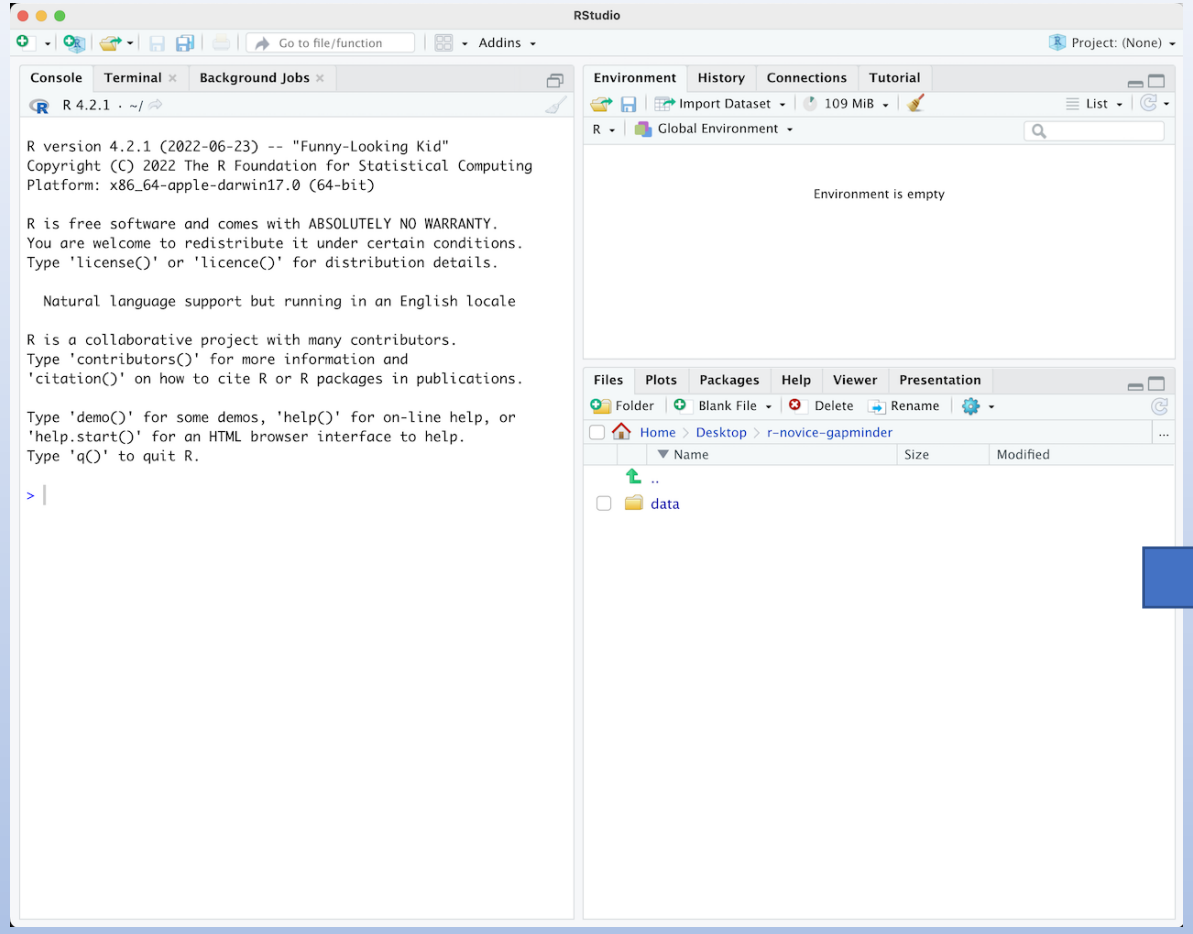
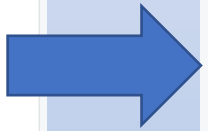
