R-powered Data Science

Allissa Dillman, PhD

Training Strategist
CIT Cloud Services | STRIDES
Outreach Coordinator
Office of Data Science Strategy

What is

A programming language and open-source software environment that can:

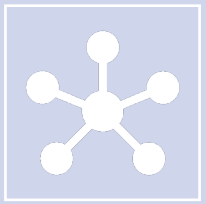
Manipulate data	Visualize	Perform statistics

Free and relatively easy to run in any environment

https://www.r-project.org/

What is RStudio

An Integrated Development Environment (IDE)



Software with R, a graphics viewer, file tracker, code notebook…etc. for interactive one stop shop R coding

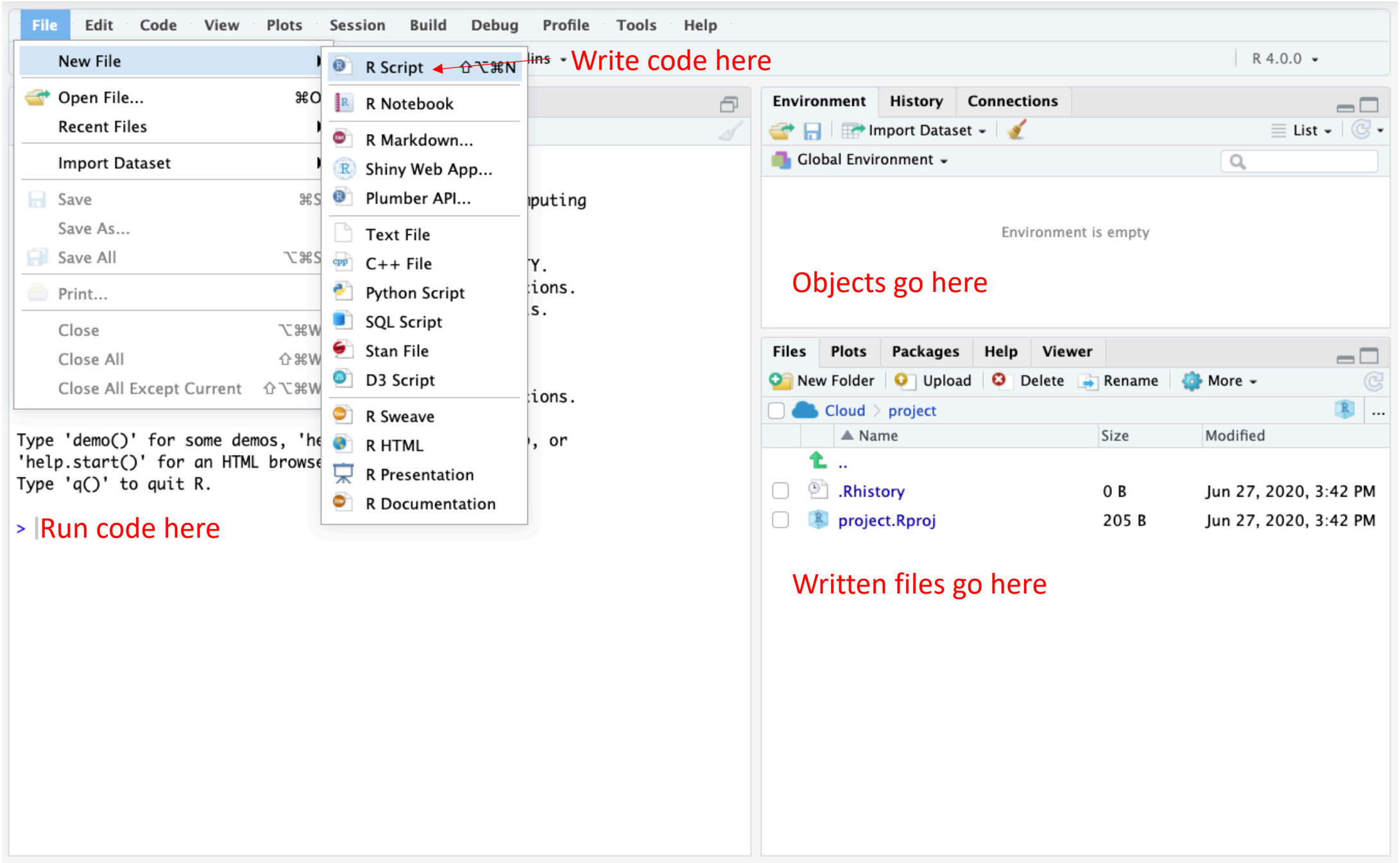
https://rstudio.com/

What we will learn today!

- A tour of R studio
- Basic R commands
- Helpful help!
- Downloading/loading R packages
- Loading data in numerous formats
- Making a table
- Manipulating/filtering data
- Writing data to a file
- Visualizing data
- Writing graphics to a file

R Script

- A text file that contains R code
- Can use # to make notes
  - # is a symbol that is recognized as “skip this line”
- There are a couple of more complex flavors that allow code, notes, and visualization
  - R Notebook
  - R Markdown
- An R Script can be saved and shared


File
Edit
Code
View
Plots
Session
Build
Debug
Profile
Tools
Help
New File
Open File...
Recent Files
Import Dataset
Save
Save As...
Save All
Print...
Close
Close All
Close All Except Current
R Script
R Notebook
R Markdown...
Shiny Web App...
Plumber API...
Text File
C++ File
Python Script
SQL Script
Stan File
D3 Script
R Sweave
R HTML
R Presentation
R Documentation
Write code here
R 4.0.0
Environment
History
Connections
Import Dataset
List
Global Environment
Environment is empty
Objects go here
Files
Plots
Packages
Help
Viewer
New Folder
Upload
Delete
Rename
More
Cloud
project
Name
Size
Modified
..
.Rhistory
0 B
Jun 27, 2020, 3:42 PM
project.Rproj
205 B
Jun 27, 2020, 3:42 PM
Written files go here
Type 'demo()' for some demos, 'help.start()' for an HTML browser
Type 'q()' to quit R.
Run code here

R is like a calculator

- Enter the following

7+7

d=5

c=3

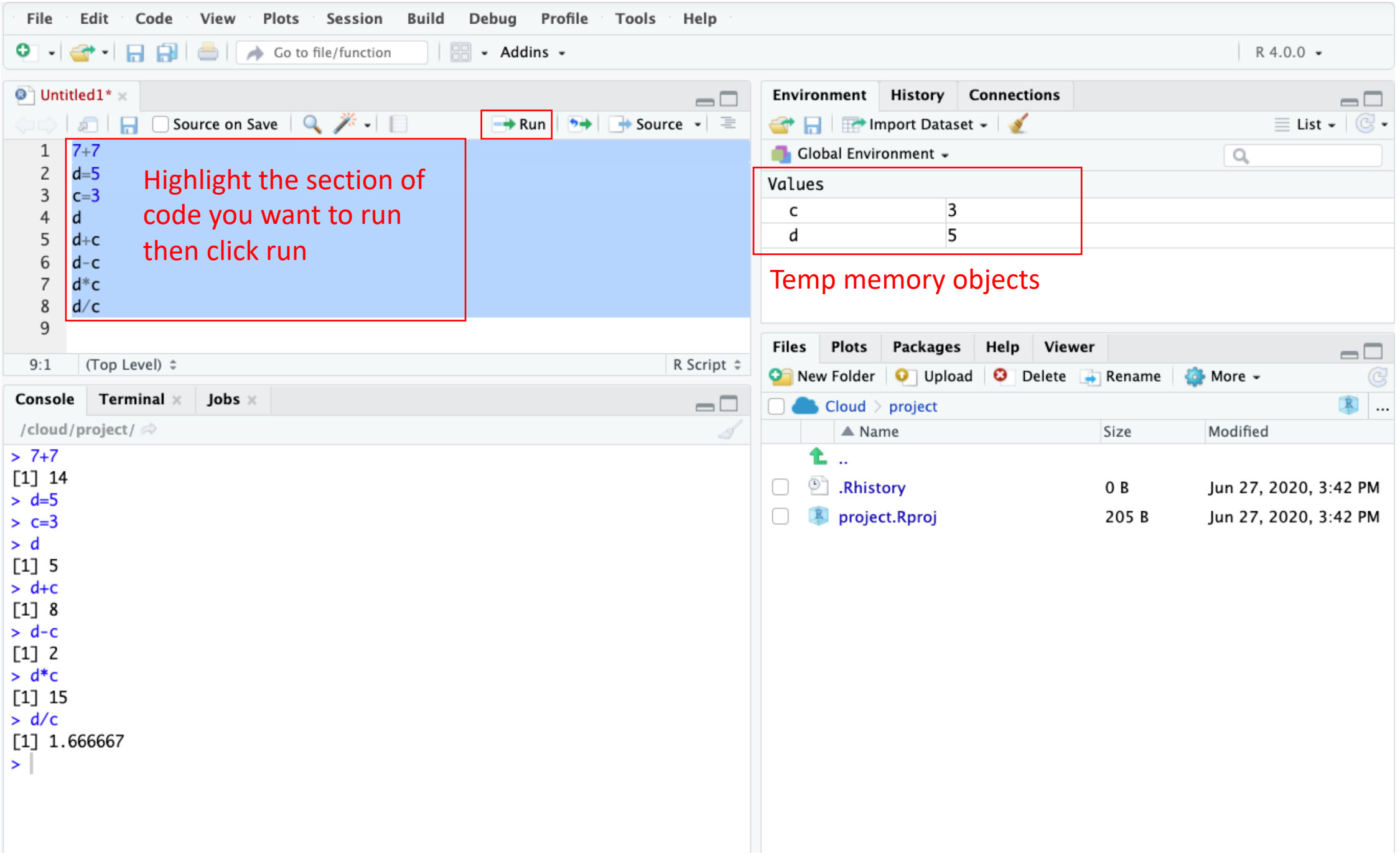
- If I type d it will print out the contents of d which is 5

d+c

d-c

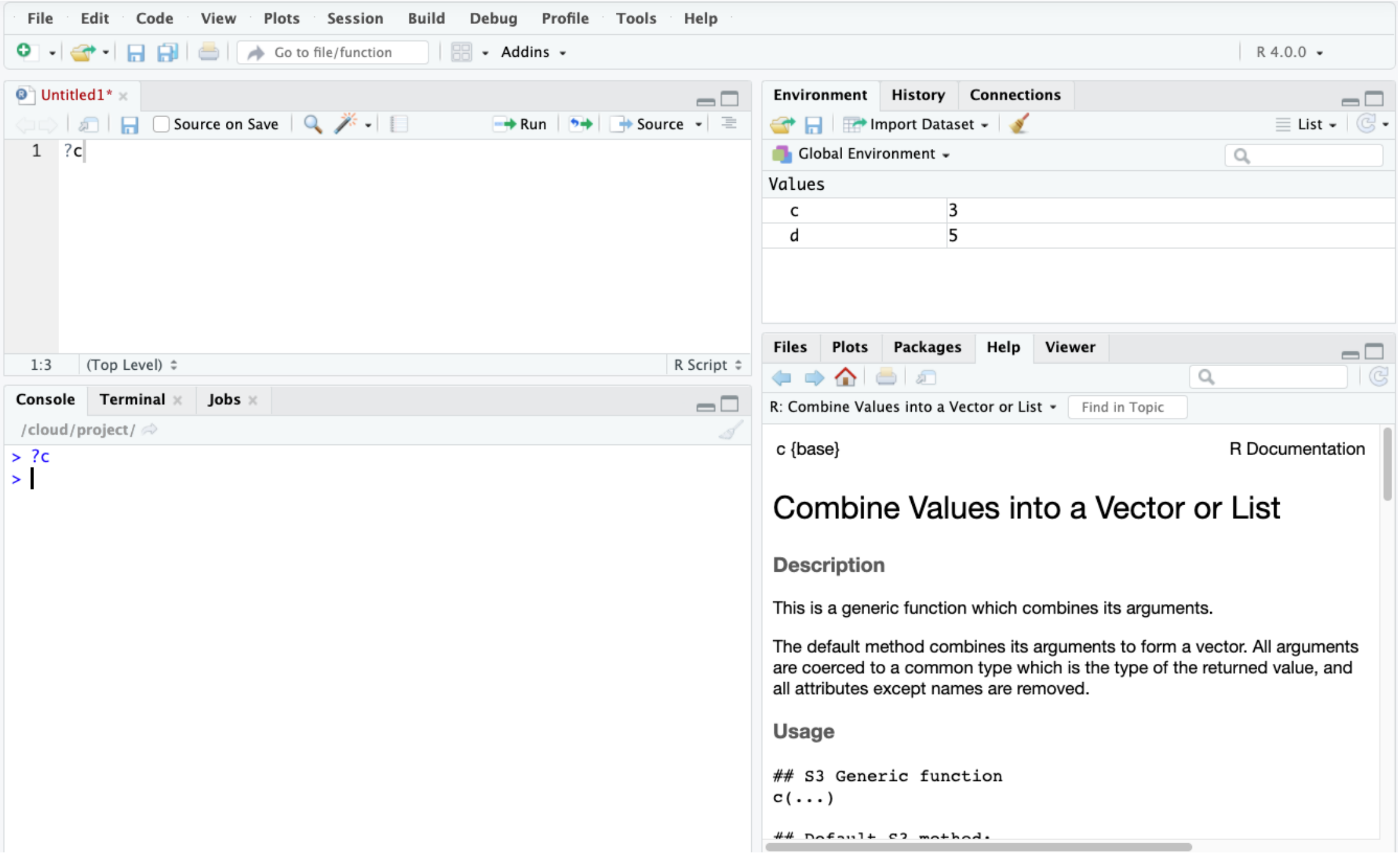
d*c

d/c


File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 4.0.0
Untitled1*
Source on Save
Run
Source
1 7+7
2 d=5
3 c=3
4 d
5 d+c
6 d-c
7 d*c
8 d/c
9
Highlight the section of code you want to run then click run
9:1
(Top Level)
R Script
Console
Terminal
Jobs
/cloud/project/
> 7+7
[1] 14
> d=5
> c=3
> d
[1] 5
> d+c
[1] 8
> d-c
[1] 2
> d*c
[1] 15
> d/c
[1] 1.666667
>
Environment
History
Connections
Import Dataset
List
Global Environment
Values
c 3
d 5
Temp memory objects
Files
Plots
Packages
Help
Viewer
New Folder
Upload
Delete
Rename
More
Cloud
project
Name
Size
Modified
..
.Rhistory 0 B Jun 27, 2020, 3:42 PM
project.Rproj 205 B Jun 27, 2020, 3:42 PM

Make a Vector

- Vectors are a data structure in R
- -list of characters
  - -list of numbers
  - -has to be same data type
- Use the c() command to enter a bunch of numbers together
- Don’t know how to use it?
- Type : ?c

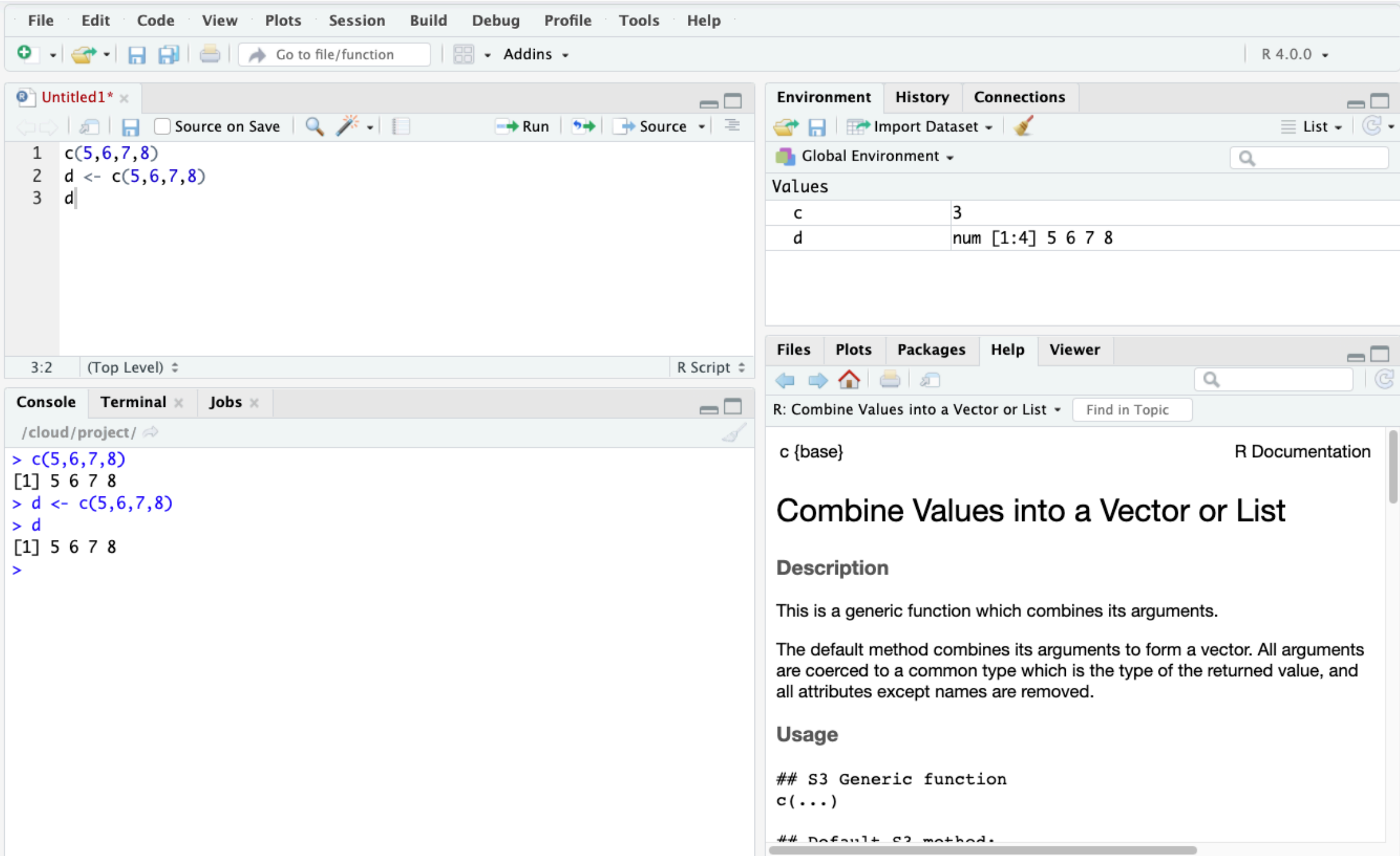

File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 4.0.0
Untitled1*
Source on Save
Run
Source
1 ?c
1:3
(Top Level)
R Script
Console
Terminal
Jobs
/cloud/project/
> ?c
>
Environment
History
Connections
Import Dataset
List
Global Environment
Values
c 3
d 5
Files
Plots
Packages
Help
Viewer
R: Combine Values into a Vector or List
Find in Topic
c {base}
R Documentation
Combine Values into a Vector or List
Description
This is a generic function which combines its arguments.
The default method combines its arguments to form a vector. All arguments are coerced to a common type which is the type of the returned value, and all attributes except names are removed.
Usage
S3 Generic function
c(...)

Typing ?

- Typing “?” then the command name will give you help onthe command name
- Typically the bottom of the sectionwill have examples that you can try
- Now let's use the c() command
- c(5,6,7,8)
- This just prints the output to the screen

Create an object

- We can create an object for R to keep in its temporary memory
- We will assign a name to the vector we used before
- Type "d <-" in front of c(5,6,7,8) : d <- c(5,6,7,8)
- We created an R object called “d” that is a vector of 5,6,7,8
- Keyboard shortcut for <-
  - -PC: Alt and - at the same time
  - -Mac: option and - at the same time

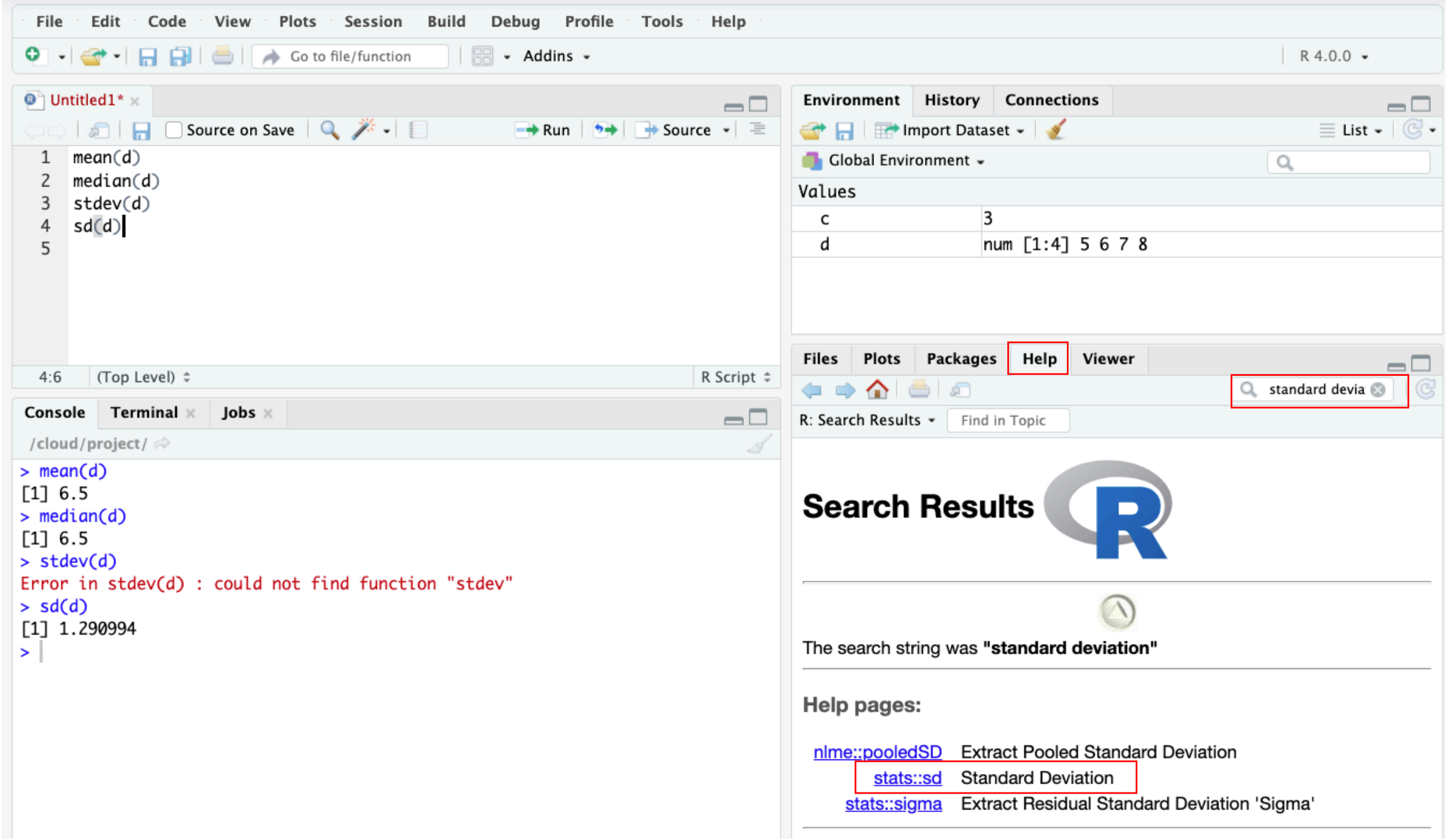

File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 4.0.0
Untitled1*
Source on Save
Run
Source
1 c(5,6,7,8)
2 d <- c(5,6,7,8)
3 d
3:2
(Top Level)
R Script
Console
Terminal
Jobs
/cloud/project/
> c(5,6,7,8)
[1] 5 6 7 8
> d <- c(5,6,7,8)
> d
[1] 5 6 7 8
>
Environment
History
Connections
Import Dataset
List
Global Environment
Values
c 3
d num [1:4] 5 6 7 8
Files
Plots
Packages
Help
Viewer
R: Combine Values into a Vector or List
Find in Topic
c {base}
R Documentation
Combine Values into a Vector or List
Description
This is a generic function which combines its arguments.
The default method combines its arguments to form a vector. All arguments are coerced to a common type which is the type of the returned value, and all attributes except names are removed.
Usage
S3 Generic function
c(...)

Basic Summary Statistics

- mean(d)
- median(d)
- stdev(d)

Error in stdev(d) : could not find function "stdev"

- How doyou find the command?
  - Use the search bar in *Rstudio*
  - Type “Standard deviation”
  - Stats::sd
  - sd(d)


File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 4.0.0
Untitled1*
Source on Save
Run
Source
1 mean(d)
2 median(d)
3 stdev(d)
4 sd(d)
5
4:6
(Top Level)
R Script
Console
Terminal
Jobs
/cloud/project/
> mean(d)
[1] 6.5
> median(d)
[1] 6.5
> stdev(d)
Error in stdev(d) : could not find function "stdev"
> sd(d)
[1] 1.290994
>
Environment
History
Connections
Import Dataset
List
Global Environment
Values
c 3
d num [1:4] 5 6 7 8
Files
Plots
Packages
Help
Viewer
standard devia
R: Search Results
Find in Topic
Search Results
The search string was "standard deviation"
Help pages:
nlme::pooledSD Extract Pooled Standard Deviation
stats::sd Standard Deviation
stats::sigma Extract Residual Standard Deviation 'Sigma'

Make more vectors

- e <- c(11,12,13,15)
- f <- c(1,2,3,4)
- g <- c(1,2,3,15)

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | R 4.0.0

Untitled1*

Source on Save | Run | Source

```
1  e <- c(11,12,13,15)
2  f <- c(1,2,3,4)
3  g <- c(1,2,3,15)
4  e
5  f
6  g
7
```

7:1 (Top Level) R Script

Console Terminal Jobs

/cloud/project/

```
> e <- c(11,12,13,15)
> f <- c(1,2,3,4)
> g <- c(1,2,3,15)
> e
[1] 11 12 13 15
> f
[1] 1 2 3 4
> g
[1]  1  2  3 15
>
```

Environment History Connections

Import Dataset | List

Global Environment

Values

c	3
d	num [1:4] 5 6 7 8
e	num [1:4] 11 12 13 15
f	num [1:4] 1 2 3 4
g	num [1:4] 1 2 3 15

Files Plots Packages Help Viewer

standard devia

R: Search Results | Find in Topic

Search Results

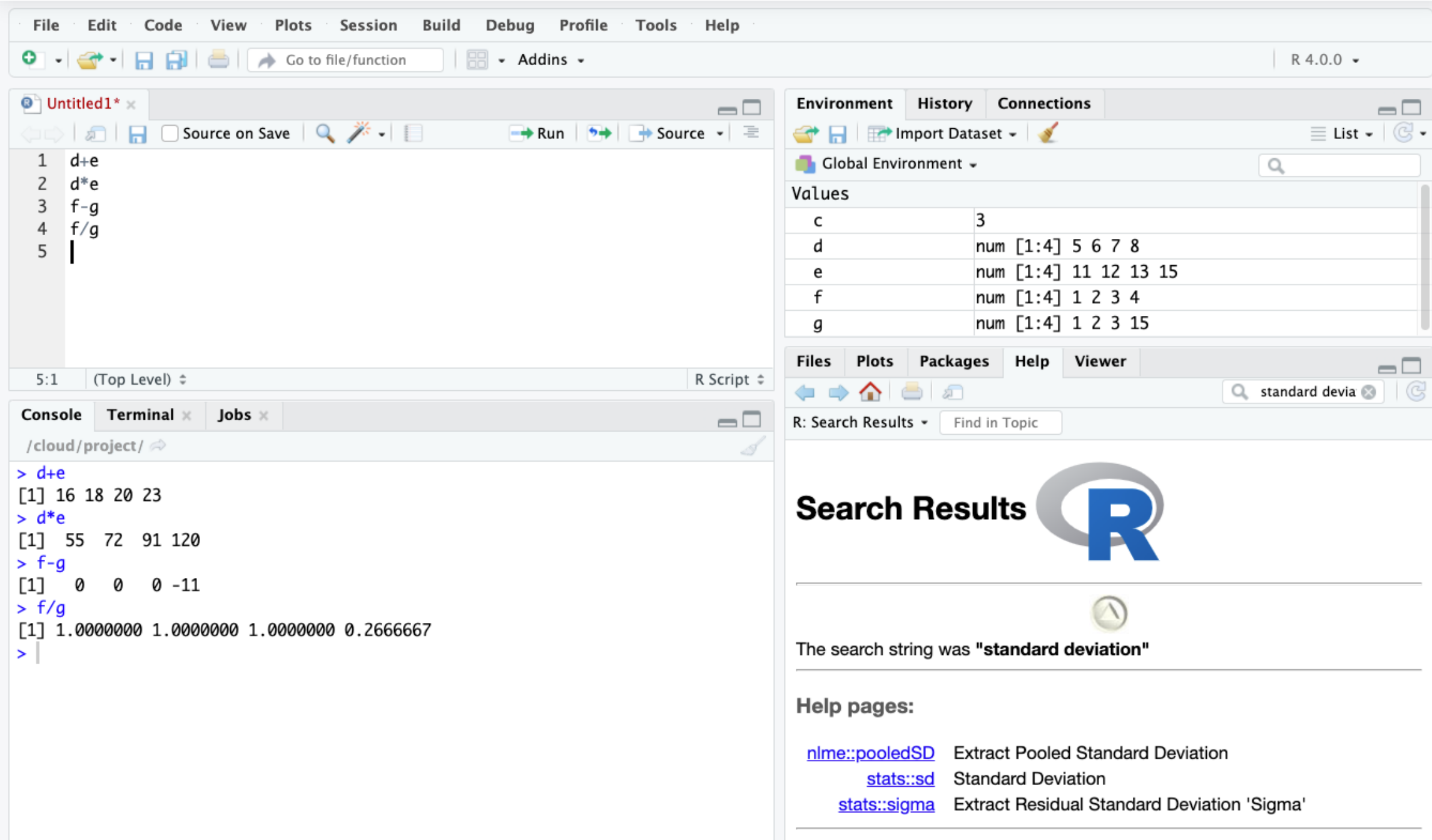
The search string was **"standard deviation"**

Help pages:

nlme::pooledSD	Extract Pooled Standard Deviation
stats::sd	Standard Deviation
stats::sigma	Extract Residual Standard Deviation 'Sigma'

Perform Vector Calculations

- d+e
- d*e
- f-g
- f/g


File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 4.0.0
Untitled1*
Source on Save
Run
Source
1 d+e
2 d*e
3 f-g
4 f/g
5
5:1
(Top Level)
R Script
Console
Terminal
Jobs
/cloud/project/
> d+e
[1] 16 18 20 23
> d*e
[1] 55 72 91 120
> f-g
[1] 0 0 0 -11
> f/g
[1] 1.0000000 1.0000000 1.0000000 0.2666667
>
Environment
History
Connections
Import Dataset
List
Global Environment
Values
c 3
d num [1:4] 5 6 7 8
e num [1:4] 11 12 13 15
f num [1:4] 1 2 3 4
g num [1:4] 1 2 3 15
Files
Plots
Packages
Help
Viewer
standard devia
R: Search Results
Find in Topic
Search Results
The search string was "standard deviation"
Help pages:
nlme::pooledSD Extract Pooled Standard Deviation
stats::sd Standard Deviation
stats::sigma Extract Residual Standard Deviation 'Sigma'

Combine Vectors

- h <- c(d,e)
- h <-rbind(d,e)
- h <- cbind(d,e)

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | R 4.0.0

Untitled1*

Source on Save | Run | Source

```
1  h <- c(d,e)
2  h
3  h <-rbind(d,e)
4  h
5  h <- cbind(d,e)
6  h
```

6:2 (Top Level) R Script

Console Terminal Jobs

/cloud/project/

```
> h <- c(d,e)
> h
[1]  5  6  7  8 11 12 13 15
> h <-rbind(d,e)
> h
  [,1] [,2] [,3] [,4]
d    5    6    7    8
e   11   12   13   15
> h <- cbind(d,e)
> h
     d  e
[1,] 5 11
[2,] 6 12
[3,] 7 13
[4,] 8 15
>
```

Environment History Connections

Import Dataset | List

Global Environment

Data	
h	num [1:4, 1:2] 5 6 7 8 11 12 13 15
Values	
c	3
d	num [1:4] 5 6 7 8
e	num [1:4] 11 12 13 15

Files Plots Packages Help Viewer

standard devia

R: Search Results | Find in Topic

Search Results

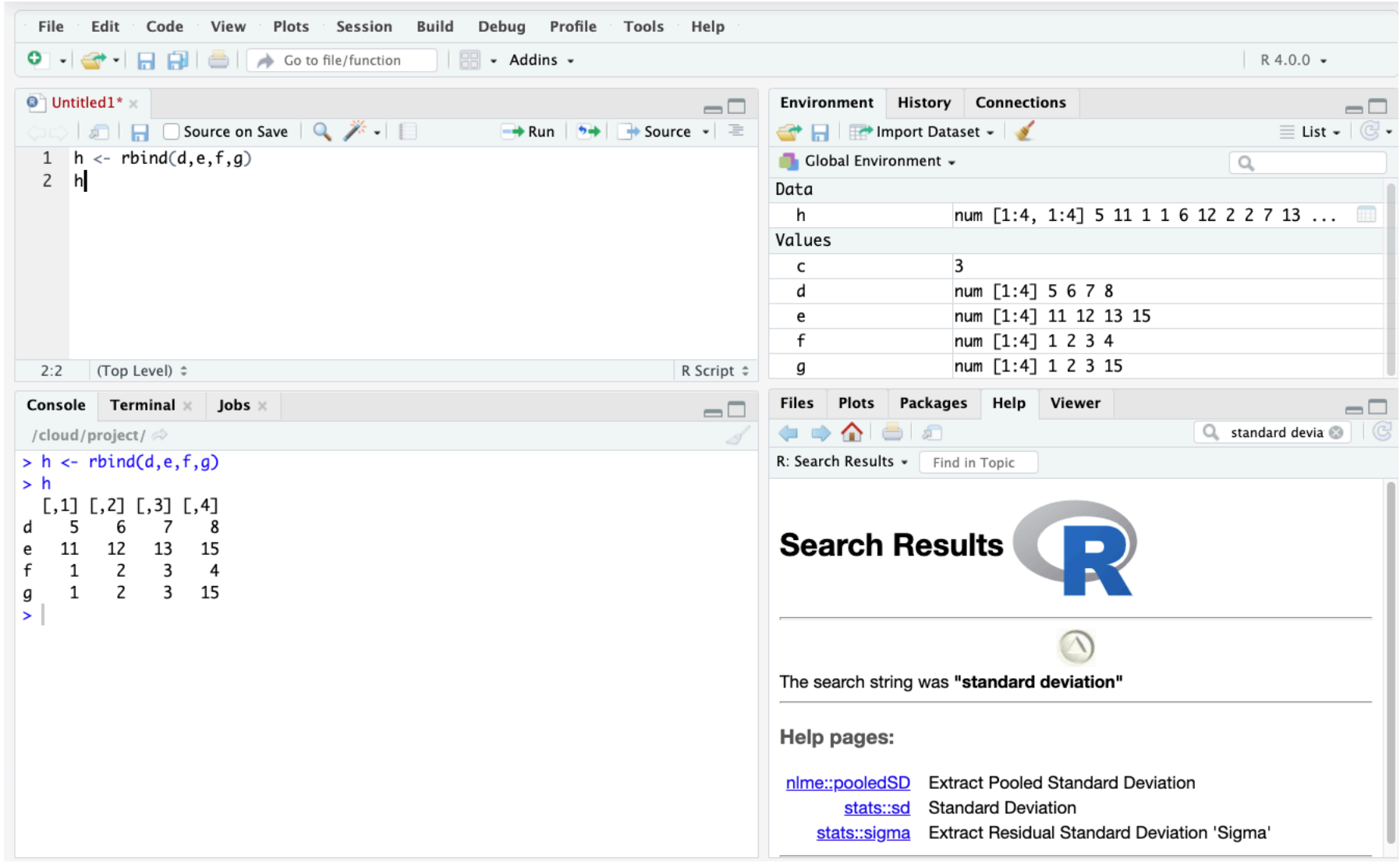
The search string was **"standard deviation"**

Help pages:

nlme::pooledSD	Extract Pooled Standard Deviation
stats::sd	Standard Deviation
stats::sigma	Extract Residual Standard Deviation 'Sigma'

Let's Make a Data Table!

h <- rbind(d,e,f,g)



Column and RowNames

- Independent of the data
- Makes it easier to work with data later
  - colnames
  - rownames
- Type the following:
  - colnames(h) <- c("Col1","Col2","Col3","Col4")
  - rownames(h) <- c("Row1","Row2","Row3","Row4")

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | R 4.0.0

Untitled1*

Source on Save | Run | Source

```
1  h
2  colnames(h) <- c("Col1","Col2","Col3","Col4")
3  h
4  rownames(h) <- c("Row1","Row2","Row3","Row4")
5  h
6
```

6:1 (Top Level) R Script

Console Terminal Jobs

/cloud/project/

```
  [,1] [,2] [,3] [,4]
d    5    6    7    8
e   11   12   13   15
f    1    2    3    4
g    1    2    3   15
> colnames(h) <- c("Col1","Col2","Col3","Col4")
> h
  Col1 Col2 Col3 Col4
d    5    6    7    8
e   11   12   13   15
f    1    2    3    4
g    1    2    3   15
> rownames(h) <- c("Row1","Row2","Row3","Row4")
> h
     Col1 Col2 Col3 Col4
Row1    5    6    7    8
Row2   11   12   13   15
Row3    1    2    3    4
Row4    1    2    3   15
>
```

Environment History Connections

Import Dataset | List

Global Environment

Data	
h	num [1:4, 1:4] 5 11 1 1 6 12 2 2 7 13 ...
Values	
c	3
d	num [1:4] 5 6 7 8
e	num [1:4] 11 12 13 15
f	num [1:4] 1 2 3 4
g	num [1:4] 1 2 3 15

