

R 통계분석

강의내용

- R입문-데이터 다루기, 그래프 그리기
- 기술통계
- 확률분포
- 표본분포
- 모평균, 모비율의 추정 검정
- 두 모평균 차이, 두 모비율 차이 검정
- 상관분석, 단순회귀분석, 다중회귀분석
- 분산분석
- 공분산분석

R에 대해 알아보기

R이란?

- 다양한 통계분석과 그래프 작성 등을 위한 프로그래밍 언어이자 개발환경
- Open Source!
 - 통계학, 컴퓨터 분야 학자들로 구성된 R Development Core Team이 유지 관리
 - 세계 곳곳의 R user들의 활동으로 발전
 - 빅데이터 시대를 맞이해 주목
 - 구글과 같은 인터넷 기업에서 분석 엔진으로 사용

Why R?

- 무료
- 다양한 형태의 데이터 이용
- 거의 모든 분야의 통계분석 가능
- 그래프 작성의 용이
- 최첨단 분석기법들의 활발한 추가 (다양한 Packages)
- 대부분의 플랫폼에서 사용가능
- 활발한 사용자 community

- 단점
 - 다양한 package들의 사용법이 제각각
 - SPSS나 Minitab 등과 같은 통계분석프로그램에 비해 직관적이지 않은 사용법

R 설치하기

- R=base system + packages
- Base 시스템의 설치
 - Comprehensive R Archive Network (CRAN)
 - [http:// www.r-project.org](http://www.r-project.org)
 - Download, Packages > CRAN
 - 가까운 Mirror 선택
 - 원하는 version 선택 후 설치

R-studio 설치

- R의 기본 인터페이스를 사용해도 무방
- R-studio
 - 더 나은 인터페이스를 통해 R을 사용
 - 무료
 - 디버깅 편리
 - Workflow 확인가능
 - <http://www.rstudio.com> 에서 다운로드 및 설치

