# 3 주차 강의

## 2. R기초

2022.03.22


Institutor: JS LEE

# WorkFlow





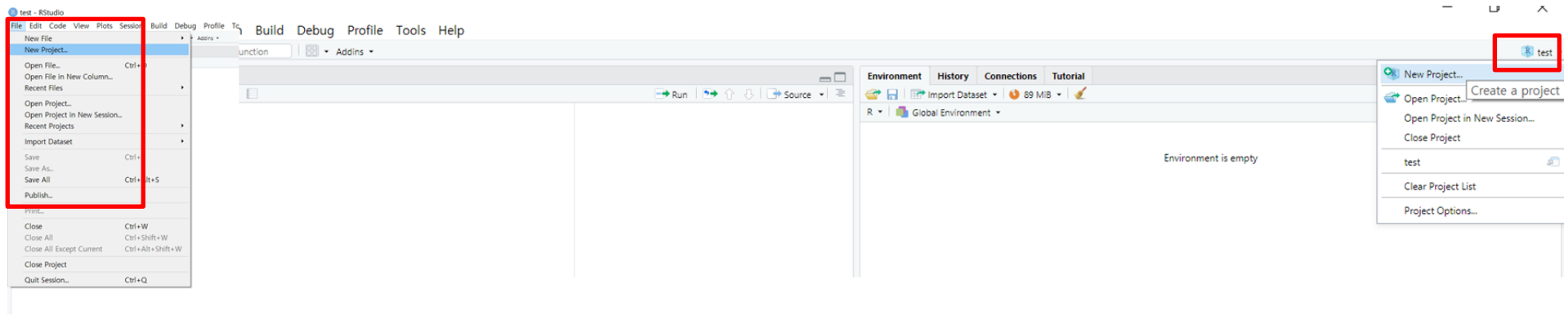
## Rstudio Project

- R 종료 후 다시 분석을 하고자 할 때,
- 데이터 분석을 위해 생성한 소스코드, 입력물, 결과물 등의 관리가 필요할 때
- R을 활용하여 여러 주제별로 분석을 진행하고  주제별로 분석업무의 관리의 필요성이 있을 때
- 내 PC에서 작업한 분석 내용 및 분석결과를 다른 곳에서 사용해야 할 때

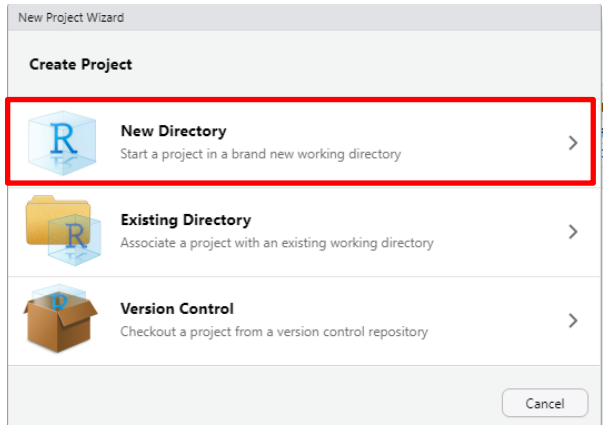
# WorkFlow

## Rstudio Project 생성하기

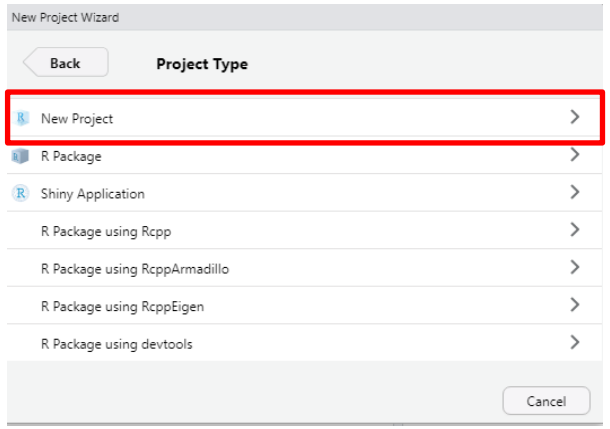
1) 메뉴 [File] – [New Project] 또는 오른쪽 상단의 육각형 모양의 R 아이콘 선택 -> [New Project] 선택




2) [ New Directory] 선택



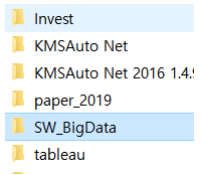

3) [ New Project] 선택




4) 프로젝트 폴더 생성

개인 PC에 새로운 폴더 생성

예> SW_BigData 폴더 생성 ( 폴더명은 자유롭게,
단, 영문 권장)



