R프로그램 기초: 외부데이터 로딩

Jinseog Kim
Dongguk University
jskim1986@gmail.com

2018-03-15

Contents

1 외부 텍스트 파일을 R에서 가져오기 3
1.1 외부 text 파일 3
1.2 read.table 함수 4
1.3 read.csv 함수 5
1.4 예제: 수강생 자료 7
1.5 readLines 10

2 Excel파일 읽고 쓰기 12
2.1 Excel파일 12
2.2 Excel관련 R 패키지 13

2.3 readxl 14
2.4 XLConnect 15

1 외부 텍스트 파일을 R에서 가져오기

1.1 외부 text 파일

 외부파일을 다음의 형식을 만족

1. 파일의 첫 번째 줄은 변수명을 지정
2. 관측치을 변수명에 대응하는 순서대로 입력
3. 예) 위의 형식에 의하여 작성된 외부파일(titanic.txt)

```
Surv N Class Age Sex
20 23 Crew Adult Female
192 862 Crew Adult Male
1 1 First Child Female
5 5 First Child Male
13 13 Second Child Female
```

1.2 read.table 함수

 예제 데이터를 데이터프레임(titanic)으로 변환

```
titanic <- read.table("titanic.txt", header=T)
head(titanic)
```

```
##   Surv   N  Class   Age    Sex
## 1   20  23   Crew Adult Female
## 2  192 862   Crew Adult   Male
## 3    1   1  First Child Female
## 4    5   5  First Child   Male
## 5   13  13 Second Child Female
```

1.3 read.csv 함수

 데이터가 Excel 파일인 경우 CSV(Comma Separated Values)포맷으로 변환 저장

```
Surv,N,Class,Age,Sex
20,23,Crew,Adult,Female
192,862,Crew,Adult,Male
1,1,First,Child,Female
5,5,First,Child,Male
13,13,Second,Child,Female
```

 read.csv 함수를 이용

```
my.table <- read.csv("titanic.csv") ## file name
my.table
```

```
##   Surv   N  Class   Age    Sex
## 1   20  23   Crew Adult Female
## 2  192 862   Crew Adult   Male
## 3    1   1  First Child Female
## 4    5   5  First Child   Male
## 5   13  13 Second Child Female
```

 read.csv 예제

1. 과목 홈페이지의 자료(mtcars.txt)의 링크를 복사

   http://datamining.dongguk.ac.kr/lectures/2018-1/R/mtcars.txt

2. 아래의 R 코드

```r
mtcars <- read.csv("http://datamining.dongguk.ac.kr/lectures/2018-1/R/mtcars.txt")
head(mtcars)
```

```
##                   X  mpg cyl disp  hp drat    wt  qsec vs am gear carb
## 1         Mazda RX4 21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## 2     Mazda RX4 Wag 21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## 3        Datsun 710 22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## 4    Hornet 4 Drive 21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## 5 Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## 6           Valiant 18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

1.4 예제: 수강생 자료

student.csv자료 읽기

```
# fileEncoding = "CP949": csv파일에서 한글 입력 코딩방식, CP949 또는 UTF-8
student <- read.csv("http://datamining.dongguk.ac.kr/lectures/2018-1/R/student.csv",
                    fileEncoding = "CP949")
head(student)
```

```
##           학과 학번   성명 이수구분
## 1 응용통계학과 2014 유현수     전공
## 2   수학교육과 2014 이진형     자선
## 3 응용통계학과 2015 김범석     전공
## 4 응용통계학과 2015 김도환     전공
## 5 응용통계학과 2015 정석환     전공
## 6 응용통계학과 2016 이재용     전공
```

```
names(student)
```

```
## [1] "학과"     "학번"     "성명"     "이수구분"
```

```r
dim(student)
```

```
## [1] 25  4
```

 student데이터프레임에서 학번별 수강생 수 구하기

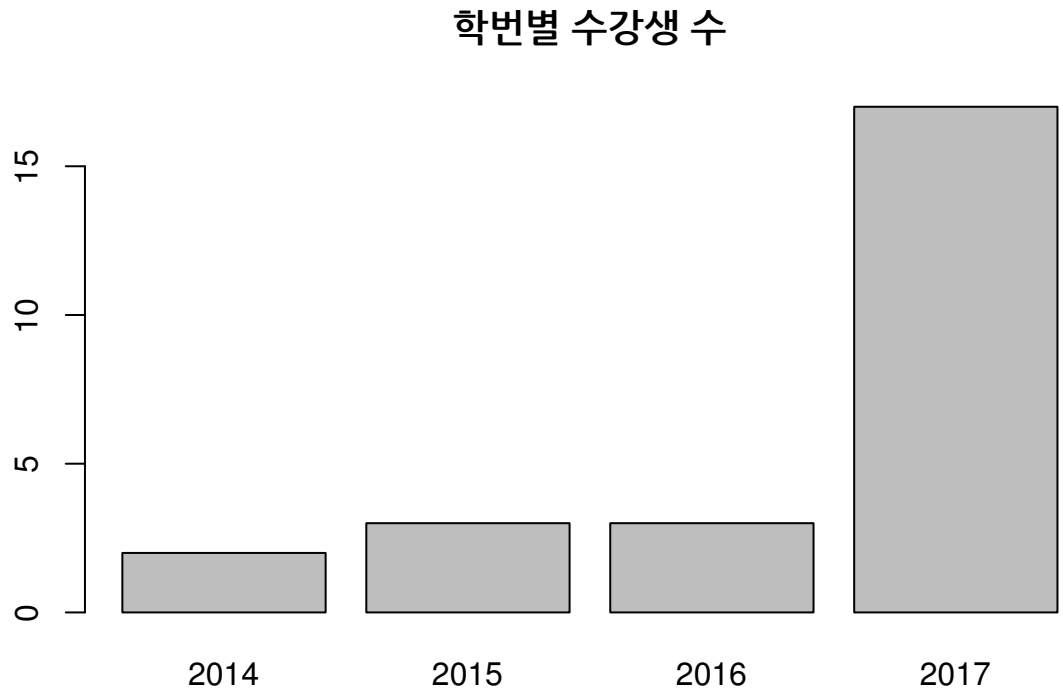
```
unique(student$학번)
```

