Welcome! The session will begin at 9am Eastern

About this session:

- Ask questions in the chat
- Please complete the survey
- The speaker has no conflicts of interest associated with this talk
- Links to the recording and slide deck will be emailed to registrants, usually within a week
- Additional topics: https://its.weill.cornell.edu/scientific-computing-training-series

About CAC https://www.cac.cornell.edu/

The Center for Advanced Computing provides a broad range of services that support the computational research needs of Cornell faculty. The Center’s role is to enable and empower research capabilities and help Cornell’s scientists, researchers, and scholars achieve new and innovative breakthrough research. Our team provides expertise on high performance compute clusters, cloud computing through our Red Cloud service, and technology and application consulting tailored to individual research needs.

Center for Advanced Computing

www.cac.cornell.edu

Center for Advanced Computing

R Basics

Christopher Cameron
Computational Scientist
Cornell University Center for Advanced Computing
https://cac.cornell.edu/Cameron/
cjc73@cornell.edu

Today’s topics

Last semester, I focused on showcasing some of R’s best features.

1. Basic concepts needed to use R
2. How to install R and RStudio
3. The essential elements of the Rstudio interface
4. Loading and summarizing data

R takes time to learn, and this could be your first step. The materials and demonstrations today will help you get started.

R is…

Good for:

- tabular data
  (or vectors or lists)
- statistical analysis
- data visualization
- Integrating custom code in C/C++, Fortran and Java.

Less suitable for

- unstructured data
- file system scripting
- data scraping, cleaning and formatting

Some people want R to do everything, so packages do exist to make some of these possible!

(Someone also wrote a web-crawler in SAS)

Motivation for R

What if we combine things we like into a statistical computing environment and make it free and open source so others could do the same?

- Two faculty members at the University of Auckland wanted a “better software environment [for] their teaching laboratory” (1990s)
  - **did not like** the commercial offerings available
  - **did like** the S statistical programming language
  - **wished** S had some of the modern language features introduced in the Lisp variant called Scheme
- R started as an S implementation with some Scheme features and was distributed via an email list
- A colleague persuaded the authors to open-source R (1995)

Ihaka, Ross. (1998) R : Past and Future History, *A Draft of a Paper for Interface ’98*. https://cran.r-project.org/doc/html/interface98-paper/paper.html

Community

- R is used and supported by a community of largely academic researchers and developers (and more recently, data scientists).
- R gains new features via *packages* developed by the community
  - Over 10,000 add-on libraries!
  - R packages can target highly specialized research areas.
  - R packages are used to implement and share cutting edge statistical methodology.
  - The official package collection is at https://cran.r-project.org
  - Other collections exist: http://www.bioconductor.org.
  - Can load packages directly from github
- Active community generating tutorials and demos:
  - https://www.r-bloggers.com
  - https://education.rstudio.com/learn/
  - https://cvw.cac.cornell.edu/R/
  - https://community.rstudio.com ← community help forum

Collective, eclectic development

- R’s developers borrow code conventions and programming styles freely.
  - “object oriented” `object.member` naming is common but has no special meaning in R
  - Many conventions mixed together: InitalCaps, camelCase, snake_case, vars.with.dots (again, R does not assign special meaning)
  - Packages tend to work well with expected input and unpredictably with incorrect input.
  - Many ways to accomplish any given task, inspired by different paradigms.
- Focus on practical, productive use
  - automatic and silent type conversion (casting)
  - convenience features can become gotchas (global namespace, attach)
    - packages can mask each other’s functions
    - variable names can have the same name as functions – mostly works, hard to read

Documentation

R has built-in help and documentation

A typical help entry includes

- *Descriptions* of each function and their arguments.
- *Examples* showing how the functions might be used.
- *References* to relevant manuals and academic papers.

Documentation for packages usually also includes:

- One or more *vignettes* demonstrating how the package can be used to perform an analysis.
- Bundled *data sets* that support the vignette and demonstrate required data formats.

Mental model for using R

- R has a *workspace* (or *environment*) that holds data tables and results.
  - Workspace is not particularly visible
  - These *objects* are held in the computer's memory
  - Objects in the workspace can be manipulated by R commands (*functions*)

- You enter commands via the R console
  - to load data into the workspace (as an object)
  - to apply statistical functions to objects in the workspace
  - to produce or display output (most commands do not produce output!)

- Most of the time, you are looking at commands describing what to do with the data and not at the data itself. (c.f. Excel)

R Concepts

- *Functions* – This is the primary way to use R!
  - Function takes input and (probably) returns output.
  - Function input is one or more values called *arguments*
  - *Calling* a function is telling the function to operate on arguments.
    - Function name followed by arguments in parenthesis :
      - `name(argument)`
      - `name(argument1, arg_name = argument2)`
      - `mean(column) # calculate mean of vector named “column”`

R Concepts

- *Variables*
  - Store and use values in the workspace
  - Variables are a name and associated value
  - Variable name represents the value in operations and function calls
  - Values or objects returned by function calls can be stored in variables
- Create a variable by *assigning* a value to a name
  - Either `<-` or `=` are assignment operators in R
    - `width = 20`
    - `width <- 20`

R Concepts - Operators

- **Mathematical**

+	addition
–	subtraction
*	multiplication
/	division
^ or **	exponentiation
x %% y	modulus
x %/% y	integer division

- **Logical Operators**

!x	not x
x || y	x or y (returns TRUE or FALSE, use in if conditions)
x && y	x and y (returns TRUE or FALSE, use in if conditions)
x | y	x OR y (compares bitwise, so it potentially returns a vector)
x & y	x AND y (compares bitwise, so it potentially returns a vector)

- **Comparison**

<	less than
<=	less than or equal to
>	greater than
>=	greater than or equal to
==	equals (comparison)
!=	not equal

