

These materials adapted by Amelia McNamara from the RStudio CC BY-SA materials Introduction to R (2014) and Master the Tidyverse (2017).

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

The screenshot shows the RStudio Cloud web interface. The browser address bar displays `https://rstudio.cloud/project/13632`. The page title is "Your Workspace / Intro to R & RStudio (day 2)". The user's name, "Amelia McNamara", is in the top right corner.

The main workspace area contains a data table with the following columns: Year, ID, LaborStatus, MaritalStatus, NumChildren, Age, and HighestSchoolCompleted. The table shows 8 rows of data, with a note indicating "Showing 1 to 8 of 2,540 entries".

The right-hand sidebar has three tabs: "Environment", "History", and "Connections". The "Connections" tab is selected and circled in red. It shows a list of data sources under the "Global Environment" section:

- Global Environment
- Data
 - babynames: 1858689 obs. of 5 variables
 - band: 3 obs. of 2 variables
 - GSS: 2540 obs. of 15 variables

Below the "Connections" tab is a "Files" panel showing the file structure:

- Home
 - project
 - R

The bottom panel is the "Console" window, which displays the following 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)
> |
```

Cleaning up

The screenshot shows the RStudio Cloud web interface. The browser address bar displays `https://rstudio.cloud/project/13632`. The page title is "Your Workspace / Intro to R & RStudio (day 2)". The user's name, "Amelia McNamara", is in the top right corner. The main workspace area shows a data table with columns: Year, ID, LaborStatus, MaritalStatus, NumChildren, Age, and HighestSchoolCompleted. The table contains 8 rows of data. A confirmation dialog box titled "Confirm Remove Objects" is overlaid on the workspace. The dialog contains a warning icon and the text: "Are you sure you want to remove all objects from the environment? This operation cannot be undone." Below this text is a checkbox labeled "Include hidden objects" which is currently unchecked. At the bottom of the dialog are two buttons: "Yes" and "No". The "Yes" button is circled in red. The background interface includes a menu bar with options like File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, and Help. The Environment pane on the right shows the Global Environment with objects: babynames (1858689 obs. of 5 variables), band (3 obs. of 2 variables), and GSS (2540 obs. of 15 variables). The Console pane at the bottom shows the following R code:

```
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)
> |
```

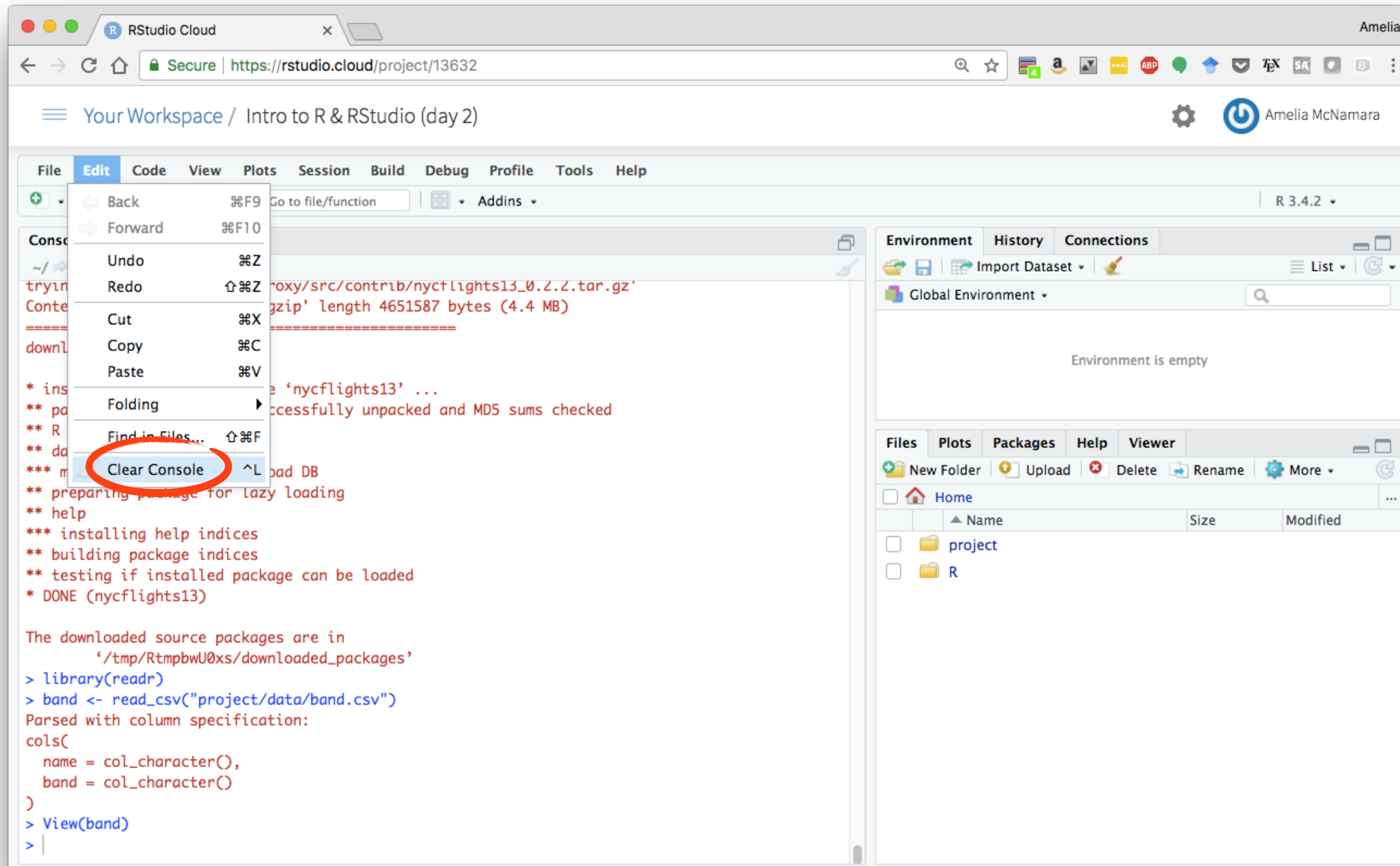

Cleaning up

The screenshot shows the RStudio Cloud web interface. The browser address bar displays <https://rstudio.cloud/project/13632>. The page title is "Your Workspace / Intro to R & RStudio (day 2)". The user's name, "Amelia McNamara", is in the top right corner. The main menu includes File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, and Help. The "File" menu is open, and the "Close All" option is highlighted with a red circle. The "band" dataset is loaded in the Environment pane, showing columns: MaritalStatus, NumChildren, Age, and HighestSchoolCompleted. The Console pane shows the R code used to load the dataset.

MaritalStatus	NumChildren	Age	HighestSchoolCompleted
Divorced	0	53.000000	16
Married	0	26.000000	16
Divorced	1	59.000000	13
Married	2	56.000000	16
Married	3	74.000000	17
Married	1	56.000000	17
Married	2	63.000000	12

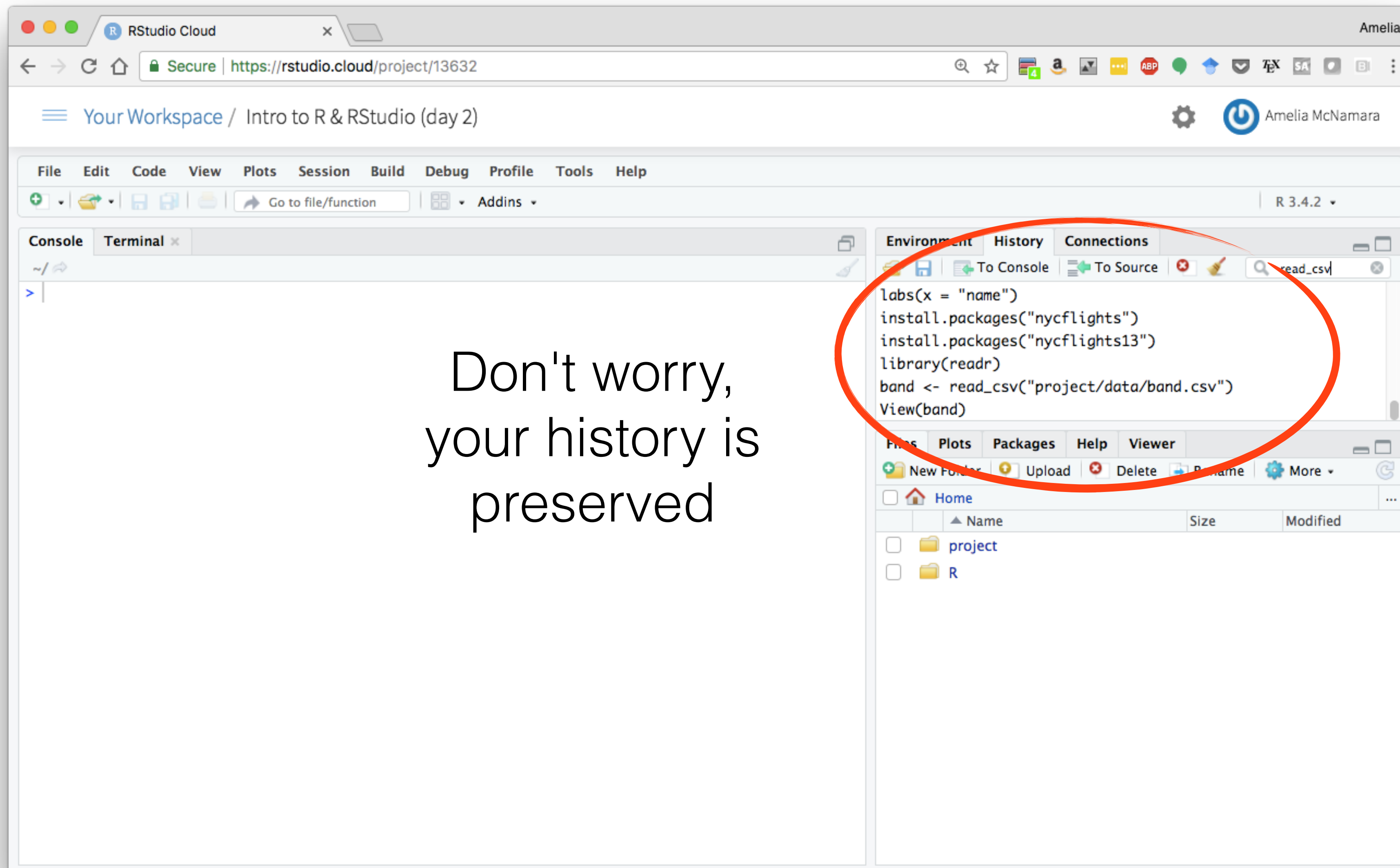
```
The downloaded source packages are in
'/tmp/RtmpbwU0xs/downloaded_packages'
