

데이터과학을 위한 **R**프로그래밍

3주차. R 데이터구조(생성, 추출)



이혜선 교수

포항공과대학교 산업경영공학과



목차

3주차. R 데이터구조(생성, 추출)

1차시

R 데이터 생성 (불러들이기)

2차시

R 데이터 활용 I (subset, 내보내기)

3차시

R 데이터 활용 II (dplyr 활용)



3주차

1차시

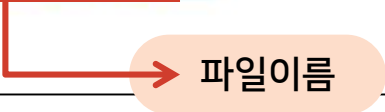
R 데이터 생성 (블러들이기)

● 데이터 불러들이기 (csv)

✓ csv파일 불러들이기 (read.csv)

➤ csv (comma separated value) 데이터 범용 형태

```
# 1. Read csv file : brain weight data  
brain<-read.csv("brain2210.csv")  
head(brain)  
dim(brain)
```



✓ xls 파일

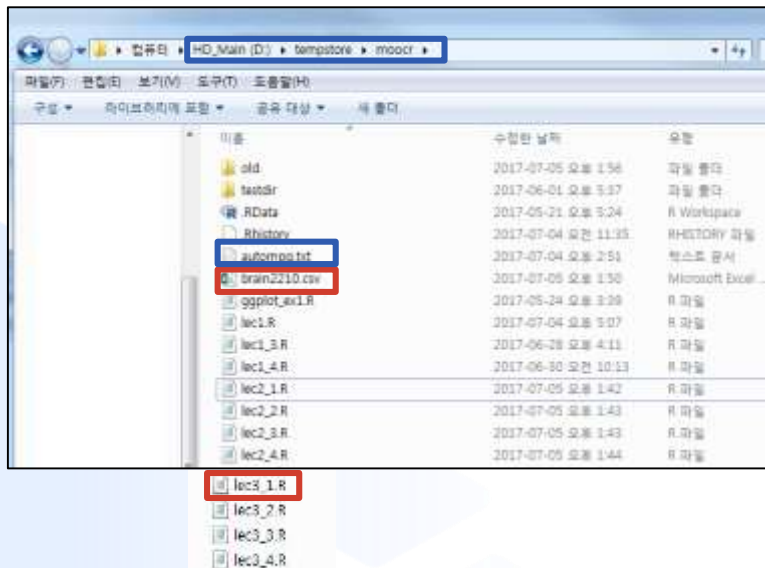
➤ *.xls파일인 경우 데이터를 .csv로 저장한 다음
read.csv 함수를 사용하여 R데이터로 불러들이는게 편리함

데이터 저장 폴더

☑ 데이터와 프로그램 저장 폴더 지정 (영문으로 폴더 이름 생성)

예 D:/tempstore/moocr

➤ brain2210.csv



	A	B	C	D
1	wt	sex		
2	1607	m		
3	1157	m		
4	1248	m		
5	1310	m		
6	1398	m		
7	1237	m		
8	1232	m		
9	1343	m		
10	1380	m		
11	1274	m		
12	1245	m		
13	1286	m		
14	1508	m		
15	1105	m		
16	1123	m		

데이터 저장 폴더

✓ 현재 프로그램 작업폴더 (setwd) :

➤ setwd : set working directory

```
# lec3_1.r
# Working directory
# Data import, frequency, histogram
# attach, detach

# set working directory
# change working directory
setwd("D:/tempstore/moocr")

# check the current working directory
getwd()

# 1. Read csv file : brain weight data
brain<-read.csv("brain2210.csv")
head(brain)
dim(brain)
```

brain2210.csv는 D:/tempstore/moocr에 들어있으므로 working directory를 여기로 설정!!

