



Introduction to Data Science in the Tidyverse

Amelia McNamara and Hadley Wickham



rstudio::conf
SAN FRANCISCO // JANUARY 27 - 30, 2020

from  RStudio



Acknowledgements

Data Science in the tidyverse is licensed under a Creative Commons Attribution 4.0 International License.

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>

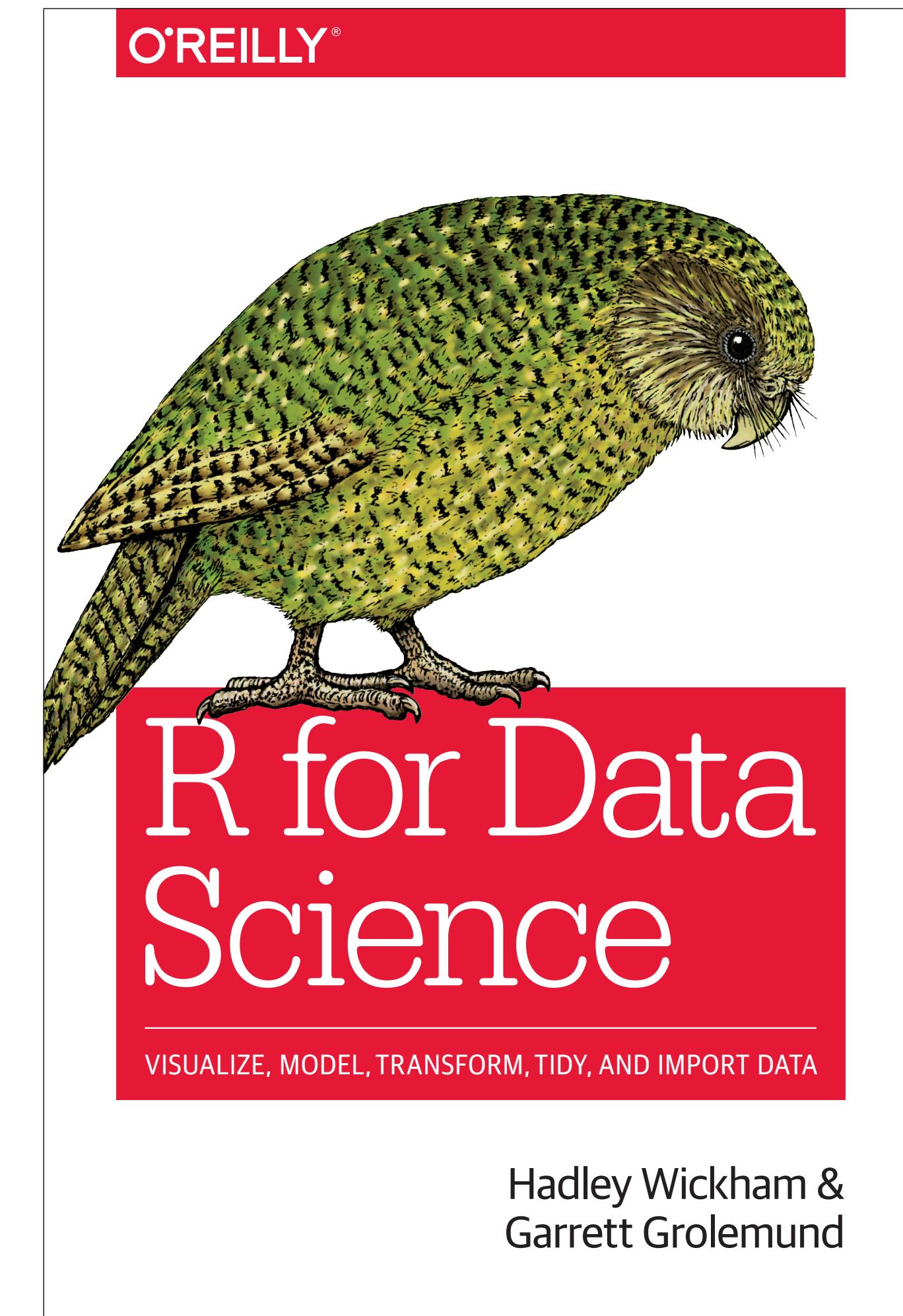
<https://github.com/AmeliaMN/IntroToR>

<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

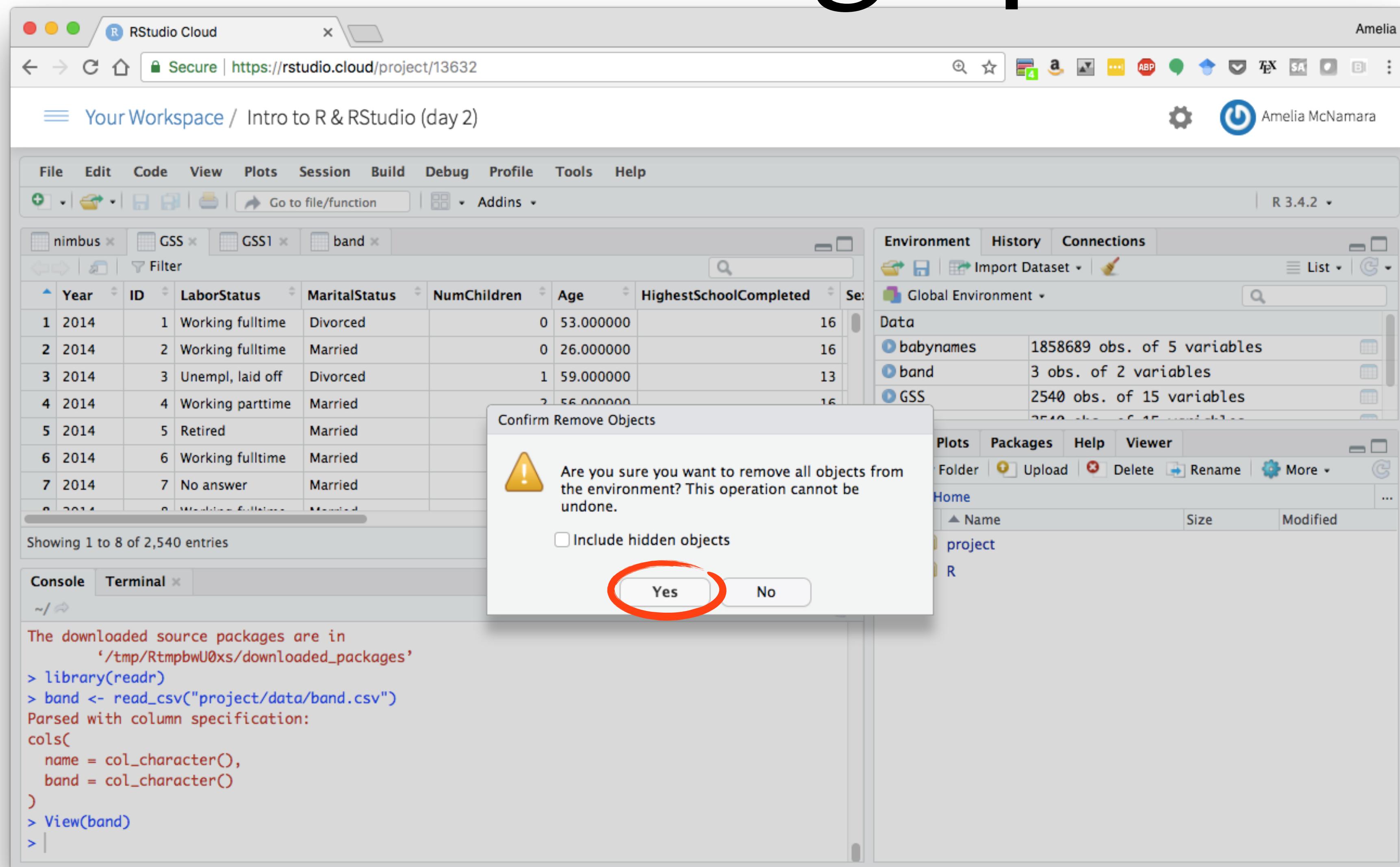


Cleaning up

The screenshot shows the RStudio Cloud interface. At the top, there's a browser header with 'RStudio Cloud' and a URL 'https://rstudio.cloud/project/13632'. Below it is the main RStudio window.

- Data Viewer:** On the left, a data frame titled 'band' is displayed with columns: Year, ID, LaborStatus, MaritalStatus, NumChildren, Age, and HighestSchoolCompleted. The data shows 2,540 entries from 2014, with various marital and labor statuses.
- Environment Pane:** On the right, the 'Environment' tab is selected. It lists global variables: babynames (185,868 obs. of 5 variables), band (3 obs. of 2 variables), GSS (2,540 obs. of 15 variables), and GSS1 (2,540 obs. of 15 variables). A red circle highlights the 'Import Dataset' button in the toolbar above the list.
- Console:** At the bottom, the console shows R code and its output. It includes commands like 'library(readr)', 'read_csv("project/data/band.csv")', and 'View(band)'. The output indicates the source packages were downloaded to '/tmp/RtmpbwU0xs/downloaded_packages'.

