

# INTRODUCTION TO R

Danny Lumian, Ph.D.  
Senior Data Scientist  
Saturday, June 24, 2023



# About me

- Ph.D. in Psychology from University of Denver (2018)
- Galvanize data science bootcamp (2018)
- Instructor @ Woz-U data science bootcamp (2021)
- NIH contractor with ODSS as a Data Science Training Specialist (2022)
- Senior Data Scientist at Daybreak LLC contracting with FDA (2023)
- Started programming in Python in 2011
- Experience with Python, R, MATLAB, SPSS, SQL, AFNI, SPM
- Neuroimaging data (fMRI), skin conductance, eye-tracking, experiment design and more
- Support free and open-source software (FOSS) like R and Python to create reproducible analysis and workflows

# Overview

- What is R
  - R platforms
    - R
    - Posit/R Studio
    - Posit Cloud/R Studio Cloud
  - R file types
1. Basic R commands
  2. R packages and built-in datasets
  3. Data manipulation
  4. Data visualization
  5. Data import and export
  6. Data analysis

Each of the 6 lessons has a separate notebook.

We will **not** be able to cover all of them today.

All notebooks remain in the repository for future reference and self-review.

---

# Tips and Tricks

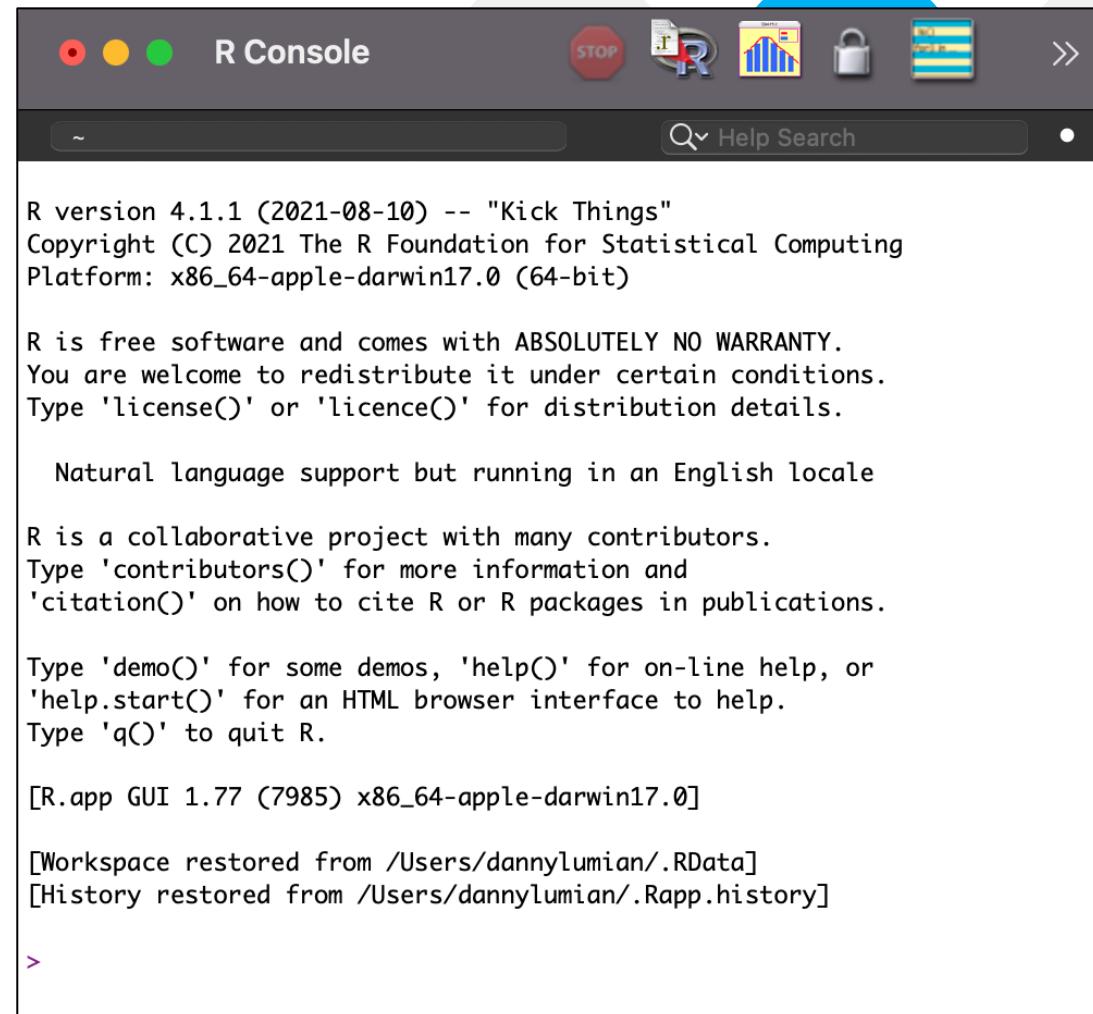
- R programming has much in common with other programming languages such as Python
  - Each has unique packages but often similar functionality
  - Packages are essential to getting the most out of your effort
- Documentation-Check it and write it!
- Error messages are designed to help with debugging
- There are MANY helpful resources available
  - Use them to level up your skills
  - Use them to help solve issues (StackOverflow)

# Tech Stack/Programming Ecosystem

- R allows for many common data operations, however, to be a complete programmer often requires additional skillsets
  - Below are common skillsets that enhance your ability to do your work as a researcher, developer or data scientist
    - Version Control: GitHub, GitLab
    - Database: SQL
    - Environments: Allows others to replicate and use your scripts
    - Remote compute and cloud services: Vital skill when data is too large for your personal workstation
    - Dashboards: R Shiny, Tableau
-

# What is R?

- An open-source programming language and software environment that can:
  - Manipulate data
  - Visualize data
  - Perform statistics
- Free and relatively easy to run in any environment (Unix, Windows, & MacOS)
- Recently changed to Posit
- <https://www.r-project.org/>



R version 4.1.1 (2021-08-10) -- "Kick Things"  
Copyright (C) 2021 The R Foundation for Statistical Computing  
Platform: x86\_64-apple-darwin17.0 (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.  
  
Natural language support but running in an English locale  
  
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.  
  
[R.app GUI 1.77 (7985) x86\_64-apple-darwin17.0]  
  
[Workspace restored from /Users/dannylumian/.RData]  
[History restored from /Users/dannylumian/.Rapp.history]  
  
>

# Why R?

- With FOSS, it is easy to get analysis tools, share and collaborate on work with few barriers to access
  - Compared with MATLAB, which requires licensing
- Use of programming language is typically required for big data
  - Excel can not handle millions of datapoints
- R is currently considered the standard for bioinformatics and biostatistics with many packages for advanced visualization and analysis
  - [Bioconductor](#)
  - [Tidyverse](#)



# What is R Studio?

- An integrated development environment (IDE) for R
  - Software that includes R, R console, file explorer, graphics viewer, code notebook and more!
  - One stop shop for coding in R
  - <https://posit.co/>
-

R\_Intro\_Fall\_2022 - main - RStudio

Basic\_Commands.Rmd Data\_Import\_and\_Export.Rmd

Source Visual

```
1 ---  
2 title: "Basic_Commands"  
3 author: "Daniel Lumian"  
4 date: '2022-08-22'  
5 output: pdf_document  
6 ---  
7  
8 ## R Markdown  
9  
10 This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.  
11  
12 When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document.  
13  
14 You can embed an R code chunk like below.  
15  
16 Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Cmd+Shift+Enter*.
```

