DATA WRANGLING AND REPRODUCIBILITY IN 

DR. COLLEEN FORTIER

BIOINFORMATICS OFFICE HOUR – JULY 24, 2025

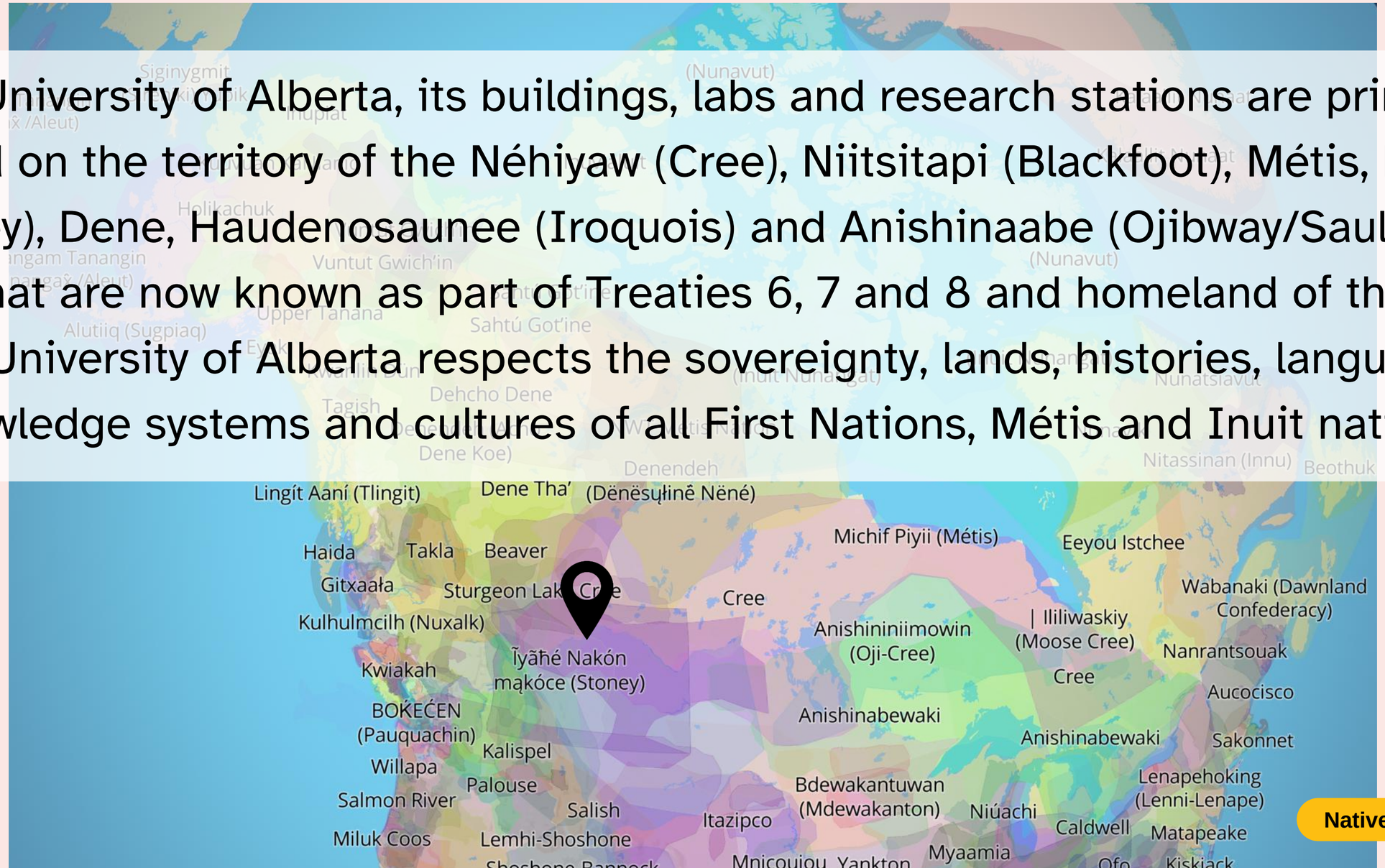
LAND ACKNOWLEDGEMENT

Native-land.ca

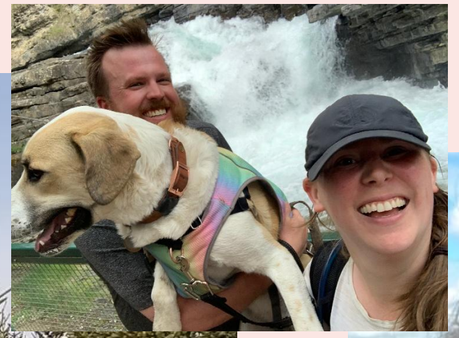
LAND ACKNOWLEDGEMENT

The University of Alberta, its buildings, labs and research stations are primarily located on the territory of the Néhiyaw (Cree), Niitsitapi (Blackfoot), Métis, Nakoda (Stoney), Dene, Haudenosaunee (Iroquois) and Anishinaabe (Ojibway/Saulteaux), lands that are now known as part of Treaties 6, 7 and 8 and homeland of the Métis.

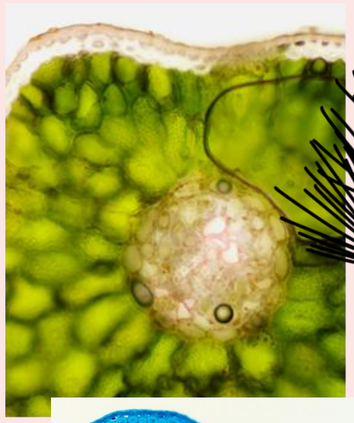
The University of Alberta respects the sovereignty, lands, histories, languages, knowledge systems and cultures of all First Nations, Métis and Inuit nations.

