



[2020 혁신성장 청년인재 집중양성 사업]

프로젝트 기반 데이터 과학자 양성과정

# 빅데이터 분석

- 2주차 -

#R #프로그래밍 #데이터전처리



# A table of Contents

- 1 R 설치 및 실행
- 2 데이터 입출력
- 3 데이터 타입
- 4 R 프로그래밍
- 5 데이터 전처리 I
- 6 Summary

# Part 1

R 설치 및 실행



# R 설치

## R 설치과정

- 홈페이지(<https://www.r-project.org/>)에 접속해서 "CRAN"메뉴로 이동
- Korea CRAN 서버를 이용하여 다운로드

The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To [download R](#), please choose your preferred [CRAN mirror](#).

If you have questions about R like how to download and install the software, or what the license terms are you send an email.

Download

**CRAN**

CRAN 메뉴를 클릭하여 이동

News

- **R version 4.0.2 (Taking Off Again) prerelease versions** will appear starting Friday 2020-06-12. Final release is scheduled for Monday 2020-06-22.
- **R version 4.0.1 (See Things Now)** has been released on 2020-06-06.
- **useR! 2020** in Saint Louis has been cancelled. The European hub planned in Munich will not be an in-person conference. Both organizing committees are working on the best course of action.
- **R version 3.6.3 (Holding the Windsock)** has been released on 2020-02-29.
- You can support the R Foundation with a renewable subscription as a [supporting member](#)

Japan

- <https://cran.ism.ac.jp/>
- <https://ftp.yz.yamagata-u.ac.jp/pub/cran/>

Korea

- <https://ftp.harukasan.org/CRAN/>
- <https://cran.yu.ac.kr/>
- <https://cran.seoul.go.kr/>**
- <https://cran.biodisk.org/>

Korea CRAN 서버를 이용

Malaysia

- <https://wbc.upm.edu.my/cran/>

Mexico

- <https://cran.itam.mx/>
- <http://www.est.colpos.mx/R-mirror/>



# R 설치

## R 설치과정

- OS에 맞는 버전 다운로드
- Base를 이용하여 설치

Download and Install R

Precompiled binary distributions of the base system and contributed packages

- [Download R for Linux](#)
- [Download R for \(Mac\) OS X](#)
- [Download R for Windows](#)

OS에 맞는 버전 다운로드

R is part of many Linux distributions, you should check with your system administrator

Subdirectories:

<a href="#">base</a>	Binaries for base distributions
Base 설치	Binaries of contributed CRAN Windows services
<a href="#">old contrib</a>	Binaries of contributed CRAN Windows services
<a href="#">Rtools</a>	Tools to build R and R packages

Please do not submit binaries to CRAN. Package developers should submit source code only.

