05-2. dplyr패키지를이용한전처리 # src: 7_R/05-2_dplyr패키지를이용한전처리 참고

1) 외부파일 read/write

(1) 엑셀파일 읽어오기: readxl 패키지 이용

(2) 데이터 쓰기: write.csv(data, ""); 파일(csv)로 저장 save(data, file=""); 변수를 파일로(rda) 저장

2) 데이터 파악하기

(1) library(dplyr); dplyr 패키지 로드 #데이터 ggplot2::mpg 이용시 - libirary(ggplot2) 필요

(2) 데이터 파악: head(), tail(), View(), dim(), str(), summary()

(3) 변수명 수정: rename()

(4) 파생변수 생성 – 기존에 존재하는 변수를 이용(계산식, 조건식 등)하여 생성

(5) 빈도 확인: table(); 빈도표, qplot(); 막대그래프 생성

3) 파악한 데이터 dplyr 패키지 이용하여 전처리(Preprocessing) 및 분석

함수	기능
filter()	행 추출; 조건에 맞는 데이터 추출
select()	열(변수) 추출
arrange()	정렬; 기본 오름차순. 내림차순: desc(변수) or -변수
mutate()	변수 추가(새필드) #실제 데이터X, dplyr코드에서 바로 사용 가능
summarise()	통계치 산출
group_by()	집단별로 나누기
left_join()	데이터 합치기(열)
bind_rows()	데이터 합치기(행)

※ 데이터 변질 없이 분석 가능(실제 데이터에 영향X, 다시 사용 예정이면 변수 등에 저장 필요)

※ 자주 사용하는 요약통계량 함수: mean(), sd(); 표준편차, sum(), median(), min(), max(), n(); 빈도

※ dplyr 패키지는 '%>%' 기호를 이용해서 함수들 나열하는 방식 # %>%단축키: ctrl+shift+m

05. 데이터 전처리 中 aggregate(), tapply(), summaryBy()

aggregate(data$col_1, by=list(data$col_2), sum) #대상열 하나이상 가능

tapply(data$col_1, data$col_2, sum) #대상열 하나만 가능

summaryBy(col_1~col_2, data=data, FUN=c(sum, mean)) #FUN 2개 이상도 가능

⇒ summaryBy는 group_by() + summarise()와 유사