

강의교안 이용 안내

- 본 강의교안의 저작권은 이윤환과 한빛아카데미(주)에 있습니다.
- 이 자료를 무단으로 전제하거나 배포할 경우 저작권법 136조에 의거하여 벌금에 처할 수 있고 이를 병과(併科)할 수도 있습니다.





제대로 알고 쓰는

R 통계분석

CHAPTER 01

통계학과 R의 시작

Contents

B

R에서의 자료구조

- 벡터, factor, 데이터 프레임
- 함수

2장을 위한 준비



부록 B. R에서의 자료구조

: 다양한 자료들의 모임

자료구조

• 자료구조

- 사전적 정의 : 전산학에서 자료를 효율적으로 이용할 수 있도록 컴퓨터에 저장하는 방법이다. (위키피디아, 자료구조)
- 통계에서 데이터 세트는 복수의 관찰대상으로부터 복수의 속성을 기록하므로 적절한 자료구조를 통해 자료를 효율적으로 관리할 필요가 있습니다.
- 예) 3명의 학생으로부터 키를 측정하고 이를 R로 저장해 봅시다.
 - 다음과 같이 3개의 변수를 마련하는 것을 생각해 볼 수 있습니다.

```
> height1 <- 170  
> height2 <- 174  
> height3 <- 162
```

- 이때 키 변수 하나에 관리해야 하는 변수는 3개로 만일 조사대상이 늘어난다면, 변수는 점점 늘어나 관리가 힘들어 질 것입니다.
- R에서 어떻게 효율적으로 이런 자료들을 관리하는지 알아보시다.

자료구조 : 벡터

- **벡터 : 동일한 자료형의 단일 값들이 한군데 모여있는 자료구조**
 - 벡터는 R의 가장 기본적인 자료 저장 방법입니다.
 - 벡터는 동일한 자료형을 갖는 값들의 집합으로, 일반적으로 하나의 속성을 저장하는 단위로 사용합니다.
- **벡터 생성하기**
 - R에서 제공하는 벡터 생성방법에 대해 알아보시다.
 - 벡터 생성 연산자 “:”(콜론)
 - ‘시작값 : 종료값’의 형태로 사용하며, 시작값부터 종료값까지 1씩 더하거나 빼서 벡터를 생성합니다. 다음의 코드는 vector 생성 연산자 ‘:’의 사용 예입니다.

```
> 1:5
[1] 1 2 3 4 5
> 5:1
[1] 5 4 3 2 1
```

1:5 는 1에서부터 5까지
1씩 증가하는 벡터를 생성합니다.

5:1 은 5에서부터 1까지
1씩 감소하는 벡터를 생성합니다.

자료구조 : 벡터

- 벡터로 생성된 출력물을 조금 더 살펴봅시다.
 - 1:1000으로 1에서부터 1000까지 1씩 증가하는 벡터의 출력물입니다.

```
> 1:1000
```

[1]	1	2	3	4	5	6	7	8	9	10	11	12
[13]	13	14	15	16	17	18	19	20	21	22	23	24
						...						
[985]	985	986	987	988	989	990	991	992	993	994	995	996
[997]	997	998	999	1000								

- 출력물의 각 첫줄에 대괄호로 둘러싼 숫자가 보입니다.
 - 이 숫자는 해당 벡터에서 위치를 나타내며(인덱스라고도 합니다.) 벡터 출력물에서 현재의 위치를 알려줍니다.
 - 위의 출력에서 2번째 줄의 [13]은 해당 줄의 첫번째 원소가 벡터에서 13번째 원소임을 나타냅니다.
- R에서 벡터의 출력은 위와 같이 대괄호 안에 숫자가 들어간 형태로 나타납니다.

자료구조 : 벡터

- ▣ 벡터생성함수 : `c()`, `seq()`, `rep()`
 - 기본 vector 생성함수입니다. 콤마로 구분된 전달인자들로 벡터를 구성할 원소를 전달합니다.

```
> c(1, 2, 3)
[1] 1 2 3
```

`c(1, 2, 3)` 은 1, 2, 3으로 구성된 벡터를 생성합니다.

```
> c(1, 2, 3, c(4, 5, 6))
[1] 1 2 3 4 5 6
```

`c()` 함수 안에 또 다른 벡터를 넣을 경우 하나의 벡터가 됩니다.

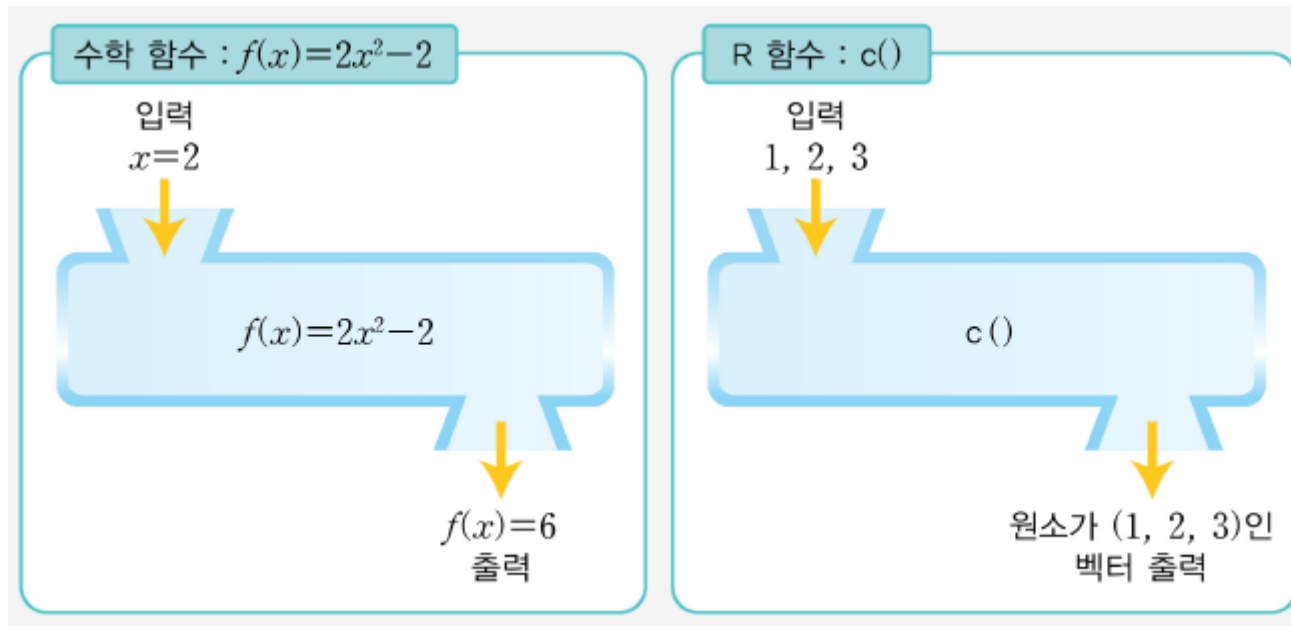
```
> x <- c(1, 2, 3)
> x [1] 1 2 3
```

일반적으로 벡터를 생성하고 변수로 저장하여 사용합니다.

자료구조 : 벡터

함수

- 프로그래밍 언어에 따라 프로시저(procedure), 메소드(method) 등으로 불리며, 수학에서의 함수와 마찬가지로 함수 작동에 필요한 입력으로 함수내부에서 계산을 함으로써 생성되는 적절한 출력을 내보내는 코드들의 모임입니다.
- c() 함수의 경우 함수명 “c” 이후 소괄호() 사이에 벡터를 구성하는 원소들을 콤마(,)로 구분하여 입력으로 사용하고, 출력으로는 입력에 사용된 자료들로 구성된 벡터를 반환합니다.



자료구조 : 벡터

- 프로그래밍에서 함수는 숫자 외에 다양한 자료들을 함수 수행시 필요한 입력으로 사용하는데, 이를 '전달인자'라고 부릅니다.
- 전달인자는 이름과 순서로 구별할 수 있습니다. 다음에 나오는 seq() 함수의 예제를 살펴봅시다.
- 예제) seq() 함수로 벡터생성 : 전달인자 사용
 - 함수가 지정한 전달인자의 이름 사용

```
> seq(from=1, to=5, by=2)
[1] 1 3 5
```

from으로 전달되는 값은 생성할 벡터의 시작값을 받습니다.

to로 전달되는 값은 생성할 벡터의 시작값을 받습니다.

By에 전달되는 값은 초기값부터 종료값까지의 변화량입니다.
1부터 5까지 2씩 증가하는 벡터

자료구조 : 벡터

- 전달인자의 위치를 통해 함수가 사용하는 값 전달

```
> seq(1, 5, 2)
[1] 1 3 5
```

첫번째 전달인자는 생성할 벡터의 시작값인 from으로 받습니다.

두번째 전달인자는 생성할 벡터의 종료값인 to로 받습니다.

세번째 전달인자는 생성할 벡터의 증감값인 by로 받습니다.
1부터 5까지 2씩 증가하는 벡터

- ▣ 전달인자는 함수에 따라 다르므로 함수 사용에 앞서 도움말 함수인 ‘help(함수명)’을 통해 필요로 하는 전달인자들을 확인도 하고 함수 설명도 들어봅시다. (예 : help(c))
 - help(함수명) 과 동일하게 ?**함수명**도 사용합니다.
 - 함수명 외에 다양한 R의 자원들에 대한 도움말을 얻을 수 있습니다.
 - help(cars) # R의 내장자료인 cars에 대한 도움말을 얻습니다.

자료구조 : 벡터

c {base}

1

R Documentation

Combine Values into a Vector or List

2

Description

This is a generic function which combines its arguments.

The default method combines its arguments to form a vector. All arguments are coerced to a common type which is the type of the returned value, and all attributes except names are removed.

3

Usage

```
c(..., recursive = FALSE)
```

4

Arguments

... objects to be concatenated.

recursive logical. If recursive = TRUE, the function recursively descends through lists (and pairlists) combining all their elements into a vector.

5

Details

The output type is determined from the highest type of the components in the hierarchy NULL < raw < logical < integer < double < complex < character < list < expression. Pairlists are treated as lists, but non-vector components (such names and calls) are treated as one-element lists which cannot be unlisted even if recursive = TRUE.

c is sometimes used for its side effect of removing attributes except names, for example to turn an array into a vector. as.vector is a more intuitive way to do this, but also drops names. Note too that methods other than the default are not required to do this (and they will almost certainly preserve a class attribute).