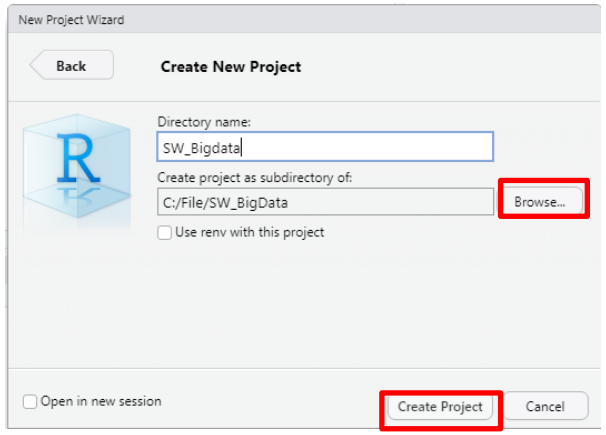

- R에서 프로젝트 설정 시, 새로운 폴더 생성이 가능함

# WorkFlow



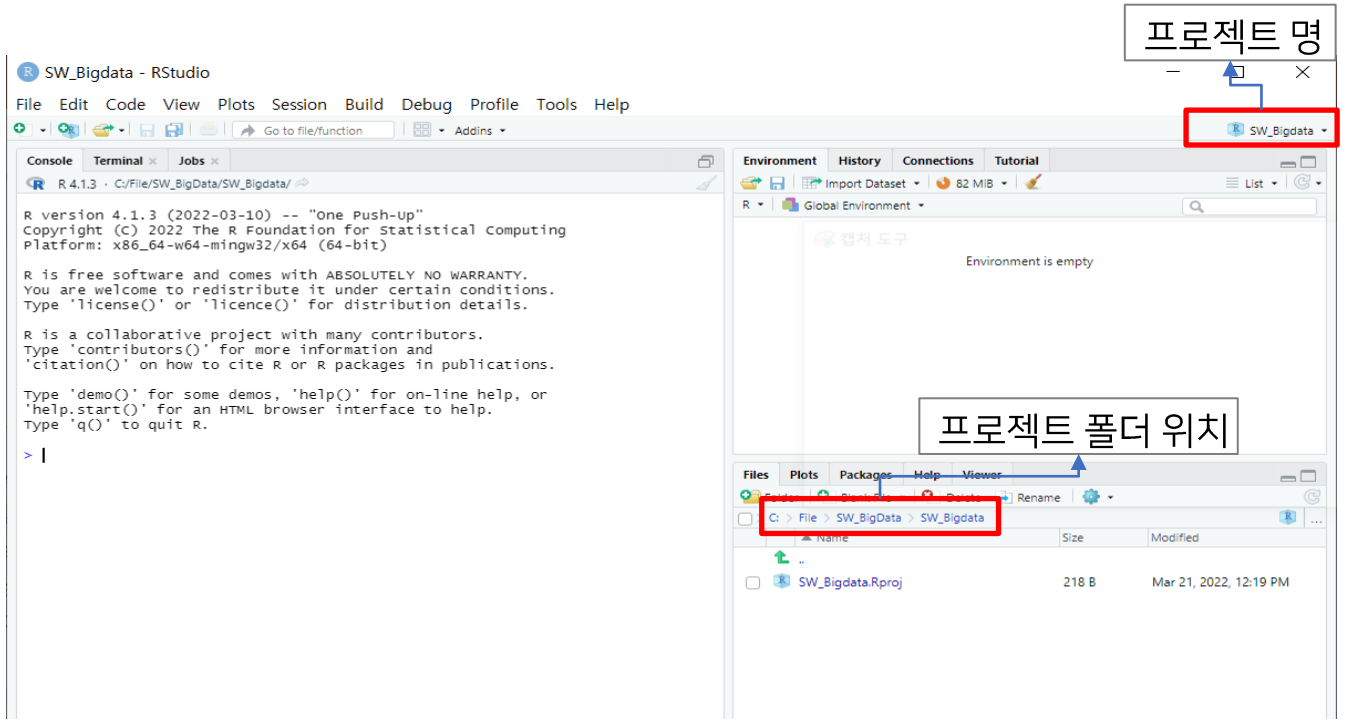
## 5) 이름 지정 및 저장할 폴더 설정

- [Directory name] 프로젝트명 입력
- SW-Bigdata로 함
- [Create project as subdirectory of] 어떤 위치에 프로젝트 폴더 생성할 지 결정
- [Browse] 선택하여 [4]에서 지정한 폴더 선택
- [Create Project] 버튼 선택




## 6) 프로젝트 생성

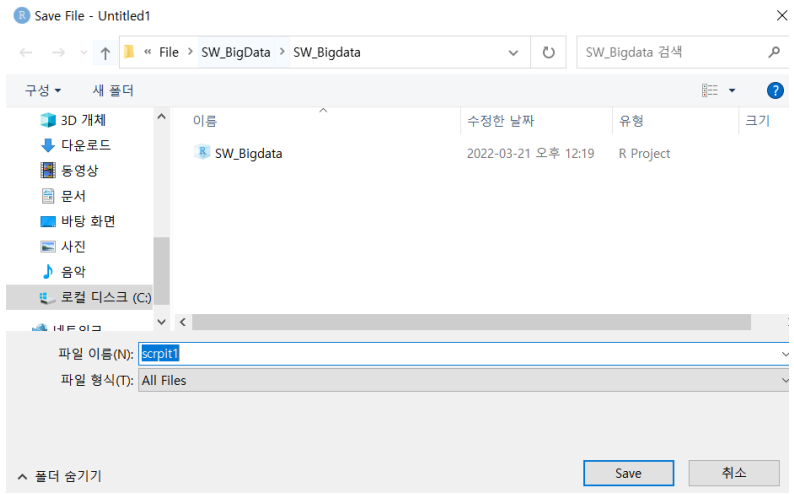
- R이 새로이 시작되면서 3개의 창이 생김
- 오른쪽 상단의 프로젝트 명과 파일 창에서 프로젝트 폴더 위치 확인



