



# 2주차: 데이터 사이언스 세계로

**ChulSoo Park**

School of Computer Engineering & Information Technology  
Korea National University of Transportation

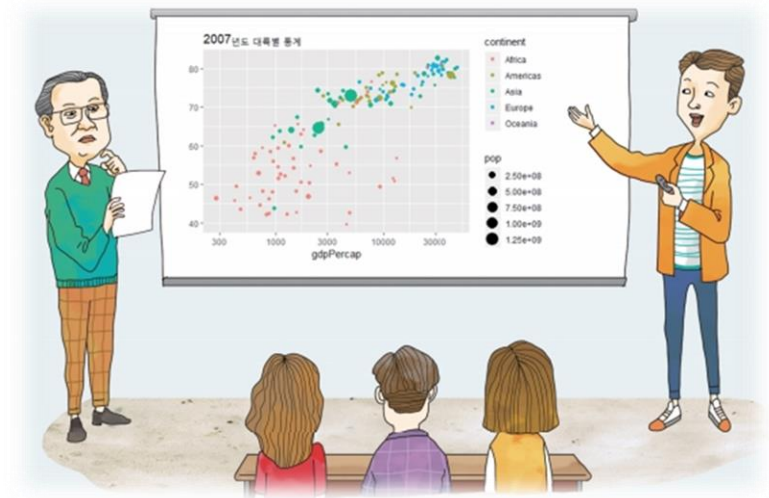
# 학습목표 (2주차)

- ❖ 분석 도구 챙기기(R, Rstudio 설치)
- ❖ 분석 도구(R) 익히기
- ❖ R 분석도구로 데이터 핸들링
- ❖ 데이터 시각화 맛보기
- ❖ 데이터 과학 Process 이해

# 02

## CHAPTER

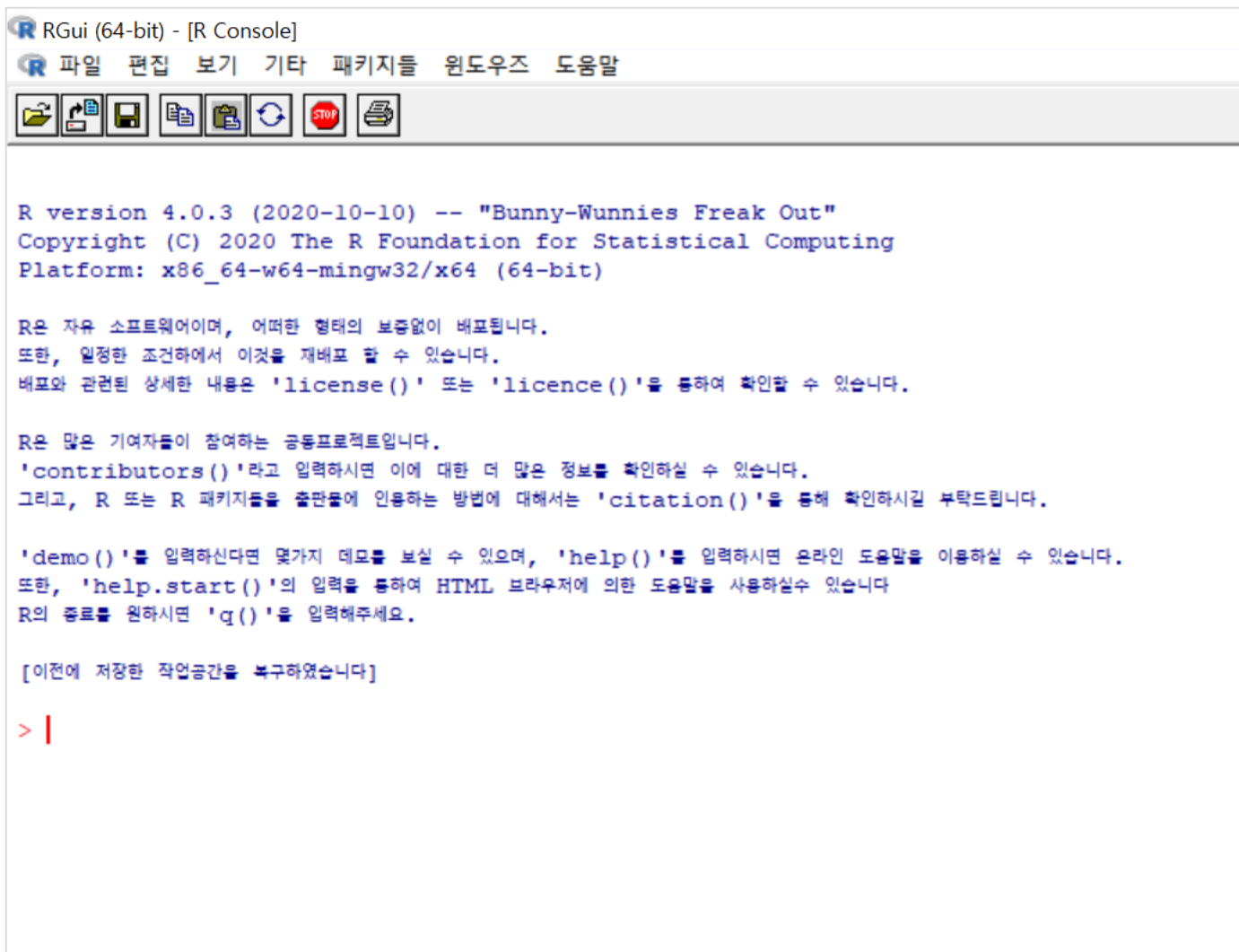
# 데이터 과학으로 풍덩



## CONTENTS

- 2.1 도구 챙기기
- 2.2 데이터와 친해지기
- 2.3 데이터 시각화 맛보기
- 2.4 데이터 과학을 위한 좋은 습관 알아보기
- 2.5 좋은 도구 익히기
- 2.6 데이터와 더 친해지기
  - 요약

## 2.5 좋은 도구 익히기



```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지들 윈도우즈 도움말

R version 4.0.3 (2020-10-10) -- "Bunny-Wunnies Freak Out"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R은 자유 소프트웨어이며, 어떠한 형태의 보증없이 배포됩니다.
또한, 일정한 조건하에서 이것을 재배포 할 수 있습니다.
배포와 관련된 상세한 내용은 'license()' 또는 'licence()'를 통하여 확인할 수 있습니다.

R은 많은 기여자들이 참여하는 공동프로젝트입니다.
'contributors()'라고 입력하시면 이에 대한 더 많은 정보를 확인하실 수 있습니다.
그리고, R 또는 R 패키지들을 출판물에 인용하는 방법에 대해서는 'citation()'을 통해 확인하시길 부탁드립니다.

'demo()'를 입력하신다면 몇가지 데모를 보실 수 있으며, 'help()'를 입력하시면 온라인 도움말을 이용하실 수 있습니다.
또한, 'help.start()'의 입력을 통하여 HTML 브라우저에 의한 도움말을 이용하실 수 있습니다.
R의 종료를 원하시면 'q()'를 입력해주세요.

[이전에 저장한 작업공간을 복구하였습니다]

> |
```

## 2.5 좋은 도구 익히기

### ① 통합 개발 환경(Rstudio)

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Project: (None)

Environment History Connections Tutorial

Import Dataset

R Global Environment

Environment is empty

Files Plots Packages Help Viewer

New Folder Delete Rename More

Home

	Name	Size	Modified
<input type="checkbox"/>	.Rhistory	5 B	Feb 4, 2021, 10:0
<input type="checkbox"/>	5. [사양] 2019년도 컨설팅 우수사례...	32.8 MB	Jan 29, 2021, 8:5
<input type="checkbox"/>	7. [경영] 2019년도 컨설팅 우수사례...	10.6 MB	Jan 29, 2021, 8:5
<input type="checkbox"/>	그림		
<input type="checkbox"/>	그림.zip	2.8 MB	Oct 16, 2020, 6:5

Console ~/

그리고, R 또는 R 패키지들을 출판물에 인용하는 방법에 대해서는 'citation()'을 통해 확인하시길 부탁드립니다.

