

Introduction to R and Reproducible Research

Melbourne Statistical Consulting Platform

University of Melbourne

April 2024

R

- free and open source software for statistics (data science, machine learning)
- also a programming language
- thousands of add-on packages via the Comprehensive R Archive Network (CRAN)

RStudio

- easy-to-use interactive environment combining R, an editor for R scripts, and general improvements to the quality of your R life

R Markdown

- an add-on to R for producing documents
- mix text, R code and R output (including tables and figures) within a document
- these slides were produced using R Markdown

Tidyverse

- an "opinionated collection of R packages" for working with data, sharing a common design philosophy
- includes `ggplot2`, `tidyr`, `dplyr` and others...

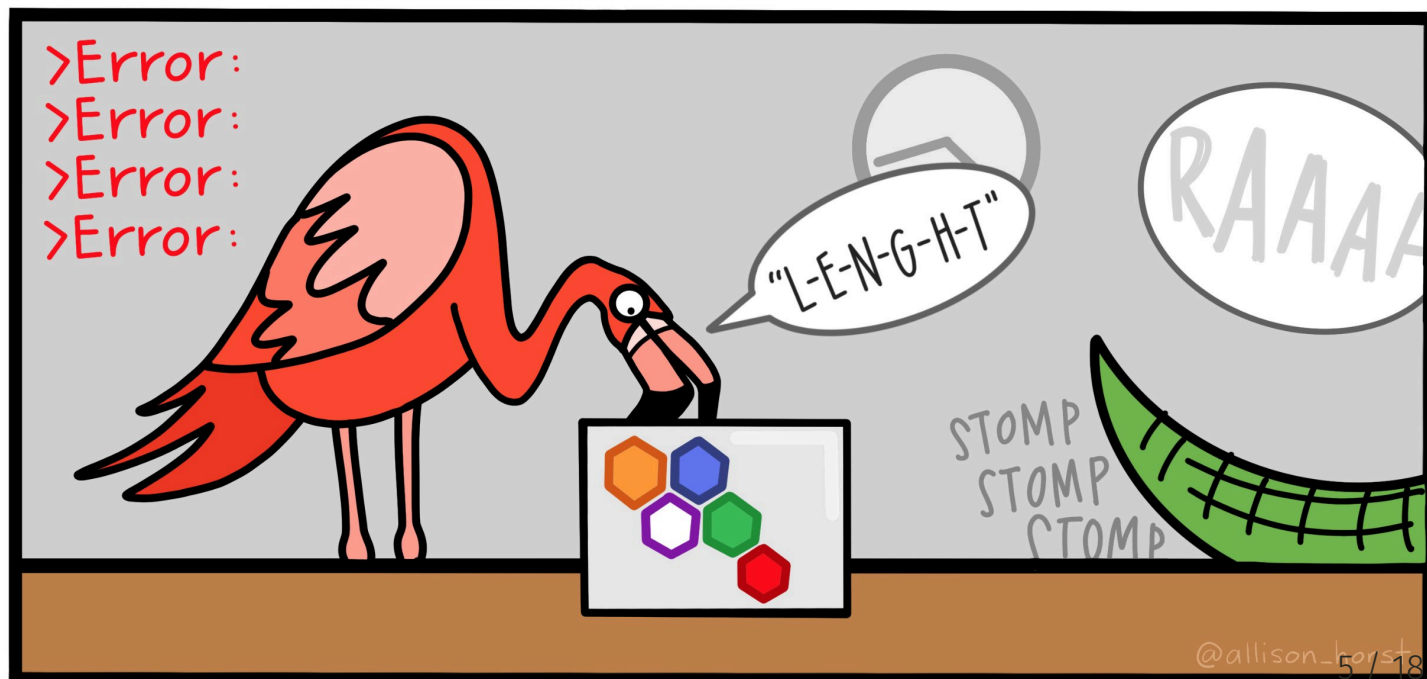
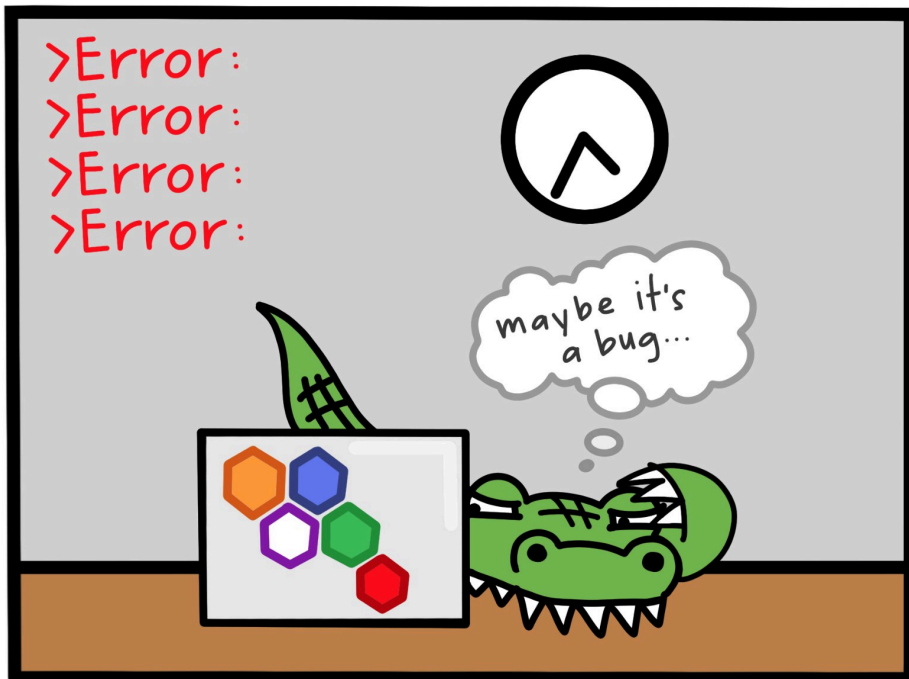
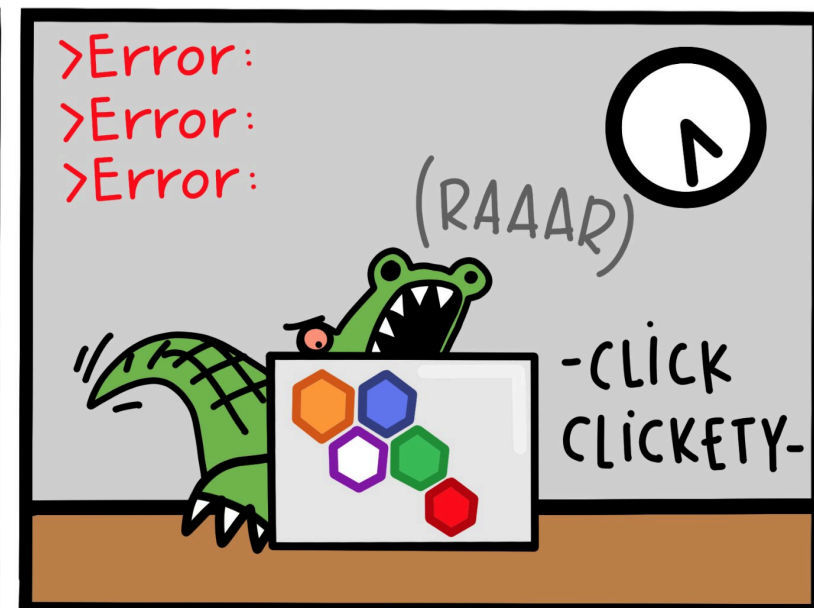
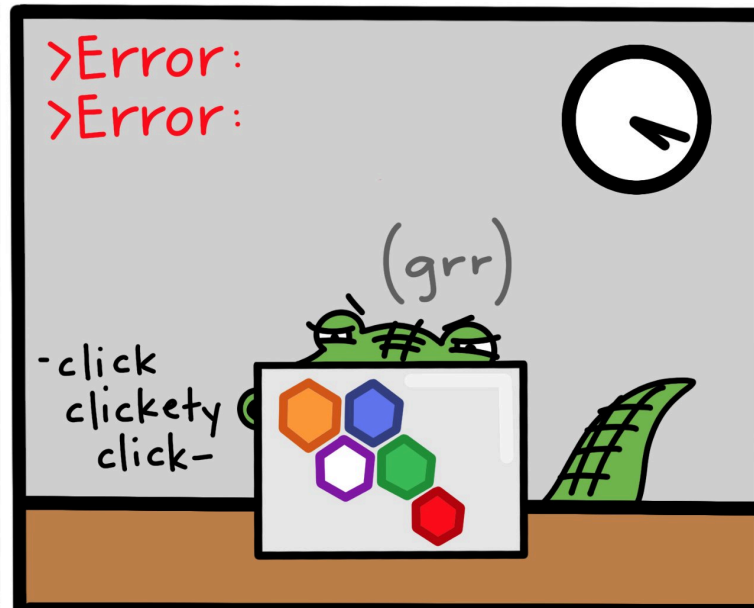
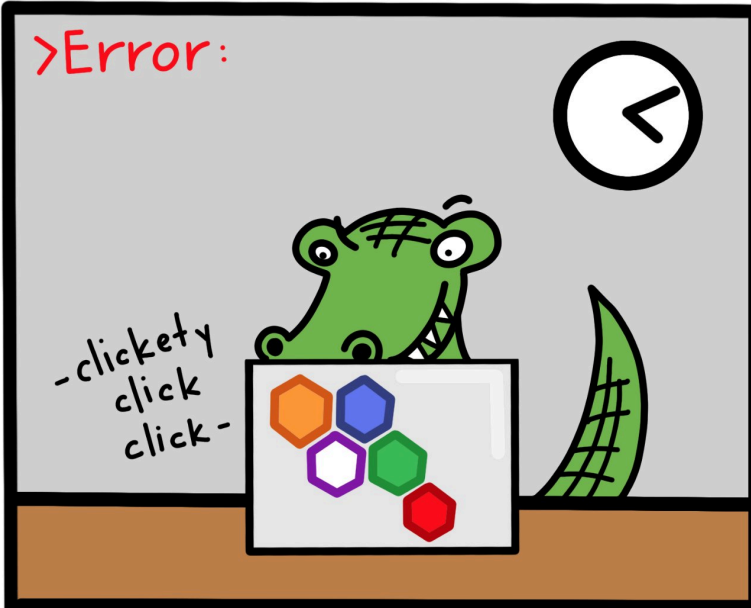
'It's easy **when you start out programming to get really frustrated** and think, "Oh it's me, I'm really stupid," or "I'm not made out to program." But, that is absolutely not the case. Everyone gets frustrated. I still get frustrated occasionally when writing R code. It's just a natural part of programming. So, **it happens to everyone and gets less and less over time**. Don't blame yourself. Just take a break, do something fun, and then come back and try again later.'

