

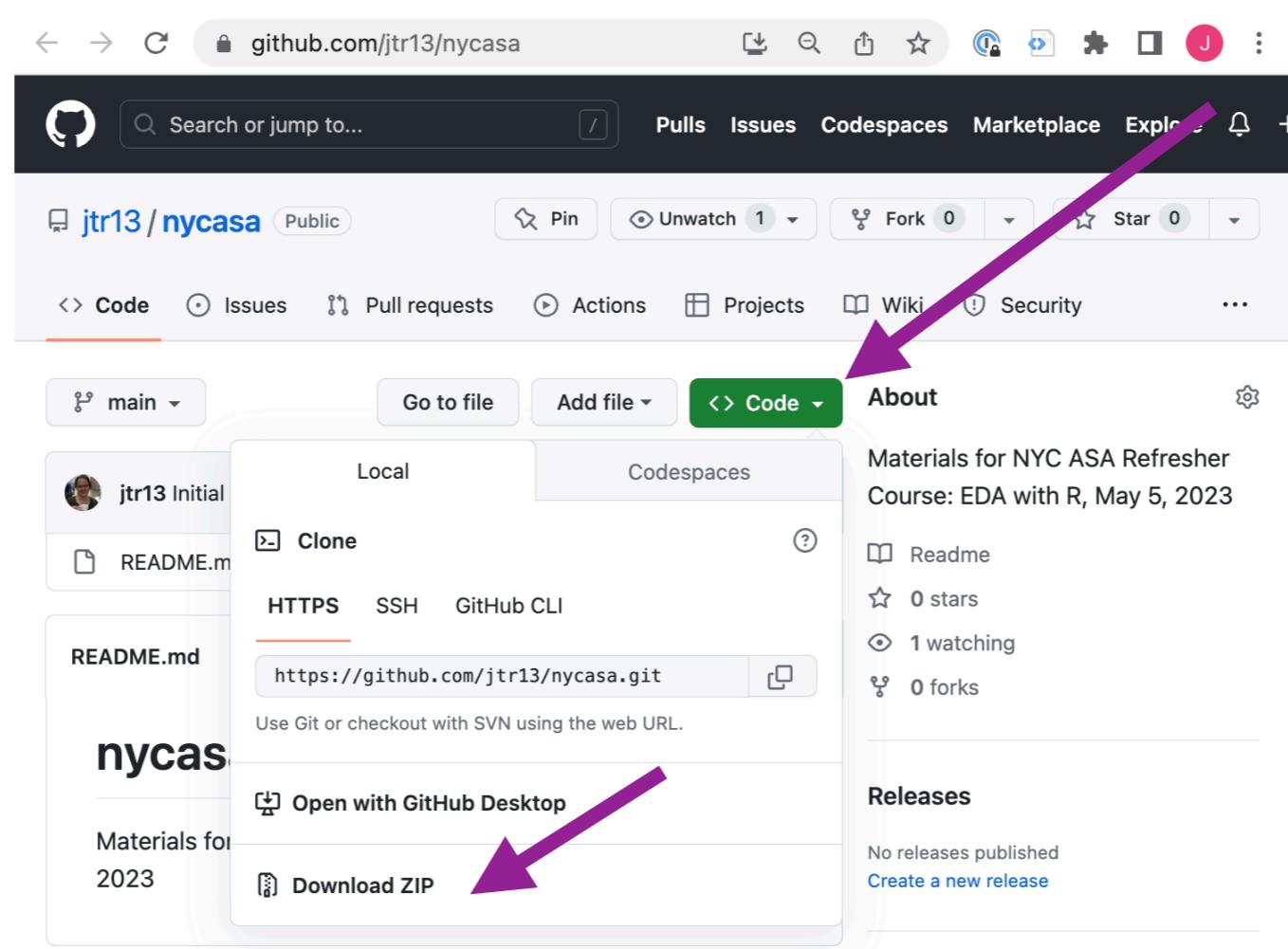
Exploratory Data Analysis with R

**ASA NYC Chapter
Friday, May 5, 2023, 3pm-5pm**

Joyce Robbins
Lecturer
Department of Statistics
Columbia University
jtr13@columbia.edu

Slides and code files

[www.github.com/jtr13/nycasa](https://github.com/jtr13/nycasa)



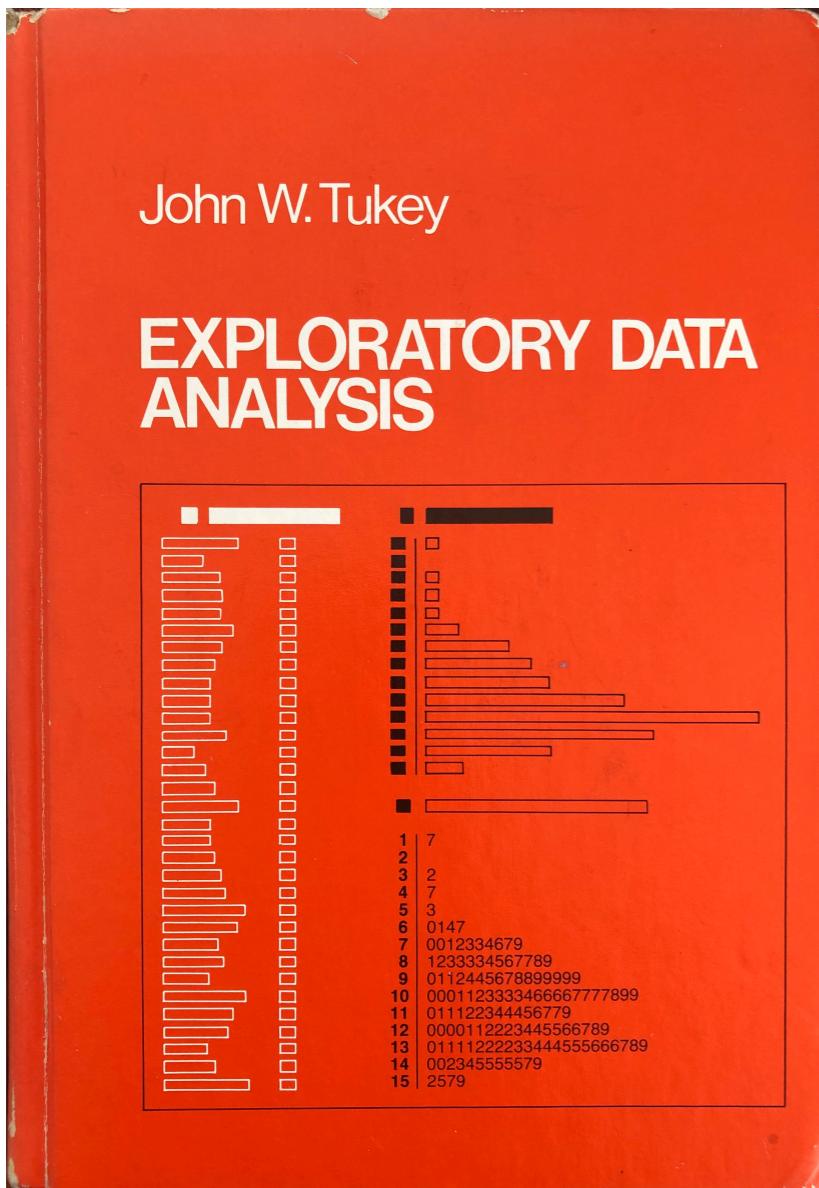
These slides: EDAwR-Slides1.pdf

Agenda

- What is EDA?
- Update / Setup R
- Finding data
- Exploring data

Exploratory Data Analysis

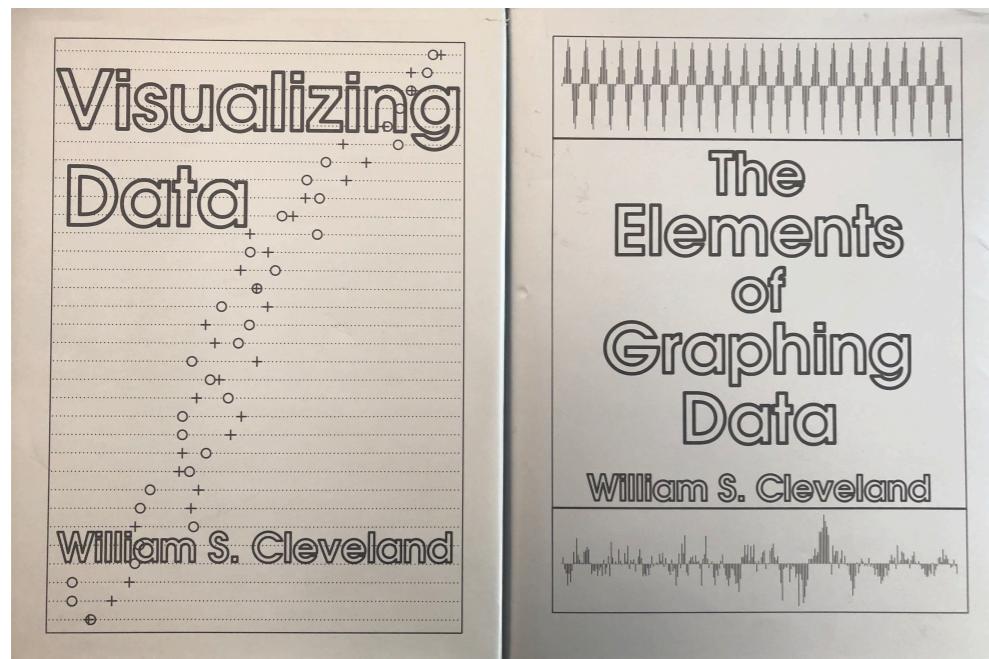
Exploratory Data Analysis



"Exploratory data analysis is detective work."

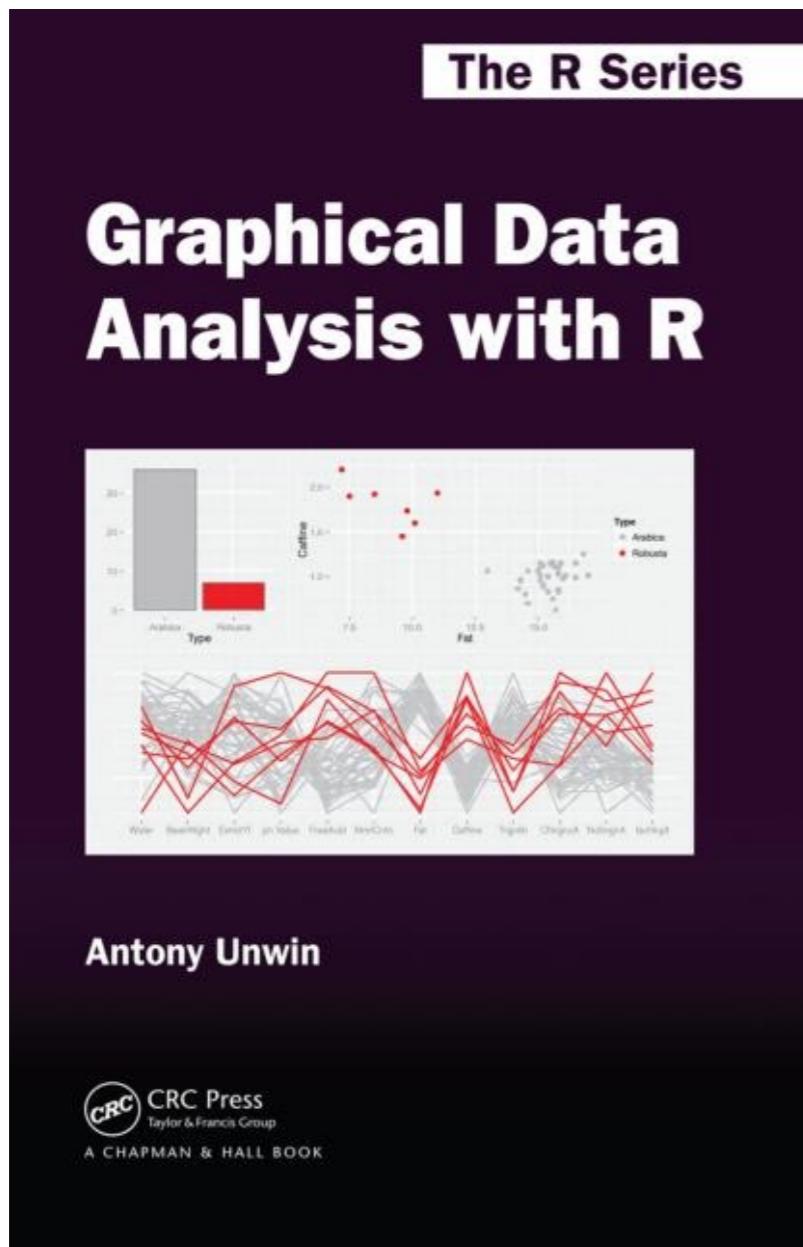
"Exploratory data analysis can never be the whole story, but nothing else can serve as the foundation stone--as the first step."