6

자료구조 : 벡터

▣ R의 도움말 구조

① 함수의 이름과 함수가 속한 패키지

- `c()` 함수의 경우 기본 패키지인 `base`에 속하고 있음을 나타냅니다. R에서 함수는 패키지 별로 존재합니다. 패키지에 대해서는 2장의 [3장을 위한 준비]를 참고합니다.

② 함수에 대한 간략한 설명

③ Description : 함수에 대한 자세한 설명

④ Usage : 함수의 사용법

- 다른 프로그래밍 언어에서 이야기하는 함수의 원형과 유사한 형태로 함수의 원형은 해당 함수가 출력으로 전달하는 자료의 유형 및 모든 전달인자를 기술하는 것을 뜻합니다.

⑤ Arguments : 함수 호출 시 전달되는 값 (전달인자) 에 대한 설명

- 위의 예에서 `recursive`로 전달되는 값은 인수로 논리값(logical)을 가짐을 나타내며, 그 값이 `TRUE`일 경우 벡터 생성의 순서를 반대로 함을 설명하고 있습니다. 이처럼 이 섹션에서는 전달인자의 유형과 값에 대해 설명합니다.

⑥ Details : 함수 사용에 대한 자세한 설명

- ▣ 도움말에서 제공하는 예제는 함수에 대한 이해와 주의점을 잘 나타내고 있습니다.

자료구조 : 벡터

Examples

```

c(1,7:9)
c(1:5, 10.5, "next")

## uses with a single argument to drop attributes
x <- 1:4
names(x) <- letters[1:4]
x
c(x)          # has names
as.vector(x)  # no names
dim(x) <- c(2,2)
x
c(x)
as.vector(x)

## append to a list:
ll <- list(A = 1, c = "C")
## do *not* use
c(ll, d = 1:3) # which is == c(ll, as.list(c(d = 1:3)))
## but rather
c(ll, d = list(1:3)) # c() combining two lists

c(list(A = c(B = 1)), recursive = TRUE)

c(options(), recursive = TRUE)
c(list(A = c(B = 1, C = 2), B = c(E = 7)), recursive = TRUE)

```

자료구조 : 벡터

- 다시 벡터로 돌아가 `seq()` 함수를 좀 더 살펴봅시다.

```
seq(from, to, by | length.out)
```

▷ from : 초깃값

▷ to : 종료값

▷ by : 증가분

▷ length.out : from부터 to 사이의 생성할 vector의 개수 지정

```
> seq(0, 1, by=0.001)
> seq(0, 1, length.out=1000)
```

증가분을 나타내는 `by` 나 생성할 벡터의 개수를 지정하는 `length.out` 두 개 중에 하나를 통해 초깃값부터 종료값까지 생성되는 벡터를 만들 수 있습니다.
`by`와 `length.out`의 결과의 차이가 무엇인지 이야기 해 봅시다.

자료구조 : 벡터

- `rep()` : 기존에 있는 벡터를 반복하여 새로운 벡터를 만듭니다.
- 반복은 벡터 전체의 반복과 벡터의 각 원소 반복의 두가지가 있습니다.

```
rep(x, times | each)
```

- ▷ `x` : 반복할 자료(vector)
- ▷ `times` : 전달된 벡터 `x`의 전체 반복 횟수
- ▷ `each` : 전달된 벡터 `x`의 개별 원소들의 반복 횟수

```
> rep(c(1, 2, 3), times=2)
[1] 1 2 3 1 2 3
```

`time`는 벡터 전체의
반복횟수를 받습니다.

```
> rep(c(1, 2, 3), each=2)
[1] 1 1 2 2 3 3
```

`each`는 벡터의 개별원소의
반복횟수를 받습니다.

자료구조 : 벡터

- 벡터내의 원소에 접근하기
 - 위치정보로 접근하기
 - 벡터에 포함되는 자료는 인덱스(Index)라는 위치정보를 가집니다.
 - 인덱스는 1부터 시작하는 정수입니다.
 - 벡터 이름 뒤에 대괄호 []를 써서 인덱스를 지정하여 벡터 내의 원하는 위치의 원소들을 추출합니다.
 - 논리값이 TRUE인 원소 접근하기
 - 벡터 이름 뒤에 대괄호 안에 논리값 벡터를 넣어 논리값이 TRUE인 자료들을 추출합니다.

자료구조 : 벡터

▣ 예제 : 위치정보로 접근하기

```
> x <- c(5, 4, 3, 2, 1)
> x[1]
[1] 5
> x[1, 2, 3]
Error in x[1, 2, 3] :
incorrect number of
dimensions
> x[c(1, 2, 3)]
[1] 5 4 3
> x[-c(1, 2, 3)]
[1] 2 1

> length(x)
[1] 5
> x[3:length(x)]
[1] 3 2 1
```

벡터 x의 첫번째 원소를 가져옵니다.

여러 위치정보를 전달할 때는 벡터로 전달합니다.

음수의 위치정보는 해당 위치는 제외하고 나머지 원소를 가져오게 합니다.

length() 함수는 전달된 벡터의 원소의 개수를 반환합니다.

자료구조 : 벡터

- 예제 : 논리값이 TRUE인 원소 접근하기

```
> ex <- c(1, 3, 7, NA, 12)
> ex < 10
[1] TRUE TRUE TRUE NA
FALSE

> ex[ex < 10]
[1] 1 3 7 NA

> ex[ex %% 2 == 0]
[1] NA 12

> ex[is.na(ex)]
[1] NA

> ex[ex %% 2 == 0 & !is.na(ex)]
[1] 12
```

벡터 ex에서 10보다 작은 원소는 TRUE를 그렇지 않은 원소는 FALSE로 구성된 벡터를 반환합니다.

대괄호에 논리연산을 넣으면 논리연산이 TRUE인 위치의 값을 벡터에서 가져옵니다. (NA는 항상 따라옵니다)

ex의 각 원소를 2로 나눈 나머지가 0인지 판별하여, 즉 ex의 원소가 짝수인 값을 가져옵니다.

is.na() 함수는 주어진 벡터의 각 값이 NA이면 TRUE를 그렇지 않으면 FALSE인 벡터를 반환합니다.

짝수이고 NA가 아닌(!is.na) 두 조건을 만족(&, AND)하는 원소를 가져옵니다.

자료구조 : factor

- factor
 - 저장 값의 크기보다 의미가 중요한 질적 자료를 위해 사용됩니다.
 - 예를 들어 숫자 1, 2, 3은 산술연산을 통해 계산되는 본래의 숫자로서의 기능을 하지만, factor로 지정된 1, 2, 3은 단지 세 개의 그룹 혹은 상태를 구별 짓는 의미로 사용됩니다.
- factor 생성함수 : factor()

```
factor(x = character(), levels, labels = levels, ordered=FALSE)
```

- ▷ x : factor로 만들 벡터
- ▷ levels : 주어진 데이터 중 factor의 각 값(수준)으로 할 값을 벡터 형태로 지정(여기서 빠진 값은 NA로 처리).
- ▷ labels : 실제 값 외에 사용할 각 수준의 이름(벡터), 예를 들어 데이터에서 1이 남자를 가리킬 경우 labels를 통해 '남자' 혹은 'M' 등으로 변경.
- ▷ ordered : 순위형 자료 여부(TRUE/FALSE)로, levels에 입력한 순서를 가짐.

자료구조 : factor

- factor() 사용하기

```
> x <- c(1, 2, 3, 4, 5)
> factor(x, levels=c(1, 2, 3, 4))
[1] 1      2      3      4      <NA>
Levels: 1 2 3 4
> factor(x, levels=c(1, 2, 3, 4), labels=c("a", "b", "c", "d"))
[1] a      b      c      d      <NA>
Levels: a b c d
> factor(x, levels=c(1, 2, 3, 4), ordered=TRUE)
[1] 1      2      3      4      <NA>
Levels: 1 < 2 < 3 < 4
```

자료구조 : factor

▣ Code 설명

- 1, 2, 3, 4, 5로 구성된 벡터를 변수 x에 저장합니다.
- factor() 함수를 이용하여 기존 벡터를 factor로 변경합니다.
 - levels는 factor에 사용될 각 범주를 구분할 기존 벡터의 값을 전달인자로 하여 벡터를 전달합니다.
 - 만일 levels 전달인자에 전달되지 않은 값은 NA가 됩니다.
 - factor 출력시 하단의 Levels를 통해 해당 factor의 수준을 표시해 줍니다.
- labels는 각 수준의 이름으로 사용자가 정한 벡터를 받습니다.
 - levels로 지정한 기존 데이터 (1, 2, 3, 4)에 대해 새로운 이름표(labels)로 ("a", "b", "c", "d ")를 사용합니다.
- ordered 에 TRUE를 넣어주면 순서 있는 factor가 됩니다.(순위형 자료)
 - 출력시 levels에 부등호를 이용하여 순위를 나타냄을 확인해 주세요.
- factor에 대한 좀 더 자세한 설명은 7장의 “8장을 위한 준비”에서 학습합니다.

