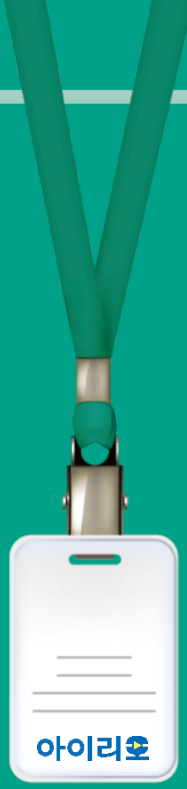


탐색적 데이터 분석 4강

- 데이터분석의 모든 것 -

강사명_이정인 저자

Contents

[04강] 학습 목표와 내용

학습 목표	탐색적 데이터 분석을 이해한다. 변수 간 관계 탐색을 학습하고, 실습한다.
학습 내용	관계 시각화, 상관계수
학습 자료	-
핵심 키워드	관계 시각화, 상관계수
참고 자료	데이터분석의 모든 것(출판 : 아이리포, 저자 : 이정인/장원중)

4. 변수 간 관계 탐색

4.1 산점도그래프

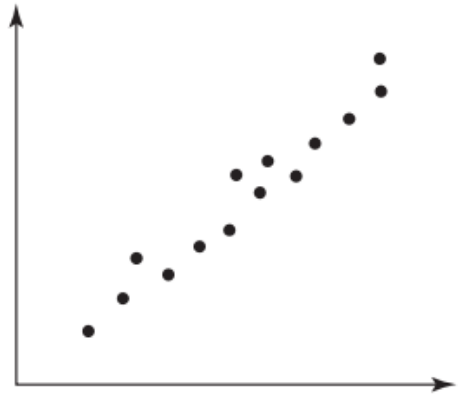
4.2 상관계수

4.3 상관행렬

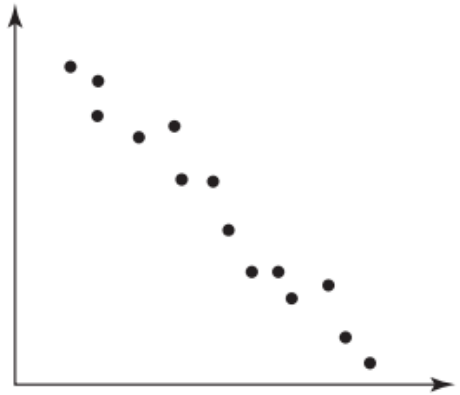
4.4 상관행렬 히트맵

4.1 산점도 그래프

- 산점도 그래프(scatter plot) : 변수와 변수 간의 관계 시각화에 유용한 그래프
  - 데이터를 X축과 Y축에 점으로 표현
    - 두 변수가 양의 선형적 상관관계를 가지고 있을 때 → 정비례
    - 반비례 : 두 변수가 음의 선형적 상관관계를 가지고 있을 때 → 반비례
    - 두 변수가 독립적일 때



[그림 3-11] 양의 선형적 상관관계



[그림 3-12] 음의 선형적 상관관계

[그림 3-13] 두 변수가 독립적

4.1 산점도 그래프

◆ 학생들의 영어점수와 수학점수가 관련성이 있는지 알아보기 (계속)

```
import matplotlib
from matplotlib import font_manager,rc
import matplotlib.pyplot as plt
%matplotlib inline

#  #한글 폰트 등록
# font_location = "c:/Windows/fonts/malgun.ttf"
# font_name = font_manager.FontProperties(fname=font_location).get_name()
# matplotlib.rc('font', family=font_name)

import warnings
warnings.filterwarnings(action='ignore')
```

