

# Data Wrangling with R and the Tidyverse

Data Wrangling: Session 1

Kieran Healy

Statistical Horizons, October 2023

# Housekeeping

# Housekeeping

**10am till 12:30pm US EST each day**

# **Housekeeping**

**10am till 12:30pm US EST each day**

**1:30pm to 3:30pm US EST each day**

# Housekeeping

**10am till 12:30pm US EST each day**

**1:30pm to 3:30pm US EST each day**

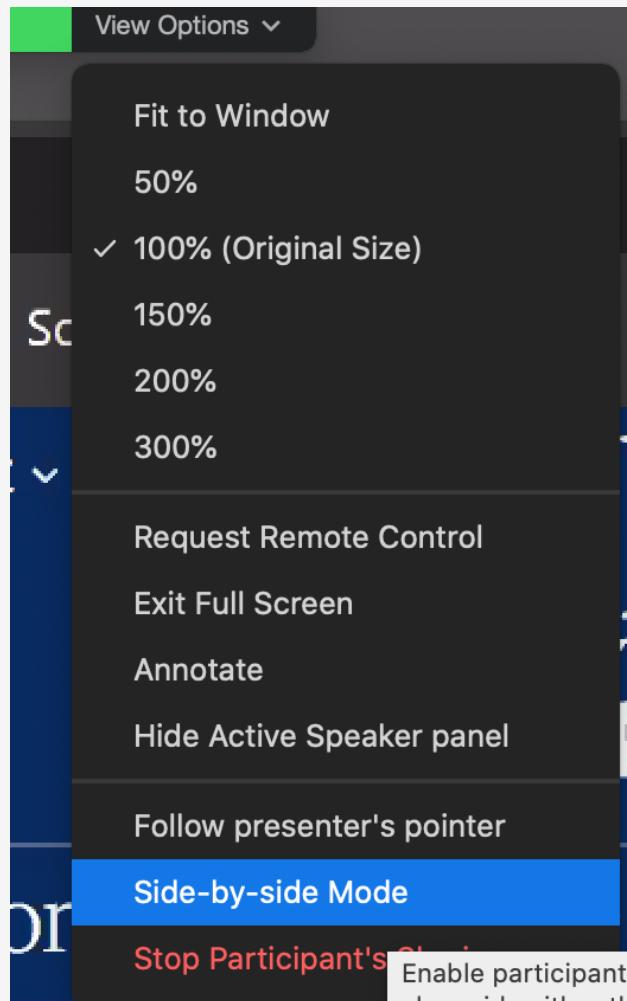
**Use the Zoom chat to ask questions, or raise a hand with**



# 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.

# My Setup and Yours

# My Setup and Yours

Talking, Slides, and Live-Coding in RStudio

Follow along with RStudio yourself if you can

The course packet is also an RStudio project and the place for your notes

# Goals for this first session

# **Goals for this first session**

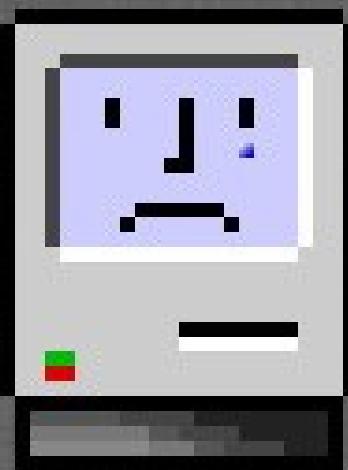
**Some big-picture motivation & perspective**

