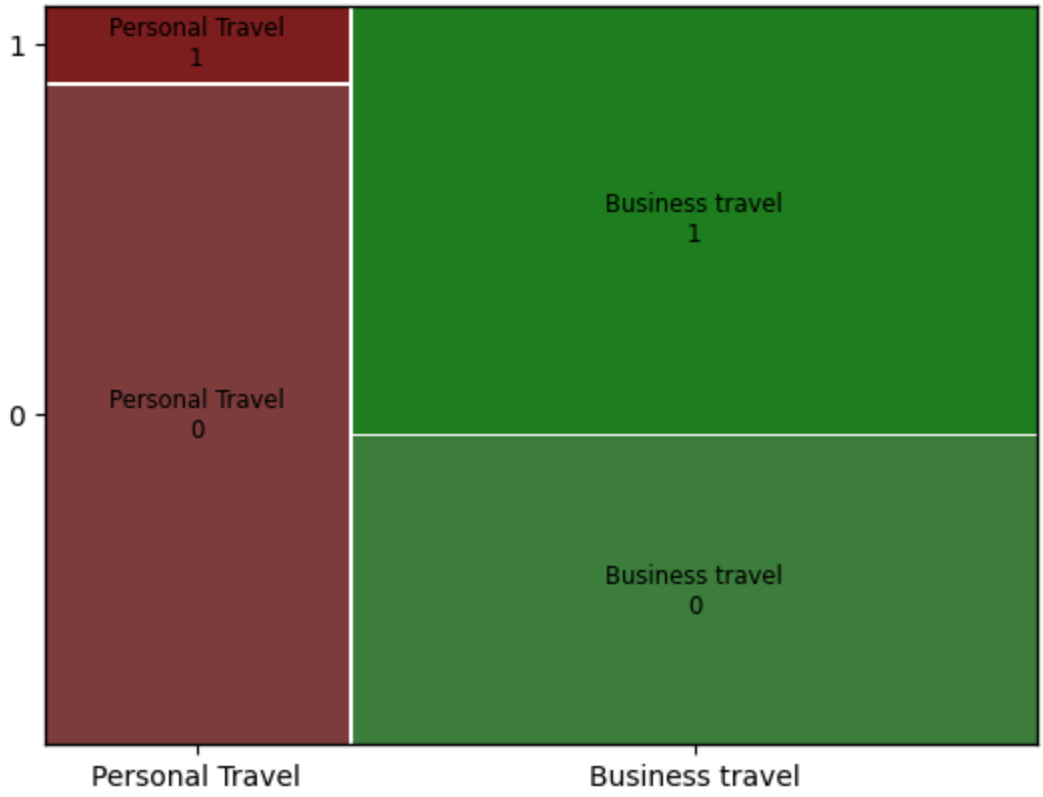
- 성별에 따라 만족도는 관계가 없는 듯 보임
- p-vale 카이제곱 통계량을 봐도 관계가 없어 보임

(2) Type of Travel --> Satisfaction

In [47]:
```
var = 'Type of Travel' # 여행타입
```

- 시각화

In [48]:
```
ed_n(var, target, data)
```



- 수치화 : 카이제곱검정

In [46]:
```
chi2(var, target, data)
```

```
satisfaction     0     1
Gender
Female        1463  1088
Male          1362  1087
---------------------------------------------------
카이제곱 통계량 : 1.461470294787199
P_value : 0.2266963263128574
자유도도 : 1
기대빈도 : [[1441.315 1109.685]
 [1383.685 1065.315]]
```

- 파악된 내용을 기술해 봅시다.

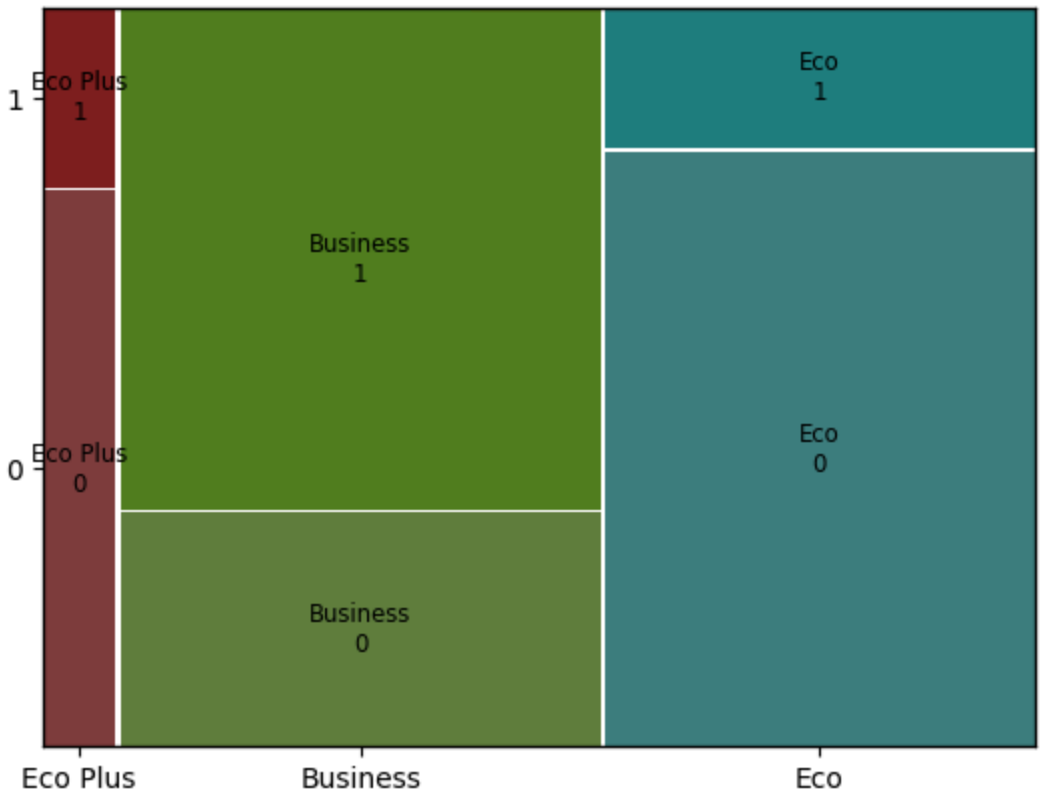
In [49]:
```
# 카이제콥과 p_value는 관계가 없다 나오지만
# 그래프로는 개인적 여행일 수록 만족도가 낮음을 보인다
```

