

Getting started with

Dr Saskia Freytag

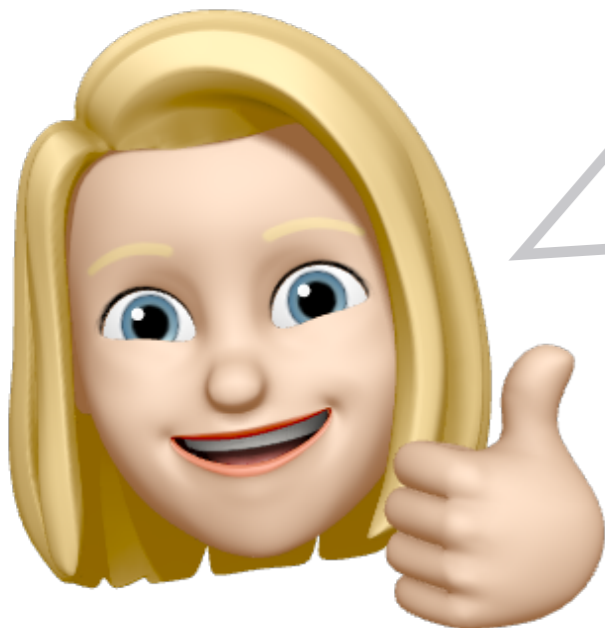
Webinar for Australian Biocommons



To me  is a:

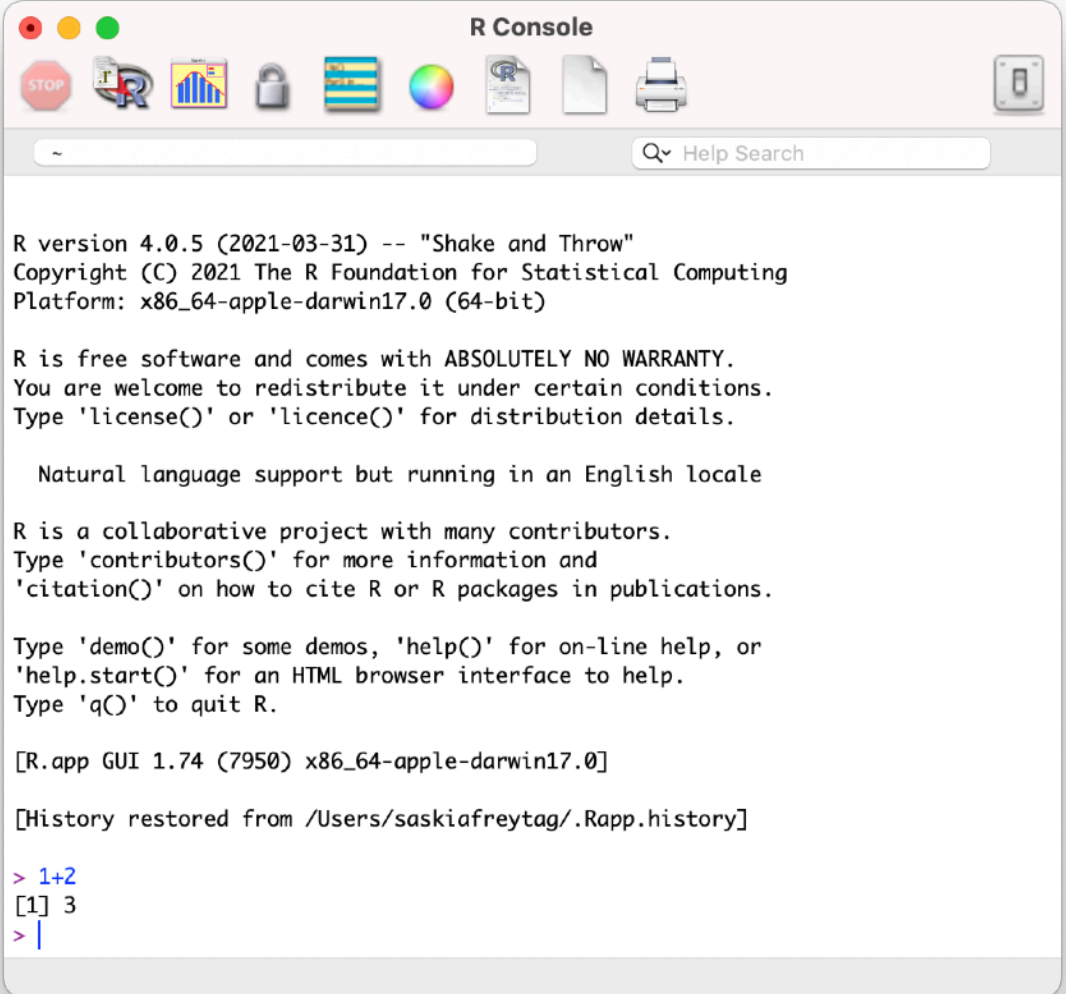
- Programming language

```
Console Terminal x Jobs x
/cloud/project/ ↗
> # if statement
> x <- -3
> if (x < 0) {
+   print("x is a negative number")
+ } else if (x == 0) {
+   print("x is zero")
+ } else {
+   print("x is a positive number")
+ }
[1] "x is a negative number"
>
> # else if statement
> x <- 0
> if (x < 0) {
+   print("x is a negative number")
+ } else if (x == 0) {
+   print("x is zero")
+ } else {
+   print("x is a positive number")
+ }
[1] "x is zero"
>
> # else statement
> x <- 5
> if (x < 0) {
+   print("x is a negative number")
+ } else if (x == 0) {
+   print("x is zero")
+ } else {
+   print("x is a positive number")
+ }
[1] "x is a positive number"
> |
```



To me  is a:

- Calculator



```
R Console
~
Q Help Search

R version 4.0.5 (2021-03-31) -- "Shake and Throw"
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

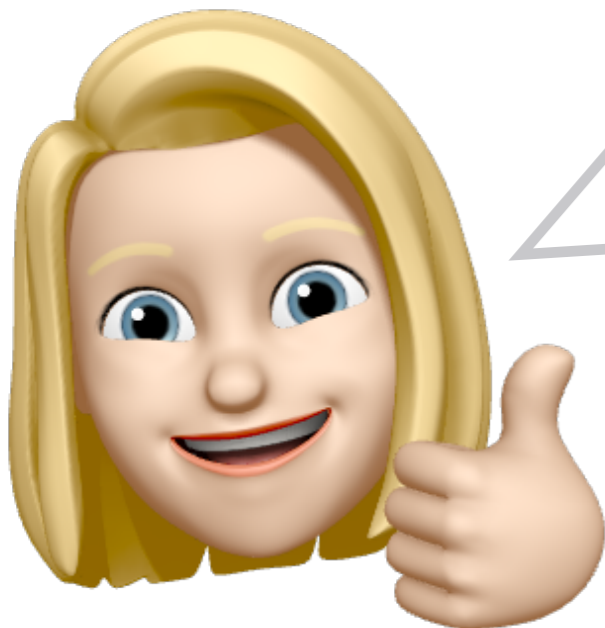
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

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 1.74 (7950) x86_64-apple-darwin17.0]

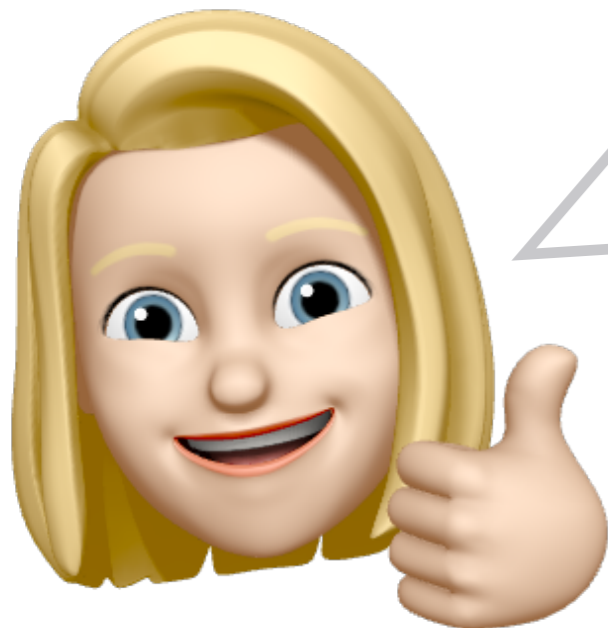
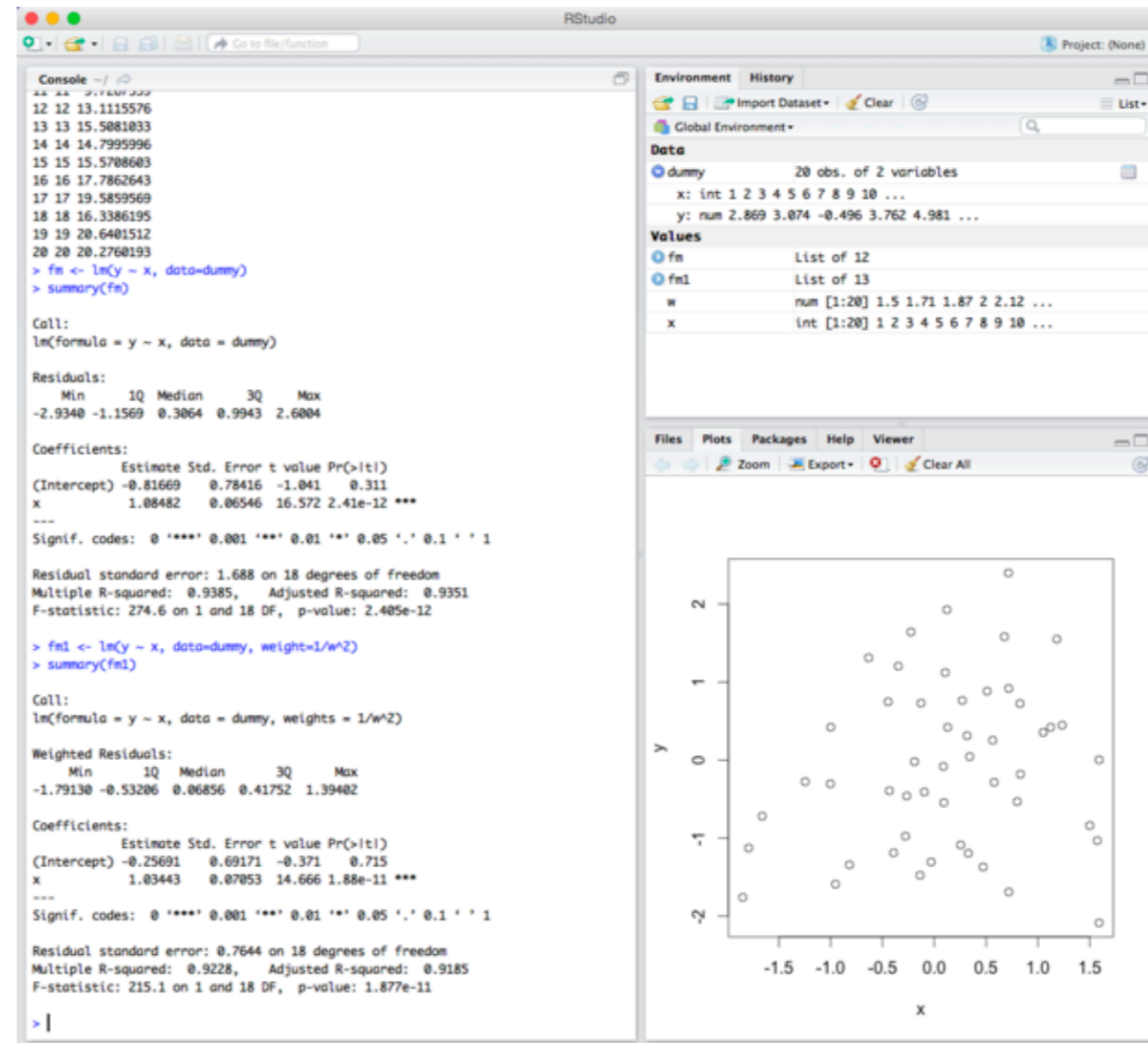
[History restored from /Users/saskiafreytag/.Rapp.history]

> 1+2
[1] 3
> |
```



