

한국한의학연구원, 구본초

통계 프로그래밍 언어



Contents

| | |
|-------------------------------|----------|
| List of Tables | v |
| List of Figures | vii |
| Course Overview | ix |
| I Get Started | 1 |
| 1 Introduction | 3 |
| 1.1 R 설치하기 | 4 |
| 1.2 R 시작 및 작동 체크 | 14 |
| 1.3 R script 편집기 사용 | 17 |
| 1.4 RStudio | 20 |
| 1.4.1 RStudio 설치하기 | 20 |



List of Tables

| | |
|---------------------------------|-----|
| 0.1 강의 계획표 | xii |
| 1.1 R help 관련 명령어 리스트 | 17 |



List of Figures

| | | |
|-----|--|----|
| 1.1 | Windows에서 R 실행화면(콘솔 창, SDI 모드) | 13 |
| 1.2 | 정규분포 100개의 히스토그램 | 16 |
| 1.3 | cars 데이터셋의 speed와 dist 간 2차원 산점도: speed는 자동차 속도(mph)이고 dist는 해당 속도에서 브레이크를 밟았을 때 멈출 때 까지 걸린 거리(ft)를 나타냄. | 19 |



Course Overview



본 문서는 2020년도 1학기 정보통계학과에서 개설한 “통계 프로그래밍 언어” 강의를 위해 개발한 강의 노트이며, Yihui Xie가 개발한 **bookdown** 패키지 (Xie, 2019)를 활용하여 생성한 문서이고 Google Chrome 또는 Firefox 브라우저에 최적화 됨. 아울러 충남대학교 정보통계학과 이상인 교수님의 2019년도 2학기 “통계패키지활용” 강의 노트와 동국대학교 ICT빅데이터 학부 김진석 교수님의 R 프로그래밍 및 실습¹ 강의 자료 내용을 본 강의노트 작성에 참고함.

본 강의 노트는 주 단위로 업데이트될 예정이며, <https://zorba78.github.io/cnu-r-programming-lecture-note/>에서 확인할 수 있고, 해당 페이지에서 pdf 파일 다운로드가 가능함.

강의소개

R은 뉴질랜드 오클랜드 대학의 Robert Gentleman 과 Ross Ihaka 가 AT&T 벨 연구소에서 개발한 S 언어를 기반으로 개발한 GNU 환경의 통계 계산 및 프로그래밍 언어이다. 현재 R 소프트웨어는 통계학 뿐 아니라 데이터 과학을 포함한 의학, 생물학 등 다양한 분야에서 활용되고 있으며 특히 통계 소프트웨어 개발과 데이터 분석에 많이 활용되고 있다. 본 강의는 데이터 분석을 위한 R의 기초 문법과 통계학 입문에서 학습한 몇 가지 중요한 통계적 이론에 대한 시뮬레이션 방법을 다룬다. 아울러 R package를 활용한 데이터 핸들링 및 시각화

그리고 Rmarkdown을 활용한 재현가능(reproducible)한 문서 작성법에 대해 학습하고자 한다.

교과 목표

- R 기초 문법 습득
- R package를 활용한 데이터 핸들링 및 자료 시각화
- R 시뮬레이션을 통한 통계학 기초 이론 확인
- R을 이용한 데이터 분석 실습
- R markdown을 이용한 재현가능(reproducible)한 보고서 작성 방법 습득

선수과목

통계학 개론

수업 방법

- 강의: 50 %
- 실험/실습: 50%

평가방법

- 중간고사: 40 %
- 기말고사: 40 %
- 출석: 10 %
- 과제: 10 %

수업 규정

- 3번 지각은 1번 결석으로 처리
- 특별한 사유 없이 수업 중간에 이탈한 경우 결석으로 처리
- 특별한 사유로 인해 결석 또는 지각을 할 경우 사유를 증빙할 수 있는 서류 제출 시 출석으로 인정
- 출결 미달, 중간 또는 기말고사 미 응시인 경우 F 학점으로 처리
- 수업 중 휴대폰 및 각종 모바일 기기 사용 금지

교재 및 참고문헌

별도의 교재 없이 본 강의 노트로 수업을 진행할 예정이며, 수업의 이해도 향상을 위해 아래 소개할 도서 및 웹 문서 등을 참고할 것을 권장함.

참고 자료

- 실리콘밸리 데이터과학자가 알려주는 따라하며 배우는 데이터 과학 ([권재명, 2017](#))
- R을 이용한 데이터 처리&분석 ([서민구, 2014](#))
- R 그래픽스 ([유충현 et al., 2005](#))
- ggplot2: elegant graphics for data analysis² ([Wickham, 2016](#))
- R for data science³ ([Wickham and Grolemund, 2016](#))
- Statistical Computing with R ([Rizzo, 2019](#))

²<https://ggplot2-book.org/>

³<https://r4ds.had.co.nz/>

강의 계획

TABLE 0.1: 강의 계획표

| 주차 | 강의 내용 | 과제 |
|---------|--|------|
| Week 1 | R 소개, R/R Studio 설치, R 패키지 설치, R 맛보기 및 markdown 문서 만들기 | |
| Week 2 | R 자료형: 스칼라, 벡터, 리스트 | |
| Week 3 | R 자료형: 행렬 및 배열 | 과제 2 |
| Week 4 | R 자료형: 팩터, 테이블, 데이터 프레임 | |
| Week 5 | R 자료형: 문자열과 정규 표현식 | 과제 3 |
| Week 6 | 데이터 프레임 가공 및 시각화 I | |
| Week 7 | 데이터 프레임 가공 및 시각화 II | 과제 4 |
| Week 8 | 중간고사 | |
| Week 9 | 데이터 프레임 가공 및 시각화 III | |
| Week 10 | R 프로그래밍: 조건문, 반복문, 함수 | 과제 5 |
| Week 11 | 통계시뮬레이션 I: 표본분포 및 중심극한정리 | |
| Week 12 | 통계시뮬레이션 2: 신뢰구간과 가설검정 | 과제 6 |
| Week 13 | R을 이용한 기초통계 분석 | |
| Week 14 | R markdown 활용 | 과제 7 |
| Week 15 | 기말고사 | |

