

Reproducible (and collaborative) science through RStudio

A whirlwind tour with R, RMarkdown, Python, LaTeX, and
more

Jenny Rieck & Derek Beaton

May 20 2019

The big outline

- ▶ Part 0: Background
- ▶ Part 1: A bit about R
- ▶ Part 2: RStudio & Project setup
- ▶ Part 3: R, RMarkdown, & more
- ▶ Part 4: Advanced, beyond, & our favorites

Part 0: Background

To dive right in

If you want to skip over the background & RStudio, go straight to
Part 2: RStudio & Project setup

Background

- ▶ This is a taste and to bring you into a bigger world
 - ▶ R, Python, SQL, and JavaScript are critical data science tools/languages
- ▶ R (language and community) strongly emphasizes
 - ▶ Centralization & standards
 - ▶ Rigor & reproducibility (packages, RMarkdown)
- ▶ An interesting language
 - ▶ Functional
 - ▶ With a sublanguage (or dialect?): the tidyverse

R is a community (actually many communities!)

- ▶ Help and resources
- ▶ Package development and distribution
- ▶ An ideal example
 - ▶ Not quite always that way
 - ▶ Strong communal presence

R: Help!

- ▶ So many websites e.g., <https://www.statmethods.net/>
- ▶ Online forums (Stack Exchange, r-lists)
- ▶ SpringerLink
 - ▶ All R books for free (pdf format) or for minimal cost (printed)
- ▶ Vignettes
 - ▶ step-by-step instruction guides for packages
- ▶ Git
 - ▶ With open books (via bookdown)
- ▶ Twitter #rstats
- ▶ RStudio (website)
 - ▶ Videos, cheat sheets

R Packages

- ▶ Packages are bundles of code made by someone (or many people) for everyone to use
 - ▶ There are packages for everything
 - ▶ We'll cover some of the diversity throughout
- ▶ Comprehensive & Reproducible
- ▶ Available primarily on CRAN
 - ▶ But also github (less so: r-forge)

Part 1: A bit about R

R Background

- ▶ Created in 1992 by Gentleman & Ihaka

[we] considered the problem of obtaining decent statistical software for our undergraduate Macintosh lab. After considering the options, we decided that the most satisfactory alternative was to write our own. [...] Finally we added some syntactic sugar to make it look somewhat like S. We call the result “R”.

What is R?

- ▶ R is general purpose programming
 - ▶ Design around & for statistics
 - ▶ “for and by statisticians”
- ▶ R is a collection of tools
 - ▶ Pre-packaged software at your disposal
- ▶ R is free (as in beer and speech)
 - ▶ No cost, no restrictions
 - ▶ E.g., Microsoft (nee Revolution) R
- ▶ R is a functional language
 - ▶ Turing complete
 - ▶ Mathematical functions
 - ▶ Pass expressions and functions to and from functions

R: Data types

- ▶ Stored as *vectors*
 - ▶ see `class()`
- ▶ numeric
 - ▶ real or decimal
 - ▶ Includes `NaN`, `Inf`, `-Inf`
- ▶ integer
- ▶ complex
- ▶ character
- ▶ logical
 - ▶ includes `NA`, `TRUE`, `FALSE`
- ▶ factor
 - ▶ factors are usually not your friends
 - ▶ generally: `stringsAsFactors = F` or convert these

R: factor disasters

```
a_numeric_vector <- c(3, 0, 1, -2, 2, 5, 5, 2, 1)
(a_numeric_vector + 1)

## [1] 4 1 2 -1 3 6 6 3 2

(a_numeric2factor_vector <- as.factor(a_numeric_vector))

## [1] 3 0 1 -2 2 5 5 2 1
## Levels: -2 0 1 2 3 5

(as.numeric(a_numeric2factor_vector))

## [1] 5 2 3 1 4 6 6 4 3

(as.numeric(as.character(a_numeric2factor_vector)))

## [1] 3 0 1 -2 2 5 5 2 1
```

R: Data structures

- ▶ Starts counting from 1
 - ▶ Not 0
- ▶ vector[1]
- ▶ matrix[1,1]
- ▶ array[1,1,1]
- ▶ list[[1]]
 - ▶ Can contain mixtures of types
 - ▶ or list\$name
- ▶ data.frame:
 - ▶ Is technically a list but access in three ways
 - ▶ data.frame[[1]][1]
 - ▶ data.frame[1,1]
 - ▶ data.frame\$name
 - ▶ tibbles: tidyverse data.frames

R

Some more about R here... Matlab cheat sheet Other cheat sheets

Tidyverse

- ▶ something here about tidy
- ▶ Learn it. But don't learn *only* the tidyverse; you'll be lost in base R

Part 2: RStudio & Project setup

RStudio

- ▶ IDE: Integrated development environment
- ▶ RStudio: Does so much
 - ▶ We scratch the surface here
- ▶ Quick walk through
- ▶ Followed by specific set up
 - ▶ Generally, but
 - ▶ Also for this workshop

RStudio Setup

- ▶ Download R and Rstudio
- ▶ Strongly recommend Microsoft R
(<https://mran.microsoft.com/open>)
 - ▶ Comes with Intel MKL
- ▶ Plain R is fine (<https://cran.r-project.org/>)
 - ▶ Can relink to faster libraries
- ▶ Download RStudio (<https://www.rstudio.com/>)

RStudio Environment

The screenshot shows the RStudio interface with the following components:

- Editor:** Displays R code for creating an ADNI data subset. The code includes library imports, data loading, cleaning, and merging steps. It also handles missing data and manually changes variable classes.
- Console:** Shows statistical summaries for various variables like APOE4, FDG, AV45, CDRSB, ADAS13, and MOCA across different groups (e.g., Mean, Min, Max, Quartiles).
- File Browser:** Shows the project structure with files like .Renviron, README.md, and Rmd files.
- Environment:** Shows the global environment with objects like `merge_subset` (665 obs. of 17 variables) and `ids` (chr vector).

```
## ~/workshops/2019_Rstudio_Magic - master - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
+ Go to file/function
+ Run Source All
Data
  merge_subset 665 obs. of 17 variables
  variable_type_map num [1:17, 1:3] 0 1 0 0 0 0 0 0 1 0 ...
  Values
  ids           chr [1:665] "2002" "2003" "2007" "2010" "2011" ...
  MOCA          num [1:665] 28 24 23 27 25 26 25 24 24 30 ...
Functions
  scatterplotter function (x, y, x.lab = NA, y.lab = NA, x.lim = ...
A Script : 4819 (Unsaved)
Console Terminal Jobs
~/workshops/2019_Rstudio_Magic/:
Mean :71.92      Mean :16.36
Min. :0.6983    Min. :0.8385   Min. :0.000   Min. :0.0   Min. :16.00
3rd Qu.:76.60    3rd Qu.:18.00   3rd Qu.:1.000   3rd Qu.:0.0   3rd Qu.:22.00
Max. :89.60      Max. :20.00    Max. :1.000   Max. :1.0   Max. :30.00
APOE4      FDG      AV45      CDRSB     ADAS13     MOCA
Min. :0.0000    Min. :0.6983  Min. :0.8385  Min. :0.000   Min. :0.0   Min. :16.00
1st Qu.:0.0000   1st Qu.:1.12213 1st Qu.:1.12213  1st Qu.:0.0000  1st Qu.:8.4   1st Qu.:22.00
Median :0.0000   Median :1.2802  Median :1.1105  Median :1.0000  Median :10.0  Median :25.00
Mean   :0.5248   Mean   :1.2682  Mean   :1.1989  Mean   :1.2020  Mean   :11.8   Mean   :23.89
3rd Qu.:1.0000   3rd Qu.:1.3620  3rd Qu.:1.3714  3rd Qu.:1.2000  3rd Qu.:18.0   3rd Qu.:26.00
Max.  :2.0000   Max.  :1.7013  Max.  :2.0256  Max.  :1.5000  Max.  :46.0   Max.  :30.00
WholeBrain  Hippocampus Midtemp nPACtrailsB  HMSCORE
Min. :14.421  Min. :1.111  Min. :1.2213  Min. :-18.6883  Min. :0.000
1st Qu.:98.8410 1st Qu.:6.510  1st Qu.:6.535  1st Qu.:0.051  1st Qu.:0.000
Median :105.1621  Median :7.223  Median :7.0186  Median :-2.5250  Median :1.000
Mean   :105.7026  Mean   :7.150  Mean   :20.302  Mean   :-3.6882  Mean   :0.588
3rd Qu.:112.0570 3rd Qu.:7.834  3rd Qu.:22.088  3rd Qu.:-0.3482  3rd Qu.:1.000
Max.  :148.6036  Max.  :10.602  Max.  :32.189  Max.  :5.3540  Max.  :3.000
> view(merge_subset)
> |
```

RStudio Environment

The screenshot displays the RStudio interface with several windows open:

- Code Editor:** Shows a script named `create_ADNI_data.R` containing R code for data manipulation. The code includes library imports, data loading, merging datasets, and manual variable class changes.
- Console:** Shows the output of the R code execution. It includes descriptive statistics for various variables like APOE4, FDG, AV45, CDRSB, ADAS13, and MOCA across different brain regions (Wholebrain, Hippocampus, Midtemp, nPACtrailsB, MSMScore).
- Environment:** Shows the global environment with objects like `merge_subset` (665 obs. of 17 variables), `ids` (chr vector), and `MOCA` (num vector).
- File Browser:** Shows the project structure with files like `README.md`, `script.R`, and `output`.

RStudio Environment

The screenshot shows the RStudio interface with the following components:

- Script Editor:** Displays the R script `create_ADNI_data.R`. The code performs the following steps:
 - Imports required packages: `tidyverse`, `lapply`, `data.table`, and `stringr`.
 - Creates a function `PTRACCAT` that takes a list of data frames and merges them into a single data frame.
 - Specifies column names for the merged data frame.
 - Loads and cleans data from `ADNI_merge` and `MOCA` datasets.
 - Specifies column names and participants for the baseline visit.
 - Removes participants with missing data.
 - Brings in modified hachimaki functions.
 - Manually changes variable classes (removing `labelled` class).
- Console:** Shows statistical summaries for various variables across different datasets (e.g., APOE4, FDG, AV45, CDRSB, ADAS13, MOCA, Hippocampus, Midtemp, nPACCtrailsB, HMSCORE). For example, the APOE4 dataset has the following summary statistics:

	Mean	3rd Qu.	Min.	Max.
APOE4	:71.92	:76.60	:69.60	:83.85
FDG	:71.92	:76.60	:69.60	:83.85
AV45	:71.92	:76.60	:69.60	:83.85
CDRSB	:71.92	:76.60	:69.60	:83.85
ADAS13	:71.92	:76.60	:69.60	:83.85
MOCA	:71.92	:76.60	:69.60	:83.85

- File Browser:** Shows the directory structure of the workspace:
 - Home
 - workshops > 2019_Rstudio_Magic
 - Renviron
 - 2019_Rstudio_Magic.Rproj
 - external
 - mac
 - output
 - R
 - README.md
 - Rmd
- Environment:** Shows the global environment with objects like `merge_subset` (665 obs. of 17 variables), `variable_type_map` (num [1:17, 1:3] 0 1 0 0 0 0 0 0 1 0 ...), and `Values` (ids, chr [1:665] "2002" "2003" "2007" "2010" "2011" "2012...").
- Plots:** Shows a scatter plot titled "FILEs, PLOTS, HELP".