William Cleveland 1980s, 1990s



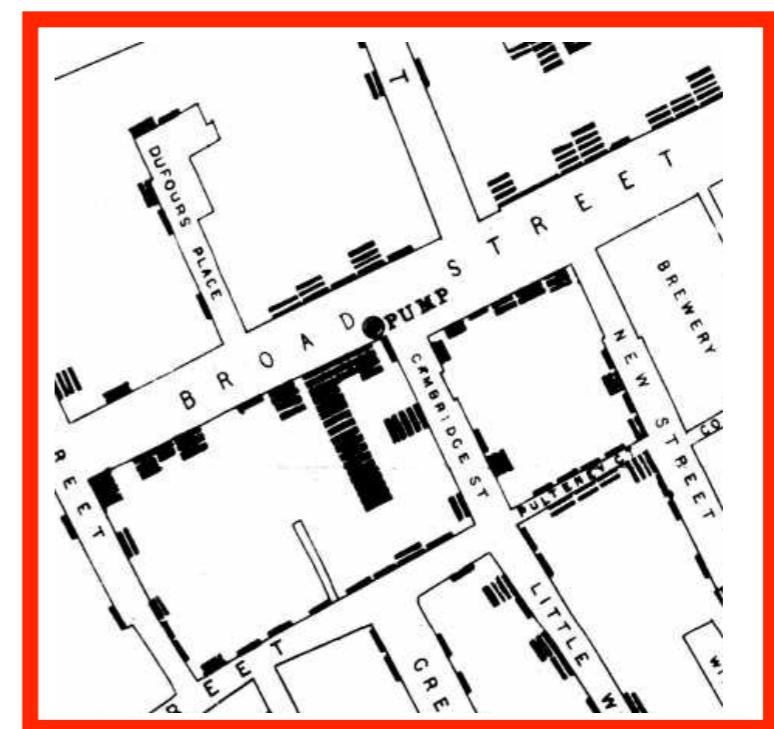
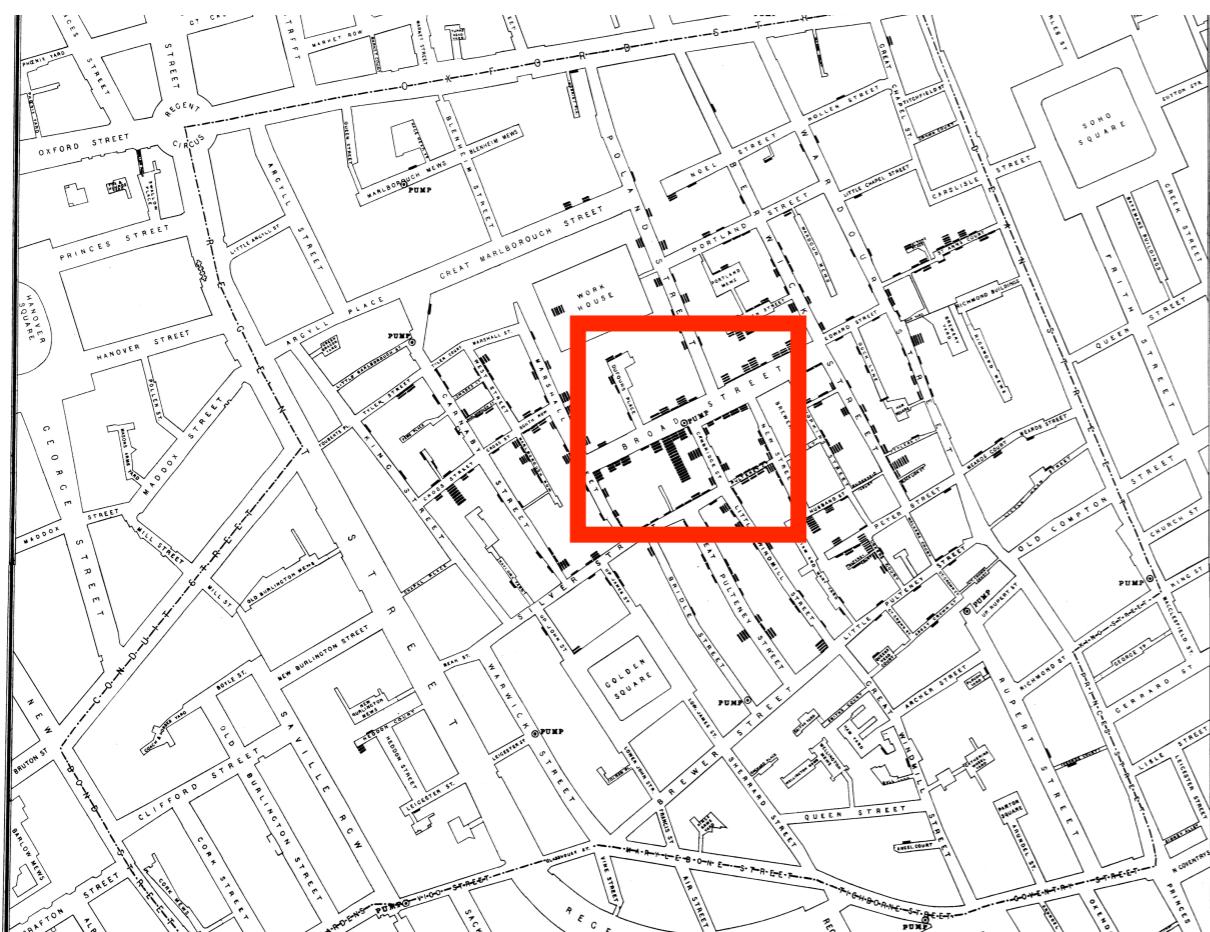
"Visualization is critical to data analysis. It provides a front line of attack, revealing intricate structure in data that cannot be absorbed in any other way. We discover unimagined effects, and we challenge imagined ones."

Graphical Data Analysis with R



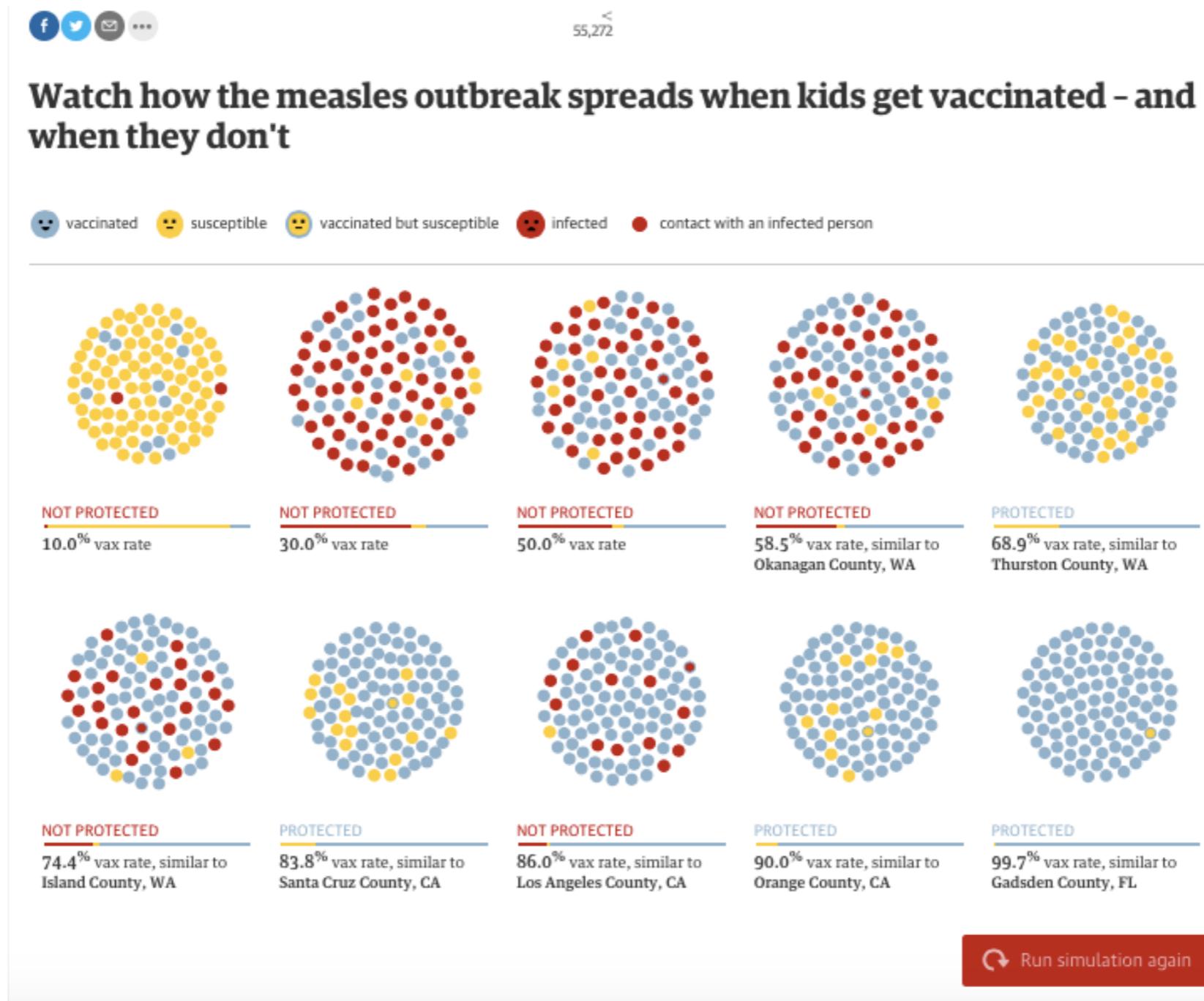
- Strong EDA focus
- Code examples
- Textbook for my EDA course
- Multivariate approaches

John Snow, Cholera Map 1854



Broad St. pump

Data Visualization



EDA vs. Data Visualization

- Data Visualization: how can I help the reader understand how vaccination affects the spread of disease?
- EDA: How does vaccination affect the spread of disease?

Setup

Update R (or install)

> R.version

```
R.version
#>
#> platform      aarch64-apple-darwin20
#> arch          aarch64
#> os            darwin20
#> system        aarch64, darwin20
#> status
#> major          4
#> minor          3.0
#> year           2023
#> month          04
#> day            21
#> svn rev        84292
#> language        R
#> version.string R version 4.3.0 (2023-04-21)
#> nickname       Already Tomorrow
```

