Biostatistics

Lecture 2: Datasets

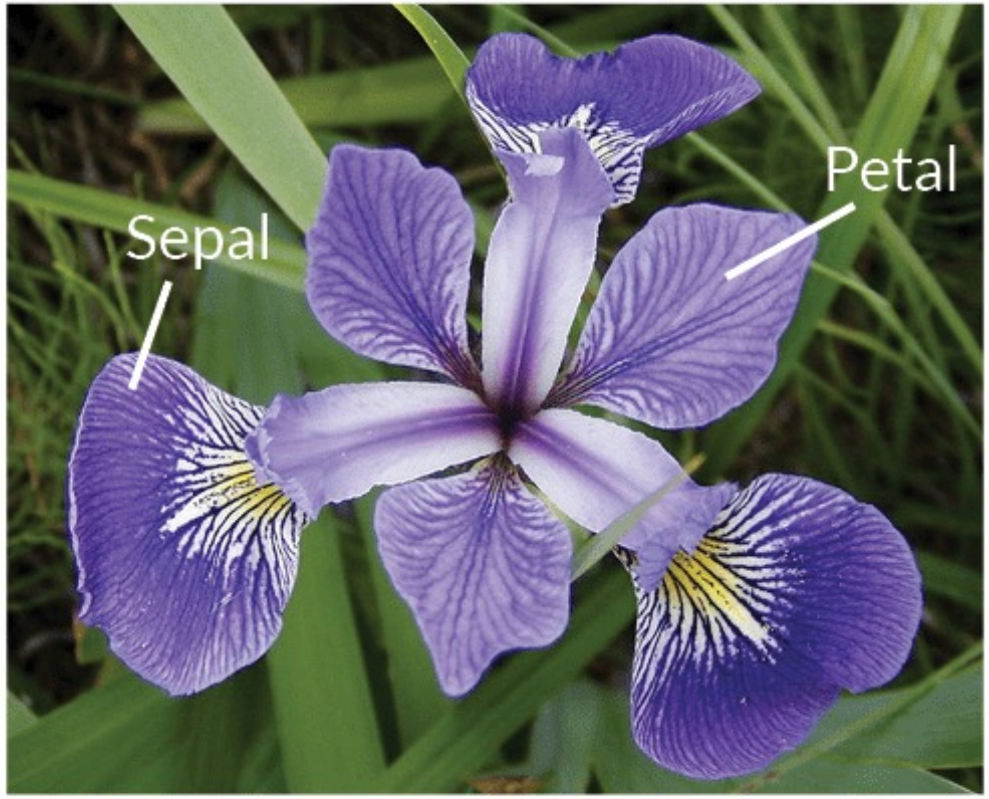
Todays goal

- Using statements to subset from datasets
- Using packages
- Create a dataset and upload it into R

We will use the iris dataset that is already included in R to explore using statements and functions

Simply type the below command in the console

iris_data <- iris



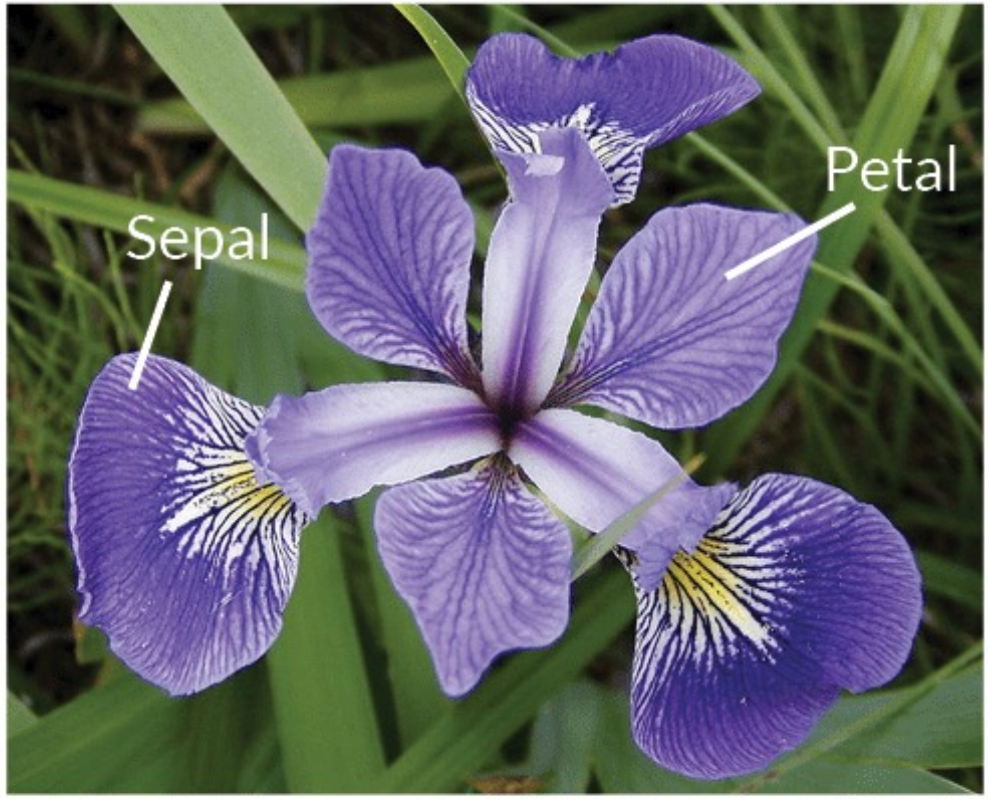
Iris Versicolor

Iris Setosa

Iris Virginica

This dataset describes the length and width of each of the Sepals and Petals of 3 species of Iris
Use the head() command to quickly look at the first 10 rows of a dataset

head(iris_data)