RStudio Environment

The screenshot shows the RStudio interface with several windows open:

- Code Editor:** Shows R code for creating an ADNI data subset. The code includes library imports, data loading, merging, and subset selection. A red box highlights the final command: `> view(merge_subset)`.
- Environment:** Shows the `merge_subset` object, which is a data frame with 665 observations and 17 variables. It lists columns like `ADAS13`, `CDRSB`, `MOCA`, and `PTEDUCAT`.
- File Browser:** Shows the project structure under `workshops > 2019_Rstudio_Magic`. It includes files like `README.md`, `environment.Rproj`, and `output`.
- Text Overlay:** A large red text overlay in the center-right area reads "VARIABLES, HISTORY, VERSION CONTROL".

RStudio Environment

The screenshot displays the RStudio interface with several panes:

- Code pane:** Shows R code for creating an ADNI data subset. The code includes library imports, data loading, cleaning, and subset selection. It also handles missing data and manually changes variable classes.
- Console pane:** Displays statistical summaries for variables like APOE4, FDG, AV45, CDRSB, ADAS13, and MOCA. For example, APOE4 has a mean of 71.92 and a median of 70.00. The FDG variable has a range from 89.60 to 208.00.
- Environment pane:** Shows the global environment with objects like `anmerge_subset`, `variable_type_map`, `Values`, and `Functions`.
- File browser pane:** Shows the project structure with files like `README.md`, `Renviron`, and `2019_Rstudio_Magic.Rproj`.

```
library(ADNImerGE)
#####
## Load and clean data
#####
## 0.1 Specify the column names and participants you want (ie, baseline visit for all participants with MOCA>=1
admin.cols <- c("RID", "VISCODE", "DX", "AGE", "PTGENDER", "PTEDUCAT", "PTETHCAT", "PTRACCAT", "APOE4", "FDG", "ADAS13", "CDRSB", "MOCA")
admin.rows <- c(adminmerge$VISCODE=="b1" & adminmerge$MOCA>=16)
anmerge_subset <- adminmerge[admin.rows,admin.cols]
#####
## remove participants with missing data
anmerge_subset <- anmerge_subset[complete.cases(anmerge_subset),]
#####
## 0.2 Bring in modified hachkins
anmerge_subset$MSMSCORE <- modhach$MSMSCORE[match(anmerge_subset$RID, modhach$RID)]
#####
## 0.3 Manually change variable classes (remove class 'labelled')
anmerge_subset$FDG <- as.numeric(as.character(anmerge_subset$FDG))
anmerge_subset$AV45 <- as.numeric(as.character(anmerge_subset$AV45))
anmerge_subset$ADAS13 <- as.numeric(as.character(anmerge_subset$ADAS13))
anmerge_subset$CDRSB <- as.numeric(as.character(anmerge_subset$CDRSB))
anmerge_subset$MOCA <- as.numeric(as.character(anmerge_subset$MOCA))
```

	APOE4	FDG	AV45	CDRSB	ADAS13	MOCA
Min.	:0.0000	Min. :0.6983	Min. :0.8385	Min. :0.0000	Min. :0.0	Min. :16.00
1st Qu.	:0.0000	1st Qu.:17.60	1st Qu.:17.60	1st Qu.:0.0000	1st Qu.:8.0	1st Qu.:22.00
Median	:0.0000	Median :28.02	Median :28.02	Median :0.0000	Median :10.0	Median :25.00
Mean	:0.5248	Mean :31.2682	Mean :31.2682	Mean :1.202	Mean :13.8	Mean :23.89
3rd Qu.	:1.0000	3rd Qu.:31.3620	3rd Qu.:31.3714	3rd Qu.:2.0000	3rd Qu.:18.0	3rd Qu.:26.00
Max.	:2.0000	Max. :31.7013	Max. :32.0256	Max. :5.500	Max. :46.0	Max. :30.00

> view(anmerge_subset)
> |

RStudio Environment

The screenshot displays the RStudio interface with several windows open:

- Data Viewer:** A central window titled "DATA VIEWER" showing a table of 665 observations across 17 variables. The variables include DX, AGE, PTGENDER, PTEDUCAT, PTRECAT, PTRACCAT, APOE4, FDG, AV45, CDRSB, ADAS13, MOCA, WholeBrain, and Hippocampus.
- Global Environment:** A window showing the global environment with objects like anerage_subset, variable_type_map, values, and functions.
- File Browser:** A window showing the file structure under "workshops > 2019_Rstudio_Magic".
- Console:** A window showing R code and its output, including descriptive statistics for variables like APOE4, FDG, AV45, CDRSB, ADAS13, MOCA, WholeBrain, Hippocampus, Midtemp, nPACCtrailsB, and HMSCore.
- Terminal:** A window showing the command line interface.
- Jobs:** A window showing the current jobs.

Some benefits of RStudio

- ▶ Built-in integration with version control (git or SVN)
- ▶ Package and documentation generation
- ▶ Reproducible science!
 - ▶ R Markdown documents
 - ▶ Save and execute code
 - ▶ Generate high quality reports that can be shared
 - ▶ Create presentations (like this one!)
 - ▶ Even write papers
 - ▶ Python, D3 (JavaScript), SQL, Shiny, LaTeX, Git/SVN, HTML/CSS, and so much more.
- ▶ This workshop
 - ▶ Will walk you through some of this (and more)
 - ▶ See https://github.com/jennyrieck/workshops/tree/master/2019_Rstudio_Magic

RStudio is more

- ▶ Not just an IDE
- ▶ A company
- ▶ A community
- ▶ A conference
- ▶ A centralized resource

RStudio Resources

The screenshot shows the RStudio website homepage. At the top, there's a navigation bar with links for Products, Resources, Pricing, About Us, Blogs, and a search icon. Below the navigation is a decorative banner featuring a colorful, abstract graphic of overlapping colored bands.

