[실습] 단변량분석 종합실습

0.환경준비

In [1]:

```
import pandas as pd
import numpy as np

import matplotlib.pyplot as plt
import seaborn as sns
```

숫자형 변수

보스톤 집값 데이터

변수	설명
medv	타운별 집값(중위수)
crim	범죄율
zn2	25,000 평방피트를 초과 거주지역 비율 (범주: 0-하, 1-중, 2-상)
indus	비소매상업지역 면적 비율
chas	찰스강변 위치(범주 : 강변1, 아니면 0)
nox	일산화질소 농도
rm	주택당 방 수
age	1940년 이전에 건축된 주택의 비율
dis	직업센터의 거리
rad	방사형 고속도로까지의 거리
tax	재산세율
ptratio	학생/교사 비율
black	인구 중 흑인 비율
lstat	인구 중 하위 계층 비율

In [2]:

```
boston = pd.read_csv('https://bit.ly/3EuWvZw')
boston.head()
```

Out[2]:

	crim	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	med
0	0.00632	2.31	0	0.538	6.575	65.2	4.0900	1	296	15.3	396.90	4.98	24.0
1	0.02731	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.90	9.14	21.6
2	0.02729	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
3	0.03237	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
4	0.06905	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.90	5.33	36.2

① crim

In [3]:

```
var= 'crim'
```

1) 변수의 비즈니스 의미

> 타운별 범죄율

2) 숫자, 범주?

> 숫자

3) NaN 존재 유무

In [ ]:

```
boston[var].isna().sum()
```

Out[ ]:

0

> NaN은 존재하지 않음

4) 기초통계량(수치화)

In [ ]:

```
boston[var].describe()
```

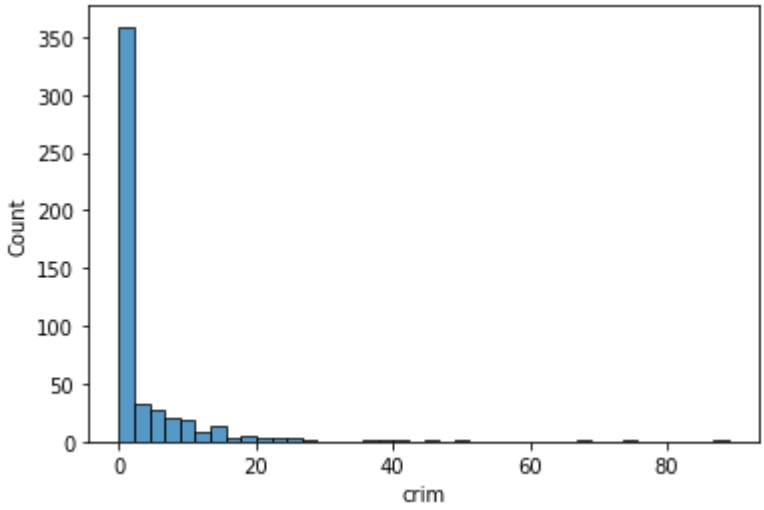
Out[ ]:

```
count    506.000000
mean       3.613524
std        8.601545
min        0.006320
25%        0.082045
50%        0.256510
75%        3.677083
max       88.976200
Name: crim, dtype: float64
```

5) 분포 확인(시각화)

In [ ]:

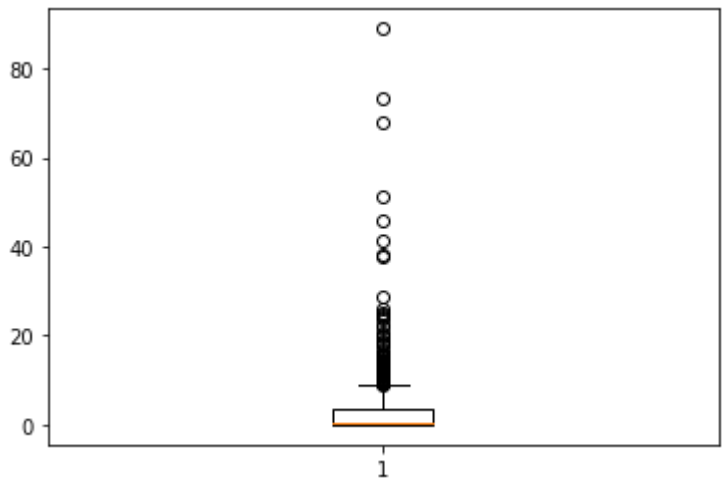
```
sns.histplot(boston[var], bins = 40)
plt.show()
```



In [ ]:

```
box = plt.boxplot(boston[var])
plt.show()

print(box['whiskers'][0].get_ydata())
print(box['whiskers'][1].get_ydata())
```




```
[0.082045 0.00632 ]
[3.6770825 8.98296  ]
```

6) 기초통계량과 분포를 통해서 파악한 내용을 적어 봅시다.