Iris Versicolor

Iris Setosa

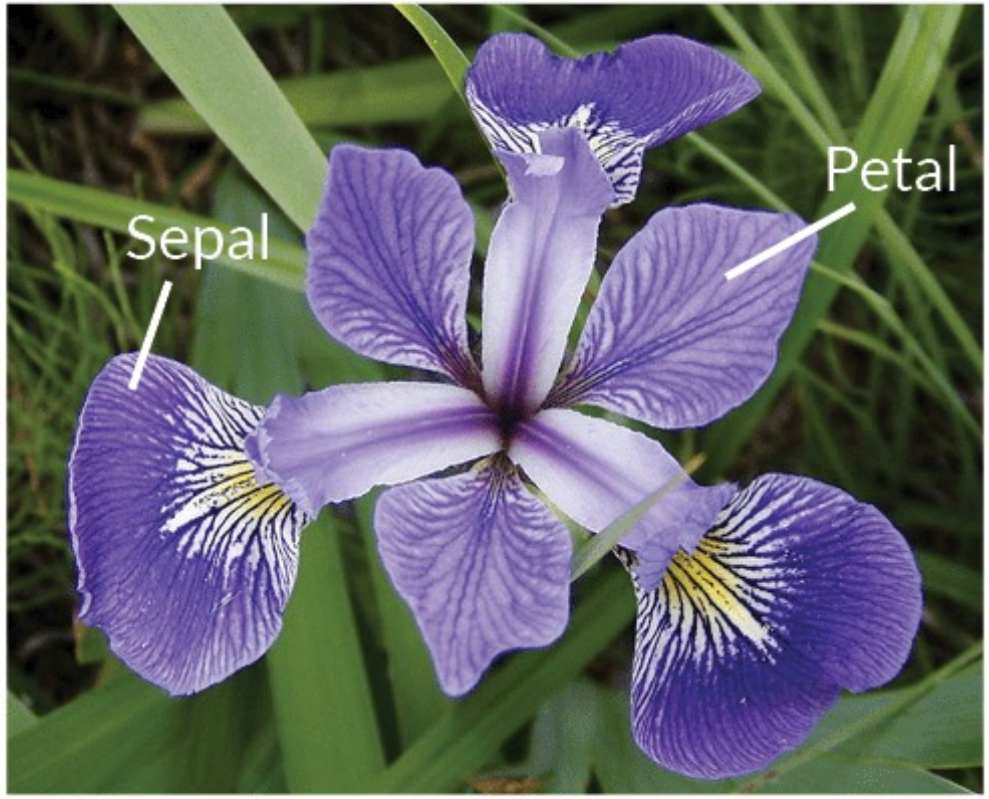
Iris Virginica

We can select columns in several ways,
The follow three lines each select the first column

```
iris_data[,1]

iris_data[ , "Sepal.Length"]

iris_data$Sepal.Length
```



Iris Versicolor

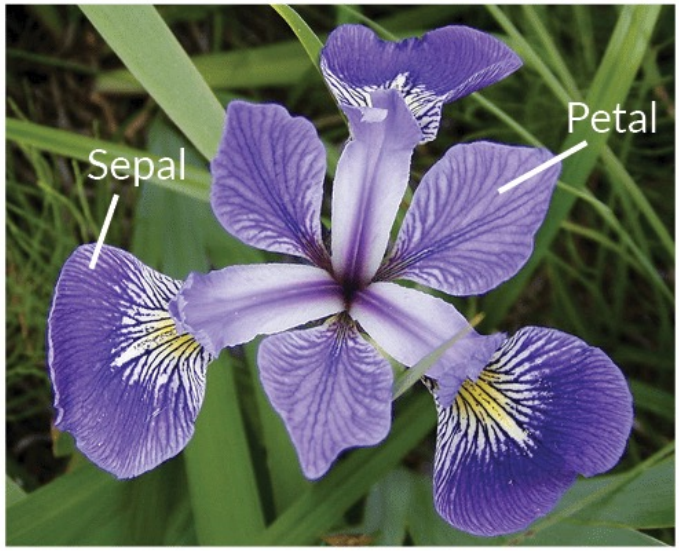
Iris Setosa

Iris Virginica

However, we might want to calculate the mean petal length of each species again there are several ways to do this,

we could ask just the rows associated with a given species

For example, Iris_data[51:100,] will give the rows 51 to 100 which correspond to the species Irish Versicolor



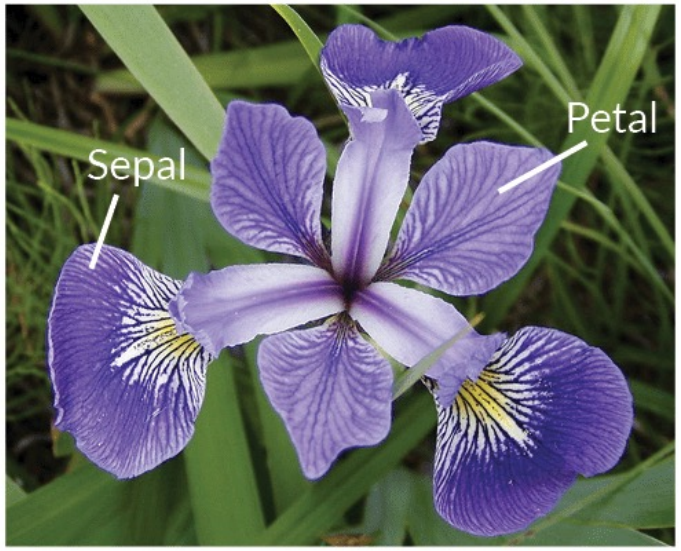
Iris Versicolor

```
>
>
>
>
>
>
>
>
>
```

However, we might want to calculate the mean petal length of each species again there are several ways to do this,

A more efficient way would be to use a statement were we ask when something is True using

==

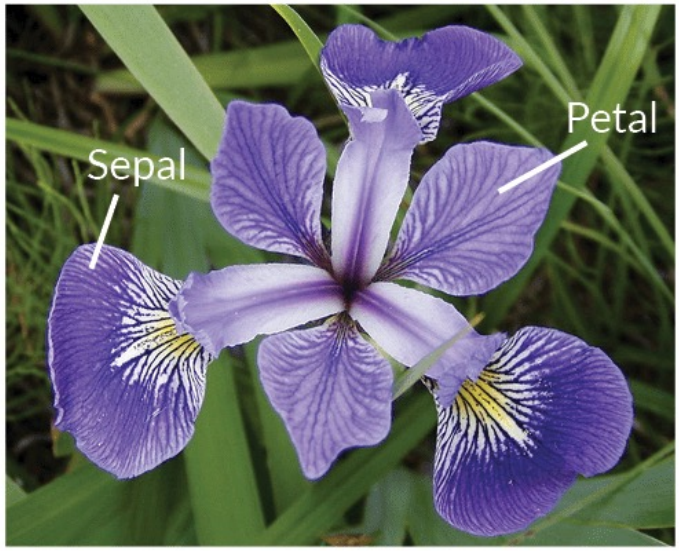


Iris Versicolor

For example

```
iris_data[, “Species”] == "versicolor"
```

asks which entries for the column “Species” in the iris_data object are called "versicolor"