R-studio

Workspace

현재 메모리에
저장된 데이터
및 변수 나열

History

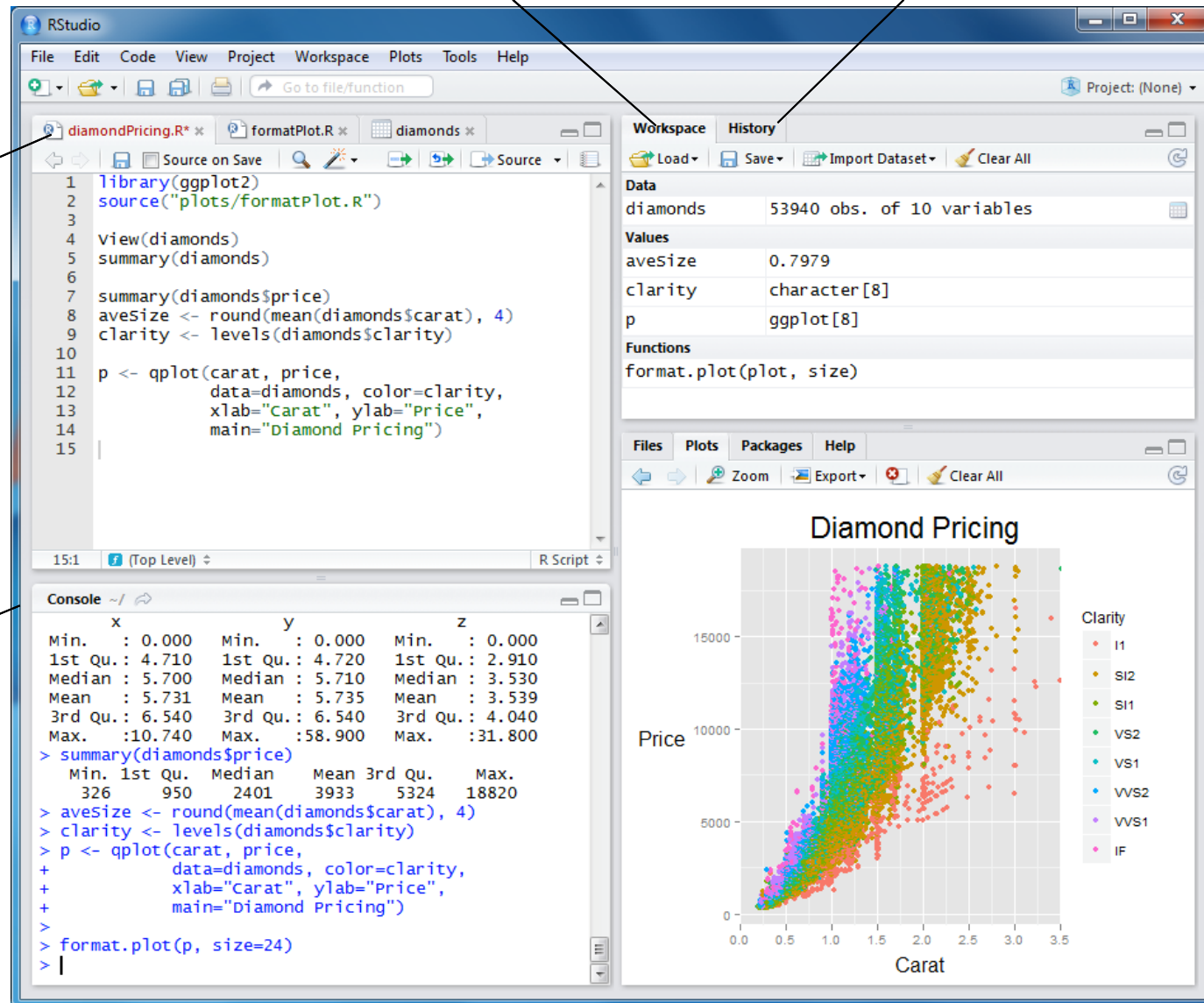
작업 history
열거

Source Editor

Batch mode로
작업수행을 위해
코드 편집 및
저장

Console

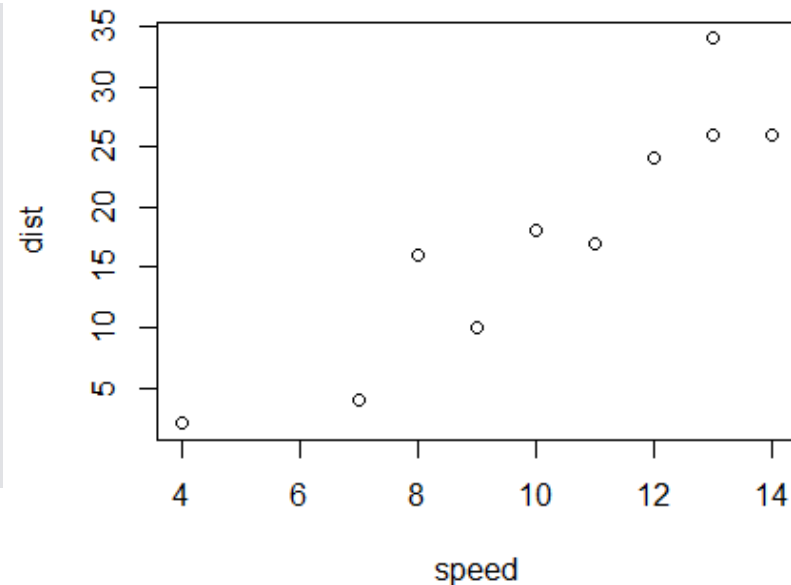
Interactive
mode로
작업 수행



예제

- Console에 입력 후 Enter
- 혹은 Editor에 타이핑 후
 - ✓ 한 줄씩 Ctrl+Enter (or Ctrl+R)
 - ✓ 실행시키고 싶은 부분 선택 후 Ctrl+Enter (or Ctrl+R)

```
> speed <- c(4, 7, 8, 9, 10, 11, 12, 13, 13, 14)
> di st=c(2, 4, 16, 10, 18, 17, 24, 34, 26, 26)
> summary(speed)
Min. 1st Qu. Median Mean 3rd Qu. Max.
4.00  8.25  10.50  10.10  12.75  14.00
> mean(di st) [1] 17.7
> sd(di st) [1] 10.22035
> cor(speed, di st)
[1] 0.9176971
> pl ot(speed, di st)
```



