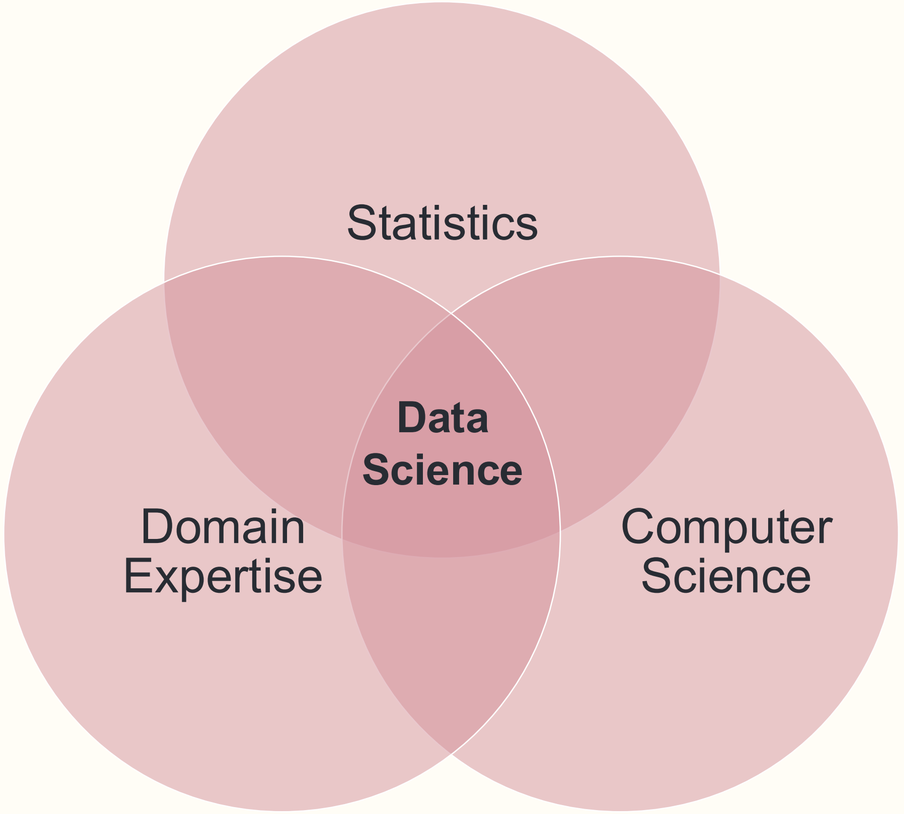
# Lecture 1: Introduction to course, R, and RStudio

**Heather Mattie, PhD**

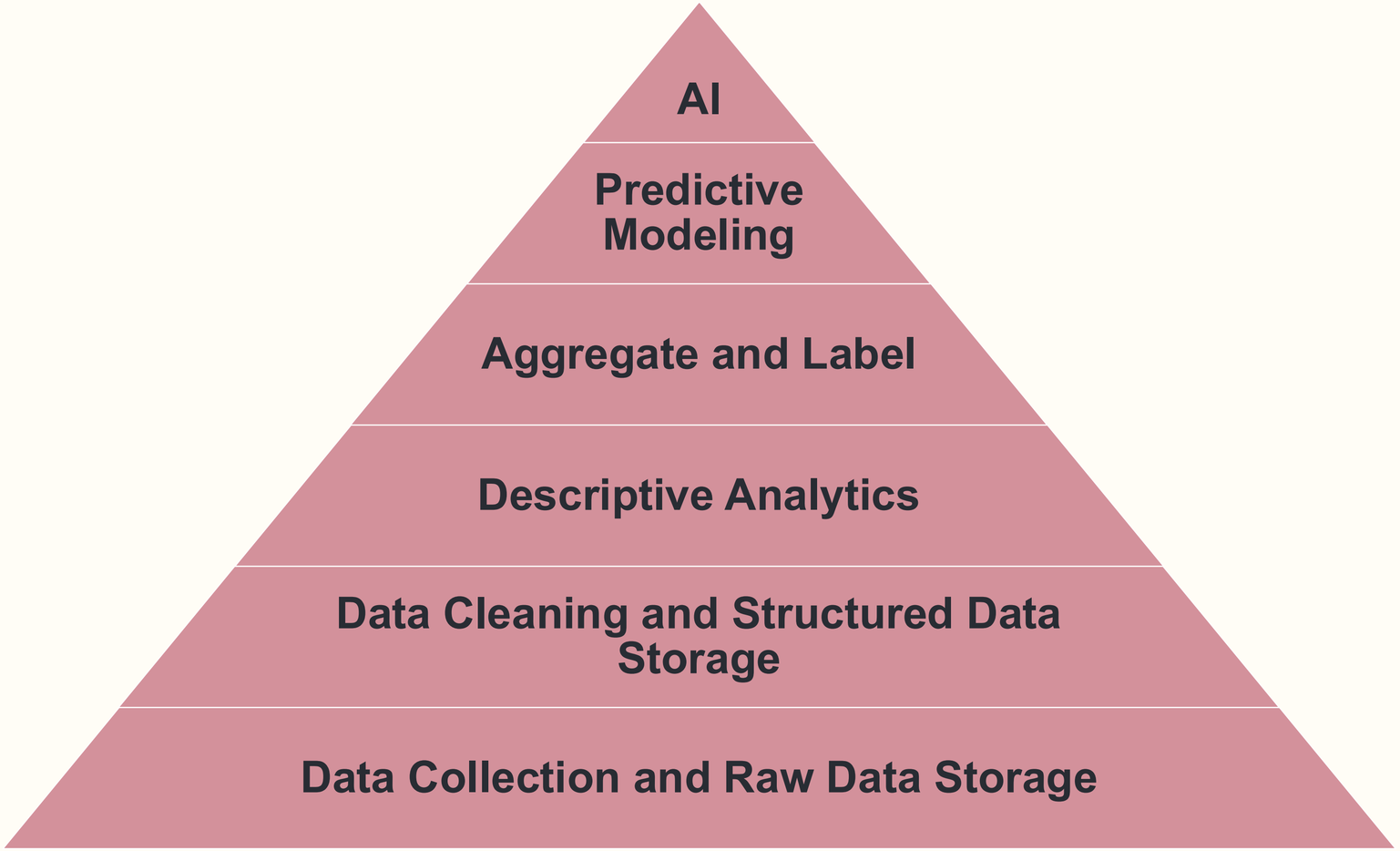
**September 2, 2025**

# What is Data Science?

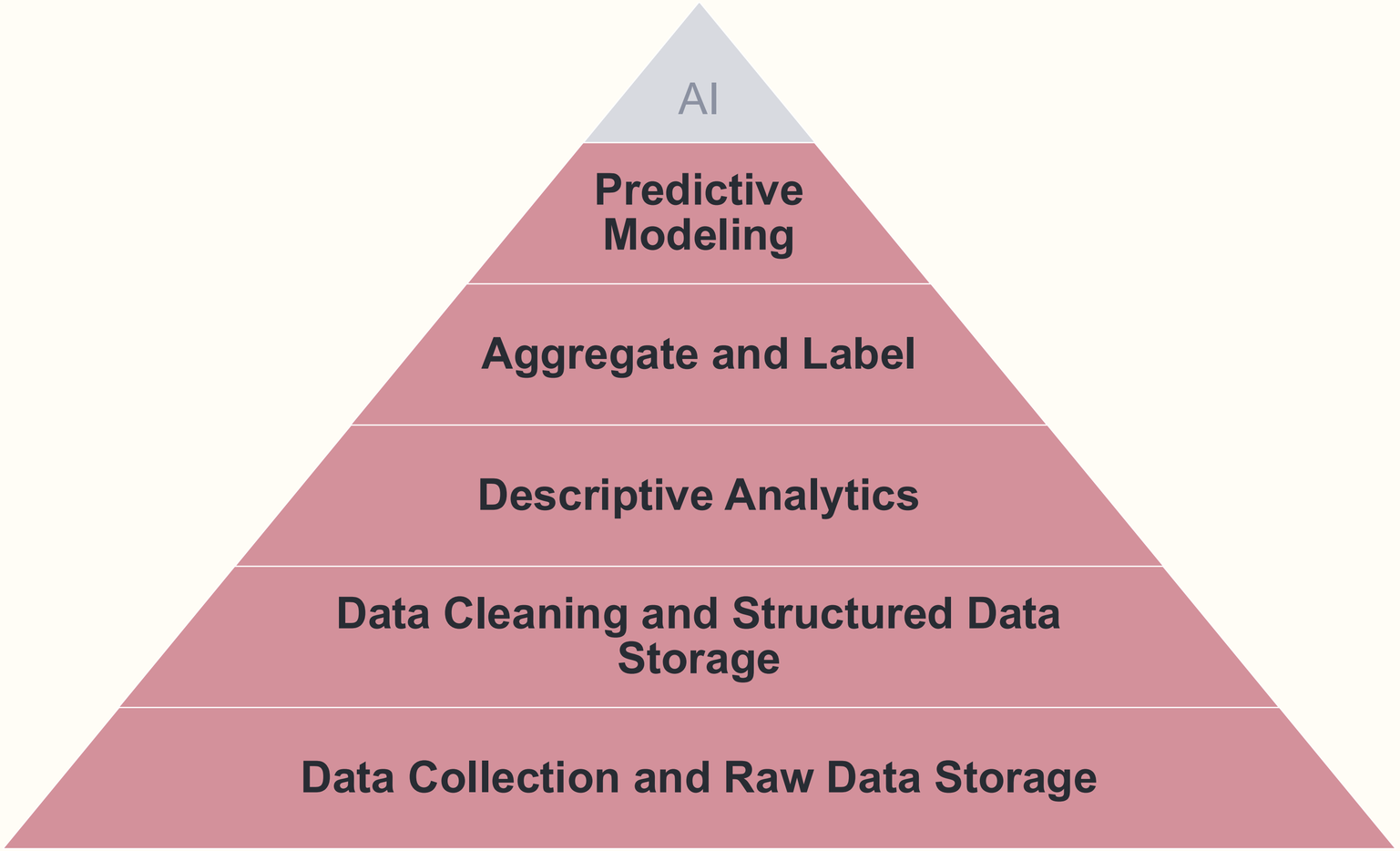
Data Science is a **multidisciplinary** field that uses scientific methods to **extract knowledge** and **insights** from **data**