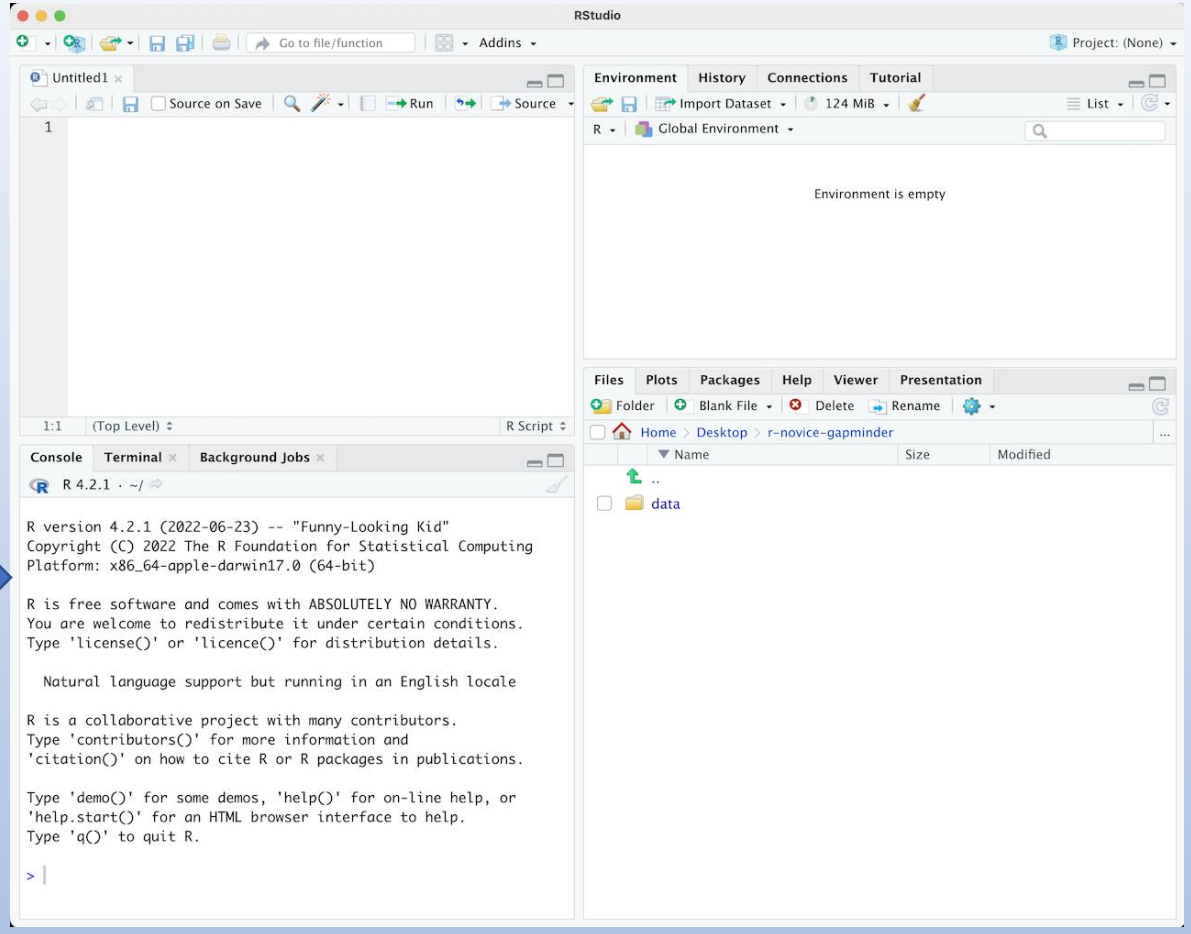
R for Reproducible Scientific Analysis

