

2022년 한국생명공학연구원 연구데이터
분석과정 R

합성생물학전문연구단 김하성

2022-05-19

Contents

1	Introduction	5
1.1	강의 개요	5
1.2	강의 계획	5
1.3	참고 자료	5
2	R/Rstudio basics	7
2.1	What is R / Rstudio	7
2.2	R / Rstudio Installation	9
2.3	Rstudio interface	10
2.4	Start a project	12
2.5	Getting help	13
2.6	R packages and Dataset	15
3	Rmarkdown	17
3.1	R markdown의 기본 작동 원리	18
3.2	코드 입력	20
3.3	Markdown 문법	22
3.4	YAML 헤더	24
3.5	Output format	24
4	R programming	25
4.1	Console calculator	25
4.2	What is a programming language	27
4.3	Data and variables	28
4.4	Object (Data structure)	30
4.5	A script in R	38
4.6	Functions	39
4.7	Flow control	43
4.8	ifelse statements	45
4.9	for, while, repeat	46
4.10	Avoiding Loops	47
4.11	Object Oriented Programming (Advanced)	47

Chapter 1

Introduction

1.1 강의 개요

- 목표: 생물 데이터 분석을 위한 R 사용법과 (Rstudio, Tidyverse, Bioconductor 포함) 프로그래밍 기술을 습득함
- 장소: 코빅 3층 전산교육장(1304호)
- 강사: 한국생명공학연구원 합성생물학전문연구단 김하성
- 연락처: 042-860-4372, haseong@kribb.re.kr
- 강의자료: <https://greendaygh.github.io/kribb2022R/>

1.2 강의 계획

1. R 사용법 및 데이터 분석 기초 5.19(목), 5.26(목)
2. R/Tidyverse 데이터 분석 중급 6.9(목), 6.16(목)
3. R/Tidyverse 활용 데이터 가시화 7.7(목), 7.14(목)
4. R/Bioconductor 활용한 바이오데이터 분석 기초 8.4(목), 8.11(목)
5. R/Bioconductor 활용한 NGS 데이터 분석 기초 9.1(목), 9.15(목)
6. R/Bioconductor 활용한 NGS 데이터 분석 및 Workflow 10.6(목), 10.13(목)

1.3 참고 자료

- R 홈페이지
- Rstudio 홈페이지
- Bioconductor
- R 기본 문서들
- R ebooks
- Cheat Sheets

- RStudio Webinars
- Shiny
- Hadley github
- R for Data Science
- Using R for Introductory Statistics by John Verzani
 - Free version of 1st Edition
 - Second edition
- Bioinformatics Data Skills by Vince Buffalo
- Introductory Statistics with R by Dalgaard
- 일반통계학 (영지문화사, 김우철 외)

Chapter 2

R/Rstudio basics

2.1 What is R / Rstudio



R은 통계나 생물통계, 유전학을 연구하는 사람들 사이에서 널리 사용되는 오픈소스 프로그래밍 언어입니다. Bell Lab에서 개발한 S 언어에서 유래했으며 많은 라이브러리 (다른 사람들이 만들어 놓은 코드)가 있어서 쉽게 가져다 사용할 수 있습니다. R은 복잡한 수식이나 통계 알고리즘을 간단히 구현하고 사용할 수 있으며 C, C++, Python 등 다른 언어들과의 병행 사용도 가능합니다. R은 IEEE에서 조사하는 Top programming languages에서 2018년 7위, 2019년 5위, 2020년 6위, 2021년 7위로 꾸준히 높은 사용자를 확보하며 빅데이터, AI 시대의 주요한 프로그래밍 언어로 사용되고 있습니다.

R은 데이터를 통계분석에 널리 사용되는데 이는 데이터를 눈으로 확인하기 위한 visualization이나 벡터 연산 등의 강력한 기능 때문에 점점 더 많은 사람들이 사용하고 있습니다. 기존에는 속도나 확장성이 다른 언어들에 비해 단점으로 지적되었으나 R 언어의 지속적인 개발과 업데이트로 이러한 단점들이 빠르게 보완되고 있습니다. R 사용을 위해서는 R 언어의 코어 프로그램을 먼저 설치하고 그 다음 R 언어용














Rank	Language	Type	Score
1	Python [▼]	  	100.0
2	Java [▼]	  	95.4
3	C [▼]	  	94.7
4	C++ [▼]	  	92.4
5	JavaScript [▼]		88.1
6	C# [▼]	   	82.4
7	R [▼]		81.7
8	Go [▼]	 	77.7
9	HTML [▼]		75.4
10	Swift [▼]	 	70.4

Figure 2.1: <https://spectrum.ieee.org/top-programming-languages/>

IDE(Integrated Development Environment)인 RStudio 설치가 필요합니다.



Rstudio는 R 언어를 위한 오픈소스 기반 통합개발환경(IDE)으로 R 프로그래밍을 위한 편리한 기능들을 제공해 줍니다. R언어가 주목을 받고 두터운 사용자 층을 확보할 수 있게된 핵심 동력이 Rstudio 입니다. 자체적으로 최고수준의 오픈소스 개발팀이 있으며 tidyverse, ‘shiny’ 등의 데이터 분석 관련 주요 패키지를 개발하였고 정기적으로 conference 개최를 하면서 기술 보급의 핵심 역할을 하고 있습니다.

Products	About RStudio		Additional Websites	
OPEN SOURCE	LEARNING	SUPPORT	ANALYSE & EXPLORE	CONNECT & INTEGRATE
RStudio Desktop	Education	Frequently Asked Questions	Tidyverse	Professional Drivers
RStudio Server	Videos & Webinars	RStudio Support	ggplot2	Launcher Plugin SDK
Shiny Server	Cheatsheets	RStudio Community	dplyr	Databases
R Packages	rstudio::conf	Certified Partners	tidyr	Environments
		Product Security	purrr	Sparklyr
HOSTED SERVICES	ABOUT US	RStudio Documentation	COMMUNICATE & INTERACT	Plumber
RStudio Academy	About the Company	Contact Us	Shiny	Reticulate
RStudio Cloud	What Makes Us Different	RStudio Legal Terms	rmarkdown	Ursa Labs
RStudio Public Package Manager	Analyst Reports	Email Subscription Management	flexdashboard	MODEL & PREDICT
shinyapps.io	RStudio Swag			Tensorflow
PROFESSIONAL	Careers			Tidymodels
RStudio Team				Spark MLlib
RStudio Workbench				
RStudio Connect				
RStudio Package Manager				

Figure 2.2: <https://www.rstudio.com/>

2.2 R / Rstudio Installation

2.2.1 R 설치

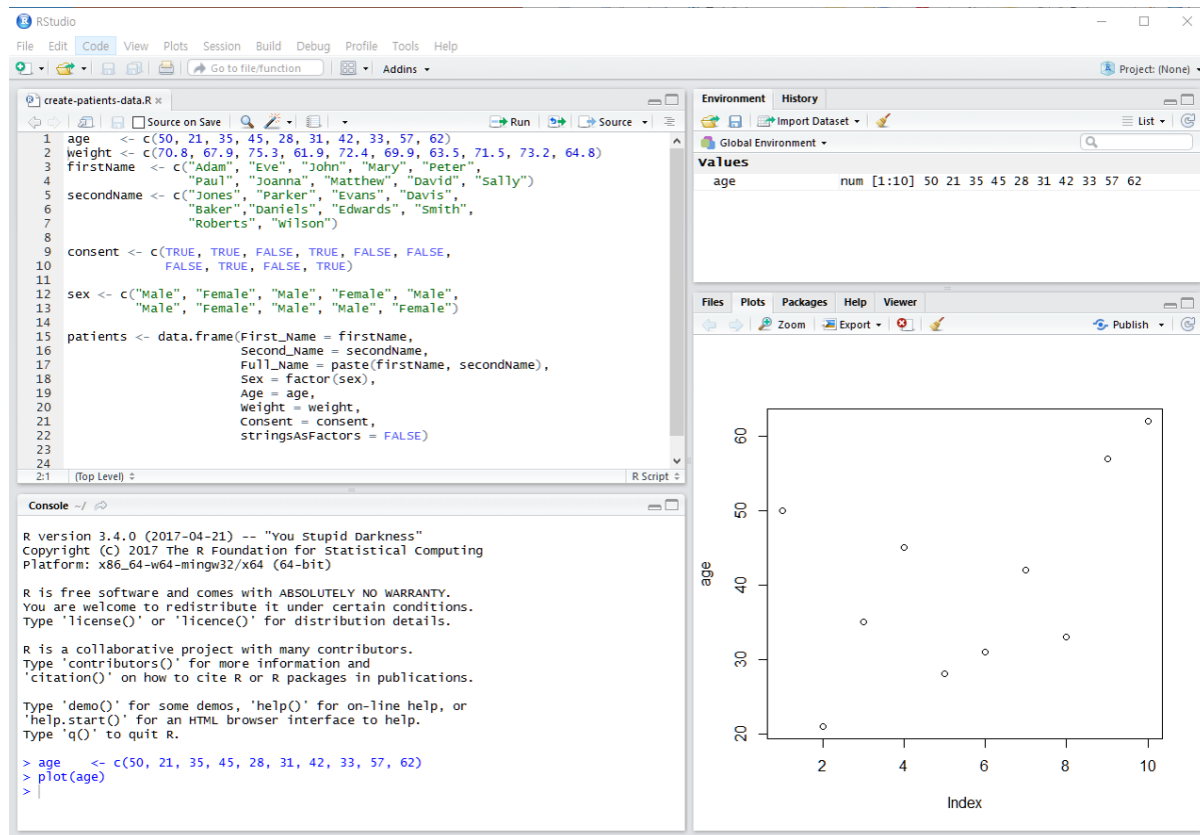
- R 사이트에 접속 후 (<https://www.r-project.org/>) 좌측 메뉴 상단에 위치한 CRAN 클릭.
- 미러 사이트 목록에서 Korea의 아무 사이트나 들어감
- Download R for Windows를 클릭 후 base 링크 들어가서
- Download R x.x.x for Windows 링크 클릭으로 실행 프로그램 다운로드

- 로컬 컴퓨터에 Download 된 R-x.x.x-win.exe 를 실행 (2022.5 현재 R 버전은 4.2.0).
- 설치 프로그램의 지시에 따라 R 언어 소프트웨어 설치를 완료