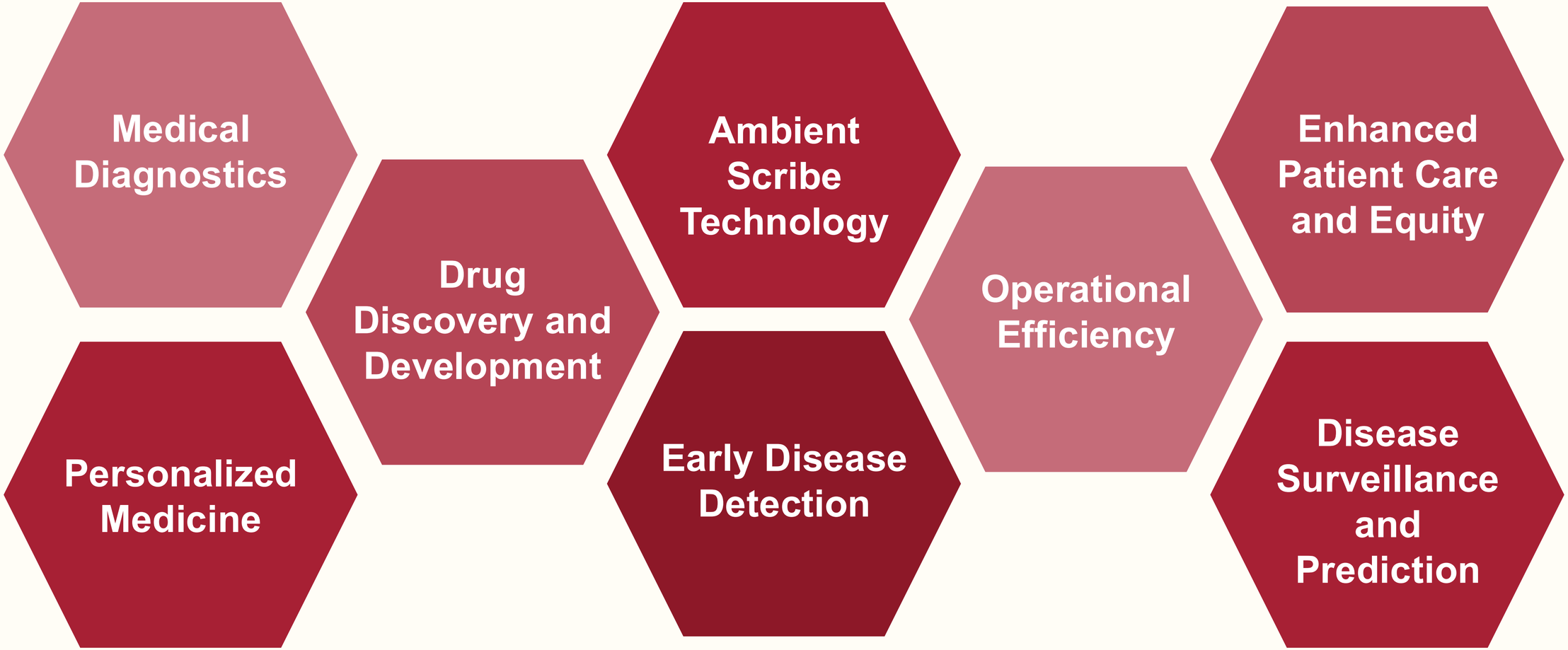
# Data Science Hierarchy of Needs



# Data Science Hierarchy of Needs



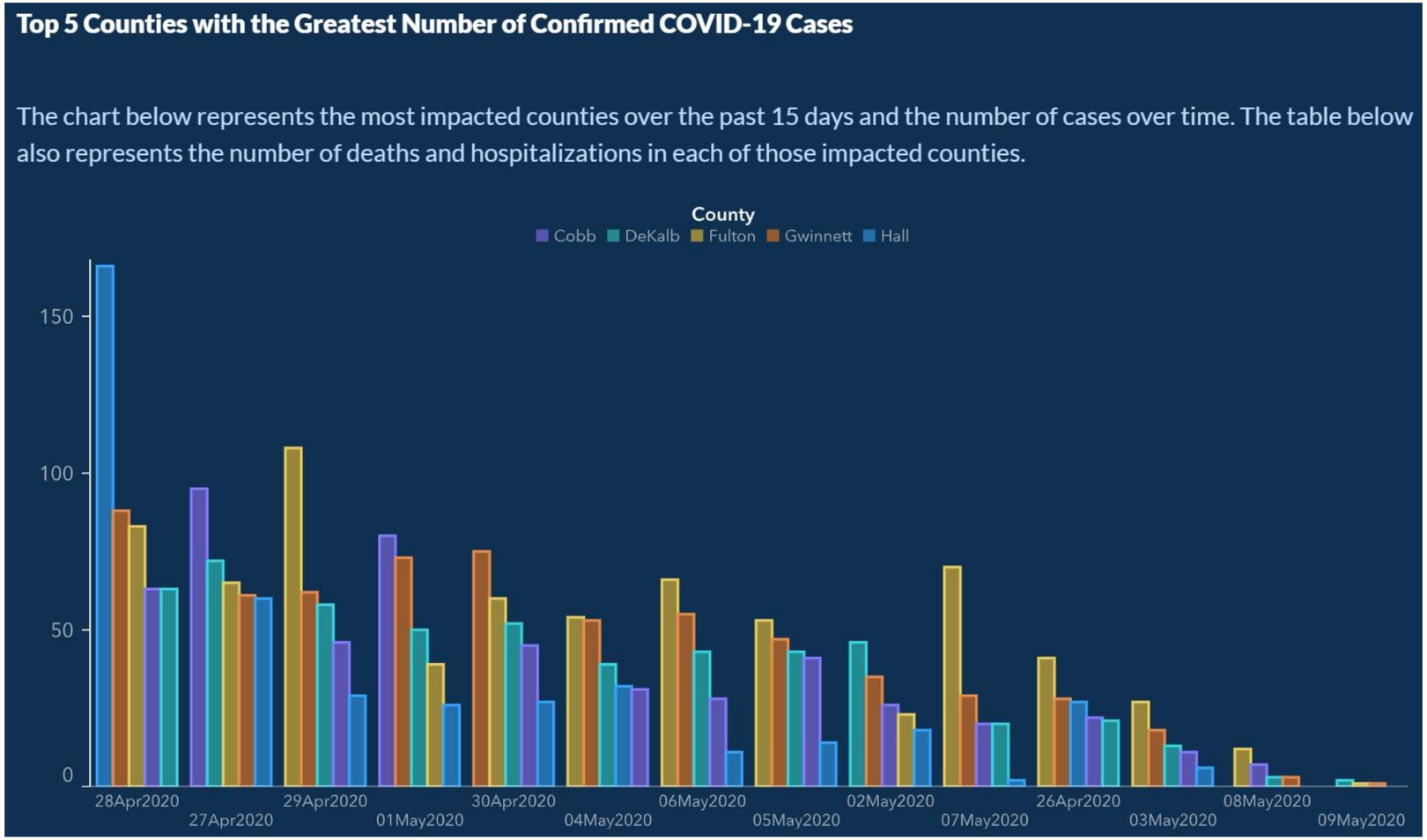
# Data Science in Public Health – Successes



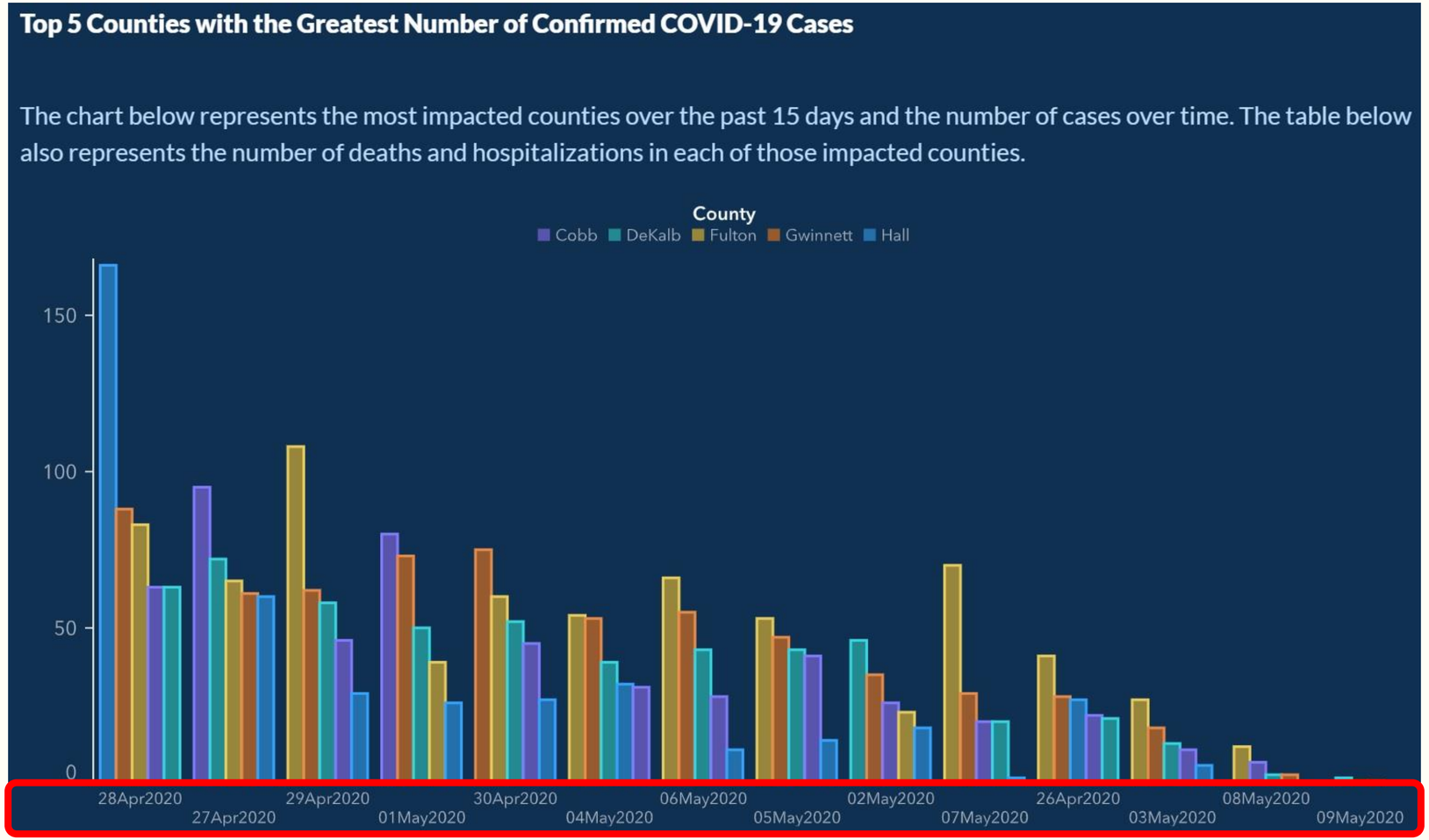
# Data Science in Public Health – Pitfalls

- Biased Algorithms
- Data Privacy and Security Risks
- Lack of Transparency
- AI Regulation Landscape
- Mis- and Disinformation