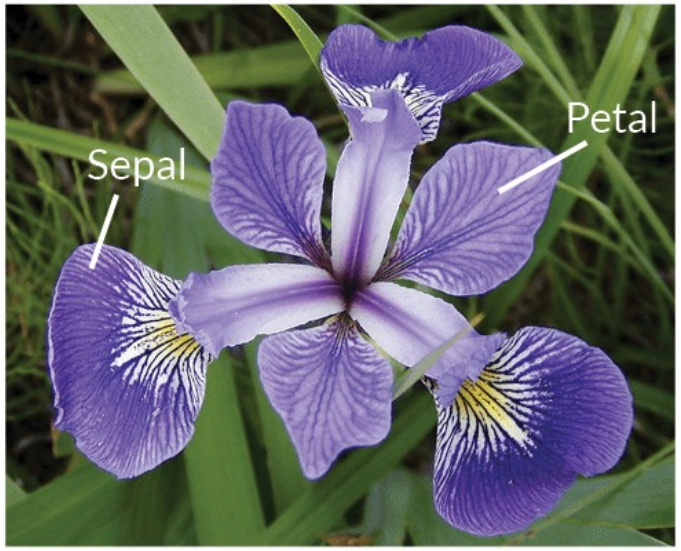


Iris Versicolor

```
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
```

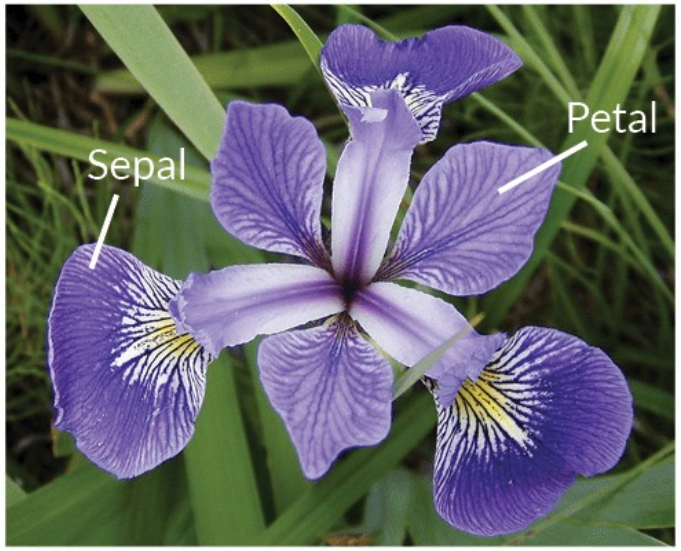
These will also do the same thing

```
iris_data[,5] == "versicolor"
iris_data$Species == "versicolor”
```



Iris Versicolor

```
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
```



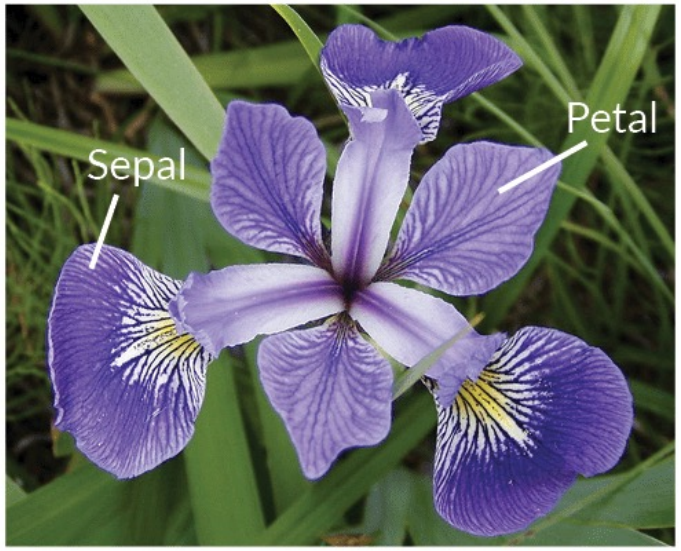
Iris Versicolor

We can use the == symbols to ask for just the rows when this argument is TRUE

```
iris_data[iris_data$Species == "versicolor", ]
```

If we put a vector of TRUE and FALSE values into a data.frame in R it will only select the rows where the statement is TRUE

In this case R will only select the rows where the Species entry is “versicolor”



Iris Versicolor

```
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
```

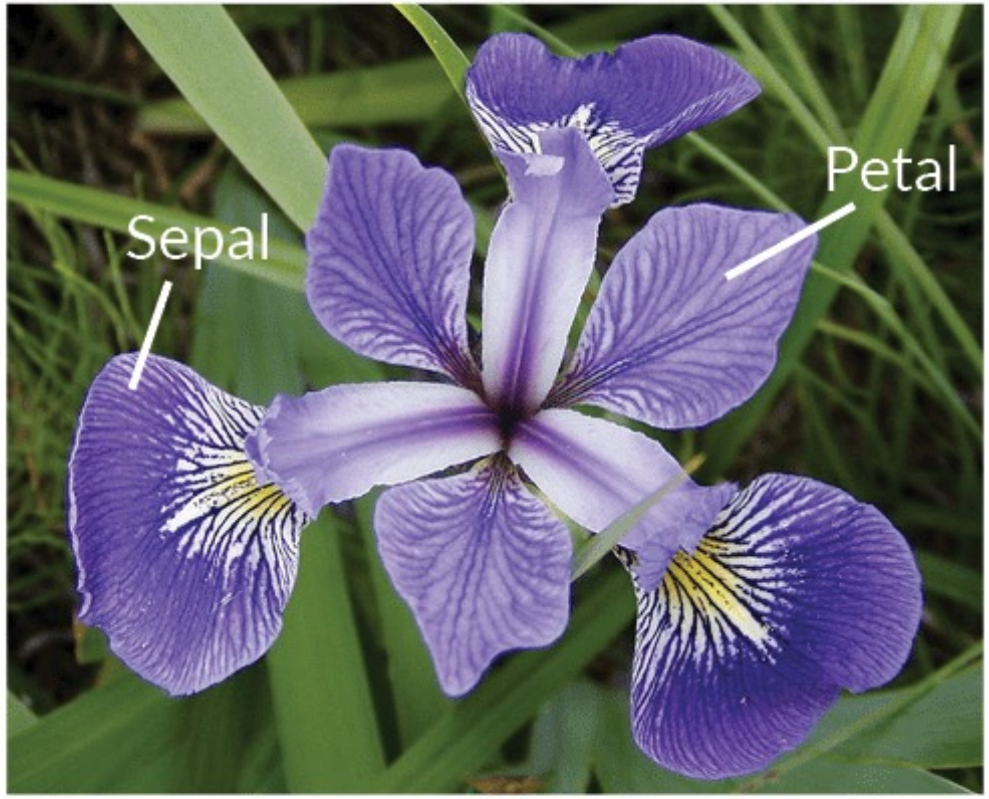
We can also use statements to give use all entries except the named one using

!=

Which means does not equal to.

