

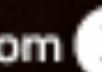


Introduction to Data Science in the Tidyverse

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>



Workshop Policies

- Identify the exits closest to you in case of emergency
- Please review the rstudio::conf code of conduct that applies to all workshops. Issues can be addressed three ways:
 - In person: contact any rstudio::conf staff member or the conference registration desk
 - By email: send a message to conf@rstudio.com
 - By phone: call 844-448-1212
- Please do not photograph people wearing red lanyards
- A chill-out room is available for neurologically diverse attendees on the 4th floor of tower 1



Introducing your teaching staff

Instructors:

Amelia McNamara (she/her) @AmeliaMN
Hadley Wickham (he/him) @hadleywickham

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

Your Turn

Introduce yourself to your neighbors:

- Who are you?
- What you do with data?
- How would you describe your experience with R?



No sticky note: "I'm happily working on it"



Blue sticky note: "I'm all done and ready to move on"



Orange sticky note: "I'm stuck, can someone help me?"

Alternatively, flag one of us down



Hopefully, color-blind friendly, let me know if not.

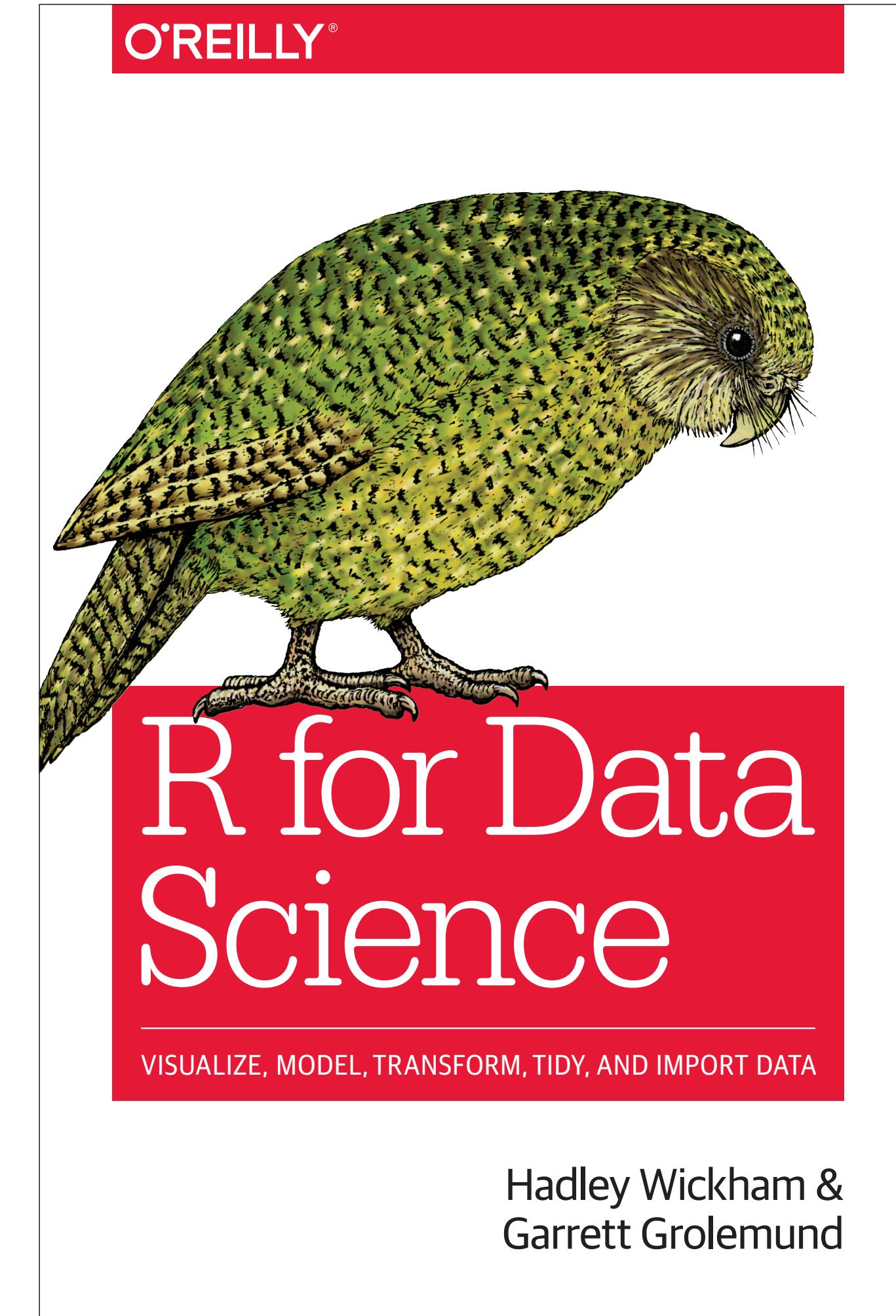
This class is heavily based on
R for Data Science

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

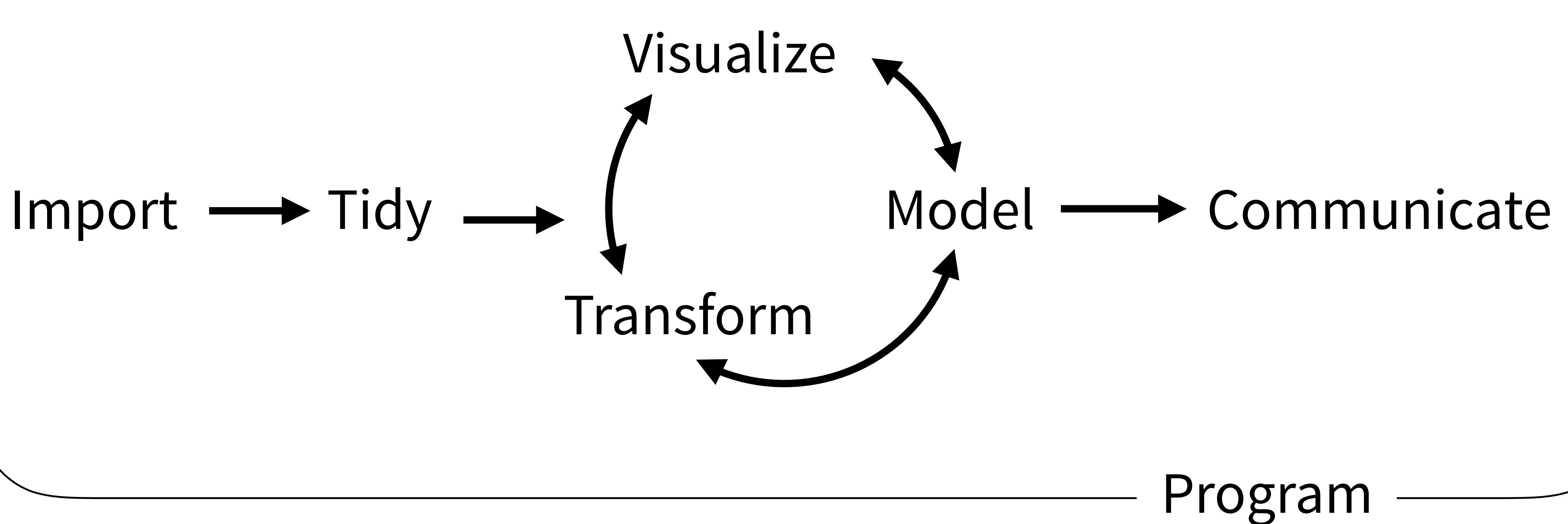
Links to the relevant
sections of the book