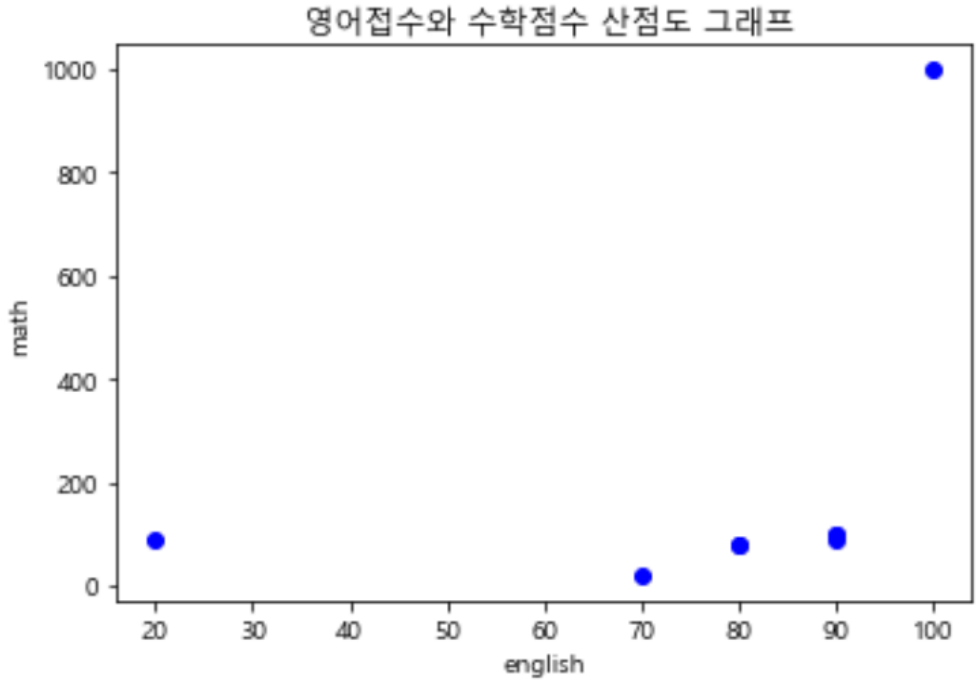
```
# 데이터 준비
import pandas as pd

students = pd.read_csv("data/students.csv")
students.head()
```

	english	math	class
0	100	999	1
1	90	90	1
2	80	80	1
3	70	20	1
4	20	90	2

4.1 산점도 그래프

◆ 학생들의 영어점수와 수학점수가 관련성이 있는지 알아보기

```
plt.plot( students["english"], students["math"], 'bo')
plt.xlabel('english')
plt.ylabel('math')
plt.title(" 영어접수와 수학점수 산점도 그래프")
plt.show()
```



4.1 산점도 그래프

- iris의 sepal_width와 petal_width가 관련성이 있는지 알아보기(계속)

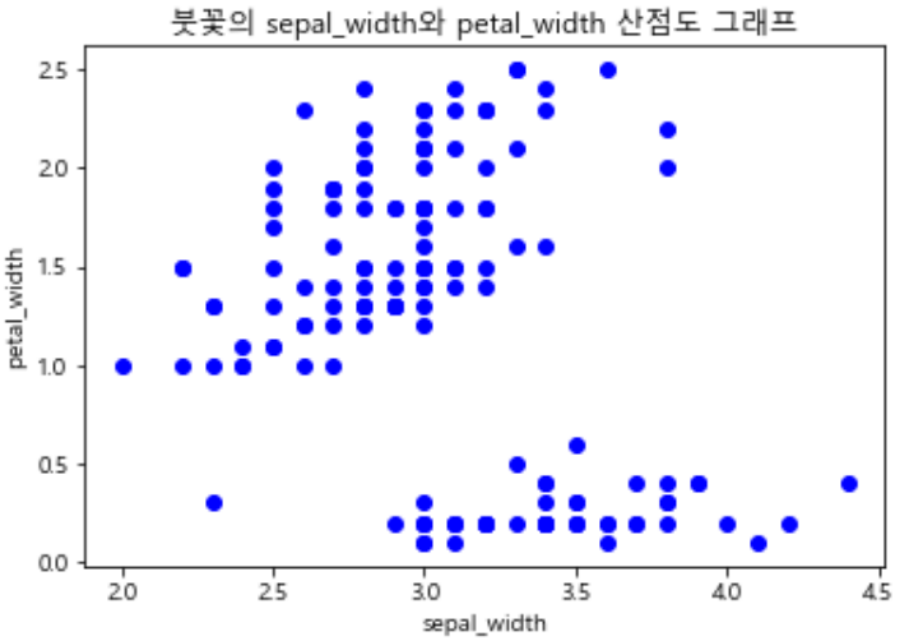
```
import seaborn as sns
# 데이터 준비
iris = sns.load_dataset("iris")  # seaborn 패키지의 샘플 데이터
iris.head()
```

