


대표적인 빅데이터 분석 모델

Description (기술)	<ul style="list-style-type: none"> ▪ 탐색적 데이터 분석 ▪ 대량 데이터의 전반적인 형태를 조사하고 데이터의 변환과 축약
Classification (분류)	<ul style="list-style-type: none"> ▪ 판별 분석 (Discriminant analysis) ▪ 계획된 기계학습 (Supervised learning) ▪ 기존 데이터를 통해 학습함으로써, 새로운 데이터가 실제로 어떤 그룹에 속해 있는지 분류하는 기법 ▪ 데이터의 실체가 어떤 그룹에 속하는지 예측 ▪ 객체를 정해놓은 범주로 분류하는데 목적
Clustering (군집)	<ul style="list-style-type: none"> ▪ 비계획된 기계학습 (Unsupervised learning) ▪ 특성에 따라 고객을 여러 개의 배타집단으로 나누는 것
Association (연관)	<ul style="list-style-type: none"> ▪ 장바구니분석 (Market Basket Analysis) ▪ 서열분석 (Sefluence Analysis), 시차분석 ▪ 데이터에서 빈발하는 속성을 찾고 그 중에서 서로 연관이 있는 규칙을 발견하는 기법
Estimation (예측)	<ul style="list-style-type: none"> ▪ Estimation (예측, 추정) <ul style="list-style-type: none"> ▪ Prediction (예측, 예상) : 미래에 발생할 값을 예측 ▪ Forecasting (예측) : 과거 또는 미래의 모르는 값을 예측 ▪ 연속적인 값을 예측하는 것 ▪ 시계열 변수를 이용한 예측, 인과관계 모형으로 예측

A decorative graphic in the bottom-left corner of the slide. It consists of a grid of squares, each divided into two triangles by a diagonal line. The colors of the triangles and squares transition from red and blue at the top-left, through light blue, green, yellow, and orange, to pink and purple at the bottom-right. The pattern is partially cut off by the edges of the slide.

R





1 장. BigData 분석 환경 구 성

BigData 분석 환경 구성

1. BigData 분석 환경 구성

- R의 공식 사이트: <https://www.r-project.org/>
- 데이터 분석을 위한 통계 및 그래픽스를 지원하는 자유 소프트웨어 환경
- 벨 연구소의 S언어에 기반
- 데이터 분석 소프트웨어
- 완성된 언어 체계
- 무료
- 멀티 프로세서에서 병렬화 실행
- Hadoop, Hive 환경에서 R 사용 가능



Version	Nickname(codename)	Released
3.1.3	Smooth Sidewalk	2015-03-09
3.2.3	Wooden Christmas-Tree	2015-12-10
3.2.5	Very, Very Secure Dishes	2016-04-14
3.3.3	Another Canoe	2017-03-06
3.4.0	You Stupid Darkness	2017-04-21
3.4.1	Single Candle	2017-06-30
3.4.2	Short Summer	2017-09-28
3.4.3	Kite-Eating Tree	2017-11-30
3.4.4	Someone to Lean On	2018-03-15
3.5.0	Joy in Playing	2018-04-23

BigData 분석 환경 구성

1. BigData 분석 환경 구성

R Language로 BigData 분석을 하기 위해 R을 설치합니다.
통합 개발 환경(IDE)에서 좀 더 편리하게 분석 작업을 하기 위해서 RStudio를 설치합니다.

교실에서는 Windows 10을 기준으로 환경 구성을 진행합니다.



Linux, Windows, Mac OS X를 지원합니다.

다운로드 : <https://cran.r-project.org/>

R의 통합 개발 환경(IDE)를 제공하며,
GNU AGPLv3를 지원하는 Open Source Edition과 상용 제품이
있습니다.



RStudio Desktop : PC 환경의 IDE 제공

RStudio Server : Server 환경의 IDE 제공, 브라우저로 접속

다운로드 : <http://www.rstudio.com/products/rstudio/>

BigData 분석 환경 구성 - R 설치

1. BigData 분석 환경 구성

• R 설치 파일 다운로드

- 다운로드 사이트 : <http://cran.rstudio.com/> 또는 <https://cran.r-project.org/>

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- [Download R for Linux](#)
- [Download R for \(Mac\) OS X](#)
- [Download R for Windows](#)

R is part of many Linux distributions, you should check with your Linux package management Subdirectories:

[base](#)

Binaries for base distribution (managed by Duncan Murdoch). This is what you want to **install R for the first time**.

[contrib](#)

Binaries of contributed packages (managed by Uwe Ligges). There is also information on [third party software](#) available for CRAN Windows services and corresponding environment and make variables.

