고급 데이터 구조

요인(factor)란 무엇인가?

- 요인은 R의 데이터 중에서 하나이며 벡터의 한 종류
- 벡터와 요인 모두 1차원의 형태로 자료가 되어 있음
  - 벡터 – 범주형 자료로 인식하지 못함
  - 요인 – 범주형 자료로 인식함
- 요인 주로 집단별로 자료를 분석하고자 할 때에 특정 자료를 범주형 자료로 변경해 줌

요인을 생성하는 방법

argument	설명
x	벡터를 지정
levels	그룹으로 지정할 문자형 벡터를 지정하며, levels를 쓰지 않으면 오름차순으로 구분하여 자체적으로 그룹을 지정
labels	levels에 대한 문자형 벡터를 지정
ordered	levels에 대해 특정한 순서를 정하고 싶으면 TRUE를 지정

요인을 생성하는 방법

- factor() 함수의 사용
  - 6명의 성별 데이터를 gender라는 벡터에 저장, gender라는 벡터를 요인(factor)로 변환
    - gender = c(“m”, “f”, “f”, “m”, “f”, “f”)
    - gender => gender라는 벡터는 범주형 자료가 아님
    - gender_factor = factor(gender) => 요인으로 변환된 gender_factor는 범주형 자료로서 남자(m) 집단과 여자(f) 집단으로 구분
    - gender_factor
    - gender

요인을 생성하는 방법

- levels() 함수의 사용
  - 요인이 가지는 집단이 몇 개 이며, 각 집단의 이름이 무엇인지를 알고자 할 경우 사용
    - levels(요인)
    - levels(gender_factor) => gender_factor가 두 개의 집단으로 구성, 각 집단의 이름은 f와 m으로 되어 있음을 알려줌
- labels argument 사용
  - 벡터에 있는 각각의 원소의 값을 다른 문자형 유형으로 변경할 경우
    - gender_factor2 = factor(gender, levels=c("m", "f"), lables=c("남자", "여자"))
      gender_factor2 => m을 남자, f를 여자로 변경, 남자 집단을 여자 집단보다 먼저 인식

- factor() 함수에 ordered=TRUE를 추가
  - 집단으로 할 뿐만 아니라 순서도 의미가 있도록 함
  - 통계에서 말하는 질적 자료이면서 순서형 자료가 됨

    - factor(벡터, ordered=TRUE)
    - gender_factor3 = factor(gender, ordered=TRUE) => 집단이 f와 m으로 구성, f < m의 순서가 의미를 갖도록 설정됨

행렬이란 무엇인가?

- 행렬(matrix)은 데이터의 형태가 2차원으로 행(row)과 열(column)로 구성
- 벡터의 확장 개념
- 벡터와 동일하게 하나의 데이터 유형만 가질 수 있음
- 행렬은 수학이나 통계에서 주로 사용

행렬을 생성하는 방법

- 행렬을 생성하는 함수
  - rbind()
  - cbind()
  - matrix()
  - rbind() : 벡터를 행 방향으로 합치는 방법
  - cbind() : 벡터를 열 방향으로 합치는 방법

행렬을 생성하는 방법

- 행렬을 생성하는 함수
  - rbind(벡터1, 벡터2, ...)
    - v1 벡터와 v2 벡터를 행으로 합쳐서 하나의 행렬을 만듦
    - 벡터의 개수가 행의 개수가 됨
    - 벡터가 가지는 원소의 개수가 열의 개수가 됨
  - cbind(벡터1, 벡터2, ...)
    - 벡터의 개수가 열의 개수가 됨
    - 벡터가 가지는 원소의 개수가 행의 개수가 됨
    - 행렬에서는 재사용 규칙이 적용
  - v1 = 1:3
  - v2 = 4:6
  - m1 = rbind(v1, v2)
  - m2 = cbind(v1, v2)
  - rbind

행렬을 생성하는 방법

- matrix() 함수

argument	설명
x	벡터를 지정
nrow	행의 개수를 지정
ncol	열의 개수를 지정
byrow	행렬에 값이 입력될 때 기본적으로 열 방향으로 먼저 입력되며, 값이 입력되는 방향을 행 방향으로 수정하고 싶으면 TRUE를 지정

행렬을 생성하는 방법

- 행렬을 생성하는 함수
  - m3 = matrix(1:4, nrow=2, ncol=2)
    - 2행 2열인 행렬이며, 값이 열 방향 우선으로 입력
  - m4 = matrix(1:4, nrow=2, ncol=2, byrow=TRUE)
    - m3과 동일하게 2행 2열이지만 값이 행 방향 우선으로 입력되어 다른 행렬이 됨

배열이란 무엇인가?