Environment		History	Connections
Import Dataset			
Global Environment			
Data			
brain	185 obs. of 2 variables		

*.csv를 R로 불러들인 후 환경창을 보면 brain이라는 이름의 R데이터가 생성되어 있음

데이터 불러들일 때 tip

☑ 데이터를 불러들일 때 몇가지 tips

- ▶ Working directory를 설정 : `setwd("데이터가 저장되어있는 폴더")`
- ▶ 데이터를 불러들이고 확인
 - `head(데이터이름)` : 첫번째 6줄을 프린트해줌
 - `dim(데이터이름)` : 데이터의 관측치수와 변수의 갯수를 알려줌

```
# lec3_1.r
# working directory
# Data import, frequency, histogram,
# attach, detach

# set working directory
# change working directory
setwd("D:/tempstore/moocr")

# check the current working directory
getwd()

# 1. Read csv file : brain weight data
brain<-read.csv("brain2210.csv")
head(brain)
dim(brain)
```



```
> head(brain)
   wt sex
1 1607  m
2 1157  m
3 1248  m
4 1310  m
5 1398  m
6 1237  m
> dim(brain)
[1] 185  2
```

데이터와 변수 이름

✓ attach 사용

- ▶ attach(데이터이름) :
데이터이름을 따로 지정하지 않아도 됨

```
# 2. example for using 'attach'

# to get frequency of male and female (brain data)
table(brain$sex)

# using the command 'attach'
attach(brain)

# get frequency of male and female
table(sex)
```

- ▶ table(변수) :
빈도 구하기 (male과 female 몇명씩?)

```
> # to get frequency of male and female (brain data)
> table(brain$sex)

  f   m 
77 108 

> attach(brain)
The following objects are masked from brain:

  sex, wt 

> # get frequency of male and female
> table(sex)
sex
  f   m 
77 108
```

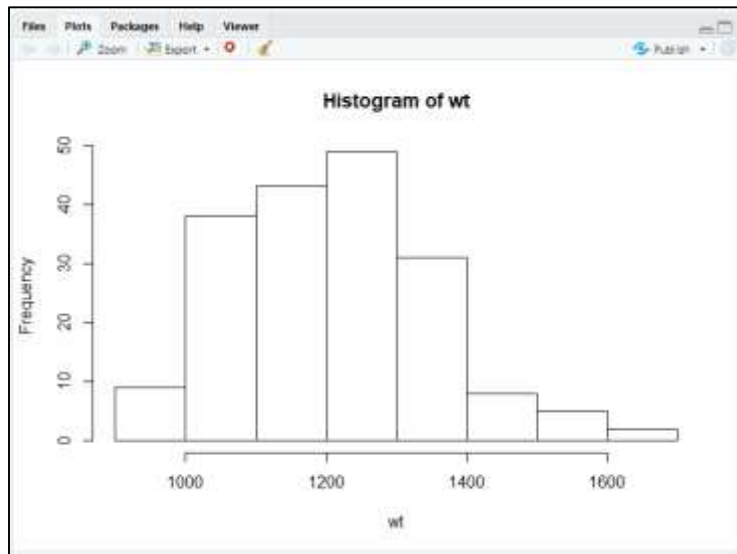

데이터분석 활용

✓ 데이터 알아보기 (히스토그램) : hist(변수이름)

➤ attach(데이터이름) :
현재 세션에서 나오는 변수들은
그 '데이터'의 변수로 인식한다는 의미

➤ detach(데이터이름) : attach를 풀어줌

```
# histogram of brain weight  
# hist(brain$wt)  
hist(wt)  
  
detach(brain)
```