보이는 그대로를 넘어, 비즈니스 관점에서 고민하며 적어 봅시다.

- 대부분(75%의 타운)의 범죄율이 3.6% 이하
- boxlplot 기준으로 볼 때, 약 9% 이상은 이상치로 검토
- 대부분의 타운은 치안 관리가 잘 되는 것 같으나, 9% 이상의 범죄율 지역은 치안상태, 외곽지역, 슬럼가/빈곤층 밀집 지역 등인지 확인이 필요하다.

7) 추가 분석해 볼 사항이 있나요?

- 범죄율을 9% 이상과 이하로 나누고(범주화), 이상인 지역과 이하의 지역에 대한 다른 변수의 차이를 비교해 본다.
- 범죄율과 집값의 관계는?

② ptratio

In [4]:

```
var = 'ptratio'
```

1) 변수의 비즈니스 의미

교사1명당 학생 비율

2) 숫자, 범주?

숫자

3) NaN 존재 유무

In [5]:

```
boston[var].isna().sum()
```

Out[5]:

0

NaN 없음

4) 기초통계량(수치화)

In [6]:

```
boston[var].describe()
```

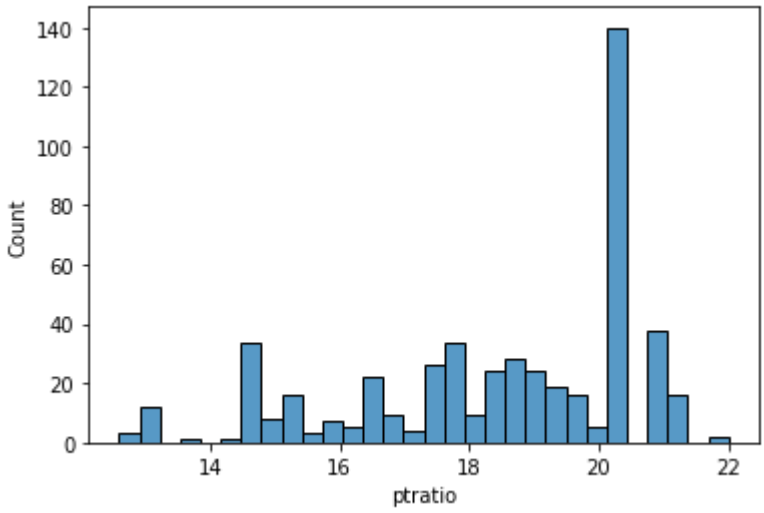
Out[6]:

```
count    506.000000
mean      18.455534
std        2.164946
min       12.600000
25%       17.400000
50%       19.050000
75%       20.200000
max       22.000000
Name: ptratio, dtype: float64
```

5) 분포 확인(시각화)

In [9]:

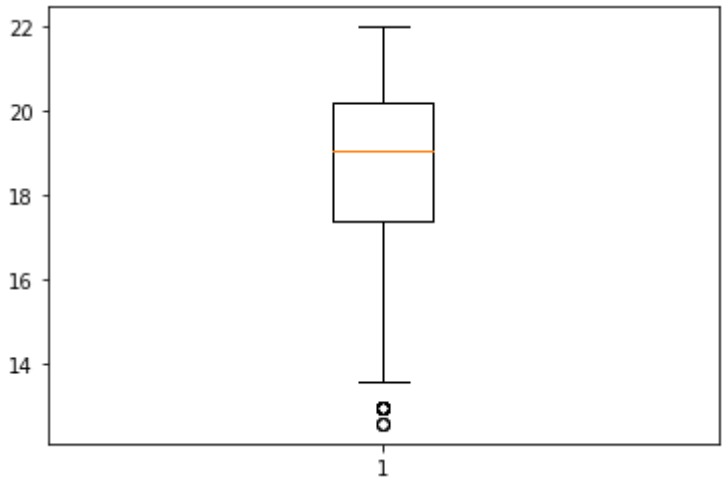
```
sns.histplot(boston[var], bins = 30)
plt.show()
```



