

Data Wrangling with R and the Tidyverse

Session 1

Kieran Healy

Statistical Horizons, September 2021

Housekeeping

Housekeeping

10am till 2pm US EST

Housekeeping

10am till 2pm US EST

Lab session from 5pm to 6pm US EST

On Thursday and Friday

Housekeeping

10am till 2pm US EST

Lab session from 5pm to 6pm US EST

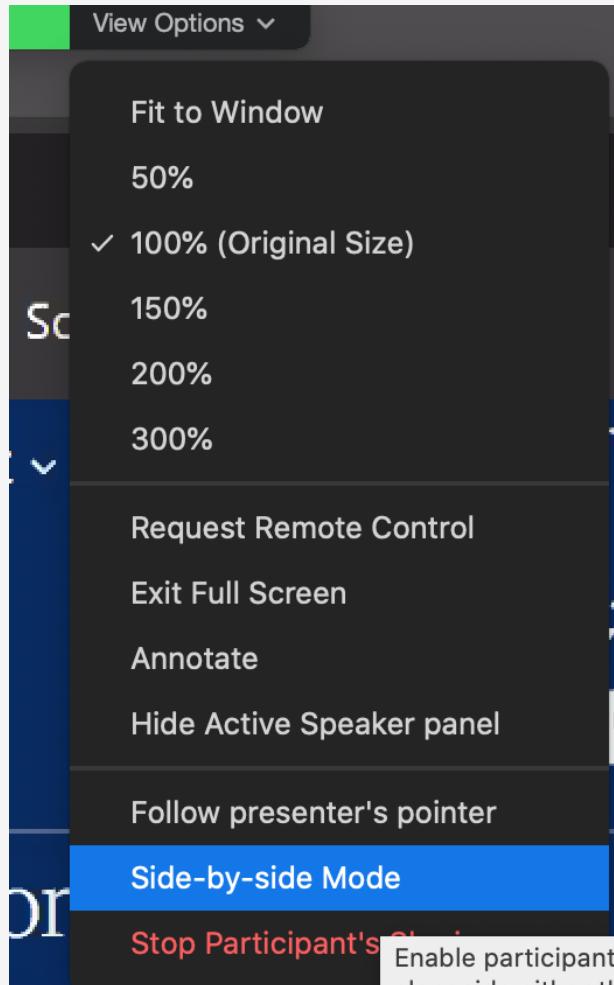
On Thursday and Friday

Use the Zoom chat to ask questions, or raise a hand with 

In between class sessions



For a better Zoom experience



If you're watching in full-screen view and I'm sharing my screen, then from Zoom's "View options" menu *turn off* "Side-by-Side" mode.

This Morning's Goals

This Morning's Goals

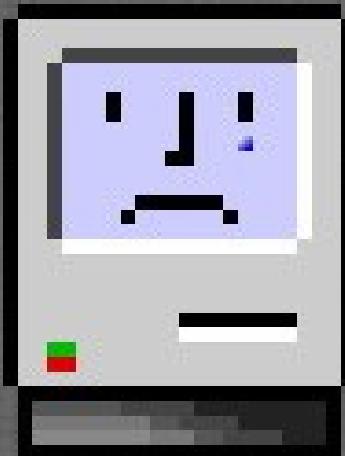
Some big-picture motivation & perspective

Getting familiar with RStudio and its relationship to R

Getting oriented to R and how it thinks

DATA ANALYSIS
is mostly
DATA WRANGLING

Wrangling data is frustrating



Can we make it **fun**?



Can we make it **fun**?



No.

Can we make it **fun**?



No.

⇒ Not *this* much fun, at any rate

OK but can we eliminate frustration?



OK but can we eliminate frustration?



Also no.

OK but can we eliminate frustration?



Also no.

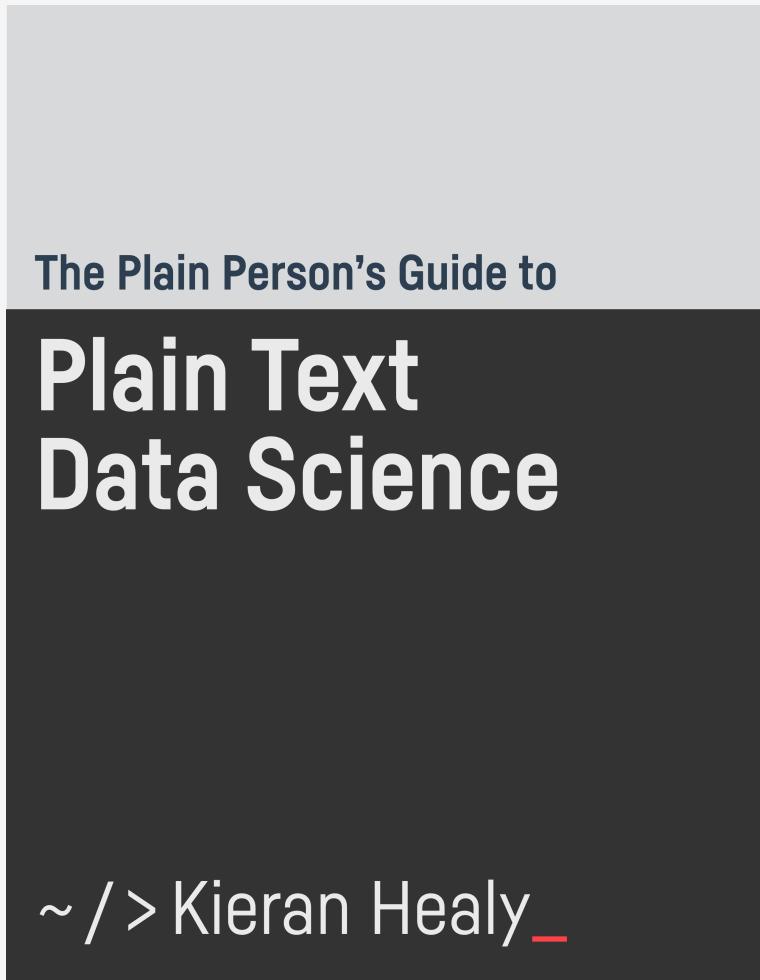
Sorry.