Files Plots Packages Help Viewer

standard devia

R: Search Results | Find in Topic

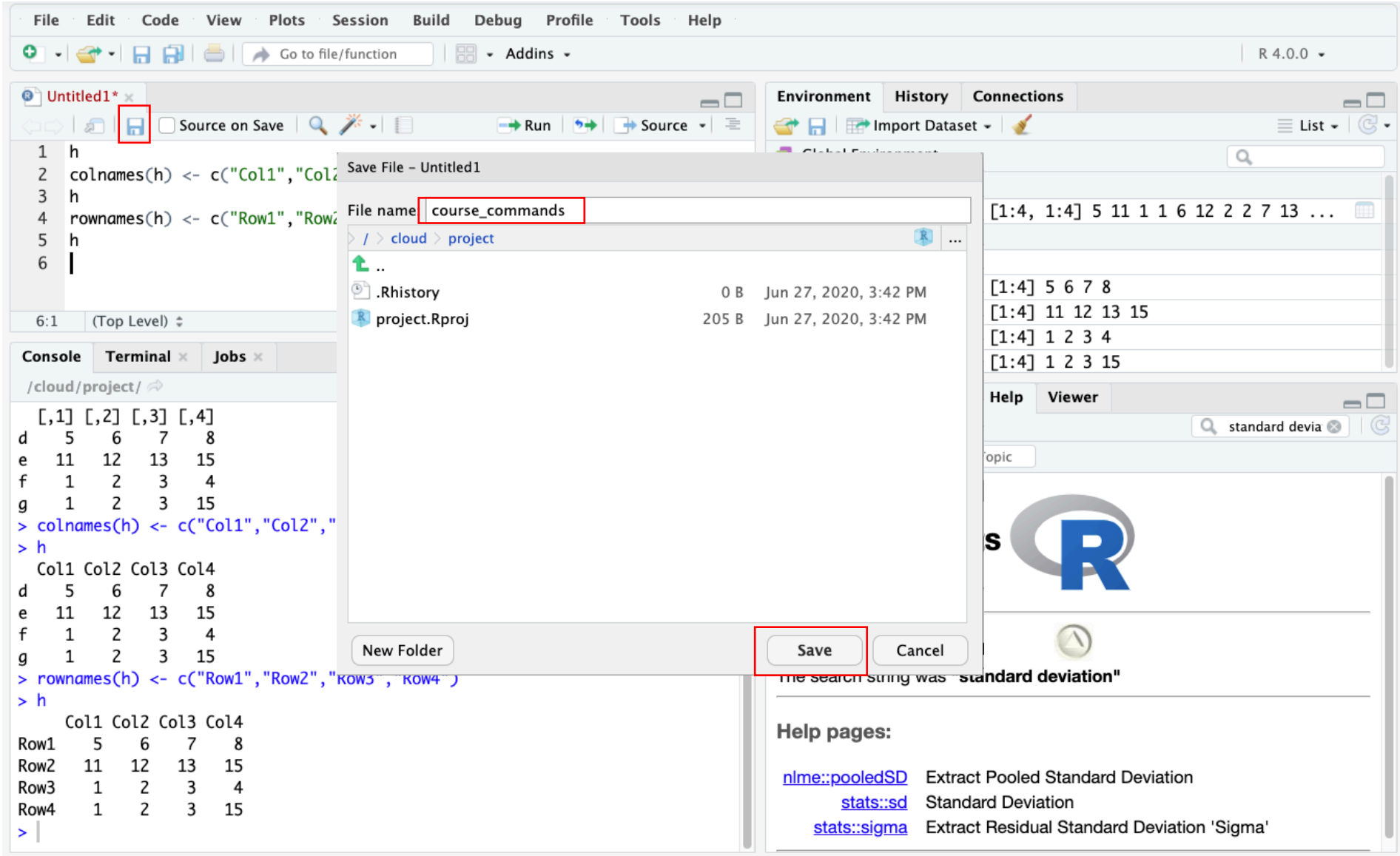
Search Results

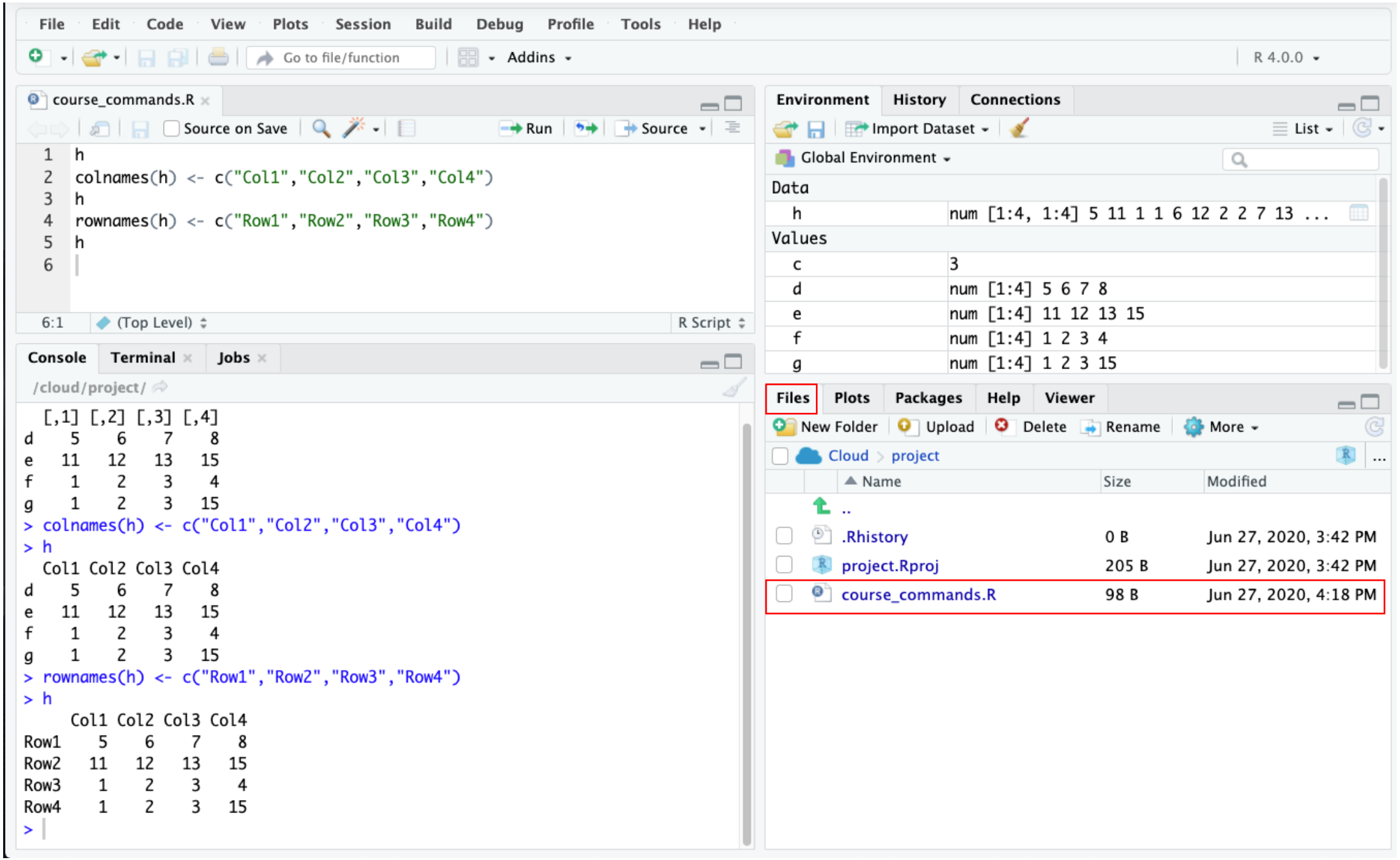
The search string was **"standard deviation"**

Help pages:

nlme::pooledSD	Extract Pooled Standard Deviation
stats::sd	Standard Deviation
stats::sigma	Extract Residual Standard Deviation 'Sigma'

Save your amazing work!

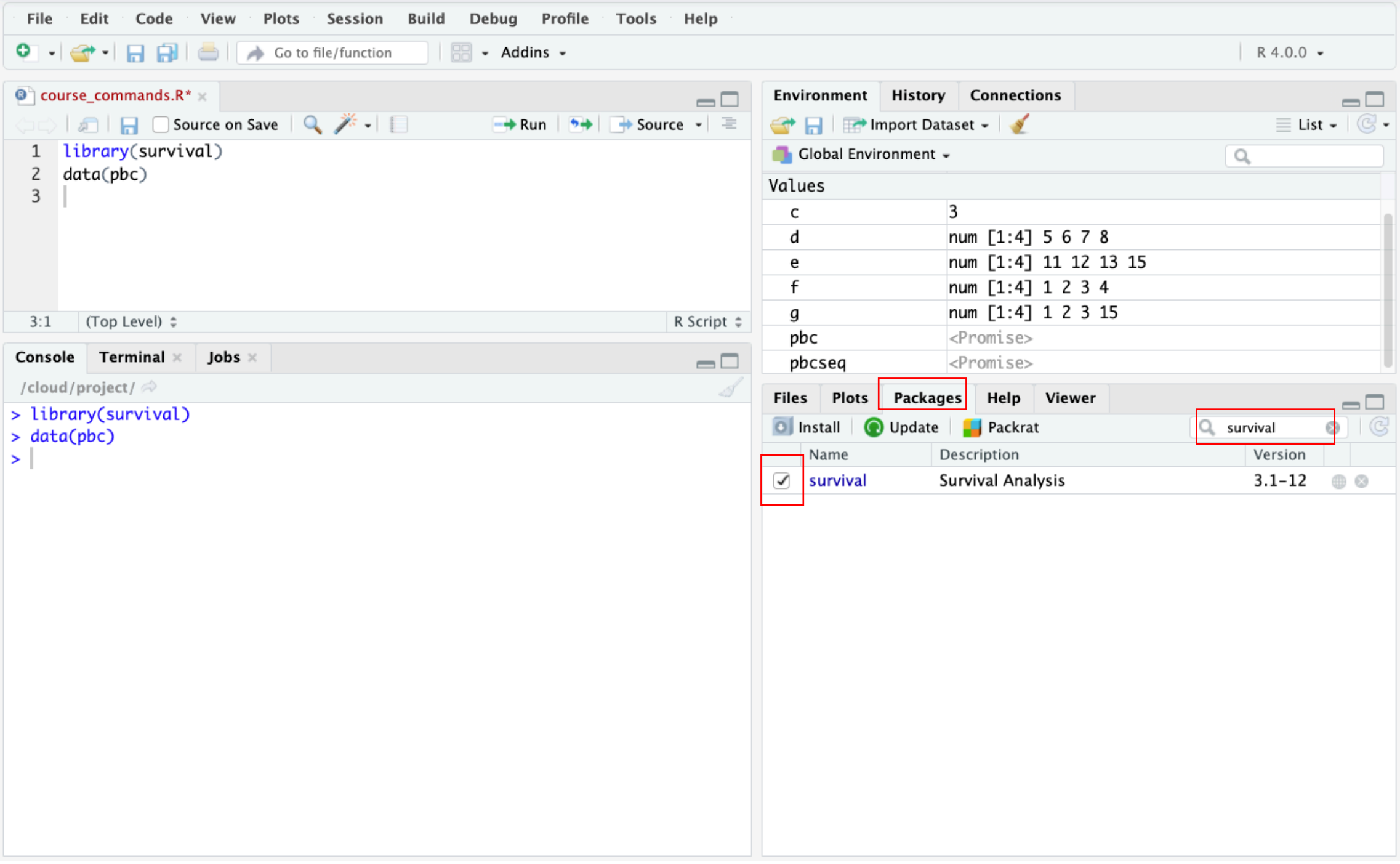


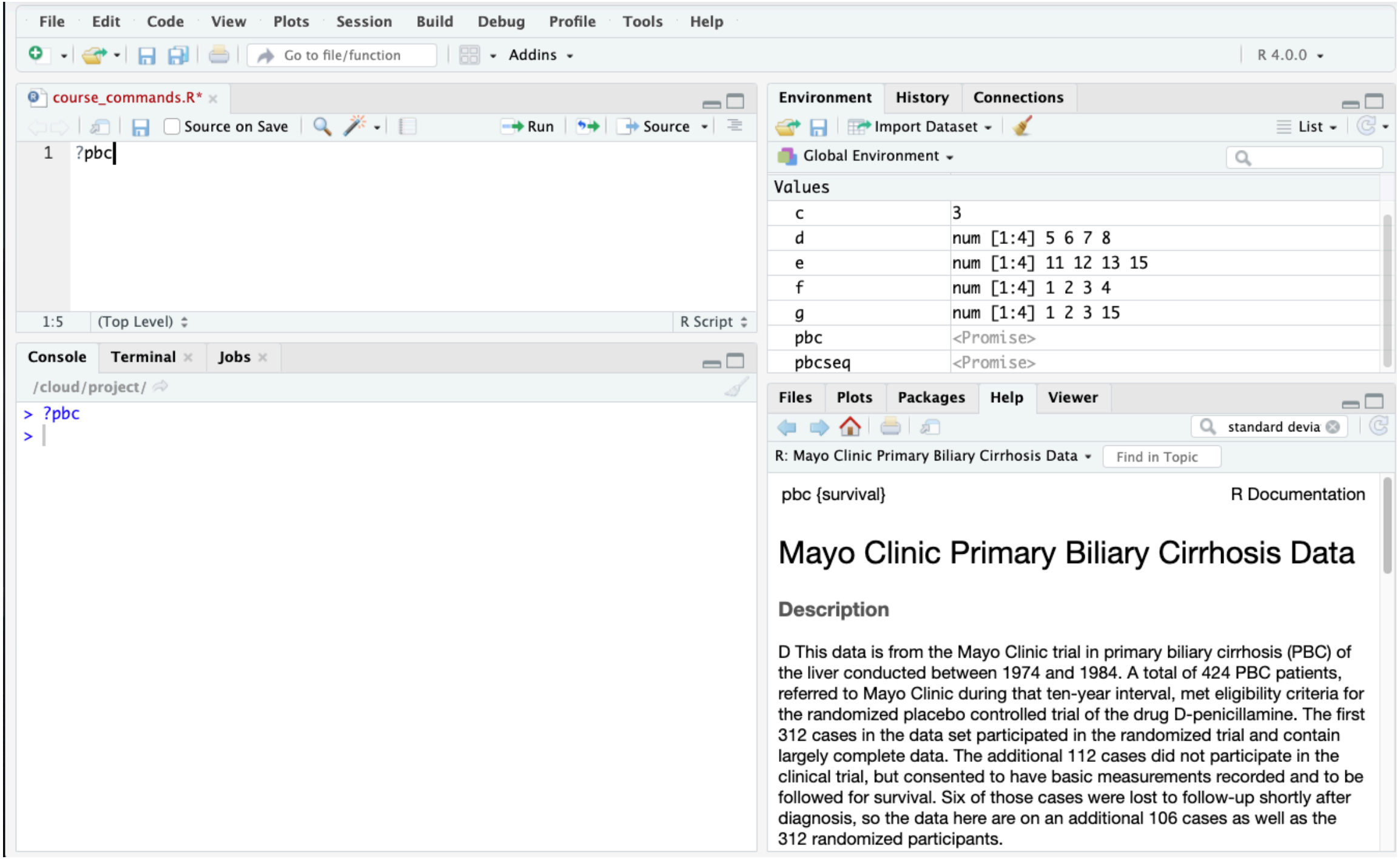

File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 4.0.0
course_commands.R
Source on Save
Run
Source
1 h
2 colnames(h) <- c("Col1","Col2","Col3","Col4")
3 h
4 rownames(h) <- c("Row1","Row2","Row3","Row4")
5 h
6
6:1
(Top Level)
R Script
Console
Terminal
Jobs
/cloud/project/
[,1] [,2] [,3] [,4]
d 5 6 7 8
e 11 12 13 15
f 1 2 3 4
g 1 2 3 15
> colnames(h) <- c("Col1","Col2","Col3","Col4")
> h
Col1 Col2 Col3 Col4
d 5 6 7 8
e 11 12 13 15
f 1 2 3 4
g 1 2 3 15
> rownames(h) <- c("Row1","Row2","Row3","Row4")
> h
Col1 Col2 Col3 Col4
Row1 5 6 7 8
Row2 11 12 13 15
Row3 1 2 3 4
Row4 1 2 3 15
>
Environment
History
Connections
Import Dataset
List
Global Environment
Data
h num [1:4, 1:4] 5 11 1 1 6 12 2 2 7 13 ...
Values
c 3
d num [1:4] 5 6 7 8
e num [1:4] 11 12 13 15
f num [1:4] 1 2 3 4
g num [1:4] 1 2 3 15
Files
Plots
Packages
Help
Viewer
New Folder
Upload
Delete
Rename
More
Cloud
project
Name
Size
Modified
..
.Rhistory 0 B Jun 27, 2020, 3:42 PM
project.Rproj 205 B Jun 27, 2020, 3:42 PM
course_commands.R 98 B Jun 27, 2020, 4:18 PM

Time to play with a real data set

R package:Survival

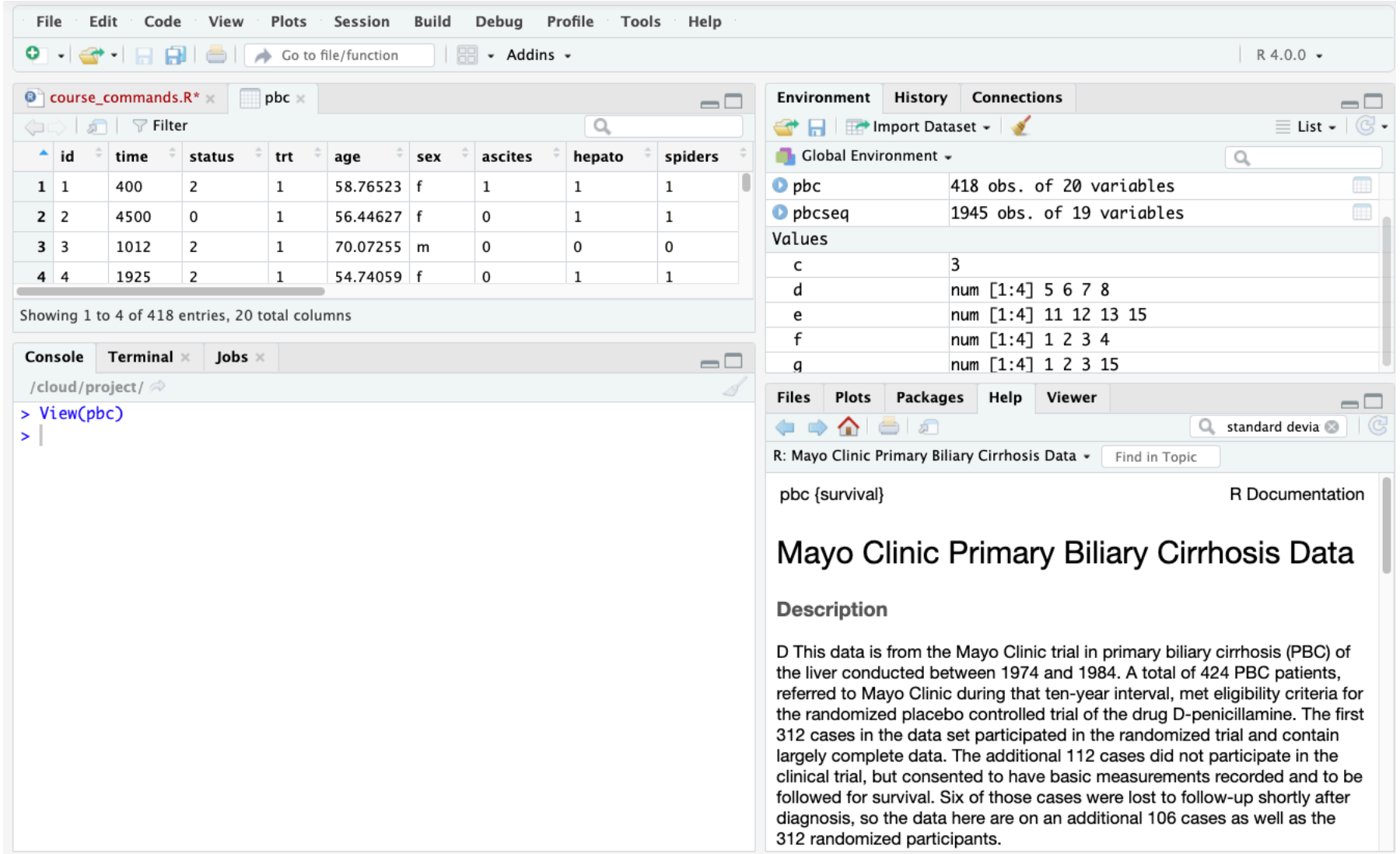
- R package: collection of R functions, complied code and data
  - Saved in a directory called “library”
  - Can be turned on and off
  - Made by different people commands may clash
  - Would also be very slow to load everything every time
- To load the package: library(survival)
- Use data from this package: data(pbc)
- This data is from the Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver
- You can get more info about the data and each field by typing:
- ?pbc


File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 4.0.0
course_commands.R*
Source on Save
Run
Source
1 library(survival)
2 data(pbc)
3
3:1
(Top Level)
R Script
Console
Terminal
Jobs
/cloud/project/
> library(survival)
> data(pbc)
>
Environment
History
Connections
Import Dataset
List
Global Environment
Values
c 3
d num [1:4] 5 6 7 8
e num [1:4] 11 12 13 15
f num [1:4] 1 2 3 4
g num [1:4] 1 2 3 15
pbc <Promise>
pbcseq <Promise>
Files
Plots
Packages
Help
Viewer
Install
Update
Packrat
survival
Name
Description
Version
survival
Survival Analysis
3.1-12


File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 4.0.0
course_commands.R*
Source on Save
Run
Source
1 ?pbc
1:5 (Top Level)
R Script
Console Terminal Jobs
/cloud/project/
> ?pbc
>
Environment History Connections
Import Dataset
List
Global Environment
Values
c 3
d num [1:4] 5 6 7 8
e num [1:4] 11 12 13 15
f num [1:4] 1 2 3 4
g num [1:4] 1 2 3 15
pbc <Promise>
pbcseq <Promise>
Files Plots Packages Help Viewer
standard devia
R: Mayo Clinic Primary Biliary Cirrhosis Data
Find in Topic
pbc {survival} R Documentation
Mayo Clinic Primary Biliary Cirrhosis Data
Description
D This data is from the Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver conducted between 1974 and 1984. A total of 424 PBC patients, referred to Mayo Clinic during that ten-year interval, met eligibility criteria for the randomized placebo controlled trial of the drug D-penicillamine. The first 312 cases in the data set participated in the randomized trial and contain largely complete data. The additional 112 cases did not participate in the clinical trial, but consented to have basic measurements recorded and to be followed for survival. Six of those cases were lost to follow-up shortly after diagnosis, so the data here are on an additional 106 cases as well as the 312 randomized participants.

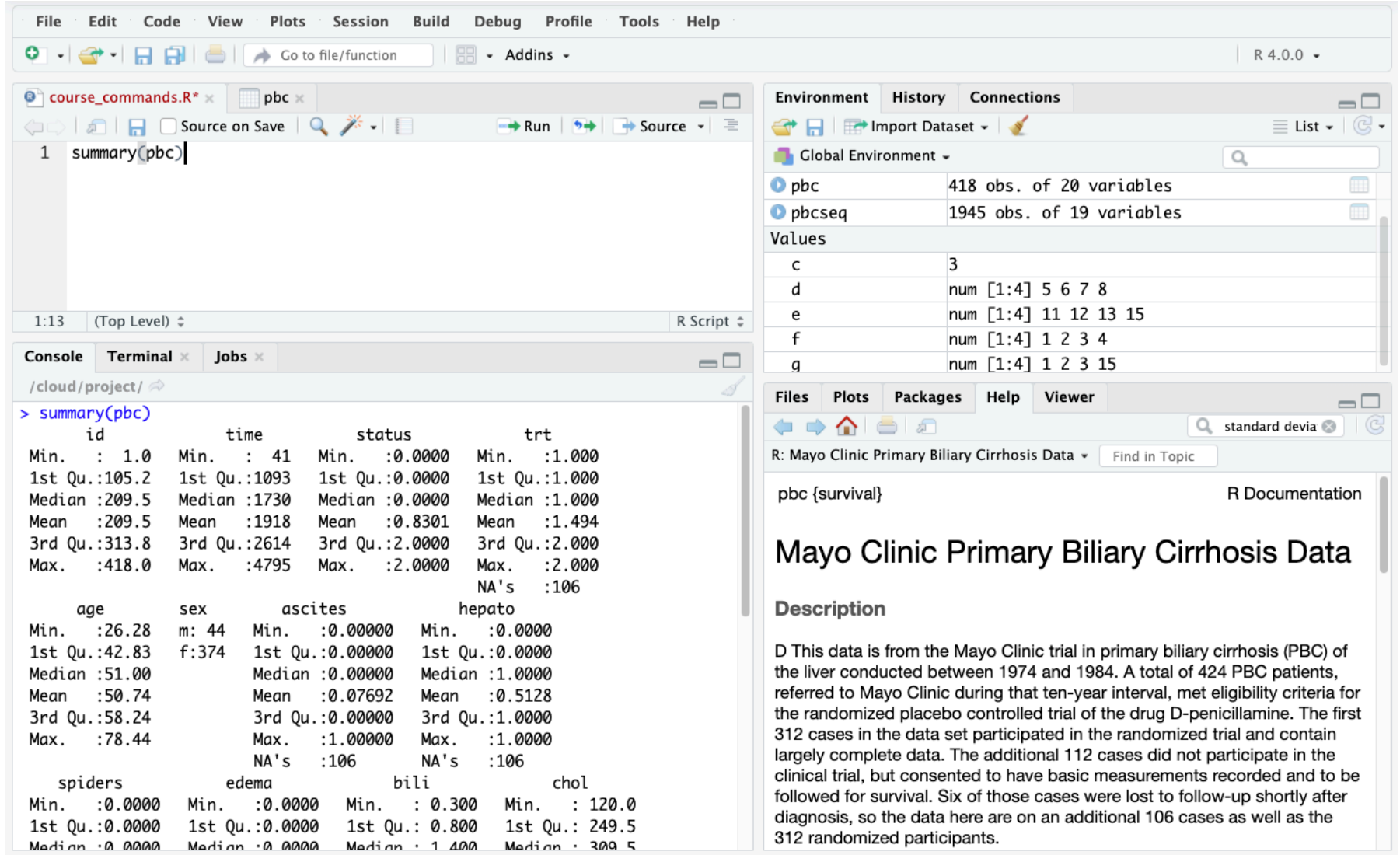
Let's examine our data set

Look at it like an excel file: View(pbc)



Survival Dataset Summary Stats

summary(pbc)



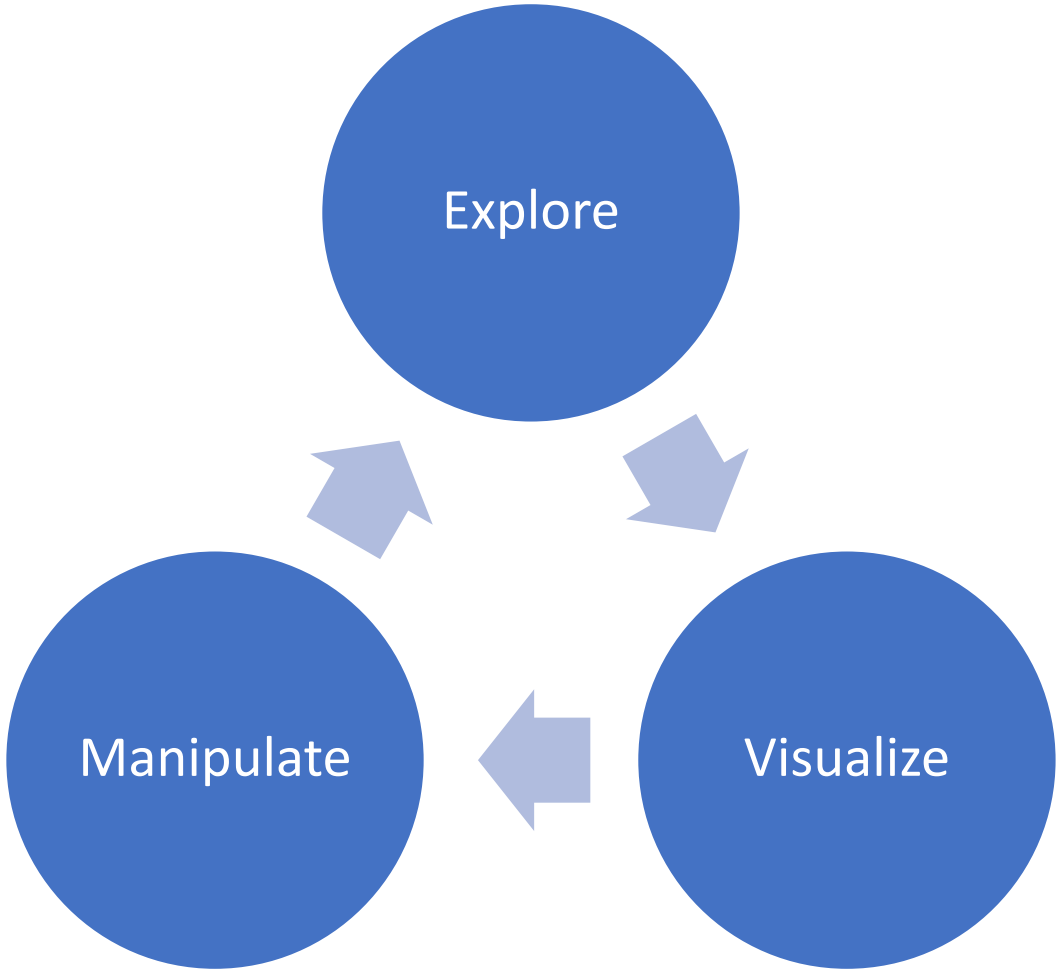
Data wrangling!

WE COULD USE BASE R

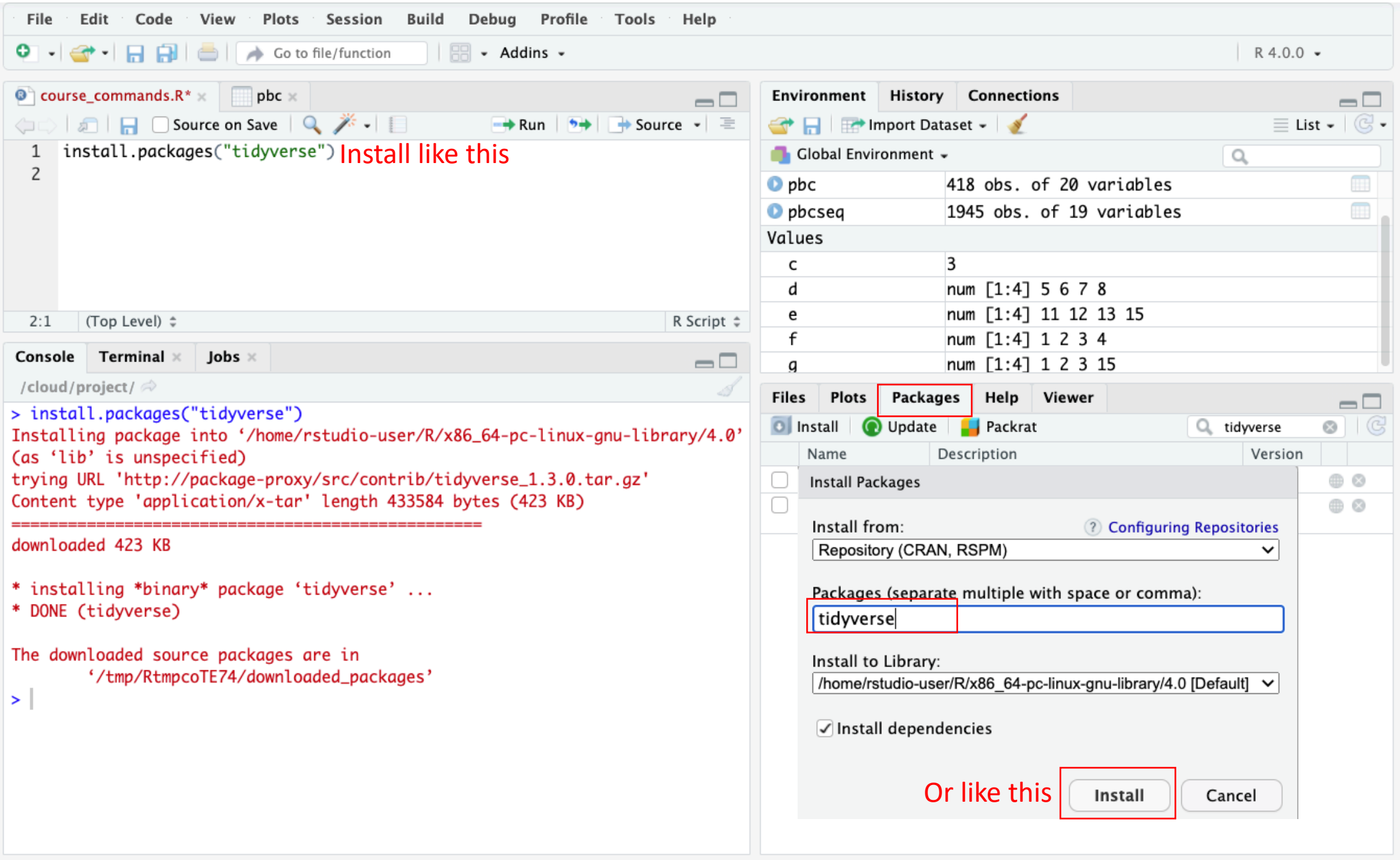
OR THE TIDYVERSE COLLECTION OF PACKAGES

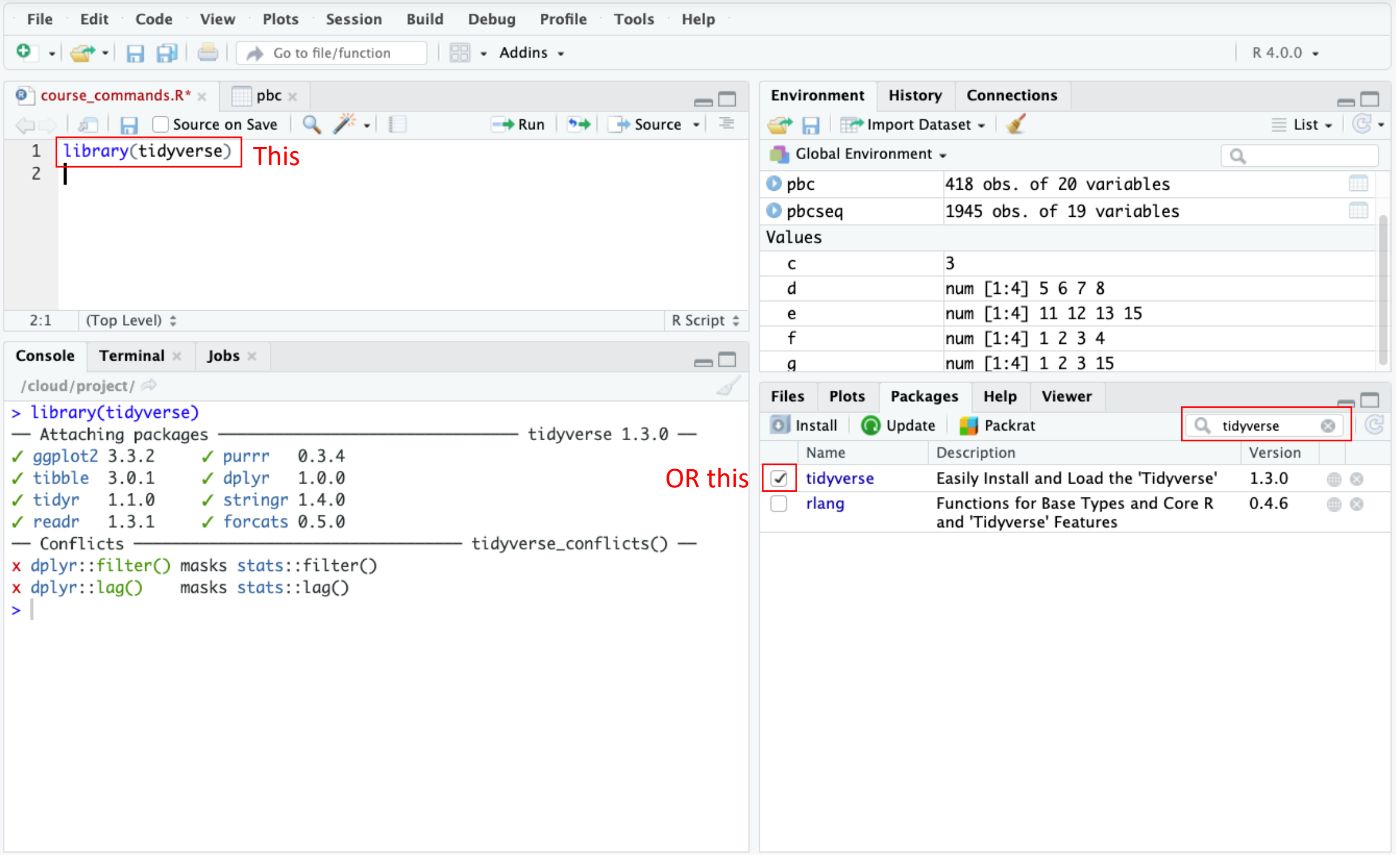
SPECIFICALLY DESIGNED FOR DATA SCIENCE

Package:Tidyverse



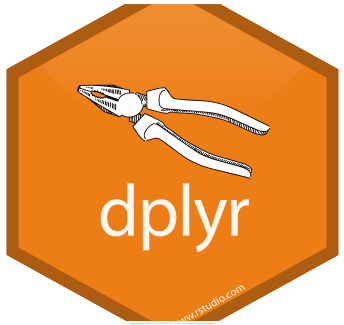
- First time install: install.packages("tidyverse")
- Saved in a directory called “library”
- Turn on package set: library(tidyverse)

