

Chapter 1

# Introduction





Sejong Oh

Bio Information technology Lab.

- ▶ 우리는 데이터의 시대(the age of data)에 살고 있음
- 우리를 둘러싼 모든 것들이 데이터 소스와 연결되고, 우리 삶의 많은 부 분들이 디지털 데이터화 되고 있음
  - 이메일, SNS, 전화사용 기록, 신용카드거래 기록, 병원 치료 기록, 성적
  - 인터넷, 주민정보, 등기정보, 판매정보, 주식거래 정보
- 데이터의 증가 속도는 점점 빨라지고 있으며, 전통적인 방법으로는 다루 기 힘든 규모가 되고 있다
  - 빅데이터
- 정치,경제, 사회, 문화 등 모든 영역이 데이터와 연관





데이터의 폭발적인 증가



100 Million Tweets are posted every day, 35 hours of video are being uploaded every minute, 6.1 x 10 $^1$ 2 text messages have been sent in 2011 and 247 x 10 $^2$ 9 E-Mails passed through the net.80 % spam and viruses. => Filtering is more and more important.

Up to 2003 the same amount of data has been produced as between 2003 and now



- "Data are becoming the new raw material of business" Craig Mundie, Microsoft
- "Data tab has been turned on and will never be turned off" Mike Hoskins, Actian CTO, 2013
- "You can have data without information, but you cannot have information without data" – Writer
- "It is a capital mistake to theorize before one has data" Sherlock Homes
- "Data is the new oil" Clive Humby, ANA Senior marketer's submit



데이터를 잘 다룰 줄 아는 기업, 개인이 성공하는 시대가 도래



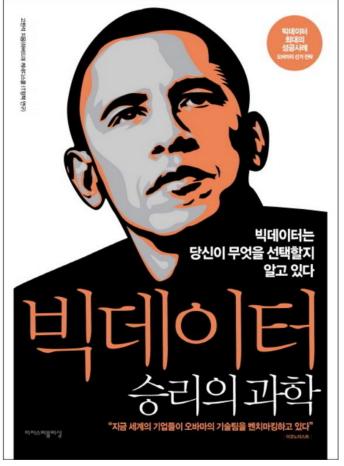
한 남성이 화가 잔뜩 난 채로 대형마트에 들어서더니 "매니저 나와"라고 소리 질렀다. 손에는 이 대형마트가 딸에게 우편으로 보낸 아기 옷과 침대 등 유아용품 할인쿠폰이 들려 있었다. 남성은 "내 딸은 아직 고등학생인데 이런 쿠폰을 보내다니. 임신을 부추기는 거냐"고 따졌다. 영문을 몰 랐던 매니저는 우선 남성을 진정시키고 돌려보냈다. 며칠 뒤 재차 사과를 하려고 전화를 건 매니 저에게 뜻밖의 반응이 돌아왔다. 이 남성은 "딸이 임신을 한 걸 뒤늦게 알게 됐다"며 사과했다. 미 성년 딸이 부모 모르게 아이를 가졌던 것이다. 대형마트 마케팅팀은 부모도 몰랐던 딸의 임신 사 실을 어떻게 알아챘을까? 답은 '빅데이터 분석'에 있다. 이 얘기는 미국 대형마트 '타깃'의 미니애 폴리스 점포에서 실제로 일어났던 일이다. (뉴욕 타임스)가 보도한 이 사례는 빅데이터 분석의 힘을 여실히 보여준다. 대형마트는 고객이 가족에게도 알리지 않은 비밀을 찾아내 앞으로 무엇이 필요할지 정확하게 예측해냈다.

타깃의 빅데이터 전문가들은 고객의 25가지 구매 행태를 분석하면 여성의 임신과 출산을 상당히 정확하게 예측할 수 있다는 사실을 확인했다. 예를 들어 향이 나는 로션을 사던 여성이 무향의 로션으로 바꾸거나, 평소 사지 않던 미네랄 영양제를 갑자기 사들이는 경우다. 타깃은 고객 데이터 베이스에 이를 적용했고, 전국적으로 수만명의 임신 추정 여성들을 가려내 관련 할인쿠폰을 보냈다. 이를 통해 고객의 구매 행동을 이끌어낼 수 있기 때문이다.

