2021.02.05

[dplyr 패키지를 이용한 데이터 전처리-I]

1. 외부파일 read / write

(1) 엑셀파일 읽어오기: read_excel('엑셀파일경로')

(2) 데이터 쓰기: write.csv(대상자료, file='저장할파일경로')

2. 데이터 파악하기

(1) view창에서 데이터 확인: View( ) / edit( )

(2) 차원 확인 / 속성 확인: dim( ) / str( )

(3) 요약 통계: summary( )

(4) 빈도표 출력: table( )

(5) 변수명 바꾸기: rename(대상자료, 바꿀변수명=기존변수명)

3. 파악한 데이터를 dplyr 패키지를 이용하여 전처리 및 분석하기

: %<% 기호를 이용하여 함수들을 나열한다.

(1) filter() 조건에 맞는 데이터 추출하기

(2) select() 필요한 변수(열) 추출하기

(3) arrange() 정렬하기

(4) mutate() 파생 변수를 추가

(5) summarise() 요약하기

(6) group_by() + summarise() 집단별로 요약하기

ex – mpg데이터에서 "회사별 suv 자동차의 도시 및 고속도로 통합 연비의 평균을 구하여 내림차순으로 정렬하고 1~5위까지 출력하기

```
mpg %>%                                  # 처음에 대상자료명(데이터명) 입력
  filter(class == 'suv') %>%             # class가 suv자동차인 행만 추출
  group_by(manufacturer) %>%             # 회사별로 그룹화
  mutate(total = (cty+hwy)/2) %>%        # 도시 연비와 고속도로 연비의 평균 파생변수 생성
  summarise(tot_mean = mean(total)) %>%  # 파생변수 total의 평균만 요약
  arrange(desc(tot_mean)) %>%            # 내림차순 정렬
  head(5)                                # 1~5위
```