A Brief Tour Through the Wonderful World of R

OCLBASA Q3 Quarterly Social
Ryan Benz · September 14, 2018

There Are LOTS of Great Tools for Doing Data Analysis

and many more…

R is a Great Choice
For Anyone Working with Data

r-project.org Home Page

The R Project for Statistical Computing

[Home]

Download

CRAN

R Project

About R
Logo
Contributors

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To **download R**, please choose your preferred CRAN mirror.

If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email

R is a Great Choice
For Anyone Working with Data

r-project.org Home Page

The R Project for Statistical Computing

[Home]

Download

CRAN

visualization

R Project

About R

Logo

Contributors

More than just a language

data analysis

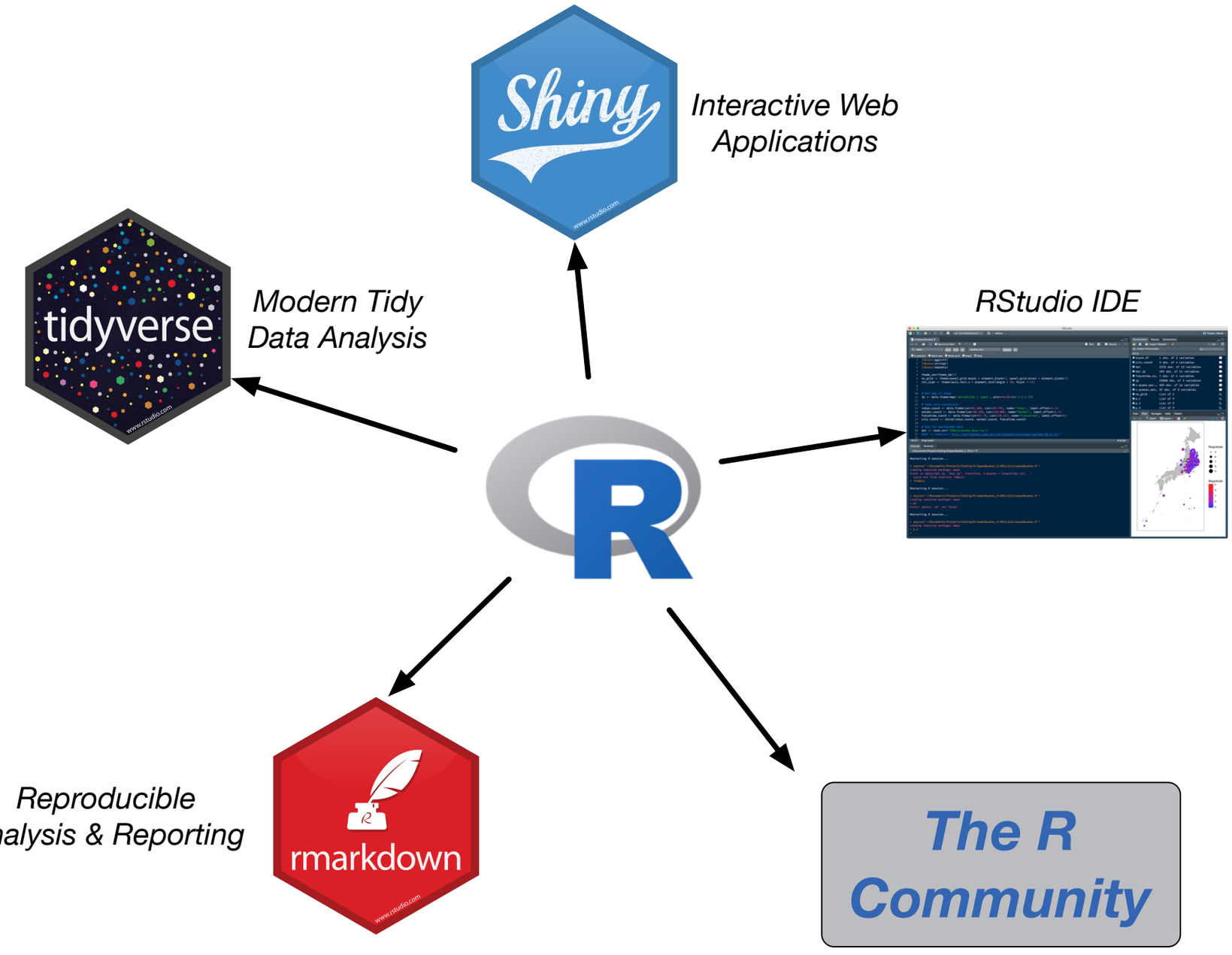
Getting Started

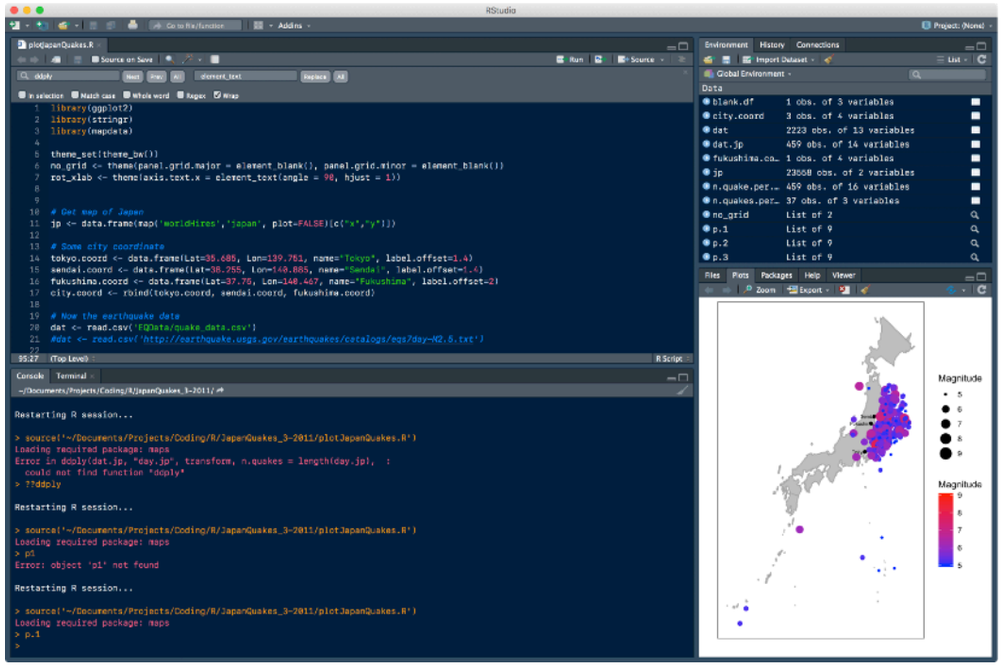
R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To **download R**, please choose your preferred CRAN mirror.

If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email

The World of R is Big!