In R4DS
Introduction

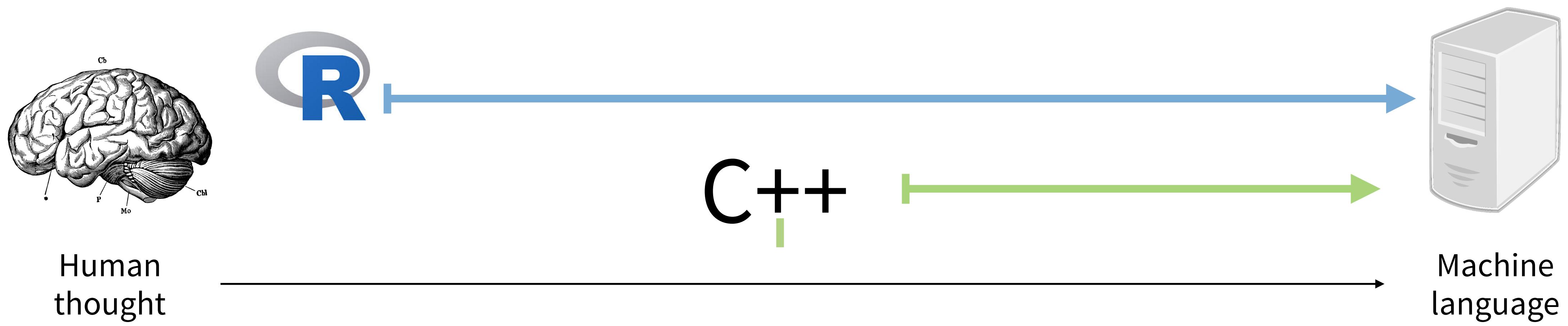


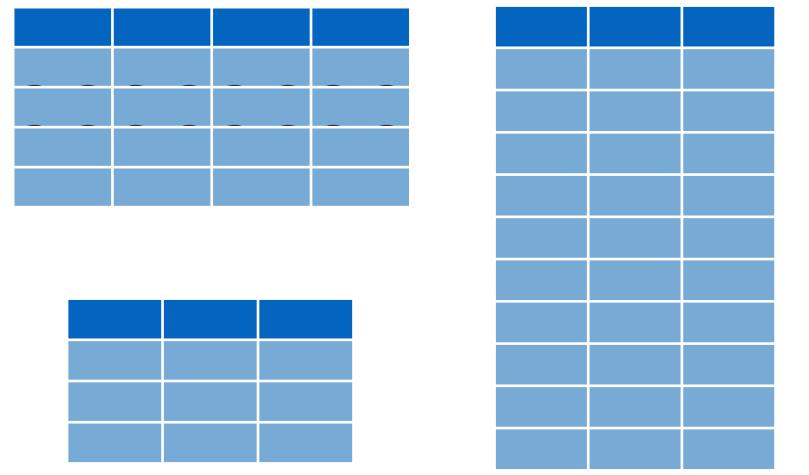
(Applied) Data Science



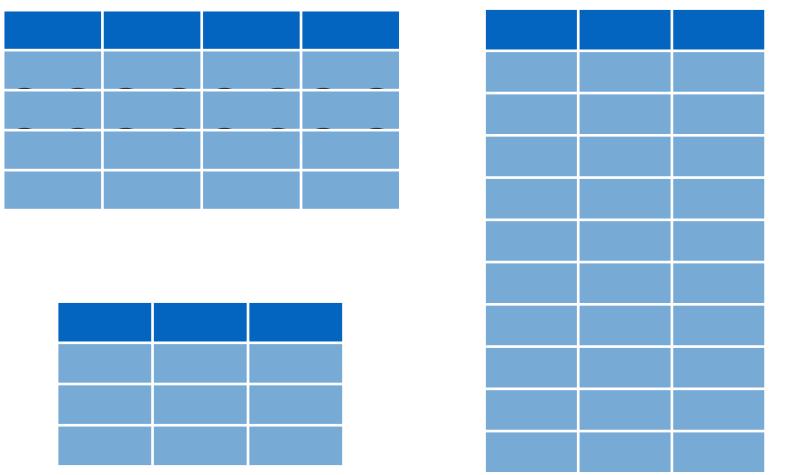


R - A computer language for scientists

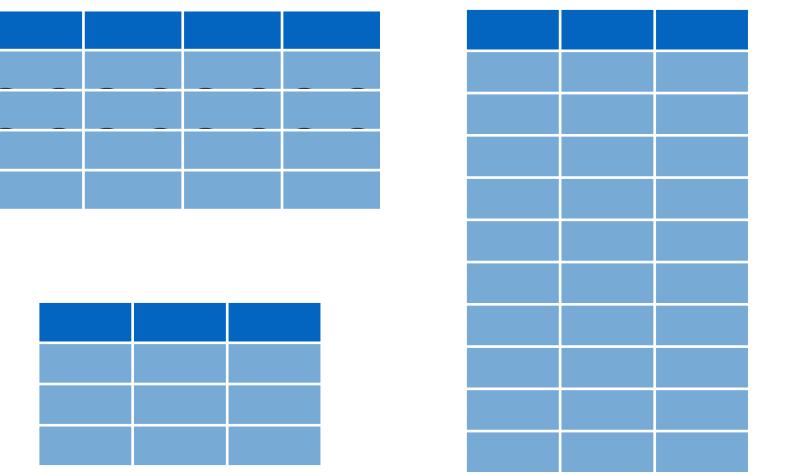
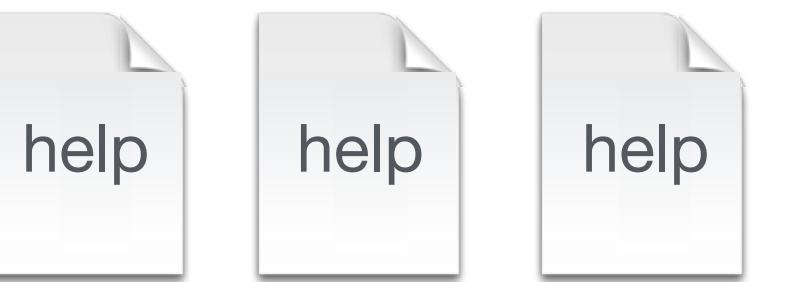




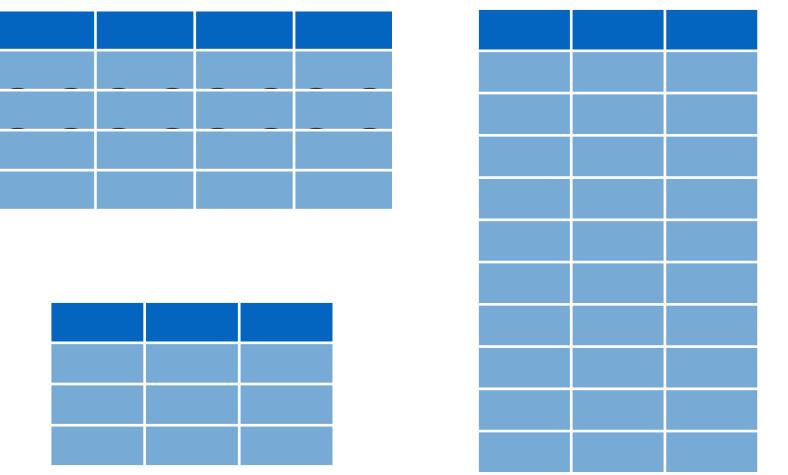
function1()
function2()
function3()
function4()



function1()
function2()
function3()
function4()



function5()
function6()
function7()
function8()

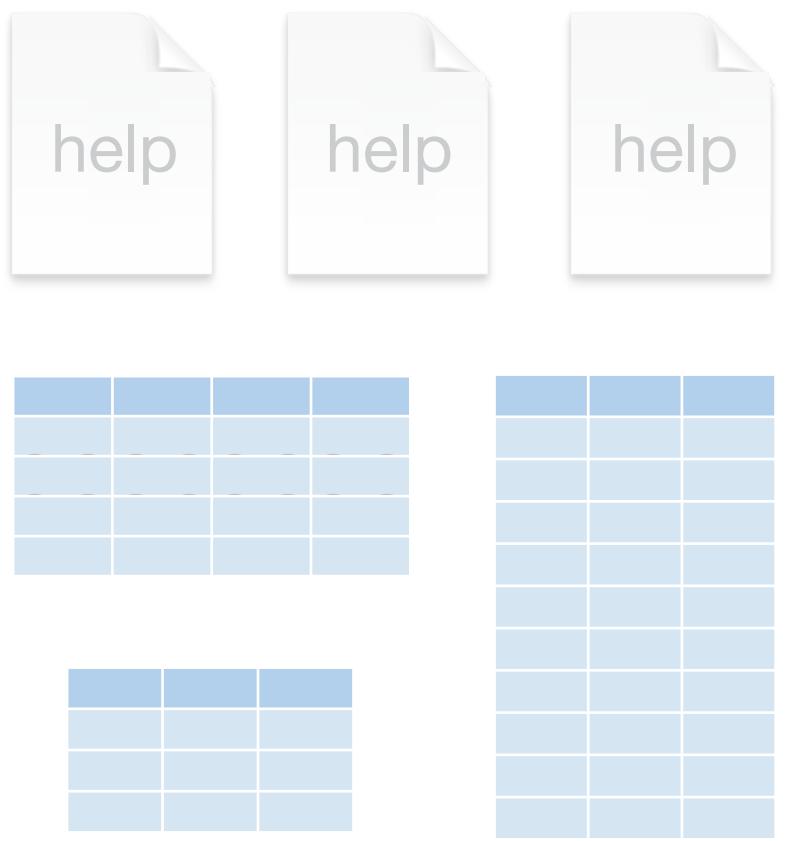


function9()
functionA()
functionB()
functionC()

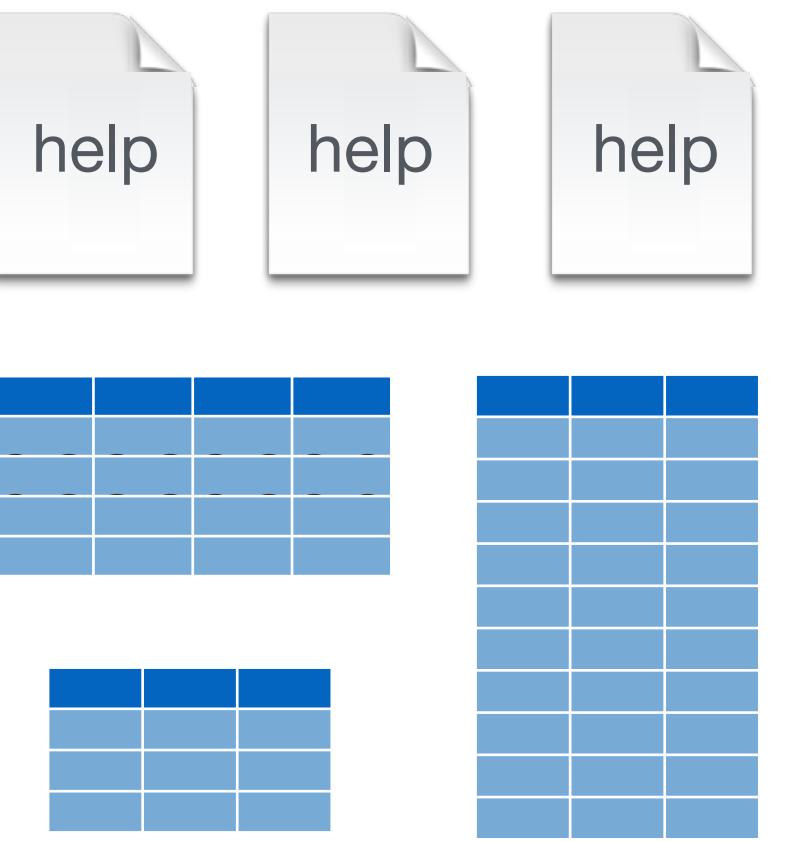


functionD()
functionE()
functionF()
functionG()