```
## [1] 2014 2015 2016 2017
```

```
(nid <- table(student$학번))
```

```
##
## 2014 2015 2016 2017
##    2    3    3   17
```

```
barplot(nid, main="학번별 수강생 수")
```



1.5 readLines

- 라인 단위로 읽어오기

```
con <- "http://datamining.dongguk.ac.kr/lectures/2017-1/R/mtcars.txt"
x <- readLines(con)
x1 <- stringr::str_split(x, ",", simplify=T)
x2 <- data.frame(x1[-1,], stringsAsFactors = F)
```

```r
  names(x2) <- x1[1,]
  head(x2, 4)
```

```
##                 "" "mpg" "cyl" "disp" "hp" "drat"  "wt" "qsec" "vs" "am"
## 1      "Mazda RX4"    21     6    160  110    3.9  2.62  16.46    0    1
## 2  "Mazda RX4 Wag"    21     6    160  110    3.9 2.875  17.02    0    1
## 3     "Datsun 710"  22.8     4    108   93   3.85  2.32  18.61    1    1
## 4 "Hornet 4 Drive"  21.4     6    258  110   3.08 3.215  19.44    1    0
##   "gear" "carb"
## 1      4      4
## 2      4      4
## 3      4      1
## 4      3      1
```

2 Excel파일 읽고 쓰기

2.1 Excel파일

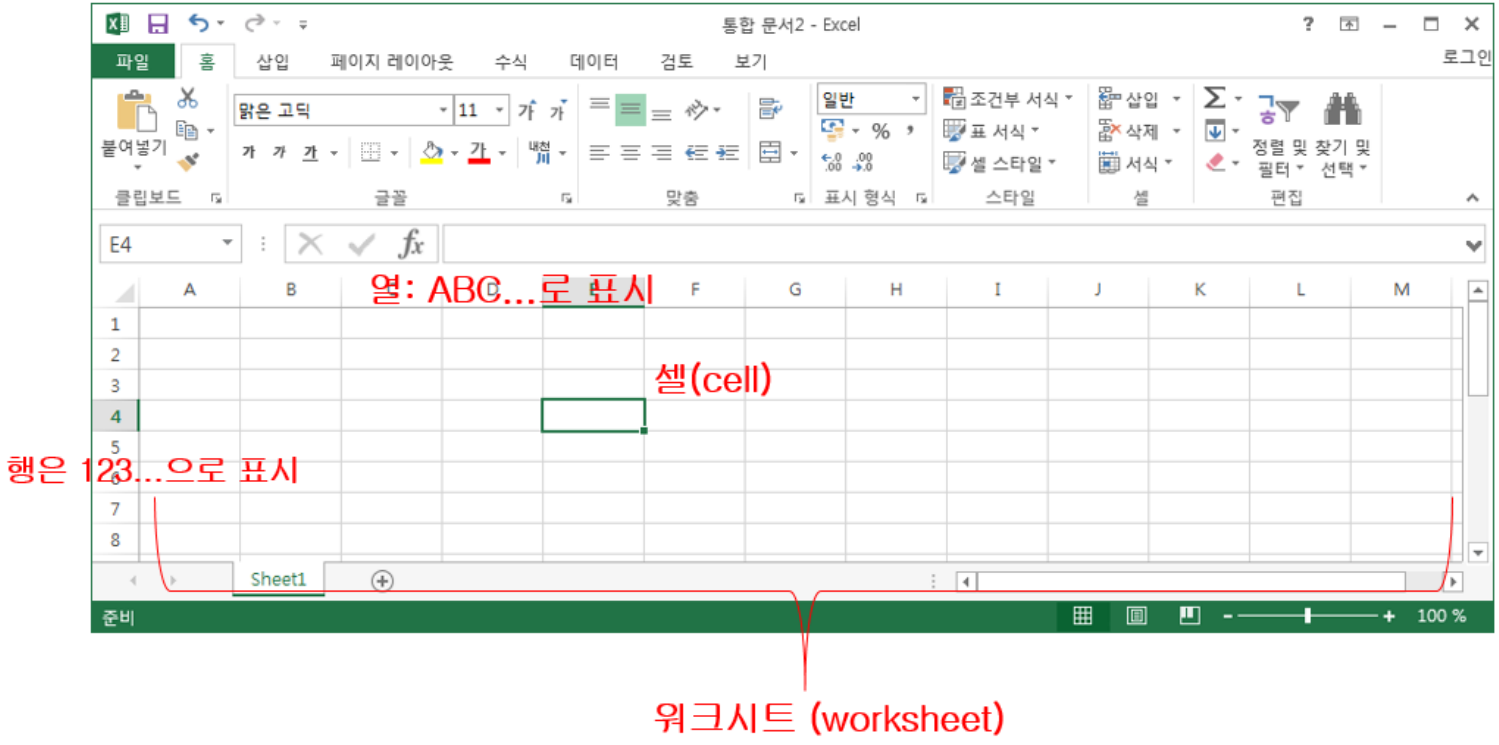


Figure 1: 엑셀파일

2.2 Excel관련 R 패키지

- XLConnect
  - java 이용
  - 많은 메모리 사용
  - 엑셀파일 읽기/쓰기 가능
- readxl
  - 적은 메모리 사용으로 빠르게 데이터를 읽어옴
  - 결과는 데이터프레임이 아니라 tibble형식으로 저장됨 - tibble형식은 데이터프레임의 확장된 데이터클래스임

2.3 readxl