In [8]:

```
box = plt.boxplot(boston[var])
plt.show()

print(box['whiskers'][0].get_ydata())
print(box['whiskers'][1].get_ydata())
```




```
[17.4 13.6]
[20.2 22. ]
```

6) 기초통계량과 분포를 통해서 파악한 내용을 적어 봅시다.

보이는 그대로를 넘어, 비즈니스 관점에서 고민하며 적어 봅시다.

- 분포가 퍼져있음
- 20%에 타운이 몰려 있는것으로 보아, 교사학생 비율에 대한 정책적인 기준이 있는 것으로 판단됨.

7) 추가 분석해 볼 사항이 있나요?

- 20% 이상과 미만을 범주로 나누고 분석할 필요 있음.
- 교사학생비율에 따른 집값의 차이는?

③ lstat

In [10]:

```
var = 'lstat'
```

1) 변수의 비즈니스 의미

> 인구 중 하위 계층 비율

2) 숫자, 범주?

> 숫자

3) NaN 존재 유무

In [11]:

```
boston[var].isna().sum()
```

Out[11]:

```
0
```

> NaN 없음

4) 기초통계량(수치화)

In [12]:

```
boston[var].describe()
```

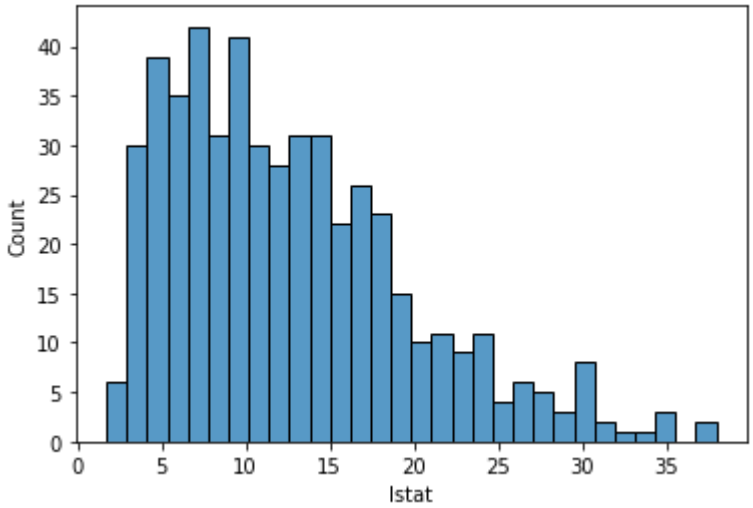
Out[12]:

```
count    506.000000
mean      12.653063
std        7.141062
min        1.730000
25%        6.950000
50%       11.360000
75%       16.955000
max       37.970000
Name: lstat, dtype: float64
```

5) 분포 확인(시각화)

In [13]:

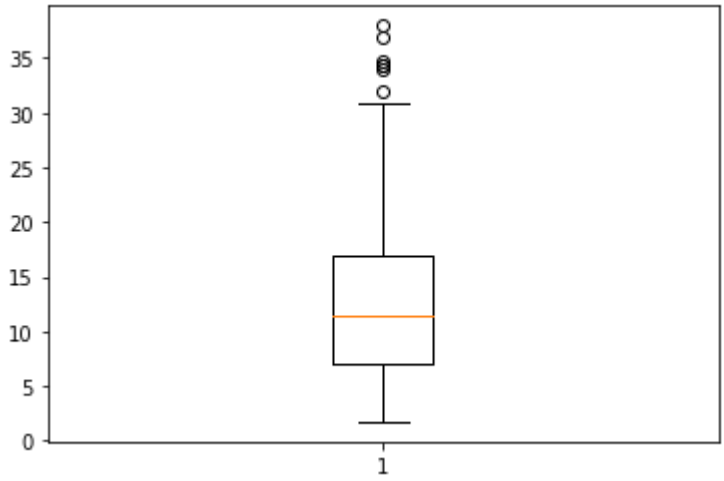
```
sns.histplot(boston[var], bins = 30)
plt.show()
```



In [14]:

```
box = plt.boxplot(boston[var])
plt.show()

print(box['whiskers'][0].get_ydata())
print(box['whiskers'][1].get_ydata())
```




```
[6.95  1.73]
[16.955 30.81 ]
```

6) 기초통계량과 분포를 통해서 파악한 내용을 적어 봅시다.