here are just a few interesting highlights…

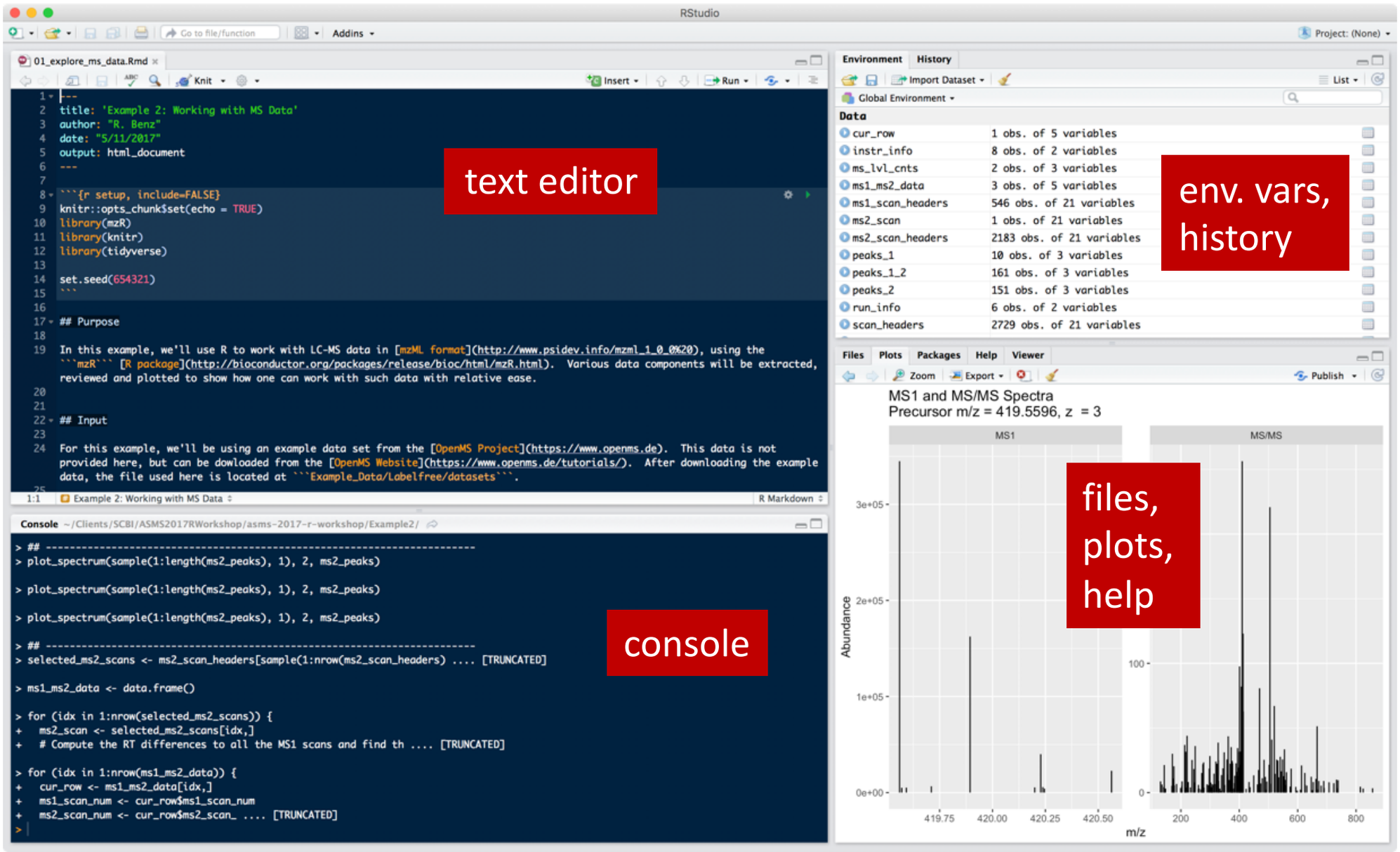




RStudio

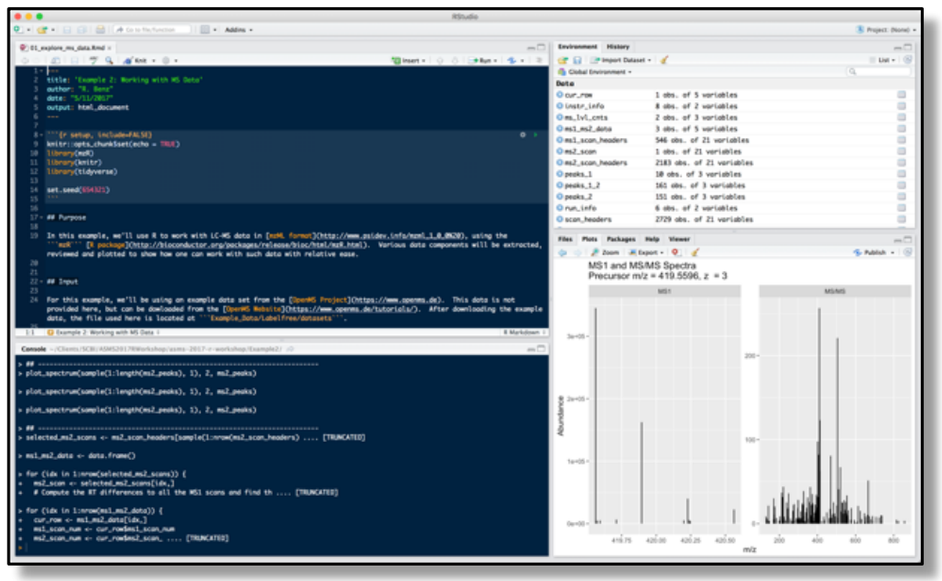
The IDE

RStudio is a Powerful IDE

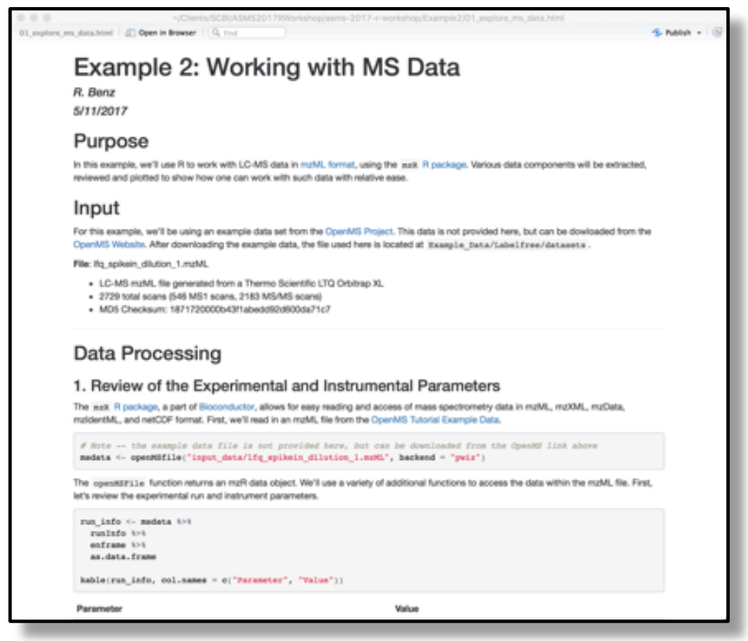


… And it Does More than Just Edit Code

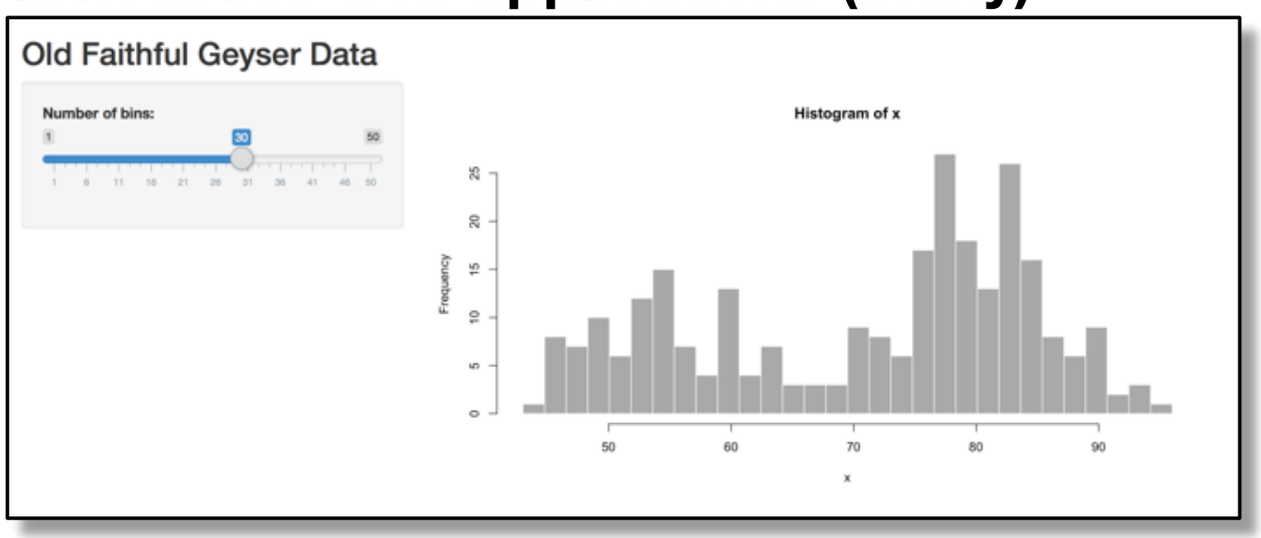
Data analysis workspace



Reproducible reporting



Interactive web applications (Shiny)



- Git & SVN integration
- Publishing to the web
- Write R packages
- Work in the cloud with RStudio Server

The tidyverse

The `tidyverse` Facilitates Modern, Streamlined Data Analysis

The `tidyverse` an opinionated collection of R packages designed for data science

- *tidyverse.org*

rstudio.com
tidyverse.org

… and LOTS more.
You can even contribute!

The Consistency of Tidy Data Opens Up a World of Possibilities

Most real world data starts out messy.

Working with messy data is hard!

All tidy data is similar in structure.

This makes it easier to work with & to develop analysis tools

See http://vita.had.co.nz/papers/tidy-data.html for more info

Data Pipelines with `dplyr`

Consider the `mtcars` dataset

```
> head(mtcars)
                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Data Pipelines with `dplyr`

Goal
Compute the mean horsepower
for each cylinder group

A simple first approach
might look like

```
hp_vals_4 <- c()
hp_vals_6 <- c()
hp_vals_8 <- c()
for (i in 1:nrow(mtcars)) {
  current_car <- mtcars[i,]

  cyl <- current_car$cyl
  hp <- current_car$hp

  if (cyl == 4) {
    #... accumulate values
    hp_vals_4 <- c(hp_vals_4, hp)
  } else if (cyl == 6) {
    #... accumulate values
  } #...
  #...
}
# compute summary stats
mean_hp_4 <- mean(hp_vals_4)
#...
```

Data Pipelines with `dplyr`

The `dplyr` approach is much simpler, easier to read, easier to understand

code

