Data Handling

데이터 import와 export

R의 내장 데이터셋

- R은 사용자들의 학습을 위해 다양한 데이터 셋을 제공
- 내장 데이터 셋 이외에도 패키지별로 통계 방법론 적용을 테스트해보기 위한 다양한 데이터 셋이 제공되기도 한다

- 내장 데이터 셋의 확인

```
> ?datasets
> library(help="datasets")
> data()
```

- R에서 직접 데이터를 입력할 수 있기는 하지만, 데이터의 입력과 기본적인 가공은 외부에서 한 후, R에서 불러오기(import) 할 것을 권장
  - R은 데이터 입력 도구라기 보다는 데이터 분석 도구임

외부 데이터 불러오기
: CSV

- CSV : Comma Separated Values
  - 데이터를 다루는 대부분의 프로그램(Excel, SAS, SPSS)에서 읽고 쓰기가 가능한 범용 데이터 파일
  - 기본적으로 값들이 콤마(,)로 구분되어 있는 형태의 파일
  - 구조가 간단하고 용량이 작아 널리 이용
- 주요 파라미터
  - header : 첫 번째 행에 변수명이 있는지 여부 (기본값 F)
  - sep : 구분자 (기본값 ',')
  - stringsAsFactors : TRUE이면 문자 타입을 Factor로 불러옴

```
read.csv({파일명}[, header = {헤더포함여부}, sep = {구분자},
    stringsAsFactors = {TRUE|FALSE}])
```

read.table 함수도 참조

외부 데이터 불러오기
: CSV

- thieves.txt를 불러와서 thieves 객체에 담아 봅시다

```
> thieves <- read.csv("thieves.txt")
> thieves
    홍길동.175.8.73.2
1 전우치\t170.2\t66.3
2 임꺽정\t186.7\t88.2
3   장길산\t188.3\t90
```

```
> thieves <- read.csv("thieves.txt", header = F, sep = "\t")
> thieves
     V1    V2   V3
1 홍길동 175.8 73.2
2 전우치 170.2 66.3
3 임꺽정 186.7 88.2
4 장길산 188.3 90.0
```

외부 데이터 불러오기
: CSV

- 컬럼에 이름 붙이기 : names()
  - 컬럼의 이름을 확인하고자 할 때에도 사용

```
> thieves
      V1    V2   V3
1 홍길동 175.8 73.2
2 전우치 170.2 66.3
3 임꺽정 186.7 88.2
4 장길산 188.3 90.0
> names(thieves)
[1] "V1" "V2" "V3"
> names(thieves) <- c("Name", "Height", "Weight")
> thieves
    Name Height Weight
1 홍길동  175.8   73.2
2 전우치  170.2   66.3
3 임꺽정  186.7   88.2
4 장길산  188.3   90.0
```

외부 데이터 불러오기
: Excel

- 엑셀 파일을 불러오기 위해서는 별도의 패키지를 이용
  - readxl 패키지를 설치하고 로드해야 함

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

  - read_excel() 함수를 이용하여 엑셀 파일을 로드할 수 있음

  - 주요 파라미터
    - col_names : 첫 번째 행을 변수명으로 불러올 것인지의 여부를 결정
    - sheet : 엑셀 파일 내 시트가 여러 개 있다면 해당 시트의 번호를 지정

```
read_excel({파일명}[, col_names = {헤더포함여부},
  stringsAsFactors = {TRUE|FALSE}, sheet = {시트 번호}])
```

외부 데이터 불러오기
: Excel

- wstudents.xlsx로부터 첫 번째 시트를 불러와 wstudents 객체에 저장해 봅시다

```
> wstudents <- read_excel("wstudents.xlsx")
> wstudents
# A tibble: 80 x 2
   height weights
    <dbl>   <dbl>
 1    151      48
 2    154      44
 3    160      48
 4    160      52
 5    163      58
 6    156      58
 7    158      62
 8    156      52
 9    154      45
10    160      55
# ... with 70 more rows
```

외부 데이터 불러오기
: from Web

- Web에는 우리가 상상할 수 있는 이상의 풍부한 데이터가 있음
- 통계 소프트웨어 개발 혹은 통계 학습에 도움이 되는 사이트
  - R 관련 많은 데이터 세트들을 제공함

사이트	URL 및 설명
R-DIR	https://r-dir.com
RDataMining	http://www.rdatamining.com
Kaggle	https://www.kaggle.com
RDatasets	https://github.com/vincentarelbundock/Rdatasets https://vincentarelbundock.github.io/Rdatasets/

```
> url <- "http://vincentarelbundock.github.com/Rdatasets/datasets.csv"
> datasets <- read.csv(url)
> datasets
```

데이터 내보내기 (export)

