데이터

데이터

데이터

- 데이터 분석

```
학생번호,학년,악력,윗몸일으키기,점수,순위
1,1,40.2,34,15,4
2,1,34.2,14,7,10
3,1,28.8,27,11,7
4,2,39.0,27,14,5
5,2,50.9,32,17,2
6,2,36.5,20,9,9
7,3,36.6,31,13,6
8,3,49.2,37,18,1
9,3,26.0,28,10,8
10,3,47.4,32,16,3
```

- 표 데이터 처리에 특화된 통계분석 라이브러리 Pandas

```
[ ]  # Pandas를 pd라는 이름으로 임포트
     import pandas as pd
```

데이터

데이터

- read_csv 함수를 사용
- 체력 테스트 결과를 담은 csv 파일을 읽어드림

```
[ ] # 학생번호를 인덱스로 csv 파일을 읽어들여, 변수 df에 저장
    df = pd.read_csv('sport_test.csv',
                     index_col='학생번호')
    # 변수 df를 표시
    df
```

```
학생번호,학년,악력,윗몸일으키기,점수,순위
1,1,40.2,34,15,4
2,1,34.2,14,7,10
3,1,28.8,27,11,7
4,2,39.0,27,14,5
5,2,50.9,32,17,2
6,2,36.5,20,9,9
7,3,36.6,31,13,6
8,3,49.2,37,18,1
9,3,26.0,28,10,8
10,3,47.4,32,16,3
```

데이터

데이터

- 체력 테스트 결과를 DataFrame으로 읽어들인 결과

학생번호	학년	악력	윗몸일으키기	점수	순위
1	1	40.2	34	15	4
2	1	34.2	14	7	10
3	1	28.8	27	11	7
4	2	39.0	27	14	5
5	2	50.9	32	17	2
6	2	36.5	20	9	9
7	3	36.6	31	13	6
8	3	49.2	37	18	1
9	3	26.0	28	10	8
10	3	47.4	32	16	3

```
학생번호,학년,악력,윗몸일으키기,점수,순위
1,1,40.2,34,15,4
2,1,34.2,14,7,10
3,1,28.8,27,11,7
4,2,39.0,27,14,5
5,2,50.9,32,17,2
6,2,36.5,20,9,9
7,3,36.6,31,13,6
8,3,49.2,37,18,1
9,3,26.0,28,10,8
10,3,47.4,32,16,3
```

데이터

데이터

- DataFrame에서 악력에 대한 열을 추출하면 1차원 데이터 구조인 Series 반환

```
[ ] df['악력']
```

```
학생번호
1     40.2
2     34.2
3     28.8
4     39.0
5     50.9
6     36.5
7     36.6
8     49.2
9     26.0
10    47.4
Name: 악력, dtype: float64
```

데이터

데이터 크기

- DataFrame의 크기는 shape라는 인스턴스 변수 참조

```
[ ] df.shape

    (10, 5)
```

- 데이터 개수 10 (인덱스(레코드) 수), 변수 개수 5 (컬럼 수)

변수의 종류

질적 변수와 양적 변수

- 질적 변수 : 선택이 필요한 변수, 종류를 구별하기 위한 변수

1. 매우 좋음　2. 좋음　3. 보통　4. 나쁨　5. 매우 나쁨

A형　B형　O형　AB형

- 양적 변수 : 양을 표현하는 변수

변수의 종류

척도 수준

- 질적 변수는 명의 척도와 순서 척도
- 양적 변수는 간격 척도와 비례 척도로 더 세분화
- 명의척도, 순서 척도, 간격 척도, 비례 척도 이 네 가지를 척도 수준

변수의 종류

척도 수준

- 명의 척도
- 명의 척도는 단순히 분류하기 위한 변수로, 학생번호나 전화번호, 성별 등
- 명의 척도의 목적은 구별하는 것이므로 변수의 동일성 여부에만 의미
  - 학생번호 4와 학생번호 8의 대소 관계는 의미가 없음
  - 합과 차를 계산하더라도 의미 있는 결과를 얻을 수 없음

변수의 종류

척도 수준

- 순서 척도
- 순서 척도는 순서 관계나 대소 관계에 의미가 있는 변수로, 성적 순위, 설문조사의 만족도 등
- 성적 순위에서 8등 은 4등보다 순위가 낮으므로 대소 관계에 의미
  - 4등 과 8등의 차이가 8등과 12등의 차이와 동일하다고 비교할 수는 없음
  - 4등은 8등의 2배라고 주장할 수도 없음

변수의 종류

척도 수준

- 간격 척도
- 간격 척도는 대소 관계 와 함께 그 차이에도 의미를 두는 변수
- 연도나 온도를 들 수 있음
- 60°C는 30°C 보다 높은 온도이므로 대소 관계에 의미가 있고
- 그 차이에 해당하는 수치도 의미가 있음
- 그러나 60°C는 30°C보다 2배 높은 온도라고 할 수 없음

변수의 종류

척도 수준

- 비례 척도
- 비례 척도는 대소 관계 , 차이, 비 모두에 의미가 있는 변수로, 길이나 무게 등
- 길이에서 50cm와 100cm의 차이 가 50cm라는 것도, 100cm는 50cm의 2배라는 것도 의미가 있음

변수의 종류

척도 수준

- 비례 척도
- 간격 척도와 비례 척도는 비슷하므로 구별하기 어려울 때가 있음
- 척도를 구별하는 요령이 있음
- 0 이 '없음'을 나타내는지 여부를 판단하면 됨
- 길이에서 0cm는 길이가 없음을 나타내지만, 온도에서 0°C는 온도가 없다는 뜻이 아님

변수의 종류

척도 수준

- 척도 수준

척도	예	대소 관계	차이	비
명의 척도	학생번호	X	X	X
순서 척도	성적 순위	O	X	X
간격 척도	온도	O	O	X
비려 척도	키	O	O	O

변수의 종류

이산형 변수와 연속성 변수

- 이산형 변수
  - 하나하나의 값을 취하는 변수
  - 서로 인접한 숫자 사이에 값이 존재하지 않음
  - 주사위의 눈, 결석 횟수, 결석 학생 수
- 연속형 변수
  - 연속적인 값을 취할 수 있는 변수
  - 어떤 두 숫자 사이에도 반드시 숫자가 존재
  - 길이, 무게, 시간

정리

체력 테스트의 예

- 체력 테스트의 예에서 변수가 어떻게 분류되는지 생각해 보기

학생번호	학년	악력	윗몸일으키기	점수	순위
1	1	40.2	34	15	4
2	1	34.2	14	7	10
3	1	28.8	27	11	7
4	2	39	27	14	5
5	2	50.9	32	17	2
6	2	36.5	20	9	9
7	3	36.6	31	13	6
8	3	49.2	37	18	1
9	3	26	28	10	8
10	3	47.4	32	16	3

정리

정리

- 데이터
- 데이터의 크기
- 변수의 종류