

INTRODUCTION



Know your Instructor



- Author "[R for Business Analytics](#)"
- Author "[R for Cloud Computing](#)"
- Founder "[Decisionstats.com](#)"
- University of Tennessee, Knoxville
MS (courses in statistics and
computer science)
- MBA (IIM Lucknow, India-2003)
- B.Engineering (DCE 2001)

<http://linkedin.com/in/ajayohri>

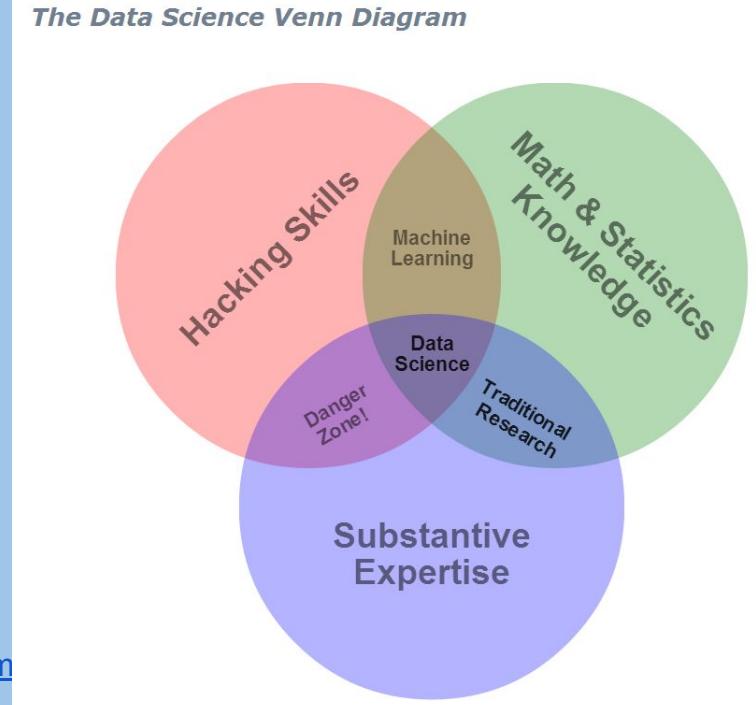
Classroom Rules

- From Instructor
- From Audience
 - mobile phones should be kindly switched off
 - Yes, this includes Whatsapp
 - Ask Questions at end of session
 - Take Notes
 - Please Take Notes



What is data science ?

Hacking (Programming) + Maths/Statistics + Domain Knowledge = Data Science



Oh really, is this a Data Scientist ?

a data scientist is simply a person who can

write code = in R,Python,Java, SQL, Hadoop (Pig,HQL,MR) etc

= **for** data storage, querying, summarization, visualization

= **how** efficiently, and in time (fast results?)

= **where** on databases, on cloud, servers

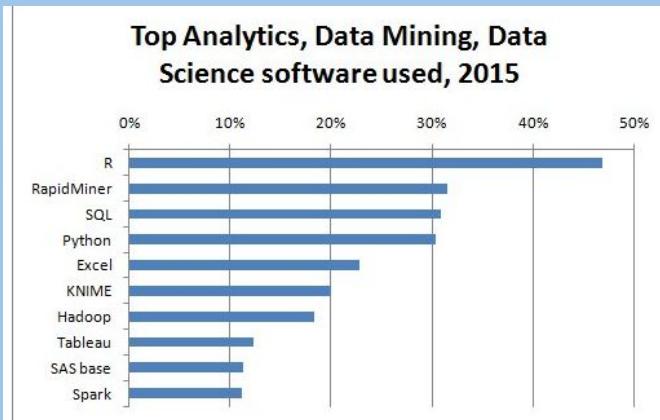
and understand **enough** statistics

to derive **insights** from data

so **business** can make **decisions**

Data Science with R

A popular language in Data Science



The top 10 tools by share of users were

Subscribe to KDnuggets News | Follow  @kdnnuggets |

R moves up to 5th place in IEEE language rankings

IEEE Spectrum has just published its third annual ranking with its [2016 Top Programming Languages](#), and the R Language is once again near the top of the list, moving up one place to fifth position.

Language Rank	Types	Spectrum Ranking
1. C		100.0
2. Java		98.1
3. Python		98.0
4. C++		95.9
5. R		87.9
6. C#		86.7
7. PHP		82.8
8. JavaScript		82.2
9. Ruby		74.5
10. Go		71.9

As I said [last year](#) (when R moved up to take sixth place), this is an extraordinary result for a domain-specific language. The other four languages in the top 5 (C, Java, Python and C++) are all general-purpose languages, suitable for just about any programming task. R by contrast is a language specifically for data science, and its high ranking here reflects both the critical importance of data science as a discipline today, and of R as the language of choice for data scientists.

What Is R

<https://www.r-project.org/about.html>

R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes

- an effective data handling and storage facility,
- a suite of operators for calculations on arrays, in particular matrices,
- a large, coherent, integrated collection of intermediate tools for data analysis,
- graphical facilities for data analysis and display either on-screen or on hardcopy, and
- a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.

Install R

<https://cran.r-project.org/bin/windows/base/>

R-3.3.1 for Windows (32/64 bit)

[Download R 3.3.1 for Windows](#) (70 megabytes, 32/64 bit)

[Installation and other instructions](#)

[New features in this version](#)

If you want to double-check that the package you have downloaded exactly matches the package distributed by R, you can compare the [md5sum](#) of the .exe to the [true fingerprint](#). You will need a version of md5sum for windows: both [graphical](#) and [command line versions](#) 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?](#)

Please see the [R FAQ](#) for general information about R and the [R Windows FAQ](#) for Windows-specific information.

[Other builds](#)

- Patches to this release are incorporated in the [r-patched snapshot build](#).
- A build of the development version (which will eventually become the next major release of R) is available in the [r-devel snapshot build](#).
- [Previous releases](#)

Note to webmasters: A stable link which will redirect to the current Windows binary release is
<<CRAN MIRROR>/bin/windows/base/release.htm>.

Install RStudio

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

The screenshot shows the RStudio website's download section. At the top, there's a navigation bar with links for Products, Resources, Pricing, About Us, and Blog, along with a search icon. Below the navigation, there's a main content area with a sidebar on the left containing links for RStudio, Shiny, R Packages, RStudio Server Pro, Shiny Server Pro, and shinyapps.io. To the right of the sidebar, there's a large blue banner with the text "CURIOUS WHO COMPANIES UPGRADED?" and a logo of a person with glasses. At the bottom of the page, there's a call-to-action button for Shiny and some footer links.

RStudio is a set of integrated tools designed to help you be more productive in R. It includes a console, syntax-highlighting editor that supports code completion and highlighting, as well as tools for plotting, history, debugging and workspace management.

If you run R on a Linux server and want to enable users to run R code from a web browser, please download RStudio Server.

Do you need support or a commercial license? Check out our [commercial offerings](#).

RStudio Desktop 0.99.903 — Release Notes

RStudio requires R 2.11.1 (or higher). If you don't already have R, you can download it [here](#).

Installers for Supported Platforms

Installers	Size	Date	MD5
RStudio 0.99.903 - Windows Vista/7/8/10	77.1 MB	2016-07-18	716f28f2143c5e21f4acea5752e284f8
RStudio 0.99.903 - Mac OS X 10.6+ (64-bit)	60 MB	2016-07-18	d14a1585b5a5ac0839507b9c04d460d6

Share your R code on the web with Shiny
Click here to learn more

Statistical Software Landscape

SAS

Python (Pandas)

IBM SPSS

R

Julia

Clojure

Octave

Matlab

JMP

E views



Using R with other software

<https://rforanalytics.wordpress.com/useful-links-for-r/using-r-from-other-software/>

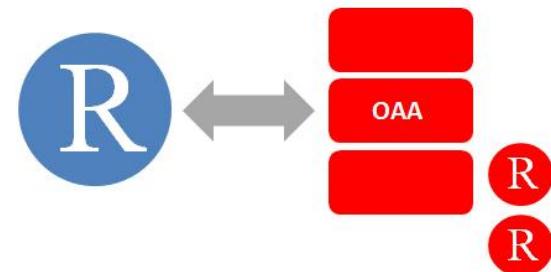
Tableau <http://www.tableausoftware.com/new-features/r-integration>

Qlik <http://qliksolutions.ru/qlikview/add-ons/r-connector-eng/>

Oracle R <http://www.oracle.com/technetwork/database/database-technologies/r/r-enterprise/overview/index.html>

Rapid Miner <https://rapid-i.com/content/view/202/206/lang,en/#r>

JMP <http://blogs.sas.com/jmp/index.php?/archives/298-JMP-Into-R!.html>



Using R with other software

<https://rforanalytics.wordpress.com/useful-links-for-r/using-r-from-other-software/>

SAS/IML <http://www.sas.com/technologies/analytics/statistics/iml/index.html>

Teradata <http://developer.teradata.com/applications/articles/in-database-analytics-with-ter>

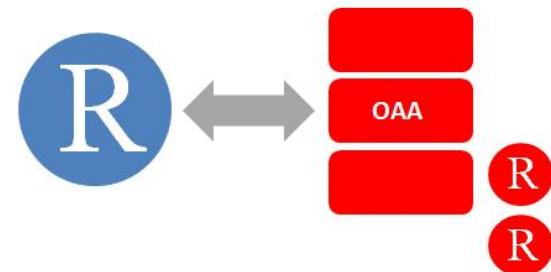
Pentaho <http://bigdatatechworld.blogspot.in/2013/10/integration-of-rweka-with-pentaho-data.html>

IBM SPSS

https://www14.software.ibm.com/webapp/iwm/web/signup.do?source=ibm-analytics&S_PKG=ov18855&S_TACT=M161003W&dy_nform=127&lang=en_US

TIBCO TERR

<http://spotfire.tibco.com/discover-spotfire/what-does-spotfire-do/predictive-analytics/tibco-enterprise-runtime-for-r-terr>



Some Advantages of R

open source

free

large number of algorithms and packages esp for statistics

flexible

very good for data visualization

superb community

rapidly growing

can be used with other software



Some Disadvantages of R

- in memory (RAM) usage
- steep learning curve
- some IT departments frown on open source
- verbose documentation
- tech support
- evolving ecosystem for corporates



Solutions for Disadvantages of R

- in memory (RAM) usage → specialized packages, in database computing
- steep learning curve → TRAINING !!!
- some IT departments frown on open source → TRAINING and education!
- verbose documentation → CRAN View , R Documentation
- tech support → expanding pool of resources
- evolving ecosystem for corporates → getting better with MS et al

R used by Government

- In the early days of the [Deepwater Horizon disaster](#), NIST used uncertainty analysis in R to harmonize spill estimates from various sources, and to provide ranges of estimates to other agencies and the media.
- Before new drugs are allowed on the market, the FDA works with pharmaceutical companies to verify safety and efficacy through clinical trials. Despite a [false perception](#) that only commercial software may be used, many pharmaceutical companies are now using open-source R to [analyze data from clinical trials](#).
- The National Weather Service uses R for research and development of [models to predict river flooding](#).
- The newly-formed [Consumer Financial Protection Bureau](#) -- freed from the restrictions of a legacy IT infrastructure -- is championing the use of open-source technologies in government.
- Local governments are also building data-based applications. The SF Estuary Institute [uses R and Google Maps](#) to provide a [tool to track pollution](#) in the San Francisco Bay area.

http://gsnmagazine.com/node/26483?c=cyber_security

R used by Telecom

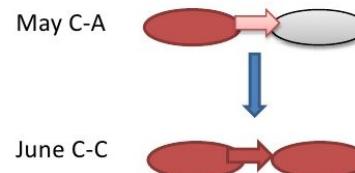
- Churn using

Social Network Analysis

<http://www.slideshare.net/dataspora/social-network-a>

Results: A Customer With a Canceller in Their Network Churns at Twice the Rate

Types of Connections (Edges)



reality	expected by chance	delta
X	Y	2.0

In essence, we are asking whether being connected to another canceller has any effect on one's rate of cancellation. It turns out that it does.

And if we only look at voluntary port-outs, we see that customers churn at 3x the rate.

R used by Insurance

a few more insurance related packages:

- [ChainLadder](#) – Reserving methods in R. The package provides Mack-, Munich-, Bootstrap, and Multivariate-chain-ladder methods, as well as the LDF Curve Fitting methods of Dave Clark and GLM-based reserving models.
- [cplm](#) – Monte Carlo EM algorithms and Bayesian methods for fitting Tweedie compound Poisson linear models
- [lossDev](#) – A Bayesian time series loss development model. Features include skewed-t distribution with time-varying scale parameter, Reversible Jump MCMC for determining the functional form of the consumption path, and a structural break in this path; by Christopher W. Laws and Frank A. Schmid
- [actuar](#): Loss distributions modelling, risk theory (including ruin theory), simulation of compound hierarchical models and credibility theory check out the [actuar](#) package by C. Dutang, V. Goulet and M. Pigeon.
- [favir](#): Formatted Actuarial Vignettes in R. FAViR lowers the learning curve of the R environment. It is a series of peer-reviewed Sweave papers that use a consistent style.
- [mondate](#): R packackge to keep track of dates in terms of months
- [lifecontingencies](#) – Package to perform actuarial evaluation of life contingencies

and

[Introduction to R for Actuaries](#) by Nigel de Silva

and <http://www.rininsurance.com/>

R in Finance

<http://www.rinfinance.com/>

R/Finance [home](#) [agenda](#) [register](#) [travel](#) [committee](#)

Friday, May 29th, 2015

08:00 - 09:00 Optional Pre-Conference Tutorials

Ross Bennett: PortfolioAnalytics: Advanced Moment Estimation & Optimization ([pdf](#))

Kris Boudt: High-frequency Price Data Analysis in R ([pdf](#))

Dirk Eddelbuettel: Hands-on Introduction to Rcpp ([pdf](#))

Guy Yollin: Getting Started with Quantstrat

Maria Belianina: An Introduction to OneTick

09:00 - 09:30 Registration (2nd floor Inner Circle) & Continental Breakfast (3rd floor by Sponsor Tables)

Transition between seminars

09:30 - 09:35 Kickoff

09:35 - 09:40 Sponsor Introduction

09:40 - 10:30 Emanuel Derman: Understanding the World

10:30 - 10:54 John Burkett: Portfolio Optimization: Price Predictability, Utility Functions, Computational Methods, and Applications ([pdf](#))

Kyle Balkissoon: A Framework for Integrating Portfolio-level Backtesting with Price and Quantity Information ([html](#))

Anthonney Tsou: Implementation of Quality Minus Junk

Ilya Kipnis: Flexible Asset Allocation With Stepwise Correlation Rank ([pptm](#))

10:54 - 11:20 Break

11:20 - 11:40 Sanjiv Das: Efficient Rebalancing of Taxable Portfolios ([pdf](#))

11:40 - 12:00 Marjan Wauters: Characteristic-based equity portfolios: economic value and dynamic style allocation ([pdf](#))

12:00 - 12:20 Bernhard Pfaff: The sequel of cccp: Solving cone constrained convex programs

12:20 - 13:40 Lunch

13:40 - 14:00 Markus Gesmann: Communicating risk - a perspective from an insurer ([pdf](#))

14:00 - 14:20 Doug Martin: Nonparametric vs Parametric Shortfall: What are the Differences?

R in Finance

<http://cran.r-project.org/web/views/Finance.html>

This CRAN Task View contains a list of packages useful for empirical work in Finance, grouped by topic.

- The Rmetrics suite of packages comprises [fArma](#), [fAsianOptions](#), [fAssets](#), [fBasics](#), [fBonds](#), [timeDate](#) (formerly: fCalendar), [fCopulae](#), [fExoticOptions](#), [fExtremes](#), [fGarch](#), [fImport](#), [fNonlinear](#), [fOptions](#), [fPortfolio](#), [fRegression](#), [timeSeries](#) (formerly: fSeries), [fTrading](#), [fUnitRoots](#) and contains a very large number of relevant functions for different aspect of empirical and computational finance.
- The [RQuantLib](#) package provides several option-pricing functions as well as some fixed-income functionality from the QuantLib project to R.
- The [quantmod](#) package offers a number of functions for quantitative modelling in finance as well as data acquisition, plotting and other utilities.
- The [portfolio](#) package contains classes for equity portfolio management; the [portfolioSim](#) builds a related simulation framework. The [backtest](#) offers tools to explore portfolio-based hypotheses about financial instruments. The [stockPortfolio](#) package provides functions for single index, constant correlation and multigroup models. The [pa](#) package offers performance attribution functionality for equity portfolios.
- The [PerformanceAnalytics](#) package contains a large number of functions for portfolio performance calculations and risk management.

R in Pharma

<http://blog.revolutionanalytics.com/2013/08/r-drug-development-and-the-fda.html>

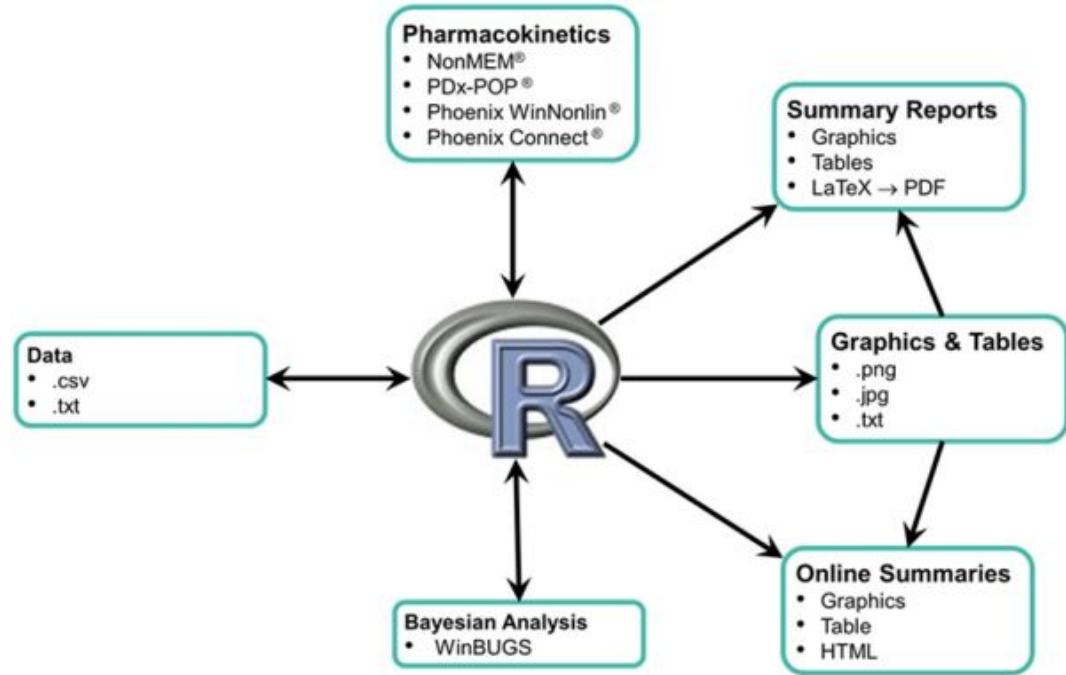
Opening the Doors to Open Source Programming in Drug Development.

R: Regulatory Compliance and Validation Issues A Guidance Document for the Use of R in Regulated Clinical Trial Environments in which he concluded that useR 2012 FDA statistician Jea Brodsky presented a poster described how FDA scientists “use R on a daily basis” and have themselves written R packages for use at various stages in the drug submission process.

Open Source Software in the Biopharma Industry: Challenges and Opportunities.

R in Pharma

<http://web.quanticate.com/bid/102741/Using-the-Statistical-Programming-Language-R-in-the-Pharma-Industry>



R in Pharma

<http://cran.r-project.org/web/views/ClinicalTrials.html>

This task view gathers information on specific R packages for design, monitoring and analysis of data from clinical trials. It focuses on including packages for clinical trial design and monitoring in general plus data analysis packages for a specific type of design.