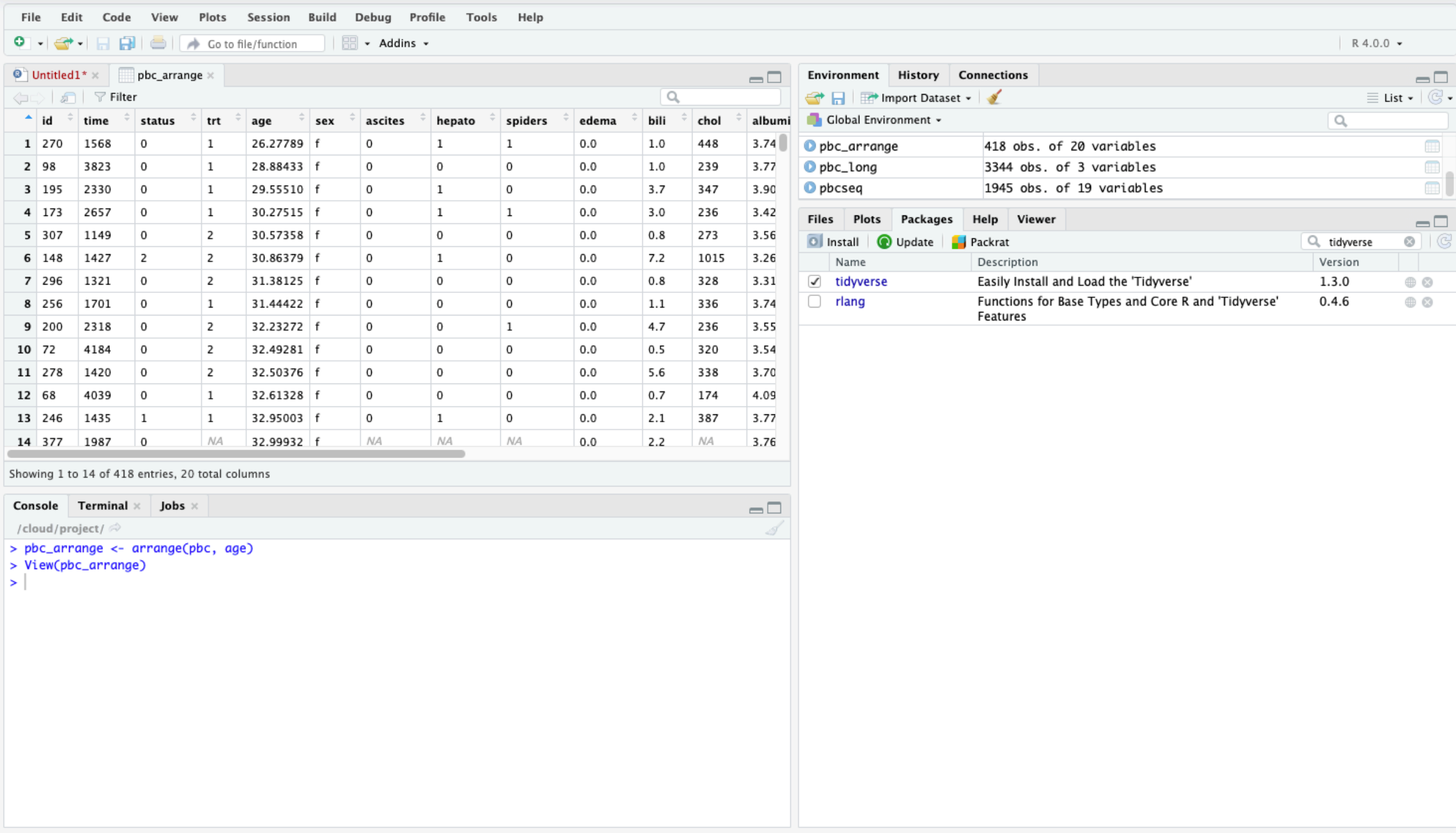

File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 4.0.0
course_commands.R*
pbc
Source on Save
Run
Source
1 install.packages("tidyverse")
2
Install like this
2:1 (Top Level)
R Script
Console
Terminal
Jobs
/cloud/project/
> install.packages("tidyverse")
Installing package into '/home/rstudio-user/R/x86_64-pc-linux-gnu-library/4.0'
(as 'lib' is unspecified)
trying URL 'http://package-proxy/src/contrib/tidyverse_1.3.0.tar.gz'
Content type 'application/x-tar' length 433584 bytes (423 KB)
==
downloaded 423 KB
* installing *binary* package 'tidyverse' ...
* DONE (tidyverse)
The downloaded source packages are in
'/tmp/RtmpcoTE74/downloaded_packages'
>
Environment
History
Connections
Import Dataset
List
Global Environment
pbc 418 obs. of 20 variables
pbcseq 1945 obs. of 19 variables
Values
c 3
d num [1:4] 5 6 7 8
e num [1:4] 11 12 13 15
f num [1:4] 1 2 3 4
g num [1:4] 1 2 3 15
Files
Plots
Packages
Help
Viewer
Install
Update
Packrat
tidyverse
Name
Description
Version
Install Packages
Install from:
Configuring Repositories
Repository (CRAN, RSPM)
Packages (separate multiple with space or comma):
tidyverse
Install to Library:
/home/rstudio-user/R/x86_64-pc-linux-gnu-library/4.0 [Default]
Install dependencies
Or like this
Install
Cancel


File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 4.0.0
course_commands.R*
pbc
Source on Save
Run
Source
1 library(tidyverse)
2
This
2:1 (Top Level)
R Script
Console
Terminal
Jobs
/cloud/project/
> library(tidyverse)
── Attaching packages ──────────────────────────────────── tidyverse 1.3.0 ──
✓ ggplot2 3.3.2 ✓ purrr 0.3.4
✓ tibble 3.0.1 ✓ dplyr 1.0.0
✓ tidyr 1.1.0 ✓ stringr 1.4.0
✓ readr 1.3.1 ✓ forcats 0.5.0
── Conflicts ─────────────────────────────────────── tidyverse_conflicts() ──
x dplyr::filter() masks stats::filter()
x dplyr::lag() masks stats::lag()
>
OR this
Environment
History
Connections
Import Dataset
List
Global Environment
pbc 418 obs. of 20 variables
pbcseq 1945 obs. of 19 variables
Values
c 3
d num [1:4] 5 6 7 8
e num [1:4] 11 12 13 15
f num [1:4] 1 2 3 4
g num [1:4] 1 2 3 15
Files
Plots
Packages
Help
Viewer
Install
Update
Packrat
tidyverse
Name
Description
Version
tidyverse
Easily Install and Load the 'Tidyverse'
1.3.0
rlang
Functions for Base Types and Core R and 'Tidyverse' Features
0.4.6

dplyr: Sort the data

- pbc_arrange <- arrange(pbc, age)
- View(pbc_arrange)
- pbc_arrange <- arrange(pbc, desc(age))
- View(pbc_arrange)




File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 4.0.0
Untitled1*
pbc_arrange
Filter
id time status trt age sex ascites hepato spiders edema bili chol albumi
1 270 1568 0 1 26.27789 f 0 1 1 0.0 1.0 448 3.74
2 98 3823 0 1 28.88433 f 0 0 0 0.0 1.0 239 3.77
3 195 2330 0 1 29.55510 f 0 1 0 0.0 3.7 347 3.90
4 173 2657 0 1 30.27515 f 0 1 1 0.0 3.0 236 3.42
5 307 1149 0 2 30.57358 f 0 0 0 0.0 0.8 273 3.56
6 148 1427 2 2 30.86379 f 0 1 0 0.0 7.2 1015 3.26
7 296 1321 0 2 31.38125 f 0 0 0 0.0 0.8 328 3.31
8 256 1701 0 1 31.44422 f 0 0 0 0.0 1.1 336 3.74
9 200 2318 0 2 32.23272 f 0 0 1 0.0 4.7 236 3.55
10 72 4184 0 2 32.49281 f 0 0 0 0.0 0.5 320 3.54
11 278 1420 0 2 32.50376 f 0 0 0 0.0 5.6 338 3.70
12 68 4039 0 1 32.61328 f 0 0 0 0.0 0.7 174 4.09
13 246 1435 1 1 32.95003 f 0 1 0 0.0 2.1 387 3.77
14 377 1987 0 NA 32.99932 f NA NA NA 0.0 2.2 NA 3.76
Showing 1 to 14 of 418 entries, 20 total columns
Console Terminal Jobs
/cloud/project/
> pbc_arrange <- arrange(pbc, age)
> View(pbc_arrange)
>
Environment History Connections
Import Dataset
List
Global Environment
pbc_arrange 418 obs. of 20 variables
pbc_long 3344 obs. of 3 variables
pbcseq 1945 obs. of 19 variables
Files Plots Packages Help Viewer
Install Update Packrat
tidyverse
Name Description Version
tidyverse Easily Install and Load the 'Tidyverse' 1.3.0
rlang Functions for Base Types and Core R and 'Tidyverse' Features 0.4.6

Untitled1* | pbc_arrange

Filter

	id	time	status	trt	age	sex	ascites	hepato	spiders	edema	bili	chol	albumi
1	253	1765	0	1	78.43943	m	1	1	1	0.0	7.1	243	3.03
2	92	388	2	1	76.70910	f	1	0	0	1.0	1.4	206	3.13
3	147	2995	0	1	75.01164	f	0	0	0	0.5	1.2	288	3.37
4	316	2071	2	NA	75.00068	f	NA	NA	NA	0.5	0.7	NA	3.96
5	260	1656	0	2	74.52430	m	0	1	0	0.0	5.6	232	3.59
6	207	2171	0	1	72.77207	f	0	0	0	0.5	0.6	NA	3.33
7	97	611	2	2	71.89322	m	0	1	0	0.5	2.0	420	3.26
8	267	179	2	1	70.90760	f	1	1	1	1.0	6.6	222	2.33
9	186	1576	2	1	70.83641	f	0	0	1	0.5	2.0	225	3.53
10	10	51	2	2	70.55989	f	1	0	1	1.0	12.6	200	2.74
11	3	1012	2	1	70.07255	m	0	0	0	0.5	1.4	176	3.48
12	178	2580	0	1	70.00411	f	0	0	0	0.0	0.6	NA	4.08
13	346	559	2	NA	70.00137	f	NA	NA	NA	0.5	0.6	NA	3.81
14	152	1152	2	1	69.94114	m	0	1	0	0.0	2.3	586	3.01

Showing 1 to 14 of 418 entries, 20 total columns

Console | Terminal | Jobs

/cloud/project/

```
> pbc_arrange <- arrange(pbc, desc(age))
> View(pbc_arrange)
>
```

Environment | History | Connections

Import Dataset | List

Global Environment

pbc_arrange	418 obs. of 20 variables
pbc_long	3344 obs. of 3 variables
pbcseq	1945 obs. of 19 variables

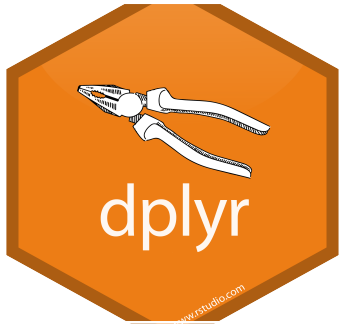
Files | Plots | Packages | Help | Viewer

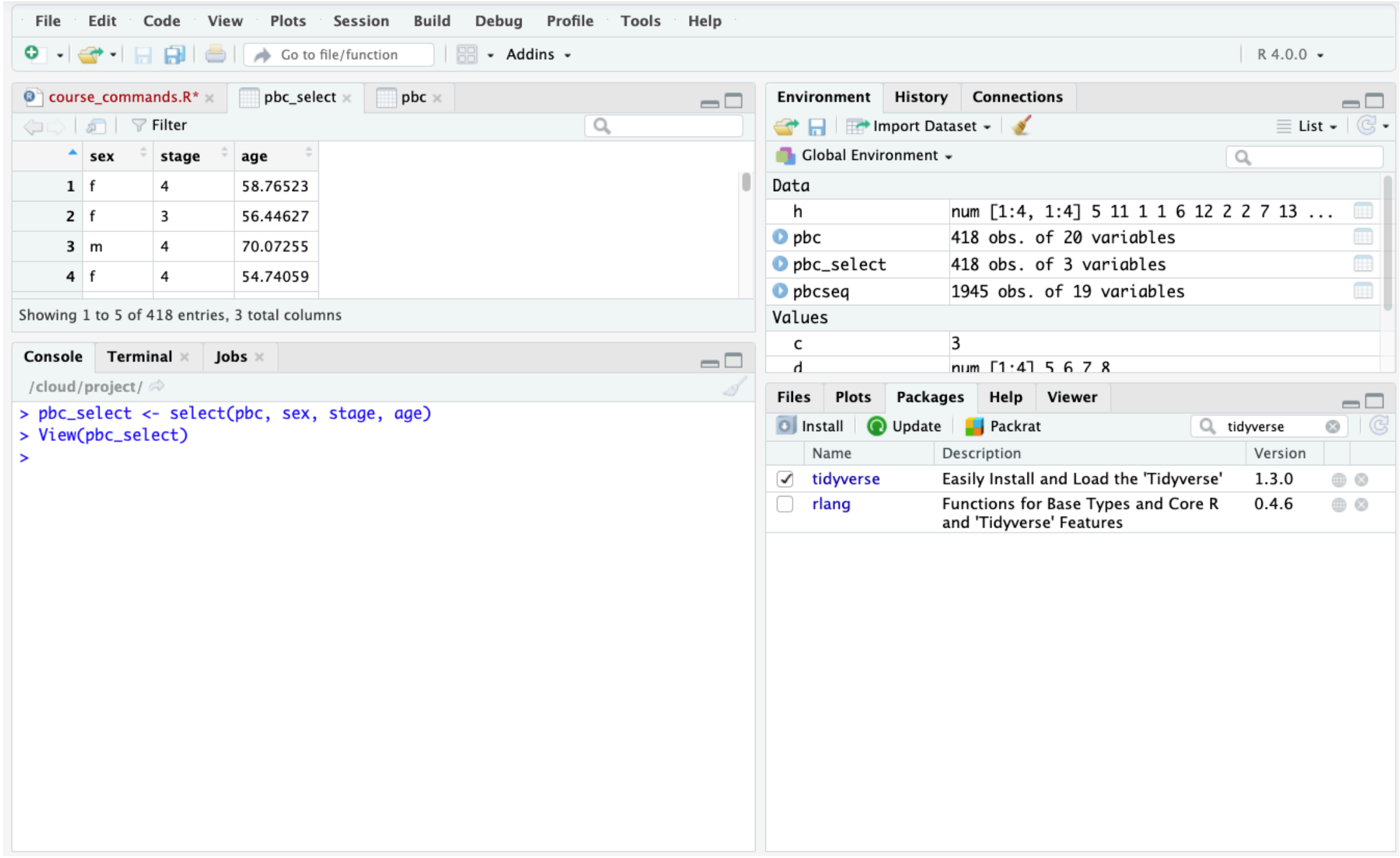
Install | Update | Packrat | tidyverse

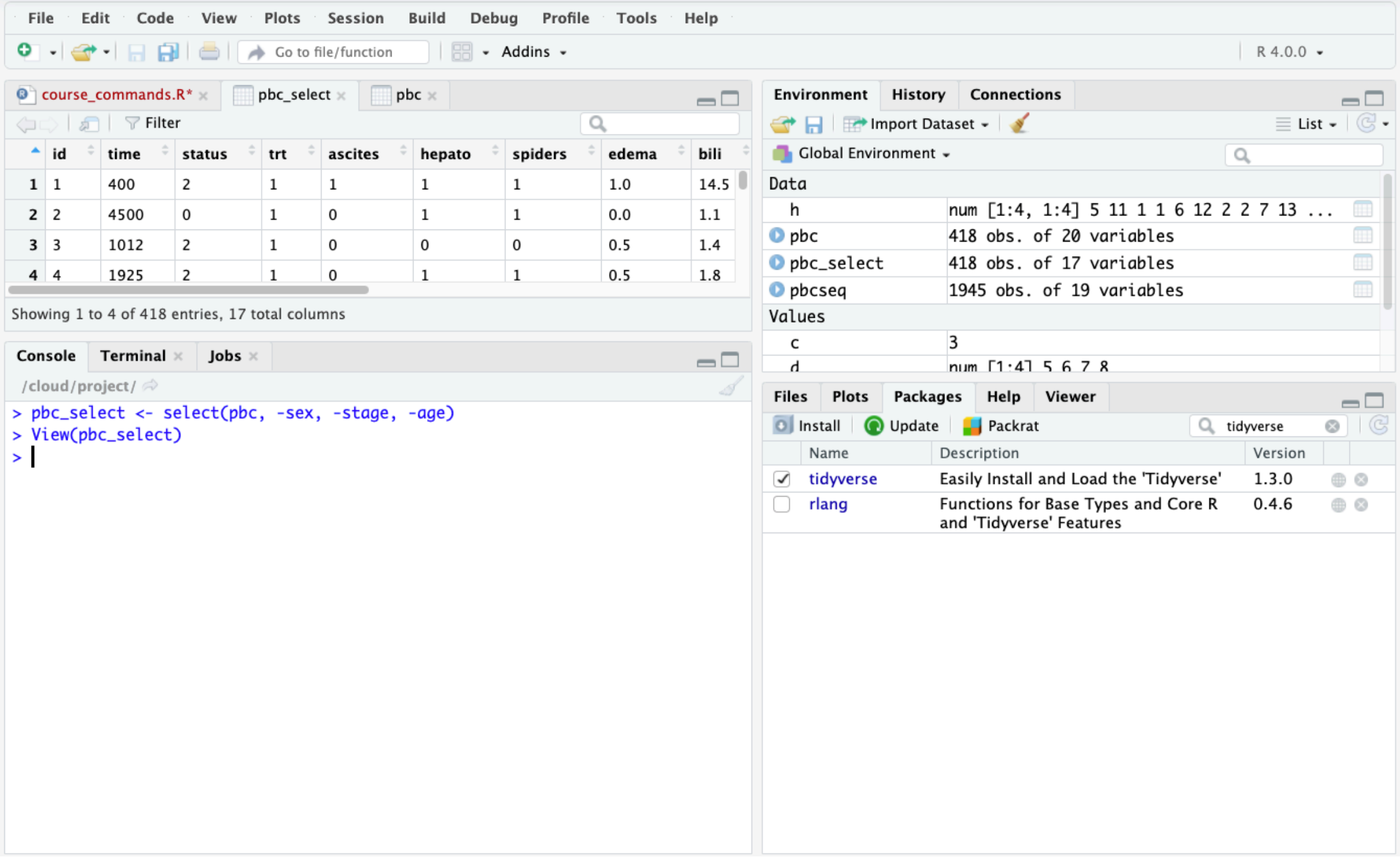
	Name	Description	Version
☑	tidyverse	Easily Install and Load the 'Tidyverse'	1.3.0
☐	rlang	Functions for Base Types and Core R and 'Tidyverse' Features	0.4.6

dplyr: Subset the data

- pbc_select <- select(pbc, sex, stage, age)
- View(pbc_select)
- pbc_select <- select(pbc, -sex, -stage, -age)
- View(pbc_select)

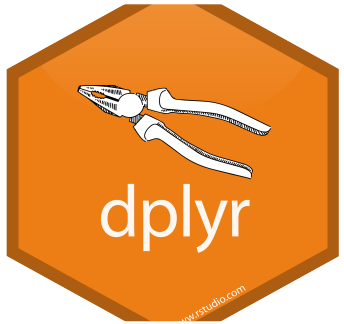


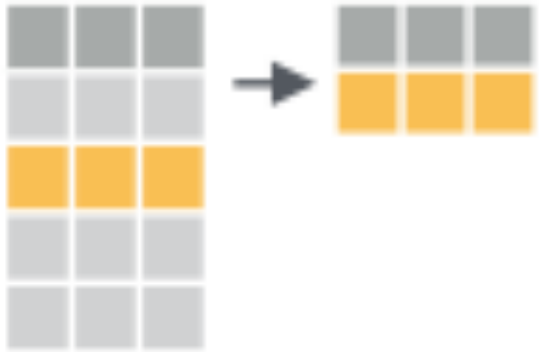

File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 4.0.0
course_commands.R*
pbc_select
pbc
Filter
sex stage age
1 f 4 58.76523
2 f 3 56.44627
3 m 4 70.07255
4 f 4 54.74059
Showing 1 to 5 of 418 entries, 3 total columns
Console Terminal Jobs
/cloud/project/
> pbc_select <- select(pbc, sex, stage, age)
> View(pbc_select)
>
Environment History Connections
Import Dataset
List
Global Environment
Data
h num [1:4, 1:4] 5 11 1 1 6 12 2 2 7 13 ...
pbc 418 obs. of 20 variables
pbc_select 418 obs. of 3 variables
pbcseq 1945 obs. of 19 variables
Values
c 3
Files Plots Packages Help Viewer
Install Update Packrat
tidyverse
Name Description Version
tidyverse Easily Install and Load the 'Tidyverse' 1.3.0
rlang Functions for Base Types and Core R and 'Tidyverse' Features 0.4.6

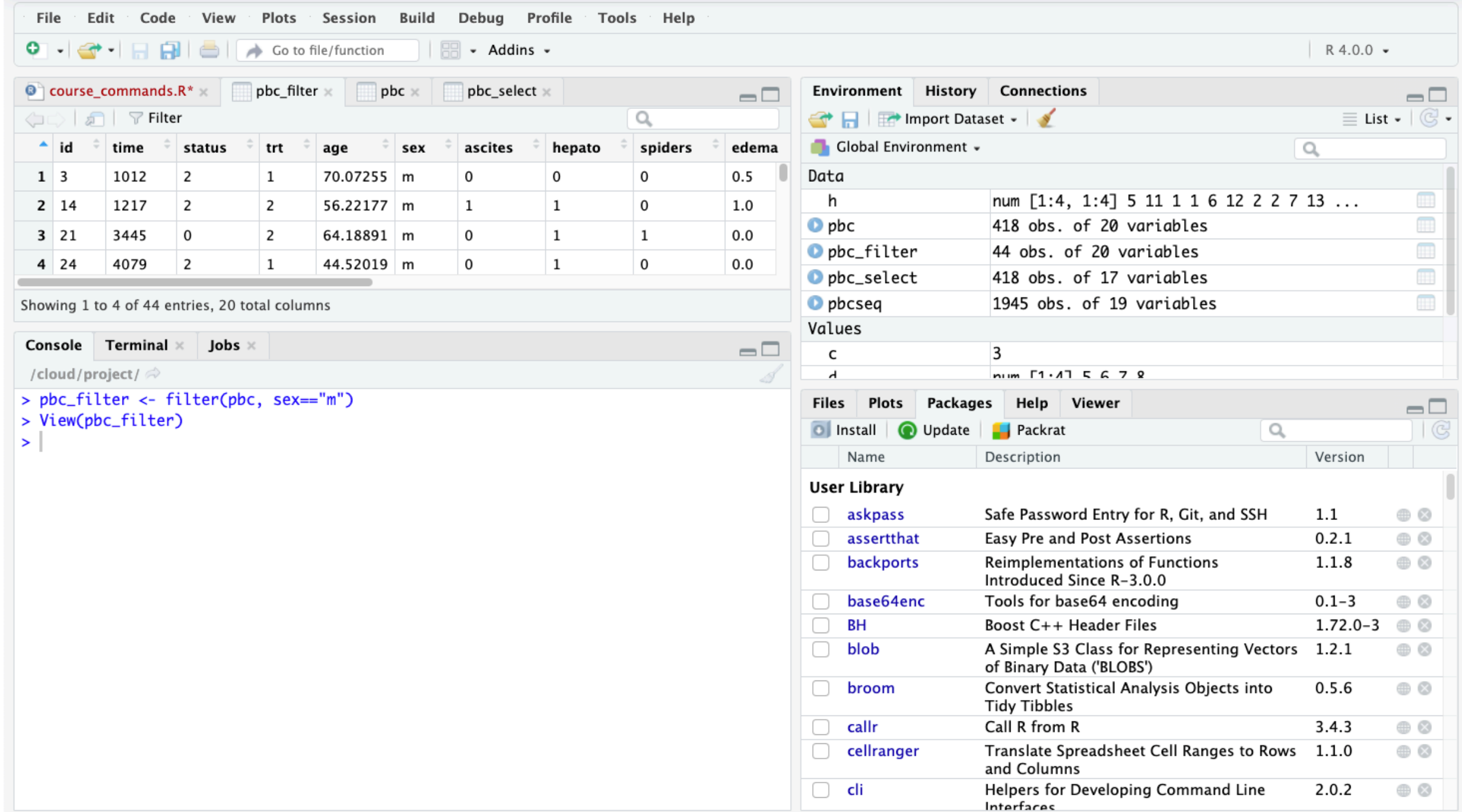

File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 4.0.0
course_commands.R*
pbc_select
pbc
Filter
id time status trt ascites hepato spiders edema bili
1 1 400 2 1 1 1 1 1.0 14.5
2 2 4500 0 1 0 1 1 0.0 1.1
3 3 1012 2 1 0 0 0 0.5 1.4
4 4 1925 2 1 0 1 1 0.5 1.8
Showing 1 to 4 of 418 entries, 17 total columns
Console Terminal Jobs
/cloud/project/
> pbc_select <- select(pbc, -sex, -stage, -age)
> View(pbc_select)
>
Environment History Connections
Import Dataset
List
Global Environment
Data
h num [1:4, 1:4] 5 11 1 1 6 12 2 2 7 13 ...
pbc 418 obs. of 20 variables
pbc_select 418 obs. of 17 variables
pbcseq 1945 obs. of 19 variables
Values
c 3
d num [1:4] 5 6 7 8
Files Plots Packages Help Viewer
Install Update Packrat
tidyverse
Name Description Version
tidyverse Easily Install and Load the 'Tidyverse' 1.3.0
rlang Functions for Base Types and Core R and 'Tidyverse' Features 0.4.6

dplyr:Filter the data

- pbc_filter <- filter(pbc, sex=="m")
- View(pbc_filter)
- pbc_filter <- filter(pbc, age > 70)
- View(pbc_filter)






File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 4.0.0
course_commands.R*
pbc_filter
pbc
pbc_select
Filter
id time status trt age sex ascites hepato spiders edema
1 3 1012 2 1 70.07255 m 0 0 0 0.5
2 14 1217 2 2 56.22177 m 1 1 0 1.0
3 21 3445 0 2 64.18891 m 0 1 1 0.0
4 24 4079 2 1 44.52019 m 0 1 0 0.0
Showing 1 to 4 of 44 entries, 20 total columns
Console
Terminal
Jobs
/cloud/project/
> pbc_filter <- filter(pbc, sex=="m")
> View(pbc_filter)
>
Environment
History
Connections
Import Dataset
List
Global Environment
Data
h num [1:4, 1:4] 5 11 1 1 6 12 2 2 7 13 ...
pbc 418 obs. of 20 variables
pbc_filter 44 obs. of 20 variables
pbc_select 418 obs. of 17 variables
pbcseq 1945 obs. of 19 variables
Values
c 3
Files
Plots
Packages
Help
Viewer
Install
Update
Packrat
Name
Description
Version
User Library
askpass Safe Password Entry for R, Git, and SSH 1.1
assertthat Easy Pre and Post Assertions 0.2.1
backports Reimplementations of Functions Introduced Since R-3.0.0 1.1.8
base64enc Tools for base64 encoding 0.1-3
BH Boost C++ Header Files 1.72.0-3
blob A Simple S3 Class for Representing Vectors of Binary Data ('BLOBS') 1.2.1
broom Convert Statistical Analysis Objects into Tidy Tibbles 0.5.6
callr Call R from R 3.4.3
cellranger Translate Spreadsheet Cell Ranges to Rows and Columns 1.1.0
cli Helpers for Developing Command Line Interfaces 2.0.2

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | R 4.0.0

course_commands.R* | pbc_filter | pbc | pbc_select

Filter

	id	time	status	trt	age	sex	ascites	hepato	spiders	edema
1	3	1012	2	1	70.07255	m	0	0	0	0.5
2	10	51	2	2	70.55989	f	1	0	1	1.0
3	92	388	2	1	76.70910	f	1	0	0	1.0
4	97	611	2	2	71.89322	m	0	1	0	0.5

Showing 1 to 4 of 13 entries, 20 total columns

Console | Terminal | Jobs

/cloud/project/

```
> pbc_filter <- filter(pbc, age > 70)
> View(pbc_filter)
>
```

Environment | History | Connections

Import Dataset | List

Global Environment

Data	
h	num [1:4, 1:4] 5 11 1 1 6 12 2 2 7 13 ...
pbc	418 obs. of 20 variables
pbc_filter	13 obs. of 20 variables
pbc_select	418 obs. of 17 variables
pbcseq	1945 obs. of 19 variables
Values	
c	3
d	num [1:4] 5 6 7 8

Files | Plots | Packages | Help | Viewer

Install | Update | Packrat

	Name	Description	Version
User Library			
☐	askpass	Safe Password Entry for R, Git, and SSH	1.1
☐	assertthat	Easy Pre and Post Assertions	0.2.1
☐	backports	Reimplementations of Functions Introduced Since R-3.0.0	1.1.8
☐	base64enc	Tools for base64 encoding	0.1-3
☐	BH	Boost C++ Header Files	1.72.0-3
☐	blob	A Simple S3 Class for Representing Vectors of Binary Data ('BLOBS')	1.2.1
☐	broom	Convert Statistical Analysis Objects into Tidy Tibbles	0.5.6
☐	callr	Call R from R	3.4.3
☐	cellranger	Translate Spreadsheet Cell Ranges to Rows and Columns	1.1.0
☐	cli	Helpers for Developing Command Line Interfaces	2.0.2

dplyr:Manipulate a variable

- pbc_mutate <- mutate(pbc, new_val=ast/1000)
- View(pbc_mutate)

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | R 4.0.0

course_commands.R* | pbc_mutate

Filter

ema	bili	chol	albumin	copper	alk.phos	ast	trig	platelet	protime	stage	new_val
)	14.5	261	2.60	156	1718.0	137.95	172	190	12.2	4	0.13795
)	1.1	302	4.14	54	7394.8	113.52	88	221	10.6	3	0.11352
;	1.4	176	3.48	210	516.0	96.10	55	151	12.0	4	0.09610
;	1.8	244	2.54	64	6121.8	60.63	92	183	10.3	4	0.06063
)	3.4	279	3.53	143	671.0	113.15	72	136	10.9	3	0.11315
)	0.8	248	3.98	50	944.0	93.00	63	*NA*	11.0	3	0.09300
)	1.0	322	4.09	52	824.0	60.45	213	204	9.7	3	0.06045
)	0.3	280	4.00	52	4651.2	28.38	189	373	11.0	3	0.02838
)	3.2	562	3.08	79	2276.0	144.15	88	251	11.0	2	0.14415
)	12.6	200	2.74	140	918.0	147.25	143	302	11.5	4	0.14725
)	1.4	259	4.16	46	1104.0	79.05	79	258	12.0	4	0.07905
)	3.6	236	3.52	94	591.0	82.15	95	71	13.6	4	0.08215

Showing 1 to 13 of 418 entries, 21 total columns

Console | Terminal | Jobs

/cloud/project/

```
> pbc_mutate <- mutate(pbc, new_val=ast/1000)
> View(pbc_mutate)
>
```

Environment | History | Connections

Import Dataset | List

Global Environment

pbc_mutate	418 obs. of 21 variables
pbcseq	1945 obs. of 19 variables

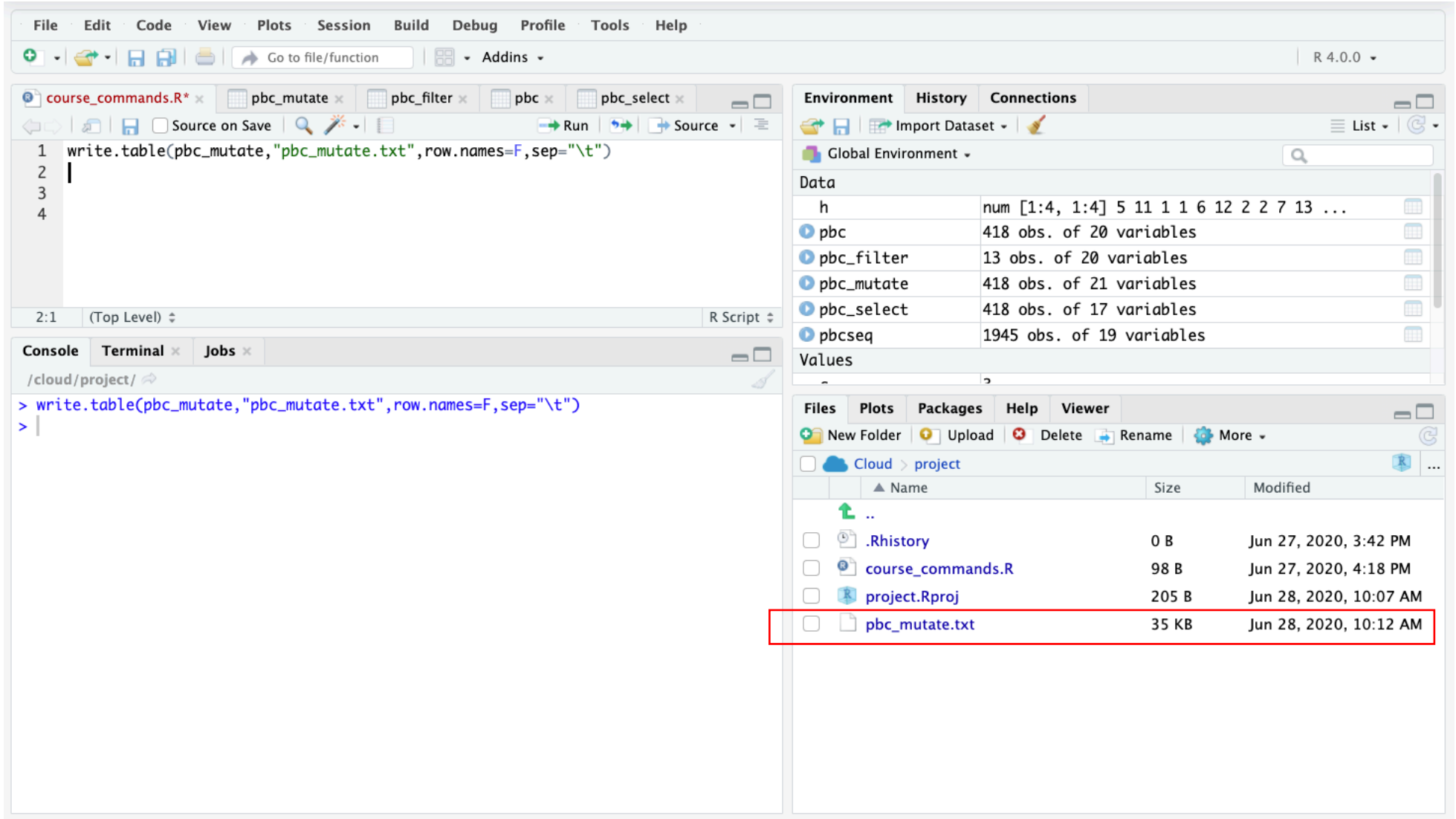
Files | Plots | Packages | Help | Viewer

Install | Update | Packrat

Name	Description	Version
User Library		
askpass	Safe Password Entry for R, Git, and SSH	1.1
assertthat	Easy Pre and Post Assertions	0.2.1
backports	Reimplementations of Functions Introduced Since R-3.0.0	1.1.8
base64enc	Tools for base64 encoding	0.1-3
BH	Boost C++ Header Files	1.72.0-3
blob	A Simple S3 Class for Representing Vectors of Binary Data ('BLOBS')	1.2.1
broom	Convert Statistical Analysis Objects into Tidy Tibbles	0.5.6
callr	Call R from R	3.4.3
cellranger	Translate Spreadsheet Cell Ranges to Rows and Columns	1.1.0
cli	Helpers for Developing Command Line Interfaces	2.0.2
clipr	Read and Write from the System Clipboard	0.7.0
colorspace	A Toolbox for Manipulating and Assessing Colors and Palettes	1.4-1
commonmark	High Performance CommonMark and Github Markdown Rendering in R	1.7
crayon	Colored Terminal Output	1.3.4
crosstalk	Inter-Widget Interactivity for HTML Widgets	1.1.0.1
curl	A Modern and Flexible Web Client for R	4.3
data.table	Extension of `data.frame`	1.12.8
DBI	R Database Interface	1.1.0
dbplyr	A 'dplyr' Back End for Databases	1.4.4
desc	Manipulate DESCRIPTION Files	1.2.0
digest	Create Compact Hash Digests of R Objects	0.6.25