Cleaning up

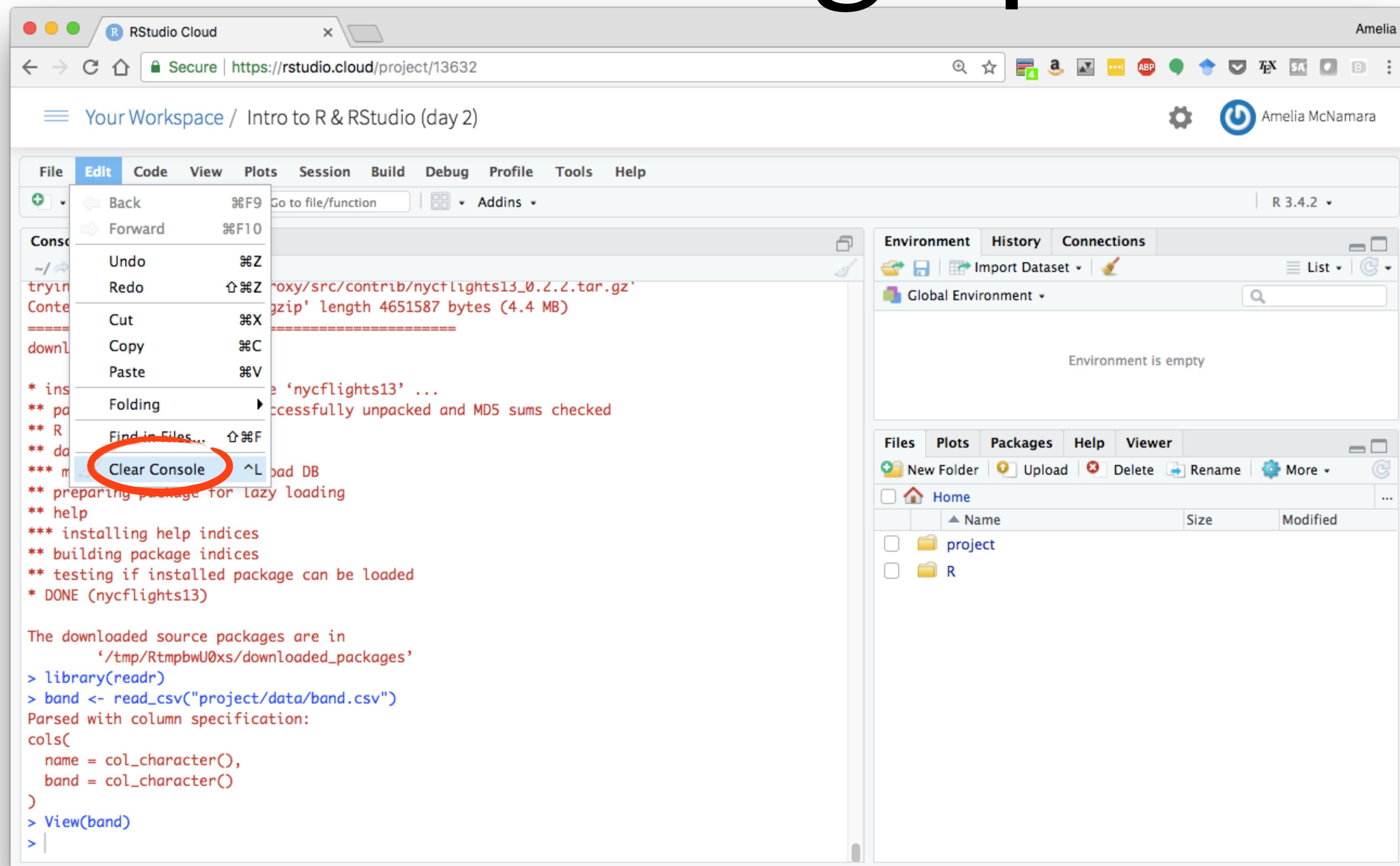


Cleaning up

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

- Header:** RStudio Cloud, Secure connection to https://rstudio.cloud/project/13632, User: Amelia McNamara.
- File Menu:** Shows options like New File, Open File..., Import Dataset, Save, Print..., Close, and Close All (which is highlighted with a red circle).
- Data View:** A data frame titled "band" is displayed with columns: MaritalStatus, NumChildren, Age, and HighestSchoolCompleted. The data shows various marital statuses (Divorced, Married) across different ages and education levels.
- Environment Tab:** Shows the Global Environment where the message "Environment is empty" is displayed.
- Files Tab:** Shows a project structure with a "project" folder and an "R" folder.
- Console Tab:** Displays R code and its output, indicating the source packages were downloaded to /tmp/RtmpbwU0xs/downloaded_packages and a CSV file was read from project/data/band.csv.