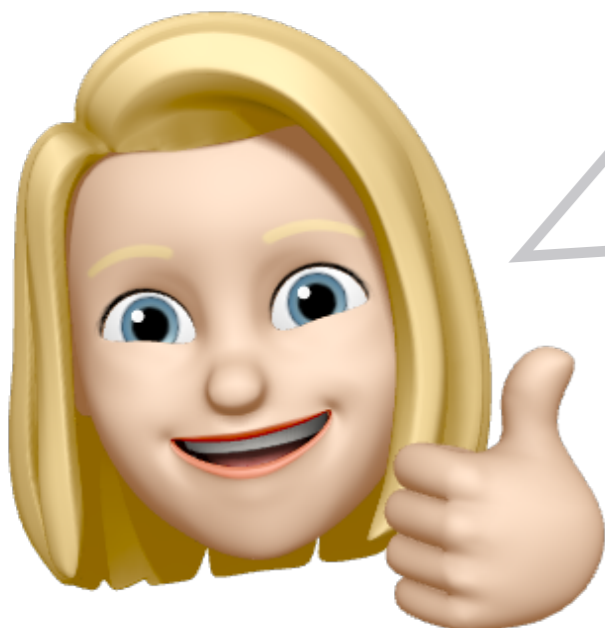
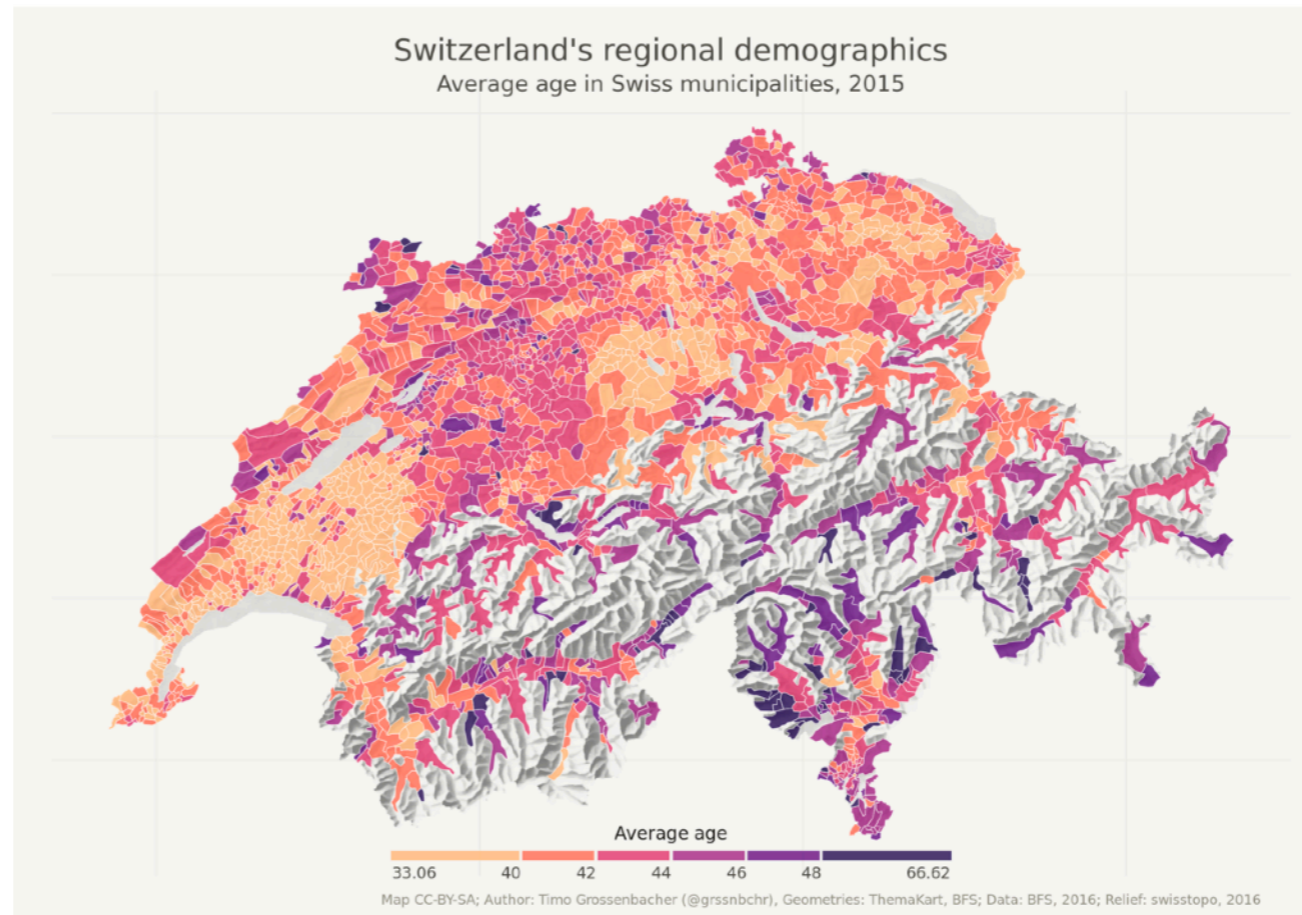
To me  is a:

- Statistical analysis tool



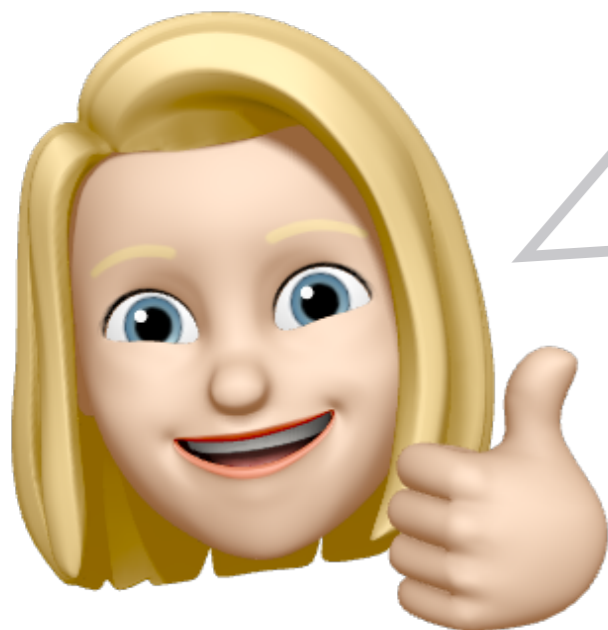
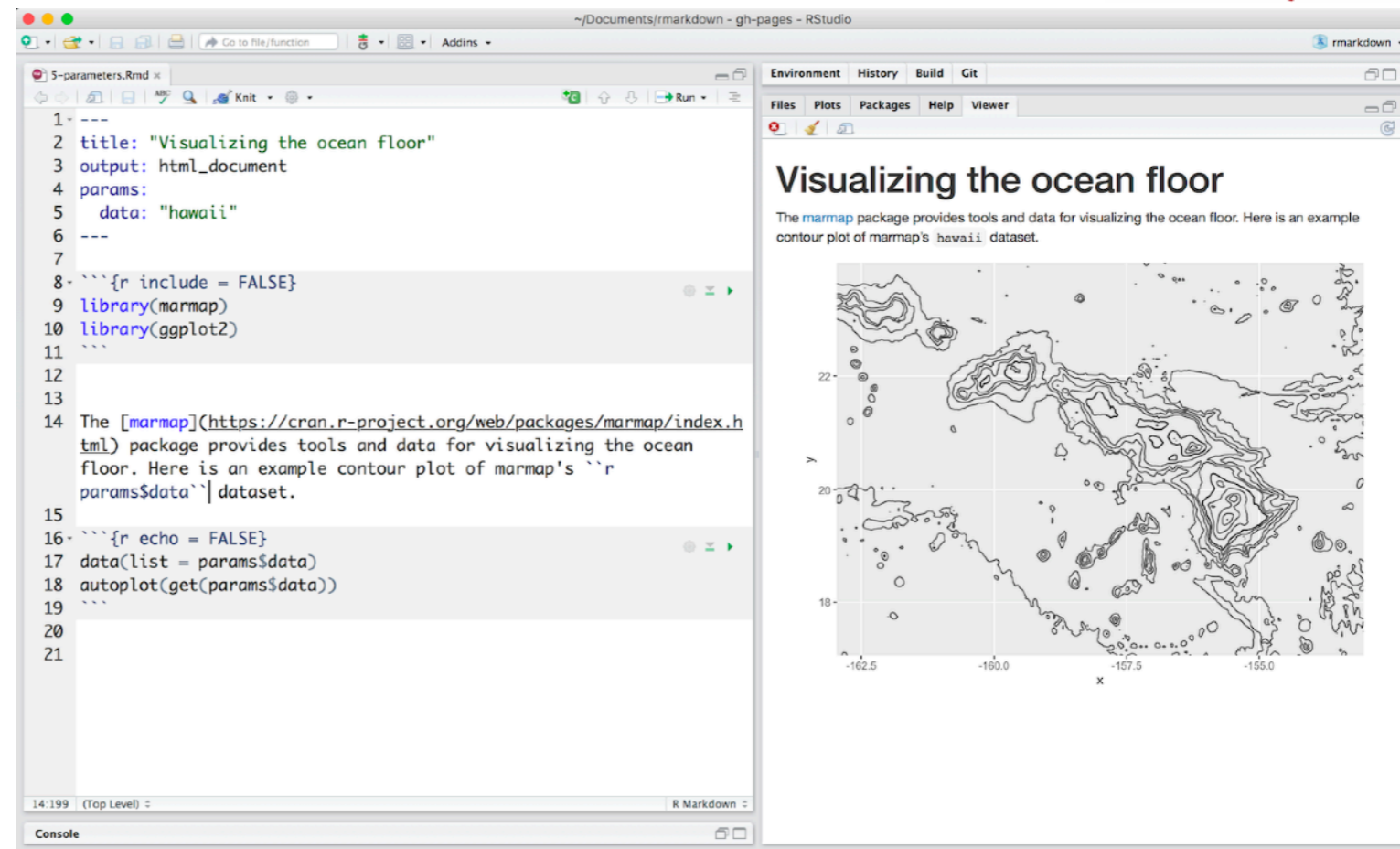
To me  is a:

- Visualisation platform



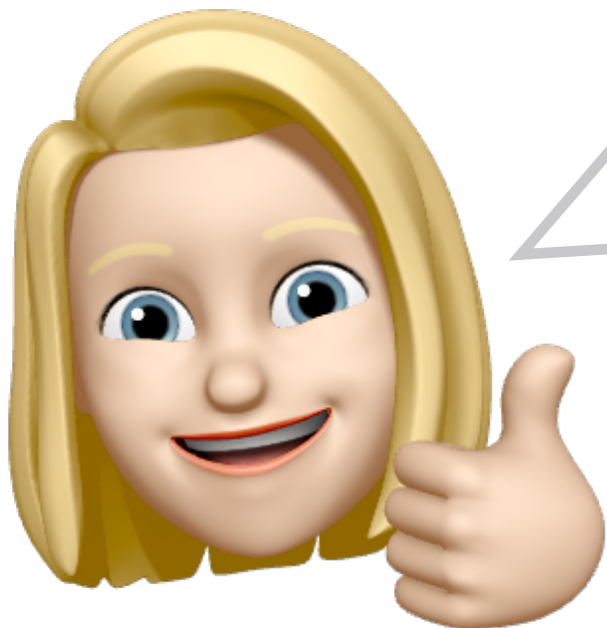
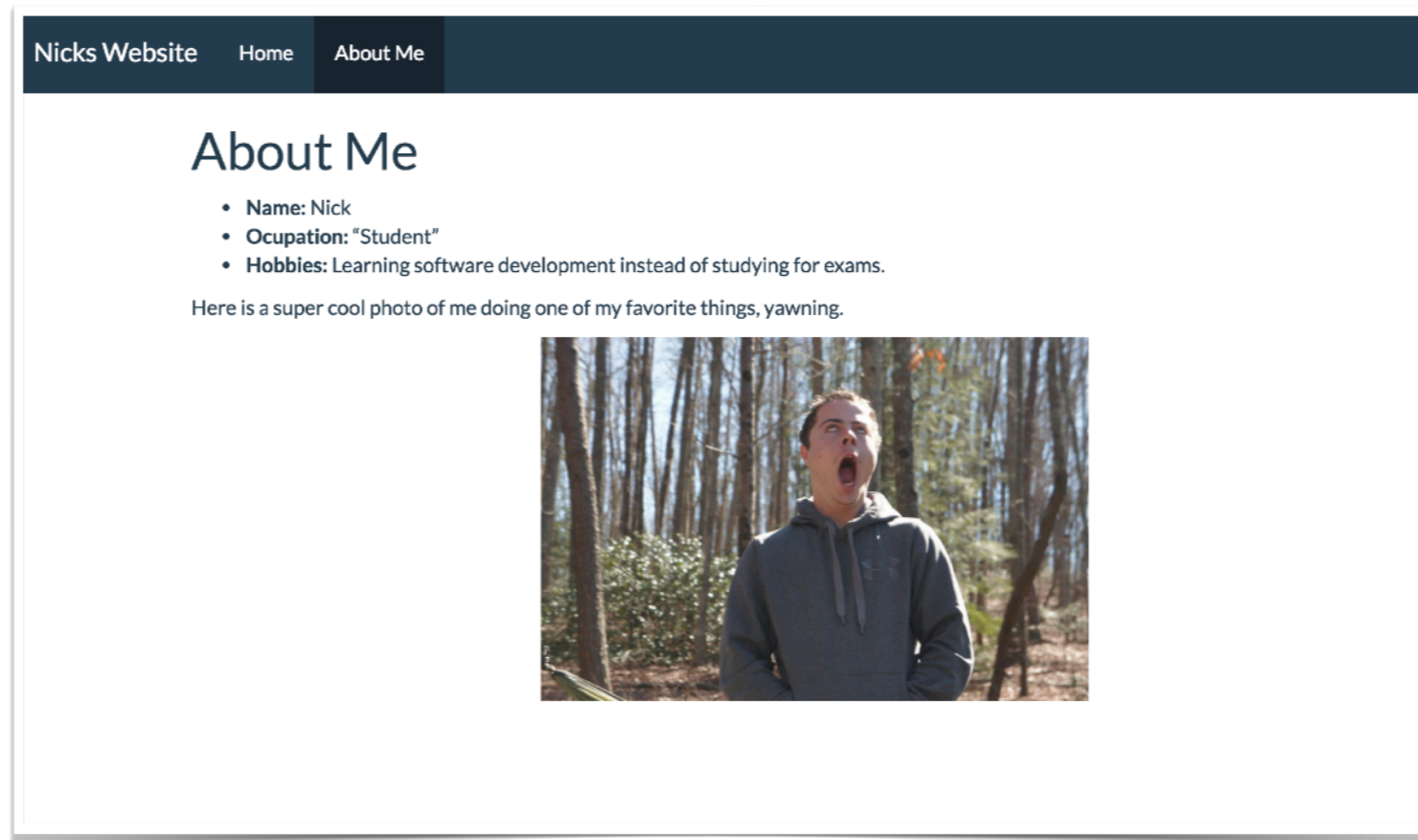
To me  is a:

- Document editor



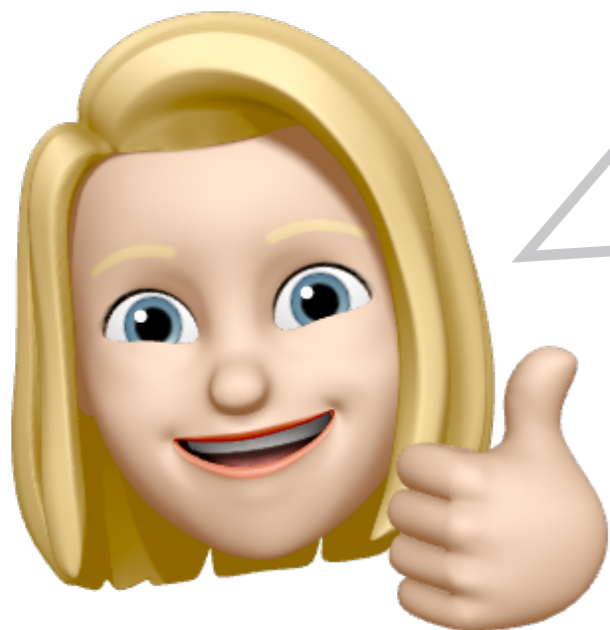
To me  is a:

- Website editor



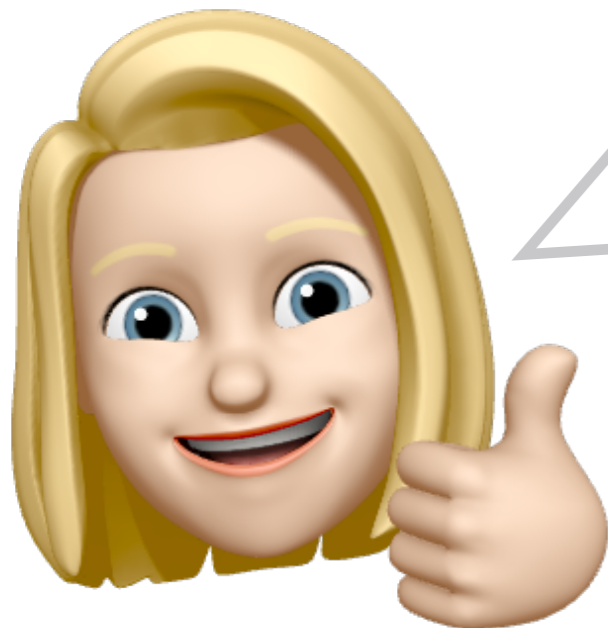
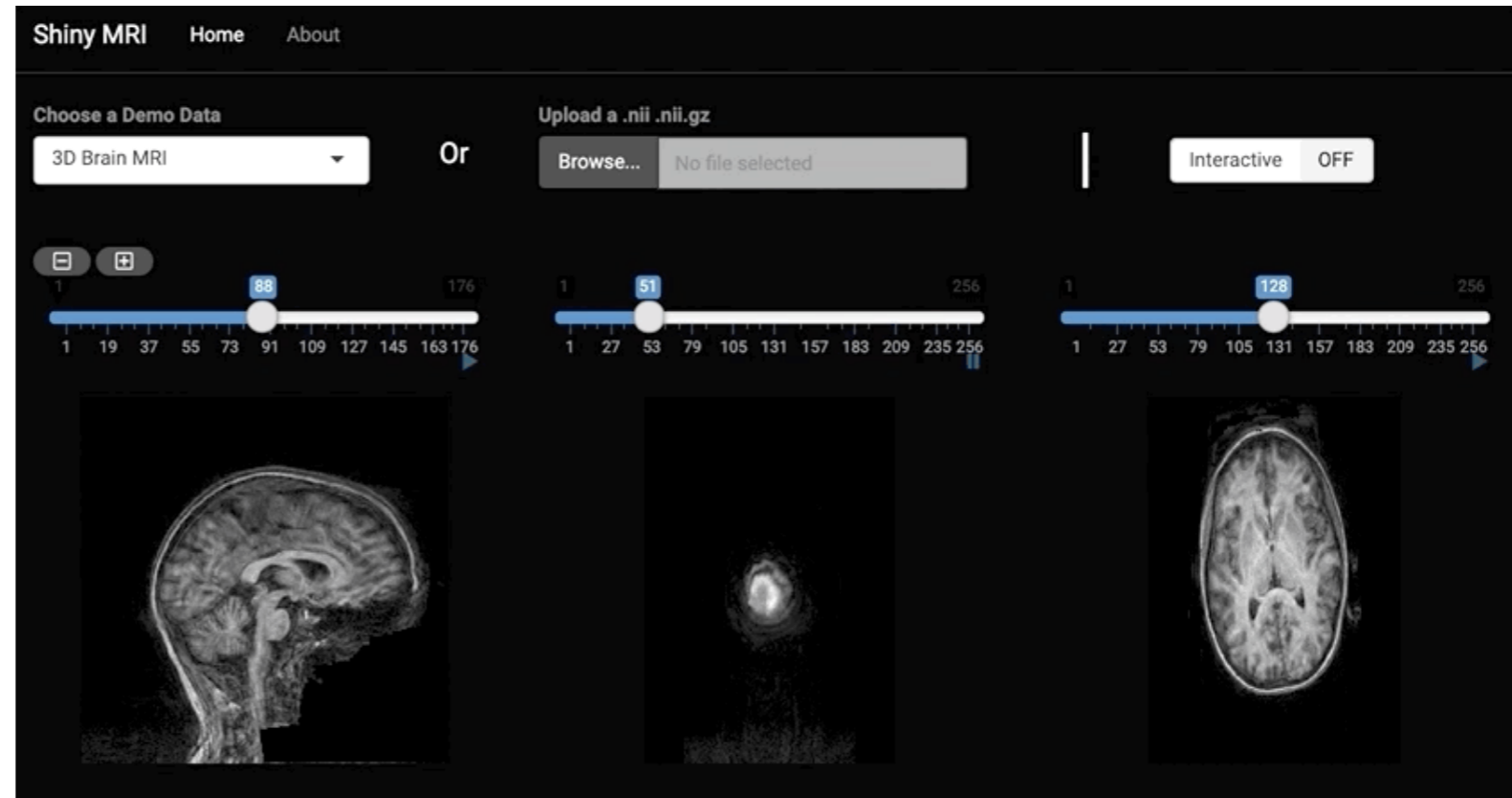
To me  is a:

- Presentation editor



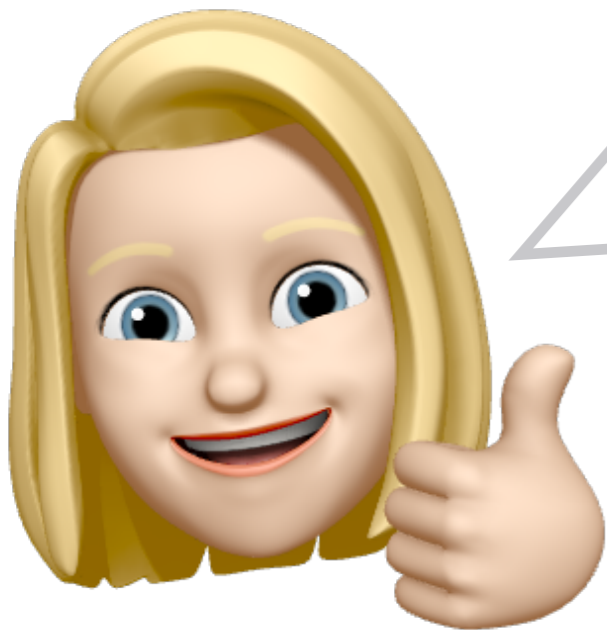
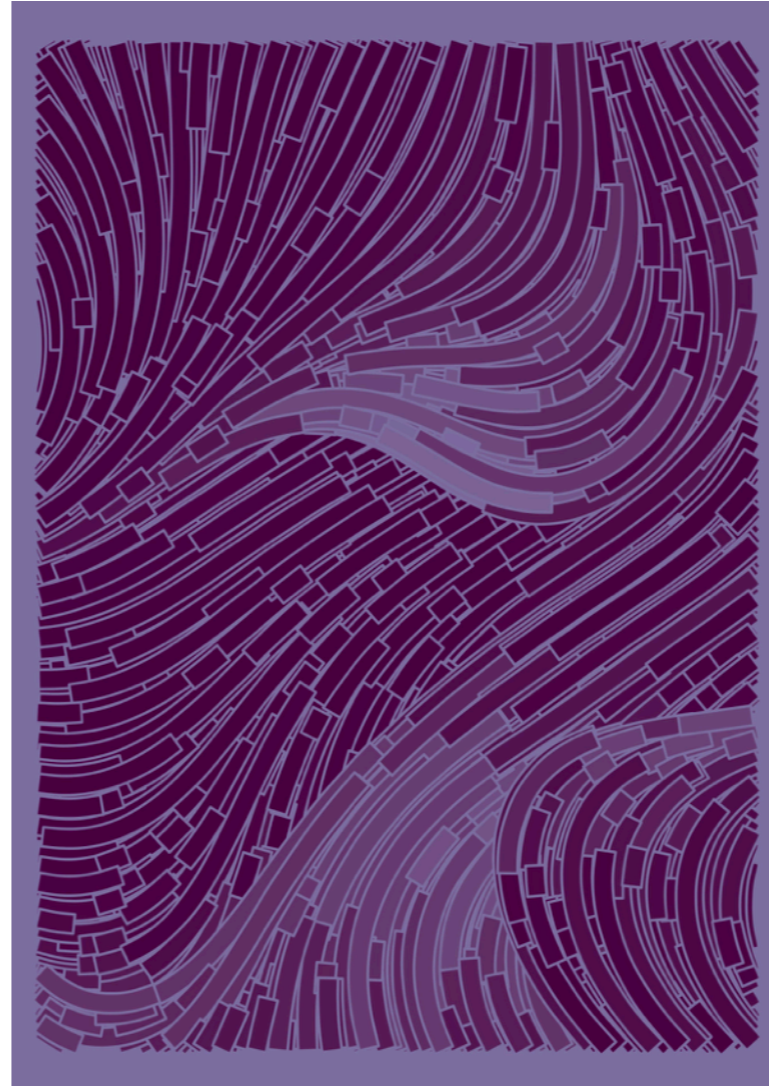
To me  is a:

- Application maker



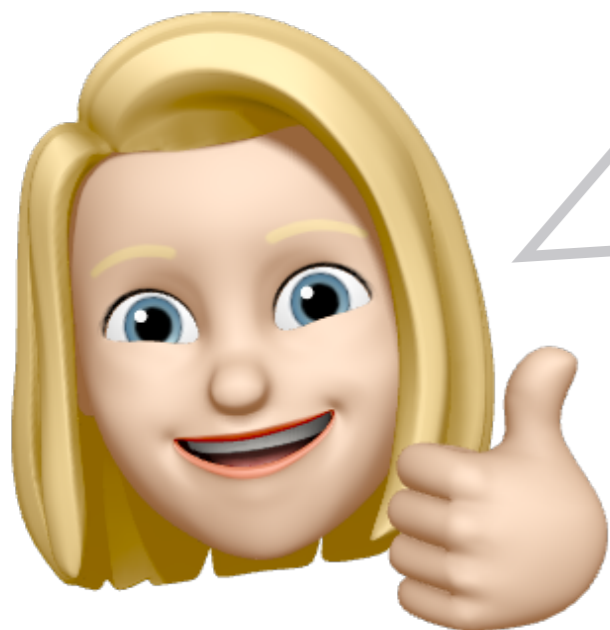
To me  is a:

- Art generator



To me  is a:

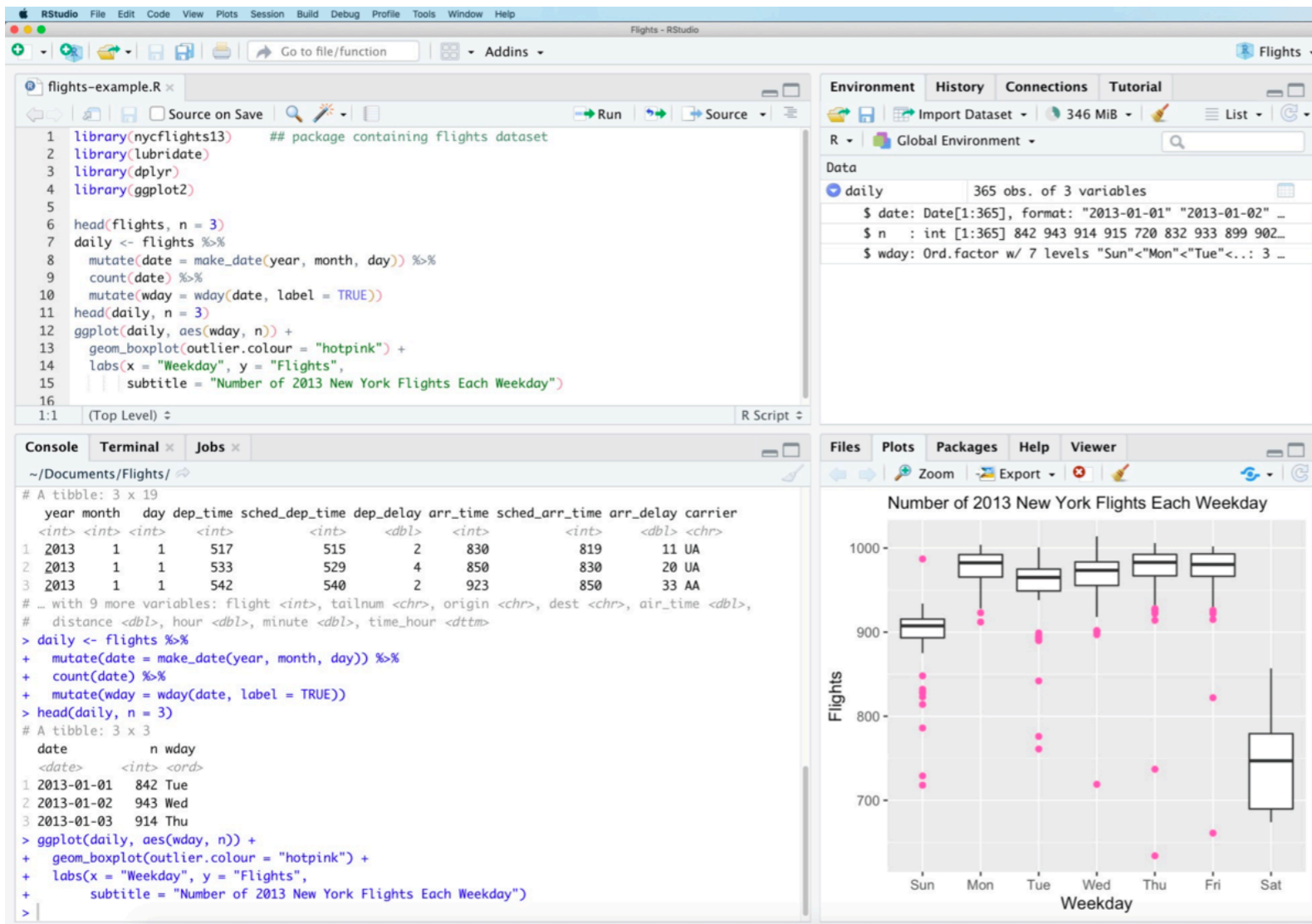
- Community



R Studio® - a natural home for and more

Source:
Recipes

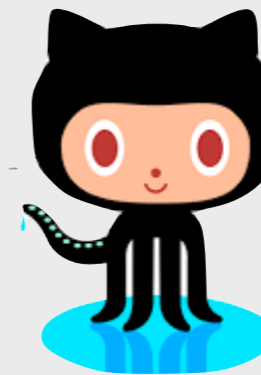
Environment/etc.:
Ingredients



The screenshot shows the R Studio interface with four main panes:

- Source:** A script named 'flights-example.R' containing R code for loading packages, creating a 'daily' dataset, and plotting a boxplot of flights by weekday.
- Environment/etc.:** Shows the 'Global Environment' with a 'daily' data object containing 365 observations of 3 variables.
- Console:** Displays the output of the R code, including a tibble of flight data and the execution of the plotting commands.
- Plotting pane:** Shows a boxplot titled 'Number of 2013 New York Flights Each Weekday' with 'Weekday' on the x-axis and 'Flights' on the y-axis.

Cool extra features



and so many extra helpful features, such as tab complete, search, code snippets, ...

Console/etc.:
Cooking stove

Plotting pane/etc.:
Plating



Extending the universe with packages

Repositories such as CRAN and Bioconductor allow you to install packages that add further functionality

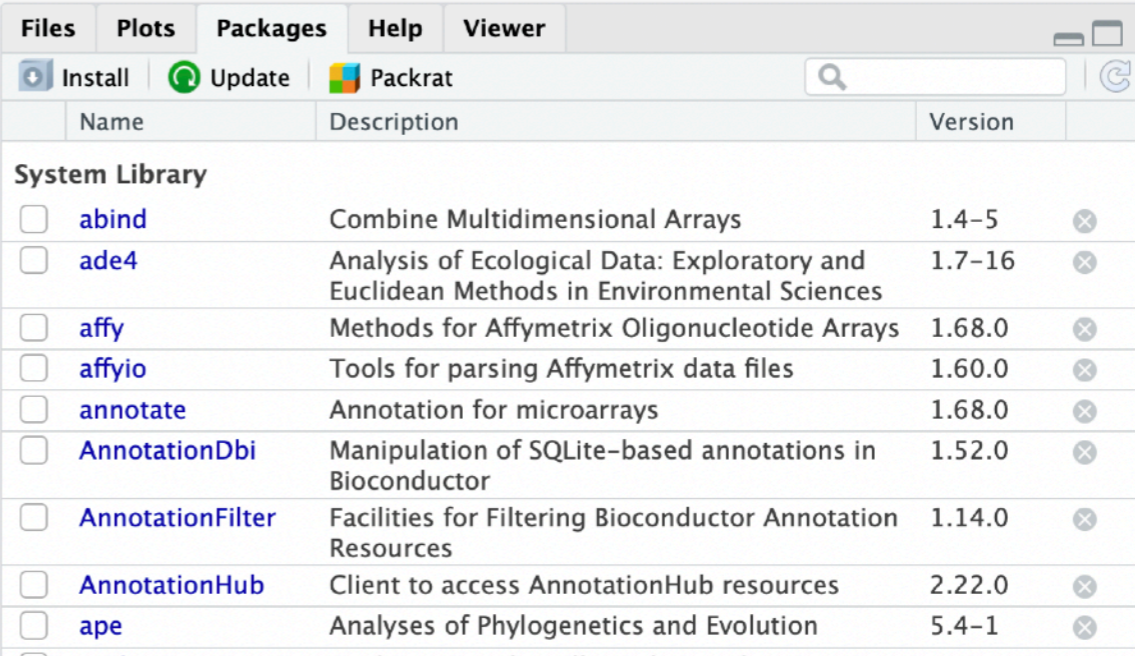
```
> install.packages("ggplot2")
trying URL 'https://cran.rstudio.com/bin/macosx/contrib/4.0/ggplot2_3.3.5.tgz'
Content type 'application/x-gzip' length 4125428 bytes (3.9 MB)
=====
downloaded 3.9 MB
```

```
The downloaded binary packages are in
  /var/folders/h9/0dch4nrd7w50dkjtx5w2jfwh0000gn/T//RtmpzBNysJ/downloaded_packages
> library(ggplot2)
> |
```

For example the reticulate package allows you to interface with python



For Bioconductor use `BiocManager::install`



Name	Description	Version
System Library		
<input type="checkbox"/> abind	Combine Multidimensional Arrays	1.4-5
<input type="checkbox"/> ade4	Analysis of Ecological Data: Exploratory and Euclidean Methods in Environmental Sciences	1.7-16
<input type="checkbox"/> affy	Methods for Affymetrix Oligonucleotide Arrays	1.68.0
<input type="checkbox"/> affyio	Tools for parsing Affymetrix data files	1.60.0
<input type="checkbox"/> annotate	Annotation for microarrays	1.68.0
<input type="checkbox"/> AnnotationDbi	Manipulation of SQLite-based annotations in Bioconductor	1.52.0
<input type="checkbox"/> AnnotationFilter	Facilities for Filtering Bioconductor Annotation Resources	1.14.0
<input type="checkbox"/> AnnotationHub	Client to access AnnotationHub resources	2.22.0
<input type="checkbox"/> ape	Analyses of Phylogenetics and Evolution	5.4-1
<input type="checkbox"/> ArchR	Analyzing single-cell regulatory chromatin in	1.0.1

Tackling your project with

Common steps in your R data analysis project:



1

Organise project



2

Read in data



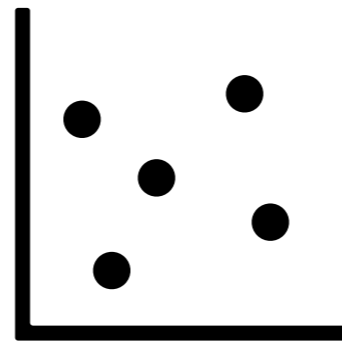
3

Data wrangling



4

Analysis



5

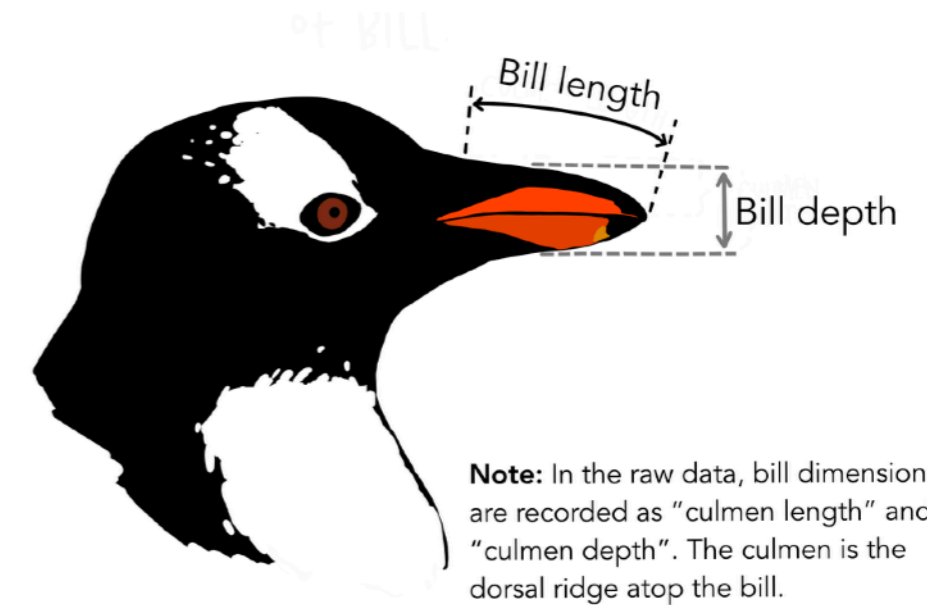
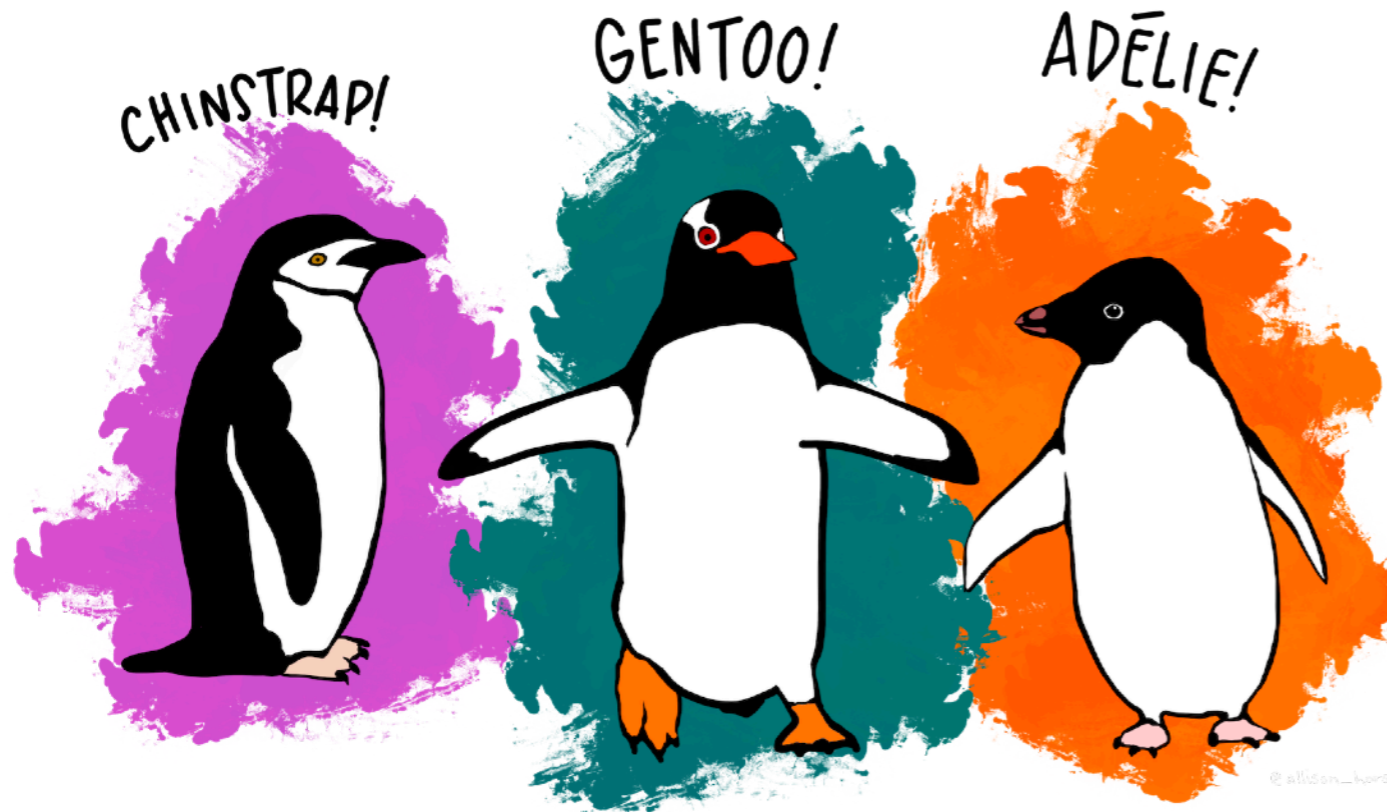
Visualisation



6

Report

Taking you through a project example



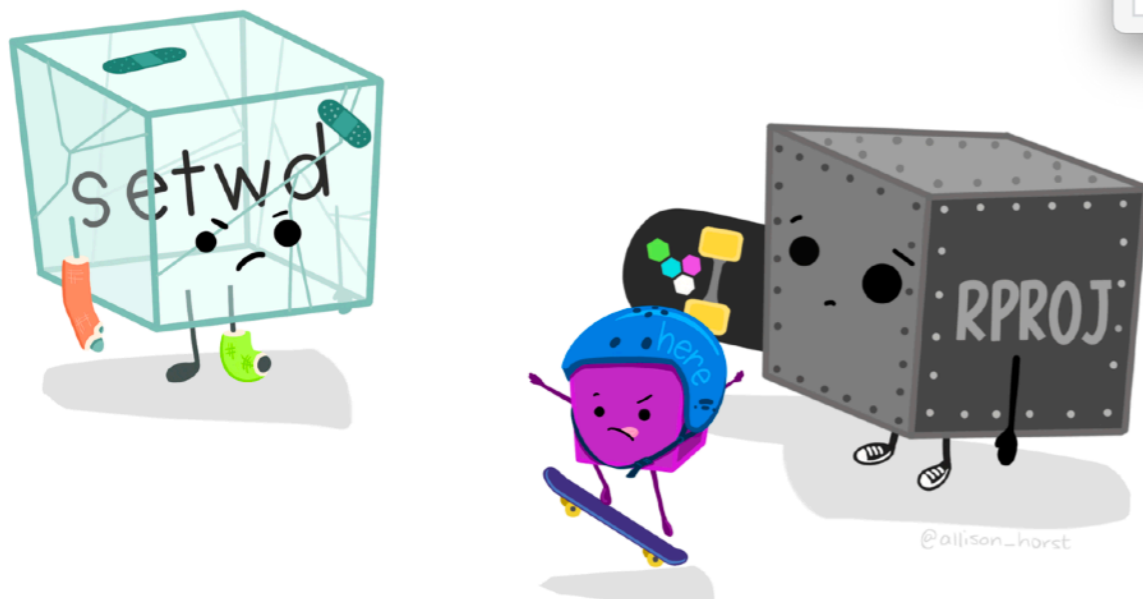
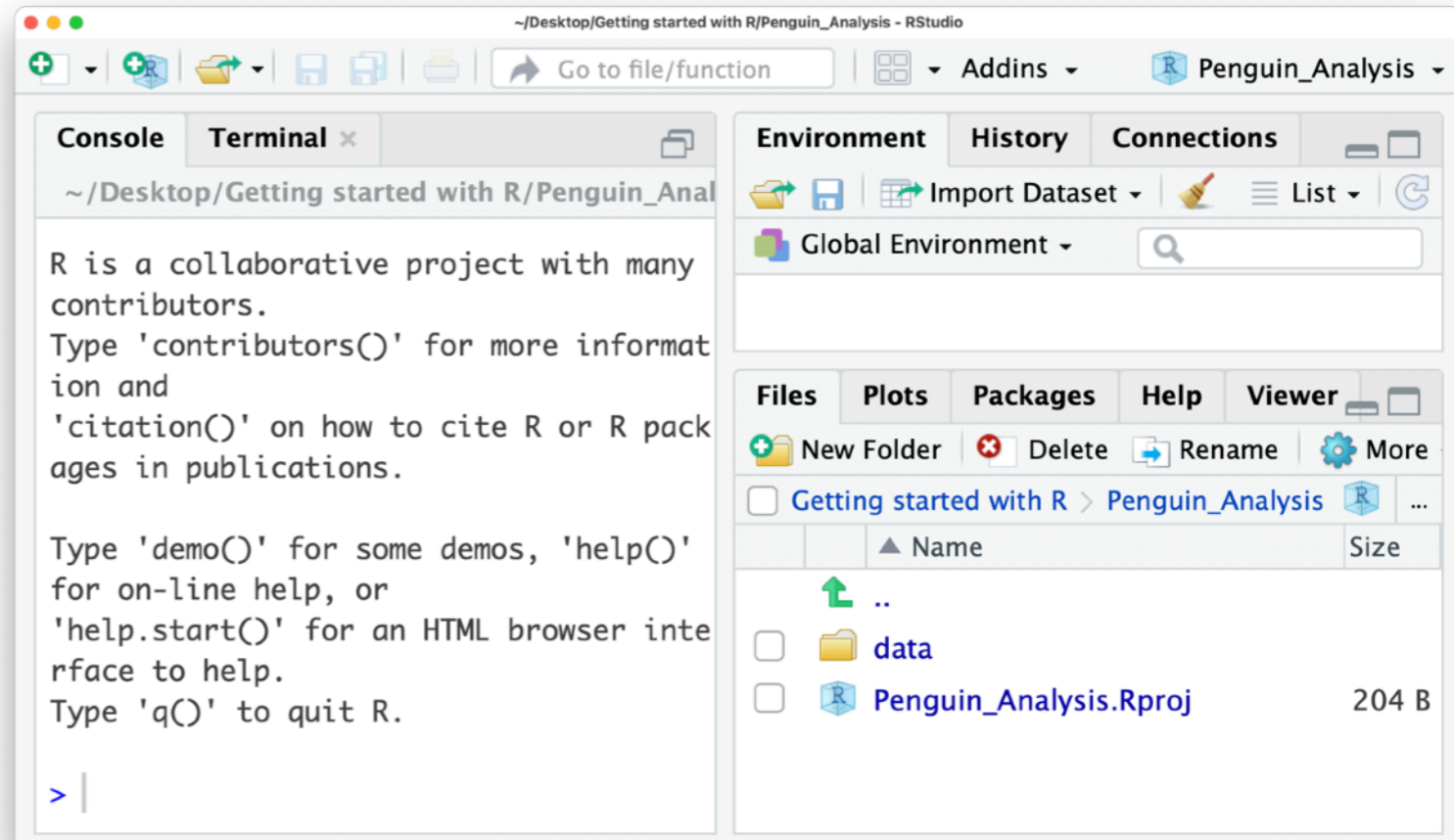
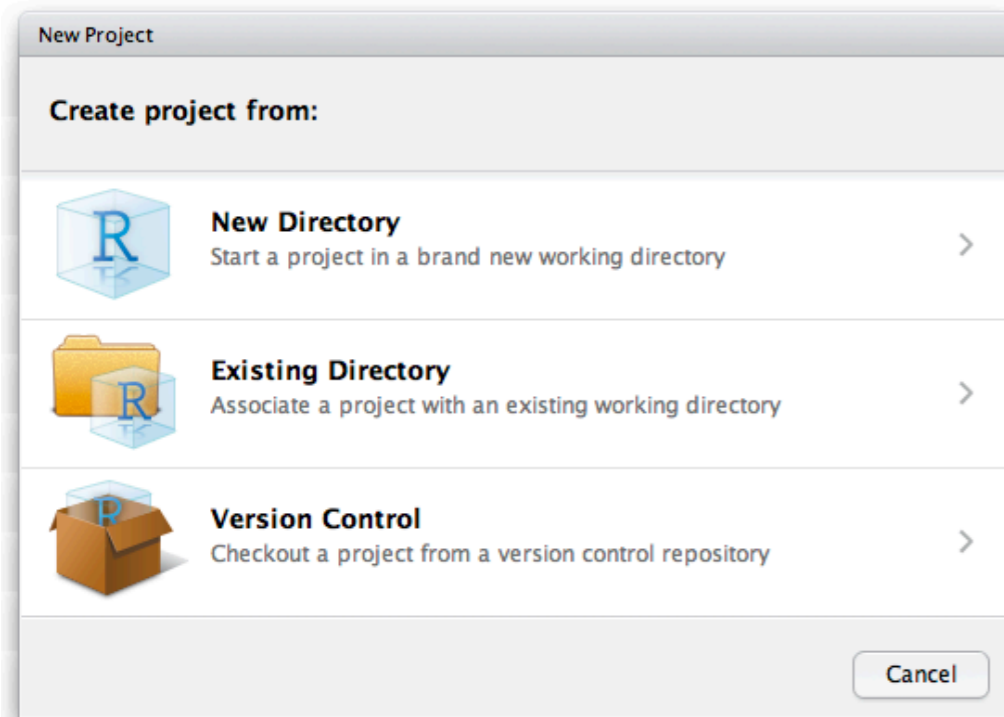
PalmerPenguins