```
mtcars %>%
  group_by(cyl) %>%
  summarize(n = n(),
            mean_hp = mean(hp),
            median_hp = median(hp),
            stdev_hp = sd(hp))
```

output

```
# A tibble: 3 x 5
    cyl     n   mean_hp median_hp stdev_hp
  <dbl> <int>     <dbl>     <dbl>    <dbl>
1     4    11  82.63636      91.0 20.93453
2     6     7 122.28571     110.0 24.26049
3     8    14 209.21429     192.5 50.97689
```

Shiny

Shiny Allows You to Build Interactive Web Applications

Define the interface

```
library(shiny)

shinyUI(fluidPage(

  # Application title
  titlePanel("Old Faithful Geyser Data"),

  # Sidebar with a slider input for number of bins
  sidebarLayout(
    sidebarPanel(
      sliderInput("bins",
                  "Number of bins:",
                  min = 1,
                  max = 50,
                  value = 30)
    ),

    # Show a plot of the generated distribution
    mainPanel(
      plotOutput("distPlot")
    )
  )
))
```

Define logic, functionality

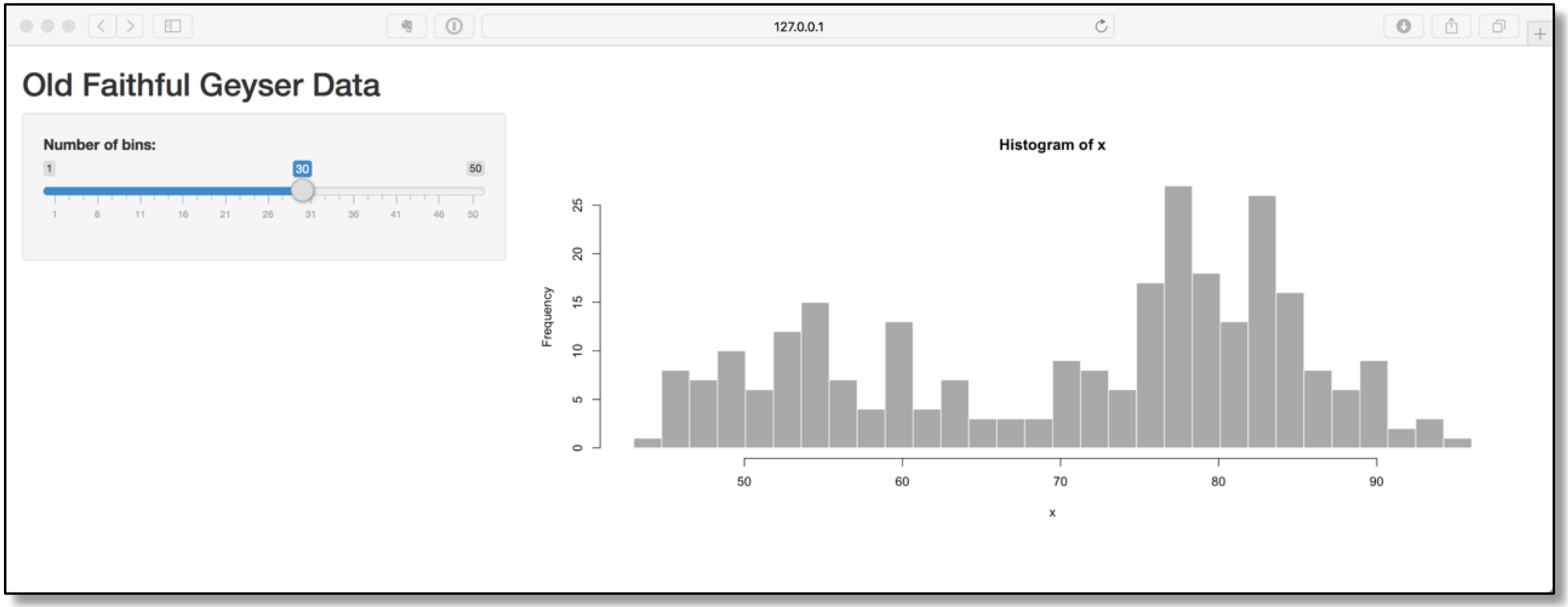
```
library(shiny)

shinyServer(function(input, output) {

  output$distPlot <- renderPlot({

    # generate bins based on input$bins from ui.R
    x    <- faithful[, 2]
    bins <- seq(min(x), max(x), length.out = input$bins + 1)

