

Big Data



About..

컴퓨터소프트웨어공학과
김 원 일



행과 열에 이름 붙이기

- matrix의 행과 열에 이름 붙이기

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지들 윈도우즈 도움말

> score <- matrix( c( 90, 80, 55, 34, 89, 96, 49, 95, 90, 80, 70, 60 ), nrow = 4 )
> score
      [,1] [,2] [,3]
[1,]   90   89   90
[2,]   80   96   80
[3,]   55   49   70
[4,]   34   95   60
> rownames( score ) <- c( 'kim', 'lee', 'park', 'choi' )
> colnames( score ) <- c( "Eng", "Math", "Com" )
> score
      Eng Math Com
kim    90   89  90
lee    80   96  80
park   55   49  70
choi   34   95  60
> |
```



행과 열에서 정보 확인

- matrix와 data frame에 동일 적용
 - 행과 열 이름 변경과 데이터 접근이 모두 동일한 형태로 구성

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지들 윈도우즈 도움말

> score[ 'kim', ]
  Eng Math  Com
    90   89   90
> score[ 'lee', c( 'Math', 'Com' ) ]
  Math  Com
    96   80
> score[ , "Eng" ]
  kim  lee park choi
    90   80   55   34
> rownames( score )
[1] "kim" "lee" "park" "choi"
> colnames( score )
[1] "Eng" "Math" "Com"
> |
```

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지들 윈도우즈 도움말

> num <- c( 1, 2, 3, 4 )
> log <- c( T, F, F, T )
> df <- data.frame( num, log )
> df
  num  log
1   1 TRUE
2   2 FALSE
3   3 FALSE
4   4 TRUE
> colnames( df ) <- c( 'seq', 'hello' )
> df
  seq hello
1   1 TRUE
2   2 FALSE
3   3 FALSE
4   4 TRUE
> |
```



• iris 데이터 셋

- R에 기본적으로 내장되어 있는 변수
- 150개의 붓꽃에 대한 측정 정보와 품종 정보를 이용하여 구성된 data frame

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지 윈도우즈 도움말

> iris
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1         5.1         3.5          1.4         0.2    setosa
2         4.9         3.0          1.4         0.2    setosa
3         4.7         3.2          1.3         0.2    setosa
4         4.6         3.1          1.5         0.2    setosa
5         5.0         3.6          1.4         0.2    setosa
6         5.4         3.9          1.5         0.2    setosa
7         4.6         3.4          1.4         0.2    setosa
8         5.0         3.4          1.5         0.2    setosa
9         4.4         2.9          1.4         0.2    setosa
10        4.9         3.1          1.5         0.2    setosa
11        5.4         3.7          1.5         0.2    setosa
12        4.8         3.4          1.4         0.1    setosa
13        4.8         3.0          1.4         0.1    setosa
```

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지 윈도우즈 도움말

> str( iris )
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species      : Factor w/ 3 levels "setosa","versicolor",...: 1 1 $
> |
```



• 데이터 처리 예제 - 1

- 데이터들을 필요에 따라 처리하는 예제
- 1, 2 열의 모든 데이터 출력
- 1, 3, 5 열의 모든 데이터 출력

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지들 윈도우즈 도움말

> iris[ , c( 1 : 2 ) ]
      Sepal.Length Sepal.Width
1             5.1           3.5
2             4.9           3.0
3             4.7           3.2
4             4.6           3.1
5             5.0           3.6
6             5.4           3.9
7             4.6           3.4
8             5.0           3.4
9             4.4           2.9
10            4.9           3.1
11            5.4           3.7
```

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지들 윈도우즈 도움말

> iris[ , c( 1, 3, 5 ) ]
      Sepal.Length Petal.Length Species
1             5.1           1.4   setosa
2             4.9           1.4   setosa
3             4.7           1.3   setosa
4             4.6           1.5   setosa
5             5.0           1.4   setosa
6             5.4           1.7   setosa
7             4.6           1.4   setosa
8             5.0           1.5   setosa
9             4.4           1.4   setosa
10            4.9           1.5   setosa
11            5.4           1.5   setosa
```



• 데이터 처리 예제 - 2

- 데이터들을 필요에 따라 처리하는 예제
- 1, 5 열의 모든 데이터를 컬럼명으로 출력
- 1 ~ 5 행의 모든 데이터 출력

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지들 윈도우즈 도움말
> iris[ , c( "Sepal.Length", "Species" ) ]
  Sepal.Length Species
1          5.1   setosa
2          4.9   setosa
3          4.7   setosa
4          4.6   setosa
5          5.0   setosa
6          5.4   setosa
7          4.6   setosa
8          5.0   setosa
9          4.4   setosa
10         4.9   setosa
11         5.4   setosa

RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지들 윈도우즈 도움말
146          6.7 virginica
147          6.3 virginica
148          6.5 virginica
149          6.2 virginica
150          5.9 virginica
> iris[ 1 : 5, ]
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1          3.5          1.4          0.2   setosa
2          4.9          3.0          1.4          0.2   setosa
3          4.7          3.2          1.3          0.2   setosa
4          4.6          3.1          1.5          0.2   setosa
5          5.0          3.6          1.4          0.2   setosa
> |
```



• 데이터 처리 예제 - 3

- 데이터들을 필요에 따라 처리하는 예제
- 1 ~ 5 행의 1, 3 열의 데이터 출력
- 1 ~ 5 행의 1 ~ 3 열의 데이터 출력

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지들 윈도우즈 도움말
[Icons]

