



4주차: 데이터 취득과 정제

ChulSoo Park

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

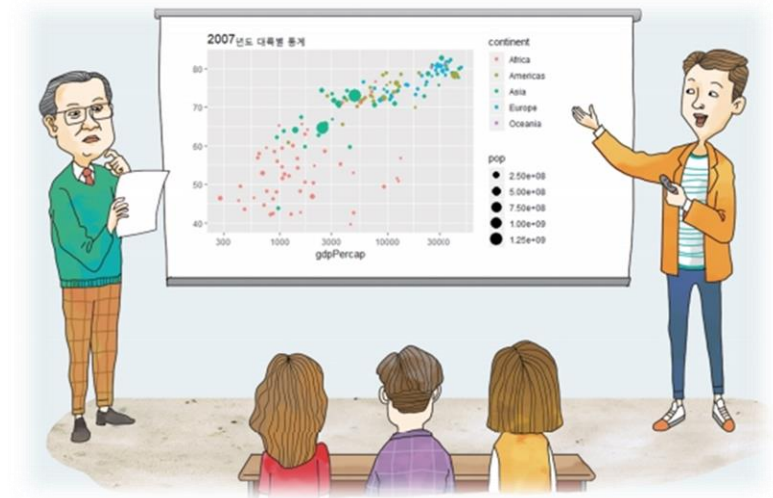
학습목표 (4주차)

- ❖ 텍스트 파일, 엑셀 파일 등 데이터 읽고 쓰기 학습
- ❖ R에서 조건문과 반복문 학습
- ❖ 사용자 함수 이해 및 만들기
- ❖ 수집한 데이터 결측값 처리 방법 숙지
- ❖ 수집한 데이터 이상값 처리 방법 숙지

