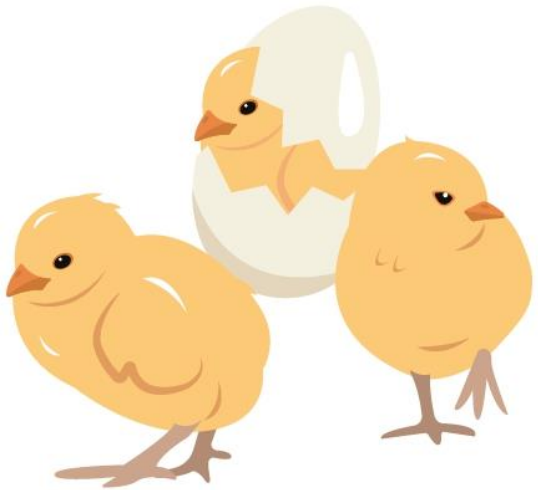
모두를 위한

R 데이터 분석 입문

2판



Chapter 01
데이터 분석과 R

목차

Section 01

데이터의 시대

1. 데이터의 비즈니스 활용

- 우리는 데이터의 시대(the age of data)에 살고 있음, 정보화 시대 → 데이터의 시대
- 우리를 둘러싼 모든 것들이 데이터 소스와 연결되고, 우리 삶의 많은 부분이 데이터에 의존하여 영위

  ex) 이메일, SNS, 전화사용 기록, 신용카드거래 기록, 병원 치료 기록, 성적, 인터넷, 주민정보, 등기정보, 판매정보, 주식거래 정보 등
- 데이터는 기업 활동에도 중요함, 대형마트들은 소비자의 구매 내역 데이터를 바탕으로 구매 패턴을 분석하고 이를 영업에 활용

그림 1-1 **판매유통 대형마트의 진열대:** 구매 패턴 데이터를 분석하여 활용

맥주를 산 고객이 견과류도 함께 구매하는 비율이 높다고 분석되면
↓
맥주 바로 옆에 견과류를 진열
↓
동반 매출 상승

1. 데이터의 시대

- 서울시는 심야교통버스, 일명 '올빼미버스'의 노선을 결정하기 위해 이동통신사로부터 심야 휴대폰 발신 데이터를 받아 분석
- 이를 통해 **사람들이 많이 모여 있는 지점**을 알아 낼 수 있었고, 이를 **올빼미버스 노선에 반영**함으로써 시민들의 만족도를 높일 수 있었음

그림 1-3 올빼미버스 노선: 심야 휴대폰 발신 데이터 분석

2. 4차 산업혁명과 데이터

- 2016년 1월, 스위스 다보스에서 열렸던 세계경제포럼(World Economic Forum)에서 클라우스 슈밥(Klaus Schwab)은 기술 혁명의 새로운 시대가 열렸음을 천명하면서 이를 **'4차 산업혁명(The Fourth Industrial Revolution)'**이라고 명명
- 4차 산업혁명이란 **인공지능**(Artificial Intelligence, AI), **빅데이터**(big data), **로봇**(robot), **사물인터넷**(Internet of Things, IoT), **생명공학기술**(Biotechnology), **3D 프린터**(3D printer) 등 새로운 과학기술이 사회, 경제, 문화 전반에 영향을 미치게 되고, 이러한 변화를 잘 수용하고 가능성을 최대화 하는 시대를 말함
- **인공지능**과 **빅데이터**가 4차 산업혁명의 핵심 기술로 인식

그림 1-4 4차 산업혁명까지의 과정과 핵심 기술

Section 02

빅데이터

1. 빅데이터의 특징

- 기존의 데이터베이스 관리도구의 데이터 수집, 저장, 관리, 분석 역량을 넘어서는 데이터
- 의료 분야의 환자 데이터, 금융 분야의 거래 데이터, 교통 분야의 대중교통 이용 데이터 등도 빅데이터에 해당

1.1 크기(volume)

- 일반적으로 수십 테라바이트(terabyte), 또는 수십 페타바이트(petabyte) 이상이 빅데이터 범위, 1페타바이트는 6기가바이트 DVD 영화를 17만 4천 편 담을 수 있는 정도의 용량

1.2 다양성(variety)

❶ 정형 데이터: 고정된 필드에 저장되는 일정한 형식의 데이터 ex) 엑셀 파일

❷ 반정형 데이터: 일정한 구조는 없으나 구조를 파악할 수 있는 데이터
ex) XML이나 HTML 같은 메타데이터

❸ 비정형 데이터: 고정된 필드에 저장되지 않는 데이터 ex) 사진, 동영상, 위치 정보 등

1.3 속도(velocity)

- 빅데이터는 빠른 증가 속도, 소비 속도를 갖음 ex) 지하철 승하차 정보, SNS 상 메시지

2. 빅데이터

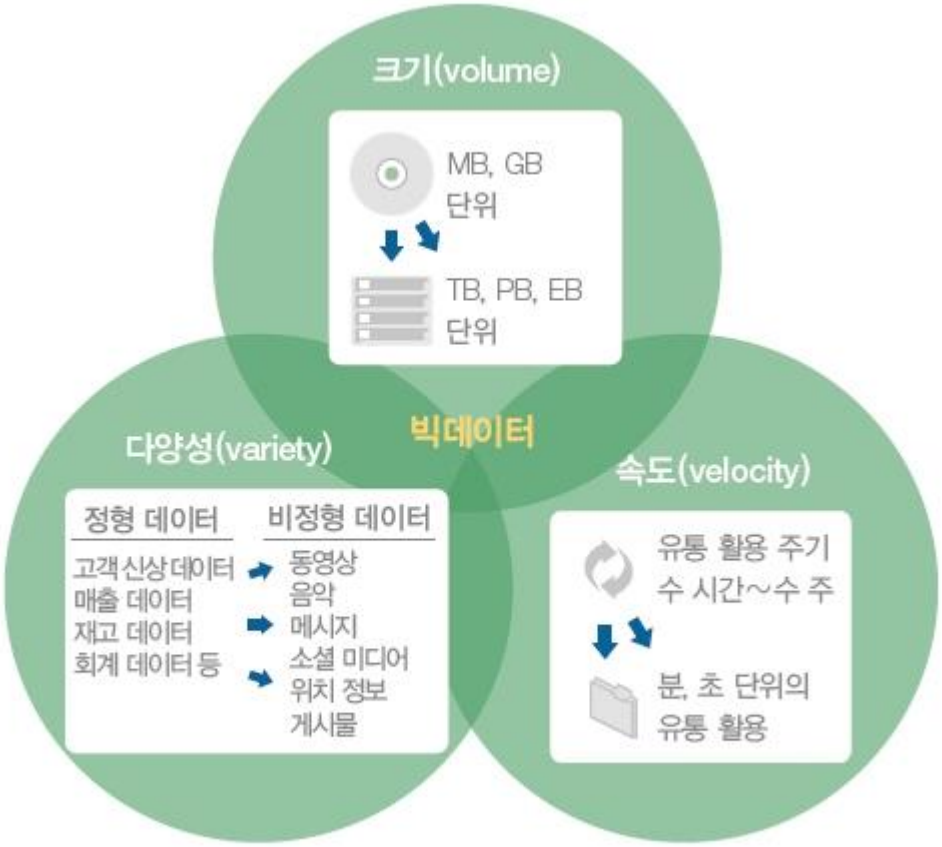


그림 1-5 빅데이터의 특징

2. 빅데이터의 성공 사례

2.1 국내 활용 사례: 아파트 관리비 적정성 평가

- 경기도는 국토교통부의 공동주택관리정보시스템에 의무적으로 등록하는 각 아파트 관리사무소의 관리비 내역과 관리비를 구성하는 37개 세부항목의 원천데이터를 비교·분석하는 방식으로 관리비 과다 청구 여부를 분석
- 분석 결과를 가지고 아파트 관리비 산출 표준 모델 및 **아파트관리비부당지수** 개발
- 556개 단지를 샘플로 조사하여 2년간 152억원의 관리비가 부당하게 징수된 사실이 적발 전국적으로 적용될 경우 **연간 1조 1000억원 정도의 관리비를 절감**할 수 있을 것으로 예상

