Data Science in the tidyverse

Amelia McNamara

amelia.mn

@AmeliaMN

@AmeliaMN

@hadleywickham

@kierisi

@i_steves

@baumerben

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

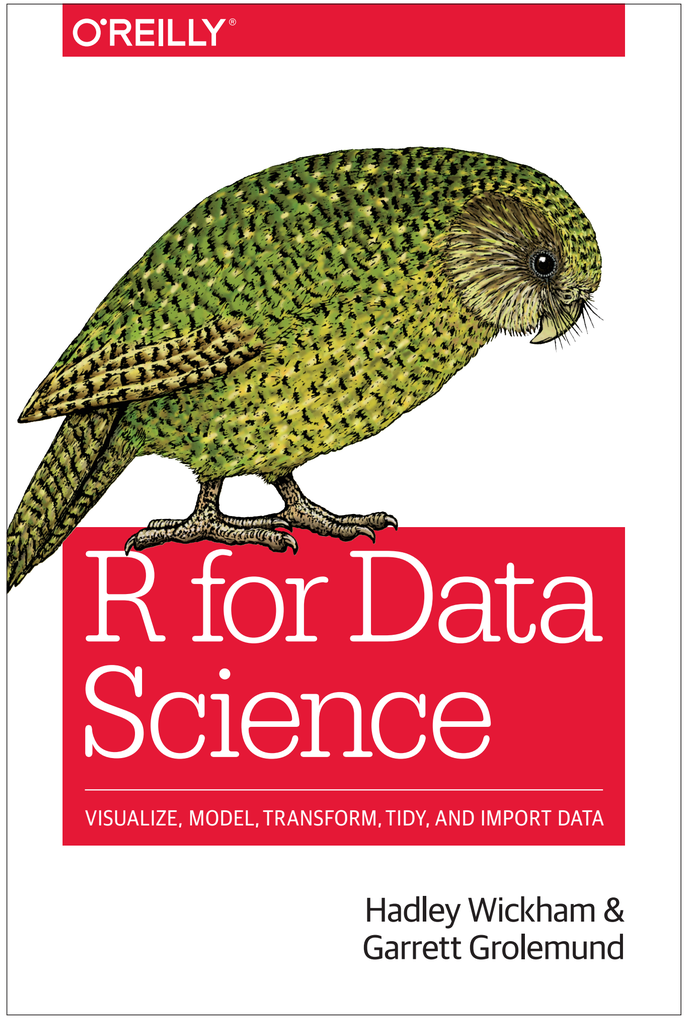
R for Data Science

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

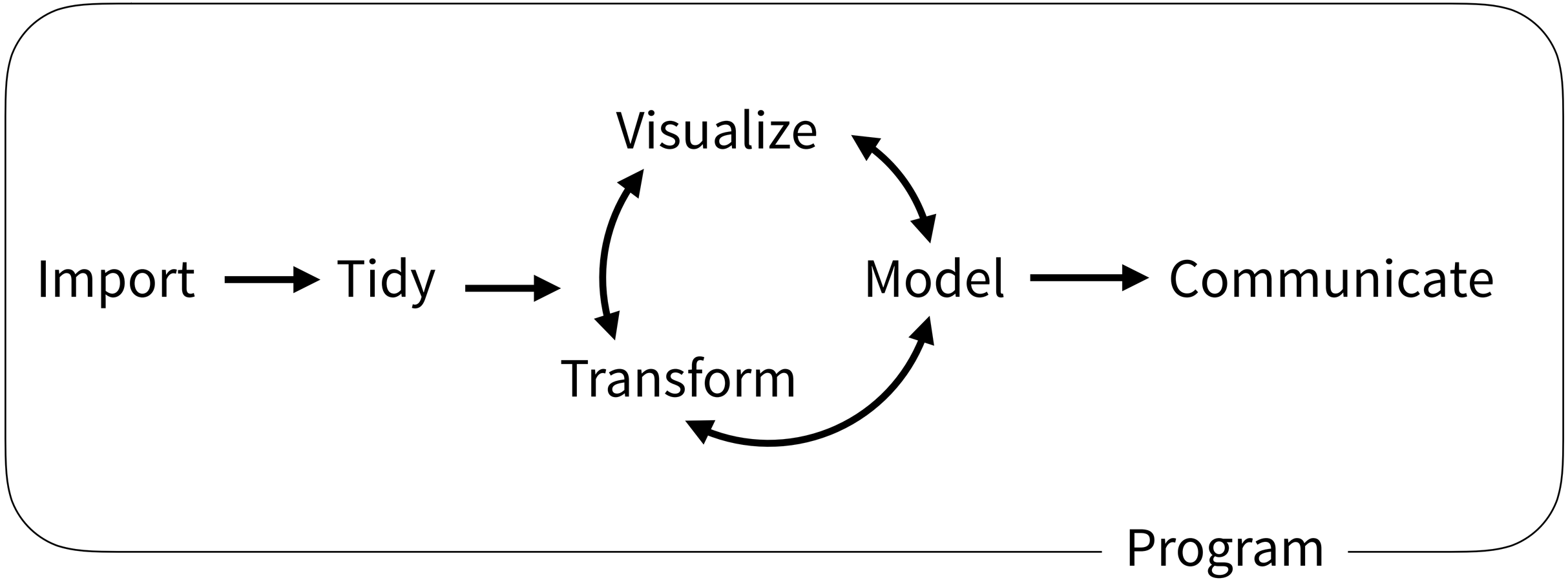
Links to the relevant sections of the book

In R4DS

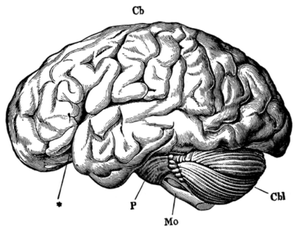
Introduction



(Applied) **Data Science**



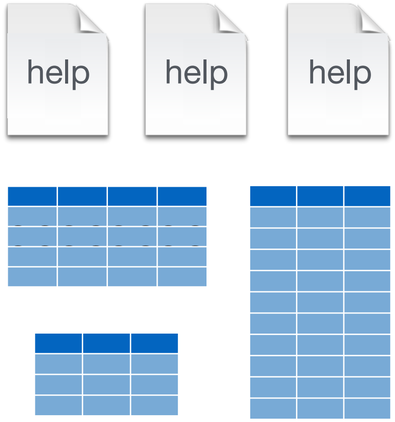
R - A computer language for scientists



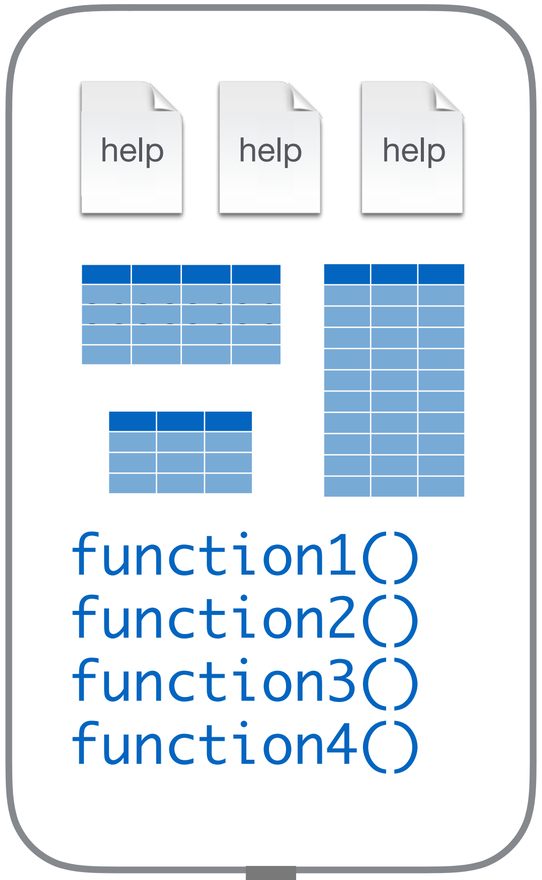
Human thought

C++

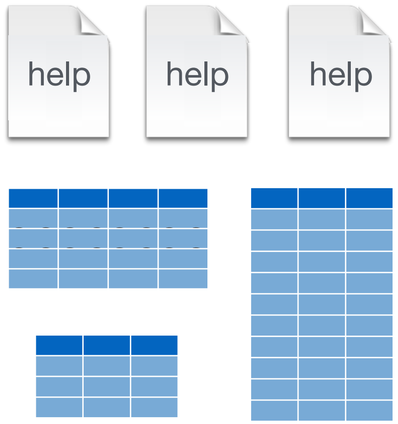
Machine language



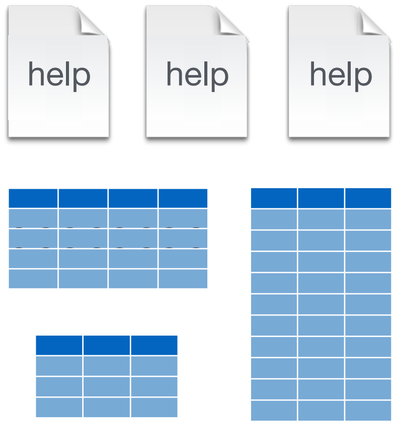

```
function1()
function2()
function3()
function4()
```



