1. **reshape2패키지  
   - 데이터 구성을 변형하는 함수이다.**

* **reshape2가 제공하는 변환은 크게 melt, cast 이며 이 두가지 변환을 사용해 데이터의 모양을 바꾸거나, 데이터를 요약할 수 있다.**

**설치방법**

> install.packages("plyr")

>library(plyr)

1. melt()

가로로 긴 데이터 모양을 세로로 전환하는 함수

인자로 데이터를 구분하는 식별자(id), 측정 대상 변수, 측정치를 받아 데이터를 간략하게 표현

**함수형태**

|  |
| --- |
| melt(데이터 세트, id.var = "기준 열", measure.vars = "변환 열")  함수 안에 결측지 제외를 하려면 na.rm = TRUE  value.name = "이름" 을 넣으면 value 열의 이름 지정 가능 |

>실습

> chart

이름 국어 영어 수학

1 주현 90 100 80

2 슬기 95 95 90

3 예리 85 85 100

> library(reshape2)

> melt1 <- melt(chart, id.vars = "이름", measure = c("국어", "영어"), value.name = "성적")

> melt1

이름 variable 성적

1 주현 국어 90

2 슬기 국어 95

3 예리 국어 85

4 주현 영어 100

5 슬기 영어 95

6 예리 영어 85

**Cast()**

- 세로로 긴 데이터 모양을 가로로 전환하는 함수

- 데이터를 요약하는 기능

2) dcast()

데이터 프레임 형태를 변환하는 함수

인자로 데이터, 데이터를 그룹 지을 변수명, 데이터 처리에 사용할 함수를 받는다.

**함수형태**

|  |
| --- |
| Dcast(데이터 세트, id.var = "기준 열", measure.vars = "변환 열") |

>실습

> melt1

이름 variable 성적

1 주현 국어 90

2 슬기 국어 95

3 예리 국어 85

4 주현 영어 100

5 슬기 영어 95

6 예리 영어 85

> dcast(melt1, 이름 ~ variable)

Using 성적 as value column: use value.var to override.

이름 국어 영어

1 슬기 95 95

2 예리 85 85

3 주현 90 100

3) acast()

벡터, 행렬, 배열 형태를 변환

**함수형태**

|  |
| --- |
| acast(데이터 세트, id.var = "기준 열"~ measure.vars = "변환 열"~분리기준열) |

>실습

> melt1

이름 variable 성적

1 주현 국어 90

2 슬기 국어 95

3 예리 국어 85

4 주현 영어 100

5 슬기 영어 95

6 예리 영어 85

> acast(melt1, 이름 ~ 성적 ~ variable)

Using 성적 as value column: use value.var to override.

, , 국어

85 90 95 100

슬기 NA NA 95 NA

예리 85 NA NA NA

주현 NA 90 NA NA

, , 영어

85 90 95 100

슬기 NA NA 95 NA

예리 85 NA NA NA

주현 NA NA NA 100

> acast(melt1, variable ~ 성적 ~ 이름)

Using 성적 as value column: use value.var to override.

, , 슬기

85 90 95 100

국어 NA NA 95 NA

영어 NA NA 95 NA

, , 예리

85 90 95 100

국어 85 NA NA NA

영어 85 NA NA NA

, , 주현

85 90 95 100

국어 NA 90 NA NA

영어 NA NA NA 100

1. **data.table패키지  
   - 데이터 프레임을 대신하여 사용할 수 있는 패키지**

* **데이터 선택, 업데이트, 병합 기능의 통합처리 및 대용량처리 속도가 개선된 패키지**

**설치방법**

> install.packages("data.table")

>library(data.table)

**함수형태**

|  |
| --- |
| acast(데이터 세트, id.var = "기준 열"~ measure.vars = "변환 열"~분리기준열) |

**실습**

1. 데이터 테이블 생성

> x <- data.table(x=c(1,2,3), y=c("a","b","c"))

> x

x y

1: 1 a

2: 2 b

3: 3 c

> head(iris,3); str(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

'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 1 1 1 1 1 1 1 1 ...

> d <- as.data.table(iris)

> d

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

---

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

> tables()

NAME NROW NCOL MB

1: d 150 5 0

2: x 3 2 0

COLS KEY

1: Sepal.Length,Sepal.Width,Petal.Length,Petal.Width,Species

2: x,y

Total: 0MB

> i\_table <- as.data.table(iris)

> head(i\_table,3)

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

1. 데이터접근과 그룹 연산

> i\_table[1.]