RStudio: A screenshot of the RStudio IDE interface, showing the code editor, workspace, and plots.

Shiny: An image of a map of the United States with a "ZIP explorer" interface overlaid.

R Packages: Icons for several popular R packages: `markdown`, `Shiny`, `tidyverse`, `knitr`, and `ggplot2`.

RStudio description: RStudio makes R easier to use. It includes a code editor, debugging & visualization tools.

Shiny description: Shiny helps you make interactive web applications for visualizing data. Bring R data analysis to life.

R Packages description: Our developers create popular packages to expand the features of R. Includes `ggplot2`, `dplyr`, `R Markdown` & more.

At the bottom, there are download and learn more buttons for each section, and a horizontal orange progress bar.

RStudio Resources

Online Learning - RStudio

https://www.rstudio.com/online-learning/

R Studio

Products Resources Pricing About Us Blogs

Online learning

A wealth of tutorials, articles, and examples exist to help you learn R and its extensions. Scroll down or click a link below for a curated guide to learning R and its extensions.

- R Programming
- Shiny
- R Markdown
- Data Science
- Books

R Programming
Read More >

Shiny
Read More >

R Markdown
Read More >

Data Science
Read More >

RStudio Resources

Cheatsheets - RStudio x + - □ x

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

R Studio Products Resources Pricing About Us Blogs SEARCH

RStudio Cheat Sheets

The cheat sheets below make it easy to learn about and use some of our favorite packages. From time to time, we will add new cheat sheets to the gallery. If you'd like us to drop you an email when we do, let us know by clicking the button to the right.

SUBSCRIBE TO CHEAT SHEET UPDATES HERE

- RStudio IDE
- R Markdown
- Shiny
- Package Development
- Data Import
- Data Transformation with dplyr
- Data Visualization with ggplot2
- Apply functions with purrr
- Deep Learning with Keras
- Data Science in Spark with Sparklyr
- String manipulation with stringr
- Dates and times with lubridate

Python with R and Reticulate Cheat Sheet

The reticulate package provides a comprehensive set of tools for interoperability between Python and R. With reticulate, you can call Python from R in a variety of ways including importing Python modules into R scripts, writing R Markdown Python chunks, sourcing Python scripts, and using Python interactively within the RStudio IDE. This cheatsheet will remind you how.
Updated 4/19.

Use Python with R with reticulate :: CHEAT SHEET

The reticulate package makes it easy to have and use Python in R. It's a Python interface, just like R itself.

Python in R Markdown

Object Conversion

Helpers



Project and Environment Setup

- ▶ Special & hidden files
- ▶ Having a structure

RStudio Setup

- ▶ See <https://jennybc.github.io/2014-05-12-ubc/r-setup.html> for a detailed guide

For safety & collaboration

- ▶ Project(s) files
 - ▶ SOMETHING!

Projects through Git

- ▶ Create a new project File

New Project

Create Project

 **New Directory**
Start a project in a brand new working directory >

 **Existing Directory**
Associate a project with an existing working directory >

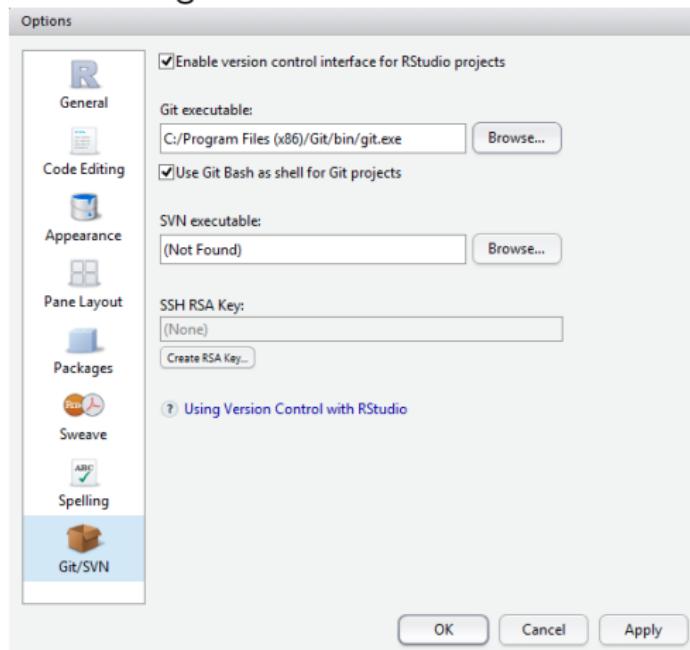
 **Version Control**
Checkout a project from a version control repository >

