

Lecture Note 01

Preview & Intro to “R”



Fall, 2021

Fall, 2021

강의 개요

COURSE DESCRIPTION

본 수업은 데이터를 이해하는 역량을 바탕으로 데이터 기반의 의사결정을 할 수 있도록 다양한 Analytics 방법론 학습을 목표로 한다. 본 수업은 Descriptive, Predictive, and Prescriptive Analytics 방법론을 학습하고, R 프로그래밍 언어를 기반으로 각 방법론을 구현 및 해석하는 데 초점을 둔다. 특히, 방법론의 기본적인 수학적 이해를 바탕으로 사회과학 현상으로서의 적용에 초점을 두고 방법론의 적용가능성을 탐구하는데 목표가 있다. 더불어 수강생들은 본 수업 간 진행되는 리서치 프로젝트를 통해 포괄적으로 수행함으로써 비즈니스 애널리틱스 기반의 의사결정을 이해할 수 있다.

TEXTBOOKS

자체 제작 강의노트 제공
자체 제작 R 실습노트 제공

REQUIREMENT

기초 확률통계

강의 개요

COURSE FORMAT

본 수업은 이론 20%, 실습 80%의 강의(Lecture) 형태로 진행되며, 별도의 Paper review는 포함되지 않는다. 평가는 격주에 1회 제공되는 코딩 과제, 학기말 Quiz 및 학기말Final research proposal로 이뤄진다.

PERFORMANCE EVALUATION

Take home assignment 30%

Quiz 20%

Final research proposal 40%

Class participation 10%

TERM PAPER PRESENTATION

2인 1조 Team Project 진행

강좌 커리큘럼

Session	Date	Topic	Note
1	9/6(월)	R Basic - R 기초 문법 학습	과제#1
2	9/13(월)	R Basic - Data Manipulation	
3	9/20(월) (추석)	<추석> (보충영상) R 기초 II - Data Manipulation	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 컨텐츠 수집) + R 을 이용한 알림 Bot 만들기	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

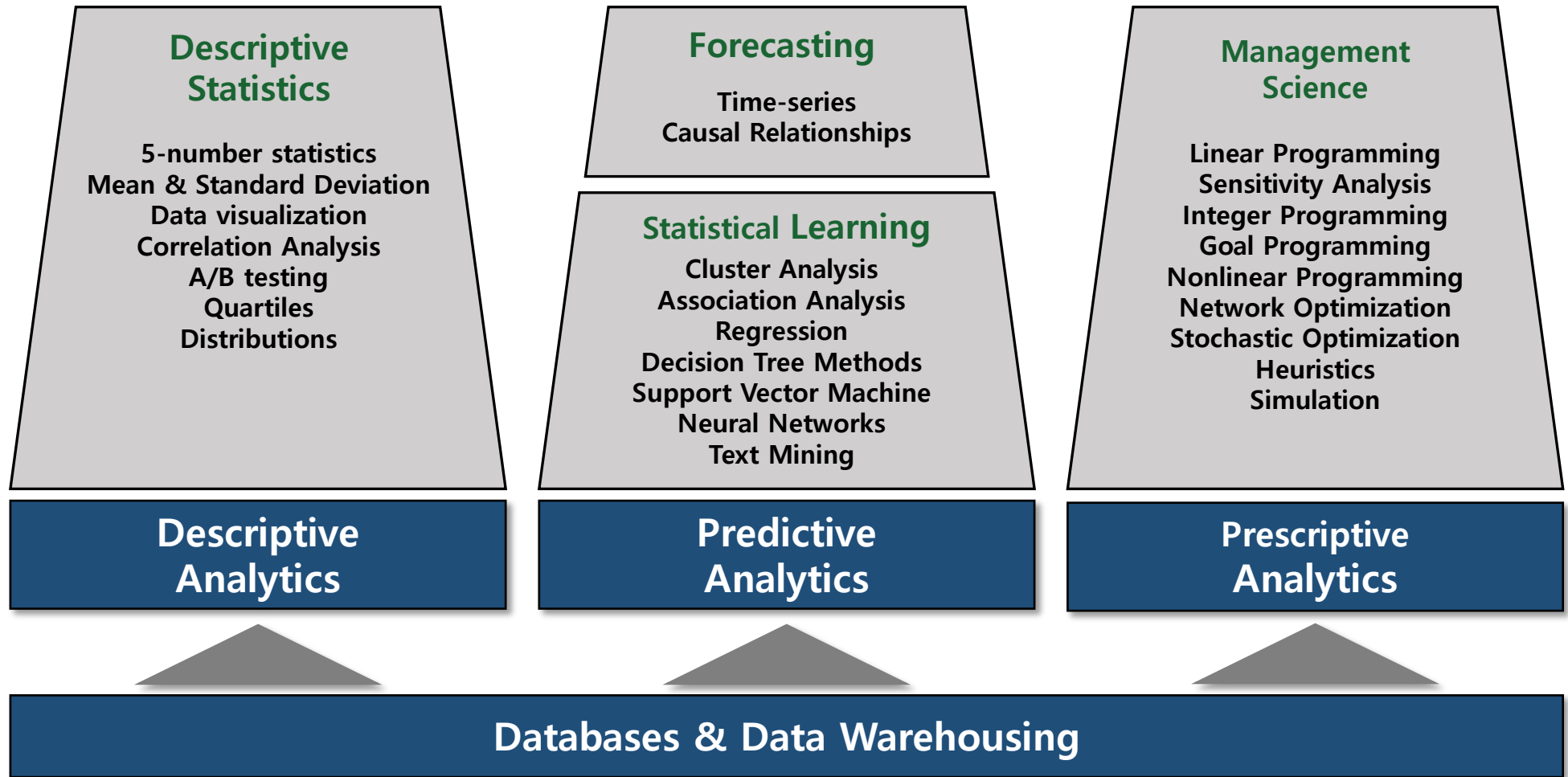
강좌 커리큘럼

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 1-1

학습 내용
Preview

What is Business Analytics?



Source : Asllani(2015). Business Analytics with Management Science Models and Methods

우리가 배울 내용은 ?