Part I

Get Started



1

Introduction

1. R 프로그램

- 데이터 분석을 위한 자료 전처리, 통계 및 시각화를 지원하는 컴퓨터 언어 및 환경
- 1980년 AT&T 벨 연구소의 John Chambers가 개발한 S 언어를 기반으로 1995년 뉴질랜드 Auckland 대학의 통계학과 교수 Robert Gentleman과 Ross Ihaka 가 개발
- GNU¹ 기반의 오픈 소스
- 통계학, 전산학, 생물학, 의학 등 거의 모든 학문분야에서 분석도구로 활용되고 있고, 최근 data science 분야에서 널리 활용

2. R 언어의 특징

- 무료 소프트웨어
- CRAN (Comprehensive R Archive Network)²에서 배포
- 특정 vendor가 아닌 전 세계 연구자들이 개발한 알고리즘 및 최신 함수 활용 가능 (packaging system)
- 범용적으로 사용되는 거의 대부분의 운영체제 (Windows, Mac, Linux)에서 작동 가능

¹https://en.wikipedia.org/wiki/GNU_Project

²<http://cran.r-project.org/web/view>

- 방대한 개발 및 사용 생태계 형성
- 강력한 그래픽 기능



유용한 웹 사이트: R과 관련한 거의 모든 문제는 Googling (구글을 이용한 검색)을 통해 해결 가능(검색주제 + “in R” or “in R software”)하고 많은 해답들이 아래 열거한 웹 페이지에 게시되어 있음.

- R 프로그래밍에 대한 Q&A: Stack Overflow³
- R 관련 웹 문서 모음: Rpubs⁴
- R package에 대한 raw source code 제공: Github⁵
- R을 이용한 통계 분석: Statistical tools for high-throughput data analysis (STHDA)⁶

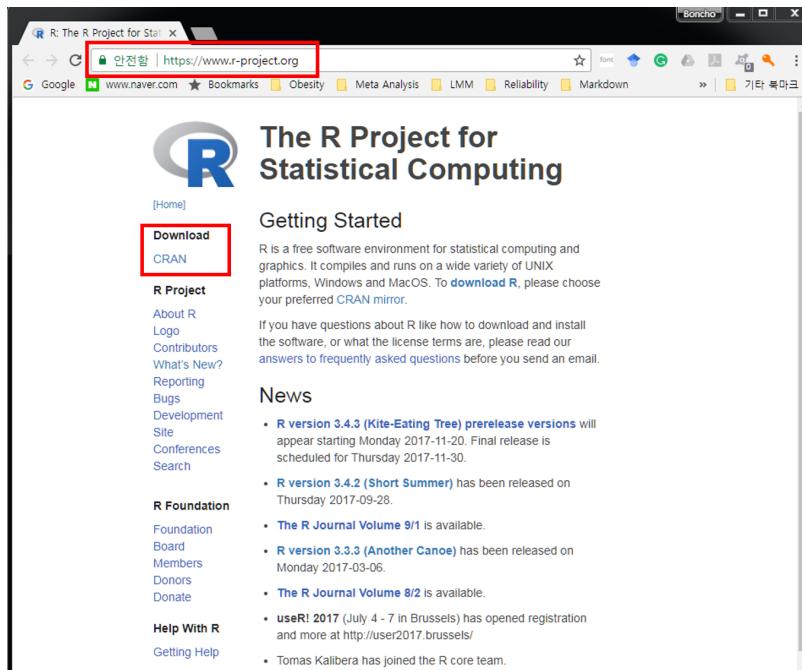
1.1 R 설치하기

R 다운로드 사이트: <https://www.r-project.org> 또는 <https://cran.r-project.org>

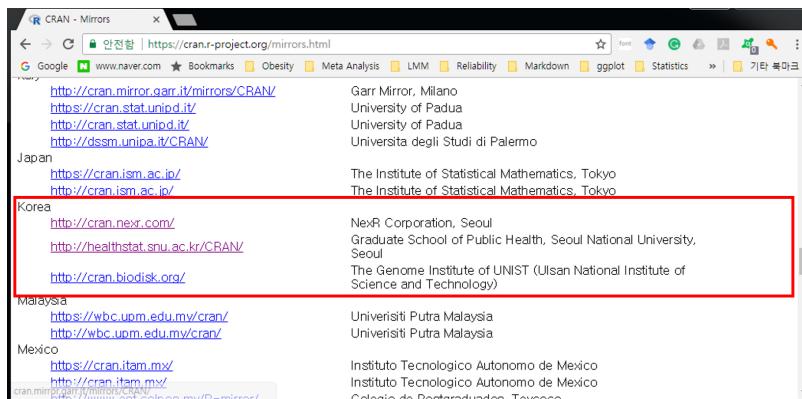
1. 웹 브라우저 (i.e. Explore, Chrome, Firefox 등)의 주소 입력창에 <https://www.r-project.org>
2. 좌측 R Logo 하단 Download 아래 CRAN 클릭