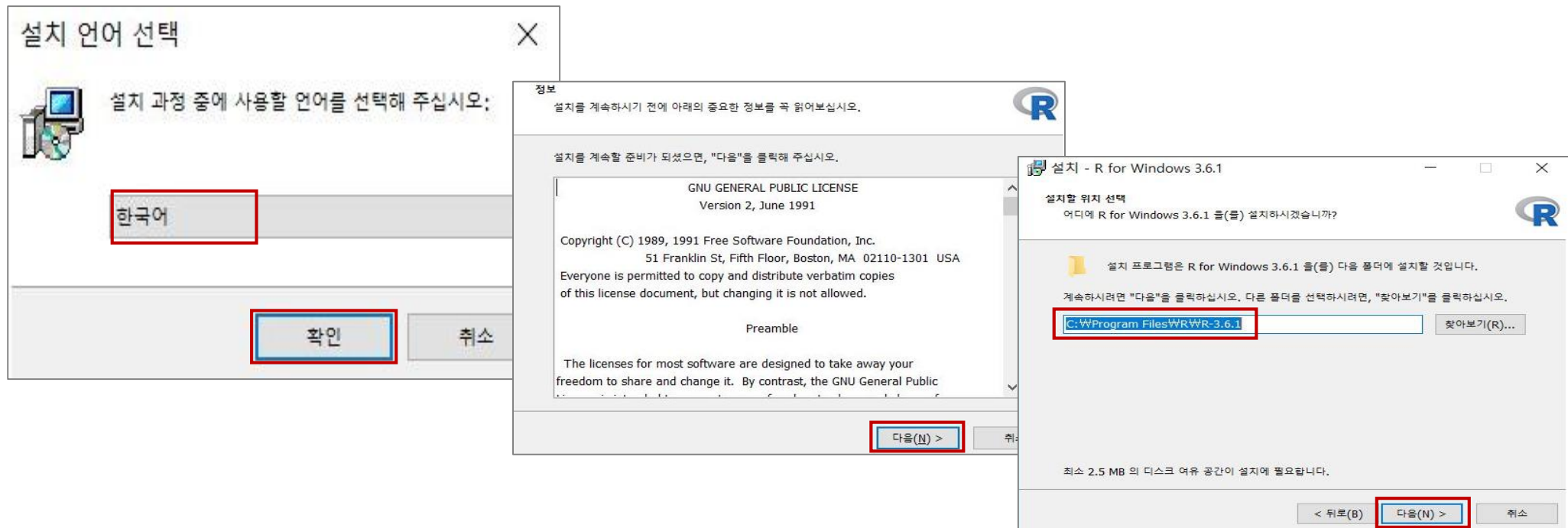
You may also want to read the [R FAQ](#) and [R for Windows FAQ](#)

Note: CRAN does some checks on these binaries for viruses, but it is not a substitute for your own security checks.

# >>>>> R 설치

## R 설치과정

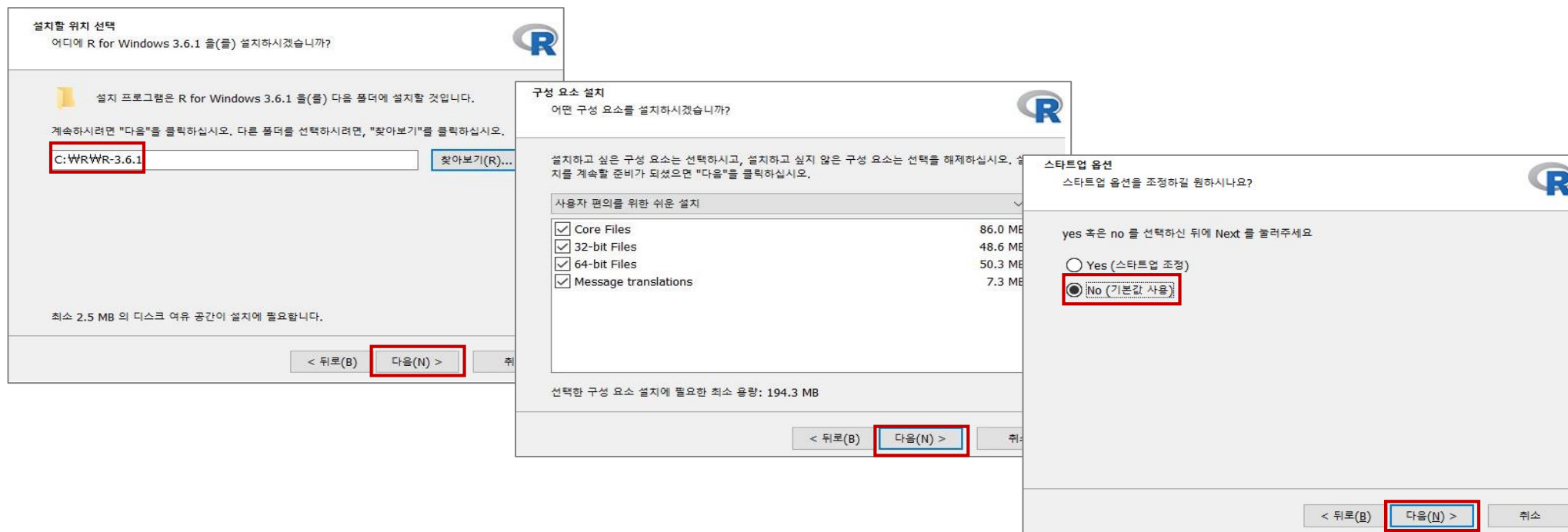
- 설치파일 실행
- 한국어 선택 > 다음 클릭 > 다음 클릭 > 경로 선택 후 다음 클릭 > 다음 클릭 > 다음 클릭 > 설치 완료



