

R for Research Scientists

Daniel Lumian, PhD
Data Scientist
Office of Data Science Strategy

Updated from work by Allissa Dillman, PhD

About me

PH.D. IN PSYCHOLOGY

DATA SCIENCE BOOTCAMP

DATA SCIENTIST

DATA SCIENCE INSTRUCTOR

DATA SCIENTIST AT NIH

What we will learn today!

- A tour of R studio and R Studio Cloud
- 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
- Saving your work

What is

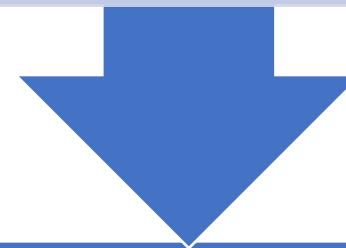


A programming language and open-source software environment that can:

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 is Rstudio Cloud



Use Rstudio in a browser, no
downloading software



Share projects easily for analysis
and teaching



<https://rstudio.cloud/>

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

Let's get started!

- You may need to log in to your Rstudio cloud account first
 - Go to <https://rstudio.cloud/>
 - Click log in
- <https://rstudio.cloud/project/4217437>
- https://github.com/dlumian/R_Intro_scRNA
- First click on Save a permanent copy
- Click on the new file icon
- Select R script



File Edit Code View Plots Session Build Debug Profile Tools Help

+ | - | ↗ | ↘ | Go to file/function | Addins | R 4.0.3

Console Terminal × Jobs ×
/cloud/project/ ↗

R version 4.0.3 (2020-10-10) -- "Bunny-Wunnies Freak Out"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |

Environment History Connections Tutorial
Import Dataset | List | C | Global Environment | Search | Environment is empty

Files Plots Packages Help Viewer
New Folder Upload Delete Rename More | Cloud > project | R ...

	Name	Size	Modified
<input type="checkbox"/>	..		
<input type="checkbox"/>	.Rhistory	0 B	Nov 3, 2020, 4:4
<input type="checkbox"/>	AMvsEM_deseq2_results.csv	2.2 MB	Nov 3, 2020, 4:5
<input type="checkbox"/>	project.Rproj	205 B	Nov 3, 2020, 8:5
<input type="checkbox"/>	qRT_PCR_val.csv	273 B	Nov 3, 2020, 4:5
<input type="checkbox"/>	RforResearchSci.Rmd	17.9 KB	Nov 3, 2020, 4:5

The screenshot shows the RStudio interface with several red annotations:

- A red arrow points to the "R Script" option in the "File" menu, with the text "Write code here" overlaid.
- The code editor area has the text "Environment is empty" and "Objects go here" overlaid.
- The "Files" browser area has the text "Written files go here" overlaid.

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

Type 'demo()' for some demos, 'help.start()' for an HTML browser. Type 'q()' to quit R.

> |Run code here

R Script Write code here

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

Environment History Connections

Import Dataset Global Environment

Environment is empty

Objects go here

Files Plots Packages Help Viewer

New Folder Upload Delete Rename More

Cloud > project

	Name	Size	Modified
<input type="checkbox"/>	..		
<input type="checkbox"/>	.Rhistory	0 B	Jun 27, 2020, 3:42 PM
<input type="checkbox"/>	project.Rproj	205 B	Jun 27, 2020, 3:42 PM

Written files go here

R is like a calculator

- Enter the following

7+7

a=3

b=5

- If I type a it will print out the contents of a which is 3

a+b

a-b

a*b

a/b

The screenshot shows the RStudio interface with several panels:

- Code Editor:** An R script named "Untitled1" is open. A red box highlights the section of code from line 2 to line 8:

```
2 a=3  
3 b=5 Highlight the section of  
4 a code you want to run  
5 a+b  
6 a-b then click run  
7 a*b  
8 a/b
```
- Toolbar:** The "Run" button is highlighted with a red box.
- Environment:** The "Global Environment" pane shows the values of variables "a" and "b":

Values
a 3
b 5

A red box highlights this table.
- Console:** The output of the executed code is shown:

```
> 7+7  
[1] 14  
> a=3  
> b=5  
> a  
[1] 3  
> a+b  
[1] 8  
> a-b  
[1] -2  
> a*b  
[1] 15  
> a/b  
[1] 0.6  
>
```
- File Explorer:** The "Cloud > project" pane lists files in the project:

Name	Size	Modified
..	0 B	Jul 11, 2020, 5:50 PM
.Rhistory	0 B	Jul 11, 2020, 5:50 PM
project.Rproj	205 B	Jul 11, 2020, 5:50 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
 - `c` stands for combine
- Don't know how to use it?
- Type : `?c`
 - Putting a `?` before a command or object will retrieve the help for that objects

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

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(...)
```

```
## Default S3 method:
```

Typing ?

- Typing “?” then the command name will give you help on the command name
- Typically, examples you can try may be available at the bottom of the help information
- 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

R Untitled1*

Source on Save Run Source

```
1 c(5,6,7,8)
2 d <- c(5,6,7,8)
3 d
```

Environment History Connections

Import Dataset 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(...)
```

```
## Default S3 method:
```

Make more vectors

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



Untitled1*

```
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

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



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

Untitled1* Go to file/function Addins R 4.0.0

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 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

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)`

Untitled1*

```
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 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

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)`

The screenshot shows the RStudio interface with the following components:

- Code Editor:** Shows the script `Untitled1*` containing the R command `h <- rbind(d,e,f,g)`. The cursor is at the end of the line.
- Environment View:** Displays the global environment with objects `h`, `c`, `d`, `e`, `f`, and `g`.
 - `h`: num [1:4, 1:4] 5 11 1 1 6 12 2 2 7 13 ...
 - `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
- Console View:** Shows the output of the command `h <- rbind(d,e,f,g)` followed by the display of matrix `h`.

```
> h <- rbind(d,e,f,g)
> h
     [,1] [,2] [,3] [,4]
d      5     6     7     8
e     11    12    13    15
f      1     2     3     4
g      1     2     3    15
```
- Search Results View:** Shows search results for "standard deviation".
 - Search Results:** A large blue "R" logo.
 - Help pages:**
 - [nlme:::pooledSD](#) Extract Pooled Standard Deviation
 - [stats::sd](#) Standard Deviation
 - [stats::sigma](#) Extract Residual Standard Deviation 'Sigma'
 - Message:** The search string was "standard deviation"

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")`

Untitled1*

```
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 Global Environment

Data

h	num [1:4, 1:4] 5 11 1 1 6 12 2 2 7 13 ...
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

Values

Files Plots Packages Help Viewer

R: Search Results Find in Topic

standard devia

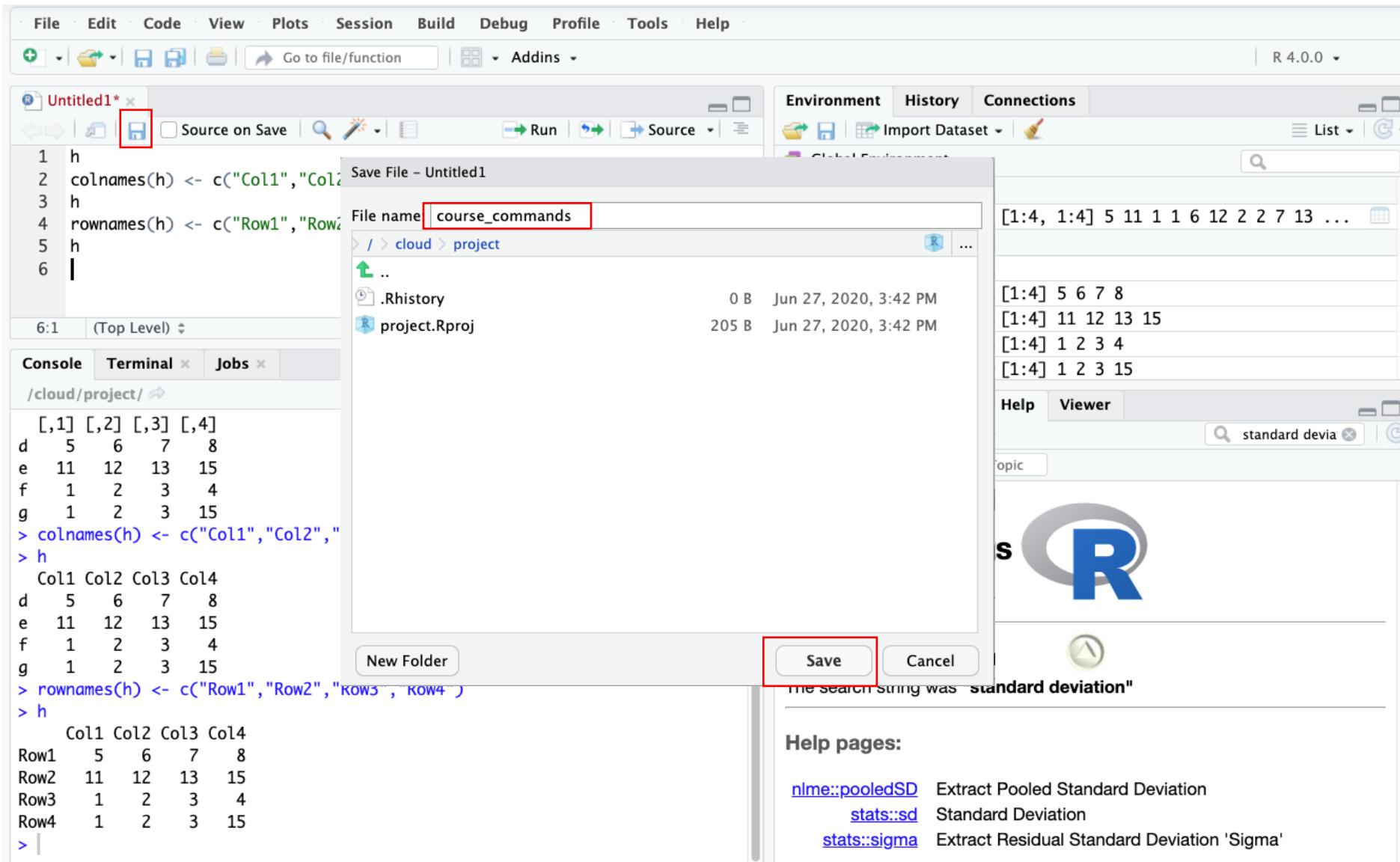
Search Results R

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

```
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 Global Environment List C

Data

h	num [1:4, 1:4]	5 11 1 1 6 12 2 2 7 13 ...
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

Values

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

Questions?

- So far, we have:
 - Accessed R via R Studio or R Studio Cloud
 - Created a script
 - Done basic math operations
 - Created and used vectors
 - Done vector math
 - Created a dataframe with named rows and cols
- Up next:
 - Data wrangling!

Data wrangling!