1      5.1      3.5      1.4
2      4.9      3.0      1.4
3      4.7      3.2      1.3
4      4.6      3.1      1.5
5      5.0      3.6      1.4
> iris[ 1 : 5, c( 1, 3 ) ]
  Sepal.Length Petal.Length
1          5.1          1.4
2          4.9          1.4
3          4.7          1.3
4          4.6          1.5
5          5.0          1.4
> |
```

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지들 윈도우즈 도움말
[Icons]

1      5.1      1.4
2      4.9      1.4
3      4.7      1.3
4      4.6      1.5
5      5.0      1.4
> iris[ 1 : 5, 1 : 3 ]
  Sepal.Length Sepal.Width Petal.Length
1          5.1          3.5          1.4
2          4.9          3.0          1.4
3          4.7          3.2          1.3
4          4.6          3.1          1.5
5          5.0          3.6          1.4
> |
```



• 데이터 셋의 정보 확인 - 1

- `dim(iris)` : 행과 열의 개수 출력
- `ncol(iris)` : 열의 개수 출력

- `nrow(iris)` : 행의 개수 출력
- `colnames(iris)` : 열 이름 출력. `names()`와 동일 출력

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지들 윈도우즈 도움말

> dim( iris )
[1] 150  5
> nrow( iris )
[1] 150
> ncol( iris )
[1] 5
> colnames( iris )
[1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"
[5] "Species"
```




• 데이터 셋의 정보 확인 - 2

- head(iris) : 시작 6줄의 정보 출력

- tail(iris) : 마지막 6줄의 정보 출력

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지를 윈도우즈 도움말

> head( iris )
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5         1.4         0.2  setosa
2          4.9         3.0         1.4         0.2  setosa
3          4.7         3.2         1.3         0.2  setosa
4          4.6         3.1         1.5         0.2  setosa
5          5.0         3.6         1.4         0.2  setosa
6          5.4         3.9         1.7         0.4  setosa

> tail( iris )
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
145          6.7         3.3         5.7         2.5 virginica
146          6.7         3.0         5.2         2.3 virginica
147          6.3         2.5         5.0         1.9 virginica
148          6.5         3.0         5.2         2.0 virginica
149          6.2         3.4         5.4         2.3 virginica
150          5.9         3.0         5.1         1.8 virginica

> |
```



• 데이터 셋의 정보 확인 - 3

- levels(iris) : 품종 문자열의 종류 보기
- 특정 값으로 설정된 경우 값 확인 가능

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지를 윈도우즈 도움말

> levels( iris )
NULL
> levels( iris[ , 5 ] )
[1] "setosa"      "versicolor" "virginica"
> levels( iris[ , 4 ] )
NULL
> levels( iris[ , 3 ] )
NULL
> levels( iris[ , 2 ] )
NULL
> levels( iris[ , 1 ] )
NULL
> |
```

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지를 윈도우즈 도움말

> iris[ , 5 ]
[1] setosa      setosa      setosa      setosa      setosa
[6] setosa      setosa      setosa      setosa      setosa
[11] setosa      setosa      setosa      setosa      setosa
[16] setosa      setosa      setosa      setosa      setosa
[21] setosa      setosa      setosa      setosa      setosa
[26] setosa      setosa      setosa      setosa      setosa
[31] setosa      setosa      setosa      setosa      setosa
```

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지를 윈도우즈 도움말

[121] virginica  virginica  virginica  virginica  virginica
[126] virginica  virginica  virginica  virginica  virginica
[131] virginica  virginica  virginica  virginica  virginica
[136] virginica  virginica  virginica  virginica  virginica
[141] virginica  virginica  virginica  virginica  virginica
[146] virginica  virginica  virginica  virginica  virginica
Levels: setosa versicolor virginica
> |
```



• 데이터 셋의 정보 확인 - 4

- `table(iris[, "Species"])` : 품종 별로 행의 개수를 출력
- 분류 가능한 값이 존재할 때, 그룹으로 나뉜 개수를 합산하여 출력

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지들 윈도우즈 도움말
[Icons]
> table( iris[ , "Species" ] )

      setosa versicolor  virginica
        50         50         50
> |
```



• 데이터 셋의 정보 확인 - 5

- colSums(iris[, col]) : 컬럼의 합계 구하기
- col은 특이하게 제외할 컬럼 번호를 의미

- colMeans(iris[, col]) : 컬럼의 평균 구하기

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지를 윈도우즈 도움말

> colSums( iris[ , 1 ] )
colSums(iris[, 1])에서 다음과 같은 에러가 발생했습니다: 'x'는 반드시 최소 2$
> colSums( iris[ , -1 ] )
colSums(iris[, -1])에서 다음과 같은 에러가 발생했습니다: 'x'는 반드시 수치$
> colSums( iris[ , -5 ] )
Sepal.Length Sepal.Width Petal.Length
      876.5      458.6      563.7
> colSums( iris[ , -4, -5 ] )
colSums(iris[, -4, -5])에서 다음과 같은 에러가
> colSums( iris[ , c( -4, -5 ) ] )
Sepal.Length Sepal.Width Petal.Length
      876.5      458.6      563.7
> |
```

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지를 윈도우즈 도움말

> colSums( iris[ , -5 ] )
Sepal.Length Sepal.Width Petal.Length Petal.Width
      876.5      458.6      563.7      179.9
> colMeans( iris[ , -5 ] )
Sepal.Length Sepal.Width Petal.Length Petal.Width
   5.843333    3.057333    3.758000    1.199333
> |
```