# >>>>> R 설치

## R 설치과정

- 설치파일 실행
- 한국어 선택 > 다음 클릭 > 다음 클릭 > 경로 선택 후 다음 클릭 > 다음 클릭 > 다음 클릭 > 설치 완료

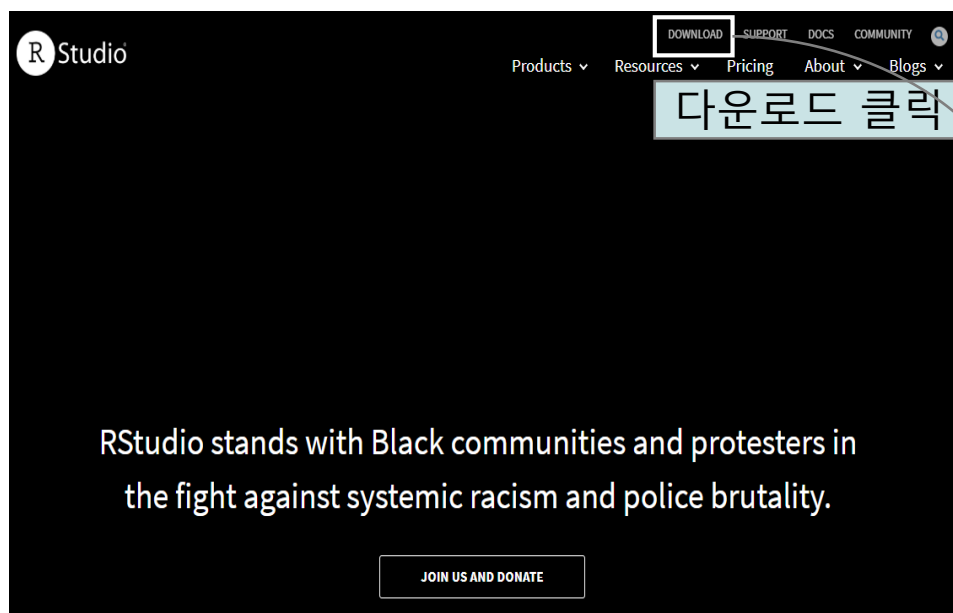




# R Studio 설치

## R Studio 설치 과정

- R Studio 홈페이지 (<https://www.rstudio.com/>)에 접속
- Download R Studio를 클릭하여 다운로드 페이지로 이동



RStudio Desktop	RStudio Desktop	RStudio Server	RStudio Server Pro
Open Source License	Commercial License	Open Source License	Commercial License
Free	\$995	Free	\$4,975
	/year		/year (5 Named Users)
<a href="#">Learn more</a>	<a href="#">Learn more</a>	<a href="#">Learn more</a>	<a href="#">Evaluation</a>   <a href="#">Learn more</a>
Desktop Free 버전으로 다운로드			
Integrated Tools for R			
Priority Support	✓		✓
Access via Web Browser		✓	✓
Enterprise Security			✓





