INTRODUCTION TO R

Danny Lumian, Ph.D.

Data Science Training Specialist

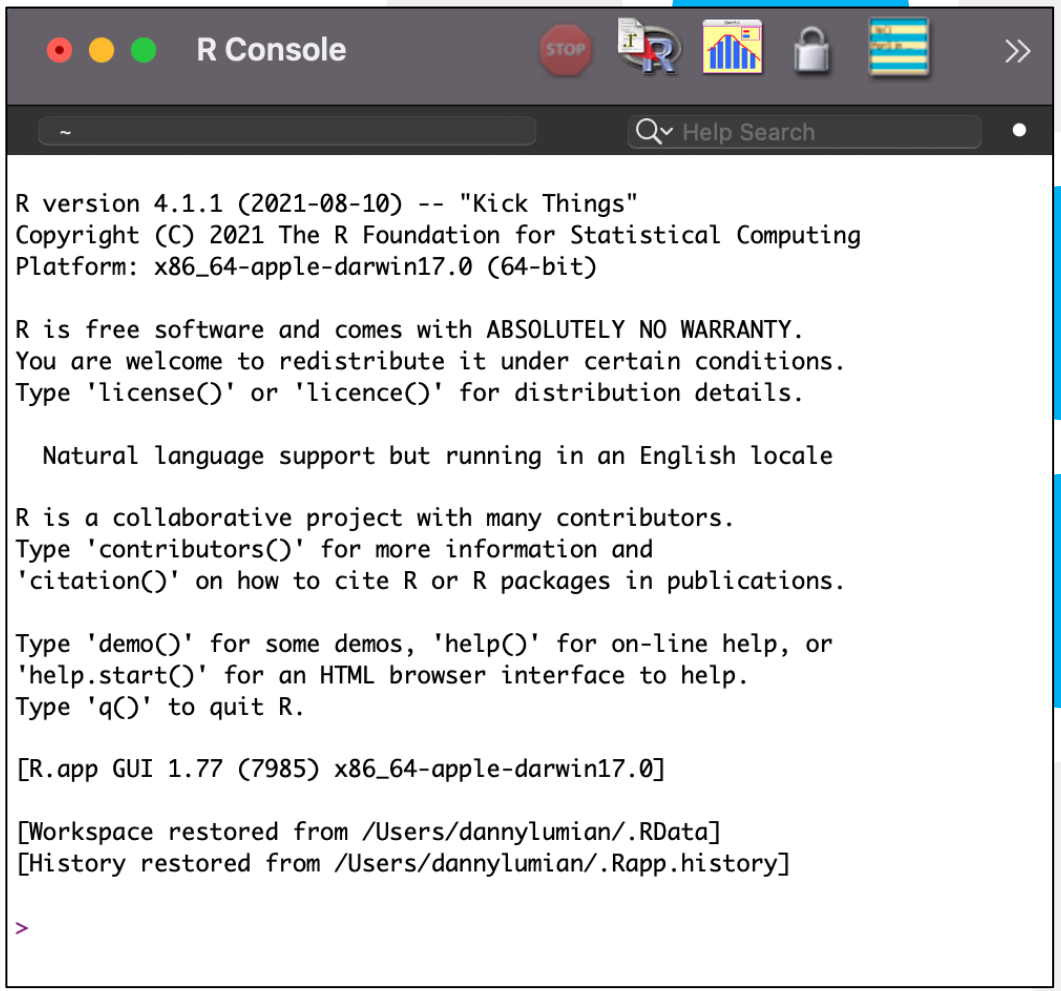
NIH Office of Data Science Strategy

Topics Covered

- R platforms
  - R
  - R Studio
  - R Studio Cloud
- R file types
  - R script
  - R markdown
  - R notebook
- Basic R commands
- R packages and built-in datasets
- Data manipulation
- Data visualization
- Data import and export
- Data analysis
- Additional Analysis Notebooks

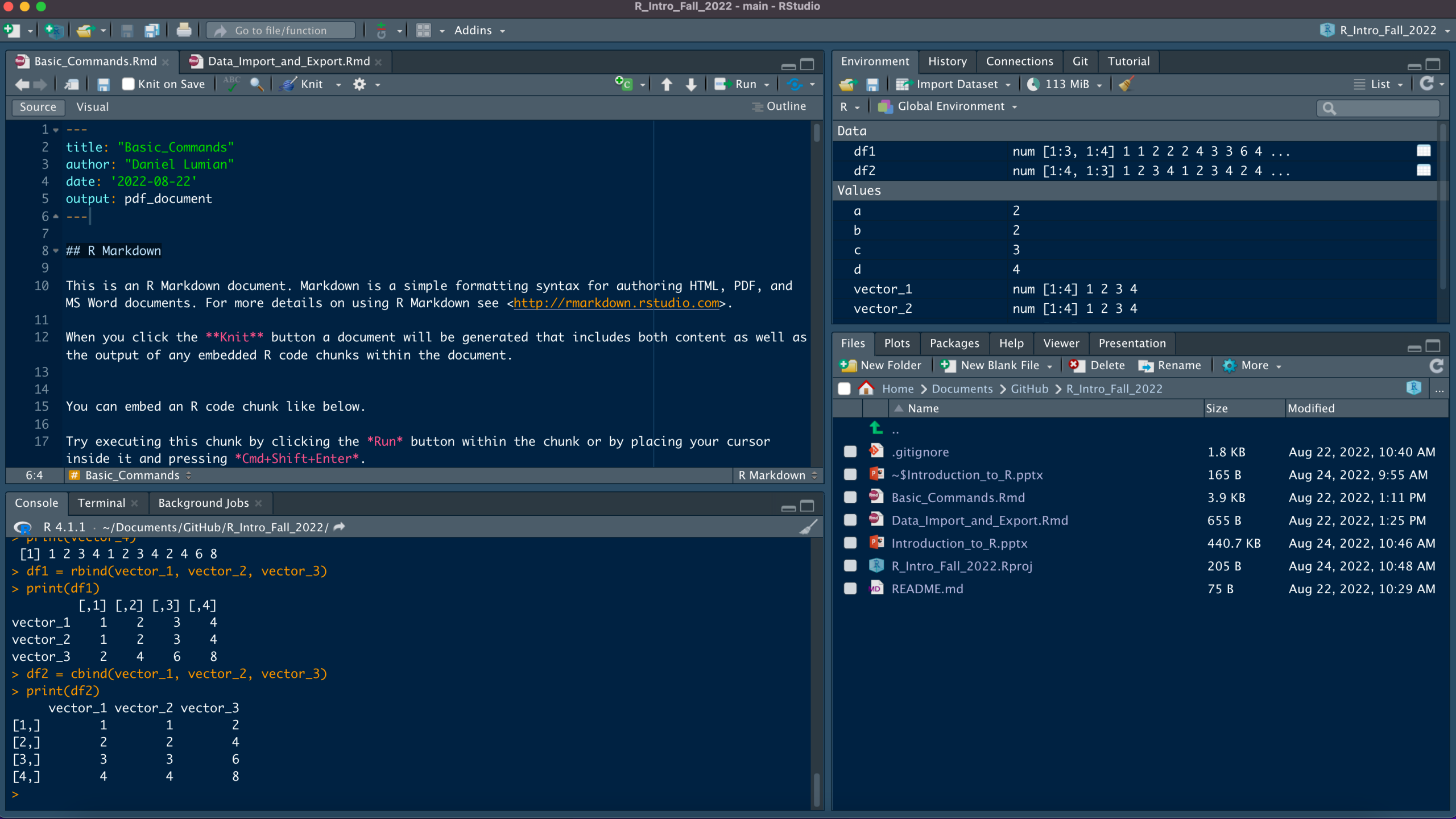
What is R?

- A programming language and open-source software environment that can
  - Manipulate data
  - Visualize data
  - Perform statistics
- Free and relatively easy to run in any environment
- https://www.r-project.org/



What is R Studio

- An integrated development environment (IDE) for R
- Software that includes R, file explorer, graphics viewer, code notebook and more!
- One stop shop for coding in R
- https://rstudio.com/


R_Intro_Fall_2022 - main - RStudio
Go to file/function
Addins
R_Intro_Fall_2022
Basic_Commands.Rmd
Data_Import_and_Export.Rmd
Knit on Save
Knit
Run
Source
Visual
Outline
1 ---
2 title: "Basic_Commands"
3 author: "Daniel Lumian"
4 date: '2022-08-22'
5 output: pdf_document
6 ---
7
8 ## R Markdown
9
10 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>.
11
12 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.
13
14
15 You can embed an R code chunk like below.
16
17 Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Cmd+Shift+Enter*.
6:4
Basic_Commands
R Markdown
Console
Terminal
Background Jobs
R 4.1.1 · ~/Documents/GitHub/R_Intro_Fall_2022/
[1] 1 2 3 4 1 2 3 4 2 4 6 8
> df1 = rbind(vector_1, vector_2, vector_3)
> print(df1)
[,1] [,2] [,3] [,4]
vector_1 1 2 3 4
vector_2 1 2 3 4
vector_3 2 4 6 8
> df2 = cbind(vector_1, vector_2, vector_3)
> print(df2)
vector_1 vector_2 vector_3
[1,] 1 1 2
[2,] 2 2 4
[3,] 3 3 6
[4,] 4 4 8
>
Environment
History
Connections
Git
Tutorial
Import Dataset
113 MiB
List
R
Global Environment
Data
df1 num [1:3, 1:4] 1 1 2 2 2 4 3 3 6 4 ...
df2 num [1:4, 1:3] 1 2 3 4 1 2 3 4 2 4 ...
Values
a 2
b 2
c 3
d 4
vector_1 num [1:4] 1 2 3 4
vector_2 num [1:4] 1 2 3 4
Files
Plots
Packages
Help
Viewer
Presentation
New Folder
New Blank File
Delete
Rename
More
Home > Documents > GitHub > R_Intro_Fall_2022
Name
Size
Modified
..
.gitignore 1.8 KB Aug 22, 2022, 10:40 AM
~$Introduction_to_R.pptx 165 B Aug 24, 2022, 9:55 AM
Basic_Commands.Rmd 3.9 KB Aug 22, 2022, 1:11 PM
Data_Import_and_Export.Rmd 655 B Aug 22, 2022, 1:25 PM
Introduction_to_R.pptx 440.7 KB Aug 24, 2022, 10:46 AM
R_Intro_Fall_2022.Rproj 205 B Aug 24, 2022, 10:48 AM
README.md 75 B Aug 22, 2022, 10:29 AM

What is R Studio Cloud

- Cloud based version of R Studio
- No need to download software
- Can share projects easily for analysis and teaching
- https://rstudio.cloud/

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | R 4.2.1

Basic_Commands.Rmd

Knit | Run | Source | Visual | Outline

Package rmarkdown required but is not installed. Install Don't Show Again

```
1  ---
2  title: "Basic_Commands"
3  author: "Daniel Lumian"
4  date: '2022-08-22'
5  output: pdf_document
6  ---
7
8  ## R Markdown
9
10 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>.
11
12 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.
13
14
15 You can embed an R code chunk like below.
16
17 Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing
   *Cmd+Shift+Enter*.
```