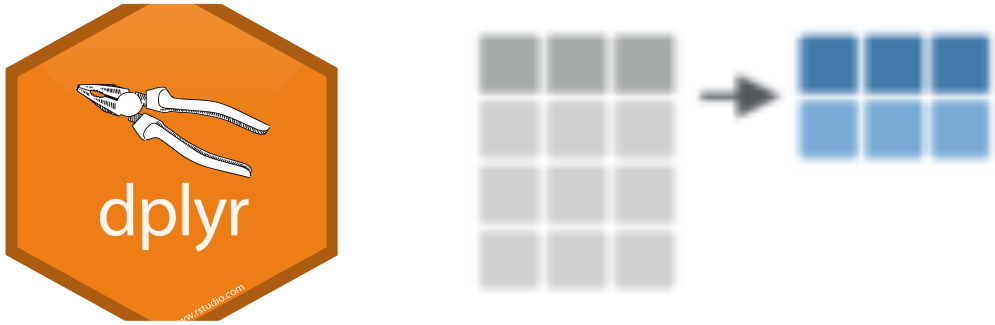
Save a table as a file

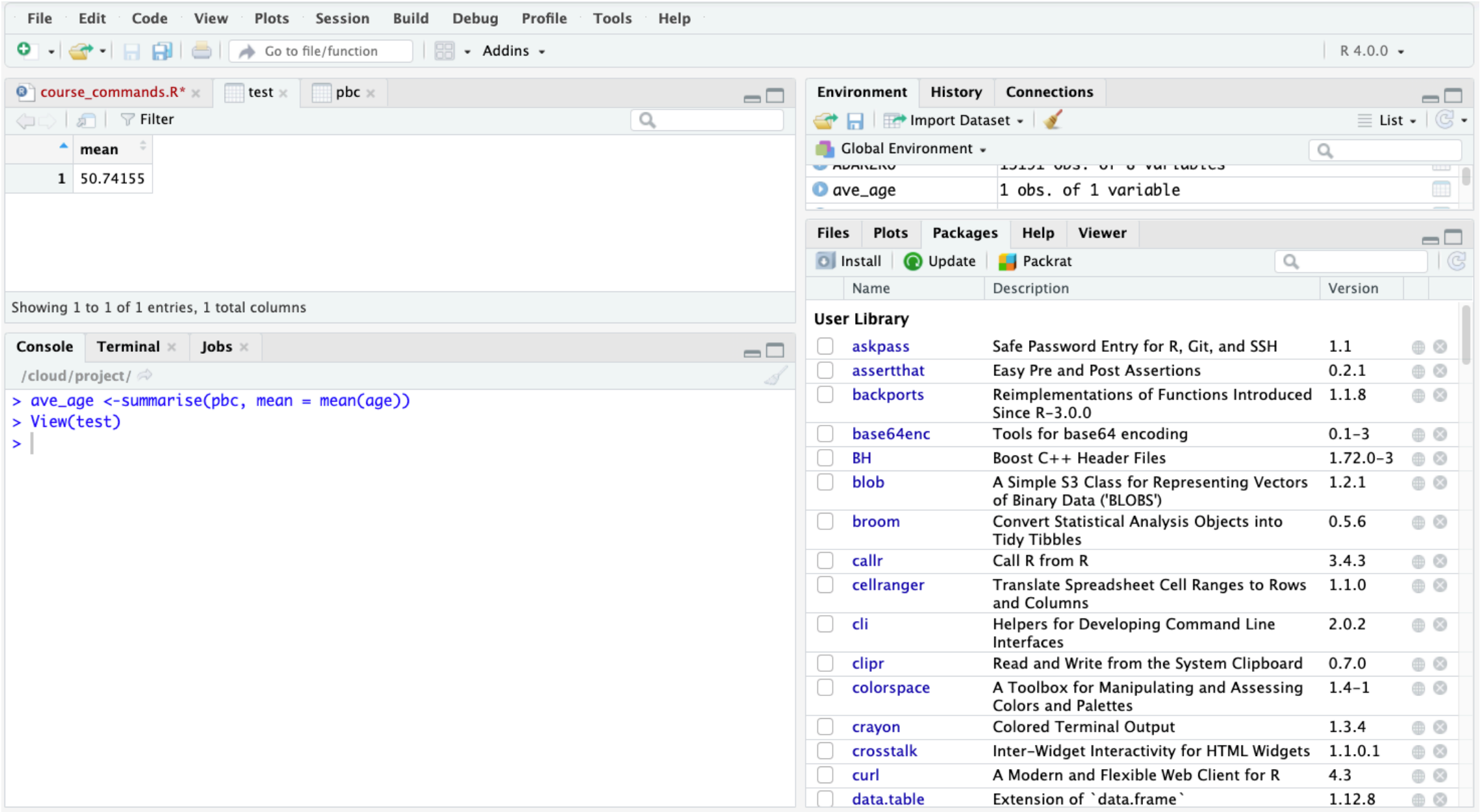
write.table(pbc_mutate,"pbc_mutate.txt",row.names=F,sep="\t")



dplyr: Summarize a variable

- ave_age <-summarise(pbc, mean = mean(age))
- View(ave_age)

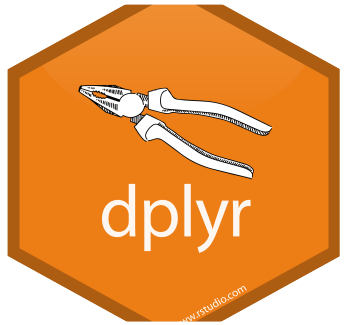


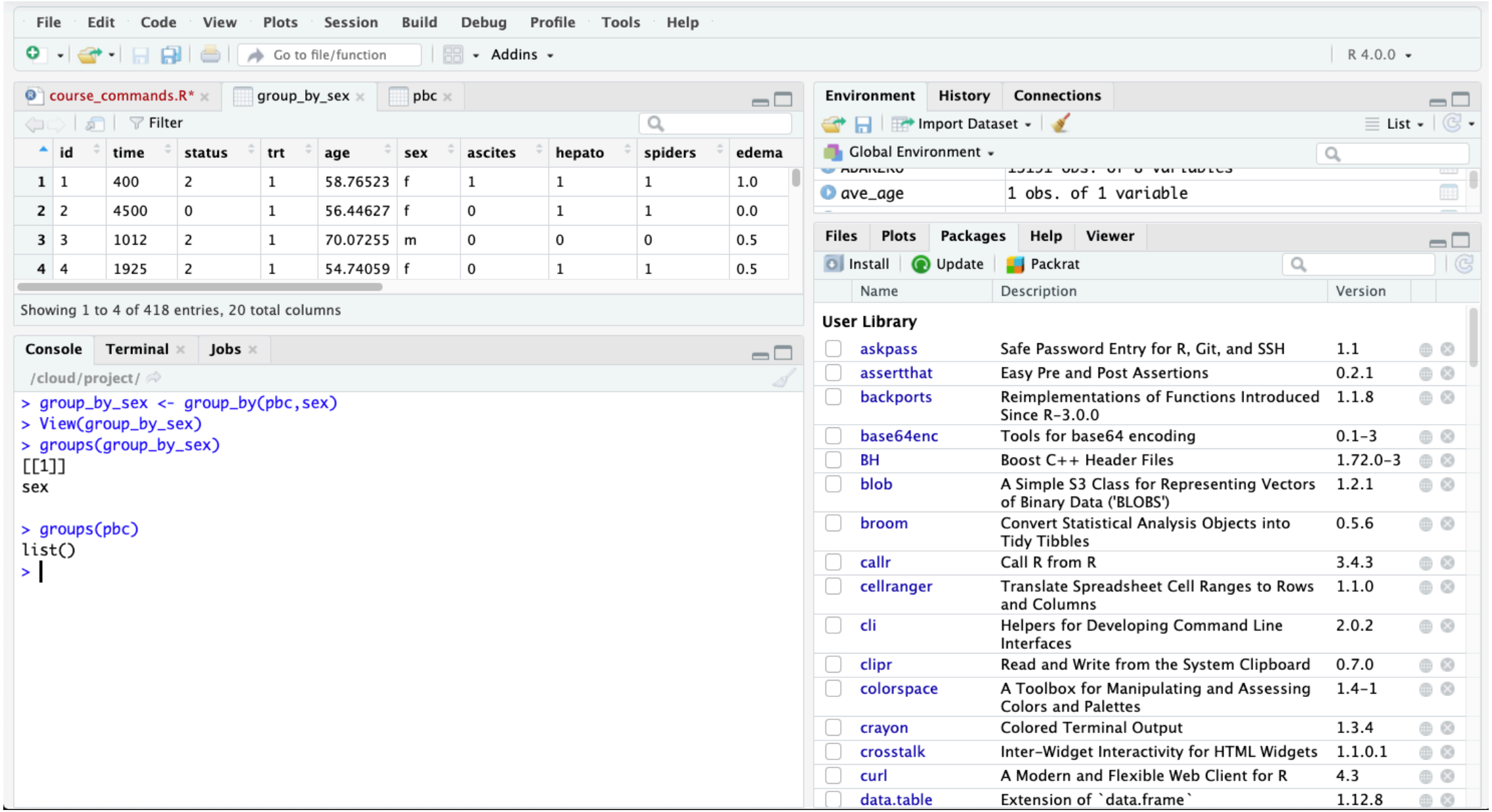

File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 4.0.0
course_commands.R*
test
pbc
Filter
mean
1 50.74155
Showing 1 to 1 of 1 entries, 1 total columns
Console
Terminal
Jobs
/cloud/project/
> ave_age <-summarise(pbc, mean = mean(age))
> View(test)
>
Environment
History
Connections
Import Dataset
List
Global Environment
ave_age
1 obs. of 1 variable
Files
Plots
Packages
Help
Viewer
Install
Update
Packrat
Name
Description
Version
User Library
askpass Safe Password Entry for R, Git, and SSH 1.1
assertthat Easy Pre and Post Assertions 0.2.1
backports Reimplementations of Functions Introduced Since R-3.0.0 1.1.8
base64enc Tools for base64 encoding 0.1-3
BH Boost C++ Header Files 1.72.0-3
blob A Simple S3 Class for Representing Vectors of Binary Data ('BLOBS') 1.2.1
broom Convert Statistical Analysis Objects into Tidy Tibbles 0.5.6
callr Call R from R 3.4.3
cellranger Translate Spreadsheet Cell Ranges to Rows and Columns 1.1.0
cli Helpers for Developing Command Line Interfaces 2.0.2
clipr Read and Write from the System Clipboard 0.7.0
colorspace A Toolbox for Manipulating and Assessing Colors and Palettes 1.4-1
crayon Colored Terminal Output 1.3.4
crosstalk Inter-Widget Interactivity for HTML Widgets 1.1.0.1
curl A Modern and Flexible Web Client for R 4.3
data.table Extension of `data.frame` 1.12.8

dplyr: Group by a variable

- group_by_sex <- group_by(pbc,sex)
- View(group_by_sex)
- Nothing changed with the data
- However there is now the data is now “ordered”

- groups(group_by_sex)
- groups(pbc)




File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 4.0.0
course_commands.R*
group_by_sex
pbc
Filter
id time status trt age sex ascites hepato spiders edema
1 1 400 2 1 58.76523 f 1 1 1 1.0
2 2 4500 0 1 56.44627 f 0 1 1 0.0
3 3 1012 2 1 70.07255 m 0 0 0 0.5
4 4 1925 2 1 54.74059 f 0 1 1 0.5
Showing 1 to 4 of 418 entries, 20 total columns
Console
Terminal
Jobs
/cloud/project/
> group_by_sex <- group_by(pbc,sex)
> View(group_by_sex)
> groups(group_by_sex)
[[1]]
sex
> groups(pbc)
list()
>
Environment
History
Connections
Import Dataset
List
Global Environment
ave_age
1 obs. of 1 variable
Files
Plots
Packages
Help
Viewer
Install
Update
Packrat
Name
Description
Version
User Library
askpass Safe Password Entry for R, Git, and SSH 1.1
assertthat Easy Pre and Post Assertions 0.2.1
backports Reimplementations of Functions Introduced Since R-3.0.0 1.1.8
base64enc Tools for base64 encoding 0.1-3
BH Boost C++ Header Files 1.72.0-3
blob A Simple S3 Class for Representing Vectors of Binary Data ('BLOBS') 1.2.1
broom Convert Statistical Analysis Objects into Tidy Tibbles 0.5.6
callr Call R from R 3.4.3
cellranger Translate Spreadsheet Cell Ranges to Rows and Columns 1.1.0
cli Helpers for Developing Command Line Interfaces 2.0.2
clipr Read and Write from the System Clipboard 0.7.0
colorspace A Toolbox for Manipulating and Assessing Colors and Palettes 1.4-1
crayon Colored Terminal Output 1.3.4
crosstalk Inter-Widget Interactivity for HTML Widgets 1.1.0.1
curl A Modern and Flexible Web Client for R 4.3
data.table Extension of `data.frame` 1.12.8

Pipes!

Let's say I want to know the average age for males versus females I could:

```
group_by_sex <- group_by(pbc,sex)
ave_age_sex <-summarise(group_by_sex, mean = mean(age))
```

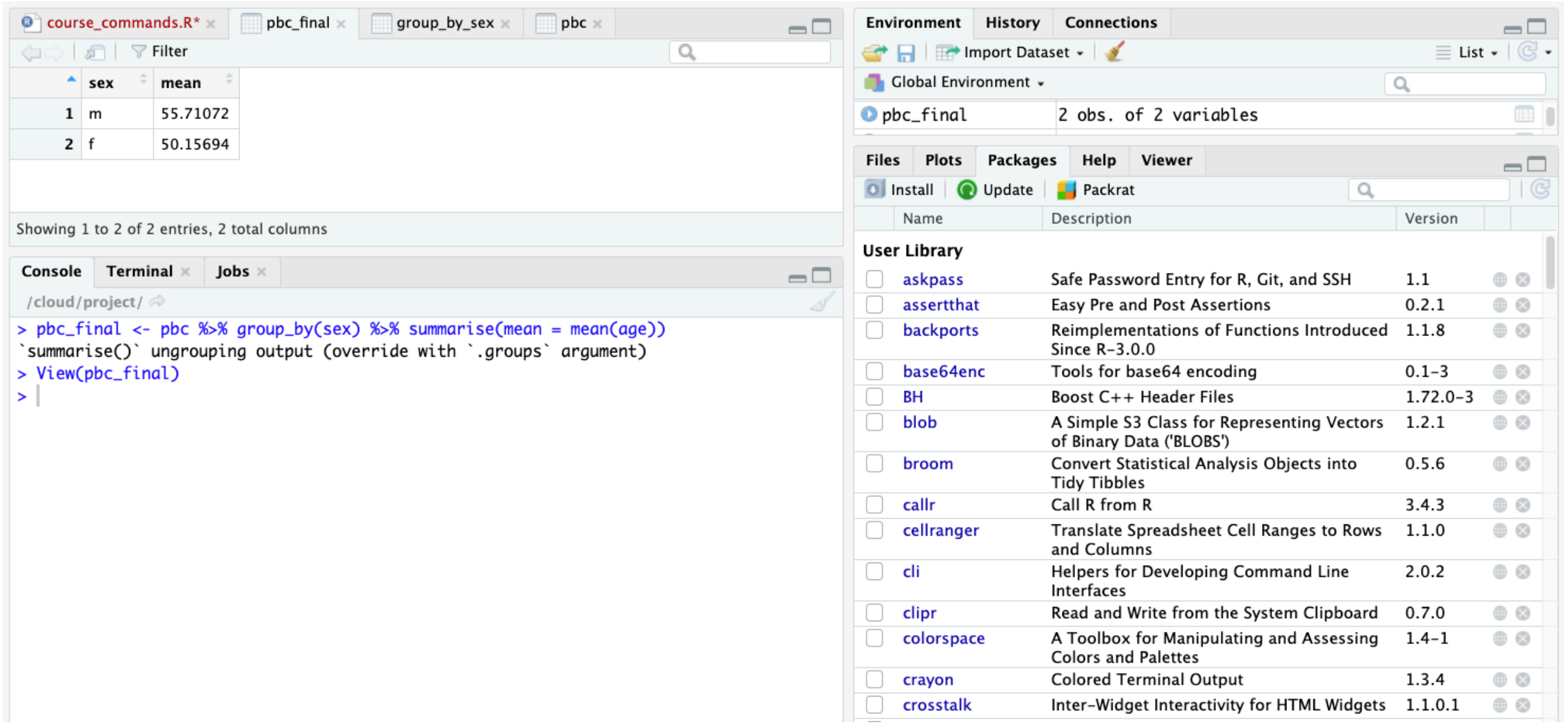
OR

Use a pipe! %>%

Keyboard Shortcut:
PC:Ctrl+Shift+M
Mac:Cmd+Shift+M

This allows sequential operations to be done on the same dataset:

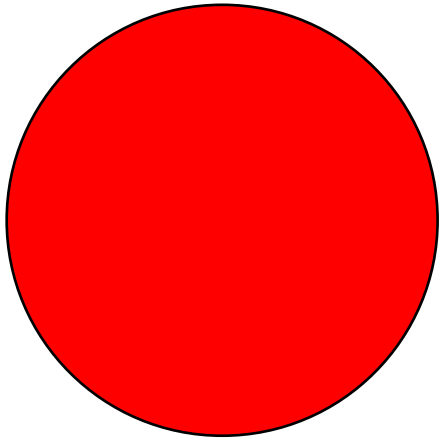
```
pbc_final <- pbc %>% group_by(sex) %>% summarise(mean = mean(age))
View(pbc_final)
```


course_commands.R*
pbc_final
group_by_sex
pbc
Filter
sex
mean
1 m 55.71072
2 f 50.15694
Showing 1 to 2 of 2 entries, 2 total columns
Console
Terminal
Jobs
/cloud/project/
> pbc_final <- pbc %>% group_by(sex) %>% summarise(mean = mean(age))
`summarise()` ungrouping output (override with `.groups` argument)
> View(pbc_final)
>
Environment
History
Connections
Import Dataset
List
Global Environment
pbc_final
2 obs. of 2 variables
Files
Plots
Packages
Help
Viewer
Install
Update
Packrat
Name
Description
Version
User Library
askpass Safe Password Entry for R, Git, and SSH 1.1
assertthat Easy Pre and Post Assertions 0.2.1
backports Reimplementations of Functions Introduced Since R-3.0.0 1.1.8
base64enc Tools for base64 encoding 0.1-3
BH Boost C++ Header Files 1.72.0-3
blob A Simple S3 Class for Representing Vectors of Binary Data ('BLOBS') 1.2.1
broom Convert Statistical Analysis Objects into Tidy Tibbles 0.5.6
callr Call R from R 3.4.3
cellranger Translate Spreadsheet Cell Ranges to Rows and Columns 1.1.0
cli Helpers for Developing Command Line Interfaces 2.0.2
clipr Read and Write from the System Clipboard 0.7.0
colorspace A Toolbox for Manipulating and Assessing Colors and Palettes 1.4-1
crayon Colored Terminal Output 1.3.4
crosstalk Inter-Widget Interactivity for HTML Widgets 1.1.0.1

Test Time!

- Find the number deceased patients
- Find the average age of the deceased patients
- Find the average age of male versus female deceased patients

10 minutes



Data visualization in R

Visualization with ggplot2 (tidyverse!)

- Easy out of box formatting
- Handles complex data quickly
- Default options are aesthetically pleasing
- Layering system = add complexity as you go
- Automatic scaling generally works well
- Great documentation and support

ggplot2

What you need:

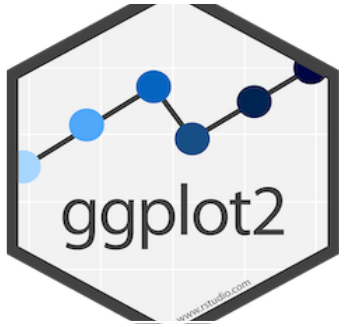
1. A data.frame object

2. Aesthetic mappings (aes) how variables in the data are assigned to visual properties
    - x- and y-direction
    - shapes, colors, lines

3. A geometry object (geom): the type of plot

Basic structure:

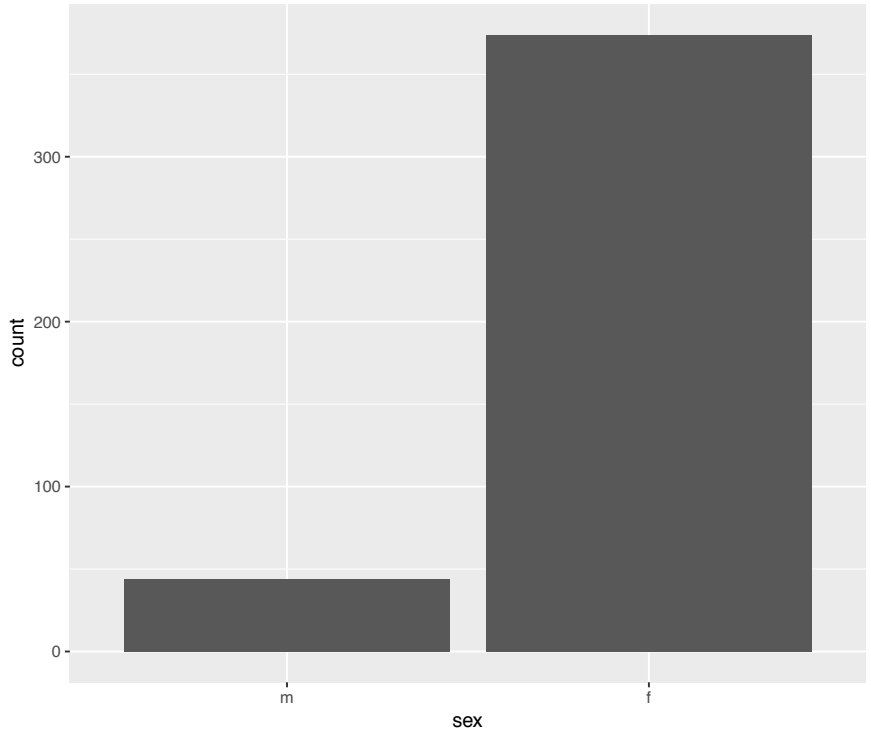
ggplot(data, aes(x=variable)) + geom_type()



Plotting one discrete variable

Bar plot

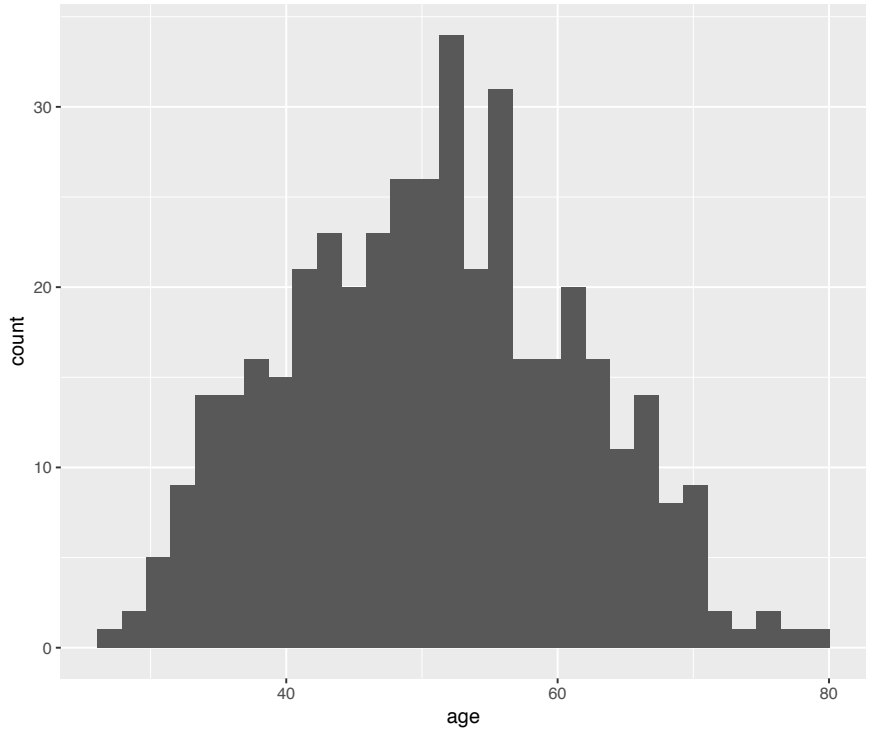
ggplot(pbc, aes(x=sex)) + geom_bar()



Plotting one continous variable

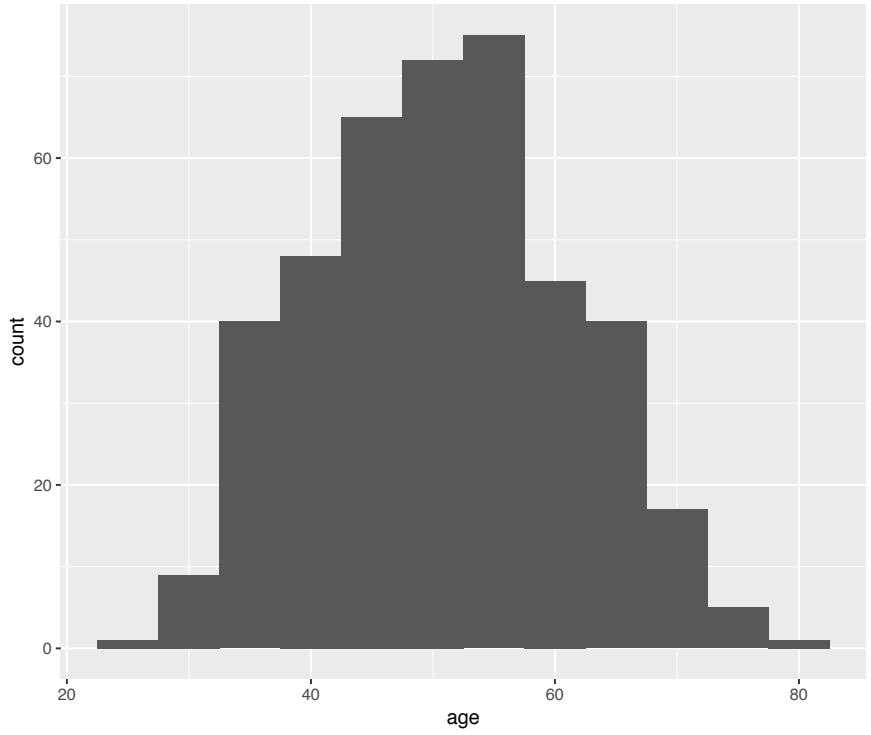
Histogram

- ggplot(pbc, aes(x=age)) + geom_histogram()
- `stat_bin()` using `bins = 30`. Pick better value with `binwidth`



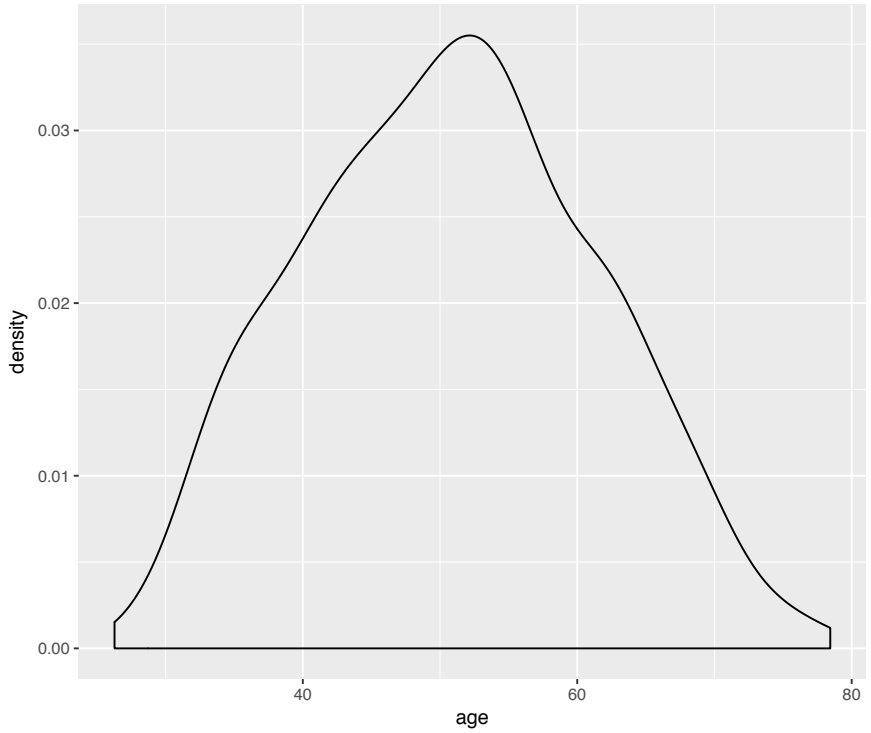
Histogram

- Define better value for binwidth
- ggplot(pbc, aes(x=age)) + geom_histogram(binwidth=5)



Density plot

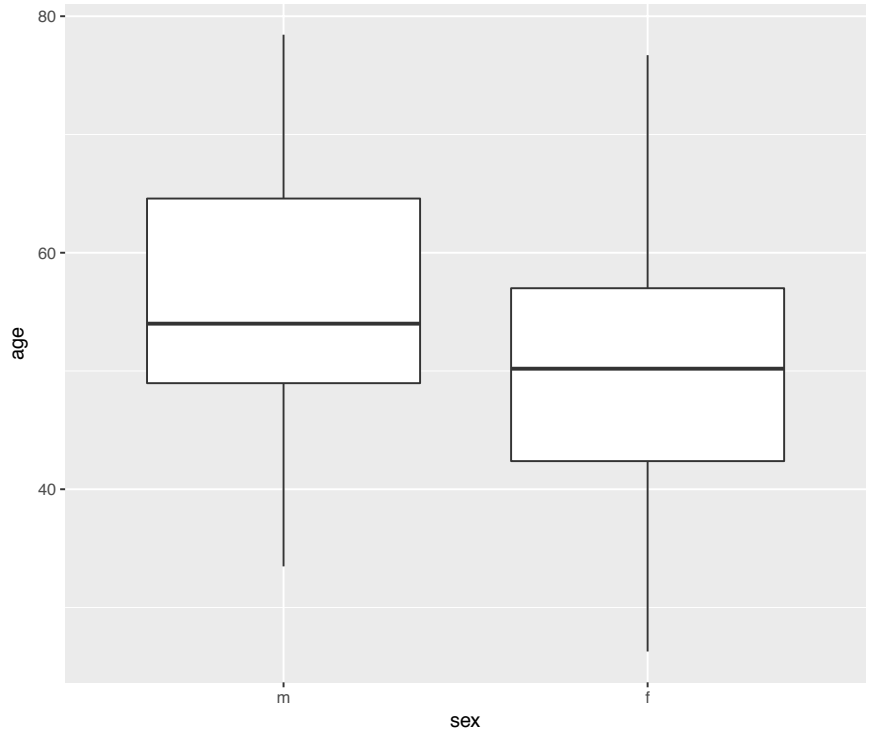
ggplot(pbc, aes(x=age)) + geom_density()



Plotting a continuous variable with a discrete variable

Box plot

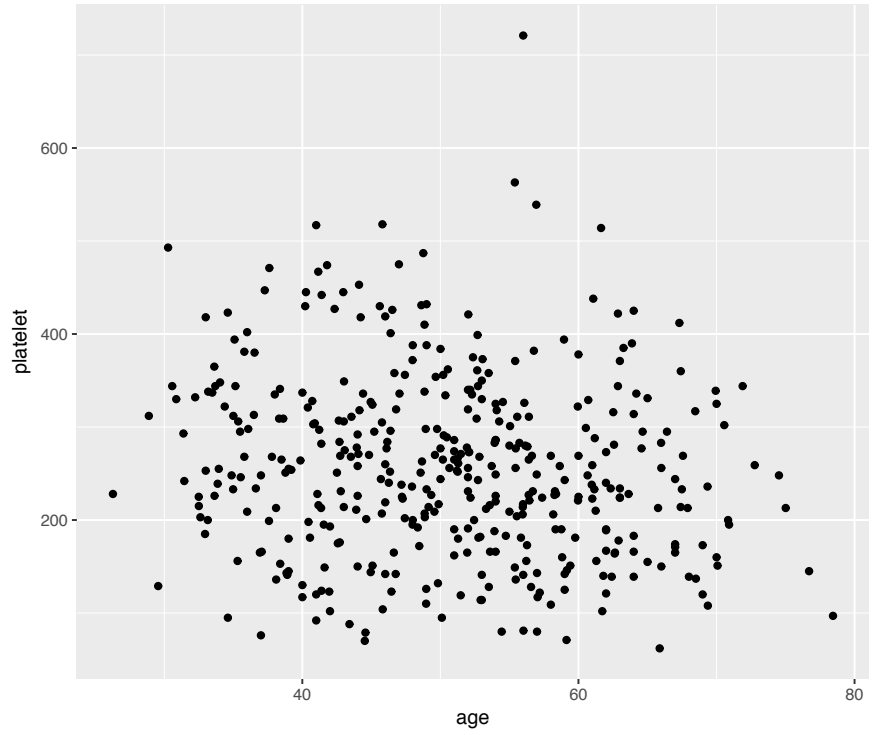
ggplot(pbc, aes(x= sex, y=age)) + geom_boxplot()



Plotting two continuous variables

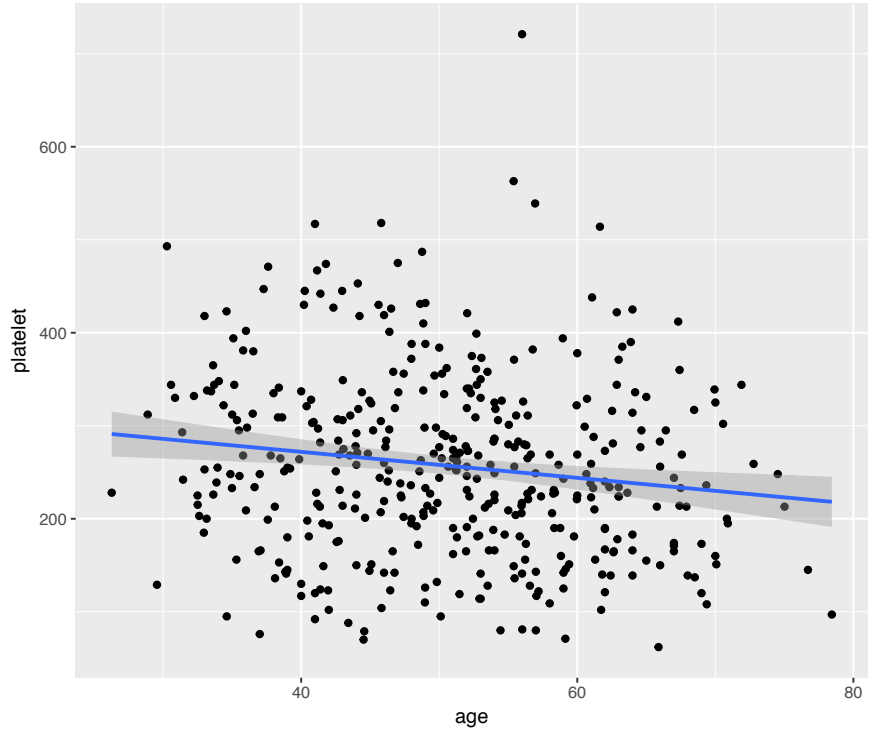
Scatter plot

- ggplot(pbc, aes(x = age, y = platelet)) + geom_point()
- Removed 11 rows containing missing values (geom_point).