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa

4.1 산점도 그래프

- iris의 sepal_width와 petal_width가 관련성이 있는지 알아보기(계속)

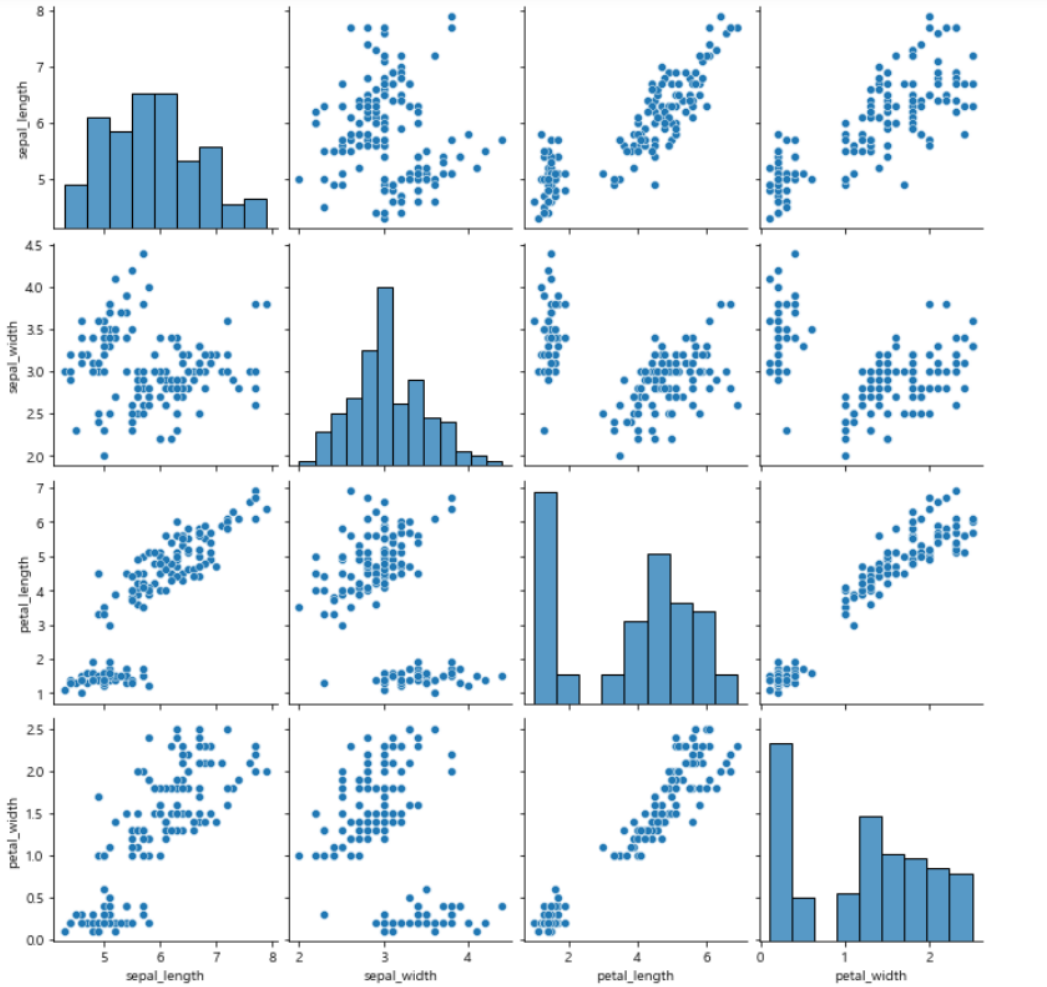
```
plt.plot( iris["sepal_width"], iris["petal_width"], 'bo')
plt.xlabel('sepal_width')
plt.ylabel('petal_width')
plt.title("붓꽃의 sepal_width와 petal_width 산점도 그래프")
plt.show()
```