R Concepts

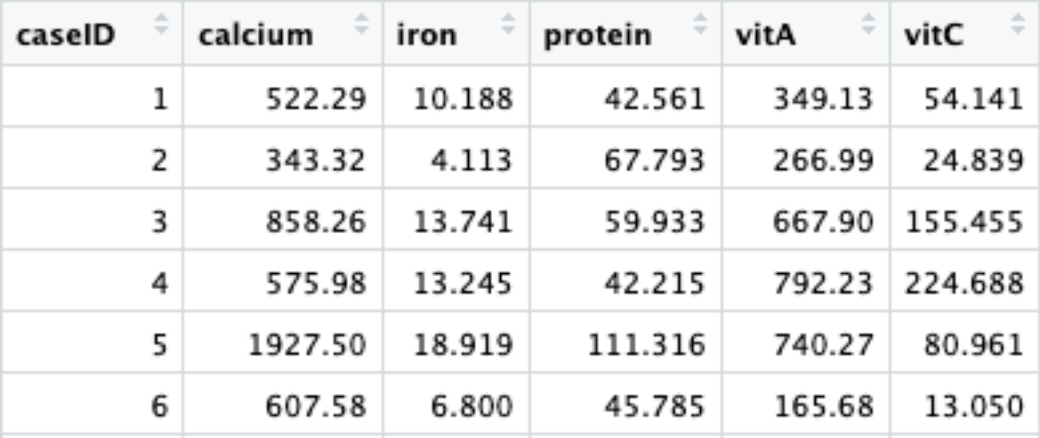
- *Working directory*
  - R always has a working directory from which it tries to read/write files.
  - You can change the working directory if needed.
  - Files outside of working directory can be accessed via full file paths
- *Packages*
  - collections of functions, data and documentation that add functionality to R.
- *Script* files
  - Sequence of commands in plain text with a .R file extension.
- *Rmarkdown* files
  - Narrative style markdown-based “R notebook” with .Rmd file extension.

R Concepts – Common Datatypes

- ***Numeric****:* represents a numeric value (integer, float, double, etc)
  - 3, 5.2, …
- ***Boolean****:* logical values
  - TRUE, FALSE
- ***Character****:* alphanumeric strings (text)
  - “cat”, “dog”, “a”
- ***Factor****:* categorical variable
  - Gender, cohort, income brackets
- ***Date[time]****:* Specific date or date and time
  - 2023-02-14 9:00:00 EST

R Concepts – Data Frames

- ***Data frame -*** a table-like arrangement of values in rows and columns.
  - Classic table of data structure like Excel or CSV files
  - Typically, rows are cases and columns are variables
  - All values in a column are same type, columns may be different types
  - **data.frame** (base R)**, data.table** (large data), **tibble** (modernized data.frame)
  - Dataframe$ColumnName



caseID	calcium	iron	protein	vitA	vitC
1	522.29	10.188	42.561	349.13	54.141
2	343.32	4.113	67.793	266.99	24.839
3	858.26	13.741	59.933	667.90	155.455
4	575.98	13.245	42.215	792.23	224.688
5	1927.50	18.919	111.316	740.27	80.961
6	607.58	6.800	45.785	165.68	13.050

R Concepts – Vectors

- ***Vector*** - 1-dimensional array of values of the same type
  - Like a single column from a data frame
  - Created by the `c()` function:
    - `my_vec = c(8, 6, 7, 5, 3, 0, 9)`
    - `my_names = c("caseID”, "calcium”, "iron”)`
  - *Indexed* by position, starting with 1:
    - `my_vec[1]` is `8`
    - `my_vec[7]` is `9`
    - `my_vec[1:4]` is `c(8, 6, 7, 5)`

R Concepts – List

- ***List*** - a collection of mixed types, optionally with names
  - Statistical functions tend to return results in list-like formats
  - Created by the `list()` function:

```
my_list2 = list(
  message = "I like R",
  yesterday =  as.Date("2023-02-13"),
  width = 20)
```

  - Retrieve by position, starting with 1:
    - `my_list2[[1]]` is `"I like R”`
  - Retrieve by name, if named (preferred):
    - `my_list2$message` is `"I like R”`
    - `names(my_list2)` is `c("message”, "yesterday”, "width”)`

R Concepts – Other data containers

- a stack of values → vector
- a stack of vectors with same data type → ***matrix***
- a stack of matrices with same data type → multidimensional ***array***

Base R

- The *R Project for Statistical Computing* is maintained by The R Foundation.
  - free and runs on Linux, Windows and MacOS.
  - https://www.r-project.org

- Command line interface via R console
  - Creates objects in memory rather than printing to screen
  - You query and manipulate these in-memory objects
  - Interactive, but not in the point-and-click GUI sense.

- Many people that “use R” do not use it directly. Instead, they use something that interfaces with the R environment.
  - RStudio IDE
  - Jupyter Lab notebooks
  - Google CoLab

