데이터 정제

- 결측치, 이상치, 중복값 관리

최수연 교수

mibm400@hanmail.net

학습목표

- 데이터 정제에 대해 설명할 수 있다.
- 결측치를 찾고 편집할 수 있다.
- 이상치를 확인하고 변환할 수 있다.
- 중복값을 찾고 제거 할 수 있다.

목차

- 데이터 정제의 정의
- 결측치 확인 및 편집
- 이상치 확인 및 편집
- 중복값 확인 및 삭제

데이터 정제란?

- 오류 데이터

✓ 데이터 수집 과정에서 누락되거나 범위에서 벗어나는 데이터

✓ 데이터 분석 결과가 왜곡되어 신뢰할 수 없음

- 데이터 정제 : 잘못된 데이터를 찾아서 오류를 수정하는 것

- 데이터 정제가 필요한 데이터 종류

✓ 결측치 : 특정 행 및 열에 누락된 데이터

✓ 이상치 : 정상 범위에서 벗어난 데이터

✓ 중복값 : 데이터 내에 동일 데이터가 존재

결측치

- 누락된 데이터, NaN(Not a Number), “?”, “-” 등으로 표시됨

- 결측치가 발생하는 경우

✓ 데이터가 수집되지 않은 경우

✓ 측정 장치의 고장 및 사고로 확보할 수 없을 때

- 결측치를 찾는 이유

✓ 함수 적용이 안되는 경우 발생

✓ 분석 결과가 왜곡됨

✓ 데이터 분석 결과를 신뢰할 수 없음

```
import pandas  as pd

score = pd.read_excel('/content/성적.xlsx')
score
```

	ID	성별	출석	프로젝트	실험
0	S1233	남	2.0	12.0	25.0
1	S1244	남	17.0	2.0	22.0
2	NaN	NaN	NaN	NaN	NaN
3	S1254	남	2.0	16.0	25.0
4	S1256	3	NaN	11.0	25.0
5	S1384	남	12.0	22.0	12.0
6	S1391	여	18.0	25.0	8.0
7	S1393	남	2.0	25.0	NaN
8	S1399	남	NaN	2.0	17.0

결측치 찾기

- 변수명.info()

✓ DataFrame의 행/열 정보를 통해서 확인

```
score.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32 entries, 0 to 31
Data columns (total 5 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   ID      31 non-null     object
 1   성별      31 non-null     object
 2   출석      25 non-null     float64
 3   프로젝트     25 non-null     float64
 4   실험      27 non-null     float64
dtypes: float64(3), object(2)
memory usage: 1.4+ KB
```

ID: 1개
성별 : 1개
출석 : 7개
프로젝트 : 7개
실험 : 5개

결측치 찾기

- 변수명['열이름'].value_counts(dropna=False)

✓ 데이터의 빈도수를 활용하여 확인

✓ dropna=False : NaN인 데이터를 포함하여 데이터 빈도수 알려줌

```
score['출석'].value_counts(dropna=False)
```

```
2.0     9
NaN     7
19.0    5
18.0    3
16.0    3
17.0    2
12.0    1
9.0     1
15.0    1
Name: 출석, dtype: int64
```

```
score['프로젝트'].value_counts(dropna=False)
```

```
NaN     7
12.0    4
2.0     4
16.0    4
25.0    4
11.0    2
17.0    2
22.0    1
18.0    1
23.0    1
24.0    1
5.0     1
Name: 프로젝트, dtype: int64
```

결측치 찾기(df.isna())

- 변수명.isna()

✓ 테이블 전체에서 결측치 찾기

✓ 결측치 유무를 True/False 값으로 반환

```
score.isna()
```

	ID	성별	출석	프로젝트	실험
0	False	False	False	False	False
1	False	False	False	False	False
2	True	True	True	True	True
3	False	False	False	False	False
4	False	False	True	False	False
5	False	False	False	False	False
6	False	False	False	False	False
7	False	False	False	False	True

