



5주차: 데이터 가공

ChulSoo Park

School of Computer Engineering & Information Technology

Korea National University of Transportation



학습목표 (5주차)

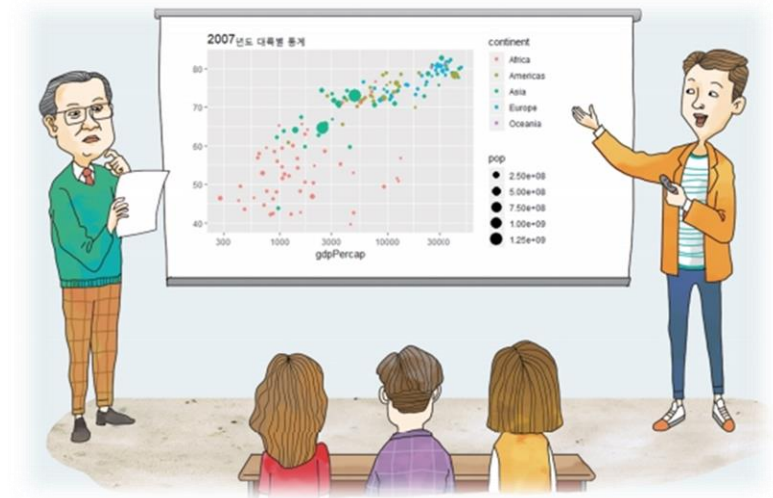
- ❖ 데이터 가공의 개념 이해
- ❖ 베이스 R을 이용한 데이터 가공
- ❖ dplyr 라이브러리를 이용한 데이터 가공
- ❖ 대량의 데이터 가공 실습
- ❖ 데이터 가공 실 사례 학습



05

CHAPTER

데이터 가공



CONTENTS

- 5.1 데이터 가공이란?
- 5.2 베이스 R을 이용한 데이터 가공
- 5.3 dplyr 라이브러리를 이용한 데이터 가공
- 5.4 대량의 데이터 가공
- 5.5 데이터 가공 사례 학습
 - 요약