— Hadley Wickham, author of `ggplot`

Troubleshooting lessons I guess I'll just relearn forever:

- take a break
- it's almost certainly not a bug
- extra eyes are awesome
- spelling

— Allison Horst, RStudio



Replication

Another researcher could conduct their own study with the same design as yours, and obtain results consistent with your research.

Reproducibility

Another researcher could take your data, run your code, and produce tables and figures identical to yours.

Reproducible research

- Record every step of your analysis in a script which can be run by a computer, with no intervention needed.
- Starting with reading in the original data (ideally in whatever form you first received it), finishing with producing tables and graphs you can include in your paper, report or thesis.
 - Taken to the extreme: write your paper, report or thesis using R Markdown.
- Tools can't do all the work, but they can help.

Reproducible research as kindness to your future self

"Oh no, we found an error in that data file!"

You're going to have to re-do every stage of that analysis from scratch and make all of those tables and figures again.

"That will take hours! I can't even remember how I did that analysis, and was my primary outcome in the variable called

`DOYM12from1stJan_2PTMETHOD` or

`DOYM12from1stJan_noextrapolatedresults?`"

"Not a problem, I'll grab the new data file, change the file name at the top of my script and run it again."

Reproducible research as kindness to your future self

Reviewer number two has asked for more detailed results from your statistical models.

"I forgot to save the full output from that analysis, and now when I try to run it again I get a different results."

"Not a problem, I'll update the script to produce what they asked for and run it again."

qualitative data

non-numerical

collected through interviews, focus groups,
case studies, personal experiences, etc

quantitative data

numerical or categorical

collected through measurement
(very broadly defined)

Categorical and ordinal data

Categories, possibly with an ordering to them.

Examples:

- country of birth
- months of the year
- gender
- Likert scale from "strongly disagree" to "strongly agree"
- species of bacteria
- cancer severity (stage I to stage IV)

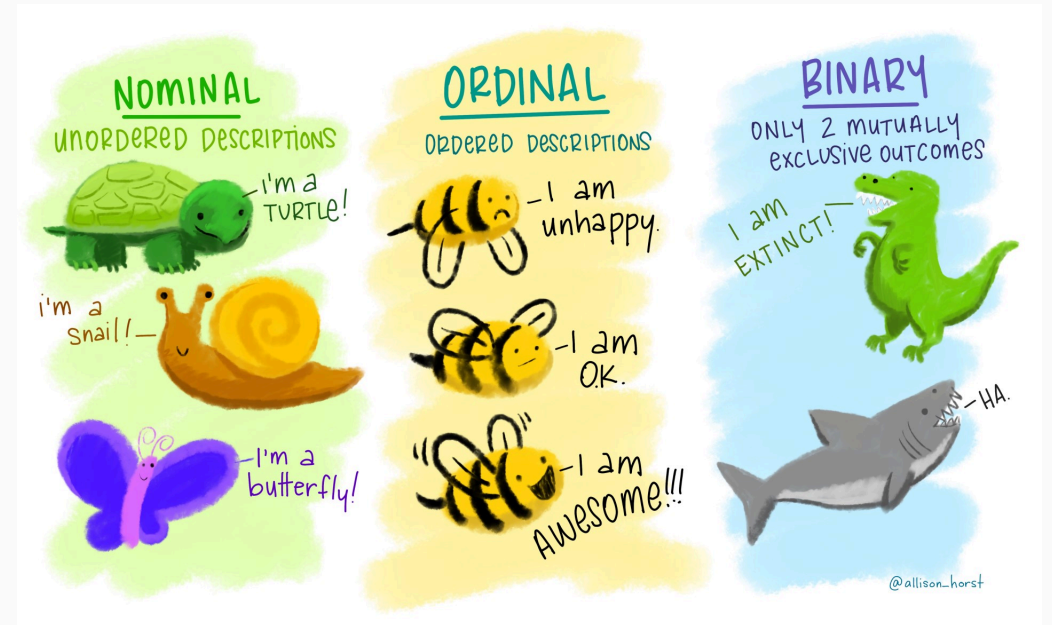


Image by Allison Horst ([Twitter](#)).

Numerical (discrete and continuous) data

Numbers, usually with some kind of units attached.

- Discrete: whole numbers of things
- Continuous: can take any value (possible within a range)

Examples:

- length of daily commute, each direction (minutes)
- number of times commuting by bicycle (per week)
- daily rainfall (mm)
- severity of depression symptoms (on the "DASS" scale, 0 to 42)
- proportion of balcony occupied by pot plants (%)
- weight of newborn baby (g)



Image by Allison Horst ([Twitter](#)).