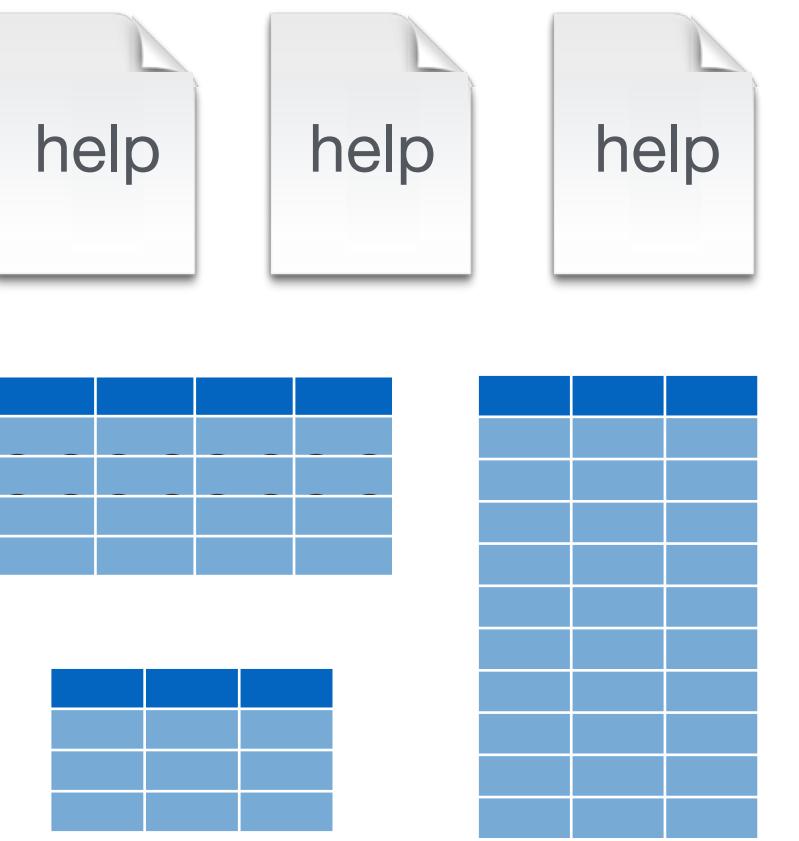
Base R



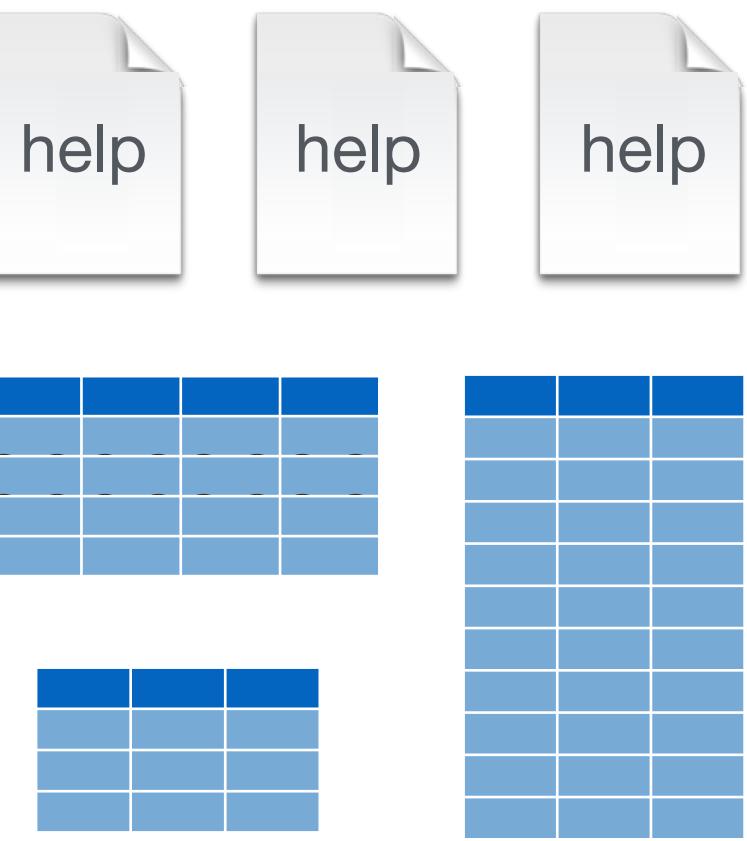
function1()
function2()
function3()
function4()



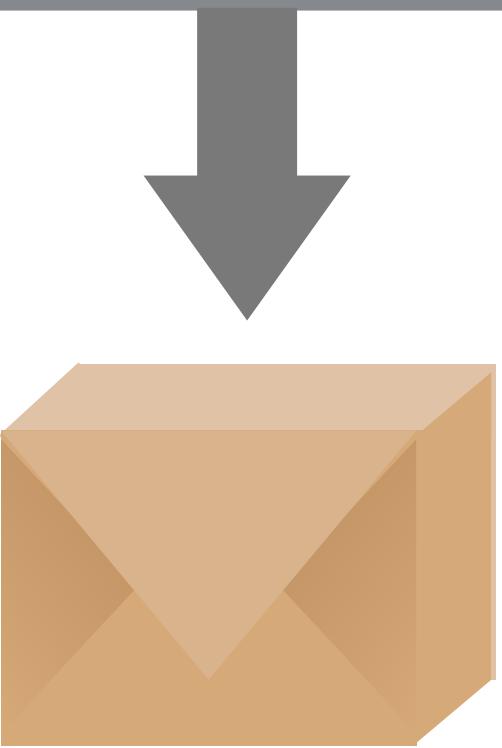
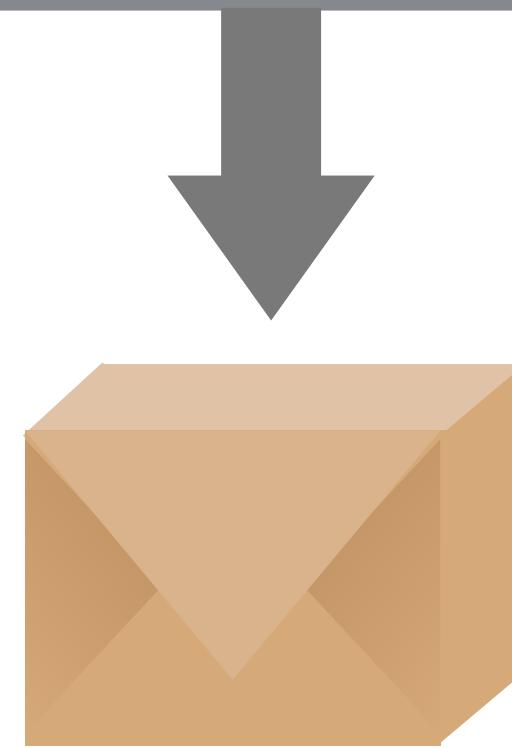
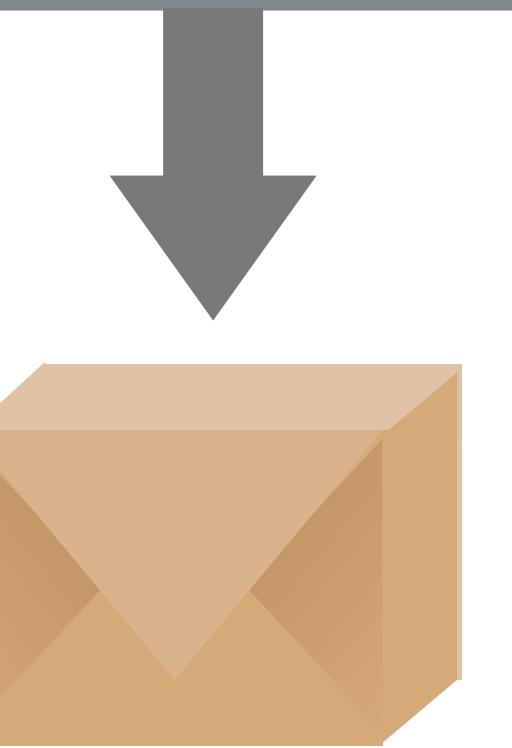
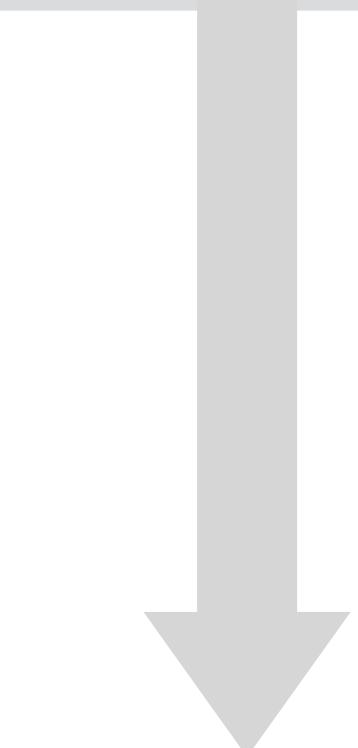
function5()
function6()
function7()
function8()



function9()
functionA()
functionB()
functionC()



functionD()
functionE()
functionF()
functionG()



Base R

R Packages

The screenshot shows a web browser window with the title "The Comprehensive R Archive". The address bar indicates a secure connection to "https://cran.r-project.org". In the top right corner, there is a user profile icon with the name "Garrett". The main content area displays the "Available CRAN Packages By Name" page, featuring a large blue "R" logo on the left. Below the logo are several navigation links: "CRAN Mirrors", "What's new?", "Task Views", "Search", "About R", "R Homepage", "The R Journal", "Software", "R Sources", "R Binaries", "Packages", and "Other". To the right, a large section titled "Available CRAN Packages By Name" lists packages grouped by letter, from A to Z. Each package name is a link to its description.



[CRAN
Mirrors](#)
[What's new?](#)
[Task Views](#)
[Search](#)

[About R](#)
[R Homepage](#)
[The R Journal](#)

[Software](#)
[R Sources](#)
[R Binaries](#)
[Packages](#)
[Other](#)

[Documentation](#)
[Manuals](#)
[FAQs](#)
[Contributed](#)

[A3](#)
[abyyR](#)
[abc](#)
[ABCAnalysis](#)
[abc.data](#)
[abcdeFBA](#)

[ABCOptim](#)
[ABCp2](#)
[ABC.RAP](#)
[abcrf](#)
[abctools](#)
[abd](#)
[abf2](#)
[ABHgenotypeR](#)
[abind](#)
[abjutils](#)
[abn](#)
[abodOutlier](#)

Available CRAN Packages By Name

[A](#) [B](#) [C](#) [D](#) [E](#) [F](#) [G](#) [H](#) [I](#) [J](#) [K](#) [L](#) [M](#) [N](#) [O](#) [P](#) [Q](#) [R](#) [S](#) [T](#) [U](#) [V](#) [W](#) [X](#) [Y](#) [Z](#)

[Accurate, Adaptable, and Accessible Error Metrics for Predictive Models](#)

[Access to Abbyy Optical Character Recognition \(OCR\) API](#)

[Tools for Approximate Bayesian Computation \(ABC\)](#)

[Computed ABC Analysis](#)

[Data Only: Tools for Approximate Bayesian Computation \(ABC\)](#)

[ABCDE_FBA: A-Biologist-Can-Do-Everything of Flux Balance Analysis with this package](#)

[Implementation of Artificial Bee Colony \(ABC\) Optimization](#)

[Approximate Bayesian Computational Model for Estimating P2](#)