'demo()'를 입력하신다면 몇가지 데모를 보실 수 있으며, 'help()'를 입력하시면 온라인 도움말을 이용하실 수 있습니다.





또한, 'help.start()'의 입력을 통하여 HTML 브라우저에 의한 도움말을 이용하실 수 있습니다.

R의 종료를 원하시면 'q()'을 입력해주세요.

> |

## 2.5 좋은 도구 익히기

### ② 스크립트 창에 있는 프로그램을 실행하는 네 가지 방법

- 한 줄만 실행 : 실행하려는 명령어에 커서를 올려놓고  Run (또는 **Ctrl** + **Enter**)를 누른다. [그림 2-12]는 3행을 실행한 경우로 실행 결과가 콘솔 창에 나타난다.
- 여러 줄을 실행 : 마우스로 명령어 여러 줄을 선택한 다음에  Run (또는 **Ctrl** + **Enter**)를 누른다.
- 전체 프로그램을 실행 :  Source 를 누른 다음 [Source with Echo]를 선택(또는 **Ctrl** + **Shift** + **Enter**)한다. [Source]를 선택하면 >프롬프트 없이 출력된다.
- 바로 이전에 실행한 부분을 다시 실행하려면  을 누른다.

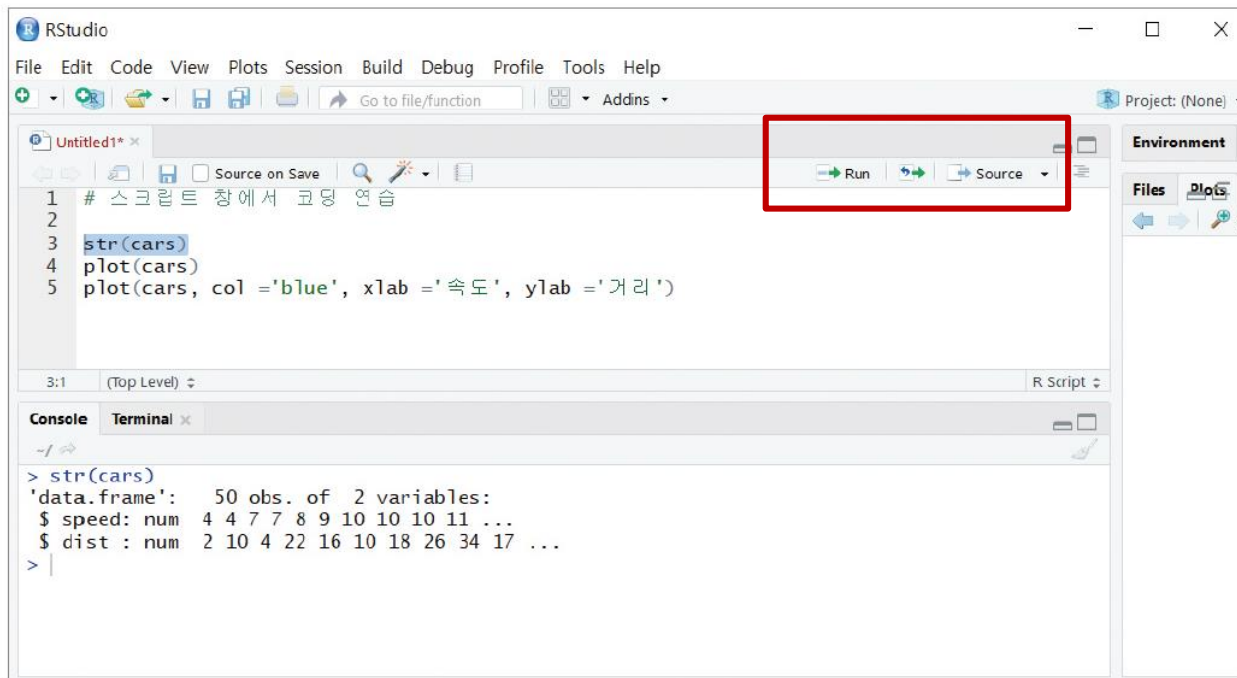


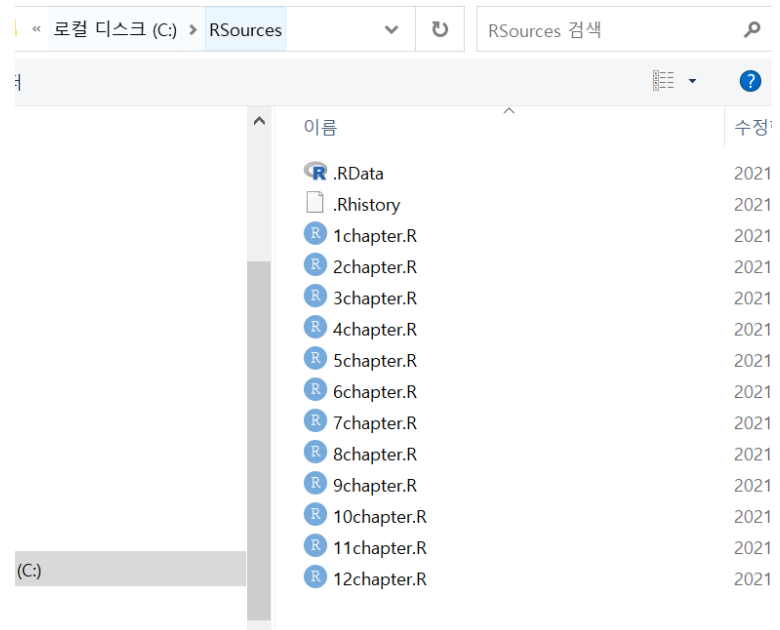
그림 2-12 스크립트 창에서 코딩하고 실행하기

## 2.5 좋은 도구 익히기

### ③ 작업 디렉토리의 지정

- 지정한 디렉토리(폴더)에 데이터 파일을 저장하는 방법
- getwd 함수는 현재 작업 디렉토리를 보여줌
- setwd로 자신이 원하는 곳으로 지정 가능 (아래 코드는 C:/Sources로 지정하는 경우)

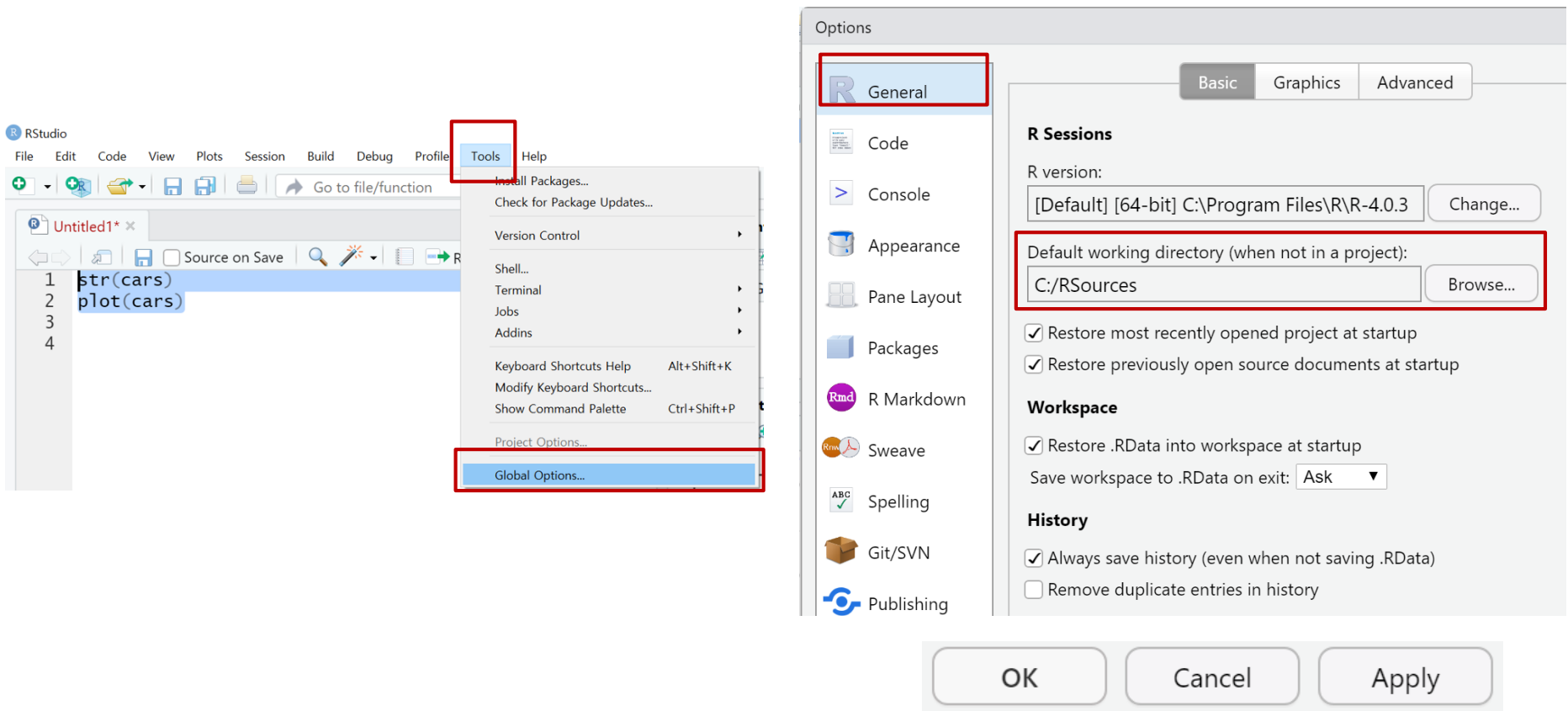
```
Console c:/RSources/ ➔  
> setwd('c:/RSources')  
> getwd()  
[1] "c:/RSources"
```



