

## *Chapter 3. Getting Started with R: data frame*

SEOUL WOMEN'S UNIVERSITY  
DEPT OF ECONOMICS

# Introduction

- 각 entity(사람, 기업, 국가 등)에 대해 다양한 characteristics(특성: 소득, 소비; 판매수입, 이윤; GDP, 인구 ...)들을 요약하는 데이터를 생각해볼 수 있다.

observation	성별	나이	소득	노동시간	...
person 1	F	26	...		
person 2	M	41	...		
person 3	F	37	...		
...					

이와 같은 형태의 데이터를 R로 표현하기 위해서 `data.frame()`을 사용할 수 있다.

- `data.frame()` 함수를 사용하여 data frame을 만들 수 있다.

	midterm	Economics	Business
person 1	90	50	
person 2	80	60	
person 3	60	100	
person 4	70	20	

## R Console

```
> econ <- c(90,80,60,70)
> econ
[1] 90 80 60 70
>
> biz <- c(50,60,100,20)
> biz
[1] 50 60 100 20
>
> midterm <- data.frame(econ,biz)
>
```

- `data.frame()`은 두 개의 numeric vector인 `econ`과 `biz`를 결합하여 하나의 data frame을 만드는 함수이다.
- 새로운 data frame의 이름은 `midterm`이다. 이어서 `midterm`을 type 해보자.

D:/MyR/CH00 - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Source

Console Terminal Jobs

D:/MyR/CH00/

```
> econ <- c(90,80,60,70)
> econ
[1] 90 80 60 70
> biz <- c(50,60,100,20)
> biz
[1] 50 60 100 20
> midterm <- data.frame(econ,biz)
> midterm
  econ biz
1   90  50
2   80  60
3   60 100
4   70  20
```

Environment History Connections

Import Dataset

Global Environment

Data

midterm 4 obs. of 2 variables

values

	num	[1:4]
biz	num	[1:4] 50 60 100 20
econ	num	[1:4] 90 80 60 70

Files Plots Packages Help Viewer

New Folder Delete Rename More

D: > MyR > CH00

	Name	Size	Modified
<input type="checkbox"/>	.Rhistory	884 B	Mar 5, 2020, 2:37 AM
<input type="checkbox"/>	CH00.Rproj	218 B	Mar 5, 2020, 3:03 PM
<input type="checkbox"/>	Example1.R	11 B	Mar 4, 2020, 11:09 PM
<input type="checkbox"/>	Example2.R	11 B	Mar 4, 2020, 11:19 PM

- data frame midterm에 있는 변수들은 econ과 biz이다. econ이란 변수만 보고 싶다면 `midterm$econ`이라고 입력하고, biz라는 변수만 보고 싶다면 `midterm$biz`라고 입력한다.

The screenshot shows the RStudio interface with the following components:

- Source Panel:** Contains R code for creating a data frame. The last two lines, `midterm$econ` and `midterm$biz`, are highlighted with a red box.
 

```
> econ <- c(90,80,60,70)
> econ
[1] 90 80 60 70
> biz <- c(50,60,100,20)
> biz
[1] 50 60 100 20
> midterm <- data.frame(econ,biz)
> midterm
  econ biz
1  90  50
2  80  60
3  60 100
4  70  20
> midterm$econ
[1] 90 80 60 70
> midterm$biz
[1] 50 60 100 20
> |
```
- Environment Panel:** Shows the 'Global Environment' with a data frame 'midterm' containing 4 observations of 2 variables.
 

Values	
biz	num [1:4] 50 60 100 20
econ	num [1:4] 90 80 60 70
- Files Panel:** Shows the file explorer for 'D:\MyR\CH00' with files like .Rhistory, CH00.Rproj, Example1.R, and Example2.R.