Section 03

데이터 분석 과정

3. 데이터 분석 과정

그림 1-7 데이터 분석 과정

1단계: 문제 정의 및 계획

- 문제가 명확해야 그 문제를 해결하기 위한 데이터가 어떤 것인지를 추정할 수 있고, 어떤 분석기법을 적용해야 할지도 계획할 수 있음

2단계: 데이터 수집

- 기존 시스템의 데이터베이스, 엑셀파일, 종이 문서, 장비내의 파일, 인터넷 등에서 필요한 자료를 수집

3단계: 데이터 정제 및 전처리

- 수집된 데이터는 바로 분석에 사용할 수 없는 경우가 대부분
- 단위의 차이, 결측값, 오류 데이터 등의 보정 필요
- 수집된 데이터를 분석이 가능한 형태로 정돈하는 과정을 데이터 정제 혹은 전처리 과정

4단계: 데이터 탐색

- 가벼운 데이터 분석
- 전반적인 데이터의 내용을 파악하는 단계

5단계: 데이터 분석

- 데이터 탐색 단계에서 파악한 정보를 바탕으로 보다 심화된 분석을 수행하는 단계
- 전통적인 통계분석을 포함하여 고급 분석 기법들이 사용됨
- 머신러닝 기술도 적용됨

6단계: 결과 보고

- 데이터의 분석과 해석이 마무리 되면 그 내용이 정리되고, 보고 되어야 함
- 결과보고 작성단계에서 중요한 기술이 바로 데이터 시각화(visualization)
  - 분석된 결과를 단순 숫자의 나열이 아니라 다양한 그래프나 그림을 통해서 결과를 쉽게 이해할 수 있도록 표현하는 것

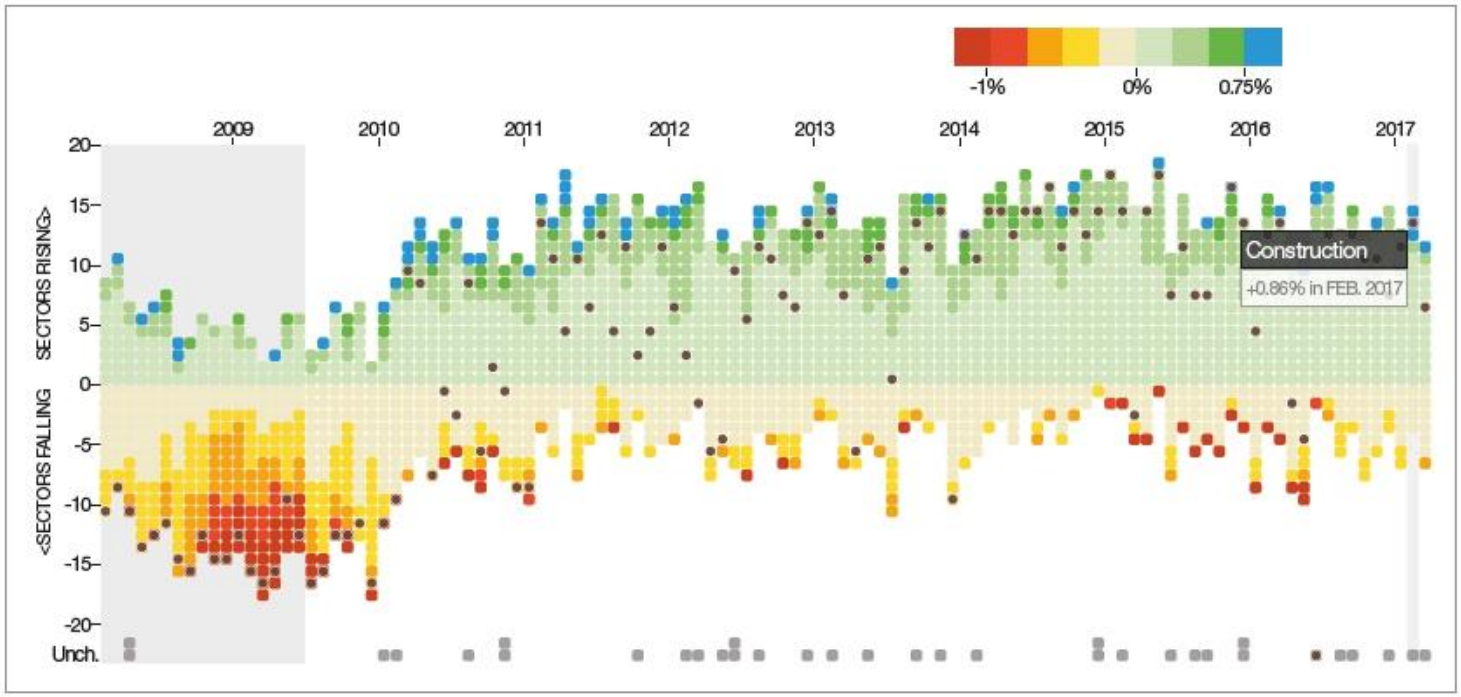


그림 1-8 데이터 시각화의 사례: 미국의 연도별 취업자와 실업자 통계 ©Toptal

여기서 잠깐! 데이터 분석의 소요 시간

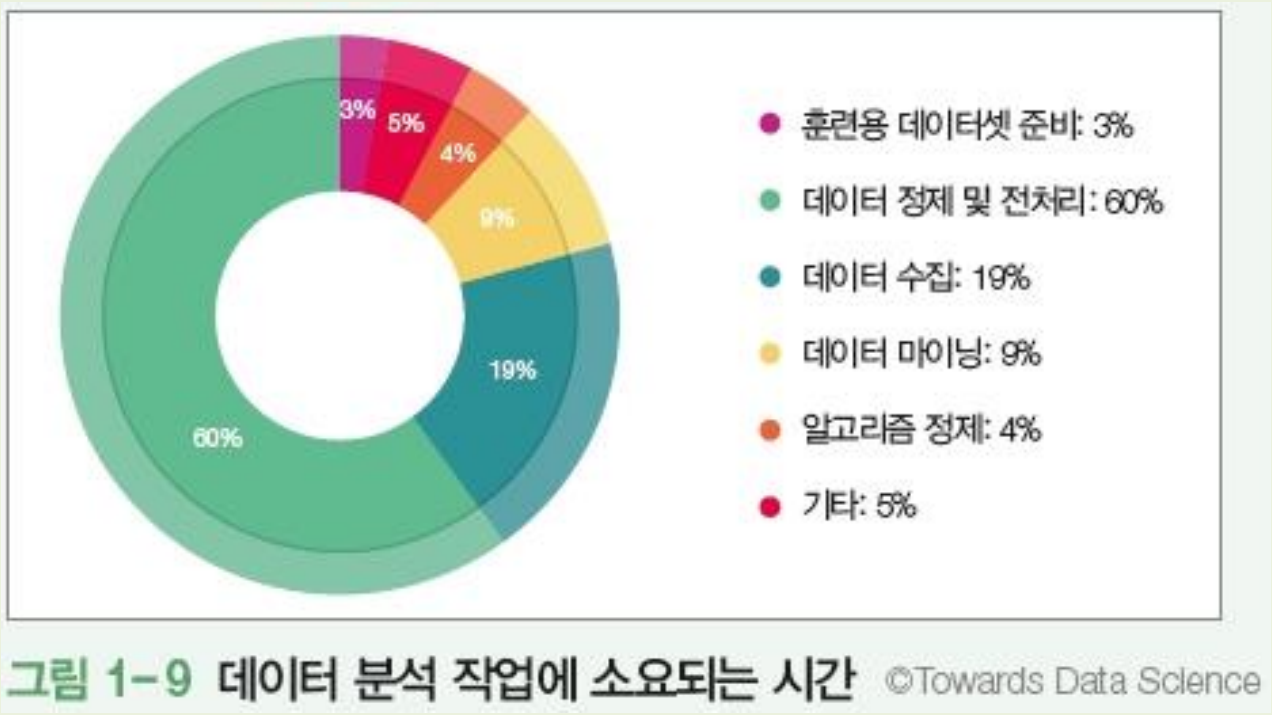


그림 1-9 데이터 분석 작업에 소요되는 시간 ©Towards Data Science

1. 데이터를 **수집**하는 일에 **19%**, 데이터를 **정제하고 전처리**하는 데 **60%**의 시간을 사용
   → 즉, 전체 분석 과정에서 약 **80%의 시간**이 분석을 위한 **데이터 준비**에 사용
2. 이러한 시간을 얼마나 줄이느냐가 전체 분석 시간을 줄이는 관건