• 데이터 셋의 정보 확인 - 6

- `rowSums(iris[, col])` : 행의 합계 구하기
- `col`은 제외할 컬럼의 번호

- `rowMeans(iris[, col])` : 행의 평균 구하기

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지를 윈도우즈 도움말

> rowSums( iris[ , -5 ] )
 [1] 10.2  9.5  9.4  9.4 10.2 11.4  9.7 10.1  8.9  9.6 10.8 10.0  9.3  8.5
[15] 11.2 12.0 11.0 10.3 11.5 10.7 10.7 10.7  9.4 10.6
[29] 10.2  9.7  9.7 10.7 10.9 11.3  9.7  9.6 10.5 10.0
[43]  9.1 10.7 11.2  9.5 10.7  9.4 10.7  9.9 16.3 15.6
[57] 15.9 11.6 15.4 13.2 11.5 14.6 13.2 15.1 13.4 15.6
[71] 15.7 14.2 15.2 14.8 14.9 15.4 15.8 16.4 14.9 12.8
[85] 14.4 15.5 16.0 14.3 14.0 13.3 13.7 15.1 13.6 11.6
[99] 11.7 13.9 18.1 15.5 18.1 16.6 17.5 19.3 13.6 18.3
[113] 17.4 15.2 16.1 17.2 16.8 20.4 19.5 14.7 18.1 15.3
[127] 15.6 15.8 16.9 17.6 18.2 20.1 17.0 15.7 15.7 19.1
[141] 17.8 17.4 15.5 18.2 18.2 17.2 15.7 16.7 17.3 15.8
> |
```

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지를 윈도우즈 도움말

> rowMeans( iris[ , -5 ] )
 [1] 2.550 2.375 2.350 2.350 2.550 2.850 2.425 2.525 2.225 2.400 2.700
[12] 2.500 2.325 2.125 2.800 3.000 2.750 2.575 2.875 2.675 2.675 2.675
[23] 2.350 2.650 2.575 2.450 2.600 2.600 2.550 2.425 2.425 2.675 2.725
[34] 2.825 2.425 2.400 2.625 2.500 2.225 2.550 2.525 2.100 2.275 2.675
[45] 2.800 2.375 2.675 2.350 2.675 2.475 4.075 3.900 4.100 3.275 3.850
[56] 3.575 3.975 2.900 3.850 3.300 2.875 3.650 3.300 3.775 3.350 3.900
[67] 3.650 3.400 3.600 3.275 3.925 3.550 3.800 3.700 3.725 3.850 3.950
[78] 4.100 3.725 3.200 3.200 3.150 3.400 3.850 3.600 3.875 4.000 3.575
[89] 3.500 3.325 3.425 3.775 3.400 2.900 3.450 3.525 3.525 3.675 2.925
[100] 3.475 4.525 3.875 4.525 4.150 4.375 4.825 3.400 4.575 4.200 4.850
[111] 4.200 4.075 4.350 3.800 4.025 4.300 4.200 5.100 4.875 3.675 4.525
[122] 3.825 4.800 3.925 4.450 4.550 3.900 3.950 4.225 4.400 4.550 5.025
[133] 4.250 3.925 3.925 4.775 4.425 4.200 3.900 4.375 4.450 4.350 3.875
[144] 4.550 4.550 4.300 3.925 4.175 4.325 3.950
> |
```



• 데이터 셋에서 정보 획득 - 1

- subset(data, condition) : 조건에 해당하는 데이터의 행과 열 값 추출

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지들 윈도우즈 도움말

> subset( iris, Species == 'setosa' )
  Sepal.Length Sepal.Width Petal.Length
1           5.1         3.5          1.4
2           4.9         3.0          1.4
3           4.7         3.2          1.5
4           4.6         3.1          1.6
5           5.0         3.6          1.4
6           5.4         3.9          1.5
7           4.6         3.4          1.6
8           5.0         3.4          1.4
9           4.4         2.9          1.4
10          4.9         3.1          1.4
11          5.4         3.7          1.5
12          4.8         3.4          1.5
13          4.8         3.0          1.4
```

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지들 윈도우즈 도움말

> subset( iris, Sepal.Length > 7.0 )
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
103          7.1         3.0          5.9         2.1 virginica
106          7.6         3.0          6.6         2.1 virginica
108          7.3         2.9          6.3         1.8 virginica
110          7.2         3.6          6.1         2.5 virginica
118          7.7         3.8          6.7         2.2 virginica
119          7.7         2.6          6.9         2.3 virginica
123          7.7         2.8          6.7         2.0 virginica
126          7.2         3.2          6.0         1.8 virginica
130          7.2         3.0          5.8         1.6 virginica
131          7.4         2.8          6.1         1.9 virginica
132          7.9         3.8          6.4         2.0 virginica
136          7.7         3.0          6.1         2.3 virginica
> |
```



• 데이터 셋에서 정보 획득 - 2

- 조건에 해당하는 행 개수 확인 등의 각종 함수 이용 가능
- 획득된 subset()을 별도의 데이터로 분리하여 보관하는 것도 가능

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지를 윈도우즈 도움말

> subset( iris, Sepal.Length > 7.0 )
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
103          7.1         3.0          5.9         2.1 virginica
106          7.6         3.0          6.6         2.1 virginica
108          7.3         2.9          6.3         1.8 virginica
110          7.2         3.6          6.1         2.5 virginica
118          7.7         3.8          6.7         2.2 virginica
119          7.7         2.6          6.9         2.3 virginica
123          7.7         2.8          6.7         2.0 virginica
126          7.2         3.2          6.0         1.8 virginica
130          7.2         3.0          5.8         1.6 virginica
131          7.4         2.8          6.1         1.9 virginica
132          7.9         3.8          6.4         2.0 virginica
136          7.7         3.0          6.1         2.3 virginica