1.1 R 설치하기

5



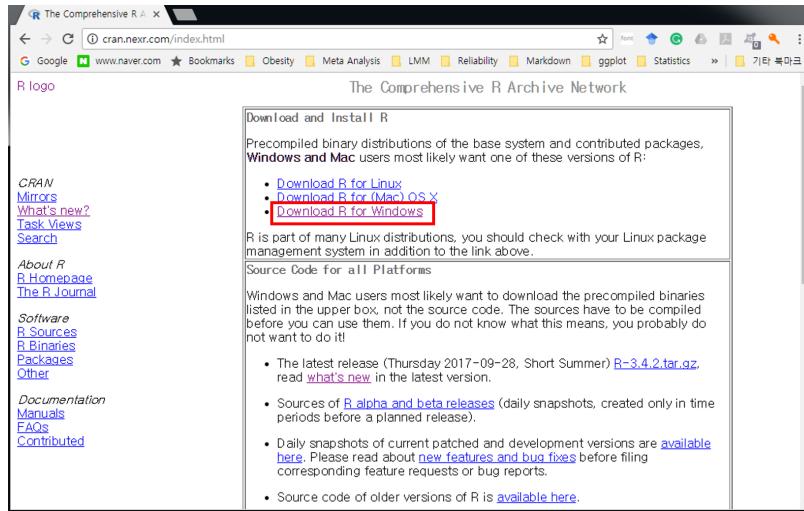
3. 클릭 후 연결한 페이지를 스크롤 후 Korea 아래 링크⁷ 클릭



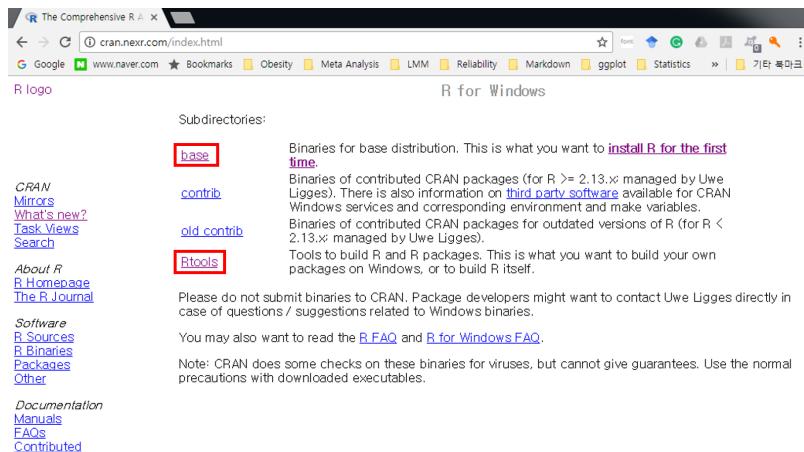
4. 클릭 후 세 가지 운영체제(Linux, Mac OS X, Windows)에 따른 R 버전 선택 가능⁸

⁷ 해당 링크들은 접속 시점에 따라 변경될 수 있음

⁸ 본 노트는 Windows 버전 설치만 다룸



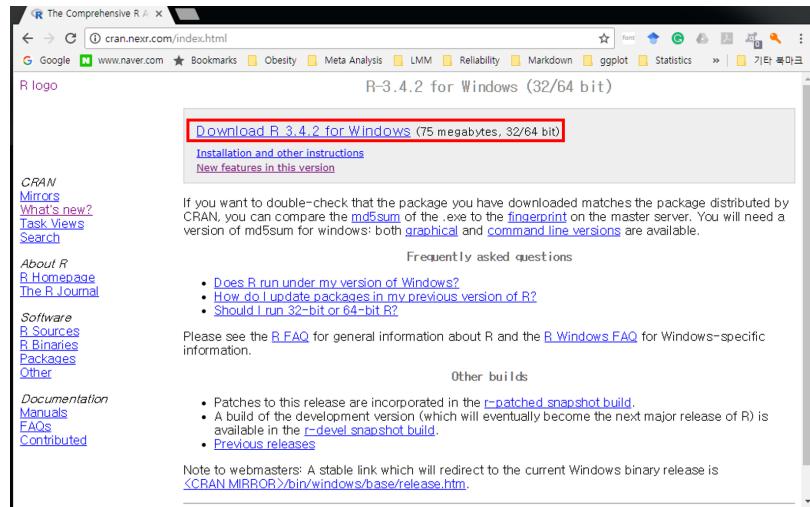
5. Downloads R for Windows 링크 클릭하면 다음과 같은 화면으로 이동



다음 하위폴더에 대한 간략 설명

- **base**: R 실행 프로그램
- **contrib**: R package의 바이너리 파일
- **Rtools**: R package 개발 및 배포를 위한 프로그램

6. 위 화면에서 **base** 링크 클릭 후 아래 화면에서 **Downloads R 3.x.x for Windows** 를 클릭 후 설치 파일을 임의의 디렉토리에 저장 및 실행