2.2.2 Rstudio 설치

- 사이트에 접속 (<https://www.rstudio.com/>), 상단의 Products > RStudio 클릭
- RStudio Desktop 선택
- Download RStudio Desktop 클릭
- RStudio Desktop Free 버전의 Download를 선택하고
- Download RStudio for Windows 클릭, 다운로드
- 로컬 컴퓨터에 다운로드된 RStudio-x.x.x.exe 실행 (2022.5 현재 RStudio Desktop 2022.02.2+485)
- 설치 가이드에 따라 설치 완료

2.3 Rstudio interface



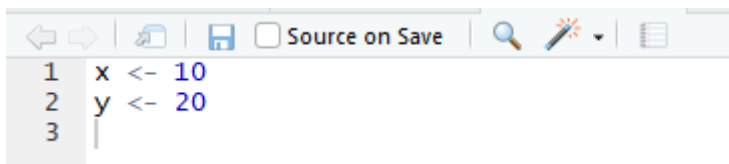
- 기본 화면에서 좌측 상단의 공간은 코드편집창, 좌측 하단은 콘솔창
- 각 위치를 기호에 따라서 바꿀 수 있음 (View -> Pane)

2.3.1 Keyboard shortcuts

- 참고사이트
 - <https://support.rstudio.com/hc/en-us/articles/200711853-Keyboard-Shortcuts>
 - Tools -> Keyboard shortcut Quick Reference (Alt + Shift + K)
- 코드편집창 이동 (Ctrl + 1) 콘솔창 이동 (Ctrl + 2)
- 한 줄 실행 (Ctrl + Enter)
- 저장 (Ctrl + S)
- 주석처리 (Ctrl + Shift + C)
 - 또는 #으로 시작하는 라인
- 탭 이동 (Ctrl + F11, Ctrl + F12)
- 코드편집창 확대 (Shift + Ctrl + 1) 콘솔창 확대 (Shift + Ctrl + 2)
- 컬럼 편집 (Alt +)
- 자동 완성 기능 (Tab completion) in RStudio

2.3.2 Exercise

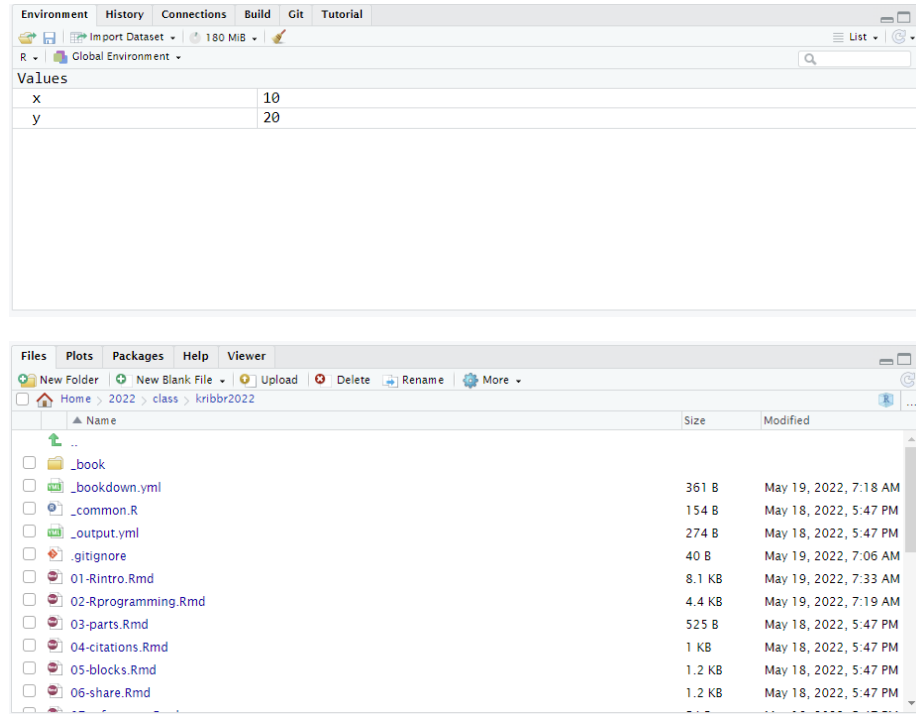
1. 코드편집창에서 다음을 입력/실행하고 단축키를 사용하여 주석을 넣으시오



- 단축키 Ctrl + enter로 코드 실행
- 단축키 Ctrl + 2로 커서 콘솔창으로 이동
- x값 x+y값 확인
- 단축키 Ctrl + 1로 코드편집창 이동
- 단축키 Ctrl + Shift + C 사용

```
# x <- 10
# y <- 20
```

2.3.3 Environment and Files



2.4 Start a project

프로젝트를 만들어서 사용할 경우 파일이나 디렉토리, 내용 등을 쉽게 구분하여 사용 가능합니다. 아래와 같이 임의의 디렉토리에 kribbR 이라는 디렉토리를 생성하고 lecture1 프로젝트를 만듭니다.

File > New Project > New Directory > New Project > “kribbR” > Create Project

시작할 때는 해당 디렉토리의 xxx.Rproj 파일을 클릭합니다. Rstudio 오른쪽 상단 프로젝트 선택을 통해서 빠르게 다른 프로젝트의 작업공간으로 이동할 수 있습니다.

2.4.1 Hello world

File > New File > R markdown > OK

```
mystring <- "Hello \n world!"
cat(mystring)
print(mystring)
```

2.5 Getting help

R은 방대한 양의 도움말 데이터를 제공하며 다음과 같은 명령어로 특정 함수의 도움말과 예제를 찾아볼 수 있습니다. ? 명령을 사용하면 되며 구글이나 웹에서도 도움을 얻을 수 있습니다.

```
help("mean")
?mean
example("mean")
help.search("mean")
??mean
help(package="MASS")
```

또한 <https://www.rstudio.com/resources/cheatsheets/>에서는 다양한 R언어의 기능을 한 눈에 알아볼 수 있게 만든 cheatsheet 형태의 문서를 참고할 수 있습니다.

Base R Cheat Sheet

Getting Help

Accessing the help files

?mean
Get help of a particular function.
help.search('weighted mean')
Search the help files for a word or phrase.
help(package = 'dplyr')
Find help for a package.

More about an object

str(iris)
Get a summary of an object's structure.
class(iris)
Find the class an object belongs to.

Using Packages

install.packages('dplyr')
Download and install a package from CRAN.

library(dplyr)
Load the package into the session, making all its functions available to use.

dplyr::select
Use a particular function from a package.

data(iris)
Load a built-in dataset into the environment.

Working Directory

getwd()
Find the current working directory (where inputs are found and outputs are sent).

setwd('C://file/path')
Change the current working directory.

Use projects in RStudio to set the working directory to the folder you are working in.

Vectors			Programming					
Creating Vectors								
c(2, 4, 6)	2 4 6	Join elements into a vector	For Loop					
2:6	2 3 4 5 6	An integer sequence						
seq(2, 3, by=0.5)	2.0 2.5 3.0	A complex sequence						
rep(1:2, times=3)	1 2 1 2 1 2	Repeat a vector						
rep(1:2, each=3)	1 1 1 2 2 2	Repeat elements of a vector	While Loop					
Vector Functions								
sort(x) Return x sorted.	rev(x) Return x reversed.							
table(x) See counts of values.	unique(x) See unique values.							
Selecting Vector Elements								
By Position								
x[4]		The fourth element.	If Statements					
x[-4]		All but the fourth.						
x[2:4]		Elements two to four.						
x[-(2:4)]		All elements except two to four.						
x[-(2:4)]		All elements except two to four.	Functions					
x[c(1, 5)]		Elements one and five.						
By Value								
x[x == 10]		Elements which are equal to 10.						
x[x < 0]		All elements less than zero.	Reading and Writing Data					
x[x %in% c(1, 2, 5)]		Elements in the set 1, 2, 5.						
Named Vectors								
x['apple']		Element with name 'apple'.						
Also see the readr package.								
Input		Output		Description				
df <- read.table('file.txt')		write.table(df, 'file.txt')		Read and write a delimited text file.				
df <- read.csv('file.csv')		write.csv(df, 'file.csv')		Read and write a comma separated value file. This is a special case of read.table/write.table.				
load('file.Rdata')		save(df, file = 'file.Rdata')		Read and write an R data file, a file type special for R.				
Conditions								
a == b	Are equal	a > b	Greater than	a >= b	Greater than or equal to			
a != b	Not equal	a < b	Less than	a <= b	Less than or equal to			
				is.na(a)	Is missing			
				is.null(a)	Is null			

Types

Converting between common data types in R. Can always go from a higher value in the table to a lower value.

<code>as.logical</code>	TRUE, FALSE, TRUE	Boolean values (TRUE or FALSE).
<code>as.numeric</code>	1, 0, 1	Integers or floating point numbers.
<code>as.character</code>	'1', '0', '1'	Character strings. Generally preferred to factors.
<code>as.factor</code>	'1', '0', '1', levels: '1', '0'	Character strings with preset levels. Needed for some statistical models.

Maths Functions

<code>log(x)</code>	Natural log.	<code>sum(x)</code>	Sum.
<code>exp(x)</code>	Exponential.	<code>mean(x)</code>	Mean.
<code>max(x)</code>	Largest element.	<code>median(x)</code>	Median.
<code>min(x)</code>	Smallest element.	<code>quantile(x)</code>	Percentage quantiles.
<code>round(x, n)</code>	Round to n decimal places.	<code>rank(x)</code>	Rank of elements.
<code>signif(x, n)</code>	Round to n significant figures.	<code>var(x)</code>	The variance.
<code>cor(x, y)</code>	Correlation.	<code>sd(x)</code>	The standard deviation.

Variable Assignment

```
> a <- 'apple'
> a
[1] 'apple'
```

The Environment




<code>ls()</code>	List all variables in the environment.
<code>rm(x)</code>	Remove x from the environment.
<code>rm(list = ls())</code>	Remove all variables from the environment.

You can use the environment panel in RStudio to browse variables in your environment.

Matrices

```
m <- matrix(x, nrow = 3, ncol = 3)
```

Create a matrix from x.