04

CHAPTER

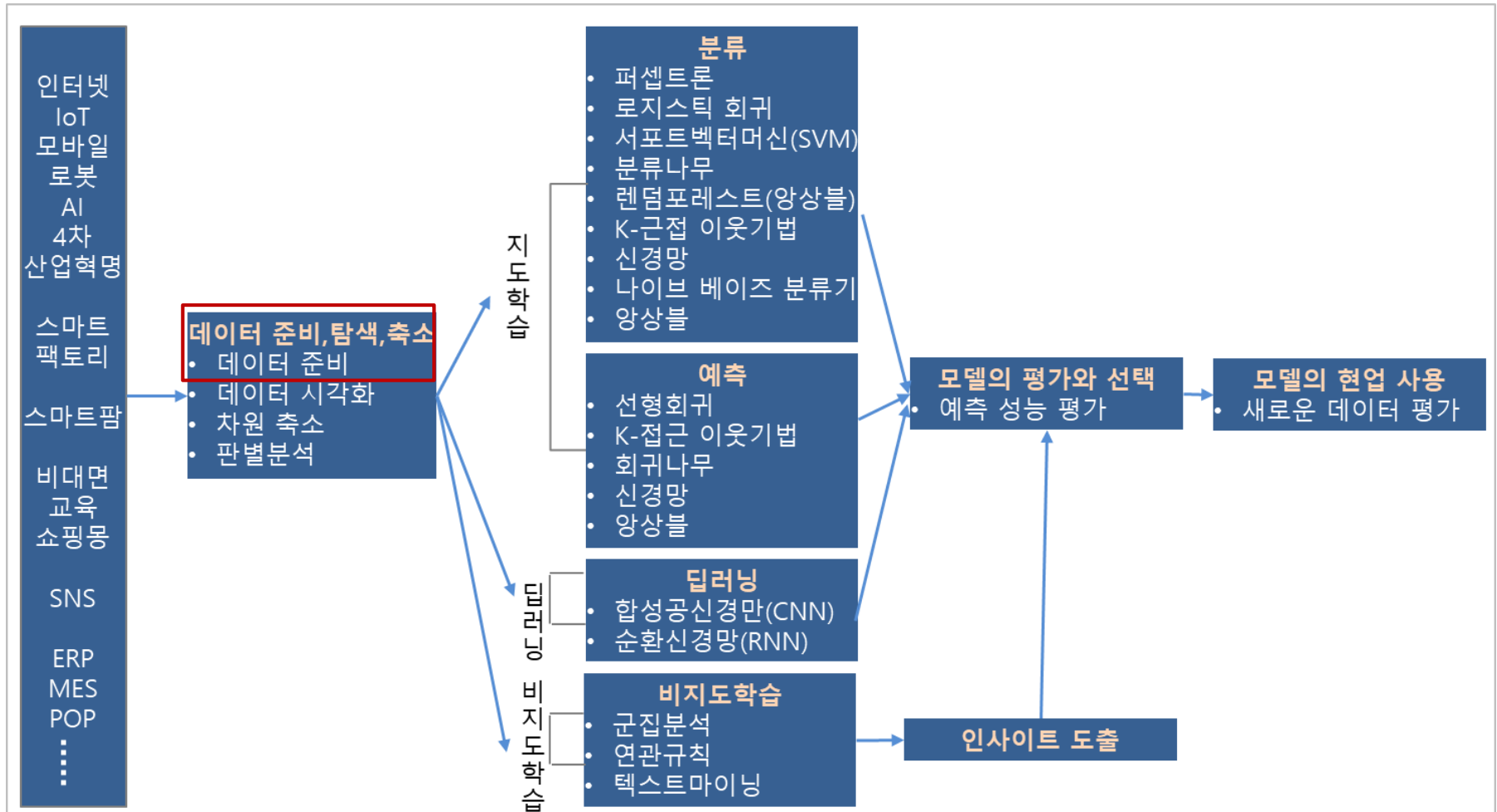
데이터 취득과 정제



CONTENTS

- 4.1 파일 일고 쓰기
- 4.2 데이터 정제를 위한 조건문과 반복문
- 4.3 사용자 정의 함수 : 원하는 기능 묶기
- 4.4 데이터 정제 예제 1 : 결측값 처리
- 4.5 데이터 정제 예제 2 : 이상값 처리
- 요약

Preview



■ 데이터 수집과 정제

- 데이터는 인터넷 서핑을 통해서, 문서를 통해서, 설문조사나 실험을 통해 얻을 수 있다.
- 수집한 자료를 데이터 과학 목적에 맞게 사용하기 위해서는 적절히 정제하여야 한다.
- 정제한 데이터를 이용하여 대부분의 데이터 가공과 처리가 이뤄질 수 있다.



4.1 파일 읽고 쓰기

- 대부분의 데이터는 파일 형태로 존재한다.
- R에서 제공하는 파일 읽고 쓰기 함수

R에서 사용할 수 있는 파일 일기와 쓰기 함수

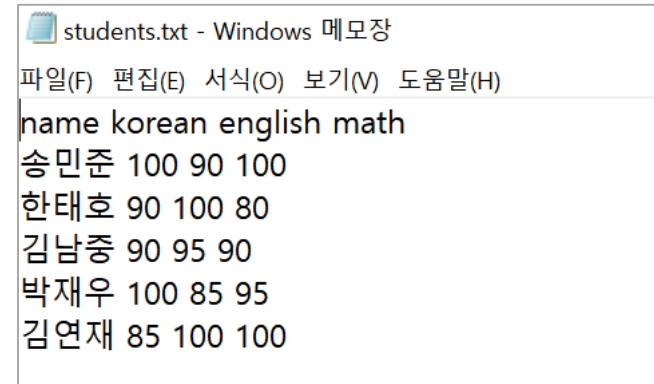
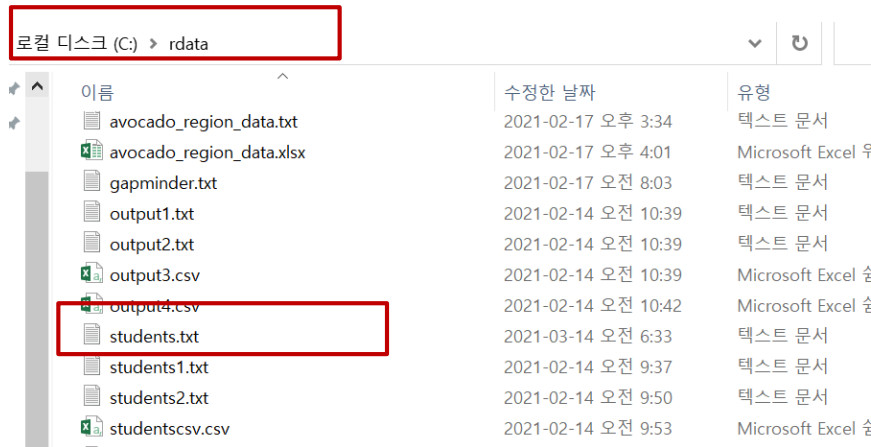
패키지	함수
Base(기본) 패키지	scan, write, write.table, read.table, save, load, write.csv, read.csv
Readr 패키지	write.csv, read.csv
Data.table 패키지	fwrite, fread
Feather 패키지	write_feather, read_feather

4.1 파일 읽고 쓰기

- read.table 함수: 일반 텍스트 파일 읽기 (c:/rdata/sudents.txt)
- read.csv 함수: 일반 텍스트 파일 읽기 (c:/rdata/sudents.csv)

```
students=read.table("C:/rdata/students.txt",header=T)
```

```
students=read.table("C:/rdata/students.txt",encoding="UTF-8",header=T)
```



```
R: Data Input ▾ Find in Topic

read.table(file, header = FALSE, sep = "", quote = "\"",
  dec = ".", numerals = c("allow.loss", "warn.loss", "no.loss"),
  row.names, col.names, as.is = !stringsAsFactors,
  na.strings = "NA", colClasses = NA, nrows = -1,
  skip = 0, check.names = TRUE, fill = !blank.lines.skip,
  strip.white = FALSE, blank.lines.skip = TRUE,
  comment.char = "#",
  allowEscapes = FALSE, flush = FALSE,
  stringsAsFactors = default.stringsAsFactors(),
  fileEncoding = "", encoding = "unknown", text, skipNul = FALSE)

read.csv(file, header = TRUE, sep = ",", quote = "\"",
  dec = ".", fill = TRUE, comment.char = "", ...)
```

