iCMS 2015
INTERNATIONAL CONFERENCE ON COMPUTING, MATHEMATICS AND STATISTICS
Bridging Research Endeavour in Computer and Mathematical Sciences
For more information, please visit http://www.icms2015.org
Organized by : FACULTY of COMPUTER & MATHEMATICAL SCIENCES UiTM KEDAH
Jointly organized by : RESEARCH & INDUSTRIAL LINKAGES
4th - 5th November 2015
Langkawi Island, MALAYSIA
UNIVERSITI TEKNOLOGI MARA CAWANGAN KEDAH

PRE-CONFERENCE WORKSHOP "INTRODUCTION TO R AND DATA VISUALIZATION"

Ciprian Alexandru

R-omania Team | www.r-project.ro

Presentation

- The R platform provides a powerful and comprehensive platform for visualizing data, understanding and evaluating statistical models, and effectively communicating research results to both technical and nontechnical audiences. This 2 days workshop will provide practical review of R's major graphing capabilities; including base functionality and new capabilities provided by the lattice and ggplot2 packages.

Date & Location

- Date: 2 - 3 November 2015
- Time: 9 am - 5 pm
- Venue: Melur 1, Langkawi Lagoon, Langkawi Island, MALAYSIA

Who should attend?

- R is widely used within the academia especially in the fields of computational biology, applied science, quantitative finance and business intelligence. R is capable of solving challenging problems and among the strengths of R are its powerful built-in tools for inferential statistics, its compact modeling syntax, and its data visualization capabilities. In addition, R's open source nature and its extensibility via add-on "packages" has allowed it to keep up with the leading edge in academic research. This workshop on R and Data Visualization is suitable and relevant for:
- Lecturers, Researchers, Engineers, Students, Industry Professionals and Scientists of any discipline who wish to explore R. Prior experience with R is not required. Interested to join??? Please register here.

Speakers

Antoniade-Ciprian Alexandru is an Associate Professor at the Ecological University of Bucharest and the dean of the Faculty of Economics. He is also attached with the National Institute of Statistics, Bucharest as an expert trainer in data analyst using R environment. Dr Alexandru is one of the six members of the R-omania team, a team that promotes R projects for statistical computing by providing a free and open source software environment for data analysis and graphics. The team acts as a user community for development of R projects among individuals, institutions, commercial entities and non-profit organizations. Dr Alexandru participated in various research projects, workshops, and, national and international conferences. His research works were published in various international databases. Currently, he is working on a project that implements the use of R as a tool for analyzing the evolution of indices on the stock market.

Nicoleta Caragea is an Associate Professor at the Faculty of Economics, Ecological University of Bucharest and a senior expert at the National Institute of Statistics. Her teaching activity is focused mainly in the field of statistics, through courses and seminars and master degree programs (statistics, economic statistics, social statistics, economic and financial analysis). Dr Caragea participated as a national expert in various projects, workshops and conferences organized by EUROSTAT, OECD, WHO, World Bank and UNICEF-UIS. She is one of the other six members of the R-omania team, a team that promotes R projects for statistical computing by providing a free and open source software environment for data analysis and graphics. She also acted as a consultant in projects in Europe. Her latest work was as a technical assistance to a consultancy work in Turkey.

R-omania Team

- Our Team acts as user community for development of R project among the Romanian individual persons, institutions and commercial and noncommercial organizations.
- The Romanian Team promote the R project for statistical computing to provide a free and open source software environment for data analysis and graphics in Romania.
- Support the further development of R and related open source software projects in Romania.
- Initiate, promote and coordinate research projects, support communication within the R user community, and organize or sponsor courses.
- Organize R-related scientific conferences and workshops, participate at relevant R conferences sponsored by others, and promote the use and development of R and R-related software in Romania.
- Publish manuals, journal articles and other R-related documents in printed and electronic form.
- Promote the using of R environment in universities and offer support for curricula development in the field of statistical software.

source: https://www.r-project.ro/

R-omania Team – Conferences & Workshops

- 2016 - Conference - New Challenges for Statistical Software - The Use of R in Official Statistics
- 2016 – Workshop - Applied R to Social Sciences
- 2015 - Conference - New Challenges for Statistical Software - The Use of R in Official Statistics
- 2014 - Workshop - New Challenges for Statistical Software - The Use of R in Official Statistics
- 2013 - Workshop - State-of-the-art statistical software commonly used in applied economics

source: https://www.r-project.ro/

R-omania Team - Courses

- 3 to 5 days Courses:
  - "Statistics with Applications in R"
  - "Data Analysis in Statistics with R"
  - "Introduction to Statistics - Applications in R"
  - "Introduction in Small Area Estimation Techniques with Applications in R"
- One-day courses:
  - "R Statistical Software – Presenting Advantages of its use for Data Analysis"
  - "Introducing Statistics, the Need for Official Statistics"
  - "Statistical Analysis – from Theory to Practice"
  - "Concepts, Models and Techniques for Data Analysis"

source: https://www.r-project.ro/

R-omania Team - Courses

- Over 400 trained people from:
  - National Institute of Statistics – over 200
  - National Bank of Romania
  - Vodafone
  - Orange
  - UniCredit Bank
  - Ministry of Finance
  - Romanian Academy
  - Universities

R-omania Team - Research

- "R cu aplicatii in statistica", [en. "R with applications in statistics"], Published in 2015
- More than 25 research papers
- "R pentru Incepatori", the Romanian version of "R for Beginners" by Emmanuel Paradis, translated by Ana-Maria Dobre, Aug, 2013

source: https://www.r-project.ro/

Course Outline 1/2

- Introduction to R Statistical Software
  - The beginning of R
  - R - Introducing the R Console
  - R - Installation, Packages, CRAN, Components
  - Graphical User Interfaces: R Console, R Studio, R Commander, R resources and online community
- Databases
  - Data manipulation
  - Queries
  - Using SQL within R
  - Data aggregation
  - Matching

Course Outline 2/2

- Data Visualization & Graphics Environments
  - Base graphics (Scatterplot, Box-and-whiskers plot, Histogram)
  - Lattice
  - ggplot2
  - Interactive graphics in R
  - Reproducibility
- Regression Analysis with R
  - Linear regression models
  - OLS-ordinary least squares method for estimating the unknown parameters in a linear regression model
  - Interpreting the regression coefficients
  - Extensions to generalized linear models. Logistic regression
  - Parameter estimates – maximum likelihood method
  - Definition of the odds and odds ratio
  - Evaluating goodness of fit

Introduction to R Statistical Software

- The beginning of R
- R - Introducing the R Console
- R - Installation, Packages, CRAN, Components
- Graphical User Interfaces: R Console, R Studio, R Commander, R resources and online community

What is R?

- R is a programming language and software environment for statistical computing and graphics
- The key point is the environment

source: https://www.r-project.org/about.html

From where R comes?