Design and Monitoring

- [TrialSize](#) This package has more than 80 functions from the book *Sample Size Calculations in Clinical Research* (Chow & Wang & Shao, 2007, 2nd ed., Chapman & Hall/CRC).
- [asd](#) This Package runs simulations for adaptive seamless designs using early outcomes for treatment selection.
- [bcm](#) This package implements a wide variety of one and two-parameter Bayesian CRM designs. The program can run interactively, allowing the user to enter outcomes after each cohort has been recruited, or via simulation to assess operating characteristics.
- [blockrand](#) creates randomizations for block random clinical trials. It can also produce a PDF file of randomization cards.
- [contDesign](#) This small package contains a series of simple tools for constructing and manipulating confounded and fractional factorial designs.
- [CRTsize](#) This package contains basic tools for the purpose of sample size estimation in cluster (group) randomized trials. The package contains traditional power-based methods, empirical smoothing (Rotondi and Donner, 2009), and updated meta-analysis techniques (Rotondi and Donner, 2011).
- [dfrm](#) This package provides functions to run the CRM and TITE-CRM in phase I trials and calibration tools for trial planning purposes.
- [experiment](#) contains tools for clinical experiments, e.g., a randomization tool, and it provides a few special analysis options for clinical trials.
- [FDR](#) This package creates regular and non-regular Factorial designs. Furthermore, analysis tools for Fractional Factorial designs with 2-level factors are offered (main effects and interaction plots for all factors simultaneously, cube plot for looking at the simultaneous effects of three factors, full or half normal plot, alias structure in a more readable format than with the built-in alias function alias). The package is currently subject to intensive development. While much of the intended functionality is already available, some changes and improvements are still to be expected.
- [GroupSeq](#) performs computations related to group sequential designs via the alpha spending approach, i.e., interim analyses need not be equally spaced, and their number need not be specified in advance.
- [gsDesign](#) derives group sequential designs and describes their properties.
- [lbd](#) and [Hausc](#) computes and plots group sequential stopping boundaries from the Lan-DeMets method with a variety of a-spending functions using the ld98 program from the Department of Biostatistics, University of Wisconsin written by DM Rebbapragada, DL DeMets, KM Kim, and KKG Lan.
- [ldbounds](#) uses Lan-DeMets Method for group sequential trial, its functions calculate bounds and probabilities of a group sequential trial.
- [longpower](#) The longpower package contains functions for computing power and sample size for linear models of longitudinal data based on the formula due to Liu and Liang (1997) and Diggle et al (2002). Either formula is expressed in terms of marginal model or Generalized Estimating Equations (GEE) parameters. This package contains functions which translate pilot mixed effect model parameters (e.g random intercept and/or slope) into marginal model parameters so that the formulas of Diggle et al or Liu and Liang formula can be applied to produce sample size calculations for two sample longitudinal designs assuming known variance.
- [PIDS](#) generates predicted interval plots, simulates and plots confidence intervals of an effect estimate given observed data and a hypothesis about the distribution of future data.
- [PowerTOST](#) contains functions to calculate power and sample size for various study designs used for bioequivalence studies. See function known.designs() for study designs covered. Moreover the package contains functions for power and sample size based on 'expected' power in case of uncertain (estimated) variability. Added are functions for the power and sample size for the ratio of two means with normally distributed data on the original scale (based on Fieller's confidence ('fiducial') interval).
- [pwr](#) has power analysis functions along the lines of Cohen (1988).
- [PwrGSD](#) is a set of tools to compute power in a group sequential design.
- [qtlDesign](#) provides tools for the design of QTL experiments.
- [seqmnu](#) is computes the probability of crossing sequential efficacy and futility boundaries in a clinical trial. It implements the Armitage-McPherson and Rowe Algorithm using the method described in Schoenfeld (2001).

Design and Analysis

- Package [AGSDes](#) This package provides tools and functions for parameter estimation in adaptive group sequential trials.
- Package [clintfun](#) has functions for both design and analysis of clinical trials. For Phase II trials, it has functions to calculate sample size, effect size, and power based on Fisher's exact test, the operating characteristics of a two-stage boundary. Optimal and Minimax 2-stage Phase II designs given by Richard Simon, the exact 1-stage Phase II design and can compute a stopping rule and its operating characteristics for toxicity monitoring based repeated significance testing. For phase III trials, it can calculate sample size for group sequential designs.

Companies using R

from <http://www.revolutionanalytics.com/companies-using-r>

ANZ, the fourth largest bank in Australia, using R for credit risk analysis

Bank of America uses R for reporting.

The Consumer Financial Protection Bureau uses R for data analysis.

Facebook

Facebook and R:

- Analysis of Facebook Status Updates
- Facebook's Social Network Graph
- How Google and Facebook are using R
- Predicting Colleague Interactions with R

Refresher in Statistics

Mean

Arithmetic Mean- the sum of the values divided by the number of values.

The [geometric mean](#) is an average that is useful for sets of positive numbers that are interpreted according to their product and not their sum (as is the case with the arithmetic mean) e.g. rates of growth.

Median

the **median** is the number separating the higher half of a data sample, a population, or a probability distribution, from the lower half

Mode-

The "mode" is the value that occurs most often.

Refresher in Statistics

Range

the **range** of a set of data is the difference between the largest and smallest values.

Variance

mean of squares of differences of values from mean

Standard Deviation

square root of its variance

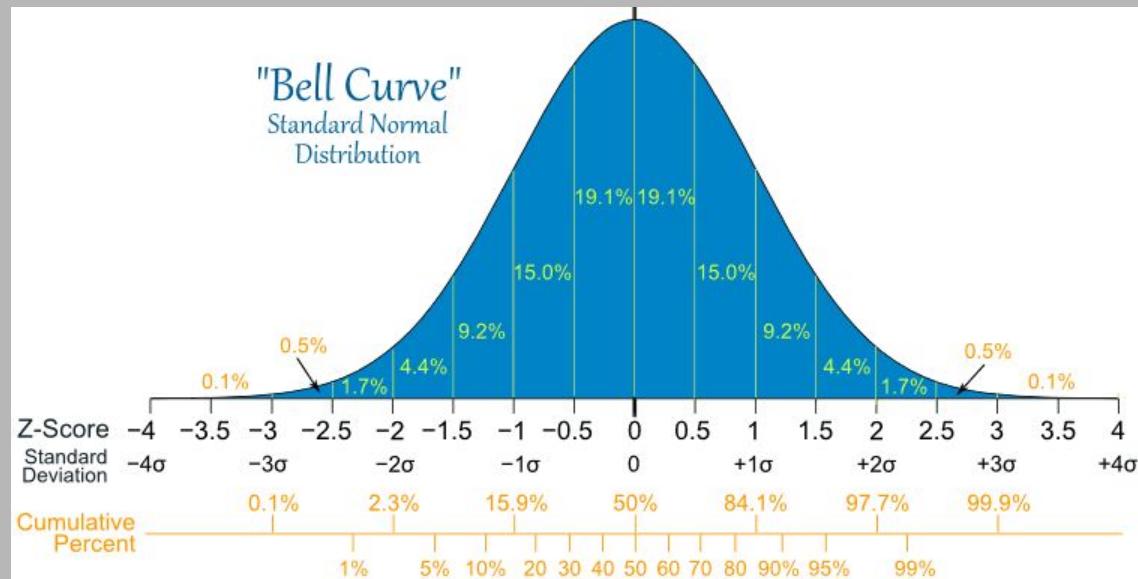
Frequency

a **frequency distribution** is a table that displays the **frequency** of various outcomes in a sample.

Distributions

Normal

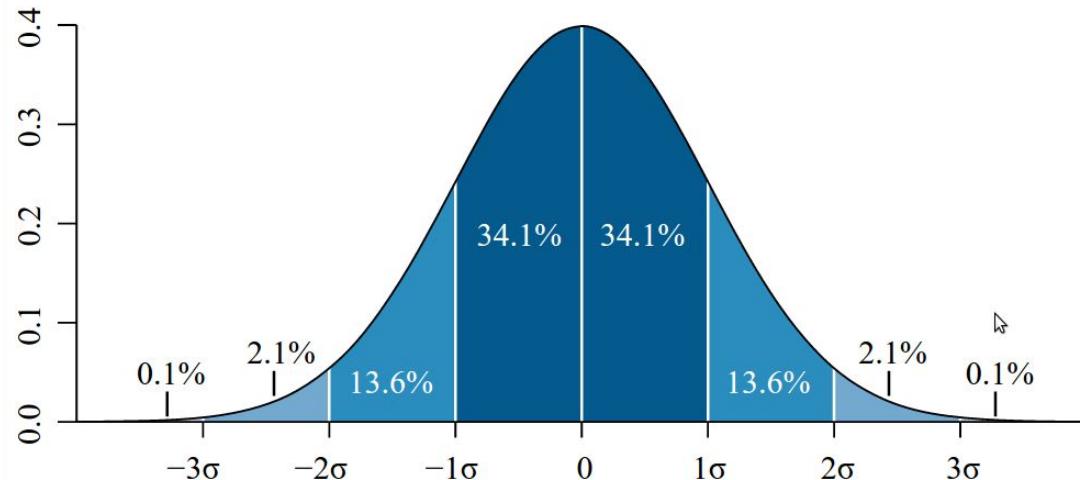
The simplest case of a normal distribution is known as the *standard normal distribution*. This is a special case where $\mu=0$ and $\sigma=1$.



Refresher in Statistics

Probability Distribution

The [probability density function](#) (pdf) of the [normal distribution](#), also called Gaussian or "bell curve", the most important continuous random distribution. As notated on the figure, the probabilities of intervals of values correspond to the area under the curve.



Pre Requisites

- Installation of R

<http://cran.rstudio.com/bin/windows/base/>



CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage
The R Journal

Software
R Sources
R binaries
Packages
Other

Documentation
Manuals
FAQs
Contributed

R-3.1.1 for Windows (32/64 bit)

[Download R 3.1.1 for Windows](#) (54 megabytes, 32/64 bit)

Installation and other instructions
[New features in this version](#)

If you want to double-check that the package you have downloaded exactly matches the package distributed by R, you can compare the [md5sum](#) of the .exe to the [true fingerprint](#). You will need a version of md5sum for windows: both [graphical](#) and [command line versions](#) are available.

Frequently asked questions

- How do I install R when using Windows Vista?
- How do I update packages in my previous version of R?
- Should I run 32-bit or 64-bit R?

Please see the [R FAQ](#) for general information about R and the [R Windows FAQ](#) for Windows-specific information.

Other builds

- Patches to this release are incorporated in the [r-patched.snapshot.build](#)
- A build of the development version (which will eventually become the next major release of R) is available in the [r-devel.snapshot.build](#)
- [Previous releases](#)

Note to webmasters: A stable link which will redirect to the current Windows binary release is
[-CRAN MIRROR-bin/windows/base/release.htm](#)

Last change: 2014-07-10, by Duncan Murdoch

- R Studio

- R Packages

Pre Requisites

- Installation of R
 - Rtools
 - <http://cran.rstudio.com/bin/windows/Rtools/>
- R Studio
- R Packages



[CRAN](#)

[Mirrors](#)

[What's new?](#)

[Data Vignettes](#)

[Search](#)

[About R](#)

[R Homepage](#)

[R Journal](#)

[Software](#)

[R Sources](#)

[R Binaries](#)

[Binaries](#)

[Other](#)

[Documentation](#)

[Manuals](#)

[FAQs](#)

[Contributed](#)

Building R for Windows

This document is a collection of resources for building packages for R under Microsoft Windows, or for building R itself (version 1.9.0 or later). The original collection was put together by Prof. Brian Ripley; it is currently being maintained by Duncan Murdoch.

The authoritative source of information for tools to work with the current release of R is the "R Administration and Installation" manual. In particular, please read the "Windows Toolkit" appendix.

Tools Downloads

With the change to gcc 4.2.1, some of the tools for 3.2 bit compiles became incompatible with obsolete versions of R. Since then we have been maintaining one actively updated version of the tools, and other "frozen" snapshots of them. We recommend that users use the latest release of Rtools with the latest release of R.

The current version of this file is recorded here: [VERSION.txt](#)

Download	R compatibility	Frozen?
Rtools31.exe	R 3.0.x to 3.1.x	No
Rtools30.exe	R 3.0.x to R 3.0.x	Yes
Rtools212.exe	R 2.12.1 to R 2.15.1	Yes
Rtools214.exe	R 2.13.x or R 2.14.x	Yes
Rtools213.exe	R 2.13.x	Yes
Rtools211.exe	R 2.12.x	Yes
Rtools2111.exe	R 2.10.x or R 2.11.x	Yes
Rtools210.exe	R 2.9.x or 2.10.x	Yes
Rtools208.exe	R 2.8.x or 2.9.x	Yes
Rtools206.exe	R 2.6.x or R 2.7.x	Yes
Rtools207.exe	R 2.6.x or R 2.7.x	Yes
Rtools206.exe	R 2.6.x, R 2.5.x or (untested) earlier	Yes

The change history to the Rtools is below:

Tools for 64 bit Windows builds

Rtools 2.12 and later include both 32 bit and 64 bit tools.

Pre Requisites

- Installation of R

- RTools

The screenshot shows the RStudio website's download section. At the top, there are navigation links for Products, Resources, Pricing, About Us, and Blog, along with a search icon. Below this, a banner says "Download RStudio". The main content area is titled "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, as well as tools for plotting, history, debugging and workspace management." A note below states: "If you run R on a Linux server and want to enable users to remotely access RStudio using a web browser please download RStudio Server." To the right, a sidebar offers support for commercial licenses. At the bottom, there's a form to subscribe to RStudio's newsletter.

Download RStudio Desktop v0.98.1074 — Release Notes

RStudio requires R 2.11.1 (or higher). If you don't already have R, you can download it [here](#).

Installers for ALL Platforms

Installers	Size	Date	MD5
RStudio 0.98.1074 - Windows XP/Vista/7/8	49 MB	2014-10-14	74d7bc76ec04287fac79cdd8dfaa8dd
RStudio 0.98.1074 - Mac OS X 10.6+ (64-bit)	38.4 MB	2014-10-14	f01c43fa29af679400c0faeae7ee33fb
RStudio 0.98.1074 - Debian 6+/Ubuntu 10.04+ (32-bit)	54.3 MB	2014-10-14	759d86599b22b28202a5aa0025e77278
RStudio 0.98.1074 - Debian 6+/Ubuntu 10.04+ (64-bit)	66.1 MB	2014-10-14	a77e51c714a27df28ffcc9b9732d04ab

- R Studio

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

- R Packages

Pre Requisites

- Installation of R

- RTools

- R Studio

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

- R Packages

about eight packages supplied with the R distribution and many more are available through the CRAN family of Internet sites covering a very wide range of modern statistics.

The screenshot shows the R Documentation homepage. At the top, there's a search bar and a navigation bar with links for Discussion, About, and documentation package. Below the search bar is a section titled "Top Ranked CRAN Packages" with a "Week" and "Month" filter. It lists the top 10 packages with their names, package IDs, and download counts:

#	Package	# Downloads
1	Rcpp	80382
2	ggplot2	69508
3	plyr	65837
4	stringr	65371
5	digest	63067
6	RColorBrewer	57602
7	reshape2	57236
8	colorspace	51693
9	labeling	49615
10	scales	47407

Below this is a "New Packages" section with a similar list:

#	Package
1	CP
2	minormpow
3	mblock
4	NB
5	BSGW
6	stabs
7	CEC
8	mvrpb
9	EhNRG
10	mdsdt

The main content area features logos for CRAN, Bioconductor, and GitHub, along with search fields for All Fields, Package Name, Function Name, Title, Description, and Author(s), and a prominent green "Start search" button.

CRAN

107 sites in 49 regions



CRAN
[Mirrors](#)
[What's new?](#)
[Task Views](#)
[Search](#)

[About R](#)
[R Homepage](#)
[The R Journal](#)

[Software](#)
[R Sources](#)
[R Binaries](#)
[Packages](#)
[Other](#)

[Documentation](#)
[Manuals](#)
[FAQs](#)
[Contributed](#)

CRAN Mirrors

The Comprehensive R Archive Network is available at the following URLs, please choose a location close to you. Some statistics on the status of the mirrors can be found here: [main page](#), [windows release](#), [windows old release](#).

0-Cloud

<http://cran.rstudio.com/>

Algeria

<http://cran.usthb.dz/>

Argentina

<http://mirror.fcaglp.unlp.edu.ar/CRAN/>

Australia

<http://cran.csiro.au/>

<http://cran.ms.unimelb.edu.au/>

Austria

<http://cran.at.r-project.org/>

Belgium

<http://www.freestatistics.org/cran/>

Brazil

<http://nbcgib.uesc.br/mirrors/cran/>

<http://cran-r.c3sl.ufpr.br/>

<http://cran.fiocruz.br/>

<http://www.vps.fmvz.usp.br/CRAN/>

<http://briger.esalq.usp.br/CRAN/>

Canada

<http://cran.stat.sfu.ca/>

<http://mirror.its.dal.ca/cran/>

<http://cran.ustat.utoronto.ca/>

<http://cran.skazkaforyou.com/>

<http://cran.parentingamerica.com/>

Chile

<http://dirichlet.mat.puc.cl/>

China

<http://ftp.ctex.org/mirrors/CRAN/>

<http://mirror.bjtu.edu.cn/cran/>

<http://mirrors.opencas.cn/cran/>

<http://jstatsoft.cntv.ac.cn/cran/> (CRAN)

Rstudio, automatic redirection to servers worldwide



University of Science and Technology Houari Boumediene

Universidad Nacional de La Plata

CSIRO

University of Melbourne

Wirtschaftsuniversitaet Wien

K.U.Leuven Association

Center for Comp. Biol. at Universidade Estadual de Santa Cruz

Universidade Federal do Parana

Oswaldo Cruz Foundation, Rio de Janeiro

University of Sao Paulo, Sao Paulo

University of Sao Paulo, Piracicaba

Simon Fraser University, Burnaby

Dalhousie University, Halifax

University of Toronto

iWeb, Montreal

iWeb, Montreal

Pontificia Universidad Catolica de Chile, Santiago

CTEX.ORG

Beijing Jiaotong University, Beijing

Chinese Academy of Sciences, Beijing

TUMA Team, Tsinghua University

Non CRAN Repositories

<http://www.rdocumentation.org/>

The screenshot shows the R Documentation website interface. At the top, there is a search bar with the placeholder "Start searching the documentation". Below it is a "TASK VIEWS" sidebar containing a list of R package categories, each preceded by a small blue triangle icon:

- Bayesian
- ChemPhys
- ClinicalTrials
- Cluster
- DifferentialEquations
- Distributions
- Econometrics
- Envirometrics
- ExperimentalDesign
- Finance
- Genetics
- gR
- Graphics
- HighPerformanceComputing
- MachineLearning
- MedicalImaging
- MetaAnalysis
- Multivariate
- NaturalLanguageProcessing
- NumericalMathematics
- OfficialStatistics
- Optimization
- Pharmacokinetics
- Phylogenetics
- Psychometrics
- ReproducibleResearch
- Robust
- SocialSciences
- Spatial
- SpatioTemporal

The main content area has a header "R Documentation" with a logo. It displays a message: "Search the R documentation of 7393 R packages and 150600 R functions". Below this is a descriptive text: "Rdocumentation is a tool that helps you easily find and browse the documentation of all current and some past packages on CRAN. Click on the search bar at the top left for instant search or fill out the forms below for advanced search!". There are six input fields for advanced search: "All Fields", "Package Name", "Function Name", "Title", "Description", and "Author(s)". A large green "Start search" button is located at the bottom of these fields. To the right of the search form is a vertical sidebar for DataCamp, featuring a DataCamp logo, a "Learn Data Science with R" section, a "\$25/month" offer, a thumbnail of a course featuring a bar chart and a gold medal, and a "Discover All Courses" button. The sidebar also mentions "Data Manipulation, Data Visualization, R Programming, Big Data, and much more". At the very bottom, it says "Aggregating packages from:" followed by logos for CRAN, Bioconductor, and GitHub.

github

<https://github.com/trending?l=R>

The screenshot shows the GitHub trending repositories page for the R programming language. The URL is <https://github.com/trending?l=R>. The page title is "Trending repositories". A sub-header says "Find what repositories the GitHub community is most excited about today." On the left, there are tabs for "Repositories" and "Developers", and a dropdown for "Trending: today". On the right, there's a sidebar with language filters: "All languages", "Unknown languages", "C", "C++", "HTML", "Java", "JavaScript", "Python", and "R" (which is selected). There's also a "Languages" dropdown and a "ProTip!" box. The main content lists four repositories:

- rdpeng/ProgrammingAssignment2**
Repository for Programming Assignment 2 for R Programming on Coursera
R + Built by [avatars]
- qinwf/awesome-R**
A curated list of awesome R frameworks, packages and software.
R + stars today + Built by [avatars]
- berndbischl/mlr**
mlr: Machine Learning in R
R + Built by [avatars]
- rstudio/shinyapps**
[repo details]

bioconductor

<http://www.bioconductor.org/>

The screenshot shows the Bioconductor website homepage. At the top, there is a dark blue header bar with the Bioconductor logo and navigation links for Home, Install, Help, Developers, and About. A search bar is also present. Below the header, there are several sections: 'BioC2015' (with information about the conference), 'About Bioconductor' (with a brief history and current status), 'Install' (with links to get started and various installation options), 'Learn' (with links to courses, support, and videos), 'Use' (with links to software, annotation, and experiment packages), and 'Develop' (with links to contribute, package guidelines, and developer resources). The main content area has a light blue background.