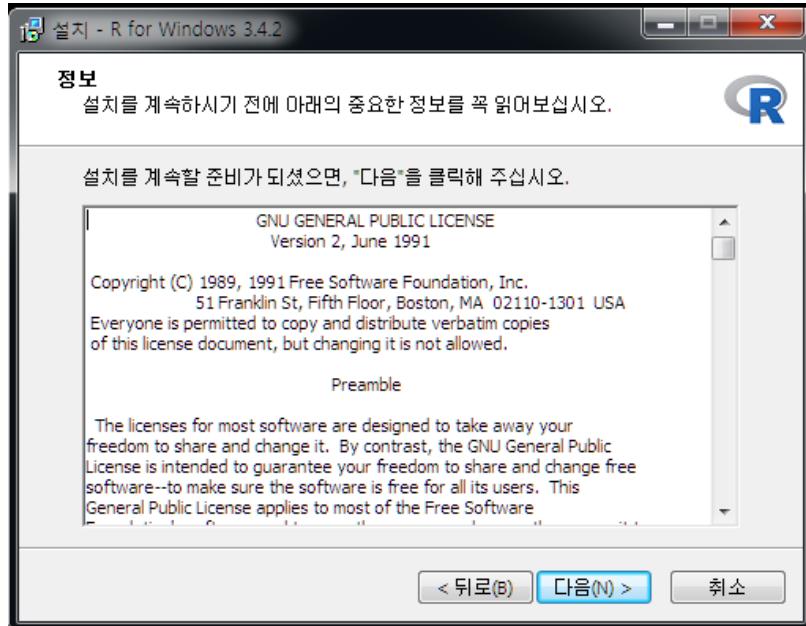


7. 다운로드한 파일을 실행하면 아래와 같은 대화창이 나타남

- 한국어 선택 → 환영 화면에서 [다음(N)>] 클릭

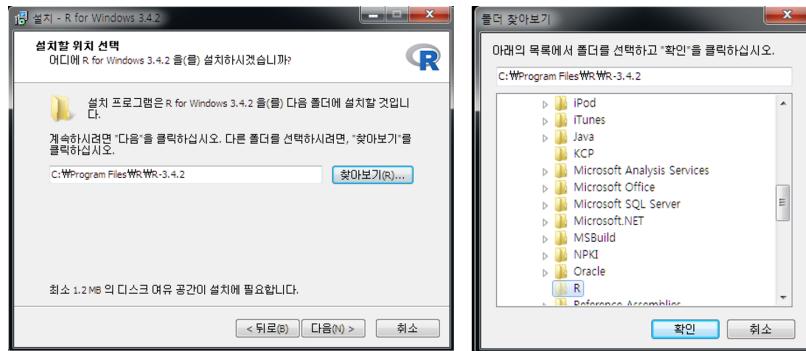


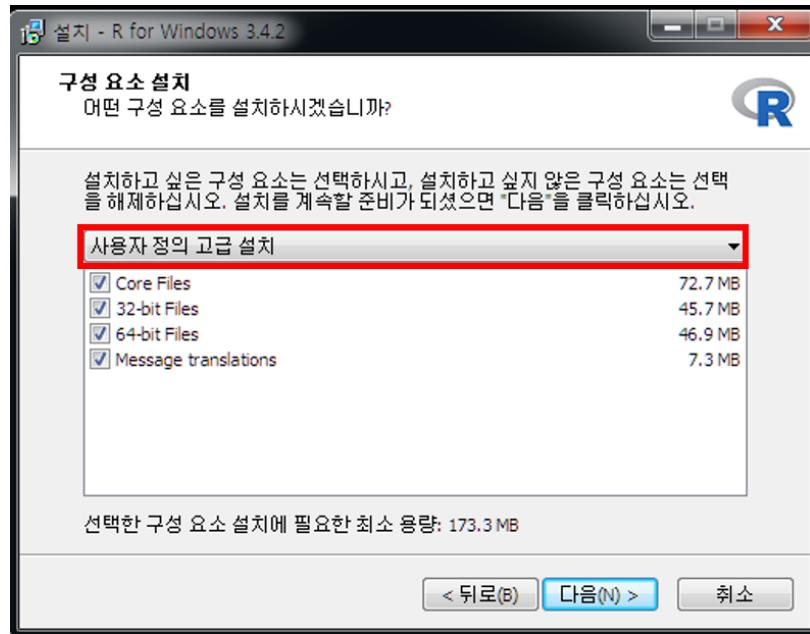
8. GNU 라이센스에 대한 설명 및 동의 여부([다음(N)>]) 클릭



9. 설치 디렉토리 설정 및 구성요소 설치 여부

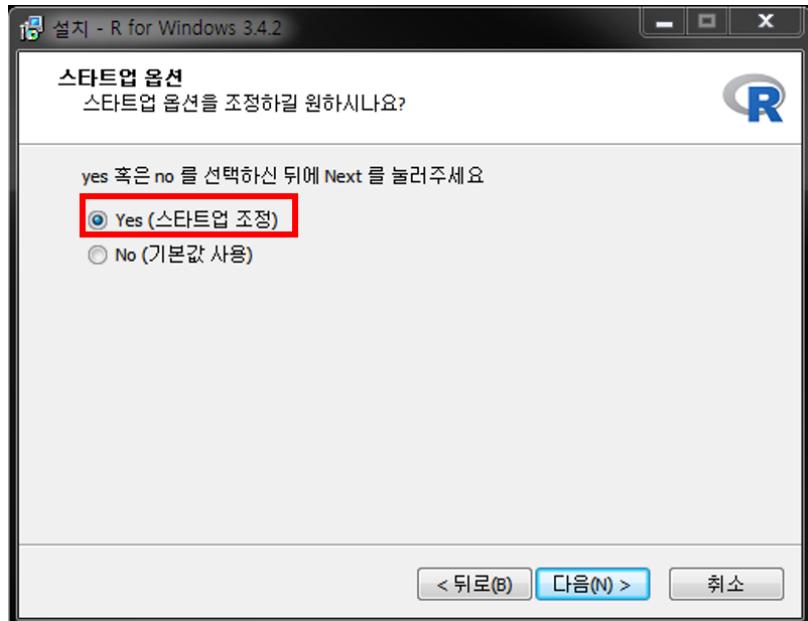
- 원하는 디렉토리 설정 (예: C:\R\R-3.x.x)
 - 기본 프로그램 (“Core Files”), 32 또는 64 bit 용 설치 파일, R console
- 한글 번역 모두 체크 뒤 [다음(N)>] 클릭





10. R 스타트업 옵션 지정

- 기본값("No" check-button)으로도 설치 진행 가능
- 본 문서에서는 스타트업 옵션 변경으로 진행

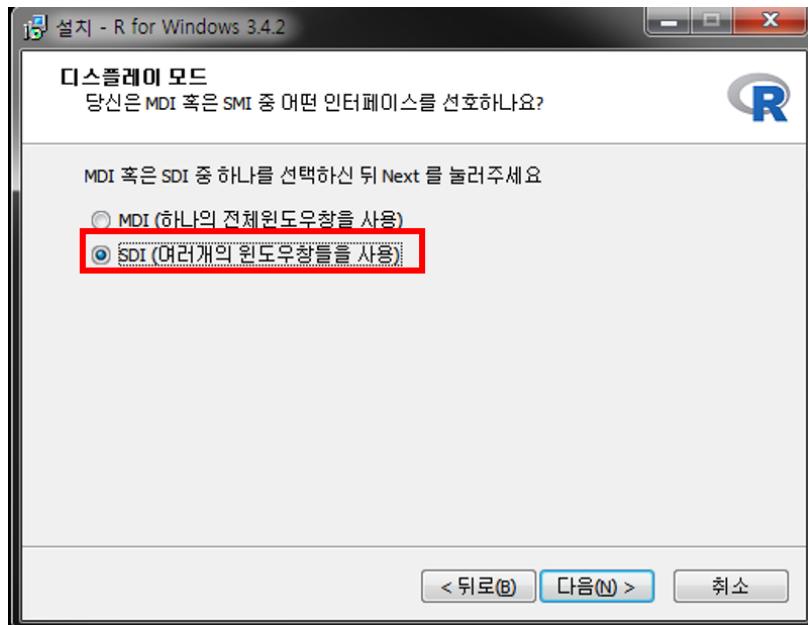


11. 화면표시방식(디스플레이) 모드 설정 변경

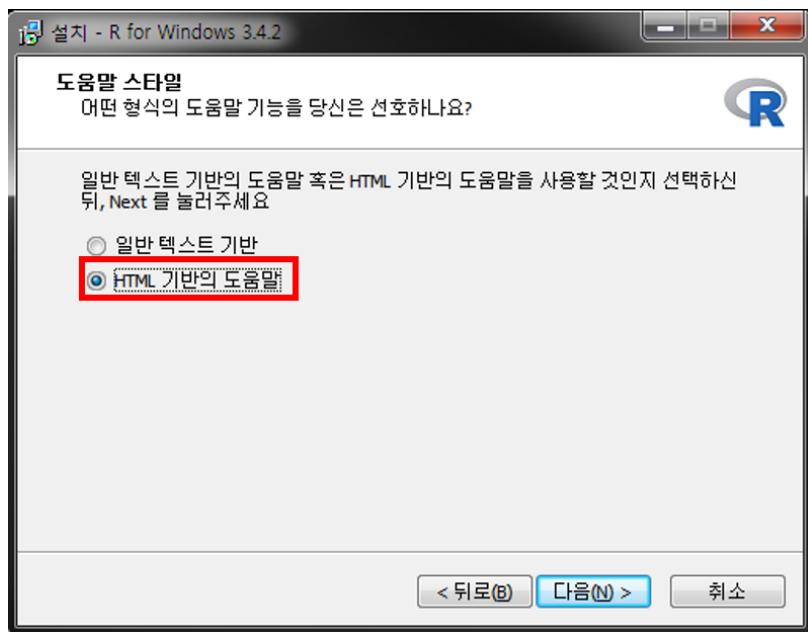
- MDI: 한 윈도우 내에서 script 편집창, 출력, 도움말 창 사용
- SDI: 다중 창에서 각각 script 편집창, 출력, 도움말 등을 독립적으로 열기

