Introduction to Data Science in the Tidyverse

Amelia McNamara and Hadley Wickham

Acknowledgements

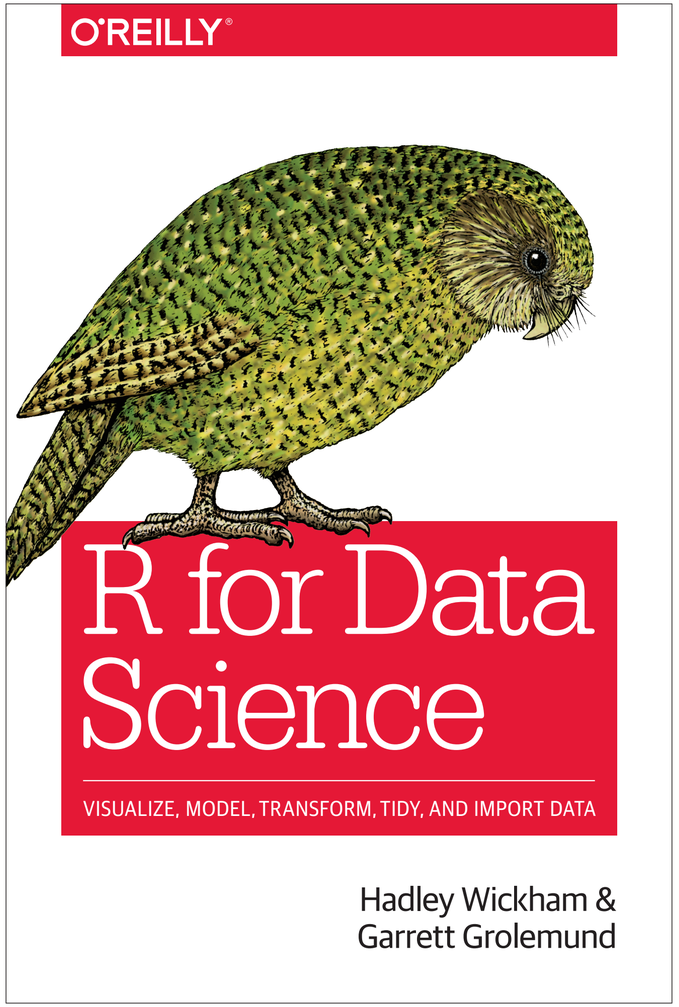
Major contributions by Garrett Grolemund, Amelia McNamara, Mike Smith, Charlotte Wickham, and Hadley Wickham.

Previous versions of this material available at

https://github.com/rstudio-education/master-the-tidyverse
https://github.com/AmeliaMN/IntroToR
https://github.com/cwickham/data-science-in-tidyverse
https://github.com/AmeliaMN/data-science-in-tidyverse

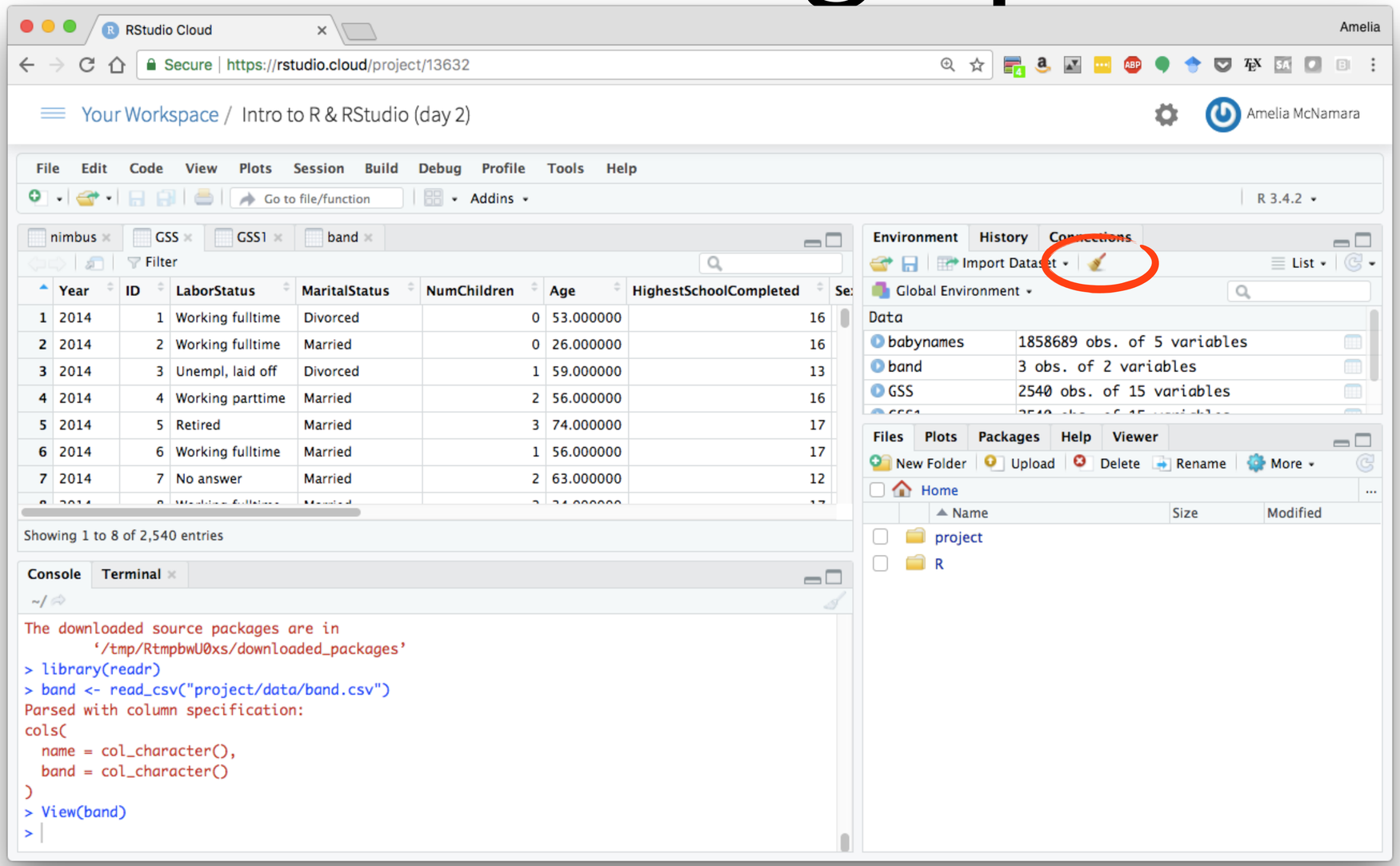
Online at:

http://r4ds.had.co.nz/

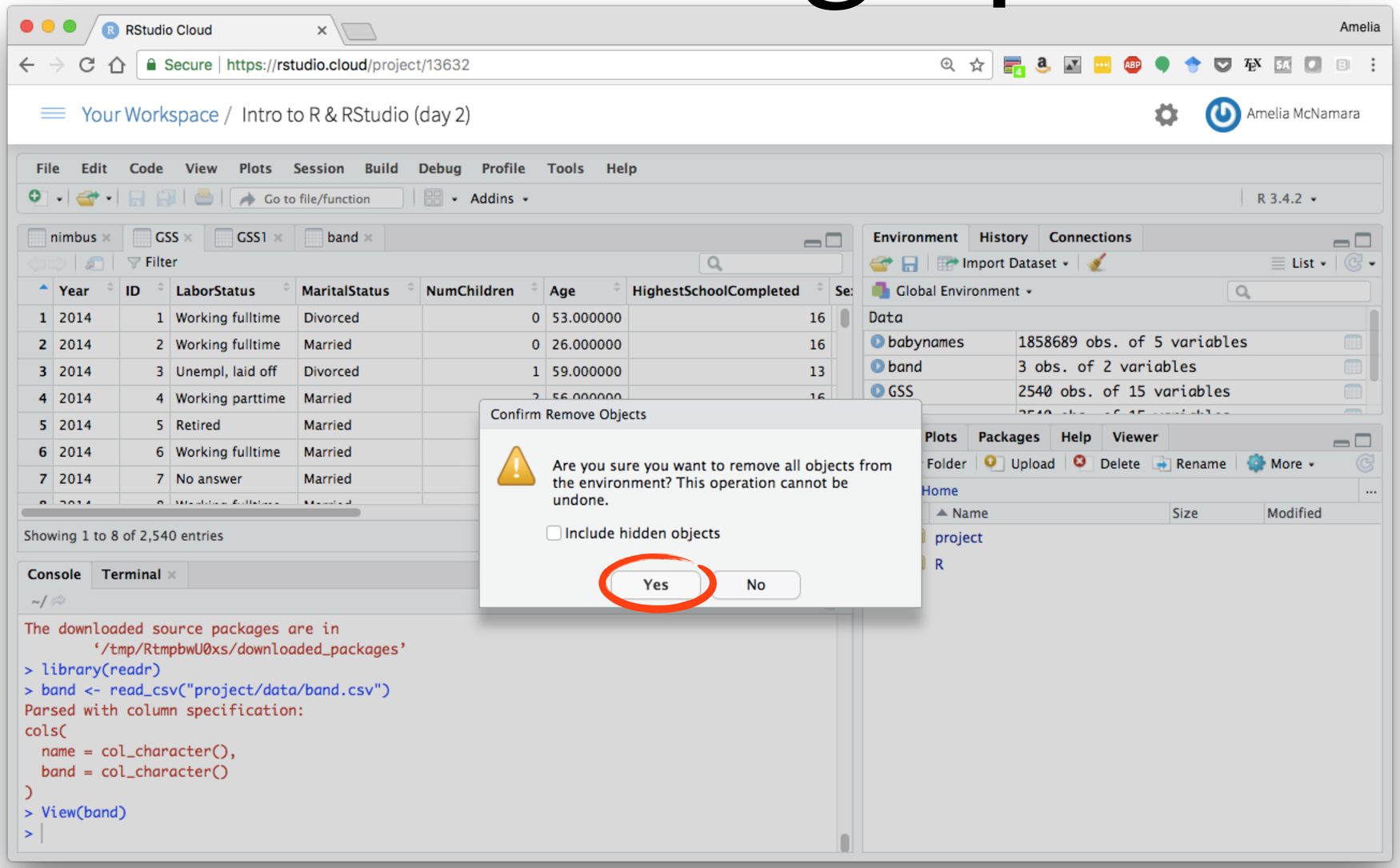


RStudio best practices

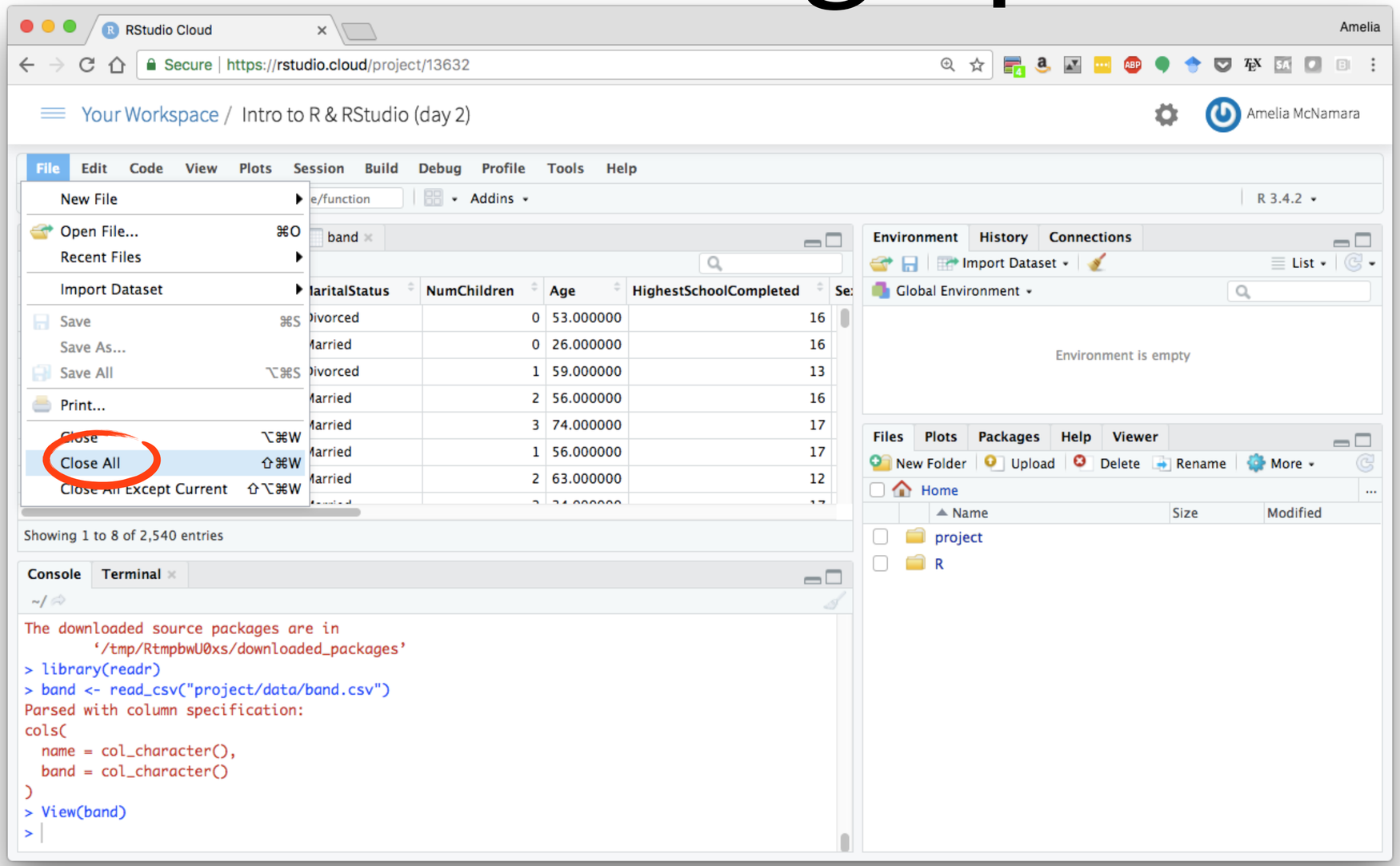
Cleaning up


RStudio Cloud
Amelia
Secure | https://rstudio.cloud/project/13632
Your Workspace / Intro to R & RStudio (day 2)
Amelia McNamara
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 3.4.2
nimbus
GSS
GSS1
band
Filter
Year ID LaborStatus MaritalStatus NumChildren Age HighestSchoolCompleted
1 2014 1 Working fulltime Divorced 0 53.000000 16
2 2014 2 Working fulltime Married 0 26.000000 16
3 2014 3 Unempl, laid off Divorced 1 59.000000 13
4 2014 4 Working parttime Married 2 56.000000 16
5 2014 5 Retired Married 3 74.000000 17
6 2014 6 Working fulltime Married 1 56.000000 17
7 2014 7 No answer Married 2 63.000000 12
Showing 1 to 8 of 2,540 entries
Environment History
Import Dataset
List
Global Environment
Data
babynames 1858689 obs. of 5 variables
band 3 obs. of 2 variables
GSS 2540 obs. of 15 variables
Files Plots Packages Help Viewer
New Folder Upload Delete Rename More
Home
Name Size Modified
project
R
Console Terminal
~/
The downloaded source packages are in
'/tmp/RtmpbwU0xs/downloaded_packages'
> library(readr)
> band <- read_csv("project/data/band.csv")
Parsed with column specification:
cols(
name = col_character(),
band = col_character()
)
> View(band)
>

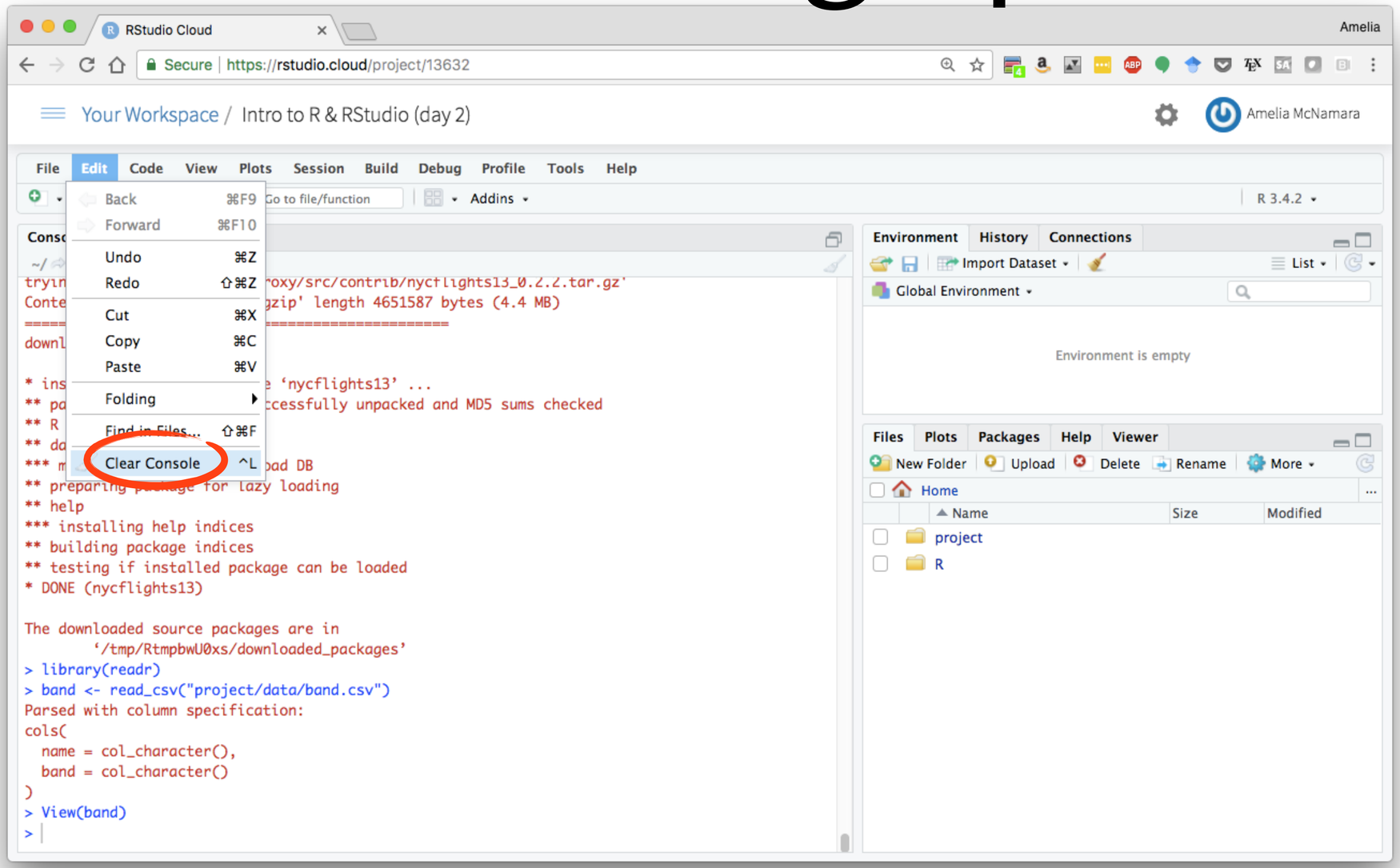
Cleaning up


RStudio Cloud
Amelia
Secure | https://rstudio.cloud/project/13632
Your Workspace / Intro to R & RStudio (day 2)
Amelia McNamara
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 3.4.2
nimbus
GSS
GSS1
band
Filter
Year ID LaborStatus MaritalStatus NumChildren Age HighestSchoolCompleted
1 2014 1 Working fulltime Divorced 0 53.000000 16
2 2014 2 Working fulltime Married 0 26.000000 16
3 2014 3 Unempl, laid off Divorced 1 59.000000 13
4 2014 4 Working parttime Married
5 2014 5 Retired Married
6 2014 6 Working fulltime Married
7 2014 7 No answer Married
Showing 1 to 8 of 2,540 entries
Environment History Connections
Import Dataset
List
Global Environment
Data
babynames 1858689 obs. of 5 variables
band 3 obs. of 2 variables
GSS 2540 obs. of 15 variables
Confirm Remove Objects
Are you sure you want to remove all objects from the environment? This operation cannot be undone.
Include hidden objects
Yes
No
Plots Packages Help Viewer
Folder Upload Delete Rename More
Home
Name Size Modified
project
R
Console Terminal
~/
The downloaded source packages are in
'/tmp/RtmpbwU0xs/downloaded_packages'
> library(readr)
> band <- read_csv("project/data/band.csv")
Parsed with column specification:
cols(
name = col_character(),
band = col_character()
)
> View(band)
>