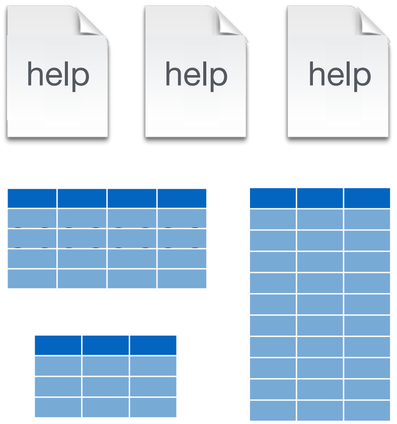

```
function1()
function2()
function3()
function4()
```



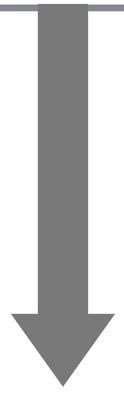

```
function5()
function6()
function7()
function8()
```



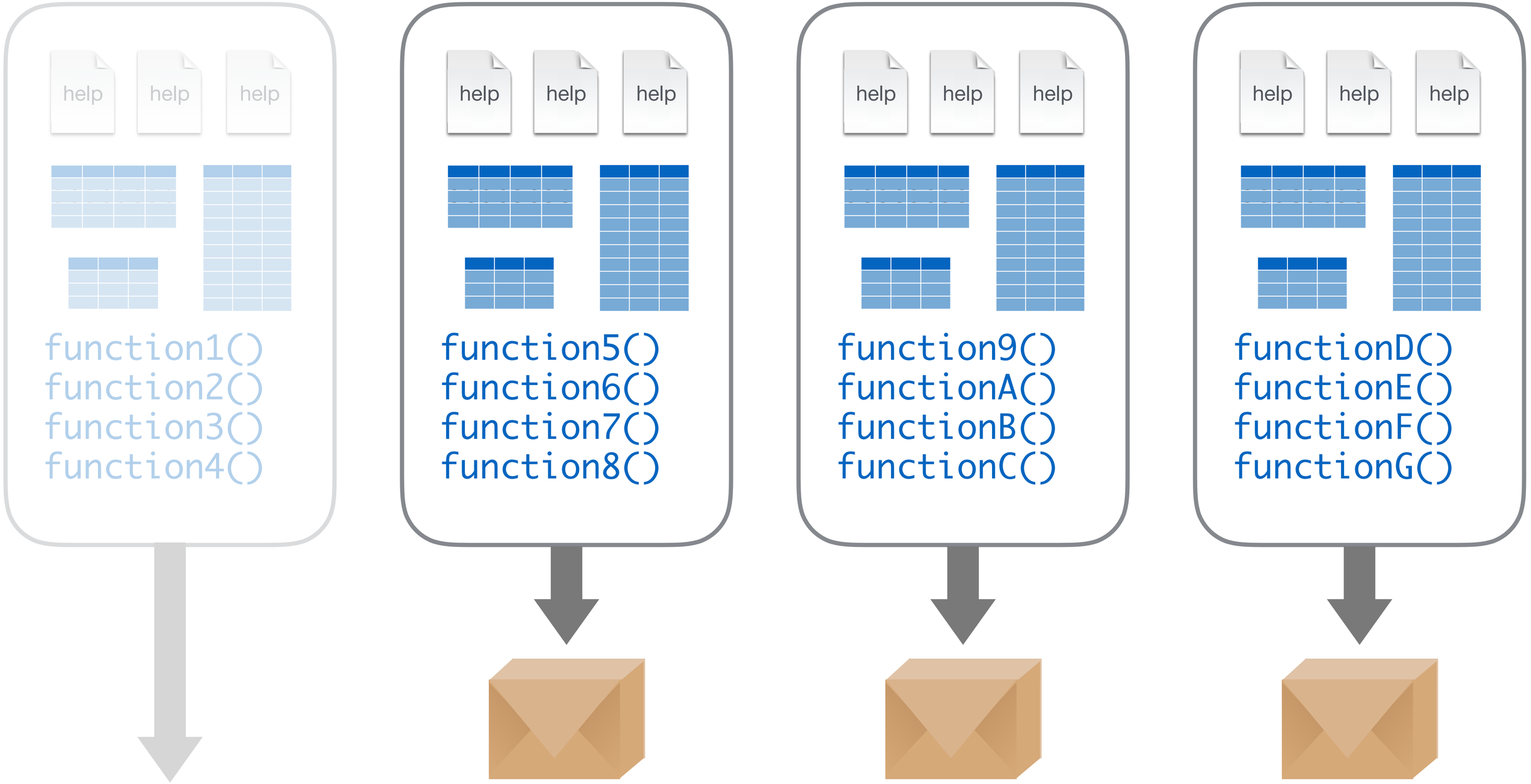

```
function9()
functionA()
functionB()
functionC()
```




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

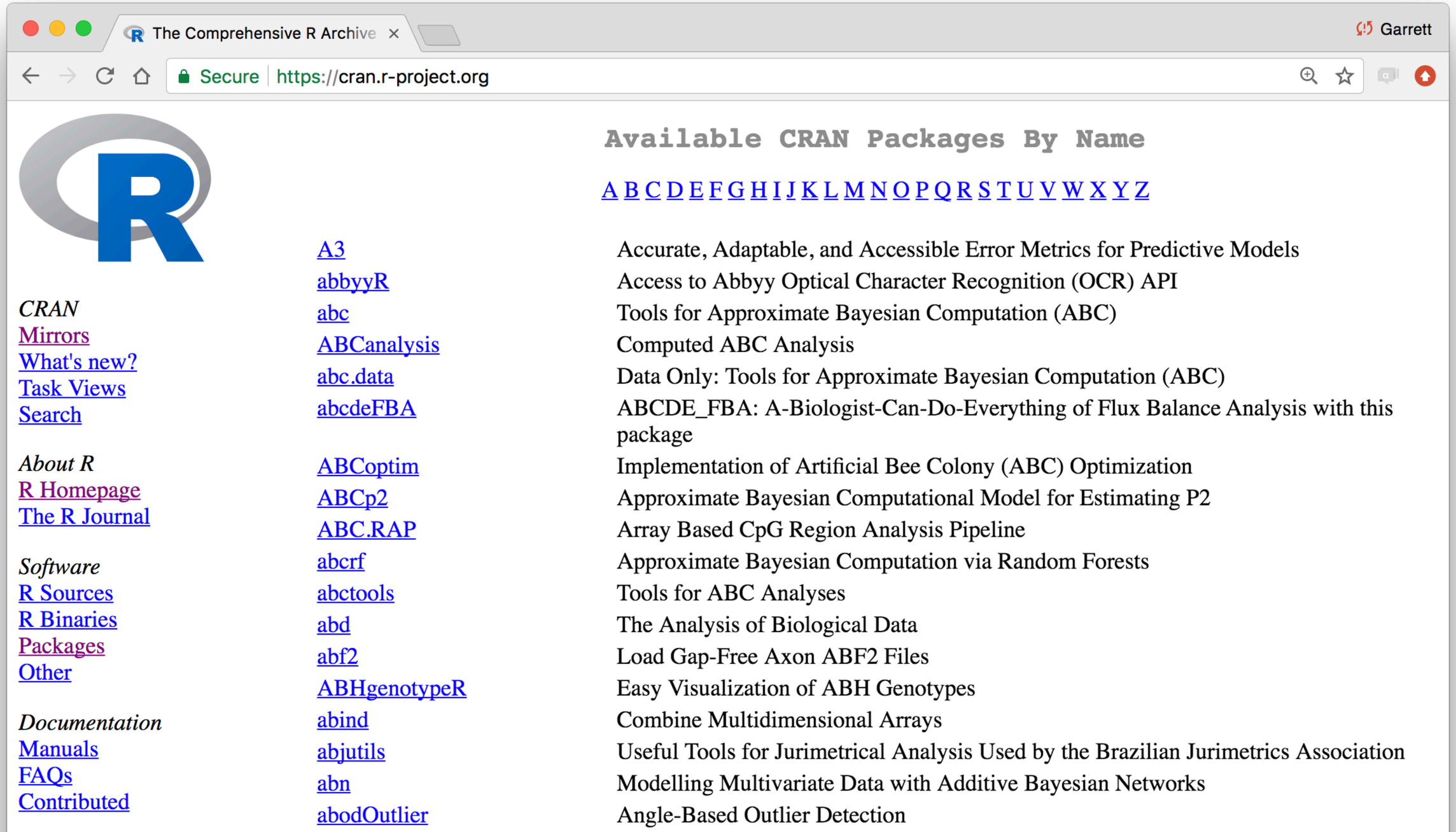


Base R



Base R

R Packages



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

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

A3	Accurate, Adaptable, and Accessible Error Metrics for Predictive Models
abbyyR	Access to Abbyy Optical Character Recognition (OCR) API
abc	Tools for Approximate Bayesian Computation (ABC)
ABCanalysis	Computed ABC Analysis
abc.data	Data Only: Tools for Approximate Bayesian Computation (ABC)
abcdeFBA	ABCDE_FBA: A-Biologist-Can-Do-Everything of Flux Balance Analysis with this package
ABCoptim	Implementation of Artificial Bee Colony (ABC) Optimization
ABCp2	Approximate Bayesian Computational Model for Estimating P2
ABC.RAP	Array Based CpG Region Analysis Pipeline
abcrf	Approximate Bayesian Computation via Random Forests
abctools	Tools for ABC Analyses
abd	The Analysis of Biological Data
abf2	Load Gap-Free Axon ABF2 Files
ABHgenotypeR	Easy Visualization of ABH Genotypes
abind	Combine Multidimensional Arrays
abjutils	Useful Tools for Jurimetrical Analysis Used by the Brazilian Jurimetrics Association
abn	Modelling Multivariate Data with Additive Bayesian Networks
abodOutlier	Angle-Based Outlier Detection