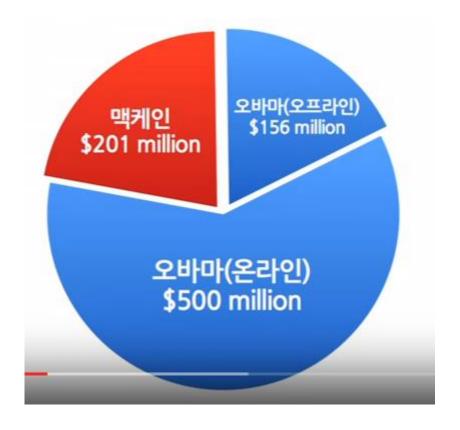


● 오바마 : 데이터베이스에 기초한 개인 맞춤형 선거 홍보

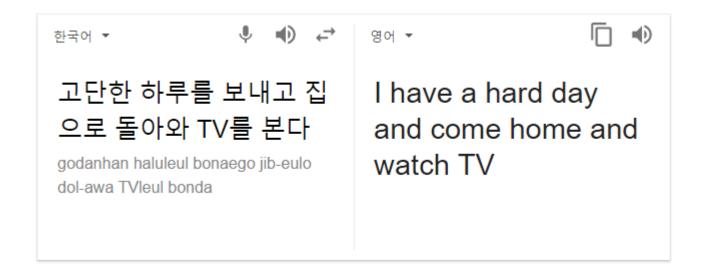




• 후원금 모금액



• 구글 자동번역



과거: 형태소 분석, 단어 by 단어로 번역

구글 : 번역된 예문 데이터를 기반으로 가장 적합한 번역 사례를 찾음

데이터가 쌓이면 쌓일수록 번역이 정교해짐





Alphago

Google car

Alphago 의 힘은 어디서 오는가: data + algorithm

구글 자율주행차: image, sensor data 의 분석, 처리를 통한 주행

• 빅데이터(유튜브) - 세계미래포럼



https://www.youtube.com/watch?v=X4hMFym0-uo

## 빅데이터\*

데이터의 규모에 초점을 맞춘 정의

기존 데이터베이스 관리도구의 데이터 수집, 저장, 관리, 분석 역량을 넘어서는 데이터 (맥킨지 2011년 5월)

• 업무 수행 방식에 초점을 맞춘 정의

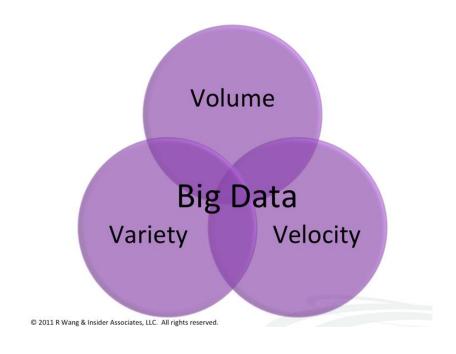
다양한 종류의 대규모 데이터로부터 저렴한 비용으로 가치를 추출하고, 데이터의 빠른 수집, 발굴, 분석을 지원하도록 고안된 차세대 기술 및 아 키텍처 (IDC 2011년 6월)

<sup>\*</sup> Yonsei University 2<sup>nd</sup> Semester, 2014 Sanghyun Park



- 빅데이터의 출현 배경
  - 스마트폰과 같은 디지털 기기의 보급과 소셜 네트워크서비스(SNS)의 부상으로 엄청난 양의 데이터가 생산되고 있음
  - 의료분야의 환자 데이터, 은행권의 거래 데이터, 교통 이용 데이터 등 도 빅데이터임
  - 빅데이터는 기업, 정부, 의료, 교육 분야에서 그 가치가 입증되고 있음
  - 빅데이터를 적극적으로 활용하는 기업은 비즈니스 성과를 개선할 수 있음

 BI/DW 리서치 기관인 TDWI 가 빅데이터의 3대 요소(3V)를 그림으로 표현



○ 3대 요소 가운데 두 가지 이상의 요소만 충족된다면 빅데이터라고 볼 수 있음



[출처:채승병 외, 빅데이터: 산업 지각변동의 진원,SERI CEO Information, 2012. 5]