> nrow( subset( iris, Sepal.Length > 7.0 ) )
[1] 12
> |
```

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지를 윈도우즈 도움말

> st1 <- subset( iris, Sepal.Length > 7.0 )
> st1
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
103          7.1         3.0          5.9         2.1 virginica
106          7.6         3.0          6.6         2.1 virginica
108          7.3         2.9          6.3         1.8 virginica
110          7.2         3.6          6.1         2.5 virginica
118          7.7         3.8          6.7         2.2 virginica
119          7.7         2.6          6.9         2.3 virginica
123          7.7         2.8          6.7         2.0 virginica
126          7.2         3.2          6.0         1.8 virginica
130          7.2         3.0          5.8         1.6 virginica
131          7.4         2.8          6.1         1.9 virginica
132          7.9         3.8          6.4         2.0 virginica
136          7.7         3.0          6.1         2.3 virginica

> |
```




• 데이터 셋에서 정보 획득 - 3

- 다중 조건을 이용한 데이터 부분 획득
- 조건을 &로 연결하여 다수의 조건을 적용한 subset을 획득 가능

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지들 윈도우즈 도움말
[Icons]
> st1 <- subset( iris, Sepal.Length > 7.0 & Sepal.Width > 3.0 )
> st1
      Sepal.Length Sepal.Width Petal.Length Petal.Width  Species
110             7.2         3.6          6.1         2.5 virginica
118             7.7         3.8          6.7         2.2 virginica
126             7.2         3.2          6.0         1.8 virginica
132             7.9         3.8          6.4         2.0 virginica
> |
```




• 데이터 셋에서 정보 획득 - 4

- `class(iris)` : 자료구조를 명시적으로 표현
- `is.vector(variable)` : vector 자료형 여부 확인
- `is.matrix(variable)` : matrix 자료형 여부 확인
- `is.data.frame(variable)` : data frame 여부 확인

```
RGGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지를 윈도우즈 도움말

> ls( )
[1] "a"      "df"      "log"      "num"      "score"
[6] "st1"
> class( iris )
[1] "data.frame"
> is.matrix( iris )
[1] FALSE
> is.data.frame( iris )
[1] TRUE
> is.vector( iris )
[1] FALSE
> is.vector( num )
[1] TRUE
> |
```



• 데이터 셋에서 정보 획득 - 5

- data.frame\$Column.Name 으로 특정 열 값만 반환 가능

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지를 윈도우즈 도움말

> iris$Species
[1] setosa      setosa      setosa
[6] setosa      setosa      setosa
[11] setosa      setosa      setosa
[16] setosa      setosa      setosa
[21] setosa      setosa      setosa
[26] setosa      setosa      setosa
[31] setosa      setosa      setosa
[36] setosa      setosa      setosa
[41] setosa      setosa      setosa
[46] setosa      setosa      setosa
[51] versicolor  versicolor  versicolor
[56] versicolor  versicolor  versicolor
[61] versicolor  versicolor  versicolor
[66] versicolor  versicolor  versicolor
```

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지를 윈도우즈 도움말

[146] virginica  virginica  virginica  virginica  virginica
Levels: setosa versicolor virginica
> iris$Sepal.Length
[1] 5.1 4.9 4.7 4.6 5.0 5.4 4.6 5.0 4.4 4.9 5.4 4.8 4.8 4.3
[15] 5.8 5.7 5.4 5.1 5.7 5.1 5.4 5.1 4.6 5.1 4.8 5.0 5.0 5.2
[29] 5.2 4.7 4.8 5.4 5.2 5.5 4.9 5.0 5.5 4.9 4.4 5.1 5.0 4.5
[43] 4.4 5.0 5.1 4.8 5.1 4.6 5.3 5.0 7.0 6.4 6.9 5.5 6.5 5.7
[57] 6.3 4.9 6.6 5.2 5.0 5.9 6.0 6.1 5.6 6.7 5.6 5.8 6.2 5.6
[71] 5.9 6.1 6.3 6.1 6.4 6.6 6.8 6.7 6.0 5.7 5.5 5.5 5.8 6.0
[85] 5.4 6.0 6.7 6.3 5.6 5.5 5.5 6.1 5.8 5.0 5.6 5.7 5.7 6.2
[99] 5.1 5.7 6.3 5.8 7.1 6.3 6.5 7.6 4.9 7.3 6.7 7.2 6.5 6.4
[113] 6.8 5.7 5.8 6.4 6.5 7.7 7.7 6.0 6.9 5.6 7.7 6.3 6.7 7.2
[127] 6.2 6.1 6.4 7.2 7.4 7.9 6.4 6.3 6.1 7.7 6.3 6.4 6.0 6.9
[141] 6.7 6.9 5.8 6.8 6.7 6.7 6.3 6.5 6.2 5.9
> |
```



자료형 변환 - 1

- vector의 matrix 생성 및 data frame 변환
 - vector는 matrix()로 간단히 matrix를 생성할 수 있음
 - matrix 자료형은 별도의 변환 과정 없이 data.frame()으로 즉시 변환이 가능

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지들 윈도우즈 도움말

> class( num )
[1] "numeric"
> num
[1] 1 2 3 4
> mtx <- matrix( num )
> mtx
      [,1]
[1,]     1
[2,]     2
[3,]     3
[4,]     4
> class( mtx )
[1] "matrix" "array"
> is.matrix( mtx )
[1] TRUE
> |
```

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지들 윈도우즈 도움말

> class( mtx )
[1] "matrix" "array"
> is.matrix( mtx )
[1] TRUE
> df <- data.frame( mtx )
> df
      mtx
1      1
2      2
3      3
4      4
> class( df )
[1] "data.frame"
> is.data.frame( df )
[1] TRUE
> |
```



자료형 변환 - 2

- data frame에서 matrix로 변환

- as.matrix(동일 자료형으로 구성된 데이터) : matrix 형태로 자료형이 변환되어 반환

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지들 윈도우즈 도움말
[Icons]

> df <- as.matrix( iris[ , 1 : 4 ] )
> df
      Sepal.Length Sepal.Width Petal.Length P
[1,]           5.1           3.5           1.4
[2,]           4.9           3.0           1.4
[3,]           4.7           3.2           1.3
[4,]           4.6           3.1           1.5
[5,]           5.0           3.6           1.4
[6,]           5.4           3.9           1.7
[7,]           4.6           3.4           1.4
[8,]           5.0           3.4           1.5
[9,]           4.4           2.9           1.4
[10,]          4.9           3.1           1.5
[11,]          5.4           3.7           1.5
[12,]          4.8           3.4           1.6
[13,]          4.8           3.0           1.4
```

