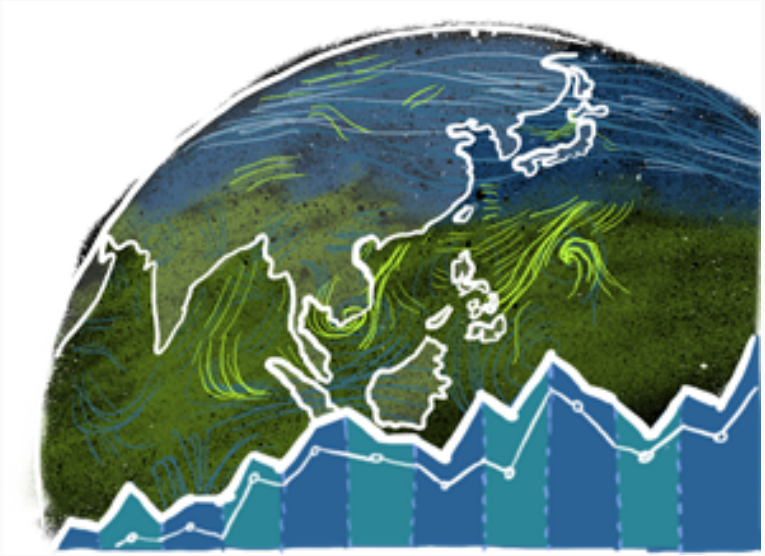
Fundamental of Data Science for ESS



R session 01 - Introduction to R

Daniel Vaulot

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

CNRS • SORBONNE UNIVERSITÉ
Station Biologique de Roscoff

Outline

- What is R and why use R ?
- Resources
- Get started
- Fundamentals of R
  - Data objects
  - Vectors
  - Operators
  - Functions
  - Packages
  - Data frames

Introduction

- Who has used R before ?

Introduction

- Who has used R before ?
- What other programming language have you used before ?

Introduction

- Who has used R before ?
- What other programming language have you used before ?
- For those who are experts in R

Introduction

- Who has used R before ?
- What other programming language have you used before ?
- For those who are experts in R
  - please refrain to answer during this session...
  - help your neighbor...

Introduction

- Who has used R before ?
- What other programming language have you used before ?
- For those who are experts in R
  - please refrain to answer during this session...
  - help your neighbor...
- Two special slide formatting

> Your turn...

Introduction

- Who has used R before ?
- What other programming language have you used before ?
- For those who are experts in R
  - please refrain to answer during this session...
  - help your neighbor...
- Two special slide formatting

Your turn...

Warning

Computer languages

Mother Tongues

Tracing the roots of computer languages through the ages

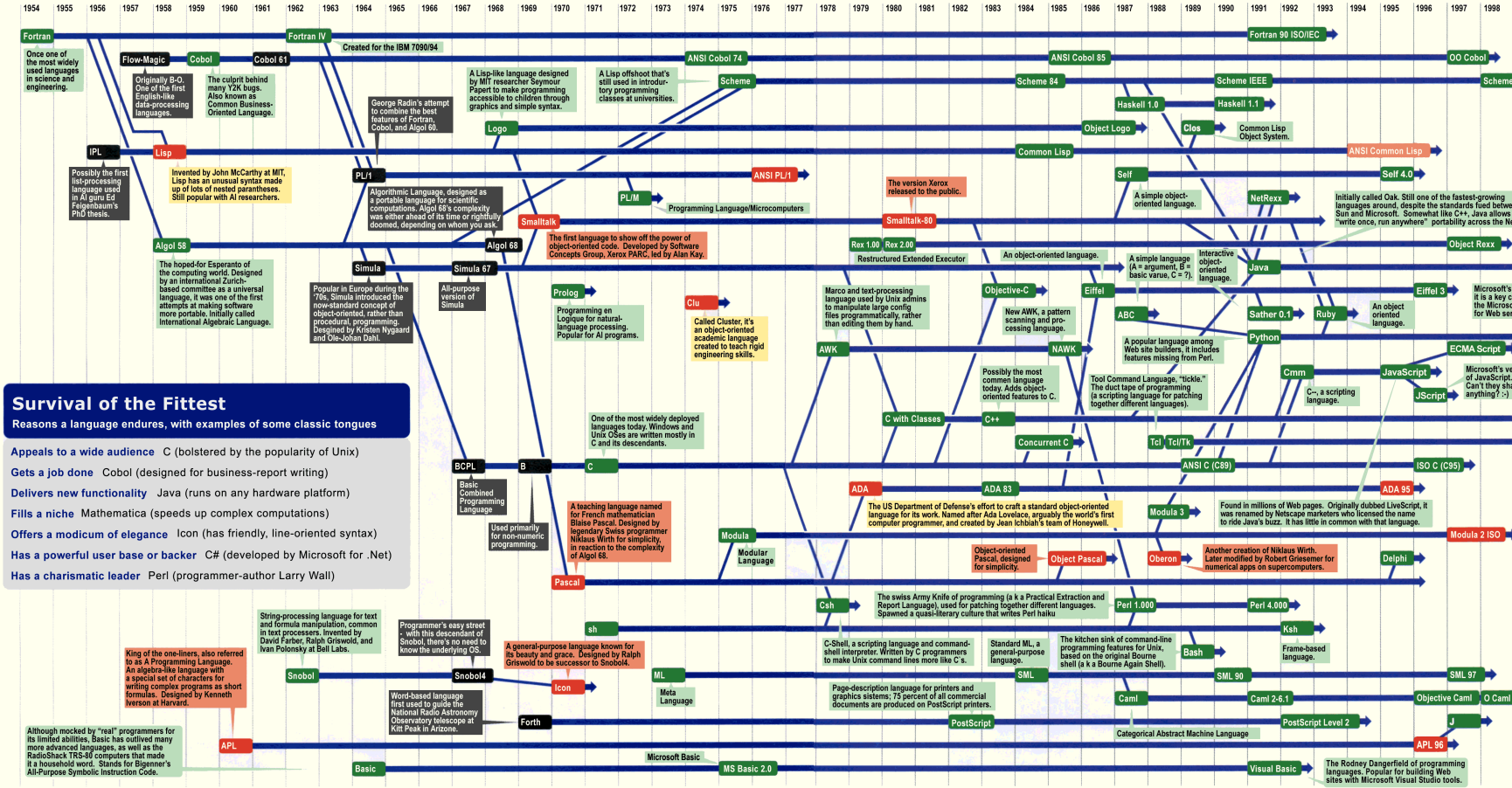
Just like half of the world's spoken tongues, most of the 2,300-plus computer programming languages are either endangered or extinct. As powerhouses C/C++, Visual Basic, Cobol, Java and other modern source codes dominate our systems, hundreds of older languages are running out of life.

An ad hoc collection of engineers-electronic lexicographers, if you will-aim to save, or at least document the lingo of classic software. They're combing the globe's 9 million developers in search of coders still fluent in these nearly forgotten lingua frangas. Among the most endangered are Ada, APL, B (the predecessor of C), Lsp, Oberon, Smalltalk, and Simula.