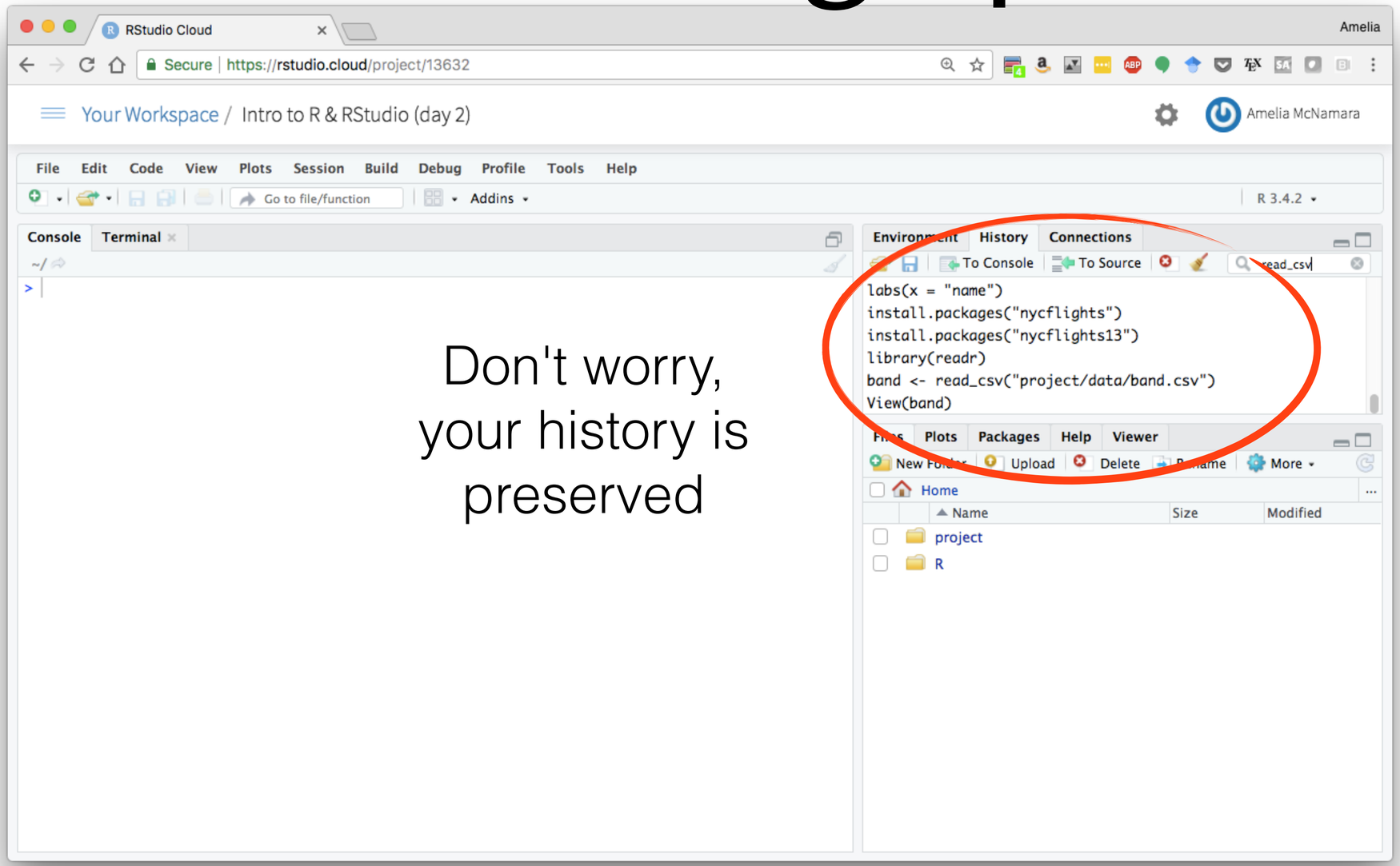
Cleaning up


RStudio Cloud
Amelia
Secure | https://rstudio.cloud/project/13632
Your Workspace / Intro to R & RStudio (day 2)
Amelia McNamara
File Edit Code View Plots Session Build Debug Profile Tools Help
New File
Open File... ⌘O
Recent Files
Import Dataset
Save ⌘S
Save As...
Save All ⌥⌘S
Print...
Close ⌥⌘W
Close All ⇧⌘W
Close All Except Current ⇧⌥⌘W
Addins
R 3.4.2
band
NumChildren Age HighestSchoolCompleted
0 53.000000 16
0 26.000000 16
1 59.000000 13
2 56.000000 16
3 74.000000 17
1 56.000000 17
2 63.000000 12
Showing 1 to 8 of 2,540 entries
Environment History Connections
Import Dataset
List
Global Environment
Environment is empty
Files Plots Packages Help Viewer
New Folder Upload Delete Rename More
Home
Name Size Modified
project
R
Console Terminal
The downloaded source packages are in
'/tmp/RtmpbwU0xs/downloaded_packages'
> library(readr)
> band <- read_csv("project/data/band.csv")
Parsed with column specification:
cols(
name = col_character(),
band = col_character()
)
> View(band)
>

Cleaning up


RStudio Cloud
Amelia
Secure | https://rstudio.cloud/project/13632
Your Workspace / Intro to R & RStudio (day 2)
Amelia McNamara
File Edit Code View Plots Session Build Debug Profile Tools Help
Back ⌘F9
Forward ⌘F10
Undo ⌘Z
Redo ⇧⌘Z
Cut ⌘X
Copy ⌘C
Paste ⌘V
Folding
Find in Files... ⇧⌘F
Clear Console ^L
Addins
R 3.4.2
Environment History Connections
Import Dataset
List
Global Environment
Environment is empty
Files Plots Packages Help Viewer
New Folder Upload Delete Rename More
Home
Name Size Modified
project
R
* installing help indices
** building package indices
** testing if installed package can be loaded
* DONE (nycflights13)
The downloaded source packages are in
'/tmp/RtmpbwU0xs/downloaded_packages'
> library(readr)
> band <- read_csv("project/data/band.csv")
Parsed with column specification:
cols(
name = col_character(),
band = col_character()
)
> View(band)