> library(readr)
> band <- read_csv("project/data/band.csv")
Parsed with column specification:
cols(
  name = col_character(),
  band = col_character()
)
> View(band)
>
```

Cleaning up



Cleaning up

Don't worry,
your history is
preserved

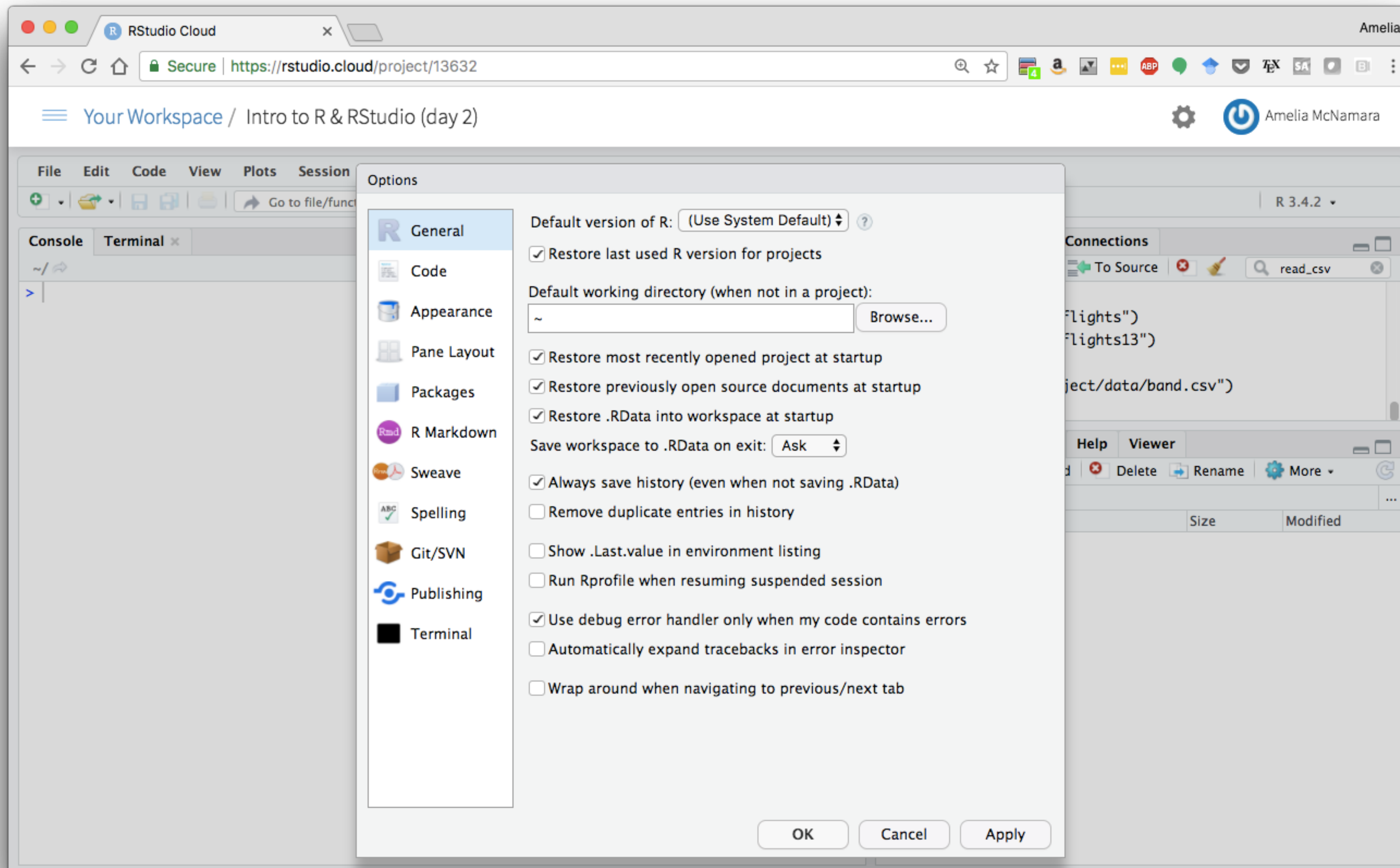


The screenshot shows the RStudio Cloud web interface. The top navigation bar includes 'Your Workspace / Intro to R & RStudio (day 2)' and the user's name 'Amelia McNamara'. The main menu bar contains 'File', 'Edit', 'Code', 'View', 'Plots', 'Session', 'Build', 'Debug', 'Profile', 'Tools', and 'Help'. Below this is a toolbar with icons for file operations and a 'Go to file/function' search bar. The left pane shows the 'Console' and 'Terminal' tabs. The right pane has tabs for 'Environment', 'History', and 'Connections'. The 'History' tab is active, displaying a list of commands executed in the session, which are circled in red. The commands are: `labs(x = "name")`, `install.packages("nycflights")`, `install.packages("nycflights13")`, `library(readr)`, `band <- read_csv("project/data/band.csv")`, and `View(band)`. Below the history pane is a file explorer showing the 'project' and 'R' folders.