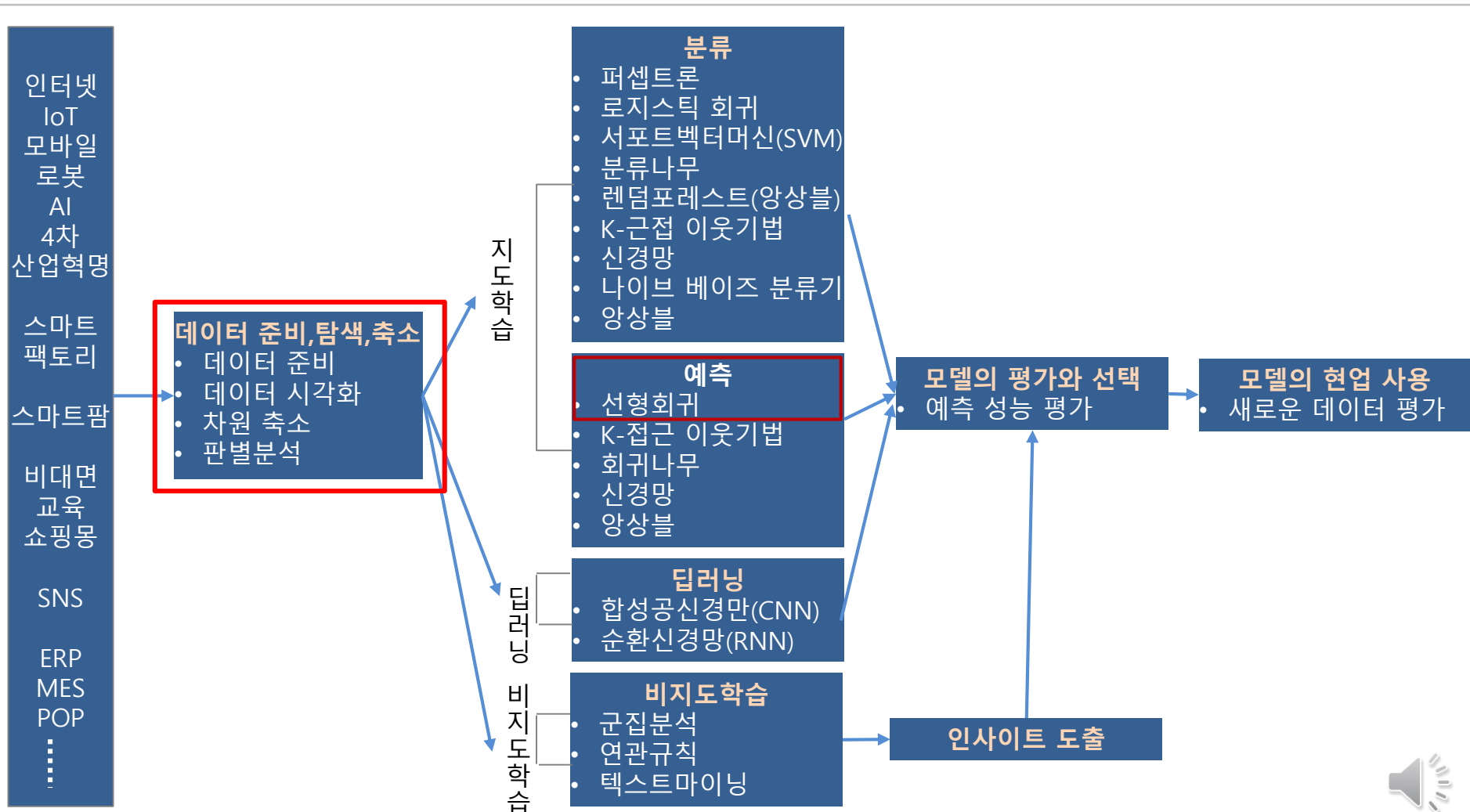


■ 4장에서는 R의 데이터 형과 연산

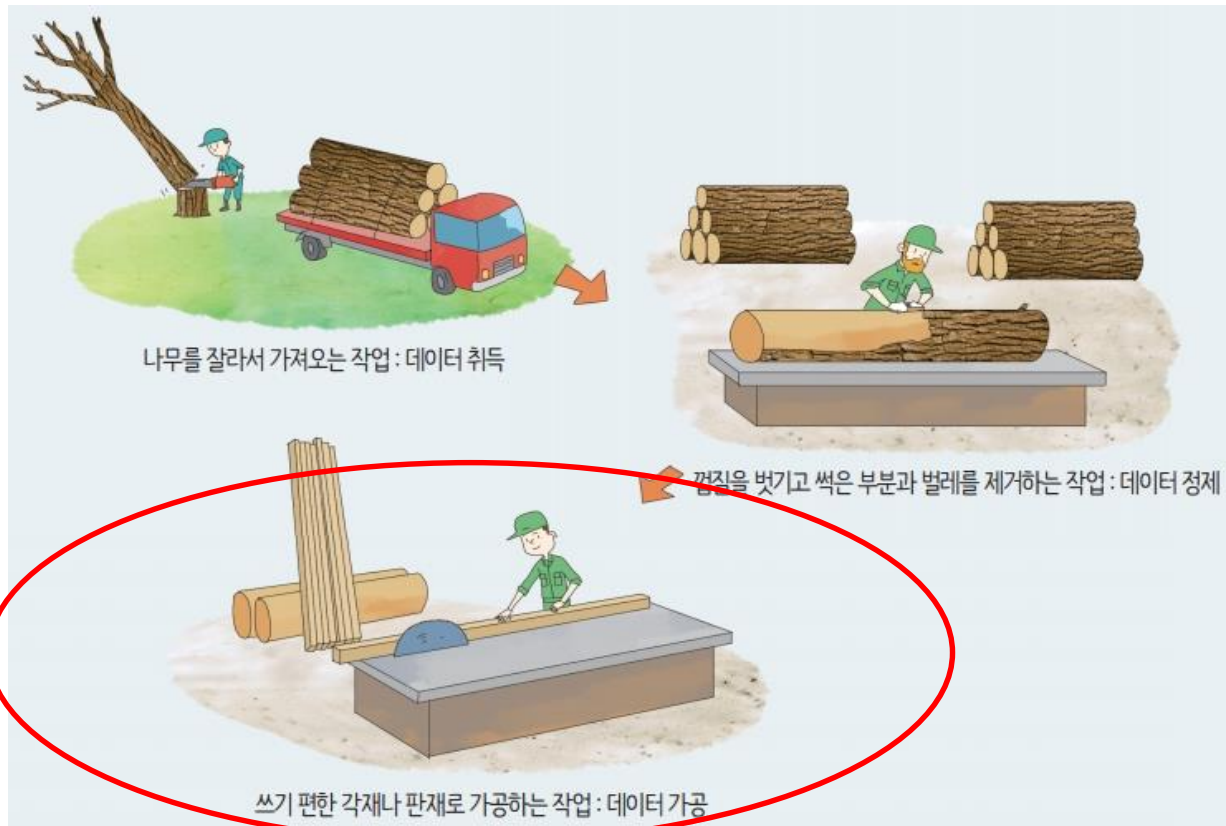
- 파일 읽고 쓰기
- 데이터 정제를 위한 조건문과 반복문
- 사용자 정의 함수
- 데이터 정제 예제(결측값 처리)
- 데이터 정제 예제(이상값 처리)



■ 데이터 분석 Process에서 이번주 교육 위치



- 잘 정제되고 다듬어진 데이터는 큰 가치가 있지만, 정리되지 않은 데이터는 의미 추출이 어려울 뿐 아니라 잘못된 결론에 이르게 할 수도 있음



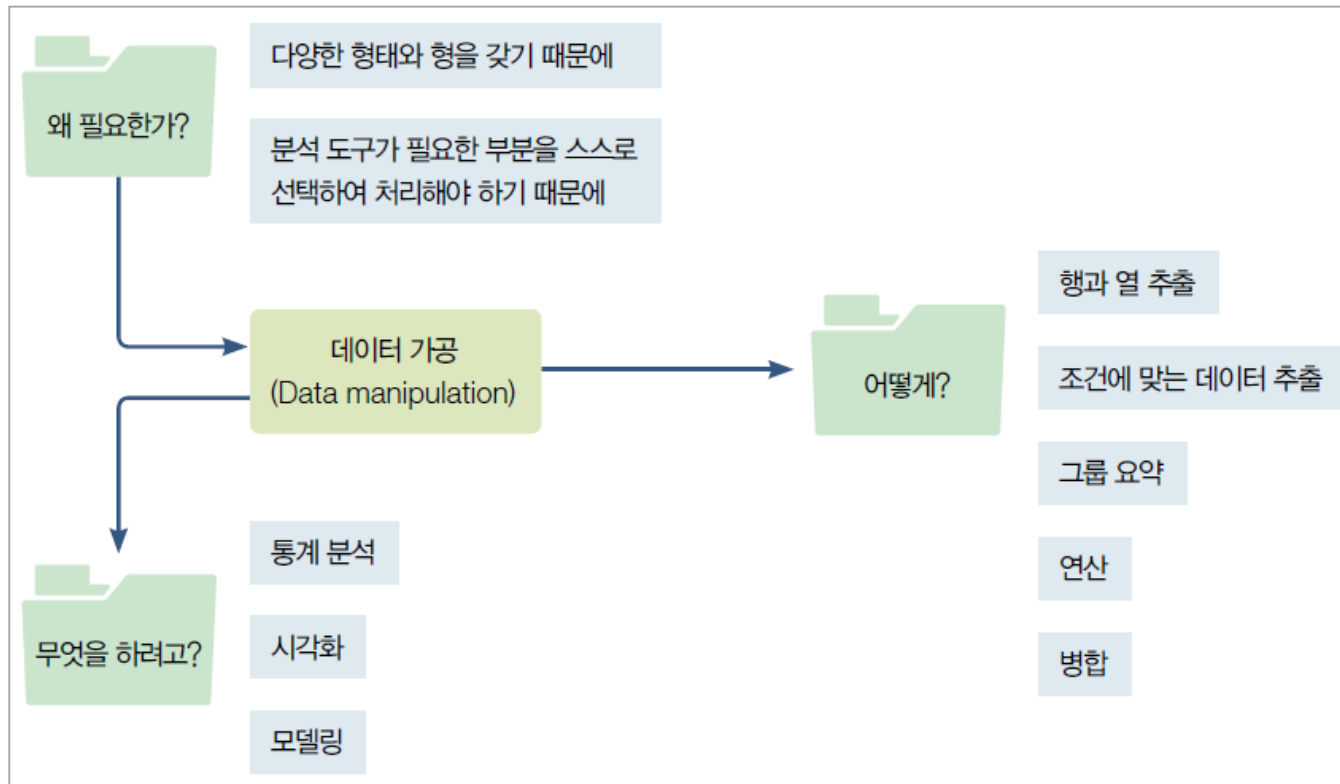
5.1 데이터 가공(data manipulation)이란?