Cleaning up



Cleaning up

A screenshot of the RStudio Cloud interface. The title bar says "RStudio Cloud" and the user is "Amelia McNamara". The URL is "https://rstudio.cloud/project/13632". The sidebar shows "Your Workspace / Intro to R & RStudio (day 2)". The main window has tabs for "Console" and "Terminal". The "Console" tab is active, showing a command line with a prompt ">". To the right of the console is the "History" tab, which contains a list of R commands. A red circle highlights this list. The commands listed are:

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

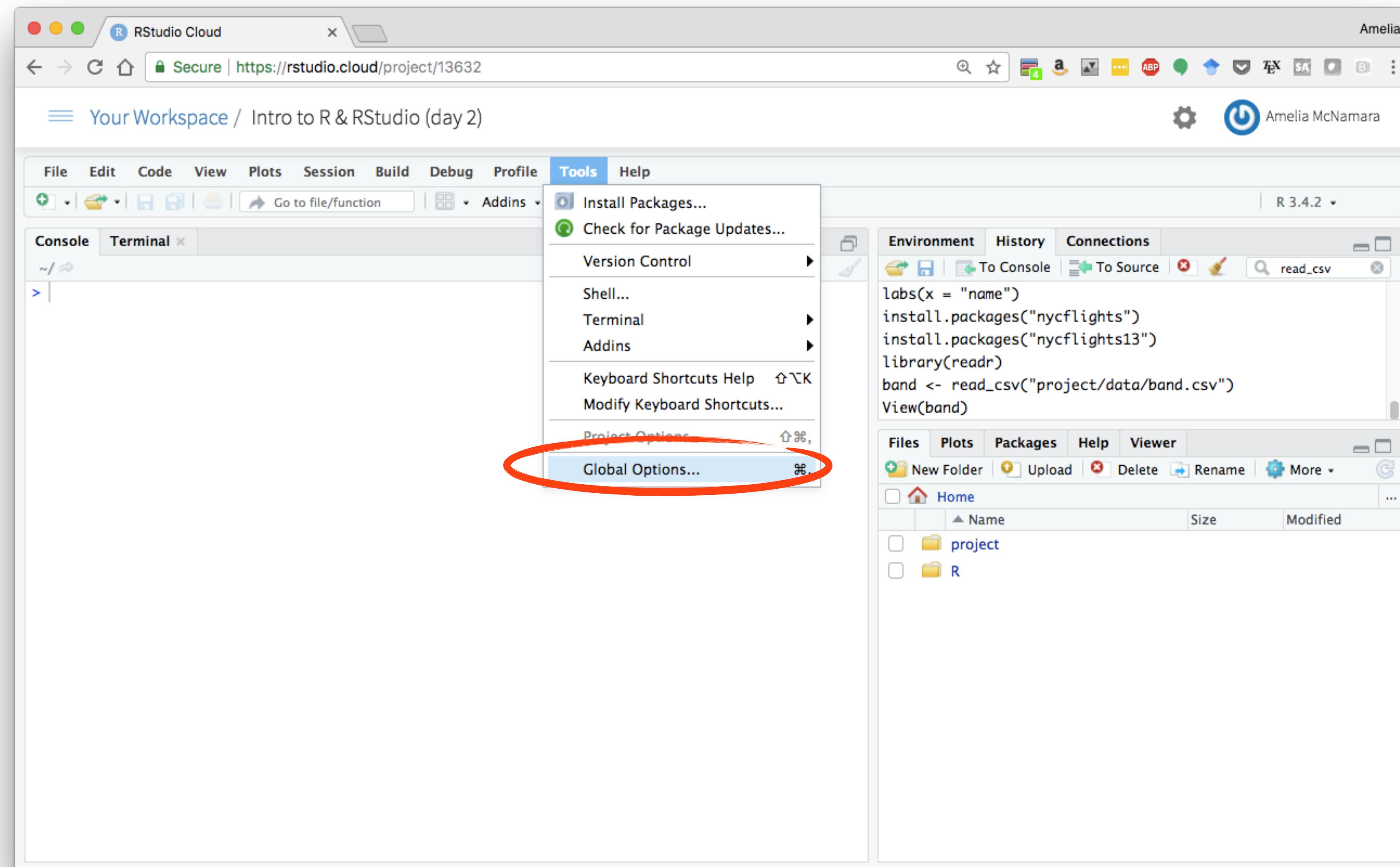
The "History" tab also has buttons for "To Console" and "To Source". Below the history tab is a file browser with tabs for "Files", "Plots", "Packages", "Help", and "Viewer". It shows a "Home" folder with two subfolders: "project" and "R".