**HOWEVER, WE CAN
MAKE IT *WORK***

HOWEVER, WE CAN MAKE IT *WORK*

Also, it's weirdly satisfying once you get into it.

We take a broadly *Plain Text* approach



We take a broadly *Plain Text* approach

The Plain Person's Guide to

Plain Text Data Science

~ /> Kieran Healy _

Using R and the Tidyverse can be understood within this broader context.

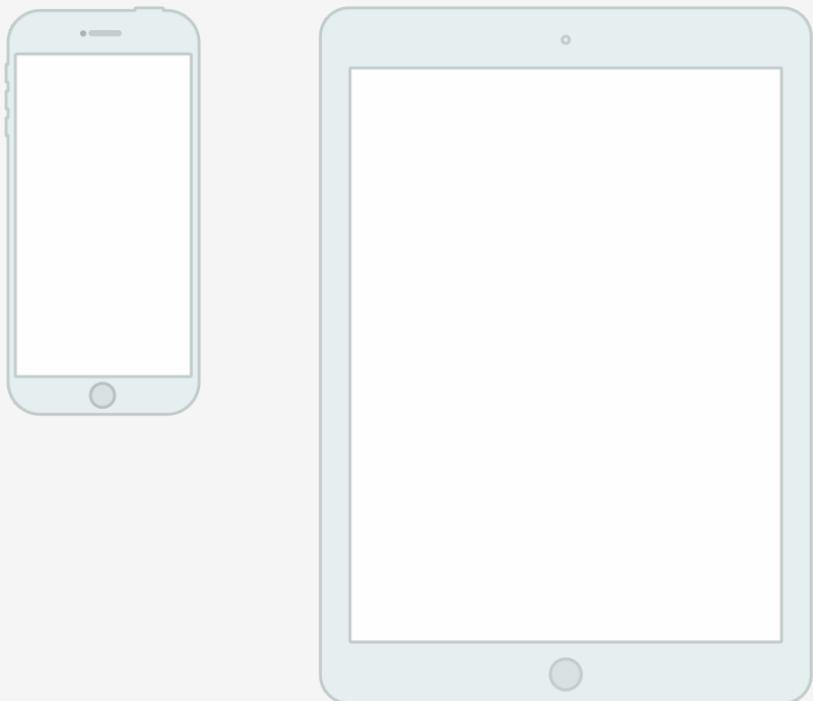
The same principles would apply to, e.g., using Python or similar tools.

