Workshop Series

Open Science Skills in R

Brought to you by

Christelinda Laureijs
Julia Riley
Elizabeth Stregger

What is Open Science?

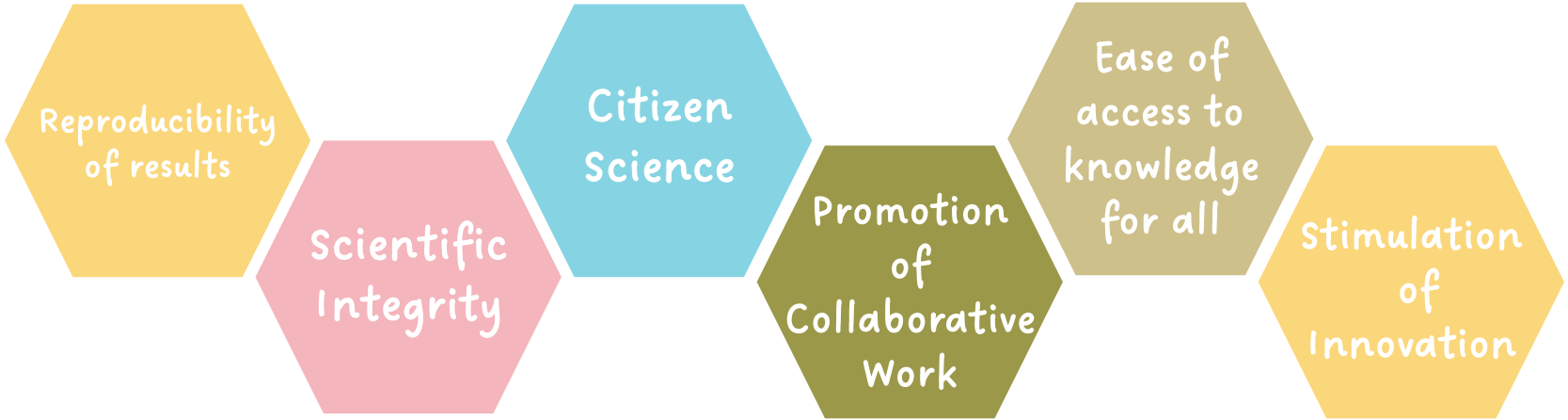
The process of making the content and process of producing evidence and claims transparent and accessible to others.

Munafo et al. 2017

What is Open Science?

The process of making the content and process of producing evidence and claims transparent and accessible to others.

SUPPORTS:



Munafo et al. 2017

About this Workshop Series

THE LEADERS

Christelinda Laureijs

M.Sc. Candidate
in Biology

Elizabeth Stregger

Data and Digital
Services Librarian

Dr. Julia Riley

Assistant
Professor

We all love coding in R and open science!!!

About this Workshop Series

THE SERIES

Wednesdays from 4:30-6:30 PM in FILL IN and fix time

Workshop # 1

29 Jan 2025

Welcome & Being Tidy

Dr. Riley

Workshop # 2

5 Feb 2025

Git with it!

Elizabeth Stregger

Workshop # 3

12 Feb 2025

Science Writing in R

Christelinda Laureijs

About this Workshop Series

WHAT WILL YOU LEARN?

 # 1
Welcome & Being Tidy

 # 2
Git with it!

 # 3
Science Writing in R

Learn about how to set up a project in R, download free tools to work with R, and organise your projects, code, and data.

About this Workshop Series

WHAT WILL YOU LEARN?

1
Welcome & Being Tidy

Learn about how to set up a project in R, download free tools to work with R, and organise your projects, code, and data.

2
Git with it!

Understand how to use a tool called Git which will help you keep track of changes to the files in your project.

3
Science Writing in R

About this Workshop Series

WHAT WILL YOU LEARN?

1
Welcome & Being Tidy

Learn about how to set up a project in R, download free tools to work with R, and organise your projects, code, and data.

2
Git with it!

Understand how to use a tool called Git which will help you keep track of changes to the files in your project.

3
Science Writing in R

Learn how to put it all together and write a paper with R with publication-quality plots and statistical tables.

About this Workshop Series

WHAT CAN YOU EXPECT?

# 1 Welcome & Being Tidy	# 2 Git with it!	# 3 Science Writing in R
Both a mix of lecture and activities.		- 45 min hybrid lecture - 45 min in-person activitu

About this Workshop Series

WHAT CAN YOU EXPECT?

 # 1
Welcome & Being Tidy

 # 2
Git with it!

 # 3
Science Writing in R

Both a mix of lecture and activities.

- 45 min hybrid lecture
- 45 min in-person activity

One person will lead each workshop, and the two others will be "floaters". If you have an issue or question, put a RED post-it note on top of your laptop. Floaters will be by to help you out!


WARNING!
LIVE
TROUBLESHOOTIN
@allison_horst

Open Science Skills in R - A Workshop Series

Welcome to R & Tidy Practices for Reproducibility

R comics by Allison Horst

In Today's Workshop...

WE WILL...

- Download free tools to work with R,
- learn about how to set up a project in R, and
- and organise your projects, code, and data...

