00 - First Tour

Data Science with R · Summer 2021

Uli Niemann · Knowledge Management & Discovery Lab

https://brain.cs.uni-magdeburg.de/kmd/DataSciR/

Photo courtesy of Ulrich Arendt

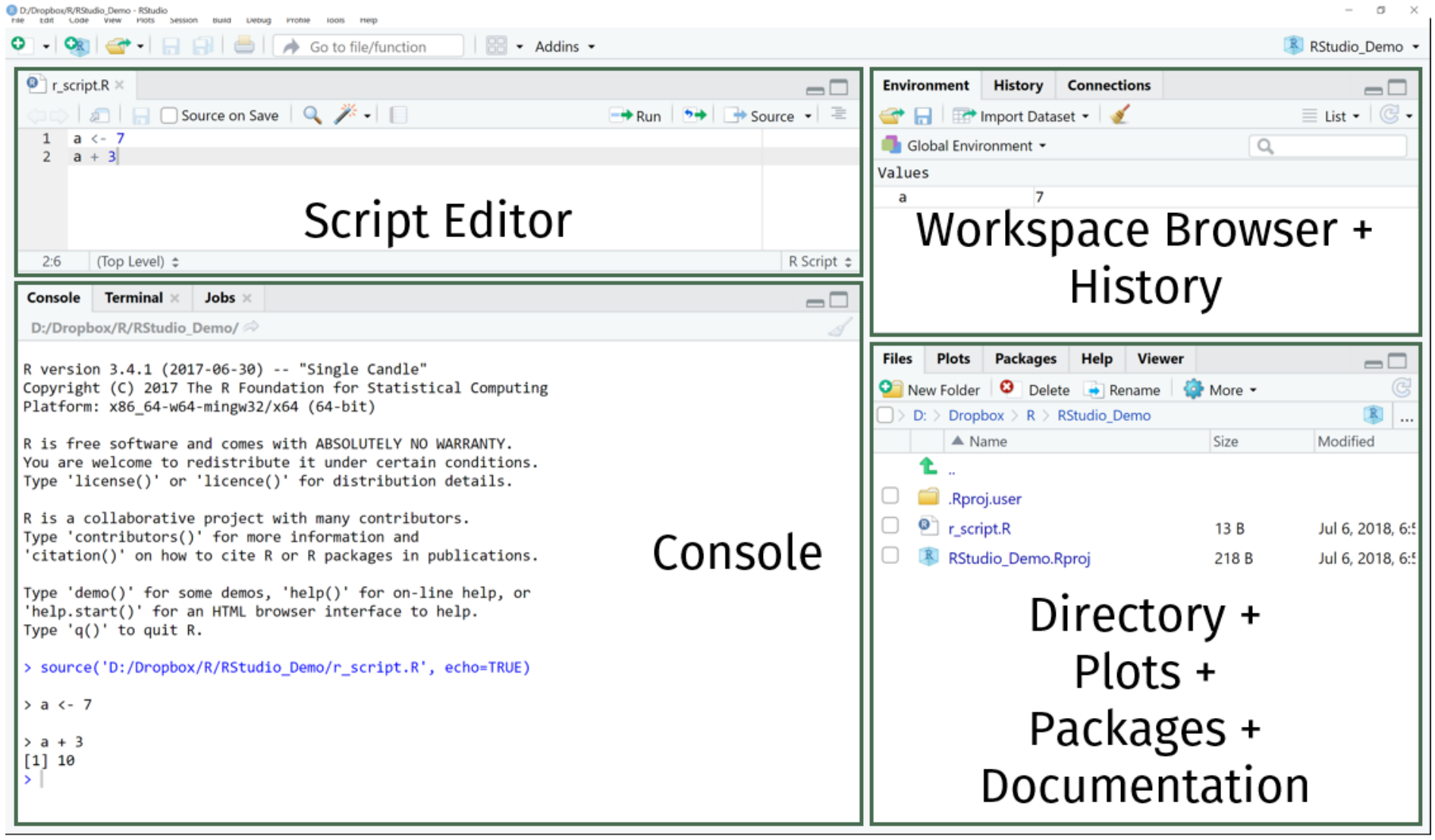
What is `R`?

- **Statistical programming language** with focus on **reproducible data analysis**
- **Free, open source** and **available for every major OS**
- Since January 2017 more than **10,000 packages** available
  - Comprehensive **statistics** and **machine learning** packages
  - Elaborate packages to create aesthetically appealing **graphics** and **charts**

What is RStudio?