Two revolutions in computing

Where the action is

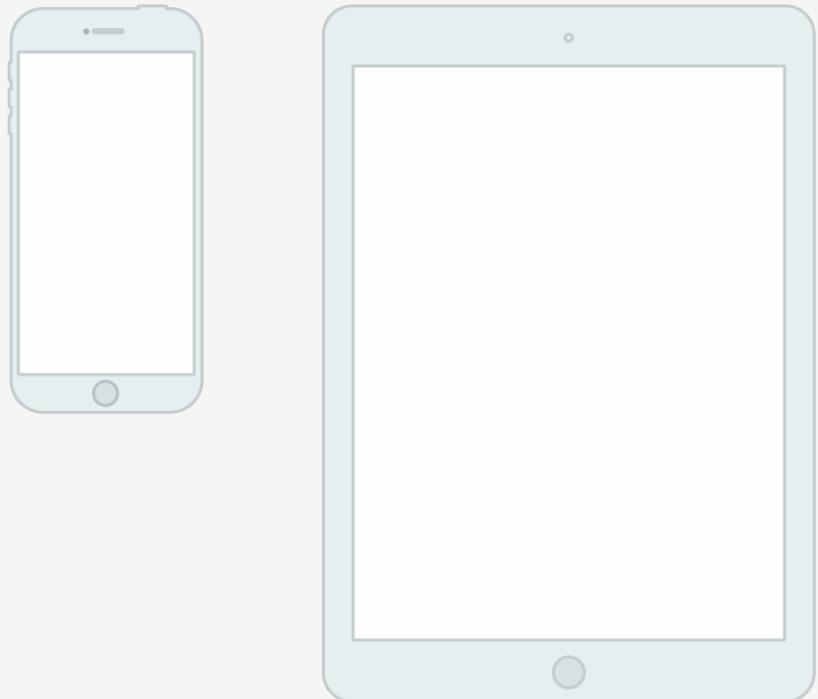


Where the action is



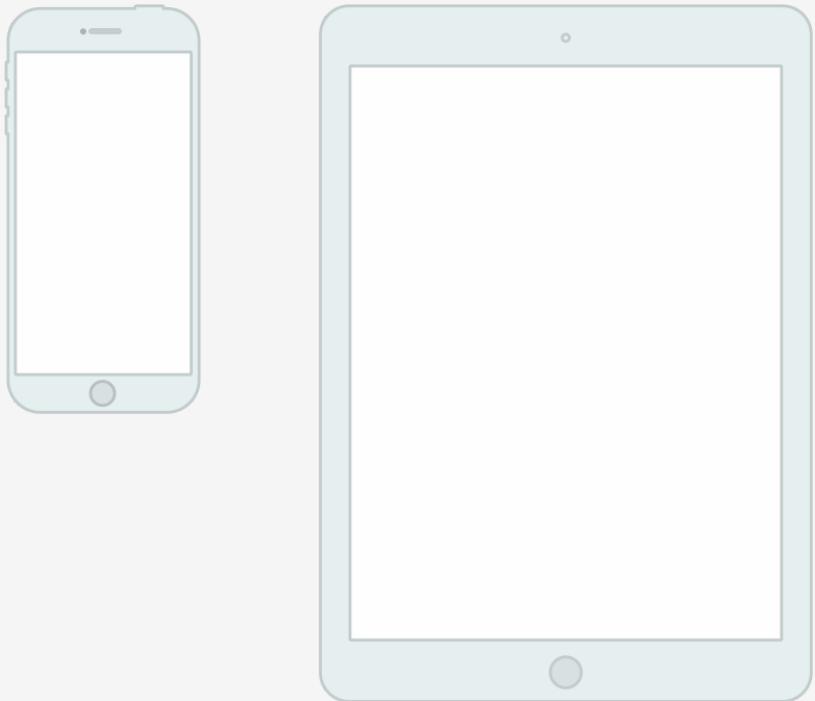
Touch-based user interface

Where the action is



Touch-based user interface
Foregrounds a single application

Where the action is



Touch-based user interface
Foregrounds a single application
Dislikes multi-tasking*

Where the action is



- Touch-based user interface
- Foregrounds a single application
- Dislikes multi-tasking*
- Hides the file system

*Multitasking

I mean, “Making different specialized applications and resources work together in the service of a single but multi-dimensional project”, not “Checking Twitter while also listening to a talk and waiting for an update from the school nurse.”

Where statistical computing lives

+

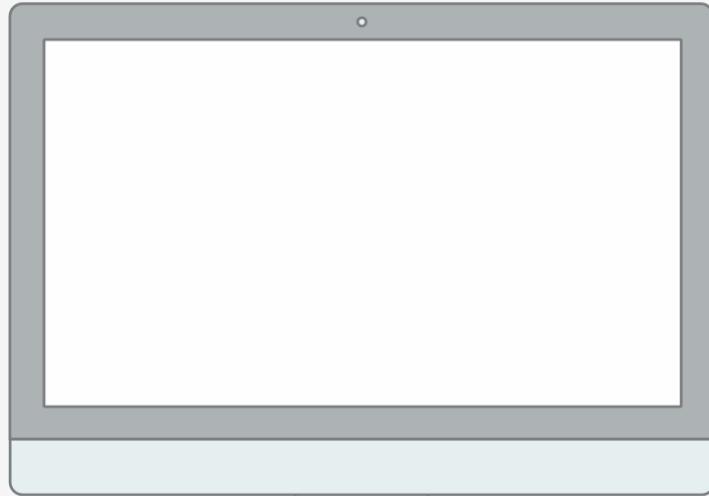


Where statistical computing lives

+



Windows and pointers.



Where statistical computing lives

+



Windows and pointers.

Multi-tasking, multiple windows.



Where statistical computing lives

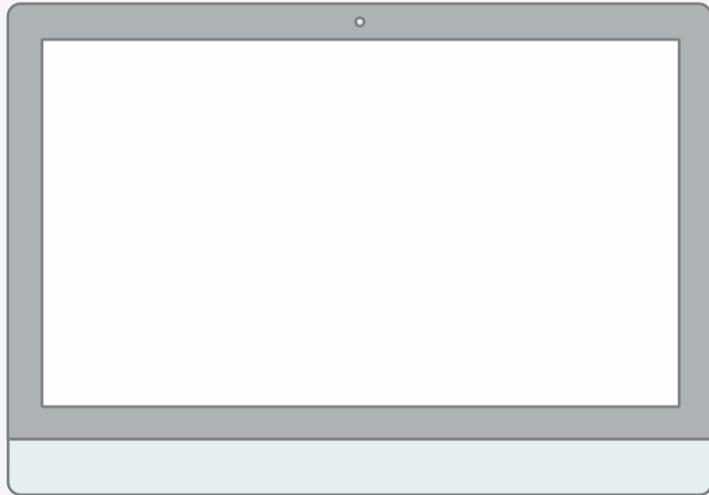
+



Windows and pointers.

Multi-tasking, multiple windows.

Exposees and leverages the file system.



