데이터 전처리(2)

숙명여자대학교 경영학부 오중산

데이터 전처리: mutate 함수

- mutate 함수 소개
  - 기존 변수를 활용하여 새로운 변수를 만들때 사용하는 함수
  - 종종 ifelse 함수와 함께 사용되기도 함
- mutate 함수를 이용한 실습
  - exam에서 세 과목(수학/영어/역사) 점수 합계인 total 변수 만들기
    - exam <- exam %>% mutate(total = math + english + history)
  - exam에서 세 과목(수학/영어/역사) 점수 평균인 average 변수 만들기
    - exam <- exam %>% mutate(average = (math + english + history)/3)
  - 동시에 여러 개의 변수를 만들 수 있음
    - exam <- exam %>% mutate(total = math + english + history, average = (math + english + history)/3)

데이터 전처리: mutate 함수

- mutate 함수와 ifelse 함수 결합한 실습
  - exam에서 test 변수 만들기
    - total 점수가 180점 이상이면 “pass”, 그렇지 않으면 “fail”로 판정하는 파생변수 test 만들기
    - exam <- exam %>% mutate(test = ifelse(total >= 180, “pass”, “fail”))
    - 합격자/불합격자 빈도수 확인하가: table(exam$test)
- mpg 데이터를 이용한 mutate 함수 실습
  - 문제11: 파생변수 sum(통합연비: 도심연비와 고속도로연비 합산)을 만드시오.
  - 문제12: 통합연비의 평균은 얼마인가?
  - 문제13: 파생변수 avg(평균연비: 도심연비와 고속도로연비 평균)를 만들고, 평균연비가 가장 높은 자동차 모델 세 개가 무엇인지 확인하시오.

데이터 전처리: group_by 함수와 summarise 함수

- group_by 함수 소개
  - 사례를 어떤 변수값의 결과를 기준으로 몇 개 집단으로 구분
    - 변수의 척도는 문자형 혹은 범주형인 게 바람직함
  - 예: 반, 성별, 주소에 따른 사례 구분
- summarise 함수 소개
  - 어떤 변수의 기술통계량에 대한 요약결과를 보여줄 때 사용
    - 기술통계량과 함께 빈도수를 보여줄 수 있음
  - 일반적으로 group_by 함수와 함께 사용됨
    - 사례를 몇 개 집단으로 구분한 후, 구분된 집단별로 관심 있는 변수의 기술통계량 제시

데이터 전처리: group_by 함수와 summarise 함수

- group_by 함수와 summarise 함수 실습
  - 반별로 학생수(빈도)와 수학점수의 평균&표준편차 제시
    - exam %>% group_by(class) %>% summarise(n(), mean(math), sd(math))
    - exam %>% group_by(class) %>% summarise(count = n(), mean_math = mean(math), sd_math = sd(math))
      - summarise를 통해 보여주는 결과값에 대해 변수명 지정
  - 반별 그리고 성별로 구분하여 학생수와 역사점수 평균을 요약하여 새로운 데이터 프레임 exam_clshist에 저장하시오.
    - exam_clshist <- exam %>% group_by(class, gender) %>% summarise(count = n(), mean_history = mean(history))
    - 세부 집단별 학생들의 구성비율을 보여주는 새로운 변수(perc) 만들기
      - exam_clshist <- exam_clshist %>% mutate(perc = count / sum(count))

데이터 전처리: group_by 함수와 summarise 함수

- mpg 데이터를 이용한 group_by 함수와 summarise 함수 실습
  - 문제14: 자동차 모델을 제조사별, 그리고 구동방식별로 구분한 후, 도심연비평균과 고속도로연비평균 요약결과를 제시하시오.
  - 문제15: 자동차 모델을 제조사별로 구분한 후, suv 모델에 대해 통합연비 평균 상위 3개 모델을 구하시오.
    - 힌트1: group_by와 summarise 사이에 다른 dplyr 함수가 들어갈 수 있음
    - 힌트2: 상위 몇 개 업체를 제시하라고 하면 arrange 함수와 head 함수를 함께 사용
  - 문제16: 평균 배기량이 가장 높은 세 개 변속기를 제시하시오.
  - 문제17: 4기통 모델을 가장 많이 생산하는 업체 세 곳을 순서대로 구하시오.
    - 힌트: summarise의 내용만 다를 뿐, 문제 15와 유사함