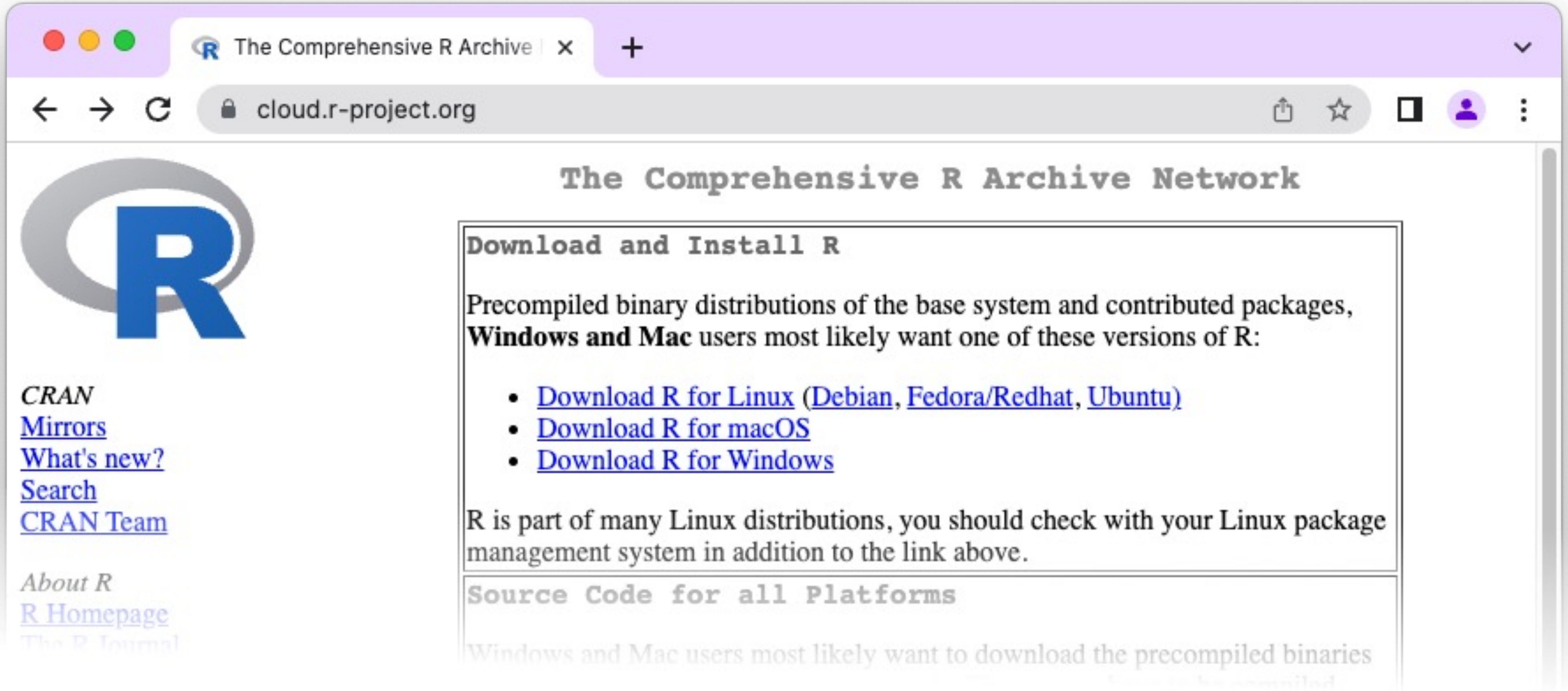
Installing R

- R is distributed via the Comprehensive R Archive Network (CRAN) which is mirrored by universities and other institutions around the world.

- ***choose a mirror:*** when installing R and packages from CRAN, the first step is to "choose a mirror". Choose an institution close to you.

- https://cloud.r-project.org is a "mirror" that automatically redirects you to a nearby mirror when installing R.

Downloading R https://cloud.r-project.org



The Comprehensive R Archive Network

CRAN
Mirrors
What's new?
Search
CRAN Team

About R
R Homepage
The R Journal

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

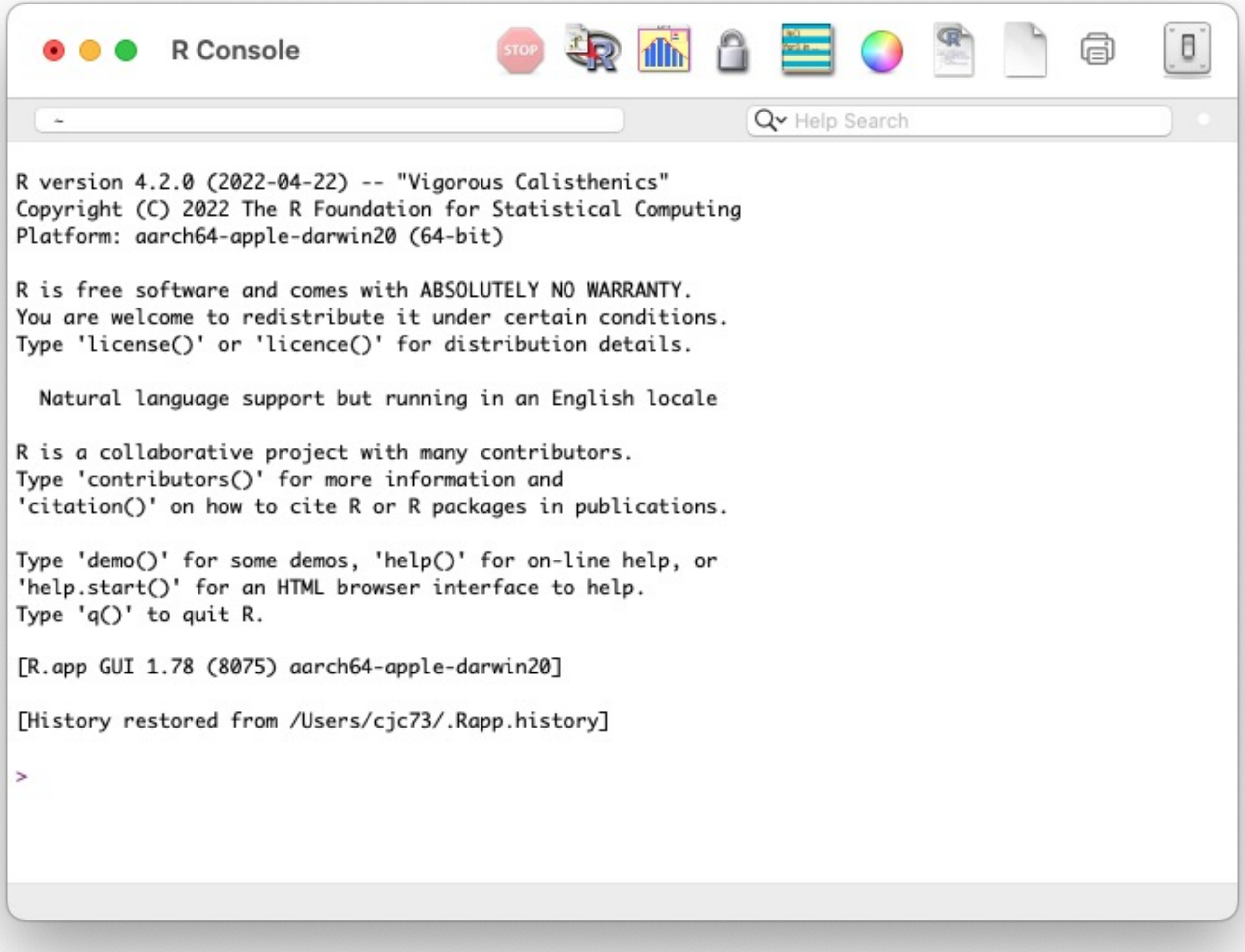
- Download R for Linux (Debian, Fedora/Redhat, Ubuntu)
- Download R for macOS
- Download R for Windows