Using packages

1

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

Downloads files to "computer"

1 x per "computer"

2

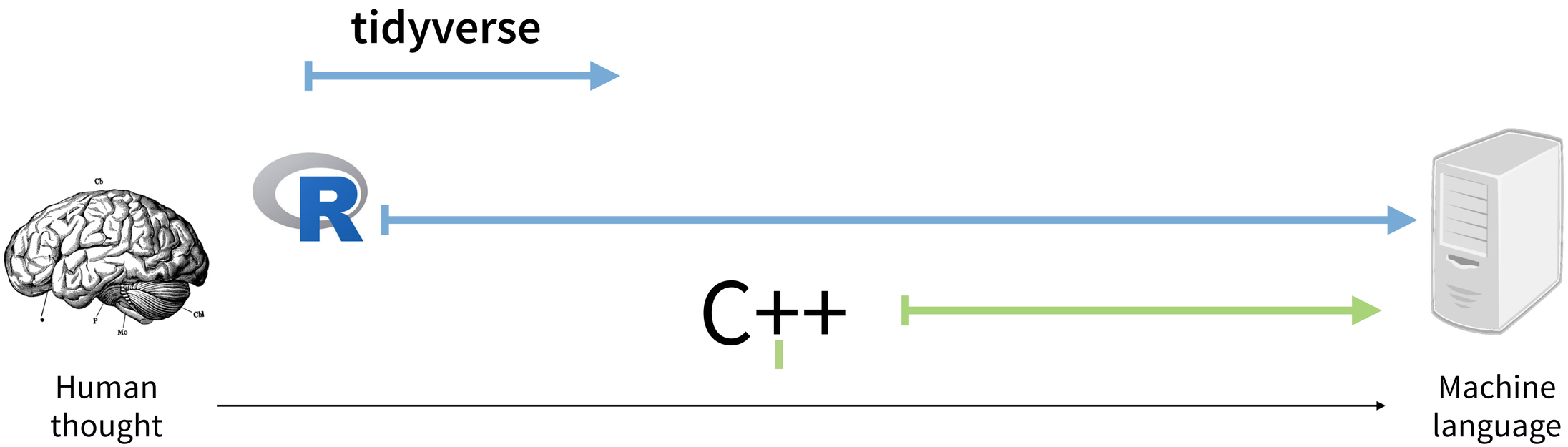
```
library("foo")
```

Loads package

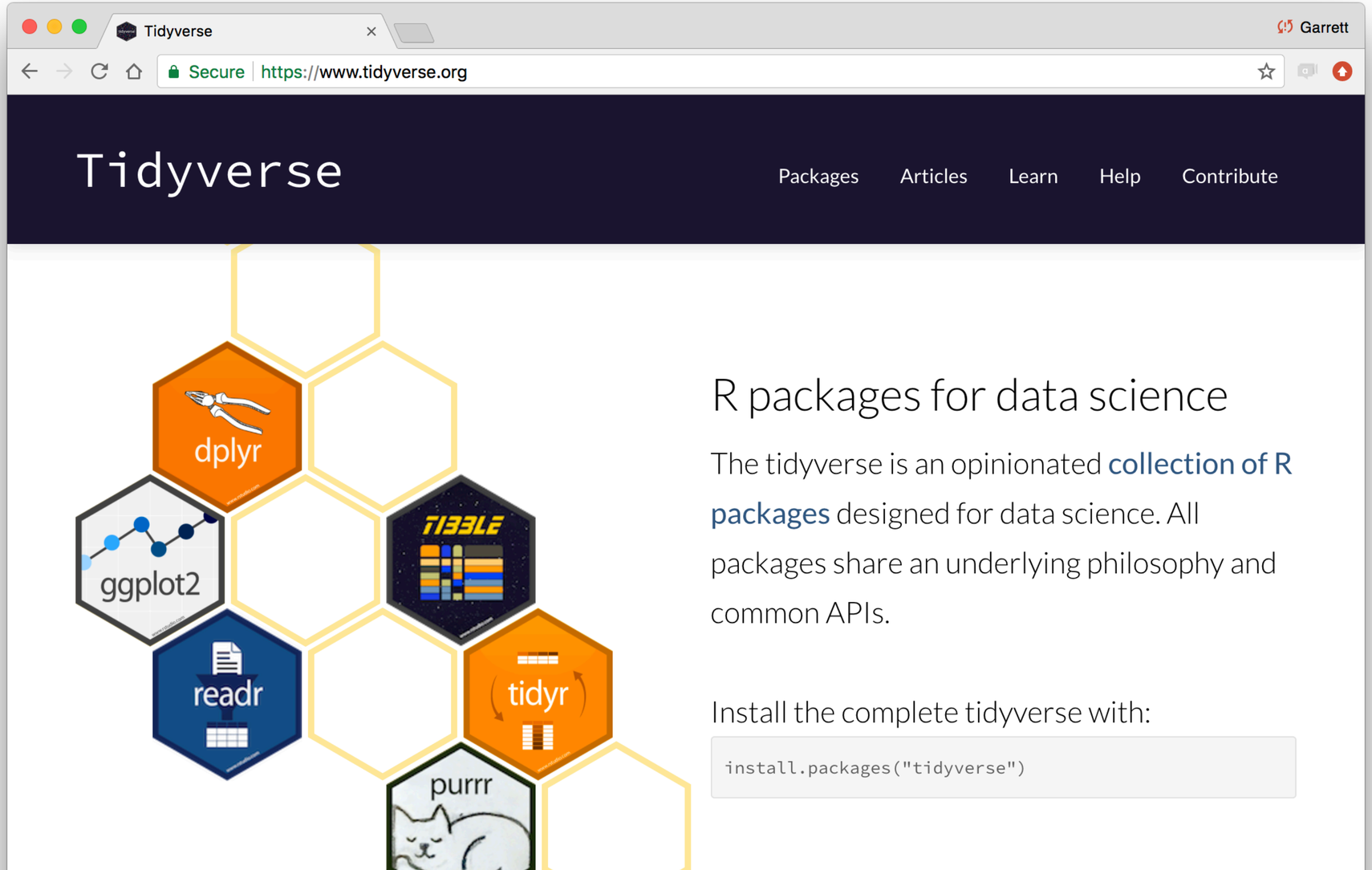
1 x per R Session

I've done this for you for this workshop

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

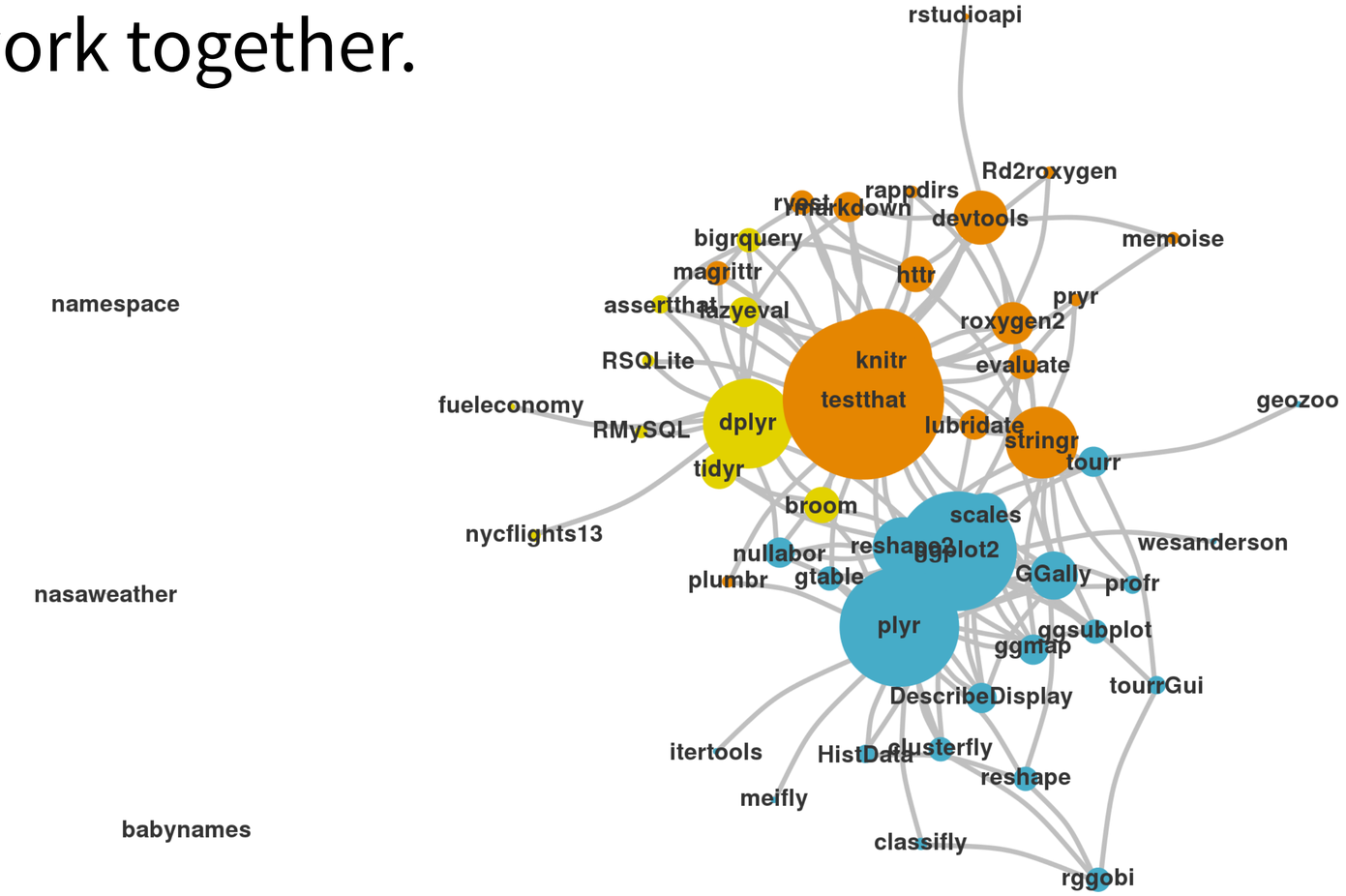


tidyverse.org



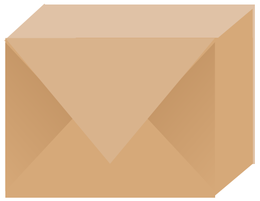
The Tidyverse

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



Display by Adolfo Álvarez

tidyverse



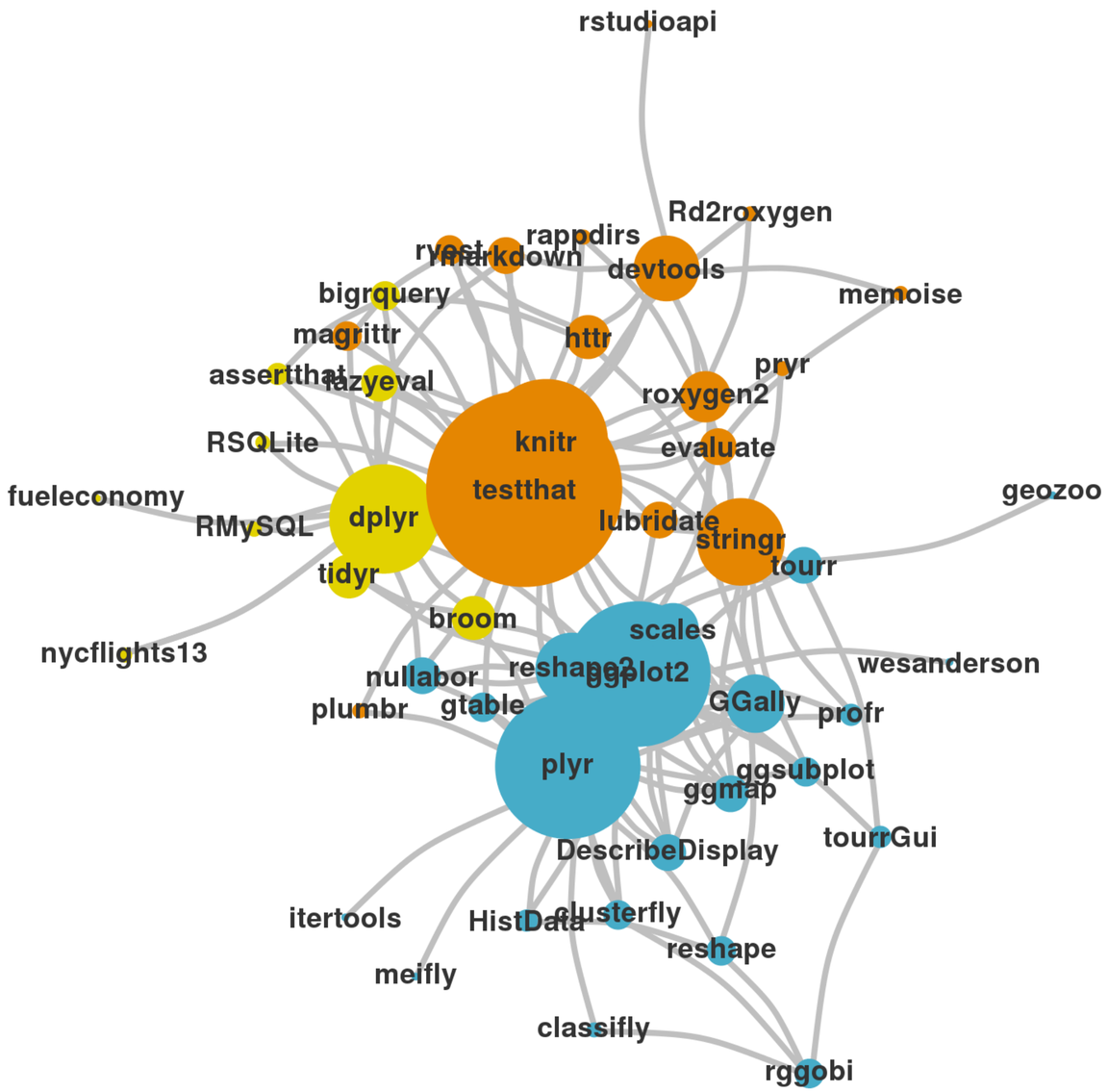
An R package that serves as a short cut for installing and loading the components of the tidyverse.

```
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	11:00 - 12:30
Lunch	12:30 - 2:00
Modeling	2:00 - 3:30
Afternoon Break	3:30 - 4:00
Organize and wrap-up	4:00 - 5:00


R

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

R Notebooks (Let's start!)