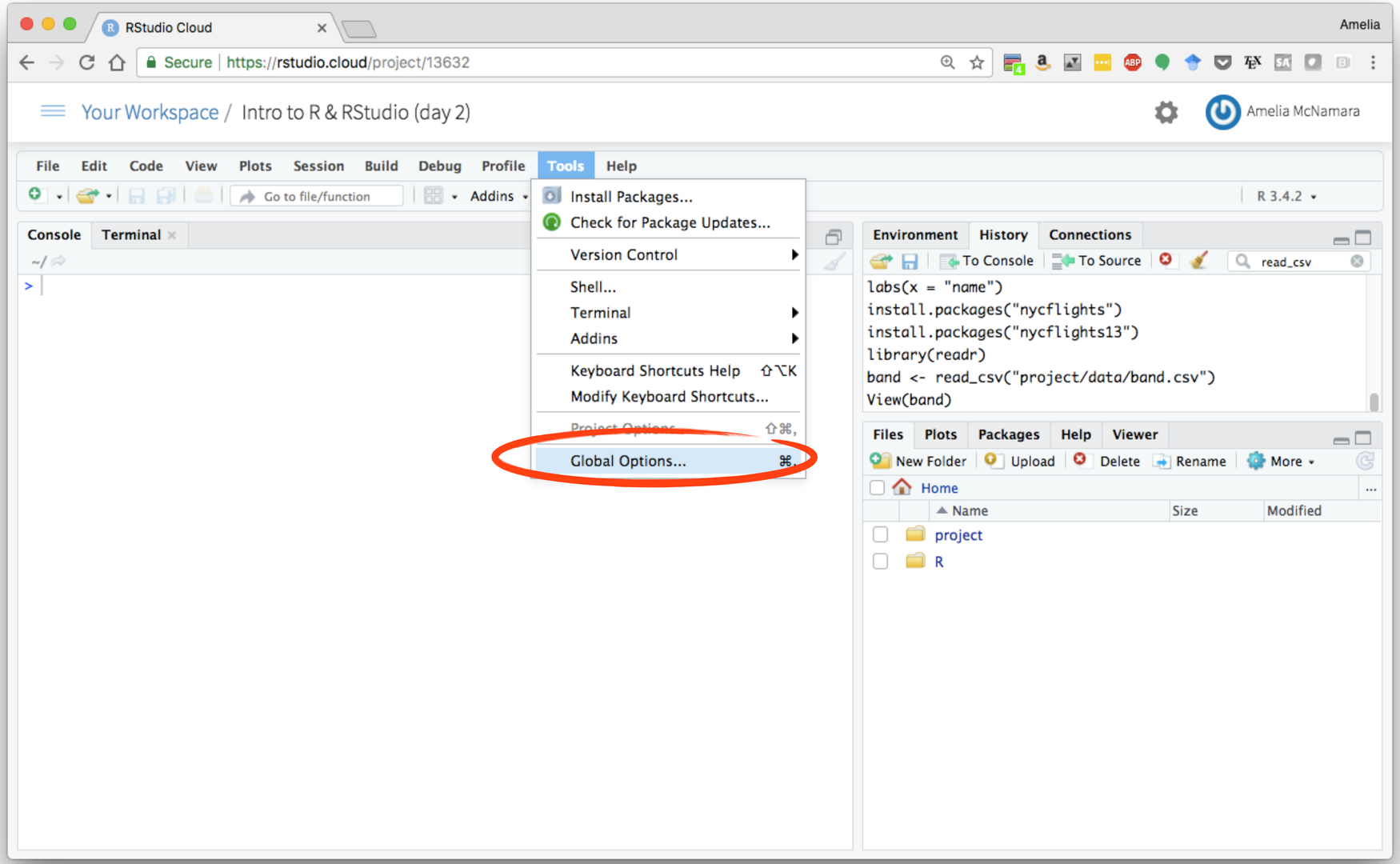
Cleaning up



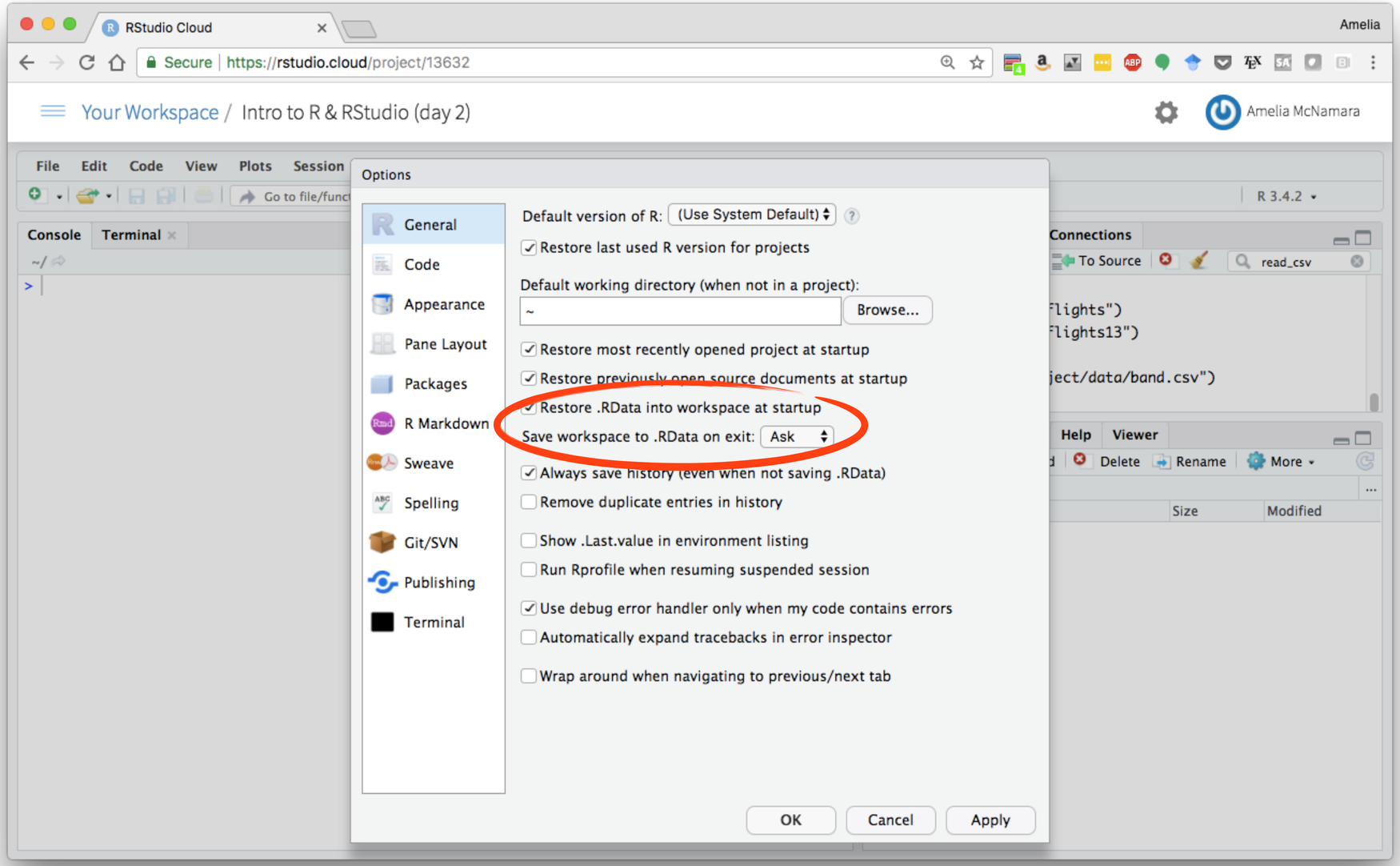
Don't worry, your history is preserved

Settings

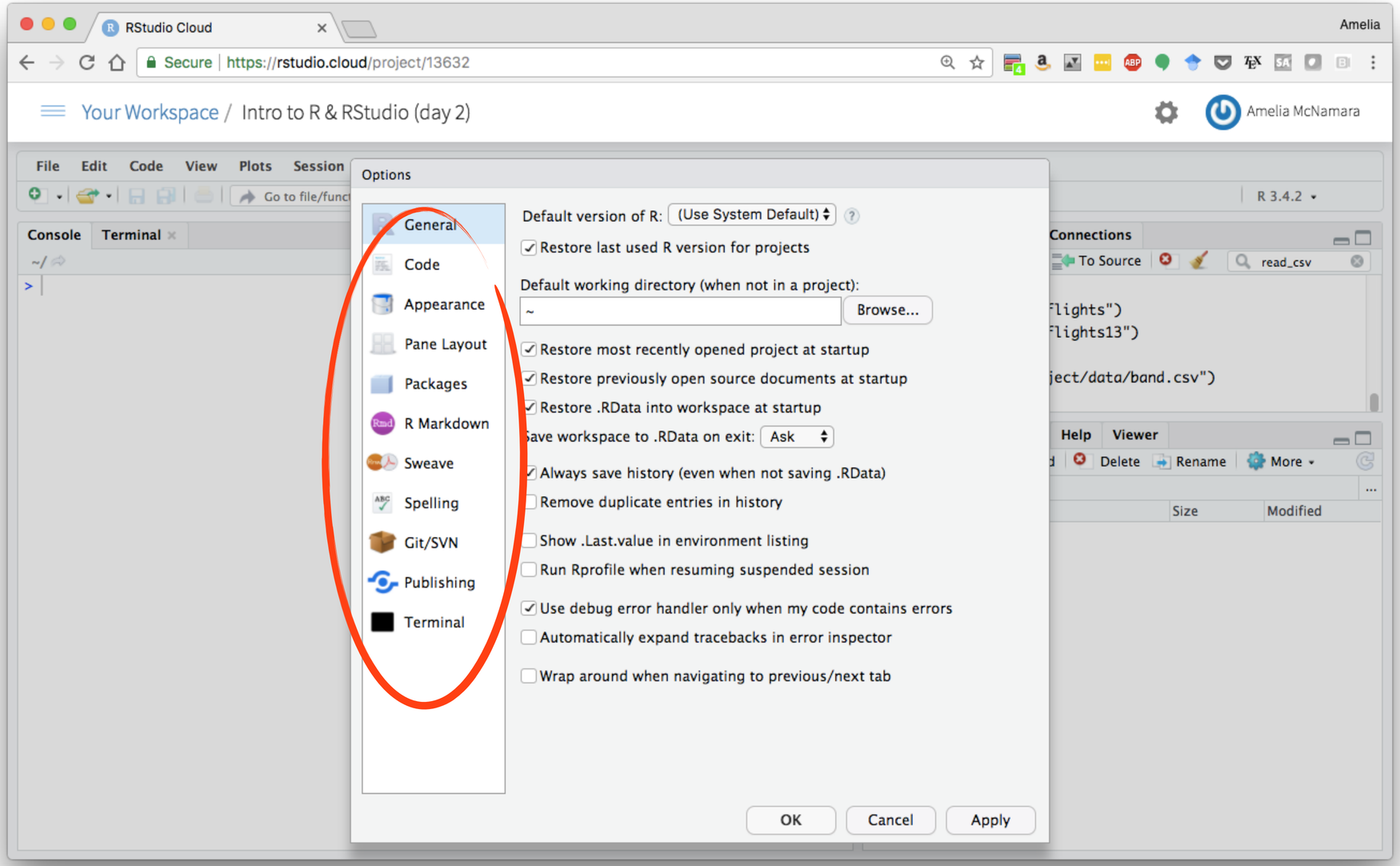
Cleaning up


RStudio Cloud
Amelia
Secure | https://rstudio.cloud/project/13632
Your Workspace / Intro to R & RStudio (day 2)
Amelia McNamara
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 3.4.2
Console Terminal
Install Packages...
Check for Package Updates...
Version Control
Shell...
Terminal
Addins
Keyboard Shortcuts Help ⇧⌥K
Modify Keyboard Shortcuts...
Global Options... ⌘,
Environment History Connections
To Console To Source read_csv
labs(x = "name")
install.packages("nycflights")
install.packages("nycflights13")
library(readr)
band <- read_csv("project/data/band.csv")
View(band)
Files Plots Packages Help Viewer
New Folder Upload Delete Rename More
Home
Name Size Modified
project
R

Cleaning up


RStudio Cloud
Amelia
Secure | https://rstudio.cloud/project/13632
Your Workspace / Intro to R & RStudio (day 2)
Amelia McNamara
File Edit Code View Plots Session
Console Terminal
R 3.4.2
Connections
To Source
read_csv
Help Viewer
Delete Rename More
Size Modified
Options
General
Code
Appearance
Pane Layout
Packages
R Markdown
Sweave
Spelling
Git/SVN
Publishing
Terminal
Default version of R: (Use System Default)
Restore last used R version for projects
Default working directory (when not in a project):
~
Browse...
Restore most recently opened project at startup
Restore previously open source documents at startup
Restore .RData into workspace at startup
Save workspace to .RData on exit: Ask
Always save history (even when not saving .RData)
Remove duplicate entries in history
Show .Last.value in environment listing
Run Rprofile when resuming suspended session
Use debug error handler only when my code contains errors
Automatically expand tracebacks in error inspector
Wrap around when navigating to previous/next tab
OK
Cancel
Apply

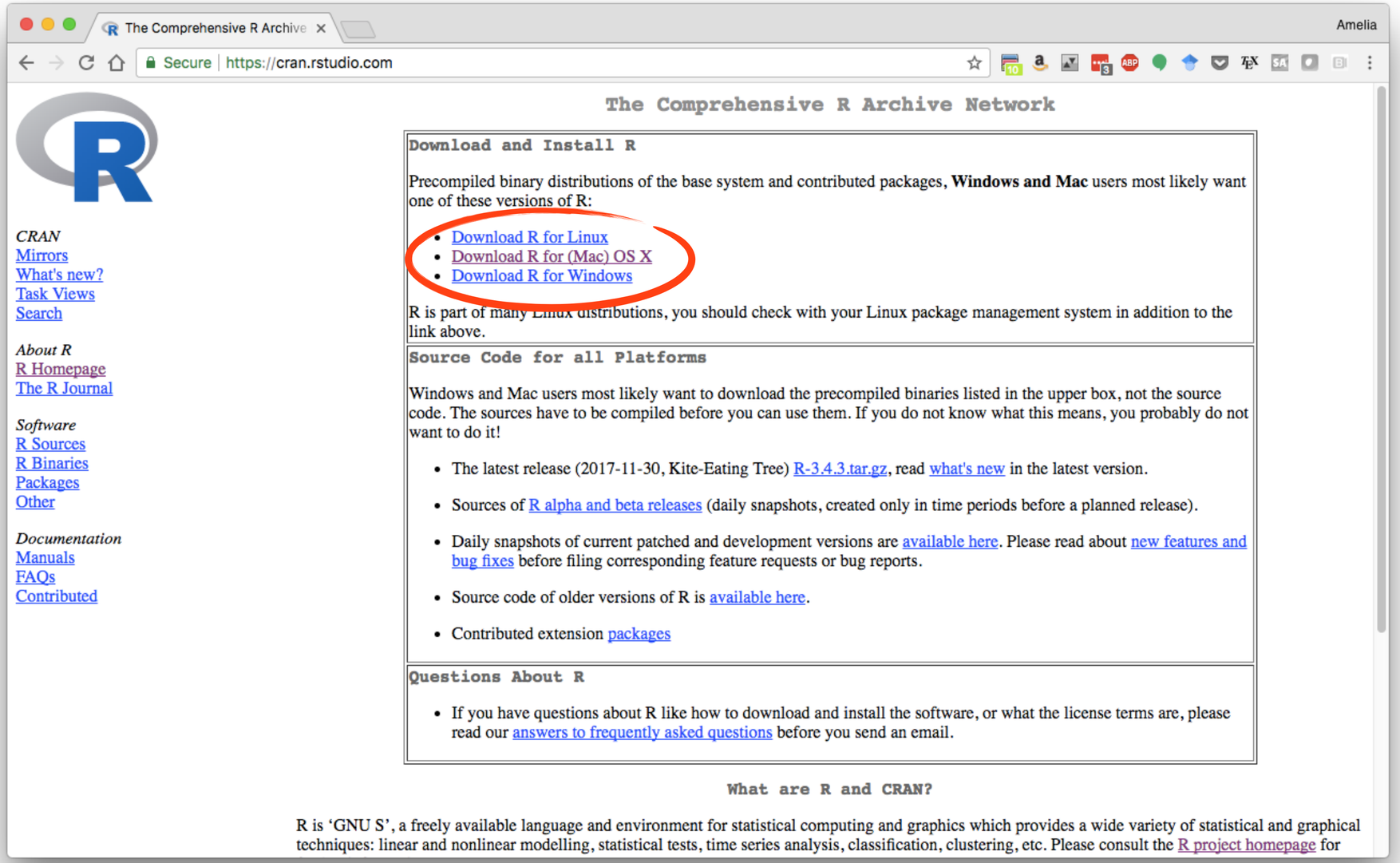
Lots more options!