- 크기(Volume)
  - 일반적으로 수십 테라바이트(terabyte) 혹은 수십 페타바이트 (petabyte) 이상이 빅데이터의 범위에 해당
  - 1 테라바이트는 1024 기가바이트(gigabyte)
  - 1 페타바이트는 1024 테라바이트
  - 1 페타바이트는 6 기가바이트 DVD 영화를 17만 4천 편 담을 수 있는 용량
  - 빅데이터는 기존 파일 시스템에 저장하기 어려울 뿐 아니라 데이터 분석을 위해 사용하는 기존 DW 같은 솔루션에서 소화하기 어려울 정 도로 급격하게 데이터의 양이 증가하고 있음



- 속도(Velocity)
  - 빅데이터의 속도적인 특징은 크게 실시간 처리와 장기적인 접근으로 나눌 수 있음
  - 오늘날 디지털 데이터는 매우 빠른 속도로 생성되기 때문에 데이터의 수집, 저장, 분석 등이 실시간으로 처리돼야 함
  - 모든 데이터가 실시간 처리만을 요구하는 것은 아님
  - 수집된 대량의 데이터를 다양한 분석 기법과 표현 기술로 분석해야 하는데, 이는 장기적이고 전략적인 차원에서 접근할 필요가 있음
  - 통계학과 전산학에서 사용하는 데이터 마이닝, 기계 학습, 자연어 처리, 패턴 인식 등이 분석 기법에 해당함



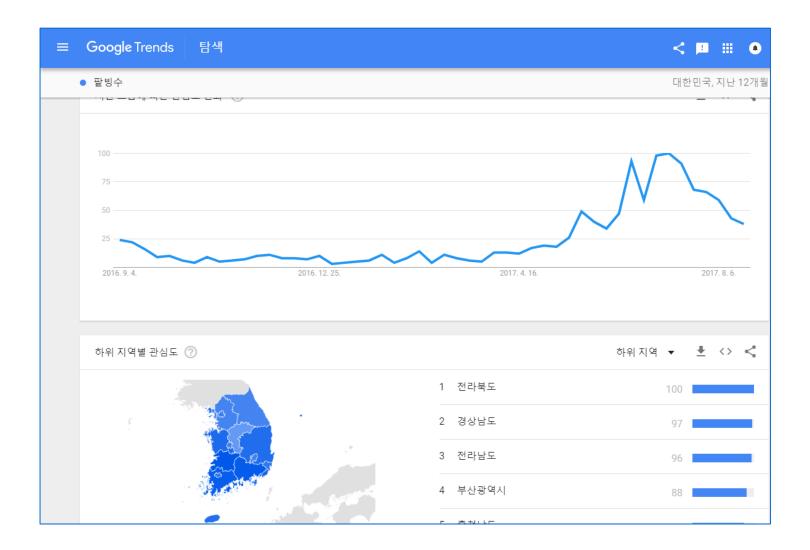
- 다양성(Variety)
  - 데이터는 정형화 정도에 따라 정형(Structured), 반정형(Semi-Structured), 비정형(Unstructured)으로 구분
  - 정형 데이터는 고정된 필드에 저장되는 데이터를 의미하며, 일정한 형식을 갖추고 있음
  - 반정형 데이터는 고정된 필드로 저장되지는 않지만, XML 이나 HTML 같이 메타데이터나 스키마 등을 포함
  - 비정형 데이터는 고정된 필드에 저장되지 않는 데이터를 의미함. 사진, 동영상, 메신저로 주고 받은 대화 내용, 스마트폰에 기록되는 위치정보, 통화 내용 등이 이에 해당

- 공공데이터 포털
  - https://www.data.go.kr/



# Google Trend

https://trends.google.co.kr/



## 강의 내용

- R 언어 기초
- 데이터 manipulation
  - 일변량 자료의 정리및 요약
  - 다변량 자료의 정리및 요약
  - 선형관계의 추정
- ▶ 데이터 시각화
  - 그래프
  - 지도상에 정보표현
  - 워드 클라우드
- 군집분석, 분류분석
- R 프로그래밍

## Syllabus

- Text /Reference
  - R을 이용한 데이터 처리 & 분석 실무, 길벗
  - 데이터 시각화, 자유 아카데미
- Evaluation

중간고사 : 30%

• 기말고사 : 30%

• 과제 : 20%

출석 : 20%

- Class materials
  - 단국대 이러닝 사이트 참조



https://www.r-project.org/



[Home]

Download

CRAN

#### R Project

About R Logo Contributors What's New? Reporting Bugs Development Site Conferences Search

# The R Project for Statistical Computing

#### Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To download R, please choose your preferred CRAN mirror.