Workspace

- R의 실행과정에서 생기는 모든 객체 및 명령문 등이 보관되어 있는 공간
- `ls()`: 작업공간에 임시 저장되어 있는 객체 리스트
- 작업 디렉토리
 - `getwd()`: 현재 작업 디렉토리 확인
 - `setwd("c:/data")`: 작업 디렉토리 변경
 - `save.image()`: 현재 작업공간을 .RData 파일로 저장
 - `load()`: 저장된 작업공간 불러오기

작업결과 저장

- Output 저장

- sink()

```
> sink("output.txt")  
> mean(di st)  
> sd(di st)  
> sink()
```

- sink 문 사이의 작업 결과를 외부 파일로 일괄저장

- 그래프 저장

- Plots 창의 Export 메뉴 이용

- pdf(), postscript(), bmp(), jpeg() 사용 후 dev.off() 실행

```
pdf("pl ot1. pdf")  
> pl ot(di st, speed)  
> dev. off()
```

R의 확장: Packages

- R Packages: R의 막강 파워의 원동력
- 사용자가 작성한 모듈
- 다양한 분야의 분석도구 제공
- 2014년 2월 현재 5246개의 패키지
- 각 패키지의 특성을 모두 파악하는 것은 불가능
- 사용자가 필요한 특정 분석기법이 어떤 패키지에 있나 알기 어려움 ➔ Google!
- 개별 사용자가 직접 작성하여 올린 것이므로 간혹 문제가 있는 패키지가 있을 수 있음

Packages

- Base packages
 - R을 설치할 때 시스템의 일부분으로 기본 설치
 - base, datasets, graphics, grid, methods, stats, utils 등
- Recommended packages
 - R 설치 시 기본 설치되지만 사용을 위해서 R 세션으로 불러들여야 함
 - MASS, foreign, lattice 등
 - library(): 패키지 불러들임

```
> ?read.spss
No documentation for 'read.spss' in specified packages and libraries: you could try
'??read.spss'
> library("foreign")
> ?read.spss
```