1.1 R 설치하기

11

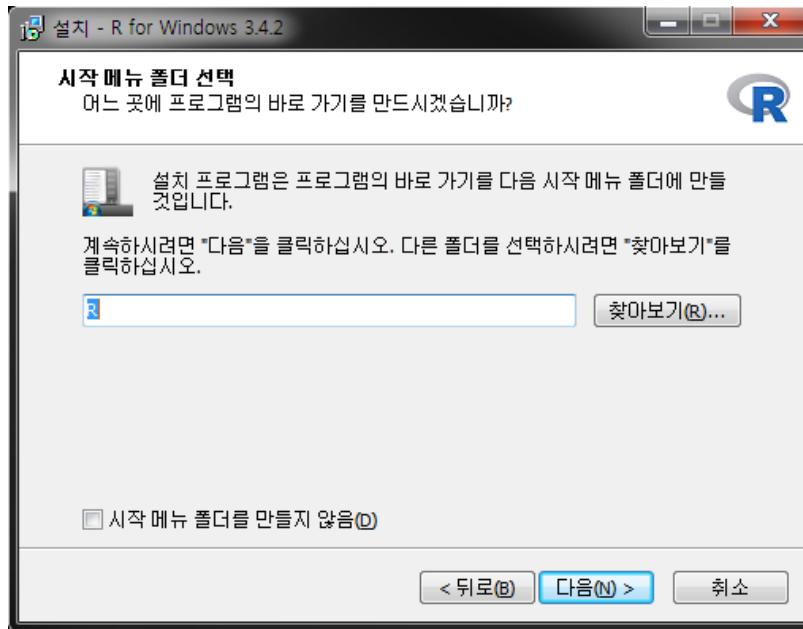


12. 도움말 형식에서 HTML 도움말 기반 선택



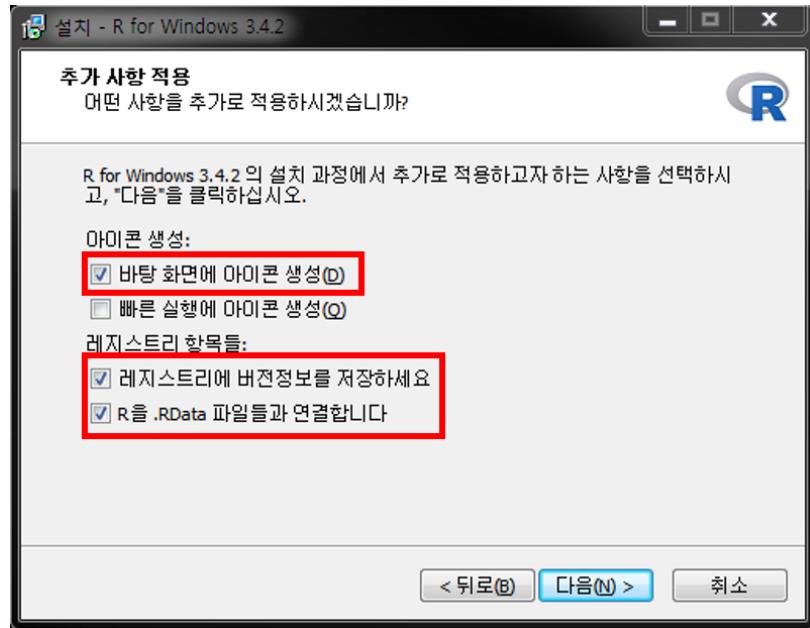
13. 시작메뉴 폴더 선택

- “바로가기”를 생성할 시작 메뉴 폴더 지정 후 [다음(N)>] 클릭 후 설치 진행
- 하단 “시작메뉴 폴더 만들지 않음” 체크박스 표시 시 시작메뉴에 “바로가기” 아이콘이 생성되지 않음(실행에 전혀 지장 없음)



14. 추가 옵션 지정 : 바탕화면 아이콘 생성 등 추가적 작업 옵션 체크 후 [다음(N)>] 클릭 → 설치 진행

- 설치된 R 버전 정보 레지스트리 저장 여부
- .Rdata 확장자를 R 실행파일과 자동 연계



15. 설치 완료 후 바탕화면의 R 아이콘을 더블클릭하면 Rgui가 실행

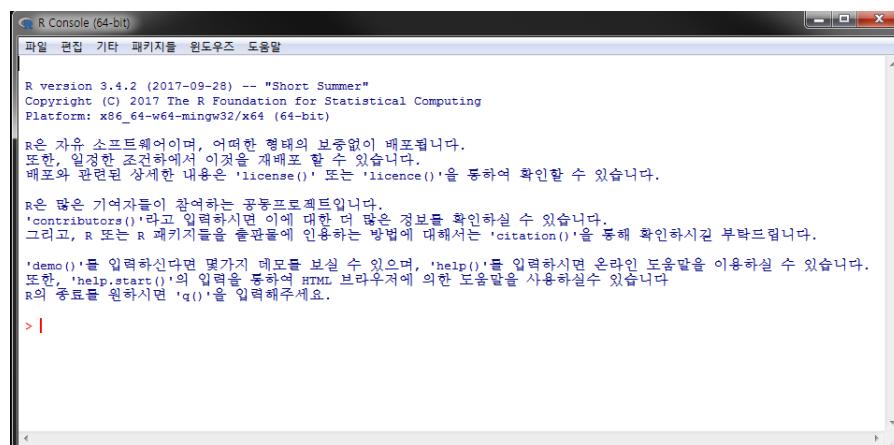


FIGURE 1.1: Windows에서 R 실행화면(콘솔 창, SDI 모드)