: csv로 데이터 저장

- 데이터 프레임을 write.csv를 이용하면 범용으로 사용할 수 있는 CSV 파일로 내보낼 수 있음 (write.table 함수도 참조)

```
> scores <- data.frame(english = c(80, 90, 70, 85),
+                      math = c(60, 70, 75, 65),
+                      kor = c(90, 95, 85, 80))

> # 아래 두 문장의 차이를 비교해 봅시다
> write.csv(scores, file = "scores.csv")
> write.csv(scores, file = "scores.csv", row.names = F)
```

데이터 내보내기 (export)

: RData 파일로 내보내기

- RData(.rda, .rdata) : R 전용 데이터 파일
  - R에서 빠르게 읽고 쓸 수 있으며 용량이 작다는 장점
  - R로 협업할 때는 RData 파일로, 타 프로그램을 이용 협업할 때는 csv를 추천
  - 저장하기 : save()
  - 불러오기 : load()

```
> scores
  english math kor
1      80   60  90
2      90   70  95
3      70   75  85
4      85   65  80

> save(scores, file = "scores.rda")
> rm(scores) # score 객체 삭제
> scores
Error: object 'scores' not found
> load("scores.rda")
> scores
  english math kor
1      80   60  90
2      90   70  95
3      70   75  85
4      85   65  80
```

Data Handling

데이터 살펴보기

데이터 살펴보기

- 데이터를 확보했으면 가장 먼저 데이터의 전반적인 구조를 이해해야 함
  - 어떤 변수들이 있는지
  - 몇 행으로 구성되어 있는지 등

- 데이터 파악을 위해 사용하는 함수들

함수	기능
head()	데이터의 앞부분 확인
tail()	데이터의 뒷부분 확인
View()	뷰어 창에서 데이터 내용 확인
dim()	데이터 차원 확인
str()	데이터의 속성 확인
summary()	요약 통계량 확인

데이터 살펴보기
: mtcars 데이터 파악

- 내장 데이터 mtcars의 데이터를 확인하고 살펴봅시다
- 데이터의 앞부분과 뒷부분 확인 : head(), tail()

```
> head(mtcars) # mtcars 앞부분을 출력(기본값 6행)
                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
> head(mtcars, n = 10) # mtcars 앞부분을 10행 출력

> tail(mtcars, n = 6) # mtcars 뒷부분을 6행 출력
                mpg cyl  disp  hp drat    wt qsec vs am gear carb
Porsche 914-2  26.0   4 120.3  91 4.43 2.140 16.7  0  1    5    2
Lotus Europa   30.4   4  95.1 113 3.77 1.513 16.9  1  1    5    2
Ford Pantera L 15.8   8 351.0 264 4.22 3.170 14.5  0  1    5    4
Ferrari Dino   19.7   6 145.0 175 3.62 2.770 15.5  0  1    5    6
Maserati Bora  15.0   8 301.0 335 3.54 3.570 14.6  0  1    5    8
Volvo 142E     21.4   4 121.0 109 4.11 2.780 18.6  1  1    4    2
> tail(mtcars, n = 10) # mtcars 뒷부분을 10행 출력
```

데이터 살펴보기
: mtcars 데이터 파악

- 내장 데이터 mtcars의 데이터를 확인하고 살펴봅시다
- 뷰어 창에서 데이터 확인 : View

```
> View(mtcars)
```

mtcars

Filter

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4

Showing 1 to 8 of 32 entries

데이터 살펴보기
: mtcars 데이터 파악

- 내장 데이터 mtcars의 데이터를 확인하고 살펴봅시다
- 데이터가 몇 열, 몇 행으로 구성되어 있는가 : dim()
- 데이터의 속성 파악 : str()

```
> dim(mtcars) # 행, 열 출력
[1] 32 11
```

mtcars는
- data frame이고
- 11개의 변수와
- 32개의 레코드를 가지고 있음

```
> str(mtcars)
'data.frame': 32 obs. of  11 variables:
 $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
 $ disp: num  160 160 108 258 360 ...
 $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
 $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
 $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
 $ qsec: num  16.5 17 18.6 19.4 17 ...
 $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
 $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
 $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
 $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

데이터 살펴보기
: mtcars 데이터 파악

- 내장 데이터 mtcars의 데이터를 확인하고 살펴봅시다
- 요약 통계량의 산출 : summary()

```
> summary(mtcars)
      mpg             cyl             disp             hp             drat
 Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0   Min.   :2.760
 1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5   1st Qu.:3.080
 Median :19.20   Median :6.000   Median :196.3   Median :123.0   Median :3.695
 Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7   Mean   :3.597
 3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0   3rd Qu.:3.920
 Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0   Max.   :4.930
       wt             qsec             vs              am              gear
 Min.   :1.513   Min.   :14.50   Min.   :0.0000   Min.   :0.0000   Min.   :3.000
 1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:3.000
 Median :3.325   Median :17.71   Median :0.0000   Median :0.0000   Median :4.000
 Mean   :3.217   Mean   :17.85   Mean   :0.4375   Mean   :0.4062   Mean   :3.688
 3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:4.000
 Max.   :5.424   Max.   :22.90   Max.   :1.0000   Max.   :1.0000   Max.   :5.000