- econ 점수의 평균과 biz 점수의 평균

## R Console

```
> mean(midterm$econ)
[1] 75
>
> mean(midterm$biz)
[1] 57.5
>
```

- data frame을 한 번에 만들 수도 있다. Source 창을 사용하는 것이 좋다. Source 창에서 새로운 스크립트 파일을 열어서 다음과 같이 입력한다.

```
midterm <- data.frame(econ=c(90,80,60,70),
                      biz=c(50,60,100,20))
midterm
```

## • Run을 클릭한다.

The screenshot shows the RStudio interface with the following components:

- Top Menu Bar:** File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help.
- Toolbar:** Includes icons for saving, running, and other functions. The 'Run' button (a green play icon) is circled in red.
- Code Editor:** Contains the following R code:
 

```
1 midterm <- data.frame(econ=c(90,80,60,70),
2                       biz=c(50,60,100,20))
3 midterm
4
```
- Environment Pane:** Shows the 'Global Environment' with a table of values:
 

Data		
midterm	4 obs. of 2 variables	
Values		
biz	num [1:4]	50 60 100 20
econ	num [1:4]	90 80 60 70
- Console:** Shows the output of the code execution:
 

```
> midterm
  econ biz
1   90  50
2   80  60
3   60 100
4   70  20
> |
```
- Files Pane:** Shows the project structure:
 

	Name	Size	Modified
<input type="checkbox"/>	..		
<input type="checkbox"/>	.Rhistory	884 B	Mar 5, 2020, 2
<input type="checkbox"/>	CH00.Rproj	218 B	Mar 5, 2020, 3
<input type="checkbox"/>	Example1.R	11 B	Mar 4, 2020, 1
<input type="checkbox"/>	Example2.R	11 B	Mar 4, 2020, 1



## How To Import Files

- 'spreadsheet' 형태의 파일들을 R로 불러와서 data frame을 만들 수 있다. 다음과 같은 excel file을 R로 불러오자.

gdp - Excel

	A	B	C	D	E	F	G	H	I	J	K	L
1	country	y2015	y2016	y2017	y2018							
2	Australia	45584.32	45929.18	46442.2	46941.79							
3	Austria	42960.97	43284.02	44109.21	45062.49							
4	Belgium	40899.68	41284.37	41800.56	42193.39							
5	Canada	42498.06	42484.75	43234.4	43430.23							
6	Chile	20789.16	20867.08	20854.69	21407.89							
7	Czech Republ	29874.27	30541.63	31798.03	32574.86							
8	Denmark	44760.23	45458.32	46180.44	46618.44							
9	Estonia	26244.8	27106.84	28429.82	29450.49							
10	Finland	37987.8	38935.65	39874.82	40744.4							
11	France	36902.42	37168.78	37875.68	38408.15							
12	Germany	42690.7	43297.04	44066.19	44562.18							
13	Greece	23649.17	23702.53	24106.9	24635.39							
14	Hungary	24103.13	24725.73	25817.2	27139.16							
15	Iceland	43726.4	45984.72	46981.15	47863.97							

준비

- 1단계: 현재 작업 중인 working directory에 해당 파일이 있어야 한다.  
R로 불러올 파일을 working directory로 옮겨놓는다.

The screenshot shows the RStudio interface with the following components:

- Source Panel:** Shows the current working directory as `D:/MyR/CH00/`. The console has a prompt `> |`.
- Environment Panel:** Displays the current environment with 4 observations of 2 variables. The data is as follows:
 

Variable	Class	Values
biz	num [1:4]	50 60 100 20
econ	num [1:4]	90 80 60 70
- Files Panel:** Shows the file explorer for the directory `D:/MyR/CH00/`. The files listed are:
 

Name	Size	Modified
..		
.Rhistory	884 B	Mar 5, 2020, 2
CH00.Rproj	218 B	Mar 5, 2020, 3
Example1.R	11 B	Mar 4, 2020, 1
Example2.R	11 B	Mar 4, 2020, 1
gdp.xlsx	10.4 KB	Mar 5, 2020, 3

 The file `gdp.xlsx` is circled in red.

- 2단계: excel file을 불러오기 위해서는 `read_excel()`이라는 함수를 사용해야 한다. 이 함수는 `readxl`이라는 패키지에 들어있다. `readxl` 패키지를 설치한다.