- 광범위하고 구체적인 목적을 가지고 이루어진다는 점에서 불필요한 요소 제거, 사용하기 편하게 정리하는 정제와 구별됨.
- 거의 모든 분야에서 적절한 데이터 가공이 필요
 - 데이터에 담긴 의미를 끄집어내기 위한 **통계 분석**
 - 효과적인 관찰을 위한 **시각화**
 - 인과관계를 추정하기 위한 **모델링** 등



5.1 데이터 가공(data manipulation)이란?

- 데이터를 보다 효과적으로 분석하기 위해 데이터를 만지고 변형하는 작업이 바로 데이터 가공(data wrangling)
- 디지털화된 데이터와 R 같은 분석 도구를 활용해 쉽고 빠르게 데이터 가공 작업을 수행할 수 있음



5.1 데이터 가공(data manipulation)이란?



5.1 데이터 가공(data manipulation)이란?

■ Data 처리 과정별 기술 영역

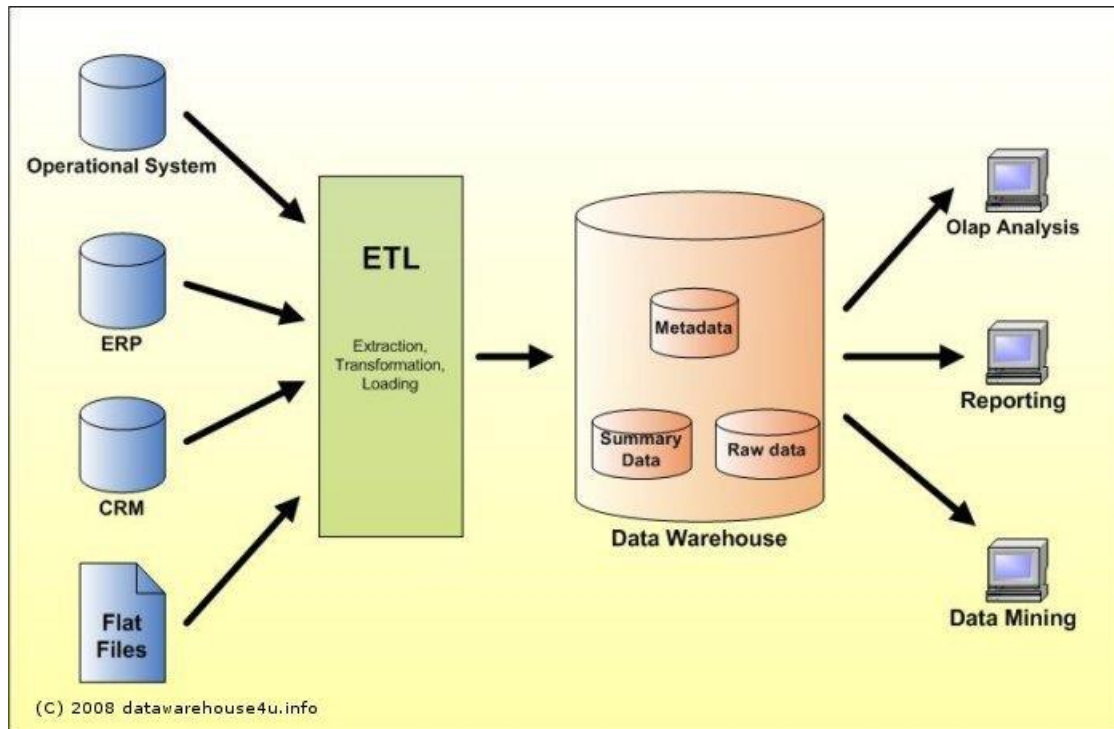
과정	영역	설명
생성	내부데이터	데이터베이스, 파일 관리 시스템 등
	외부데이터	인터넷으로 연결된 파일, 멀티 미디어, 스트림
수집	크롤링	검색 엔진의 로봇, HTML 크롤링 소프트웨어를 사용한 데이터 수집
	ETL (Extraction, Transformation, Loading)	소스 데이터의 추출, 전송, 변환, 적재
저장	NOSQL 데이터 베이스	비정형데이터 관리
	스토리지 Storage	빅데이터 저장, 저장소
	서버 Server	초경량 서버
처리	맵리듀스 MapReduce	데이터 추출
	프로세싱 Processing	다중 업무 처리
분석	NLP Natural Language Processing	자연어 처리
	기계학습 Machine Learning	머신러닝, 딥러닝을 이용한 데이터의 패턴 인식
	직렬화 Serialization	데이터 간의 순서화
표현	시각화 Visualization	데이터를 도표나 그래픽 등으로 표현
	획득 Acquisition	데이터의 획득, 재해석



5.1 데이터 가공(data manipulation)이란?

■ ETL(Extraction, Transformation, Loading) 기능

- **Extraction(추출)** : 다양한 데이터 원천(Source)에서 부터 데이터 획득
- **Transformation(변형)** : 데이터 클린징, 형식 변환, 표준화, 통합 또는 다수 애플리케이션에 내장된 비즈니스 룰 적용
- **Loading(적재)** : 변형 단계 완료 후 특정 목표 시스템에 적재



5.2 베이스 R을 이용한 데이터 가공

■ gapminder 라이브러리

- 세계 각국의 기대 수명 1인당 국내총생산 인구 데이터 등을 집계한 gapminder 데이터 셋의 일부를 담고 있음
- 이 데이터는 R을 배우고 통계학을 연습하는데 매우 유용한 기초 자료이고 여러 유형의 데이터가 데이터 프레임 형식으로 저장되어 있어 데이터 과학을 학습하는 우리들에게 매우 좋은 자료이다.

