Research projects

& programming with R

TropBio

Costa Rica June 4th 2018

Organization for Tropical Studies

where science and nature converge

Ronny

HEREDIA
R User
Group

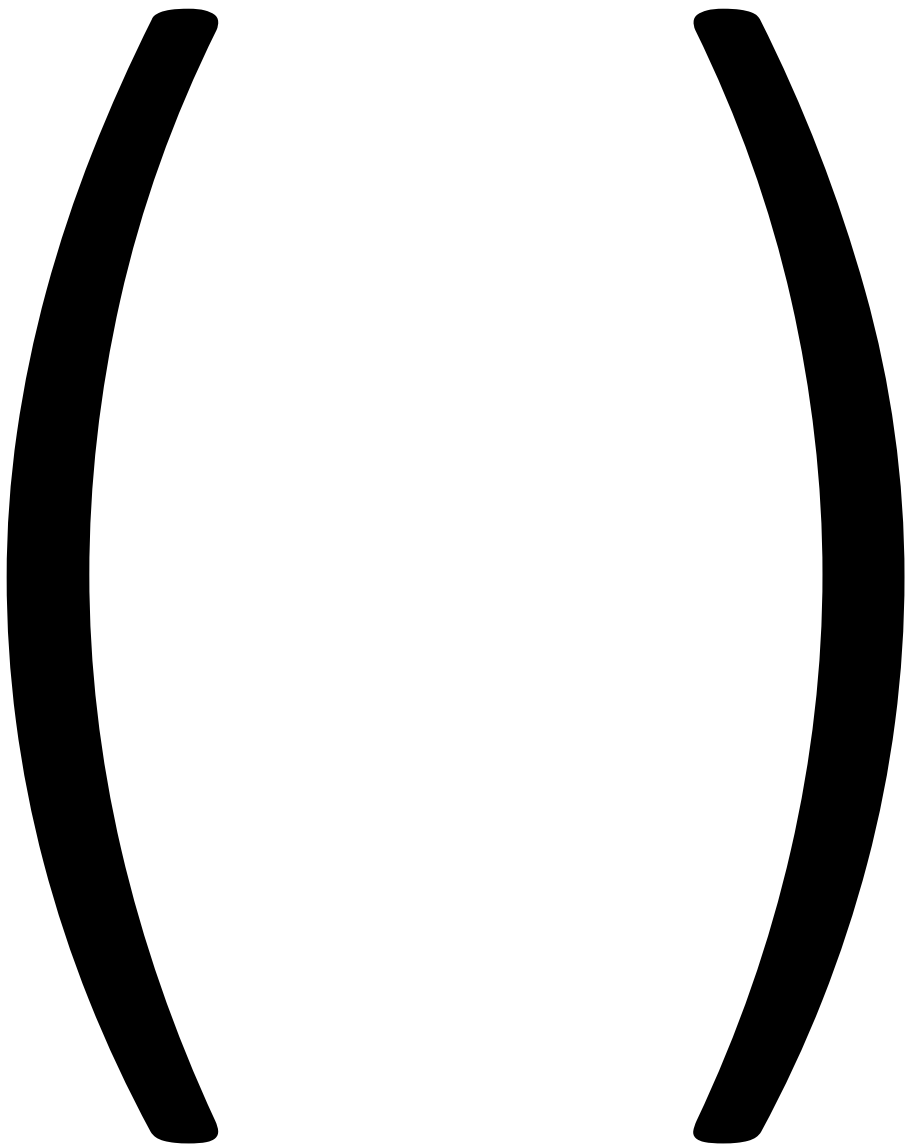
consortium

@RonnyHdezM

ronnyhdez

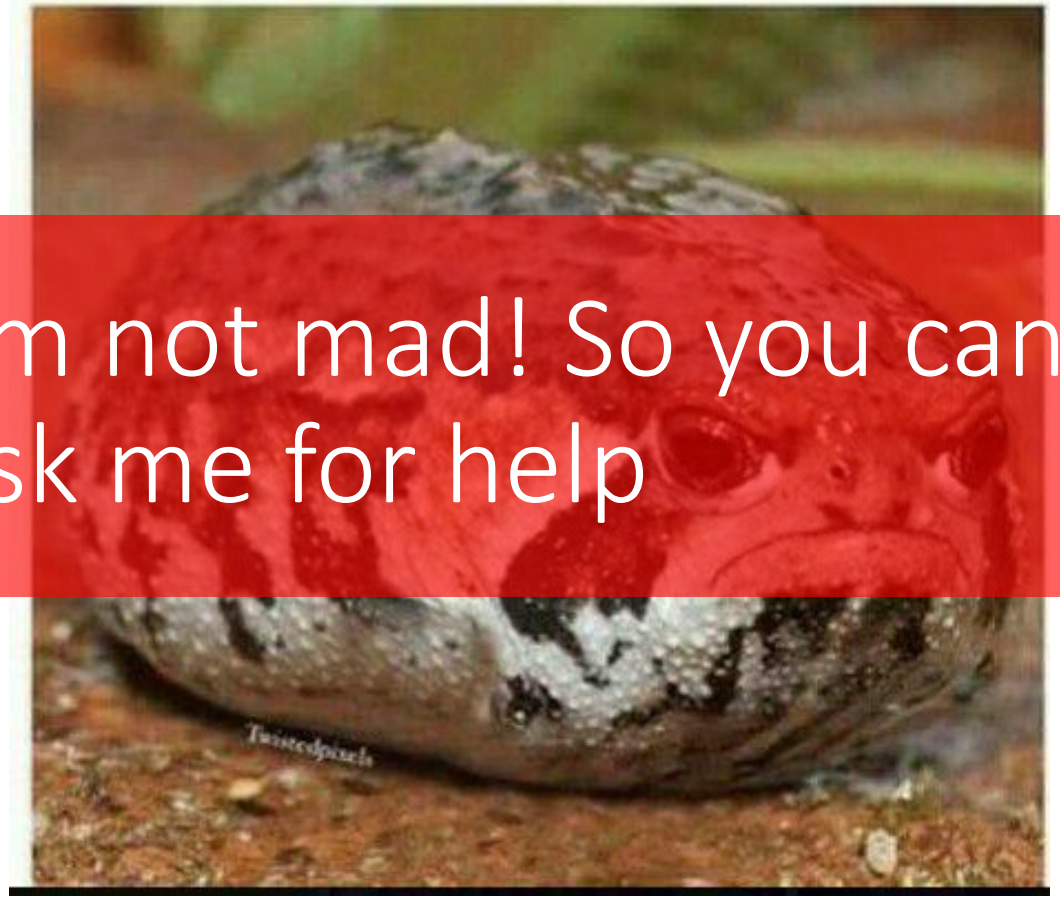
ronny.hernandezm@gmail.com

My life with R:

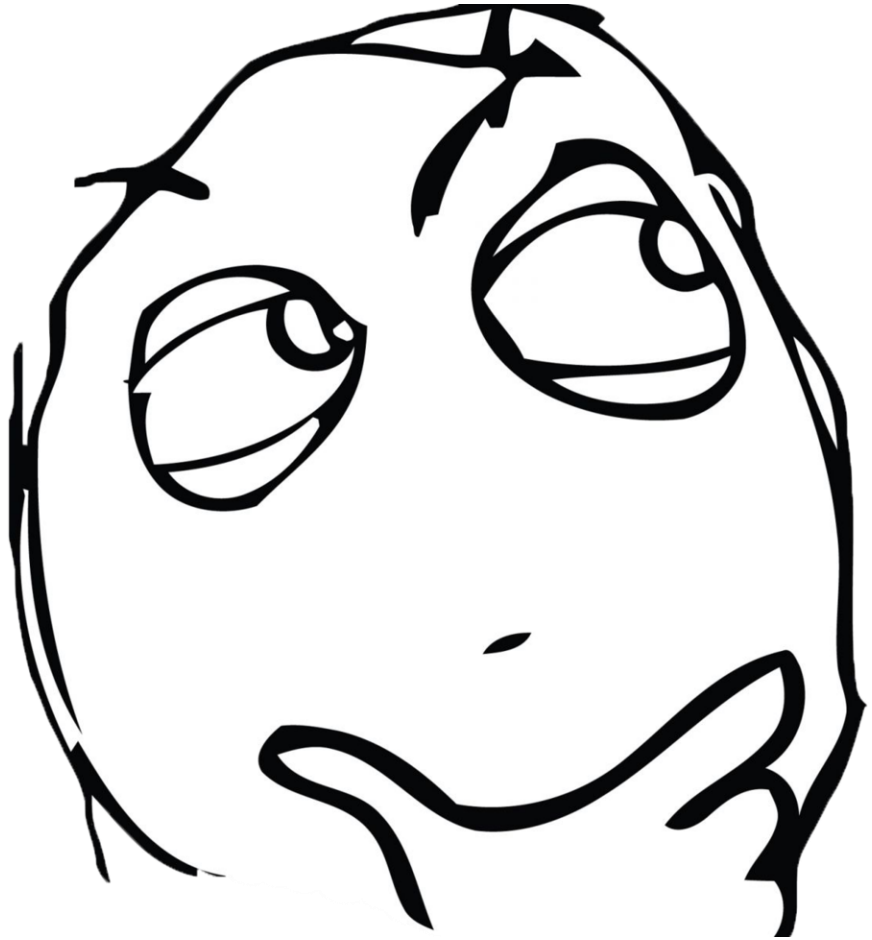


Me: Why do people think I'm mad all the time?