4.1 파일 읽고 쓰기

② 파일 쓰기

- write.table 함수: 일반 텍스트 파일로 저장할 때 사용 : .txt
- Write.csv 함수 : csv 파일로 저장 : .csv

Console C:/RSources/ ↗

> ?write.table

R: Data Output ▾ Find in Topic

Usage

```
write.table(x, file = "", append = FALSE, quote = TRUE, sep = " ",  
            eol = "\n", na = "NA", dec = ".", row.names = TRUE,  
            col.names = TRUE, qmethod = c("escape", "double"),  
            fileEncoding = "")
```

```
write.csv(...)  
write.csv2(...)
```


4.1 파일 읽고 쓰기

- write.table 함수: 일반 텍스트 파일로 저장할 때 사용 : .txt

```
students=read.table("C:/rdata/students.txt",header=T)
```

```
write.table(students, file="c:/rdata/output.txt")
```

```
write.table(students, file="c:/rdata/output1.txt",quote=F)
```

```
write.table(students, file="c:/rdata/output2.txt",quote=F,row.names = FALSE)
```

output.txt - Windows 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

```
"name" "korean" "english" "math"  
"1" "송민준" 100 90 100  
"2" "한태호" 90 100 80  
"3" "김남중" 90 95 90  
"4" "박재우" 100 85 95  
"5" "김연재" 85 100 100
```

output1.txt - Windows 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

```
name korean english math  
1 송민준 100 90 100  
2 한태호 90 100 80  
3 김남중 90 95 90  
4 박재우 100 85 95  
5 김연재 85 100 100
```

output2.txt - Windows 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

```
name korean english math  
송민준 100 90 100  
한태호 90 100 80  
김남중 90 95 90  
박재우 100 85 95  
김연재 85 100 100
```

4.2 데이터 정제를 위한 조건문과 반복문

① 조건문

- 데이터 정제를 위해 특정 조건에 맞는 값을 찾아내거나 일부 구간의 값을 추출하여 연산하는 등 다양한 목적에 맞게 작업할 수 있다.
- R에서 제공하는 조건 탐색 기능을 살펴보고, 조건문과 반복문 사용법에 대해 학습해 보자.
- 조건문 형식

조건에 맞는 요소를 추출하는 방법	형식
[]에 행/열 조건 명시	변수명 [행 조건식, 열 조건식]
If 문 활용 (if/ else if/else)	If(조건식) 표현식
Ifelse 문 활용	Ifelse(조건식, 참인 경우 반환값, 거짓인 경우 반환값)

4.3 사용자의 함수 : 반복문 원하는 기능 묶기

■ 함수

- 입력과 출력간의 관계식을 함수라고 할 수 있다.
- 사용자의 목적에 맞는 다양한 함수를 만들어 보자.

■ 사용자 정의 함수의 구조

```
함수명 = function(전달자1, 전달자2, 전달자3, ...) {  
    함수 동작 시 수행 프로그램  
    return(반환값)  
}
```

4.4 데이터 정제 예제1 : 결측값 처리

- 우리가 수집한 데이터에는 결측값(missing value)이 존재할 수 있다.
- 결측값은 데이터 중 고의 또는 실수로 누락된 값을 의미한다.
- 결측값이 존재한 채 데이터 가공을 하면 결과값에 오류가 뜨거나 잘못된 연산이 수행될 수 있으므로 정제과정에서 적절한 처리가 필요하다.
- 결측값 처리 대표적인 방법 3가지

방법	설명
is.na 함수 이용	NA인 데이터가 있으면 T, 없으면 F로 나타낸다.
na.omit 함수 이용	NA인 데이터를 제거한다. 즉, NA가 포함된 행을 지운다.
함수의 속성 이용	Na.rm=T로하여 함수 수행 시, NA를 제외한다.

4.4 데이터 정제 예제1 : 결측값 처리

■ airquality data 살펴보기

```
Console C:/RSources/ ↗
> str(airquality)
'data.frame': 153 obs. of 6 variables:
 $ Ozone   : int  41 36 12 18 NA 28 23 19 8 NA ...
 $ Solar.R: int  190 118 149 313 NA NA 299 99 19 194 ...
 $ Wind    : num  7.4 8 12.6 11.5 14.3 14.9 8.6 13.8 20.1 8.6 ...
 $ Temp    : int  67 72 74 62 56 66 65 59 61 69 ...
 $ Month   : int  5 5 5 5 5 5 5 5 5 5 ...
 $ Day     : int  1 2 3 4 5 6 7 8 9 10 ...
> head(airquality,10)
   Ozone Solar.R Wind Temp Month Day
1     41     190  7.4   67     5   1
2     36     118  8.0   72     5   2
3     12     149 12.6   74     5   3
4     18     313 11.5   62     5   4
5     NA      NA 14.3   56     5   5
6     28      NA 14.9   66     5   6
7     23     299  8.6   65     5   7
8     19      99 13.8   59     5   8
9      8      19 20.1   61     5   9
10    NA     194  8.6   69     5  10
```




4.4 데이터 정제 예제1 : 결측값 처리

■ airquality data 살펴보기

```
Console C:/Rsources/ ↗  
> table(is.na(airquality))  
  
FALSE  TRUE  
  985    86  
> table(is.na(airquality$Temp))  
  
FALSE  
  153  
> table(is.na(airquality$Ozone))  
  
FALSE  TRUE  
  116    37  
> mean(airquality$Temp)  
[1] 77.88235  
> mean(airquality$Ozone)  
[1] NA
```