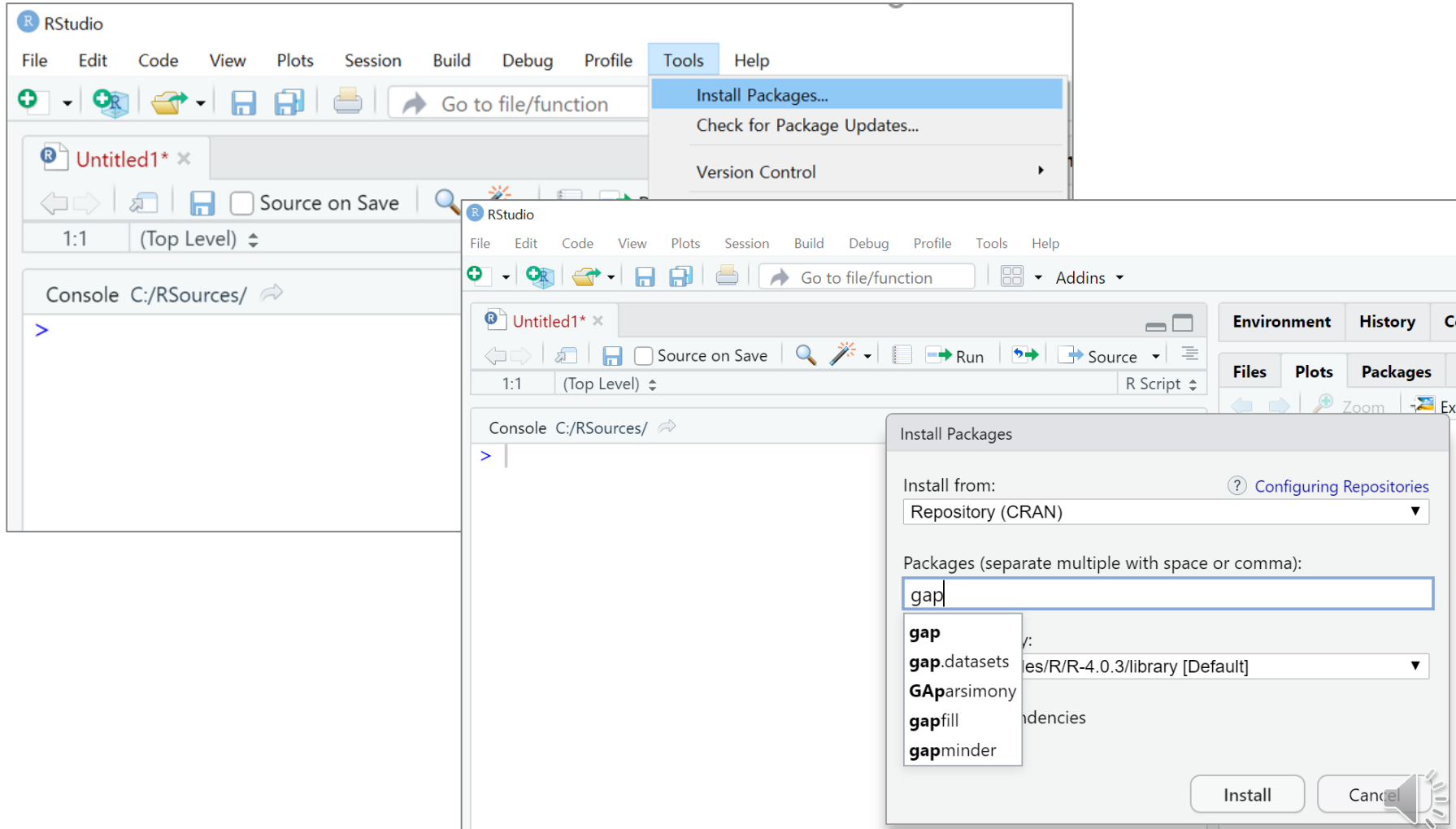
표 5-1 gapminder 데이터 프레임의 구성 항목

열이름(변수명)	변수형	내용
country	142개 레벨의 범주형	국가명
continent	5개 레벨의 범주형	국가가 속한 대륙
year	int	1952~2007 관측 연도(5년 단위)
lifeExp	num	기대 수명(평균 수명)
pop	int	인구
gdpPercap	num	1인당 국내총생산(물가 상승 반영)