Cancel

Git & Projects

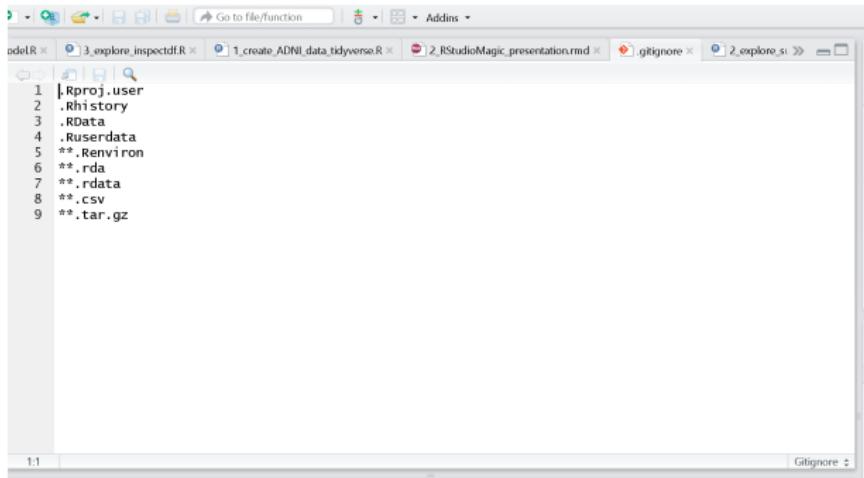
► Git

- Download git and link executable within RStudio



Format .gitignore

- ▶ File types to ignore via version control
 - ▶ ** before each extentions will match directories anywhere in the repo

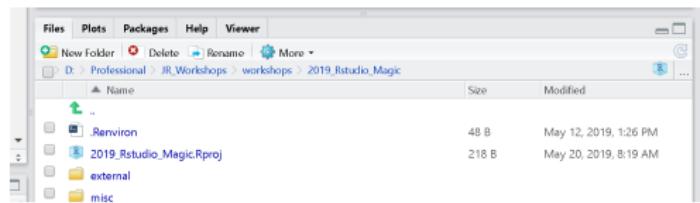


The screenshot shows the RStudio interface with the .gitignore tab selected in the top navigation bar. The code editor window displays the following content:

```
1 |Rproj.user
2 .Rhistory
3 .RData
4 .Ruserdata
5 **.Renvironment
6 **.rds
7 **.rdata
8 **.csv
9 **.tar.gz
```

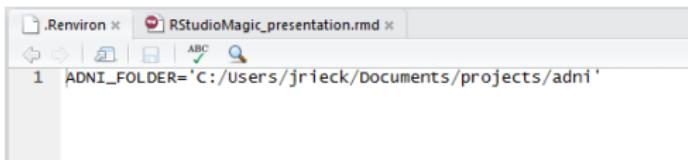
The code editor has a status bar at the bottom showing "1:1" and "Gitignore".

Environmental variables



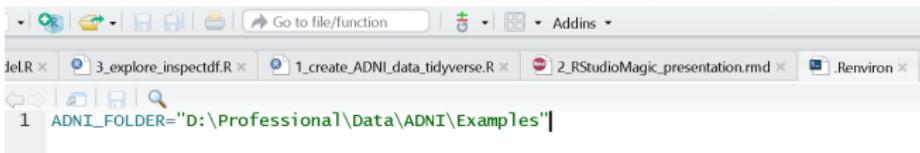
Format environmental variables

- ▶ Set environmental variables (ie, directory location of data) to make code generalizable across computers
 - ▶ Don't commit or share these
- ▶ In **your** project folder create a `.Renvironment` file and define variables
 - ▶ Jenny's:



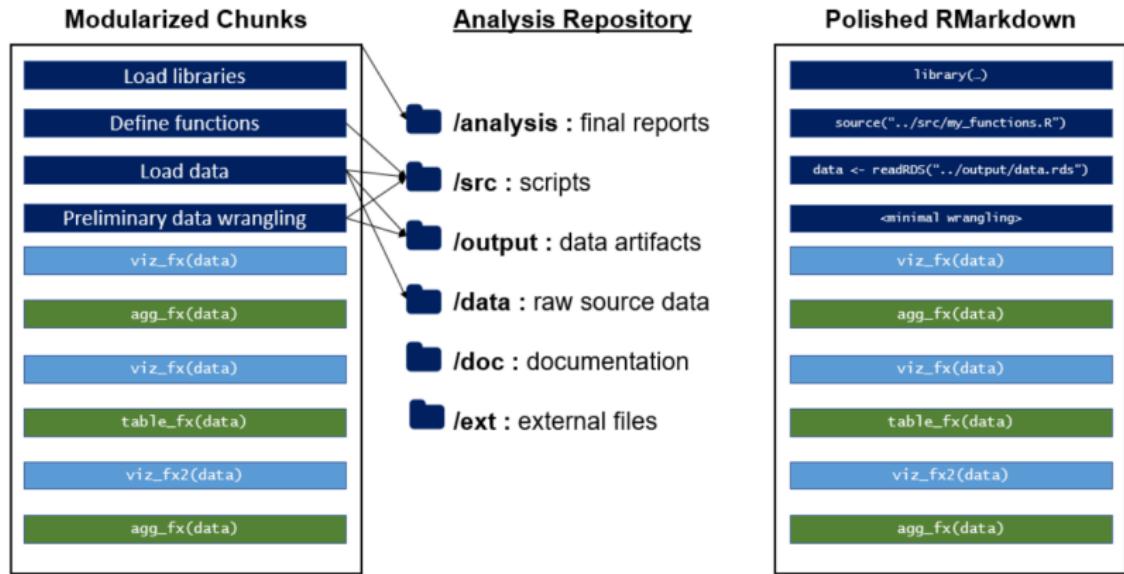
```
1 ADNI_FOLDER='C:/Users/jrieck/Documents/projects/adni'
```

- ▶ Derek's:



```
1 ADNI_FOLDER="D:\Professional\Data\ADNI\Examples"
```

Organize your project folders and markdown



<https://emilyriederer.netlify.com/post/rmarkdown-driven-development/>

Organize your project folders and markdown

- ▶ What works for you?
- ▶ What works for your organization or team?
- ▶ Maximize utility, minimize complexity

This works for us

 [jennyrieck / workshops](#)

 Watch ▾ 1  Star 0  Fork 0

 Code  Issues 0  Pull requests 0  Projects 0  Wiki  Insights  Settings

Branch: master ▾ [workshops / 2019_Rstudio_Magic /](#)

 Create new file  Upload files  Find file  History

 jennyrieck added our favoRite things ... Latest commit d818f26 6 hours ago

..