```
# install.packages("readxl", dependencies=T)
```

- Excel파일 읽기
  - 1980~2015년까지 연도별/성별 고용율 자료

```
library(readxl)
o <- read_excel("고용률.xlsx", sheet=1); head(o)
```

```
## # A tibble: 6 x 4
##    연도  전체  남자  여자
##   <dbl> <dbl> <dbl> <dbl>
## 1  1980  55.9  71.7  41.3
## 2  1981  55.9  71.5  41.2
## 3  1982  56.1  70.9  42.3
## 4  1983  55.3  69.8  41.9
## 5  1984  53.7  68.7  39.8
## 6  1985  54.3  68.7  40.9
```

2.4 XLConnect

```
# install.packages("XLConnect")
```

1. XLConnect를 이용한 Excel파일 읽기

    1. loadWorkbook, readWorksheet

```
# install.packages("XLConnect")
# 패키지 로드
library(XLConnect)
# 파일 로딩
wb <- loadWorkbook("고용률.xlsx")
# 첫번째 워크시트 읽기
df <- readWorksheet(wb, sheet=1)
head(df)
```

```
##   연도 전체 남자 여자
## 1 1980 55.9 71.7 41.3
## 2 1981 55.9 71.5 41.2
## 3 1982 56.1 70.9 42.3
## 4 1983 55.3 69.8 41.9
```

```
## 5 1984 53.7 68.7 39.8
## 6 1985 54.3 68.7 40.9
```

2. `readWorksheetFromFile`: 파일에서 지정된 워크시트를 직접 읽어오기

```r
sheet=1, # 시트번호
endRow=10, # 읽어 올 시트의 마지막 행
startCol=1, # 시작 열번호
endCol=2    # 마지막 열번호
```

```r
  df2 <- readWorksheetFromFile("고용률.xlsx", sheet=1,
            endRow=10, startCol=1, endCol=2)
  head(df2)
```

```
##   연도 전체
## 1 1980 55.9
## 2 1981 55.9
## 3 1982 56.1
## 4 1983 55.3
## 5 1984 53.7
```

```
## 6 1985 54.3
```

2. R data.frame을 Excel파일에 쓰기

 연도별 고용률 자료(df)를 읽어서 기간을 2000년 이전과 2000년 이후로 나누고, 각 기간동안 전체 및 성별 고용률의 평균을 구하라.

1. 기간을 2000년 이전과 2000년 이후로 구분하는 변수 생성

```
df$period <- "2000년 이전";
df$period[df$연도 >= 2000] <- "2000년 이후"
```

2. 2000년 이전과 2000년 이후의 평균 계산

```
tot <- rbind(tapply(df$전체, df$period, mean),
             tapply(df$남자, df$period, mean),
             tapply(df$여자, df$period, mean))
rownames(tot) <- names(df)[2:4]
tot
```

```
##      2000년 이전 2000년 이후
## 전체      57.465    59.44375
## 남자      71.425    71.10625
## 여자      44.445    48.36250
```

3. 결과를 엑셀 파일로 저장

```r
# 파일 생성 및 열기
wb <- loadWorkbook("ex.xlsx", create = TRUE)
# 시트 생성
createSheet(wb, name = "summary")
# 시트에 결과 저장
writeWorksheet(wb, tot, sheet = 1, header=TRUE)
# 파일 닫기
saveWorkbook(wb)
```