Getting Started

Mandy Vogel

May 31, 2015

Table of Contents

1. R
   - What is R?
   - Why R?
   - Getting R
2. Packages
   - What are packages for?
   - How do I find the one I want
3. First Session
   - Starting
   - The Workspace
   - Help
   - Quitting
4. Citation/License
   - Citation
   - License
   - Exercises

Table of Contents

What's R?

- R is a high-level language and an environment for data analysis and graphics
- influenced by S (Becker, Chamber, Wilks) and Scheme (Sussman)
- and created by Ross Ihaka and Robert Gentleman at the university of Auckland
- R is free.
- R is open source.
- R is a dialect of S system.

What R can do...

R provides a wide variety of statistical and graphical techniques including

- linear and nonlinear modelling
- classical statistical tests
- time-series analysis
- classification
- clustering and many more

What R can do...

R provides a wide variety of statistical and graphical techniques including

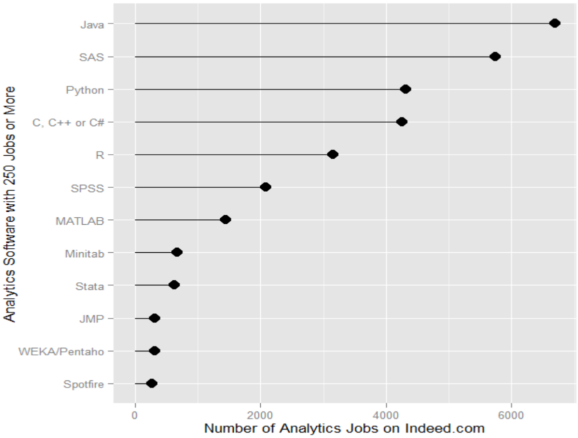
- linear and nonlinear modelling
- classical statistical tests
- time-series analysis
- classification
- clustering and many more

R is easily extensible, can produce publication-quality graphs including mathematical symbols; dynamic and interactive graphics are available through additional packages.

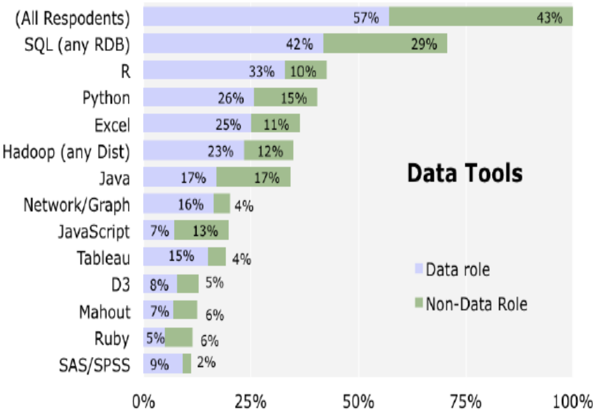
Pros

- R is free and R is open source
- there is a lot of material and books available
- there is a lot of help on the web, including developers who are active in mailing lists
- most of your problems are already solved and with a high probability the solution is available from one of the repositories (as package)
- there are a lot of intuitive GUIs
- the language is easy to learn and also intuitive
- the graphics capabilities are impressive

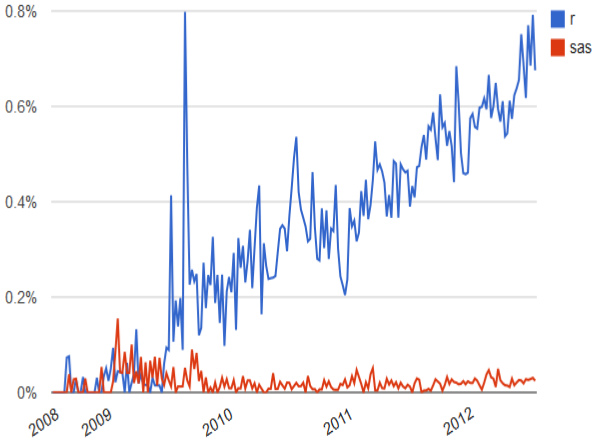
The number of analytics jobs for the more popular software (250 jobs or more, 2/2014).



OReilly Data Science Survey results for 2012 and 2013 combined.



OReilly Data Science Survey results for 2012 and 2013 combined



Cons

- there is a LOT of help on the web
- with a high probability there is more than one solution for your problem
- there are a lot of intuitive GUIs so you have to decide what you want (so first you have to *know* what you want)
- the real power of R (i.e. high flexibility) is not entirely available through GUIs
- and therefore the learning curve can be lengthy in the beginning (but soon accelerating ;)

Use it!