자료구조 : 데이터 프레임

- 데이터 프레임
 - 자료 처리를 위해 가장 많이 사용된 자료구조 입니다.
 - 서로 다른 벡터로 구성된 자료들을 열로 배치한 자료구조입니다.
 - 설명을 위해 지금은 자료를 직접 입력해서 만드는 데이터 프레임에 대해 알아보지만, 직접 입력하는 경우보다 외부 데이터로부터 가져오는 경우가 많습니다.
 - 6장의 “7장을 위한 자료준비”에서 더 자세히 알아보겠습니다.
- 생성함수 : `data.frame()`

```
data.frame(..., row.names = NULL,
           stringsAsFactors = default.stringsAsFactors())
```

- ▷ ... : 데이터 프레임을 구성할 열 정의 (값 혹은 열이름 = 자료의 형태)
- ▷ `row.names` : 행의 이름으로 사용할 값 저장. 기본값은 NULL로 각 행의 번호 저장
- ▷ `stringsAsFactors` : 문자열로 구성된 자료를 factor로 변환할지 여부로 기본값은 문자열을 factor로 변환

자료구조 : 데이터 프레임

- 생성 예제

```
> name <- c("철수", "영희", "길동")
> age <- c(21, 20, 31)
> gender <- factor(c("M", "F", "M"))
> character <- data.frame(name, age, gender)
> str( character )
'data.frame': 3 obs. of 3 variables:
 $ name   : Factor w/ 3 levels "길동","영희",...: 3 2 1
 $ age    : num  21 20 31
 $ gender : Factor w/ 2 levels "F","M": 2 1 2
> character
  name age gender
1 철수  21      M
2 영희  20      F
3 길동  31      M
```


자료구조 : 데이터 프레임

▣ Code 설명

- 데이터 프레임으로 구성할 벡터를 준비합니다. 각 벡터의 크기는 동일해야 하며, 각 벡터내의 위치정보가 동일한 값이 관찰대상이 됩니다.
 - name, age, gender는 관찰대상의 이름, 나이, 성별을 저장한 벡터입니다.
- data.frame() 함수에 위의 세 벡터를 전달하면, 세 개의 열로 구성된 데이터 프레임을 생성합니다. 이렇게 생성된 데이터 프레임을 변수 character에 저장합니다.
- str() 함수를 이용하여 데이터 프레임의 구조를 살펴봅니다.
 - 3개의 변수(3 variables)를 갖는 3개의 관찰치가 있음(obs.)을 알려줍니다.
 - 각 변수에 대한 구조를 보여줍니다.
 - ▣ \$ 이후에 데이터 프레임 내에서의 이름을 보여줍니다.
 - ▣ 각 벡터의 자료형을 보여줍니다. (num : 수치형 자료, Factor : factor형 자료)
 - ▣ 각 자료들의 앞의 일부 자료를 보여줍니다.
- 데이터 프레임 character를 출력합니다.

자료구조 : 데이터 프레임

- 데이터 프레임의 각 요소에 접근하기
 - 가져온 자료구조가 데이터 프레임인지 벡터인지 잘 구분합시다.

```
> character$name
[1] 철수 영희 길동
Levels: 길동 영희 철수
```

```
> character[1, ]
  name age gender
1  철수  21      M
```

```
> character[ , 2]
[1] 21 20 31
```

```
> character[3, 1]
[1] 길동
Levels: 길동 영희 철수
```

“데이터프레임명\$열이름”의 구조로 각 열에 접근합니다. 이 때 반환되는 자료구조는 벡터입니다.

대괄호 안에逗를 이용하여 가져오고자 하는 자료의 행과 열 위치정보를 넣습니다.

[1,]는 1행의 모든 열을 가져옵니다.
(데이터 프레임으로 가져옵니다.)

[, 2]는 모든 관찰치의 2번째 변수를 가져옵니다.
(벡터로 가져옵니다. 복수의 열일 경우 데이터 프레임이 됩니다.)

[3, 1]은 3번째 관찰치의 1번째 변수 값을 가져옵니다.



2장을 위한 준비

: 외부로부터 자료 가져오기

외부로부터 자료 가져오기

- **통계청에서 제공하는 자료 활용하기**

- 자료 수집의 시작은 통계청입니다.
- 통계청은 기획재정부 산하의 외청으로 1948년 공보처의 통계국으로부터 시작되어 “국가통계발전을 선도하고 신뢰받는 통계를 생산한다”는 미션을 두고 국가 수준의 각종 통계자료들을 생산, 관리 및 인증하고 있으며, 다양한 서비스를 통해 각종 통계자료들을 배포하고 있습니다. [표 1-6]은 통계청이 제공하는 서비스 중 제가 자주 사용하는 서비스들을 정리한 것입니다.

외부로부터 자료 가져오기

[표 1-6] 통계청이 제공하는 다양한 서비스

서비스명	주소	설명
국가통계포털	http://kosis.kr	주제별로 다양한 통계들을 제공하며, 통계를 얻기 위한 조사들에 대한 설명들도 함께 제공하는 서비스
국가지표체계	http://www.index.go.kr	국가주요지표, e-나라지표, 국민 삶의 지표, 녹색성장 지표 등 우리나라의 각종 상황들을 지표화 및 시각화하여 한눈에 알아볼 수 있도록 한 서비스
SGIS + 통계지리정보서비스	http://sgis.kostat.kr	각종 통계들을 지리정보와 결합하여, 지도를 통해 정보들을 탐색할 수 있도록 하였으며, 지도 제작을 위한 각종 지리정보를 제공하는 서비스
마이크로데이터 통합서비스	https://mdis.kostat.go.kr	통계자료가 아닌 원자료에서 입력오류 등을 제거한 마이크로데이터(microdata, 통계기초자료)를 제공하는 서비스로서 무료로 제공하는 공공용 마이크로데이터, 유료로 제공하는 인가된 마이크로데이터를 사용할 수 있는 서비스

- 통계자료가 아닌 수집된 자료 자체에 대한 요구가 점점 발생하는 요즘 통계청의 “마이크로데이터 통합서비스”는 이런 요구를 일부 해소해 주는 서비스로 ‘2010년 인구주택 총조사’의 일부 자료를 받아오고 R에서 읽는 과정을 함께 해 보시다.

외부로부터 자료 가져오기

실습 개요

파일 형태로 존재하는 외부 자료를 R에서 불러옵니다.



mdis로부터 자료추출



csv로 확장자 변환
(필수는 아님)



R로 읽기 : read.csv()

```
> str( data )  
'data.frame':  468284 obs. of  5 variables:  
 $ V1: Factor w/  2 levels "남자","여자": 1 1 1 1 1 1 1 1 1 1 ...  
 $ V2: int   0 0 0 0 0 0 0 0 0 0 ...  
 $ V3: Factor w/ 14 levels "가구주","가구주의 배우자",...: 3 3 3 3 3 3  
 3 3 3 3 ...  
 $ V4: Factor w/  8 levels "안 받았음","초등학교",...: 1 1 1 1 1 1 1 1 1  
 1 ...  
 $ V5: int  NA NA NA NA NA NA NA NA NA NA ...
```



읽은 온 자료 정리

외부로부터 자료 가져오기

예제 1-3 통계청이 제공하는 마이크로데이터 받아오기

• 실습내용

- ▣ 통계청의 통계기초자료(마이크로데이터) 제공 서비스인 '마이크로데이터 통합서비스'로부터 2010년 인구주택총조사 데이터에서 '성별', '나이', '가구주와의 관계', '교육정도', '총 출생아수'를 받아옵니다.
- ▣ 먼저 마이크로데이터 통합서비스를 원활히 이용하기 위해서는 통계청의 통합 아이디(통합 One ID)를 생성합니다.
 - <https://kosis.kr/oneid/cmmn/login/LoginView.do>
 - “14세 이상 통합회원 가입하기”를 통해 통합아이디를 생성합니다.
 - 휴대폰 및 아이핀 인증이 필요합니다.
 - 통합아이디를 생성하면 통계청에서 제공하는 각종 서비스에 별도의 회원가입 없이 이용할 수 있습니다.
 - 통합아이디를 갖고 있는 경우로 가정하고 실습을 진행합니다.

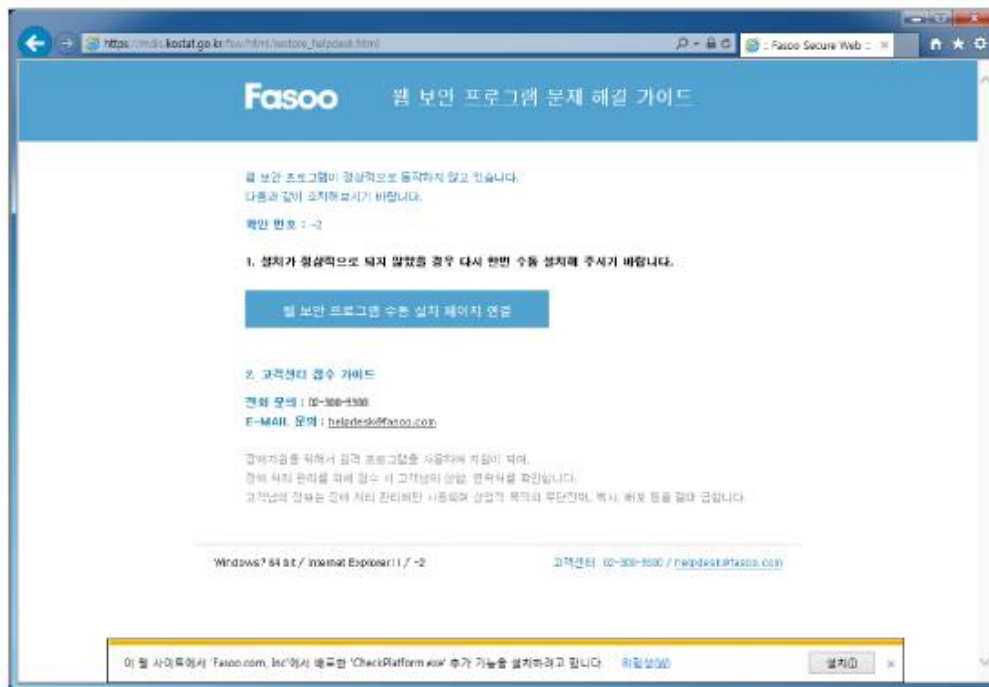
외부로부터 자료 가져오기

- 웹 브라우저를 열고 주소창에 통계청의 '마이크로데이터 통합서비스'의 주소 (<https://mdis.kostat.go.kr>)를 입력합니다.
- 첫 화면에서 좌측의 로그인을 클릭하고 통계청 통합 아이디로 로그인합니다.