Don't worry,
your history is
preserved

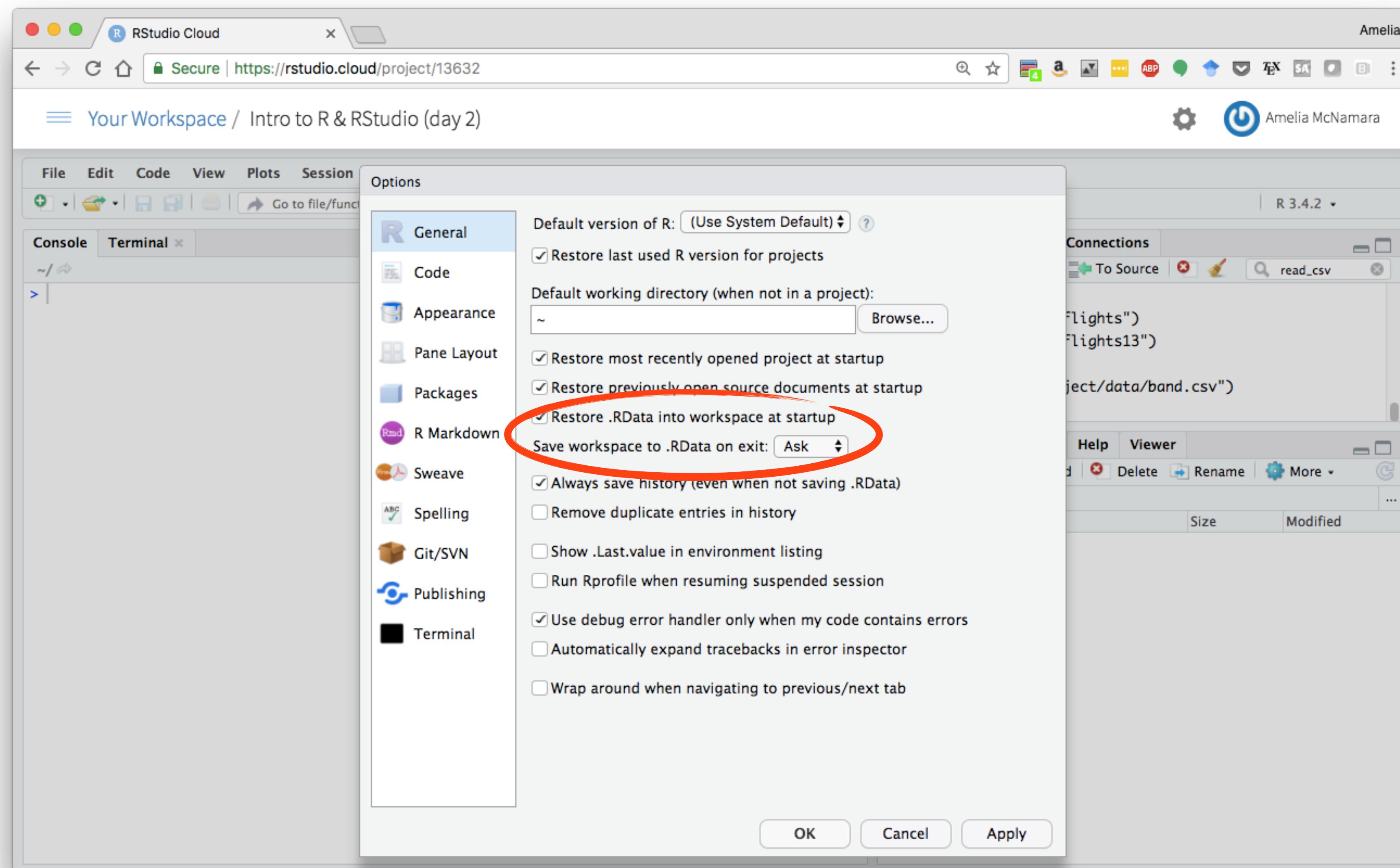
Settings



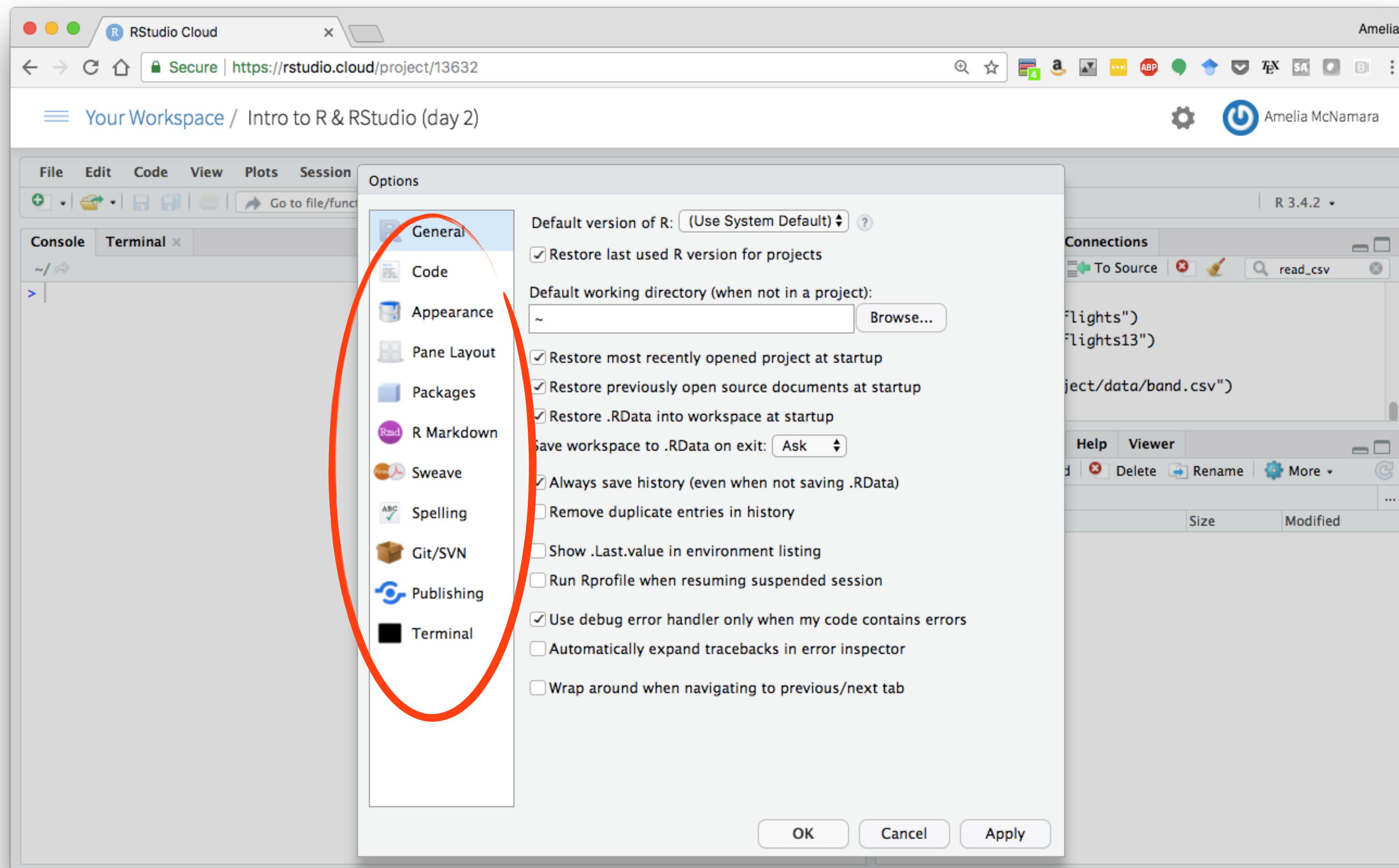
Cleaning up



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 a link to 'Learn More about RStudio features'. There are five license options listed:

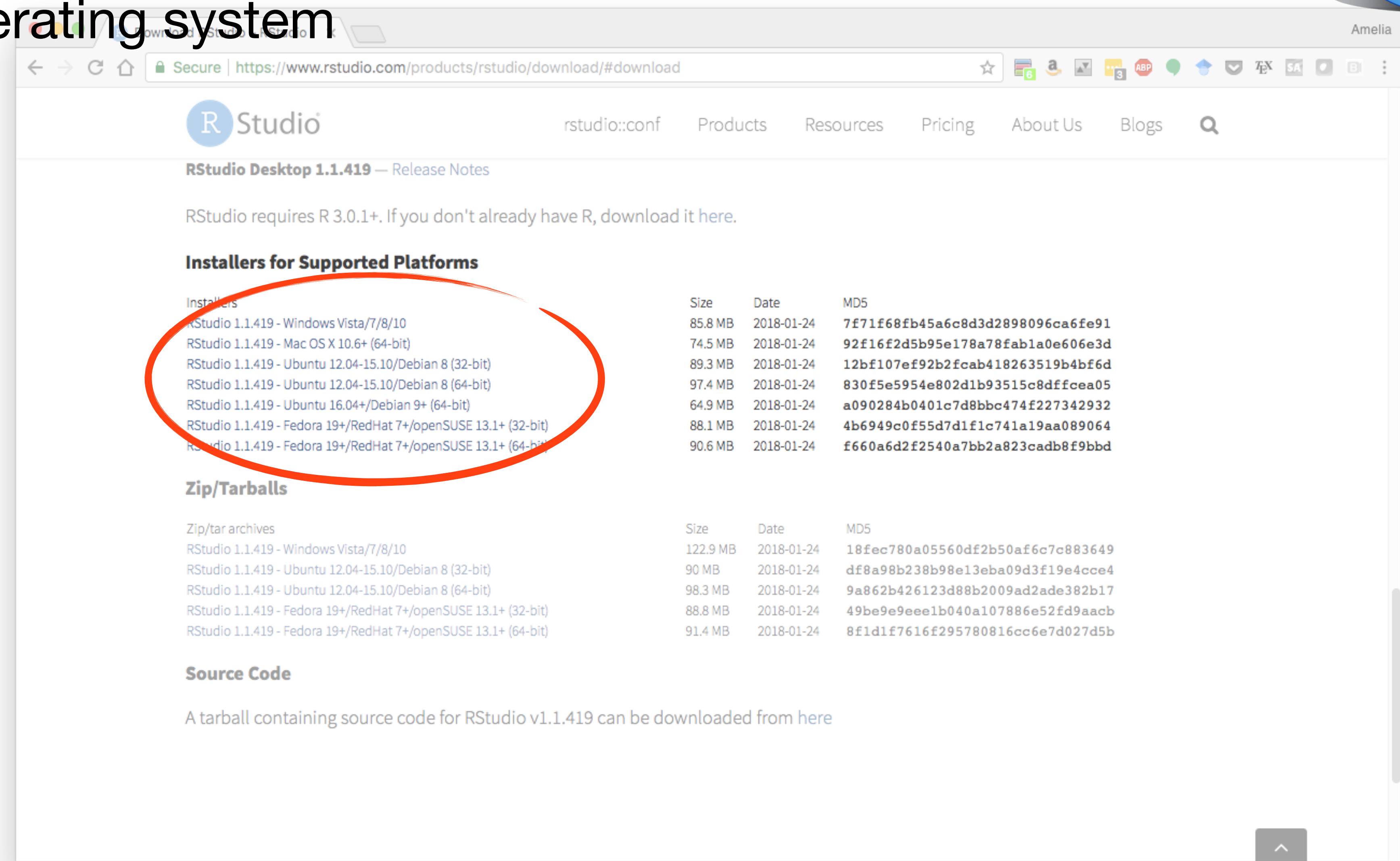
RStudio Desktop Open Source License	RStudio Desktop Commercial License	RStudio Server Open Source License	RStudio Server Pro Commercial License	RStudio Server Pro + RStudio Connect Commercial License
FREE	\$995 per year	FREE	\$9,995 per year	\$29,995 per year
DOWNLOAD Learn More	BUY Learn More	DOWNLOAD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

Below the license options, there's a section titled 'Integrated Tools for R' with a list of tools: RStudio IDE, R Markdown, R Studio Server, R Studio Connect, and R Studio Pro. Each tool has a green dot next to it.