Packages

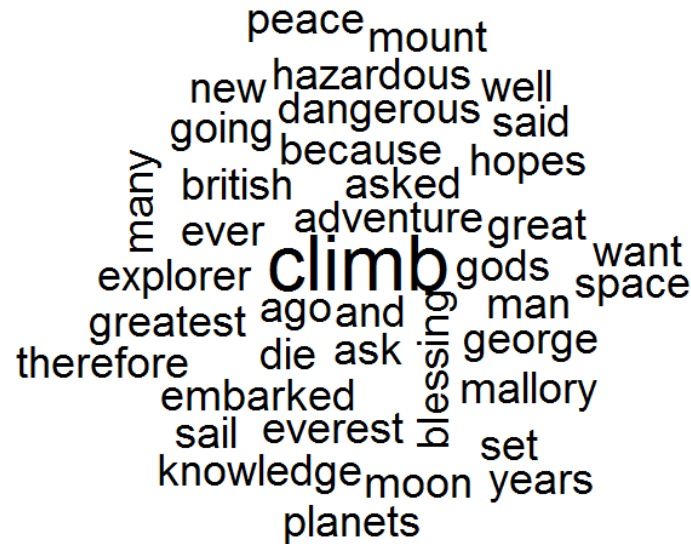
Files	Plots	Packages	Help	Viewer	
<div><div> Install Packages</div><div> Check for Updates</div><div></div></div>					<div><input type="text" value=""/></div>
<input type="checkbox"/>	boot	Bootstrap Functions (originally by Angelo Canty for S)	1.3-9		
<input type="checkbox"/>	class	Functions for Classification	7.3-9		
<input type="checkbox"/>	cluster	Cluster Analysis Extended Rousseeuw et al.	1.14.4		
<input type="checkbox"/>	codetools	Code Analysis Tools for R	0.2-8		
<input type="checkbox"/>	compiler	The R Compiler Package	3.0.2		
<input checked="" type="checkbox"/>	datasets	The R Datasets Package	3.0.2		
<input type="checkbox"/>	foreign	Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, dBase, ...	0.8-55		
<input checked="" type="checkbox"/>	graphics	The R Graphics Package	3.0.2		
<input checked="" type="checkbox"/>	grDevices	The R Graphics Devices and Support for Colours and Fonts	3.0.2		
<input type="checkbox"/>	grid	The Grid Graphics Package	3.0.2		
<input type="checkbox"/>	KernSmooth	Functions for kernel smoothing for Wand & Jones (1995)	2.23-10		
<input type="checkbox"/>	lattice	Lattice Graphics	0.20-23		
<input type="checkbox"/>	manipulate	Interactive Plots for RStudio	0.98.501		
<input type="checkbox"/>	MASS	Support Functions and Datasets for Venables and Ripley's MASS	7.3-29		
<input type="checkbox"/>	Matrix	Sparse and Dense Matrix Classes and Methods	1.0-14		
<input checked="" type="checkbox"/>	methods	Formal Methods and Classes	3.0.2		
<input type="checkbox"/>	mgcv	Mixed GAM Computation Vehicle with GCV/AIC/REML smoothness estimation	1.7-26		
<input type="checkbox"/>	nlme	Linear and Nonlinear Mixed Effects Models	3.1-111		
<input type="checkbox"/>	nnet	Feed-forward Neural Networks and Multinomial Log-Linear Models	7.3-7		
<input type="checkbox"/>	parallel	Support for Parallel computation in R	3.0.2		
<input type="checkbox"/>	plyr	Tools for splitting, applying and combining data	1.8		
<input type="checkbox"/>	rpart	Recursive Partitioning	4.1-3		
<input type="checkbox"/>	rstudio	Tools and Utilities for RStudio	0.98.501		
<input type="checkbox"/>	spatial	Functions for Kriging and Point Pattern Analysis	7.3-7		
<input type="checkbox"/>	splines	Regression Spline Functions and Classes	3.0.2		
<input checked="" type="checkbox"/>	stats	The R Stats Package	3.0.2		
<input type="checkbox"/>	stats4	Statistical Functions using S4 Classes	3.0.2		
<input type="checkbox"/>	survival	Survival Analysis	2.37-4		
<input type="checkbox"/>	tcltk	Tcl/Tk Interface	3.0.2		
<input type="checkbox"/>	tools	Tools for Package Development	3.0.2		
<input type="checkbox"/>	translations	The R Translations Package	3.0.2		
<input checked="" type="checkbox"/>	utils	The R Utils Package	3.0.2		

Package의 설치 및 사용

- `install.packages()`: 패키지 설치
- `library()`: 설치된 패키지를 R 세션으로 불러들임
- `installed.packages()`: 설치된 패키지들의 정보

