CNU 데이터 분석 교육

5th lecture

"Visualization"

2022 - 10 - 31

수학과 최명수

지난시간?

1. EDA with Pandas ★ ★ ★

오늘은 무엇을?

1. matplotlib, seaborn 라이브러리를 이용한 시각화 실습 ★ ★ ★ ★ ★
   -plot, scatter, box plot

데이터 분석 절차

1. 문제정의
-사회현상, 기업이 마주친 문제를
  데이터의 문제로 정의
-데이터 선택

2. 데이터 수집
-선택한 데이터가 잘 확보되어 있는가?

3. 데이터 전처리
-노가다... 개노가다....
-노이즈 제거, 결측치 처리, 구조변경, 중복제거

4. 데이터 모델링, 마이닝
- 데이터간의 관계를 설정
- 데이터의 패턴을 찾거나 분류 또는 예측

5. 데이터 분석
- 의미있는 결과를 도출하는 과정
- 머신러닝... 예측, 분류 등등

6. 데이터 시각화 및 평가
- 분석결과에 대한 정보를 요약, 압축
- 결과를 해석, 분석 목적과 일치하는지 평가

matplotlob

Plot : 선 그래프 그리기
Scatter : 산점도 그리기

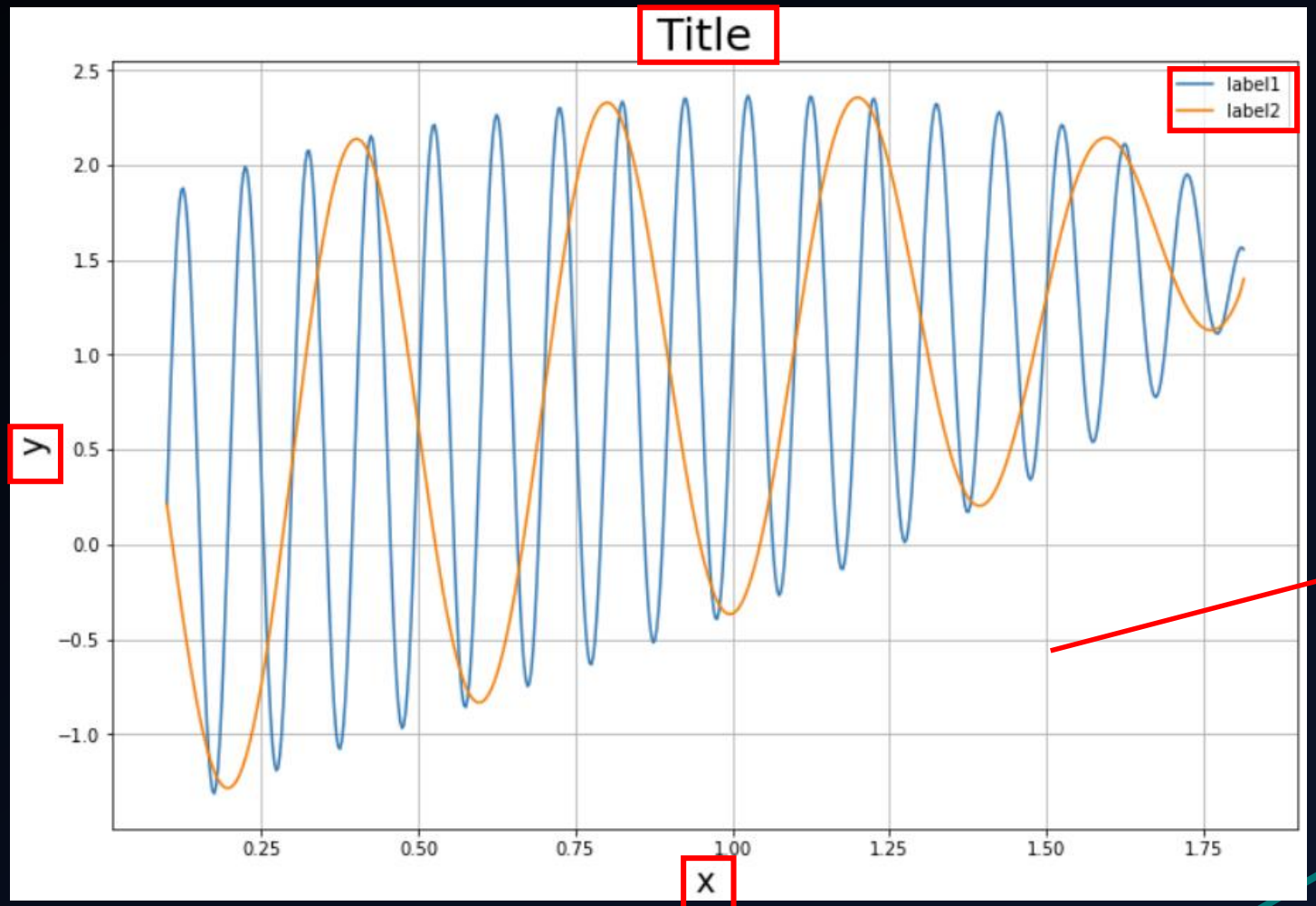
- 그래프는 무수히 많은 점의 집합이다!
  -> $graph = \{(x, f(x)) \mid x \in X, f : X \to Y\}$

시각화의 포인트 ★ ★ ★ ★ ★

1. 설명하고자 하는 대상을 직관적으로 명확하게 표현해줘야 한다.
2. 시각화 자료를 보고 다른 이에게 전달하고자 하는것을 정확히 설명할 수 있어야 한다.