6:4 # Basic\_Commands R Markdown

Console Terminal Background Jobs

```
R 4.1.1 . ~/Documents/GitHub/R_Intro_Fall_2022/  
> print(vector_1)  
[1] 1 2 3 4 1 2 3 4 2 4 6 8  
> df1 = rbind(vector_1, vector_2, vector_3)  
> print(df1)  
     [,1] [,2] [,3] [,4]  
vector_1  1    2    3    4  
vector_2  1    2    3    4  
vector_3  2    4    6    8  
> df2 = cbind(vector_1, vector_2, vector_3)  
> print(df2)  
   vector_1 vector_2 vector_3  
[1,]      1      1      2  
[2,]      2      2      4  
[3,]      3      3      6  
[4,]      4      4      8  
>
```

Environment History Connections Git Tutorial

Import Dataset 113 MiB

R Global Environment

Data

df1	num [1:3, 1:4] 1 1 2 2 2 4 3 3 6 4 ...
df2	num [1:4, 1:3] 1 2 3 4 1 2 3 4 2 4 ...

Values

a	2
b	2
c	3
d	4
vector_1	num [1:4] 1 2 3 4
vector_2	num [1:4] 1 2 3 4

Files Plots Packages Help Viewer Presentation

New Folder New Blank File Delete Rename More

Home Documents GitHub R\_Intro\_Fall\_2022

Name	Size	Modified
..		
.gitignore	1.8 KB	Aug 22, 2022, 10:40 AM
~\$Introduction_to_R.pptx	165 B	Aug 24, 2022, 9:55 AM
Basic_Commands.Rmd	3.9 KB	Aug 22, 2022, 1:11 PM
Data_Import_and_Export.Rmd	655 B	Aug 22, 2022, 1:25 PM
Introduction_to_R.pptx	440.7 KB	Aug 24, 2022, 10:46 AM
R_Intro_Fall_2022.Rproj	205 B	Aug 24, 2022, 10:48 AM
README.md	75 B	Aug 22, 2022, 10:29 AM

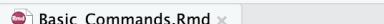
# What is Posit Cloud?

- Cloud based version of R Studio
- No need to download software
- Can share projects easily for analysis and teaching
- Allows for pre-installation of required packages
- <https://posit.cloud/>
- Recommended platform if you have not used R/R Studio before
  - Not recommended for larger projects due to limitation on resources

File Edit Code View Plots Session Build Debug Profile Tools Help

    Go to file/function  Addins

R 4.2.1

 Basic\_Commands.Rmd 

Source Visual

⚠ Package rmarkdown required but is not installed. [Install](#) [Don't Show Again](#)

```

1 ---  

2 title: "Basic_Commands"  

3 author: "Daniel Lumian"  

4 date: '2022-08-22'  

5 output: pdf_document  

6 ---  

7  

8 ## R Markdown  

9  

10 This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For  

more details on using R Markdown see <http://rmarkdown.rstudio.com>.  

11  

12 When you click the **Knit** button a document will be generated that includes both content as well as the output of any  

embedded R code chunks within the document.  

13  

14 You can embed an R code chunk like below.  

15  

16 Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing  

*Cmd+Shift+Enter*.

