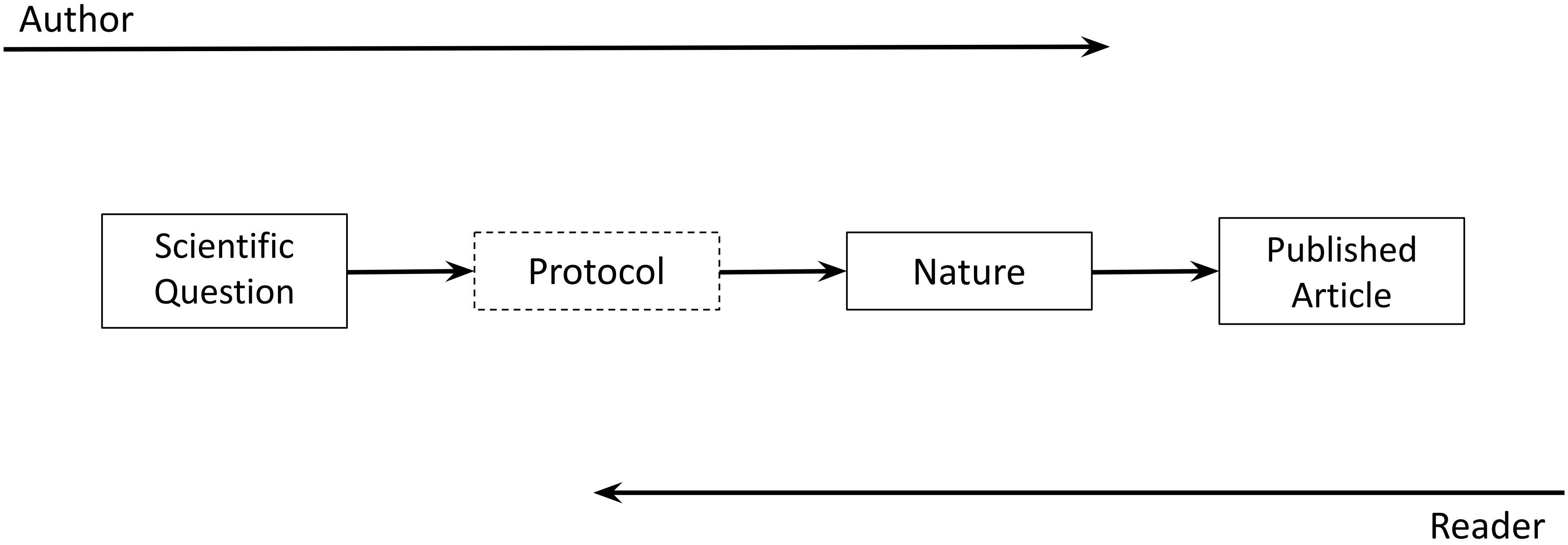
Data Wrangling in R

Reproducible Research

What is Reproducible Research?



Author

Scientific Question

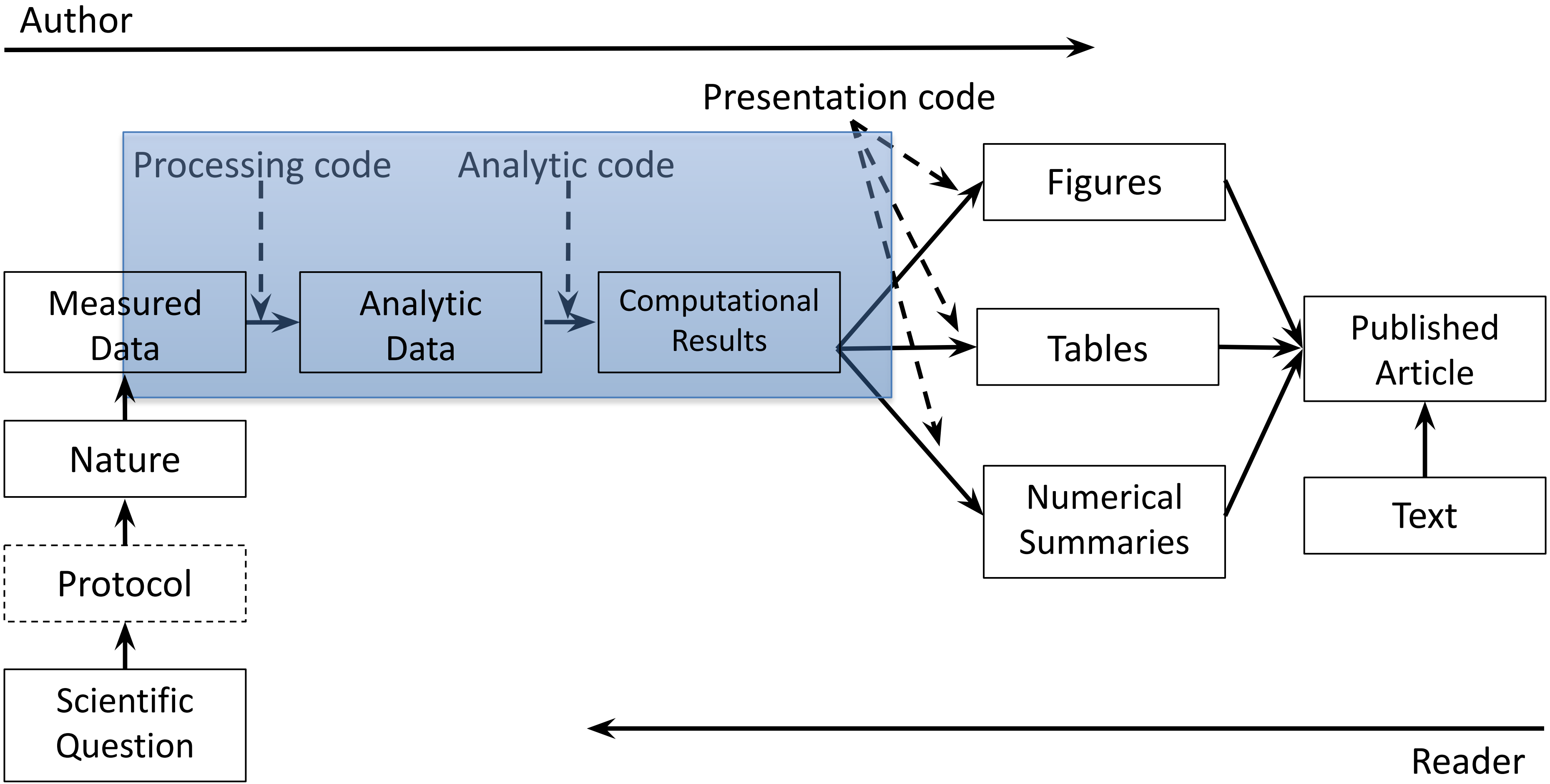
Protocol

Nature

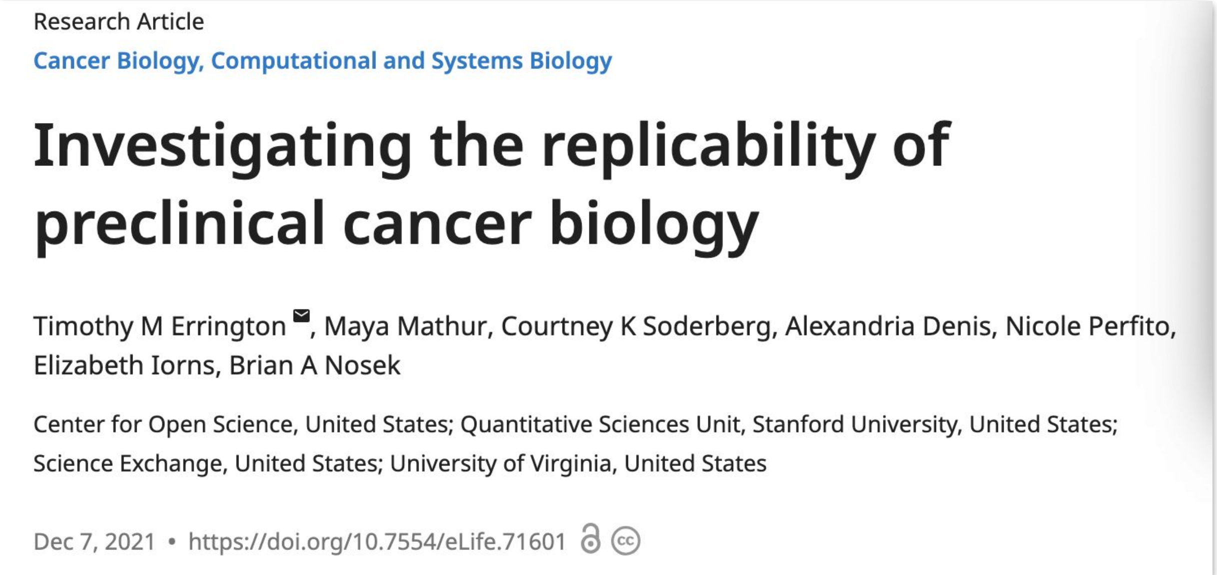
Published Article

Reader

What is Reproducible Research?



Why is this important?



Research Article

Cancer Biology, Computational and Systems Biology

Investigating the replicability of preclinical cancer biology

Timothy M Errington ✉, Maya Mathur, Courtney K Soderberg, Alexandria Denis, Nicole Perfito, Elizabeth Iorns, Brian A Nosek

Center for Open Science, United States; Quantitative Sciences Unit, Stanford University, United States; Science Exchange, United States; University of Virginia, United States

Dec 7, 2021 • https://doi.org/10.7554/eLife.71601

Edited by

Roger J Davis et al.

Reproducibility Project: Cancer Biology

Collection • Dec 10, 2014

https://elifesciences.org/articles/71601

Reproducibility in Cancer Biology: Challenges for assessing replicability in preclinical cancer biology

Timothy M Errington, Alexandria Denis, Nicole Perfito, Elizabeth Iorns, Brian A Nosek

Center for Open Science, United States; Science Exchange, United States; University of Virginia, United States

Dec 7, 2021 • https://doi.org/10.7554/eLife.67995

“We report the challenges confronted during a large-scale effort to replicate findings in cancer biology, and describe how improving **transparency and sharing** can make it easier to assess rigor and replicability and, therefore, to **increase research efficiency.**”

https://elifesciences.org/articles/67995

Your closest collaborator is you six months ago, but you don’t reply to emails

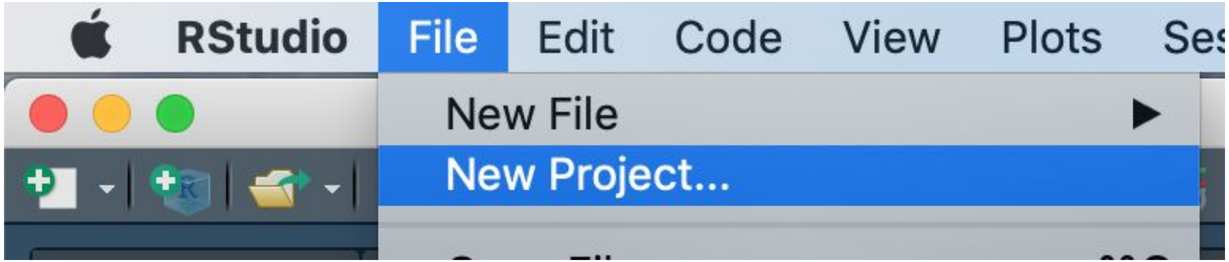
- Karl Broman

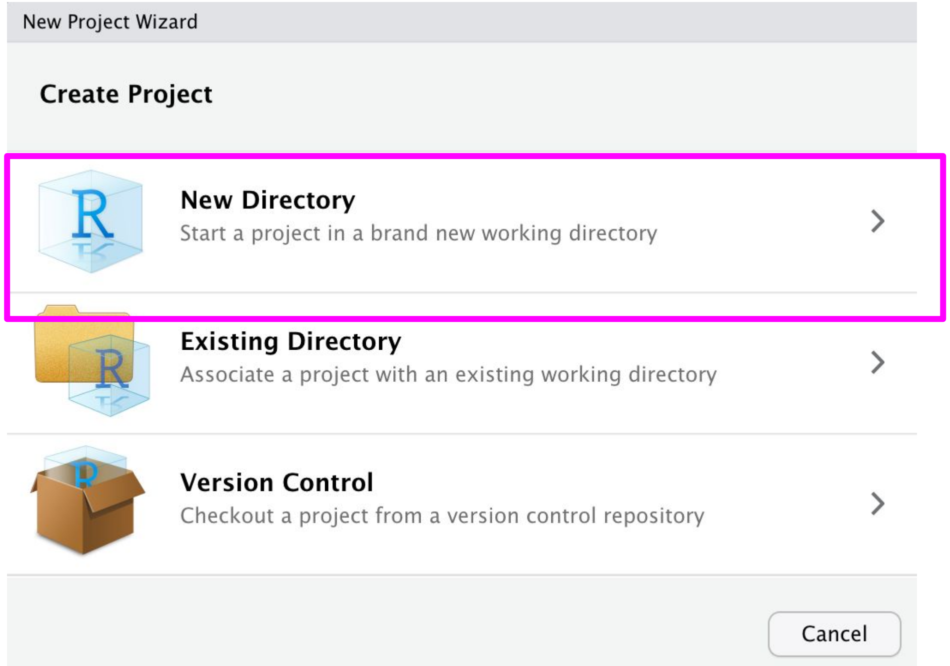
(http://kbroman.org/Tools4RR/assets/lectures/06_org_eda.pdf)

