Introduction to R & RStudio:

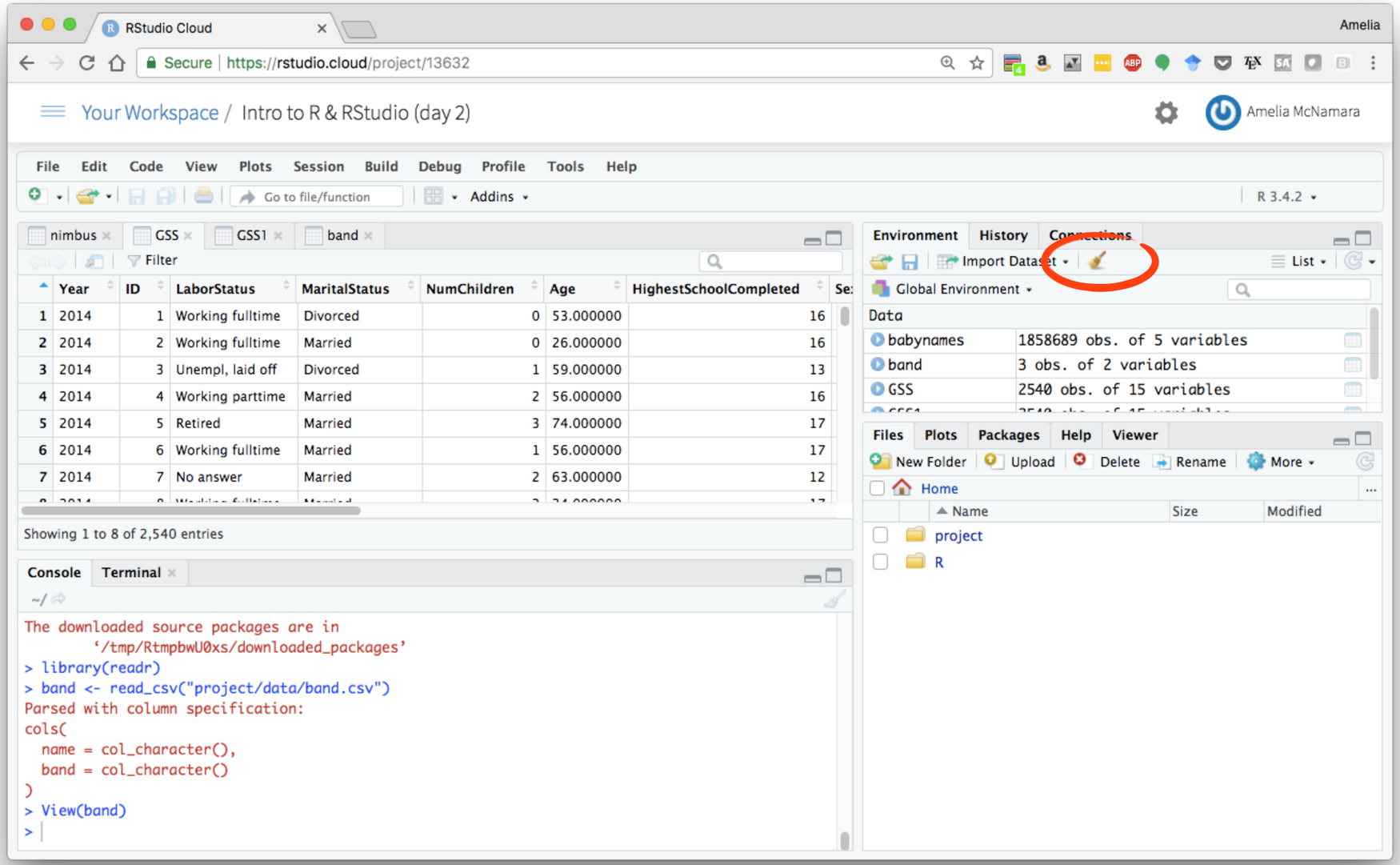
deck 08: Best practices

Amelia McNamara
Visiting Assistant Professor of Statistical and Data Sciences
Smith College