## 2.5 좋은 도구 익히기

### ③ 작업 디렉토리의 지정

- 디폴트 작업 디렉토리의 기본 값 지정 방법
- R 스튜디오 → [Tools] → [Global Options] → [Option] → [General] → [Default working directory] 항목에 지정하고자하는 디렉토리 지정(예 C:/Rsources) 입력 → [ OK ]





## 2.5 좋은 도구 익히기

### ④ 라이브러리(패키지)의 활용

- 라이브러리는 특정 분야를 위해 개발된 R 함수를 모아둔 소프트웨어
  - ✓예) ggplot2는 데이터를 깔끔하고 일관성 있게 시각화하는 함수 모음
  - ✓예) gapminder는 1952~2007년까지 5년 간격으로 여러 나라의 인구, 1인당 GDP, 기대 수명 등을 모은 갭마인더 데이터를 활용하는데 필요한 함수 모음
- R이 강력하고 널리 쓰이는 이유는 방대한 라이브러리 덕분
- CRAN 사이트에 접속하면 현재 계속 추가됨을 알 수 있음
  - ✓[Packages] 메뉴: R이 제공하는 모든 라이브러리 확인 가능
  - ✓[Task Views] 메뉴: 라이브러리를 분야별로 묶어 소개

## ④ 라이브러리(패키지)의 활용

The screenshot shows the RStudio interface with the following components:

- Source Editor:** Contains R code for installing and loading the `gapminder` package.
 

```

22 class(x)
23
24
25 library(gapminder)
26 ?gapminder
27
      
```
- Console:** Shows the output of the commands:
 

```

> ?gapminder
No documentation for 'gapminder' in specified packages and libraries:
you could try '??gapminder'
> library(gapminder)
> ?gapminder
      
```
- Environment/History/Connections/Tutorial Panel:** Shows the `gapminder` package loaded in the Environment pane.
- Files/Plots/Packages/Help/Viewer Panel:** Shows the R documentation for `gapminder`.
 

**Gapminder data.**

**Description**

Excerpt of the Gapminder data on life expectancy, GDP per capita, and population by country.

**Usage**

```
gapminder
```

**Format**

The main data frame `gapminder` has 1704 rows and 6 variables:

## 2.5 좋은 도구 익히기

### ④ 라이브러리(패키지)의 활용

- 설치 방법(1. 명령어 실행, 2. 메뉴 선택 방법)
  - 예) R에서 가장 널리 활용되는 dplyr(디플라이어)와 ggplot2(지지도플롯투)

The screenshot shows the RStudio interface with the following components:

- Source Editor:** Contains R code for installing packages. Lines 29 and 30 are highlighted with a red box:
 

```
install.packages("dplyr")
install.packages("ggplot2")
```
- Console:** Shows the output of the installation process:
 

```
https://cran.rstudio.com/bin/windows/Rtools/
also installing the dependency 'lifecycle'

URL 'https://cran.rstudio.com/bin/windows/contrib/4.0/lifecycle_1.0.0.zip'을 시도합니다
Content type 'application/zip' length 111126 bytes (108 KB)
downloaded 108 KB

URL 'https://cran.rstudio.com/bin/windows/contrib/4.0/dplyr_1.0.5.zip'을 시도합니다
Content type 'application/zip' length 1331233 bytes (1.3 MB)
downloaded 1.3 MB

package 'lifecycle' successfully unpacked and MD5 sums checked
package 'dplyr' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
C:\Users\Administrator\AppData\Local\Temp\RtmpCeIQoV\downloaded_packages
> |
```
- Environment:** Displays the Global Environment with variables:
 

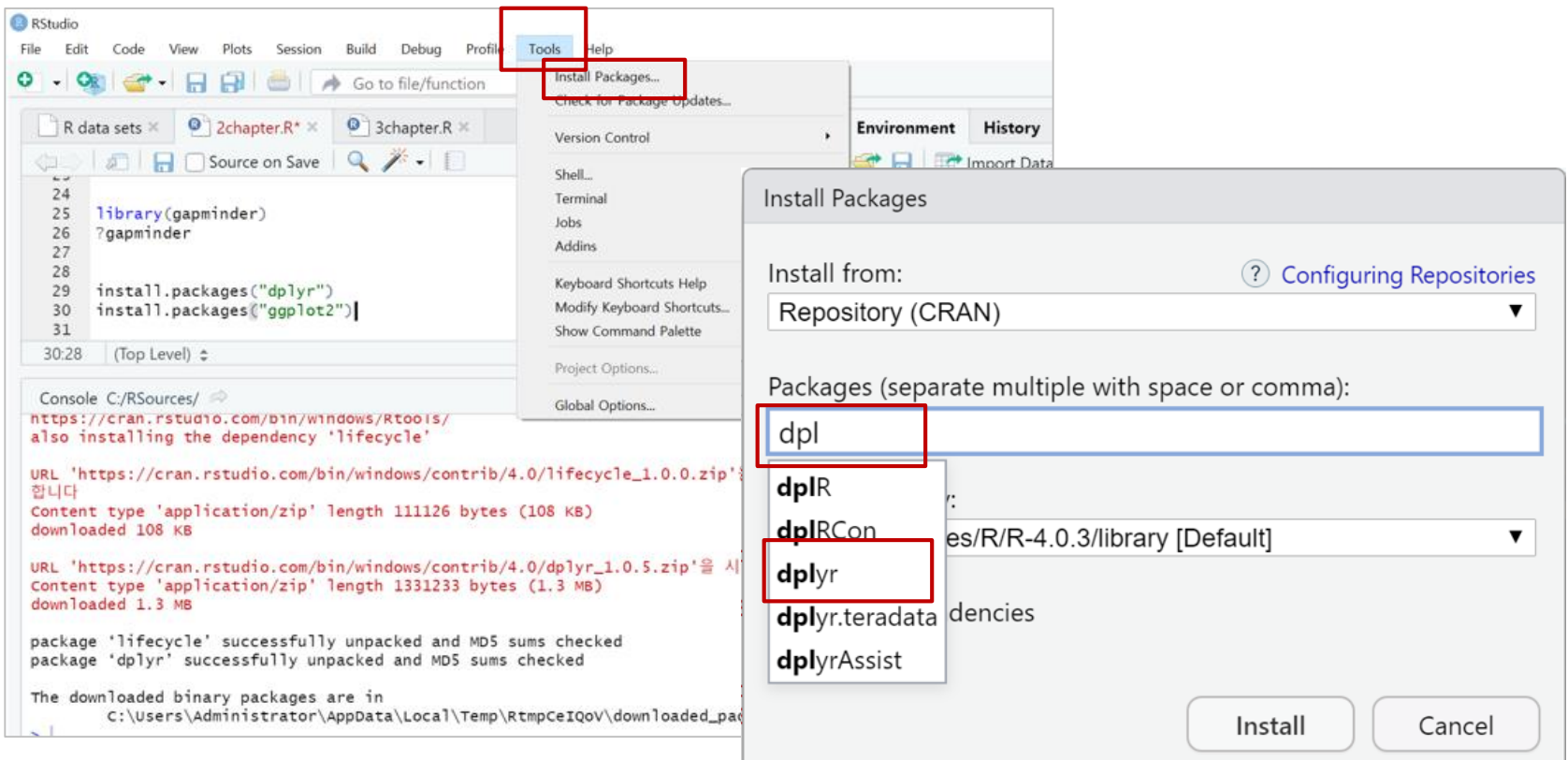
Variable	Value
a	Factor w/ 3 levels "A","B","O": 1 2 3
aa	-25
b	-5
bb	25
cc	25
x	1L
- Packages:** Shows the System Library with a list of installed and available packages:
 