If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

#### News

- useR! 2017 (July 4 7 in Brussels) has opened registration and more at http://user2017.brussels/
- Tomas Kalibera has joined the R core team.

## • 다운로드 사이트 선택

http://cran.stat.unipd.it/	University of Padua		
http://dssm.unipa.it/CRAN/	Universita degli Studi di Palermo		
Japan			
https://cran.ism.ac.jp/	The Institute of Statistical Mathematics, Tokyo		
http://cran.ism.ac.jp/	The Institute of Statistical Mathematics, Tokyo		
Korea			
http://cran.nexr.com/	NexR Corporation, Seoul		
http://healthstat.snu.ac.kr/CRAN/	Graduate School of Public Health, Seoul National		
http://cran.biodisk.org/	The Genome Institute of UNIST (Ulsan National In and Technology)		
Lebanon			
http://rmirror.lau.edu.lb/	Lebanese American University, Byblos		
Malaysia			
https://wbc.upm.edu.my/cran/	Univerisiti Putra Malaysia		
http://wbc.upm.edu.my/cran/	Univerisiti Putra Malaysia		

#### • 윈도우용 버전 선택

#### The Comprehensive R Archive Network

#### 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 system in addition to the link above.

Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

• The latest release (Monday 2016-10-31, Sincere Pumpkin Patch) R-3.3.2.tar.gz, read

#### base 버전 선택

#### R for Windows

#### Subdirectories:

Binaries for base distribution (managed by Duncan Murdoch). This is what you want

to install R for the first time.

Binaries of contributed CRAN packages (for R >= 2.11.x; managed by Uwe Ligges).

<u>contrib</u> There is also information on <u>third party software</u> available for CRAN Windows

services and corresponding environment and make variables.

old contrib

Binaries of contributed CRAN packages for outdated versions of R (for R < 2.11.x;

managed by Uwe Ligges).

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

want to build your own packages on Windows, or to build R itself.

Please do not submit binaries to CRAN. Package developers might want to contact Duncan Murdoch or Uwe Ligges directly in case of questions / suggestions related to Windows binaries.

현재 최신 버전은 R3.4.3

Download R 3.4.3 for Windows (62 megabytes, 32/64 bit)

<u>Installation and other instructions</u> <u>New features in this version</u>

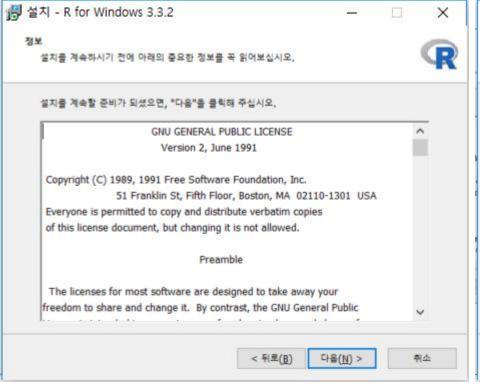
If you want to double-check that the package you have downloaded matches the package distributed by CRA md5sum of the .exe to the <u>fingerprint</u> on the master server. You will need a version of md5sum for windows: <u>command line versions</u> are available.

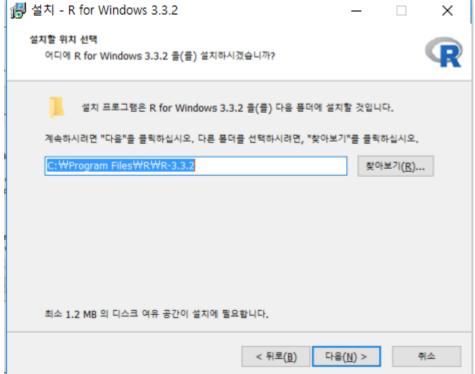
Frequently asked questions

- Does R run under my version of Windows?
- How do I update packages in my previous version of R?
- Should I run 32-bit or 64-bit R?