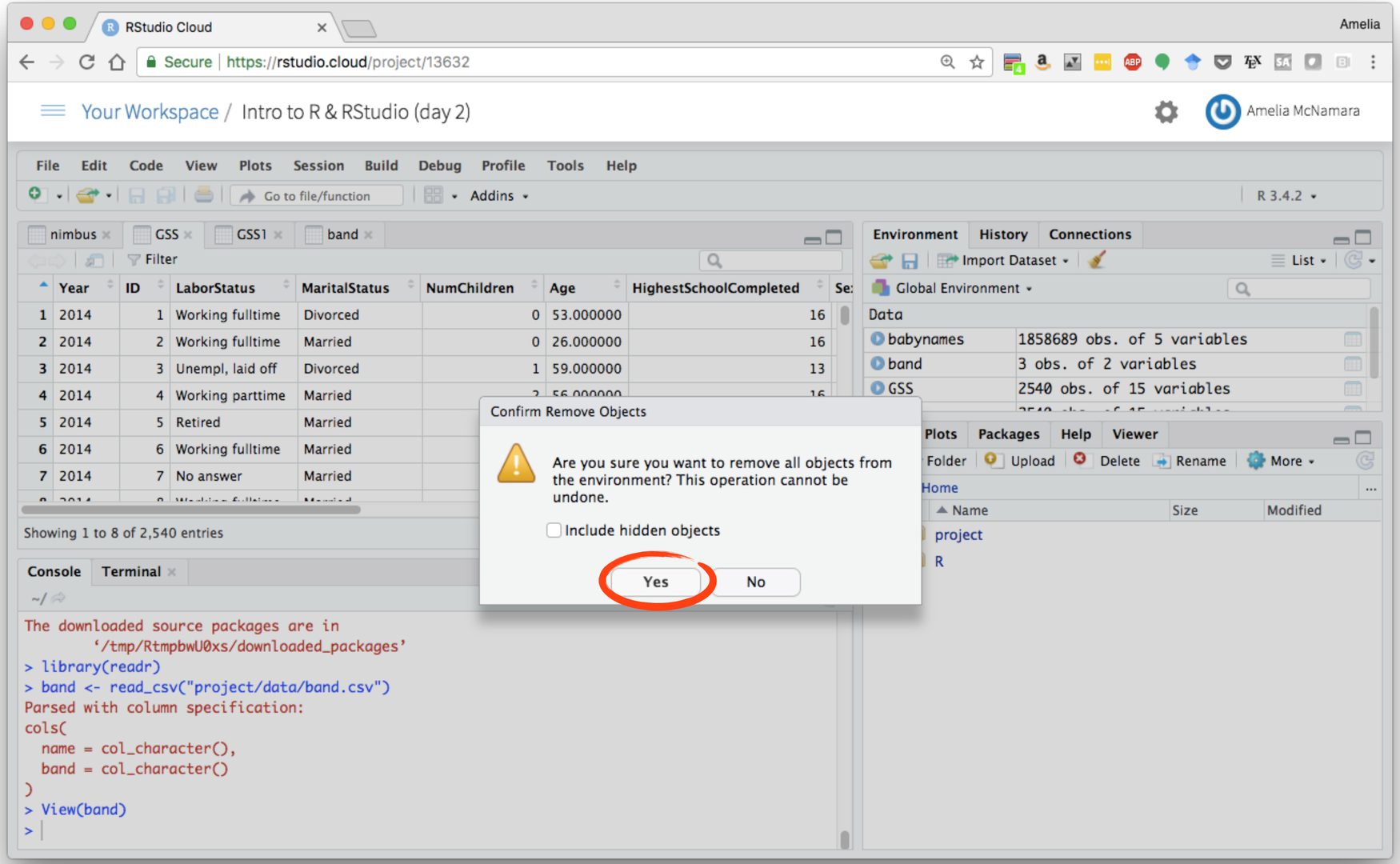
January 2018

RStudio

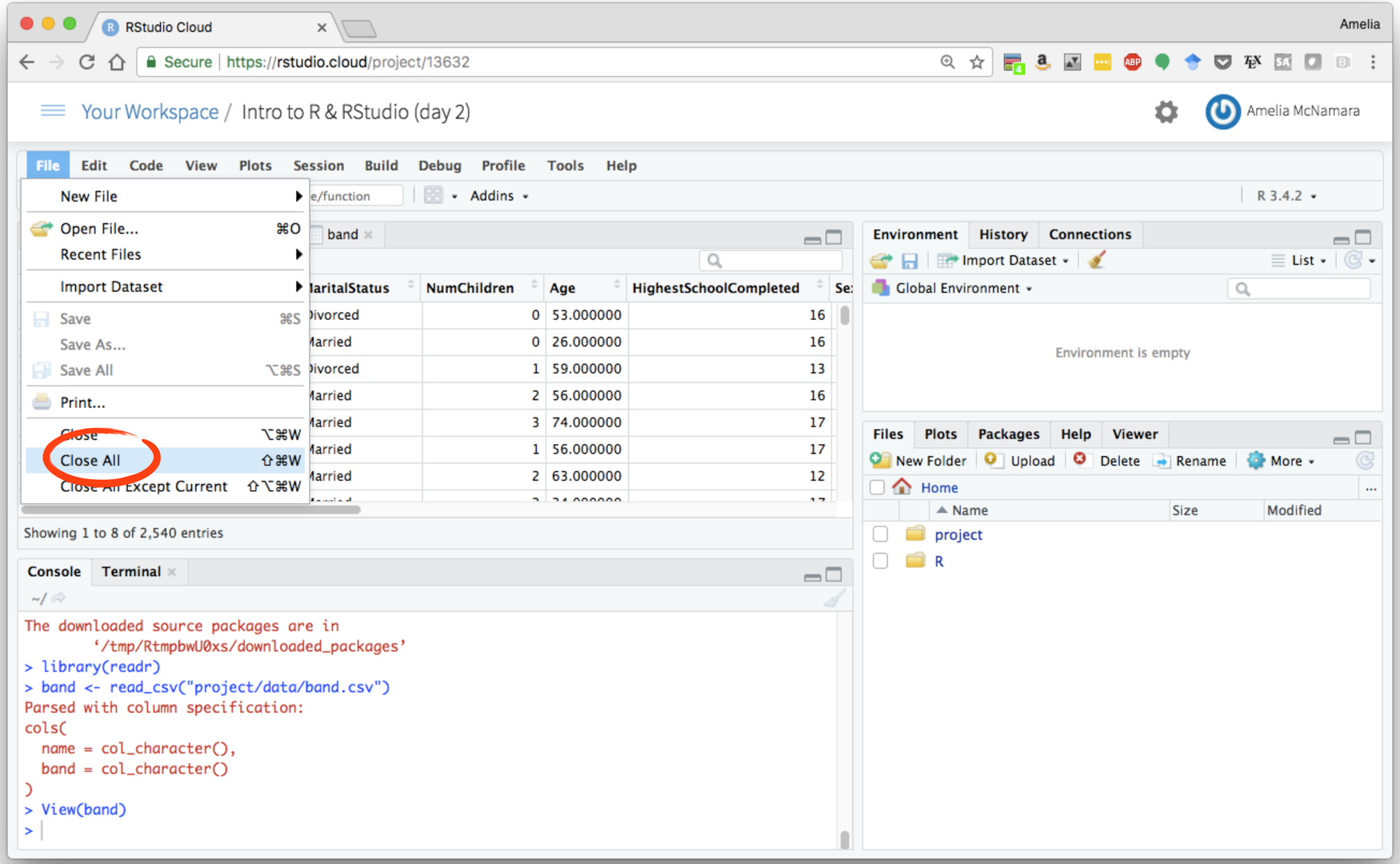
Cleaning up


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
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
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