Update R (or install)

cran.r-project.org

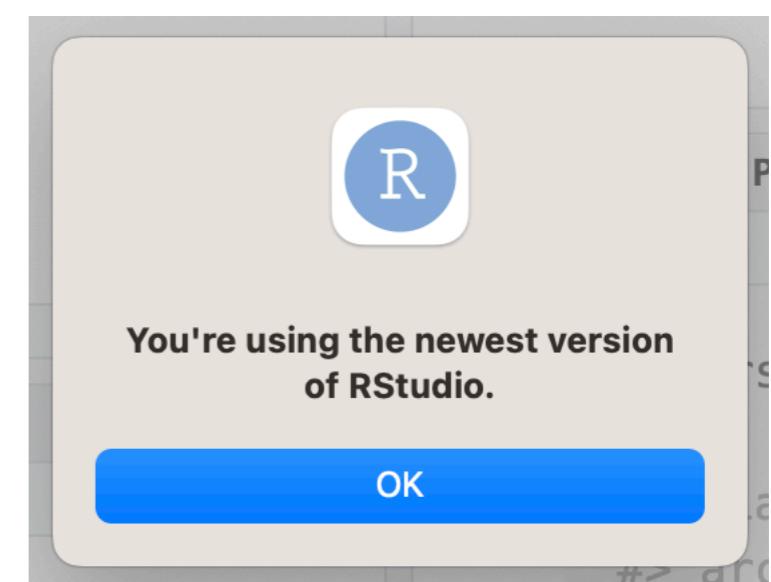
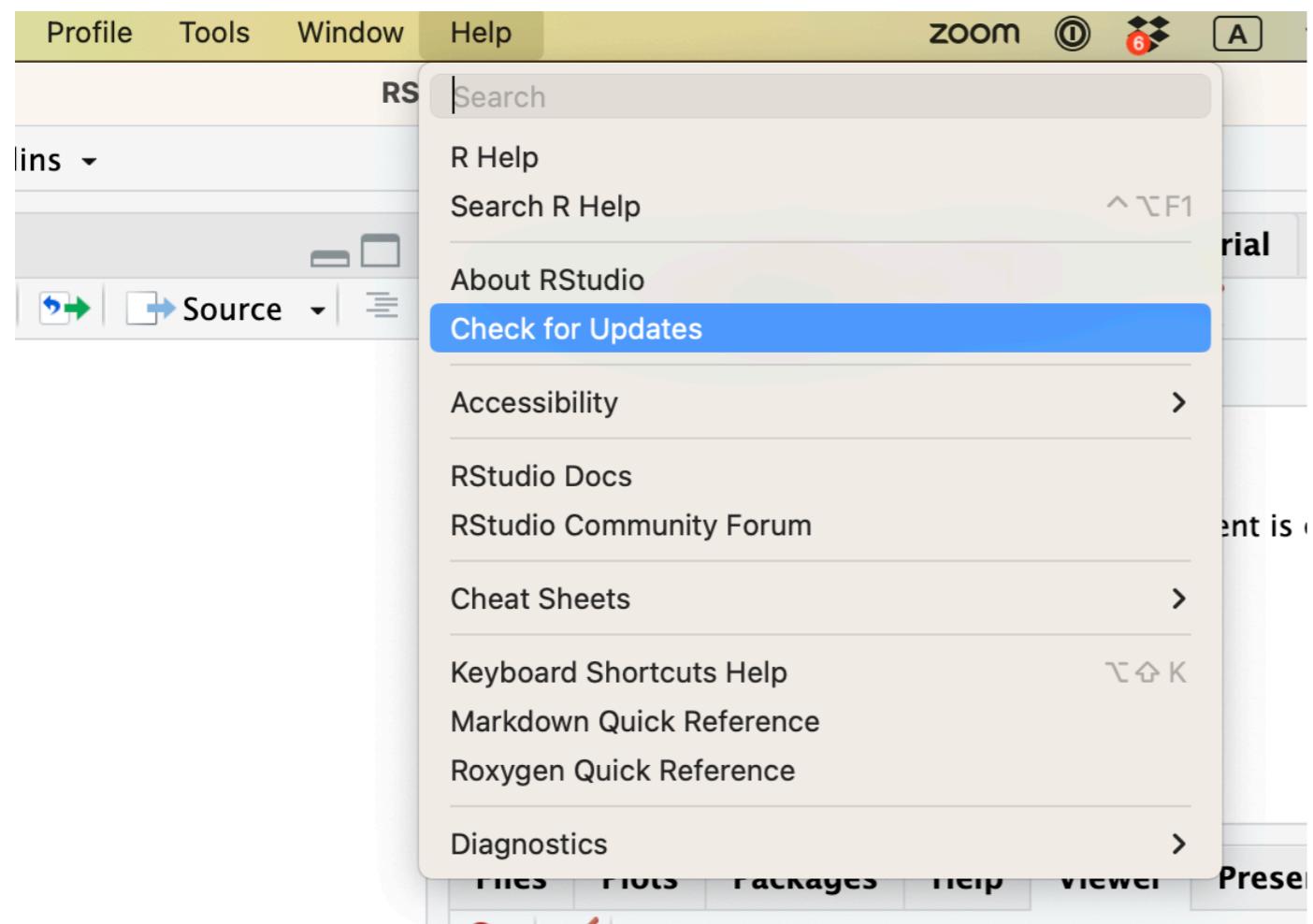
The screenshot shows a web browser displaying the CRAN homepage. The title bar reads "The Comprehensive R Archive Network". The main content area is divided into several sections:

- Download and Install R**: Precompiled binary distributions for Windows and Mac users. It includes links for Linux (Debian, Fedora/Redhat, Ubuntu), macOS, and Windows.
- Source Code for all Platforms**: Information for Windows and Mac users about downloading source code, which needs to be compiled. It lists releases, alpha/beta versions, daily snapshots, and older versions.
- Questions About R**: A section for users with questions about R, pointing them to frequently asked questions.

On the left sidebar, there are links for CRAN, Mirrors, What's new?, Search, CRAN Team, About R, R Homepage, The R Journal, Software, R Sources, R Binaries, Packages, Task Views, Other, Documentation, Manuals, FAQs, and Contributed.

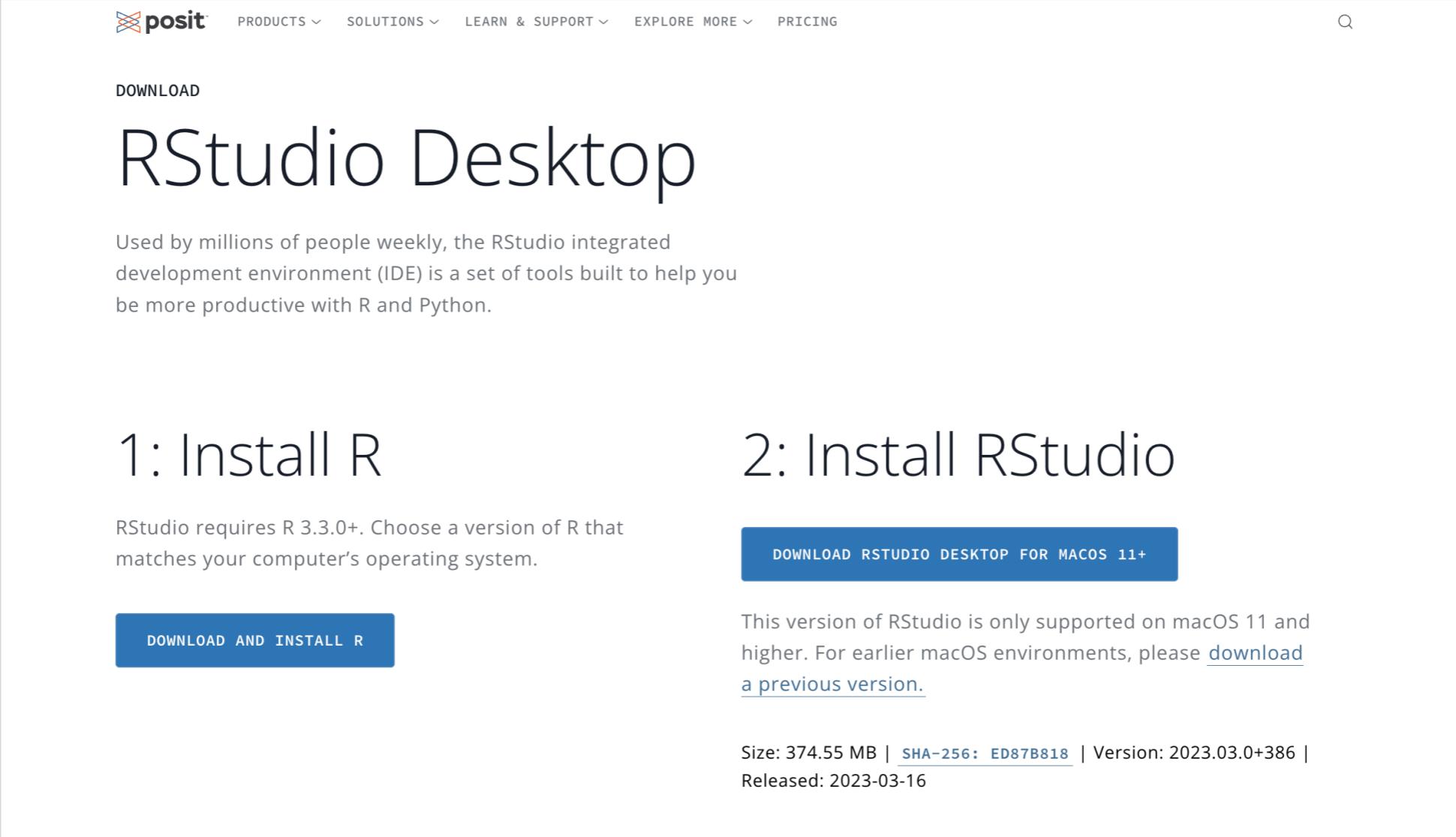
What are R and CRAN?

Update RStudio



Or install RStudio

<https://posit.co/download/rstudio-desktop/> (free)



The screenshot shows the RStudio Desktop download page on the posit.co website. At the top, there's a navigation bar with links for PRODUCTS, SOLUTIONS, LEARN & SUPPORT, EXPLORE MORE, and PRICING. A search icon is also present. Below the navigation, a "DOWNLOAD" button is visible. The main title is "RStudio Desktop". A descriptive paragraph explains that it's used by millions weekly and is an integrated development environment for R and Python. Two sections are shown: "1: Install R" and "2: Install RStudio". The "1: Install R" section has a "DOWNLOAD AND INSTALL R" button. The "2: Install RStudio" section has a "DOWNLOAD RSTUDIO DESKTOP FOR MACOS 11+" button. Below this, a note states that the version is only supported on macOS 11 and higher, with a link to download a previous version. At the bottom, there's information about the file size (374.55 MB), SHA-256 hash (ED87B818), version (2023.03.0+386), and release date (2023-03-16).

USED BY MILLIONS OF PEOPLE WEEKLY, THE RSTUDIO INTEGRATED DEVELOPMENT ENVIRONMENT (IDE) IS A SET OF TOOLS BUILT TO HELP YOU BE MORE PRODUCTIVE WITH R AND PYTHON.

1: Install R

RStudio requires R 3.3.0+. Choose a version of R that matches your computer's operating system.

DOWNLOAD AND INSTALL R

2: Install RStudio

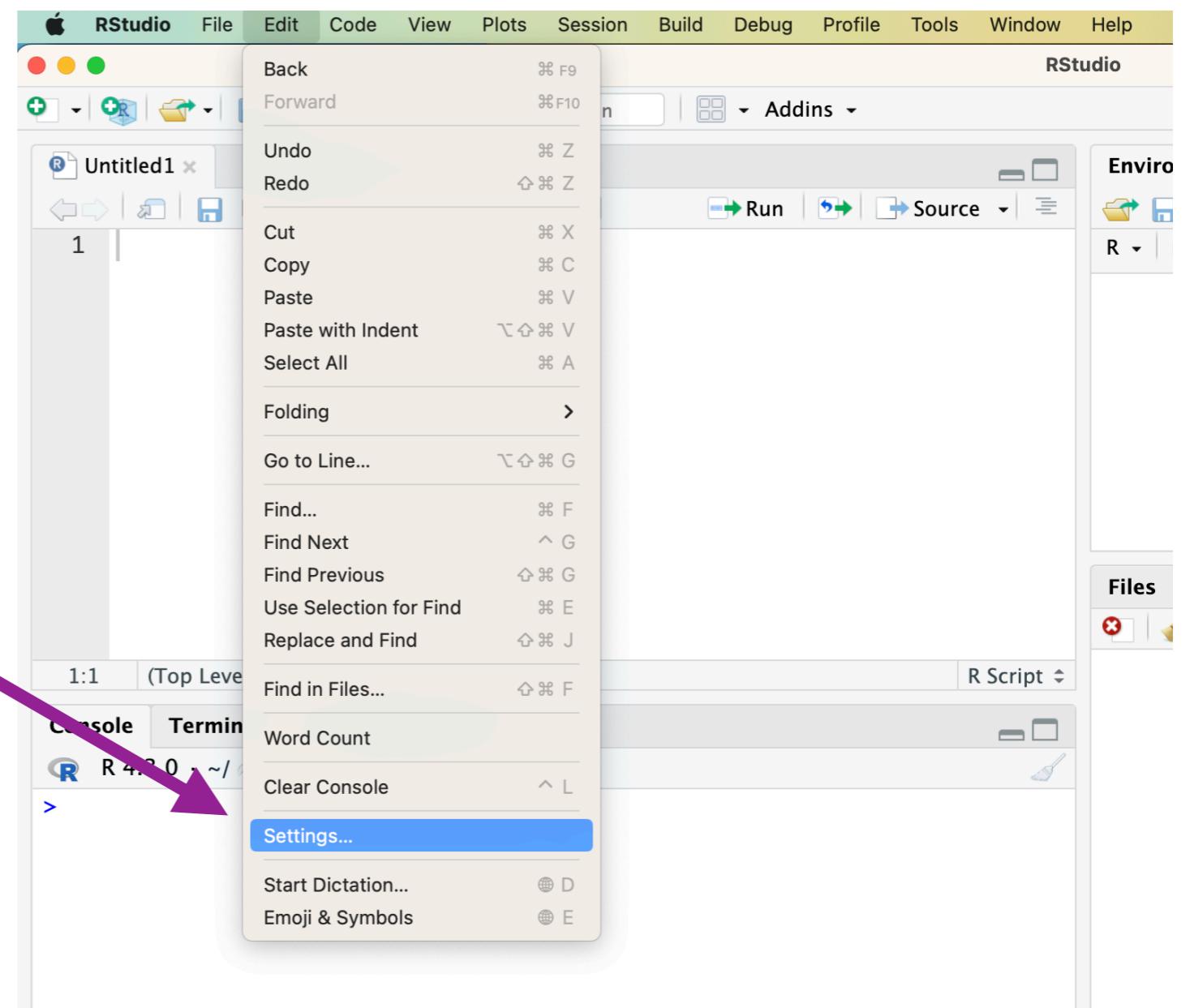
DOWNLOAD RSTUDIO DESKTOP FOR MACOS 11+

This version of RStudio is only supported on macOS 11 and higher. For earlier macOS environments, please [download a previous version](#).

