



# 2주차: 데이터 사이언스 세계로

**ChulSoo Park**

School of Computer Engineering & Information Technology

Korea National University of Transportation



# 학습목표 (2주차)

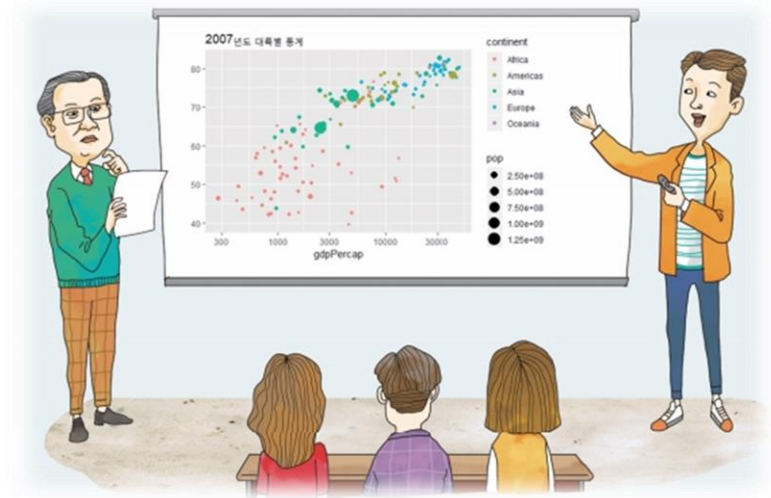
- ❖ 분석 도구 챙기기(R, Rstudio 설치)
- ❖ 분석 도구(R) 익히기
- ❖ R 분석도구로 데이터 핸들링
- ❖ 데이터 시각화 맛보기
- ❖ 데이터 과학 Process 이해



# 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



## 2.2 데이터와 친해지기

```

RGui (64-bit)
파일 편집 윈도우즈

R Console

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 확인할 수 있습니다.
그리고, R 또는 R 패키지를 출판물에 인용하는 방법에 대해서는 'citation()'을 입력하시면 됩니다.

'demo()'를 입력하신다면 몇가지 예제를 보실 수 있으며, 'help()'를 입력
또한, 'help.start()'의 입력을 통하여 HTML 브라우저에 의한 도움말을
R의 종류를 원하시면 'q()'를 입력해주세요.

> data()
> |

```

R data sets	
Data sets in package 'datasets':	
AirPassengers	Monthly Airline Passenger Numbers 1949-1960
BJsales	Sales Data with Leading Indicator
BJsales.lead (BJsales)	Sales Data with Leading Indicator
BOD	Biochemical Oxygen Demand
CO2	Carbon Dioxide Uptake in Grass Plants
ChickWeight	Weight versus age of chicks on different diets
DNase	Elisa assay of DNase
EuStockMarkets	Daily Closing Prices of Major European Stock Indices, 1991-1998
Formaldehyde	Determination of Formaldehyde
HairEyeColor	Hair and Eye Color of Statistics Students
Harman23.cor	Harman Example 2.3
Harman74.cor	Harman Example 7.4
Indometh	Pharmacokinetics of Indomethacin
InsectSprays	Effectiveness of Insect Sprays
warpbreaks	The Number of Breaks in Yarn during Weaving
women	Average Heights and Weights for American Women

## 2.2 데이터와 친해지기

### ■ 데이터의 내용 확인하기

- 데이터 이름을 명령어로 간주하여 입력함

> women →

```

> women
  height weight
1     58    115
2     59    117
3     60    120
4     61    123
5     62    126
6     63    129
7     64    132
8     65    135
9     66    139
10    67    142
11    68    146
12    69    150
13    70    154
14    71    159
15    72    164
  
```

그림 2-4 베이스 R의 women 데이터 확인 예

### ■ 행<sub>row</sub>과 열<sub>column</sub>

- 행을 샘플<sub>sample</sub> 또는 관측<sub>observation</sub>이라 부름
- 열을 속성<sub>attribute</sub>, 특징<sub>feature</sub>, 또는 변수<sub>variable</sub>라 부름

### ■ R은 그림 2-4와 같은 구조를 데이터 프레임<sub>data frame</sub>이라 부름



## 2.2 데이터와 친해지기

### ■ 데이터의 내용 확인하기(ChickWeight)

#### 데이터프레임은 무엇인가?

데이터프레임(dataframe)은 가장 대중적인 표형식 데이터에 대한 사실상 표준으로, 통계 및 시각화에 활용하는 자료구조다.

데이터프레임은 동일한 길이를 갖는 벡터 집합이다. 벡터 각각은 칼럼을 표현하지만 각 벡터는 서로 다른 자료형이 될 수 있다(예를 들어, 문자형, 정수형, 요인형). `str()` 함수를 사용해서 각 칼럼별 자료형을 조사한다.

```
> ChickWeight
  weight Time Chick Diet
1      42   0     1    1
2      51   2     1    1
3      59   4     1    1
4      64   6     1    1
5      76   8     1    1
6      93  10     1    1
7     106  12     1    1
8     125  14     1    1
9     149  16     1    1
10     171  18     1    1
11     199  20     1    1
12     205  21     1    1
13      40   0     2    1
14      49   2     2    1
15      58   4     2    1
16      72   6     2    1
17      84   8     2    1
18     103  10     2    1
```

```
> str(ChickWeight)
Classes 'nfnGroupedData', 'nfGroupedData', 'groupedData' and 'data.frame':
 $ weight: num  42 51 59 64 76 93 106 125 149 171 ...
 $ Time  : num  0 2 4 6 8 10 12 14 16 18 ...
 $ Chick : Ord.factor w/ 50 levels "18"<"16"<"15"<...: 15 15 15 15 15 15 15 15 15 15 ...
 $ Diet  : Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 1 1 1 1 1 ...
```

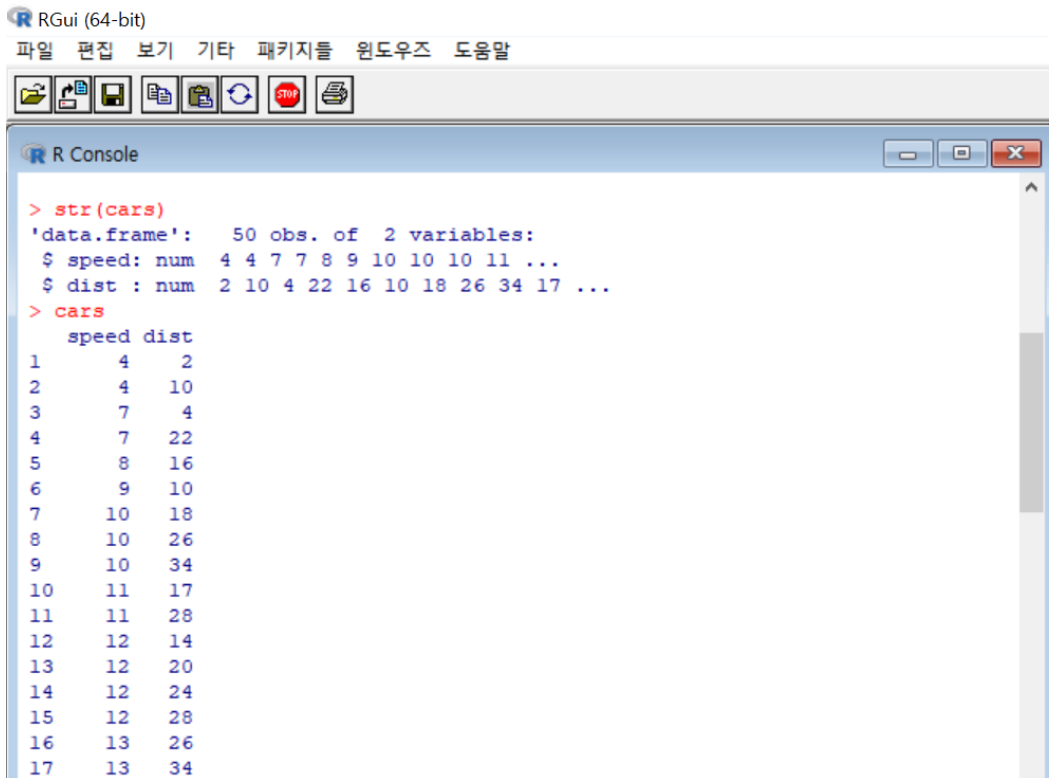
578 obs. of 4 variables:



## 2.2 데이터와 친해지기

### ■ cars 데이터로 반복 연습

- str 함수: 데이터의 내용을 요약해서 보여주는 함수
- str은 내부 구조(structure)을 의미임



RGui (64-bit)