University of Oslo Library
Apr 24, 2023
9:00 am - 4:00 pm CET
Instructors: Mohamed Abdelhalim







There are two main ways one can work within RStudio:

1. Test and play within the interactive R console then copy code into a .R file to run later.
    - This works well when doing small tests and initially starting off.
    - It quickly becomes laborious
2. Start writing in a .R file and use RStudio's short cut keys for the Run command to push the current line, selected lines or modified lines to the interactive R console.
    - This is a great way to start; all your code is saved for later
    - You will be able to run the file you create from within RStudio or using R's `source()` function.

Challenge 1

Which of the following are valid R variable names?

R

```
min_height
max.height
_age
.mass
MaxLength
min-length
2widths
celsius2kelvin
```

Challenge 5

Install the following packages: `ggplot2`, `plyr`, `gapminder`

Best practices for project organization

Although there is no “best” way to lay out a project, there are some general principles to adhere to that will make project management easier:

Treat data as read only

This is probably the most important goal of setting up a project. Data is typically time consuming and/or expensive to collect. Working with them interactively (e.g., in Excel) where they can be modified means you are never sure of where the data came from, or how it has been modified since collection. It is therefore a good idea to treat your data as “read-only”.

Data Cleaning

In many cases your data will be “dirty”: it will need significant preprocessing to get into a format R (or any other programming language) will find useful. This task is sometimes called “data munging”. Storing these scripts in a separate folder, and creating a second “read-only” data folder to hold the “cleaned” data sets can prevent confusion between the two sets.

Treat generated output as disposable

Anything generated by your scripts should be treated as disposable: it should all be able to be regenerated from your scripts.

There are lots of different ways to manage this output. Having an output folder with different sub-directories for each separate analysis makes it easier later. Since many analyses are exploratory and don’t end up being used in the final project, and some of the analyses get shared between projects.

📌 Tip: Good Enough Practices for Scientific Computing

Good Enough Practices for Scientific Computing gives the following recommendations for project organization:

1. Put each project in its own directory, which is named after the project.
2. Put text documents associated with the project in the `doc` directory.
3. Put raw data and metadata in the `data` directory, and files generated during cleanup and analysis in a `results` directory.
4. Put source for the project’s scripts and programs in the `src` directory, and programs brought in from elsewhere or compiled locally in the `bin` directory.
5. Name all files to reflect their content or function.

Good enough practices in scientific computing

Greg Wilson, Jennifer Bryan, Karen Cranston, Justin Kitzes, Lex Nederbragt, Tracy K. Teal

Published: June 22, 2017 • https://doi.org/10.1371/journal.pcbi.1005510

Article | Authors | Metrics | Comments | Media Coverage

Author summary

Computers are now essential in all branches of science, but most researchers are never taught the equivalent of basic lab skills for research computing. As a result, data can get lost, analyses can take much longer than necessary, and researchers are limited in how effectively they can work with software and data. Computing workflows need to follow the same practices as lab projects and notebooks, with organized data, documented steps, and the project structured for reproducibility, but researchers new to computing often don't know where to start. This paper presents a set of good computing practices that every researcher can adopt, regardless of their current level of computational skill. These practices, which encompass data management, programming, collaborating with colleagues, organizing projects, tracking work, and writing manuscripts, are drawn from a wide variety of published sources from our daily lives and from our work with volunteer organizations that have delivered workshops to over 11,000 people since 2010.

PLOS Computational Biology 13(6): e1005510.

Challenge 3

Download the gapminder data from here.

1. Download the file (right mouse click on the link above -> “Save link as” / “Save file as”, or click on the link and after the page loads, press Ctrl + S or choose File -> “Save page as”)
2. Make sure it’s saved under the name `gapminder_data.csv`
3. Save the file in the `data/` folder within your project.

We will load and inspect these data later.

https://raw.githubusercontent.com/swcarpentry/r-novice-gapminder/gh-pages/_episodes_rmd/data/gapminder_data.csv

How do I ask a good question?

We're happy to help you, but in order to **improve your chances** of getting an answer, here are some **guidelines to follow**:

Make sure your question is on-topic and suitable for this site

Stack Overflow only accepts certain types of questions about programming and software development, and your question must be written in English. If your question is not on-topic or is otherwise unsuitable for this site, then it will likely be closed.

Closure is not the end of the road for questions; it is intended to be a temporary state until the question is revised to meet our requirements. However, if you fail to do that, or it is impossible to do so, then the question will stay closed and will not be answered.

Since you're reading this page, hopefully you will post a suitable, on-topic question from the outset, thus eliminating the need for the closure and reopening process!