Installing locally

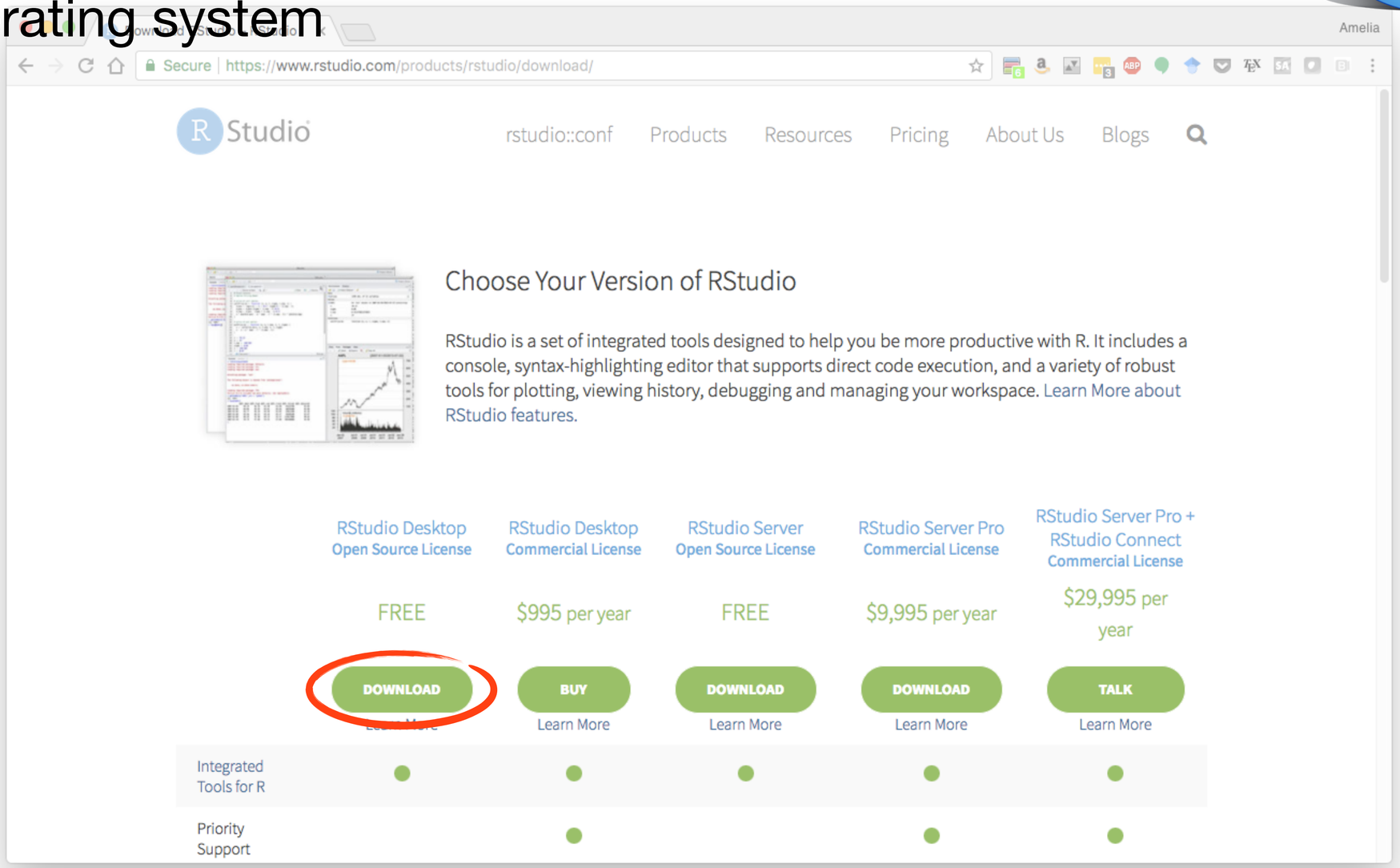
First, you will need to install R (the programming language).

1. Go to https://cran.rstudio.com/
2. Select your operating system



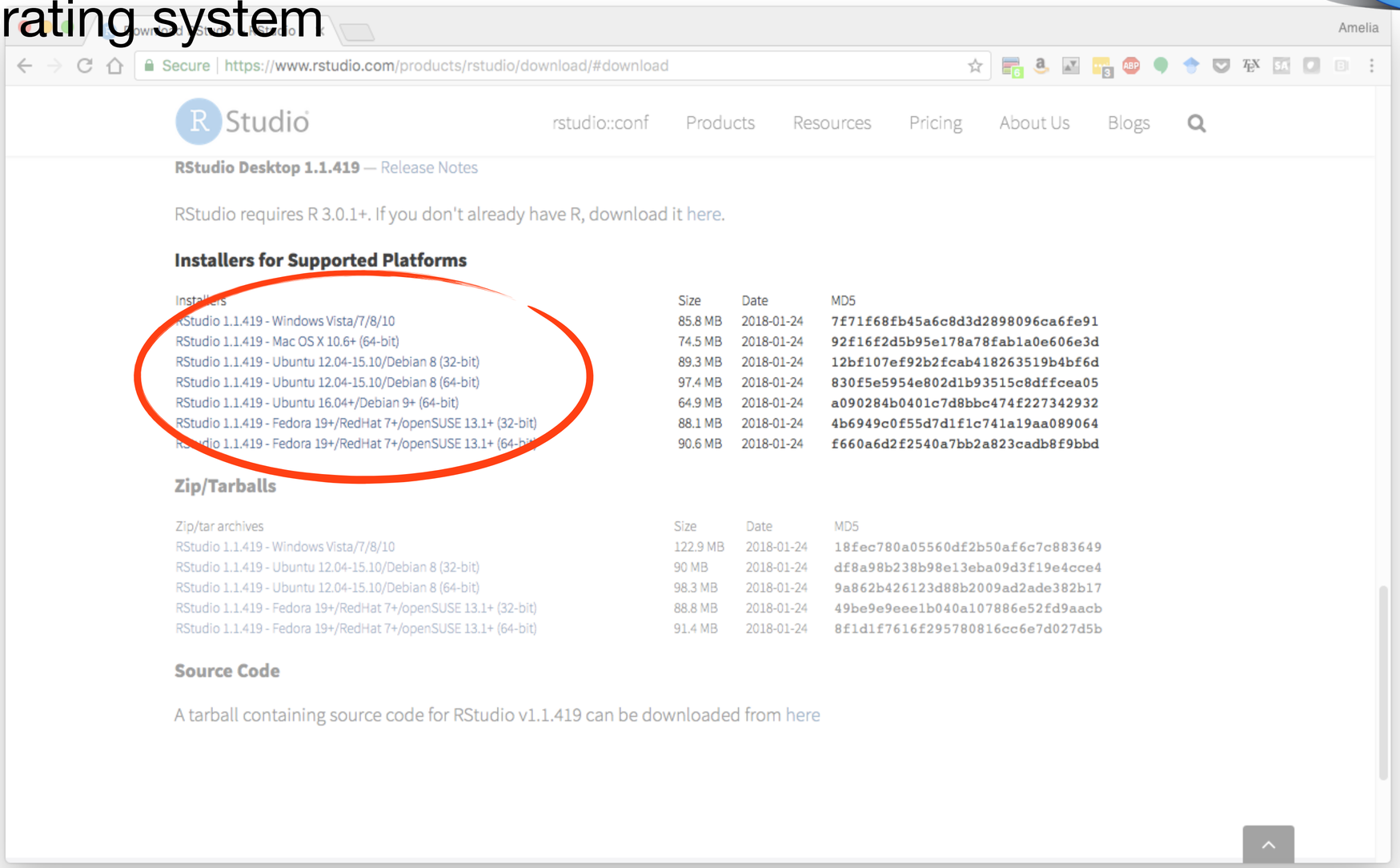
Then, install RStudio (the application).

1. Go to https://www.rstudio.com/products/rstudio/download/
2. Select RStudio desktop
3. Select your operating system



Then, install RStudio (the application).