파일 편집 보기 기타 패키지들 윈도우즈 도움말

R Console

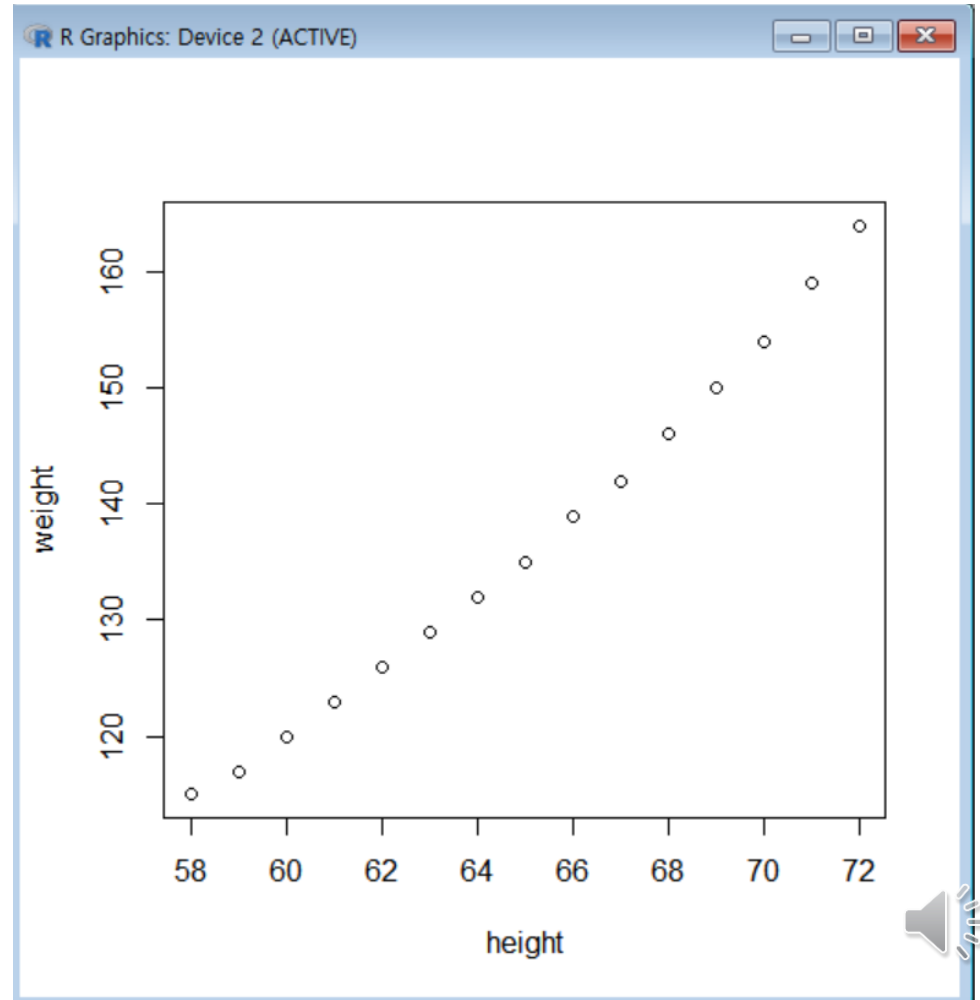
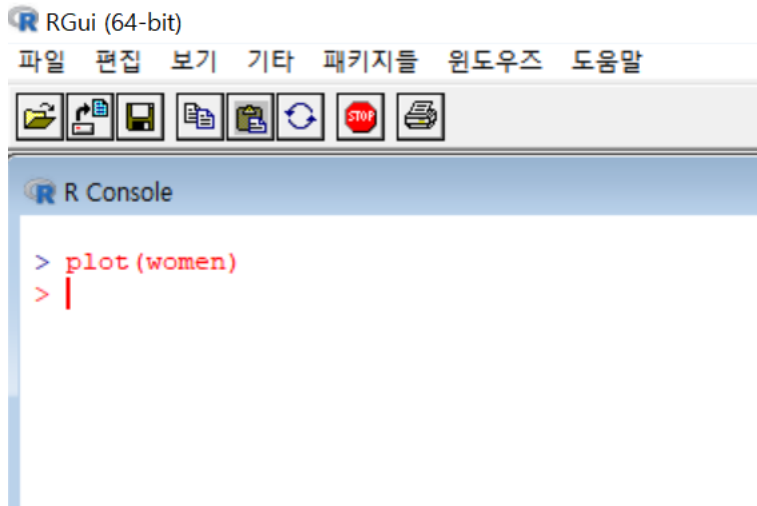
```
> str(cars)
'data.frame':  50 obs. of  2 variables:
 $ speed: num  4 4 7 7 8 9 10 10 10 11 ...
 $ dist : num  2 10 4 22 16 10 18 26 34 17 ...

> cars
  speed dist
1     4    2
2     4   10
3     7    4
4     7   22
5     8   16
6     9   10
7    10   18
8    10   26
9    10   34
10   11   17
11   11   28
12   12   14
13   12   20
14   12   24
15   12   28
16   13   26
17   13   34
```



## 2.3 데이터 시각화 맛보기

- 다양한 시각화<sub>visualization</sub> 함수
  - 베이스 R에서 가장 널리 쓰이는 plot 함수



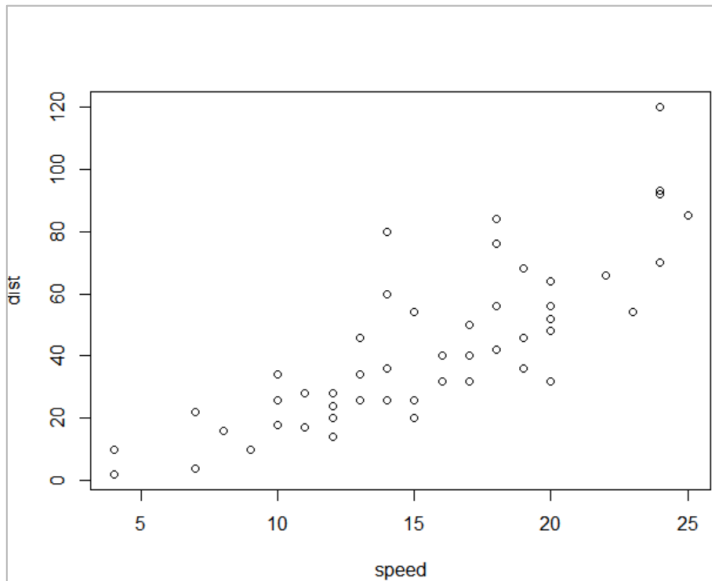


## 2.3 데이터 시각화 맛보기

### ■ 여러 가지 시각화 옵션을 적용

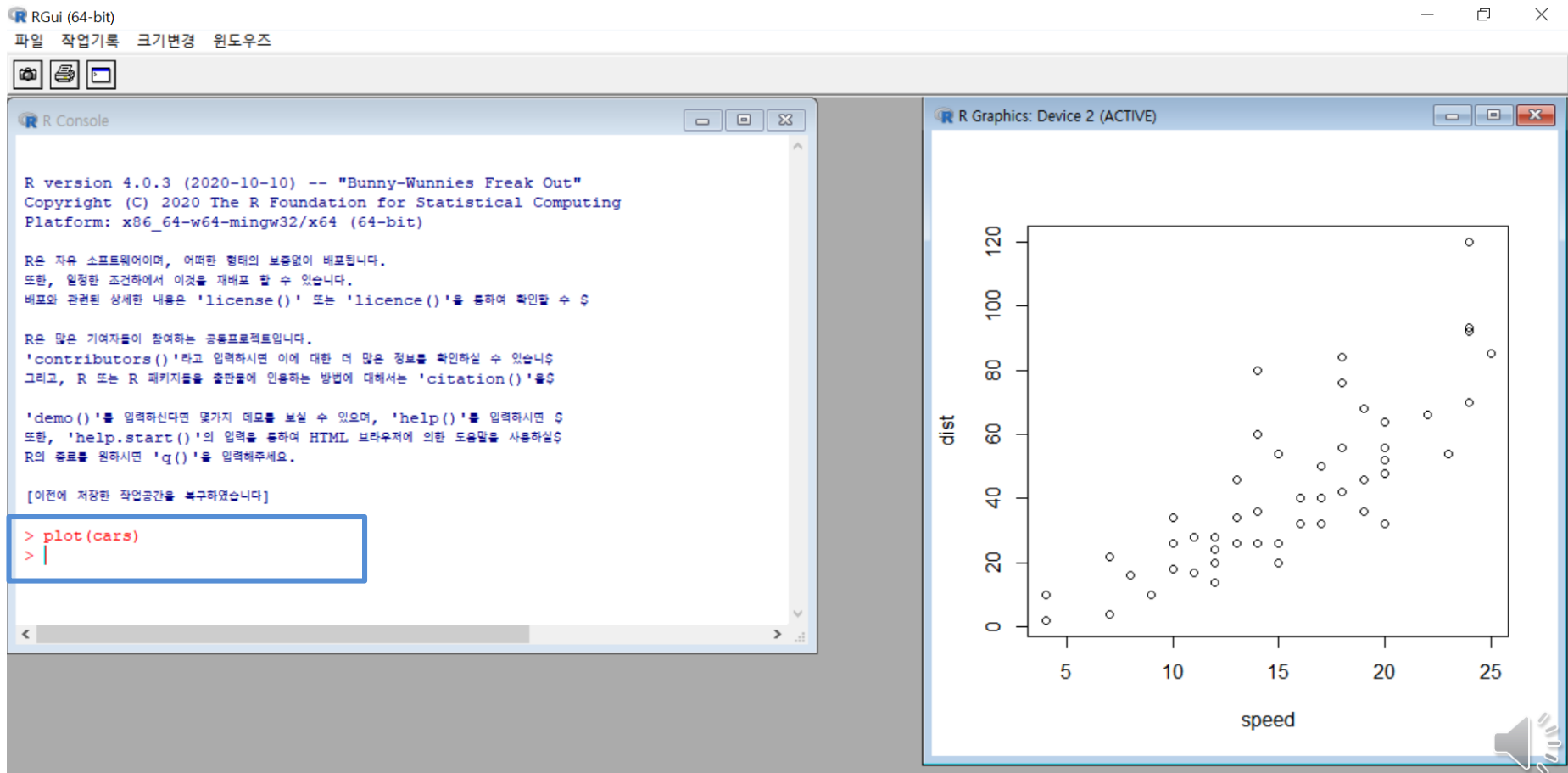
- 색깔 지정하는 옵션(매개변수) col, 축의 이름을 지정하는 xlab과 ylab, 기호 모양을 지정하는 pch

```
> plot(cars) # 그림 2-6(a)
> plot(cars, col='blue') # 그림 2-6(b)
> plot(cars, col='blue', xlab='속도', ylab='거리') # 그림 2-6(c)
> plot(cars, col='blue', xlab='속도', ylab='거리', pch=18) # 그림 2-6(d)
```