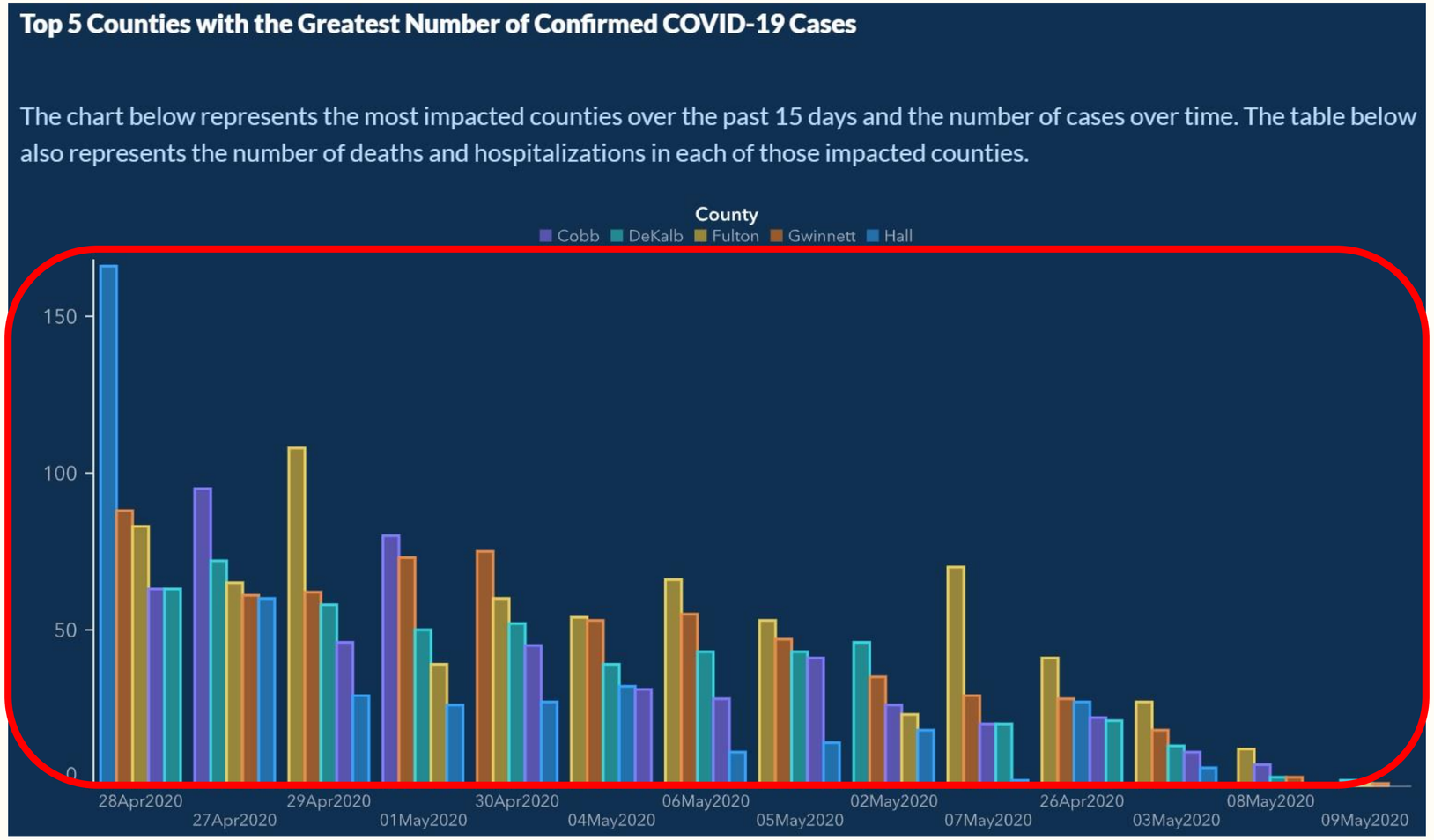
# Misinformation Example



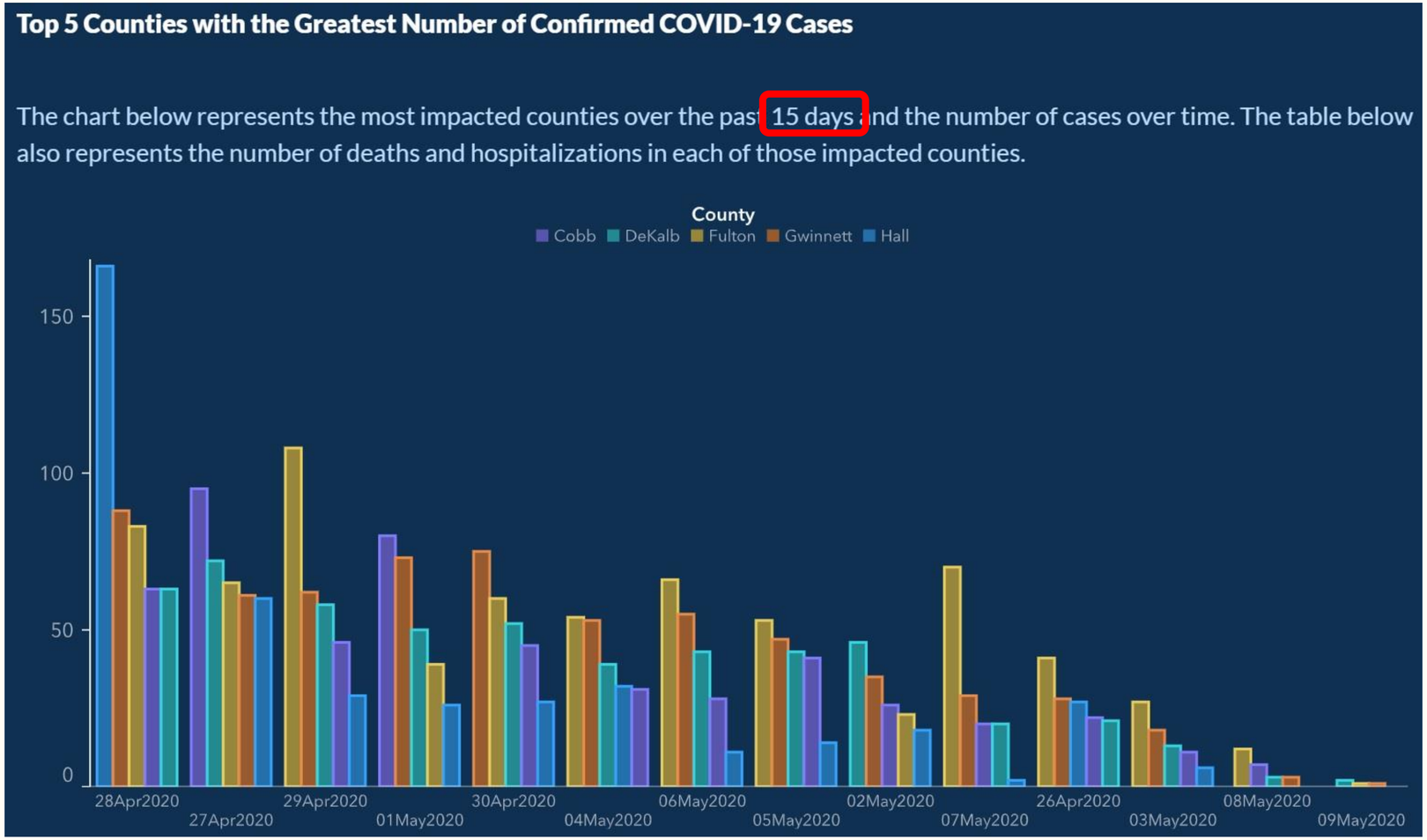
# Misinformation Example



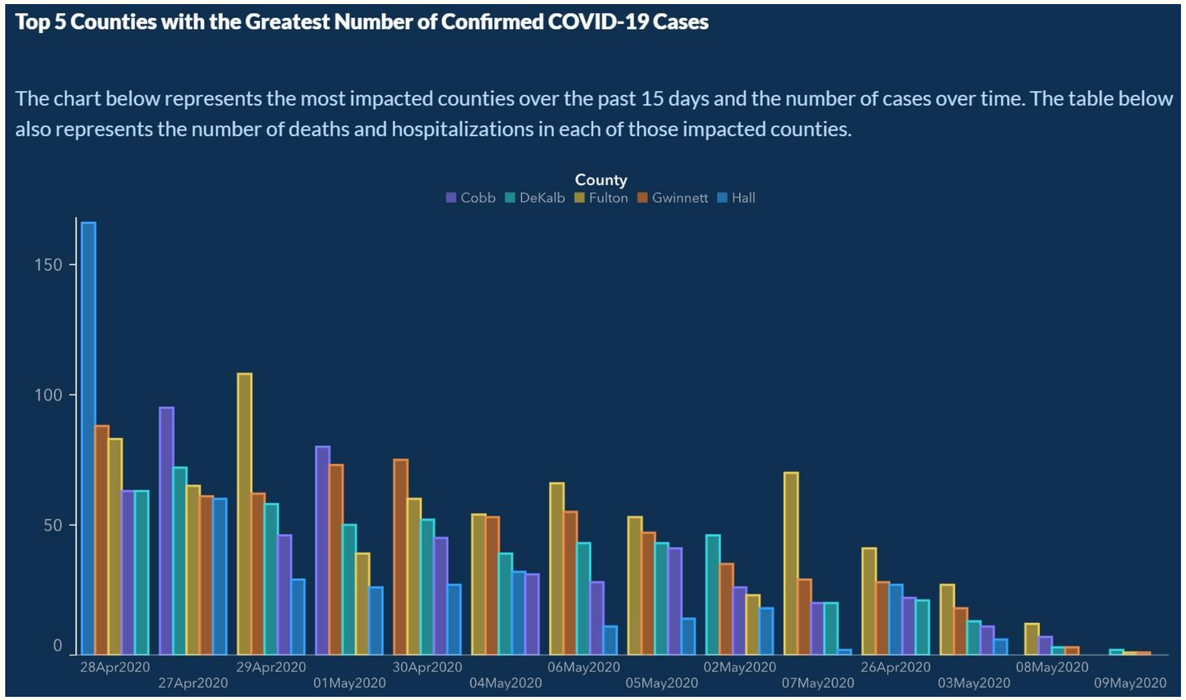
# Misinformation Example



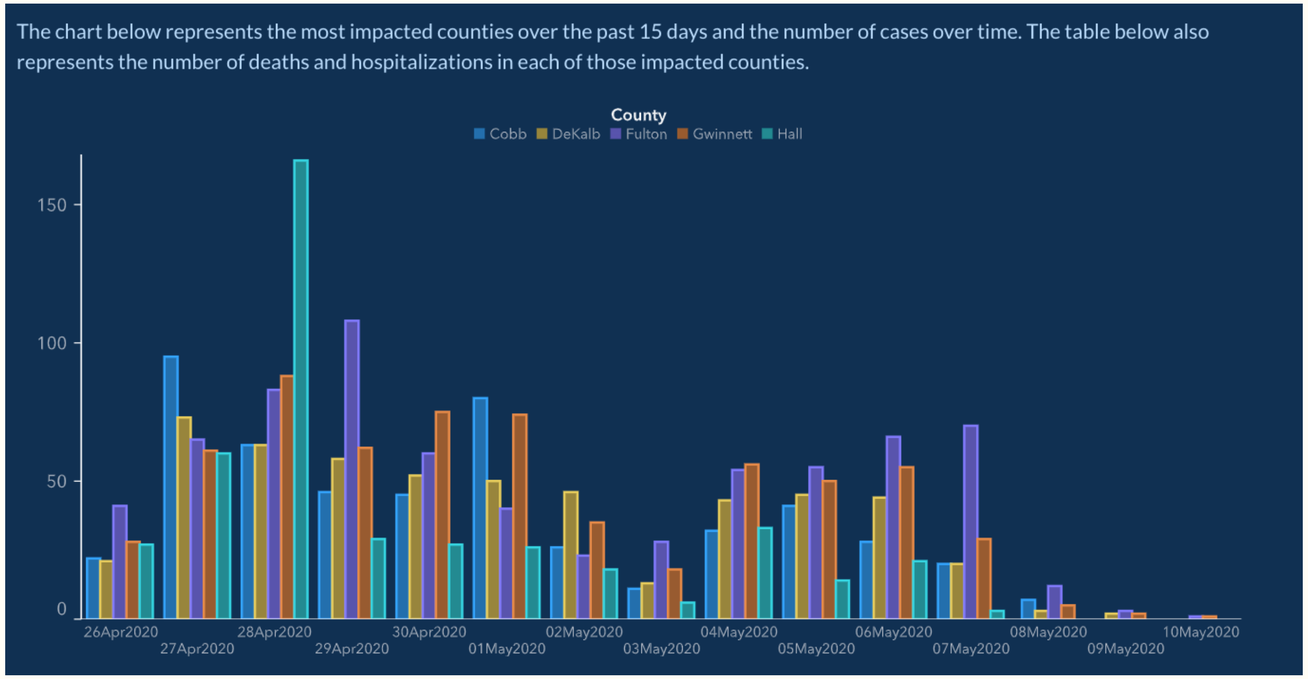
# Misinformation Example



# Misinformation Example




## Corrected version



As data
[scientists, analysts, creators, consumers],
we all play a role in **understanding** and **confronting**
**ethics**, **equity**, and **bias** in data, algorithms, clinical
AI, and healthcare

# The Data Science Pipeline

Collecting and importing (loading) the data

Processing (cleaning, wrangling) the data

Visualizing and summarizing the data

Building models (statistical and ML)

Interpretation and communication of results

# Who is helping you learn data science?

## Instructor

Heather Mattie
Lecturer on Biostatistics
hemattie@hsph.harvard.edu

## Teaching Fellows

Carmen Rodriguez Cabrera

crodriguezcabrera@g.harvard.edu

Claire Chu

clairechu@hsph.harvard.edu

Sajia Darwish

sajiadarwish@g.harvard.edu

# Office Hours and Lab

## Office Hours

| Day | Time | Staff | Location |
|---|---|---|---|
| TBD | TBD | Heather | Building 1, 4th floor, Room 421A |
| TBD | TBD | Carmen | TBD |
| TBD | TBD | Claire | TBD |
| TBD | TBD | Sajia | TBD |

## Lab

| Day | Time | Location |
|---|---|---|
| TBD | TBD | Zoom |

***Lab will not be held every week**

# Grading

## Homework

- 5 assignments
- 60% of final grade
- You are welcome to discuss the course material and homework questions with others, but the work you turn in must be your own. Be sure to cite any sources you use, including generative AI.