1:1 Basic_Commands R Markdown

Console Terminal Background Jobs

R 4.2.1 · /cloud/project/

```
R version 4.2.1 (2022-06-23) -- "Funny-Looking Kid"
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (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.

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 Git Tutorial

Import Dataset | 135 MiB | List

R | Global Environment

Environment is empty

Files Plots Packages Help Viewer Presentation

New Folder | New Blank File | Upload | Delete | Rename | More

Cloud > project

Name	Size	Modified
..		
.gitignore	1.8 KB	Aug 25, 2022, 8:33 AM
.Rhistory	0 B	Aug 25, 2022, 8:33 AM
Basic_Commands.Rmd	3.9 KB	Aug 25, 2022, 8:33 AM
Data_Import_and_Export.Rmd	655 B	Aug 25, 2022, 8:33 AM
Introduction_to_R.pptx	1.3 MB	Aug 25, 2022, 8:33 AM
R_Intro_Fall_2022.Rproj	205 B	Aug 25, 2022, 8:33 AM
README.md	75 B	Aug 25, 2022, 8:33 AM

R File Types

R Script File

- Basic file for saving code
- Can add comments using `#` symbol
- Useful for testing code snippets and work that will not be shared
- Sections can be selected and run

R Markdown and Notebook

- More advanced files that allow inclusion of markdown and style elements for documentation
- Code lives in chunks within the file
- Better for sharing and publishing code
- Notebooks allow a preview feature, while Markdown files are knit

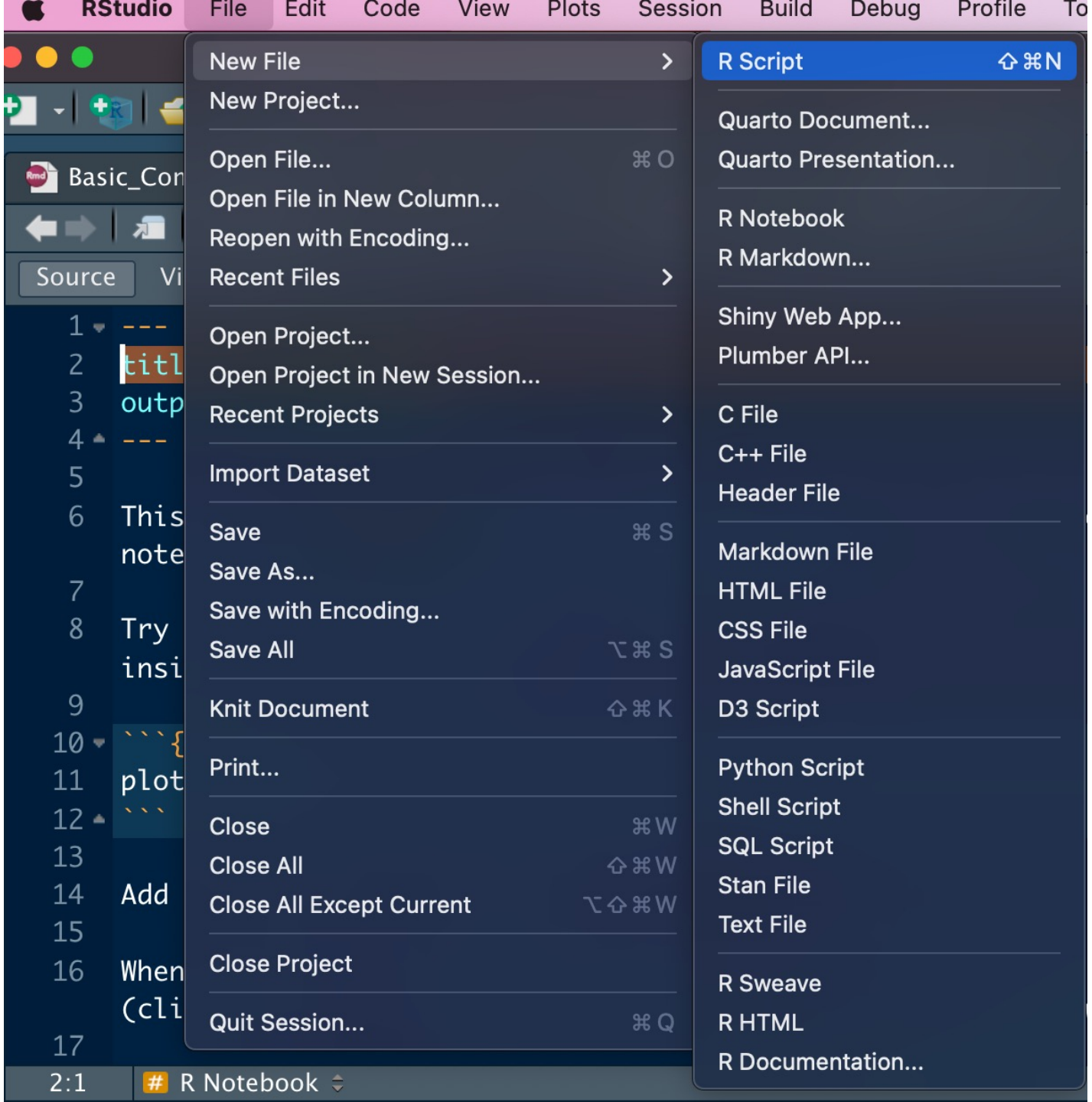
---

https://stackoverflow.com/questions/43820483/difference-between-r-markdown-and-r-notebook

- To open a new file, select File from the top panel
- Select new file from the dropdown
- Select appropriate file type
  - R script
  - R Notebook
  - R Markdown

- To open an existing file, select File from the top panel
- Select Open File...

- To save a file, select File from the top panel
- Select Save or Save as... and specify the save location



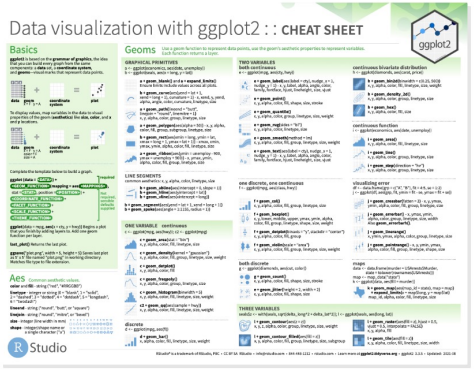
R Studio Cheat sheets:

https://www.rstudio.com/resources/cheatsheets/

Data visualization with ggplot2 cheatsheet

The ggplot2 package lets you make beautiful and customizable plots of your data. It implements the grammar of graphics, an easy to use system for building plots. Updated August 2021.

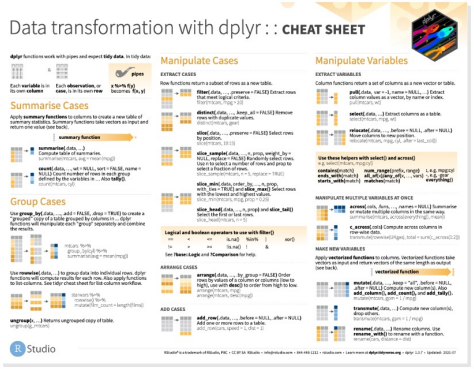
DOWNLOAD



Data transformation with dplyr cheatsheet

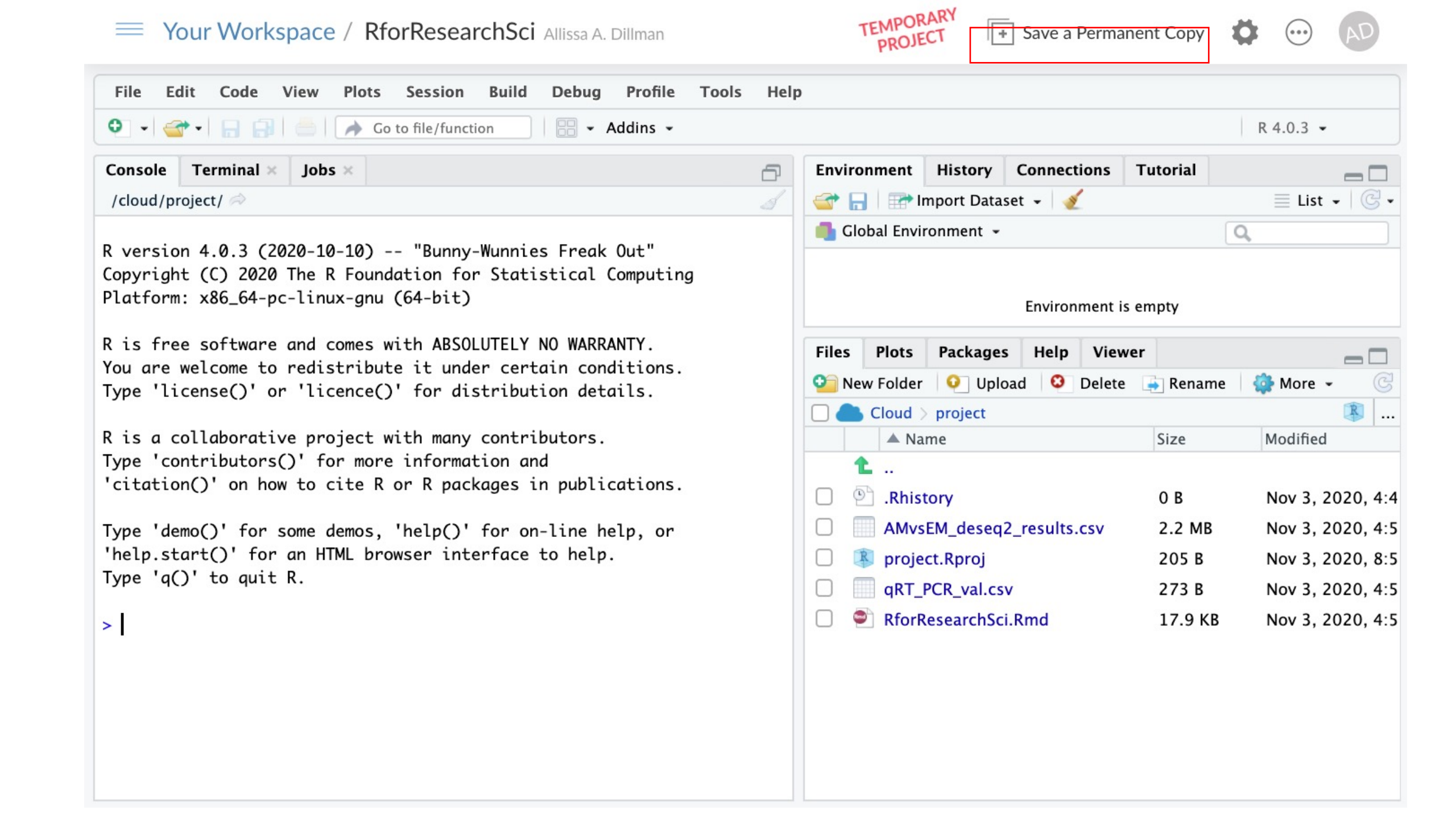
The dplyr package provides a grammar for manipulating tables in R. This cheatsheet will guide you through the grammar, reminding you how to select, filter, arrange, mutate, summarise, group, and join data frames and tibbles. Updated July 2021.

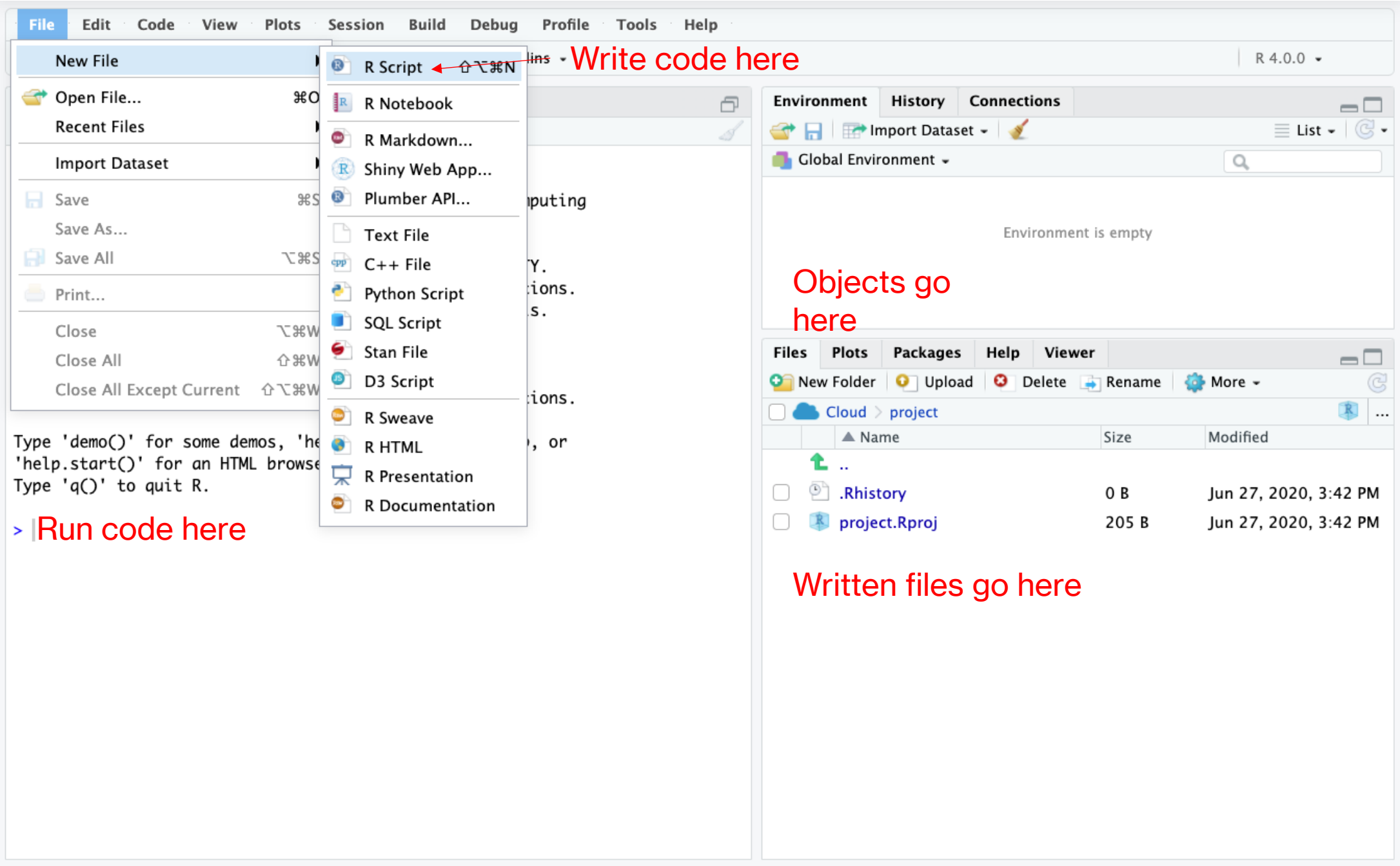
DOWNLOAD



Let’s get started!

- You may need to log in to your RStudio cloud account first
  - Go to https://rstudio.cloud/
  - Click log in
- [Need to add r cloud studio link when ready]
  - First click on Save a permanent copy
- https://github.com/dlumian/R_Intro_Fall_2022
- Click on the new file icon
- Select R script


Your Workspace / RforResearchSci Allissa A. Dillman
TEMPORARY PROJECT
Save a Permanent Copy
AD
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 4.0.3
Console Terminal Jobs
/cloud/project/
R version 4.0.3 (2020-10-10) -- "Bunny-Wunnies Freak Out"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (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.
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
Import Dataset
List
Global Environment
Environment is empty
Files Plots Packages Help Viewer
New Folder
Upload
Delete
Rename
More
Cloud project
Name
Size
Modified
..
.Rhistory 0 B Nov 3, 2020, 4:4
AMvsEM_deseq2_results.csv 2.2 MB Nov 3, 2020, 4:5
project.Rproj 205 B Nov 3, 2020, 8:5
qRT_PCR_val.csv 273 B Nov 3, 2020, 4:5
RforResearchSci.Rmd 17.9 KB Nov 3, 2020, 4:5


File
Edit
Code
View
Plots
Session
Build
Debug
Profile
Tools
Help
New File
Open File...
Recent Files
Import Dataset
Save
Save As...
Save All
Print...
Close
Close All
Close All Except Current
R Script
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 Presentation
R Documentation
Write code here
R 4.0.0
Type 'demo()' for some demos, 'he
'help.start()' for an HTML browse
Type 'q()' to quit R.
> Run code here
Environment
History
Connections
Import Dataset
List
Global Environment
Environment is empty
Objects go here
Files
Plots
Packages
Help
Viewer
New Folder
Upload
Delete
Rename
More
Cloud
project
Name
Size
Modified
..
.Rhistory
0 B
Jun 27, 2020, 3:42 PM
project.Rproj
205 B
Jun 27, 2020, 3:42 PM
Written files go here

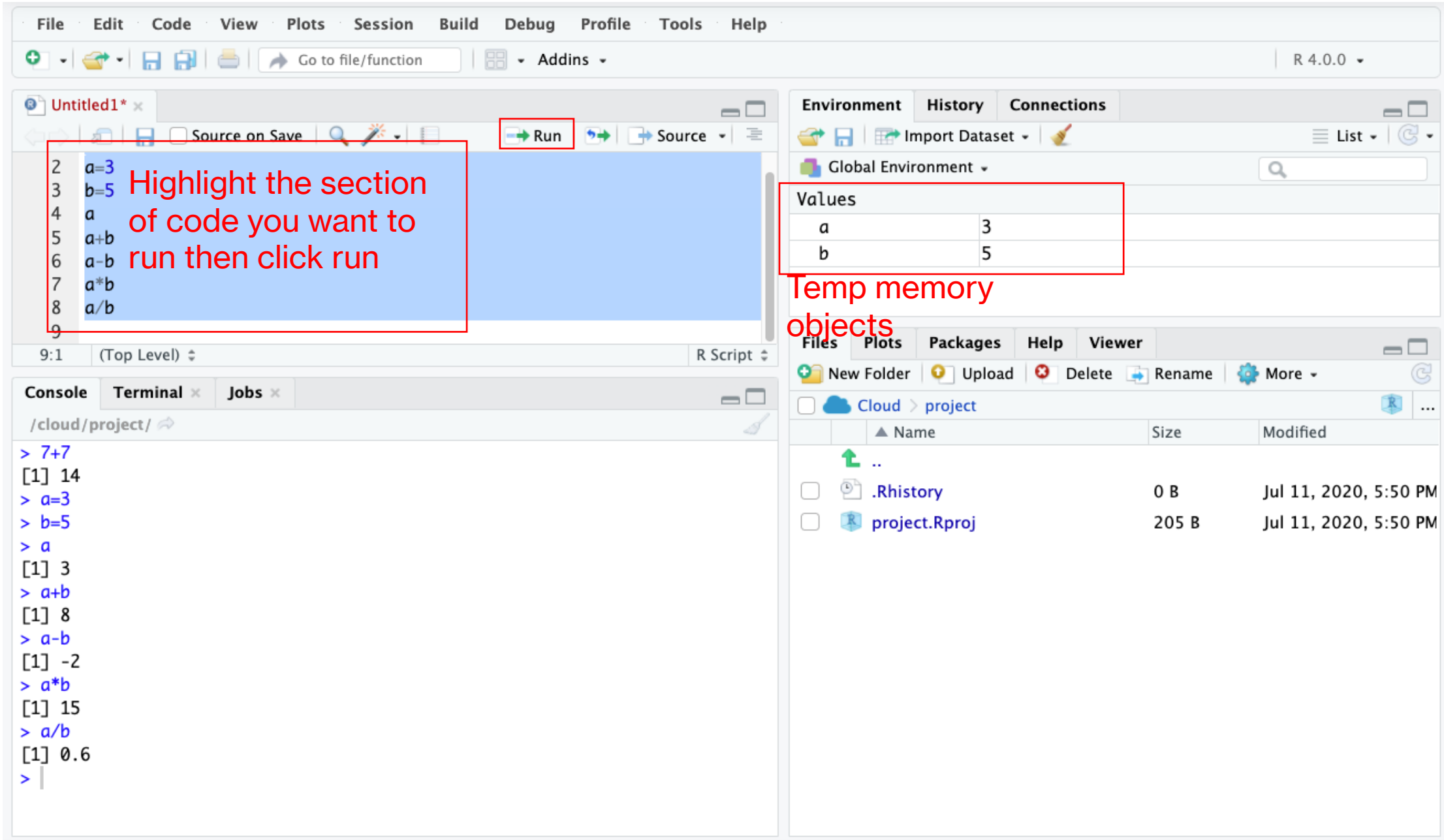
R is like a calculator

- Enter the following