## 2.3 데이터 시각화 맛보기

### ■ 여러 가지 시각화 옵션을 적용(plot의 기본 실행)



## 2.3 데이터 시각화 맛보기

- 여러 가지 시각화 옵션을 적용(plot의 기본 실행 + 기호 색상 표시 방법)

RGui (64-bit)

파일 편집 보기 기타 패키지들 윈도우즈 도움말

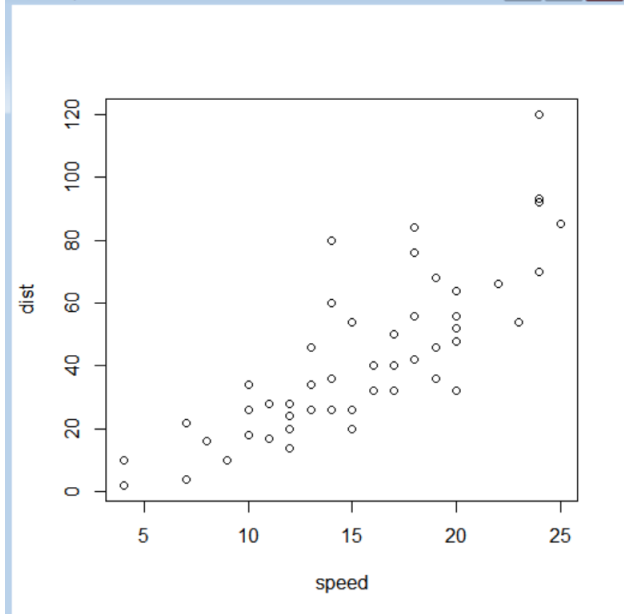


R Console

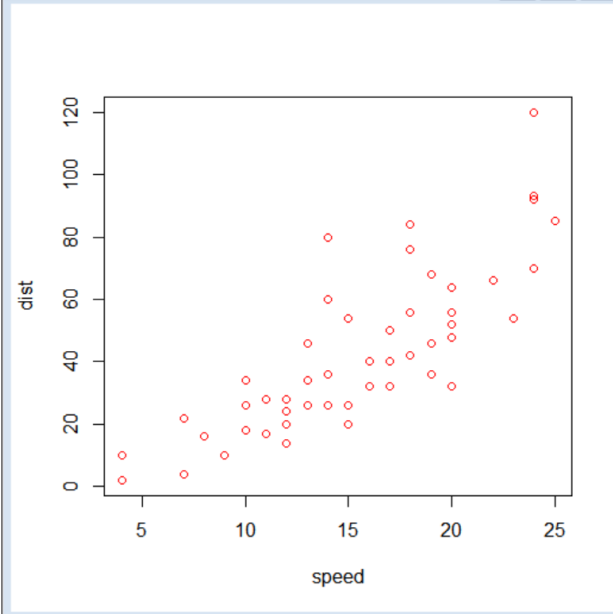
```
> plot(cars)
> plot(cars,col='red')
> plot(cars,col='blue')
> |
```

Ctrl + L

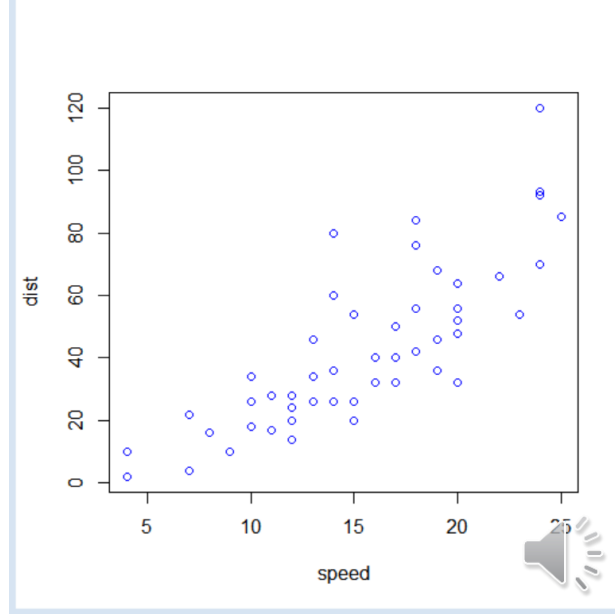
R Graphics: Device 2 (ACTIVE)



R Graphics: Device 2 (ACTIVE)



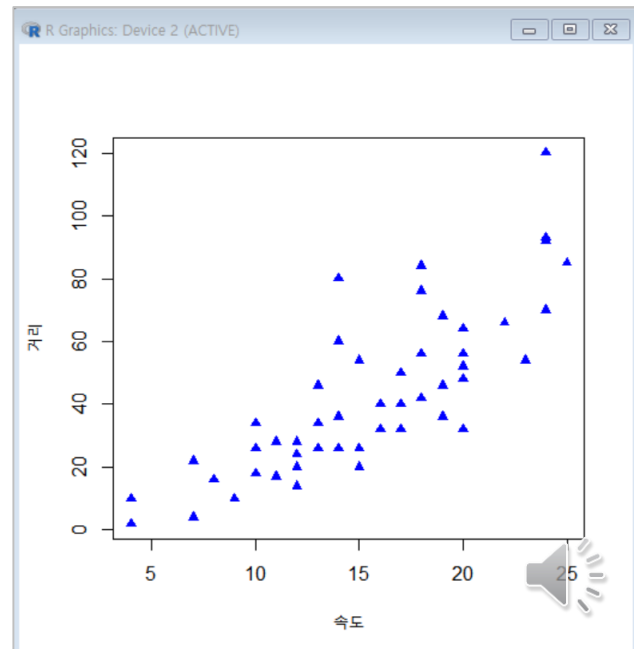
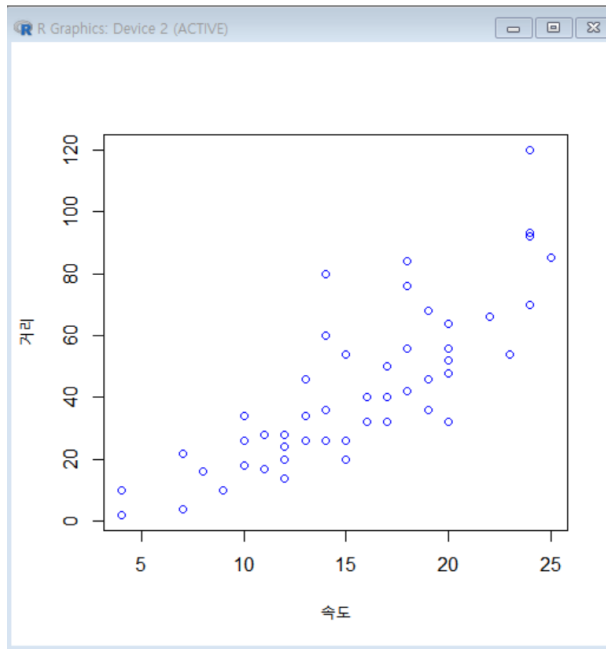
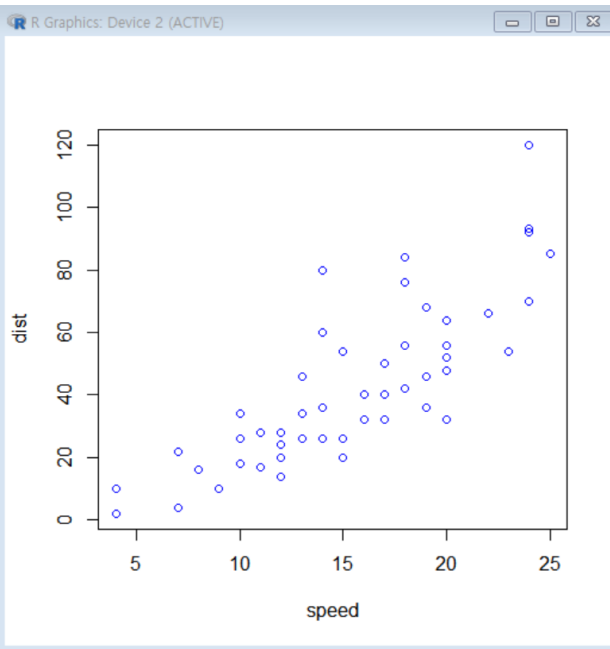
R Graphics: Device 2 (ACTIVE)