- 1997 - *R*oss Ihaka and *R*obert Gentleman, professors of statistical at the Auckland University from New Zeeland, starts to build a new software for statistical analysis and data graphical visualizations
- R is a dialect of S language (S was built by AT&T Bell Laboratories as a software for data analysis, statistical modeling, simulation and graphics)

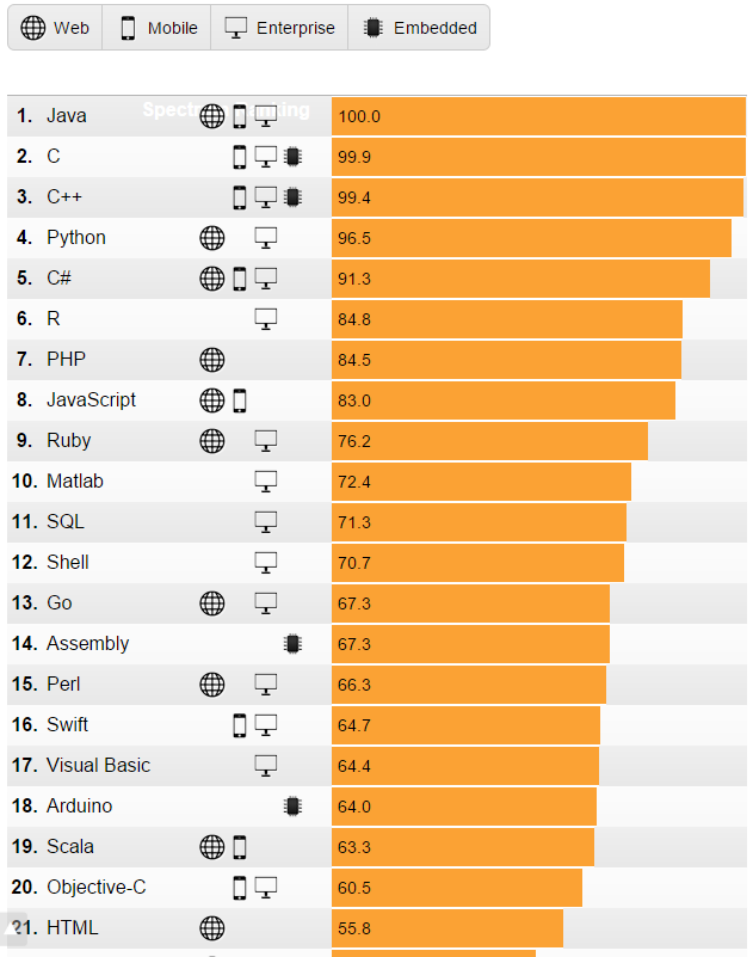
source: http://www.burns-stat.com/documents/tutorials/why-use-the-r-language/

Why R?

- R is supported by academia
- R is an open source initiative, similar with the Linux operating system or LaTeX markup language
- R is not just a statistics package, it's a statistical programming language
- R is designed to overcome the data scientist problems
- R is both flexible, powerful and endless

R – a top language

IEEE Spectrum ranking



source: http://spectrum.ieee.org/static/interactive-the-top-programming-languages-2015#index/

R's features

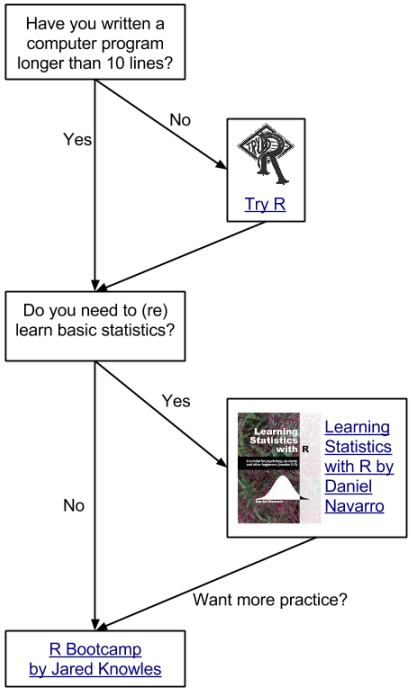
- mirror of thinking – (ie. weight ~ height + girth)
- abstraction – simple to teach R the abstraction
- interactive language – data analysis is an interactive process by excellence
- object-oriented language – all commands are loaded and executed into memory without temporary files, the results are stored into objects with possibility to reuse them
- vector-oriented => vectorization: objects are generally treated as a whole (ie. height.cm <- 2.54 * height.inches)
- data structures – creating freely any kind of data structures, sometime without well structured (but tricky some times), interesting data types: list, factor, data frame
- mixing tools – call C or C++ functionality from R
- graphics – a central point of analysis, because life is more visual than other senses
- missing values – this is the real data, sometime people don't answers to all questions in a survey)
- functions as first class objects
- packages – personalized solution for specific problem;
- community – in your analysis you are not alone, the community is very strong and very professional

sources:
http://www.burns-stat.com/documents/tutorials/why-use-the-r-language/
https://www.packtpub.com/big-data-and-business-intelligence/r-object-oriented-programming/

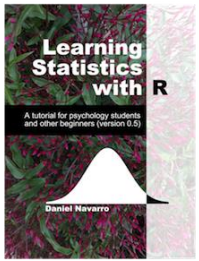
Who are the R users?

- statisticians
- data miners
- academia
- financial analyst
- data analyst
- mathematicians
- engineers
- students
- industry professionals
- any other researchers

Starting learning R?



- http://tryr.codeschool.com/



- http://health.adelaide.edu.au/psychology/ccs/teaching/lsr/
- http://health.adelaide.edu.au/psychology/ccs/docs/lsr/lsr-0.5.pdf

- R bootcamp: http://jaredknowles.com/r-bootcamp/

source: http://www.r-bloggers.com/how-to-learn-r-a-flow-chart/

R-ide the Wave

- R-ide the Wave

Photo source: http://bigthink.com/disrupt-education/how-stanford-wants-to-ride-the-wave-of-online-learning

What you really want to do?

Data scientist paid $500k can barely code!

