

Lecture Note 02

Data Handling & Manipulation



Dohyung Bang

Fall, 2021

Syllabus

Week	Date	Topic	Note
1	9/6(월)	R Basic - R 기초 문법 학습	
2	9/13(월)	R Basic – Data Manipulation I	과제#1
3	9/20(월) (추석)	<추석> (보충영상) R Basic - Data Manipulation II	과제#2
4	9/27(월)	Descriptive Analytics I - 데이터 요약하기/상관관계/차이검증	
5	10/4(월) (대체공휴일)	<대체공휴일> (보충영상) Descriptive Analytics II - 데이터 시각화	과제#3
6	10/11(월) (대체공휴일)	<대체공휴일> (보충영상) Supplementary Topic I - 외부 데이터 수집 (정적 콘텐츠 수집)	과제#4
7	10/18(월)	Predictive Analytics I – Linear regression & Logistic Regression	
8	10/25(월)	Predictive Analytics II – Clustering & Latent Class Analysis	시험 대체 수업
9	11/1(월)	Predictive Analytics III – Tree-based Model and Bagging (Random Forest)	
10	11/8(월)	Predictive Analytics IV – Association Rules	
11	11/15(월)	Supplementary Topic II - 외부 데이터 수집 (동적 콘텐츠 수집)	
12	11/22(월)	Prescriptive Analytics I – Linear Programming	과제#5
13	11/29(월)	Prescriptive Analytics II – Data Envelopment Analysis (DEA)	
14	12/6(월)	Prescriptive Analytics III – Integer Programming	과제#6
15	12/13(월)	Prescriptive Analytics IV – Simulation	Quiz
16	12/20(월)	Final Presentation	


Lecture 2-1

데이터, 변수,
그리고 “R”

자료(data)의 분류



자료(data)는 정형자료(Structured data)와 비정형자료(Unstructured data)로 나뉨

Structured Data



0.103	0.176	0.387	0.300	0.379
0.333	0.384	0.564	0.587	0.857
0.421	0.309	0.654	0.729	0.228
0.266	0.750	1.056	0.936	0.911
0.225	0.326	0.643	0.337	0.721
0.187	0.586	0.529	0.340	0.829
0.153	0.485	0.560	0.428	0.628

Unstructured Data



- 우리가 일상적으로 접하는 자료는 대개 정형자료이나 전체 자료 중 정형자료의 비중은 20%가 채 되지 않으며, 나머지 80%에 해당하는 비정형자료를 의미있게 분석하는 것이 매우 중요함

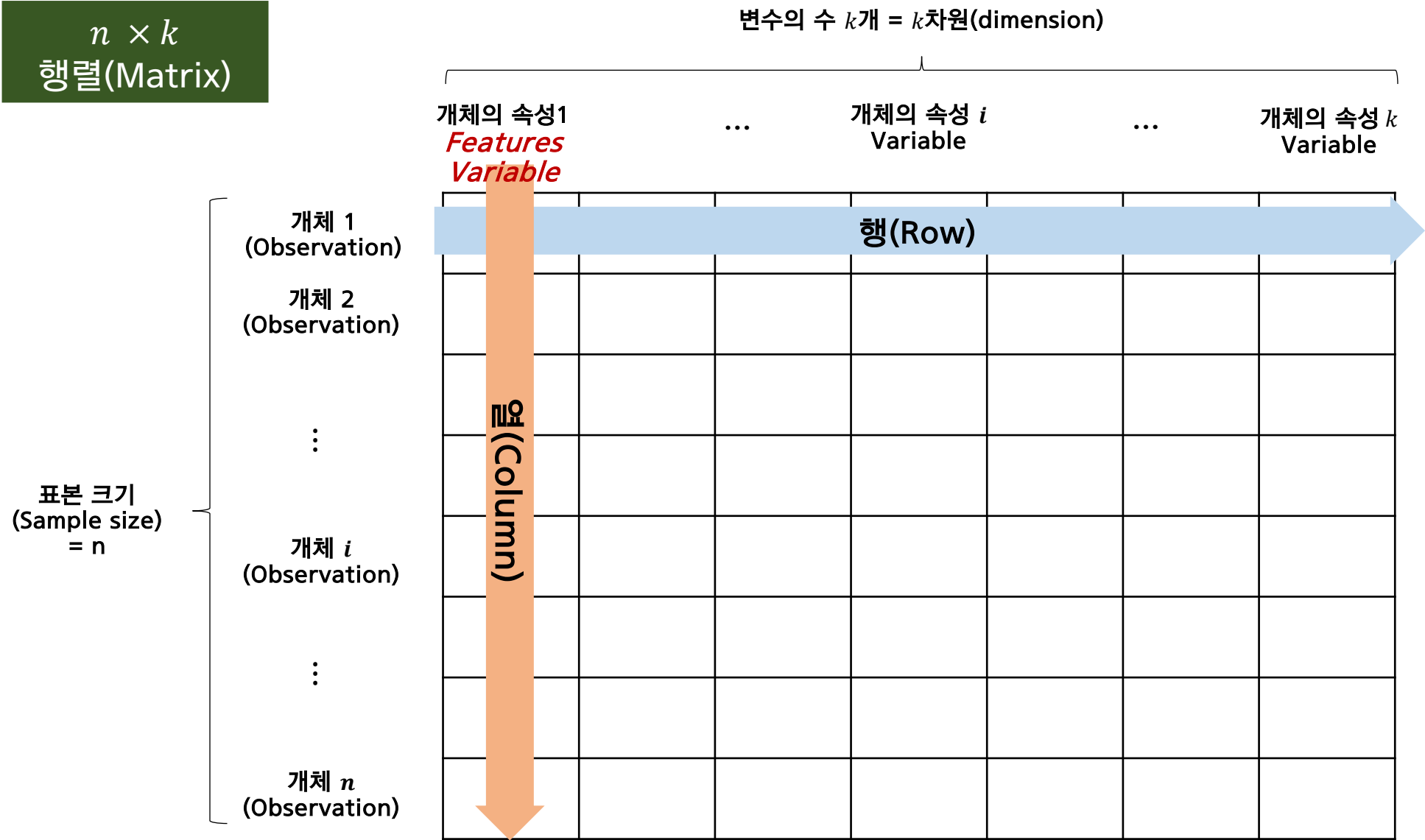
변수(Variable)란 무엇인가?