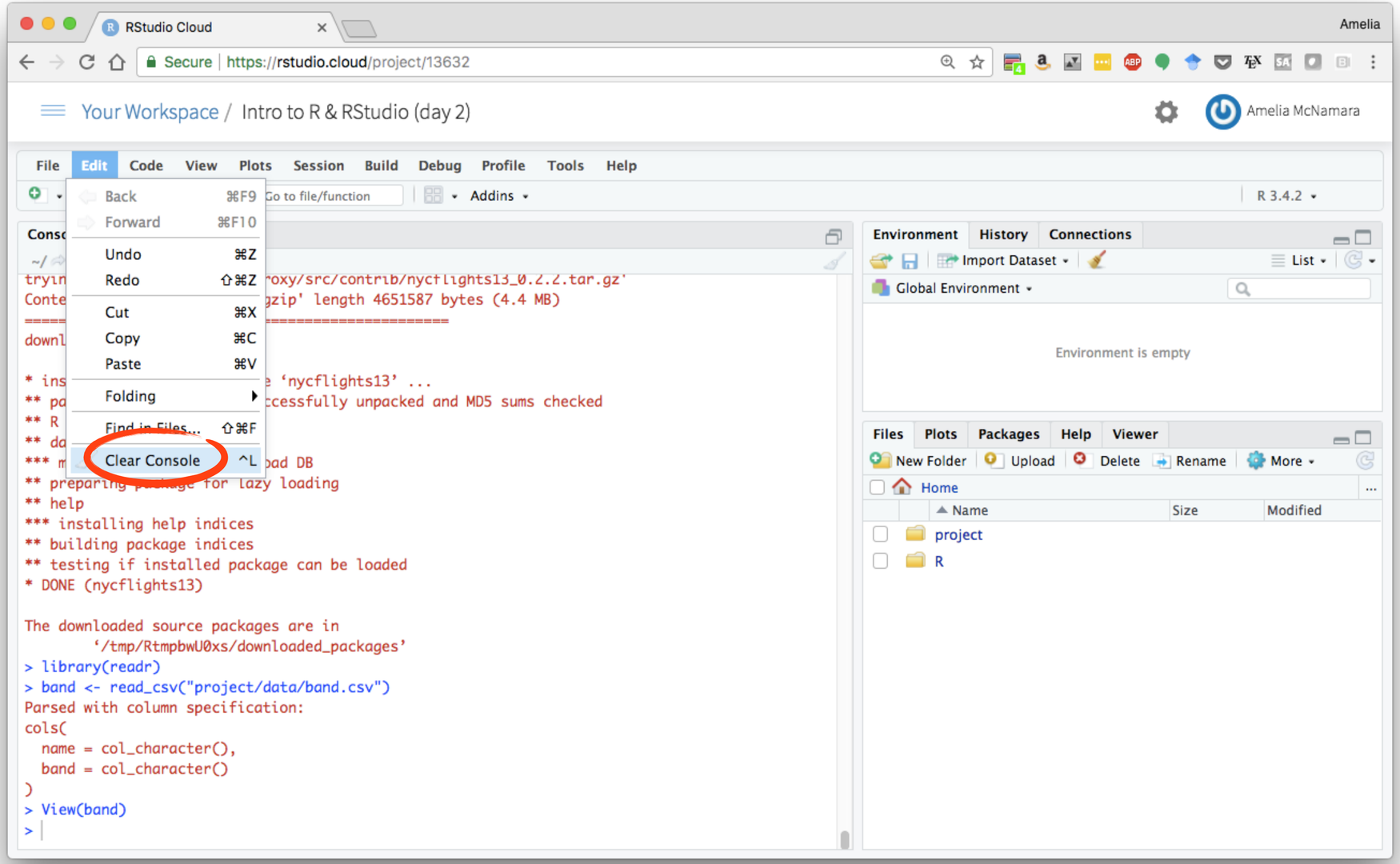
Cleaning up


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
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
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
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)
>

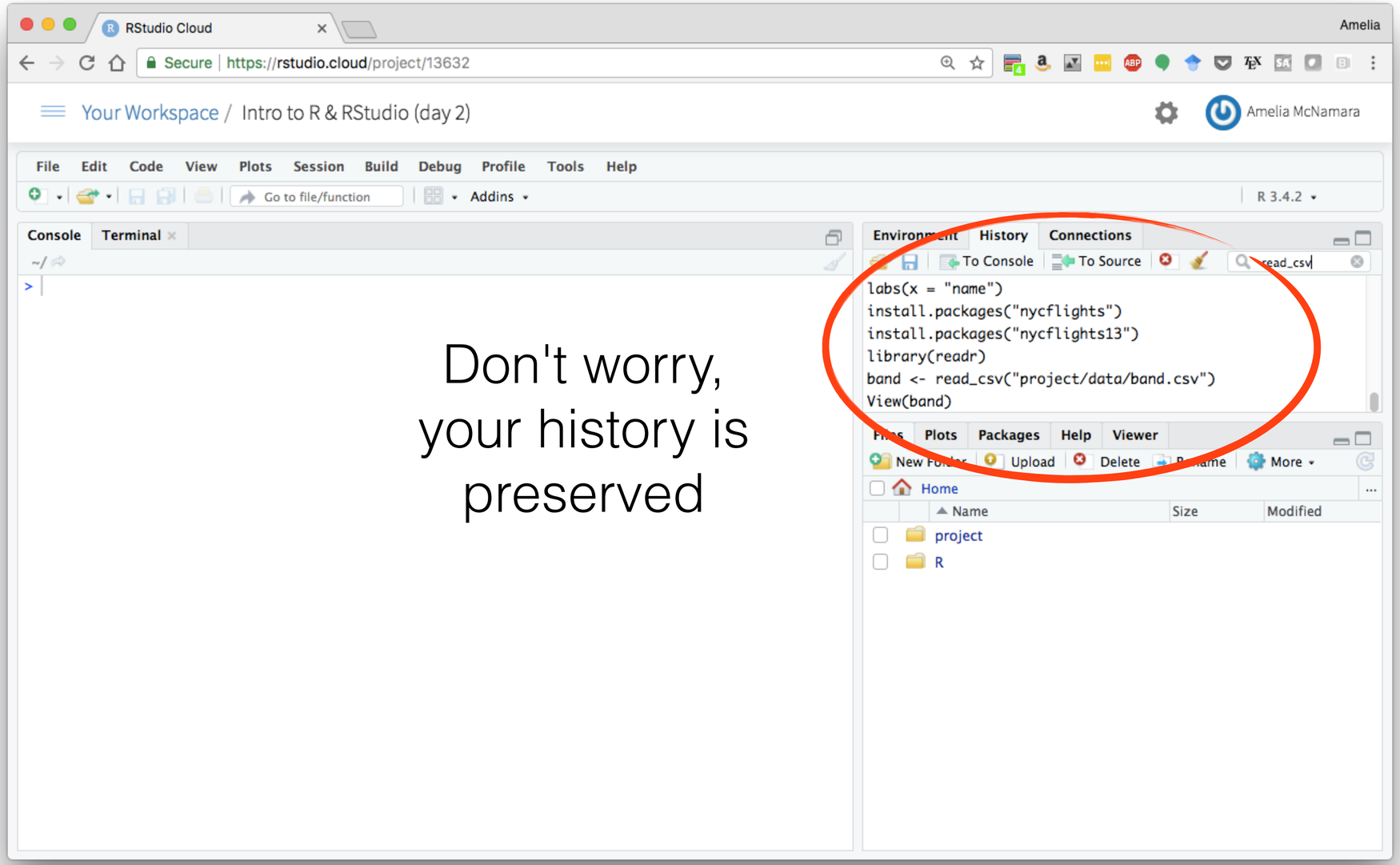
Cleaning up


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
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
Go to file/function
Addins
R 3.4.2
roxy/src/contrib/nycflights13_0.2.2.tar.gz'
gzip' length 4651587 bytes (4.4 MB)
* ins
** pa
** R
** da
*** m
** preparing package for lazy loading
** help
*** 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)
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