- Open source **Integrated Development Environment (IDE)** for `R`
- First release in 2011
- Two major versions: **RStudio Desktop** and **RStudio Server**
- Download & Installation of `R` and RStudio: https://www.rstudio.com/products/rstudio/download/

RStudio IDE

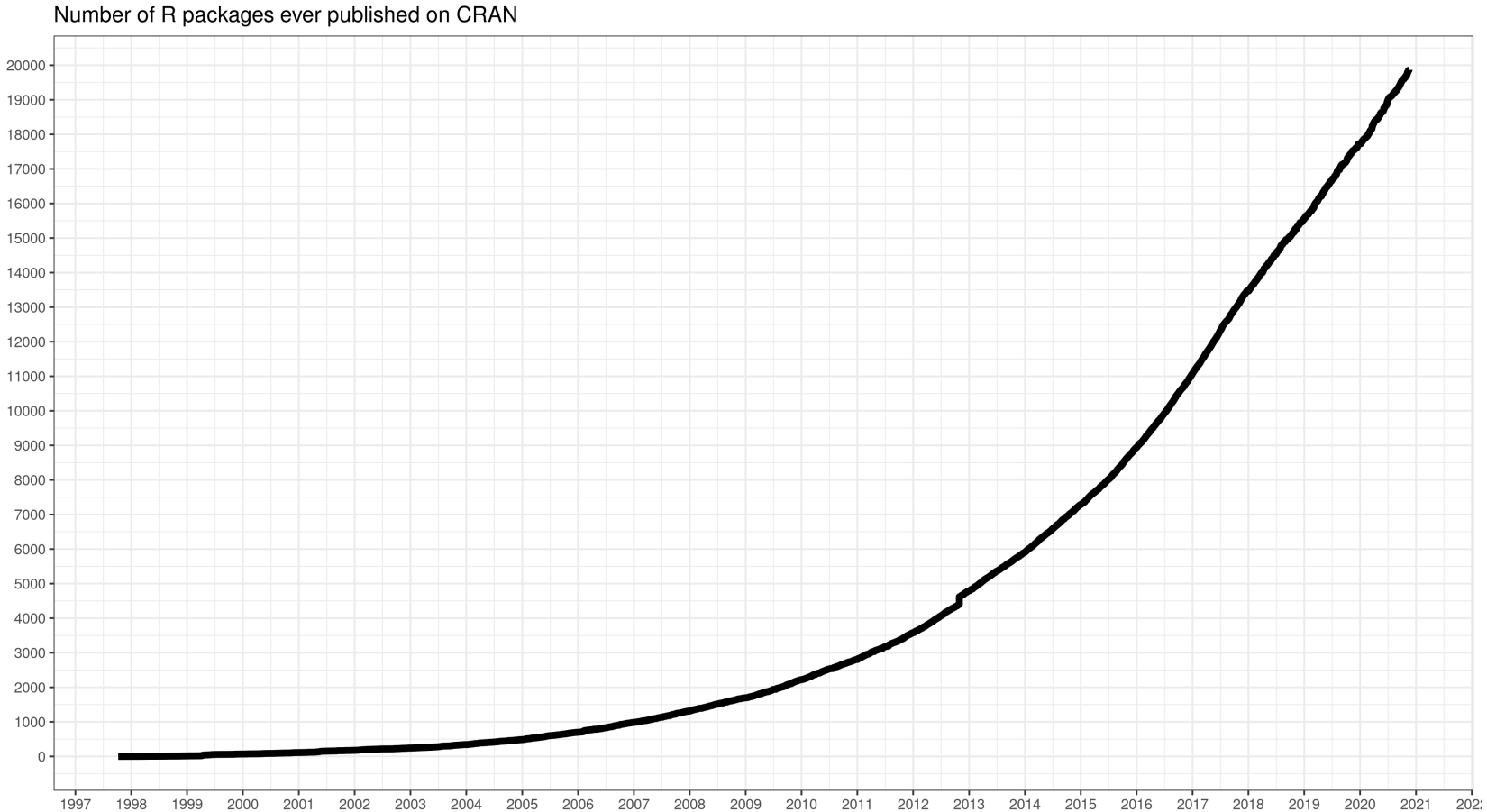

Go to file/function
Addins
RStudio_Demo
r_script.R
Source on Save
Run
Source
1 a <- 7
2 a + 3
Script Editor
2:6
(Top Level)
R Script
Environment
History
Connections
Import Dataset
List
Global Environment
Values
a
7
Workspace Browser + History
Console
Terminal
Jobs
D:/Dropbox/R/RStudio_Demo/
R version 3.4.1 (2017-06-30) -- "Single Candle"
Copyright (C) 2017 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (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.
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.
> source('D:/Dropbox/R/RStudio_Demo/r_script.R', echo=TRUE)
> a <- 7
> a + 3
[1] 10
>
Console
Files
Plots
Packages
Help
Viewer
New Folder
Delete
Rename
More
D: Dropbox R RStudio_Demo
Name
Size
Modified
..
.Rproj.user
r_script.R
13 B
Jul 6, 2018, 6:5
RStudio_Demo.Rproj