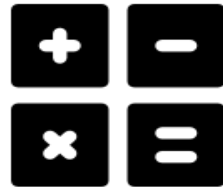
Sourcing & Data Preparation



“ 어떤 자료를 쓸 것인가? ”

- ✓ Data 발생 주체
- ✓ DB 확인
- ✓ Data 수집
- ✓ Data 전처리

Modeling & Analysis



“ 당면 문제를 어떻게 풀어낼 것인가? ”

- ✓ 분석목표 선정
- ✓ 가용 Data 확인
- ✓ 분석 알고리즘 선정
- ✓ 최적 모델 도출

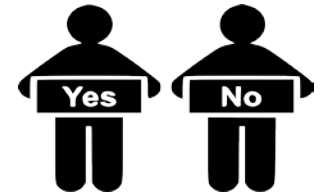
Create report & Dashboard



“ 분석결과를 어떻게 표현할 것인가? ”

- ✓ 결과 Visualization
- ✓ Report 작성
ex) PPT, Markdown
- ✓ Dashboard 구성

Supporting Decision Making



“ 어떤 의사결정을 지지할 것인가? ”

- ✓ Best Decision 도출
- ✓ 예상결과 제시

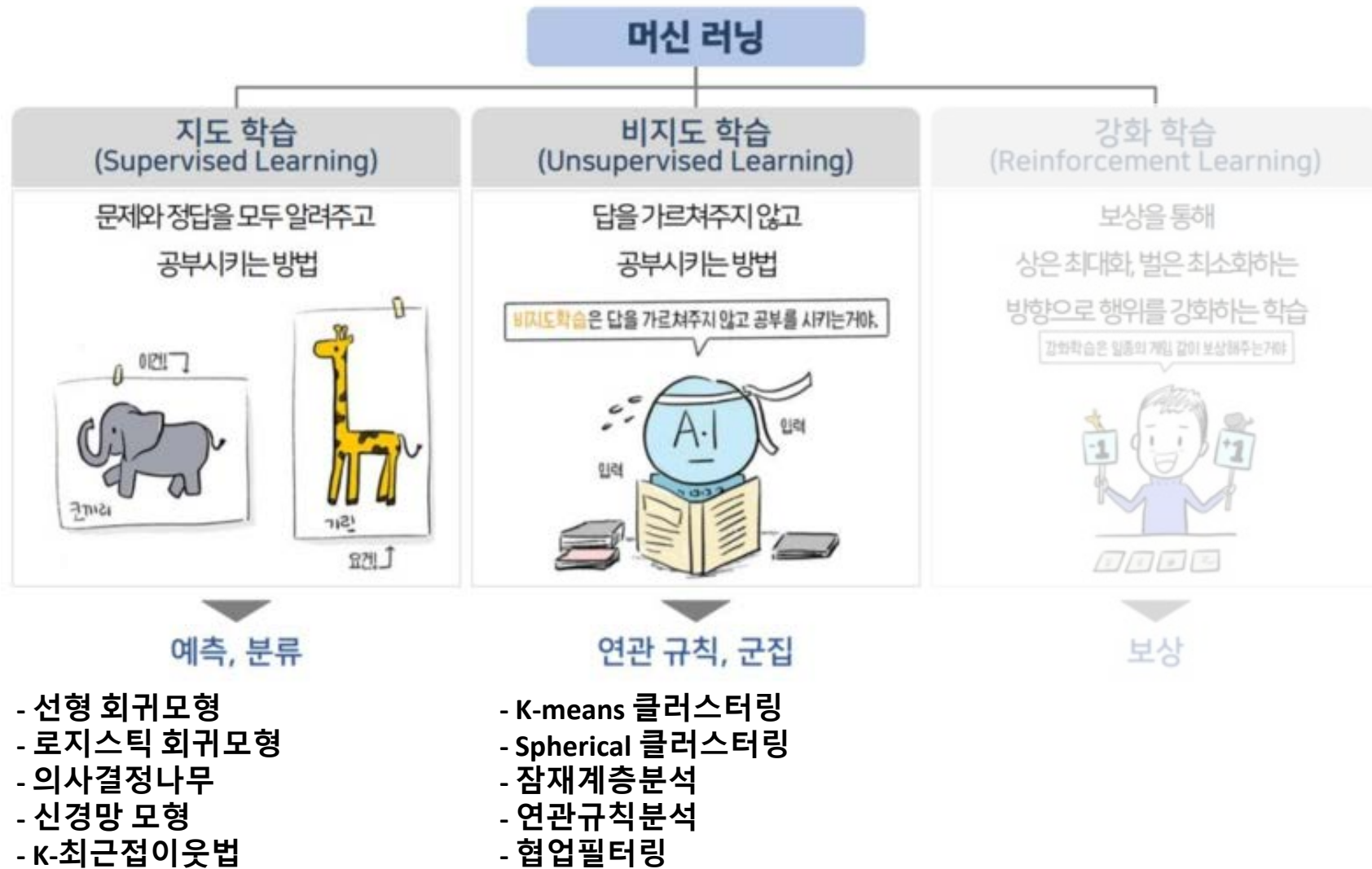
“Fitting”

“Predicting”

기술(Description)	설명(Explanation)	예측(Explanation)
<ul style="list-style-type: none">▪ 데이터가 지니고 있는 특성을 요약, 정리, 및 시각화▪ 주로 데이터의 분포 특성을 나타낼 때 “기술한다”라고 표현▪ 주로 기술통계분석을 이용해 데이터를 기술할 수 있음	<ul style="list-style-type: none">▪ 데이터 간 관계를 정의하고, 그 관계의 “방향성”이나 “정도”에 대한 정보를 제공할 때, “설명한다”라고 표현▪ 주로 상관관계, 인과관계 등을 표현할 때 현상 간 관계의 “설명”이 이뤄짐▪ 현상에 대한 논리적 과정이 더욱 중요	<ul style="list-style-type: none">▪ 주어진 데이터로부터 신뢰할 수 있는 “모형”을 만들어 불확실한 미래를 예측▪ 과정보다는 “결과”가 더 중요

대부분의 사회과학 실증연구(Empirical study)가 “설명”에 초점을 맞추기 때문에
현업과의 연계성이 떨어질 수 밖에 없음

지도학습 ? 비지도학습 ?



Source : <https://m.blog.naver.com/k0sm0s1/221863569856>

우리가 쓸 데이터는 ?