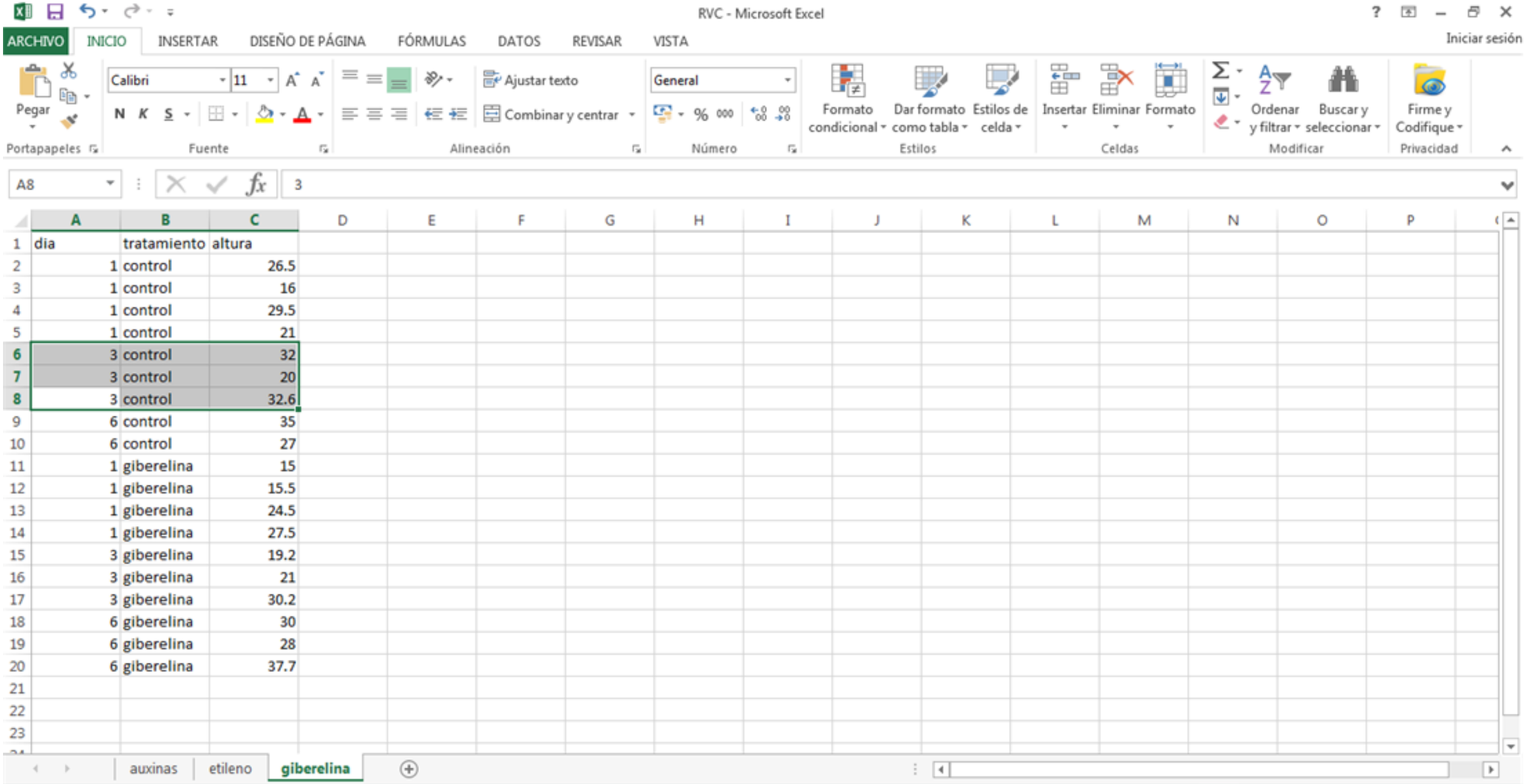
Also me:



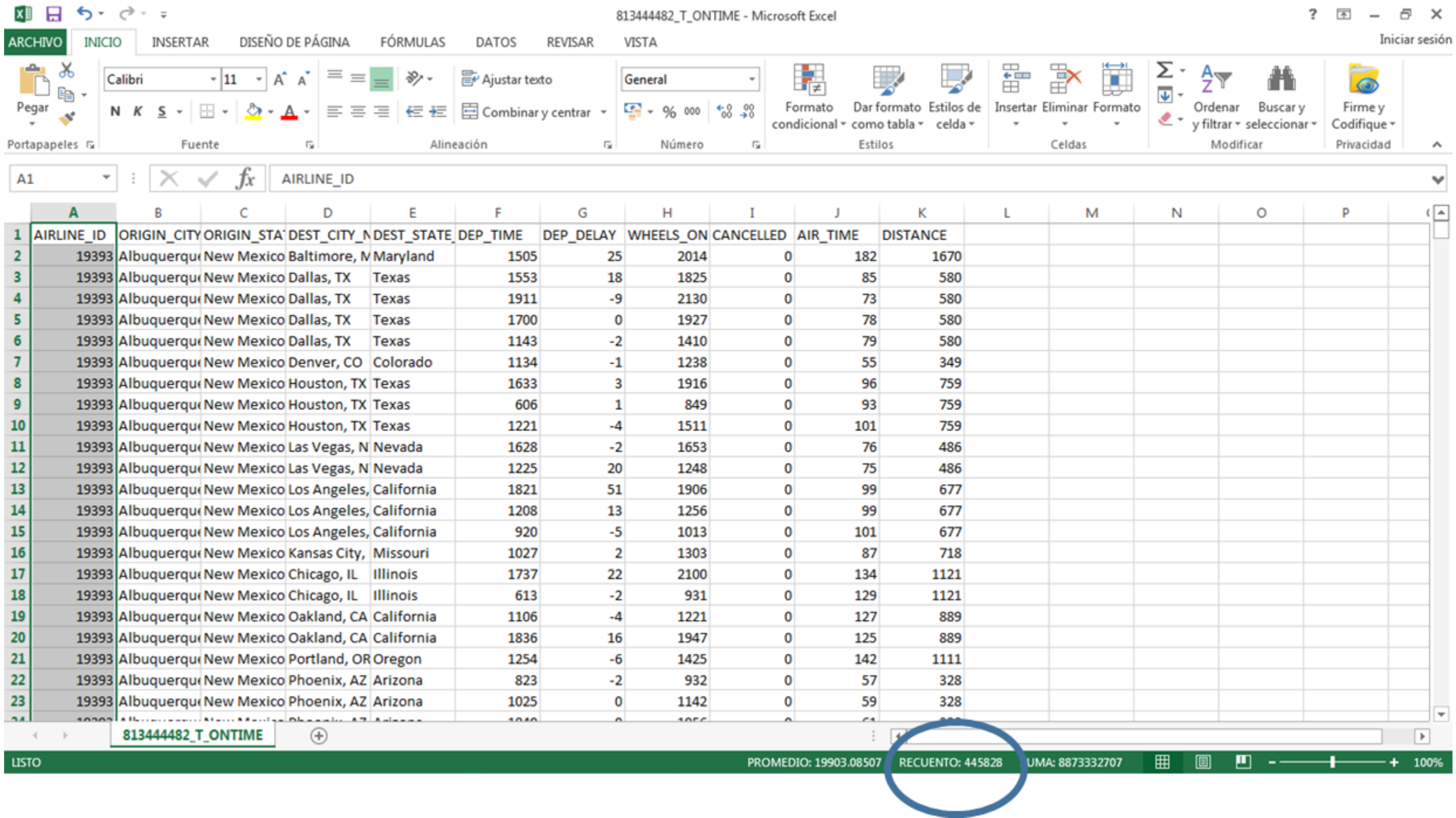
I’m not mad! So you can ask me for help



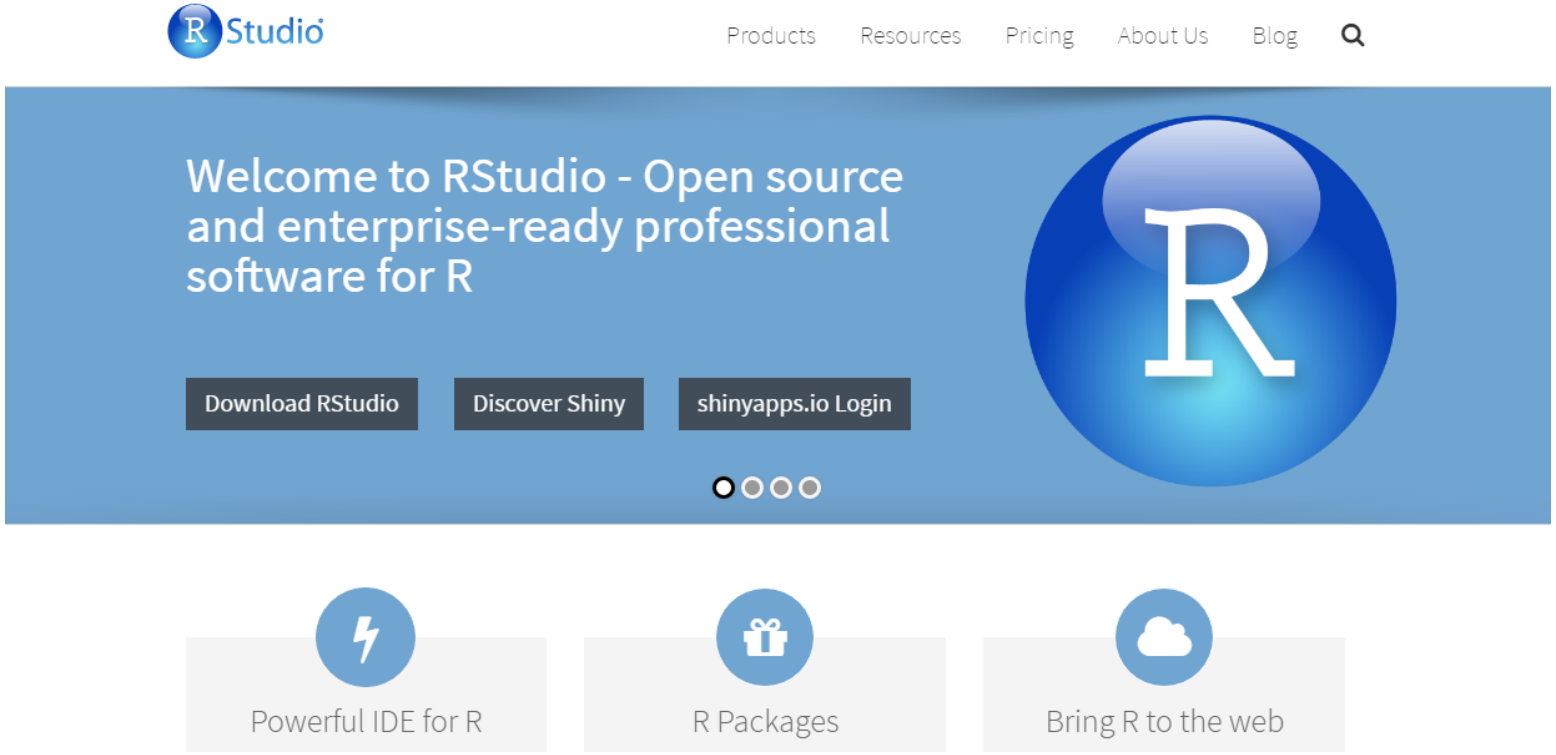
Why R and not a GUI?