Size: 374.55 MB | [SHA-256: ED87B818](#) | Version: 2023.03.0+386 | Released: 2023-03-16

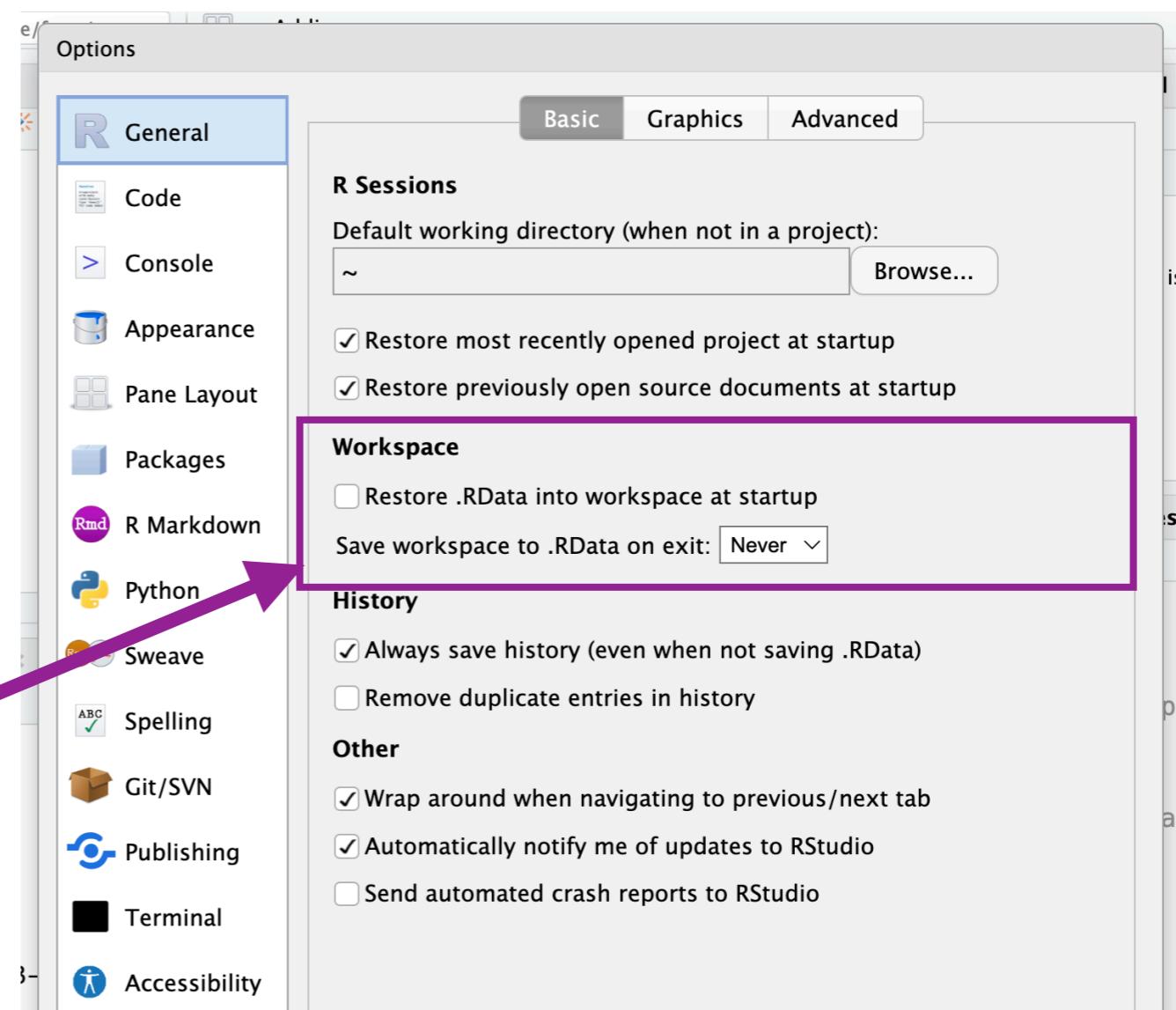
Setup RStudio

Choose
"Edit",
"Settings..."



Setup RStudio

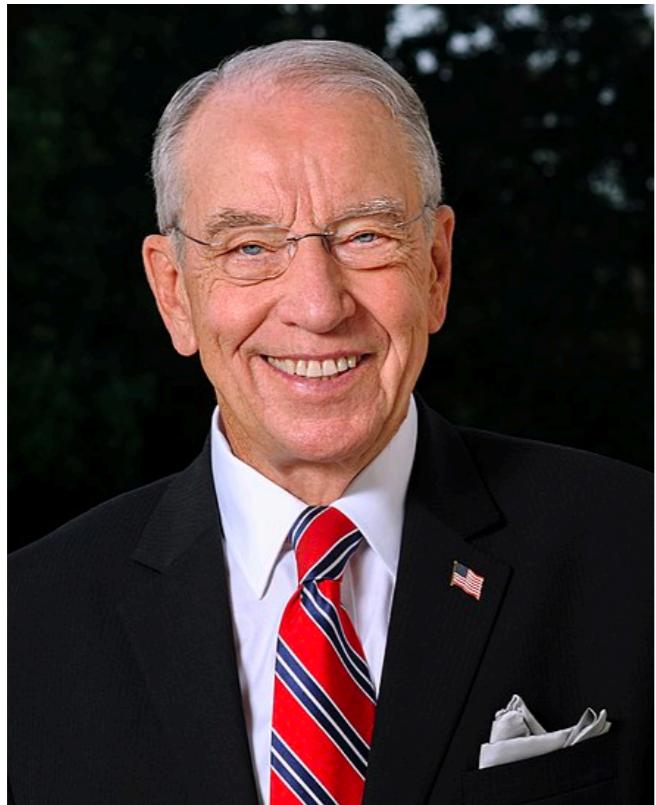
In the
"General"
"Workspace"
section,
choose these
settings.



Work in projects

- Your files will be organized and in the right place.
- You will never have to use `getwd()` or `setwd()` ever again!
- It will be easy to switch between projects.
- It will be easy to sync with GitHub.
- It is simple -- a project is just a folder (`project=folder=repo`)