☰

kaggle

🏠 Home

🏆 Compete

📊 Data

📓 Notebooks

🗨 Communities

🎓 Courses

⌵ More

🔍 Search

Sign In

Register

Datasets

Explore, analyze, and share quality data. [Learn more](#) about data, types creating and collaborating.

+ New Dataset

Your Work

🔍 Search datasets

Filters

Datasets

Tasks

Computer Science

Education

Classification

Computer Vision

NLP

Data Visualization

🔥 Trending Datasets

See All

NETFLIX

Netflix Movies and TV Shows

Shivam Bansal · Updated a month ago

Usability 10.0 · 1 MB

5 Tasks · 1 File (CSV)

3844

COVID-19 World Vaccination Progress

Gabriel Preda · Updated 6 hours ago

Usability 10.0 · 111 KB

3 Tasks · 1 File (CSV)

946

Restaurant Business Rankings 2020

Michal Bogacz · Updated a month ago

Usability 10.0 · 16 KB

2 Tasks · 3 Files (CSV)

158

HR Analytics: Job Change of Data Scientists

Möbius · Updated 3 months ago

Usability 10.0 · 295 KB

2 Tasks · 3 Files (CSV)

632

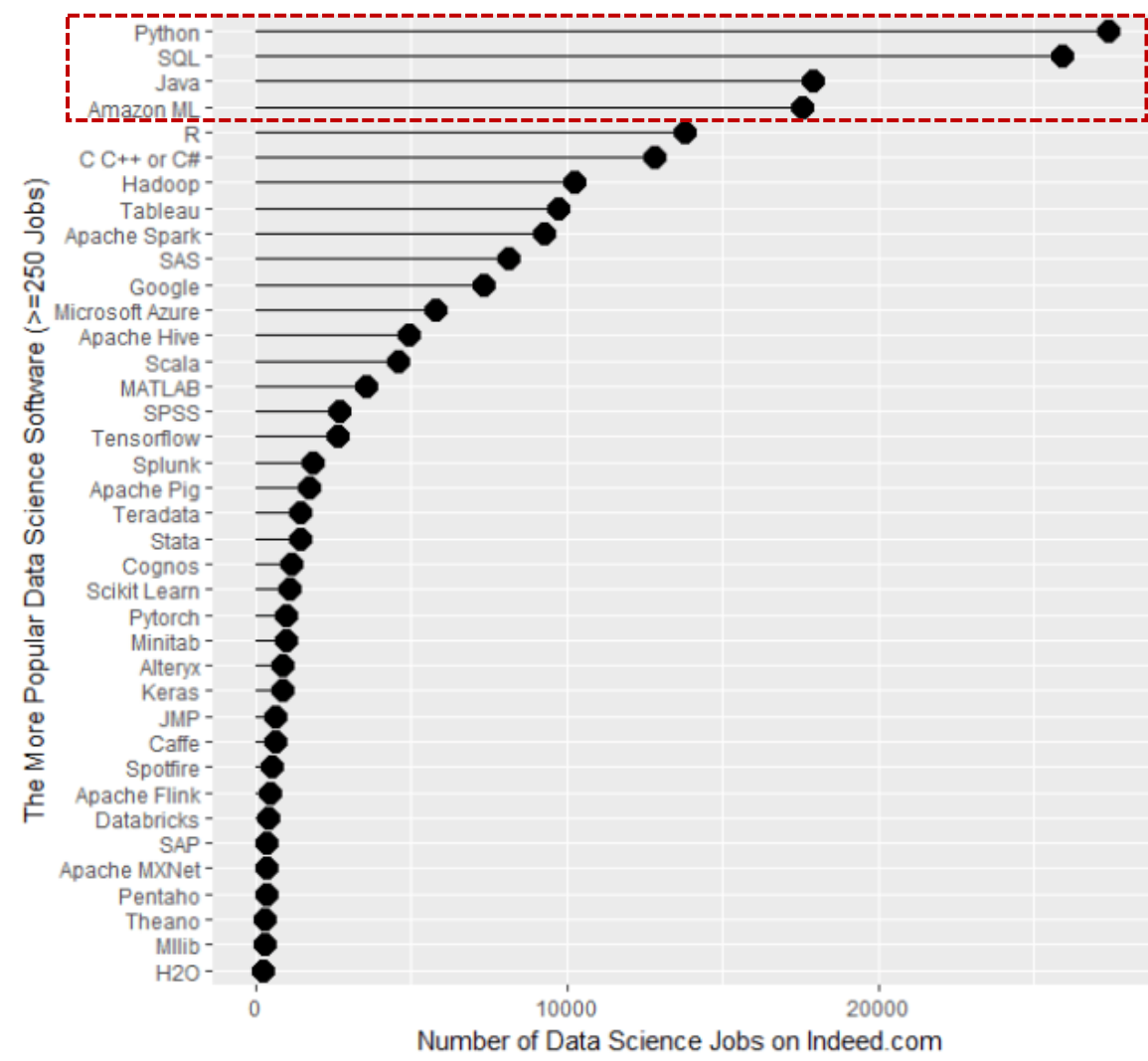
Fall, 2021 Business Analytics (CSFSM7053-00)

11

Lecture 1-2

왜 “R” 인가?

What is R?



다소 진입장벽이 있는 Tool

Source : <https://www.r-bloggers.com/2019/05/data-science-jobs-report-2019-python-way-up-tensorflow-growing-rapidly-r-use-double-sas/>

What is R?

R과 Python은 개발 목적이 다르지만, 최근 경계가 많이 허물어져 있으나, 여전히 데이터를 다루는 데 있어서는 R이 접근성과 활용도가 높음



Features	R	Python
Scope	Used mainly for statistical modeling	Used for a variety of purposes like web-application development and data analysis
Used By	Statisticians, Analyst & Data Scientist	Developer, Data Engineers & Data Scientist
Suitable For	People with no prior experience in programming	Newbies to experienced IT professionals
Package Distribution	CRAN	PyPi
Visualization Tools	ggplot2, plotly, ggiraph	Matplotlib, bokkeh, seaborn

Source : <https://data-flair.training/blogs/r-vs-python-for-data-science/>

What is R?