4.1 산점도 그래프

- pairplot() 함수 : 여러 가지 변수의 산점도 그래프를 한눈에 볼 수 있도록 작성

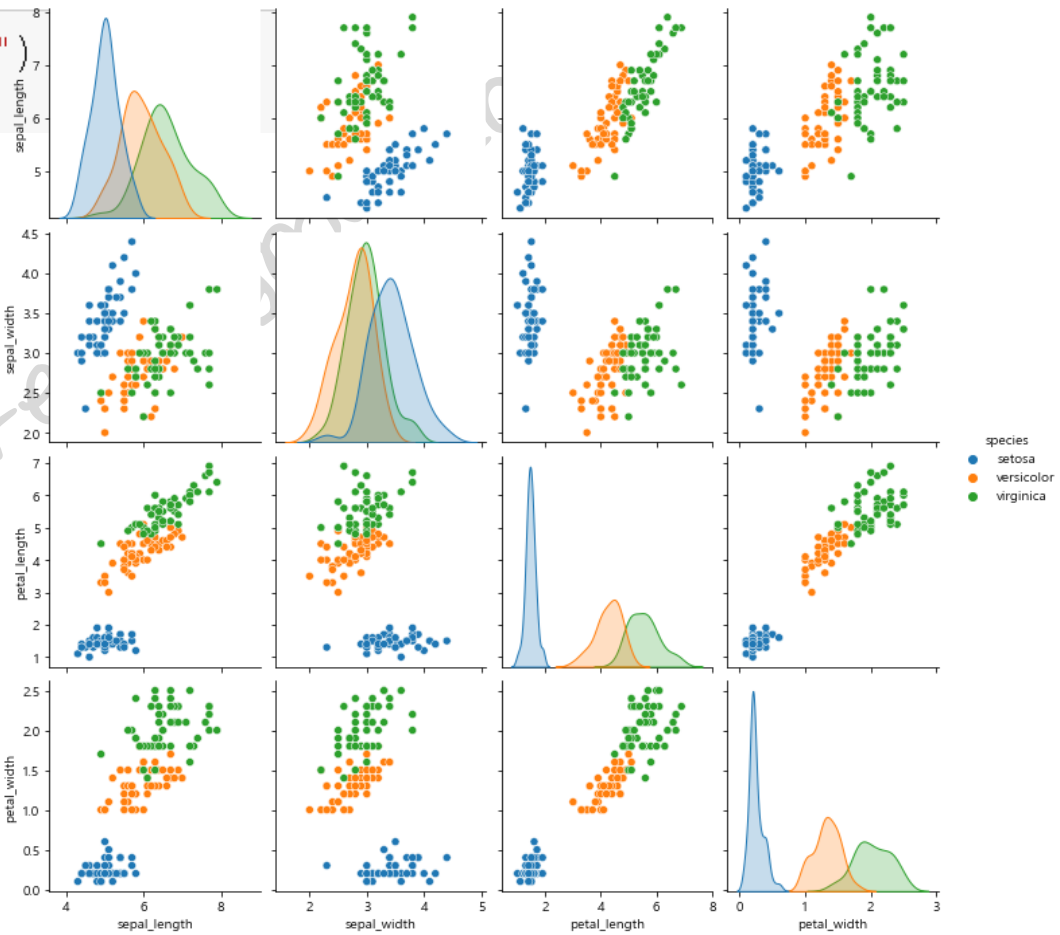
```
sns.pairplot(iris)
plt.show()
```



4.1 산점도 그래프

- pairplot() 함수 : 여러 가지 변수의 산점도 그래프를 한눈에 볼 수 있도록 작성

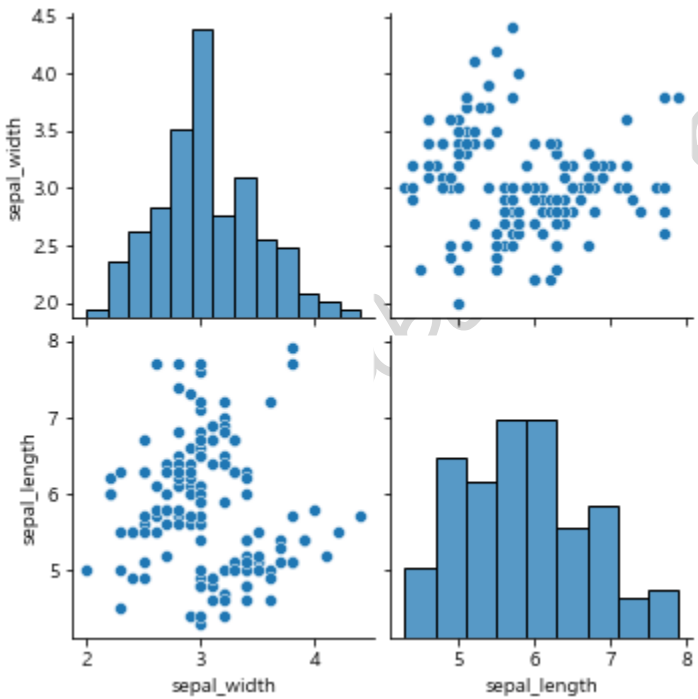
```
sns.pairplot(iris, hue="species")
plt.show()
```



4.1 산점도 그래프

- pairplot() 함수 : 여러 가지 변수의 산점도 그래프를 한눈에 볼 수 있도록 작성

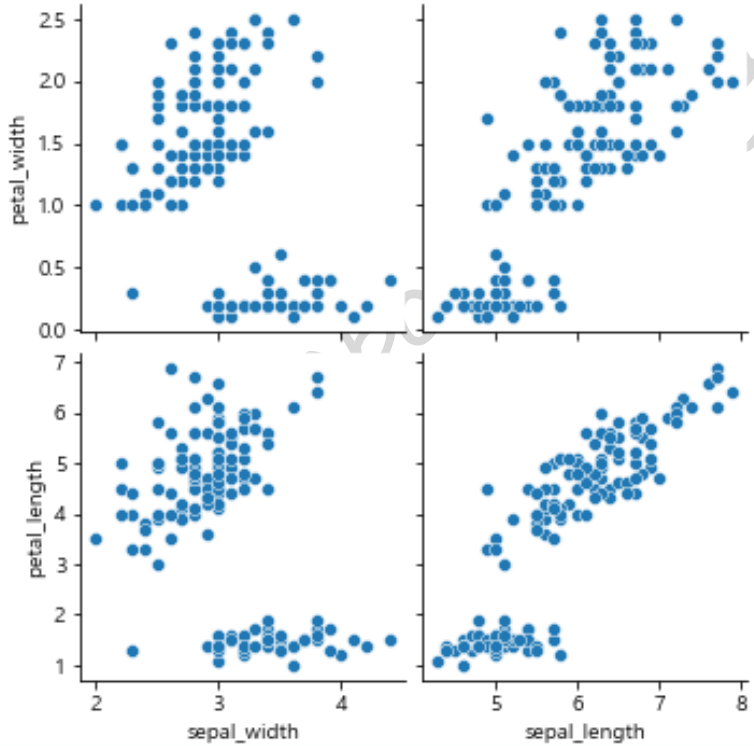
```
sns.pairplot(iris, vars=["sepal_width", "sepal_length"])
plt.show()
```



4.1 산점도 그래프

- pairplot() 함수 : 여러 가지 변수의 산점도 그래프를 한눈에 볼 수 있도록 작성

```
sns.pairplot(iris, x_vars=["sepal_width", "sepal_length"],
             y_vars=["petal_width", "petal_length"])
plt.show()
```



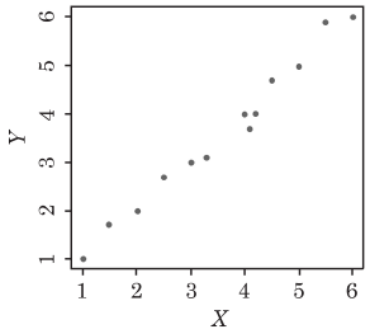
4.2 상관계수

- 상관계수 : 변수 간의 관련성을 수치로 계산
  - 가장 많이 사용하는 계산법 → 피어슨 상관계수
  - 피어슨 상관계수는 -1에서 1사이의 값을 가짐
  - 피어슨 상관계수값이 1에 가까울수록 양의 상관관계
  - -1에 가까울수록 음의 상관관계
  - 두 변수가 독립일 때 0에 가까운 수를 갖음