WE COULD USE BASE R



OR THE TIDYVERSE
COLLECTION OF PACKAGES



SPECIFICALLY DESIGNED
FOR DATA SCIENCE

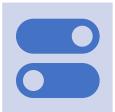
What is a package



A collection of R functions, complied code and data



Saved in a directory called “library”



Can be turned on and off

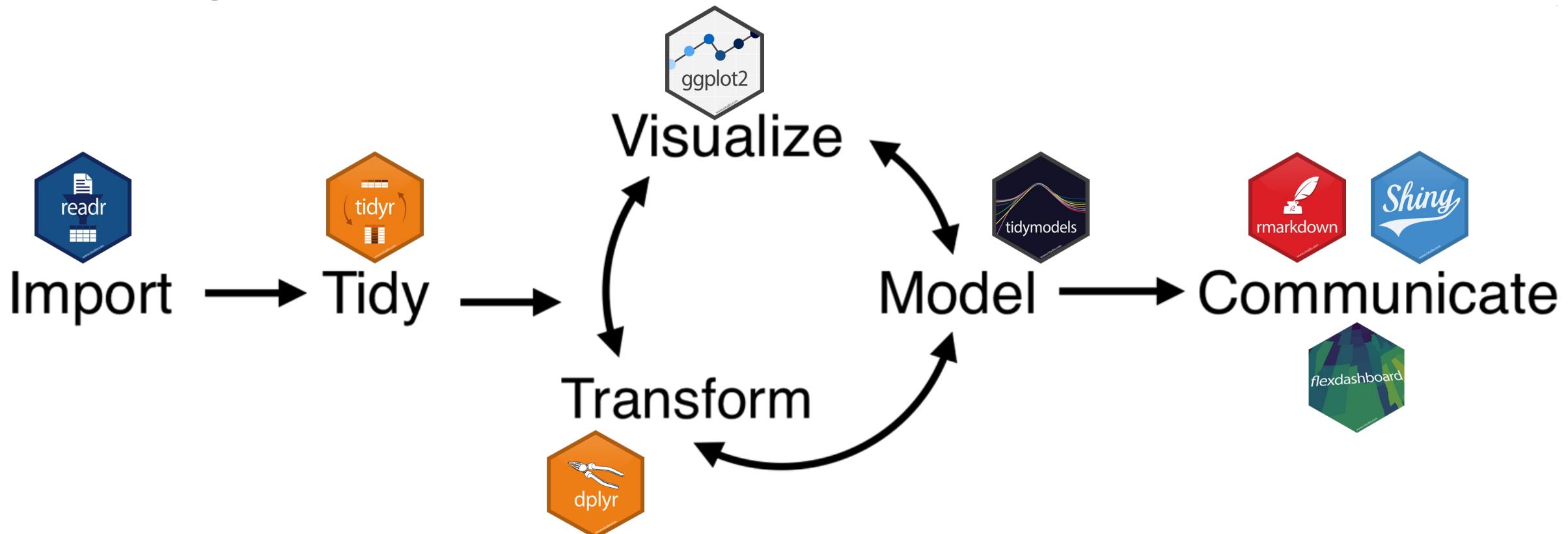


Made by different people commands may clash



Would also be very slow to load everything every time

Tidyverse



First time install:
`install.packages("tidyverse")`

Turn on package set: `library(tidyverse)`

Let's get started!

- Click on the Intro_R.Rmd
- The first dataset we will use is from the Mayo Clinic trial in primary biliary cholangitis (PBC) of the liver
- It is found within the survival package

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") **Install like this**

2

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

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

library(tidyverse) 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 OR this

✓ 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()

Environment History Connections

Import Dataset 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
<input checked="" type="checkbox"/> tidyverse	Easily Install and Load the 'Tidyverse'	1.3.0
<input type="checkbox"/> 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*

Source on Save Run Source

```
1 library(survival)
2
3
```

3:1 (Top Level) R Script

Console Terminal Jobs

/cloud/project/

```
> library(survival)
```

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>

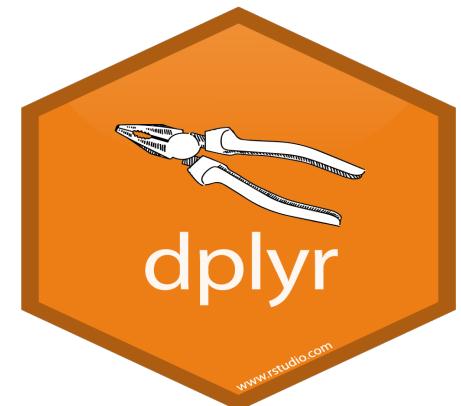
Packages

Install Update Packrat

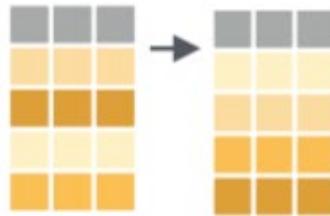
survival

Name	Description	Version
<input checked="" type="checkbox"/> survival	Survival Analysis	3.1-12

dplyr: Transform your data



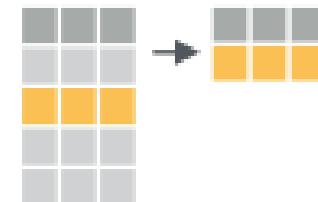
`arrange()`



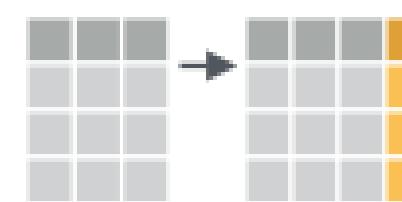
`select()`



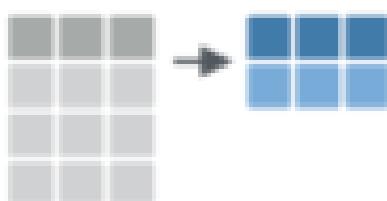
`filter()`



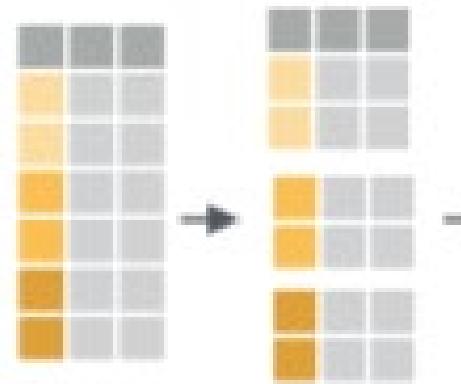
`mutate()`



`summarize()`



`group_by()`



Basic structure:
dpylr function(data, specifics)
Example:
`arrange(pbc, age)`

Save a table as a file

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

The screenshot shows the RStudio interface with the following components:

- Code pane:** Displays the R script with the command `write.table(pbc_mutate,"pbc_mutate.txt",row.names=F,sep="\t")`.
- Environment pane:** Shows the global environment with objects `h`, `pbc`, `pbc_filter`, `pbc_mutate`, `pbc_select`, and `pbcseq`.
- Console pane:** Displays the output of the command being run.
- Files pane:** Shows the contents of the project directory, including `.Rhistory`, `course_commands.R`, `project.Rproj`, and `pbc_mutate.txt`, with `pbc_mutate.txt` highlighted by a red box.

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:

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

File Edit Code View Plots Session Build Debug Profile Tools Help

Untitled1* pbc_final Go to file/function Addins R 4.0.0

sex new_col
1 m 55.71072
2 f 50.15694

Filter

Showing 1 to 2 of 2 entries, 2 total columns

Console Terminal Jobs

```
> pbc_final <- pbc %>% group_by(sex) %>% summarise(new_col = mean(age))  
`summarise()` ungrouping output (override with `.`groups` argument)  
> View(pbc_final)  
>
```

Environment History Connections

Import Dataset Global Environment

ave_age 1 obs. of 1 variable
pbc 418 obs. of 20 variables
pbc_final 2 obs. of 2 variables

Files Plots Packages Help Viewer

Install Update Packrat

Name	Description	Version
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 Objects into Tidy Tibbles	0.7.0
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

R Live Demo-Data Wrangling

- Using the pre-made R Notebook `Intro_R.Rmd` we will explore the following data manipulations:
 - Arrange
 - Select
 - Filter
 - Mutate
 - Summarize
 - GroupBy

Questions?

- What data manipulation options are available?
- Which operation would you use to create a new column of data?
- Which two options are available for arranging data?
- What symbol is used to join multiple commands?



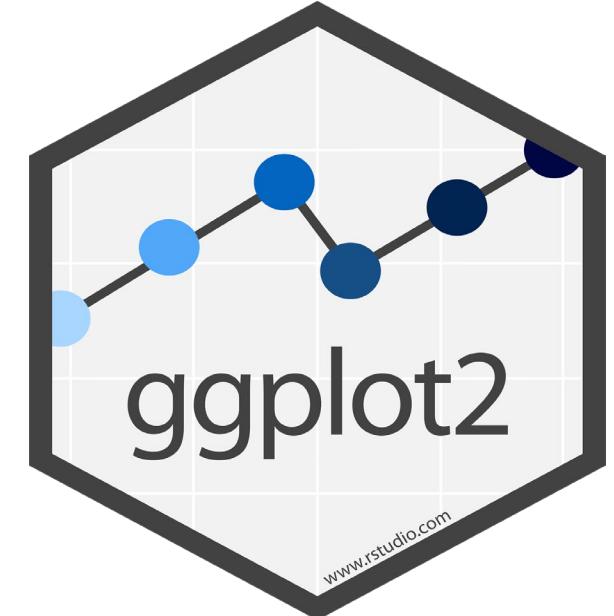
Data visualization in R

Visualizations for Today

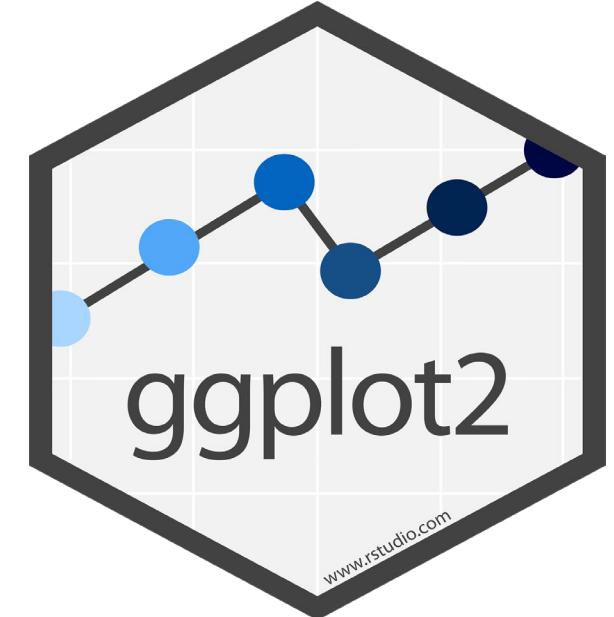
- Bar plot
- Histogram
- Density Plot
- Scatter plot
- Faceted plots (multiple plots in the same figure)
- And more!

ggplot2: Visualize your data

- 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: Visualize your data



What you need:

1. A data 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()
```

Example:

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

Aesthetic mapping options

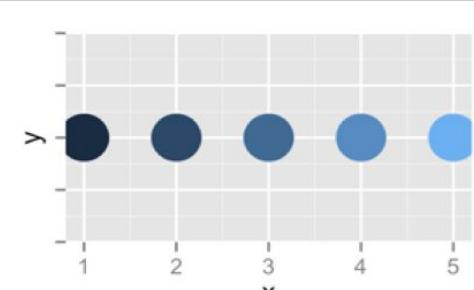
Color

Discrete

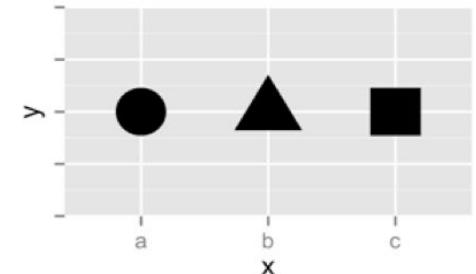
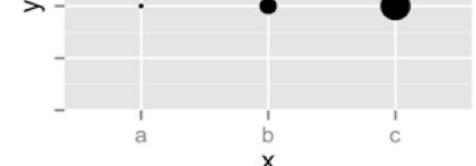


Size

Continuous



Shape



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



`t + facet_grid(cols = vars(f1))`
facet into columns based on f1



`t + facet_grid(rows = vars(year))`
facet into rows based on year



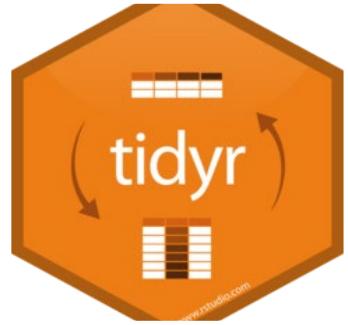
`t + facet_grid(rows = vars(year), cols = vars(f1))`