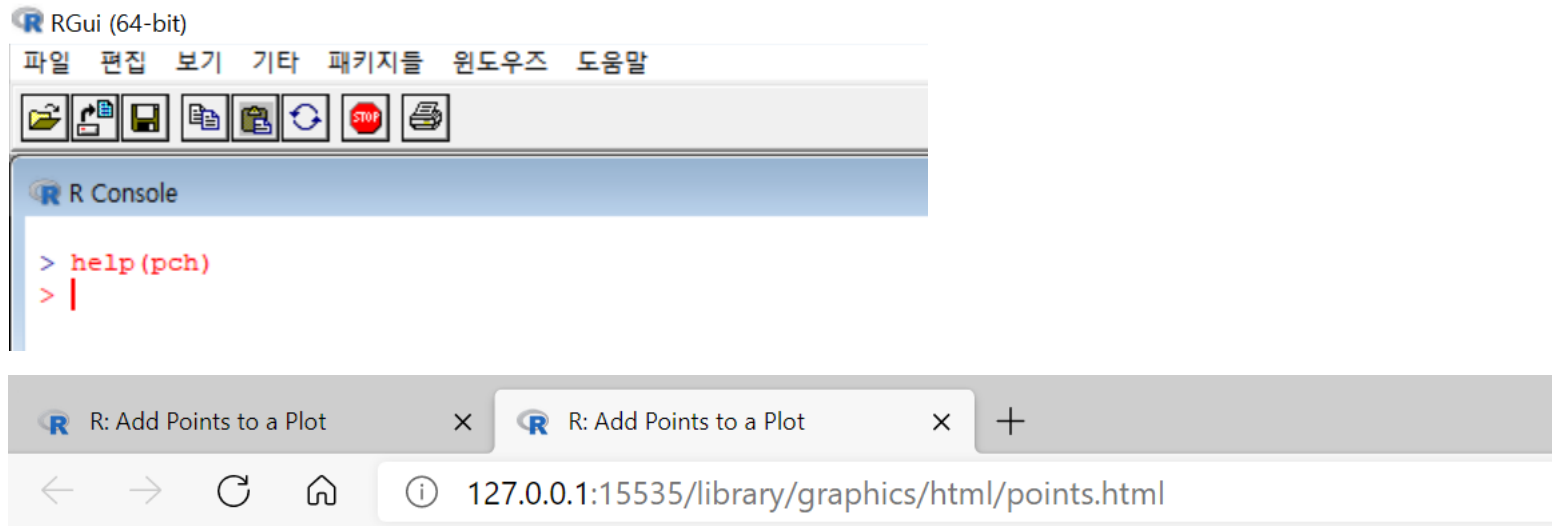
## 2.3 데이터 시각화 맛보기

- 여러 가지 시각화 옵션을 적용(plot의 기본 실행 + 기호 색상 표시 방법)  
+ x,y축 표시 변경, 기호 표시 방법

```
R Console
> plot(cars)
> plot(cars, col='blue')
> plot(cars, col='blue',xlab='속도',ylab='거리') # x,y 축 한글
> plot(cars, col='blue',xlab='속도',ylab='거리',pch=17) # x,y 축 한글, 기호 $
~ !
```



## 2.3 데이터 시각화 맛보기



포인트 {그래픽}

플롯에 점 추가

설명

`points` 지정된 좌표에서 점 시퀀스를 그리는 일반 함수입니다. 지정된 문자는 좌표를 중심으로 플롯됩니다.