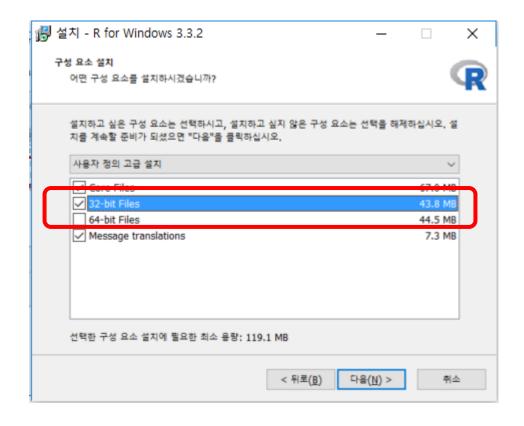
• 다운로드 받은 설치 프로그램 실행



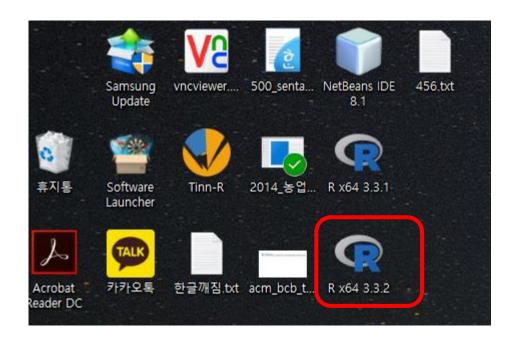




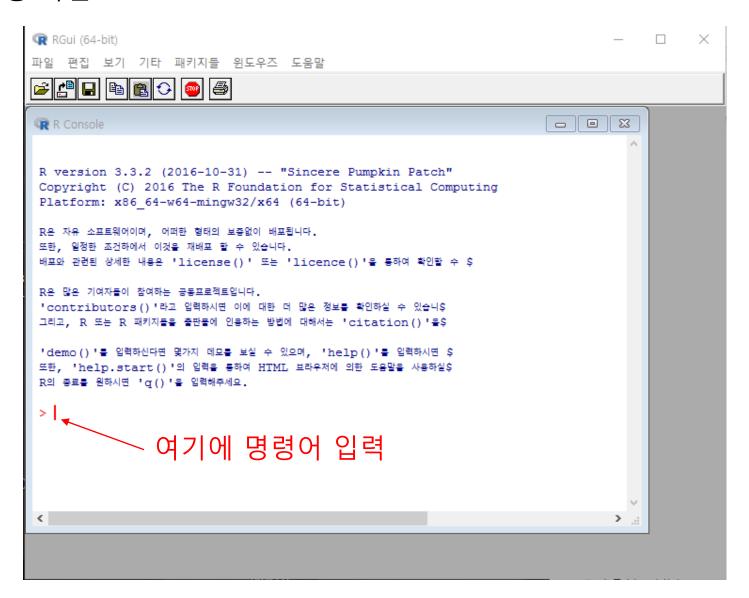
- 32 bit 또는 64 bit 버전 중 하나를 설치한다.
- 운영체제의 버전에 따라 선택. 64 bit 버전이 더 큰 데이터를 더 빠르게 처리할 수 있다



• 설치가 완료되면 바탕화면에 실행 아이콘 생성

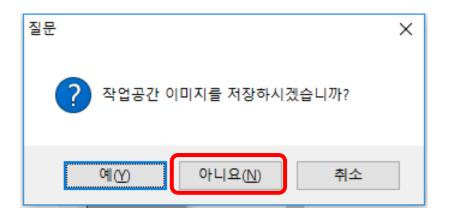


R 을 실행 화면

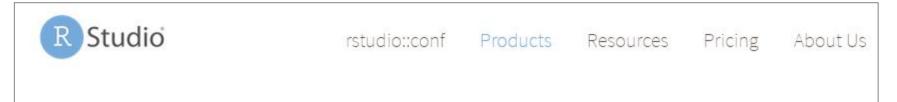


```
> 20+4
[1] 24
> a<-10
> b<-20
> a+b
[1] 30
> barplot(1:10)
                                                     > 20+4
                                                     [1] 24
                                                     > a<-10
                                                     > b<-20
                                                     > a+b
                                                     [1] 30
                                                     >
```