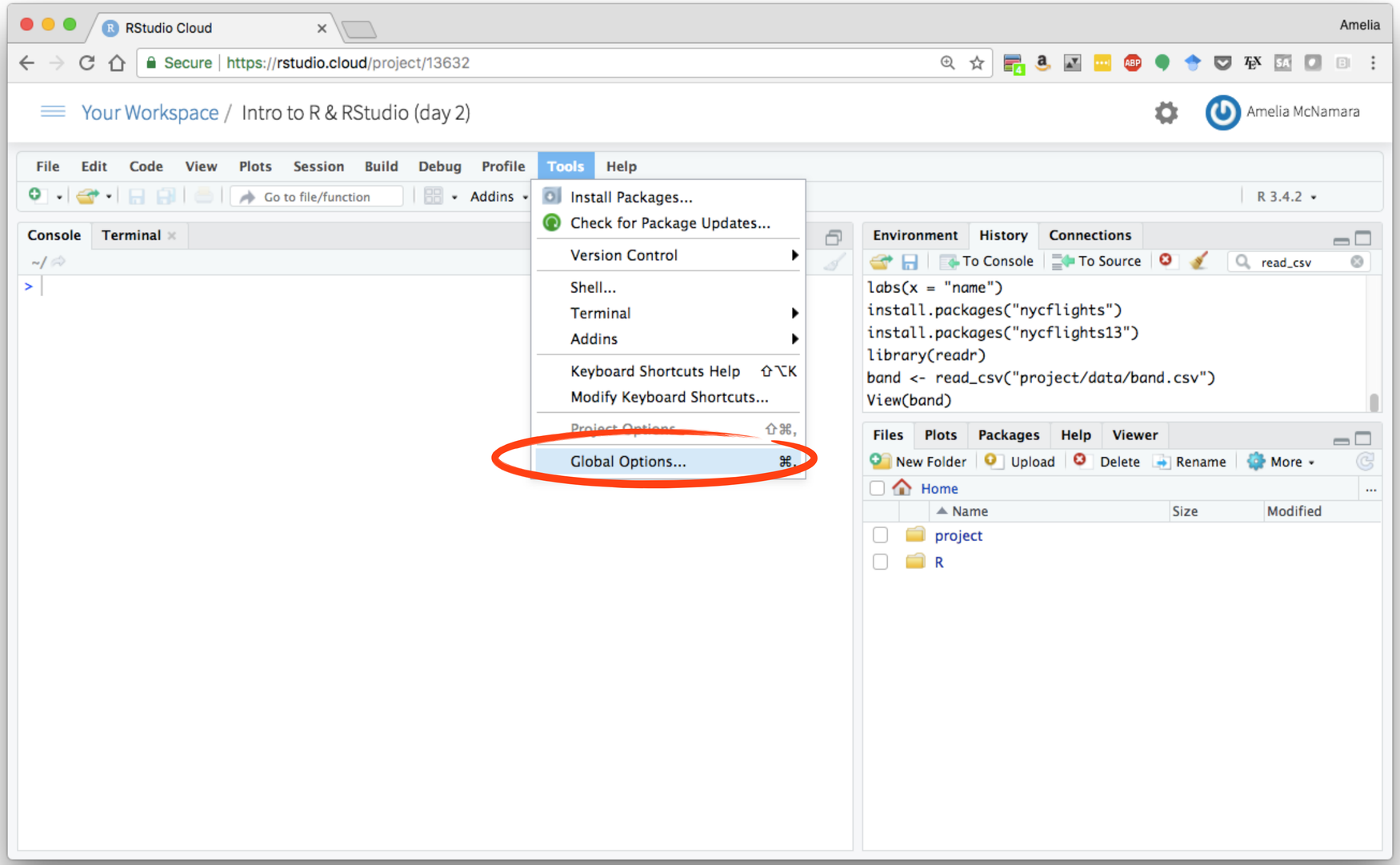
Cleaning up



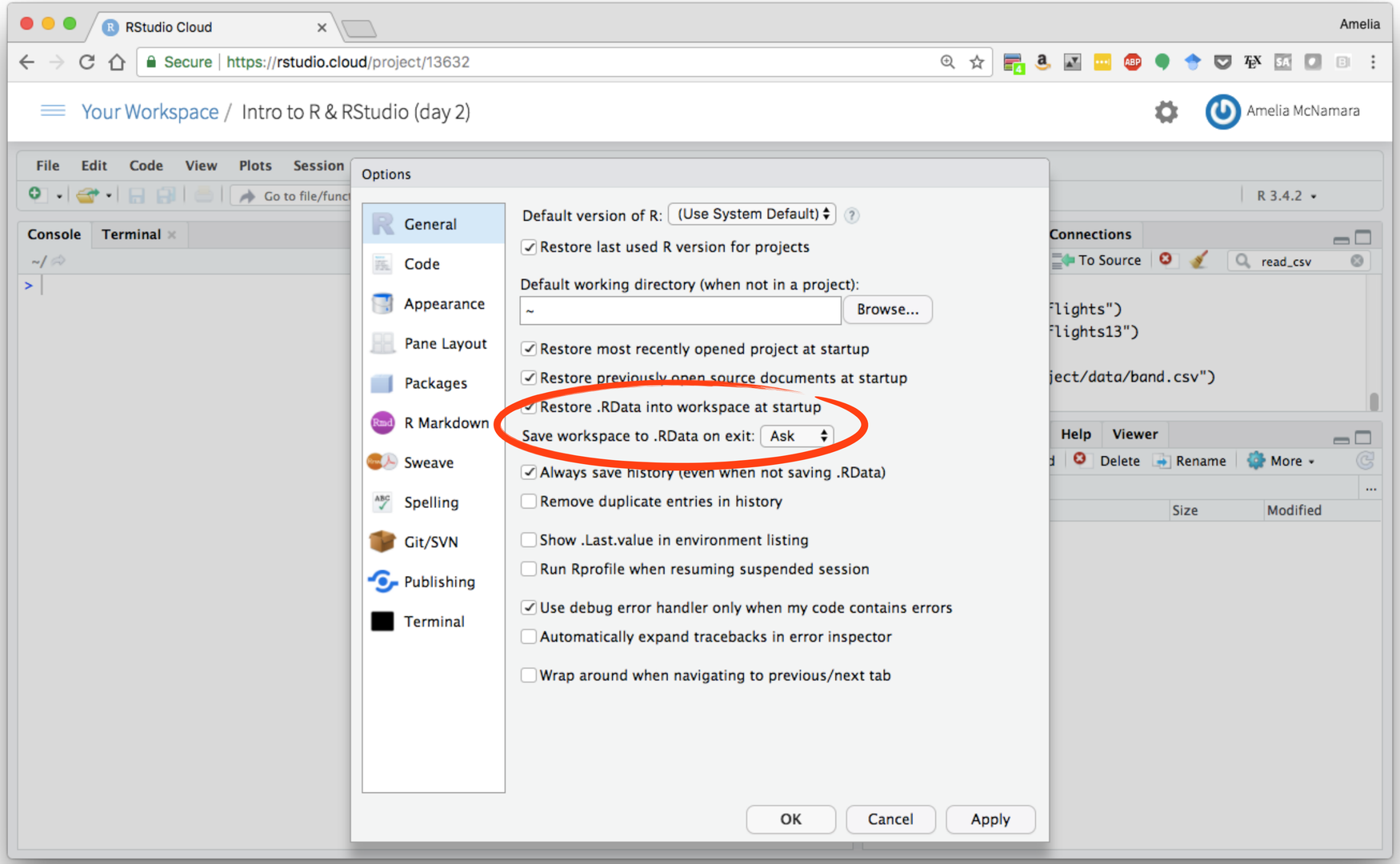
Don't worry, your history is preserved

Settings

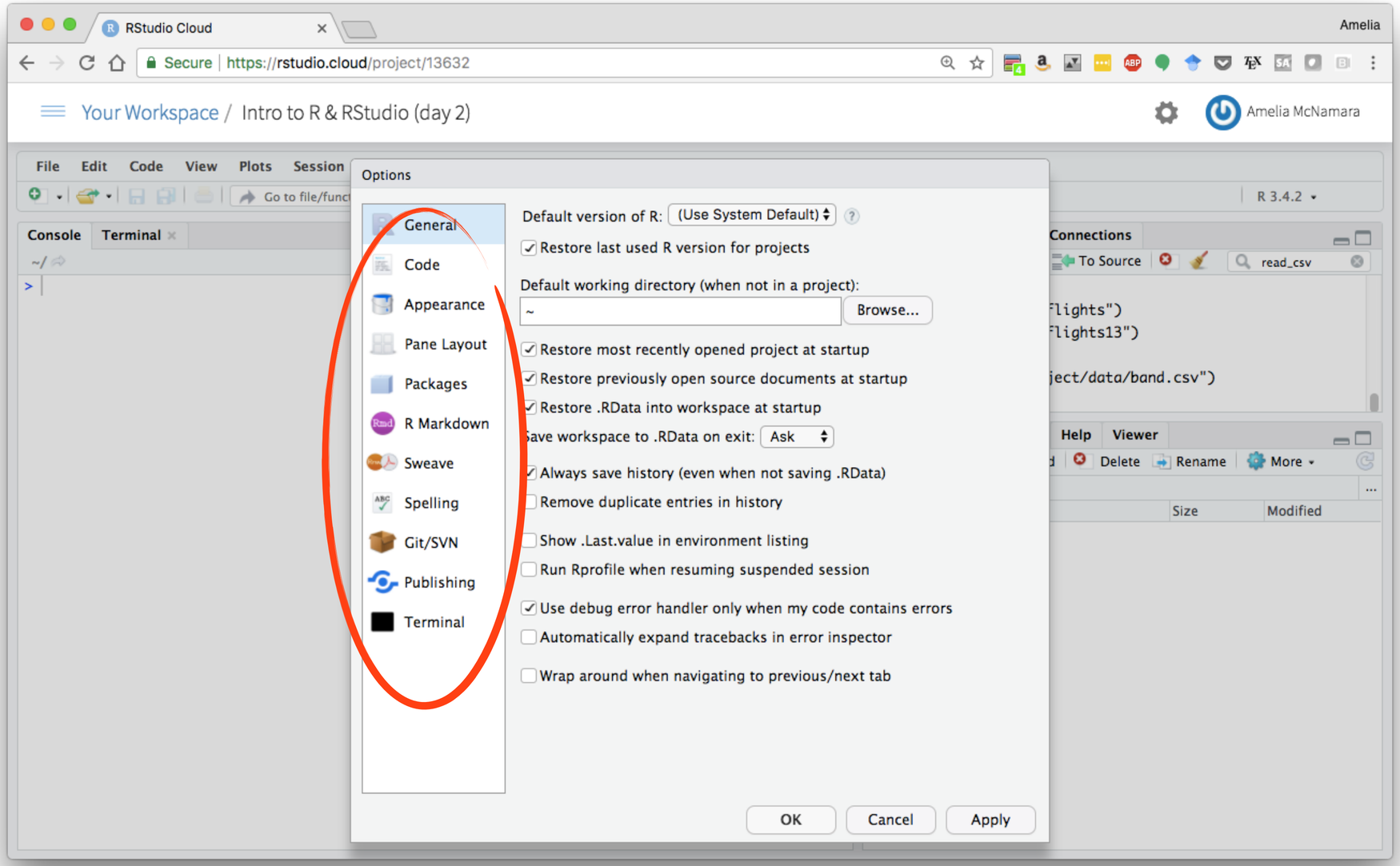
Cleaning up


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
Install Packages...
Check for Package Updates...
Version Control
Shell...
Terminal
Addins
Keyboard Shortcuts Help ⇧⌥K
Modify Keyboard Shortcuts...
Project Options... ⇧⌘,
Global Options... ⌘,
Console Terminal
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!



Collaborating

Spreadsheets

‹ All Collections

Collection idea for us?

Practical Data Science for Stats - a PeerJ Collection

Data Science | Statistics | Scientific Computing and Simulation | Computer Education | Computational Science

Social Computing | Software Engineering | Science and Medical Education | Computational Biology

Human-Computer Interaction | Anthropology | Programming Languages | Visual Analytics | Graphics

Data Mining and Machine Learning

Practical Data Science for Stats

The "Practical Data Science for Stats" Collection contains preprints focusing on the practical side of data science workflows and statistical analysis. Curated by Jennifer Bryan and Hadley Wickham.

Karl Broman, Kara Woo. Data organization in spreadsheets. PeerJ preprint and The American Statistician.

https://peerj.com/preprints/3183/

Abstract: Spreadsheets are widely used software tools for data entry, storage, analysis, and visualization. Focusing on the data entry and storage aspects, this paper offers practical recommendations for organizing spreadsheet data to reduce errors and ease later analyses. The basic principles are: be consistent, write dates like YYYY-MM-DD, don't leave any cells empty, put just one thing in a cell, organize the data as a single rectangle (with subjects as rows and variables as columns, and with a single header row), create a data dictionary, don't include calculations in the raw data files, don't use font color or highlighting as data, choose good names for things, make backups, use data validation to avoid data entry errors, and save the data in plain text file.

https://github.com/jtleek/datasharing

jtleek/datasharing: The Leek g

Amelia

GitHub, Inc. [US] | https://github.com/jtleek/datasharing

This repository Search Pull requests Issues Marketplace Explore

jtleek / datasharing

Watch 576 Star 4,252 Fork 182,181

Code Issues 18 Pull requests 408 Projects 0 Wiki Insights

The Leek group guide to data sharing

29 commits 1 branch 0 releases 10 contributors

Branch: master New pull request Create new file Upload files Find file Clone or download

jtleek committed on Nov 8, 2016 Merge pull request #464 from Amherst-Statistics/master Latest commit df97230 on Nov 8, 2016