	<code>m[2,]</code>	- Select a row
	<code>m[, 1]</code>	- Select a column
	<code>m[2, 3]</code>	- Select an element

`t(m)`
Transpose
`m %*% m`
Matrix Multiplication
`solve(m, x)`
Find x in: $m \cdot x$

Lists

```
l <- list(x = 1:5, y = c('a', 'b'))
```

A list is a collection of elements which can be of different types.

<code>l[[2]]</code>	<code>l[[1]]</code>	<code>l\$x</code>	<code>l['y']</code>
Second element of l.	New list with only the first element.	Element named x.	New list with only element named y.

Also see the **dplyr** package.



Data Frames

```
df <- data.frame(x = 1:3, y = c('a', 'b', 'c'))
```



A special case of a list where all elements are the same length.

x	y
1	a
2	b
3	c

List subsetting

<code>df\$x</code>	<code>df[[2]]</code>
	
Understanding a data frame	
<code>View(df)</code>	See the full data frame.
<code>head(df)</code>	See the first 6 rows.

Matrix subsetting

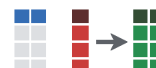
<code>df[, 2]</code>	
<code>df[2,]</code>	
<code>df[2, 2]</code>	

`nrow(df)`
Number of rows.

`ncol(df)`
Number of columns.

`dim(df)`
Number of columns and rows.

cbind - Bind columns



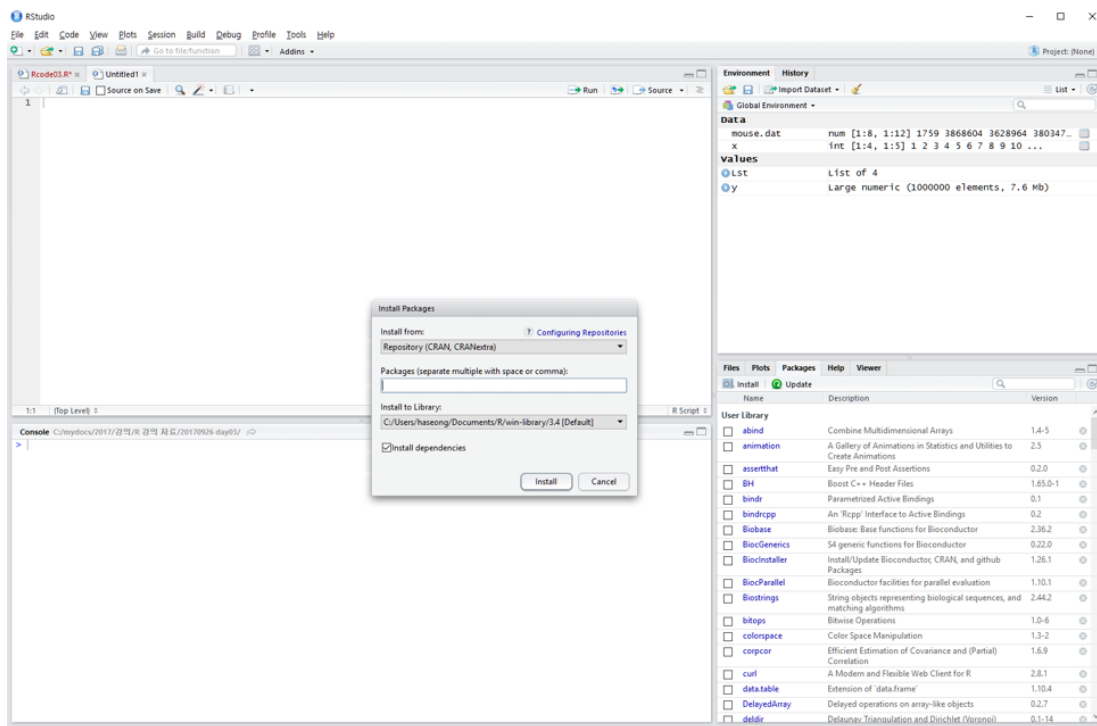
rbind - Bind rows



2.6 R packages and Dataset

R 패키지는 함수와 데이터셋의 묶음으로 다른 사람들이 만들어 놓은 코드나 기능을 가져와서 사용함으로써 코드 작성의 수고로움을 줄이고 편리하고 검증된 함수(기능)를 빠르게 도입하여 사용할 수 있다는 장점이 있습니다. 예를 들어 `sd()` 함수는 `stats` package에서 제공하는 함수로써 표준편차 계산을 위한 별도의 함수를 만들어서 사용할 필요가 없이 바로 (`stats` 패키지는 R 기본 패키지로) 별도 설치 없이 바로 사용 가능합니다.

이러한 패키지는 인터넷의 repository에서 구할 수 있으며 대표적인 repository는 The Comprehensive R Archive Network (CRAN) (<http://cran.r-project.org/web/views/>) 와 생물학자를 위한 Bioconductor (<http://www.bioconductor.org/>) 가 있습니다. 이러한 패키지의 설치에 아래와 같이 RStudio를 이용하거나 콘솔창에서 `install.packages()` 함수를 이용할 수 있습니다.



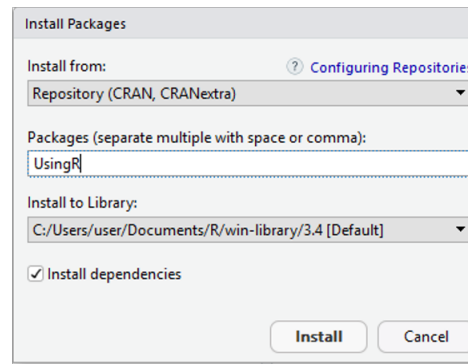
패키지를 설치하고 사용하기 위해서는 `library()` 함수를 사용해서 관련 명령어를 사용하기 전에 미리 loading 해 두어야 합니다. 한 번 로딩으로 작업 세션이 끝날때까지 관련된 함수를 사용할 수 있으나 R 세션이나 RStudio를 재시작 할 경우 다시 로딩해야 사용할 수 있습니다.

`library(UsingR)`

- R 설치 디렉토리
- R 패키지 설치 디렉토리

```
.libPaths()
path.package()
```

Packages → Install



```
> install.packages("UsingR")
Installing package into 'C:/Users/user/Documents/R/win-library/3.4'
(as 'lib' is unspecified)
also installing the dependency 'HistData'

trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.4/HistData_
0.8-4.zip'
Content type 'application/zip' length 359785 bytes (351 KB)
downloaded 351 KB

trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.4/UsingR_2.
0-6.zip'
Content type 'application/zip' length 2081603 bytes (2.0 MB)
downloaded 2.0 MB

package 'HistData' successfully unpacked and MD5 sums checked
package 'UsingR' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
C:\Users\user\AppData\Local\Temp\RtmpwdxEQ7\downloaded_package
> |
```

Files	Plots	Packages	Help	Viewer
Install	Update			
Name	Description			
<input type="checkbox"/> UsingR	Data Sets, Etc. for the Text "Using R for Introductory Statistics", Second Edition			
<input type="checkbox"/> ggplot2	Create Elegant Data Visualisations Using the Grammar of Graphics			
<input type="checkbox"/> munsell	Utilities for Using Munsell Colours			
<input type="checkbox"/> rsbml	R support for SBML, using libsbml			
<input type="checkbox"/> stats4	Statistical Functions using S4 Classes			

```
> library("UsingR", lib.loc="~/R/win-library/3.4")
필요한 패키지를 로드합니다: MASS
필요한 패키지를 로드합니다: HistData
필요한 패키지를 로드합니다: Hmisc
필요한 패키지를 로드합니다: lattice
필요한 패키지를 로드합니다: survival
필요한 패키지를 로드합니다: Formula
필요한 패키지를 로드합니다: ggplot2
```

다음의 패키지를 부착합니다: 'Hmisc'

The following objects are masked from 'package:base':
format.pval, units

다음의 패키지를 부착합니다: 'UsingR'

일반적으로 패키지 안에 관련된 데이터도 같이 저장되어 있으며 `data()` 함수를 이용해서 패키지 데이터를 사용자 작업공간에 복사해서 사용 가능합니다.

```
head(rivers)
length(rivers)
class(rivers)
data(rivers)
data(package="UsingR")
library(HistData)
head(Cavendish)
str(Cavendish)
head(Cavendish$density2)
```

이 저작물은 크리에이티브 커먼즈 저작자표시-비영리-변경금지 4.0 국제 라이선스에 따라 이용할 수 있습니다.

Chapter 3

Rmarkdown

R markdown은 데이터를 분석하는 코드와 리포트를 동시에 수행할 수 있는 일종의 통합 문서입니다. 워드나 아래한글에서 프로그래밍과 데이터분석을 위한 코드를 작성할 수 있는 경우라고 생각해도 됩니다. Plain-text 기반의 markdown 문법을 사용하며 R markdown으로 작성된 문서는 HTML, PDF, MS word, Beamer, HTML5 slides, books, website 등 다양한 포맷의 출력물로 변환할 수 있습니다.

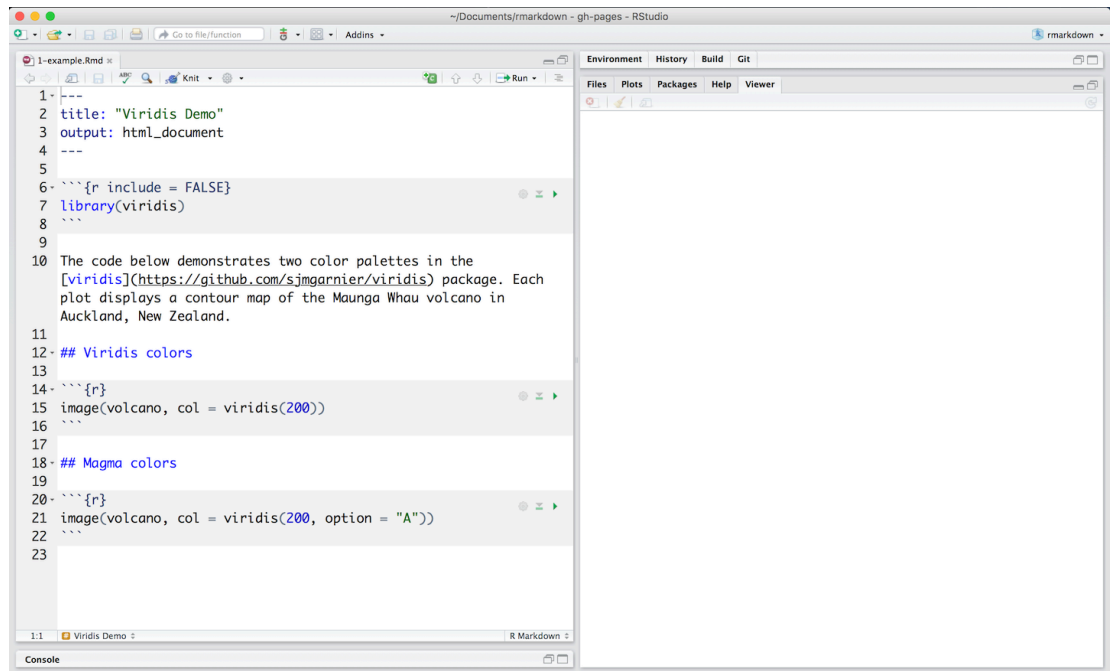


Figure 3.1: Image from rmarkdown.rstudio.com

Rmarkdown 웹사이트에 R markdown 소개 동영상과 R markdown 공식 사이트 메뉴얼 관련 서적 R markdown: The Definitive Guide를 참고하세요. 또한 R markdown을 사용할 때 cheatsheet를 옆에 두고 수시로 보면서 사용하시면 많은 도움이 될 수 있습니다.