For example, the follow line will give all species except versicolor

```
iris_data[ iris_data$Species != "versicolor" ,]
```



Iris Versicolor

Iris Setosa

Iris Virginica

Excel is excellent for building datasets

Column are variables

	A	B	C
1	species	class	longevity
2	Dolichotis_patagonum	Mammalia	-0.149004113
3	Eidolon_helvum	Mammalia	0.468611104
4	Elephas_maximus	Mammalia	2.107128571
5	Equus_asinus	Mammalia	1.612802371
6	Equus_burchellii	Mammalia	1.296219408
7	Equus_caballus	Mammalia	1.900107602
8	Equus_grevyi	Mammalia	0.992984918
9	Gazella_dorcas	Mammalia	0.593070554
10	Gazella_gazella	Mammalia	0.207957814
11	Gazella_subgutturosa	Mammalia	0.03558398
12	Giraffa_camelopardalis	Mammalia	1.353879701
13	Glossophaga_soricina	Mammalia	-0.550140921
14	Gorilla_gorilla	Mammalia	1.857702701
15	Acinonyx_jubatus	Mammalia	0.37703697

Rows are observations

(i.e. the measurements)

Excel is excellent for building datasets

	A	B	C
1	species	class	longevity
2	Dolichotis_patagonum	Mammalia	-0.149004113
3	Eidolon_helvum	Mammalia	0.468611104
4	Elephas_maximus	Mammalia	2.107128571
5	Equus_asinus	Mammalia	1.612802371
6	Equus_burchellii	Mammalia	1.296219408
7	Equus_caballus	Mammalia	1.900107602
8	Equus_grevyi	Mammalia	0.992984918
9	Gazella_dorcas	Mammalia	0.593070554
10	Gazella_gazella	Mammalia	0.207957814
11	Gazella_subgutturosa	Mammalia	0.03558398
12	Giraffa_camelopardalis	Mammalia	1.353879701
13	Glossophaga_soricina	Mammalia	-0.550140921
14	Gorilla_gorilla	Mammalia	1.857702701
15	Acinonyx_jubatus	Mammalia	0.37703697

Avoid including spaces

What looks like an empty space to you might look like something else to the computer.

For example, in R the species names “Gorilla_gorilla”, “Gorilla_gorilla ” and “Gorilla gorilla” will all be seen as different because of the spaces.

Excel is excellent for building datasets

	A	B	C
1	species	class	longevity
2	Dolichotis_patagonum	Mammalia	-0.149004113
3	Eidolon_helvum	Mammalia	0.468611104
4	Elephas_maximus	Mammalia	2.107128571
5	Equus_asinus	Mammalia	1.612802371
6	Equus_burchellii	Mammalia	1.296219408
7	Equus_caballus	Mammalia	1.900107602
8	Equus_grevyi	Mammalia	0.992984918
9	Gazella_dorcas	Mammalia	0.593070554
10	Gazella_gazella	Mammalia	0.207957814
11	Gazella_subgutturosa	Mammalia	0.03558398
12	Giraffa_camelopardalis	Mammalia	1.353879701
13	Glossophaga_soricina	Mammalia	-0.550140921
14	Gorilla_gorilla	Mammalia	1.857702701
15	Acinonyx_jubatus	Mammalia	0.37703697

Avoid using special characters like $, :, ‘, % @ etc.

These characters are often special instructions in computer language.

Excel is excellent for building datasets

	A	B	C
1	species	class	longevity
2	Dolichotis_patagonum	Mammalia	-0.149004113
3	Eidolon_helvum	Mammalia	0.468611104
4	Elephas_maximus	Mammalia	2.107128571
5	Equus_asinus	Mammalia	1.612802371
6	Equus_burchellii	Mammalia	1.296219408
7	Equus_caballus	Mammalia	1.900107602
8	Equus_grevyi	Mammalia	0.992984918
9	Gazella_dorcas	Mammalia	0.593070554
10	Gazella_gazella	Mammalia	0.207957814
11	Gazella_subgutturosa	Mammalia	0.03558398
12	Giraffa_camelopardalis	Mammalia	1.353879701
13	Glossophaga_soricina	Mammalia	-0.550140921
14	Gorilla_gorilla	Mammalia	1.857702701
15	Acinonyx_jubatus	Mammalia	0.37703697

Keep it simple - The dataset should never look any more complex than columns and row.

Don’t have empty column in between data, or floating data to the side.

Don’t use multiple sheets. (create a new excel file instead)

Avoid calculations in Excel.

(That’s what R is for)

Don’t use colors, this data will be lost once saved,

For the same data in excel

What we see

	A	B	C
1	species	class	longevity
2	Dolichotis_patagonum	Mammalia	-0.149004113
3	Eidolon_helvum	Mammalia	0.468611104
4	Elephas_maximus	Mammalia	2.107128571
5	Equus_asinus	Mammalia	1.612802371
6	Equus_burchellii	Mammalia	1.296219408
7	Equus_caballus	Mammalia	1.900107602
8	Equus_grevyi	Mammalia	0.992984918
9	Gazella_dorcas	Mammalia	0.593070554
10	Gazella_gazella	Mammalia	0.207957814
11	Gazella_subgutturosa	Mammalia	0.03558398
12	Giraffa_camelopardalis	Mammalia	1.353879701
13	Glossophaga_soricina	Mammalia	-0.550140921
14	Gorilla_gorilla	Mammalia	1.857702701
15	Acinonyx_jubatus	Mammalia	0.37703697

