Part 1. R 프로그래밍 (데이터 분석 전문가 양성과정)

08

데이터 전처리 (2)

선택, 집계, 분리, 결합, 정렬

경북대학교 배준현 교수

(joonion@knu.ac.kr)

08. 데이터 전처리 (2): 선택, 집계, 분리, 결합, 정렬

- **데이터 다루기:** *Wrangling Data*
  - **데이터의 형태를 통계분석에 적합한 형태로 변환하기 위한 R 함수들:**
    - **선택**: subset()
    - **집계**: aggregate()
    - **분리**: split()
    - **결합**: rbind(), cbind(), merge()
    - **정렬**: sort(), order()

08. 데이터 전처리 (2): 선택, 집계, 분리, 결합, 정렬

- subset() 함수: 인덱싱이나 필터링보다 간편하게 필요한 데이터를 추출

```
> subset(iris, subset = Species == "setosa")
   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
......(이하 생략)

> subset(iris, select = c(1, 2, 5))
    Sepal.Length Sepal.Width    Species
1            5.1         3.5     setosa
2            4.9         3.0     setosa
......(이하 생략)

> subset(iris, subset = Sepal.Length > 7.5)
> subset(iris, subset = Sepal.Length > 7.5 & Sepal.Width > 3.0)
> subset(iris,
         subset = Sepal.Length > 7.5 & Sepal.Width > 3.0,
         select = c(1, 2, 5))
```

08. 데이터 전처리 (2): 선택, 집계, 분리, 결합, 정렬

- split() 함수: 데이터 프레임을 범주형 변수를 기준으로 여러 개로 분할

```
> split(iris, f = iris$Species)
$setosa
   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

> df <- split(iris, f = iris$Species)
> names(df)
[1] "setosa"     "versicolor" "virginica"

> summary(df)
           Length           Class Mode
setosa     5      data.frame list
versicolor 5      data.frame list
virginica  5      data.frame list

> df$setosa
> df$versicolor
```

08. 데이터 전처리 (2): 선택, 집계, 분리, 결합, 정렬

- aggregate() 함수: 범주별로 통계량을 확인하고 싶을 때 주로 활용

```
> df <- subset(iris, select = c(1, 2))
> aggregate(df, by = list(Species=iris$Species), FUN = length)
     Species Sepal.Length Sepal.Width
1     setosa           50          50
2 versicolor           50          50
3  virginica           50          50

> aggregate(df, by = list(Species=iris$Species), FUN = mean)
     Species Sepal.Length Sepal.Width
1     setosa        5.006       3.428
2 versicolor        5.936       2.770
3  virginica        6.588       2.974

> aggregate(df, by = list(Species=iris$Species), FUN = sd)
     Species Sepal.Length Sepal.Width
1     setosa    0.3524897   0.3790644
2 versicolor    0.5161711   0.3137983
3  virginica    0.6358796   0.3224966
```

- aggregate() 함수: 여러 개의 범주를 사용해서 분할할 수도 있음

```
> str(mtcars)
> df <- subset(mtcars, select = c("mpg", "cyl", "gear"))
> aggregate(df,
+           by = list(cyl=mtcars$cyl, gear=mtcars$gear),
+           FUN = mean)
  cyl gear    mpg       hp       wt
1   4    3 21.500  97.0000 2.465000
2   6    3 19.750 107.5000 3.337500
3   8    3 15.050 194.1667 4.104083
4   4    4 26.925  76.0000 2.378125
5   6    4 19.750 116.5000 3.093750
6   4    5 28.200 102.0000 1.826500
7   6    5 19.700 175.0000 2.770000
8   8    5 15.400 299.5000 3.370000
```

- rbind() 함수: 행(row)을 기준으로 여러 개의 데이터 프레임을 결합

```
> df.split <- split(iris, f = iris$Species)
> df.rbind <- rbind(df.split$setosa, df.split$virginica)
> str(df.rbind)
'data.frame':  100 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 1 1 1 1 1 1 1 1 ...
```

08. 데이터 전처리 (2): 선택, 집계, 분리, 결합, 정렬

- cbind() 함수: 열(column)을 기준으로 여러 개의 데이터 프레임을 결합

```
> df.sepal <- subset(iris, select = c(1, 2))
> str(df.sepal)
'data.frame':  150 obs. of  2 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 ...
> df.petal <- subset(iris, select = c(3, 4))
> str(df.petal)
'data.frame':  150 obs. of  2 variables:
 $ 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 ...
> df.cbind <- cbind(df.sepal, df.petal)
> str(df.cbind)
'data.frame':  150 obs. of  4 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 ...
```

- merge() 함수: 특정 변수의 값이 같은 행을 기준으로 여러 개의 데이터 프레임을 병합

```
> x <- data.frame(name = c('A', 'B', 'C'), kor = c(60, 70, 80))
> y <- data.frame(name = c('A', 'B', 'D'), eng = c(65, 75, 85))
> cbind(x, y)
> merge(x, y)
  name kor eng
1    A  60  65
2    B  70  75

> merge(x, y, all = T)
  name kor eng
1    A  60  65
2    B  70  75
3    C  80  NA
4    D  NA  85

> merge(x, y, all.x = T)
> merge(x, y, all.y = T)
```

