데이터 전처리(3)

숙명여자대학교 경영학부 오중산

데이터 전처리: left_join 함수

- left_join 함수 소개
  - 기존 데이터 프레임에 새로운 데이터 프레임을 합칠 때 사용하는 함수
  - 두 데이터 프레임에서 공통의 변수가 존재해야 함
- left_join 함수 실습
  - 현재 exam 데이터 프레임 검토하여 id 변수를 삭제했다면 복원
    - 1~30까지 열벡터 생성: v1 <- c(1:30)
    - 생성한 열벡터를 id라는 변수로 데이터 프레임에 저장: exam <- exam %>% mutate(id = v1)
    - id의 위치를 맨 앞으로 변경: exam <- exam %>% relocate(id, .before = address)
      - 참고로 어떤 변수 뒤로 이동 시킬 경우에는 relocate(v1, .after = v2)와 같이 지정

데이터 전처리: left_join 함수

- left_join 함수 실습
  - ◆ exam_science.csv 파일 불러와서 같은 이름의 데이터 프레임 생성
  - ◆ exam과 exam_science 합치기
    - exam <- left_join(exam, exam_science, by = "id")
    - by = “var”와 같이 공통 변수를 따옴표로 지정해야 함
- mpg 데이터를 이용한 left_join 함수 실습
  - ◆ 문제18-1: 연료별 가격 데이터 프레임(fuel_price) 만들기
    - fuel_price <- data.frame(fuel = c("CNG", "diesel", "ethanol", "premium", "regular"), fuel_price = c(2.35, 2.38, 2.11, 2.76, 2.22))
  - ◆ 문제18-2: 공통변수인 fuel을 기준으로 mpg와 fuel_price 합치기
    - 공통변수의 척도가 가급적 일치하는 게 바람직함

fuel	fuel_price($/gallon)
CNG	2.35
diesel	2.38
ethanol	2.11
premium	2.76
regular	2.22

데이터 전처리: left_join 함수

- mpg 데이터를 이용한 left_join 함수 실습
  - ◆ 문제19-1: 구동방식별 가격 데이터 프레임(drv_price) 만들기
  - ◆ 문제19-2: mpg와 drv_price 합치기
    - 공통변수가 없는 상황에서 어떻게 합칠 수 있을까?
    - 기본 파라미터: by = c("변수1" = "변수2")
  - ◆ 문제20: fuel_price는 fuel 뒤로, drv_price는 city 앞으로 이동시키기
    - 하나의 명령문으로 작성하기 어렵고 변수별로 나누어서 실행

driving	driving_price($)
4	40000
forward	30000
rear	50000

데이터 전처리: bind_rows 함수

- bind_rows 함수 소개
  - 새로운 사례를 추가할 때 사용하는 함수
  - 새로운 사례(들) 역시 기존 데이터 프레임에 속한 변수로 측정되어야 함
- bind_rows 함수 실습
  - exam_add.csv를 불러와서 동일한 명칭의 데이터 프레임을 만든 후, exam과 통합
    - exam_add에는 science와 average를 제외한 9개 변수에 대한 6개 사례(id = 30~35) 관련 데이터가 저장되어 있음
    - exam_add <- read.csv("exam_add.csv", stringsAsFactors = F)
    - exam <- bind_rows(exam, exam_add)
  - 중복된 id = 30에 대해 하나를 제외할 수 있는 방법
    - exam <- exam %>% distinct(id, total, .keep_all = T)

데이터 전처리: bind_rows 함수

- bind_rows 함수 실습
  - 문제21: 통합된 exam 데이터 프레임에서 id = 34는 id만 다를 뿐, id = 29와 동일한 사례이므로 제거하시오.
  - 문제22: average 변수 측정결과가 NA인 사례에 대해 세 과목에 대한 실제 평균값을 구해서 이 값으로 대체하시오.
  - 문제23: science 변수 측정결과가 NA인 사례에 대해 다른 사례들의 sciene 평균값으로 대체하시오.