1. Code + documentation
2. Versions of software
3. Data provenance

https://tenor.com/view/struggle-cant-move-over-it-hard-no-gif-4734482

RStudio Projects



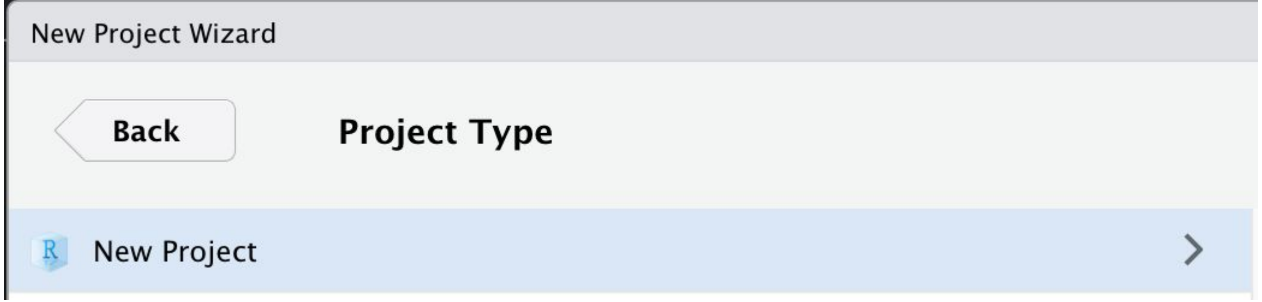

New Project Wizard

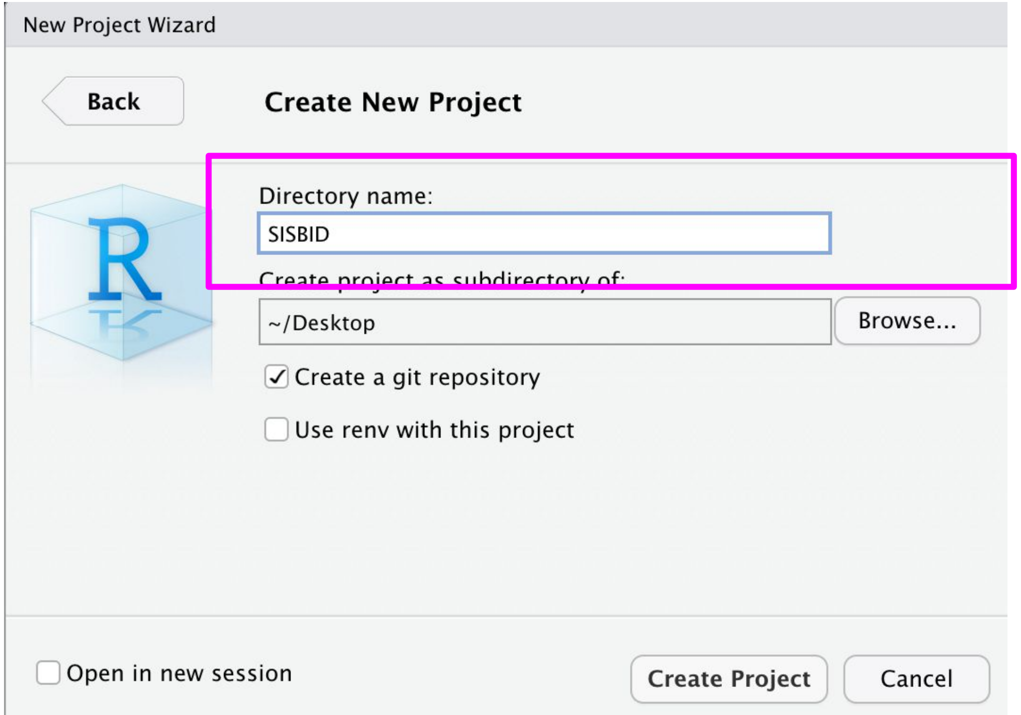
Create Project

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





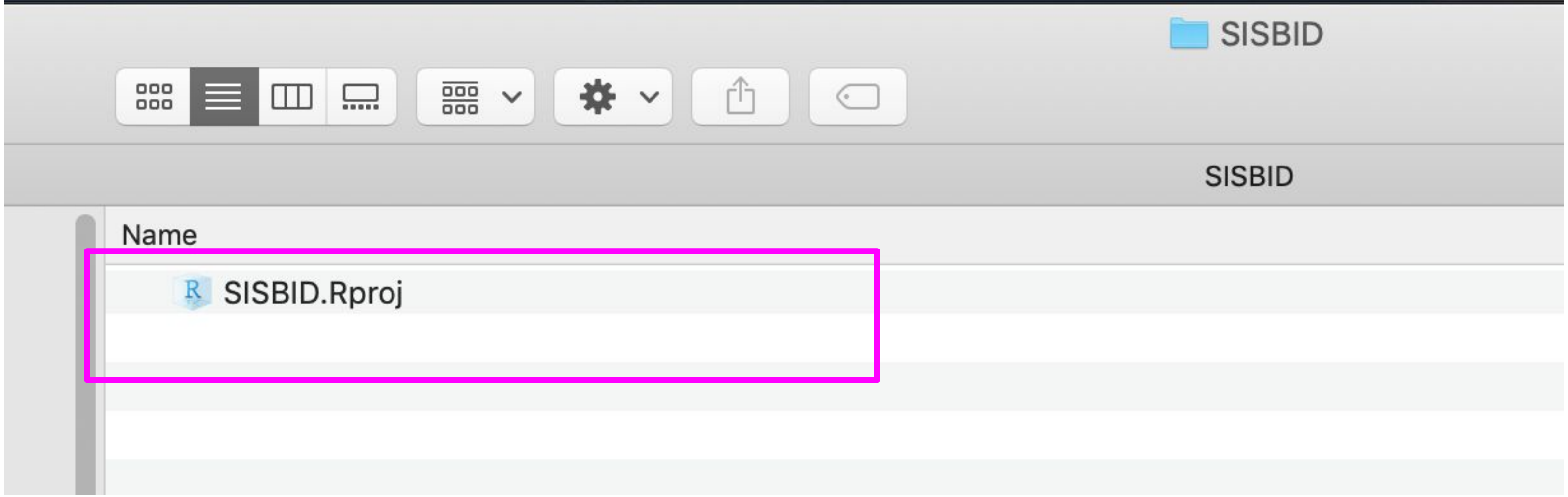
Directory name: SISBID

Create project as subdirectory of: ~/Desktop

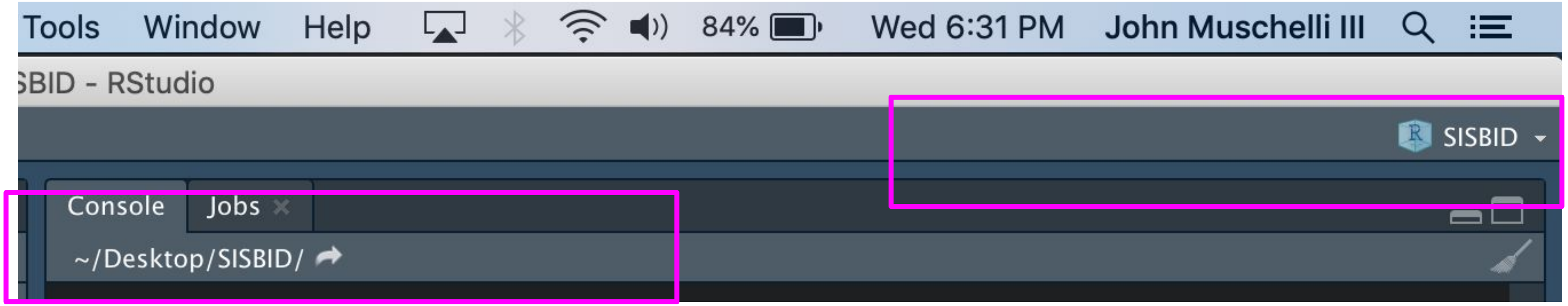
☑ Create a git repository

☐ Use renv with this project

☐ Open in new session



Double click on the Rproj file - opens RStudio



RStudio Projects

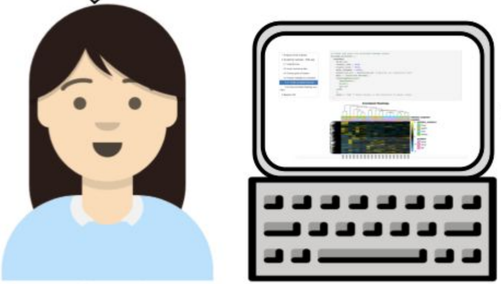
- Rproj - Open to directory of Project
- Everything is *relative* to working directory
- Zip up whole folder and send to someone else

`read_csv("C:/terrible/path/john/blah.csv")`

`read_csv("data/blah.csv")`

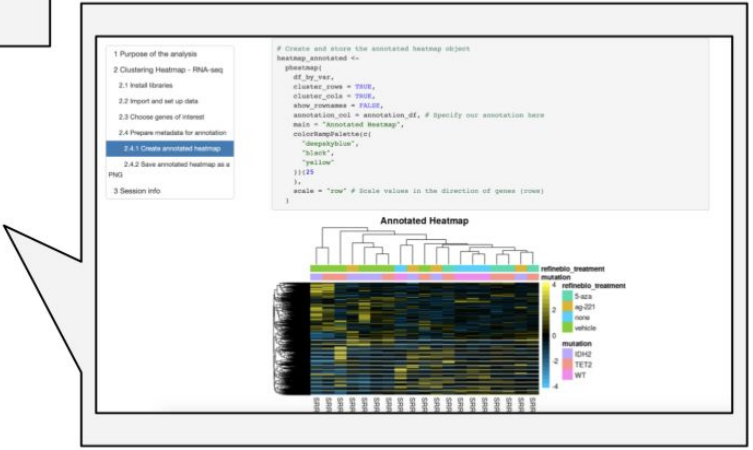
R Markdown lets you test your work

Working from this notebook allows me to interactively develop on my data analysis and write down my thoughts about the process all in one place!



Ruby the Researcher

RMarkdown is conducive to interactive development!