- 배열(array)은 데이터의 형태가 3차원 이상으로 구성될 수 있음
- 행렬의 확장 개념
- 배열은 차원을 어떻게 지정하느냐에 따라 1차원, 2차원, 3차원, 4차원 등으로 구성할 수 있음
- 벡터와 행렬처럼 데이터의 유형은 하나만 가질 수 있음

배열을 생성하는 방법

- array() 함수

argument	설명
x	벡터를 지정
dim	원하는 차원을 지정하며, 지정하는 숫자의 개수에 따라 차원이 결정됨

배열을 생성하는 방법

- 배열을 생성하는 함수
  - a1 = array(1:10, dim=10)
    - a1 배열은 dim에 하나의 수치를 지정 => 1차원 형태의 데이터를 가짐. 즉, 벡터가 됨
  - a2 = array(1:10, dim=c(2, 5))
    - a2 배열은 dim에 두 개의 수치를 지정 => 2차원 형태가 되며 2행 5열인 행렬이 됨
  - a3 = array(1:10, dim=c(3, 3, 4))
    - a3 배열은 dim에 세 개의 수치를 지정 => 3차원 형태가 됨

리스트란 무엇인가?

- 리스트(list)는 R의 데이터 형태인 벡터(vector), 요인(factor), 행렬(matrix), 배열(array), 데이터 프레임(data frame)과 리스트 자체까지 원소로 가질 수 있음
- 리스트 구조로 데이터를 저장해서 분석
- 많은 경우에는 데이터를 분석한결과의 형태가 리스트인 경우가 많음
  - 초보 단계에서는 데이터 분석의 결과를 저장하는 데이터 형태로 리스트를 기억하는 것이 더 좋음

리스트를 생성하는 방법

- list() 함수 사용방법
  - 하나의 벡터와 하나의 행렬을 가지는 리스트를 생성
  - list(벡터, 요인, 행렬, 배열, 데이터 프레임, 리스트)
  - v1 = 1:5
  - m1 = matrix(1:6, nrow=2, ncol=3)
  - d1 = list(v1, m1)
    - d1 이라는 리스트는 5개의 원소를 가지는 수치형 벡터와 2행 3열로 행렬을 원소로 가짐

리스트를 생성하는 방법

- 리스트의 원소 중에서 일부를 추출하는 방법
  - [] 사용법
    - 리스트명[index]
      - 리스트명[index]의 경우 지정된 index의 위치에 있는 원소를 가져오며 최종적인 형태는 리스트가 됨
  - [[]] 사용법
    - 리스트명[[index]]
      - 리스트명[[index]]의 경우 지정된 index의 위치에 있는 원소를 가져오며, 최종적인 결과는 index 위치에 있는 원소의 데이터 형태가 됨

리스트를 생성하는 방법

- 리스트의 원소 중에서 일부를 추출하는 방법
  - d1[1]
    - d1[1]은 리스트에 있는 첫 번째 원소를 가져오며, => 리스트가 됨
  - d1[[1]]
    - d1[[1]은 리스트에 있는 첫 번째 원소를 가져오지만, 첫 번째 원소가 벡터이기 때문에 => 벡터가 됨

데이터 프레임은 무엇인가?

- 데이터 프레임(data frame)은 R의 대표적인 데이터 형태
- 2차원 구조로 행렬처럼 행과 열로 구성
- 행렬은 하나의 데이터 유형만 가질 수 있지만 데이터 프레임은 여러 가지 데이터 유형을 가질 수 있음
- 일반적으로 R에서 데이터를 분석할 때에는 데이터 프레임 형태로 되어 있어야 하지만 데이터 프레임에서 하나의 열은 벡터처럼 하나의 데이터 유형만 가짐

데이터 프레임을 생성하는 방법

- 텍스트, CSV, 엑셀, DB 형태로 되어 있는 외부 데이터를 R에서 읽어오면 그 데이터 형태는 데이터 프레임이 됨
  - data.frame() 함수를 이용하여 데이터 프레임을 생성하는 것을 학습하고자 함

argument	설명
...	**벡터나 행렬을 지정**
stringsAsFactors	**데이터의 유형이 문자형인 경우는 데이터 프레임을 생성할 때 기본적으로 요인(factor)으로 변경되며, 이것을 원하지 않을 경우 FALSE를 지정하면 문자형 그대로 유지됨**

데이터 프레임을 생성하는 방법

- data.frame() 함수의 사용
  - 5명에 대한 나이, 성별, 키를 조사하여 age, gender, height라는 벡터에 저장하고, 이 벡터들을 data.frame() 함수를 이용하여 하나의 데이터 프레임을 만드는 과정
    - id 벡터는 5명을 식별하기 고유한 값으로 사용

  - id = 1:5 age = c(29, 32, 47, 35, 23)
  - gender = c("f", "m", "m", "f", "f")
  - height = c(163, 177, 172, 157, 169)
  - DF1 = data.frame(id, age, gender, height)
  - DF2 = data.frame(id, age, gender, height, stringsAsFactors=FALSE)
    - DF1, DF2
      - 모두 데이터 프레임 임
      - 5행 4열의 2차원 구조
      - 수치형과 문자형의 두 가지 데이터 유형을 모두 가지고 있음