4.4 데이터 정제 예제1 : 결측값 처리

■ airquality data 살펴보기

```
Console C:/RSources/   
```

```
> airnotozone=airquality[!is.na(airquality$Ozone), ]  
> head(airnotozone,10)
```

	Ozone	Solar.R	Wind	Temp	Month	Day	total
1	41	190	7.4	67	5	1	311.4
2	36	118	8.0	72	5	2	241.0
3	12	149	12.6	74	5	3	255.6
4	18	313	11.5	62	5	4	413.5
6	28	NA	14.9	66	5	6	NA
7	23	299	8.6	65	5	7	407.6
8	19	99	13.8	59	5	8	203.8
9	8	19	20.1	61	5	9	122.1
11	7	NA	6.9	74	5	11	NA
12	16	256	9.7	69	5	12	367.7

```
> mean(airnotozone$Ozone)  
[1] 42.12931  
> airnotna=na.omit(airquality)  
> mean(airnotna$Ozone)  
[1] 42.0991  
> mean(airquality$Ozone,na.rm=T) # 함수 속성인 na.rm을 이용해 결측값 처리  
[1] 42.12931
```

4.4 데이터 정제 예제1 : 결측값 처리

■ airquality data 살펴보기

Console C:/RSources/ ↗

```
> airquality$total = airquality$Ozone + airquality$Solar.R + airquality$Wind + airquality$Temp + airquality$Month + airquality$Day  
> head(airquality,10)
```

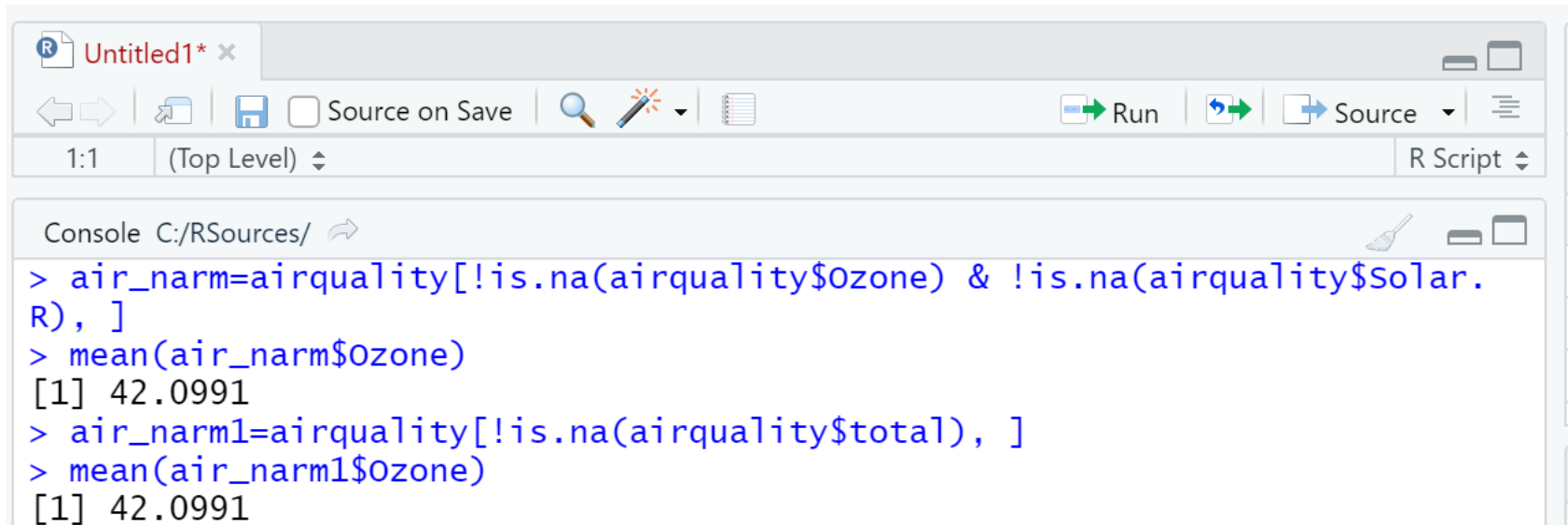
	Ozone	Solar.R	wind	Temp	Month	Day	total
1	41	190	7.4	67	5	1	311.4
2	36	118	8.0	72	5	2	241.0
3	12	149	12.6	74	5	3	255.6
4	18	313	11.5	62	5	4	413.5
5	NA	NA	14.3	56	5	5	NA
6	28	NA	14.9	66	5	6	NA
7	23	299	8.6	65	5	7	407.6
8	19	99	13.8	59	5	8	203.8
9	8	19	20.1	61	5	9	122.1
10	NA	194	8.6	69	5	10	NA