IN WAYS THAT PROMOTE OPEN SCIENCE

Let's Get Set Up in R

OUR SOFTWARE TOOLKIT FOR THESE WORKSHOPS

☑ Excel
(or other spreadsheet software)

Let's Get Set Up in R

OUR SOFTWARE TOOLKIT FOR THESE WORKSHOPS

☑ Excel
(or other spreadsheet software)

☑ R Software

- Open-source statistical programming language
- Also an environment for statistical computing and graphics that is easily extendable using *packages*

Let's Get Set Up in R

OUR SOFTWARE TOOLKIT FOR THESE WORKSHOPS

☑ R Studio

- R Studio is a convenient interface for R called an IDE (integrated development environment; e.g., *"I write R code in the R Studio IDE"*)
- It is not a requirement for programming with R, but it is very commonly used by data scientists

Let's Get Set Up in R

OUR SOFTWARE TOOLKIT FOR THESE WORKSHOPS

☑ R Studio

- R Studio is a convenient interface for R called an IDE (integrated development environment; e.g., *"I write R code in the R Studio IDE"*)
- It is not a requirement for programming with R, but it is very commonly used by data scientists

depending on your operating system you also may need to download..

☑ R Tools

Let's Get Set Up in R

OUR SOFTWARE TOOLKIT FOR THESE WORKSHOPS

☑ tinytex

- This facilitates creation of PDF documents in R

Let's Get Set Up in R

OUR SOFTWARE TOOLKIT FOR THESE WORKSHOPS

☑ tinytex

- This facilitates creation of PDF documents in R

RESTART

Let's Get Set Up in R

OUR SOFTWARE TOOLKIT FOR THESE WORKSHOPS

☑ tinytex

- This facilitates creation of PDF documents in R

RESTART

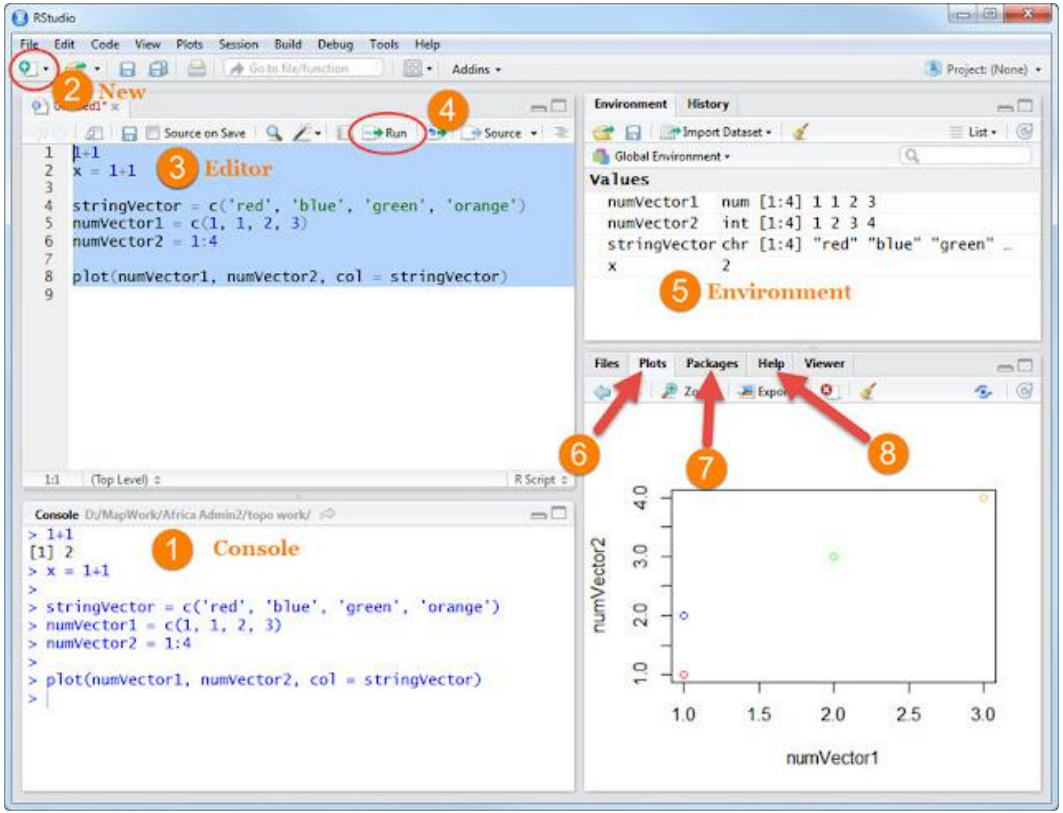
☑ A variety of R Packages

These are the fundamental units of reproducible R code that are made up of functions, documentation on how to use them, and sample data.

Whickham and Bryan 2021


R
#rstats
#RLadies
WELCOME
Horst '18

Quick Tour of RStudio



Introducing Rmarkdown

- **rmarkdown** and the packages that support it enable R users to write their code and prose in reproducible computational documents
- We will generally refer to R Markdown documents (with .Rmd extension; e.g., *"Do this in your R Markdown document"*) and rarely discuss loading this package

Why are we using Rmarkdown?

- **rmarkdown** can be easily used to generate fully reproducible reports - each time you 'knit' the document the analysis is run from the beginning

Why are we using Rmarkdown?

- **rmarkdown** can be easily used to generate fully reproducible reports - each time you 'knit' the document the analysis is run from the beginning

WAIT! **What is reproducibility?**

Why are we using Rmarkdown?