BioC2015

Join us for morning talks from distinguished speakers and community members, afternoon workshops to hone your skills, and poster sessions and social activities to get to know members of the Bioconductor community at our [Annual Conference](#), July 20 (Developer Day), 21 and 22 in Seattle, WA.

About Bioconductor

Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. Bioconductor uses the R statistical programming language, and is open source and open development. It has two releases each year, [1024 software packages](#), and an active user community. Bioconductor is also available as an [AMI](#) (Amazon Machine Image) and a series of [Docker](#) images.

News

- Bioconductor [3.1](#) is available.
- Orchestrating high-throughput genomic analysis with Bioconductor ([abstract](#)) and other [recent literature](#).
- Read our latest [newsletter](#) and [course](#)

Install »

Get started with Bioconductor

- [Install Bioconductor](#)
- [Explore packages](#)
- [Get support](#)
- [Latest newsletter](#)
- [Follow us on twitter](#)
- [Install R](#)

Learn »

Master Bioconductor tools

- [Courses](#)
- [Support site](#)
- [Package vignettes](#)
- [Literature citations](#)
- [Common work flows](#)
- [FAQ](#)
- [Community resources](#)
- [Videos](#)

Use »

Create bioinformatic solutions with Bioconductor

- [Software, Annotation, and Experiment packages](#)
- [Amazon Machine Image](#)
- [Latest release announcement](#)
- [Support site](#)

Develop »

Contribute to Bioconductor

- [Use BioC 'devel'](#)
- ['Devel' Software, Annotation and Experiment packages](#)
- [Package guidelines](#)
- [New package submission](#)
- [Developer resources](#)
- [Build reports](#)

Install R

<https://cran.r-project.org/bin/windows/base/>

R-3.3.1 for Windows (32/64 bit)

[Download R 3.3.1 for Windows](#) (70 megabytes, 32/64 bit)

[Installation and other instructions](#)

[New features in this version](#)

If you want to double-check that the package you have downloaded exactly matches the package distributed by R, you can compare the [md5sum](#) of the .exe to the [true fingerprint](#). You will need a version of md5sum for windows: both [graphical](#) and [command line versions](#) 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?](#)

Please see the [R FAQ](#) for general information about R and the [R Windows FAQ](#) for Windows-specific information.

[Other builds](#)

- Patches to this release are incorporated in the [r-patched snapshot build](#).
- A build of the development version (which will eventually become the next major release of R) is available in the [r-devel snapshot build](#).
- [Previous releases](#)

Note to webmasters: A stable link which will redirect to the current Windows binary release is
<<CRAN MIRROR>/bin/windows/base/release.htm>.

Install RStudio

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

The screenshot shows the RStudio website's download section. At the top, there's a navigation bar with links for Products, Resources, Pricing, About Us, and Blog, along with a search icon. Below the navigation, there's a main content area with a sidebar on the left containing links for RStudio, Shiny, R Packages, RStudio Server Pro, Shiny Server Pro, and shinyapps.io. To the right of the sidebar, there's a large blue banner with the text "CURIOUS WHO COMPANIES UPGRADED?" and a logo of a person with glasses. At the bottom of the page, there's a call-to-action button for Shiny and some footer links.

RStudio is a set of integrated tools designed to help you be more productive in R. It includes a console, syntax-highlighting editor that supports code completion and highlighting, as well as tools for plotting, history, debugging and workspace management.

If you run R on a Linux server and want to enable users to run R code from a web browser, please download RStudio Server.

Do you need support or a commercial license? Check out our [commercial offerings](#).

RStudio Desktop 0.99.903 — Release Notes

RStudio requires R 2.11.1 (or higher). If you don't already have R, you can download it [here](#).

Installers for Supported Platforms

Installers	Size	Date	MD5
RStudio 0.99.903 - Windows Vista/7/8/10	77.1 MB	2016-07-18	716f28f2143c5e21f4acea5752e284f8
RStudio 0.99.903 - Mac OS X 10.6+ (64-bit)	60 MB	2016-07-18	d14a1585b5a5ac0839507b9c04d460d6

Share your R code on the web with Shiny
Click here to learn more

Pre Requisites

- R Packages

`install.packages()` INSTALLS

`update.packages()` UPDATES

`library()` LOADS

- Packages are **installed** once, updated periodically, but **loaded** every time

Interfaces to R

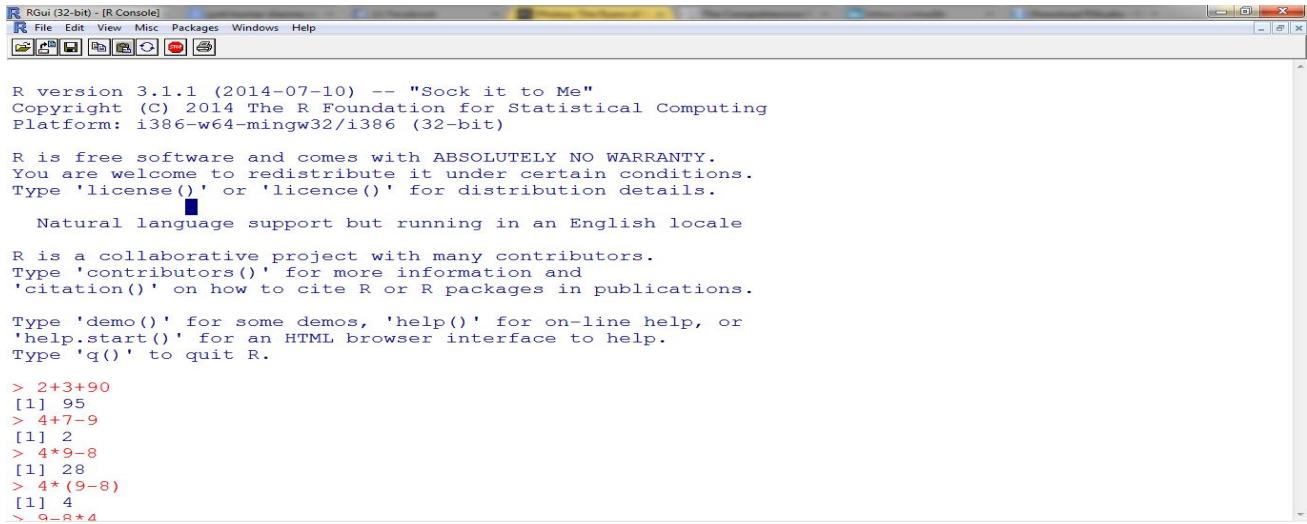
- Console

Default

Customization

- IDE

- GUI



R version 3.1.1 (2014-07-10) -- "Sock it to Me"
Copyright (C) 2014 The R Foundation for Statistical Computing
Platform: i386-w64-mingw32/i386 (32-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

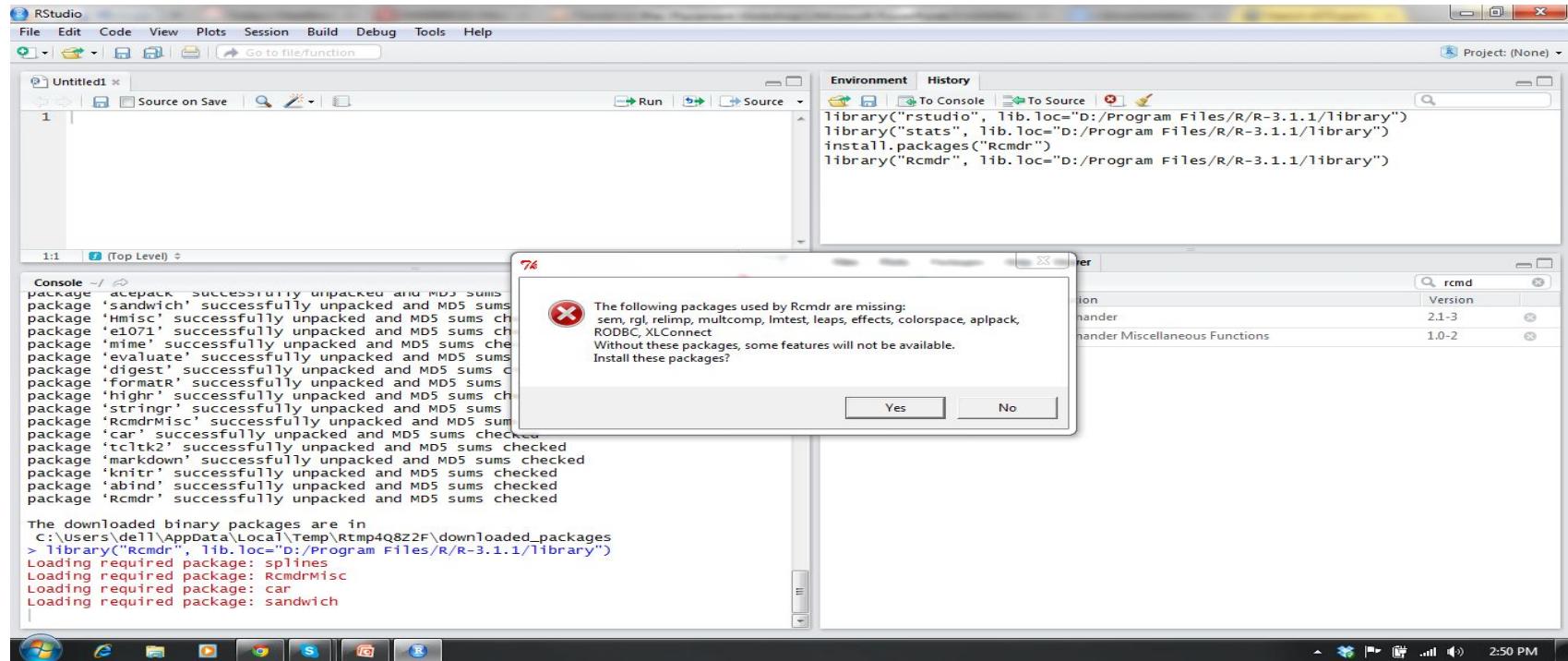
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

```
> 2+3+90
[1] 95
> 4+7-9
[1] 2
> 4*9-8
[1] 28
> 4*(9-8)
[1] 4
> a-a*4
```

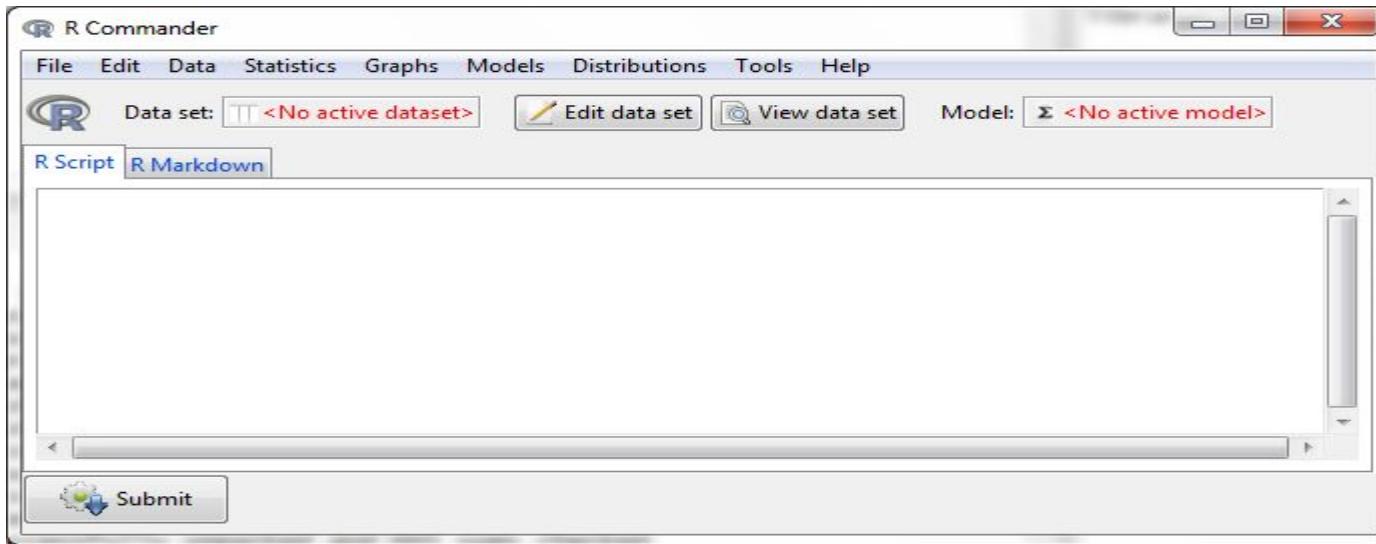
Graphical Interfaces to R

- R Commander
- Rattle
- Deducer

Installation of R Commander

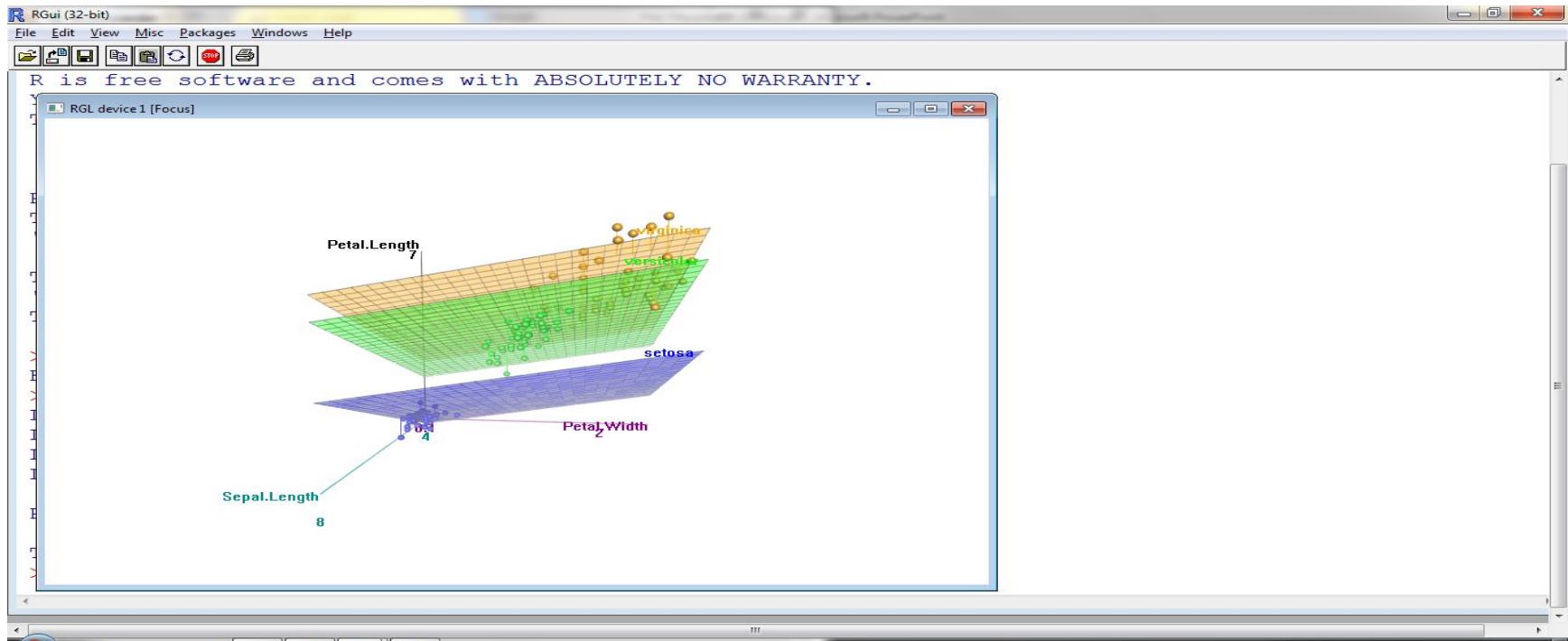


Overview of R Commander



Demo

R Commander – 3D Graphs



Installation of Rattle

The screenshot shows the RStudio interface with the following details:

- Top Bar:** RStudio, File, Edit, Code, View, Plots, Session, Build, Debug, Tools, Help.
- Project Bar:** Project: (None)
- Console:** Untitled1, showing R code and its execution output. The output includes:

 - Library imports: library("rstudio"), library("stats"), install.packages("Rcmdr")
 - Library loading: library("Rcmdr")
 - Message: "The downloaded binary packages are in C:/Users/dell/AppData/Local/Temp/Rtmp4Q8Z2F/downloaded_packages"
 - RcmdrMsg: [1] NOTE: R Commander Version 2.1-3: Fri Oct 17 14:54:03 2014

- Environment Tab:** Shows the current environment variables.
- Install Packages Dialog:** A modal window titled "Install Packages".
 - Install from: Repository (CRAN, CRANextra)
 - Packages (separate multiple with space or comma): rattle
 - Library: D:/Program Files/R/R-3.1.1/library [Default]
 - Install dependencies:
 - Buttons: Install, Cancel
- Viewer Tab:** Shows a table of installed packages:

Description	Version
R Commander	2.1-3
R Commander Miscellaneous Functions	1.0-2
- Taskbar:** Windows taskbar showing various application icons.
- System Tray:** Shows the date and time: 2:54 PM.

Installation of Rattle

The screenshot shows the RStudio interface with the following details:

- Console Output:**

```
library('rattle', lib.loc="D:/Program Files/R/R-3.1.1/library")
library("stats", lib.loc="D:/Program Files/R/R-3.1.1/library")
install.packages("Rcmdr")
library("Rcmdr", lib.loc="D:/Program Files/R/R-3.1.1/library")
install.packages("rattle")
library("rattle", lib.loc="D:/Program Files/R/R-3.1.1/library")
```



```
package 'rattle' successfully unpacked and MD5 sums checked
package 'leaps' successfully unpacked and MD5 sums checked
package 'effects' successfully unpacked and MD5 sums checked
package 'colorspace' successfully unpacked and MD5 sums checked
package 'alppack' successfully unpacked and MD5 sums checked
package 'RODBC' successfully unpacked and MD5 sums checked
package 'XLConnect' successfully unpacked and MD5 sums checked
```

The downloaded binary packages are in
C:\Users\dell\AppData\Local\Temp\Rtmp4Q8Z2F\downloaded_packages

RcmdrMsg: [1] NOTE: R Commander Version 2.1-3: Fri Oct 17 14:54:03 2014

Rcmdr version 2.1-3

```
> install.packages("rattle")
trying URL 'http://cran.rstudio.com/bin/windows/contrib/3.1/rattle_3.3.0.zip'
Content type 'application/zip' length 3211375 bytes (3.1 Mb)
opened URL
downloaded 3.1 Mb
```

package 'rattle' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
C:\Users\dell\AppData\Local\Temp\Rtmp4Q8Z2F\downloaded_packages

```
> library("rattle", lib.loc="D:/Program Files/R/R-3.1.1/library")
```
- Environment Tab:** Shows the current R environment with the 'rattle' package installed.
- Packages Tab:** Displays a list of installed packages, including rattle, parallel, multcomp, mvtnorm, nlme, nnet, RColorBrewer, relimp, rgl, rJava, RODBC, rpart, rstudio, sandwich, and nla.
- Dropbox Screenshot:** A tooltip from the Dropbox icon in the taskbar indicates "Screenshot Added" and "A screenshot was added to your Dropbox."

Installation of Rattle

The screenshot shows the RStudio interface with the following components:

- Top Bar:** RStudio, File, Edit, Code, View, Plots, Session, Build, Debug, Tools, Help.
- Console Tab:** Shows the command-line output of the R session.
- Code Editor Tab:** Untitled1, showing R code for library loading and package installation.
- Environment Tab:** Shows the current environment with loaded packages: rstudio, stats, Rcmdr, Rcmdr, rattle, and rattle.
- Packages Tab:** Shows a list of installed packages with their descriptions and versions.