1.2 R 시작 및 작동 체크



실습: 설치된 R을 실행 후 보이는 R 콘솔(console) 창에서 명령어를 실행하고 결과 확인

그림 Figure 1.1에서 > 기호는 R의 명령 프롬프트(prompt) 임

1. 현재 R session 정보(R 설치 버전, locale, 로딩 packages) 출력

```
# R의 설치 버전 및 현재 설정된 locale(언어, 시간대) 및 로딩된 R package 정보 출력  
sessionInfo()
```

```
R version 3.6.2 (2019-12-12)  
Platform: x86_64-w64-mingw32/x64 (64-bit)  
Running under: Windows 10 x64 (build 18363)

Matrix products: default

locale:  
[1] LC_COLLATE=Korean_Korea.949  LC_CTYPE=Korean_Korea.949  
[3] LC_MONETARY=Korean_Korea.949 LC_NUMERIC=C  
[5] LC_TIME=Korean_Korea.949

attached base packages:  
[1] stats      graphics   grDevices utils      datasets   methods    base

loaded via a namespace (and not attached):  
[1] compiler_3.6.2  magrittr_1.5    bookdown_0.16  tools_3.6.2  
[5] htmltools_0.4.0 yaml_2.2.1     Rcpp_1.0.3     stringi_1.4.5  
[9] rmarkdown_2.1   knitr_1.28    stringr_1.4.0  xfun_0.12
```

```
[13] digest_0.6.25     rlang_0.4.5      evaluate_0.14
```

2. 문자열 출력

```
#문자열 출력  
print("Hello R") #문자열
```

```
[1] "Hello R"
```

기호는 주석의 시작을 의미하고 실제로 실행되지 않음 같은 행에서 # 뒤 내용의 코드 역시 실행되지 않음

3. a라는 변수에 숫자 9, b라는 변수에 숫자 7를 할당 후 출력

```
# 수치형 값(scalar)을 변수에 할당(assign)  
# 여러 명령어를 한줄에 입력할 때에는 세미콜론(;)으로 구분  
a = 9; b = 7  
a
```

```
[1] 9
```

```
b
```

```
[1] 7
```

4. 변수 a와 b의 사칙연산

```
a+b; a-b; a*b; a/b
```

```
[1] 16
```

```
[1] 2
```

```
[1] 63
```

```
[1] 1.285714
```

5. R 그래픽 맛보기: 정규분포로부터 난수 100개 생성 후 생성된 데이터에 대한 히스토그램 작성

```
# 난수 생성 시 값은 매번 달라지기 때문에 seed를 주어 일정값이 생성되도록 고정
# "="과 "<-"는 모두 동일한 기능을 가진 할당 연산자임
#평균이 0이고 분산이 1인 정규분포에서 난수 100개 생성
set.seed(12345) # random seed 지정
x <- rnorm(100) # 난수 생성
hist(x) # 히스토그램
```

Histogram of x

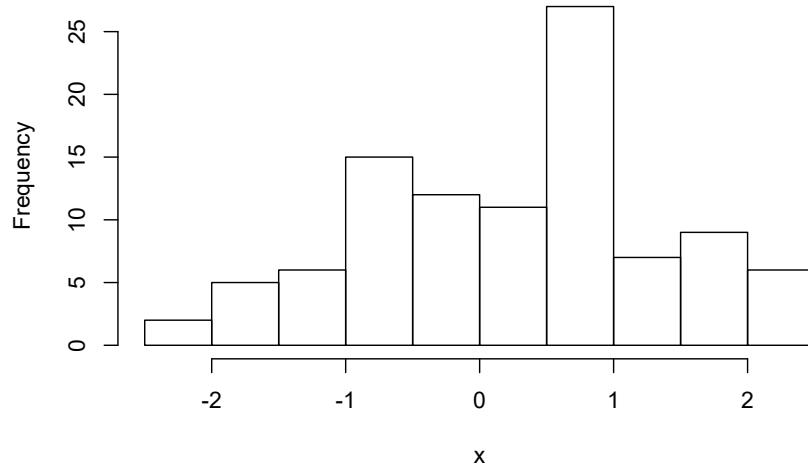


FIGURE 1.2: 정규분포 100개의 히스토그램

 R 명령어 또는 전체 프로그램 소스 실행 시 매우 빈번히 오류가 나타나는데, 이를 해결할 수 있는 가장 좋은 방법은 앞에서 언급한 Google을 이용한 검색 또는 R 설치 시 자체적으로 내장되어 있는 도움말을 참고하는 것이 가장 효율적임.

TABLE 1.1: R help 관련 명령어 리스트

| 도움말 보기 명령어 | 설명 | 사용법 |
|-----------------------|--|------------------------|
| ‘help’ 또는 ‘?’ | 도움말 시스템 호출 | ‘help(함수명)’ |
| ‘help.search’ 또는 ‘??’ | 주어진 문자열을 포함한 문서 검색 | ‘help.search(pattern)’ |
| ‘example’ | topic의 도움말 페이지에 있는 examples section 실행 | ‘example(함수명)’ |
| ‘vignette’ | topic의 pdf 또는 html 레퍼런스 메뉴얼 불러오기 | ‘vignette(패키지명 또는 패턴)’ |



Vignette 의 활용

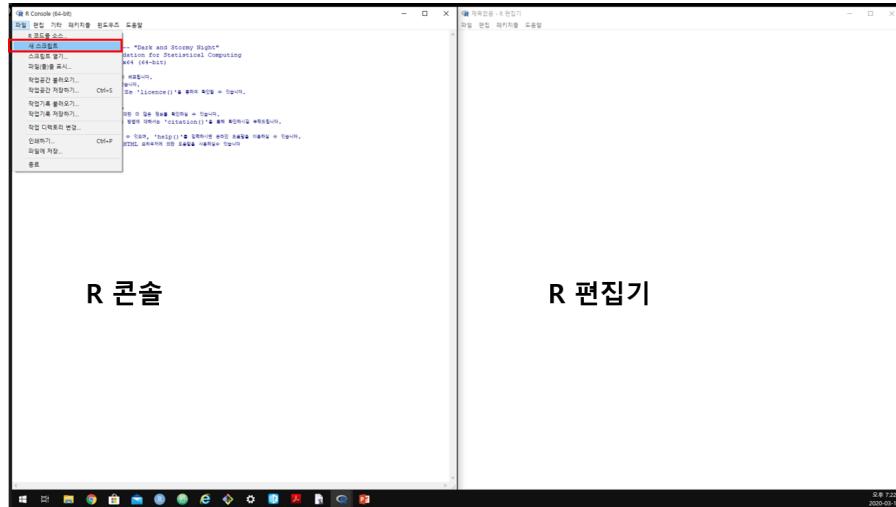
- vignette()에서 제공하는 문서는 데이터를 기반으로 사용하고자 하는 패키지의 실제 활용 예시를 작성한 문서이기 때문에 초보자들이 R 패키지 활용에 대한 접근성을 높혀줌.
- browseVignettes() 명령어를 통해 vignette을 제공하는 R 패키지 및 해당 vignette 문서 확인 가능

