4회차

**dplyr**

**# dplyr 패키지 설치**

install.packages('dplyr')

library(dplyr)

**# 데이터 불러오기**

delivery = read.csv('SKT.csv', fileEncoding='UTF-8')

head(delivery)

**## 1. Slice()**

행 번호를 활용해서 특정 행 불러오기

slice(delivery, c(1,3,5:10))

**##1.2 filter( )**

조건에 맞는 데이터 행만 불러오기

구문형식 : filter(데이터명, 조건)

filter(delivery, 시군구=='성북구')

filter(delivery, delivery$시군구=='성북구')

**## 복수의 조건 사용**

구문형식 : filter(데이터명, 조건1, 조건2, 조건3, ...)

filter(delivery, 시군구=='성북구', 요일 %in% c('토', '일'), 업종=='피자' | 통화건수>=100)

업종=='피자' | 통화건수>=100 부분 조건이 OR임으로 두 가지 조건이 적용된 결과값이 나온다.

**##1.3 arrange( )**

# 정렬하기 (오름차순)

arrange(delivery, 시군구, 요일, 업종)

구문형식 : arrange(데이터명, 정렬기준변수1, 정렬기준변수2, ...)

## 내림차순(Descending)으로 정렬할 때는 'desc' 옵션 활용

arrange(delivery, desc(시군구), 요일, 업종)

##**1.4 select( )**

# 변수를 선택하거나 제외하기

select(delivery, 통화건수)

# ":"를 활용한 순서대로 여러 변수 선택하기

select(delivery, 시간대:통화건수)

# "-"를 활용한 변수제외

select(delivery, -요일)

**##1.5 distinct( )**

# 반복내용 제거하기 (컬럼 내 값 종류 확인하기)

distinct( delivery, 업종)

“족발보쌈, 중국음식, 치킨, 피자” 로 분류됨

**#### 연습문제**

## 1.iris 데이터중 1부터 50행중 홀수, 100부터 150행중 짝수 선택

slice(iris,c(seq(1,50,2),seq(100,150,2)))

## 2.iris 데이터중 Species가 "setosa"이면서 Sepal.Length가 5보다 큰 값을 추출하시오

filter(iris,Species=="setosa",Sepal.Length>5)

## 3.iris 데이터중 Sepal.Length는 내림차순 Sepal.Width는 오름차순으로 출력하시오

arrange(iris,desc(Sepal.Length),Sepal.Width)

## ## 4.iris 데이터중 "Sepal.Width" 와 "Species" 열을 선택하시오

dplyr::select(iris,Sepal.Width,Species)

## 5.iris 데이터중 "Species"의 종류를 확인하시오

distinct(iris, Species)

**##1.6 mutate( )**

# 기존 변수를 활용한 임시 변수 만들기 (2가지 방법)

mutate(delivery, 새요일=paste0(요일, '요일'))

delivery$새요일 = paste0(delivery$요일, '요일')

또다른 기본 예시

> a=c(1,2,3,4,5)

> paste(a,"hi")

[1] "1 hi" "2 hi" "3 hi" "4 hi" "5 hi"

**##1.7 count( )**

#그룹별 개수 세기 (그룹별로 개수를 센다는게 중요)

count(delivery, 시군구)

**## 1.8 group\_by( )**

# 그룹 지정해주기

delivery\_grp = group\_by(delivery, 시군구)

**##1.9 summarize( )로 요약 하기**

## 원본 데이터는 전체 요약, 그룹이 지정된 데이터는 그룹별 요약

summarise(delivery, mean=mean(통화건수), m=min(통화건수), M=max(통화건수))

summarise(delivery\_grp, mean=mean(통화건수), m=min(통화건수), M=max(통화건수))

summarise(delivery\_grp, length(통화건수))

## "delivery %>% count(통화건수)"와 동일

## 아래위 결과값이 동일하기 때문에 중요!

**##1.10 top\_n( )**

# 상위 관측치 확인하기

설정해준 변수를 기준으로 가장 값이 높은 n개의 데이터를 가져옵니다.