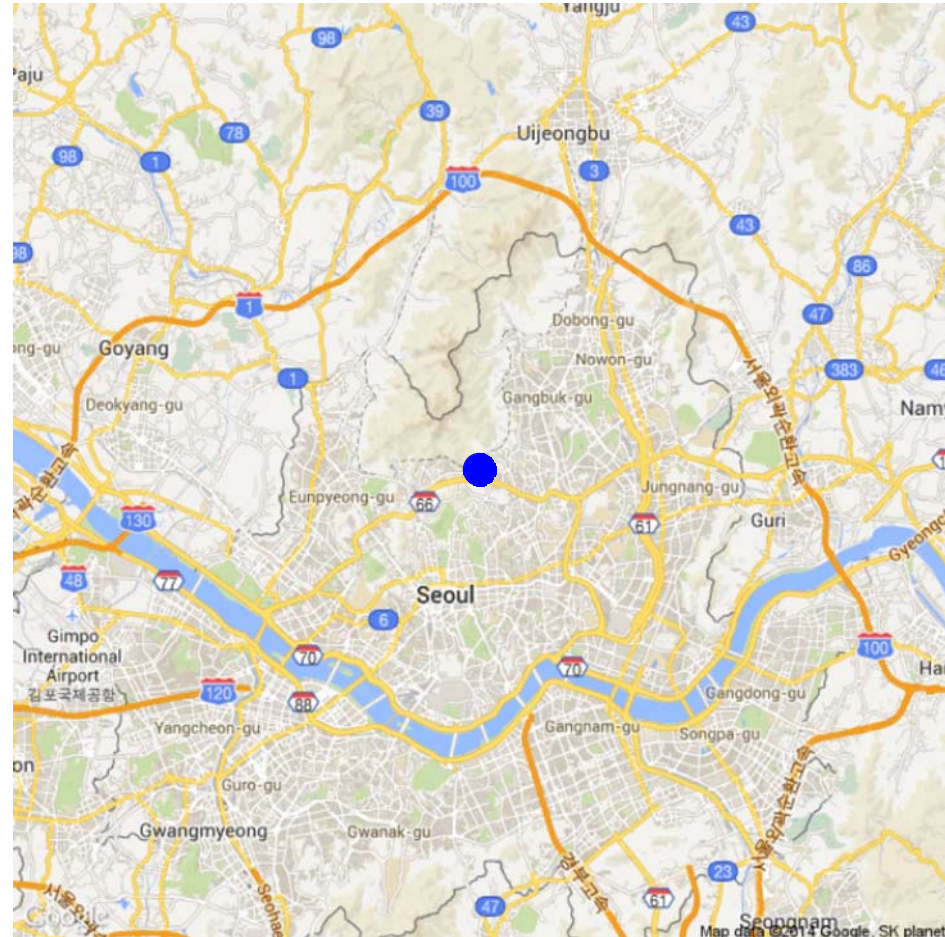
Package 활용의 예: Word Cloud

```
>install.packages("wordcloud")
>install.packages("tm")
>library("wordcloud")
필요한 패키지를 로딩중입니다: Rcpp 필요한 패키지를 로딩중입니다: RColorBrewer
>library("tm")
>wordcloud( + "Many years ago the great British explorer George Mallory, who
+ was to die on Mount Everest, was asked why did he want to climb
+ it. He said, ...)
```



Package 활용의 예: Google Map

```
> library(RgoogleMaps)
> map.center.loc <- c(37.6121866,
126.9953410)
> input.zoom <- 11 > win.graph()
> mymap <- GetMap(center =
map.center.loc,
+ zoom = input.zoom, maptype =
"road", format
+ = "roadmap", destfile =
"mymap.png")
[1]
"http: //maps. google. com/maps/api /st
aticmap?center=37.6121866,126.99534
1&zoom=11&size=640x640&maptype=road
&format=roadmap&sensor=true"
> PlotOnStaticMap(mymap, lat =
37.6121866, lon = 126.9953410, +
destfile = "mymap.poi nt. png", cex =
5, pch=20, col ="bl ue")
```



유용한 package 찾기 및 도움말

- R Project (<http://www.r-project.org>)
- Korean R Users Group (<http://www.r-project.kr>)
- R seek (<http://rseek.org>)
- R wiki (<http://wiki.r-project.org>)
- GOOGLE!
- R Console에서 ? 명령어, ?? 키워드

R에서 데이터 준비하기

데이터 유형 및 구조

- 데이터 유형
 - numeric: 1, 0.1, 10^{10} 등 숫자
 - character: “a”, “Yejin”, “Kookmin” 등 문자
 - logical: TRUE / FALSE
- 데이터 구조
 - vector
 - factor
 - matrix
 - array
 - data frame
 - list

Vector

- c()로 생성
- numeric, character, logical 유형의 벡터
- 하나의 벡터에는 동일한 유형만 입력 가능

```
> x=c(1,3,5,-4)
> y=c("a","b","c")
> z=c(TRUE,FALSE, FALSE, TRUE)
> c(3,TRUE,FALSE)
[1] 3 1 0
> c(3,"1",TRUE,FALSE)
[1] "3"      "1"      "TRUE"   "FALSE"
```

- 벡터의 인덱싱
 - 벡터의 일부분만 선택
 - 대괄호 []를 이용

```
> x[2]
[1] 3
> x[c(1,3)]
[1] 1 5
> x[c(TRUE,FALSE,FALSE,TRUE)]
[1] 1 -4
> x[5]
[1] NA
> x[1:3]
[1] 1 3 5
```

Factor (요인)