Search, and research

Before posting a question, we strongly recommend that you spend a reasonable amount of time researching the problem and searching for existing questions on this site that may provide an answer. (Stack Overflow has been around for a long time now, so many common questions have already been answered.)

Make sure to keep track of what you find when researching, even if it doesn't help! If you ultimately aren't able to find the answer to your question elsewhere on this site, then including links to related questions (as well as an explanation of why they didn't help in your specific case) will help prevent your question from being marked as a duplicate when you ultimately do ask it.

Write a title that summarizes the specific problem

The title is the first thing that potential answerers will see. If your title isn't interesting, they won't read the rest. Also, without a good title, people may not even be able to find your question. So, *make the title count*:

- **Pretend you're talking to a busy colleague** and have to sum up your entire question in one sentence:

https://stackoverflow.com/help/how-to-ask

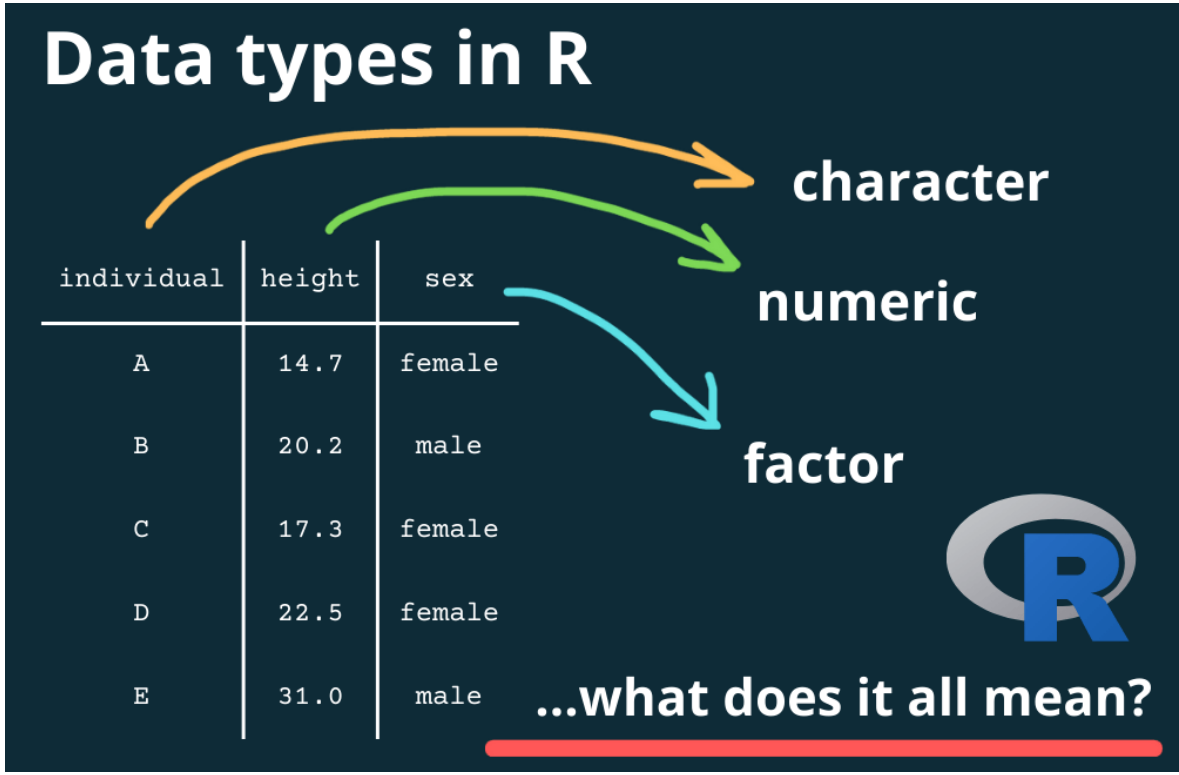
Keyboard Shortcuts

- Insert the <- assignment operator with `Option + -` on a Mac, or `Alt + -` on Linux and Windows.
- Insert the pipe operator `%>%` with `Command + Shift + M` on a Mac, or `Ctrl + Shift + M` on Linux and Windows.
- Run the current line of code with `Command + Enter` on a Mac or `Control + Enter` on Linux and Windows.
- Run all lines of code with `Command + A + Enter` on a Mac or `Control + A + Enter` on Linux and Windows.
- Restart the current R session and start fresh with `Command + Shift + F10` on a Mac or `Control + Shift + F10` on Linux and Windows.