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

	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

	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

R

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

R

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.
- Environment Browser:** A pane showing the global environment with objects like "band", "GSS", "GSS1", and "nimbus".
- File Manager:** A pane titled "Files" showing a directory structure with a "project" folder selected. A context menu is open over the "project" folder, listing options such as "Copy...", "Copy To...", "Move...", "Export...", "Set As Working Directory", and "Go To Working Directory".
- Console:** A terminal window showing R code and its output. The output includes:

```
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)
>
```

Text Overlay: Overlaid on the bottom right of the interface is the text: "You can export an entire directory from RStudio cloud".

You can export an
entire directory from
RStudio cloud

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

Your repository details have been saved.

[rstudio-conf-2020 / data-science-tidy](#)

Materials for Introduction to Data Science in the Tidyverse, a two-day workshop @ rstudio::conf 2020

Manage topics

25 commits 2 branches 0 packages 0 releases 2 contributors CC-BY-SA-4.0

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

Clone with HTTPS Use SSH
Use Git or checkout with SVN using the web URL.
<https://github.com/rstudio-conf-2020/data-science-tidy>

Open in Desktop Download ZIP
15 hours ago

6 months ago

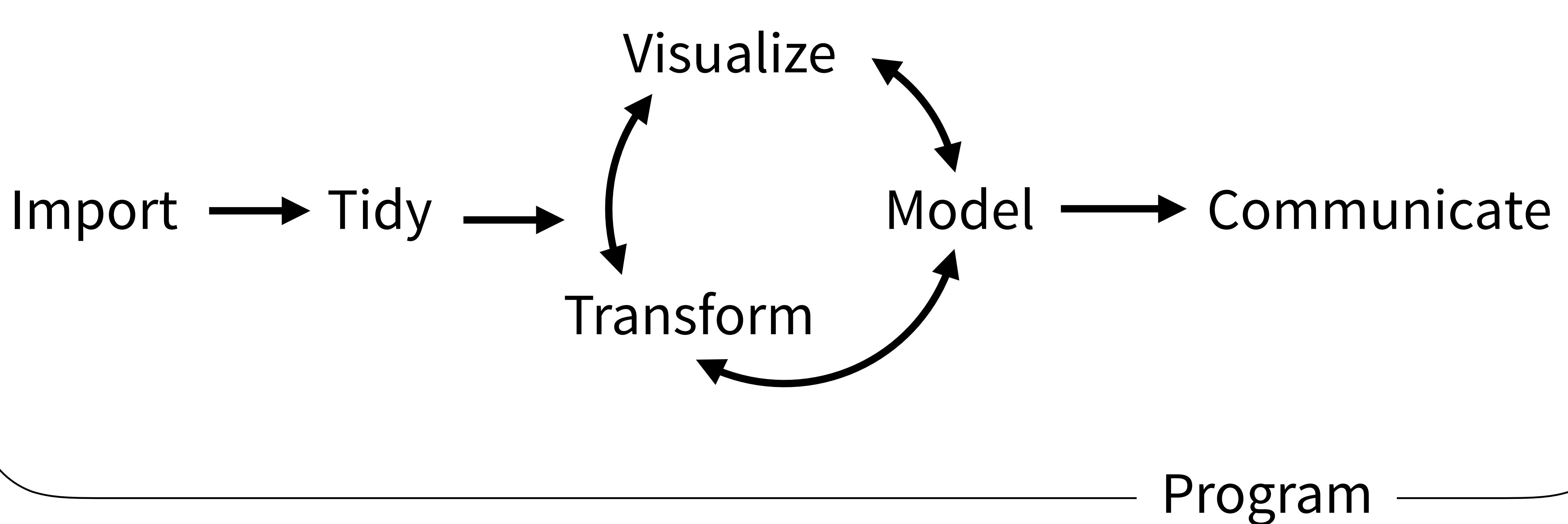
AmeliaMN 08-organize
materials
.gitignore
LICENSE.md
README.md
workshop-conf-2020.Rproj

08-organize
Add workshop repo template and license
Add workshop repo template and license
add David
Add workshop repo template and license

More to learn

R

(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

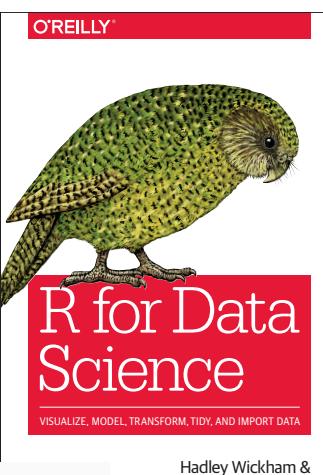
27 R Markdown

28 Graphics for communication

29 R Markdown formats

30 R Markdown workflow

**Review things
we've covered**





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

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 dropdown for 'Amelia'. 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', and 'svjour3.cls'. The latest commit was made by 'AmeliaMN' on June 20, 2016, at 445fcde. The commit message for 'data' was 'add code for making popdata.rObj'. The commit message for 'README.md' was 'update readme'. The commit message for 'SoCbib.bib' was 'bibliography'. The commit message for 'SoulOfCommunity.Rproj' was 'Add Rproj to close #1'. The commit message for 'svjour3.cls' was 'removing extra LaTeX files'.