결측치 찾기(df.isna())

- 변수명['열이름'].**isna**()

✓ 특정열에서 결측치 찾기

```
score['프로젝트'].isna()
```

```
0    False
1    False
2     True
3    False
4    False
5    False
6    False
7    False
8    False
9    False
```

```
score['실험'].isna()
```

```
0    False
1    False
2     True
3    False
4    False
5    False
6    False
7     True
8    False
9    False
```

```
score[['출석','실험']].isna()
```

	출석	실험
0	False	False
1	False	False
2	True	True
3	False	False
4	True	False
5	False	False
6	False	False
7	False	True
8	True	False
9	False	False

결측치 찾기(df.isna())

- pd.isna(변수명).sum()

✓ 각 열의 결측치 개수 확인

```
score.isna().sum()
```

```
ID        1
성별        1
출석        7
프로젝트      7
실험        5
dtype: int64
```

결측값 제거하기(df.dropna())

```
import pandas  as pd

score2 = pd.read_excel('/content/성적2.xlsx')
score2
```

	ID	성별	출석	프로젝트	실험
0	S1233	남	2.0	12.0	25.0
1	S1244	남	17.0	2.0	22.0
2	NaN	NaN	NaN	NaN	NaN
3	S1254	남	2.0	16.0	25.0
4	S1256	3	NaN	11.0	25.0
5	S1384	남	12.0	22.0	12.0
6	S1391	여	18.0	25.0	8.0
7	S1393	남	NaN	25.0	NaN
8	S1399	남	NaN	2.0	17.0
9	S1411	여	19.0	12.0	17.0
10	S1411	여	19.0	12.0	17.0
11	S1414	여	18.0	NaN	23.0
12	S1421	남	9.0	12.0	1.0
13	S1424	여	19.0	2.0	15.0
14	S1428	남	19.0	NaN	2.0

```
score2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 15 entries, 0 to 14
Data columns (total 5 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   ID      14 non-null     object
 1   성별      14 non-null     object
 2   출석      11 non-null     float64
 3   프로젝트     12 non-null     float64
 4   실험      13 non-null     float64
dtypes: float64(3), object(2)
memory usage: 728.0+ bytes
```

결측값 제거하기(df.dropna())

- df.dropna(subset=['열이름1','열이름2'])

✓ subset=['열이름'] : 특정열에 NaN이 존재하는 행만 선택하여 삭제

✓ subset=['열이름1','열이름2']) :열이름1 또는 열이름2에 NaN이 있는 모든 행 삭제

– 단점 : 분석에 필요한 데이터도 삭제될 수 있음

```
score2.dropna(subset=['출석'])
```

	ID	성별	출석	프로젝트	실험
0	S1233	남	2.0	12.0	25.0
1	S1244	남	17.0	2.0	22.0
3	S1254	남	2.0	16.0	25.0
5	S1384	남	12.0	22.0	12.0
6	S1391	여	18.0	25.0	8.0
9	S1411	여	19.0	12.0	17.0
10	S1411	여	19.0	12.0	17.0
11	S1414	여	18.0	NaN	23.0
12	S1421	남	9.0	12.0	1.0
13	S1424	여	19.0	2.0	15.0
14	S1428	남	19.0	NaN	2.0

```
score2.dropna(subset=['프로젝트','실험'])
```

	ID	성별	출석	프로젝트	실험
0	S1233	남	2.0	12.0	25.0
1	S1244	남	17.0	2.0	22.0
3	S1254	남	2.0	16.0	25.0
4	S1256	3	NaN	11.0	25.0
5	S1384	남	12.0	22.0	12.0
6	S1391	여	18.0	25.0	8.0
8	S1399	남	NaN	2.0	17.0
9	S1411	여	19.0	12.0	17.0
10	S1411	여	19.0	12.0	17.0
12	S1421	남	9.0	12.0	1.0
13	S1424	여	19.0	2.0	15.0

결측값 제거하기(df.dropna())