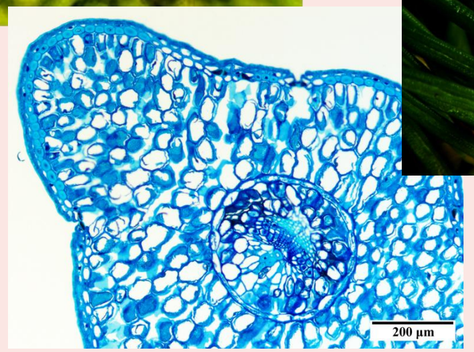


A BIT ABOUT ME



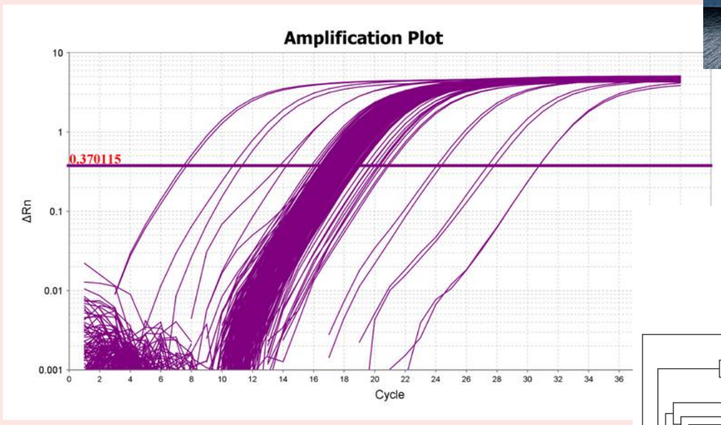
A BIT ABOUT ME

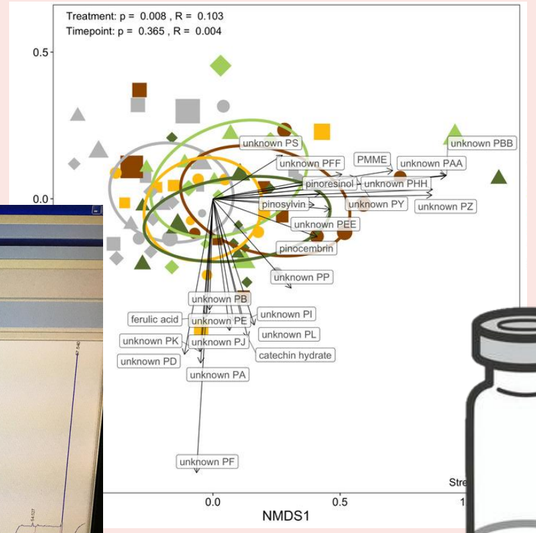


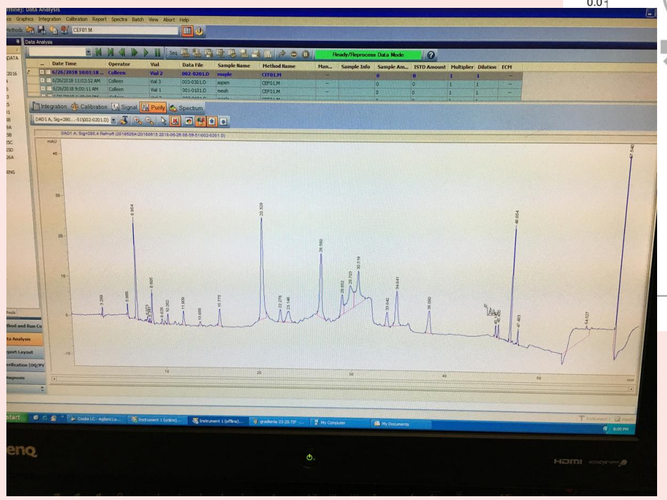


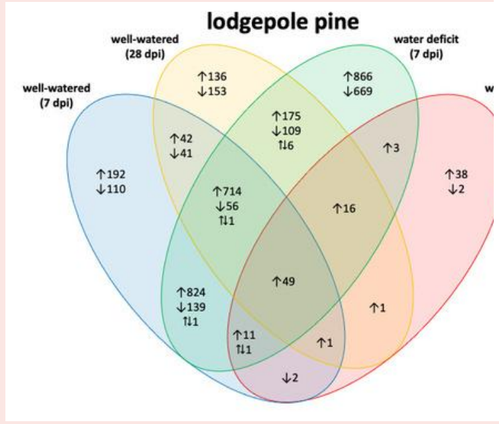
microscopy

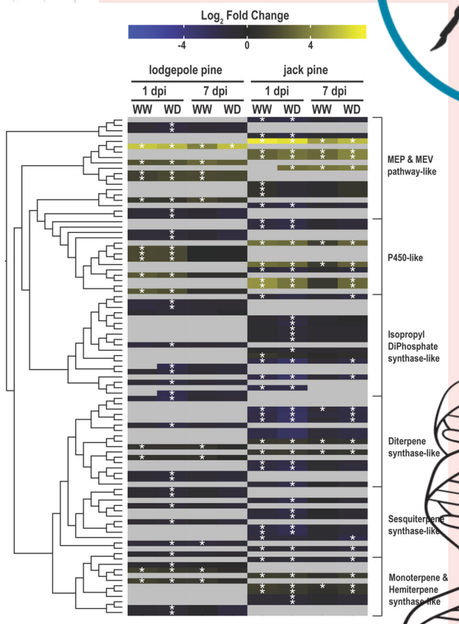
transcriptomics

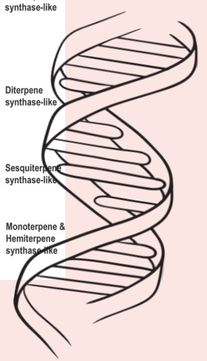












metabolomics

BEST DATA PRACTICES

Findable

Accessible

Interoperable

Reusable

Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

DOI

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

R is great for this!

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. *Sci Data* **3** (2016) https://doi.org/10.1038/sdata.2016.1818

HAVING A PLAN

Tri-Agency Research Data Management Policy

4. Implementation dates

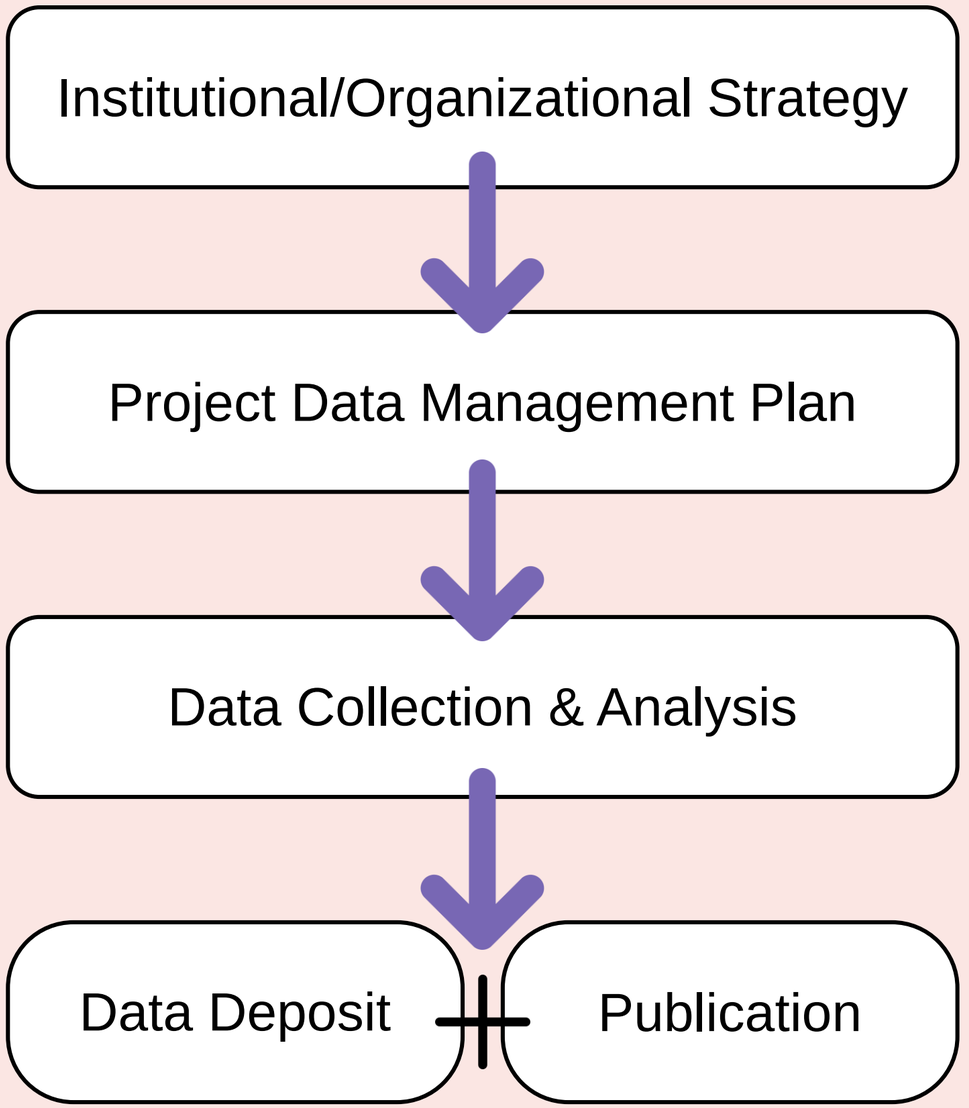
The agencies plan to implement the policy incrementally, as determined through ongoing engagement with the research community and other stakeholders, and in step with the continuing development of research data practices and capacity in Canada and internationally.