What your computer sees


```
"","species","class","longevity","mass","volant"
"1","Dolichotis_patagonum","Mammalia",-0.149004113,1.087544642,"nonvolant"
"2","Eidolon_helvum","Mammalia",0.468611104,-0.274833698,"volant"
"3","Elephas_maximus","Mammalia",2.107128571,3.122033965,"nonvolant"
"4","Equus_asinus","Mammalia",1.612802371,2.035276413,"nonvolant"
"5","Equus_burchellii","Mammalia",1.296219408,2.229529943,"nonvolant"
"6","Equus_caballus","Mammalia",1.900107602,2.254871628,"nonvolant"
"7","Equus_grevyi","Mammalia",0.992984918,2.345545495,"nonvolant"
"8","Gazella_dorcas","Mammalia",0.593070554,1.142060609,"nonvolant"
"9","Gazella_gazella","Mammalia",0.207957814,1.315487817,"nonvolant"
"10","Gazella_subgutturosa","Mammalia",0.03558398,1.58555476,"nonvolant"
"11","Giraffa_camelopardalis","Mammalia",1.353879701,2.615138797,"nonvolant"
"12","Glossophaga_soricina","Mammalia",-0.550140921,-1.531696908,"volant"
"13","Gorilla_gorilla","Mammalia",1.857702701,1.974516149,"nonvolant"
"14","Acinonyx_jubatus","Mammalia",0.37703697,1.621594307,"nonvolant"
"15","Acomys_cahirinus","Mammalia",-1.477934381,-0.97760725,"nonvolant"
"16","Addax_nasomaculatus","Mammalia",0.84139267,1.822705763,"nonvolant"
"17","Aepyceros_melampus","Mammalia",0.707926871,1.614663748,"nonvolant"
"18","Ailurus_fulgens","Mammalia",0.263865723,0.697713801,"nonvolant"
"19","Ammotragus_lervia","Mammalia",0.461763407,1.822705763,"nonvolant"
"20","Antidorcas_marsupialis","Mammalia",0.325291811,1.505480662,"nonvolant"
"21","Antilope_cervicapra","Mammalia",0.605586355,1.491074551,"nonvolant"
"22","Aonyx_cinerea","Mammalia",0.548417955,0.563352314,"nonvolant"
"23","Artibeus_jamaicensis","Mammalia",0.279461379,-1.002832864,"volant"
"24","Atelerix_albiventris","Mammalia",-0.496943425,-0.09477662,"nonvolant"
"25","Ateles_fusciceps","Mammalia",1.580771641,0.874572234,"nonvolant"
"26","Ateles_geoffroyi","Mammalia",1.61596788,0.888347122,"nonvolant"
"27","Axis_axis","Mammalia",0.39867471,1.476080275,"nonvolant"
"28","Bison_bison","Mammalia",1.108497945,2.527391708,"nonvolant"
"29","Bos_grunniens","Mammalia",0.748105033,2.548354102,"nonvolant"
"30","Boselaphus_tragocamelus","Mammalia",0.461763407,2.067240906,"nonvolant"
"31","Callimico_goeldii","Mammalia",0.495691333,-0.056444273,"nonvolant"
"32","Callithrix_geoffroyi","Mammalia",0.14986911,-0.234279515,"nonvolant"
"33","Callithrix_jacchus","Mammalia",0.535410261,-0.341813995,"nonvolant"
```

Why do this?

The first row is called the header and is not treated as a row but usually donates the names of the columns.

Notice how all the data is separated by a comma

This is how the computer knows how to sperate the data Into columns.

When we save our excel data we save it as a comma separated value or .csv file

We are in effect telling the computer exactly how we want it to store our data.

What your computer sees

```
"","species","class","longevity","mass","volant"
"1","Dolichotis_patagonum","Mammalia",-0.149004113,1.087544642,"nonvolant"
"2","Eidolon_helvum","Mammalia",0.468611104,-0.274833698,"volant"
"3","Elephas_maximus","Mammalia",2.107128571,3.122033965,"nonvolant"
"4","Equus_asinus","Mammalia",1.612802371,2.035276413,"nonvolant"
"5","Equus_burchellii","Mammalia",1.296219408,2.229529943,"nonvolant"
"6","Equus_caballus","Mammalia",1.900107602,2.254871628,"nonvolant"
"7","Equus_grevyi","Mammalia",0.992984918,2.345545495,"nonvolant"
"8","Gazella_dorcas","Mammalia",0.593070554,1.142060609,"nonvolant"
"9","Gazella_gazella","Mammalia",0.207957814,1.315487817,"nonvolant"
"10","Gazella_subgutturosa","Mammalia",0.03558398,1.58555476,"nonvolant"
"11","Giraffa_camelopardalis","Mammalia",1.353879701,2.615138797,"nonvolant"
"12","Glossophaga_soricina","Mammalia",-0.550140921,-1.531696908,"volant"
"13","Gorilla_gorilla","Mammalia",1.857702701,1.974516149,"nonvolant"
"14","Acinonyx_jubatus","Mammalia",0.37703697,1.621594307,"nonvolant"
"15","Acomys_cahirinus","Mammalia",-1.477934381,-0.97760725,"nonvolant"
"16","Addax_nasomaculatus","Mammalia",0.84139267,1.822705763,"nonvolant"
"17","Aepyceros_melampus","Mammalia",0.707926871,1.614663748,"nonvolant"
"18","Ailurus_fulgens","Mammalia",0.263865723,0.697713801,"nonvolant"
"19","Ammotragus_lervia","Mammalia",0.461763407,1.822705763,"nonvolant"
"20","Antidorcas_marsupialis","Mammalia",0.325291811,1.505480662,"nonvolant"
"21","Antilope_cervicapra","Mammalia",0.605586355,1.491074551,"nonvolant"
"22","Aonyx_cinerea","Mammalia",0.548417955,0.563352314,"nonvolant"
"23","Artibeus_jamaicensis","Mammalia",0.279461379,-1.002832864,"volant"
"24","Atelerix_albiventris","Mammalia",-0.496943425,-0.09477662,"nonvolant"
"25","Ateles_fusciceps","Mammalia",1.580771641,0.874572234,"nonvolant"
"26","Ateles_geoffroyi","Mammalia",1.61596788,0.888347122,"nonvolant"
"27","Axis_axis","Mammalia",0.39867471,1.476080275,"nonvolant"
"28","Bison_bison","Mammalia",1.108497945,2.527391708,"nonvolant"
"29","Bos_grunniens","Mammalia",0.748105033,2.548354102,"nonvolant"
"30","Boselaphus_tragocamelus","Mammalia",0.461763407,2.067240906,"nonvolant"
"31","Callimico_goeldii","Mammalia",0.495691333,-0.056444273,"nonvolant"
"32","Callithrix_geoffroyi","Mammalia",0.14986911,-0.234279515,"nonvolant"
"33","Callithrix_jacchus","Mammalia",0.535410261,-0.341813995,"nonvolant"
```