여러 형태의 데이터 불러들이기

☑ 통합 Excel 파일 (여러 worksheet가 있을때 : readxl 패키지 설치)

```
# several sheets in Excel file
install.packages("readxl")
library(readxl)

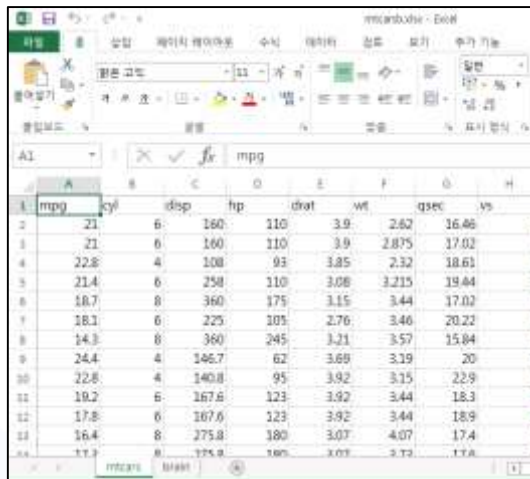
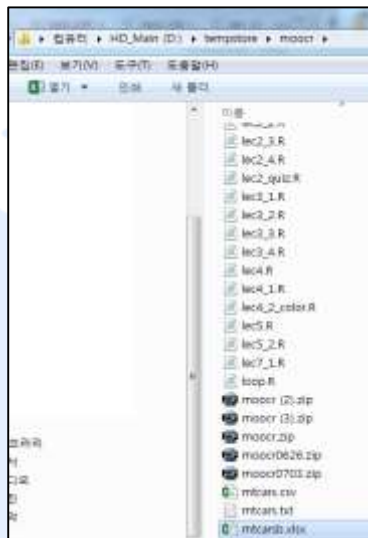
mtcars1 <- read_excel("D:/tempstore/moocr/mtcarsb.xlsx",
                      sheet = "mtcars")
```

```
> mtcars1 <- read_excel("D:/tempstore/moocr/mtcarsb.xlsx",
+                       sheet = "mtcars")
> head(mtcars1)
# A tibble: 6 x 11
   mpg   cyl  disp    hp  drat    wt  qsec    vs  am  gear  carb
<dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1  21.0     6   160   110   3.90  2.620  16.46     0     1     4     4
2  21.0     6   160   110   3.90  2.875  17.02     0     1     4     4
3  22.8     4   108    93   3.85  2.320  18.61     1     1     4     1
4  21.4     6   258   110   3.08  3.215  19.44     1     0     3     1
5  18.7     8   360   175   3.15  3.440  17.02     0     0     3     2
6  18.1     6   225   105   2.76  3.460  20.22     1     0     3     1
```

여러 형태의 데이터 불러들이기

☑ 통합 Excel 파일 (여러 worksheet가 있을 때)

➤ 데이터 확인 (mtcars, brain이라는 두개의 sheets)



mpg	cyl	disp	hp	drat	wt	qsec	vs
21.0	6	160	110	3.9	2.62	16.46	0
21.0	6	160	110	3.9	2.875	17.02	0
22.8	4	108	93	3.85	2.32	18.61	1
21.4	6	258	110	3.08	3.215	19.44	0
18.7	8	360	175	3.15	3.44	17.02	0
18.1	6	225	101	2.76	3.46	20.22	0
14.3	8	360	245	3.21	3.57	15.84	0
24.4	4	146.7	62	3.69	3.19	20	1
22.8	4	140.8	95	3.92	3.15	22.9	1
19.2	6	167.6	123	3.92	3.44	18.3	1
17.8	6	167.6	123	3.92	3.44	18.9	1
16.4	8	275.8	180	3.07	4.07	17.4	0

여러 형태의 데이터 불러들이기

- ☑ 통합 Excel 파일 (여러 worksheet가 있을때 : readxl 패키지 설치)
read_excel("파일이름(폴더패스 포함)", sheet="이름")

➤ sheet 이름 혹은 번호를 기입

```
install.packages("readxl")
library(readxl)

mtcars1 <- read_excel("D:/tempstore/moocr/mtcarsb.xlsx",
  sheet = "mtcars")
mtcars1 <- read_excel("D:/tempstore/moocr/mtcarsb.xlsx",
  sheet = 1)
head(mtcars1)

brain1<-read_excel("D:/tempstore/moocr/mtcarsb.xlsx",
  sheet = "brain")
head(brain1)

brainm<-read_excel("D:/tempstore/moocr/mtcarsb.xlsx",
  sheet = 2)
head(brainm)
```



```
> head(brainm)
# A tibble: 6 x 2
  wt    sex
<dbl> <chr>
1  1607    m
2  1157    m
3  1248    m
4  1310    m
5  1398    m
6  1237    m
```