- 범주형 데이터를 위한 구조
- level (수준): 요인이 가질 수 있는 값들
- factor(): 숫자형 혹은 문자형 벡터를 요인으로 변환

```
> gender=c("male","female","male","male","female")
> f.gender=factor(gender)
> gender
[1] "male" "female" "male" "male" "female"
> f.gender
[1] male female male male female
Levels: female male
> gender2=c(1,0,1,1,0)
> f.gender2=factor(gender2)
> gender2
[1] 1 0 1 1 0
> f.gender2
[1] 1 0 1 1 0
Levels: 0 1
```

Matrix (행렬)

- 동일한 유형의 데이터로 이루어진 2차원 배열
- `matrix(데이터 벡터, nrow=행의 수, ncol=열의 수, byrow=TRUE 또는 FALSE, dimnames=list(행이름 벡터, 열이름 벡터))`
- `byrow`: 데이터를 행단위로 채울 것인지 (TRUE) 열단위로 채울 것인지 (FALSE: default)

```
> x=matrix(1:9,nrow=3)
> x
      [,1] [,2] [,3]
[1,]    1    4    7
[2,]    2    5    8
[3,]    3    6    9
> y=matrix(1:9, nrow=3, byrow=TRUE)
> y
      [,1] [,2] [,3]
[1,]    1    2    3
[2,]    4    5    6
[3,]    7    8    9
> rnames=c("R1","R2","R3")
> cnames=c("C1","C2","C3")
> z=matrix(1:9, nrow=3,dimnames=list(rnames, cnames))
> z
      C1 C2 C3
R1    1  4  7
R2    2  5  8
R3    3  6  9
```

Matrix (행렬)

- 인덱싱
 - $x[i,j]$: x 의 i 행 j 열 원소
 - $x[i,]$: x 의 i 행
 - $x[,j]$: x 의 j 열

```
> x
      [,1] [,2] [,3]
[1,]    1    4    7
[2,]    2    5    8
[3,]    3    6    9
> x[2,3]
[1] 8
> x[1,]
[1] 1 4 7
> x[,2]
[1] 4 5 6
> x[c(1,2),2]
[1] 4 5
> x[,-1]
      [,1] [,2]
[1,]    4    7
[2,]    5    8
[3,]    6    9
> x[c(-1,-2),]
[1] 3 6 9
```


Array (배열)

- 행렬과 유사
- 2차원 이상의 구조
- array(데이터 벡터, 차원의 정의, dimnames=list(각 차원에 대한 라벨 벡터))

```
> y=array(1:24,c(4,3,2))
```

```
> y
```

```
, , 1
```

	[,1]	[,2]	[,3]
[1,]	1	5	9
[2,]	2	6	10
[3,]	3	7	11
[4,]	4	8	12

```
, , 2
```

	[,1]	[,2]	[,3]
[1,]	13	17	21
[2,]	14	18	22
[3,]	15	19	23
[4,]	16	20	24

Data Frame (데이터 프레임)

- 행렬과 같은 2차원 구조
- 모든 데이터가 동일한 유형일 필요가 없음
- SAS, SPSS 등에서 접하는 데이터셋의 개념
- data.frame(변수1, 변수2, 변수3, ...)

```
> x1=c(24,28,31,25)
> y1=c("F","M","F","F")
> xy=data.frame(x1,y1)
> xy
  x1 y1
1 24  F
2 28  M
3 31  F
4 25  F
> xy=data.frame(age=x1,gender=y1)
> xy
  age gender
1  24      F
2  28      M
3  31      F
4  25      F
```

Data Frame 행, 열 추가

- `rbind()`: 기존의 데이터 프레임에 행 벡터를 추가 (즉, 관찰치 추가)
- `cbind()`: 기존의 데이터 프레임에 열 벡터를 추가 (즉, 변수 추가)

```
> xy_add=data.frame(age=32, gender="M")
> xy=rbind(xy,xy_add)
> xy
  age gender
1  24      F
2  28      M
3  31      F
4  25      F
5  32      M

> xyz=cbind(xy,income=c(2000,3100,3800,2800,3000))
> xyz
  age gender income
1  24      F  2000
2  28      M  3100
3  31      F  3800
4  25      F  2800
5  32      M  3000
```