Name	Description	Version
<input type="checkbox"/> abind	Combine Multidimensional Arrays	1.4-5
<input type="checkbox"/> assertthat	Easy Pre and Post Assertions	0.2.1
<input checked="" type="checkbox"/> base	The R Base Package	4.0.3
<input type="checkbox"/> base64enc	Tools for base64 encoding	0.1-3
<input type="checkbox"/> BH	Boost C++ Header Files	1.75.0-0
<input type="checkbox"/> boot	Bootstrap Functions (Originally by Angelo Canty for S)	1.3-25

## 2.5 좋은 도구 익히기

### ④ 라이브러리(패키지)의 활용

- 설치 방법(1. 명령어 실행, 2. 메뉴 선택 방법)
  - 예) R에서 가장 널리 활용되는 dplyr(디플라이어)와 ggplot2(지지플롯투)



## 2.5 좋은 도구 익히기

### ④ 라이브러리(패키지)의 활용

- 사용할 때는 library 함수를 이용하여 부착해야 함

The screenshot illustrates the RStudio interface during the execution of R code to load packages. The script editor on the left shows the following code:

```

31
32 library(dplyr)
33 library(ggplot2)
34 search()
35

```

Two callouts highlight the code: "라이브러리 부착" (Library attachment) points to the `library(dplyr)` and `library(ggplot2)` lines, and "부착 목록 확인" (Check attachment list) points to the `search()` line.

The console output shows the results of these commands:

```

> library(dplyr)
다음의 패키지를 부착합니다: 'dplyr'

The following objects are masked from 'package:stats':
  filter, lag

The following objects are masked from 'package:base':
  intersect, setdiff, setequal, union

경고메시지(들):
패키지 'dplyr'는 R 버전 4.0.4에서 작성되었습니다
> library(ggplot2)
> search()
[1] ".GlobalEnv"          "package:ggplot2"      "package:dplyr"
[4] "tools:rstudio"        "package:stats"        "package:graphics"
[7] "package:grDevices"    "package:utils"        "package:datasets"
[10] "package:methods"     "AutoLoads"            "package:base"
>

```

The Environment pane on the right shows the Global Environment with the following variables:

Variable	Value
a	Factor w/ 3 levels "A","B","O": 1 2 3
aa	-25
b	-5
bb	25
cc	25
x	1L

The Files pane on the right shows the project structure in the `C:/RSources` directory:

Name	Size	Modified
..		
.RData	851.7 KB	Mar 11
.Rhistory	21.6 KB	Mar 11
1chapter.R	33 B	Mar 4
2chapter.R	307 B	Mar 12
3chapter.R	591 B	Mar 11
4chapter.R	33 B	Mar 4
5chapter.R	98 B	Feb 18

## 2.5 좋은 도구 익히기

### ④ 라이브러리(패키지)의 활용

- 라이브러리 설치는 라이브러리 파일을 하드 디스크에 저장
- 라이브러리 부착은 하드 디스크에서 주기억 장치로 적재

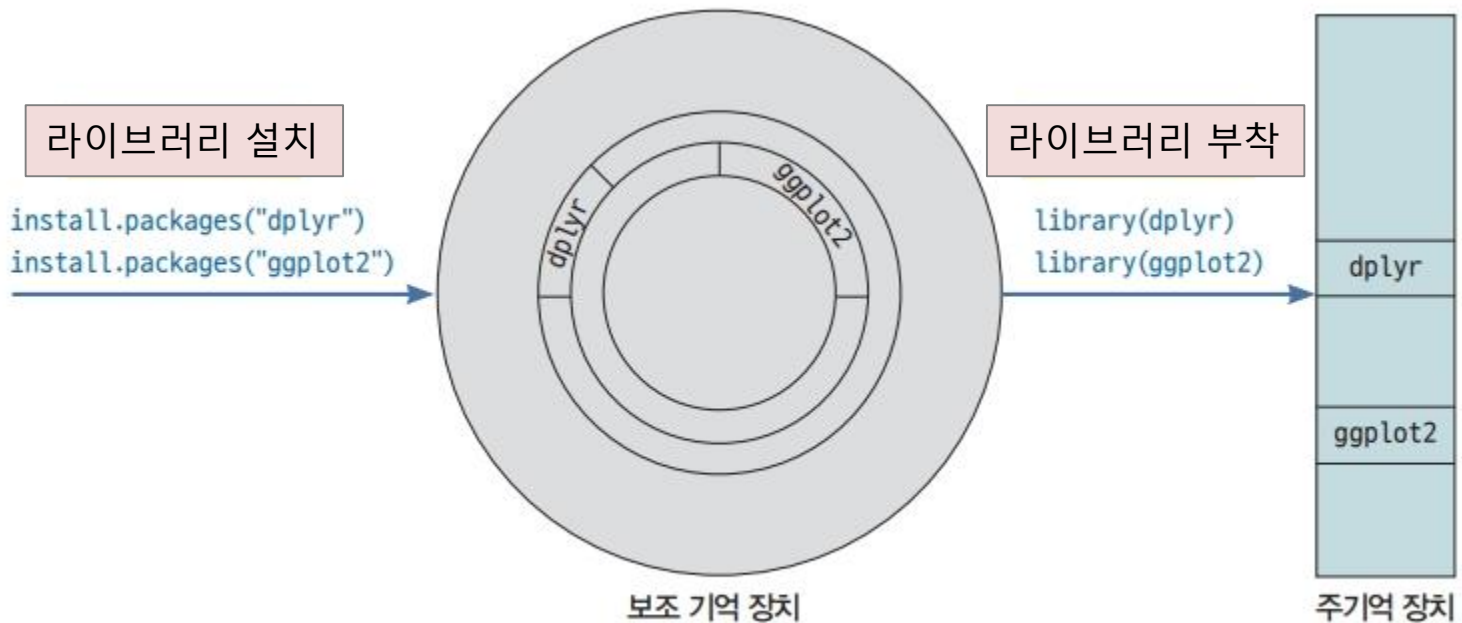


그림 2-15 라이브러리의 설치와 부착에 따른 기억 장치 변화

## 2.6 데이터와 더 친해지기

### ① 사랑스런 iris 데이터

- 1936년 엡저 앤더슨이 캐나다 동부 지역의 가스페 반도에 서식하는 붓꽃을 수집
- 같은 날에 세 가지 품종 (setosa, versicolor, virginica) 각각을 50송이씩 채취
- 같은 사람이 같은 자를 사용하여 꽃잎의 너비와 길이, 꽃받침의 너비와 길이를 측정
- 통계학자인 로널드 피셔 교수가 논문으로 발표하여 유명해졌으며, 지금도 널리 사용됨



setosa



versicolor



virginica

그림 2-16 Iris 데이터에 있는 붓꽃의 세 가지 품종

- 데이터에 대한 관심과 애정, 직관을 키워보자.