4.4 데이터 정제 예제1 : 결측값 처리

■ airquality data 살펴보기

```
Console C:/RSources/ ↗  
> airquality[is.na(airquality$total), ]  
   Ozone Solar.R Wind Temp Month Day total  
5      NA      NA 14.3   56     5   5    NA  
6      28      NA 14.9   66     5   6    NA  
10     NA     194  8.6   69     5  10    NA  
11      7      NA  6.9   74     5  11    NA  
25     NA      66 16.6   57     5  25    NA  
26     NA     266 14.9   58     5  26    NA  
27     NA      NA  8.0   57     5  27    NA  
32     NA     286  8.6   78     6   1    NA  
33     NA     287  9.7   74     6   2    NA  
34     NA     242 16.1   67     6   3    NA  
35     NA     186  9.2   84     6   4    NA  
36     NA     220  8.6   85     6   5    NA  
37     NA     264 14.3   79     6   6    NA  
39     NA     273  6.9   87     6   8    NA  
42     NA     259 10.9   93     6  11    NA  
43     NA     250  9.2   92     6  12    NA  
45     NA     332 13.8   80     6  14    NA  
46     NA     322 11.5   79     6  15    NA  
52     NA     150  6.3   77     6  21    NA  
53     NA      59  1.7   76     6  22    NA
```

4.4 데이터 정제 예제1 : 결측값 처리



The image shows a screenshot of the RStudio interface. The top toolbar includes icons for navigation, saving, and running code. The console window at the bottom displays the following R commands and their output:

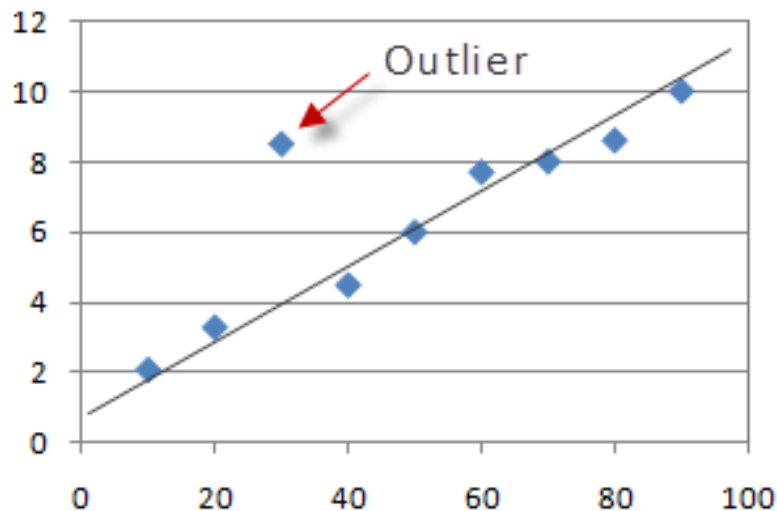
```
> air_narm=airquality[!is.na(airquality$Ozone) & !is.na(airquality$Solar.
R), ]
> mean(air_narm$Ozone)
[1] 42.0991
> air_narm1=airquality[!is.na(airquality$total), ]
> mean(air_narm1$Ozone)
[1] 42.0991
```

4.5 데이터 정제 예제2 : 이상값 처리

- 결측값과 더불어 데이터에는 논리적 혹은 통계학적으로 이상한 데이터가 입력되어 있을 수 있다. 이러한 데이터를 이상값(outlier)이라 한다.
- "통계학에서 이상값이란 다른 관측값과 멀리 떨어진 관측값"




Outliers can occur in the dataset due to one of the following reasons,

- 1.Genuine extreme high and low values in the dataset
- 2.Introduced due to human or mechanical error
- 3.Introduced by replacing missing values



4.5 데이터 정제 예제2 : 이상값 처리

■ 성별에서 이상값 제거

```
Console C:/Rsources/   
```

```
> # 이상값이 포함된 환자 데이터
> patients
에러: 객체 'patients'를 찾을 수 없습니다
> patients=data.frame(name=c("환자1","환자2","환자3","환자4","환자5"), age=c(2
2,20,25,30,27), gender=factor(c("M","F","M","K","F")),blood.type=factor(c
("A","O","B","AB","C")))
> patients
  name age gender blood.type
1 환자1  22      M         A
2 환자2  20      F         O
3 환자3  25      M         B
4 환자4  30      K        AB
5 환자5  27      F         C
> patients_outrm=patients[patients$gender=="M"|patients$gender=="F", ]
> patients_outrm # 성별에서 이상값 제거 후
  name age gender blood.type
1 환자1  22      M         A
2 환자2  20      F         O
3 환자3  25      M         B
5 환자5  27      F         C
```

4.5 데이터 정제 예제2 : 이상값 처리

■ 성별에서 이상값 제거

```
Console C:/RSources/
> patients_outrm1=patients[(patients$gender=="M"|patients$gender=="F") & (patients$blood.type=="A"|patients$blood.type=="B"|patients$blood.type=="O"|patients$blood.type=="AB"), ]
> patients_outrm1
  name age gender blood.type
1 환자1  22      M         A
2 환자2  20      F         O
3 환자3  25      M         B
> patients
  name age gender blood.type
1 환자1  22      M         A
2 환자2  20      F         O
3 환자3  25      M         B
4 환자4  30      K        AB
5 환자5  27      F         C
```