File	Commit Message	Date
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

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

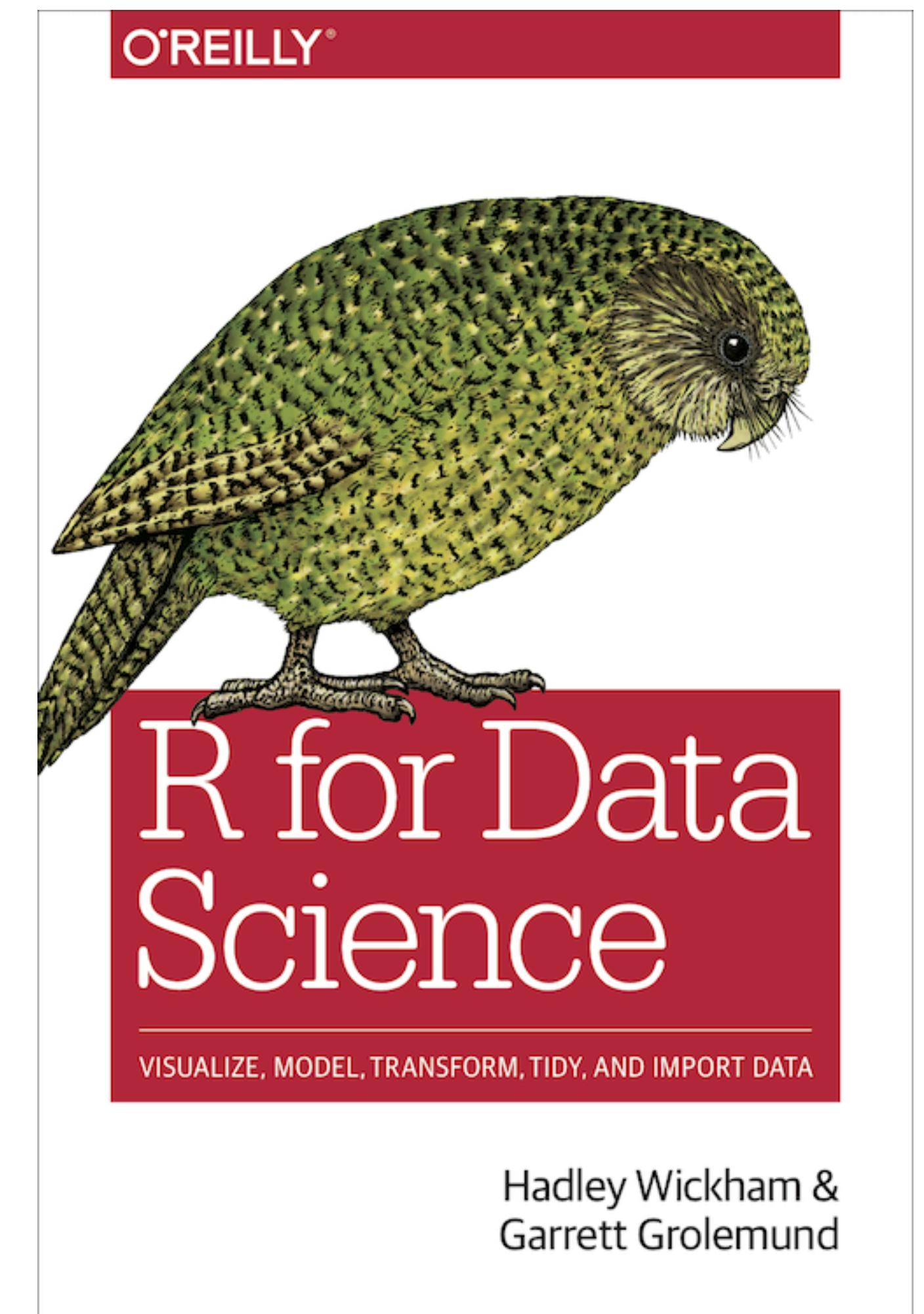
Recent commits:

File	Message	Time
<code>analysis</code>	add corresponding author email'	5 months ago
<code>data</code>	make examples match	9 months ago
<code>reviews</code>	last of Mine's comments	9 months ago
<code>.gitignore</code>	spaces around ==	10 months ago
<code>README.md</code>	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