Data sci

datascienc

mine - who

- https://www.linkedin.com/grp/post/77616-6043795029238050816

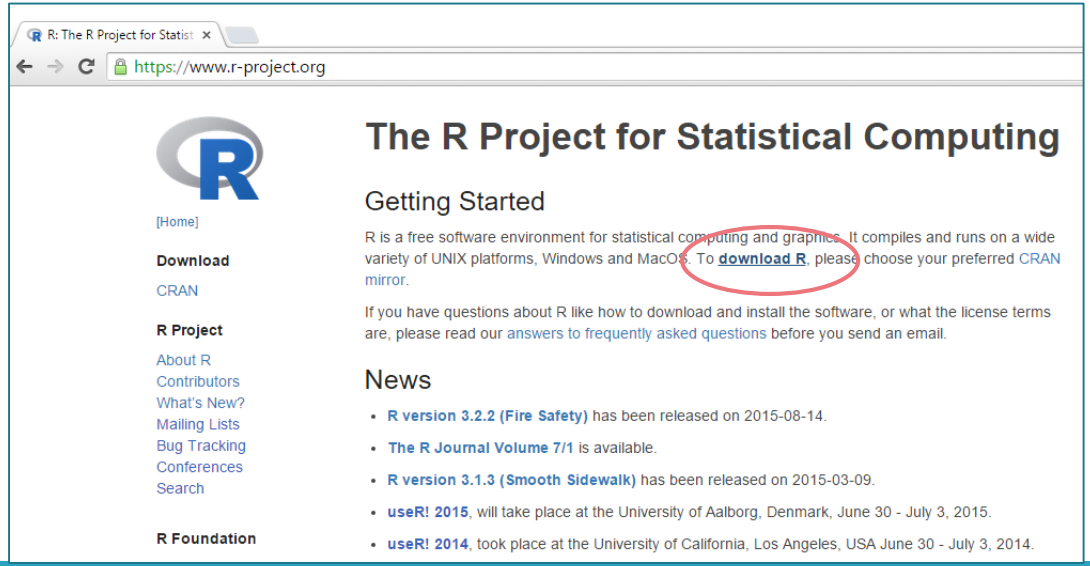
Define your goal!

Introduction to R Statistical Software

- The beginning of R
- R - Introducing the R Console
- R - Installation, Packages, CRAN, Components
- Graphical User Interfaces: R Console, R Studio, R Commander, R resources and online community

R Quick installation

- Install R for UNIX platforms, Windows and MacOS from https://www.r-project.org/
- The Windows users just clicks, other users know better than others.



R Console

- R Console is the GUI included in R core
- > is the command prompt followed by a flashing cursor, meaning that R is waiting your reaction
- commands => functions

Quit function

> q()

Help function

> help()

Or:

> ?

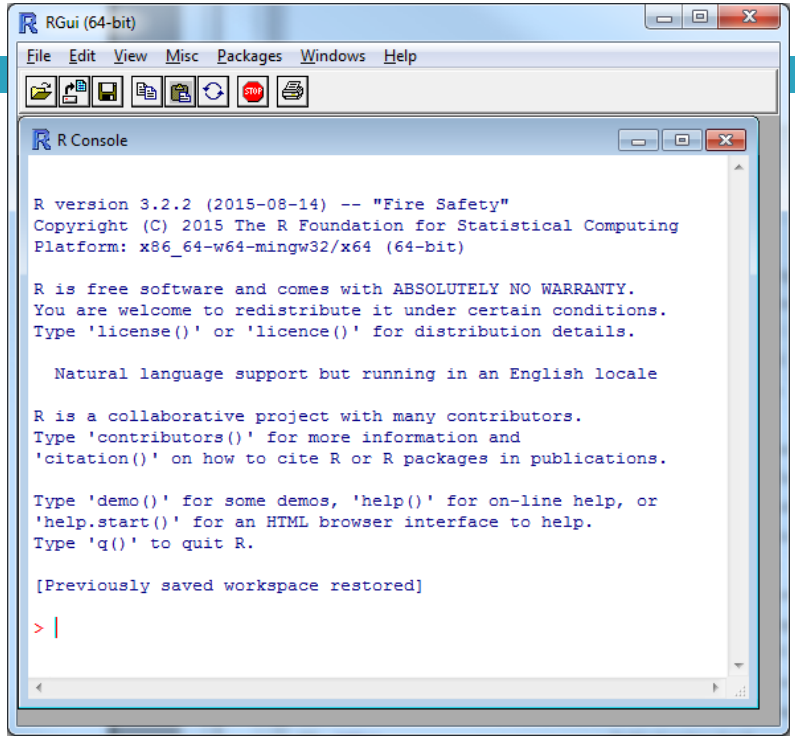
> log(10)

[1] 2.302585

> ?log

Or:

> help(log)



Working directory

- In working directory "lives" all R files for writing or reading, also your files, if you don't specify the path.
- "/" or double "\\" instead of "\"

```
> getwd()
[1] "C:/Users/PCG/Documents"
> setwd("c:/")
> getwd()
[1] "c:/"
> setwd("c:\\Users")
> getwd()
[1] "c:/Users"
```

Always useful…

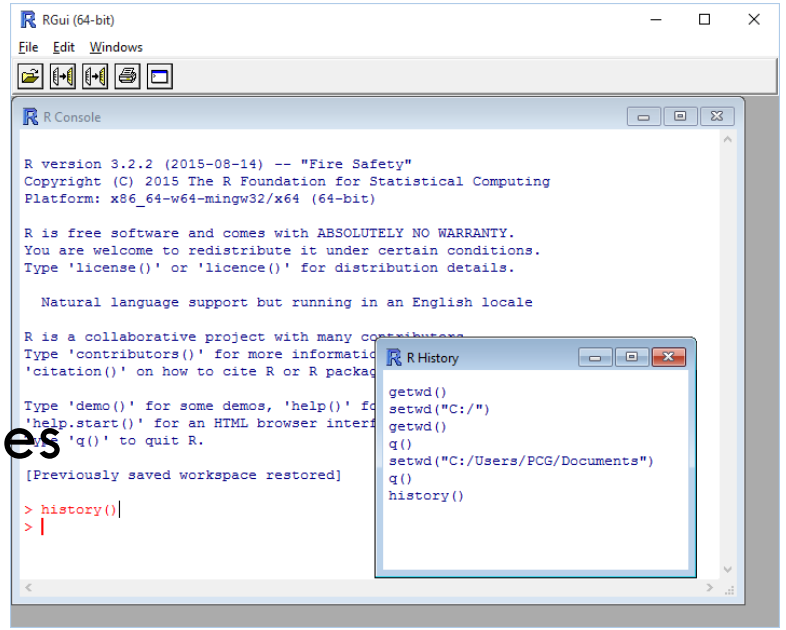
- R is case sensitive: variable *a* is different from *A*

```
> a <- 2
> a
[1] 2
> A
Error: object 'A' not found
> A <- 3
> A
[1] 3
```

- Keywords: if, else, repeat, while, function, for, in, next, break, TRUE, FALSE, NULL, Inf, NaN, NA, NA_integer_, NA_real_, NA_complex_, and finally, NA_character_
- navigation commands executed: arrow Up and Down
- Ctrl+L clear console

Sometimes useful…

- history()
- savehistory() #default .RHistory
- loadhistory()
- dir() or list.files()
- list.dirs() #including subdirectories
- file.exists()
- file.remove()



Environment…view, load & save

- ls()
- rm()
- environment()
- save.image() or save(list = ls(), file = ".Rdata")
- load(".Rdata")

- save(variable, file = "myvariable.rda")
- load(file = "myvariable.rda")

```
> a <- sqrt(2)
> a
[1] 1.414214
> ls()
[1] "a"
> b <- a^12
> ls()
[1] "a" "b"
> rm(list = ls())
> ls()
character(0)
```