## 2.6 데이터와 더 친해지기

### ① 사랑스런 iris 데이터

- str과 head 함수로 내용 살펴보기
- >head(iris, 15)에서 15의 의미는 15개 행을 의미함

RStudio interface showing the execution of R code to explore the iris dataset.

**Source Editor:**

```

30 install.packages("ggplot2")
31
32 library(dplyr)
33 library(ggplot2)
34 search()
35
36 str(iris)
37 head(iris)
38

```

**Console Output:**

```

> str(iris)
'data.frame': 150 obs. of 5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",...: 1 1 1 1 1 1 1 1 1 1 ...
...
> head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1          3.5          1.4          0.2  setosa
2          4.9          3.0          1.4          0.2  setosa
3          4.7          3.2          1.3          0.2  setosa
4          4.6          3.1          1.5          0.2  setosa
5          5.0          3.6          1.4          0.2  setosa
6          5.4          3.9          1.7          0.4  setosa

```

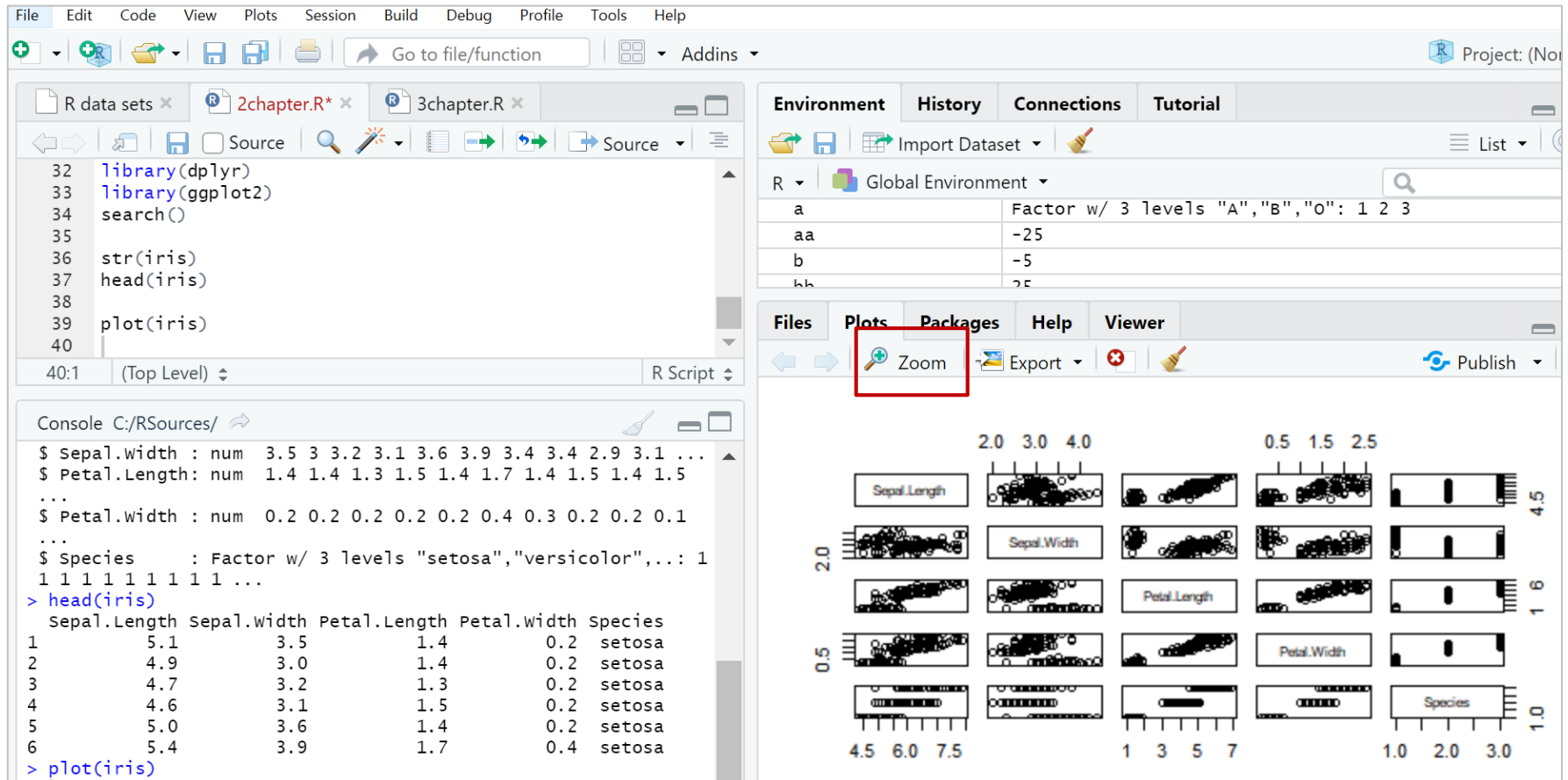
**Environment Pane:**

Variable	Value
a	Factor w/ 3 levels "A","B","O": 1 2 3
aa	-25
b	-5
bb	25
cc	25
x	1L
xx	NULL
xy	chr [1:3] "1" "2" "AA"
y	NULL