Save you data using a sensible name that describes what it is and with the date

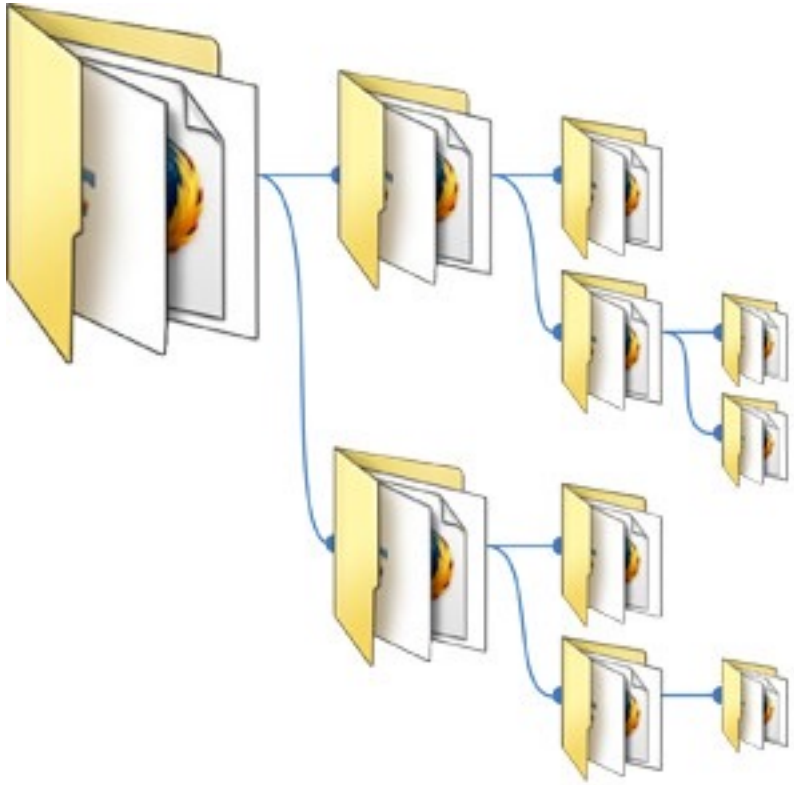
For example

ZO208_data_23_11_2023.csv

Working directory

Computer term for a folder and the location of the folder.

For every project you do in R you should make a new folder and put all data and R files in it.



Computers are organized
As hierarchical folders

Your desktop for example is just a folder

Working directory

Computer term for a folder and the location of the folder.

For every project you do in R you should make a new folder and put all data and R files in it.

In R you can ask what folder it’s currently is in
By typing getwd()

You can also manually set what folder it’s currently is in by typing setwd()

GIPHY CAPTURE File Edit Window

RStudio

Go to file/function Addins

Console Terminal Background Jobs

R 4.2.2 · ~/

```
R version 4.2.2 (2022-10-31) -- "Innocent and Trusting"
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: aarch64-apple-darwin20 (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.

[Workspace loaded from ~/.RData]

>
```

Working directory

R can only see the files in its current directory
When uploading data the saved csv file needs to be in the current working directory otherwise R won’t find it.

One of the most common errors is not setting the correct directory

Before Importing data into R check

- Data is saved as a .csv file
- Data file is named appropriately
- File is in specific folder
- Working directory is set to folder

Run the following line of code

This saves the data into R as an object

This is the name of the data file you wan to import

This part tells R that the first row is the column names

```
New_data <- read.csv("data_file_name.csv", header = T, sep = ",")
```

This is the function that will read in the data

This part tells R that the data is separated by commas

Run the following line of code

This is the name of the
data file you wan to import

```
New_data <- read.csv(“data_file_name.csv”, header = T, sep = “,”)
```

*Make sure the file name is spelt exactly the same as
how you saved it, its also case sensitive.
**
It also needs to be in quotation marks “”

Run the following line of code

This part tells R that the first row is the column names

```
New_data <- read.csv("data_file_name.csv", header = T, sep = ",")
```

*This is a default setting already, but its good practice to explicitly state things

Run the following line of code

This part tells R that the data is separated by commas

```
New_data <- read.csv(“data_file_name.csv”, header = T, sep = “,”)
```

*This is a default setting already, but its good practice to explicitly state things

Example

Download the file from CANVAS called

lifespan_data_22_7_2023.xls

in a new folder called “First stats” on your desktop

Change the directory in R and upload the new file using

```
New_data <- read.csv(”lifespan_data_22_7_2023.csv”, header = T, sep = “,”)
```

Check it worked using

```
head(New_data)
```

If you get an error saying "No such file or directory" go back and check

- Data is saved as a .csv file
- Data file is named appropriately
- File is in specific folder
- Working directory is set to that folder

Packages

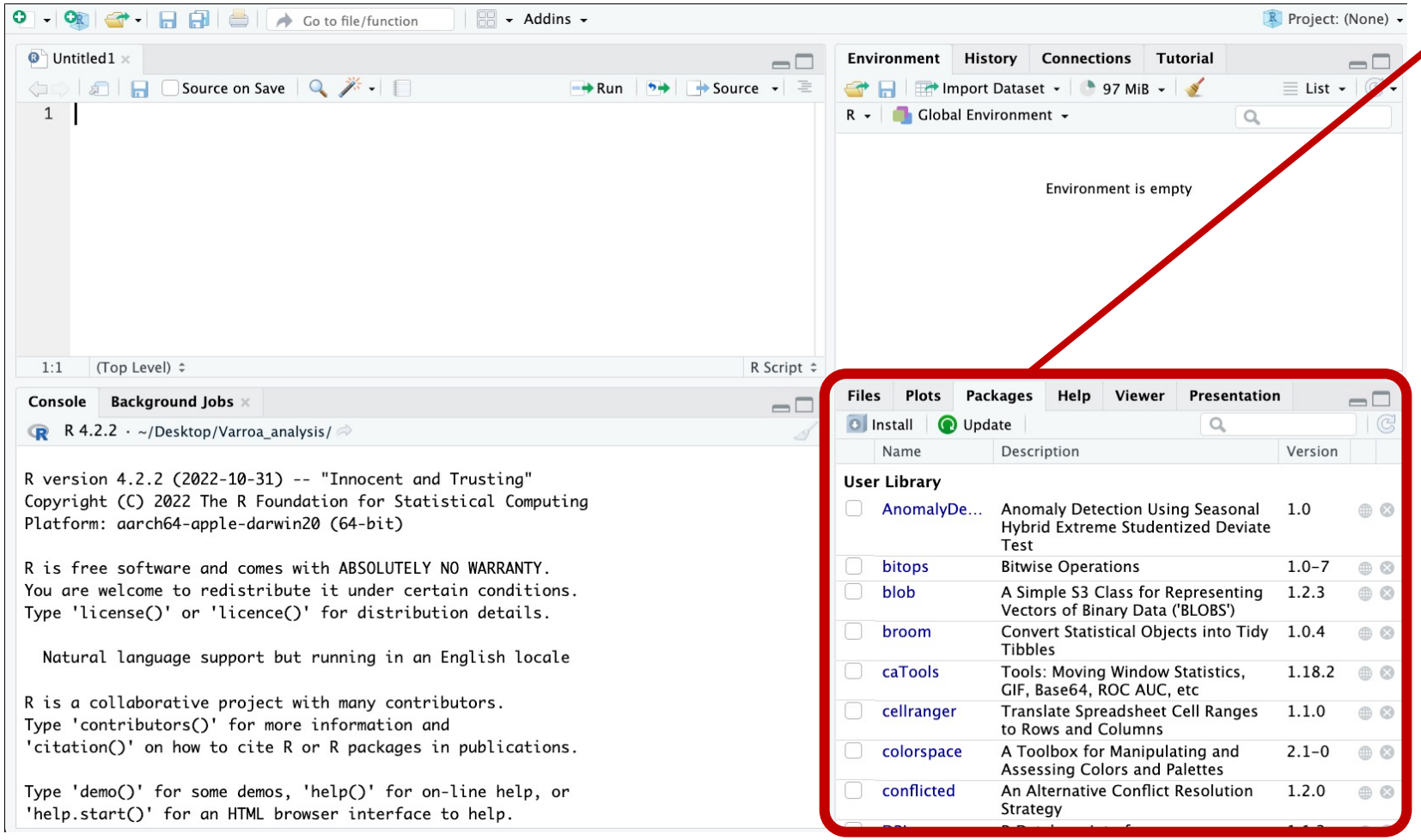
One of the true strength so R is that packages can offer a enormous variety of functions and test you can use