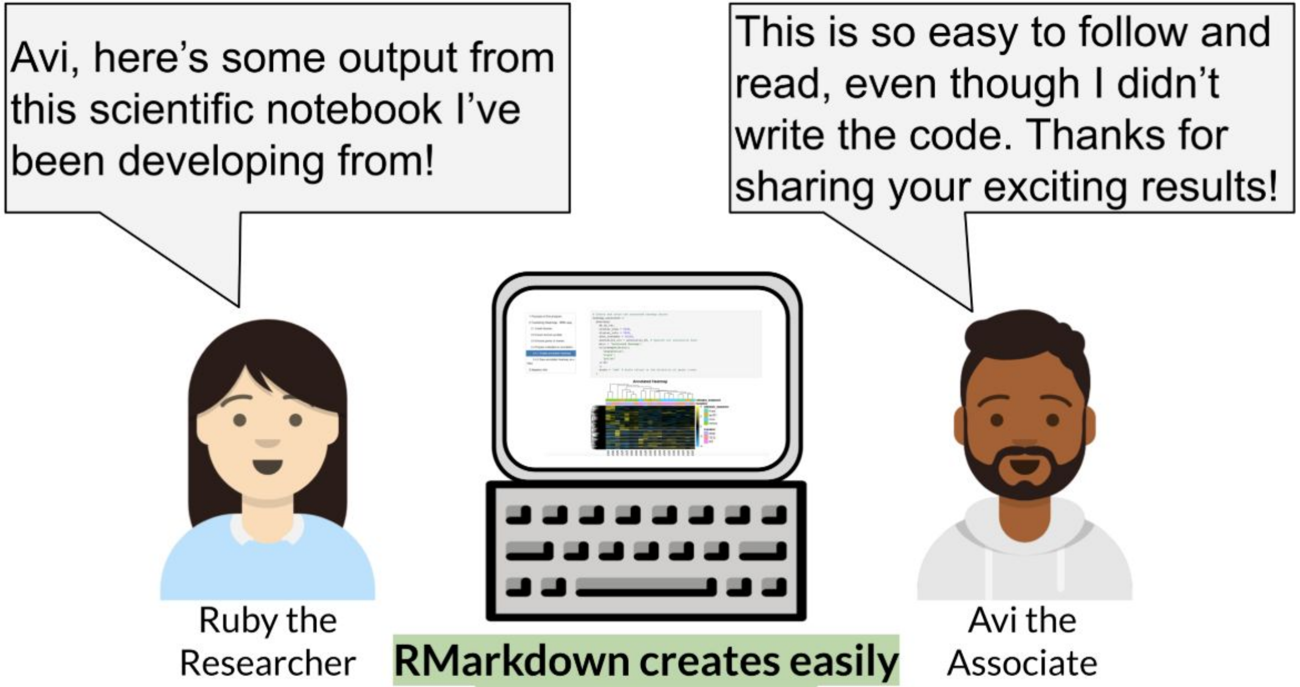
1 Purpose of the analysis
2 Clustering Heatmap - RNA-seq
  2.1 Install libraries
  2.2 Import and set up data
  2.3 Choose genes of interest
  2.4 Prepare metadata for annotation
    2.4.1 Create annotated heatmap
    2.4.2 Save annotated heatmap as a PNG
3 Session info

```
# Create and store the annotated heatmap object
heatmap_annotated <-
  pheatmap(
    df_by_var,
    cluster_rows = TRUE,
    cluster_cols = TRUE,
    show_rownames = FALSE,
    annotation_col = annotation_df, # Specify our annotation here
    main = "Annotated Heatmap",
    colorRampPalette(c(
      "deepskyblue",
      "black",
      "yellow"
    ))(25
    ),
    scale = "row" # Scale values in the direction of genes (rows)
  )
```

Annotated Heatmap

Image created by Candace Savonen using Avataars.

R Markdown allows you to more clearly show what you did



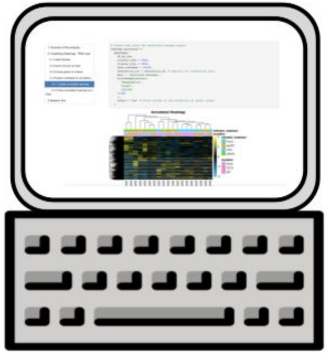
RMarkdown creates easily shareable output!

Image created by Candace Savonen using Avataars.

R Markdown makes it easier to update code and see results

Yay! I just got the data for 5 more samples. Because of my handy notebook set up, I can easily call one command and re-run the analysis so it is updated with the new samples included!

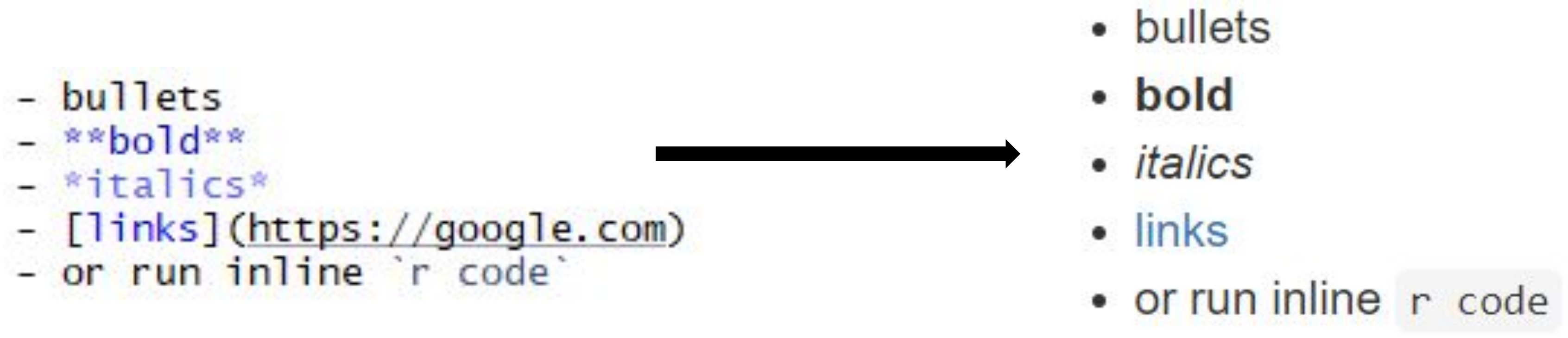
Ruby the Researcher



RMarkdown is handy for creating updateable reports!

Image created by Candace Savonen using Avataars.

The magic of Markdown

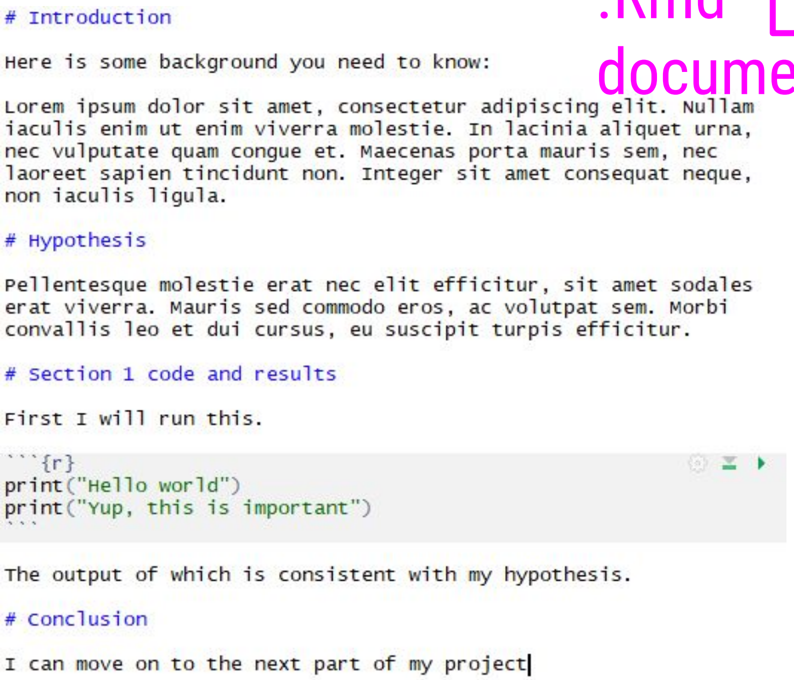


```
- bullets
- **bold**
- *italics*
- [links](https://google.com)
- or run inline `r code`
```

- bullets
- **bold**
- *italics*
- [links](https://google.com)
- or run inline `r code`

https://rmarkdown.rstudio.com/authoring_basics.html

.Rmd document



~~~
# Introduction

Here is some background you need to know:

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nullam
iaculis enim ut enim viverra molestie. In lacinia aliquet urna,
nec vulputate quam congue et. Maecenas porta mauris sem, nec
laoreet sapien tincidunt non. Integer sit amet consequat neque,
non iaculis ligula.

# Hypothesis

Pellentesque molestie erat nec elit efficitur, sit amet sodales
erat viverra. Mauris sed commodo eros, ac volutpat sem. Morbi
convallis leo et dui cursus, eu suscipit turpis efficitur.

# Section 1 code and results

First I will run this.

```{r}
print("Hello world")
print("Yup, this is important")
```

The output of which is consistent with my hypothesis.

# Conclusion

I can move on to the next part of my project
~~~

PDF, HTML or Word document

Introduction

Here is some background you need to know:

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nullam iaculis enim ut enim viverra molestie. In lacinia aliquet urna, nec vulputate quam congue et. Maecenas porta mauris sem, nec laoreet sapien tincidunt non. Integer sit amet consequat neque, non iaculis ligula.

Hypothesis

Pellentesque molestie erat nec elit efficitur, sit amet sodales erat viverra. Mauris sed commodo eros, ac volutpat sem. Morbi convallis leo et dui cursus, eu suscipit turpis efficitur.

Section 1 code and results

First I will run this.

```
print("Hello world")
```

```
## [1] "Hello world"
```

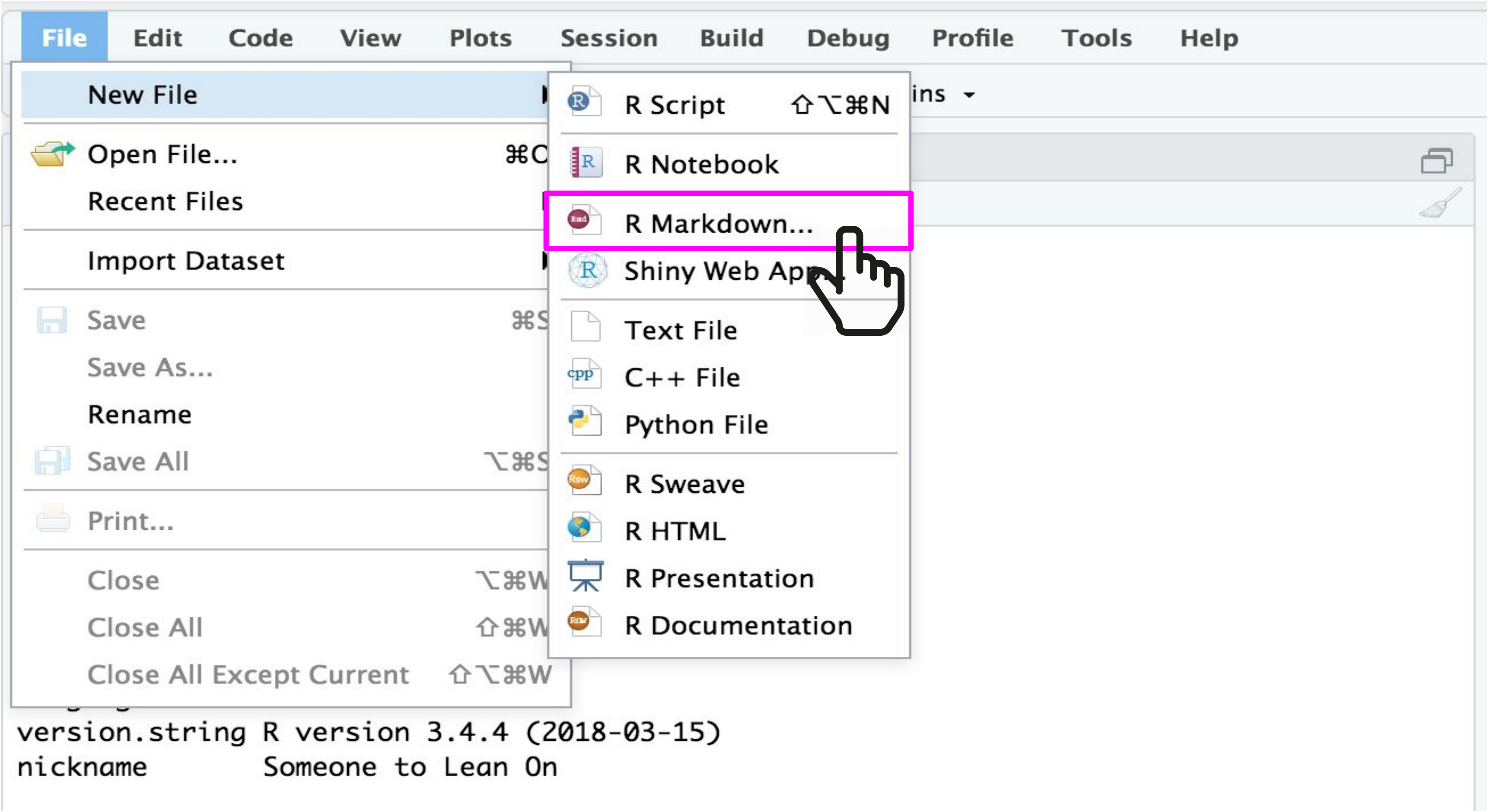
```
print("Yup, this is important")
```

```
## [1] "Yup, this is important"
```

