

Rbasics

PhD toolbox - 40th PhD cycle



Part IV - Tips and Tricks

Summary

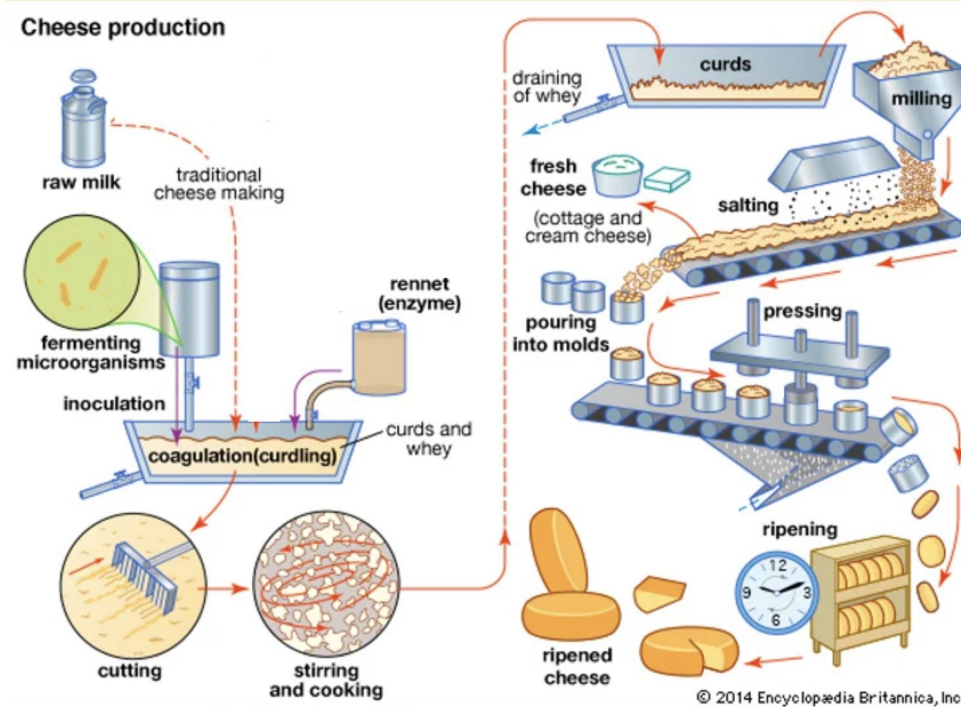
- Organize your working directory and your script
- Manage files & folders
- Tips on Warning/errors
- R documentation and manuals

Practical tips for organizing scripts, files and folders

- working directory concept
 - `setwd("path/to/folder/")`
- a working directory for each research project/experiment/analysis ... I suggest to avoid use the default R working directory. After a bit you will have an incomprehensible multitude of input files, outputs, from different projects, analysis, errors and reconsiderations ... trust me I already lived it XD.
- long analysis need order (teutonic strengthness!) it is easy to create bugs in mis-organized scripts.
 - `script "chapterization" using`
`# 1. Chapter 1 -----`
`## 1.1 Sub-chapter -----`
`or`
`#### 1. Chapter 1 ####`
`##### 1.1 Sub-chapter #####`

this will create tracking points in your script

practical tips for organizing scripts, files and folders



- long analysis needs order (tip2). Long scripts (I mean > 500 rows) are good for braggarts: several scripts generating intermediate results are better.

- only you know where to stop! my suggestion (actually is an example)

script 1: data loading + basic transformations + preliminary data surveys

script 2: main analysis with transformed data

script 3: cool figures

Your project folder

README file

explains the content of the project and the folders/files contained in the project's main directory

RAW data folder

contains the untouched raw data. Ideally should have read-only permissions.

They must be stored in a very safe place, and possibly have a recovery backup.

Have you made them available after publication? This prevents to have your materials lost and is increasingly required by journals

Scripts folder

The tidy collection of all your script used in the several steps of data analysis

Analysis (or Results) folder

Contains the outputs of your analyses (elaborated tables, figures, etc..)

