**<중복 데이터 처리하기>**

df.duplicated() : 한 개 컬럼을 기준으로 중복 체크

# keep 생략하면 기본 "first"

# keep = 'first' 중복값 중 처음값만 표시

# keep = 'last' 중복값 중 마지막값만 표시

# keep = False 중복되는게 존재하면 모두 True

df.drop\_duplicates(['이용거리'], keep='first') # 첫번째 중복값만 남기고 나머지 중복 데이터 제거하기 # 실행결과는 제거된 후의 정상데이터만 조회시킴

df.drop\_duplicates(['이용거리'], keep='last') # 마지막 중복값들만 남기고 나머지 제거

df.drop\_duplicates(['이용거리'], keep=False) # 중복된 데이터 전체 삭제하기

df2.value\_counts() : 빈도분석

**< 데이터 탐색하기 >**

**1. 데이터 그룹 분석하기**

df.groupby(['class']) # 반별그룹화하기

df.groupby(['class']).head() # 그룹 데이터를 확인하기 위해서 head() 사용

df1.get\_group('A') # 특정 그룹의 데이터 조회하기

df.groupby('class').mean() # 그룹별 평균

df.temp.get\_group(('A','m')) # 여러개 그룹 조회하기. get\_group()을 통해 그룹별 조회를 하기 위해 # 값을 넘겨 줄 떄는 튜플 형태로 괄호() 안에 그룹변수를 지정

## 반별 수학점수의 평균 구하기

df.groupby(['class'])['math'].mean()

df.groupby(['class')].mean()[['math']]

df['math'].groupby(df['class']).mena()

sex\_group.groups # 실제 데이터가 어떻게 구성되어 있는지 확인

sex\_group.get\_group('m') # 그룹 중에 남자인 학생들만 조회

**2. 재구조화**

pd.cut(df['math'],3) # cut() : 동일한 길이(값의 범위)로 3구간으로 나눈다.

df.agg(['count','mean','std','min','max']) # df에 대한 기초통계

pd.qcut(df['math'],3) # qcut() : 동일개수로 나누기

pd.qcut(df['math'], 3, labels=False) # labels=False : 범주의 번호 표시하기

# range(), np.arange()

# 첫번째값 : 시작값

# 두번째값 : 끝 값 -1

# 세번째값 : 증가값(+) or 감소값(-)

for i in range(0, 3, 1):

print(i)

for j in range(3, 0, -1):

print(j)

**3. 원-핫 인코딩**

# 범주형 컬럼에 대해서만 가능

# 범주형이 아닌 컬럼에 대해서 범주형으로 만들면 가능

# get\_dummies() : 숫자값을 제외한 object(문자열) 데이터 타입만 가능

# 해당 one-hot 인코딩 되는 원본 컬럼은 없어지고,

# 해당 '컬럼명\_범주명'으로 컬럼이 만들어짐

df\_onehot = pd.get\_dummies(df)

**4. 데이터 전치**

df.T

# 행과 열의 위치를 바꿈

# 컬럼 -> 행 인덱스로 바뀌고

# 행 인덱스 번호 -> 컬럼명