4주차:
데이터 취득과 정제

ChulSoo Park

School of Computer Engineering & Information Technology

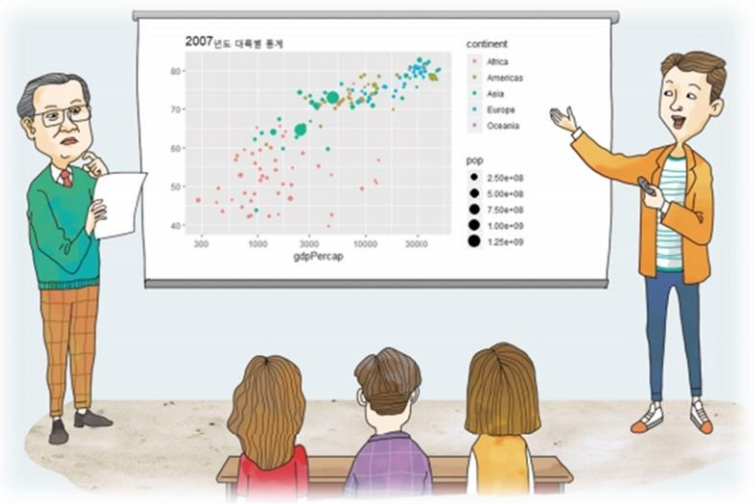
Korea National University of Transportation

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

Review

- 3장에서는 R의 데이터 형과 연산

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.

Review

배열(행렬)

R data sets | 3chapter.R*

Source on Save | Run | Source

```
57  x=rep(c(1:3),each=3)
58  x
59
60  x=array(1:12,c(3,4))
```

66:1 (Top Level) R Script

Console C:/RSources/

```
> x=array(1:12,c(3,4))
> x
     [,1] [,2] [,3] [,4]
[1,]    1    4    7   10
[2,]    2    5    8   11
[3,]    3    6    9   12
> ?array
> y=1:20
> matrix(y,nrow=4,byrow=T)
     [,1] [,2] [,3] [,4] [,5]
[1,]    1    2    3    4    5
[2,]    6    7    8    9   10
[3,]   11   12   13   14   15
[4,]   16   17   18   19   20
> matrix(y,ncol=4,byrow=F)
     [,1] [,2] [,3] [,4]
[1,]    1    6   11   16
[2,]    2    7   12   17
[3,]    3    8   13   18
[4,]    4    9   14   19
[5,]    5   10   15   20
```

Environment | History

Import Datas

R | Global Environm

b	-5
bb	25
cc	25
xx	NULL
xy	chr
y	int

Files | Plots | Packages

R: Multi-way Arrays

An array in R can have
simply a vector which is
giving the dimensions (
names for those dimen

A two-dimensional arra

One-dimensional array
handled differently by a
them in recent version

Review

데이터프레임

환자 데이터

name	age	gender	blood.type
철수	22	M	A
춘향	20	F	O
길동	25	M	B

속성

요소

```
Console C:/RSources/
> name=c("철수","춘향","길동")
> age=c(22,20,25)
> gender=factor(c("M","F","M"))
> blood.type=factor(c("A","O","B"))
> patients1=data.frame(name,age,gender,blood.type)
> patients1
  name age gender blood.type
1 철수  22      M          A
2 춘향  20      F          O
3 길동  25      M          B
> patients2=data.frame(name=c("철수","춘향","길동"),
+ age=c(22,20,25),gender=factor(c("M","F","M")),blood.type=fa
ctor(c("A","O","B")))
> patients2
  name age gender blood.type
1 철수  22      M          A
2 춘향  20      F          O
3 길동  25      M          B
```

Review

연산자 우선순위

연산자	설명	우선순위
^,**	자승수	높음 ↑
+,-	단항 플러스,마이너스	
%any%%	%%, %/% 등 연산자	
*, /	곱셈, 나눗셈	
+,-	덧셈, 뺄셈	
==,!=,<,>,<=,>=	비교 연산자	낮음 ↓
!	논리 부정(not)	
&, &&	논리 and	
|, ||	논리 or	