The best way to learn R is to use it!

Where can I get it?

For the basic installation CRAN is a good place to start

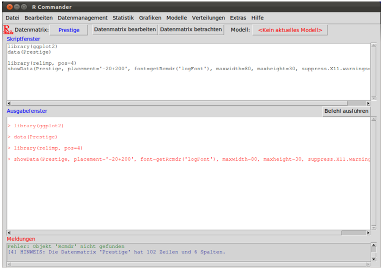
- CRAN stands for Comprehensive R Archive Network
- `http://cran.r-project.org`
  - Microsoft Windows: :: http://cran.r-project.org/bin/windows/base/
  - MacOS: :: http://cran.r-project.org/bin/macosx/
  - Linux: :: http://cran.r-project.org/bin/linux/
- for mac and pc users: just download and install the precompiled binaries
- for ubuntu users: add
  `deb http://ftp5.gwdg.de/pub/misc/cran/bin/linux/ubuntu oneiric/`
  to
  `/etc/apt/sources.list`;
  detailed howto:
  `http://cran.r-project.org/bin/linux/ubuntu/`

The R-Commander

The R commander, developed by John Fox is a complete GUI for R. It is implemented in the package `Rcmdr`:

- `Rcmdr` has a comprehensive menu, which includes data reading, summaries, statistical analyses, etc.
- When the menu is activated, the `Rcmdr` will generate an R script. This script can be used as a log for documentation or for self learning.
- It has excellent graphical tools.

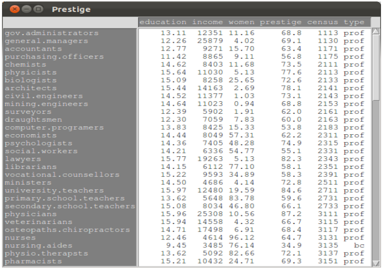
The R-Commander



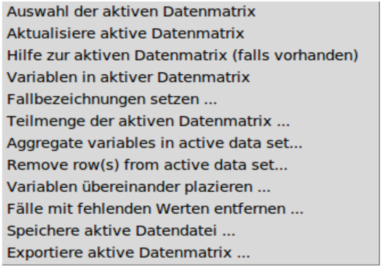
```
library(ggplot2)
data(Prestige)

library(relimp, pos=4)
showData(Prestige, placement='-20+200', font=getRcmdr('logFont'), maxwidth=80, maxheight=30, suppress.X11.warnings=
```

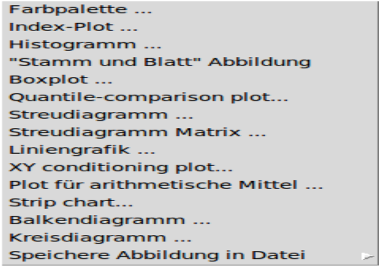
```
> library(ggplot2)

> data(Prestige)

> library(relimp, pos=4)

> showData(Prestige, placement='-20+200', font=getRcmdr('logFont'), maxwidth=80, maxheight=30, suppress.X11.warning
```

```
Fehler: Objekt 'Rcmdr' nicht gefunden
[4] HINWEIS: Die Datenmatrix 'Prestige' hat 102 Zeilen und 6 Spalten.
```



```
                            education income women prestige census type
gov.administrators              13.11  12351 11.16     68.8   1113 prof
general.managers                12.26  25879  4.02     69.1   1130 prof
accountants                     12.77   9271 15.70     63.4   1171 prof
purchasing.officers             11.42   8865  9.11     56.8   1175 prof
chemists                        14.62   8403 11.68     73.5   2111 prof
physicists                      15.64  11030  5.13     77.6   2113 prof
biologists                      15.09   8258 25.65     72.6   2133 prof
architects                      15.44  14163  2.69     78.1   2141 prof
civil.engineers                 14.52  11377  1.03     73.1   2143 prof
mining.engineers                14.64  11023  0.94     68.8   2153 prof
surveyors                       12.39   5902  1.91     62.0   2161 prof
draughtsmen                     12.30   7059  7.83     60.0   2163 prof
computer.programers             13.83   8425 15.33     53.8   2183 prof
economists                      14.44   8049 57.31     62.2   2311 prof
psychologists                   14.36   7405 48.28     74.9   2315 prof
social.workers                  14.21   6336 54.77     55.1   2331 prof
lawyers                         15.77  19263  5.13     82.3   2343 prof
librarians                      14.15   6112 77.10     58.1   2351 prof
vocational.counsellors          15.22   9593 34.89     58.3   2391 prof
ministers                       14.50   4686  4.14     72.8   2511 prof
university.teachers             15.97  12480 19.59     84.6   2711 prof
primary.school.teachers         13.62   5648 83.78     59.6   2731 prof
secondary.school.teachers       15.08   8034 46.80     66.1   2733 prof
physicians                      15.96  25308 10.56     87.2   3111 prof
veterinarians                   15.94  14558  4.32     66.7   3115 prof
osteopaths.chiropractors        14.71  17498  6.91     68.4   3117 prof
nurses                          12.46   4614 96.12     64.7   3131 prof
nursing.aides                    9.45   3485 76.14     34.9   3135   bc
physio.therapsts                13.62   5092 82.66     72.1   3137 prof
pharmacists                     15.21  10432 24.71     69.3   3151 prof
```