Environment… customize

- options()

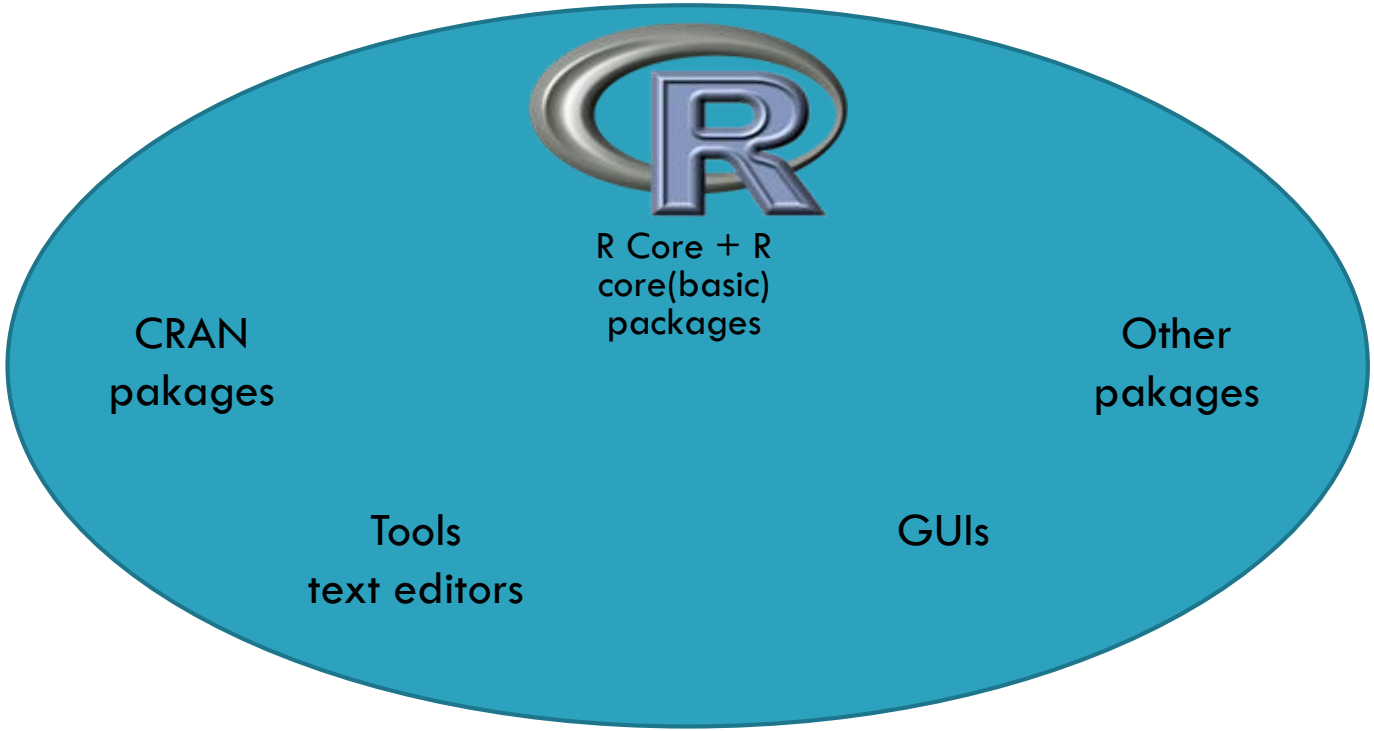
- digits = 7 (including the decimal separator)
- OutDec = "."
- prompt = "> "
- papersize = "a4"
- editor = "internal"
- continue="+ "
- width = 80

```
> (a <- sqrt(2))
[1] 1.414214
> options(digits = 3)
> (a <- sqrt(2))
[1] 1.41
> options(OutDec = ",")
> (a <- sqrt(2))
[1] 1,41
```

Introduction to R Statistical Software

- The beginning of R
- R - Introducing the R Console
- R - Installation, Packages, CRAN, Components
- Graphical User Interfaces: R Console, R Studio, R Commander, R resources and online community

R - Environment



R - Packages

R Core + R core(basic) packages

Installed and loaded initialy:
- stats; graphics; grDevices; utils; datasets; methods; base

Installed but not loaded:
- base; boot; class; cluster; codetools; compiler; datasets; foreign; graphics; grDevices; grid; KernSmooth; lattice; MASS; matrix; methods; mgcv; nlme; nnet; parallel; rpart; spatial; splines; stats; stats4; survival; tcltk; tools; utils

CRAN pakages

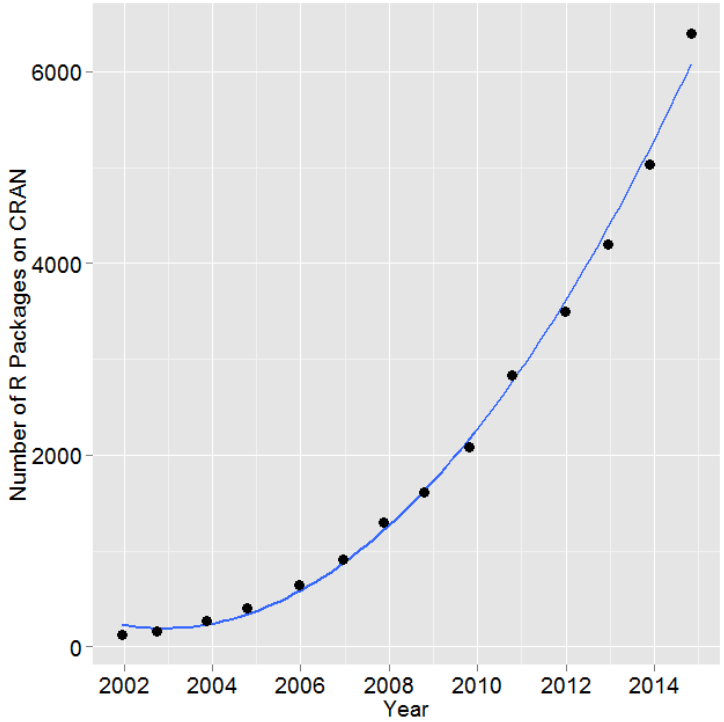
Contributed packages: ggplot2; zoo; ggmap; 7.168+

Other pakages

not included in CRAN

source: https://cran.r-project.org/web/packages/, [15-sep-2015]