Where statistical computing lives

+

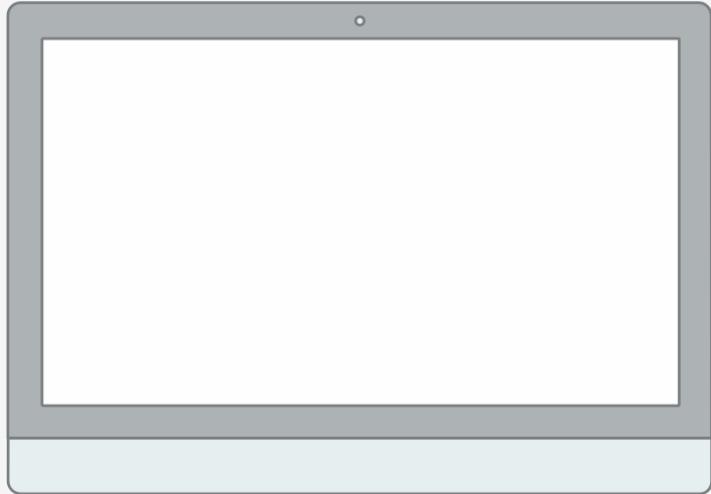


Windows and pointers.

Multi-tasking, multiple windows.

Exposees and leverages the file system.

Many specialized tools in concert.



Where statistical computing lives

+



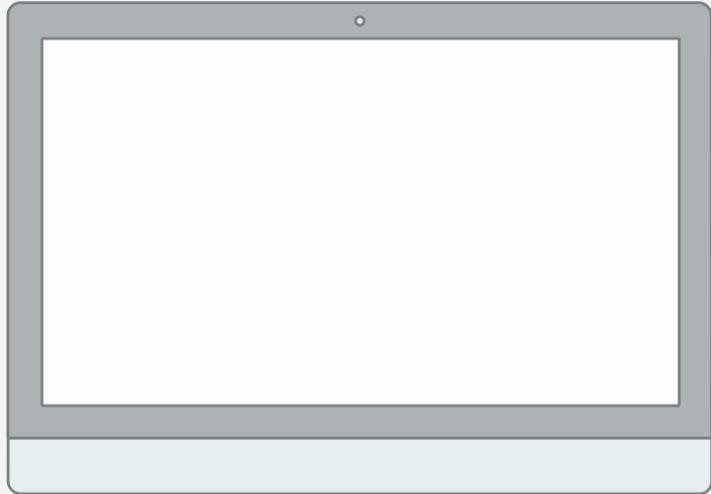
Windows and pointers.

Multi-tasking, multiple windows.

Exposees and leverages the file system.

Many specialized tools in concert.

Underneath, it's the 1970s, UNIX, and the command-line.



Plain-Text Tools for Data Analysis



Plain-Text Tools for Data Analysis



Better than they've ever been!



Plain-Text Tools for Data Analysis



Better than they've ever been!

Free! Open! Powerful!



Plain-Text Tools for Data Analysis



Better than they've ever been!

Free! Open! Powerful!

Friendly community! Many resources!



Plain-Text Tools for Data Analysis



Better than they've ever been!

Free! Open! Powerful!

Friendly community! Many resources!

But grounded in a UI paradigm that is increasingly far away from the everyday use of computing devices



Plain-Text Tools for Data Analysis



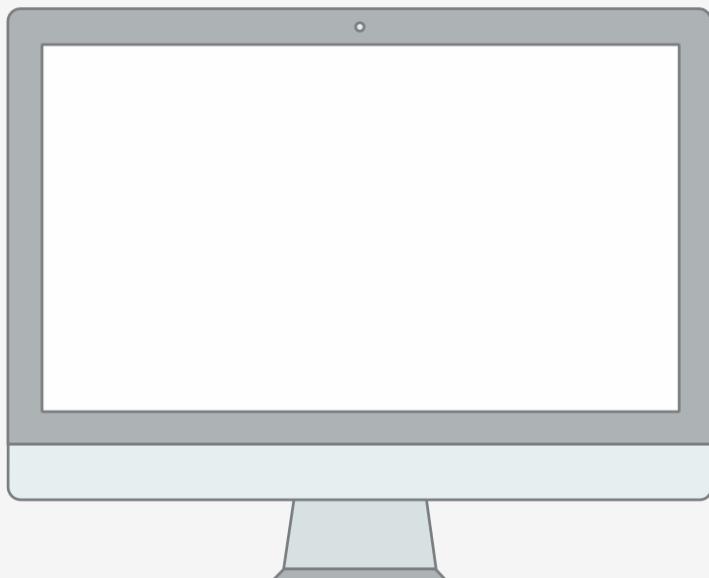
Better than they've ever been!

Free! Open! Powerful!

Friendly community! Many resources!

But grounded in a UI paradigm that is increasingly far away from the everyday use of computing devices

So why do we use these tools?



The research process is
intrinsically messy

The research process is
intrinsically messy

A rough distinction: "Office" vs "Engineering" approaches

Questions

What is "real" in your project?

What is the final output?

How is it produced?

How are changes managed?

Different Answers

In the Office model

Formatted documents are real.

Intermediate outputs are cut and pasted
into documents.

Changes are tracked inside files.

Final output is often in the same format
you've been working in, e.g. a Word file, or
perhaps a PDF.

Different Answers

In the Office model

Formatted documents are real.

Intermediate outputs are cut and pasted into documents.

Changes are tracked inside files.

Final output is often in the same format you've been working in, e.g. a Word file, or perhaps a PDF.

In the Engineering model

Plain-text files are real.

Intermediate outputs are produced via code, often inside documents.

Changes are tracked outside files.

Final outputs are assembled programatically and converted to a desired output format.

Different strengths and weaknesses

Everyone knows Word, Excel, or Google Docs.

"Track changes" is powerful and easy.