 ...
```

데이터 살펴보기
: mtcars 데이터 파악

- 내장 데이터 mtcars의 데이터를 확인하고 살펴봅시다
- 범위를 좁혀 요약 통계량을 산출해 봅시다

```
> summary(mtcars[c("mpg", "wt")])
      mpg              wt
 Min.   :10.40   Min.   :1.513
 1st Qu.:15.43   1st Qu.:2.581
 Median :19.20   Median :3.325
 Mean   :20.09   Mean   :3.217
 3rd Qu.:22.80   3rd Qu.:3.610
 Max.   :33.90   Max.   :5.424
```

출력값	통계량	설명
Min	최소값	가장 작은 값
1st Qu.	1사분위수	하위 25% 지점에 위치하는 값
Median	중앙값	중앙에 위치하는 값
Mean	평균	산술평균
3rd Qu.	3사분위수	하위 75% 지점에 위치하는 값
Max	최대값	가장 큰 값

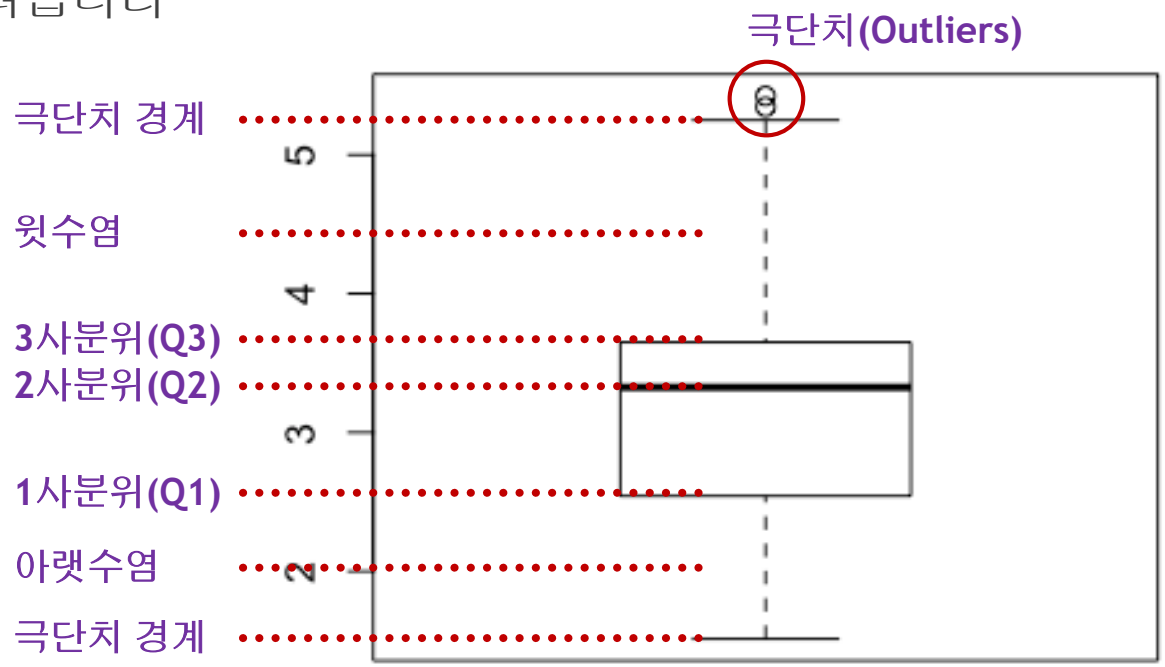
```
> quantile(mtcars$mpg)
    0%    25%    50%    75%   100%
10.400 15.425 19.200 22.800 33.900
> quantile(mtcars$wt)
     0%      25%      50%      75%     100%
1.51300 2.58125 3.32500 3.61000 5.42400
```

데이터 살펴보기
: mtcars 데이터 파악

- 내장 데이터 mtcars의 데이터를 확인하고 살펴봅시다
- boxplot으로 데이터의 분포와 구성 요소들을 살펴봅니다

```
> boxplot(mtcars$wt)
```

값	설명
아랫수염	하위 0~25%
1사분위수(Q1)	하위 25% 위치의 값
2사분위수(Q2)	하위 50% 위치의 값 중앙값(Median)
3사분위수(Q3)	상위 75% 위치의 값
윗수염	상위 75~100%
극단치 경계	
극단치(Outlier)	극단적으로 크거나 작은 값



데이터 살펴보기
: IQR

- boxplot의 값들은 $stat 변수를 참조하여 얻을 수 있음

```
> boxplot(mtcars$wt)$stat
       [,1]
[1,] 1.5130
[2,] 2.5425
[3,] 3.3250
[4,] 3.6500
[5,] 5.2500
```

- IQR(Interquartile Range)
  - 1사분위 ~ 3사분위 사이의 범위 :
    전체 데이터의 50%가 분포
  - 극단치(Outliers)를 찾아내는데 자주 사용

```
> IQR(mtcars$wt)
[1] 1.02875
```