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지들 윈도우즈 도움말
[Icons]

[138,]           6.4           3.1           5.5           1.8
[139,]           6.0           3.0           4.8           1.8
[140,]           6.9           3.1           5.4           2.1
[141,]           6.7           3.1           5.6           2.4
[142,]           6.9           3.1           5.1           2.3
[143,]           5.8           2.7           5.1           1.9
[144,]           6.8           3.2           5.9           2.3
[145,]           6.7           3.3           5.7           2.5
[146,]           6.7           3.0           5.2           2.3
[147,]           6.3           2.5           5.0           1.9
[148,]           6.5           3.0           5.2           2.0
[149,]           6.2           3.4           5.4           2.3
[150,]           5.9           3.0           5.1           1.8

> class( df )
[1] "matrix" "array"
> |
```



데이터 찾기 - 1

• 특정 조건 데이터 찾기

- `which(condition)` : 조건에 해당하는 값의 위치를 출력

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지를 윈도우즈 도움말

> which( iris$Sepal.Length == 5.9 )
[1] 62 71 150
> iris[ 62, ]
      Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
62              5.9          3         4.2         1.5 versicolor
> iris[ 71, ]
      Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
71              5.9          3.2         4.8         1.8 versicolor
> iris[ 150, ]
      Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
150              5.9          3         5.1         1.8 virginica
> |
```



• 최대 값 데이터 찾기

- `which.max(condition)` : 조건에 해당하는 값의 위치를 출력
- 조건은 vector 형태 또는 행, 열을 기준으로 구별이 가능한 값만을 허용

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지들 윈도우즈 도움말

> which.max( iris )
which.max(iris)에서 다음과 같은 에러가 발생했습니다:
'list' object cannot be coerced to type 'double'
> which.max( iris$Sepal.Length )
[1] 132
> iris[ 132, ]
      Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
132           7.9         3.8         6.4           2 virginica
> |
```



데이터 찾기 - 3

• 최소 값 데이터 찾기

- `which.min(condition)` : 조건에 해당하는 값의 위치를 출력
- 조건은 vector 형태 또는 행, 열을 기준으로 구별이 가능한 값만을 허용

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지들 윈도우즈 도움말

> iris[ 132, ]
      Sepal.Length Sepal.Width Petal.Length Petal.Width  Species
132           7.9         3.8         6.4           2 virginica

> which.min( iris$Sepal.Length )
[1] 14

> iris[ 14, ]
      Sepal.Length Sepal.Width Petal.Length Petal.Width Species
14           4.3         3         1.1         0.1   setosa

> |
```

Big Data - file



About..

컴퓨터소프트웨어공학과
김 원 일



현재 작업 디렉터리

• 작업 디렉터리 확인 및 설정

- getwd() : get working directory로 현재 작업 중인 디렉터리 정보 반환
- setwd() : set working directory로 작업 디렉터를 지정한 디렉터리로 변경
 - 경로는 단일 '/'로 설정

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지들 윈도우즈 도움말
[85] 5.4 6.0 6.7 6.3 5.6 5.5 5.5 6.1 5.8 5$
[99] 5.1 5.7 6.3 5.8 7.1 6.3 6.5 7.6 4.9 7$
[113] 6.8 5.7 5.8 6.4 6.5 7.7 7.7 6.0 6.9 5$
[127] 6.2 6.1 6.4 7.2 7.4 7.9 6.4 6.3 6.1 7$
[141] 6.7 6.9 5.8 6.8 6.7 6.7 6.3 6.5 6.2 5$
> getwd( )
[1] "C:/Users/YUHAN/Documents"
> setwd( "D:/tmp" )
> getwd( )
[1] "D:/tmp"
> |
```



일반 파일 입출력 - 1

- sink()로 파일에 기록 - 1

- sink("filename") 실행 후부터는 콘솔 대신 파일에 출력이 이루어짐
- print(), cat() 등의 출력 명령어가 모두 아무 출력이 없는 것을 확인할 수 있음
- 출력을 중단하려면 sink() 로 출력 중단을 명시

The screenshot shows two windows. The left window is 'RGui (64-bit) - [R Console]' with a menu bar (파일, 편집, 보기, 기타, 패키지들, 윈도우즈, 도움말) and a toolbar. The console contains the following R code:

```
> getwd( )  
[1] "D:/tmp"  
> sink( 'data.txt' )  
> print( "Result data file" )  
> cat( '3 + 7 =', 3 + 7, '\n' )  
> sink( )  
> |
```

The right window is 'data.txt - Windows 메모장' with a menu bar (파일(F), 편집(E), 서식(O), 보기(V), 도움말(H)). It contains the output of the R code:

```
[1] "Result data file"  
3 + 7 = 10
```

The status bar at the bottom of the text editor shows 'Ln 1, Col 1', '100%', 'Windows (CRLF)', and 'UTF-8'.



일반 파일 입출력 - 2

• sink()로 파일에 기록 - 2

- sink("filename", append = T) 명령은 존재하는 파일 뒤에 추가 기록할 때 사용
- "append = F"가 기본 설정이며, 이전 파일 내용을 지우고 기록할 때 사용
- sink()로 파일 출력을 중단하는 것은 동일

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지들 윈도우즈 도움말

> print( "Result data file" )
> cat( '3 + 7 =', 3 + 7, '\n' )
> sink( )
> sink( 'data.txt', append = T )
> print( df )
> sink( )
> |
```

```
data.txt - Windows 메모장
파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

[1] "Result data file"
3 + 7 = 10

      Sepal.Length Sepal.Width
[1,]          5.1          3.5
[2,]          4.9          3.0
[3,]          4.7          3.2
[4,]          4.6          3.1
[5,]          5.0          3.6
[6,]          5.4          3.9

Ln 1, Col 1    100%    Windows (CRLF)    UTF-8
```