Section 04

R과 R스튜디오의 설치 및 사용

1. R과 R스튜디오의 소개

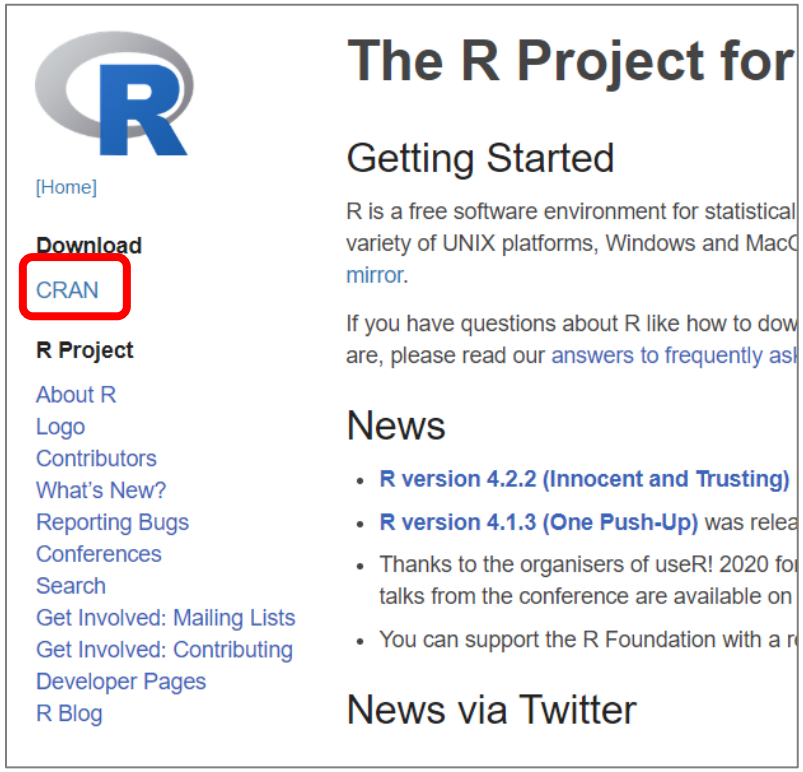
- Python : 프로그래밍 언어로서의 특성이 강함
- R : 데이터 분석을 목적으로 개발, R로 SW를 만들지는 못함
  - RStudio라는 훌륭한 작업환경 제공
  - 풍부한 패키지 제공
  - 미려한 데이터 시각화 패키지 제공

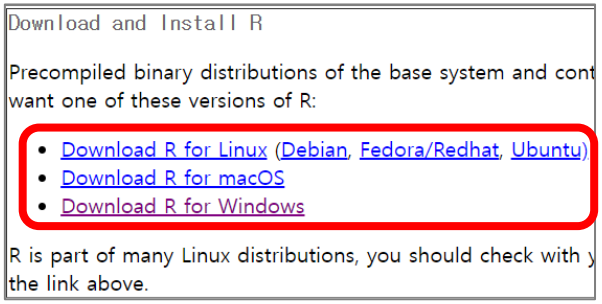
그림 1-10 R과 파이썬

그림 1-11 R을 쉽게 사용할 수 있는 환경을 제공하는 R스튜디오

2. R의 설치

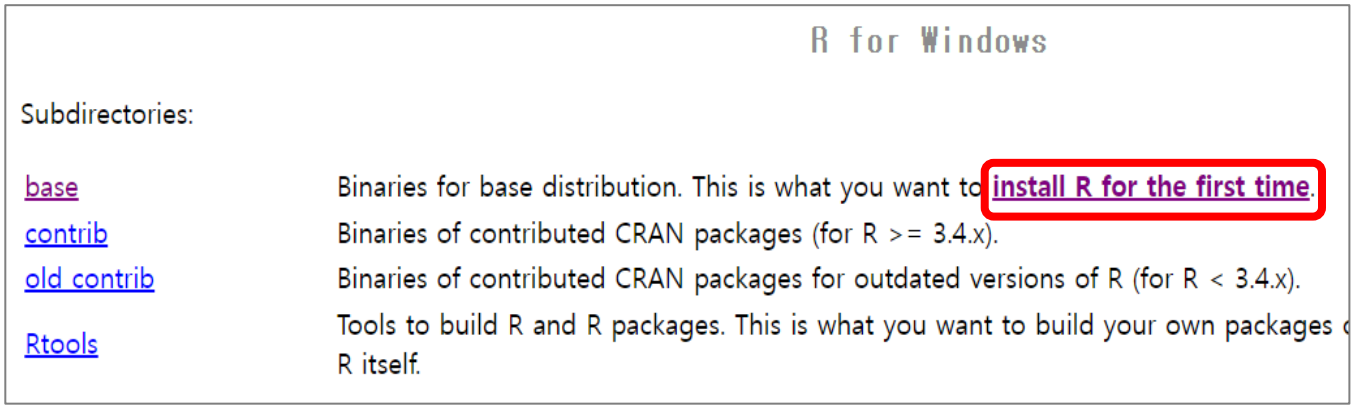
01 https://www.r-project.org/ 에 접속하여 설치 진행