Hm, why can't I remember how I made this figure?

Where did this table of results come from?

Paper_Submitted_Final_edits_FINAL_kh-1.docx

Different strengths and weaknesses

Everyone knows Word, Excel, or Google Docs.

"Track changes" is powerful and easy.

Hm, why can't I remember how I made this figure?

Where did this table of results come from?

Plain text is universally portable.

Push button, recreate analysis.

Why can't I make R do this simple thing?

This version control stuff is a pain.

Object of type 'closure' is not subsettable

Paper_Submitted_Final_edits_FINAL_kh-1.docx

Different strengths and weaknesses

Everyone knows Word, Excel, or Google Docs.

"Track changes" is powerful and easy.

Hm, why can't I remember how I made this figure?

Where did this table of results come from?

Plain text is universally portable.

Push button, recreate analysis.

Why can't I make R do this simple thing?

This version control stuff is a pain.

Object of type 'closure' is not subsettable

Each approach generates solutions to its own problems, too

Paper_Submitted_Final_edits_FINAL_kh-1.docx

INTO THE KITCHEN



RStudio is an IDE for R



A kitchen is an IDE for Meals



R & RStudio

The screenshot shows the RStudio interface with the following components:

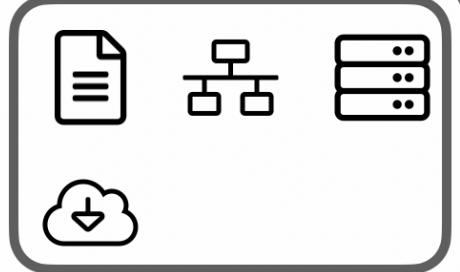
- Code Editor:** Displays the file `covdata.Rmd` containing R Markdown code. The code includes a header section and a code block starting with ````{r setup}`. A note about the `tidyverse` package is present.
- Console:** Shows the R startup message, the availability of the `knitr` hook, and the loading of the `testthat` package.
- Environment:** Shows the Global Environment pane with a single entry for `set`.
- Plots:** No plots are currently displayed.
- Packages:** Shows the contents of the `covdata` directory, which includes files like `.github`, `.gitignore`, `.Rbuildignore`, `.Rhistory`, `_pkgdown.yml`, `_sinewconfig.yml`, `covdata.Rproj`, `data`, `data-raw`, `DESCRIPTION`, `inst`, `LICENSE`, `LICENSE.md`, `man`, and `NAMESPACE`.
- Help:** No help pages are currently selected.
- Viewer:** No files are currently being viewed.
- File Browser:** Shows the file structure of the `covdata` project.

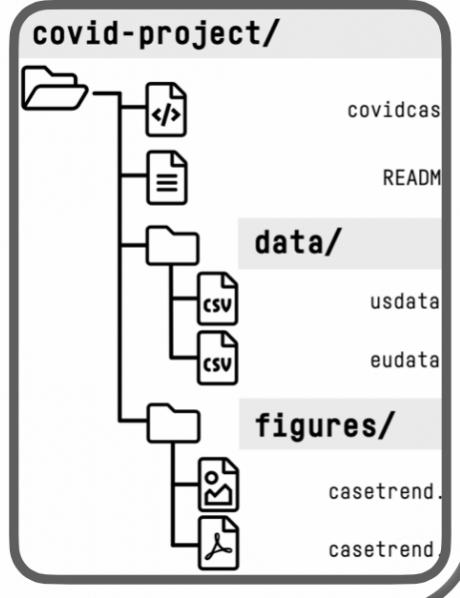
R & RStudio

```
# COVID      covidcases.Rmd

## Get data from ECDC
```{r get-data}
covid_raw <- get_ecdc[url]
```

## Get data from the US
```{r get-data}
us_raw <- get_us[url]
```
>_
```