변수(Variable)란 **모형에 전달되는 정보나 그 밖의 상황에 따라 바뀔 수 있는 값**을 의미

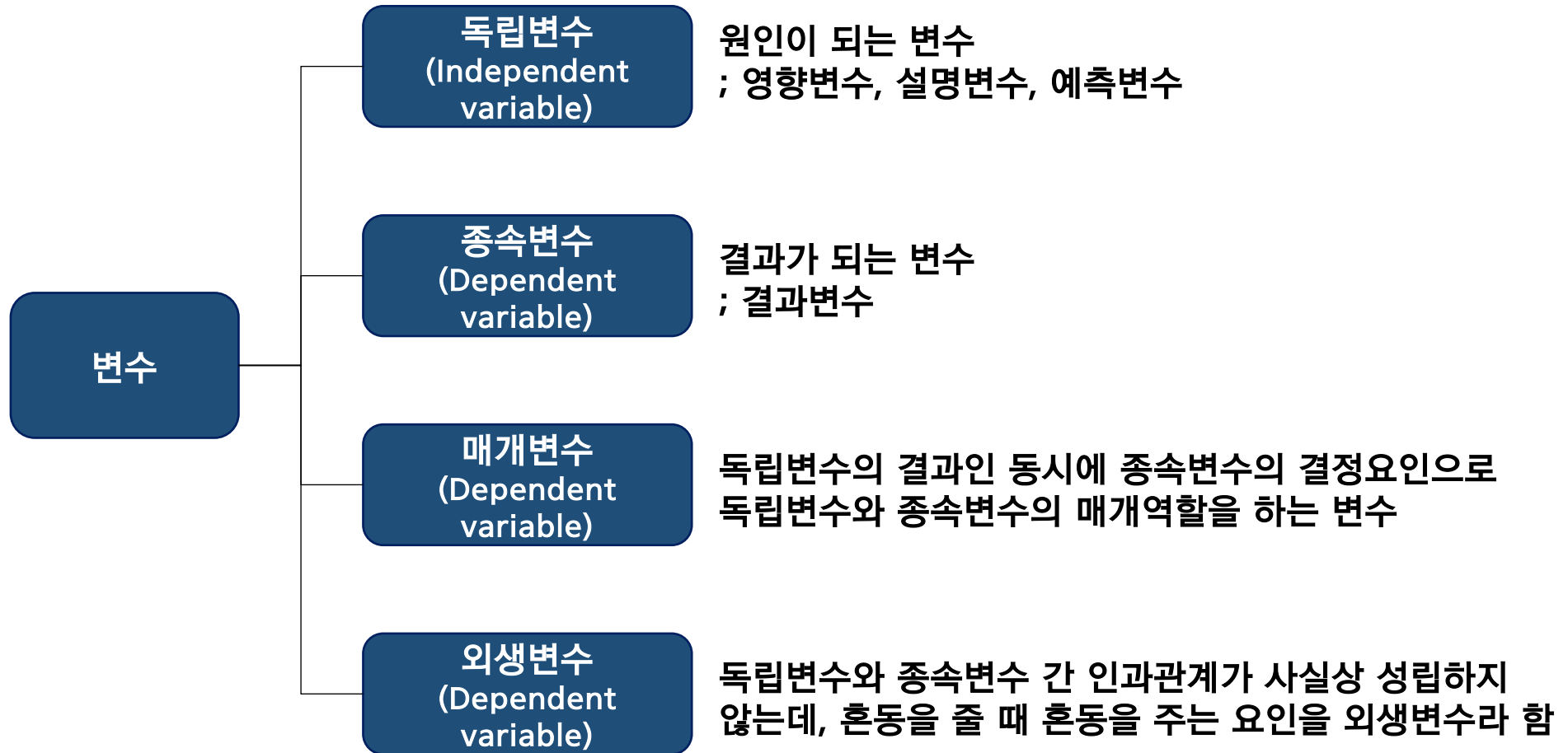
변수(Variable) = 개체의 속성(Feature)