1.3 R script 편집기 사용



실습: R 설치 후 Rgui에서 제공하는 편집기(R editor)에 명령어를 입력하고 실행

설치된 R을 실행 후 상단 pull-down 메뉴에서 [File] → [새 스크립트]를 선택하면 아래 그림과 같이 편집창(R 인스톨 시 SDI 옵션 기준)이 나타남



편집기 창에 다음 명령어 입력

```
# R에 내장된 cars 데이터셋 불러오기 cars dataset에 포함된 변수들의 기초통계량
# 출력 2차원 산점도
data(cars)
help(cars) # cars 데이터셋에 대한 설명 help 창에 출력
head(cars) # cars 데이터셋 처음 6개 행 데이터 출력
summary(cars) # cars 데이터셋 요약
plot(cars) # 변수가 2개인 경우 산점도 출력
```

- 편집창에서 한 줄을 실행시키려면 명령어가 입력된 줄에서 [Ctrl] + [R] 입력
- 편집창에 입력한 모든 명령어를 실행시키려면 모든 줄을 선택(마우스 또는 [Shift] + ↓)

```
speed dist
1     4    2
2     4   10
3     7    4
4     7   22
```

```
5      8    16  
6      9    10  
  
       speed          dist  
Min.   : 4.0   Min.   : 2.00  
1st Qu.:12.0   1st Qu.: 26.00  
Median :15.0   Median : 36.00  
Mean   :15.4   Mean   : 42.98  
3rd Qu.:19.0   3rd Qu.: 56.00  
Max.   :25.0   Max.   :120.00
```

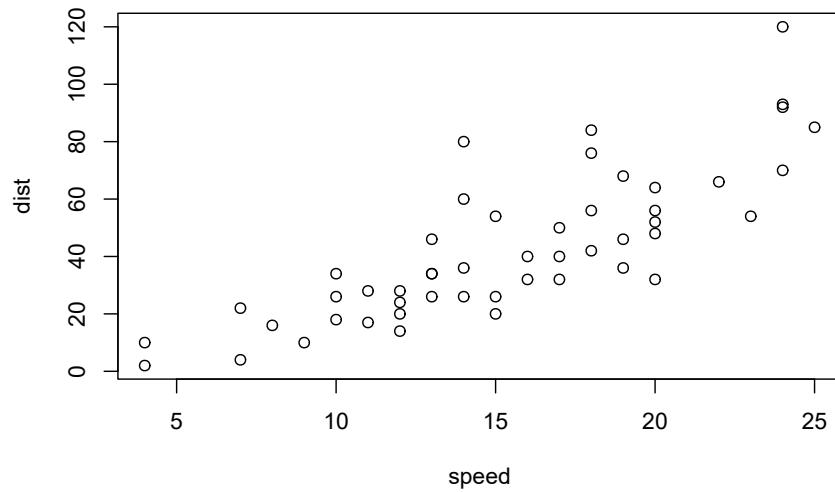


FIGURE 1.3: cars 데이터셋의 speed와 dist 간 2차원 산점도: speed는 자동차 속도(mph)이고 dist는 해당 속도에서 브레이크를 밟았을 때 멈출 때 까지 걸린 거리(ft)를 나타냄.

- R은 명령어를 입력하고 실행결과를 확인하는 대화형(interpreter) 방식
- 콘솔창에서 ↑/↓를 누르면 이전/이후 실행 명령 기록 확인 가능

- 여러 줄 이상 R 명령어라든가 반복적, 장기간 작업을 수행해야 할 경우 R 명령어로 구성된 스크립트 작성 후 일괄 실행하는 것이 일반적
 - 여러 다중 명령 코딩 시 콘솔창에 직접 입력하는 것은 비효율적이므로 스크립트 에디터를 사용
 - 위 예시처럼 R 에디터 사용할 수 있으나 가독성 및 코딩 효율이 떨어짐
 - 과거 많이 사용됐던 R 에디터: WinEdt⁹, Tinn-R¹⁰, Vim¹¹
 - 현재 가장 범용적 R 에디터: **Rstudio**
-

1.4 RStudio

- RStudio¹²: R 통합 분석/개발 환경(integrated development environment, IDE)으로 현재 가장 대중적으로 사용되고 있는 R 사용 환경
- 명령 콘솔 외 파일 편집, 데이터 객체, 명령 기록(.history), 그래프 등에 쉽게 접근 가능
- RStudio 독자적인 개발 환경 제공: Rmarkdown, Rnotebook, Shiny Web Application 등 다양한 R 환경을 제공
- 버전관리(git, subversion)를 통해 project 관리 가능
- 무료 및 유료 소프트웨어 제공

1.4.1 RStudio 설치하기

1. 웹 브라우저를 통해 <https://rstudio.com> 접속 후 상단 DOWNLOAD¹³ 링크 클릭

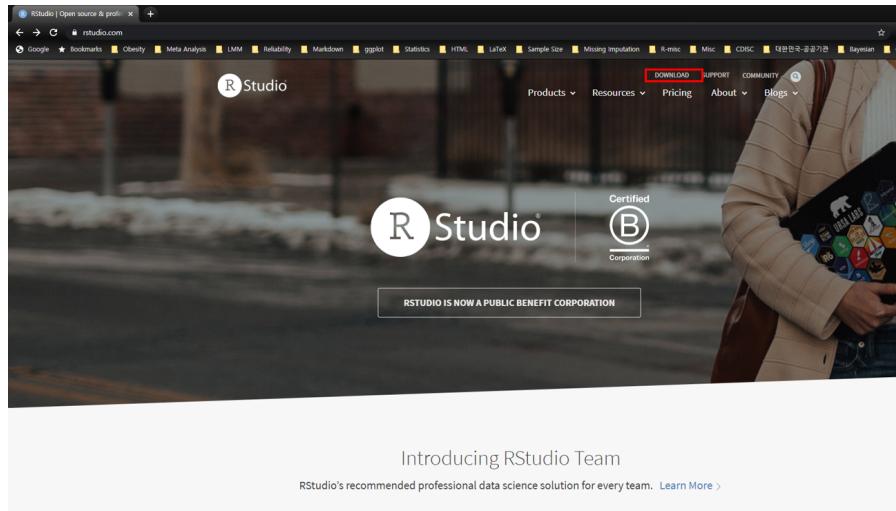
⁹<http://www.winedt.com>

¹⁰<https://sourceforge.net/projects/tinn-r/>

¹¹http://www.vim.org/scripts/script.php?script_id=2628

¹²<https://rstudio.com/>

¹³<https://rstudio.com/products/rstudio/download/>



2. Desktop 또는 Server 버전 중 택일

- 서버용 설치를 위해서는 Server 클릭 → 소규모 자료 분석용으로는 불필요
- 여기서는 **Desktop** 버전 선택 후 다음 링크로 이동

The screenshot shows the RStudio website's download page. At the top, there's a navigation bar with links for DOWNLOAD, SUPPORT, COMMUNITY, Products, Resources, Pricing, About, and Blogs. Below the navigation is a large blue banner with the text "Download RStudio" and a background of hexagonal icons representing various R packages like dplyr, testthat, stringr, devtools, etc.