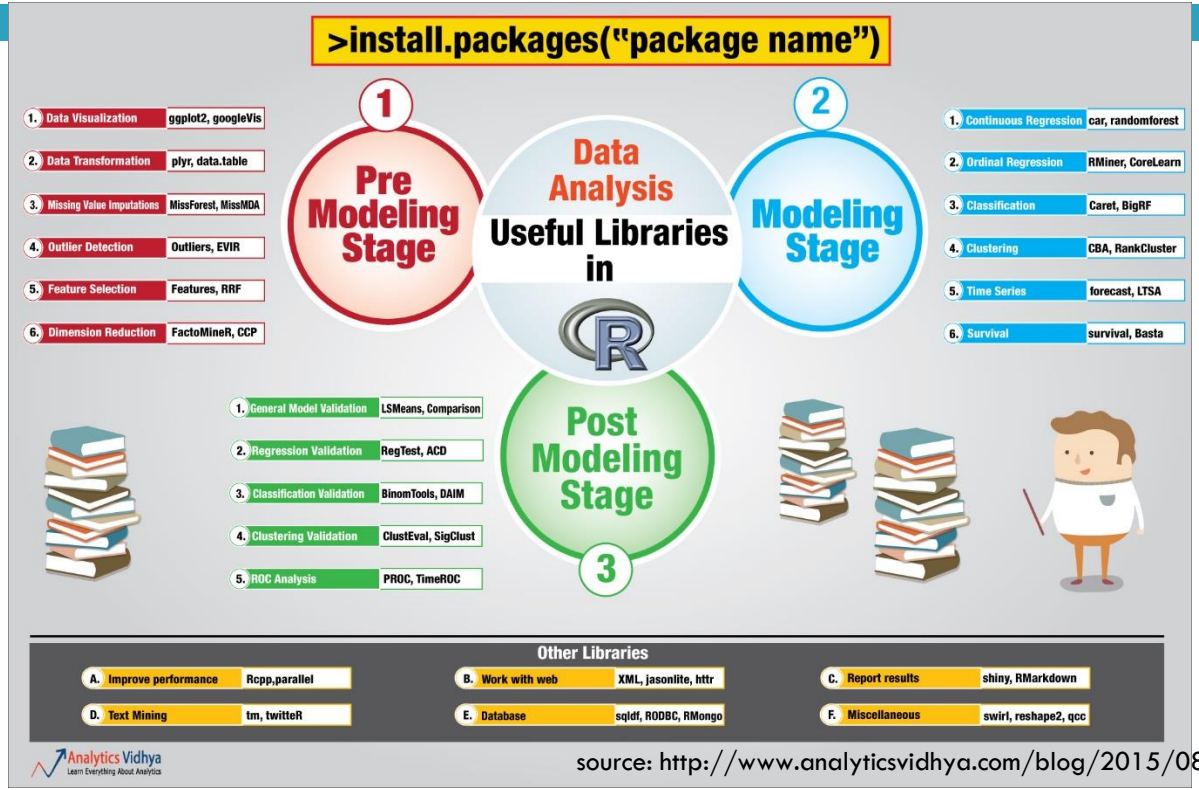
R - Contributed packages 7.168+



Number of R packages available on its main distribution site for the last version released in each year.

source: http://r4stats.com/articles/popularity/, [03-oct-2015]

R - Most commonly used packages



source: http://www.analyticsvidhya.com/blog/2015/08/list-r-packages-data-analysis/, [03-oct-2015]

R – Managing the package

- search() – list all attached (loaded) packages
- getOption("defaultPackages")

- library() – list all available packages
- library(lattice) – attach the package *lattice*
- require(lattice) – the same like library
- detach("package:lattice") – remove the package from the memory

- install.packages("zoo")
  - install.packages("zoo", pos = 2)
  - install.packages("zoo", pos = "package:base")
- remove.packages("zoo")
- update.packages()

R – CRAN

- Comprehensive R Archive Network (CRAN)
  https://cran.r-project.org/web/views/
- CRAN Mirrors
  https://cran.r-project.org/mirrors.html
- CRAN Task Views
  https://cran.r-project.org/web/views/

```
...
Distributions             Probability Distributions
Econometrics              Econometrics
Environmetrics            Analysis of Ecological and Environmental Data
...
Finance                   Empirical Finance
Genetics                  Statistical Genetics
Graphics                  Graphic Displays & Dynamic Graphics & Graphic Devices & Visualization
NumericalMathematics      Numerical Mathematics
OfficialStatistics        Official Statistics & Survey Methodology
Optimization              Optimization and Mathematical Programming
ReproducibleResearch      Reproducible Research
SocialSciences            Statistics for the Social Sciences
TimeSeries                Time Series Analysis
...
```

R – CRAN Task Views

- install.packages("ctv") – necessary for install.views() and update.views() functions
- library("ctv")
- install.views("SocialSciences")
- update.views("SocialSciences")

```
> install.packages("ctv")
Installing package into 'C:/Users/PCG/Documents/R/win-library/3.2'
...
> library("ctv")
> install.views("SocialSciences")
Installing packages into 'C:/Users/PCG/Documents/R/win-library/3.2'
(as 'lib' is unspecified)
also installing the dependencies 'quadprog', 'ash', 'misc3d', 'multicool', 'DEoptimR', 'pcaPP', 'timeDate',
'tseries', 'fracdiff', 'hdrcde', 'ks', 'sde', 'fda', 'iterators', 'miscTools', 'xtable', 'svd', 'chron',
'RcppArmadillo', 'flexmix', 'modeltools', 'DBI', 'robustbase', 'inline', 'rrcov', 'profileModel',
'forecast', 'rainbow', 'ftsa', 'cobs', 'strucchange', 'RCurl', 'statnet.common', 'trust', 'lpSolve',
'foreach', 'estimability', 'ape', 'corpcor', 'tensorA', 'cubature', 'randomForest', 'maxLik', 'statmod',
'rgenoud', 'RItools', 'polspline', 'networkDynamic', 'tergm', 'ergm.count', 'data.table'
...
```

Introduction to R Statistical Software

- The beginning of R
- R - Introducing the R Console
- R - Installation, Packages, CRAN, Components
- Graphical User Interfaces: R Console, R Studio, R Commander, R resources and online community

GUI: R Console, R Studio, R Commander

- R Console: default GUI, include: the default multipledocument interface (MDI) and the single-document interface (SDI)
- R Studio: probably most complex GUI or integrated development environment (IDE) for R [https://www.rstudio.com/products/rstudio/features/, 2015-09-20]
- R Commander: contributed package Rcmdr: basic statistics GUI