```
7+7
a=3
b=5
```

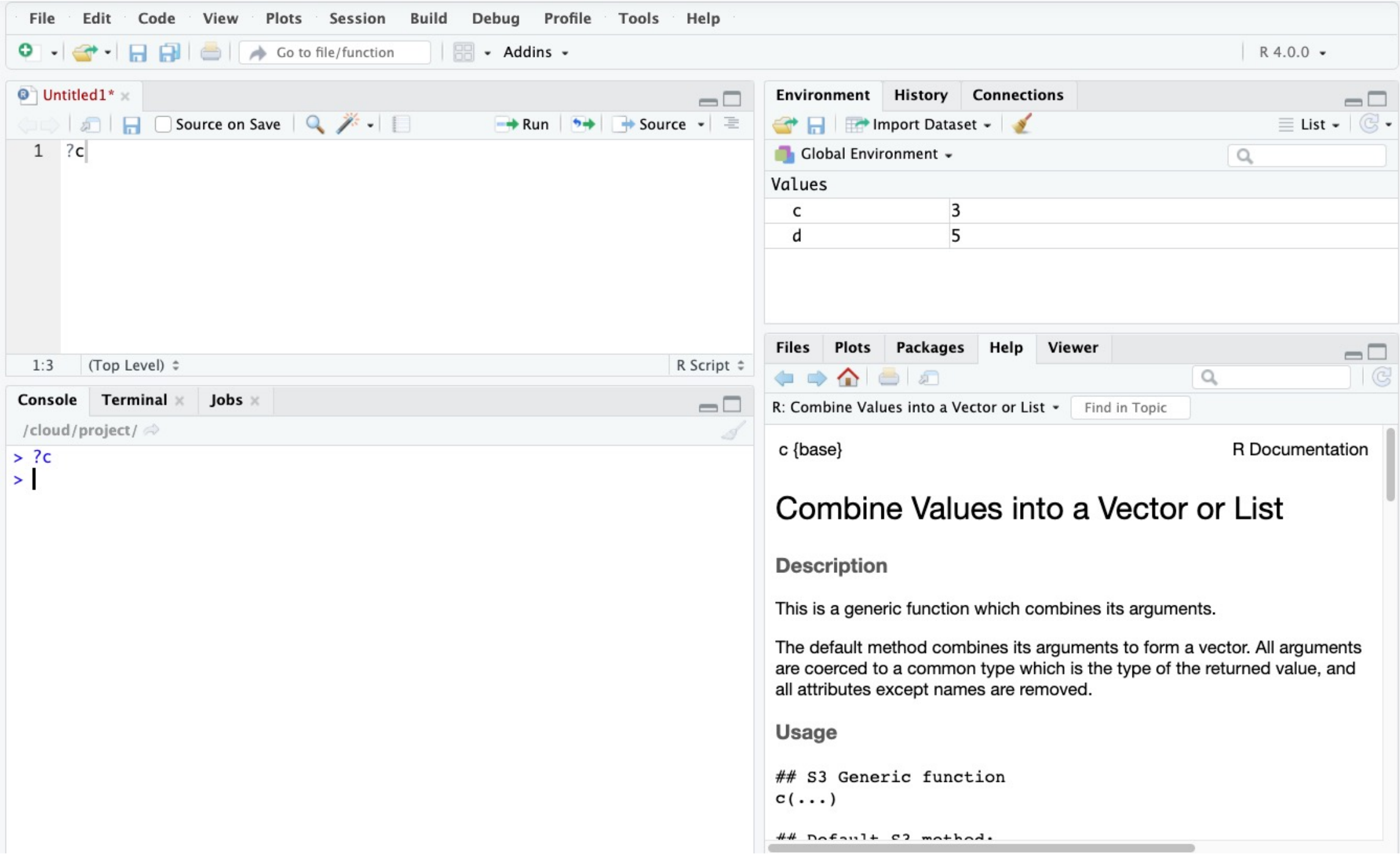
- If I type a it will print out the contents of a which is 3

```
a+b
a-b
a*b
a/b
```


File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 4.0.0
Untitled1*
Source on Save
Run
Source
a=3
b=5
a
a+b
a-b
a*b
a/b
Highlight the section of code you want to run then click run
9:1
(Top Level)
R Script
Console
Terminal
Jobs
/cloud/project/
> 7+7
[1] 14
> a=3
> b=5
> a
[1] 3
> a+b
[1] 8
> a-b
[1] -2
> a*b
[1] 15
> a/b
[1] 0.6
>
Environment
History
Connections
Import Dataset
List
Global Environment
Values
a 3
b 5
Temp memory objects
Files
Plots
Packages
Help
Viewer
New Folder
Upload
Delete
Rename
More
Cloud
project
Name
Size
Modified
..
.Rhistory 0 B Jul 11, 2020, 5:50 PM
project.Rproj 205 B Jul 11, 2020, 5:50 PM

Make a Vector

- Vectors are a data structure in R
  - -list of characters
  - -list of numbers
  - -must be same data type
- Use the c() command to enter a bunch of numbers together
  - c stands for combine
- Don't know how to use it?
- Type: ?c
  - Putting a `?` before a command or object will retrieve the help for that object

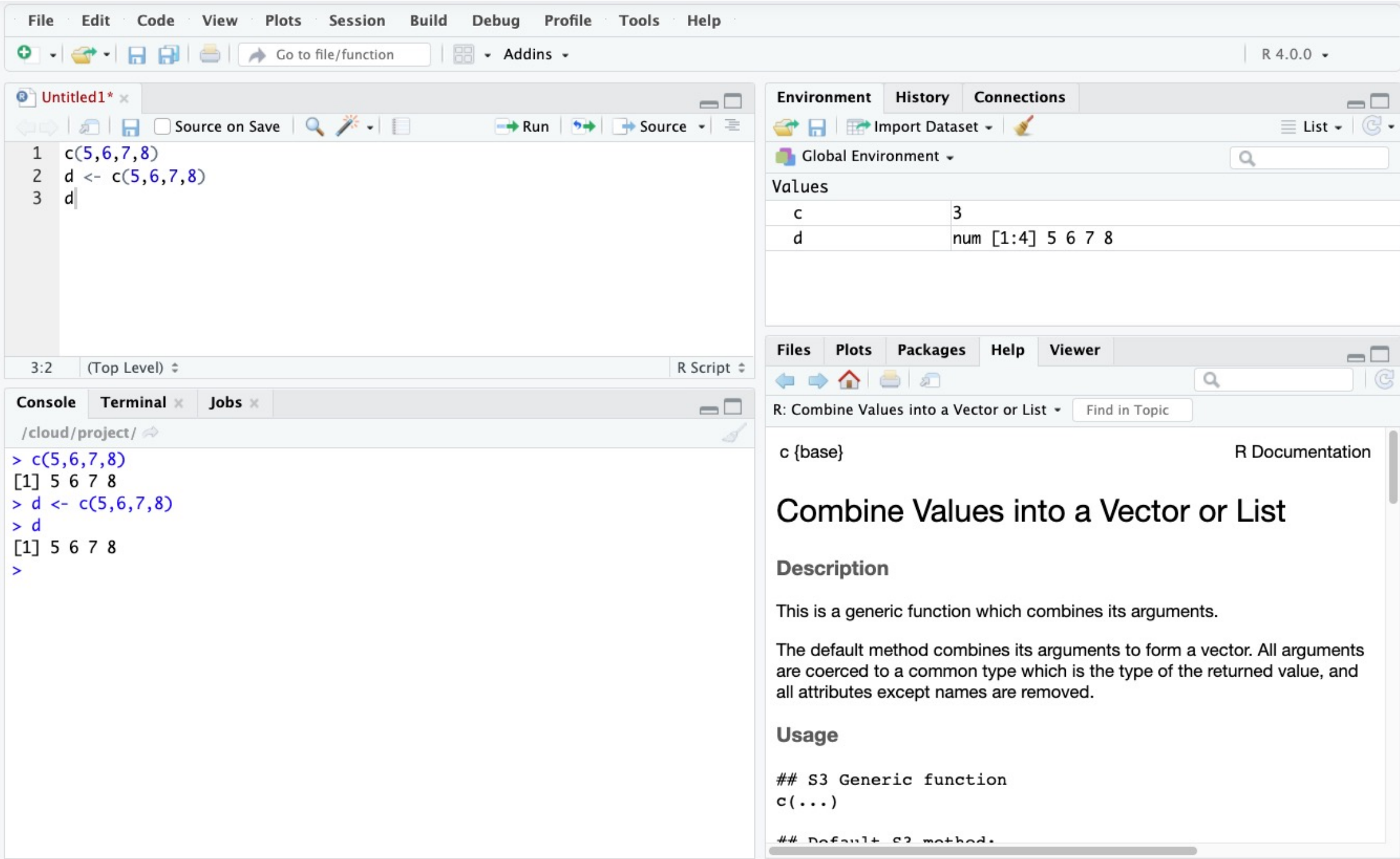

File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 4.0.0
Untitled1*
Source on Save
Run
Source
1 ?c
1:3
(Top Level)
R Script
Console
Terminal
Jobs
/cloud/project/
> ?c
>
Environment
History
Connections
Import Dataset
List
Global Environment
Values
c 3
d 5
Files
Plots
Packages
Help
Viewer
R: Combine Values into a Vector or List
Find in Topic
c {base}
R Documentation
Combine Values into a Vector or List
Description
This is a generic function which combines its arguments.
The default method combines its arguments to form a vector. All arguments are coerced to a common type which is the type of the returned value, and all attributes except names are removed.
Usage
S3 Generic function
c(...)

Typing ?

- Typing “?” then the command name will give you help on the command name
- Typically, examples you can try may be available at the bottom of the help information
- Now let's use the c() command
- c(5,6,7,8)
- This just prints the output to the screen

Create an object

- We can create an object for R to keep in its temporary memory
- We will assign a name to the vector we used before
- Type "d <-" in front of c(5,6,7,8) : d <- c(5,6,7,8)
- We created an R object called “d” that is a vector of 5,6,7,8
- Keyboard shortcut for <-
  - -PC: Alt and - at the same time
  - -Mac: option and - at the same time


File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 4.0.0
Untitled1*
Source on Save
Run
Source
1 c(5,6,7,8)
2 d <- c(5,6,7,8)
3 d
3:2
(Top Level)
R Script
Console
Terminal
Jobs
/cloud/project/
> c(5,6,7,8)
[1] 5 6 7 8
> d <- c(5,6,7,8)
> d
[1] 5 6 7 8
>
Environment
History
Connections
Import Dataset
List
Global Environment
Values
c 3
d num [1:4] 5 6 7 8
Files
Plots
Packages
Help
Viewer
R: Combine Values into a Vector or List
Find in Topic
c {base}
R Documentation
Combine Values into a Vector or List
Description
This is a generic function which combines its arguments.
The default method combines its arguments to form a vector. All arguments are coerced to a common type which is the type of the returned value, and all attributes except names are removed.
Usage
S3 Generic function
c(...)

Make more vectors

- e <- c(11,12,13,15)
- f <- c(1,2,3,4)
- g <- c(1,2,3,15)

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | R 4.0.0

Untitled1*

Source on Save | Run | Source

```
1  e <- c(11,12,13,15)
2  f <- c(1,2,3,4)
3  g <- c(1,2,3,15)
4  e
5  f
6  g
7
```

7:1 (Top Level) R Script

Console Terminal Jobs

/cloud/project/