218 B
Jul 6, 2018, 6:5
Directory + Plots + Packages + Documentation

R packages 

- Base `R` contains functions that are needed by the majority of users.
- Additional functions can be used on demand by loading **packages**.

A **package** is a **collection of functions, datasets and documentation** that extends the capabilities of base R.

Packages on CRAN[1]



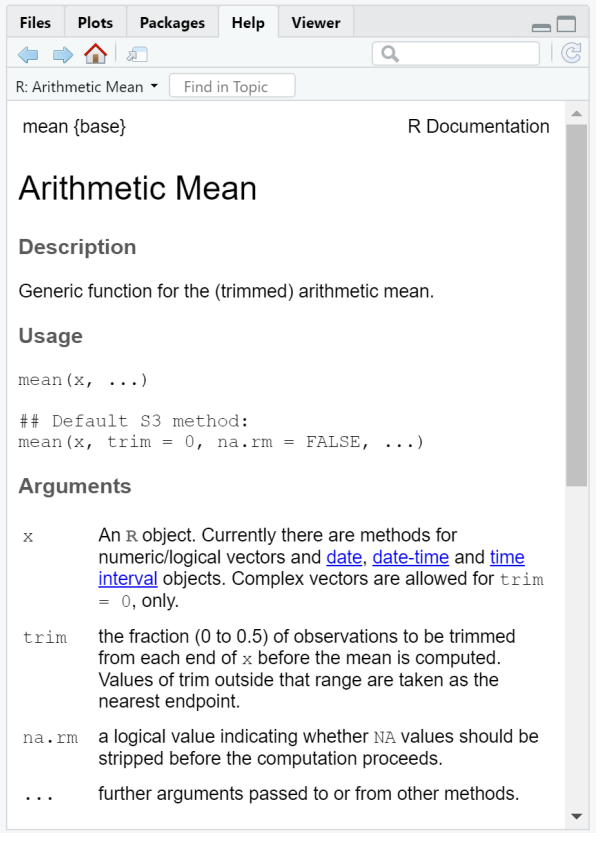
[1] CRAN, the *The Comprehensive R Archive Network*, consists of multiple worldwide mirror servers, used to distribute `R` and `R` packages.
Figure source: https://gist.github.com/daroczig/3cf06d6db4be2bbe3368

Tour: R and RStudio

Calling functions

Basic function call scheme:

```
some_function(arg_1 = val_1,
              arg_2 = val_2,
              ...)
```



Example: the `mean()` function:

- `x` is the only mandatory argument
- arguments `trim` and `na.rm` have default values

```
x <- 1:10
mean(x) # trim = 0 and na.rm = FALSE
```

```
## [1] 5.5
```

```
x <- c(1:10, NA)
mean(x)
```

```
## [1] NA
```

```
mean(x, na.rm = TRUE) # NA's will be ignored
```

```
## [1] 5.5
```

