Data Management With R: Markup languages

Matthias Haber

13 November 2017

Prerequisites

Packages

```r
library(tidyverse)
library(stringr)
library(rvest)
library(knitr) # install.packages("knitr")
```

Last week's homework

Scraping journal editorial team

```r
url <- "https://www.jstatsoft.org/about/editorialTeam"
download.file(url, destfile = "data/editorial.html")
url_parsed <- read_html("data/editorial.html")
names <- html_nodes(url_parsed, "#group a") %>%
  html_text()
affiliations <- html_nodes(url_parsed, ".member li") %>%
  html_text() %>%
  str_replace("^[^,]*$", "") %>%
  str_replace("^[^,]*,", "") %>%
  str_trim()
df <- data.frame(names, affiliations)
str_detect(affiliations, "tatisti|athemati") %>%
  table
```

```
## .
## FALSE  TRUE
##    38    29
```

A note on forms

```
url <-"https://connect.hertie-school.org/login/"
session <-html_session(url)
form    <-html_form(session)[[1]]
username <- "haber.matthias@gmail.com"
password <- "HSOGDS2017!"

filled_form <- set_values(form,
                          `loginname` = username,
                          `loginpwd` = password)
img <- submit_form(session, filled_form) %>%
  jump_to("https://connect.hertie-school.org/directory/") %>%
  read_html() %>%
  html_nodes(".rounded") %>%
  html_attr("src") %>%
  .[1]
```

```
## Submitting with 'loginform'
```

Be careful with your data

Markup language and literate programming

Objective for today

What is literate programming? Why is it important for reproducible research?

- Introduction to Markdown
- Introduction to R Markdown:
  - Simple pages
  - PDF papers
  - Presentations
  - References
- Slides adapted from Christopher Gandrud's course on Collaborative Social Science Data Analysis

What is literate programming?

Literate programming: a program using natural language interspersed with code snippets that are compilable by a computer. Donald Knuth (1992):

This produces two representations of the program:

- A formatted easily human readable document (e.g. a paper).
- Source code that can be compiled by a computer.

General benefits

Creates better programs by forcing programmers to explicitly state thoughts

Clear documentation so that others can understand and build on the program

Benefits related to research

Quantitative social science is computer programming.

- You are creating a program that gathers and analyses data.
- You then advertise this work (a paper) in a way that is completely understandable to others.

Implementing literate programming

In addition to the computer language, we need:

1. Natural language part formatted using a markup language. Markup language: typesetting instructions. E.g. Markdown, LaTeX, HTML.
2. A way to tangle or weave the computer language part into the natural language part.

`knitr` for R

In R you can use the `knitr` package.

Two parts:

- Natural language part written in intended markup language.
- R code (or almost any other language on your system) written in code chunks.

Latex Output

```
\documentclass{article}
\begin{document}
\section{This is a section heading}
Here is some text. Followed by an R code chunk to create a plot:
<< >>=
library(ggplot2)
ggplot(mtcars, aes(hp, mpg)) + geom_point()
@
Then some more text.
\end{document}
```

In a knitr-LaTeX document (also known as R Sweave and has the file extension `.Rnw`) code chunks are delimited with `<< >>=` and `@`.