## Take-home Midterm

- 15% of final grade
- Questions that require writing code and short answers
- You are not allowed to work on or discuss this assignment with other students or use generative AI

# Grading

## Final Project

- 25% of final grade
- Will work in teams of 4-5 people
- Will choose 1 of 3 possible projects

## Pass/Fail Threshold

- If taking course pass/fail, must earn final grade of 70% or more to pass

## Auditing

- Auditors do not need to submit any assignments

# Generative AI Policy

As a tool, generative AI can provide valuable assistance in understanding concepts, troubleshooting issues, or even offering general guidance (use discretion!). However, to ensure true learning in an introduction to data science course, you should not rely on generative AI to complete your homework/midterm/final project. Make sure you understand the course content. **See the syllabus for more details on the generative AI policy.**

# Homework Assignments

- Real-world/public health/medical focus
- Scrape and wrangle/clean messy data
- Explore data
- Visualize data
- Perform statistical analyses
- Make predictions (build ML models)
- Communicate results

- Will be written in R using RMarkdown and submitted via private Github repositories
- One repository per student per assignment
- Only you and the teaching staff will have access to files in your repository
- Must also submit knitted html file
- Points will be deducted if we are unable to knit your RMarkdown file when grading
- Can use **2 late days per assignment**

# Midterm Exam

- You will have 1 week to complete the midterm exam
- Very similar to a homework assignment
  - Similar length
  - Submit via GitHub repositoy
  - R Markdown and html files must be submitted
- No collaboration of any kind is permitted
- Use of generative AI is not permitted

# Final Project

- Teams of 4-5 students
- Choose a project from list of 3
  - Will be given a dataset and prompt and will be tasked with data wrangling, data visualization, statistical analyses, building ML models and communicating your findings
- A TF will be assigned to each team to give advice throughout the project
  - Assigned in beginning of November

# Course Communication

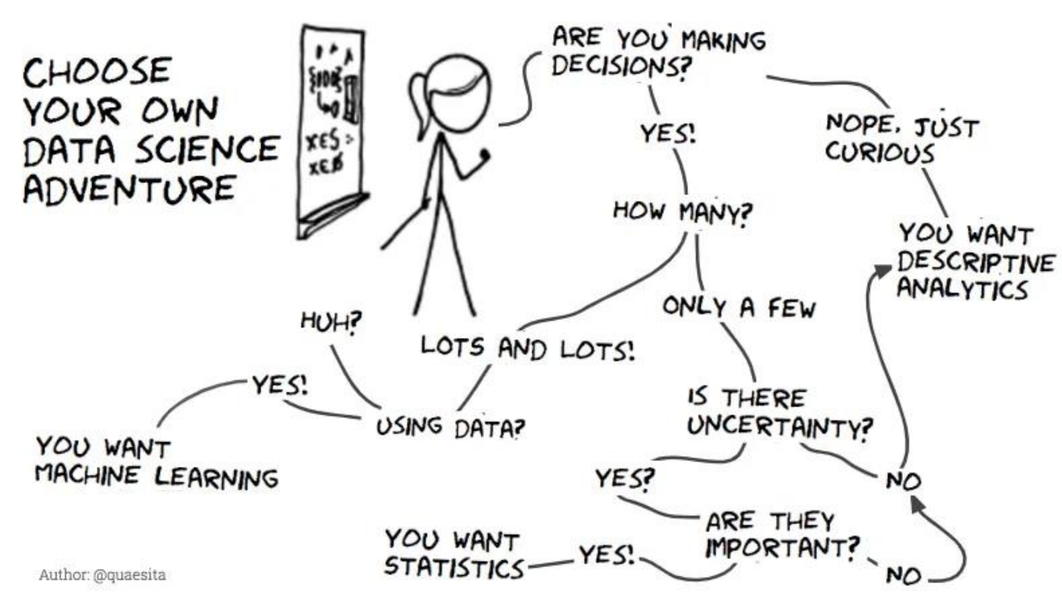
- Email me or the teaching fellows
- Do **not** send a Canvas message – I rarely check them!
- Office hours
- Lab sessions
- See me during the class break or after lecture

# Course Expectations

- You are encouraged, but not required, to attend lecture
  - Each lecture will be recorded and available on the Canvas course site
- Participation is not required or included as part of your final grade, but is highly encouraged
  - We all learn from each other
  - After a while I start to dislike the sound of my own voice
- Attending a weekly lab session is highly recommended but not required
  - The session will be recorded and available on the Canvas course site
- Break time - we'll take a 5- minute break around the middle of each lecture (45-50 minutes in)

# This course IS…

- An introduction to R and RStudio
- An introduction to creating publishable visualizations using **`ggplot2`**
- An introduction to using git and GitHub for version control and reproducibility
- Focused on foundational knowledge for coding, the data science project pipeline, machine learning, and communication of process and results
- A chance for you to collaborate with and learn from other students
- A safe space to “get your hands dirty” and write and debug code



# This course is NOT…

- A machine learning course
  - You will get a great foundation in ML from this course, but ML is only about 25% of the course
- Meant to be a silver bullet for all data science tasks / projects
  - Passing this course does not mean you won't have more to learn - data science is constantly evolving and we need to be lifelong learners

# Action Items

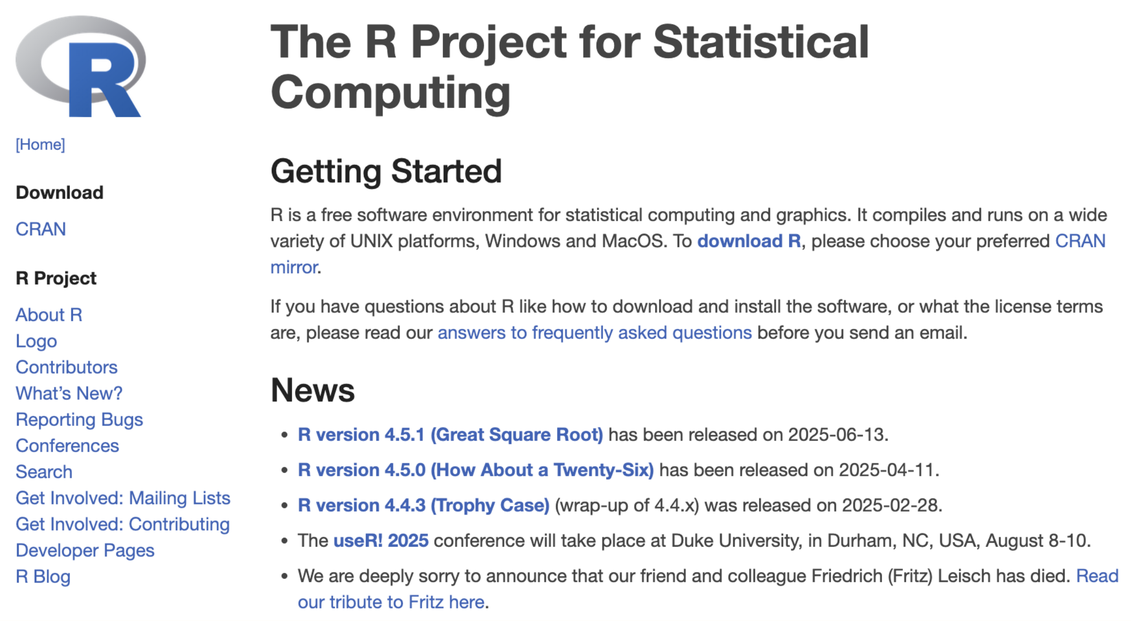
- Download and install R and RStudio
  - Make sure you **download R first**
- Create a GitHub account
  - Remember what your username is - you'll need it to complete the survey below
- Complete this survey
  - **We need this information in order to create homework repositories for you**

# Introduction to R

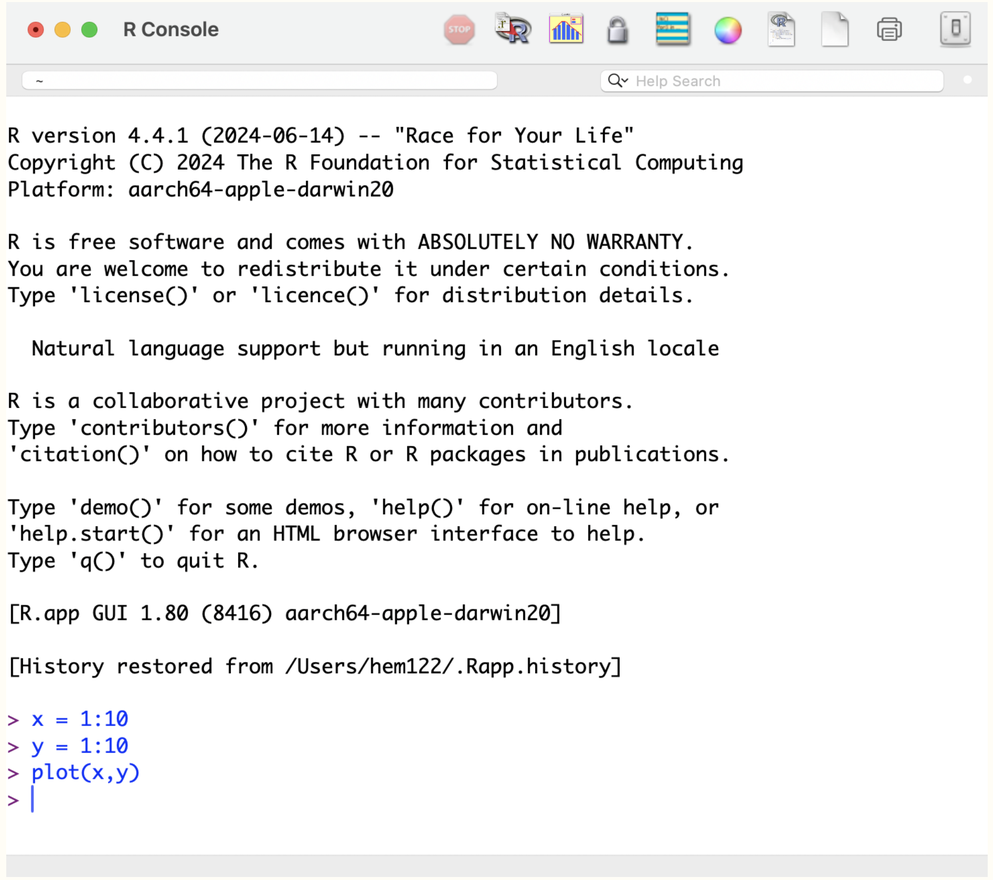
# What is R?

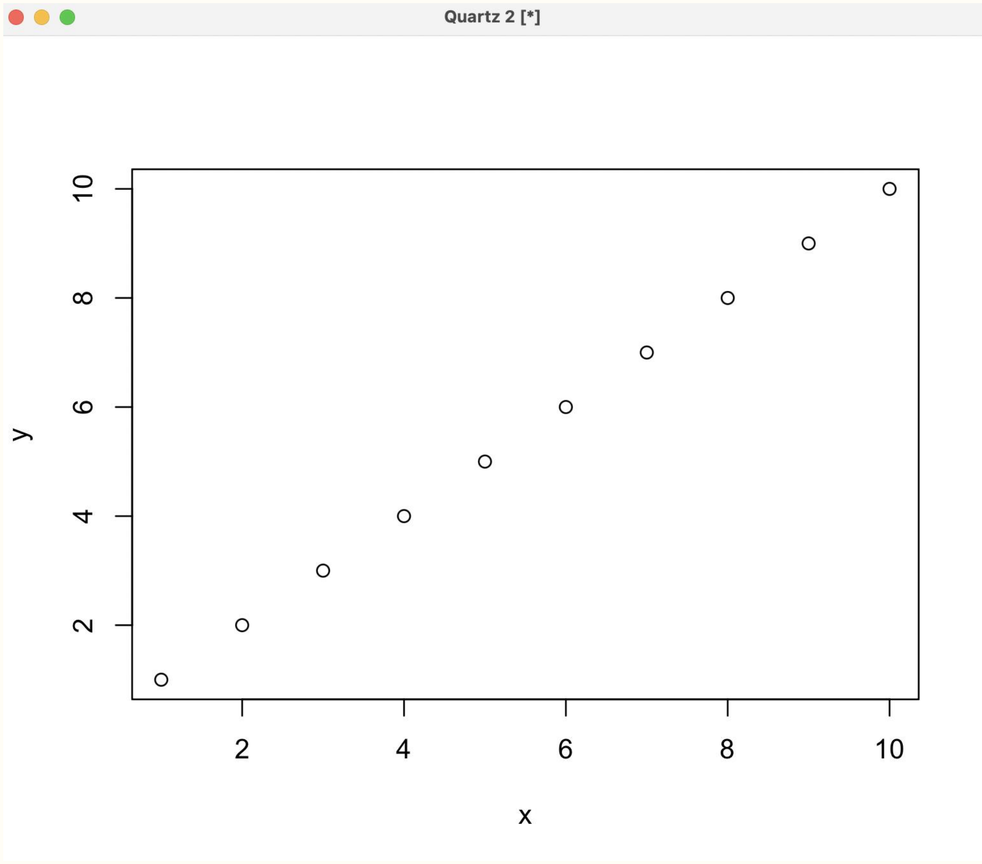
**R** is a programming language and software environment specifically designed for **statistical computing** and **graphics**