보이는 그대로를 넘어, 비즈니스 관점에서 고민하며 적어 봅시다.

- 분포가 왼쪽으로 치우침
- 5~15%에 대부분이 몰려 있음.
- boxplot 기준으로 30% 이상은 도시 외곽, 슬럼가 등이 아닐까 확인 필요.

7) 추가 분석해 볼 사항이 있나요?

- 하위계층 비율과 범죄율에 대한 상관관계 분석
- 하위계층 비율과 집값의 관계는?

④ medv

In [15]:

```
var = 'medv'
```

1) 변수의 비즈니스 의미

타운별 집값의 중위수. Target!

2) 숫자, 범주?

숫자

3) NaN 존재 유무

In [16]:

```
boston[var].isna().sum()
```

Out[16]:

0

NaN 없음
(당연히!, Target은 NaN는 없어야 함.)

4) 기초통계량(수치화)

In [17]:

```
boston[var].describe()
```

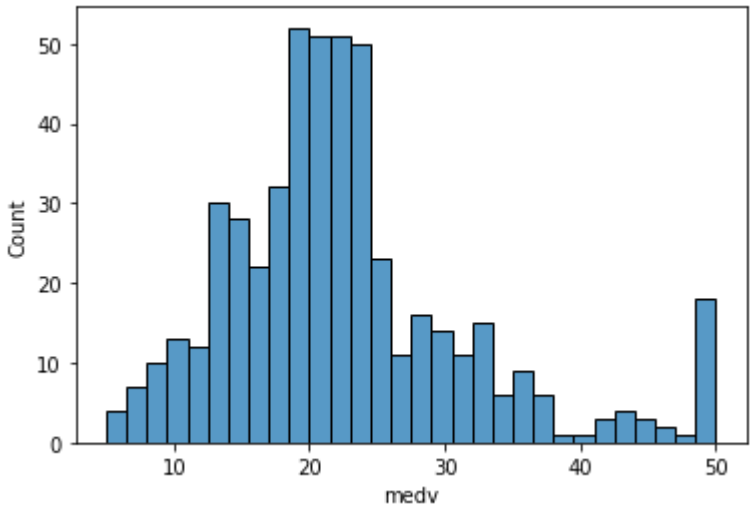
Out[17]:

```
count    506.000000
mean      22.532806
std        9.197104
min        5.000000
25%       17.025000
50%       21.200000
75%       25.000000
max       50.000000
Name: medv, dtype: float64
```

5) 분포 확인(시각화)

In [18]:

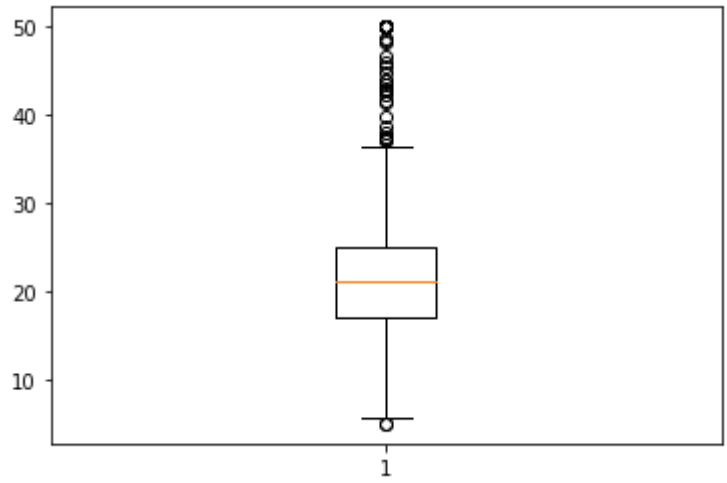
```
sns.histplot(boston[var], bins = 30)
plt.show()
```



In [19]:

```
box = plt.boxplot(boston[var])
plt.show()

print(box['whiskers'][0].get_ydata())
print(box['whiskers'][1].get_ydata())
```




```
[17.025  5.6  ]
[25.  36.5]
```

6) 기초통계량과 분포를 통해서 파악한 내용을 적어 봅시다.