Possible Data Loss Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file format.

	A	B	C	D	E	F	G	H	I	J
1		species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year	
2	1	Adelie	Torgersen	39.1	18.7	181	3750	male	2007	
3	2	Adelie	Torgersen	39.5	17.4	186	3800	female	2007	
4	3	Adelie	Torgersen	40.3	18	195	3250	female	2007	
5	4	Adelie	Torgersen	NA	NA	NA	NA	NA	2007	
6	5	Adelie	Torgersen	36.7	19.3	193	3450	female	2007	
7	6	Adelie	Torgersen	39.3	20.6	190	3650	male	2007	
8	7	Adelie	Torgersen	38.9	17.8	181	3625	female	2007	
9	8	Adelie	Torgersen	39.2	19.6	195	4675	male	2007	
10	9	Adelie	Torgersen	34.1	18.1	193	3475	NA	2007	
11	10	Adelie	Torgersen	42	20.2	190	4250	NA	2007	
12	11	Adelie	Torgersen	37.8	17.1	186	3300	NA	2007	
13	12	Adelie	Torgersen	37.8	17.3	180	3700	NA	2007	
14	13	Adelie	Torgersen	41.1	17.6	182	3200	female	2007	
15	14	Adelie	Torgersen	38.6	21.2	191	3800	male	2007	
16	15	Adelie	Torgersen	34.6	21.1	198	4400	male	2007	



Organise your project



Read in your data

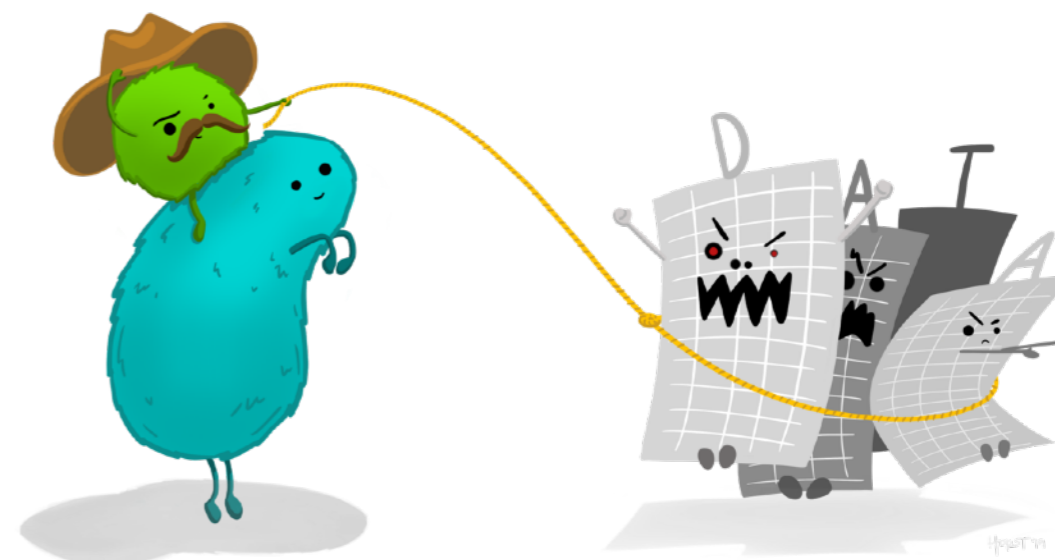


```
> data <- read.csv("data/PalmerPenguins.csv")
> head(data)
  X species      island bill_length_mm bill_depth_mm
1 1  Adeline Torgersen      39.1          18.7
2 2  Adeline Torgersen      39.5          17.4
3 3  Adeline Torgersen      40.3          18.0
4 4  Adeline Torgersen      NA            NA
5 5  Adeline Torgersen      36.7          19.3
6 6  Adeline Torgersen      39.3          20.6
  flipper_length_mm body_mass_g      sex year
1                181        3750  male 2007
2                186        3800 female 2007
3                195        3250 female 2007
4                 NA          NA  <NA> 2007
5                193        3450 female 2007
6                190        3650  male 2007
```

NA - not available, i.e. missing data



... and many more specialised packages



Wrangle your data into shape



```
> data <- data %>% filter(!is.na(bill_length_mm))
> head(data)
  X species      island bill_length_mm bill_depth_mm flipper_length_mm
1 1  Adelie Torgersen      39.1           18.7             181
2 2  Adelie Torgersen      39.5           17.4             186
3 3  Adelie Torgersen      40.3           18.0             195
4 5  Adelie Torgersen      36.7           19.3             193
5 6  Adelie Torgersen      39.3           20.6             190
6 7  Adelie Torgersen      38.9           17.8             181
  body_mass_g      sex year
1          3750   male 2007
2          3800 female 2007
3          3250 female 2007
4          3450 female 2007
5          3650   male 2007
6          3625 female 2007
```



... and many more specialised packages

dplyr : go wrangling



Analysis time



Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data.



Search:

[Home](#) [Install](#) [Help](#) [Developers](#) [About](#)

[Home](#) » [BiocViews](#)

All Packages

Bioconductor version 3.13 (Release)

Autocomplete biocViews search:

- ▼ Software (2041)
 - ▶ AssayDomain (819)
 - ▶ BiologicalQuestion (866)
 - ▶ Infrastructure (480)
 - ▶ ResearchField (953)
 - ▶ StatisticalMethod (762)
 - ▶ Technology (1301)
 - ▶ WorkflowStep (1121)
- ▶ AnnotationData (976)
- ▶ ExperimentData (406)
- ▶ Workflow (29)

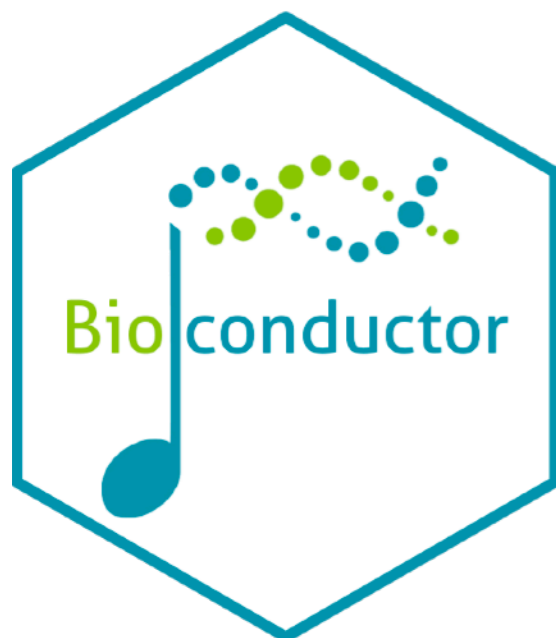
Packages found under Software:

Rank based on number of downloads: lower numbers are more frequently downloaded.