Console Output (Top Level):

```
1:1 | (Top Level) + R Script +
```

```
Console ~/ ~ 5.1 MD
DOWNLOADED 5.1 MD
package 'rattle' successfully unpacked and MD5 sums checked
The downloaded binary packages are in
  C:/Users/dell/AppData/Local/Temp/Rtmp4Q8Z2F/downloaded_packages
> library("rattle", lib.loc="D:/Program Files/R/R-3.1.1/library")
Rattle: A free graphical interface for data mining with R.
Version 3.3.0 copyright (c) 2006-2014 Togaware Pty Ltd.
Type 'rattle()' to shake, rattle, and roll your data.
> rattle()
The package 'RGtk2' is required to display the Rattle GUI. It does not
appear to be installed. This package (and its dependencies) can be
installed using the following R command:
install.packages('RGtk2')

This one-time install will allow access to the full functionality of
Rattle.

Would you like Rattle to install the package now?
(yes/NO) yes
trying URL 'http://cran.rstudio.com/bin/windows/contrib/3.1/RGtk2_2.20.31.zip'
Content type 'application/zip' length 13884133 bytes (13.2 Mb)
opened URL
```

Code Editor (Untitled1):

```
library("rstudio", lib.loc="D:/Program Files/R/R-3.1.1/library")
library("stats", lib.loc="D:/Program Files/R/R-3.1.1/library")
install.packages("Rcmdr")
library("Rcmdr", lib.loc="D:/Program Files/R/R-3.1.1/library")
install.packages("rattle")
library("rattle", lib.loc="D:/Program Files/R/R-3.1.1/library")
rattle()
```

Packages Tab:

Name	Description	Version
multcomp	Simultaneous Inference in General Parametric Models	1.3-7
mvtnorm	Multivariate Normal and t Distributions	1.0-0
nlme	Linear and Nonlinear Mixed Effects Models	3.1-117
nnet	Feed-forward Neural Networks and Multinomial Log-Linear Models	7.3-8
parallel	Support for Parallel computation in R	3.1.1
<input checked="" type="checkbox"/> rattle	Graphical user interface for data mining in R	3.3.0
Rcmdr	R Commander	2.1-3
RcmdrMisc	R Commander Miscellaneous Functions	1.0-2
RColorBrewer	ColorBrewer palettes	1.0-5
relimp	Relative Contribution of Effects in a Regression Model	1.0-3
rgl	3D visualization device system (OpenGL)	0.94.1143
rJava	Low-level R to Java interface	0.9-6
RODBC	ODBC Database Access	1.3-10
rpart	Recursive Partitioning and Regression Trees	4.1-8
rstudio	Tools and Utilities for RStudio	0.98.1074
sandwich	Robust Covariance Matrix Estimators	2.3-2

Installation of Rattle

The screenshot shows the RStudio interface with the following details:

- Console:** Displays the R command to install the rattle package and its dependencies.
- Progress Bar:** Shows "90% downloaded" for the file URL: ... //cran.rstudio.com/bin/windows/contrib/3.1/RGtk2_2.20.31.zip.
- Environment:** Shows the current library path: D:/Program Files/R/R-3.1.1/library.
- File Explorer:** Shows the contents of the library directory, including packages like nlme, nnet, parallel, rattle, Rcmdr, RcmdrMisc, RColorBrewer, relimp, rgl, rJava, RODBC, rpart, rstudio, and sandwich.

```
library("rstudio", lib.loc="D:/Program Files/R/R-3.1.1/library")
library("stats", lib.loc="D:/Program Files/R/R-3.1.1/library")
install.packages("Rcmdr")
library("Rcmdr", lib.loc="D:/Program Files/R/R-3.1.1/library")
install.packages("rattle")
```

```
1:1 | (Top Level) +
```

```
Console ~/ ~
```

```
DOWNLOADED 5.1 MB
```

```
package 'rattle' successfully
```

```
The downloaded binary packages were:
C:/Users/dell/AppData/Local/Temp/Rtmp4Q8Z2F/downloaded_packages
```

```
> library('rattle', lib.loc="D:/Program Files/R/R-3.1.1/library")
```

```
Rattle: A free graphical interface for data mining with R.
```

```
Version 3.3.0 copyright (c) 2006-2014 Togaware Pty Ltd.
```

```
Type 'rattle()' to shake, rattle, and roll your data.
```

```
> rattle()
```

```
The package 'RGtk2' is required to display the Rattle GUI. It does not appear to be installed. This package (and its dependencies) can be installed using the following R command:
```

```
install.packages('RGtk2')
```

```
This one-time install will allow access to the full functionality of Rattle.
```

```
would you like Rattle to install the package now?
```

```
(yes/NO) yes
```

```
trying URL 'http://cran.rstudio.com/bin/windows/contrib/3.1/RGtk2_2.20.31.zip'
```

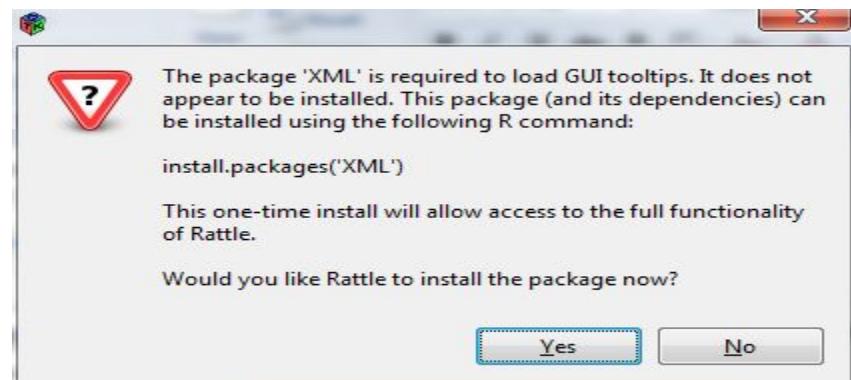
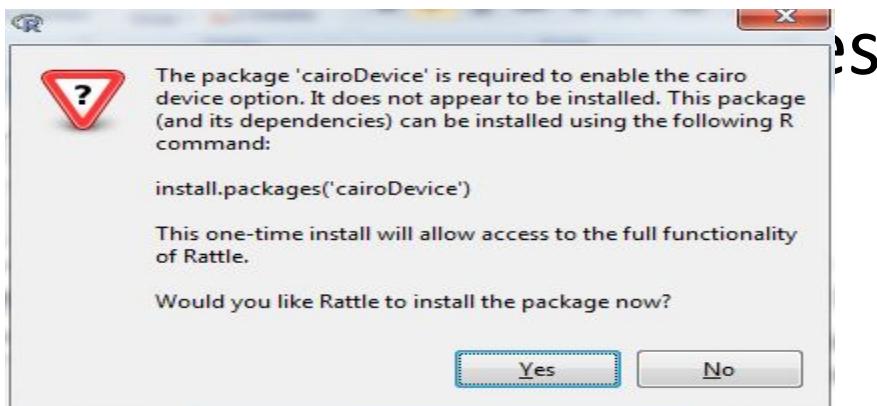
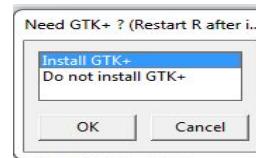
```
Content type 'application/zip' length 13884133 bytes (13.2 Mb)
```

```
opened URL
```

Name	Description	Version
nlme	Linear and Nonlinear Mixed Effects Models	3.1-7
nnet	Feed-forward Neural Networks and Multinomial Log-Linear Models	1.0-0
parallel	Support for Parallel computation in R	3.1-117
<input checked="" type="checkbox"/> rattle	Graphical user interface for data mining in R	3.3.0
Rcmdr	R Commander	2.1-3
RcmdrMisc	R Commander Miscellaneous Functions	1.0-2
RColorBrewer	ColorBrewer palettes	1.0-5
relimp	Relative Contribution of Effects in a Regression Model	1.0-3
rgl	3D visualization device system (OpenGL)	0.94.1143
rJava	Low-level R to Java interface	0.9-6
RODBC	ODBC Database Access	1.3-10
rpart	Recursive Partitioning and Regression Trees	4.1-8
rstudio	Tools and Utilities for RStudio	0.98.1074
sandwich	Robust Covariance Matrix Estimators	2.3-2

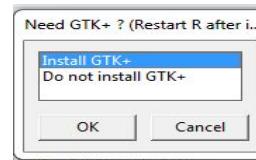
Installation of Rattle

- GTK+ Installation Necessary

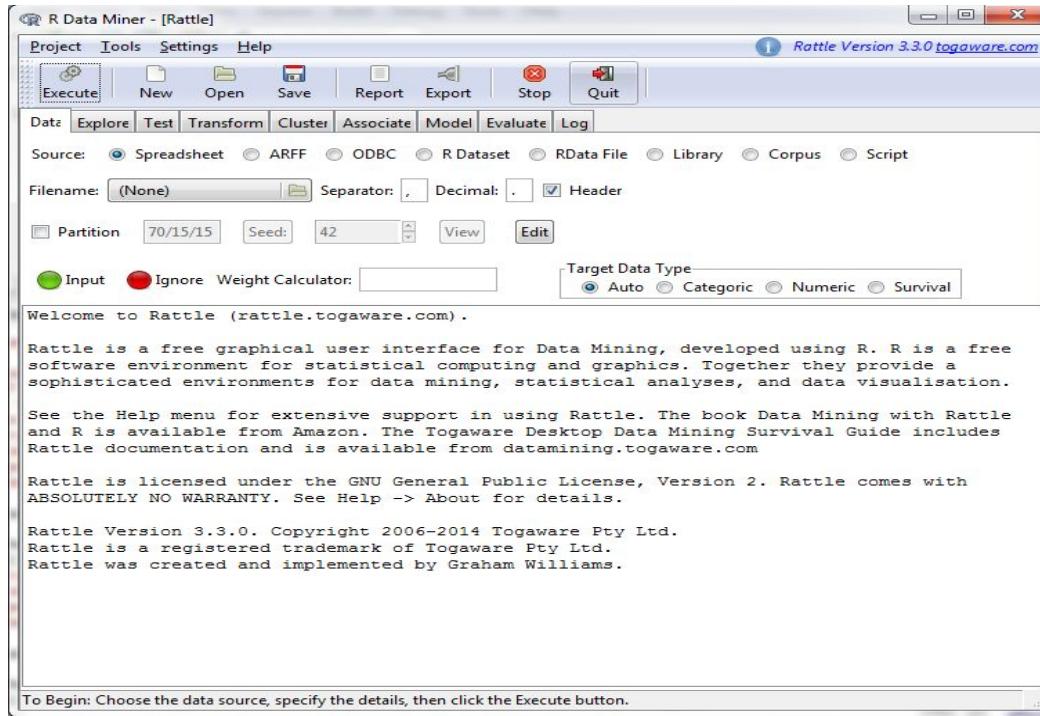


Installation of Rattle

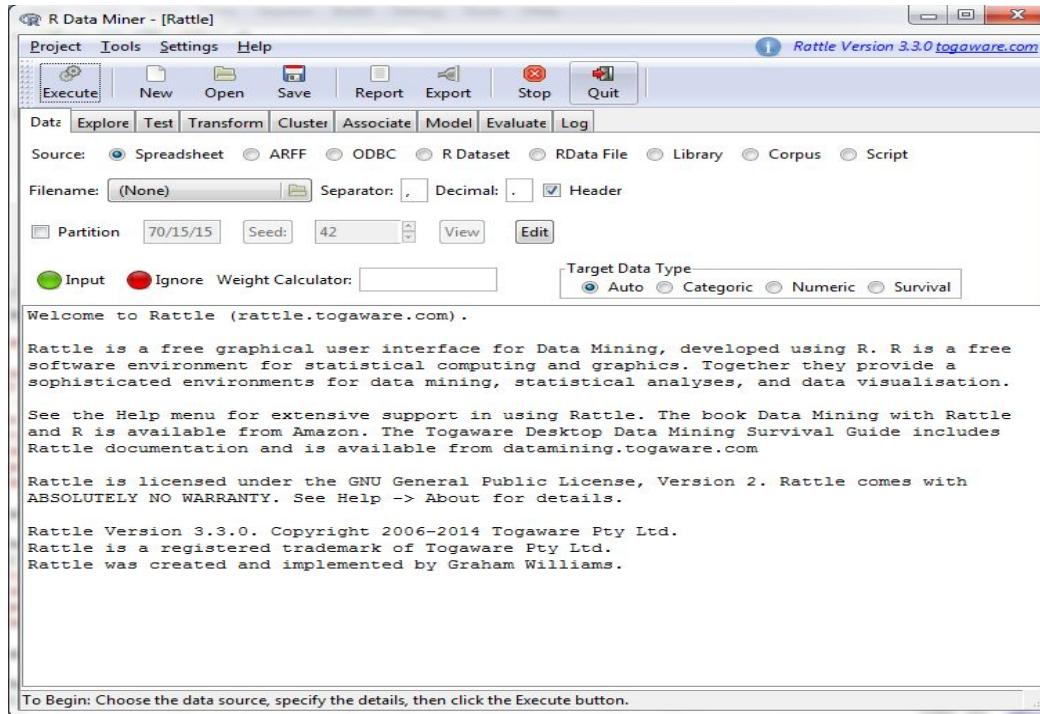
- GTK+ Installation Necessary



Overview of Rattle



Demo Rattle



RStudio

RStudio Desktop enables you with following advantages of native R console

- Syntax highlighting, code completion, and smart indentation
- Execute R code directly from the source editor
- Quickly jump to function definitions
- Easily manage multiple working directories using projects
- Integrated R help and documentation
- Interactive debugger to diagnose and fix errors quickly
- Extensive package development tools

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

RStudio

RStudio Server enables you to provide a browser based interface (the RStudio IDE) to a version of R running on a remote Linux server. Deploying R and RStudio on a server has a number of benefits, including:

- The ability to access your R workspace from any computer in any location;
- Easy sharing of code, data, and other files with colleagues;
- Allowing multiple users to share access to the more powerful compute resources (memory, processors, etc.) available on a well equipped server; and
- Centralized installation and configuration of R, R packages, TeX, and other supporting libraries.

new2.R x packages.R x chapter1.Rmd x Untitled1* x

Run Source

```

1 library(ggplot2)
2 data(diamonds)
3 barplot(diamonds$price)
4 plot(diamonds$price)
5 plot(diamonds$price,diamonds$carat)
6 pie(table(diamonds$cut))
7 boxplot(diamonds$price)
8 boxplot(diamonds$price-diamonds$cut)
9 boxplot(diamonds$price-diamonds$color)
10 plot(diamonds$cut,diamonds$color)
11 hist(diamonds$price)
12
13
12:1 (Top Level) ▾

```

RStudio - Interface

Console ~/ ↵

>

kmeans

```

function (x, centers, iter.max = 10, nstart = 1, algorithm = c("Hartigan-Wong",
  "Lloyd", "Forgy", "MacQueen"), trace = FALSE)
{
  do_one <- function(nmeth) {
    switch(nmeth, {
      isteps.Qtran <- 50 * m
      iTran <- c(as.integer(isteps.Qtran), integer(max(0,
        k - 1)))
      Z <- .Fortran(C_kmns, x, m, p, centers = centers,
        as.integer(k), c1 = integer(m), c2 = integer(m),
        nc = integer(k), double(k), double(k), ncp = integer(k),
        D = double(m), iTran = iTran, live = integer(k),
        iter = iter.max, wss = double(k), ifault = as.integer(trace))
      switch(Z$ifault, stop("empty cluster: try a better set of initial centers",
        call. = FALSE), Z$iter <- max(Z$iter, iter.max +
          1L), stop("number of cluster centres must lie between 1 and nrow(x)",
        call. = FALSE), warning(gettextf("Quick-TRANSFER stage steps exceeded maximum (%d)",
          isteps.Qtran), call. = FALSE))
    }, {
      Z <- .C(C_kmeans_Lloyd, x, m, p, centers = centers,
        k, c1 = integer(m), iter = iter.max, nc = integer(k),
        wss = double(k))
    })
  }
}
```

Environment History

Import Dataset Clear

Global Environment

Data

diamonds	53940 obs. of 10 variables
iris3	50 obs. of 12 variables

Values

a	NULL (empty)
i	90L

Files Plots Packages Help Viewer

R: Search Results Find In Topic

Search Results



The search string was "kmeans"

Vignettes:

[broom::kmeans](#) kmeans with dplyr+broom

[HTML](#) [source](#) [R code](#)

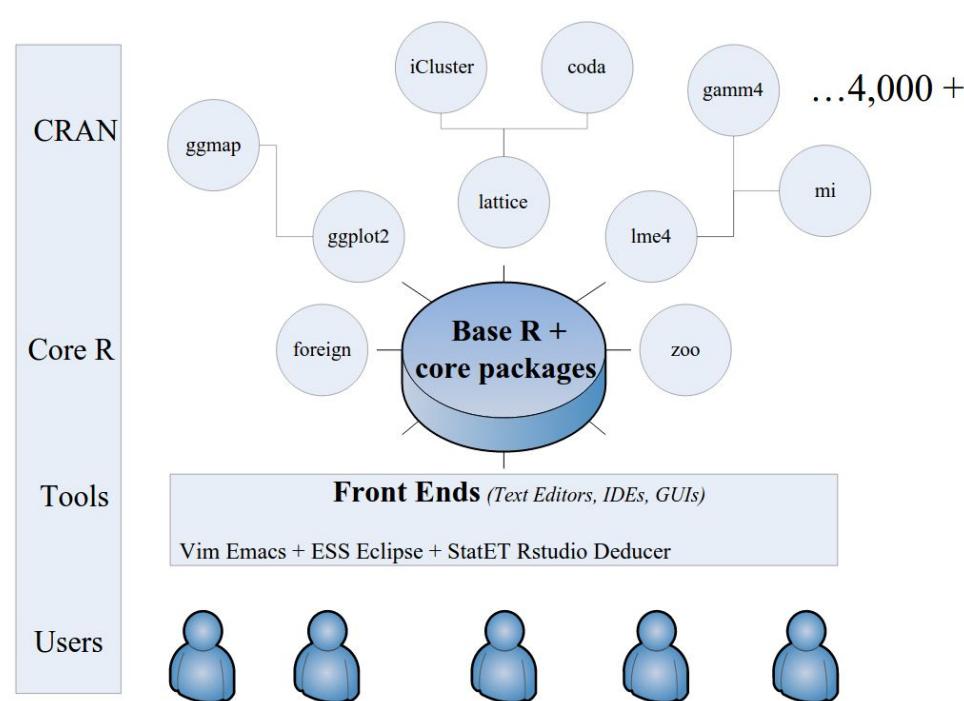
Help pages:

[amap::Kmeans](#) K-Means Clustering

[broom::augment.kmeans](#) Tidying methods for kmeans objects

[e1071::cmeans](#) Fuzzy C-Means Clustering

R Landscape



R Documentation

<http://cran.r-project.org/manuals.html>

Manuals

edited by the R Development Core Team.

The following manuals for R were created on Debian Linux and may differ from the manuals for Mac or Windows on platform. Version of the manuals for each platform are part of the respective R installations. The manuals change with R, hence we provide version for the patched release version (R-patched) and finally a version for the forthcoming R version that is still in development.

Here they can be downloaded as PDF files, EPUB files, or directly browsed as HTML:

Manual	R-release
An Introduction to R is based on the former "Notes on R", gives an introduction to the language and how to use R for doing statistical analysis and graphics.	HTML PDF EPUB
R Data Import/Export describes the import and export facilities available either in R itself or via packages which are available from CRAN.	HTML PDF EPUB
R Installation and Administration	HTML PDF EPUB
Writing R Extensions covers how to create your own packages, write R help files, and the foreign language (C, C++, Fortran, ...) interfaces.	HTML PDF EPUB
A draft of The R language definition documents the language <i>per se</i> . That is, the objects that it works on, and the details of the expression evaluation process, which are useful to know when programming R functions.	HTML PDF EPUB
R Internals : a guide to the internal structures of R and coding standards for the core team working on R itself.	HTML PDF EPUB
The R Reference Index : contains all help files of the R standard and recommended packages in printable form. (9MB, approx. 3500 pages)	PDF

Translations of manuals into other languages than English are available from the [contributed documentation](#) section (only a few are currently available).

The LaTeX or Texinfo sources of the latest version of these documents are contained in every R source distribution (in the `doc/manuals` directory). They can be found in the respective [archives of the R sources](#). The HTML versions of the manuals are also part of most R installation distributions.

Please check the manuals for R-devel before reporting any issues with the released versions.

R

Documentation

Vignettes

ggplot2: An Implementation of the Grammar of Graphics

An implementation of the grammar of graphics in R. It combines the advantages of both base and lattice graphics by step from multiple data sources. It also implements a sophisticated multidimensional conditioning system for information, documentation and examples.