**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.

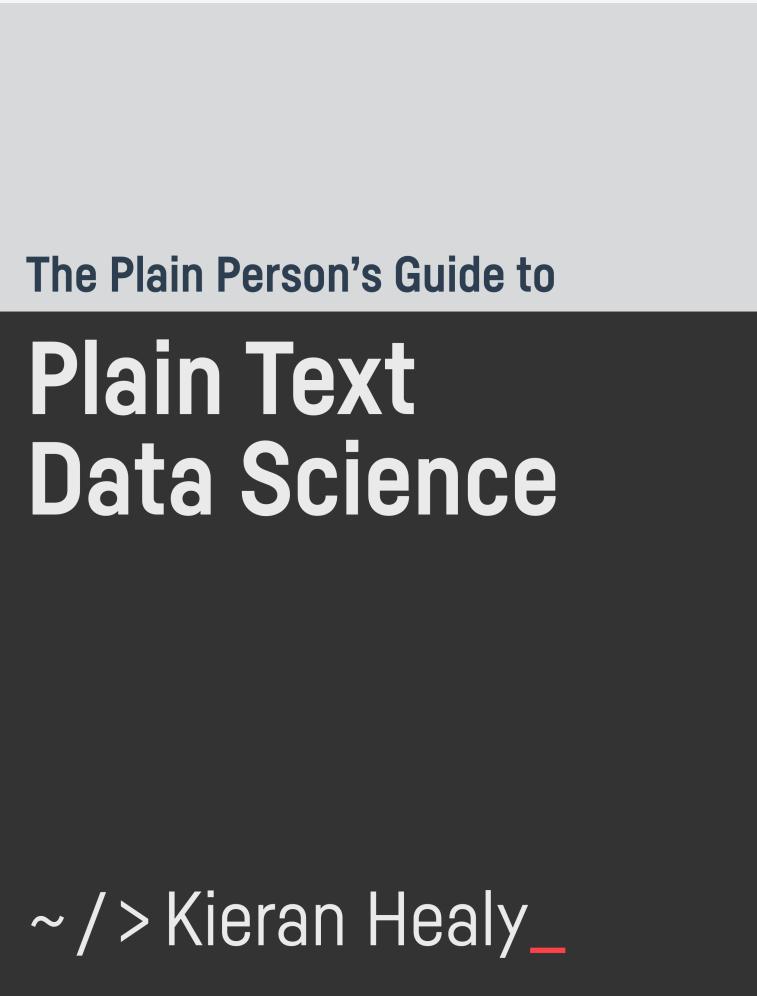
Sorry.

**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.

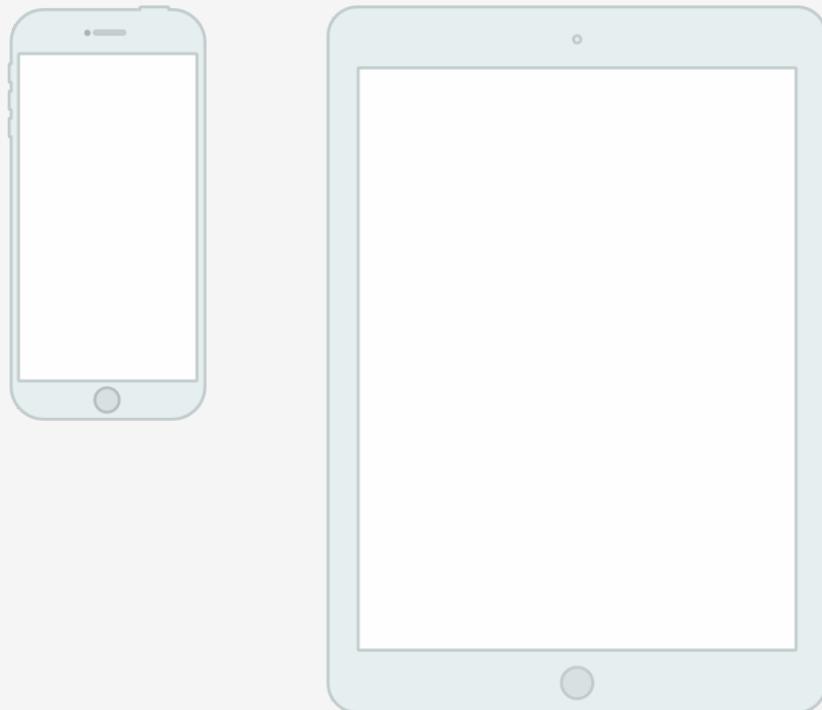
The same principles would apply to, e.g., using Python or similar tools.