- **Institutional strategies:** By March 1, 2023, research institutions subject to this requirement must post their RDM strategies and notify the agencies when they have been completed.
- **Data management plans:** By spring 2022, the agencies will identify the initial set of funding opportunities subject to the DMP requirement. The agencies will pilot the DMP requirement in targeted funding opportunities before this date.
- **Data deposit:** After reviewing the institutional strategies and in line with the readiness of the Canadian research community, the agencies will phase in the deposit requirement.
  - **CIHR-funded researchers:** Since January 1, 2008, recipients of CIHR funding have had to comply with the limited data deposit requirements included in the Tri-Agency Open Access Policy on Publications. They must continue to comply with these requirements, which are specific to bioinformatics, atomic, and molecular coordinate data.

The policy will be updated as the implementation dates for requirements are further specified.

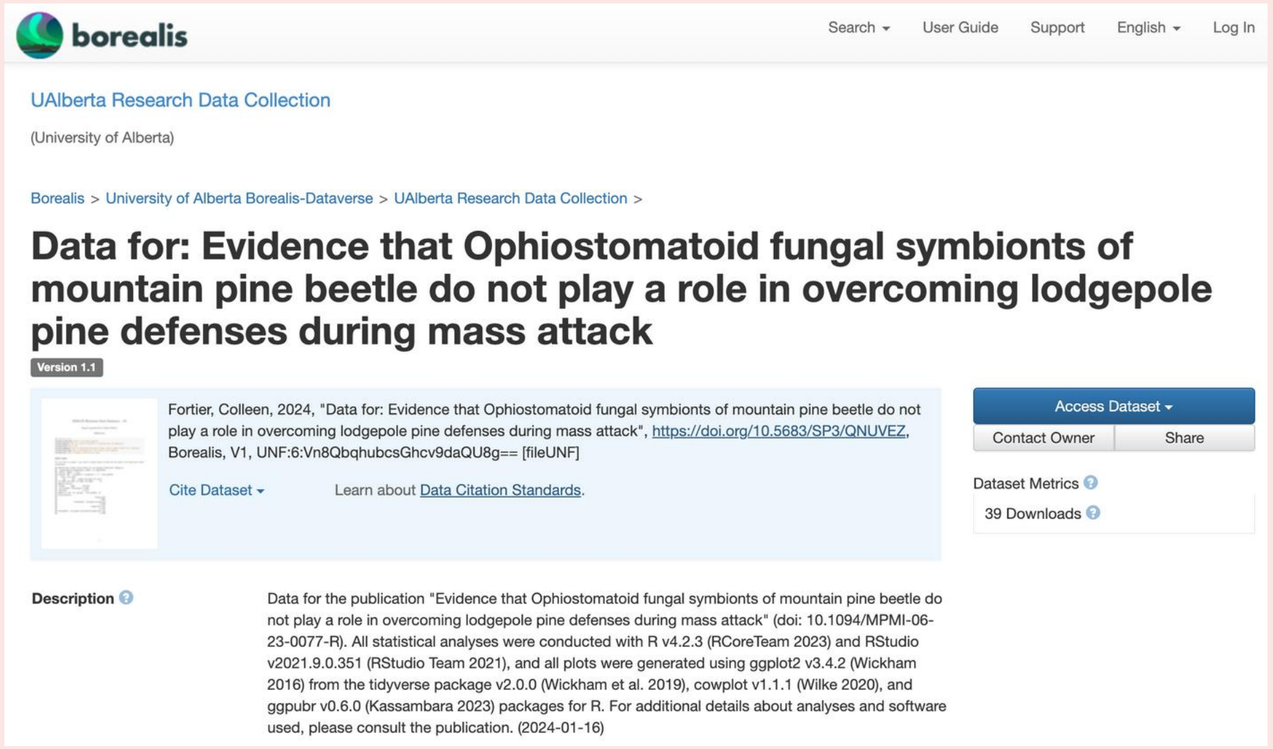
NSERC CRSNG

Genome Canada



MAKING DATA *FINDABLE*

borealis



FRDR
Federated Research Data Repository

NIH NLM
National Center for Biotechnology Information

GitHub

MAKING DATA *ACCESSIBLE*

METADATA!

README file(s): description of the dataset, how was it generated, sharing and access info, acknowledgements

Data Dictionary: defines data terms, column names, codes, abbreviations, acronyms, etc.

File Naming: short and relevent, consider versioning, be consistent

UBC Library Research Commons "How to Create a README File"

Harvard Medical School "File Naming Conventions"

York U "Naming for Electronic Files and Folders"

