



# 3주차: R의 데이터형과 연산

**ChulSoo Park**

School of Computer Engineering & Information Technology

Korea National University of Transportation



# 학습목표 (3주차)

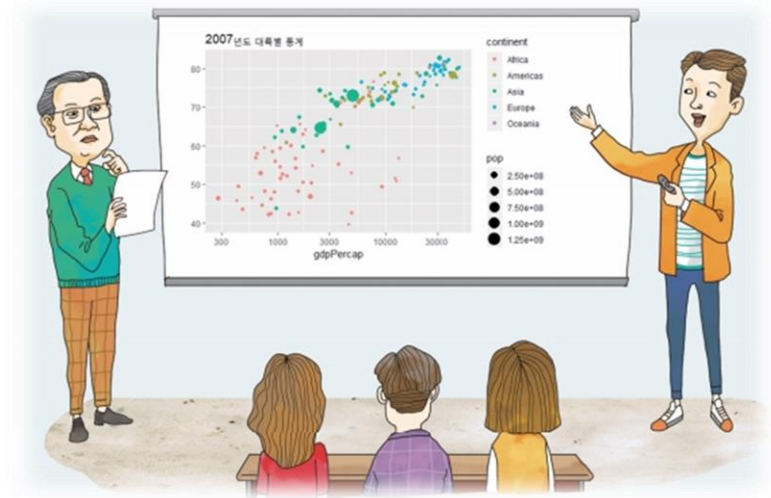
- ❖ R에서 데이터의 처리와 저장
- ❖ 변수, 연사자, 벡터, 행렬 등 Data 형태에 대한 이해
- ❖ R활용 문법 숙지



# 03

## CHAPTER

# R의 데이터형과 연산



## CONTENTS

- 3.1 데이터 저장과 처리
- 3.2 변수
- 3.3 데이터형
- 3.4 연산자
- 3.5 벡터
- 3.6 배열(행렬)
- 3.7 데이터 프레임
- 3.8 리스트
  - 요약



## 1. R과 RStudio를 설치

The screenshot displays the RStudio interface with the following components:

- Source Editor:** Contains R code for a linear regression model.
 

```

22 m=lm(y ~ x) # 모델 적합(학습)
23 m
24
25 plot(x,y)
26 abline(m, col='blue')
27
28 loef(m) # 매개변수(계수) 값을 알려줌
29 fitted(m) # 훈련 집합에 대한 예측값
30 residuals(m) # 잔차를 알려줌
31 deviance(m)/length(x) # 잔차 제곱합을 평균 제곱 오차로 계산
32
28:1 (Top Level) R Script
      
```
- Environment:** Shows the Global Environment with objects 'm' (List of 12) and 'tmp1' (338 obs. of 14 variables).
- Console:** Displays the output of the R script, including masked objects and the execution of the model and plot commands.
 

```

filter, lag

The following objects are masked from 'package:base':

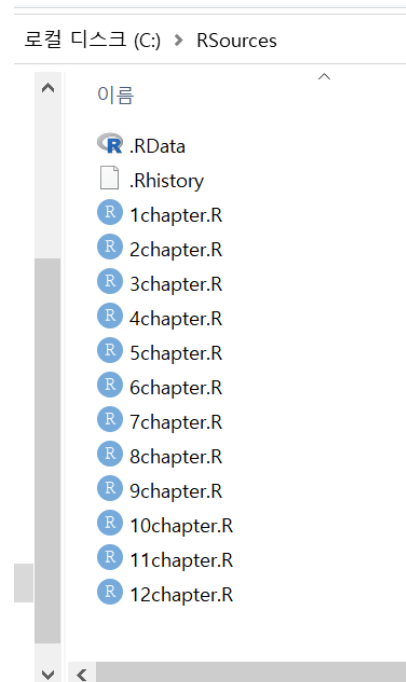
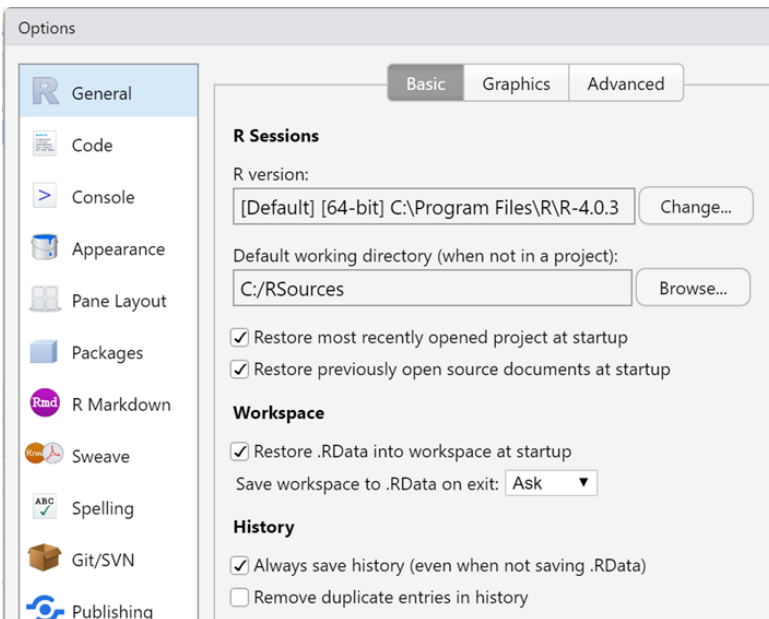
intersect, setdiff, setequal, union

> library(ggplot2)
> library(RColorBrewer)
> x=c(3.0,6.0,9.0,12.0) # 설명 변수 x
> y=c(3.0,4.0,5.5,6.5) # 반응 변수 y
> m=lm(y ~ x) # 모델 적합(학습)
> plot(x,y)
> abline(m, col='blue')
>
      
```
- Plots:** A scatter plot of y vs x with a blue regression line. The x-axis ranges from 4 to 12, and the y-axis ranges from 3.0 to 6.0. Data points are at (3, 3), (6, 4), (9, 5.5), and (12, 6.5).