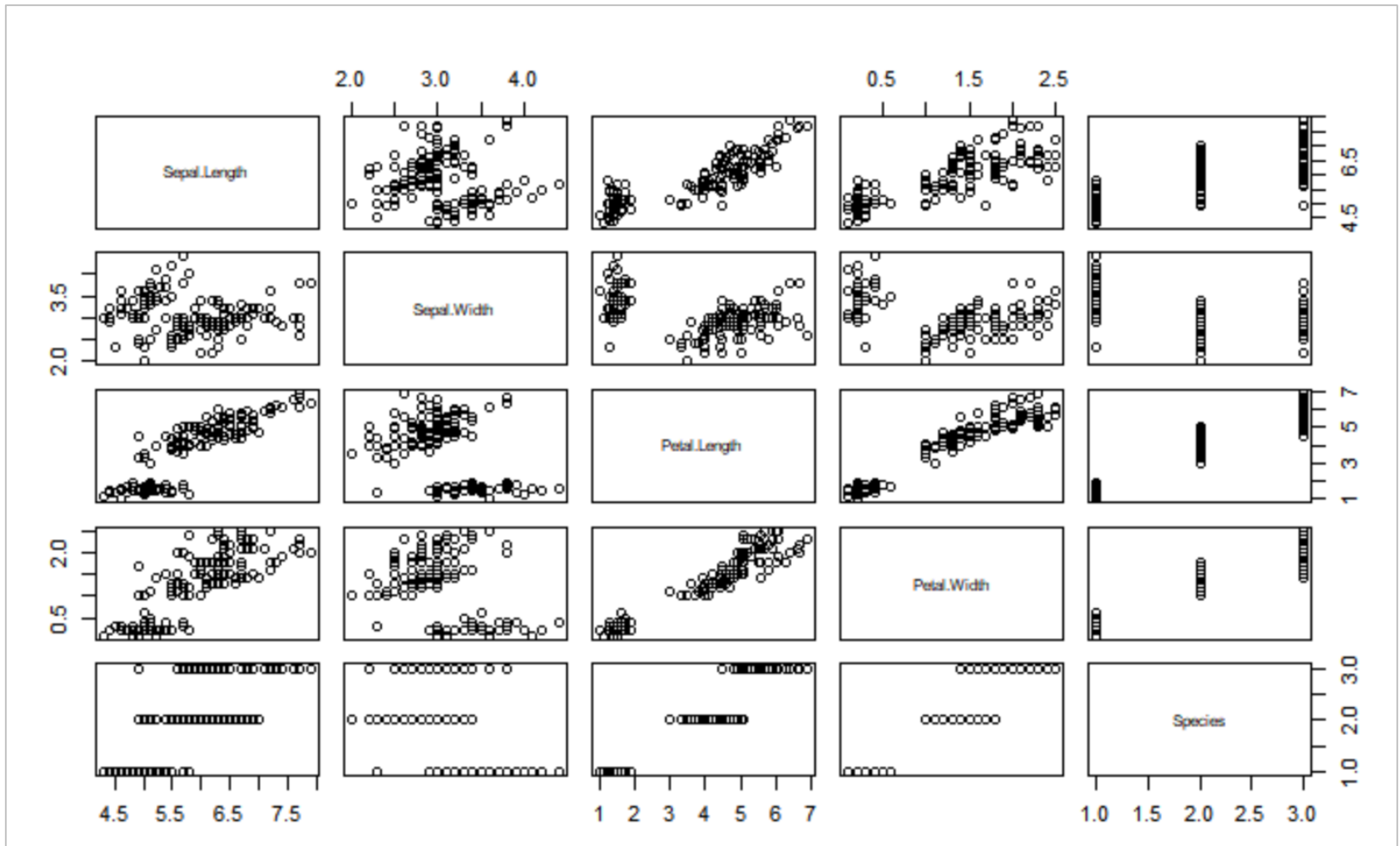
## 2.6 데이터와 더 친해지기

### ① 사랑스런 iris 데이터(plot 함수로 시각화 : > plot(iris))



## 2.6 데이터와 더 친해지기

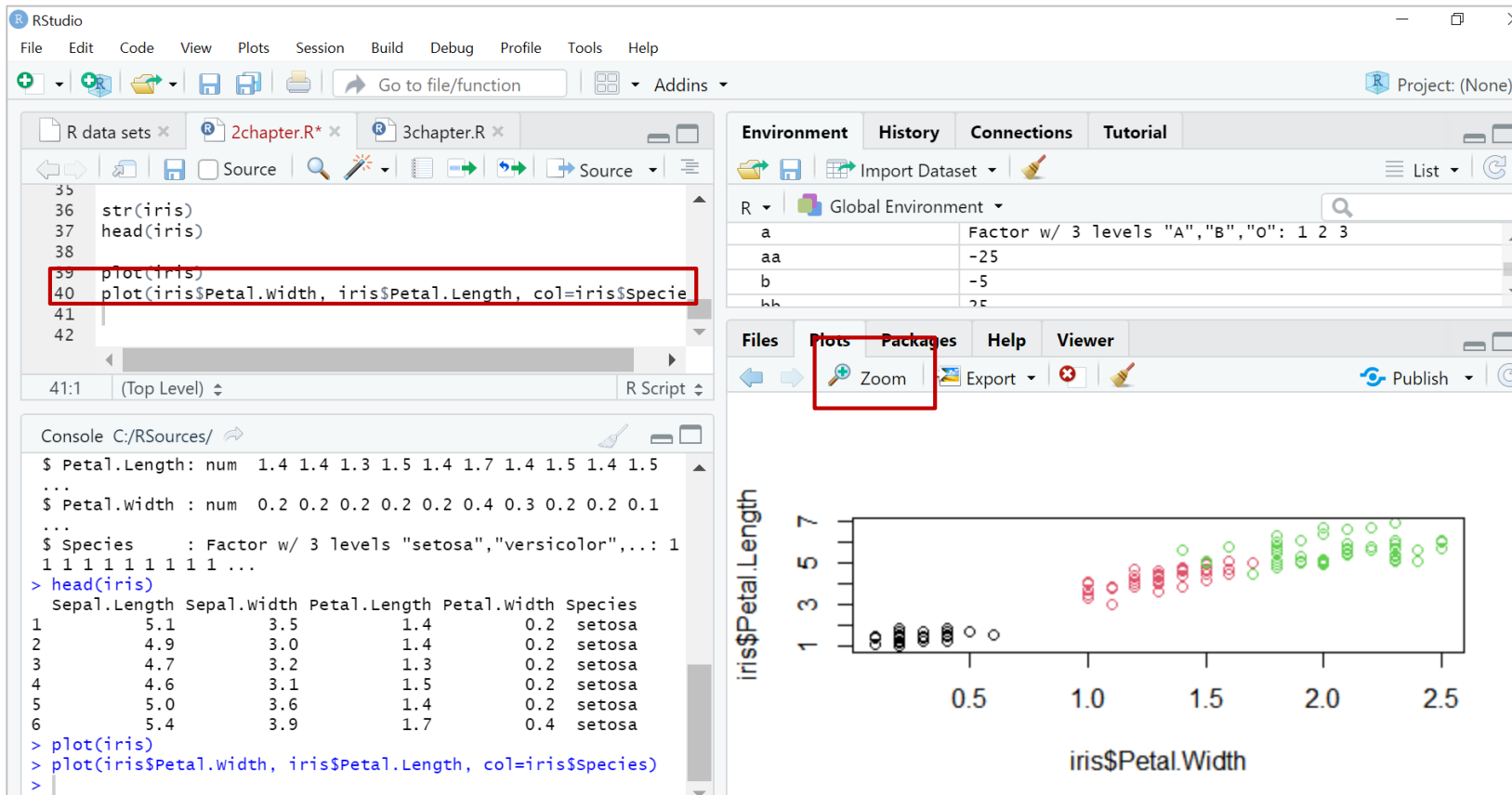
① 사랑스런 iris 데이터(plot 함수로 시각화 : `> plot(iris)`)



## 2.6 데이터와 더 친해지기

### ① 사랑스런 iris 데이터

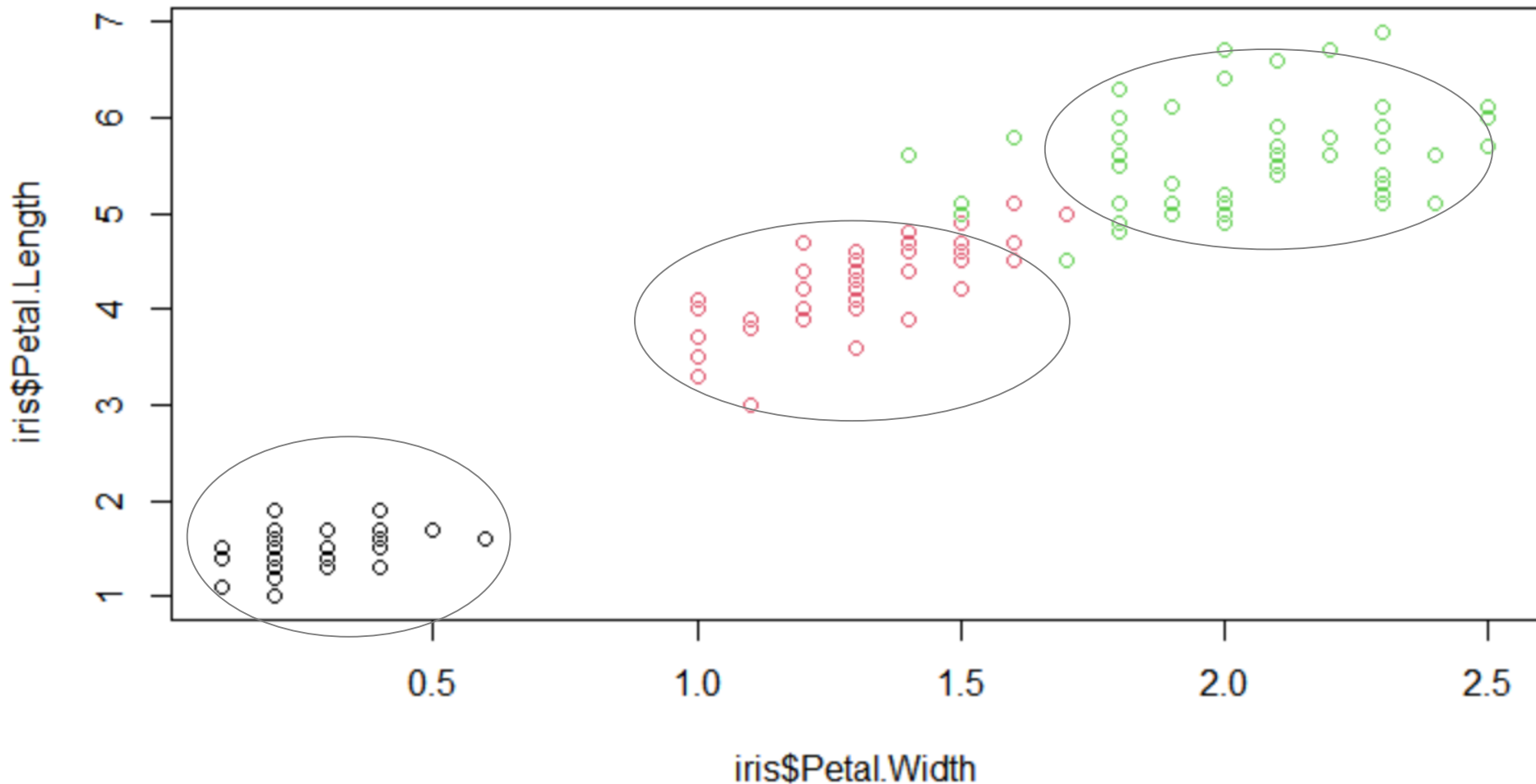
- 두 속성의 상관관계를 그림
- `col=iris$Species`는 iris 데이터의 Species 값에 따라 서로 다른 색상으로 그리라는 옵션
- `plot(iris$Petal.Width, iris$Petal.Length, col=iris$Species)`