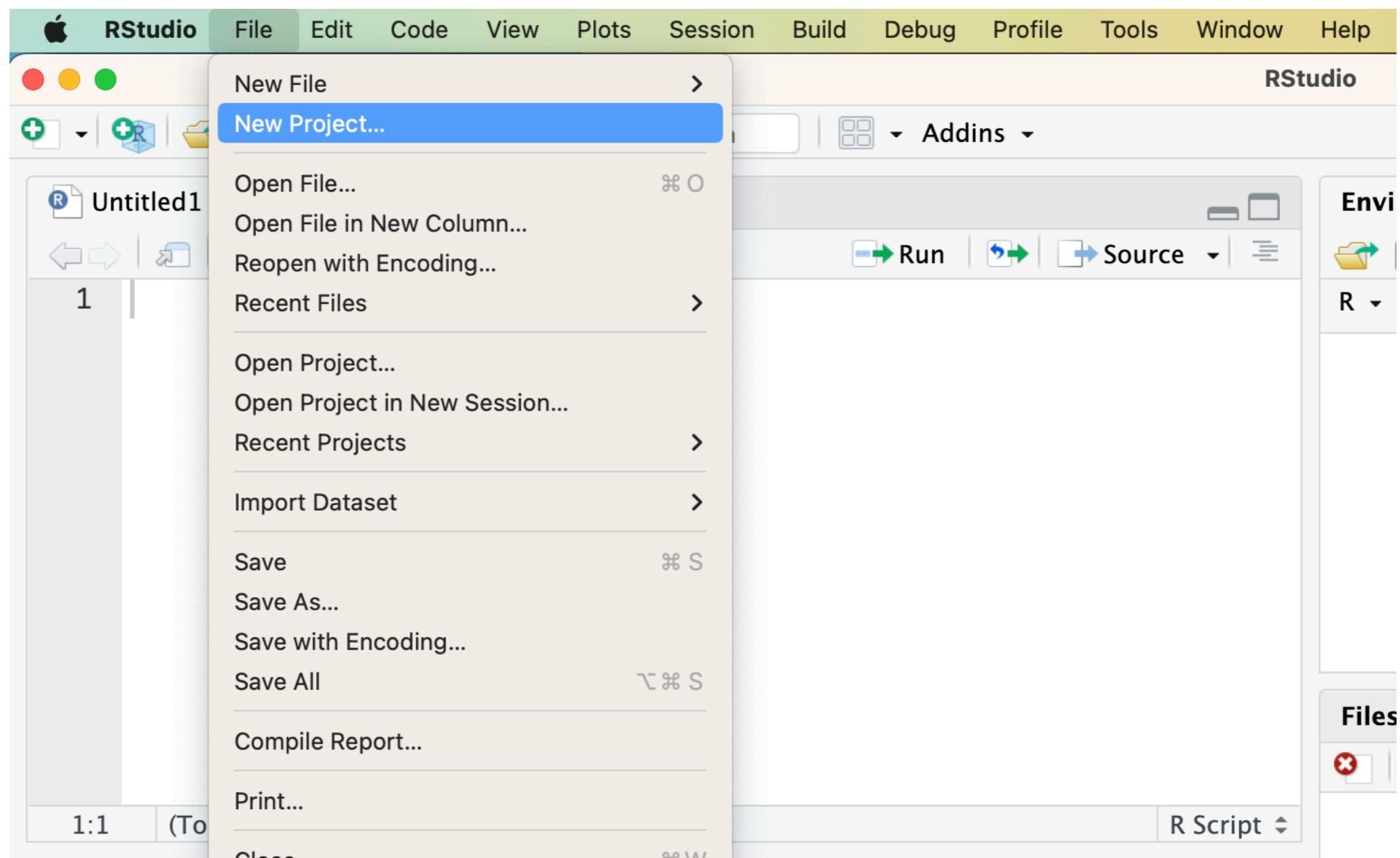
Senator Chuck Grassley's advice



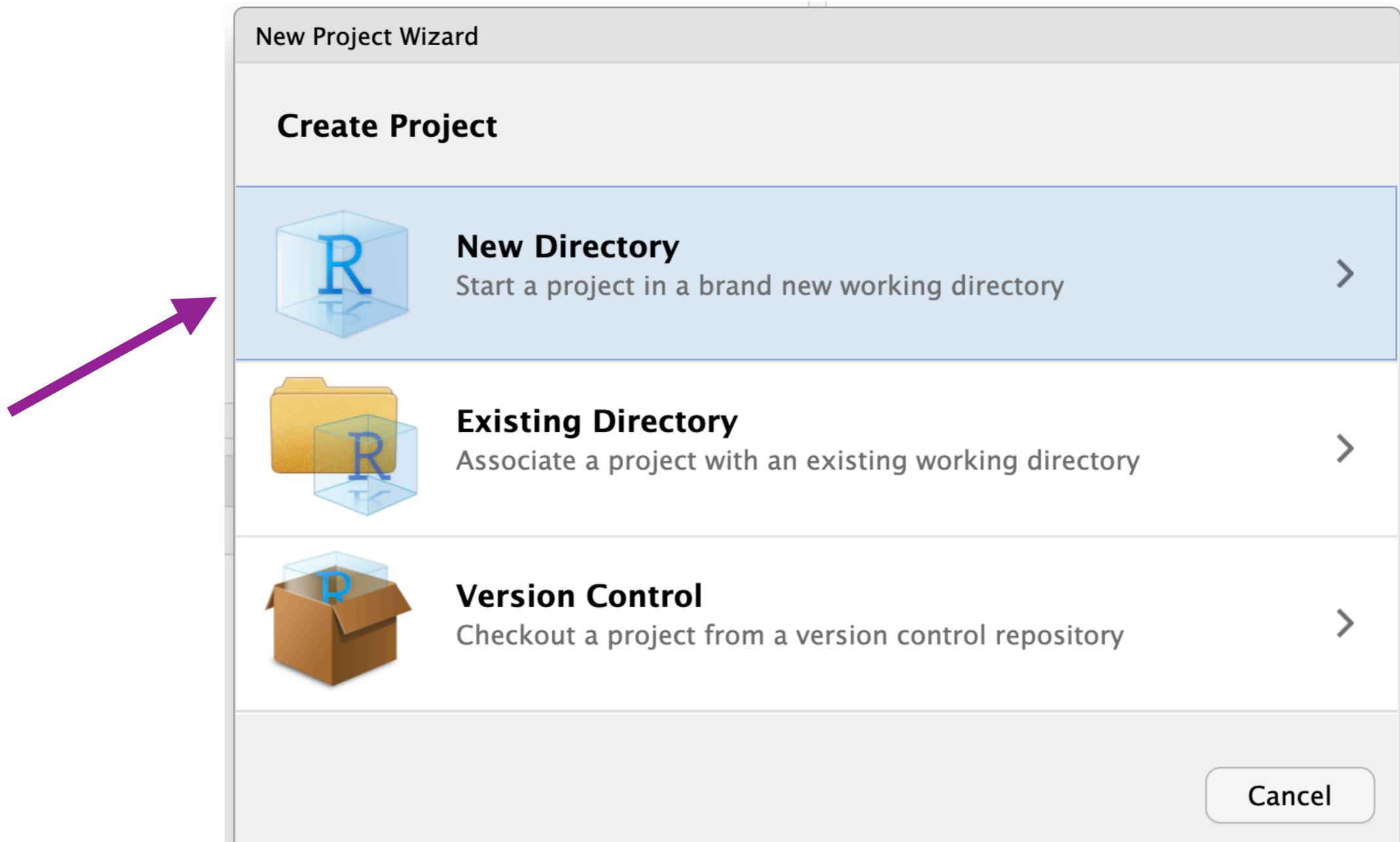
Wikipedia

- 2nd oldest senator (DOB: 9-17-33)
- *Master the Latest Tools*
"I like to be on the forefront of technology that I can use to better serve Iowans. Whether by fax, satellite, Instagram, or Twitter, I'd use whatever platform works. In the digital age, tweets are an instant, unfiltered way to get my work and my message [out]."