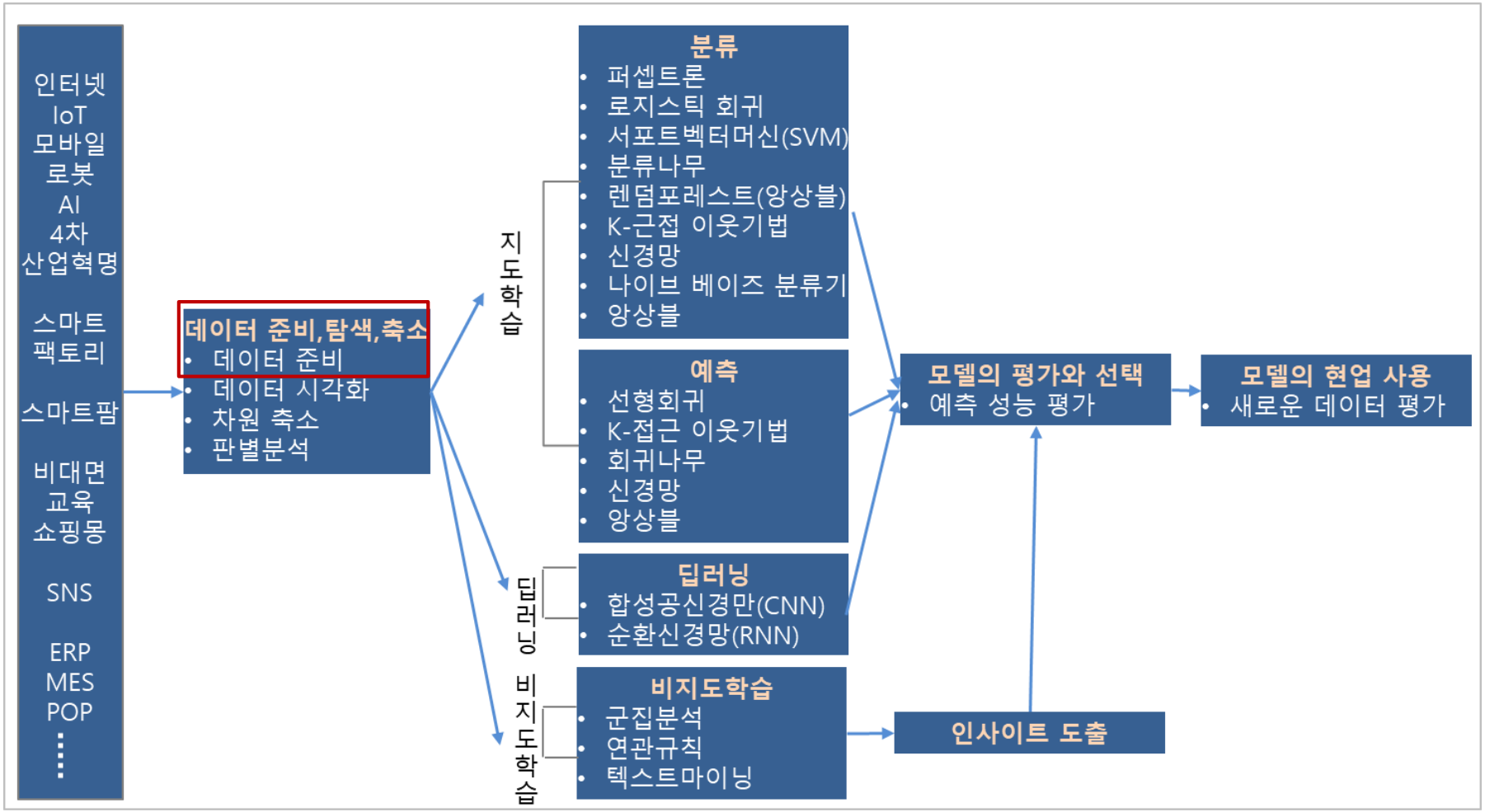
```
Console  C:/RSources/
> aa=-5**2
> b=-5
> bb=b**2
> cc=(-5)**2
>
>
```

aa = ??

bb = ??

cc = ??