- Open source
- Specialized for statistics
- Active community
- Reproducible research



# What is R?

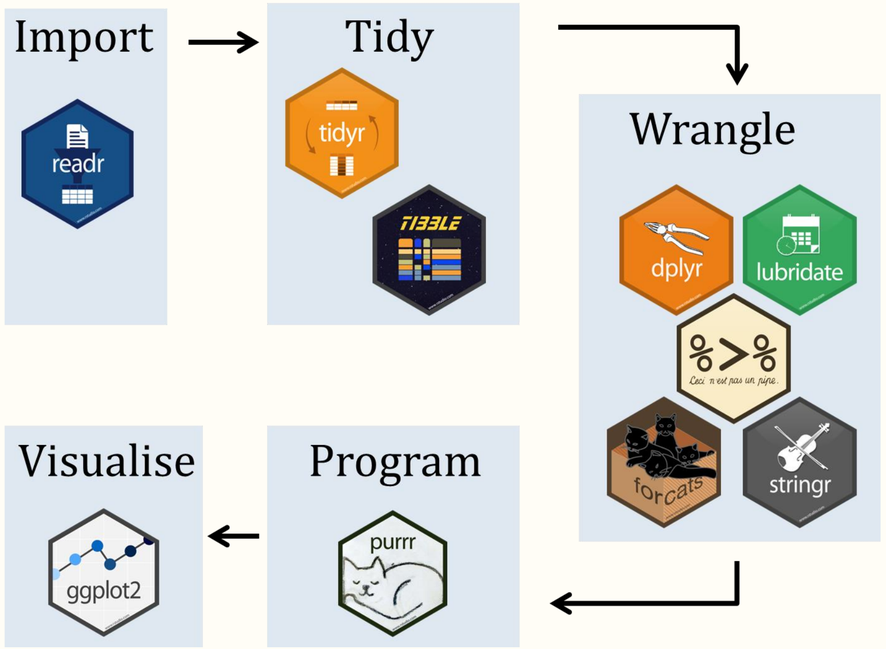

R Console
~
Help Search
R version 4.4.1 (2024-06-14) -- "Race for Your Life"
Copyright (C) 2024 The R Foundation for Statistical Computing
Platform: aarch64-apple-darwin20
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
Natural language support but running in an English locale
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
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.
[R.app GUI 1.80 (8416) aarch64-apple-darwin20]
[History restored from /Users/hem122/.Rapp.history]
> x = 1:10
> y = 1:10
> plot(x,y)
>


Quartz 2 [*]
y
x
2
4
6
8
10

# Packages

In R, **packages** are collections of functions, data, and documentation that extend the base functionality of the language.

- Allow users to perform specific tasks or analyze particular types of data
- Essentially bundles of reusable code that can be easily shared and installed
- Must first install a package, and then load it when you want to use it
- All have documentation available online



# The R Community

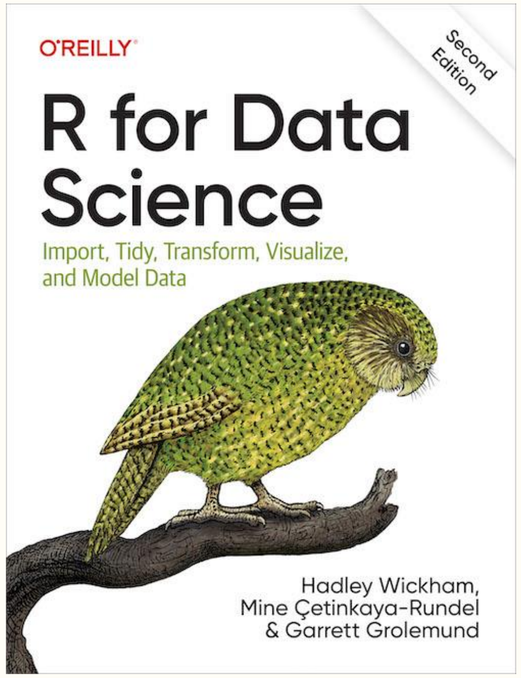
One of the best things about R is the **R community**. The R Community is a global network of individuals who use, contribute to, and promote the R programming language.

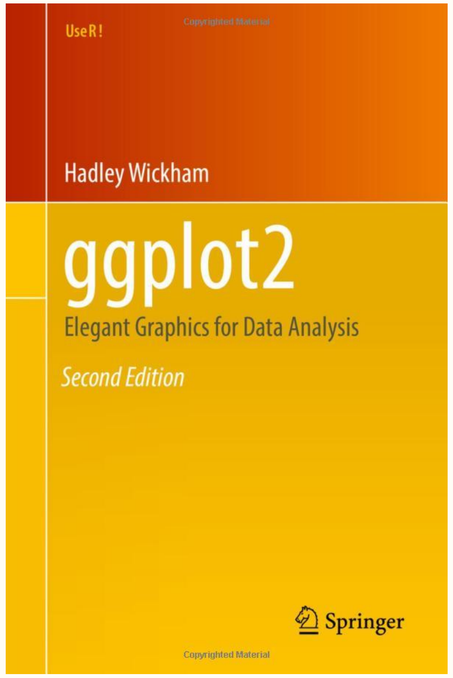
- Active engagement through online forums, blogs, social media, and conferences
- Members range from beginners to experts in various fields
- The community fosters collaboration, knowledge sharing, and support for users of all levels

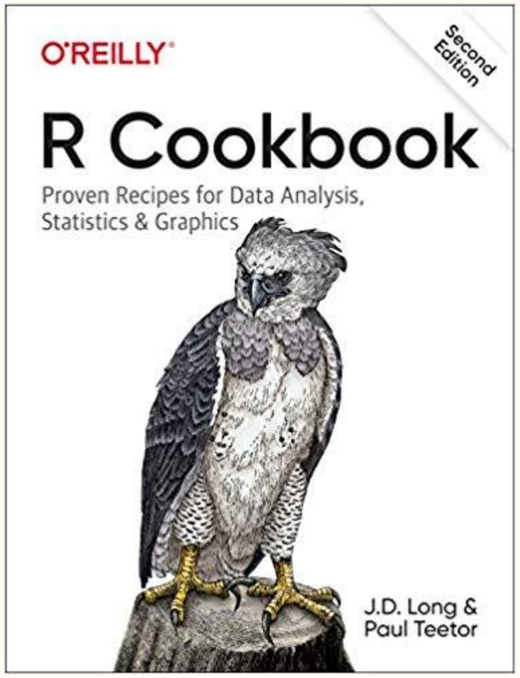
# R Resources

For help **writing** code:

- R for Data Science
- ggplot2
- R Cookbook



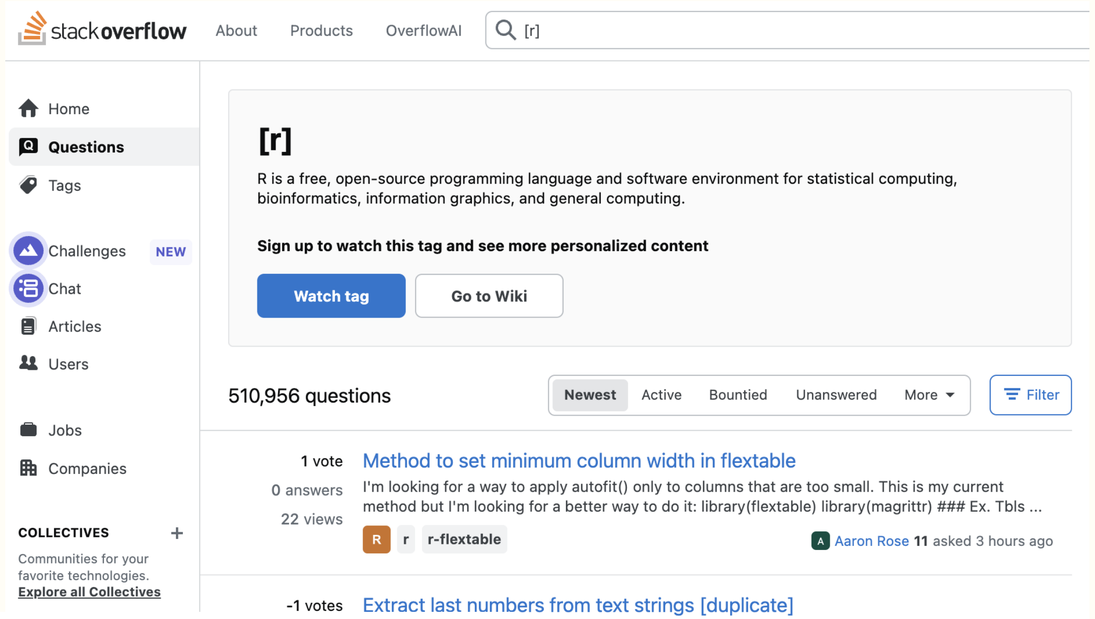



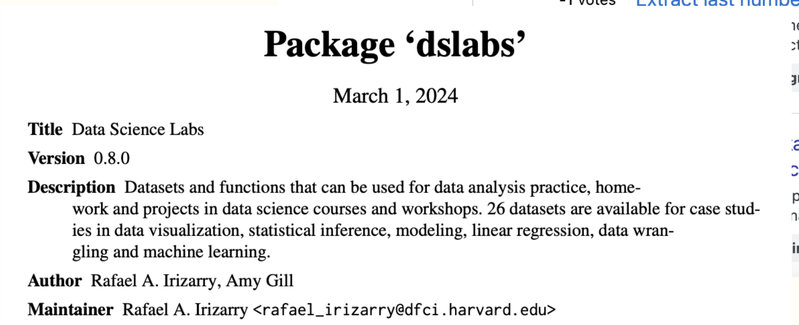



# R Resources

For help **debugging** code:

- R Help: help() and ?
- Vignettes and code demonstrations
- Package documentation
- Stack Overflow



**Package 'dslabs'**

March 1, 2024

**Title** Data Science Labs

**Version** 0.8.0

**Description** Datasets and functions that can be used for data analysis practice, homework and projects in data science courses and workshops. 26 datasets are available for case studies in data visualization, statistical inference, modeling, linear regression, data wrangling and machine learning.