Packages

One of the true strength so R is that packages can offer a enormous variety of functions and test you can use

R comes with lots of packages as standard

All current the packages can be found here.



Packages

One of the true strength so R is that packages can offer a enormous variety of functions and test you can use

We can also install new packages using the code below

```
install.packages(“name_of_package”)
```

EXAMPLE

```
#We can install the package hdrcde which contains a function to calculate the mode
install.packages("hdrcde")

#the function library tells R you want to use it now and loads it up
#This needs to be done every time you start R again
library("hdrcde")

#hdr is the function that will calculate the mode
hdr(iris_data[iris_data$Species == "versicolor","Petal.Length"])
```

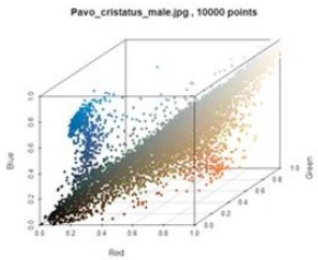
Packages can do a range of things

And are built by the scientists that use them

Some examples in my lab

`colordistance`

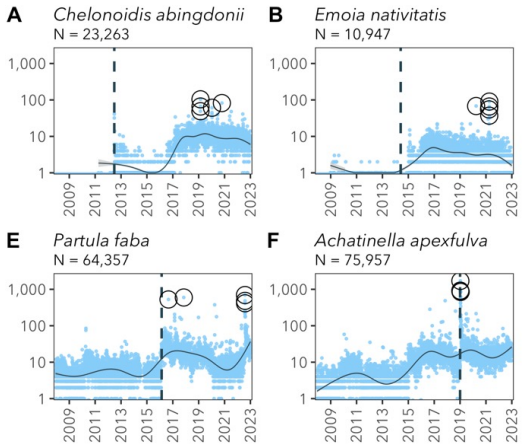
Quantifies colors from images
So that we can test why some animals evolve bright displays



Link to package

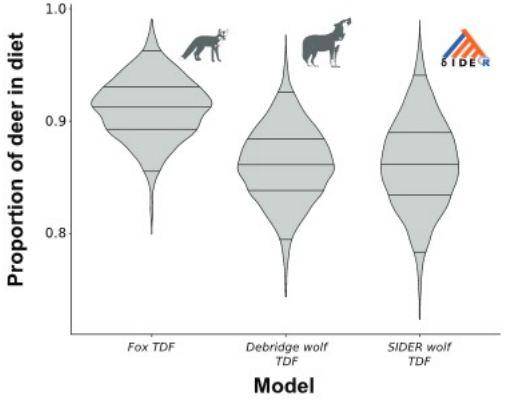
`pageviews`

Page views for on Wikipedia.
We use it to test how the public reacts to the announcement of species extinctions



`SIDER`

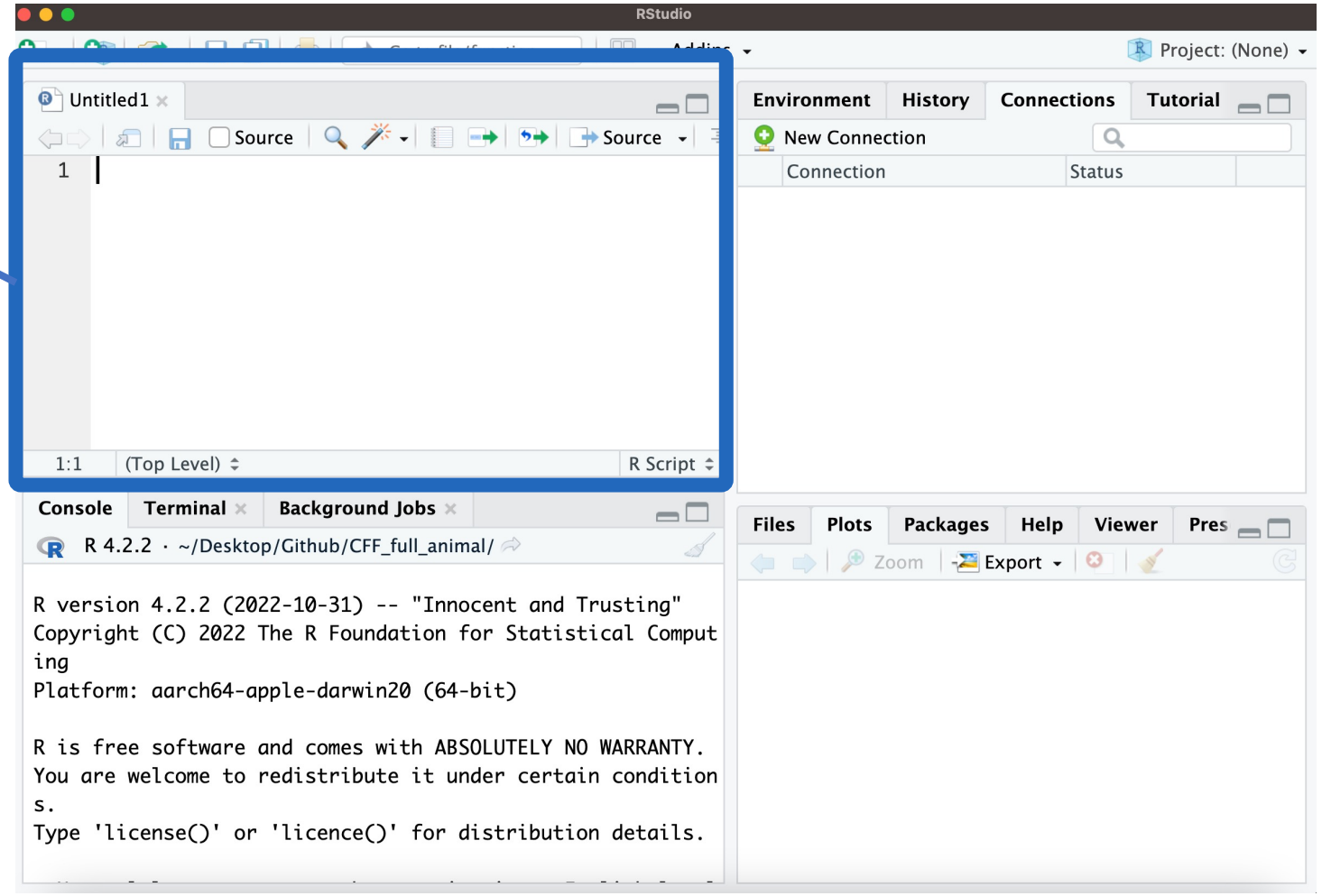
Packaged my lab wrote which estimates parameters needed to estimate animal diets when using stable isotopes