Auswahl der aktiven Datenmatrix
Aktualisiere aktive Datenmatrix
Hilfe zur aktiven Datenmatrix (falls vorhanden)
Variablen in aktiver Datenmatrix
Fallbezeichnungen setzen ...
Teilmenge der aktiven Datenmatrix ...
Aggregate variables in active data set...
Remove row(s) from active data set...
Variablen übereinander plazieren ...
Fälle mit fehlenden Werten entfernen ...
Speichere aktive Datendatei ...
Exportiere aktive Datenmatrix ...



The R-Commander

To install `Rcmdr` go to Packages → Install package(s) (or simply type `install.packages("Rcmdr")`), then choose a CRAN mirror close to you, than OK. A window with a list of packages will pop-up, on this list choose `Rcmdr` and OK. A bundle of packages will be automatically installed.
To run the "R Commander" GUI type at the prompt line:

```
> library(Rcmdr)
```

This will start a GUI similar to other statistical software. Therefore, any typical process, like read data, produce plots, make statistical analyses, etc. will be made by clicking the appropriate menu.

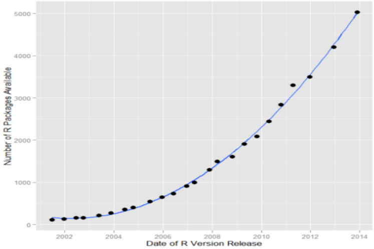
Table of Contents

1. R
   - What is R?
   - Why R?
   - Getting R
2. Packages
   - What are packages for?
   - How do I find the one I want
3. First Session
   - Starting
   - The Workspace
   - Help
   - Quitting
4. Citation/License
   - Citation
   - License
   - Exercises

Packages

The capabilities of R are extended through user-created packages, which allow specialized statistical techniques, graphical devices, import/export capabilities, reporting tools, etc.

These packages are developed primarily in R, and sometimes in Java, C and Fortran. A core set of packages is included with the installation of R, with more than 6381 (as of Feb 2015) available at the Comprehensive R Archive Network (CRAN), 2095 (936 Software packages) on Bioconductor, and more on other repositories (e.g. R-Forge).



R taskviews

- of course: you can google your problem: but you should use `http://www.rseek.org/` instead of `www.google.com`; `rseek` is a google custom search, can easily be added to the toolbar of popular browsers
- `http://cran.r-project.org/web/views/`
- before you install a new package: `help.search()` allows for searching the help system for documentation matching a given character string in the (file) name, alias, title, concept or keyword entries (or any combination thereof), using either fuzzy matching or regular expression matching.(installed help system)

Packages

- An R installation contains a library of packages. Some of these packages are part of the basic installation. These packages have the *recommended* status.
- Others (over 6000) can be downloaded from CRAN.
- A package is loaded into R using the `library()` or the `require()` command. For example to load the survival package you should enter

```
> library(survival)
```

- The loaded packages are not considered part of the workspace. You need to load a package when you start a new R session.

Getting Packages

- you can download a package from CRAN and install by using the *package menu* (bottom right corner)
- another effective way to download and install a package is by command line. For example the following line install the R commander package with all its dependencies:

  ```
  > install.packages("Rcmdr", dependencies=TRUE)
  ```

- install now the packages `ggplot2` and `faraway`

Table of Contents

1. R
   - What is R?
   - Why R?
   - Getting R
2. Packages
   - What are packages for?
   - How do I find the one I want
3. First Session
   - Starting
   - The Workspace
   - Help
   - Quitting
4. Citation/License
   - Citation
   - License
   - Exercises

First Session I

- start R (depending on you OS and UI) by double-clicking the R icon, typing R in a console or starting you favorite UI
- choose your working directory (via a menu or by typing `setwd('/your/directory/')`
- R works fundamentally by the question-and-answer model: you enter a line with a command and press Enter ($\hookleftarrow$). Then the program does something, prints the results, and asks for more input. When R is ready for input, it prints out its prompt, a ”$>$”. It is possible to use R as a text-only application, and also in batch mode.