**Author** Rafael A. Irizarry, Amy Gill

**Maintainer** Rafael A. Irizarry <rafael_irizarry@dfci.harvard.edu>

# Summary

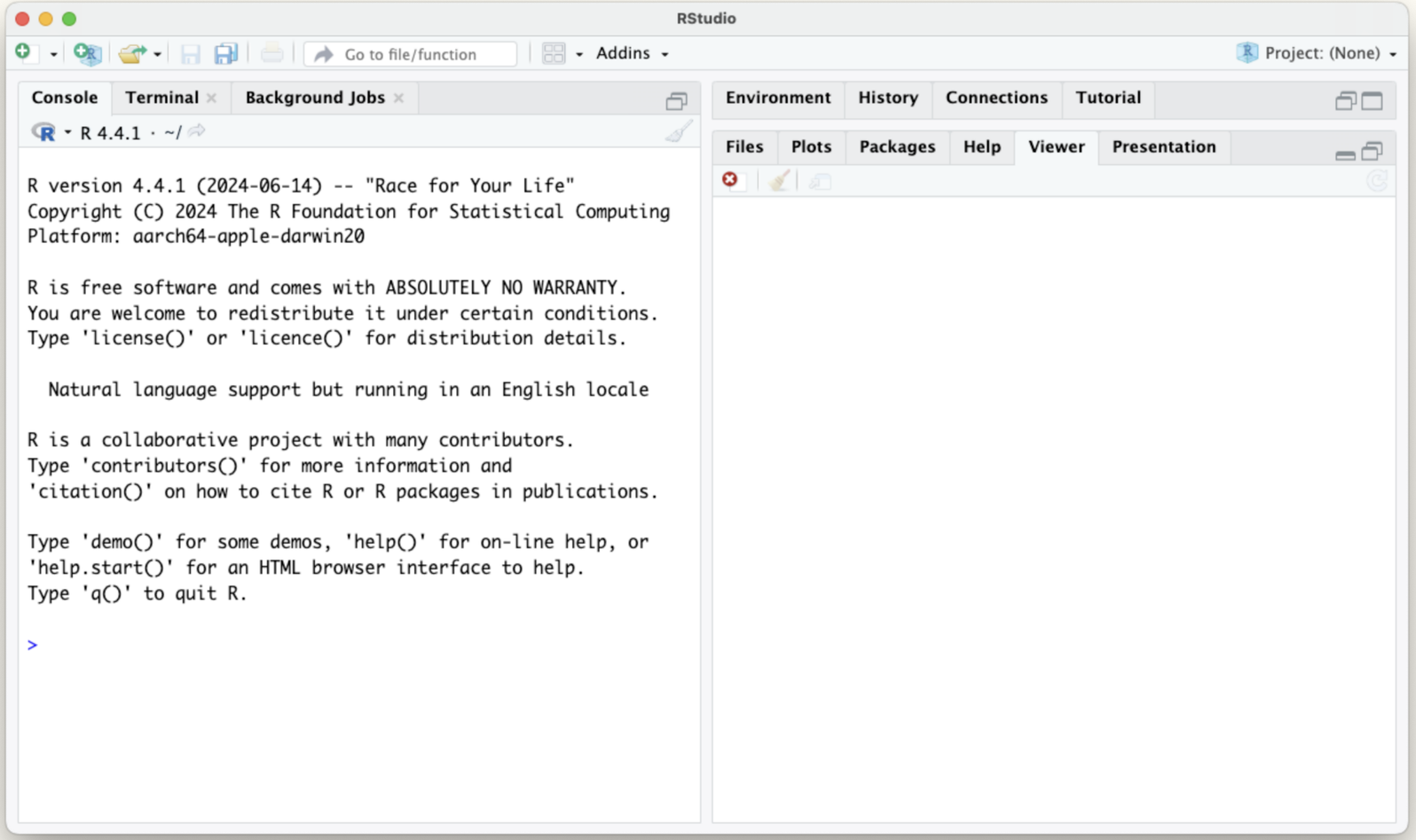
- R is a programming language designed for statistical computing and visualizations
- R is open source and has a large, active, and diverse community that contributes to and promotes R
- R packages are collections of functions, data, and documentation that extend the base functionality of R
- There are many resources for writing and debugging R code

# Introduction to RStudio

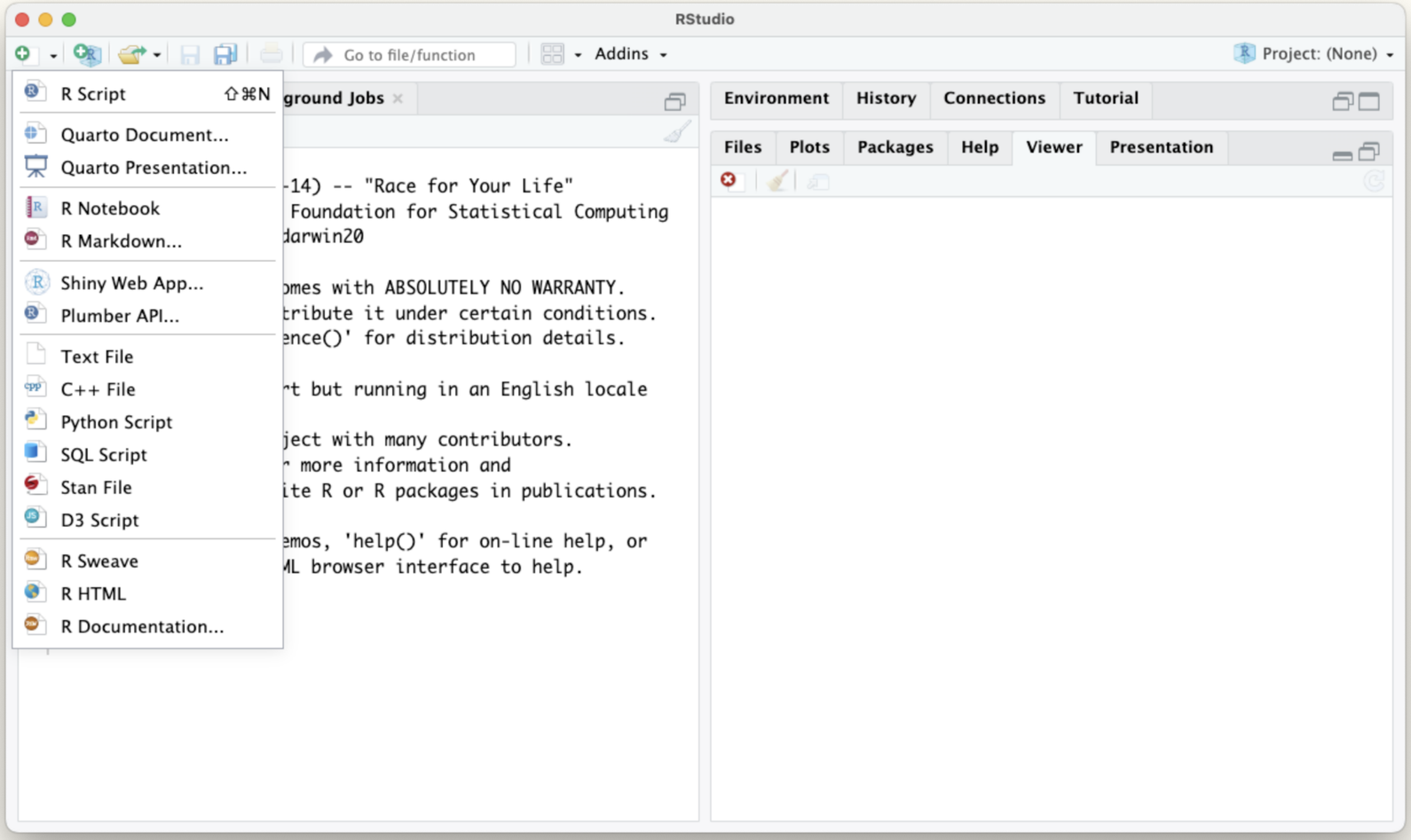
# What is RStudio?

- **RStudio** is an integrated development environment (IDE) specifically created for R
  - RStudio desktop
  - User-friendly interface

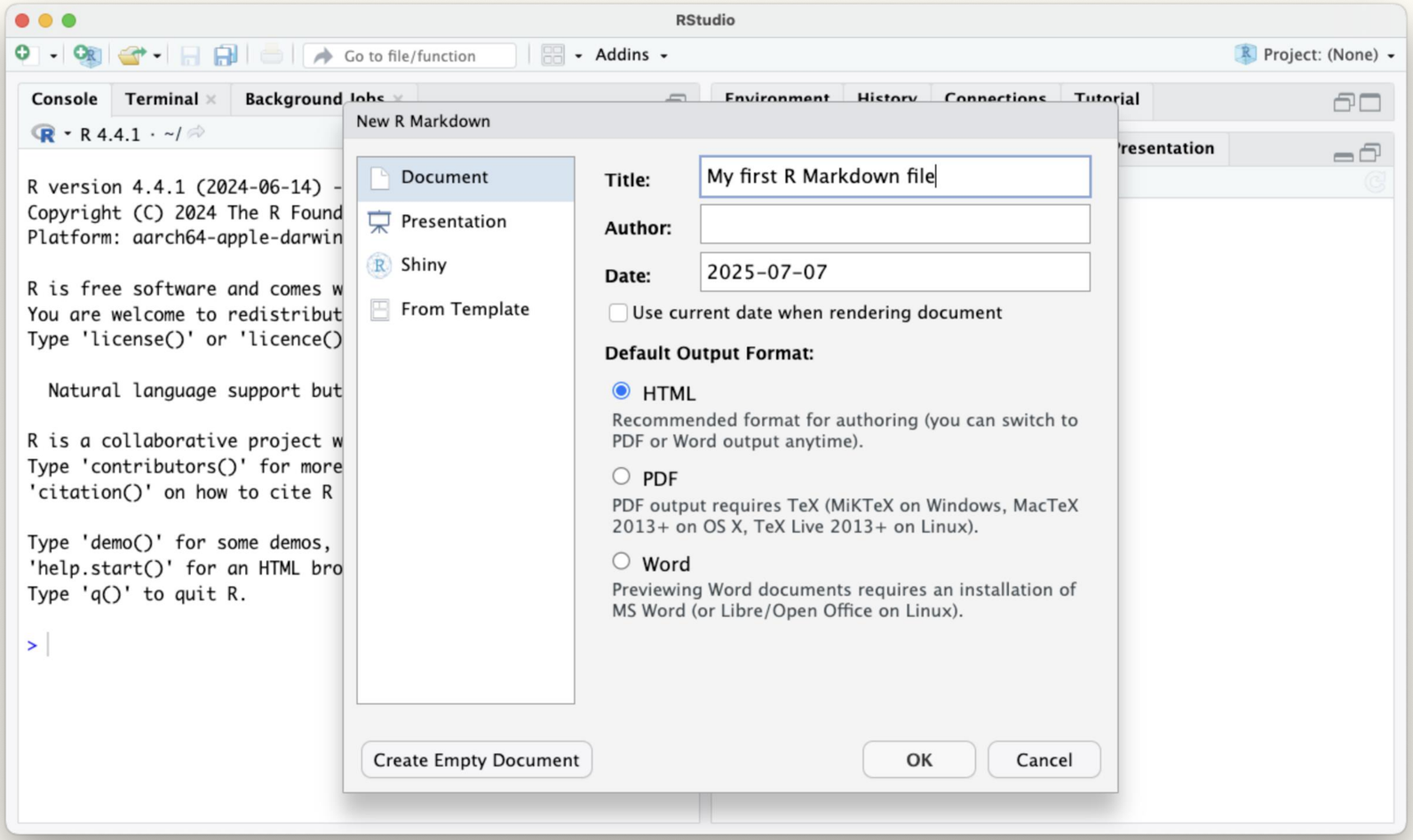
# Layout


RStudio
Go to file/function
Addins
Project: (None)
Console
Terminal
Background Jobs
R 4.4.1 · ~/
R version 4.4.1 (2024-06-14) -- "Race for Your Life"
Copyright (C) 2024 The R Foundation for Statistical Computing
Platform: aarch64-apple-darwin20
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
Natural language support but running in an English locale
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
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.
>
Environment
History
Connections
Tutorial
Files
Plots
Packages
Help
Viewer
Presentation

