

외부 데이터 파일 불러오기

1. csv 파일 읽기

CSV 파일이란 **Comma Separated Values**의 약자로 콤마(쉼표 ,)로 구분된 텍스트 파일을 의미한다.

read.csv() 함수에서 sep 옵션이 콤마(,)로 기본 지정되어 있으며, header 옵션의 기본값이 TRUE로 지정되어 있다.

```
read.csv( file, # 가져올 파일의 경로 혹은 URL (필수)
          header=T # 첫 번째 행을 헤더로 사용할 것인지 여부 (기본값=T)
          sep="," # 각 열(또는 데이터)을 구분하는 글자 (기본값=콤마',')
          row.names = c(...), # 행 이름 설정. 데이터를 불러온 후 지정하는게 좋음
          col.names = c(...), # 열 이름 설정. 데이터를 불러온 후 지정하는게 좋음
          na.strings = "NA", # NA 처리를 하고자 하는 문자열 지정
          fileEncoding="utf-8 또는 euc-kr") # 파일 인코딩 (Mac을 사용할 경우 필수)
```

```
separator: one blank(띄어쓰기) : sep = " "
separator: comma(,) : sep = " , "
separator: tab : sep = " \t "
na.strings=" " : 빈칸을 결측값(NA)으로 인식
```

옵션(**header=F**)을 추가하면 첫 줄이 컬럼 이름으로 인식되지 않고 일반 데이터로 인식된다.

```
KOSPI <- read.csv("~/R-Programming/data/KOSPI_20230330.csv",
                  fileEncoding="euc-kr") ; KOSPI # 23.3.30일 데이터1)
data.frame(KOSPI)
str(KOSPI)
```



View() 함수는 데이터셋 내용을 **source** 창 화면에 별도로 출력한다. 데이터셋이

1) 자료: <http://data.krx.co.kr/contents/MDC/MDI/mdiLoader/index.cmd?menuId=MDC0301>

어떤 데이터 구조인지 상시 확인해야 하는 경우 별도로 소스 창에 띄워 놓고 참조하면 유용하게 사용할 수 있다. **V**는 반드시 대문자로 작성해야 함.

View(KOSPI)

View(iris)

2. 엑셀파일 가져오기

(1) excel 파일(xls: 2003 이하 버전, xlsx: 2007 이상 버전)을 처리하기 위한 패키지 설치 및 로드

R의 기본 기능에서는 엑셀 파일을 읽어 올 수 없다. 따라서 추가 **package**의 설치가 필요하다. package란 쉽게 말해서, R의 추가기능을 사용하기 위해 설치하는 파일이라고 보면 된다. package를 설치하기 위해서는 인터넷 연결이 필요하며, 동일 컴퓨터에서 package 설치한 한 번만 하면 된다.

install.packages("package")

install.packages("readxl") # 설치할 때 패키지명에 쌍따옴표가 필요하다.



참고. RStudio 상에서 R 패키지를 설치하는 방법은 메뉴 상단의

Tools > Install Packages...

를 클릭하면 Install Packages 팝업 메뉴가 나오고 원하는 패키지명을 입력하면 된다.

① **library()** 명령어를 실행시키면 설치된 전체 패키지가 검색된다.

② **.libPaths()** 명령어로 현재 패키지가 설치된 디렉토리를 확인할 수 있다.

package를 설치하고 나서, R과의 연결을 위해 package의 내용을 RAM(메모리)에 올려야 한다. 이 과정은 R을 재시작하고, 다른 필요한 패키지를 설치할 때마다 이 과정을 실행해야 한다. 라이브러리는 도서관이라고 생각하면 되고, 패키지는 도서관에서 대여하는 책이라고 생각할 수 있다. 즉, 책(package)가 필요하면 **library()** 함수로 대여해서 데이터 분석과 작업에 사용한다.

library(package)

```
library(readxl) # 책(readxl)를 도서관(library)에 등록
```

(2) 엑셀파일 가져오기

- 특별한 처리를 하지 않는 이상 텍스트나 csv와 달리 현재 컴퓨터에 저장되어 있는 파일만 가져올 수 있다.
- xlsx 패키지에서 제공되는 기본적인 엑셀 불러오기 함수는 `read_excel()` 이다.

```
read_excel( file,                # 엑셀파일의 경로 (필수)
            sheet="...",        # 시트의 이름 (필수), sheet가 첫 번째 있는 경우에는 sheet=1
            range = "...",      # 읽어 들일 파일의 범위 (ex: A1:D19)
            col_names=T ) # 엑셀의 첫 행을 컬럼 이름으로 사용할지 여부(기본값=T)
```

```
KOSDAQ <- read_excel("~/R-Programming/data/KOSDAQ_20230330.xlsx",
                    sheet=1) ; KOSDAQ # 23.3.30일 데이터2)
```

```
data.frame(KOSDAQ)
```

```
str(KOSDAQ)
```

```
tibble [52 × 10]
```

tibble은 data.frame 자료 형태와 근본적으로 큰 차이가 없다.