    # draw the histogram with the specified number of bins
    hist(x, breaks = bins, col = 'darkgray', border = 'white')

  })

})
```

default RStudio Shiny example

All R code!
No need to know JavaScript, CSS, XML

A Shiny Example

Default Shiny Demo App



A full, interactive (albeit simple) web app created in <50 lines of code

All in R, no need to understand web technologies

RMarkdown

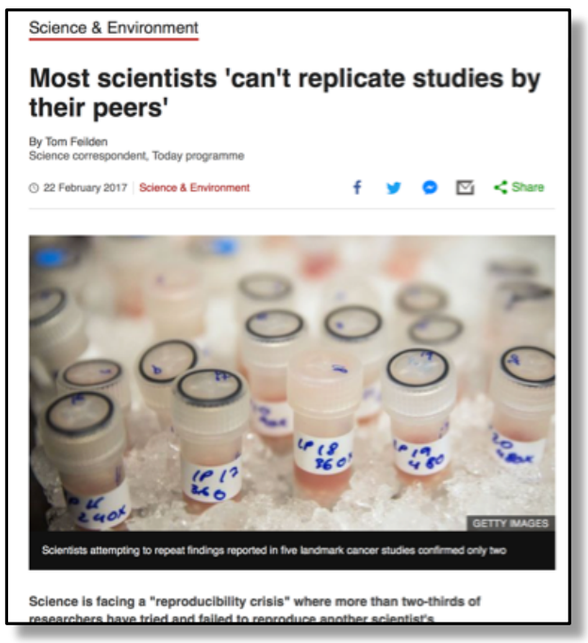
Reproducible Data Analysis is an Important Topic!

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

PLoS Medicine, Aug 2005 2(8)



Science & Environment

Most scientists 'can't replicate studies by their peers'

By Tom Feilden
Science correspondent, Today programme

22 February 2017 | Science & Environment

GETTY IMAGES

Scientists attempting to repeat findings reported in five landmark cancer studies confirmed only two

Science is facing a "reproducibility crisis" where more than two-thirds of researchers have tried and failed to reproduce another scientist's

Slatest YOUR NEWS COMPANION MARCH 3 2016 2:10 PM

Psychologists Call Out the Study That Called Out the Field of Psychology