Collecting and organising data

Experimental units: the things we gather data on.

Variables: the things we measure.

Experimental units	Variables
people	age (years)
trees	diameter at breast height (cm)
plots of land	area (hectares)
samples in a petri dish	relative abundance of bacteria
episodes of Brooklyn 99	number of scenes starring Terry Crews

Organising data in a spreadsheet

General principles:

- One row per experimental unit
- One column per variable
- First row has variable names (be kind to yourself, use a consistent naming scheme)
- Avoid annotations, e.g. 1.58? (ask Josh)
- Don't record important information as formatting, e.g. colours

More detailed recommendations we endorse:

Broman, K. W., & Woo, K. H. (2018). Data organization in spreadsheets. *The American Statistician*, 72(1), 2-10.

<https://www.tandfonline.com/doi/full/10.1080/00031305.2017.1375989>

Specific recommendations for easy import to R:

- Missing values should be encoded consistently in a way that couldn't be confused with a real value (e.g. NA rather than 999 - unless your column records continents)
- Categorical data should be recorded as short category labels, with consistent spelling, punctuation and capitalisation (no need to covert into numbers)
- Column names use letters, numbers and underscore _ only

	A	B	C	D	E	F	G	H
1	Gender	Event	Location	Year	Medal	Name	Nationality	Result
2	M	100M Men	Rio	2016	G	Usain BOLT	JAM	9.81
3	M	100M Men	Rio	2016	S	Justin GATLIN	USA	9.89
4	M	100M Men	Rio	2016	B	Andre DE GRASSE	CAN	9.91
5	M	100M Men	Beijing	2008	G	Usain BOLT	JAM	9.69
6	M	100M Men	Beijing	2008	S	Richard THOMPSON	TTO	9.89
7	M	100M Men	Beijing	2008	B	Walter DIX	USA	9.91
8	M	100M Men	Sydney	2000	G	Maurice GREENE	USA	9.87
9	M	100M Men	Sydney	2000	S	Ato BOLDON	TTO	9.99
10	M	100M Men	Sydney	2000	B	Obadele THOMPSON	BAR	10.04
11	M	100M Men	Barcelona	1992	G	Linford CHRISTIE	GBR	9.96
12	M	100M Men	Barcelona	1992	S	Frank FREDERICKS	NAM	10.02
13	M	100M Men	Barcelona	1992	B	Dennis MITCHELL	USA	10.04
14	M	100M Men	Los Angeles	1984	G	Carl LEWIS	USA	9.99
15	M	100M Men	Los Angeles	1984	S	Sam GRADDY	USA	10.19
16	M	100M Men	Los Angeles	1984	B	Ben JOHNSON	CAN	10.22

Olympic_100m_results

Types of data in R

- numeric: e.g. `20.5`, `4`, `0.15`, `-5`
sometimes further sub-categorised as "double" (real numbers)
and "integer" (whole numbers)
- logical: `TRUE` or `FALSE`
- character: e.g. `"February"`, `"Australia"`, `"Male"`
- factor: a 'character' with an ordered list of possible values
- missing: `NA` (can be any of the above types)

Types of data in R

- vector: a one-dimensional column of values, all of the same type. In R, unlike most programming languages, vectors are the fundamental data type. Individual values are vectors with length one.
- data frame: a collection of vectors, each with an associated column name. All columns have the same number of rows. Also sometimes called a *tibble* (a Tidyverse extension to built-in R data frames).

The next two are mostly used for programming, not data analysis. You might see them mentioned elsewhere, but we won't be talking about them:

- list: similar to a vector, but can contain values of different types, including other lists.
- matrix: similar to a data frame, but all values are of the same type (most commonly all numeric).

Exercise 1.1 (RStudio demo).