- **rmarkdown** can be easily used to generate fully reproducible reports - each time you 'knit' the document the analysis is run from the beginning

WAIT! What is reproducibility?

Reproducible :

Replicable

Why are we using Rmarkdown?

- rmarkdown can be easily used to generate fully reproducible reports - each time you 'knit' the document the analysis is run from the beginning

WAIT! What is reproducibility?

Reproducible =
Same result can be independently reached given the same data and analytical pipeline

Replicable

g

Why are we using Rmarkdown?

- **rmarkdown** can be easily used to generate fully reproducible reports - each time you 'knit' the document the analysis is run from the beginning

WAIT! What is reproducibility?

Reproducible =
Same result can be independently reached given the same data and analytical pipeline

Replicable =
Same result can be reached given a different, independent data and analytical pipeline

Why are we using Rmarkdown?

- **rmarkdown** can be easily used to generate fully reproducible reports - each time you 'knit' the document the analysis is run from the beginning

WAIT! What is reproducibility?

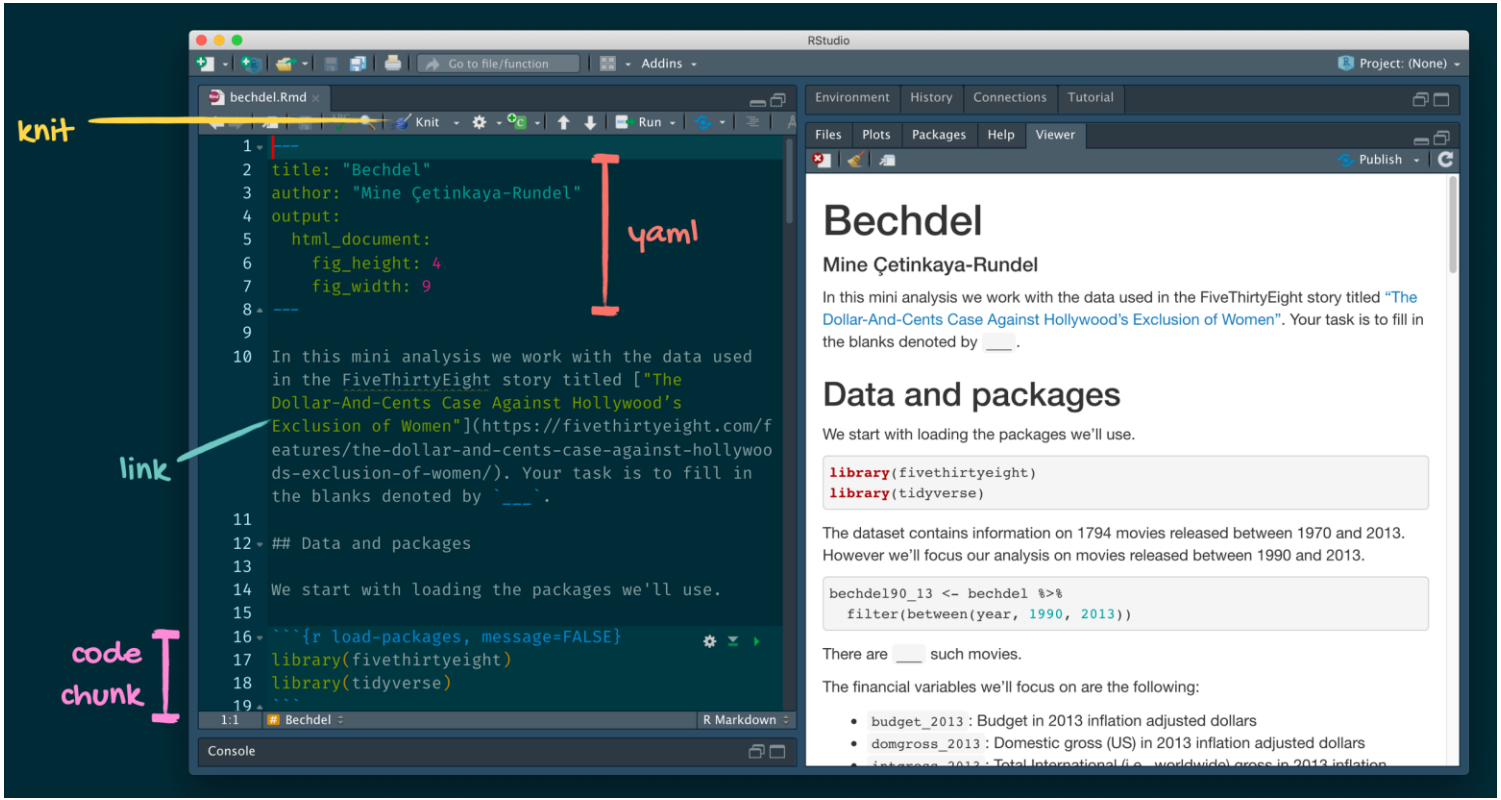
Reproducible =
Same result can be independently reached given the same data and analytical pipeline

Replicable =
Same result can be reached given a different, independent data and analytical pipeline

Why do we care about reproducibility?