Code-raker Grady Booch, Rational Software's chief scientist, is working with the Computer History Musuem in Silicon Valley to record and, in some cases, maintain languages by writing new compilers so our ever-changing hardware can grok the code. Why bother? "They tell us about the state of software practice, the minds of their inventors, and the technical, social, and economic forces that shaped history at the time," Booch explains. "They'll provide the raw material for software archaeologists, historians, and developers to learn what worked, what was brilliant, and what was an utter failure." Here's a peek at the strongest branches of programming's family tree. For a nearly exhaustive rundown, check out the Language List at *HTTP://www.informatik.uni-freiburg.de/Java/misc/lang_list.html*. - Michael Mendeno

Key

- 1954 Year Introduced
- Active: thousands of users
- Protected: taught at universi available
- Endangered: usage droppin
- Extinct: no known active use compilers
- Lineage continues



Survival of the Fittest

Reasons a language endures, with examples of some classic tongues

- **Appeals to a wide audience** C (bolstered by the popularity of Unix)
- **Gets a job done** Cobol (designed for business-report writing)
- **Delivers new functionality** Java (runs on any hardware platform)
- **Fills a niche** Mathematica (speeds up complex computations)
- **Offers a modicum of elegance** Icon (has friendly, line-oriented syntax)
- **Has a powerful user base or backer** C# (developed by Microsoft for .Net)
- **Has a charismatic leader** Perl (programmer-author Larry Wall)

Sources: Paul Boutin; Brent Hailpern, associate director of computer science at IBM Research; The Retrocomputing Museum; Todd Proebsting, senior researcher at Microsoft; Gio Wiederhold, computer scientist, Stanford University

Introduction

History of R

- **Mid 1970s** - S Language for Statistical Computing conceived by John Chambers, Rick Becker, Trevor Hastie, Allan Wilks and others at Bell Labs
- **Early 1990's** - R was first implemented in the early 1990's by Robert Gentleman and Ross Ihaka, both faculty members at the University of Auckland.
- **1995** - Open Source Project
- **1997** - Managed by the R Core Group
- **2000** - First release of R
- **2011** - First release of R studio
- Historical notes - Paper from 1998

Introduction

Why use R ?

- **Script vs. Menu driven software (e.g. Excel)**
  - Can be re-rerun with new data
  - Reproducible workflow

Introduction

Why use R ?

- **Script vs. Menu driven software (e.g. Excel)**
  - Can be re-rerun with new data
  - Reproducible workflow
- **Open source**
  - Huge number of libraries
  - Tidy "universe" : tidyverse and ggplot2
    - Very easy to manipulate tables (select columns, create new variables)
    - High quality graphics

Introduction

Why use R ?

- **Script vs. Menu driven software (e.g. Excel)**
  - Can be re-rerun with new data
  - Reproducible workflow
- **Open source**
  - Huge number of libraries
  - Tidy "universe" : tidyverse and ggplot2
    - Very easy to manipulate tables (select columns, create new variables)
    - High quality graphics
- **Work environment**
  - R studio

Introduction

Why use R ?

- **Script vs. Menu driven software (e.g. Excel)**
  - Can be re-rerun with new data
  - Reproducible workflow
- **Open source**
  - Huge number of libraries
  - Tidy "universe" : tidyverse and ggplot2
    - Very easy to manipulate tables (select columns, create new variables)
    - High quality graphics
- **Work environment**
  - R studio
- **Document your data processing**
  - R markdown
  - Create HTML, pdf, presentations

Introduction

Why use R ?

- **Script vs. Menu driven software (e.g. Excel)**
  - Can be re-rerun with new data
  - Reproducible workflow
- **Open source**
  - Huge number of libraries
  - Tidy "universe" : tidyverse and ggplot2
    - Very easy to manipulate tables (select columns, create new variables)
    - High quality graphics
- **Work environment**
  - R studio
- **Document your data processing**
  - R markdown
  - Create HTML, pdf, presentations
- **Share your data and workflow**
  - GitHub

Introduction

What can you do with R ?

Introduction

What can you do with R ?

- **Science**
  - Statistics of course...
  - Data processing
  - Graphics
  - Time series analyses
  - Maps
  - Bioinformatics

Introduction

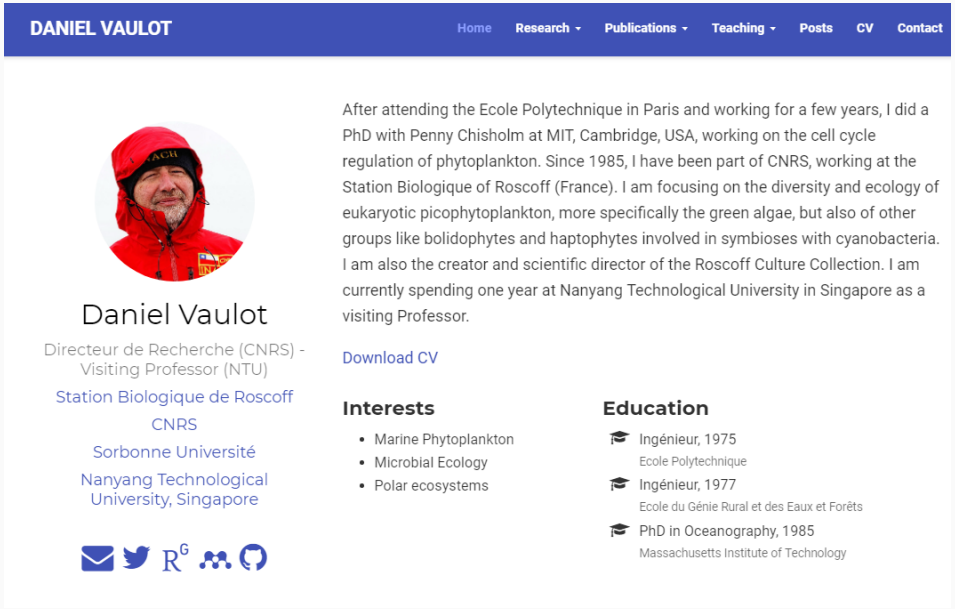
What can you do with R ?

- **Science**
  - Statistics of course...
  - Data processing
  - Graphics
  - Time series analyses
  - Maps
  - Bioinformatics
- **But also**
  - Teach
  - Do a presentation
  - Write your CV
  - Build a web site
  - Write a book
  - Much more...