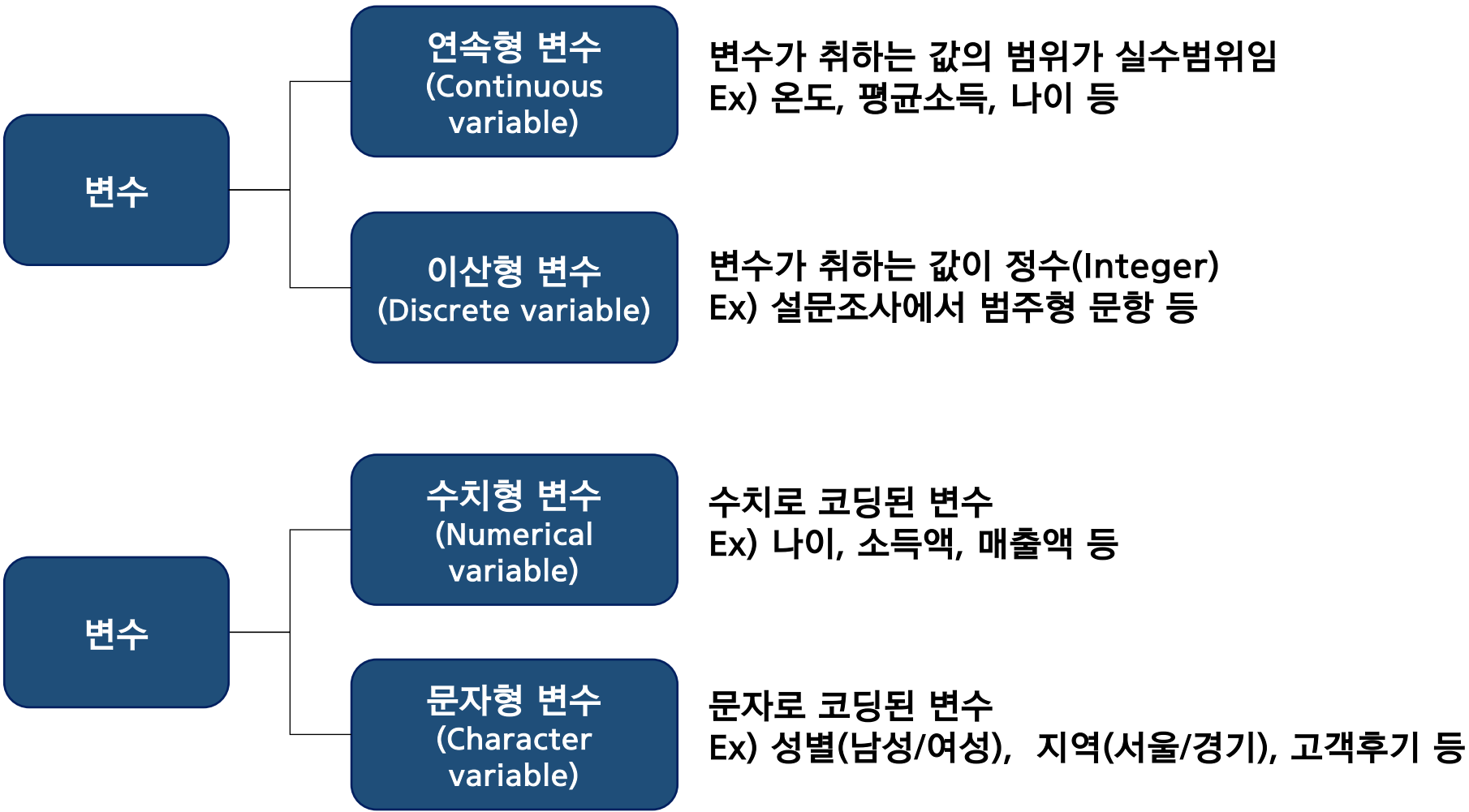
테이블 데이터의 구조



기능에 따른 변수(Variable) 분류



속성 및 형태에 따른 변수(Variable) 분류



척도(Scale)의 종류

명목척도

Nomial scale

데이터 항목의 속성을 단지
숫자로 식별하기 위한 목적
숫자의 크기 의미 X
서열화 X
특정 범주만을 의미함

Ex) 성별, 산업분류

서열척도

Ordinal scale

명목척도의 특성을 포함
크기 순의 서열화 O
순서에 관한 정보를 나타냄