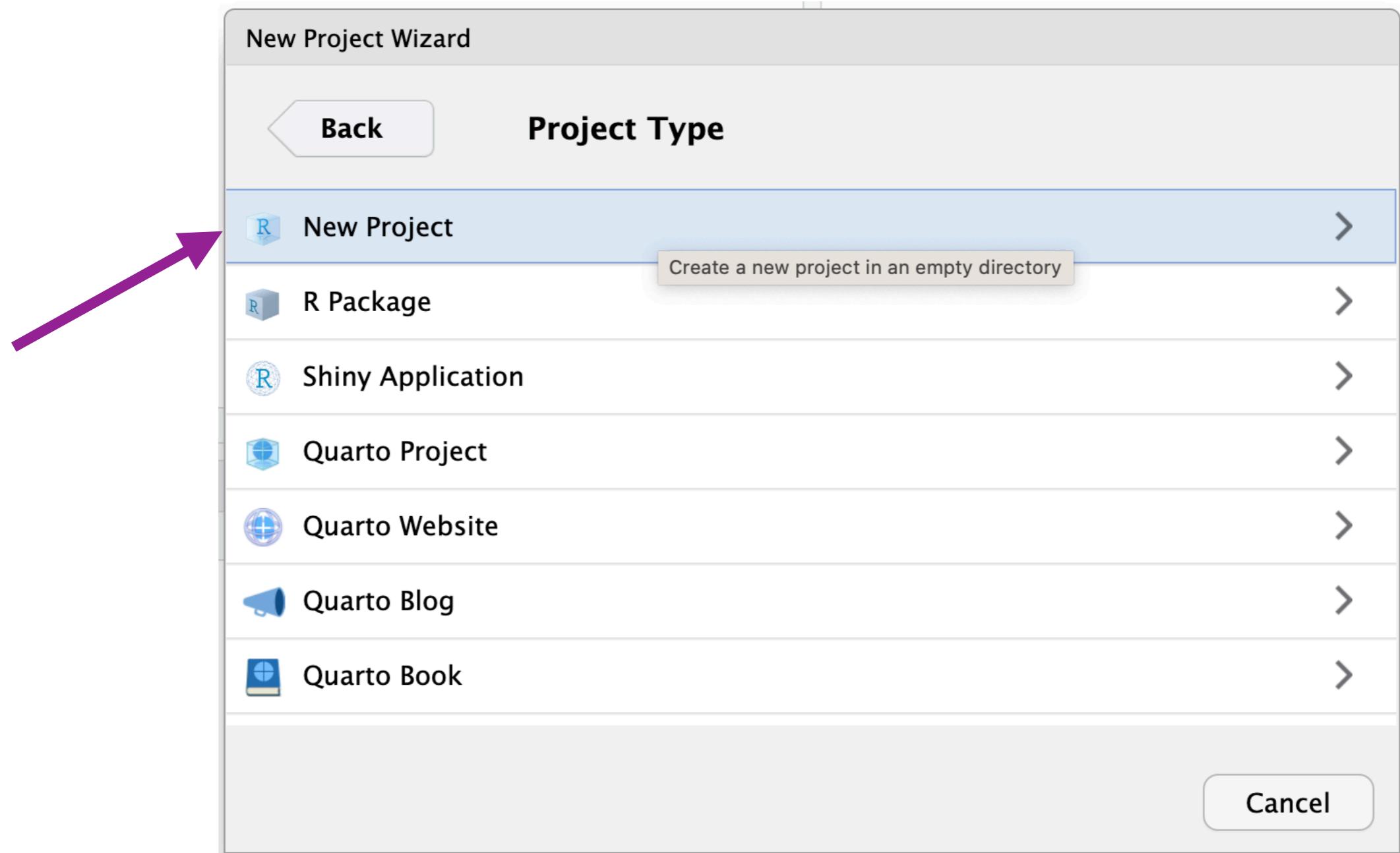
Create a project



Create a project

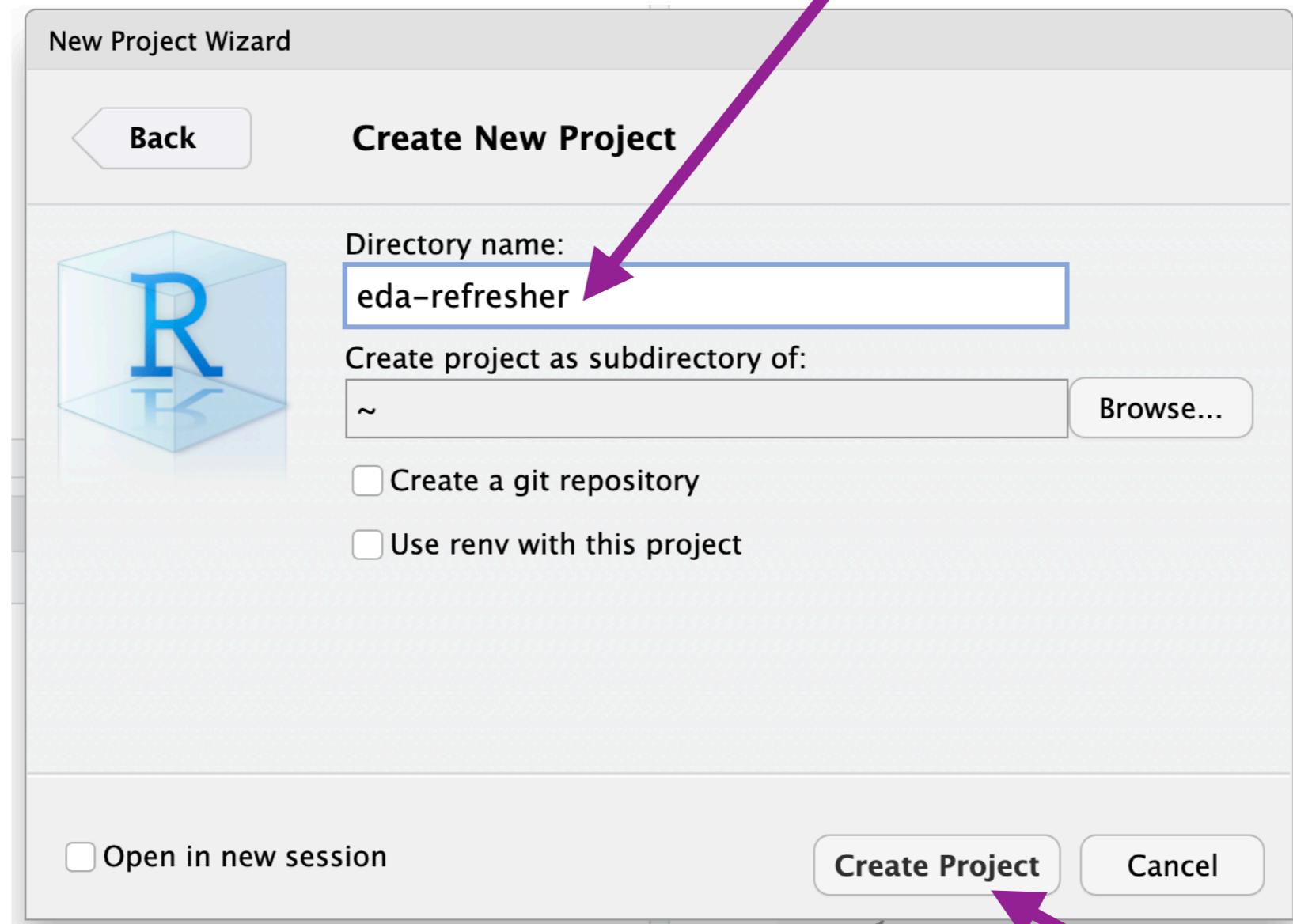


Create a project

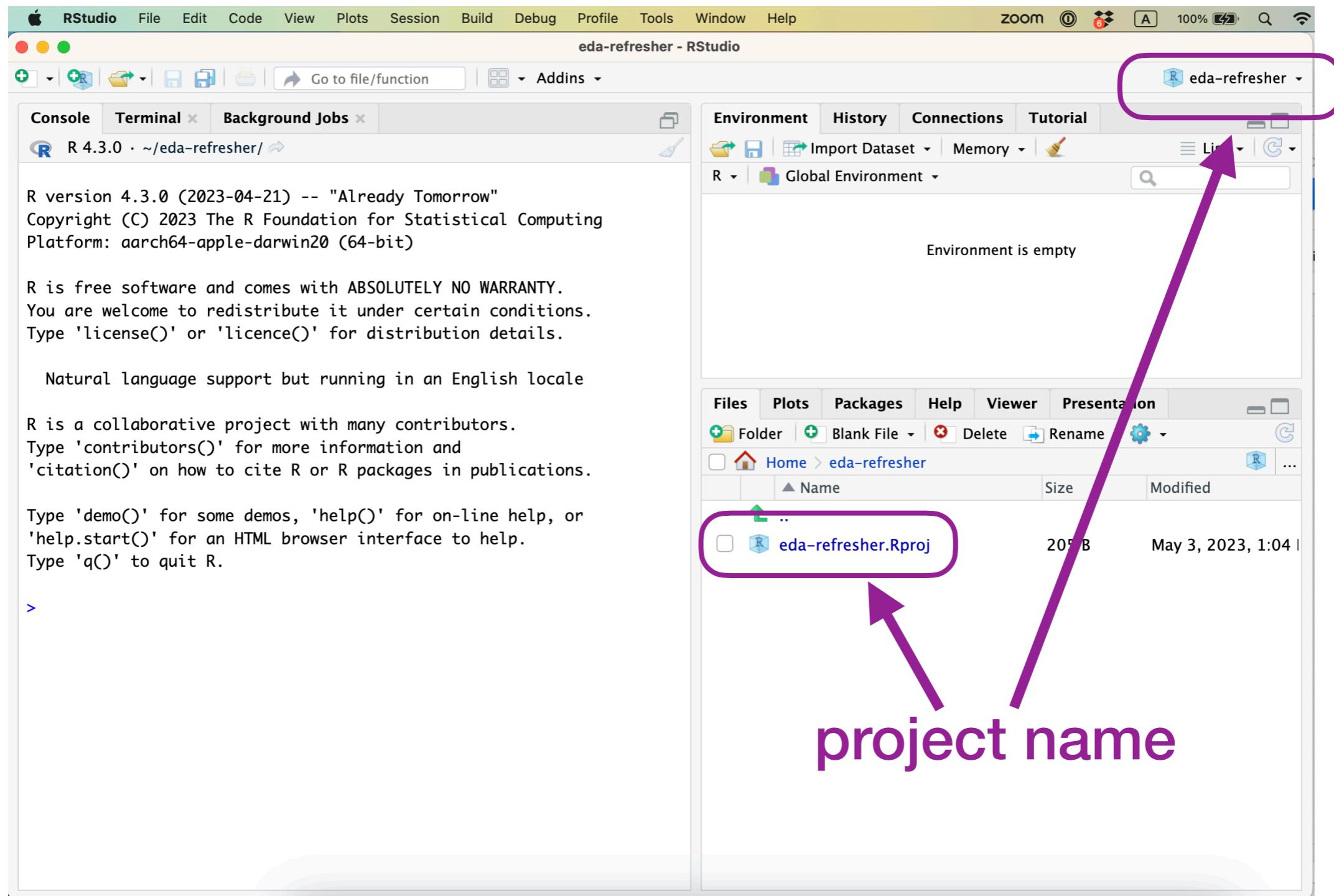


Create a project

choose a name



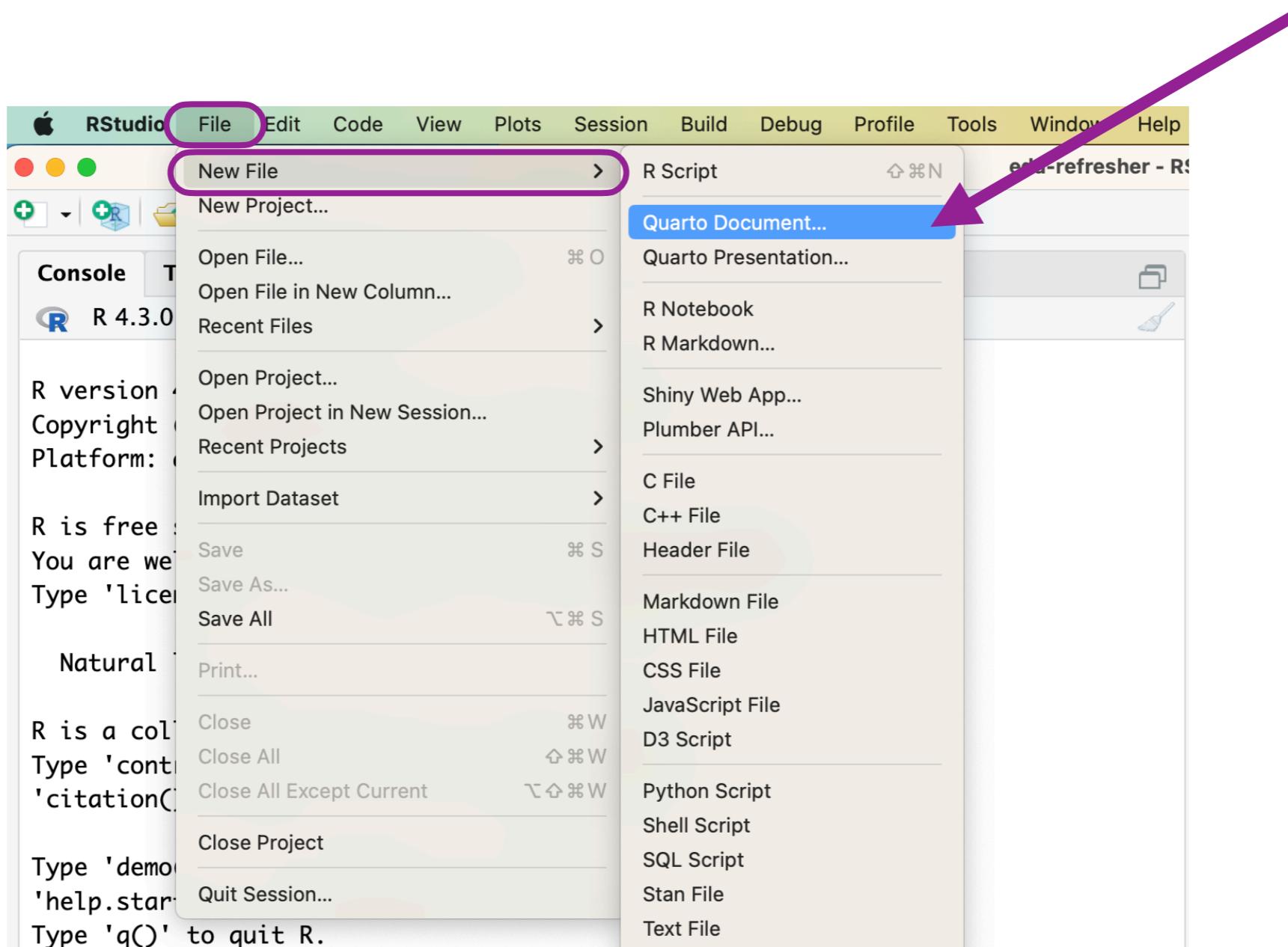
Create a project



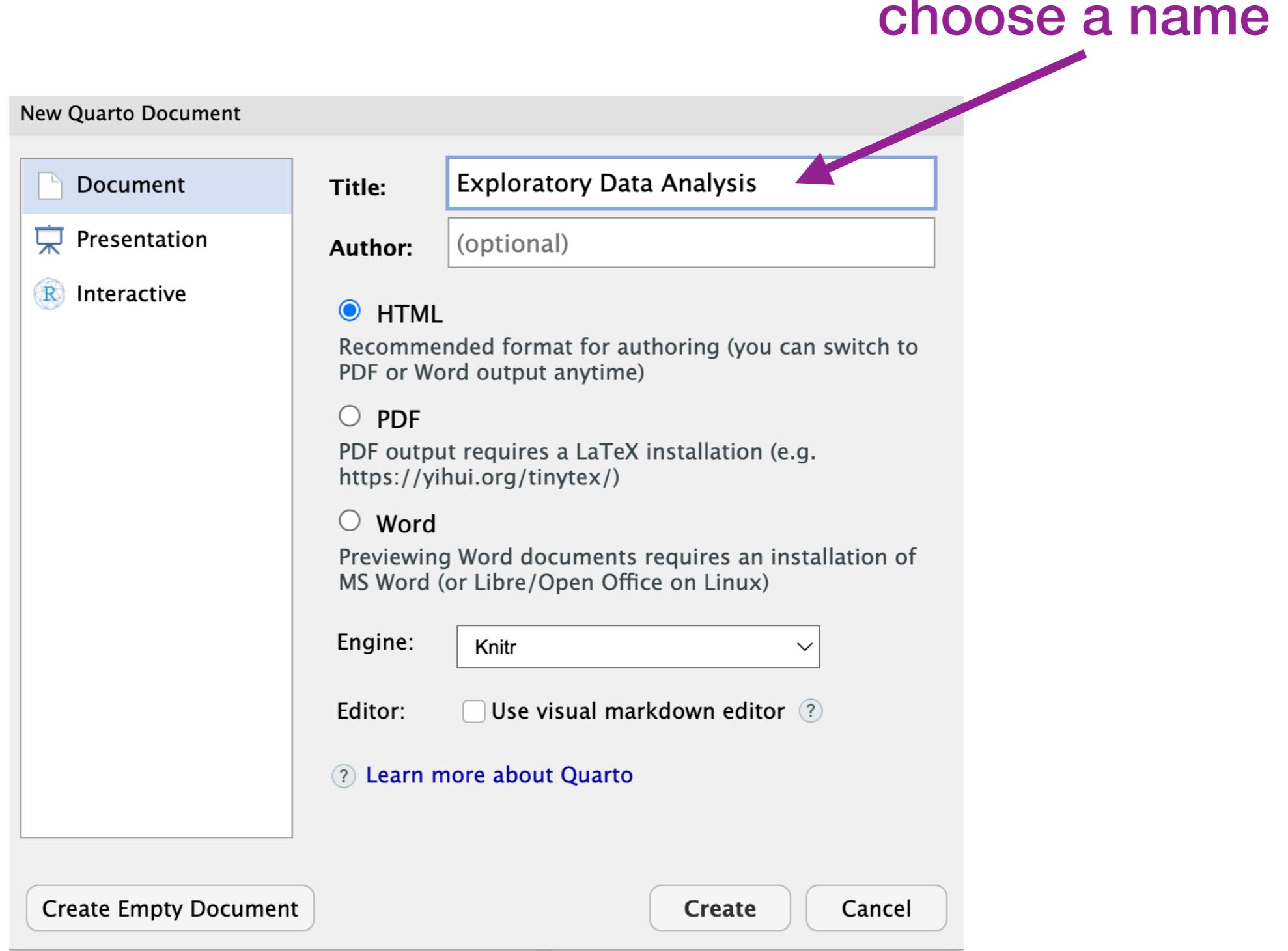
Use Quarto or RMarkdown

- Your workflow will be reproducible -- change the data file and you will have a revised output with one click
- No more copying and pasting text, code, output, and plots.
- You will have fewer mistakes and mismatches between stages of analysis.
- It will be easy to share your work.

Create a Quarto document



Create a Quarto document



Parts of a Quarto document

The screenshot shows the RStudio interface with a Quarto document open. The document content is as follows:

```
1 ---  
2 title: "Exploratory Data Analysis"  
3 format: html  
4 ---  
5 ## Quarto  
6  
7 Quarto enables you to weave together content and executable code into a  
finished document. To learn more about Quarto see <https://quarto.org>.  
8  
9 ## Running Code  
10  
11 When you click the **Render** button a document will be generated that  
includes both content and the output of embedded code. You can embed code  
like this:  
12  
13 ```{r}  
14 1 + 1  
15 ...  
16  
17  
18 You can add options to executable code like this  
19  
20 ```{r, echo=FALSE}  
19:1 # Running Code
```

Annotations with arrows point to specific parts of the code:

- A purple arrow points to the YAML header (lines 1-4) with the label "YAML".
- A purple arrow points to the heading (line 5) with the label "heading".
- A purple arrow points to the code chunk (lines 13-16) with the label "code chunk".
- A purple arrow points to the text "Quarto enables you to weave together content and executable code into a finished document." with the label "text".

Base R vs. tidyverse

- Not either / or
- Statistical functions and packages are the same
- Tidyverse fills in gaps to make R more versatile
- I find tidyverse more intuitive and convenient
- At least know what it's about and make an informed decision

A screenshot of a web browser window displaying the tidyverse.org homepage. The title bar shows the page is titled "Tidyverse". The main content area features the word "Tidyverse" in large white letters. Below it is a grid of hexagonal icons representing various R packages: dplyr (top left), ggplot2 (bottom left), readr (bottom left), purrr (bottom center), stringr (bottom right), tidyverse (center), forcats (middle right), and lubridate (top right). To the right of the grid, there is descriptive text about the tidyverse and a code block for installing it.



R packages for data science

The tidyverse is an opinionated **collection of R packages** designed for data science. All packages share an underlying design philosophy, grammar, and data structures.

Install the complete tidyverse with:

```
install.packages("tidyverse")
```

What is the philosophy?