# WorkFlow

## 1) Script 저장하기

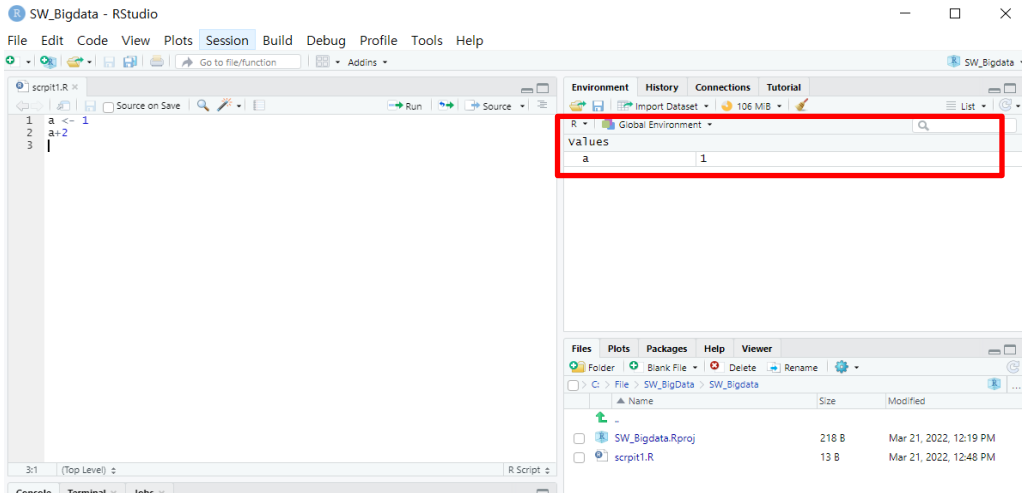
1) 새 스크립트 ([File] –[New File] –[R script]) 생성 후
2) 두 줄의 코드 입력
a <- 1
a+2
3) scrpit1로 저장하기 (Ctrl +S, 또는 [File]-[Save])




## 2) 결과물 입력

1) Script 실행, 두 줄 각각 Ctrl + Enter

a <- 1
a+2



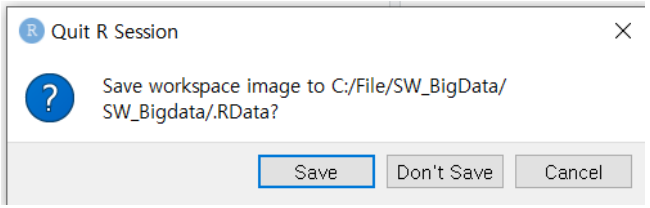
# WorkFlow

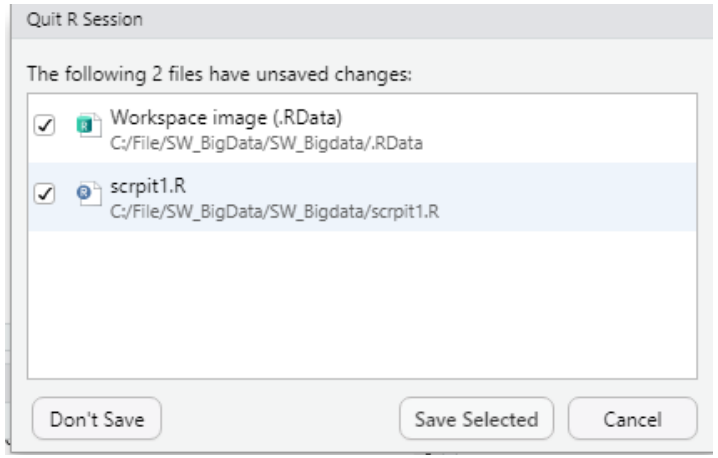


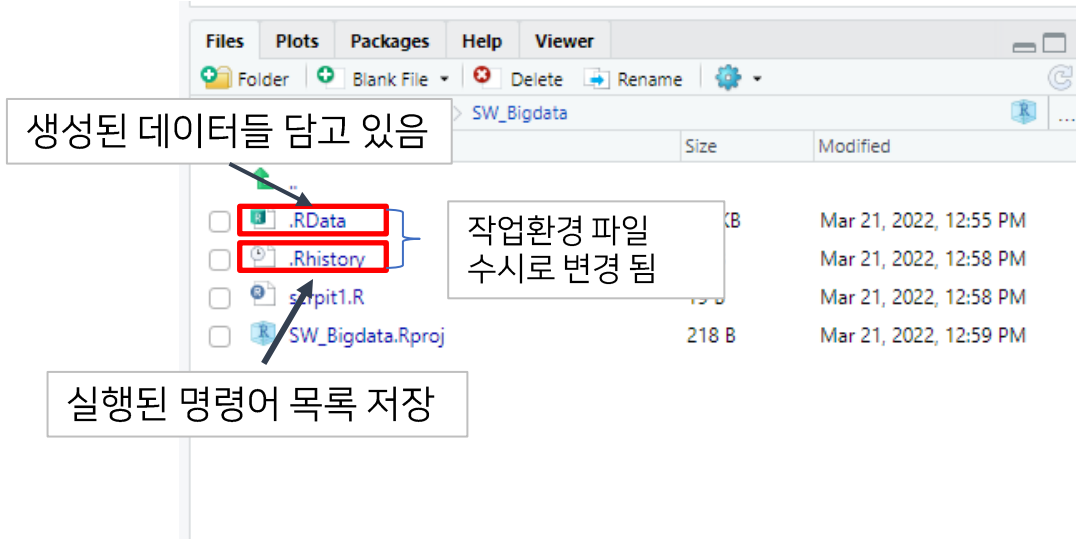


## 3) 프로젝트 저장하기

1) Rstudio는 자동 저장 기능을 갖추고 있음
2) R Studio 종료 시 저장 여부를 묻는 메시지 창이 나타남
3) [Save]버튼 선택 하면 Script와 데이터를 모두 저장할 것인지 묻는 창 나옴
4) 선택하여 저장