[Rtools](#)

Tools to build R and R packages (managed by Duncan Murdoch). This is

R-3.2.1 for Windows (32/64 bit)

[Download R 3.2.1 for Windows](#) (62 megabytes, 32/64 bit)

[Installation and other instructions](#)

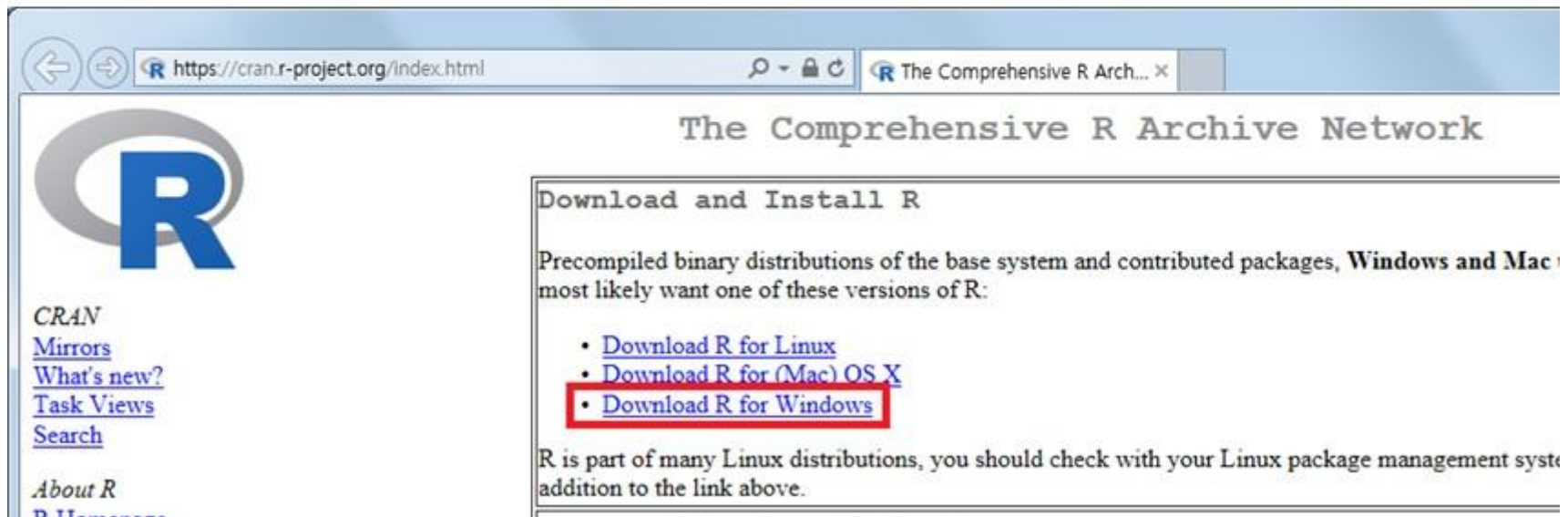
[New features in this version](#)

R 다운로드

1. BigData 분석 환경 구성

R은 Linux(Debian, Red Hat, Suse, Ubuntu), (Mac)OS X, Windows를 지원합니다.

R 다운로드는 <https://cran.r-project.org/index.html>에서 합니다.

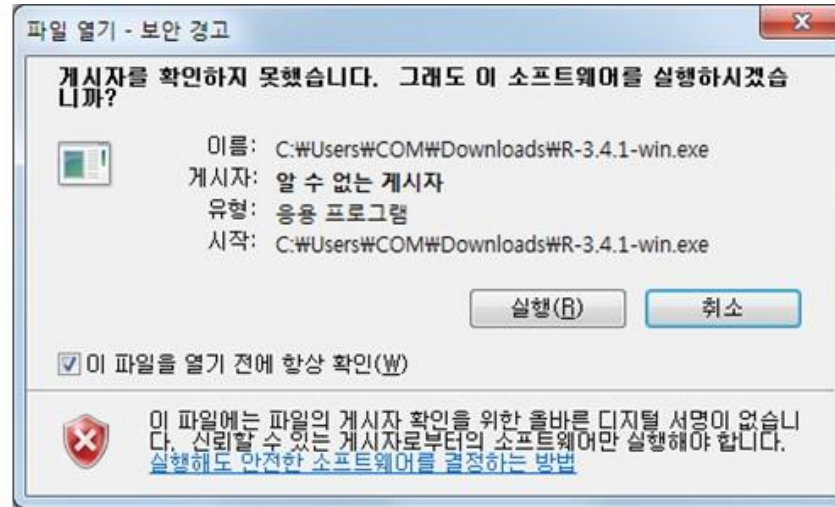
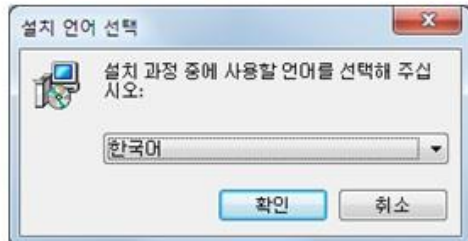


R 다운로드

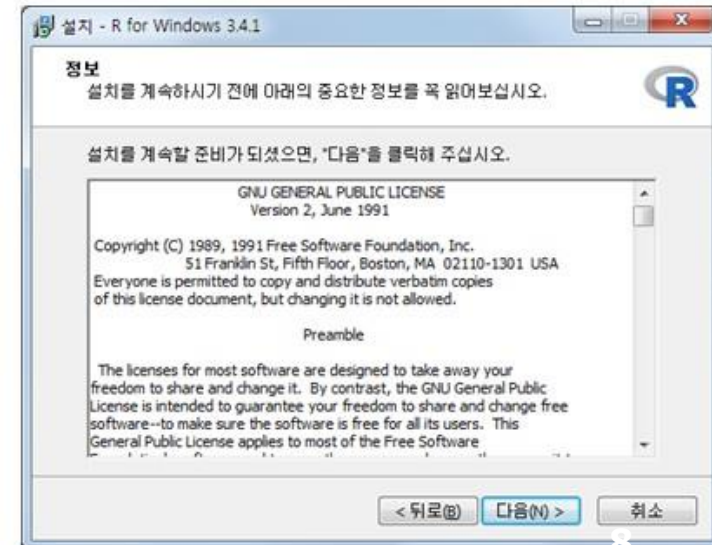
1. BigData 분석 환경 구성

1. 다운로드 받은 설치파일을 더블클릭
만 하면 설치를 시작합니다.