Scatter plot with linear regression

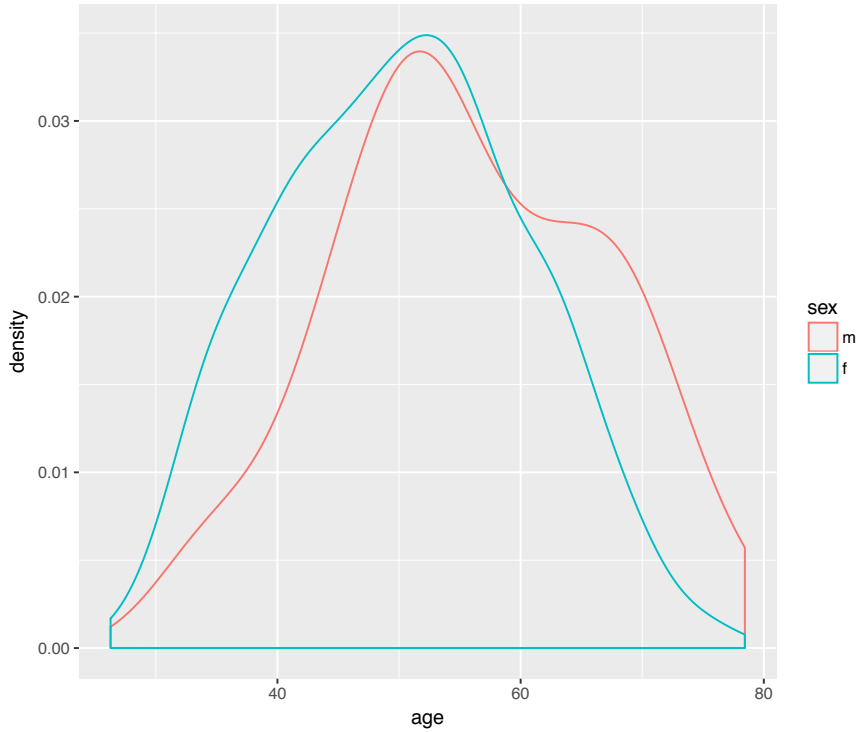
ggplot(pbc, aes(x = age, y = platelet)) + geom_point() + geom_smooth(method=lm)



Grouped visualization

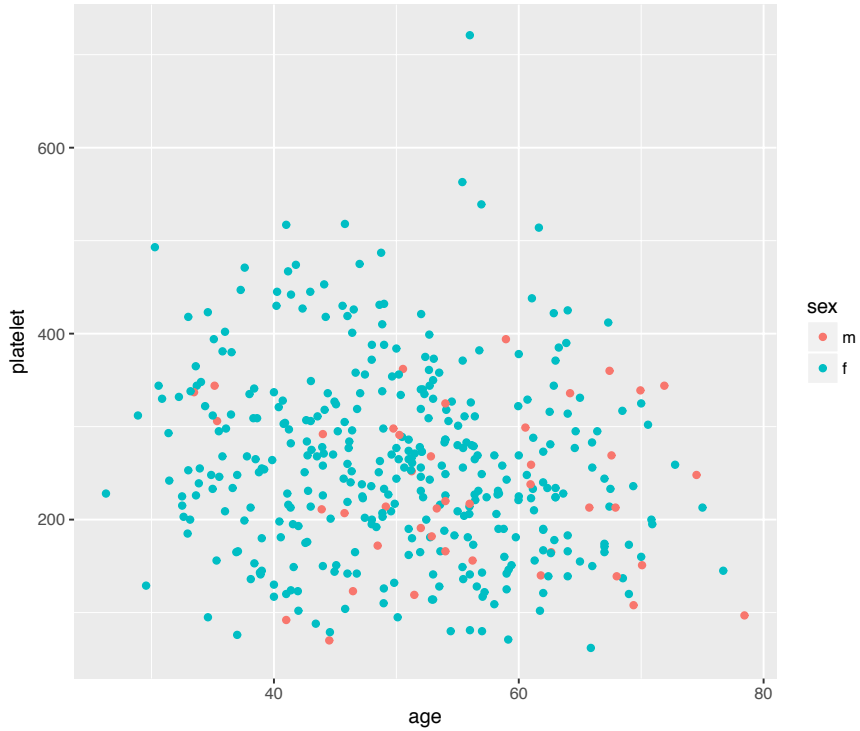
Density plot

ggplot(pbc, aes(x=age, color=sex)) + geom_density()



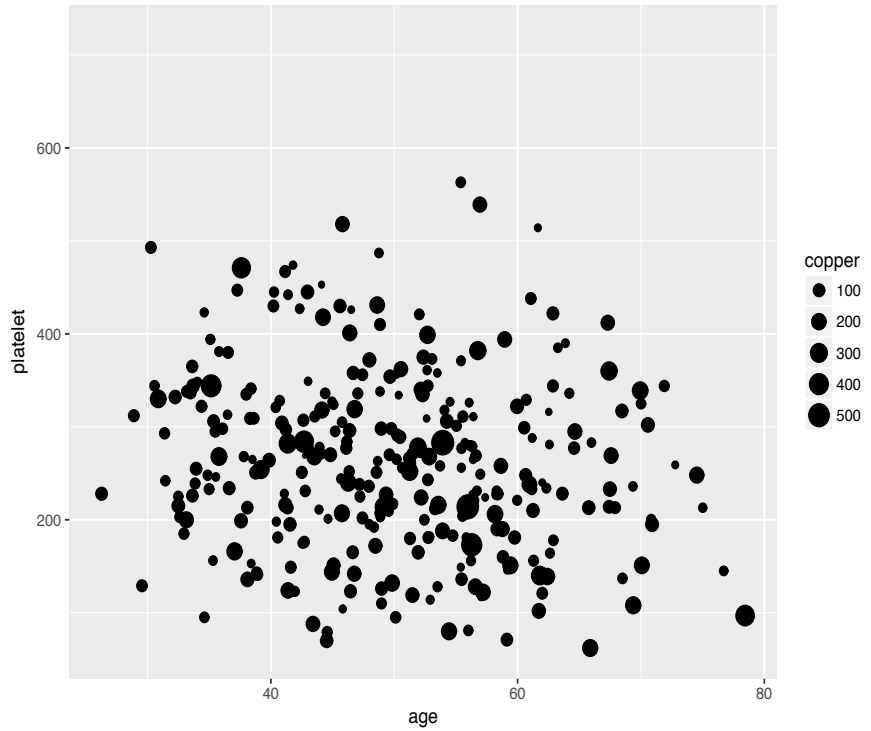
Scatter plot with discrete variable as color

ggplot(pbc, aes(x=age, y=platelet, color=sex)) + geom_point()



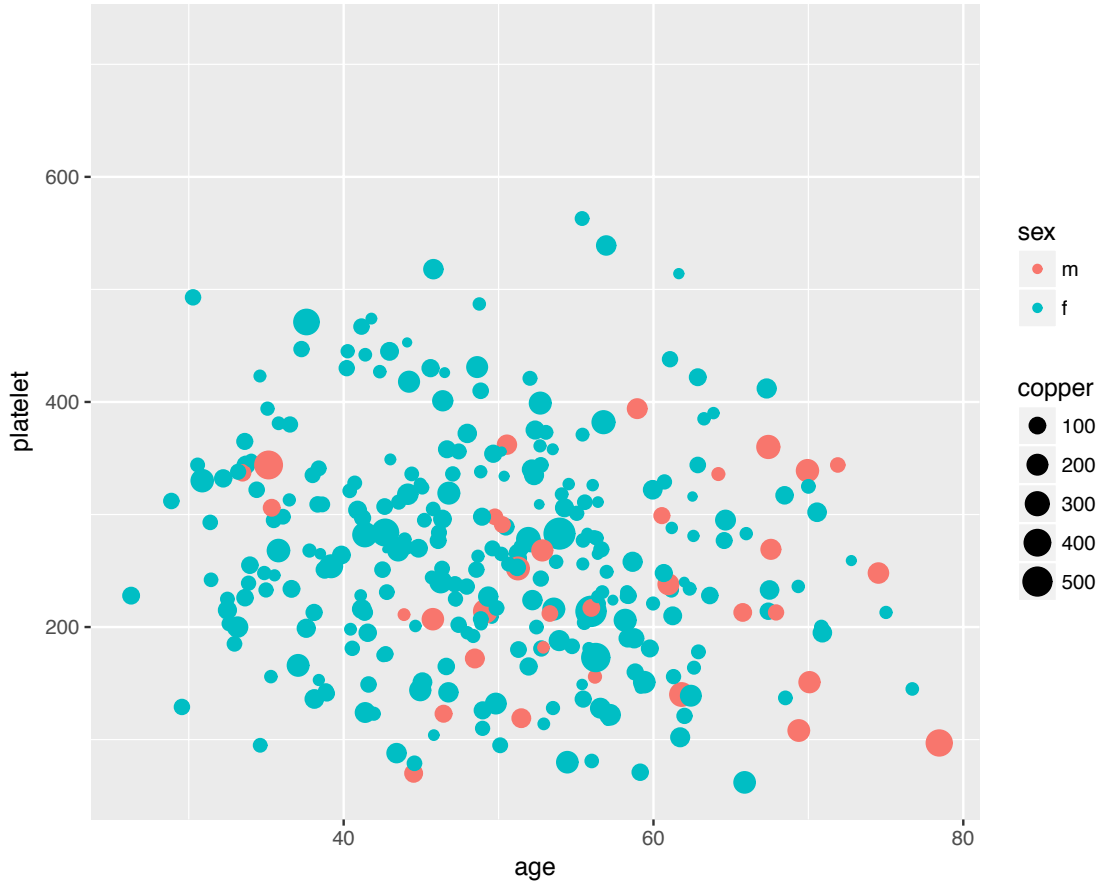
Scatter plot with continuous variable as size

ggplot(pbc, aes(x=age, y=platelet, size=copper)) + geom_point()

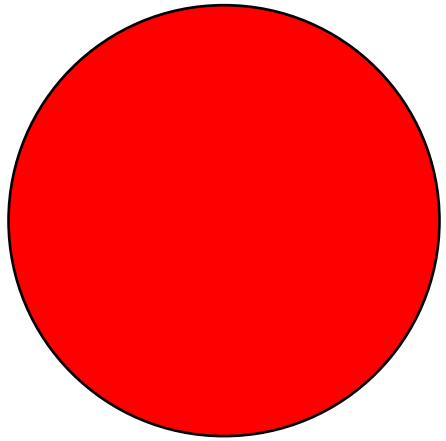


Scatter plot with continuous and discrete extra variables

How would I plot age versus platelet count with points colored by sex and sized by copper values?



2 minutes



The ridiculous scatter plot

```
ggplot(pbc, aes(x=age, y=platelet, color=sex, size=copper, shape=(status)) +
geom_point()
```

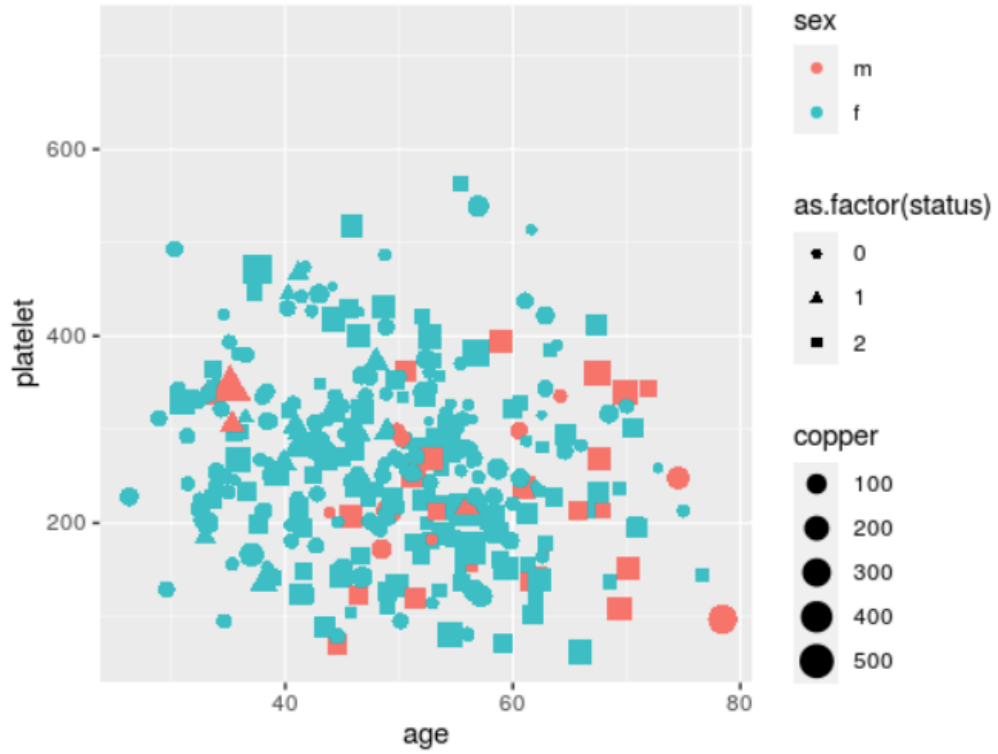
Error: A continuous variable can not be mapped to shape

What happened?

- Status is a discrete variable with the values 0,1,2
- R sees these as numeric values
- We need to convert them using as.factor
- Factors are used to represent categorical data

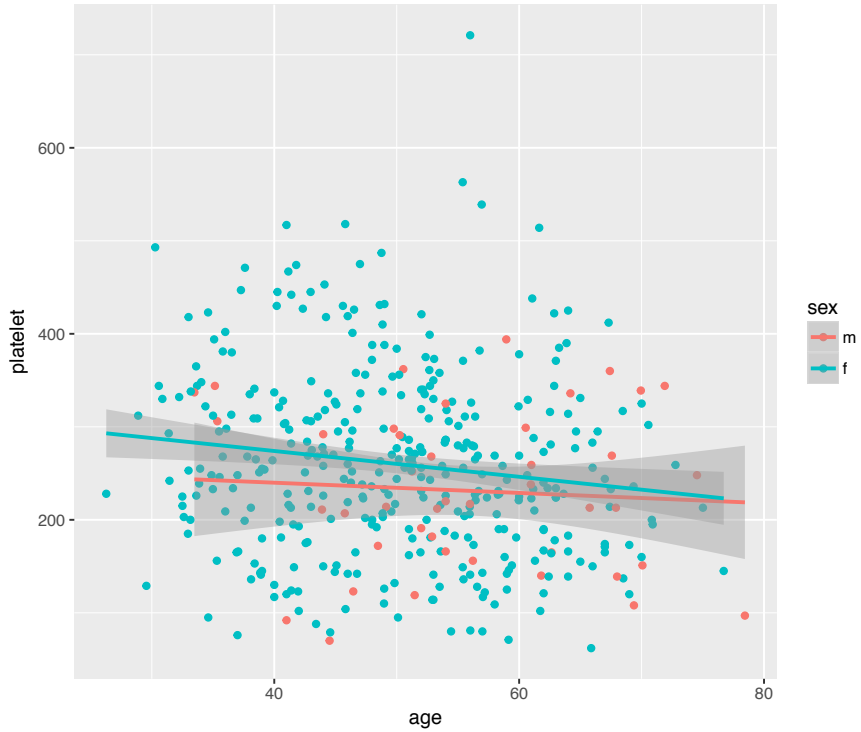
The ridiculous scatter plot

```
ggplot(pbc, aes(x=age, y=platelet, color=sex, size=copper, shape=as.factor(status))) +
geom_point()
```



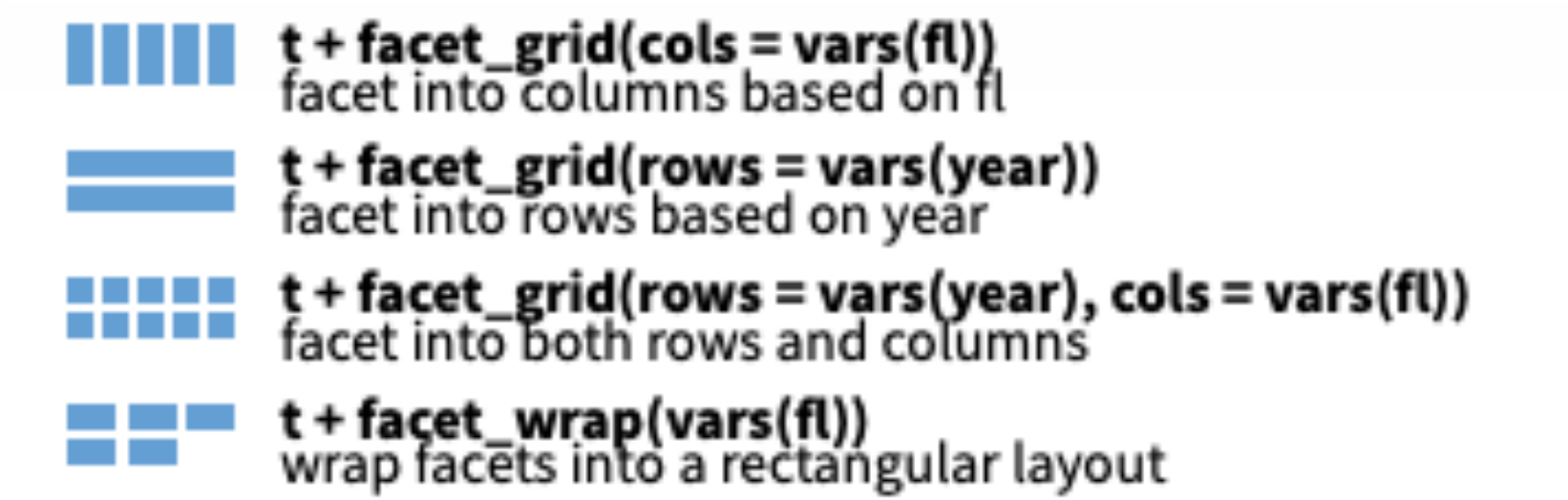
Scatter plot with linear regression

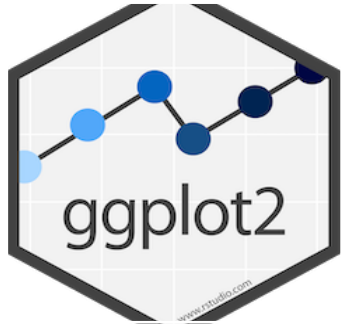
ggplot(pbc, aes(x=age, y=platelet, color=sex)) + geom_point() + geom_smooth(method=lm)



ggplot2:Faceting

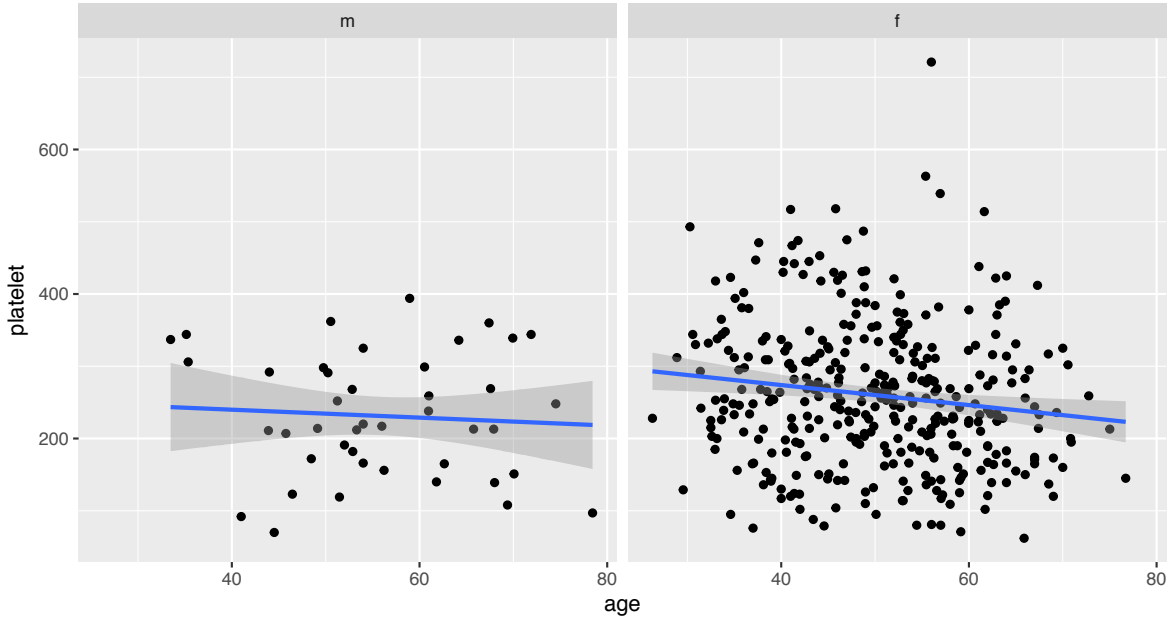
- Divide a plot into subplots based on one or more discrete variable
- Can be used with a variety of plot types
- There are a couple of facet flavors
- We will use facet wrap





Facetted scatter plot with linear regression

```
ggplot(pbc, aes(x=age, y=platelet)) + geom_point() +
geom_smooth(method = lm) + facet_wrap(vars(sex), nrow=1)
```



tidr:gather

- Typically each variable is in its own column and each observation/case is its own row
- Transform data from wide to long format
- **gather(**data, key, value**)**
- Moves column names into a **key** column, gathering the column values into a single **value** column

country	1999	2000
A	0.7K	2K
B	37K	80K
C	212K	213K

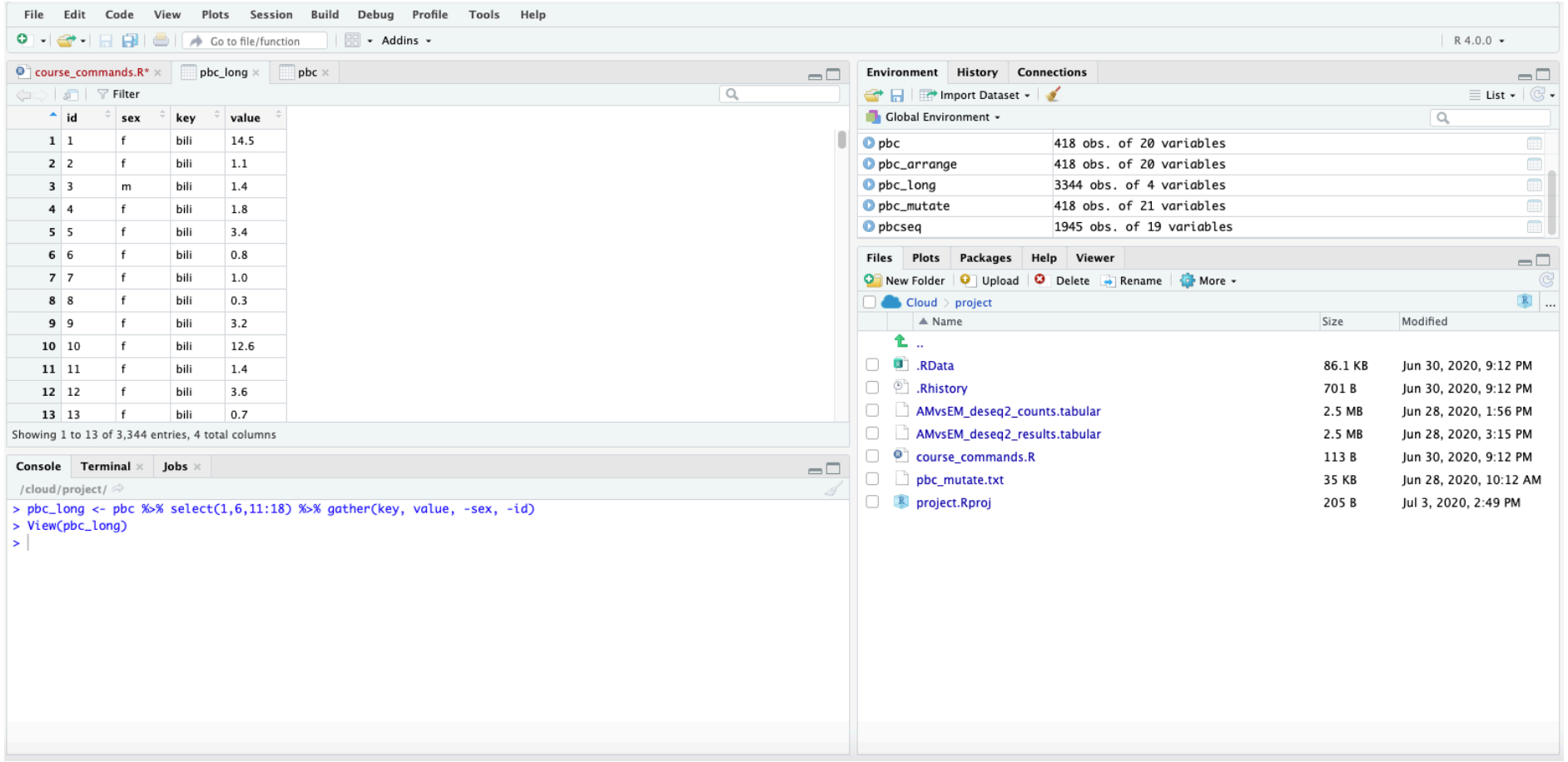
→

country	year	cases
A	1999	0.7K
B	1999	37K
C	1999	212K
A	2000	2K
B	2000	80K
C	2000	213K

key value

Data manipulation

pbc_long <- pbc %>% select(1,6,11:18) %>% gather(key, value, -sex, -id)

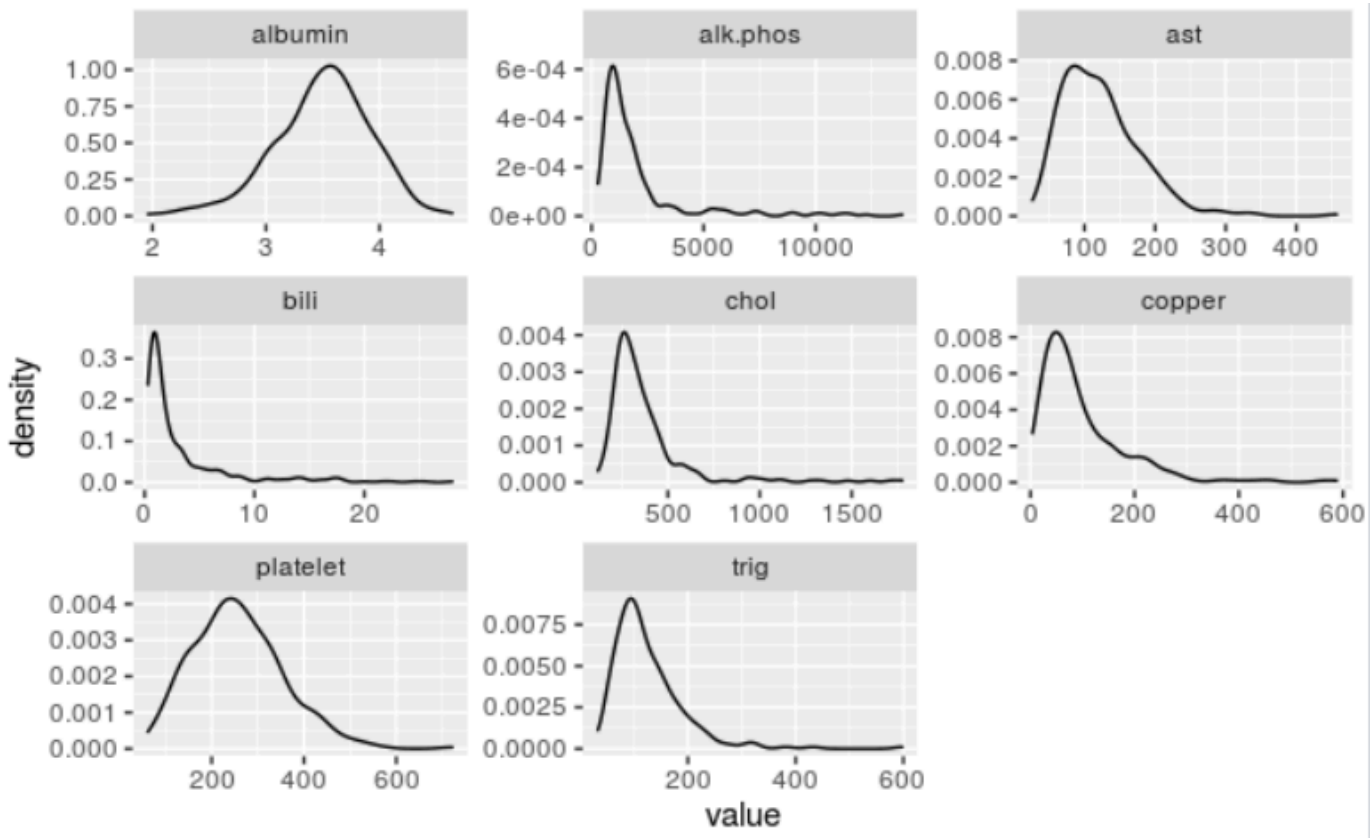


	id	sex	key	value
1	1	f	bili	14.5
2	2	f	bili	1.1
3	3	m	bili	1.4
4	4	f	bili	1.8
5	5	f	bili	3.4
6	6	f	bili	0.8
7	7	f	bili	1.0
8	8	f	bili	0.3
9	9	f	bili	3.2
10	10	f	bili	12.6
11	11	f	bili	1.4
12	12	f	bili	3.6
13	13	f	bili	0.7

The reciprocal command: pbc_wide <- spread(pbc_long, key =key, value = value)

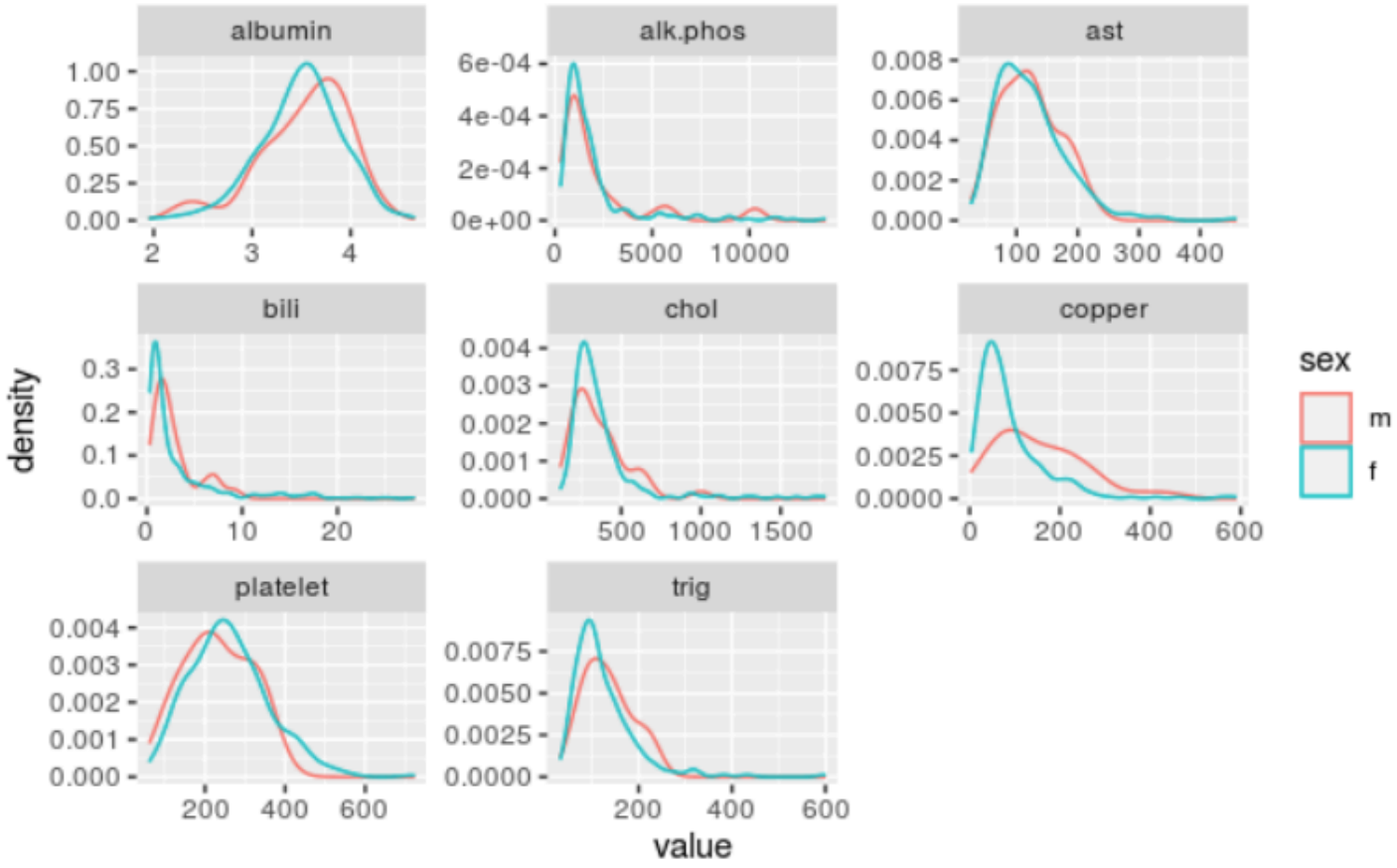
Plotting manipulated data with facet wrap

```
ggplot(pbc_long , aes(value)) + geom_density() +
facet_wrap(vars(key), scales = "free")
```



Adding second variable

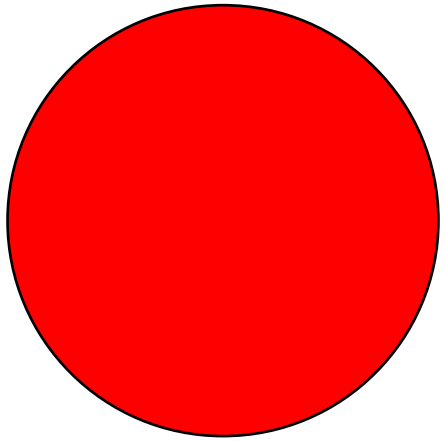
ggplot(pbc_long , aes(value, color=sex)) + geom_density() + facet_wrap(vars(key), scales = "free")



Test Time!

- How would I create similar facetted plots for disease stage rather than sex?

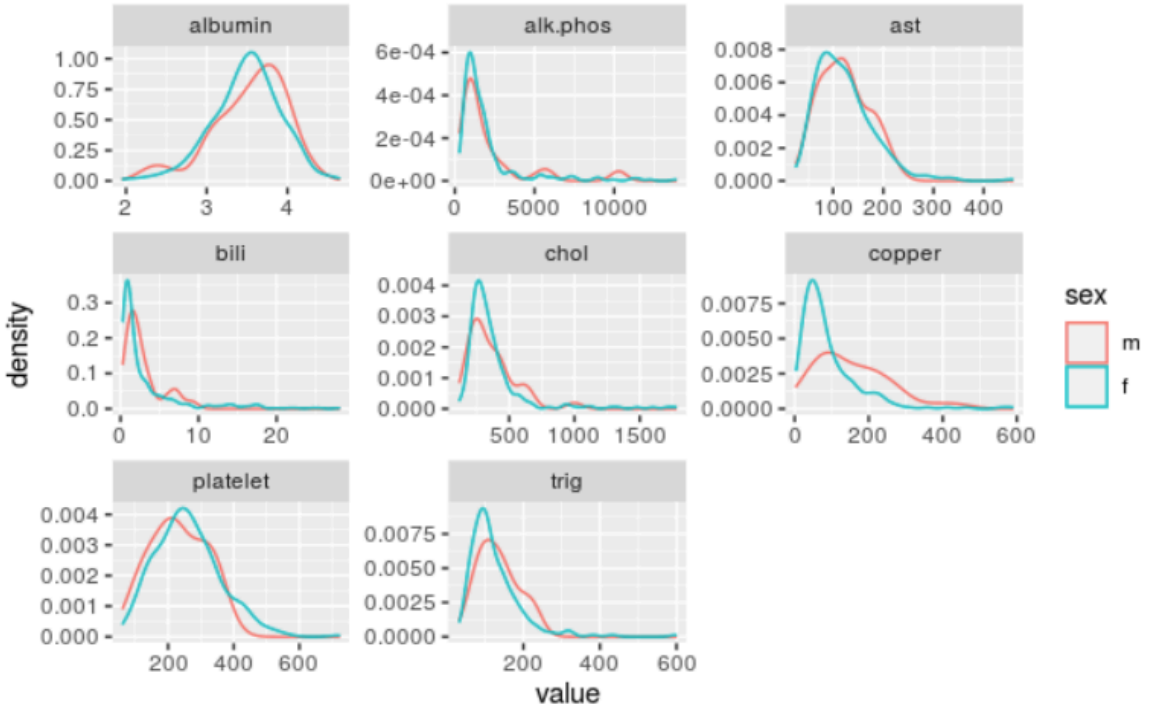
15 minutes



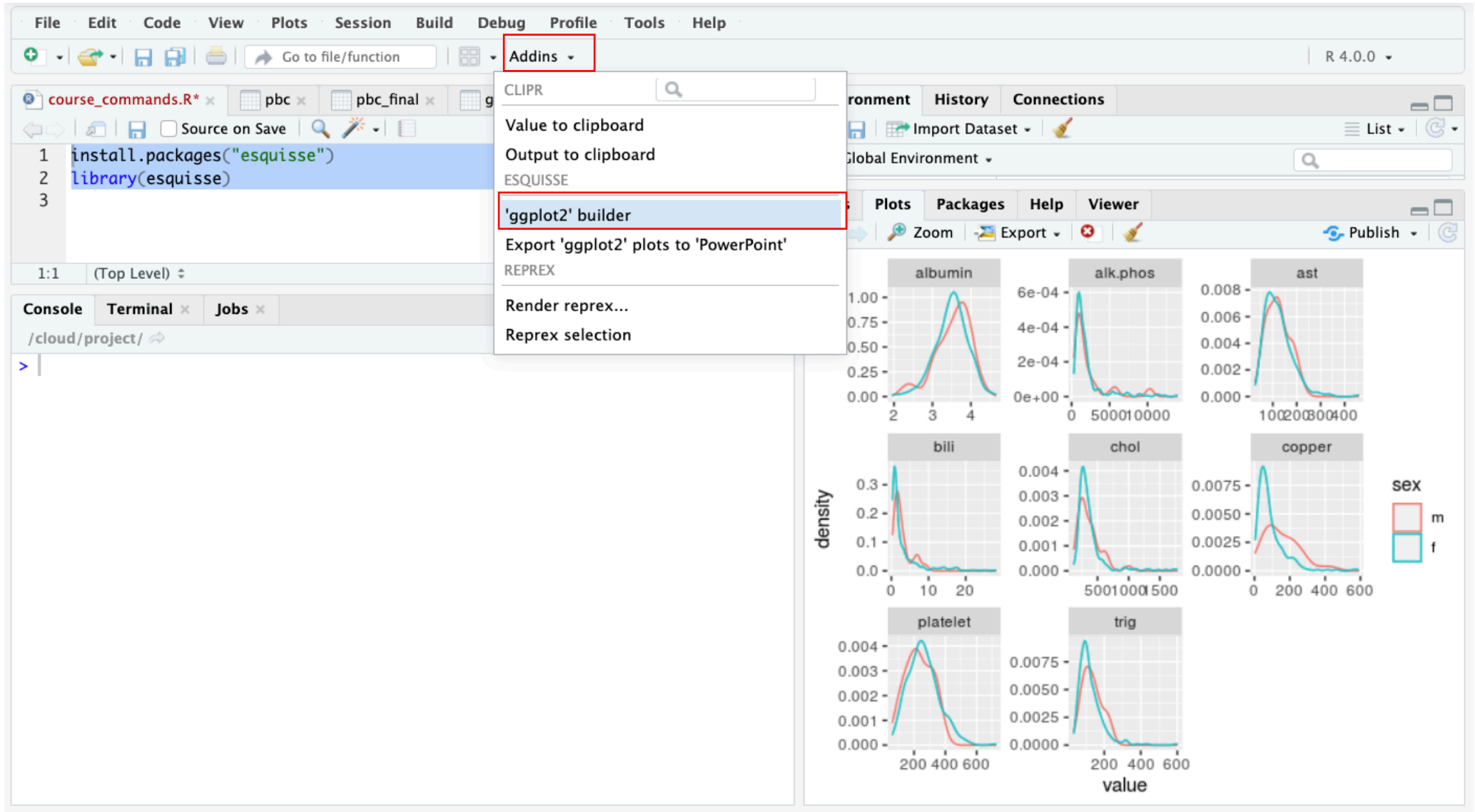
Turn Rstudio into an interactive graphic program!

Package:Esquisse

- French for “sketch”
- install.packages("esquisse")
- library(esquisse)
- Click on “Addins”
- Select “ggplot2 builder”
- Let’s recreate this graph!