- R 실행 종료
  - 종료버튼 클릭시 다음과 같은 화면이 나오는데, 현재까지 작업한 결과를 저 장할 것인지를 묻는 것이다.
  - 대부분의 경우는 실행한 명령어만 저장하고 작업공간 이미지는 저장하지 않는다

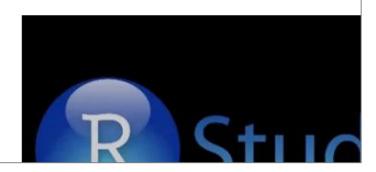


- https://www.rstudio.com/products/rstudio/
  - R 을 보다 편리하게 사용할 수 있는 환경 제공

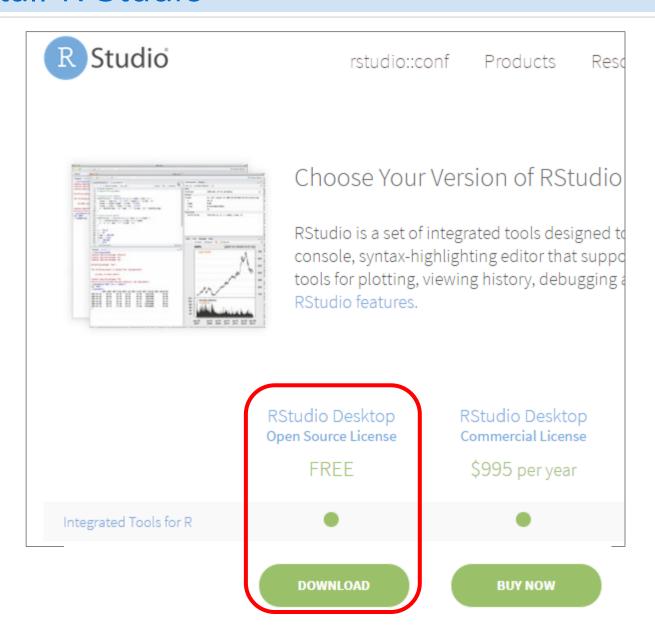


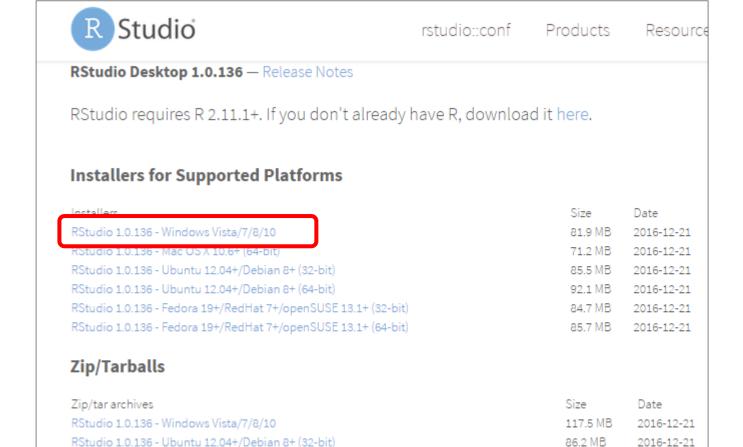
## Take control of your R code

RStudio is an integrated development environment (IDE) for R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management. Click here to see more RStudio features.



License	AGPL v3	RStudio License Agreement
Pricing	Free	\$995/year
	DOWNLOAD RSTUDIO DESKTOP	BUY NOW