(구체적인 차이에 관한 정
보는 포함 X)

Ex) 석차, 모스 경도

간격척도

Interval scale

값 간 간격이 고정된 척도
서열척도의 특성 포함
값 간의 차이가 의미 있음
값 간의 비율계산은 의미 X

Ex) 온도, 토익성적

비율척도

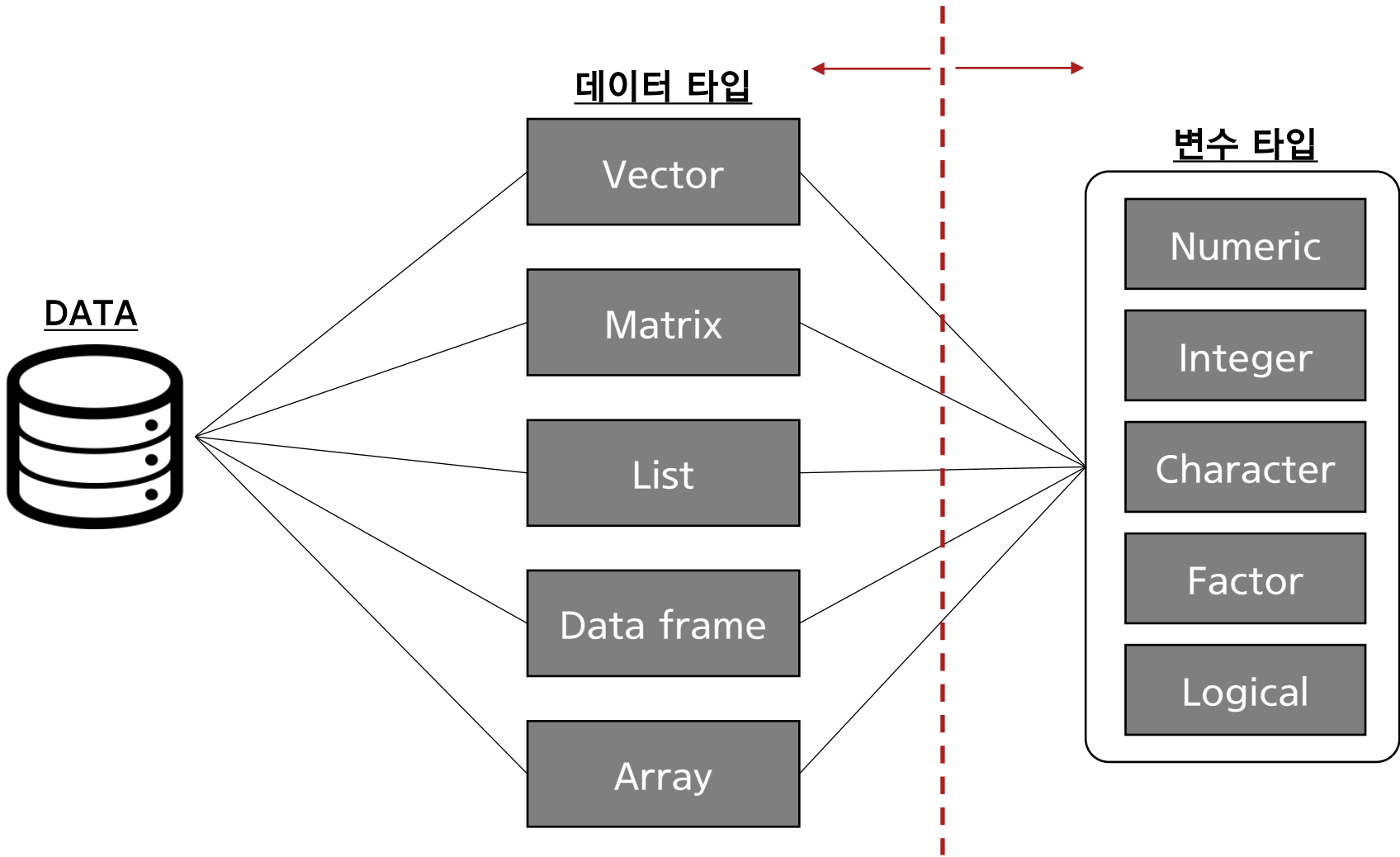
Ratio scale

크기의 비교가 가능
간격척도의 특성 포함
값 간 간격이 동일
비율계산이 가능

Ex) 길이, 무게, 시간, 나이

R에서 정의하는 “데이터”와 “변수”