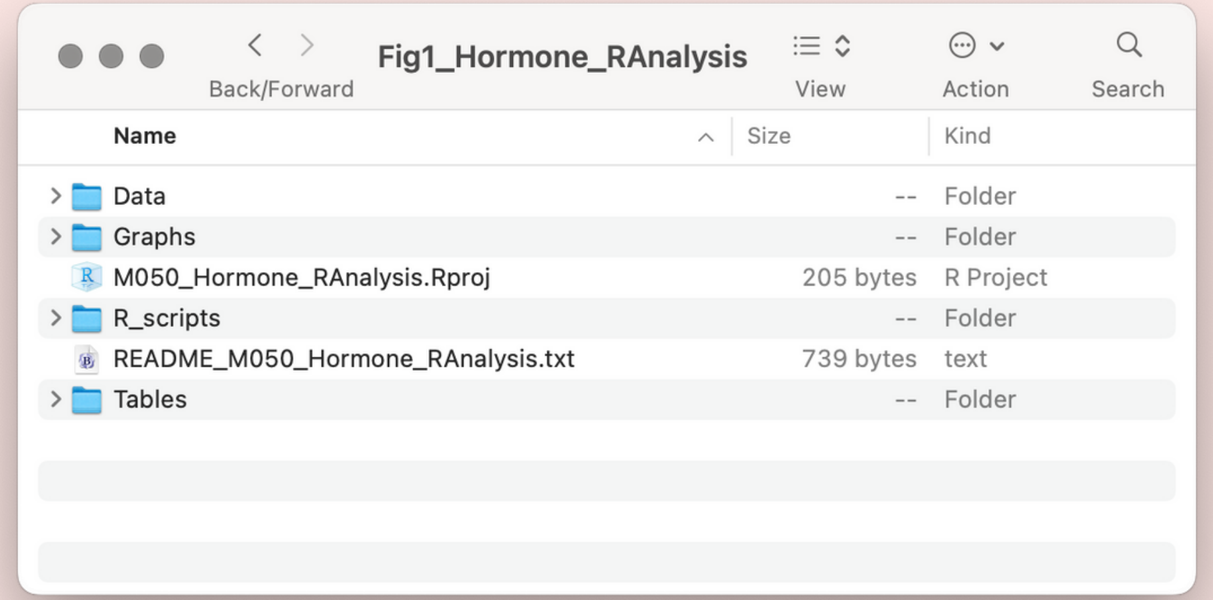
MAKING DATA *ACCESSIBLE*

METADATA!

Directory Structure

Think about this as soon as you collect your data

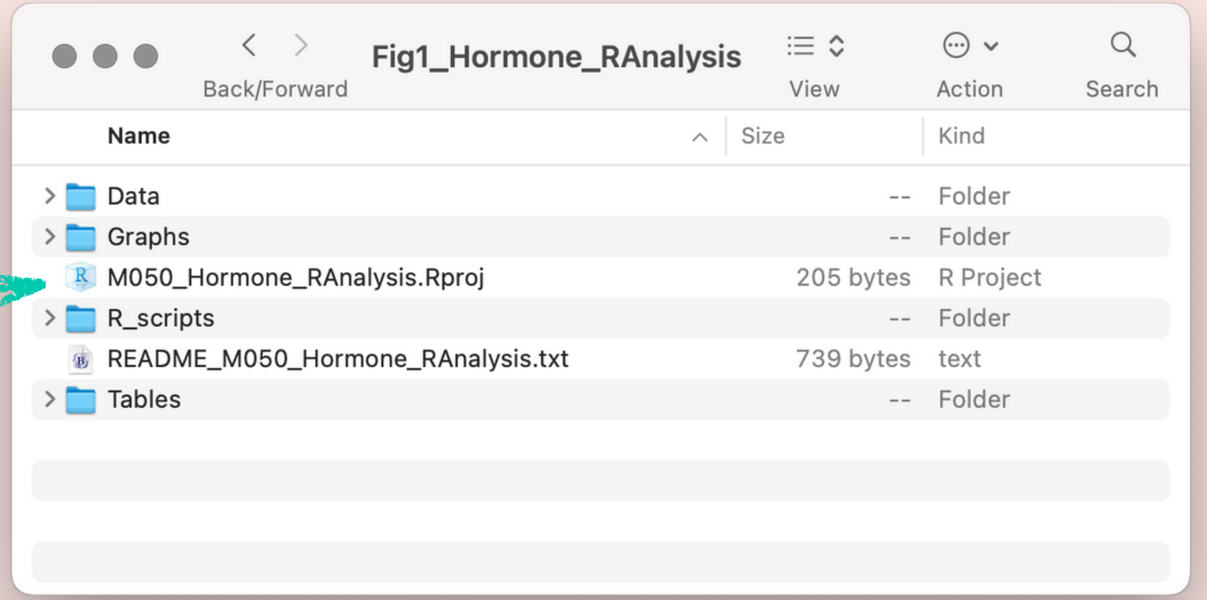
- keep raw data separate, work with copies
- having designated locations for things will prevent data loss!



MAKING DATA *ACCESSIBLE* AND *INTEROPERABLE*

Directory Structure

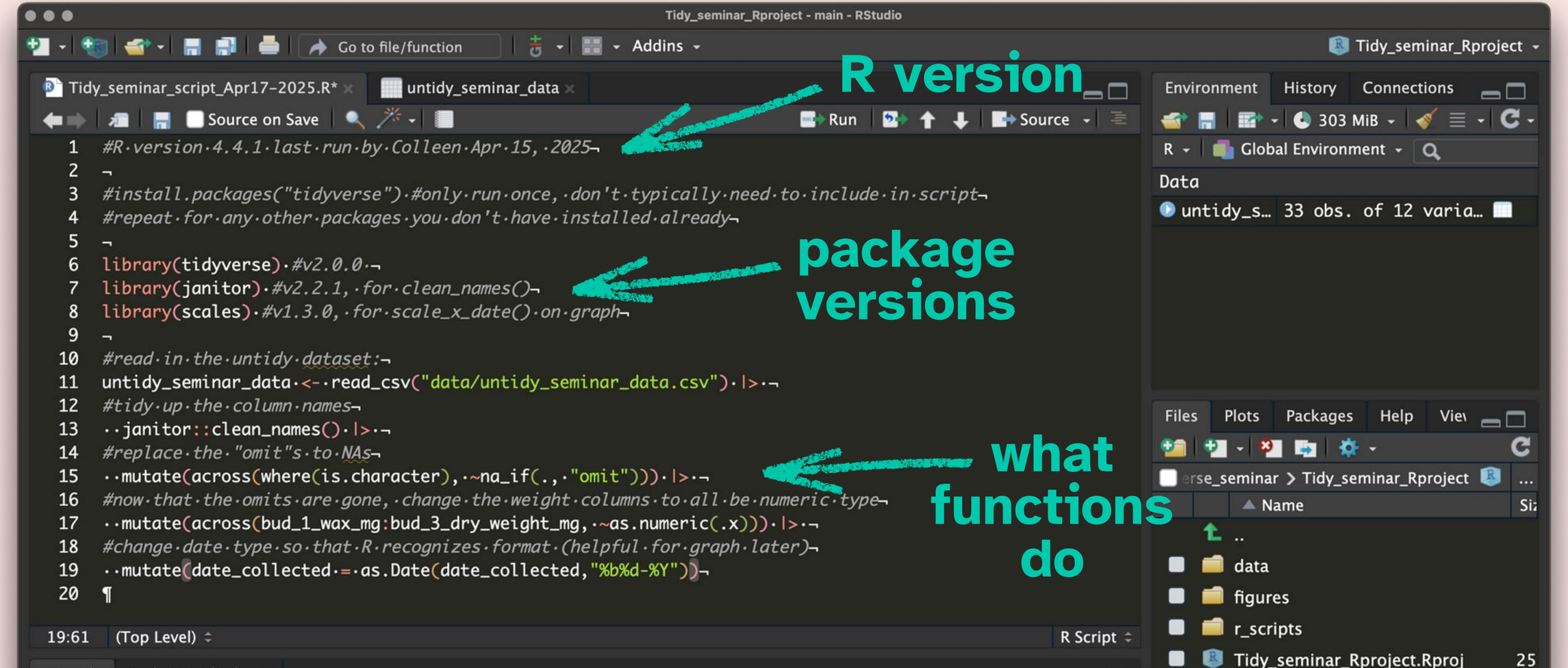
Consider using R Projects!
(require RStudio)