First Session

One of the simplest possible tasks in R is to enter an arithmetic expression and receive a result.

```
> 2 + 2
[1] 4
> exp(-2)
[1] 0.1353353
> round(exp(-2),3)
[1] 0.135
>
```

First Session

- During a session you create a workspace. The workspace contains all variables created, for example typing

  ```
  > x <- rnorm(100, mean=2, sd=4)
  > y <- 2 * x  + rnorm(100, mean=0, sd=0.5)
  ```

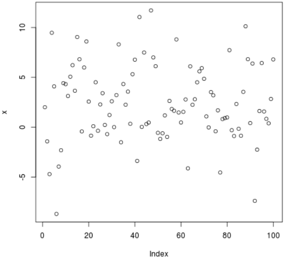
  creates a vector variable with 100 random numbers from a normal distribution with mean 2 and standard deviation 4 and a second vector containing also 100 numbers dependend on x
- to see the contents of this variables just type its names, e.g.

  ```
  > x
   [1]  2.663558 2.187709 -1.849147
                5.566364 2.5016523.046095  ...
  ```

First Session

- To plot these values type

```
> plot(x)
```



Nothing is lost or hidden

- statistical packages provide *canned* procedures to address common statistical problems
- canned procedures are useful for routine analysis, but they are also limiting - you can only do what the programmer lets you do
- in R, the result of statistical calculation are always accessible, so
  - you can use them for further calculations
  - you can always see how calculations were done
- what you see in the first place is most of the time only a small part of the result

Nothing is lost or hidden

- for example building a linear model gives you the following result

```
> lm(y˜x)

Call:
lm(formula = y ˜ x)

Coefficients:
(Intercept)            x
    0.07101      1.98083
```

Nothing is lost or hidden

- save the model in an object mm so you can run functions on it, e.g. `summary()` or `plot()`

```
> mm <- lm(y~x)
> summary(mm)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.67162 -0.36329  0.02206  0.29193  1.36333 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.07101    0.06353   1.118    0.266    
x            1.98083    0.01404 141.043   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.5385 on 98 degrees of freedom
Multiple R-squared:  0.9951, Adjusted R-squared:  0.995 
F-statistic: 1.989e+04 on 1 and 98 DF,  p-value: < 2.2e-16
```

Nothing is lost or hidden

- to explore the object you can use the object browser or `str()`

```
> str(mm)
List of 12
 $ coefficients : Named num [1:2] 0.071 1.981
  ..- attr(*, "names")= chr [1:2] "(Intercept)" "x"
 $ residuals    : Named num [1:100] 0.6828 -0.1417 0.0992 0.1057 ...
  ..- attr(*, "names")= chr [1:100] "1" "2" "3" "4" ...
 $ effects      : Named num [1:100] -48.23926 -75.95182 0.04414 ...
  ..- attr(*, "names")= chr [1:100] "(Intercept)" "x" "" "" ...
 $ rank         : int 2
 $ fitted.values: Named num [1:100] 9.18 13.41 8.12 15.11 5.96 ...
  ..- attr(*, "names")= chr [1:100] "1" "2" "3" "4" ...
 $ assign       : int [1:2] 0 1
 $ qr           :List of 5
  ..$ qr   : num [1:100, 1:2] -10 0.1 0.1 0.1 0.1 0.1 0.1  ...
  .. ..- attr(*, "dimnames")=List of 2
  .. .. ..$ : chr [1:100] "1" "2" "3" "4" ...
  .. .. ..$ : chr [1:2] "(Intercept)" "x"
  .. ..- attr(*, "assign")= int [1:2] 0 1
...
```