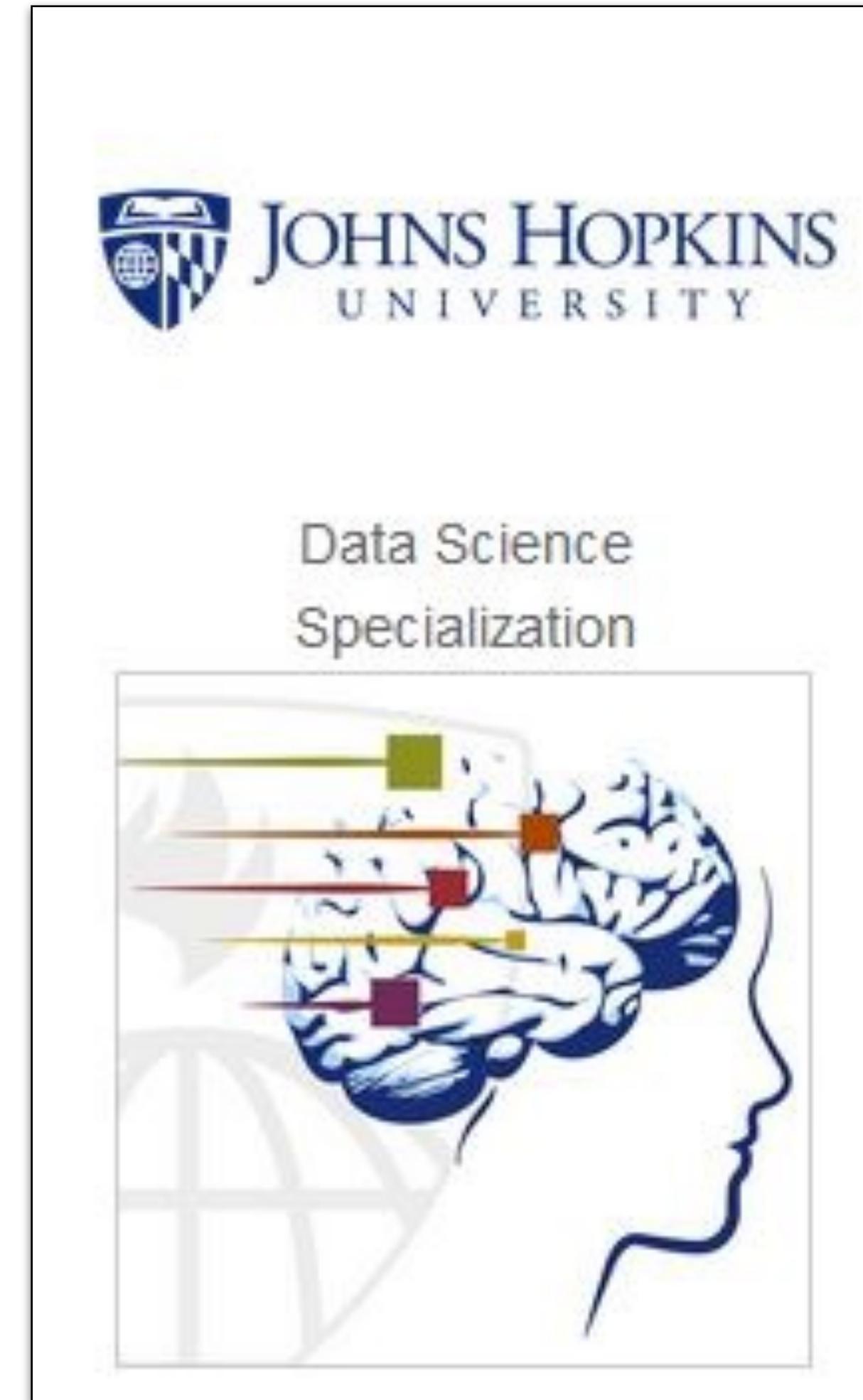
Books

- 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.



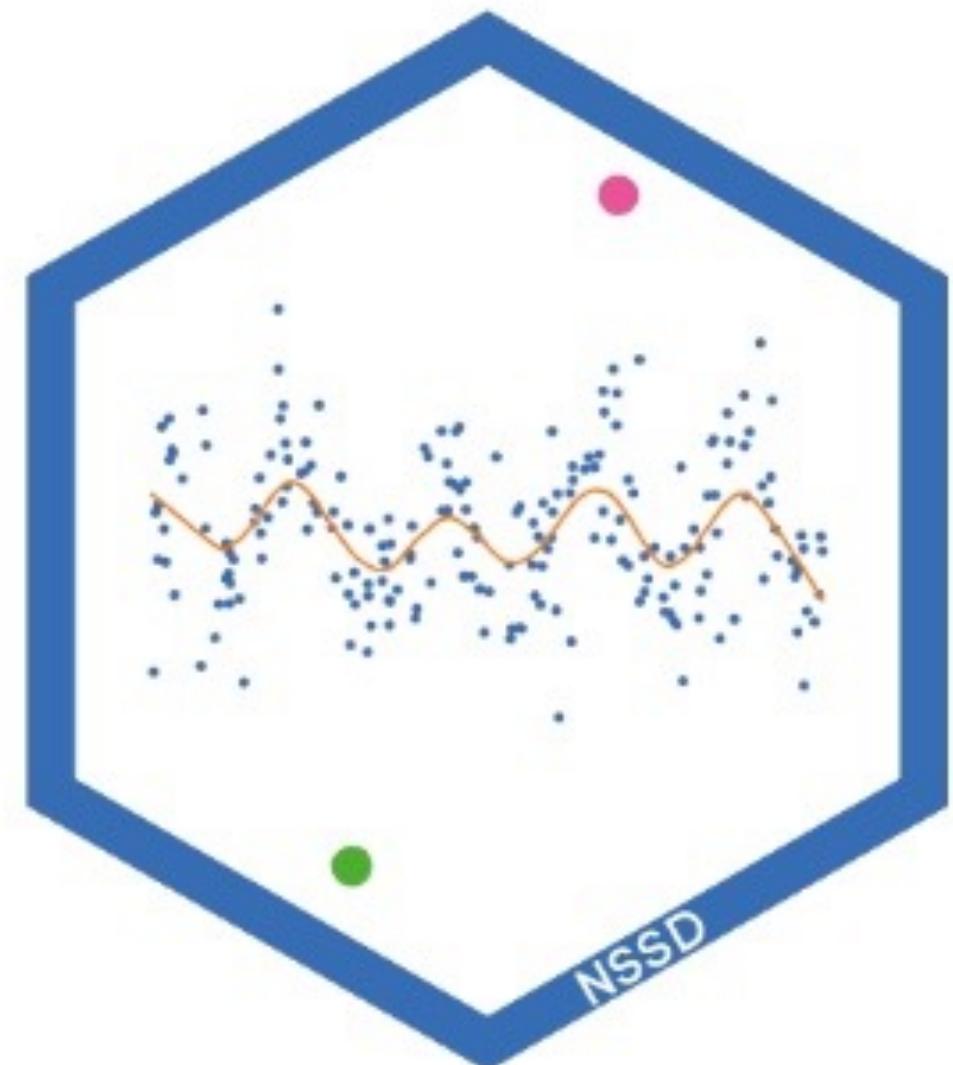
Online learning/courses

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



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

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 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

An aerial photograph of the San Francisco skyline during sunset. The city is bathed in a warm, golden light from the setting sun, which is visible on the horizon. The Transamerica Pyramid is prominent on the left, and the Golden Gate Bridge is visible in the distance across the water. The city's dense grid of buildings and streets stretches towards the horizon.

Thank you!