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries

R Console



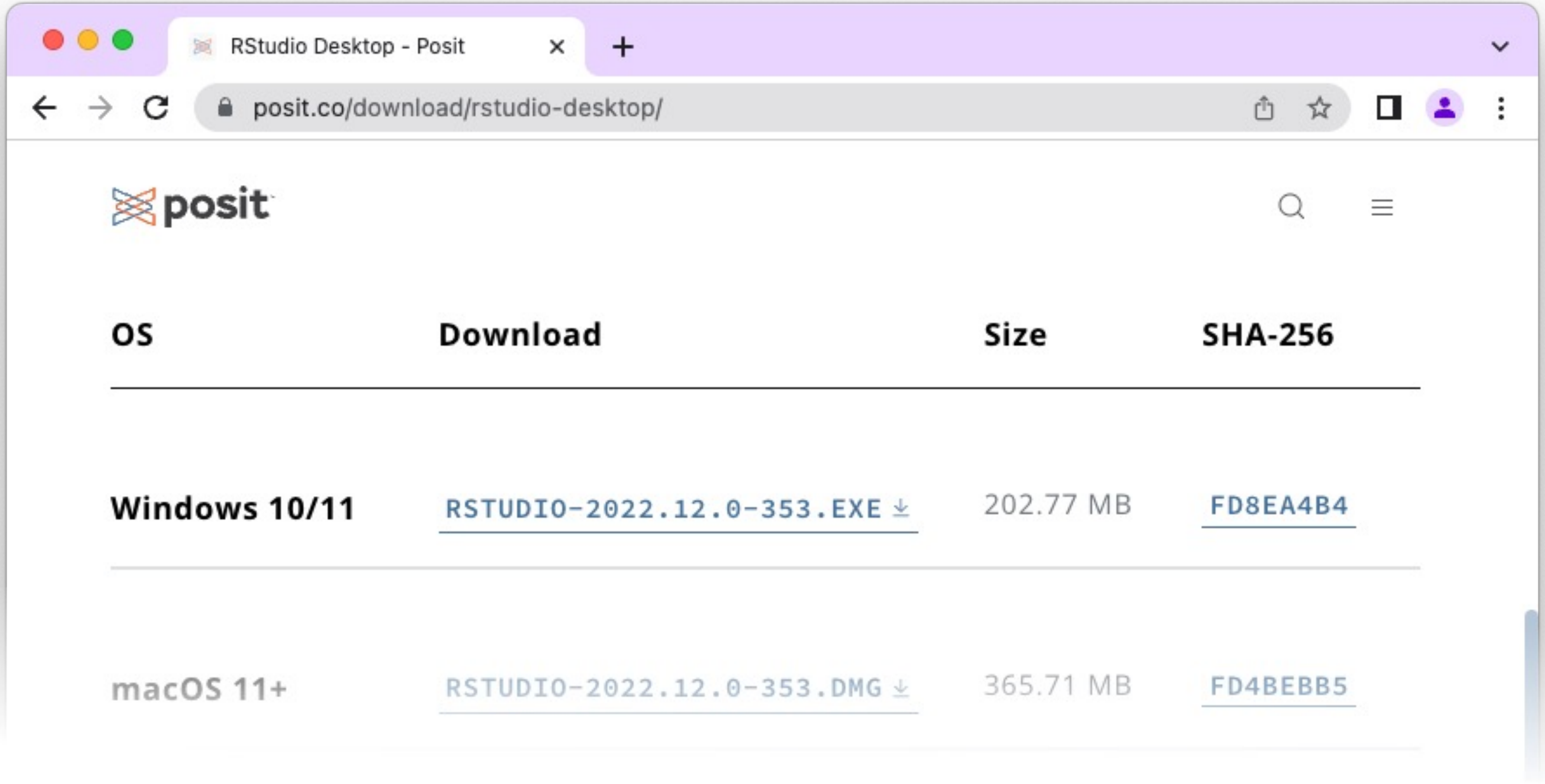
```
R version 4.2.0 (2022-04-22) -- "Vigorous Calisthenics"
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: aarch64-apple-darwin20 (64-bit)

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 1.78 (8075) aarch64-apple-darwin20]

[History restored from /Users/cjc73/.Rapp.history]

>
```

RStudio

- RStudio is an integrated development environment for R
  - developed by RStudio Public Benefit Corporation (now Posit)
  - depends on installed R version
  - adds useful development, analysis and authoring features
- RStudio interface incorporates the R Console
  - Posit will incorporate Python compatibility
- Tip: If you want to install RStudio locally, install R and *then* install RStudio
- RStudio Cloud (soon to be Posit Cloud) https://rstudio.cloud is a hosted version of RStudio with the same interface as the desktop application.

Downloading Rstudio

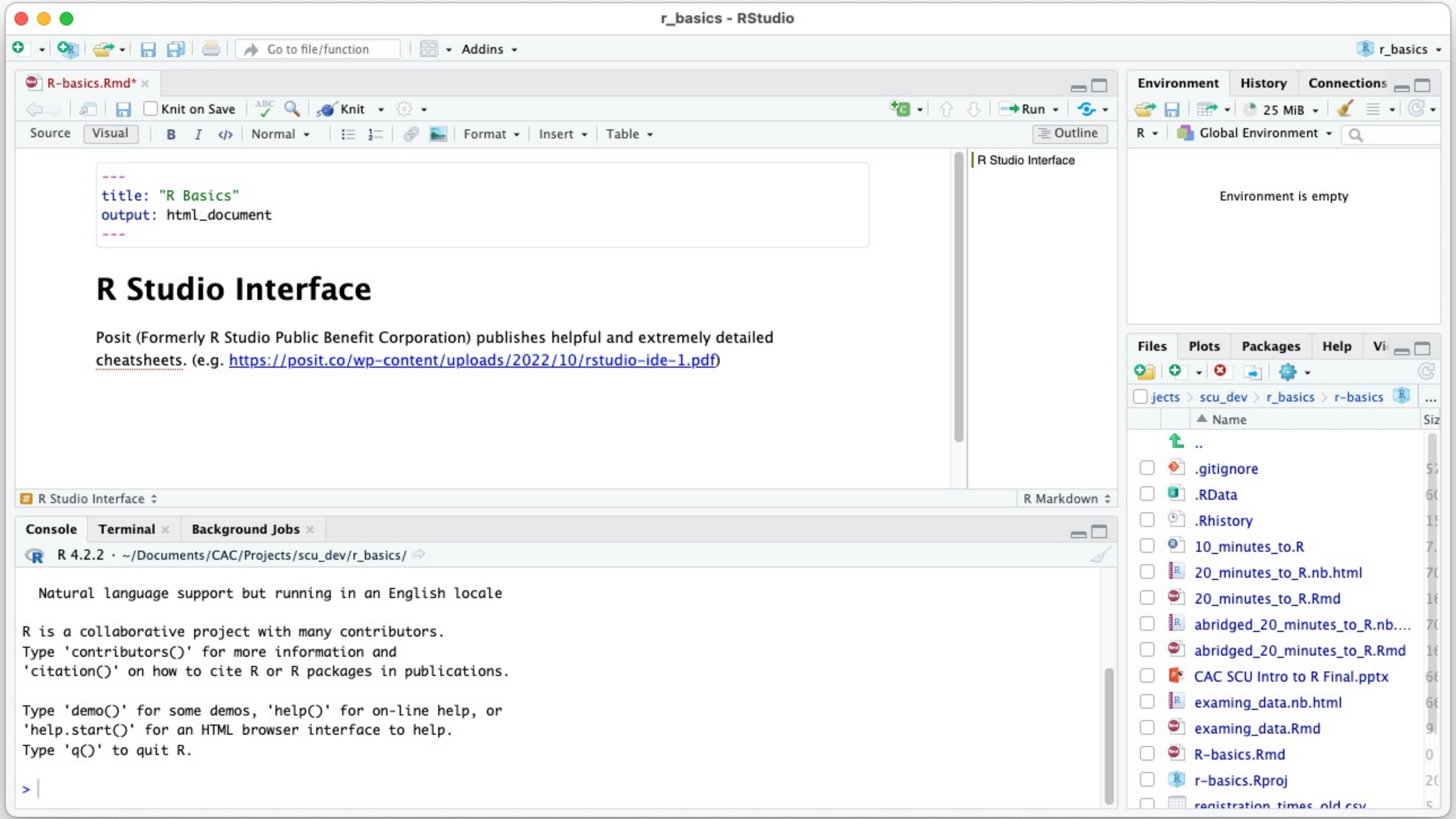
https://posit.co/download/rstudio-desktop/