# Two revolutions in computing

# Where the action is

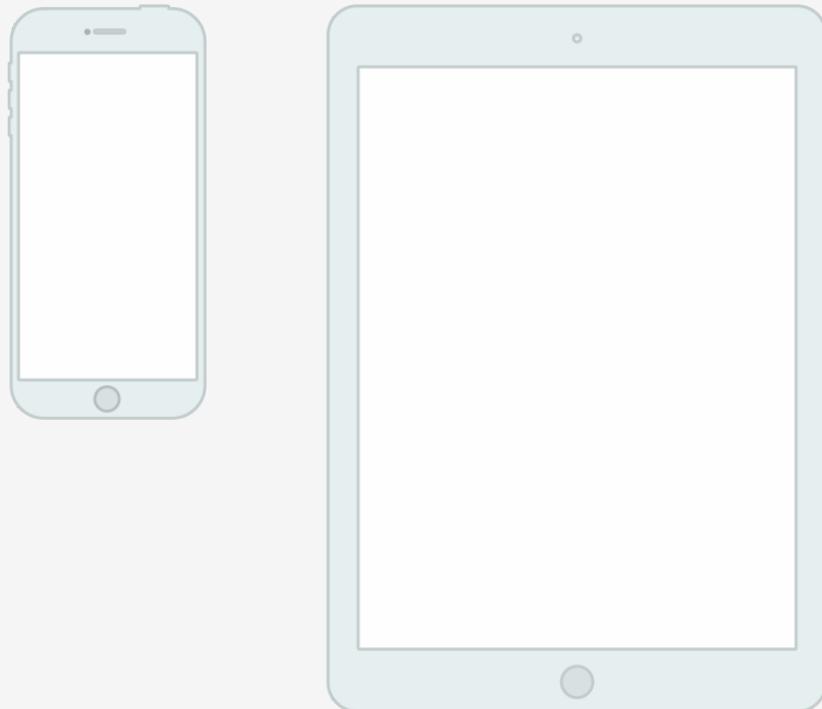


# Where the action is



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.

Exposes and leverages the file system.



# Where statistical computing lives



Windows and pointers.

Multi-tasking, multiple windows.

Exposes and leverages the file system.

Many specialized tools in concert.



# Where statistical computing lives



Windows and pointers.

Multi-tasking, multiple windows.

Exposes 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 programmatically 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?.

Paper\_Submitted\_Final\_edits\_FINAL\_kh-1.docx

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

# 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

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

# **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, a note about the covdata package, and a setup block for tidyverse.
- Environment Pane:** Shows the global environment with a single entry: `set function (name, value)`.
- Console:** Displays R startup messages, including the R logo, a note about contributors, help information, and the loading of the `testthat` package.
- File Browser:** Shows the directory structure of the `covdata` project, listing files like `.github`, `.gitignore`, `.Rbuildignore`, `.Rhistory`, `_pkgdown.yml`, `_sinewconfig.yml`, `covdata.Rproj`, `data`, `data-raw`, `DESCRIPTION`, `inst`, `LICENSE`, `LICENSE.md`, `man`, and `NAMESPACE`.