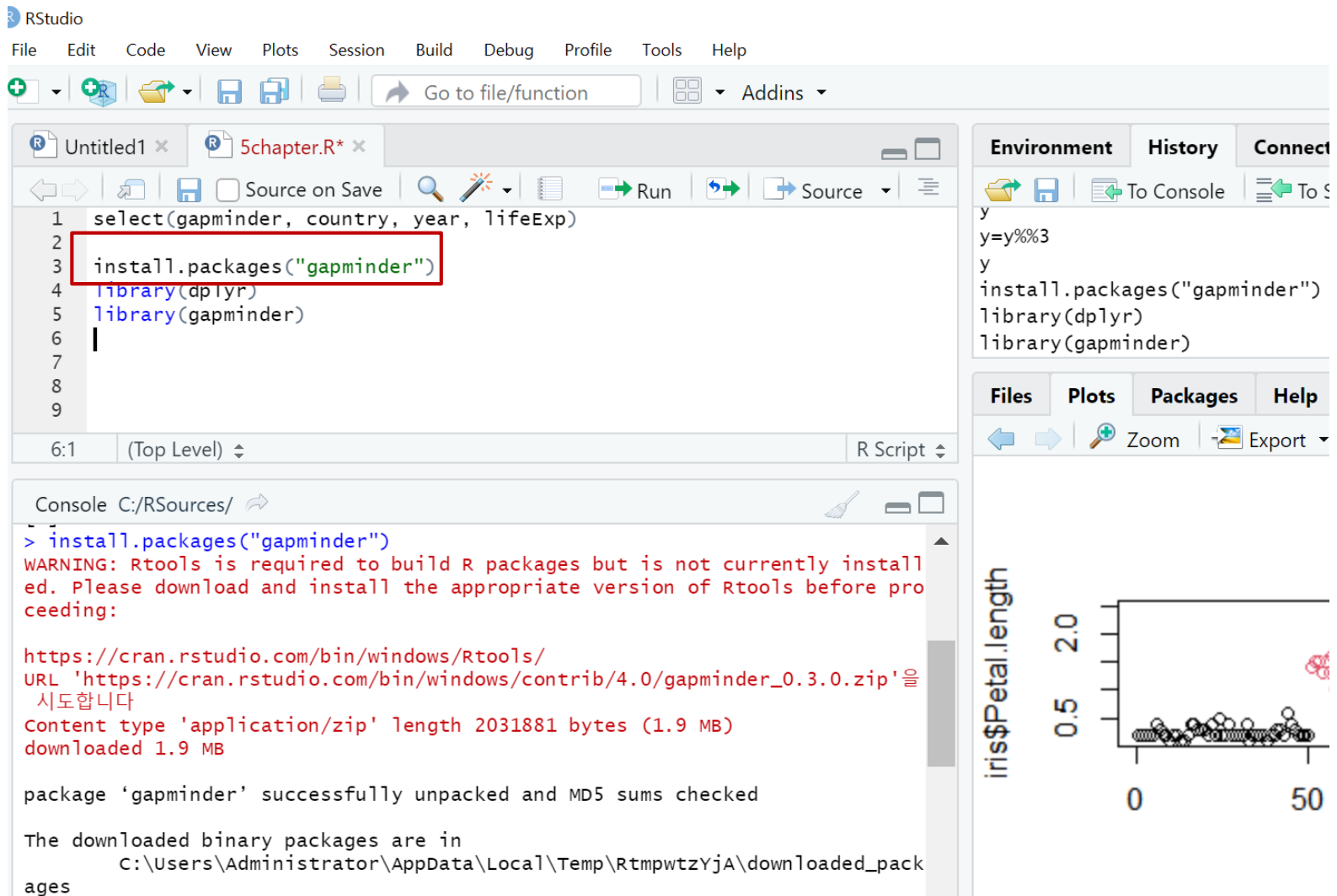
5.2 베이스 R을 이용한 데이터 가공

■ gapminder 라이브러리 install



5.2 베이스 R을 이용한 데이터 가공

■ gapminder 라이브러리 install (2가지 방법)



RStudio interface showing the installation of the `gapminder` package.

Code Editor:

```
1 select(gapminder, country, year, lifeExp)
2
3 install.packages("gapminder")
4 library(dplyr)
5 library(gapminder)
6
7
8
9
```

Console:

```
> install.packages("gapminder")
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/
URL 'https://cran.rstudio.com/bin/windows/contrib/4.0/gapminder_0.3.0.zip'을 시도합니다
Content type 'application/zip' length 2031881 bytes (1.9 MB)
downloaded 1.9 MB

package 'gapminder' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
C:\Users\Administrator\AppData\Local\Temp\RtmpwtzYjA\downloaded_packages
```

Environment:

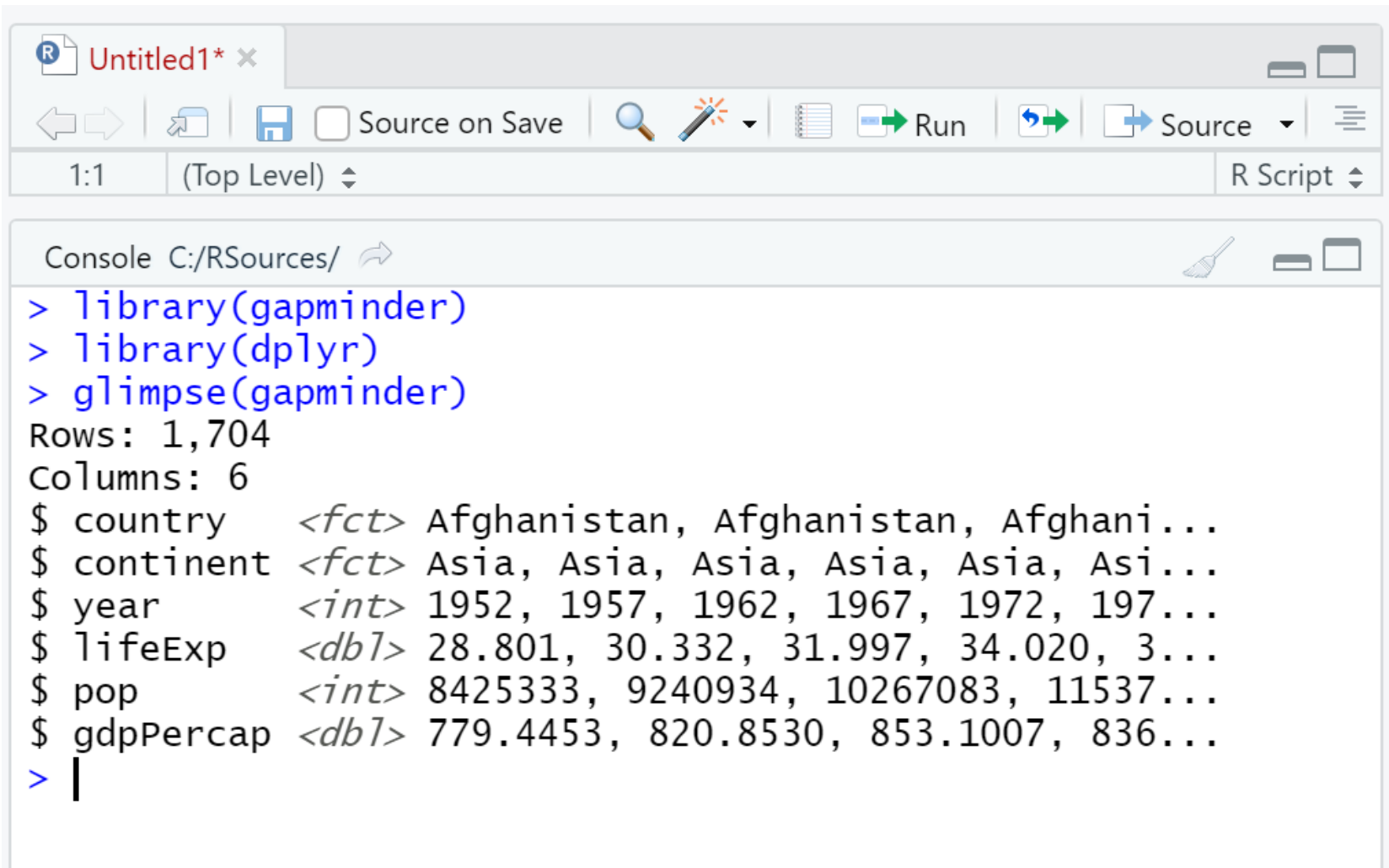
```
y
y=y%%3
y
install.packages("gapminder")
library(dplyr)
library(gapminder)
```

Plots:

iris\$Petal.length vs iris\$Sepal.length

5.2 베이스 R을 이용한 데이터 가공

■ gapminder 라이브러리



The screenshot shows an RStudio interface with a script editor and a console. The script editor has a tab titled 'Untitled1*' and contains the following R code:

```
> library(gapminder)
> library(dplyr)
> glimpse(gapminder)
```

The console output shows the result of the `glimpse(gapminder)` command, displaying the structure of the `gapminder` dataset:

```
Rows: 1,704
Columns: 6
$ country    <fct> Afghanistan, Afghanistan, Afghani...
$ continent  <fct> Asia, Asia, Asia, Asia, Asia, Asi...
$ year       <int> 1952, 1957, 1962, 1967, 1972, 197...
$ lifeExp    <dbl> 28.801, 30.332, 31.997, 34.020, 3...
$ pop        <int> 8425333, 9240934, 10267083, 11537...
$ gdpPercap  <dbl> 779.4453, 820.8530, 853.1007, 836...
```

The console window title is 'Console C:/RSources/'.

