shirleywang.rbind.io
shirleywang@g.harvard.edu
@shirleybwang

Introduction to R

Day 1: Data Cleaning & Wrangling

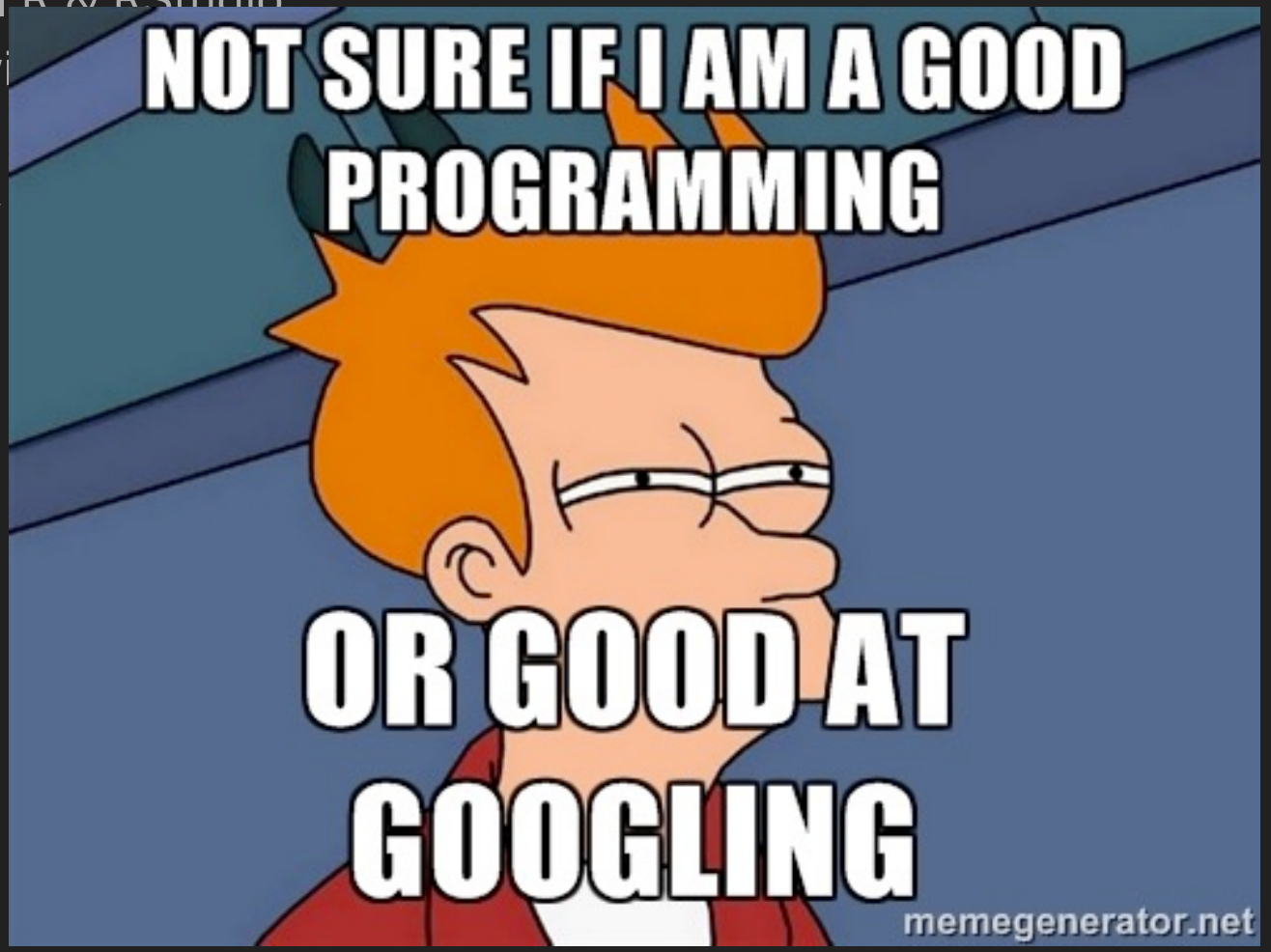
Shirley Wang
July 27, 2021

Day 1 Goals

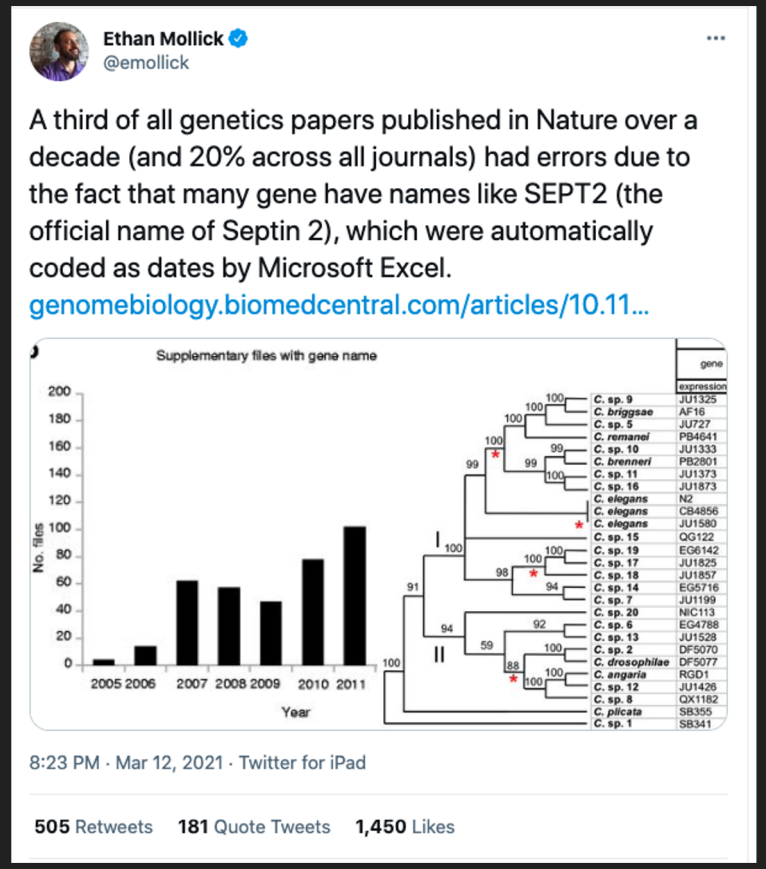
- Download & install R & RStudio
- Become familiar with the RStudio environment
- Operations
- Objects (what they are & how to create them)
- Data types
- Functions

Day 1 Goals

- Download & install R & RStudio
- Become familiar wi
- Operations
- Objects (what they
- Data types
- Functions



Why use R?