[Array Based CpG Region Analysis Pipeline](#)

[Approximate Bayesian Computation via Random Forests](#)

[Tools for ABC Analyses](#)

[The Analysis of Biological Data](#)

[Load Gap-Free Axon ABF2 Files](#)

[Easy Visualization of ABH Genotypes](#)

[Combine Multidimensional Arrays](#)

[Useful Tools for Jurimetrical Analysis Used by the Brazilian Jurimetrics Association](#)

[Modelling Multivariate Data with Additive Bayesian Networks](#)

[Angle-Based Outlier Detection](#)

Using packages

1

```
install.packages("foo")
```

Downloads files to "computer"

1 x per "computer"

2

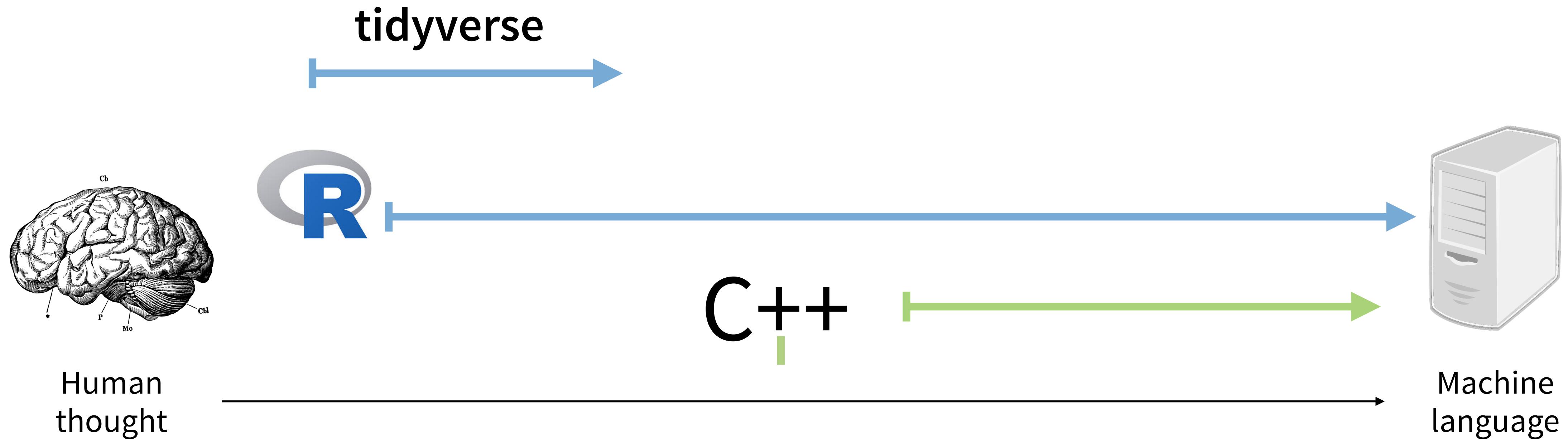
```
library("foo")
```

Loads package

1 x per R Session

I've done this
for you for this
workshop

The tidyverse - A set of R packages to unify some data science tasks



tidyverse.org

The screenshot shows the homepage of tidyverse.org. The top navigation bar includes links for Packages, Articles, Learn, Help, and Contribute. Below the navigation, there's a large graphic featuring hexagonal icons for various R packages: dplyr (orange, with pliers icon), ggplot2 (grey, with line plot icon), readr (blue, with document icon), purrr (white with cat icon), tibble (dark blue, with grid icon), and tidyr (orange, with arrows icon). To the right of the graphic, the text reads: "R packages for data science. The tidyverse is an opinionated **collection of R packages** designed for data science. All packages share an underlying philosophy and common APIs." At the bottom, there's a code block:

```
install.packages("tidyverse")
```

Tidyverse

Packages Articles Learn Help Contribute

R packages for data science

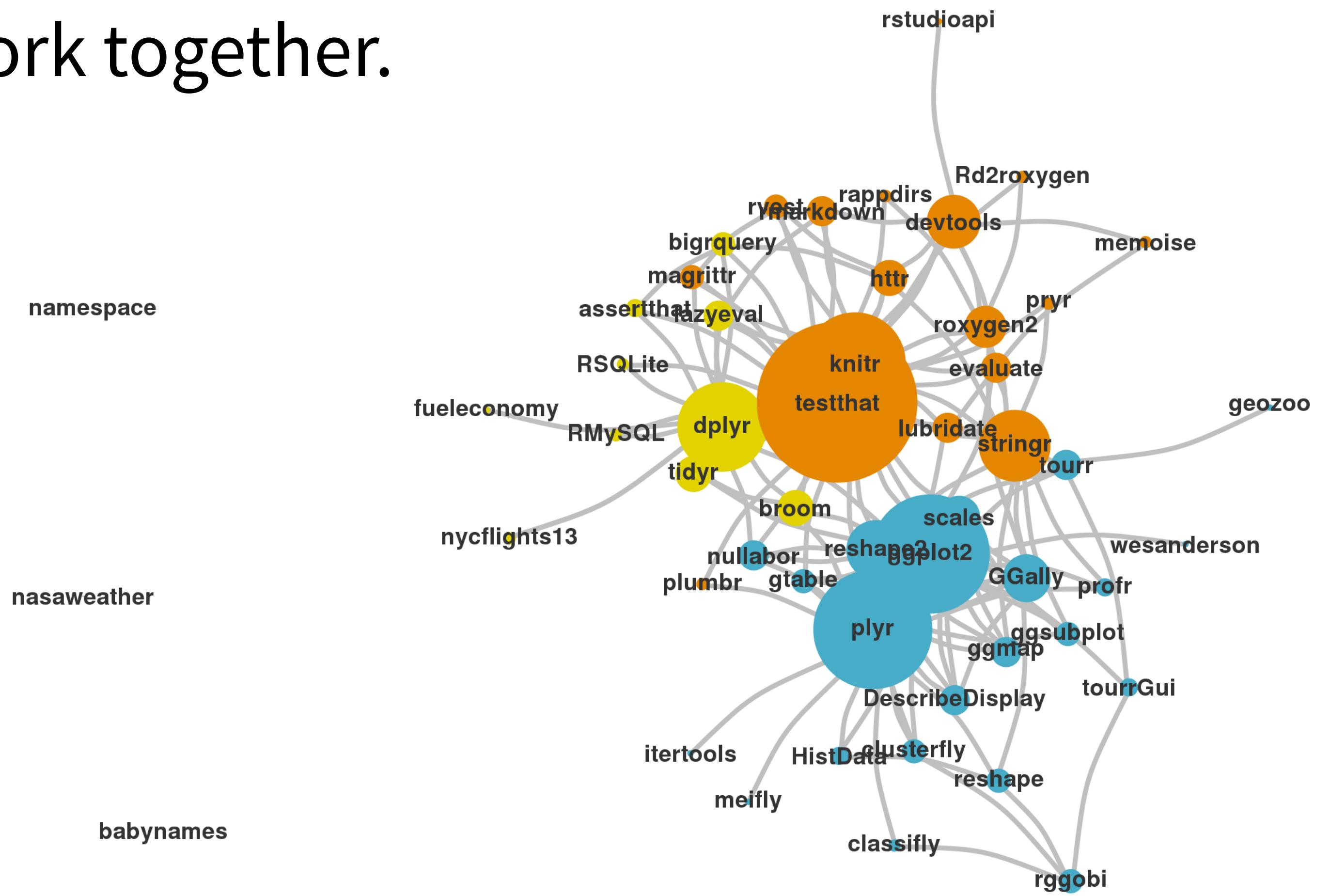
The tidyverse is an opinionated **collection of R packages** designed for data science. All packages share an underlying philosophy and common APIs.

Install the complete tidyverse with:

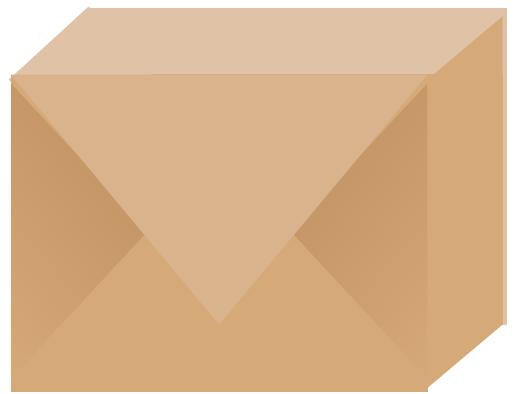
```
install.packages("tidyverse")
```

The Tidyverse

A collection of modern R packages that share common philosophies, embed best practices, and are designed to work together.