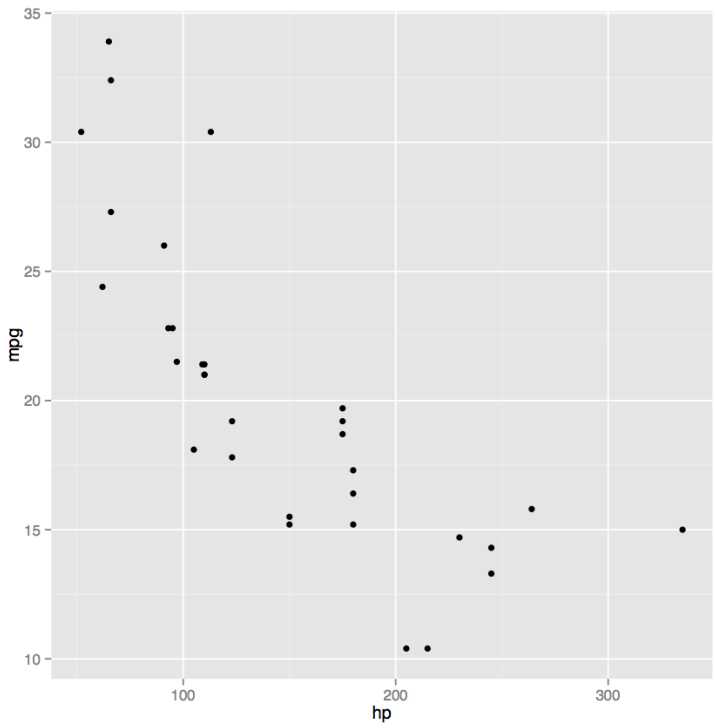
PDF Output

1 This is a section heading

Here is some text. Followed by an R code chunk to create a plot:

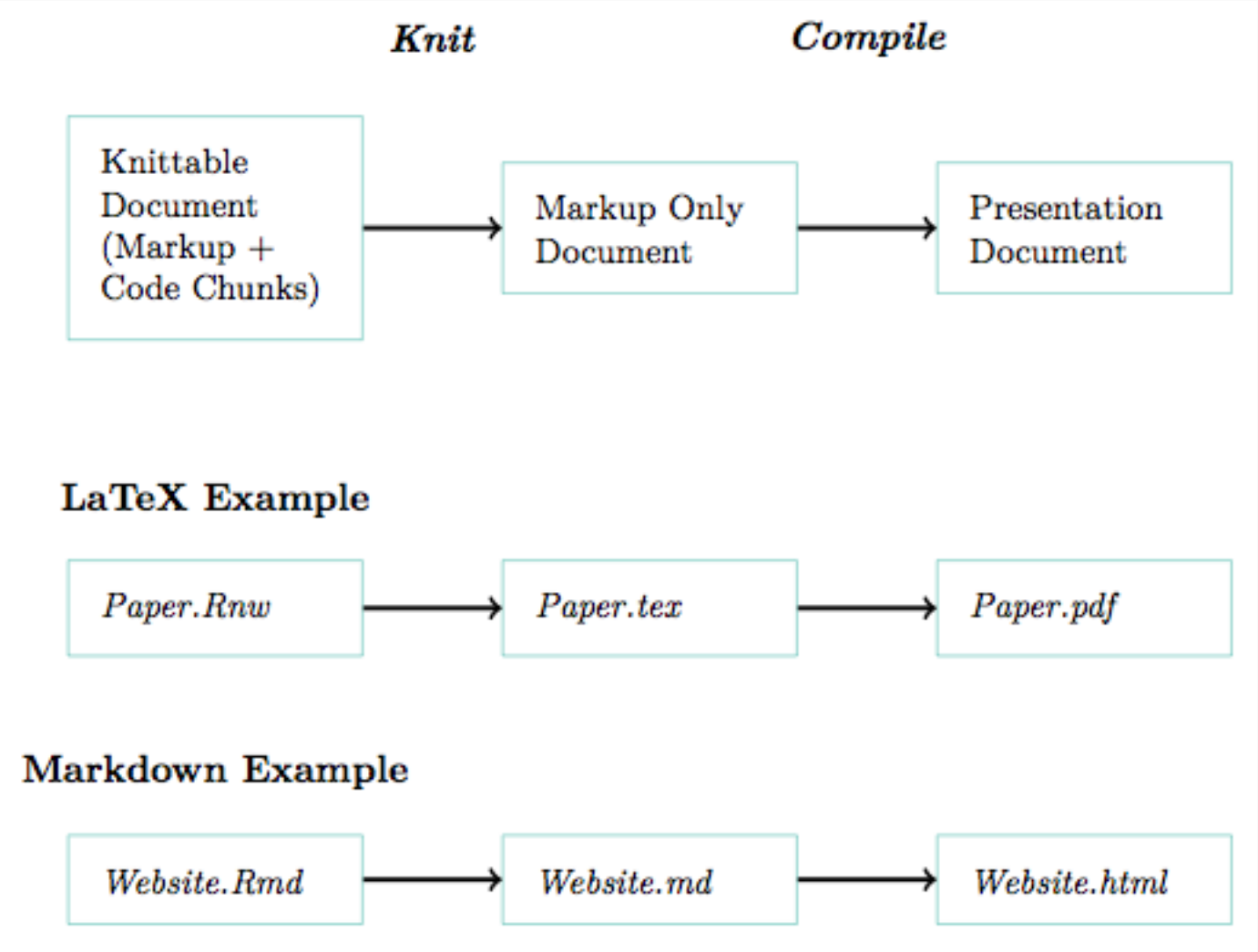
```
library(ggplot2)

ggplot(mtcars, aes(hp, mpg)) + geom_point()
```



Then some more text.