보이는 그대로를 넘어, 비즈니스 관점에서 고민하며 적어 봅시다.

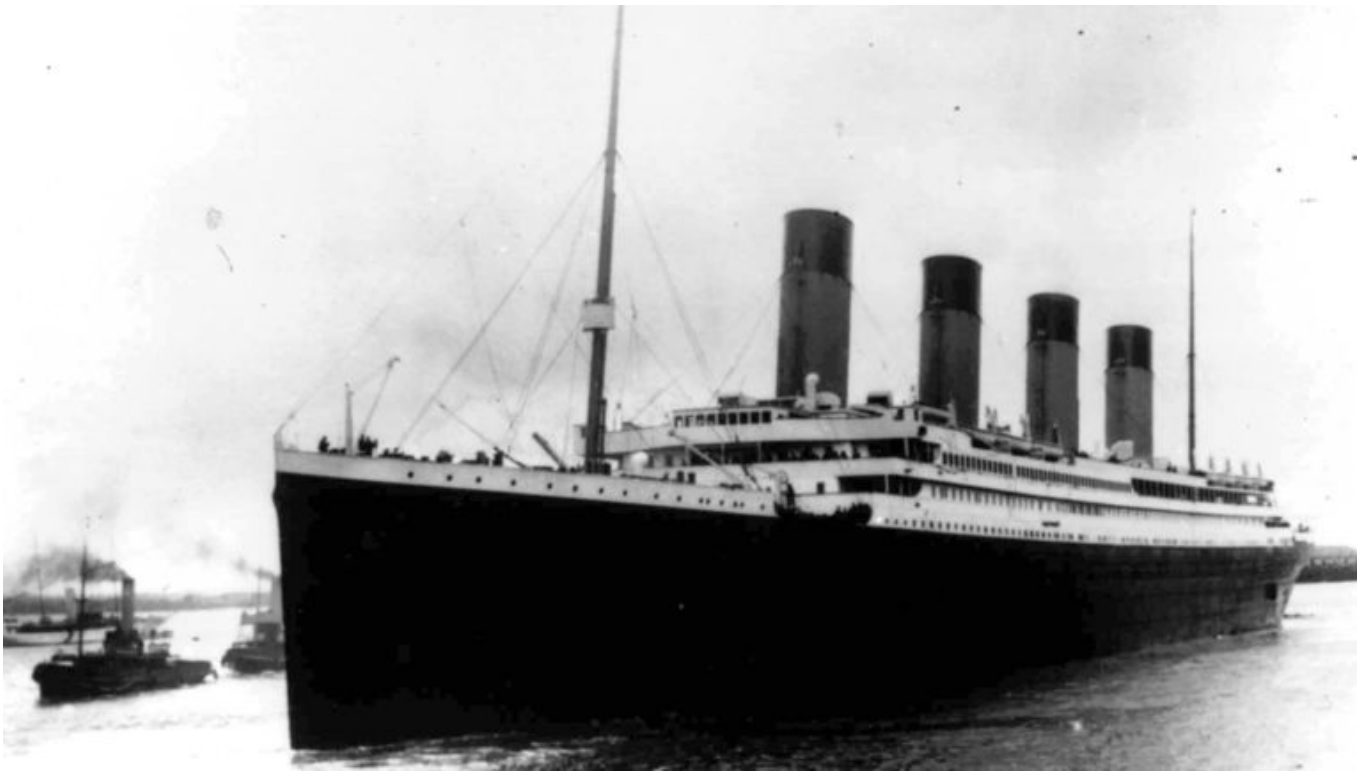
- 4만달러 이상과 이후로 분포가 구분됨.
- 5만달러에 타운들이 몰려 있음 --> 왜?
  - 집값 상한제?
  - 이상치에 대한 데이터 전처리?

7) 추가 분석해 볼 사항이 있나요?

집값의 분포를 4만 달러 이전과 이후에 대해서 구분해서 분석할 필요 있음.