# R Studio 설치

## R Studio 설치 과정

- Install버전에서 OS 선택 후, 다운로드

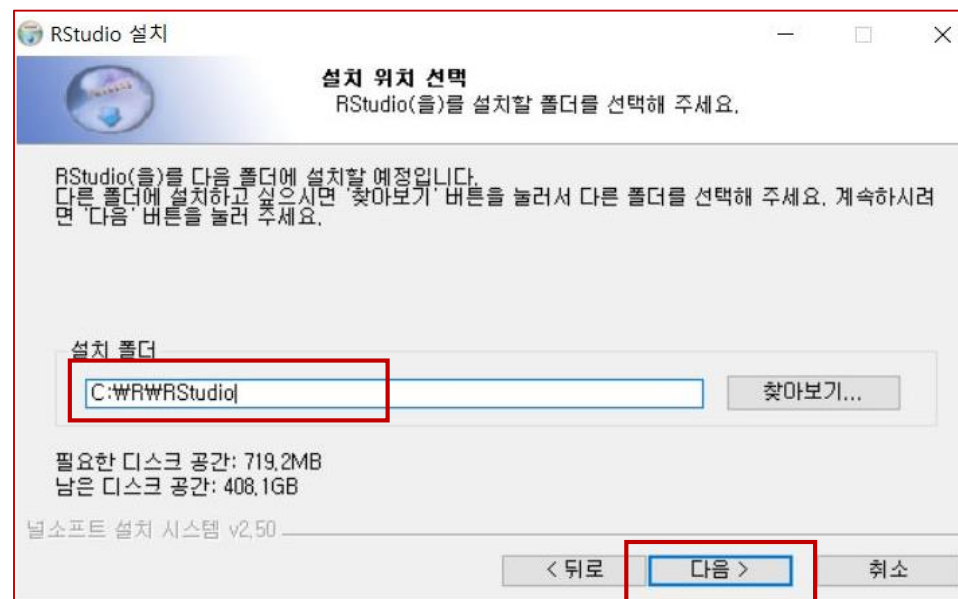
All Installers			
Linux users may need to <a href="#">import RStudio's public code-signing key</a> prior to installation, depending on the operating system's security policy.			
RStudio requires a 64-bit operating system. If you are on a 32 bit system, you can use an <a href="#">older version of RStudio</a> .			
OS	Download	Size	SHA-256
Windows 10/8/7	<a href="#">RStudio-1.3.959.exe</a>	171.41 MB	3d493ae5
			7c5b695d
			c2931495
Ubuntu 18/Debian 10	<a href="#">rstudio-1.3.959-amd64.deb</a>	126.11 MB	411ab500
Fedora 19/Red Hat 7	<a href="#">rstudio-1.3.959-x86_64.rpm</a>	146.24 MB	a144e4e6
Fedora 28/Red Hat 8	<a href="#">rstudio-1.3.959-x86_64.rpm</a>	150.32 MB	57169bee
Debian 9	<a href="#">rstudio-1.3.959-amd64.deb</a>	126.42 MB	b2d9366f
SLES/OpenSUSE 12	<a href="#">rstudio-1.3.959-x86_64.rpm</a>	119.02 MB	bbc9397e
OpenSUSE 15	<a href="#">rstudio-1.3.959-x86_64.rpm</a>	127.59 MB	a4f404f0