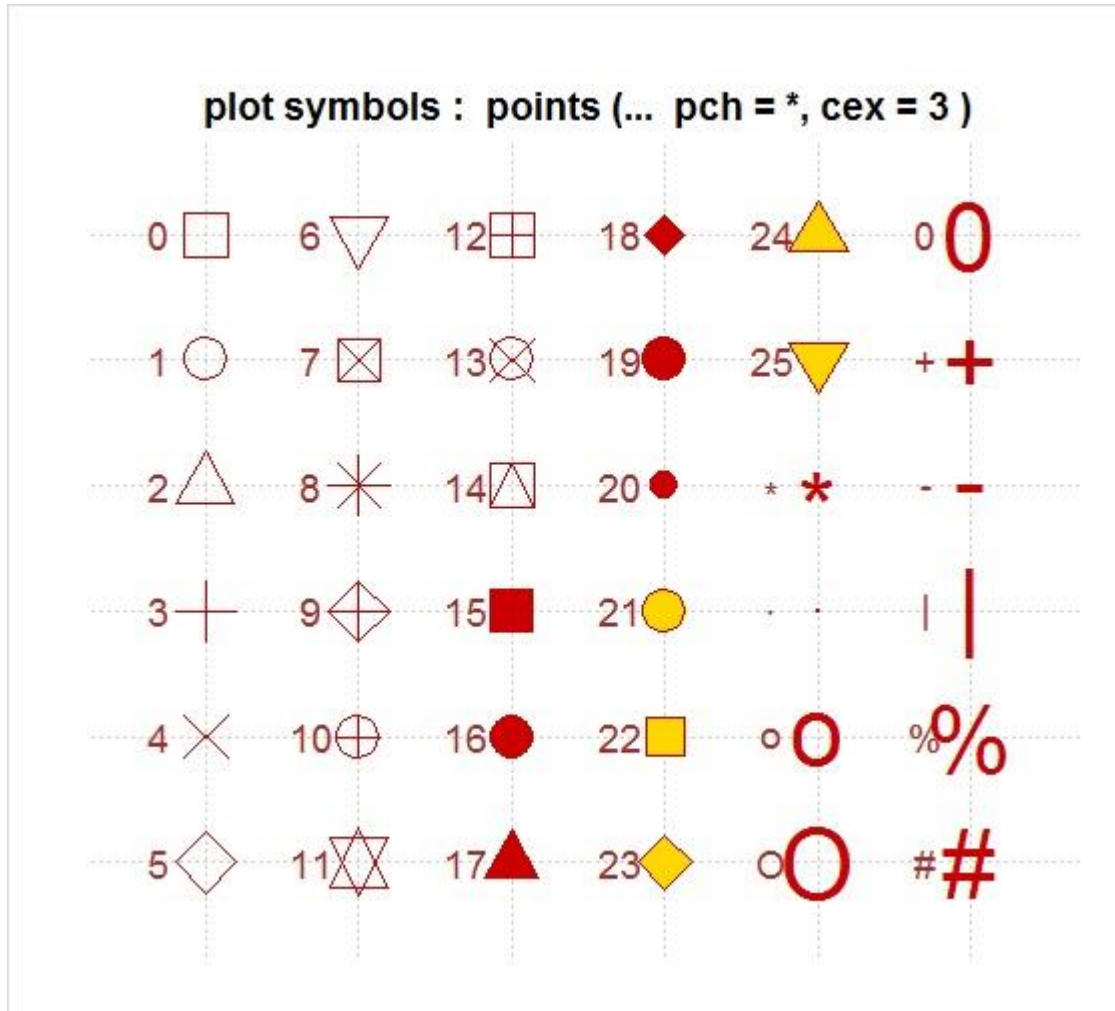
사용

`s`(팔각형사용)와 는 달리 기호를 사용하고 원을 사용합니다. 채워진 셰이프에는 테두리가 포함되어 있지 않습니다. 110131615:18

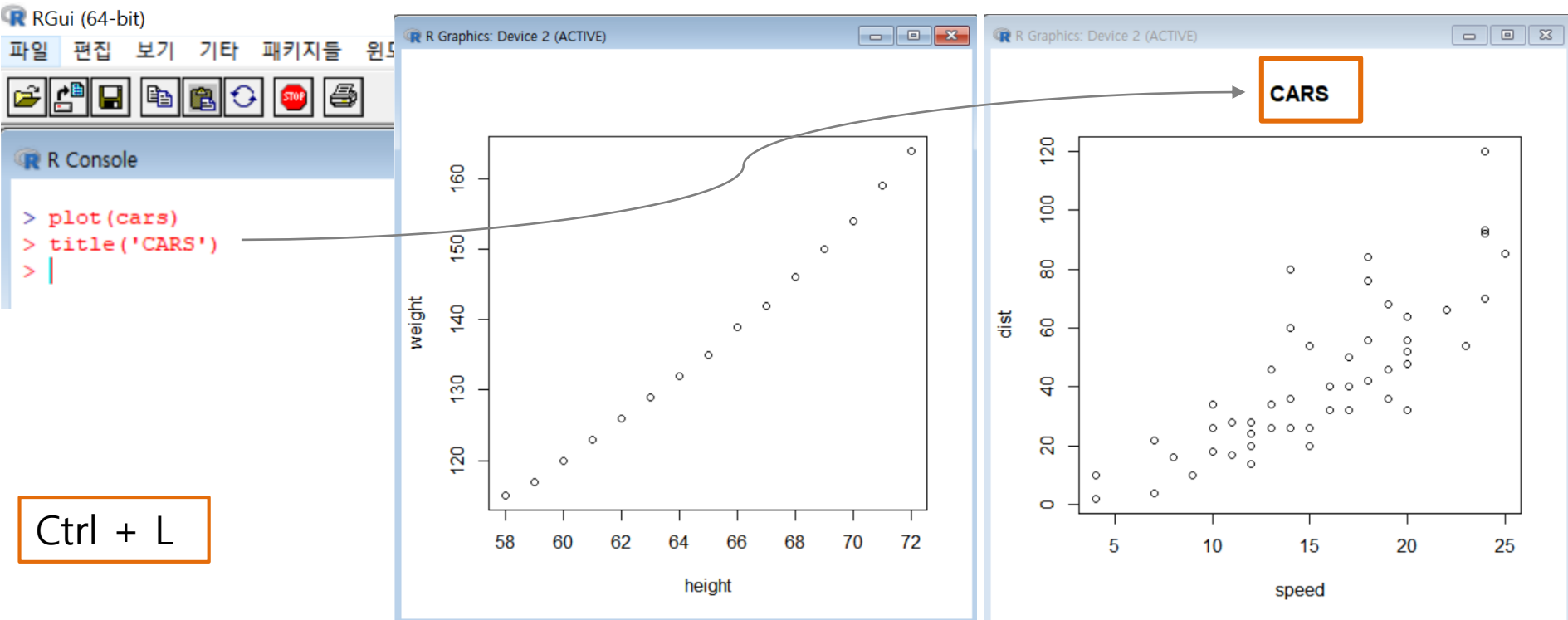


## 2.3 데이터 시각화 맛보기

기호 (plotting symbols, characters) : **pch=** , **cex=** , **lwd=**

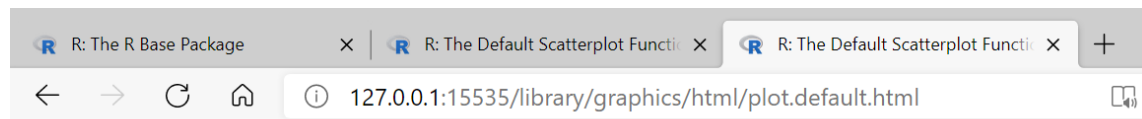
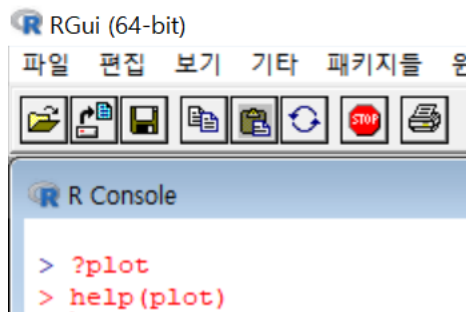


## 2.3 데이터 시각화 맛보기



## 2.4 데이터 과학 학습을 위한 좋은 습관 알아보기

### ① 도움말 청하기



플롯.기본 {그래픽}

기본 분산도 함수

설명

활성 그래픽 창에 축 및 제목과 같은 방식으로 분산 플롯을 그립니다.

사용

```
## Default S3 method:
plot(x, y = NULL, type = "p", xlim = NULL, ylim = NULL,
     log = "", main = NULL, sub = NULL, xlab = NULL, ylab = NULL,
     ann = par("ann"), axes = TRUE, frame.plot = axes,
     panel.first = NULL, panel.last = NULL, asp = NA,
     xgap.axis = NA, ygap.axis = NA,
     ...)
```

인수

x, y

및 인수는 플롯에 대한 x 및 y 좌표를 제공합니다. 좌표를 정의하는 모든 합리적인 방법은 허용됩니다. 별도로 제공된 경우 동일한 길이어야 합니다. [xy.coords](#)

주제 '플롯'에 대한 도움말은 다음 패키지에서 발견되었습니다.

[기본 분산도 함수](#)

(라이브러리 C:/프로그램 파일/R/R-4.0.3/라이브러리의 패키지)

[일반 X-Y 플롯](#)

(라이브러리 C:/PROGRA~1/R/R-40~1.3/라이브러리의 패키지)





## 2.4 데이터 과학 학습을 위한 좋은 습관 알아보기

### ② 익숙해지기

#### ■ 반복 학습의 중요성

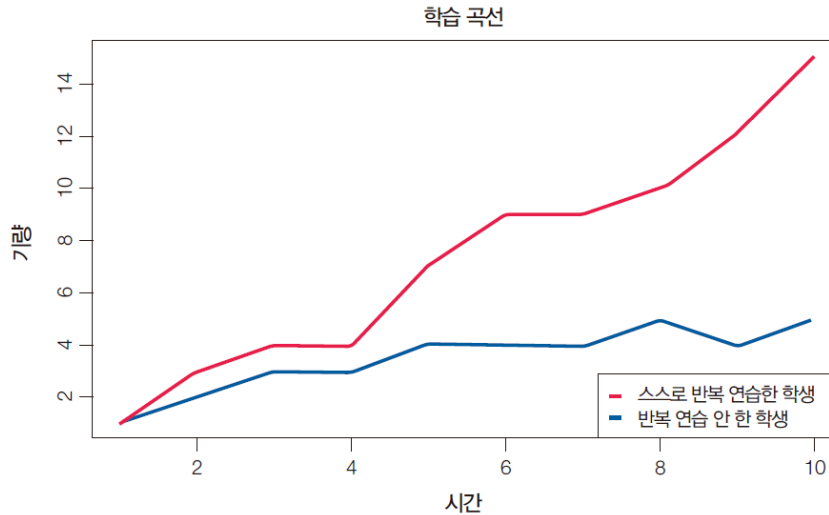
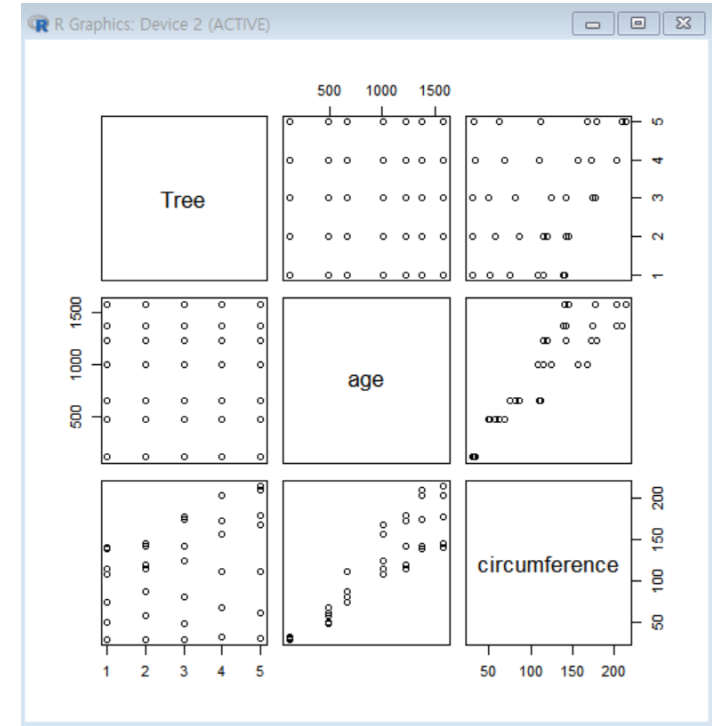