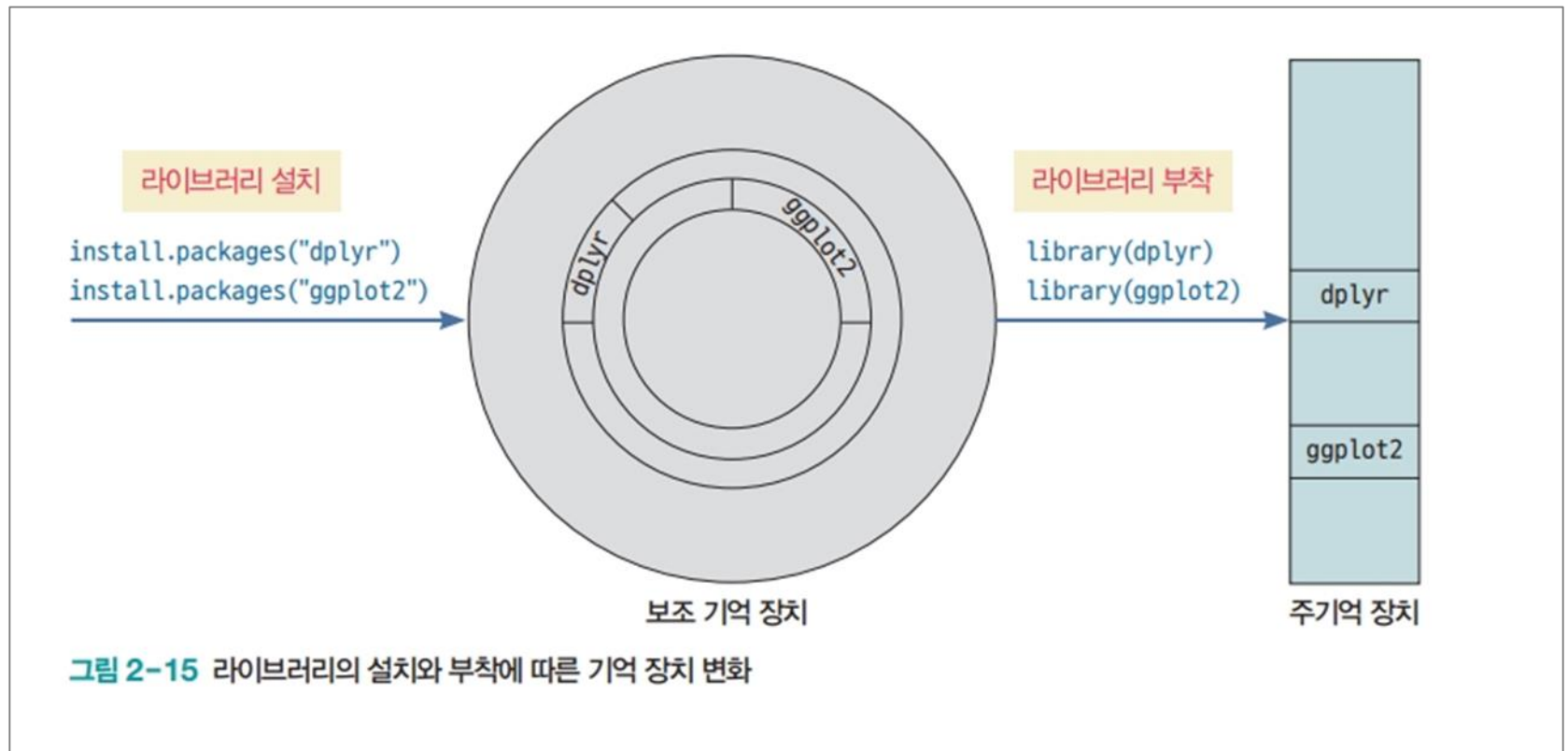
## 2. 작업 디렉토리의 지정

```
Console c:/Rsources/ ↗
> setwd('c:/Rsources')
> getwd()
[1] "c:/Rsources"
```