# 설정하기 : “<- “

R의 문법 중에서 가장 기본은 “<“와 “-” 합쳐진 “<- “ 기호 임, R은 모든 변수는 <- 를 통해 설정됨

R에서 이 기능들을 확인해 봅시다.

```
> print(lecture)
Error in print(lecture) : object 'lecture' not found
```

Lecture 에 “빅데이터처리” 를 지정

```
> lecture <-"빅데이터처리"
> print(lecture)
[1] "빅데이터처리"
```

```
> lecture("과목명은 빅데이터처리입니다")
Error in lecture("과목명은 빅데이터처리입니다") :
  could not find function "lecture"
> lecture <- Print
```

lecture에 print함수 지정도 가능

```
> lecture <- print
> lecture("과목명은 빅데이터처리입니다")
[1] "과목명은 빅데이터처리입니다"
```

# R 기본 – 데이터 구조





다음 중 분석 대상이 아닌 것은?

| 대학교 | 대학 | 결석 | 학점 |
|---|---|---|---|
| 삼육대학교 | 인문사회대학 | 0 | 4.3 |
| 삼육대학교 | 미래융합대학 | 0 | 3.4 |
| 삼육대학교 | 인문사회대학 | 1 | 2.3 |
| 삼육대학교 | 미래융합대학 | 2 | 4.5 |
| 삼육대학교 | 인문사회대학 | 1 | 3.7 |

# R 기본 – 데이터 구조



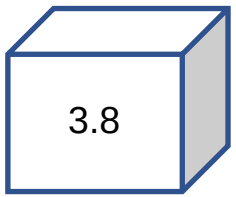


- 변수 (Variable)

| 학번 | 대학교 | 단과대학 | 결석 | 평점 | 등급 |
|---|---|---|---|---|---|
| 111111 | 삼육대학교 | 인문사회대학 | 0 | 4.3 | A+ |
| 111112 | 삼육대학교 | 미래융합대학 | 0 | 3.4 | B |
| 111113 | 삼육대학교 | 인문사회대학 | 1 | 2.3 | C |
| 111114 | 삼육대학교 | 미래융합대학 | 2 | 4.5 | A+ |
| 111115 | 삼육대학교 | 인문사회대학 | 1 | 3.7 | B+ |
| 111116 | 삼육대학교 | 과학기술대학 | 3 | 2.4 | C |

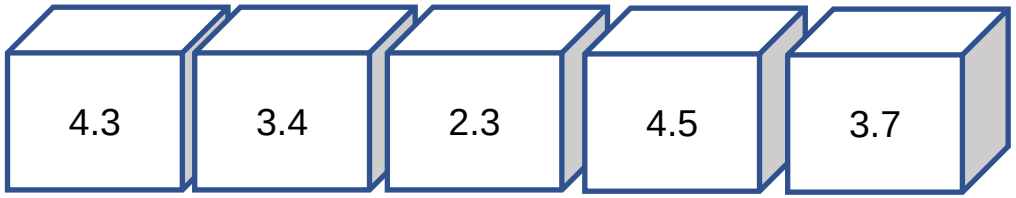
- 학점이라는 변수에 3.8이라는 데이터(값)이 있음

  “학점” 이라는 상자에 4.3 이라는 값을 넣음




- 벡터(Vector)

“학점” 이라는 상자에 여러 개의 값이 들어가 있음




- 같은 데이터 타입이어야 함
  (수치이면 모두 수치. 문자이면 모두 문자)
- R의 가장 기본적인 데이터 구조
- 일반적인 1차배열의 형태임
- 벡터는 숫자, 문자열, 논리(T/F), 팩터(범주형 데이터)의 데이터 타입을 다룰 수 있음
- 변수명, 벡터명 등은 영어로 명명할 것을 권장함

# R 기본 – 데이터 구조





- 벡터(Vector)

1) 숫자형 벡터 : 양수, 음수, 소수점등을 모두 담을 수 있음

2) 논리형 벡터 : 합격여부, 성공여부와 같이 논리 타입 (TRUE, FALSE, T/F 로 표현)