Introduction

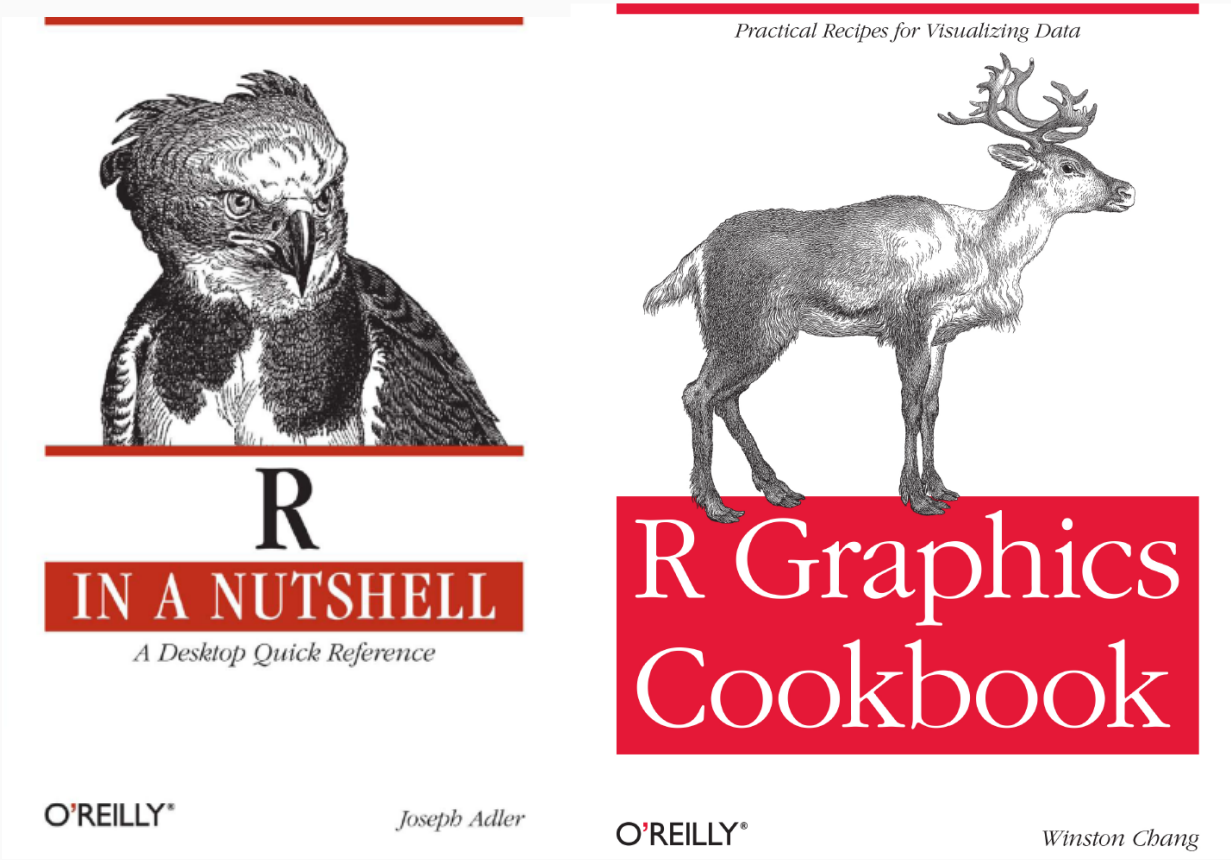
What can you do with R ?

- **Science**
  - Statistics of course...
  - Data processing
  - Graphics
  - Time series analyses
  - Maps
  - Bioinformatics
- **But also**
  - Teach
  - Do a presentation
  - Write your CV
  - Build a web site
  - Write a book
  - Much more...



Resources

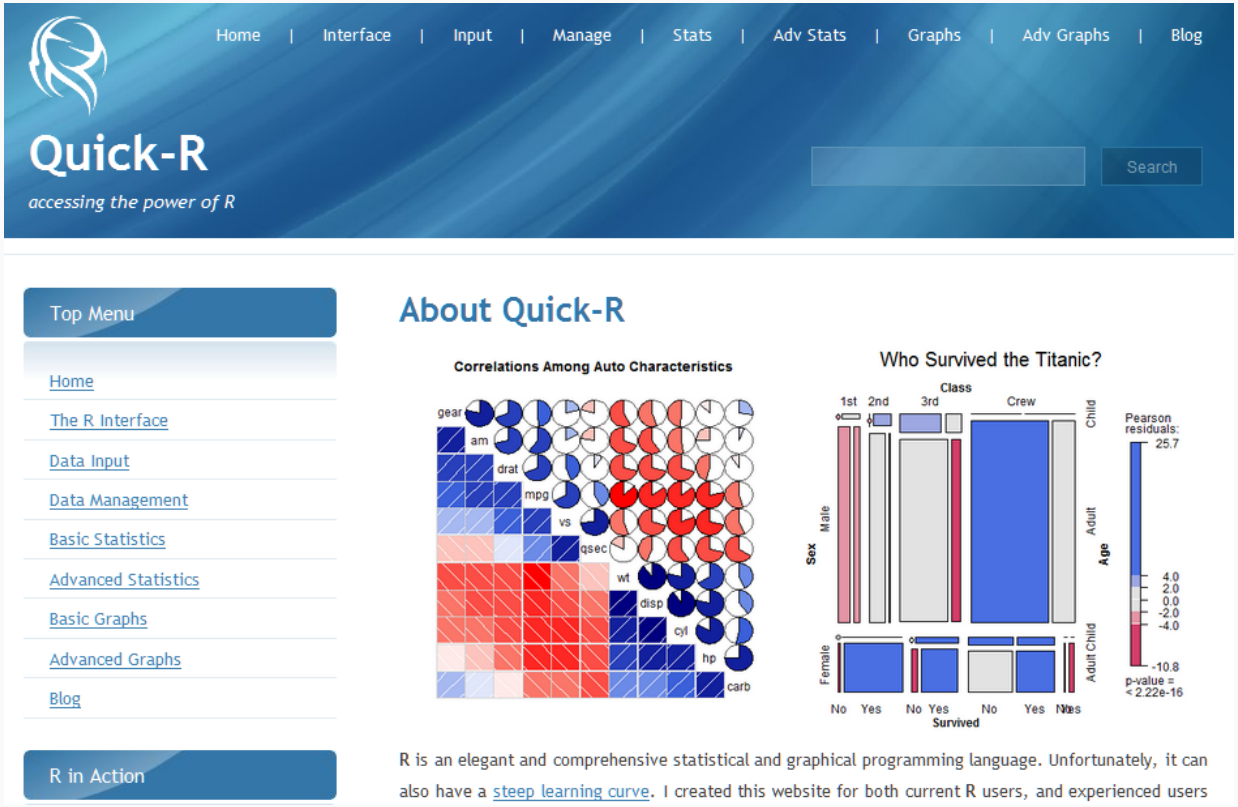
Books and Manuals



- R intro : Very good introduction to R, short and clear
- R in a nutshell : Many many receipes to solve all your questions
- R graphics cook book : very good for graphics