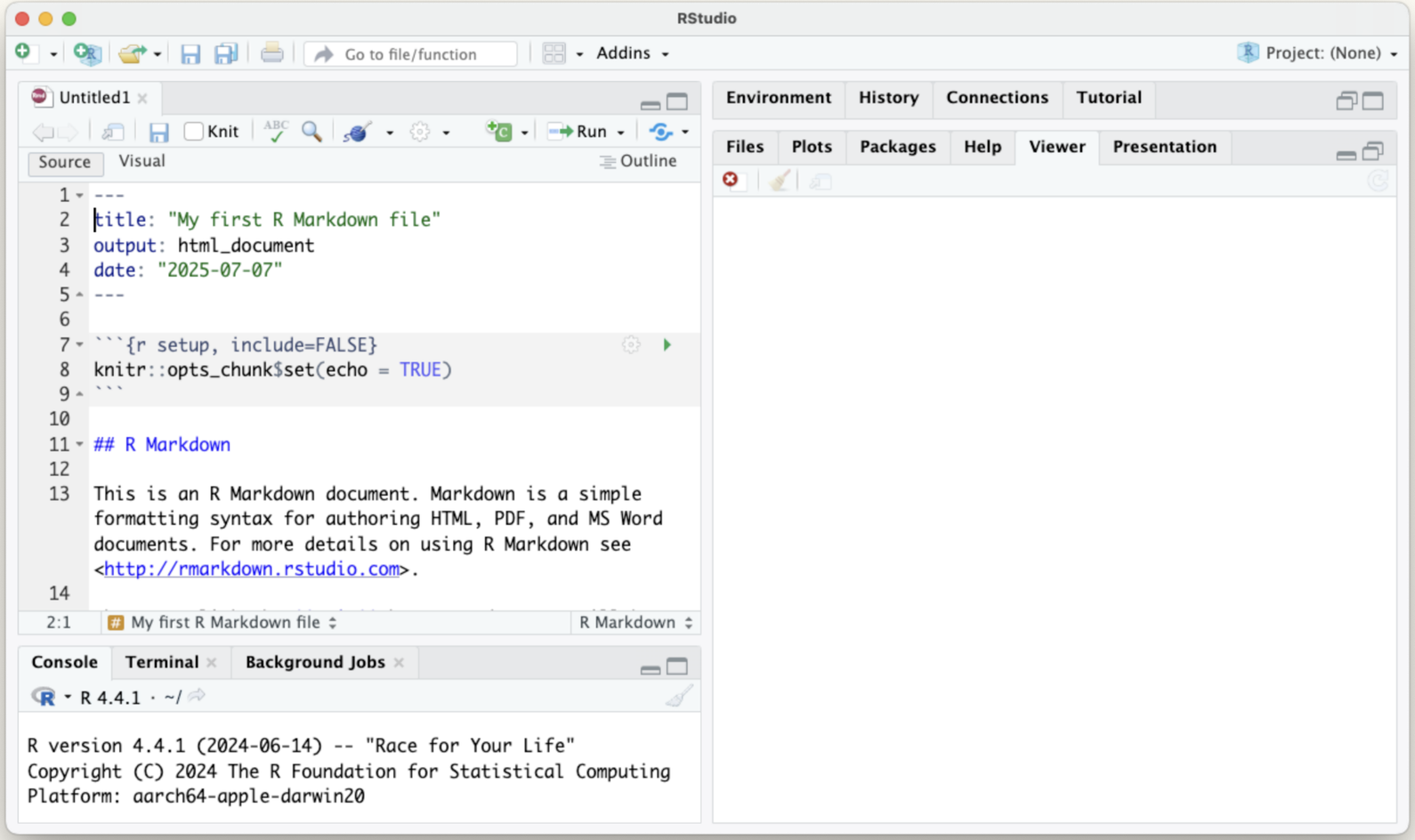
# Layout


RStudio
Go to file/function
Addins
Project: (None)
R Script ⇧⌘N
Quarto Document...
Quarto Presentation...
R Notebook
R Markdown...
Shiny Web App...
Plumber API...
Text File
C++ File
Python Script
SQL Script
Stan File
D3 Script
R Sweave
R HTML
R Documentation...
ground Jobs
-14) -- "Race for Your Life"
Foundation for Statistical Computing
darwin20
omes with ABSOLUTELY NO WARRANTY.
tribute it under certain conditions.
ence()' for distribution details.
rt but running in an English locale
ject with many contributors.
r more information and
ite R or R packages in publications.
emos, 'help()' for on-line help, or
ML browser interface to help.
Environment
History
Connections
Tutorial
Files
Plots
Packages
Help
Viewer
Presentation

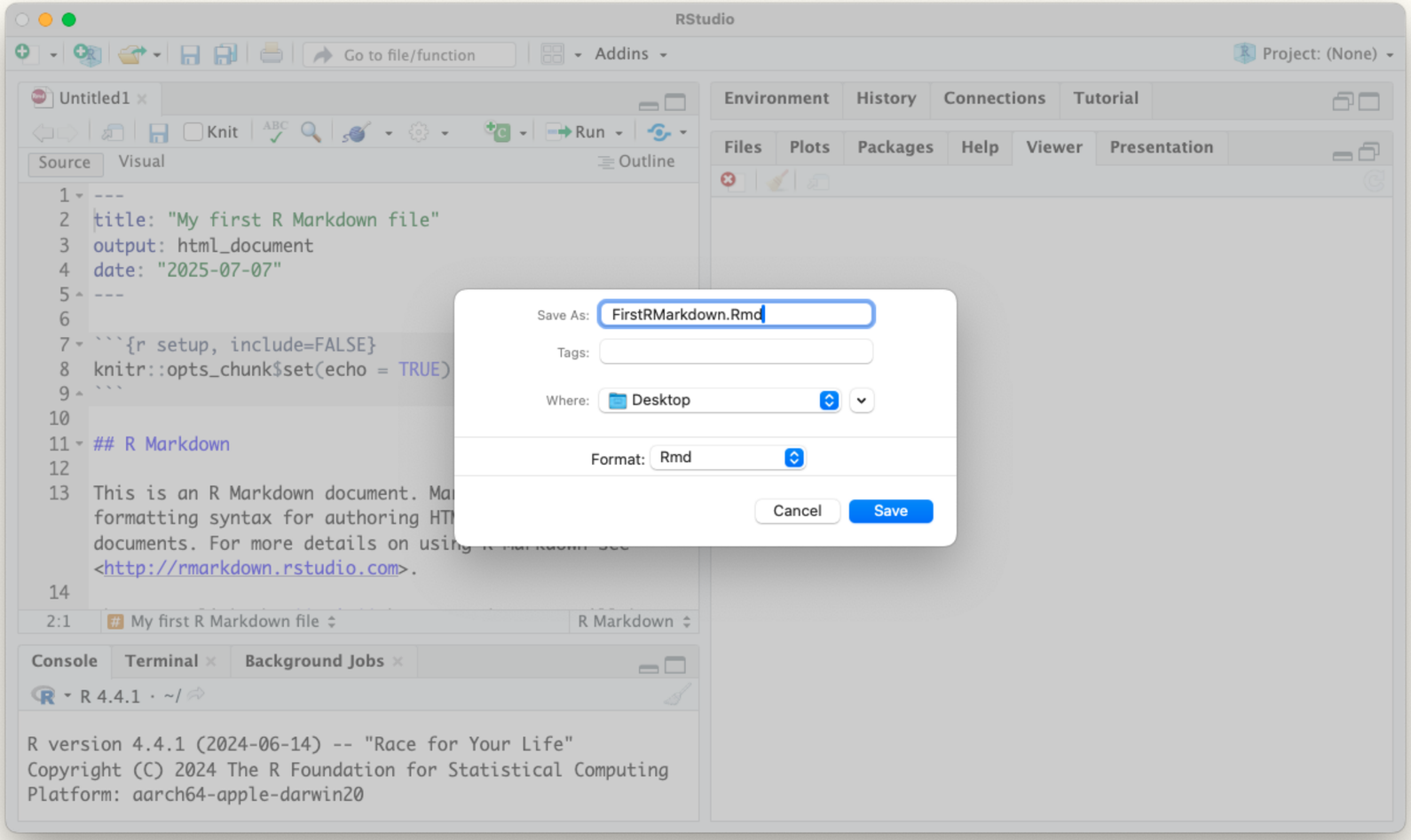
# Layout


RStudio
Go to file/function
Addins
Project: (None)
Console
Terminal
Background Jobs
Environment
History
Connections
Tutorial
R 4.4.1 · ~/
R version 4.4.1 (2024-06-14) -
Copyright (C) 2024 The R Found
Platform: aarch64-apple-darwin
R is free software and comes w
You are welcome to redistribut
Type 'license()' or 'licence()
Natural language support but
R is a collaborative project w
Type 'contributors()' for more
'citation()' on how to cite R
Type 'demo()' for some demos,
'help.start()' for an HTML bro
Type 'q()' to quit R.
>
New R Markdown
Document
Presentation
Shiny
From Template
Title:
My first R Markdown file
Author:
Date:
2025-07-07
Use current date when rendering document
Default Output Format:
HTML
Recommended format for authoring (you can switch to PDF or Word output anytime).
PDF
PDF output requires TeX (MiKTeX on Windows, MacTeX 2013+ on OS X, TeX Live 2013+ on Linux).
Word
Previewing Word documents requires an installation of MS Word (or Libre/Open Office on Linux).
Create Empty Document
OK
Cancel

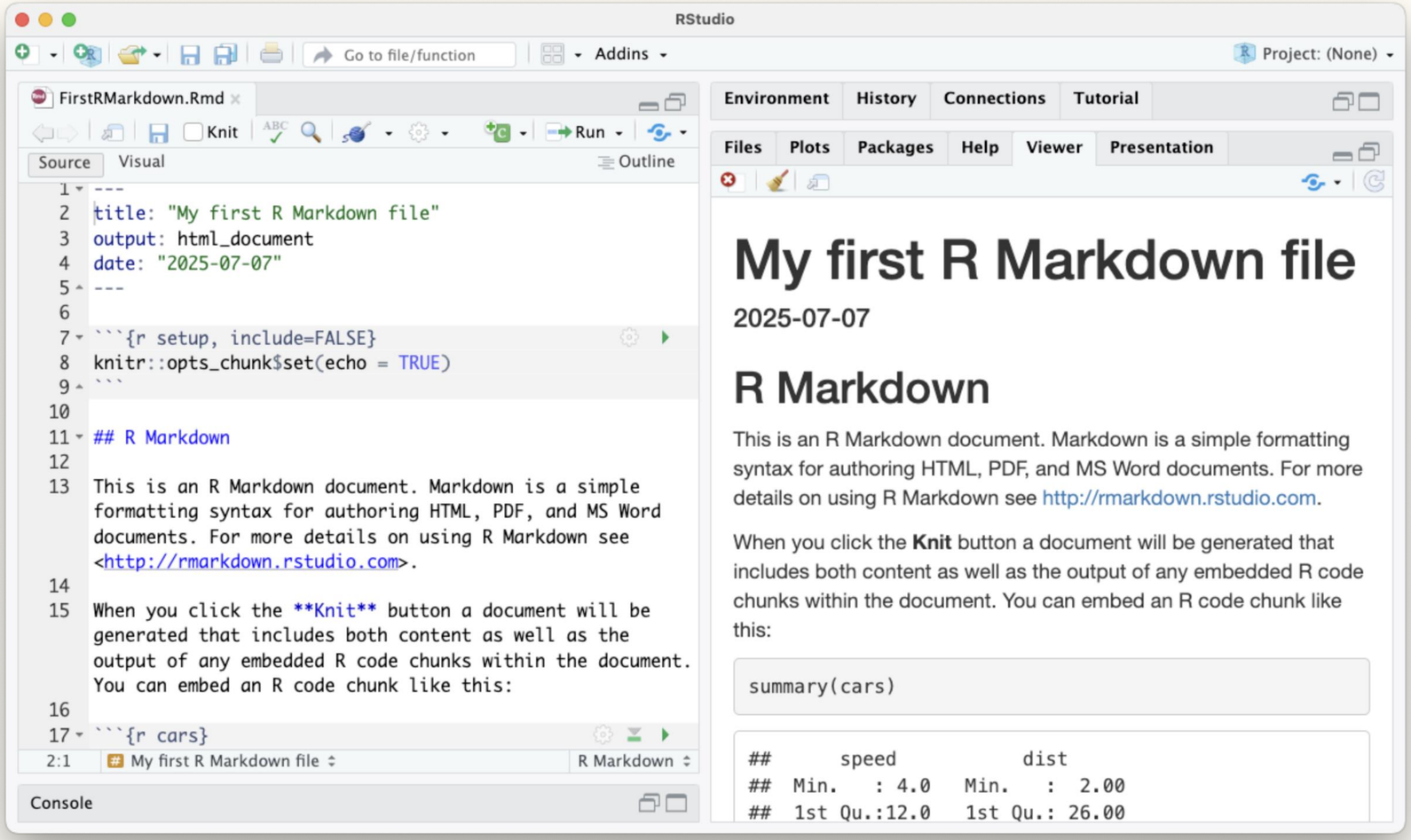
# Layout


RStudio
Go to file/function
Addins
Project: (None)
Untitled1
Knit
Run
Source
Visual
Outline
1 ---
2 title: "My first R Markdown file"
3 output: html_document
4 date: "2025-07-07"
5 ---
6
7 ```{r setup, include=FALSE}
8 knitr::opts_chunk$set(echo = TRUE)
9 ```
10
11 ## R Markdown
12
13 This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.
14
2:1
My first R Markdown file
R Markdown
Console
Terminal
Background Jobs
R 4.4.1 · ~/
R version 4.4.1 (2024-06-14) -- "Race for Your Life"
Copyright (C) 2024 The R Foundation for Statistical Computing
Platform: aarch64-apple-darwin20
Environment
History
Connections
Tutorial
Files
Plots
Packages
Help
Viewer
Presentation