5.2 베이스 R을 이용한 데이터 가공

■ gapminder 라이브러리

Console C:/RSources/

```
> library(gapminder)
> library(dplyr)
> glimpse(gapminder)
```

Rows: 1,704

Columns: 6

\$ country <fct> Afghanistan, Afghanistan, Afghani...

\$ continent <fct> Asia, Asia, Asia, Asia, Asia, Asi...

\$ year <int> 1952, 1957, 1962, 1967, 1972, 197...

\$ lifeExp <dbl> 28.801, 30.332, 31.997, 34.020, 3...

\$ pop <int> 842

\$ gdpPercap <dbl> 779

```
> ?glimpse
```

```
> |
```

glimpse ▾

Find in Topic

Help on topic 'glimpse' was found in the following packages:

[Objects exported from other packages](#)

(in package [dplyr](#) in library C:/Program Files/R/R-4.0.3/library)

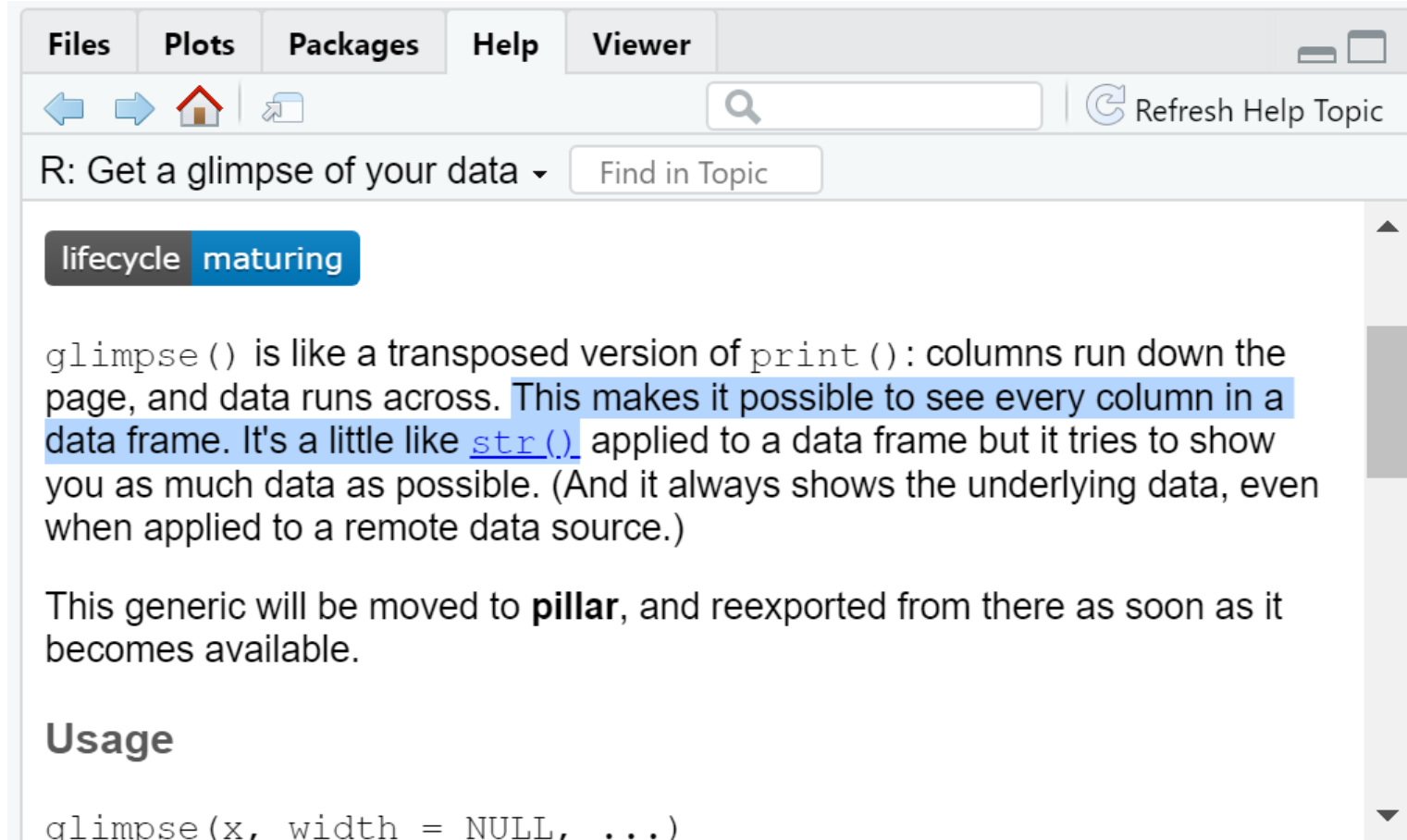
[Get a glimpse of your data](#)

(in package [tibble](#) in library C:/Program Files/R/R-4.0.3/library)



5.2 베이스 R을 이용한 데이터 가공

- gapminder 라이브러리
 - `>?glimpse`



The screenshot shows the R help interface for the `glimpse` function. The top navigation bar includes tabs for Files, Plots, Packages, Help, and Viewer. Below the navigation bar is a search bar and a 'Refresh Help Topic' button. The main content area displays the title 'R: Get a glimpse of your data' with a 'Find in Topic' search box. Two tabs, 'lifecycle' and 'maturing', are visible. The text describes `glimpse()` as a transposed version of `print()` for data frames, highlighting that it allows viewing every column. It also mentions that the function will be moved to the **pillar** package. The 'Usage' section shows the function signature: `glimpse(x, width = NULL, ...)`.

Files Plots Packages Help Viewer

← → Home ↗ Search Refresh Help Topic

R: Get a glimpse of your data ▾ Find in Topic

lifecycle maturing

`glimpse()` is like a transposed version of `print()`: columns run down the page, and data runs across. This makes it possible to see every column in a data frame. It's a little like `str()` applied to a data frame but it tries to show you as much data as possible. (And it always shows the underlying data, even when applied to a remote data source.)

This generic will be moved to **pillar**, and reexported from there as soon as it becomes available.