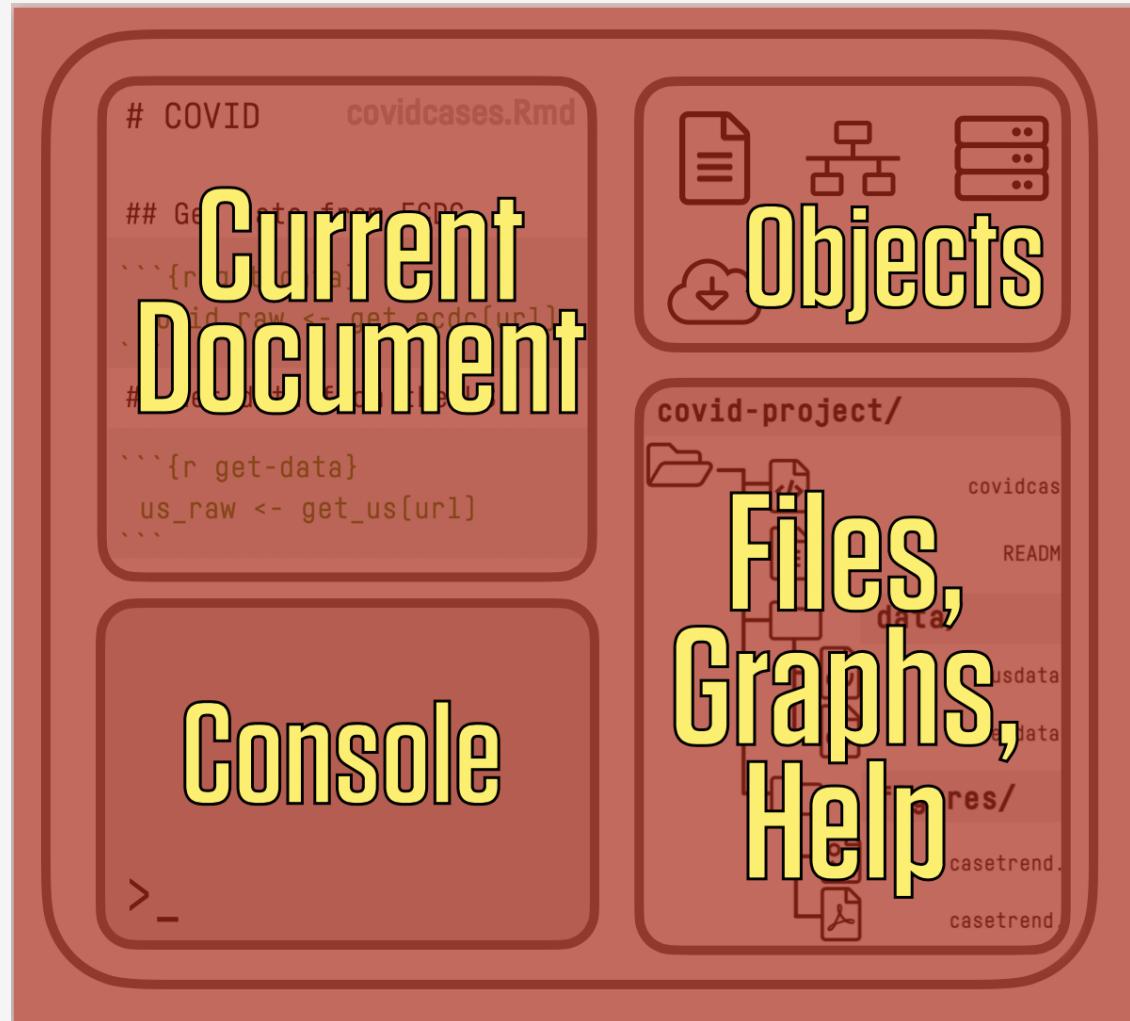




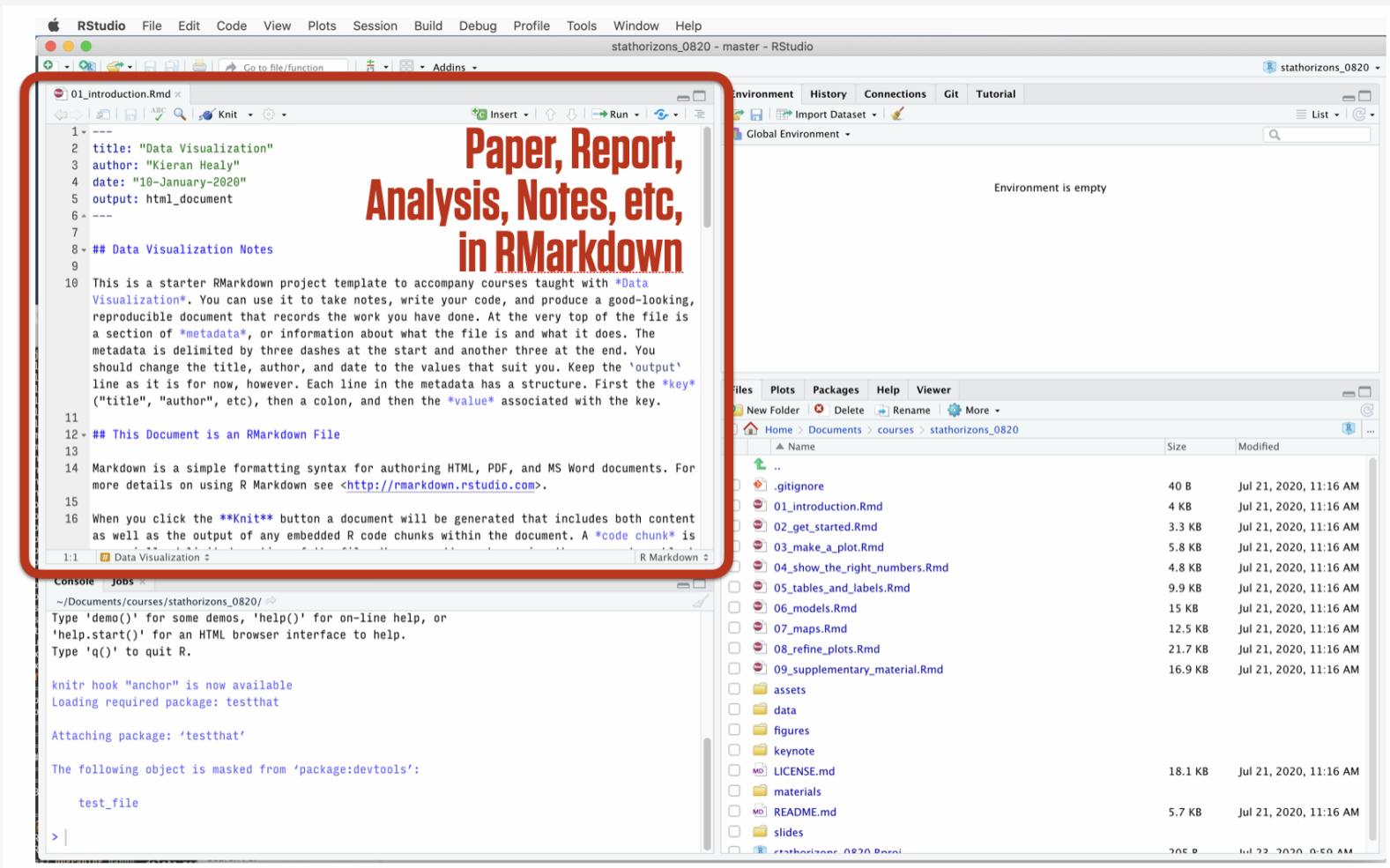
covid-project/

- covidcases
- README
- data/**
 - usdata
 - eudata
- figures/**
 - casetrend.
 - casetrend.

R & RStudio



RStudio



R & RStudio

The screenshot shows the RStudio desktop application interface. The main window has the following components:

- Top Bar:** RStudio, File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Window, Help.
- Title Bar:** stathorizons_0820 - master - RStudio.
- Left Panel:** A code editor showing a R Markdown file named "01_introduction.Rmd". The content includes metadata (title, author, date) and a section titled "## Data Visualization Notes".
- Middle Panel:** The Environment pane displays "Environment is empty".
- Right Panel:** The Files pane shows a directory tree for "stathorizons_0820" containing various R Markdown files (e.g., 01_introduction.Rmd, 02_get_started.Rmd, etc.) and other project files like LICENSE.md and README.md.
- Bottom Panel:** The Console pane contains R command history and output, including the loading of the testthat package and the creation of a test_file object.

A large red box highlights the Console pane, which contains the text:

Console: Type or send code here, see results

```
~/Documents/courses/stathorizons_0820/ 
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

knitr hook "anchor" is now available
Loading required package: testthat

Attaching package: 'testthat'

The following object is masked from 'package:devtools':

  test_file

>
```

R & RStudio

The screenshot shows the RStudio interface with a project titled "stathorizons_0820". The left pane displays an R Markdown file named "01_introduction.Rmd" containing metadata and introductory text. The right pane shows the "Environment" tab with a message "Environment is empty". A red box highlights the "Files" tab in the bottom right corner, which lists the contents of the project directory:

| Name | Size | Modified |
|-------------------------------|---------|------------------------|
| .. | | |
| .gitignore | 40 B | Jul 21, 2020, 11:16 AM |
| 01_introduction.Rmd | 4 KB | Jul 21, 2020, 11:16 AM |
| 02_get_started.Rmd | 3.3 KB | Jul 21, 2020, 11:16 AM |
| 03_make_a_plot.Rmd | 5.8 KB | Jul 21, 2020, 11:16 AM |
| 04_show_the_right_numbers.Rmd | 4.8 KB | Jul 21, 2020, 11:16 AM |
| 05_tables_and_labels.Rmd | 9.9 KB | Jul 21, 2020, 11:16 AM |
| 06_models.Rmd | 15 KB | Jul 21, 2020, 11:16 AM |
| 07_maps.Rmd | 12.5 KB | Jul 21, 2020, 11:16 AM |
| 08_refine_plots.Rmd | 21.7 KB | Jul 21, 2020, 11:16 AM |
| 09_supplementary_material.Rmd | 16.9 KB | Jul 21, 2020, 11:16 AM |
| assets | | |
| data | | |
| figures | | |
| keynote | | |
| LICENSE.md | 18.1 KB | Jul 21, 2020, 11:16 AM |
| materials | | |
| README.md | 5.7 KB | Jul 21, 2020, 11:16 AM |
| slides | | |
| stathorizons_0820.Rproj | 205 B | Jul 22, 2020, 8:50 AM |

Project files, Plots, Help

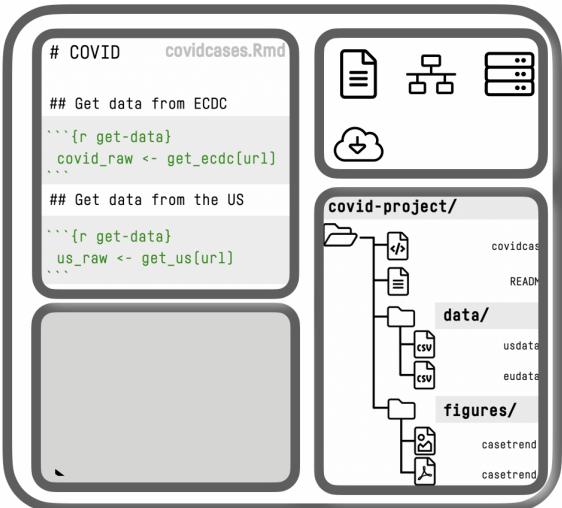
R & RStudio

The screenshot shows the RStudio interface with several panes:

- Code Pane:** Displays the file `01_introduction.Rmd` containing R Markdown code. The code includes metadata (title, author, date, output) and a section titled "## Data Visualization Notes".
- Environment Pane:** Shows the Global Environment, which is currently empty.
- File Explorer:** Shows the project structure under `stathorizons_0820`, including files like `01_introduction.Rmd`, `02_get_started.Rmd`, and `03_make_a_plot.Rmd`.
- Console Pane:** Displays R session logs, including the loading of the `testthat` package and the creation of a `test_file`.
- Knitr Hook:** A message indicates that the `"anchor"` knitr hook is now available.
- Text at the bottom:** The text "Inspect objects you create" is overlaid in red within a red-bordered box.