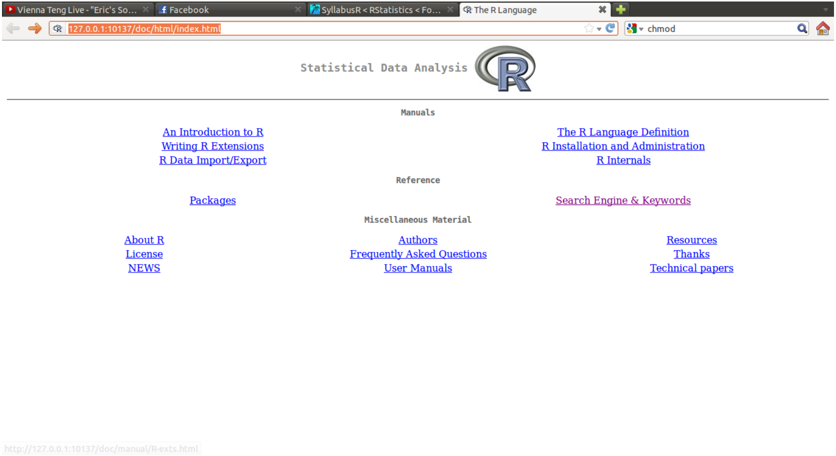
First Session

- Entering the command

  ```
  > help.start()
  ```

  at the command line, will launch an extensive online help that can be read using a Web browser such as Firefox or Internet Explorer. Another way to access to these ”help” pages is the help tab in the bottom right corner. Notice that the HTML version of the help system has a very useful ”Search Engine and Keywords”.

First Session


Vienna Teng Live - "Eric's So...
Facebook
SyllabusR < RStatistics < Fo...
The R Language
127.0.0.1:10137/doc/html/index.html
chmod

Statistical Data Analysis

Manuals

An Introduction to R
Writing R Extensions
R Data Import/Export

The R Language Definition
R Installation and Administration
R Internals

Reference

Packages

Search Engine & Keywords

Miscellaneous Material

About R
License
NEWS

Authors
Frequently Asked Questions
User Manuals

Resources
Thanks
Technical papers

http://127.0.0.1:10137/doc/manual/R-exts.html

First Session

- All variables, functions and diverse objects can be seen by the `ls()` and the newer version of it `objects()` function. Thus in our example we will have

  ```
  > ls()
  [1] "x"
  ```

- quitting R is done with the `q()` function

  ```
  > q()
  ```

  at the command prompt. You will be asked to save your "workspace image". Give a name for your workspace for example "project1", if you want to save your workspace. Or use `save()` to save only the objects you want to keep.

- you can load this workspace in a new R session. On windows just click directly on the workspace file and R will be opened and automatically load the workspace

Table of Contents

1. R
   - What is R?
   - Why R?
   - Getting R
2. Packages
   - What are packages for?
   - How do I find the one I want
3. First Session
   - Starting
   - The Workspace
   - Help
   - Quitting
4. Citation/License
   - Citation
   - License
   - Exercises

Citation

Input

```
> citation()
```

```
To cite R in publications use:

  R Development Core Team (2012). R: A language and environment for
  statistical computing. R Foundation for Statistical Computing,
  Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

A BibTeX entry for LaTeX users is

  @Manual,
    title = R: A Language and Environment for Statistical Computing,
    author = R Development Core Team,
    organization = R Foundation for Statistical Computing,
    address = Vienna, Austria,
    year = 2012,
    note = ISBN 3-900051-07-0,
    url = http://www.R-project.org/,


We have invested a lot of time and effort in creating R, please cite it
when using it for data analysis. See also citation("pkgname") for
citing R packages.
```

Licence

Licence

R is mainly distributed under the terms of the GNU General Public License, either Version 2, June 1991 or Version 3, June 2007. Core Bioconductor packages are typically licensed under Artistic-2.0.

You get detailed information with: `license()`, `RShowDoc("COPYING")`, `packageDescription("packagename")$License`

Exercises I

1. start R in your favourite UI!
2. try to load the MASS package (`libary()`)! The MASS package belongs to the recommended packages and should be included in every standard installation of R. If - contrary to expactations - the MASS package is not installed, install it using the `install.packages()` command or the appropriate menue of your ui.
3. after loading the package try to load some example data: the `data()` command loads data contained in packages. Here we want to load and inspect the Pima.tr data set. Type the following lines. What are they for?

Exercises II

```
> library(MASS)
> ## require(MASS)
> data(Pima.te)
> ?Pima.te
> names(Pima.te)
> head(Pima.te)
> summary(Pima.te)
```

So Pima.te is here the name of the table. `names()` gives you the names of the columns. Remember: To access one particular column type

```
> table_name$column_name
```

So to get the column `skin` out of the `Pima.te` data frame type

```
> Pima.te$skin
```