Resources

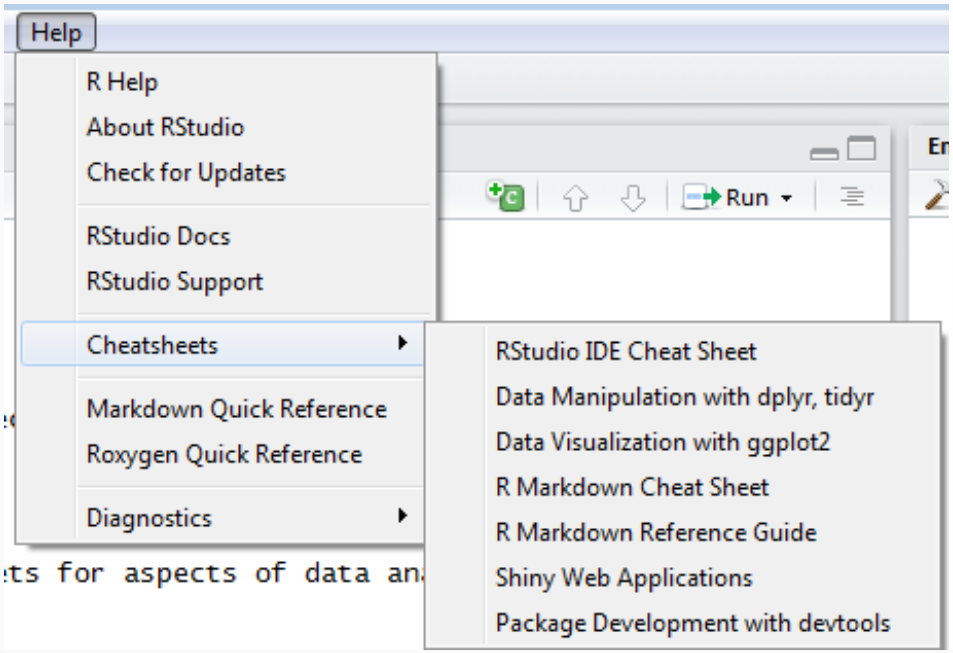
On line courses and web sites

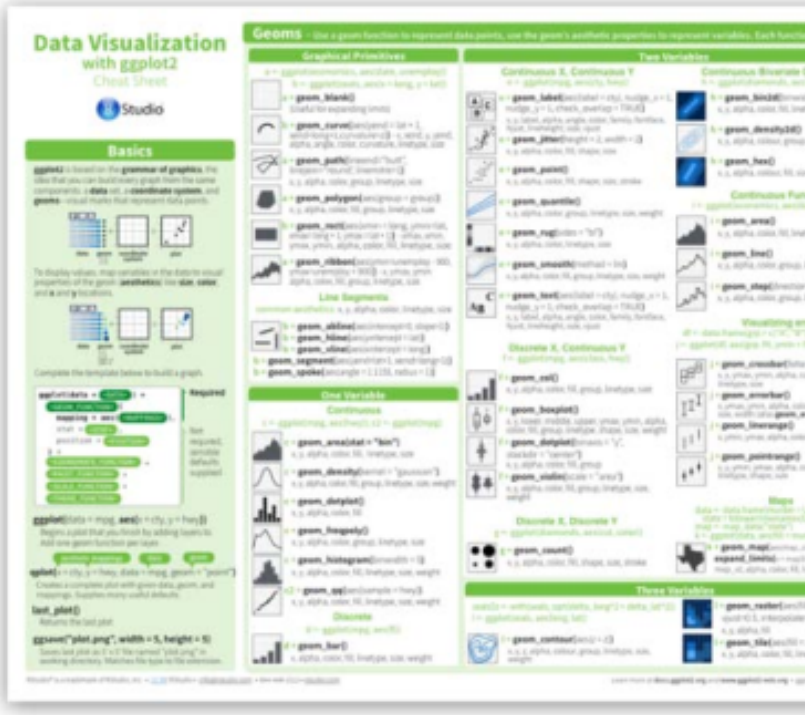


- Coursera
- Pluralsight
- Quick-R, very simple

Resources

Cheat sheets

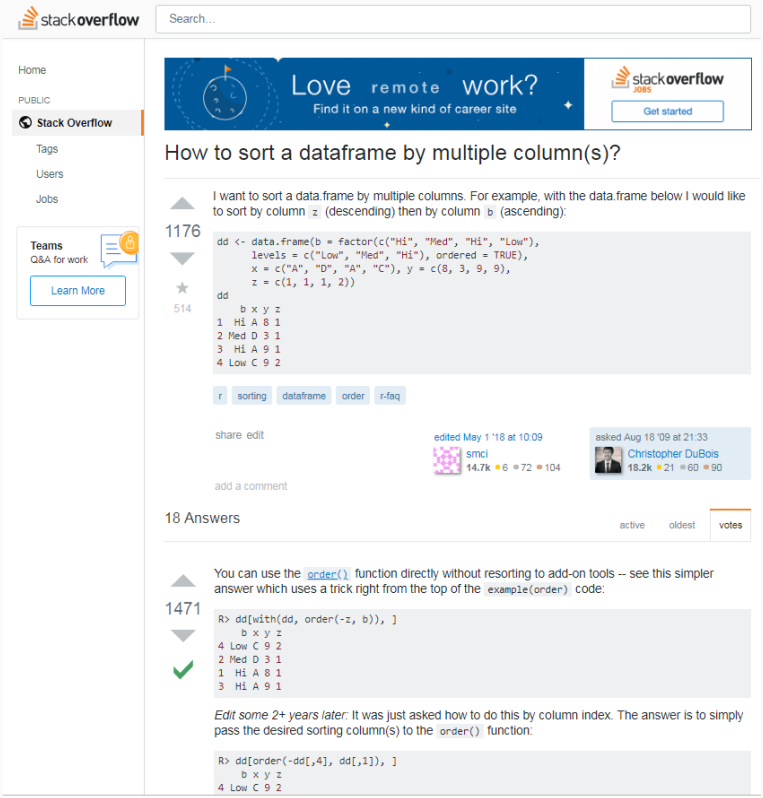




- R basics
- ggplot2
- dplyr

Resources

Forum

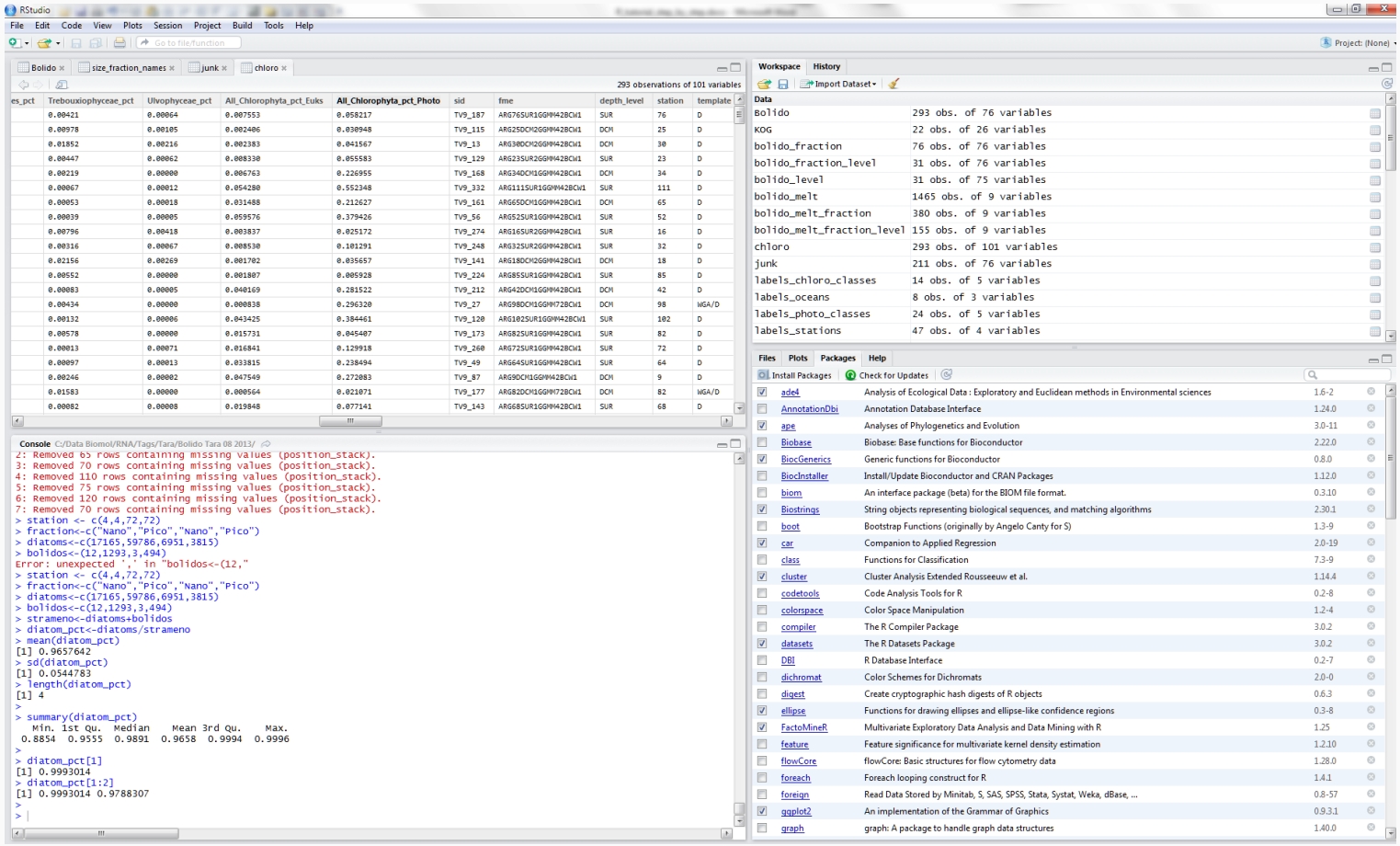


- https://stackoverflow.com/
- http://r-statistics.co/
- https://stats.stackexchange.com/
- https://www.r-bloggers.com/

Let's get started

Setup

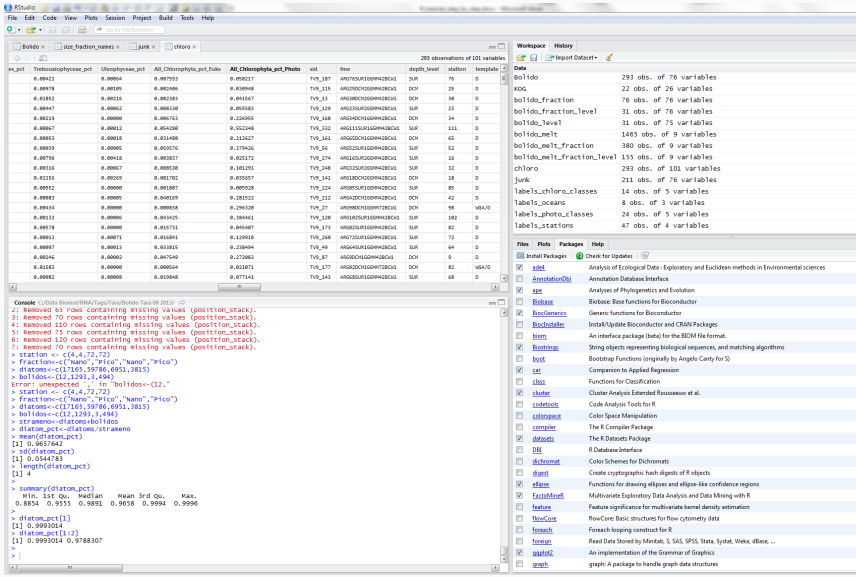
- Install R
- Install R studio



Let's get started

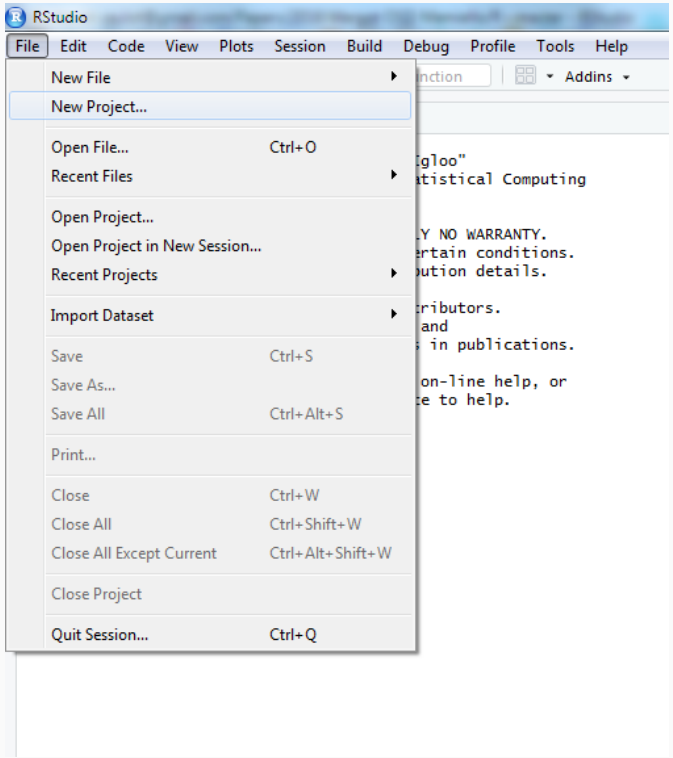
The R studio interface