Also an Excel error cost led to an estimated 1,500 deaths in the UK last year. The UK COVID contact tracing effort saved to an old Excel file format (.XLS) rather than the new one (.XLSX) & smaller row limits resulted in names being dropped. They were never contacted by tracers.

Does Contact Tracing Work? Quasi-Experimental Evidence from an Excel Error in England*

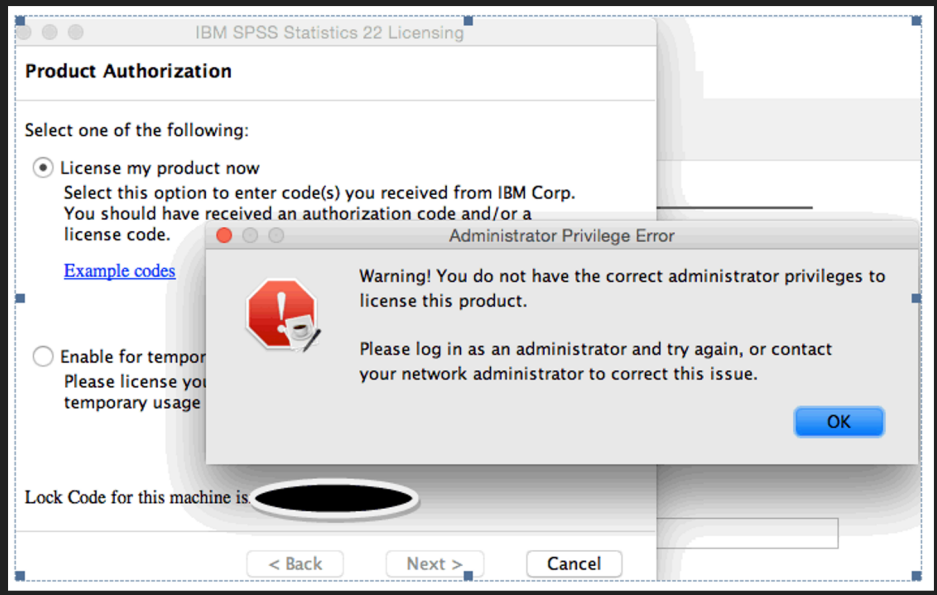
Thiemo Fetzer† Thomas Graeber‡

November 24, 2020

Abstract

Contact tracing has been a central pillar of the public health response to the COVID-19 pandemic. Yet, contact tracing measures face substantive challenges in practice and well-identified evidence about their effectiveness remains scarce. This paper exploits quasi-random variation in COVID-19 contact tracing. Between September 25 and October 2, 2020, a total of 15,841 COVID-19 cases in England (around 15 to 20% of all cases) were not immediately referred to the contact tracing system due to a data processing error. Case information was truncated from an Excel spreadsheet after the row limit had been reached, which was discovered on October 3. There is substantial variation in the degree to which different parts of England areas were exposed – by chance – to delayed referrals of COVID-19 cases to to the contact tracing system. We show that more affected areas subsequently experienced a drastic rise in new COVID-19 infections and deaths alongside an increase in the positivity rate and the number of test performed, as well as a decline in the performance of the contact tracing system. Conservative estimates suggest that the failure of timely contact tracing due to the data glitch is associated with more than 125,000 additional infections and over 1,500 additional COVID-19-related deaths. Our findings provide strong quasi-experimental evidence for the effectiveness of contact tracing.

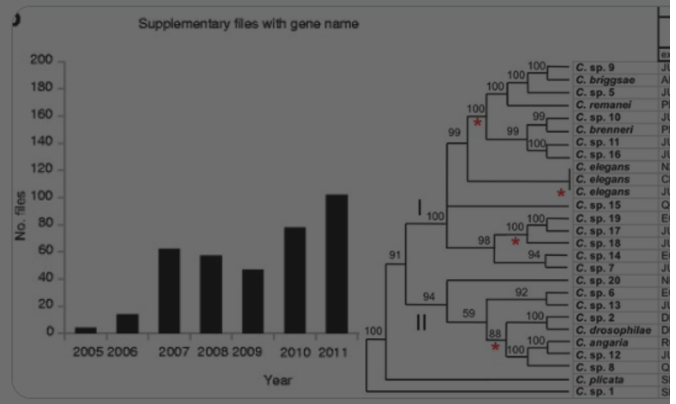
Keywords: Health, Coronavirus
JEL Classification: I31, Z18



Why use R?

Ethan Mollick
@emollick

A third of all genetics papers published in Nature over a decade (and 20% across all journals) had errors due to the fact that many gene have names like SEPT2 (the official name of Septin 2), which were automatically coded as dates by Microsoft Excel.
genomebiology.biomedcentral.com/articles/10.11...

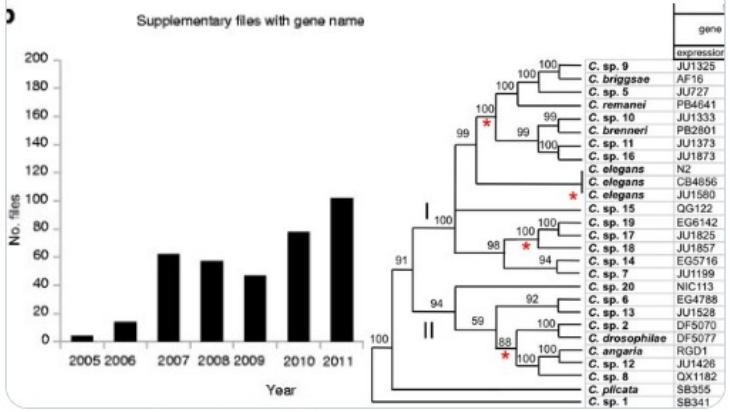


8:23 PM · Mar 12, 2021 · Twitter for iPad

505 Retweets **181** Quote Tweets **1,450** Likes

Shirley B. Wang @ShirleyBWang · Mar 16
This is horrifying 😳 I wonder how many papers in psych have errors due to data cleaning by hand in excel...copy/paste mistakes, accidental deletions, etc

> **Ethan Mollick** @emollick · Mar 12
> A third of all genetics papers published in Nature over a decade (and 20% across all journals) had errors due to the fact that many gene have names like SEPT2 (the official name of Septin 2), which were automatically coded as dates by Microsoft Excel.
> genomebiology.biomedcentral.com/articles/10.11...
> Show this thread