- df.dropna(axis=0, how='any', thresh=개수, inplace=True)

✓ 행과 열에 존재하는 결측치를 선택하여 삭제

✓ axis=0 : 행과 열을 선택하여 삭제

  - axis=0 : 행 삭제
  - axis=1 : 열 삭제

✓ how='any': 결측치의 포함 정도에 따라 삭제

  - how='any': 하나라도 포함하면 행/열 삭제
  - how='all' : 모두 포함하면 행/열 삭제

✓ thresh=개수 : 유효한 데이터가 존재하는'개수'이상만 남기고 삭제

✓ inplace=True : 원본 데이터에 반영

결측값 제거하기(df.dropna())

	ID	성별	출석	프로젝트	실험
0	S1233	남	2.0	12.0	25.0
1	S1244	남	17.0	2.0	22.0
2	NaN	NaN	NaN	NaN	NaN
3	S1254	남	2.0	16.0	25.0
4	S1256	3	NaN	11.0	25.0
5	S1384	남	12.0	22.0	12.0
6	S1391	여	18.0	25.0	8.0
7	S1393	남	NaN	25.0	NaN
8	S1399	남	NaN	2.0	17.0
9	S1411	여	19.0	12.0	17.0
10	S1411	여	19.0	12.0	17.0
11	S1414	여	18.0	NaN	23.0
12	S1421	남	9.0	12.0	1.0
13	S1424	여	19.0	2.0	15.0
14	S1428	남	19.0	NaN	2.0

```
score2.dropna(axis=0)  score2.dropna(axis=0, how='any')
```

	ID	성별	출석	프로젝트	실험
0	S1233	남	2.0	12.0	25.0
1	S1244	남	17.0	2.0	22.0
3	S1254	남	2.0	16.0	25.0
5	S1384	남	12.0	22.0	12.0
6	S1391	여	18.0	25.0	8.0
9	S1411	여	19.0	12.0	17.0
10	S1411	여	19.0	12.0	17.0
12	S1421	남	9.0	12.0	1.0
13	S1424	여	19.0	2.0	15.0

결측값 제거하기(df.dropna())

	ID	성별	출석	프로젝트	실험
0	S1233	남	2.0	12.0	25.0
1	S1244	남	17.0	2.0	22.0
2	NaN	NaN	NaN	NaN	NaN
3	S1254	남	2.0	16.0	25.0
4	S1256	3	NaN	11.0	25.0
5	S1384	남	12.0	22.0	12.0
6	S1391	여	18.0	25.0	8.0
7	S1393	남	NaN	25.0	NaN
8	S1399	남	NaN	2.0	17.0
9	S1411	여	19.0	12.0	17.0
10	S1411	여	19.0	12.0	17.0
11	S1414	여	18.0	NaN	23.0
12	S1421	남	9.0	12.0	1.0
13	S1424	여	19.0	2.0	15.0
14	S1428	남	19.0	NaN	2.0

```
score2.dropna(axis=0, how='any')
```

	ID	성별	출석	프로젝트	실험
0	S1233	남	2.0	12.0	25.0
1	S1244	남	17.0	2.0	22.0
3	S1254	남	2.0	16.0	25.0
5	S1384	남	12.0	22.0	12.0
6	S1391	여	18.0	25.0	8.0
9	S1411	여	19.0	12.0	17.0
10	S1411	여	19.0	12.0	17.0
12	S1421	남	9.0	12.0	1.0
13	S1424	여	19.0	2.0	15.0

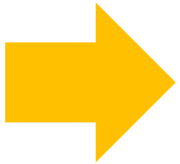
```
score2.dropna(axis=0, how='all')
```

