Data Analysis and Unsupervised Learning
Introduction to R

MAP573 – Julien Chiquet

École Polytechnique, Autumn semester, 2020-2021

https://jchiquet.github.io/MAP573

ACCÉLÉRATEUR &
INCUBATEUR DE START-UP
ÉCOLE POLYTECHNIQUE

la science pour la vie, l'humain, la terre

Outline

1. What is R?
2. Getting started
3. First steps
4. R markdown
5. Interfacing to other languages

Outline

1. What is R?

2. Getting started

3. First steps

4. R markdown

5. Interfacing to other languages

R ?

In a nutshell

R is a scientific software specialized in calculation and statistical analysis.

It is also

- a programming language,
- a environment/interpreter,
- an open-source project (GNU-R)
- a multi-plateform software (Linux, Mac, Windows)

A bit of history

- 1970s: S-language developed at Bell labs (Chambers, Beckers)
- 1980s: S−PLUS developed at AT&T. Lab
- 1990s: R is developed as a GNU/GPL open-source counterpart to S by Gentleman and Ihaka (Auckland university)
- 1997: The R-core team now leads the development
- 2002: The R foundation is created and chaired by Gentleman and Ihaka
- 2011: first public of R-studio (JJ Allaire)
- 2019: Rstudio lead scientist H. Wickham receives COPSS Award (statistician Nobel price)

Remarkable basics features

Scientific Computing

- linear algebra
- statistical models and data analysis

Data manipulation an visualization

- import, export, transformation
- great, versatile plotting system

Interfacing is easy

- to most programming languages (C/C++, Python)
- to most database systems (SQL, postgrey)
- for distributed computing (Hadoop, H20, spark)

Package manager

- Extremely versatile

Why R ?

Community

- SatRDay, R user groups, meet-up, conference
- CRAN community https://cran.r-project.org/
- Rstudio community https://community.rstudio.com/
- R dev/package well integrated on github

Packages manager

- more than 13,000 community-based libraries
- cutting-edges statistical methods
- easy to learn even for non-statistician/data scientist

Reproducibility

- Rmarkdown is not just notebook
- Great for interfacing, plotting, scientific reports

Why not R?

- Easy to make dirty code (less and less true)
- Not typed language, not compilation by default: may be slow
- Many ways to do the same things
- Less well interfaced to ML/Deep-learning library than Python

Why R again?

The Rstudio group

Even if it is a company...

- Rstudio IDE is a great all-in-one tool for data analysis and development
- Cleaner implementation (tidyverse and co)
- New functionalities (unitary test, github integration)
- Interface to deep learning tools (Tensor Flow, Keras, Torch, etc.)
- Interface with Python (reticulate)
- Nice surrogate Oriented-Object programming with R6

⇝ Rstudio basically saved `R` from `Python`

Outline

1. What is R?
2. Getting started
3. First steps
4. R markdown
5. Interfacing to other languages

Setup instructions I

R and **RStudio** are separate downloads and installations

- R is the underlying statistical computing environment
- RStudio is a graphical integrated development environment (IDE)

Windows

1. Download R from the CRAN website and
2. Run the `.exe` file that was just downloaded
3. Go to the RStudio download page
4. Under *Installers* select **RStudio x.yy.zzz - Windows Vista/7/8/10**

Setup instructions II

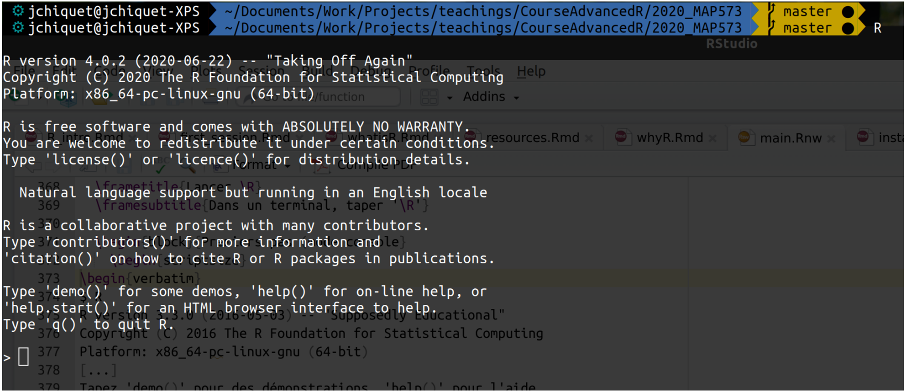
MacOS

1. Download R from the CRAN website.
2. Select the `.pkg` file for the latest R version and double click
3. Go to the RStudio download page
4. Under *Installers* select **RStudio x.yy.zzz - Mac OS X 10.6+ (64-bit)**