Usage

```
glimpse(x, width = NULL, ...)
```



5.2 베이스 R을 이용한 데이터 가공

■ gapminder 라이브러리

```
Console C:/RSources/
> str(gapminder)
tibble [1,704 x 6] (S3: tbl_df/tbl/data.frame)
 $ country   : Factor w/ 142 levels "Afghanistan",...: 1 1 1
 1 1 1 1 1 1 1 ...
 $ continent : Factor w/ 5 levels "Africa","Americas",...: 3
 3 3 3 3 3 3 3 ...
 $ year      : int [1:1704] 1952 1957 1962 1967 1972 1977 19
82 1987 1992 1997 ...
 $ lifeExp   : num [1:1704] 28.8 30.3 32 34 36.1 ...
 $ pop       : int [1:1704] 8425333 9240934 10267083 1153796
6 13079460 14880372 12881816 13867957 16317921 22227415 ...
 $ gdpPercap: num [1:1704] 779 821 853 836 740 ...
> head(gapminder,6)
# A tibble: 6 x 6
  country      continent  year lifeExp      pop gdpPercap
  <fct>        <fct>    <int>   <dbl>    <int>    <dbl>
1 Afghanistan Asia      1952    28.8  8425333    779.
2 Afghanistan Asia      1957    30.3  9240934    821.
3 Afghanistan Asia      1962    32.0 10267083    853.
4 Afghanistan Asia      1967    34.0 11537966    836.
5 Afghanistan Asia      1972    36.1 13079460    740.
6 Afghanistan Asia      1977    38.4 14880372    786.
```



5.2 베이스 R을 이용한 데이터 가공

■ 샘플과 속성의 추출

- 각 나라(country)의 기대 수명(lifeExp) 및 년도(year)

Console C:/RSources/ ↗

```
> gapminder[, c("country", "lifeExp")]  
# A tibble: 1,704 x 2  
  country    lifeExp  
  <fct>      <dbl>  
1 Afghanistan 28.8  
2 Afghanistan 30.3  
3 Afghanistan 32.0  
4 Afghanistan 34.0  
5 Afghanistan 36.1  
6 Afghanistan 38.4  
7 Afghanistan 39.9  
8 Afghanistan 40.8  
9 Afghanistan 41.7  
10 Afghanistan 41.8  
# ... with 1,694 more rows  
> |
```

Console C:/RSources/ ↗




```
> gapminder[, c("country", "lifeExp", "year")]  
# A tibble: 1,704 x 3  
  country    lifeExp  year  
  <fct>      <dbl> <int>  
1 Afghanistan 28.8  1952  
2 Afghanistan 30.3  1957  
3 Afghanistan 32.0  1962  
4 Afghanistan 34.0  1967  
5 Afghanistan 36.1  1972  
6 Afghanistan 38.4  1977  
7 Afghanistan 39.9  1982  
8 Afghanistan 40.8  1987  
9 Afghanistan 41.7  1992  
10 Afghanistan 41.8  1997  
# ... with 1,694 more rows  
> |
```



5.2 베이스 R을 이용한 데이터 가공

■ 샘플과 속성의 추출

- Head 함수 활용


```
Console C:/Rsources/     
> head(gapminder, 15)  
# A tibble: 15 x 6  
  country      continent year lifeExp      pop gdpPercap  
  <fct>        <fct>    <int>   <dbl>   <int>   <dbl>  
1 Afghanistan Asia      1952    28.8  8.43e6    779.  
2 Afghanistan Asia      1957    30.3  9.24e6    821.  
3 Afghanistan Asia      1962    32.0  1.03e7    853.  
4 Afghanistan Asia      1967    34.0  1.15e7    836.  
5 Afghanistan Asia      1972    36.1  1.31e7    740.  
6 Afghanistan Asia      1977    38.4  1.49e7    786.  
7 Afghanistan Asia      1982    39.9  1.29e7    978.  
8 Afghanistan Asia      1987    40.8  1.39e7    852.  
9 Afghanistan Asia      1992    41.7  1.63e7    649.  
10 Afghanistan Asia      1997    41.8  2.22e7    635.  
11 Afghanistan Asia      2002    42.1  2.53e7    727.  
12 Afghanistan Asia      2007    43.8  3.19e7    975.  
13 Albania     Europe    1952    55.2  1.28e6   1601.  
14 Albania     Europe    1957    59.3  1.48e6   1942.  
15 Albania     Europe    1962    64.8  1.73e6   2313.  
> |
```



5.2 베이스 R을 이용한 데이터 가공

■ 샘플과 속성의 추출

- 국가 이름이 "Croatia"인 샘플을 조건식을 사용해 추출





```
Console C:/RSources/   
> gapminder[gapminder$country=="Croatia", ]  
# A tibble: 12 x 6  
  country continent year lifeExp      pop gdpPercap  
  <fct>    <fct>    <int>   <dbl>   <int>    <dbl>  
1 Croatia Europe    1952    61.2 3882229    3119.  
2 Croatia Europe    1957    64.8 3991242    4338.  
3 Croatia Europe    1962    67.1 4076557    5478.  
4 Croatia Europe    1967    68.5 4174366    6960.  
5 Croatia Europe    1972    69.6 4225310    9164.  
6 Croatia Europe    1977    70.6 4318673   11305.  
7 Croatia Europe    1982    70.5 4413368   13222.  
8 Croatia Europe    1987    71.5 4484310   13823.  
9 Croatia Europe    1992    72.5 4494013    8448.  
10 Croatia Europe    1997    73.7 4444595    9876.  
11 Croatia Europe    2002    74.9 4481020   11628.  
12 Croatia Europe    2007    75.7 4493312   14619.  
> |
```



5.2 베이스 R을 이용한 데이터 가공

■ 샘플과 속성의 추출




- 국가 이름이 "Korea", "Japan"인 샘플을 조건식을 사용해 추출

```
Console C:/RSources/      
> gapminder[gapminder$country=="Korea", ]  
# A tibble: 0 x 6  
# ... with 6 variables: country <fct>, continent <fct>,  
#   year <int>, lifeExp <dbl>, pop <int>,  
#   gdpPercap <dbl>  
> gapminder[gapminder$country=="Japan", ]  
# A tibble: 12 x 6  
   country continent  year lifeExp      pop gdpPercap  
   <fct>    <fct>    <int>   <dbl>    <int>    <dbl>  
1 Japan    Asia      1952    63.0  86459025    3217.  
2 Japan    Asia      1957    65.5  91563009    4318.  
3 Japan    Asia      1962    68.7  95831757    6577.  
4 Japan    Asia      1967    71.4 100825279    9848.  
5 Japan    Asia      1972    73.4 107188273   14779.  
6 Japan    Asia      1977    75.4 113872473   16610.  
7 Japan    Asia      1982    77.1 118454974   19384.  
8 Japan    Asia      1987    78.7 122091325   22376.  
9 Japan    Asia      1992    79.4 124329269   26825.  
10 Japan   Asia      1997    80.7 125956499   28817.  
11 Japan   Asia      2002    82   127065841   28605.  
12 Japan   Asia      2007    82.6 127467972   31656.  
> |
```