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



## 3. 라이브러리(패키지)의 활용



## 4. 데이터와 친해지기(data 확인)

RGui (64-bit)

파일 편집 보기 기타 패키지들 윈도우즈 도움말

R Console

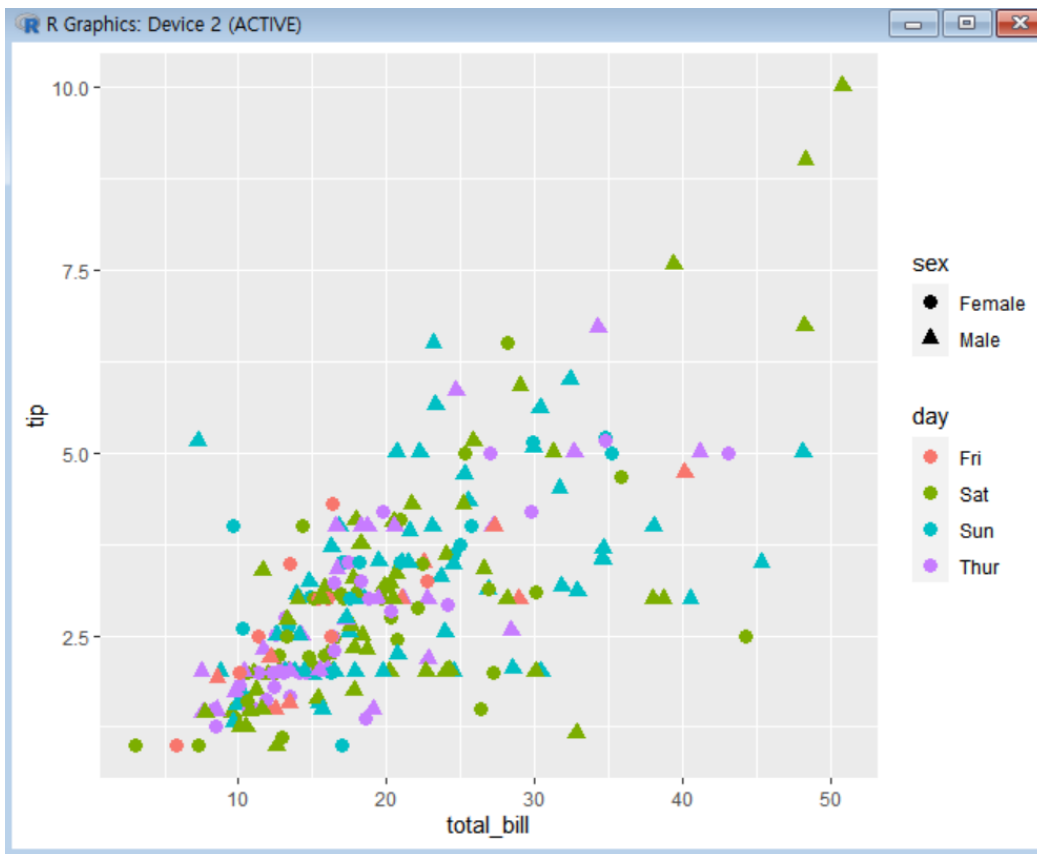
```
> tips=read.csv('https://raw.githubusercontent.com/mwaskom/seaborn-data/master/tips.csv')
> str(tips)
'data.frame': 244 obs. of 7 variables:
 $ total_bill: num 17 10.3 21 23.7 24.6 ...
 $ tip       : num 1.01 1.66 3.5 3.31 3.61 4.71 2 3.12 1.96 3.23 ...
 $ sex       : chr "Female" "Male" "Male" "Male" ...
 $ 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
> |
```



# Review

## 5. 패키지를 활용한 그래프 그리기

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





- 데이터 과학의 문법에 대해 학습해 보자.
  - 변수: 데이터 저장 공간
  - 데이터형: 숫자형, 문자형, 범주형, 논리형, 특수 상수 등
  - 연산자: 산술, 비교, 논리 연산자
  - 벡터: 단일값들의 모임
  - 배열: 열과 행을 가지는 데이터 집합. 벡터의 요소들이 다시 벡터로 구성된 형태.
  - 데이터 프레임: 서로 다른 데이터 형이 표 형태로 정리된 구조. 각 속성의 크기가 같음.
  - 리스트: 데이터 프레임과 유사한 표 형태의 구조. 각 속성의 크기가 달라도 됨.