# R & RStudio

```
# COVID      covidcases.Rmd

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

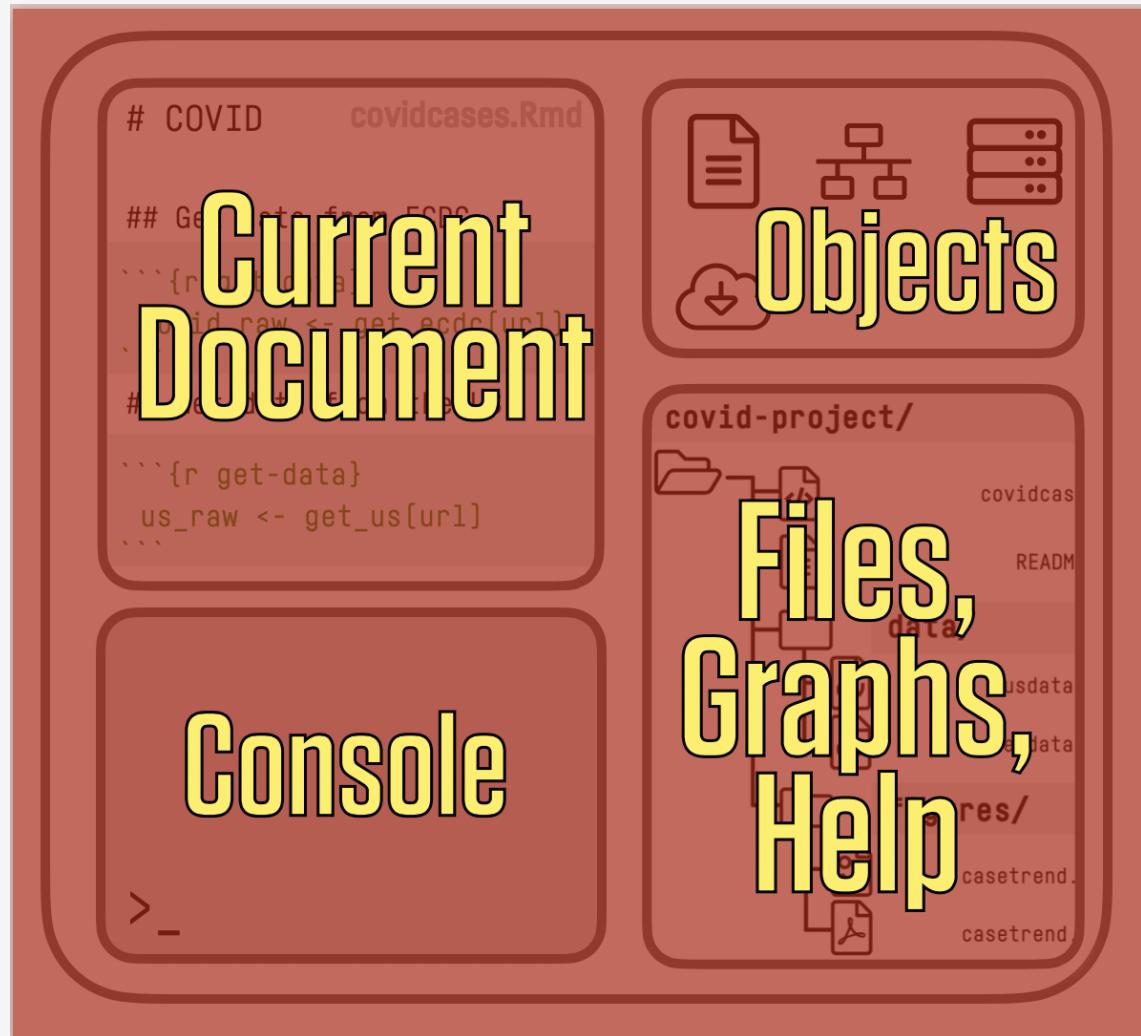
## Get data from the US
```{r get-data}
us_raw <- get_us[url]
```

```

The diagram illustrates a project structure named "covid-project". It contains several sub-directories and files:

- covidcas**: A folder containing a file named **README**.
- data/**: A folder containing two sub-folders: **usdata** and **eudata**. Each of these sub-folders contains two CSV files.
- figures/**: A folder containing two files: **casetrend.** and **casetrend.**

# R & RStudio



# RStudio

The screenshot shows the RStudio interface. The main area is divided into several panes:

- Code Editor (highlighted by a red box):** Displays the R Markdown file `01_introduction.Rmd`. The content includes metadata (title, author, date) and a section titled "## Data Visualization Notes". A large red box highlights this section.
- Preview Pane:** Shows the rendered content of the R Markdown file, featuring a large red title: "Paper, Report, Analysis, Notes, etc, in RMarkdown".
- 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:** Displays R code and its output, including the loading of the `testthat` package and the creation of a `test_file`.

# R & RStudio

The screenshot shows the RStudio IDE interface. The top menu bar includes RStudio, File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Window, and Help. The title bar indicates the project is "stathorizons\_0820 - master - RStudio".

The left pane displays the code editor for "01\_introduction.Rmd". The content includes R Markdown metadata at the top, followed by sections like "## Data Visualization Notes" and "## This Document is an R Markdown File". A red box highlights the "Console" tab in the bottom-left corner.