top\_n(delivery, 5, 통화건수)

top\_n(delivery\_grp, 5, 통화건수)

**#### 연습문제**

## 6.iris 데이터중 "Sepal.Length" 와 "Sepal.Width" 두변수의 합을 Sepal\_sum이라는 변수에 저장하시오

mutate(iris, Sepal\_sum=Sepal.Length+Sepal.Width)

## 7.iris 데이터중 "Species"의 종별 개수를 확인하시오

count(iris,Species)

## 8.iris 데이터중 Sepal.Length의 합과 Sepal.Width의 평균을 구하시오

summarise(iris,sum=sum(Sepal.Length),mean=mean(Sepal.Width))

## 9.iris 데이터중 "Petal.Width"의 상위 5개의 값을 출력

top\_n(iris,5,Petal.Width)

**##1.11 파이프라인( %>% )을 활용한 연속작업**

delivery %>%

filter(업종=='중국음식') %>%

group\_by(시군구) %>%

summarise(mean\_call = mean(통화건수)) %>%

arrange(desc(mean\_call))

# 데이터 저장

new\_data = delivery %>%

filter(업종=='중국음식') %>%

group\_by(시군구) %>%

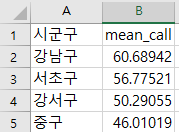
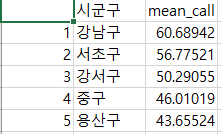
summarise(mean\_call = mean(통화건수)) %>%

arrange(desc(mean\_call)) ## 마지막에 arrange 내림차순 정리 중요!!

# 결과를 csv파일로 저장

write.csv(new\_data, 'result.csv', row.names=FALSE)

write.csv(new\_data, 'result.csv', row.names=TRUE)



**##1.12 ungroup( )의 활용**

만약 구분된 집단을 삭제하고 싶다면 ungroup()함수를 이용합니다. 집단이 사라짐을 알 수있습니다.

기준을 다시 잡거나 새롭게 특정 상위를 뽑고 싶을 때 사용을 한다.

**# 시군구별 상위 3대 시간대 확인**

delivery %>%

filter(업종=='중국음식') %>%

group\_by(시간대, 시군구) %>% **## 시간대와 시군구의 모든 경우의 수 나옴**

summarise(mean\_call = mean(통화건수)) %>%

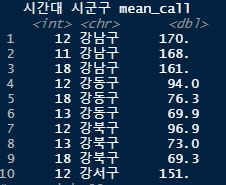
ungroup() %>%

group\_by(시군구) %>% **## 시군구 기준으로 상위 3개의 시간대만 보고 싶다**

top\_n(3, mean\_call) %>% **## mean\_call 기준으로 상위 3개만 가져와줘**

arrange(시군구, desc(mean\_call))

**시군구별로 상위 3개의 시간대까지 결과값이 나옴**

****

**## 이해가 안될수 있는 케이스**

delivery %>%

filter(업종=='중국음식') %>%

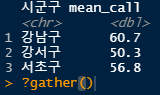
group\_by(시군구) %>% ## ungroup 없애고 “시군구”로만 지정해줘도

되는 것 아닌가?

summarise(mean\_call = mean(통화건수)) %>%

top\_n(3, mean\_call) %>%

arrange(시군구, desc(mean\_call))

****

**\* 시간대 컬럼 자체가 없음 (위 구문과 차이점 비교)**

**##2 (실습) 보험료 데이터 요약하기**

# 예제 데이터 불러오기

ins = read.csv('insurance.csv')

head(ins)

**#1. 데이터 ins에서 sex가 female인 관측치로 region별 관측치 수 계산 (count함수 개념 이해!)**

ins %>%

filter(sex=='female') %>%

count(region)

**#2 charges가 10000이상인 관측치 중에서 smoker별 평균 age 계산**

ins %>%

filter(charges >= 10000) %>%

group\_by(smoker) %>%

summarise(mean(age))