# R Studio 설치

## R Studio 설치 과정

- R Studio 설치 파일을 시작
- 다음 클릭 > 경로 선택 후, 다음 클릭

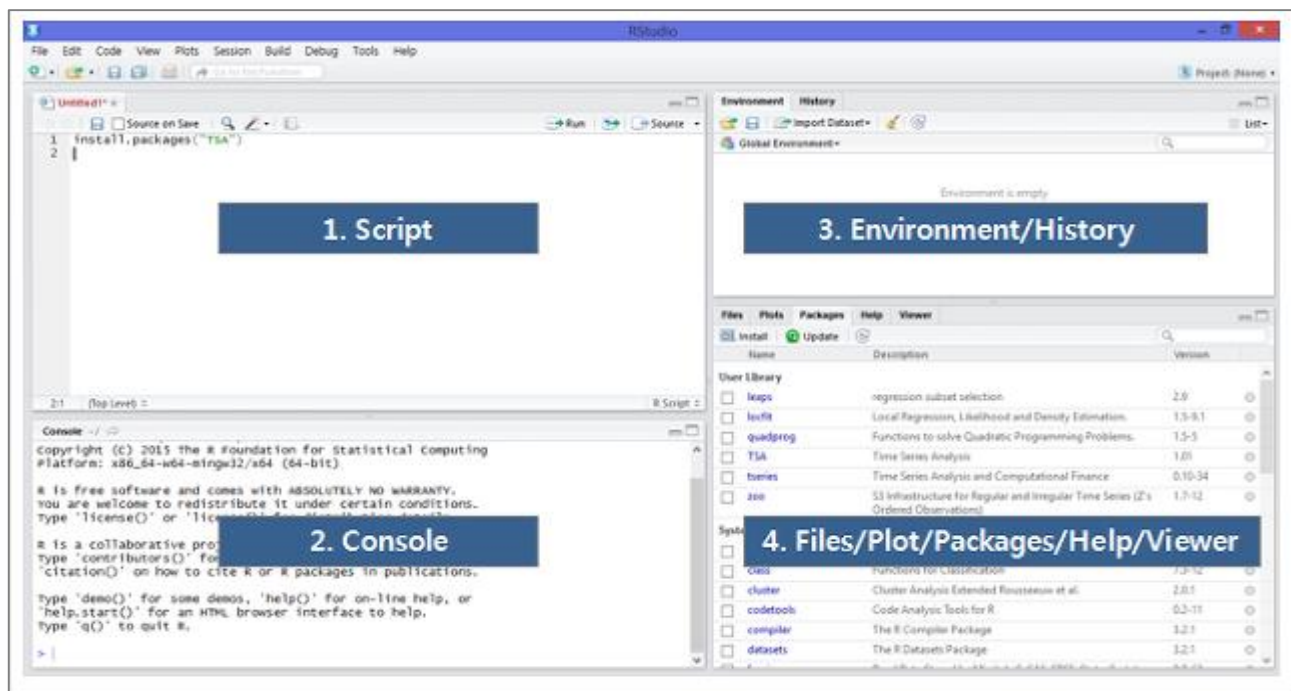




# R Studio 설치

## R Studio 화면 설명

- R Studio는 크게 Script, Console, Environment/History, File/Plot/Packages/Help/Viewer로 나뉨



- 1) R Script 창에서 작성한 Script를 실행하려면 Ctrl + R / Ctrl + Enter / Run을 클릭
- 2) Console 창은 Interactive 하게 R 프로그램을 짜고 실행하기 / R Script 창 혹은 Console 창에서 작성한 프로그램의 실행(계산) 결과 보기 / 패키지 설치, 에러/오류 메시지 등의 로그 보기
- 3) Environment탭에는 데이터셋이 확인 가능 / History탭에는 R Script 의 이력 확인 가능
- 4) Files/Plot/Packages/Help/Viewer
  - Files: 탐색기 기능
  - Plot: 그래프 확인
  - Packages: Package 설치 확인
  - Help: 도움말 검색 기능
  - Viewer : 웹으로 출력했을 때 확인 가능



# R 기본 활용

## R 기본 활용

- R에서 변수명은 특정한 규칙을 따름
- 리눅스의 명령어와 유사

- R의 변수명은 알파벳, 숫자, \_(언더스코어), .(마침표)로 구성되며, -(하이픈)은 사용할 수 없다.
- 이름의 첫 글자로 숫자와 '\_'은 사용할 수 없음
- 대문자와 소문자는 서로 구분
- 변수값은 <-, =을 사용함
- 대부분 <-을 사용하여 변수값 할당함
- ;(세미콜론)은 명령문과 명령문을 구분 짓는 역할
- ls() : 생성된 변수의 리스트 출력
- rm() : 생성된 변수를 삭제
- rm(list=ls()) : 모든 변수 삭제



# R 기본 활용

## R 패키지

- R에는 다양한 사용자들이 구축해 높은 방대한 양의 패키지가 존재
- 인터넷인 연결되어 있는 환경에서는 아래와 같은 명령어로 다운로드가 가능함

```
> install.packages("randomForest")
--- Please select a CRAN mirror for use in this session ---
Loading Tcl/Tk interface ... done
trying URL 'http://cran.nexr.com/bin/macosx/leopard/contrib/2.15/randomForest_4.6
Content type 'application/x-gzip' length 206287 bytes (201 Kb)
opened URL
=====
downloaded 201 Kb

The downloaded binary packages are in
/var/folders/3t/kmb3l9cn5bxf6m020rhnpshm0000gp/T//RtmpdmgXpE/downloaded_packages
```

```
> library(randomForest)
randomForest 4.6-7
Type rfNews() to see new features/changes/bug fixes.
```