```

1:1  Basic\_Commands.Rmd 

Console Terminal Background Jobs

 R 4.2.1 · /cloud/project/ 

```

R version 4.2.1 (2022-06-23) -- "Funny-Looking Kid"
Copyright (C) 2022 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 Git Tutorial

 Import Dataset  135 MiB 

R Global Environment

Environment is empty

Files Plots Packages Help Viewer Presentation

 New Folder  New Blank File  Upload  Delete  Rename  More

 Cloud > project

	Name	Size	Modified
 ..			
<input type="checkbox"/>  .gitignore	.gitignore	1.8 KB	Aug 25, 2022, 8:33 AM
<input type="checkbox"/>  .Rhistory	.Rhistory	0 B	Aug 25, 2022, 8:33 AM
<input type="checkbox"/>  Basic_Commands.Rmd	Basic_Commands.Rmd	3.9 KB	Aug 25, 2022, 8:33 AM
<input type="checkbox"/>  Data_Import_and_Export.Rmd	Data_Import_and_Export.Rmd	655 B	Aug 25, 2022, 8:33 AM
<input type="checkbox"/>  Introduction_to_R.pptx	Introduction_to_R.pptx	1.3 MB	Aug 25, 2022, 8:33 AM
<input type="checkbox"/>  R_Intro_Fall_2022.Rproj	R_Intro_Fall_2022.Rproj	205 B	Aug 25, 2022, 8:33 AM
<input type="checkbox"/>  README.md	README.md	75 B	Aug 25, 2022, 8:33 AM

# R File Types

## R Script File

- Basic file for saving code
- Can add comments using `#` symbol
  - Documenting code with comments is a best practice
- Useful for testing code snippets and work that will not be shared
- Sections can be selected and run

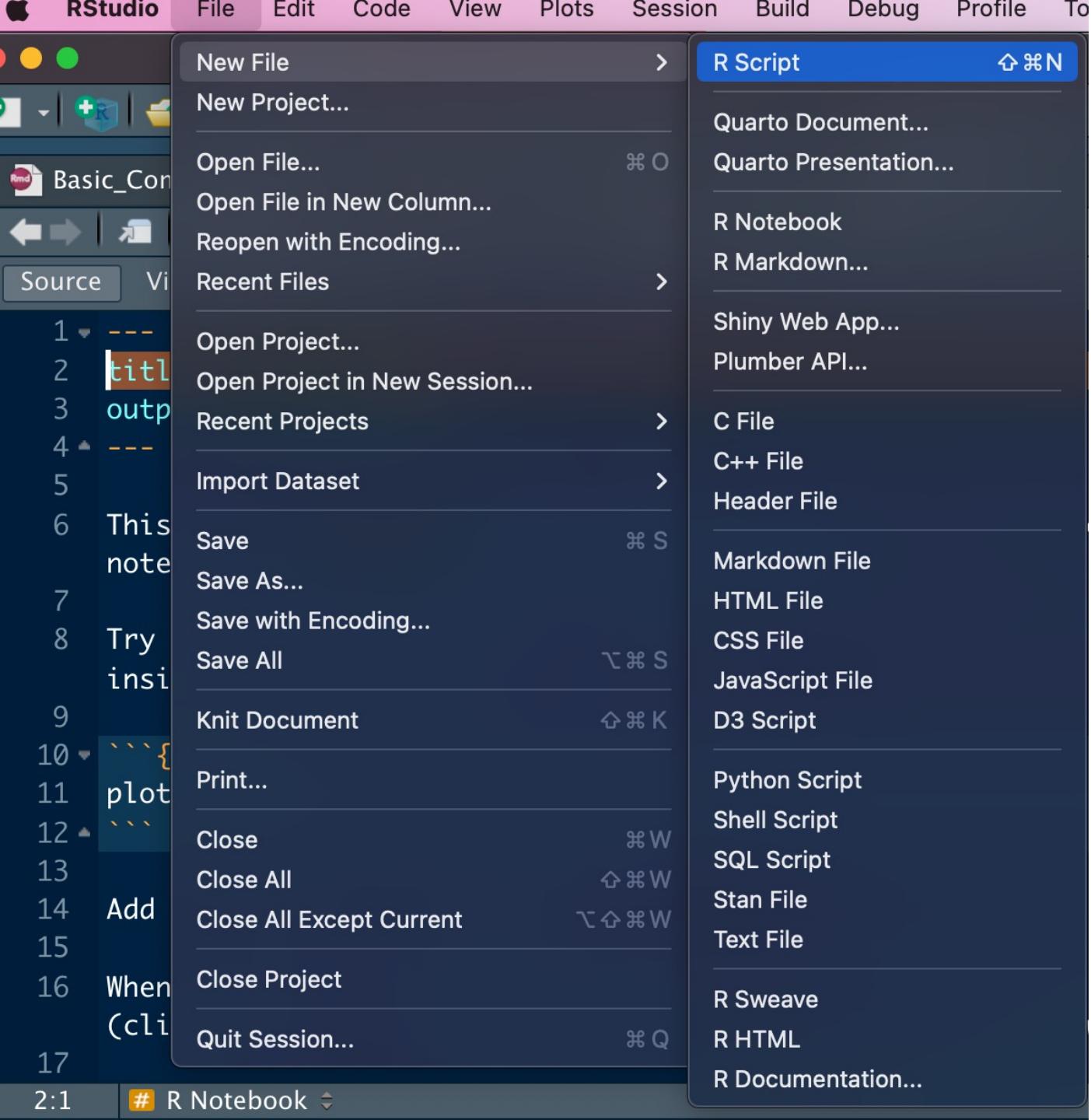
## R Markdown and Notebook

- More advanced files that allow inclusion of markdown and style elements for documentation
- Code lives in chunks within the file
- Better for sharing and publishing code
- Notebooks allow a preview feature, while Markdown files are knit

<https://stackoverflow.com/questions/43820483/difference-between-r-markdown-and-r-notebook>

---

- Open a new file
  - Select File from the top panel
  - Select New File from the dropdown
  - Select appropriate file type
    - R script
    - R Notebook
    - R Markdown
- Open an existing file
  - Select File from the top panel
  - Select Open File
  - Select the file you wish to open
- Save a file
  - Select File from the top panel
  - Select Save or Save as
  - Specify the save location

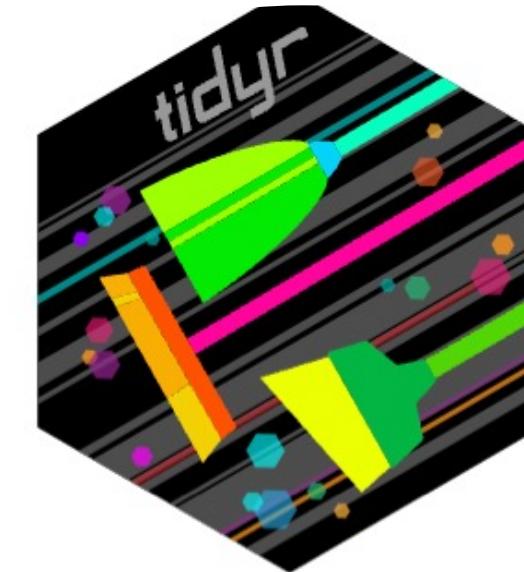


# R and R Studio Cheat Sheets

## Cheat Sheets

There are many capabilities, packages and tools available in R which can not be covered in a single session

This resource has helpful guides to some of the most popular packages and common uses of these packages



CHEATSHEET |

Data tidying with tidyr cheatsheet



The `tidyverse` package provides a framework for creating and shaping tidy data, the data format that works the most seamlessly with R and the `tidyverse`. The front page of this cheatsheet provides an overview of the `tidyverse`.

# Let's get started!

## R Studio Cloud

- Log in or sign up for Posit Cloud
  - Go to: <https://posit.cloud/>
- Open the following project
  - [https://posit.cloud/content/613\\_6578](https://posit.cloud/content/613_6578)
  - Click on Save a permanent copy
- Click on the new file icon
- Select R script

## Local R Studio

- Clone GitHub repo
  - <https://github.com/dlumian/NIH-R-Introduction>
- Navigate to repo directory
  - Open NIH-R-Introduction.Rproj file with RStudio
- 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 | Global Environment

List | C

Environment is empty

Files Plots Packages Help Viewer

New Folder Upload Delete Rename More

Cloud > project

	Name	Size	Modified
	..		
<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.
- The sidebar on the left has the text "Type 'demo()' for some demos, 'help.start()' for an HTML browser, Type 'q()' to quit R." and "Run code 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

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

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

> Run code here

Write code here

Environment is empty

Objects go here

Written files go here

R 4.0.0

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

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, which defines variables `a` and `b`, and performs arithmetic operations on them.
- Toolbar:** The "Run" button is highlighted with a red box.
- Environment Panel:** Shows the "Global Environment" with two objects:

Values
a 3
b 5

A red box highlights this table.
- Console Panel:** Displays the R session history:

```
> 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:** Shows a project structure in the cloud:

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

Annotations in red text:

- "Highlight the section of code you want to run then click run"
- "Temp memory objects"

# Basic R Arithmetic Operations

- Basic addition: **7 + 7**
- Assign variables:
  - **a=3**
  - **b=5**
- See contents of a variable:
  - Typing “**a**” returns “**3**”
  - Typing “**b**” returns “**5**”

Run operations on variables:

**a+b**

**a-b**

**a\*b**

**a/b**

Two ways to run these operations:

- Enter each line in the console and push enter
- Enter all lines into the script, select which lines to run and push run

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, which defines variables `a` and `b`, and performs arithmetic operations on them.
- Toolbar:** The "Run" button is highlighted with a red box.
- Environment Panel:** Shows the "Global Environment" with two objects:

Values
a 3
b 5

A red box highlights this table.
- Console Panel:** Displays the R session history:

```
> 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:** Shows a project structure in the cloud:

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

Annotations in red text are overlaid on the code editor and environment panel:

- "Highlight the section of code you want to run then click run"
- "Temp memory objects"

# Vectors and Functions

- Vectors are a data structure in R
  - must be same data type
  - list of characters
  - list of numbers
- Use the "c" function to combine items
  - `c(5,6,7,8)` creates a vector
- c stands for combine
- Putting a `?` before a command or object will retrieve the help for that object
- Type: `?c`

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

← → Home

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:
```

# Help Details

- 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

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:
c(..., recursive = FALSE, use.names = TRUE)
```

# Create an object

- Running `c(5,6,7,8)` will only return the output to console
- To store objects in R memory assign a variable name
  - This allows for reuse of the variable
- Type "d<- " in front of `c(5,6,7,8)` : `d <- c(5,6,7,8)`
- This creates 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)



```
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



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 |

Source on Save Run Source

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)`
- Not sure how `c` will combine vectors?
  - Run the example and see!
- Not sure what `rbind` and `cbind` do?
  - Check their help documentation!