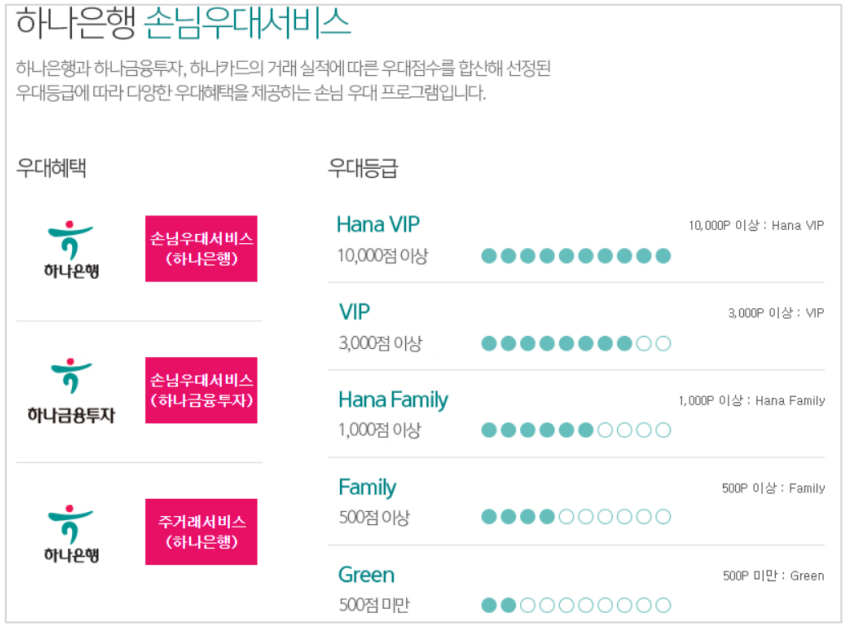
소문자 true, false나 , “” 함께 이용하면 논리형으로 인식하지 않음

3) 문자열 벡터 : “ “와 같이 사용

4) 팩터 : 범주형 데이터 이용시

성적 : A,B,C,D, F

고객등급 ; HanaVIP, VIP, Hana Family, Green



# R 기본 – 데이터 구조





- 벡터(Vector)

| ID 변수 | | 범주형 (팩터 벡터) | 수치형 (숫자형벡터) | 범주형 (팩터벡터) |
|---|---|---|---|---|
| 학번 | 대학교 | 단과대학 | 결석 | 평점 | 등급 |
| 111111 | 삼육대학교 | 인문사회대학 | 0 | 4.3 | A+ |
| 111112 | 삼육대학교 | 미래융합대학 | 0 | 3.4 | B |
| 111113 | 삼육대학교 | 인문사회대학 | 1 | 2.3 | C |
| 111114 | 삼육대학교 | 미래융합대학 | 2 | 4.5 | A+ |
| 111115 | 삼육대학교 | 인문사회대학 | 1 | 3.7 | B+ |
| 111116 | 삼육대학교 | 과학기술대학 | 3 | 2.4 | C |

순서가 있는 범주형 변수
: 순서형 또는 서열형

단순분류 범주형 변수
: 명목형

# R 기본 – 데이터 구조





- 벡터(Vector)

| 결석 |
| --- |
| 0 |
| 0 |
| 1 |
| 2 |
| 1 |
| 3 |

벡터는 “c” (combine) 함수를 이용해 생성됨

absence < -c(0,0,1,2,1,3)

데이터 타입 확인하는 함수 mode

> mode(absence)
> numeric

# R 기본 – 데이터 구조





| 단과대학 |
| --- |
| 인문사회대학 |
| 미래융합대학 |
| 인문사회대학 |
| 미래융합대학 |
| 인문사회대학 |
| 과학기술대학 |

Factor_name <- factor (x, levels,ordered)

x: 팩터로 변환할 값들

Levels ; 범주 정의 (범주 순서 정의 가능)

특별히 지정하지 않은 팩터 값 입력 순서대로 정의

Ordered : 서열형인 경우 TRUE로 설정

(기본은 FALSE / 명목형)

College1 <-"factor(c("인문사회대학","미래융합대학",
"인문사회대학", "미래융합대학",인문사회대학","과학기술대학"),
levels =c("인문사회대학","미래융합대학","과학기술대학"),
ordered="FALSE" )

- College <- c("인문사회대학","미래융합대학", "인문사회대학",
"미래융합대학",인문사회대학","과학기술대학")
College1 <- factor(College,
levels =c("인문사회대학","미래융합대학","과학기술대학"),
ordered="FALSE" )

- College2 <- as.factor(College)
: level 순서대로(ㄱ,ㄴ,ㄷ 순) 사용, 명목형일때
=> 가장 많이 사용

생각해 봅시다

순서형인 경우는 어떻게 해야 할까요?

