

Data Science in the

Amelia McNamara

amelia.mn

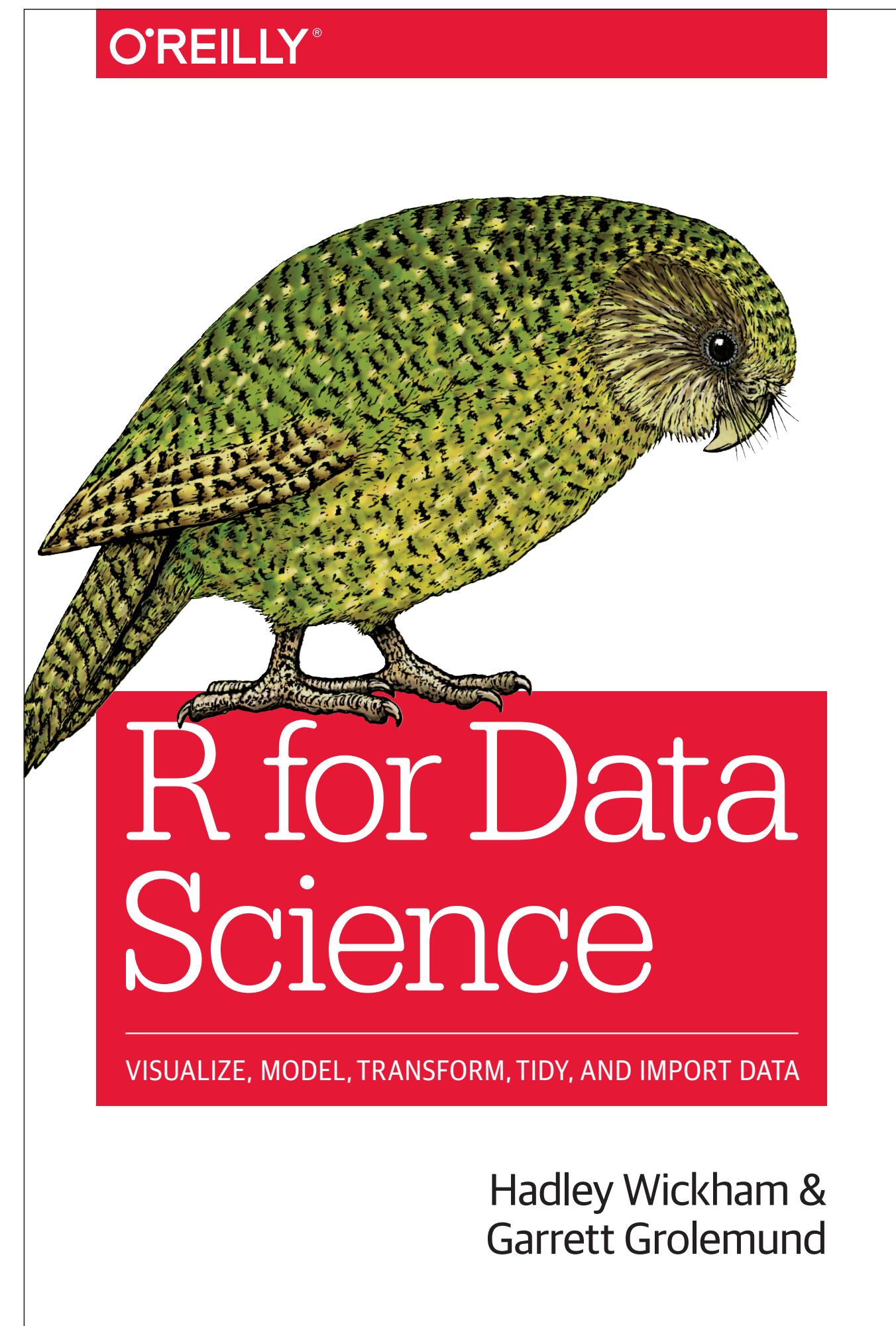
@AmeliaMN



Data Science in the tidyverse is licensed under a [Creative Commons Attribution 4.0 International License](#). Based on work at <https://github.com/cwickham/data-science-in-tidyverse> and <https://github.com/rstudio-education/master-the-tidyverse>

Online at:

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



RStudio best practices

Cleaning up

The screenshot shows the RStudio Cloud interface. The top bar indicates it's a secure connection to https://rstudio.cloud/project/13632. The user is named Amelia McNamara.

The main area features a data viewer displaying a table with columns: Year, ID, LaborStatus, MaritalStatus, NumChildren, Age, and HighestSchoolCompleted. The data shows various rows for the year 2014, with different labor and marital statuses, ages, and education levels.

To the right of the data viewer is the Environment pane, which lists global variables and datasets:

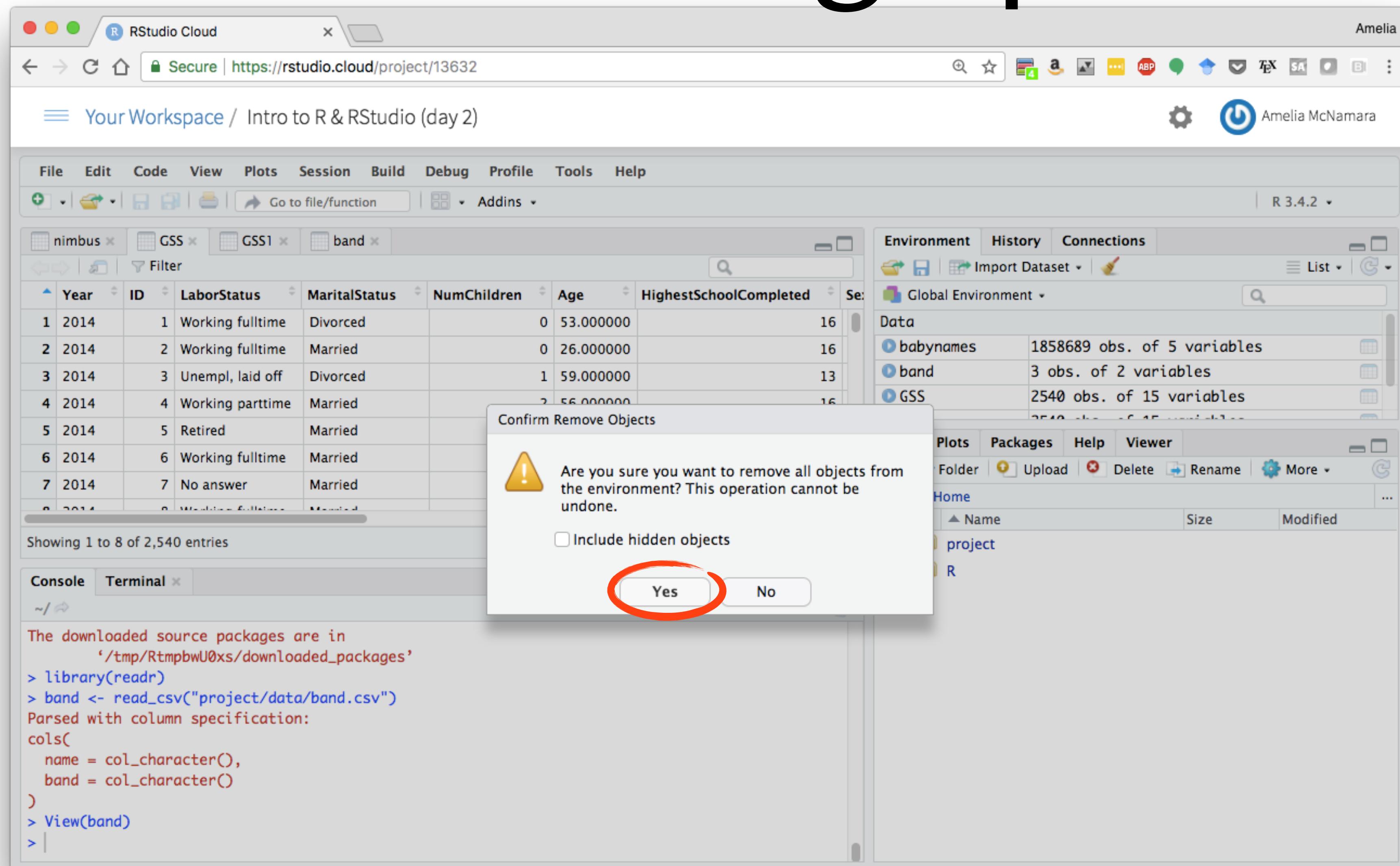
- babynames: 1858689 obs. of 5 variables
- band: 3 obs. of 2 variables
- GSS: 2540 obs. of 15 variables
- GGG1: 2540 obs. of 15 variables