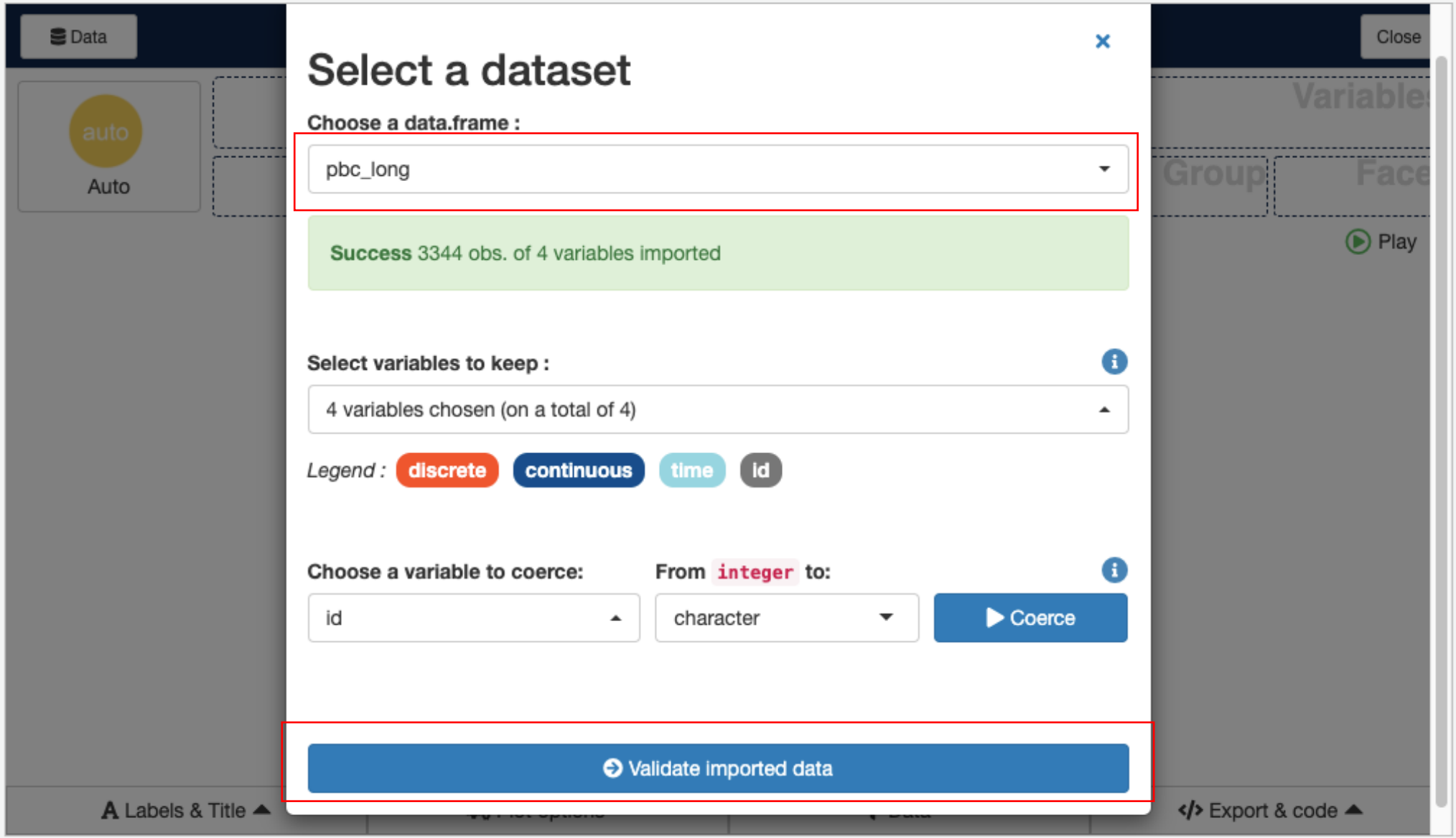

```
ggplot(pbc_long , aes(value, color=sex)) +
geom_density() + facet_wrap(vars(key),
scales = "free")
```

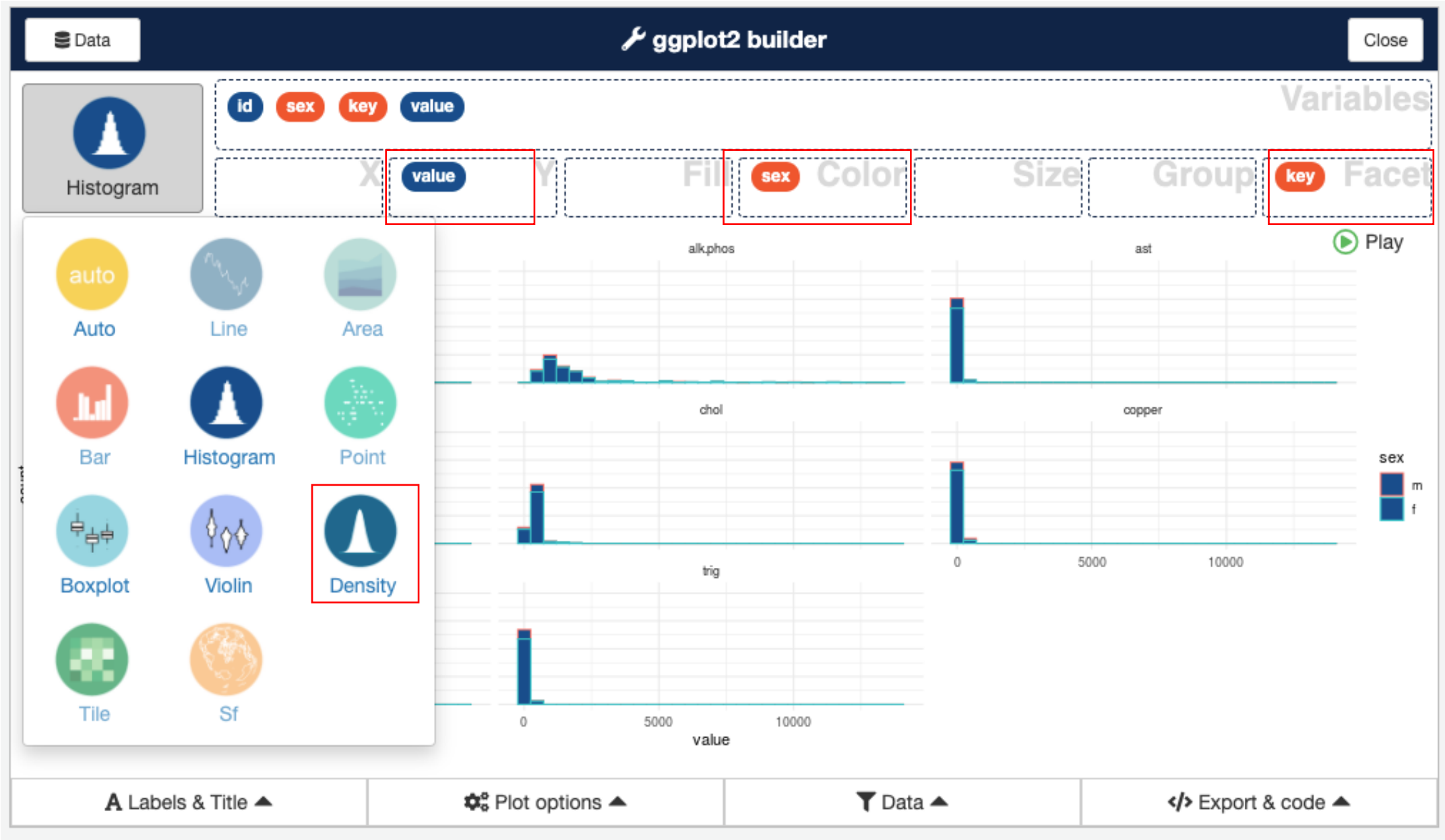

File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 4.0.0
course_commands.R*
pbc
pbc_final
Source on Save
1 install.packages("esquisse")
2 library(esquisse)
3
1:1 (Top Level)
CLIPR
Value to clipboard
Output to clipboard
ESQUISSE
'ggplot2' builder
Export 'ggplot2' plots to 'PowerPoint'
REPREX
Render reprex...
Reprex selection
Console
Terminal
Jobs
/cloud/project/
>
History
Connections
Import Dataset
List
Global Environment
Plots
Packages
Help
Viewer
Zoom
Export
Publish
albumin
alk.phos
ast
bili
chol
copper
platelet
trig
density
value
sex
m
f

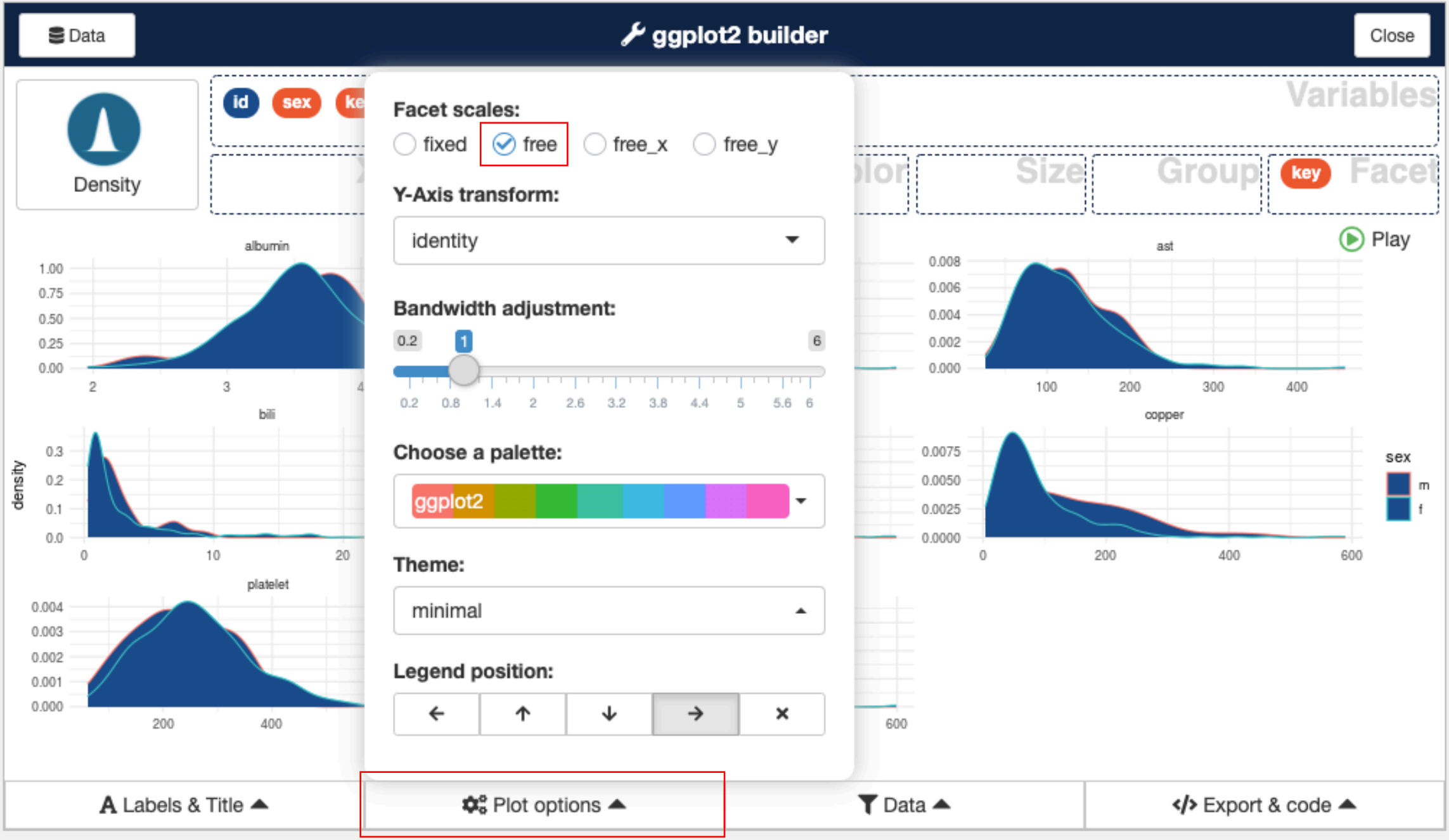
Let’s build our plot

- First let’s select the data frame we will use: pbc_long
- Then we click Validate imported data
- We can drag variables:
  - ”value” into Y
  - “sex” into Color
  - ”key” into Facet
- Click on “histogram” to change the plot type to “Density”
- Click on “Plot options” on the bottom tab
- Select Facet scales: free

C'est le temps que tu as perdu pour ta rose qui rend ta rose importante.



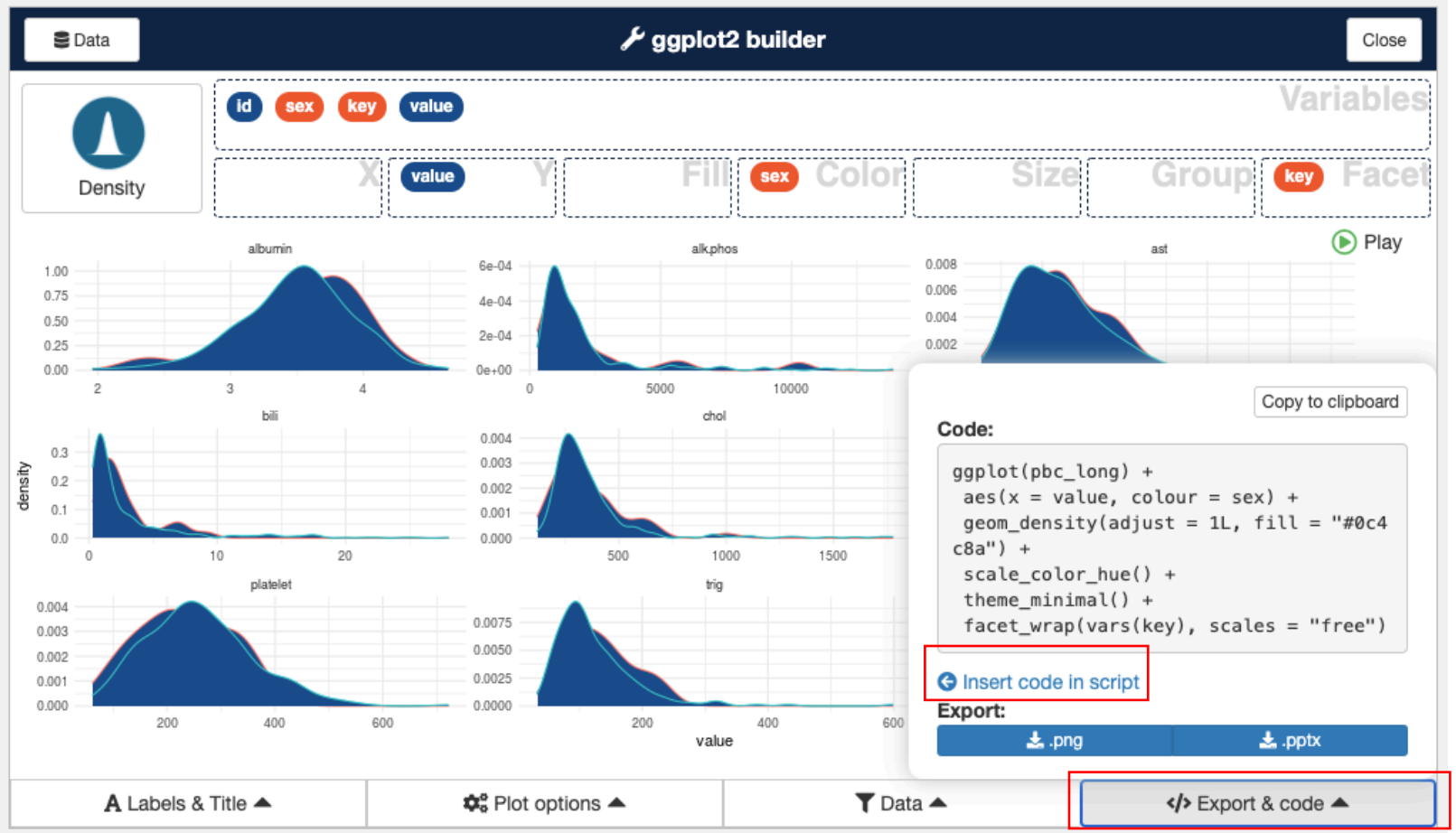

Data
ggplot2 builder
Close
Histogram
id
sex
key
value
Variables
X
value
Y
Fill
sex
Color
Size
Group
key
Facet
Play
auto
Auto
Line
Area
Bar
Histogram
Point
Boxplot
Violin
Density
Tile
Sf
alk.phos
ast
chol
copper
trig
0
5000
10000
value
sex
m
f
Labels & Title
Plot options
Data
Export & code


Data
ggplot2 builder
Close
Density
id
sex
key
Variables
Size
Group
Facet
Play
Facet scales:
fixed
free
free_x
free_y
Y-Axis transform:
identity
Bandwidth adjustment:
0.2
1
6
Choose a palette:
ggplot2
Theme:
minimal
Legend position:
albumin
ast
bili
copper
platelet
density
sex
m
f
Labels & Title
Plot options
Data
Export & code

We can export the code and the graph

- We can export the graph by selecting Export: .png
- Better yet let's push the code to our script
- Select "Insert code in script"

- This graph doesn't look quite the same as the one we coded earlier…
- Our density plots are filled in making it difficult to see the groups
- What part of the code is different?
- Try modifying the code, how does it change the plots
- Can you edit the code so it matched our original plots?

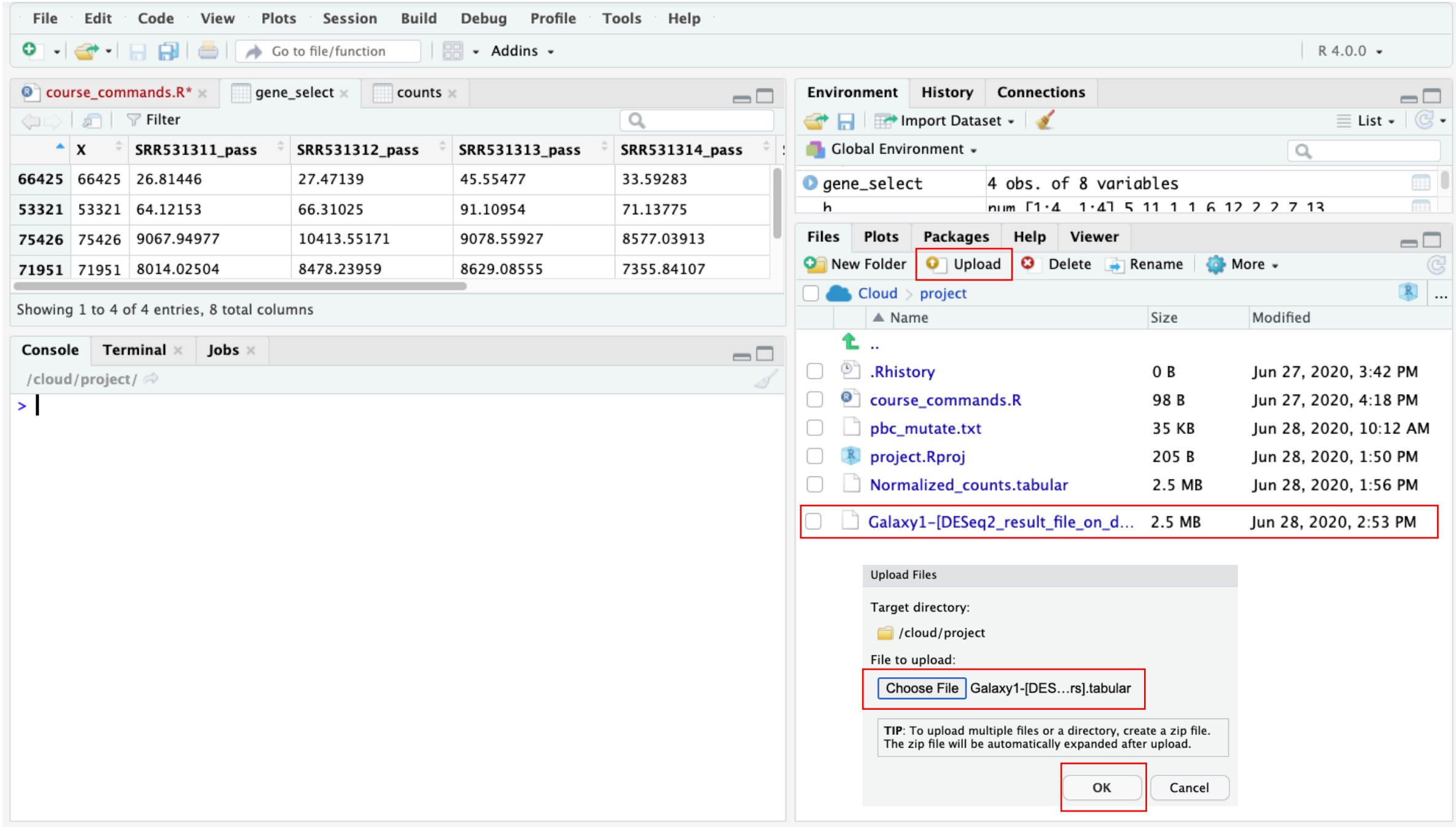


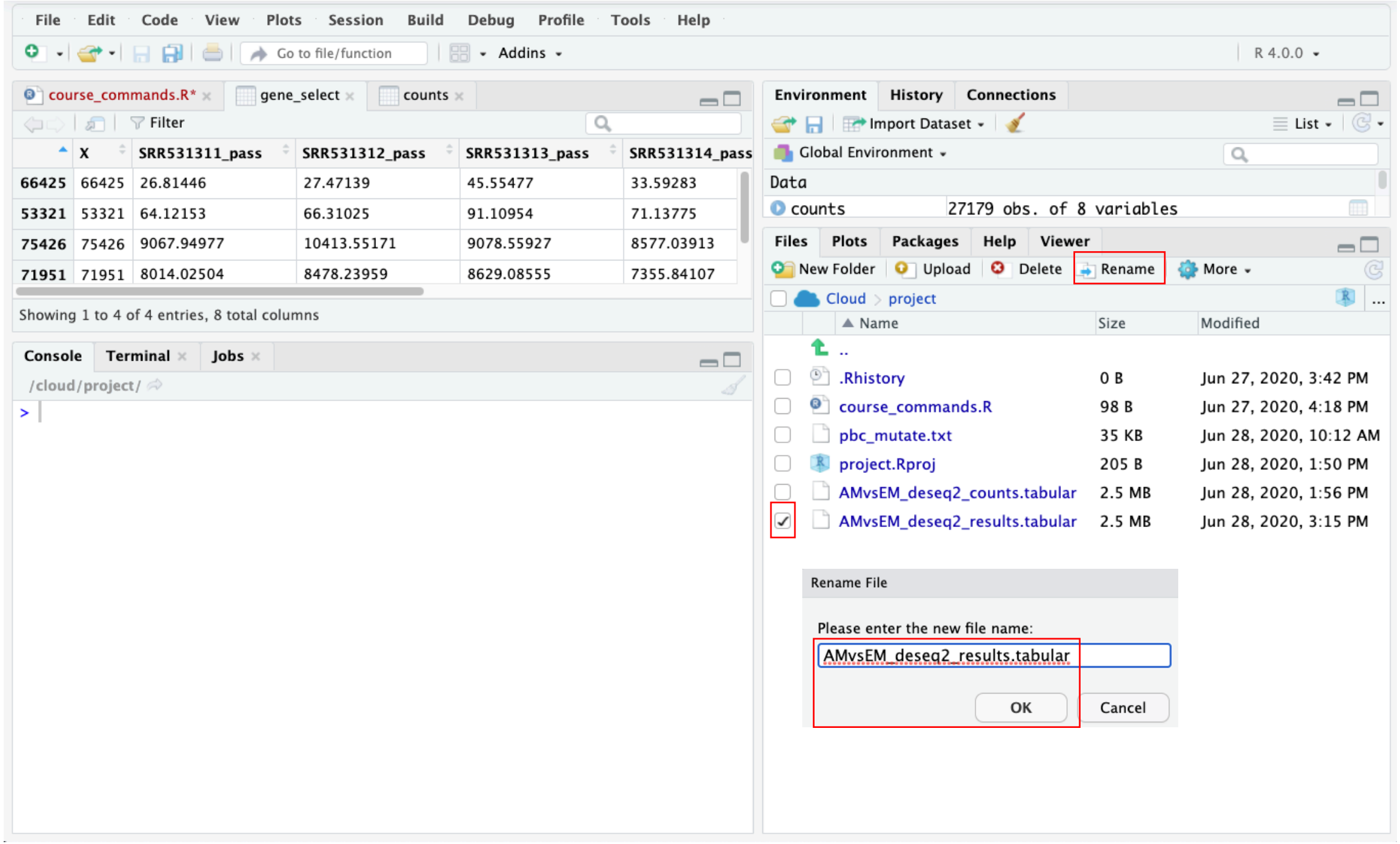

```
ggplot(pbc_long) +
 aes(x = value, colour = sex) +
 geom_density(adjust = 1L, fill = "#0c4c8a") +
 scale_color_hue() +
 theme_minimal() +
 facet_wrap(vars(key), scales = "free")
```

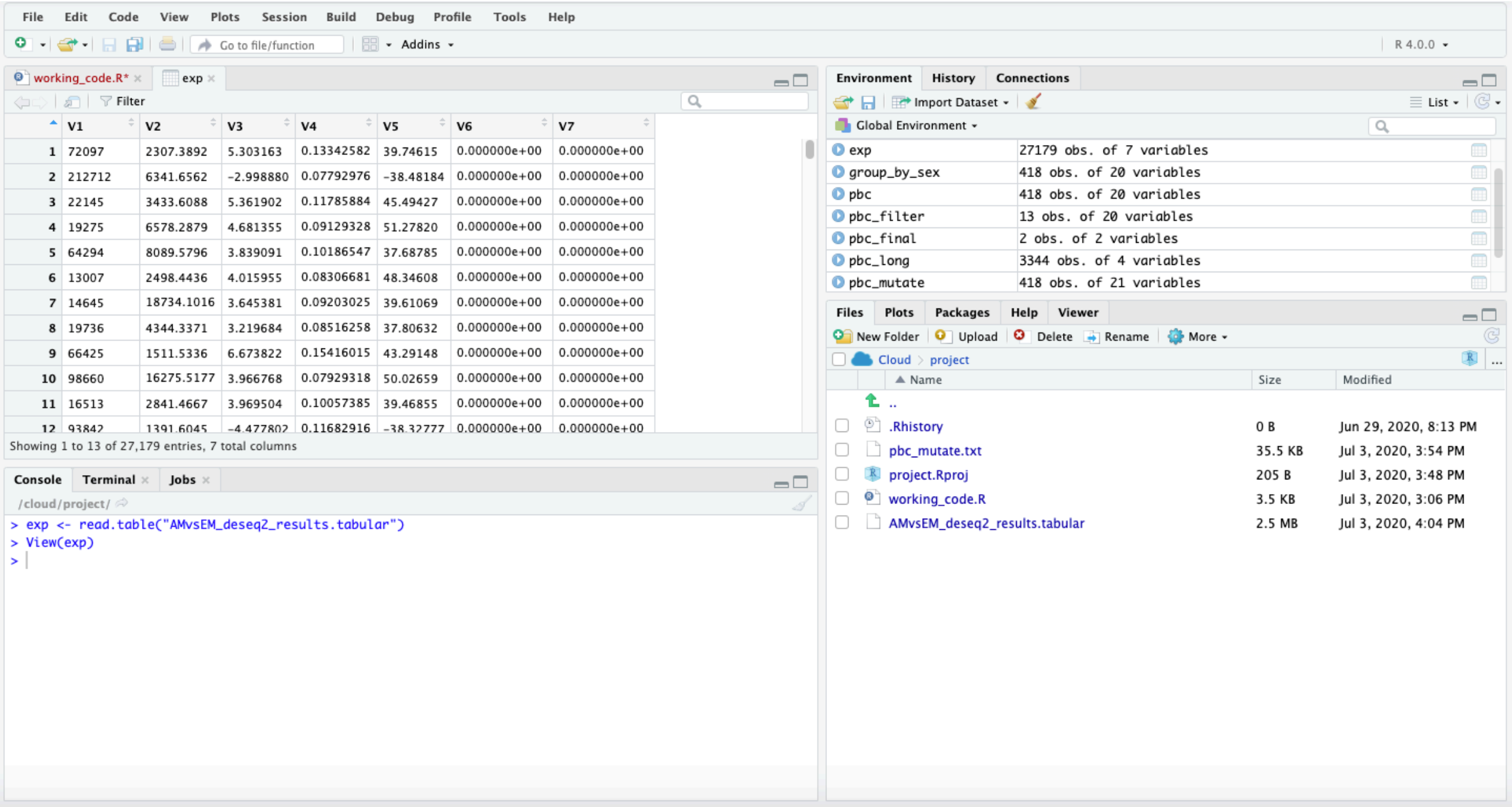
Volcano Plot

Loading DEseq2 results into R

- We will load the Deseq2 Normalized counts table
  - This is from our galaxy exercise
  - Can also be found in the github repo in the “data” folder
- Click the upload button
- Click “Choose File”
- We will need to find the file path for our data
  - Default: found in the download folder on your computer
- Click “ok”
- exp <- read.table("AMvsEM_deseq2_results.tabular")
- View(exp)


File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 4.0.0
course_commands.R*
gene_select
counts
Filter
X SRR531311_pass SRR531312_pass SRR531313_pass SRR531314_pass
66425 66425 26.81446 27.47139 45.55477 33.59283
53321 53321 64.12153 66.31025 91.10954 71.13775
75426 75426 9067.94977 10413.55171 9078.55927 8577.03913
71951 71951 8014.02504 8478.23959 8629.08555 7355.84107
Showing 1 to 4 of 4 entries, 8 total columns
Console
Terminal
Jobs
/cloud/project/
>
Environment
History
Connections
Import Dataset
List
Global Environment
gene_select 4 obs. of 8 variables
Files
Plots
Packages
Help
Viewer
New Folder
Upload
Delete
Rename
More
Cloud
project
Name
Size
Modified
..
.Rhistory 0 B Jun 27, 2020, 3:42 PM
course_commands.R 98 B Jun 27, 2020, 4:18 PM
pbc_mutate.txt 35 KB Jun 28, 2020, 10:12 AM
project.Rproj 205 B Jun 28, 2020, 1:50 PM
Normalized_counts.tabular 2.5 MB Jun 28, 2020, 1:56 PM
Galaxy1-[DESeq2_result_file_on_d... 2.5 MB Jun 28, 2020, 2:53 PM
Upload Files
Target directory:
/cloud/project
File to upload:
Choose File
Galaxy1-[DES…rs].tabular
TIP: To upload multiple files or a directory, create a zip file.
The zip file will be automatically expanded after upload.
OK
Cancel


File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 4.0.0
course_commands.R*
gene_select
counts
Filter
X SRR531311_pass SRR531312_pass SRR531313_pass SRR531314_pass
66425 66425 26.81446 27.47139 45.55477 33.59283
53321 53321 64.12153 66.31025 91.10954 71.13775
75426 75426 9067.94977 10413.55171 9078.55927 8577.03913
71951 71951 8014.02504 8478.23959 8629.08555 7355.84107
Showing 1 to 4 of 4 entries, 8 total columns
Console Terminal Jobs
/cloud/project/
>
Environment History Connections
Import Dataset
List
Global Environment
Data
counts 27179 obs. of 8 variables
Files Plots Packages Help Viewer
New Folder Upload Delete Rename More
Cloud project
Name Size Modified
..
.Rhistory 0 B Jun 27, 2020, 3:42 PM
course_commands.R 98 B Jun 27, 2020, 4:18 PM
pbc_mutate.txt 35 KB Jun 28, 2020, 10:12 AM
project.Rproj 205 B Jun 28, 2020, 1:50 PM
AMvsEM_deseq2_counts.tabular 2.5 MB Jun 28, 2020, 1:56 PM
AMvsEM_deseq2_results.tabular 2.5 MB Jun 28, 2020, 3:15 PM
Rename File
Please enter the new file name:
AMvsEM_deseq2_results.tabular
OK
Cancel


File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 4.0.0
working_code.R*
exp
Filter
V1 V2 V3 V4 V5 V6 V7
1 72097 2307.3892 5.303163 0.13342582 39.74615 0.000000e+00 0.000000e+00
2 212712 6341.6562 -2.998880 0.07792976 -38.48184 0.000000e+00 0.000000e+00
3 22145 3433.6088 5.361902 0.11785884 45.49427 0.000000e+00 0.000000e+00
4 19275 6578.2879 4.681355 0.09129328 51.27820 0.000000e+00 0.000000e+00
5 64294 8089.5796 3.839091 0.10186547 37.68785 0.000000e+00 0.000000e+00
6 13007 2498.4436 4.015955 0.08306681 48.34608 0.000000e+00 0.000000e+00
7 14645 18734.1016 3.645381 0.09203025 39.61069 0.000000e+00 0.000000e+00
8 19736 4344.3371 3.219684 0.08516258 37.80632 0.000000e+00 0.000000e+00
9 66425 1511.5336 6.673822 0.15416015 43.29148 0.000000e+00 0.000000e+00
10 98660 16275.5177 3.966768 0.07929318 50.02659 0.000000e+00 0.000000e+00
11 16513 2841.4667 3.969504 0.10057385 39.46855 0.000000e+00 0.000000e+00
12 93842 1391.6045 -4.477802 0.11682916 -38.32777 0.000000e+00 0.000000e+00
Showing 1 to 13 of 27,179 entries, 7 total columns
Console Terminal Jobs
/cloud/project/
> exp <- read.table("AMvsEM_deseq2_results.tabular")
> View(exp)
>
Environment History Connections
Import Dataset
List
Global Environment
exp 27179 obs. of 7 variables
group_by_sex 418 obs. of 20 variables
pbc 418 obs. of 20 variables
pbc_filter 13 obs. of 20 variables
pbc_final 2 obs. of 2 variables
pbc_long 3344 obs. of 4 variables
pbc_mutate 418 obs. of 21 variables
Files Plots Packages Help Viewer
New Folder Upload Delete Rename More
Cloud project
Name Size Modified
.. 0 B Jun 29, 2020, 8:13 PM
.Rhistory 35.5 KB Jul 3, 2020, 3:54 PM
pbc_mutate.txt 205 B Jul 3, 2020, 3:48 PM
project.Rproj 3.5 KB Jul 3, 2020, 3:06 PM
working_code.R 2.5 MB Jul 3, 2020, 4:04 PM
AMvsEM_deseq2_results.tabular

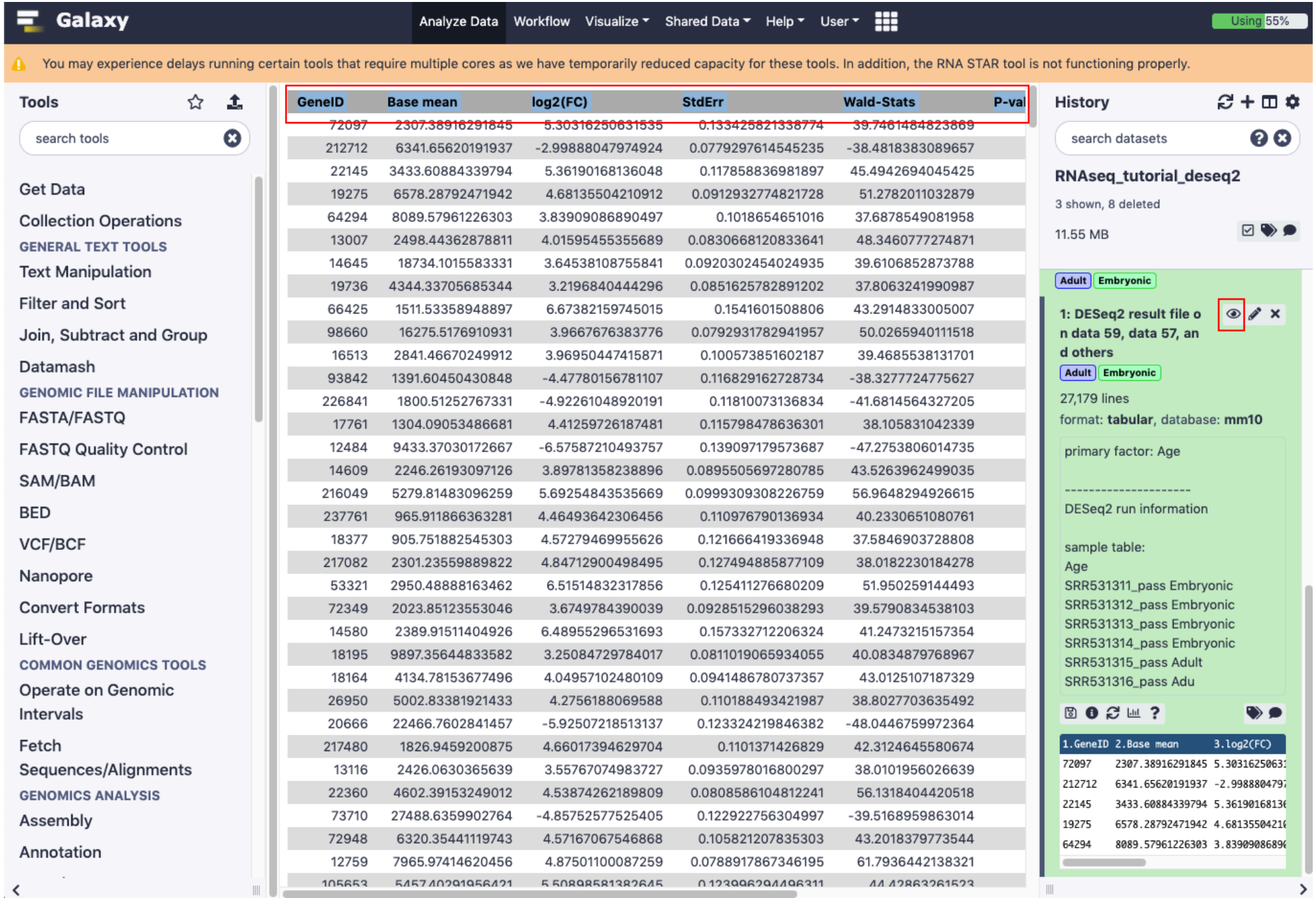
What happened to our column names?