## 2.6 데이터와 더 친해지기

### ① 사랑스런 iris 데이터

- `plot(iris$Petal.Width, iris$Petal.Length, col=iris$Species)`



## 2.6 데이터와 더 친해지기

### ② 돈을 벌어주는 tips 데이터

- 식당에서 팁을 받아 생계를 꾸리는 상황
- 데이터 과학을 활용하면 더 많은 팁을 챙길 수 있음

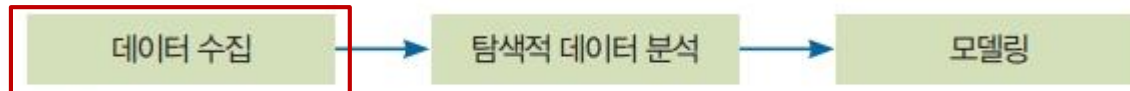


그림 1-5 데이터 과학의 절차

### ■ 단계 1: 데이터 수집

- 일곱 가지 변수에 대해 값을 수집하기로 결정
  - 계산서 금액(total\_bill)
  - 팁 액수(tip)
  - 계산한 사람의 성별(sex)
  - 흡연자 포함 여부(smoker)
  - 요일(day)
  - 시간(time)
  - 동석자 수(size)
- 몇 주일 고생하여 244건을 모아 tips.csv 파일에 저장함

## 2.6 데이터와 더 친해지기

### ■ 단계 1: 데이터 수집(tips.csv) 파일 연동

```
tips = read.csv('https://raw.githubusercontent.com/mwaskom/seaborn-data/master/tips.csv')
```

The image shows the RStudio interface. The top-left pane displays the R script with the following code:

```
41  
42 tips = read.csv('https://raw.githubusercontent.com/mwaskom/seaborn-data/master/tips.csv')  
43 str(tips)  
44 head(tips)  
45
```

The bottom-left pane shows the console output:

```
> tips = read.csv('https://raw.githubusercontent.com/mwaskom/seaborn-data/master/tips.csv')  
> str(tips)  
'data.frame': 244 obs. of 7 variables:  
 $ total_bill: num 17 10.3 21 23.7 24.6 ...  
 $ tip : num 1.01 1.66 3.5 3.31 3.61 4.71 2 3.12 1.96 3.23 ...  
 $ sex : chr "Female" "Male" "Male" "Male" ...  
 $ smoker : chr "No" "No" "No" "No" ...  
 $ day : chr "Sun" "Sun" "Sun" "Sun" ...  
 $ time : chr "Dinner" "Dinner" "Dinner" "Dinner" ...  
 $ size : int 2 3 3 2 4 4 2 4 2 2 ...  
> head(tips)  
 total_bill tip sex smoker day time size  
1 16.99 1.01 Female No Sun Dinner 2  
2 10.34 1.66 Male No Sun Dinner 3  
3 21.01 3.50 Male No Sun Dinner 3  
4 23.68 3.31 Male No Sun Dinner 2  
5 24.59 3.61 Female No Sun Dinner 4  
6 25.29 4.71 Male No Sun Dinner 4
```

The right-hand pane shows the Environment tab with the following data objects:

Object	Size
tips	244 obs. of 7 variables
tmp1	338 obs. of 14 variables
v.c.non	60 obs. of 3 variables

The bottom-right pane shows a scatter plot of iris data with 'Petal.Length' on the y-axis and 'Petal.Width' on the x-axis. The plot shows three distinct clusters of points, colored red, green, and blue, representing different species of iris flowers.

**CSV이란 ?**([영어](#): comma-separated values) 몇 가지 필드를 [선표\(,\)](#)로 구분한 [텍스트](#) 데이터 및 텍스트 파일이다. 확장자는 .csv이며 [MIME 형식](#)은 text/csv이다. **comma-separated variables**라고도 한다.

## 2.6 데이터와 더 친해지기

```
> head(tips)
```

	total_bill	tip	sex	smoker	day	time	size
1	16.99	1.01	Female	No	Sun	Dinner	2
2	10.34	1.66	Male	No	Sun	Dinner	3
3	21.01	3.50	Male	No	Sun	Dinner	3
4	23.68	3.31	Male	No	Sun	Dinner	2
5	24.59	3.61	Female	No	Sun	Dinner	4
6	25.29	4.71	Male	No	Sun	Dinner	4

- 첫 번째 샘플을 해석하면, 일요일에 2명이 저녁 식사를 했고 흡연자는 없었고 16.99달러를 여자가 계산한 테이블에서 1.01달러의 팁을 받았다는 사실을 나타냄

## 2.6 데이터와 더 친해지기

### ② 단계 2: 탐색적 데이터 분석 (EDA)

- summary 함수로 요약 통계 summary statistics 확인
- 아래 요약 통계에서 어떤 분석이 가능한가?

```

45
46 summary(tips)
47

```

47:1 (Top Level) R Script

Console C:/RSources/

```

> summary(tips)
  total_bill    tip          sex      smoker
Min.   : 3.07  Min.   : 1.000  Length:244  Length:244
1st Qu.:13.35  1st Qu.: 2.000  Class :character  Class :character
Median :17.80  Median : 2.900  Mode  :character  Mode  :character
Mean   :19.79  Mean   : 2.998
3rd Qu.:24.13  3rd Qu.: 3.562
Max.   :50.81  Max.   :10.000

   day      time      size
Length:244  Length:244  Min.   :1.00
Class :character  Class :character  1st Qu.:2.00
Mode  :character  Mode  :character  Median :2.00
                                Mean   :2.57
                                3rd Qu.:3.00
                                Max.   :6.00

```

- 이 통계 요약은 요일이나 성별이 팁에 미치는 영향을 알 수 없으므로 시각화를 통해 더욱 깊숙이 탐색해 보자.