“**R**은 **통계 계산**과 **그래픽**을 위한 프로그래밍 언어이자
소프트웨어 환경이다” - Wikipedia



계산기
CALCULATOR



소프트웨어 환경
SOFTWARE ENVIRONMENT



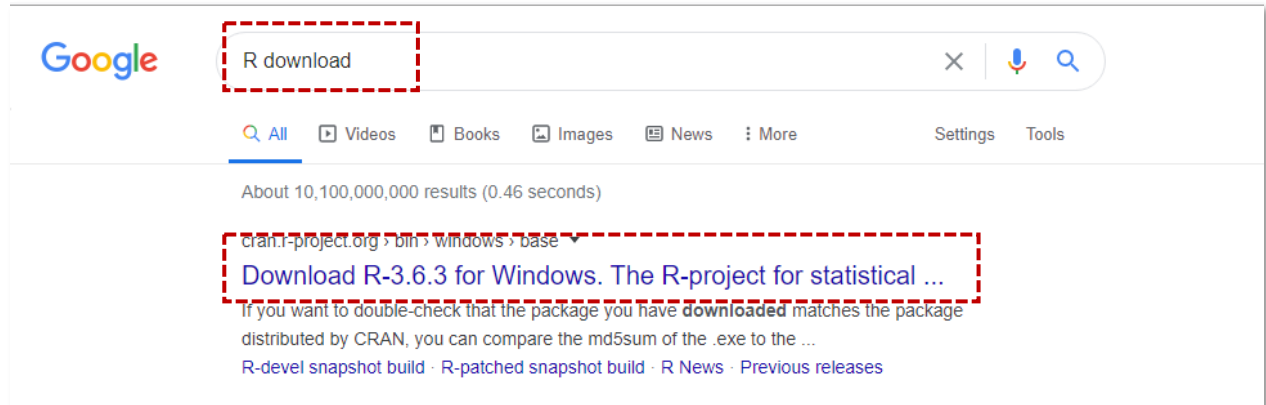
그래픽 도구
GRAPHIC TOOL

실 습 준 비

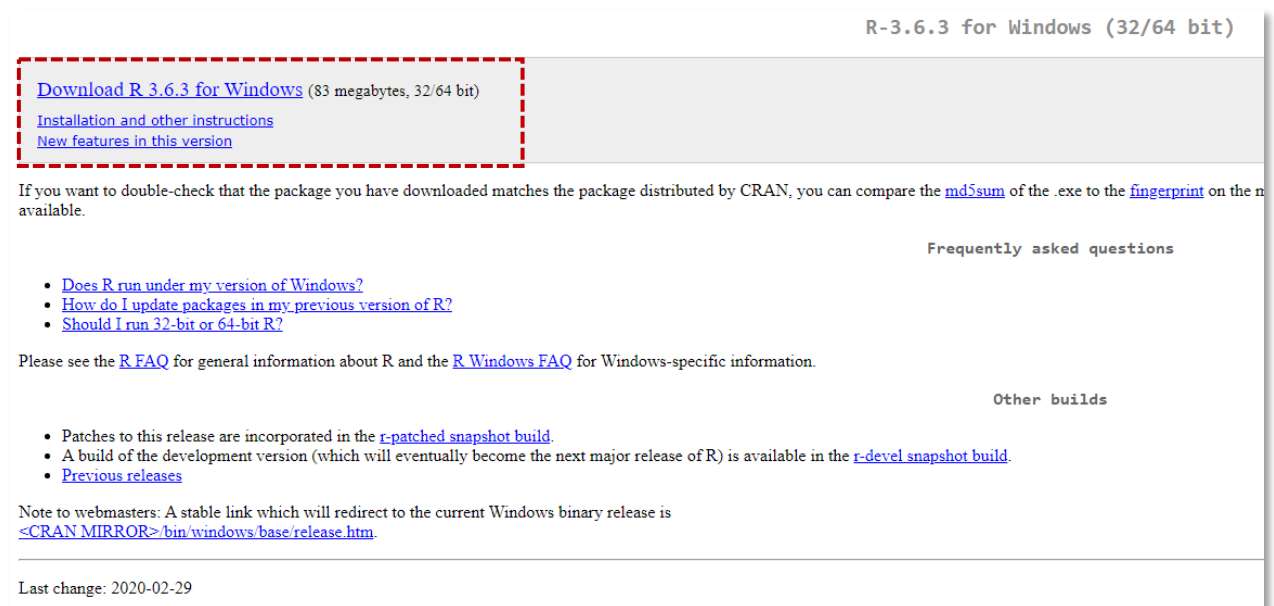
How to Install “R”

<https://cran.r-project.org/>

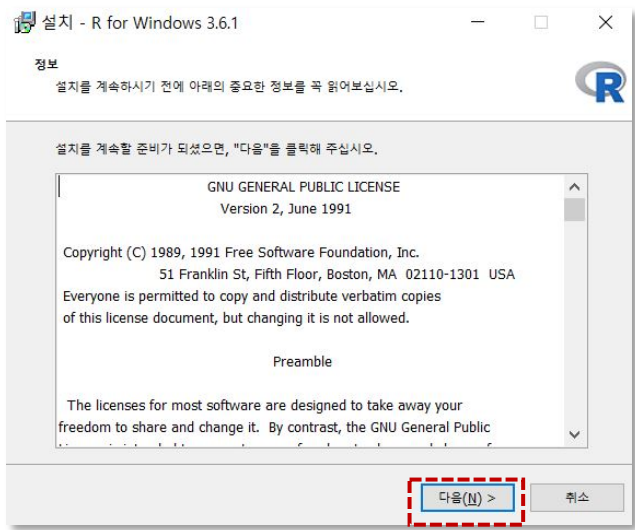
1. Cran 링크를 따라가서
설치하거나 Google에
검색하면 쉽게 링크를 찾을
수 있음