OS	Download	Size	SHA-256
Windows 10/11	RSTUDIO-2022.12.0-353.EXE ⤓	202.77 MB	FD8EA4B4
macOS 11+	RSTUDIO-2022.12.0-353.DMG ⤓	365.71 MB	FD4BEBB5

Scroll down

RStudio Interface



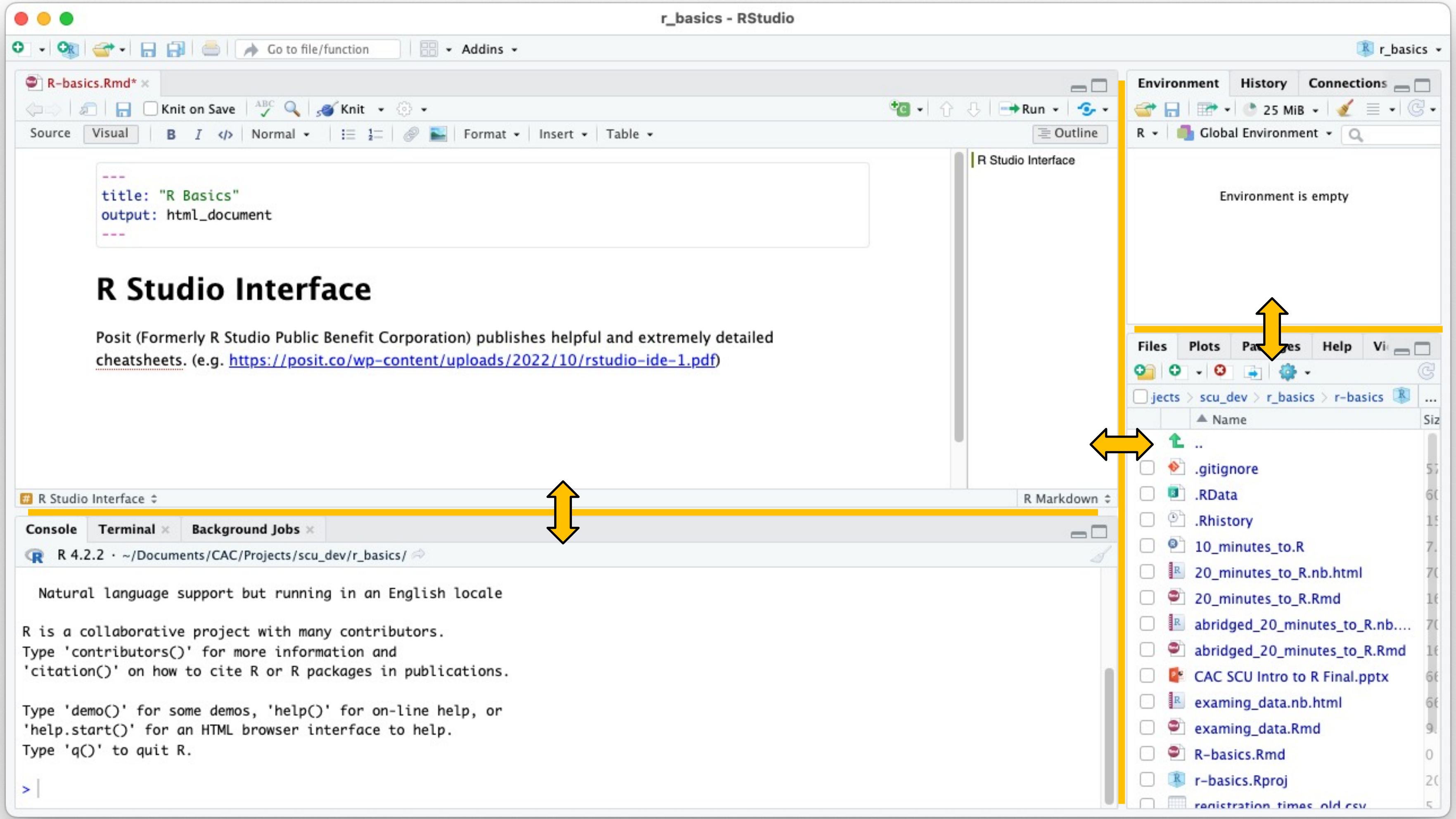
```
---
title: "R Basics"
output: html_document
---
```

R Studio Interface

Posit (Formerly R Studio Public Benefit Corporation) publishes helpful and extremely detailed cheatsheets. (e.g. https://posit.co/wp-content/uploads/2022/10/rstudio-ide-1.pdf)

```
  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.

>
```

RStudio Interface – Resizable Panes



```
---
title: "R Basics"
output: html_document
---
```