- Ethical; science is a public good
- Maximizes translation and utility of your work
- Sets your science on a strong foundation that works against fraud & retraction
- Huge practical benefits for collaboration
  - Easier to share and resuse your work
  - Mistakes are easier to find
  - Analyses and manuscripts easier to update
  - Minimizes duplication of your efforts

Back to Rmarkdown: Tour & Tips



From Data Science in a Box by Dr. Çetinkaya-Rundel

Back to Rmarkdown: Tour & Tips

Simple R markdown syntax for writing text:

Input:

```
This is a sentence in R Markdown, containing `code`,
**bold text**, and *italics*.
```

Output:

This is a sentence in R Markdown, containing code, **bold text**, and *italics*.

Back to Rmarkdown: Tour & Tips

Code is placed in 'chunks', defined by three backticks, and the narrative goes outside of those 'chunks':

Input:

````
The function 'rnorm()' creates normal variates.
```{R}
rnorm(5) # creates 5 normal variates
```
````

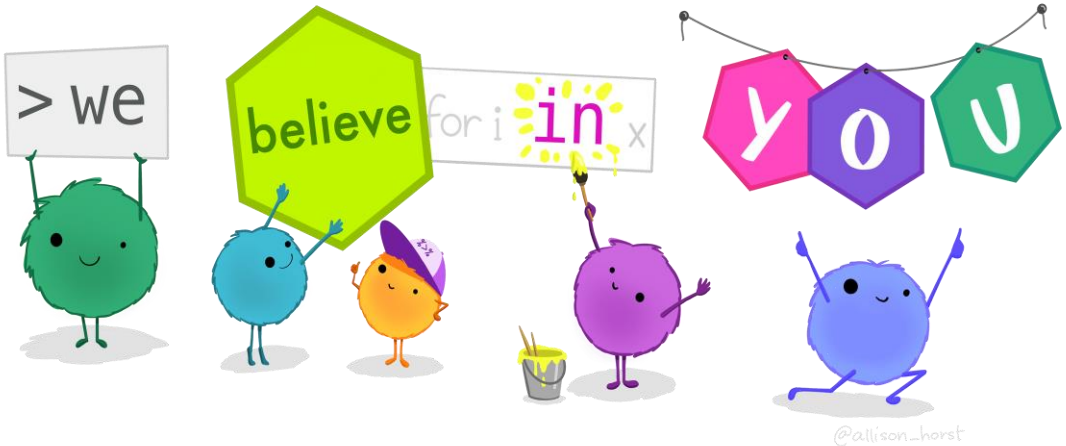
Output:

The function `rnorm()` creates normal variates.

```
rnorm(5) # creates 5 normal variates
```

```
[1] 1.1281735 1.7376142  0.7629712  1.1308147  0.9969855
```

Any Questions?



Break Time

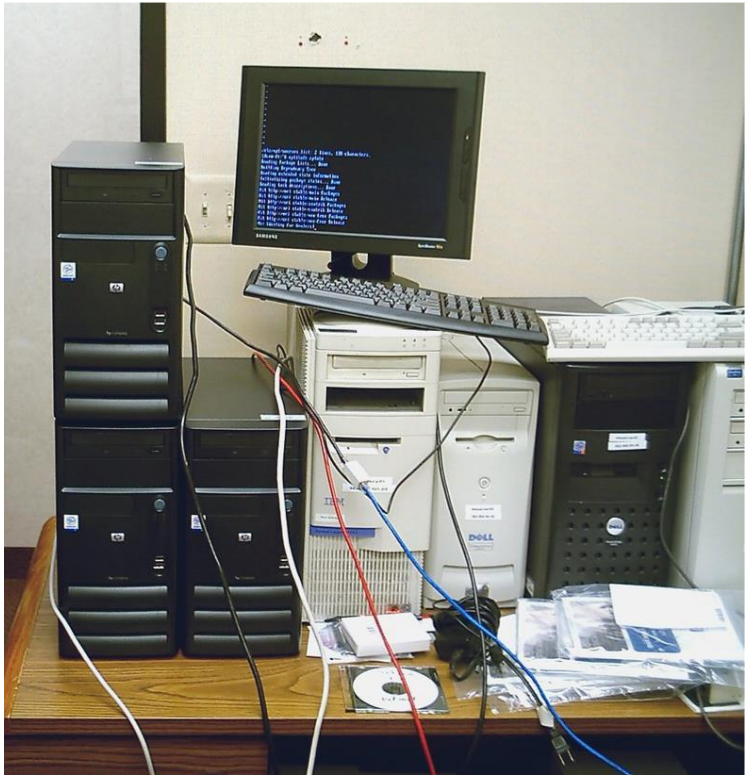
TIDY PRACTICES

for efficiency, reproducibility, & collaboration

"Tidy" organized projects are the foundation of reproducible, efficient, and open science!

- ☐ Tidy projects
- ☐ Tidy code
- ☐ Tidy data

What is a Tidy Project?

