2021.02.05

[dplyr 패키지를 이용한 데이터 전처리-II]

4. 데이터 합치기

(1) 열 합치기(가로 합치기): 합치는 열들의 각 행의 수가 다르면 실행이 불가능

1) cbind( ): 합치는 열들의 데이터들이 그대로 합쳐진다. (중복된 열도 그대로 표시)

2) left_join( ): 중복되는 열 중심으로 병합된다. (중복되는 열은 두 번 이상 표시되지 않는다.)

(2) 행 합치기(세로 합치기): rbind( ), bind_rows( )

5. 데이터 정제하기

(1) 결측치 정제하기

1) table(is.na( )): 결측치 수(TRUE) / 결측가 아닌 요소 수(FALSE)

2) na.omit( ): 결측치가 존재하는 행 모두 제거(분석이 필요한 해당 행의 열까지 손실될 우려가 있다.)

3) 요소 위치 **<- NA:** 결측치 인위적으로 설정

4) 결측치를 다른 값으로 대체하기

ex) – exam 데이터의 영어 점수 일부를 결측치로 설정 후, 그 자리를 전체 영어점수의 중앙값으로 대체하기

```
exam[1:2, 'english'] <- NA   # 1,2행의 영어점수를 결측치로 설정
median <- round(apply(exam[3:5], 2, median, na.rm=T))
                                        # median변수에 결측치를 제외한 과목별 점수의 중앙값
exam %>% mutate( english = ifelse(is.na(english), median['english'], english))
                                        # dplyr함수와 ifelse함수로 결측치를 중앙값으로 대체하는 식 만들기
```

(2) 이상치 정제하기

1) 논리적인 이상치

ex) 성별 데이터에서 남자도 여자도 아닌 값, 0~100사이의 점수 데이터에서 100보다 큰 점수 ...

2) 정상범위 기준에서 많이 벗어난 이상치: 상하위 0.3% 또는 평균+3*표준편차

ex) – mpg 데이터에서 고속도로 연비의 이상치 구하기

```
mean(mpg$hwy)+3*sd(mpg$hwy)
mean(mpg$hwy)-3*sd(mpg$hwy)
```