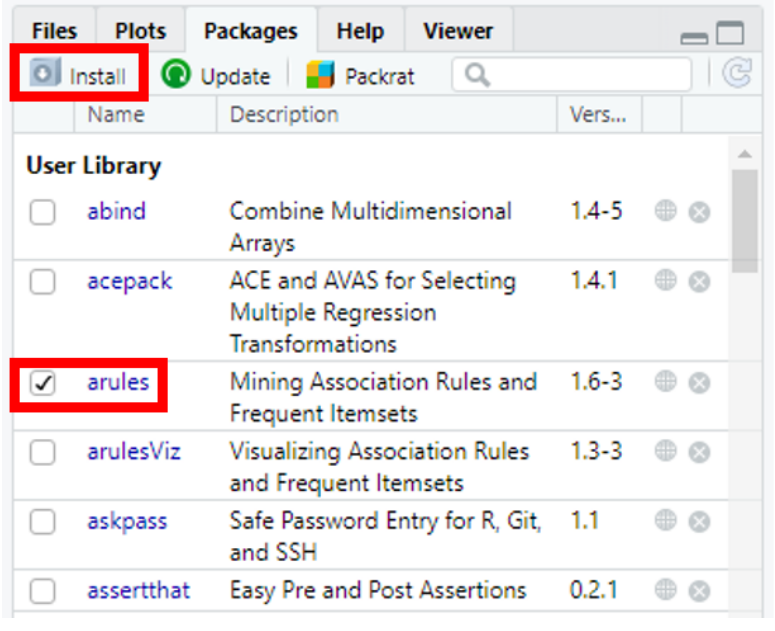
```
mean(x, TRUE) # match unnamed args to their position
```

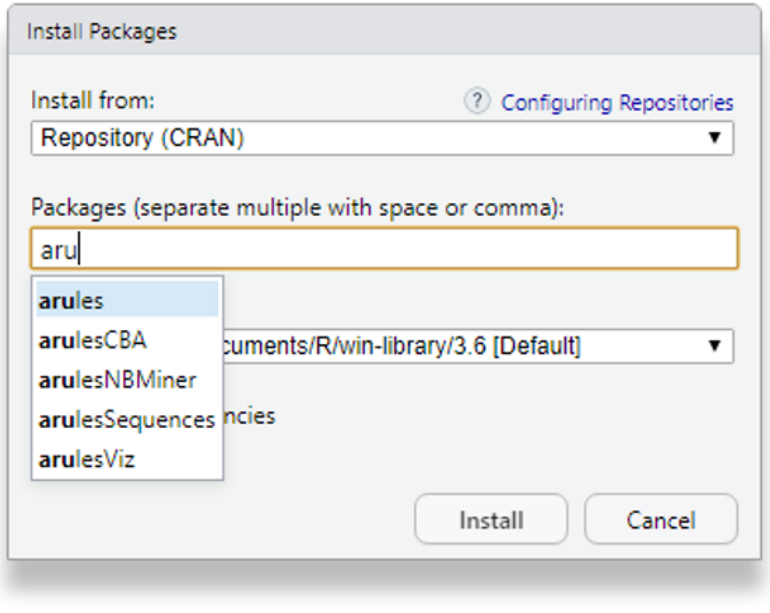
```
## Error in mean.default(x, TRUE): 'trim' must be numeric of length one
```

Downloading, installing and loading packages

```
# A package has to be installed only once:
install.packages("dplyr") # -> install package "dplyr" from CRAN
# Load the package once per session:
library(dplyr)
```

Alternatively, you can install and load packages in RStudio using the Packages tab.





To install a package that is hosted on GitHub use `remotes::install_github("<REPOSITORY>")`.

Data Frames

- `data.frame` is `R`'s data structure for a **table**.
- A data frame is a rectangular collection of data, arranged by **variables** (**columns**) and **observations** (**rows**).

Columns of data frames are accessed with `$`:

```
dataframe$var_name
```

The Tidyverse

Quote from the Tidyverse website:

> "**R packages for data science.** The tidyverse is an **opinionated collection of R packages designed for data science**. All packages share an underlying **design philosophy, grammar, and data structures**."

→ collection of open-source `R` packages mainly for data wrangling and visualization
→ shared conventions and common APIs across all Tidyverse packages

R Markdown

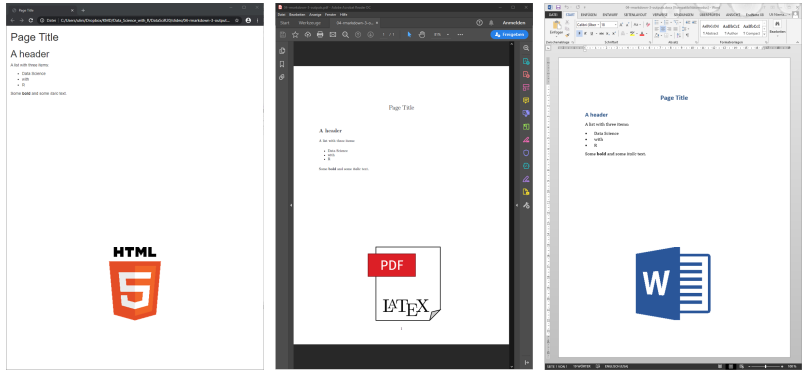
R Markdown is a file format to combine **code**, the associated **results** and **narrative text** in a simple text file, to create **reproducible reports** which can be **flexibly distributed in multiple ways**.