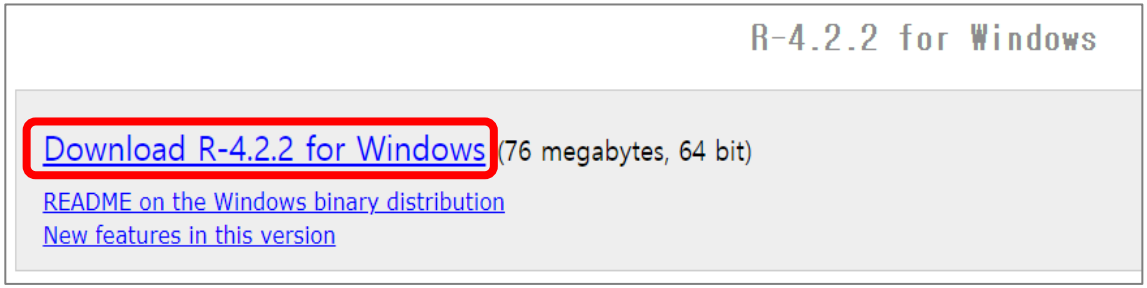




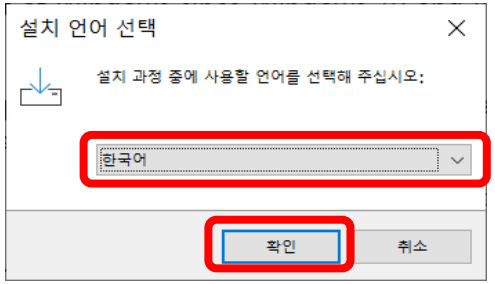
자신의 운영체제에 맞는 버전을 설치한다

02 [install R for the first time] 링크 클릭 → [Download R-4.2.2 for Windows] 클릭



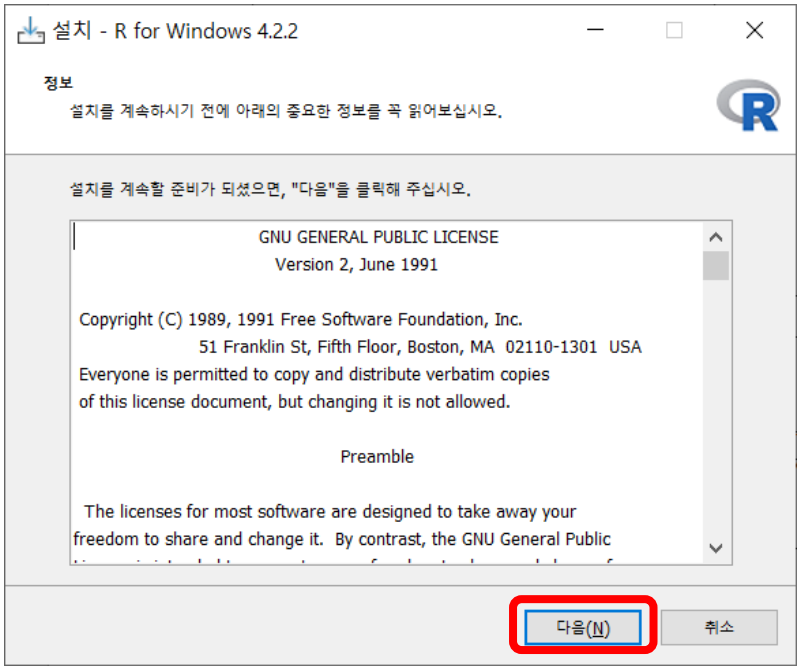


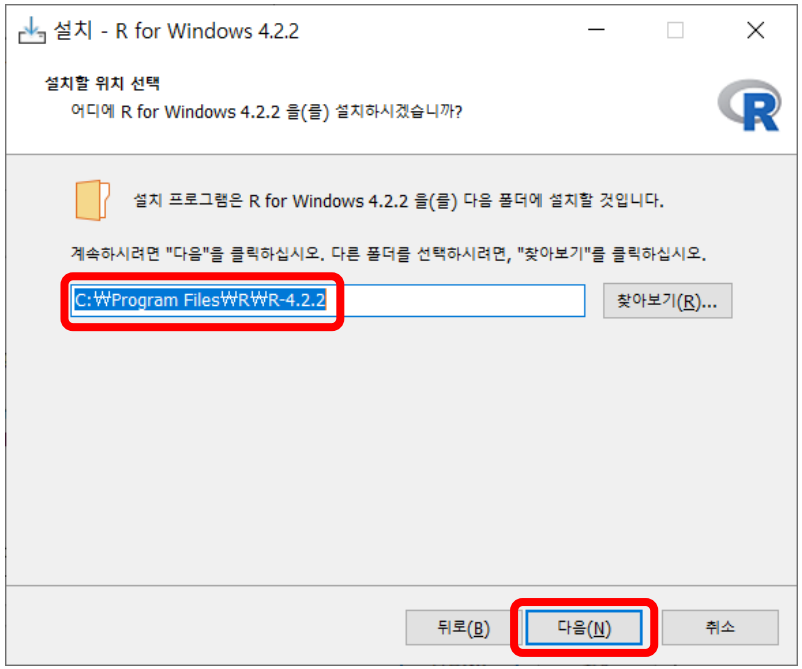
03 [한국어] 선택하고 [확인] 클릭



04 설치 정보가 나타나면 내용 확인하고 [다음] 버튼 클릭

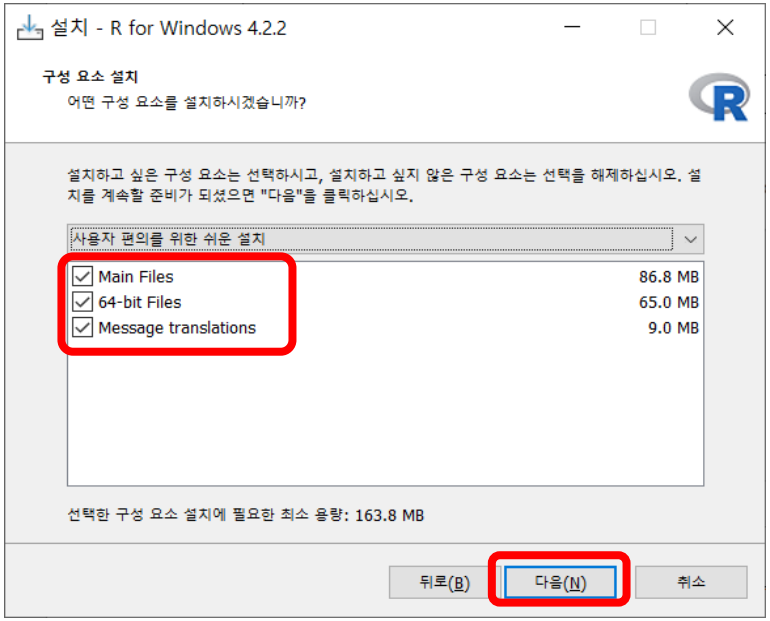
→ 설치할 위치 선택에서 경로 변경하거나 유지한 채 [다음] 버튼 클릭

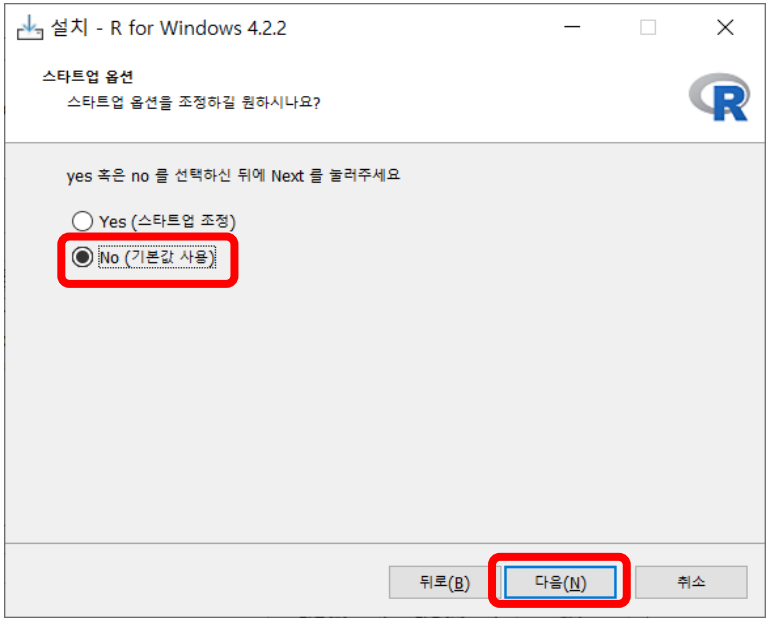