범주형 변수

타이타닉 탑승객 데이터



변수	설명	값 설명
survived	생존여부	0 - 사망, 1- 생존
Pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
Sex	성별	
Age	Age in years	
Sibsp	# of siblings / spouses aboard the Titanic	
Parch	# of parents / children aboard the Titanic	
Ticket	Ticket number	
Fare	Passenger fare	
Cabin	Cabin number	
Embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

In [20]:

```
titanic = pd.read_csv('https://bit.ly/3FsgwkJ')
titanic.head()
```

Out[20]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	7
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	5
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8

① Survived

In [21]:

```
var = 'Survived'
```

1) 변수의 비즈니스 의미

> 탑승객의 생존여부

2) 숫자, 범주?

In [ ]:

```
titanic[var].unique()
```

Out[ ]:

```
array([0, 1])
```

- 범주형 데이터
- 범주 종류 1, 0
  - 1: 생존
  - 0: 사망

3) NaN 존재 유무

In [ ]:

```
titanic[var].isna().sum()
```

Out[ ]:

```
0
```

NA 없음

4) 기초통계량(수치화)

In [ ]:

```
print(titanic[var].value_counts())
print(titanic[var].value_counts()/ len(titanic[var]))
```

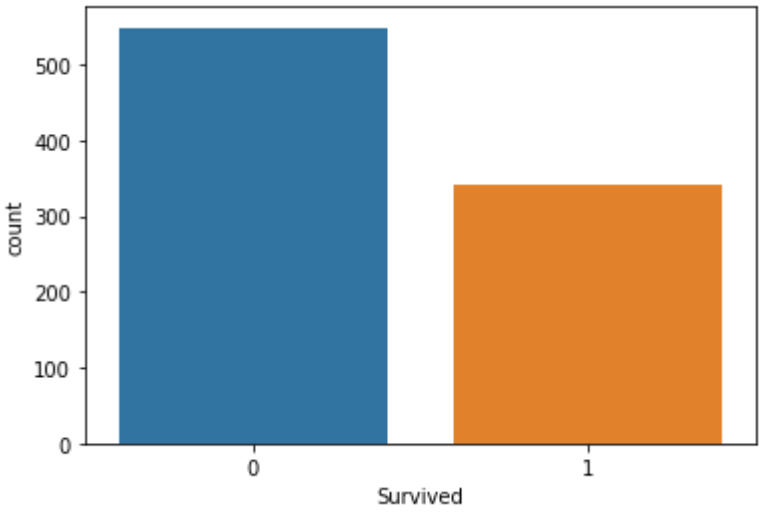
```
0    549
1    342
Name: Survived, dtype: int64
0    0.616162
1    0.383838
Name: Survived, dtype: float64
```

5) 분포 확인(시각화)

In [ ]:

```
sns.countplot(titanic[var])
plt.show()
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning: P
ass the following variable as a keyword arg: x. From version 0.12, the only valid
positional argument will be `data`, and passing other arguments without an explici
t keyword will result in an error or misinterpretation.
  FutureWarning
```



6) 기초통계량과 분포를 통해서 파악한 내용을 적어 봅시다.

보이는 그대로를 넘어, 비즈니스 관점에서 고민하며 적어 봅시다.

- 전체 891명
- 생존자
  - 생존자의 수는 342명
  - 생존율 0.384
- 사망율이 높은 이유는 무엇일까?

7) 추가 분석해 볼 사항이 있나요?

- Survived가 Target 이므로, feature들과 Target 과의 관계를 살펴보게 될 것.
- 그러므로 추가 분석하고자 하는 사항은 feature들을 살펴볼 때 도출하게 될 것이라 생각됨.

② Pclass

In [22]:

```
var = 'Pclass'
```

1) 변수의 비즈니스 의미

> 탑승객의 객실 등급

2) 숫자, 범주?

In [23]:

```
titanic[var].unique()
```

Out[23]:

```
array([3, 1, 2])
```

> - 범주형 데이터
> - 범주 종류 1,2,3
>   - 1: 1등급
>   - 2: 2등급
>   - 3: 3등급

3) NaN 존재 유무

In [24]:

```
titanic[var].isna().sum()
```

Out[24]:

```
0
```

> NA 없음

4) 기초통계량(수치화)

In [25]:

```
print(titanic[var].value_counts())
print(titanic[var].value_counts()/ len(titanic[var]))
```