# R 기본 활용

## R 도움말

- 활용하고자 하는 함수에 대한 도움말을 얻고자 할 때 help나 ? 명령어를 사용

```
> ?print
```

```
print      package:base      R Documentation
```

```
Print Values
```

```
Description:
```

```
'print' prints its argument and returns it _invisibly_ (via  
'invisible(x)'). It is a generic function which means that new  
printing methods can be easily added for new 'class'es.
```

```
Usage:
```

```
print(x, ...)
```

```
## S3 method for class 'factor'
```

```
print(x, quote = FALSE, max.levels = NULL,  
      width = getOption("width"), ...)
```

# Part 2

데이터 입출력



# 데이터 입출력

## 데이터 입출력

- CSV 파일을 데이터 프레임으로 읽으려면 `read.csv()`
- 데이터 프레임을 CSV파일로 저장하려면 `write.csv()`

- `ls();rm(list=ls())`
- `data(iris)`
- `head(iris)` # 3개 레코드를 확인하려면?
- `write.csv(iris,file="newiris.csv",row.names=FALSE)`
- `newiris<-read.csv("newiris.csv")`
- `head(newiris)`
- `newiris2<-read.table("newiris.csv",sep=";",header=T)`
- `save(newiris2, file="newiris2.RData")`
- `load("newiris2.RData")`



# Part 3

데이터 타입





# 벡터

## 벡터의 정의

- 벡터는 한 개 이상의 원소로 구성된 자료구조로서 R의 자료 객체 중에서 가장 기본이 되는 자료 객체를 의미

속성	설명
length	자료의 개수
mode	자료의 형태
dim	각 차원 벡터의 크기
dimnames	각 차원 리스트의 이름



# 행렬

## 행렬의 정의

- 행렬은 동일한 형태로 구성된 2차원의 데이터 구조
- 행의 차원과 열의 차원을 갖고 있으며 벡터와 마찬가지로 하나의 행렬은 수치형, 문자형, 논리형 중 한 가지 형태의 원소만 갖는 점에 유의

속성	설명
length	자료의 개수
mode	자료의 형태
dim	행과 열의 개수
dimnames	행과 열의 이름



# 배열

## 배열의 정의

- 배열(Array)은 행렬을 2차원 이상으로 확장시킨 객체를 의미
- 2차원 구조로 이루어진 행렬도 일종의 배열이라고 할 수 있으며 일반 적으로는 3차원 이상의 데이터 객체를 배열이라고 함

속성	설명
length	자료의 개수
mode	자료의 형태
dim	각 차원 벡터의 크기
dimnames	각 차원 리스트의 이름



# 리스트

## 리스트의 정의

- 서로 다른 형태(mode)의 데이터로 구성된 객체를 의미
- 행렬과 배열 등이 동일한 형태의 원소로 이루어진 객체인 반면 리스트를 구성하는 성분(component)은 서로 다른 형태의 원소를 가질 수 있고, 길이도 다를 수도 있음

속성	설명
length	자료의 개수
mode	자료의 형태
names	각 구성요소의 이름



# 데이터 프레임

## 데이터 프레임의 정의

- 행렬은 차원으로 표시되며 같은 형태의 객체를 가지는 반면, 데이터 프레임은 각 열들이 서로 다른 형태의 객체를 가질 수 있음

- 데이터 프레임은 형태(mode)가 일반화된 행렬(matrix)
- 데이터 프레임이라는 하나의 객체에 여러 종류의 자료가 들어갈 수 있음
- 데이터 프레임의 각 열은 각각 변수와 대응
- 분석이나 모형 설정에 적합한 자료 객체