3.1 R markdown의 기본 작동 원리

R markdown은 plain text 기반으로 작성되며 Rmd 라는 확장자를 갖는 파일로 저장됩니다. 다음과 같은 텍스트 파일이 Rmd 파일의 전형적인 예 입니다.



위 예제에서 네 가지 다른 종류의 콘텐츠를 볼 수 있습니다. 하나는 --- 으로 둘러싸인 내용으로 YAML 이라고 하며 JSON과 같은 데이터 직렬화를 수행하는 하나의 데이터 저장 포맷입니다. 백틱(`) 으로 둘러싸인 코드청그(Code Chunks)라고 하는 부분에는 R이나 python 등의 다양한 코드(실제 작동하는)를 넣어서 사용합니다. 그리고 ### 으로 표시된 글은 제목 글을 나타내며 나머지는 일반적인 텍스트를 나타냅니다.

이러한 R markdown 파일은 render라는 명령어로 원하는 포맷의 문서로 변환할 수 있습니다. 다음 예의 파일을 pdf 형식으로 rendering 하기 위해서는 YAML에 pdf 임을 명시하고 아래와 같이 render함수를 사용하면 됩니다. 또는 Rstudio 코드 입력창 상단의 Knit 버튼으로 pdf나 html 문서를 생성할 수 있습니다.

```
1 * ---
2 title: "My R markdown example"
3 output:
4   pdf_document: default
5 * ---
6
7 ```{r setup, include=FALSE}
8 library(tidyverse)
9 ```
10
11 * ## R Markdown
12
13 This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML,
14 PDF, and MS Word documents. When you click the **Knit** button a document will be generated
15 that includes both content as well as the output of any embedded R code chunks within the
16 document. You can embed an R code chunk like this:
17
18 ```{r}
19 cars %>% head
20 cars %>%
21   ggplot(aes(x=speed, y=dist)) +
22     geom_point()
23 ```
```

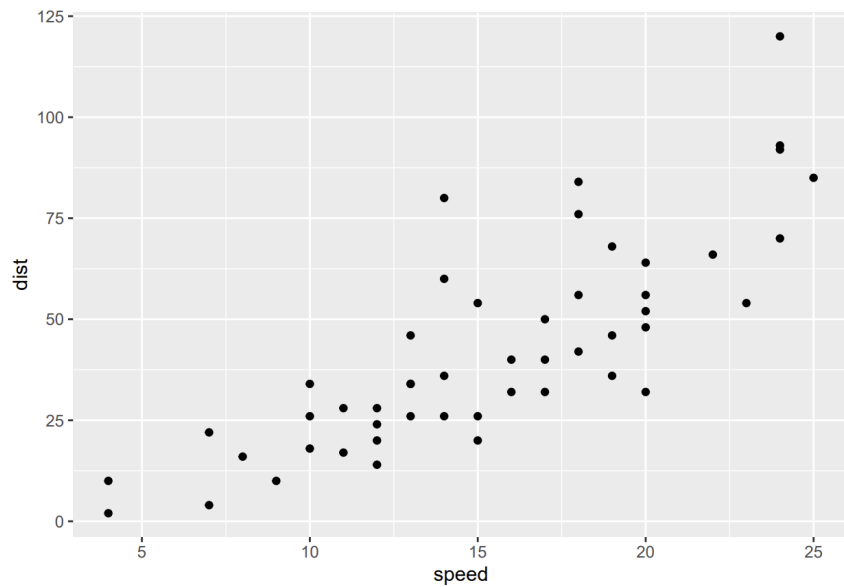
`render("examples/test.Rmd", output_format = "pdf_document")`

My R markdown example

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
cars %>% head
cars %>%
  ggplot(aes(x=speed, y=dist)) +
  geom_point()
```



R markdown의 작동 원리는 Rmd 파일을 만든 후 `render` 함수를 부르면 knitr 소프트웨어가 R 코드를 실행시킨 후 markdown (.md) 파일을 생성합니다. 이 후 .md 파일을 pandoc 이라는 문서변환기가 원하는 문서 형태로 전환해 줍니다.

3.2 코드 입력

R markdown에서 사용하는 코드청크는 CTRL+ALT+I 단축키를 사용해서 넣을 수 있으며 다음과 같은 몇 가지 옵션으로 코드 스니펫들의 실행/숨김 여부를 결정할 수 있습니다.

- `include = FALSE` : 코드는 실행되지만 보고서에 결과와 코드가 보여지지 않음
- `echo = FALSE` : 코드는 실행되고 보고서에 결과가 포함되지만 코드는 보여지지 않음

- `eval = FALSE` : 코드가 실행되지 않지만 보고서에 코드는 보여짐
- `message = FALSE, warning=FALSE, error=FALSE` : 코드에 의해서 발생하는 메시지/경고/에러가 보고서에 보여지지 않음
- `fig.cap = "..."` : 코드로 그려지는 그래프에 캡션을 붙일 수 있음

```

43 ▾ ```{r}
44 # default
45 n <- c(1, 2, 3)
46 mean(n)
47 ▲ ```
48
49 ▾ ```{r, eval=F}
50 # eval=FALSE
51 n <- c(1, 2, 3)
52 mean(n)
53 ▲ ```
54
55 ▾ ```{r, echo=F}
56 # echo=FALSE
57 n <- c(1, 2, 3)
58 mean(n)
59 ▲ ```

```

Figure 3.2: 코드청크 옵션 예시

```

# default
n <- c(1, 2, 3)
mean(n)
#> [1] 2

```

```

# eval=FALSE
n <- c(1, 2, 3)
mean(n)

```

```

#> [1] 2

```

R markdown에서는 `r` 을 사용해서 코드청크가 아닌 곳에 R 코드를 넣을 수 있습니다. 예를 들어 `n` 은 1, 2, 3 값을 가지는 벡터 입니다. 또한 R 언어 외에도 Python, SQL, Bash, Rcpp, Stan, JavaScript, CSS 등의 다양한 프로그래밍 언어에 대해서도 지원합니다. 그런데 이러한 언어들이 사용 가능해지기 위해서는 해당 언어들을 실행해주는 엔진이 있어야 하며 python의 경우 `reticulate` 라는 패키지가 이러한 기능을 담당합니다. 이 패키지를 설치할 경우 `miniconda` 라는 가상환경 및 데이터 분석을 위한 오픈소스 패키지가 자동으로 설치됩니다.

```

x = "hello, python in R"
print(x.split(' '))

```

3.3 Markdown 문법

마크다운은 plain text 기반의 마크업 언어로서 마크업 언어는 태그 등을 이용해서 문서의 데이터 구조를 명시하는데 이러한 태그를 사용하는 방법 체계를 마크업 언어라고 합니다. 가장 대표적으로 html 이 있습니다.

```
<html>
  <head>
    <title> Hello HTML </title>
  </head>
  <body>
    Hello markup world!
  </body>
</html>
```

마크다운도 몇 가지 태그를 이용해서 문서의 구조를 정의하고 있으며 상세한 내용은 Pandoc 마크다운 문서를 참고하시기 바랍니다. 마크다운언어의 철학은 쉽게 읽고 쓸 수 있는 문서입니다. plain text 기반으로 작성되어 쓰기 쉬우며 (아직도 사람들이 메모장 많이 사용하는 이유와 같습니다) 태그가 포함되어 있어도 읽는데 어려움이 없습니다. html 언어와 rmd 파일의 예를 보시면 그 철학을 어렵지 않게 알 수 있습니다.

마크다운에서는 Enter를 한 번 입력해서 줄바꿈이 되지 않습니다.
 또는 문장 마지막에 공백을 두 개 입력하면 되겠습니다.

이 문장은 줄바꿈이 되지 않습니다

이 문장은 줄바꿈이
됩니다

마크다운 태그를 몇 가지 살펴보면 먼저 # 을 붙여서 만드는 header 가 있습니다.

```
# A level-one header
## A level-two header
### A level-three header

# A level-one header {#l1-1}
## A level-two header {#l2-1}
### A level-three header {#l3-1}

# A level-one header {#l1-2}
## A level-two header {#l2-2}
### A level-three header {#l3-2}
```

Block quotations

This is block quote. This paragraph has two lines

This is a block quote. This paragraph has two lines.

This is a block quote.

A block quote within a block quote.

code with five spaces

Italic

Bold

Naver link

이미지를 직접 삽입하고 가운데 정렬합니다.

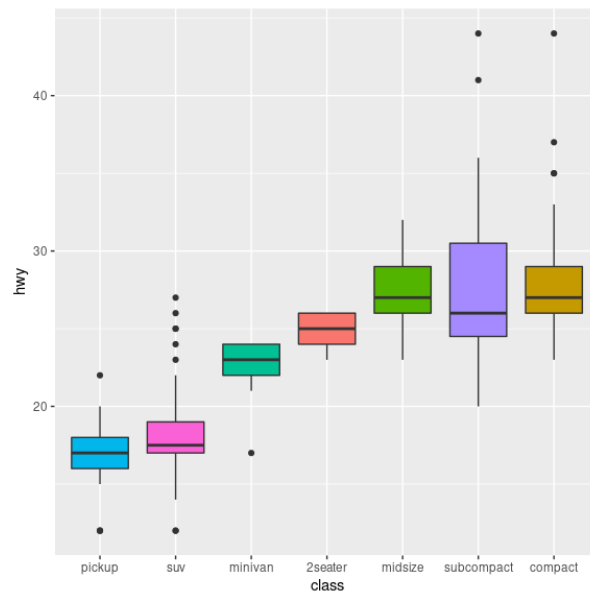


Figure 3.3: 자동차 모델에 따른 고속도로 연비 분포

1. 첫 번째
2. 두 번째
3. 세 번째

- 아이템 1
- 아이템 2
- 아이템 3
 - 아이템 3-1
 - 아이템 3-2

소스코드 그대로 표현하기 위해서는 ~~~ 를 사용합니다.