The `knitr` process



R Markdown

R Markdown

Most of the focus is on RStudio's R Markdown.

- Directly builds on knitr (developer works at RStudio now).
  - Code chunk syntax is almost identical to Markdown in knitr.
- But uses Pandoc to be more output agnostic.
  - You can write in R Markdown and output to many different formats.

Markdown

Originally created by John Gruber to be an easy way to:

- write HTML files
- that are human readable as text files

Markdown: HTML less painful

HTML:

```
<h1>A header</h1>

<p>Some text with a <a href="http://www.example.com">link</a></p>

<p>Here is some <strong>bold</strong> text.</p>
```

Markdown:

```
# A header

This is some text with a [link](http://www.example.com).

Here is some **bold** text.
```

Markdown syntax: Headers

```
# Header 1

## Header 2

### Header 3
```

And so on.

Markdown syntax

Horizontal lines:

```
---
```

Bold text:

```
**bold**
```

Italics:

```
*italics*
```

Markdown syntax

Links:

```
[link](http://www.example.com)
```

Images:

```
![text description](FILE/PATH.png)
```

Markdown syntax

Unordered Lists:

```
- An item

- An item

- An item
```

Ordered Lists:

```
1. Item one

2. Item two

3. Item three
```

Tables

```
| Name   | Something |
| ------ | --------- |
| Stuff  | Things    |
| Things | Stuff     |
```

Name	Something
Stuff	Things
Things	Stuff

Math

R Markdown from RStudio supports MathJax. So, you can write any LaTeX math with R Markdown.

```
Inline equations have one dollar sign $s^2 =
\frac{\sum(x - \bar{x})^2}{n - 1}$.
```

Inline equations have one dollar sign $s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$.

Display equations have two dollar signs:

```
$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$$
```

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$$

Code chunks Inline

To use syntax highlighting on code chunks inline with the text, surround your text with ‘ ‘

For example:

```
Two plus two equals `r 2 + 2`.
```

Produces:

Two plus two equals 4.

Code chunks in Display

Use three ticks ```` to start and end a code chunk that is not run.

Create a knit-able code chunk begin the chunk with ```{r}

```
# This is a section heading

Here is some text. Followed by an R code chunk to create a plot:

```{r}
library(ggplot2)

ggplot(mtcars, aes(hp, mpg)) + geom_point()
```

Then some more text, followed by a table.

```{r echo=FALSE, results='asis'}
knitr::kable(mtcars)
```
```

Automatic table generation

You can turn any matrix or data frame into a well formatted table with the knitr function `kable`.

```
knitr::kable(mtcars)
```

Make sure that the code chunk option `results='asis'`.

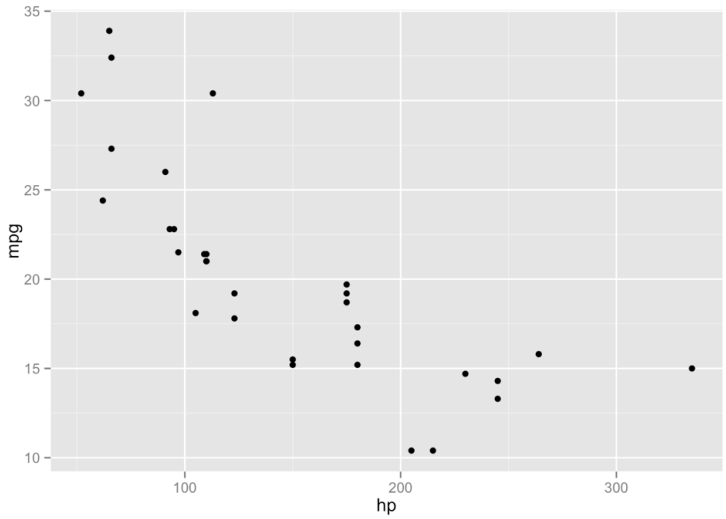
Output Web

This is a section heading

Here is some text. Followed by an R code chunk to create a plot:

```
library(ggplot2)

ggplot(mtcars, aes(hp, mpg)) + geom_point()
```