matplotlib

title



legend

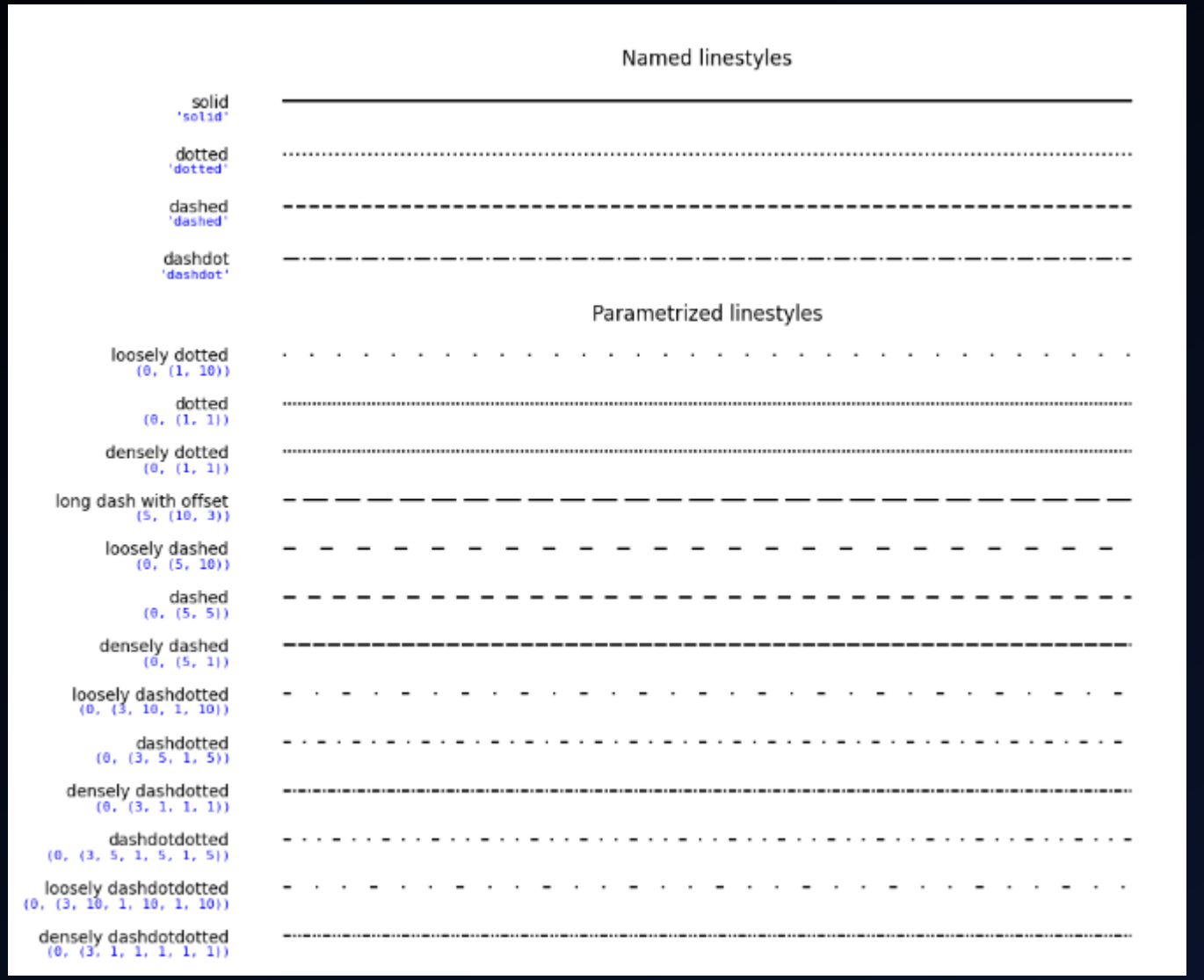
ylabel

grid

xlabel

matplotlob

linestyle

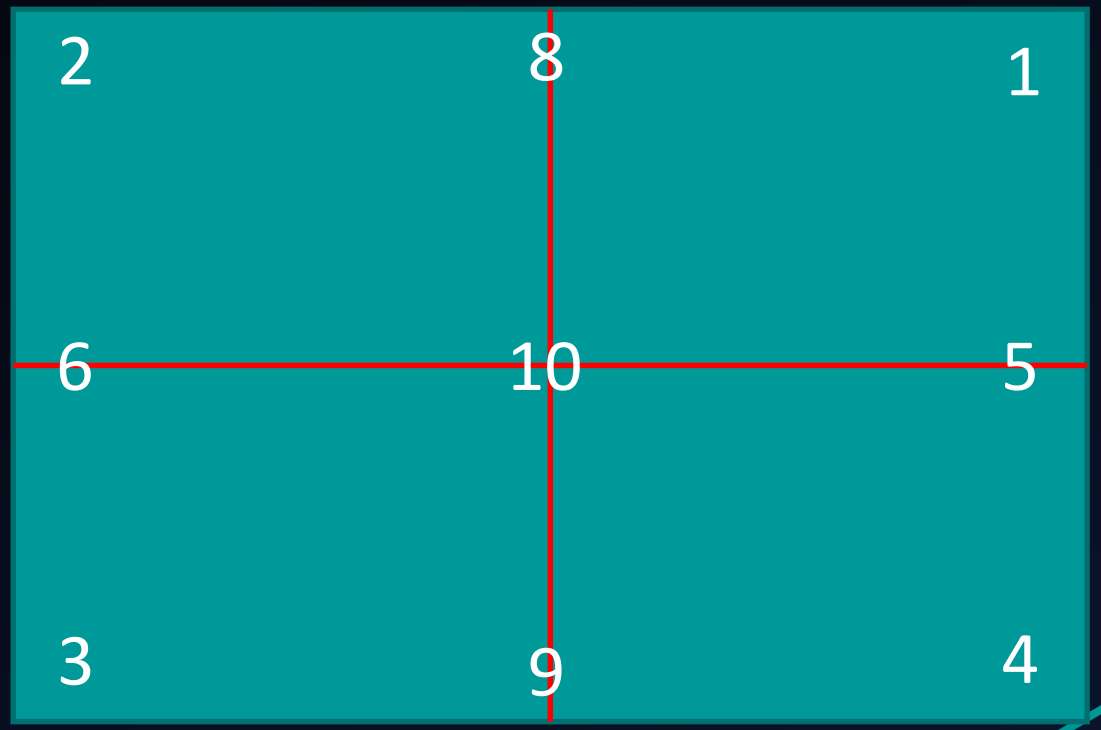


https://matplotlib.org/stable/gallery/lines_bars_and_markers/linestyles.html

matplotlib

Legend 위치

Location String	Location Code
'best'	0
'upper right'	1
'upper left'	2
'lower left'	3
'lower right'	4
'right'	5
'center left'	6
'center right'	7
'lower center'	8
'upper center'	9
'center'	10