## 2.6 데이터와 더 친해지기

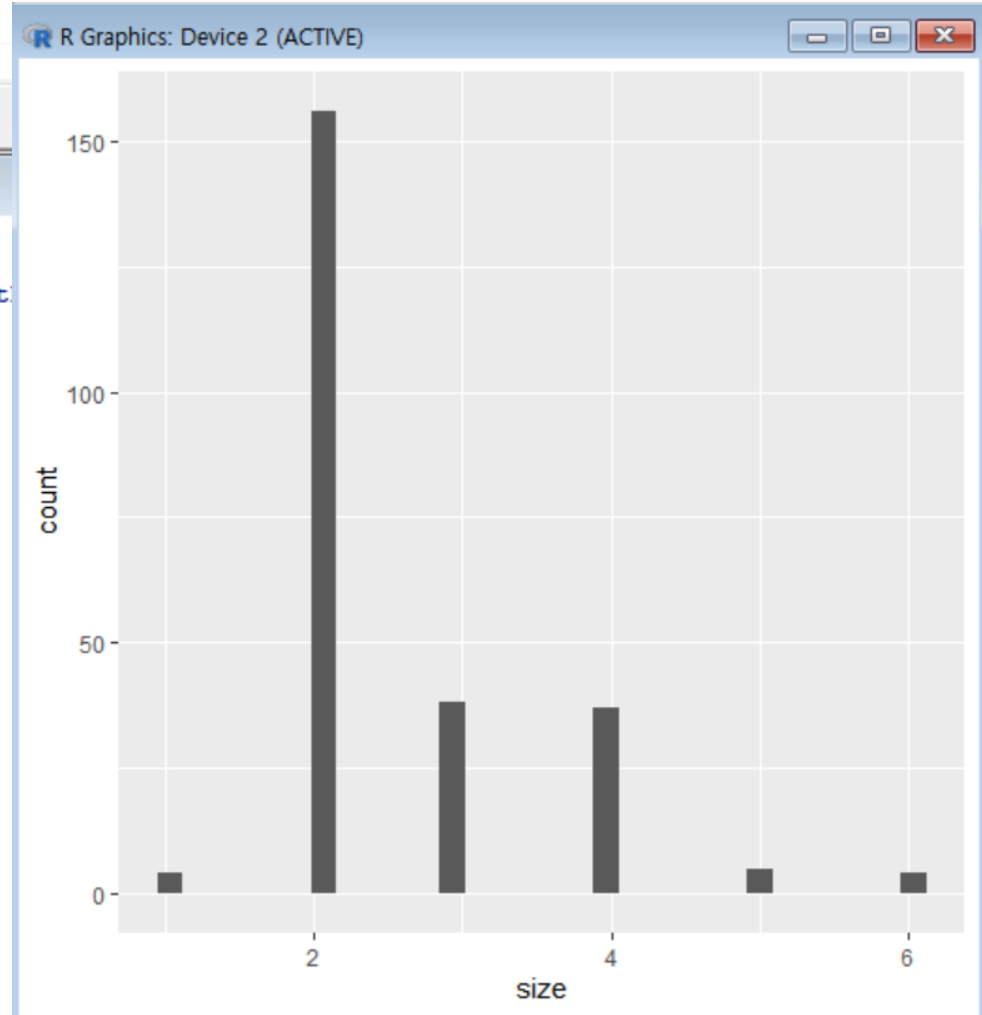
### ■ 단계 2: 탐색적 데이터 분석

- dplyr와 ggplot2 라이브러리 활용 (지금은 그냥 실행해 보고 의미는 5~6장에서 공부)

```
RGui (64-bit)
파일 작업기록 크기변경 윈도우즈

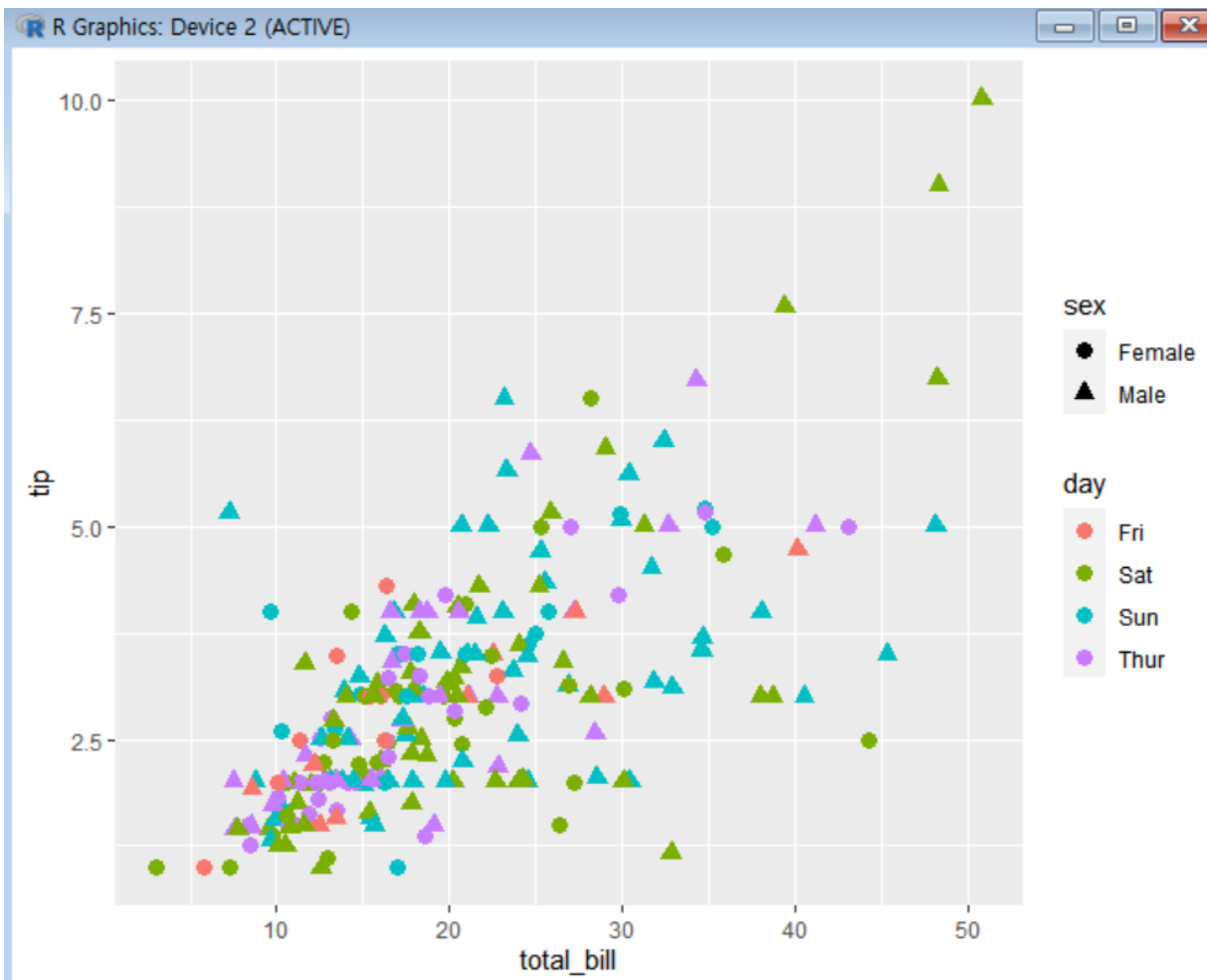
R Console

> tips%>%ggplot(aes(size))+geom_histogram()
`stat_bin()` using `bins = 30`. Pick better value with
> |
```



## 2.6 데이터와 더 친해지기

```
R Console
> tips%>%ggplot(aes(total_bill,tip))+geom_point(aes(col=day,pch=sex),size=3)
> |
```



### ■ 단계 2: 탐색적 데이터 분석

- 탐색적 데이터 분석의 효능 (돈을 더 벌기 위한 전략 수립에 도움)
  - 유리한 요일과 시간대에 자신의 근무시간을 맞추거나,
  - 성별이나 동석자 수를 보고 어느 테이블을 차지할지 결정하거나,
  - 프로모션 행사 때 어느 손님에게 할인 쿠폰을 제공할 지 결정 등

## 2.6 데이터와 더 친해지기

### ■ 단계 3: 모델링

- 탐색적 데이터 분석의 한계: 돈을 더 버는 전략을 짤 수는 있어도, 새로운 전략에서 수입이 얼마나 더 늘지 정확히 예측할 수는 없음
- 모델링을 하면 예측 (prediction)이 가능 → 미래 재정 포트폴리오를 짤 수 있음
  - 예) 동석자 수가 늘면 수입이 얼마 늘지, 성별 분포가 바뀌면 수입이 얼마나 늘지 미리 알 수 있음
- 모델링은 7~11장에서 공부

1. R 분석 도구 설치
2. R 패키지 Install, Library, 실행의 이해
3. R에서 데이터 활용 방법 습득
4. R에서 시각화 습득
5. 데이터 사이언스 Process 이해

# Thank you