A red circle highlights the "Import Dataset" button in the Environment pane toolbar.

The bottom section contains a Console tab showing R code and its output:

```
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



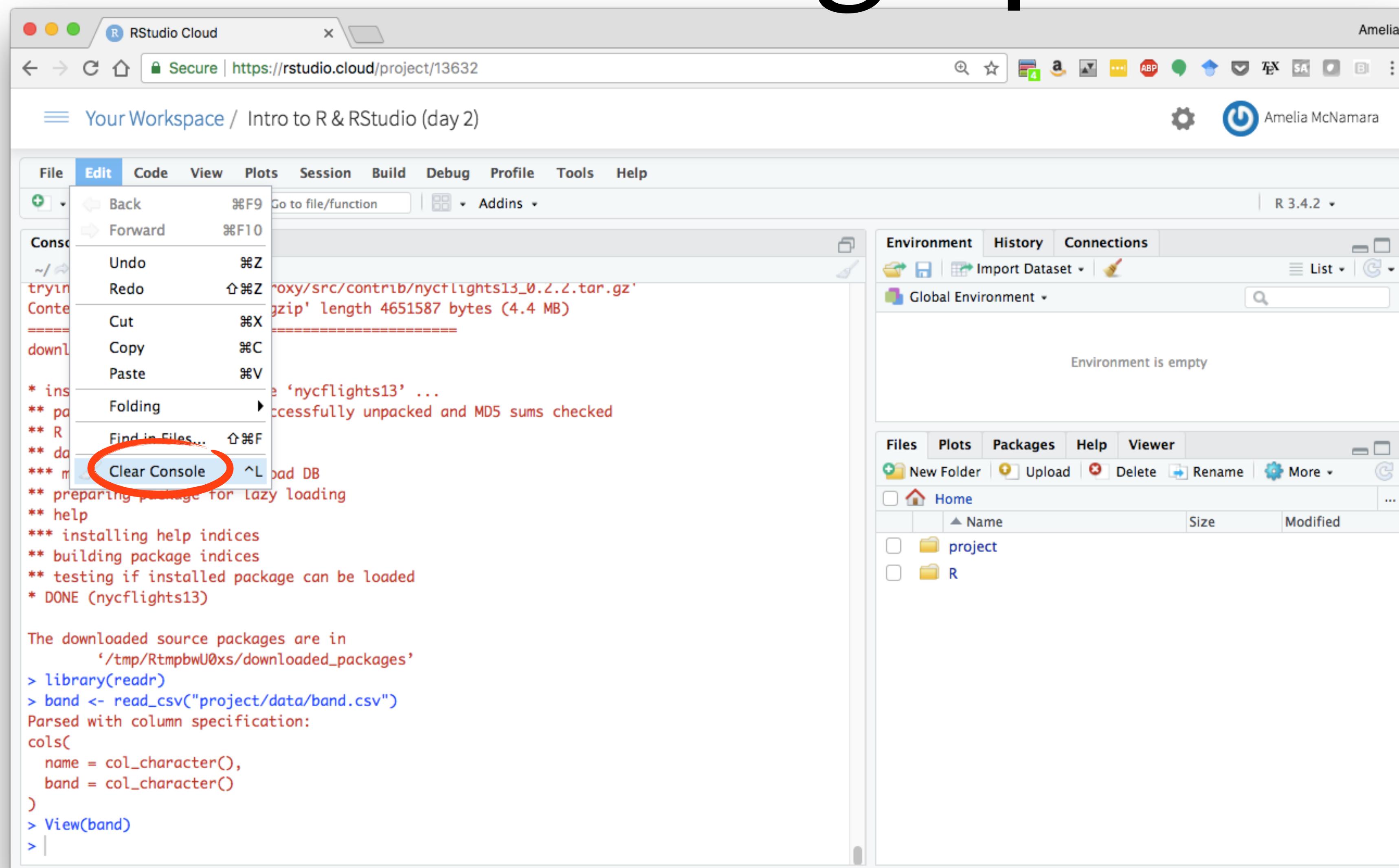
Cleaning up

The screenshot shows the RStudio Cloud interface with the following components:

- Header:** RStudio Cloud, Secure connection to https://rstudio.cloud/project/13632, User: Amelia McNamara.
- File Menu:** New File, Open File..., Import Dataset, Save, Save As..., Save All, Print..., Close, Close All (highlighted with a red circle), Close All Except Current.
- Data View:** Displays the "band" dataset with columns: MaritalStatus, NumChildren, Age, HighestSchoolCompleted. The data shows various marital statuses (Divorced, Married) across different ages and education levels.
- Environment:** Shows the Global Environment pane which is currently empty, indicating "Environment is empty".
- Console:** Displays R code and output:

```
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)
> |
```
- File Explorer:** Shows the project structure with a "project" folder and an "R" folder.

Cleaning up



Cleaning up

A screenshot of the RStudio Cloud interface. The title bar says "RStudio Cloud" and the user is "Amelia". The URL is "Secure | https://rstudio.cloud/project/13632". The sidebar shows "Your Workspace / Intro to R & RStudio (day 2)". The main window has a menu bar with File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help. Below the menu is a toolbar with various icons. The console tab is selected, showing a command line with a single character '>'. To the right of the console is the History tab of the Environment panel, which contains the following R code:

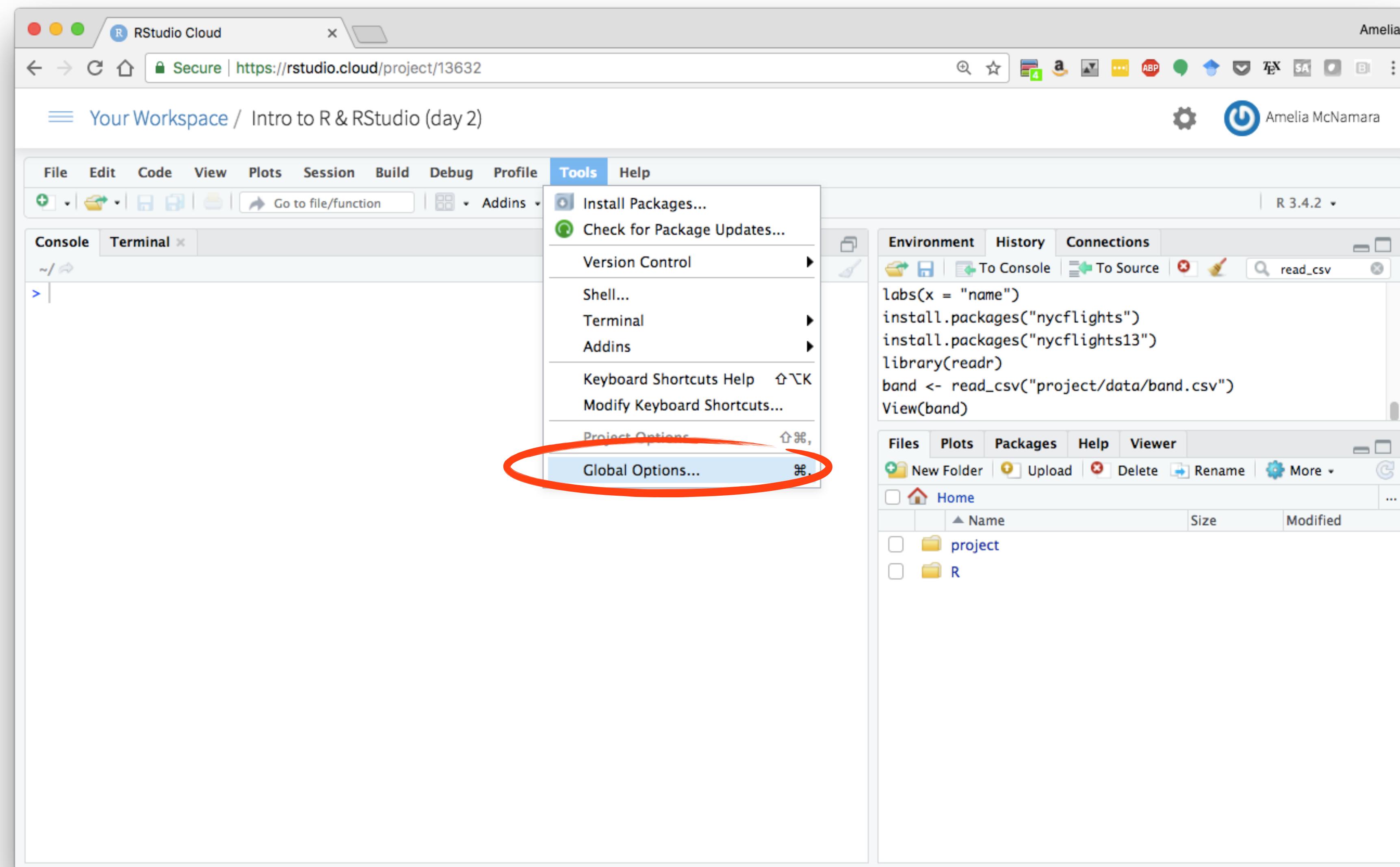
```
labs(x = "name")
install.packages("nycflights")
install.packages("nycflights13")
library(readr)
band <- read_csv("project/data/band.csv")
View(band)
```

A red circle highlights this code. Below the History tab is the Connections tab. At the bottom of the interface is a file browser with tabs for Files, Plots, Packages, Help, and Viewer. It shows a "Home" folder with two items: "project" and "R".

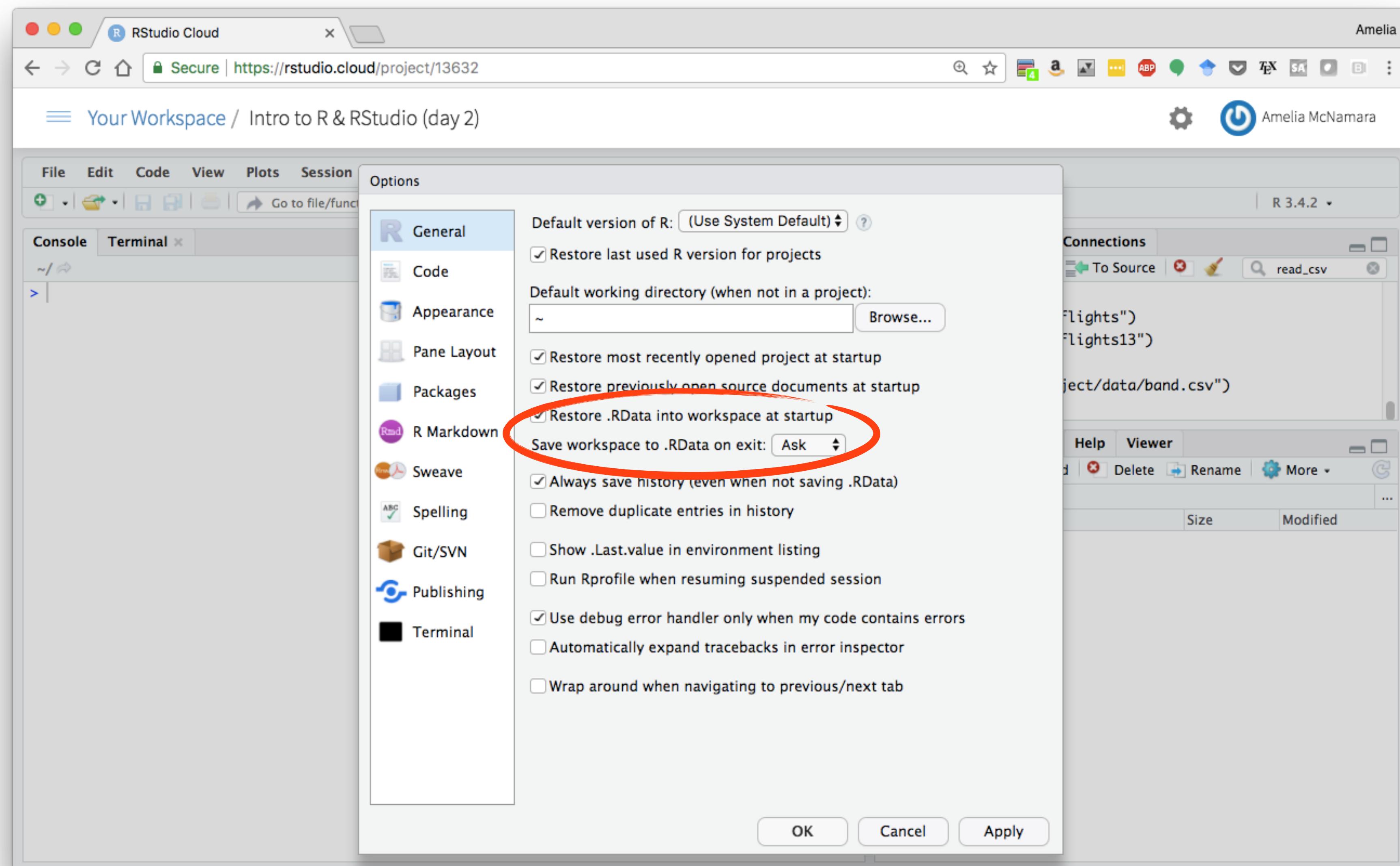
Don't worry,
your history is
preserved

Settings

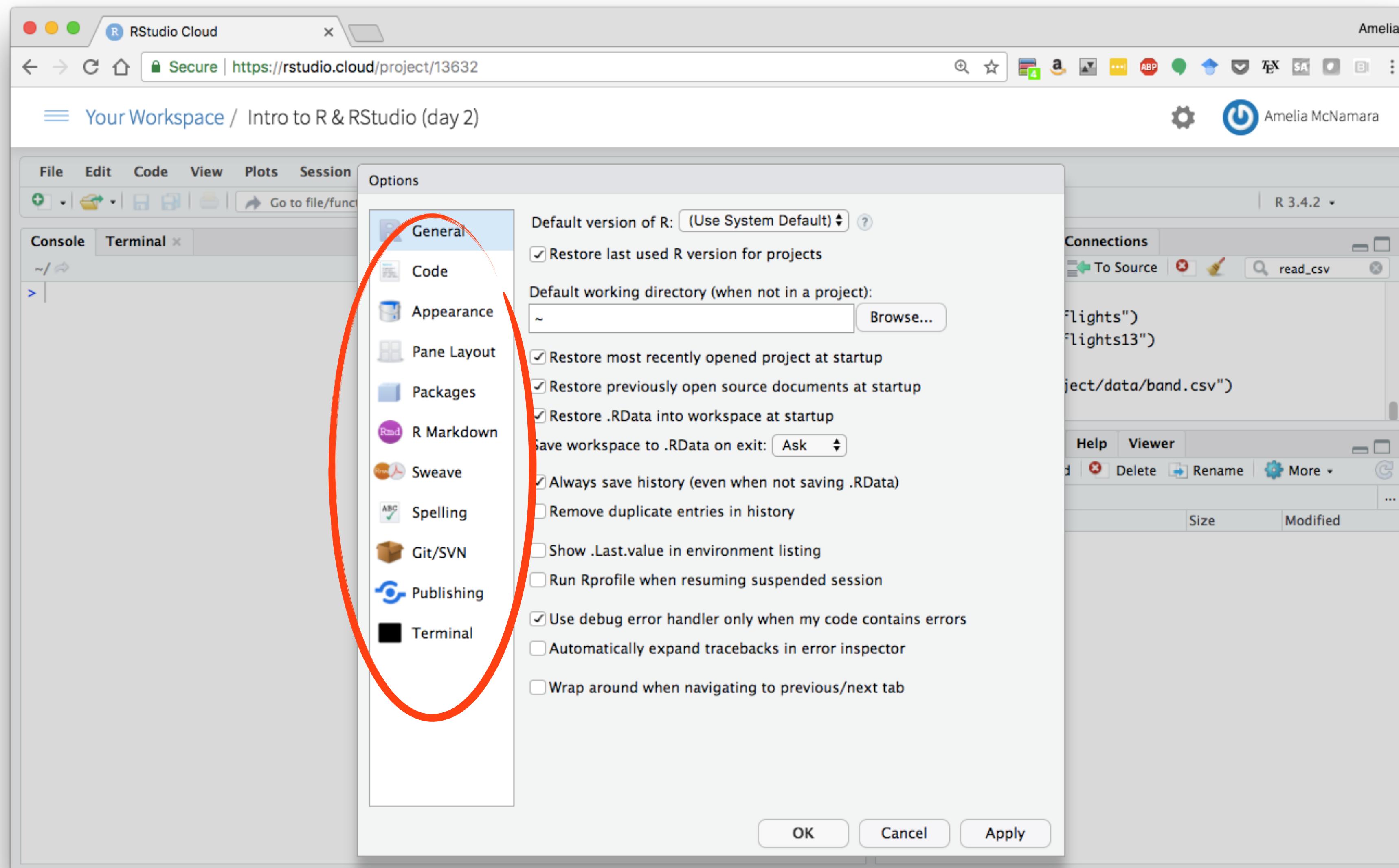
Cleaning up



Cleaning up



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



The Comprehensive R Archive Network

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](#)
- [Download R for \(Mac\) OS X](#)
- [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 listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2017-11-30, Kite-Eating Tree) [R-3.4.3.tar.gz](#), read [what's new](#) in the latest version.
- Sources of [R alpha and beta releases](#) (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are [available here](#). Please read about [new features and bug fixes](#) before filing corresponding feature requests or bug reports.
- Source code of older versions of R is [available here](#).
- Contributed extension [packages](#)

Questions About R

- If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

What are R and CRAN?