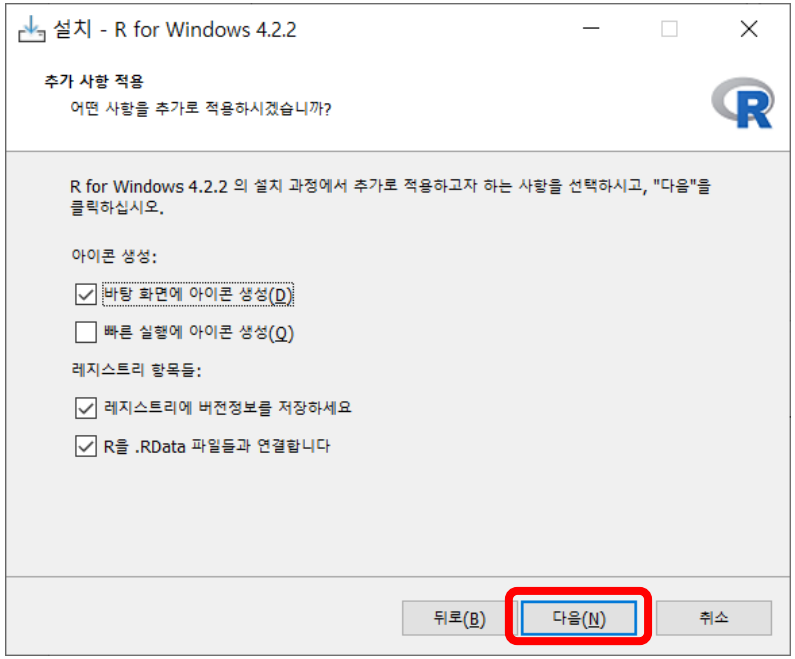
05 구성 요소 설치에서 필요한 항목 체크하고 [다음] 버튼 클릭

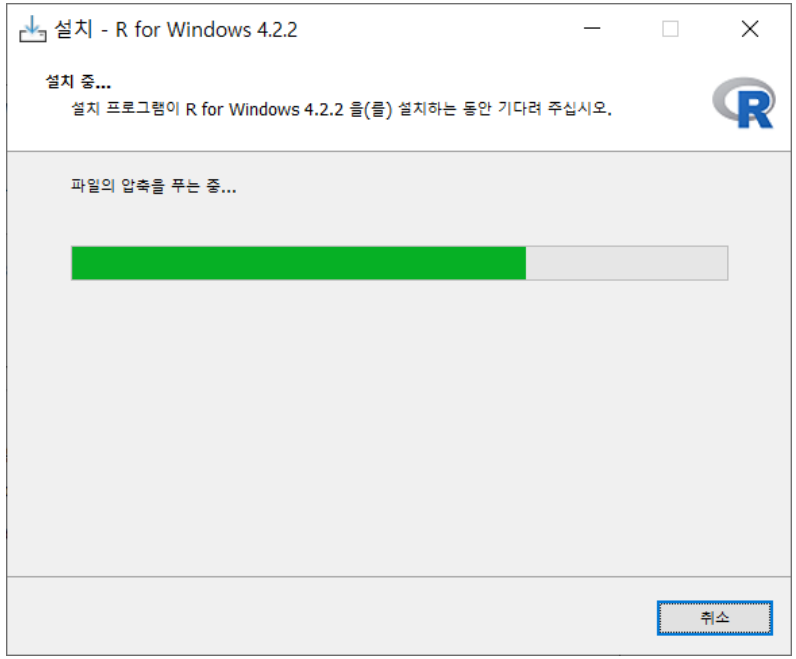
→ 스타트업 옵션에서 [No]를 선택 후, [다음] 버튼 클릭



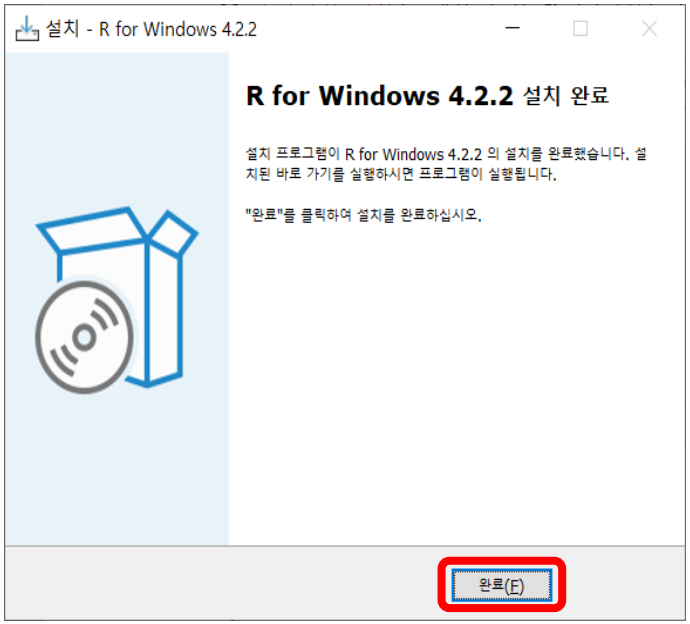


06 추가 사항 적용에서 내용 변경 없이 [다음] 버튼 클릭 → 설치가 진행됨





07 설치 완료 창 열리면 [완료] 버튼을 눌러 설치 완료

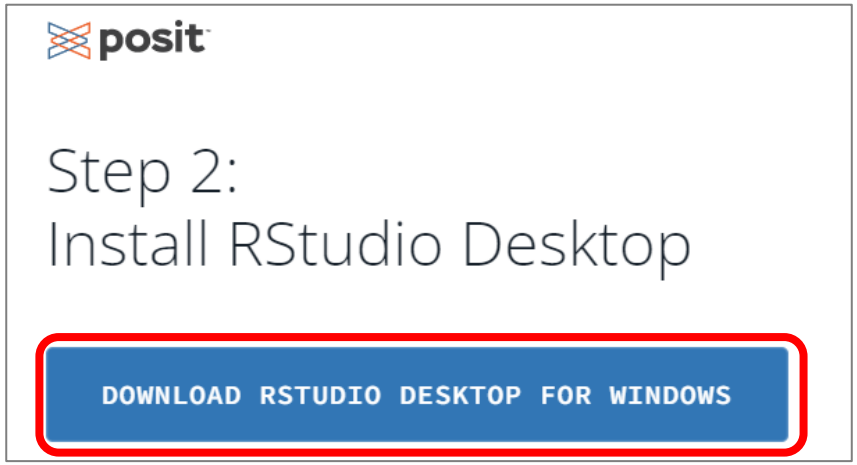


3. R스튜디오의 설치

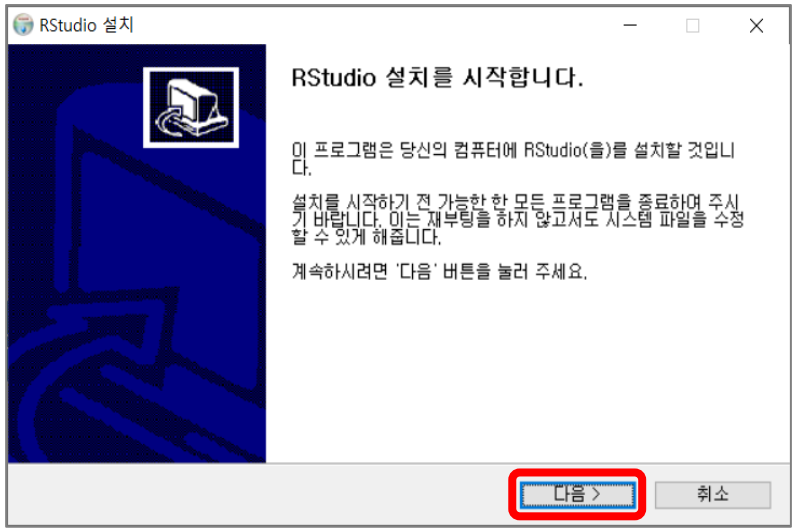
01 https://posit.co/download/rstudio-desktop/에 접속