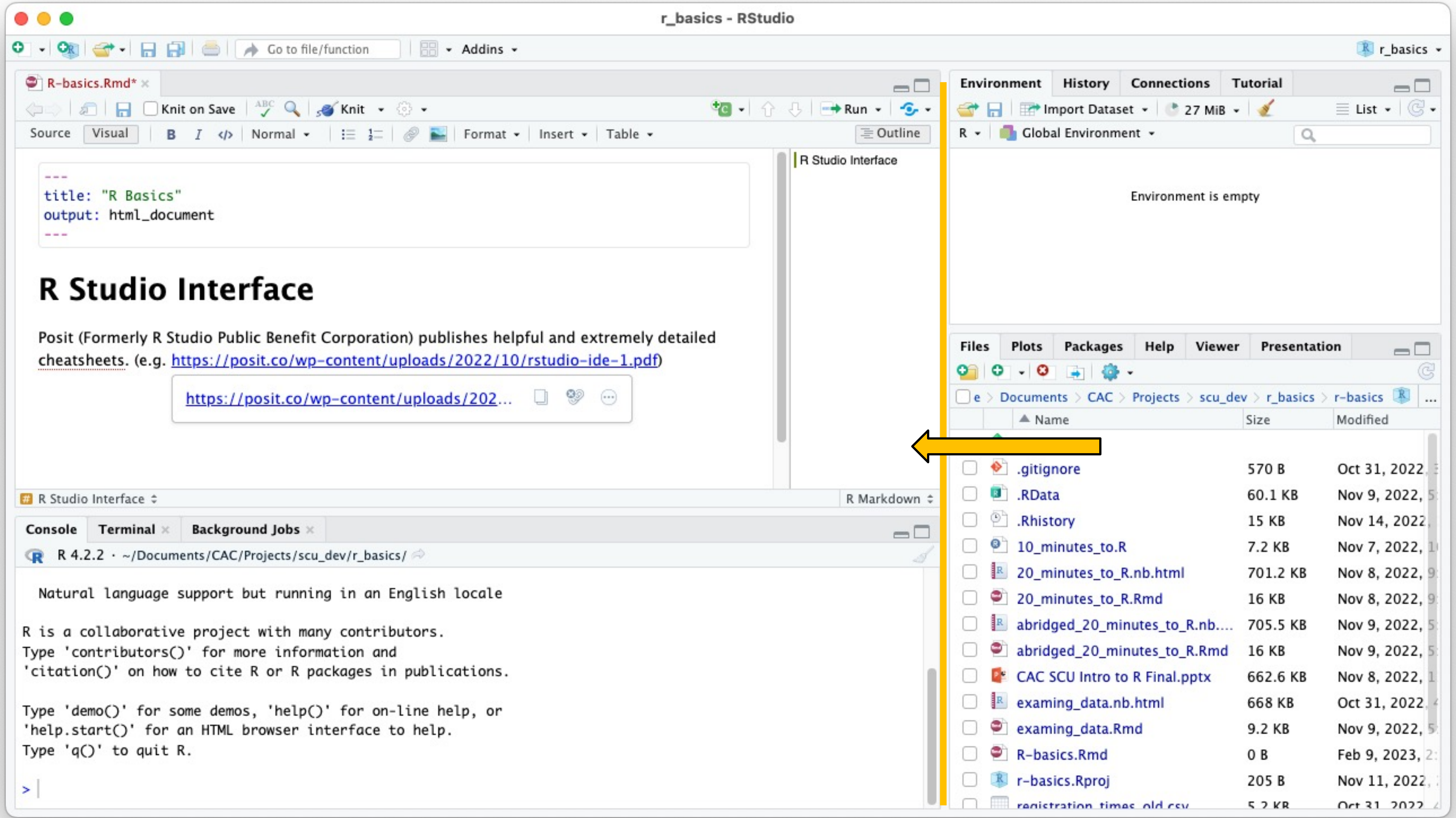
R Studio Interface

Posit (Formerly R Studio Public Benefit Corporation) publishes helpful and extremely detailed cheatsheets. (e.g. https://posit.co/wp-content/uploads/2022/10/rstudio-ide-1.pdf)

```
 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.

> 
```

```
..
.gitignore
.RData
.Rhistory
10_minutes_to.R
20_minutes_to_R.nb.html
20_minutes_to_R.Rmd
abridged_20_minutes_to_R.nb....
abridged_20_minutes_to_R.Rmd
CAC SCU Intro to R Final.pptx
examing_data.nb.html
examing_data.Rmd
R-basics.Rmd
r-basics.Rproj
registration_times_old.csv
```

RStudio Interface - Resized



```
---
title: "R Basics"
output: html_document
---
```

R Studio Interface

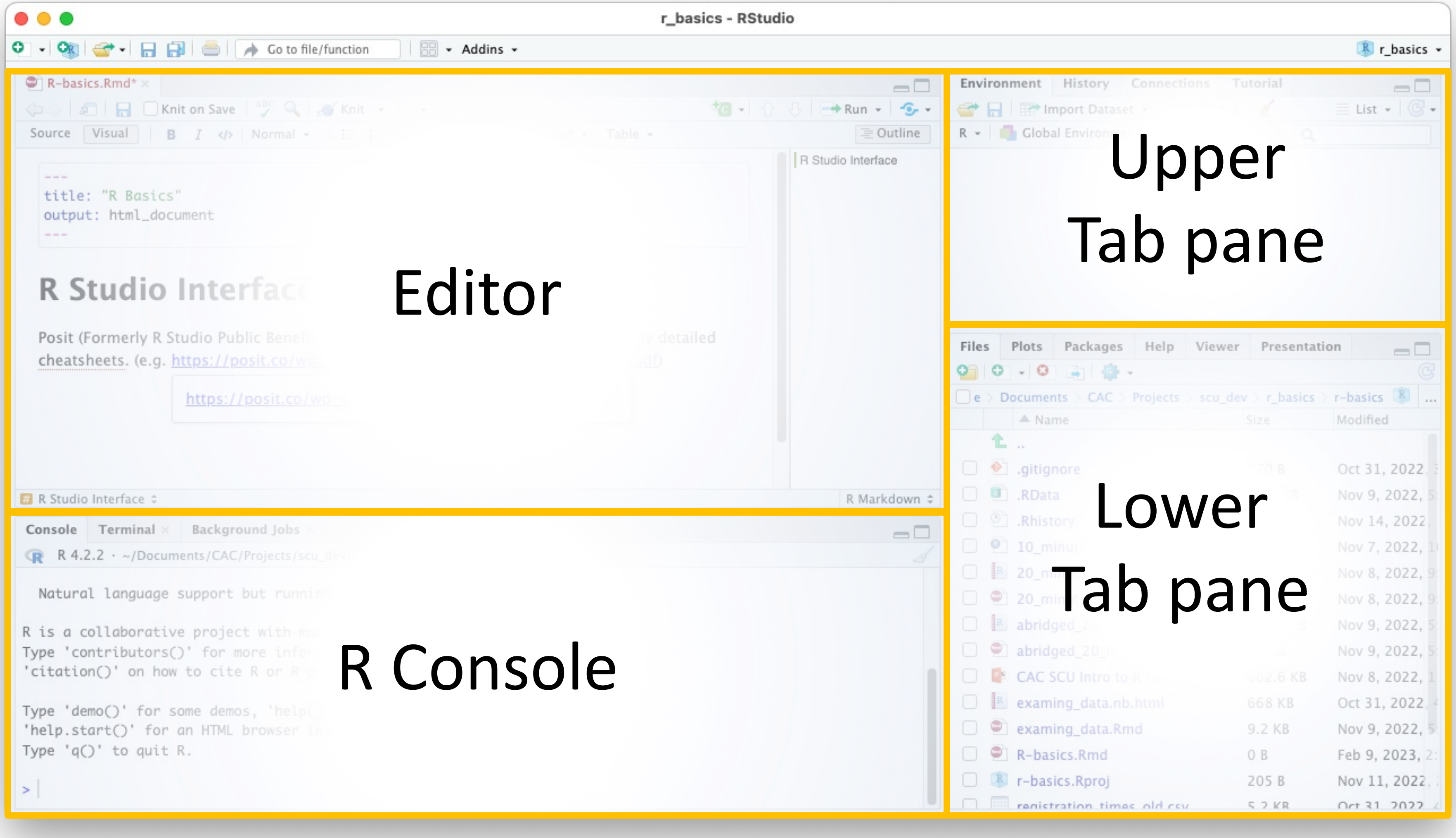
Posit (Formerly R Studio Public Benefit Corporation) publishes helpful and extremely detailed cheatsheets. (e.g. https://posit.co/wp-content/uploads/2022/10/rstudio-ide-1.pdf)