```
> e <- c(11,12,13,15)
> f <- c(1,2,3,4)
> g <- c(1,2,3,15)
> e
[1] 11 12 13 15
> f
[1] 1 2 3 4
> g
[1]  1  2  3 15
>
```

Environment History Connections

Import Dataset | List

Global Environment

Values

c	3
d	num [1:4] 5 6 7 8
e	num [1:4] 11 12 13 15
f	num [1:4] 1 2 3 4
g	num [1:4] 1 2 3 15

Files Plots Packages Help Viewer

standard devia

R: Search Results | Find in Topic

Search Results

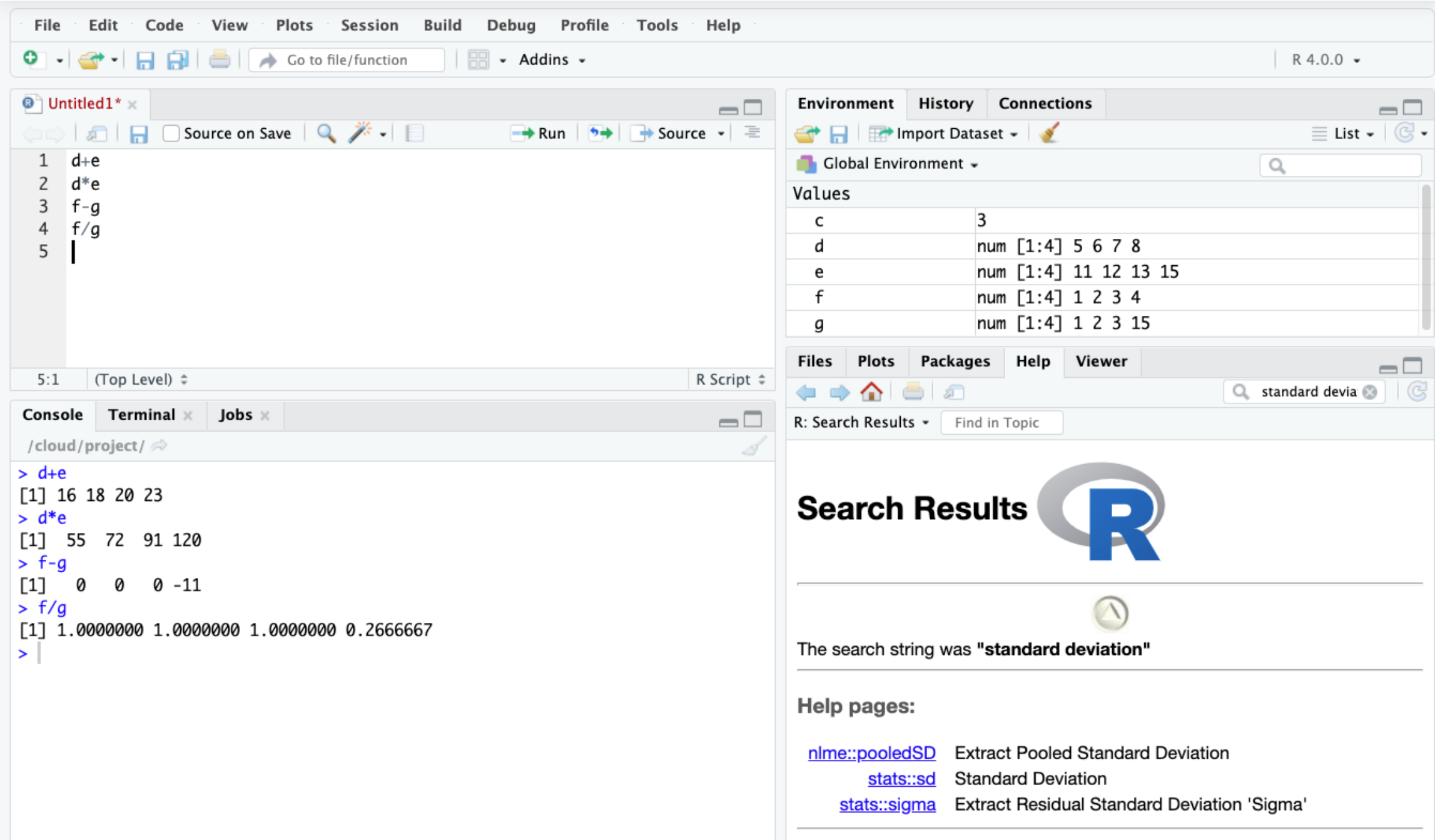
The search string was **"standard deviation"**

Help pages:

nlme::pooledSD	Extract Pooled Standard Deviation
stats::sd	Standard Deviation
stats::sigma	Extract Residual Standard Deviation 'Sigma'

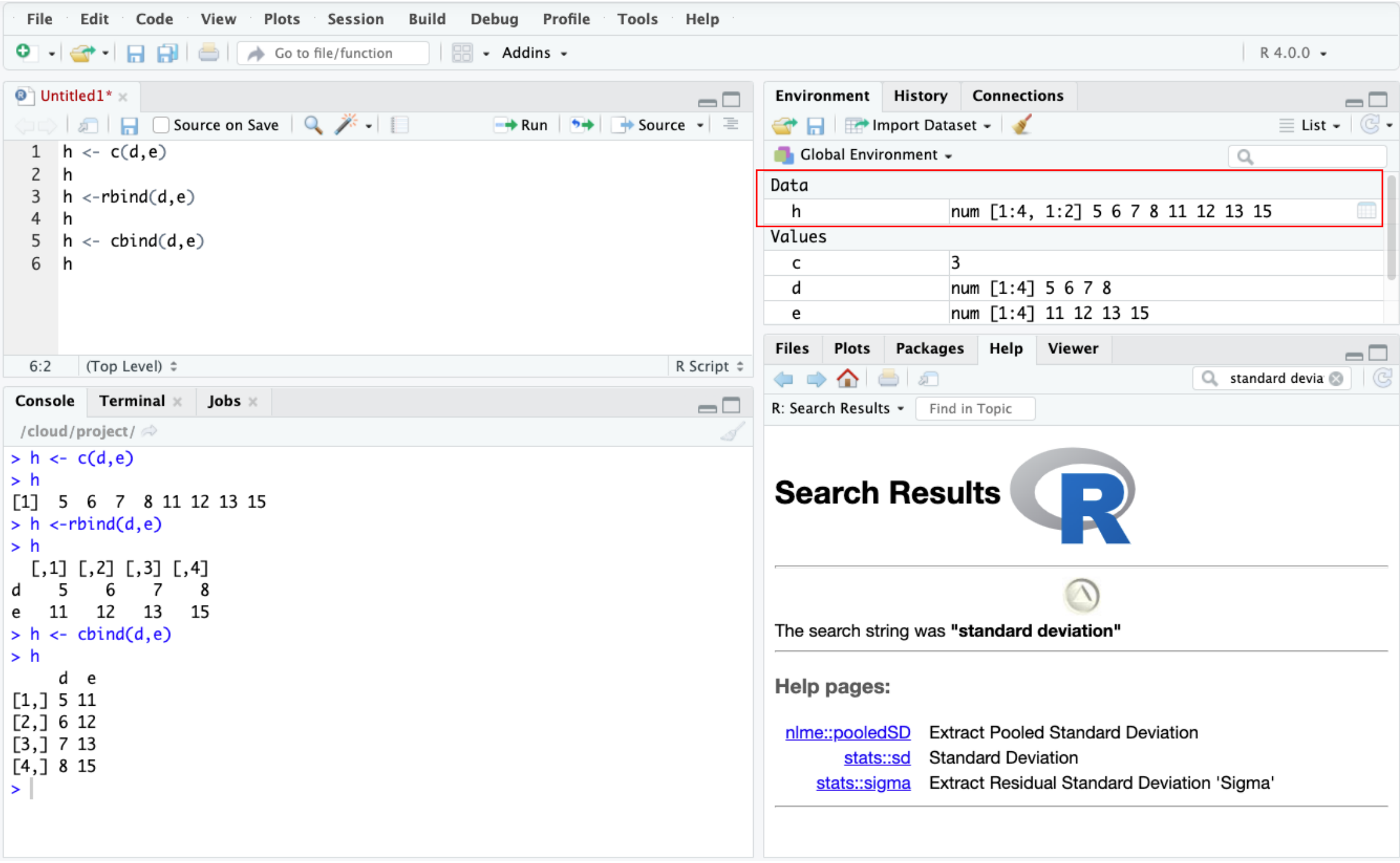
Perform Vector Calculations

- d+e
- d*e
- f-g
- f/g


File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 4.0.0
Untitled1*
Source on Save
Run
Source
1 d+e
2 d*e
3 f-g
4 f/g
5
5:1
(Top Level)
R Script
Console
Terminal
Jobs
/cloud/project/
> d+e
[1] 16 18 20 23
> d*e
[1] 55 72 91 120
> f-g
[1] 0 0 0 -11
> f/g
[1] 1.0000000 1.0000000 1.0000000 0.2666667
>
Environment
History
Connections
Import Dataset
List
Global Environment
Values
c 3
d num [1:4] 5 6 7 8
e num [1:4] 11 12 13 15
f num [1:4] 1 2 3 4
g num [1:4] 1 2 3 15
Files
Plots
Packages
Help
Viewer
standard devia
R: Search Results
Find in Topic
Search Results
The search string was "standard deviation"
Help pages:
nlme::pooledSD Extract Pooled Standard Deviation
stats::sd Standard Deviation
stats::sigma Extract Residual Standard Deviation 'Sigma'

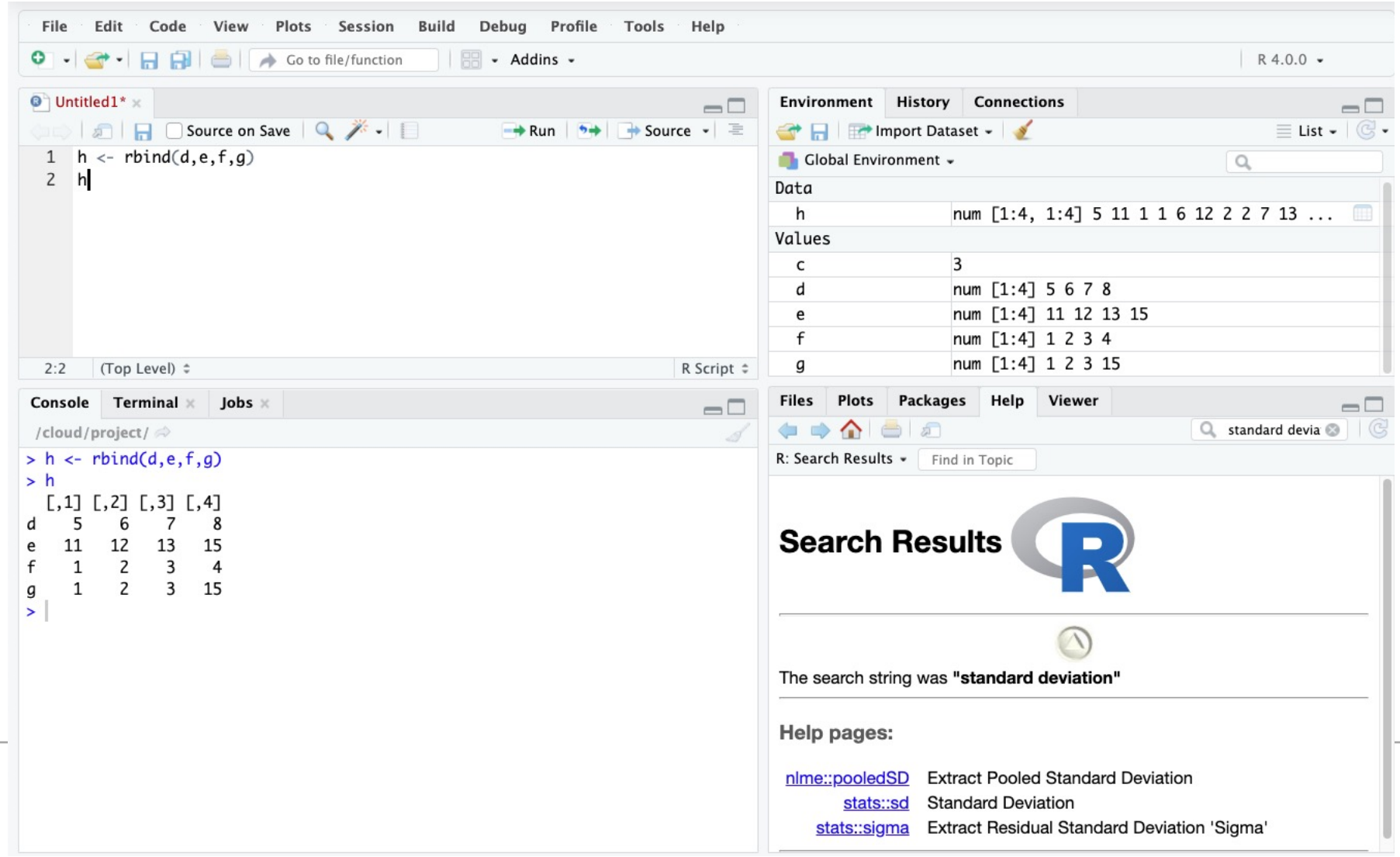
Combine Vectors

- h <- c(d,e)
- h <-rbind(d,e)
- h <- cbind(d,e)


File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 4.0.0
Untitled1*
Source on Save
Run
Source
1 h <- c(d,e)
2 h
3 h <-rbind(d,e)
4 h
5 h <- cbind(d,e)
6 h
6:2
(Top Level)
R Script
Console
Terminal
Jobs
/cloud/project/
> h <- c(d,e)
> h
[1] 5 6 7 8 11 12 13 15
> h <-rbind(d,e)
> h
[,1] [,2] [,3] [,4]
d 5 6 7 8
e 11 12 13 15
> h <- cbind(d,e)
> h
d e
[1,] 5 11
[2,] 6 12
[3,] 7 13
[4,] 8 15
>
Environment
History
Connections
Import Dataset
List
Global Environment
Data
h
num [1:4, 1:2] 5 6 7 8 11 12 13 15
Values
c
3
d
num [1:4] 5 6 7 8
e
num [1:4] 11 12 13 15
Files
Plots
Packages
Help
Viewer
standard devia
R: Search Results
Find in Topic
Search Results
The search string was "standard deviation"
Help pages:
nlme::pooledSD Extract Pooled Standard Deviation
stats::sd Standard Deviation
stats::sigma Extract Residual Standard Deviation 'Sigma'

Let's Make a Data Table!

h <- rbind(d,e,f,g)



Column and Row Names

- Independent of the data
- Makes it easier to work with data later
  - colnames
  - rownames
- Type the following:
  - colnames(h) <- c("Col1","Col2","Col3","Col4")
  - rownames(h) <- c("Row1","Row2","Row3","Row4")

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | R 4.0.0

Untitled1*

Source on Save | Run | Source

```
1  h
2  colnames(h) <- c("Col1","Col2","Col3","Col4")
3  h
4  rownames(h) <- c("Row1","Row2","Row3","Row4")
5  h
6
```

6:1 (Top Level) R Script

Console Terminal Jobs

/cloud/project/