practical tips for organizing scripts, files and folders

- R scripts should be published to address reproducible research requirements. Don't do the assholes, sharing is good for scientific community!
- You can add a nice header to your scripts including:
 - Paper title
 - Script author(s)
 - A brief explanation on its usage, if it is needed
 - R and Rstudio version you used
 - List of packages and their version

practical tips for organizing scripts, files and folders

```
# Herbonaute
# Adamo M., Marmesse R.
# R script by M. Adamo

# R version 4.2.2 Patched (2022-11-10 r83330) -- "Innocent and Trusting"
# Copyright (C) 2022 The R Foundation for Statistical Computing
# Platform: x86_64-pc-linux-gnu (64-bit)

# used libraries -----
library("raster")          # Geographic data analysis and modeling v3.5-29
library("ggbiplot")        # A ggplot2 based biplot v0.55
library("factoextra")      # Multivariate data analysis v1.0.7
...
```

Repositories

Some good places where to deposit your data (raw, elaborated or scripts)



DRYAD



panic! - warnings vs errors -

Warnings: Your function was able to run but there are some conditions that needs to be checked to avoid possible issues

```
> matrix(data = sample(seq(1,30)), nrow = 10, ncol = 5, byrow = F)
```

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	3	19	9	3	19
[2,]	7	27	25	7	27
[3,]	13	1	22	13	1
[4,]	17	16	6	17	16

Warning message:

```
In matrix(data = sample(seq(1, 30)), nrow = 10, ncol = 5, byrow = F) :  
  data length differs from size of matrix: [30 != 10 x 5]
```

panic! - warnings vs errors -

Errors: Your command cannot be executed

```
> library(unicorns)
```

```
Error in library(unicorns) : there is no package called 'unicorns'
```


Several reasons behind:

- there is a typo in your command
- the function you typed does not exist or the library containing it has not be loaded
- the object name you typed has some typos or does not still exist
- your command does not comply with the R language syntax
 - e.g. there are unclosed brackets/quotation marks

panic! - warnings vs errors

Some useful tips:

- RStudio most of the times tells you that something is wrong with the syntax

A screenshot of the RStudio code editor. The toolbar at the top includes navigation arrows, a copy icon, a save icon, a checkbox labeled 'Source on Save', a search icon, a help icon, and a notebook icon. The code editor shows a script with line numbers 1 through 6. Line 1 contains 'library(ggplot2)'. Line 2 is empty. Line 3 contains 'data<-read.table("my_table.tsv, header=T,)' and has a red 'x' icon to its left, indicating a syntax error. Line 4 is empty. Line 5 is empty. Line 6 is empty.

```
1 library(ggplot2)
2
3 data<-read.table("my_table.tsv, header=T, )
4
5
6
```

- sometimes function names overlap between different packages.

You need to get aware of that and force the function from the right package:

```
dplyr::filter
```

- functions can be updated over the time and their usage can have changed from the last time you used it. No way back: the best solution is to integrate the new function in the script.

panic! - Other strategies

- bug fixing
 - it is usually possible when problems are simple and/or you're experienced
 - check the presence of NAs or other lacks in the data
 - check the format of your input data
- manuals
 - almost all packages have the help page
 - some package have dedicated web-pages and online tutorials. *e.g.*:
<http://www.ggtern.com/docs/>

panic! - Other strategies

- forums
 - "Google knows!"
 - you will probably ground in one of the several R-specialized forums
- AI chat
 - ChatGPT is able to create and explain good parts of R code ... Ask it! ...
Otherwise you must know R to understand if the reply is good or not

I get an error message that I don't understand:

- Googling the error message
- Check Stack Overflow forum pages
- Ask your skilled colleagues (do not abuse of this solution)
- Ask on R-help mailing list

Note: Most of the time some other else already encountered your same issue, or a bug is already known (and not fixed, sure!)

How to make easier to help you

- Use the **correct terms** to describe your problem
- Try to **generalize** what you are doing
- Include the output of **sessionInfo()** in your requests
- Most of the forums allow user to add **example code sections** that can be easily handled by other participants.
- Try to share a **reproducible example**, including your input data!


see here for further ideas: <http://adv-r.had.co.nz/Reproducibility.html>

Cheatsheets


Data Wrangling

with dplyr and tidyr

Cheat Sheet




Tidy Data - A foundation for wrangling in R




In a tidy data set:

Each **variable** is saved in its own **column**



Each **observation** is saved in its own **row**

Tidy data complements R's **vectorized operations**. R will automatically preserve observations as you manipulate variables. No other format works as intuitively with R.