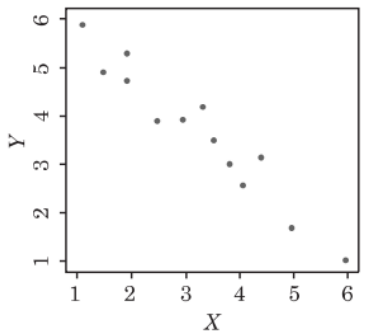
4.2 피어슨 상관계수

- 두 변수 간의 선형적인 관계를 측정하여 [-1, 1] 사이 의 값을 갖음
  - 1 또는 -1에 가까울수록 뚜렷한 선형적인 상관관계
  - 양의 상관관계 : 1에 가까울수록 한 변수의 값이 커지면 다른 변수의 값도 커짐
  - 음의 상관관계 : -1에 가까울수록 한 변수의 값이 커지면 다른 변 수의 값은 작아지짐
  - 선형적인 상관관계가 없는 경우 : 0에 가까운 값
  - 피어슨 상관계수의 정의 : $\rho(X, Y) = \frac{\cos(X, Y)}{\sigma_x \sigma_y}$

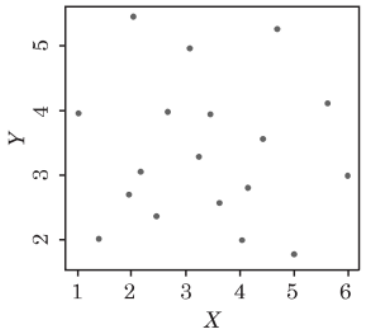
$\cos(X, Y)$ 는 X와 Y의 공분산, σ_x, σ_y 는 X와 Y의 표준편차



[그림 5-1] 양(+)의 상관관계일 때 산점도 그래프



[그림 5-2] 음(-)의 상관관계일 때 산점도 그래프



[그림 5-3] 독립관계일 때 산점도 그래프

4.2 스피어만 상관계수

- 스피어만 상관계수 : 데이터가 서열척도인 경우 즉, 자료의 값 대신 순위를 이용하는 경우의 상관계수
  - 장점 : 스피어만 상관계수는 비선형 관계의 연관성을 파악할 수 있음
  - 데이터를 작은 것부터 차례로 순위를 매겨 서열순서(순위)를 이용해 상관계수를 구함
  - -1과 1 사이의 값을 가짐

두 변수 안의 순위가 완전히 일치하면 +1

두 변수의 순위가 완전히 반대이면 -1

- 두 변수 간 연관관계 유무 확인
- 데이터에 이상점이 있거나 표본 크기가 작을 때 유용
- 두 변수 간의 스피어만 상관계수는 두 변수의 순위값 사이의 피어슨 상관계수와 같음

4.3 상관행렬

- 상관행렬(Correlation Matrix Heatmap) : 여러 변수 간의 상관계수 값으로 생성한 행렬
- pandas의 dataframe 객체의 corr() 함수로 간단히 상관행렬 구하기

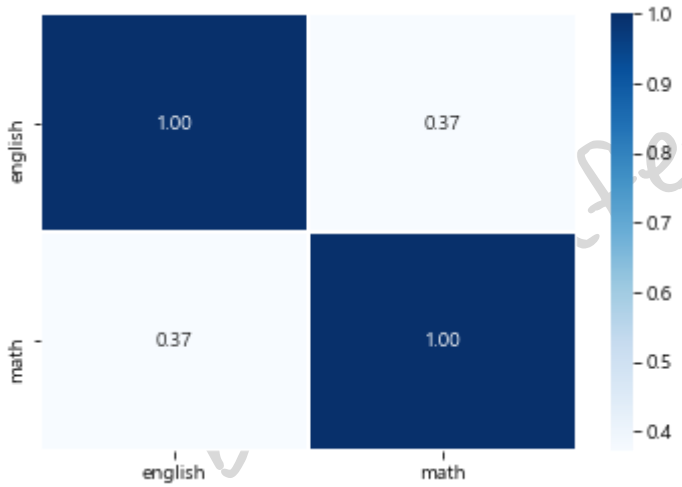
```
students.loc[:, ("english" ,"math" ) ].corr(method='pearson')
```