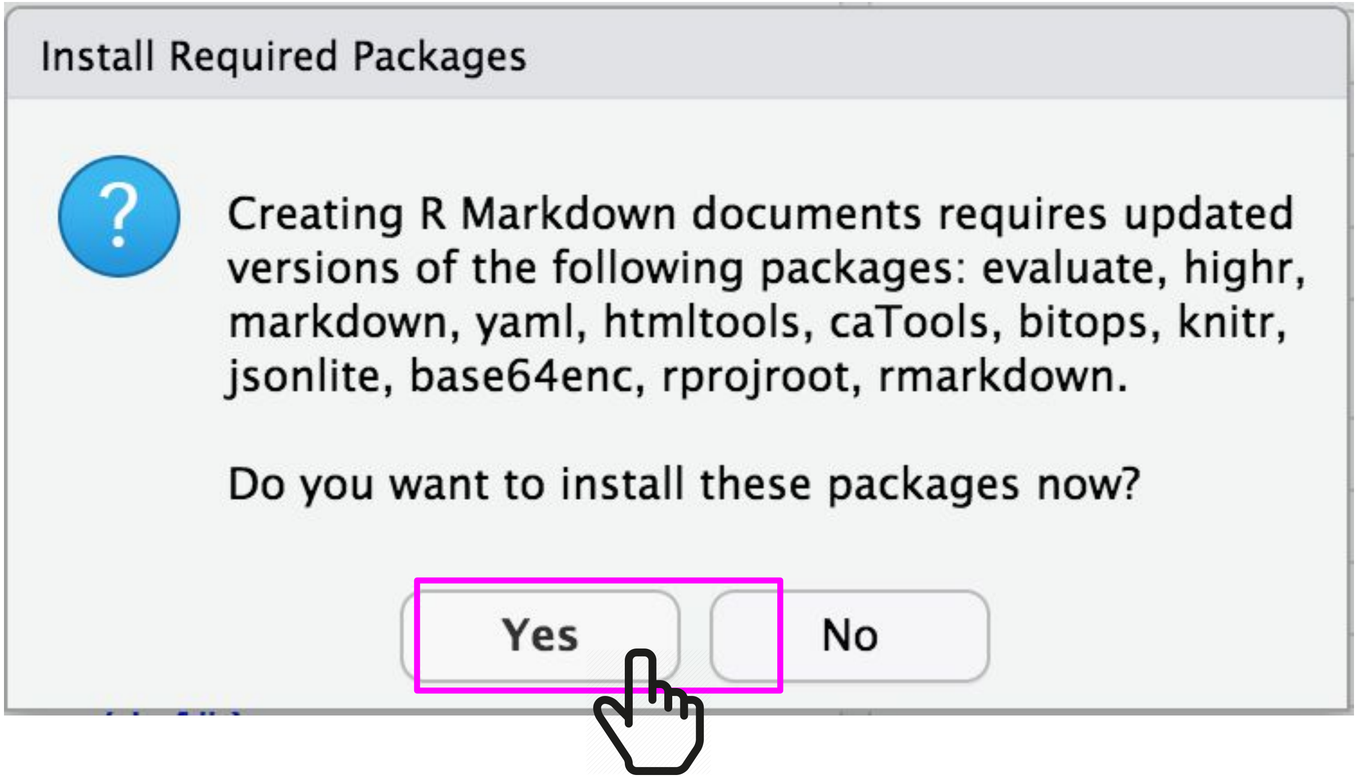
The output of which is consistent with my hypothesis.

Conclusion

I can move on to the next part of my project



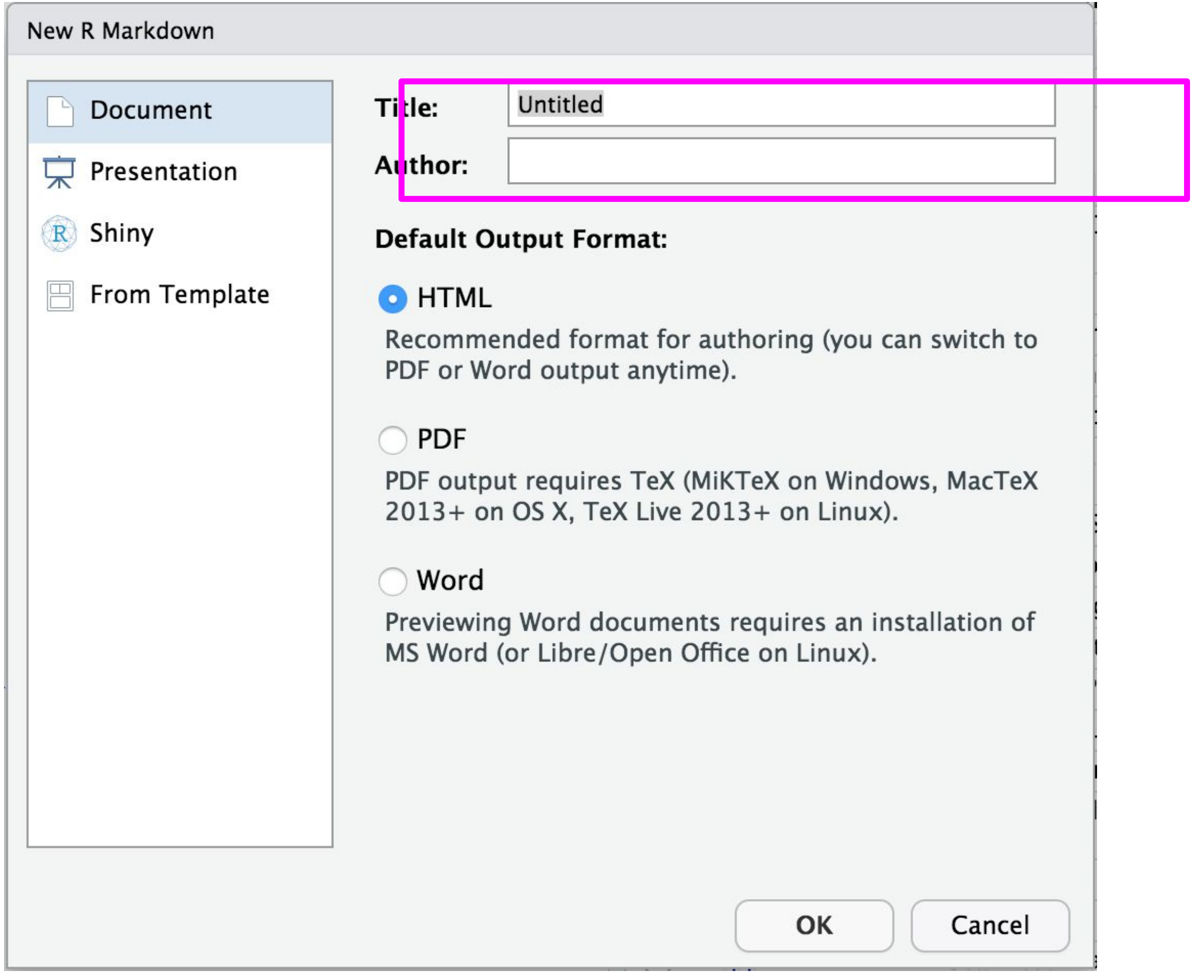
```
version.string R version 3.4.4 (2018-03-15)
nickname       Someone to Lean On
```



Install Required Packages

Creating R Markdown documents requires updated versions of the following packages: evaluate, highr, markdown, yaml, htmltools, caTools, bitops, knitr, jsonlite, base64enc, rprojroot, rmarkdown.

Do you want to install these packages now?


New R Markdown

Title: Untitled

Author:

Default Output Format:

● HTML

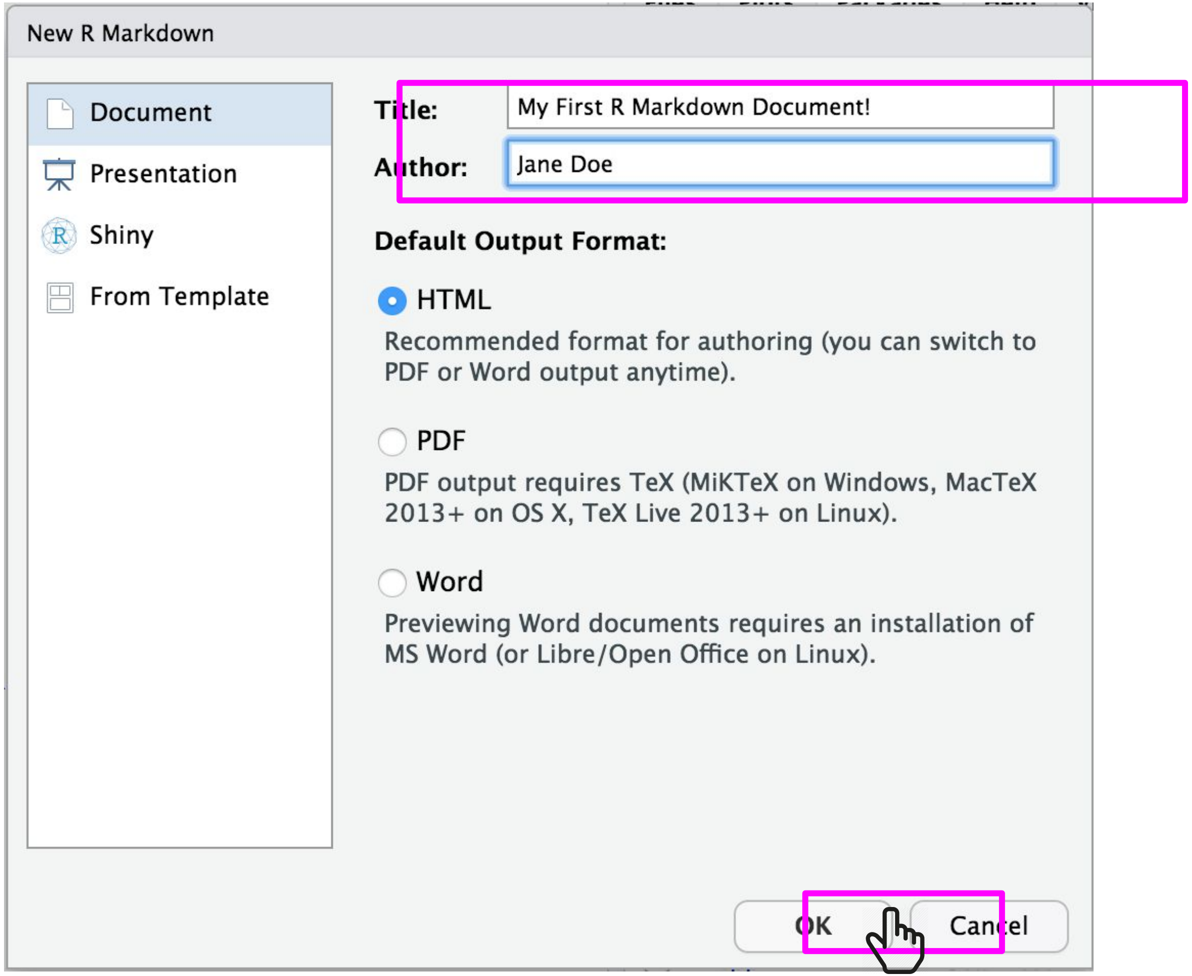
Recommended format for authoring (you can switch to PDF or Word output anytime).