출처 : [www.ncs.go.kr](http://www.ncs.go.kr)

## NCS 및 학습모듈검색

NCS 및 학습모듈 검색

이전 NCS및학습모듈('15년)

직업계고 교육과정

직업기초능력

구 사이트자료(구NCS)

구 사이트자료(구모듈교재)

## NCS 및 학습모듈 검색

HOME / NCS 및 학습모듈검색 / NCS 및 학습모듈 검색

NCS 분류보기 메뉴따라하기

분야별검색

키워드검색

코드검색

< 이전화면

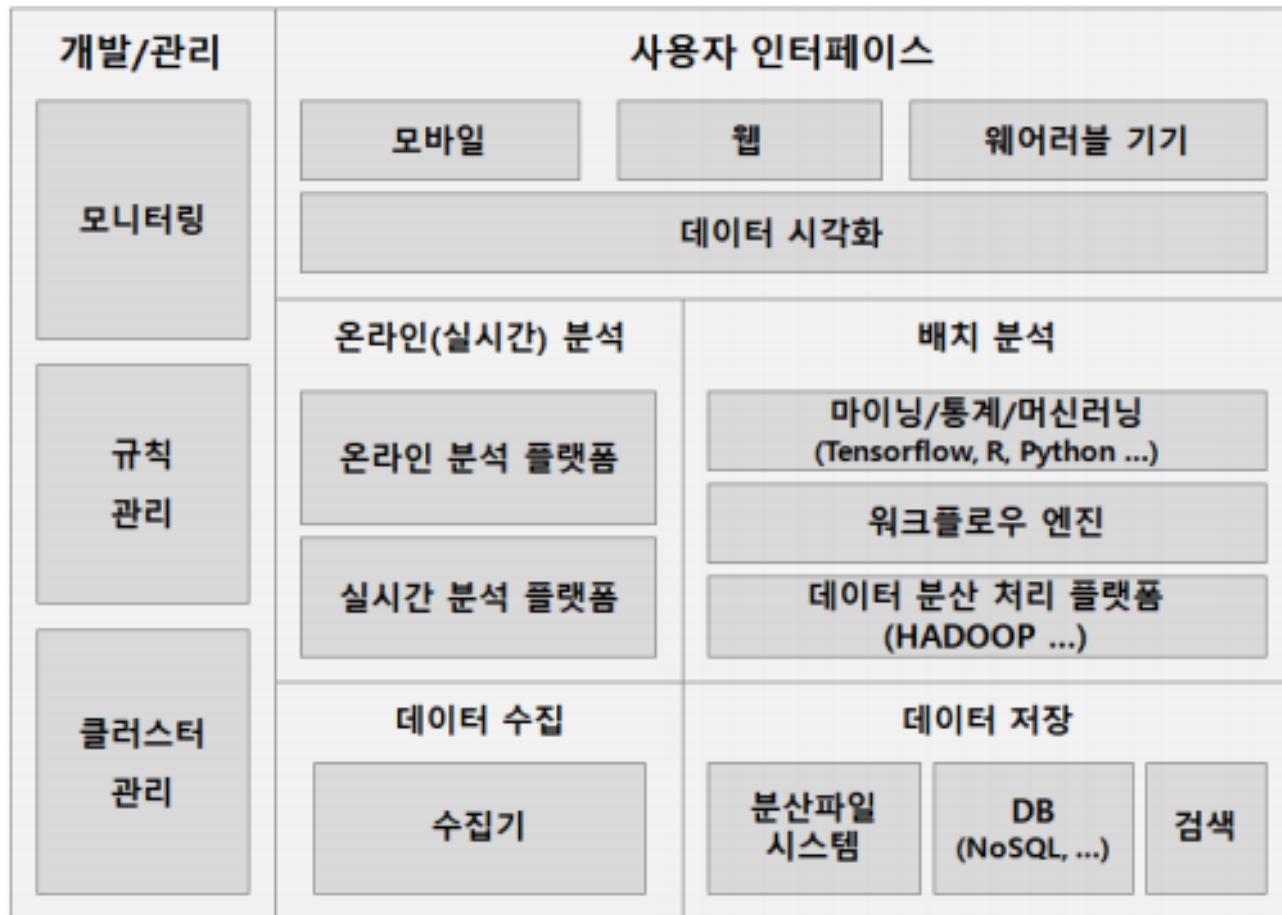
### 20. 정보통신

중분류	소분류	세분류	능력단위
01. 정보기술	01. 정보기술전략·계획	01. 정보기술전략	01. 빅데이터 서비스 기획
02. 통신기술	02. 정보기술개발	02. 정보기술컨설팅	02. 빅데이터 환경분석
03. 방송기술	03. 정보기술운영	03. 정보기술기획	03. 빅데이터 분석 기획
	04. 정보기술관리	04. SW제품기획	04. 빅데이터 기술 플랫폼 기획
	05. 정보기술영업	05. 빅데이터 분석	05. 빅데이터 성과 관리 기획
	06. 정보보호	06. IoT융합서비스기획	
	07. 인공지능	07. 빅데이터기획	



# Preview

출처 : [www.ncs.go.kr](http://www.ncs.go.kr):정보통신>정보기술>정보기술 전략.계획>빅데이터 기획>빅데이터 환경 분석



[그림 1-5] 서비스 모델 관련한 빅데이터 기술 플랫폼 분석 사례



## 데이터사이언스 대학원을 1학기를 마치고(2020.07.18)

데이터사이언스학과라는 특성상, 기존 대학의 전기과, 컴퓨터공학과, 통계학과 (바이오 포함), 인문대학 언어학과(NLP) 등 다양한 백그라운드의 교수님들이 한 데 모여 계신다.