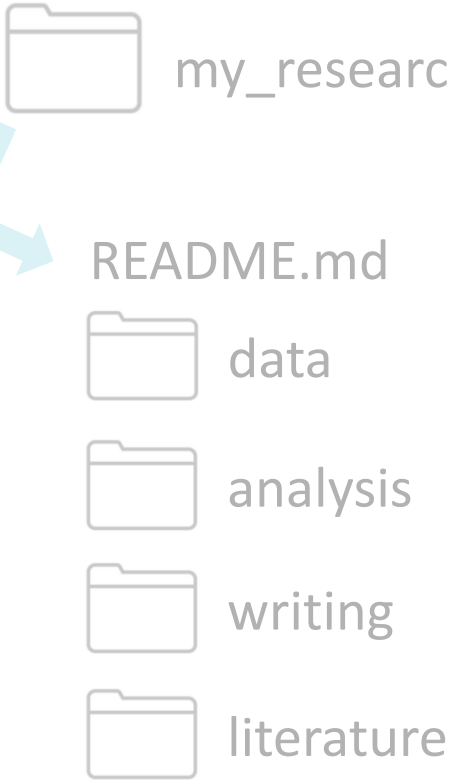


R transcect

Shared Folder

Name	Date Modified	Size	Kind
function for plotting outlines.R	22/04/2013 1:44 PM	4 KB	R Source File
area_output_example2222.csv	22/04/2013 1:44 PM	15 KB	comm…values
area_output_example.csv	22/04/2013 1:41 PM	15 KB	comm…values
image_outlines_complete	22/04/2013 1:40 PM	--	Folder
area_output_march13.txt	22/04/2013 1:27 PM	13 KB	Plain…cument
outlines_area_calculations (#1) function.R	22/04/2013 11:58 AM	4 KB	R Source File
LIZARD 1_COMPLETE	18/04/2013 11:58 AM	--	Folder
database_previous versions	18/04/2013 11:52 AM	--	Folder
Concept sheets	18/04/2013 11:51 AM	--	Folder
area_output.csv	17/04/2013 4:16 PM	15 KB	comm…values
sc_database_areas.csv	17/04/2013 4:14 PM	46 KB	comm…values
fragmentation w histograms.pdf	03/04/2013 9:02 AM	364 KB	Portab…(PDF)
summary plots and glms.docx	02/04/2013 4:04 PM	1.5 MB	Micro…ument
outlines_area_calculations (#1).R	28/03/2013 4:16 PM	4 KB	R Source File
outline_files	28/03/2013 3:16 PM	--	Folder
area_script_current.R	14/03/2013 11:46 AM	11 KB	R Source File
size histograms 1989-1992.pdf	14/03/2013 10:56 AM	329 KB	Portab…(PDF)
sc_database_areas_march13.txt	13/03/2013 7:55 AM	27 KB	Plain…cument
do corals lie about their age.pdf	27/02/2013 8:27 AM	2.1 MB	Portab…(PDF)
Fitting a power midel.R	21/02/2013 3:55 PM	2 KB	R Source File
R scripts for maps	20/02/2013 9:45 AM	--	Folder
trial_photos_outline_maps	20/02/2013 8:24 AM	--	Folder
Technical note/ Cu…g pdfkeywords.pdf	19/02/2013 10:55 AM	396 KB	Portab…(PDF)
old files	15/02/2013 6:08 PM	--	Folder
trial outline files_first files	13/02/2013 8:53 AM	--	Folder
image_data	05/02/2013 1:54 PM	--	Folder
images	05/02/2013 1:53 PM	--	Folder

What is a Tidy Project?



THREE GUIDELINES

1. Make it self-contained
2. Create a consistent, sensibly-named directory structure
3. Include a README file that includes information about the layout and contents of your project

What is a Tidy Project?

ANOTHER EXAMPLE

- **informative_project_name**
  - **README.txt** (a text file at the top-level of the directory that outlines the broad structure/details of the project)
  - **/data** (raw data, such as images, videos, datasheets, as well as the processed products for analysis)
  - **/doc** (all notes and the draft manuscript associated with the project)
  - **/figs** (figures to be included in the manuscript, typically generated via code)
  - **/output** (items generated from data handling and analysis, like tables of statistical results, which can be regenerated at any time)
  - **/code** (code for processing and analysing data)

TIDY PROJECT

DISCUSSION

1) Who is a stakeholder that may have an interest in outcomes of your research (aka. who will benefit from open science)?

2) What is one positive and one negative consequence of conducting open, reproducible research? Let's discuss.

TIDY PROJECTS

DISCUSSION

1) Who is a stakeholder that may have an interest in outcomes of your research (aka. who will benefit from open science)?

2) What is one positive and one negative consequence of conducting open, reproducible research? Let's discuss.

ACTIVITY

Create a tidy project template for a research project. Do this in Rmarkdown and make use of their 'list' syntax to keep it clear.

Four Steps to Tidy Code

1) Choose good names and be consistent

Good:

```
dat_mass_2020 <- read.csv('2020_field_data_mass.csv')
```

Less good (maybe):

```
dat_field <- read.csv('2020_field_data_mass.csv')
```

Bad:

```
dat <- read.csv('2020_field_data_mass.csv')
```

Four Steps to Tidy Code

1) Choose good names and be consistent

2) Write human-readable code

Space it out

Good:

```
height <- cm * 6 + mm
mean(x, na.rm = TRUE)
```

Bad:

```
height<-cm*6+mm
mean(x,na.rm=TRUE)
```

Four Steps to Tidy Code

1) Choose good names and be consistent

2) Write human-readable code

Space it out

Good:

```
do_something_complicated(
    something = “that”,
    requires = many,
    arguments = “that are very long”
)
```

Bad:

```
do_something_complicated(“that”, needs, many, arguments, “that are very long”)
```

Four Steps to Tidy Code

1) Choose good names and be consistent
2) Write human-readable code
3) Make use of your Tidy Project set-up

Use relative file paths, never absolute.