By Rachel E. Gross

Remember that study that found that most psychology studies were wrong? Yeah, that study was wrong. That's the conclusion of four researchers who recently interrogated the methods of that study, which itself interrogated the methods of 100 psychology studies to find that very few could be replicated. (Whoa.) Their damning commentary will be published Friday in the journal Science. (The scientific body that publishes the journal sent *Slate* an early copy.)

In case you missed the hullabaloo: A key feature of the scientific method is that scientific results should be reproducible—that is, if you run an experiment again, you should get the same results. If you don't, you've got a problem. And a problem is exactly what 270 scientists found last August, when they decided to try to reproduce 100 peer-reviewed journal studies in the field of social psychology. Only around 39 percent of the reproduced studies, they found, came up with similar results to the originals.

FUTURE TENSE THE CITIZEN'S GUIDE TO THE FUTURE. APRIL 15 2016 7:08 AM

FROM SLATE, NEW AMERICA, AND ASU

The Reproducibility Crisis Is Good for Science

Weak statistics are getting called out, and replication is gaining respect.

By Monya Baker

After reports of widespread problems in psychology and biomedicine, scientists have become increasingly anxious that many published studies do not stand up.

http://www.slate.com/blogs/the_slatest/2016/03/03/psychology_study_that_induced_the_reproducibility_crisis_was_wrong.html

http://www.bbc.com/news/science-environment-39054778

http://www.slate.com/articles/technology/future_tense/2016/04/the_reproducibility_crisis_is_good_for_science.html

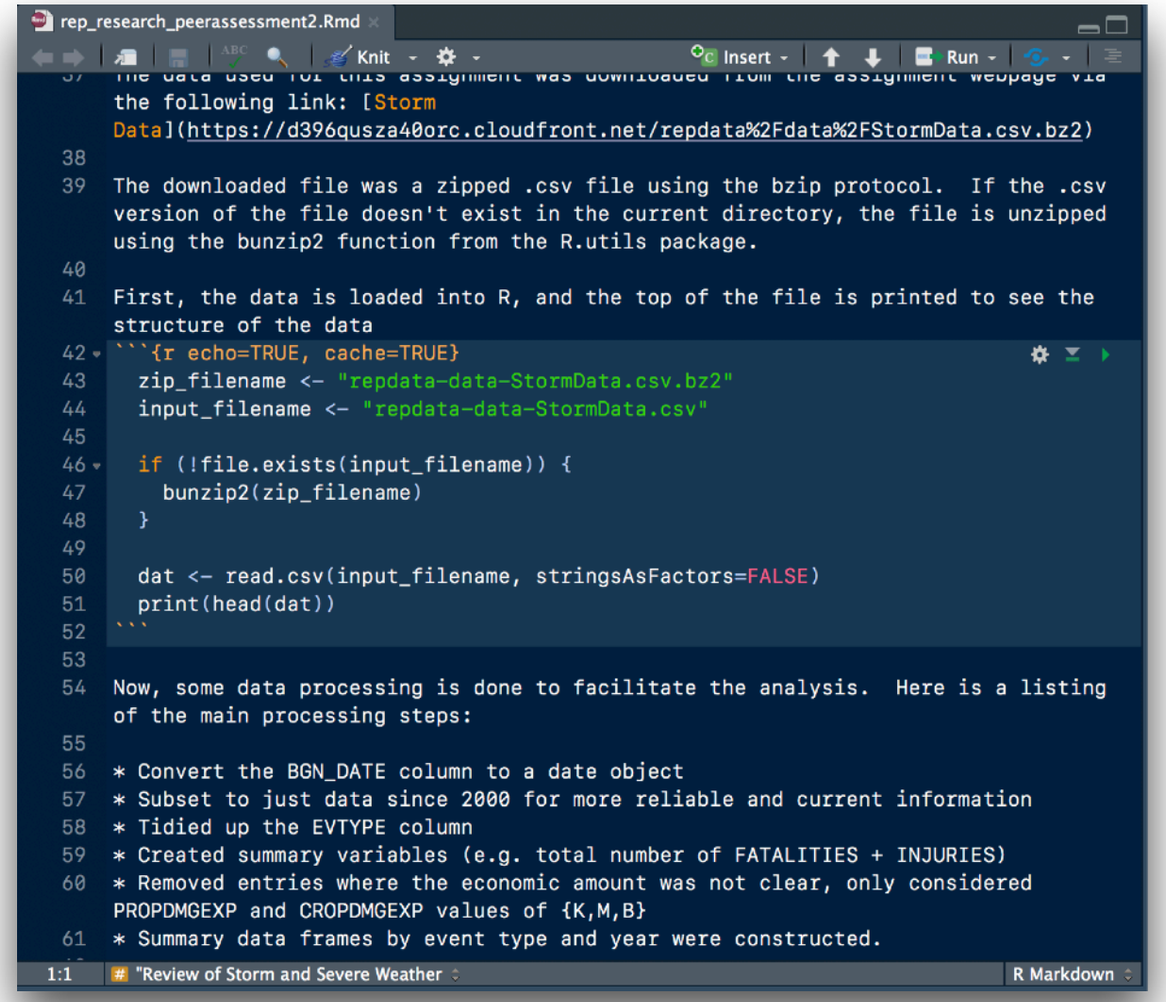
RMarkdown Provides a Way to Create Reproducible Analyses & Reports

Literate Programming: a style of programming that intermixes explanation of the code with the code itself

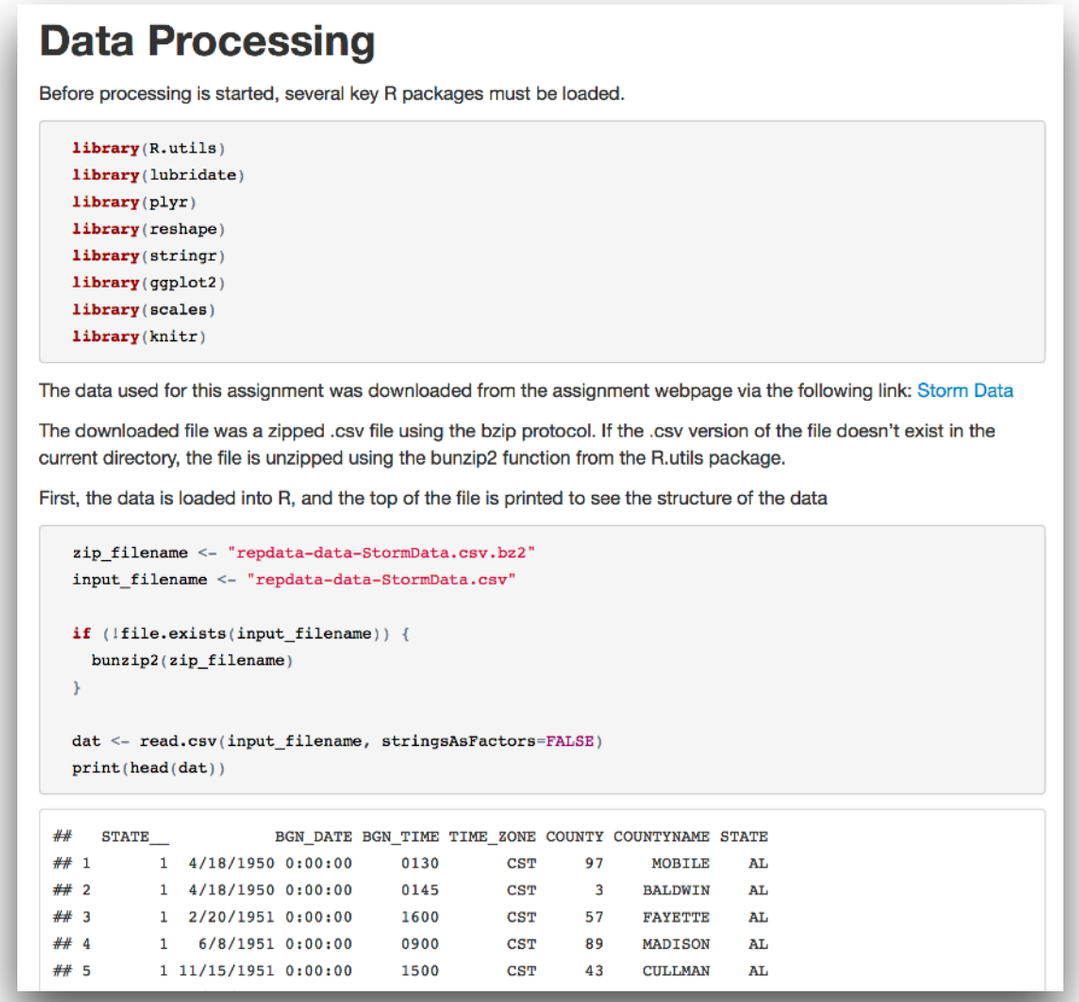
R + Markdown = RMarkdown

RMarkdown Provides a Way to Create Reproducible Analyses & Reports

Source Document
plain english (Markdown) and code (R)



Compiled HTML Output
Human readable, ready for distribution



Data Processing

Before processing is started, several key R packages must be loaded.