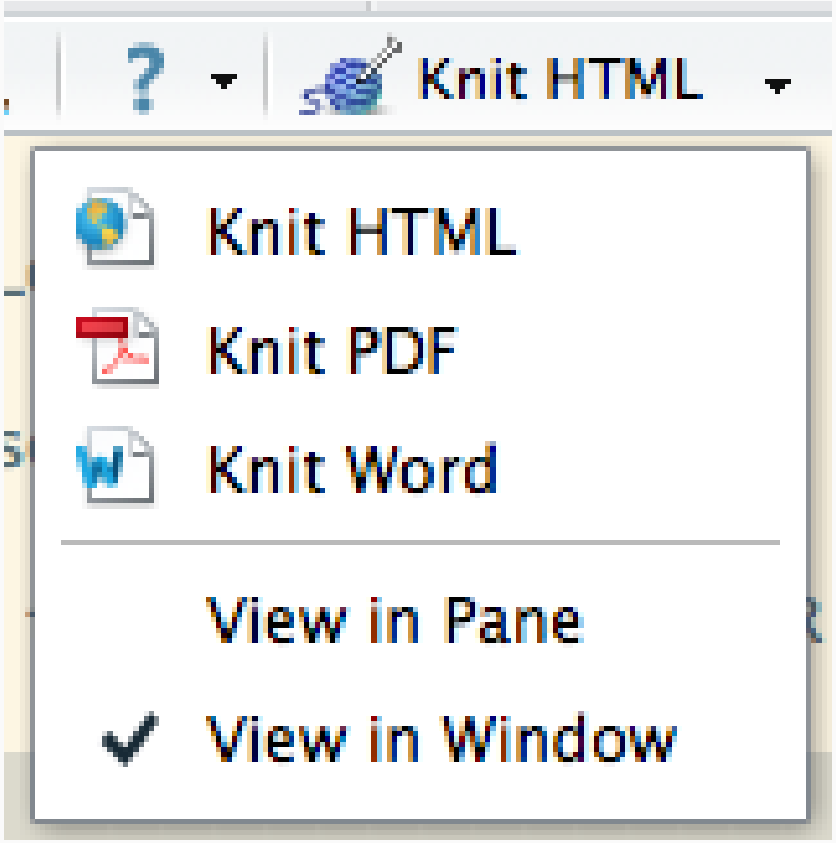


Then some more text, followed by a table.

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4

Output PDF or Word

A R Markdown file can be compiled to PDF (via LaTeX) or MS Word with RStudio.



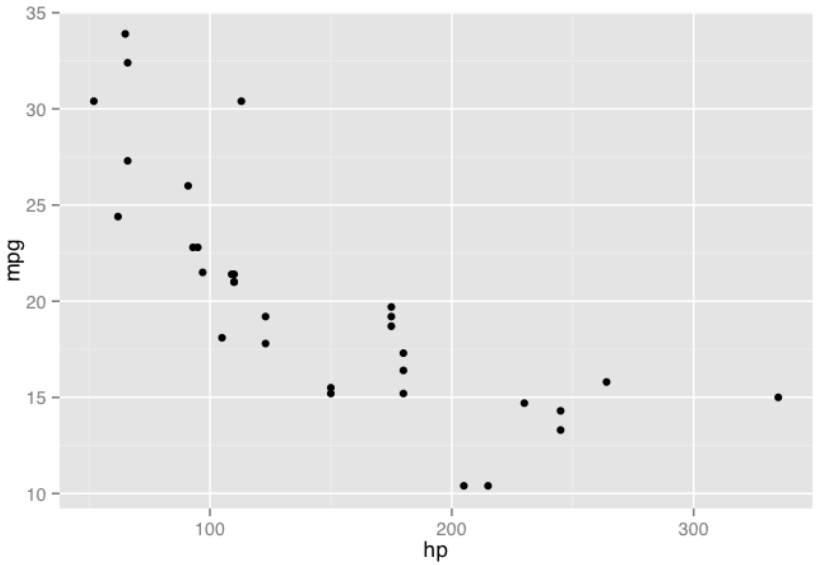
Output PDF

This is a section heading

Here is some text. Followed by an R code chunk to create a plot:

```
library(ggplot2)

ggplot(mtcars, aes(hp, mpg)) + geom_point()
```



Then some more text, followed by a table.