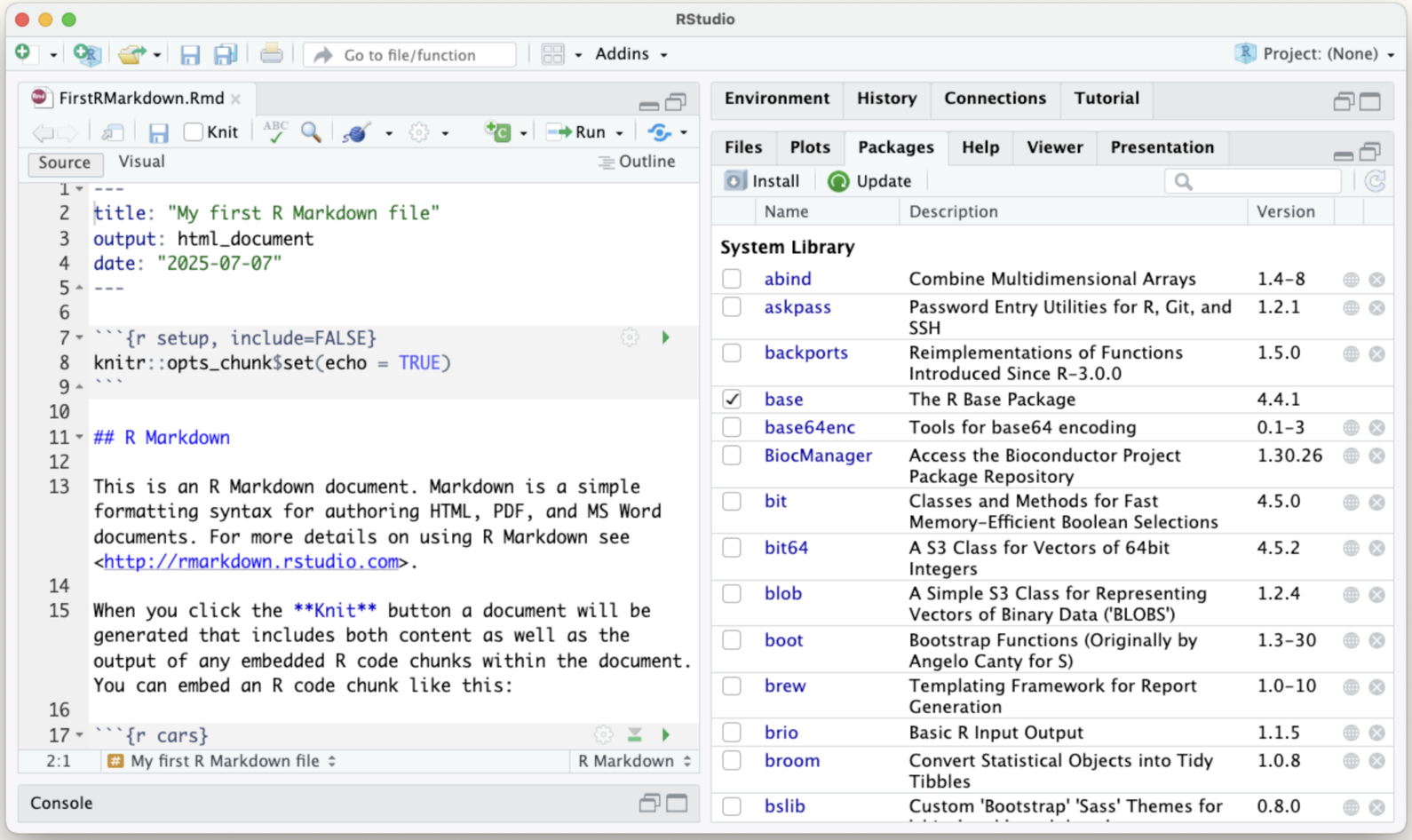
# Layout


RStudio
Go to file/function
Addins
Project: (None)
Untitled1
Knit
Run
Source
Visual
Outline
Environment
History
Connections
Tutorial
Files
Plots
Packages
Help
Viewer
Presentation
1 ---
2 title: "My first R Markdown file"
3 output: html_document
4 date: "2025-07-07"
5 ---
6
7 ```{r setup, include=FALSE}
8 knitr::opts_chunk$set(echo = TRUE)
9 ```
10
11 ## R Markdown
12
13 This is an R Markdown document. Ma
formatting syntax for authoring HT
documents. For more details on usi
<http://rmarkdown.rstudio.com>.
14
Save As: FirstRMarkdown.Rmd
Tags:
Where: Desktop
Format: Rmd
Cancel
Save
2:1
My first R Markdown file
R Markdown
Console
Terminal
Background Jobs
R 4.4.1 · ~/
R version 4.4.1 (2024-06-14) -- "Race for Your Life"
Copyright (C) 2024 The R Foundation for Statistical Computing
Platform: aarch64-apple-darwin20

# Layout

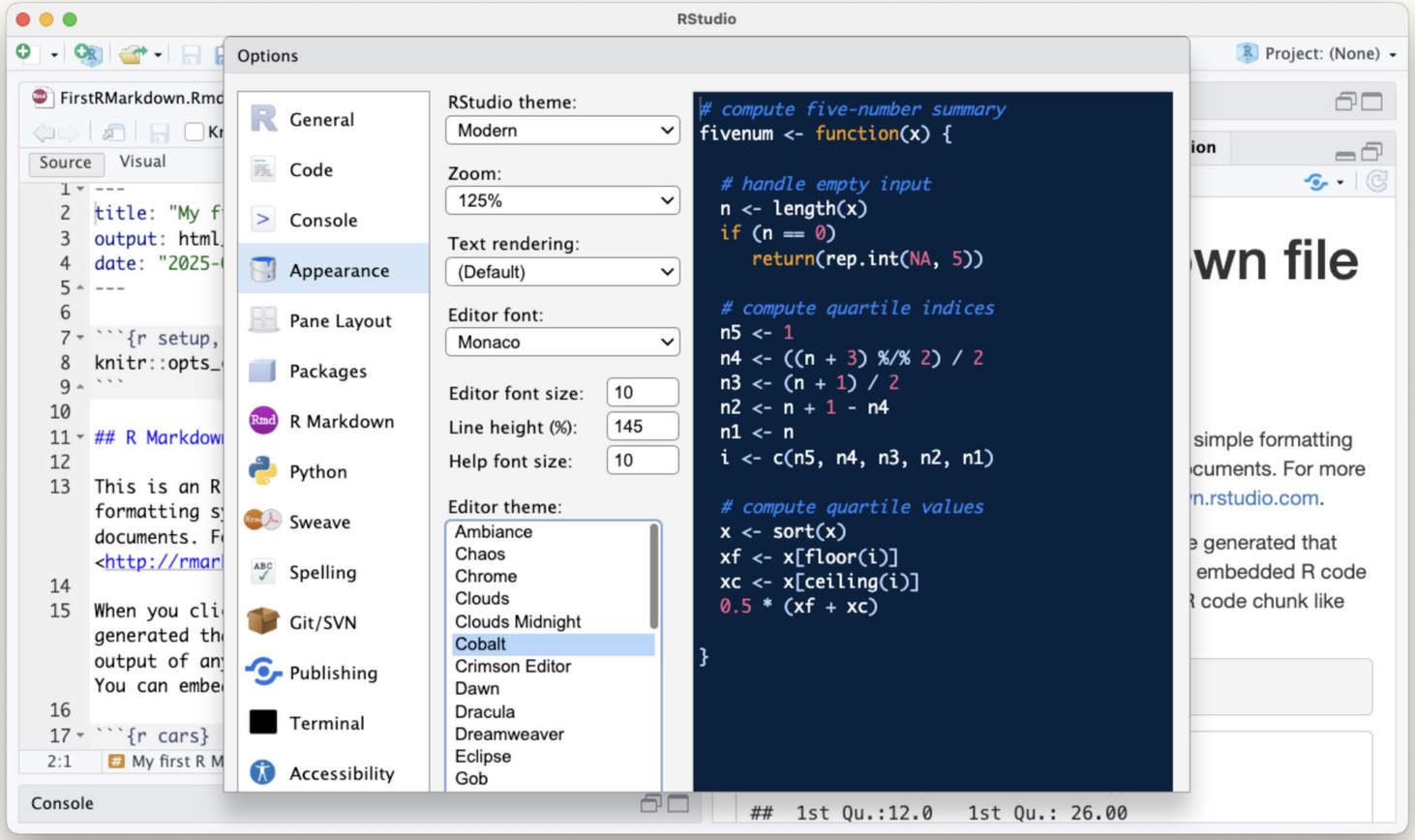

RStudio
Go to file/function
Addins
Project: (None)
FirstRMarkdown.Rmd
Knit
Run
Source
Visual
Outline
1 ---
2 title: "My first R Markdown file"
3 output: html_document
4 date: "2025-07-07"
5 ---
6
7 ```{r setup, include=FALSE}
8 knitr::opts_chunk$set(echo = TRUE)
9 ```
10
11 ## R Markdown
12
13 This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.
14
15 When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:
16
17 ```{r cars}
2:1
My first R Markdown file
R Markdown
Console
Environment
History
Connections
Tutorial
Files
Plots
Packages
Help
Viewer
Presentation
My first R Markdown file
2025-07-07
R Markdown
This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.
When you click the Knit button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:
summary(cars)
## speed dist
## Min. : 4.0 Min. : 2.00
## 1st Qu.:12.0 1st Qu.: 26.00

# Layout


RStudio
Go to file/function
Addins
Project: (None)
FirstRMarkdown.Rmd
Knit
Run
Source
Visual
Outline
1 ---
2 title: "My first R Markdown file"
3 output: html_document
4 date: "2025-07-07"
5 ---
6
7 ```{r setup, include=FALSE}
8 knitr::opts_chunk$set(echo = TRUE)
9 ```
10
11 ## R Markdown
12
13 This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.
14
15 When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:
16
17 ```{r cars}
2:1
My first R Markdown file
R Markdown
Console
Environment
History
Connections
Tutorial
Files
Plots
Packages
Help
Viewer
Presentation
Install
Update
Name
Description
Version
System Library
abind
Combine Multidimensional Arrays
1.4-8
askpass
Password Entry Utilities for R, Git, and SSH
1.2.1
backports
Reimplementations of Functions Introduced Since R-3.0.0
1.5.0
base
The R Base Package
4.4.1
base64enc
Tools for base64 encoding
0.1-3
BiocManager
Access the Bioconductor Project Package Repository
1.30.26
bit
Classes and Methods for Fast Memory-Efficient Boolean Selections
4.5.0
bit64
A S3 Class for Vectors of 64bit Integers
4.5.2
blob
A Simple S3 Class for Representing Vectors of Binary Data ('BLOBS')
1.2.4
boot
Bootstrap Functions (Originally by Angelo Canty for S)
1.3-30
brew
Templating Framework for Report Generation
1.0-10
brio
Basic R Input Output
1.1.5
broom
Convert Statistical Objects into Tidy Tibbles
1.0.8
bslib
Custom 'Bootstrap' 'Sass' Themes for
0.8.0

# Appearance

Tools → Global Options → Appearance



# Summary

- RStudio is an integrated development environment specifically created for R
- RStudio provides a user-friendly interface
- RStudio windowpanes showcase code, an opened file, output, etc.
- The appearance of RStudio can be changed to your preferred style