	A	B	C
1	dia	tratamiento	altura
2	1	control	26.5
3	1	control	16
4	1	control	29.5
5	1	control	21
6	3	control	32
7	3	control	20
8	3	control	32.6
9	6	control	35
10	6	control	27
11	1	giberelina	15
12	1	giberelina	15.5
13	1	giberelina	24.5
14	1	giberelina	27.5
15	3	giberelina	19.2
16	3	giberelina	21
17	3	giberelina	30.2
18	6	giberelina	30
19	6	giberelina	28
20	6	giberelina	37.7
21			
22			
23			


813444482_T_ONTIME - Microsoft Excel
ARCHIVO
INICIO
INSERTAR
DISEÑO DE PÁGINA
FÓRMULAS
DATOS
REVISAR
VISTA
Iniciar sesión
Pegar
Calibri
11
Ajustar texto
Combinar y centrar
General
Formato condicional
Dar formato como tabla
Estilos de celda
Insertar
Eliminar
Formato
Ordenar y filtrar
Buscar y seleccionar
Firme y Codifique
Portapapeles
Fuente
Alineación
Número
Estilos
Celdas
Modificar
Privacidad
A1
AIRLINE_ID
AIRLINE_ID ORIGIN_CITY ORIGIN_STA DEST_CITY_N DEST_STATE_ DEP_TIME DEP_DELAY WHEELS_ON CANCELLED AIR_TIME DISTANCE
19393 Albuquerque New Mexico Baltimore, M Maryland 1505 25 2014 0 182 1670
19393 Albuquerque New Mexico Dallas, TX Texas 1553 18 1825 0 85 580
19393 Albuquerque New Mexico Dallas, TX Texas 1911 -9 2130 0 73 580
19393 Albuquerque New Mexico Dallas, TX Texas 1700 0 1927 0 78 580
19393 Albuquerque New Mexico Dallas, TX Texas 1143 -2 1410 0 79 580
19393 Albuquerque New Mexico Denver, CO Colorado 1134 -1 1238 0 55 349
19393 Albuquerque New Mexico Houston, TX Texas 1633 3 1916 0 96 759
19393 Albuquerque New Mexico Houston, TX Texas 606 1 849 0 93 759
19393 Albuquerque New Mexico Houston, TX Texas 1221 -4 1511 0 101 759
19393 Albuquerque New Mexico Las Vegas, N Nevada 1628 -2 1653 0 76 486
19393 Albuquerque New Mexico Las Vegas, N Nevada 1225 20 1248 0 75 486
19393 Albuquerque New Mexico Los Angeles, California 1821 51 1906 0 99 677
19393 Albuquerque New Mexico Los Angeles, California 1208 13 1256 0 99 677
19393 Albuquerque New Mexico Los Angeles, California 920 -5 1013 0 101 677
19393 Albuquerque New Mexico Kansas City, Missouri 1027 2 1303 0 87 718
19393 Albuquerque New Mexico Chicago, IL Illinois 1737 22 2100 0 134 1121
19393 Albuquerque New Mexico Chicago, IL Illinois 613 -2 931 0 129 1121
19393 Albuquerque New Mexico Oakland, CA California 1106 -4 1221 0 127 889
19393 Albuquerque New Mexico Oakland, CA California 1836 16 1947 0 125 889
19393 Albuquerque New Mexico Portland, OR Oregon 1254 -6 1425 0 142 1111
19393 Albuquerque New Mexico Phoenix, AZ Arizona 823 -2 932 0 57 328
19393 Albuquerque New Mexico Phoenix, AZ Arizona 1025 0 1142 0 59 328
813444482_T_ONTIME
LISTO
PROMEDIO: 19903.08507
RECUENTO: 445828
UMA: 8873332707
100%

RStudio

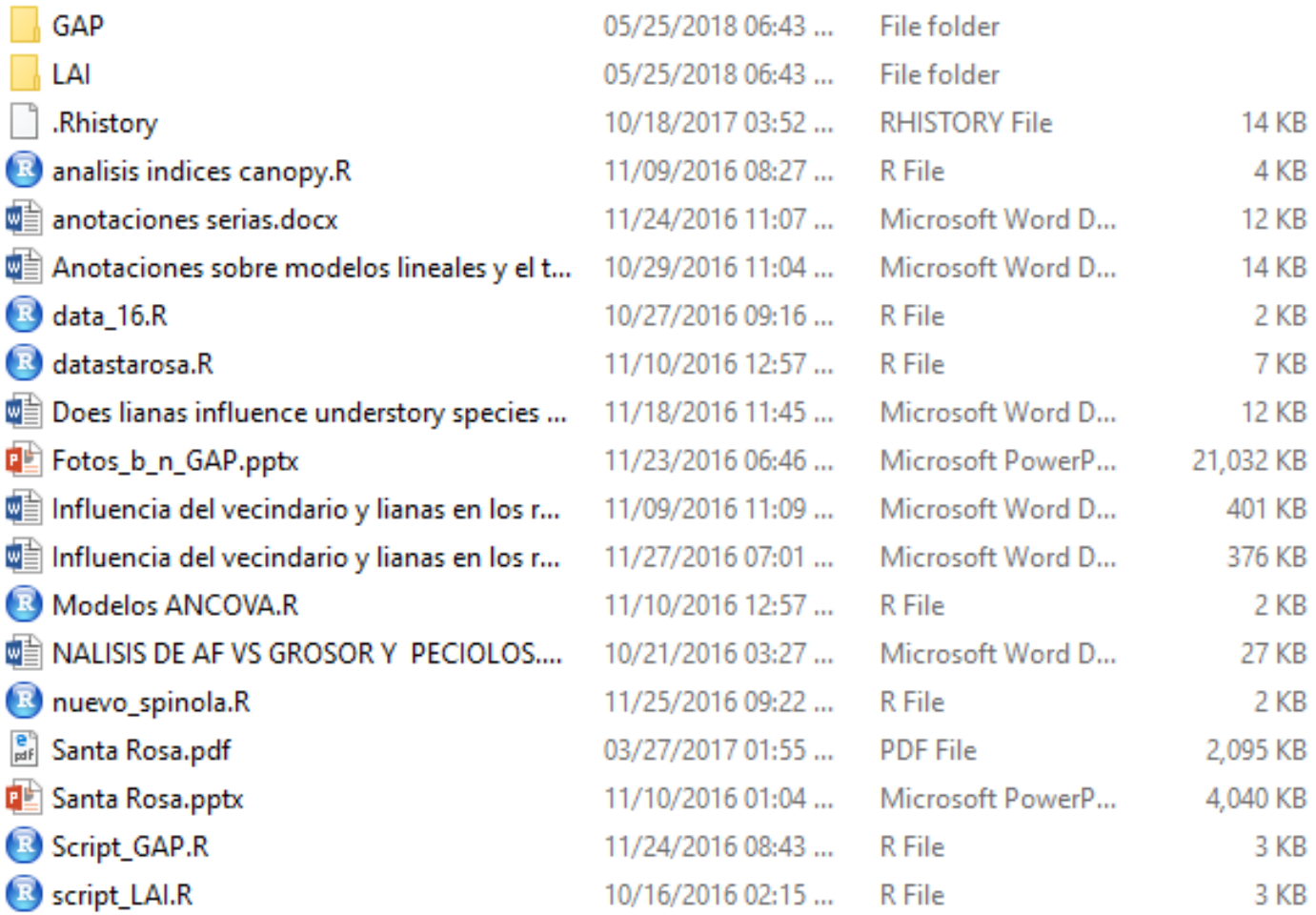


Integrated Development Environment

Starting a project

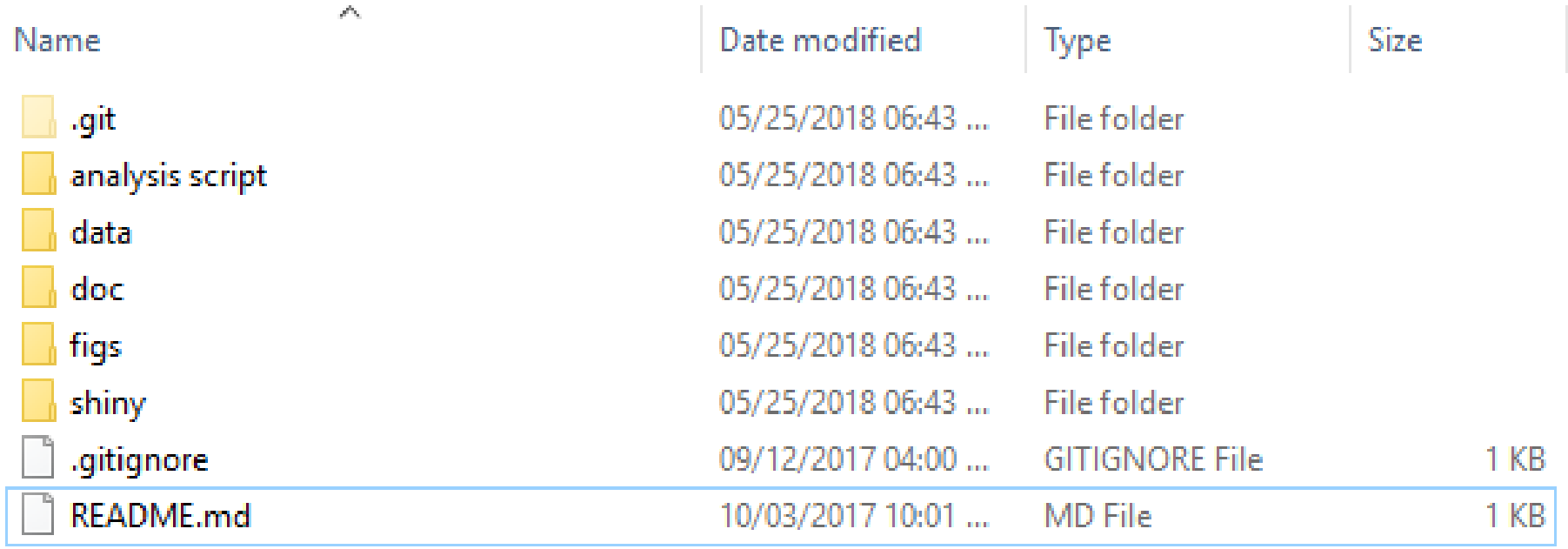
Have you ever felt like this looking at an old analysis?

Bad project structure example:



GAP	05/25/2018 06:43 ...	File folder	
LAI	05/25/2018 06:43 ...	File folder	
.Rhistory	10/18/2017 03:52 ...	RHISTORY File	14 KB
analisis indices canopy.R	11/09/2016 08:27 ...	R File	4 KB
anotaciones serias.docx	11/24/2016 11:07 ...	Microsoft Word D...	12 KB
Anotaciones sobre modelos lineales y el t...	10/29/2016 11:04 ...	Microsoft Word D...	14 KB
data_16.R	10/27/2016 09:16 ...	R File	2 KB
datastarosa.R	11/10/2016 12:57 ...	R File	7 KB
Does lianas influence understory species ...	11/18/2016 11:45 ...	Microsoft Word D...	12 KB
Fotos_b_n_GAP.pptx	11/23/2016 06:46 ...	Microsoft PowerP...	21,032 KB
Influencia del vecindario y lianas en los r...	11/09/2016 11:09 ...	Microsoft Word D...	401 KB
Influencia del vecindario y lianas en los r...	11/27/2016 07:01 ...	Microsoft Word D...	376 KB
Modelos ANCOVA.R	11/10/2016 12:57 ...	R File	2 KB
NALISIS DE AF VS GROSOR Y PECIOLOS....	10/21/2016 03:27 ...	Microsoft Word D...	27 KB
nuevo_spinola.R	11/25/2016 09:22 ...	R File	2 KB
Santa Rosa.pdf	03/27/2017 01:55 ...	PDF File	2,095 KB
Santa Rosa.pptx	11/10/2016 01:04 ...	Microsoft PowerP...	4,040 KB
Script_GAP.R	11/24/2016 08:43 ...	R File	3 KB
script_LAI.R	10/16/2016 02:15 ...	R File	3 KB