R is 'GNU S', a freely available language and environment for statistical computing and graphics which provides a wide variety of statistical and graphical techniques: linear and nonlinear modelling, statistical tests, time series analysis, classification, clustering, etc. Please consult the [R project homepage](#) for

Then, install RStudio (the application).

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

The screenshot shows the RStudio download page. At the top, there's a navigation bar with links for rstudio::conf, Products, Resources, Pricing, About Us, Blogs, and a search icon. Below the navigation, there's a heading "Choose Your Version of RStudio" with a subtext explaining what RStudio is and linking to more features. There are five options listed:

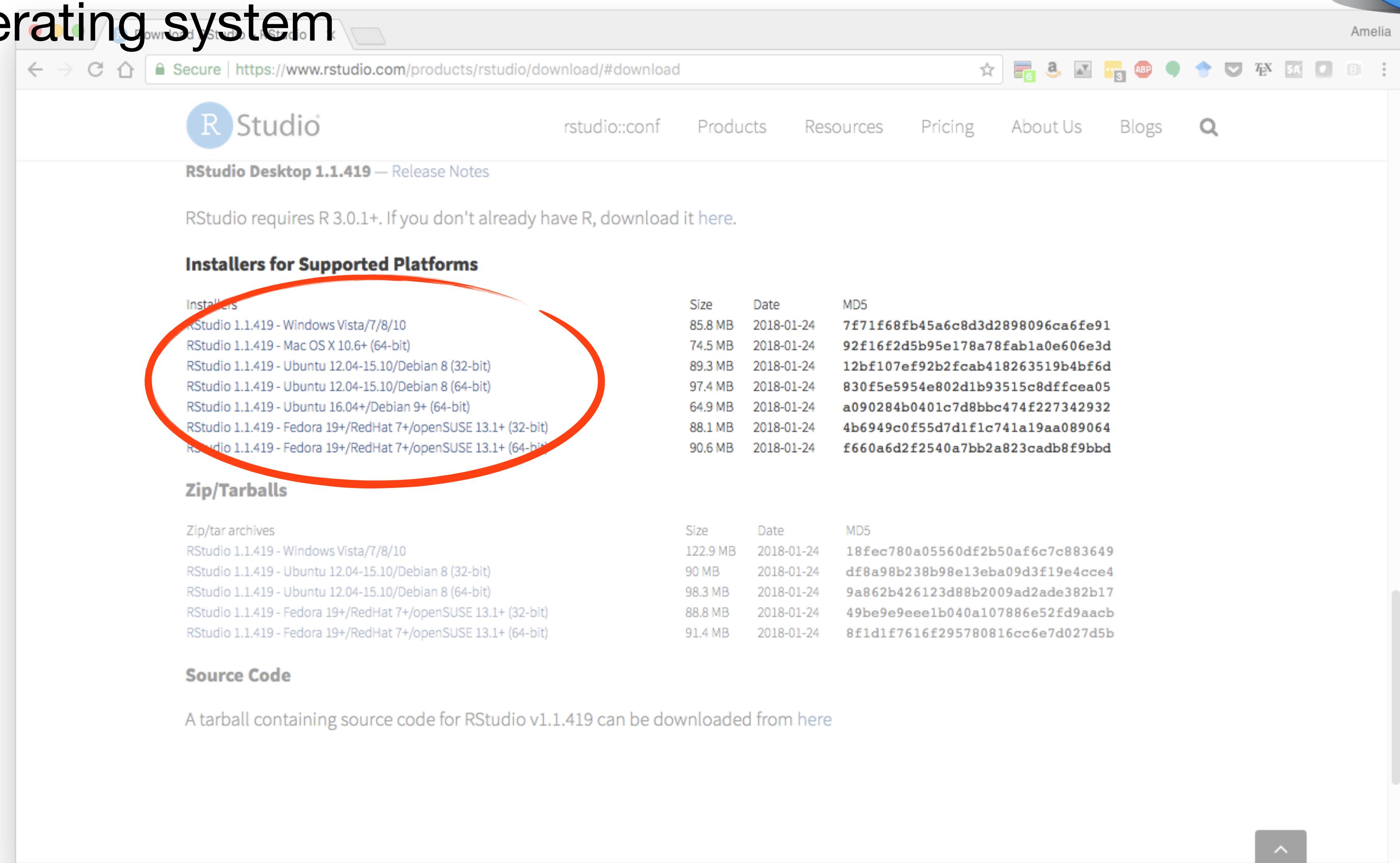
RStudio Desktop Open Source License	RStudio Desktop Commercial License	RStudio Server Open Source License	RStudio Server Pro Commercial License	RStudio Server Pro + RStudio Connect Commercial License
FREE	\$995 per year	FREE	\$9,995 per year	\$29,995 per year
DOWNLOAD Learn More	BUY Learn More	DOWNLOAD Learn More	DOWNLOAD Learn More	TALK Learn More

Each row has a green button for "DOWNLOAD" or "BUY" and a "Learn More" link. Below the table, there's a section titled "Integrated Tools for R" with a list of tools and a "Priority Support" section.



Then, install RStudio (the application).

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



The screenshot shows the RStudio download page for version 1.1.419. A red oval highlights the 'Installers for Supported Platforms' section. Below it are sections for 'Zip/Tarballs' and 'Source Code'.

Installers for Supported Platforms

Installer	Size	Date	MD5
RStudio 1.1.419 - Windows Vista/7/10	85.8 MB	2018-01-24	7f71f68fb45a6c8d3d2898096ca6fe91
RStudio 1.1.419 - Mac OS X 10.6+ (64-bit)	74.5 MB	2018-01-24	92f16f2d5b95e178a78fab1a0e606e3d
RStudio 1.1.419 - Ubuntu 12.04-15.10/Debian 8 (32-bit)	89.3 MB	2018-01-24	12bf107ef92b2fcab418263519b4bf6d
RStudio 1.1.419 - Ubuntu 12.04-15.10/Debian 8 (64-bit)	97.4 MB	2018-01-24	830f5e5954e802d1b93515c8dffcea05
RStudio 1.1.419 - Ubuntu 16.04+/Debian 9+ (64-bit)	64.9 MB	2018-01-24	a090284b0401c7d8bbc474f227342932