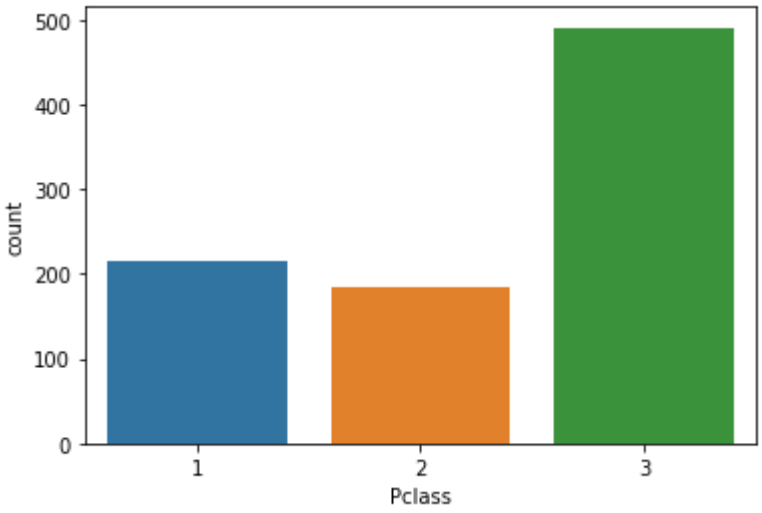
```
3    491
1    216
2    184
Name: Pclass, dtype: int64
3    0.551066
1    0.242424
2    0.206510
Name: Pclass, dtype: float64
```

5) 분포 확인(시각화)

In [29]:

```
sns.countplot(titanic[var])
plt.show()
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning: P
ass the following variable as a keyword arg: x. From version 0.12, the only valid
positional argument will be `data`, and passing other arguments without an explici
t keyword will result in an error or misinterpretation.
  FutureWarning
```



6) 기초통계량과 분포를 통해서 파악한 내용을 적어 봅시다.

보이는 그대로를 넘어, 비즈니스 관점에서 고민하며 적어 봅시다.

- 전체 891명
- 객실등급 비율
  - 3 : 0.551066
  - 1 : 0.242424
  - 2 : 0.206510
- 탑승지역이 주로 2차산업혁명의 중심 도시 --> 아메리칸드림을 갖고 탑승한 노동자들이 주류 --> 3등급 객실에 주로 탑승
- 혹은, 원래 타이타닉호의 객실수는 정해져 있는것.

7) 추가 분석해 볼 사항이 있나요?

- 객실 등급별 생존여부
- 객실 등급, 탑승지역 비교

③ Sex

In [30]:

```
var = 'Sex'
```

1) 변수의 비즈니스 의미

> 탑승객의 성별

2) 숫자, 범주?

In [31]:

```
titanic[var].unique()
```

Out[31]:

```
array(['male', 'female'], dtype=object)
```

> - 범주형 데이터
> - 범주 종류
>   - male
>   - female

3) NaN 존재 유무

In [32]:

```
titanic[var].isna().sum()
```

Out[32]:

```
0
```

> NA 없음

4) 기초통계량(수치화)

In [33]:

```
print(titanic[var].value_counts())
print(titanic[var].value_counts()/ len(titanic[var]))
```

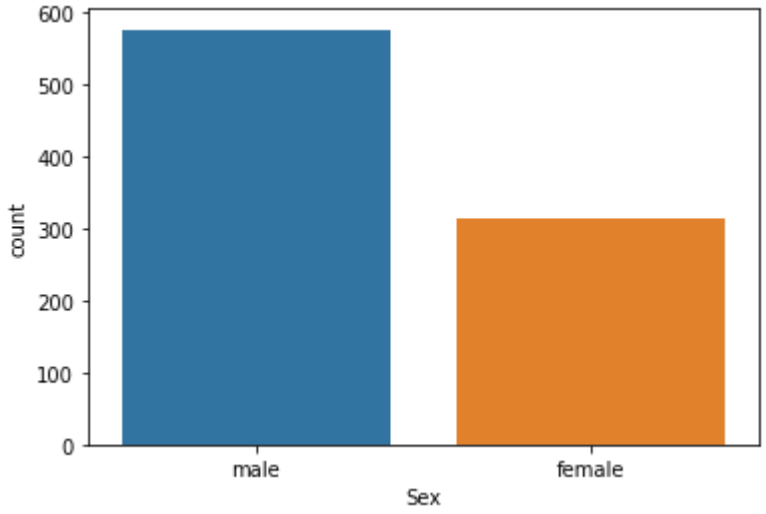
```
male      577
female    314
Name: Sex, dtype: int64
male      0.647587
female    0.352413
Name: Sex, dtype: float64
```