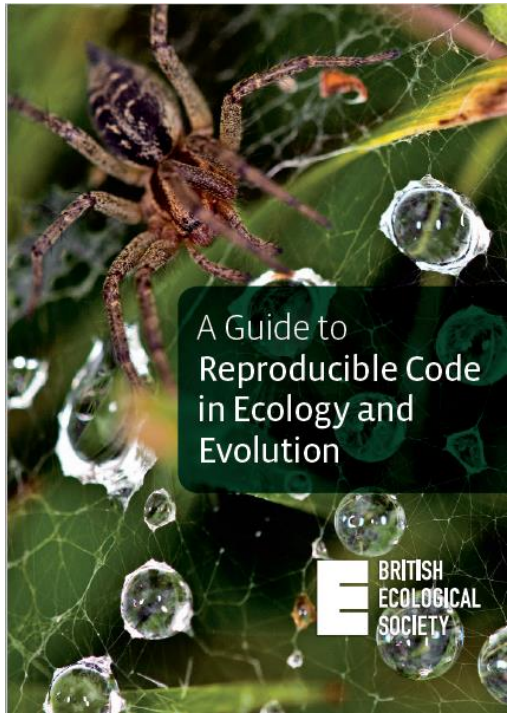
Good project structure example:



Name	Date modified	Type	Size
.git	05/25/2018 06:43 ...	File folder	
analysis script	05/25/2018 06:43 ...	File folder	
data	05/25/2018 06:43 ...	File folder	
doc	05/25/2018 06:43 ...	File folder	
figs	05/25/2018 06:43 ...	File folder	
shiny	05/25/2018 06:43 ...	File folder	
.gitignore	09/12/2017 04:00 ...	GITIGNORE File	1 KB
README.md	10/03/2017 10:01 ...	MD File	1 KB

Here is an example of a basic project directory structure:

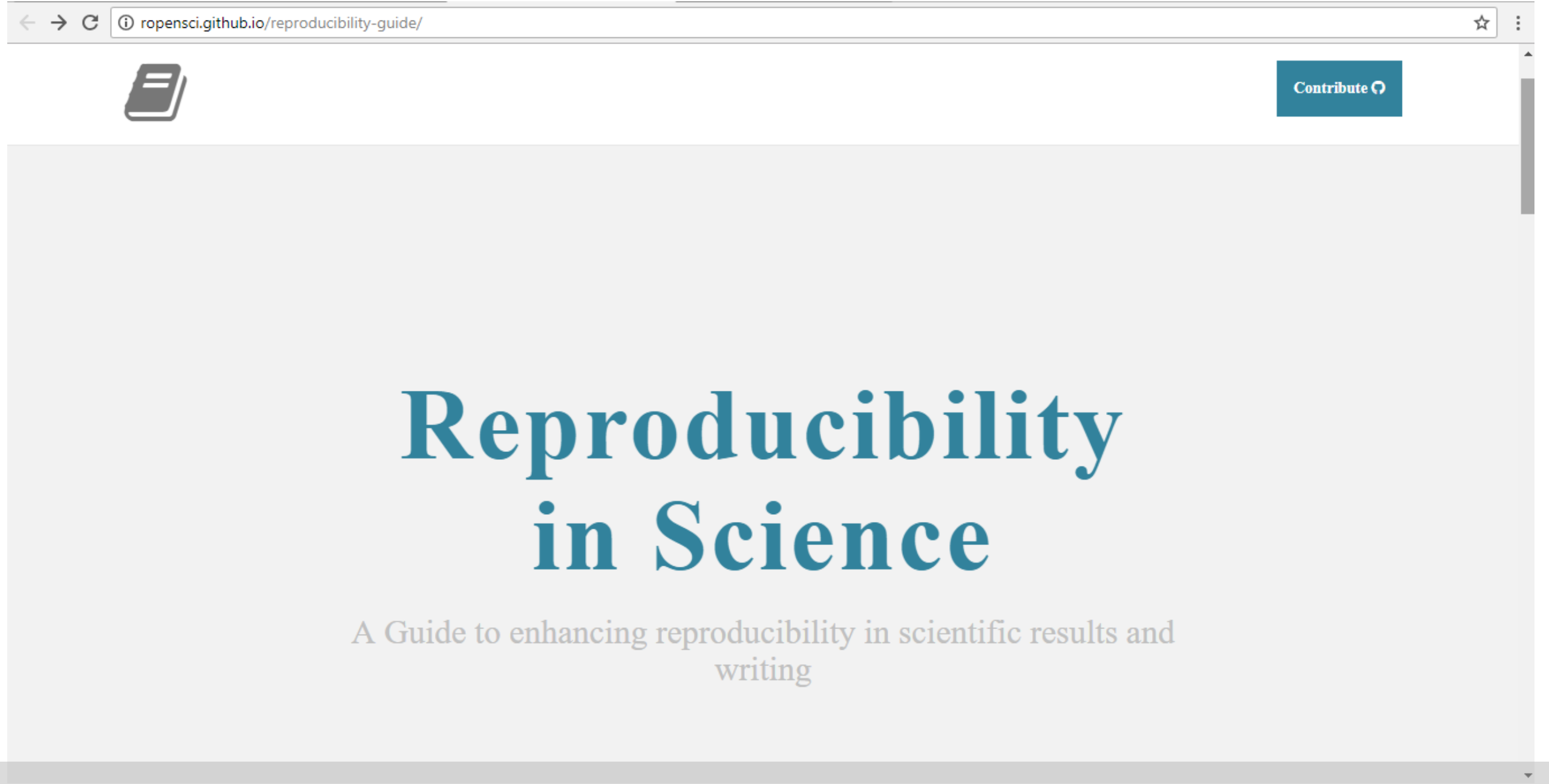
- The `data` folder contains all input data (and metadata) used in the analysis.
- The `doc` folder contains the manuscript.
- The `figs` directory contains figures generated by the analysis.
- The `output` folder contains any type of intermediate or output files (e.g. simulation outputs, models, processed datasets, etc.). You might separate this and also have a `cleaned-data` folder.
- The R directory contains R scripts with function definitions.
- The `reports` folder contains RMarkdown files that document the analysis or report on results.



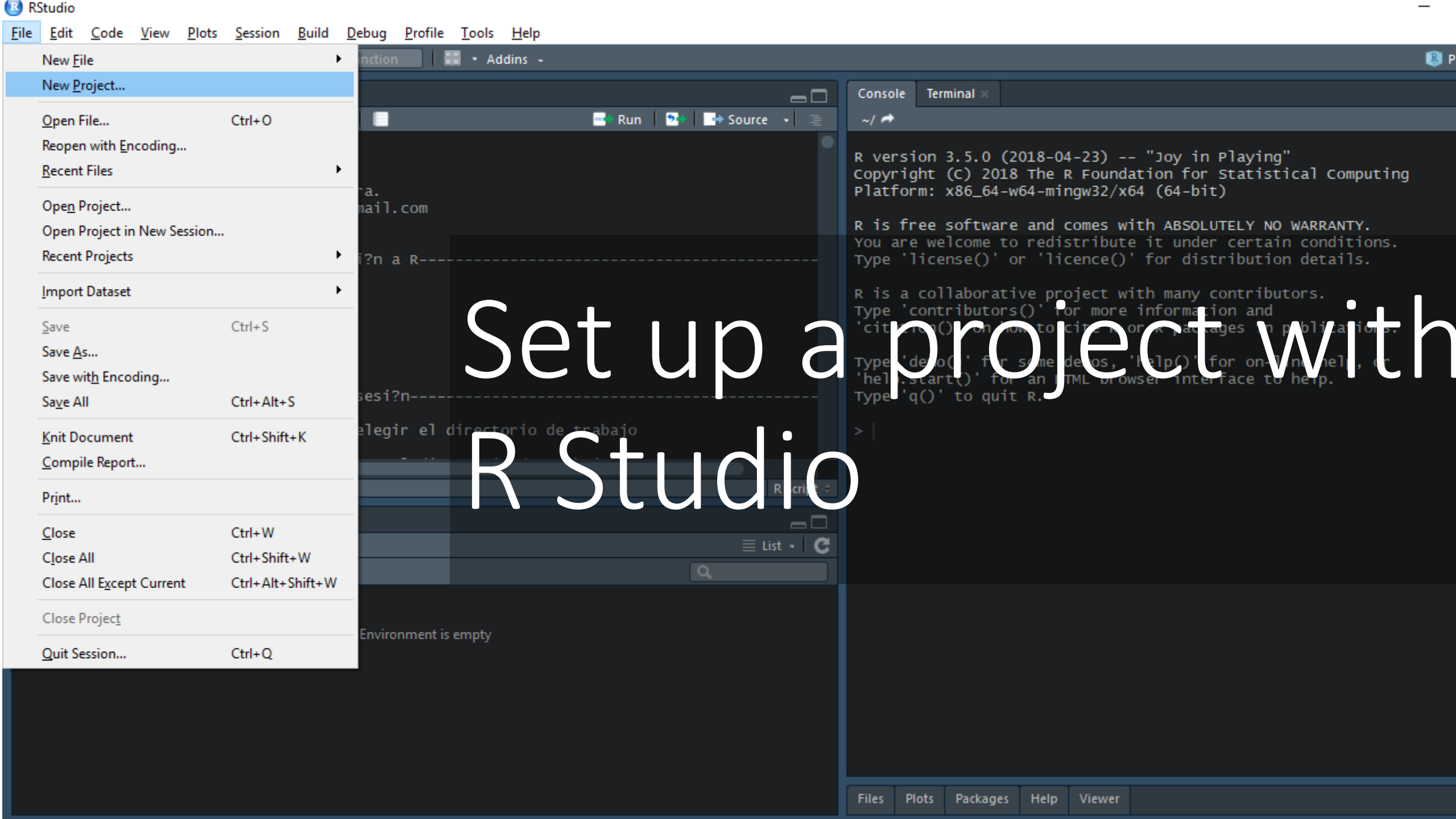
A Guide to Reproducible Code in Ecology and Evolution

Why do all this?