일반 파일 입출력 - 3

• 파일 읽기 - 1

- readLines("filename")를 통해 파일 읽기
- 파일을 그대로 읽어 들여 화면에 출력. 변수에 직접 입력도 가능

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지들 윈도우즈 도움말

> readLines( "data.txt" )
[1] "[1] \"Result data file\""
[2] "3 + 7 = 10 "
[3] "          Sepal.Length Sepal.Width"
[4] " [1,]           5.1           3.5"
[5] " [2,]           4.9           3.0"
[6] " [3,]           4.7           3.2"
[7] " [4,]           4.6           3.1"
[8] " [5,]           5.0           3.6"
[9] " [6,]           5.4           3.9"
```

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지들 윈도우즈 도움말

> data.txt <- readLines( "data.txt" )
> data.txt
[1] "[1] \"Result data file\""
[2] "3 + 7 = 10 "
[3] "          Sepal.Length Sepal.Width"
[4] " [1,]           5.1           3.5"
[5] " [2,]           4.9           3.0"
[6] " [3,]           4.7           3.2"
[7] " [4,]           4.6           3.1"
[8] " [5,]           5.0           3.6"
[9] " [6,]           5.4           3.9"
```



일반 파일 입출력 - 4

• 파일 읽기 - 2

- 파일 객체를 이용하여 다이얼로그로 파일을 선택하기
- "file.choose()"를 통해 파일을 선택하여 경로를 획득할 수 있음
- sink(), readLines()에서 동일하게 파일명을 입력하여 처리 가능

The screenshot displays the RGui (64-bit) interface with two windows. The top window shows the R console with the command `file.choose()` executed, returning the file path `"C:\\Users\\unangel\\Desktop\\test.txt"`. The bottom window shows the R console with the command `sink(file.choose())` executed, followed by `df` and another `sink()` command. A file selection dialog box is open, showing the contents of the `tmp` folder on the D: drive. The file `data.txt` is selected, and the file type is set to `All files (*.*)`. The file name `data.txt` is entered in the text field at the bottom of the dialog.

RGui (64-bit) - [R Console]

파일 편집 보기 기타 패키지를 윈도우즈 도움말

```
> file.choose()  
[1] "C:\\Users\\unangel\\Desktop\\test.txt"  
> |
```

RGui (64-bit) - [R Console]

파일 편집 보기 기타 패키지를 윈도우즈 도움말

```
> sink( file.choose() )  
> df  
> sink( )  
> |
```

파일선택

내 PC > 로컬 디스크 (D:) > tmp

tmp 검색

구성	새 폴더	이름	수정된 날짜	유형	크기
카카오톡 받은 파일					
OneDrive - Personal		data.txt	2021-11-08 오후 4:31	텍스트 문서	10KB
내 PC					

N: data.txt

All files (*.*)

열기(O) 취소



일반 파일 입출력 - 5

• 읽은 변수 확인

- 단순 문자 데이터로 구성되며, 행과 열로 획득되지 않음
- vector로 한 줄이 하나의 인덱스로 구성

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지를 윈도우즈 도움말

> data.txt <- readLines( "data.txt" )
> data.txt
[1] "[1] \"Result data file\""
[2] "3 + 7 = 10 "
[3] "      Sepal.Length Sepal.Width"
[4] " [1,]           5.1           3.5"
[5] " [2,]           4.9           3.0"
[6] " [3,]           4.7           3.2"
[7] " [4,]           4.6           3.1"
```

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지를 윈도우즈 도움말

> str( data.txt )
chr [1:304] "[1] \"Result data file\"" ...
> class( data.txt )
[1] "character"
> nrow( data.txt )
NULL
> ncol( data.txt )
NULL
> head( data.txt )
[1] "[1] \"Result data file\""
[2] "3 + 7 = 10 "
[3] "      Sepal.Length Sepal.Width"
[4] " [1,]           5.1           3.5"
[5] " [2,]           4.9           3.0"
[6] " [3,]           4.7           3.2"
> |
```