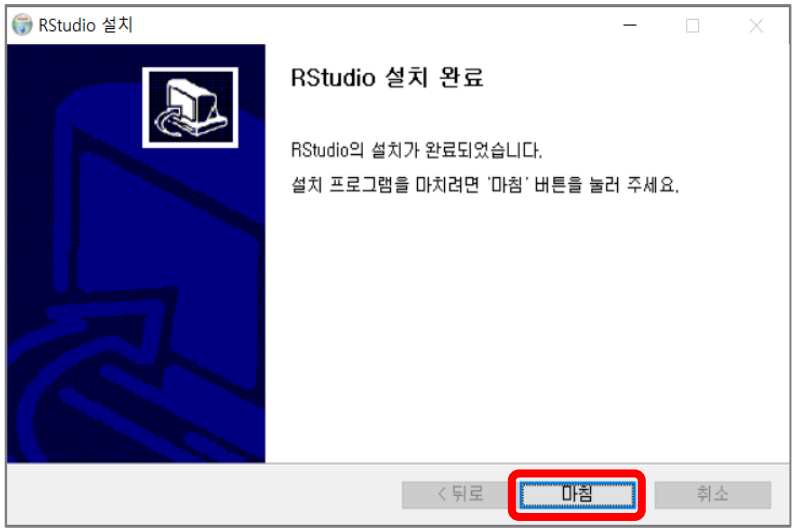
Step 1: Install R ▶ 이미 설치 했으므로 Skip

Step 2: Install RStudio Desktop ▶ 설치 파일을 다운로드

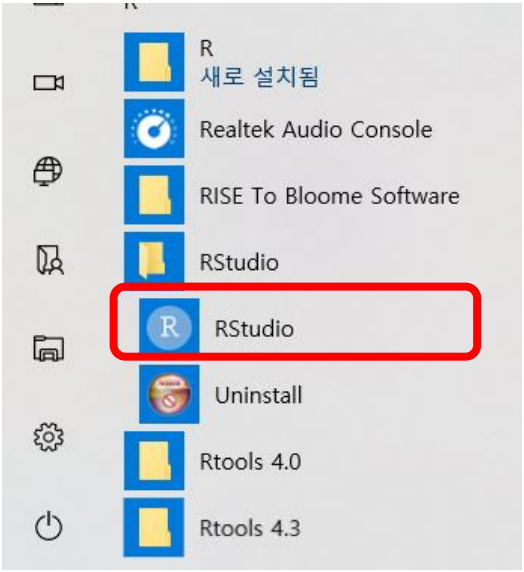


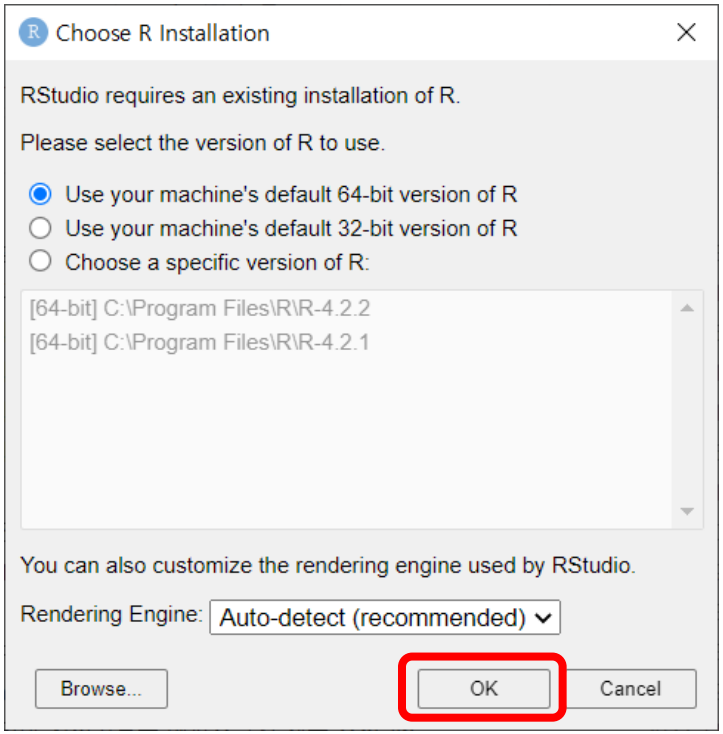
02 설치 파일 더블클릭 → 계속해서 [다음] 버튼을 클릭 → R스튜디오 설치가 완료되면 [마침] 버튼 클릭





04 설치가 완료되면 윈도우 시작 메뉴에서 [RStudio]-[RStudio] 클릭하여 R스튜디오 실행





설치된 R 버전이 여러 개인 경우 어떤 버전을 사용할지 묻는다. 선택사항 변경 없이 [OK] 클릭

4. R스튜디오의 화면 구성

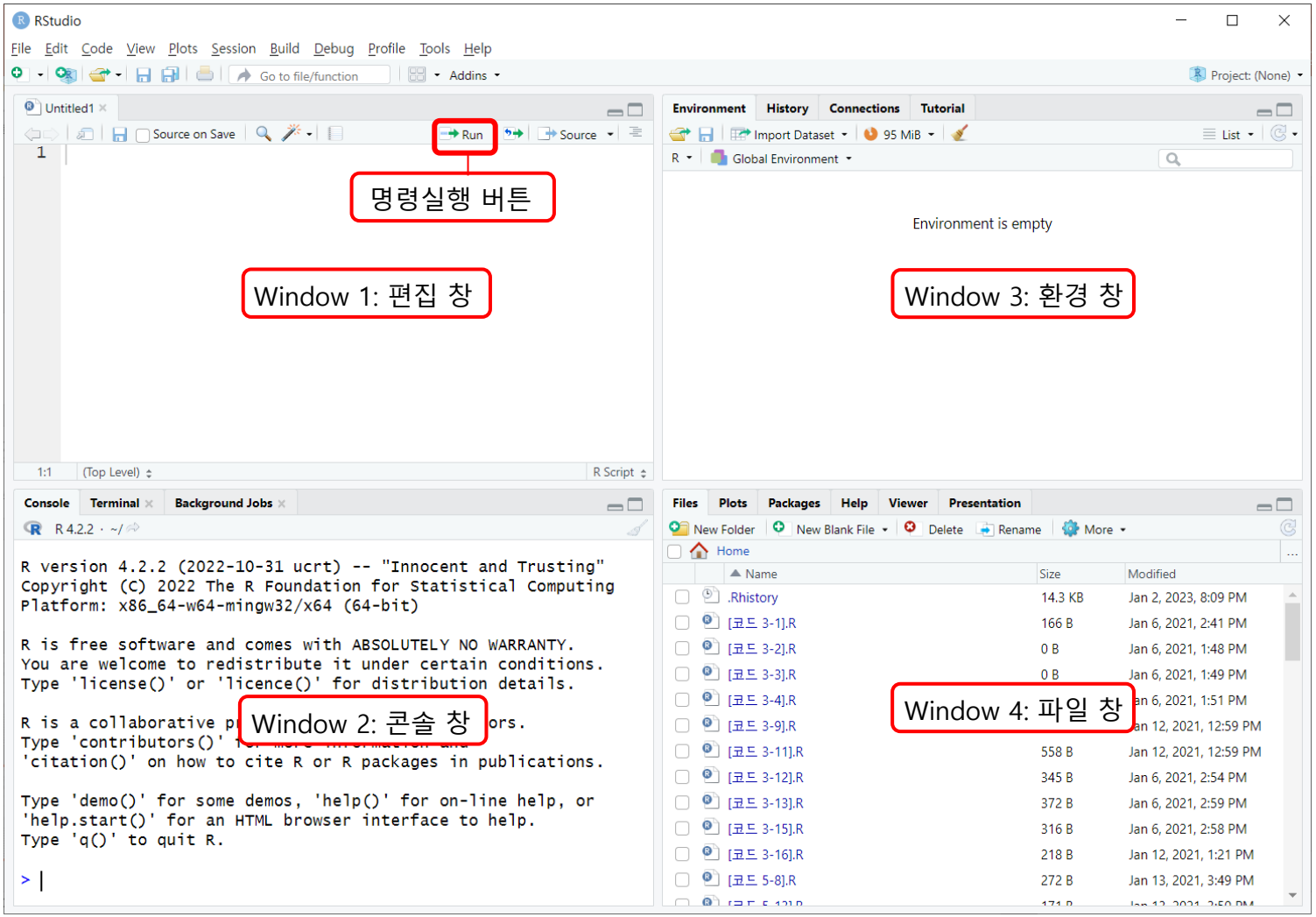


그림 1-12 R스튜디오 초기 화면

4.1 편집(Script) 창

- R 명령문('R 스크립트' 라고도 한다.)들을 작성하고 실행하는 영역

4.2 콘솔(Console) 창

- 편집 창에서 R 명령문을 편집하고 실행 버튼을 클릭했을 때, 명령문의 실행 과정 및 결과를 표시하는 창

4.3 환경(Environment) 창

- R 명령문이 실행하는 동안 만들어지는 각종 변수나 자료구조의 내용을 보여주는 영역

4.4 파일(Files) 창

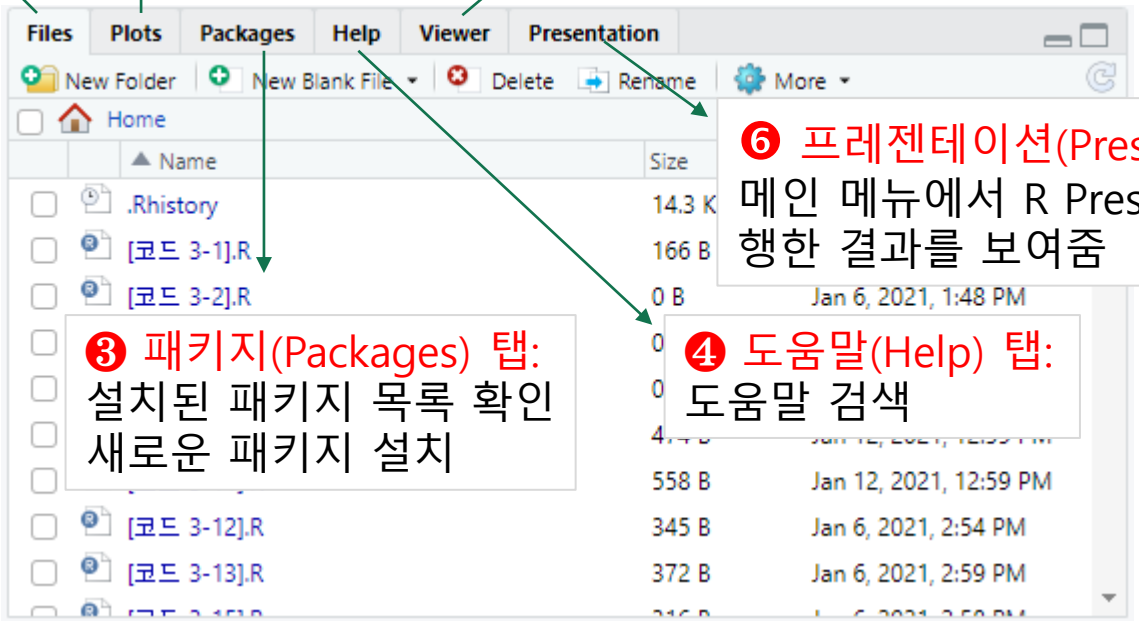
- 도움말, 패키지 설치 및 조회, 그래프 실행 내용 조회 등 유용한 기능을 제공하는 창