4.5 데이터 정제 예제2 : 이상값 처리

- 공기 질을 측정하기 위한 풍속, 온도, 오존량, 태양복사량 등의 측정 값의 이상치 처리(airquality)

```
Console C:/RSources/
> head(airquality,10)
  Ozone Solar.R Wind Temp Month Day
1    41    190  7.4   67     5    1
2    36    118  8.0   72     5    2
3    12    149 12.6   74     5    3
4    18    313 11.5   62     5    4
5    NA     NA 14.3   56     5    5
6    28     NA 14.9   66     5    6
7    23    299  8.6   65     5    7
8    19     99 13.8   59     5    8
9     8     19 20.1   61     5    9
10   NA    194  8.6   69     5   10

> str(airquality)
'data.frame':   153 obs. of  6 variables:
 $ Ozone   : int  41 36 12 18 NA 28 23 19 8 NA ...
 $ Solar.R: int  190 118 149 313 NA NA 299 99 19 194 ...
 $ Wind    : num  7.4 8 12.6 11.5 14.3 14.9 8.6 13.8 20.1 8.6 ...
 $ Temp    : int  67 72 74 62 56 66 65 59 61 69 ...
 $ Month   : int  5 5 5 5 5 5 5 5 5 5 ...
 $ Day     : int  1 2 3 4 5 6 7 8 9 10 ...
```

4.5 데이터 정제 예제2 : 이상값 처리

■ 박스 플롯(box plot) 설명

박스 플롯을 사용하는 이유는 많은 데이터를 눈으로 확인하기 어려울 때 그림을 이용해 데이터 집합의 범위와 중앙값을 빠르게 확인할 수 있는 목적으로 사용한다. 또한 통계적으로 이상치(outlier)가 있는지도 확인이 가능하다.

기술 통계학에서 박스 플롯은 수치적 자료를 표현하는 그래프이다. 이 그래프는 자료에서 얻은 다섯 수치 요약(five number summary)을 가지고 그린다.

다섯 수치 요약은 아래와 같다.

1. 최솟값 : 제 1사분위에서 1.5 IQR을 뺀 위치이다.
 2. 제 1사분위(Q1) : 25%의 위치를 의미한다.
 3. 제 2사분위(Q2) : 50%의 위치로 중앙값(median)을 의미한다.
 4. 제 3사분위(Q3) : 75%의 위치를 의미한다.
 5. 최댓값 : 제 3사분위에서 1.5 IQR을 더한 위치이다.
- 최솟값과 최댓값을 넘어가는 위치에 있는 값을 이상치(Outlier)라고 부른다.

4.5 데이터 정제 예제2 : 이상값 처리

■ 박스 플롯(box plot) 설명

박스 플롯은 박스와 박스 바깥의 선(whisker)으로 이루어져 있다.

구분	설명
whisker	상자의 좌우 또는 상하로 뻗어나간 선
박스 내부의 가로선	중앙 값을 나타냄
Lower whisker	최소값 중앙값 - 1.5 X IQR보다 큰 데이터 중 가장 작은 값
Upper whisker	최대값 중앙값 + 1.5 X IQR보다 작은 데이터 중 가장 큰 값
IRQ	Inter Quartile Range 제3사분위수 - 제1사분위수 실수 값 분포에서 1사분위(Q1)와 3사분위(Q3)를 뜻하고 이 3사분위수와 1사분위의 차이(Q3 - Q1)를 IRQ(Inter Quartile Range)라 함
점	이상치(outlier) ≡ 특이점 Lower whisker보다 작은 데이터 또는 upper whisker보다 큰 데이터가 여기에 해당됨.