README.md Offered suggestions to Jeff for the TAS DSS submission. a year ago

README.md

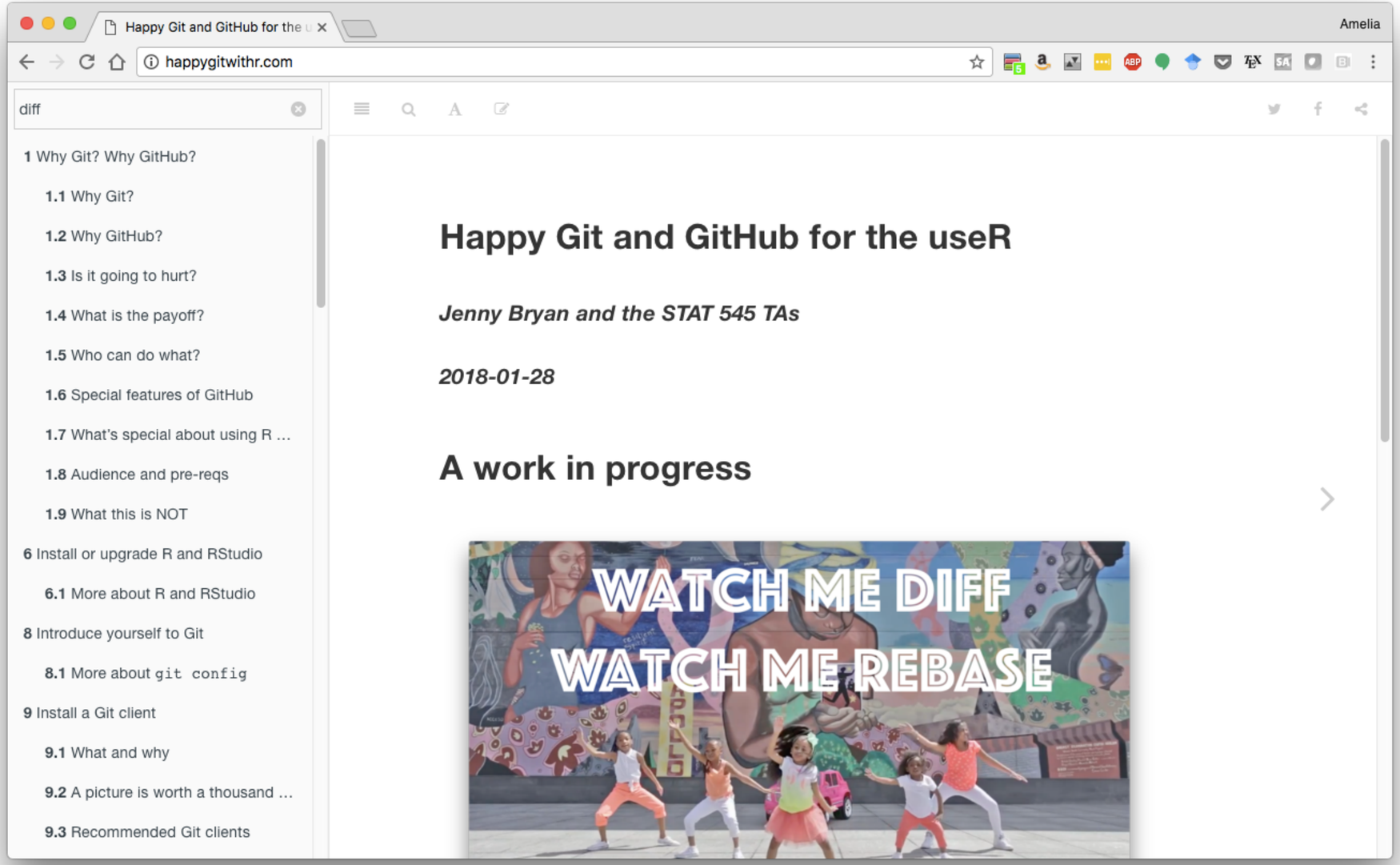
How to share data with a statistician

This is a guide for anyone who needs to share data with a statistician or data scientist. The target audiences I have in mind are:

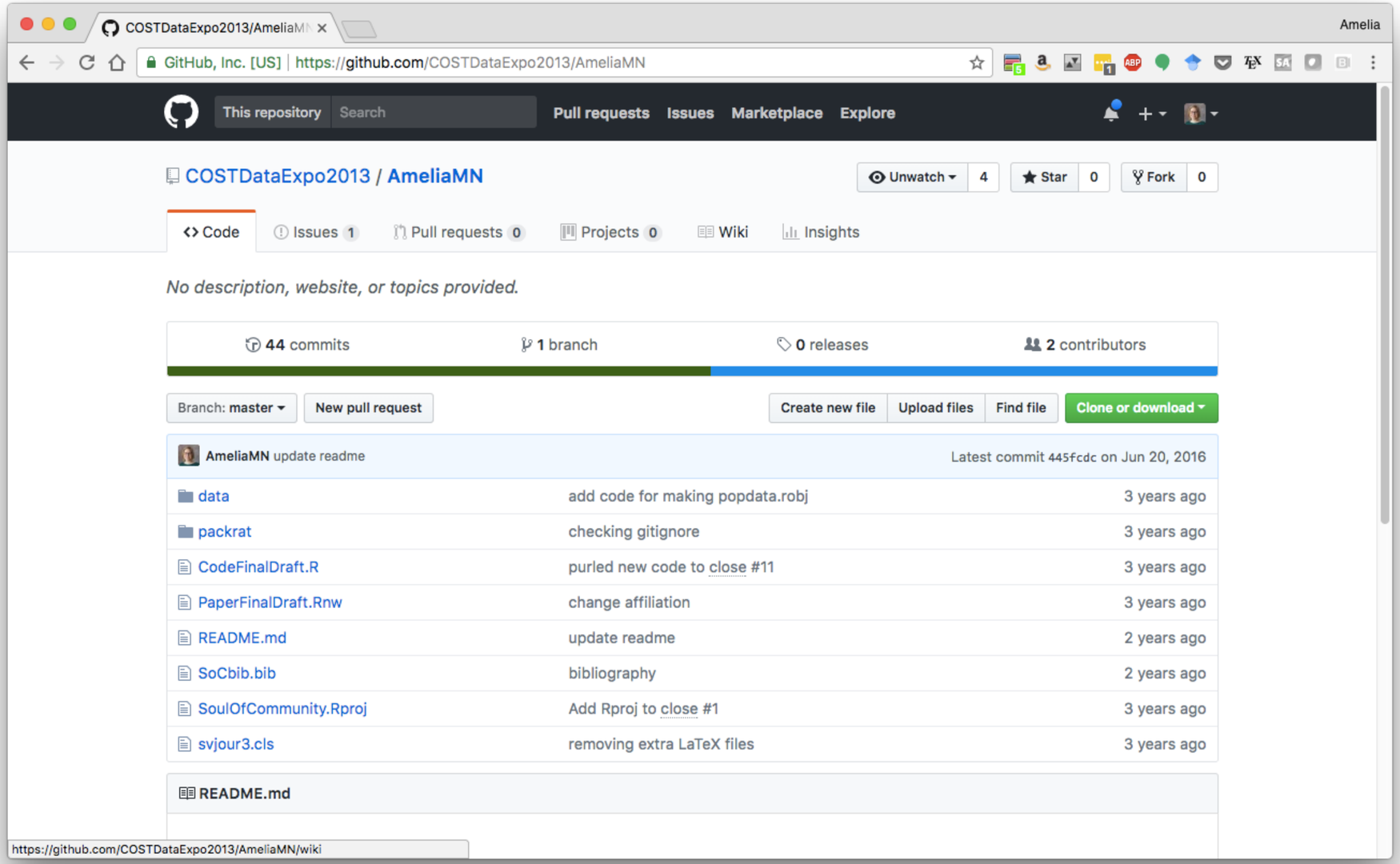
- Collaborators who need statisticians or data scientists to analyze data for them
- Students or postdocs in various disciplines looking for consulting advice
- Junior statistics students whose job it is to collate/clean/wrangle data sets