학문의 융합이라는 측면에서 벽이 없는 자유로운 대화와 협업이 이를 가속화 할 수 있기에 이러한 교수진 체제는 매우 바람직하다고 생각된다. 특히 학문의 융합이라는 기치로 탄생한 자유전공학부 출신이 많은 교수진 체제인 경계를 허무는 것보다 오히려 어느정도 학과 단계를 가지고 있는 단계에서, 아는 것과 모르는 것을 아날 수 있다는 점에서 교수진과 이 대학원이 된다고 생각한다.

### 수업

대학원 개설이후 처음으로 열린 이번학기에는

- 기초 통계
- 기초 프로그래밍
- 빅데이터 및 데이터베이스
- 소프트웨어 플랫폼
- 머신러닝 / 딥러닝
- 세미나 수업

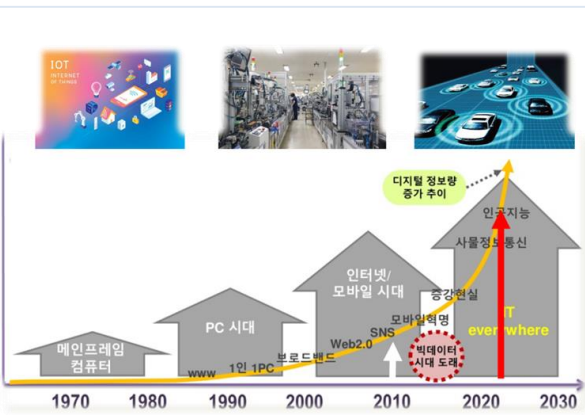


## 3.1 데이터의 저장과 처리

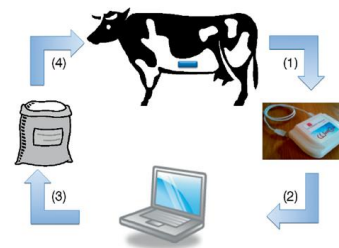
- 데이터를 저장하고 연산 등 처리하려면 문법 공부는 필수
  - $a=1$
  - $b=2$
  - $c=a+b$
- 학생 성적 처리 등 데이터가 많은 경우
  - 단일 변수로 표현하기에는 너무 많은 변수가 필요함.(student1=90, student2=100, ....)
  - 벡터, 행렬, 데이터 프레임, 리스트 등을 활용하면 하나의 변수명으로 많은 데이터 저장 가능 students=c(90,100,95,.....)
  - 우리 주변에서 보는 출석부, 성적 관리부, 회원 관리부 등이 데이터를 관리하기 편한 표 형태로 구성되어 있음.



### 3.1 데이터의 저장과 처리

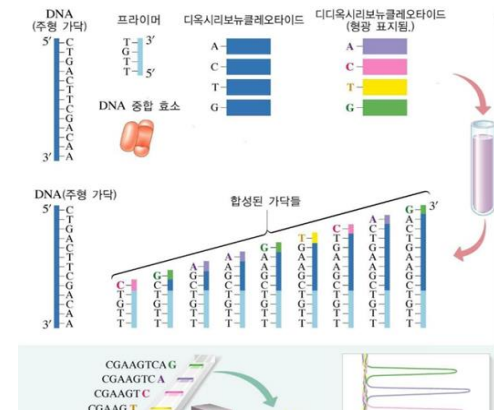


### The 'CowNet' System



- (1) Once the Well Cow Smart Sensor has been ingested it starts relaying information wirelessly to the Smart Reader
- (2) The information can then be viewed on a phone or computer for continual "on the spot" monitoring
- (3) Analysis of the data using Biopara-Milk informs feed management
- (4) Cow health and feed efficiency are improved through diet optimisation

미국 지놈 분석 비용 (단위=달러)