- merge() 함수: 해당 컬럼이 모두 존재하지 않는 경우에는 NA 값으로 병합

```
> x <- data.frame(name = c('A', 'B', 'C'), kor = c(60, 70, 80))
> y <- data.frame(NAME = c('A', 'B', 'D'), eng = c(65, 75, 85))
> merge(x, y)
  name kor NAME eng
1    A  60    A  65
......(이하 생략)

> merge(x, y, by.x = "name", by.y = "NAME")
  name kor eng
1    A  60  65
2    B  70  75

> merge(x, y, by.x = "name", by.y = "NAME", all = T)
  name kor eng
1    A  60  65
2    B  70  75
3    C  80  NA
4    D  NA  85
```

08. 데이터 전처리 (2): 선택, 집계, 분리, 결합, 정렬

- sort()와 order() 함수: 벡터의 값을 오름차순/내림차순으로 정렬하거나 정렬 위치를 반환

```
> sort(mtcars$mpg)
 [1] 10.4 10.4 13.3 14.3 14.7 15.0 15.2 15.2 15.5 15.8 16.4 17.3 17.8 18.1 18.7 19.2
[17] 19.2 19.7 21.0 21.0 21.4 21.4 21.5 22.8 22.8 24.4 26.0 27.3 30.4 30.4 32.4 33.9

> sort(mtcars$mpg, decreasing = T)
 [1] 33.9 32.4 30.4 30.4 27.3 26.0 24.4 22.8 22.8 21.5 21.4 21.4 21.0 21.0 19.7 19.2
[17] 19.2 18.7 18.1 17.8 17.3 16.4 15.8 15.5 15.2 15.2 15.0 14.7 14.3 13.3 10.4 10.4

> order(mtcars$mpg)
 [1] 15 16 24  7 17 31 14 23 22 29 12 13 11  6  5 10 25 30  1  2  4 32 21  3  9  8 27
[28] 26 19 28 18 20

> order(mtcars$mpg, decreasing = T)
[1] 20 18 19 28 26 27  8  3  9 21  4 32  1  2 30 10 25  5  6 11 13 12 29 22 14 23 31
[28] 17  7 24 15 16
```

08. 데이터 전처리 (2): 선택, 집계, 분리, 결합, 정렬

- order() 함수: 정렬 후의 위치를 알려주므로, 데이터 프레임을 변수의 순서대로 정렬할 때 유용

```
> ord <- order(mtcars$mpg, decreasing = T)
> mtcars[ord, 1:6]
                     mpg cyl  disp  hp drat    wt
Toyota Corolla      33.9   4  71.1  65 4.22 1.835
Fiat 128            32.4   4  78.7  66 4.08 2.200
Honda Civic         30.4   4  75.7  52 4.93 1.615
......(이하 생략)

> mtcars[ord[1:10], 1:6]
> n <- length(ord)
> mtcars[ord[(n-10):n], 1:6]
```

08. 데이터 전처리 (2): 선택, 집계, 분리, 결합, 정렬

- order() 함수: 여러 개의 변수로 정렬할 때에도 유용하게 사용할 수 있음

```
> ord <- order(iris$Petal.Length, iris$Sepal.Length)
> head(iris[ord, c(3, 1)])
   Petal.Length Sepal.Length
23          1.0          4.6
14          1.1          4.3
15          1.2          5.8

> ord <- order(iris$Petal.Length, -iris$Sepal.Length)
> head(iris[ord, c(3, 1)])
   Petal.Length Sepal.Length
23          1.0          4.6
14          1.1          4.3
15          1.2          5.8
```

- **연습문제 8.1:**
  - state.x77 데이터셋에 대하여, 다음 R 코드를 작성하시오.
    - Population을 기준으로 오름차순으로 정렬하시오.
    - Income을 기준으로 내림차순으로 정렬하시오.
    - Illiteracy를 기준으로 오름차순으로 정렬하되,
      - 문맹률이 같은 주에 대해서는 Population의 내림차순으로 정렬하시오.

- **연습문제 8.2:**
  - mtcars 데이터셋에 대하여, 다음 R 코드를 작성하시오.
    - mtcars 데이터셋을 gear의 개수에 따라 그룹을 나누시오.
      - split() 함수를 이용하여 df.split에 저장
    - mtcars 데이터셋에서 gear의 개수가 3인 그룹과 4인 그룹을 합치시오.
      - rbind() 함수를 이용하여 df.34에 저장

- **연습문제 8.3:**
  - airquality 데이터셋에 대하여, 다음 R 코드를 작성하시오.
    - airquality에서 1, 2, 3, 4번 column을 추출하여 df에 저장: subset() 함수
    - 위에서 추출한 변수에 대해 월별(Month)로 평균을 구하시오.
      - aggregate() 함수로 mean() 함수를 범주를 Month로 하여 구할 수 있음.
      - NA 값에 대해서는 na.rm = T로 매개변수값을 지정
    - 위에서 추출한 변수에 대해 일별(Day)로 표준편차를 구하시오.
      - aggregate() 함수로 sd() 함수를 적용하여 df.day 에 저장

Any Questions?