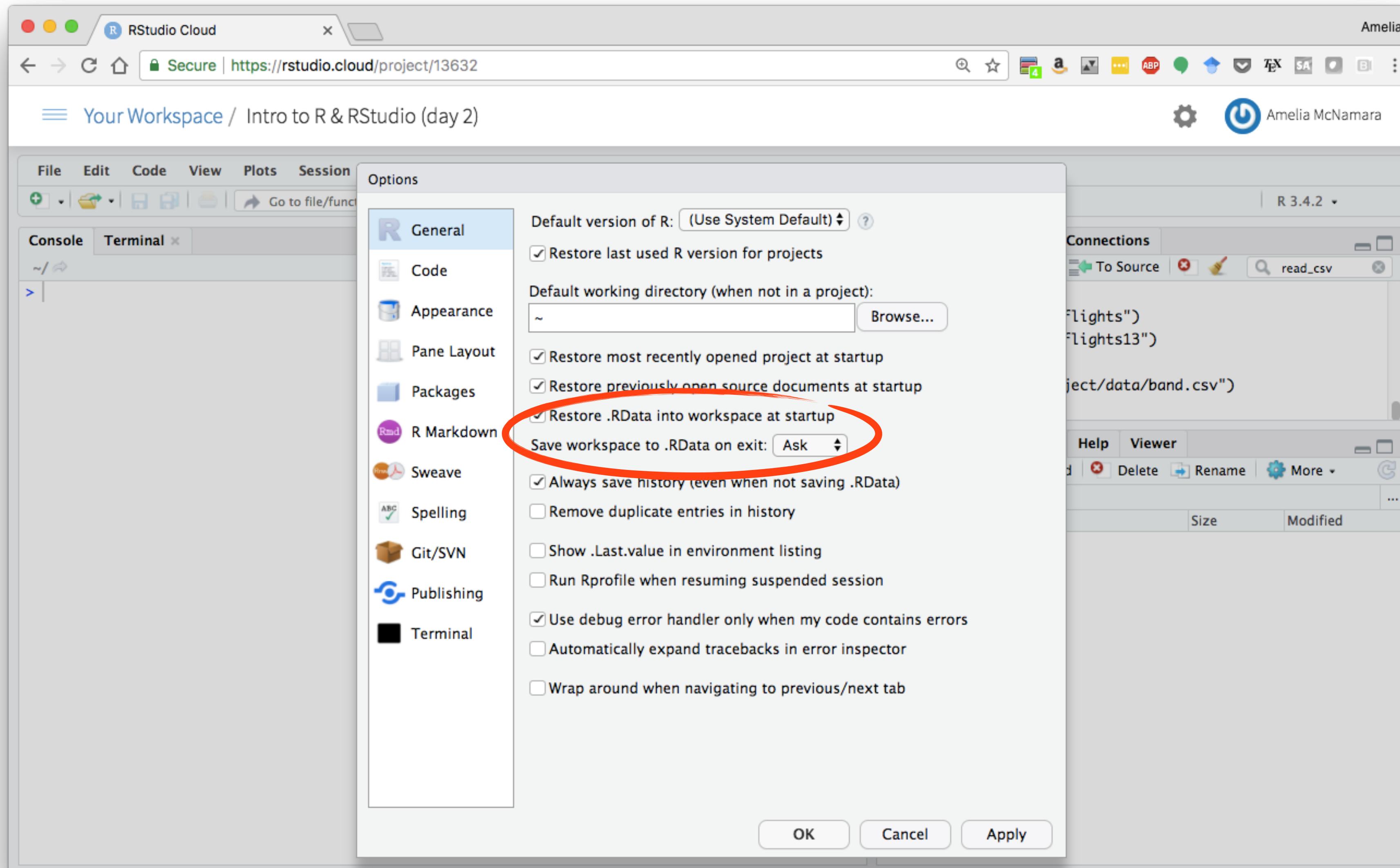
```
labs(x = "name")
install.packages("nycflights")
install.packages("nycflights13")
library(readr)
band <- read_csv("project/data/band.csv")
View(band)
```


Settings

Cleaning up



Cleaning up



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

Karl Broman, Kara Woo. Data organization in spreadsheets.