	R	more updates to manuscript example!	23 hours ago
	Rmd	added our favoRite things	6 hours ago
	external/images	reorganizing pngs	6 hours ago
	misc	reorganizing pngs	6 hours ago
	2019_Rstudio_Magic.Rproj	initial folder structure	5 days ago
	README.md	create readme	5 days ago
 README.md			
Rstudio magic for BrainHack Toronto 2019			

This works for us

Screenshot of a GitHub repository interface showing a list of commits.

Branch: master workshops / 2019_Rstudio_Magic / R /

Create new file Upload files Find file History

derekbeaton almost done now we hope Latest commit e87b65d 16 hours ago

..

File	Message	Time Ago
0_create_ADNI_data_base.R	fixed rownames/race coding	8 days ago
1_create_ADNI_data_tidyverse.R	fixed rownames/race coding	8 days ago
2_explore_summarytools.R	almost done now we hope	16 hours ago
3_explore_inspectdf.R	almost done now we hope	16 hours ago
4_explore_DataExplorer_one_liner.R	almost done now we hope	16 hours ago
5_linear_model.R	almost done now we hope	16 hours ago
6_covstatis_example.R	please don't collide.	2 days ago

This works for us

Code Issues 0 Pull requests 0 Projects 0 Wiki Insights

Branch: master workshops / 2019_Rstudio_Magic / Rmd / Create new file Upload files Find file History

derekbeaton small update Latest commit 74c5384 14 hours ago

1_a_Simple_RMarkdown_PDF_files/figure-latex more updates to manuscript example! 3 days ago

3_RMarkdown APA Manuscript_files updated numbers & structures 16 hours ago

1_a_Simple_RMarkdown_PDF.Rmd almost done now we hope 16 hours ago

1_a_Simple_RMarkdown_PDF.log whatever 2 days ago

1_a_Simple_RMarkdown_PDF.pdf more updates to manuscript example! 3 days ago

1_a_Simple_RMarkdown_PDF.tex tons of bells-and-whistles via the manuscript. 4 days ago

2_RStudioMagic_presentation.pdf small update 14 hours ago

2_RStudioMagic_presentation.rmd small update 14 hours ago

2_RStudioMagic_presentation.tex small update 14 hours ago

3_RMarkdown APA Manuscript.Rmd updated numbers & structures 16 hours ago

3_RMarkdown APA Manuscript.docx updated numbers & structures 16 hours ago

3_RMarkdown APA Manuscript.pdf updated numbers & structures 16 hours ago

3_RMarkdown APA Manuscript.tex updated numbers & structures 16 hours ago

r-references.bib updated numbers & structures 16 hours ago

Get the packages you need

```
#to install from CRAN
install.packages('devtools', dependencies = TRUE)

#to install from a git (requires the devtools package)
devtools::install_github(Gibbsdavidl/CatterPlots)

#to install from a file
install.packages('/mypath/to/package/ADNIMERGE.tar.gz',
                 type='source', repos=NULL)
```

Some transition?

Here we dive into our stuff. We show a bit of the code here and there and explain syntax

Read in and create your dataframe

- ▶ ADNI Dataset adnimerge package
 - ▶ Reduce full dataset to only those participants (rows) and variables (columns) you're interested in
- ▶ Two methods to create your dataframe
 - ▶ using base R functions: `0_create_ADNI_data_base.R`
 - ▶ Using tidyverse functions:
`1_create_ADNI_data_tidyverse.R`

Create data 2 ways:

Base R

```
library(ADNIMERGE)

#####
## Load and clean data
#####

## 0.1 Specify the column names and participants you want (ie, baseline visit for all participants with MOCA>=16)
adni.cols <- c("RID", "VISCODE", "DX", "AGE", "PTGENDER", "PTEDUCAT", "PTETHCAT", "PTRACCAT", "APOE4", "FDG", "AV45",
adni.rows <- c(adnimerge$VISCODE=="b1" & adnimerge$MOCA>=16)
amerge_subset <- adnimerge[adni.rows,adni.cols]

#### remove participants with missing data
amerge_subset <- amerge_subset[complete.cases(amerge_subset),]

## 0.2 Bring in modified hachinks
amerge_subset$HMSCORE <- modhach$HMSCORE[match(amerge_subset$RID, modhach$RID)]

## 0.3 Manually change variable classes (remove class 'labelled')
amerge_subset$RID <- as.character(amerge_subset$RID)
```

tidyverse

```
1 library(ADNIMERGE)
2 library(tidyverse)
3 library(magrittr)
4
5 #####
6 ## Load and clean data
7 -
8
9 ## 1.1 Specify the column names you want and filter those participants at baseline with MOCA>=16 and complete data
10 adnimerge %>%
11   dplyr::select(RID, VISCODE, DX, AGE, PTGENDER, PTEDUCAT, PTETHCAT, PTRACCAT, APOE4, FDG, AV45, CDRSB, ADAS13, MOCA,
12   filter(VISCODE == "b1") %>%
13   filter(MOCA >= 16) %>%
14   drop_na() -> amerge_subset
15
16 ## 1.2 Bring in modified hachinks
17 amerge_subset %>% inner_join(modhach[,c("RID","HMSCORE")])
18
19 ## 1.3 Manually change variable classes (remove class 'labelled')
20 char.cols<-c("RID", "VISCODE", "DX", "PTGENDER", "PTETHCAT", "PTRACCAT")
21 amerge_subset[char.cols] %>% lapply(function(x) as.character(x))
22 num.cols<-c("AGE", "PTEDUCAT", "APOE4", "FDG", "AV45", "CDRSB", "ADAS13", "MOCA", "WholeBrain", "Hippocampus", "MidTemp",
23 amerge_subset[,num.cols] %>% lapply(function(x) as.numeric(x))
```