데이터 프레임을 생성하는 방법

- data.frame() 함수의 사용
  - 5명에 대한 나이, 성별, 키를 조사하여 age, gender, height라는 벡터에 저장하고, 이 벡터들을 data.frame() 함수를 이용하여 하나의 데이터 프레임을 만드는 과정
    - id 벡터는 5명을 식별하기 고유한 값으로 사용
  - id = 1:5
  - age = c(29, 32, 47, 35, 23)
  - gender = c("f", "m", "m", "f", "f")
  - height = c(163, 177, 172, 157, 169)
  - DF1 = data.frame(id, age, gender, height)
    - DF1
      - stringsAsFactors라는 argument를 지정하지 않음
  - DF2 = data.frame(id, age, gender, height, stringsAsFactors=FALSE)

데이터 프레임을 생성하는 방법

- data.frame() 함수의 사용
  - 5명에 대한 나이, 성별, 키를 조사하여 age, gender, height라는 벡터에 저장하고, 이 벡터들을 data.frame() 함수를 이용하여 하나의 데이터 프레임을 만드는 과정
    - id 벡터는 5명을 식별하기 고유한 값으로 사용
  - id = 1:5
  - age = c(29, 32, 47, 35, 23)
  - gender = c("f", "m", "m", "f", "f")
  - height = c(163, 177, 172, 157, 169)
  - DF1 = data.frame(id, age, gender, height)
  - DF2 = data.frame(id, age, gender, height, stringsAsFactors=FALSE)
    - DF2
      - stringsAsFactors라는 argument를 FALSE로 지정

데이터 프레임을 생성하는 방법

- DF1, DF2 두 결과의 차이

```
> str(DF1)
'data.frame':   5 obs. of  4 variables:
 $ id    : int  1 2 3 4 5
 $ age   : num  29 32 47 35 23
 $ gender: Factor w/ 2 levels "f","m": 1 2 2 1 1
 $ height: num  163 177 172 157 169
> str(DF2)
'data.frame':   5 obs. of  4 variables:
 $ id    : int  1 2 3 4 5
 $ age   : num  29 32 47 35 23
 $ gender: chr  "f" "m" "m" "f" ...
 $ height: num  163 177 172 157 169
```

DF1는 stringsAsFactors라는 argument를 FALSE로 지정하지 않았기 때문에 데이터 유형이 문자형인 gender는 자동으로 요인(factor)으로 변경 gender는 두 개의 집단으로 인식된 범주형 자료가 됨

DF2 stringsAsFactors라는 argument를 FALSE로 지정하였기 때문에 데이터 유형이 문자형인 gender를 그대로 유지

데이터 프레임의 속성

- 데이터 프레임의 속성에는 행의 개수, 열의 개수, 행의 이름, 열의 이름, 차원, 차원의 이름이 있음

- 행의 개수
  - 행의 개수를 알려주는 속성
    - nrow(DF1)
    - NROW(DF1)
      - DF1의 행의 개수는 5임을 알려줌

데이터 프레임의 속성

- 열의 개수
  - 열의 개수를 알려주는 속성
    - ncol(DF1)
    - NCOL(DF1)
      - DF1의 열의 개수는 4임을 알려줌

데이터 프레임의 속성

- 행의 이름을 알려주는 속성
- 기본적으로 행의 이름은 문자형이며 1부터 시작하는 일련번호로 되어 있음
- 행의 이름을 변경하고 싶으면 c() 함수나 paste() 함수 등을 이용하여 수정할 수 있음

  - rownames(DF1)
  - 행의 이름을 “R1”, “R2”, “R3”, “R4”, “R5”로 변경하고자 할 경우
    - rownames(DF1) = paste(“R”, 1:5, sep=“”)

데이터 프레임의 속성

- 열의 이름
  - 열의 이름을 알려주는 속성
  - 열의 이름도 문자형으로 되어 있음
  - 열의 이름을 변경하려면 행의 이름과 동일하게 c() 함수나 paste() 함수 등을 이용
    - colnames(DF1)

데이터 프레임의 속성

- 차원
  - 차원(dimension)은 행과 열이 몇 개로 구성되어 있는지를 의미
  - dim() 함수를 사용하면 행의 개수와 열의 개수를 한꺼번에 알려 줌
  - 첫 번째 나오는 숫자가 행의 개수, 두 번째로 나오는 숫자가 열의 개수를 의미함
    - dim(DF1)

데이터 프레임의 속성

- 차원의 이름
  - 차원의 이름은 행의 이름과 열의 이름을 의미함
  - dimnames() 함수를 사용하면 행의 이름과 열의 이름을 한꺼번에 알려줌
    - 차원의 이름은 리스트(list) 형태로 되어 있음
  - 첫 번째로 나오는 이름이 행의 이름, 두 번째로 나오는 이름이 열의 이름을 의미함
    - dimnames(DF1)

데이터의 구조

- str() 함수를 사용하면 지정된 데이터가 어떠한 구조로 형성되어 있는지 알 수가 있음
- 데이터의 구조로 데이터의 형태, 행의 개수, 열의 개수, 변수명, 데이터의 유형 등을 알려
  - DF1이라는 데이터 프레임이 어떤 구조로 되어 있는지 확인하는 방법
  - str(DF1)
  - 데이터를 분석하기 전에 데이터가 어떤 구조로 이루어졌는지 확인하는 것이 중요함!