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:** An untitled R script with the following code:

```
1 h <- rbind(d,e,f,g)
2 h
```
- Environment View:** Shows the global environment with variables d, e, f, g, and h. The variable h is a numeric matrix of size 4x4 with values: 5, 11, 1, 1; 6, 12, 2, 2; 7, 13, 15, ...
- Console View:** Displays the output of the R code:

```
> 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:** A search results page for "standard deviation". It shows the R logo and a message: "The search string was 'standard deviation'". It also lists help pages for `nlme:::pooledSD`, `stats::sd`, and `stats::sigma`.

# Colnames 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\*

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 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

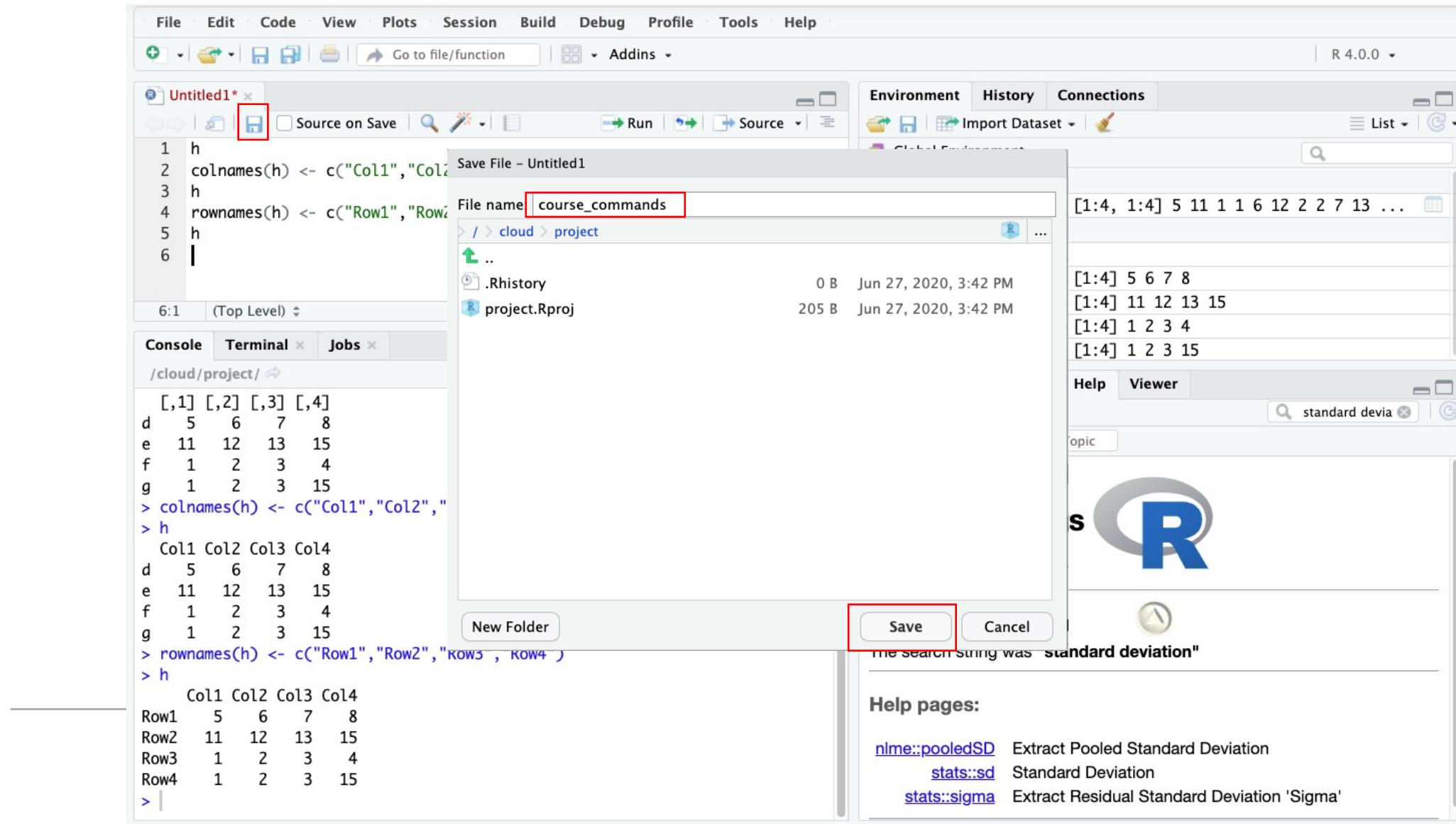


The search string was "standard deviation"

## 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

course\_commands.R | Go to file/function | Addins | R 4.0.0

course\_commands.R x

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

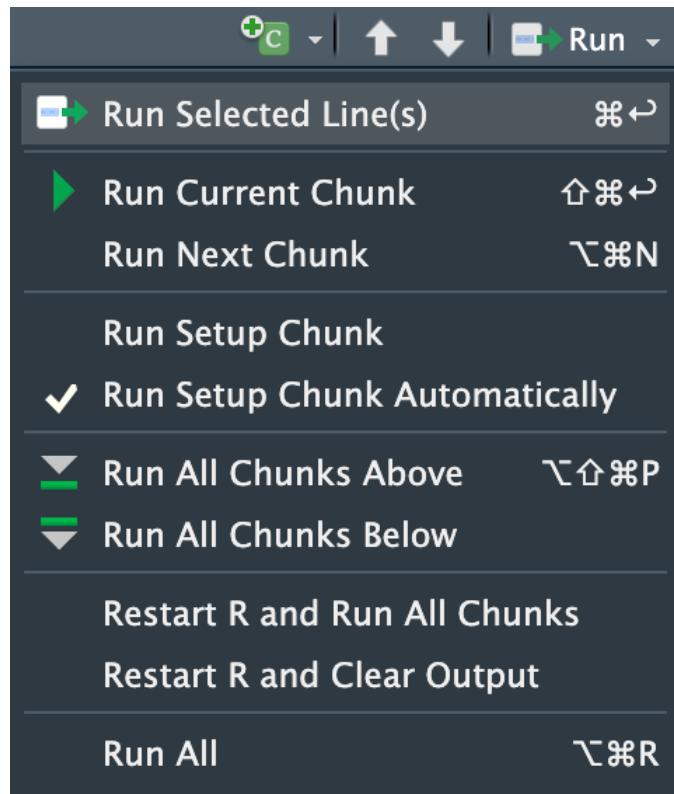
Files Plots Packages Help Viewer

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

# Activity 1: 1\_Basic\_Commands.Rmd



- We just learned about creating R scripts to manipulate basic data (vectors, data frames)
- Run 1\_Basic\_Command.Rmd notebook to reinforce your skills and learn about running R Markdown
- Please let me know if you run into any issues or have any questions

	Name	Size	Modified
	..		
	1_Basic_Commands.Rmd	5 KB	Jun 21, 2023, 4:22 PM
	2_Datasets.Rmd	4.3 KB	Jun 21, 2023, 4:22 PM
	3_Data_Manipulation.Rmd	3.7 KB	Jun 21, 2023, 4:22 PM
	4_Data_Import_and_Export.Rmd	3.2 KB	Jun 21, 2023, 4:22 PM
	5_Data_Visualization.Rmd	7.1 KB	Jun 21, 2023, 4:22 PM
	6_Data_Analysis.Rmd	6.8 KB	Jun 21, 2023, 4:22 PM

# Questions?

- Recap
    - 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 data frame with named rows and cols
    - Ran a markdown notebook reinforcing these concepts
  - Up next:
    - R Built-In Datasets
-

# Datasets

- In the previous section, we created datasets by inputting and combining vectors
  - It is more common to import existing data
  - R has functionality to import many kinds of data
    - Tabular data like csv, excel and text delimited files
    - Outputs from other analysis packages (e.g., SPSS, Stata)
    - Images and other complex data formats
    - Specific packages for importing bioinformatics data
-

# Built-in Datasets

- Datasets can be imported to or exported from R,
- R and some packages also include built-in datasets
- Built-in datasets are helpful for learning about data wrangling and testing functionality
  - Do not require complex imports
  - Often have example analysis and results for comparison

---

# New Operations

- `data()`: shows available datasets
  - `summary(df)`: summary data for all cols in df, where df is a data frame
  - `$`: `df$col_of_interest` will return the column specified for that df
  - Basic plots: `hist`, `boxplot` and `scatterplot`
-

# **Activity 2:**

## **Built-in Datasets**

- Open 2\_Datasets.Rmd file
  - This file explores working with built-in datasets and base R plotting
  - We will work with the Iris dataset from Base R
-

# Data Manipulation

USE BASE R FUNCTIONS

USE 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

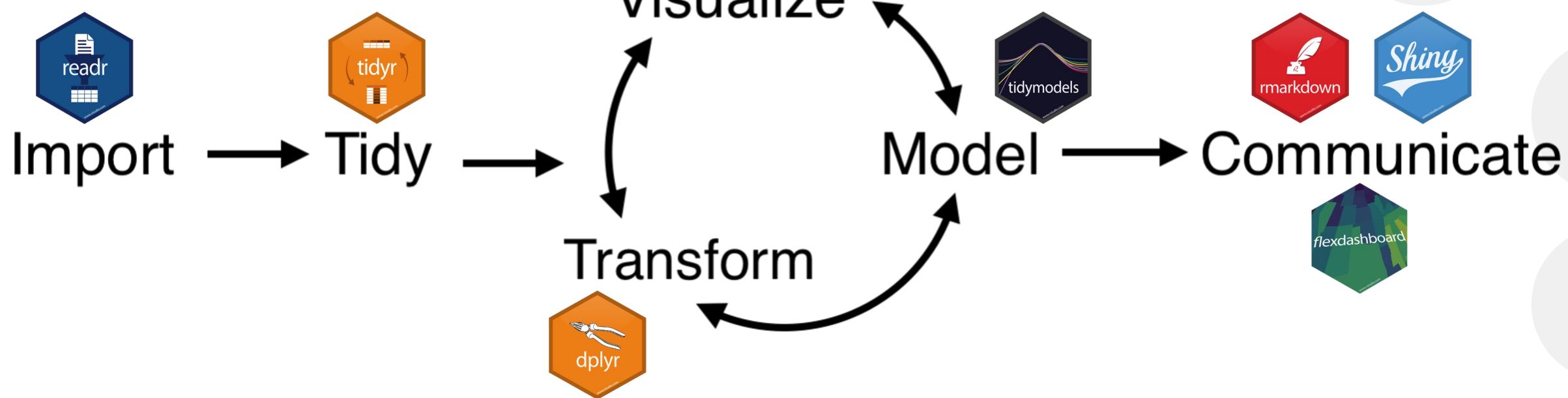


Made by different people so commands may clash



Would be very slow to load everything every time, so load only what you need

# Tidyverse



First time install:

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

Turn on package:

```
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") **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

The screenshot shows the RStudio interface with three main panels:

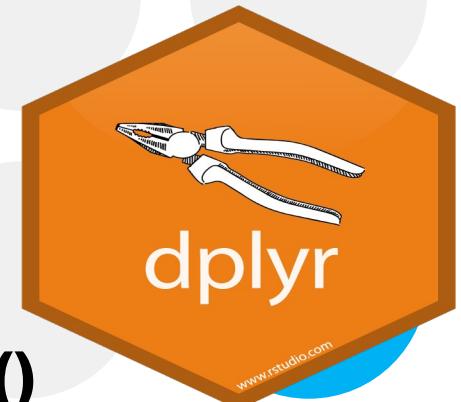
- Code Editor (Left Panel):** Displays the file `course_commands.R*`. The first line contains `library(tidyverse)`, which is highlighted with a red box. To the right of the code, the word "This" is written in red.
- Environment (Top Right Panel):** Shows the global environment with objects `pbc` and `pbcseq`. Below that, it lists variables `c`, `d`, `e`, `f`, and `g`, each with its corresponding value and type.
- Packages (Bottom Right Panel):** Shows the `Packrat` interface. The search bar at the top contains "tidyverse". A red box highlights the checkbox next to `tidyverse`, which is checked. The table below lists the package details:

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

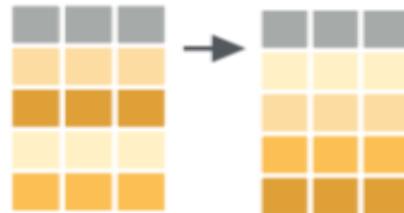
Centered between the Code Editor and the Packages panel is the text "OR this" in red.

```
library(tidyverse)
#> 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()
#>
```

# dplyr: Transform your data



**arrange()**



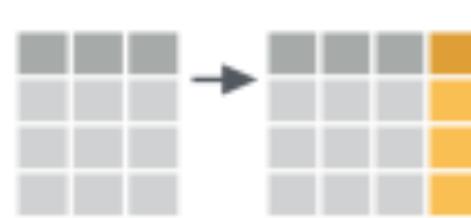
**select()**



**filter()**



**mutate()**



**summarize()**



**group\_by()**



**Basic structure:**

`functionName(data, specifics)`

**Example:**

`arrange(iris, Sepal.Length)`

**Keyboard Shortcut:**

**PC:**Ctrl+Shift+M

**Mac:**Cmd+Shift+M

# Pipes!

Pipes allow for sequential operations to be done on the same dataset

- Output from one step is used as input to the next step
- Avoids having to store all intermediate steps

Basic way to chain actions

- `group_by_species <- group_by(iris, Species)`
- `iris_sep_len <- summarize(group_by_species, mean = mean(Sepal.Length))`
- `View(iris_sep_len)`

Pipe sequence (%>%)

- `iris_sep_len <- iris %>% group_by(Species) %>% summarize(new_col = mean(Sepal.Length))`
- `View(iris_sep_len)`

	Species	new_col
1	setosa	5.006
2	versicolor	5.936
3	virginica	6.588

# **Activity 3:**

## **Data Manipulation**

- 3\_Data\_Manipulation.Rmd will explore the following data manipulations using the Tidyverse library:
  - Arrange
  - Select
  - Filter
  - Mutate
  - Summarize
  - GroupBy

---

# Questions?

- Activity 3 used the Tidyverse packages to explore data manipulation techniques
  - Process used the built-in Iris dataset
  - Final result was a grouped by aggregation
-

# Data Import and Export

- Data comes in many formats
  - There are a variety of tools for importing and exporting data
  - R can handle data in csv, tab-delimited, excel formats, images and more
-

# Save a table as a file

- `write.table(iris_mutate,"./data/iris_mutate.txt",row.names=F,sep="\t")`
  - `write.table` is the command
  - `iris_mutate` is the data to be written
  - “`./data/iris_mutate.txt`” is the file path to write to
    - .. means go up a directory
    - Then go into the data directory and name the file `iris_mutate.txt`
  - `row.names=F` (False, do not write row names)
  - Sep = “`\t`” for tab delimited data
-