○ PDF

PDF output requires TeX (MiKTeX on Windows, MacTeX 2013+ on OS X, TeX Live 2013+ on Linux).

○ Word

Previewing Word documents requires an installation of MS Word (or Libre/Open Office on Linux).



Title: My First R Markdown Document!

Author: Jane Doe

Default Output Format:

● HTML

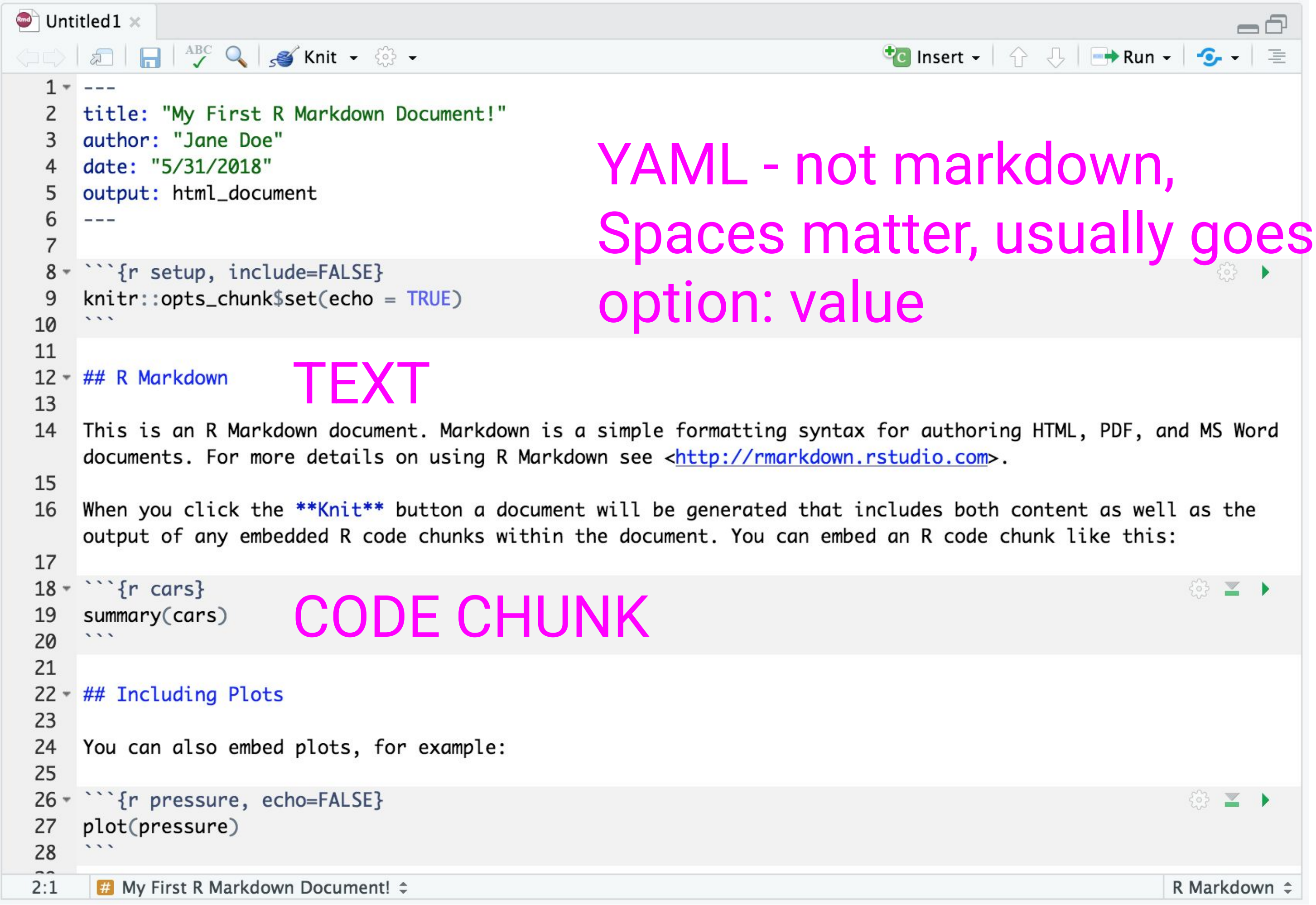
Recommended format for authoring (you can switch to PDF or Word output anytime).

○ PDF

PDF output requires TeX (MiKTeX on Windows, MacTeX 2013+ on OS X, TeX Live 2013+ on Linux).

○ Word

Previewing Word documents requires an installation of MS Word (or Libre/Open Office on Linux).



~~~
---
title: "My First R Markdown Document!"
author: "Jane Doe"
date: "5/31/2018"
output: html_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
~~~

R Markdown

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>.

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. You can embed an R code chunk like this:

```{r cars}
summary(cars)
```

Including Plots

You can also embed plots, for example:

```{r pressure, echo=FALSE}
plot(pressure)
```
```

YAML - not markdown, Spaces matter, usually goes option: value

TEXT

CODE CHUNK
```

~~~
---
title: "My First R Markdown Document!"
author: "Jane Doe"
date: "5/31/2018"
output: html_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
~~~

## R Markdown

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>.

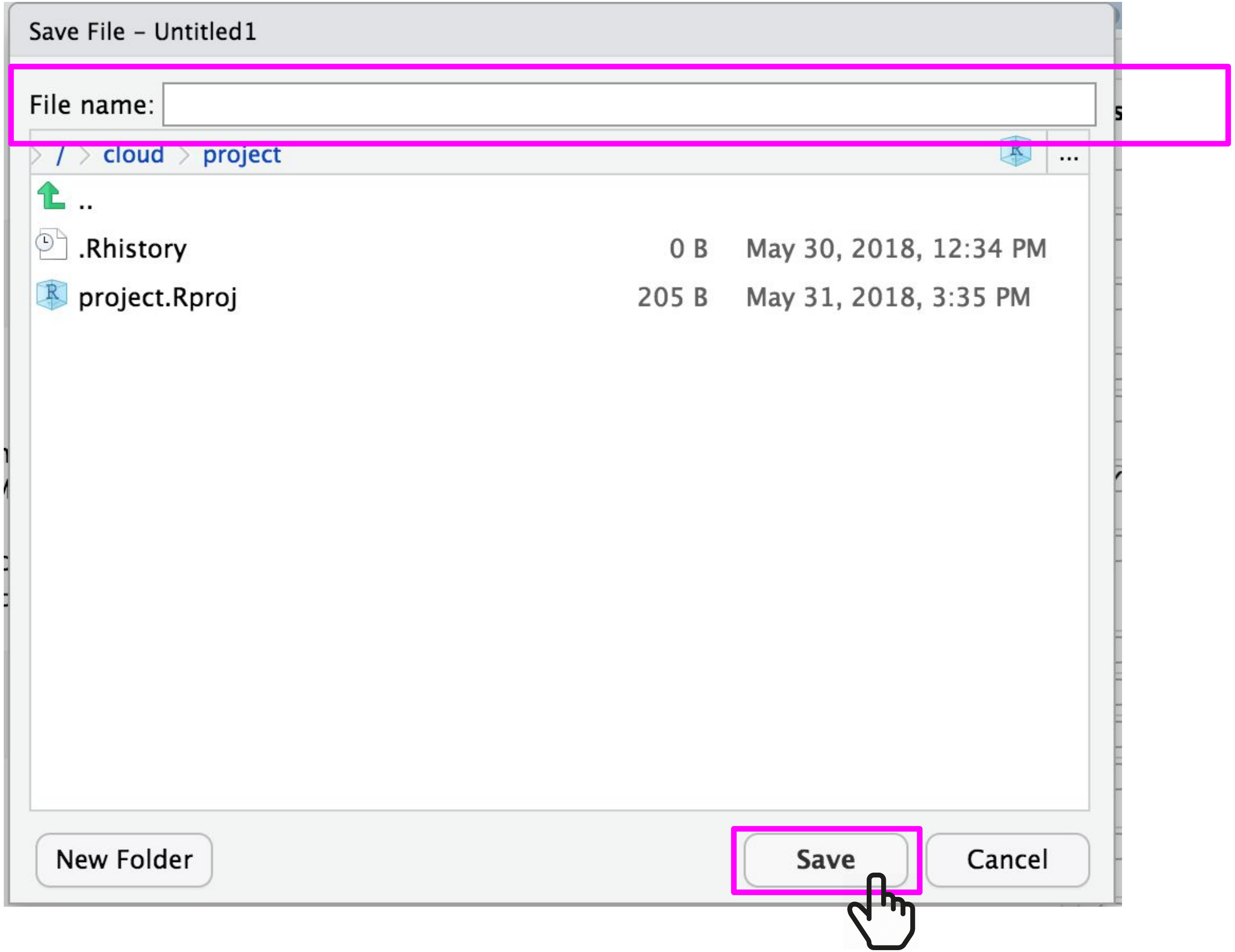
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```{r cars}
summary(cars)
```

Including Plots

You can also embed plots, for example:

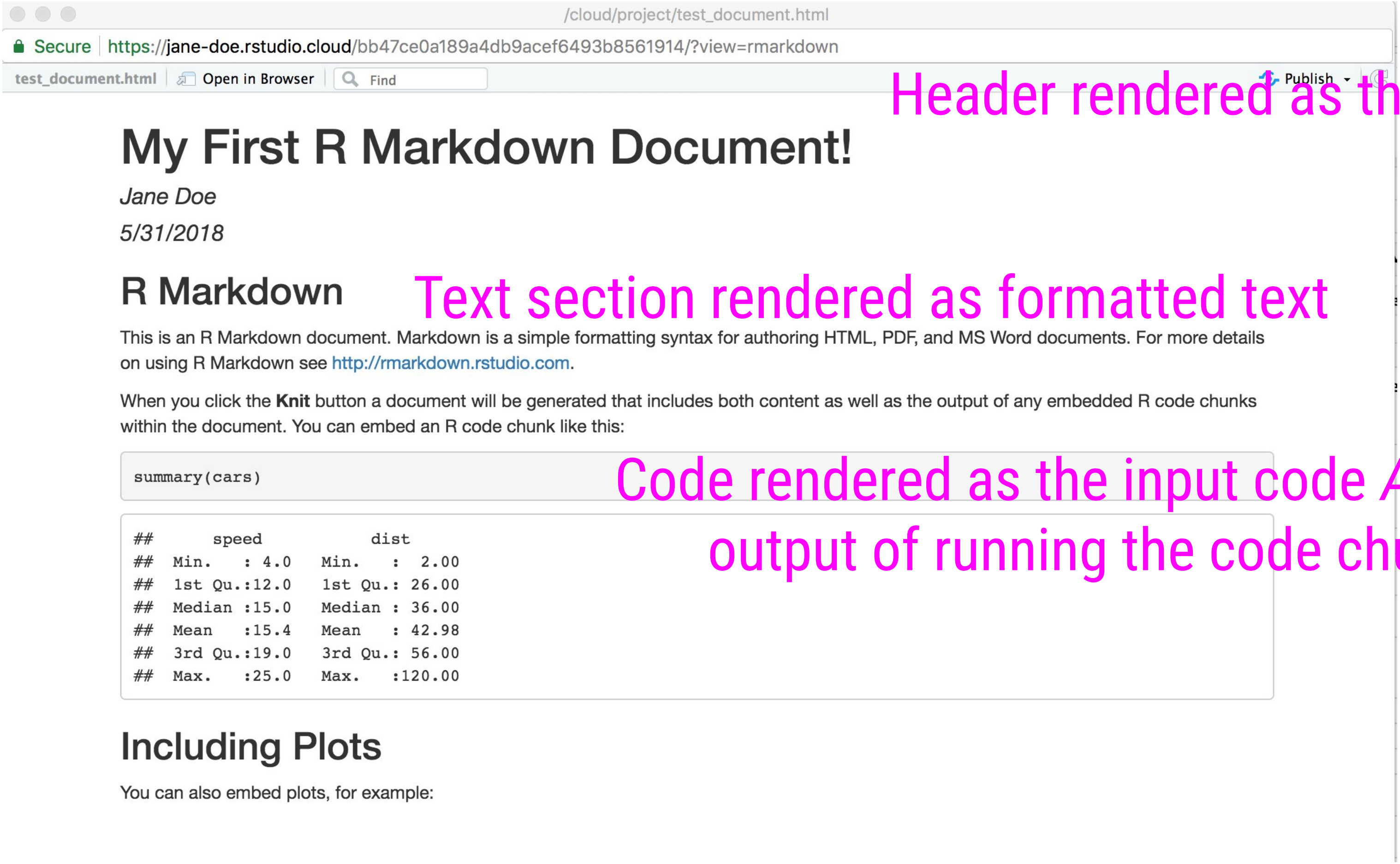
```{r pressure, echo=FALSE}
plot(pressure)
```
```
```


Save File – Untitled1

File name:

/ › cloud › project

.Rhistory	0 B	May 30, 2018, 12:34 PM
project.Rproj	205 B	May 31, 2018, 3:35 PM



My First R Markdown Document!

Jane Doe

5/31/2018

R Markdown

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.

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. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00  
```