Linux

1. Follow the CRAN instructions, to update your /etc/sources.list
2. On Debian/Ubuntu, run `sudo apt-get install r-base`
3. Go to the RStudio download page
4. Under *Installers* select the version that matches your distribution

The R console



```
jchiquet@jchiquet-XPS  ~/Documents/Work/Projects/teachings/CourseAdvancedR/2020_MAP573  master
jchiquet@jchiquet-XPS  ~/Documents/Work/Projects/teachings/CourseAdvancedR/2020_MAP573  master  R

R version 4.0.2 (2020-06-22) -- "Taking Off Again"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-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.

> 
```

Figure 1: Screenshot of the R console

- `help(str)`, `?str`: launch dedicated help for command `str`,
- `help.search("factorial")`, `??factorial`: look for command with key word `factorial`,
- `help.start()`, `???`: launch the HTML help pages in a browser

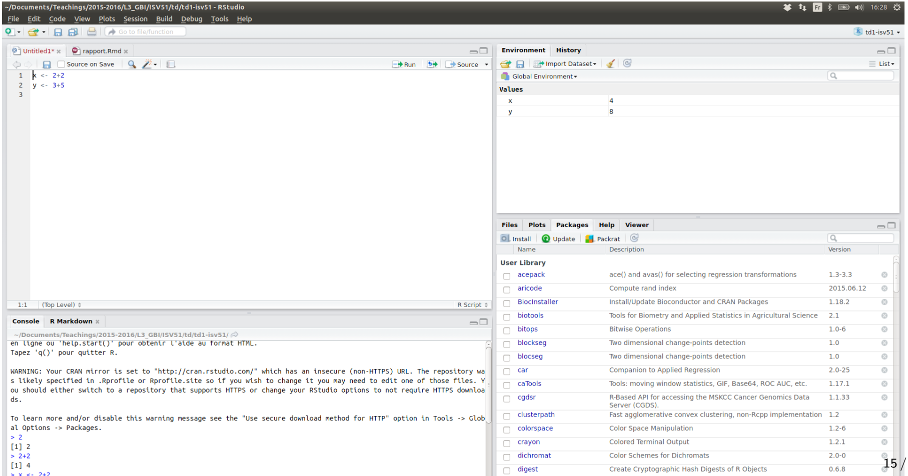
The Rstudio IDE

- A full IDE with code, interpreter, workspace and plots
- Package development and external code integration are easier
- Notebooks integration with Rmarkdown
- Interface with github

⇝ Rstudio is a state-of-the-art tool for efficient development in `R`

My favorites shortcuts

- `ctrl + return`: execute current selection in console
- `ctrl + 1/2/3/4`: navigate between panels
- `ctrl + down/up`: navigate between tabs
- `ctrl + shift + k`: knit current doccument



```r
x <- 2+2
y <- 3+5
```

Values

x	4
y	8

```
en ligne ou 'help.start()' pour obtenir l'aide au format HTML.
Tapez 'q()' pour quitter R.

WARNING: Your CRAN mirror is set to "http://cran.rstudio.com/" which has an insecure (non-HTTPS) URL. The repository wa
s likely specified in .Rprofile or Rprofile.site so if you wish to change it you may need to edit one of those files. Y
ou should either switch to a repository that supports HTTPS or change your RStudio options to not require HTTPS downloa
ds.

To learn more and/or disable this warning message see the "Use secure download method for HTTP" option in Tools -> Glob
al Options -> Packages.
> 2
[1] 2
> 2+2
[1] 4
> x <- 2+2
```

	Name	Description	Version
	User Library		
□	acepack	ace() and avas() for selecting regression transformations	1.3-3.3
□	aricode	Compute rand index	2015.06.12
□	BiocInstaller	Install/Update Bioconductor and CRAN Packages	1.18.2
□	biotools	Tools for Biometry and Applied Statistics in Agricultural Science	2.1
□	bitops	Bitwise Operations	1.0-6
□	blockseg	Two dimensional change-points detection	1.0
□	blocseg	Two dimensional change-points detection	1.0
□	car	Companion to Applied Regression	2.0-25
□	caTools	Tools: moving window statistics, GIF, Base64, ROC AUC, etc.	1.17.1
□	cgdsr	R-Based API for accessing the MSKCC Cancer Genomics Data Server (CGDS).	1.1.33
□	clusterpath	Fast agglomerative convex clustering, non-Rcpp implementation	1.2
□	colorspace	Color Space Manipulation	1.2-6
□	crayon	Colored Terminal Output	1.2.1
□	dichromat	Color Schemes for Dichromats	2.0-0
□	digest	Create Cryptographic Hash Digests of R Objects	0.6.8