11 22 142

Erin Reilly
@ee_reilly

Replying to @ShirleyBWang

If I think about this for too long I enter into an existential crisis

1:38 PM · Mar 16, 2021 · Twitter Web App

17 Likes

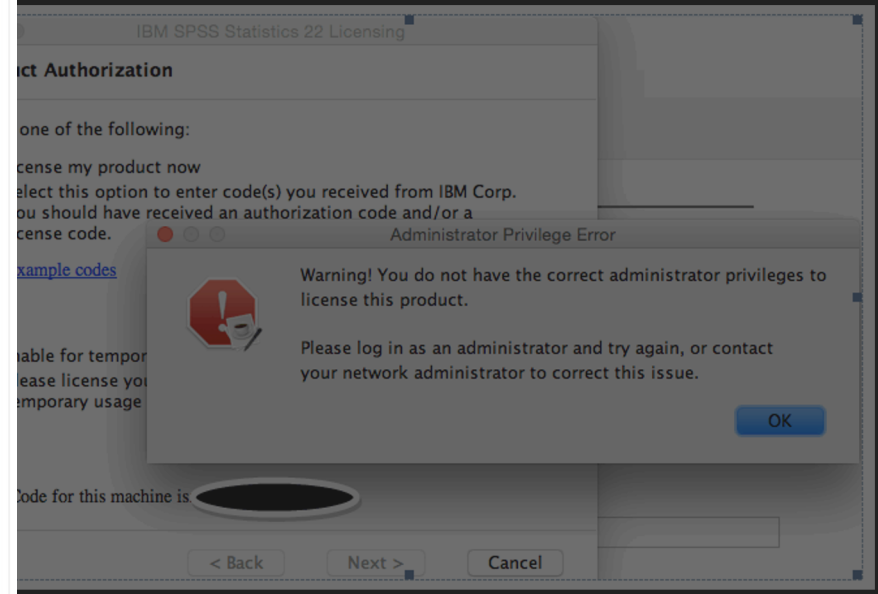
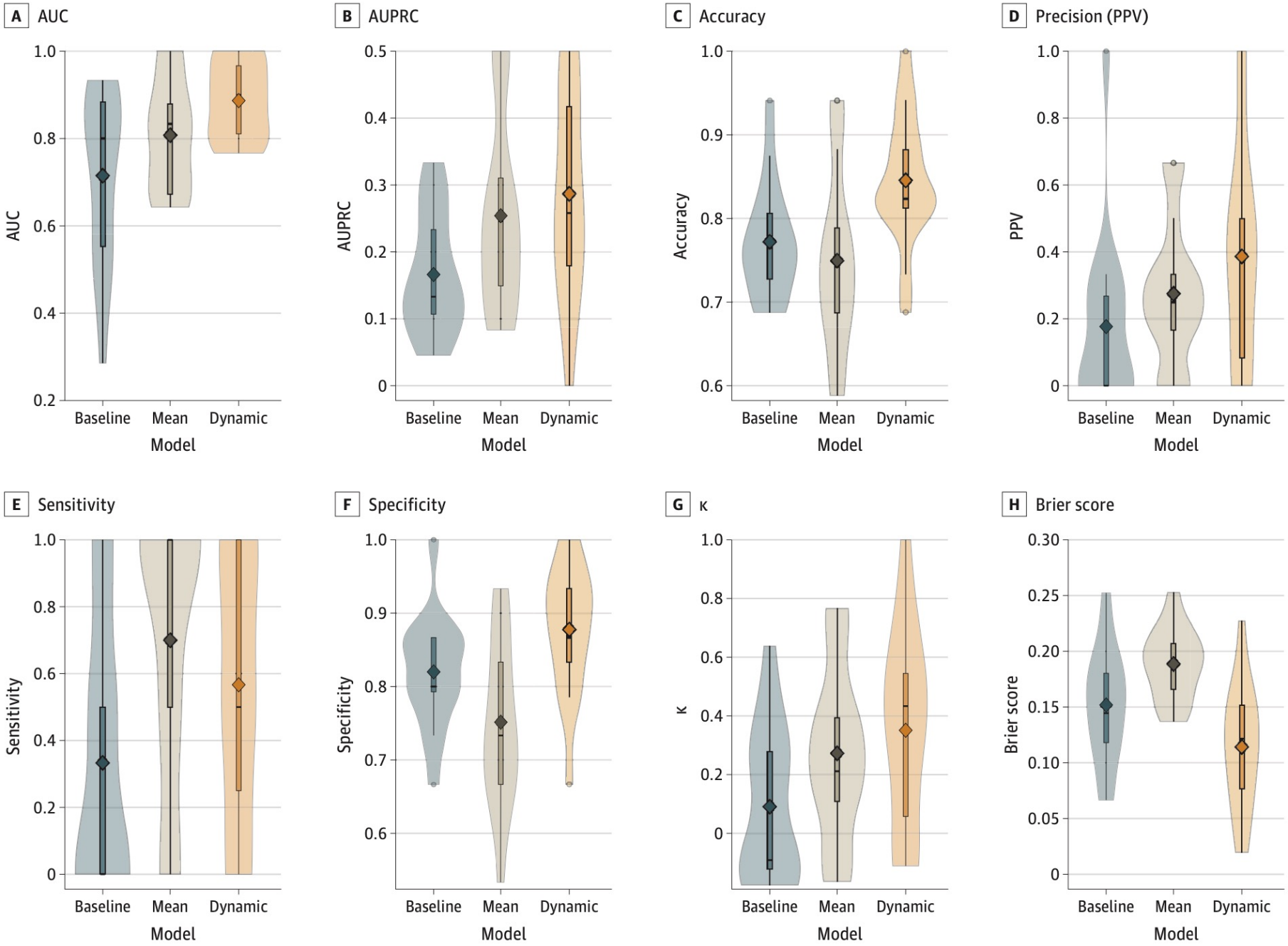
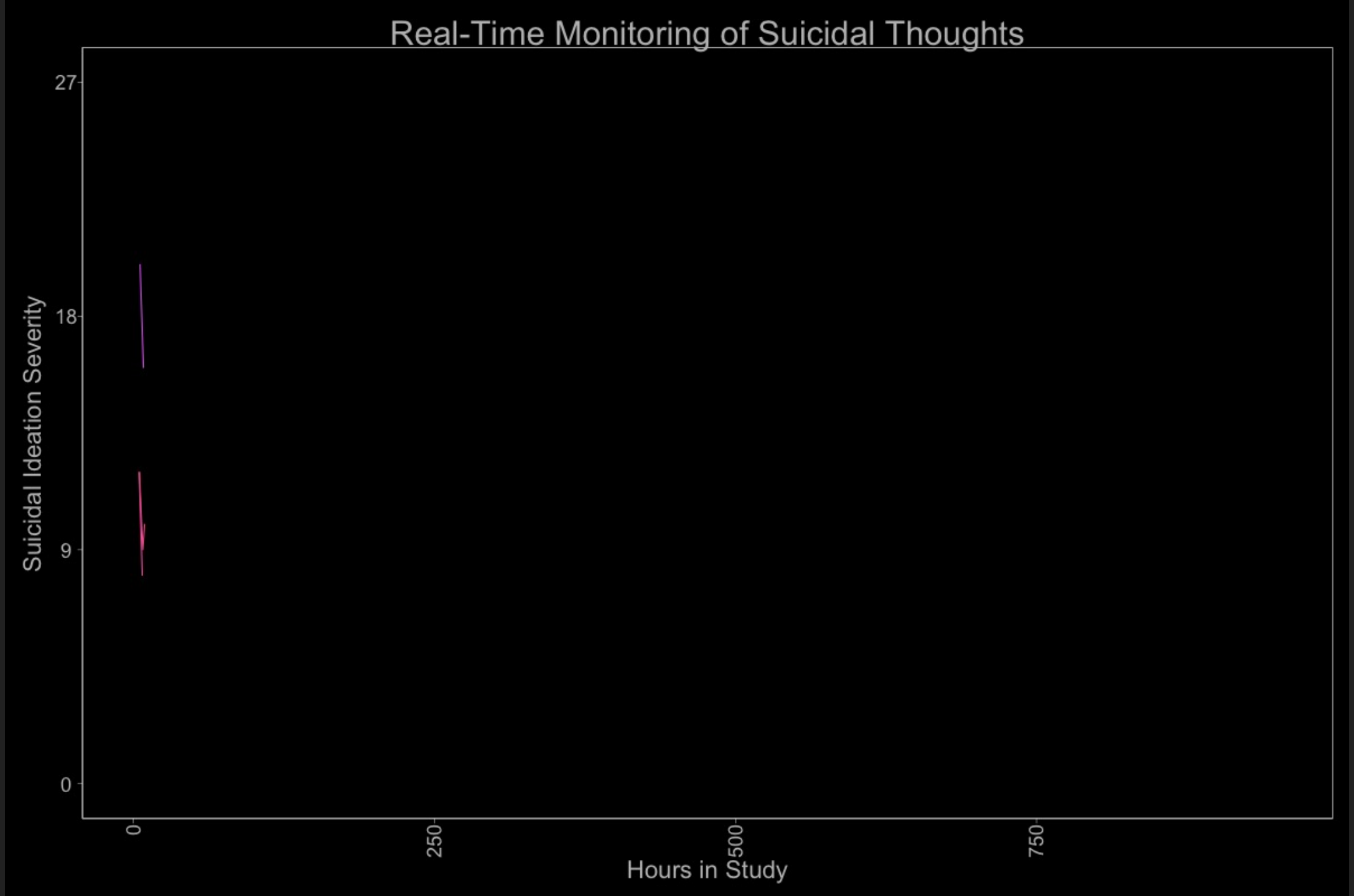
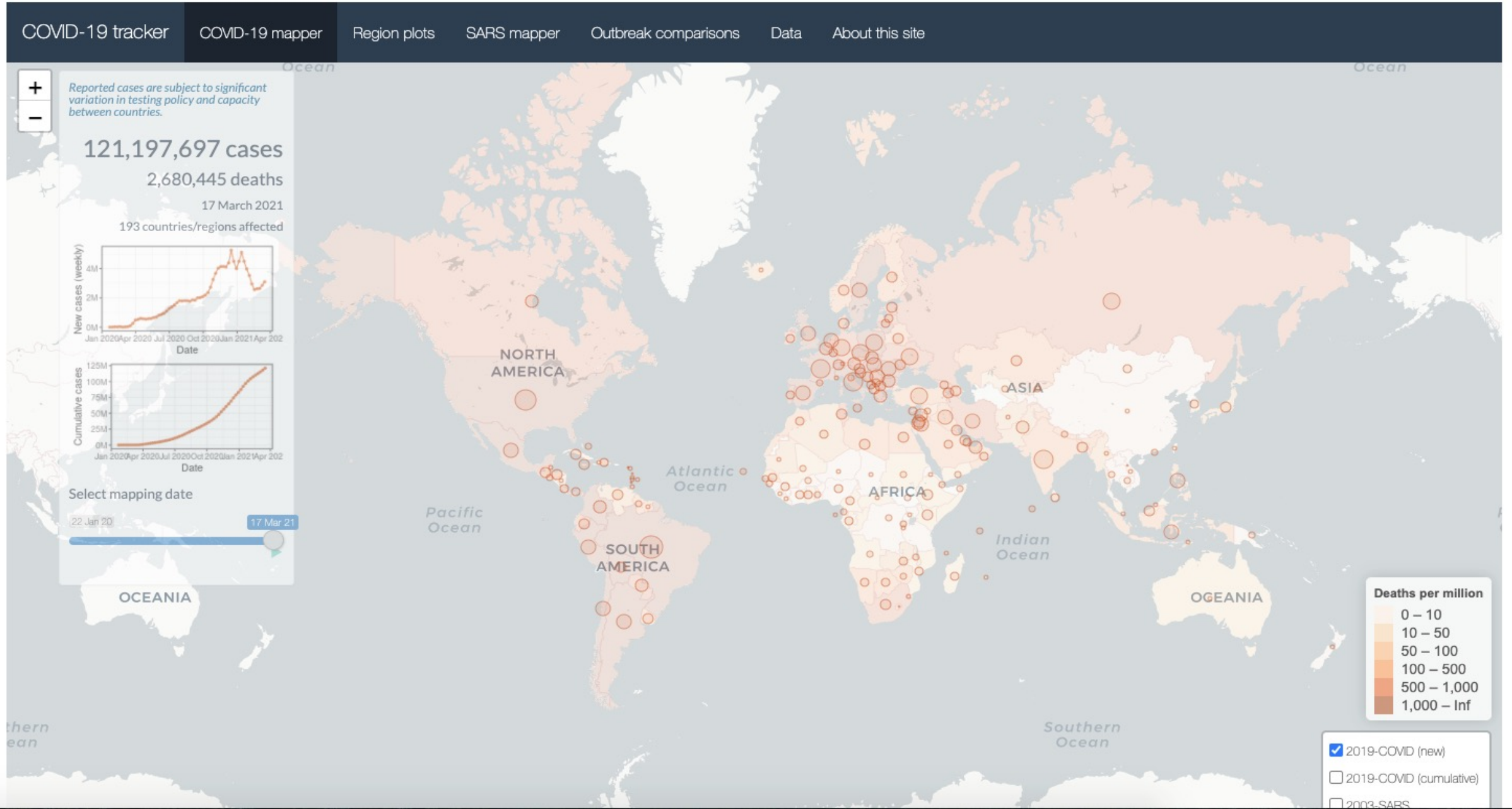