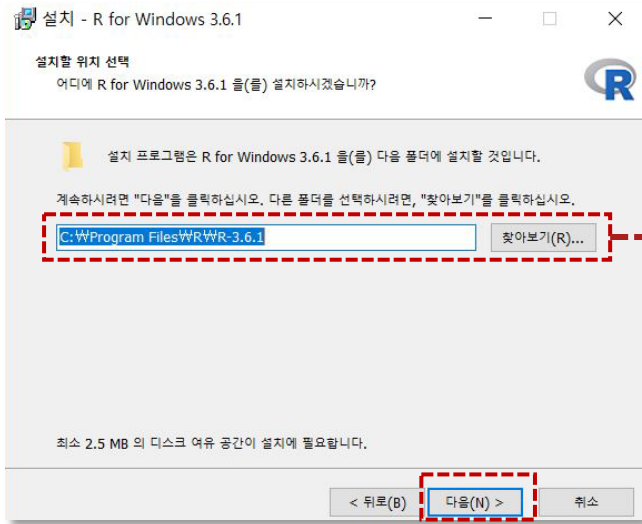


2. Download 링크를 통해
파일을 다운받음

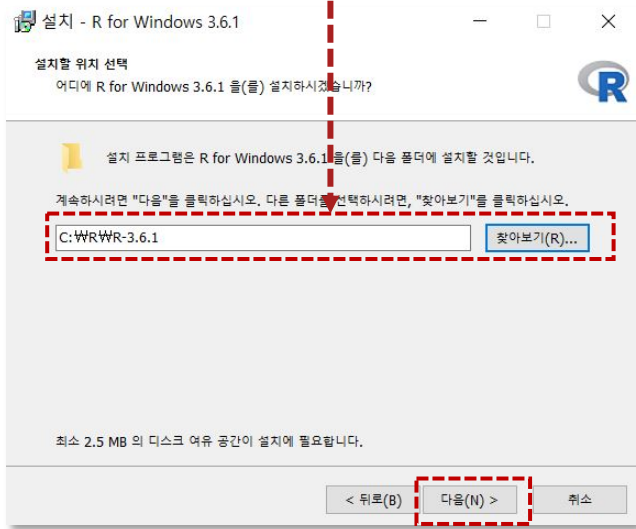


3. 다운받은 파일 실행 후 설치 디렉토리를 변경하고, 다음으로 넘어간다.

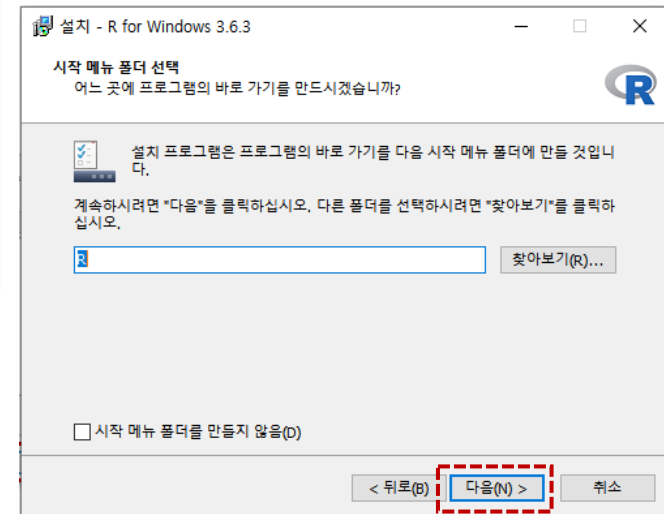
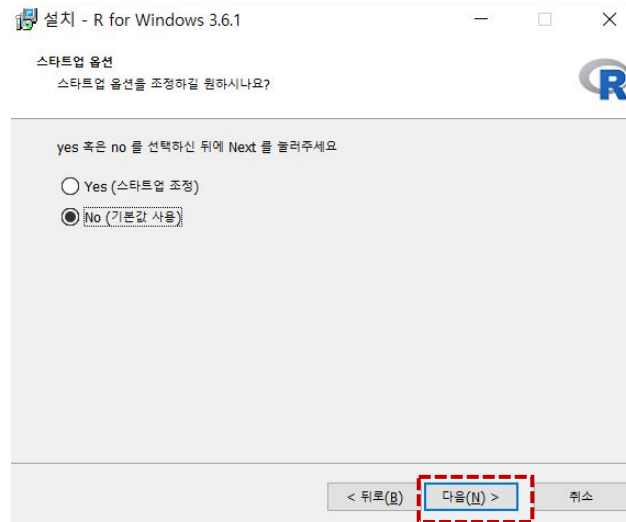
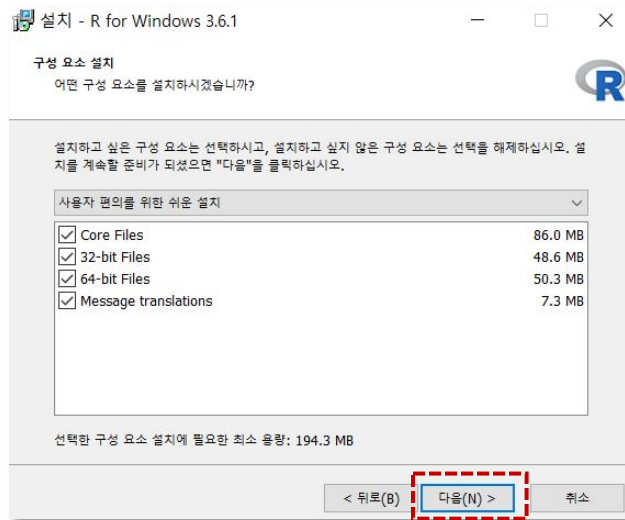




