Wk3-4 : 데이터핸들링 - dplyr 활용 -



ⓒ포항공대 산업경영공학과 이혜선

2. 데이터핸들링을 위한 라이브러리

3-4. R 데이터핸들링

• dplyr : 데이터핸들링을 편리하게 수행할수 있는 패키지

dplyr 패키지의 주요 함수

select : 일부변수를 선택

filter: 필터링 기능 (조건에 맞는 데이터 추출)

mutate: 새로운 변수 생성

group_by : 그룹별 통계량을 얻을때

summarize : 요약통계량 (mean, min, max, sum)

arrange: 행 정렬시 사용

1. 데이터 핸들링 : dplyr 활용

• dplyr 패키지 설치 및 기본 설정

프로그램 편집창

```
lec3_4.r
# Data handling
# Data analysis with autompg.csv
# data manipulation package
# select, filter, group by, summarise in dplyr
install.packages("dplyr")
library(dplyr)
# set working directory
# change working directory
setwd("D:/tempstore/moocr")
# Read txt file with variable name
# http://archive.ics.uci.edu/ml/datasets/Auto+M
# Data reading in R
car<-read.csv(file="autompg.csv")</pre>
attach(car)
# data checking
str(car)
```

Step0 : 분석을 위한 설정 (install, library, setwd)

Step1 : 데이터핸들링 (csv파일 불러들이기, ...)



3

2. 변수 추출 (select)

3-4. R 데이터핸들링

• 변수 추출 : select(데이터, 변수이름, ...)

car 데이터에서 mpg, hp 변수만 추출

```
# Data handling using "dplyr"
# 1 subset data : selecting a few variables
set1<-select(car, mpg, hp)
head(set1)</pre>
```

```
> head(set1)
  mpg hp
1  18 17
2  15 35
3  18 29
4  16 29
5  17 24
6  15 42
```



2. 변수 추출 (select)

• 변수 추출 : select(데이터, 변수이름, ...)

car 데이터에서 mpg로 시작하는 변수를 제외하고 set2 라는 데이터를 생성

2 subset data : Drop variables with set2<-select(car, -starts_with("mpg"))
head(set2)</pre>

starts with(): 변수 시작



```
> set2<-select(car, -starts_with("mpg"))</pre>
> head(set2)
               wt accler year origin
 cyl disp hp
                                                      carname
     307 17 3504
                    12.0 70 1 chevrolet chevelle malibu
      350 35 3693
                    11.5
                           70
                                  1
                                      buick skylark 320
                         70
                                          plymouth satellite
3
      318 29 3436
                    11.0
                                  1
                         70
      304 29 3433
                    12.0
                                  1
                                                amc rebel sst
                          70
5
   8 302 24 3449
                    10.5
                                  1
                                                  ford torino
   8 429 42 4341
                    10.0
                           70
                                             ford galaxie 500
```



5

3. 데이터 추출 (filter)

3-4. R 데이터핸들링

• 조건식에 맞는 데이터 추출 : filter(데이터, 변수조건, ...)

car 데이터에서 mpg가 30보다 큰 행 추출

```
# 3. subset data : filter mpg>50
set3<-filter(car, mpg>30)
head(set3)
```



```
head(set3)
tibble: 6 x 9
     cyl disp
                   hp
                         wt accler year origin carname
     int> <db1> <db1> <int>
                               <db1> <int> <int> <fct>
 31
              71
                                19
                    62 <u>1</u>773
                                                 3 toyota corolla 1200
             72
79
 35
                    66 <u>1</u>613
                                18
                                        71
                                                 3 datsun 1200
31
                    64 <u>1</u>950
                                19
                                                 3 datsun b210
             71
32
                    62 <u>1</u>836
                                        74
                                                 3 toyota corolla 1200
                                21
                                16.5
                        <u>1</u>649
31
              76
                    53
                                         74
                                                 3 toyota corona
                    58
                        <u>2</u>003
                                19
                                                 3 datsun 710
```



4. 변수생성 (mutate)

• 변수 생성 : mutate(새로운 변수이름=기존변수 활용)

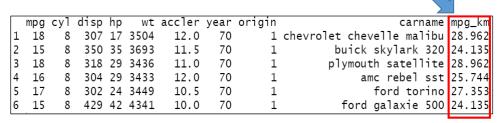
```
%>%(파이프 연산자) 연산자 사용하여 연결
# 4. create a derived variable
set4<-car %>% ①
filter(!is.na(mpg)) %>% ②
mutate(mpg_km = mpg*1.609) ◀③
head(set4)
```

파이프연산자 : 앞에서부터 ①,②,③ 순서 대로 수행하여 데이터전처리를 하고 set4 라는 이름으로 저장

• filter car데이터 mpg열의 NA가 아닌 모든 데이터 추출

is.na() NA여부 판단하는 함수 (! 기호는 부정하는 기호)

• mutate 기존의 mpg열 사용하여 새로운 mpg_km열 생성





-

5. 데이터 요약통계치

3-4. R 데이터핸들링

•데이터 요약통계치(평균 구하기): summarize(mean(변수이름))

```
mpg, hp, wt의 평균값 구하기
```

몇 개 변수들의 평균값 한번에 구하기

```
# mean of some variables
                                        > select(car, 1:6) %>%
                                                                 1~6열 추출함
select(car, 1:6) %>%
                                           colMeans()
 colMeans()
                                                 mpg
                                                             cyl
                                                                        disp
                                           23.514573
                                                        5.454774
                                                                  193.425879
                                                                      accler
                                                              wt
                                           51.389447 2970.424623
                                                                   15.568090
```

• colMeans 데이터를 열로 재구성하여 평균값 구함



5. 데이터 요약통계치

• 벡터화 요약치 : summarize_all(FUN)

열 추출하여 기술통계치 구하고 요약치 보여줌

```
# table with descriptive statistics
a1 <- select(car, 1:6) %>% summarize_all(mean)
a2 <- select(car, 1:6) %>% summarize_all(sd)
a3 <- select(car, 1:6) %>% summarize_all(min)
a4 <- select(car, 1:6) %>% summarize_all(max)
table1 <- data.frame(rbind(a1,a2,a3,a4))|
rownames(table1) <- c("mean","sd","min","max")
table1</pre>
```

data.frame을 tbl_df로 전환시켰으므로 data.frame으로 원상복귀시켜서 행 이름을 바꾼다.



9

6. 그룹별 기술통계치

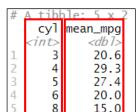
3-4. R 데이터핸들링

• 그룹별 통계량 얻기: group_by(변수), summarize(___=FUN())

그룹별 요약통계량 구하기

```
# summary statistics by group variable
tar %>%
    group_by(cyl) %>%
    summarize(mean_mpg = mean(mpg, na.rm = TRUE))
```





- group_by car데이터의 cyl열을 그룹으로 묶음
- summarize() cyl그룹의 mpg 평균을 구함

요약통계량을 구할 때 group_by와 summarize 함께 사용하는 경우 많음

함수	요약통계량
mean	평균
min	최솟값
max	최댓값
sum	합계
var	분산
sd	표준편차
median	중앙값
n	빈도