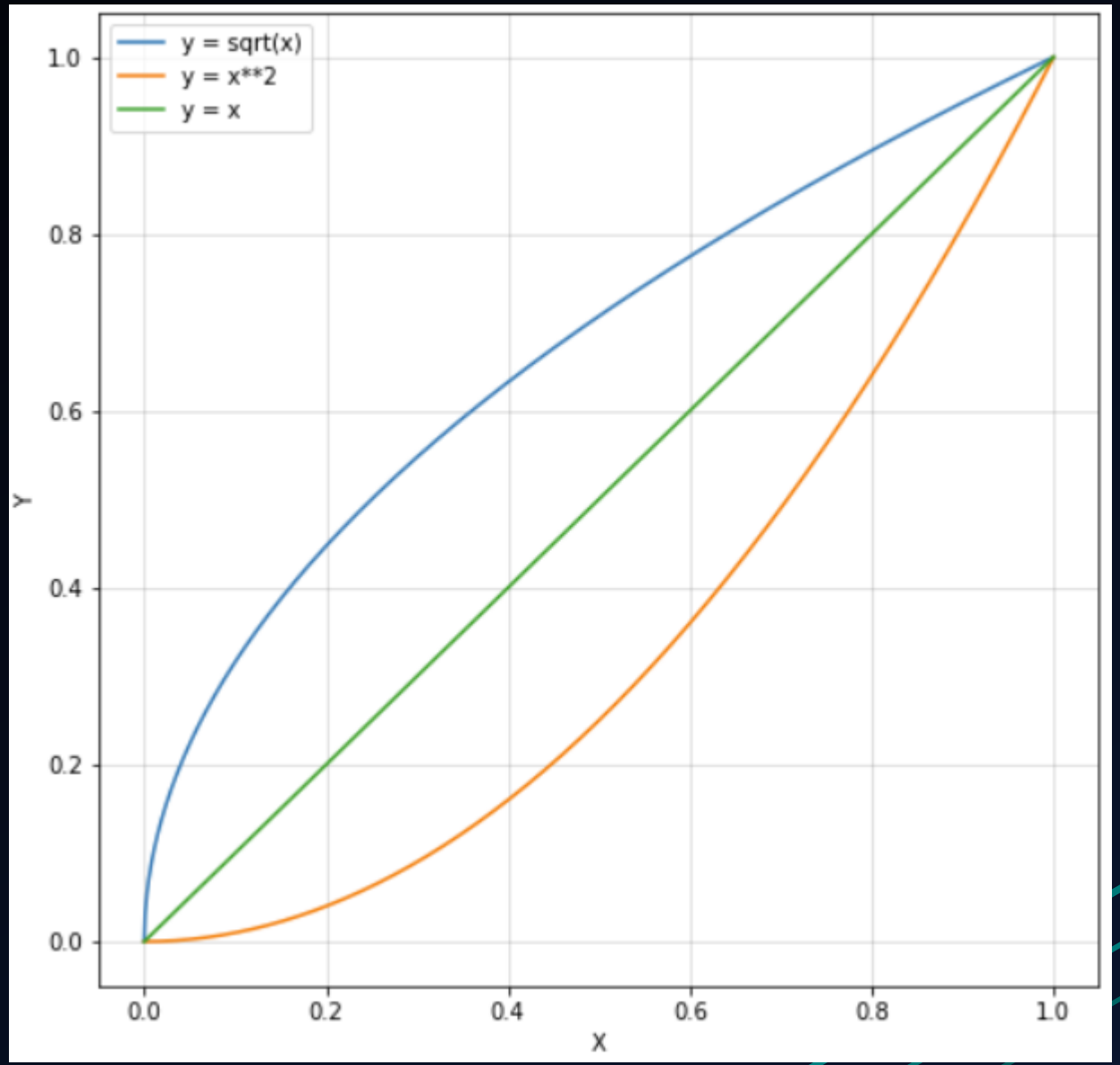


문제 1

다음 그래프를 그리시오

$$y_1 = \sqrt{x},$$
$$y_2 = x^2,$$
$$y_3 = x$$

x는 0부터 1까지.

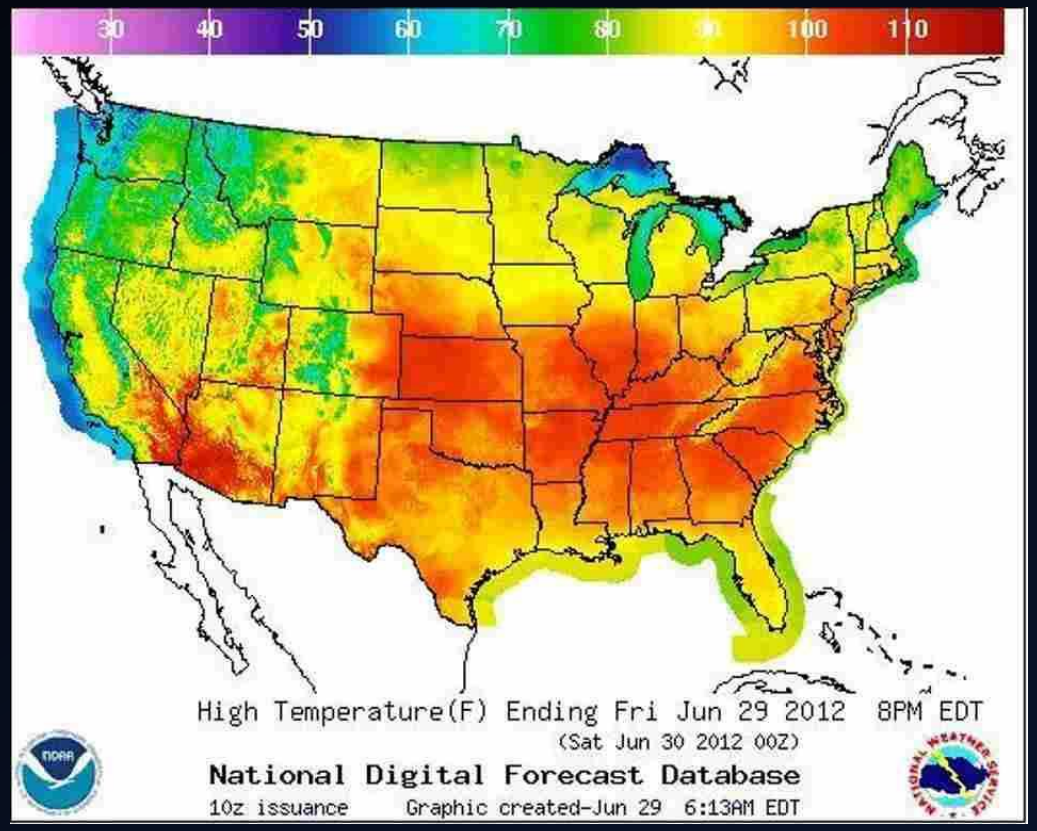


정답

```
# 문제 1
plt.figure(figsize = (8,8))

x = np.linspace(0,1,1001)
y1 = np.sqrt(x)
y2 = x**2
y3 = x

plt.plot(x, y1, label = 'y = sqrt(x)')
plt.plot(x, y2, label = 'y = x**2')
plt.plot(x, y3, label = "y = x")

plt.xlabel("X")
plt.ylabel("Y")

plt.grid(color = 'gray', alpha = 0.3, linestyle = '-')
plt.legend(loc = "best")
plt.show()
```