Exploring your data

- ▶ Many packages to help explore and describe your data:
 - ▶ `summarytools`: `2_explore_summarytools.R`
 - ▶ `inspectdf`: `3_explore_inspectdf.R`
 - ▶ `DataExplorer`: `4_explore_DataExplorer_one_liner.R`

Code w/ eval=F

Hard Break

- ▶ DataExplorer is dangerous
- ▶ Blind analyses can be *criminal*
 - ▶ de Leeuw paper quote
 - ▶ DEREK RANTS, PER USUAL.

Analyze your data

- ▶ Linear models: 5_linear_model.R

Screenshots / Code w/ eval=F

Get experimental

- ▶ Explain motivation, not method
- ▶ covSTATIS: 6_covstatis_example.R

Part 3: RMarkdown

RMarkdown

- ▶ What it is /why to use it
- ▶ A short deviation for LaTeX, and new helpers: kable & kableExtra
 - ▶ A taxonomy and how to approach this *Tying it all together through here 1: simple RMD Plot-based visuals*
 - ▶ Base, gt, ggplot, grobTable()/grid/gridExtra
 - ▶ 2: Slides (these ones here)
 - ▶ 3: Manuscripts!!
- ▶ Reporting/presentin

RMarkdown Don(u)'ts

- ▶ Don't hardcode values
- ▶ Don't hardcode absolute file paths
- ▶ Don't do complicated database queries
- ▶ Don't litter
 - ▶ avoid eval=FALSE
 - ▶ reduce repeated code by making functions
- ▶ Don't load unnecessary libraries
- ▶ More at: <https://emilyriederer.netlify.com/post/rmarkdown-driven-development/>

Part 4: Advanced R

Some advanced/other things we're not covering

- ▶ package development
- ▶ Shiny
- ▶ SQL
- ▶ C/C++
- ▶ R2D3

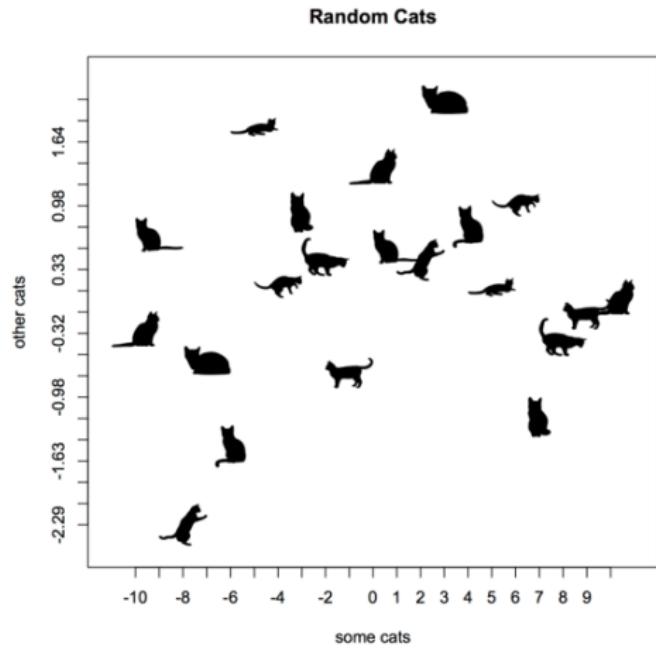
A few of our favorite things

- ▶ Fun R do-dads

CatterPlot for feline based graphics:

► <https://github.com/Gibbsdavidl/CatterPlots>

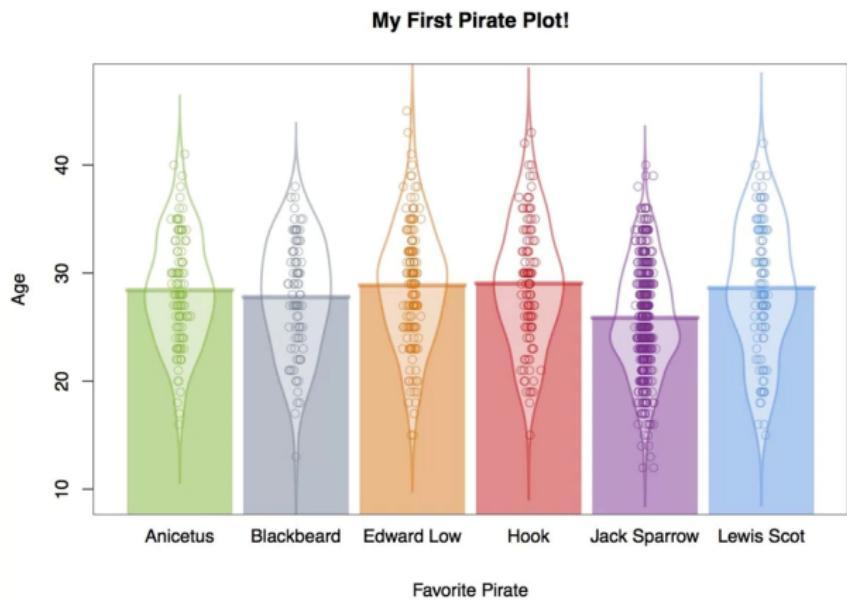
```
devtools::install_github(Gibbsdavidl/CatterPlots)
```



What's a pirate's favorite programming language?