	ID	성별	출석	프로젝트	실험
0	S1233	남	2.0	12.0	25.0
1	S1244	남	17.0	2.0	22.0
3	S1254	남	2.0	16.0	25.0
4	S1256	3	NaN	11.0	25.0
5	S1384	남	12.0	22.0	12.0
6	S1391	여	18.0	25.0	8.0
7	S1393	남	NaN	25.0	NaN
8	S1399	남	NaN	2.0	17.0
9	S1411	여	19.0	12.0	17.0
10	S1411	여	19.0	12.0	17.0
11	S1414	여	18.0	NaN	23.0
12	S1421	남	9.0	12.0	1.0
13	S1424	여	19.0	2.0	15.0
14	S1428	남	19.0	NaN	2.0

결측값 제거하기(df.dropna())

```
score2.dropna(axis=0, thresh=4)
```

	ID	성별	출석	프로젝트	실험
0	S1233	남	2.0	12.0	25.0
1	S1244	남	17.0	2.0	22.0
2	NaN	NaN	NaN	NaN	NaN
3	S1254	남	2.0	16.0	25.0
4	S1256	3	NaN	11.0	25.0
5	S1384	남	12.0	22.0	12.0
6	S1391	여	18.0	25.0	8.0
7	S1393	남	NaN	25.0	NaN
8	S1399	남	NaN	2.0	17.0
9	S1411	여	19.0	12.0	17.0
10	S1411	여	19.0	12.0	17.0
11	S1414	여	18.0	NaN	23.0
12	S1421	남	9.0	12.0	1.0
13	S1424	여	19.0	2.0	15.0
14	S1428	남	19.0	NaN	2.0



	ID	성별	출석	프로젝트	실험
0	S1233	남	2.0	12.0	25.0
1	S1244	남	17.0	2.0	22.0
3	S1254	남	2.0	16.0	25.0
4	S1256	3	NaN	11.0	25.0
5	S1384	남	12.0	22.0	12.0
6	S1391	여	18.0	25.0	8.0
8	S1399	남	NaN	2.0	17.0
9	S1411	여	19.0	12.0	17.0
10	S1411	여	19.0	12.0	17.0
11	S1414	여	18.0	NaN	23.0
12	S1421	남	9.0	12.0	1.0
13	S1424	여	19.0	2.0	15.0
14	S1428	남	19.0	NaN	2.0

결측치 치환하기(df.fillna())

- 데이트의 수가 적고 결측치가 많을 때 활용
- df.interpolate(): 앞/뒤 행의 중간값으로 치환

```
import pandas  as pd
score3 = pd.read_excel('/content/성적3.xlsx')
score3
```

	ID	성별	출석	프로젝트	실험
0	S1233	1	2.0	12.0	25.0
1	S1244	1	17.0	2.0	22.0
2	S1254	1	2.0	16.0	25.0
3	S1256	3	NaN	11.0	25.0
4	S1384	1	12.0	22.0	12.0
5	S1391	2	18.0	25.0	8.0
6	S1393	1	2.0	25.0	NaN
7	S1399	1	NaN	2.0	17.0
8	S1411	2	19.0	12.0	17.0
9	S1411	2	19.0	12.0	17.0
10	S1414	2	18.0	NaN	23.0
11	S1421	3	9.0	12.0	1.0
12	S1424	2	19.0	2.0	15.0
13	S1428	1	19.0	NaN	2.0
14	S1433	1	19.0	16.0	5.0

score3.interpolate()

	ID	성별	출석	프로젝트	실험
0	S1233	1	2.0	12.0	25.0
1	S1244	1	17.0	2.0	22.0
2	S1254	1	2.0	16.0	25.0
3	S1256	3	7.0	11.0	25.0
4	S1384	1	12.0	22.0	12.0
5	S1391	2	18.0	25.0	8.0
6	S1393	1	2.0	25.0	12.5
7	S1399	1	10.5	2.0	17.0
8	S1411	2	19.0	12.0	17.0
9	S1411	2	19.0	12.0	17.0
10	S1414	2	18.0	12.0	23.0
11	S1421	3	9.0	12.0	1.0
12	S1424	2	19.0	2.0	15.0
13	S1428	1	19.0	9.0	2.0
14	S1433	1	19.0	16.0	5.0