Version:	1.0.1
Depends:	R (≥ 2.14), stats, methods
Imports:	plyr (≥ 1.7.1), digest , grid , gttable (≥ 0.1.1), reshape2 , scales (≥ 0.2.3), proto , MASS
Suggests:	quantreg , Hmisc , mapproj , maps , hexbin , maptools , multcomp , nlme , testthat , knitr , mgcv
Enhances:	sp
Published:	2015-03-17
Author:	Hadley Wickham [aut, cre], Winston Chang [aut]
Maintainer:	Hadley Wickham <h.wickham at gmail.com>
BugReports:	https://github.com/hadley/ggplot2/issues
License:	GPL-2
URL:	http://ggplot2.org , https://github.com/hadley/ggplot2
NeedsCompilation:	no
Citation:	ggplot2 citation info
Materials:	README NEWS
In views:	Graphics , Phylogenetics
CRAN checks:	ggplot2 results

Downloads :

Reference manual:	ggplot2.pdf
Vignettes:	Contributing to ggplot2 development ggplot2 release process
Package source:	ggplot2_1.0.1.tar.gz
Windows binaries:	r-devel: ggplot2_1.0.1.zip , r-release: ggplot2_1.0.1.zip , r-oldrel: ggplot2_1.0.1.zip
OS X Snow Leopard binaries:	r-release: not available, r-oldrel: ggplot2_1.0.1.tgz
OS X Mavericks binaries:	r-release: ggplot2_1.0.1.tgz
Old sources:	ggplot2 archive

Reverse dependencies :

Reverse depends: [alphahull](#), [AmpliconDuo](#), [aoristic](#), [apsimr](#), [bcrm](#), [bde](#), [benchmark](#), [biomod2](#), [bootnet](#), [brms](#)

CRAN Views

<http://cran.r-project.org/web/views/>

Bayesian	Bayesian Inference
ChemPhys	Chemometrics and Computational Physics
ClinicalTrials	Clinical Trial Design, Monitoring, and Analysis
Cluster	Cluster Analysis & Finite Mixture Models
DifferentialEquations	Differential Equations
Distributions	Probability Distributions
Econometrics	Econometrics
Environmetrics	Analysis of Ecological and Environmental Data
ExperimentalDesign	Design of Experiments (DoE) & Analysis of Experimental Data
Finance	Empirical Finance
Genetics	Statistical Genetics
Graphics	Graphic Displays & Dynamic Graphics & Graphic Devices & Visualization
HighPerformanceComputing	High-Performance and Parallel Computing with R
MachineLearning	Machine Learning & Statistical Learning
MedicalImaging	Medical Image Analysis
MetaAnalysis	Meta-Analysis
Multivariate	Multivariate Statistics
NaturalLanguageProcessing	Natural Language Processing
NumericalMathematics	Numerical Mathematics
OfficialStatistics	Official Statistics & Survey Methodology
Optimization	Optimization and Mathematical Programming
Pharmacokinetics	Analysis of Pharmacokinetic Data
Phylogenetics	Phylogenetics, Especially Comparative Methods
Psychometrics	Psychometric Models and Methods
ReproducibleResearch	Reproducible Research
Robust	Robust Statistical Methods
SocialSciences	Statistics for the Social Sciences
Spatial	Analysis of Spatial Data
SpatioTemporal	Handling and Analyzing Spatio-Temporal Data
Survival	Survival Analysis
TimeSeries	Time Series Analysis
WebTechnologies	Web Technologies and Services
gR	gRaphical Models in R

R Community

- email groups <http://www.r-project.org/mail.html>

R-announce

R-help

R-package-devel

R-devel

R-packages

Special Interest Groups

- Stack Overflow [r]
- Twitter #rstats
- Blogs at <http://www.r-bloggers.com/> (573 blogs)

Stack Overflow

<http://stackoverflow.com/questions/tagged/r>

The screenshot shows the Stack Overflow homepage with the search bar set to 'r'. The main content area displays 'Tagged Questions' for the 'r' tag. There are three visible posts:

- R Count number of rows in one column of a data frame?**
I just want to know how to get r to list the number of occupied rows of a specific column of a data frame. My guess was nrow(dataframe\$column) though that didn't work.
asked 2 mins ago by RyanMe321
1 answer, 3 views
- Create interactive webmap with markers in R using Shiny, Leaflet and rCharts**
I am trying to create an interactive webmap in R to display storms using Shiny, Leaflet and rCharts (the structure is loosely based on the <http://ramnathv.github.io/bikeshare app>). The idea is that ...
asked 5 mins ago by Louise
0 votes, 0 answers, 2 views
- R - gsub a specific character of a specific position**
I would like to delete the last character of a variable. I was wondering if it is possible to select the position with gsub and delete the character of this particular position. In this example, I ...
asked 15 mins ago by giacomoV
0 votes, 0 answers, 7 views

On the right side, there are sidebar sections:

- 90,861 questions tagged**
- [about »](#)
- Featured on Meta**
 - April 2015 Community Moderator Election Results
 - Hot Meta Posts
 - Failed edit to a question says: "Your answer couldn't be submitted"
 - The Font Awesome child tags are too specific - are they even necessary?
 - Flagging questions with details only in comments
- Favorite Tags** [edit](#)
[Add a favorite tag](#)
- Looking for a job?**
 - Chief Software Architect - Java - \$100K
 - Crossover
 - Bengaluru, India / remote
 - adobe, netbeans

Twitter

<https://twitter.com/search?q=rstats&src=spqr>

...

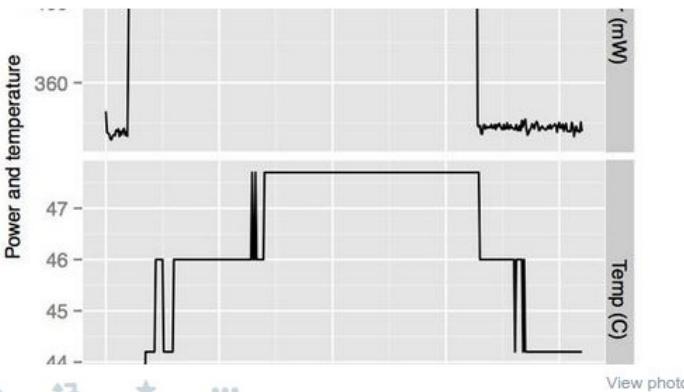
Results for #rstats

Top / All



Mark Benson @markbenson · 5m

Power and heat are related. Here's an R plot I did that proves it on the Kindle Fire. [#rstats vanilladraft.com/stmes/](http://vanilladraft.com/stmes/)



Stéphane Fréchette @sfrechette · 8m

How to get your very own RStudio Server and Shiny Server with DigitalOcean r-bloggers.com/how-to-get-you... #datascience #feedly #rstats #shiny



Ankit kansal @sinisterinankit · 9m

Interesting post on configuring parallel computing on #r #rstudio #rstats #dataprocessing #data



Learn R @R_Programming

How to do parallel computing with R? rstatistics.net/parallel-compu...
#rstats #datascience

Help within R

? "keyword"

?? "keyword"

Example-

```
> ?kmeans
```

```
> ??kmeans
```

Functions Used in this Lesson

function(x)

for

library

install.packages

update.packages

ls

rm

print

Citations and References

> citation()

To cite R in publications use:

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>.

Introductory R

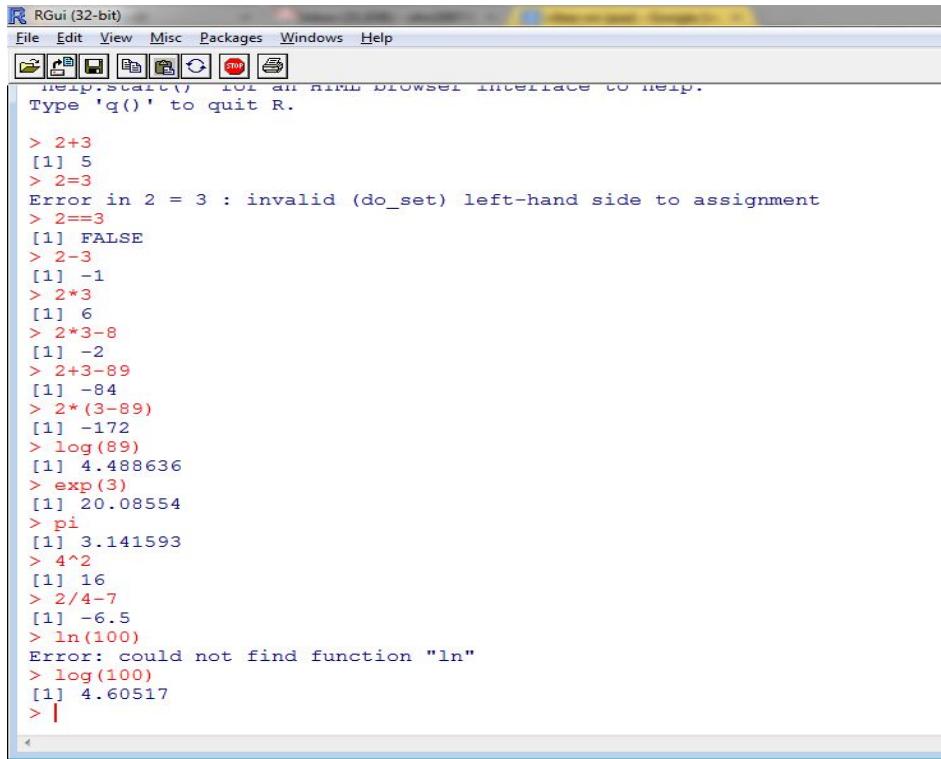
```
> Sys.Date()  
[1] "2015-05-10"  
> Sys.time()  
[1] "2015-05-10 18:28:32 IST"
```

R as a Calculator

Basic Math on R Console

- +
 - -
 - Log
 - Exp
 - *
 - /
 - ()
- mean
 - sum
 - sd
 - log
 - median
 - exp

Demo- Basic Math on R Console



R Gui (32-bit)

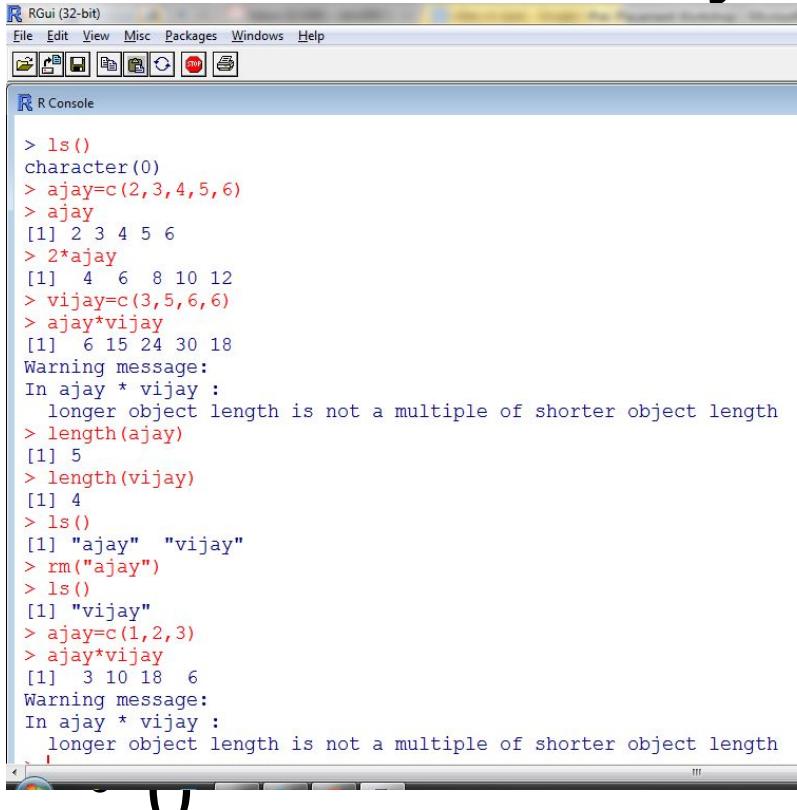
File Edit View Misc Packages Windows Help

help.start() for an html browser interface to help.
Type 'q()' to quit R.

```
> 2+3
[1] 5
> 2=3
Error in 2 = 3 : invalid (do_set) left-hand side to assignment
> 2==3
[1] FALSE
> 2-3
[1] -1
> 2*3
[1] 6
> 2*3-8
[1] -2
> 2+3-89
[1] -84
> 2*(3-89)
[1] -172
> log(89)
[1] 4.488636
> exp(3)
[1] 20.08554
> pi
[1] 3.141593
> 4^2
[1] 16
> 2/4-7
[1] -6.5
> ln(100)
Error: could not find function "ln"
> log(100)
[1] 4.60517
> |
```

Hint- Ctrl +L clears screen

Demo- Basic Objects on R Console



```
R Gui (32-bit)
File Edit View Misc Packages Windows Help
R Console

> ls()
character(0)
> ajay=c(2,3,4,5,6)
> ajay
[1] 2 3 4 5 6
> 2*ajay
[1] 4 6 8 10 12
> vijay=c(3,5,6,6)
> ajay*vijay
[1] 6 15 24 30 18
Warning message:
In ajay * vijay :
  longer object length is not a multiple of shorter object length
> length(ajay)
[1] 5
> length(vijay)
[1] 4
> ls()
[1] "ajay" "vijay"
> rm("ajay")
> ls()
[1] "vijay"
> ajay=c(1,2,3)
> ajay*vijay
[1] 3 10 18 6
Warning message:
In ajay * vijay :
  longer object length is not a multiple of shorter object length
I
```

Functions-

ls() – what objects are here

rm("foo") removes object named foo

Assignment

Using = or -> assigns object names to values

Hint- Up arrow ↑ gives you last typed command

Functions and Loops

- Loops

```
for (number in 1:5){ print (number) }
```

```
> for (number in 1:5){ print (number) }
[1] 1
[1] 2
[1] 3
[1] 4
[1] 5
> for (i in 1:5){ print (i) }
[1] 1
[1] 2
[1] 3
[1] 4
[1] 5
> for (i in 1:5){ rnorm(i,10,10) }
> for (i in 1:5){ print(rnorm(i,10,10)) }
[1] 1.090406
[1] 8.611727 16.670168
[1] 10.84623 13.13938 11.56230
[1] 6.068250 -18.723389 33.174107 -1.320091
[1] 13.939702 -9.037375 13.755986 9.459680 9.625309
> |
```

Functions and Loops

- Function

```
functionajay=function(a)(a^2+2*a+1)
```

```
> -----  
> functionajay=function(a) (a^2+2*a+1)  
[1] 4  
> functionajay(2)  
[1] 9  
> for (i in 1:5){ print(rnorm(i) )  
Error: unexpected '}' in "for (i in 1:5){ print(rnorm(i) )"  
>  
> for (i in 1:5){ print(functionajay(i)) }  
[1] 4  
[1] 9  
[1] 16  
[1] 25  
[1] 36  
> |
```

Hint: Always match brackets

Each (deserves a)

Each { deserves a }

Each [deserves a]

Other sources to learn R

swirlstats

<http://swirlstats.com/>

datacamp

<https://www.datacamp.com/>

codeschool

<http://tryr.codeschool.com/>

coursera

<https://www.coursera.org/course/compdata>

<https://www.coursera.org/course/rprog>



Good coding practices

- Use # for comment
- Use git for version control
- Use Rstudio for multiple lines of code

Functions in R

- custom functions
- source code for a function

```
Console ~/ ◁
> kmeans
function (x, centers, iter.max = 10, nstart = 1, algorithm = c("Hartigan-Wong",
  "Lloyd", "Forgy", "MacQueen"), trace = FALSE)
{
  do_one <- function(nmeth) {
    switch(nmeth, {
      isteps.Qtran <- 50 * m
      iTran <- c(as.integer(isteps.Qtran), integer(max(0,
        k - 1)))
      Z <- .Fortran(C_kmeans, x, m, p, centers = centers,
        as.integer(k), c1 = integer(m), c2 = integer(m),
        nc = integer(k), double(k), double(k), ncp = integer(k),
        D = double(m), iTran = iTran, live = integer(k),
        iter = iter.max, wss = double(k), ifault = as.integer(trace))
      switch(Z$ifault, stop("empty cluster: try a better set of initial centers",
        call. = FALSE), Z$iter <- max(Z$iter, iter.max +
          1L), stop("number of cluster centres must lie between 1 and nrow(x)",
        call. = FALSE), warning(gettextf("Quick-TRANSFER stage steps exceeded maximum (= %d",
        isteps.Qtran), call. = FALSE))
    }, {
      Z <- .C(C_kmeans_Lloyd, x, m, p, centers = centers,
        k, c1 = integer(m), iter = iter.max, nc = integer(k),
        wss = double(k))
    }, {
      Z <- .C(C_kmeans_MacQueen, x, m, p, centers = as.double(centers),
        k, c1 = integer(m), iter = iter.max, nc = integer(k),
        wss = double(k))
    })
    if (m23 <- any(nmeth == c(2L, 3L))) {
      if (any(Z$nc == 0))
        warning("empty cluster: try a better set of initial centers",
          call. = FALSE)
    }
  }
}
```

HOMEWORK TIME !



Learning Objectives

- how to input data in R using various ways
- how to check for correct data input
- how to use special packages for fast data input
- how to input data from statistical file formats
- how to input data from databases
- how to input data from web (web scraping)

What will you learn from this lesson

- data input from various kinds of format
- efficient data input via various packages
- sql to R
- web scraping
- piping in R
- using json in R

Environment

ls() -lists objects

rm()-removes an object

gc() -does garbage collection and frees up
memory

Console ~/ ↗

Copyright (C) 2015 The R Foundation for Statistical Computing
Platform: i686-pc-linux-gnu (32-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from ~/.RData]

```
> ls()
[1] "a"      "i"      "iris3"
> rm(a)
> ls()
[1] "i"      "iris3"
> gc()
    used (Mb) gc trigger (Mb) max used (Mb)
Ncells 334887  9.0      597831 16.0    407500 10.9
Vcells 624499  4.8     1215808  9.3    1215802  9.3
>
```



Environment History

Import Dataset Clear

Global Environment



List

Data

iris3 50 obs. of 12 variables

Values

i 90L

Files Plots Packages Help Viewer

← → Home ↗ ↘

R: Search Results Find In Topic

Search Results



The search string was "kmeans"

Vignettes:

[broom::kmeans](#) kmeans with dplyr+broom[HTML](#) [source](#) [R code](#)

Help pages:

[amap::Kmeans](#) K-Means Clustering[broom::augment.kmeans](#) Tidying methods for kmeans objects[e1071::cmeans](#) Fuzzy C-Means Clustering

Console ~/

Copyright (C) 2015 The R Foundation for Statistical Computing
Platform: i686-pc-linux-gnu (32-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from ~/.RData]

```
> ls()
[1] "a"      "i"      "iris3"
> rm(a)
> ls()
[1] "i"      "iris3"
> gc()
    used (Mb) gc trigger (Mb) max used (Mb)
Ncells 334887  9.0      597831 16.0   407500 10.9
Vcells 624499  4.8     1215808  9.3   1215802  9.3
>
```

The screenshot shows the RStudio IDE interface. On the left, the R console displays the standard R startup message and a workspace summary. On the right, the Environment pane shows the 'iris3' dataset with 50 observations and 12 variables. A large oval highlights the 'iris3' entry. Below the Environment pane, the 'Search Results' panel is open, showing results for the search term 'kmeans'. The results include links to 'broom::kmeans', 'amap::Kmeans', 'broom::augment.kmeans', and 'e1071::cmeans'. There are also tabs for 'HTML', 'source', and 'R code' for each result.