# Part 4

R 프로그래밍



# 연산자

## 연산자

- R은 반복문, 조건문 등을 이용하여 다양한 프로그래밍이 가능한 언어
- 산술, 비교, 논리 연산자

연산자와 함수	의미
$+$ , $-$ , $*$ , $/$	사칙 연산
$n \% m$	$n$ 을 $m$ 으로 나눈 나머지
$n \ \%/\% m$	$n$ 을 $m$ 으로 나눈 몫
$n^m$	$n$ 의 $m$ 승
$\exp(n)$	$e$ 의 $n$ 승
$\log(x, \text{base}=\exp(1))$	$\log_{\text{base}}(x)$ . 만약 $\text{base}$ 가 지정되지 않으면 $\log_e(x)$ 를 계산
$\log_2(x)$ , $\log_{10}(x)$	각각 $\log_2(x)$ , $\log_{10}(x)$ 를 계산
$\sin(x)$ , $\cos(x)$ , $\tan(x)$	삼각 함수





# 기본 함수

## 기본 함수

- R base에 기본으로 탑재되어 있는 함수 목록

함수	예
• pi	<pre>&gt; pi [1] 3.141593 &gt; options(digits=20) &gt; pi [1] 3.141592653589793</pre>

함수	예
• sin(x): sin 함수	<pre>&gt; sin(10) [1] -0.54</pre>
• cos(x): cosine 함수	<pre>&gt; cos(10) [1] -0.84</pre>
• tan(x): tangent 함수	<pre>&gt; tan(10) [1] 0.65</pre>
• asin(x) : arcsin 함수	<pre>&gt; asin(1) [1] 1.6</pre>
• acos(x): arc cosine 함수	<pre>&gt; acos(0) [1] 1.6</pre>
• atan(x): arc tangent 함수	<pre>&gt; atan(0.6) [1] 0.54</pre>

함수	예	
• log(x) : 자연로그 함수	예1) <pre>&gt; log(2) [1] 0.7</pre>	예2) <pre>&gt; x&lt;-3 &gt; y&lt;-4 &gt; log(x+y) [1] 1.9</pre>
• log10(x): 상용 로그 함수	예1) <pre>&gt; log10(10) [1] 1</pre>	예2) <pre>&gt; x&lt;-3 &gt; y&lt;-14 &gt; log(x+y) [1] 1.2</pre>
• exp(x): 지수 로그 함수	<pre>&gt; exp(10) [1] 22026</pre>	
• sqrt(x) : 루트함수	<pre>&gt; sqrt(8) [1] 2.8</pre>	



# 기본 함수

## 기본 함수

- R base에 기본으로 탑재되어 있는 함수 목록

함수	예
• min(x) : 벡터에서 최소값	<pre>&gt; x&lt;-c(1,2,-3,4) &gt; min(x) [1] -3</pre>
• max(x): 벡터에서 최대값	<pre>&gt; x&lt;-c(1,2,-3,4) &gt; max(x) [1] 4</pre>
• min(x1, x2,...) : 전체 벡터 원소 중에서 최소값	<pre>&gt; x1&lt;-c(1,2,-3,4) &gt; x2&lt;-c(2,4,-6,7) &gt; min(x1,x2) [1] -6</pre>
• range(x): 벡터의 범위 -> c(min(x), max(x))	<pre>&gt; x&lt;-c(1,2,-3,4) &gt; range(x) [1] -3 4 &gt; c(min(x), man(x)) [1] -3 4</pre>

함수	예
• pmin(x1,x2) : 두 벡터의 상응하는 원소들 중 작은 값	<pre>&gt; x1&lt;-c(1,2,-3,4) &gt; x2&lt;-c(2,4,-6,7) &gt; pmin(x1,x2) [1] 1 2 -6 4</pre>
• pmax(x1,x2) : 두 벡터의 상응하는 원소들 중 큰 값	<pre>&gt; x1&lt;-c(1,2,-3,4) &gt; x2&lt;-c(2,4,-6,7) &gt; pmin(x1,x2) [1] 2 4 -3 7</pre>