https://onlinelibrary.wiley.com/doi/epdf/10.1111/ecog.03371

Reproducibility using scripts

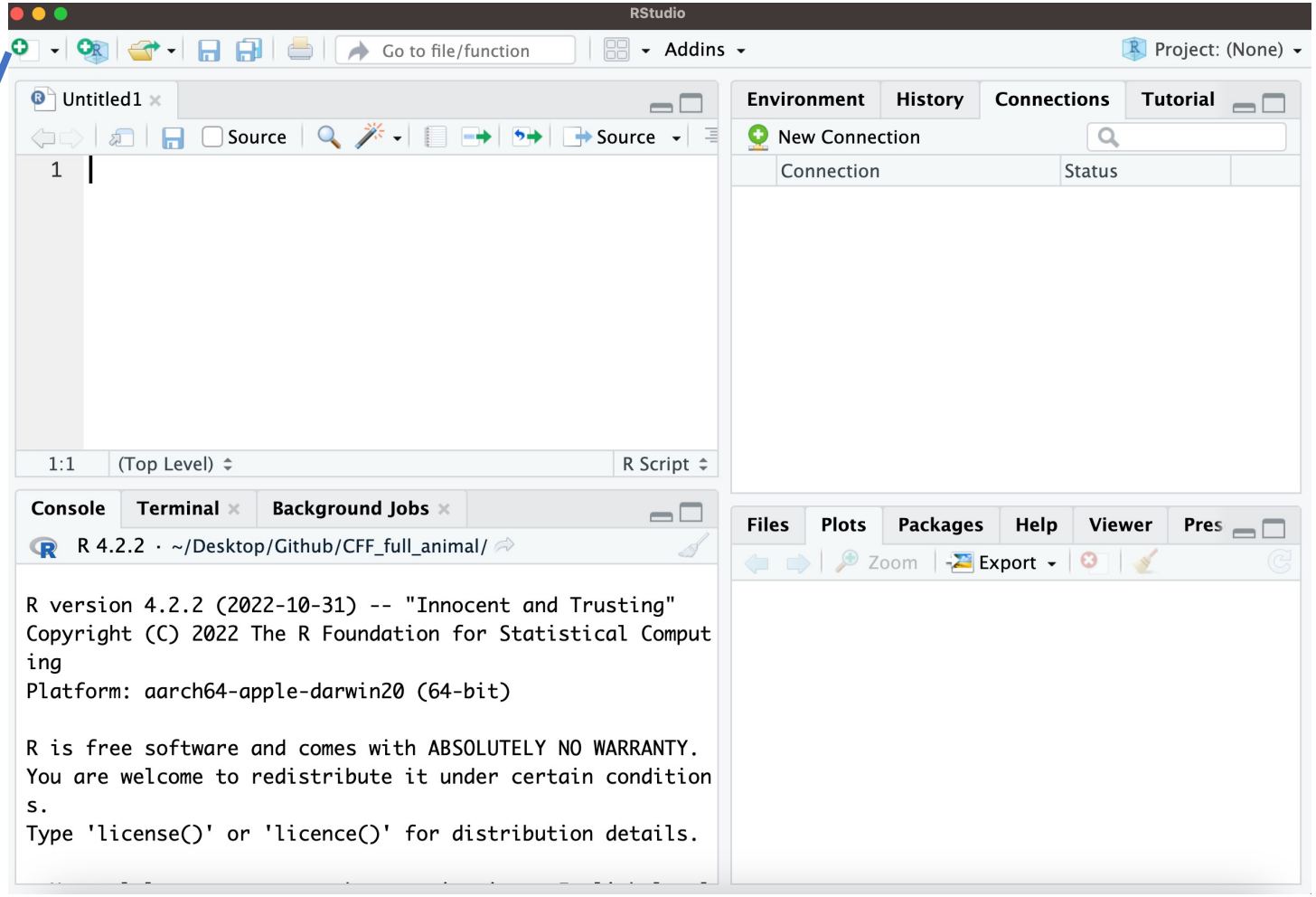
The top left is called the script. This is where you write code that can be saved just like a plain txt file



Reproducibility using scripts

We will write the code we want to save in scripts
Saves progress and allows others to easier repeat what we did

We can start a new script by clicking here and selecting R script



Reproducibility using scripts

We use hash tags before lines that are not code

Its good to use these to give the script a title, data and author

You should also use these to describe what the code is doing

This is called annotating

Go to file/function Addins

My_first_script_24_7_2023.R*

Source on Save

```
#My first script
#Kevin Healy
#24 July 2023

#Use hash tags to make comments on what you are doing

#This line installs the package hdrcde
install.packages("hdrcde")

#This line loads the package hdrcde
library("hdrcde")

#This line reads in my data
new_data <- read.csv("new_data_24_7_2023.csv", header = T, sep = ",")
```

Reproducibility using scripts

Call the document a sensible name, with a date

When we are finished our session we simply save it

Go to file/function Addins

My_first_script_24_7_2023.R*

Source on Save

```
#My first script
#Kevin Healy
#24 July 2023

#Use hash tags to make comments on what you are doing

#This line installs the package hdrcde
install.packages("hdrcde")

#This line loads the package hdrcde
library("hdrcde")

#This line reads in my data
new_data <- read.csv("new_data_24_7_2023.csv", header = T, sep = ",")
```

Congratulations you are now able to load in data, upload a package and share you code