# R 기본 – 데이터 구조

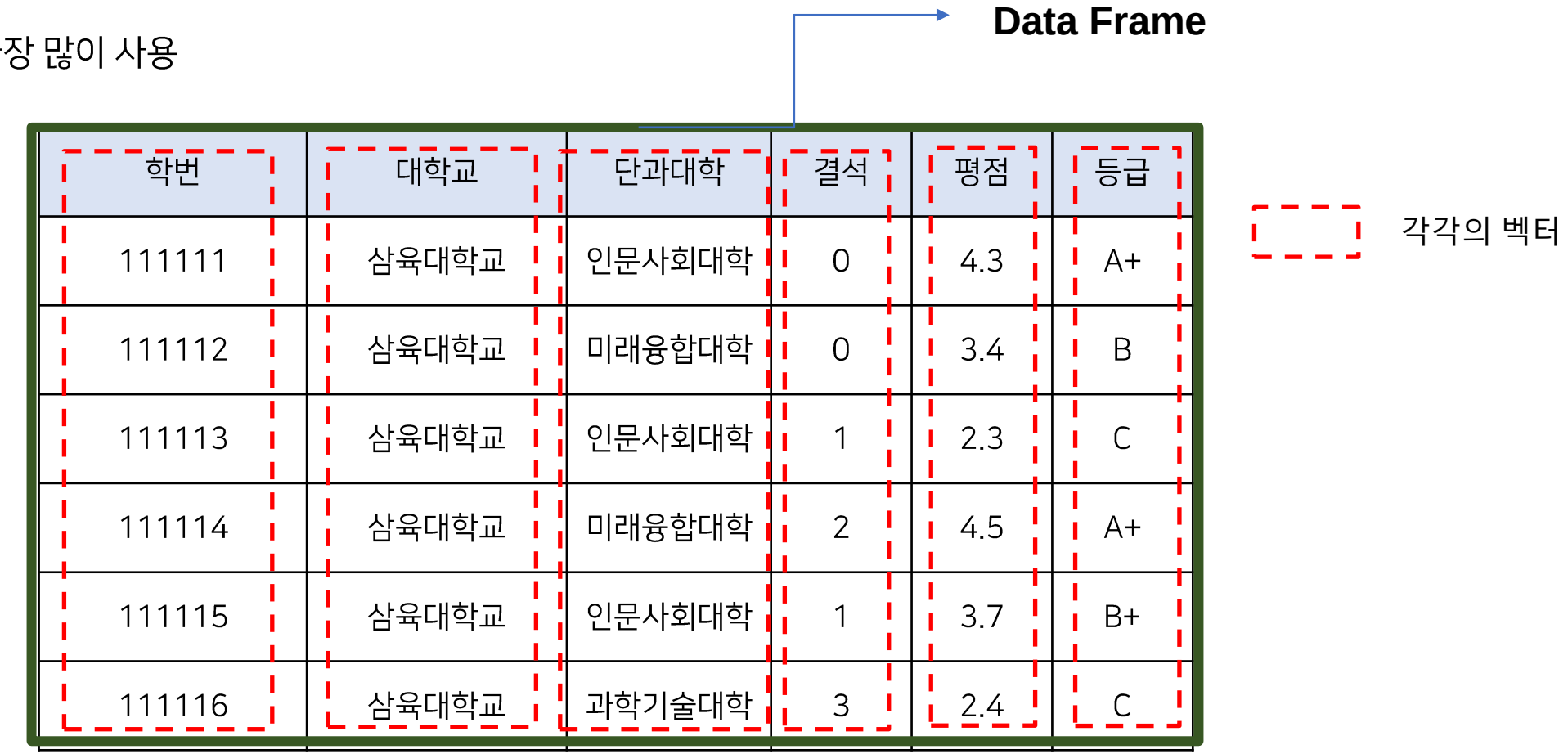




- Data Frame

  " 여러 데이터 유형을 한 번에 "

  분석 시 가장 많이 사용




| 학번 | 대학교 | 단과대학 | 결석 | 평점 | 등급 |
|---|---|---|---|---|---|
| 111111 | 삼육대학교 | 인문사회대학 | 0 | 4.3 | A+ |
| 111112 | 삼육대학교 | 미래융합대학 | 0 | 3.4 | B |
| 111113 | 삼육대학교 | 인문사회대학 | 1 | 2.3 | C |
| 111114 | 삼육대학교 | 미래융합대학 | 2 | 4.5 | A+ |
| 111115 | 삼육대학교 | 인문사회대학 | 1 | 3.7 | B+ |
| 111116 | 삼육대학교 | 과학기술대학 | 3 | 2.4 | C |

# R 기본 – 데이터 구조





- Data Frame

  " 여러 데이터 유형을 한 번에 "
  분석 시 가장 많이 사용

학번 -> ID,
대학교 -> university
대학 -> college
결석 -> absence
평점 -> gpa
등급 -> grade

| | | | | | | |
|---|---|---|---|---|---|---|
| 학번 | 111111 | 111112 | 111113 | 111114 | 111115 | 111116 |
| 대학교 | 삼육대학교 | 삼육대학교 | 삼육대학교 | 삼육대학교 | 삼육대학교 | 삼육대학교 |
| 대학 | 인문사회대학 | 미래융합대학 | 인문사회대학 | 미래융합대학 | 인문사회대학 | 과학기술대학 |
| 결석 | 0 | 0 | 1 | 2 | 1 | 3 |
| 평점 | 4.3 | 3.4 | 2.3 | 4.5 | 3.7 | 2.4 |
| 등급 | A+ | B | C | A+ | B+ | C |

Vector

조합

| 학번 | 대학교 | 대학 | 결석 | 평점 | 등급 |
|---|---|---|---|---|---|
| 111111 | 삼육대학교 | 인문사회대학 | 0 | 4.3 | A+ |
| 111112 | 삼육대학교 | 미래융합대학 | 0 | 3.4 | B |
| 111113 | 삼육대학교 | 인문사회대학 | 1 | 2.3 | C |
| 111114 | 삼육대학교 | 미래융합대학 | 2 | 4.5 | A+ |
| 111115 | 삼육대학교 | 인문사회대학 | 1 | 3.7 | B+ |
| 111116 | 삼육대학교 | 과학기술대학 | 3 | 2.4 | C |

Data Frame

DataFrame 생성 코드

Dataframe 명 <- data.frame(vector1,vector2,vector3,…)