Environment History

Global Environment

Data

iris3 50 obs. of 12 variables

Values

i 90L

Files Plots Packages Help Viewer

R: Search Results Find In Topic

Search Results

The search string was "kmeans"

Vignettes:

broom::kmeans kmeans with dplyr+broom [HTML](#) [source](#) [R code](#)

Help pages:

amap::Kmeans K-Means Clustering [HTML](#)

broom::augment.kmeans Tidying methods for kmeans objects [HTML](#)

e1071::cmeans Fuzzy C-Means Clustering [HTML](#)

File System

`getwd()`- get working directory

`setwd()`- set or change working directory

`dir()` - lists files in working directory



Console ~/Desktop/new/ ↻

```
> getwd()  
[1] "/home/ajay/Desktop"  
> setwd("/home/ajay/Desktop/new")  
> dir()  
[1] "obama"  
> |
```

Environment History

Import Dataset Clear

Global Environment

Data

iris3 50 obs. of 12 variables

Values

i 90L

Files Plots Packages Help Viewer

New Folder Delete Rename More

Home

	Name	Size	Modified
<input type="checkbox"/>	.RData	3.8 KB	May 2, 2015, 11:47 AM
<input type="checkbox"/>	.Rhistory	10.7 KB	May 10, 2015, 2:20 PM
<input type="checkbox"/>	17811636-Brain-function-as-gears-and-cogs-in-the-shape-of-a-human-head-as-a-medical-symbol-of-mental-health-c-Stock-Photo.jpg	139.8 KB	Apr 16, 2015, 9:41 AM
<input type="checkbox"/>	21.png	352.7 KB	May 4, 2015, 5:18 PM
<input type="checkbox"/>	2167434.jpg	32.8 KB	May 4, 2015, 5:36 PM
<input type="checkbox"/>	a.out	7.7 KB	May 2, 2015, 2:26 PM
<input type="checkbox"/>	anaconda		
<input type="checkbox"/>	animation		
<input type="checkbox"/>	animation2		
<input type="checkbox"/>	backports-3.18.1-1		
<input type="checkbox"/>	backports-3.18.1-1.tar.xz	8.6 MB	Dec 22, 2014, 3:14 AM
<input type="checkbox"/>	Call R and Python from base SAS.html	48.4 KB	May 5, 2015, 12:53 PM

Console ~/Desktop/new/ ↻

```
> getwd()  
[1] "/home/ajay/Desktop"  
> setwd("/home/ajay/Desktop/new")  
> dir()  
[1] "obama"  
> |
```

Environment History

 Import Dataset  Clear

Global Environment

Data

iris3 50 obs. of 12 variables

Values

i 90L

Files Plots Packages Help Viewer

 New Folder  Delete  Rename  More

Home

Name	Size	Modified
.RData	3.8 KB	May 2, 2015, 11:47 AM
.Rhistory	10.7 KB	May 10, 2015, 2:20 PM
17811636-Brain-function-as-gears-and-cogs-in-the-shape-of-a-human-head-as-a-medical-symbol-of-mental-health-c-Stock-Photo.jpg	139.8 KB	Apr 16, 2015, 9:41 AM
21.png	352.7 KB	May 4, 2015, 5:18 PM
2167434.jpg	32.8 KB	May 4, 2015, 5:36 PM
a.out	7.7 KB	May 2, 2015, 2:26 PM
anaconda		
animation		
animation2		
backports-3.18.1-1		
backports-3.18.1-1.tar.xz	8.6 MB	Dec 22, 2014, 3:14 AM
Call R and Python from base SAS.html	48.4 KB	May 5, 2015, 12:53 PM



Console ~/Desktop/new/ ↻

```
> getwd()  
[1] "/home/ajay/Desktop"  
> setwd("/home/ajay/Desktop/new")  
> dir()  
[1] "obama"  
> |
```

Environment History

Import Dataset Clear
Global Environment

Data

iris3 50 obs. of 12 variables

Values

i 90L

Files Plots Packages Help Viewer

New Folder Delete Rename More

Home

	Name	Size	Modified
	.RData	3.8 KB	May 2, 2015, 11:47 AM
	.Rhistory	10.7 KB	May 10, 2015, 2:20 PM
	17811636-Brain-function-as-gears-and-cogs-in-the-shape-of-a-human-head-as-a-medical-symbol-of-mental-health-c-Stock-Photo.jpg	139.8 KB	Apr 16, 2015, 9:41 AM
	21.png	352.7 KB	May 4, 2015, 5:18 PM
	2167434.jpg	32.8 KB	May 4, 2015, 5:36 PM
	a.out	7.7 KB	May 2, 2015, 2:26 PM
	anaconda		
	animation		
	animation2		
	backports-3.18.1-1		
	backports-3.18.1-1.tar.xz	8.6 MB	Dec 22, 2014, 3:14 AM
	Call R and Python from base SAS.html	48.4 KB	May 5, 2015, 12:53 PM

Assigning

objectname=read.csv(filepath,parameters)

OR

objectname<-read.csv(filepath,parameters)

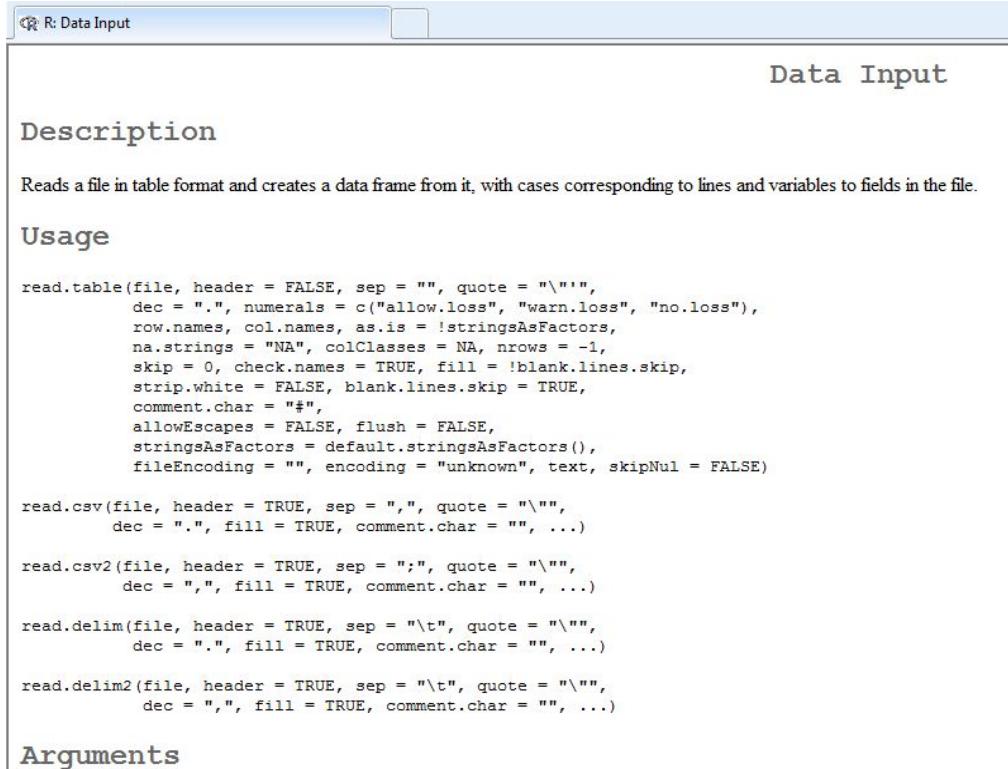
Data Input

`read.table()` or `read.csv()`

`read.spss()`

`read.sas7bdat()`

read.table()



R: Data Input

Data Input

Description

Reads a file in table format and creates a data frame from it, with cases corresponding to lines and variables to fields in the file.

Usage

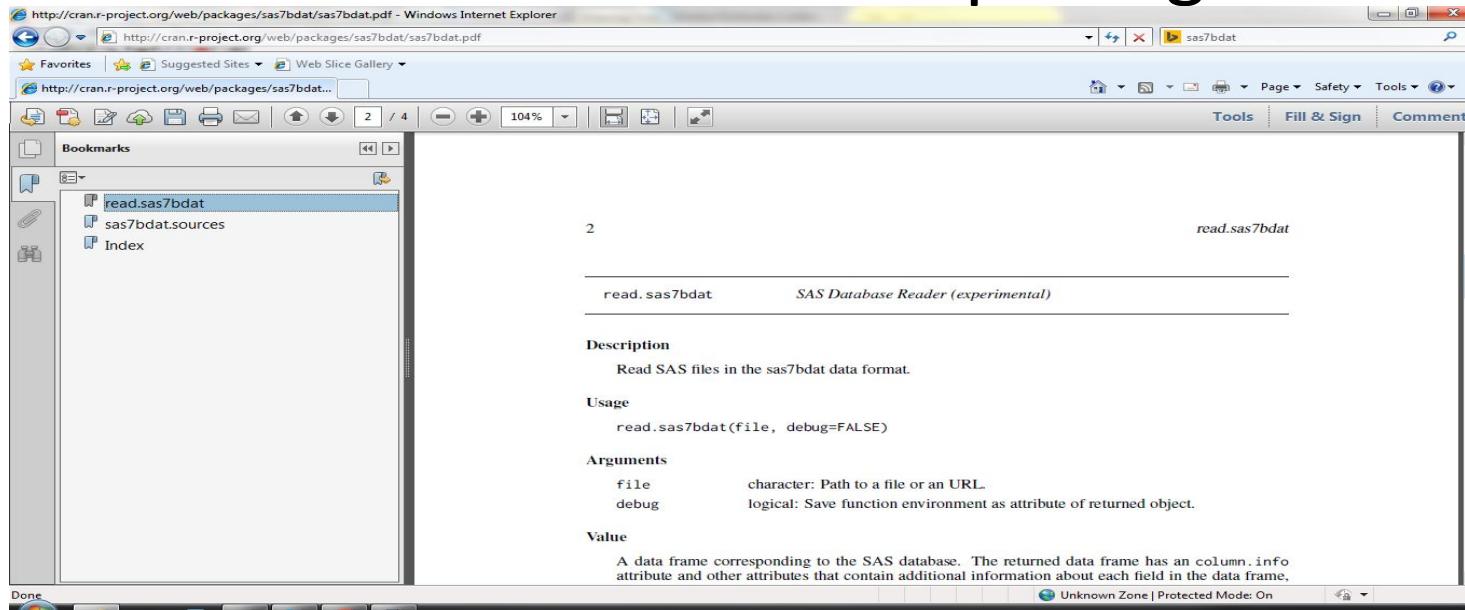
```
read.table(file, header = FALSE, sep = "", quote = "\"\"",  
          dec = ".", numerals = c("allow.loss", "warn.loss", "no.loss"),  
          row.names, col.names, as.is = !stringsAsFactors,  
          na.strings = "NA", colClasses = NA, nrow = -1,  
          skip = 0, check.names = TRUE, fill = !blank.lines.skip,  
          strip.white = FALSE, blank.lines.skip = TRUE,  
          comment.char = "#",  
          allowEscapes = FALSE, flush = FALSE,  
          stringsAsFactors = default.stringsAsFactors(),  
          fileEncoding = "", encoding = "unknown", text, skipNul = FALSE)  
  
read.csv(file, header = TRUE, sep = ",", quote = "\"\"",  
        dec = ".", fill = TRUE, comment.char = "", ...)  
  
read.csv2(file, header = TRUE, sep = ";", quote = "\"\"",  
          dec = ",", fill = TRUE, comment.char = "", ...)  
  
read.delim(file, header = TRUE, sep = "\t", quote = "\"\"",  
          dec = ".", fill = TRUE, comment.char = "", ...)  
  
read.delim2(file, header = TRUE, sep = "\t", quote = "\"\"",  
            dec = ",", fill = TRUE, comment.char = "", ...)
```

Arguments

<https://stat.ethz.ch/R-manual/R-devel/library/utils/html/read.table.html>

Statistical formats

- `read.spss` from `foreign` package
- `read.sas7bdat` from `sas7bdat` package



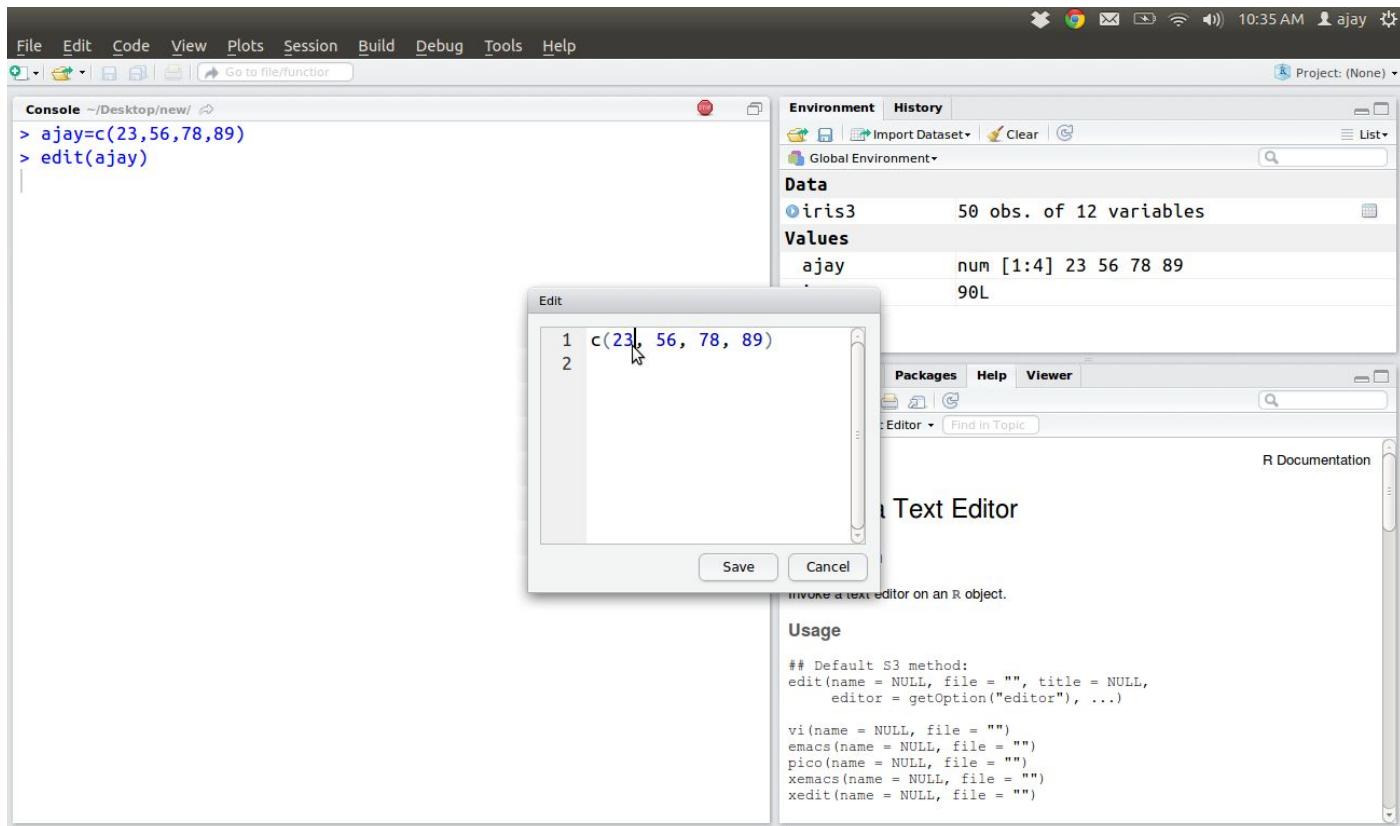
Manual Entry

The screenshot shows the RStudio interface with the following components:

- File Edit Code View Plots Session Build Debug Tools Help**: The main menu bar.
- Project: (None) ▾**: A dropdown menu for projects.
- Environment History**: A tabbed pane showing the environment and history.
- Data**: A section displaying the `iris3` dataset, which has 50 observations and 12 variables.
- Values**: A section showing the values of variables `ajay` and `i`.
- Files Plots Packages Help Viewer**: A tabbed pane for navigating between different types of files and plots.
- Console ~/Desktop/new/ ↵**: The console window showing the command history:

```
> ajay=c(22,56,78,89)
> View(ajay)
> |
```
- Go to file/function**: A search bar at the top of the interface.

Manual Editing



Manual Editing

The screenshot shows the R Data Editor interface. In the top-left, the R console displays the command `> edit(iris)`. The main area is a data grid titled "R Data Editor" showing the Iris dataset. The grid has columns: Sepal.Length, Sepal.Width, Petal.Length, Petal.Width, and Species. The first row is selected, indicated by a red border around the entire row. The bottom right corner of the grid contains the buttons "Copy", "Paste", and "Quit".

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa
13	4.8	3	1.4	0.1	setosa
14	4.3	3	1.1	0.1	setosa
15	5.8	4	1.2	0.2	setosa
16	5.7	4.4	1.5	0.4	setosa
17	5.4	3.9	1.3	0.4	setosa
18	5.1	3.5	1.4	0.3	setosa
19	5.7	3.8	1.7	0.3	setosa
20	5.1	3.8	1.5	0.3	setosa
21	5.4	3.4	1.7	0.2	setosa
22	5.1	3.7	1.5	0.4	setosa
23	4.6	3.6	1	0.2	setosa
24	5.1	3.3	1.7	0.5	setosa
25	4.8	3.4	1.9	0.2	setosa

The R environment pane shows the iris dataset has 50 observations and 12 variables. The global environment pane lists objects like `is3`, `ay`, `90L`, and `is`. The help pane shows the `edit` function documentation, which describes it as a text editor for R objects.

```
edit
Description
  Create a text editor on an R object.

Usage
  edit(x, ..., editor = getOption("editor"))

Default S3 method:
  (name = NULL, file = "", title = NULL,
   editor = getOption("editor"), ...)

  name = NULL, file = ""))
  emacs(name = NULL, file = "")
  pico(name = NULL, file = "")
  xemacs(name = NULL, file = "")
  xedit(name = NULL, file = "")
```

readr from Hadley

The goal of `readr` is to provide a fast and friendly way to read tabular data into R. The most important functions are:

- Read delimited files: `read_delim()`, `read_csv()`, `read_tsv()`, `read_csv2()`.
- Read fixed width files: `read_fwf()`, `read_table()`.
- Read lines: `read_lines()`.
- Read whole file: `read_file()`.
- Re-parse existing data frame: `type_convert()`.

<https://github.com/hadley/readr>

readr from Hadley

Source Data - <https://bit.ly/dsdata>

```
> library(readr)
> system.time(read_csv("BigDiamonds.csv"))
|=====| 100%   49 MB
  user  system elapsed
 2.396   0.068   2.448
Warning message:
597311 problems parsing 'BigDiamonds.csv'. See problems(...) for more details.
```

readxl from Hadley

Readxl supports both the legacy .xls format and the modern xml-based .xlsx format. .xls support is made possible with [libxls](#) C library, which abstracts away many of the complexities of the underlying binary format. To parse .xlsx, we use the [RapidXML](#) C++ library.

```
read_excel("my-old-spreadsheet.xls")
read_excel("my-new-spreadsheet.xlsx")

read_excel("my-spreadsheet.xls", sheet = "data")
read_excel("my-spreadsheet.xls", sheet = 2)

read_excel("my-spreadsheet.xls", na = "NA")
```

<https://github.com/hadley/readxl>

data.table

fread is the fastest way to read data

```
> b=fread("BigDiamonds.csv")
Read 598024 rows and 13 (of 13) columns from 0.049 GB file in 00:00:04
```

data.table

fread is the fastest way to read data

```
> b=fread("BigDiamonds.csv")
Read 598024 rows and 13 (of 13) columns from 0.049 GB file in 00:00:04
```

data.table

fread is the fastest way to read data

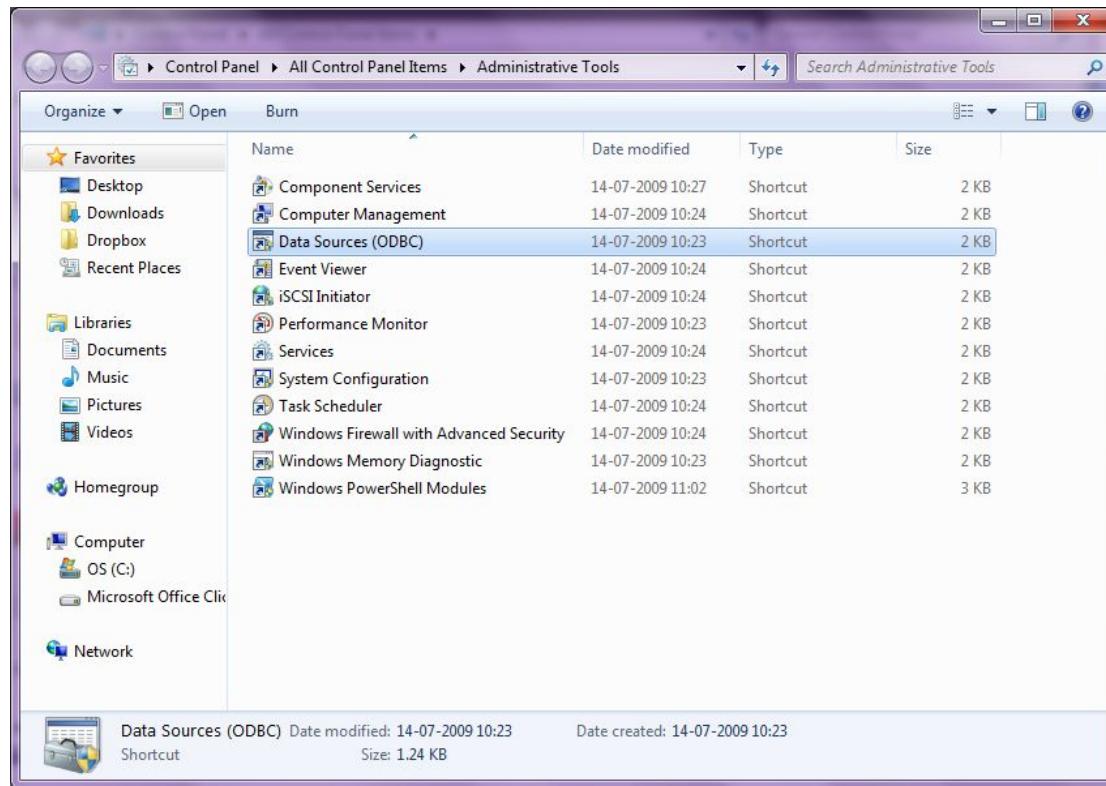
```
> system.time(read_csv("BigDiamonds.csv"))
  user  system elapsed
 2.552   0.028   2.581
Warning message:
597311 problems parsing 'BigDiamonds.csv'. See problems(...) for more details.
> system.time(fread("BigDiamonds.csv"))
  user  system elapsed
 1.532   0.012   1.540
> system.time(read.csv("BigDiamonds.csv"))
  user  system elapsed
10.892   0.032  10.922
```