```
  [,1] [,2] [,3] [,4]
d    5    6    7    8
e   11   12   13   15
f    1    2    3    4
g    1    2    3   15
> colnames(h) <- c("Col1","Col2","Col3","Col4")
> h
  Col1 Col2 Col3 Col4
d    5    6    7    8
e   11   12   13   15
f    1    2    3    4
g    1    2    3   15
> rownames(h) <- c("Row1","Row2","Row3","Row4")
> h
     Col1 Col2 Col3 Col4
Row1    5    6    7    8
Row2   11   12   13   15
Row3    1    2    3    4
Row4    1    2    3   15
>
```

Environment History Connections

Import Dataset | List

Global Environment

Data	
h	num [1:4, 1:4] 5 11 1 1 6 12 2 2 7 13 ...
Values	
c	3
d	num [1:4] 5 6 7 8
e	num [1:4] 11 12 13 15
f	num [1:4] 1 2 3 4
g	num [1:4] 1 2 3 15

Files Plots Packages Help Viewer

standard devia

R: Search Results | Find in Topic

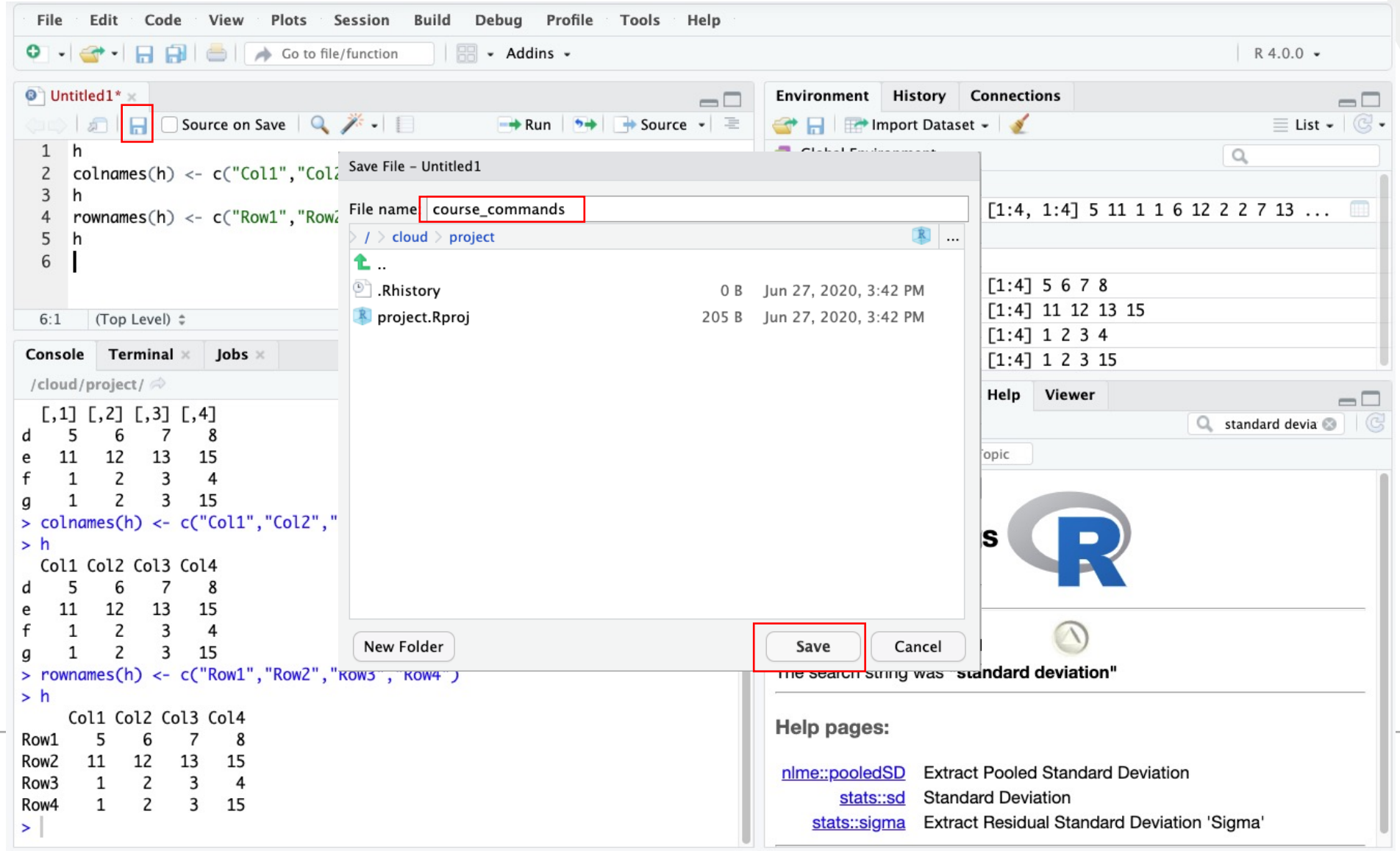
Search Results

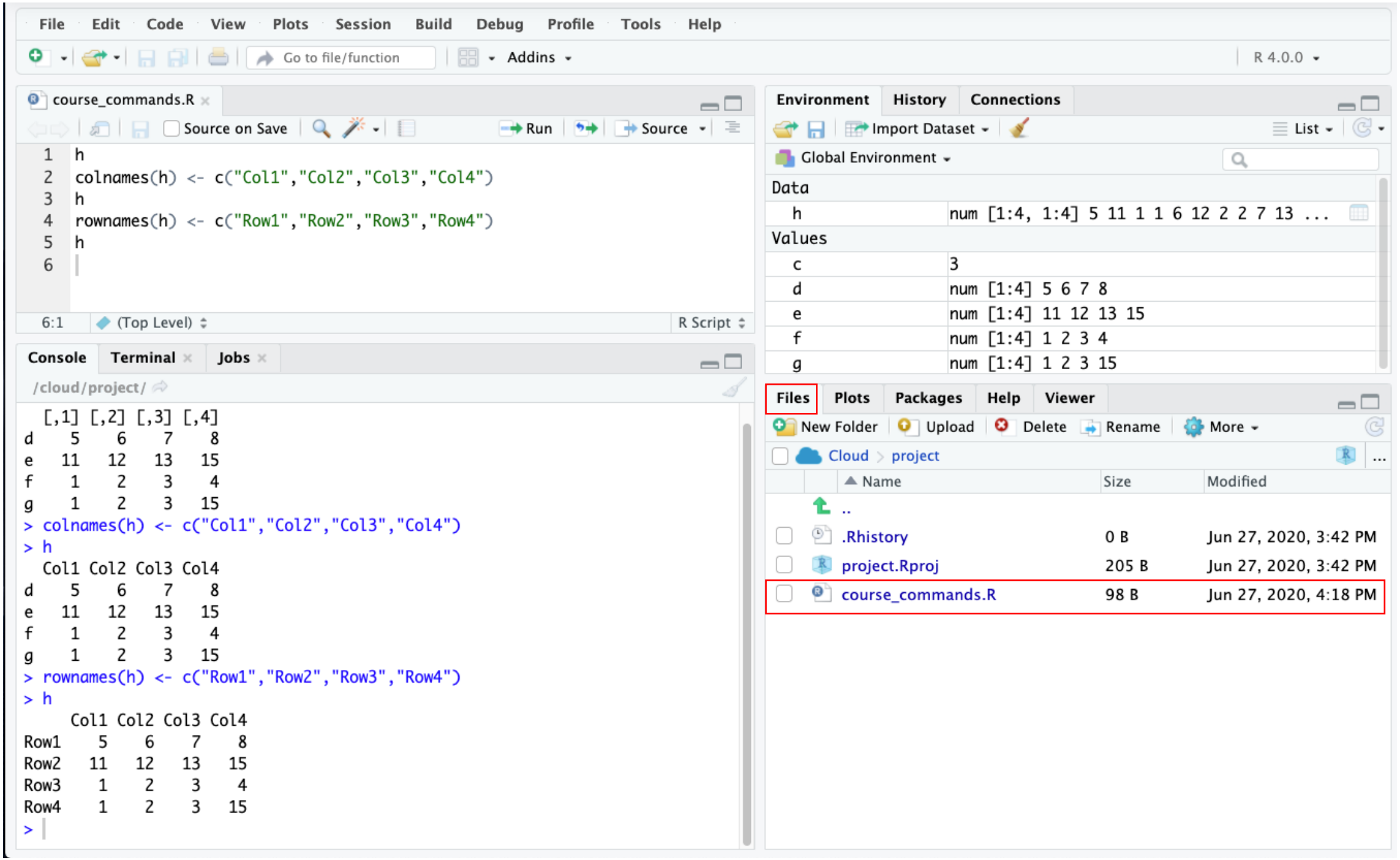
The search string was **"standard deviation"**

Help pages:

nlme::pooledSD	Extract Pooled Standard Deviation
stats::sd	Standard Deviation
stats::sigma	Extract Residual Standard Deviation 'Sigma'

Save your amazing work!




File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 4.0.0
course_commands.R
Source on Save
Run
Source
1 h
2 colnames(h) <- c("Col1","Col2","Col3","Col4")
3 h
4 rownames(h) <- c("Row1","Row2","Row3","Row4")
5 h
6
6:1
(Top Level)
R Script
Console
Terminal
Jobs
/cloud/project/
[,1] [,2] [,3] [,4]
d 5 6 7 8
e 11 12 13 15
f 1 2 3 4
g 1 2 3 15
> colnames(h) <- c("Col1","Col2","Col3","Col4")
> h
Col1 Col2 Col3 Col4
d 5 6 7 8
e 11 12 13 15
f 1 2 3 4
g 1 2 3 15
> rownames(h) <- c("Row1","Row2","Row3","Row4")
> h
Col1 Col2 Col3 Col4
Row1 5 6 7 8
Row2 11 12 13 15
Row3 1 2 3 4
Row4 1 2 3 15
>
Environment
History
Connections
Import Dataset
List
Global Environment
Data
h num [1:4, 1:4] 5 11 1 1 6 12 2 2 7 13 ...
Values
c 3
d num [1:4] 5 6 7 8
e num [1:4] 11 12 13 15
f num [1:4] 1 2 3 4
g num [1:4] 1 2 3 15
Files
Plots
Packages
Help
Viewer
New Folder
Upload
Delete
Rename
More
Cloud
project
Name
Size
Modified
..
.Rhistory 0 B Jun 27, 2020, 3:42 PM
project.Rproj 205 B Jun 27, 2020, 3:42 PM
course_commands.R 98 B Jun 27, 2020, 4:18 PM

Activity 1: Basic_Commands.Rmd

- We just practiced creating a R script to manipulate basic data.
- Try running the Basic_Command.Rmd Notebook to reinforce your skills and learn about running R Markdown.
- Please let me know if you run into any issues or have any questions.

Questions?

- So far, we have:
  - Accessed R via R Studio or R Studio Cloud
  - Created a script
  - Done basic math operations
  - Created and used vectors
  - Done vector math
  - Created a data frame with named rows and cols
  - Run a markdown notebook reinforcing these concepts
- Up next:
  - R Packages and Built-In Datasets

Packages and Built-in Datasets

- Packages are add-ons to base R that increase functionality or ease of use
- Datasets can be imported to or exported from R, however base R and some packages include built-in datasets
- Built-in datasets are helpful for learning about datasets and testing functionality
- Since they do not require complex import, they are ideal for many situations

What is a package?

A collection of R functions, complied code and data

Saved in a directory called “library”

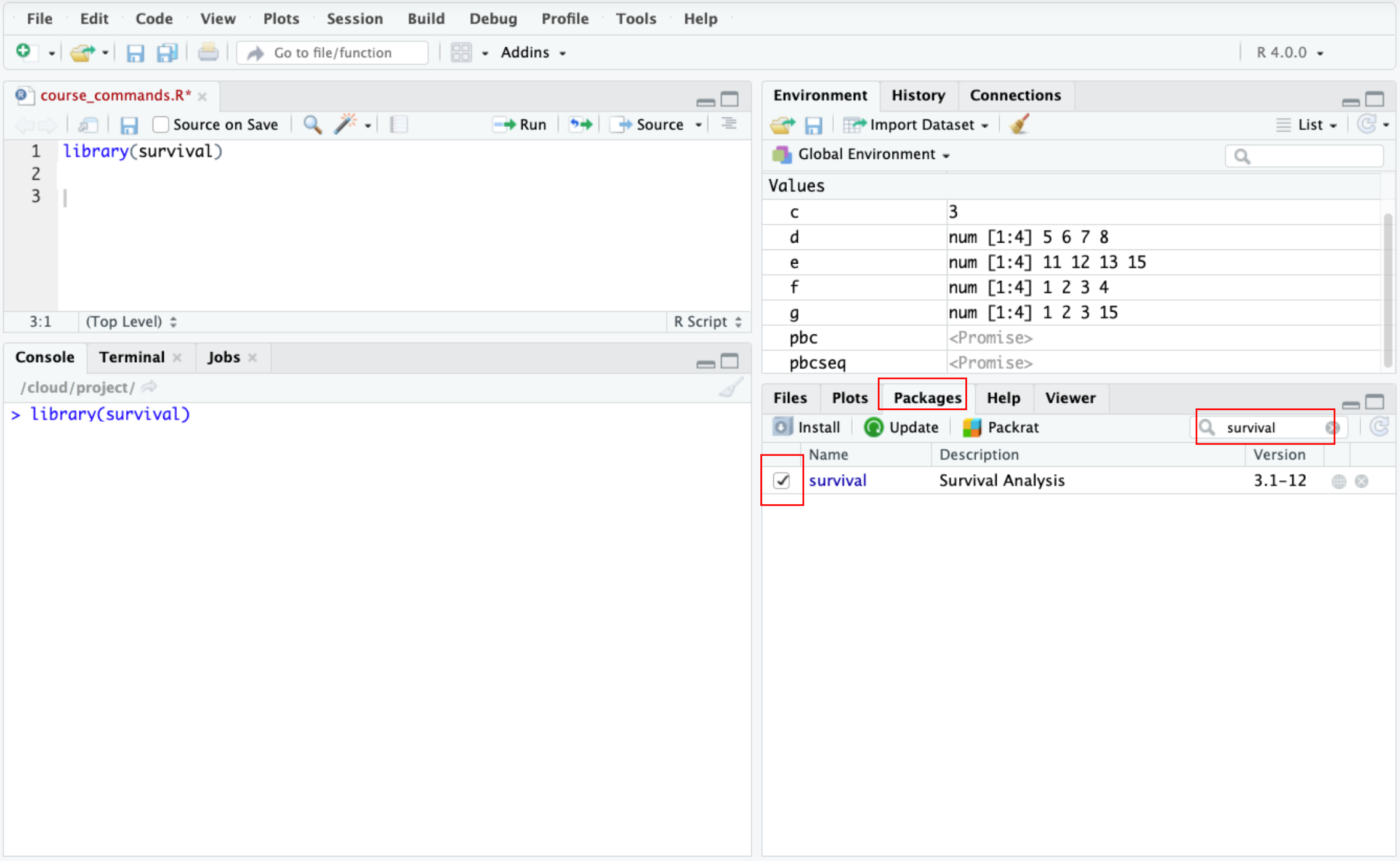
Can be turned on and off

Made by different people commands may clash

Would also be very slow to load everything every time

Survival Package

- https://cran.r-project.org/web/packages/survival/index.html
- PBC Dataset
  - Has a built-in dataset called PBC which is the Mayo Clinic Primary Biliary Cholangitis Data
- We will have to download and load the package to use this dataset


File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 4.0.0
course_commands.R*
Source on Save
Run
Source
1 library(survival)
2
3
3:1
(Top Level)
R Script
Console
Terminal
Jobs
/cloud/project/
> library(survival)
Environment
History
Connections
Import Dataset
List
Global Environment
Values
c 3
d num [1:4] 5 6 7 8
e num [1:4] 11 12 13 15
f num [1:4] 1 2 3 4
g num [1:4] 1 2 3 15
pbc <Promise>
pbcseq <Promise>
Files
Plots
Packages
Help
Viewer
Install
Update
Packrat
survival
Name
Description
Version
survival
Survival Analysis
3.1-12

Activity 2:

Packages and Built-in Datasets

- Click on the Packages_and_Datasets.Rmd file
- This file explores working with built-in datasets and importing packages
- We will work with two datasets
  - Iris
    - Base R dataset
  - PBC
    - Found within the survival package

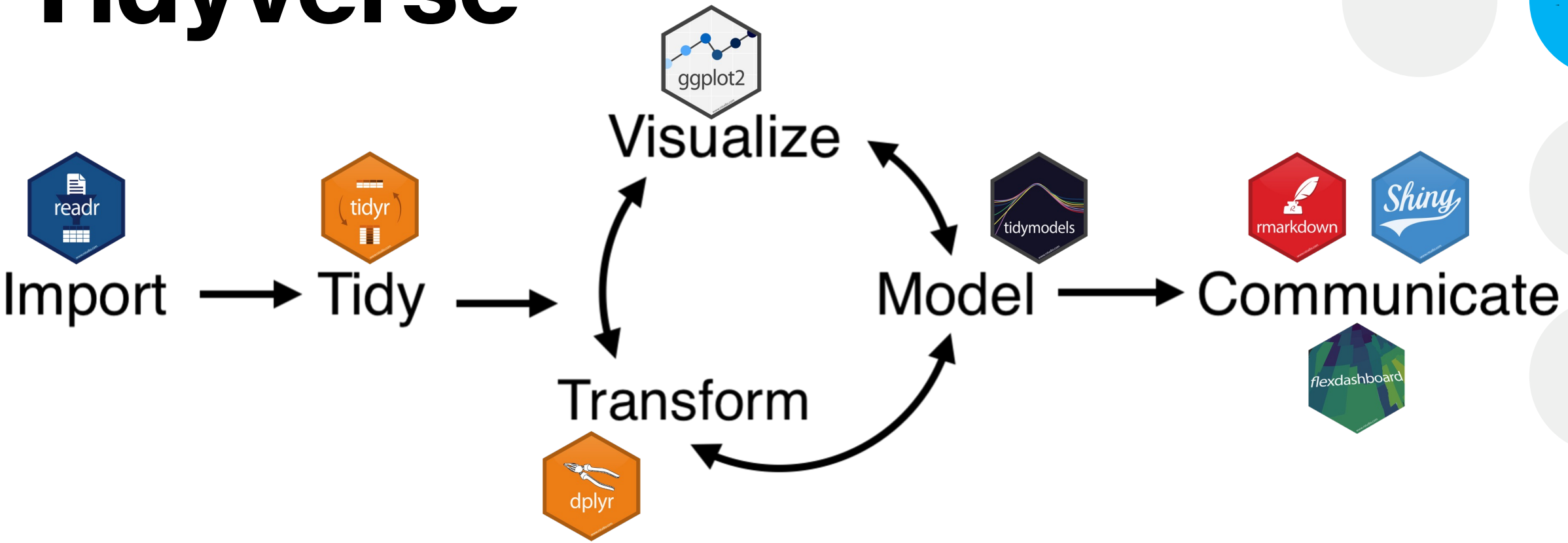
Data Manipulation

WE COULD USE BASE R

OR THE TIDYVERSE COLLECTION OF PACKAGES

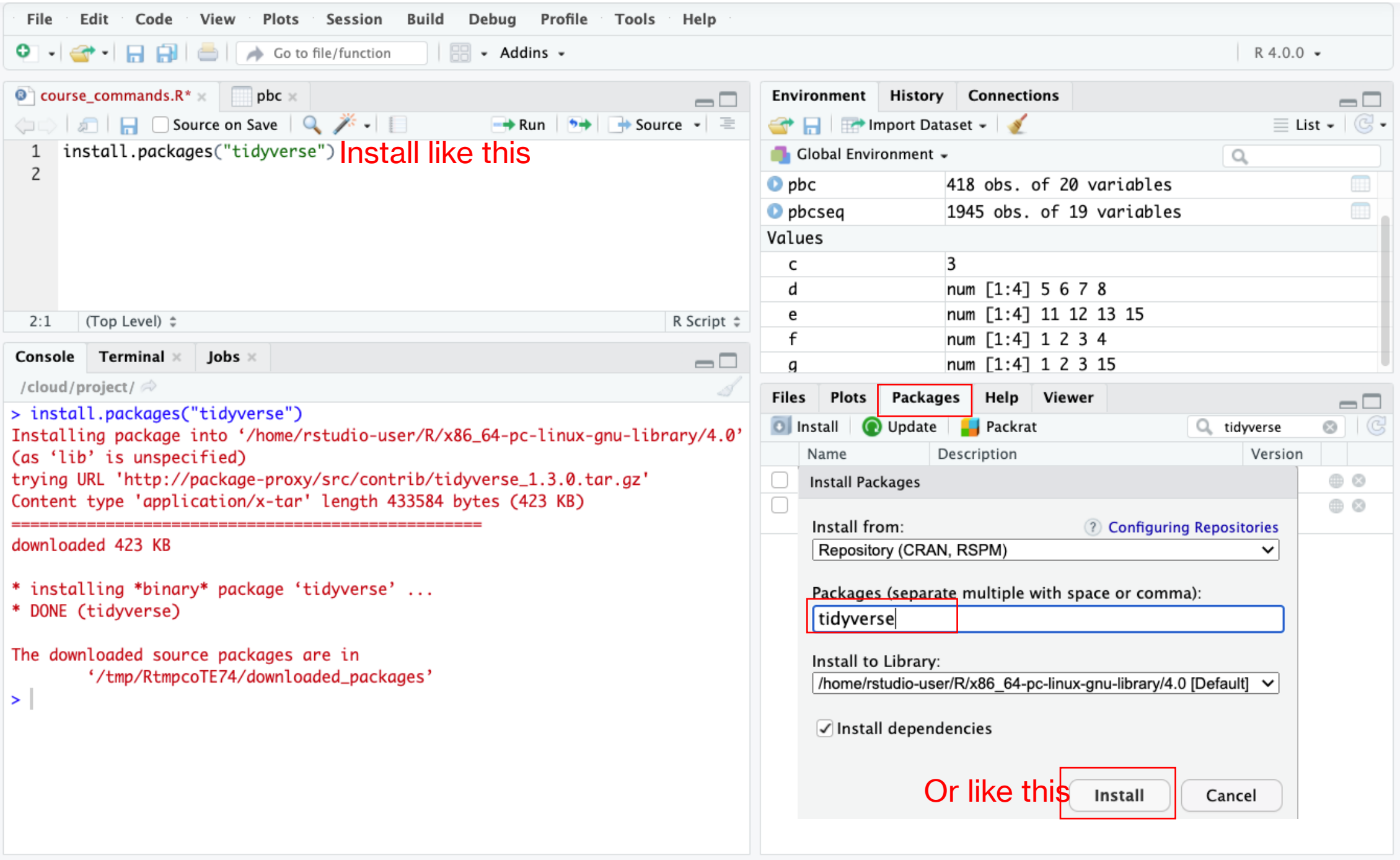
SPECIFICALLY DESIGNED FOR DATA SCIENCE

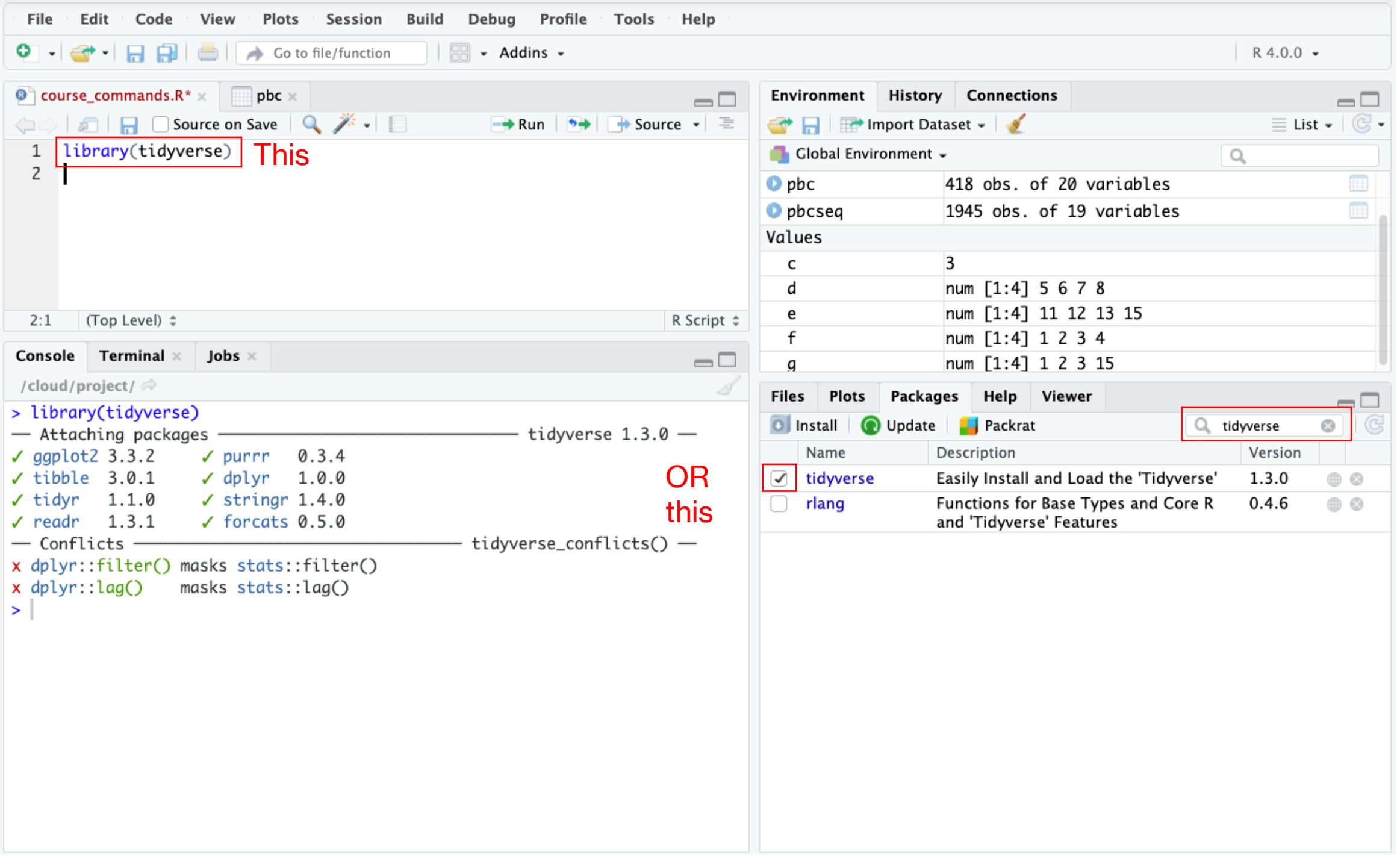
Tidyverse



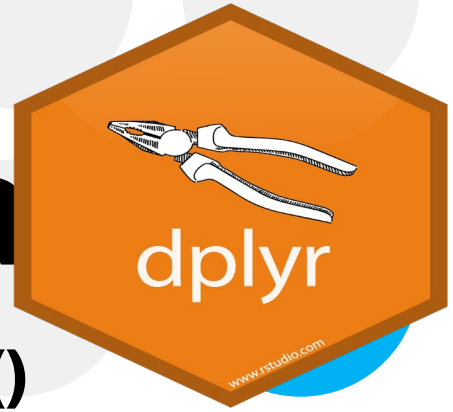
First time install:
install.packages("tidyverse")

Turn on package set: library(tidyverse)


File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 4.0.0
course_commands.R*
pbc
Source on Save
Run
Source
1 install.packages("tidyverse")
2
Install like this
2:1 (Top Level)
R Script
Console
Terminal
Jobs
/cloud/project/
> install.packages("tidyverse")
Installing package into '/home/rstudio-user/R/x86_64-pc-linux-gnu-library/4.0'
(as 'lib' is unspecified)
trying URL 'http://package-proxy/src/contrib/tidyverse_1.3.0.tar.gz'
Content type 'application/x-tar' length 433584 bytes (423 KB)
==
downloaded 423 KB
* installing *binary* package 'tidyverse' ...
* DONE (tidyverse)
The downloaded source packages are in
'/tmp/RtmpcoTE74/downloaded_packages'
>
Environment
History
Connections
Import Dataset
List
Global Environment
pbc 418 obs. of 20 variables
pbcseq 1945 obs. of 19 variables
Values
c 3
d num [1:4] 5 6 7 8
e num [1:4] 11 12 13 15
f num [1:4] 1 2 3 4
g num [1:4] 1 2 3 15
Files
Plots
Packages
Help
Viewer
Install
Update
Packrat
tidyverse
Name
Description
Version
Install Packages
Install from:
Configuring Repositories
Repository (CRAN, RSPM)
Packages (separate multiple with space or comma):
tidyverse
Install to Library:
/home/rstudio-user/R/x86_64-pc-linux-gnu-library/4.0 [Default]
Install dependencies
Or like this
Install
Cancel


File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 4.0.0
course_commands.R*
pbc
Source on Save
Run
Source
1 library(tidyverse)
2
This
2:1 (Top Level)
R Script
Console
Terminal
Jobs
/cloud/project/
> library(tidyverse)
── Attaching packages ──────────────────────────────────── tidyverse 1.3.0 ──
✓ ggplot2 3.3.2 ✓ purrr 0.3.4
✓ tibble 3.0.1 ✓ dplyr 1.0.0
✓ tidyr 1.1.0 ✓ stringr 1.4.0
✓ readr 1.3.1 ✓ forcats 0.5.0
── Conflicts ─────────────────────────────────────── tidyverse_conflicts() ──
x dplyr::filter() masks stats::filter()
x dplyr::lag() masks stats::lag()
>
OR
this
Environment
History
Connections
Import Dataset
List
Global Environment
pbc 418 obs. of 20 variables
pbcseq 1945 obs. of 19 variables
Values
c 3
d num [1:4] 5 6 7 8
e num [1:4] 11 12 13 15
f num [1:4] 1 2 3 4
g num [1:4] 1 2 3 15
Files
Plots
Packages
Help
Viewer
Install
Update
Packrat
tidyverse
Name
Description
Version
tidyverse
Easily Install and Load the 'Tidyverse'
1.3.0
rlang
Functions for Base Types and Core R and 'Tidyverse' Features
0.4.6

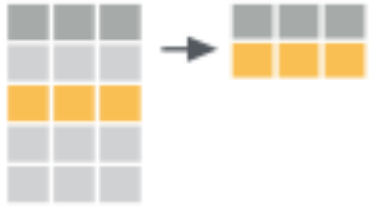
dplyr: Transform your da



arrange()

select()

filter()



mutate()

summarize()

group_by()

Basic structure:
dpylr function(data, specifics)
Example:
arrange(pbc, age)

Pipes!

Let's say I want to know the average age for males versus females I could:

```
group_by_sex <- group_by(pbc,sex)
ave_age_sex <-summarize(group_by_sex, mean = mean(age))
```

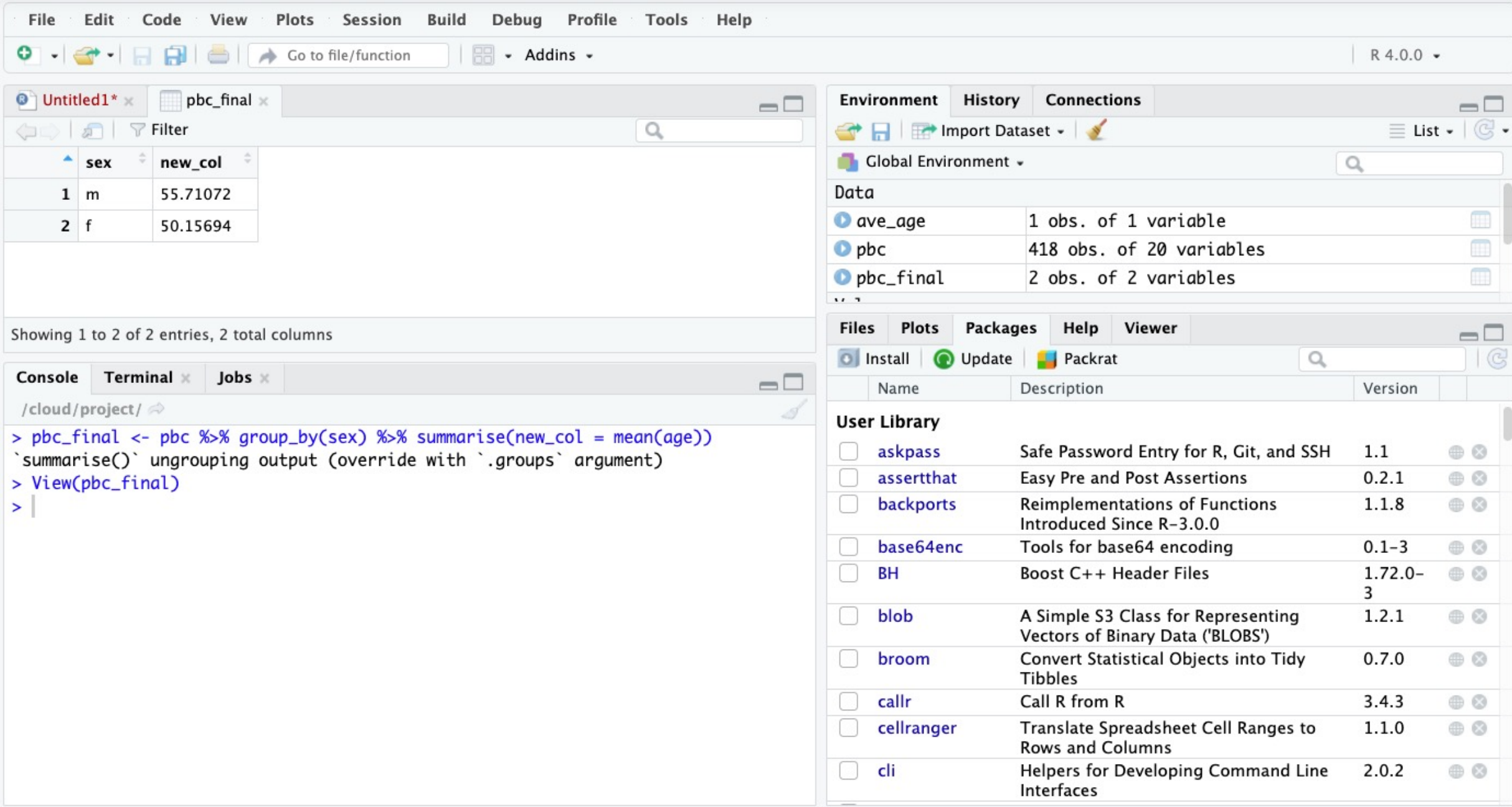
OR.

Use a pipe! %>%

Keyboard Shortcut:
PC:Ctrl+Shift+M
Mac:Cmd+Shift+M

This allows sequential operations to be done on the same dataset:

```
pbc_final <- pbc %>% group_by(sex) %>% summarize(new_col = mean(age))
View(pbc_final)
```


File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 4.0.0
Untitled1*
pbc_final
Filter
sex
new_col
1 m 55.71072
2 f 50.15694
Showing 1 to 2 of 2 entries, 2 total columns
Console
Terminal
Jobs
/cloud/project/
> pbc_final <- pbc %>% group_by(sex) %>% summarise(new_col = mean(age))
`summarise()` ungrouping output (override with `.groups` argument)
> View(pbc_final)
>
Environment
History
Connections
Import Dataset
List
Global Environment
Data
ave_age 1 obs. of 1 variable
pbc 418 obs. of 20 variables
pbc_final 2 obs. of 2 variables
Files
Plots
Packages
Help
Viewer
Install
Update
Packrat
Name
Description
Version
User Library
askpass Safe Password Entry for R, Git, and SSH 1.1
assertthat Easy Pre and Post Assertions 0.2.1
backports Reimplementations of Functions Introduced Since R-3.0.0 1.1.8
base64enc Tools for base64 encoding 0.1-3
BH Boost C++ Header Files 1.72.0-3
blob A Simple S3 Class for Representing Vectors of Binary Data ('BLOBS') 1.2.1
broom Convert Statistical Objects into Tidy Tibbles 0.7.0
callr Call R from R 3.4.3
cellranger Translate Spreadsheet Cell Ranges to Rows and Columns 1.1.0
cli Helpers for Developing Command Line Interfaces 2.0.2

Activity 3:
Data Manipulation

- Using the pre-made R Notebook Data_Manipulation.Rmd we will explore the following data manipulations:
  - Arrange
  - Select
  - Filter
  - Mutate
  - Summarize
  - GroupBy

Questions?

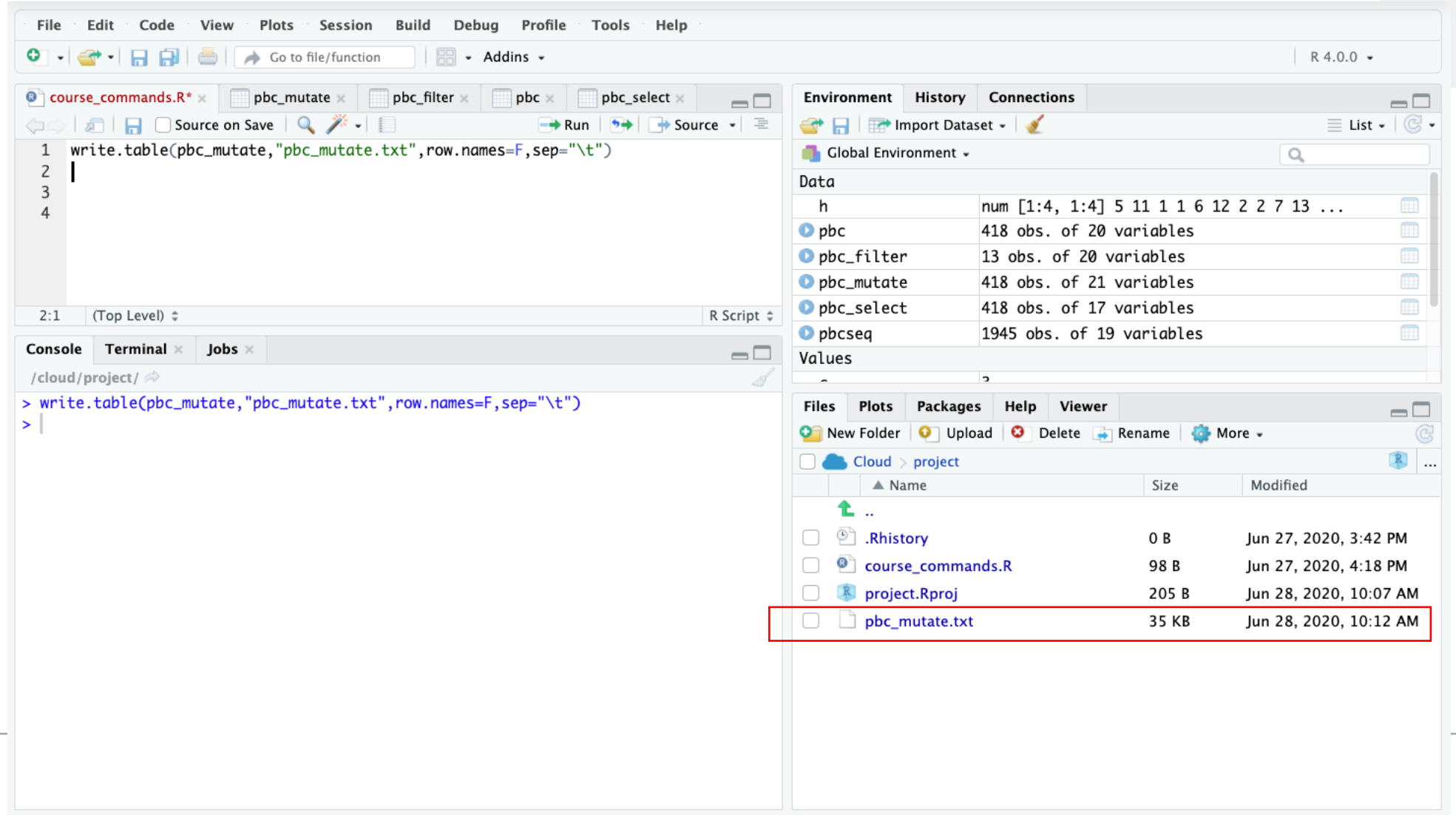
- What data manipulation options are available?
- Which operation would you use to create a new column of data?
- Which two options are available for arranging data?
- What symbol is used to join multiple commands?

Data Import and Export

- Data comes in many formats
- There are a variety of tools for importing and exporting data
- R can handle data in csv, tab-delimited, excel formats and more

Save a table as a file

write.table(pbc_mutate,"pbc_mutate.txt",row.names=F,sep="\t")



Read Data From Website

- read_csv() and read_tsv() are special cases of the more general read_delim(). They're useful for reading the most common types of flat file data, comma separated values and tab separated values, respectively.

- File can be a url where data is stored
- Run `?read_tsv` for more help information

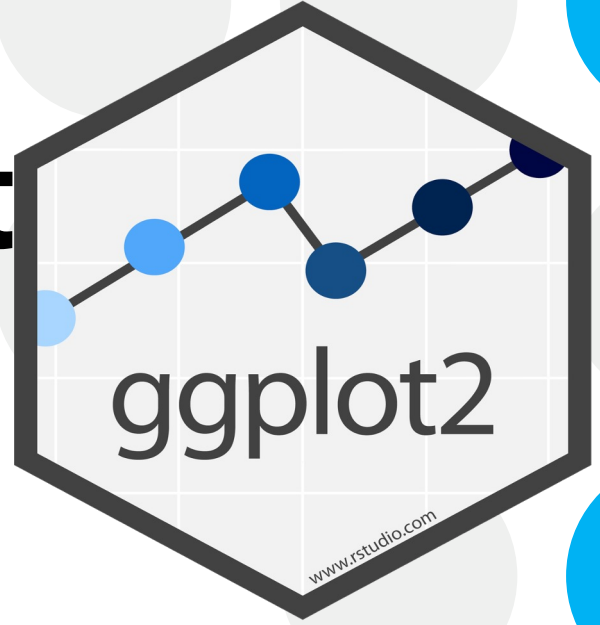
Activity 4:
Data Import and Export

- Using the pre-made R Notebook Data_Import_and_Export.Rmd we will explore the following:
  - Save as tab delimited text
  - Save as a csv file
  - Load both tab delimited and csv files
  - Load data from a url

Visualizations for Today

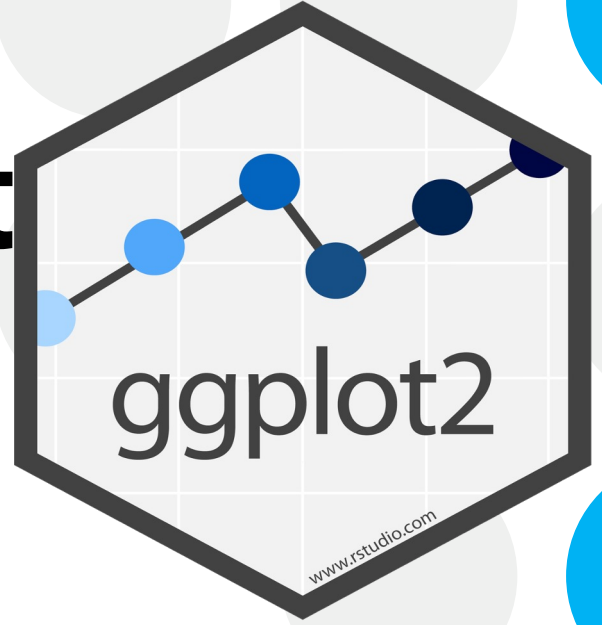
- Bar plot
- Histogram
- Density Plot
- Scatter plot
- Faceted plots (multiple plots in the same figure)

ggplot2: Visualize your dat



- Easy out of box formatting
- Handles complex data quickly
- Default options are aesthetically pleasing
- Layering system = add complexity as you go
- Automatic scaling generally works well
- Great documentation and support

ggplot2: Visualize your dat



What you need:

1. A data object
2. Aesthetic mappings (aes): how variables in the data are assigned to visual properties
    - x- and y-direction
    - shapes, colors, lines
3. A geometry object (geom): the type of plot

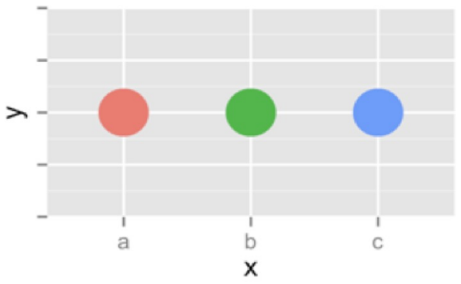
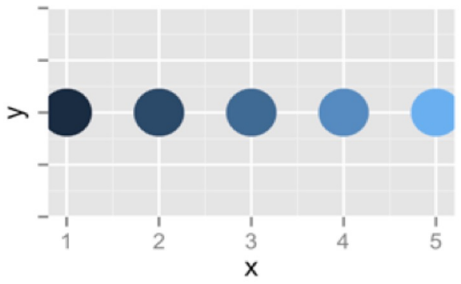
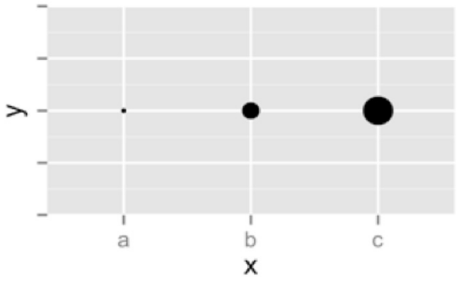
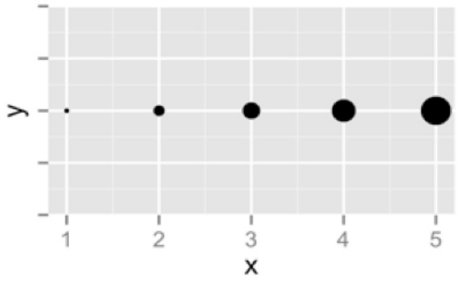
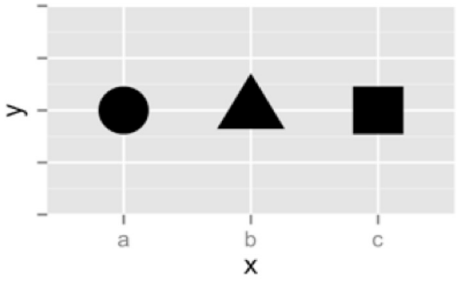
Basic structure:

ggplot(data, aes(x=variable)) + geom_type()

Example:

ggplot(pbc, aes(x=sex)) + geom_bar()

Aesthetic mapping options

	Discrete	Continuous
Color		
Size		
Shape		

ggplot2:Faceting

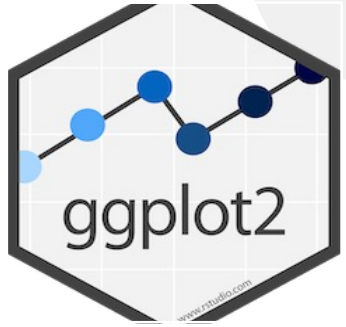
- Divide a plot into subplots based on one or more discrete variable
- Can be used with a variety of plot types
- There are a couple of facet flavors

t + facet_grid(cols = vars(fl))
facet into columns based on fl

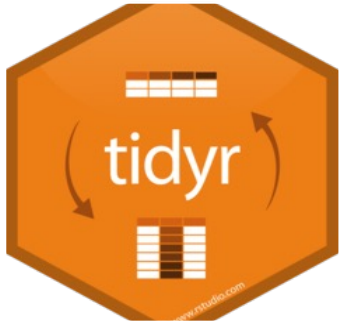
t + facet_grid(rows = vars(year))
facet into rows based on year

t + facet_grid(rows = vars(year), cols = vars(fl))
facet into both rows and columns

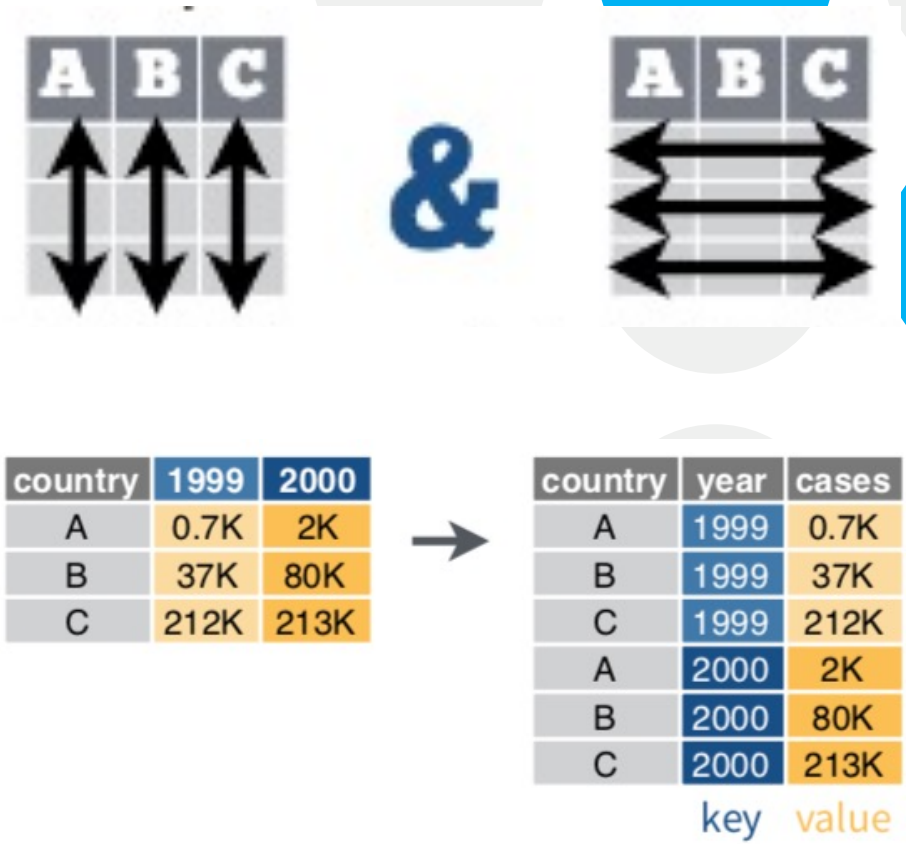
t + facet_wrap(vars(fl))
wrap facets into a rectangular layout



tidr:gather



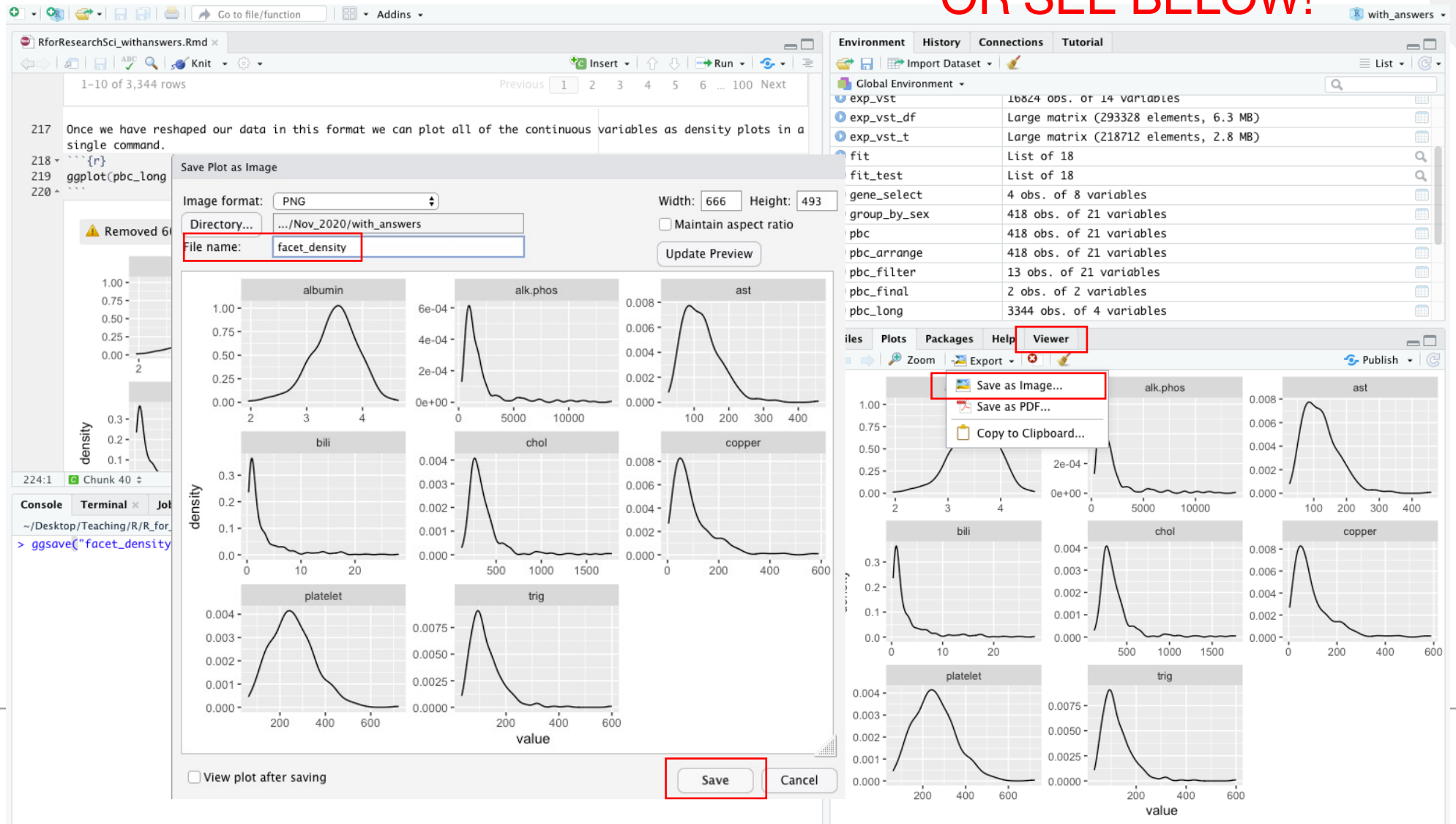
- Typically, each variable is in its own column and each observation/case is its own row
- Transform data from wide to long format
- **gather(**data, key, value**)**
- Moves column names into a **key** column, gathering the column values into a single **value** column



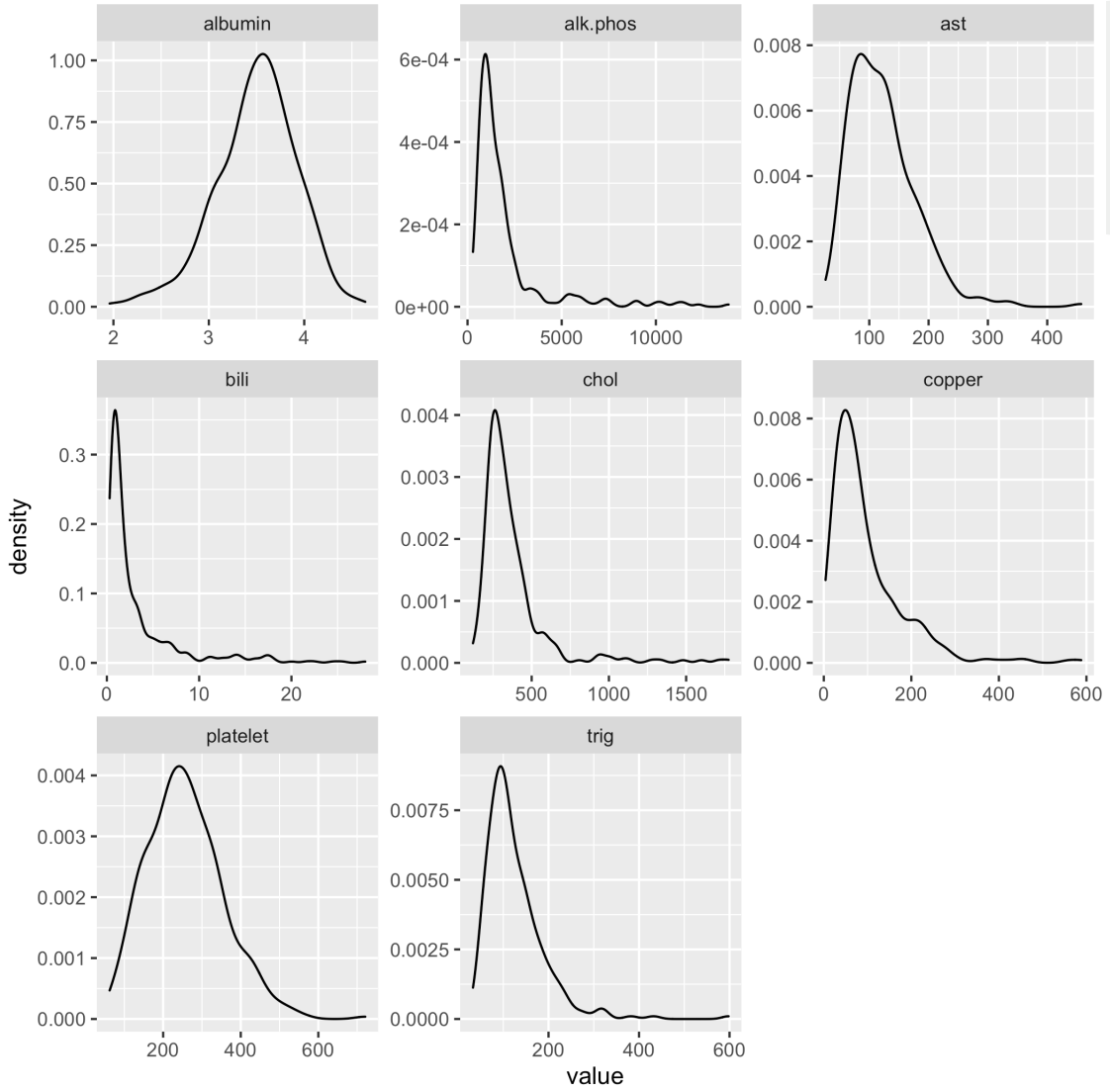
Saving your beautiful graph!

ggsave("facet_density.png")

OR SEE BELOW!



Questions on Basic Visualizations



Activity 5:
Data Visualization

- Open the notebook Data_Visualization.Rmd
- Explore plotting and data visualizations

Data Analysis in R

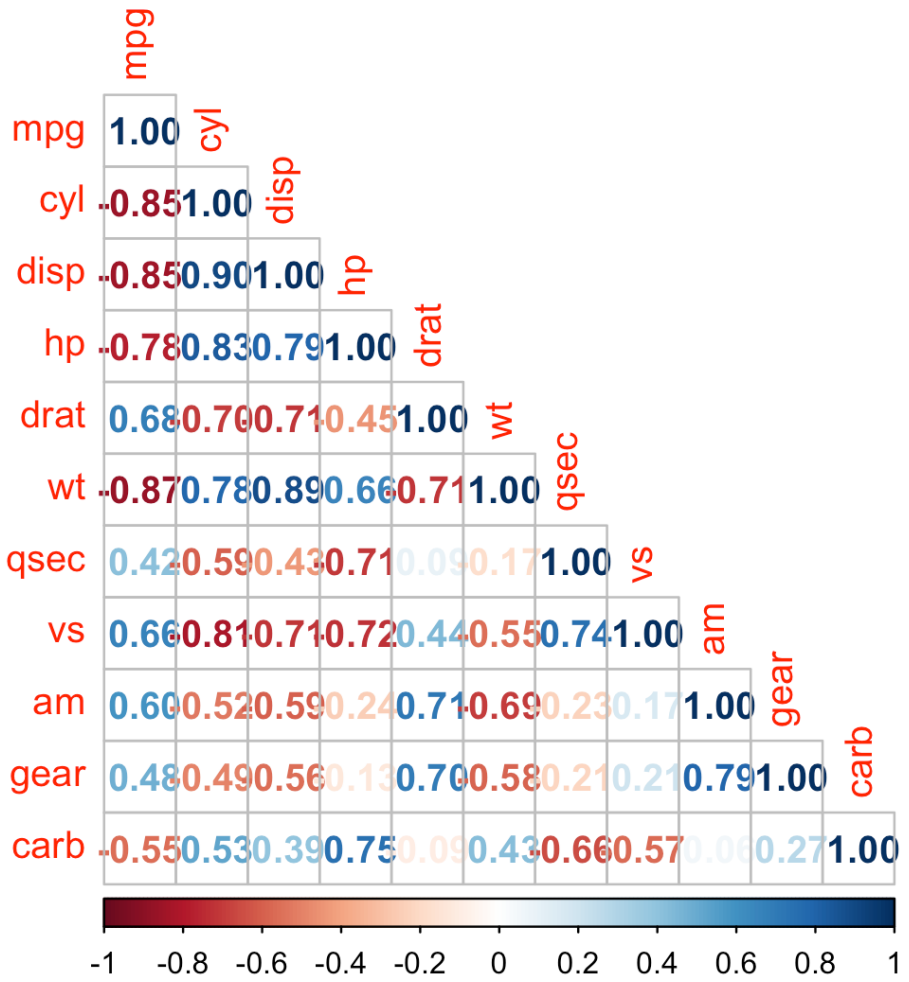
- Correlations
  - Pearson
- t-tests
  - Single sample
  - Independent
  - Paired

Correlation

- Measure of how well two variables hang together
- Range from 1 to -1
- 1 is a perfect positive correlation
- -1 is an inverse correlation

Correlation

- Diagonal is all 1's
- Positive and negative correlations are color coded
- Strength is shown in darker colors



Correlation Test

Cor.test

P-value shows significance

Correlation coefficient is shown at bottom