Syntax - Helpful conventions for wrangling

dplyr::tbl_dfr(iris)

Converts data to tbl class. tbl's are easier to examine than data frames. R displays only the data that fits onscreen:

```
Source: local data frame [50 x 5]
  Sepal.Length Sepal.Width Petal.Length
1      5.1         3.5         1.4
2      4.9         3.0         1.4
3      5.4         3.7         1.5
4      4.3         3.1         1.0
5      5.0         3.6         1.4
...
# Variables not shown: Petal.Width (dbl), Species (fctr)
```


dplyr::glimpse(iris)

Information dense summary of tbl data.

utils::View(iris)


View data set in spreadsheet-like display (note capital V).

Reshaping Data - Change the layout of a data set




tidyr::gather(cases, "year", "n", 2:4)

Gather columns into rows.




tidyr::spread(pollution, size, amount)

Spread rows into columns.



tidyr::separate(storms, date, c("y", "m", "d"))


Separate one column into several.



tidyr::unite(data, col, ..., sep)

Unite several columns into one.

Subset Observations (Rows)



dplyr::filter(iris, Sepal.Length > 7)

Extract rows that meet logical criteria.

dplyr::distinct(iris)

Remove duplicate rows.

dplyr::sample_frac(iris, 0.5, replace = TRUE)

Randomly select fraction of rows.

dplyr::sample_n(iris, 10, replace = TRUE)

Randomly select n rows.


dplyr::slice(iris, 10:15)

Select rows by position.

dplyr::top_n(storms, 2, date)

Select and order top n entries (by group if grouped data).

Subset Variables (Columns)



dplyr::select(iris, Sepal.Width, Petal.Length, Species)

Select columns by name or helper function.

Helper Functions for select - select

selections, contains("x")

Select columns whose name contains a character string

selections, ends_with("length")

Select columns whose name ends with a character string

selections, everything()

Select every column

selections, matches("x")

Select columns whose name matches a regular expression.

selections, num_range("x", 1:5)

Select columns named x1, x2, x3, x4, x5

selection, one_of("Species", "Species1")

Select columns whose names are in a group of names.

selections, starts_with("Sepal")

Select columns whose name starts with a character string

selections, Sepal.Length:Petal.Width()

Select all columns between Sepal.Length and Petal.Width (inclusive)

selections, !Species()

Select all columns except Species

Logic in R - Comparison, Boolean Logic

	Less than	Equal	Not equal to
<	Less than	Equal	Not equal to
>	Greater than	Equal	Not equal to
<=	Less than or equal to	Equal	Not equal to
>=	Greater than or equal to	Equal	Not equal to

Logic in R - Comparison, Boolean Logic

	AND	OR	NOT
&	AND	OR	NOT
	AND	OR	NOT
!	AND	OR	NOT

Logic in R - Comparison, Boolean Logic

iris %>% group_by(Species) %>% summarise(avg = mean(Sepal.Width)) %>% arrange(avg)

Logic in R - Comparison, Boolean Logic

iris %>% group_by(Species) %>% summarise(avg = mean(Sepal.Width)) %>% arrange(avg)

Logic in R - Comparison, Boolean Logic

iris %>% group_by(Species) %>% summarise(avg = mean(Sepal.Width)) %>% arrange(avg)

Logic in R - Comparison, Boolean Logic

iris %>% group_by(Species) %>% summarise(avg = mean(Sepal.Width)) %>% arrange(avg)

Logic in R - Comparison, Boolean Logic

iris %>% group_by(Species) %>% summarise(avg = mean(Sepal.Width)) %>% arrange(avg)

Logic in R - Comparison, Boolean Logic

iris %>% group_by(Species) %>% summarise(avg = mean(Sepal.Width)) %>% arrange(avg)

Logic in R - Comparison, Boolean Logic

iris %>% group_by(Species) %>% summarise(avg = mean(Sepal.Width)) %>% arrange(avg)

Logic in R - Comparison, Boolean Logic

iris %>% group_by(Species) %>% summarise(avg = mean(Sepal.Width)) %>% arrange(avg)

Logic in R - Comparison, Boolean Logic

iris %>% group_by(Species) %>% summarise(avg = mean(Sepal.Width)) %>% arrange(avg)

Logic in R - Comparison, Boolean Logic

iris %>% group_by(Species) %>% summarise(avg = mean(Sepal.Width)) %>% arrange(avg)

Logic in R - Comparison, Boolean Logic

iris %>% group_by(Species) %>% summarise(avg = mean(Sepal.Width)) %>% arrange(avg)

Logic in R - Comparison, Boolean Logic

iris %>% group_by(Species) %>% summarise(avg =

You cannot know in detail
all the function of a given
package:

Cheatsheets ("bigliettino") help you in summarizing the main functions and their usage!