데이터 전처리(1)

숙명여자대학교 경영학부 오중산

데이터 전처리의 정의와 함수 소개

- 데이터 전처리란?
  - ◆ 분석에 적합하도록 원자료(raw data)를 가공하는 작업
  - ◆ 실제 데이터 분석 과정에서 가장 많은 시간이 소요되기도 함
- 데이터 전처리에 자주 사용되는 dplyr 함수

dplyr 함수	기능
rename()	변수 이름 바꾸기
filter()	행 추출
select()	열(변수) 추출
arrange()	정렬
mutate()	변수 추가
summarise()	기초 통계량 계산
group_by()	집단별로 나눔
left_join()	열 데이터 합치기
bind_rows()	행 데이터 합치기

데이터 전처리: filter 함수

- filter 함수 소개
  - 조건에 부합하는 사례를 추출할 때 사용
  - 예: gender 변수에서 남학생만 추출
- filter 함수를 이용한 실습
  - exam 데이터 프레임에서 1반 학생만 추출하기
    - exam %>% filter(class == 1)
    - %>%는 파이프 연산자로서 데이터 프레임과 함수를 연결할 때 사용(ctrl + shift + M)
  - 남학생들의 영어시험 분산 구하기
    - 1단계(데이터 프레임 만들기): exam_male <- exam %>% filter(gender == "Male")
    - 2단계(분산 구하기): var(exam_male$english)

데이터 전처리: filter 함수

- exam 데이터 프레임을 활용한 filter 함수 실습
  - 문제1: 1반, 2반, 3반 학생들의 수학시험 평균은 얼마인가?
  - 문제2: 4반이 아닌 학생들 중에서 수학시험이 90점 이상이거나, 역사시험이 95점 이상인 학생들을 추출하여 새로운 데이터 프레임(exam_N4)을 만드시오.
  - 문제3: 영어시험 성적이 상위 10%인 학생들만 추출하시오.
- mpg 데이터 프레임을 활용한 filter 함수 실습
  - 문제4: 배기량이 4이하인 자동차의 고속도로 연비평균과 배기량이 5이상인 자동차의 고속도로연비 평균을 비교하시오.
  - 문제5: 아우디와 도요타의 도심연비 평균을 비교하시오.
  - 문제6: 세 회사(쉐보레, 포드, 혼다 자동차)의 고속도로연비 평균을 구하시오.

데이터 전처리: select 함수

- select 함수 소개
  - 원하는 변수들만 추출할 때 사용
  - 예: 반, 수학점수, 영어점수만 추출하기
- select 함수를 이용한 실습
  - 반, 수학점수, 영어점수만 추출하기
    - exam %>% select(class, math, english) / 여러 변수를 쉼표로 연결함
  - 주소만 제외하고 다른 모든 변수 추출하기
    - exam %>% select(-address) / -표시를 하면 해당 변수를 제외한다는 의미
  - 참고: 특정 단어가 포함된 변수 추출하기
    - df %>% select(contains("특정 단어"))

데이터 전처리: select 함수

- exam 데이터 프레임을 활용한 select 함수 실습
  - 문제7: 1반 학생들만을 대상으로 성별과 수학점수를 추출
- mpg 데이터 프레임을 활용한 select 함수 실습
  - 문제8: mpg에서 class와 city 두 개 변수만을 추출하여 새로운 데이터 프레임(mpg_cc)을 만드시오.
  - 문제9: 위에서 만든 데이터 프레임에서 suv 자동차와 compact 자동차의 도시연비 평균을 비교하시오.

데이터 전처리: arrange 함수

- arrange 함수 소개
  - 계량척도로 측정된 변수에 대해 오름차순 혹은 내림차순으로 정렬할 때 사용
- arrange 함수를 이용한 실습
  - 수학점수를 오름차순으로 정렬하기
    - exam %>% arrange(math)
  - 수학점수를 내림차순으로 정렬하기
    - exam %>% arrange(desc(math))
  - 반을 오름차순으로 정렬한 후, 수학점수를 기준으로 내림차순으로 정렬하기
    - exam %>% arrange(class, -math)
  - 문제10: 아우디 모델 중에서 고속도로연비가 가장 높은 상위 3개 모델은?