- work with git for version control
- easily shareable with collaborators

MAKING DATA *INTEROPERABLE*

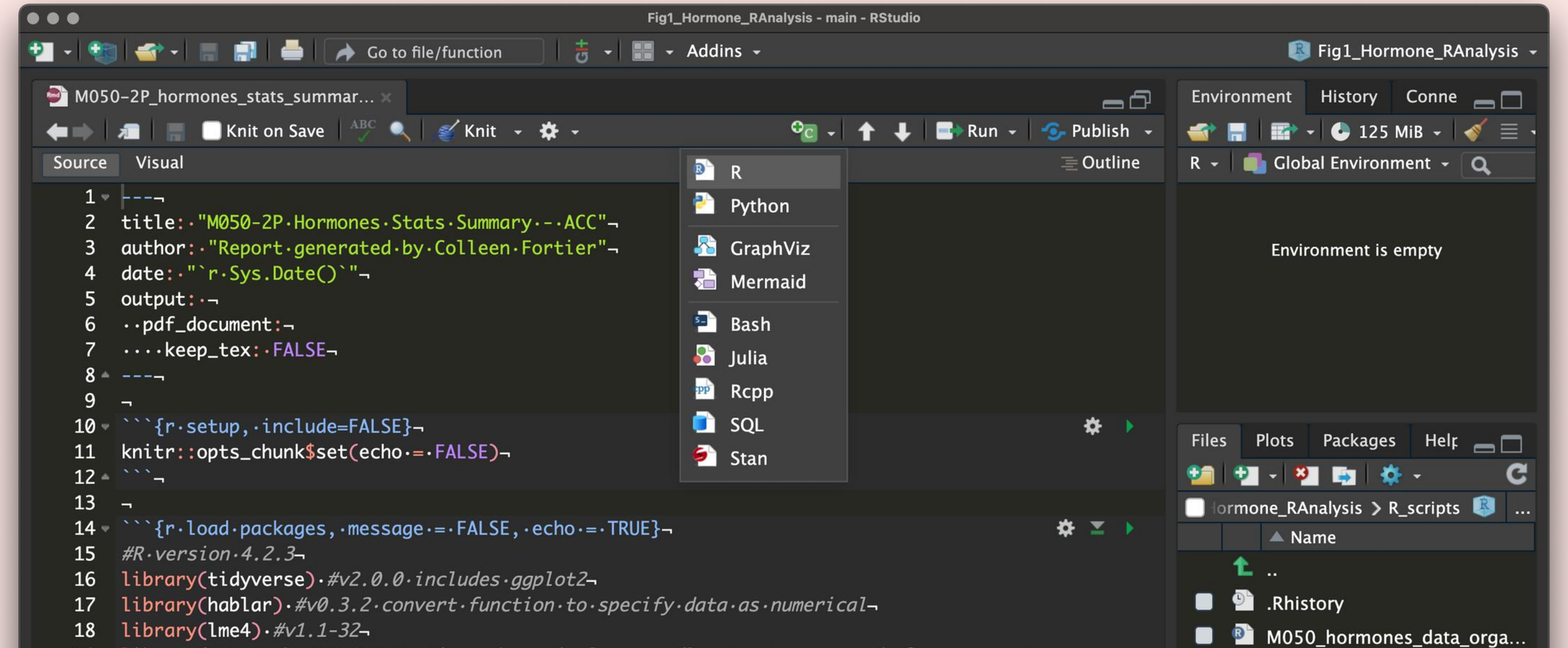
Annotations



MAKING DATA *INTEROPERABLE*

Consider using RMarkdown (in RStudio)

- Can run multiple languages (R, python, bash, SQL) as “chunks”



MAKING DATA *INTEROPERABLE*

Consider using RMarkdown (in RStudio)

- Can “knit” into html or pdf report, showing all code and output

M050-2P_hormones_s... × + Create

Subscribe

All tools Edit Convert E-Sign Find text or tools Share Ask AI Assistant

M050-2P Hormones Stats Summary - ACC

Report generated by Colleen Fortier

2023-04-04

```
#R version 4.2.3
library(tidyverse) #v2.0.0 includes ggplot2
library(hablar) #v0.3.2 convert function to specify data as numerical
library(lme4) #v1.1-32
library(emmeans) #v1.8.5 #estimated marginal means (least square means) for contrasts
library(multcomp) #v1.4-23 #needed for cld results from emmeans
library(car) #v3.1-2 #Anova function for glms
```

INOC GLM

Fit your data to a glmm - may need to change family to better fit the model to the data/meet model assumptions

```
## Generalized linear mixed model fit by maximum likelihood (Adaptive
##   Gauss-Hermite Quadrature, nAGQ = 0) [glmerMod]
##  Family: Gamma  ( log )
## Formula: QTY ~ treatment * timepoint + (1 | tree_number)
##    Data: inoc_data
##       AIC       BIC    logLik  deviance  df.resid
##  285.7404  292.8087 -136.8702  273.7404        18
## Random effects:
##  Groups      Name        Std.Dev.
##  tree_number (Intercept) 0.3531
##  Residual                0.3614
## Number of obs: 24, groups:  tree_number, 16
## Fixed Effects:
```

CHECK THAT YOUR DATA MEETS ASSUMPTION OF YOUR MODEL

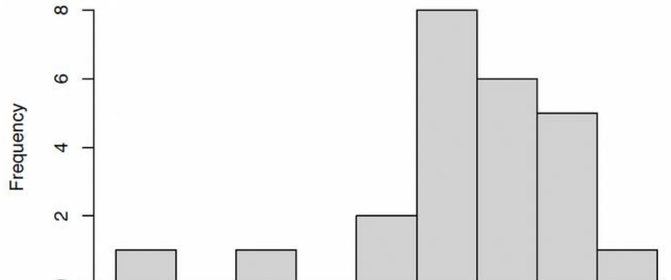
```
##
##  Shapiro-Wilk normality test
##
## data:  inoc_model_residuals
## W = 0.89898, p-value = 0.02048

##
##  Bartlett test of homogeneity of variances
##
## data:  inoc_model_residuals and inoc_data$timepoint
## Bartlett's K-squared = 5.1333, df = 1, p-value = 0.02347

##
##  Bartlett test of homogeneity of variances
##
## data:  inoc_model_residuals and inoc_data$treatment
## Bartlett's K-squared = 0.87047, df = 1, p-value = 0.3508
```

Plots examining distribution of model residuals

These plots use the model residuals rather than the data values to determine fit of the model.