Preview


인터넷
IoT
모바일
로봇
AI
4차 산업혁명
스마트 팩토리
스마트팜
비대면 교육
쇼핑몽
SNS
ERP
MES
POP
.....
데이터 준비,탐색,축소
데이터 준비
데이터 시각화
차원 축소
판별분석
지도학습
분류
퍼셉트론
로지스틱 회귀
서포트벡터머신(SVM)
분류나무
렌덤포레스트(앙상블)
K-근접 이웃기법
신경망
나이브 베이즈 분류기
앙상블
예측
선형회귀
K-접근 이웃기법
회귀나무
신경망
앙상블
딥러닝
딥러닝
합성공신경만(CNN)
순환신경망(RNN)
비지도학습
비지도학습
군집분석
연관규칙
텍스트마이닝
모델의 평가와 선택
예측 성능 평가
모델의 현업 사용
새로운 데이터 평가
인사이트 도출

Preview

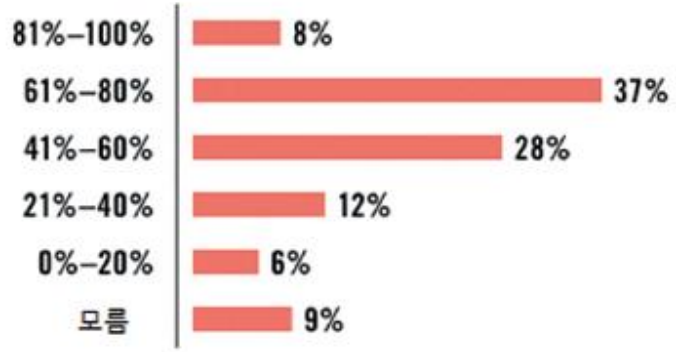
데이터 수집과 정제

- 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.

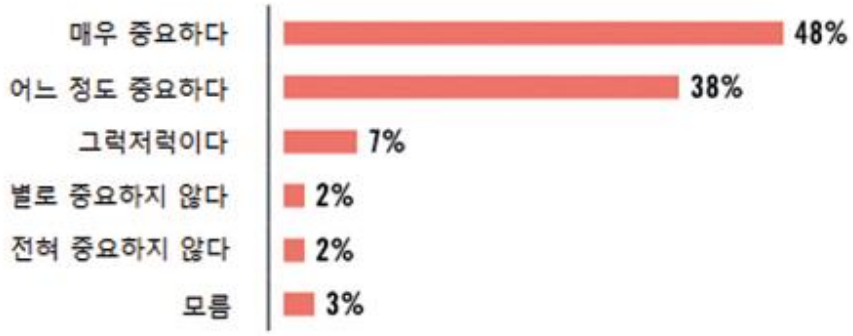
Preview

기업의 의사결정권자들은 흔히 DB(Database)로 관리되는 데이터를 분석에 그대로 활용할 수 없다는 것에 의문을 표하곤 한다. 비즈니스 상에서 발생한 데이터를 일정한 기준에 맞춰 축적해왔다면 이를 데이터 분석에 그대로 활용할 수 있지 않느냐는 얘기다. 하지만 DB에 축적된 데이터에는 얼마든지 오류가 발생할 수 있으며 축적된 데이터끼리 완벽히 같은 형태나 구조를 갖추고 있을 가능성도 낮다.

데이터 전처리를 위해 필요한 시간은 전체 BI 및 분석 프로젝트에 걸린 시간의 어느 정도를 차지 합니까?



데이터 전처리를 과정에 소요되는 시간과 자원을 줄이는 것이 귀사에 얼마나 중요합니까?