# Read Data From Website

- `read_csv()` and `read_tsv()` are special cases of the more general `read_delim()`. They're useful for reading the most common types of flat file data, comma separated values and tab separated values, respectively.
- File can be a url where data is stored
- Run `'?read_tsv'` for more help information

# **Activity 4:**

## **Data Import and Export**

- 4\_Data\_Import\_and\_Export.Rmd will explore the following:
  - Save data as tab delimited text
  - Save data as a csv file
  - Load both tab delimited and csv files
  - Load data from a url

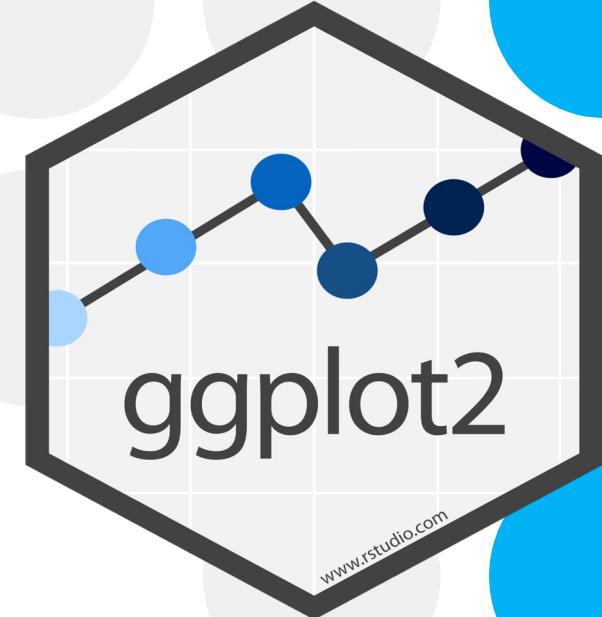
---

# Visualizations for Today

- Pie chart
  - Bar plot
  - Histogram
  - Density Plot
  - Scatter plot
  - Box Plot
  - Heat Map
  - Faceted plots (multiple plots in the same figure)
-

# ggplot2: Visualize Data

- Part of 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: Visualize Data

## What you need:

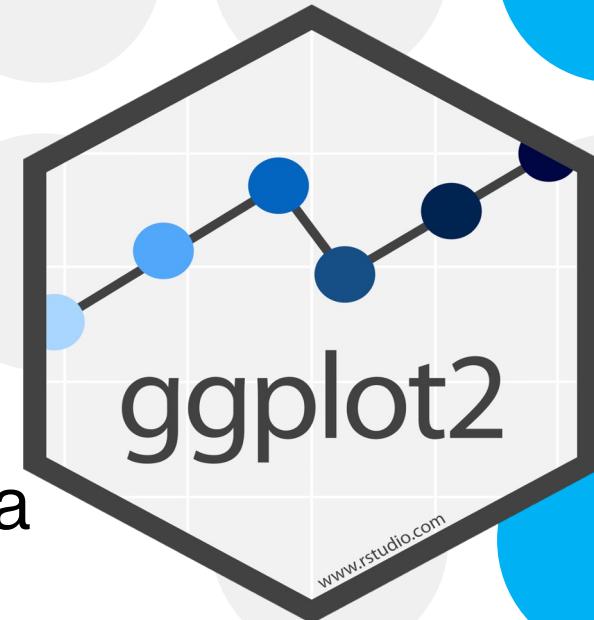
1. A data object
2. Aesthetic mappings (aes): how variables in the data are assigned to visual properties
  - x- and y-direction
  - shapes, colors, lines
3. A geometry object (geom): the type of plot

## Basic structure:

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

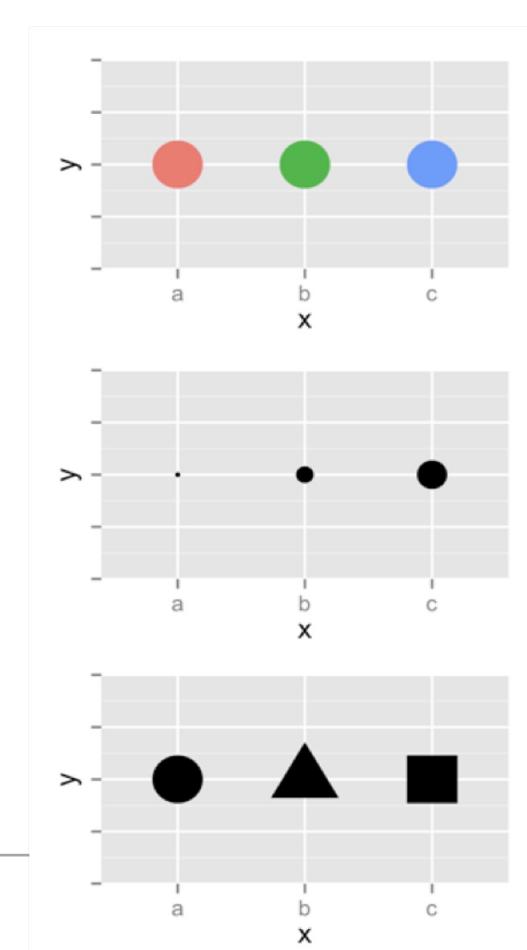
## Example:

```
ggplot(iris, aes(x=Sepal.Length)) + geom_bar()
```



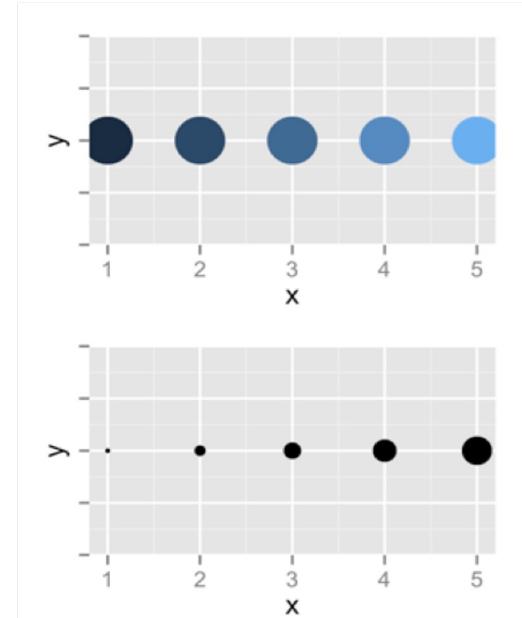
# Aesthetic mapping options

Color



Size

Continuous

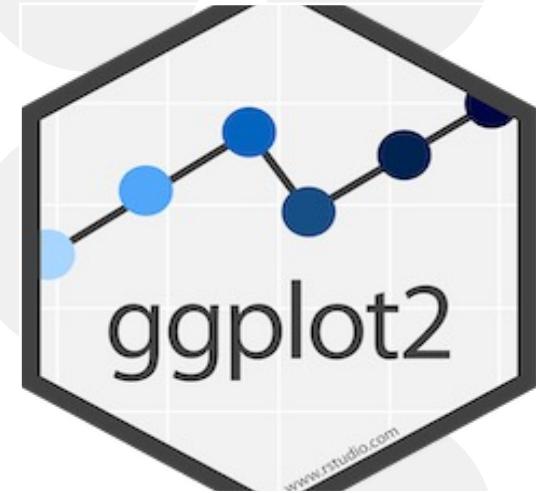


Shape

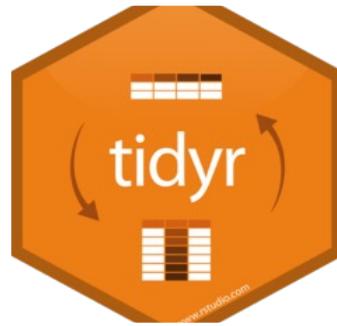


# ggplot2: Faceting

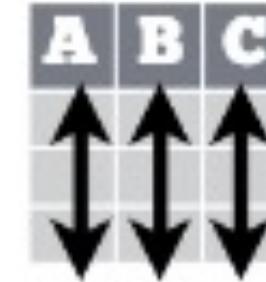
- Divide a plot into subplots based on one or more discrete variable
- Can be used with a variety of plot types
- There are a couple of facet flavors
  -  **`t + facet_grid(cols = vars(f1))`**  
facet into columns based on f1
  -  **`t + facet_grid(rows = vars(year))`**  
facet into rows based on year
  -  **`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