1. Reproducibility
2. Automation
3. Communication

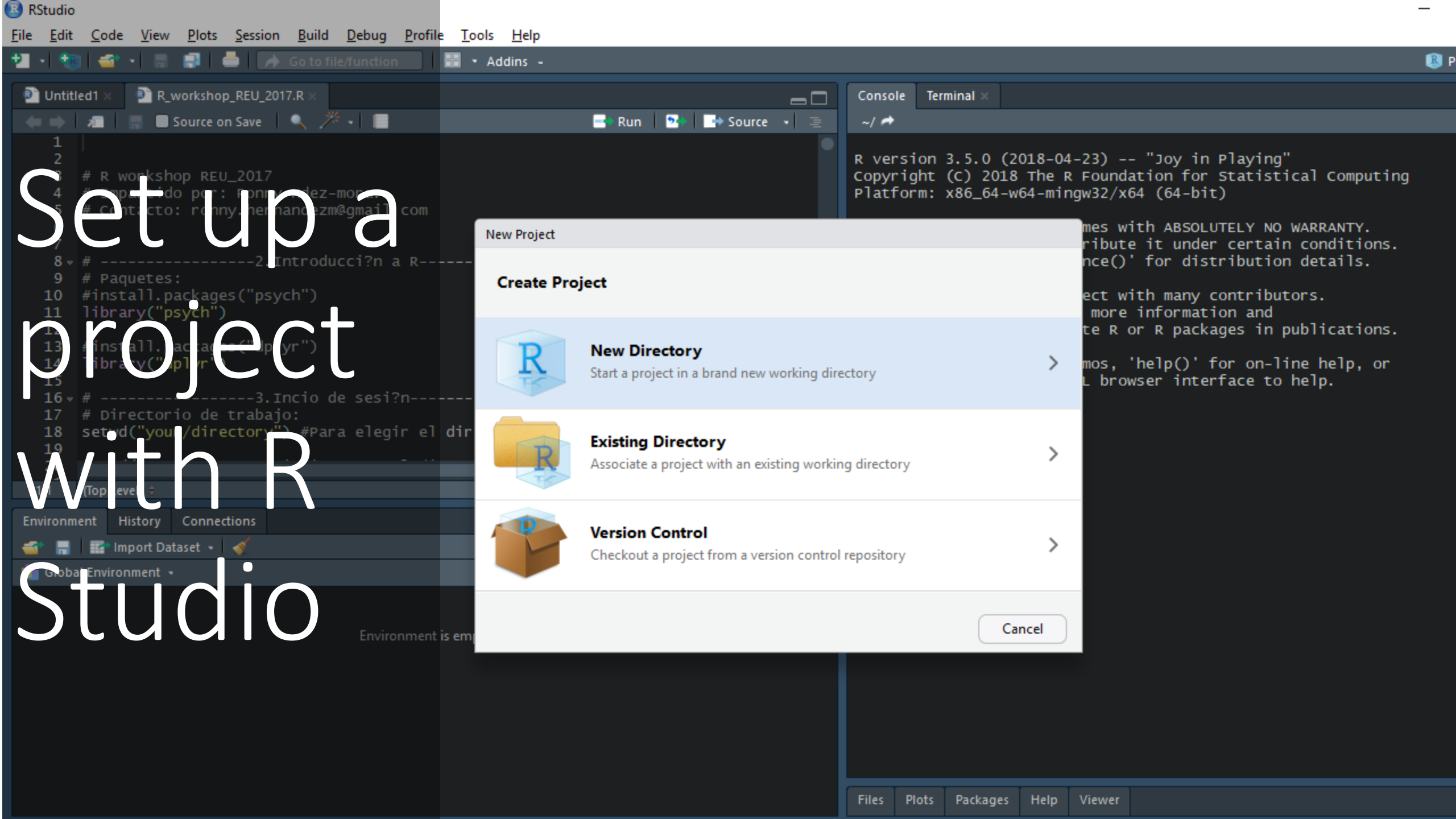


http://ropensci.github.io/reproducibility-guide/

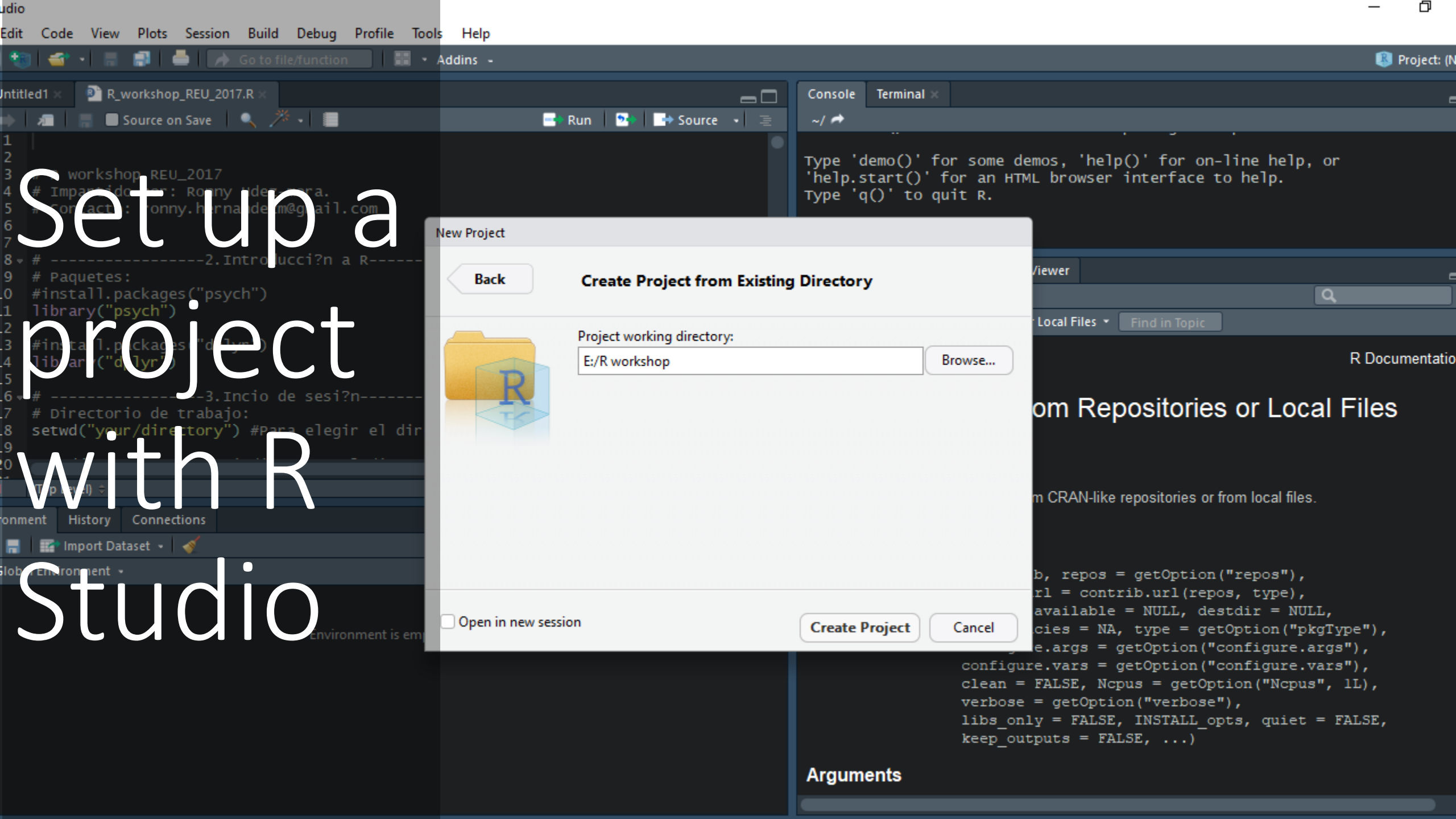


Set up a project with R Studio

Set up a project with R Studio



Set up a project with R Studio



Names:

File names better without spaces: my_super_pro_data.csv
Case sensitivity: My_Super_Pro_data.csv
Not informative names: data_1.csv

Scripts:

Separate scripts: wrangling – analysis - figures

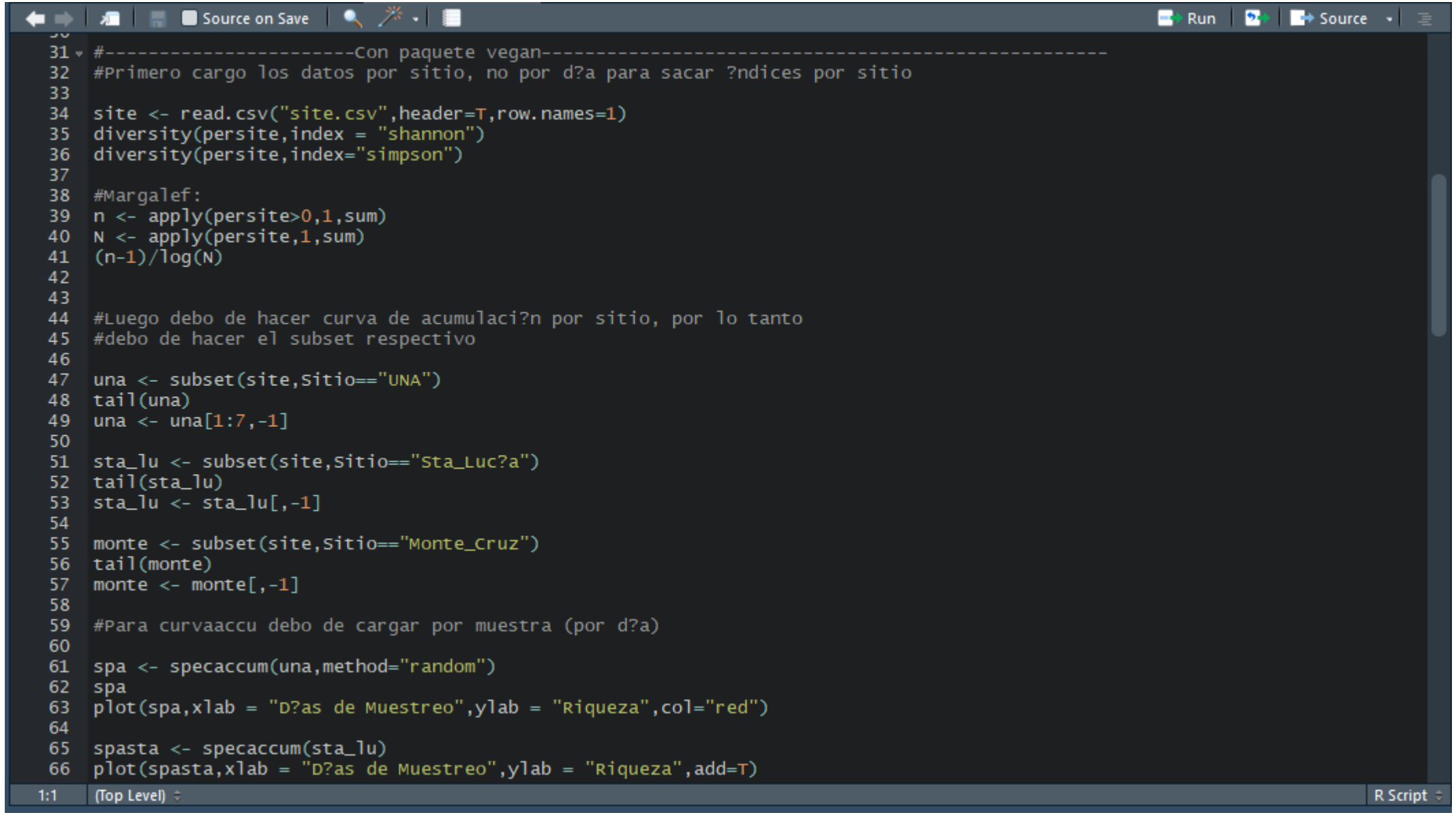
How easy is it to read this?