RStudio 1.1.419 - Fedora 19+/RedHat 7+/openSUSE 13.1+ (32-bit)	88.1 MB	2018-01-24	4b6949c0f55d7d1f1c741a19aa089064
RStudio 1.1.419 - Fedora 19+/RedHat 7+/openSUSE 13.1+ (64-bit)	90.6 MB	2018-01-24	f660a6d2f2540a7bb2a823cadb8f9bbd

Zip/Tarballs

Zip/tar archives	Size	Date	MD5
RStudio 1.1.419 - Windows Vista/7/10	122.9 MB	2018-01-24	18fec780a05560df2b50af6c7c883649
RStudio 1.1.419 - Ubuntu 12.04-15.10/Debian 8 (32-bit)	90 MB	2018-01-24	df8a98b238b98e13eba09d3f19e4cce4
RStudio 1.1.419 - Ubuntu 12.04-15.10/Debian 8 (64-bit)	98.3 MB	2018-01-24	9a862b426123d88b2009ad2ade382b17
RStudio 1.1.419 - Fedora 19+/RedHat 7+/openSUSE 13.1+ (32-bit)	88.8 MB	2018-01-24	49be9e9eee1b040a107886e52fd9aacb
RStudio 1.1.419 - Fedora 19+/RedHat 7+/openSUSE 13.1+ (64-bit)	91.4 MB	2018-01-24	8f1d1f7616f295780816cc6e7d027d5b

Source Code

A tarball containing source code for RStudio v1.1.419 can be downloaded from here



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

The screenshot shows the RStudio Cloud interface. At the top, there's a header bar with the title "RStudio Cloud" and a user profile for "Amelia McNamara". Below the header is a toolbar with various icons for navigation and project management. The main workspace contains several panes:

- Data Viewer:** A grid view showing data from four datasets: "nimbus", "GSS", "GSS1", and "band". The "band" dataset is currently selected, displaying columns like Year, ID, LaborStatus, MaritalStatus, NumChildren, Age, and HighestSchoolCompleted. The data shows 2,540 entries from 2014.
- File Browser:** An "Environment" pane showing the global environment with objects like "band", "GSS", "GSS1", and "nimbus".
- Console:** A terminal window showing R code and its output. The user has run commands to download packages, load the "readr" library, read the "band.csv" file, and view the resulting "band" data frame.

A large text overlay in the center-right of the screen reads:

You can export an entire directory from RStudio cloud

Or, download a clean version from <https://github.com/AmeliaMN/data-science-in-tidyverse>