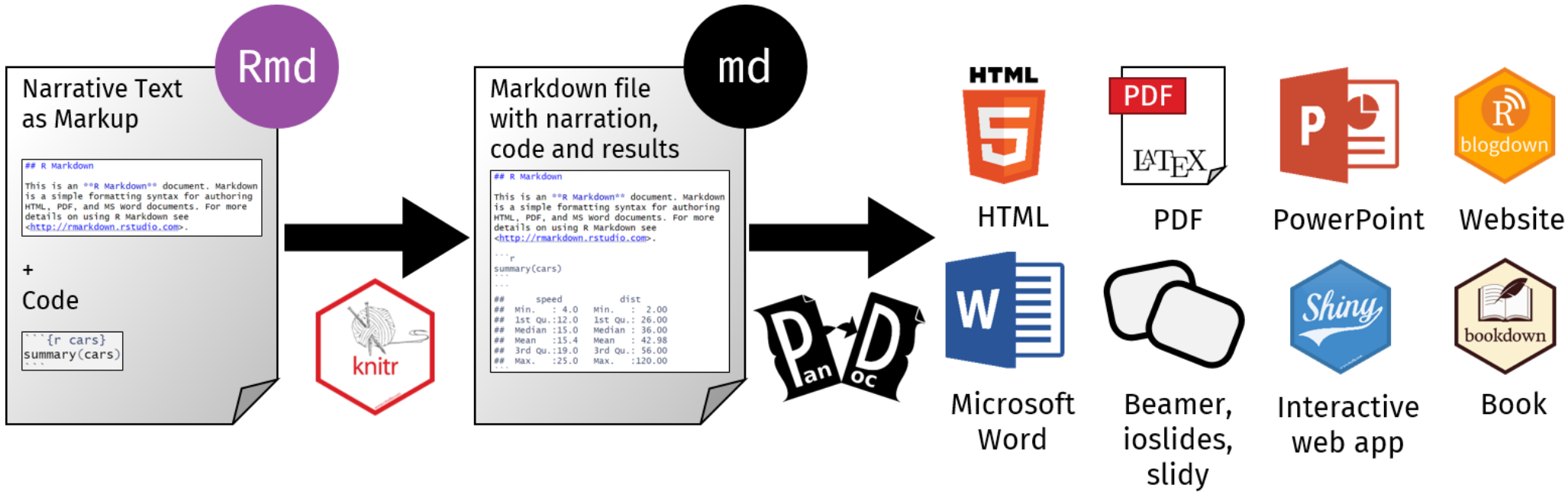
An R Markdown document is saved as `.Rmd` file. It contains both the (`R`) code and a prose description of a data analysis task. The R Markdown document can be rendered as HTML, PDF, Word and various other output formats.