데이터 안에 변수가 포함된 개념으로
“데이터” 타입과 “변수” 타입을 잘 구분할 필요가 있음



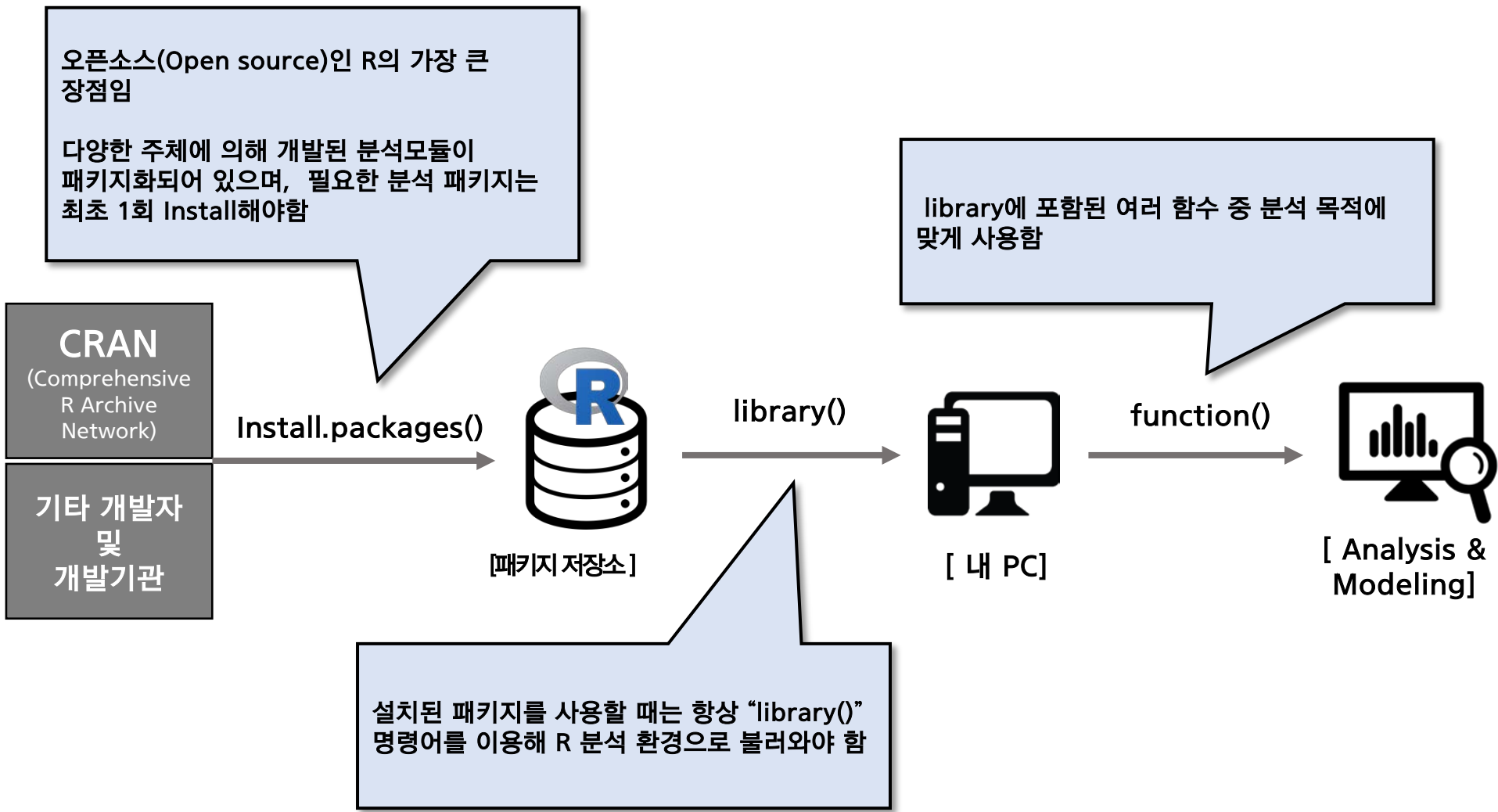
R에서 정의하는 “변수”