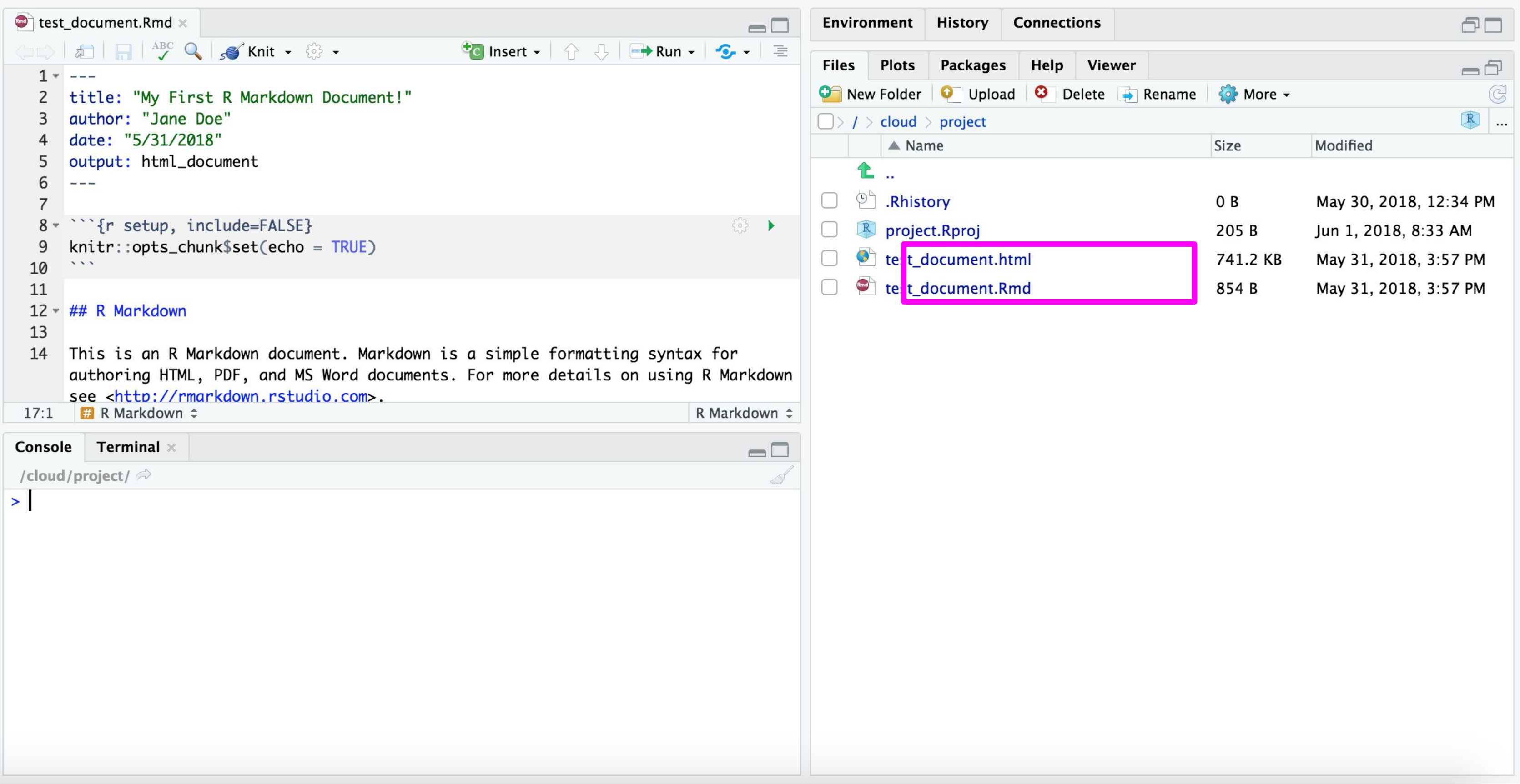
Including Plots

You can also embed plots, for example:

Header rendered as the title

Text section rendered as formatted text

Code rendered as the input code *AND* the output of running the code chunk



~~~
---
title: "My First R Markdown Document!"
author: "Jane Doe"
date: "5/31/2018"
output: html_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
~~~

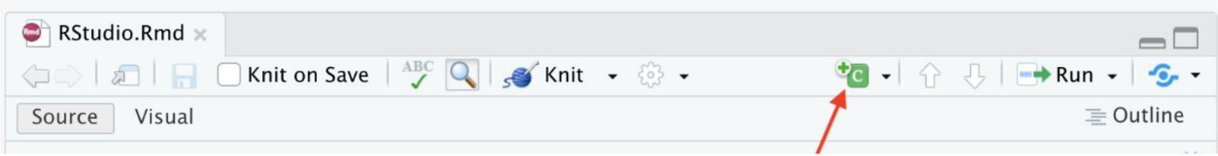
R Markdown

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>.
```

/ > cloud > project

Name	Size	Modified
..		
.Rhistory	0 B	May 30, 2018, 12:34 PM
project.Rproj	205 B	Jun 1, 2018, 8:33 AM
test_document.html	741.2 KB	May 31, 2018, 3:57 PM
test_document.Rmd	854 B	May 31, 2018, 3:57 PM

```
/cloud/project/
>
```
```

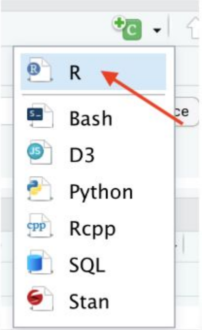
# Create Chunks

To create a new R code chunk:

- Use the insert code chunk button at the top of RStudio.

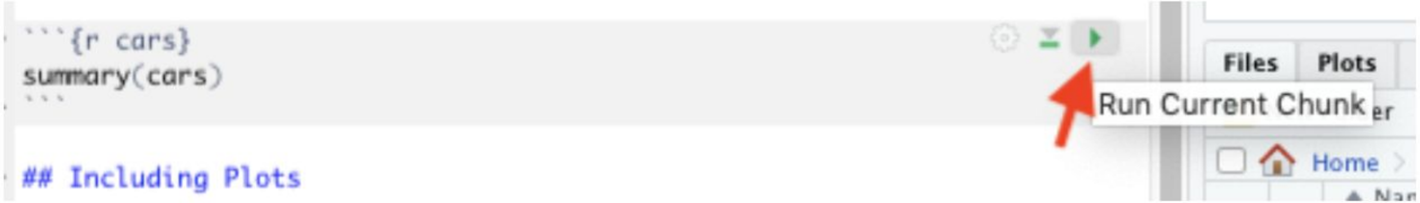


- Select R (default) as the language:



# Run code in a chunk

Clicking the run (play) button runs the code in the chunk.

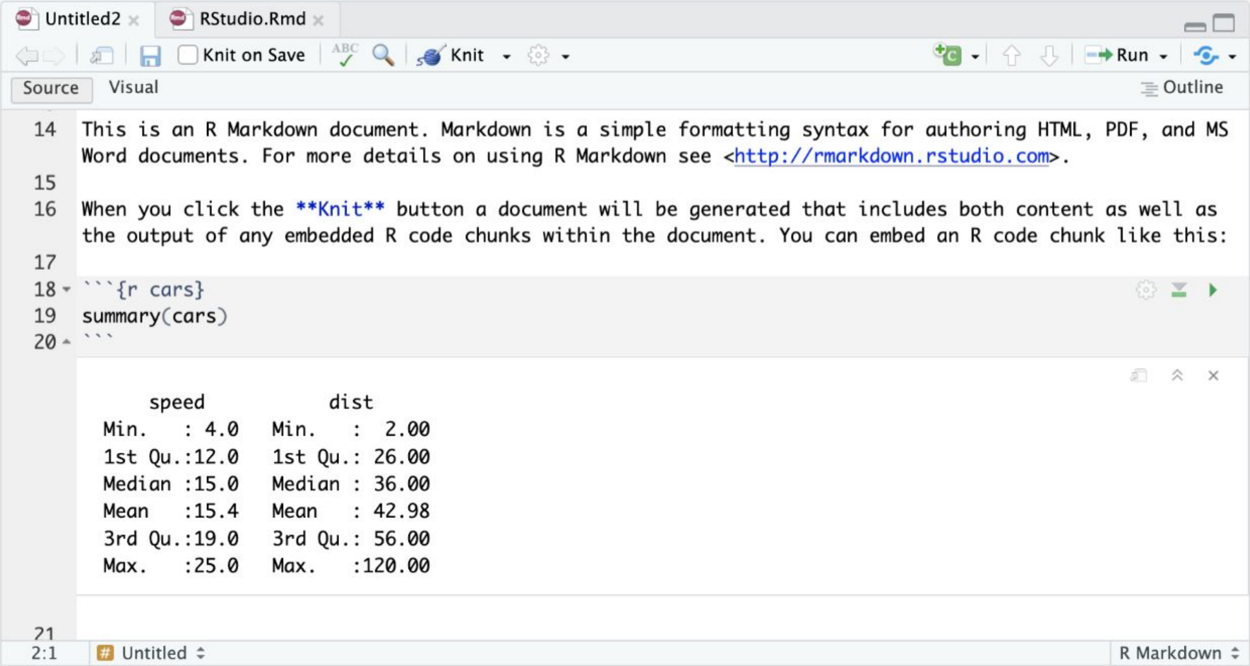


~~~
```{r cars}
summary(cars)
```

Including Plots
~~~

`Ctrl + Enter` on Windows or `Command + Enter` on Mac in your script evaluates that line of code

Running a chunk executes the code

- generally see a preview of the output of the code just below the chunk
- see the code in the console



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>.