- **Bottom left**
  - Console
- **Top left**
  - File editor
- **Top right**
  - Environment (i.e. R objects)
  - History
- **Bottom right**
  - Files
  - Plots
  - Packages
  - Help



Let's get started

Create a new project



- Open R studio
- Create new project for the course in a new directory
  - e.g. Data Science Class

Let's get started

Your first script

Two ways to proceed

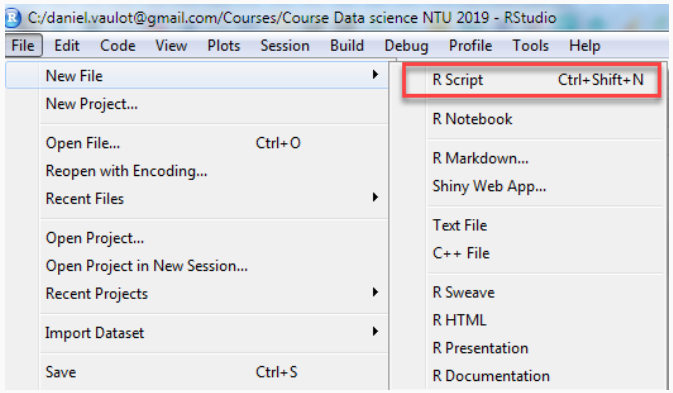
1. Type directly in command window

Let's get started

Your first script

Two ways to proceed

1. Type directly in command window
2. Create a new script



Type in script window, select and execute (CTRL-R)