그림 2-9 데이터 과학을 공부하는 두 학생의 학습 곡선 비교

```
> plot(sleep)
> plot(Orange)
```

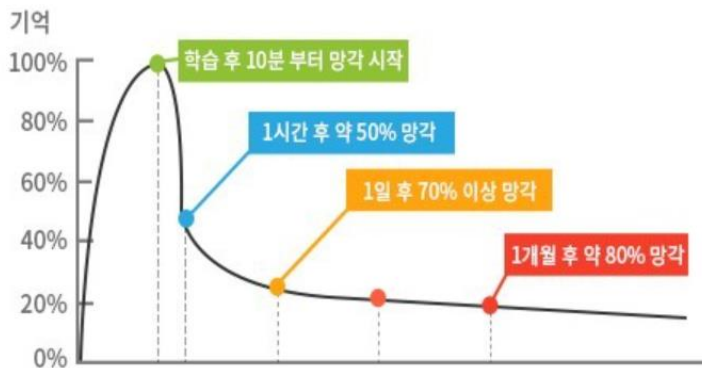


- 실전을 빨리 치러 봄: 다른 과목의 과제 수행 또는 프레젠테이션에 데이터 과학에서 배운 것 활용
- 각 절이 제공하는 연습문제 풀어보기
- 추가적인 데이터로 반복해 봄

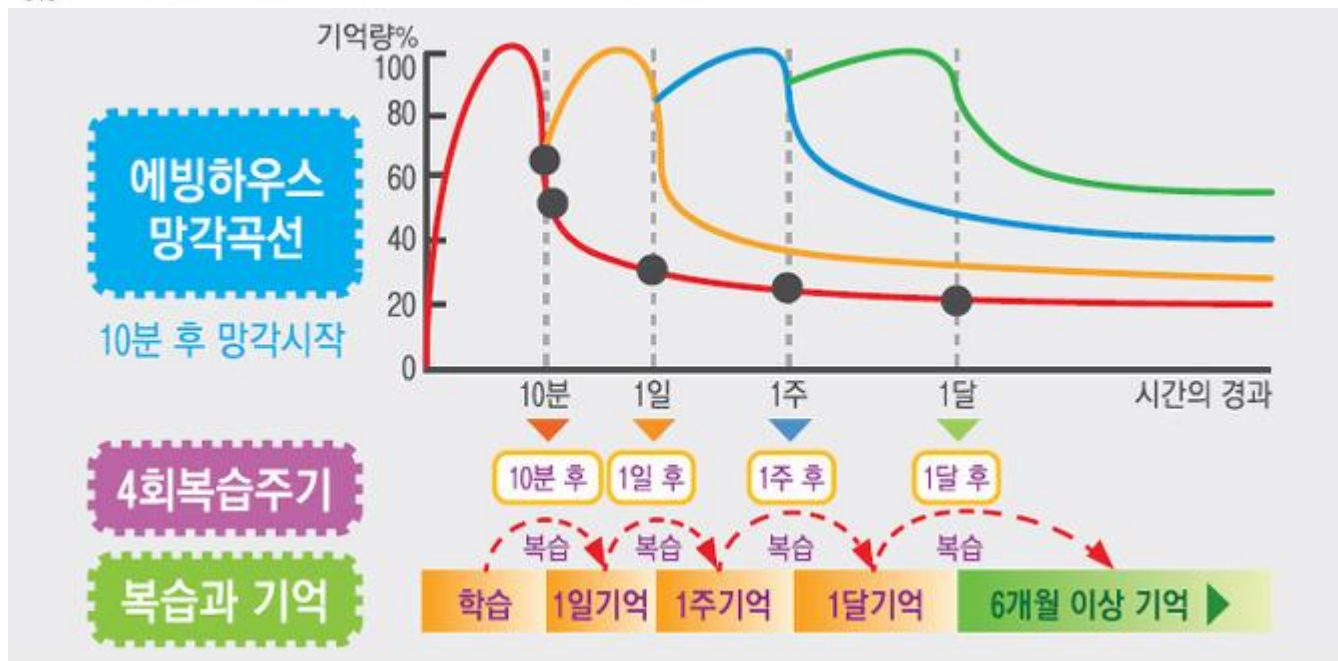


## 2.4 데이터 과학 학습을 위한 좋은 습관 알아보기

에빙하우스는 망각의 속도



사무자동화 교육 사례 : 엑셀, 파워포인트



### ③ 점증적으로 생각하기

- 가장 기본적인 기능을 만든 다음에 동작을 확인하고, 여기에 새로운 기능 하나를 추가해서 확인하고, 또 다른 기능 하나를 추가해서 확인하는 방식
  - 모든 것을 만든 다음에 확인하는 방식에서는 나중에 문제가 발생하면 어느 곳이 원인인지 찾기가 어려움
- [그림 2-6]의 예) 가장 기본적인 plot 함수를 확인하고, col 옵션을 추가해서 확인하고, xlab과 ylab 옵션을 추가해서 확인하고, pch 옵션을 추가해서 확인함


```
> plot(cars) # 그림 2-6(a)
> plot(cars, col='blue') # 그림 2-6(b)
> plot(cars, col='blue', xlab='속도', ylab='거리') # 그림 2-6(c)
> plot(cars, col='blue', xlab='속도', ylab='거리', pch=18) # 그림 2-6(d)
```

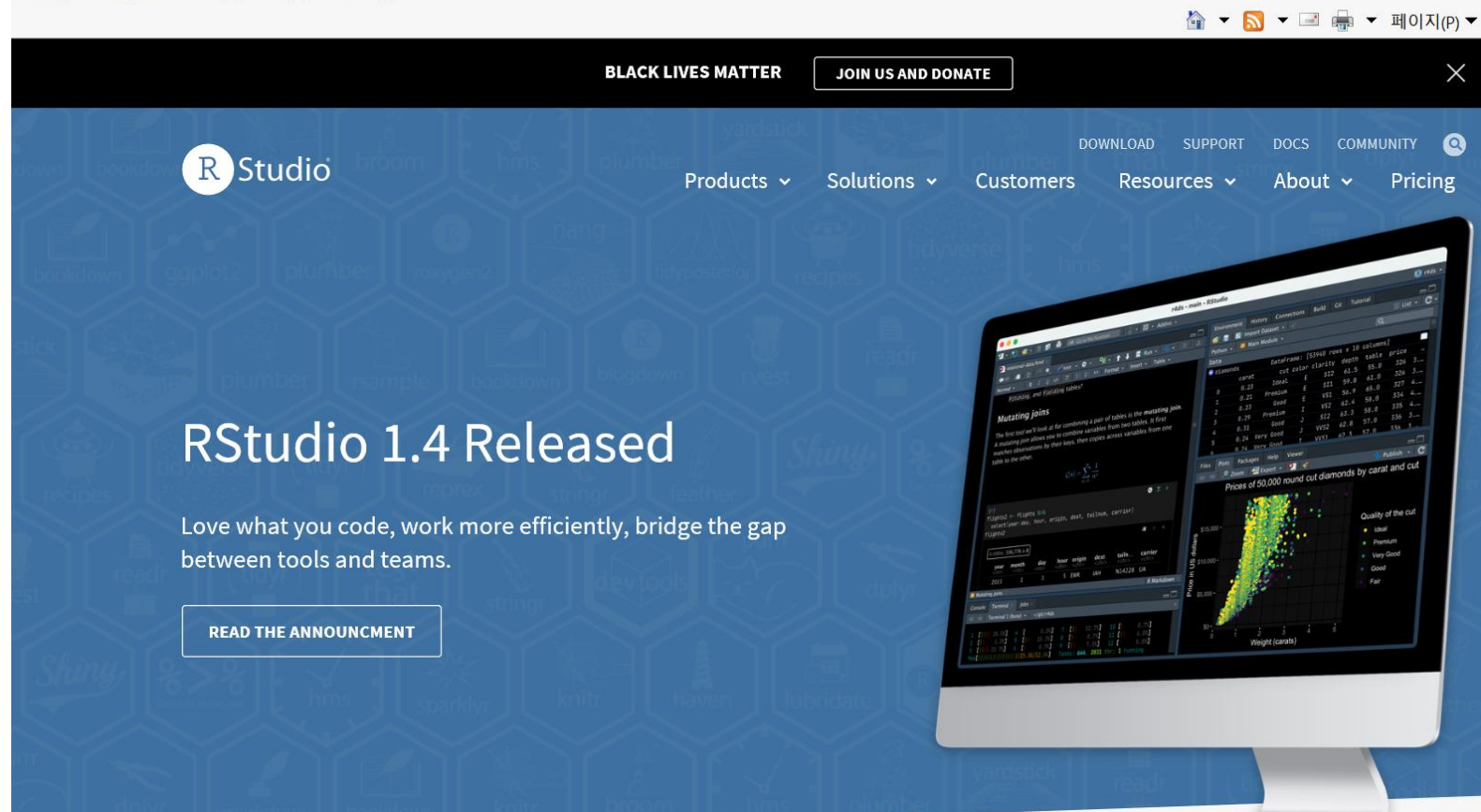


## 2.5 좋은 도구 익히기

### ① 통합 개발 환경

- 프로그래머가 편리하게 작업할 수 있게 해주는 통합 개발 환경 → IDE 또는 IDLE이라 부름
- R에서 가장 널리 쓰이는 IDE는 R 스튜디오 (<https://www.rstudio.com>)

 <https://rstudio.com/> RStudio | Open source & pr...  
편집(E) 보기(V) 즐겨찾기(A) 도구(T) 도움말(H) IDE(Integrated DeveLopment Environment)



## 2.5 좋은 도구 익히기

The image shows a screenshot of the RStudio website. A blue rectangular box highlights the main content area of the page, which includes the RStudio logo, navigation links, and the main heading 'RStudio'. A blue arrow points from the 'RStudio' link in the left sidebar to the main heading in the highlighted area.