Academic resources

Conferences

- UseR, annual conference of the R foundation conference
- SatRday, community-led, regional conferences
- Rstudio-conf, annual conference of the Rstudio community

Journals

- The R journal http://journal.r-project.org/
- The Journal of Statistical software https://www.jstatsoft.org/

Important web resources

Institutional

- R fondation web site: http://www.r-project.org/
- CRAN (Comprehensive R Arxiv Network): http://cran.r-project.org/
- Rstudio Community https://rstudio.com

Community

- https://ropensci.org/ promotes reproducible science
- R user groups, meet-up, conference
- Stackoverflow https://stackoverflow.com/

Blogs and plateforms

- Datacamp, online teaching plateform https://www.datacamp.com/
- Rstudio eduction program https://education.rstudio.com/
- Blogs community-driven http://www.inside-r.org/, http://www.r-statistics.com/, http://www.r-bloggers.com/
- Twitter #rstats

Outline

1. What is R?
2. Getting started
3. First steps
4. R markdown
5. Interfacing to other languages

Data Structures in base R

1. Atomic vector (integer, double, logical, character)
2. Recursive vector (list)
3. Factor
4. Matrix and array
5. Data Frame

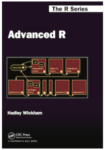
⇝ Creation, Basic Operation, Manipulation, Representation

Resources

- Advanced R, chapters I.2, I.3 (Wickham, 2014, http://adv-r.had.co.nz/)
- An introduction to R programming http://julien.cremeriefamily.info/teachings_L3BI_ISV51.html

Going further

Advanced R (Wickham, 2014), http://adv-r.had.co.nz/



A Language and Environment for Statistical Computing (R Core Team, 2017), https://www.R-project.org/

Basics plotting

⇝ Creation, Basic Operation, Manipulation, Representation

Resources

- tutorial

Outline

1. What is R?

2. Getting started

3. First steps

4. R markdown

5. Interfacing to other languages

R markdown

Figure 3: an authoring framework for data science

R Markdown?

- Markdown is a *lightweight markup language* with plain text formatting syntax that can be converted to HTML. It is completely independent from R. The extension is typically `.md`.
- R Markdown is an *extension of the markdown syntax* that enables R code to be executed. The extention is typically `.Rmd`.
- `rmarkdown` is a library/package which processes and converts `.Rmd` files into a number of different formats, including HTML or `.pdf`. The core function is `rmarkdown::render()`.
- `knitr` is a library/package which processes plain text document with embedded code, executes the code and ‘knits’ the results back into the document. The core function is `knitr::knit()`.

```r
install.packages("rmarkdown")
install.packages("knitr")
```

How does it work?

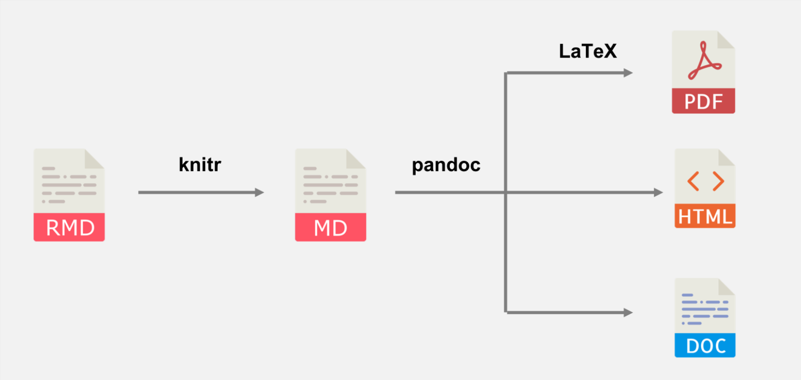
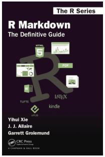


Figure 4: R Markdow workflow

References

Rmarkdown: Dynamic Documents for R (Allaire et al., 2020), https://bookdown.org/yihui/rmarkdown/



Knitr: A General-Purpose Package for Dynamic Report Generation in R (Xie, 2020), https://yihui.name/knitr/

Rstudio doc

See https://rmarkdown.rstudio.com/

R Markdown possibilities

See https://rmarkdown.rstudio.com/

Handle various inputs

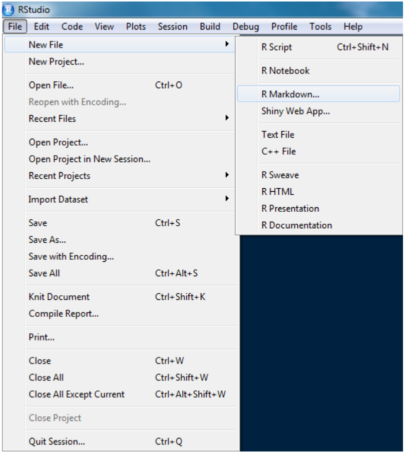
- Markdown Syntax (Markdown reference cheat sheet)
- LaTeX(Advanced mathematical expressions)
- HTML/javascript
- Code chunks (R, Python, Julia and more)

