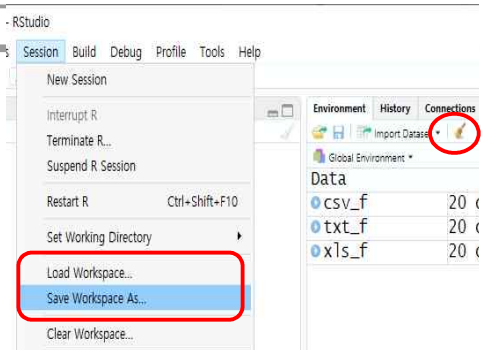
06 외부데이터 사용 및 데이터 탐색

학습목표

1. 외부 데이터인 TXT, CSV, 엑셀 데이터를 불러오고, R 데이터를 외부 데이터로 저장할 수 있다.
2. R 데이터를 R로 저장 및 불러오기 할 수 있고, R 데이터를 삭제할 수 있다.
3. 원본데이터 관찰, 요약통계값, 시각화를 통해 분석하고자 하는 데이터를 전체적으로 살펴볼 수 있다.

강의내용



R 데이터 관리

- R 데이터로 저장 : save() 사용하여 .RData인 파일로 저장
- R 데이터 불러오기 : load()사용하여 .RData 파일을 불러오기
- 현재 사용 중인 데이터 목록 확인 : ls() 함수
- 사용 중인 데이터 삭제 : rm() 함수 ※ rm(list=ls())

외부 데이터 사용

종류	읽어오기	저장하기
TXT 데이터	read.table()	write.table()
CSV 데이터	read.csv()	write.csv()
엑셀 데이터	readxl::read_excel()	writexl::read_excel()

[참고] 패키지 별 파일 read/write 함수

패키지	함수
base 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[1]

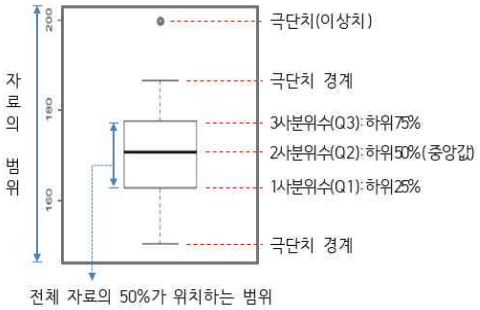
```
#  외부 데이터 읽어오기
csv_r1 <- read.csv("csv_exam.csv")                    # file.choose() 함수로 파일 선택 가능
csv_r2 <- read.csv("csv_exam.csv", header = FALSE)       # 첫 행을 머리글로 인식하지 않음
txt_r1 <- read.table("c:/CodingR/txt_exam.txt",header=TRUE)    # 첫 행을 머리글로 인식
txt_r1 <- read.table("txt_exam.txt",header=TRUE, sep=",")      # 텍스트에서 (,)로 열 구분
xls_r1 <- read_xlsx("xls_exam.xlsx" )                          # readxl패키지를 이용
xls_r2 <- read_xlsx("xls_exam.xlsx", sheet="final" )           # "final" 시트 읽어오기
xls_r3 <- read_xlsx("xls_exam.xlsx", sheet=3, skip=2 )         #  3번째 시트, 2줄 건너뛰고 읽어오기

#  읽어온 자료 확인하기
head(csv_r1)
str(csv_r1)
View(csv_r1)       # 뷰어창에서 데이터 확인

#  외부 데이터로 저장하기
write.csv(csv_r1, "csv_write1.csv")                   # 기본적으로 행이름 추가 저장
write.csv(csv_r1, "csv_ewrite2.csv", row.names=FALSE) # 행이름 저장 안함
write.table(txt_r1, "txt_write1.txt")                      # 기본적으로 행이름 추가 저장
write.table(txt_r1, "txt_write2.txt", row.names=FALSE)     # 행이름 저장 안함
write.table(txt_r1, "txt_write2.txt", append=TRUE)         # 기존 파일 뒤에  데이터 추가 저장
writexl::write_xlsx(xls_r1, "xls_write1.xlsx")         # readxl패키지를 이용
writexl::write_xlsx(xls_r1, "xls_write2.xlsx", col_names = FALSE)       # 머리글 행 저장 안함

#  저장된 파일 확인하기 : 파일 목록에서 [View File] 서브메뉴 선택
```

1) 참고사이트 : R에서 빠르게 데이터 읽기/쓰기 (http://bit.ly/2NFVR4C)

06 외부데이터 사용 및 데이터 탐색

데이터 개관(전체적으로 탐색)

- 데이터의 구성 형태를 파악하기 위해 전체적으로 살펴보기
  . 항목 개수 확인 nrow(), 행과 열 개수 확인 dim()
  . 데이터 앞쪽/뒤쪽 확인하기 head()와 tail(), 전체적으로 데이터 관찰 View()
  . 데이터 속성 확인 str()
- 데이터 특성을 파악하기 위해 각 속성값 살펴보기
  . 요약 통계와 시각화를 활용하여 데이터의 특성을 파악할 수 있다.
  . 요약 통계량 확인
    summary() : 숫자형 변수의 경우, 최소값, 최대값, 1사분위수, 3사분위수, 중앙값, 평균 계산
    팩터형 변수의 경우, 출현 회수 계산
    quantile() : 데이터를 순서대로 정렬할 때 25%(Q1), 50%(Q2), 75%(Q3)에 해당하는 분위수 계산
    median(), mean(), max(), min() : 중앙값, 평균, 최대, 최소
  . 빈도나 분포 확인
    table() : 빈도표
    hist() : 히스토그램, 값(숫자)의 빈도도를 막대그래프로 표현
    qplot() : 막대그래프로 빈도 표현 qplot()
  . 값의 범주 확인
    boxplot() : 상자그림을 통해 데이터의 분포 및 범주 확인,
    이상값 여부 확인 가능
    boxplot.stats() : 상자그림의 통계치 확인
    (사분위수, 관측값 개수, 신뢰구간, 이상값)




```
#  외부데이터 읽어와서 전체적으로 살펴보기
st <- read.csv("studentlist.csv", header = TRUE)

head(st, 10)            # 데이터 앞부분 확인
tail(st, 10)            # 데이터 뒷부분 확인
View(st)                # 뷰어창에서 확인

nrow(st)                # 데이터의 행 수 확인
dim(st)                 # 행과 열 수 확인
str(st)                 # 자료의 속성 확인 : 관측치, 변수, 자료형

summary(st)             # 요약 통계량 : 변수들의 특성 파악(예: 혈액형별 인원수, 키의 평균 등)
summary(st$키)          # 키에 대한 특성 파악

st <- read.csv("studentlist.csv", header = TRUE, stringsAsFactors = F)
summary(st)             # 요약 통계량 => 형액형별 인원수 파악할 수 없다 (문자형 변수)

quantile(st$키)         # 분위수
median(st$키);mean(st$키); max(st$키); min(st$키) # 중앙값, 평균, 최대, 최소

st$성별                 # 160개 출력
distinct(st, 성별)      # 2개 출력

boxplot(st$키)          # 상자그림
boxplot.stats(st$키)    # 상자그림 통계치 출력

table(st$성별)          # 빈도표
hist(st$키)             # 히스토그램
qplot(st$성별)          # 값의 빈도를 막대그래프로 표현
```