The screenshot shows a GitHub repository page for 'AmeliaMN / data-science-in-tidyverse'. The repository has 11 commits, 1 branch, 0 releases, and 1 contributor. The 'Clone or download' button is highlighted with a red circle. A modal window is open, showing the 'Clone with HTTPS' URL (<https://github.com/AmeliaMN/data-science-in-tidyverse>) and a 'Download ZIP' button, which is also highlighted with a red circle.

Materials for Data Science in the Tidyverse, a two-day workshop @ rstudio:conf(2019)

Manage topics

11 commits 1 branch 0 releases 1 contributor

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

AmeliaMN add solutions, update cheatsheets

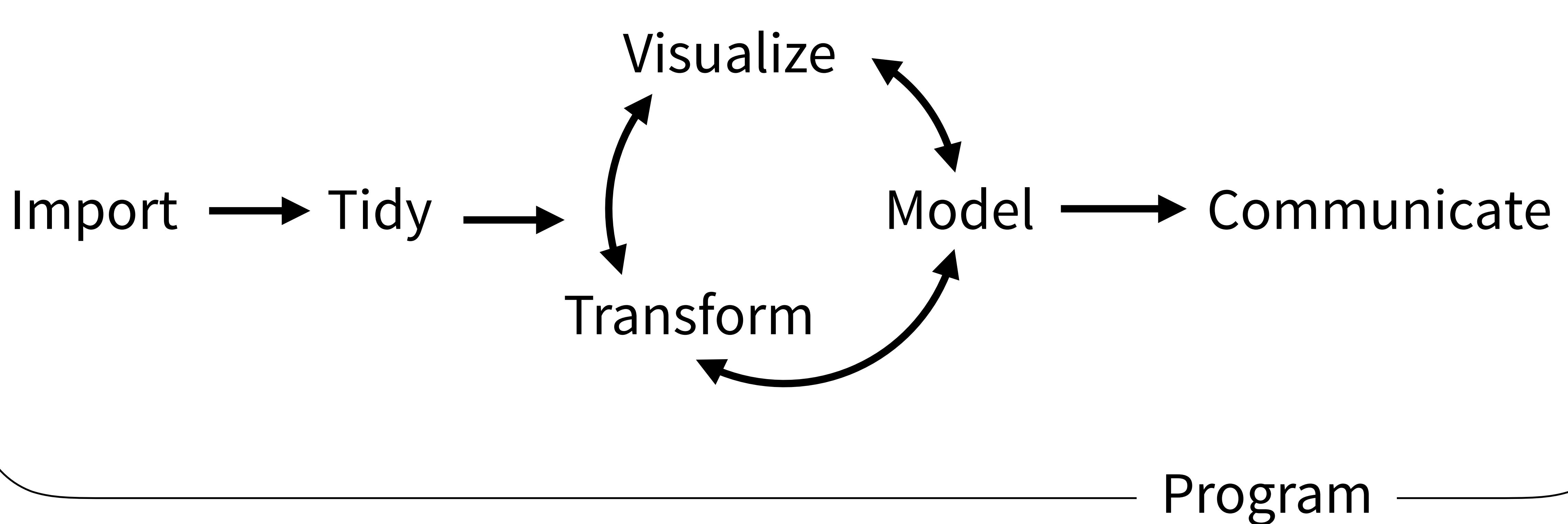
- cheatsheets add solutions, update cheatsheets
- resources remove 3rd screenshot
- slides add solutions, update cheatsheets
- solutions add solutions, update cheatsheets
- .gitignore ignore keynotes
- 00-Getting-started.Rmd update README
- 01-Visualize.Rmd add solutions, update cheatsheets
- 02-Transform.Rmd many changes
- 03-Tidy.Rmd add solutions, update cheatsheets

Open in Desktop Download ZIP

<https://github.com/AmeliaMN/data-science-in-tidyverse/archive/master.zip>

More to
learn

(Applied) 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

Table of contents

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

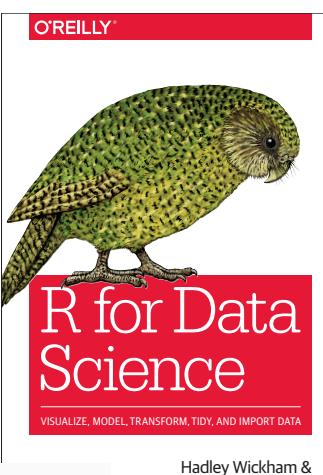
27 R Markdown

28 Graphics for communication

29 R Markdown formats

30 R Markdown workflow

**Review things
we've covered**



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

Table of contents

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

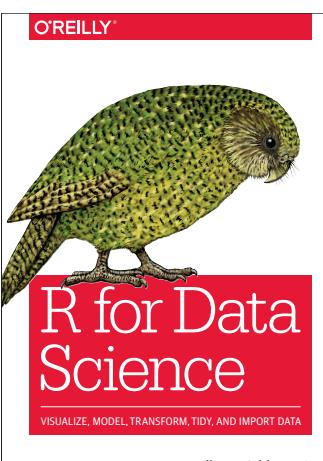
27 R Markdown

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>

The screenshot shows a GitHub repository page for the user 'Amelia' with the repository name 'COSTDataExpo2013 / AmeliaMN'. The page includes a navigation bar with links for 'This repository', 'Search', 'Pull requests', 'Issues', 'Marketplace', and 'Explore'. Below the navigation bar, there are buttons for 'Unwatch', 'Star', 'Fork', and a profile picture. The main content area displays the repository's details: 'No description, website, or topics provided.' It shows 44 commits, 1 branch, 0 releases, and 2 contributors. A green 'Clone or download' button is prominent. The commit history lists several files: 'data', 'packrat', 'CodeFinalDraft.R', 'PaperFinalDraft.Rnw', 'README.md', 'SoCbib.bib', 'SoulOfCommunity.Rproj', 'svjour3.cls', and 'README.md'. The latest commit was made by 'AmeliaMN' on June 20, 2016.