2. 설치 언어는 한국어를 선택합니다.



3. 이후 [다음(N)>] 버튼만 클릭하면 쉽게 설치됩니다.



RStudio 다운로드

1. BigData 분석 환경 구성

RStudio는 R을 위한 무료 및 오픈 소스 통합 개발 환경(IDE)입니다. RStudio는 ColdFusion 프로그래밍 언어의 창시자 JJ 앨라이어(Allaire)에 의해 설립되었습니다.



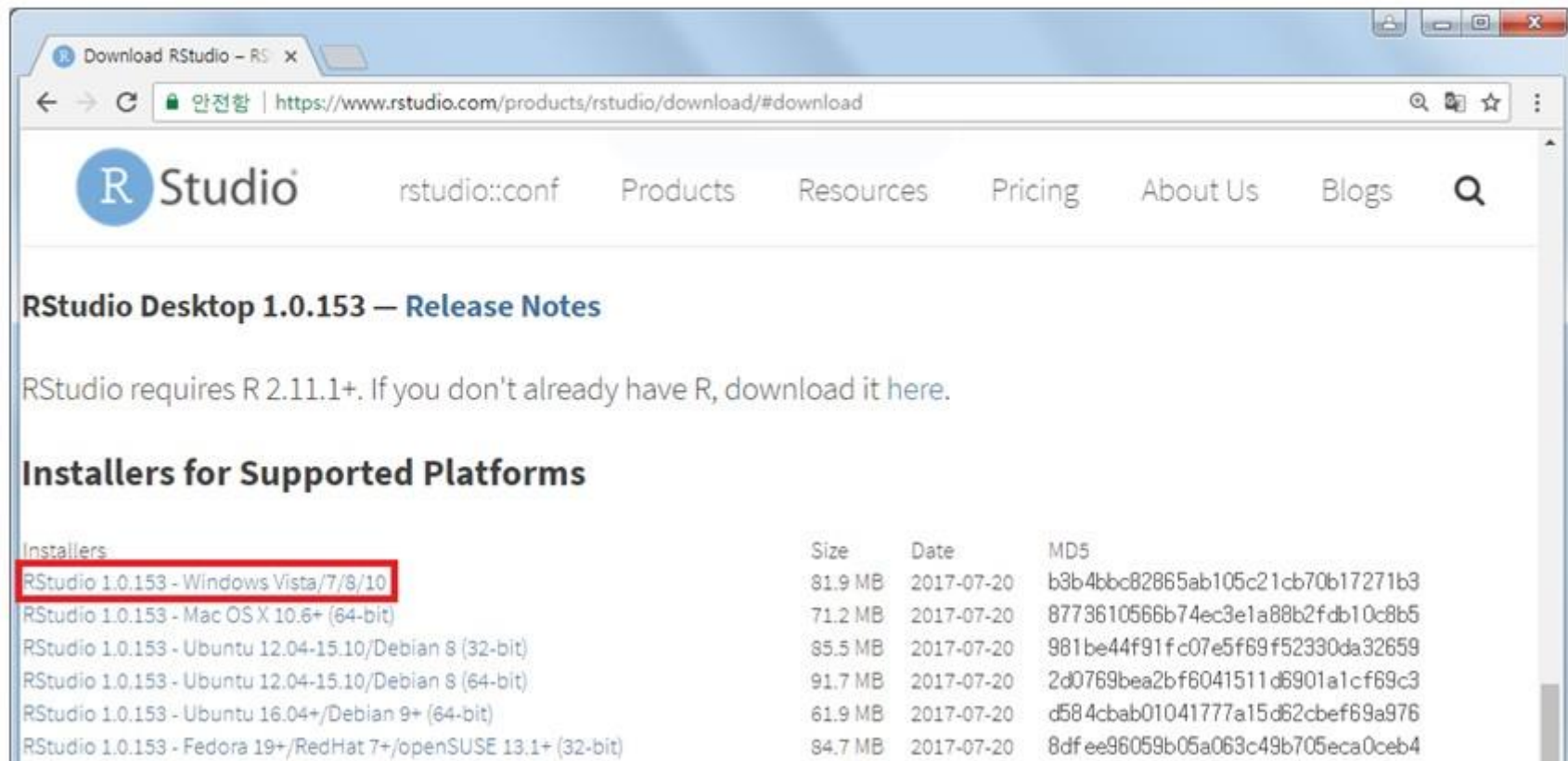
RStudio는 두 가지 버전으로 제공됩니다. RStudio Desktop은 프로그램이 로컬 데스크톱 응용 프로그램으로 실행됩니다. RStudio Server는 원격 Linux 서버에서 실행되는 동안 웹 브라우저를 사용하여 RStudio에 액세스 할 수 있게 합니다. RStudio Desktop의 배포판은 Windows, MacOS 및 Linux에서 사용할 수 있습니다 .