Data Science in the Tidyverse

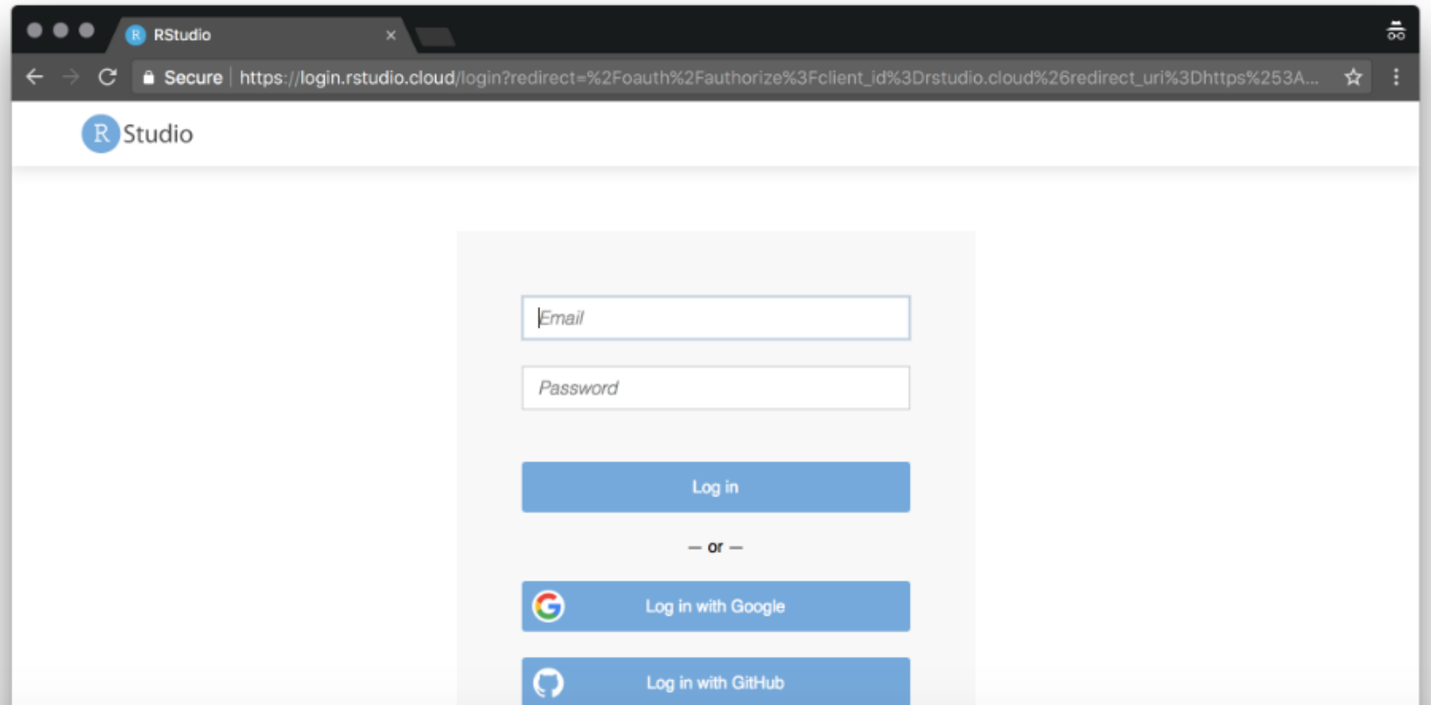
Go to http://bit.ly/rstudio2019setup

and follow instructions

To get started:

To get set up follow these steps:

1. Visit the project at https://rstudio.cloud/project/163983
2. Log in using google, github, shinyapps.io or "Sign Up".



Your Turn

Instructions with screenshots at bit.ly/rstudio2019setup

First, if you haven't already

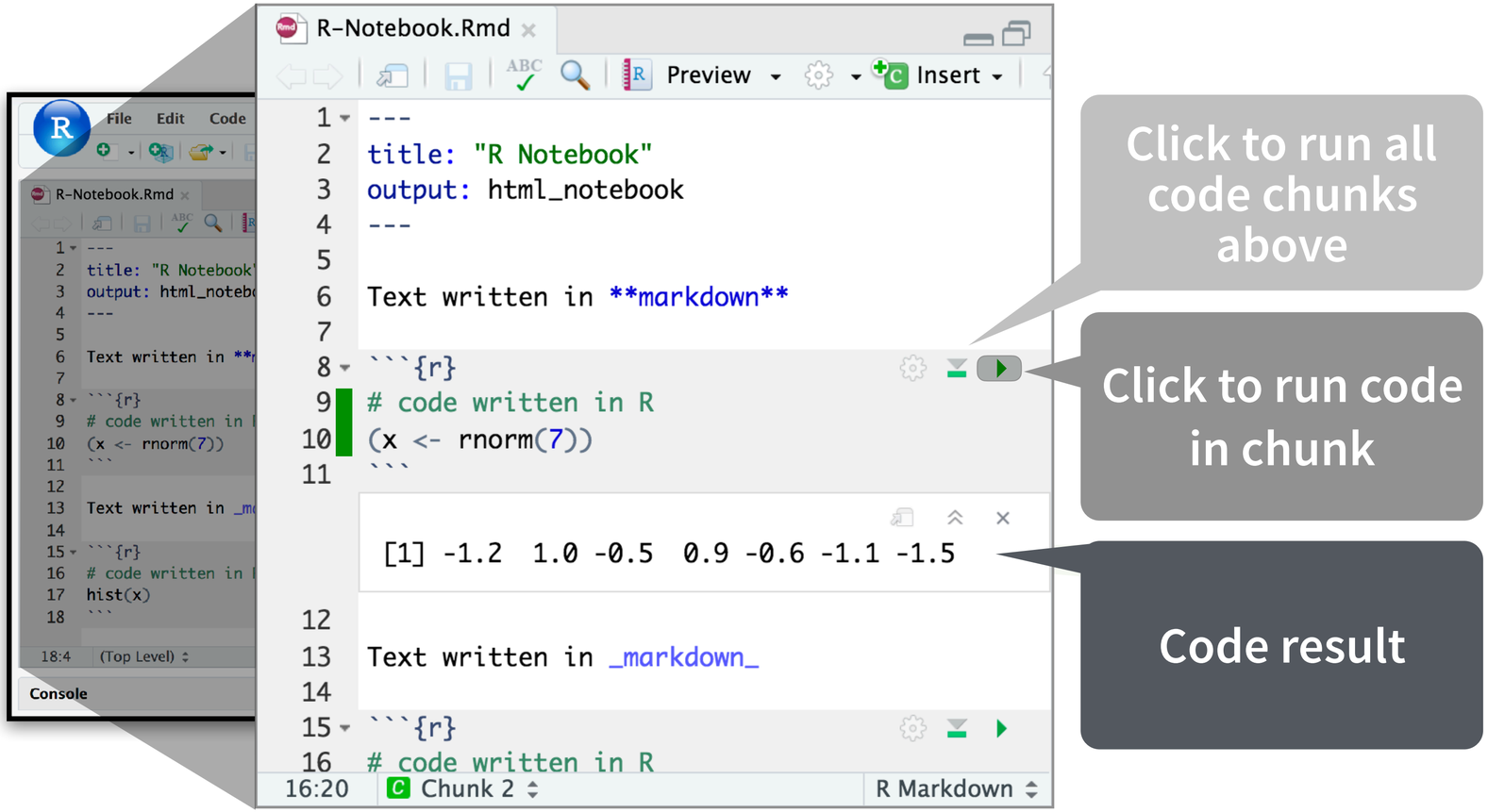
- visit https://rstudio.cloud/project/163983
- Log In / Sign Up
- "Save a copy" of the project
- Open data-science-in-the-tidyverse.Rproj

When you have your copy of the project, let us know by putting up the **Blue** post-it.

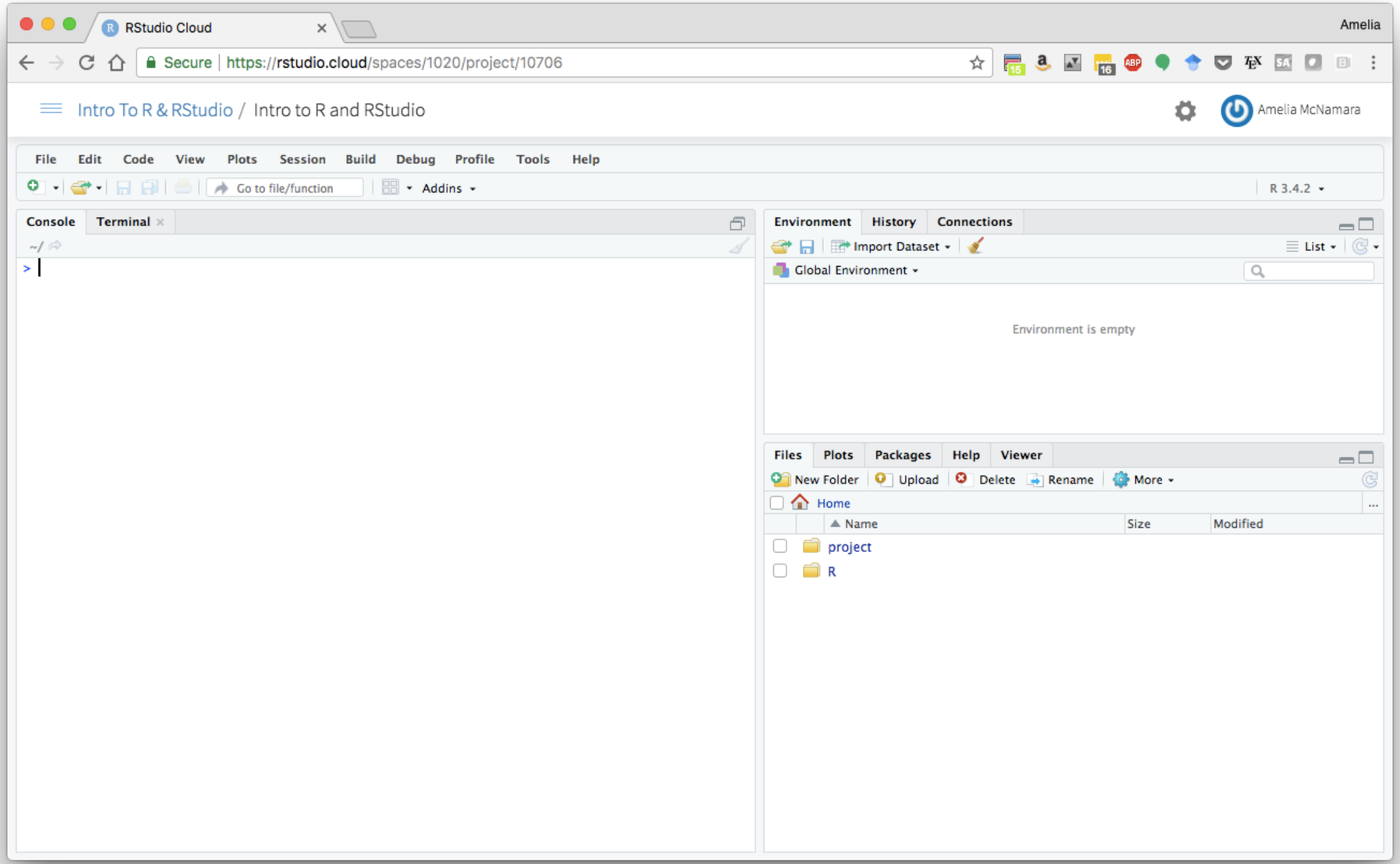
Then, open **00-Getting-started.Rmd** and look around

R Notebooks

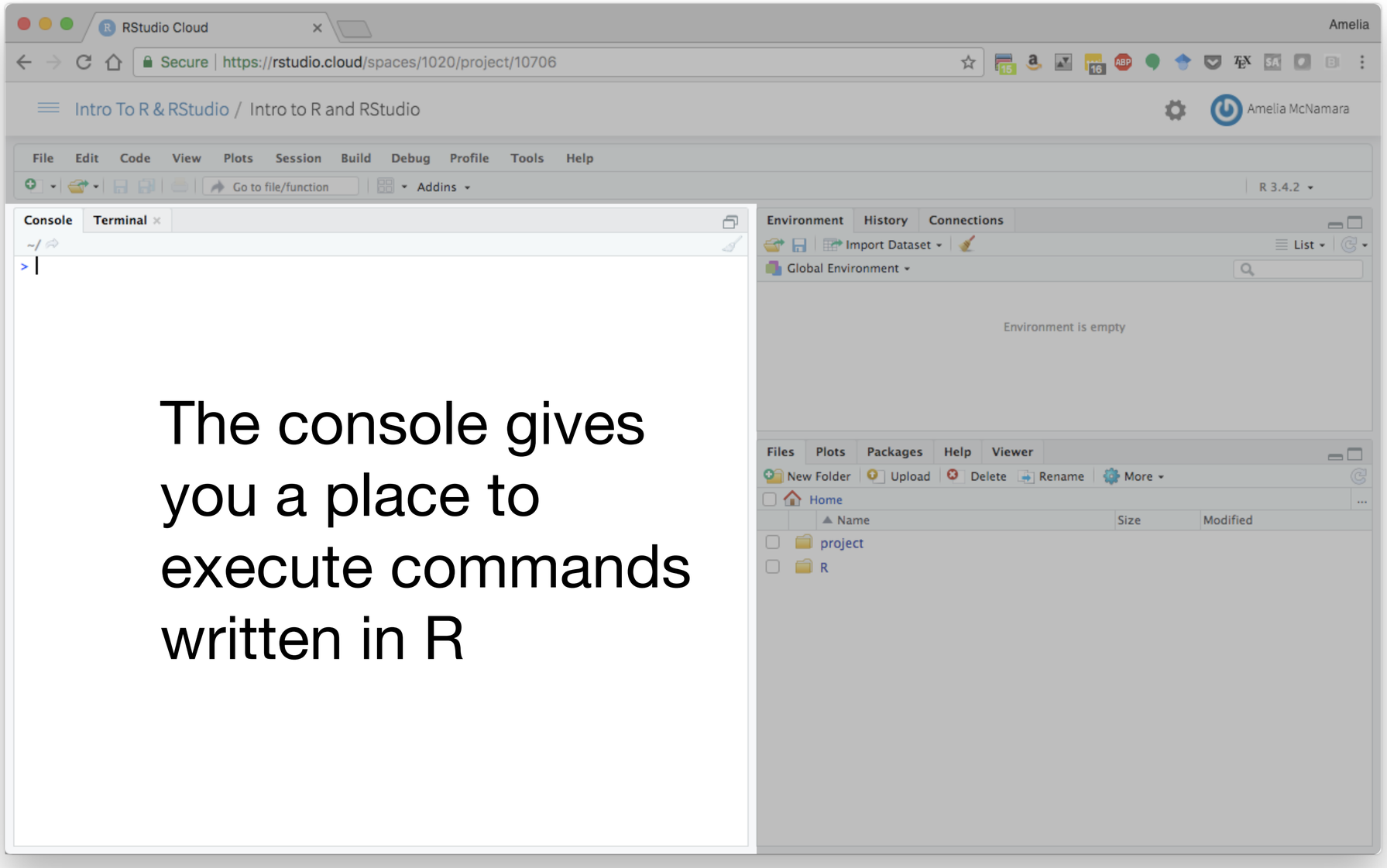
An authoring format for Data Science.



RStudio

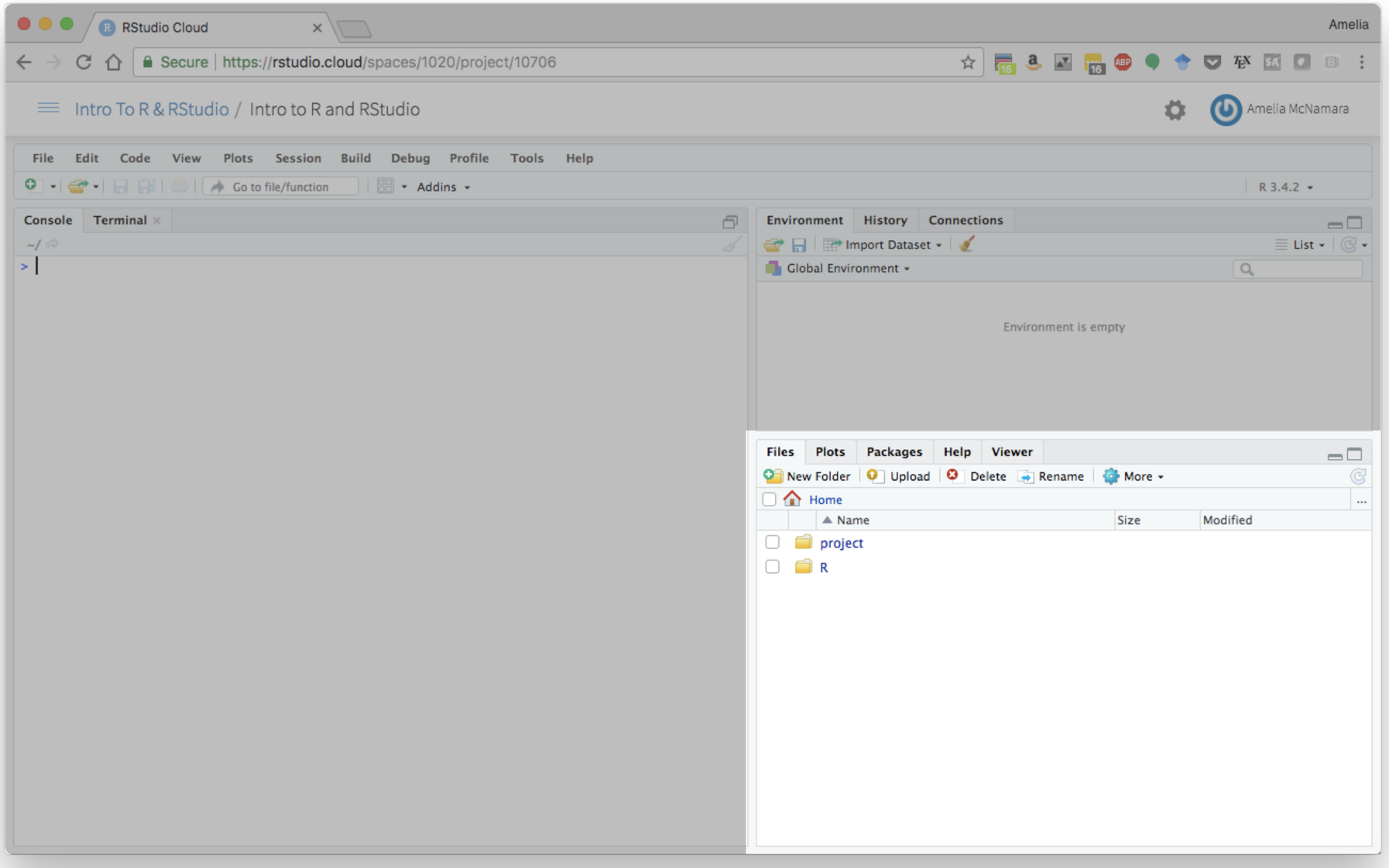

RStudio Cloud
Amelia
Secure | https://rstudio.cloud/spaces/1020/project/10706