facet into both rows and columns



`t + facet_wrap(vars(f1))`
wrap facets into a rectangular layout



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



→

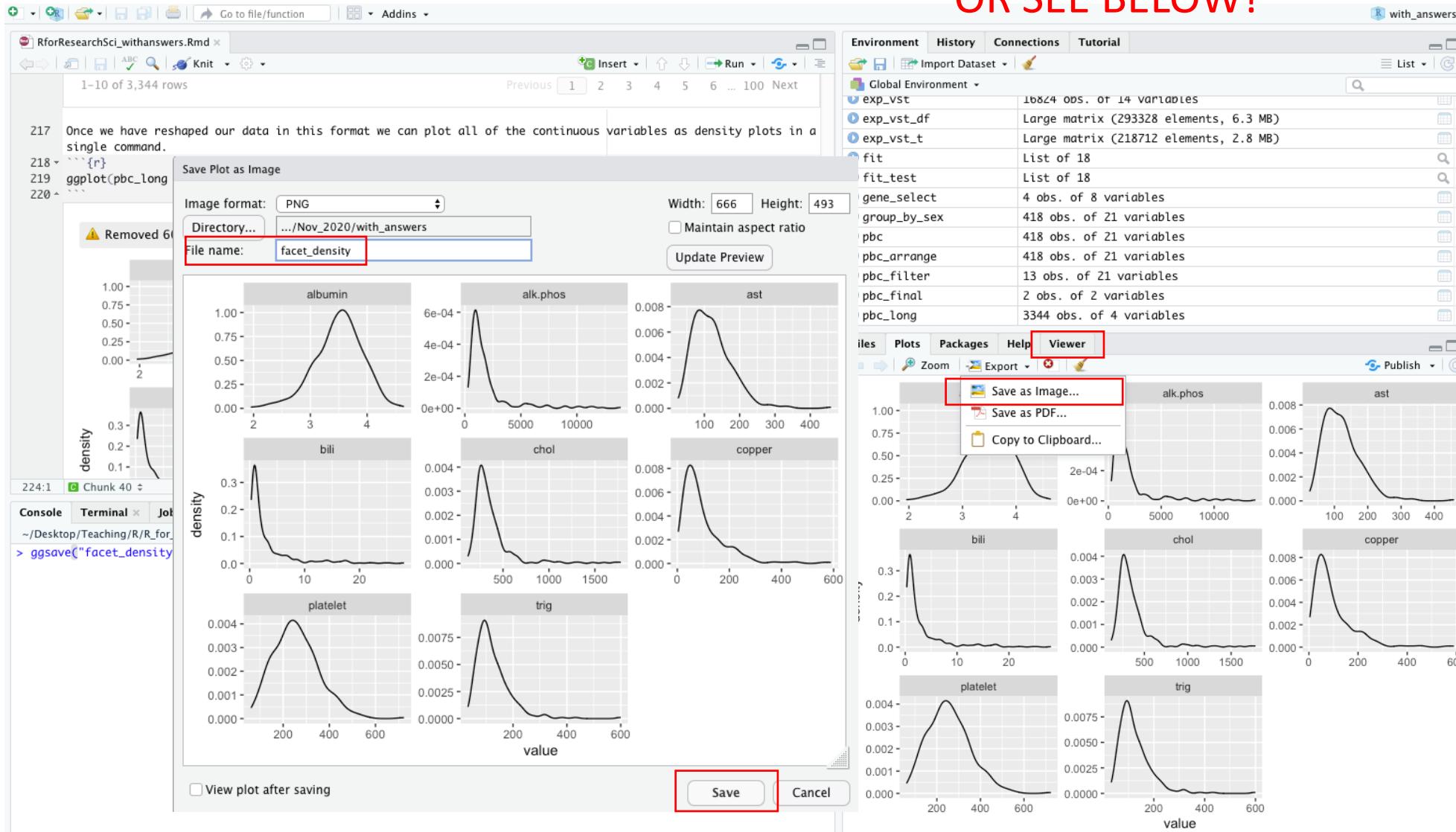
country	1999	2000
A	0.7K	2K
B	37K	80K
C	212K	213K

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

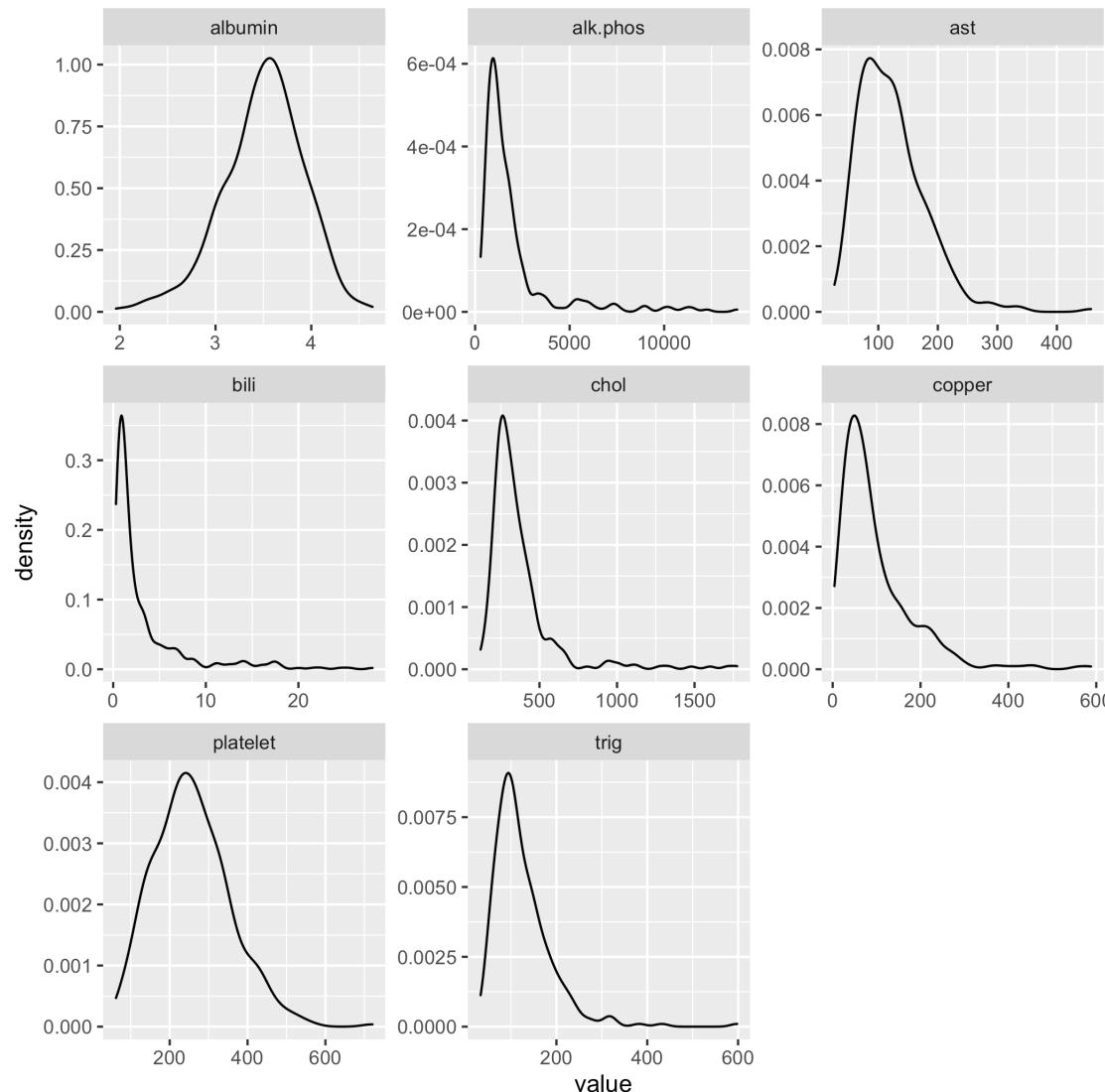
Saving your beautiful graph!

`ggsave("facet_density.png")`

OR SEE BELOW!



Questions on Basic Visualizations



Transferring to your local R instance

1. Select all the files in the workspace
2. Click more on the lower right panel