# R 기본 – 데이터 구조





- Data Frame

DataFrame 생성 코드

Dataframe 명 <- data.frame(vector1,vector2,vector3,…, StringsAsFactors)

| 입력항목 | 설명 |
|---|---|
| vetor | 데이터 프레임에서 열이 될 벡터 |
| StringsAsFactor | 문자인 경우, 팩터로 변환여부 (기본값은 팩터변환)<br>StringsAsFactor = FALSE => 문자열 |

- 데이터 타입 변경 하는 방법 : as.데이터타입

  예> 수치형으로 변경 a <- as.numeric(a)

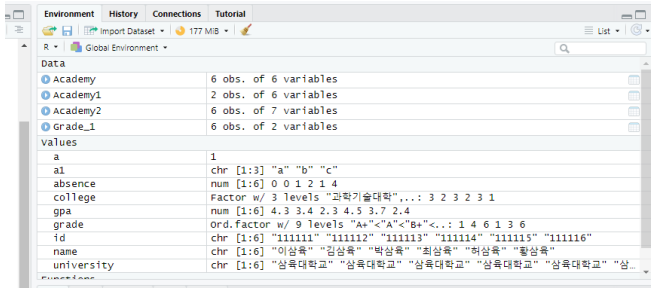
- 데이터 프레임 구조 확인 ; str(dataframe명)

- 데이터 셋 보기

  1> script창에서 데이터 셋명 쓰고 실행

```
> Academy
     id university      college absence gpa grade
1 111111 삼육대학교 인문사회대학       0 4.3    A+
2 111112 삼육대학교 미래융합대학       0 3.4     B
3 111113 삼육대학교 인문사회대학       1 2.3     C
4 111114 삼육대학교 미래융합대학       2 4.5    A+
5 111115 삼육대학교 인문사회대학       1 3.7    B+
6 111116 삼육대학교 과학기술대학       4 2.4     C
>
```

2> Environment에서 dataframe 클릭



# R 기본 – 데이터 구조





- R 실습

### 1) 데이터 타입에 맞게 벡터 생성

- 벡터 명 변경

학번 -> ID,
대학교 -> university
대학 -> college
결석 -> absence
평점 -> gpa
등급 -> grade

### 2) 생성한 벡터들을 dataframe화 하기

- Dataframe명 : Academic

### 3) 데이터 타입 확인하기

학번: 문자열
단과대학 : 범주형(명목형)
등급 : 범주형(순서형)

| 학번 | 대학교 | 단과대학 | 결석 | 평점 | 등급 |
|---|---|---|---|---|---|
| 111111 | 삼육대학교 | 인문사회대학 | 0 | 4.3 | A+ |
| 111112 | 삼육대학교 | 미래융합대학 | 0 | 3.4 | B |
| 111113 | 삼육대학교 | 인문사회대학 | 1 | 2.3 | C |
| 111114 | 삼육대학교 | 미래융합대학 | 2 | 4.5 | A+ |
| 111115 | 삼육대학교 | 인문사회대학 | 1 | 3.7 | B+ |
| 111116 | 삼육대학교 | 과학기술대학 | 3 | 2.4 | C |

# R 기본 – 데이터 구조





- 데이터 접근 하기

[ ]를 이용하여 특정 위치의 데이터를 선택할 수 있음

데이터프레임명 [행 위치 벡터, 열 위치 벡터]

| 학번 | 대학교 | 단과대학 | 결석 | 평점 | 등급 |
|---|---|---|---|---|---|
| 111111 | 삼육대학교 | 인문사회대학 | 0 | 4.3 | A+ |
| 111112 | 삼육대학교 | 미래융합대학 | 0 | 3.4 | B |
| 111113 | 삼육대학교 | 인문사회대학 | 1 | 2.3 | C |
| 111114 | 삼육대학교 | 미래융합대학 | 2 | 4.5 | A+ |
| 111115 | 삼육대학교 | 인문사회대학 | 1 | 3.7 | B+ |
| 111116 | 삼육대학교 | 과학기술대학 | 3 | 2.4 | C |

- 3행에 4열 값 출력

Academy [3,4]

=> 1

| 학번 | 대학교 | 단과대학 | 결석 | 평점 | 등급 |
|---|---|---|---|---|---|
| 111111 | 삼육대학교 | 인문사회대학 | 0 | 4.3 | A+ |
| 111112 | 삼육대학교 | 미래융합대학 | 0 | 3.4 | B |
| 111113 | 삼육대학교 | 인문사회대학 | 1 | 2.3 | C |
| 111114 | 삼육대학교 | 미래융합대학 | 2 | 4.5 | A+ |
| 111115 | 삼육대학교 | 인문사회대학 | 1 | 3.7 | B+ |
| 111116 | 삼육대학교 | 과학기술대학 | 3 | 2.4 | C |