- Comment or uncomment lines with `Command + Shift + C` on a Mac or `Control + Shift + C` on Linux and Windows.
- Trying to remember a command you submitted earlier? Search the command history from the Console with `Command + [up arrow]` on a Mac or `Control + [up arrow]` on Linux and Windows.

- To access shortcuts, type Option + Shift + K on a Mac, or Alt + Shift + K on Linux and Windows

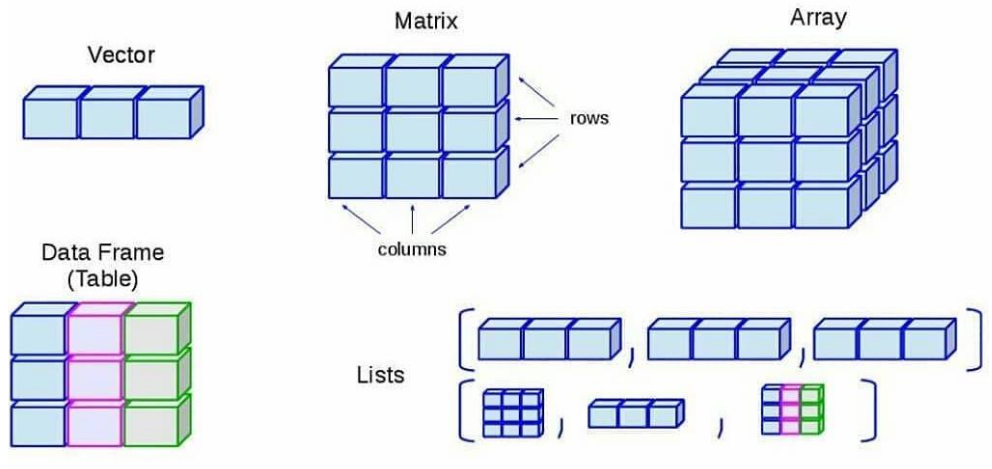
Data Types in R

Numeric	Contains decimal as well as whole numbers
Integer	Contains only whole numbers
Character	Holds character strings
Factor	A vector that can contain only predefined values, and is used to store categorical data.
Logical	Can only take on two values, TRUE or FALSE.



Data Structures in R:

Vector	1 dimension	Sequence of data elements of the same basic type.
Matrix	2 dimensions	Like a Vector but additionally contains the dimension attribute.
Array	2 or more dimensions	Hold multidimensional data. Matrices are a special case of two-dimensional arrays.
Data frame	2 dimensions	Table-like data object allowing different data types for different columns.
List	Collection of data objects, each element of a list is a data object.	



Logical statements

Operator	Function	Usage
<	less than	expression1 < expression2
<=	less than or equal to	expression1 <= expression2
>	greater than	expression1 > expression2
>=	greater than or equal to	expression1 >= expression2
==	equality	expression1 == expression2
!=	inequality	expression1 != expression2
&	logical AND	expression1 & expression2
||	logical OR	expression1 || expression2
!	logical NOT	!expression1
isTrue()	test if a variable is true	isTRUE(x)
is.na()	return TRUE if missing value	is.na(x)