Some learnings

1. Multiple packages can do the same thing faster or slower in R
2. Knowing the right package is the essential difference as a data scientist
3. Putting code within system.time() helps measure speed

also see <http://adv-r.had.co.nz/Profiling.html> for advanced ways to speed up code

Creating DSN (Optional)



Creating DSN (in Windows)

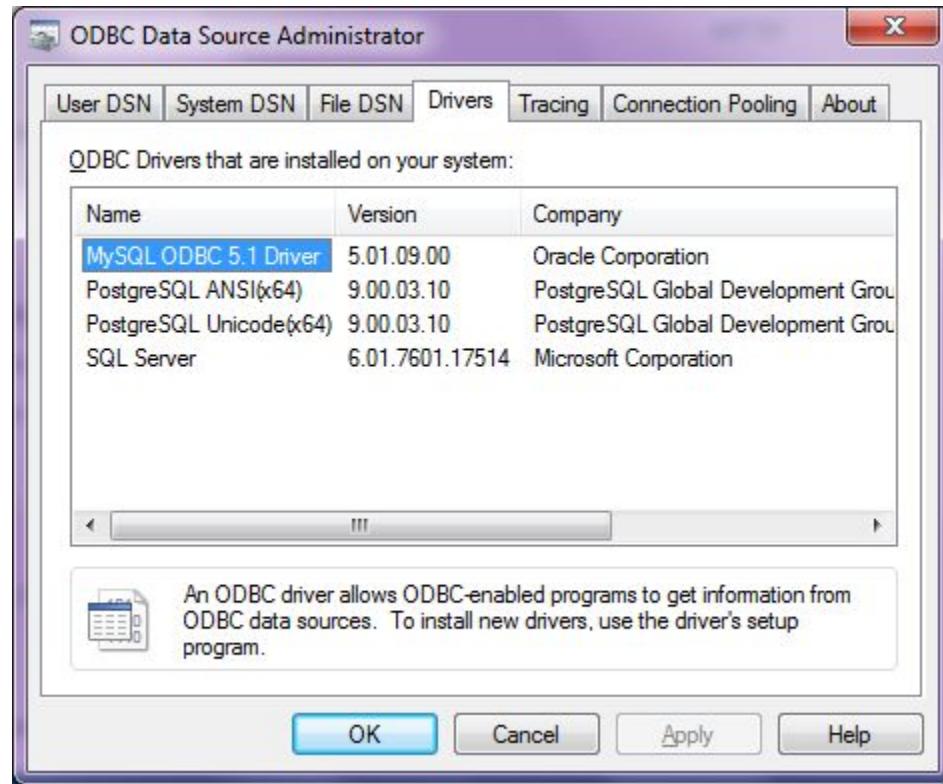
A Data Source Name (DSN) is the logical name that is used by Open Database Connectivity (ODBC) to refer to the drive and other information that is required to access data. The name is used by Internet Information Services for a connection to an ODBC data source, such as a Microsoft SQL Server database.

<https://support.microsoft.com/en-us/kb/kbview/300596>

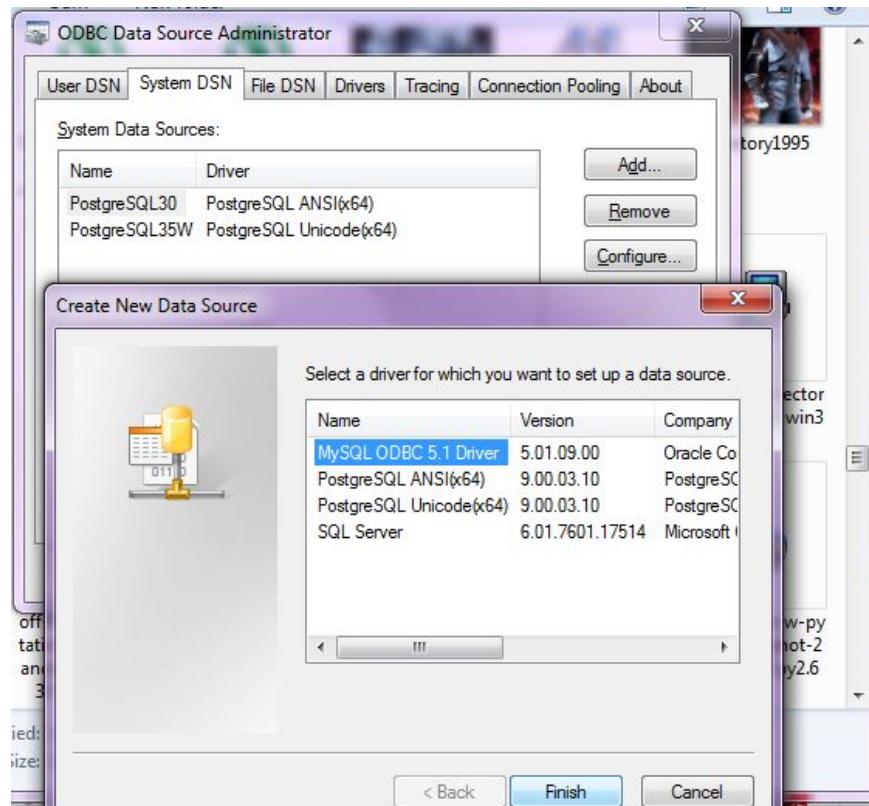
Creating DSN (in Windows)

1. Click **Start**, point to **Control Panel**, double-click **Administrative Tools**, and then double-click **Data Sources(ODBC)**.
2. Click the **System DSN** tab, and then click **Add**.
3. Click the database driver that corresponds with the database type to which you are connecting, and then click **Finish**.
4. Type the data source name. Make sure that you choose a name that you can remember. You will need to use this name later.
5. Click **Select**.
6. Click the correct database, and then click **OK**.
7. Click **OK**, and then click **OK**.

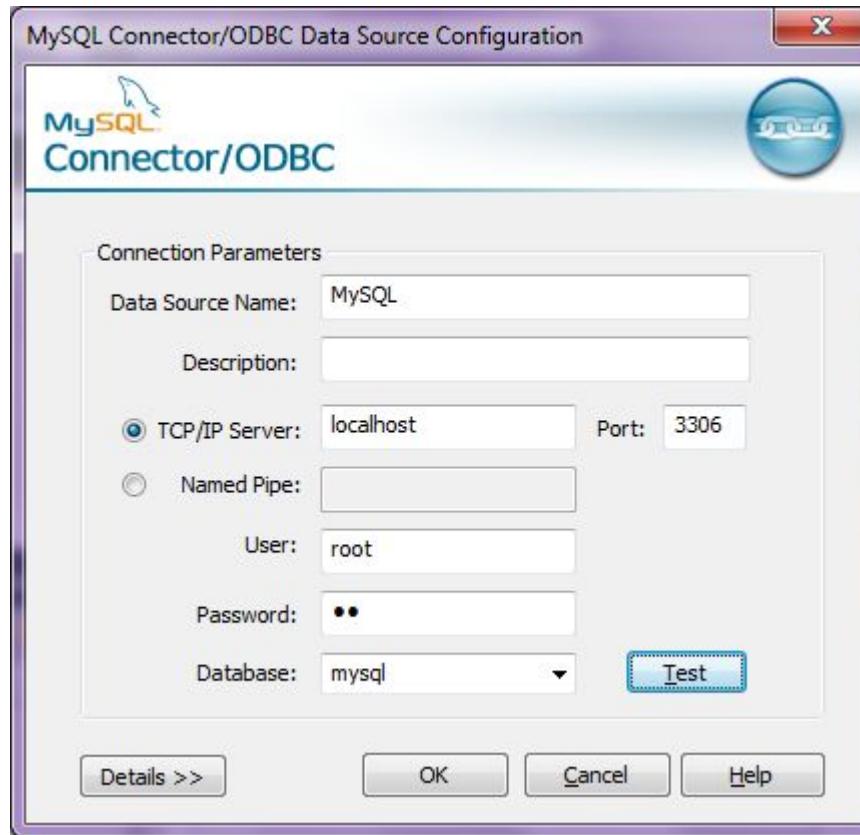
Creating DSN



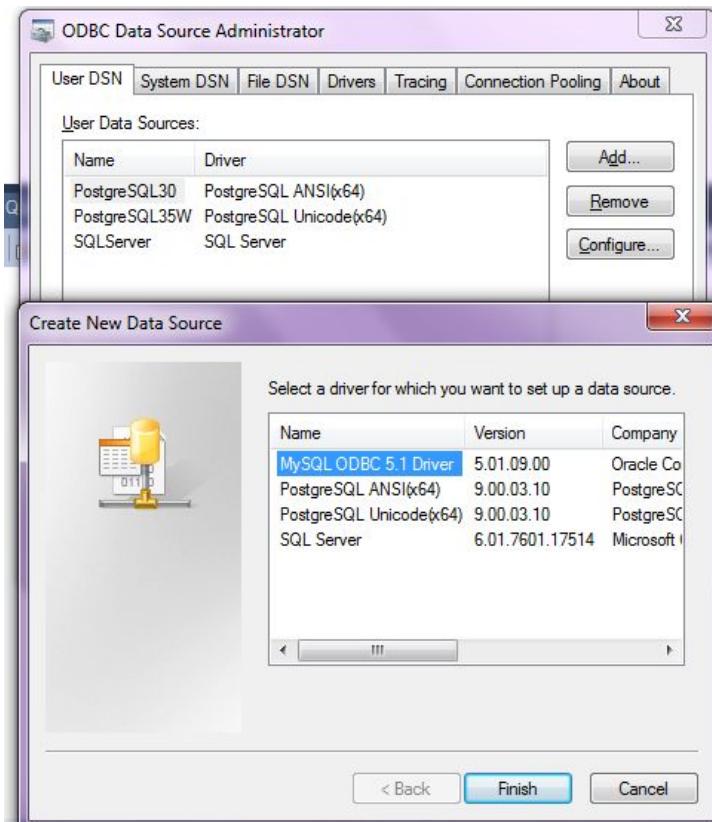
Creating DSN



Creating DSN



Creating DSN



RODBC

```
> library(RODBC)
> odbcDataSources()
> ajay=odbcConnect("MySQL",uid="root",pwd="XX")
> ajay
> sqlTables(ajay)
> tested=sqlFetch(ajay,"host")
```

From Databases

The **RODBC** package provides access to databases through an **ODBC** interface.

The primary functions are

- **`odbcConnect(ds, uid="", pwd="")`** Open a connection to an ODBC database
- **`sqlFetch(channel, sqltable)`** Read a table from an ODBC database into a data frame

Hint- a good site to revise R
<http://www.statmethods.net>

sqlite

<http://cran.r-project.org/web/packages/RSQLite/RSQLite.pdf> embeds the SQLite database engine in R and provides an interface compliant with the DBI package.

SQLite is a software library that implements a self-contained, serverless, zero-configuration, transactional SQL database engine.

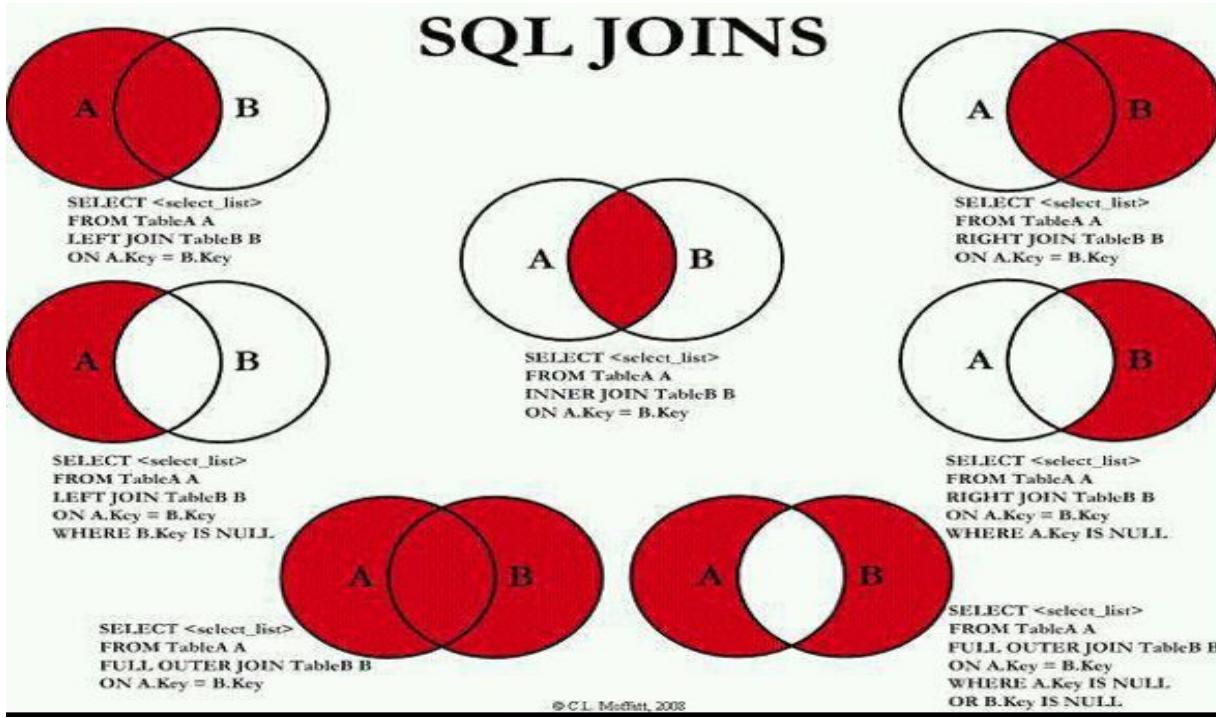
SQLite is the most widely deployed database engine in the world

```
library(RSQLite)
con <- dbConnect("SQLite", dbname = "sample_db")
# read csv file into sql database
dbWriteTable(con, name="sample_data", value="sample_data.csv", row.names=FALSE, header=TRUE, sep = ",")
```

<http://cran.r-project.org/web/packages/sqldf/index.html> Manipulate R data frames using SQL

`read.csv.sql` in the `sqldf` package imports data into a temporary SQLite database and then reads it into R.

A Detour to SQL Joins (Optional)



RMySQL

```
install.packages("RMySQL")
library(RMySQL)

mydb = dbConnect(MySQL(), user='user', password='password', dbname='database_name', host='host')

dbListTables(mydb)

dbListFields(mydb, 'some_table')

dbSendQuery(mydb, 'drop table if exists some_table, some_other_table')

dbWriteTable(mydb, name='table_name', value=data.frame.name)
```

Other databases

Teradata <https://github.com/Teradata/teradataR>

PostgreSQL <http://cran.r-project.org/web/packages/RPostgreSQL/>

MongoDB <http://cran.r-project.org/web/packages/mongolite/index.html>

couchDB <http://cran.r-project.org/web/packages/couchDB/index.html>

MonetDB <http://cran.r-project.org/web/packages/MonetDB.R/index.html>

Other data sources

Cassandra with R <http://cran.r-project.org/web/packages/RCassandra/RCassandra.pdf>

Neo4j with R

<http://things-about-r.tumblr.com/post/47392314578/venue-recommendation-a-simple-use-case-connecting-r>

R with Hadoop Stack <https://github.com/RevolutionAnalytics/RHadoop/wiki>

- NEW! `ravro` - read and write files in avro format
- `plyrnr` - higher level plyr-like data processing for structured data, powered by `rnr`
- `rnr` - functions providing Hadoop MapReduce functionality in R
- `rhdfs` - functions providing file management of the HDFS from within R
- `rbase` - functions providing database management for the HBase distributed database from within R

<https://amplab-extras.github.io/SparkR-pkg/> SparkR is an R package to use Spark from R.

Web Scraping

Web scraping (**web** harvesting or **web** data extraction) is a computer software technique of extracting information from websites.

example - python (scrapy and beautiful soup)

You didn't write that awful page. You're just trying to get some data out of it. Beautiful Soup is here to help. Since 2004, it's been saving programmers hours or days of work on quick-turnaround screen scraping projects.

Beautiful Soup
"A tremendous boon." – Python411 Podcast
[Download | Documentation | Hall of Fame | Source | Discussion group]

If Beautiful Soup has saved you a lot of time and money, the best way to pay me back is to check out *Constellation Games*, my sci-fi novel about alien video games. You can read the first two chapters for free, and the full novel starts at 5 USD. Thanks!

If you have questions, send them to the discussion group. If you find a bug, file it.

Beautiful Soup is a Python library designed for quick turnaround projects like screen-scraping. Three features make it powerful:

- Beautiful Soup provides a few simple methods and Pythonic idioms for navigating, searching, and modifying a parse tree: a toolkit for dissecting a document and extracting what you need. It doesn't take much code to write an application.
- Beautiful Soup automatically converts incoming documents to Unicode and outgoing documents to UTF-8. You don't have to think about encodings, unless the document doesn't specify one. Then you just have to specify the original encoding.
- Beautiful Soup sits on top of popular Python parsers like `lxml` and `html5lib`, so you can use it to try out different parsing strategies or trade speed for flexibility.

Beautiful Soup parses anything you give it, and does the tree traversal stuff for you. You can tell it "Find all the links", or "Find all the links of class `external_link`", or "Find all the links whose urls match "`foo.com`", or "Find the table heading that's got bold text, then give me that text."

Valuable data that was once locked up in poorly-designed websites is now within your reach. Projects that would have taken hours take only minutes with Beautiful Soup.

Interested? [Read more](#).

Download Beautiful Soup

The current release is **Beautiful Soup 4.3.2** (October 2, 2013). You can install it with `pip install beautifulsoup4` or `easy_install beautifulsoup4`. It's also available as the `python-beautifulsoup` package in recent versions of Debian, Ubuntu, and Fedora.

Beautiful Soup 4 works on both Python 2 (2.6+*) and Python 3.

Beautiful Soup is licensed under the MIT license, so you can also download the tarball, drop the `bs4` directory into almost any Python application (or into your library path) and start using it immediately. (If you want to do this under Python 3, you will need to manually convert the code using `2to3`.)

Meet Scrapy

An open source and collaborative framework for extracting the data you need from websites. In a fast, simple, yet extensible way.

Install latest version:
 **Scrapy 0.24**
\$ pip install scrapy

[PyPI](#) [Ubuntu Package](#) [Tarball](#) [Zip](#)

Sample Scrapy Code

```
$ pip install scrapy
$ cut > myspider.py <<EOF
from scrapy import Spider, Item, Field

class PostItem(Item):
    title = Field()

class BlogSpider(Spider):
    name = 'myspider'
    start_urls = ['http://blog.scrapinghub.com']

    def parse(self, response):
        return [PostItem(text=response.css("h2 a |text"))]
EOF
$ scrapy runspider myspider.py
```

 Build your own webcrawlers

 Scrapy

Web Scraping

- **readlines**

```
> url="http://nytimes.com"
> ajay=readlines(url)
Error: could not find function "readlines"
> ajay=readLines(url)
> head(ajay)
[1] "<!DOCTYPE html>"                                                 $ 
[2] "<!--[if gt IE 9) ||!(IE)]> <!--> <html lang=\"en\" class=\"no-js edition-domestic app-homepage\" itemscope xmlns:og$ 
[3] "<!--[if IE 9]> <html lang=\"en\" class=\"no-js ie9 lt-ie10 edition-domestic app-homepage\" xmlns:og=\"http://opengraph.$ 
[4] "<!--[if IE 8]> <html lang=\"en\" class=\"no-js ie8 lt-ie10 lt-ie9 edition-domestic app-homepage\" xmlns:og=\"http://$ 
[5] "<!--[if (lt IE 8)]> <html lang=\"en\" class=\"no-js lt-ie10 lt-ie9 lt-ie8 edition-domestic app-homepage\" xmlns:og=$ 
[6] "<head>"                                                       $ 
> tail(ajay)
[1] "<div id=\"ab3\" class=\"ad ab3-ad hidden\"></div>"           $ 
[2] "<div id=\"prop1\" class=\"ad prop1-ad hidden\"></div>"          $ 
[3] "<div id=\"prop2\" class=\"ad prop2-ad hidden\"></div>"          $ 
[4] "<div id=\"Anchor\" class=\"ad anchor-ad hidden\"></div>"         $ 
[5] "<script type=\"text/javascript\">window.NREUM|| (NREUM={});NREUM.info={"beacon": \"beacon-6.newrelic.com\", \"licens$ 
[6] "</html>"                                                       $ 
> |
```

Hint : R is case sensitive
readlines is not the same as readLines

Hint : Use head() and tail() to inspect objects

Other packages are XML and Curl

Case Study- <http://decisionstats.com/2013/04/14/using-r-for-cricket-analysis-rstats/>

curl

cURL is a computer software project providing a library and command-line tool for transferring data using various protocols. The **cURL** project produces two products, **libcurl** and **cURL**.