Source on Save | Run | Source

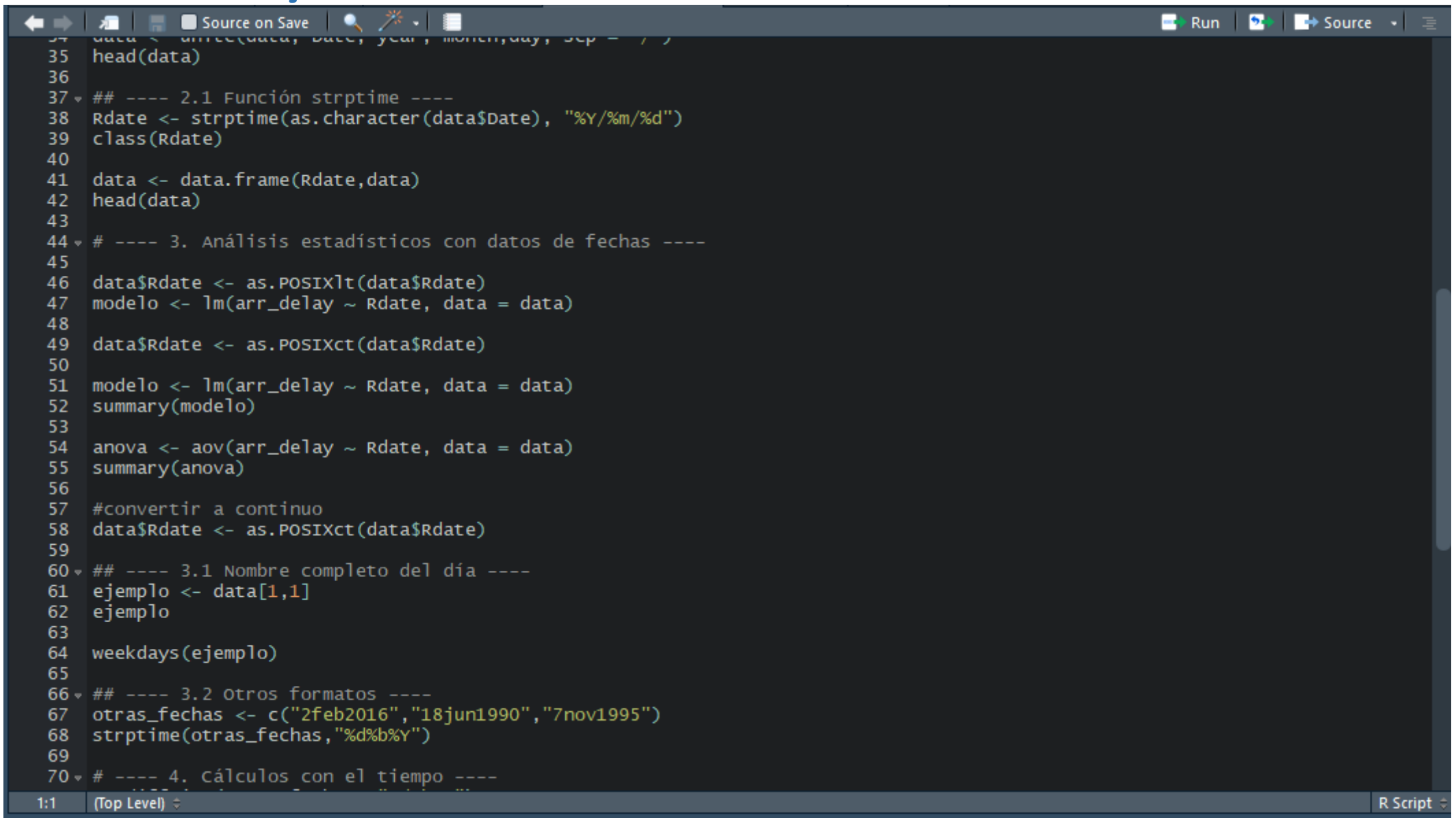
```
eco <- read.csv("~/eco.csv")
View(eco)
class(eco$Temperatura)
cobertura <- subset(eco,Sitio=="cobertura")
library(psych)
desnudo <- subset(eco,Sitio=="sincobertura")
describeBy(desnudo,desnudo$Medici.n)
describeBy(cobertura,cobertura$Medici.n,na.rm=T)
describeBy(eco,eco$Sitio,na.rm=T)
sitio <- c("cobertura","cobertura","cobertura","sin","sin","sin")
lux <- c(8800,13300,12800,71000,104600,118400)
iluminancia <- data.frame(sitio,lux)
humedad <- c(37.1,37.9,38.7,39.9,39,36.8)
hume <- data.frame(sitio,humedad)
iluminancia
class(iluminancia$lux)
describeBy(iluminancia,iluminancia$sitio)
describeBy(hume,hume$sitio)
z <- read.table(file="clipboard",sep="\t",header=T)
plot(x=z$?rea.acumulada,y=z$Spp.acumulada,pch=16,xlab=expression(paste("?rea(",cm^2,")",sep="")),ylab="Especies acumuladas")
lines.default(z$?rea.acumulada,z$Spp.acumulada)
```

1:1 (Top Level) R Script

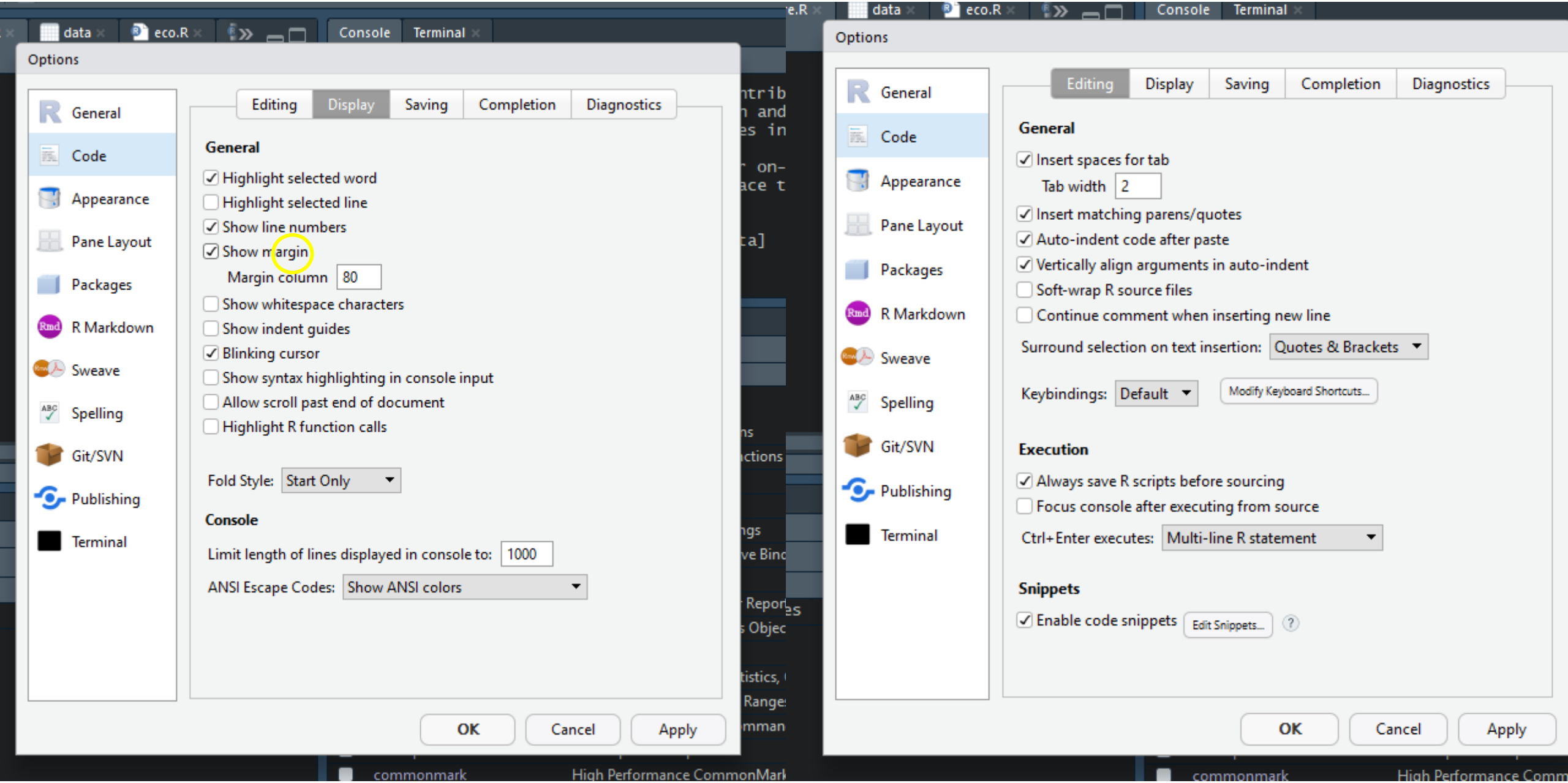
How easy is it to read this?