외부로부터 자료 가져오기

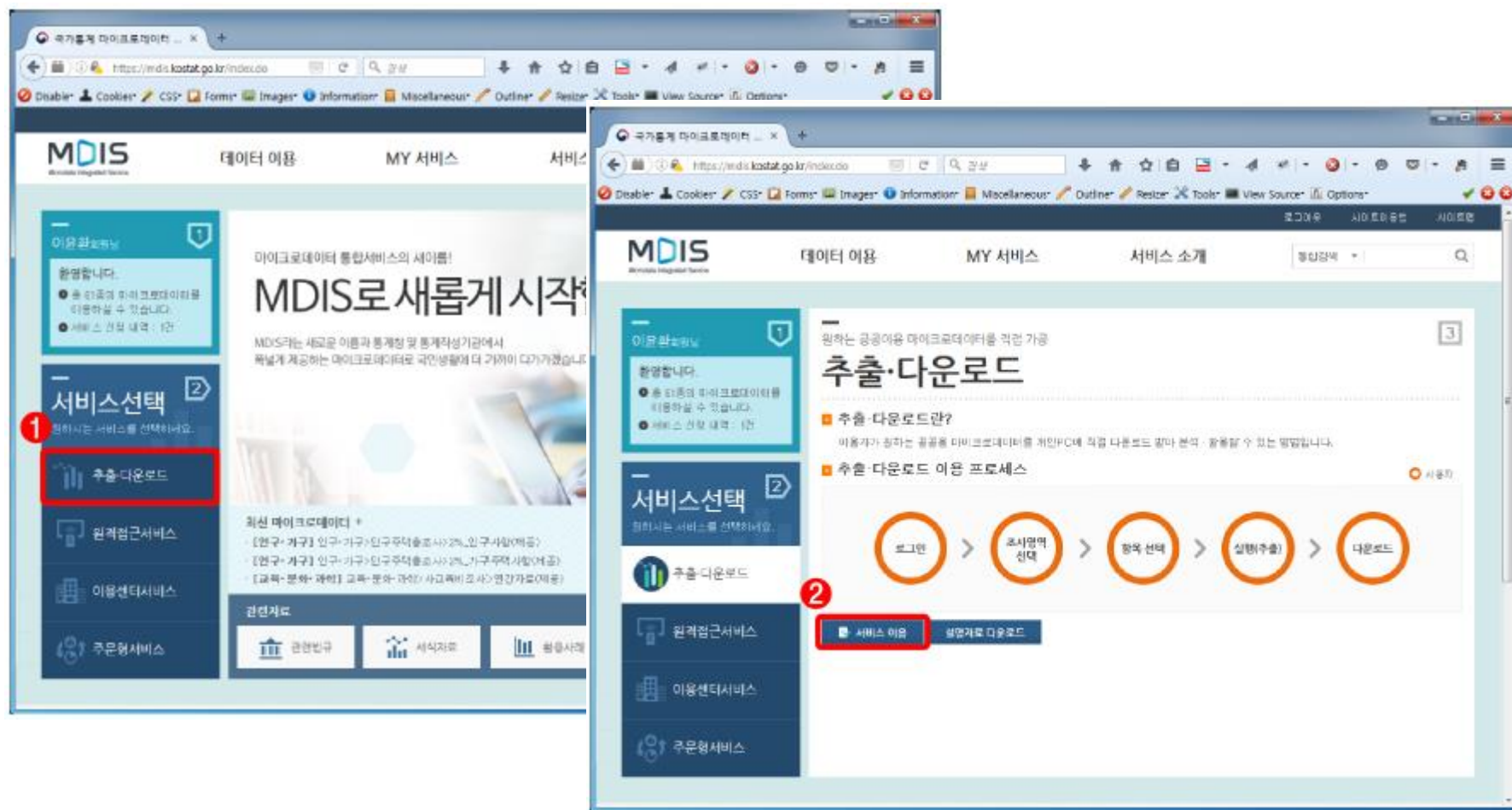
- 처음 접속 시 Fasoo 웹 보안프로그램 설치를 필요로 할 수 있습니다.
 - Fasoo 웹 보안프로그램은 기업용 문서보안 솔루션으로 기업, 관공서 사용합니다.
 - 본 서비스는 Fasoo 웹 보안프로그램 설치가 되어야 로그인 할 수 있습니다.
 - 문제점 : 사용시 컴퓨터의 점유율이 높아지고 설치된 엑셀(2013버전) 등과 충돌이 보고 되기도 합니다.



마이크로 소프트의 문제해결
<https://support.microsoft.com/ko-kr/kb/2529286>

외부로부터 자료 가져오기

- ❶ '서비스선택'에서 '추출·다운로드'를 클릭합니다.
- '추출·다운로드' 프로세스가 나오면 하단의 ❷ '서비스 이용'을 클릭합니다.



외부로부터 자료 가져오기

- 데이터 이용화면의 '분야별' 메뉴에서 원하는 정보 추출하기

- ① 대분류 중 '인구·가구'를 클릭합니다.
- ② 하위분류 중 '인구주택총조사'를 클릭합니다.
- ③ 조사년도에서 '2010'을 선택합니다.
- ④ 하위분류에서 '1%_인구사항(제공)'을 클릭합니다.

데이터 이용

이용안내

서비스 수수료

- 산정기준
- 비용계산

추출·다운로드

- 추출·다운로드 안내
- 추출·다운로드 범위 조회
- 추출
- 다년도
- 집계
- 설명자료

원격접근서비스

- 원격접근서비스 안내
- 원격접근서비스 신청
- 개인정액제신청

이용센터서비스


추출

통계항목을 '단어'로 입력

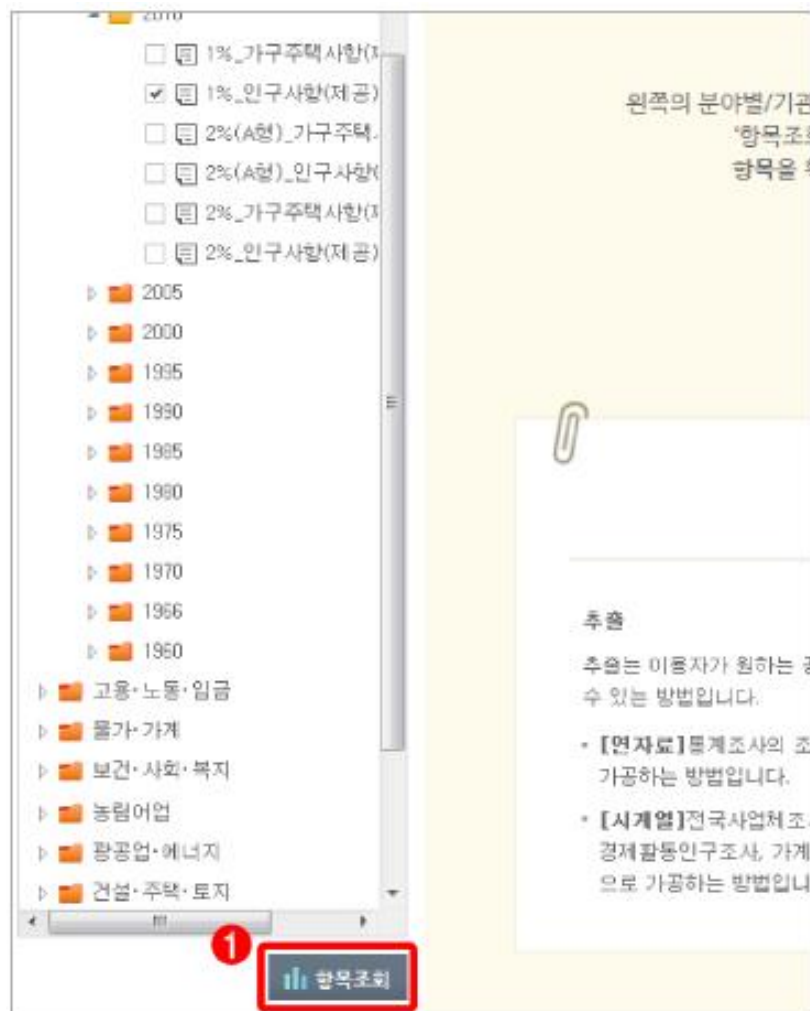
분야별 기관별 검색결과

- 1 인구·가구
 - ▷ 가족실태조사
 - ▷ 국내인구이동통계
 - ▷ 인구동향조사
- 2 인구주택총조사
 - 3 2010
 - 4 ☐ 1%_가구주택사항(제공)
 - ☐ 2%(A형)_가구주택...
 - ☐ 2%(A형)_인구사항(제...
 - ☐ 2%_가구주택사항(제...
 - ☐ 2%_인구사항(제공)
 - ▷ 2005
 - ▷ 2000
 - ▷ 1995
 - ▷ 1990
 - ▷ 1985

외부로부터 자료 가져오기

- ▣ '항목조회'에서 원하는 측정 대상(변수) 선택하기
 - ① 화면 하단의 '항목조회'를 클릭합니다.
 - ② 화면 오른쪽에 나타나는 '설명자료' 창의 '전체항목'에서 '성별', '나이(만나이)', '가구주와의관계', '교육정도'를 선택합니다.
 - ③ 화면을 내려 '총출생아수1'도 선택하고, 화면 중간의  아이콘을 클릭하여 '선택항목'으로 옮깁니다.
 - ④ '선택항목'에 있는 모든 변수들을 선택한 후 '데이터 추출'을 클릭하여 다음 단계인 다운로드 실행으로 진행합니다.

외부로부터 자료 가져오기



[그림 1-33] '항목조회' 클릭



[그림 1-34] 원하는 변수 선택

외부로부터 자료 가져오기

설명자료 ☐ (2015)자료 이용시 주의사항.hwp ☒ 2010 파일설계서 및 코드.xls ☐ 2010 인구주택총조사 표본조사표.pdf

전체항목

<input type="checkbox"/>	번호	형태	항목	<input type="checkbox"/>	번호	형태	항목
<input type="checkbox"/>	47	문자	근로장소	<input type="checkbox"/>	48	문자	혼인상태
<input type="checkbox"/>	49	문자	초혼연령	<input checked="" type="checkbox"/>	50	문자	출생아수1
<input type="checkbox"/>	51	문자	출생아수2	<input type="checkbox"/>	52	문자	출생아수1
<input type="checkbox"/>	53	문자	출생아수2	<input type="checkbox"/>	54	문자	추가 계획 자녀여부
<input type="checkbox"/>	55	문자	추가 계획 자녀여부	<input type="checkbox"/>	56	문자	추가 계획 자녀여부

선택항목

<input type="checkbox"/>	번호	형태	항목	조건설정
선택항목이 없습니다.				