The screenshot shows the RStudio IDE with the following components:

- Source Editor:** Contains the R script `install.packages("readxl")`, which is highlighted with a red box.
- Console:** Displays the output of the installation process, including a warning about Rtools, the download of the package from CRAN, and the successful unpacking of the binary package.
- Environment:** Shows the current data environment with 4 observations and 2 variables: `biz` and `econ`.
- Files Panel:** Shows the file explorer for the project directory `D:\MyR\CH00`, listing files such as `.Rhistory`, `CH00.Rproj`, `Example1.R`, `Example2.R`, and `gdp.xlsx`.

**Console Output:**

```

D:\MyR\CH00/
install.packages("readxl")
Warning: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate version of Rtools before proceeding:
https://cran.rstudio.com/bin/windows/Rtools/
Installing package into 'C:/Users/JOY/Documents/R/win-library/3.6'
(as 'lib' is unspecified)
URL 'https://cran.rstudio.com/bin/windows/contrib/3.6/readxl_1.3.1.zip'을 시도합니다
Content type 'application/zip' length 1524522 bytes (1.5 MB) downloaded 1.5 MB

package 'readxl' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
C:\Users\JOY\AppData\Local\Temp\Rtmpspncwd\downloaded_packages
>
  
```

**Environment Data:**

Variable	Type	Values
biz	num [1:4]	50 60 100 20
econ	num [1:4]	90 80 60 70

**Files Panel:**

Name	Size	Modified
..		
.Rhistory	884 B	Mar 5, 2020, 2
CH00.Rproj	218 B	Mar 5, 2020, 3
Example1.R	11 B	Mar 4, 2020, 1
Example2.R	11 B	Mar 4, 2020, 1
gdp.xlsx	10.4 KB	Mar 5, 2020, 3

- 3단계: `library(readxl)`를 입력해서 설치한 패키지를 load한다.
- 4단계: `read_excel("파일이름.xlsx")`를 입력해서 excel 파일을 불러온다. R로 불러들인 excel 파일은 data frame 형태로 자동 변환된다.

```
read_excel( "파일이름.xlsx" )
```

#### R Console

```
> library(readxl)
> gdp <- read_excel("gdp.xlsx")
> gdp
```

D:/MyR/CH00 - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins CH00

**Source**

Console Terminal Jobs

D:/MyR/CH00/

downloaded 1.5 MB

package 'readxl' successfully unpacked and MD5 sums checked

The downloaded binary packages are in  
C:\Users\JOY\AppData\Local\Temp\RtmpsxhIZT\download  
ed\_packages

```
> library(readxl)
> gdp <- read_excel("gdp.xlsx")
> gdp
# A tibble: 36 x 5
  country      y2015 y2016 y2017 y2018
  <chr>      <dbl>  <dbl>  <dbl>  <dbl>
1 Australia  45584.  45929.  46442.  46942.
2 Austria   42961.  43284.  44109.  45062.
3 Belgium   40900.  41284.  41801.  42193.
4 Canada    42498.  42485.  43234.  43430.
5 Chile     20789.  20867.  20855.  21408.
6 Czech Republic 29874. 30542. 31798. 32575.
7 Denmark   44760. 45458. 46180. 46618.
8 Estonia   26245. 27107. 28430. 29450.
9 Finland   37988. 38936. 39875. 40744.
10 France    36902. 37169. 37876. 38408.
# ... with 26 more rows
```

**Environment** History Connections

Import Dataset

Global Environment

**Data**

Variable	Obs.	Variables
gdp	36 obs.	of 5 variables
interim	4 obs.	of 2 variables

**Values**

Variable	Values
biz	num [1:4] 50 60 100 20
econ	num [1:4] 90 80 60 70

**Files** Plots Packages Help Viewer

New Folder Delete Rename More

D: > MyR > CH00

Name	Size	Modified
..		
.Rhistry	884 B	Mar 5, 2020, 2
CH00.Rproj	218 B	Mar 5, 2020, 3
Example1.R	11 B	Mar 4, 2020, 1
Example2.R	11 B	Mar 4, 2020, 1
gdp.xlsx	10.4 KB	Mar 5, 2020, 3