1. Go to https://www.rstudio.com/products/rstudio/download/
2. Select RStudio desktop
3. Select your operating system



Installing packages

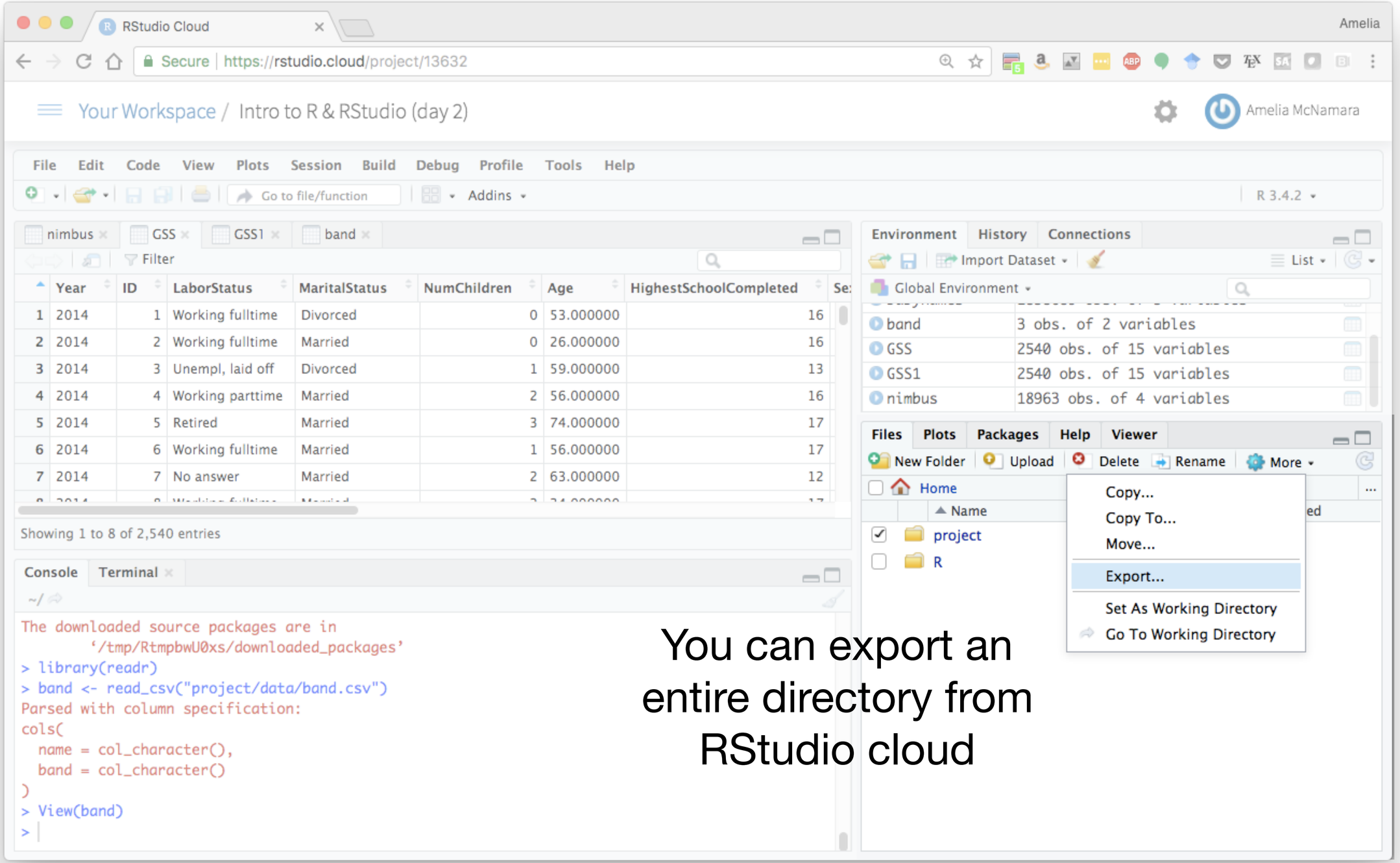
Shortcut to install

- ggplot2, for data visualisation.
- dplyr, for data manipulation.
- tidyr, for data tidying.
- readr, for data import.
- purrr, for functional programming.
- tibble, for tibbles, a modern re-imagining of data frames.

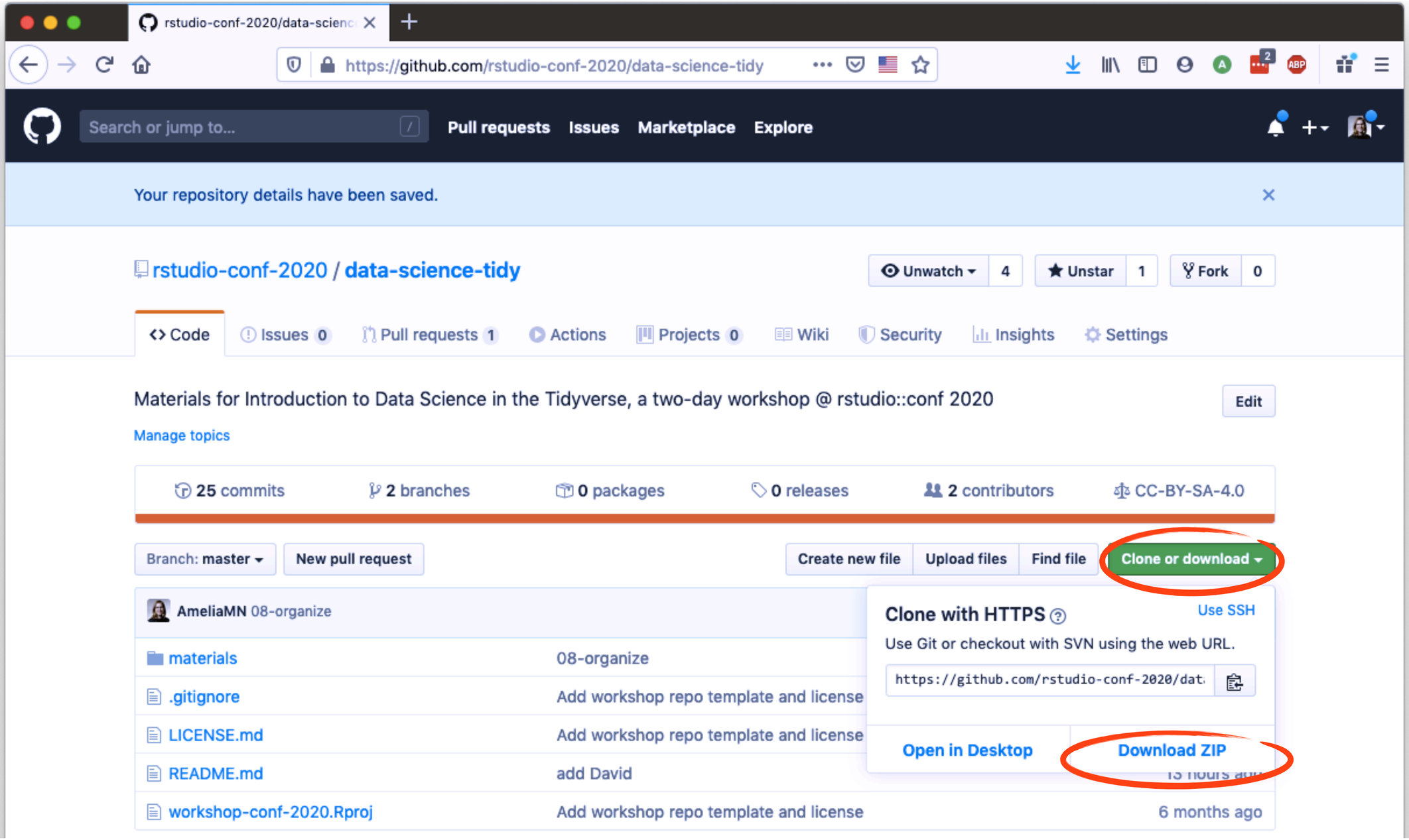
And more

```
install.packages(c("babynames", "fivethirtyeight", "formatR",
"gapminder", "hexbin", "mgcv", "maps", "mapproj", "nycflights13",
"rmarkdown", "skimr", "tidyverse", "viridis"))
```

Getting our code

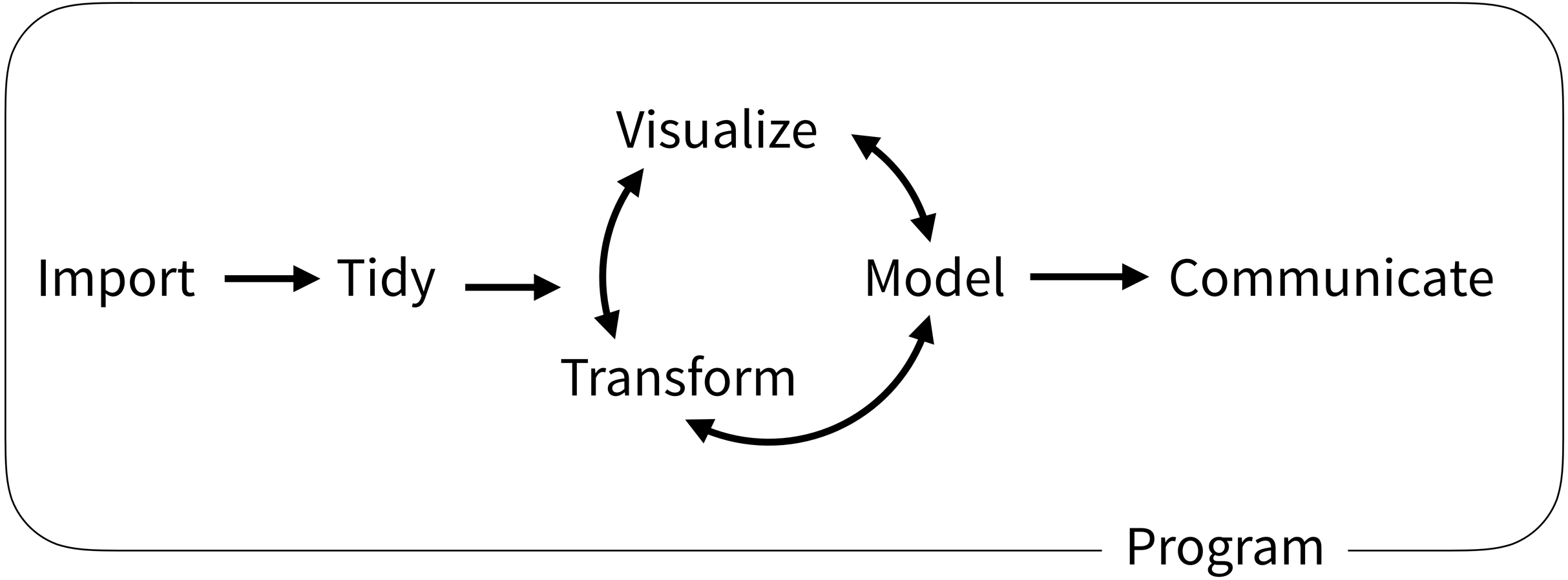

RStudio Cloud
Amelia
Secure | https://rstudio.cloud/project/13632
Your Workspace / Intro to R & RStudio (day 2)
Amelia McNamara
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 3.4.2
nimbus
GSS
GSS1
band
Filter
Year ID LaborStatus MaritalStatus NumChildren Age HighestSchoolCompleted
1 2014 1 Working fulltime Divorced 0 53.000000 16
2 2014 2 Working fulltime Married 0 26.000000 16
3 2014 3 Unempl, laid off Divorced 1 59.000000 13
4 2014 4 Working parttime Married 2 56.000000 16
5 2014 5 Retired Married 3 74.000000 17
6 2014 6 Working fulltime Married 1 56.000000 17
7 2014 7 No answer Married 2 63.000000 12
Showing 1 to 8 of 2,540 entries
Console
Terminal
~/
The downloaded source packages are in
'/tmp/RtmpbwU0xs/downloaded_packages'
> library(readr)
> band <- read_csv("project/data/band.csv")
Parsed with column specification:
cols(
name = col_character(),
band = col_character()
)
> View(band)
>
Environment
History
Connections
Import Dataset
List
Global Environment
band 3 obs. of 2 variables
GSS 2540 obs. of 15 variables
GSS1 2540 obs. of 15 variables
nimbus 18963 obs. of 4 variables
Files
Plots
Packages
Help
Viewer
New Folder
Upload
Delete
Rename
More
Home
Name
project
R
Copy...
Copy To...
Move...
Export...
Set As Working Directory
Go To Working Directory
You can export an entire directory from RStudio cloud