The RCurl package is an R-interface to the [libcurl](#) library that provides HTTP facilities. This allows us to download files from Web servers, post forms, use HTTPS (the secure HTTP), use persistent connections, upload files, use binary content, handle redirects, password authentication, etc.

The primary top-level entry points are

- [getURL\(\)](#)
- [getURLContent\(\)](#)
- [getForm\(\)](#)
- [postForm\(\)](#)

<http://www.omegahat.org/RCurl/RCurlJSS.pdf>



Untitled1*

Source on Save

```
1 library(RCurl)
2 h = getCurlHandle()
3 getURI("http://www.omegahat.org/RCurl/index.html", curl = h)
4 names(getCurlInfo(h))
```

1:1 (Top Level) ▾

Console ~/ ↻

```
oop>">REventLoop</a>).\nWe can potentially turn them into regular
).\\n\\n\\n<h2>License</h2>\nThis is distributed under the <a href='
e</a>\\nin the same spirit as libcurl itself.\\n\\n<hr>\\n<address><a
Temple Lang</a>\\n<a href='mailto:duncan@wald.ucdavis.edu'>dunc
t -->\\nLast modified: Mon May 25 11:35:38 PDT 2009\\n!-- hhmts en
> names(getCurlInfo(h))
[1] "effective.url"           "response.code"          "total.t
[5] "connect.time"            "pretransfer.time"        "size.up
[9] "speed.download"          "speed.upload"           "header.
[13] "ssl.verifyresult"         "filetime"                "content
[17] "starttransfer.time"      "content.type"             "redirec
[21] "private"                  "http.connectcode"        "httpauth
[25] "os errno"                 "num.connects"            "ssl.en
[29] "lastsocket"               "ftp.entry.path"          "redirec
[33] "appconnect.time"          "certinfo"                "conditi
```

XML

The screenshot shows an RStudio interface with the following components:

- Header Bar:** File, Edit, Code, View, Plots, Session, Build, Debug, Tools, Help.
- Toolbar:** Includes icons for file operations like Open, Save, and Run, along with a "Go to file/function" search bar.
- Code Editor:** Untitled1* tab. The code is written in R, using the XML library to read a cricket statistics table from a URL.

```
library(XML)
theurl="http://stats.espncricinfo.com/ci/engine/stats/index.html?class=1;team=6;template=results;type=batting"
#Note I can also break the url string and use paste command to modify this url with parameters
table2 <- readHTMLTable(theurl)
table_cricket=table2$"Overall figures"
head(table_cricket)
```
- Run Buttons:** Run, Stop, and Refresh buttons located in the top right corner of the code editor.
- Console:** Shows the R commands run and their output. The output table is displayed below.
- Output Table:** A data frame showing the top 6 rows of cricket player statistics. The columns are Player, Span, Mat, Inns, NO, Runs, HS, Ave, 100, 50, and 0.

	Player	Span	Mat	Inns	NO	Runs	HS	Ave	100	50	0
1	SR Tendulkar	1989-2013	200	329	33	15921	248*	53.78	51	68	14
2	R Dravid	1996-2012	163	284	32	13265	270	52.63	36	63	7
3	SM Gavaskar	1971-1987	125	214	16	10122	236*	51.12	34	45	12
4	VVS Laxman	1996-2012	134	225	34	8781	281	45.97	17	56	14
5	V Sehwag	2001-2013	103	178	6	8503	319	49.43	23	31	16
6	SC Ganguly	1996-2008	113	188	17	7212	239	42.17	16	35	13

json format

jsonlite for json data

<http://arxiv.org/abs/1403.2805>

```
> library(jsonlite)
Attaching package: 'jsonlite'

The following object is masked from 'package:utils':

```

View



```
> library(httr)
> library(curl)
> zips <- stream_in(curl("https://media.mongodb.org/zips.json"))
opening curl input connection.
Found 29353 lines...
binding pages together (no custom handler).
closing curl input connection.
>
> head(zips)
  _id      city          loc  pop state
1 01001    AGAWAM -72.62274, 42.07021 15338   MA
2 01002    CUSHMAN -72.51564, 42.37702 36963   MA
3 01005     BARRE -72.10835, 42.40970  4546   MA
4 01007 BELCHERTOWN -72.41095, 42.27510 10579   MA
5 01008   BLANDFORD -72.93611, 42.18295  1240   MA
6 01010 BRIMFIELD -72.18846, 42.11654  3706   MA

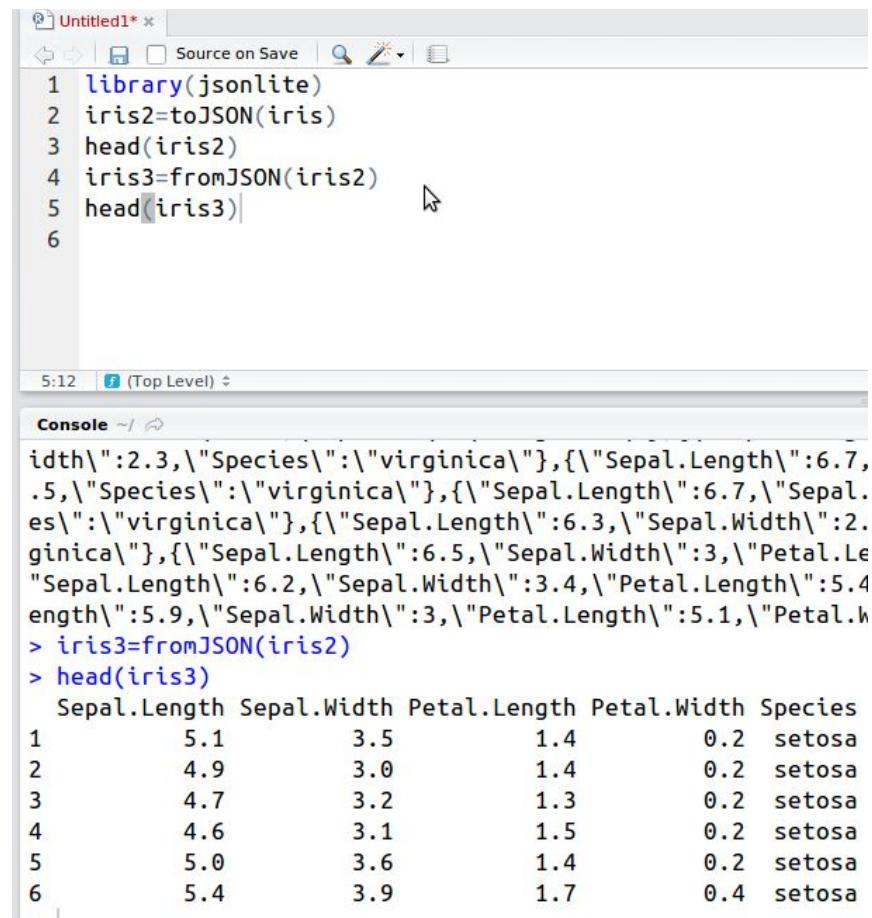
```

```
{ "_id" : "01001", "city" : "AGAWAM", "loc" : [ -72.622739, 42.070206 ], "pop" : 15338, "state" : "MA" }
{ "_id" : "01002", "city" : "CUSHMAN", "loc" : [ -72.51564999999999, 42.377017 ], "pop" : 36963, "state" : "MA" }
{ "_id" : "01005", "city" : "BARRE", "loc" : [ -72.10835400000001, 42.409698 ], "pop" : 4546, "state" : "MA" }
{ "_id" : "01007", "city" : "BELCHERTOWN", "loc" : [ -72.41095300000001, 42.275103 ], "pop" : 10579, "state" : "MA" }
{ "_id" : "01008", "city" : "BLANDFORD", "loc" : [ -72.936114, 42.182949 ], "pop" : 1240, "state" : "MA" }
{ "_id" : "01010", "city" : "BRIMFIELD", "loc" : [ -72.188455, 42.116543 ], "pop" : 3706, "state" : "MA" }
{ "_id" : "01011", "city" : "CHESTER", "loc" : [ -72.988761, 42.279421 ], "pop" : 1688, "state" : "MA" }
{ "_id" : "01012", "city" : "CHESTERFIELD", "loc" : [ -72.833309, 42.38167 ], "pop" : 177, "state" : "MA" }
{ "_id" : "01013", "city" : "CHICOPEE", "loc" : [ -72.607962, 42.162046 ], "pop" : 23396, "state" : "MA" }
{ "_id" : "01020", "city" : "CHICOPEE", "loc" : [ -72.576142, 42.176443 ], "pop" : 31495, "state" : "MA" }
{ "_id" : "01022", "city" : "WESTOVER AFB", "loc" : [ -72.558657, 42.196672 ], "pop" : 1764, "state" : "MA" }
{ "_id" : "01026", "city" : "CUMMINGTON", "loc" : [ -72.905767, 42.435296 ], "pop" : 1484, "state" : "MA" }
{ "_id" : "01027", "city" : "MOUNT TOM", "loc" : [ -72.67992099999999, 42.264319 ], "pop" : 16864, "state" : "MA" }
{ "_id" : "01028", "city" : "EAST LONGMEADOW", "loc" : [ -72.505565, 42.067203 ], "pop" : 13367, "state" : "MA" }
{ "_id" : "01030", "city" : "FEEDING HILLS", "loc" : [ -72.675077, 42.07182 ], "pop" : 11985, "state" : "MA" }
{ "_id" : "01031", "city" : "GILBERTVILLE", "loc" : [ -72.19858499999999, 42.332194 ], "pop" : 2385, "state" : "MA" }
{ "_id" : "01032", "city" : "GOSHEN", "loc" : [ -72.844092, 42.466234 ], "pop" : 122, "state" : "MA" }
```

json format

jsonlite for json data

<http://arxiv.org/abs/1403.2805>



The screenshot shows an RStudio interface. The top panel is an 'Untitled1' script editor with the following code:

```
1 library(jsonlite)
2 iris2<-toJSON(iris)
3 head(iris2)
4 iris3<-fromJSON(iris2)
5 head(iris3)
6
```

The bottom panel is a 'Console' window with the following output:

```
5:12 [f] (Top Level) ▾
Console ~/ ↵
idth\":2.3,\"Species\":\"virginica\"}, {"Sepal.Length\":6.7,
.5,\"Species\":\"virginica\"}, {"Sepal.Length\":6.7,\"Sepal.
es\":[\"virginica\"}, {"Sepal.Length\":6.3,\"Sepal.Width\":2.
ginica\"}, {"Sepal.Length\":6.5,\"Sepal.Width\":3,\"Petal.Le
\"Sepal.Length\":6.2,\"Sepal.Width\":3.4,\"Petal.Length\":5.4
ength\":5.9,\"Sepal.Width\":3,\"Petal.Length\":5.1,\"Petal.W
> iris3<-fromJSON(iris2)
> head(iris3)
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa

Using APIs for data

<https://ropensci.org/>

CRAN Task View: Web Technologies and Services

Maintainer: Scott Chamberlain, Thomas Leeper, Patrick Mair, Karthik Ram, Christopher Gandrud

Contact: scott at ropensci.org

Version: 2015-03-20

This task view contains information about using R to obtain and parse data from the web. The base version of R does not ship with many tools for interacting with the web. Thankfully, there are an increasingly large number of tools for interacting with the web. A list of available packages and functions is presented below, grouped by the type of activity. If you have any comments or suggestions for additions or improvements for this taskview, go to GitHub and [submit an issue](#), or make some changes and [submit a pull request](#). If you can't contribute on GitHub, [send Scott an email](#). If you have an issue with one of the packages discussed below, please contact the maintainer of that package. If you know of a web service, API, data source, or other online resource that is not yet supported by an R package, consider adding it to [the package development to do list on GitHub](#).

Tools for Working with the Web from R

Parsing Data from the Web

- **downloading files** : `download.file()` is in base R and commonly used way to download a file. However, downloading files over HTTPS is not supported in R's internal method for `download.file()`. The `downloader` function in the package `downloader` wraps `download.file()`, and takes all the same arguments, but works for https across platforms.
- **tabular data as txt, csv, etc.** : You can use `read.table()`, `read.csv()`, and friends to read a table directly from a URL, or after acquiring the csv file from the web via e.g., `getURL()` from RCurl. `read.csv()` works with http but not https, i.e.: `read.csv("http://...")`, but not `read.csv("https://...")`. You can download a file first before reading the file in R, and you can use `downloader` to download over https. `read.table()` and friends also have a `text` parameter so you can read a table if a table is encoded as a string with line breaks, etc.
- **JSON I/O** : JSON is *javascript object notation* . There are three packages for reading and writing JSON: `rjson`, `RJSONIO`, and `jsonlite`. `jsonlite` includes a different parser from `RJSONIO` called `yajl`. We recommend using `jsonlite`. Check out the paper describing jsonlite by Jeroen Ooms <http://arxiv.org/abs/1403.2805>.
- **XML/HTML I/O** : The package `XML` contains functions for parsing XML and HTML, and supports xpath for searching XML (think regex for strings). A helpful function to read data from one or more HTML tables is `readHTMLTable()`. `XML` also includes `XPATH` parsing ability, see `xpathApply()` and `xpathSApply()`. The `XML2R` package is a collection of convenient functions for coercing XML into data frames (development version [on GitHub](#)). An alternative to `XML` is `selectr`, which parses CSS Selectors and translates them to XPath 1.0 expressions. `XML` package is often used for parsing xml and html, but `selectr` translates CSS selectors to XPath, so can use the CSS selectors instead of XPath. The `selectorgadget browser extension` can be used to identify page elements. `RHTMLForms` reads HTML documents and obtains a description of each of the forms it contains, along with the different elements and hidden fields. `scraper` provides additional tools for scraping data from HTML and XML documents.
- **rvest** : rvest scrapes html from web pages, and is designed to work with `magrittr` to make it easy to express common web scraping tasks.
- The `tidyextract` package extract top level domains and subdomains from a host name. It's a port of [a Python library of the same name](#).
- **webutils**: Utility functions for developing web applications. Parsers for application/x-www-form-urlencoded as well as multipart/form-data. [Source on Github](#)
- **urllibs**: URL encoding, decoding, parsing, and parameter extraction. [Source on Github](#)
- The `repmis` package contains a `source_data()` command to load and cache plain-text data from a URL (either http or https). It also includes `source_Dropbox()` for downloading/caching plain-text data from non-public Dropbox folders and `source_XlsxData()` for downloading/caching Excel xlsx sheets.
- **rsdmx** provides tools to read data and metadata documents exchanged through the Statistical Data and Metadata Exchange (SDMX) framework. The package currently focuses on the SDMX XML standard format (SDMX-ML). [project website \(Github\)](#).

Curl, HTTP, FTP, HTML, XML, SOAP

- **RCurl**: A low level curl wrapper that allows one to compose general HTTP requests and provides convenient functions to fetch URIs, get/post forms, etc. and process the results returned by the Web server. This provides a great deal of control over the HTTP/FTP connection and the form of the request while providing a higher-level interface than is available just using R socket connections. It also provide tools for Web authentication.

ff package

<http://cran.r-project.org/web/packages/ff/index.html>

The ff package provides data structures that are stored on disk but behave (almost) as if they were in RAM by transparently mapping only a section (pagesize) in main memory - the effective virtual memory consumption per ff object.

<http://cran.r-project.org/web/packages/ffbase/index.html>

Basic (statistical) functionality for package ff

Example- <http://www.bnosa.be/index.php/blog/22-if-you-are-into-large-data-and-work-a-lot-package-ff>

```
> require(ffbase)
> hhp <- 
read.table.ffdf(file="/home/jan/Work/RForgeBNOSAC/github/RBelgium_HeritageHealthPrize/Data/Claims.csv", FUN =
"read.csv", na.strings = "")
```

Also see <http://cran.r-project.org/web/packages/bigmemory/index.html>

Create, store, access, and manipulate massive matrices. Matrices are allocated to shared memory and may use memory-mapped files. Packages biganalytics, bigtabulate, synchronicity, and bigalgebra provide advanced functionality

RevoScaleR package

RevoScaleR has its own file format, XDF, which is able to rapidly access data by row or by column and to read some data sequentially. XDF file data is stored in the same binary format used in memory, which eliminates the need for conversion when it is brought into memory.

<http://www.revolutionanalytics.com/revolution-r-enterprise-scaler>

r hdf5

This R/Bioconductor package provides an interface between HDF5 and R. HDF5's main features are the ability to store and access very large and/or complex datasets and a wide variety of metadata on mass storage (disk) through a completely portable file format.

<http://www.bioconductor.org/packages/release/bioc/html/rhdf5.html>

HDF5 is a data model, library, and file format for storing and managing data. It supports an unlimited variety of datatypes, and is designed for flexible and efficient I/O and for high volume and complex data. HDF5 is portable and is extensible, allowing applications to evolve in their use of HDF5.

<https://www.hdfgroup.org/HDF5/>

HDF5 simplifies the file structure to include only two major types of object:

- Datasets, which are multidimensional arrays of a homogeneous type
- Groups, which are container structures which can hold datasets and other groups



HDF5 interface to R

Bioconductor version: Release (3.1)

This R/Bioconductor package provides an interface between HDF5 and R. HDF5's main features are the ability to store and access very large and/or complex datasets and a wide variety of metadata on mass storage (disk) through a completely portable file format. The rhdf5 package is thus suited for the exchange of large and/or complex datasets between R and other software package, and for letting R applications work on datasets that are larger than the available RAM.

Author: Bernd Fischer, Gregoire Pau

Maintainer: Bernd Fischer <b.fischer at dkfz.de>

Citation (from within R, enter `citation("rhdf5")`):

Fischer B and Pau G. *rhdf5: HDF5 interface to R*. R package version 2.12.0.

Installation

To install this package, start R and enter:

```
source("http://bioconductor.org/biocLite.R")
biocLite("rhdf5")
```

Documentation

To view documentation for the version of this package installed in your system, start R and enter:

```
browseVignettes("rhdf5")
```

Documentation »

Bioconductor

- Package [vignettes](#) and manuals.
- [Workflows](#) for learning and use.
- [Course and conference](#) material.
- [Videos](#).
- Community [resources](#) and [tutorials](#).

R / [CRAN](#) packages and [documentation](#)

Support »

Please read the [posting guide](#). Post questions about Bioconductor to one of the following locations:

- [Support site](#) - for questions about Bioconductor packages
- [Bioc-devel](#) mailing list - for package developers

Functions Used in this lesson

- `toJSON` and `fromJSON`

Packages

- `data.table`
- `jsonlite`
- `rvest`

Revision

getwd

setwd

dir

ls

rm

Install.packages

library

fread vs read.csv

Df[i,j]

Df\$column

str

summary

table

citation

help

Revision

mean

std

median

length

Vector

data.frame

Indexing

class

nrow

ncol

head

tail

Citations and References

M Dowle, T Short, S Lianoglou, A Srinivasan with contributions from R Saporta and E Antonyan (2014) `data.table`: Extension of `data.frame`. R package version 1.9.4.

<http://CRAN.R-project.org/package=data.table>

Jeroen Ooms (2014). The `jsonlite` Package: A Practical and Consistent Mapping Between JSON Data and R Objects. arXiv:1403.2805 [stat.CO] URL <http://arxiv.org/abs/1403.2805>

Hadley Wickham (2015). `rvest`: Easily Harvest (Scrape) Web Pages. R package version 0.2.0.

<http://CRAN.R-project.org/package=rvest>