결측치 치환하기(df.fillna())

- df.fillna(method='ffill') : 결측치가 있는 앞/뒤 행의 값으로 치환

✓ method='ffill': 앞 행의 값으로 치환

✓ method='bfill': 뒤 행의 값으로 치환

	ID	성별	출석	프로젝트	실험
0	S1233	1	2.0	12.0	25.0
1	S1244	1	17.0	2.0	22.0
2	S1254	1	2.0	16.0	25.0
3	S1256	3	NaN	11.0	25.0
4	S1384	1	12.0	22.0	12.0
5	S1391	2	18.0	25.0	8.0
6	S1393	1	2.0	25.0	NaN
7	S1399	1	NaN	2.0	17.0
8	S1411	2	19.0	12.0	17.0
9	S1411	2	19.0	12.0	17.0
10	S1414	2	18.0	NaN	23.0
11	S1421	3	9.0	12.0	1.0
12	S1424	2	19.0	2.0	15.0
13	S1428	1	19.0	NaN	2.0
14	S1433	1	19.0	16.0	5.0

```
score3.fillna(method='ffill')
```

	ID	성별	출석	프로젝트	실험
0	S1233	1	2.0	12.0	25.0
1	S1244	1	17.0	2.0	22.0
2	S1254	1	2.0	16.0	25.0
3	S1256	3	2.0	11.0	25.0
4	S1384	1	12.0	22.0	12.0
5	S1391	2	18.0	25.0	8.0
6	S1393	1	2.0	25.0	8.0
7	S1399	1	2.0	2.0	17.0
8	S1411	2	19.0	12.0	17.0
9	S1411	2	19.0	12.0	17.0
10	S1414	2	18.0	12.0	23.0
11	S1421	3	9.0	12.0	1.0
12	S1424	2	19.0	2.0	15.0
13	S1428	1	19.0	2.0	2.0
14	S1433	1	19.0	16.0	5.0

```
score3.fillna(method='bfill')
```

	ID	성별	출석	프로젝트	실험
0	S1233	1	2.0	12.0	25.0
1	S1244	1	17.0	2.0	22.0
2	S1254	1	2.0	16.0	25.0
3	S1256	3	12.0	11.0	25.0
4	S1384	1	12.0	22.0	12.0
5	S1391	2	18.0	25.0	8.0
6	S1393	1	2.0	25.0	17.0
7	S1399	1	19.0	2.0	17.0
8	S1411	2	19.0	12.0	17.0
9	S1411	2	19.0	12.0	17.0
10	S1414	2	18.0	12.0	23.0
11	S1421	3	9.0	12.0	1.0
12	S1424	2	19.0	2.0	15.0
13	S1428	1	19.0	16.0	2.0
14	S1433	1	19.0	16.0	5.0

결측치 치환하기(df.fillna())

- df.fillna(값, inplace=True)

✓ Table에 있는 전체 NaN을 동일값으로 변경

✓ 값 : 변환할 데이터, 데이터가 저장된 변수명, 열의 평균이나 중간값

  - df['math'].mean(): math 열의 평균값으로 대체
  - df.fillna({'학과':'영어영문학과'}) : 학과열의 NaN을 모두 '영어영문학과'로 변환
  - df.fillna({'학과':'영어영문학과', '참여횟수':15})
    - 학과열의 NaN을 모두 '영어영문학과'로 변환하고 '참여횟수'를 15로 변환

결측치 치환하기(df.fillna())

	ID	성별	출석	프로젝트	실험
0	S1233	1	2.0	12.0	25.0
1	S1244	1	17.0	2.0	22.0
2	S1254	1	2.0	16.0	25.0
3	S1256	3	NaN	11.0	25.0
4	S1384	1	12.0	22.0	12.0
5	S1391	2	18.0	25.0	8.0
6	S1393	1	2.0	25.0	NaN
7	S1399	1	NaN	2.0	17.0
8	S1411	2	19.0	12.0	17.0
9	S1411	2	19.0	12.0	17.0
10	S1414	2	18.0	NaN	23.0
11	S1421	3	9.0	12.0	1.0
12	S1424	2	19.0	2.0	15.0
13	S1428	1	19.0	NaN	2.0
14	S1433	1	19.0	16.0	5.0

```
score3.fillna(50)
```