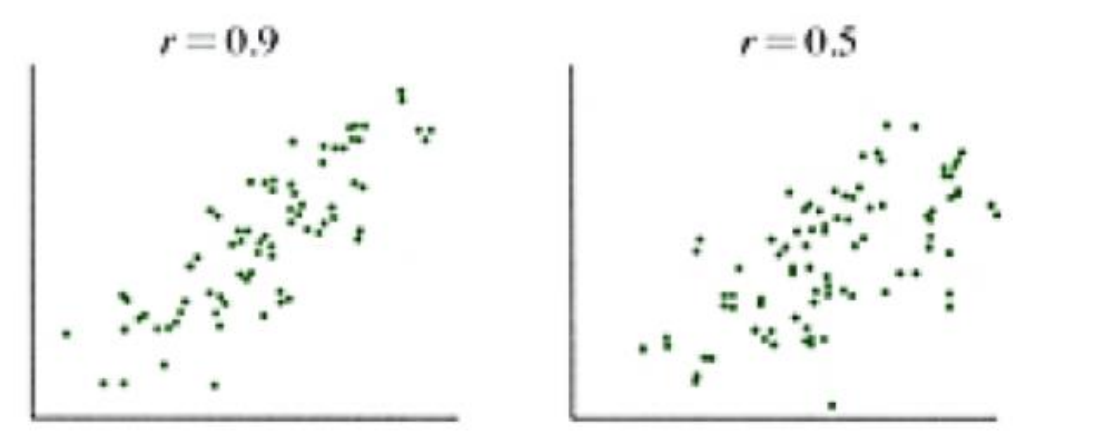
Seaborn

Heatmap
-2차원 데이터를 색으로 표현해줌으로써 데이터의 높고 낮음을 표현

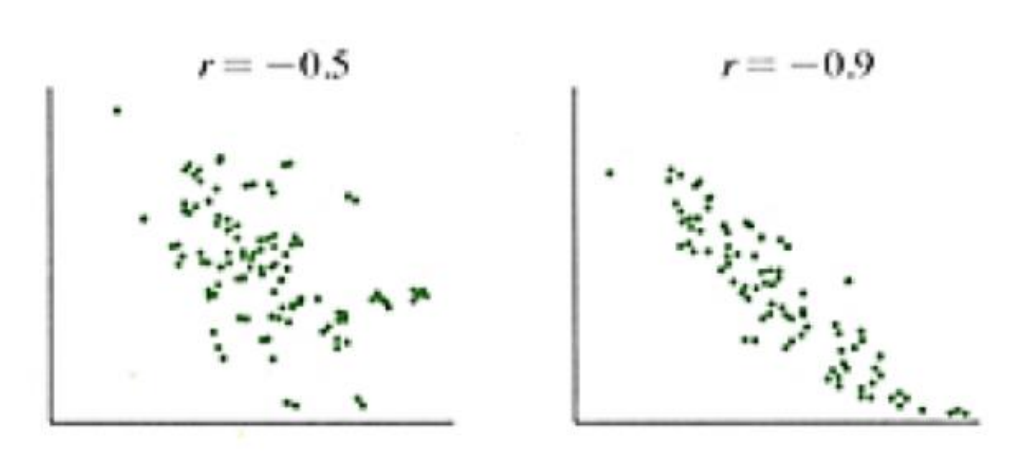


상관계수

양의 상관관계



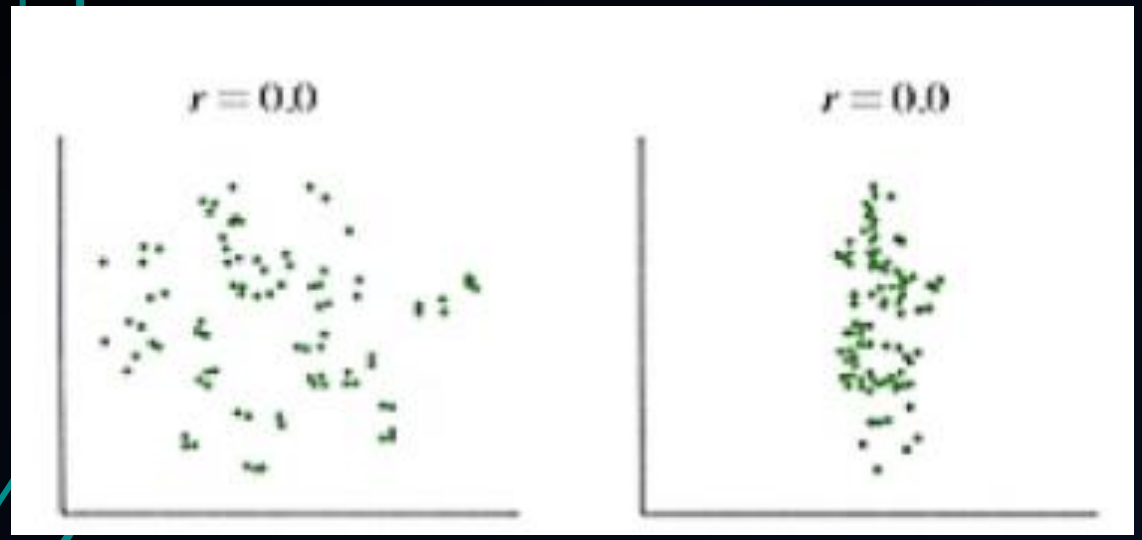
음의 상관관계

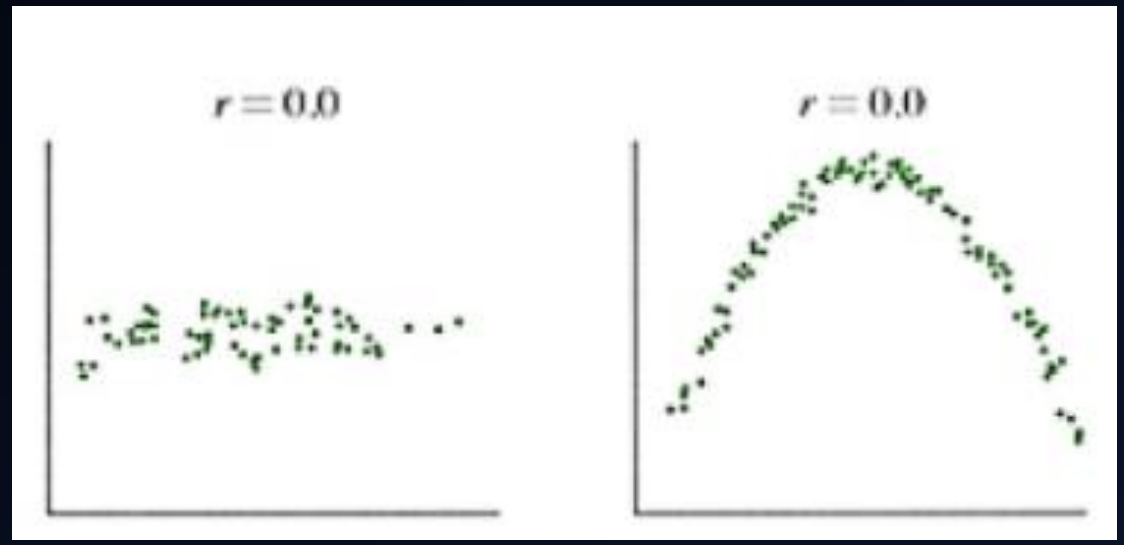


https://leedakyeong.tistory.com/entry/%EA%B8%B0%EC%B4%88%ED%86%B5%EA%B3%84-%EC%83%81%EA%B4%80%EA%B3%84%EC%88%98%EB%9E%80-What-is-correlation-coefficient