tidyverse



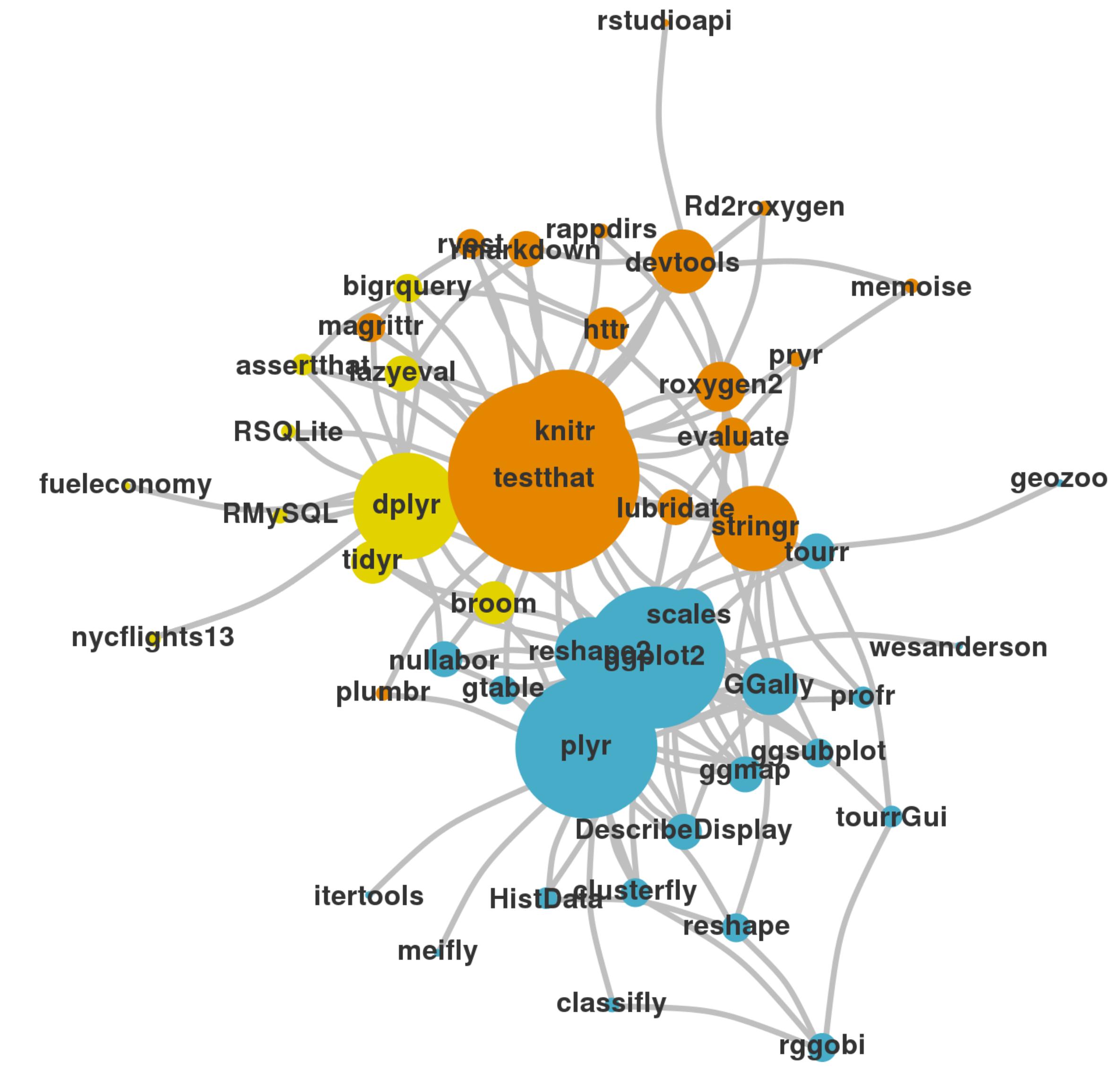
An R package that serves as a short cut for installing and loading the components of the tidyverse.

```
library("tidyverse")
```

```
install.packages("tidyverse")
```

does the equivalent of

```
install.packages("ggplot2")
install.packages("dplyr")
install.packages("tidyr")
install.packages("readr")
install.packages("purrr")
install.packages("tibble")
install.packages("hms")
install.packages("stringr")
install.packages("lubridate")
install.packages("forcats")
install.packages("DBI")
install.packages("haven")
install.packages("httr")
install.packages("jsonlite")
install.packages("readxl")
install.packages("rvest")
install.packages("xml2")
install.packages("modelr")
install.packages("broom")
```



```
install.packages("tidyverse")
```

does the equivalent of

```
install.packages("ggplot2")
install.packages("dplyr")
install.packages("tidyr")
install.packages("readr")
install.packages("purrr")
install.packages("tibble")
install.packages("hms")
install.packages("stringr")
install.packages("lubridate")
install.packages("forcats")
install.packages("DBI")
install.packages("haven")
install.packages("httr")
install.packages("jsonlite")
install.packages("readxl")
install.packages("rvest")
install.packages("xml2")
install.packages("modelr")
install.packages("broom")
```

```
library("tidyverse")
```

does the equivalent of

```
library("ggplot2")
library("dplyr")
library("tidyr")
library("readr")
library("purrr")
library("tibble")
```

Day 1

Introduction and
Visualize Data

9:00 - 10:30

Morning Break

10:30 - 11:00

Visualize and Transform

11:00 - 12:30

Lunch

12:30 - 2:00

Transform

2:00 - 3:30

Afternoon Break

3:30 - 4:00

Tidy Data/
Case Study

4:00 - 5:00

Day 2

Data types

9:00 - 10:30

Morning Break

10:30 - 11:00

Iteration with purr

11:00 - 12:30

Lunch

12:30 - 2:00

Modeling and organization

2:00 - 3:30

Afternoon Break

3:30 - 4:00

Organize and wrap-up

4:00 - 5:00



RStudio: a software program

1. like Microsoft Word, Excel, etc.
2. built to help you write R code, run R code, and analyze data with R
3. text editor, version control, keyboard shortcuts, debugging tools, and much more

RMarkdown

(let's start!)

Your Turn

If you haven't already

- visit http://bit.ly/DS_TV_2020 (all caps after the slash, underscores)
- Log In / Sign Up
- Join Space
- Go to Projects and then click on **workshop**
- open **00-Getting-started.Rmd**

When you are finished, let us know by putting up the  icon.



RMarkdown

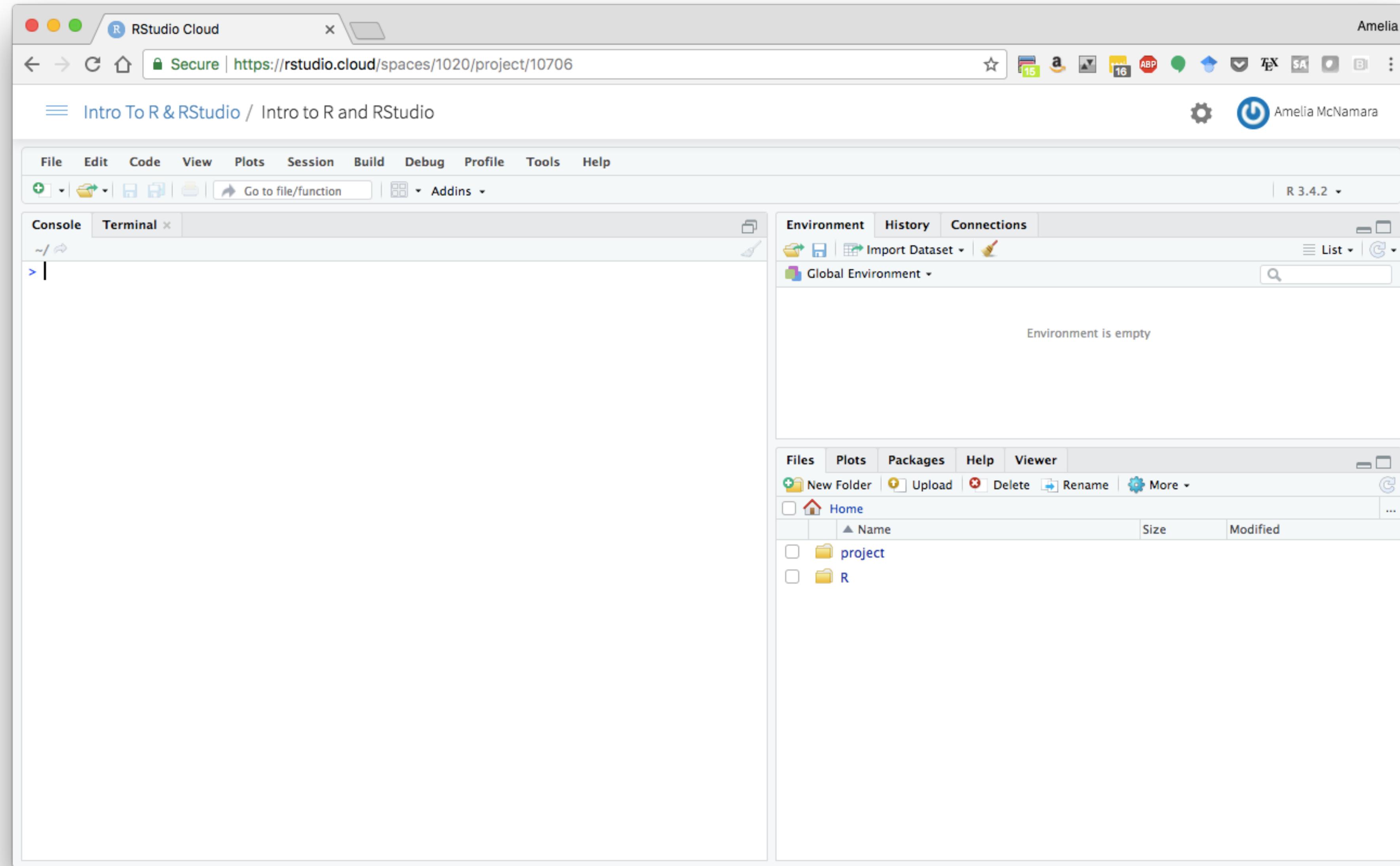
An authoring format for Data Science.

A screenshot of the RStudio interface showing an R Markdown document titled "Untitled1". The code editor pane displays the following R Markdown code:

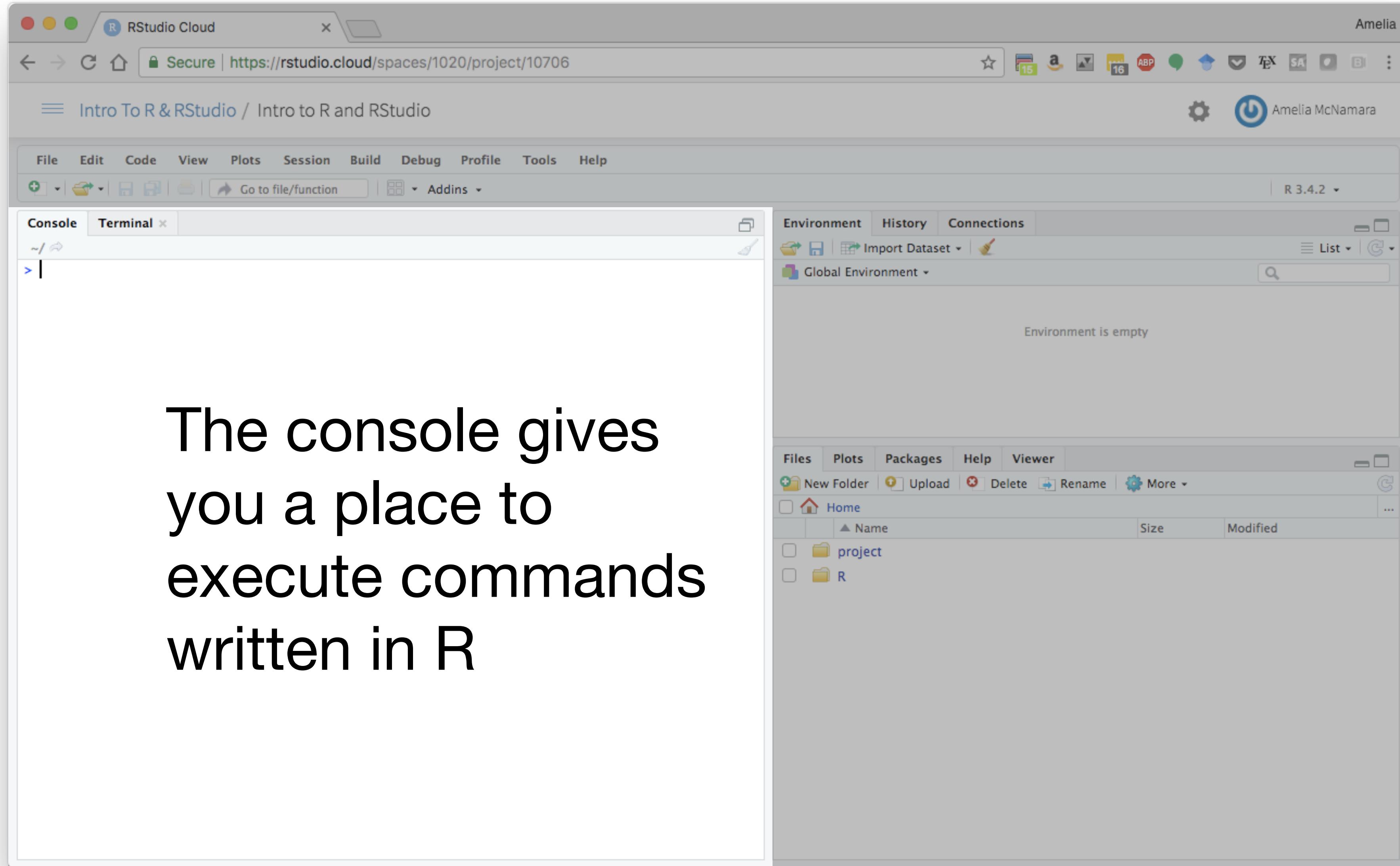
```
1 ---  
2 title: "My document"  
3 author: "Amelia McNamara"  
4 date: "1/23/2020"  
5 output: html_document  
---  
8 ```{r setup, include=FALSE}  
9 knitr::opts_chunk$set(echo = TRUE)  
10 ```  
11  
12 ## R Markdown  
13  
14 This is an R Markdown document. Markdown is a simple formatting syntax for authoring  
HTML, PDF, and MS Word documents.  
http://rmarkdown.rstudio.com.  
15  
16 When you click the **Knit** button a document will be generated that includes both  
content as well as the output of any embedded R code chunks within the document. You  
can embed an R code chunk like this:  
17  
18 ```{r cars}  
19 summary(cars)  
20 ````
```

The RStudio interface includes toolbars, a file menu, and a status bar at the bottom.

RStudio



RStudio

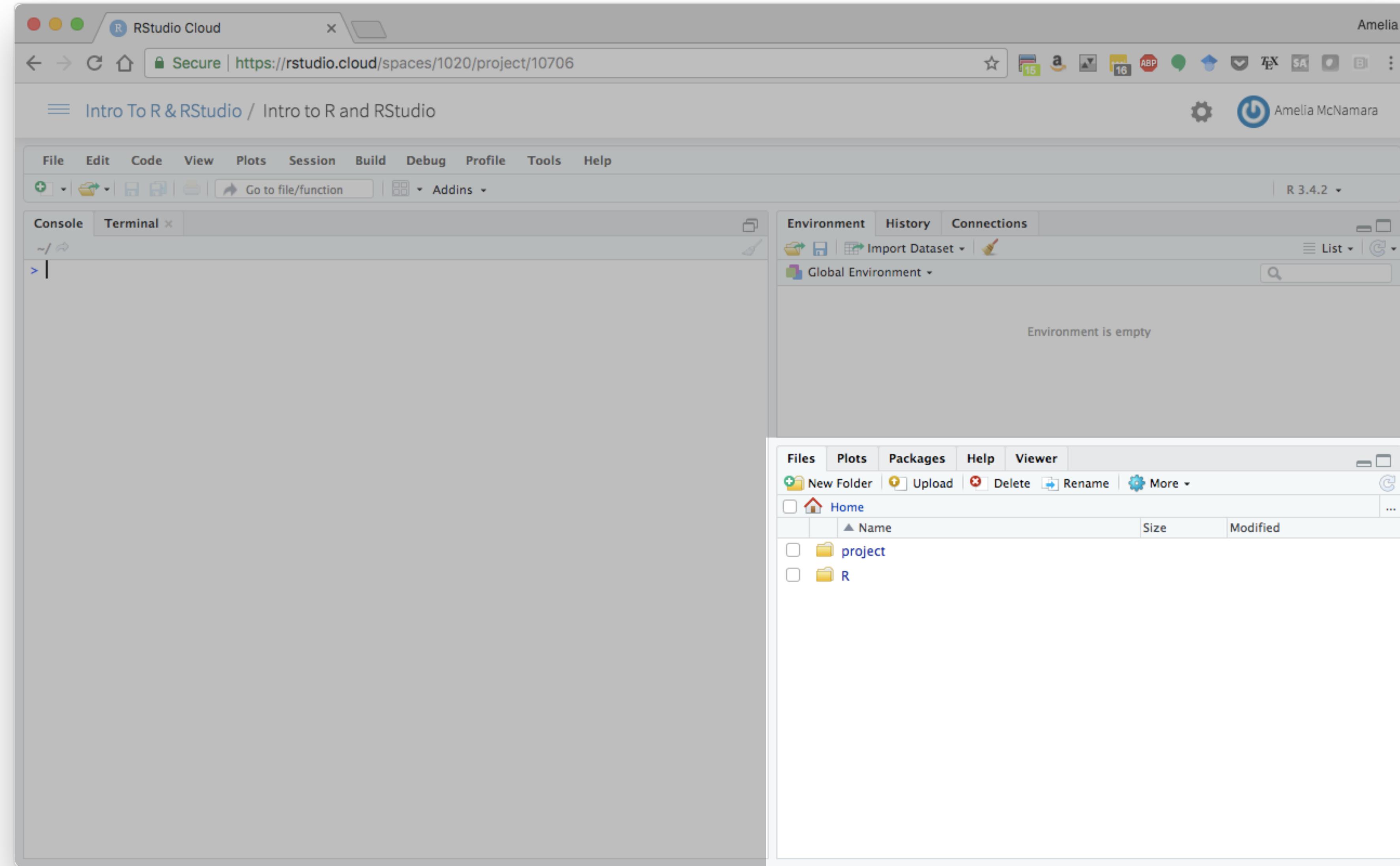


The screenshot shows the RStudio Cloud interface. The top navigation bar includes a back button, forward button, refresh button, and a secure connection indicator. The title bar says "RStudio Cloud" and "Amelia". The main menu bar has options: File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help. Below the menu is a toolbar with icons for file operations like New, Open, Save, Print, and Addins. The version "R 3.4.2" is shown. The interface is divided into several panes:

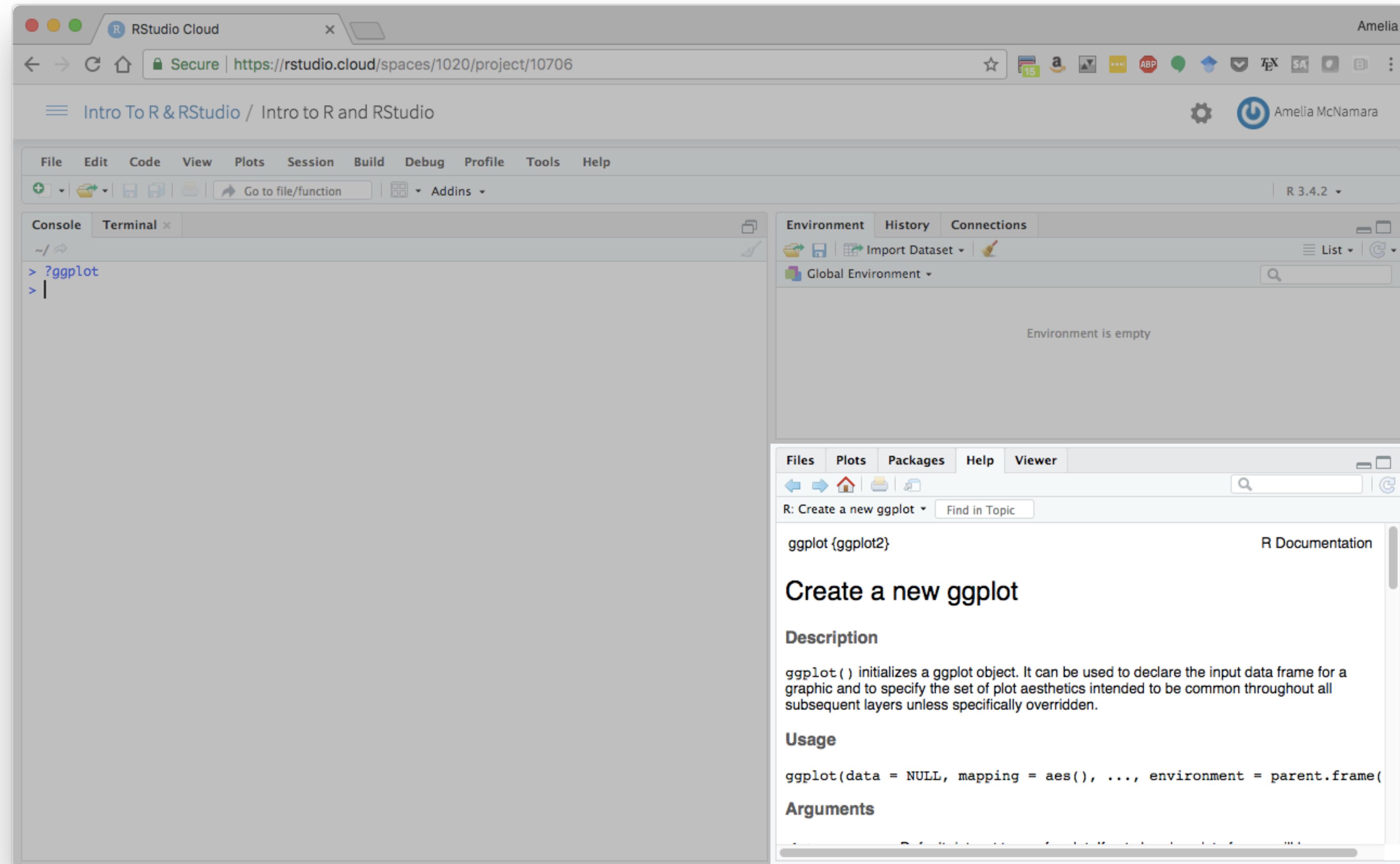
- Console:** Shows a command prompt starting with "> |".
- Environment:** Shows the Global Environment pane which is currently empty.
- Files:** Shows a file browser with a "Home" folder containing "project" and "R" subfolders. Actions like New Folder, Upload, Delete, Rename, and More are available.
- Plots:** This pane is currently empty.
- Packages:** This pane is currently empty.
- Help:** This pane is currently empty.
- Viewer:** This pane is currently empty.