- Forget `setwd()` ever existed
- Assume the file is being run from the folder it is sitting in
- Use paths *relative* to that

Four Steps to Tidy Code

1) Choose good names and be consistent
2) Write human-readable code
3) Make use of your Tidy Project set-up

Use relative file paths, never absolute.

Bad

```
data <-read.csv('C:/juliascomupter/projects/toadexp/data/behaviour_data.csv')
```

- Won't run on any computer
- Also won't run on ANY computer if I move the file or change the path to it!

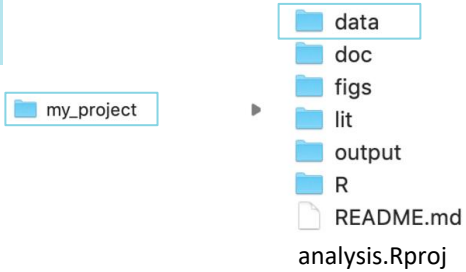
Four Steps to Tidy Code

1) Choose good names and be consistent
2) Write human-readable code
3) Make use of your Tidy Project set-up

Use relative file paths, never absolute.

Good

```
data <- read.csv("my_project/data/behaviour_data.csv")
```



Four Steps to Tidy Code

1) Choose good names and be consistent
2) Write human-readable code
3) Make use of your Tidy Project set-up
4) Keep it well-styled and use help

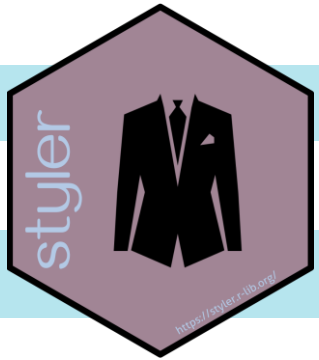
R package - styler

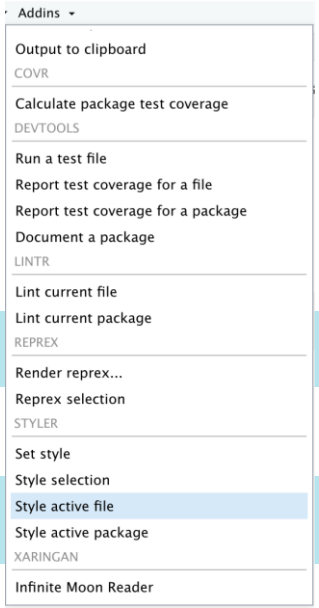
Before

```
height<-cm*6+mm+2; mean(x,na.rm=TRUE)
```

After

```
height <- cm * 6 + mm + 2
mean(x, na.rm = TRUE)
```





TIDY CODE

ACTIVITY

In the Rmarkdown file for this workshop, please work through the activity using the R package `styler`.

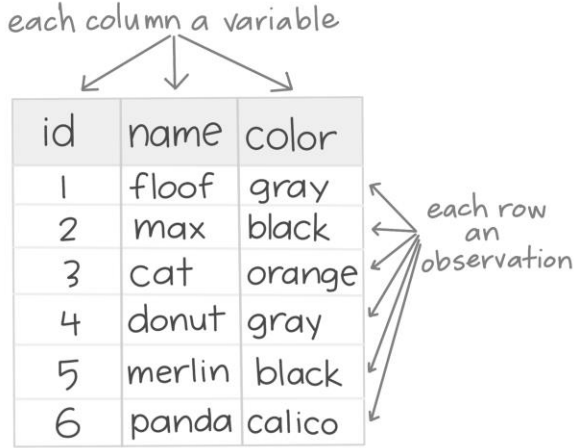
What is different about the code after it is reformatted using `styler`?

What is Tidy Data?

> "TIDY DATA is a standard way of mapping the meaning of a dataset to its structure."
>
> –HADLEY WICKHAM

In tidy data:

- each variable forms a column
- each observation forms a row
- each cell is a single measurement



Whickham 2014; Lowndes and Horst 2020

An Example

How could we tidy this messy data?

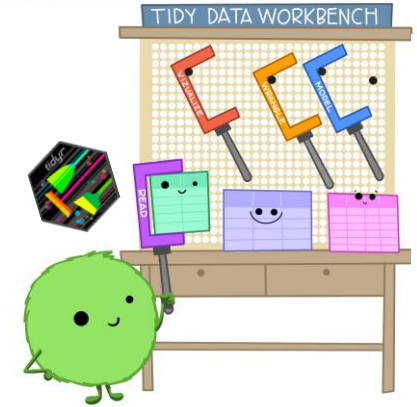
religion	<$10k	$10-20k	$20-30k	$30-40k	$40-50k	$50-75k
Agnostic	27	34	60	81	76	137
Atheist	12	27	37	52	35	70
Buddhist	27	21	30	34	33	58
Catholic	418	617	732	670	638	1116
Don't know/refused	15	14	15	11	10	35
Evangelical Prot	575	869	1064	982	881	1486
Hindu	1	9	7	9	11	34
Historically Black Prot	228	244	236	238	197	223
Jehovah's Witness	20	27	24	24	21	30
Jewish	19	19	25	25	30	95

religion	income	freq
Agnostic	<$10k	27
Agnostic	$10-20k	34
Agnostic	$20-30k	60
Agnostic	$30-40k	81
Agnostic	$40-50k	76
Agnostic	$50-75k	137
Agnostic	$75-100k	122
Agnostic	$100-150k	109
Agnostic	>150k	84
Agnostic	Don't know/refused	96