Intro To R & RStudio / Intro to R and RStudio
Amelia McNamara
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 3.4.2
Console
Terminal
~/
>
Environment
History
Connections
Import Dataset
List
Global Environment
Environment is empty
Files
Plots
Packages
Help
Viewer
New Folder
Upload
Delete
Rename
More
Home
Name
Size
Modified
project
R

RStudio

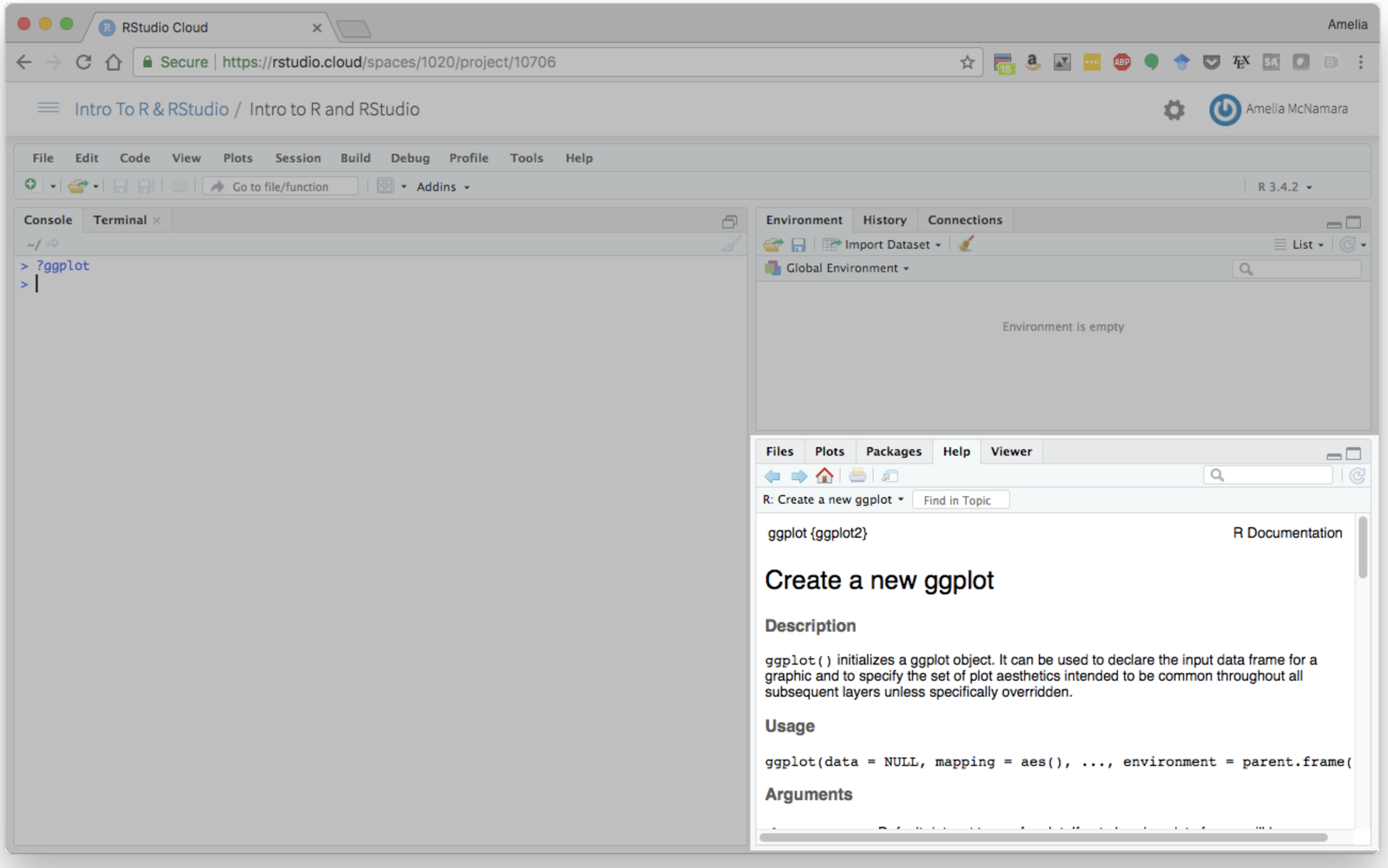


The console gives you a place to execute commands written in R

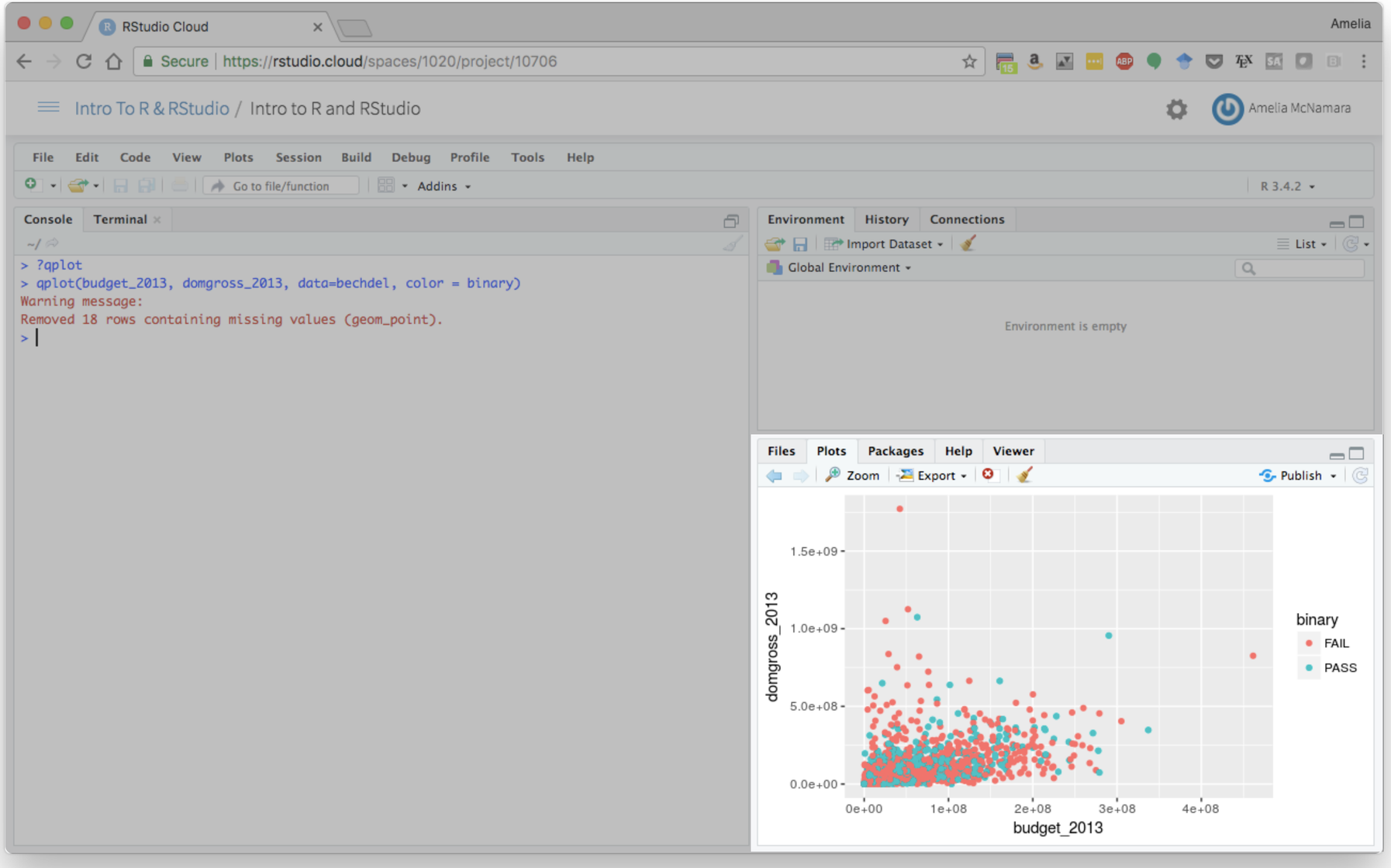
RStudio


RStudio Cloud
Amelia
Secure | https://rstudio.cloud/spaces/1020/project/10706
Intro To R & RStudio / Intro to R and RStudio
Amelia McNamara
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 3.4.2
Console
Terminal
~/
>
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

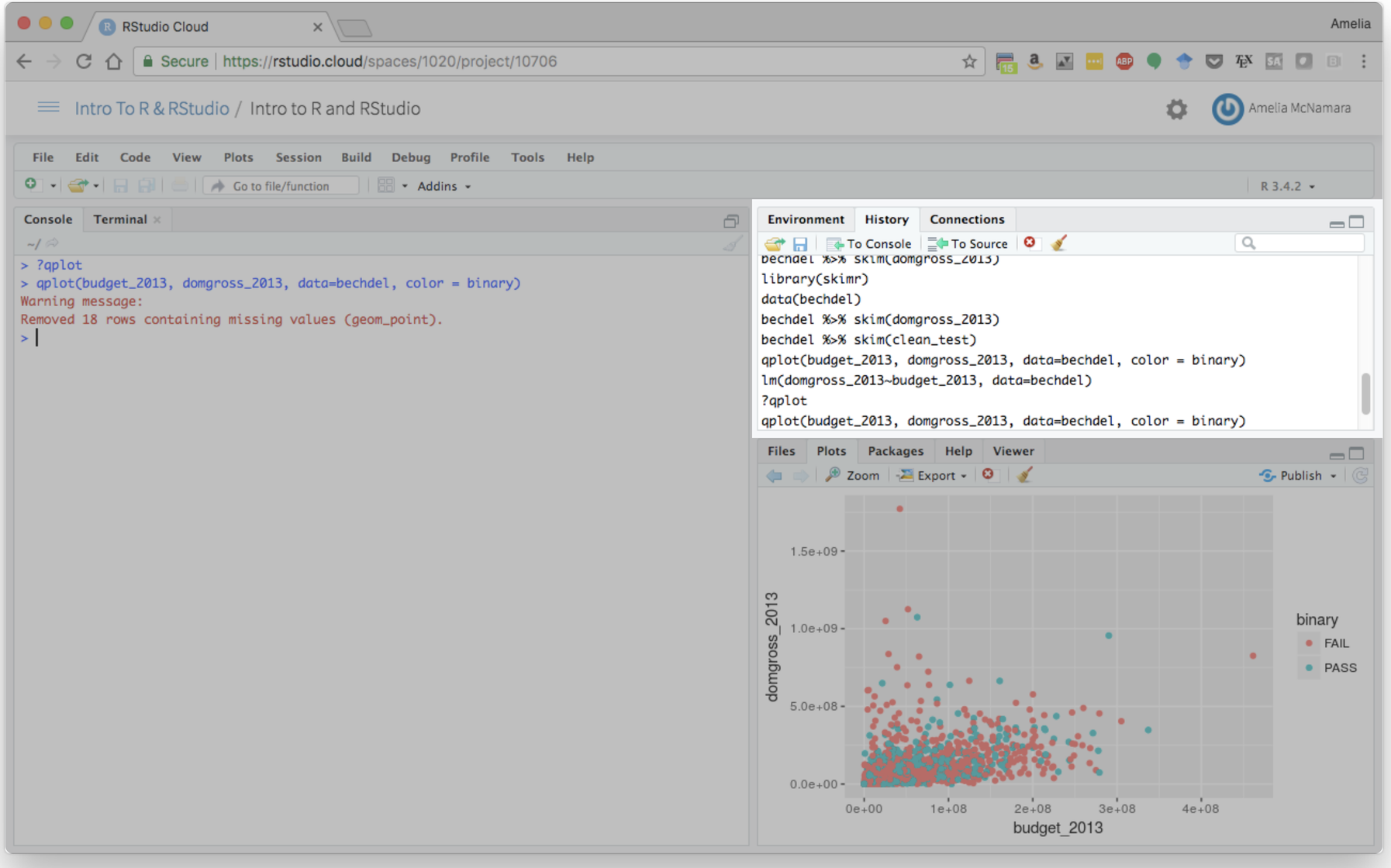
RStudio


RStudio Cloud
Amelia
Secure | https://rstudio.cloud/spaces/1020/project/10706
Intro To R & RStudio / Intro to R and RStudio
Amelia McNamara