Figure 1. Suicide Attempt Prediction Model Metrics




Real-Time Monitoring of Suicidal Thoughts
Suicidal Ideation Severity
27
18
9
0
0
250
500
750
Hours in Study

covid19, epidemiology



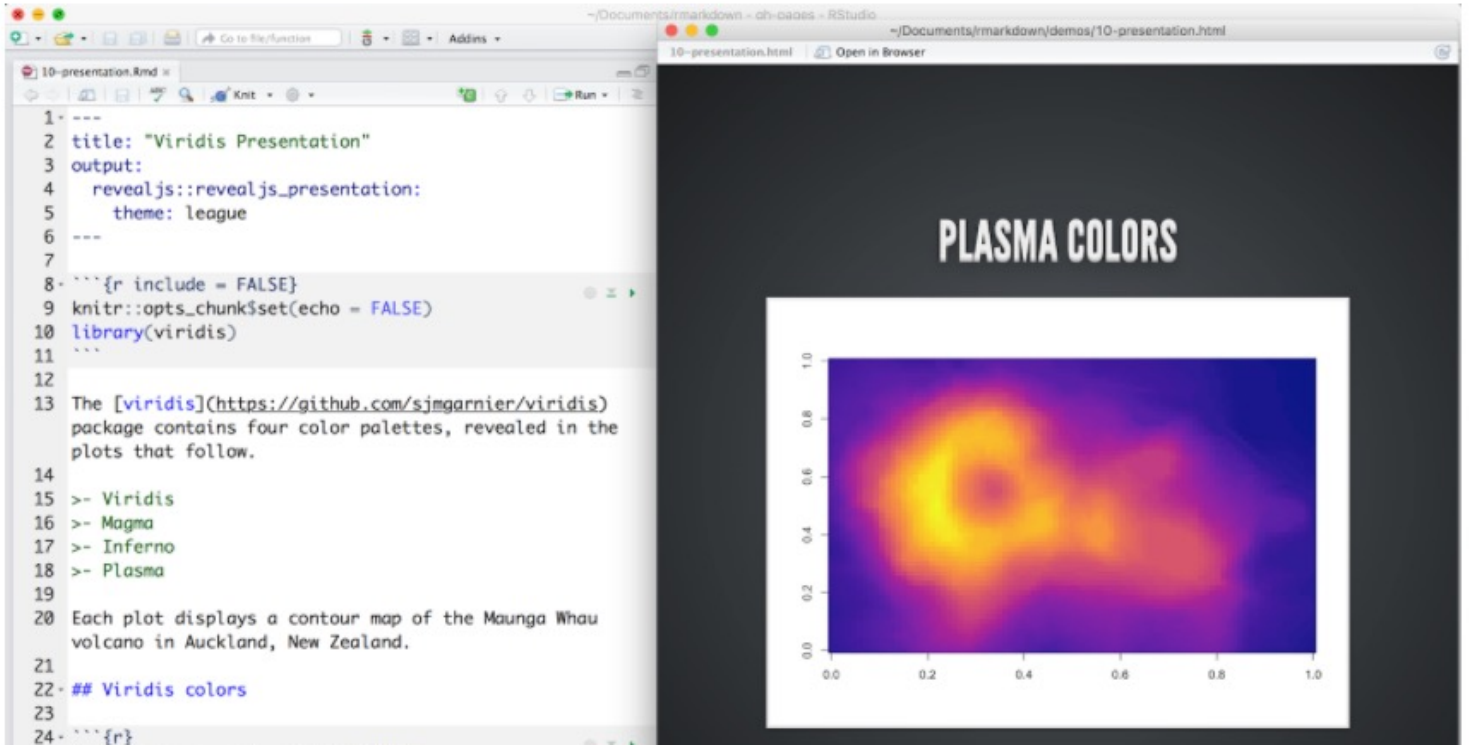
Slide Presentations

R Markdown renders to four presentation formats:

- beamer_presentation - PDF presentations with beamer
- ioslides_presentation - HTML presentations with ioslides
- slidy_presentation - HTML presentations with slidy
- powerpoint_presentation - PowerPoint presentation
- revealjs::revealjs_presentation - HTML presentations with reveal.js

Each format will intuitively divide your content into slides, with a new slide beginning at each first or second level header.

Insert a horizontal rule (`***`) into your document to create a manual slide break. Create incremental bullets with `>-` , as in the .Rmd file below, which is available here on RStudio Cloud.



Shirley Wang

Ph.D. candidate

Harvard University

About

I am a PhD candidate in clinical psychology with a secondary in computational science and engineering at Harvard University. I work with Matt Nock and am funded by the National Science Foundation Graduate Research Fellowship Program. My research examines why people engage in behaviors that are harmful to themselves, including eating disorder behaviors, nonsuicidal self-injury, and suicide. I use a wide range of methods to study these problems, including laboratory-based behavioral experiments, real-time monitoring, and large-scale longitudinal studies. I am particularly interested in using mathematical and computational modeling to formalize theories in psychopathology.

I value diversity, inclusion, and belonging, and am committed to promoting these values in research, academia, clinical practice, and beyond. I believe that increasing representation of people from marginalized and historically disadvantaged backgrounds is critical for conducting ethical, comprehensive, and innovative clinical science research.

Interests

- eating disorders
- suicide
- computational modeling
- real-time monitoring

Education

Ph.D. in Clinical Science, expected 2023
Harvard University

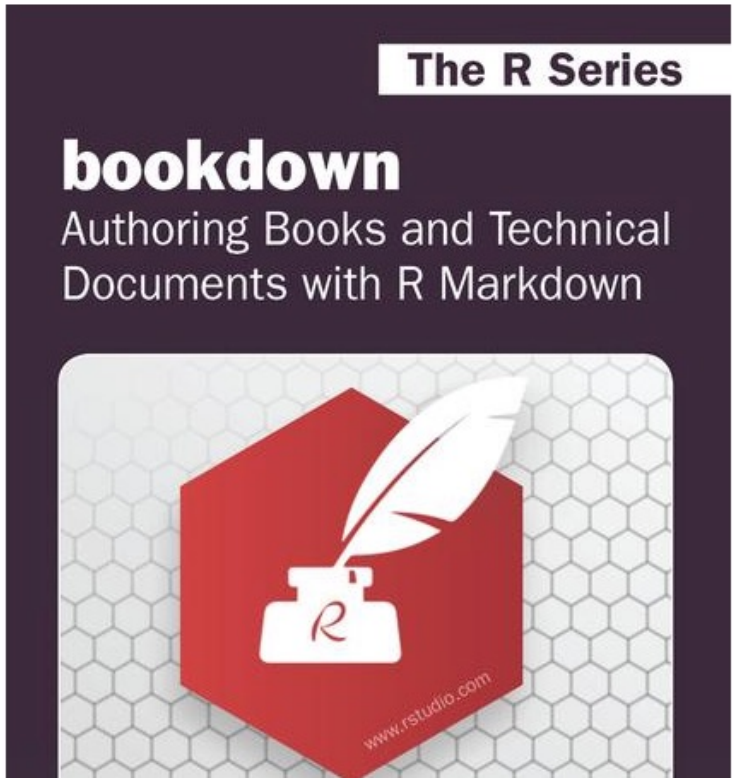
A.M. in Clinical Science, 2019
Harvard University