- 불러오고 싶은 파일이 working directory가 아닌 다른 곳에 있는 경우, path를 지정하면 된다.

```
read_excel( ".../.../파일이름.xlsx" )
```

#### R Console

```
> library(readxl)
> gdp <- read_excel("D:/WEHSOL_TEACHING/BigData2020/BD00/C/gdp.xlsx")
> gdp
```

D:/MyR/CH00 - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins CH00

**Source**

Console Terminal Jobs

```
D:/MyR/CH00/
> CH11e 20/89. 20867. 20855. 21408.
6 Czech Republic 29874. 30542. 31798. 32575.
7 Denmark 44760. 45458. 46180. 46618.
8 Estonia 26245. 27107. 28430. 29450.
9 Finland 37988. 38936. 39875. 40744.
10 France 36902. 37169. 37876. 38408.
# ... with 26 more rows

> gdp <- read_excel("D:/WEHSOL_TEACHING/BigData2020/BD00/C/
gdp.xlsx")
> gdp
# A tibble: 36 x 5
  country      y2015 y2016 y2017 y2018
  <chr>      <dbl> <dbl> <dbl> <dbl>
1 Australia  45584. 45929. 46442. 46942.
2 Austria   42961. 43284. 44109. 45062.
3 Belgium   40900. 41284. 41801. 42193.
4 Canada    42498. 42485. 43234. 43430.
5 Chile     20789. 20867. 20855. 21408.
6 Czech Republic 29874. 30542. 31798. 32575.
7 Denmark   44760. 45458. 46180. 46618.
8 Estonia   26245. 27107. 28430. 29450.
9 Finland   37988. 38936. 39875. 40744.
10 France    36902. 37169. 37876. 38408.
# ... with 26 more rows
> |
```

**Environment** History Connections

Import Dataset

Global Environment

Data

gdp	36 obs. of 5 variables
midterm	4 obs. of 2 variables

Values

biz	num [1:4] 50 60 100 20
econ	num [1:4] 90 80 60 70

**Files** Plots Packages Help Viewer

New Folder Delete Rename More

D: > MyR > CH00

Name	Size	Modified
..		
.Rhistory	884 B	Mar 5, 2020, 2
CH00.Rproj	218 B	Mar 5, 2020, 3
Example1.R	11 B	Mar 4, 2020, 1
Example2.R	11 B	Mar 4, 2020, 1
gdp.xlsx	10.4 KB	Mar 5, 2020, 3



- gdp가 어떻게 생겼는지 직접 확인할 수 있다.

The screenshot shows the RStudio interface with the 'gdp' dataset loaded. The Environment pane on the right shows 'gdp' with 36 observations of 5 variables. The console on the bottom left shows the command 'View(gdp)' being executed, which opens a viewer window for the dataset.

**Environment Pane:**

Variable	Description
gdp	36 obs. of 5 variables
midterm	4 obs. of 2 variables

**Console Output:**

```

D:/MyR/CH00/ > View(gdp)
# ... with 26 more rows
  
```

**Viewer Window (gdp):**

	country	y2015	y2016	y2017	y2018
1	Australia	45584.32	45929.18	46442.20	46941.79
2	Austria	42960.97	43284.02	44109.21	45062.49
3	Belgium	40899.68	41284.37	41800.56	42193.39
4	Canada	42498.06	42484.75	43234.40	43430.23
5	Chile	20789.16	20867.08	20854.69	21407.89
6	Czech Republic	29874.27	30541.63	31798.03	32574.86
7	Denmark	44760.23	45458.32	46180.44	46618.44

이번에는 다음과 같은 csv(comma separated values) file을 R로 불러오자.

- 1단계: 현재 작업 중인 working directory에 해당 파일이 있어야 한다.  
R로 불러올 파일을 working directory로 옮겨놓는다.