테이블 출력 결과를 해석하자면, “# A tibble: 52x10”는 데이터 구조를 설명한다. 즉 행렬형태의 데이터로 52개의 행과 10개의 열을 갖는 것을 의미한다.

```
View(KOSDAQ)
```

3. 데이터에 접근하기

(1) 데이터의 개요 확인

① 각 열의 이름 확인

```
컬럼이름 <- names(KOSPI) ; 컬럼이름
```

[1] "지수명"	"종가"	"대비"	"등락률"
[5] "시가"	"고가"	"저가"	"거래량"

2) 자료: <http://data.krx.co.kr/contents/MDC/MDI/mdiLoader/index.cmd?menuId=MDC0301>

[9] "거래대금" "상장시가총액"

② 데이터의 크기 확인 (행과 열의 수)

```
크기 <- dim(KOSPI) ; 크기
```

[1] 47 10 # 47행 10열(10개의 변수에 각각 47개의 객체)

③ 상위 6개 항목 확인하기: head()

head() 는 데이터의 상위 6개 항목을 보여주는 명령어이다. 보고 싶은 데이터의 개수는 사용자 임의로 지정할 수 있으며, 지정하지 않으면 기본값으로 6개만 출력이 된다. 예를 들어 10개의 데이터를 보고 싶다면 head(파일명, 8) 형태로 입력한다.

```
상위데이터 <- head(KOSPI, 8) ; 상위데이터 # 8×10 행렬
```

④ 하위 6개 항목 확인하기

tail() 는 하위 6개 항목(기본값)을 보여주는 명령어.

```
하위데이터 <- tail(KOSPI, 3) ; 하위데이터 # 3×10 행렬
```

⑤ View() 명령어를 이용하시면 데이터 전체를 source 창에서 확인할 수 있다.

View(KOSPI)

[실습] 통계청, 교육부, 「초중고사교육비조사」, 2022년 학교급별 사교육비 총액

https://kosis.kr/statHtml/statHtml.do?orgId=101&tblId=DT_1PE003&conn_path=I2

- (1) (다운로드) 학교급별_사교육비_총액.csv 파일을 R에서 불러오기를 하여라.
- (2) source 창에 전체 데이터를 확인하여라.
- (3) 데이터의 상위 10개 항목을 출력하여라.

(2) 열 단위로 접근하기

데이터프레임이름\$열이름

- ① R 자체 기능으로서의 출력 기능: print(KOSPI\$등락률) 또는 KOSPI\$등락률

[실습] KOSPI에서 **지수명**과 **등락률**을 데이터 프레임 형태로 출력하는 R-코드를 작성하라

(3) 행 단위로 접근하기

① 특정 행 하나만 가져오기

데이터프레임이름['인덱스이름(문자형)',] 혹은 **데이터프레임이름**[인덱스번호,]
KOSPI[1,]

② 복수의 행을 가져오기

가져올 행 이름이나 인덱스 번호를 벡터로 구성한다.

KOSPI[c(1, 3),]

(4) 행을 기준으로 열의 일부만 가져오기

KOSPI[1, c('시가', '종가')] # 1행에서 시작일 기준가와 종료일 종가 열만 조회

(5) 단일 값에 접근하기

데이터프레임이름['인덱스이름', '컬럼이름'] 혹은 **데이터프레임이름**[인덱스번호, '컬럼이름']

KOSPI[1, 3]



작업 공간(Working Directory) 설정

불러오는 파일과 작업 중인 R파일의 폴더(경로)가 같으면 file 경로 설정이 편하다.

파일경로 지정방법 1	파일 경로 확인하기 : <code>getwd()</code> 파일 경로 지정하기 : <code>setwd("c:/폴더이름")</code>
파일경로 지정방법 2	RStudio의 메뉴 ' Session - Set Working Directory - To Source File Location '을 클릭하면 console 창에 R파일 있는 경로가 지정된다. (예: <code>setwd("~/R-Programming/data")</code>)