# 기본 함수

## 기본 함수

- R base에 기본으로 탑재되어 있는 함수 목록


함수	예	함수	예
• mean(x1): 평균	<pre>&gt; x1&lt;-c(1,2,3,4,5,6) &gt; mean(x1) [1] 3.5</pre>	• quantile(x,p): (100*p)%에 해당하는 값	<pre>&gt; x1&lt;-c(1,2,3,4,5,6,7,8,9,10) &gt; quantile(x1, 0.5) [1] 50% [1] 5.5</pre>
• sd(x1): 표준 편차	<pre>&gt; x1&lt;-c(1,2,3,4,5,6) &gt; sd(x1) [1] 1.9</pre>	• cor(x,y) : 상관 계수	<pre>&gt; x&lt;-c(1,2,3,4,5,6,7,8,9,10) &gt; y&lt;-c(10,9,8,7,6,5,4,3,2,5) &gt; cor(x,y) [1] -0.91</pre>
• var(x1): 분산	<pre>&gt; x1&lt;-c(1,3,6,9,12,3,2) &gt; var(x1) [1] 16</pre>		
• median(x1): 중앙값(중위수)	<pre>&gt; x1&lt;-c(1,3,6,9,12,3,2) &gt; median(x1) [1] 3</pre>		



# 조건문

## 조건문

- 조건문이라는 것은 특정한 조건을 만족했을 경우에만 프로그램 코드를 수행하는 제어 구문을 의미
- 조건문에는 항상 논리 연산이 수반되며 조건문의 구체적인 표현 식은 조건의 개수, 조건문의 위치, 조건에 따른 명령수행 방식 등 에 따라 구분

문법	의미
<pre>if (cond) {     cond가 참일 때 실행할 문 } else {     cond가 거짓일 때 실행할 }</pre> 	조건 cond가 참, 거짓인 경우에 따라 {} 블록을 실행한다. 필요한 경우 else 블록을 지정할 수 있다.



# 반복문

## 반복문

- 조건문이라는 것은 특정한 조건을 만족했을 경우에만 프로그램 코드를 수행하는 제어 구문을 의미
- 조건문에는 항상 논리 연산이 수반되며 조건문의 구체적인 표현 식은 조건의 개수, 조건문의 위치, 조건에 따른 명령수행 방식 등에 따라 구분

문법	의미
<pre>for (i in data) {   i를 사용한 문장 }</pre>	data에 들어 있는 각각의 값을 변수 i에 할당하면서 각각에 대해 블록 안의 문장을 수행한다.
<pre>while (cond) {   조건이 참일 때 수행할 문장 }</pre> <div></div>	조건 cond가 참일 때 블록 안의 문장을 수행한다.
<pre>repeat {   반복해서 수행할 문장 }</pre>	블록 안의 문장을 반복해서 수행한다. repeat은 다른 언어의 do-while에 해당한다.



# 함수

## 함수 정의

- 함수(function)란 특정한 작업을 독립적으로 수행하는 프로그램 코드의 집합체
- R의 내장 함수에 사용자가 원하는 특정한 기능이 구현되어 있지 않다면 사용자 스스로 직접 함수를 생성하여 원하는 기능을 수행할 수 있음

```
function_name <- function(인자, 인자, ...) {  
  함수 본문  
  return(반환 값) # 반환 값이 없다면 생략  
}
```



# Part 5

## 데이터 전처리 I



# 데이터 전처리

## 데이터 분리 / 병합 / 정렬

- 주어진 데이터를 조건에 따라 분리 : `split()`, `subset()`,
- 주어진 데이터를 조건에 따라 병합 : `merge()`
- 주어진 데이터를 직접 정렬해주는 함수 : `sort()`
- 데이터를 정렬했을 때의 순서를 반환 : `order()`

함수	특징
<code>split()</code>	주어진 조건에 따라 데이터를 분리한다.
<code>subset()</code>	주어진 조건을 만족하는 데이터를 선택한다.
<code>merge()</code>	데이터를 공통된 값에 기준해 병합한다.



A grayscale photograph of a workspace. In the background, a laptop is partially visible. In the foreground, a silver pen lies diagonally across an open, lined notebook. To the right of the notebook, a smartphone is lying flat. A solid teal vertical bar is positioned on the far left side of the image. The text 'Part 6' is written in a large, teal, sans-serif font in the upper left area, and the word 'Summary' is written in a smaller, teal, sans-serif font directly below it.

# Part 6

## Summary