- ▶ <https://cran.r-project.org/web/packages/yarrr/vignettes/pirateplot.html>

```
install.packages('yarrr')
```

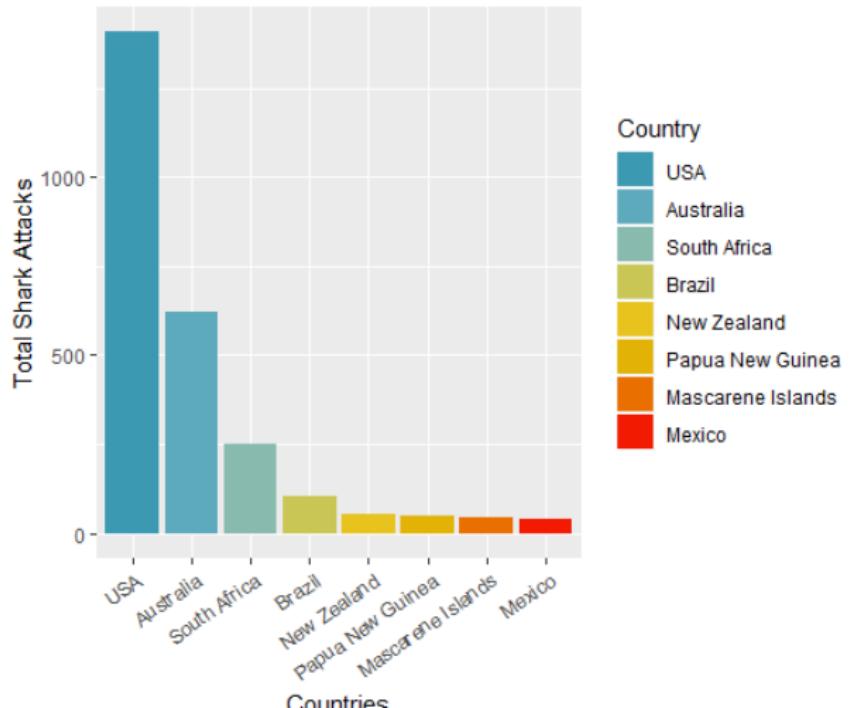


Color palettes to fit your mood

► <https://github.com/karthik/wesanderson>

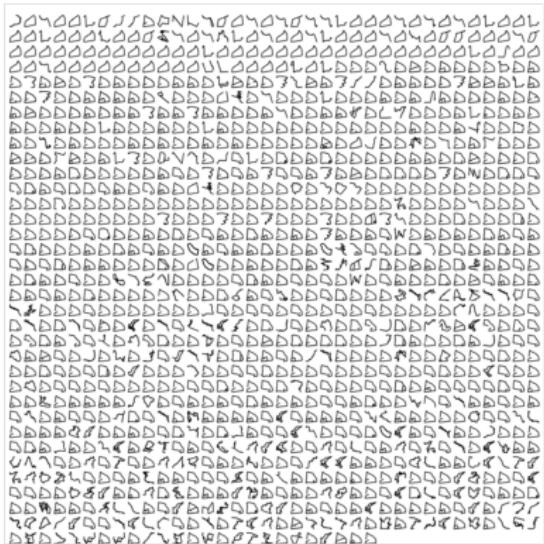
```
devtools::install_github(karthik/wesanderson)
```

Top countries with shark attacks
(Esteban was eaten)



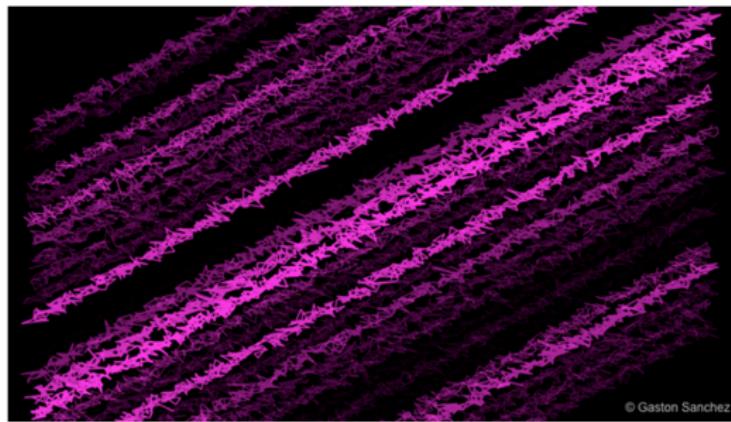
Mapping your Strava routes

- ▶ <https://www.r-bloggers.com/strava-rides-map-in-r/>
- ▶ ALSO <https://marcusvolz.com/?p=4068>
 - ▶ `devtools::install_github(marcusvolz/strava)`



Make aRt!

- ▶ R Graph Gallery
 - ▶ <http://www.r-graph-gallery.com/>
- ▶ Rtist: Gaston Sanchez
 - ▶ <http://gastonsanchez.com/Rtist/>



© Gaston Sanchez

```
# Pink Barbs
# -----
# generate pairs of x-y values
x <- seq(1, 100, length = 1000)
y <- x + rnorm(1000)

# pink_barbs.png
# set graphical parameters
op <- par(bg = "black", mar = rep(0, 4))
# plot
plot(x, y, type = "n")
for (i in seq(-80, 70, by = 5)) {
  lines(x + rnorm(1000), x + i + rnorm(1000, 2), pch = 19,
        col = hsv(0.85, 1, 1, runif(1000)),
        lwd = sample(seq(0.3, 2, length = 20), 1))
}
# signature
legend("bottomright", legend = "@ Gaston Sanchez", bty = "n",
       text.col = "gray70")
# reset par
par(op)
dev.off()
```