RStudio는 GNU AGPLv3를 지원하는 오픈소스 버전 및 상용 버전으로 제공되며 데스크톱 (Windows, MacOS 및 Linux) 또는 RStudio Server 또는 RStudio Server Pro (Debian, Ubuntu, Red Hat Linux, CentOS, openSUSE 및 SLES(SUSE Linux Enterprise Server))에 연결 된 브라우저에서 실행됩니다.

RStudio 다운로드

1. BigData 분석 환경 구성

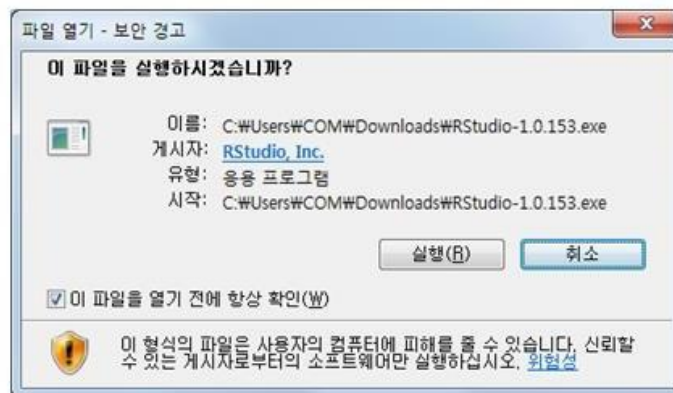
<https://www.rstudio.com/products/rstudio/download/#download>에 접속하여 RStudio 1.0.x Windows Vista/7/8/10을 다운로드 받습니다.



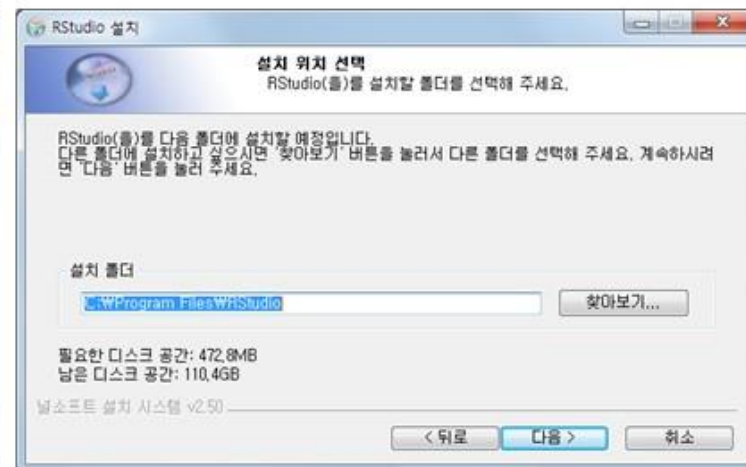
RStudio 다운로드

1. BigData 분석 환경 구성

다운로드 한 설치파일을 더블클릭하면 쉽게 설치할 수 있습니다. 파일 열기 보안 경고창이 뜨면 [실행]버튼을 클릭하세요.



RStudio는 쉽게 설치할 수 있습니다. 설치 위치와 시작 메뉴 폴더의 위치 등은 기본값으로 두고 설치하세요.



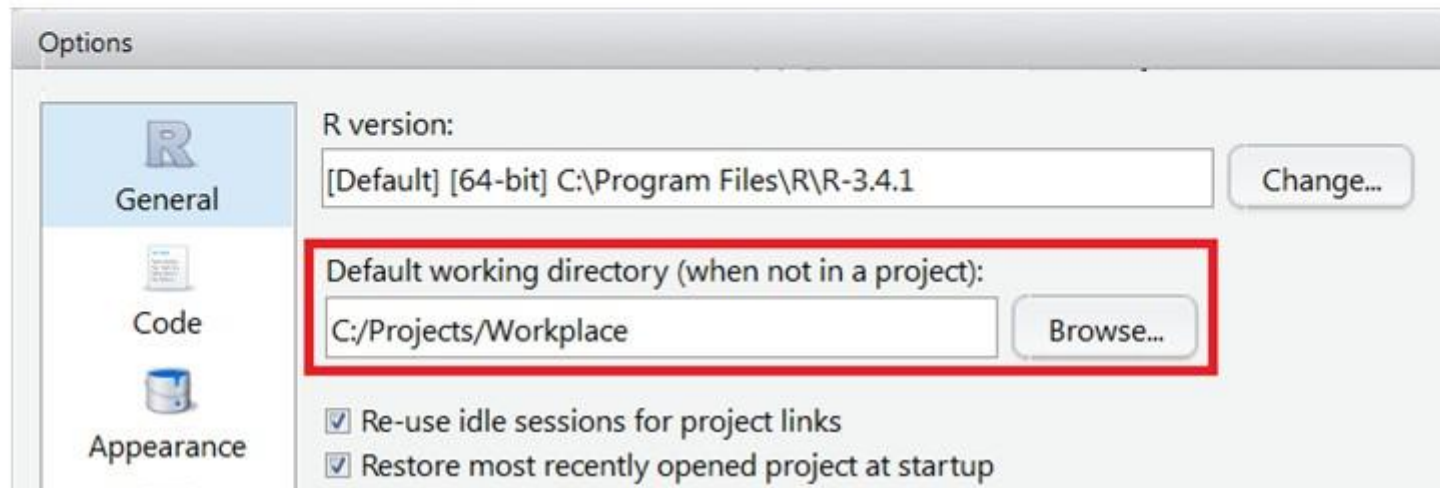
Rstudio 설정

1. BigData 분석 환경 구성

RStudio의 Tools -> Global Options... 메뉴를 선택하고 몇 가지 설정을 할 필요가 있습니다.