**BLACK LIVES MATTER** [JOIN US AND DONATE](#)

RStudio

Products ^ Solutions v Customers R

**OPEN SOURCE**  
*Get started with R*

- RStudio**  
The premier IDE for R
- RStudio Server**  
RStudio anywhere using a web browser
- Shiny Server**  
Put Shiny applications online
- R Packages**  
Shiny, R Markdown, Tidyverse and more

[READ THE ANNOUNCEMENT](#)

**BLACK LIVES MATTER** [JOIN US AND DONATE](#)

RStudio

Products v Solutions v Cu

# RStudio

## Take control of your R code

RStudio is an integrated development environment (IDE) for R. It includes a code editor that supports direct code execution, as well as tools for plotting, history, and package management. [Click here to see more RStudio features.](#)

RStudio is available in **open source** and **commercial** editions and runs on Windows, macOS, Linux, or in a browser connected to RStudio Server or RStudio Server Pro (Docker and SUSE Linux).

## 2.5 좋은 도구 익히기

### Overview

- Quickly jump to function definitions
- View content changes in real-time with the Visual Markdown Editor
- Easily manage multiple working directories using projects
- Integrated R help and documentation
- Interactive debugger to diagnose and fix errors
- Extensive package development tools

### Support

Community forums only

### License

AGPL v3

### Pricing

Free

[DOWNLOAD RSTUDIO DESKTOP](#)

### RStudio Desktop

Open Source License

**Free**

[DOWNLOAD](#)

[Learn more](#)

### RStudio Desktop Pro

Commercial License

**\$995**

/year

[BUY](#)

[Learn more](#)

### RStudio Server

Open Source License

**Free**

[DOWNLOAD](#)

[Learn more](#)

### RStudio Server Pro

Commercial License

**\$4,975**

/year

(5 Named Users)

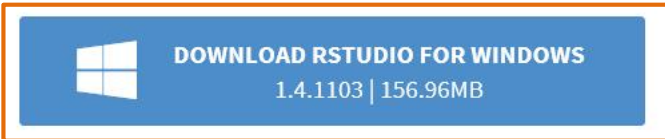
[BUY](#)

[Evaluation](#) | [Learn more](#)

## 2.5 좋은 도구 익히기

### RStudio Desktop 1.4.1103 - [Release Notes](#)

1. Install R. RStudio requires R 3.0.1+.
2. Download RStudio Desktop. Recommended for your system:



Requires Windows 10/8/7 (64-bit)



### All Installers

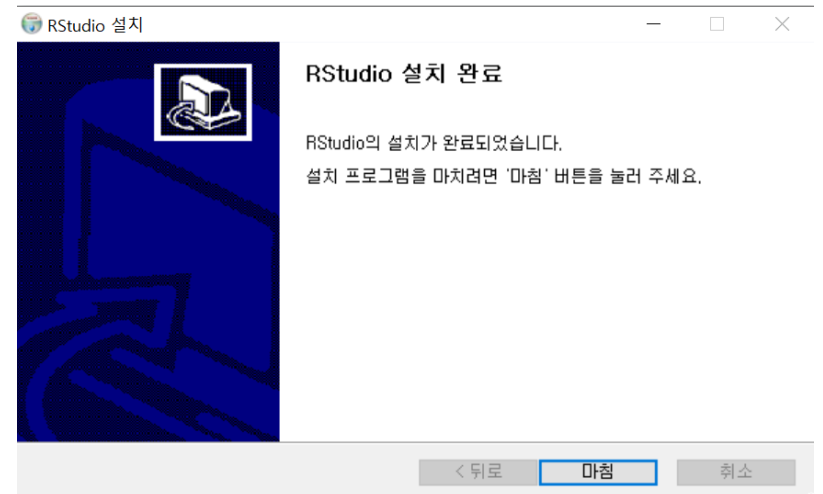
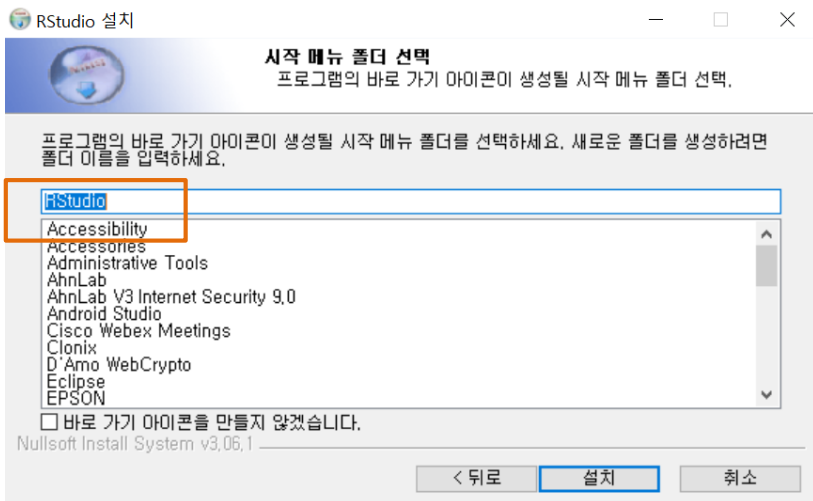
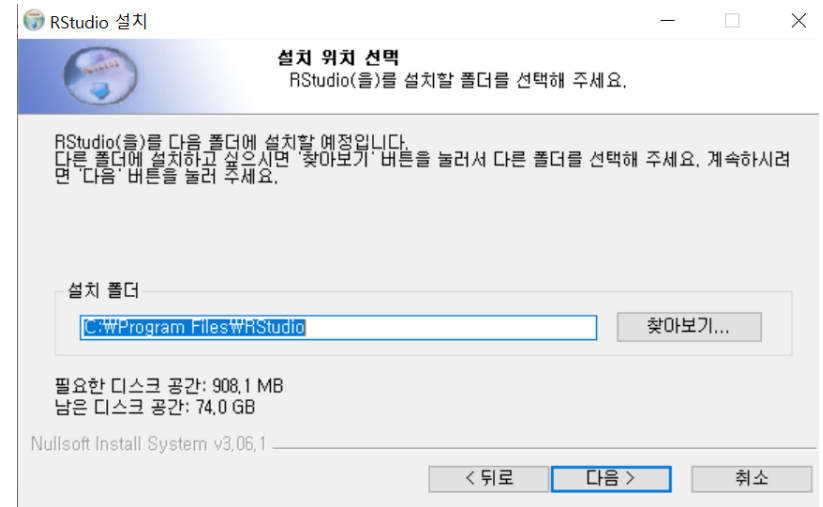
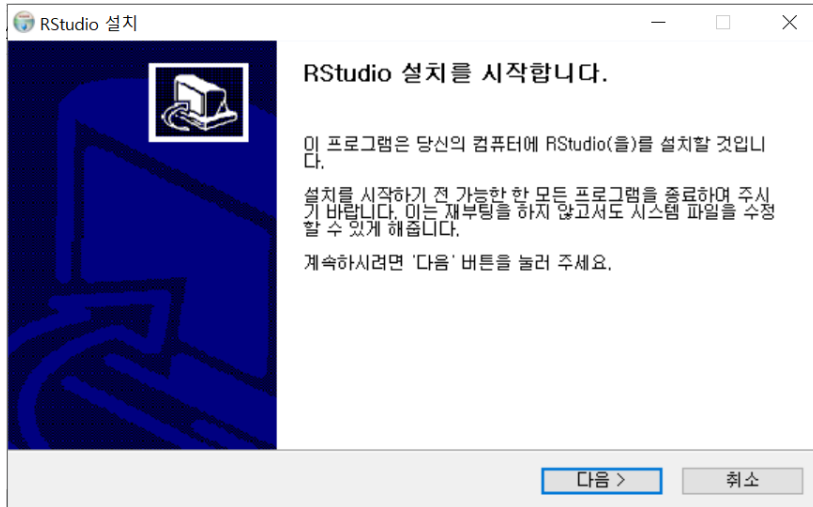
Linux users may need to [import RStudio's public code-signing key](#) prior to installation, depending on the operating system's security policy.

RStudio requires a 64-bit operating system. If you are on a 32 bit system, you can use an [older version of RStudio](#).

