# RIOIEI 의한 RIPARE 1

3**주차**. R 데이터구조(생성, 추출)



이혜선 교수

포항공과대학교 산업경영공학과



# 3주차. R 데이터구조(생성, 추출)

1차시 R 데이터 생성 (불러들이기)

2차시 R 데이터 활용 I (subset, 내보내기)

3차시 R 데이터 활용 II (dplyr 활용)





# ● 데이터핸들링

☑ dplyr: 데이터 핸들링을 편리하게 하는 라이브러리

### dplyr 패키지의 주요 함수

★ select : 일부변수를 선택

★ filter : 필터링 기능 (조건에 맞는 데이터 추출)

★ mutate : 새로운 변수 생성

★ group\_by : 그룹별 통계량을 얻을 때

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

★ arrange : 행 정렬시 사용

### ☑ dplyr 설치 및 설정

# Data handling using "dplyr" install.packages("dplyr") library(dplyr)

앞에서의 subset과 같은 기능

# ● 데이터 핸들링

### ☑ dplyr 패키지 설치 및 기본 설정

### 프로그램 편집 창

```
# lec3_3.r
# Data handling
# Data analysis with autompg.csv
# data manipulation package
# select, filter, group by, summarise in dplyr package
                                                                   Step0 : 분석을 위한 설정
install.packages("dplyr")
library(dplyr)
                                                                   (install, library, setwd)
# set working directory
# change working directory
setwd("D:/tempstore/moocr")
# Read txt file with variable name
# http://archive.ics.uci.edu/ml/datasets/Auto+MPG
# Data reading in R
                                                                     Step1: 데이터핸들링
car<-read.csv(file="autompg.csv")</pre>
                                                                   (csv파일 불러들이기, ···)
attach(car)
```

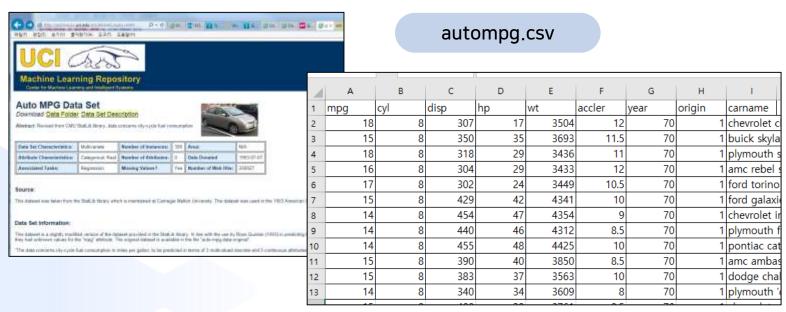
▶ tbl\_df(): 빅데이터에 유용함. df는 데이터프레임을 의미





## ○ 오픈 데이터 활용

✓ UCI repository data





# ● 오픈 데이터 활용

- ☑ autompg 데이터
  - 🔰 1. mpg: continuous (연비 : 연속형변수)
    - 2. cylinders: multi-valued discrete (실린더: 정수값)
    - 3. displacement: continuous (배기량 : 연속형변수)
    - 4. horsepower: continuous (마력: 연속형변수)
    - 5. weight: continuous (무게: 연속형변수)
    - 6. acceleration: continuous (가속: 연속형변수)
    - 7. year: multi-valued discrete (모델연도 : 정수값)
    - 8. origin: multi-valued discrete (정수값)
    - 9. car name: string (unique for each instance) (차종류 이름)

☑ 변수특성 변경 (as.numeric, as.integer, factor)

## ● 데이터 활용 Ⅱ

### ☑ 데이터 불러들이기

```
# Data reading in R
car<-read.csv(file="autompg.csv")</pre>
attach(car)
head(car)
dim(car)
str(car)
```

R 데이터 이름은? : car

car 데이터의 수는?: 398

car 는 몇개의 변수? : 9

```
> head(car)
                   wt accler year origin
                                                           carname
                        12.0
                                       1 chevrolet chevelle malibu
          350 35 3693
                        11.5 70
                                                 buick skylark 320
                        11.0 70
12.0 70
          318 29 3436
                                                plymouth satellite
          304 29 3433
                                                     amc rebel sst
       8 302 24 3449
                              70
                        10.5
                                                       ford torino
       8 429 42 4341
                        10.0
                               70
                                                  ford galaxie 500
 dim(car)
```