상관계수

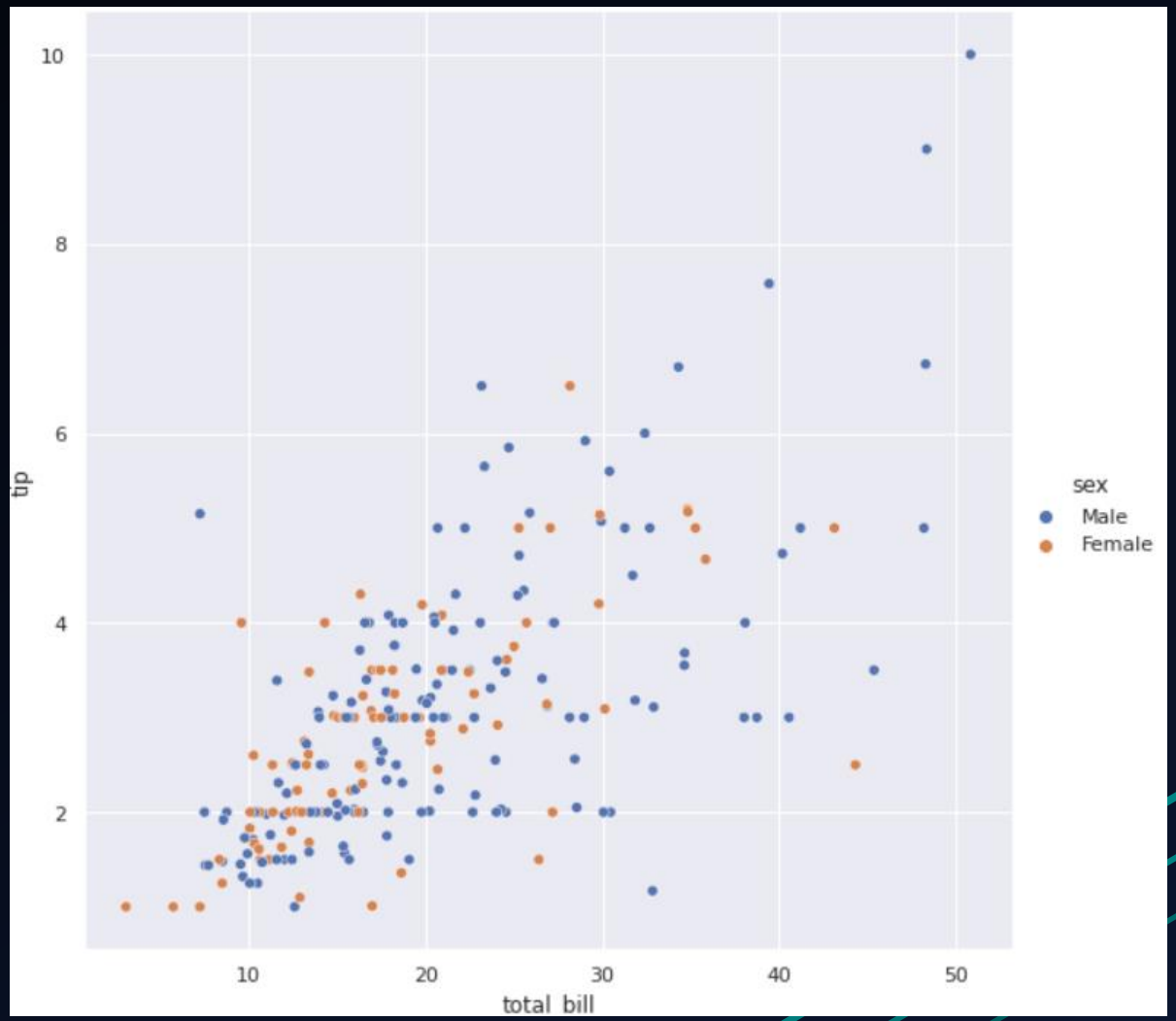
노 상관관계





Seaborn

Relplot
-seaborn 에서의 scatter plot

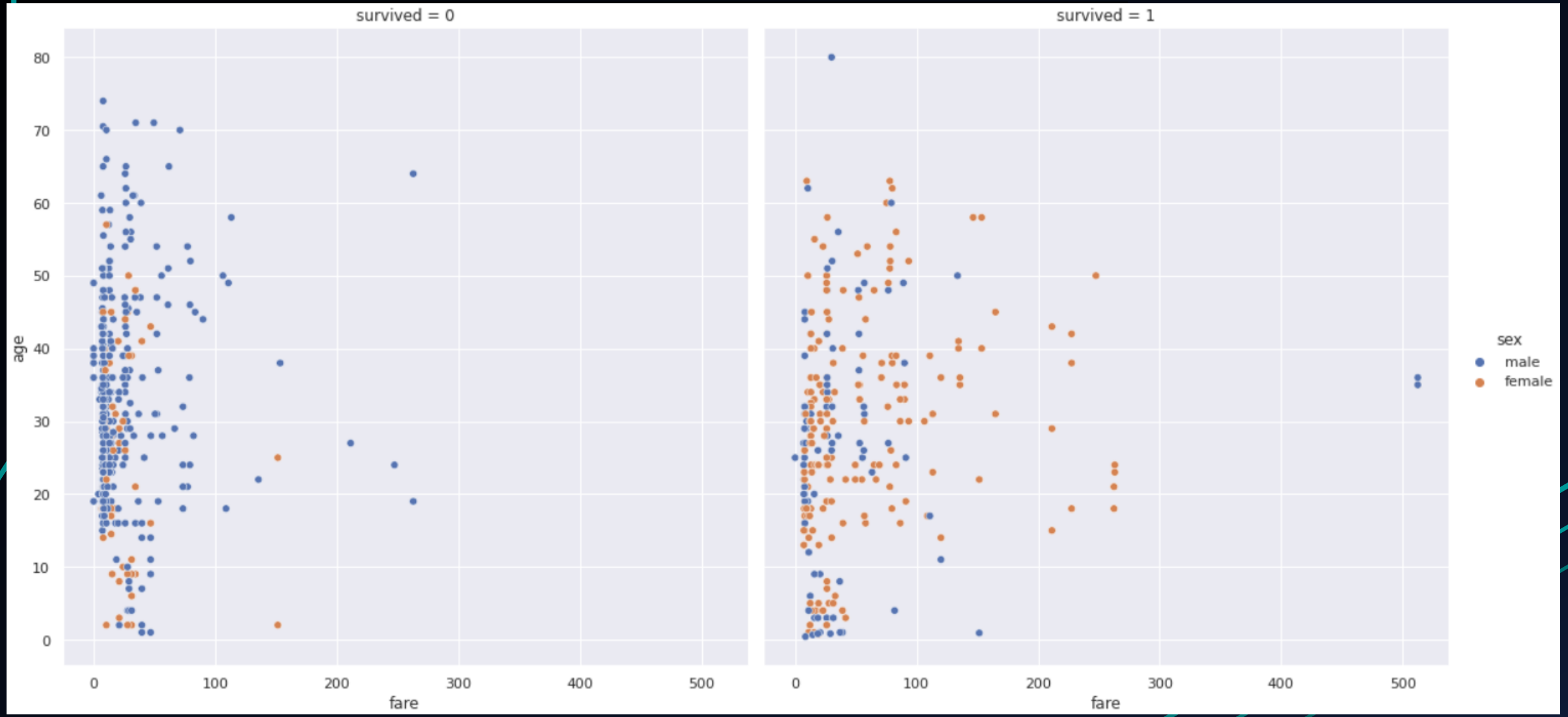


문제2

문제 2
titanic 데이터를 운임요금(fare)에 따른 연령(age)를 아래 조건하에 relplot을 이용하여 산점도를 나타내고 분석하라

- 생사여부(survived)를 각각 구분할 것.
- 성별(sex)에 따라 산점도를 표시할 것.