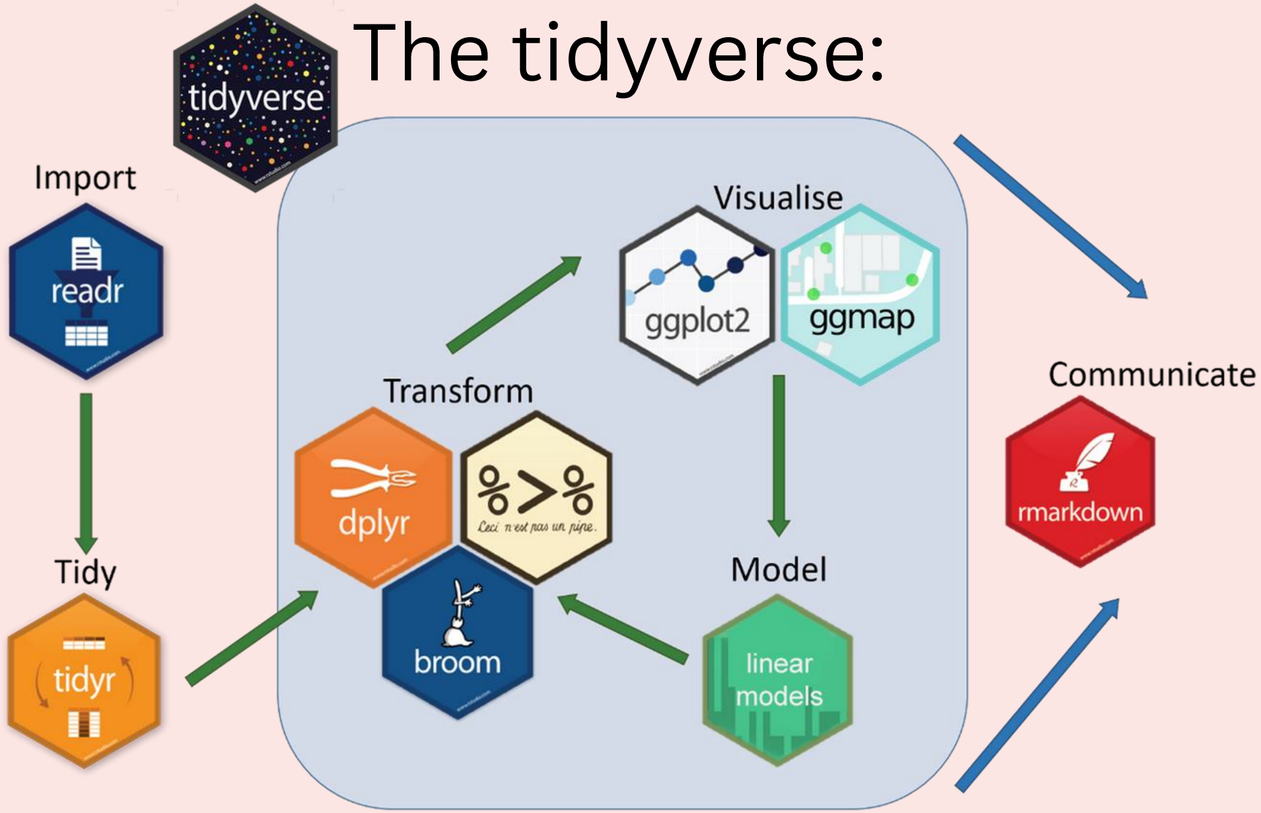


MAKING DATA *RESUABLE*

Use Smarter Code and TIDY YOUR DATA!

- why use many lines when few line do trick? (Kevin, *The Office*)

Slides and RProject with data and scripts to try out tidyverse!



MAKING DATA *RESUABLE*

Use Smarter Code and TIDY YOUR DATA!

- why use many lines when few line do trick? (Kevin, *The Office*)

#Base R:

>names(df)[names(df) == “old_name”] <- “new_name”

#tidyr:

>rename(df, new_name = old_name)

#if your old name has any spaces, make sure to include “ ”

Slides and RProject with data and scripts to try out tidyverse!

MAKING DATA *RESUABLE*

Use Smarter Code and TIDY YOUR DATA!

- why use many lines when few line do trick? (Kevin, *The Office*)

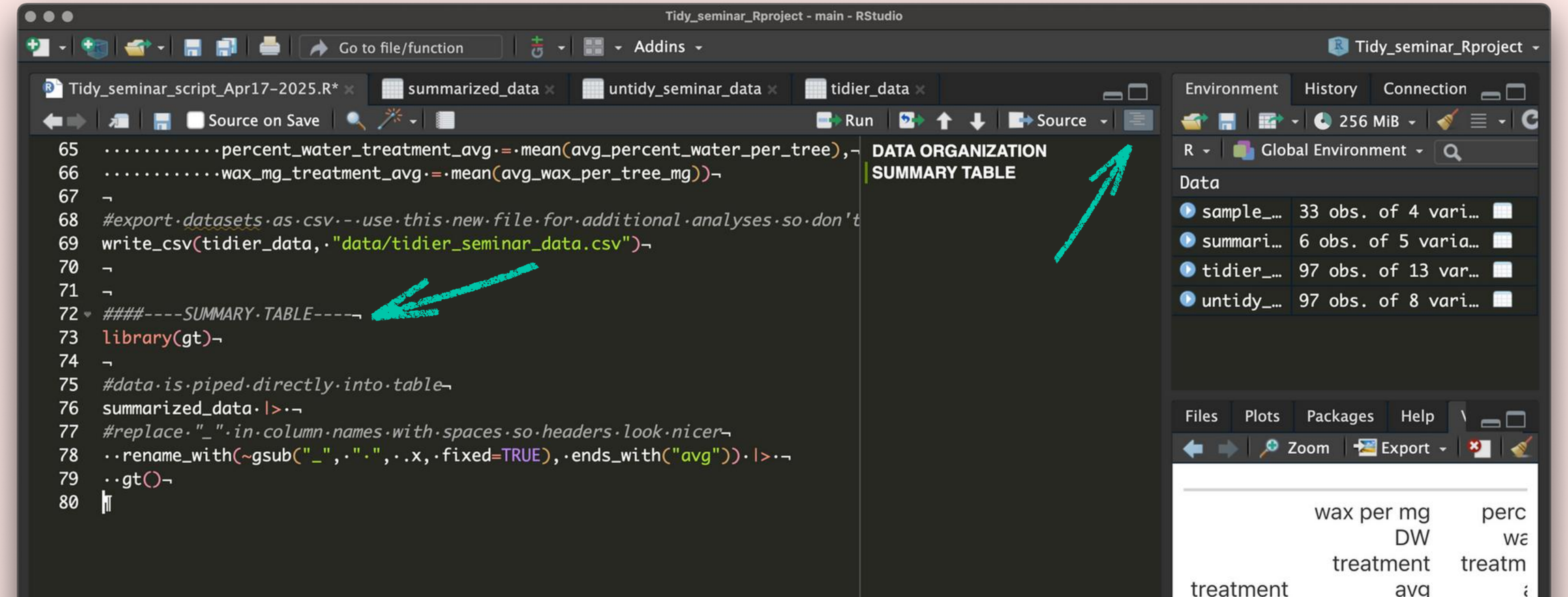
1. Each variable must have its own column
2. Each observation must have its own row
3. Each value must have its own cell

Slides and RProject with data and scripts to try out tidyverse!