미리보기 순서 변경 항목저장 DDI 조회 데이터다운로드

[그림 1-35] 선택된 변수 적용

설명자료 ☐ (2015)자료 이용시 주의사항.hwp ☒ 2010 파일설계서 및 코드.xls ☐ 2010 인구주택총조사 표본조사표.pdf

전체항목

<input type="checkbox"/>	번호	형태	항목	<input type="checkbox"/>	번호	형태	항목
<input type="checkbox"/>	47	문자	근로장소	<input type="checkbox"/>	48	문자	혼인상태
<input type="checkbox"/>	49	문자	초혼연령	<input checked="" type="checkbox"/>	50	문자	출생아수1
<input type="checkbox"/>	51	문자	출생아수2	<input type="checkbox"/>	52	문자	출생아수1
<input type="checkbox"/>	53	문자	출생아수2	<input type="checkbox"/>	54	문자	추가 계획 자녀여부
<input type="checkbox"/>	55	문자	추가 계획 자녀여부	<input type="checkbox"/>	56	문자	추가 계획 자녀여부

선택항목


<input checked="" type="checkbox"/>	번호	형태	항목	조건설정
<input type="checkbox"/>	1	문자	성별	입력
<input type="checkbox"/>	2	문자	나이(만나이)	입력
<input type="checkbox"/>	3	문자	가구주와의관계	입력
<input type="checkbox"/>	4	문자	교육정도	입력
<input type="checkbox"/>	5	문자	출생아수1	입력

UI_POR_P1070#layer_popDataExtractInfoInp DDI 조회 데이터다운로드

[그림 1-36] 다운로드 실행으로 이동

외부로부터 자료 가져오기

▣ 파일 추출 신청하기

- ① 자료의 이용 용도를 입력하는 창이 나오면,
 - '제목'은 나타내기 쉬운것으로 정하고,
 - '이용목적'은 '교육자료'를 선택합니다.
 - '이용목적 내용'은 학습과 관련된 내용으로 입력하고,
 - '구분자'는 '구분자_кома'를 선택한 후 '확인'을 클릭합니다. ([그림 1-37])
 - 그러면 몇 개의 확인창이 나오는데 모두 '확인'을 클릭합니다.
- ② 파일은 바로 다운로드가 되지 않고 요청한 자료들을 처리한 후에 다운로드가 가능해집니다. 소요되는 시간은 상황에 따라 다르지만 그리 오래 걸리지 않습니다(추출 상태 '실행중', [그림 1-38]).
- ③ 잠시 후 파일을 받을수 있게 되면(추출상태 '완료') [그림 1-39]와 같이 데이터파일, SAS와 SPSS의 파일 추출 구문을 받을 수 있는 상태가 됩니다. 여기서 파일 아이콘을() 눌러 다운로드 받습니다. ([그림 1-39])

외부로부터 자료 가져오기

데이터다운로드 실행

통계항목명	1%_인구사할(제공)
제목	2010년 인구사할
이용목적	교육자료
이용목적 내용	학업에 활용
구분자	<input type="radio"/> 고정길이 <input checked="" type="radio"/> 구분자_콤마 <input type="radio"/> 구분자_세미콜론 <input type="radio"/> 구분자_탭
공동연구자	

미리보기

확인

취소

• 60,000건 이상의 데이터는 Excel로 변환되지 않습니다.
 • 제공데이터는 비밀유지를 위하여 통계적 노출관리 기법이 적용될 수 있습니다.

[그림 1-37] 입력 후 '확인' 클릭

외부로부터 자료 가져오기

데이터 다운로드 및 이용현황 홈 > MY서비스 > 데이터 다운로드 및 이용현황1

추출·다운로드 주문형서비스 원격질문서비스 메타데이터서비스 연구결과공유서비스

추출형태: 전체 추출상태: 전체

자료명:

[검색](#)

Total 2 | Page 1 / 1 [excel](#) | [txt](#) | [sas](#) | [spss](#)

번호	추출형태	자료명(상세정보)	이용년도	데이터포맷	다운로드	추출상태	추출일
<input type="checkbox"/> 2	추출	인구주택총조사 > 1%-인구사항(제공)[20...		보기	파일 없음	실행중	2016-06-03
<input type="checkbox"/> 1	추출	인구주택총조사 > 무료_1%-인구사항(제...	2010	보기	파일 이미지 문서	완료	2016-02-02

[«](#)
[<](#)
[1](#)
[>](#)
[»](#)

[삭제](#)

[그림 1-38] 요청한 파일 추출중

외부로부터 자료 가져오기

데이터 다운로드 및 이용현황 홈 > MY서비스 > 데이터 다운로드 및 이용현황1

추출-다운로드 추출형 서비스 일괄접근서비스 이용센터서비스 연구결과공유서비스

추출형태: 전체 추출상태: 전체

자료명:

Total 2 | Page 1 / 1

☒ excel | ☐ txt | ☐ sas | ☐ spss

<input type="checkbox"/>	번호	추출형태	자료명(상세정보)	이용년도	데이터포함	다운로드	추출상태	추출일
<input type="checkbox"/>	2	추출	인구주택총조사 > 1%_인구사항(제공)[20...	2010	보기	   	완료	2016-06-03
<input type="checkbox"/>	1	추출	인구주택총조사 > 무로_1%_인구사항(제...	2010	보기	   	완료	2016-02-02

« < 1 > »

[그림 1-39] 요청한 파일 다운로드 가능

외부로부터 자료 가져오기

예제 1-4 파일 확인 및 파일명 변경

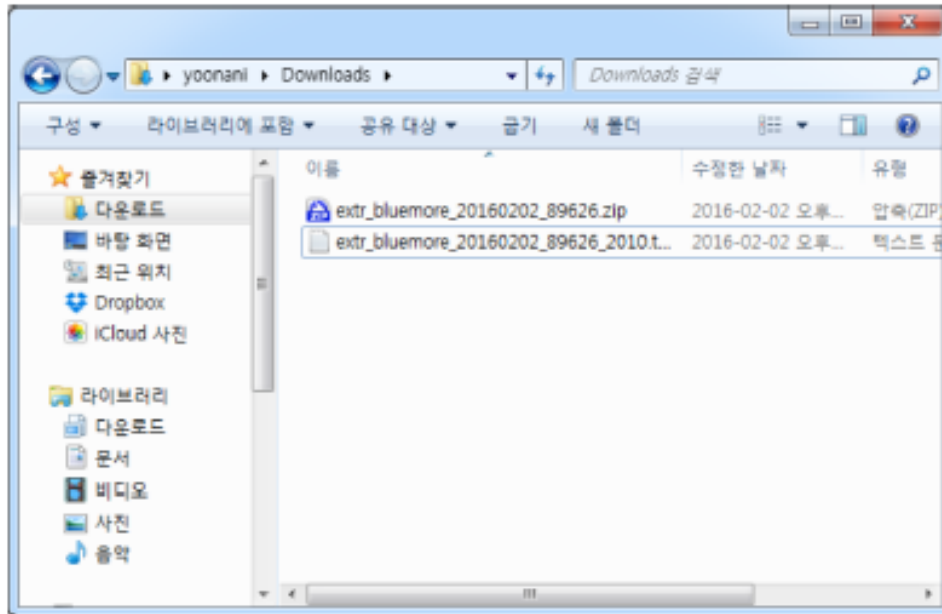
- **실습내용**

- 다운로드 받은 파일의 압축을 해제하고, 그 내용을 확인합니다. 그리고 원하는 이름으로 지정한 후 원하는 위치로 옮깁니다.

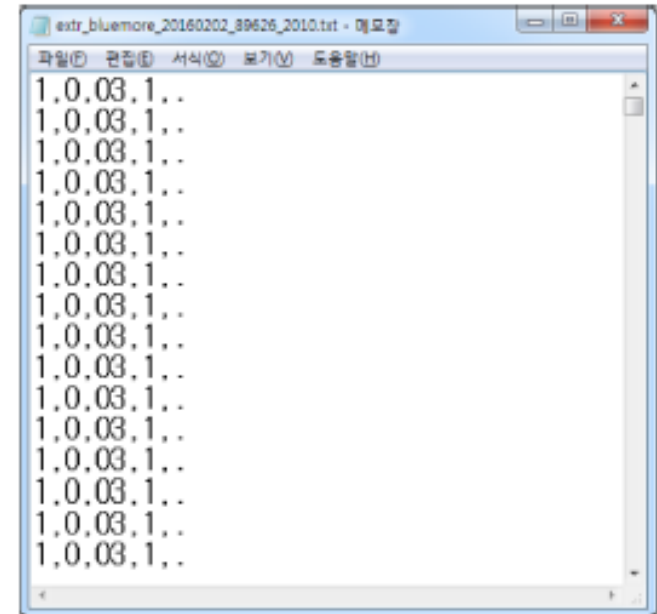
- **파일 확인**

- 다운로드 받은 파일은 압축파일입니다. ([그림 1-40])
- 압축을 풀었을 때 나타나는 파일은 일반 텍스트 파일로 '메모장' 같은 프로그램을 열어볼 수 있습니다.
- 메모장을 통해 파일을 열어보면 행 구분은 새 줄로, 열 구분은 콤마(,)로 되어 있음을 확인할 수 있습니다.
- 화면에 보이는 각 줄의 마지막을 보면 점(.)으로 끝나는데, 이 파일에서의 점(.)은 관측되지 않은 결측값을 나타냅니다. ([그림 1-41])

외부로부터 자료 가져오기



[그림 1-40] 압축 해제

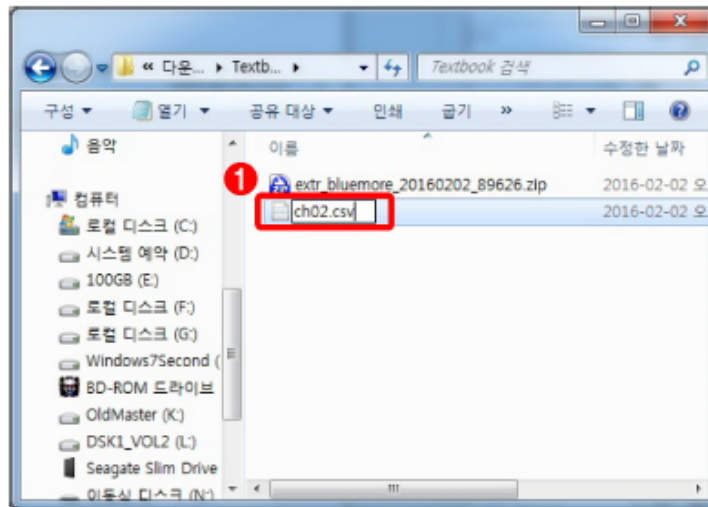


[그림 1-41] 다운로드 받은 파일 내용

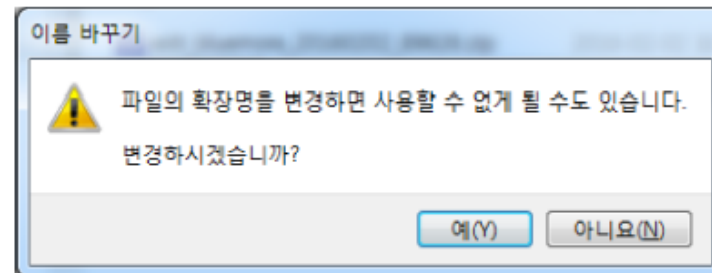
외부로부터 자료 가져오기

- 파일 이름과 확장명 변경하기
 - 콤마로 열을 구분한 파일을 csv(comma separated value) 파일이라 부르고 확장명으로 'csv'를 사용합니다.
 - 텍스트 파일 중에서도 콤마로 열이 구분된 csv 파일임을 이름만 보고도 알 수 있도록 하기 위해 다운로드 받은 파일(확장자는 txt)의 확장자를 csv로 변경해봅시다(확장명이 보이지 않을 경우 다음 실습 참고).
- ① 파일을 선택한 후, 키보드 상의 F2키를 누른 후 확장명을 'cho2.csv'로 변경하고, 엔터키를 누릅니다. ([그림 1-42])
 - 확장명을 변경하면 응용프로그램 연결이 변경되어 더블클릭해서 실행할 사용자의 예상과 다르게 실행될 수 있음을 경고합니다. ([그림 1-43])
- ② 확장자를 변경하고 나면 아이콘이 엑셀(Excel) 아이콘으로 변경됩니다(엑셀이 설치되어 있을 경우, [그림 1-44])
 - 엑셀이 설치될 때 확장자가 csv일 경우 자신과 연결되도록 하기 때문입니다.