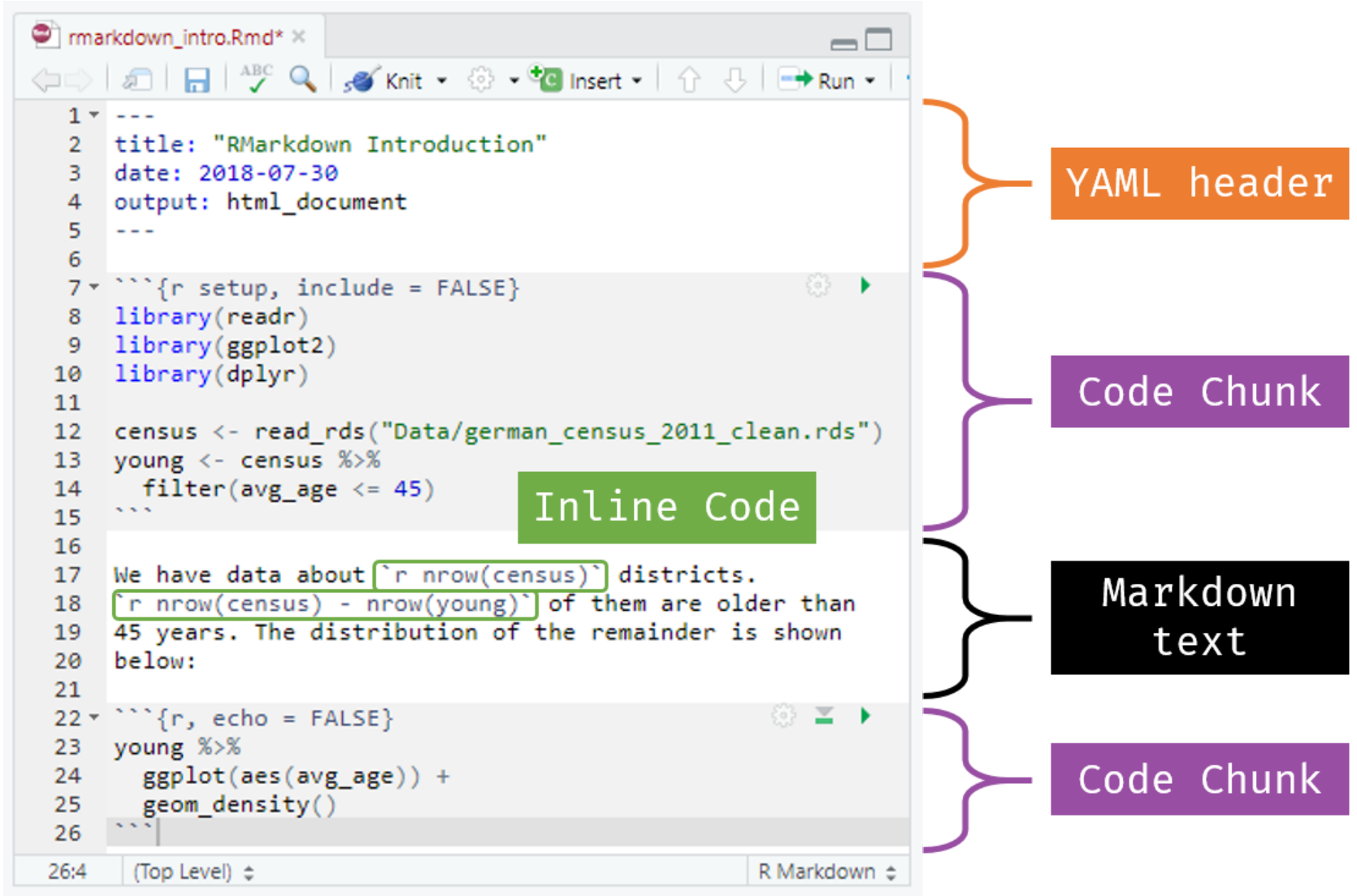
Pros of R Markdown documents:

- **reproducibility** (reduce **copy&paste**)
- **simple Markdown syntax** for text
- a **single source document** that can be rendered for different target audiences and purposes
- **simple, future-proof file format** that can be managed by a version control system like Git or SVN