Choose Your Version

RStudio is a set of integrated tools designed to help you be more productive with R. It includes a console, syntax-highlighting editor that supports direct code execution, and a variety of robust tools for plotting, viewing history, debugging and managing your workspace.

[LEARN MORE ABOUT RSTUDIO FEATURES](#)

RStudio Desktop

- Open Source License
- Free**
- [DOWNLOAD](#) (button highlighted with a red box)
- [Learn more](#)

RStudio Desktop

- Commercial License
- \$995 /year**
- [BUY](#)
- [Learn more](#)

RStudio Server

- Open Source License
- Free**
- [DOWNLOAD](#)
- [Learn more](#)

RStudio Server Pro

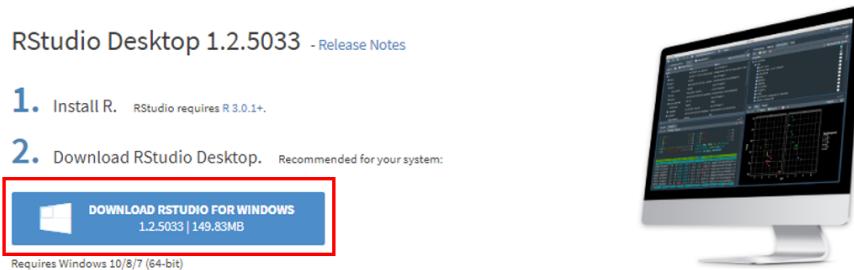
- Commercial License
- \$4,975 /year
(5 Named Users)**
- [BUY](#)
- [Evaluation | Learn more](#)

RStudio Team

RStudio's new solution for every professional data science team. RStudio Team includes RStudio Server Pro, RStudio Connect and RStudio Package Manager.

[LEARN MORE](#)

3. 운영체제에 맞는 Rstudio installer 다운로드(여기서는 Windows 버전 다운로드)



All Installers

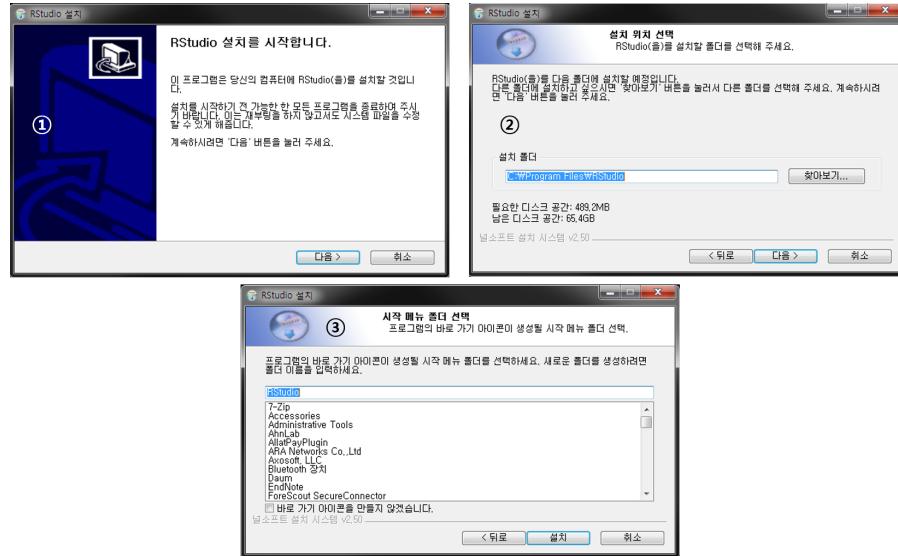
Linux users may need to [import RStudio's public code-signing key](#) prior to installation, depending on the operating system's security policy.
RStudio 1.2 requires a 64-bit operating system. If you are on a 32 bit system, you can use an [older version of RStudio](#).

| OS | Download | Size | SHA-256 |
|---------------------|---|-----------|----------|
| Windows 10/8/7 | RStudio-1.2.5033.exe | 149.83 MB | 7fd0b01b |
| macOS 10.12+ | RStudio-1.2.5033.dmg | 126.89 MB | b67e9875 |
| Ubuntu 14/Debian 8 | rstudio-1.2.5033-amd64.deb | 98.18 MB | 89cc2e22 |
| 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 |
| Fedor 19/Red Hat 7 | rstudio-1.2.5033-x86_64.rpm | 120.23 MB | 580f45c6 |
| Fedor 28/Red Hat 8 | rstudio-1.2.5033-x86_64.rpm | 120.87 MB | 452bc0d0 |

4. RStudio installer 다운로드 시 파일이 저장된 폴더에서 보통

RStudio-xx.xx.xxx.exe 형식의 파일명 확인

- 더블 클릭 후 실행
- [다음>] 몇 번 클릭 후 설치 종료



Bibliography

Rizzo, M. L. (2019). *Statistical computing with R*. CRC Press.

Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. Springer.

Wickham, H. and Grolemund, G. (2016). *R for data science: import, tidy, transform, visualize, and model data*. ” O'Reilly Media, Inc.”.

Xie, Y. (2019). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.16.

권재명 (2017). 실리콘밸리 데이터 과학자가 알려주는 따라하며 배우는 데이터 과학. 제이펍, 1st edition. ISBN 979-1185890869.

서민구 (2014). *R을 이용한 데이터 처리 & 분석*. 길벗, 1st edition. ISBN 978-8966188260.

유충현, 이상호, and 김정일 (2005). *R 그래픽스*. 자유아카데미, 1st edition. ISBN 978-8973385539.