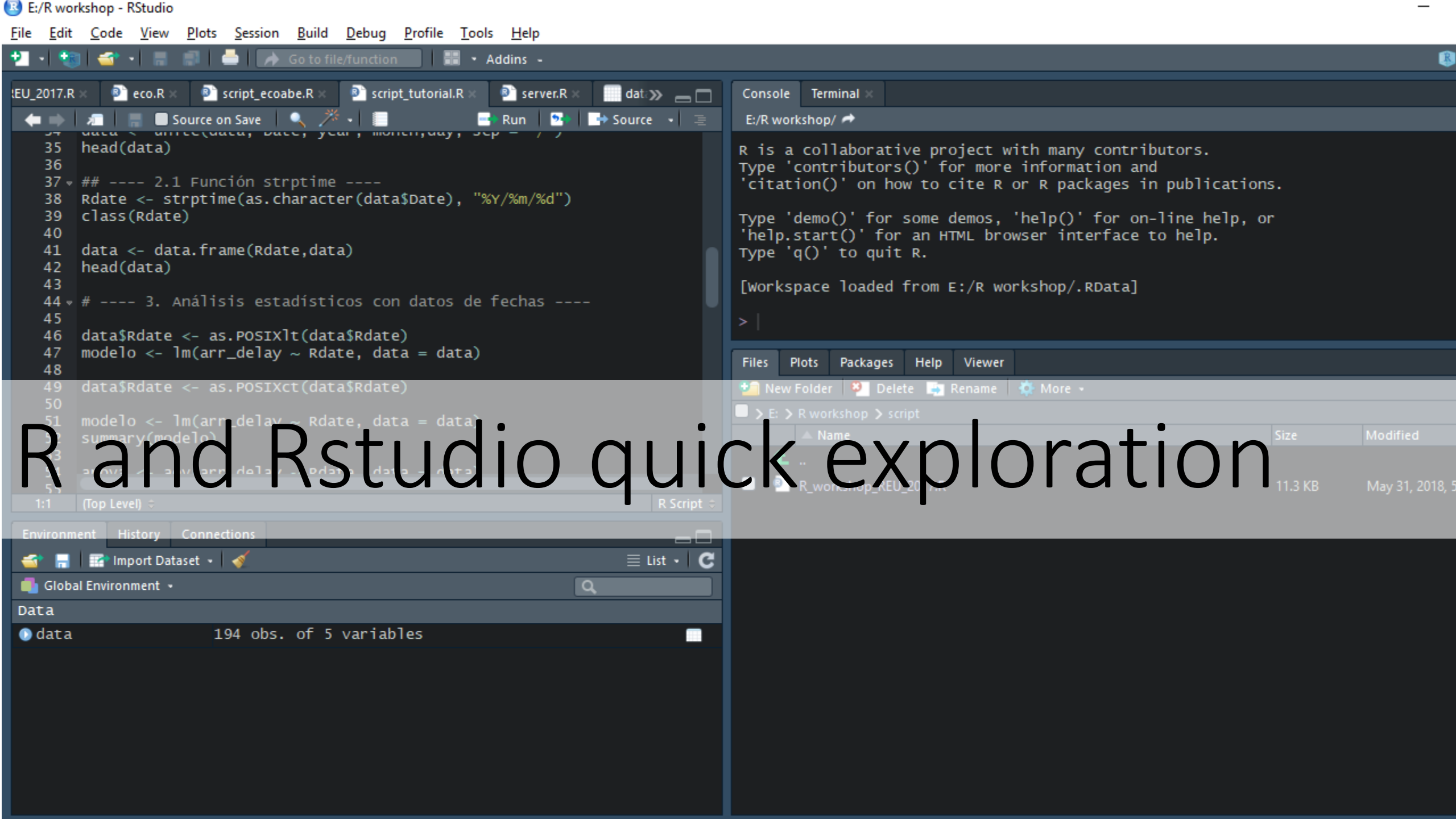


How easy is it to read this?



```
35  head(data)
36
37  ## ---- 2.1 Función strptime ----
38  Rdate <- strptime(as.character(data$Date), "%Y/%m/%d")
39  class(Rdate)
40
41  data <- data.frame(Rdate,data)
42  head(data)
43
44  # ---- 3. Análisis estadísticos con datos de fechas ----
45
46  data$Rdate <- as.POSIXlt(data$Rdate)
47  modelo <- lm(arr_delay ~ Rdate, data = data)
48
49  data$Rdate <- as.POSIXct(data$Rdate)
50
51  modelo <- lm(arr_delay ~ Rdate, data = data)
52  summary(modelo)
53
54  anova <- aov(arr_delay ~ Rdate, data = data)
55  summary(anova)
56
57  #convertir a continuo
58  data$Rdate <- as.POSIXct(data$Rdate)
59
60  ## ---- 3.1 Nombre completo del día ----
61  ejemplo <- data[1,1]
62  ejemplo
63
64  weekdays(ejemplo)
65
66  ## ---- 3.2 Otros formatos ----
67  otras_fechas <- c("2feb2016","18jun1990","7nov1995")
68  strptime(otras_fechas,"%d%b%Y")
69
70  # ---- 4. Cálculos con el tiempo ----
```


data
eco.R
Console
Terminal
Options
General
Code
Appearance
Pane Layout
Packages
R Markdown
Sweave
Spelling
Git/SVN
Publishing
Terminal
Editing
Display
Saving
Completion
Diagnostics
General
Highlight selected word
Highlight selected line
Show line numbers
Show margin
Margin column
80
Show whitespace characters
Show indent guides
Blinking cursor
Show syntax highlighting in console input
Allow scroll past end of document
Highlight R function calls
Fold Style:
Start Only
Console
Limit length of lines displayed in console to:
1000
ANSI Escape Codes:
Show ANSI colors
OK
Cancel
Apply
Insert spaces for tab
Tab width
2
Insert matching parens/quotes
Auto-indent code after paste
Vertically align arguments in auto-indent
Soft-wrap R source files
Continue comment when inserting new line
Surround selection on text insertion:
Quotes & Brackets
Keybindings:
Default
Modify Keyboard Shortcuts...
Execution
Always save R scripts before sourcing
Focus console after executing from source
Ctrl+Enter executes:
Multi-line R statement
Snippets
Enable code snippets
Edit Snippets...



R and Rstudio quick exploration


Data and structures in R

Types of data:

-Characters “biología” , “A” , “estadística”

-Numerics 18.6 , 90

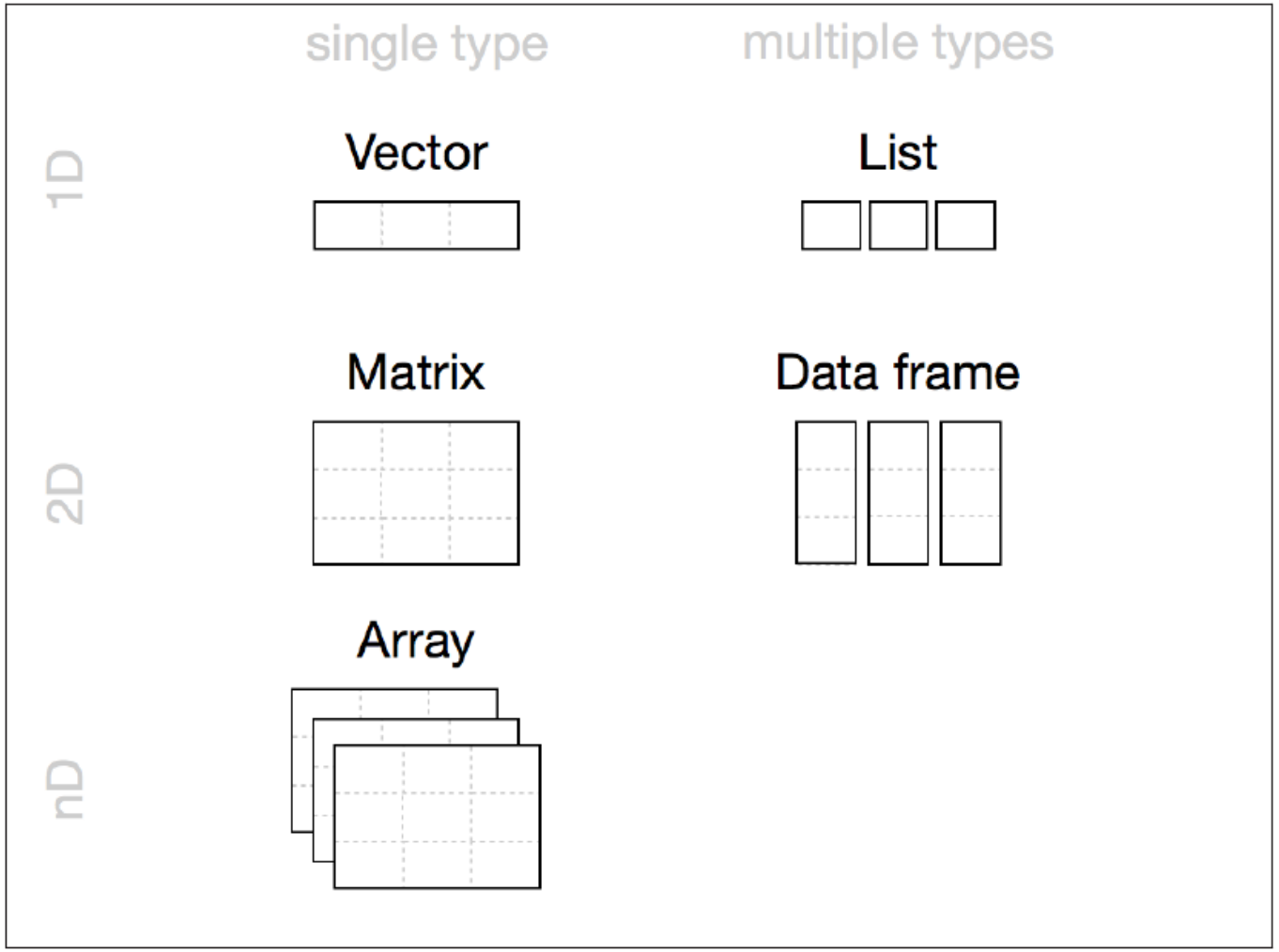
-Integers 18L , 26L

-Complex 2+4i

-Logical TRUE , FALSE

Para verificar su clase: class( )

Objetos se almacenan en estructuras:



>Check_the_code::