R Studio – features 1

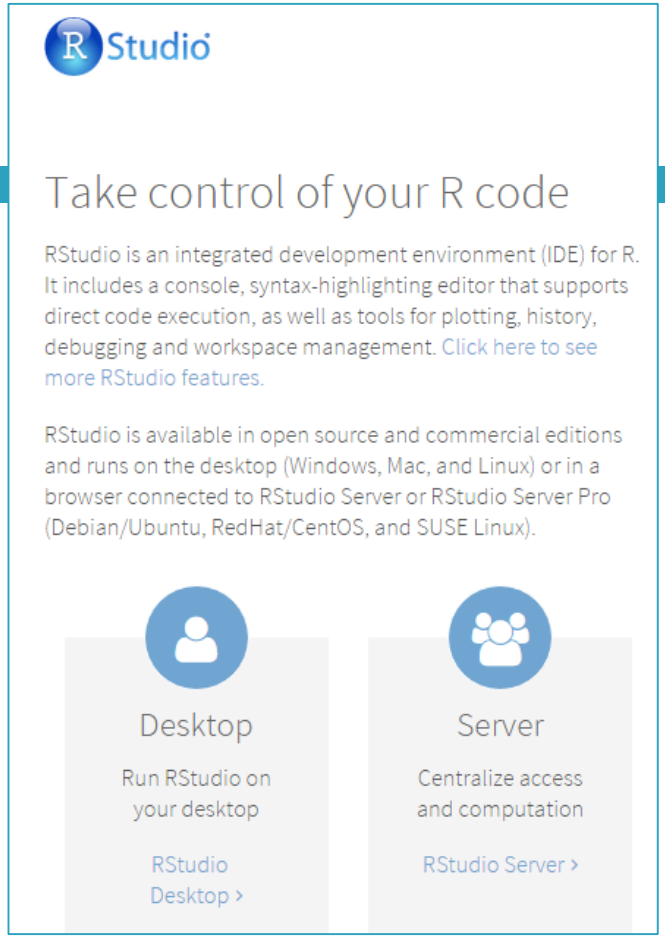
- open source and commercial editions integrates the tools you use with R into a single environment
- available for Windows, Mac and Linux
- running on desktop, web browser and server
- efficient navigation to files and functions
- structure your work into projects
- integrated support for Git and subversion
- authoring HTML, PDF, Word Documents, and slide shows
- supports interactive graphics with Shiny and ggvis

R Studio – features 2

- Integrated Development Environment (IDE):
  - syntax highlighting
  - code completion
  - smart indentation
  - execute R code directly from the source editor
  - quickly jump to function definitions
- Bring your workflow together:
  - Integrated R help and documentation
  - Easily manage multiple working directories using projects
  - Workspace browser and data viewer
- Authoring & Debugging:
  - Interactive debugger to diagnose and fix errors quickly
  - Extensive package development tools
  - Authoring with Sweave and R Markdown