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1

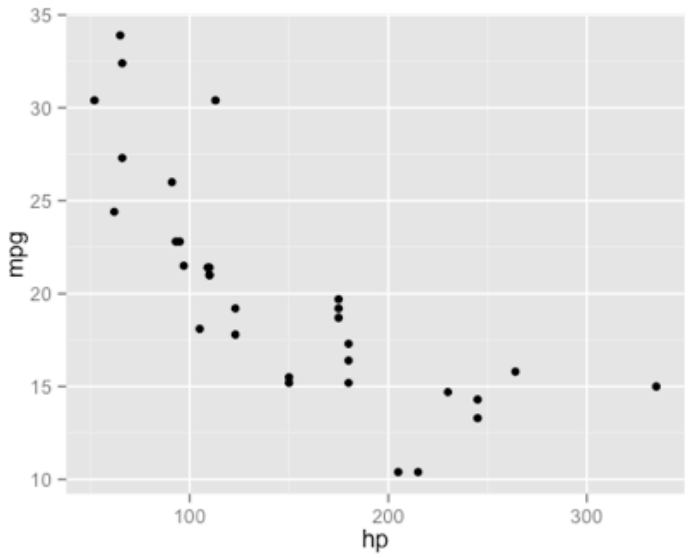
Output Word

This is a section heading

Here is some text. Followed by an R code chunk to create a plot:

```
library(ggplot2)

ggplot(mtcars, aes(hp, mpg)) + geom_point()
```



Then some more text, followed by a table.

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1

Chunk options

Change how R Markdown chunks behave with options. Place options in the chunk head: ```{r echo=FALSE, error=FALSE}