PeerJ preprint and The American Statistician.
<https://peerj.com/preprints/3183/>



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.

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>

The screenshot shows the GitHub repository page for `jtleek/datasharing`. The browser address bar displays the URL `https://github.com/jtleek/datasharing`. The repository page header includes the repository name, a search bar, and navigation links for Pull requests, Issues, Marketplace, and Explore. The repository statistics are shown as 576 Watchers, 4,252 Stars, and 182,181 Forks. The navigation tabs include Code, Issues (18), Pull requests (408), Projects (0), Wiki, and Insights. The main content area displays the README for "The Leek group guide to data sharing". The README includes a title "How to share data with a statistician" and a list of target audiences: Collaborators who need statisticians or data scientists to analyze data for them, Students or postdocs in various disciplines looking for consulting advice, and Junior statistics students whose job it is to collate/clean/wrangle data sets.

jtleek / datasharing 576 Watchers 4,252 Stars 182,181 Forks

[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

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/>

diff

1 Why Git? Why GitHub?

- 1.1 Why Git?
- 1.2 Why GitHub?
- 1.3 Is it going to hurt?
- 1.4 What is the payoff?
- 1.5 Who can do what?
- 1.6 Special features of GitHub
- 1.7 What's special about using R ...
- 1.8 Audience and pre-reqs
- 1.9 What this is NOT

6 Install or upgrade R and RStudio

- 6.1 More about R and RStudio

8 Introduce yourself to Git

- 8.1 More about `git config`

9 Install a Git client


- 9.1 What and why
- 9.2 A picture is worth a thousand ...
- 9.3 Recommended Git clients

Happy Git and GitHub for the useR

Jenny Bryan and the STAT 545 TAs

2018-01-28

A work in progress



WATCH ME DIFF
WATCH ME REBASE

Example paper and file structure:

<https://github.com/COSTDataExpo2013/AmeliaMN>

The screenshot displays the GitHub interface for the repository **COSTDataExpo2013 / AmeliaMN**. The repository has 44 commits, 1 branch, 0 releases, and 2 contributors. The latest commit is 445fcdc on Jun 20, 2016.

The repository structure is as follows:

File/Folder	Description	Time
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

The repository also includes a **README.md** file.

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>

September 1, 2017 **preprint**



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/>

The screenshot shows a web browser window with the URL <https://github.com/dsscollection/>. The page displays the GitHub organization profile for "dsscollection". The organization's logo is a teal square with a white cross-like pattern. The name "dsscollection" is in a large, bold, black font. Below the name, there are tabs for "Repositories" (16), "People" (35), "Teams" (1), and "Projects" (0). A search bar is present with the placeholder text "Search repositories...". To the right of the search bar are filters for "Type: All" and "Language: All". The main content area lists three repositories:

- dsscollection** (Private): Repo for coordination and admin. Language: HTML. 6 forks. Updated on Dec 10, 2017.
- stat-comp-curriculum** (Private): Language: TeX. Updated on Oct 15, 2017.
- git-github-for-stats**: Git and GitHub evangelism for the practicing statistician. Language: TeX. 1 star. Updated on Oct 5, 2017.

Each repository entry includes a green line graph showing activity over time. On the right side of the page, there is a section titled "Top languages" showing a list of languages: TeX, HTML, Python, and R. Below this is a section titled "People" showing 35 members, with a grid of 12 profile pictures and a "35 >" link to view all members.

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

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