출처 : 컴퓨터월드 : https://www.comworld.co.kr/news/articleView.html?idxno=49333

Preview

구분		종류 예시
기업 내부 데이터	비즈니스, 웹 서비스 활동	매출 기록, 거래 내역, 웹 로그, 고객 정보 등
	각종 센서 장비	센서 데이터(스마트 팩도리,스마트 물류,스마트팜,환경 센서 등)
	과학 기술 연구 활동	연구 실험 및 측정 데이터
기업 외부 데이터	정부, 공공 기관	공공 인프라(공공 데이터 포털,통계청 KOSIS, 한국은행(ECOS)기상,정부 예산 등
	소셜 네트워크 활동	다양한 SNS 데이터
데이터 형태별		정형,반정형,비정형 (숫자, 텍스트, 오디오, 비디오 등)

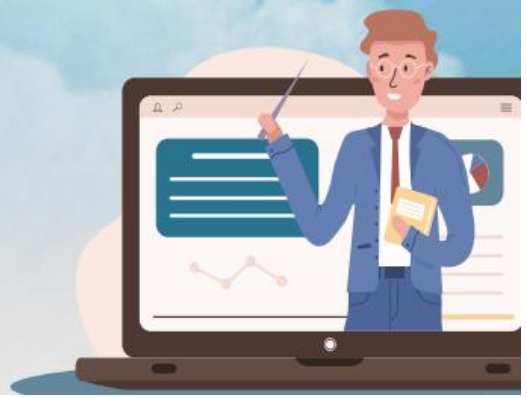
출처 : 데이터 사이언스 p.86~88

Preview

축산 ICT융복합 확산사업

컨설턴트 교육

2021.02.10.



농림축산식품부 EPIS 농림수산식품교육문화정보원

Preview

양돈, 양계, 낙농분야 주요 내용

분야	양돈	양계(산란계/육계)	양계(종계)	낙농
지원대상	축산업등록 경영체	축산업등록 경영체	축산업등록 경영체	축산업등록 경영체
지원내용	▪ 군사급이기 ▪ 자동급이기 ▪ 사료믹스급이기 ▪ 컴퓨터액상급이기 ▪ 돈선별기 ▪ 사료빈관리기 ▪ 음수관리기 ▪ 모돈발정체크기 ▪ 환경관리기 (온도, 습도, 정전, 화재 등) (팬, 쿨링패드, 냉방기, 난방기) 농장 기상대 (온도, 습도, 풍향, 풍속 등) ▪ CCTV, 차량출입장치 ▪ 악취저감장치 ▪ **분뇨처리장치** ▪ 양돈생산경영관리 프로그램 ▪ PC	▪ 부화기 ▪ 자동급이기(사료빈관리기 포함) ▪ 자동급수기(음수관리기 포함) ▪ 난선별기 ▪ 체중기 ▪ 사료빈관리기 ▪ 음수관리기 ▪ 계사환경관리기 (온도, 습도, 정전, 화재 등) (팬, 쿨링패드, 냉방기, 난방기 등) ▪ 농장 기상대 (온도, 습도, 풍향, 풍속) ▪ CCTV, 차량출입장치 ▪ 악취저감장치 ▪ **분뇨처리장치** ▪ 양계생산경영관리 프로그램 ▪ PC		▪ 자동포유기 ▪ 체중측정기 ▪ 발정탐지기/분만알리미 ▪ 착유기 ▪ 로봇착유기 ▪ 자동급이기(개체급이정보 포함) ▪ 음수관리기 ▪ 사료빈관리기 ▪ 조사료분석기/유성분 분석기 ▪ TMR배합기 ▪ 환경관리기, 농장기상대 (온도, 습도, 강우, 풍속, 화재 등) (팬, 안개분무기, 천장개폐기, 윈치커텐) ▪ CCTV ▪ 차량출입장치 ▪ **분뇨처리장치** ▪ 낙농생산경영관리 프로그램 ▪ PC