	ID	성별	출석	프로젝트	실험
0	S1233	1	2.0	12.0	25.0
1	S1244	1	17.0	2.0	22.0
2	S1254	1	2.0	16.0	25.0
3	S1256	3	50.0	11.0	25.0
4	S1384	1	12.0	22.0	12.0
5	S1391	2	18.0	25.0	8.0
6	S1393	1	2.0	25.0	50.0
7	S1399	1	50.0	2.0	17.0
8	S1411	2	19.0	12.0	17.0
9	S1411	2	19.0	12.0	17.0
10	S1414	2	18.0	50.0	23.0
11	S1421	3	9.0	12.0	1.0
12	S1424	2	19.0	2.0	15.0
13	S1428	1	19.0	50.0	2.0
14	S1433	1	19.0	16.0	5.0

```
score3.fillna(score3['출석'].mean())
```

	ID	성별	출석	프로젝트	실험
0	S1233	1	2.000000	12.000000	25.000000
1	S1244	1	17.000000	2.000000	22.000000
2	S1254	1	2.000000	16.000000	25.000000
3	S1256	3	13.461538	11.000000	25.000000
4	S1384	1	12.000000	22.000000	12.000000
5	S1391	2	18.000000	25.000000	8.000000
6	S1393	1	2.000000	25.000000	13.461538
7	S1399	1	13.461538	2.000000	17.000000
8	S1411	2	19.000000	12.000000	17.000000
9	S1411	2	19.000000	12.000000	17.000000
10	S1414	2	18.000000	13.461538	23.000000
11	S1421	3	9.000000	12.000000	1.000000
12	S1424	2	19.000000	2.000000	15.000000
13	S1428	1	19.000000	13.461538	2.000000
14	S1433	1	19.000000	16.000000	5.000000

결측치 치환하기(df.fillna())

```
score3.fillna({'출석':10, '프로젝트':20})
```

	ID	성별	출석	프로젝트	실험
0	S1233	1	2.0	12.0	25.0
1	S1244	1	17.0	2.0	22.0
2	S1254	1	2.0	16.0	25.0
3	S1256	3	NaN	11.0	25.0
4	S1384	1	12.0	22.0	12.0
5	S1391	2	18.0	25.0	8.0
6	S1393	1	2.0	25.0	NaN
7	S1399	1	NaN	2.0	17.0
8	S1411	2	19.0	12.0	17.0
9	S1411	2	19.0	12.0	17.0
10	S1414	2	18.0	NaN	23.0
11	S1421	3	9.0	12.0	1.0
12	S1424	2	19.0	2.0	15.0
13	S1428	1	19.0	NaN	2.0
14	S1433	1	19.0	16.0	5.0

	ID	성별	출석	프로젝트	실험
0	S1233	1	2.0	12.0	25.0
1	S1244	1	17.0	2.0	22.0
2	S1254	1	2.0	16.0	25.0
3	S1256	3	10.0	11.0	25.0
4	S1384	1	12.0	22.0	12.0
5	S1391	2	18.0	25.0	8.0
6	S1393	1	2.0	25.0	NaN
7	S1399	1	10.0	2.0	17.0
8	S1411	2	19.0	12.0	17.0
9	S1411	2	19.0	12.0	17.0
10	S1414	2	18.0	20.0	23.0
11	S1421	3	9.0	12.0	1.0
12	S1424	2	19.0	2.0	15.0
13	S1428	1	19.0	20.0	2.0
14	S1433	1	19.0	16.0	5.0