5) 분포 확인(시각화)

In [34]:

```
sns.countplot(titanic[var])
plt.show()
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning: P
ass the following variable as a keyword arg: x. From version 0.12, the only valid
positional argument will be `data`, and passing other arguments without an explici
t keyword will result in an error or misinterpretation.
  FutureWarning
```



6) 기초통계량과 분포를 통해서 파악한 내용을 적어 봅시다.

보이는 그대로를 넘어, 비즈니스 관점에서 고민하며 적어 봅시다.

- 전체 891명
- 성별 비율
  - male : 0.647587
  - female : 0.352413

7) 추가 분석해 볼 사항이 있나요?

- 성별 별 생존여부

④ Embarked

In [35]:

```
var = 'Embarked'
```

1) 변수의 비즈니스 의미

탑승객의 승선지역

2) 숫자, 범주?

In [36]:

```
titanic[var].unique()
```

Out[36]:

```
array(['S', 'C', 'Q', nan], dtype=object)
```

- 범주형 데이터
- 범주 종류
  - C : Cherbourg
  - Q : Queenstown
  - S : Southampton
  - nan : NaN 데이터가 있다는 의미!

3) NaN 존재 유무

In [38]:

```
print(titanic[var].isna().sum())
print(titanic[var].isna().sum()/len(titanic[var]))
```

```
2
0.002244668911335578
```

- NA 2건, 0.22%
- 어떻게 조치를 하는것이 좋을까요?
  - 1) 삭제한다.
  - 2) 값을 채운다.
  - 3) 추정해서 넣는다.

4) 기초통계량(수치화)

In [39]:

```
print(titanic[var].value_counts())
print(titanic[var].value_counts()/ len(titanic[var]))
```

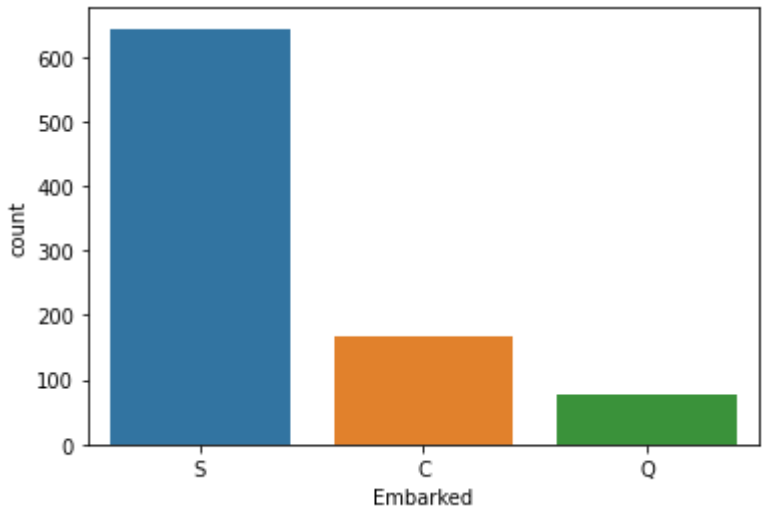
```
S    644
C    168
Q     77
Name: Embarked, dtype: int64
S    0.722783
C    0.188552
Q    0.086420
Name: Embarked, dtype: float64
```

5) 분포 확인(시각화)

In [40]:

```
sns.countplot(titanic[var])
plt.show()
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning: P
ass the following variable as a keyword arg: x. From version 0.12, the only valid
positional argument will be `data`, and passing other arguments without an explici
t keyword will result in an error or misinterpretation.
  FutureWarning
```



6) 기초통계량과 분포를 통해서 파악한 내용을 적어 봅시다.

보이는 그대로를 넘어, 비즈니스 관점에서 고민하며 적어 봅시다.

- 전체 891명
- 탑승지 비율
  - Cherbourg : 0.188552
  - Queenstown : 0.086420
  - Southampton : 0.722783

7) 추가 분석해 볼 사항이 있나요?

- 탑승지 별 생존 여부