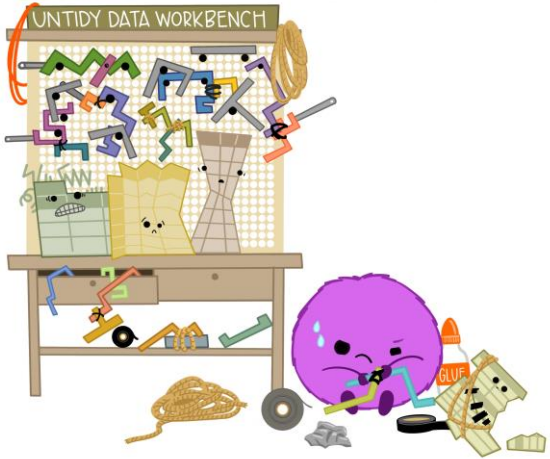
Why use a TIDY format?

1) Allows you to be more efficient

When working with tidy data, we can use the same tools in similar ways for different datasets...



...but working with untidy data often means reinventing the wheel with one-time approaches that are hard to iterate or reuse.

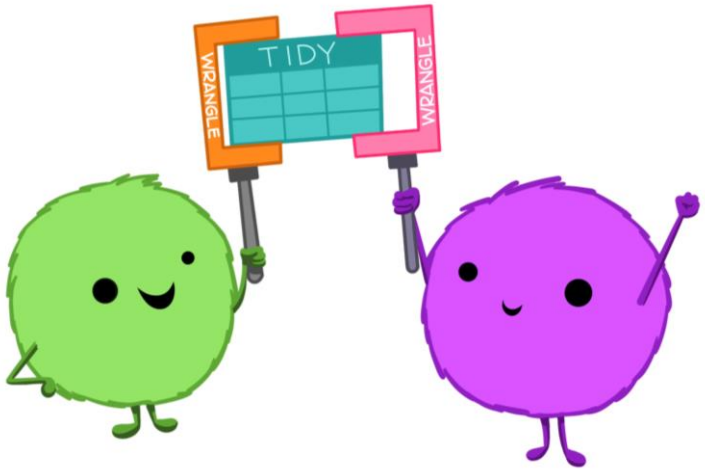


Whickham 2014; Lowndes and Horst 2020

Why use a TIDY format?

1) Allows you to be more efficient

2) Easier to collaborate because it is known what to expect & can make use of the same tools



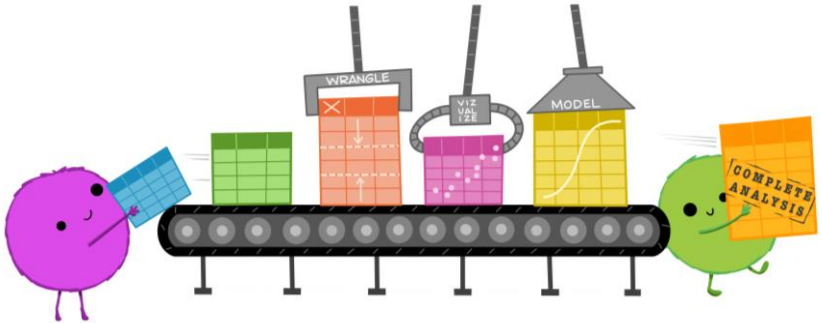
Whickham 2014; Lowndes and Horst 2020

Why use a TIDY format?

1) Allows you to be more efficient

2) Easier to collaborate because it is known what to expect & can make use of the same tools

3) Makes it easier to reproduce analyses



Whickham 2014; Lowndes and Horst 2020

Seven Tips to Keep Data Tidy

- Use plain text

Versions across the ages...

Microsoft Excel	Text
.xls	.txt
.xlt	
.xlm	
.xlam	
.xltm	
.xlsx	
...	

Seven Tips to Keep Data Tidy

- Use plain text

Types of Text File

.csv – Comma-separated values. Great all-purpose format.

.txt or **.tsv** – Plain-text/tab-deliminated. Future proof.

Can all be opened with anything/anywhere.

Seven Tips to Keep Data Tidy

- Use plain text
- Choose good names

Messy

myabstract.docx
Julia's best ideas.docx
figure 1.png
newNEWv2_dontdelete_forREAL.xlsx
FINALfinal_v2_5.xlsx

Tidy

2020_abstract_hons_conf.docx
julias_ideas.docx
fig_01_scatterplot_length_width.png2019-08-07_raw_data_hons.xlsx

Seven Tips to Keep Data Tidy

- Use plain text
- Choose good names

  These are: machine-readable

Don't use: Special characters or formatting

`! @ # $ % ^ & * ( ) ~ + =`

Do use: Underscores and dashes for `separating_metadata` and `splitting-up-words`

Seven Tips to Keep Data Tidy

- Use plain text
- Choose good names

  These are: machine-readable

  human-readable

Make sure that names contain information about **content**

Nay: `data 1.csv`

Yay: `2020-08-09_field-data_morphology.csv`

Seven Tips to Keep Data Tidy

- Use plain text
- Choose good names

  These are: machine-readable
  human-readable
  nicely ordered 

Think about how your file names will sort.