이상치 확인 및 치환하기

- 이상치

✓ 정상 범위에서 벗어난 존재할 수 없는 값 또는 극단적인 값

✓ 예) 성별은 1(남)과 2(여) 값만을 갖는데 그외의 다른 데이터

- 이상치 확인하기

✓ df['열이름'].value_counts().sort_index()

- 이상치 데이터 치환하기

✓ df['열이름'] = df['열이름'].replace('찾는 데이터', '변환할 데이터')

이상치 확인 및 치환하기

```
score4['성별'] = score4['성별'].replace(3, 2)
```

```
score4['성별'].value_counts().sort_index()
```

```
1    8
2    5
3    2
Name: 성별, dtype: int64
```

	ID	성별	출석	프로젝트	실험
0	S1233	1	2.0	12.0	25.0
1	S1244	1	17.0	2.0	22.0
2	S1254	1	2.0	16.0	25.0
3	S1244	3	NaN	11.0	25.0
4	S1384	1	12.0	22.0	12.0
5	S1391	2	18.0	25.0	8.0
6	S1393	1	2.0	25.0	NaN
7	S1399	1	NaN	2.0	17.0
8	S1411	2	19.0	12.0	17.0
9	S1411	2	19.0	12.0	17.0
10	S1414	2	18.0	NaN	23.0
11	S1421	3	9.0	12.0	1.0
12	S1424	2	19.0	2.0	15.0
13	S1428	1	19.0	NaN	2.0
14	S1433	1	19.0	16.0	5.0

	ID	성별	출석	프로젝트	실험
0	S1233	1	2.0	12.0	25.0
1	S1244	1	17.0	2.0	22.0
2	S1254	1	2.0	16.0	25.0
3	S1244	2	NaN	11.0	25.0
4	S1384	1	12.0	22.0	12.0
5	S1391	2	18.0	25.0	8.0
6	S1393	1	2.0	25.0	NaN
7	S1399	1	NaN	2.0	17.0
8	S1411	2	19.0	12.0	17.0
9	S1411	2	19.0	12.0	17.0
10	S1414	2	18.0	NaN	23.0
11	S1421	2	9.0	12.0	1.0
12	S1424	2	19.0	2.0	15.0
13	S1428	1	19.0	NaN	2.0
14	S1433	1	19.0	16.0	5.0

중복값 확인하기

- df.duplicated(['열이름'])

✓ '열이름'을 기준으로 중복되는 행 검출

	ID	성별	출석	프로젝트	실험
0	S1233	1	2.0	12.0	25.0
1	S1244	1	17.0	2.0	22.0
2	S1254	1	2.0	16.0	25.0
3	S1244	3	NaN	11.0	25.0
4	S1384	1	12.0	22.0	12.0
5	S1391	2	18.0	25.0	8.0
6	S1393	1	2.0	25.0	NaN
7	S1399	1	NaN	2.0	17.0
8	S1411	2	19.0	12.0	17.0
9	S1411	2	19.0	12.0	17.0
10	S1414	2	18.0	NaN	23.0
11	S1421	3	9.0	12.0	1.0
12	S1424	2	19.0	2.0	15.0
13	S1428	1	19.0	NaN	2.0
14	S1433	1	19.0	16.0	5.0

```
score4.duplicated(['ID'])
```

```
0     False
1     False
2     False
3      True
4     False
5     False
6     False
7     False
8     False
9      True
10    False
11    False
12    False
13    False
14    False
dtype: bool
```

중복값 제거하기

- df.drop_duplicates(['열이름']) : 열이름을 기준으로 중복 데이터 행 삭제
- df.drop_duplicates(subset=['열이름', '열이름'], keep="last")