문제2


survived = 0
survived = 1
age
fare
fare
sex
male
female
0
10
20
30
40
50
60
70
80
0
100
200
300
400
500

문제 2

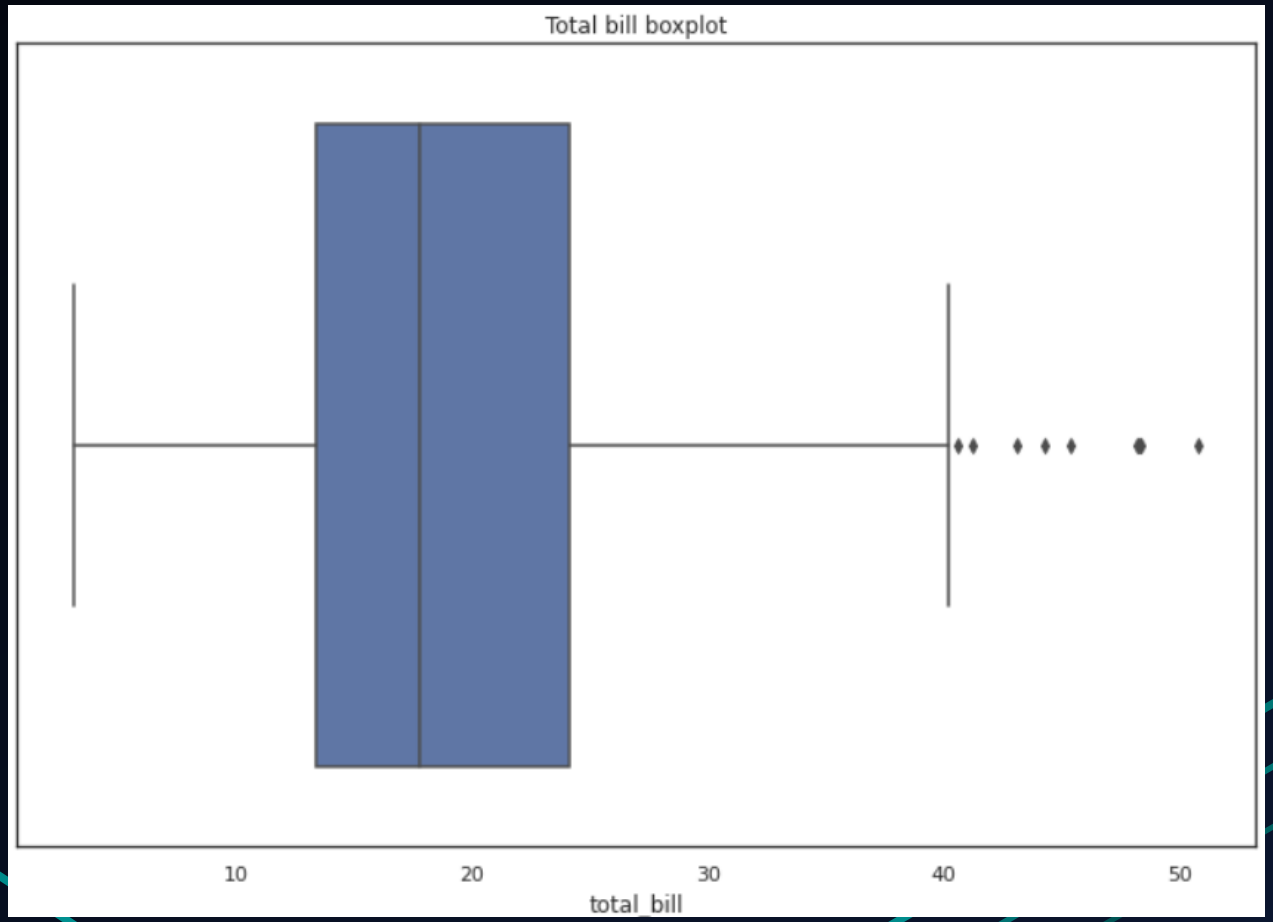
정답

```
# 문제 2
titanic = sns.load_dataset("titanic")
titanic

sns.relplot(x = "fare", y = "age",
            hue = 'sex', height = 8, col = 'survived', data = titanic);
```

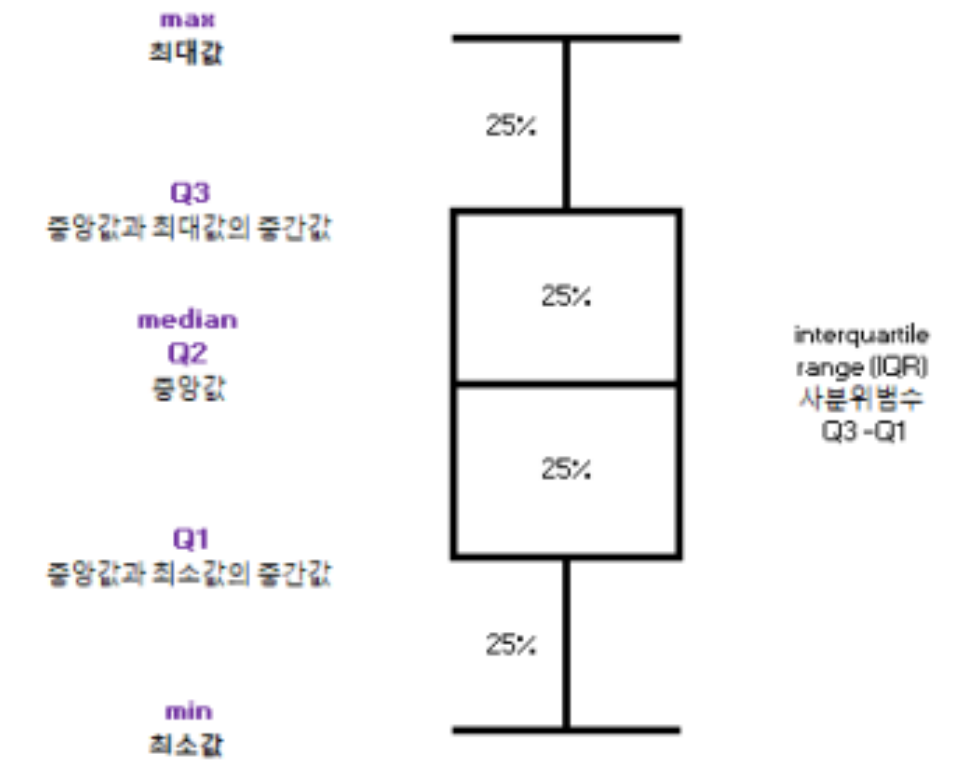
- 운임요금과 연령간의 상관관계는 전혀 없다.
- 생존자는 주로 여성이 더 많다.
- 운임요금을 많이 낸 사람일수록 운임요금이 적은 사람보다 생존한 사례가 더 많다.