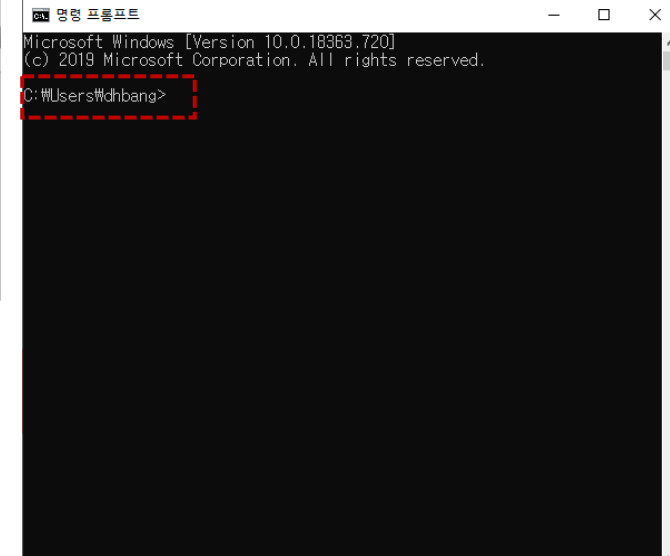
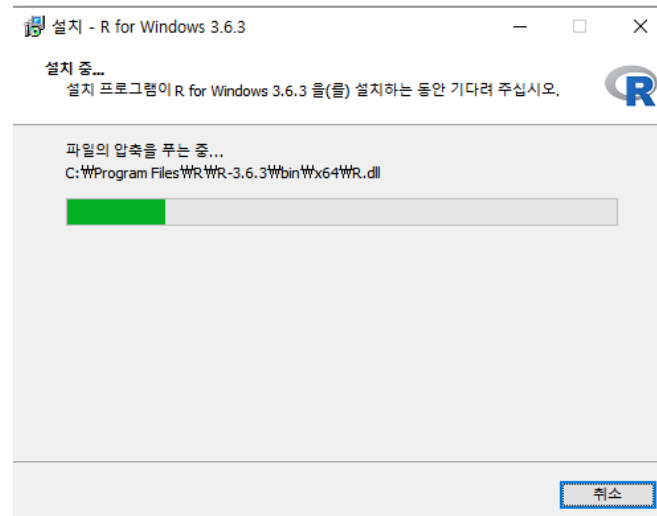
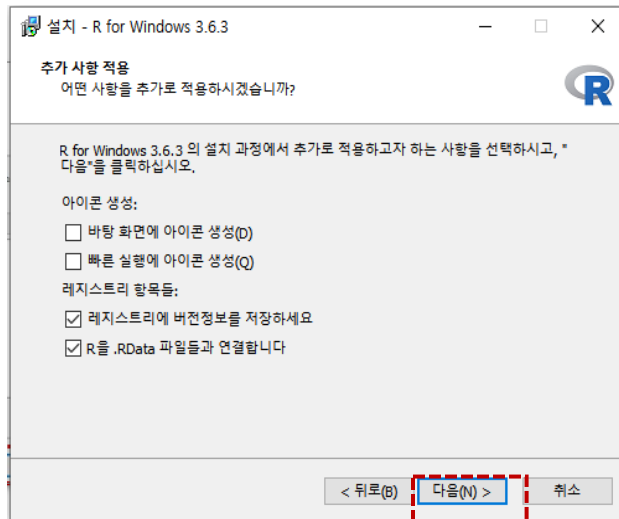
“Program Files”에는 공백(띄어쓰기)가 있어 프로그래밍 시 오류가 발생할 가능성이 있으므로 c:\로 바꿔줌



4. 별도의 설정 변경 없이 “다음” 실행



4. 별도의 설정 변경 없이 “다음” 실행

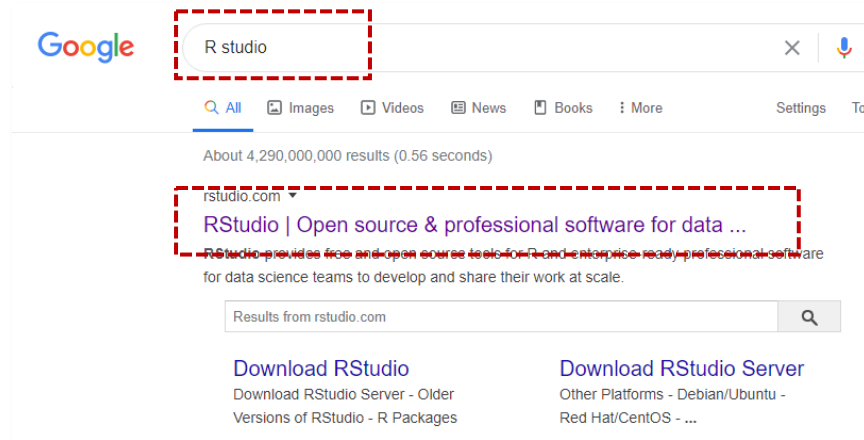


★주의사항

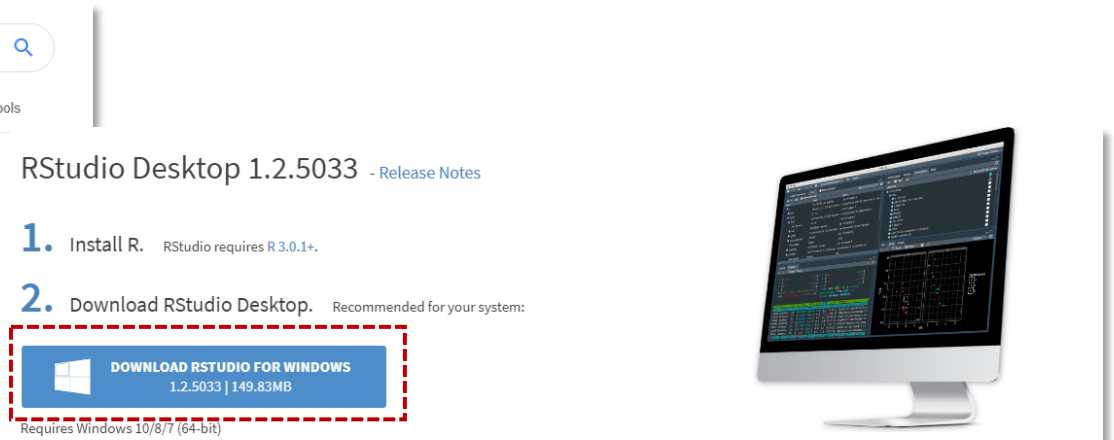
- 윈도우 계정명이 한글인 경우 영어로 변경하기
ex) “C:\홍길동\Document” -> “C:\Gildong\Document”

How to Install “R STUDIO”

1. 마찬가지로 구글에서 검색, 다운로드 한다.




Google search results for "R studio". The search bar contains "R studio". The first result is from rstudio.com, titled "RStudio | Open source & professional software for data ...". Below the title, it says "RStudio provides free and open source tools for R and enterprise-ready, professional software for data science teams to develop and share their work at scale." There are two download links: "Download RStudio" and "Download RStudio Server".