```
<div>
```

```
</div>
```

3.4 YAML 헤더

R markdown 파일에서 YAML의 가장 중요한 기능은 output 포맷을 지정하는 것이며 title, author, date, 등을 설정할수도 있습니다.

```
---
layout: page
title: "R "
subtitle: "R markdown "
output:
  html_document:
    css: style.css
    includes:
      in_header: header.html
      after_body: footer.html
    theme: default
    toc: yes
    toc_float: true
    highlight: tango
    code_folding: show
    number_sections: TRUE
mainfont: NanumGothic
---
```

3.5 Output format

주요 문서 포맷으로 다음과 같은 몇 가지가 있습니다. 상세한 내용은 Rmarkdown output format을 참고하시기 바랍니다.

- html_document - HTML document w/ Bootstrap CSS
- pdf_document - PDF document (via LaTeX template)
- word_document - Microsoft Word document (docx)
- ioslides_presentation - HTML presentation with ioslides
- beamer_presentation - PDF presentation with LaTeX Beamer
- powerpoint_presentation: PowerPoint presentation

이 저작물은 크리에이티브 커먼즈 저작자표시-비영리-변경금지 4.0 국제 라이선스에 따라 이용할 수 있습니다.

Chapter 4

R programming

4.1 Console calculator

콘솔에서 바로 계산을 수행할 수 있습니다. 참고로 이전에 수행한 명령은 콘솔에 커서가 있는 상태에서 위 아래 화살표를 누르면 볼 수 있고 엔터를 눌러 재사용 할 수 있습니다. ;을 사용하면 두 개의 명령을 동시에 수행할 수 있습니다.

$$2 + 2$$

$$((2 - 1)^2 + (1 - 3)^2)^{1/2}$$

```
2 + 2
((2 - 1)^2 + (1 - 3)^2)^(1/2)
2 + 2; 2 - 2
```

Exercise

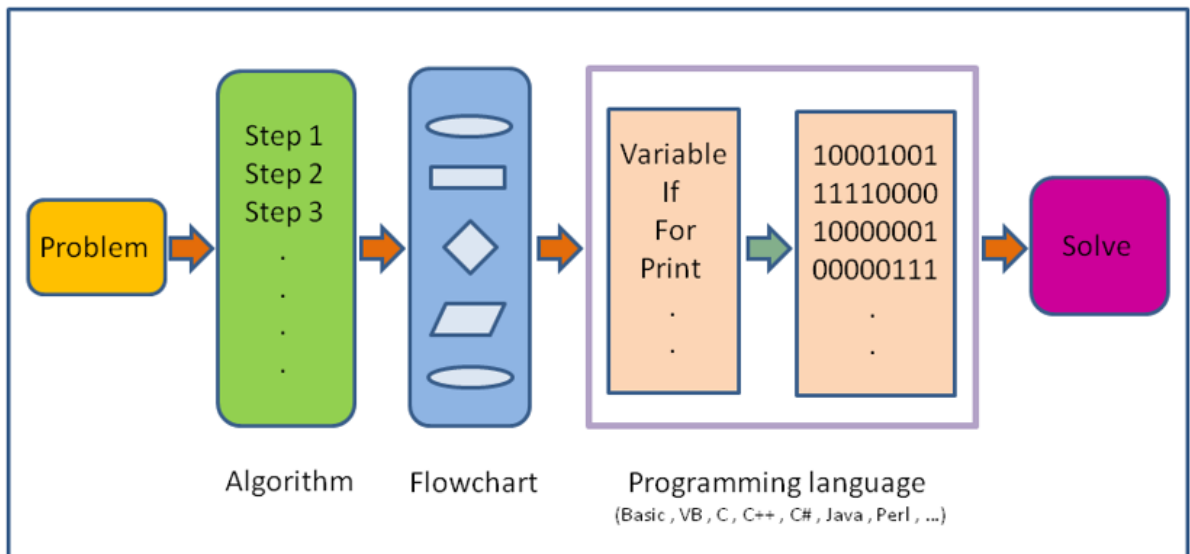
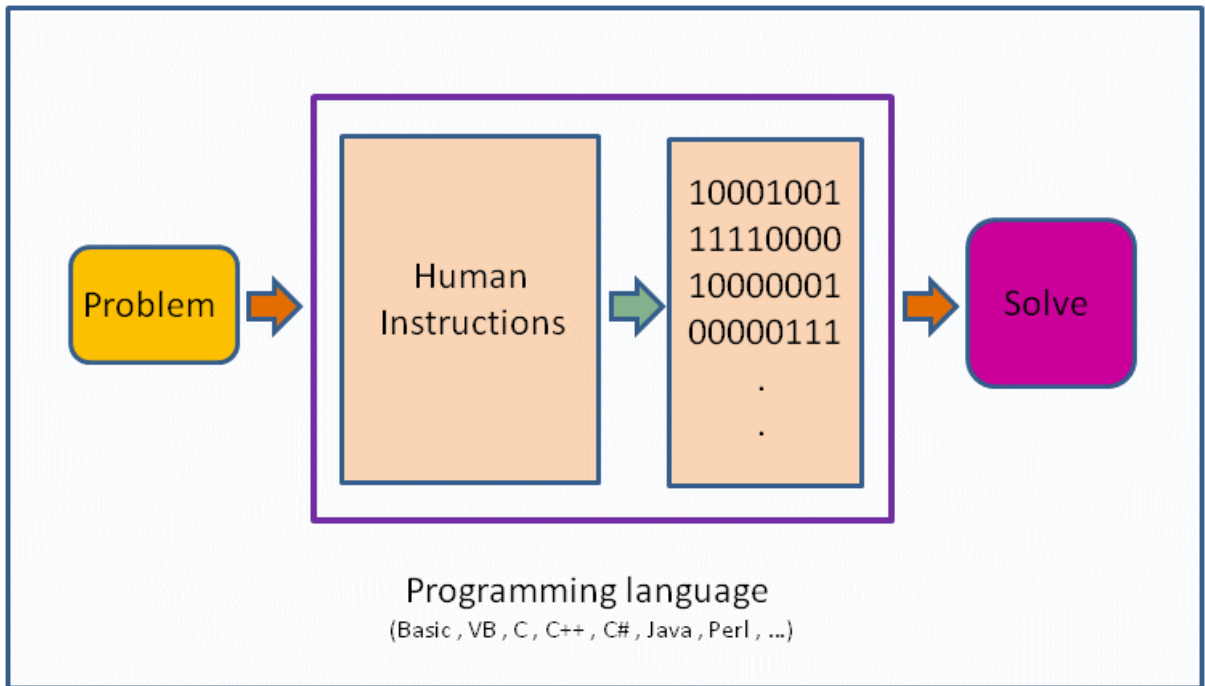
다음 공식들을 계산하는 R 코드를 작성하시오

$$\sqrt{(4 + 3)(2 + 1)}$$

$$2^3 + 3^2$$

$$\frac{0.25 - 0.2}{\sqrt{0.2(1 - 0.2)/100}}$$

4.2 What is a programming language



Problem

두 값 451와 224 중 높은 값을 출력하라

만약 451가 224보다 크면
452 출력
아니면
224 출력



```

If 451>224
  print 452
else
  print 224

```

```

0101000101010
0010101001010
0011110101101
1010110101010

```

R은 programming language로서 다른 프로그래밍 언어와 같이 몇 가지 공통적 개념을 가집니다 (, , , ,)

4.2.1 Terminology

- Session: R 언어 실행 환경
- Console: 명령어 입력하는 창
- Code: R 프로그래밍 변수/제어문 모음
- Object: 변수, 함수 등 프로그래밍에서 사용되는 모든 객체 (Data structure)
 - array: 1D, 2D, 3D, ... 형태 값들의 모임
 - vector: 1차원 형태 값들의 모임 combine function c() EX: c(6, 11, 13, 31, 90, 92)
 - matrix: 2차원 형태 값들의 모임 (같은 타입 값으로 구성)
 - data frame: 2차원 형태 값들의 모임 (다른 타입 값 구성 가능)
 - list: vector, matrix, data.frame 및 list 등 다양한 객체를 원소로 가집
- function: 특정 기능 수행, [함수이름, 입력값 (arguments), 출력값 (return)] 으로 구성
- Data (value): 값 - 자료형 (Data type)
 - Integers
 - doubles/numerics
 - logicals
 - characters
 - factor: 범주형
- Conditionals (조건, 제어):
 - if, ==, & (AND), | (OR) Ex: (2 + 1 == 3) & (2 + 1 == 4)
 - for, while: 반복 수

4.3 Data and variables

4.3.1 Data

일반적으로 데이터의 의미는 사실을 나타내는 수치입니다.

- 맥도너 정보경제학 (1963)
 - 지혜 (wisdom) : 패턴화된 지식
 - 지식 (knowledge) : 가치있는 정보
 - 정보 (information) : 의미있는 데이터
 - 데이터 (data) : 단순한 사실의 나열

```
library(UsingR)
exec.pay
?exec.pay
```

데이터는 속성에 따라서 다음과 같이 분류할 수 있습니다.

- 범주형 - 질적 데이터, 숫자로 나타낼 수 있으나 의미 없음

- 명목형 (Nominal) - 사람 이름
- 순서형 (Ordinal) - 달리기 도착 순서
- 수치형 - 숫자로 나타내며 데이터 속성을 그대로 지님
 - 구간형 (Interval) - 선수1, 선수2 종점통과 시간
 - 비율형 (Ratio) - 출발시간 기준 종점 통과 시간

이름	등수	도착	걸린시간
둘리	1	13:12	1:12
희동	5	14:30	2:30
길동	2	13:30	1:30
철수	4	14:00	2:00
영희	3	13:50	1:50

- Data type in R
 - Numeric (수치형)
 - * Discrete (이산형) data - 카운트, 횟수
 - * Continuous (연속형) data - 키, 몸무게, Cannot be shared
 - * Date and time
 - Factors (범주형)
 - * Categories to group the data
 - * Character data - Identifiers (범주형)

4.3.2 Variables

변수는 데이터를 저장하는 공간으로 이해할 수 있습니다.

- Assignment operator (<- OR =)
 - Valid object name <- value
 - 단축키: Alt + - (the minus sign)
- 내장 변수 Built-in variables