✓ subset=['열이름', '열이름']: 2개 열의 데이터가 일치하는 행 삭제

✓ keep=" ": 데이터를 유지할 행 설정

- 기본은 첫 번째 데이터를 유지
- keep="last": 마지막 데이터 유지
- keep=False : 모두 삭제

중복값 제거하기

```
score4.drop_duplicates(subset=['ID','성별'])
```

	ID	성별	출석	프로젝트	실험
0	S1233	1	2.0	12.0	25.0
1	S1244	1	17.0	2.0	22.0
2	S1254	1	2.0	16.0	25.0
3	S1244	3	NaN	11.0	25.0
4	S1384	1	12.0	22.0	12.0
5	S1391	2	18.0	25.0	8.0
6	S1393	1	2.0	25.0	NaN
7	S1399	1	NaN	2.0	17.0
8	S1411	2	19.0	12.0	17.0
9	S1411	2	19.0	12.0	17.0
10	S1414	2	18.0	NaN	23.0
11	S1421	3	9.0	12.0	1.0
12	S1424	2	19.0	2.0	15.0
13	S1428	1	19.0	NaN	2.0
14	S1433	1	19.0	16.0	5.0

```
score4.drop_duplicates(['ID'])
```

	ID	성별	출석	프로젝트	실험
0	S1233	1	2.0	12.0	25.0
1	S1244	1	17.0	2.0	22.0
2	S1254	1	2.0	16.0	25.0
4	S1384	1	12.0	22.0	12.0
5	S1391	2	18.0	25.0	8.0
6	S1393	1	2.0	25.0	NaN
7	S1399	1	NaN	2.0	17.0
8	S1411	2	19.0	12.0	17.0
10	S1414	2	18.0	NaN	23.0
11	S1421	2	9.0	12.0	1.0
12	S1424	2	19.0	2.0	15.0
13	S1428	1	19.0	NaN	2.0
14	S1433	1	19.0	16.0	5.0

	ID	성별	출석	프로젝트	실험
0	S1233	1	2.0	12.0	25.0
1	S1244	1	17.0	2.0	22.0
2	S1254	1	2.0	16.0	25.0
3	S1244	2	NaN	11.0	25.0
4	S1384	1	12.0	22.0	12.0
5	S1391	2	18.0	25.0	8.0
6	S1393	1	2.0	25.0	NaN
7	S1399	1	NaN	2.0	17.0
8	S1411	2	19.0	12.0	17.0
10	S1414	2	18.0	NaN	23.0
11	S1421	2	9.0	12.0	1.0
12	S1424	2	19.0	2.0	15.0
13	S1428	1	19.0	NaN	2.0
14	S1433	1	19.0	16.0	5.0

중복값 제거하기

```
score4.drop_duplicates(subset=['ID','성별'], keep='last')
```

	ID	성별	출석	프로젝트	실험
0	S1233	1	2.0	12.0	25.0
1	S1244	1	17.0	2.0	22.0
2	S1254	1	2.0	16.0	25.0
3	S1244	2	NaN	11.0	25.0
4	S1384	1	12.0	22.0	12.0
5	S1391	2	18.0	25.0	8.0
6	S1393	1	2.0	25.0	NaN
7	S1399	1	NaN	2.0	17.0
9	S1411	2	19.0	12.0	17.0
10	S1414	2	18.0	NaN	23.0
11	S1421	2	9.0	12.0	1.0
12	S1424	2	19.0	2.0	15.0
13	S1428	1	19.0	NaN	2.0
14	S1433	1	19.0	16.0	5.0

```
score4.drop_duplicates(subset=['ID','성별'], keep=False)
```

	ID	성별	출석	프로젝트	실험
0	S1233	1	2.0	12.0	25.0
1	S1244	1	17.0	2.0	22.0
2	S1254	1	2.0	16.0	25.0
3	S1244	2	NaN	11.0	25.0
4	S1384	1	12.0	22.0	12.0
5	S1391	2	18.0	25.0	8.0
6	S1393	1	2.0	25.0	NaN
7	S1399	1	NaN	2.0	17.0
10	S1414	2	18.0	NaN	23.0
11	S1421	2	9.0	12.0	1.0
12	S1424	2	19.0	2.0	15.0
13	S1428	1	19.0	NaN	2.0
14	S1433	1	19.0	16.0	5.0

수고하셨습니다.