Data Frame에서 변수선택

```
> xyz[[1]]  
[1] 24 28 31 25 32  
> xyz[1]  
  age  
1  24  
2  28  
3  31  
4  25  
5  32  
> xyz[["age"]]  
[1] 24 28 31 25 32  
> xyz$age  
[1] 24 28 31 25 32  
> xyz["age"]  
  age  
1  24  
2  28  
3  31  
4  25  
5  32
```

대괄호 두 개 `[[]]`: 벡터로 출력

대괄호 한 개 `[]`:
데이터프레임으로 출력

Data Frame에서 여러 개 변수 선택

```
> xyz[c(1,3)]
  age income
1  24   2000
2  28   3100
3  31   3800
4  25   2800
5  32   3000
> xyz[c("age","income")]
  age income
1  24   2000
2  28   3100
3  31   3800
4  25   2800
5  32   3000
> xyz[,c(1,3)]
  age income
1  24   2000
2  28   3100
3  31   3800
4  25   2800
5  32   3000
> xyz[c(2,4,5),]
  age gender income
2  28      M   3100
4  25      F   2800
5  32      M   3000
```

Data frame도 2차원
배열이므로 Matrix의
인덱싱 적용 가능

List

- R에서 가장 포괄적인 형태의 데이터 구조
- 구성요소로서 벡터, 배열, 데이터 프레임, 함수, 다른 리스트 등 가능
- 다른 유형의 객체를 한데 묶은 또 다른 객체
- `list(객체1, 객체2, 객체3, ...)`

```
> a=c("a","b","c")
> b=1:10
> c=matrix(1:9,3,3)
> L=list(vec=a,b,mat=c,xyz)
> L
$vec
[1] "a" "b" "c"

[[2]]
[1] 1 2 3 4 5 6 7 8 9 10

$mat
      [,1] [,2] [,3]
[1,]    1    4    7
[2,]    2    5    8
[3,]    3    6    9

[[4]]
  age gender income
1  24      F   2000
2  28      M   3100
3  31      F   3800
4  25      F   2800
5  32      M   3000
```

객체의 구조에 대한 정보를 얻는 함수

- `length()`: 객체를 구성하는 요소의 개수
- `dim()`: 객체의 차원
- `names()`: 객체 구성요소들의 이름
- `str()`: 객체 구조
- `head()`: 객체의 처음 부분
- `tail()`: 객체의 끝 부분

```
> airquality
```

	Ozone	Solar.R	wind	Temp	Month	Day
1	41	190	7.4	67	5	1
2	36	118	8.0	72	5	2
3	12	149	12.6	74	5	3
4	18	313	11.5	62	5	4
5	NA	NA	14.3	56	5	5
6	28	NA	14.9	66	5	6
7	23	299	8.6	65	5	7
8	19	99	13.8	59	5	8
9	8	19	20.1	61	5	9
10	NA	194	8.6	69	5	10
11	7	NA	6.9	74	5	11
12	16	256	9.7	69	5	12

```
> str(airquality)
```

```
'data.frame': 153 obs. of 6 variables:
 $ Ozone   : int  41 36 12 18 NA 28 23 19 8 NA ...
 $ Solar.R: int  190 118 149 313 NA NA 299 99 19 194 ...
 $ wind    : num  7.4 8 12.6 11.5 14.3 14.9 8.6 13.8 20.1 8.6 ...
 $ Temp    : int  67 72 74 62 56 66 65 59 61 69 ...
 $ Month   : int   5 5 5 5 5 5 5 5 5 5 ...
 $ Day     : int   1 2 3 4 5 6 7 8 9 10 ...
```

```
> length(airquality)
```

```
[1] 6
```

```
> dim(airquality)
```

```
[1] 153 6
```

```
> names(airquality)
```

```
[1] "Ozone" "Solar.R" "wind" "Temp" "Month" "Day"
```

```
> head(airquality)
```

	Ozone	Solar.R	wind	Temp	Month	Day
1	41	190	7.4	67	5	1
2	36	118	8.0	72	5	2
3	12	149	12.6	74	5	3
4	18	313	11.5	62	5	4
5	NA	NA	14.3	56	5	5
6	28	NA	14.9	66	5	6

```
> tail(airquality)
```