Seaborn

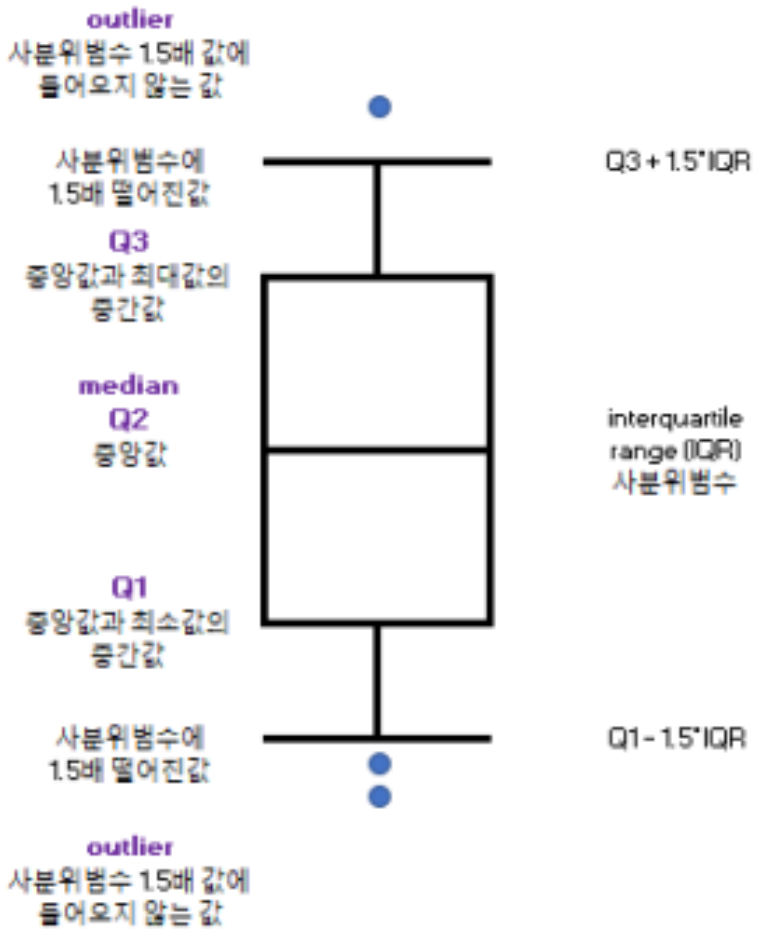
Boxplot
-자료의 분포를 박스형태로 나타내는 시각화 기법



Seaborn



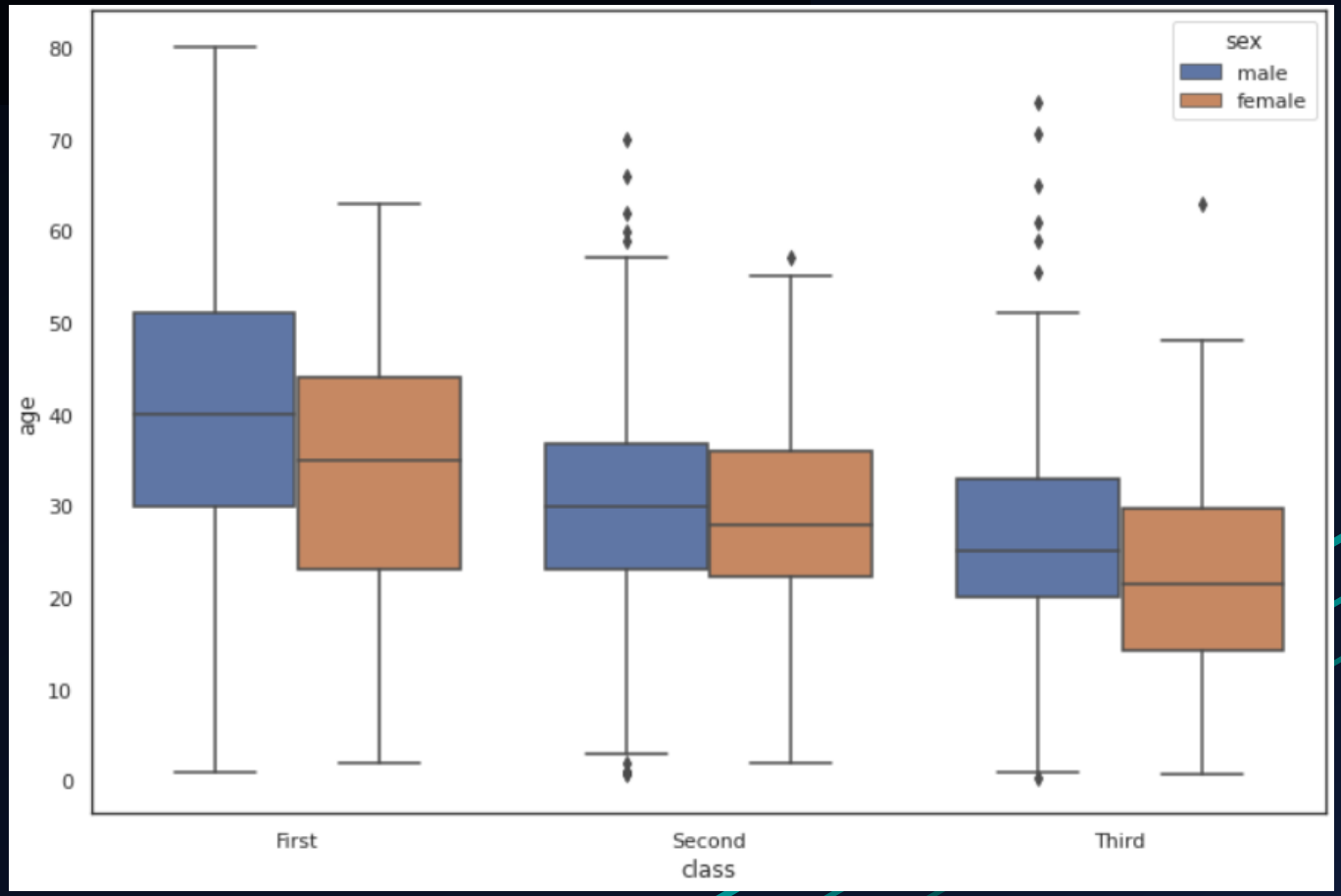
<이상치 예측하지 않았을 때 상자그림: 상자수염의 끝은 최대값과 최소값>