git and GitHub— happy git with R, Jenny Bryan

http://happygitwithr.com/



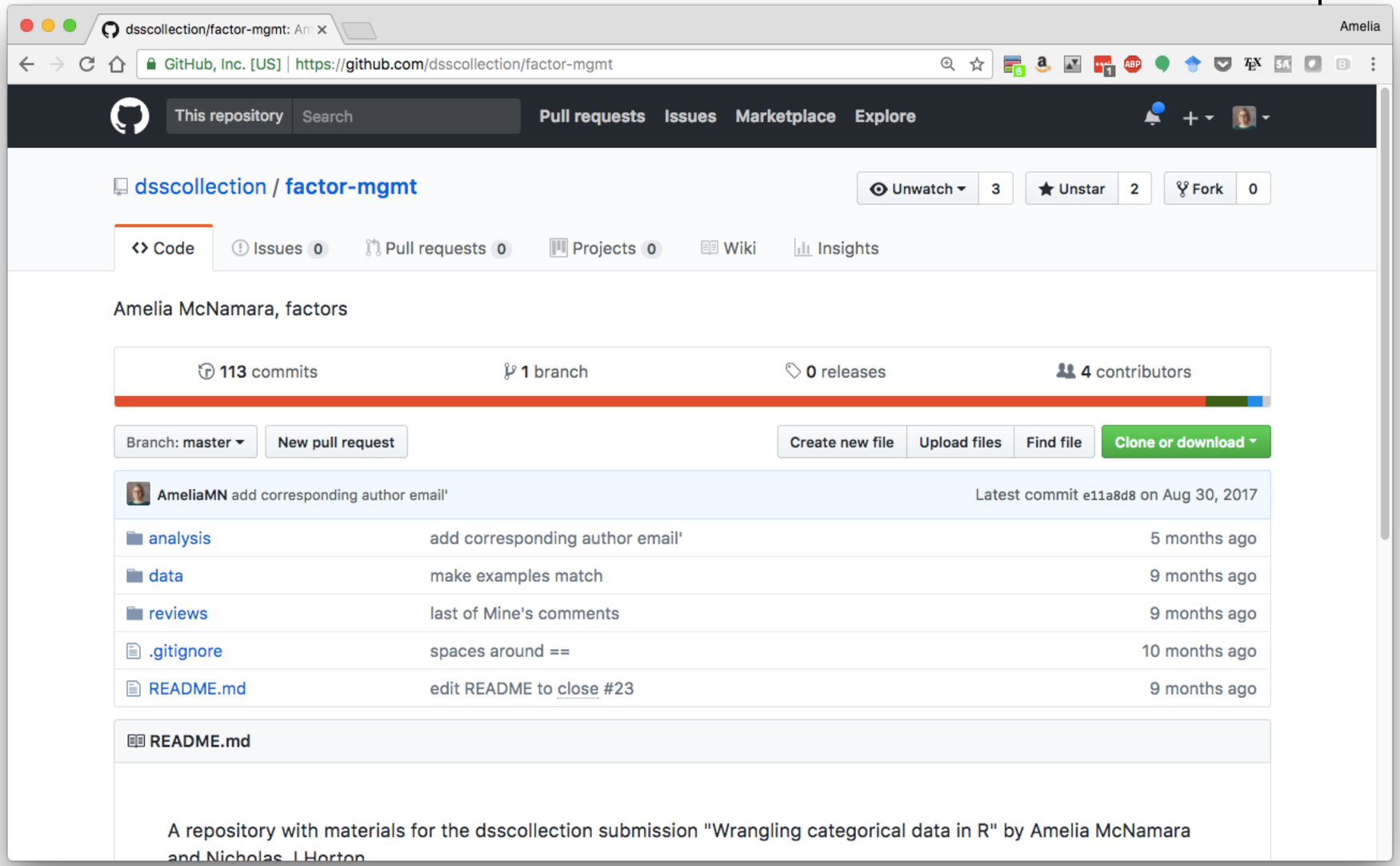
Example paper and file structure: https://github.com/COSTDataExpo2013/AmeliaMN



Another example

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

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



Resources

Practical Data Science for Stats PeerJ Collection. Curated by Jenny Bryan and Hadley Wickham. https://peerj.com/collections/50-practicaldatascistats/

‹ All Collections

Collection idea for us?

Practical Data Science for Stats - a PeerJ Collection

Data Science | Statistics | Scientific Computing and Simulation | Computer Education | Computational Science

Social Computing | Software Engineering | Science and Medical Education | Computational Biology

Human-Computer Interaction | Anthropology | Programming Languages | Visual Analytics | Graphics