Preview

대분류	정보통신
중분류	정보기술
소분류	정보기술 전략 · 계획

세분류	능력단위	학습모듈명
정보기술전략	빅데이터 분석 기획	빅데이터 분석 기획
정보기술 컨설팅	**빅데이터 수집**	**빅데이터 수집**
정보기술기획	빅데이터 저장	빅데이터 저장
SW제품기획	빅데이터 처리	빅데이터 처리
빅데이터분석	분석용 데이터 탐색	분석용 데이터 탐색
	통계 기반 데이터 분석	통계 기반 데이터 분석
	머신러닝 기반 데이터 분석	머신러닝 기반 데이터 분석
	텍스트 마이닝 기반 데이터 분석	텍스트 마이닝 기반 데이터 분석
	빅데이터 분석 결과 시각화	빅데이터 분석 결과 시각화

Preview

1-1. 기초 데이터 수집

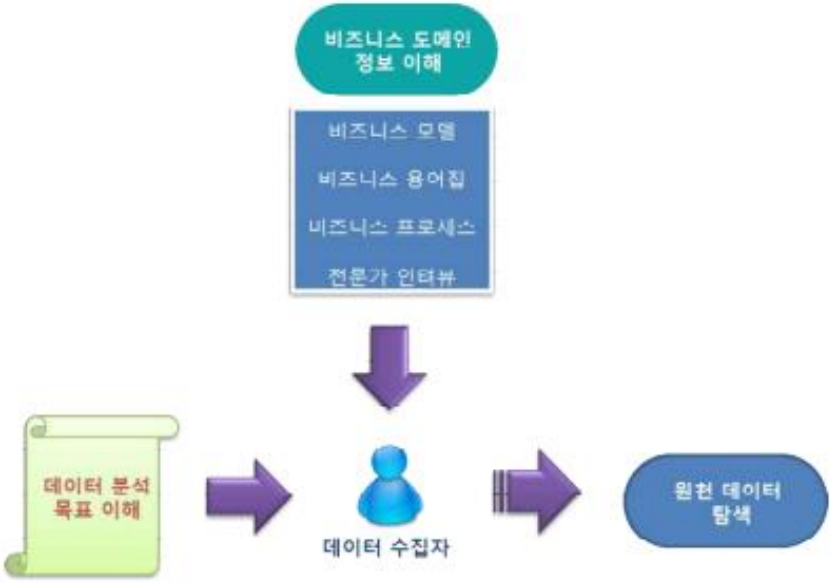
학습 목표

- 분석 목표 달성을 위한 데이터 수집을 위해 비즈니스 프로세스 영역으로부터 기초 자료를 수집할 수 있다.

필요 지식

[1] 비즈니스 도메인과 원천데이터 정보

데이터 수집을 위해서 데이터 수집자는 데이터 분석 목표를 이해하고, 비즈니스 도메인에 대한 이해를 바탕으로 원천 데이터를 탐색해야 한다.



[그림 1-1] 비즈니스 도메인의 이해를 통한 원천데이터 탐색

데이터 사이언스 프로세스

문제 정의	해결하려는 문제를 명확이 정의하는 것
전략 수립	문제 해결을 위해 어떤 데이터를 어떻게 분석할 것인지 결정하는 것
데이터 수집	분석에 필요한 데이터를 수집하는 것 (데이터 분석 전체 과정의 70~80% 시간 소요)
데이터 분석	패턴 찾기, 분류, 예측 등 분석 모델을 구현하는 것
결과 적용	분석 결과를 실제 상황에 적용하고 성능을 개선하는 것