Challenge 1

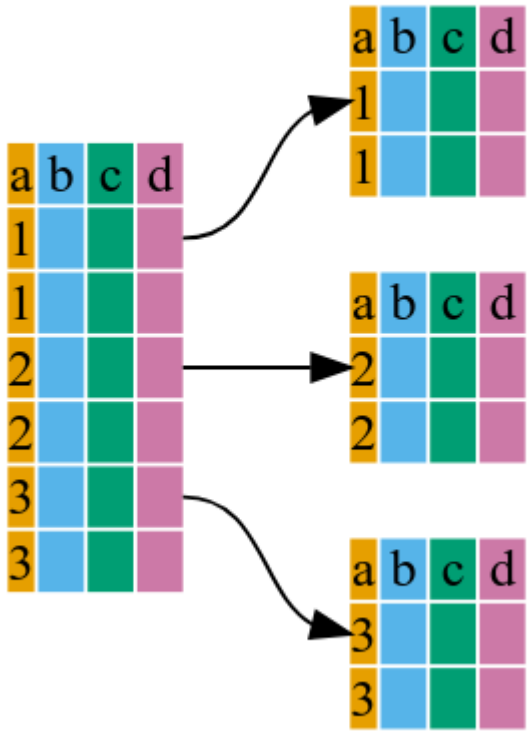
Start by making a vector with the numbers 1 through 26. Multiply the vector by 2, and give the resulting vector names A through Z (hint: there is a built in vector called `LETTERS`)

Challenge 1

Let's imagine that 1 cat year is equivalent to 7 human years.

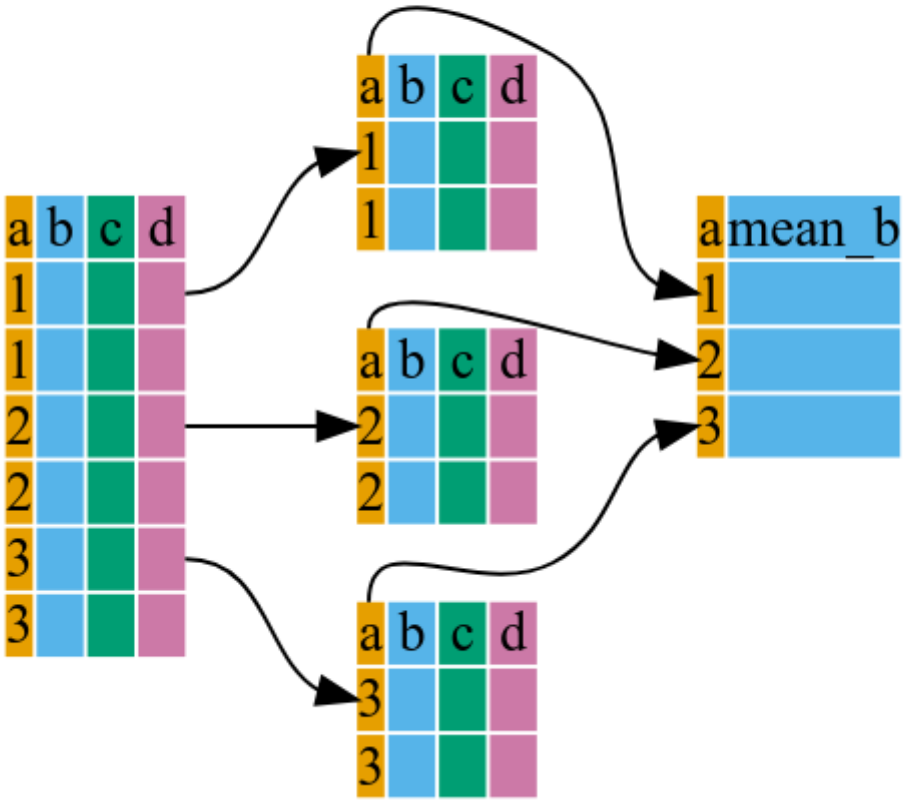
1. Create a vector called `human_age` by multiplying `cats$age` by 7.
2. Convert `human_age` to a factor.
3. Convert `human_age` back to a numeric vector using the `as.numeric()` function. Now divide it by 7 to get the original ages back. Explain what happened.

```
gapminder %>%
 group_by(a)
```



A grouped_df can be thought of as a list where each item in the listis a data.frame which contains only the rows that correspond to the a particular value

```
gapminder %>%
group_by(a) %>%
summarize(mean_b=mean(b))
```



Challenge 2

Calculate the average life expectancy per country. Which has the longest average life expectancy and which has the shortest average life expectancy?