The right pane contains several panes:

- Environment:** Shows "Global Environment" with the message "Environment is empty".
- File Browser:** Shows the directory structure of the project "stathorizons\_0820".
- Console:** Displays R session output, including the loading of packages and testthat, and a message about a knitr hook.

A large red box highlights the "Console" tab and its output area. Overlaid on the right side of the console area is the text:

**Console: Type or send code here, see results**

# 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-left 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       |         | Jul 22, 2020, 8:50 AM  |

**Project files, Plots, Help**

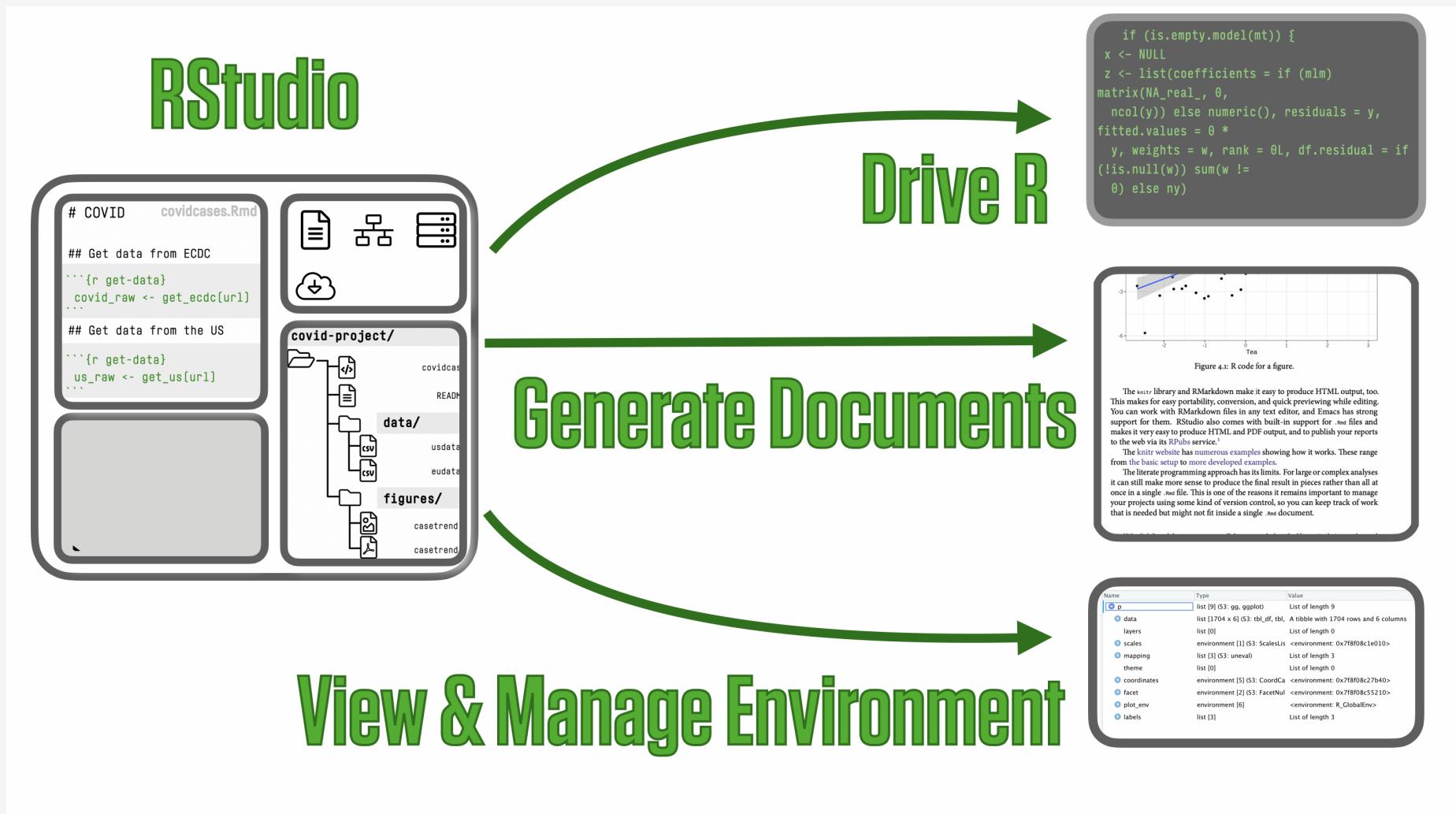
# R & RStudio

The screenshot shows the RStudio interface. On the left, the code editor displays an R Markdown file named '01\_introduction.Rmd'. The file contains metadata and introductory text about R Markdown. The right side of the interface includes the Environment pane, which is currently empty, and a file browser showing a directory structure for a project named 'stathorizons\_0820'.