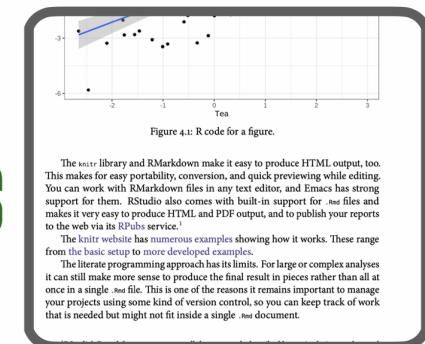
R & RStudio

RStudio



Drive R

```
if (is.empty.model(mt)) {  
  x <- NULL  
  z <- list(coefficients = if (m1m)  
    matrix(NA_real_, 0,  
    ncol(y)) else numeric(), residuals = y,  
    fitted.values = 0 *  
    y, weights = w, rank = 0L, df.residual = if  
    (!is.null(w)) sum(w !=  
    0) else ny)
```

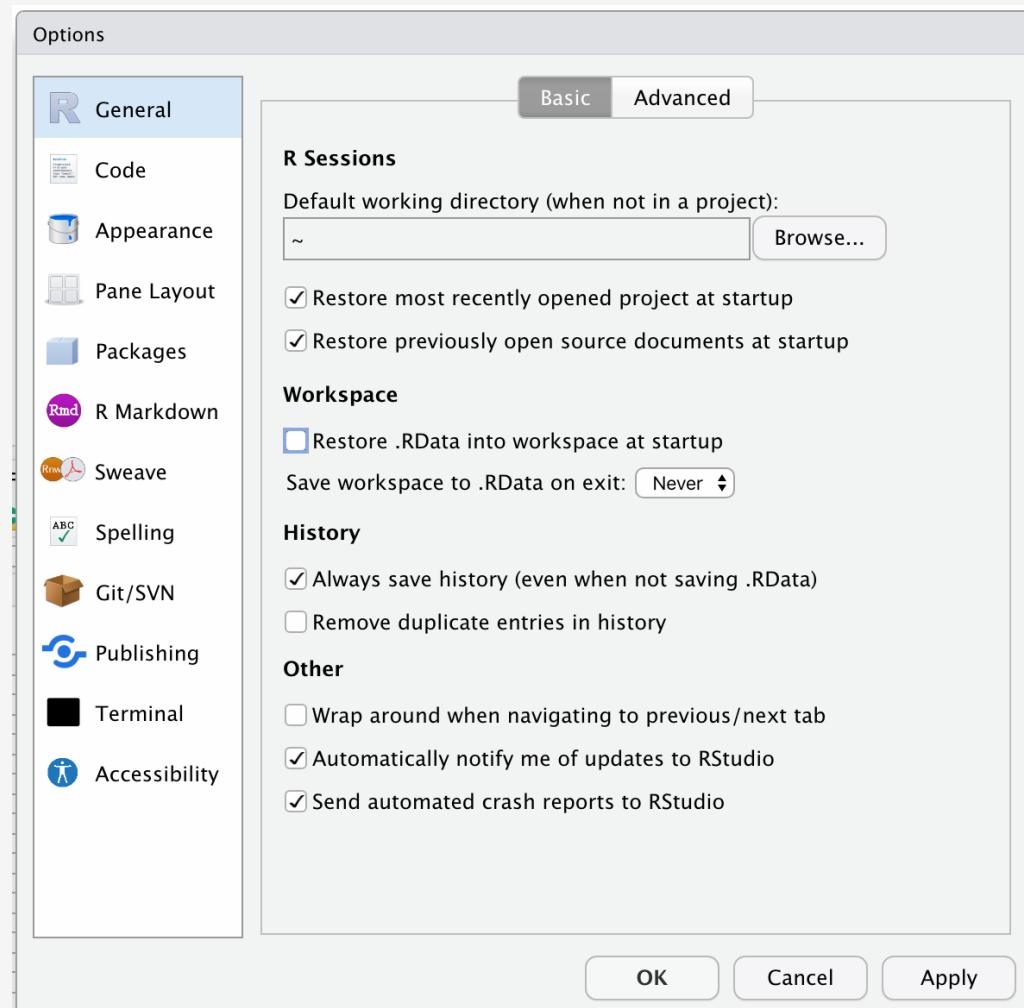


Generate Documents

View & Manage Environment

| Name | Type | Value |
|-------------|-----------------------------------|-------------------------------------|
| p | list [9] (S3: gg, ggplot) | List of length 9 |
| data | list [1704 x 6] (S3: tbl_df, tbl) | Aibble with 1704 rows and 6 columns |
| layers | list [0] | List of length 0 |
| scales | environment [1] (S3: ScalesList) | <environment: 0x7f8f08c1e010> |
| mapping | list [3] (S3: uneval) | List of length 3 |
| theme | list [0] | List of length 0 |
| coordinates | environment [5] (S3: CoordCa) | <environment: 0x7f8f08c27b40> |
| facet | environment [2] (S3: FacetNul) | <environment: 0x7f8f08c55210> |
| plot_env | environment [6] | <environment: R_GlobalEnv> |
| labels | list [3] | List of length 3 |

Your code is what's real in your project



Consider not showing your output inline