```r
cor.test(mtcars$mpg, mtcars$hp)
```

```
        Pearson's product-moment correlation

data:  mtcars$mpg and mtcars$hp
t = -6.7424, df = 30, p-value = 1.788e-07
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.8852686 -0.5860994
sample estimates:
       cor
-0.7761684
```

t-tests

- Single sample
  - Compares a sample to a known value (mu)
- Independent
  - Compares two samples which are not related
- Paired
  - Compares two samples which are related

One Sample t-test

- Compares a vector of data against a know value (mu)
- p-value indicates significance
- Mean of vector is shown at bottom

```r
t.test(mtcars$wt, mu=3)
```

```
        One Sample t-test

data:  mtcars$wt
t = 1.256, df = 31, p-value = 0.2185
alternative hypothesis: true mean is not equal to 3
95 percent confidence interval:
 2.864478 3.570022
sample estimates:
mean of x
  3.21725
```

Two Sample Independent t-test

- Compares two vectors of numeric data
- Determines if means are different
- p-value indicates significance
- Means of each group displayed at bottom

```{r}
t.test(mtcars$mpg~mtcars$vs)
```

```
        Welch Two Sample t-test

data:  mtcars$mpg by mtcars$vs
t = -4.6671, df = 22.716, p-value = 0.0001098
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 -11.462508  -4.418445
sample estimates:
mean in group 0 mean in group 1
       16.61667        24.55714
```

Paired-Samples t-test

- Compares two related samples
- P-value indicates significance
- Average of the differences is shown at the bottom

```r
t.test(ChickWeightWide$weight.0, ChickWeightWide$weight.2, paired = TRUE)
```

```
        Paired t-test

data:  ChickWeightWide$weight.0 and ChickWeightWide$weight.2
t = -17.409, df = 44, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -9.496402 -7.525820
sample estimates:
mean of the differences
              -8.511111
```

Activity 6:
Data Analysis in R

- Open the notebook Data_Analysis.Rmd
- Basic analysis techniques in R
  - Correlations
  - t-tests

This concludes the Intro to R

Additional materials and examples can be found in the supplemental topics folder of this repository.