4. R과 R스튜디오의 설치 및 사용

❶ 파일(Files) 탭:
현재 작업 폴더의 내용을 탐색기처럼 보여준다

❷ 플롯(Plots) 탭:
그래프가 표시되는 영역

❺ 뷰어(Viewer) 탭:
결과가 웹브라우저에 나타나는 경우 여기에 표시



❻ 프레젠테이션(Presentation) 탭:
메인 메뉴에서 R Presentation을 실행한 결과를 보여줌

❸ 패키지(Packages) 탭:
설치된 패키지 목록 확인
새로운 패키지 설치

❹ 도움말(Help) 탭:
도움말 검색

5. R스튜디오 다루기

5.1 R스튜디오 화면 재구성하기

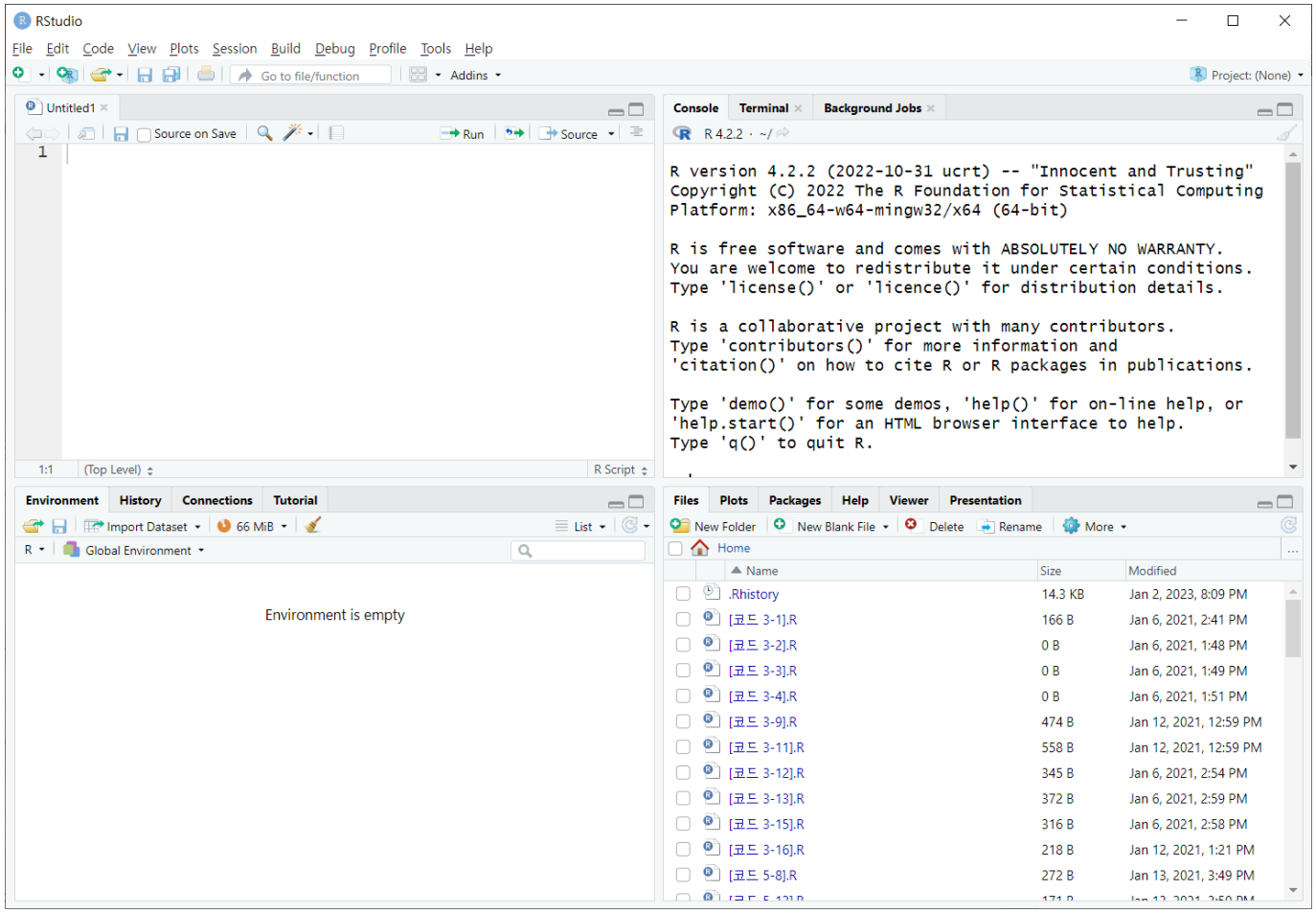


그림 1-13 콘솔 창 재배치 후의 R스튜디오

5.2 R스튜디오에서 명령문의 실행

```
5+8
3+(4*5)
a <- 10
print(a)
```

```
> 5+8
[1] 13
```

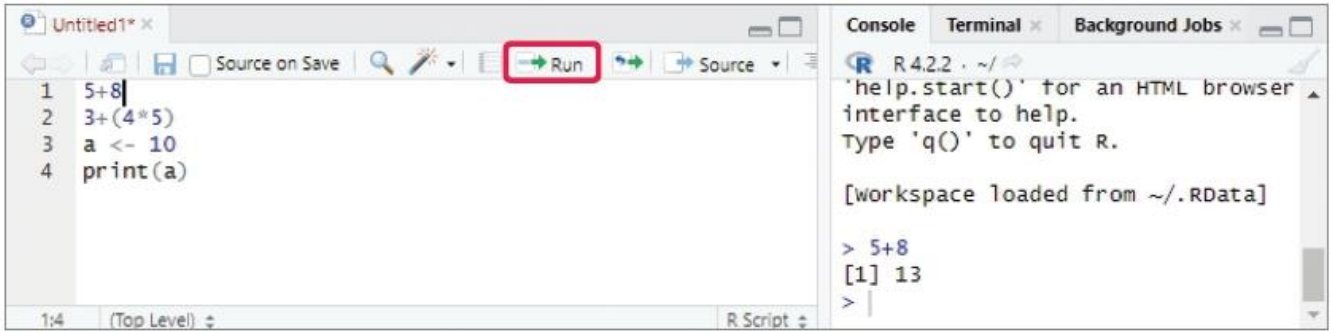


그림 1-14 편집 창 1행에 커서를 놓고 실행 아이콘을 클릭했을 때 콘솔 창에서의 실행 결과

4. R과 R스튜디오의 설치 및 사용

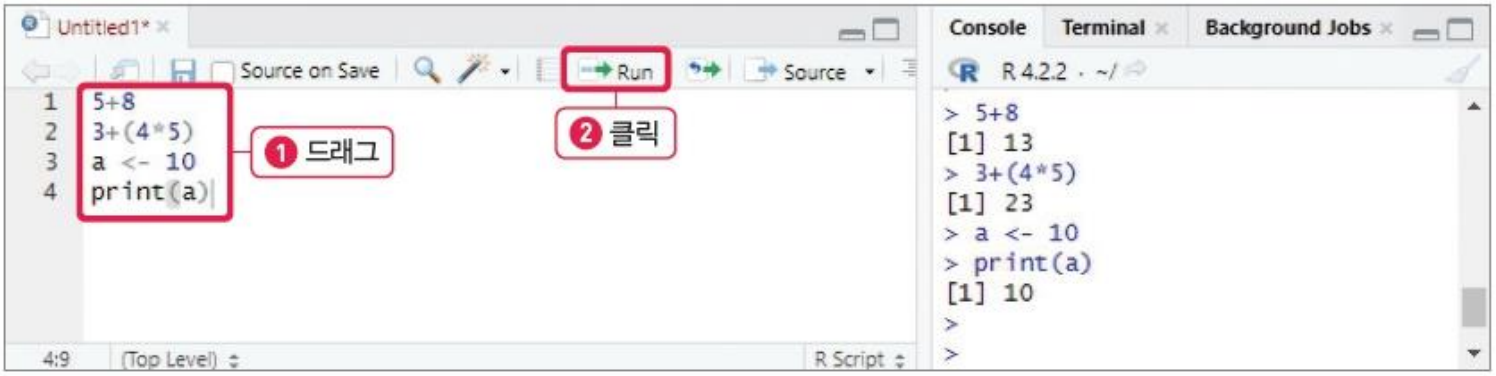
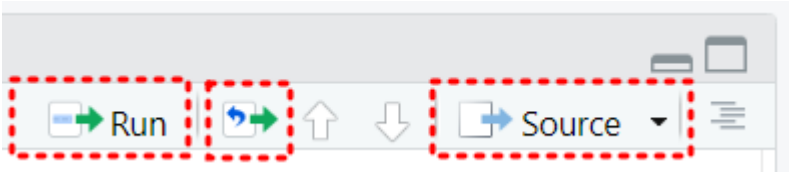
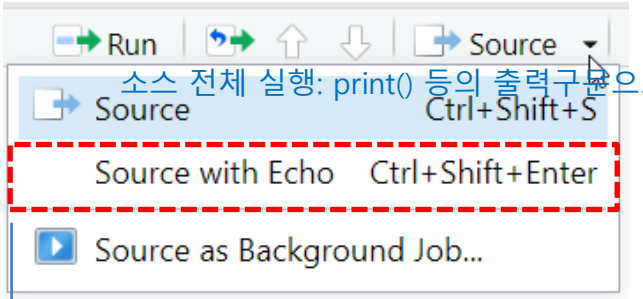


그림 1-15 편집 창 1~4행을 블록 선택하고 실행 아이콘을 클릭했을 때 콘솔 창에서의 실행 결과





소스 전체 실행: print() 등의 출력구문으로만 결과 표시

표 1-1 명령어 실행과 관련된 단축키

동일한 기능: 파일 전체를 콘솔에서 코드와 실행 표시

명령어 실행	단축키
한 줄만 실행할 때	명령어가 있는 줄에서 Ctrl + Enter
여러 줄을 실행할 때	명령어들을 드래그하여 블록을 만든 후 Ctrl + Enter
편집된 모든 명령문들을 실행할 때	Ctrl + Alt + R