Numeric	<ul style="list-style-type: none">실수 범위의 값으로 표현되는 연속형(continuous)인 숫자형 변수	Ex) 매출액, 온도, 확률값 등
Integer	<ul style="list-style-type: none">정수 범위의 값으로 표현되는 숫자형 변수로, 모든 Integer는 Numeric에 속함	Ex) 사람 수, 박스 수량, 주문 량 등
Character	<ul style="list-style-type: none">문자형 변수로 문자 외에 의미를 지니지 않음	Ex) “Male”, “Female”, “Apple” 등
Factor	<ul style="list-style-type: none">문자형이든, 숫자형이든 그 형태가 아니라 구분을 위한 별도의 의미를 지니는 변수로, 명목척도와 성격이 같음	Ex) “Male”, “Female” or 1,2,3,4, 등번호
Logical	<ul style="list-style-type: none">논리형 변수는 TRUE와 FALSE 두가지로 구성되어 종종 쓰이며 Factor와 마찬가지로 TRUE이면 1, FALSE면 0과 같은 의미를 일반적으로 내포함	Ex) 구매 여부에 대해 구매했으면 TRUE, 구매 안했으면 FALSE

R에서 정의하는 “데이터”

Vector	<ul style="list-style-type: none">동일한 유형의 데이터로 구성된 1차원 데이터로, 하나의 행 또는 열만 있는 경우 Vector data라 부름한 Vector에는 동일한 타입의 데이터만 포함될 수 있음
Matrix	<ul style="list-style-type: none">둘 이상의 행 또는 열이 모이면 행렬이라고 부름수학적/대수적 연산이 가능
Data Frame	<ul style="list-style-type: none">Matrix와 동일한 형태이나 오로지 데이터 분석을 위해 정의되는 형태이므로 수학적/대수적 연산은 불가하며, “변수명”이 붙음
List	<ul style="list-style-type: none">여러 형태의 값 또는 데이터를 묶어 놓을 수 있는 데이터 형태로, 변수와 매트릭스, 매트릭스와 매트릭스, 데이터 프레임과 또 다른 리스트 등 여러 형태의 데이터 및 변수를 담을 수 있는 데이터 형태
Array	<ul style="list-style-type: none">3차원 이상의 데이터를 정의할 때 나타내는 데이터 타입으로, 이미지/영상 데이터의 경우 3,4차원 이상이기 때문에 Array로 주로 정의됨

R은 어떻게 작동하는가?



R script(.r) vs Markdown(.rmd)

➤ R script(.r)

- R의 코드를 실행 및 작성하기 위한 가장 기본적인 코드 작성 단위
- 주석 처리 '#'를 해주지 않으면 모든 Line을 코드로 인식함

➤ R Markdown (.rmd)

- R을 이용한 상호작용 활동을 위해 만들어진 작성 확장자로 Markdown을 이용해 html 문서, Words 문서, PDF 문서 등을 생성할 수 있음
- 따로 코드 Chunk 처리를 해줘야만 코드가 실행되고, 나머지 부분은 문서 처럼 작성이 가능함
- 코드에 대한 주석, 상호작용을 위해 코드북은 Markdown을 주로 활용할 예정

R 함수에 대한 이해

함수 소괄호 안에 들어가는 부분을 함수의
인자(Parameter)라 부름. 함수 인자는 콤마로 구분되며,
함수를 만들 때, 순서를 인식하거나 혹은 인자를
지정해줌으로써 인자를 구분하도록 되어있음

Function(**data, option1, option 2, ...**)



R 함수는 소괄호 '(' ')' 를 이용함

ex)

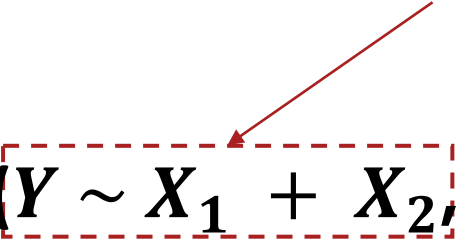
```
Function(my_data, "option1_name", "option2_name")
```

```
Function(data = my_data, option2 = "option2_name", option1 = "option1_name")
```

R 함수의 Formula에 대한 이해

$$Y = aX_1 + bX_2 + c$$

Function($Y \sim X_1 + X_2$, data = data)