The R language

Everything in R is an **object**

- Assignement done with **<-**

The R language

Everything in R is an **object**

- Assignement done with **<-**

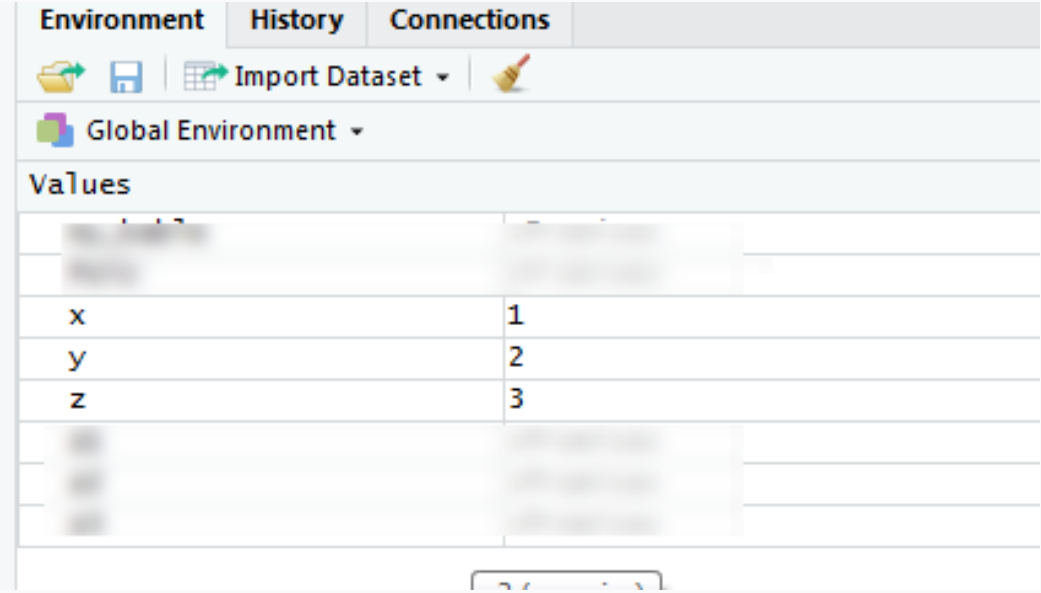
The R language

= can be used instead of **<-** but refrain from it (not good style)

The R language

= can be used instead of **<-** but refrain from it (not good style)

You can view the values of the objects in R-studio environment window (top-right)



The R language

R is **case sensitive**

The R language

R is **case sensitive**

Rules for naming objects

- Use
  - letters
  - numbers
  - the dot
  - the underscore (not the minus sign !)

- Start always with a letter
  - , , , are OK
  - , , are **not** OK

Use consistent naming

Five conventions

- alllowercase: e.g. adjustcolor
- period.separated: e.g. plot.new
- **underscore_separated**: e.g. numeric_version
- lowerCamelCase: e.g. addTaskCallback
- UpperCamelCase: e.g. SignatureMethod

Prefer third one, much more easy to read

- Use **names** for objects : **last_name**
- Use **verbs** for function : **build_name**
- Think about best order
  - e.g. prefer maybe **name_last** because then you can have name_first, name_full...
  - and you identify that all these objects are related to a name...

R objects

Data types

- **character**: "Daniel", "This is a course in R", 'Donald'
- **numeric**: 2, 15.5, 10e-3
- **integer**: 2L (the L tells R to store this as an integer)
- **date**: 2018-02-25
- **logical**: TRUE, FALSE
- **complex**: 1+4i (complex numbers with real and imaginary parts)

R objects

Data types

- **character**: "Daniel", "This is a course in R", 'Donald'
- **numeric**: 2, 15.5, 10e-3
- **integer**: 2L (the L tells R to store this as an integer)
- **date**: 2018-02-25
- **logical**: TRUE, FALSE
- **complex**: 1+4i (complex numbers with real and imaginary parts)
- **No data** "NA"
- **Not a number** "NaN" (e.g. division by zero)

R objects

Data structures

- **Vector**
- **List**
- **Matrix**
- **Data frames**
- **Function**

Vectors

The basic R structure is a vector:

Vectors

The basic R structure is a vector:

A vector can with a single element only

Vectors

The basic R structure is a vector:

A vector can with a single element only

Assign a value to a vector

Vectors

Assign several elements

Vectors

Assign several elements

Assign range

Vectors

Assign characters

Assign logical

Vectors

Access specific elements of a vector

First

Vectors

Access specific elements of a vector

First

Range

Vectors

Access specific elements of a vector

First

Range

Remove one element

Vectors

Determine object properties

Apply functions (we will come back to functions latter)

- **typeof()** - what is the object's data type (low-level)?
- **length()** - how long is it? What about two dimensional objects?

Vectors

Determine object properties

Apply functions (we will come back to functions latter)

- **typeof()** - what is the object's data type (low-level)?
- **length()** - how long is it? What about two dimensional objects?

Vectors

Determine object properties

Apply functions (we will come back to functions latter)

- **typeof()** - what is the object's data type (low-level)?
- **length()** - how long is it? What about two dimensional objects?

> What is the type and length of **PoTU** ?

Operators

Arithmetic Operators

Operator	Description
+	addition
-	subtraction
*	multiplication
/	division
^ or **	exponentiation
x %% y	modulus (x mod y) 5%%2 is 1
x %/% y	integer division 5%/%2 is 2

Operators

Arithmetic Operators

We are performing vector operations !

Operators

Arithmetic Operators

Vector one element

Operators

Arithmetic Operators

Vector several elements

Operators

Arithmetic Operators

Vector several elements

- Several instructions on same line separate by **;**
- The hastag **#** indicate a comment -> Use heavily to document your code

Operators

Arithmetic Operators

Vector several elements

- Several instructions on same line separate by **;**
- The hastag **#** indicate a comment -> Use heavily to document your code

Use the other operators

Operators

Arithmetic Operators

What happens when the vectors have different number of elements ?

Operators

Arithmetic Operators

What happens when the vectors have different number of elements ?

Operators

Arithmetic Operators

What happens when the vectors have different number of elements ?

Equivalent to

The recycling rule...

Operators

Can we add logical ?

Operators

Can we add logical ?

Operators

Can we add logical ?

It does not give an error but...

The resulting variable is transformed to a **numeric**

> How you would show that ?

Operators

Can we add logical ?

It does not give an error but...

The resulting variable is transformed to a **numeric**

> How you would show that ?

Operators

Logical Operators

Operator	Description
<	less than
<=	less than or equal to
>	greater than
>=	greater than or equal to
==	exactly equal to
!=	not equal to
!x	Not x
x | y	x OR y
x & y	x AND y
isTRUE(x)	test if X is TRUE

Operators

Logical Operators

Operators

Logical Operators

Do not mix

- == which is logical operator
- = which is assignement

Operators

Can we add characters ?

Operators

Can we add characters ?

Generates an error

Operators

Can we add characters ?

Generates an error

> What can we do ?

Functions

Function perform specific task on objects

- e.g. to concatanate strings we use **paste0()**

Functions

Function perform specific task on objects

- e.g. to concatanate strings we use **paste0()**

Functions

Function perform specific task on objects

- e.g. to concatanate strings we use **paste0()**

- Functions take **arguments** and return an object called **result**
- To know the arguments use ?

Functions

Function perform specific task on objects

- e.g. to concatanate strings we use **paste0()**

- Functions take **arguments** and return an object called **result**
- To know the arguments use ?

What happened ?

Functions

Function perform specific task on objects

- e.g. to concatanate strings we use **paste0()**

- Functions take **arguments** and return an object called **result**
- To know the arguments use ?

> What happened ?

- Can go directly to Help panel and type function name

Functions

Help

paste {base} R Documentation

Concatenate Strings

Description

Concatenate vectors after converting to character.