The console gives you a place to execute commands written in R

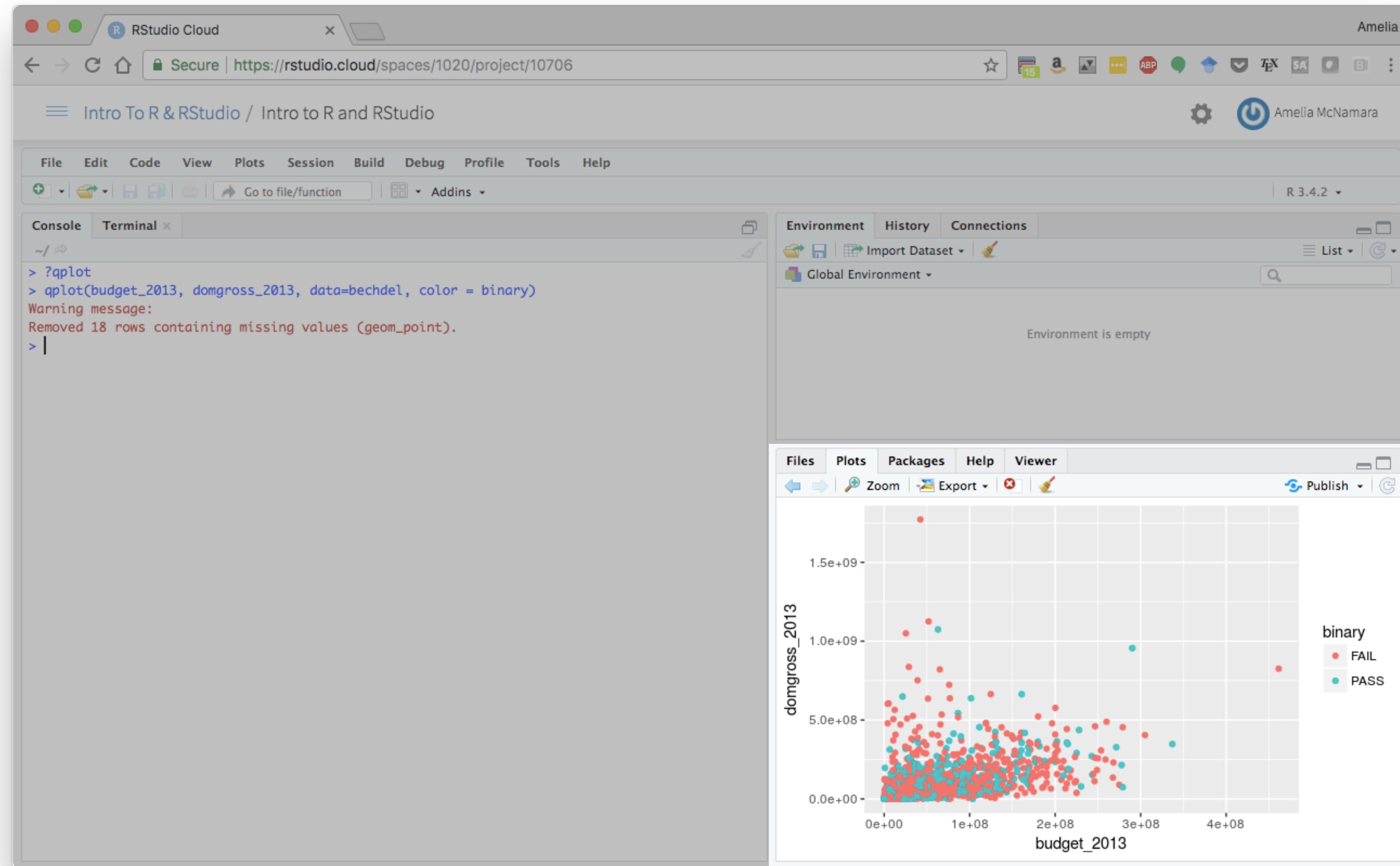
RStudio



RStudio



RStudio



RStudio

RStudio Cloud | Secure | https://rstudio.cloud/spaces/1020/project/10706

Intro To R & RStudio / Intro to R and RStudio

Console Terminal

```
> ?qplot
> qplot(budget_2013, domgross_2013, data=bechdel, color = binary)
Warning message:
Removed 18 rows containing missing values (geom_point).
> |
```

Environment History Connections

```
dechdel %>% skim(domgross_2013)
library(skimr)
data(bechdel)
bechdel %>% skim(domgross_2013)
bechdel %>% skim(clean_test)
qplot(budget_2013, domgross_2013, data=bechdel, color = binary)
lm(domgross_2013~budget_2013, data=bechdel)
?qplot
qplot(budget_2013, domgross_2013, data=bechdel, color = binary)
```

Files Plots Packages Help Viewer

domgross_2013

budget_2013

binary

- FAIL
- PASS

RStudio

RStudio Cloud | Secure | https://rstudio.cloud/spaces/1020/project/10706

Intro To R & RStudio / Intro to R and RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Untitled1 x Go to file/function Addins R 3.4.2

```
1 ---  
2 title: "Untitled"  
3 output: html_document  
4 ---  
5  
6 ```{r setup, include=FALSE}  
7 knitr::opts_chunk$set(echo = TRUE)  
8  
9  
10 ## R Markdown  
11  
12 This is an R Markdown document. Markdown is a simple formatting syntax for authoring  
HTML, PDF, and MS Word documents. For more details on using R Markdown see  
http://rmarkdown.rstudio.com.  
13  
14 When you click the **Knit** button a document will be generated that includes both  
2:1 # Untitled
```

Console Terminal

```
> ?qplot  
> qplot(budget_2013, domgross_2013, data=bechdel, color = binary)  
Warning message:  
Removed 18 rows containing missing values (geom_point).  
>
```