OS	Download	Size	SHA-256
----	----------	------	---------



## 2.5 좋은 도구 익히기





## 2.5 좋은 도구 익히기

### ■ R 스튜디오 개발 환경

- 콘솔 창과 스크립트 창
- 환경 창과 파일 창

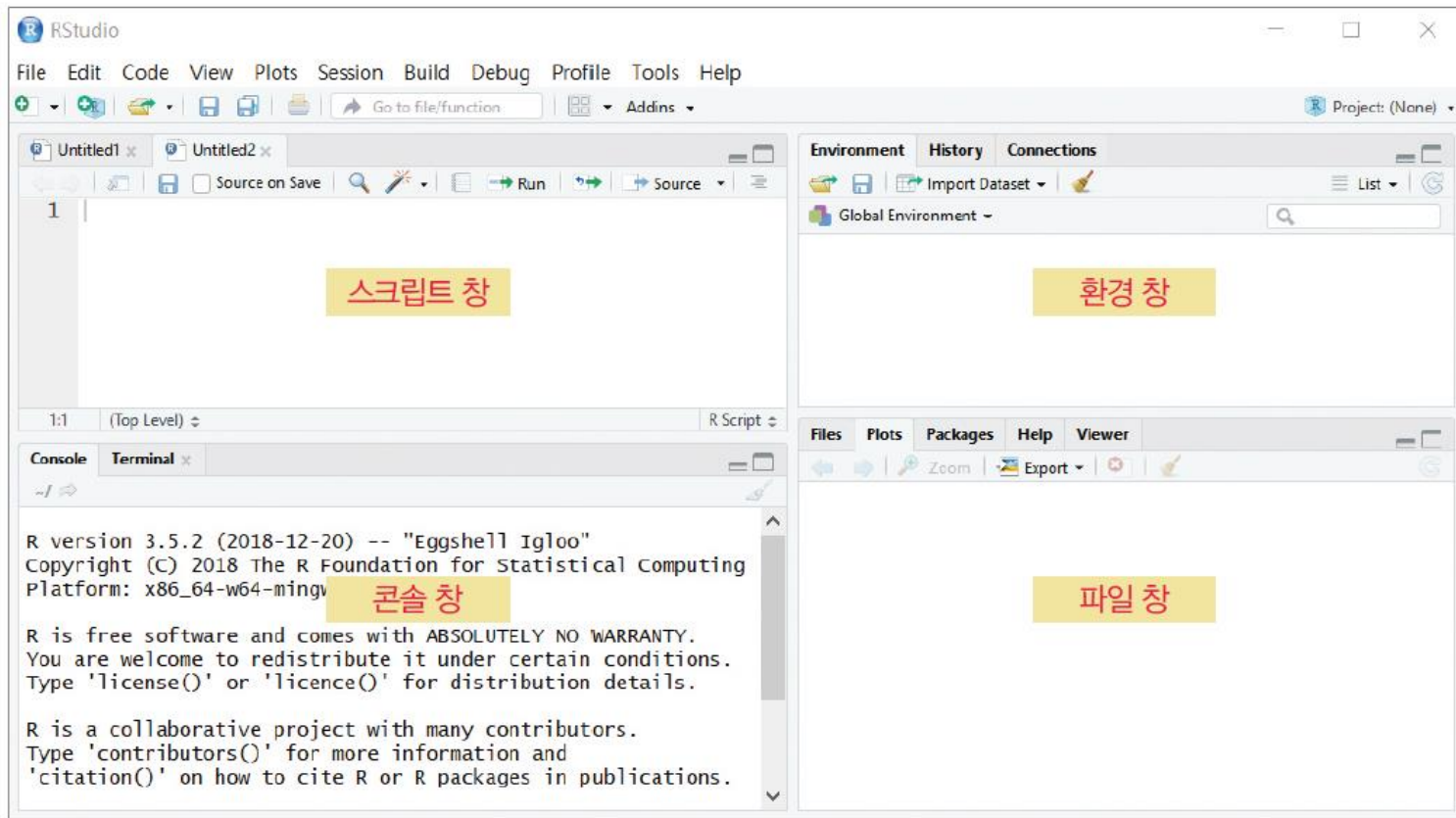


그림 2-11 R 스튜디오 개발 환경



## 4개의 창 익히기

The screenshot displays the RStudio IDE interface. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, and Help. Below the menu is a toolbar with icons for file operations and a search bar. The main workspace is divided into four panes:

- Source**: The top-left pane shows a script named 'Untitled1' with a single line of code '1'.
- Environment**: The top-right pane shows the 'Global Environment' with the message 'Environment is empty'.
- Console**: The bottom-left pane shows the R console output. It contains text about citing R and R packages, and instructions on how to use 'demo()', 'help()', 'help.start()', and 'q()'.
- Files**: The bottom-right pane shows a file explorer view of the 'Home' directory. It lists files and folders including '.Rhistory', '5. [사양] 2019년도 컨설팅 우수사례...', '7. [경영] 2019년도 컨설팅 우수사례...', '그림', and '그림.zip'.

A red box highlights the 'Run' button (a green play icon) in the Source pane toolbar.

## 4개의 창 익히기

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Project: (None)

R data sets 2chapter.R\* elec\_gen\_use

Source on Save Run Source

```

1 setwd('c:/Rsources')
2
3 getwd()
4 Chickweight
5 dim(Chickweight)
6 names(Chickweight)
7 rownames(Chickweight)
8
9 library(dplyr)
10 library(ggplot2)
11
12 cars
13 plot(cars)
14 plot(cars, col='blue')
15 plot(cars, col='blue', xlab='속도', ylab='거리') # x,y 축
16 plot(cars, col='blue', xlab='속도', ylab='거리', pch=17) #
17
18

```

15:1 (Top Level) R Script

Console

Environment History Connections Tutorial

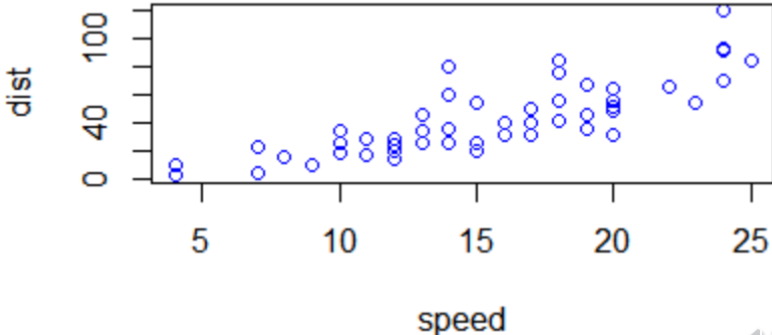
Import Dataset

R Global Environment

elec_gen_use	2310 obs. of 4 variables
elec_gen_use...	750 obs. of 4 variables

Files Plots Packages Help Viewer

Zoom Export



dist

speed

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Project: (None)

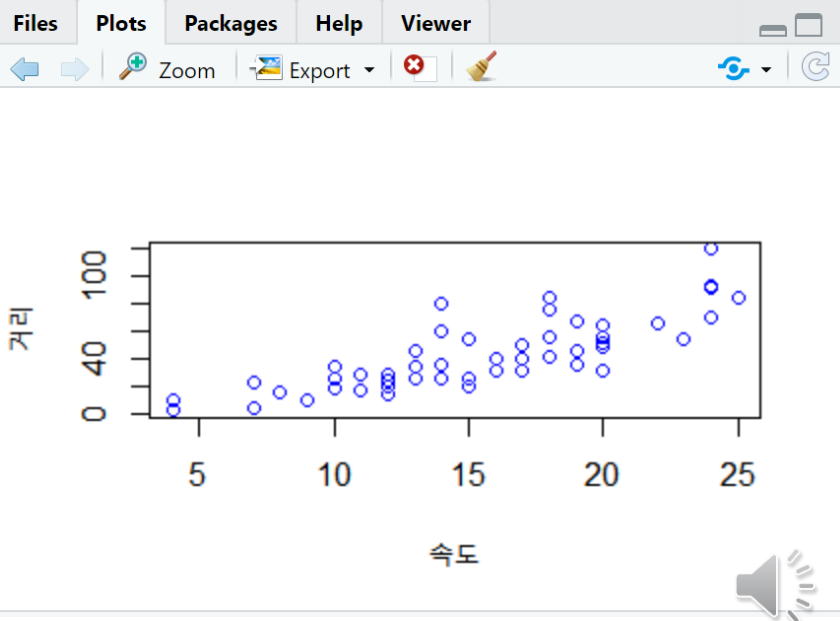
```
8  
9 library(dplyr)  
10 library(ggplot2)  
11  
12 cars  
13 plot(cars)  
14 plot(cars, col='blue')  
15 plot(cars, col='blue', xlab='속도', ylab='거리') # x,y  
16 plot(cars, col='blue', xlab='속도', ylab='거리', pch=17)  
17  
18
```

15:5 (Top Level) R Script

Console C:/RSources/

```
> library(dplyr)  
> library(ggplot2)  
> plot(cars)  
> View(elec_gen_use)  
> plot(cars, col='blue', xlab='속도', ylab='거리') # x,y 축 한글  
>
```

Environment	History	Connections	Tutorial
Import Dataset			
Global Environment			
elec_gen_use	2310 obs. of 4 variables		
elec_gen_use...	750 obs. of 4 variables		



# Thank you