Usage

```
paste (..., sep = " ", collapse = NULL)
paste0(..., collapse = NULL)
```

Arguments

`...`	one or more `R` objects, to be converted to character vectors.
`sep`	a character string to separate the terms. Not `NA_character_`.
`collapse`	an optional character string to separate the results. Not `NA_character_`.

Details

`paste` converts its arguments (*via* `as.character`) to character strings, and concatenates them (separating them by the string given by `sep`). If the arguments are vectors, they are concatenated term-by-term to give a character vector result. Vector arguments are recycled as needed, with zero-length arguments being recycled to `""`.

Note that `paste()` coerces `NA_character_`, the character missing value, to `"NA"` which may seem undesirable, e.g., when pasting two character vectors, or very desirable, e.g. in `paste("the value of p is ", p)`.

`paste0(..., collapse)` is equivalent to `paste(..., sep = "", collapse)`, slightly more efficiently.

If a value is specified for `collapse`, the values in the result are then concatenated into a single string, with the elements being separated by the value of `collapse`.

Functions

Help

Examples

```
## When passing a single vector, paste0 and paste work like as.character.
paste0(1:12)
paste(1:12)         # same
as.character(1:12) # same

## If you pass several vectors to paste0, they are concatenated in a
## vectorized way.
(nth <- paste0(1:12, c("st", "nd", "rd", rep("th", 9))))

## paste works the same, but separates each input with a space.
## Notice that the recycling rules make every input as long as the longest input.
paste(month.abb, "is the", nth, "month of the year.")
paste(month.abb, letters)

## You can change the separator by passing a sep argument
## which can be multiple characters.
paste(month.abb, "is the", nth, "month of the year.", sep = "_*_")

## To collapse the output into a single string, pass a collapse argument.
paste0(nth, collapse = ", ")

## For inputs of length 1, use the sep argument rather than collapse
paste("1st", "2nd", "3rd", collapse = ", ") # probably not what you wanted
paste("1st", "2nd", "3rd", sep = ", ")

## You can combine the sep and collapse arguments together.
paste(month.abb, nth, sep = ": ", collapse = "; ")

## Using paste() in combination with strwrap() can be useful
## for dealing with long strings.
(title <- paste(strwrap(
    "Stopping distance of cars (ft) vs. speed (mph) from Ezekiel (1930)",
    width = 30), collapse = "\n"))
plot(dist ~ speed, cars, main = title)
```

Functions

Getting what you want

We would like to write "Donald Trump" but we have :

> Can you read the help and suggest a change in the way we call the function ?

Functions

Getting what you want

We would like to write "Donald Trump" but we have :

Can you read the help and suggest a change in the way we call the function ?

Functions

Write your own function

Functions

Write your own function

> If you write 3 times the same piece of code write a function...

Functions

Examples of functions

Most of the time you do not have to write functions because someone has already written one for what you want to do...

- Sum

Functions

Examples of functions

Most of the time you do not have to write functions because someone has already written one for what you want to do...

- Sum

- Normal distribution

Functions

Statistics

Functions

Statistics

- Is the mean close to expected mean ?
- What can be done ?

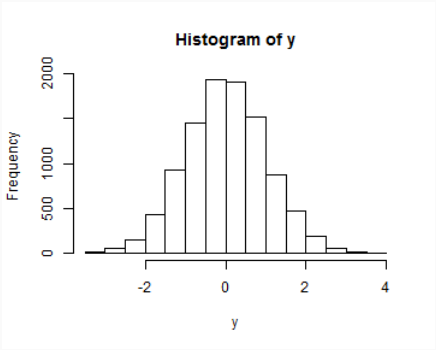
Functions

Sample more points... 10,000 instead of 100

Functions

Plot

Histogram

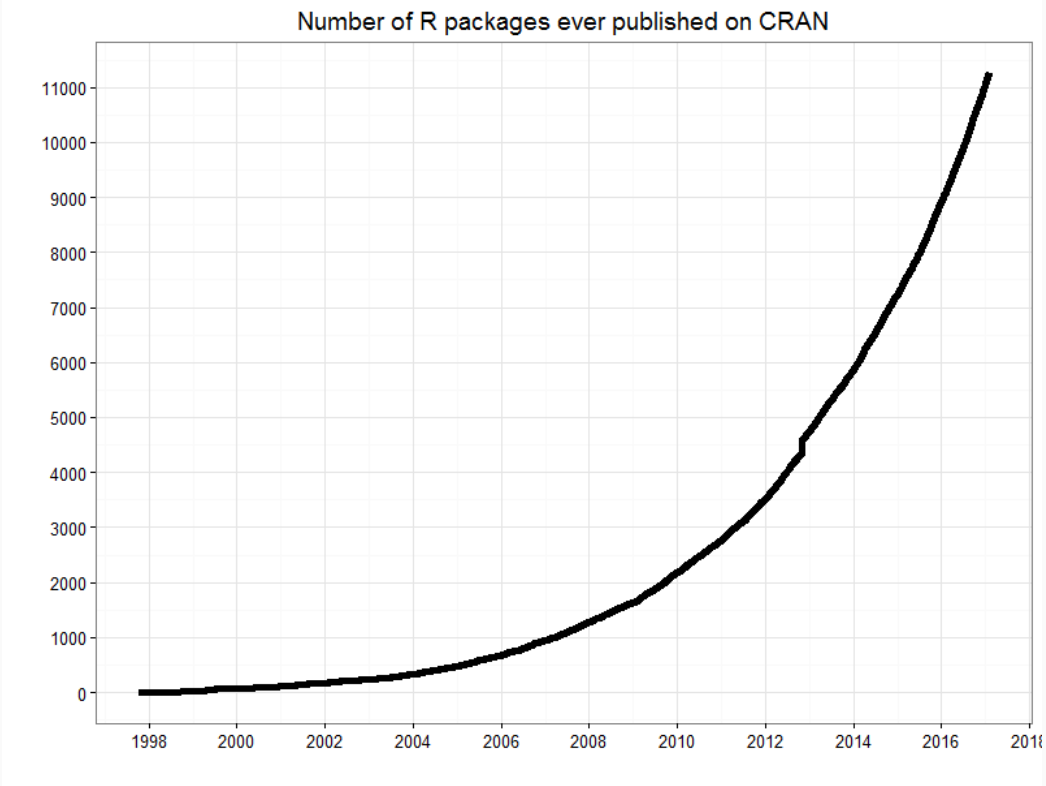


- What is this "library()"