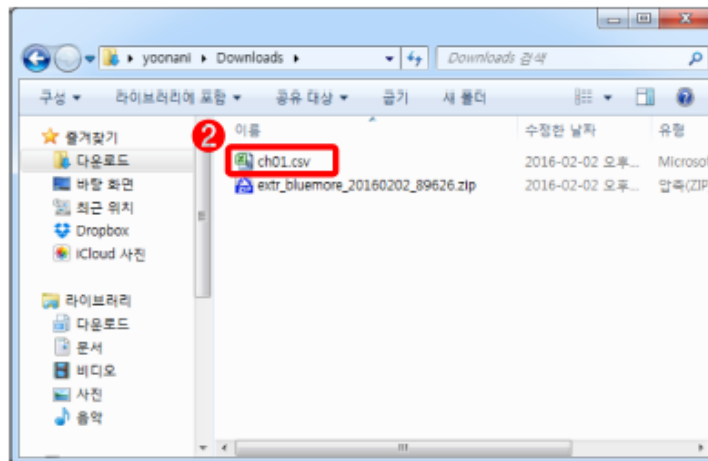
외부로부터 자료 가져오기



[그림 1-42] F2 키를 누르고 준비파일 변경

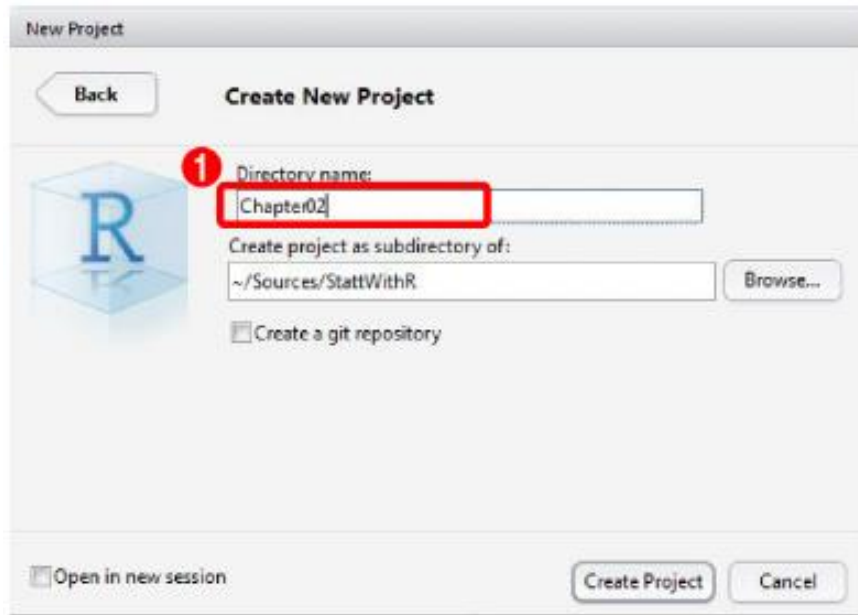


[그림 1-43] 확장명 변경 시 경고 메시지

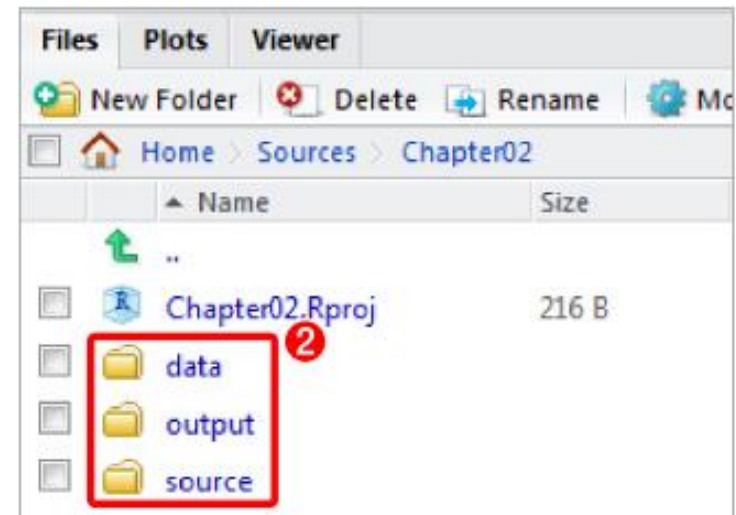


[그림 1-44] csv로 변경 시 엑셀 아이콘으로 변경

외부로부터 자료 가져오기

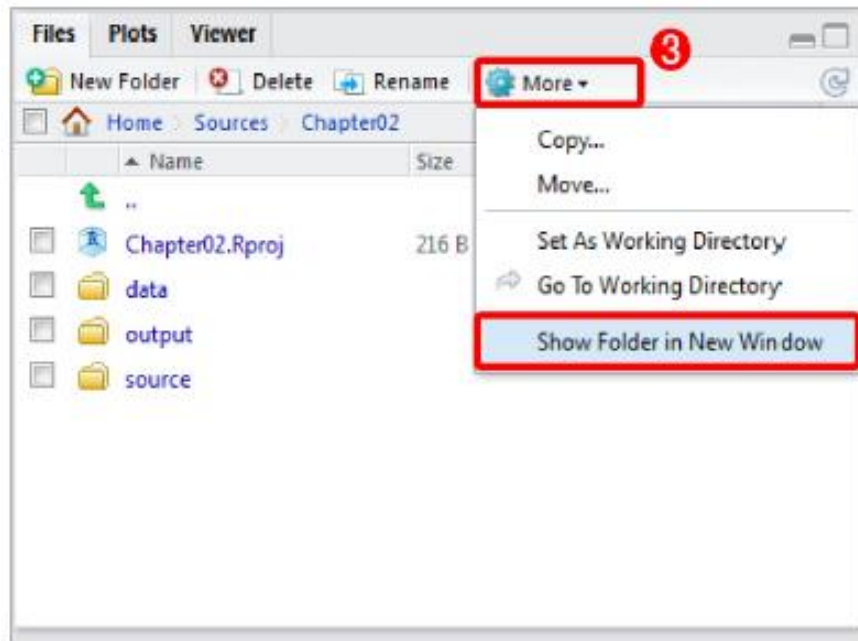


[그림 1-45] 'Chapter02' 프로젝트 생성

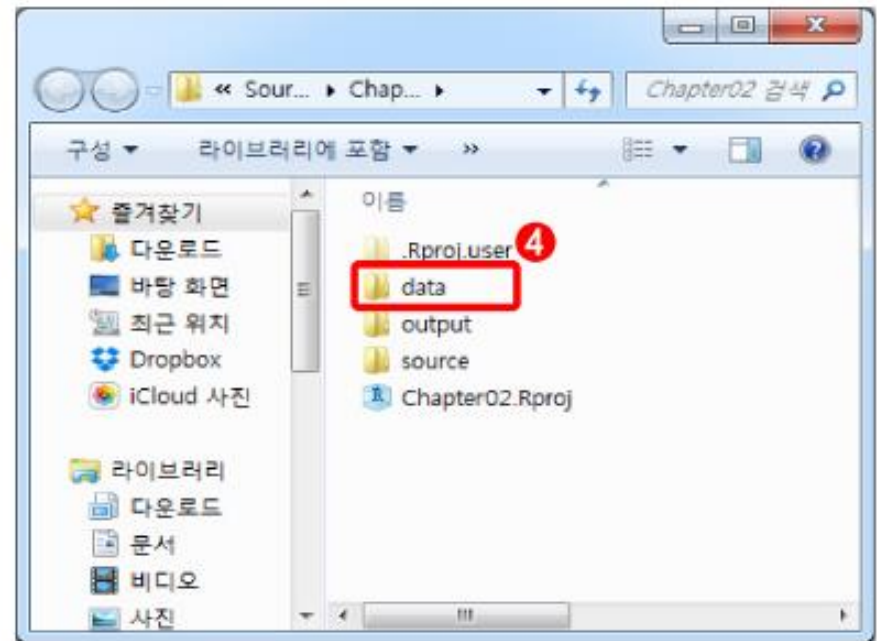


[그림 1-46] 각종 디렉토리 생성

외부로부터 자료 가져오기



[그림 1-47] 프로젝트를 탐색기에서 열기 위한 메뉴

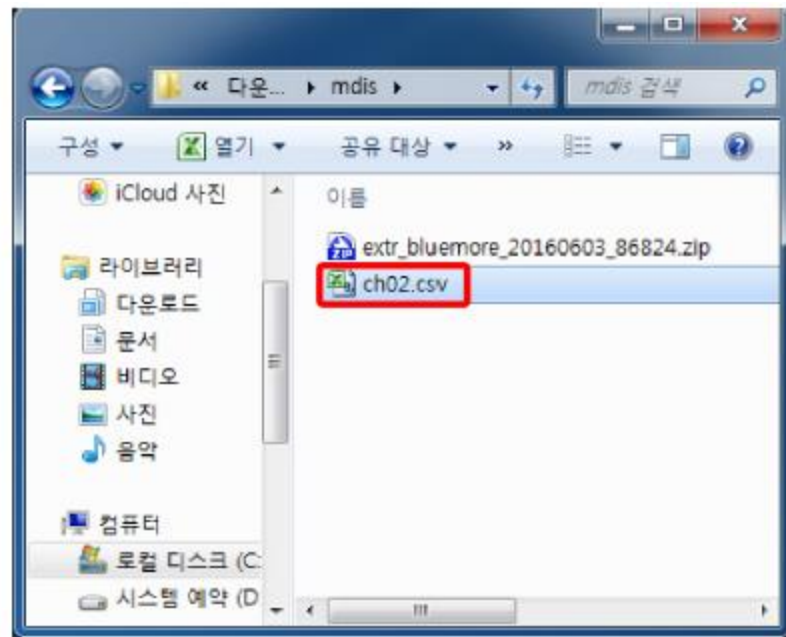


[그림 1-48] 탐색기에서 열린 프로젝트 폴더

외부로부터 자료 가져오기



(a) 'ch02.ssv'를 'data' 폴더로 이동



(b) 생성된 'ch02.csv'