Or, download a clean version from https://github.com/rstudio-conf-2020/data-science-tidy



More to learn

(Applied) **Data Science**



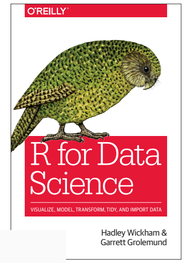


Table of contents

R for Data Science

Welcome

1 Introduction

I Explore

2 Introduction

3 Data visualisation

4 Workflow: basics

5 Data transformation

6 Workflow: scripts

7 Exploratory Data Analysis

8 Workflow: projects

II Wrangle

9 Introduction

10 Tibbles

11 Data import

12 Tidy data

13 Relational data

14 Strings

15 Factors

16 Dates and times

III Program

17 Introduction

18 Pipes

19 Functions

20 Vectors

21 Iteration

IV Model

22 Introduction

23 Model basics

24 Model building

25 Many models

V Communicate

26 Introduction

27 R Markdown

28 Graphics for communication

29 R Markdown formats

30 R Markdown workflow

Review things we've covered

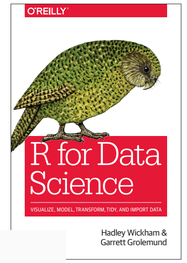


Table of contents

R for Data Science

Welcome

1 Introduction

I Explore

2 Introduction

3 Data visualisation

4 Workflow: basics

5 Data transformation

6 Workflow: scripts

7 Exploratory Data Analysis

8 Workflow: projects

II Wrangle

9 Introduction

10 Tibbles

11 Data import

12 Tidy data

13 Relational data

14 Strings

15 Factors

16 Dates and times

III Program

17 Introduction

18 Pipes

19 Functions

20 Vectors

21 Iteration

IV Model

22 Introduction

23 Model basics

24 Model building

25 Many models

V Communicate

26 Introduction

27 R Markdown

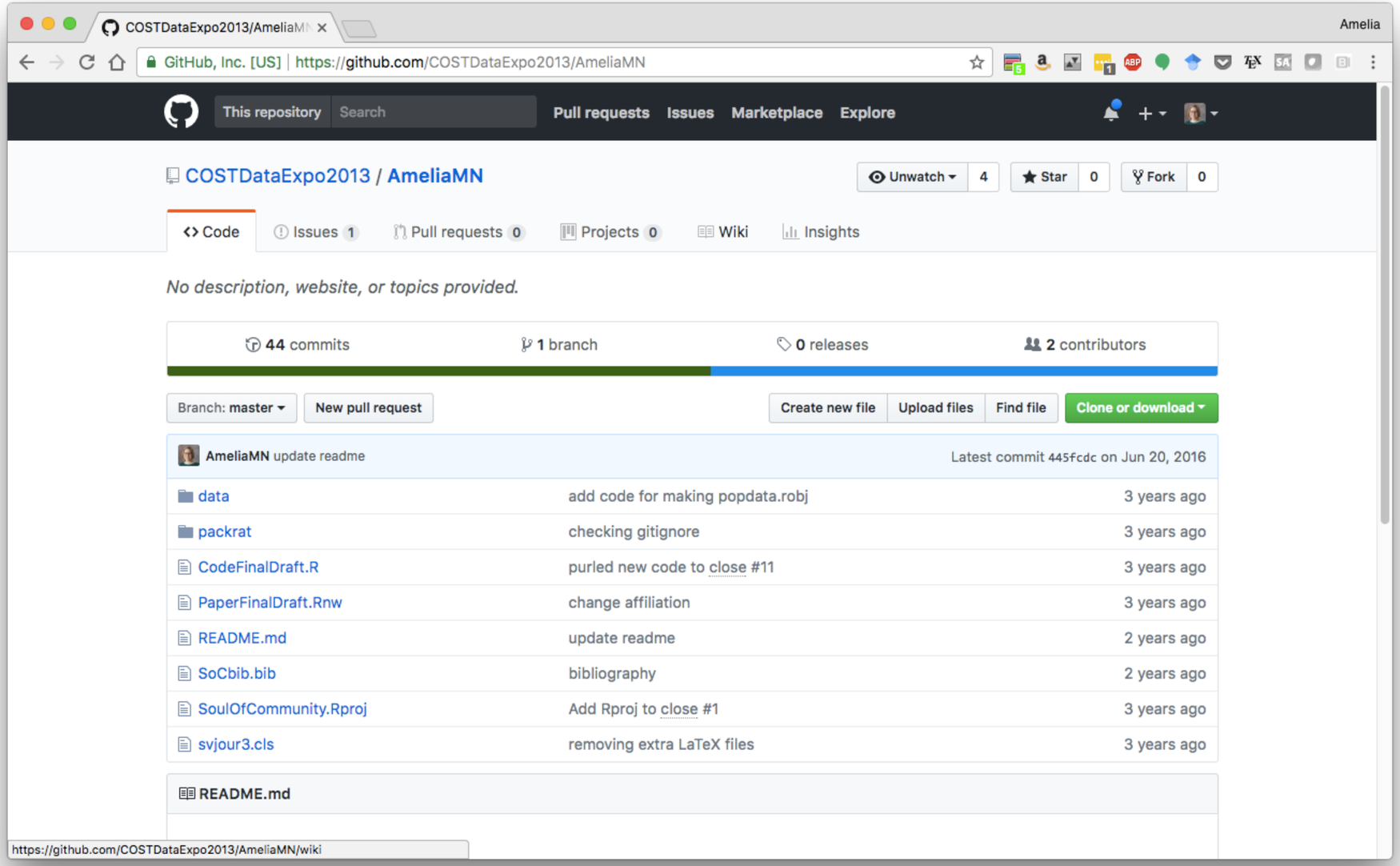
28 Graphics for communication

29 R Markdown formats

30 R Markdown workflow

Generally useful things

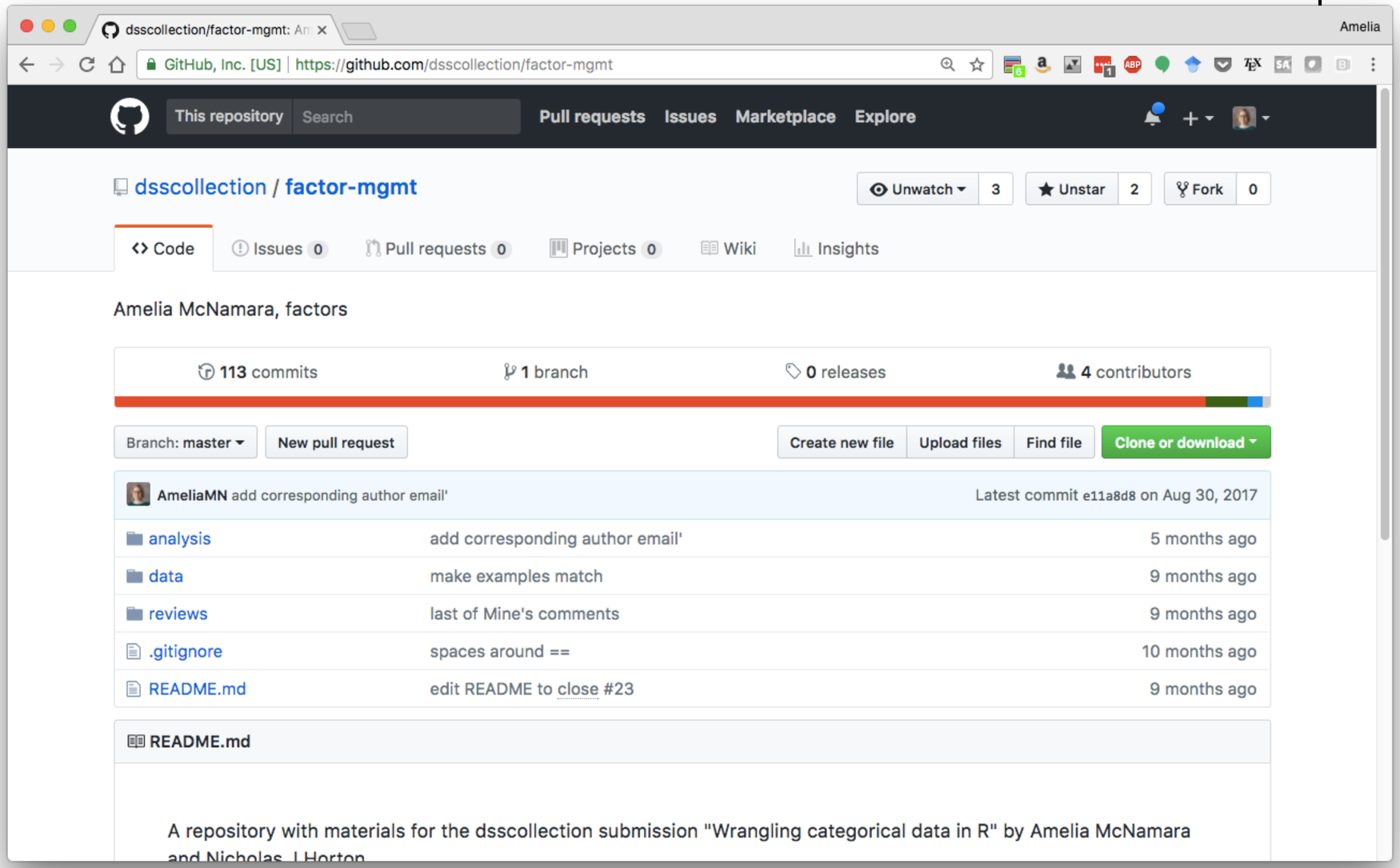
Example paper and file structure: https://github.com/COSTDataExpo2013/AmeliaMN