bookdown: Authoring Books and Technical Documents with R Markdown

Yihui Xie

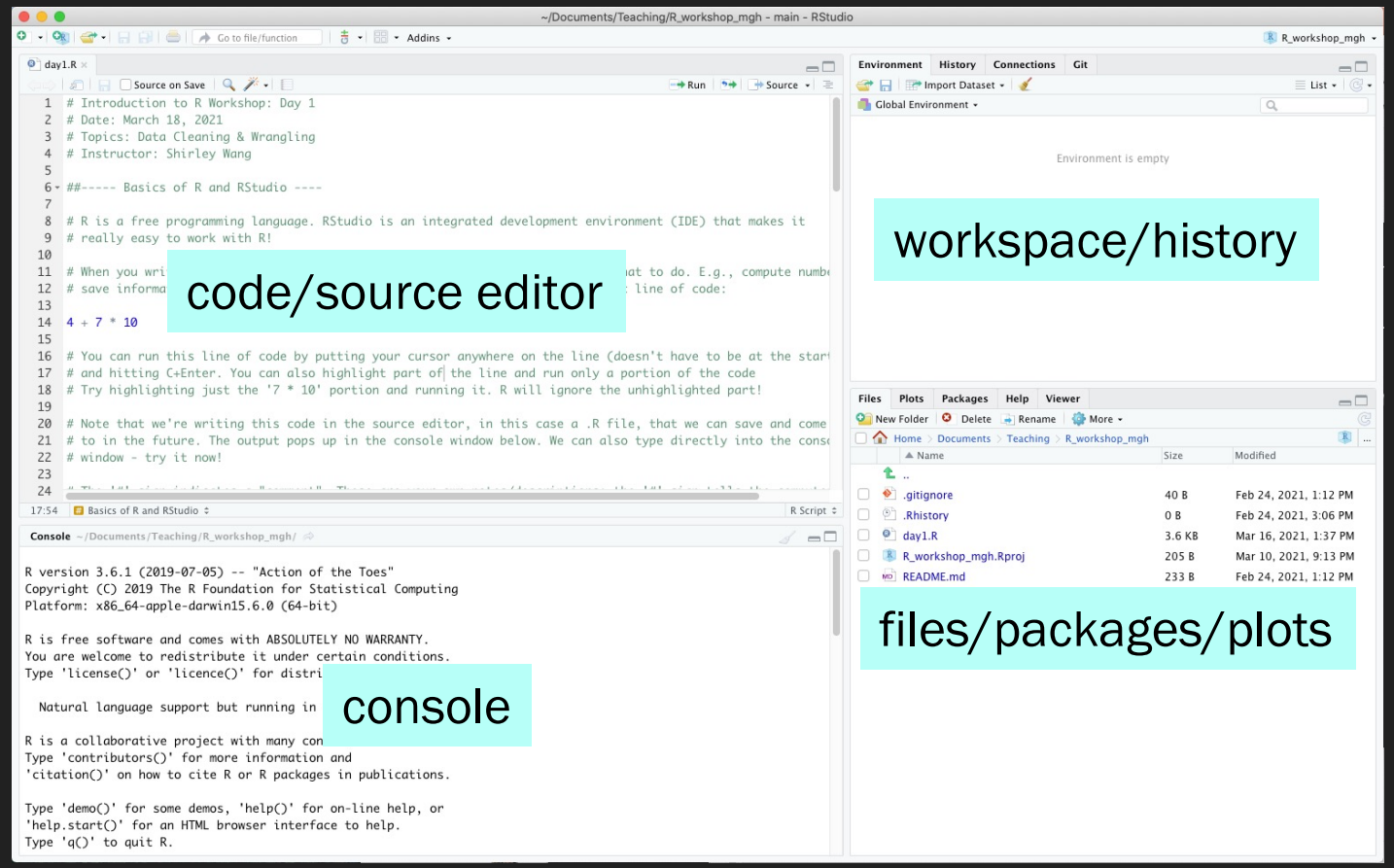
2021-03-15

Preface



R & RStudio Orientation

- R is a free programming language
- RStudio is an IDE that makes it easier write, run, and save R code!