https://rmarkdown.rstudio.com/

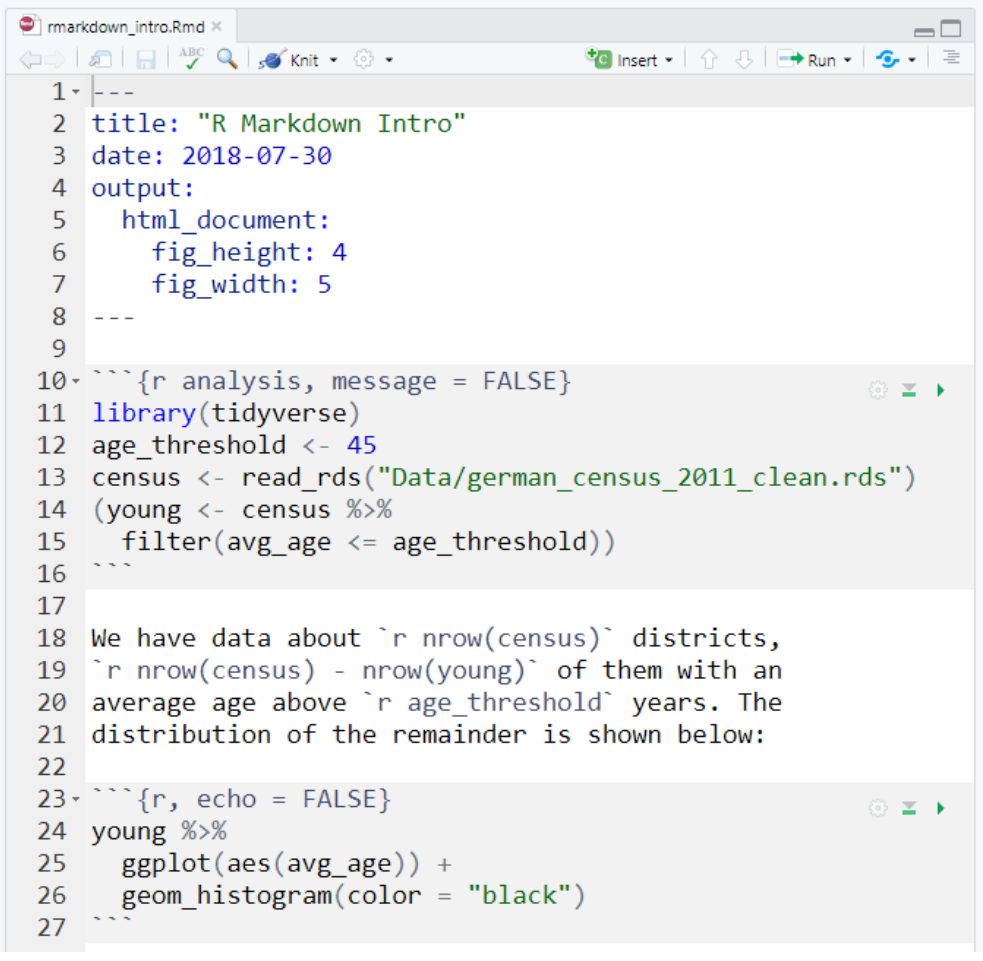



Rmd
Narrative Text as Markup
R Markdown
This is an **R Markdown** document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.
+
Code
```{r cars}
summary(cars)
```
knitr
md
Markdown file with narration, code and results
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>.
```r
summary(cars)
```
speed   dist
Min.   : 4.0   Min.   : 2.00
1st Qu.:12.0   1st Qu.: 26.00
Median :15.0   Median : 36.00
Mean   :15.4   Mean   : 42.98
3rd Qu.:19.0   3rd Qu.: 56.00
Max.   :25.0   Max.   :120.00
Pan Doc
HTML
HTML
PDF
LaTeX
PDF
PowerPoint
R
blogdown
Website
W
Microsoft Word
Beamer, ioslides, slidy
Shiny
Interactive web app
bookdown
Book


rmarkdown_intro.Rmd*
Knit
Insert
Run
1 ---
2 title: "RMarkdown Introduction"
3 date: 2018-07-30
4 output: html_document
5 ---
6
7 ```{r setup, include = FALSE}
8 library(readr)
9 library(ggplot2)
10 library(dplyr)
11
12 census <- read_rds("Data/german_census_2011_clean.rds")
13 young <- census %>%
14 filter(avg_age <= 45)
15 ```
16
17 We have data about `r nrow(census)` districts.
18 `r nrow(census) - nrow(young)` of them are older than
19 45 years. The distribution of the remainder is shown
20 below:
21
22 ```{r, echo = FALSE}
23 young %>%
24 ggplot(aes(avg_age)) +
25 geom_density()
26 ```
26:4
(Top Level)
R Markdown
Inline Code
YAML header
Code Chunk
Markdown text
Code Chunk