출처 : 데이터 사이언스 개론, p27~33

Preview

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

① 파일 읽기

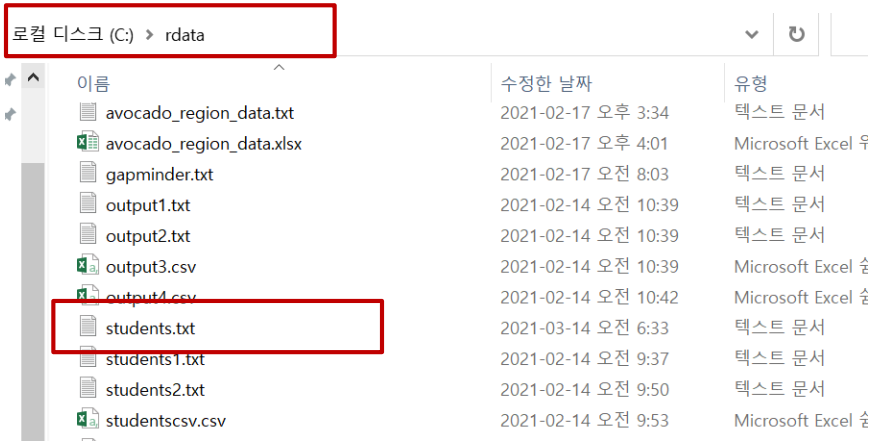
- 파일을 읽을 때는 read.table이나 read.csv 함수를 사용한다.
- read.table 함수: 일반 텍스트 파일을 읽을 때 사용
- Read.csv함수는 CSV(Comma-Separated Values)파일을 읽을 때 사용

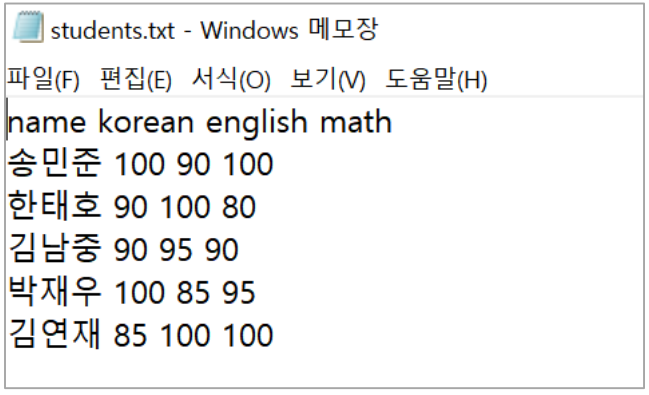
```
>?read.table


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)
```