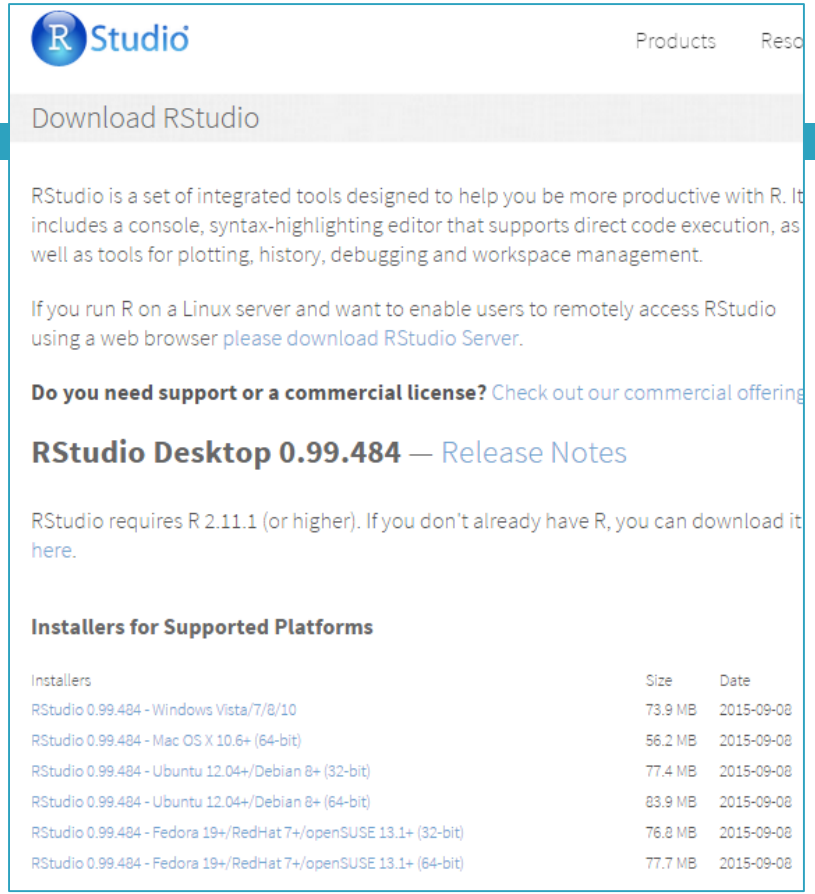
R Studio – Installation

- https://www.rstudio.com/



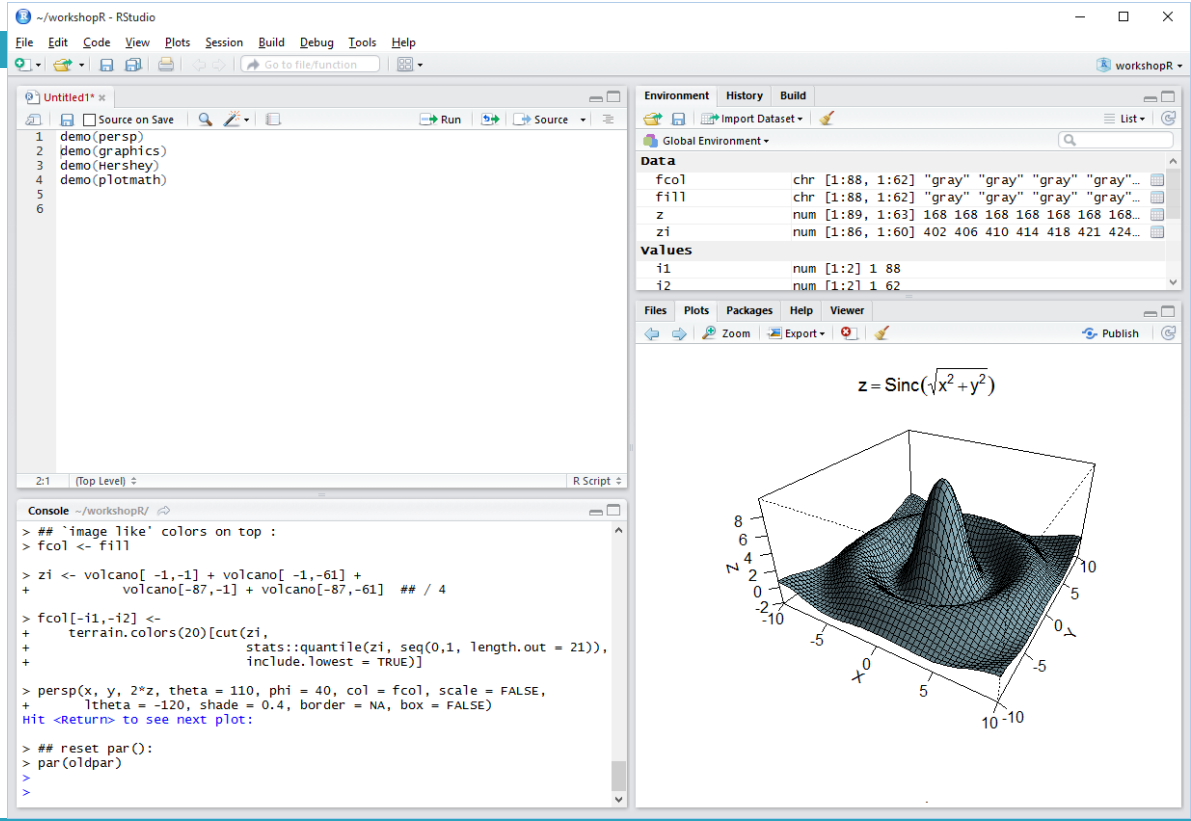
R Studio – Version

- RStudio Desktop 0.99.484 – available from 2015-09-08
- RStudio requires R version 2.11.1 (or higher)



R Studio – IDE

- 4 working area:
  - Text/Commands /Script editor
  - Console
  - Environment, History, Build
  - Files, Plots, Packages, Help, Viewer



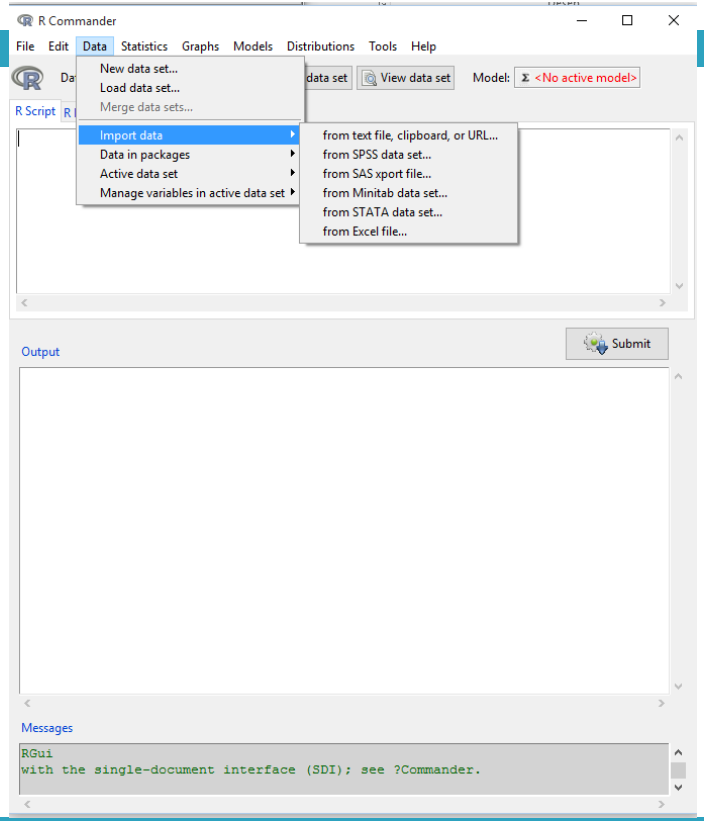
R Commander – Installation

- install.packages("Rcmdr")
- library(Rcmdr)

- Features:
  - Data manipulation
  - Statistics – basic statistical analyses
  - Graphs – simple statistical graphs
  - Models – numerical summaries, confidence intervals, hypothesis tests, diagnostics, and graphs for a statistical model, and for adding diagnostic quantities (eg, residuals) to the data set
  - Distributions – probabilities, quantiles, and graphs of standard statistical distributions

R Commander – other

- Script/editor window shows the command generates by user interactions with the menus.
- Edit/View data sets (reasonable small data sets)
- Data import from plain-text, Minitab, SPSS, or STATA
- Output window
- Messages window
- Graphics Device windows (appear separately)
- Save Graphs to different file type: bitmap, PDF, Postscript, EPS…
- Submit or Ctrl+r will run your commands



R resources

- Vignette – specific documentation for packages
- CRAN – The R Manuals [https://cran.r-project.org/manuals.html]
- CRAN – Contributed Documentation [https://cran.r-project.org/other-docs.html]
- CRAN – Reference Card, eg R reference card v2 [https://cran.r-project.org/doc/contrib/Baggott-refcard-v2.pdf]
- CRAN - Non-English Documents, eg “R pentru Incepatori” , the Romanian version of “R for Beginners” by Emmanuel Paradis, translated by Ana-Maria Dobre (PDF, 2013-08-10) [https://cran.r-project.org/doc/contrib/Paradis-rdebuts_RO.pdf]

R online community

- R-FAQ - [https://cran.r-project.org/doc/FAQ/R-FAQ.html]
- Search engine – [http://rseek.org/]
- R news and tutorials contributed by (573) R bloggers – [http://www.r-bloggers.com/]
- Google Groups
- LinkedIn Groups
- Quick-R – [http://www.statmethods.net/]
- R for SPSS and SAS users – [http://r4stats.com/]
- R Journal – [http://journal.r-project.org]
- MRAN (Managed R Archive Network) – [http://mran.revolutionanalytics.com]

R – Who can we ask?

- R-help mailing list – [http://www.r-project.org/mail.html]
- Stackoverflow – [http://stackoverflow.com/questions/tagged/r]
- Talk Stats – [http://www.talkstats.com]
- CrossValidated – [http://stats.stackexchange.com/]
- StatExchange – [http://stats.stackexchange.com]

R – But before asking…

- Read carefully online help
- Read R-FAQ
- Archives `r-help` of R mailing list
- Just use search engines or Google RSeek
- Posting Guide: How to ask good questions that prompt useful answers [https://www.r-project.org/posting-guide.html]

R – What to keep in mind when asked a question online community?

- How came the problem?
- What is the expected output?
- What is wrong in output? / What is the output obtained?
- What version of R / package use?
- What is your operating system?