	V1	V2	V3	V4	V5	V6	V7
1	72097	2307.3892	5.303163	0.13342582	39.74615	0.000000e+00	0.000000e+00
2	212712	6341.6562	-2.998880	0.07792976	-38.48184	0.000000e+00	0.000000e+00
3	22145	3433.6088	5.361902	0.11785884	45.49427	0.000000e+00	0.000000e+00
4	19275	6578.2879	4.681355	0.09129328	51.27820	0.000000e+00	0.000000e+00
5	64294	8089.5796	3.839091	0.10186547	37.68785	0.000000e+00	0.000000e+00
6	13007	2498.4436	4.015955	0.08306681	48.34608	0.000000e+00	0.000000e+00

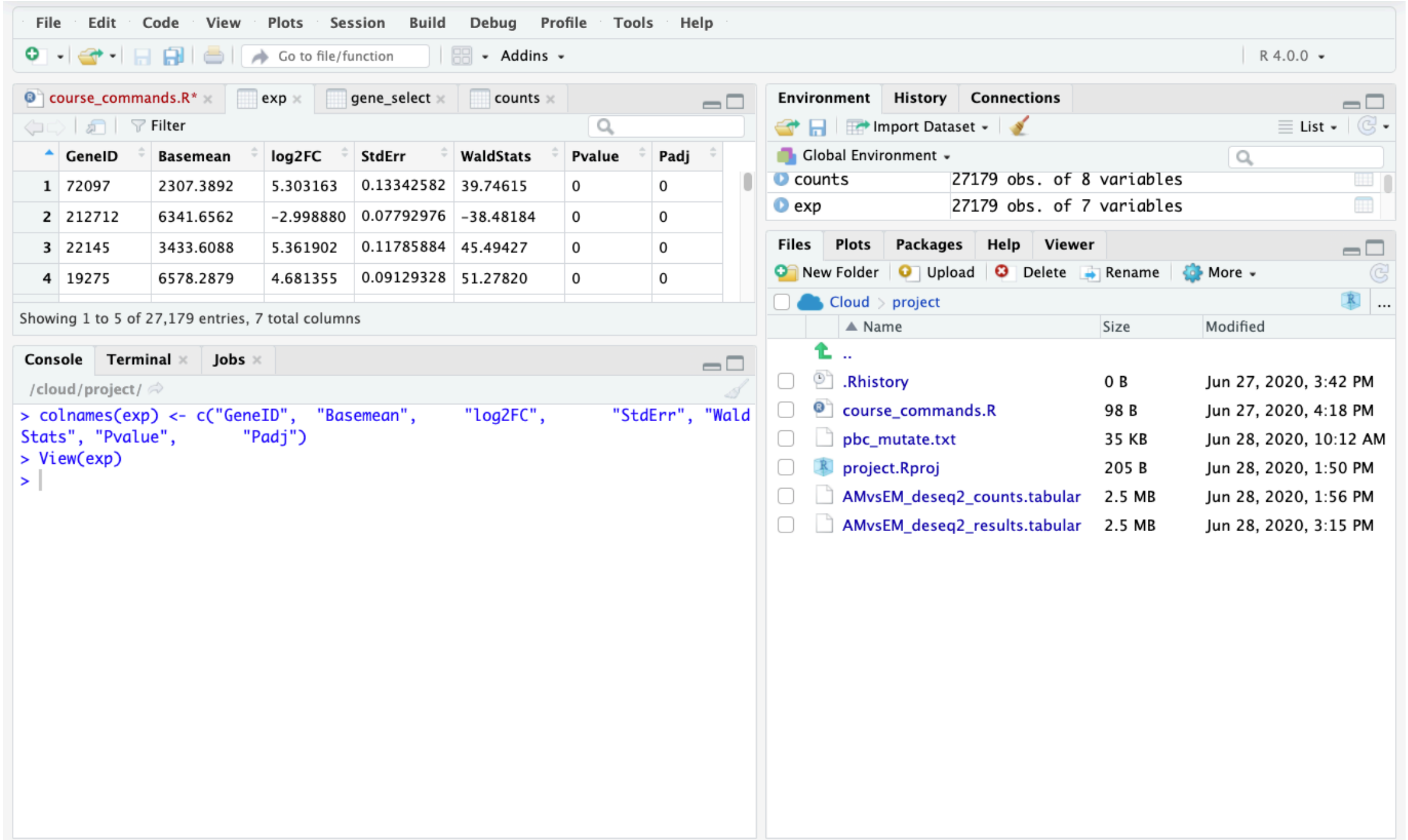
- Inspecting the downloaded galaxy: no column names
- Turns out galaxy does not save column headers
- We will need to add them

Let’s get those column names!

- We can go back to galaxy and click View for the DESeq2 results file
- We can copy and paste the names
- We need to add quotes around each name and add commas to separate the names
- Spaces and special characters will not behave well as column names
  - Remove the space between Base mean
  - Remove the parentheses around log2(FC)
  - Remove hyphens in Wald-Stats, P-value, and P-adj
- colnames(exp) <- c("GeneID", "Basemean", "log2FC", "StdErr", "WaldStats", "Pvalue", "Padj")
- View(exp)

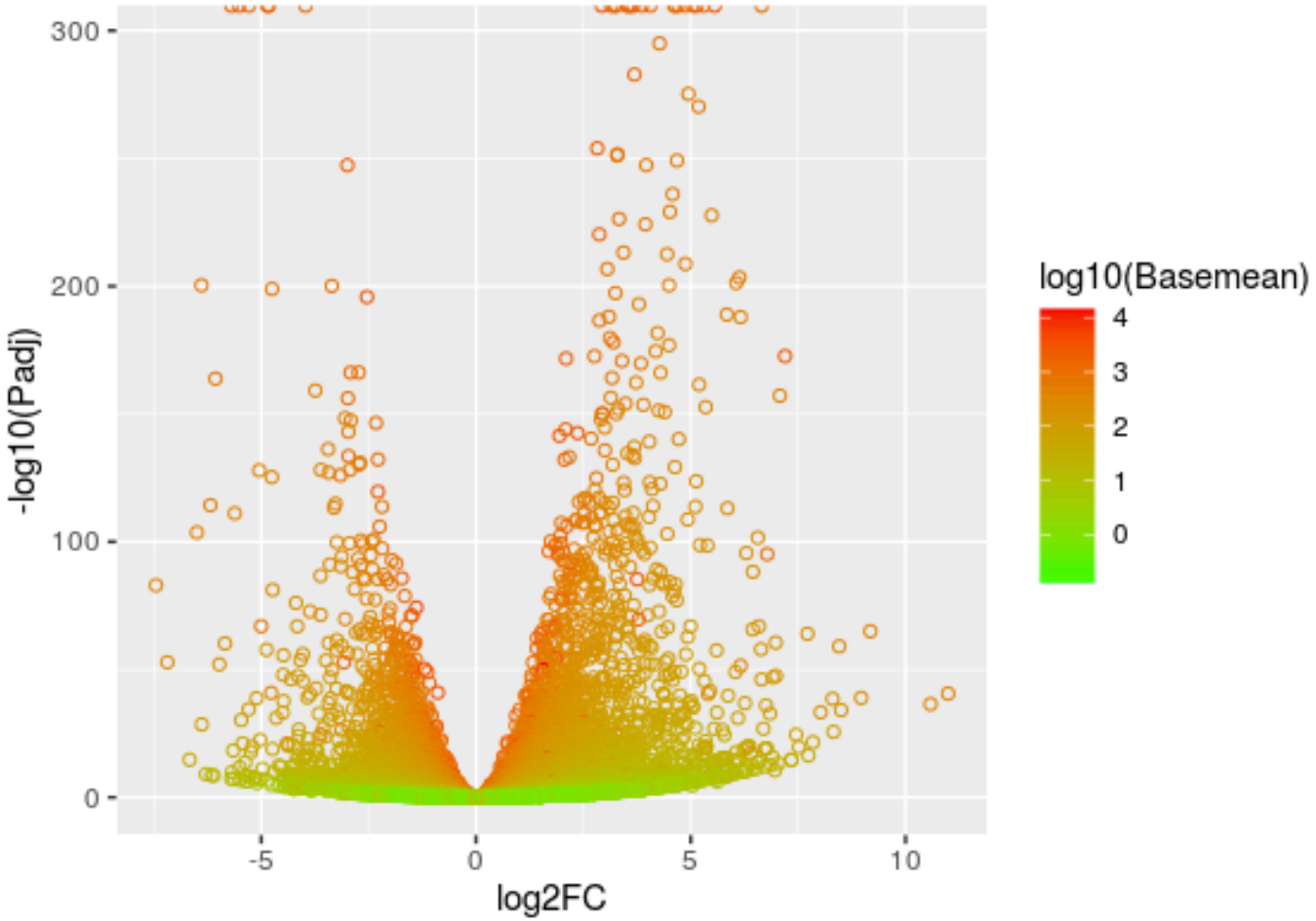

Galaxy
Analyze Data
Workflow
Visualize
Shared Data
Help
User
Using 55%
You may experience delays running certain tools that require multiple cores as we have temporarily reduced capacity for these tools. In addition, the RNA STAR tool is not functioning properly.
Tools
search tools
Get Data
Collection Operations
GENERAL TEXT TOOLS
Text Manipulation
Filter and Sort
Join, Subtract and Group
Datamash
GENOMIC FILE MANIPULATION
FASTA/FASTQ
FASTQ Quality Control
SAM/BAM
BED
VCF/BCF
Nanopore
Convert Formats
Lift-Over
COMMON GENOMICS TOOLS
Operate on Genomic Intervals
Fetch Sequences/Alignments
GENOMICS ANALYSIS
Assembly
Annotation
GeneID Base mean log2(FC) StdErr Wald-Stats P-val
72097 2307.38916291845 5.30316250631535 0.133425821338774 39.7461484823869
212712 6341.65620191937 -2.99888047974924 0.0779297614545235 -38.4818383089657
22145 3433.60884339794 5.36190168136048 0.117858836981897 45.4942694045425
19275 6578.28792471942 4.68135504210912 0.0912932774821728 51.2782011032879
64294 8089.57961226303 3.83909086890497 0.1018654651016 37.6878549081958
13007 2498.44362878811 4.01595455355689 0.0830668120833641 48.3460777274871
14645 18734.1015583331 3.64538108755841 0.0920302454024935 39.6106852873788
19736 4344.33705685344 3.2196840444296 0.0851625782891202 37.8063241990987
66425 1511.53358948897 6.67382159745015 0.1541601508806 43.2914833005007
98660 16275.5176910931 3.9667676383776 0.0792931782941957 50.0265940111518
16513 2841.46670249912 3.96950447415871 0.100573851602187 39.4685538131701
93842 1391.60450430848 -4.47780156781107 0.116829162728734 -38.3277724775627
226841 1800.51252767331 -4.92261048920191 0.11810073136834 -41.6814564327205
17761 1304.09053486681 4.41259726187481 0.115798478636301 38.105831042339
12484 9433.37030172667 -6.57587210493757 0.139097179573687 -47.2753806014735
14609 2246.26193097126 3.89781358238896 0.0895505697280785 43.5263962499035
216049 5279.81483096259 5.69254843535669 0.0999309308226759 56.9648294926615
237761 965.911866363281 4.46493642306456 0.110976790136934 40.2330651080761
18377 905.751882545303 4.57279469955626 0.121666419336948 37.5846903728808
217082 2301.23559889822 4.84712900498495 0.127494885877109 38.0182230184278
53321 2950.48888163462 6.51514832317856 0.125411276680209 51.950259144493
72349 2023.85123553046 3.6749784390039 0.0928515296038293 39.5790834538103
14580 2389.91511404926 6.48955296531693 0.157332712206324 41.2473215157354
18195 9897.35644833582 3.25084729784017 0.0811019065934055 40.0834879768967
18164 4134.78153677496 4.04957102480109 0.0941486780737357 43.0125107187329
26950 5002.83381921433 4.2756188069588 0.110188493421987 38.8027703635492
20666 22466.7602841457 -5.92507218513137 0.123324219846382 -48.0446759972364
217480 1826.9459200875 4.66017394629704 0.1101371426829 42.3124645580674
13116 2426.0630365639 3.55767074983727 0.0935978016800297 38.0101956026639
22360 4602.39153249012 4.53874262189809 0.0808586104812241 56.1318404420518
73710 27488.6359902764 -4.85752577525405 0.122922756304997 -39.5168959863014
72948 6320.35441119743 4.57167067546868 0.105821207835303 43.2018379773544
12759 7965.97414620456 4.87501100087259 0.0788917867346195 61.7936442138321
History
search datasets
RNAseq_tutorial_deseq2
3 shown, 8 deleted
11.55 MB
Adult Embryonic
1: DESeq2 result file on data 59, data 57, and others
Adult Embryonic
27,179 lines
format: tabular, database: mm10
primary factor: Age

DESeq2 run information
sample table:
Age
SRR531311_pass Embryonic
SRR531312_pass Embryonic
SRR531313_pass Embryonic
SRR531314_pass Embryonic
SRR531315_pass Adult
SRR531316_pass Adu
1.GeneID 2.Base mean 3.log2(FC)
72097 2307.38916291845 5.3031625063
212712 6341.65620191937 -2.9988804797
22145 3433.60884339794 5.3619016813
19275 6578.28792471942 4.6813550421
64294 8089.57961226303 3.8390908689


File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 4.0.0
course_commands.R*
exp
gene_select
counts
Filter
GeneID Basemean log2FC StdErr WaldStats Pvalue Padj
1 72097 2307.3892 5.303163 0.13342582 39.74615 0 0
2 212712 6341.6562 -2.998880 0.07792976 -38.48184 0 0
3 22145 3433.6088 5.361902 0.11785884 45.49427 0 0
4 19275 6578.2879 4.681355 0.09129328 51.27820 0 0
Showing 1 to 5 of 27,179 entries, 7 total columns
Console
Terminal
Jobs
/cloud/project/
> colnames(exp) <- c("GeneID", "Basemean", "log2FC", "StdErr", "Wald
Stats", "Pvalue", "Padj")
> View(exp)
>
Environment
History
Connections
Import Dataset
List
Global Environment
counts 27179 obs. of 8 variables
exp 27179 obs. of 7 variables
Files
Plots
Packages
Help
Viewer
New Folder
Upload
Delete
Rename
More
Cloud
project
Name
Size
Modified
..
.Rhistory 0 B Jun 27, 2020, 3:42 PM
course_commands.R 98 B Jun 27, 2020, 4:18 PM
pbc_mutate.txt 35 KB Jun 28, 2020, 10:12 AM
project.Rproj 205 B Jun 28, 2020, 1:50 PM
AMvsEM_deseq2_counts.tabular 2.5 MB Jun 28, 2020, 1:56 PM
AMvsEM_deseq2_results.tabular 2.5 MB Jun 28, 2020, 3:15 PM

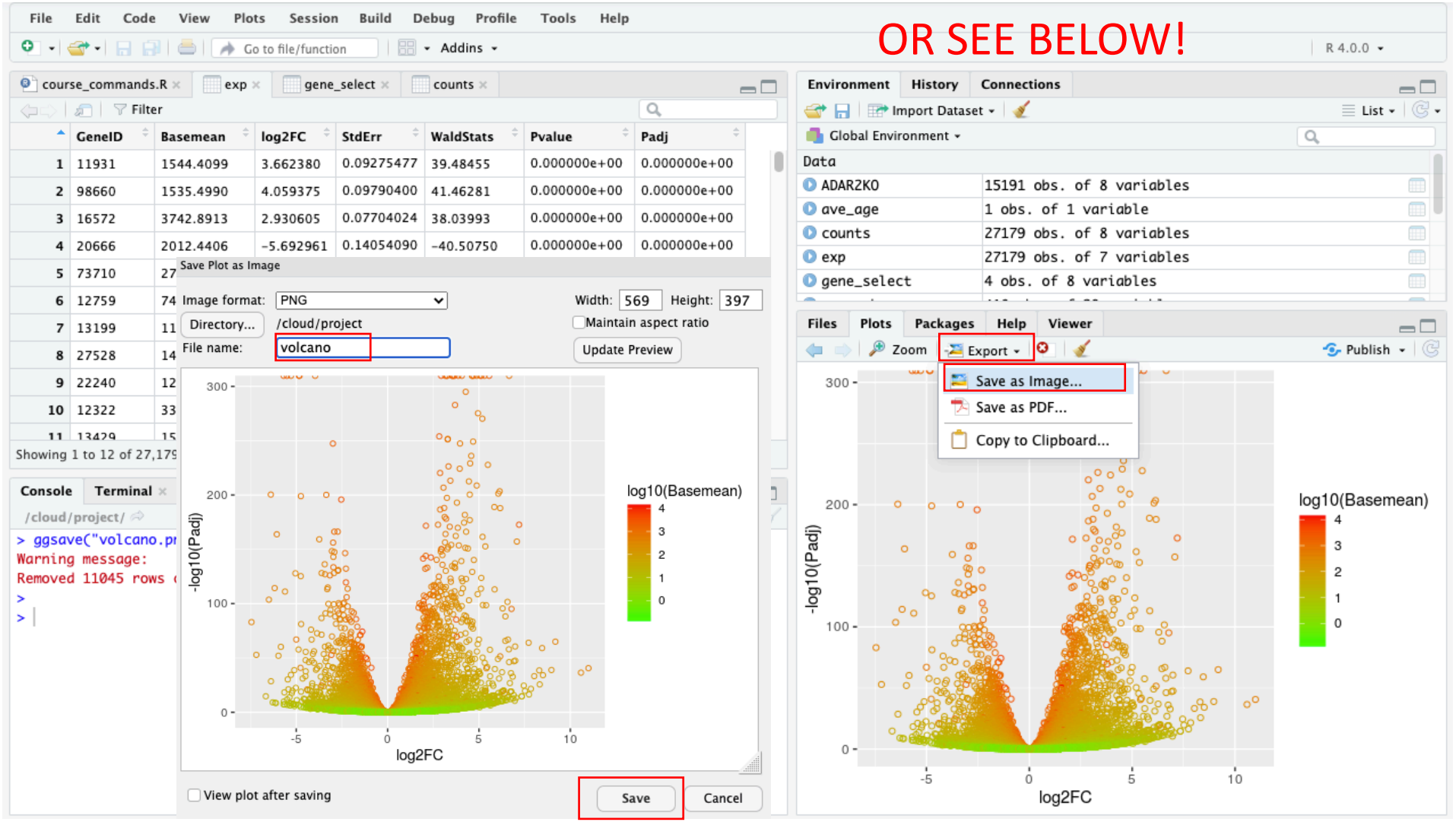
Volcano Plot

```
ggplot(exp, aes(x=log2FC, y=-log10(Padj), color=log10(Basemean)))  +
geom_point(shape=1, size=1.5) + scale_color_gradient(low="green",high="red")
```



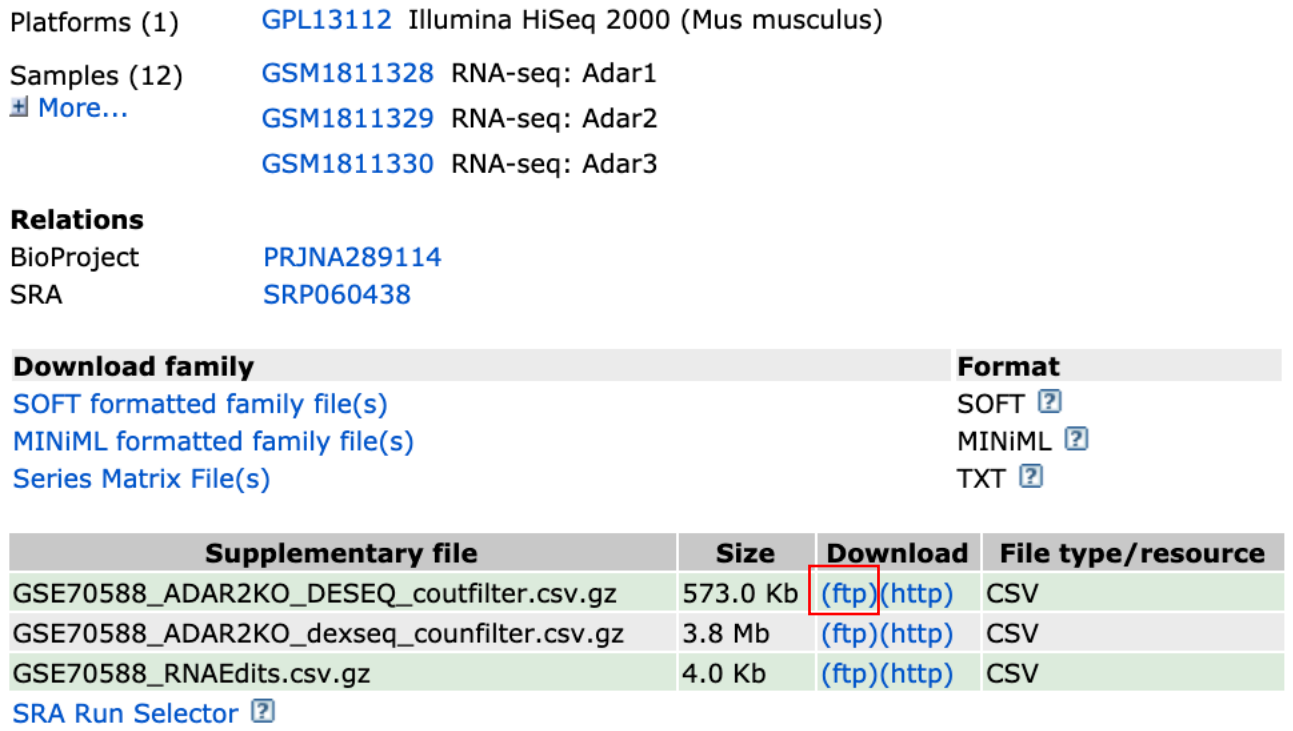
Saving your beautiful graph!

ggsave("volcano.png", width = 6, height = 4)



Let’s get some data!

- https://www.ncbi.nlm.nih.gov/geo/
- In the search bar: GSE70588
- Scroll down to the bottom
- Find the *DESES_countfilter.csv.gz
- Right click on the ftp download
- Select “Copy link address”
- We will paste this address in R



Platforms (1) GPL13112 Illumina HiSeq 2000 (Mus musculus)

Samples (12) GSM1811328 RNA-seq: Adar1
More... GSM1811329 RNA-seq: Adar2
GSM1811330 RNA-seq: Adar3

Relations

BioProject PRJNA289114
SRA SRP060438

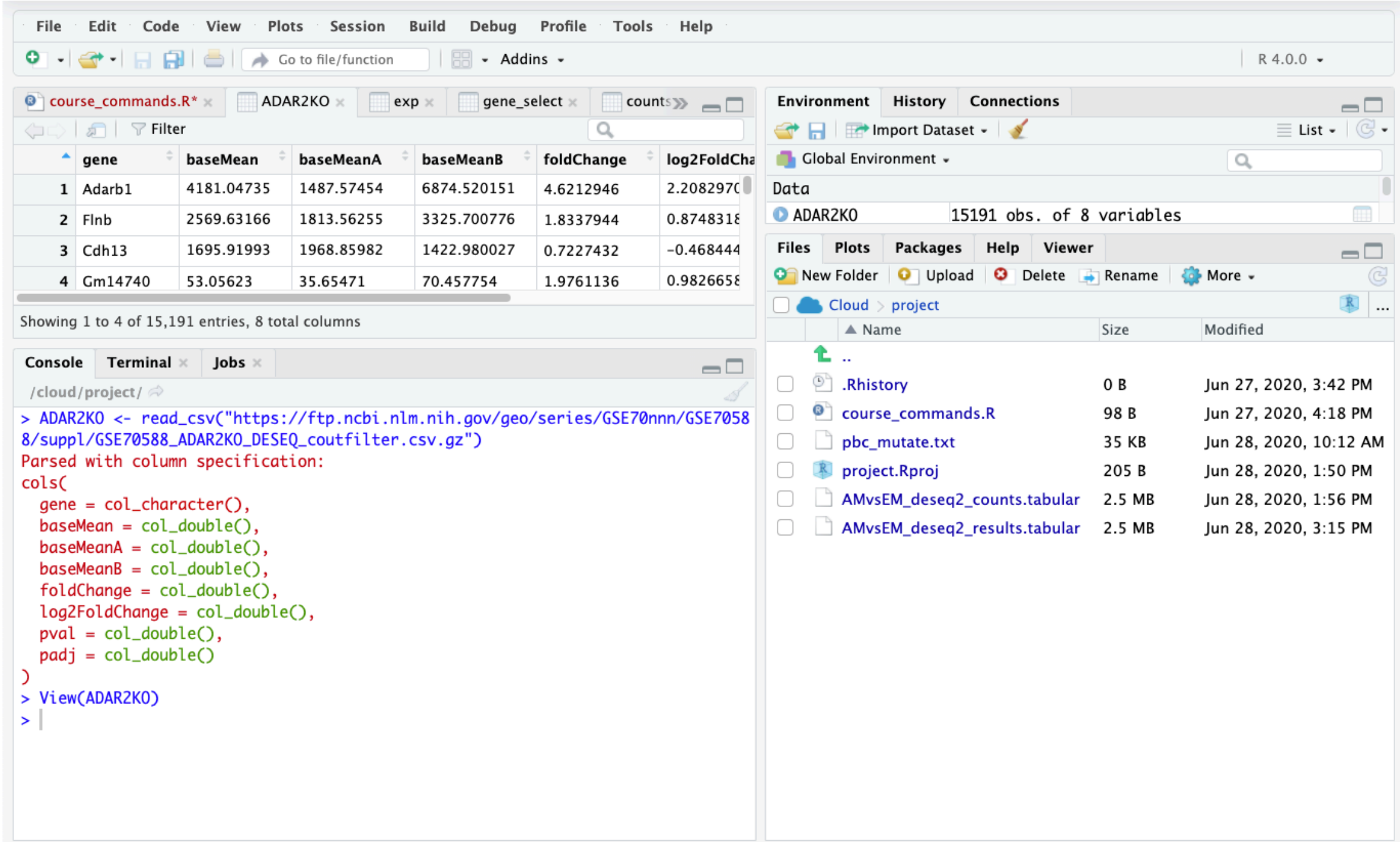
Download family	Format
SOFT formatted family file(s)	SOFT
MINiML formatted family file(s)	MINiML
Series Matrix File(s)	TXT

Supplementary file	Size	Download	File type/resource
GSE70588_ADAR2KO_DESEQ_coutfilter.csv.gz	573.0 Kb	(ftp)(http)	CSV
GSE70588_ADAR2KO_dexseq_counfilter.csv.gz	3.8 Mb	(ftp)(http)	CSV
GSE70588_RNAEdits.csv.gz	4.0 Kb	(ftp)(http)	CSV

SRA Run Selector

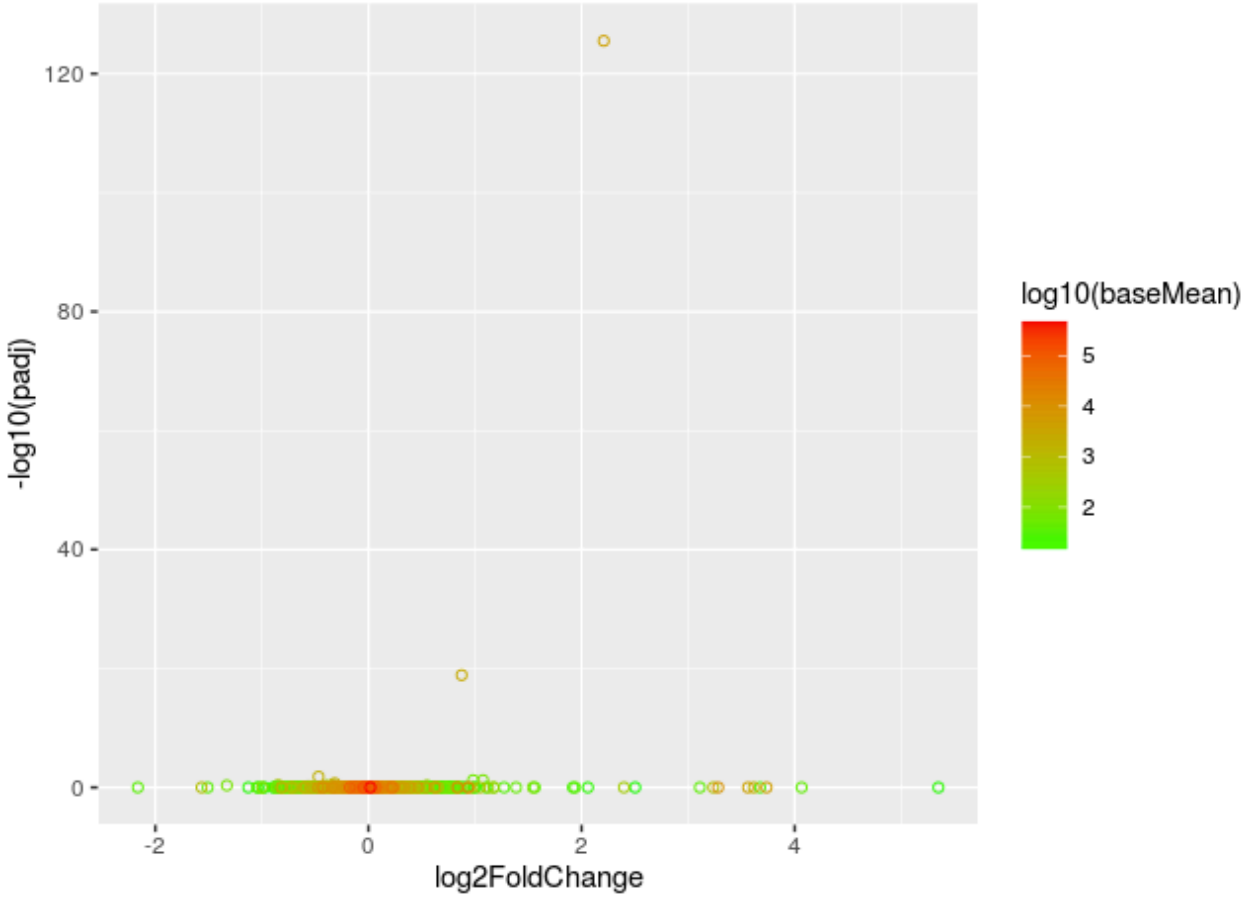
Load your table

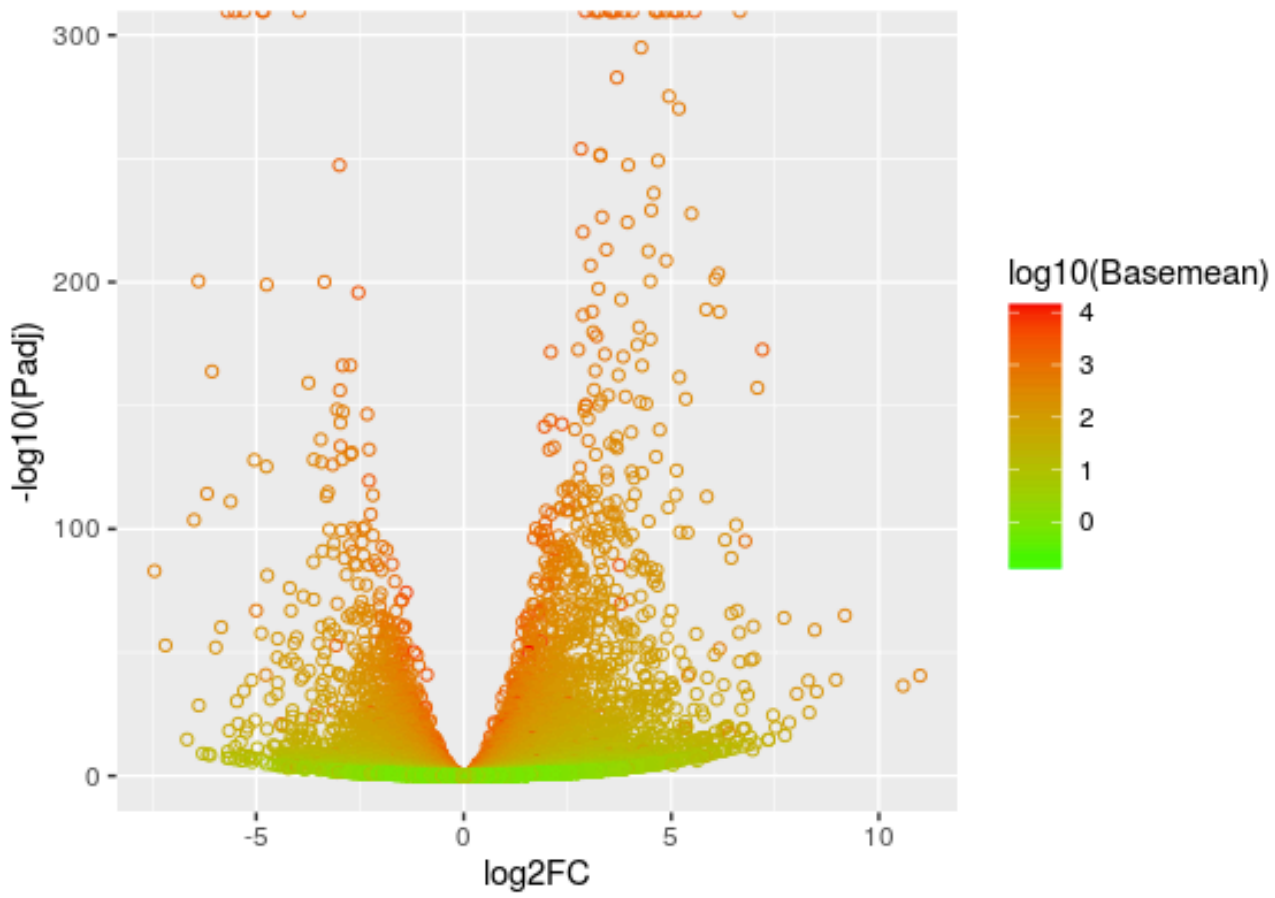
- ADAR2KO <- read_csv("ftp://ftp.ncbi.nlm.nih.gov/geo/series/GSE70nnn/GSE70588/suppl/GSE70588_ADAR2KO_DESEQ_coutfilter.csv.gz")
- View(ADAR2KO)


File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 4.0.0
course_commands.R*
ADAR2KO
exp
gene_select
counts
Filter
gene baseMean baseMeanA baseMeanB foldChange log2FoldCha
1 Adarb1 4181.04735 1487.57454 6874.520151 4.6212946 2.208297
2 Flnb 2569.63166 1813.56255 3325.700776 1.8337944 0.8748318
3 Cdh13 1695.91993 1968.85982 1422.980027 0.7227432 -0.468444
4 Gm14740 53.05623 35.65471 70.457754 1.9761136 0.9826658
Showing 1 to 4 of 15,191 entries, 8 total columns
Console
Terminal
Jobs
/cloud/project/
> ADAR2KO <- read_csv("https://ftp.ncbi.nlm.nih.gov/geo/series/GSE70nnn/GSE70588/suppl/GSE70588_ADAR2KO_DESEQ_coutfilter.csv.gz")
Parsed with column specification:
cols(
gene = col_character(),
baseMean = col_double(),
baseMeanA = col_double(),
baseMeanB = col_double(),
foldChange = col_double(),
log2FoldChange = col_double(),
pval = col_double(),
padj = col_double()
)
> View(ADAR2KO)
>
Environment
History
Connections
Import Dataset
List
Global Environment
Data
ADAR2KO
15191 obs. of 8 variables
Files
Plots
Packages
Help
Viewer
New Folder
Upload
Delete
Rename
More
Cloud
project
Name
Size
Modified
..
.Rhistory 0 B Jun 27, 2020, 3:42 PM
course_commands.R 98 B Jun 27, 2020, 4:18 PM
pbc_mutate.txt 35 KB Jun 28, 2020, 10:12 AM
project.Rproj 205 B Jun 28, 2020, 1:50 PM
AMvsEM_deseq2_counts.tabular 2.5 MB Jun 28, 2020, 1:56 PM
AMvsEM_deseq2_results.tabular 2.5 MB Jun 28, 2020, 3:15 PM

Volcano plot

ggplot(data=ADAR2KO, aes(x=log2FoldChange, y=-log10(padj), color=log10(baseMean))) +
geom_point(shape=1, size=1.5) + scale_color_gradient(low="green",high="red")





Did we do something wrong?