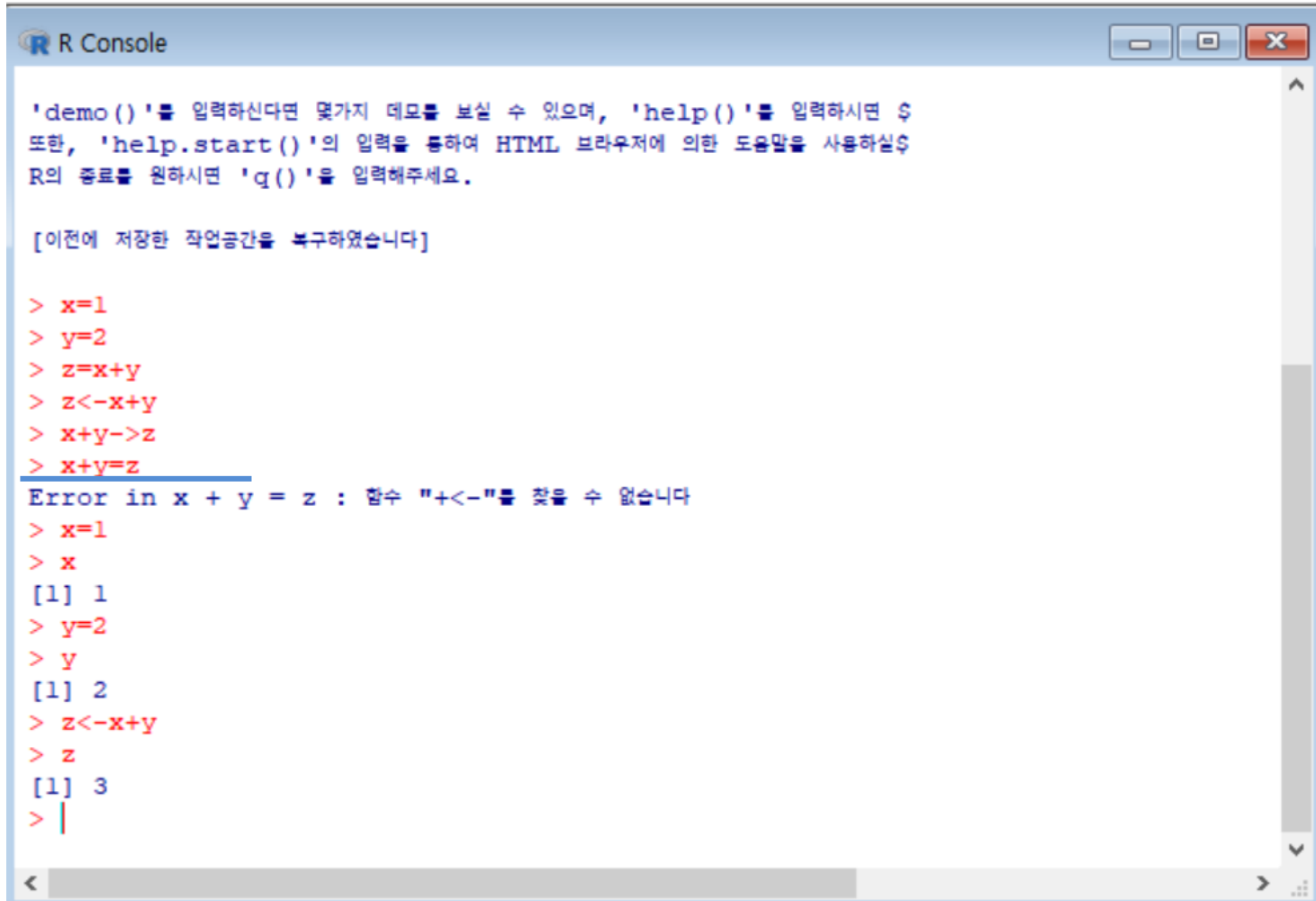
❖ 공장별 Point(Tag)

TAG 유형	TAG 유형_DESC	제일두부	제일생면	열가리품	제일생면
AI	산도, 악물놀이, 음도, PH	-	15	-	-
AL	Alarm	13	-	-	-
BI	Brix	-	3	-	-
CT	Cycle Time	3	5	-	-
DE	두유농도	3	-	-	-
FT	가수할, 매두투입공간유량, 두유생산량, 용고제투입량	9	-	-	-
HI	습도	-	3	3	-
PI	Reset	3	-	-	-
PI	진공압	-	5	-	-
PN	제품 Selector	16	51	32	13
QT	박산량	28	47	38	28
SL	제품 Selector(TOPIA)	3	-	-	-
ST	가동상태(급속검출기)	3	4	8	8
TI	로도	45	47	18	4
WI	유질량	-	-	-	-
WT	On/Off, 비가량, Start/Stop/종료시간	16	42	30	25
	합계	142	226	129	101



## ① 변수 사용법

- `=`, `<-`, `->`을 이용한 값 대입에서 차이점



```
R Console

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

[이전에 저장한 작업공간을 복구하였습니다]

> x=1
> y=2
> z=x+y
> z<-x+y
> x+y->z
> x+y=z
Error in x + y = z : 함수 "+<-"를 찾을 수 없습니다
> x=1
> x
[1] 1
> y=2
> y
[1] 2
> z<-x+y
> z
[1] 3
> |
```



## 3.2 변수

### ■ 두 값을 교환하는 예제

- 임시 저장공간을 만들어 미리 하나의 값을 빼 두어야 함.
- 위에서 아래로 순차적으로 프로그램이 실행되는 구조임.