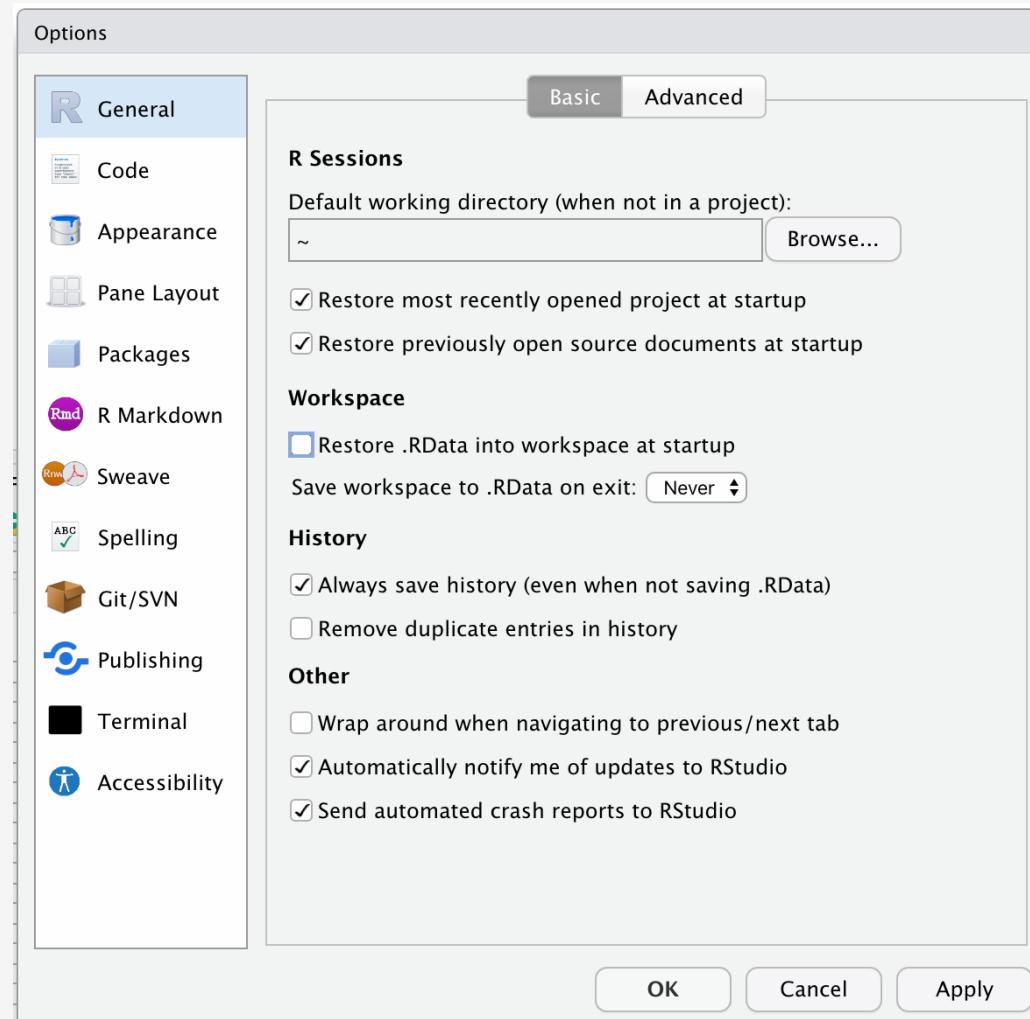
**Inspect objects you create**

| 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, 0:50 AM  |

# R & RStudio



# Your code is what's real in your project



# Consider not showing output inline