	english	math
english	1.000000	0.372093
math	0.372093	1.000000

4.4 상관행렬 히트맵

- 많은 변수로 상관행렬을 만들면 상관계수 값을 일일이 확인하기 어려운 경우

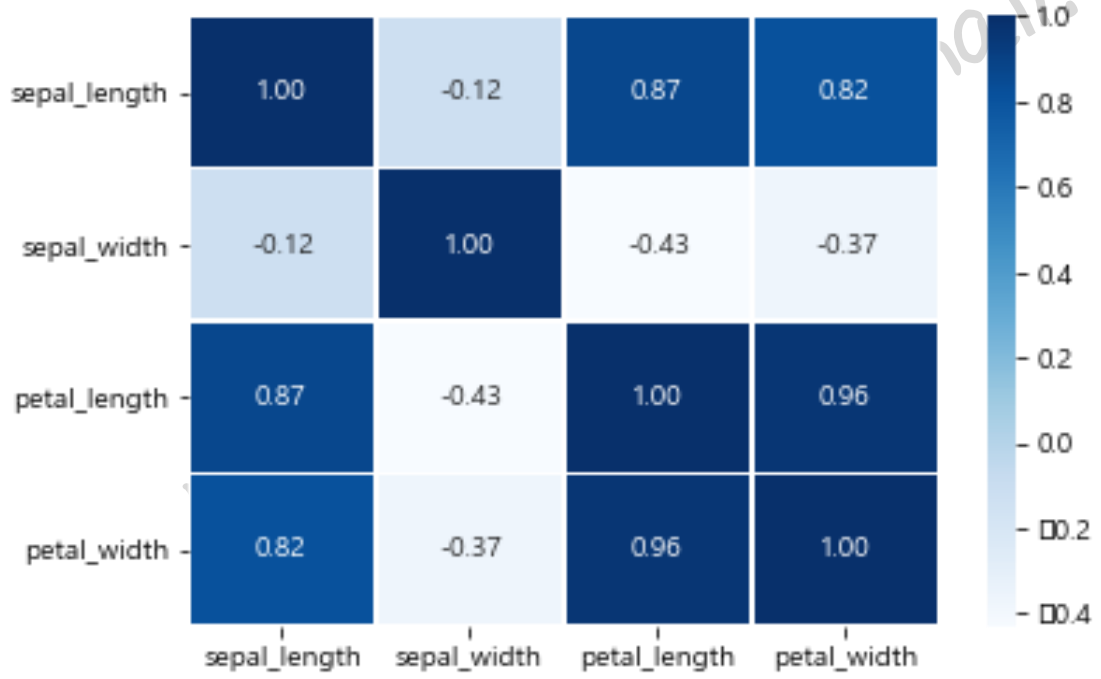
```
sns.heatmap(data = students.corr(), annot=True,
            fmt = '.2f', linewidths=.5, cmap='Blues')
```



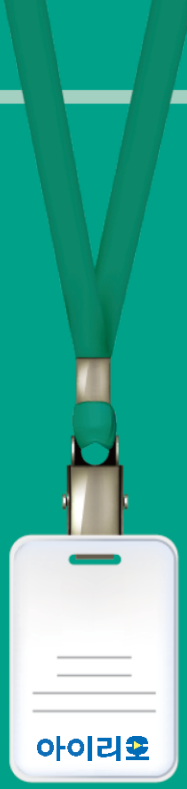
4.4 상관행렬 히트맵

- 많은 변수로 상관행렬을 만들면 상관계수 값을 일일이 확인하기 어려운 경우

```
sns.heatmap(data = iris.corr(), annot=True,
            fmt = '.2f', linewidths=.5, cmap='Blues')
```



Summary



수고하셨습니다.