```
library(R.utils)
library(lubridate)
library(plyr)
library(reshape)
library(stringr)
library(ggplot2)
library(scales)
library(knitr)
```

The data used for this assignment was downloaded from the assignment webpage via the following link: Storm Data

The downloaded file was a zipped .csv file using the bzip protocol. If the .csv version of the file doesn't exist in the current directory, the file is unzipped using the bunzip2 function from the R.utils package.

First, the data is loaded into R, and the top of the file is printed to see the structure of the data

```
zip_filename <- "repdata-data-StormData.csv.bz2"
input_filename <- "repdata-data-StormData.csv"

if (!file.exists(input_filename)) {
  bunzip2(zip_filename)
}

dat <- read.csv(input_filename, stringsAsFactors=FALSE)
print(head(dat))
```

```
##   STATE__           BGN_DATE BGN_TIME TIME_ZONE COUNTY COUNTYNAME STATE
## 1       1  4/18/1950 0:00:00     0130       CST     97     MOBILE    AL
## 2       1  4/18/1950 0:00:00     0145       CST      3    BALDWIN    AL
## 3       1  2/20/1951 0:00:00     1600       CST     57    FAYETTE    AL
## 4       1   6/8/1951 0:00:00     0900       CST     89    MADISON    AL
## 5       1 11/15/1951 0:00:00     1500       CST     43    CULLMAN    AL
```

The R Community

R Has An Amazing Community!

- The R community has come a long way in the last ~5 years
- Active, diverse, welcoming, eager to help
- Supported by:
  R Core Team, R Foundation, R Consortium,
  RStudio, Microsoft, many others
- Conferences: UserR!, rstudio::conf,
  regional conferences throughout the world

R Users Groups Throughout the World

R Users Groups on Meet-up (51) and there are many more

Some of My Favorite Resources for Getting Started

- R for Data Science
  http://r4ds.had.co.nz

- RStudio Community
  https://community.rstudio.com

- R Tutorials and Presentations (via R Consortium)
  https://www.youtube.com/channel/UC_R5smHVXRYGhZYDJsnXTwg

- Online learning resource (via RStudio)
  https://www.rstudio.com/online-learning/