NCBI

Gene Expression Omnibus

HOME | SEARCH | SITE MAP | GEO Publications | FAQ | MIAME | Email GEO

NCBI > GEO > **Accession Display**

Not logged in | Login

Scope: Self | Format: HTML | Amount: Quick | GEO accession: GSE70588 | GO

Series GSE70588 Query DataSets for GSE70588

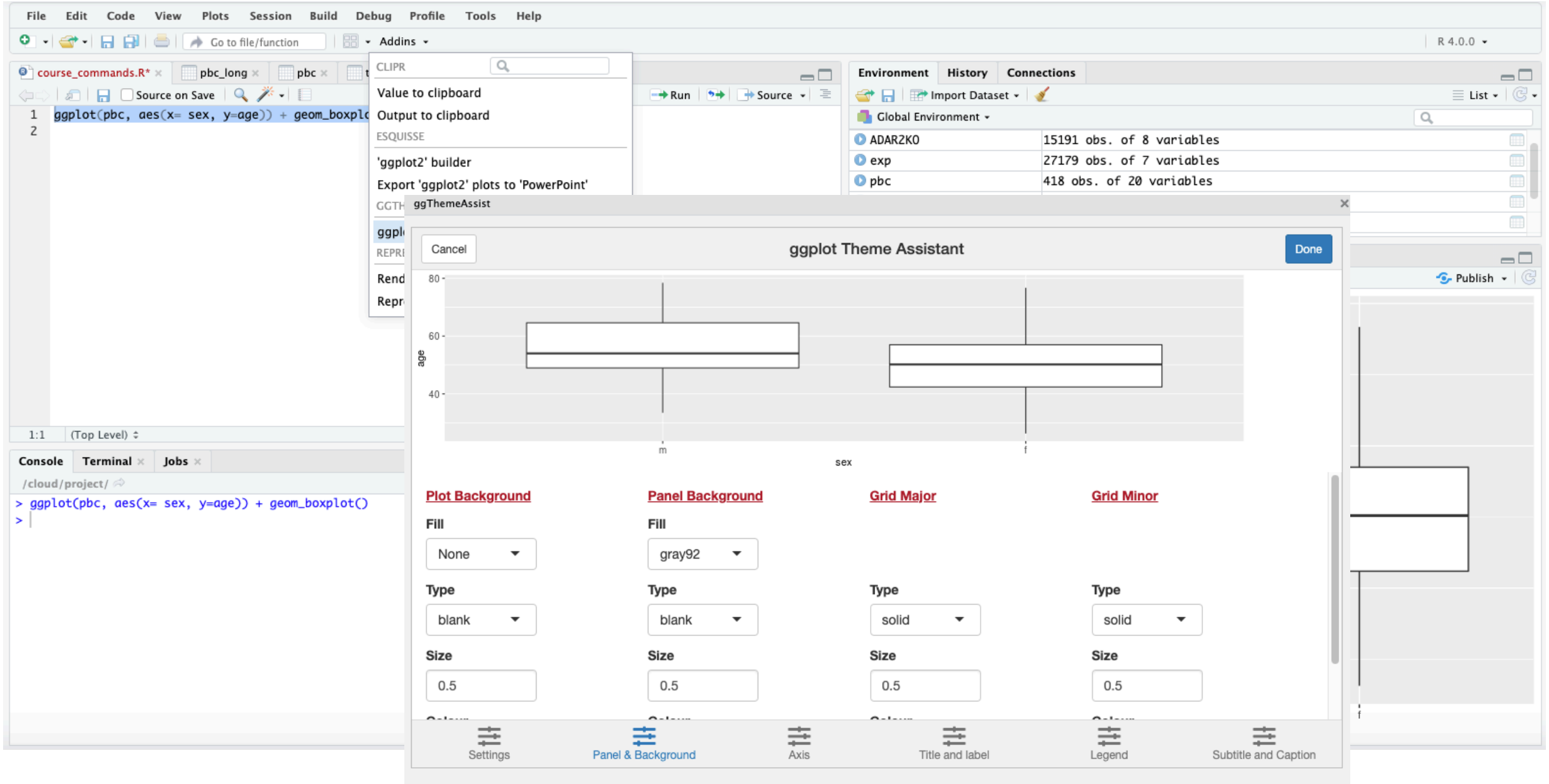
Status	Public on Jan 01, 2016
Title	ADAR2 affects mRNA coding sequence edits but not gene expression or splicing in vivo
Organism	Mus musculus
Experiment type	Expression profiling by high throughput sequencing
Summary	Adenosine deaminases (ADARs) are RNA binding proteins that bind to double stranded RNA and convert adenosine to inosine. Editing creates multiple isoforms of neurotransmitter receptors, including AMPA-subtype Glutamate channels such as Gria2 that is edited to introduce a Q to R amino acid change. Adar2 knock out mice die of seizures shortly after birth, but if the Gria2 Q/R editing site is mutated to mimic the edited version then the animals are viable. We performed RNA-Seq on the frontal cortices of Adar2-/- Gria2R/R mice and their Adar2+/+ Gria2R/R littermates, quantifying overall gene expression, splicing, and A to I editing in a transcriptome-wide fashion. We found 56 editing sites whose level of editing was significantly diminished in the Adar2 deficient animals. The majority of Adar2 responsive editing sites were in the coding regions of genes. Interestingly, other than Adar2 expression, there were only two additional statistically significant differentially expressed genes, Flnb and Cdh13. There were also only three exons that showed statistically significant differences in expression levels between the two genotypes. This work illustrates that ADAR2 is important in site-specific changes of protein coding sequences but has only modest effects on gene expression or splicing in vivo.

Extra Bits

Interactive plot customization!

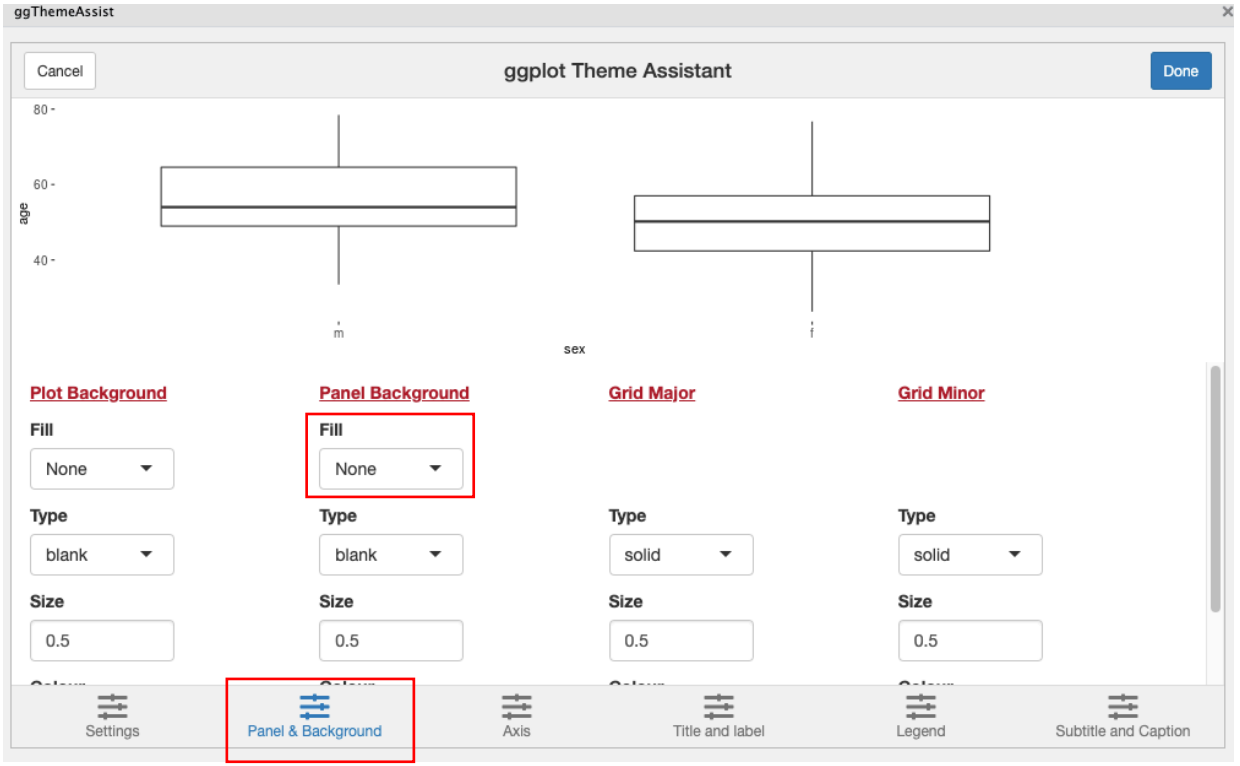
Interactive plot customization with ggThemeAssist

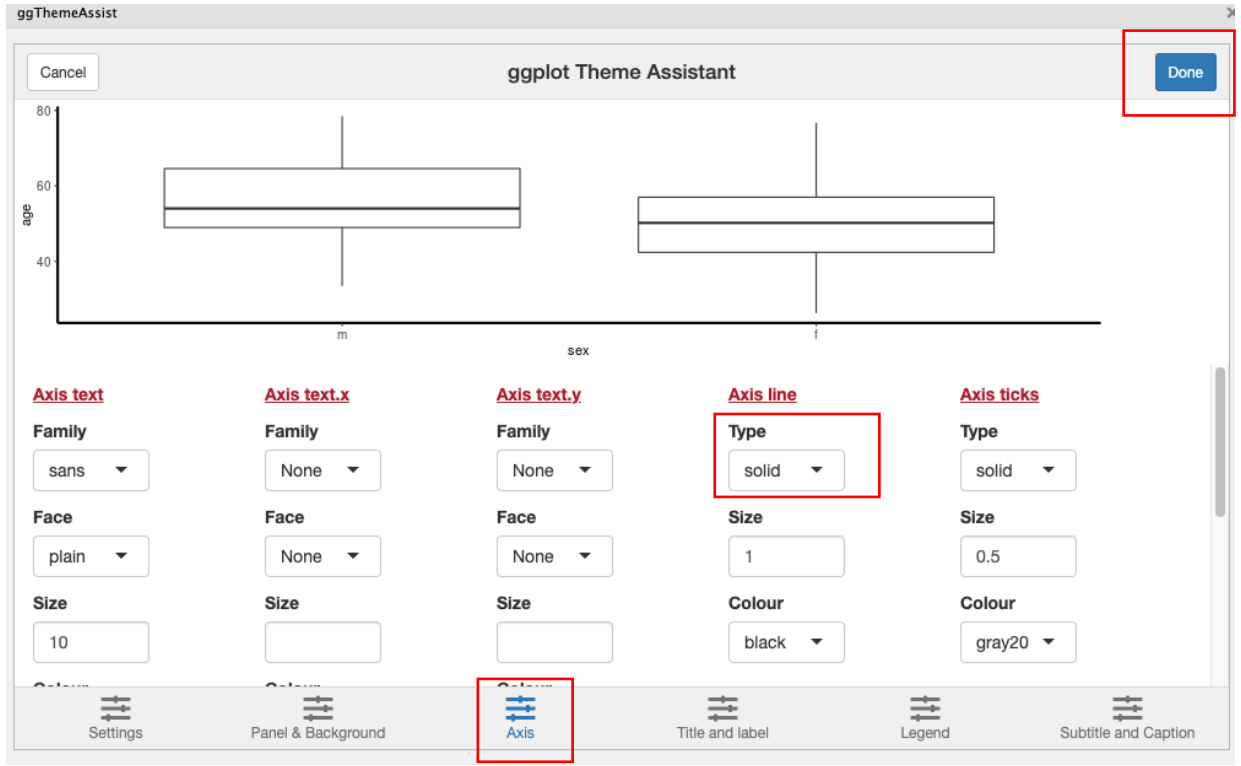
- install.packages("ggThemeAssist")
- library(ggThemeAssist)
- ggplot(pbc, aes(x= sex, y=age)) + geom_boxplot()
- Highlight this code in your Rscript
- Go to Addins and select ggplot Theme assist


File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 4.0.0
course_commands.R*
pbc_long
pbc
Source on Save
1 ggplot(pbc, aes(x= sex, y=age)) + geom_boxplo
2
Run
Source
CLIPR
Value to clipboard
Output to clipboard
ESQUISSE
'ggplot2' builder
Export 'ggplot2' plots to 'PowerPoint'
Environment
History
Connections
Import Dataset
List
Global Environment
ADAR2KO 15191 obs. of 8 variables
exp 27179 obs. of 7 variables
pbc 418 obs. of 20 variables
ggThemeAssist
Cancel
ggplot Theme Assistant
Done
80
60
40
age
m
f
sex
Plot Background
Panel Background
Grid Major
Grid Minor
Fill
None
gray92
Type
blank
blank
solid
solid
Size
0.5
0.5
0.5
0.5
Settings
Panel & Background
Axis
Title and label
Legend
Subtitle and Caption
1:1
(Top Level)
Console
Terminal
Jobs
/cloud/project/
> ggplot(pbc, aes(x= sex, y=age)) + geom_boxplot()
>
Publish
f

Interactive plot customization with ggThemeAssist

- Select the Panel & Background tab at the bottom
- In the Panel Background column choose None for the fill drop-down
- Select the Axis tab at the bottom
- In the Axis line column choose Solid for the type drop-down
- Click Done
- The package will modify your code in the Rscript






```
ggplot(pbc, aes(x= sex, y=age)) + geom_boxplot() +
theme(axis.line = element_line(linetype = "solid"), panel.background = element_rect(fill = NA))
```

Making an interactive plot!

Interactive Volcano with Plotly

- install.packages("plotly")
- library(plotly)
- plot_ly(data = exp, x = ~log2FC, y = ~-log10(Padj), color=~log10(Basemean), type = 'scatter', mode='markers', text= ~GeneID)

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | R 4.0.0

Untitled1* | exp | gene_select | counts

Source on Save | Run | Source

```
1  install.packages("plotly")
2  library(plotly)
3  plot_ly(data = exp, x = ~log2FC, y = ~-log10(Padj), color=~log10(Basemean), type = 'scat
4
```

4:1 (Top Level) R Script

Console | Terminal | Jobs

/cloud/project/

```
> plot_ly(data = exp, x = ~log2FC, y = ~-log10(Padj), color=~log10(Basemean), type = 'scatte
r', mode='markers', text= ~GeneID)
Warning messages:
1: Ignoring 11045 observations
2: `arrange_()` is deprecated as of dplyr 0.7.0.
Please use `arrange()` instead.
See vignette('programming') for more help
This warning is displayed once every 8 hours.
Call `lifecycle::last_warnings()` to see where this warning was generated.
>
>
```

Environment | History | Connections

Import Dataset | List

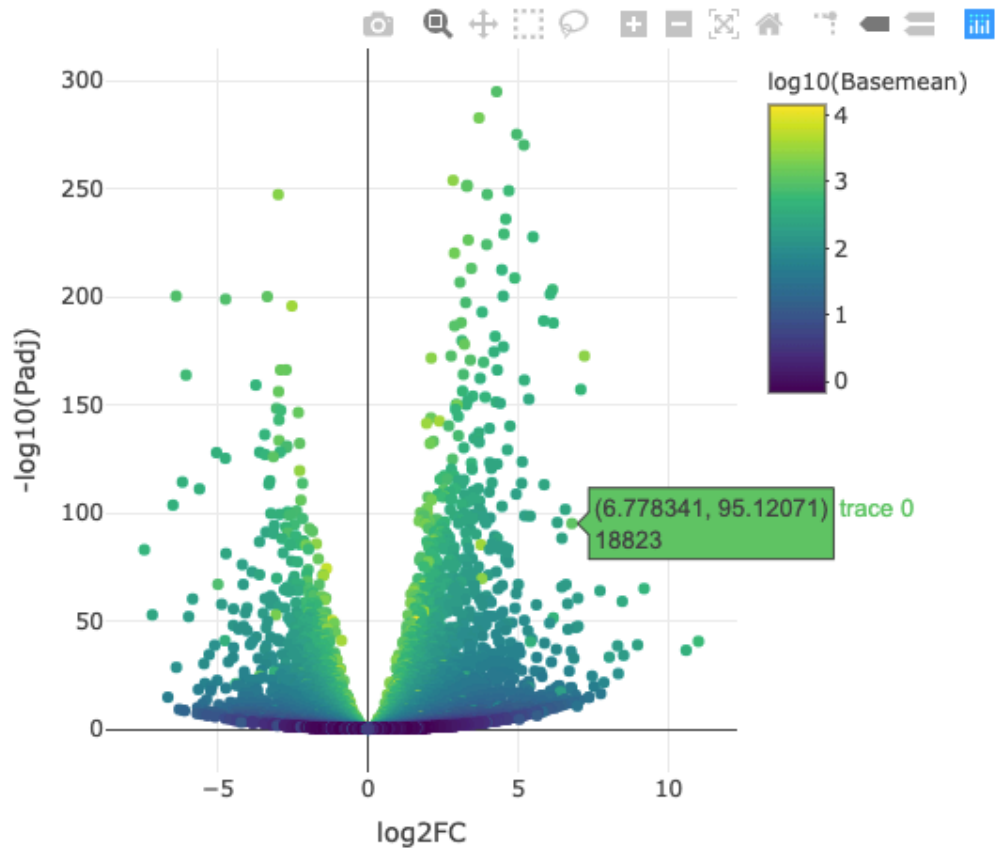
Global Environment

Data

ADAR2KO	15191 obs. of 8 variables

Files | Plots | Packages | Help | Viewer

Zoom | Export | Publish



Heatmaps

Package:pheatmap

- This is a package for making nice heatmaps
- install.packages("pheatmap")
- library(pheatmap)

Uploading data into R

- We will load the Deseq2 Normalized counts table
  - This is from our galaxy exercise
  - Can also be found in the github repo in the “data” folder
- Click the upload button
- Click “Choose File”
- We will need to find the file path for our data
  - Default: found in the download folder on your computer
- Click “ok”

You may experience delays running certain tools that require multiple cores as we have temporarily reduced capacity for these tools. In addition, the RNA STAR tool is not functioning properly.

Tools

search tools

- Get Data
- Collection Operations

GENERAL TEXT TOOLS

- Text Manipulation
- Filter and Sort
- Join, Subtract and Group
- Datamash

GENOMIC FILE MANIPULATION

- FASTA/FASTQ
- FASTQ Quality Control
- SAM/BAM
- BED
- VCF/BCF
- Nanopore
- Convert Formats
- Lift-Over

COMMON GENOMICS TOOLS

- Operate on Genomic Intervals
- Fetch Sequences/Alignments

GENOMICS ANALYSIS

- Assembly
- Annotation

1	2	3	4	5	6
	SRR531311_pass	SRR531312_pass	SRR531313_pass	SRR531314_pass	SRR
497097	811.428778836481	709.51971540483	509.20108281565	778.563165000904	581.42
100503874	36.1412243447283	15.1566294345491	24.2958767148621	36.556899879484	14.487
100038431	2.33169189320828	0	2.02465639290517	0.988024321067136	2.8974
19888	1.16584594660414	0.947289339659319	2.02465639290517	1.97604864213427	0.96582
20671	54.7947594903946	67.2575431158117	65.8013327694181	50.389240374424	138.11
27395	489.655297573739	467.013644452044	549.694210673754	523.652890165582	508.98
18777	258.817800146119	222.61299481994	196.391670111802	232.185715450777	401.78
100503730	3.49753783981242	5.68373603795592	2.02465639290517	4.94012160533568	
21399	823.087238302523	738.885684934269	666.111953265801	865.509305254812	734.9
58175	202.85719470912	245.347938971764	222.712203219569	259.850396440657	606.53
108664	1955.12365245514	1944.78501432058	1949.74410636768	2207.24633326398	3205.5
18387	24.4827648786869	17.0512081138677	16.1972511432414	27.6646809898798	260.77
226304	6.99507567962484	7.57831471727456	10.1232819645259	11.8562918528056	36.701
12421	2093.85932010103	1635.02140025199	1561.01007892989	2087.69539041486	2685.9
620393	1.16584594660414	0	1.01232819645259	0	
240690	290.295640704431	388.388629260321	216.638234040853	243.053982982516	230.83
319263	1449.14651162895	1215.37222278291	1159.11578493821	1511.67721123272	2747.7
71096	127.077208179851	119.358456797074	121.47938357431	171.916231865682	536.03
59014	577.093743569049	539.007634266153	521.349021173081	543.413376586925	487.74
76187	55.9606054369987	54.9427817002405	49.6040816261767	59.2814592640282	330.31
72481	6.99507567962484	3.78915735863728	8.09862557162068	10.8682675317385	20.28
76982	672.693111190589	567.426314455932	717.740691284883	744.970338084621	7373.0
17864	135.23812980608	118.411167457415	68.8383173587758	105.718602354184	75.334
70675	1240.4600871868	1153.79841570505	868.577592556318	1372.36578196225	1579
73331	0	0	1.01232819645259	0	3.8632
170755	116.584594660414	117.463878117756	120.467055377858	148.20364816007	471.32
620986	15.1559973058538	11.3674720759118	14.1725947503362	15.8083891370742	9.6582
240697	26.8144567718952	26.5241015104609	47.5794252332715	35.5688755584169	31.872
73824	361.412243447283	308.816324728938	313.821740900301	349.760609657766	143.90
266793	5.8292297330207	9.47289339659319	10.1232819645259	12.8443161738728	1.9316
100038398	1.16584594660414	0.947289339659319	3.03698458935776	2.96407296320141	10.624
100039596	1.16584594660414	0	0	0.988024321067136	0.96582
69312	13.9901513592497	2.84186801897796	6.07396917871551	3.95209728426855	13.521

History

search datasets

RNAseq_tutorial_deseq2

3 shown, 8 deleted

11.55 MB

3: Normalized counts file on data 59, data 57, and others

Adult | Embryonic

27,180 lines

format: **tabular**, database: **mm10**

primary factor: Age

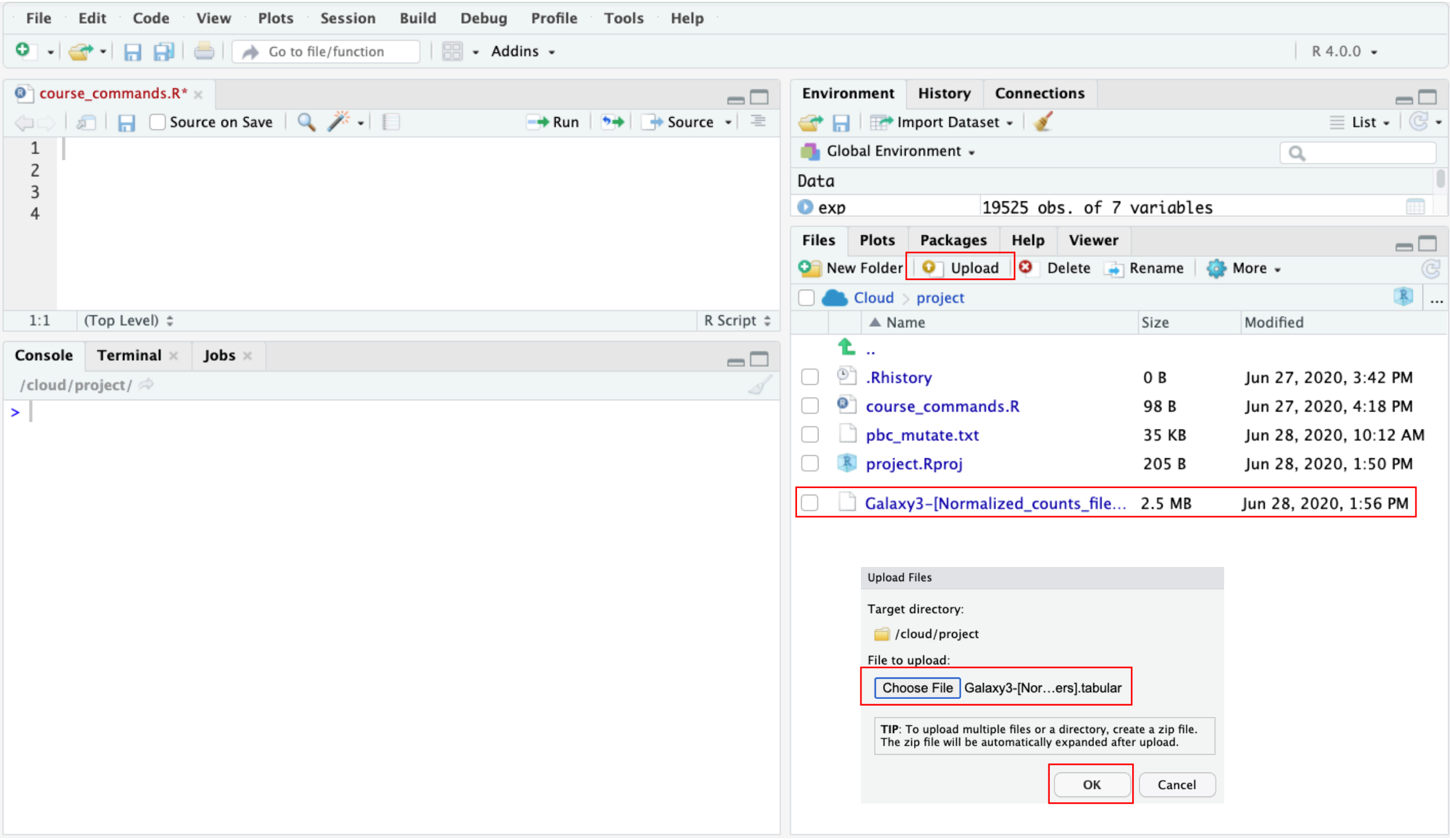
DESeq2 run information

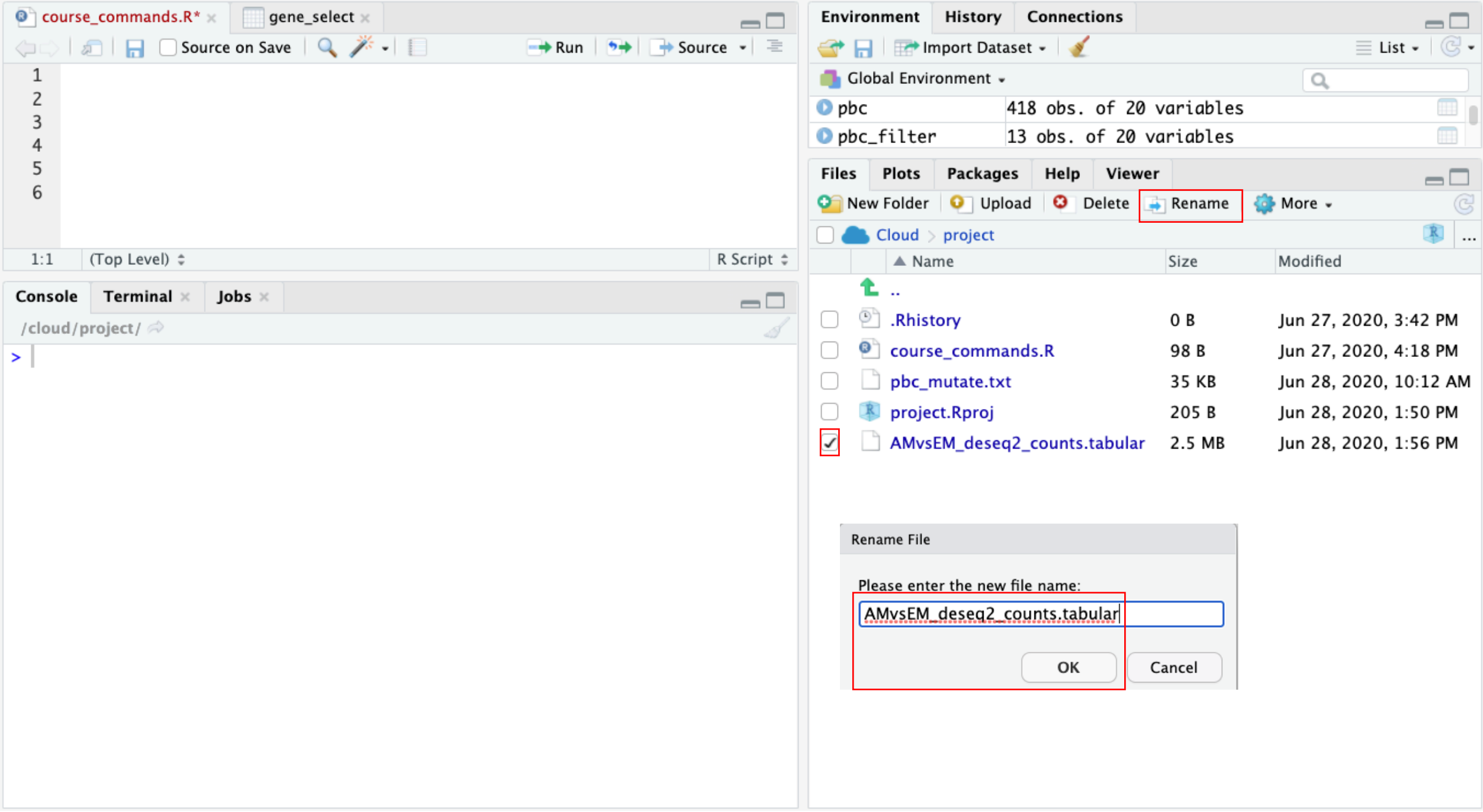
sample table:

Age
SRR531311_pass Embryonic
SRR531312_pass Embryonic
SRR531313_pass Embryonic
SRR531314_pass Embryonic
SRR531315_pass Adult
SRR531316_pass Adu

1	2	3
	SRR531311_pass	SRR531312_pc
497097	811.428778836481	709.5197154
100503874	36.1412243447283	15.156629434
100038431	2.33169189320828	0
19888	1.16584594660414	0.947289339

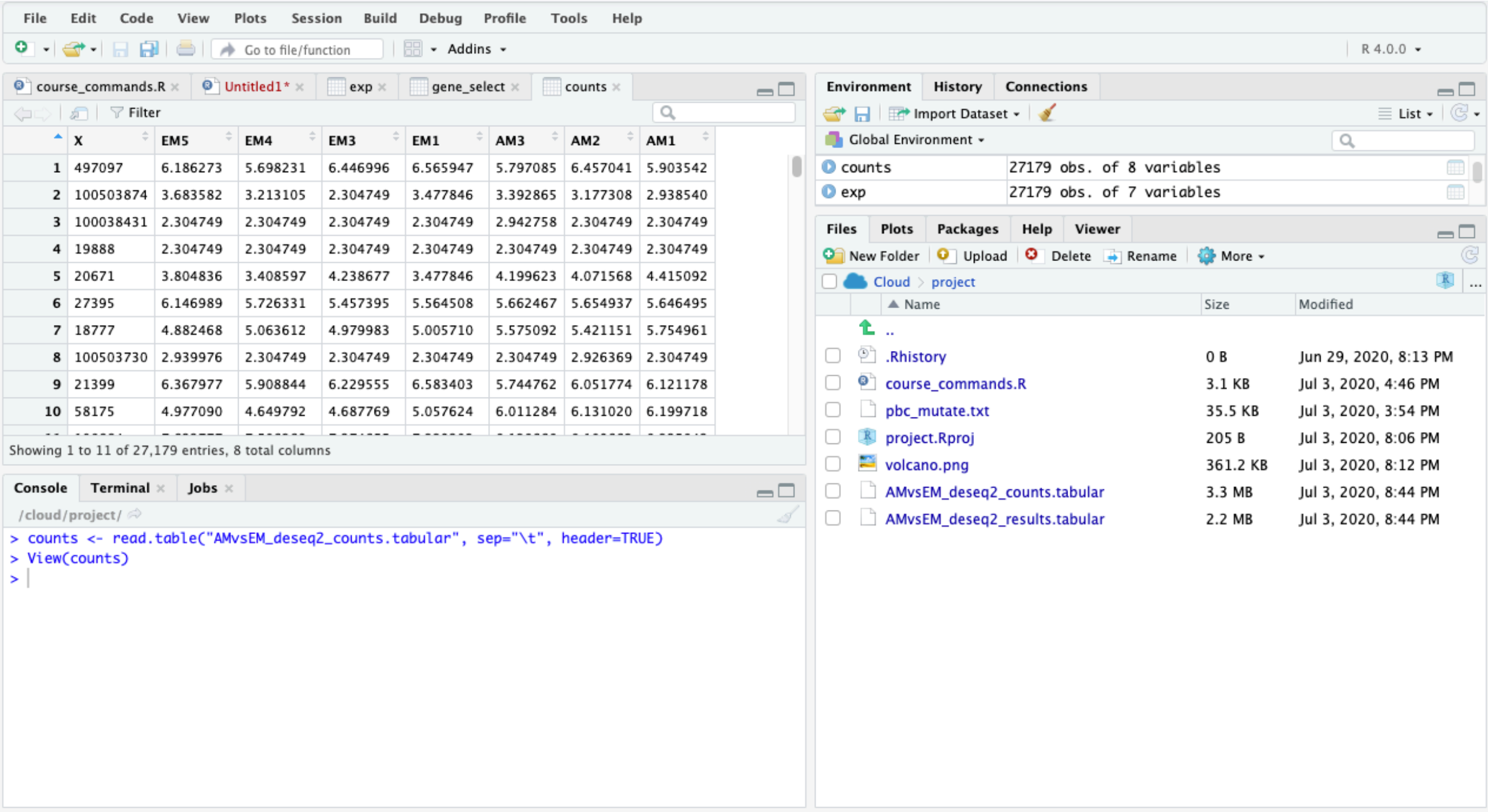
2: DESeq2 plots on dat


File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 4.0.0
course_commands.R*
Source on Save
Run
Source
1
2
3
4
1:1
(Top Level)
R Script
Console
Terminal
Jobs
/cloud/project/
>
Environment
History
Connections
Import Dataset
List
Global Environment
Data
exp
19525 obs. of 7 variables
Files
Plots
Packages
Help
Viewer
New Folder
Upload
Delete
Rename
More
Cloud
project
Name
Size
Modified
..
.Rhistory
0 B
Jun 27, 2020, 3:42 PM
course_commands.R
98 B
Jun 27, 2020, 4:18 PM
pbc_mutate.txt
35 KB
Jun 28, 2020, 10:12 AM
project.Rproj
205 B
Jun 28, 2020, 1:50 PM
Galaxy3-[Normalized_counts_file...
2.5 MB
Jun 28, 2020, 1:56 PM
Upload Files
Target directory:
/cloud/project
File to upload:
Choose File
Galaxy3-[Nor...ers].tabular
TIP: To upload multiple files or a directory, create a zip file.
The zip file will be automatically expanded after upload.
OK
Cancel


course_commands.R*
gene_select
Source on Save
Run
Source
1
2
3
4
5
6
1:1
(Top Level)
R Script
Console
Terminal
Jobs
/cloud/project/
>
Environment
History
Connections
Import Dataset
List
Global Environment
pbc
418 obs. of 20 variables
pbc_filter
13 obs. of 20 variables
Files
Plots
Packages
Help
Viewer
New Folder
Upload
Delete
Rename
More
Cloud
project
Name
Size
Modified
..
.Rhistory
0 B
Jun 27, 2020, 3:42 PM
course_commands.R
98 B
Jun 27, 2020, 4:18 PM
pbc_mutate.txt
35 KB
Jun 28, 2020, 10:12 AM
project.Rproj
205 B
Jun 28, 2020, 1:50 PM
AMvsEM_deseq2_counts.tabular
2.5 MB
Jun 28, 2020, 1:56 PM
Rename File
Please enter the new file name:
AMvsEM_deseq2_counts.tabular
OK
Cancel

Reading in data table into R

- Use the read.csv command
- counts <- read.table("AMvsEM_deseq2_counts.tabular", sep="\t", header=TRUE)
- View(counts)


File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
R 4.0.0
course_commands.R
Untitled1*
exp
gene_select
counts
Filter
X EM5 EM4 EM3 EM1 AM3 AM2 AM1
1 497097 6.186273 5.698231 6.446996 6.565947 5.797085 6.457041 5.903542
2 100503874 3.683582 3.213105 2.304749 3.477846 3.392865 3.177308 2.938540
3 100038431 2.304749 2.304749 2.304749 2.304749 2.942758 2.304749 2.304749
4 19888 2.304749 2.304749 2.304749 2.304749 2.304749 2.304749 2.304749
5 20671 3.804836 3.408597 4.238677 3.477846 4.199623 4.071568 4.415092
6 27395 6.146989 5.726331 5.457395 5.564508 5.662467 5.654937 5.646495
7 18777 4.882468 5.063612 4.979983 5.005710 5.575092 5.421151 5.754961
8 100503730 2.939976 2.304749 2.304749 2.304749 2.304749 2.926369 2.304749
9 21399 6.367977 5.908844 6.229555 6.583403 5.744762 6.051774 6.121178
10 58175 4.977090 4.649792 4.687769 5.057624 6.011284 6.131020 6.199718
Showing 1 to 11 of 27,179 entries, 8 total columns
Console Terminal Jobs
/cloud/project/
> counts <- read.table("AMvsEM_deseq2_counts.tabular", sep="\t", header=TRUE)
> View(counts)
>
Environment History Connections
Import Dataset
List
Global Environment
counts 27179 obs. of 8 variables
exp 27179 obs. of 7 variables
Files Plots Packages Help Viewer
New Folder Upload Delete Rename More
Cloud project
Name Size Modified
..
.Rhistory 0 B Jun 29, 2020, 8:13 PM
course_commands.R 3.1 KB Jul 3, 2020, 4:46 PM
pbc_mutate.txt 35.5 KB Jul 3, 2020, 3:54 PM
project.Rproj 205 B Jul 3, 2020, 8:06 PM
volcano.png 361.2 KB Jul 3, 2020, 8:12 PM
AMvsEM_deseq2_counts.tabular 3.3 MB Jul 3, 2020, 8:44 PM
AMvsEM_deseq2_results.tabular 2.2 MB Jul 3, 2020, 8:44 PM

Selecting genes of interest for heatmap

- gene_select <- as.data.frame(filter(counts, X %in% c("75426","71951","53321", "66425")))
- View(gene_select)
- rownames(gene_select) <- gene_select$X
- pheatmap(gene_select[2:8])

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | R 4.0.0

course_commands.R | exp | gene_select | counts

Filter

	X	EM5	EM4	EM3	EM1	AM3	AM2	AM1
66425	66425	3.388259	3.569766	3.367829	3.271022	8.524712	8.593079	8.285984
53321	53321	3.804836	3.831882	3.885372	3.794239	9.344727	9.337682	9.171662
75426	75426	9.566874	9.699907	9.901679	9.672746	3.392865	3.882366	3.385881
71951	71951	9.484966	9.644426	9.582395	9.504735	4.021039	3.882366	4.102799

Showing 1 to 4 of 4 entries, 8 total columns

Console | Terminal | Jobs

/cloud/project/

```
> gene_select <- as.data.frame(filter(counts, X %in% c("75426","71951","53321", "66425")))
> View(gene_select)
> rownames(gene_select) <- gene_select$X
> pheatmap(gene_select[2:8])
>
```

Environment | History | Connections

Import Dataset | List

Global Environment

exp	27179 obs. of 7 variables
gene_select	4 obs. of 8 variables

Files | Plots | Packages | Help | Viewer

Zoom | Export | Publish