File	Commit Message	Time Ago
data	add code for making popdata.rObj	3 years ago
packrat	checking gitignore	3 years ago
CodeFinalDraft.R	purled new code to close #11	3 years ago
PaperFinalDraft.Rnw	change affiliation	3 years ago
README.md	update readme	2 years ago
SoCbib.bib	bibliography	2 years ago
SoulOfCommunity.Rproj	Add Rproj to close #1	3 years ago
svjour3.cls	removing extra LaTeX files	3 years ago
README.md		

Another example

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

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

The screenshot shows the GitHub repository page for `dsscollection/factor-mgmt`. The repository was created by Amelia McNamara and has 113 commits, 1 branch, 0 releases, and 4 contributors. The latest commit was made on August 30, 2017. The repository description states: "A repository with materials for the dsscollection submission 'Wrangling categorical data in R' by Amelia McNamara and Nicholas J Horton".

Key statistics:

- 113 commits
- 1 branch
- 0 releases
- 4 contributors

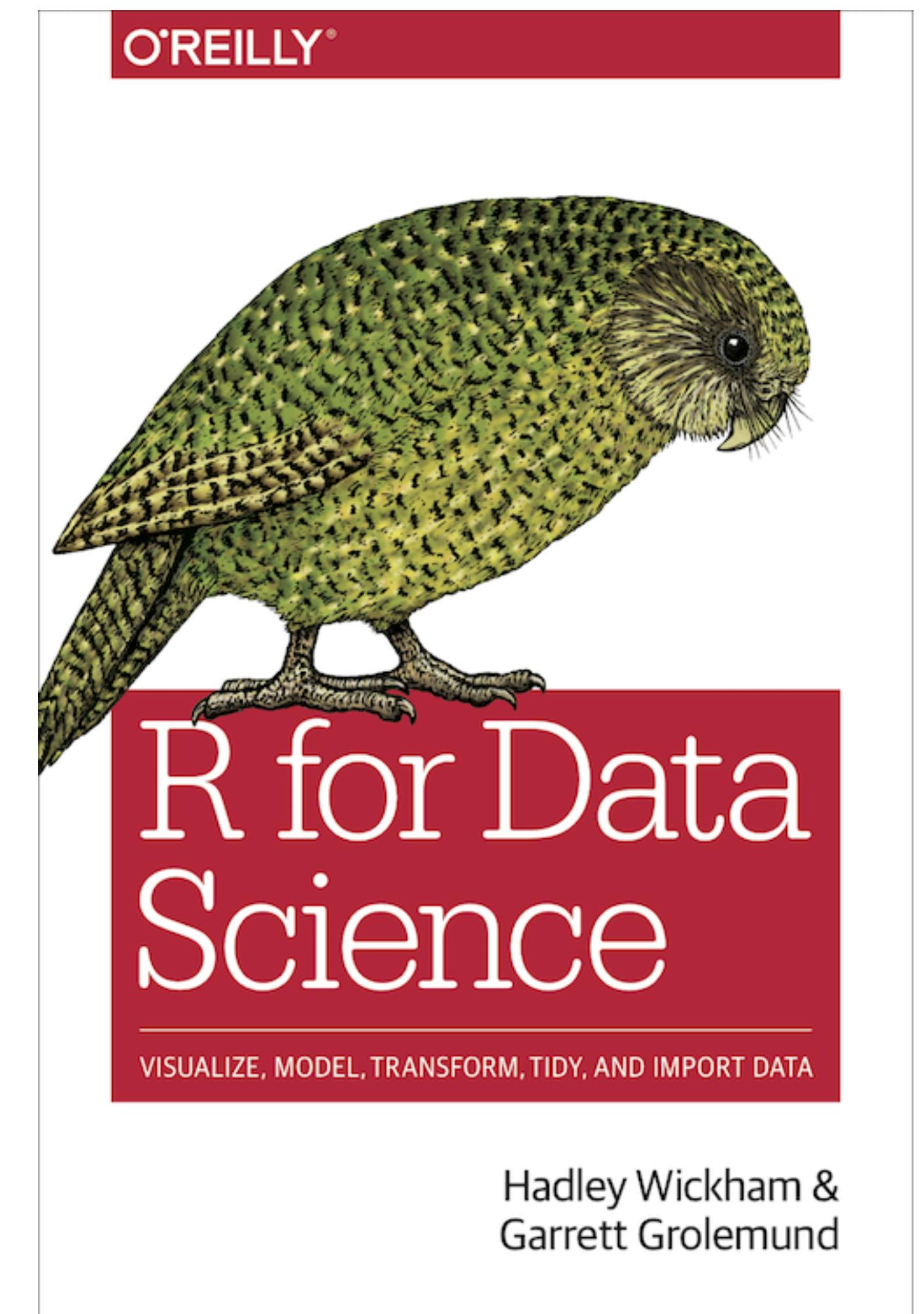
Recent commits:

File	Message	Time
analysis	add corresponding author email'	5 months ago
data	make examples match	9 months ago
reviews	last of Mine's comments	9 months ago
.gitignore	spaces around ==	10 months ago
README.md	edit README to close #23	9 months ago

A repository with materials for the dsscollection submission "Wrangling categorical data in R" by Amelia McNamara and Nicholas J Horton

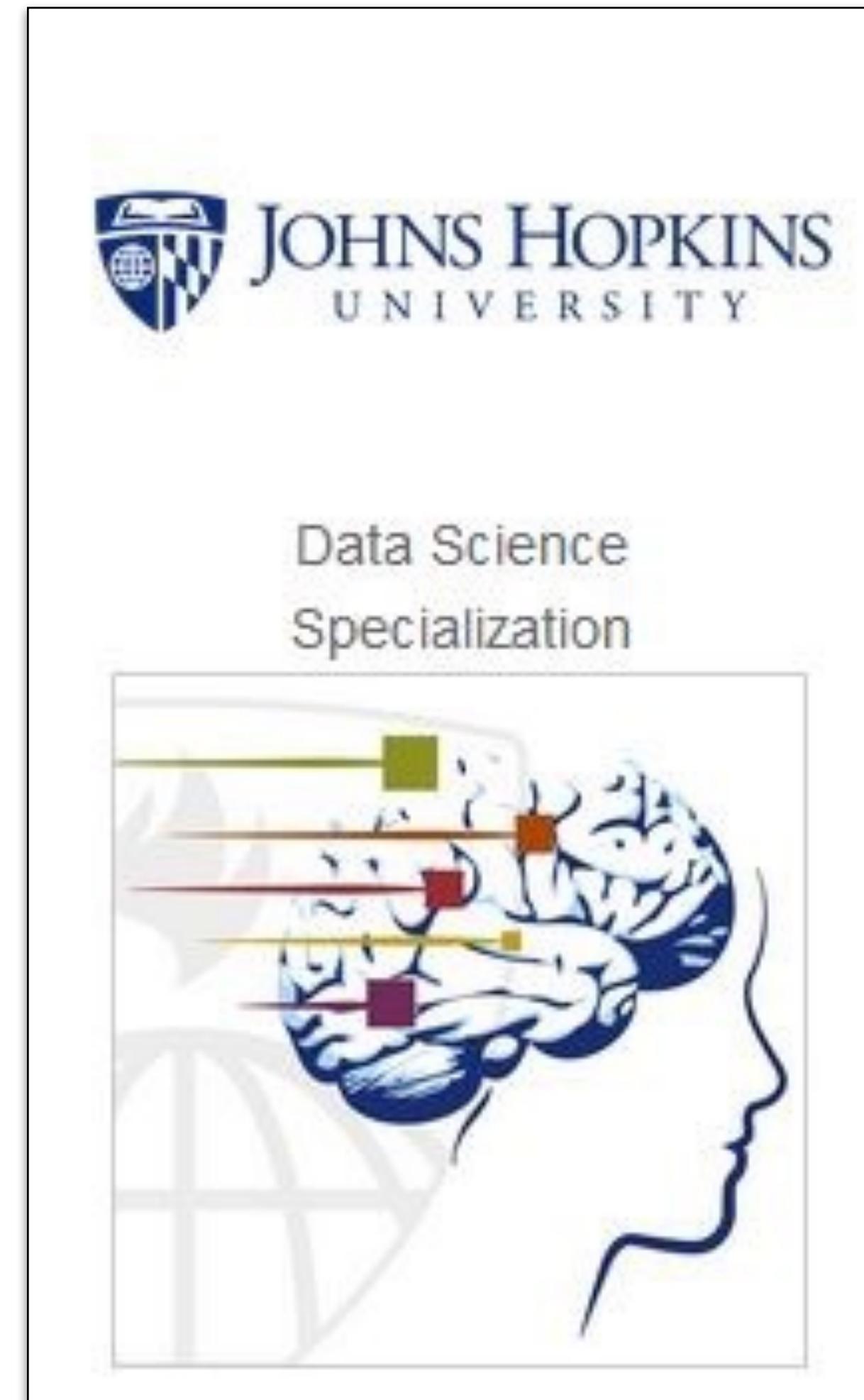
Books

- 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



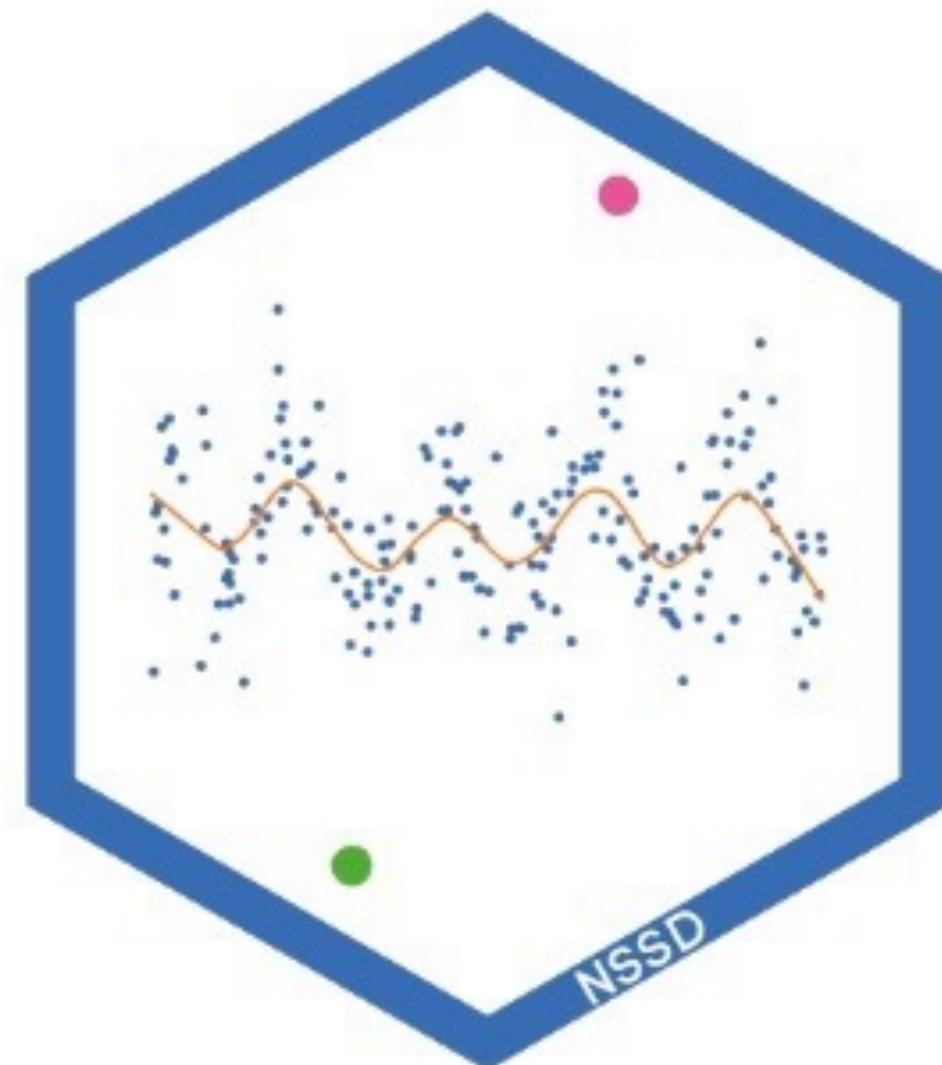
(Brains?)



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





Who to follow

- me! [Amelia McNamara](#), University of St Thomas
- [Hadley Wickham](#), RStudio
- [Jenny Bryan](#), on leave from UBC, at RStudio
- [Hillary 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 X