## ●데이터 활용 Ⅱ

☑ 데이터의 전체 구조 파악하기 : str(데이터이름)

```
# Data reading in R
car<-read.csv(file="autompg.csv")</pre>
attach(car)
head(car)
dim(car)
str(car)
```

car 변수들 중 실수값으로 저장된 변수들은? 4개

car 변수들 중 정수값으로 저장된 변수들은? 4개

car 변수들 중 문자변수는 ? 1개

```
> str(car)
'data.frame':
               398 obs. of 9 variables:
         : num 18 15 18 16 17 15 14 14 14 15 ...
$ mpa
 $ cy1
         : int 88888
                307 350 318 304 302 429 454 440 455 390 ...
         : num
 $ hp
         : num
$ wt
         : int
               3504 3693 3436 3433 3449 4341 4354 4312 4425 3850
$ accler : num
               12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
         : int 70 70 70 70 70 70 70 70 70 70 ...
$ vear
$ origin : int 1 1 1 1 1 1 1 1 1 ...
$ carname: chr "chevrolet chevelle malibu" "buick skylark 320" "p"
th satellite" "amc rebel sst" ...
```

## ● 데이터 활용 Ⅱ

☑ 데이터 요약하기 : summary(데이터이름)

```
# summary
summary(car)
```

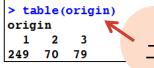
최소값 25%값 중위수 평균 75%값 최대값

```
summary(car)
                                       disp
                      cyl
                                                        hp
     mpg
       : 9.00
                Min.
                        :3.000
                                 Min.
                                        : 68.0
                                                  Min.
                                                          : 1.00
1st Qu.:17.50
                1st Qu.:4.000
                                 1st Qu.:104.2
                                                  1st Qu.:26.00
Median :23.00
                Median :4.000
                                 Median :148.5
                                                  Median :60.50
       :23.51
                        :5.455
                                         :193.4
                                                         :51.39
Mean
                Mean
                                 Mean
                                                  Mean
3rd ou.:29.00
               3rd Qu.:8.000
                                3rd Ou.:262.0
                                                  3rd Qu.:79.00
       :46.60
                Max.
                        :8.000
                                 Max.
                                         :455.0
                                                  Max.
                                                          :94.00
Max.
                   accler
                                                     origin
      wt
                                     year
       :1613
               Min.
                       : 8.00
                                        :70.00
                                                         :1.000
Min.
                                Min.
                                                 Min.
1st Ou.: 2224
               1st Qu.:13.82
                                1st Ou.:73.00
                                                 1st Ou.:1.000
Median :2804
               Median :15.50
                                Median :76.00
                                                 Median :1.000
       :2970
                      :15.57
                                        :76.01
                                                        :1.573
Mean
               Mean
                                Mean
                                                 Mean
3rd Qu.:3608
               3rd Qu.:17.18
                                3rd Qu.:79.00
                                                 3rd Qu.:2.000
       :5140
               Max.
                       :24.80
                                Max.
                                        :82.00
                                                 Max.
                                                        :3.000
Max.
  carname
Length: 398
class :character
```

## ● 데이터 활용 Ⅱ

☑ 데이터의 요약통계치 (빈도 구하기) : table(데이터이름)

```
table(origin)
table(year)
```



attach(데이터이름): 현재 세션에서 나오는 변수들은 그 '데이터'의 변수로 인식 => 데이터이름을 안써줘도 됨!

```
> table(year)
    year
Year 70 | 71 | 72 73 74 75 76 77 78 79 80 81 82
freg 29 28 28 40 27 30 34 28 36 29 29 29 31
```

1970년도부터 1982년까지의 차량

☑ 데이터의 요약통계치 (평균, 표준편차) 구하기

### 개별변수의 평균 구하기

```
# mean and standard deviation
mean(mpg)
mean(hp)
mean(wt)
```



```
mean(mpg)
[1] 23.51457
> mean(hp)
[1] 51.38945
> mean(wt)
[1] 2970.425
```

- 데이터핸들링- 데이터 추출 (select)
- ☑ 변수 추출: select(데이터, 변수이름, …)

car 데이터에서 mpg, hp 변수만 추출

# 1. subset data : selecting a few variables
set1<-select(car, mpg, hp)</pre>

### car 데이터에서 mpg로 시작하는 변수를 제외하고 set2 라는 데이터를 생성

# 2. subset data : Drop variables with set2<-select(car, -starts\_with("mpg"))</pre>

starts\_with(): 변수 시작

- **데이터핸들링 데이터 추출** (filter)
- ☑ 조건식에 맞는 데이터 추출 : filter(데이터, 변수조건, ··· )