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 3.4.2
Console Terminal
~/
> ?ggplot
>
Environment History Connections
Import Dataset
List
Global Environment
Environment is empty
Files Plots Packages Help Viewer
R: Create a new ggplot
Find in Topic
ggplot {ggplot2}
R Documentation
Create a new ggplot
Description
ggplot() initializes a ggplot object. It can be used to declare the input data frame for a graphic and to specify the set of plot aesthetics intended to be common throughout all subsequent layers unless specifically overridden.
Usage
ggplot(data = NULL, mapping = aes(), ..., environment = parent.frame(
Arguments

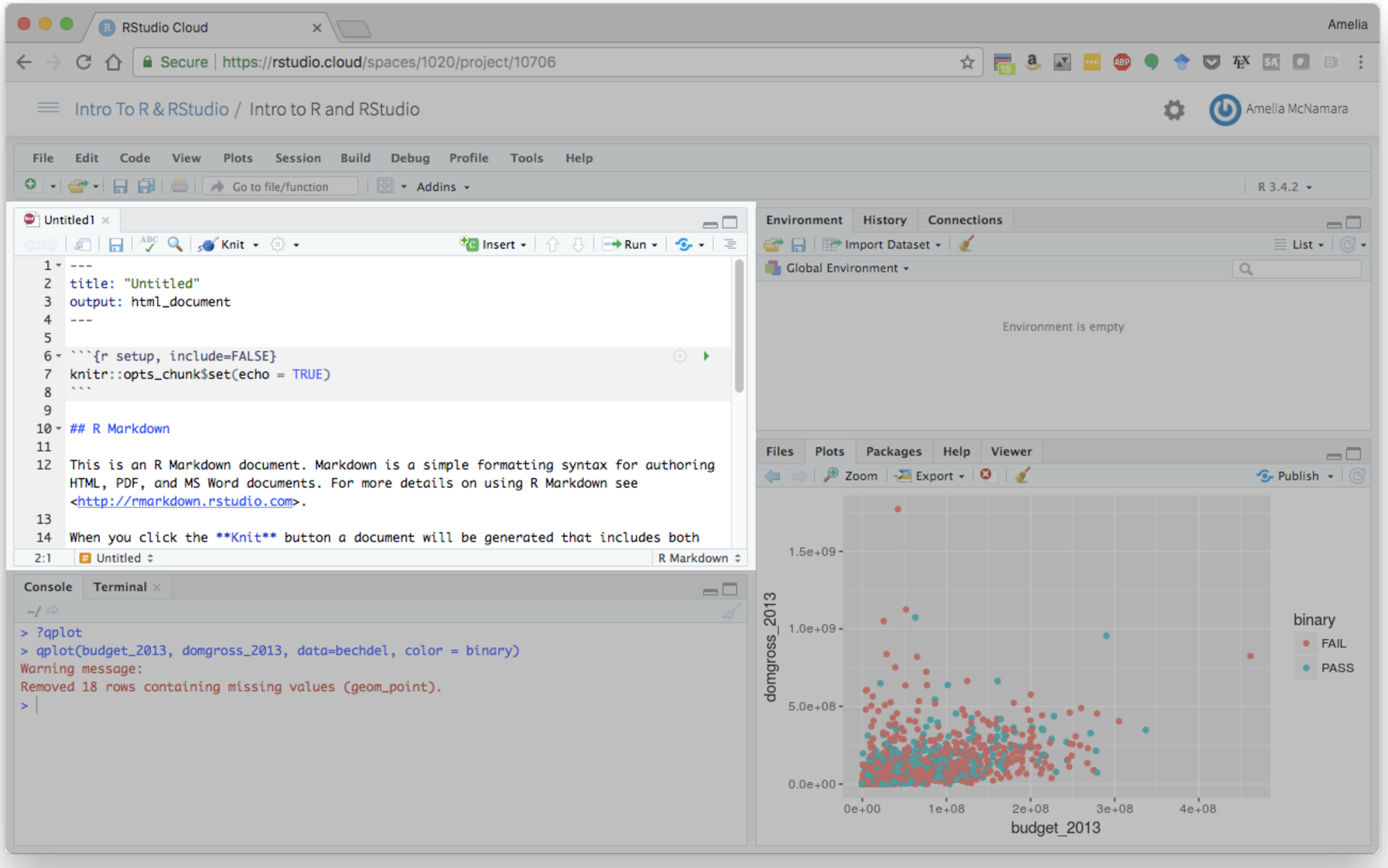
RStudio


RStudio Cloud
Amelia
Secure | https://rstudio.cloud/spaces/1020/project/10706
Intro To R & RStudio / Intro to R and RStudio
Amelia McNamara
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 3.4.2
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
List
Global Environment
Environment is empty