https://posit.co/wp-content/uploads/202...

```
R 4.2.2 · ~/Documents/CAC/Projects/scu_dev/r_basics/

 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.

>
```

Name	Size	Modified
.gitignore	570 B	Oct 31, 2022
.RData	60.1 KB	Nov 9, 2022,
.Rhistory	15 KB	Nov 14, 2022
10_minutes_to.R	7.2 KB	Nov 7, 2022,
20_minutes_to_R.nb.html	701.2 KB	Nov 8, 2022,
20_minutes_to_R.Rmd	16 KB	Nov 8, 2022,
abridged_20_minutes_to_R.nb....	705.5 KB	Nov 9, 2022,
abridged_20_minutes_to_R.Rmd	16 KB	Nov 9, 2022,
CAC SCU Intro to R Final.pptx	662.6 KB	Nov 8, 2022,
examing_data.nb.html	668 KB	Oct 31, 2022
examing_data.Rmd	9.2 KB	Nov 9, 2022,
R-basics.Rmd	0 B	Feb 9, 2023,
r-basics.Rproj	205 B	Nov 11, 2022
registration_times_old.csv	5.2 KB	Oct 31, 2022

RStudio Interface



Editor

```
---
title: "R Basics"
output: html_document
---
```

R Studio Interfac

Posit (Formerly R Studio Public Benef ... y detailed cheatsheets. (e.g. https://posit.co/w ... df)

https://posit.co/w

Upper Tab pane

R Console

R 4.2.2 · ~/Documents/CAC/Projects/scu_dev

```
Natural language support but runnin

R is a collaborative project with ma
Type 'contributors()' for more info
'citation()' on how to cite R or R

Type 'demo()' for some demos, 'help(
'help.start()' for an HTML browser i
Type 'q()' to quit R.

>
```

Lower Tab pane

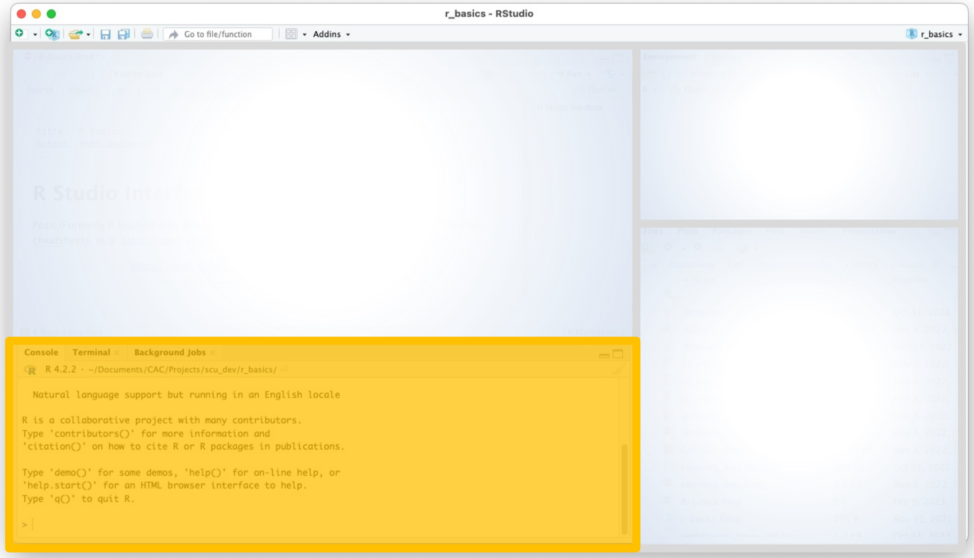
Documents › CAC › Projects › scu_dev › r_basics › r-basics

Name	Size	Modified
..		
.gitignore		Oct 31, 2022
.RData		Nov 9, 2022
.Rhistory		Nov 14, 2022
10_minut		Nov 7, 2022
20_m		Nov 8, 2022
20_min		Nov 8, 2022
abridged_2		Nov 9, 2022
abridged_20		Nov 9, 2022
CAC SCU Intro to R	682.6 KB	Nov 8, 2022
examing_data.nb.html	668 KB	Oct 31, 2022
examing_data.Rmd	9.2 KB	Nov 9, 2022
R-basics.Rmd	0 B	Feb 9, 2023
r-basics.Rproj	205 B	Nov 11, 2022
registration_times_old.csv	5.2 KB	Oct 31, 2022

R Console

- Type and run commands
- View output
- Up/Down arrows to navigate command history

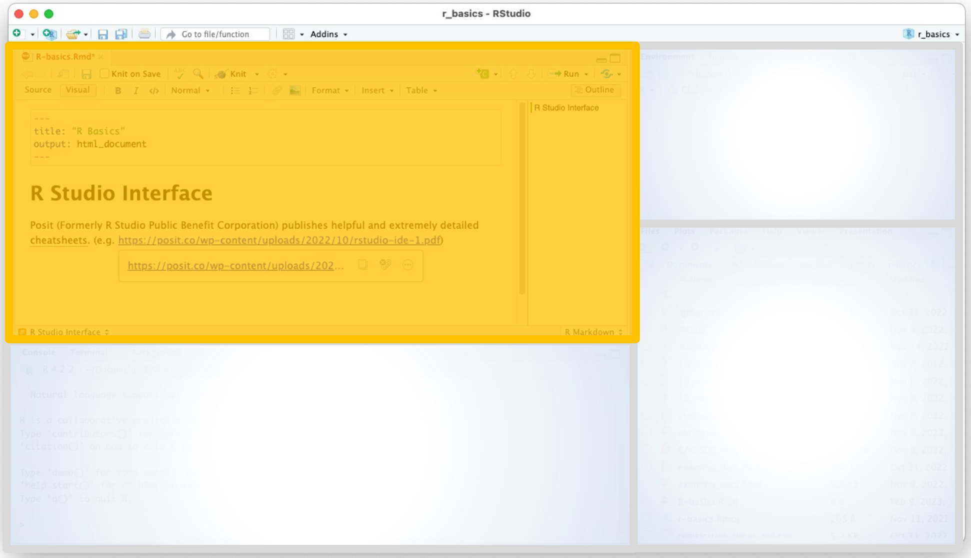
- Good for:
  - Exploratory commands
  - Testing commands
- Bad for:
  - Documenting analysis
  - Repeatable analysis



```
 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 Editor

- Write scripts. (.R)
- Create Rmarkdown documents (.Rmd)
- Limited data viewer
- Good for:
  - Documenting analysis
  - Repeatable analysis
- For most people, writing the commands in the editor is the best workflow
  - Execute script commands in Console
  - Execute commands within Rmarkdown documents