<이상치 예측할 때 상자그림: 상자수염의 끝은 사분위범수에 1.5배 떨어진 값>

문제 3

Titanic 데이터를 이용하여 등급(class)에 따른 연령(age)을 성별(sex)을 구분하여 boxplot으로 시각화 하고, 분석하라



문제 3

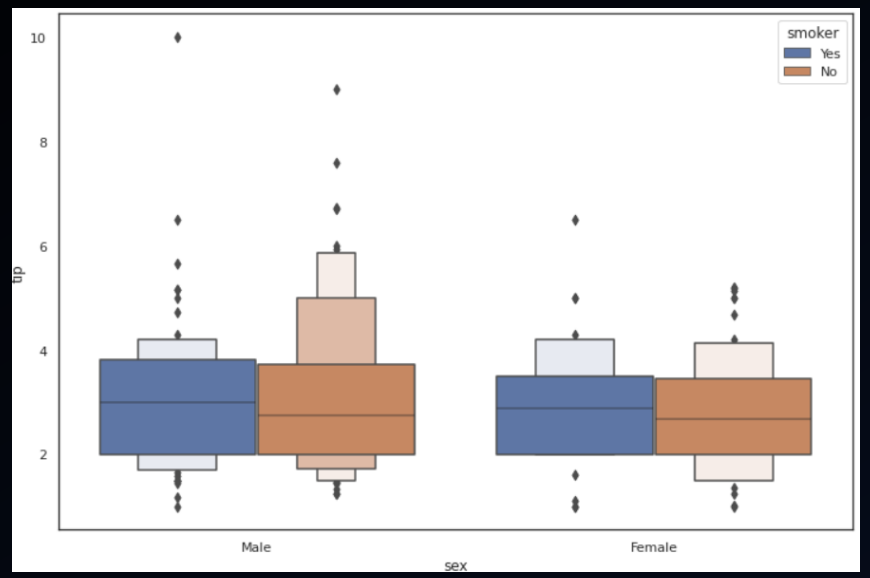
정답

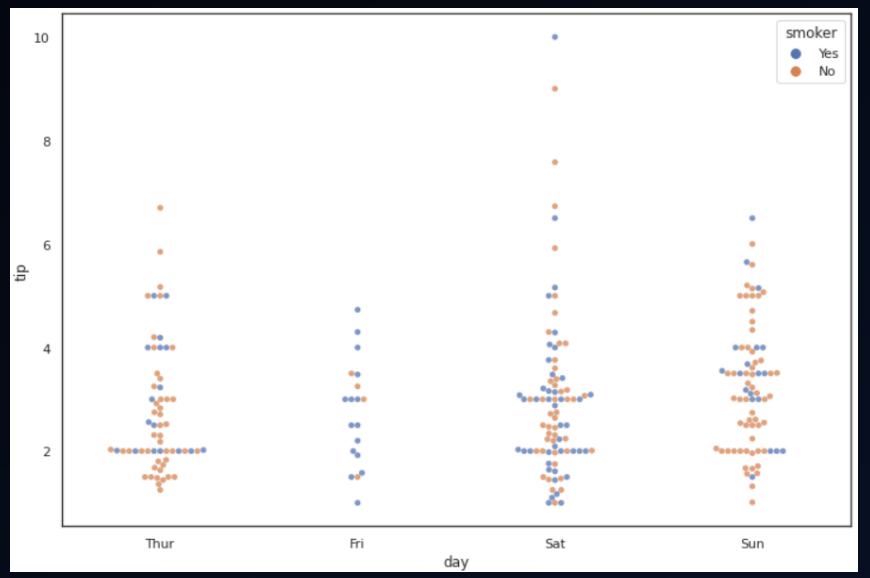
```
# 문제 3
titanic = sns.load_dataset("titanic")
titanic


sns.boxplot(x = "class", y = "age", hue = "sex", data = titanic);
```

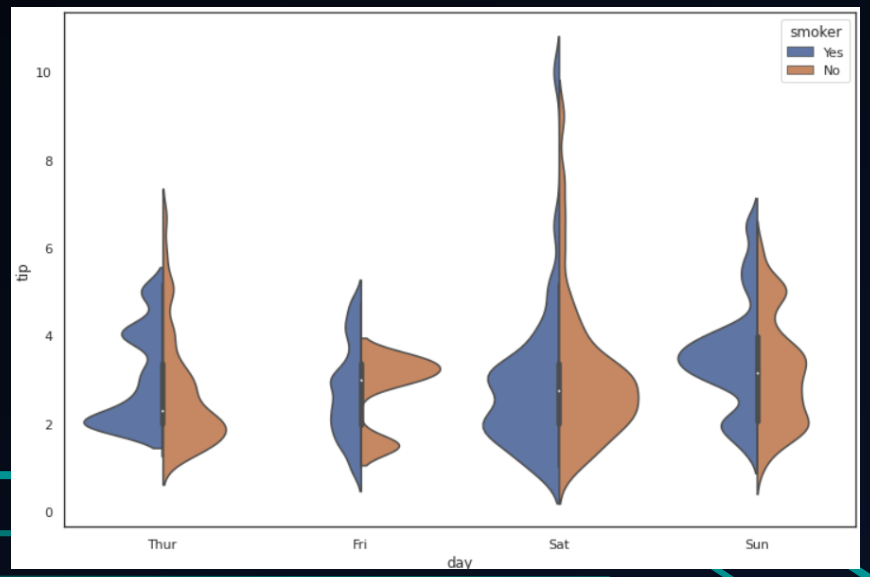
- 연령이 높을수록 높은 등급으로 탑승하였다.
- 같은 등급에서는 대체로 남성의 연령이 여성보다 더 높다.

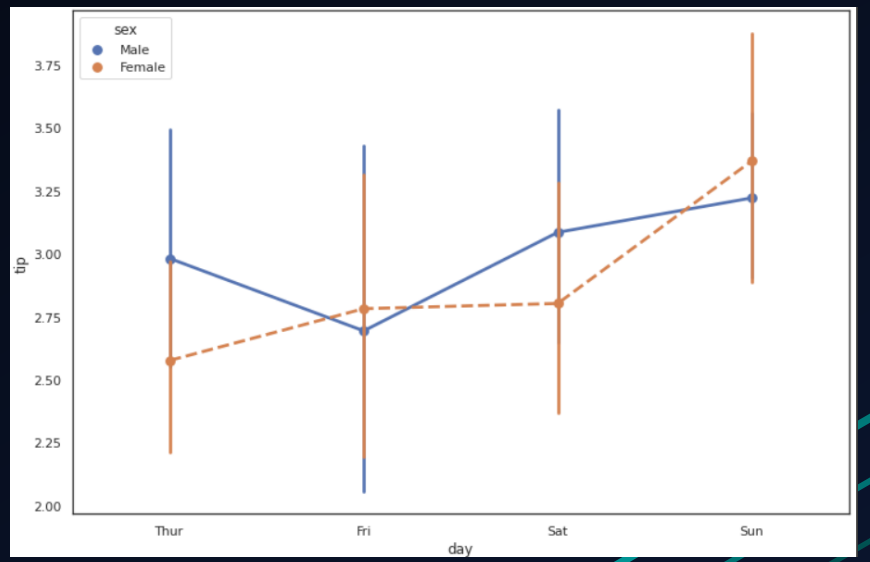
Seaborn

여러가지 boxplot





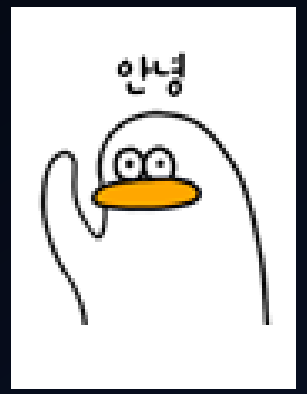




공지

다음시간에는 barplot, pieplot, histogram 실습

끝



담에봅시당