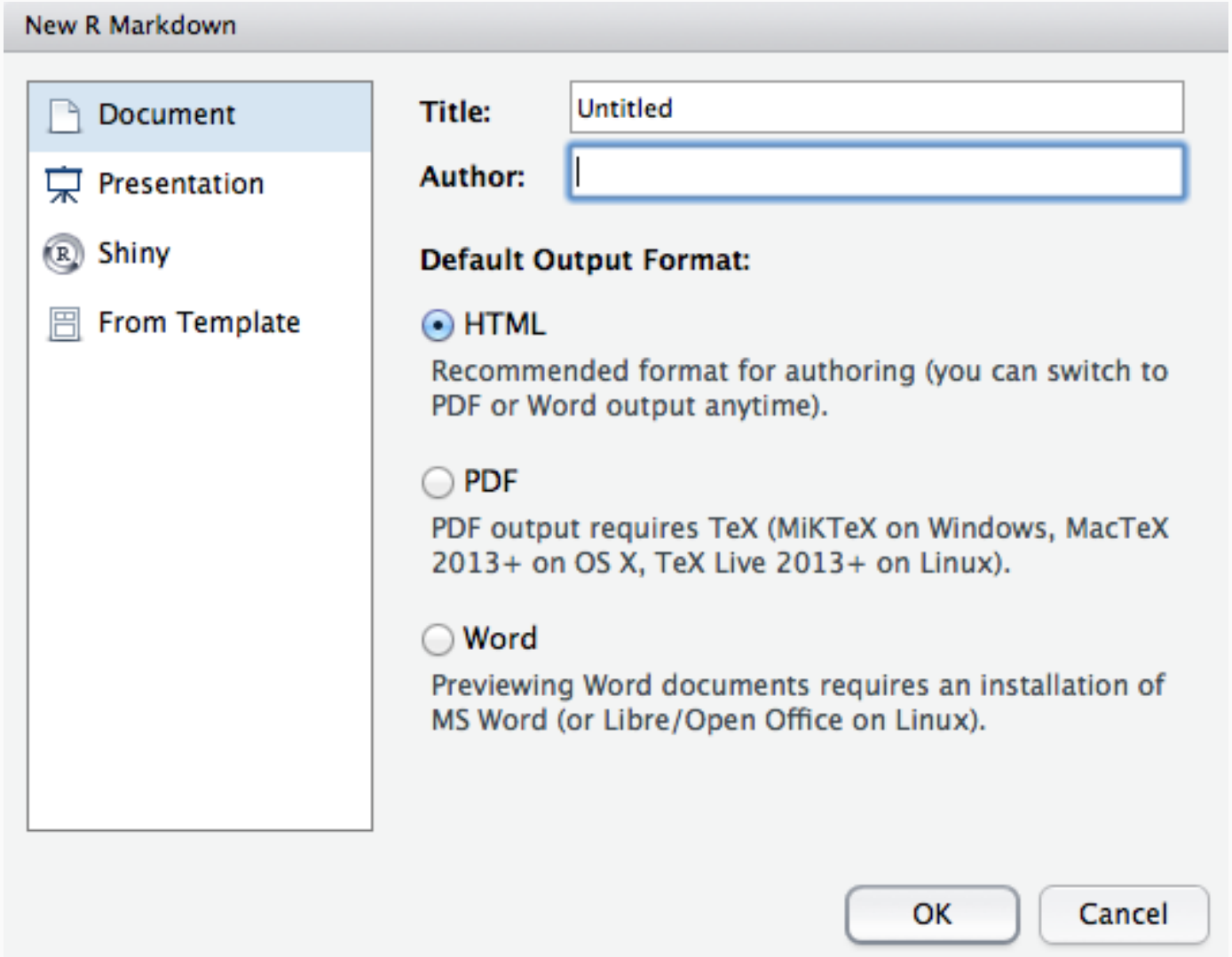
Option	What it Does
`echo=FALSE`	Does not print the code only the output
`error=FALSE`	Does not print errors
`include=FALSE`	Does not include the code or output, but does run the code
`fig.width`	Sets figure width
`cache=TRUE`	Cache the chunk. It is only run when the contents change.

Many others at `http://yihui.name/knitr/options`

RMarkdown Presentations

RMarkdown Presentations

These lecture slides are created using R Markdown.



RMarkdown Presentations

All of the syntax is the same, except:

Does not mean Header 2. It is creates a new slide and title.

You can create a slide with no title using ---.

R Markdown Header

An R Markdown file is just a text file with markup instructions that RStudio understands. The key to document-consistent formatting is the header.

It is at the start of a file and comes between `---`.

The header is written in YAML.

The header lets you make changes to the whole document.

YAML

YAML is a human read-able data format ("YAML Ain't Markup Language").

Elements are separated from values with a colon (`:`).

Each element is separated by new lines.

Hierarchy is maintained with tabs.

YAML

This presentation's head is:

```
---
title: "Data Management With R: Markup languages"
author: "Matthias Haber"
date: "13 November 2017"
output:
  beamer_presentation:
    theme: "metropolis"
    colortheme: "default"
    fonttheme: "default"
    fig_caption: false
    df_print: default
    toc: false
  ioslides_presentation:
  slidy_presentation: default
---
```

Different Presentation Styles

By defualt, R Markdown uses the ioslides HTML presentation slides style.

You can also use `beamer_presentation` style that you may know from LaTeX by changing the output in the YAML header:

```
output: revealjs::revealjs_presentation
```

Table of Contents & Numbered Sections

You can add a table of contents and numbered sections to your PDF output:

```
output:
  pdf_document:
    toc: true
    number_sections: true
```

To do the same for `HTML` also include the information under `html_document`.

Figure Options

Create consistent figure formatting:

```
output:
  pdf_document:
    fig_width: 7
    fig_height: 6
    fig_caption: true
```

`fig_caption: true` attaches captions to figures.

To set the actual caption label, use `fig.cap='SOME CAPTION'`.

Pandoc footnotes

R Markdown can use Pandoc footnotes.

In-text: In the text place a unique footnote key in the format:

- `[^KEY]`

End: At the end your document put the full footnote starting with the key, e.g.