https://community.rstudio.com Press **tab** to search RStudio Community Search ☆ 6 a A ABP 3 TEX SA B :

R Studio Community

[https://community.rstudio.com/](#)

[+ New Topic](#)

all categories ▾ all tags ▾ Categories Latest New (7) Unread Top

Category	Topics	Latest
rstudio::conf 2018  This category is for anything and everything related to rstudio::conf.	8 / week	 Welcome to the RStudio Community! 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 17 1h RStudio IDE
Teaching  For discussions about teaching.	3 / week	 Devtools::document Index Page •new 2 2h Package development documentation
shiny  Please ask your questions about shiny here.	29 / week 3 new	 Remove helpText() from panel 1 2h shiny
R Markdown  Please ask your questions about R Markdown here.	5 / week	 NB Classifier with Priors and Likelihoods 2 2h

Welcome to the RStudio Community!

This category is for anything and everything related to rstudio::conf.

Memory usage and R's global string pool

This category is for anything and everything about the tidyverse.

Missing value function

This category is for discussing the RStudio IDE, both desktop and server versions.

Devtools::document Index Page •new

For discussions about teaching.

Remove helpText() from panel

Please ask your questions about shiny here.

NB Classifier with Priors and Likelihoods

Please ask your questions about R Markdown here.

Welcome to the RStudio Community!

This category is for anything and everything related to rstudio::conf.

Memory usage and R's global string pool

This category is for anything and everything about the tidyverse.

Missing value function

This category is for discussing the RStudio IDE, both desktop and server versions.

Devtools::document Index Page •new

For discussions about teaching.

Remove helpText() from panel

Please ask your questions about shiny here.

NB Classifier with Priors and Likelihoods

Please ask your questions about R Markdown here.

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](mailto:r-help@r-project.org)
- Official word on asking for help: <https://www.tidyverse.org/help/>

Thanks to my fantastic TAs

HELLO

my name is

Jesse

 @kierisi

HELLO

my name is

Irene

 @i_stevies

HELLO

my name is

Ben

 @baumerben

Thank you!