**#3 age가 40 미만인 관측치 중에서 sex, smoker별 charges의 평균과 최댓값 계산**

ins %>%

filter(age < 40) %>%

group\_by(sex, smoker) %>%

summarise(mean(charges),

max(charges))

**##3 (실습) 국민건강보험공단 데이터 요약**

# 성 / 연령대별 진료건수 계산 (진료건수 기준 내림차순 정렬)

**CASE1**

NHIS %>%

group\_by(SEX, AGE\_GROUP) %>%

summarise(n=length(SEX)) %>%

arrange(desc(n))

**CASE2**

NHIS %>%

dplyr::count(SEX, AGE\_GROUP) %>%

arrange(desc(n))

SEX, AGE\_GROUP에서 행(데이터 개수)를 구해주는 거라서 length()함수를 사용해서 데이터 개수 파악, count()와 같은 느낌

**# 성별/연령대별 환자 분포 확인**

## distinct() : 중복값 제거

NHIS %>%

dplyr::select(IDV\_ID, SEX, AGE\_GROUP) %>%

unique() %>%

dplyr::count(SEX, AGE\_GROUP)

############################ 연습 문제 ################################

**# 문제1**

# 성 / 연령대 / 진료과목별 환자수 계산 (환자수 기준 내림차순 정렬)

NHIS %>%

dplyr::select(SEX, AGE\_GROUP,DSBJT\_CD, IDV\_ID) %>%

unique() %>%

count(SEX, AGE\_GROUP,DSBJT\_CD) %>%

arrange(desc(n))

**# 문제2**

# 성별/연령대별 평균(요양일수/입내원일수/급여비용/본인부담금액) 계산 후 급여비용 내림차

순으로 정렬

NHIS %>%

group\_by(SEX, AGE\_GROUP) %>%

summarise(m1=mean(VSCN), m2=mean(RECN), m3=mean(EDEC\_TRAMT),

m4=mean(EDEC\_SBRDN\_AMT)) %>%

arrange(desc(m3))

**# 문제3**

# 성별/연령대별 3개 최고빈도 주상병코드

## top\_n(n=k, wt=기준변수) : 기준변수를 기준으로 상위 k개 관측치 선택

## 주상병코드 조회 (MAIN\_SICK)

**TYPE1**

NHIS %>%

count(SEX, AGE\_GROUP, MAIN\_SICK) %>%

group\_by(SEX, AGE\_GROUP) %>%

top\_n(n=3 , wt=n) %>%

arrange(SEX, AGE\_GROUP, desc(n))

**TYPE2**

NHIS %>%

group\_by(SEX, AGE\_GROUP, MAIN\_SICK) %>%

summarise(n=n()) %>%

top\_n(n=3 , wt=n) %>%

arrange(SEX, AGE\_GROUP, desc(n))

* num = n() 국룰처럼 사용된다는 그 내용처럼 count를 해주는 느낌!

**###### mutate의 확장**

**## 변수의 group별 비율을 알고싶다면??**

**## 도전해보자**

**## Species 별로 Sepal.Length의 비중을 알고싶다면??**

<정답>

data(iris)

iris %>%

dplyr::group\_by(Species) %>%

dplyr::mutate(sum = sum(Sepal.Length)) %>%

dplyr::mutate(prop = Sepal.Length/sum) %>%

dplyr::select(Species,Sepal.Length,prop)

**###### mutate의 확장2**

**## group별로 번호를 매기고 싶다면??**

**# 도전해보자**

**# iris 데이터에서 Species별 Sepal.Width가 3번쨰로 큰 값들의 합은?**

iris %>%

arrange(Species,Sepal.Width) %>%

dplyr::group\_by(Species) %>%

dplyr::mutate(seq\_num = row\_number()) %>%

filter(seq\_num==3)

row\_number() : 순위(ranking) index 반환, 동일값에 대해서는 '1, 2, 3, ...' 처리

* 알겠는데 확 이해하진 못함.