### car 데이터에서 mpg가 30보다 큰 행 추출

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



```
head(set3)
# A tibble: 6 x 9
                           wt accler year origin carname
                               <dbl> <int> <int> <fct>
                                               3 toyota corolla 1200
                         1773
    35
               72
                         1613
                                               3 datsun 1200
                         1950
                                               3 datsun b210
                                   74
               71
                               21
                     62
                        1836
                                               3 toyota corolla 1200
               76
                               16.5
                                       74
                         1649
                                               3 tovota corona
                                       74
                                               3 datsun 710
                         2003
                                19
```

- **데이터핸들링 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
```

파이프연산자: 앞에서부터 ①,②,③ 순서대로 수행하여 데이터전처리를 하고 set4라는 이름으로 저장

☑ filter car데이터 mpg열의 NA가 아닌 모든 데이터 추출

is.na() NA여부 판단하는 함수

☑ mutate 기존의 mpg 사용하여 새로운 변수 mpg\_km 생생 기호는 부정하는 기호)

*	mpg	cyl		disp	hp	wt	accler	year	origin	carname	mpg_km
1	16.0		b	307.0	17	3504	12.0	70		chevrolet chevelle mélibu	26,9620
2	150		8	350.0	35	3693	115	70	7	buck skywik 320	24.1350
3	18.0		5	318.0	29	3436	11.0	70	1	prymouth satellite	25.9620
4	16.0		5	304.0	29	3433	12.0	70	- 1	amc rebel sat	25,7440
5	17.0		8	302.0	24	3449	10.5	70	-	ford torino	27.3530
6	150		8	429.0	42	4341	100	70	1	ford galaxie 500	24.1330
,	14.0			454.0	47	4554	9.0	70	1	cheumet impala	22.5260



- 데이터핸들링 데이터 요약통계치
- ☑ 데이터 요약통계치(평균 구하기) : summarize(mean(변수이름))

### mpg, hp, wt의 평균값 구하기

```
mean and standard deviation
car %>%
  summarize(mean(mpg), mean(hp), mean(wt))
```



```
summarize(mean(mpg), mean(hp), mean(wt))
'mean(mpg)` `mean(hp)` `mean(wt)`
                       <db1>
                                       \langle db 1 \rangle
        \langle db 1 \rangle
        23.5
                        51.4
                                       <u>2</u>970.
```

1~6열 추출함

### 몇 개 변수들의 평균값 한번에 구하기

```
# mean of some variables
select(car, 1:6) %>%
  colMeans()
```

```
> select(car, 1:6) %>%
    colMeans()
                                disp
                    cyl
  23.514573
               5.454774
                         193.425879
                              accler
  51.389447 2970.424623
                          15.568090
```

colMeans 데이터를 열로 재구성하여 평균값 구함

## ● 데이터핸들링 - 데이터 요약통계치

✓ 벡터화 요약치: summarize all(FUN)

### 변수를 추출해 기술통계치 구하고 요약치 보여줌

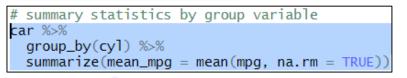
```
# table with descriptive statistics
al <- select(car, 1:6) %>% summarize_all(mean)
a2 <- select(car, 1:6) %>% summarize_all(sd)
a3 <- select(car, 1:6) %>% summarize_all(min)
a4 <- select(car, 1:6) %>% summarize_all(max)
table1 <- rbind(a1,a2,a3,a4)
rownames(table1) <- c("mean", "sd", "min", "max"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



```
table1
                                                        accler
                    cyl
                            disp
           mpq
mean 23.514573 5.454774 193.4259 51.38945 2970.4246 15.568090
     7.815984 1.701004 104.2698 29.93236
                                            846.8418
                                                      2.757689
     9.000000 3.000000
                         68.0000
                                  1.00000 1613.0000
                                                      8.000000
     46.600000 8.000000 455.0000 94.00000 5140.0000 24.800000
```

- 데이터핸들링 그룹별 기술통계치
- ☑ 그룹별 통계량 얻기 : group\_by(변수), summarize( \_\_\_=FUN())

### 그룹별 요약통계량 구하기





#	<u>A tibble: 5 x 2</u>					
	cyl	mean_mpg				
	<int></int>	<db1></db1>				
1	3	20.6				
2	4	29.3				
3	5	27.4				
4	6	20.0				
5	8	15.0				

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

- ▶ group\_by car데이터의 cyl열을 그룹으로 묶음
- 🔰 summarize() cyl그룹의 mpg 평균을 구함

▶ 요약통계량을 구할 때 group\_by와 summarize 함께 사용하는 경우 많음