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. You can embed an R code chunk like this:

~~~
```{r cars}
summary(cars)
```
~~~

```
     speed           dist       
 Min.   : 4.0   Min.   :  2.00  
 1st Qu.:12.0   1st Qu.: 26.00  
 Median :15.0   Median : 36.00  
 Mean   :15.4   Mean   : 42.98  
 3rd Qu.:19.0   3rd Qu.: 56.00  
 Max.   :25.0   Max.   :120.00  
```

Run previous chunks button

You can run all chunks above a specific chunk using this button:

~~~
```{r, out.width = "80%", echo = FALSE, fig.align='center'}
knitr::include_graphics("images/chunk.png")
~~~
```
```

~~~
```{r eval}
```
~~~

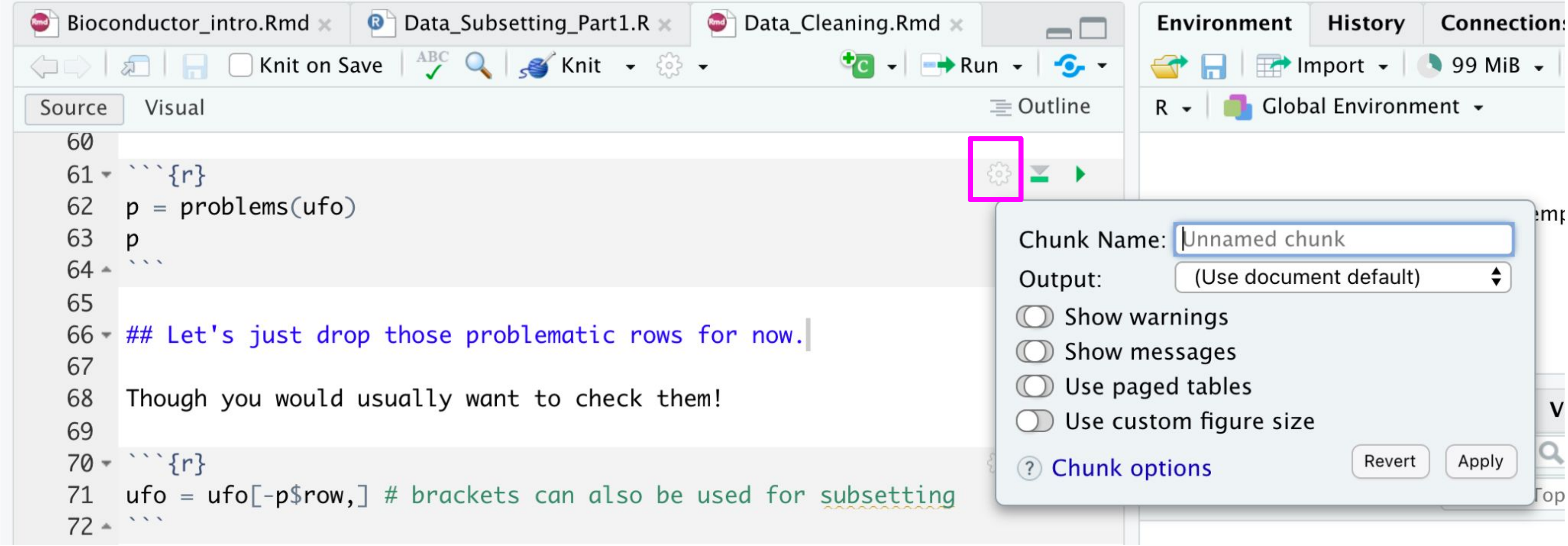
eval = FALSE will show but not run the code

~~~
```{r ec}
```
~~~

echo = FALSE will not show but run the code

Using both set to FALSE will hide the code and will not run the code!

The default is TRUE - do not need to write it out if you want code to be shown and run.



~~~
```{r}
p = problems(ufo)
p
```

## Let's just drop those problematic rows for now.

Though you would usually want to check them!

```{r}
ufo = ufo[-p$row,] # brackets can also be used for subsetting
```
~~~

Chunk Name: Unnamed chunk

Output: (Use document default)

○ Show warnings

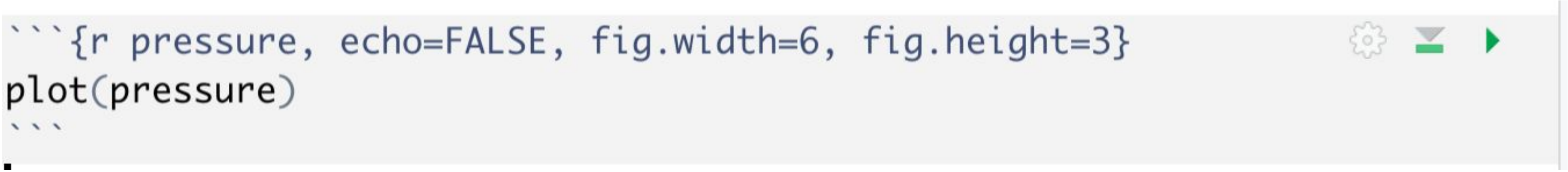
○ Show messages

○ Use paged tables

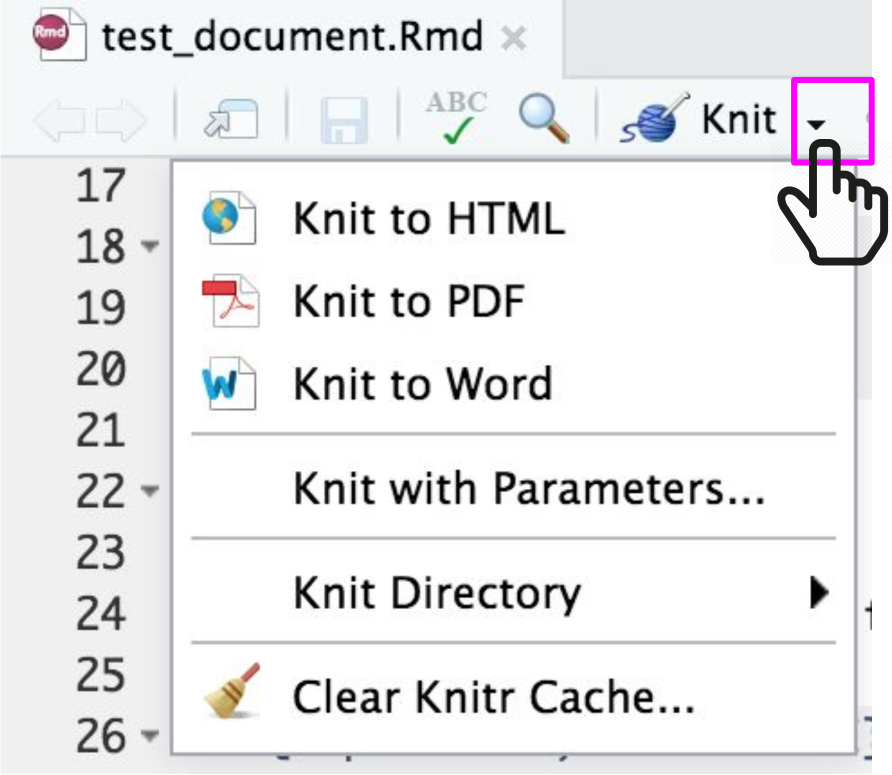
○ Use custom figure size

Can Also influence how a plot is displayed with a chunk!

fig.align = “center”,
fig.height = 3, fig.width = 6

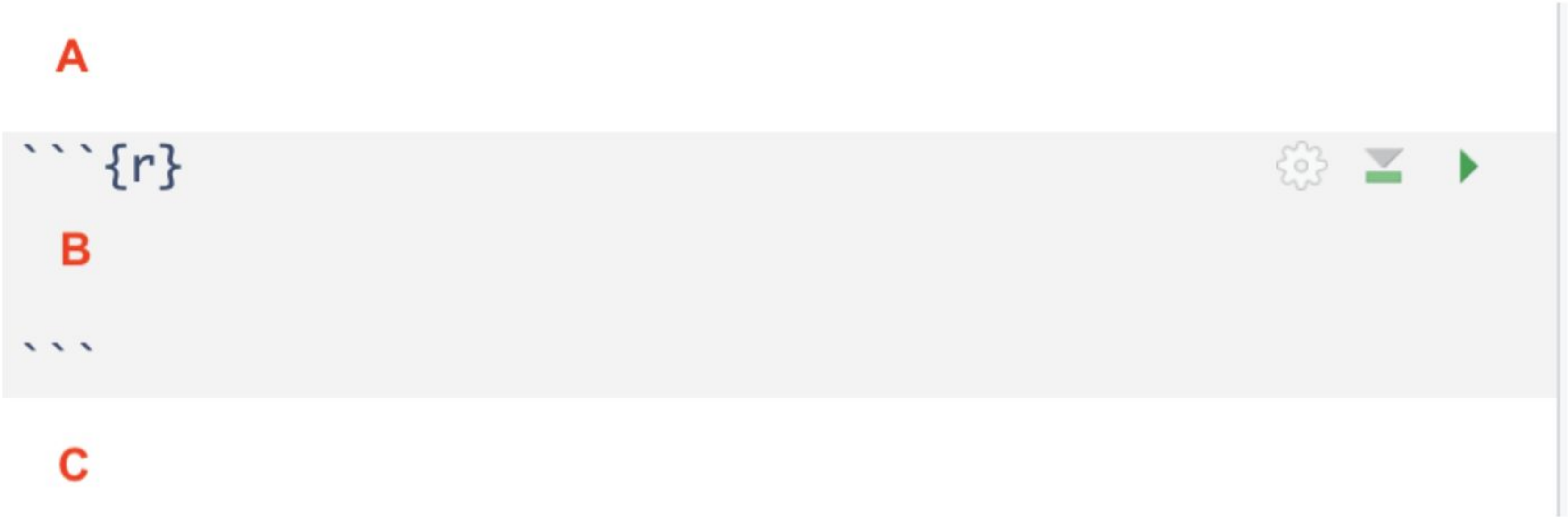


~~~
```{r pressure, echo=FALSE, fig.width=6, fig.height=3}
plot(pressure)
```
~~~


test_document.Rmd
ABC
Knit
17
18
19
20
21
22
23
24
25
26
Knit to HTML
Knit to PDF
Knit to Word
Knit with Parameters...
Knit Directory
Clear Knitr Cache...

Gut Check

Where does code go typically in an Rmd file?



A

~~~
```{r}
B

```
~~~

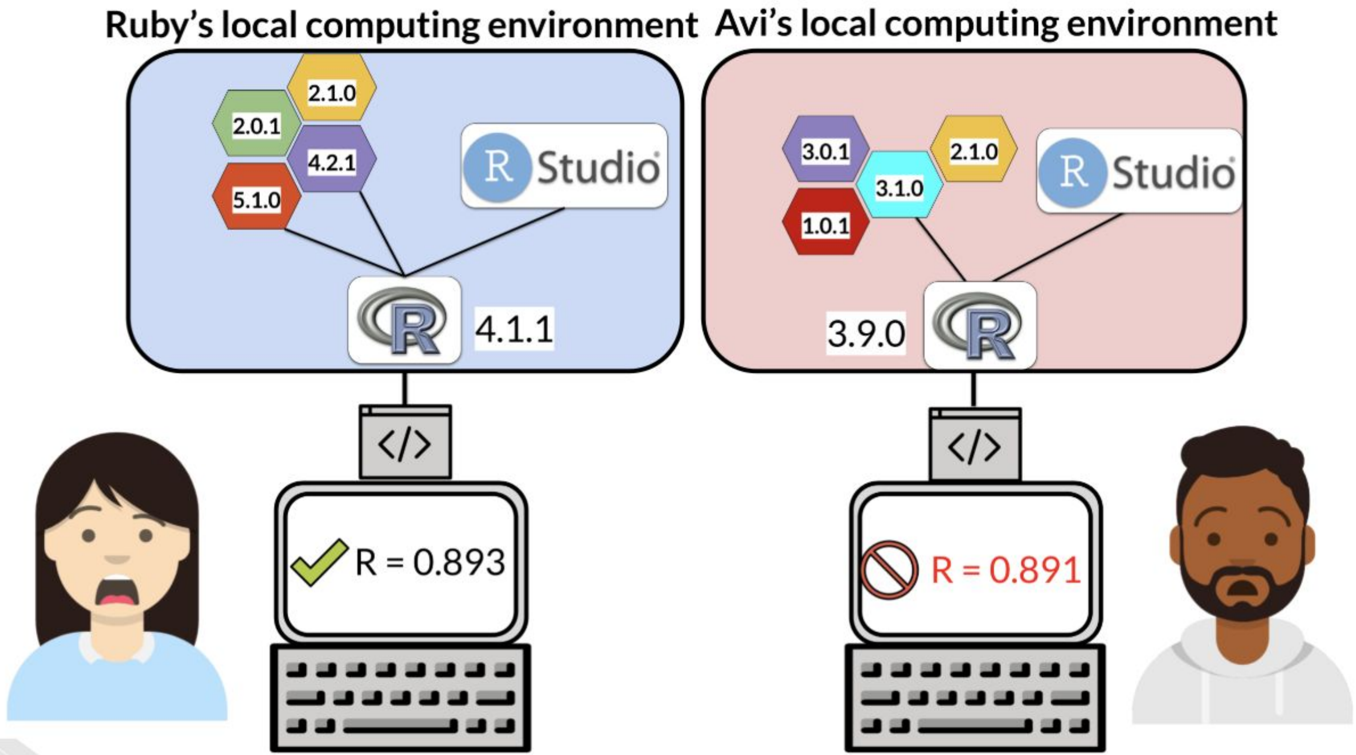
C

https://support.rstudio.com/hc/en-us/articles/200711853-Keyboard-Shortcuts

Rendering in R

```
library(rmarkdown)
render("Untitled.Rmd")
```

Versions matter



Created by Candace Savonen

Session information - what's loaded?