- 현재 설치되어 있는 [R version]을 확인하세요.
- [Default working directory] 기본값은 ~(내 문서)로 되어 있습니다.

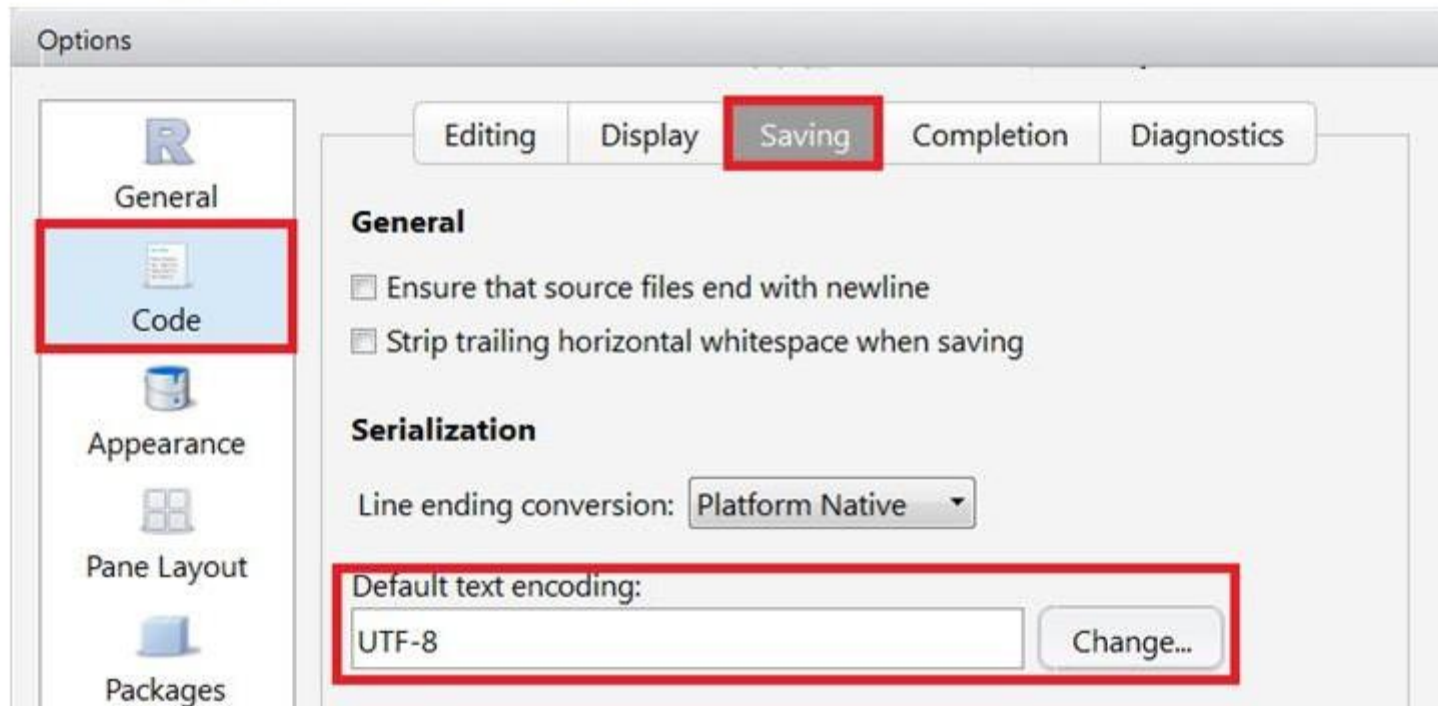
필자는 C:/Projects/Workplace 디렉토리로 변경했습니다.(반드시 작업 디렉토리를 변경할 필요는 없습니다. 기본값을 그대로 사용해도 됩니다.)



RStudio 설정

1. BigData 분석 환경 구성

[Code] 옵션의 [Saving] 탭에서 [Default text encoding:]을 UTF-8로 바꿔주세요. 제공되는 소스코드의 인코딩이 UTF-8로 되어 있습니다.



RStudio 화면 구성

1. BigData 분석 환경 구성

• RStudio 화면 구성

The screenshot displays the RStudio integrated development environment (IDE) with four main panels highlighted by red boxes:

- Source:** The top-left panel shows the R script editor. It contains code for reading CSV files and checking for missing values. The code is as follows:

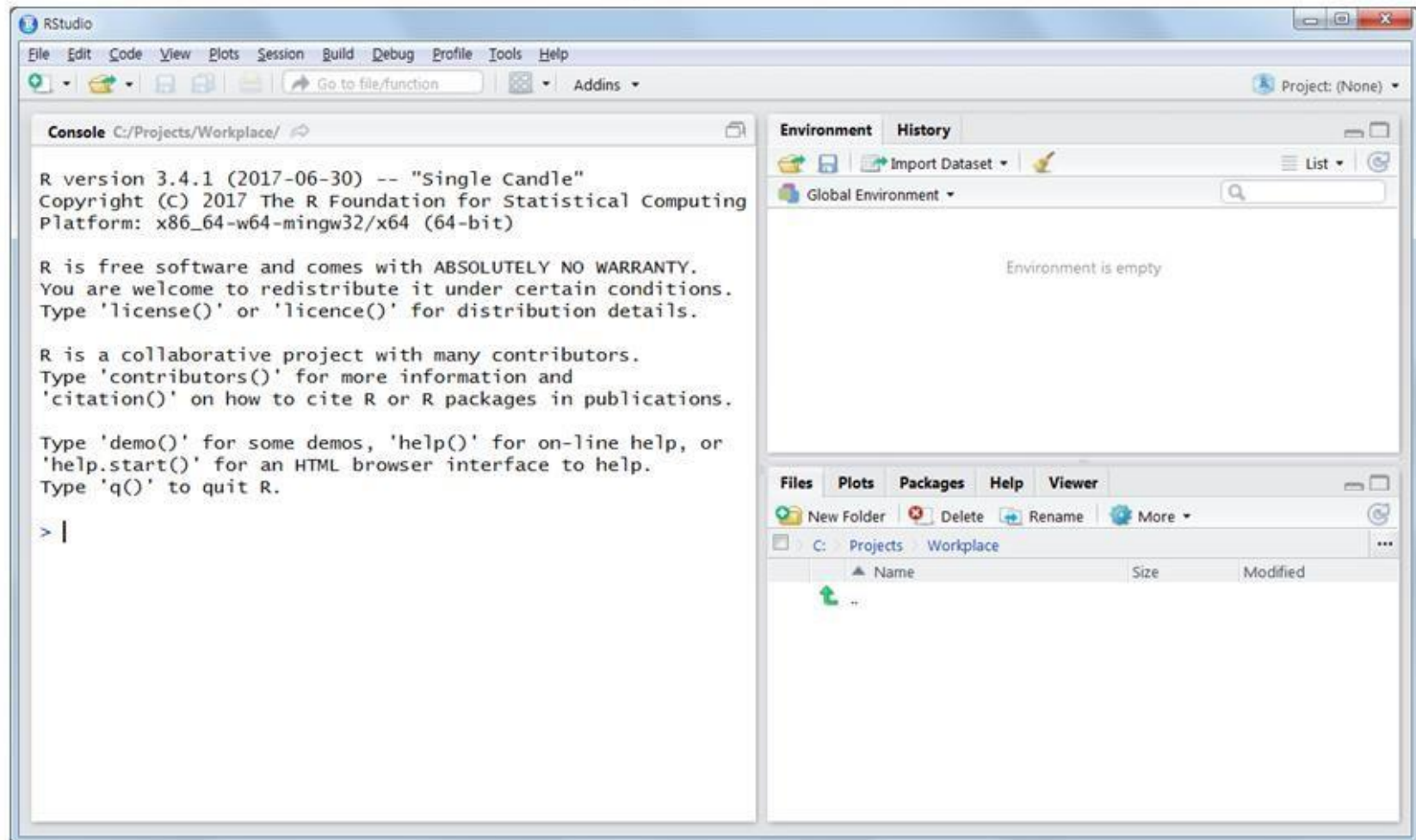
```
1 # -----  
2 # Difference among 0, missing-value and NA (만 사용)  
3 # -----  
4  
5 #--- incomeshort.csv  
6 income = read.csv("C:/001_work/RStudioworkplace/incomeshort.csv",  
7                   header=TRUE, sep="\t")  
8 income  
9 str(income)  
10 colSums(is.na(income))  
11 ? read.csv  
12  
13 income = read.csv("C:/001_work/RStudioworkplace/incomeshort.csv",  
14                  header=TRUE, sep="\t", na.strings="")  
15 income  
16 str(income)  
17 colSums(is.na(income))  
18  
19 #--- incomeshort_Missing.csv  
20 income = read.csv("C:/001_work/RStudioworkplace/incomeshort_Missing.csv",
```
- Environment:** The top-right panel shows the Global Environment. It displays a data frame named 'income' with 4 observations and 2 variables.
- Console:** The bottom-left panel shows the R console output. It displays the results of the commands entered in the Source panel, including the structure of the 'income' data frame and the summary of missing values.

```
$ age: int  20 20 55 0  
$ sex: Factor w/ 2 levels "Female","Male": 1 2 2 1  
> colSums(is.na(income))  
age sex  
0 0  
> ? read.csv  
> income = read.csv("C:/001_work/RStudioworkplace/incomeshort.csv",  
+ header=TRUE, sep="\t", na.strings="")  
> income  
  age sex  
1  20 Female  
2  20  Male  
3  55  Male  
4   0 Female  
> str(income)  
'data.frame': 4 obs. of  2 variables:  
 $ age: int  20 20 55 0  
 $ sex: Factor w/ 2 levels "Female","Male": 1 2 2 1  
> colSums(is.na(income))  
age sex  
0 0  
> view(income)  
> view(income)  
>
```
- Help:** The bottom-right panel shows the R Documentation for the 'read.table' function. It includes a description and usage examples.