바로 직전에 실행한 명령을 다시 실행할 때	Ctrl + Alt + P

Ctrl + Shift + Enter == Ctrl + Alt + R

5.3 R스튜디오에서의 저장과 종료

- 메뉴에서 [File]-[Save] 또는 [File]-[Save As]
- R 스크립트 파일의 확장자 이름은 일반적으로 'test.R'과 같이 '.R'을 붙임
- 아래와 같은 메시지가 출력되면 [Save] 클릭

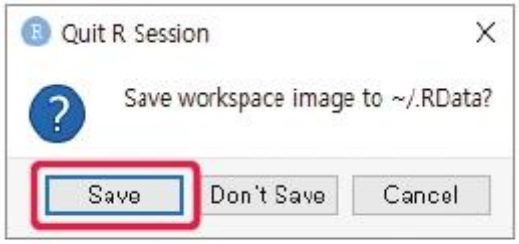


그림 1-16 R스튜디오 종료 대화상자

5.4 패키지의 설치

- R에서는 데이터 분석을 위해서 매우 다양한 함수들을 제공
- 패키지(package) 는 이러한 함수들을 기능별로 묶어놓은 '꾸러미'
- 어떤 함수를 이용하기 위해서는 그 함수를 포함하고 있는 패키지를 사전에 설치해야 함

그림 1-17 현재 설치된 패키지 목록

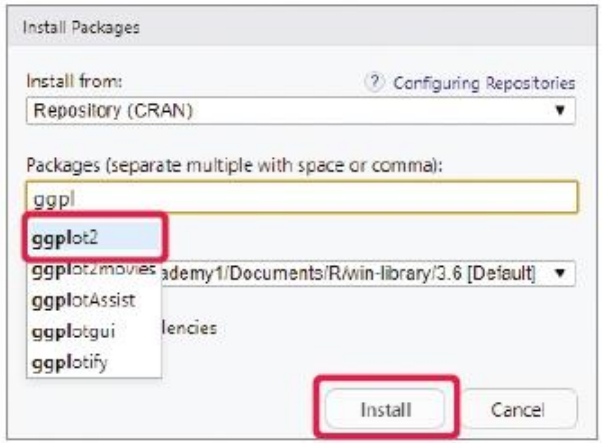


그림 1-18 패키지 설치 윈도우

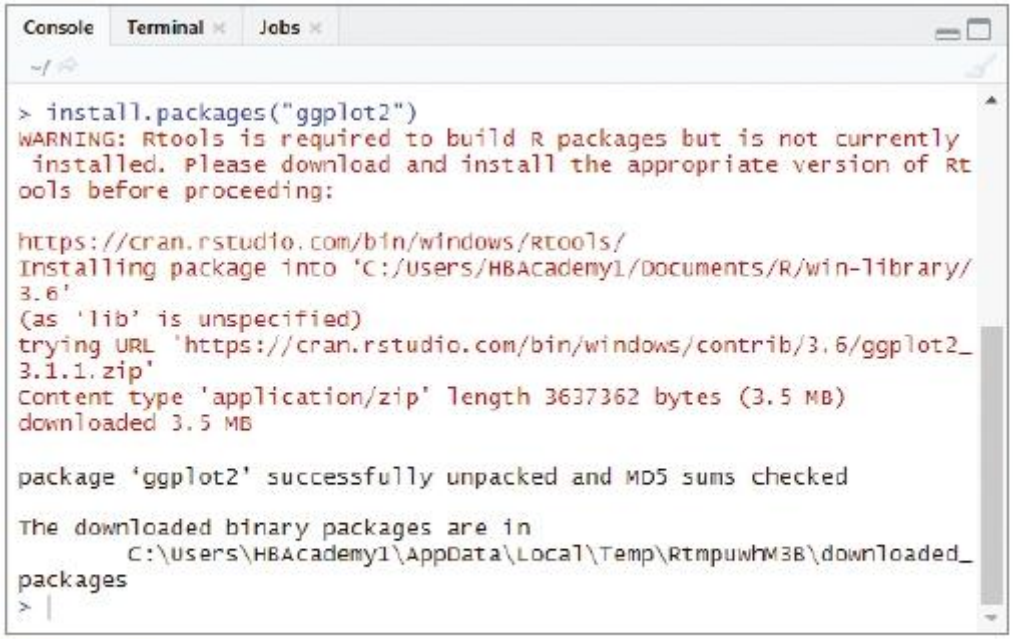


그림 1-19 패키지 설치가 성공한 경우의 일반적인 화면

- 설치한 패키지 불러오기

```
library(ggplot2)
```

여기서 잠깐! R과 R스튜디오 설치 시 주의 사항

- R과 R스튜디오 설치 시 보통 문제없이 설치되지만 가끔 에러가 발생함.
  이러한 에러를 방지하기 위해 다음 사항들을 사전에 점검.

① 현재 로그인한 윈도우의 계정 이름은 한글이 아니어야 함.
한글 계정명을 사용하는 경우는 영문 계정을 새로 만들고,
영문 계정으로 로그인 후 설치 작업을 해야 함.

② R이나 R스튜디오를 설치하는 폴더의 경로에 한글이 포함되지 않도록 함.

③ 원 드라이브나 구글 드라이브와 같이 클라우드와 연동된 폴더에 설치하는 것은 피하기.

Thank you!

한빛아카데미 Hanbit Academy, Inc.