Environment History Connections

Import Dataset Global Environment

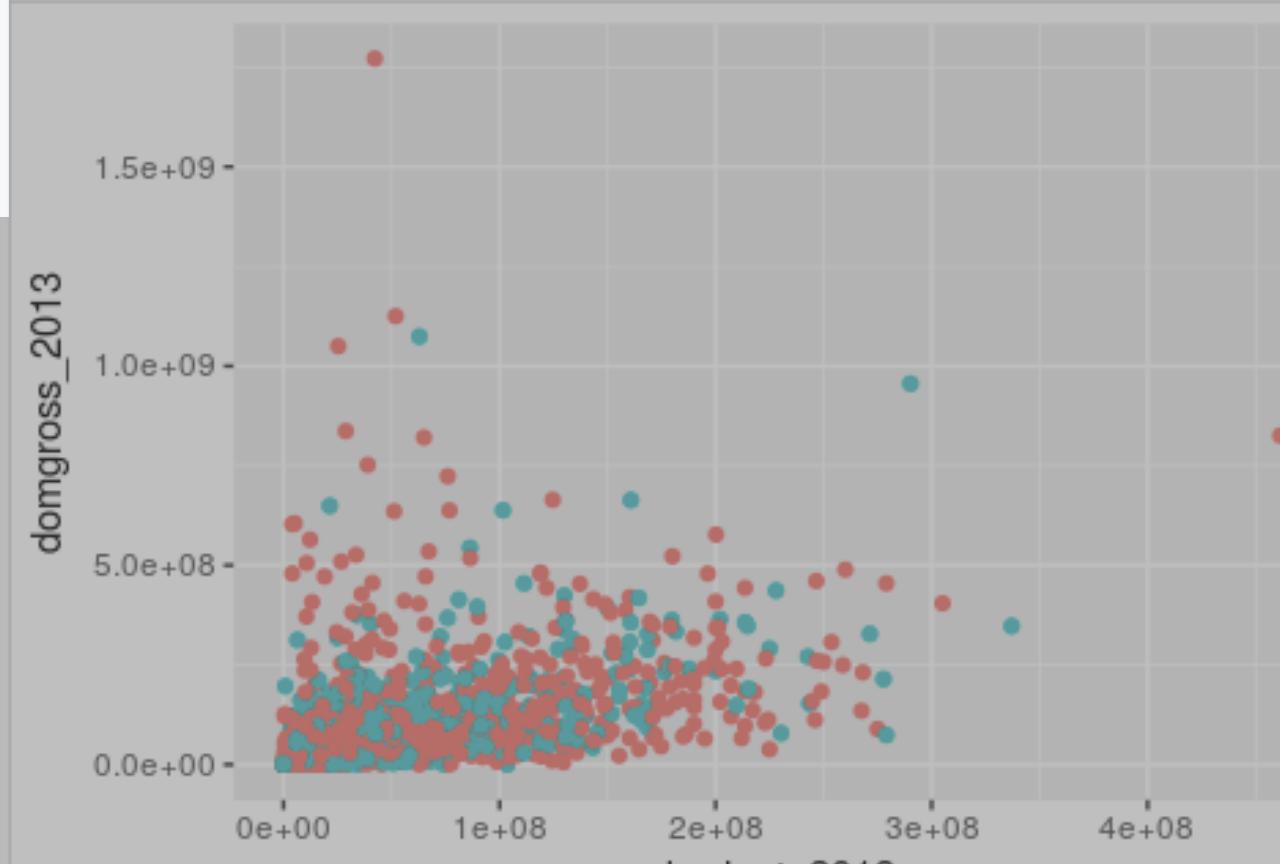
Environment is empty

Files Plots Packages Help Viewer

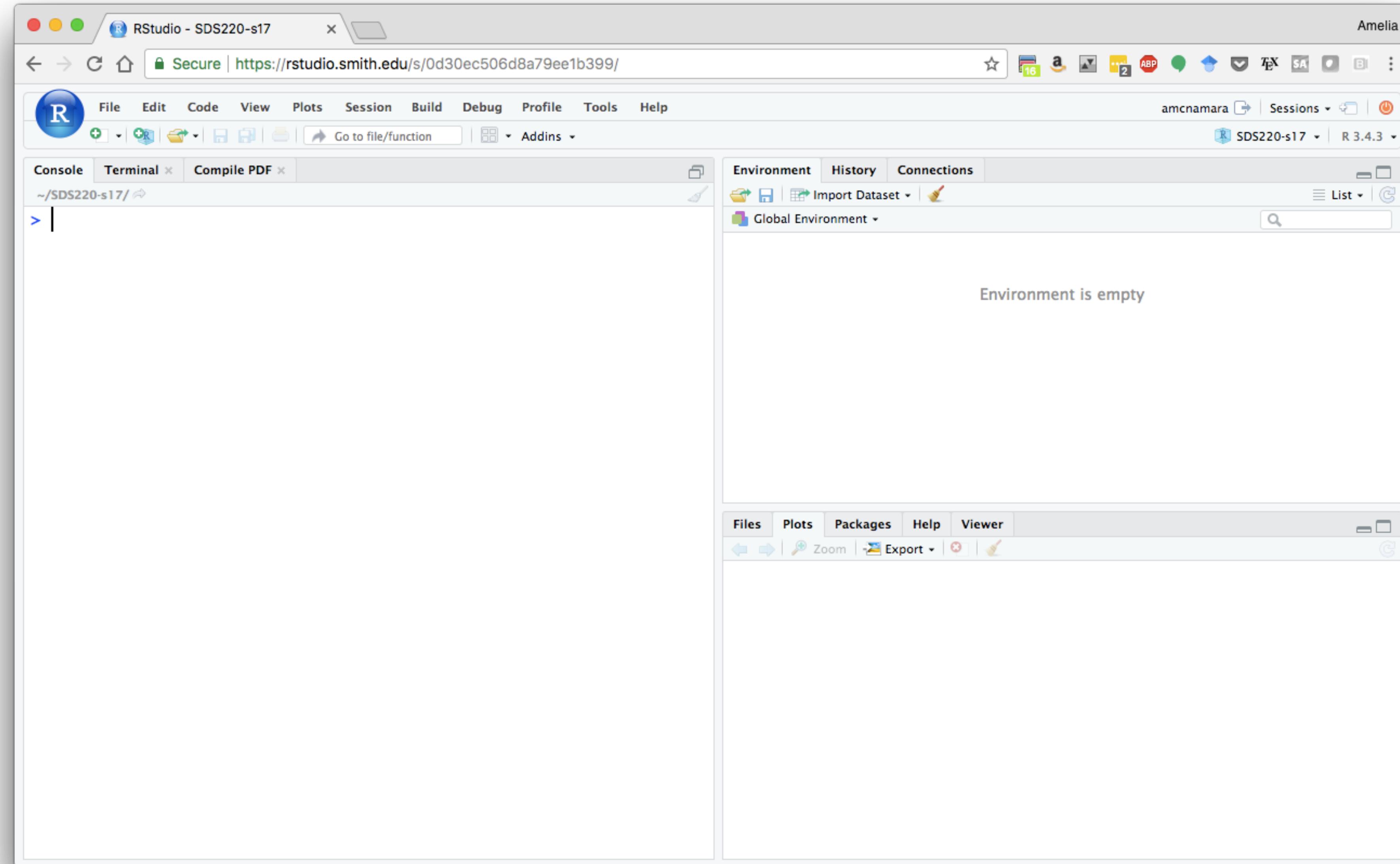
domgross_2013 budget_2013

binary

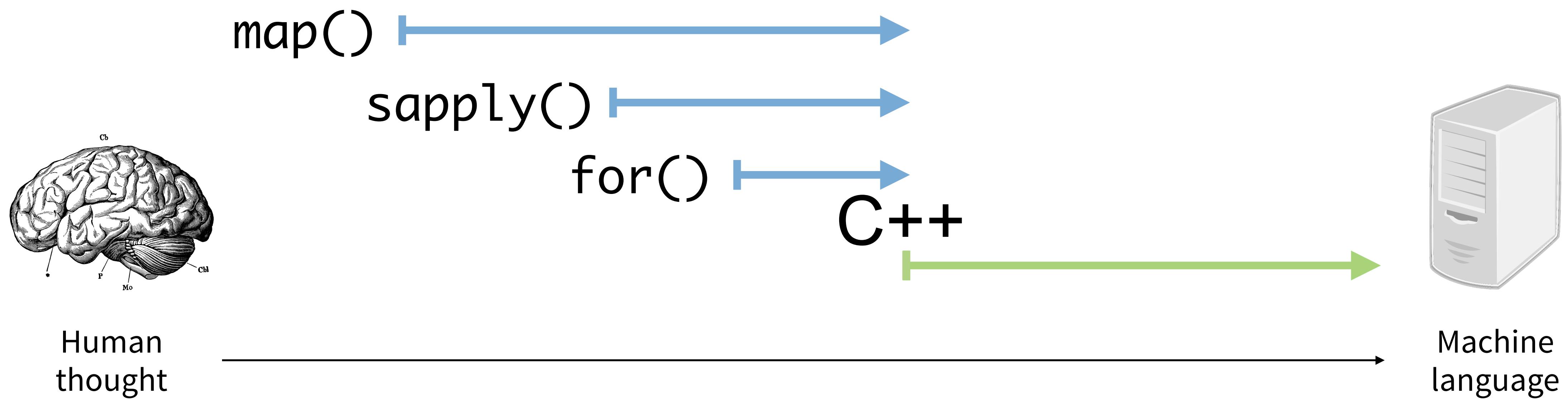
- FAIL
- PASS



RStudio



R - A computer language for scientists



R Syntax Comparison :: CHEAT SHEET

Dollar sign syntax

```
goal(data$x, data$y)
```

SUMMARY STATISTICS:

one continuous variable:
`mean(mtcars$mpg)`

one categorical variable:
`table(mtcars$cyl)`

two categorical variables:
`table(mtcars$cyl, mtcars$am)`

one continuous, one categorical:
`mean(mtcars$mpg[mtcars$cyl==4])`
`mean(mtcars$mpg[mtcars$cyl==6])`
`mean(mtcars$mpg[mtcars$cyl==8])`

PLOTTING:

one continuous variable:
`hist(mtcars$disp)`

boxplot(mtcars\$disp)

one categorical variable:
`barplot(table(mtcars$cyl))`

two continuous variables:
`plot(mtcars$disp, mtcars$mpg)`

two categorical variables:
`mosaicplot(table(mtcars$am, mtcars$cyl))`

one continuous, one categorical:
`histogram(mtcars$disp[mtcars$cyl==4])`
`histogram(mtcars$disp[mtcars$cyl==6])`
`histogram(mtcars$disp[mtcars$cyl==8])`

boxplot(mtcars\$disp[mtcars\$cyl==4])
boxplot(mtcars\$disp[mtcars\$cyl==6])
boxplot(mtcars\$disp[mtcars\$cyl==8])

WRANGLING:

subsetting:
`mtcars[mtcars$mpg>30,]`

making a new variable:
`mtcars$efficient[mtcars$mpg>30] <- TRUE`
`mtcars$efficient[mtcars$mpg<30] <- FALSE`

Formula syntax

```
goal(y~x|z, data=data, group=w)
```

SUMMARY STATISTICS:

one continuous variable:
`mosaic::mean(~mpg, data=mtcars)`

one categorical variable:
`mosaic::tally(~cyl, data=mtcars)`

two categorical variables:
`mosaic::tally(cyl~am, data=mtcars)`

one continuous, one categorical:
`mosaic::mean(mpg~cyl, data=mtcars)`

tilde

PLOTTING:

one continuous variable:
`lattice::histogram(~disp, data=mtcars)`

`lattice::bwplot(~disp, data=mtcars)`

one categorical variable:
`mosaic::bargraph(~cyl, data=mtcars)`

two continuous variables:
`lattice::xyplot(mpg~disp, data=mtcars)`

two categorical variables:
`mosaic::bargraph(~am, data=mtcars, group=cyl)`

one continuous, one categorical:
`lattice::histogram(~disp|cyl, data=mtcars)`

`lattice::bwplot(cyl~disp, data=mtcars)`

The variety of R syntaxes give
you many ways to “say” the
same thing

read across the cheatsheet to see how different
syntaxes approach the same problem

Tidyverse syntax

```
data %>% goal(x)
```

SUMMARY STATISTICS:

one continuous variable:
`mtcars %>% dplyr::summarize(mean(mpg))`

one categorical variable:
`mtcars %>% dplyr::group_by(cyl) %>%
dplyr::summarize(n())`

the pipe

two categorical variables:
`mtcars %>% dplyr::group_by(cyl, am) %>%
dplyr::summarize(n())`

one continuous, one categorical:
`mtcars %>% dplyr::group_by(cyl) %>%
dplyr::summarize(mean(mpg))`

PLOTTING:
one continuous variable:
`ggplot2::qplot(x=mpg, data=mtcars, geom = "histogram")`

`ggplot2::qplot(y=disp, x=1, data=mtcars, geom="boxplot")`

one categorical variable:
`ggplot2::qplot(x=cyl, data=mtcars, geom="bar")`

two continuous variables:
`ggplot2::qplot(x=disp, y=mpg, data=mtcars, geom="point")`

two categorical variables:
`ggplot2::qplot(x=factor(cyl), data=mtcars, geom="bar") +
facet_grid(.~am)`

one continuous, one categorical:
`ggplot2::qplot(x=disp, data=mtcars, geom = "histogram") +
facet_grid(.~cyl)`

`ggplot2::qplot(y=disp, x=factor(cyl), data=mtcars,
geom="boxplot")`

WRANGLING:
subsetting:
`mtcars %>% dplyr::filter(mpg>30)`

making a new variable:
`mtcars <- mtcars %>%
dplyr::mutate(efficient = if_else(mpg>30, TRUE, FALSE))`





Introduction to Data Science in the Tidyverse

Amelia McNamara and Hadley Wickham



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

from  RStudio