RStudio 설정


1. BigData 분석 환경 구성

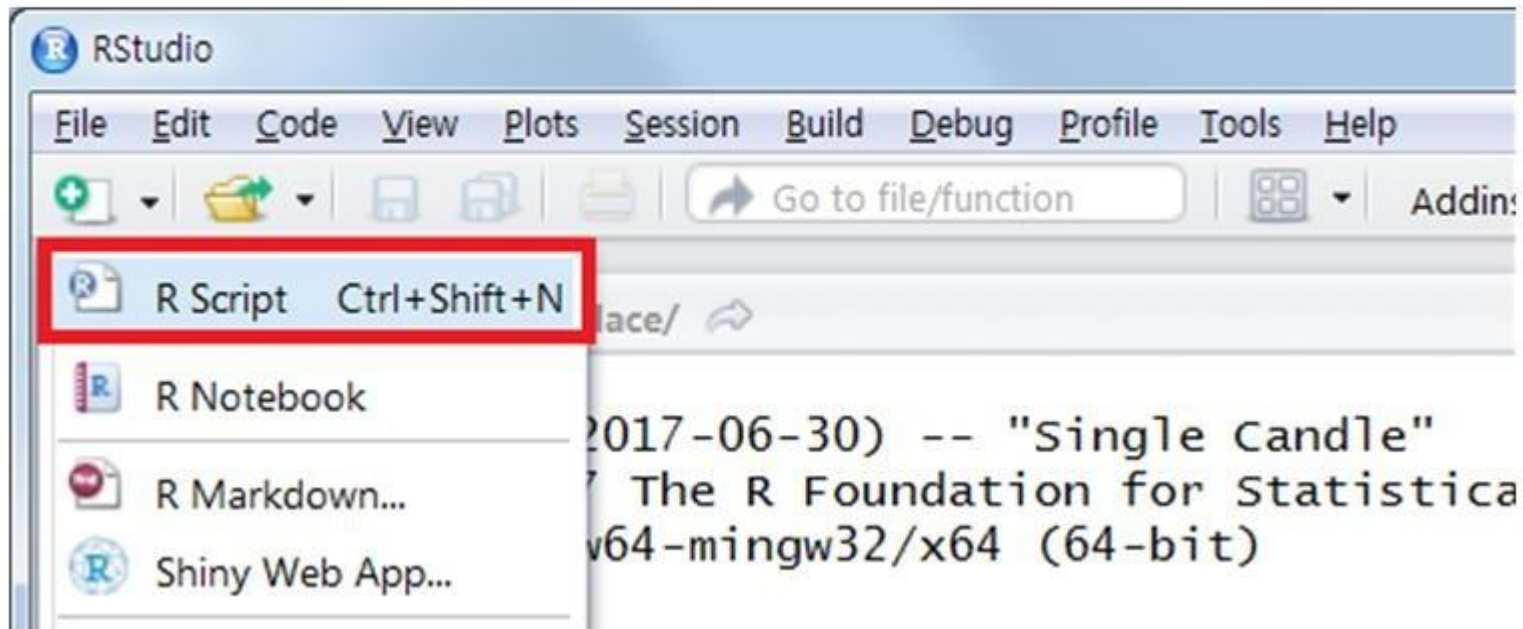
다음 그림은 R을 처음 실행시켰을 때의 화면입니다.



RStudio 설정

1. BigData 분석 환경 구성

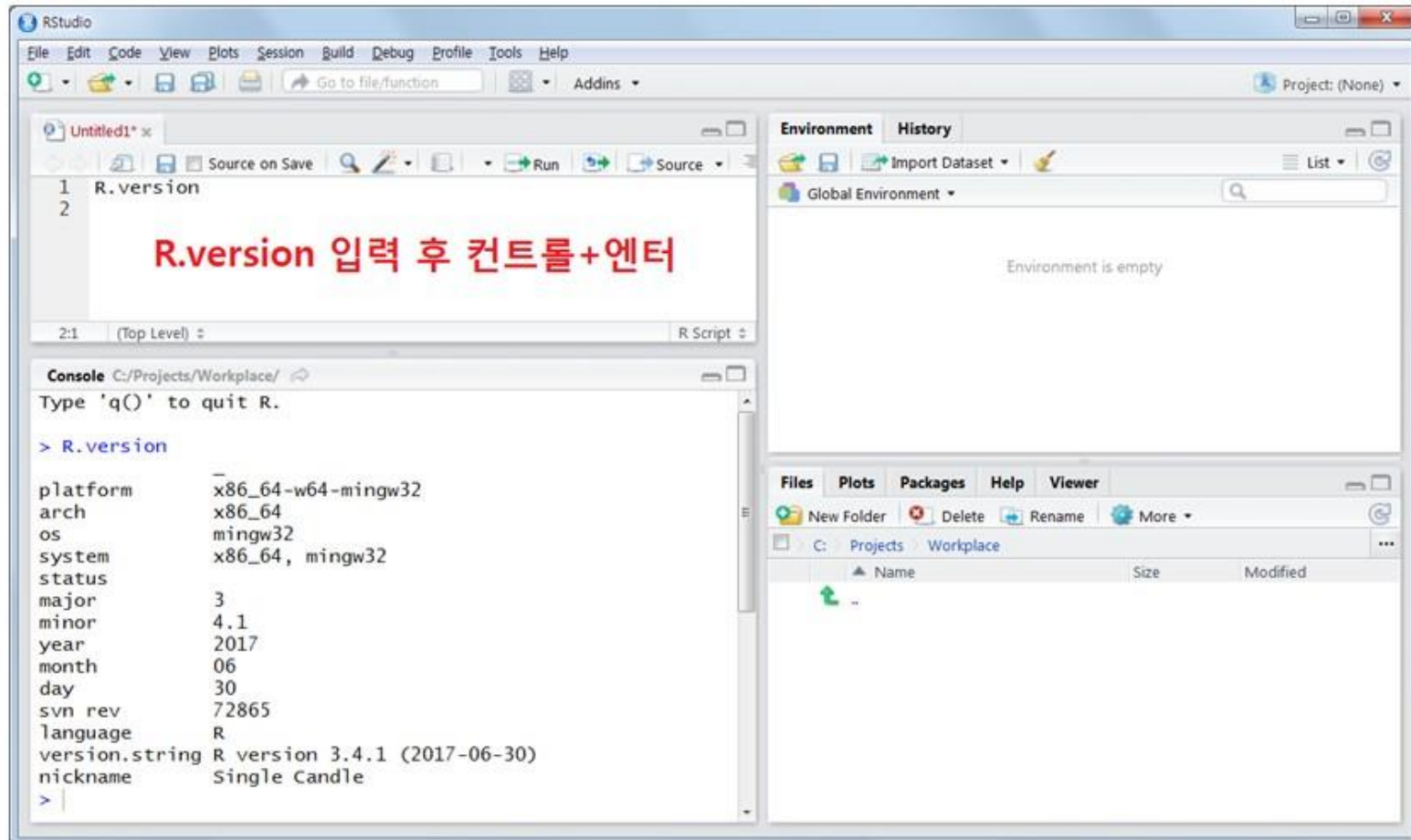
RStudio를 처음 실행시킨 후  아이콘을 클릭하고 [R Script] 메뉴를 선택하면 R 스크립트를 입력할 수 있는 새로운 레이아웃이 나타납니다.