4.1 파일 읽고 쓰기

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

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



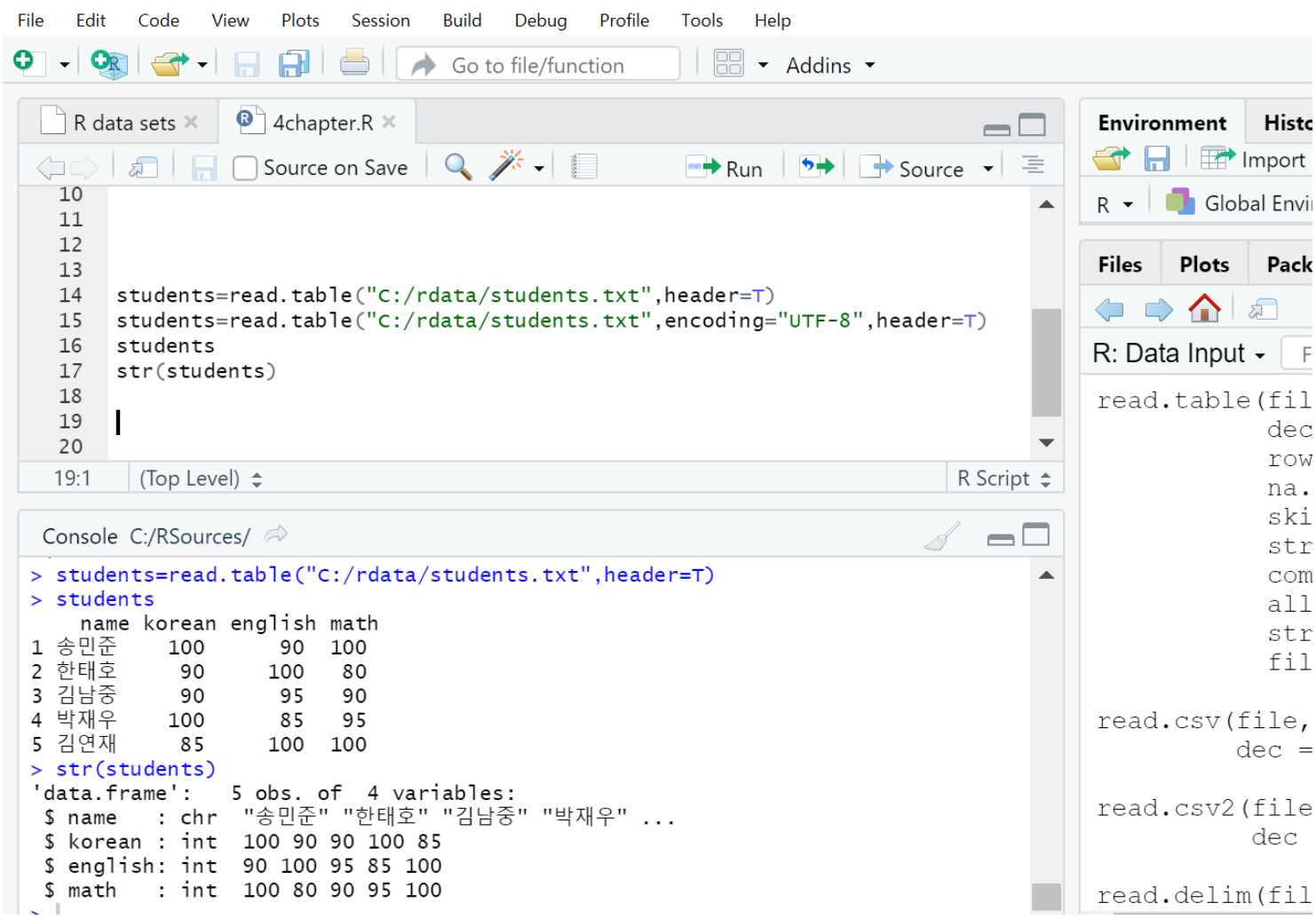


R: Data Input — Find in Topic

```
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

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



4.1 파일 읽고 쓰기

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

```
Console C:/RSources/
> students = read.table("c:/Rdata/students.txt",he
ader=T)
Error in make.names(col.names, unique = TRUE) :
  '<ec><9d><b4>由<84>'에서 유효하지 않은 멀티바이트 문
자열이 있습니다
```

```
> str(students)
'data.frame':   5 obs. of  4 variables:
 $ name   : chr  "송민준" "한태호" "김남중" "박재우" ...
 $ korean : int  100 90 90 100 85
 $ english: int  90 100 95 85 NA
 $ math   : int  100 80 90 95 100
```

4chapter.R* | R data sets

Source on Save | Run | Source

```
12
13
14 students=read.table("C:/rdata/students.txt",header=T)
15 students=read.table("C:/rdata/students.txt",encoding="UTF-8",header=T)
16 students=read.table("C:/rdata/students1.txt",sep=",",header=T)
17 students=read.table("C:/rdata/students1.txt",sep=",",header=T,as.is=T)
18 students=read.table("C:/rdata/students1.txt",sep=",",header=T,as.is=T,na.strings="NA")
```

Data를 읽은 후에 데이터 형에 대해 확인하는 과정이 필요함

4.1 파일 읽고 쓰기

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

```
students=read.table("C:/rdata/students1.txt",sep=",",header=T)
students=read.table("C:/rdata/students1.txt",sep=",",header=T,as.is=T)
students=read.table("C:/rdata/students1.txt",sep=",",header=T,as.is=T,na.strings="NA")
```

students1.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,NA,100
```

Console C:/RSources/

```
> students
    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      NA  100
> str(students)
'data.frame':   5 obs. of  4 variables:
 $ name   : chr  "송민준" "한태호" "김남중" "박재우" ...
 $ korean : int  100 90 90 100 85
 $ english: int  90 100 95 85 NA
 $ math   : int  100 80 90 95 100
```

Thank you


2007년도 대륙별 통계
80
70
60
50
40
300
1000
3000
10000
30000
gdpPercap
continent
Africa
Americas
Asia
Europe
Oceania
pop
2.50e+08
5.00e+08
7.50e+08
1.00e+09
1.25e+09