Data Mining and Machine Learning

September 27, 2017 preprint

Forecasting at scale

8,126 downloads 12,545 views

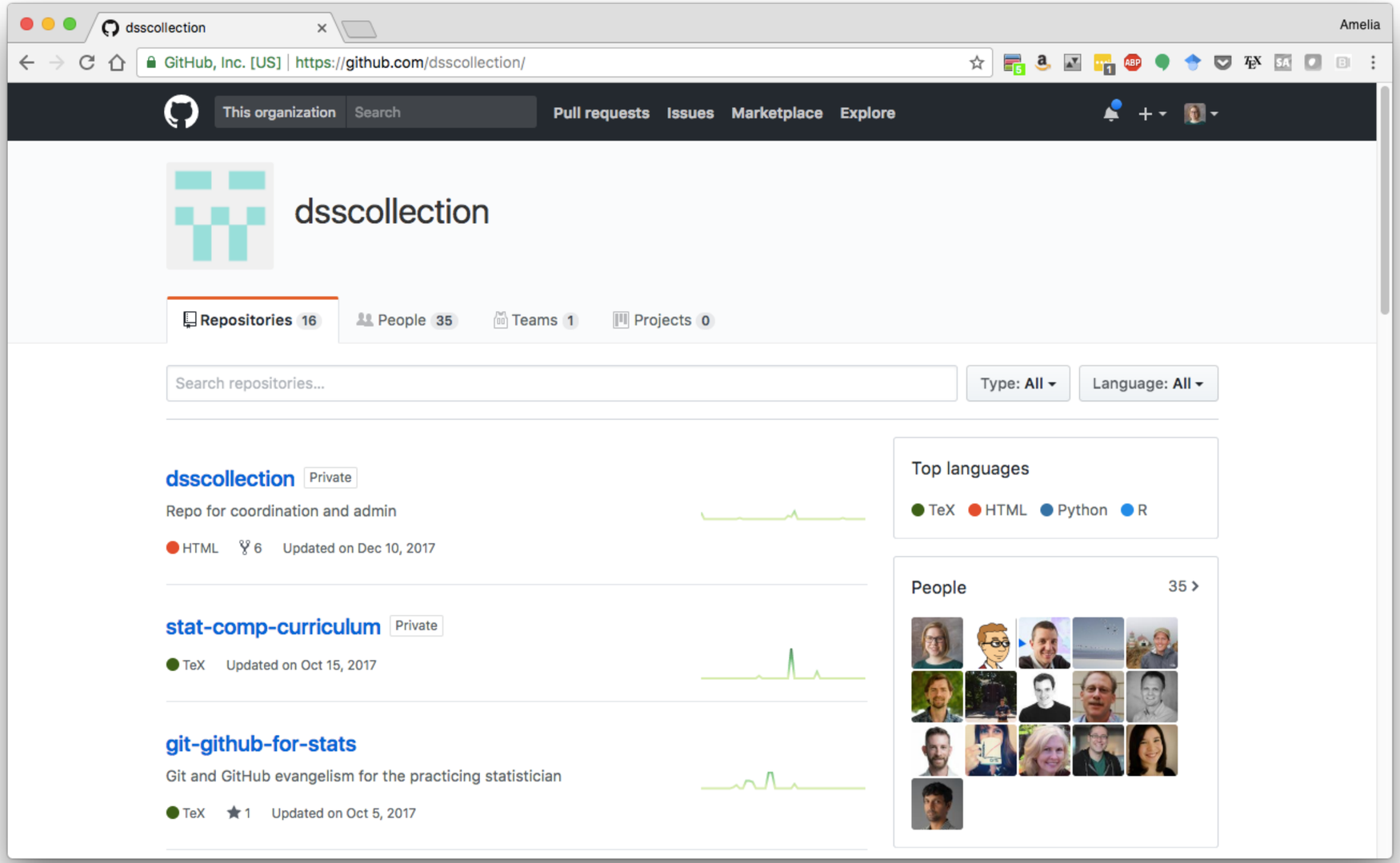
Sean J Taylor, Benjamin Letham

https://doi.org/10.7287/peerj.preprints.3190v2

Practical Data Science for Stats

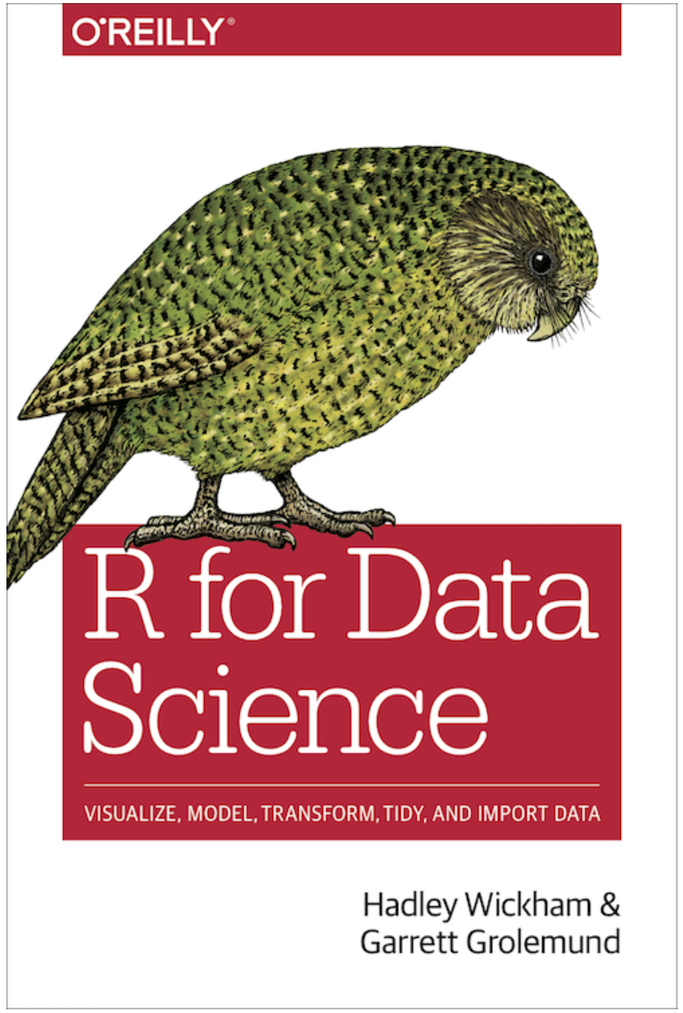
The "Practical Data Science for Stats" Collection contains preprints focusing on the practical side of data science workflows and statistical analysis. Curated by Jennifer Bryan and Hadley Wickham.

More at https://github.com/dsscollection/



Books

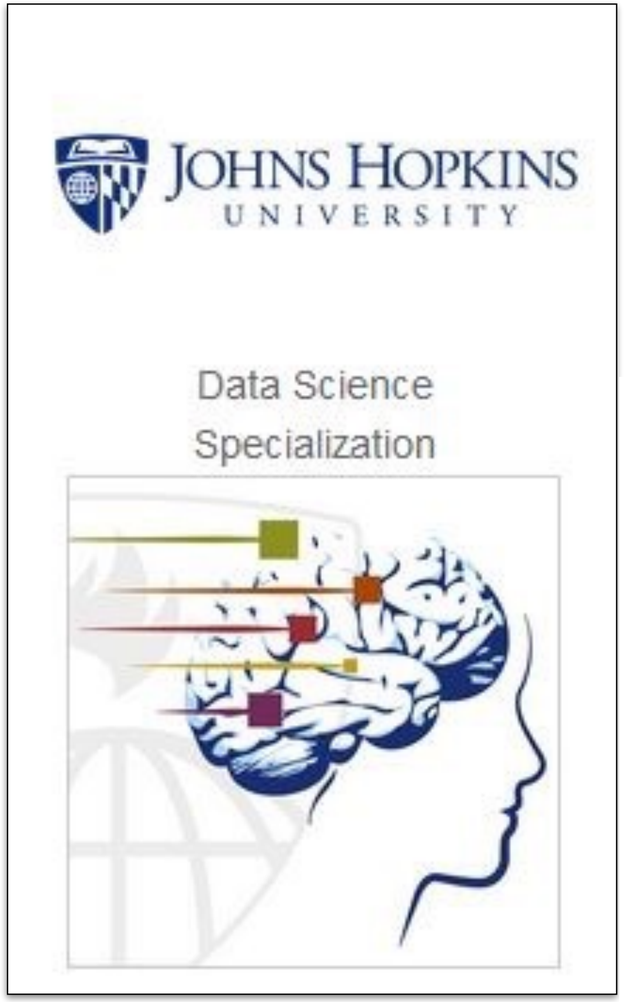
- Elements of Data Analytic Style, Jeff Leek
- R Programming for Data Science, Roger Peng
-The Art of Data Science, Roger Peng
- R Cookbook. Both a website, and a book, Winston Chang
- R for Data Science. Both a website and a book. Hadley Wickham and Garrett Grolemund.



Online learning/courses

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

- DataCamp. Interactive way to learn R in the browser. Free to start, pretty cheap to continue, discounts for education



(Brains?)

DataCamp

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, Smith College
- 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, Studio
- Julia Silge, tidytext, StackOverflow

Hashtags:
- #rstats
- #tidyverse
- #rcatladies

RStudio Community

https://community.rstudio.com

Press tab to search RStudio Community Search

Amelia

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	Welcome to the RStudio Community! meta 0 Aug '17
tidyverse This category is for anything and everything about the tidyverse.	17 / week 1 new	Memory usage and R's global string pool 2 43m
RStudio IDE This category is for discussing the RStudio IDE, both desktop and server versions.	20 / week 1 new	Missing value function RStudio IDE 17 1h
Teaching For discussions about teaching.	3 / week	Devtools::document Index Page • new Package development documentation 2 2h
shiny Please ask your questions about shiny here.	29 / week 3 new	Remove helpText() from panel shiny 1 2h
R Markdown Please ask your questions about R Markdown here.	5 / week	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