[그림 1-49] 다운로드 받은 파일 이동

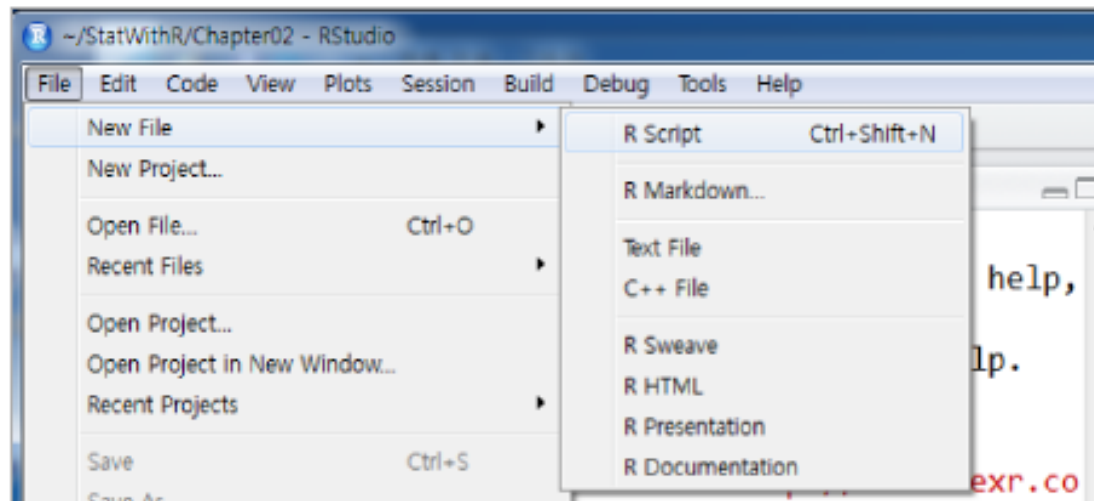
외부로부터 자료 가져오기

예제 1-5 R로 데이터 불러오기

준비파일 | 00.data_preparation.R

• 실습내용

- ‘data’ 폴더에 저장한 다운로드 받은 파일을 R로 불러오고, 분석에 사용할 수 있도록 값을 변경해봅시다.
- 새로운 R 스크립트 파일을 만듭니다.
 - File -> New File -> R Script



외부로부터 자료 가져오기

- ▣ 생성된 R 스크립트 창에 다음과 같이 입력하고 실행합니다.

```

1: data <- read.csv("./data/ch02.csv", header=F, na.strings=c("."))
2: str(data)
3: data$V1 <- factor(data$V1, levels=c(1, 2),
                     labels=c("남자", "여자"))
4: data$V3 <- factor(data$V3, levels=1:14,
                     labels=c("가구주", "가구주의 배우자", "자녀",
                               "자녀의 배우자", "가구주의 부모", 배우자의 부모",
                               "손자녀, 그 배우자", "증손자녀, 그 배우자",
                               "조부모", "형제자매, 그 배우자",
                               "형제자매의 자녀, 그 배우자", "부모의 형제자매, 그 배우자",
                               "기타 친인척", "그외같이사는사람") )
5: data$V4 <- factor(data$V4, levels=1:8,
                     labels=c("안 받았음", "초등학교", "중학교",
                               "고등학교", "대학-4년제 미만", "대학-4년제 이상",
                               "석사과정", "박사과정") )
6: str( data )
7: save.image("data.rda")

```

외부로부터 자료 가져오기

▣ Code 설명

- 1줄 : 프로젝트 내의 data 폴더에서 cho2.csv를 read.csv() 함수로 불러옵니다.
 - header=F : 자료의 첫줄부터 데이터로 읽어옵니다.
 - na.strings(c(".")) : 읽어온 데이터에서 점(.)은 결측(NA)로 처리합니다.
 - 읽어온 내용을 변수 data에 저장합니다.
- 2줄 : read.csv()로 자료를 읽으면 데이터 프레임으로 저장합니다.
 - 5개의 변수를 468,284명으로부터 관찰했음을 알려줍니다.
 - 각각의 변수는 R이 임의로 V1부터 V5까지로 하였으며 각 벡터는 정수형(int)로 되어 있음을 알 수 있습니다.

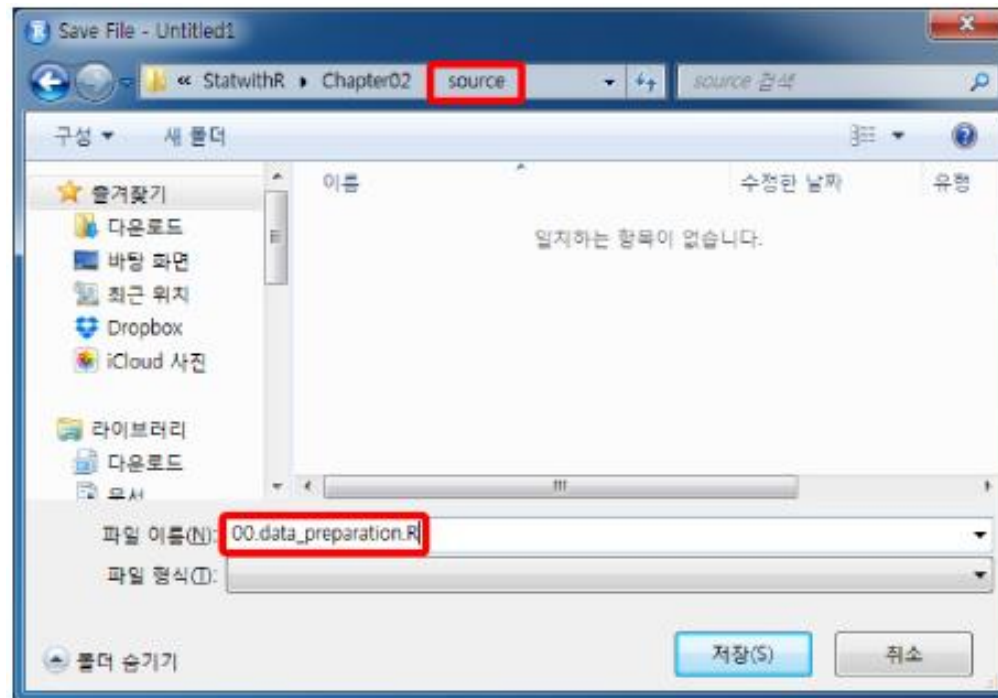
```
> str(data)
'data.frame':      468284 obs. of  5 variables:
 $ V1: int  1 1 1 1 1 1 1 1 1 1 ...
 $ V2: int  0 0 0 0 0 0 0 0 0 0 ...
 $ V3: int  3 3 3 3 3 3 3 3 3 3 ...
 $ V4: int  1 1 1 1 1 1 1 1 1 1 ...
 $ V5: int  NA NA NA NA NA NA NA NA NA ...
```

외부로부터 자료 가져오기

- ▣ 3~5줄 : 각 변수별로 저장된 값으로 'levels'를 지정하고, 각 값에 맞는 문자열을 'labels'에 지정한 후 'factor'로 만듭니다.
 - 3번째 줄에서 'levels=c(1, 2)'로 저장된 값은 1과 2이고, 각 값을 'labels=c("남자", "여자")'를 통해 1은 '남자'로, 2는 '여자'로 표시되는 'factor'를 만듭니다.
- ▣ 6줄 : str()을 통해 앞서 처음 읽어왔을 때와 factor로 변경했을 때 어떻게 달라지는지 확인해 봅시다.
- ▣ 7줄 : 코드에서 생성한 객체들을 'data.rda'로 저장합니다.
 - 본 교재에서는 한번만 사용하지만, 현재 작업환경에서 생성한 각종 R에서 사용하는 자료들을 저장하는 함수는 save.image("파일명") 입니다.
 - save.image() 로 저장한 내용은 load() 함수를 이용하여 언제든지 불러와서 사용할 수 있습니다.

외부로부터 자료 가져오기

- 위에서 작성한 R 코드를 'sources' 폴더에 'oo.data_preparation.R'로 저장합니다.
- R 스크립트 창 상단의 저장 아이콘, File메뉴의 Save, Ctrl+s 로 저장합니다.
- 저장할 폴더는 source 입니다. 앞으로 각 챕터에서 작성하는 R 스크립트는 source 폴더에 저장합니다.