- `[^KEY]: This is a footnote.`

Citations in R Markdown

BibTeX citations

BibTex allows you to create a database of all of the literature/packages you cite.

You can then insert them into your text and they will:

- Be automatically formatted consistently.
- Generate an appropriately ordered, consistently formatted reference list at the end of your document with only the works you actually cited.

The BibTeX Database

A BibTeX database is just a text file with the extension `.bib`.

Each entry follows a standard format depending on the type of media.

```
@DOCUMENT_TYPE{CITE_KEY,
    title = {TITLE},
    author = {AUTHOR},
    . . . = {. . .},
}
```

Note: Commas are very important!

The Cite Key

The cite key links a specific citation in your presentation document to a specific BibTeX database entry. They must be unique.

It does not matter what order your BibTeX entries are in the `.bib` file.

BibTeX Articles

```
@article{Acemoglu2000,
    author = {Acemoglu, Daron and Robinson, James A.},
    title = {Why Did the West Extend the Franchise? Democracy,
            Inequality,and Growth in Historical Perspective},
    journal = {The Quarterly Journal of Economics},
    year = {2000},
    volume = {115},
    number = {4},
    pages = {1167--1199},
}
```

BibTeX Books

```
@book{Cox1997,
    title={Making Votes Count: Strategic Coordination in the World's
            Electoral Systems},
    author={Gary W. Cox},
    year={1997},
    volume = {7},
    publisher={Cambridge University Press},
    address = {Cambridge}
}
```

More

For more media types and entry fields see
`http://en.wikipedia.org/wiki/BibTeX`.

Tip: Google Scholar

Google scholar generates BibTeX entries.

On an entry click `Cite > BibTeX`.

For a YouTube how-to see
`https://www.youtube.com/watch?v=SsJSR2b4_qc`.

Sometimes they need to be cleaned a little.

Linking your .bib file.

To link your `.bib` file to your RMarkdown document add to the header:

```
bibliography:
    - BIB_FILE_NAME.bib
    - ANOTHER_BIB_FILE_NAME.bib
```

Note: The files should be in the same directory as your R Markdown file.

Including BibTeX citations in RMarkdown

R Markdown uses Pandoc syntax to include a citation in-text.

General format: `@CITE_KEY`.

So if the cite key is `Box1973` then `@Box1973` will return Box and Tiao (1973) in the text of the presentation document.

Formatting In-Text Citations

Markup	Result
`[@Box1973]`	(Box and Tiao 1973)
`[see @Box1973]`	(see Box and Tiao 1973)
`[see @Box1973, 33-40]`	(see Box and Tiao 1973, 33–40)
`[@Box1973; @Acemoglu2000]`	(Box and Tiao 1973; Acemoglu and Robinson 2000)
`@Box1973 [33-40]`	Box and Tiao (1973, 33–40)

Reference List

A reference list with the full bibliographic details of all cited documents will be automatically created at the end of your document.

Tip: Put `# References` at the very end of your R Markdown document to have a section heading before the reference list.

Need more help

Cheatsheet

For a really good RMarkdown cheatsheet see:
`https://www.rstudio.com/wp-content/uploads/2015/03/rmarkdown-reference.pdf`.

Homework Exercises

Homework Exercises

1. Convert what work you have done on your first assignment for this course (or any other work that involved R) to R Markdown and output it to a `.pdf` AND a `.docx` document. Your document has to contain contain the code chunks used to produce e.g. the plots.
2. Create a basic R Markdown presentation with 5 slides that includes at least one plot generated in R (again display the code chunks). Save the presentation as a `.pdf` and `.html`.

Submit: one `.rmd` file with code to produce a the .pdf and the `.docx` file, the `.pdf` and `.docx` files and the `.rmd`, the `.pdf` and the `.html` files for the presentation.

Deadline: Sunday, November 19 before midnight.

References

Acemoglu, Daron, and James A. Robinson. 2000. "Why Did the West Extend the Franchise? Democracy, Inequality, and Growth in Historical Perspective." *The Quarterly Journal of Economics* 115 (4): 1167–99.

Box, G. E. P., and G. C. Tiao. 1973. *Bayesian Inference in Statistical Analysis*. New York: Wiley Classics.