- 3,4행 중 4,5열 값

Academy [c(3,4),c(4,5)]

| absence | gpa |
|---|---|
| 3 | 1 2.3 |
| 4 | 2 4.5 |

# R 기본 – 데이터 구조





- 데이터 접근 하기
  - 4,5 열 전체 행 출력

Academy [ ,c(4,5)]

```
  absence gpa
1       0 4.3
2       0 3.4
3       1 2.3
4       2 4.5
5       1 3.7
6       4 2.4
```

| 학번 | 대학교 | 단과대학 | 결석 | 평점 | 등급 |
|---|---|---|---|---|---|
| 111111 | 삼육대학교 | 인문사회대학 | 0 | 4.3 | A+ |
| 111112 | 삼육대학교 | 미래융합대학 | 0 | 3.4 | B |
| 111113 | 삼육대학교 | 인문사회대학 | 1 | 2.3 | C |
| 111114 | 삼육대학교 | 미래융합대학 | 2 | 4.5 | A+ |
| 111115 | 삼육대학교 | 인문사회대학 | 1 | 3.7 | B+ |
| 111116 | 삼육대학교 | 과학기술대학 | 3 | 2.4 | C |

  - 3,4행 전체 열 출력

Academy [c(3,4), ]

```
      id university      college absence gpa grade
3 111113 삼육대학교 인문사회대학       1 2.3     C
4 111114 삼육대학교 미래융합대학       2 4.5    A+
> |
```

| 학번 | 대학교 | 단과대학 | 결석 | 평점 | 등급 |
|---|---|---|---|---|---|
| 111111 | 삼육대학교 | 인문사회대학 | 0 | 4.3 | A+ |
| 111112 | 삼육대학교 | 미래융합대학 | 0 | 3.4 | B |
| 111113 | 삼육대학교 | 인문사회대학 | 1 | 2.3 | C |
| 111114 | 삼육대학교 | 미래융합대학 | 2 | 4.5 | A+ |
| 111115 | 삼육대학교 | 인문사회대학 | 1 | 3.7 | B+ |
| 111116 | 삼육대학교 | 과학기술대학 | 3 | 2.4 | C |

선정한 값만 저장하려면:Academy1 <- Academy [c(3,4), ]

# R 기본 – 데이터 구조





- 데이터 접근 하기

  이름이라는 새로운 벡터 추가하려면

  1) 벡터 생성 (name)

  name < -c("이삼육","김삼육","박삼육","최삼육","허삼육","황삼육")

  2) cbind이용 => Column 추가

  Academy2 <- cbind(Academy,name)

| 학번 | 대학교 | 단과대학 | 결석 | 평점 | 등급 |
|---|---|---|---|---|---|
| 111111 | 삼육대학교 | 인문사회대학 | 0 | 4.3 | A+ |
| 111112 | 삼육대학교 | 미래융합대학 | 0 | 3.4 | B |
| 111113 | 삼육대학교 | 인문사회대학 | 1 | 2.3 | C |
| 111114 | 삼육대학교 | 미래융합대학 | 2 | 4.5 | A+ |
| 111115 | 삼육대학교 | 인문사회대학 | 1 | 3.7 | B+ |
| 111116 | 삼육대학교 | 과학기술대학 | 3 | 2.4 | C |

| | id | university | college | absence | gpa | grade | name |
|---|---|---|---|---|---|---|---|
| 1 | 111111 | 삼육대학교 | 인문사회대학 | 0 | 4.3 | A+ | 이삼육 |
| 2 | 111112 | 삼육대학교 | 미래융합대학 | 0 | 3.4 | B | 김삼육 |
| 3 | 111113 | 삼육대학교 | 인문사회대학 | 1 | 2.3 | C | 박삼육 |
| 4 | 111114 | 삼육대학교 | 미래융합대학 | 2 | 4.5 | A+ | 최삼육 |
| 5 | 111115 | 삼육대학교 | 인문사회대학 | 1 | 3.7 | B+ | 허삼육 |
| 6 | 111116 | 삼육대학교 | 과학기술대학 | 4 | 2.4 | C | 황삼육 |

# R 기본 – 데이터 구조





- 데이터 접근 하기
  - 열 이름으로도 접근 가능

| | id | university | college | absence | gpa | grade | name |
|---|---|---|---|---|---|---|---|
| 1 | 111111 | 삼육대학교 | 인문사회대학 | 0 | 4.3 | A+ | 이삼육 |
| 2 | 111112 | 삼육대학교 | 미래융합대학 | 0 | 3.4 | B | 김삼육 |
| 3 | 111113 | 삼육대학교 | 인문사회대학 | 1 | 2.3 | C | 박삼육 |
| 4 | 111114 | 삼육대학교 | 미래융합대학 | 2 | 4.5 | A+ | 최삼육 |
| 5 | 111115 | 삼육대학교 | 인문사회대학 | 1 | 3.7 | B+ | 허삼육 |
| 6 | 111116 | 삼육대학교 | 과학기술대학 | 4 | 2.4 | C | 황삼육 |

```
> Academy [, c("college","gpa")]
       college gpa
1 인문사회대학 4.3
2 미래융합대학 3.4
3 인문사회대학 2.3
4 미래융합대학 4.5
5 인문사회대학 3.7
6 과학기술대학 2.4
>
```

- 데이터프레임명$열이름 접근가능

```
> Academy$gpa
[1] 4.3 3.4 2.3 4.5 3.7 2.4
```

# R 기본 – 데이터 구조





- R 실습

1) 111115 학번의 모든 정보를 출력하시오

2) 단과대학별 등급분포를 알고자 한다
: 단과대학정보와 등급 모든 정보만을 선정하여, Grade_1이라는 새로운 셋을 생성하시오

3) 미래융합대학의 결석과 평점(헴) 정보를 가져오시오