RStudio 설정

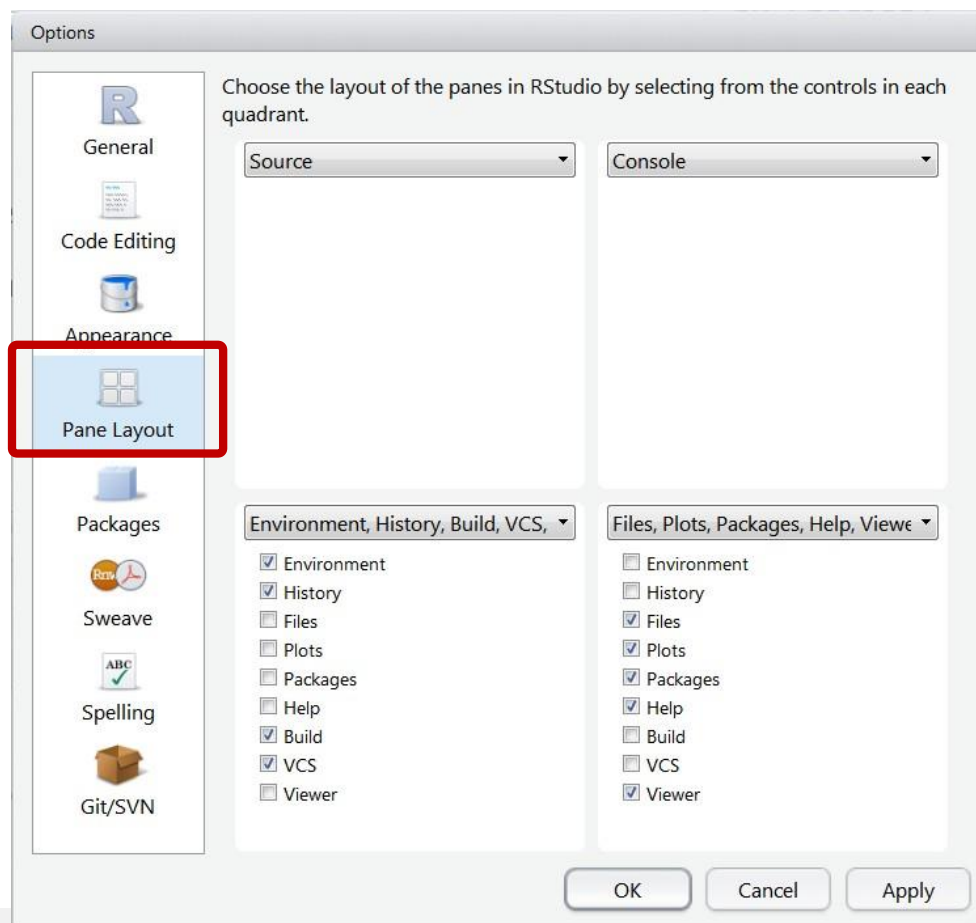
1. BigData 분석 환경 구성

R 스크립트 창에 입력한 스크립트를 실행시키기 위해서는 실행시키고자 하는 행에 커서를 두거나 블록을 설정한 다음 Ctrl+Enter키를 누르거나 Ctrl+R키를 누르면 됩니다.



RStudio 화면 레이아웃 변경

- RStudio 화면 레이아웃을 변경할 수 있다.
 - 메뉴 -> Tools -> Global Options 에서 Pane Layout을 선택하고 각 패널에 보여질 항목을 선택한다.



BigData 분석 환경 구성 실습

- R을 설치하고 환경을 구성합니다.
- RStudio를 설치하고 환경을 구성합니다.