Another example

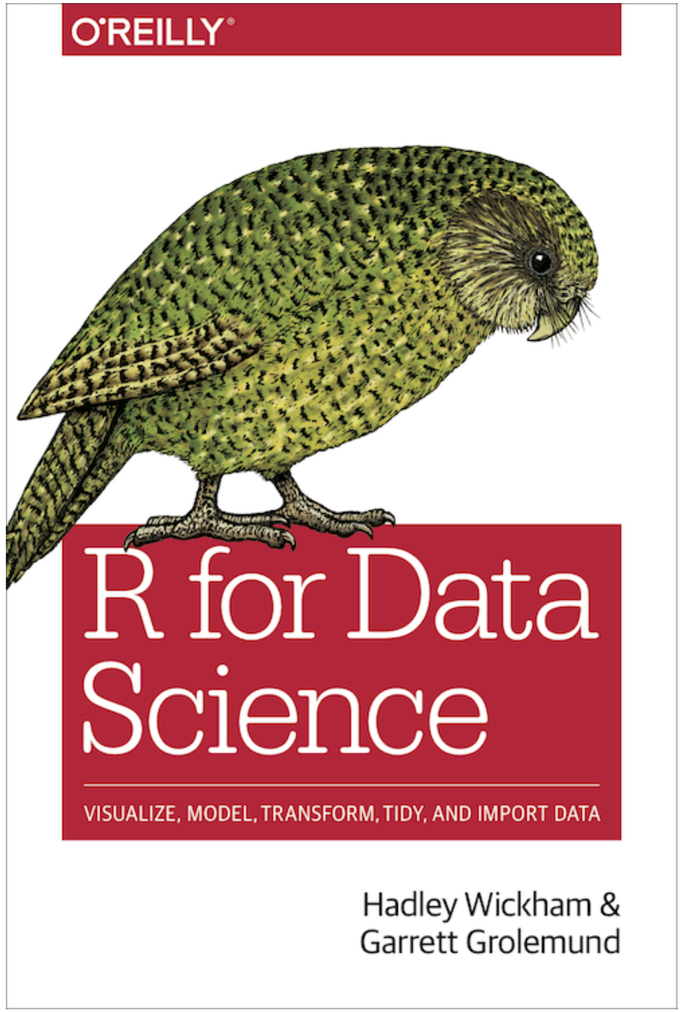
https://github.com/dsscollection/factor-mgmt

Bonus— this has a ton of info on factor variables and their pitfalls!



Books

- Elements of Data Analytic Style, Jeff Leek
- R Programming for Data Science, Roger Peng
-The Art of Data Science, Roger Peng
- R Cookbook. Both a website, and a book, Winston Chang
- R for Data Science. Both a website and a book. Hadley Wickham and Garrett Grolemund.



Online learning/courses

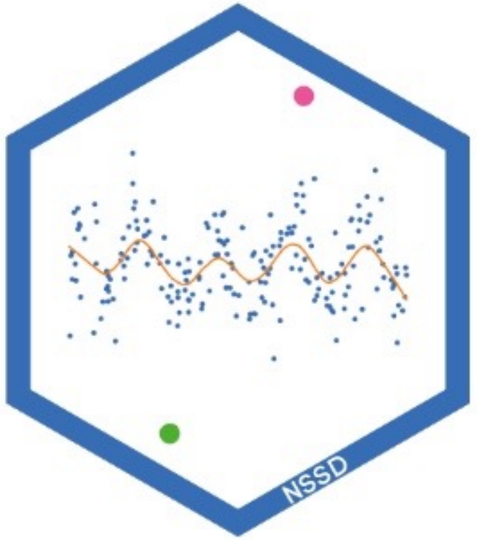
- Johns Hopkins Coursera Course on R. Part of the Data Science specialization. Courses are free, but the certificate costs money.

-

Blogs, etc.

- Simply statistics, blog by Roger Peng, Jeff Leek, and Rafa Irizarry

- Not so standard deviations podcast by Hilary Parker and Roger Peng



- https://rweekly.org/, open-sourced aggregator of all things R

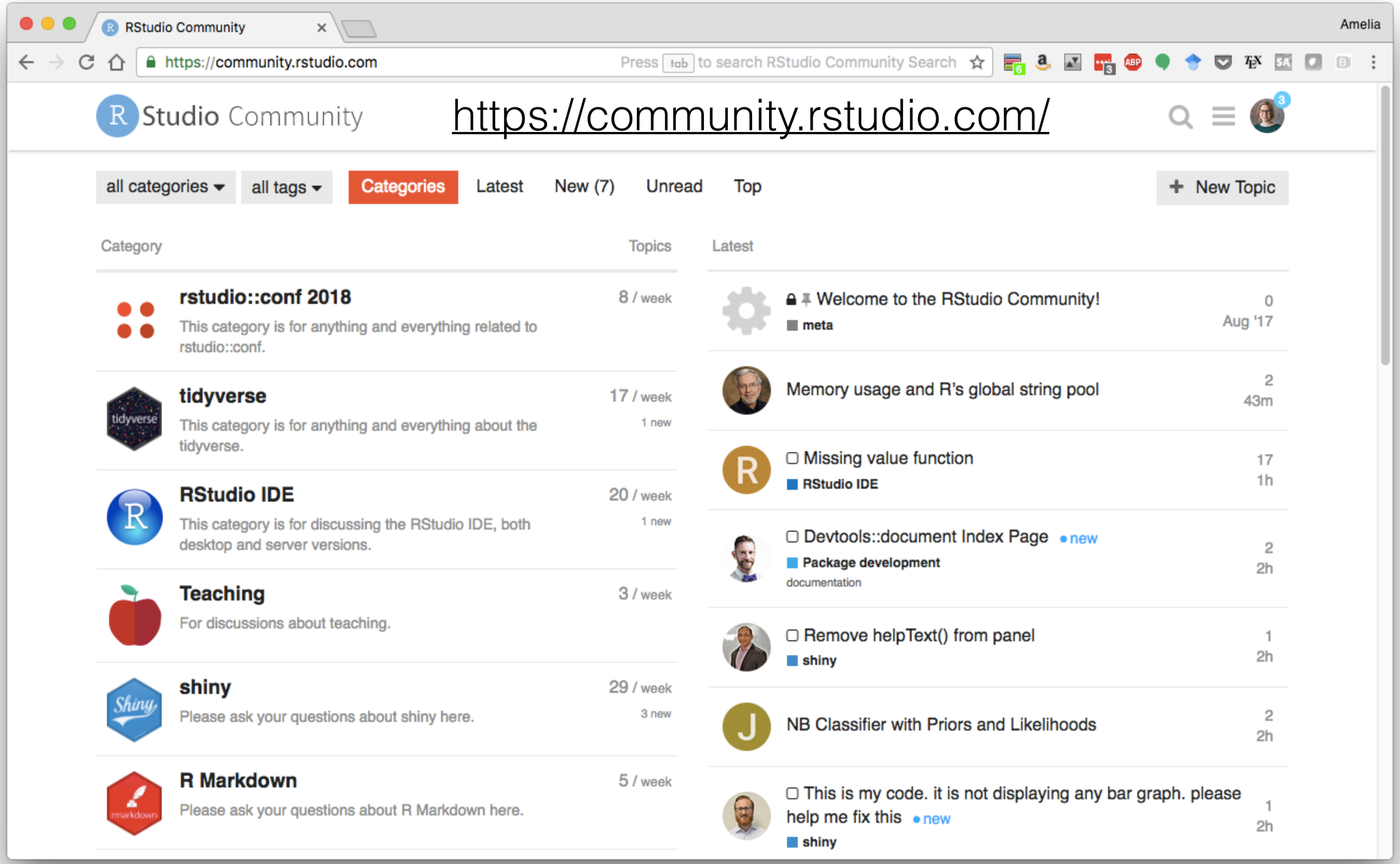
Twitter

Who to follow

- me! Amelia McNamara, University of St Thomas
- Hadley Wickham, RStudio
- Jenny Bryan, on leave from UBC, at RStudio
- Hilary Parker, data scientist at StitchFix
- Roger Peng, biostatistician at JHU
- Jeff Leek, biostatistician at JHU
- David Robinson, formerly of StackOverflow, now DataCamp
- Karl Broman, biostatistician at UW
- Karthik Ram, rOpenSci
- Renee Teate, BecomingDataSci
- Mine Cetinkaya-Rundel, Duke, RStudio
- Julia Silge, tidytext, StackOverflow

Hashtags:

- #rstats
- #tidyverse
- #rcatladies


RStudio Community
Amelia
https://community.rstudio.com
Press tab to search RStudio Community Search
RStudio Community
https://community.rstudio.com/
all categories
all tags
Categories
Latest
New (7)
Unread
Top
New Topic
Category
Topics
Latest
rstudio::conf 2018
This category is for anything and everything related to rstudio::conf.
8 / week
tidyverse
This category is for anything and everything about the tidyverse.
17 / week
1 new
RStudio IDE
This category is for discussing the RStudio IDE, both desktop and server versions.
20 / week
1 new
Teaching
For discussions about teaching.
3 / week
shiny
Please ask your questions about shiny here.
29 / week
3 new
R Markdown
Please ask your questions about R Markdown here.
5 / week
Welcome to the RStudio Community!
meta
0
Aug '17
Memory usage and R's global string pool
2
43m
Missing value function
RStudio IDE
17
1h
Devtools::document Index Page
new
Package development
documentation
2
2h
Remove helpText() from panel
shiny
1
2h
NB Classifier with Priors and Likelihoods
2
2h
This is my code. it is not displaying any bar graph. please help me fix this
new
shiny
1
2h

Getting help

Searching for help

- Official word on learning more: https://www.tidyverse.org/learn/
- We've seen the R help functions `?` and `help()`
- Google, putting in R as a search term (Google recognizes it now!)
- Search on http://stackoverflow.com/ (add keywords like tidyverse)

Physical communities

- There are R meetups in many major cities
- If you are a gender minority, check out R-ladies meetups

Online communities

- R4DS learning community: https://medium.com/@kierisi/r4ds-the-next-iteration-d51e0a1b0b82
- https://community.rstudio.com/ is intentionally friendly to beginners!
- Asking on http://stackoverflow.com/ is perhaps an intermediate skill
- I don't recommend asking on R-help
- Official word on asking for help: https://www.tidyverse.org/help/

Thanks to our fantastic TAs!

Jesse Mostipak (she/her)	@kierisi
Ben Baumer (he/him)	@BaumerBen
Matthew Flickinger (he/him)	@EmEmEff
Mike Smith (he/him)	@MikeKSmith
David Keyes (he/him)	@dgkeyes


Thank you!