3. Click export
4. Click download
5. Unzip this file on your local computer
6. Open the folder and click on the file ending in “.Rproj”
 - This will automatically open Rstudio with all of the files needed
7. Once in Rstudio click on the file ending in “.Rmd” for the notebook
8. You will need to install all of the packages ahead of time
 - `install.packages("tidyverse")`
 - `install.packages("ggfortify")`
 - `install.packages("plotly")`
 - `install.packages("ggThemeAssist")`

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins R 4.0.3

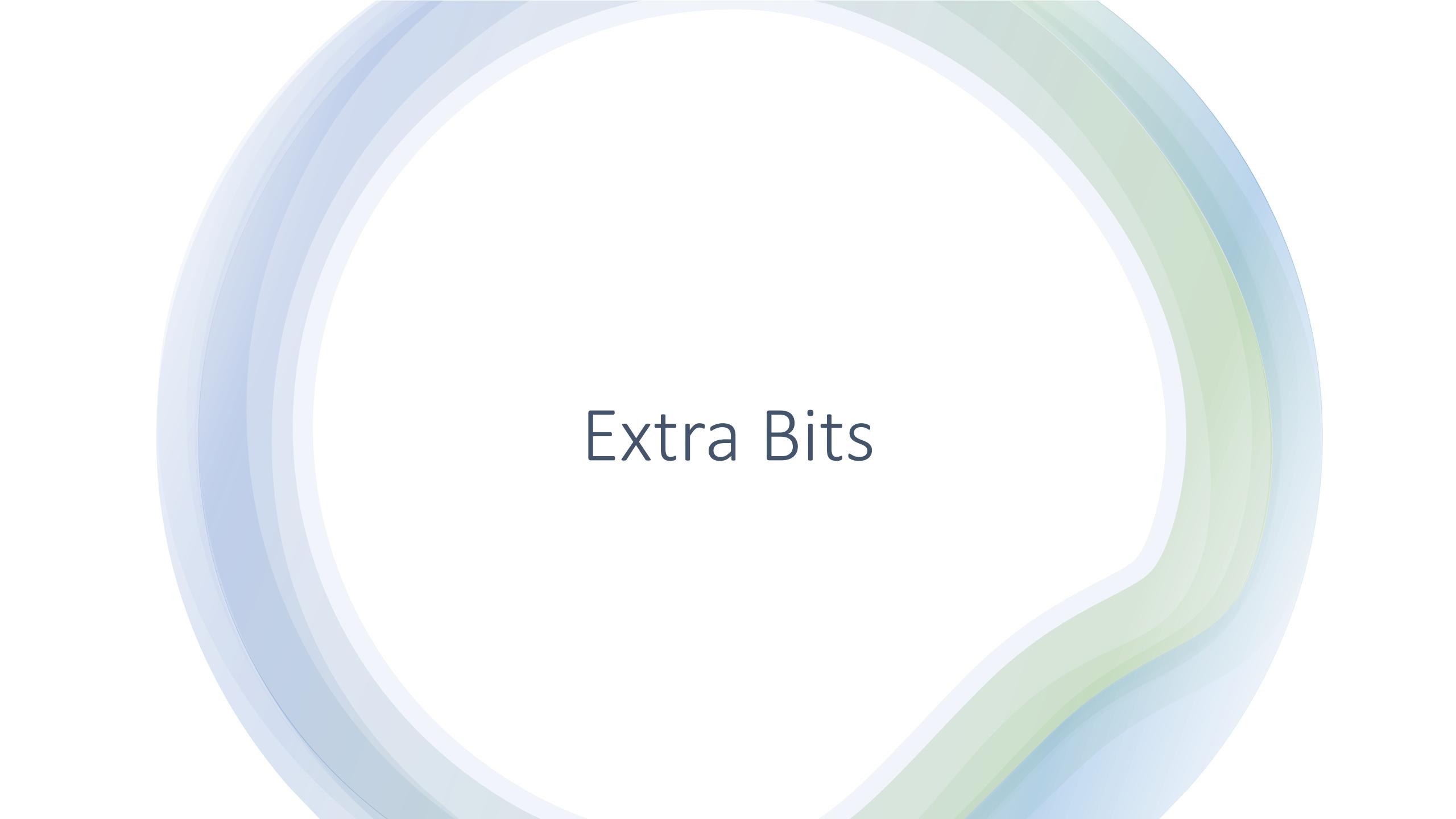
RforResearchSci.Rmd x Knit