The screenshot shows the RStudio interface with the following components:

- Source Pane:** Shows the current working directory as `D:/MyR/CH00/`.
- Environment Pane:** Displays the loaded data objects:
 

Object	Size
gdp	36 obs. of 5 variables
midterm	4 obs. of 2 variables

 Below this, the 'Values' section shows:
 

Variable	Type	Range	Values
biz	num	[1:4]	50 60 100 20
econ	num	[1:4]	90 80 60 70
- Files Pane:** Shows the file explorer for the directory `D:\MyR\CH00`. The file `gdp.csv` is highlighted with a red box.

- 2단계: csv file을 불러오기 위해서는 `read.csv()`라는 함수를 사용해야 한다. `read.csv()`함수는 R에 내장되어 있다.
- 3단계: `read.csv("파일이름.csv")`를 입력해서 csv 파일을 불러온다. R로 불러들인 csv 파일은 data frame 형태로 자동 변환된다.

```
read.csv( ".../.../파일이름.csv", stringsAsFactors=F )
```

- `read.csv("파일이름.csv")`를 사용하여 문자(character)로 이루어진 data를 불러오면 자동으로 범주형(factor) 변수로 변환된다. 범주형 (factor) 변수로 변환될 경우 data 분석을 할 때 error가 발생하기 쉽다.
- `stringsAsFactors=F`라고 설정하면, 문자를 'character'로 불러온다.

#### R Console

```
> gdp <- read.csv("gdp.csv", stringsAsFactors=F)
> gdp
```

D:/MyR/CH00 - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins CH00

**Source**

Console Terminal Jobs

```
D:/MyR/CH00/ > gdp <- read.csv("gdp.csv", stringsAsFactors=F)
> gdp
```

	country	y2015	y2016	y2017	y2018
1	Australia	45584.32	45929.18	46442.20	46941.79
2	Austria	42960.97	43284.02	44109.21	45062.49
3	Belgium	40899.68	41284.37	41800.56	42193.39
4	Canada	42498.06	42484.75	43234.40	43430.23
5	Chile	20789.16	20867.08	20854.69	21407.89
6	Czech Republic	29874.27	30541.63	31798.03	32574.86
7	Denmark	44760.23	45458.32	46180.44	46618.44
8	Estonia	26244.80	27106.84	28429.82	29450.49
9	Finland	37987.80	38935.65	39874.82	40744.40
10	France	36902.42	37168.78	37875.68	38408.15
11	Germany	42690.70	43297.04	44066.19	44562.18
12	Greece	23649.17	23702.53	24106.90	24635.39
13	Hungary	24103.13	24725.73	25817.20	27139.16
14	Iceland	43726.40	45984.72	46981.15	47863.97
15	Ireland	60263.86	62560.46	66334.23	69898.18
16	Israel	31345.41	31967.07	32442.54	32867.66
17	Italy	33162.32	33590.40	34207.36	34557.46
18	Japan	37035.85	37313.39	38101.22	38481.31
19	Korea	34192.45	35035.41	35968.09	36778.42
20	Latvia	22146.74	22811.31	24092.42	25437.80
21	Lithuania	25784.38	26728.63	28288.35	29508.55
22	Luxembourg	87496.36	87347.42	86788.14	87316.63
23	Mexico	16659.51	16967.78	17143.90	17314.83

**Environment** History Connections

Import Dataset

Global Environment

Data

gdp	36 obs. of 5 variables
midterm	4 obs. of 2 variables

Values

biz	num [1:4] 50 60 100 20
econ	num [1:4] 90 80 60 70

**Files** Plots Packages Help Viewer

New Folder Delete Rename More

D: > MyR > CH00

	Name	Size	Modified
	..		
	.Rhistory	884 B	Mar 5, 2020, 2
	CH00.Rproj	218 B	Mar 5, 2020, 3
	Example1.R	11 B	Mar 4, 2020, 1
	Example2.R	11 B	Mar 4, 2020, 1
	gdp.xlsx	10.4 KB	Mar 5, 2020, 3
	gdp.csv	2.1 KB	Mar 5, 2020, 4