93.2 MB

85.4 MB

86.6 MB

2016-12-21

2016-12-21

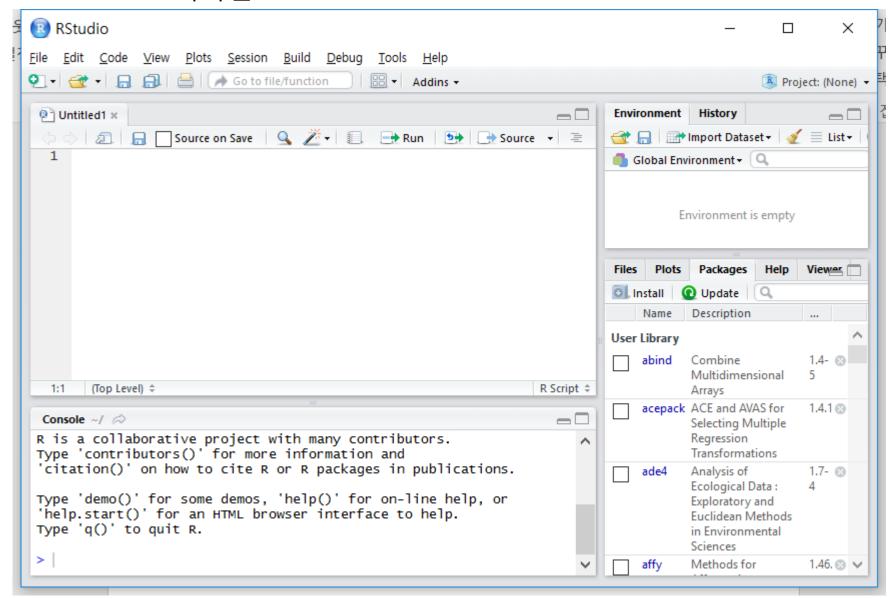
2016-12-21

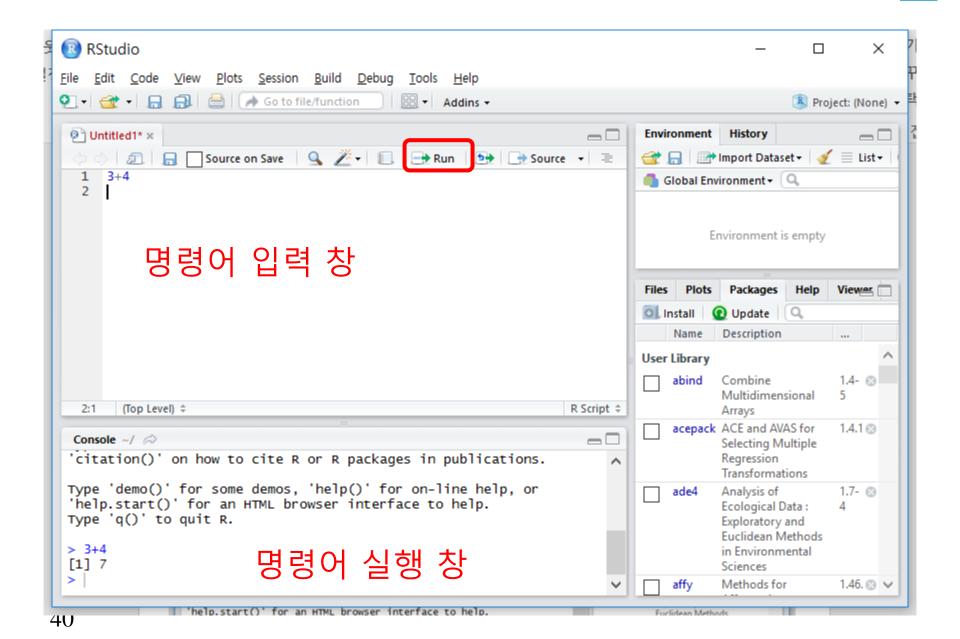
RStudio 1.0.136 - Ubuntu 12.04+/Debian 8+ (64-bit)

RStudio 1.0.136 - Fedora 19+/RedHat 7+/openSUSE 13.1+ (32-bit)

RStudio 1.0.136 - Fedora 19+/RedHat 7+/openSUSE 13.1+ (64-bit)