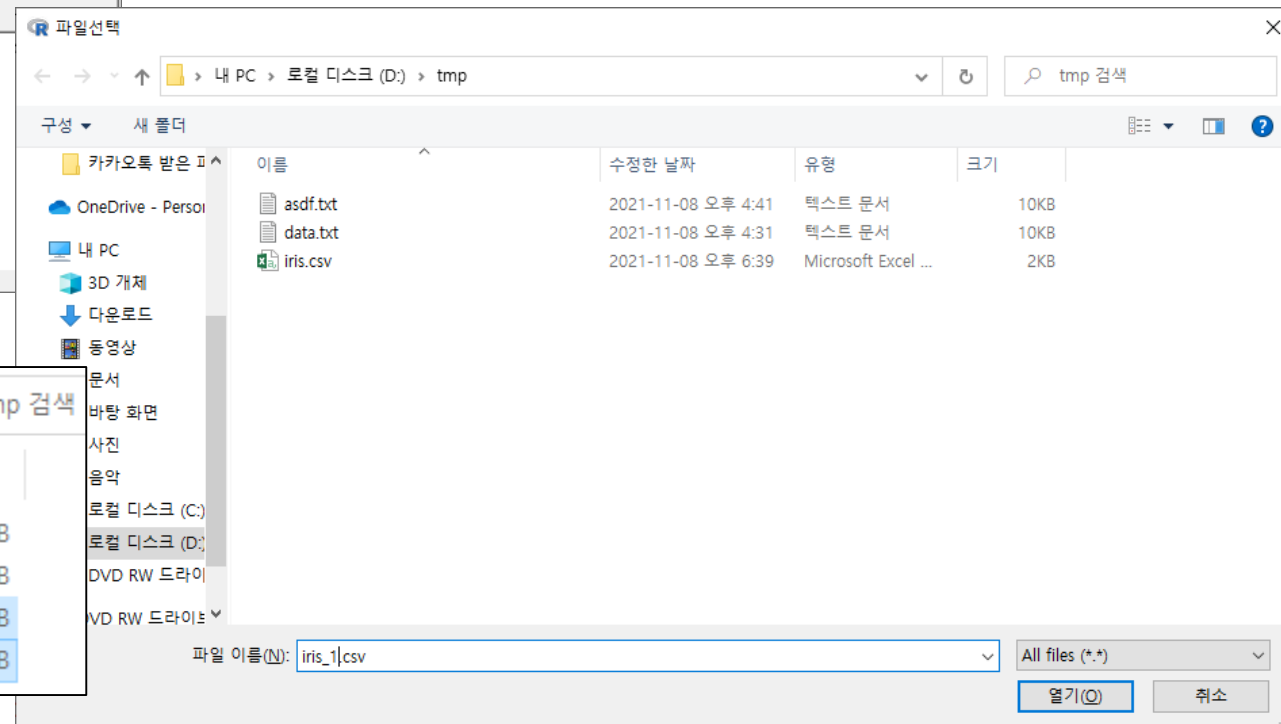
CSV 파일 입출력 - 1

- CSV(Comma Separated Value) 파일 사용
 - R에서 가장 간단하게 파일을 읽거나 쓰며, 엑셀과 호환되는 파일 형태
 - "file.choose()"로 파일 이름을 설정하고 기록도 가능