```
install.packages(“devtools”)
library(devtools)
session_info()
```

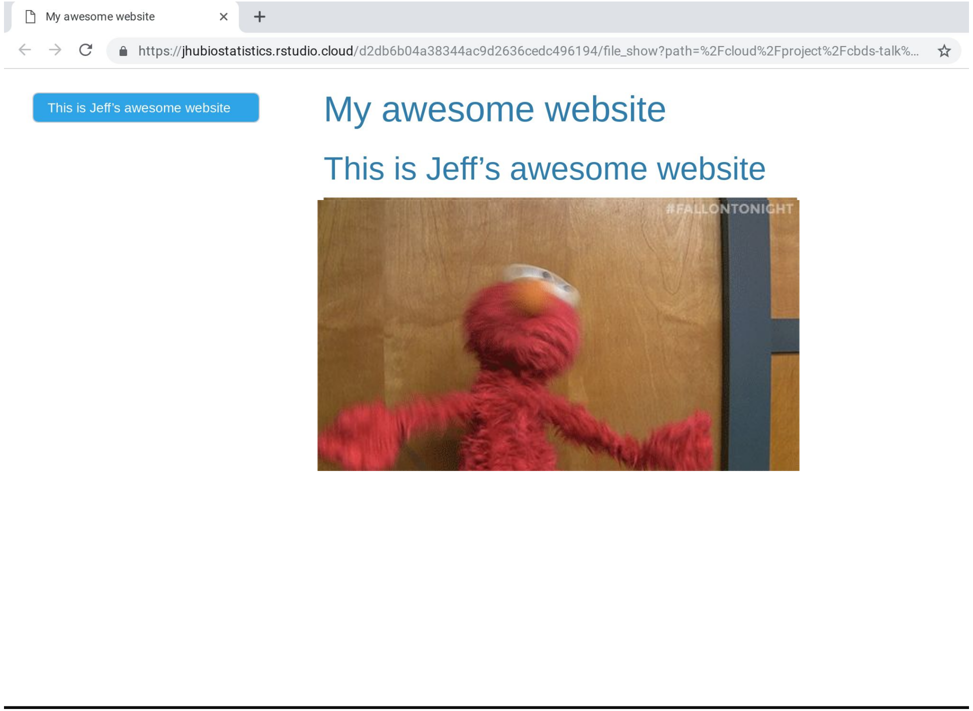
Rmarkdown lab

Download the file from https://github.com/SISBID/data-wrangling/raw/gh-pages/labs/rmarkdown-lab.Rmd

Barely scratching the surface

rmarkdown

```
---
title: "My awesome website"
output:
    html_document:
        toc: true
        toc_float: true
        theme: cerulean
---
# This is Jeff's awesome website

![](https://media.giphy.com/media/d
rXGoW1iudhKw/giphy.gif)
```

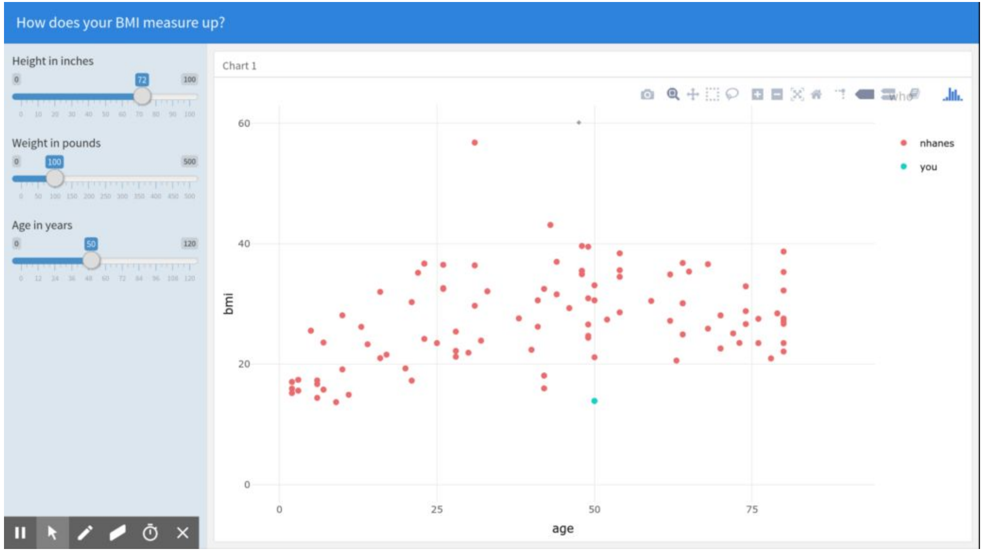


My awesome website

This is Jeff’s awesome website

flexdashboard

~~~
---
title: "How does your BMI measure up?"
output: flexdashboard::flex_dashboard
runtime: shiny
---

Inputs {.sidebar}
-------------------------------------

```{r}
library(flexdashboard); library(NHANES); library(plotly);library(dplyr)
sliderInput("height", "Height in inches",0,100,72)
sliderInput("weight", "Weight in pounds",0,500,100)
sliderInput("age", "Age in years",0,120,50)

```

Column
-------------------------------------

### Chart 1

```{r}
nhanes = sample_n(NHANES,100)
renderPlotly({
  df = data.frame(bmi = c(nhanes$BMI,input$weight*0.45/(input$height*0.025)^2),
                  age = c(nhanes$Age,input$age),
                  who = c(rep("nhanes",100),"you"))
  ggplotly(ggplot(df) +
             geom_point(aes(x=age,y=bmi,color=who)) +
             scale_x_continuous(limits=c(0,90)) +
             scale_y_continuous(limits=c(0,60)) +
             theme_minimal()
           )
})
```
~~~



How does your BMI measure up?

Height in inches: 72 (0–100)

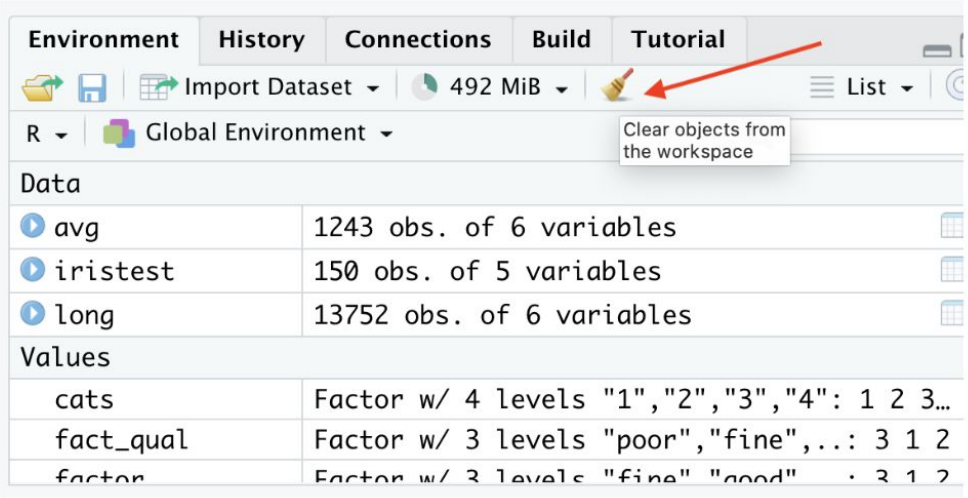
Weight in pounds: 100 (0–500)

Age in years: 50 (0–120)

Clean your environment

Regularly cleaning your environment and trying your code again, can help ensure that your code is running as expected.

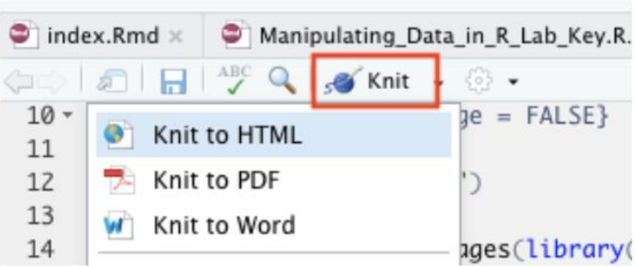
Occasionally we might forget to save a step of our code in our R Markdown file that we ran only in the console. This will help us figure that out.

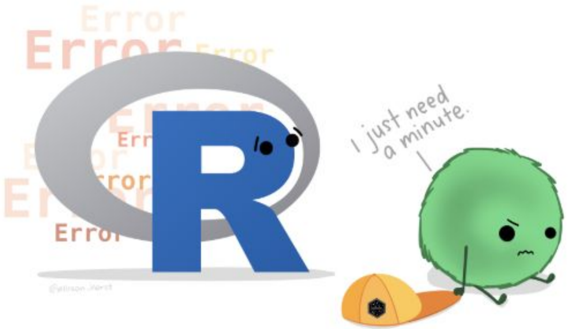


Data	
avg	1243 obs. of 6 variables
iristest	150 obs. of 5 variables
long	13752 obs. of 6 variables
Values	
cats	Factor w/ 4 levels "1","2","3","4": 1 2 3…
fact_qual	Factor w/ 3 levels "poor","fine",..: 3 1 2
factor	Factor w/ 3 levels "fine","good",..: 3 1 2

Check if your file knits regularly

Regularly checking if your file knits will help you spot a missing step or error earlier when you have less code to try to identify where your code might have gone wrong.





GUT CHECK

Why is reproducibility so important?

A. It helps to ensure that your code is working consistently and it helps others understand what you did

B. It ensures that your code is correct

GUT CHECK

What is NOT a practice to improve the reproducibility of our work?

A. Using R Markdown files to describe what your code is doing

B. Using scripts instead of R Markdown files

C. Testing your code with R Markdown files or the run previous button

D. Regularly cleaning the environment

Summary

To help make your work more reproducible:

- Use R Projects
- RMarkdown
- Clean your environment regularly
- Check the knit of your RMarkdown regularly
- Tell your future self and others what you did!
- Print session info!

More resources

These are just some quick tips, for more information:

- Reproducibility in Cancer Informatics course
- Advanced Reproducibility in Cancer Informatics course
- The RMarkdown book
- Jenny Bryan's organizational strategies
- Write efficient R code for science
- Tools for reproducible workflows in R

Link to lab: https://sisbid.github.io/Data-Wrangling/labs/rmarkdown-lab.Rmd