Files Plots Packages Help Viewer
Zoom
Export
Publish
1.5e+09
1.0e+09
5.0e+08
0.0e+00
domgross_2013
0e+00
1e+08
2e+08
3e+08
4e+08
budget_2013
binary
FAIL
PASS

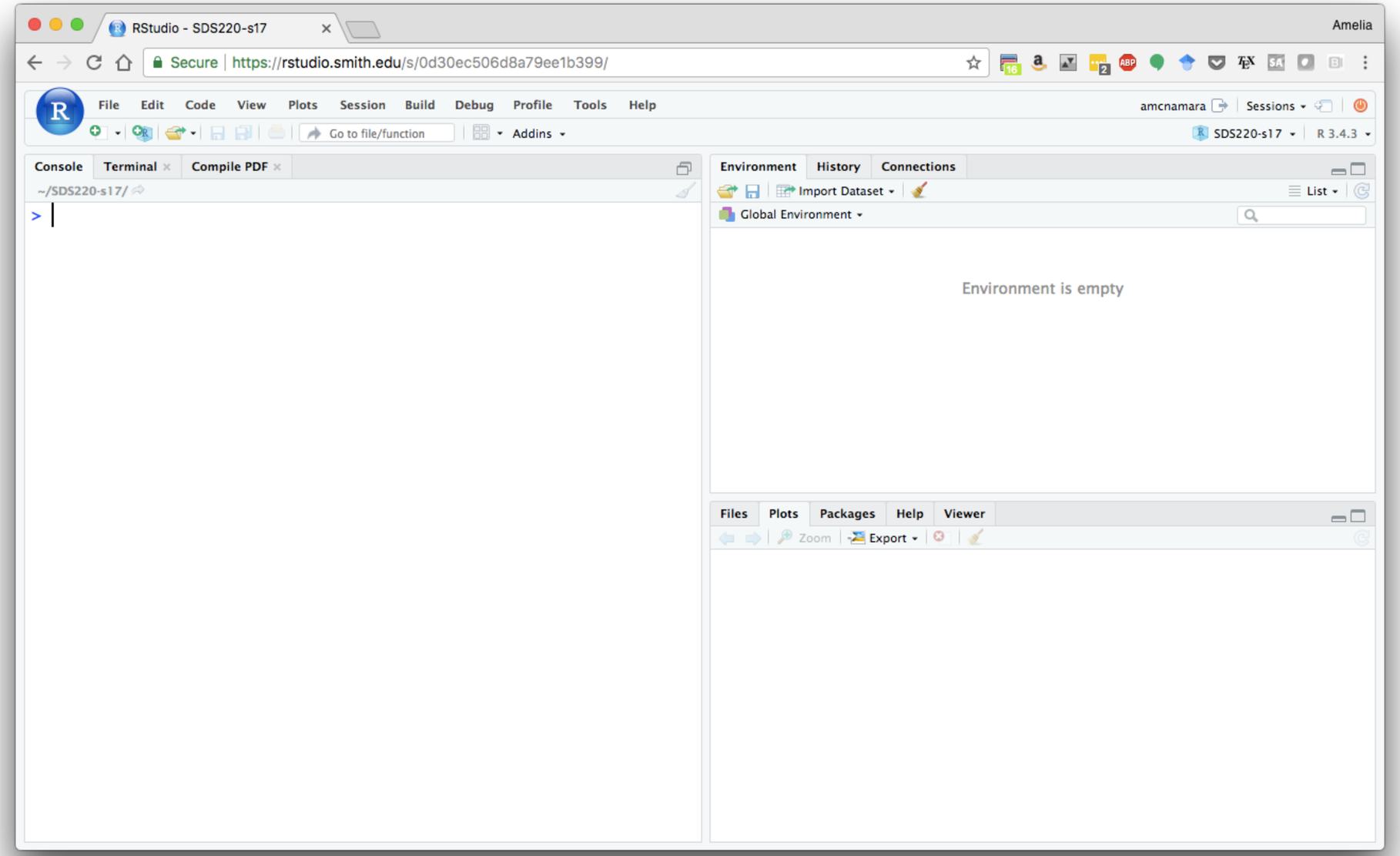
RStudio


RStudio Cloud
Amelia
Secure | https://rstudio.cloud/spaces/1020/project/10706
Intro To R & RStudio / Intro to R and RStudio
Amelia McNamara
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 3.4.2
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
To Console
To Source
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
Zoom
Export
Publish
1.5e+09
1.0e+09
5.0e+08
0.0e+00
domgross_2013
0e+00
1e+08
2e+08
3e+08
4e+08
budget_2013
binary
FAIL
PASS

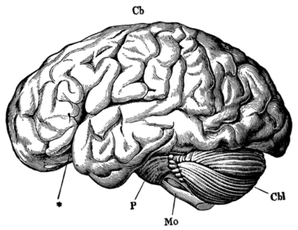
RStudio


RStudio Cloud
Amelia
Secure | https://rstudio.cloud/spaces/1020/project/10706
Intro To R & RStudio / Intro to R and RStudio
Amelia McNamara
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 3.4.2
Untitled1
Knit
Insert
Run
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
R Markdown
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
List
Global Environment
Environment is empty
Files Plots Packages Help Viewer
Zoom
Export
Publish
1.5e+09
1.0e+09
5.0e+08
0.0e+00
domgross_2013
0e+00
1e+08
2e+08
3e+08
4e+08
budget_2013
binary
FAIL
PASS

RStudio


RStudio - SDS220-s17
Amelia
Secure | https://rstudio.smith.edu/s/0d30ec506d8a79ee1b399/
File Edit Code View Plots Session Build Debug Profile Tools Help
amcnamara | Sessions
Go to file/function | Addins
SDS220-s17 | R 3.4.3
Console Terminal Compile PDF
~/SDS220-s17/
>
Environment History Connections
Import Dataset | List
Global Environment
Environment is empty
Files Plots Packages Help Viewer
Zoom | Export

R - A computer language for scientists



Human thought

`map()`

`sapply()`

`for()`

C++

Machine language

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))
```

SMITH COLLEGE

https://www.rstudio.com/resources/cheatsheets/

Data Science in the tidyverse

Amelia McNamara

amelia.mn

@AmeliaMN