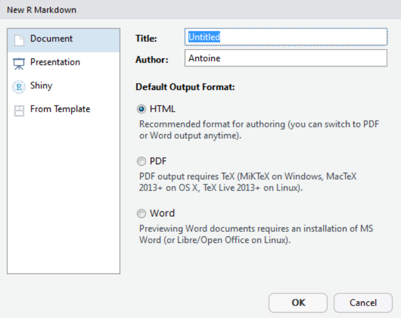
Handle various output

- Rstudio Notebook
- HTML report (static, dynamic)
- HTML website (static, dynamic)
- PDF document
- Doc documents

⇝ More than a Jupyter notebook

Create a new .Rmd





Title: Untitled

Author: Antoine

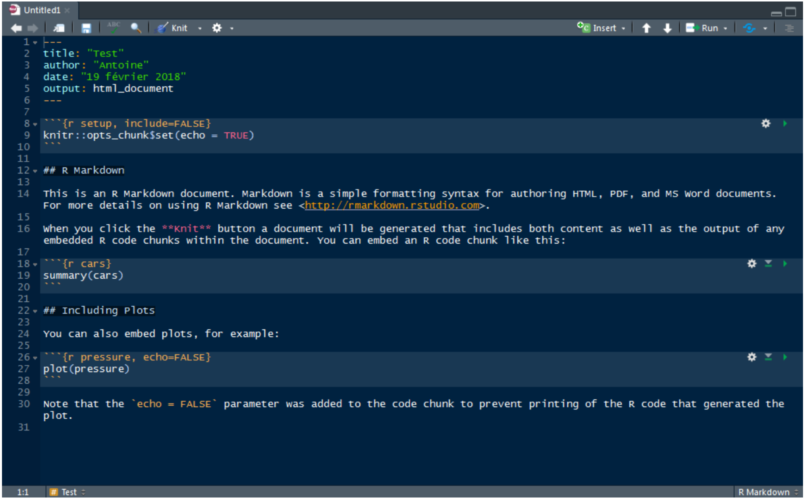
Default Output Format:

● HTML
Recommended format for authoring (you can switch to PDF or Word output anytime).

○ PDF
PDF output requires TeX (MiKTeX on Windows, MacTeX 2013+ on OS X, TeX Live 2013+ on Linux).

○ Word
Previewing Word documents requires an installation of MS Word (or Libre/Open Office on Linux).

New .Rmd



~~~
---
title: "Test"
author: "Antoine"
date: "19 février 2018"
output: html_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
~~~

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents.
For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any
embedded R code chunks within the document. You can embed an R code chunk like this:

```{r cars}
summary(cars)
```

Including Plots

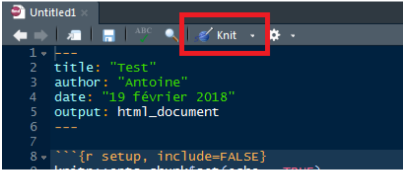
You can also embed plots, for example:

```{r pressure, echo=FALSE}
plot(pressure)
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the
plot.
```
```

Compile `.Rmd`

Use the `Knit` button to produce a HTML file



~~~
---
title: "Test"
author: "Antoine"
date: "19 février 2018"
output: html_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
~~~

Shortcut: Ctrl + Maj + K

Outline

1. What is R?

2. Getting started

3. First steps

4. R markdown

5. Interfacing to other languages

TODO

References

Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., ... Chang, W. (2020). *Rmarkdown: Dynamic documents for R*. Retrieved from https://bookdown.org/yihui/rmarkdown

Eddelbuettel, D. (2013). *Seamless R and C++ integration with Rcpp*. Springer. Retrieved from http://dirk.eddelbuettel.com

Gandrud, C. (2016). *Reproducible research with R and Rstudio*. Chapman; Hall/CRC. Retrieved from https://github.com/christophergandrud/Rep-Res-Book

Gillespie, C., & Lovelace, R. (2016). *Efficient R programming*. " O'Reilly Media, Inc.". Retrieved from https://bookdown.org/csgillespie/efficientR/

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Wickham, H. (2014). *Advanced r*. CRC Press. Retrieved from http://adv-r.had.co.nz/

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer. Retrieved from http://ggplot2.tidyverse.org/reference/

Wilkinson, L. (2006). *The grammar of graphics*. Springer Science & Business