```
---
title: "R Basics"
output: html_document
---
```

R Studio Interface

Posit (Formerly R Studio Public Benefit Corporation) publishes helpful and extremely detailed cheatsheets. (e.g. https://posit.co/wp-content/uploads/2022/10/rstudio-ide-1.pdf)

https://posit.co/wp-content/uploads/202...

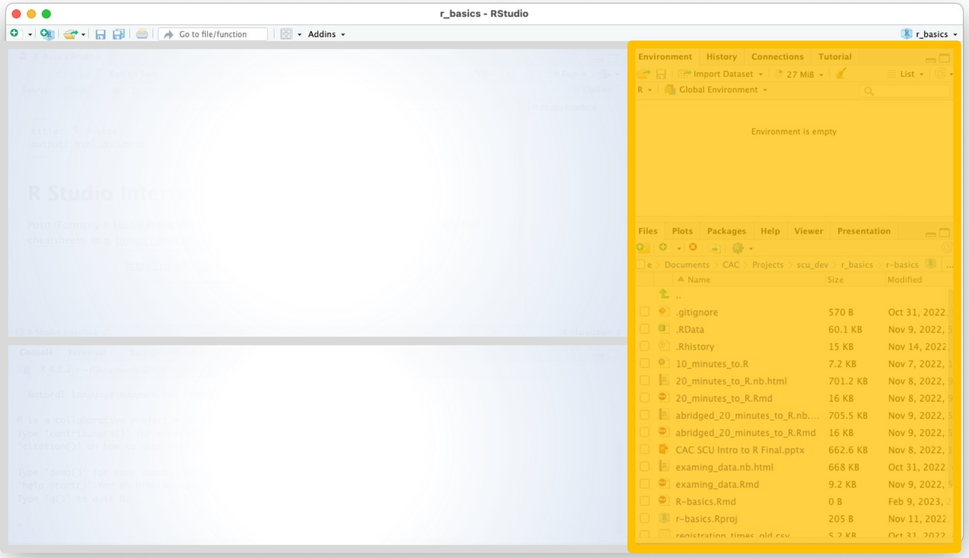
RStudio Tab Panes

Upper:

- Workspace / Environment Viewer
- Data import wizard
- Command history
- Tutorials

Lower:

- File browser
- Package Manager
- Help
- Plots



Environment is empty

Name	Size	Modified
..		
.gitignore	570 B	Oct 31, 2022
.RData	60.1 KB	Nov 9, 2022,
.Rhistory	15 KB	Nov 14, 2022
10_minutes_to.R	7.2 KB	Nov 7, 2022,
20_minutes_to_R.nb.html	701.2 KB	Nov 8, 2022,
20_minutes_to_R.Rmd	16 KB	Nov 8, 2022,
abridged_20_minutes_to_R.nb....	705.5 KB	Nov 9, 2022,
abridged_20_minutes_to_R.Rmd	16 KB	Nov 9, 2022,
CAC SCU Intro to R Final.pptx	662.6 KB	Nov 8, 2022,
examing_data.nb.html	668 KB	Oct 31, 2022
examing_data.Rmd	9.2 KB	Nov 9, 2022,
R-basics.Rmd	0 B	Feb 9, 2023,
r-basics.Rproj	205 B	Nov 11, 2022
registration_times_old.csv	5.2 KB	Oct 31, 2022

Packages

- Many packages from many disciplines
  → many ways to accomplish most tasks

- No need to be an R “purist”
  - Packages were created to streamline common operations

Tidyverse https://tidyverse.tidyverse.org

- `install.packages("tidyverse”, dependencies=TRUE)`
- `library(tidyverse)` will load the core tidyverse packages:
  - **ggplot2**, for data visualization
  - **Dplyr**, for data manipulation, merging
  - **Tidyr**, for data tidying
  - **Readr**, for data import
  - **Purrr**, for functional programming
  - **Tibble**, a modern replacement for data frames
  - **stringr**, for strings
  - **forcats**, for factors

Follow along online

- RStudio Cloud signup:
  - https://login.posit.cloud/register

- Materials from today in posit Cloud:
  - https://posit.cloud/content/5417403

More information

- Cornell Virtual Workshop in R: https://cvw.cac.cornell.edu/R/
  - CVW offers free self-paced, text-based modules covering a variety of computational focused topics. The CVW R topic complements today's workshop and covers using R on multiple cores and on supercomputer infrastructure.

- RStudio Cheatsheets:
  - https://www.rstudio.com/resources/cheatsheets/
  - Thoughtfully designed, single-page, double-sided reference sheets for major R packages.

More information

- Using R for teaching and research:
  - https://www.chrisbail.net/teaching
  - Chris Bail's work is a good example of incorporating R into teaching and research at undergraduate and graduate levels. Dr. Bail uses R for most aspects of his data collection and analysis.

- eBooks:
  - R for Data Science, Hadley Wickam and Garrett Grolemund - https://r4ds.had.co.nz
  - Advanced R (Programming), Hadley Wickam - https://adv-r.hadley.nz
  - R for Epidemiology - https://www.r4epi.com

More information

- Installing R for Jupyter Notebooks:
  - If you already use Jupyter, you can install the R jupyter kernel to use R in the familiar notebook environment. If you are on macOS, read the yellow warning box on the linked page. https://irkernel.github.io/installation/

- R packages on CRAN by area:
  - https://cran.r-project.org/web/views/

Thank you for attending!

- Please complete the survey

- Links to the recording and slide deck will be emailed to registrants, usually within a week

- Additional topics: https://its.weill.cornell.edu/scientific-computing-training-series