MAKING DATA *RESUABLE*

Use Headers and Annotations to break up script

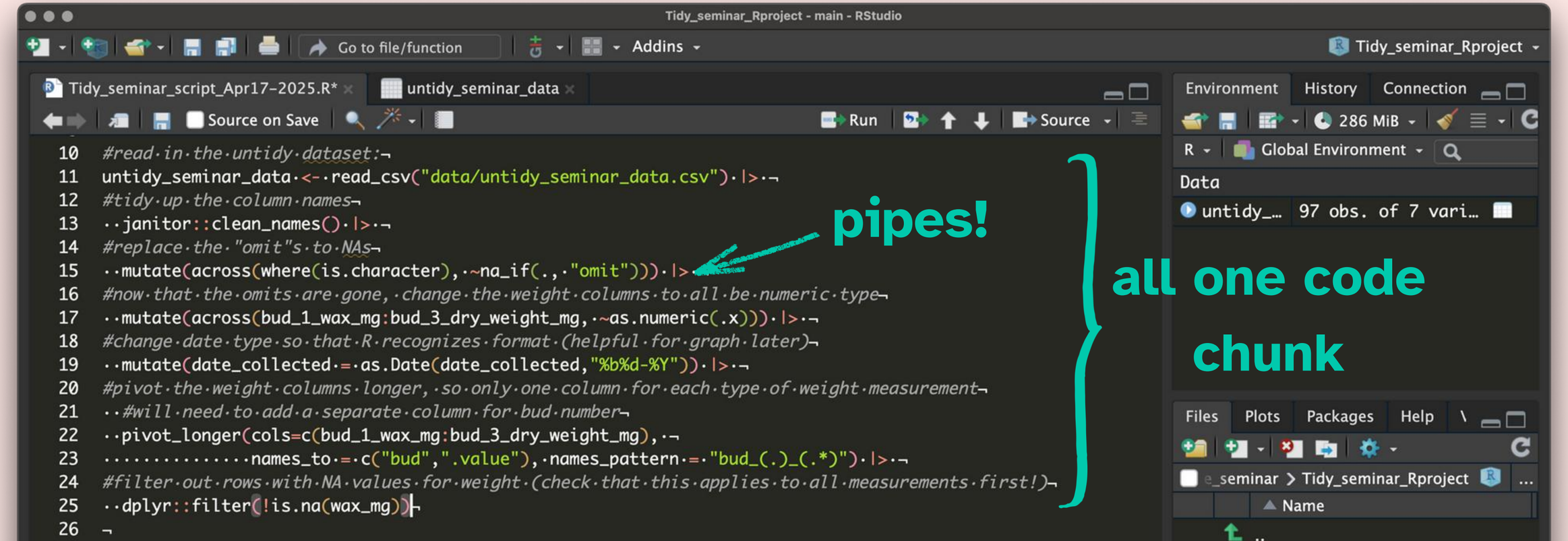
- RStudio recognizes 4 dashes (----) as cues for headers



MAKING DATA *RESUABLE*

Use Headers and Annotations to break up script

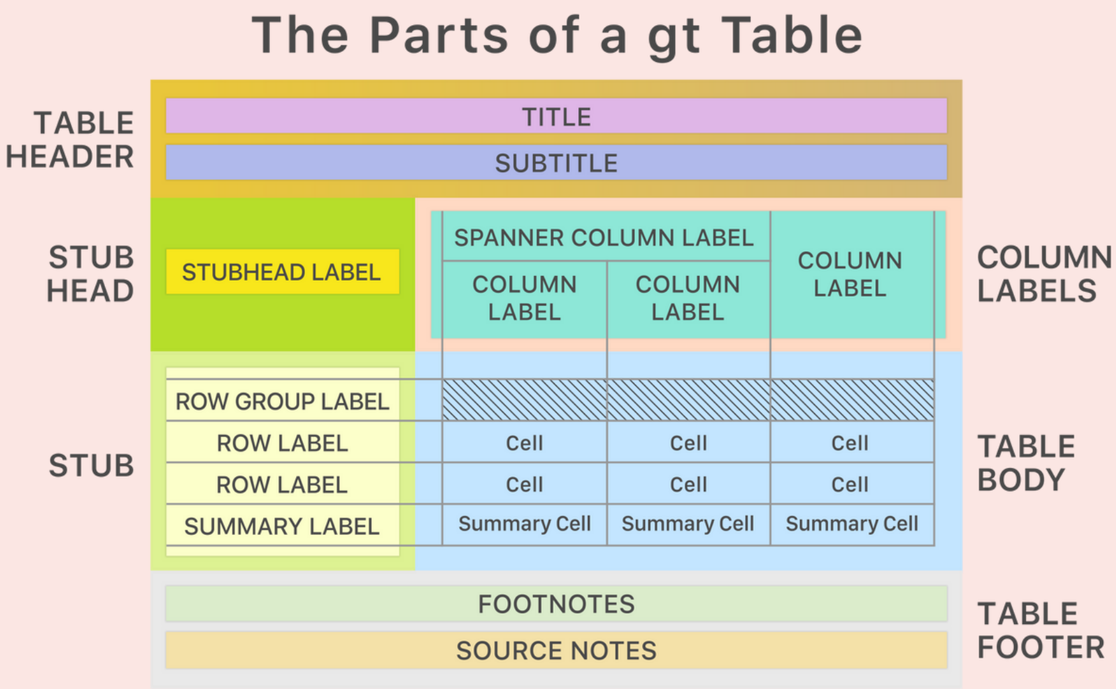
- continuous code lines will still run even if broken up by spaces or annotations - just don't put a # in front of the code you want to run



MAKING DATA *RESUABLE*

Try to code as much as you can

- avoid copy + paste between programs, or editing outputs
- use packages like gt to create results tables IN YOUR SCRIPT



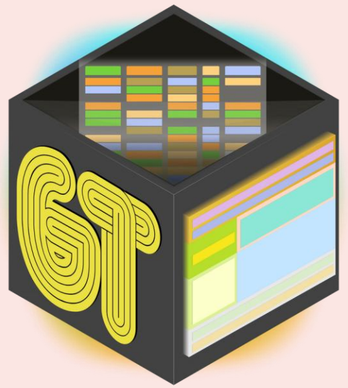
The title of the table

The table's subtitle

	num	char	fctr	date	time	datetime	currency
grp_a							
row_1	[1] 1.111e-01	apricot	one	2015-01-15	13:35	2018-01-01 02:22	49.950
row_2	2.222e+00	banana	two	2015-02-15	14:40	2018-02-02 14:33	17.950
row_3	3.333e+01	coconut	three	2015-03-15	15:45	2018-03-03 03:44	1.390
row_4	4.444e+02	durian	four	2015-04-15	16:50	2018-04-04 15:55	65100.000
min	0.11	—	—	—	—	—	1.39
max	444.40	—	—	—	—	—	65,100.00
grp_b							
row_5	5.550e+03	NA	five	2015-05-15	17:55	2018-05-05 04:00	1325.810
row_6	NA	fig	six	2015-06-15	NA	2018-06-06 16:11	13.255
row_7	7.770e+05	grapefruit	seven	NA	19:10	2018-07-07 05:22	NA
row_8	8.880e+06	honeydew	eight	2015-08-15	20:20	NA	0.440
total	—	—	—	—	—	—	66,508.79

[1] This is a footnote.

This is a source note.