외부로부터 자료 가져오기

예제 1-6 Windows에서 파일 확장명을 볼 수 없는 경우 해결법

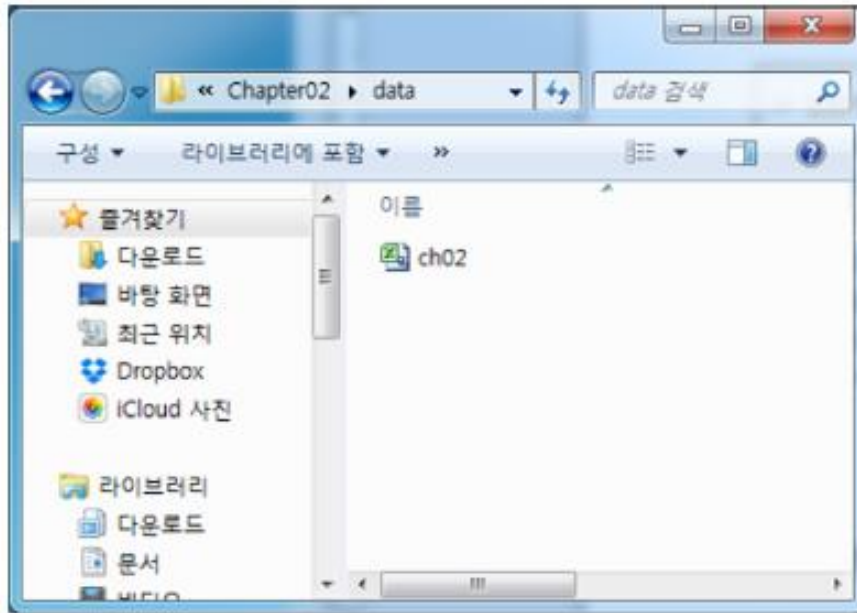
• 실습내용

- Windows는 기본적으로 알려진 확장명을 숨기게 되어 있는데, 이를 해제하는 방법을 알아보시다.
- Windows에서는 프로그램과 파일의 연결을 나타내는 아이콘으로 파일을 구별할 수 있지만 이는 파일에 대한 명쾌한 구분이 힘들 수 있습니다.
 - 동영상 파일은 .avi, .mp4, .mov 등으로 끝나지만 동영상 플레이어 프로그램이 자신의 아이콘으로 모두 덮어써서 어떤 파일인지 한눈에 알아보기 힘듭니다.
- 확장자는 파일의 종류를 나타내는 암묵적 동의하에 사용하므로, 확장자를 통해 명확히 해당 파일의 종류를 구분할 수 있어 편리한 경우도 많습니다.

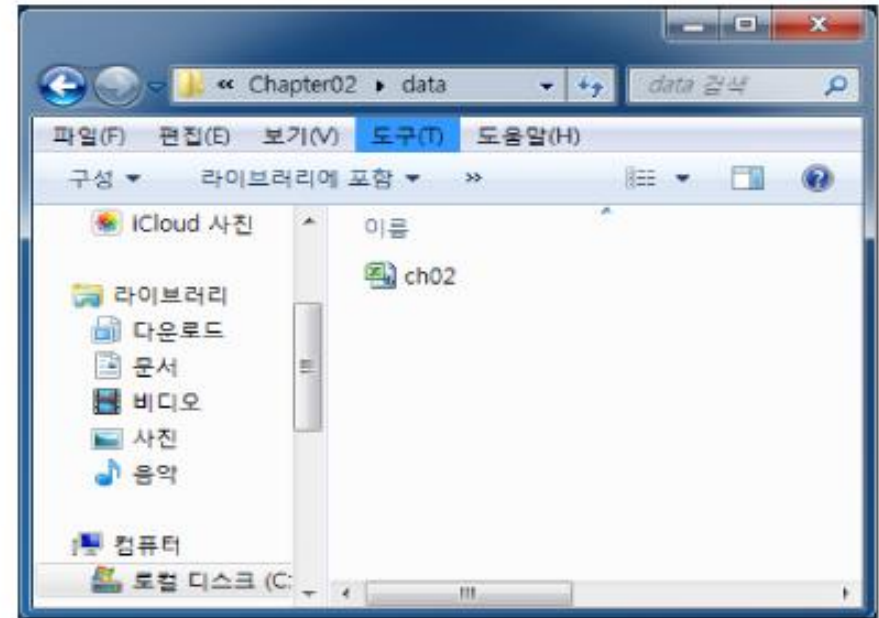
• 파일 확장자 표시하기

- 탐색기를 열어 [그림 1-52]처럼 확장명이 안 나올 경우에 해당합니다.
- 키보드의 Alt키를 누르면 [그림 1-53]과 같이 숨겨져 있던 메뉴가 나옵니다.

외부로부터 자료 가져오기



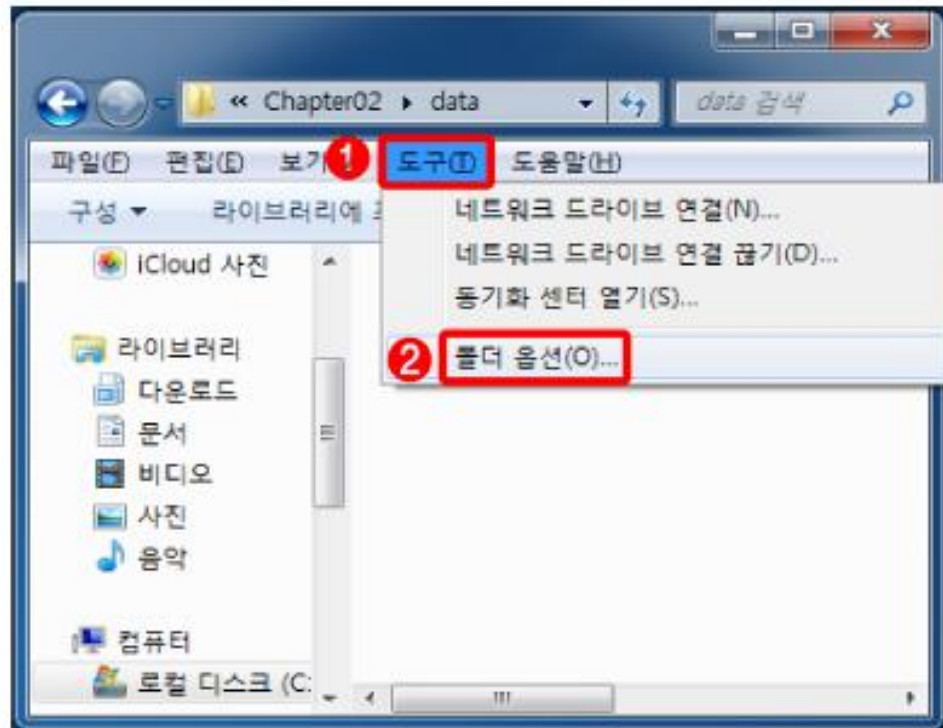
[그림 1-52] 확장명이 숨겨진 경우



[그림 1-53] 숨겨진 메뉴를 [Alt]로 열기

외부로부터 자료 가져오기

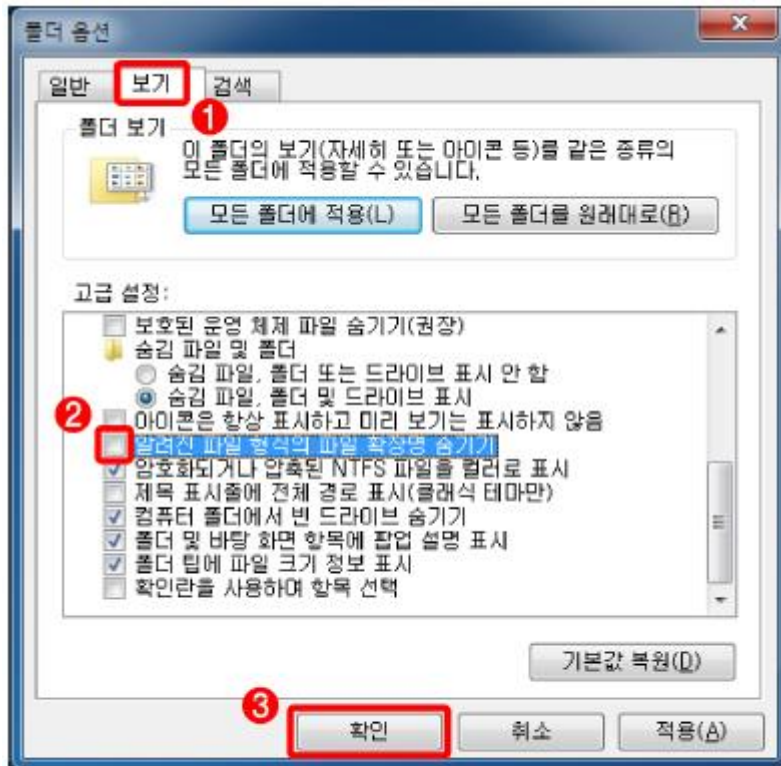
- 숨겨진 메뉴에서 ❶ '도구'를 클릭하고, 하단의 ❷ '폴더 옵션'을 클릭합니다.



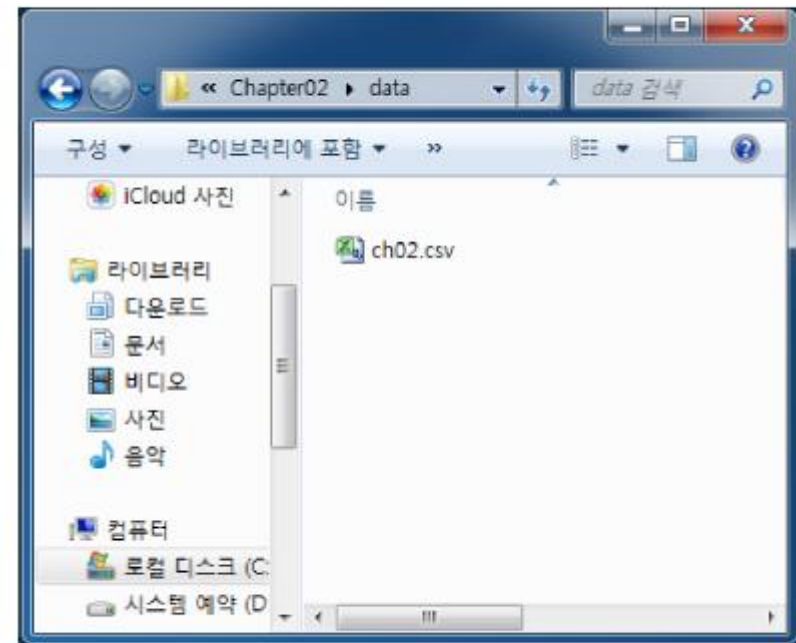
외부로부터 자료 가져오기

- ▣ 다음과 같이 진행합니다.
 - ① '폴더 옵션' 창이 열리면 '보기' 탭을 선택합니다.
 - ② '고급 설정'의 스크롤을 내려 '알려진파일 형식의 파일 확장명 숨기기'의 선택을 해제합니다.
 - ③ '확인'을 클릭하여 설정을 마칩니다.
- 이후 [그림 1-56]과 같이 탐색기에서 확장명이 나타나는 것을 확인할 수 있습니다.

외부로부터 자료 가져오기



[그림 1-55] 설정 변경



[그림 1-56] 숨겨진 확장명이 나타남



Q & A



수고하셨습니다.