- *(partial, unofficial list)*
- **consistency**
- **human readability**
 - `grepl()` --> `str_detect()`
 - `gsub()` --> `str_replace_all()`
- **fail faster**
- **focus on data transformation**

Data transformation

```
# base R
crime.by.state <- read.csv("CrimeStatebyState.csv")
crime.ny.2005 <- crime.by.state[crime.by.state$Year==2005 &
                                crime.by.state$State=="New York",
                                c("Type.of.Crime", "Count")]
crime.ny.2005 <- crime.ny.2005[order(crime.ny.2005$Count,
                                decreasing=TRUE), ]
crime.ny.2005$Proportion <- crime.ny.2005$Count /
                                sum(crime.ny.2005$Count)
summary1 <- aggregate(Count ~ Type.of.Crime,
                        data=crime.ny.2005, FUN=sum)
summary2 <- aggregate(Count ~ Type.of.Crime,
                        data=crime.ny.2005, FUN=length)
final <- merge(summary1, summary2, by="Type.of.Crime")
```

<https://www.r-bloggers.com/2014/02/how-dplyr-replaced-my-most-common-r-idioms/>

Data transformation

dplyr package

```
# dplyr
crime.by.state <- read.csv("CrimeStatebyState.csv")
final <- crime.by.state |>
  filter(State=="New York", Year==2005) |>
  arrange(desc(Count)) |>
  select(Type.of.Crime, Count) |>
  mutate(Proportion=Count/sum(Count)) |>
  group_by(Type.of.Crime) |>
  summarise(num.types = n(), counts = sum(Count))
```

base R pipe (R 4.1)

<https://www.r-bloggers.com/2014/02/how-dplyr-replaced-my-most-common-r-idioms/>

ggplot2

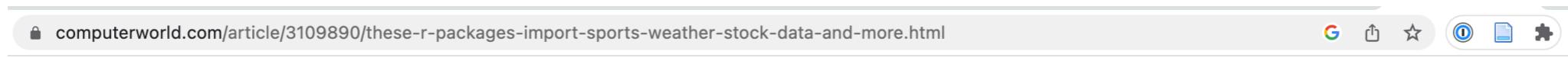
- underlying logic (gg = grammar of graphics)
- faceting
- automatic legends
- scales
- themes

Finding / importing data

Finding data

- Look for reputable sources
- Don't necessarily go with the first find
- Consider pros and cons of different options
 - R API package
 - Read from URL (or download file(s))
 - Web interface, then download
 - Web scraping

R API package



The screenshot shows a web browser window with a title bar containing the URL "computerworld.com/article/3109890/these-r-packages-import-sports-weather-stock-data-and-more.html". Below the title bar is a toolbar with various icons. The main content is a table titled "R packages to import public data". The table has columns for PACKAGE, CATEGORY, DESCRIPTION, SAMPLE CODE, and MORE INFO. It lists four packages: blscrapeR, quantmod, Bureau of Economic Analysis, and edgarWebR.

R packages to import public data				
PACKAGE	CATEGORY	DESCRIPTION	SAMPLE CODE	MORE INFO
blscrapeR	Economics, Government	For specific information about U.S. salaries and employment info, the Bureau of Labor Statistics offers a wealth of data available via this new package. blsAPI package is another option. CRAN.	<code>bls_api(c("LEU0254530800", "LEU0254530600"), startyear = 2000, endyear = 2015)</code>	Package vignettes
quantmod	Finance, Government	This package is designed for financial modelling but also has functions to easily pull data from Google Finance, Yahoo Finance and the St. Louis Federal Reserve (FRED). CRAN.	<code>getSymbols("DEXJPUS",src="FRED")</code>	Intro on getting data
Bureau of Economic Analysis	Economics, Government	Maintained by Andera Batch at BEA, this taps into the bureau's API to download data sets. CRAN.	<code>beaSpecs <- list('UserID' = beaKey , 'Method' = 'GetData', 'datasetname' = 'NIPA', 'TableName' = 'T20305', 'Frequency' = 'Q', 'Year' = 'X', 'ResultFormat' = 'json'); beaPayload <- beaGet(beaSpecs);</code>	See the GitHub repo , including info about recent project changes
edgarWebR	Finance, Government	This package is designed to let you search and download data from the U.S. Securities and Exchange Commission, including corporate and mutual-fund financial filings. CRAN.	<code>getSymbols("DEXJPUS",src="FRED")</code>	See the package vignette

<https://www.computerworld.com/article/3109890/these-r-packages-import-sports-weather-stock-data-and-more.html>

R API package: `quantmod`

```
library(quantmod)
getSymbols("AAPL")
#> [1] "AAPL"
tail(AAPL)
#>          AAPL.Open AAPL.High AAPL.Low AAPL.Close AAPL.Volume AAPL.Adjusted
#> 2023-04-25    165.19    166.31   163.73     163.77  48714100      163.77
#> 2023-04-26    163.06    165.28   162.80     163.76  45498800      163.76
#> 2023-04-27    165.19    168.56   165.19     168.41  64902300      168.41
#> 2023-04-28    168.49    169.85   167.88     169.68  55209200      169.68
#> 2023-05-01    169.28    170.45   168.64     169.59  52472900      169.59
#> 2023-05-02    170.09    170.35   167.54     168.54  48329100      168.54
```

```
plot(AAPL$AAPL.Close)
```



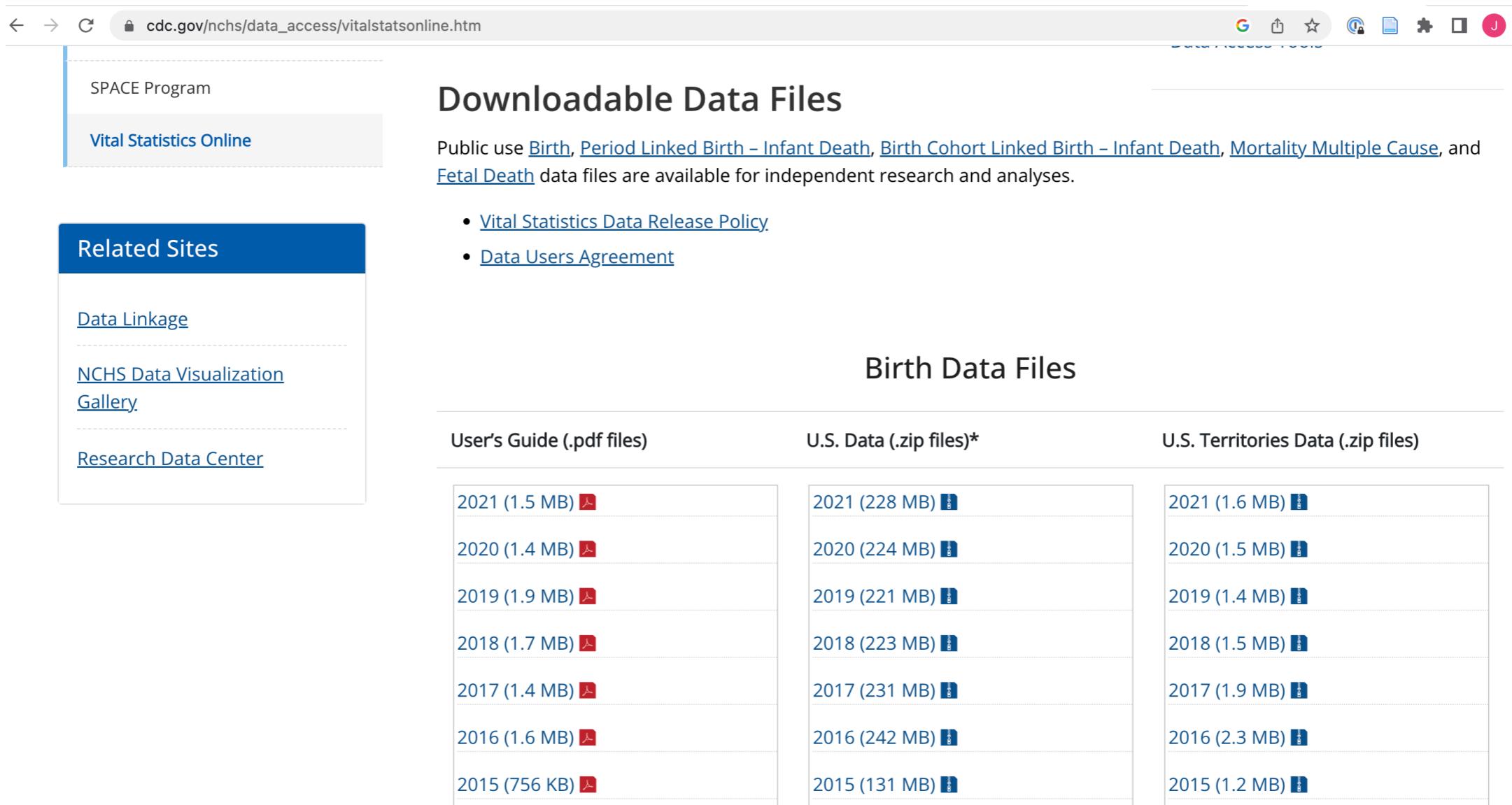
R API Package

- Particularly useful if data has frequent updates
- For some analyses you may want to set an end date so results don't change every time you run the code
- Add "R package" to internet searches for data to find these packages
- Many require API keys, but not hard to obtain
- Check for data transformations, etc.

Read from URL or download file(s)

- Very common, convenient way to get data
- Often data is available in multiple formats
- Content is not always the same
- Consider reading directly rather than downloading first, at least for small files

CDC Natality Database



The screenshot shows a web browser displaying the CDC Natality Database download page. The URL in the address bar is [cdc.gov/nchs/data_access/vitalstatsonline.htm](https://www.cdc.gov/nchs/data_access/vitalstatsonline.htm). The page has a header with the "SPACE Program" and "Vital Statistics Online" sections. A sidebar on the left lists "Related Sites" including "Data Linkage", "NCHS Data Visualization Gallery", and "Research Data Center". The main content area is titled "Downloadable Data Files" and describes public use data files for research. It includes links to the "Vital Statistics Data Release Policy" and "Data Users Agreement". Below this, there are three columns of data files: "User's Guide (.pdf files)", "U.S. Data (.zip files)*", and "U.S. Territories Data (.zip files)". Each column lists data files from 2015 to 2021, with file sizes and download icons.

Downloadable Data Files

Public use [Birth](#), [Period Linked Birth – Infant Death](#), [Birth Cohort Linked Birth – Infant Death](#), [Mortality Multiple Cause](#), and [Fetal Death](#) data files are available for independent research and analyses.

- [Vital Statistics Data Release Policy](#)
- [Data Users Agreement](#)

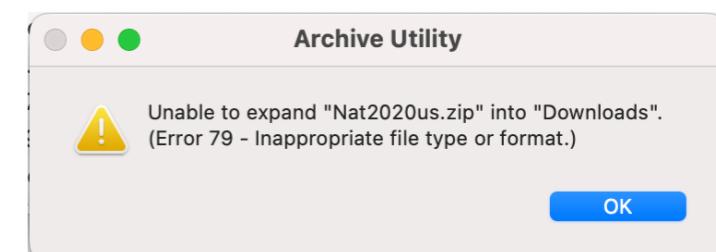
Birth Data Files

User's Guide (.pdf files)	U.S. Data (.zip files)*	U.S. Territories Data (.zip files)
2021 (1.5 MB) 	2021 (228 MB) 	2021 (1.6 MB)
2020 (1.4 MB) 	2020 (224 MB) 	2020 (1.5 MB)
2019 (1.9 MB) 	2019 (221 MB) 	2019 (1.4 MB)
2018 (1.7 MB) 	2018 (223 MB) 	2018 (1.5 MB)
2017 (1.4 MB) 	2017 (231 MB) 	2017 (1.9 MB)
2016 (1.6 MB) 	2016 (242 MB) 	2016 (2.3 MB)
2015 (756 KB) 	2015 (131 MB) 	2015 (1.2 MB)

https://www.cdc.gov/nchs/data_access/vitalstatsonline.htm

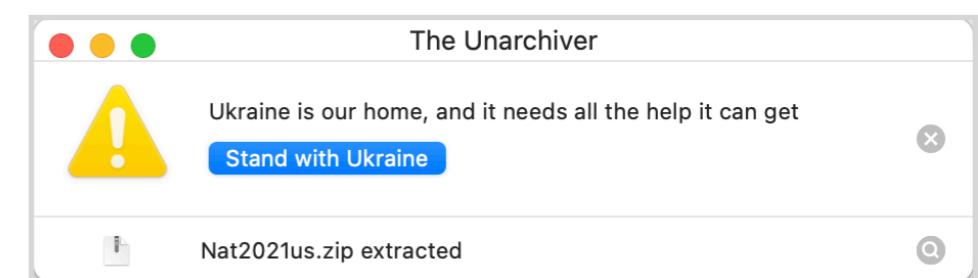
CDC Natality Database ZIP files

- 2021 ZIP file (228MB) downloaded in less than a minute



- Couldn't unzip with default utility

- Worked with another app,
The Unarchiver



- Unzipped .txt file is almost 5GB, but could read the file in less than a minute

https://www.cdc.gov/nchs/data_access/vitalstatsonline.htm

CDC Natality Database .txt files

BUT .txt file is fixed width, need to specify column widths:

Position	Length	Field	Description	Values	Definition
1-8	6	FILLER	Filler	Blank	
9-12	4	DOB_YY	Birth Year	2021	Year of birth
13-14	2	DOB_MM	Birth Month	01	January
				02	February
				03	March
				04	April
				05	May
				06	June
				07	July
				08	August
				09	September
				10	October
				11	November
				12	December

[https://ftp.cdc.gov/pub/Health_ Statistics/NCHS/Dataset_Documentation/DVS/nativity/UserGuide2021.pdf](https://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/DVS/nativity/UserGuide2021.pdf)

CDC Natality Database .txt files

to:

Position	Length	Field	Description	Values	Definition
				U	Unknown or not stated
569	1	BFED	Infant Breastfed at Discharge	Y N U	Yes No Unknown or not stated
570	1	F_BFED	Reporting Flag for Breastfed at Discharge	0 1	Non-Reporting Reporting
571-1330	760	FILLER_X	Filler	Blank	

[https://ftp.cdc.gov/pub/Health_ Statistics/NCHS/Dataset_Documentation/
DVS/nativity/UserGuide2021.pdf](https://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/DVS/nativity/UserGuide2021.pdf)

CDC Natality Database .txt files

```
library(readr)
read_fwf("~/Downloads/Nat2021us.txt",
         col_positions = fwf_widths(c(8, 4, 2, 4, 4, 1)),
         n_max = 8, show_col_types = FALSE)
#> # A tibble: 8 × 6
#>   X1      X2  X3    X4      X5      X6
#>   <lgl> <dbl> <chr> <lgl> <chr> <dbl>
#> 1 NA     2021 01    NA    0636      7
#> 2 NA     2021 01    NA    0259      7
#> 3 NA     2021 01    NA    0223      1
#> 4 NA     2021 01    NA    0241      1
#> 5 NA     2021 01    NA    0503      1
#> 6 NA     2021 01    NA    2341      7
#> 7 NA     2021 01    NA    1800      7
#> 8 NA     2021 01    NA    0652      1
```

CDC Wonder web interface

The screenshot shows a web browser window for the CDC Wonder web interface. The title bar reads "CDC Natality, 2016-2021 expanded" and the URL is "wonder.cdc.gov/controller/datarequest/D149". The page header includes the CDC logo, a search bar, and links for "A-Z Index", "Search", "FAQs", "Help", "Contact Us", and "WONDER Search". Below the header is a navigation bar with links for "Request Form", "Results", "Map", "Chart", and "About". The main content area is titled "Nativity, 2016-2021 expanded Request". It contains tabs for "Nativity Information", "Dataset Documentation", "Other Data Access", "Data Use Restrictions", and "How to Use WONDER", along with "Save" and "Reset" buttons. A note at the top says "Make all desired selections and then click any **Send** button one time to send your request." The first section, "1. Organize table layout:", includes dropdown menus for "Group Results By" (Census Region of Residence, And By: None) and a "Notes" section stating "Default query is limited to most recent year of data, 2021. You can change this in section 10 below." Below this are sections for "Measures" (Default measures always checked and included. Check box to include any others.) and "Notes" (Default query is limited to most recent year of data, 2021. You can change this in section 10 below.). The "Measures" section lists various options like Births, Birth Rate, Fertility Rate, Percent of Total Births, etc., each with a checkbox and a "Standard Deviation" checkbox.

<https://wonder.cdc.gov/controller/datarequest/D149>

NBER Natality Birth Data

The screenshot shows a web browser window for the NBER Vital Statistics Natality Birth Data. The title bar reads "Vital Statistics Natality Birth Da". The address bar shows the URL "nber.org/research/data/vital-statistics-nativity-birth-data". The page header includes the NBER logo, navigation links for "Subscribe", "Media", "Open Calls", and "Login", and a search icon. The main navigation menu at the top has links for "Research", "Programs & Projects", "Conferences", "Affiliated Scholars", "NBER News", "Career Resources", and "About". Below the menu, a breadcrumb trail shows "Home > Public Use Data Archive > Vital Statistics Natality Birth Data". The main content area features a large section titled "Vital Statistics Natality Birth Data" with a decorative orange horizontal bar below it. To the right of this title are social media sharing icons for Twitter, LinkedIn, and Email.

[Natality Data](#) from the [National Vital Statistics System](#) of the [National Center for Health Statistics](#) provides demographic and health data for births occurring during the calendar year. The microdata is based on information abstracted from birth certificates filed in vital statistics offices of each State and District of Columbia.

Other available birth data are [Birth Cohort Linked Birth/Infant Death Data](#), [Period Linked Birth/Infant Death Data](#) from the Perinatal Mortality Data, and [Matched Multiple Birth Data](#).

The use of this data signifies agreement with [NCHS's data use rules](#). Works referring to the datasets or codebooks should contain a [citation to NCHS](#). Published material derived from this data should include a citation such as this at the bottom of the table: "Source: National Center for Health Statistics (span of years used)"

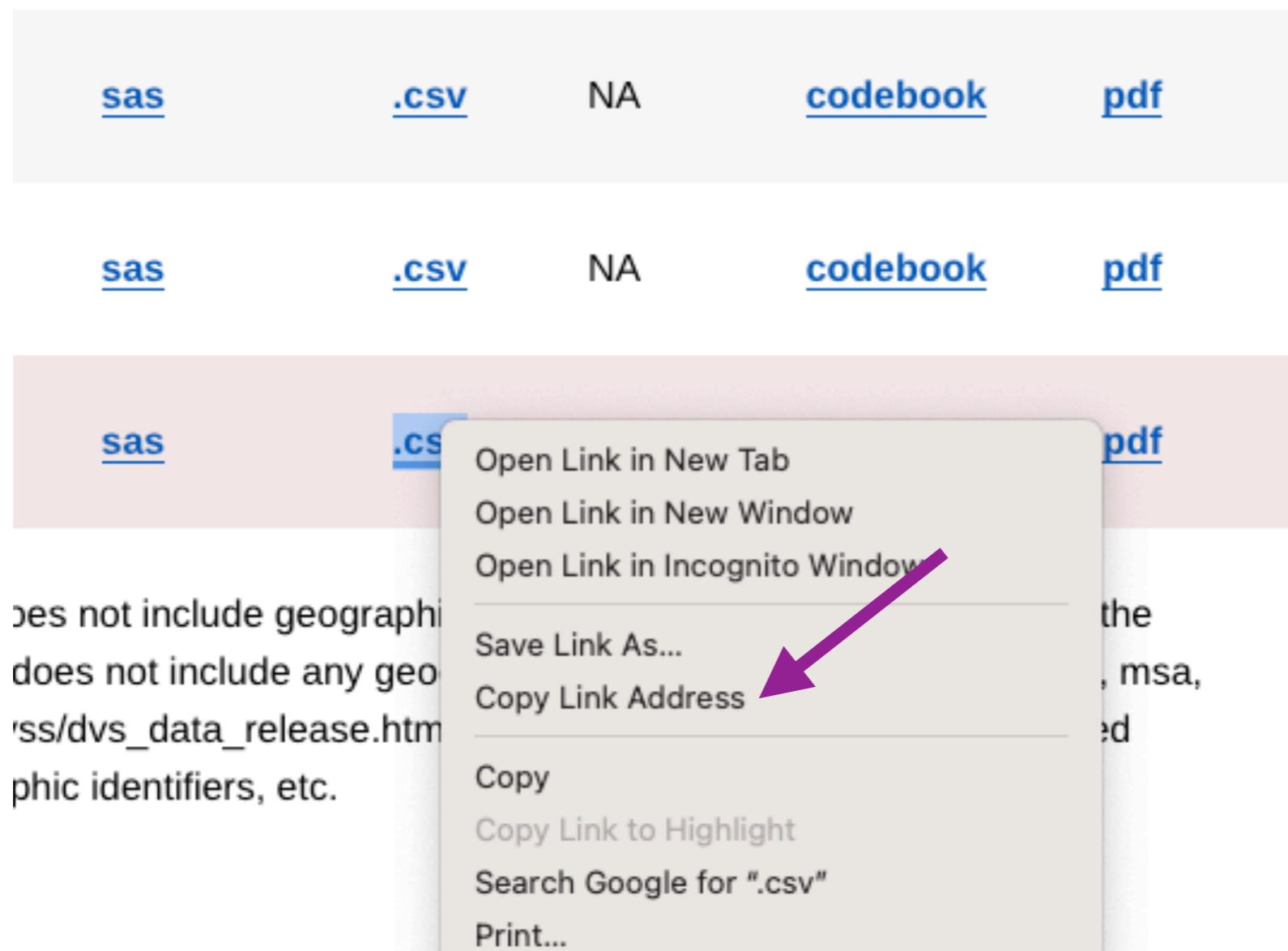
<https://www.nber.org/research/data/vital-statistics-nativity-birth-data>

NBER Natality Birth Data

2013	.zip	.dta , .do , .dct	sas , code	.csv	.spss	desc	pdf
2014	.zip	.dta , .do , .dct	sas , code	.csv	.spss	desc	pdf
2015	.zip	.dta , .do , .dct	sas , code	.csv	.spss	desc	pdf
2016	.zip	.dta , .do , .dct	sas , code	.csv	.spss	desc	pdf
2017	.zip	.dta , .do , .dct	sas , code	.csv	.spss	desc	pdf
2018	.zip	.dta	sas	.csv	NA	codebook	pdf
2019	.zip	.dta	sas	.csv	NA	codebook	pdf
2020	.zip	.dta	sas	.csv	NA	codebook	pdf
2021	.zip	.dta	sas	.csv	NA	codebook	pdf

<https://www.nber.org/research/data/vital-statistics-nativity-birth-data>

NBER Natality Birth Data



<https://www.nber.org/research/data/vital-statistics-nativity-birth-data>

NBER Natality Birth Data

```
df <- read.csv("https://data.nber.org/nvss/nativity/csv/  
nat2021us.csv", nrows = 10)  
df[,1:7]  
#>   dob_yy dob_mm dob_tt dob_wk bfacil f_facility bfacil3  
#> 1 2021     1     636      7       1           1       1  
#> 2 2021     1     259      7       1           1       1  
#> 3 2021     1     223      1       1           1       1  
#> 4 2021     1     241      1       1           1       1  
#> 5 2021     1     503      1       1           1       1  
#> 6 2021     1    2341      7       1           1       1  
#> 7 2021     1    1800      7       1           1       1  
#> 8 2021     1     652      1       1           1       1  
#> 9 2021     1     227      6       1           1       1  
#> 10 2021    1    2151      6       1           1       1
```

<https://www.nber.org/research/data/vital-statistics-nativity-birth-data>

Web scraping

The screenshot shows a web browser window with the URL <https://data.nber.org/nvss/nativity/code/nat2021us.html>. The page displays two tables of data.

Number of observations: 3,669,928

dob_yy		Num	Birth Year 2015 Year of birth
Frequency	Percent	Value	Label
3669928	100%	2021	

dob_mm		Num	Birth Month 01 January NOTE: Smallest 5 and largest 5 values are displayed.
Frequency	Percent	Value	Label
277533	8%	1	
266725	7%	2	
303139	8%	3	
293630	8%	4	
301343	8%	5	
640635	17%	6-7	NOTE: Range of values omitted from display
330740	9%	8	
326280	9%	9	
315909	9%	10	
302309	8%	11	
311685	8%	12	

<https://data.nber.org/nvss/nativity/code/nat2021us.html>

Web scraping with rvest

```
library(rvest)
df <- read_html("https://data.nber.org/nvss/nativity/code/
nat2021us.html") |>
  html_table()
df[[5]]
#> # A tibble: 9 × 4
#>   X1           X2           X3     X4
#>   <chr>        <chr>        <chr>  <chr>
#> 1 dob_wk      ""          Num    "Birth Day of Week 1 Sunday"
#> 2 Frequency  "Percent"  Value  "Label"
#> 3 355442     "10%"      1      ""
#> 4 548910     "15%"      2      ""
#> 5 593276     "16%"      3      ""
#> 6 593664     "16%"      4      ""
#> 7 591182     "16%"      5      ""
#> 8 579740     "16%"      6      ""
#> 9 407714     "11%"      7      ""
```

<https://data.nber.org/nvss/nativity/code/nat2021us.html>

The winner is...

- Download .csv file from:

<https://data.nber.org/nvss/nativity/csv/nat2021us.csv>

- Read once and save small versions:
 - One with 4 of 225 variables (all rows)
 - One with a sample of 1000 rows