## 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 mean(), sd(), sum(), median(), min(), max(), n(); 빈도 등 |
| **group_by()** | 집단별로 나누기 |
| **left_join()** | 데이터 합치기(열) #cbind(), merge()와 비교 |
| **bind_rows()** | 데이터 합치기(행) #rbind(), merge( ,na.rm=T)와 비교 |

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

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

### 4) 데이터 정제; 이상치, 결측치(NA) 처리 #이상치->결측치로 변경->중앙값or평균으로 대체

(1) 이상치 ①논리적 이상치, ②정상 범위 벗어난 이상치: boxplot(객체)$stats 이용

(2) 결측치 na.rm=T 이용 └#상하위0.3% or 평균±3*표준편차 └#boxplot 통계치

## 05. 데이터 전처리 中 tapply(), by(), summaryBy(), aggregate() 비교

| | 비교 열 수 | 비교 그룹 | 결과 함수(mean, sum, ...) |
|---|---|---|---|
| **tapply()** | 1개 | 1개 | 1개 |
| **by()** | 1개 이상<br>(단, 1개 이상은 mean, sd 등X) | 1개 | 1개 |
| **summaryBy()** | 1개 이상 | 1개 이상 | 1개 이상 |
| **aggregate()** | 1개 이상 | 1개 이상 | 1개 |

⇒ summaryBy는 group_by() + summarise()와 유사