▶ R Studio 초기화면

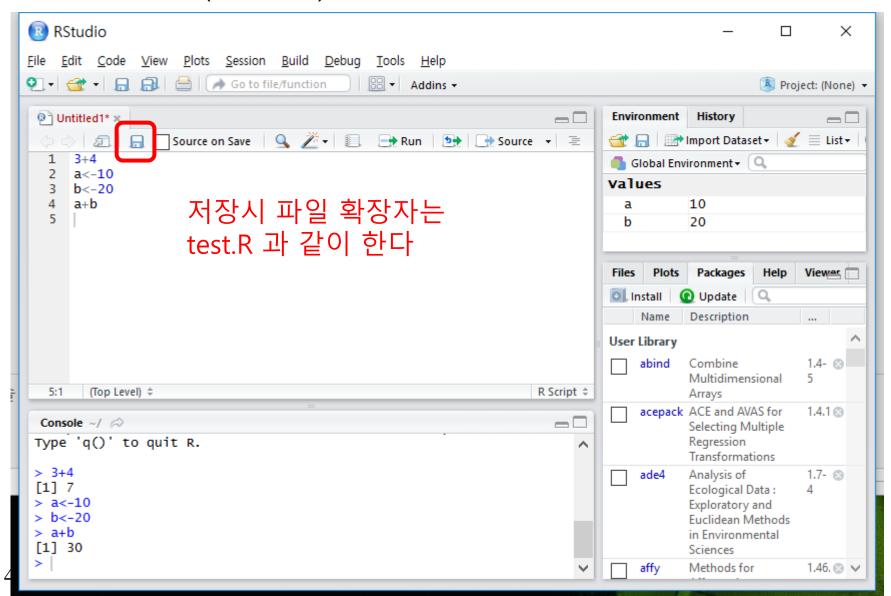




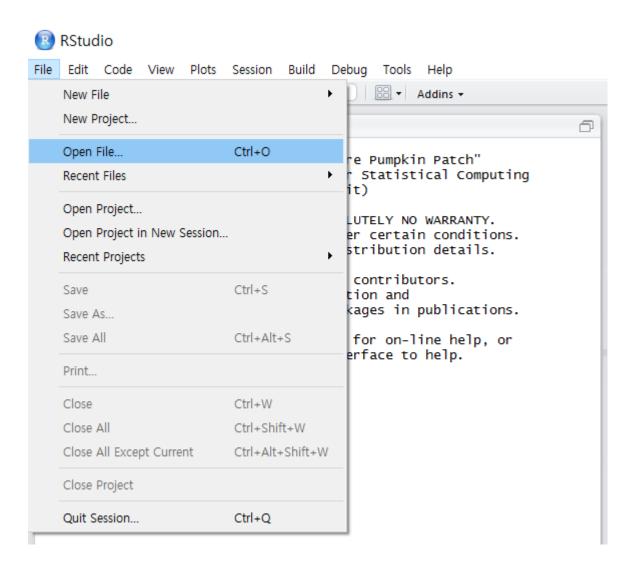


- Run : 명령어 실행 버튼
  - 현재 커서가 위치한 라인의 명령어 실행
  - 여러 명령어 라인을 선택한 상태이면 선택한 명령어들을 실행
  - 명령어 실행 버튼을 클릭하는 대신 <Ctrl>+<Enter> 를 해도 된다

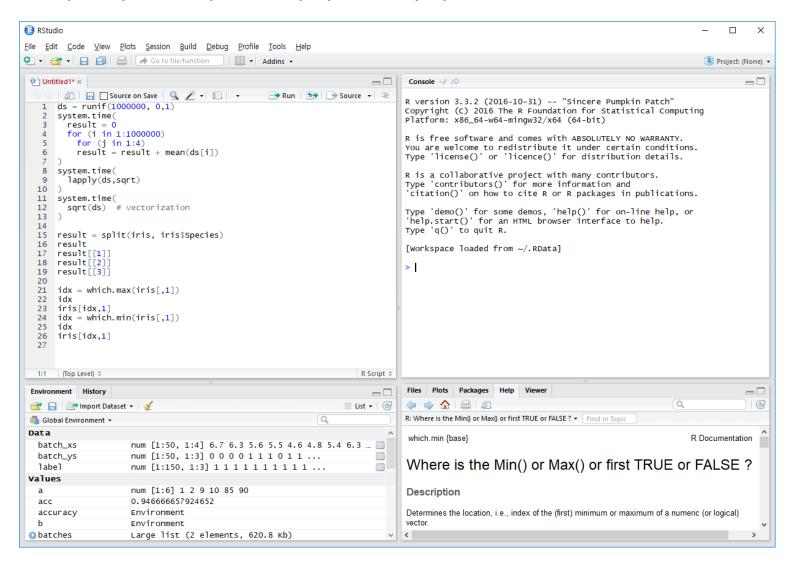
▶ 현재 작업 내용 (명령어들) 파일에 저장하기



• R 명령어 파일 불러오기



▶ 콘솔 화면이 오른쪽으로 가게 조정하기



콘솔 화면이 오른쪽으로 가게 조정하기