Packages

Packages are set of functions that have a common goal

They are really the strength of R

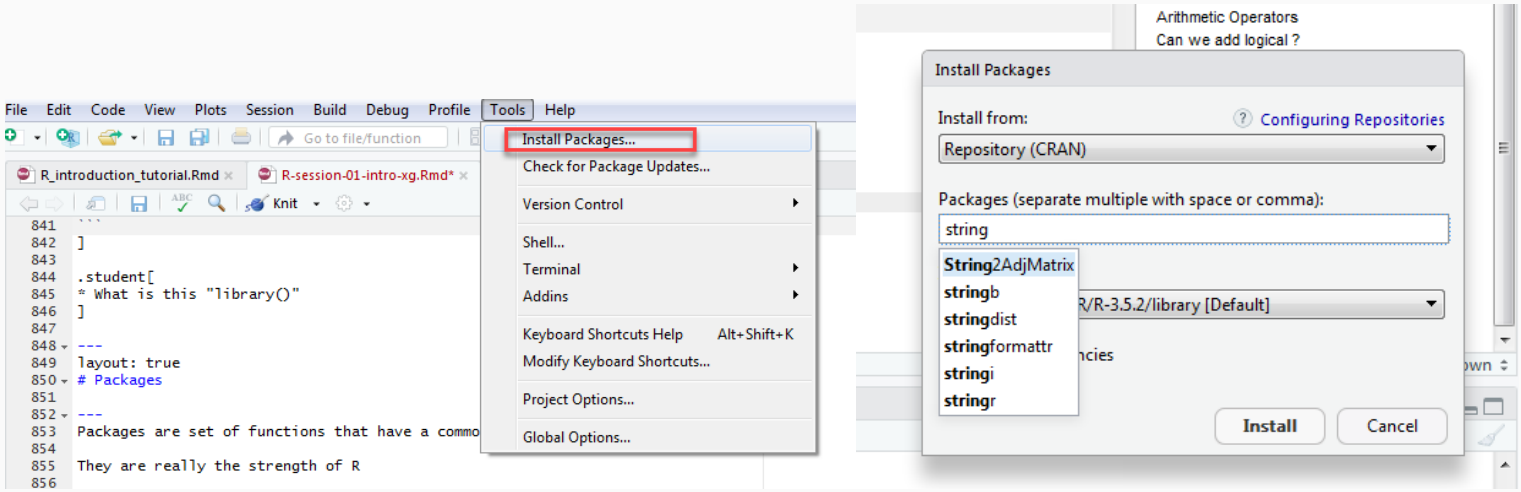


And these are only the "official"" packages. You can find more on GitHub

Packages

Installing a package

Download on your computer the package you need



> Install package **stringr** (to manipulate strings of characters)

Packages

Using a package

To use functions from the package

- use the syntax `                    `

Packages

Using a package

To use functions from the package

- use the syntax `package_name::function_name()`

```
```

- load the package with the library function

```
```

Packages

Using a package

To use functions from the package

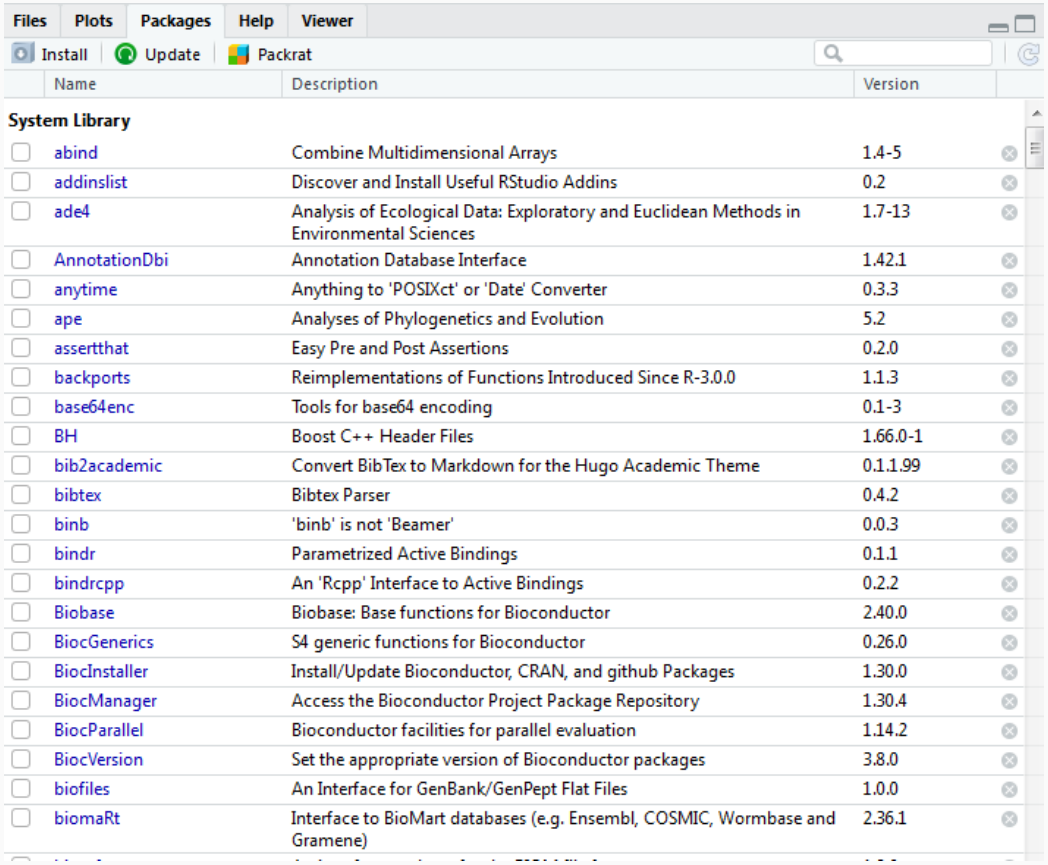
- use the syntax `` `

- load the package with the library function

Sometimes functions from different libraries have similar names

Packages

List installed packages



Other objects

- List
- Matrix
- Factors
- **Data frames**

Data frames

What is it ?

- Table mixing different types of columns (an Excel table...)
- However within a column all values are similar, e.g. numeric, logical, character

Data frames

What is it ?

- Table mixing different types of columns (an Excel table...)
- However within a column all values are similar, e.g. numeric, logical, character

- We will not get into the factor at this time, why we set
- Notice the recycling rule ?

Data frames

Useful functions

Data frames

Useful functions

Data frames

Access specific column

- Use `$` notation

Data frames

Access specific column

- Use `[]` notation

```
```

- Use the `$` notation

```
```

```
```

Data frames

Access specific value

- Use `[]` notation

Data frames

Access specific value

- Use `[]` notation

- Use the `$` notation

Data frames

Filter the data

Select lines for which the label is c

Data frames

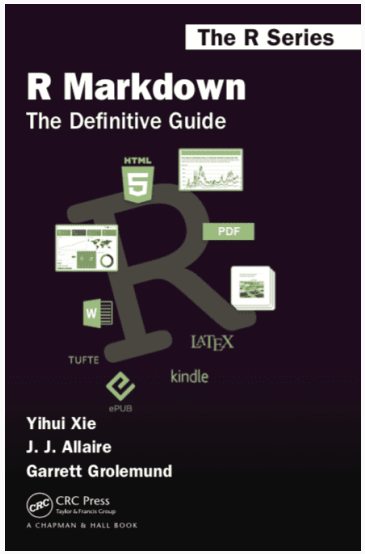
Filter the data

Select lines for which the label is c

Next time: Markdown

What you will learn :

- Mix text, R code and R output in a single document
- Produce documents as HTML, pdf or even Word from the same template



- Please install the following packages and their dependencies
  - rmarkdown (will install also knitr)
  - tinytex (Latex)
- Read the installation instruction : https://bookdown.org/yihui/rmarkdown/installation.html
- Have a look at the book https://bookdown.org/yihui/rmarkdown