	Ozone	Solar.R	wind	Temp	Month	Day
148	14	20	16.6	63	9	25
149	30	193	6.9	70	9	26
150	NA	145	13.2	77	9	27
151	14	191	14.3	75	9	28
152	18	131	8.0	76	9	29
153	20	223	11.5	68	9	30

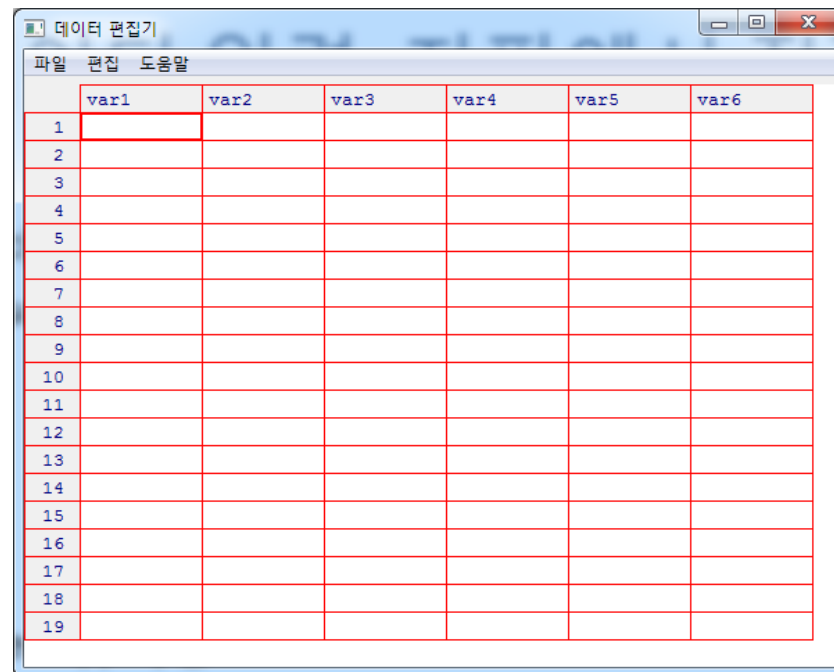
데이터 입력: 자판에서 직접 입력

- `c()`, `matrix()` 사용
- `scan()` 사용

```
> x=scan()  
1: 24  
2: 35  
3: 28 21  
5:  
Read 4 items  
> x  
[1] 24 35 28 21  
> y=scan(what="character")  
1: park kim lee  
4: chung  
5:  
Read 4 items  
> y  
[1] "park" "kim" "lee" "chung"
```

- `edit()` 사용

```
> xyz=data.frame()  
> edit(xyz)
```



	var1	var2	var3	var4	var5	var6
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						
11						
12						
13						
14						
15						
16						
17						
18						
19						

데이터 입력: 외부파일 불러오기

- 다양한 형태의 데이터 import 가능
 - 내장함수: read.table, read.csv, read.delim을 사용해 텍스트 파일 import
 - foreign 패키지: Minitab, S, SAS, SPSS, Stata, Systat, dBase 등의 데이터 import
 - xlsReadWrite 패키지, xlsx 패키지: Excel파일 import
 - RODBC 패키지: MS access 데이터 import
- csv (comma separated values) 형식으로 저장해 불러들이는 것 추천

데이터 입력: csv 파일

- Excel 등의 데이터를 csv 형식으로 저장
- 영어로 된 파일 명, 변수 명, 데이터 사용 추천

```
> setwd("c:/Rdata")
> data=read.csv("chickwts.csv")
> head(data)
  chick weight    feed
1     1   179 horsebean
2     2   160 horsebean
3     3   136 horsebean
4     4   227 horsebean
5     5   217 horsebean
6     6   168 horsebean
> dim(data)
[1] 71  3
```

데이터 저장

- R에서 작업한 데이터 개체를 외부 파일로 저장
- `write.table()`, `write.csv()`
 - > `write.table(women, "women1.txt")`
 - > `write.table(women, "women2.txt", row.names=FALSE)`
 - > `write.csv(women, "women.csv", row.names=FALSE)`