```
x <- 2
y <- x^2 - 2*x + 1
y
x <- "two"
some_data <- 9.8
pi
```

- 변수이름 작명법
 - Characters (letters), numbers, “_”, “.”

- A and a are different symbols
- Names are effectively unlimited in length

```
i_use_snake_case <- 1
otherPeopleUseCamelCase <- 2
some.people.use.periods <- 3
And_aFew.People.RENOUNCEconvention <- 4
```

4.4 Object (Data structure)

변수, 함수 등 프로그래밍에서 사용되는 모든 개체를 말합니다.

4.4.1 vector

vector는 R의 기본 데이터 구조입니다. numeric vector, logical vector, character vector 등 저장되는 값의 타입에 따라 크게 세가지로 나눌 수 있습니다. `class()` 함수를 이용해서 값의 타입을 알아낼 수 있습니다. Combine function인 `c()`를 활용하여 만들어 값을 순차적으로 붙여갈 수 있습니다. 다음과 같은 Univariate (단변량, Single variable)을 표현할 때 사용됩니다.

$$x_1, x_2, \dots, x_n$$

```
x <- c(10.4, 5.6, 3.1, 6.4, 21.7)
class(x)
y <- c("X1", "Y2", "X3", "Y4")
class(y)
z <- c(T, F, F, T)
class(z)
```

4.4.1.1 numeric

numeric 형식의 벡터는 다음과 같은 다양한 편의 함수들을 사용해서 만들 수 있습니다.

```
1:5
seq(1, 5, by=1)
seq(0, 100, by=10)
seq(0, 100, length.out=11)
?seq

rep(5, times=10)
rep(1:3, times=4)
rep(1:3, each=3)
```

Exercise

odds라는 이름의 변수에 1부터 100까지의 홀수만을 저장하시오 (seq() 함수 사용)

4.4.1.2 logical

Logical 벡터는 True 또는 False를 원소로 갖는 벡터입니다. 앞글자가 대문자로 시작하는 것을 기억하시고 T 또는 F와 같이 한 문자로 표현할 수도 있습니다. 특정 조건에 대한 판단 결과를 반환할 경우에도 논리값을 사용합니다. 이 경우 조건을 판단 후 인덱싱 방법으로 해당 값들을 뽑아내기도 합니다.

```
is.na(1)
is.numeric(1)
is.logical(TRUE)

x <- 1:20
x > 13
temp <- x > 13
class(temp)

ages <- c(66, 57, 60, 41, 6, 85, 48, 34, 61, 12)
ages < 30
which(ages < 30)
i <- which(ages < 30)
ages[i]
any(ages < 30)
all(ages < 30)
```

Exercise

1부터 100까지의 수를 n이라는 이름의 변수에 저장하고 이 중 짝수만을 뽑아내서 출력하시오 (which() 함수 사용)

4.4.1.3 character

Character(문자형) 벡터의 경우 문자열을 다루는데 자주 쓰이는 paste() 함수의 사용법을 알아두면 편리합니다. paste() 함수는 서로 다른 문자열을 붙이는데 주로 사용됩니다. 참고로 문자열을 나누는 함수는 strsplit() 입니다. paste()에서 붙이는 문자 사이에 들어가는 문자를 지정하는 파라미터는 sep 이고 strsplit() 함수에서 자르는 기준이 되는 문자는 split 파라미터로 지정해 줍니다 (?split 또는 ?paste 확인).

```
paste("X", "Y", "Z", sep="_")
paste(c("Four", "The"), c("Score", "quick"), c("and", "fox"), sep="_")
paste("X", 1:5, sep="")
paste(c("X", "Y"), 1:10, sep="")
```

```
x <- c("X1", "Y2", "X3", "Y4", "X5")
paste(x[1], x[2])
paste(x[1], x[2], sep="")
paste(x, collapse="_")

strsplit("XYZ", split="")
```

Exercise

m이라는 변수에 “Capital of South Korea is Seoul” 문자열을 저장하고 “Capital of South Korea”를 따로 뽑아내 m2에 저장하시오 (substr() 사용)

4.4.1.4 factor

Factor형은 범주형데이터를 저장하기 위한 object이며 R 언어에서 특별히 만들어져 사용되고 있습니다. factor() 함수를 이용해 생성하며 생성된 객체는 다음과 같이 level이라는 범주를 나타내는 특성값을 가지고 있습니다.

```
x <- c("Red", "Blue", "Yellow", "Green", "Blue", "Green")
y <- factor(x)
y
```

새로운 범주의 데이터를 추가할 경우 다음과 같이 해당되는 level을 먼저 추가하고 값을 저장해야 합니다.

```
levels(y)
y[1] <- "Gold"
y

levels(y) <- c(levels(y), "Gold")
levels(y)
y
y[1] <- "Gold"
y
```

factor는 기본적으로 level에 표시된 순서가 위치 (정렬) 순서입니다. 이를 바꾸기 위해서는 다음과 같이 levels 함수를 이용해서 순서를 바꿀 수 있습니다.

```
#library(UsingR)
str(Cars93)
x <- Cars93$Origin
plot(x)
levels(x) <- c("non-USA", "USA")
levels(x)
plot(x)
```


4.4.1.5 Attribute

vector 들은 다음과 같은 builtin 함수들을 사용해서 해당 변수의 attribute를 알아낼 수 있습니다. attribute에는 원소 이름, 타입, 길이 등 vector형 변수가 가질 수 있는 특성을 말합니다.

```
head(precip)
class(precip)
length(precip)
names(precip)

test_scores <- c(100, 90, 80)
names(test_scores) <- c("Alice", "Bob", "Shirley")
test_scores
```

4.4.1.6 indexing

인덱싱은 vector 데이터의 일부 데이터를 참조할 때 사용하는 방법입니다.

```
x[1]
x[1:3]
i <- 1:3
x[i]
x[c(1,2,4)]
y[3]

head(precip)
precip[1]
precip[2:10]
precip[c(1,3,5)]
precip[-1]
precip["Seattle Tacoma"]
precip[c("Seattle Tacoma", "Portland")]
precip[2] <- 10
```

4.4.1.7 Missing values

특정 값이 “Not available” 이거나 “Missing value” 일 경우 벡터의 해당 원소 자리에 데이터의 이상을 알리기 위해 NA를 사용합니다. 따라서 일반적인 연산에서 NA가 포함되어 있는 경우 데이터의 불완전성을 알리기 위해 연산의 결과는 NA가 됩니다. is.na() 함수는 해당 변수에 NA 값이 있는지를 검사해주는 함수이며 R에는 이 외에도 다음과 같은 특수 값들이 사용되고 있습니다.

- NA: Not available, The value is missing
- NULL: a reserved value
- NaN: Not a number (0/0)
- Inf: (1/0)

```
hip_cost <- c(10500, 45000, 74100, NA, 83500)
sum(hip_cost)
sum(hip_cost, na.rm=TRUE)
?sum
```

4.4.1.8 Useful functions

다음은 벡터형 변수와 같이 쓰이는 유용한 함수들입니다.

```
z <- sample(1:10, 100, T)
head(z)
sort(z)
order(z)
table(z)
p <- z/sum(z)
round(p, digits=1)
digits <- as.character(z)
n <- as.numeric(digits)
d <- as.integer(digits)
```

4.4.2 matrix

매트릭스는 2차원 행렬로 같은 형식의 데이터 값 (numeric, character, logical) 으로서 채워진 행렬을 말합니다. 매트릭스를 만드는 방법은 아래와 같으며 `nrow` 와 `ncol` 파라미터에 행과 열의 수를 넣고 각 셀에 들어갈 값은 가장 앞에 위치한 `data` 파라미터에 넣어 줍니다 (`?matrix`로 파라미터 이름 확인). 매트릭스 인덱싱은 매트릭스 안의 값을 저장하거나 참조할때 (빼올때) 사용하는 방법입니다. 매트릭스 변수이름 바로 뒤에 대괄호를 이용해서 제어를 하며 대괄호 안에 콤마로 구분된 앞쪽은 `row`, 뒷쪽은 `column` 인덱스를 나타냅니다.

```
mymat <- matrix(0, nrow=100, ncol=3) # 1
mymat[,1] <- 1:100 # 2
mymat[,2] <- seq(1,200,2) # 3
mymat[,3] <- seq(2,200,2) # 4
```

매트릭스의 `row`나 `column`에 이름이 주어져 있을 경우 이름을 따옴표(")로 묶은 후 참조가 가능합니다. `row`나 `column`의 이름은 `rownames()` 또는 `colnames()`로 생성하거나 변경할 수 있습니다. `row`나 `column`의 개수는 `nrow()` 또는 `ncol()` 함수를 사용합니다.

```
colnames(mymat)
colnames(mymat) <- c("A", "B", "C")
colnames(mymat)
colnames(mymat)[2] <- "D"
colnames(mymat)
rownames(mymat) <- paste("No", 1:nrow(mymat), sep="")
```

```
rownames(mymat)
```

여러 row나 column을 참조할 경우 아래와 같이 combine 함수를 사용하여 묶어줘야 하며 스칼라값을 (임의의 숫자 하나) 더하거나 뺄 경우 vector / matrix 연산을 기본으로 수행합니다.