등호(=)는 물결(~)로 대체

항상 종속변수가 물결 좌변, 독립변수가 물결 우변

➤ Formula가 적용되는 함수

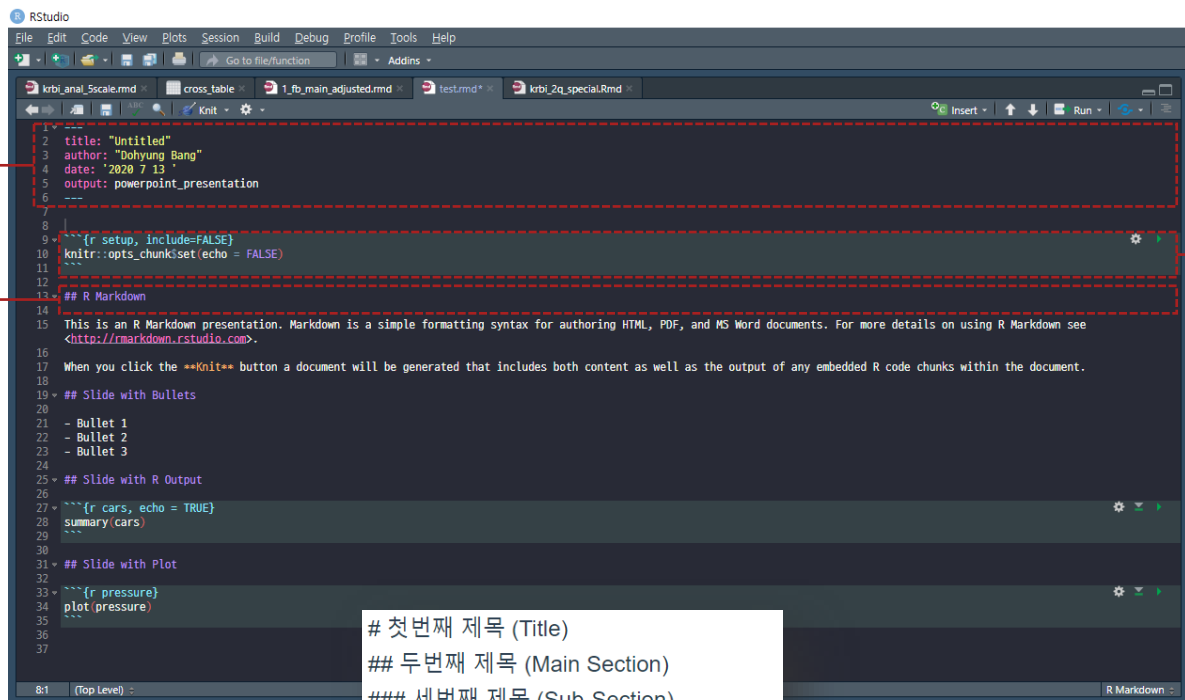
- 인과모형(선형 회귀모형 및 분류모형)
- 시각화 시 Group에 따른 차이 표현

Lecture 2-2

R 마크다운
문법 요소

R 마크다운 문법 기초(1/2)

- 1 문서 제목(title) / 저자(author) / 날짜(date) 정보와
문서 템플릿 형태(word / html / pdf / slide etc.) 정보 포함



```
1 title: "Untitled"
2 author: "Dohyung Bang"
3 date: "2020 7 13"
4 output: powerpoint_presentation
5 ---
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
```

- 2 설정 덩어리(set up chunk)
: 문서를 처음 생성하면 최초로 보이는 코드 덩어리로, 전체 문서의 기본 설정값 (Default)을 설정하는 코드 chunk임.