```
1 ---  
2 title: "R Notebook"  
3 output:  
4   html_document  
5     df_print:  
6 ---  
7 ---  
8 title: "Data"  
9 output:  
10  html_document  
11    df_print:  
12 ---  
13 ---  
14 This is an [F]  
notebook, the  
15 Try executing this chunk by clicking the Run button within the chunk or by placing your cursor  
inside it and pressing *Cmd+Shift+Enter*.  
16  
17 # Data Transformation  
18 Running this in a new project or on your own computer for the first time you would need to run  
install.packages("tidyverse"). Thereafter you only need to turn on the package. You will also  
be using data from the survival package for this exercise, so you need to load this package as  
well.  
19 ``{r}  
20 library(survival)  
21 library(tidyverse)  
22 ``# R Notebook
```

Export Files
The selected file(s) will be downloaded to your computer. Please specify a name for the downloaded file:
rstudio-export.zip 4
Download Cancel

Environment History Connections Tutorial
Import Dataset Global Environment
Environment is empty

Files Plots Packages Help Viewer
New Folder Upload Delete Rename More 2
Cloud > project
Name
..
.Rhistory
AMvsEM_deseq2_results.csv
project.Rproj
qRT_PCR_val.csv
RforResearchSci_nov_2020.pptx
RforResearchSci.Rmd
1
1
Copy...
Copy To...
Move...
Copy Folder Path to Clipboard
Export... 3
Set As Working Directory
Go To Working Directory
Open New Terminal Here
Show Hidden Files



Extra Bits

Common biodata visualizations

- ggfortify: a ggplot extension for making PCA, survival curves, clustering and more
- pheatmap: a package for making nice heatmaps
- Volcano plots: just need ggplot
- Gene selected box plots

Streaming DEseq2 normalized counts from GEO to R

- We will load the Deseq2 normalized counts table
 - Data will be obtained from the gene expression omnibus :
<https://www.ncbi.nlm.nih.gov/geo/>
 - RNAseq of the developing mouse cerebral cortex
 - 3 Adult, 4 embryonic brains
- `vst <-
read_tsv("ftp://ftp.ncbi.nlm.nih.gov/geo/series/GSE39nnn/GSE
39866/suppl/GSE39866_Mouse.VarianceStabilized.Datagene.txt.
gz")`
- `View(vst)`

Volcano Plot

Loading DEseq2 results into R

- The data has already been uploaded to the shared workspace
- This is from the same dataset of embryonic and adult mouse brains
 - This is differential expression results created using the DEseq2 package in R
- Click the upload button
- Click “Choose File”
- We will need to find where the file lives on your computer
- Default: found in the download folder
- Click “ok”
- `exp <- read_csv("AMvsEM_deseq2_results.csv")`
- `exp`

Screenshot of RStudio interface showing the process of uploading a CSV file to a project.

The interface includes:

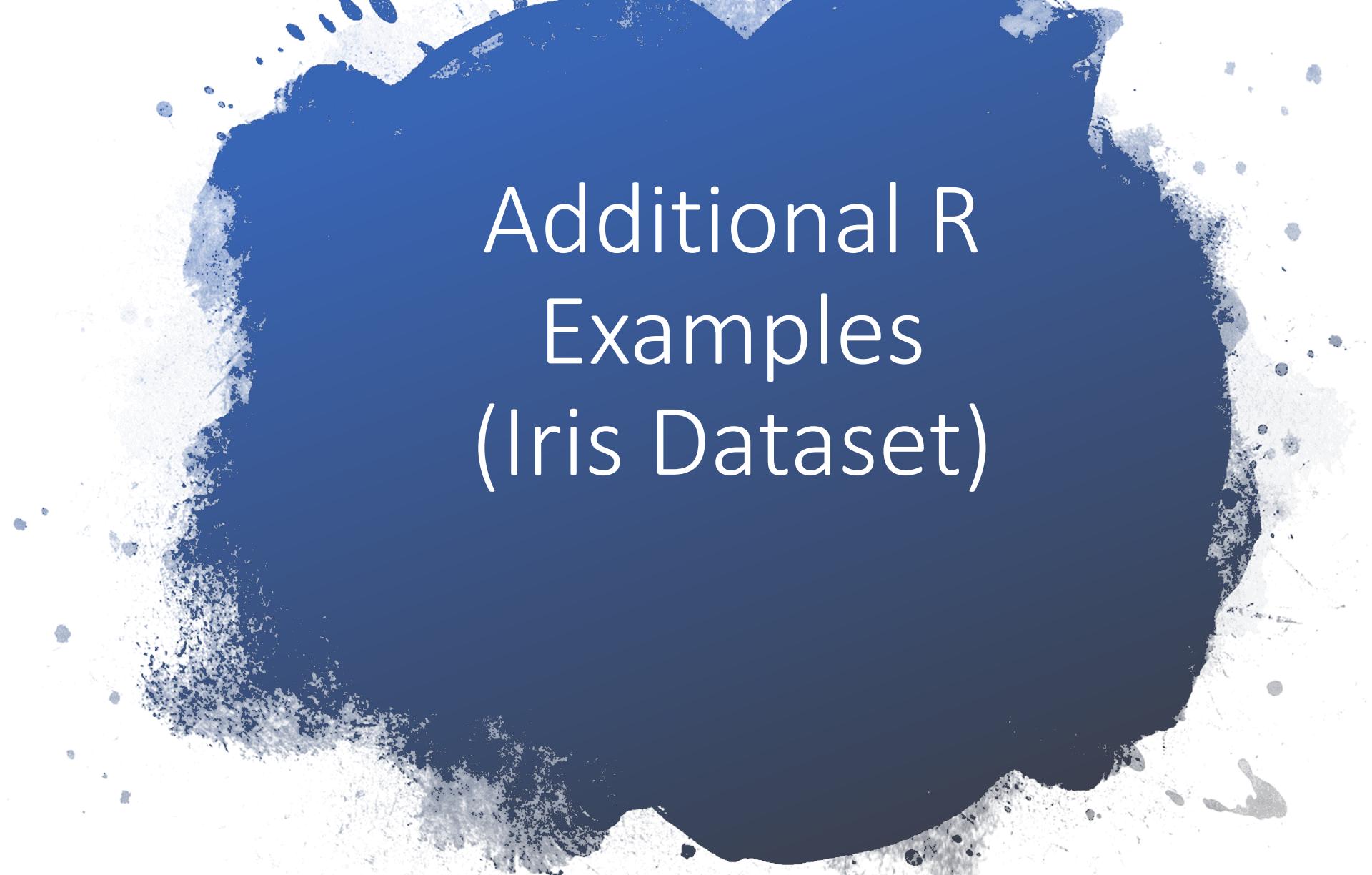
- File**, **Edit**, **Code**, **View**, **Plots**, **Session**, **Build**, **Debug**, **Profile**, **Tools**, **Help** menus.
- A toolbar with icons for New Project, Open Project, Save Project, Print, Go to file/function, and Addins.
- A status bar showing "R 3.6.3".
- An environment pane showing datasets: pbc (418 obs. of 20 variables), pbc_long (3344 obs. of 3 variables), and pbcseq (1945 obs. of 19 variables).
- A global environment pane.
- A code editor with tabs for pbc, pbc_long, and Untitled1.
- A console tab showing the path /cloud/project/.
- An R Script tab.
- A terminal tab.
- A jobs tab.
- An "Upload" button in the project pane, highlighted with a red box.
- A file list in the project pane showing .Rhistory, project.Rproj, and AMvsEM_deseq2_results.csv (2.2 MB, modified Jul 23, 2020, 8:14 PM).
- An "Upload Files" dialog box:
 - Target directory: /cloud/project (Browse... button)
 - File to upload: Choose File (highlighted with a red box) containing AMvsEM_deseq2_results.csv
 - TIP: To upload multiple files or a directory, create a zip file. The zip file will be automatically expanded after upload.
 - OK and Cancel buttons (highlighted with red boxes).



Making an
interactive plot!

Interactive Volcano with Plotly

- `install.packages("plotly")`
- `library(plotly)`
- `plot_ly(data = exp, x = ~log2FoldChange, y = ~-log10(padj), color=~log10(baseMean), type = 'scatter', mode='markers', text= ~gene_id)`



Additional R Examples (Iris Dataset)