```
mymat[c(2,3,4,5),2] # 5
mymat-1 # 6
mysub <- mymat[,2] - mymat[,1] #7
sum(mysub) #8
sum(mysub^2) #8
```

Exercise

- score 라는 변수에 1부터 100까지 중 랜덤하게 선택된 20개의 수로 10 x 2 matrix를 만드시오 (sample() 사용)
- score의 row 이름을 문자형으로 Name1, Name2, ..., Name10으로 지정하시오 (paste() 사용)
- score의 column 이름을 문자형으로 math와 eng로 지정하시오
- 이 matrix의 첫번째 컬럼과 두 번째 컬럼의 수를 각각 더한 후 total_score라는 변수에 저장하시오
- total_score의 오름차순 순서를 나타내는 인덱스 (order()함수 사용)를 o라는 변수에 저장하시오
- score를 o순서로 재배치하고 score_ordered 변수에 저장하시오

4.4.3 data.frame

데이터프레임은 형태는 매트릭스와 같으나 컬럼 하나가 하나의 변수로서 각 변수들이 다른 모드의 값을 저장할 수 있다는 차이가 있습니다. \$ 기호를 이용하여 각 구성 변수를 참조할 수 있습니다. 컬럼 한 줄이 하나의 변수 이므로 새로운 변수도 컬럼 형태로 붙여 넣을 수 있습니다. 즉, 각 row는 샘플을 나타내고 각 column은 변수를 나타내며 각 변수들이 갖는 샘플의 개수 (row의 길이, vector 의 길이)는 같아야 합니다. R 기반의 데이터 분석에서는 가장 선호되는 데이터 타입이라고 볼 수 있습니다.

```
## data.frame
ids <- 1:10
ids
idnames <- paste("Name", ids, sep=" ")
idnames
students <- data.frame(ids, idnames)
students
class(students$ids)
class(students$idnames)
students$idnames
str(students)
```

```
students <- data.frame(ids, idnames, stringsAsFactors = F)
class(students$idnames)
students$idnames
students[1,]
str(students)
```

데이터프레임에서도 변수 이름으로 인덱싱이 가능합니다.

```
## data frame indexing
students$ids
students[,1]
students[, "ids"]
```

Exercise

- math라는 변수에 1부터 100까지 중 랜덤하게 선택된 10개의 수를 넣으시오
- eng라는 변수에 1부터 100까지 중 랜덤하게 선택된 10개의 수를 넣으시오
- students라는 변수에 문자형으로 Name1, Name2, ..., Name10으로 지정하시오 (paste() 사용)
- math와 eng라는 벡터에 저장된 값들의 이름을 students 변수에 저장된 이름으로 지정하시오
- math와 eng 벡터를 갖는 score 라는 data.frame을 만드시오
- math와 eng 변수를 지우시오 (rm()사용)
- score data frame의 math와 eng를 각각 더한 후 total_score라는 변수에 저장 하시오

4.4.4 list

리스트는 변수들의 모임이라는 점에서 데이터프레임과 같으나 구성 변수들의 길이가 모두 같아야 하는 데이터프레임과는 달리 다른 길이의 변수를 모아둘 수 있는 점이 다릅니다. 즉, R언어에서 두 변수를 담을 수 있는 데이터 타입은 list와 data frame 두 종류가 있는데 list 변수 타입은 vector 형태의 여러개의 element를 가질 수 있으며 각 vector 길이가 모두 달라도 됩니다. list의 인덱싱에서 []는 리스트를 반환하고 [[]]는 vector element들을 반환합니다.

Lists

```
l <- list(x = 1:5, y = c('a', 'b'))
```

A list is a collection of elements which can be of different types.

<code>l[[2]]</code>	<code>l[1]</code>	<code>l\$x</code>	<code>l['y']</code>
Second element of l.	New list with only the first element.	Element named x.	New list with only element named y.

```
## list
parent_names <- c("Fred", "Mary")
number_of_children <- 2
child_ages <- c(4, 7, 9)
data.frame(parent_names, number_of_children, child_ages)
lst <- list(parent_names, number_of_children, child_ages)
lst[1]
lst[[1]]
class(lst[1])
class(lst[[1]])
lst[[1]][1]
lst[[1]][c(1,2)]
```

Also see the
dplyr package.

Data Frames

```
df <- data.frame(x = 1:3, y = c('a', 'b', 'c'))
```

A special case of a list where all elements are the same length.

x	y
1	a
2	b
3	c

Matrix subsetting

`df[, 2]`

`df[2,]`

`df[2, 2]`

List subsetting

`df$x`

`df[[2]]`

Understanding a data frame

`View(df)`

See the full data frame.

`head(df)`

See the first 6 rows.

`nrow(df)`

Number of rows.

`ncol(df)`

Number of columns.

`dim(df)`

Number of columns and rows.

cbind - Bind columns.

rbind - Bind rows.

4.5 A script in R

R 프로그래밍을 통해서 사용자가 원하는 기능을 수행하는 방법은 다음과 같이 스크립트를 만들어서 실행하는 것입니다. 일반적으로 R을 이용한 스크립트

명령을 어떻게 실행하는지 알아보겠습니다. 다음 예제는 입력 값들의 평균을 계산해서 출력해 주는 스크립트 명령입니다. R base 패키지에서 기본으로 제공되는 `mean()`이라는 함수가 있지만 사용하지 않고 `sum()`과 `length()` 함수를 사용했습니다.

```
numbers <- c(0.452, 1.474, 0.22, 0.545, 1.205, 3.55)
cat("Input numbers are", numbers, "\n")
numbers_mean <- sum(numbers)/length(numbers)
out <- paste("The average is ", numbers_mean, ".\n", sep="")
cat(out)
```

상황에 따라 다르긴 하지만 보통 위 스크립트를 실행할 때 R 파일을 하나 만들고 `source()`라는 함수를 사용해서 파일 전체를 한번에 읽어들이고 실행을 시킵니다. 위 코드를 `myscript.R`이라는 새로운 R 파일을 하나 만들고 저장 후 다음과 같이 실행할 수 있습니다. 참고로 위 파일은 현재 Working directory와 같은 위치에 저장해야 합니다.

```
source("myscript.R")
```

그러나 위와 같은 식으로 실행할 경우 다음 몇 가지 문제가 있습니다. 하나는 입력 값이 바뀔 때마다 파일을 열어 바뀐 값을 저장해 줄 필요가 있습니다. 결과 값에 대해서 다른 처리를 하고 싶을 경우 또한 파일을 직접 수정해 주어야 합니다. 또한 모든 변수들이 전역변수로 사용되어 코드가 복잡해질 경우 변수간 간섭이 생길 가능성이 높습니다.

4.6 Functions

함수(Function)란 사용자가 원하는 기능을 수행하는 코드의 모음으로서 반복적으로 쉽게 사용할 수 있도록 만들어 놓은 코드입니다. 특정 데이터를 입력으로 받아 원하는 기능을 수행한 후 결과 데이터를 반환하는 구조를 가집니다. 함수는 일반적으로 다음과 같은 포맷으로 구현할 수 있습니다.

```
my_function_name <- function(parameter1, parameter2, ... ){
  ##any statements
  return(object)
}
```

예를 들어 다음과 같은 `my_sine` 함수를 만들 수 있으며 `parameter` (매개변수)는 `x`이고 `y`는 반환값을 저장하는 지역변수입니다.

```
my_sine <- function(x){
  y <- sin(x)
  return(y)
}
```

만들어진 함수는 다음과 같이 사용할 수 있습니다. 만들어진 함수는 처음에 한 번 실행해 주어 실행중인 R session에 등록한 후 사용할 수 있습니다. 여기서 함수로

전달되는 값 `pi`는 argument (전달인자) 라고 합니다. 전달인자는 함수에서 정의된 매개변수의 갯수와 같은 수의 전달인자를 입력해 주어야 합니다.

```
my_sine(pi)
my_sine(90)
sin(90)
```

- Terminology
 - function name: `my_sine`
 - parameter: `x`
 - argument: `pi`
 - return value: `y`

이제 위 스크립트 ('myscript.R') 에서 사용된 코드를 함수로 바꿔봅니다. `numbers` (전달인자)를 받는 매개변수를 `x`로 하고 함수 이름은 `mymean` 이고 평균값 (`numbers_mean`)을 반환하는 함수입니다.

```
numbers <- c(0.452, 1.474, 0.22, 0.545, 1.205, 3.55)

mymean <- function(x){
  cat("Input numbers are", x, "\n")
  numbers_mean <- sum(x)/length(x)
  out <- paste("The average is ", numbers_mean, ".\n", sep="")
  cat(out)
  return(numbers_mean)
}

retval <- mymean(numbers)
cat(retval)
```

`myscript.R`이라는 파일을 열고 작성된 스크립트에 더해서 아래처럼 함수 코드를 만들 경우 `source()` 함수로 함수를 세션으로 읽어오고 바로 사용할 수 있습니다. 위와 같이 함수를 만들 경우 입력 값을 언제든지 바꿔서 사용할 수 있고 반환값에 대한 추가적인 연산도 쉽게 수행 할 수 있습니다.

```
new_values <- c(1:10)
retval <- mymean(new_values)
retval
```

Exercise

1. 변수 `x`에 1, 3, 5, 7, 9를, 변수 `y`에 2, 4, 6, 8, 10을 저장하는 코드를 작성하시오
2. `x`와 `y`를 더한 값을 `z`에 저장하는 코드를 작성하시오
3. `mysum`이라는 이름의 함수를 작성하되 두 변수를 입력으로 받아 더한 후 결과를 반환하는 코드를 작성하시오

4. `mymean`이라는 이름의 함수를 작성하되 두 변수를 입력으로 받아 평균을 구한 후 결과를 반환하는 코드를 작성하시오

Exercise

- 1) `mysd`라는 이름의 (표본)표준편차를 구하는 함수를 `myscript.R` 파일에 구현하시오 (`sd()` 함수 사용하지 않고, 다음 표준편차 공식 이용)

$$\sigma = \sqrt{\frac{\sum (x - \text{mean}(x))^2}{\text{length}(x) - 1}}$$

코드는 아래와 같음

```
mysd <- function(x){
  numbers_sd <- sqrt(sum((x - mymean(x))^2)/(length(x)-1))
  return(numbers_sd)
}
```

- 2) 1부터 100까지의 값을 `x`에 저장하고 `mysd` 함수를 사용해서 표준편차를 구하시오

```
x <- 1:100
mysd(x)
```

- 3) 앞서 작성한 `mymean` 함수와 `mysd` 함수를 같이 사용하여 `x`를 표준화 하고 `z`로 저장하시오. 표준화 공식은 다음과 같음

$$z = \frac{x - \text{mean}(x)}{\text{sd}(x)}$$

```
z <- (x - mymean(x))/mysd(x)
```

- 4) `x`와 `z`를 갖는 `y`라는 이름의 `data.frame`을 생성하시오

4.6.1 local and global variables

다음 코드를 보면 전역변수 `x`, `y`는 지역변수 `x`, `y`와 독립적으로 사용됨을 알 수 있습니다.