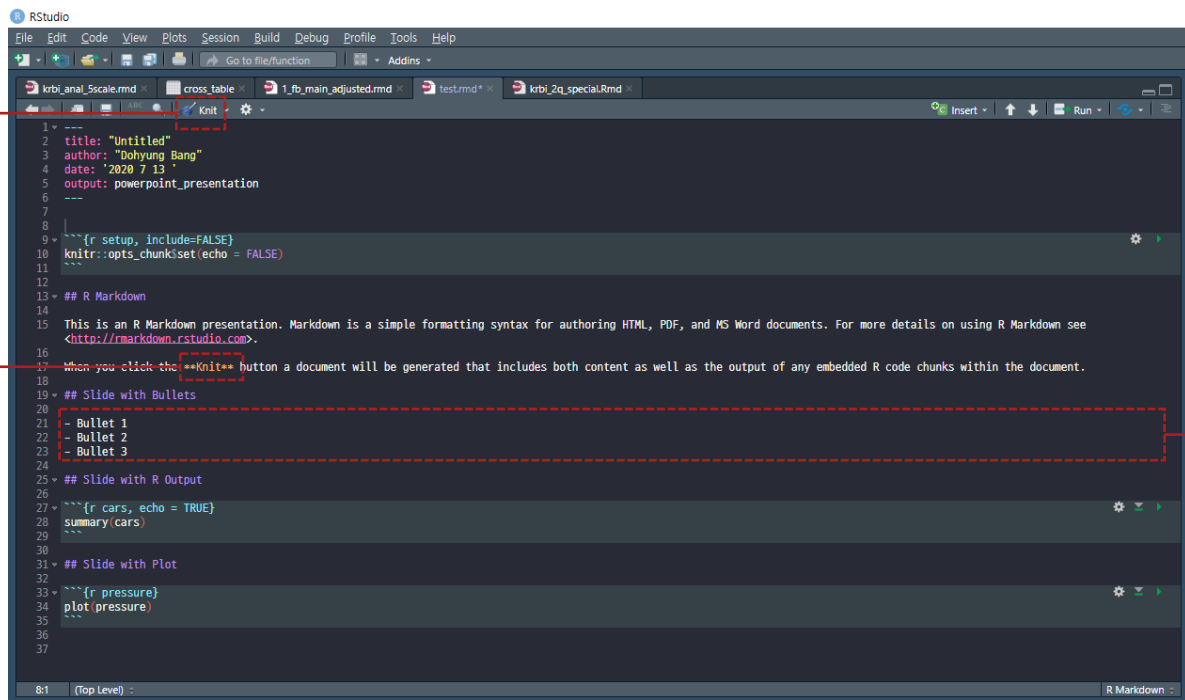
- eval = FALSE / TRUE
 - 코드를 실행하지 않는다. / 실행한다.
- echo = FALSE / TRUE
 - 코드를 보여주지 않는다. / 보여준다.
- Include = FALSE / TRUE
 - 실행 결과를 보여주지 않는다.
- message = FALSE / TRUE
 - 실행 때 나오는 메시지를 보여주지 않는다.
- warning = FALSE / TRUE
 - 실행 때 나오는 경고를 보여주지 않는다.
- error = TRUE / FALSE
 - 에러가 있어도 실행하고 에러코드를 보여준다.
- fig.height = 10
 - 그림 높이, R로 그린 그림에만 해당한다.
- fig.width = 12
 - 그림 너비, R로 그린 그림에만 해당한다.
- fig.align = 'center'
 - 그림 위치, R로 그린 그림에만 해당한다.

- 3 코드 chunk 밖의 문자는 모두 text로 인식되며, 이때 `#`의 수에 따라 제목 수준이 달라진다.

- # 첫번째 제목 (Title)
 - ## 두번째 제목 (Main Section)
 - ### 세번째 제목 (Sub-Section)
 - #### 네번째 제목 (Sub-sub section)
- 첫번째 제목 (Title)
- 두 번째 제목 (Main Section)
- 세 번째 제목 (Sub-Section)
- 네 번째 제목 (Sub-sub Section)

R 마크다운 문법 기초(2/2)

- 4 작업이 끝난 마크다운 문서를 지정해놓은 문서 템플릿으로 생성할 땐, '니트(Knit)'를 실행한다.



```
1 ---
2 title: "Untitled"
3 author: "Dohyung Bang"
4 date: "2020 7 13"
5 output: powerpoint_presentation
6 ---
7
8
9 ```{r setup, include=FALSE}
10 knitr::opts_chunkset(echo = FALSE)
11 ```
12
13 ## R Markdown
14
15 This is an R Markdown presentation. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see
16 <http://rmarkdown.rstudio.com>.
17
18 When you click the Knit button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document.
19
20 ## Slide with Bullets
21 - Bullet 1
22 - Bullet 2
23 - Bullet 3
24
25 ## Slide with R Output
26
27 ```{r cars, echo = TRUE}
28 summary(cars)
29 ```
30
31 ## Slide with Plot
32
33 ```{r pressure}
34 plot(pressure)
35 ```
36
37
```

- 6 불릿(Bullet) 만들기
: 일반적으로 하이픈(-), 별표(*), 더하기(+)
혹은 숫자를 적용하면 텍스트 구분자가
생성됨

```
21 - 위 그래프는 ...
22 - 위 표는 ...
23 - 따라서, ...
24
25 * 위 그래프는 ...
26 * 위 표는 ...
27 * 따라서, ...
28
29 + 위 그래프는 ...
30 + 위 표는 ...
31 + 따라서, ...
```



- 위 그래프는 ...
- 위 표는 ...
- 따라서, ...

```
40 1. 위 그래프는 ...
41 2. 위 표는 ...
42 3. 따라서, ...
```

1. 위 그래프는 ...
2. 위 표는 ...
3. 따라서, ...

- 5 텍스트를 강조하고자 할 때, 다음과
같이 *을 이용하여 **Bold** 혹은 *Italic*을
표현할 수 있다.

```
20 **굵게(Bold)**
21 *이탤릭(Italic)*
22
23
24 ~강조(Highlight)~
```



굵게(Bold)
이탤릭(Italic)
강조(Highlight)