Pandas 연습문제

1. Iris - sns.load_dataset('iris')

a. 붓꽃 종(species)별로 꽃잎길이(sepal_length), 꽃잎폭(sepal_width),
꽃받침길이(petal_length), 꽃받침폭(petal_width)의 평균, 표준편차 등
기초통계량(describe())을 구하시오.

b. 3분위수(Q3)와 1분위수(Q1)의 차이보다 1.5배가 크거나 작은 데이터는 이상치이다.
즉,

Q1 - 1.5 * (Q3 - Q1) 보다 작은 데이터
Q3 + 1.5 * (Q3 - Q1) 보다 큰 데이터

이 이상치를 제거하고 위의 4가지 항목에 대해서 평균, 표준편차를 구하시오.

2. Titanic - sns.load_dataset('titanic')

a. 타이타닉호의 승객에 대해 나이와 성별에 의한 카테고리 열인 category1 열을 만드시오. category1 카테고리는 다음과 같이 정의됨

   1) 20살이 넘으면 성별을 그대로 사용한다.
   2) 20살 미만이면 성별에 관계없이 “child”라고 한다.

b. 타이타닉호의 승객 중 나이를 명시하지 않은 고객은 나이를 명시한 고객의 평균 나이 값이 되도록 titanic 데이터프레임을 고치시오.

c. 성별, 선실(class)별, 출발지(embark_town)별 생존율을 구하시오.

d. 타이타닉호 승객을 ‘미성년자’, ‘청년’, ‘중년’, ‘장년’, ‘노년’ 나이 그룹으로 나누고, 각 그룹별 생존율을 구하시오.

```
bins = [1, 20, 30, 50, 70, 100]
labels = ["미성년자", "청년", "중년", "장년", "노년"]
```

e. qcut 명령으로 세 개의 나이 그룹을 만들고, 나이 그룹별 남녀 성비와 생존율을 구하시오.

3. Mile Per Gallon - sns.load_dataset('mpg')

a. 배기량(displacement) 대비 마력(horsepower) 열(hp_per_cc)을 추가하시오.

b. name으로부터 manufacturer(제조사)와 모델을 추출하여
새로운 열 manufacturer와 model을 추가하고, name 열은 삭제하시오.

c. 엔진의 실린더(cylinders) 갯수별 연비(mpg)의 평균을 구하시오.

d. 생산지(origin)별 배기량 대비 마력(hp_per_cc)의 평균을 구하시오.

e. 모델이 5개 이상인 제조사에 대하여
연비(mpg)의 평균이 가장 좋은 제조사 Top 5를 구하시오.