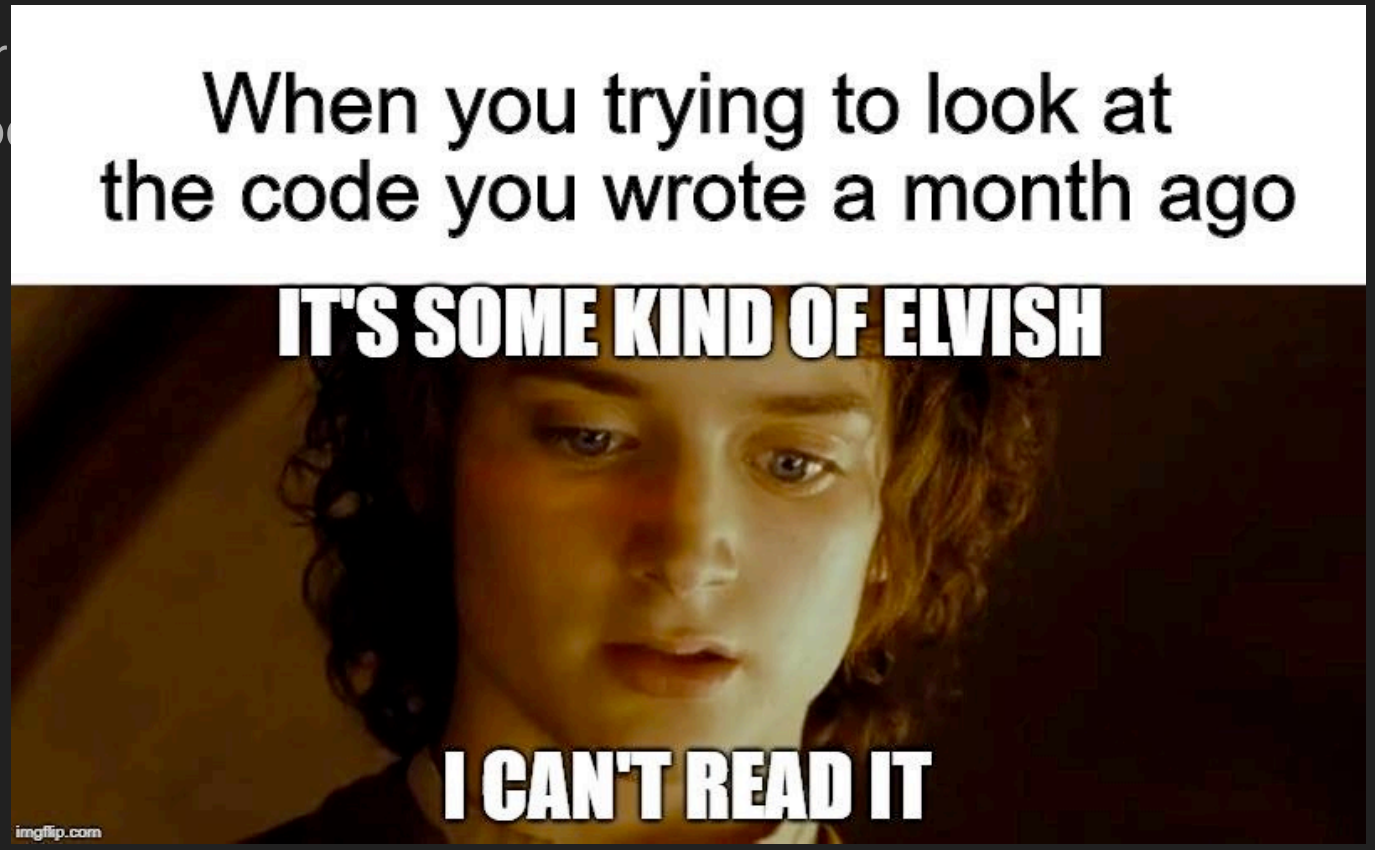


R Basics

- Case sensitive
- Run lines of code with cmd+Enter
- Get help with `?` or `help()`
- Comment your code with #

R Basics

- Case sensitive
- Run lines of code with cmd+Enter
- Get help with ? or
- Comment your co



Operations in R

Arithmetic Operators

Operator	Description
+	addition
-	subtraction
*	multiplication
/	division
^ or **	exponentiation
x %% y	modulus (x mod y) 5%%2 is 1
x %/% y	integer division 5%/%2 is 2

Logical Operators

Operator	Description
<	less than
<=	less than or equal to
>	greater than
>=	greater than or equal to
==	exactly equal to
!=	not equal to
!x	Not x
x | y	x OR y
x & y	x AND y
isTRUE(x)	test if X is TRUE

https://www.statmethods.net/management/operators.html

Objects in R

- Objects store information (can be an entire dataset, a single value, a string of text, an image, etc)
- Assign values using '`<-`'
- Value on the right-hand side gets assigned to object name on the left-hand side (e.g., `myname <- 'shirley'; myage <- 26`)
- Rules for object names:
  - Must start with a letter
  - Can only contain letters, numbers, periods, underscores
  - Case-sensitive

Data Types

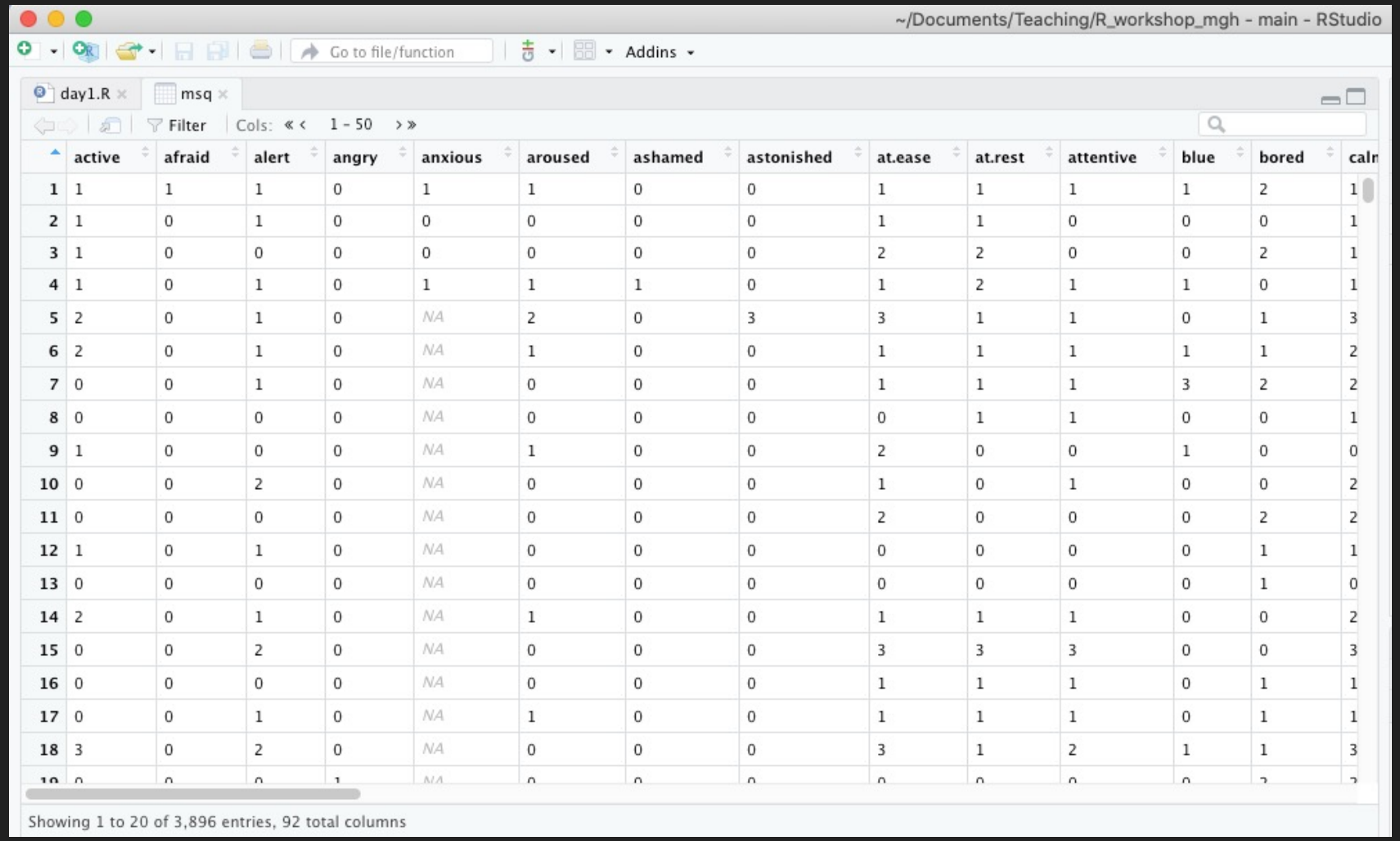
- Numeric
- Integer
- Character
- Factor
- Logical
- Check data types in R with class()