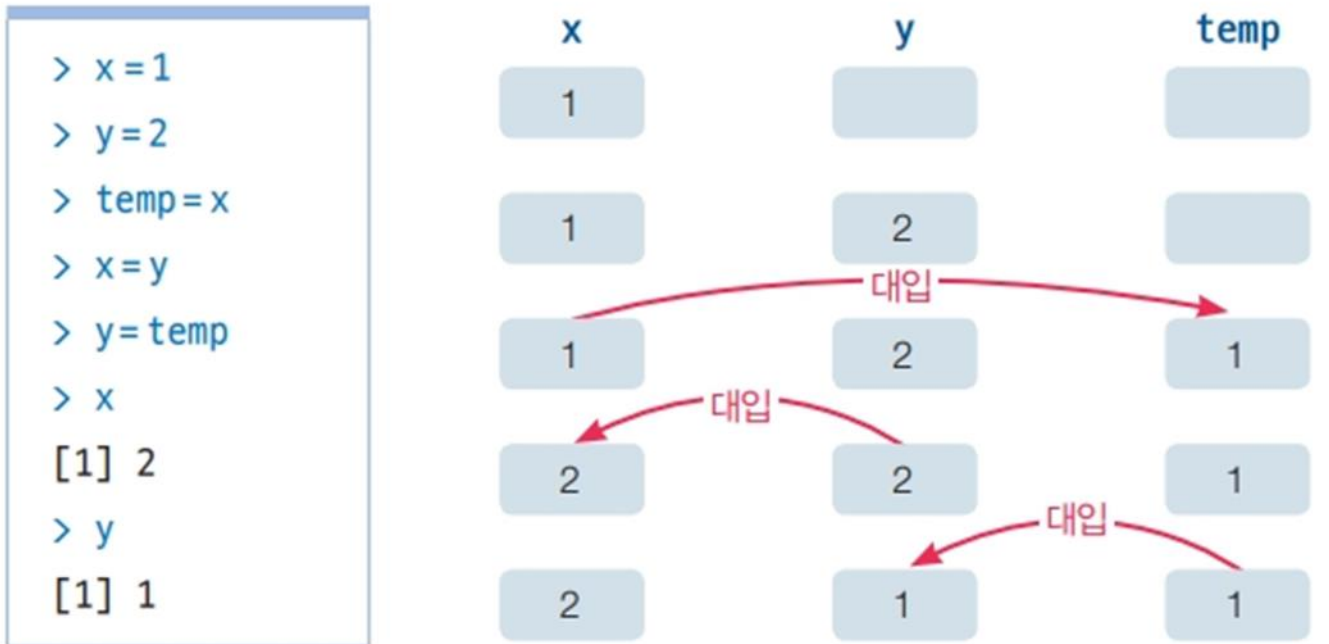


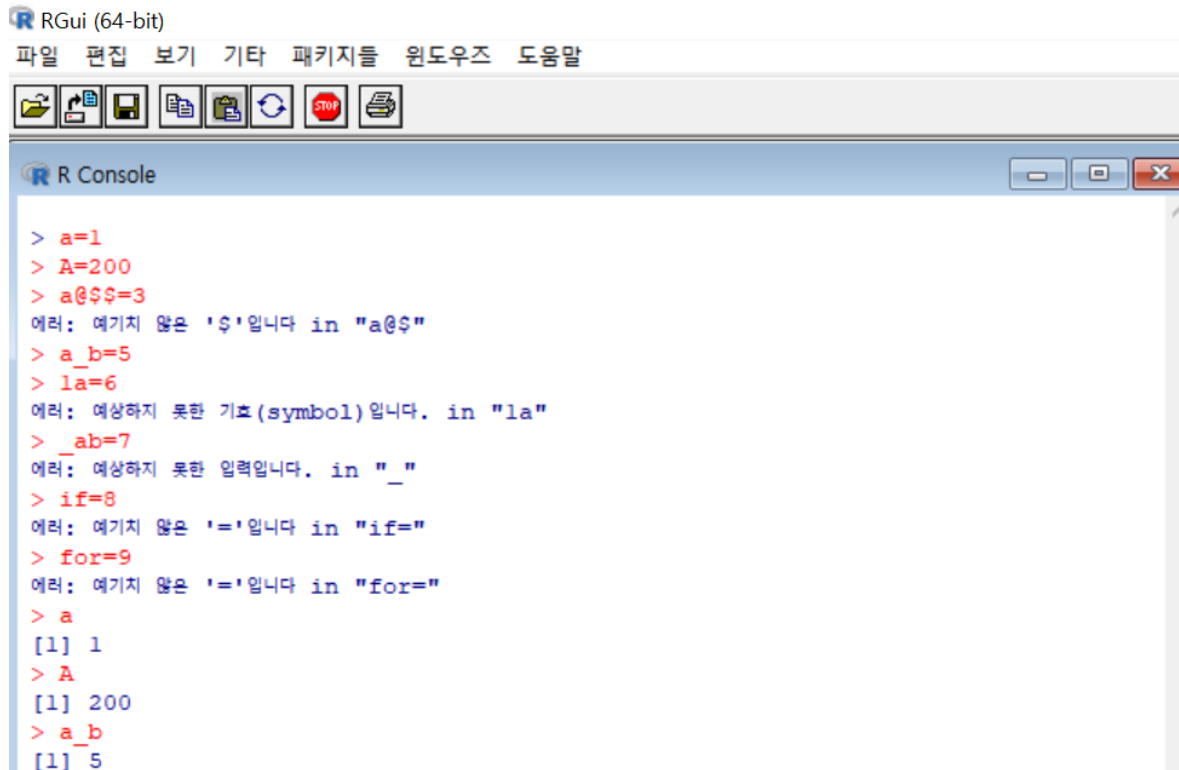
그림 3-1 두 값의 교환 소스 코드와 동작 원리 예





## ② 변수 이름 규칙

- 대소문자를 구별
- 특수 문자 쓸 수 없음(예외 `_` , `.` 는 가능)
- 숫자나 밑줄 `'_'` 을 변수의 첫 글자로 쓸 수 없음
- `if`, `for` , `while` 등 예약어는 사용할 수 없음



RGui (64-bit)

파일 편집 보기 기타 패키지들 윈도우즈 도움말

R Console

```
> a=1
> A=200
> a@@@=3
예러: 여기치 않은 '@'입니다 in "a@@"
> a_b=5
> la=6
예러: 예상하지 못한 기호(symbol)입니다. in "la"
> _ab=7
예러: 예상하지 못한 입력입니다. in "_"
> if=8
예러: 여기치 않은 '='입니다 in "if="
> for=9
예러: 여기치 않은 '='입니다 in "for="
> a
[1] 1
> A
[1] 200
> a_b
[1] 5
```



## 3.3 데이터형

### ■ R의 기본 데이터형

R에서는 데이터형을 지정하지 않고, 변수에 어떤 값을 저장하느냐에 따라 데이터 형이 결정 된다.

데이터형	종류
숫자형	정수(integer), 실수(numeric), 복소수(complex)
문자형	character : 작은따옴표나 큰따옴표로 묶어 표기
범주형	factor : 레벨(level)에 따라 분류된 형태
논리형	TRUE(T), FALSE(F)
특수 상수	NULL : 정의 되지 않은 값 NA(Not Available) : 결측 값 -Inf(음의 무한대), Inf(양의 무한대) NaN(Not a Number : 0/0, Inf/Inf 등과 같이 연산불가능함 값 표시)



## 3.3 데이터형

## ■ R의 기본 데이터형 학습 예제

RGui (64-bit) - [R Console]

파일 편집 보기 기타 패키지들 윈도우즈 도움말



```

> a=1
> class(a)
[1] "numeric"
> a='11'
> class(a)
[1] "character"
> a=factor(c('A', 'AB', 'B', 'O'))
> a
[1] A  AB B  O
Levels: A AB B O
> class(a)
[1] "factor"
> df <- data.frame(sex = c("M", "F", NA, "M"),