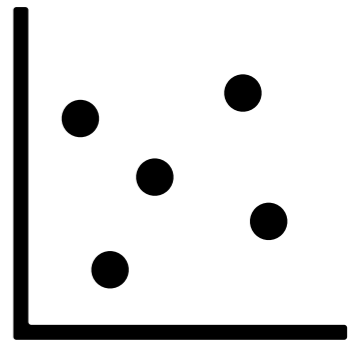
Show entries

Search table:

Package	Maintainer	Title	Rank
BiocGenerics	Bioconductor Package Maintainer	S4 generic functions used in Bioconductor	1
BiocVersion	Bioconductor Package Maintainer	Set the appropriate version of Bioconductor packages	2
S4Vectors	Bioconductor Package Maintainer	Foundation of vector-like and list-like containers in Bioconductor	3
IRanges	Bioconductor Package Maintainer	Foundation of integer range manipulation in Bioconductor	4
Biobase	Bioconductor Package Maintainer	Biobase: Base functions for Bioconductor	5
zlibbioc	Bioconductor Package Maintainer	An R packaged zlib-1.2.5	6
GenomeInfoDb	Bioconductor Package Maintainer	Utilities for manipulating chromosome names, including modifying them to follow a particular naming style	7

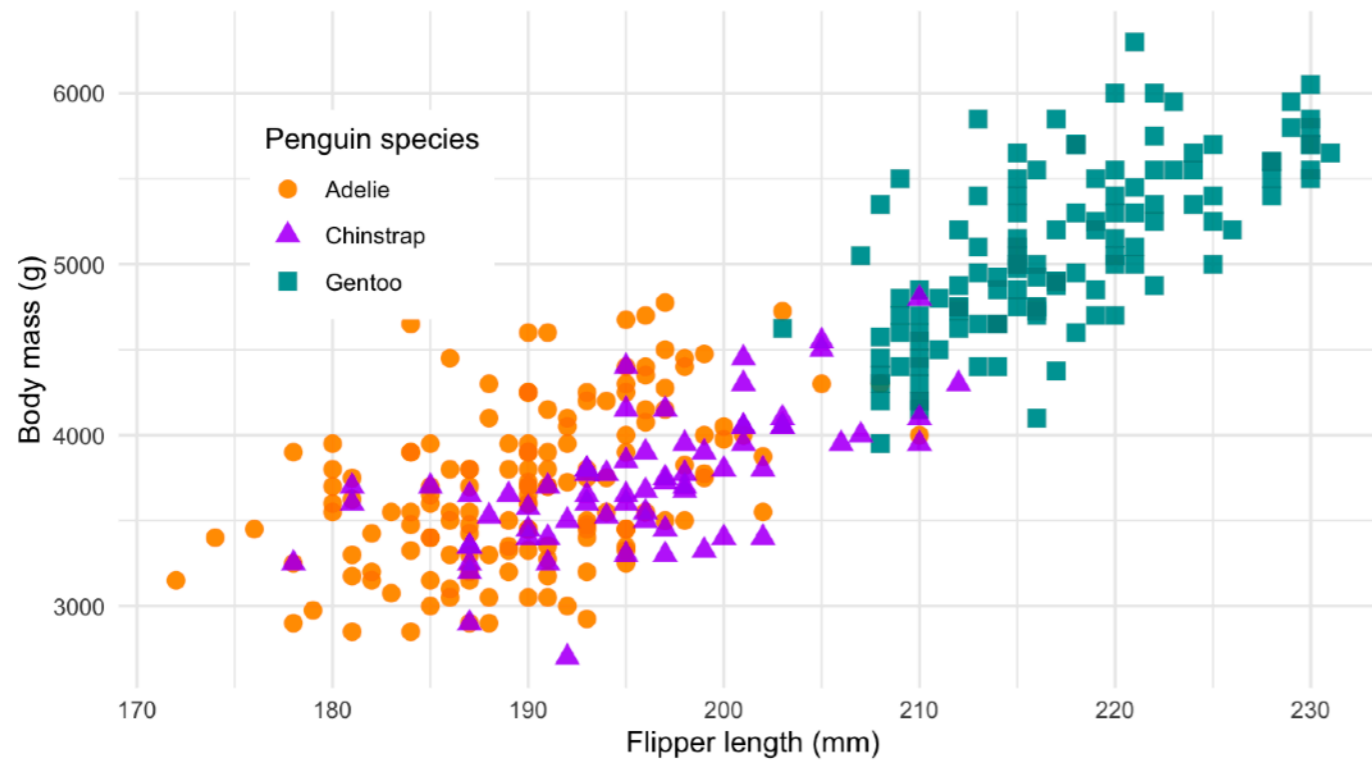


Visualise your results



Penguin size, Palmer Station LTER

Flipper length and body mass for Adelie, Chinstrap and Gentoo Penguins



```
ggplot(data, aes(x, y, color)) +  
geom_something() + theme()
```



Summarise it in a report



Penguins

Saskia Freytag
01/08/2021

Read in data

```
data <- read.csv("data/PalmerPenguins.csv")
```

Exploring factors

The penguins data has three factor variables:

```
data %>%  
  dplyr::select(where(is.factor)) %>%  
  glimpse()
```

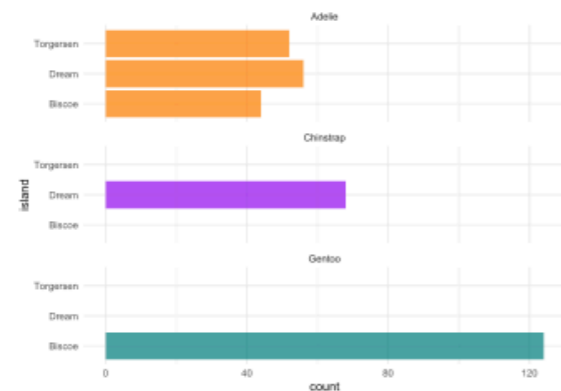
```
## Rows: 344  
## Columns: 0
```

```
# Count penguins for each species / island  
data %>%  
  count(species, island, .drop = FALSE)
```

```
##   species  island  n  
## 1  Adelle  Biscoe  44  
## 2  Adelle  Dream   56  
## 3  Adelle  Torgesen 52  
## 4 Chinstrap Dream   68  
## 5  Gentoo  Biscoe 124
```

```
ggplot(data, aes(x = island, fill = species)) +  
  geom_bar(alpha = 0.8) +  
  scale_fill_manual(values = c("darkorange", "purple", "cyan4"),  
                    guide = FALSE) +  
  theme_minimal() +  
  facet_wrap(~species, ncol = 1) +  
  coord_flip()
```

```
## Warning: It is deprecated to specify `guide = FALSE` to remove a guide. Please  
## use `guide = "none"` instead.
```

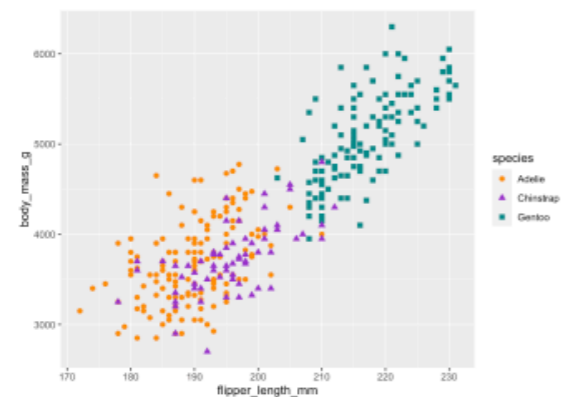


Exploring scatterplots

The penguins data also has four continuous variables, making six unique scatterplots possible!

```
ggplot(data = data, aes(x = flipper_length_mm, y = body_mass_g)) +  
  geom_point(aes(color = species,  
                 shape = species),  
            size = 2) +  
  scale_color_manual(values = c("darkorange", "darkorchid", "cyan4"))
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```



Debugging in



1.
I got this.



2.
Huh. Really
thought that
was it.



3.
(...)



4.
Fine. Restarting.



5.
OH WTF.



6.
Zombie
meltdown



7.



8.
A NEW HOPE!



9.
[insert awesome
theme song]



10.
I ♥ CODING!

Debugging in



Error messages can be helpful, google them.

Use  's interactive capabilities to your advantage.

Make it repeatable.

Remember  really, really wants to complete any call.

```
> 1+c(1,3,4)
[1] 2 4 5
```

Finding help for

- stackoverflow
- #rstats twitter
- rOpenScience forum
- RStudio community
- Bioconductor support forum
- RLadies slack



Best resources for beginners



Jesse Mostipak is in the #SLICED playoffs!
@kierisi

over in Slack-land a colleague asked for resources on learning [#rstats](#), with a particular emphasis on resources aimed at beginners.

🌟so🌟

here is a thread of my personal favorites:



1:10 AM · Oct 9, 2018 · Twitter Web Client

- MODERN DIVE
- R for Data Science
- Stat545
- Chromebook Data Science
- swirl