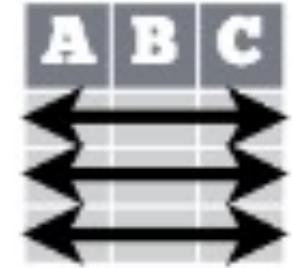
# tidr: gather



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



&



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



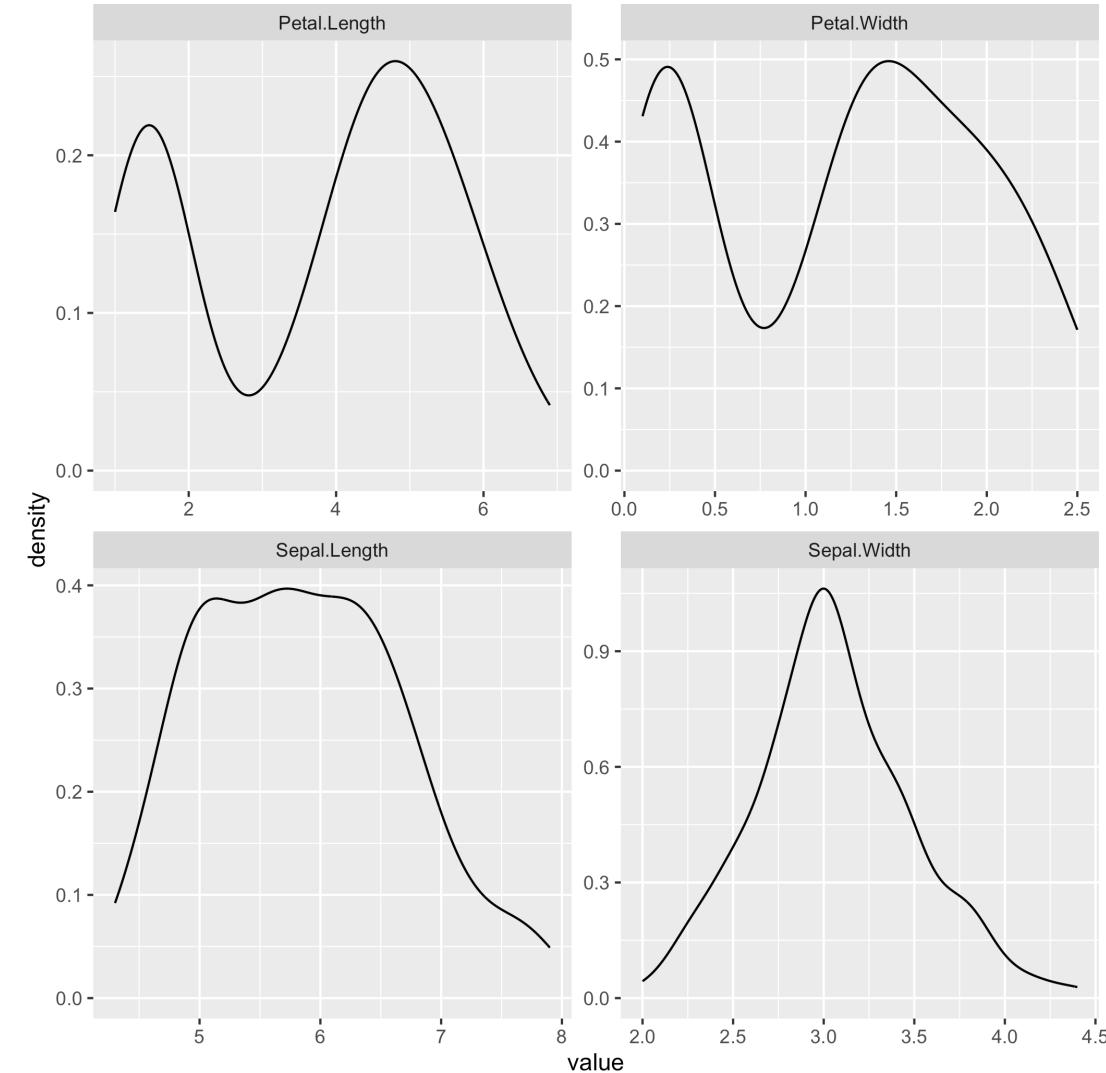
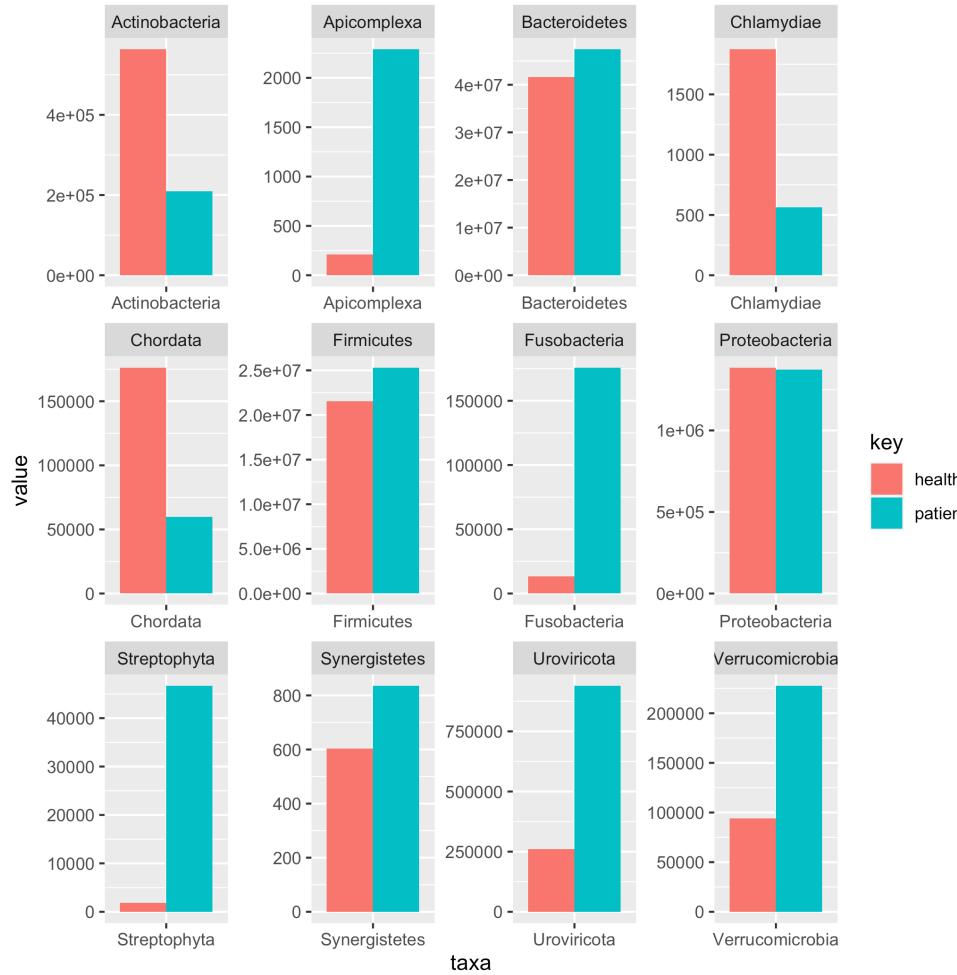
country	year	cases	key	value
A	1999	0.7K		
B	1999	37K		
C	1999	212K		
A	2000	2K		
B	2000	80K		
C	2000	213K		