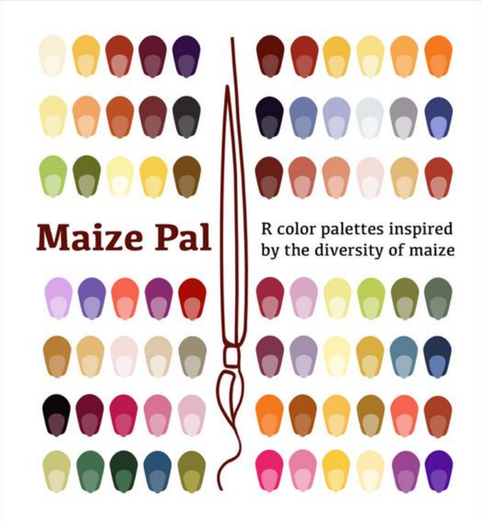


{gt} 0.7.0

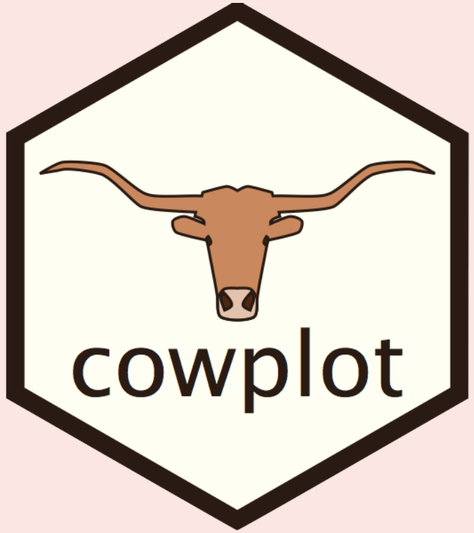
MAKING DATA *RESUABLE*

Try to code as much as you can

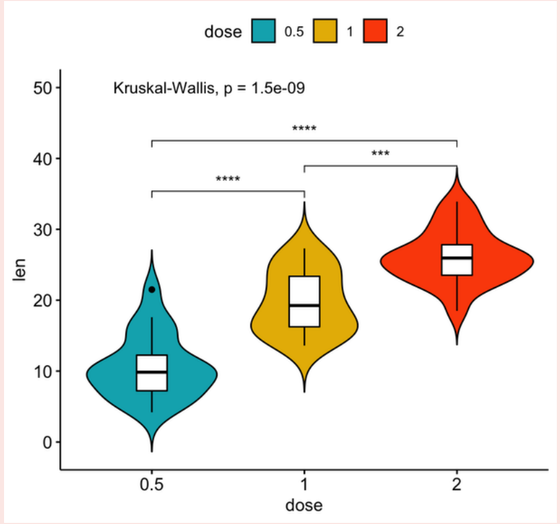
- avoid copy + paste between programs, or editing outputs
- explore packages that build on ggplot to do even more



Maize Pal
color palettes



combine plots, add axes and labels with **cowplot**



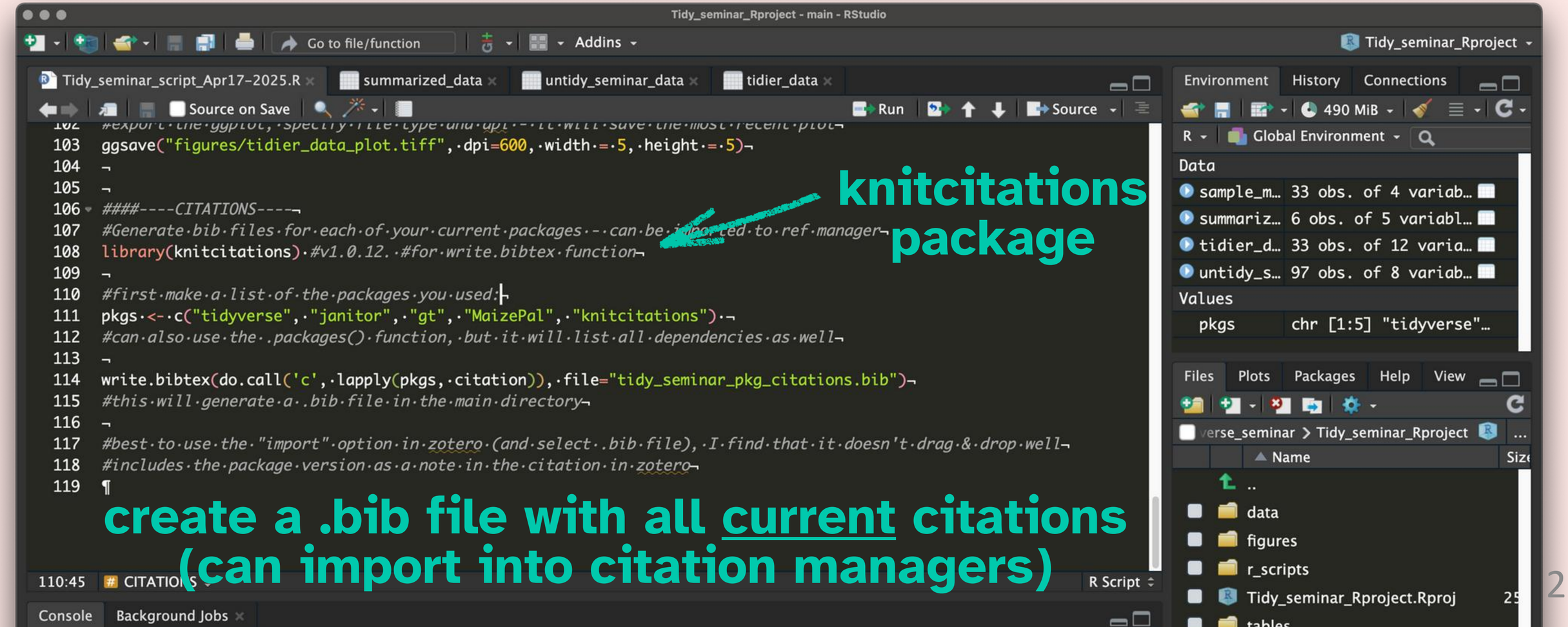
add statistics and regression lines with **ggpubr**

more ideas!

MAKING DATA *RESUABLE*

more on citing packages

Don’t forget to cite your packages (and versions)!



LAST THOUGHTS

- Make a plan - get everyone involved
- REVISIT THE PLAN REGULARLY
- Share your code with others - they are the best at finding errors/where more annotation is needed
- Schedule time periodically to curate your data

ADDITIONAL RESOURCES

My **Github**, where I post presentations + materials

UBC Library Research Commons FREE **Workshops**

these have great resources

including a whole section on Data Management

Digital Research Alliance **YouTube**

and newsletter signup for workshops

ADDITIONAL RESOURCES

- UBC Advanced Research Computing **Workshops**
- UAlberta Research Bootcamp **Workshop Recordings**
- Compute Ontario **Courses**
- SHARCNET **Training**

THANKS FOR LISTENING!

ANY QUESTIONS?

You Have A
New Messages!

OK

Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod

OK

LOADING...

OK