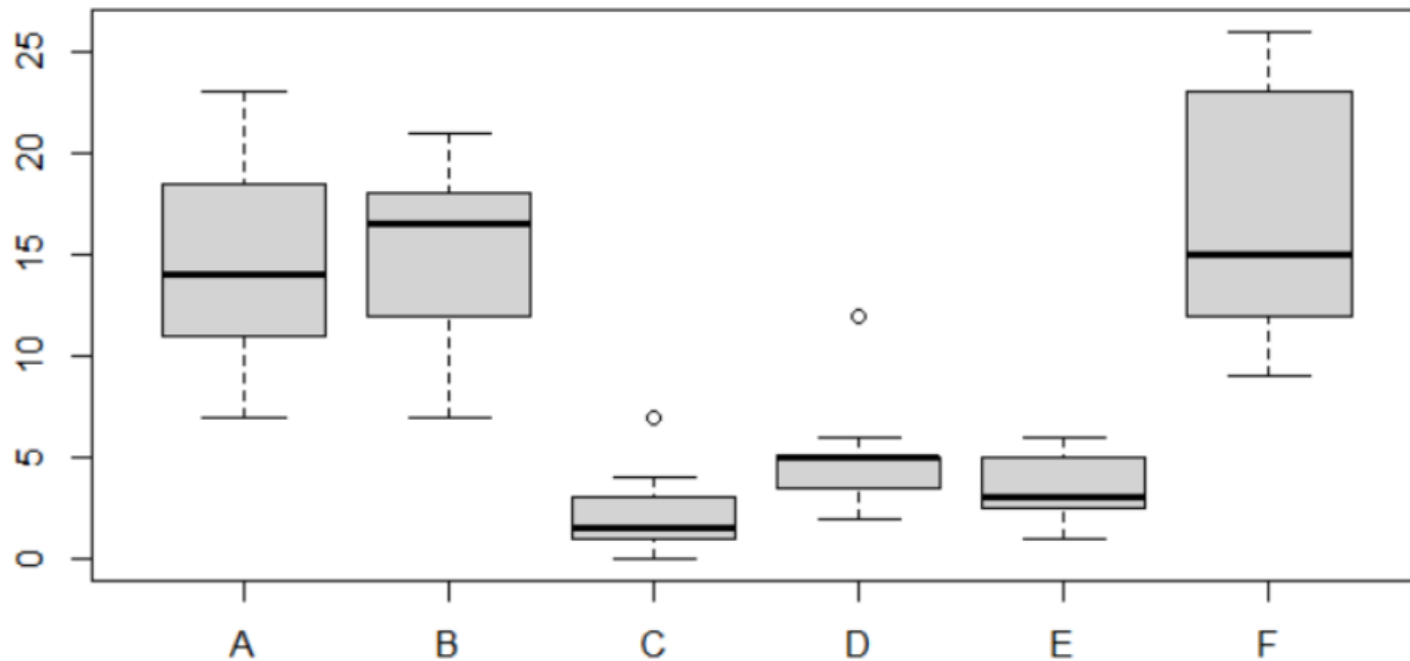
4.5 데이터 정제 예제2 : 이상값 처리

박스 플롯을 그리는 방법은 아래와 같다.

1. 주어진 데이터에서 각 사분위수를 계산한다.
2. 그래프에서 제 1사분위수와 제 3사분위수를 기준으로 박스를 그린다.
3. 제 2사분위수에 해당하는 위치에 선을 긋는다.
4. 제 3사분위수에서 $1.5IQR$ 을 더한 위치에 가로 선을 긋고 제 3사분위수부터 가로선까지 세로선을 긋는다.
5. 제 1사분위수에서 $1.5IQR$ 을 뺀 위치에 가로 선을 긋고 제 1사분위수부터 가로선까지 세로선을 긋는다.
6. 4,5번에 그은 직선을 넘어서는 위치에 존재하는 값은 동그라미와 같은 기호로 표시한다.(이상치 의미)