- Question title: R 3.2.2 LM() FUNCTION ON WINDOWS 7 -- SEG FAULT ON LARGE DATA FRAME

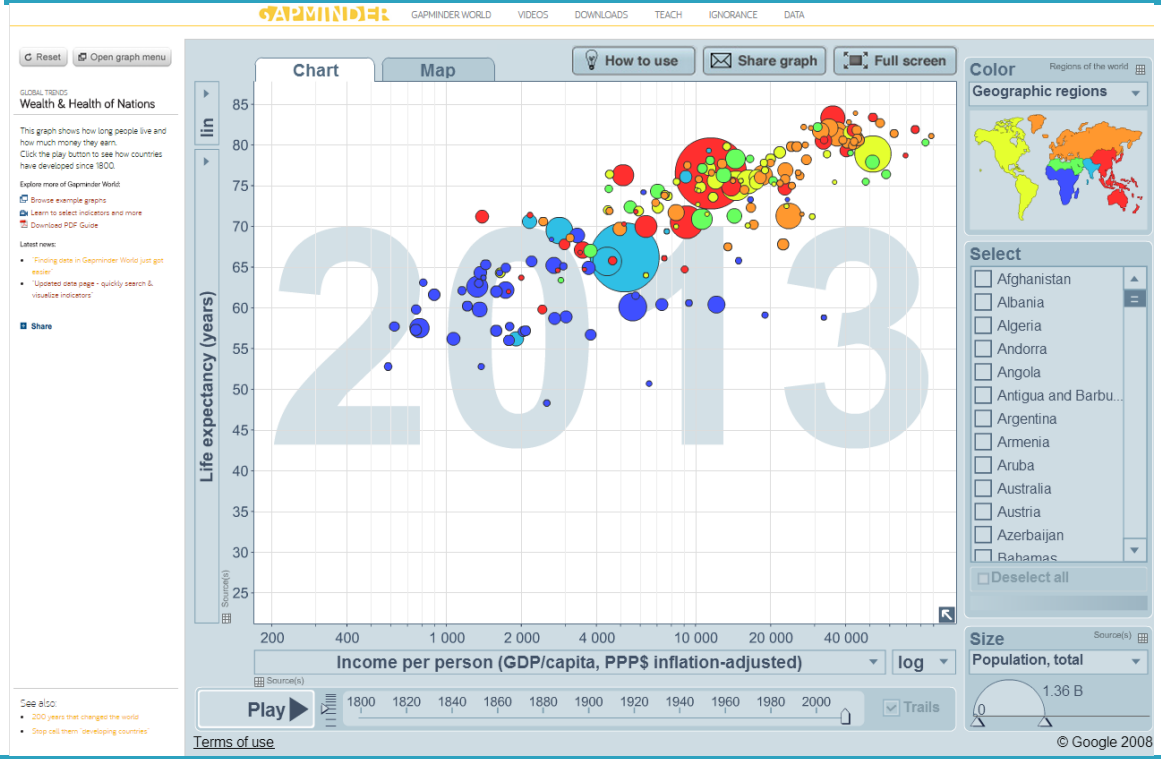
R code style guide

- Google's R Style Guide [http://google-styleguide.googlecode.com/svn/trunk/Rguide.xml]
- R code style guide [http://4dpiecharts.com/r-code-style-guide/]
- Learning R: Index of Online R Courses [http://www.r-bloggers.com/learning-r-index-of-online-r-courses-october-2015/]

R - What are the limits?

source: https://www.facebook.com/notes/facebook-engineering/visualizing-friendships/469716398919, [03-oct-2015]

Your imagination!



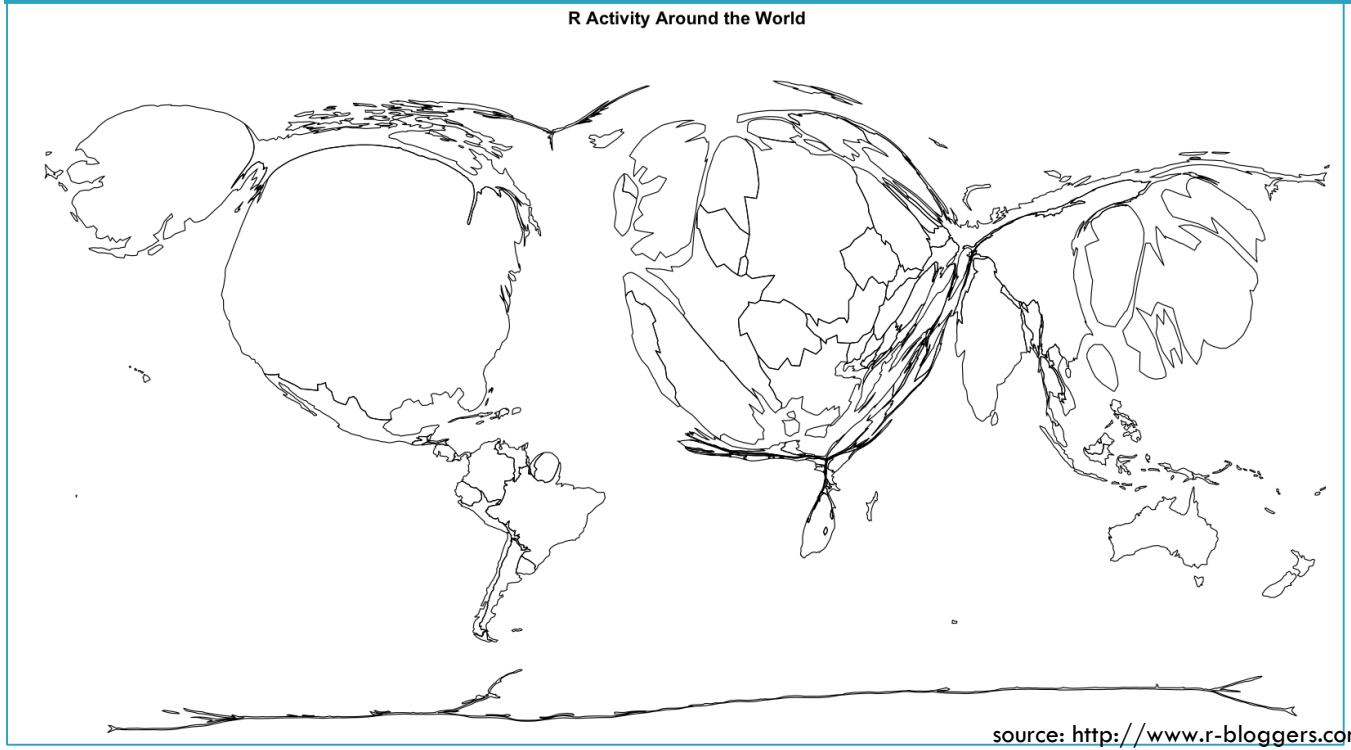
source: http://www.gapminder.org/, [03-oct-2015]

Millions of data!

The world become larger than ever!

source: https://www.quandl.com/platform/for-institutions, [03-oct-2015]

R lives!



- http://www.r-bloggers.com/r-activity-around-the-world/
- http://cran-logs.rstudio.com/
- http://www.r-bloggers.com/analyzing-package-dependencies-and-download-logs-from-rstudio-and-a-start-towards-building-an-r-recommendation-engine/
- http://www.r-bloggers.com/where-in-the-world-is-r-and-rstudio/?utm_source=feedburner&utm_medium=email&utm_campaign=Feed%3A+RBloggers+%28R+bloggers%29
- http://www.omegahat.org/Rcartogram/

source: http://www.r-bloggers.com/where-is-the-r-activity/, [03-oct-2015]

Thank you!

Ciprian Alexandru
alexcipro@yahoo.com