Tour: R Markdown

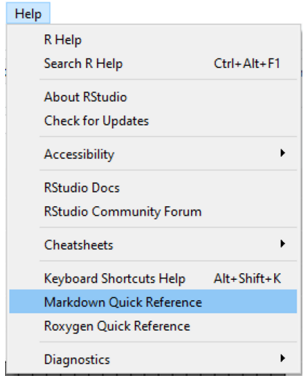
Structure of an `.Rmd` file

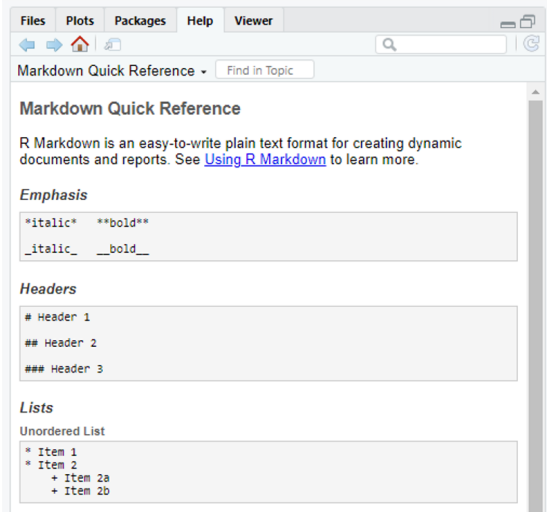


- **YAML header** (metadata)
  - enclosed by lines with three dashes
  - collection of key-value pairs separated by colons
- narrative text as **Markdown markup**
- `R` **code**
  - as code chunks surrounded by ```` ```{r} ```` and ```` ``` ````
  - as inline code surrounded by `` `r `` and `` ` ``

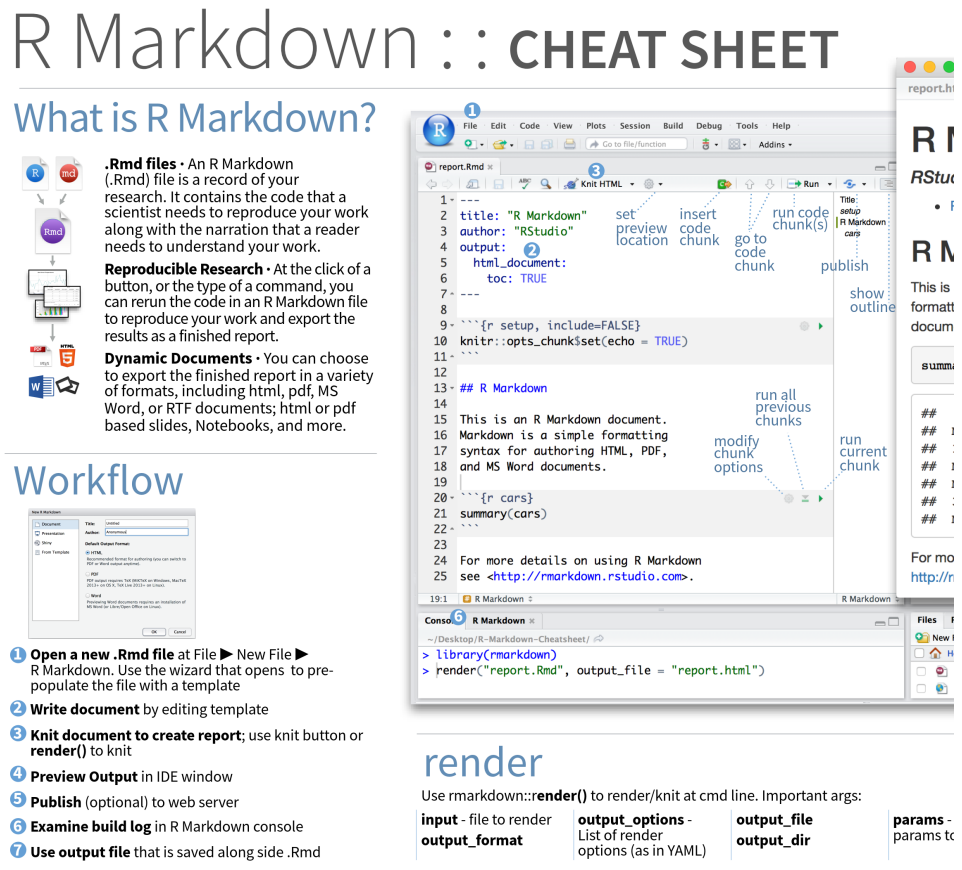
R Markdown help

Markdown Quick Reference:
Help → Markdown Quick Reference





R Markdown Cheat Sheet:
Help → Cheatsheets → R Markdown Cheat Sheet



Environments

⚠ Each R Markdown document has its own environment.

→ Consequence: Objects from the global environment are unavailable in the .Rmd document when knitting.

Example:

1. Run the following code in the console:

```
x <- 5
x * 3
```

😊 *All good!*

2. Add the following code within a code chunk to your .Rmd file and knit the document:

```
x * 3
```

😱 *Knitting fails!*

How will we use R Markdown?

- Every exercise comes with an R Markdown document which contains some code scaffolding.
- You submit your project report as an R Markdown document.
- The course slides are made with R Markdown.


Thank you! Questions?
Photo courtesy of Stefan Berger