```
my_half <- function(x){
  y <- x/z
  cat("local variable x:", x, "\n")
  cat("local variable y:", y, "\n")
  cat("global variable z:", z, "\n")
  return(y)
}
```

```

y <- 100
x <- 20
z <- 30
cat("Global variable x:", x, "\n")
cat("Global variable y:", y, "\n")
cat("Global variable z:", z, "\n")
my_half(5)

my_half <- function(x, z){
  y <- x/z
  cat("local variable x:", x, "\n")
  cat("local variable y:", y, "\n")
  cat("local variable z:", z, "\n")
  return(y)
}

my_half(5, 10)

```

log, sin 등의 함수들은 Built-in function으로 같은 이름의 함수를 만들지 않도록 주의합니다.

```

x <- pi
sin(x)
sqrt(x)
log(x)
log(x, 10)
x <- c(10, 20, 30)
x + x
mean(x)
sum(x)/length(x)

```

4.6.2 Vectorized functions

초기에 R이 다른 프로그래밍 언어에 비해서 경쟁력을 갖는 이유 중 하나가 바로 이 벡터 연산 기능이었습니다. vector 변수에 들어있는 각 원소들에 대해서 특정 함수나 연산을 적용하고 싶을 경우 전통 방식의 C나 Java 등의 언어에서는 원소의 개수만큼 반복문을 돌면서 원하는 작업을 수행 했습니다. 그러나 R의 벡터 연산 기능은 별도의 반복문 없이 vector 안에 있는 원소들에 대한 함수 실행 또는 연산을 수행할 수 있습니다.

```

x <- c(10, 20, 30)
x + x
sqrt(x)
sin(x)
log(x)
x-mean(x)

```

```
length(x)
test_scores <- c(Alice = 87, Bob = 72, James= 99)
names(test_scores)
```

Exercise

다음은 한 다이어트 프로그램의 수행 전 후의 다섯 명의 몸무게이다.

Before	78	72	78	79	105
after	67	65	79	70	93

- 1) 각각을 before 와 after 이름의 변수에 저장 후 몸무게 값의 변화량을 계산하여 diff 라는 변수에 저장하시오
- 2) diff에 저장된 값들의 합, 평균, 표준편차를 구하시오

Exercise

다음 네 학생이 있으며 “John”, “James”, “Sara”, “Lilly” 각 나이는 21, 55, 23, 53 이다. ages 라는 변수를 생성하고 각 나이를 저장한 후 who라는 이름의 함수를 만들어서 50살 이상인 사람의 이름을 출력하는 함수를 만드시오.

- ages라는 변수에 나이 저장, c() 함수 이용, vector 형태 저장
- names() 함수 이용해서 각 ages 벡터의 각 요소에 이름 붙이기
- which() 함수 사용해서 나이가 50보다 큰 인덱스 찾고 해당 인덱스 값들을 idx에 저장
- ages에서 idx에 해당하는 인덱스를 갖는 값을 sel_ages에 저장
- names() 함수를 이용해서 sel_ages의 이름을 sel_names에 저장
- 위 설명을 참고해서 input이라는 파라미터를 갖고 sel_names라는 50살 이상인 사람의 이름을 반환하는 who50이라는 이름의 함수 만들기
- who50 함수의 사용법은 who50(ages) 임

4.7 Flow control

4.7.1 if statements

R에서의 제어문의 사용은 다른 프로그래밍 언어와 거의 유사합니다. 먼저 if 는 다음과 같은 형식으로 사용되며 () 안에 특정 조건 판단을 위한 표현이 들어갑니다.

```
if(condition){
  expr_1
}else{
  expr_2
}
```

특히 `condition`은 하나의 원소에 대한 조건 판단문으로 T 또는 F 값 하나만을 반환하는 문장이어야 합니다. 위 코드는 만약 `condition` 조건이 True 이면 `expr_1`를 실행하고 False이면 `expr_2`를 실행하라는 명령입니다. `condition` 안에서 사용되는 비교 연산자들은 다음과 같습니다.

<code>! x</code>	logical negation, NOT x
<code>x & y</code>	elementwise logical AND
<code>x && y</code>	vector logical AND
<code>x y</code>	elementwise logical OR
<code>x y</code>	vector logical OR
<code>xor(x, y)</code>	elementwise exclusive OR
<code><</code>	Less than, binary
<code>></code>	Greater than, binary
<code>==</code>	Equal to, binary
<code>>=</code>	Greater than or equal to, binary
<code><=</code>	Less than or equal to, binary

```
x <- 2
if(x%%2 == 1){
  cat("Odd")
}else{
  cat("Even")
}

x <- 5
if(x > 0 & x < 4){
  print("Positive number less than four")
}

if(x > 0) print("Positive number")

x <- -5
if(x > 0){
  print("Non-negative number")
} else if(x <= 0 & x > -5){
  print("Negative number greater than -5")
} else {
  print("Negative number less than -5")
}
```

```

}

if(x > 0)
  print("Non-negative number")
else
  print("Negative number")

```

4.8 ifelse statements

if는 하나의 조건만 비교하는데 사용할 수 있습니다. 그러나 변수에는 여러 값이 벡터형식으로 들어가고 벡터연산을 수행할 경우의 결과도 벡터형식으로 나오지만 if문은 이들을 한 번에 처리하기 어렵습니다. ifelse는 이러한 단점을 보완하여 여러 값을 한번에 처리할 수 있습니다.

```
ifelse (condition, True , False )
```

```

x <- c(1:10)
if(x>10){
  cat("Big")
}else{
  cat("Small")
}

ifelse(x>10, "Big", "Small")

```

그러나 출력만 가능하며 조건별로 다른 명령 수행은 불가능하다는 단점이 있습니다.

Exercise

다음은 median (중간값)을 구하는 공식이며 x의 길이가 (n)이 홀수일 경우와 짝수일 경우에 따라서 다른 공식이 사용된다. 다음 공식과 코드를 이용하여 mymedian 이라는 이름의 함수를 만들고 입력 값들의 중간값을 구해서 반환하는 함수를 만드시오. (%% 나머지 연산, if문 사용, 아래 중간값 코드 참고)

$$median(X) = \begin{cases} \frac{1}{2}X[\frac{n}{2}] + \frac{1}{2}X[1 + \frac{n}{2}] & \text{if } n \text{ is even} \\ X[\frac{n+1}{2}] & \text{if } n \text{ is odd} \end{cases}$$

```

sorted_x <- sort(x)
#
retval <- sort_x[n/2]/2 + sort_x[1+(n/2)]/2
#
retval <- sort_x[(n+1)/2]

```

4.9 for, while, repeat

for 문은 반복적으로 특정 코드를 실행하고자 할 때 사용됩니다. 다음과 같은 형식으로 사용할 수 있습니다.

```
for(var in seq){
  expression
}
```

var는 반복을 돌 때마다 바뀌는 변수로 {} 안에서 사용되는 지역 변수입니다. seq는 vector 형식의 변수로 반복을 돌 때마다 순차적으로 var에 저장되는 값들입니다.

```
x <- 1:10
for(i in x){
  cat(i, "\n")
  flush.console()
}

sum_of_i <- 0
for(i in 1:10){
  sum_of_i <- sum_of_i + i
  cat(i, " ", sum_of_i, "\n");flush.console()
}
```

while문도 for문과 같이 반복적으로 특정 코드를 수행하고자 할 때 사용합니다. 사용하는 문법은 다음과 같으며 cond는 True 또는 False 로 반환되는 조건문을 넣고 True 일 경우 계속해서 반복하면서 expressions를 수행하며 이 반복은 cond가 False로 될 때 까지 계속됩니다.

```
while(cond){
  expression
}
```

while문을 사용할 경우 다음과 같이 indicator라 불리우는 변수를 하나 정해서 반복 할 때마다 값이 바뀌도록 해 주어야 합니다. 그렇지 않으면 무한 루프를 돌게 되는 문제가 발생합니다.

```
i <- 10
f <- 1
while(i>1){
  f <- i*f
  i <- i-1
  cat(i, f, "\n")
}
f
factorial(10)
```

repeat 명령은 조건 없이 블록 안에 있는 코드를 무조건 반복하라는 명령입니다. 따라서 블록 중간에 멈추기 위한 코드가 필요하고 이 명령이 break 입니다.

```
repeat{
  expressions
  if(cond) break
}

i <- 10
f <- 1
repeat {
  f <- i*f
  i <- i-1
  cat(i, f, "\n")
  if(i<1) break
}
f
factorial(10)
```

4.10 Avoiding Loops

R에서는 가능하면 loop문을 사용하지 않는 것이 좋습니다. 이는 다른 언어들 보다 반복문이 느리게 수행된다는 이유 때문이기도 합니다. 그러나 R에서는 반복문을 수행하는 것 보다 훨씬 더 빠르게 반복문을 수행 한 것과 같은 결과를 얻을 수 있는 다양한 방법들이 제공되고 있습니다. 차차 그런 기법들에 대한 학습을 진행하도록 하겠습니다.

```
x <- 1:1E7
sum(x)
system.time(sum(x))

st <- proc.time()
total <- 0
for(i in 1:length(x)){
  total <- total + x[i]
}
ed <- proc.time()
ed-st
```

4.11 Object Oriented Programming (Advanced)

OOP는 객체지향 프로그래밍 이라고 합니다. OOP를 이용해서 프로그래밍으로 풀고자 하는 문제를 좀 더 명확하게 개념을 수립하고 복잡한 코드를 명료하게 만들 수 있습니다. 그런데 R에서 OOP는 다른 언어보다는 좀 더 어려운 개념적인 이해가 필요합니다. S3, S4, 그리고 Reference class 가 있으며 S3, S4는 Generic function을 이용하여 다른 언어에서 사용하는 OOP 개념과는 다릅니다.

`Reference class`는 다른 언어에서 사용하는 OOP 개념과 유사하며 R6 패키지를 이용해서 사용할 수 있습니다.

이 저작물은 크리에이티브 커먼즈 저작자표시-비영리-변경금지 4.0 국제 라이선스에 따라 이용할 수 있습니다.