패키지

모듈 : 함수들이 묶여진 하나의 .py 파일
패키지 : 여러개의 모듈을 그룹화한 것

패키지 설치

패키지/모듈을 설치할 때는 import 구문을 사용한다.

ex) import numpy
import pandas

별칭

↳ 패키지의 활용을 쉽게 하기 위해 사용

ex) import numpy as np (numpy → 데이터 분석용 패키지)
import seaborn as sns (seaborn → 시각화 패키지)

Seaborn 패키지

import seaborn as sns

데이터 불러오기

sns.load_dataset()
↳ parameter의 데이터를 반환한다.

ex) df = sns.load_dataset('titanic')

데이터 시각화

sns.countplot(data = dataframe명, x = x축 [, hue = 추가할 범주형 자료])
↳ 범주형 자료를 시각화 한다.

ex) sns.countplot(data = df, x = 'sex')

hue 파라미터를 사용하면, 색을 더 세분화 가능하다.
ex) sns.countplot(data = df, x = 'class', hue = 'alive')

pydataset

```
import pydataset

df = pydataset.data('mtcars')
```

pandas

```
import pandas as pd

df = pd.DataFrame({'name' : ['주재원', '000', '000'],
                   'Age'  : [21, 20, 19],
                   'math' : [90, 80, 70]})
```

변수(column)가 된다.

Dictionary 형태로 만든다.

	name	Age	math
0	주재원	21	90
1	000	20	80
2	000	19	70

데이터 읽어오기

```
pd.read_csv('파일명')
```

파일명을

↳ csv 파일을 읽어오는 함수

ex) exam = pd.read_csv('exam.csv')

데이터 분석 명령어

데이터프레임.head([행수])

↳ 데이터의 앞(default : 5행)부분만 출력

ex) exam.head() → 0~4행 출력
exam.head(10) → 0~9행 출력

데이터프레임.tail([행수])

↳ 데이터의 뒤(default : 5행)부분만 출력

데이터프레임.shape

↳ 데이터의 (행수, 열수)를 튜플로 저장하고 있는 attribute

데이터프레임.info()

↳ 변수 속성을 알려줄
⊥ (columns)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20 entries, 0 to 19
Data columns (total 5 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   id       20 non-null     int64
 1   nclass   20 non-null     int64
 2   math     20 non-null     int64
 3   english  20 non-null     int64
 4   science  20 non-null     int64
dtypes: int64(5)
memory usage: 928.0 bytes
```

정수 64bit로 이루어져 있다.

결측치(누락된 값)는 없다.

데이터프레임.describe([include='all'])

↳ 8개의 요약 통계량을 출력함

출력값	통계량	설명
count	빈도(frequency)	값의 개수
mean	평균(mean)	모든 값을 더해 값의 개수로 나눈 값
std	표준편차(standard deviation)	변수의 값들이 평균에서 떨어진 정도를 나타낸 값
min	최소값(minimum)	가장 작은 값
25%	1사분위수(1st quantile)	하위 25%(4분의 1) 지점에 위치한 값
50%	중앙값(median)	하위 50%(중앙) 지점에 위치한 값
75%	3사분위수(3rd quantile)	하위 75%(4분의 3) 지점에 위치한 값
max	최대값(maximum)	가장 큰 값

	id	nclass	math	english	science
count	20.00000	20.000000	20.000000	20.000000	20.000000
mean	10.50000	3.000000	57.450000	84.900000	59.450000
std	5.91608	1.450953	20.299015	12.875517	25.292968
min	1.00000	1.000000	20.000000	56.000000	12.000000
25%	5.75000	2.000000	45.750000	78.000000	45.000000
50%	10.50000	3.000000	54.000000	86.500000	62.500000
75%	15.25000	4.000000	75.750000	98.000000	78.000000
max	20.00000	5.000000	90.000000	98.000000	98.000000

- include='all' 파라미터를 사용시, unique, top, freq 통계량을 제공한다.

unique : 고유값 빈도
top : 최빈값
freq : 최빈값 빈도

프레임이름[변수명].unique()

↳ unique한 값을 반환

ex) df['Age'].unique() → [45, 20] 반환

프레임이름[변수명].value_counts()

↳ unique한 값의 개수까지 출력