RStudio Desktop 1.2.5033 - Release Notes

1. Install R. RStudio requires R 3.0.1+.
2. Download RStudio Desktop. Recommended for your system:

DOWNLOAD RSTUDIO FOR WINDOWS
1.2.5033 | 149.83MB
Requires Windows 10/8/7 (64-bit)



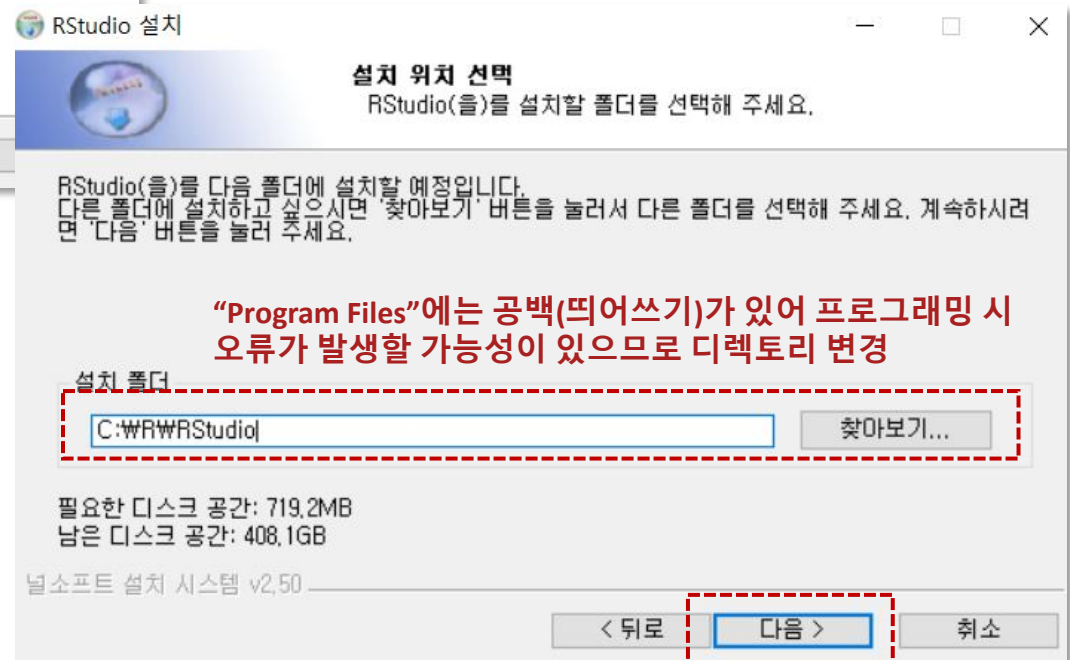
All Installers

Linux users may need to [import RStudio's public code-signing key](#) prior to installation, depending on the operating system's security policy.

RStudio 1.2 requires a 64-bit operating system. If you are on a 32 bit system, you can use an [older version of RStudio](#).

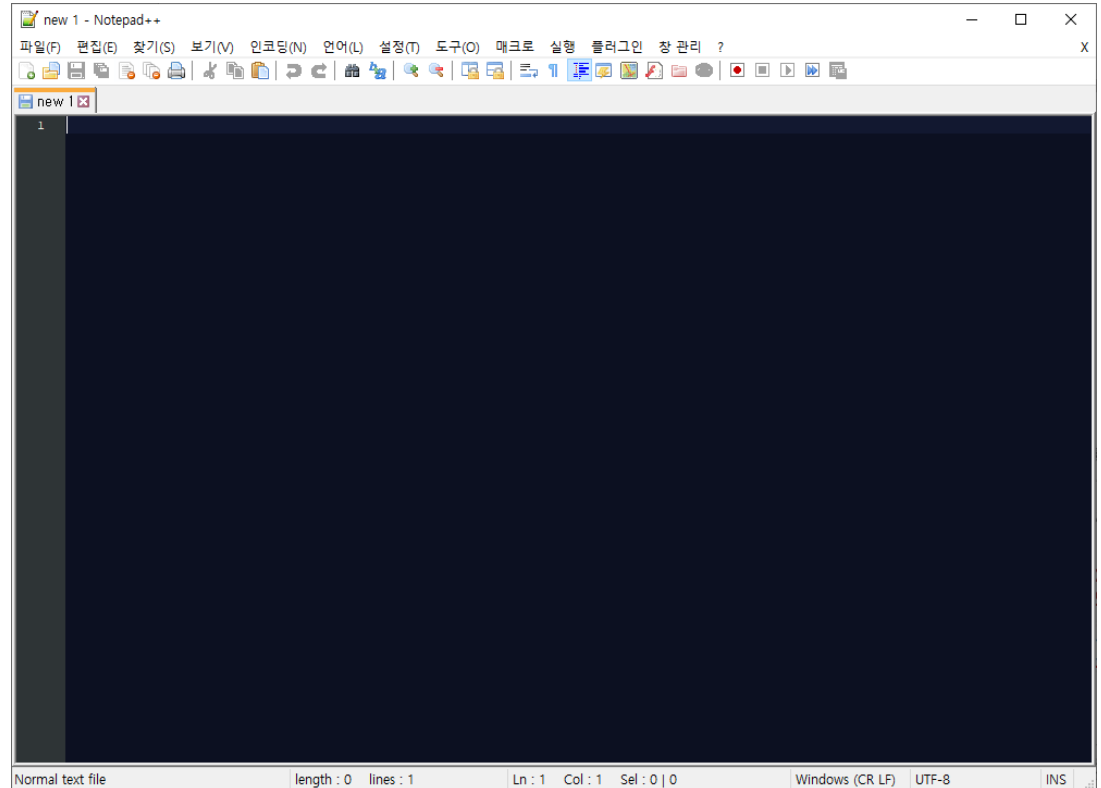
OS	Download	Size	SHA-256
Windows 10/8/7	RStudio-1.2.5033.exe	149.83 MB	7fd3bc1b
macOS 10.13+	RStudio-1.2.5033.dmg	126.89 MB	b67c9875
Ubuntu 14/Debian 8	rstudio-1.2.5033-amd64.deb	96.18 MB	89dc2e22
Ubuntu 16	rstudio-1.2.5033-amd64.deb	104.14 MB	a1591ed7
Ubuntu 18/Debian 10	rstudio-1.2.5033-amd64.deb	105.21 MB	08eaa295
Fedora 19/Red Hat 7	rstudio-1.2.5033-x86_64.rpm	120.23 MB	38cf43c6

How to Install “R STUDIO”



노트패드 ++

노트패드는 다양한 타입의 데이터 관리가 가능한 데이터 편집기이자 동시에 간단한 코드 에디팅도 가능함



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

➤ R script(.r)