■ 샘플과 속성의 추출

- Write.table을 활용하여 .txt 파일 생성

```
Console C:/Rsources/     
> write.table(gapminder, file="c:/rdata/gapminder.txt", quote=  
F)  
> |
```

gapminder - Windows 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

```
837 Korea, Dem. Rep. Asia 1992 69.978 20711375 3726.063507  
838 Korea, Dem. Rep. Asia 1997 67.727 21585105 1690.756814  
839 Korea, Dem. Rep. Asia 2002 66.662 22215365 1646.758151  
840 Korea, Dem. Rep. Asia 2007 67.297 23301725 1593.06548  
841 Korea, Rep. Asia 1952 47.453 20947571 1030.592226  
842 Korea, Rep. Asia 1957 52.681 22611552 1487.593537  
843 Korea, Rep. Asia 1962 55.292 26420307 1536.344387  
844 Korea, Rep. Asia 1967 57.716 30131000 2029.228142  
845 Korea, Rep. Asia 1972 62.612 33505000 3030.87665  
846 Korea, Rep. Asia 1977 64.766 36436000 4657.22102  
847 Korea, Rep. Asia 1982 67.123 39326000 5622.942464  
848 Korea, Rep. Asia 1987 69.81 41622000 8533.088805  
849 Korea, Rep. Asia 1992 72.244 43805450 12104.27872  
850 Korea, Rep. Asia 1997 74.647 46173816 15993.52796  
851 Korea, Rep. Asia 2002 77.045 47969150 19233.98818  
852 Korea, Rep. Asia 2007 78.623 49044790 23348.13973  
853 Kuwait Asia 1952 55.565 160000 108382.3529  
854 Kuwait Asia 1957 58.033 212846 113523.1329
```



5.2 베이스 R을 이용한 데이터 가공

■ 샘플과 속성의 추출

- 조건을 활용한 샘플 추출(대한민국, 일본)

```
Console C:/Rsources/
> gapminder[gapminder$country=="Korea, Rep.",]
# A tibble: 12 x 6
  country      continent year lifeExp      pop gdpPercap
  <fct>      <fct>    <int>   <dbl>   <int>   <dbl>
1 Korea, Rep. Asia      1952    47.5  2.09e7    1031.
2 Korea, Rep. Asia      1957    52.7  2.26e7    1488.
3 Korea, Rep. Asia      1962    55.3  2.64e7    1536.
4 Korea, Rep. Asia      1967    57.7  3.01e7    2029.
5 Korea, Rep. Asia      1972    62.6  3.35e7    3031.
6 Korea, Rep. Asia      1977    64.8  3.64e7    4657.
7 Korea, Rep. Asia      1982    67.1  3.93e7    5623.
8 Korea, Rep. Asia      1987    69.8  4.16e7    8533.
9 Korea, Rep. Asia      1992    72.2  4.38e7   12104.
10 Korea, Rep. Asia      1997    74.6  4.62e7   15994.
11 Korea, Rep. Asia      2002    77.0  4.80e7   19234.
12 Korea, Rep. Asia      2007    78.6  4.90e7   23348.
> |
```



```
country=="Japan", ]
year lifeExp      pop gdpPercap
<int>   <dbl>   <int>   <dbl>
1952    63.0  86459025    3217.
1957    65.5  91563009    4318.
1962    68.7  95831757    6577.
1967    71.4 100825279    9848.
1972    73.4 107188273   14779.
1977    75.4 113872473   16610.
1982    77.1 118454974   19384.
1987    78.7 122091325   22376.
1992    79.4 124329269   26825.
1997    80.7 125956499   28817.
2002    82   127065841   28605.
2007    82.6 127467972   31656.
```



5.2 베이스 R을 이용한 데이터 가공

■ 샘플과 속성의 추출

- 국가 이름이 "Korea, Rep."인 샘플을 조건식을 사용해 추출 + 인구 속성만 추출 + 등등

```
Console C:/Rsources/    
> gapminder[gapminder$country=="Korea, Rep.", c("pop")]  
# A tibble: 12 x 1  
      pop  
  <int>  
1 20947571  
2 22611552  
3 26420307  
4 30131000  
5 33505000  
6 36436000  
7 39326000  
8 41622000  
9 43805450  
10 46173816  
11 47969150  
12 49044790  
> |  
  
> gapminder[gapminder$country=="Korea, Rep.", c("lifeExp", "pop")]  
# A tibble: 12 x 2  
  lifeExp      pop  
  <dbl>    <int>  
1  47.5 20947571  
2  52.7 22611552  
3  55.3 26420307  
4  57.7 30131000  
5  62.6 33505000  
6  64.8 36436000  
7  67.1 39326000  
8  69.8 41622000  
9  72.2 43805450  
10 74.6 46173816  
11 77.0 47969150  
12 78.6 49044790  
> |
```



5.2 베이스 R을 이용한 데이터 가공

■ 샘플과 속성의 추출

- 국가 이름이 "Korea, Rep."인 샘플을 조건식을 사용해 추출

Console C:/RSources/ ↗

```
> gapminder[gapminder$country=="Korea, Rep."&gapminder$year>1970, c("lifeExp", "pop")]
```

```
# A tibble: 8 x 2
```

	lifeExp <dbl>	pop <int>
1	62.6	33505000
2	64.8	36436000
3	67.1	39326000
4	69.8	41622000
5	72.2	43805450
6	74.6	46173816
7	77.0	47969150
8	78.6	49044790

```
> apply(gapminder[gapminder$country=="Korea, Rep.", c("lifeExp", "pop", "gdpPercap")], 2, mean)
```

lifeExp	pop	gdpPercap
65.001	36499386.333	8217.318

```
> |
```



5.2 베이스 R을 이용한 데이터 가공

- 데이터 가공은 데이터 프레임을 중심으로 R이 제공하는 다양한 연산자와 함수를 이용해 이루어지는 작업
- 보다 정교하게 추출하려면 조건식 여러 개를 논리 연산자로 결합
- 데이터를 탐색하는 과정에서, 샘플들의 요약 통계 혹은 행/열 단위의 빠른 연산이 필요한 때가 있음
- R에서 제공하는 apply 함수를 이용하면 데이터 프레임을 구성하는 여러 항목을 한꺼번에 연산 가능



Thank you