(3) Class --> Satisfaction

In [50]:
```
var = 'Class' # 객실등급
```

- 시각화

In [52]:
```
ed_n(var, target, data)
```



- 수치화 : 카이제곱검정

In [53]:
```
chi2(var, target, data)
```

```
satisfaction     0     1
Class
Business       777  1663
Eco           1765   420
Eco Plus       283    92
--------------------------------------------------
카이제곱 통계량 : 1182.4142005723843
P_value : 1.745897261154762e-257
자유도도 : 2
기대빈도 : [[1378.6   1061.4  ]
 [1234.525  950.475]
 [ 211.875  163.125]]
```

- 파악된 내용을 기술해 봅시다.

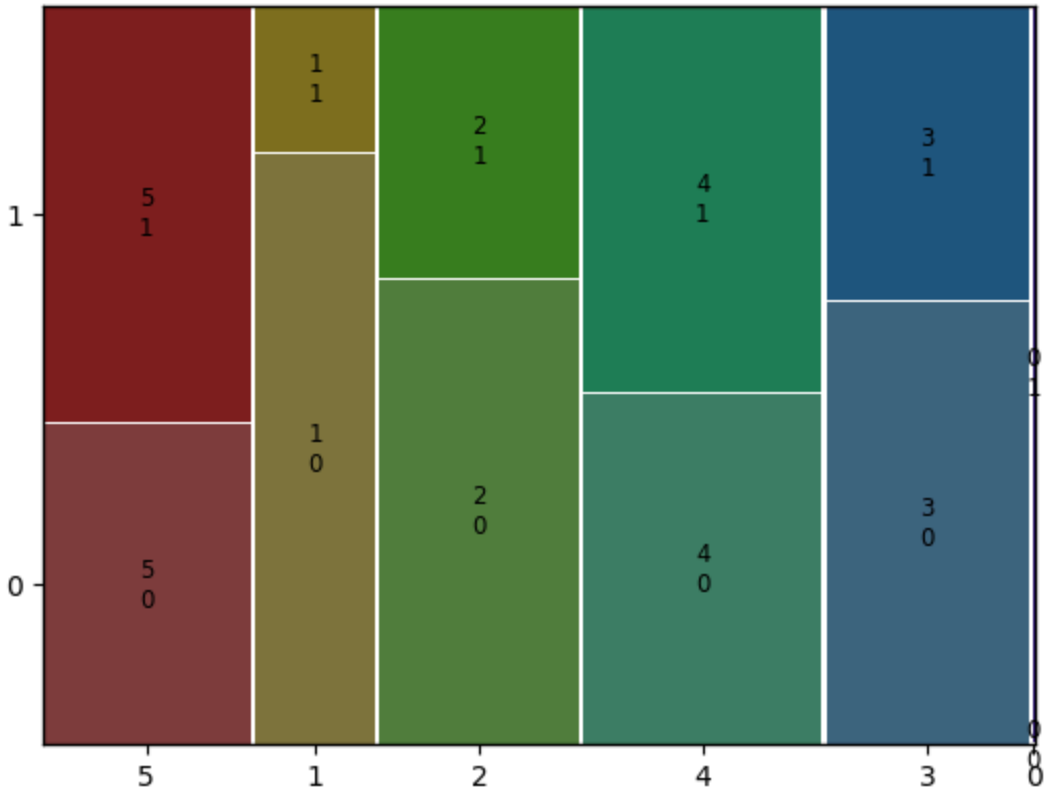
In [54]:
```
# 그래프 및 카이제곱을 봐도 좌석등급에 대해 만족도가 차이가 있어 보임
```

(4) Food and drink --> Satisfaction

In [56]:
```
var = 'Food and drink' # 식음료 만족도
```

- 시각화

In [57]:
```
ed_n(var, target, data)
```



- 수치화 : 카이제곱검정

In [58]:
```
chi2(var, target, data)
```

```
satisfaction      0    1
Food and drink
0                 0    4
1               497  123
2               650  380
3               628  418
4               583  643
5               467  607
--------------------------------------------------
카이제곱 통계량 : 284.02977867350586
P_value : 2.7111955246469l4e-59
자유도도 : 5
기대빈도 : [[  2.26   1.74]
 [350.3  269.7 ]
 [581.95 448.05]
 [590.99 455.01]
 [692.69 533.31]
 [606.81 467.19]]
```

- 파악된 내용을 기술해 봅시다.

In [59]:
```
# 식음료에 따라 만족도가 차이가 있어 보임
```