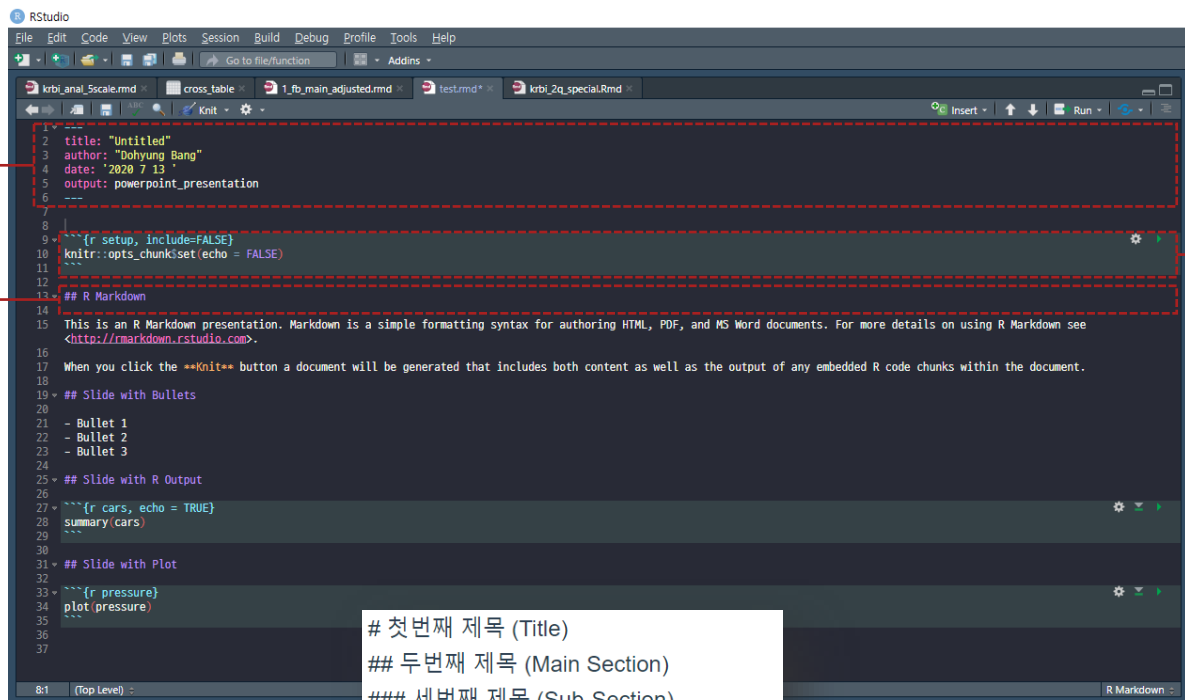
- R의 코드를 실행 및 작성하기 위한 가장 기본적인 코드 작성 단위
- 주석 처리 '#'를 해주지 않으면 모든 Line을 코드로 인식함

➤ 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 문서처럼 작성이 가능함
- 코드에 대한 주석, 상호작용을 위해 코드북은 Markdown을 주로 활용할 예정

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.

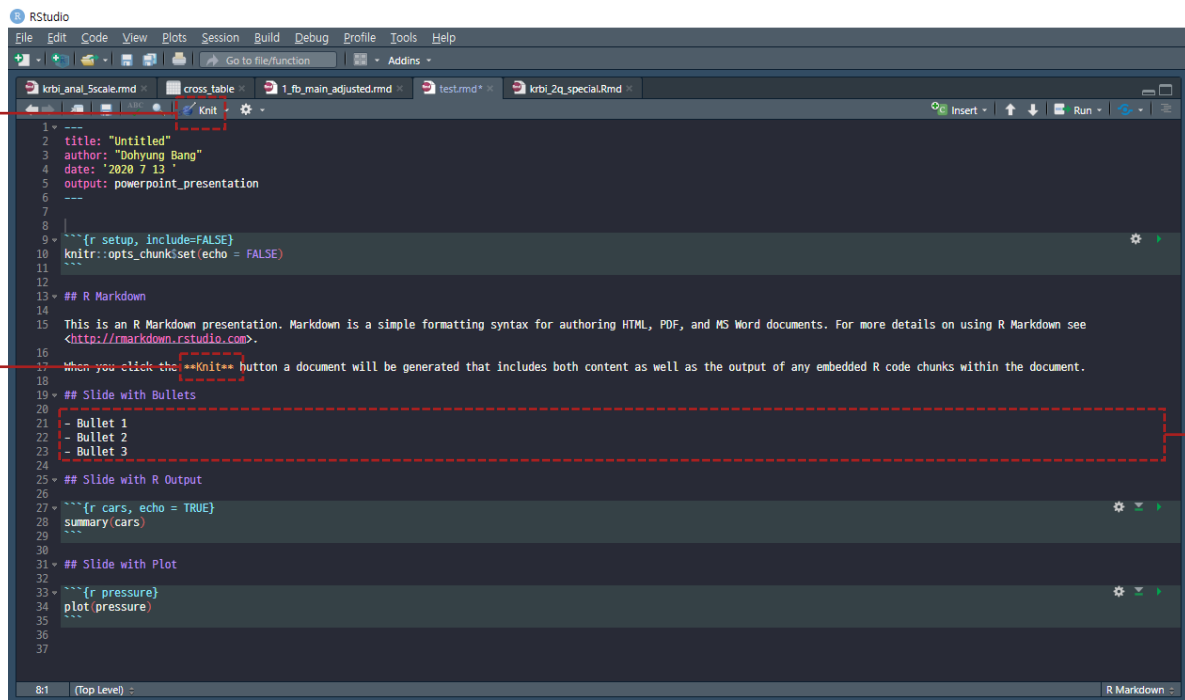
- eval = FALSE / TRUE
 - 코드를 실행하지 않는다. / 실행한다.
- echo = FALSE / TRUE
 - 코드를 보여주지 않는다. / 보여준다.
- Include = FALSE / TRUE
 - 실행 결과를 보여주지 않는다.
- message = FALSE / TRUE
 - 실행 때 나오는 메시지를 보여주지 않는다.
- warning = FALSE / TRUE
 - 실행 때 나오는 경고를 보여주지 않는다.
- error = TRUE / FALSE
 - 에러가 있어도 실행하고 에러코드를 보여준다.
- fig.height = 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)