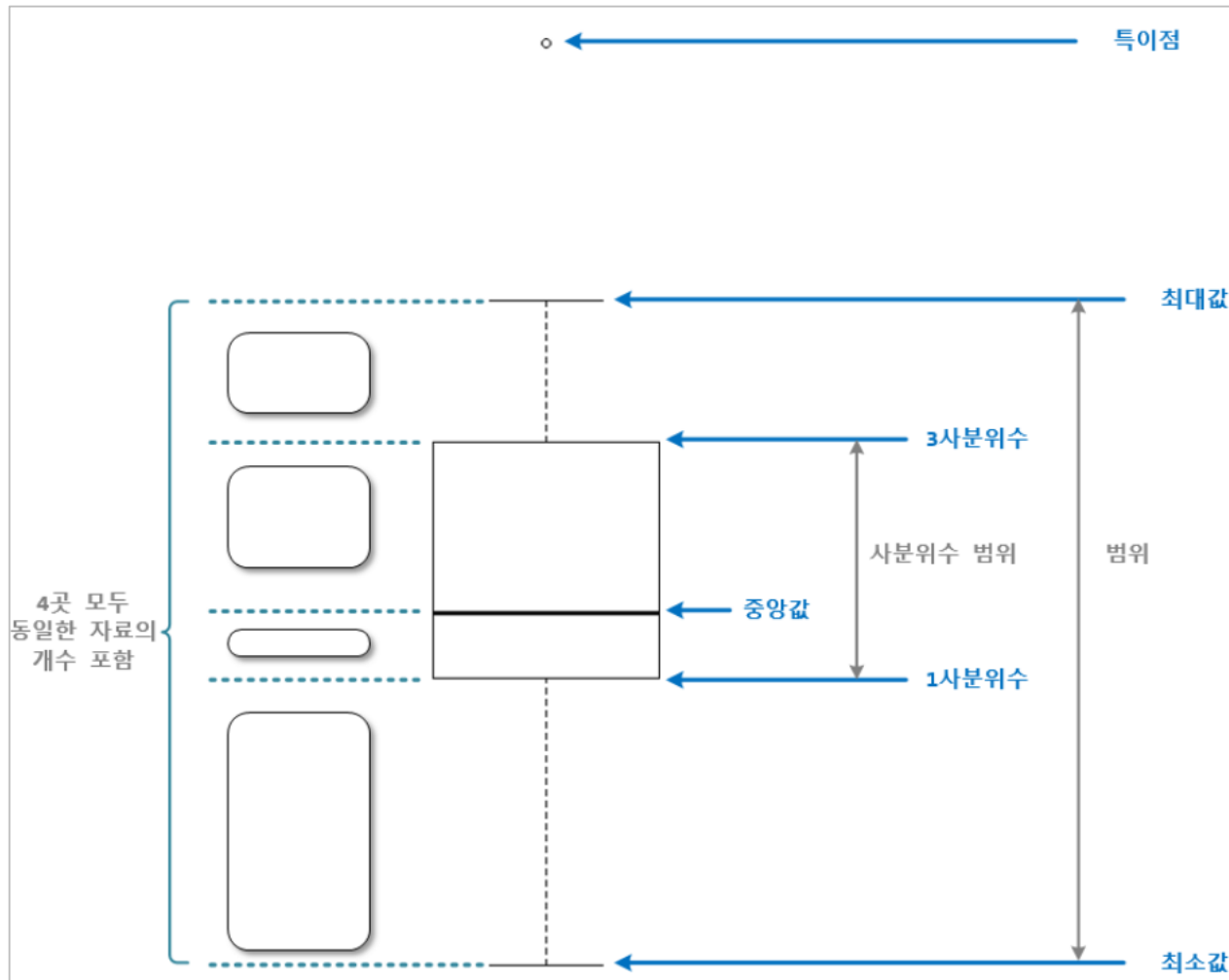
4.5 데이터 정제 예제2 : 이상값 처리

■ 박스 플롯(box plot) (예)



4.5 데이터 정제 예제2 : 이상값 처리

■ 박스 플롯(box plot) (예)



4.5 데이터 정제 예제2 : 이상값 처리

- 공기 질을 측정하기 위한 풍속, 온도, 오존량, 태양복사량 등의 측정 값의 이상치 처리(airquality)

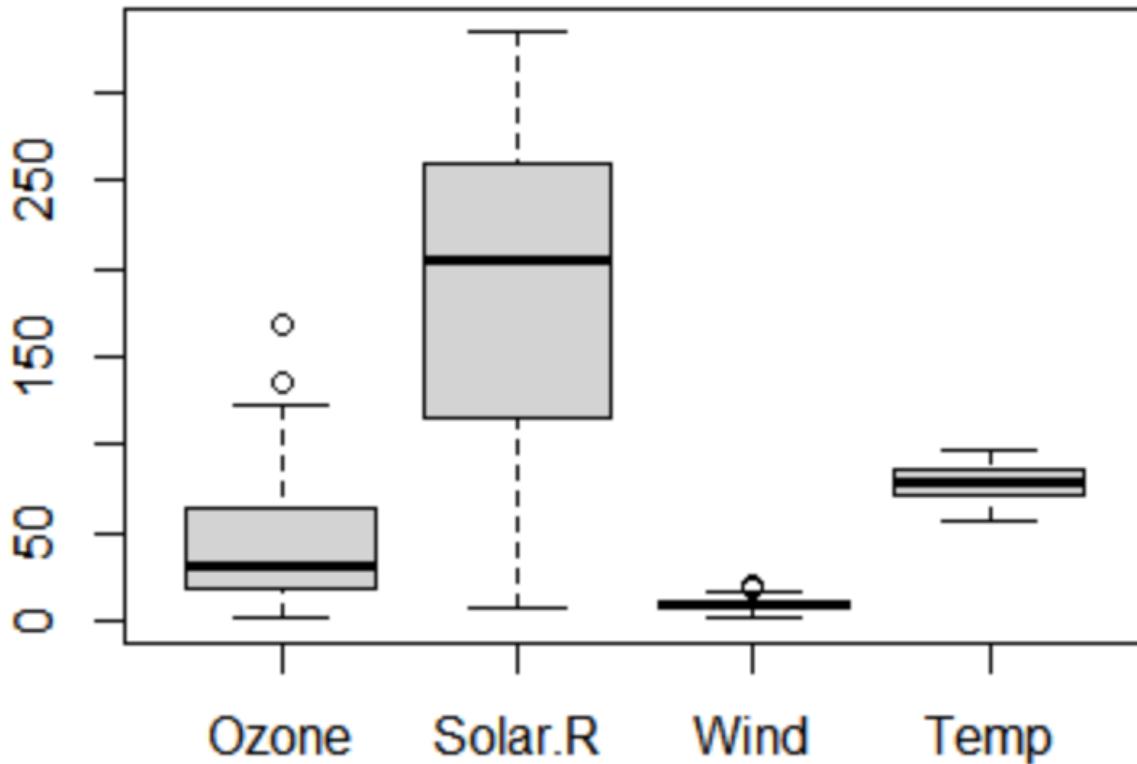
```
Console C:/RSources/
> boxplot(airquality[, 1])$stats
      [,1]
[1,]  1.0
[2,] 18.0
[3,] 31.5
[4,] 63.5
[5,] 122.0
attr(,"class")
      1
"integer"
> airquality[airquality$Ozone>122&!is.na(airquality$Ozone), ]
      Ozone Solar.R Wind Temp Month Day
62      135      269  4.1   84     7   1
117     168      238  3.4   81     8  25
```

이 값 미만은 이상 값으로 분류 할 수 있음

이 값 이상은 이상 값으로 분류 할 수 있음

4.5 데이터 정제 예제2 : 이상값 처리

- 공기 질을 측정하기 위한 풍속, 온도, 오존량, 태양복사량 등의 측정 값의 이상치 처리(airquality)



1. 파일 읽고 쓰기
2. 데이터 정제를 위한 조건문과 반복문
3. 사용자 정의 함수
4. 데이터 정제(결측값 처리)
5. 데이터 정제(이상값 처리)
6. Boxplot 이해

Thank you