Sepal.Length Sepal.Width Petal.Length Petal.Width Species

1: 5.1 3.5 1.4 0.2 setosa

> i\_table[i\_table$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

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

7: 4.6 3.4 1.4 0.3 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 0.1 setosa

11: 5.4 3.7 1.5 0.2 setosa

12: 4.8 3.4 1.6 0.2 setosa

13: 4.8 3.0 1.4 0.1 setosa

14: 4.3 3.0 1.1 0.1 setosa

15: 5.8 4.0 1.2 0.2 setosa

16: 5.7 4.4 1.5 0.4 setosa

17: 5.4 3.9 1.3 0.4 setosa

18: 5.1 3.5 1.4 0.3 setosa

19: 5.7 3.8 1.7 0.3 setosa

20: 5.1 3.8 1.5 0.3 setosa

21: 5.4 3.4 1.7 0.2 setosa

22: 5.1 3.7 1.5 0.4 setosa

23: 4.6 3.6 1.0 0.2 setosa

24: 5.1 3.3 1.7 0.5 setosa

25: 4.8 3.4 1.9 0.2 setosa

26: 5.0 3.0 1.6 0.2 setosa

27: 5.0 3.4 1.6 0.4 setosa

28: 5.2 3.5 1.5 0.2 setosa

29: 5.2 3.4 1.4 0.2 setosa

30: 4.7 3.2 1.6 0.2 setosa

31: 4.8 3.1 1.6 0.2 setosa

32: 5.4 3.4 1.5 0.4 setosa

33: 5.2 4.1 1.5 0.1 setosa

34: 5.5 4.2 1.4 0.2 setosa

35: 4.9 3.1 1.5 0.2 setosa

36: 5.0 3.2 1.2 0.2 setosa

37: 5.5 3.5 1.3 0.2 setosa

38: 4.9 3.6 1.4 0.1 setosa

39: 4.4 3.0 1.3 0.2 setosa

40: 5.1 3.4 1.5 0.2 setosa

41: 5.0 3.5 1.3 0.3 setosa

42: 4.5 2.3 1.3 0.3 setosa

43: 4.4 3.2 1.3 0.2 setosa

44: 5.0 3.5 1.6 0.6 setosa

45: 5.1 3.8 1.9 0.4 setosa

46: 4.8 3.0 1.4 0.3 setosa

47: 5.1 3.8 1.6 0.2 setosa

48: 4.6 3.2 1.4 0.2 setosa

49: 5.3 3.7 1.5 0.2 setosa

50: 5.0 3.3 1.4 0.2 setosa

> i\_table[1,Sepal.Length]

[1] 5.1

> i\_table[, mean(Sepal.Length)]

[1] 5.843333

> i\_table [ , mean ( Sepal.Length - Sepal.Width ) ]

[1] 2.786

3) 데이터를 그룹 지을 변수 지정 (by)  
> result <- i\_table[,mean(Sepal.Length), by="Species"]

> colnames(result)[2]<-"mean"

> result

Species mean

1: setosa 5.006

2: versicolor 5.936

3: virginica 6.588

**Key를 사용한 탐색**

* 데이터 양이 많고, 데이터 검색 작업의 횟수가 많으면 긴 수행시간이 걸리므로 이를 해결하기 위해 미리 색인을 만들어두었다가 검색 시 색인을 사용하여 검색한다.

> d <- data.frame ( x = runif (260000) , y = rep ( LETTERS , each =10000) )

> head(d,4)

x y

1 0.0727432 A

2 0.1767064 A

3 0.2085195 A

4 0.9171219 A

> system.time(x<-d[d$y=="c",])

사용자 시스템 elapsed

0 .034 0 .000 0 .033

색인 정의 후

> d <- as.data.table(d)

> setkey ( d , y )

> system.time ( x <- d [ J ( " C " ) , ])

사용자 시스템 elapsed

0 .002 0 .000 0 .002

* 수행시간이 단축되었다

**Key를 이용한 연산**

실습

> d[d$y=="C", mean(x)]

[1] 0.5041991

> d[J("C"), mean(x)]

[1] 0.5041991

> system.time(d[d$y=="C", mean(x)])

사용자 시스템 elapsed

0 0 0

> system.time(d[J("C"), mean(x)])

사용자 시스템 elapsed

0 0 0

> d[J("C"), list(mean=mean(x), sd=sd(x))]

mean sd

1: 0.5041991 0.2857317

**key를 사용한 데이터 테이블 병합**

* key는 여러 데이터 테이블의 병합에도 사용할 수 있다
* DT1, DT2의 두개 데이터 테이블이 있을때 DT1[DT2, 표현식]는 DT1에서 DT2에 대응되는 데이터를 찾는 방식으로 데 이터를 병합3

실습

> DT1 <- data.table ( x = runif (260000) , y = rep ( LETTERS , each =10000) )

> DT2 <- data.table ( y = c ( " A " , " B " , " C " ) , z = c ( " a " , " b " , " c " ) )

> setkey ( DT1 , y )

> DT1 [ DT2 , ]

x y z

1: NA A a

2: NA B b

3: NA C c

> setkey ( DT2 , y )

> DT2 [ DT1 , ]

y z x

1: A <NA> 0.260808500

2: A <NA> 0.010805749

3: A <NA> 0.077678092

4: A <NA> 0.004991856