+                  score = c(5, 4, 3, NA))
> df
  sex score
1   M     5
2   F     4
3 <NA>     3
4   M    NA
> df[4,2]
[1] NA
> df[3,2]
[1] 3
> class(df[4,2])
[1] "numeric"
> class(df[1,2])
[1] "numeric"
> class(df[1,1])
[1] "character"

```



## 3.3 데이터형

### ■ 데이터형 확인 함수 및 변환 함수

함수	설명
<code>class(x)</code>	R 객체지향 관점에서 x의 데이터형
<code>typeof(x)</code>	R 언어 자체 관점에서 x의 데이터형
<code>is.integer(x)</code>	x가 정수형이면 TRUE, 아니면 FALSE
<code>is.numeric(x)</code>	x가 실수형이면 TRUE, 아니면 FALSE
<code>is.complex(x)</code>	x가 복소수형이면 TRUE, 아니면 FALSE
<code>is.character(x)</code>	x가 문자형이면 TRUE, 아니면 FALSE
<code>is.na(x)</code>	x가 NA이면 TRUE, 아니면 FALSE

```

RGui (64-bit)
파일 편집 보기 기타 패키지를 윈도우즈 도움

R Console

> x=1
> class(x)
[1] "numeric"
> typeof(x)
[1] "double"
> mode(x)
[1] "numeric"
> is.integer(x)
[1] FALSE
> x <- as.integer(x)
> class(x)
[1] "integer"
> typeof(x)
[1] "integer"
> mode(x)
[1] "numeric"
> is.integer(x)
[1] TRUE
  
```



## 3.3 데이터형

### ■ 데이터형 확인 함수 및 변환 함수

함수	설명
as.factor(x)	x를 범주형으로 변환
as.integer(x)	x를 정수형으로 변환
as.numeric(x)	x를 숫자형으로 변환
as.character(x)	x를 문자형으로 변환
as.matrix(x)	x를 행렬로 변환
as.array(x)	x를 배열로 변환

Console C:/RSources/ 

```
> x="1"
> class(x)
[1] "character"
> x<-as.integer(x)
> class(x)
[1] "integer"
```



# Thank you