Chronological

```
2020-08-09_field-data_morphology.csv
2020-08-12_field-data_morphology.csv
2020-08-18_field-data_morphology.csv
```

Seven Tips to Keep Data Tidy

- Use plain text
- Choose good names

  These are: machine-readable
  human-readable
  nicely ordered

Think about how your file names will sort.

Logical

```
01_load_functions.R
02_clean_data.R
03_analysis.R
```

Seven Tips to Keep Data Tidy

- Use plain text
- Choose good names
- No empty cells or special characters!

More questions than answers!

cow_ID	milk_volume	weight
moo	12	1100
bumbo	2	1201
spot	?	1084
jeffrey		1044
holy	16	1244
daisy	-	1093

Seven Tips to Keep Data Tidy

- Use plain text
- Choose good names
- No empty cells or special characters!

So, use NA if NA or 0 if 0.

cow_ID	milk_volume	weight
moo	12	1100
bumbo	2	1201
spot	NA	1084
jeffrey	0	1044
holy	16	1244
daisy	0	1093

Seven Tips to Keep Data Tidy

- Use plain text
- Choose good names
- No empty cells or special characters!
- Use metadata (aka. a data dictionary)

DATA

emlployee_id	first_name	last_name	nin	department_id
44	Simon	Martinez	HH 45 09 73 D	1
45	Thomas	Goldstein	SA 75 35 42 B	2
46	Eugene	Cornelsen	NE 22 63 82	2
47	Andrew	Petculescu	XY 29 87 61 A	1
48	Ruth	Stadick	MA 12 89 36 A	15
49	Barry	Scardelis	AT 20 73 18	2
50	Sidney	Hunter	HW 12 94 21 C	6
51	Jeffrey	Evans	LX 13 26 39 B	6
52	Doris	Berndt	YA 49 88 11 A	3
53	Diane	Eaton	BE 08 74 68 A	1
54	Bonnie	Hall	WW 53 77 68 A	15
55	Taylor	Li	ZE 55 22 80 B	1

Seven Tips to Keep Data Tidy

METADATA

Column	Data Type	Description
emlployee_id	int	Primary key of a table
first_name	nvarchar(50)	Employee first name
last_name	nvarchar(50)	Employee last name
nin	nvarchar(15)	National Identification Number
position	nvarchar(50)	Current postion title, e.g. Secretary
department_id	int	Employee depamntnet. Ref: Departmetns
gender	char(1)	M = Male, F = Female, Null = unknown
employment_start_date	date	Start date of employment in organization.
employment_end_date	date	Employment end date. Null if employee st

- Metadata = data about data
- A file describing the contents and structure of a data file
- The more detail the better
- Essential to reproducibility!

DATA

emlployee_id	first_name	last_name	nin	department_id
44	Simon	Martinez	HH 45 09 73 D	1
45	Thomas	Goldstein	SA 75 35 42 B	2
46	Eugene	Comelsen	NE 22 63 82	2
47	Andrew	Petculescu	XY 29 87 61 A	1
48	Ruth	Stadick	MA 12 89 36 A	15
49	Barry	Scardelis	AT 20 73 18	2
50	Sidney	Hunter	HW 12 94 21 C	6
51	Jeffrey	Evans	LX 13 26 39 B	6
52	Doris	Berndt	YA 49 88 11 A	3
53	Diane	Eaton	BE 08 74 68 A	1
54	Bonnie	Hall	WW 53 77 68 A	15
55	Taylor	Li	ZE 55 22 80 B	1

Seven Tips to Keep Data Tidy

- Use plain text
- Choose good names
- No empty cells or special characters!
- Use metadata (aka. a data dictionary)
- Treat raw data as read-only

HANDS OFF!

Modification by hand:

- Avoid as much as possible
- If you have to, create a copy to work on & document every change you make in a separate file

Preferentially modify by code!

Seven Tips to Keep Data Tidy

- Use plain text
- Choose good names
- No empty cells or special characters!
- Use metadata (aka. a data dictionary)
- Treat raw data as read-only
- Be consistent

...WITH YOUR NAMING CONVENTIONS

- snake_case
- camelCase
- SCREAMING_SNAKE_CASE
- kebab-case
- Train-Case

PICK-one_AndUse_ItCONSISTENTLY

Seven Tips to Keep Data Tidy

- Use plain text
- Choose good names
- No empty cells or special characters!
- Use metadata (aka. a data dictionary)
- Treat raw data as read-only
- Be consistent
- Dates are awful

...SO MANY WAYS TO BE CODED

- MM/DD/YY
- DD/MM/YY
- YY/MM/DD
- DD-MM-YYYY
- MM-YY

What can we do with this mess?

Seven Tips to Keep Data Tidy

- Use plain text
- Choose good names
- No empty cells or special characters!
- Use metadata (aka. a data dictionary)
- Treat raw data as read-only
- Be consistent
- Dates are awful

OPTION # 1

Split up the variables:

day	month	year
4	01	2015
25	01	2014
27	11	2010

OPTION # 2

Use the ISO standard: **YYYY-MM-DD**

TIDY DATA

ACTIVITY

We will be spending the remainder of the workshop, making a messy dataset tidy. There is a description of the data and a link to download it in this workshop's Rmarkdown file.

The mission is simple - focus all of your tidy skills on the catastrophe that is **messy survey.xls** to parse it into its cleanest, tidiest, and most useful self.

Document any adjustments you make. We will chat about them to wrap up this workshop.

ANY LAST QUESTIONS?

Thank you.