# **Activity 5:** **Data Visualization**

- 5\_Data\_Visualization.Rmd explores plotting of various data types
- Explore plotting and data visualizations

---

# Questions on Basic Visualizations?



# Data Analysis in R

- Correlations
  - Pearson
- t-tests
  - Single sample
  - Independent
  - Paired

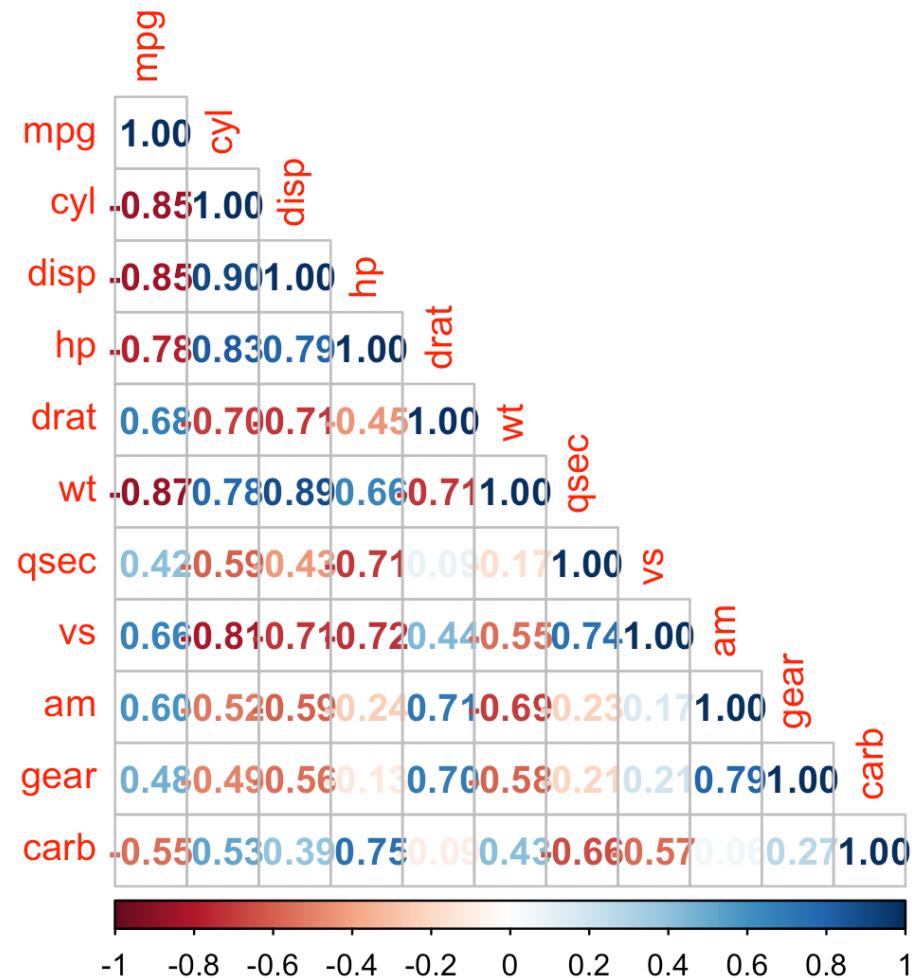
---

# Correlation

- Measure of how well two variables hang together
  - Range from -1 to 1
  - 1 is a perfect positive correlation
  - -1 is an inverse correlation
-

# Correlation Matrix

- Diagonal is all 1's
  - Each variable is a perfect correlation with itself
- Positive and negative correlations are color coded
  - Blue is positive
  - Red is negative
- Strength is shown in darker colors



# Correlation Test

Cor.test

P-value shows significance

Correlation coefficient is shown at bottom

```
```{r}
cor.test(mtcars$mpg, mtcars$hp)
```
```

Pearson's product-moment correlation

```
data: mtcars$mpg and mtcars$hp
t = -6.7424, df = 30, p-value = 1.788e-07
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.8852686 -0.5860994
sample estimates:
cor
-0.7761684
```

# t-tests

- Single sample
    - Compares a sample to a known value ( $\mu$ )
  - Independent
    - Compares two samples which are not related
  - Paired
    - Compares two samples which are related
-

# One Sample t-test

- Compares a vector of data against a known value ( $\mu$ )
- p-value indicates significance
- Mean of vector is shown at bottom

```
```{r}
t.test(mtcars$wt, mu=3)
```

One Sample t-test

data: mtcars$wt
t = 1.256, df = 31, p-value = 0.2185
alternative hypothesis: true mean is not equal to 3
95 percent confidence interval:
2.864478 3.570022
sample estimates:
mean of x
3.21725
```

# Two Sample Independent t-test

- Compares two vectors of numeric data
- Determines if means are different
- p-value indicates significance
- Means of each group displayed at bottom

```
```{r}
t.test(mtcars$mpg~mtcars$vs)
```

Welch Two Sample t-test

data: mtcars$mpg by mtcars$vs
t = -4.6671, df = 22.716, p-value = 0.0001098
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
-11.462508 -4.418445
sample estimates:
mean in group 0 mean in group 1
16.61667      24.55714
```

# Paired-Samples t-test

- Compares two related samples
- p-value indicates significance
- Average of the differences is shown at the bottom

```
```{r}
t.test(ChickWeightWide$weight.0, ChickWeightWide$weight.2, paired = TRUE)
```
Paired t-test

data: ChickWeightWide$weight.0 and ChickWeightWide$weight.2
t = -17.409, df = 44, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-9.496402 -7.525820
sample estimates:
mean of the differences
-8.511111
```

# **Activity 6:**

## **Data Analysis in R**

- 6\_Data\_Analysis.Rmd explores two basic analysis techniques
    - Correlations
    - t-tests
  - MANY more analysis tools exist, this is only an introduction to the concept
-

# Concluding Remarks

- This introduction to R previewed many common functions and analysis tools
- Importing and using packages often prevents you from having to “reinvent the wheel”
- Statistical programming is greatly enhanced when combined with version control and data sharing techniques
- Don’t be afraid of errors and be sure to explore the available resources for your work

**This  
concludes  
the Intro to R**

Thank you for your attention and participation!

---