Reproducibility

r markdown

Federica Fusi

MSCA, UIC

Updated: 2021-03-09

R markdown check

- Open R Studio
- Click on File -> New File -> R Markdown.
    - If you have never used R markdown before, it should prompt you to install a set of packages. Please say "yes" and install them.
    - If R markdown is alredy installed, it should prompt you to name a new file and open a new document similar to a R script.
- In either cases, if you have never knit a R markdown document into a PDF document, please install tinytex by running both these commands:

```
install.packages('tinytex')
tinytex::install_tinytex()
```

REPRODUCIBILITY

Reproducibility

What does reproducibility mean?

It is about setting up all your processes in a way that is repeatable by others and well documented.

Why do we care about reproducibility?

- Can I take your script and run it again without making any change?
- Can I slightly change the original data, run the script again and not have any warning messages?
- Can I take your script, not run it, and still understand what you are doing?
- Can you take your script in one year and still perfectly understand what you did?

Reproducibility - how?

(1) At the most basic level, it means that you are using a repeatable method for every actions that you are doing - your R script!

R script can be saved, annotated, and shared with others so that they can run the same workflow and get to the same output. Avoid any manual step when working with your data.

(2) Document your process. Write comments and annotations at each step of the way. Other people or you after one year should be able to quickly understand what you did and WHY.

In too many cases, I still need to carefully read the code or run it to understand what you are doing.

Reproducibility - how?

(3) The best R script are simple and well annotated. You have to re-check your script after you are done with it to clean it up a bit. In some cases, you might realize that some steps could have been done more efficiently.

A couple of things I have noticed:

- Avoid reporting number manually: if you are setting the hline or vline to be equal to the mean or setting the abline, put the formula in plot.
- Avoid merging datasets unless it is necessary. Use pipes instead.

```
ggplot(election_turnout) +
  geom_density(mapping = aes(x = turnoutho)) +
  geom_vline(xintercept = 23.5)

ggplot(election_turnout) +
  geom_density(mapping = aes(x = turnoutho)) +
  geom_vline(xintercept = mean(election_turnout$turnoutho))
```

Reproducibility - how?

(4) Version control means keep old versions of your codes and data so that you can track down errors but also go back to old versions of the data.

Always keep the original dataset but save intermediate versions and the final one. Use progressive names (e.g., numbers or dates)

- data_1, data_2, data_3 OR data0307, data0315, data0324

(5) Keep your work on a cloud system rather than your own computer as much as possible. A few consideration:

- Pay attention to privacy when saving online sensitive data (e.g., data containing personal information)
- You can use password-protected files or remove personal identifiable information.

(6) When possible, provide public access to your data and code (and encourage your organization to do it).

- Again, privacy concerns: can individuals be identified?

Replicability

Replicability and reproducibility are often used interchangeably.

Replicability is different from reproducibility. Replicability means that your results can be replicated by others even if they use a different dataset or a slightly different method.

- For instance, you find that speed cameras do not predict the number of car accidents in Chicago. Does the same correlation hold in Seattle?

Reproducibility is needed to ensure replicability - another individual should know what you did in order to check if they can find the same result.

Transparency

Transparency means to disclose your thinking process and any discover that you make from your own data. Report all information that you extracted from the data, including limitations or issues with the data collection.

Some of these information is reported in your final report - the level of detail depends on the purpose of your report.

In most cases, you should create a methodological note at the end of the your report all these details: data used, missing value issues, limitations of your data collection, and so on.