Data Structures

- Vector
- Matrix
- Array
- Data frame
- List

Data Structures

- Vector
- Matrix
- Array
- **Data frame**
- List



	active	afraid	alert	angry	anxious	aroused	ashamed	astonished	at.ease	at.rest	attentive	blue	bored	calm
1	1	1	1	0	1	1	0	0	1	1	1	1	2	1
2	1	0	1	0	0	0	0	0	1	1	0	0	0	1
3	1	0	0	0	0	0	0	0	2	2	0	0	2	1
4	1	0	1	0	1	1	1	0	1	2	1	1	0	1
5	2	0	1	0	NA	2	0	3	3	1	1	0	1	3
6	2	0	1	0	NA	1	0	0	1	1	1	1	1	2
7	0	0	1	0	NA	0	0	0	1	1	1	3	2	2
8	0	0	0	0	NA	0	0	0	0	1	1	0	0	1
9	1	0	0	0	NA	1	0	0	2	0	0	1	0	0
10	0	0	2	0	NA	0	0	0	1	0	1	0	0	2
11	0	0	0	0	NA	0	0	0	2	0	0	0	2	2
12	1	0	1	0	NA	0	0	0	0	0	0	0	1	1
13	0	0	0	0	NA	0	0	0	0	0	0	0	1	0
14	2	0	1	0	NA	1	0	0	1	1	1	0	0	2
15	0	0	2	0	NA	0	0	0	3	3	3	0	0	3
16	0	0	0	0	NA	0	0	0	1	1	1	0	1	1
17	0	0	1	0	NA	1	0	0	1	1	1	0	1	1
18	3	0	2	0	NA	0	0	0	3	1	2	1	1	3

Data Frames

- Access columns with ‘$’ (e.g., `df$age` accesses the column named ‘age’ in the data frame ‘df’)
- `View(df)` to view data frame in excel-like format
- `dim(df)` to get dimensions (# rows, # columns)
  - `nrow(df)`
  - `ncol(df)`
- `str(df)` to view structure of data frame
- `colnames(df)` to view all column (variable) names

Functions

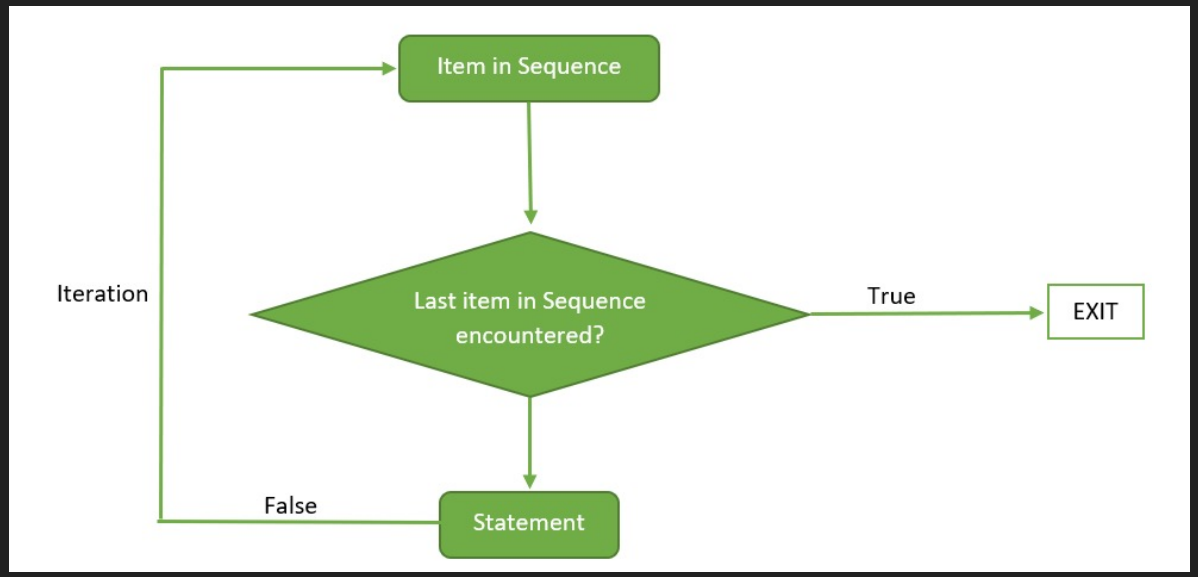
- Sets of instructions/statements to complete a specific task; takes input (e.g., numbers) and returns an output (e.g., mean of those numbers)
- R comes with many built-in functions (e.g., `mean()`, `min()`, `max()`, `print()`)
- You can also write your own functions to perform specific tasks!
- `function(arguments)`
- `myFunction <- function(input) {# do something; return output}`

Packages

- R packages are bundles of functions, documentation, and data to perform tasks (e.g., machine learning, natural language processing, data visualization)
- Anyone can build an R package to share with other R users!
- Hosted on the central R repository CRAN: https://cran.r-project.org/
- Currently over 17,000 packages available
- Packages need to be installed on your computer, and loaded at the start of each new R session
- `install.packages('ggplot2')`
- `library(ggplot2)`

Loops

- Loops are useful to repeat a block of code multiple times
- Rather than copy/pasting code several times, you can iterate through a loop! This can shorten hundreds of lines of code to just a few lines.



shirleywang.rbind.io
shirleywang@g.harvard.edu
@shirleybwang

Thank you!

Helpful resources:

- https://rstudio.com/resources/cheatsheets/
- https://www.r-bloggers.com/
- https://psyr.djnavarro.net/
- https://swirlstats.com/
- https://r4ds.had.co.nz/
- https://paulvanderlaken.com/2017/08/10/r-resources-cheatsheets-tutorials-books/
- https://www.learnr4free.com/en/index.html
- https://docs.google.com/document/d/1qtdiLbU32F_AVNRlF7d23wvPSgTCYqr5TFeBwbncPcw/edit