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지들 윈도우즈 도움말

> iris.sub <- subset( iris, Species == 'setosa' )
> write.csv( iris.sub, "iris.csv", row.names = F )
> write.csv( iris.sub, file.choose( ), row.names = F )
> |
```

이름	수정한 날짜	유형	크기
asdf.txt	2021-11-08 오후 4:41	텍스트 문서	10KB
data.txt	2021-11-08 오후 4:31	텍스트 문서	10KB
iris.csv	2021-11-08 오후 6:39	Microsoft Excel ...	2KB
iris_1.csv	2021-11-08 오후 6:40	Microsoft Excel ...	2KB

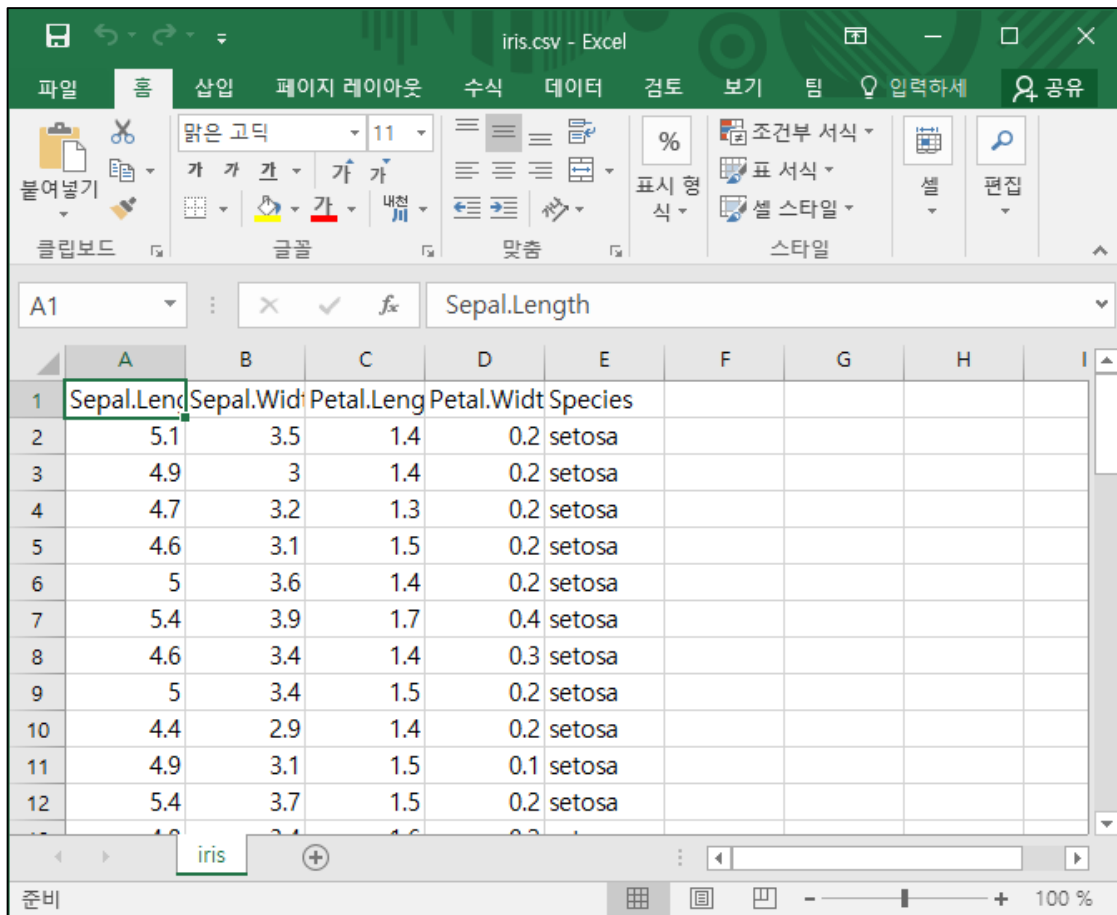




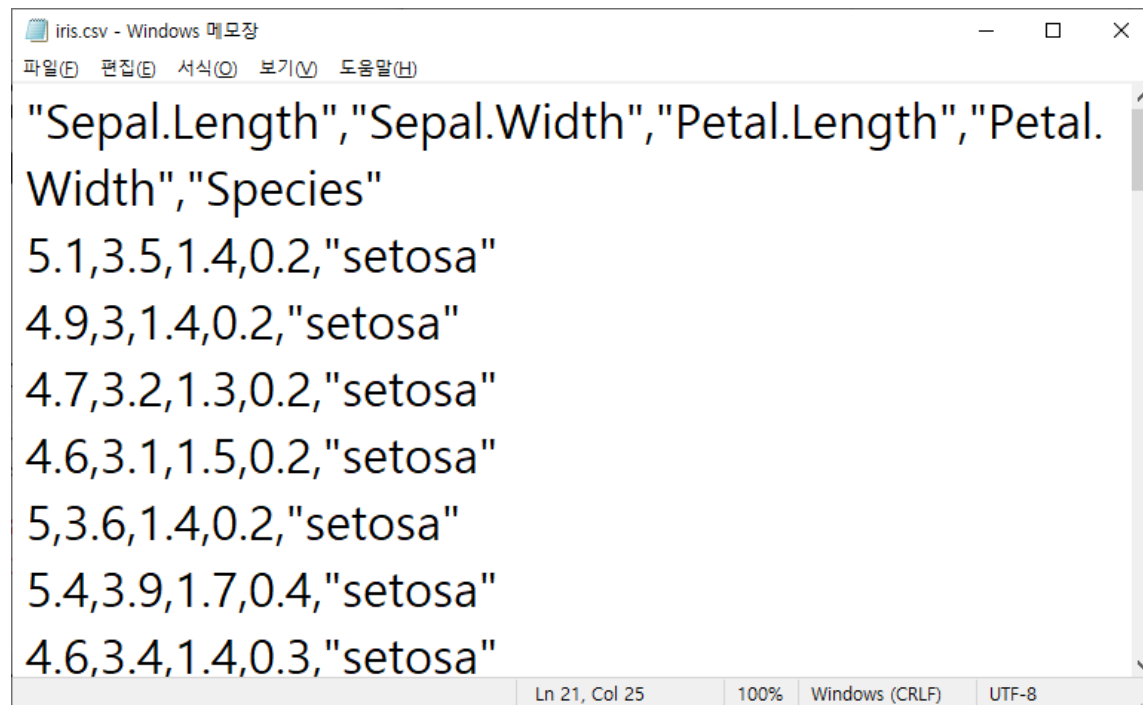
CSV 파일 입출력 - 2

• 기록된 파일 내용

- 엑셀로도 열기 가능하며, 메모장에서는 “,”로 분리되어 기록된 내용 확인 가능



	A	B	C	D	E	F	G	H	I
1	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species				
2	5.1	3.5	1.4	0.2	setosa				
3	4.9	3	1.4	0.2	setosa				
4	4.7	3.2	1.3	0.2	setosa				
5	4.6	3.1	1.5	0.2	setosa				
6	5	3.6	1.4	0.2	setosa				
7	5.4	3.9	1.7	0.4	setosa				
8	4.6	3.4	1.4	0.3	setosa				
9	5	3.4	1.5	0.2	setosa				
10	4.4	2.9	1.4	0.2	setosa				
11	4.9	3.1	1.5	0.1	setosa				
12	5.4	3.7	1.5	0.2	setosa				



```
"Sepal.Length","Sepal.Width","Petal.Length","Petal.Width","Species"
5.1,3.5,1.4,0.2,"setosa"
4.9,3,1.4,0.2,"setosa"
4.7,3.2,1.3,0.2,"setosa"
4.6,3.1,1.5,0.2,"setosa"
5,3.6,1.4,0.2,"setosa"
5.4,3.9,1.7,0.4,"setosa"
4.6,3.4,1.4,0.3,"setosa"
```




CSV 파일 입출력 - 3

• CSV 파일을 data frame으로 읽기

- read.table()로 data frame 형태로 읽기 가능
- "read.table("filename", header = T/F, sep = ' \"/>' data-bbox="16 357 584 843" data-label="Code-Block">

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지를 윈도우즈 도움말

> iris.sub <- read.table( "iris.csv", header = T, sep = ',' )
> iris.sub
  Sepal.Length Sepal.Width Petal.Length Petal.Width
1          5.1          3.5          1.4          0.2
2          4.9          3.0          1.4          0.2
3          4.7          3.2          1.3          0.2
4          4.6          3.1          1.5          0.2
5          5.0          3.6          1.4          0.2
6          5.4          3.9          1.7          0.4
7          4.6          3.4          1.4          0.3
8          5.0          3.4          1.5          0.2
9          4.4          2.9          1.4          0.2
```

```
RGui (64-bit) - [R Console]
파일 편집 보기 기타 패키지를 윈도우즈 도움말

> class( iris.sub )
[1] "data.frame"
> nrow( iris.sub )
[1] 50
> ncol( iris.sub )
[1] 5
> head( iris.sub )
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1          3.5          1.4          0.2  setosa
2          4.9          3.0          1.4          0.2  setosa
3          4.7          3.2          1.3          0.2  setosa
4          4.6          3.1          1.5          0.2  setosa
5          5.0          3.6          1.4          0.2  setosa
6          5.4          3.9          1.7          0.4  setosa
> |
```