R MARKDOWN

R markdown

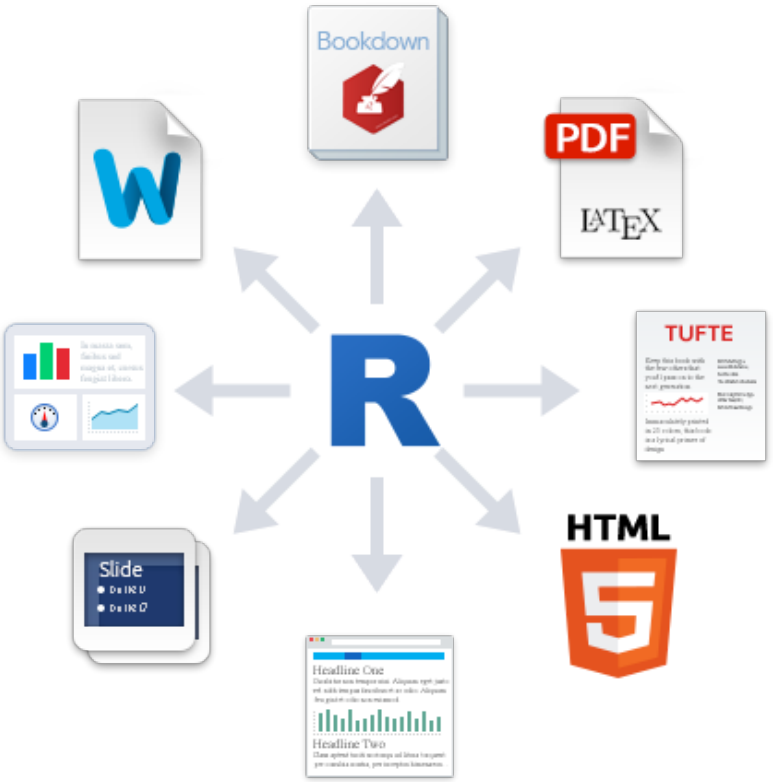
R markdown files are data-driven docs that combine TEXT, CODES, and RESULTS.

They promote transparency, reproducibility, and replicability as they allow others to see your annotations + codes + outputs. The balance of these three elements depends on the purpose of your work.

They are a great way to share your results to non-R users.

R markdown

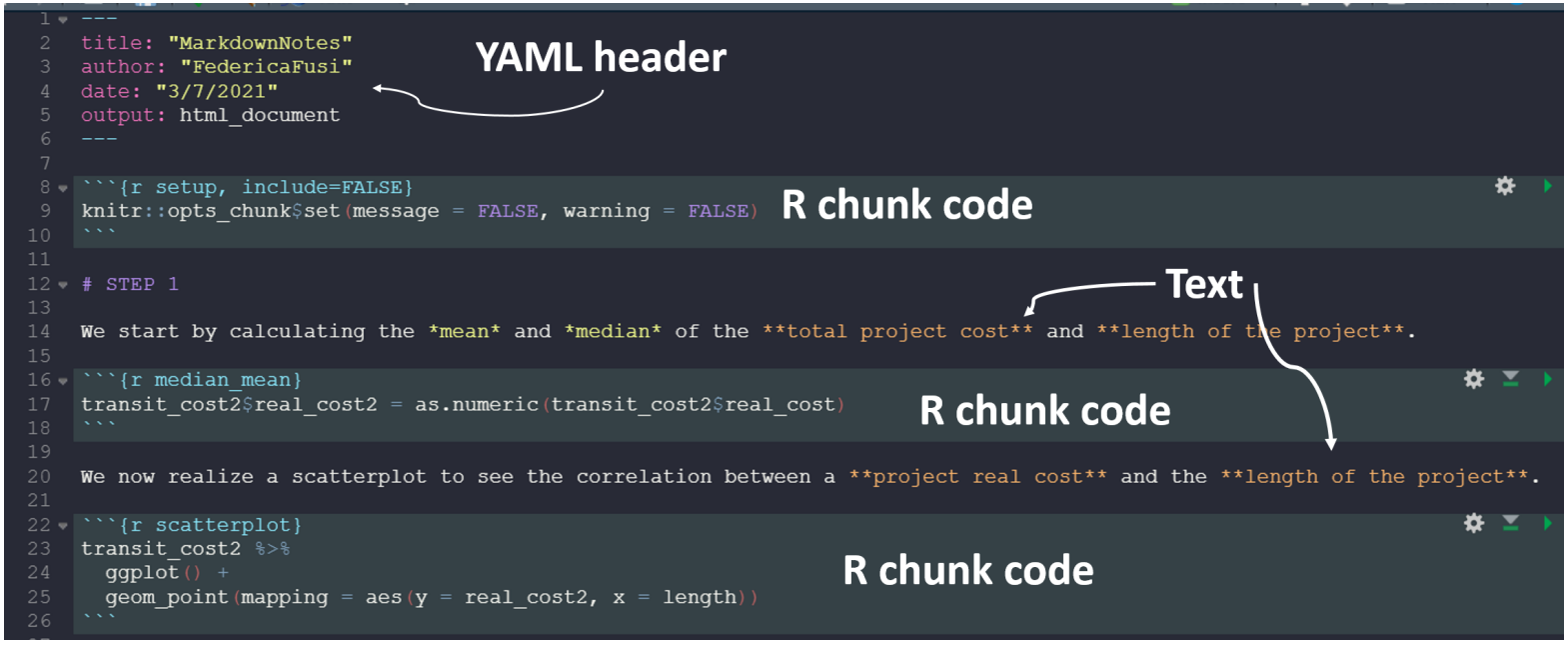
R markdown is a set of rules to format these data-driven documents. It upports dozens of output formats, like PDFs, Word files, slideshows, and more.



Type of R markdown documents

How it works - writing

```
knitr::include_graphics("rmarkdown.png")
```



How it works - knitting

When your script is ready, you can knit it and produce a complete report containing all text, code, and results.

```
knitr::include_graphics("knitting.png")
```

Text

There are few conventions on how to write text in R markdown. For instance:

```
# indicates the main Title

## Subtitle

### Header 3

**Bold text**
*Italic text*
```

R chunks

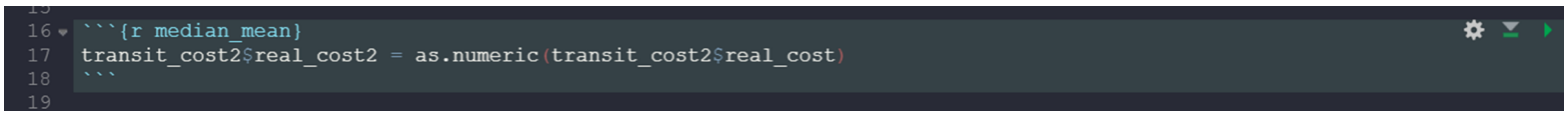
You can create a new R chunk by:

1. using the keyboard shortcut Cmd/Ctrl + Alt + I (recommended option)

OR

1. by manually typing the chunk delimiters at the beginning and end.

```
knitr::include_graphics("chunk.png")
```



R chunks

Let's look at the breakdown of a R chunk

```
{r NAME_OF_THE_CHUNK, OPTION1, OPTION2}
```

You can give the chunk any name (use something meaningful).

```
{r project_mean, OPTIONS}
```

R chunk options

There are several options that you can put in a r chunk. You can see the full list here

eval = F
Show the code but don't run it

echo = FALSE
Code doesn't appear in the output file. Results appers. When including an image, you generally want to use this.

message = FALSE
Prevent messages from appearing in the output file

warning = FALSE
Prevents warnings from appearing in the output file

```
{r project_mean, echo = F, warning = F, message = F}
```

R setup

The first R chunk in a document is generally called 'setup'. The name 'setup' is used only for the very first chunk where you can set up settings to be applied to the entire document.

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

This setup, for instance, is great for reports where you might not want to show your code but only the results.

linline codes

If needed, you can also set text into your inline codes simply using the format `` `mean(XXX)` `` .

This will report your results into the text and save you time when updating your work!

R markdown resources

R markdown the definitive guide

Gallery with multiple examples of R markdown galley

Multiple R markdown formats

Intro to R markdown

R markdown quick explanation

Assignment

Go back to assignment 7 (your transit project plot) or assingment 8.
Incorporate some of the suggestions that you received on how to improve your graph.

Write a RMD document and knit into PDF or html.

Submit both your output file and .rmd document.

Group discussion

Present your work to others!

Briefly describe the question that you were interested in exploring and your key results.

Show your data visualization and ask others to comment on it.

- Do they get your main message?
- How can you improve the visualization of your data?

Your pitch should last only a couple of minutes! The scope is to receive feedback not presenting!