Coercing rules

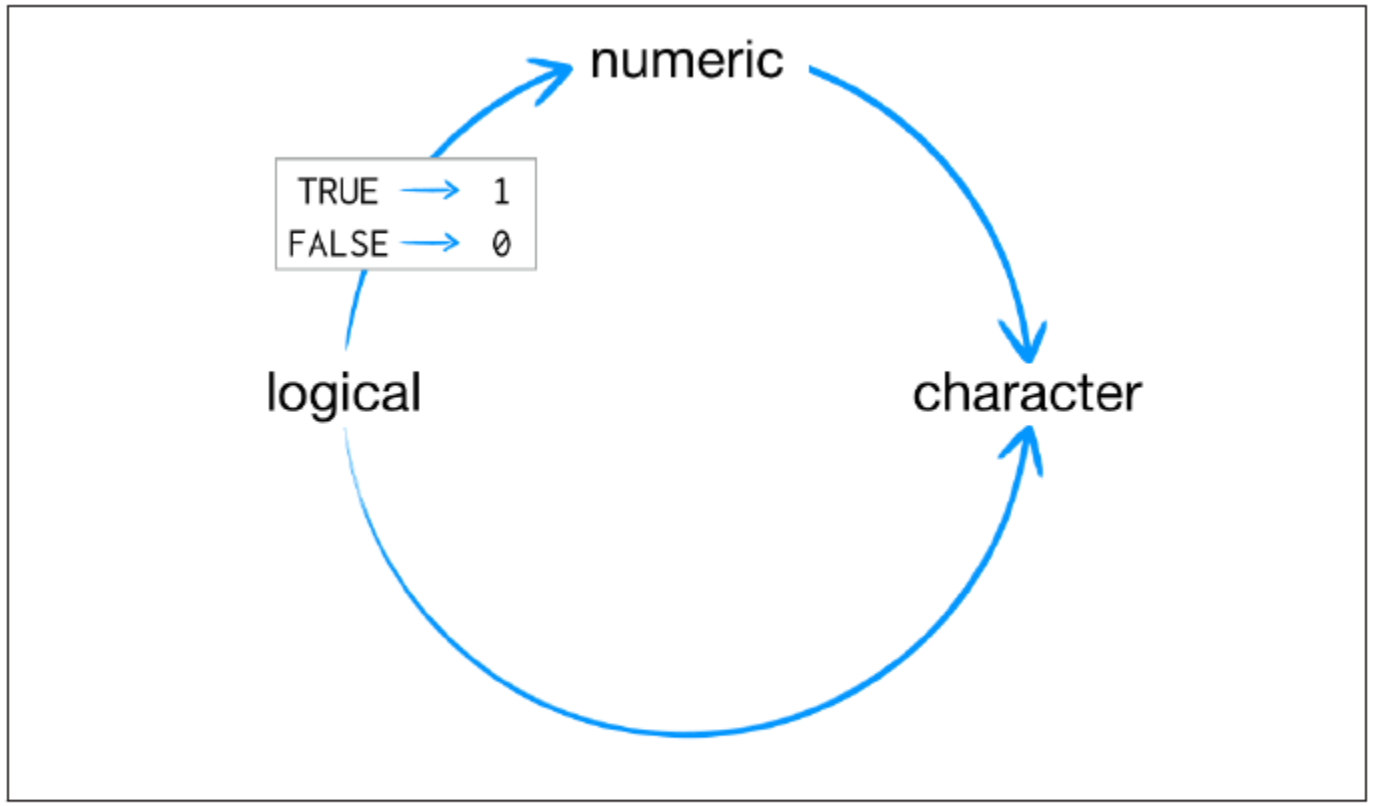
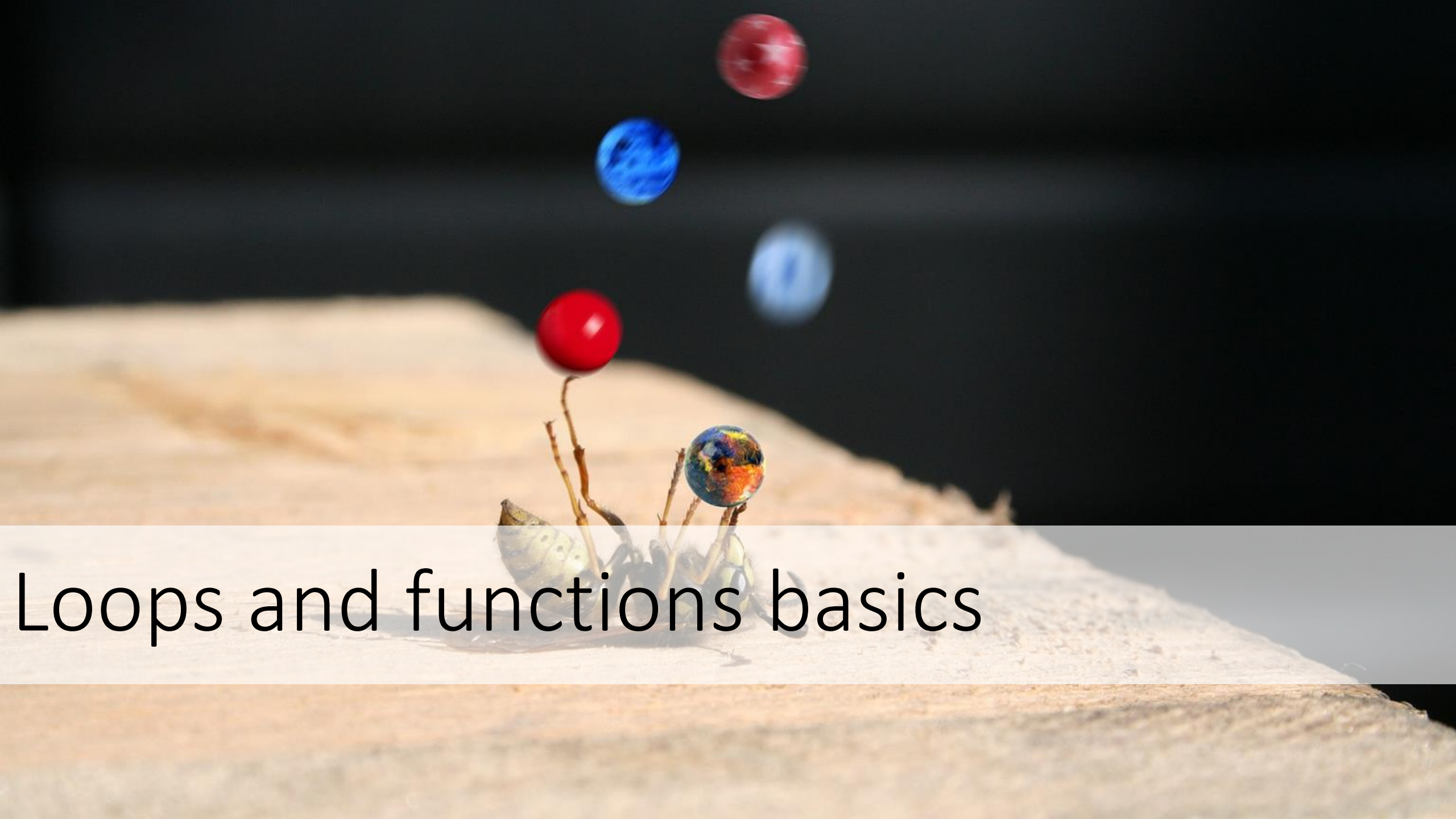


Figure 3-1. R always uses the same rules to coerce data to a single type. If character strings are present, everything will be coerced to a character string. Otherwise, logicals are coerced to numerics.

>Check_the_code::

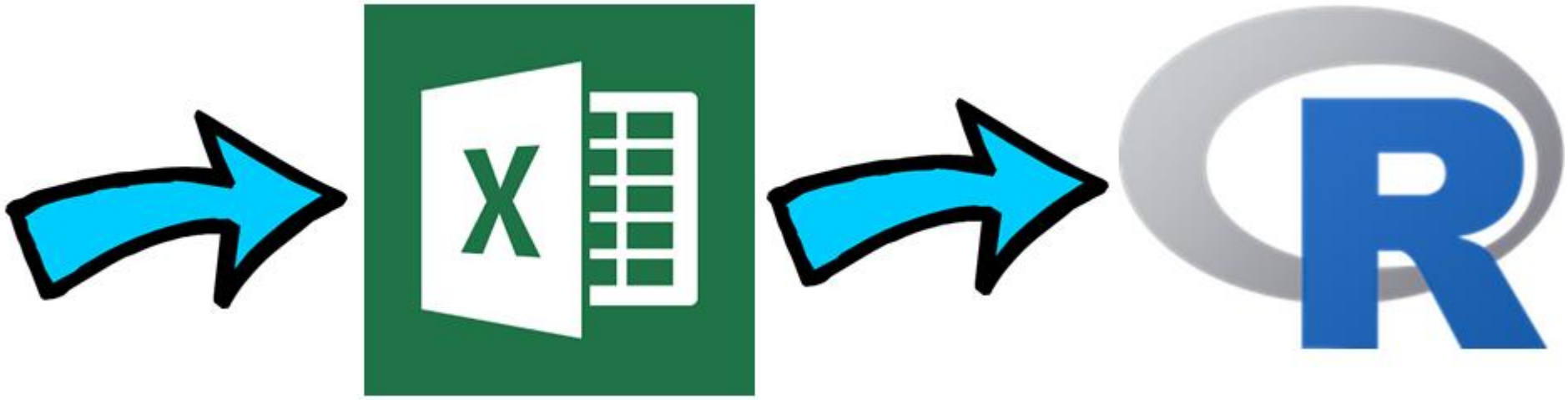


Loops and functions basics

Reduce duplication!

- Code easier to read
- Easier to respond to changes in requirements
- Fewer bugs

Data wrangling



Don’t mix characters and numbers in the same cell

Put easy names to write and remember

Don’t mix lower case and upper case

Experimental design:

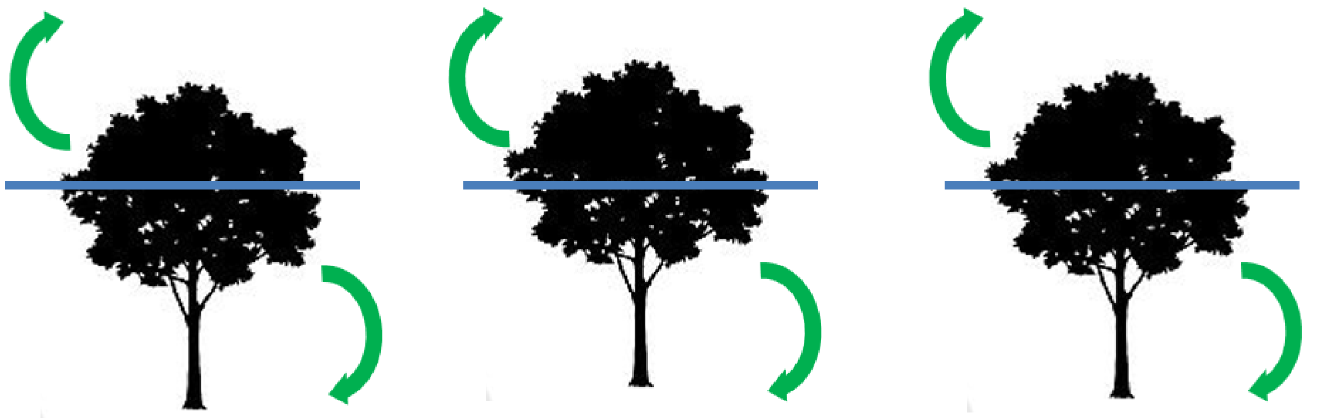
-3 Individuals

-2 measurement per leaf

-5 treatments applied to 2 strata

How should the data set be structured?

10 HD	27 °C	10 HD	27 °C	10 HD	27 °C
10 HD	34°C	10 HD	34°C	10 HD	34°C
10 HD	38 °C	10 HD	38 °C	10 HD	38 °C
10 HD	45 °C	10 HD	45 °C	10 HD	45 °C
10 HD	EH	10 HD	EH	10 HD	EH



10 HS	27 °C	10 HS	27 °C	10 HS	27 °C
10 HS	34°C	10 HS	34°C	10 HS	34°C
10 HS	38 °C	10 HS	38 °C	10 HS	38 °C
10 HS	45 °C	10 HS	45 °C	10 HS	45 °C
10 HS	EH	10 HS	EH	10 HS	EH

DATA_FIXb - Microsoft Excel

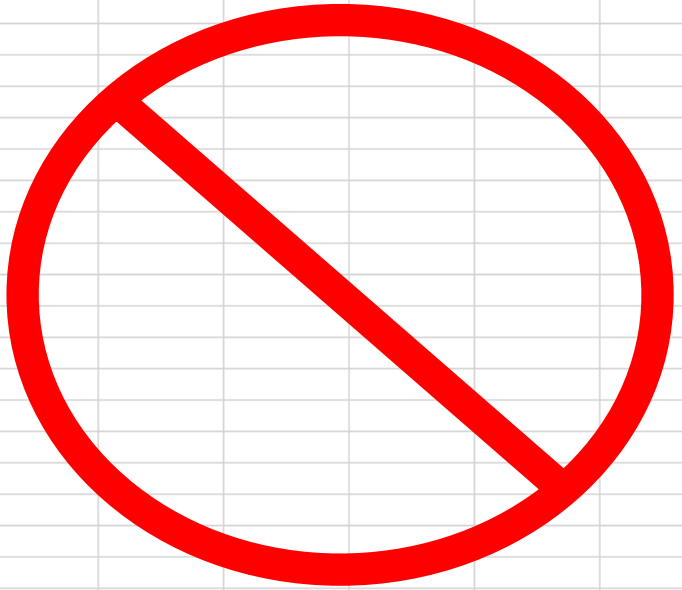
ARCHIVO | INICIO | INSERTAR | DISEÑO DE PÁGINA | FÓRMULAS | DATOS | REVISAR | VISTA | Iniciar sesión

Pegar | Calibri | 11 | Ajustar texto | General | Formato condicional | Dar formato como tabla | Estilos de celda | Insertar | Eliminar | Formato | Ordenar y filtrar | Buscar y seleccionar | Firme y Codifique

Portapapeles | Fuente | Alineación | Número | Estilos | Celdas | Modificar | Privacidad

I19

	A	B	C	D	E	F	G	H	I
1	Inicial					Final			
2	ID	IND	POS	TRT		ID	IND	POS	TRT
3	1	A	D	27		170	B	S	34
4	2	A	D	27		261	C	S	34
5	3	A	D	27		262	C	S	34
6	4	A	D	27		263	C	S	34
7	5	A	D	27		264	C	S	34
8	6	A	D	27		265	C	S	34
9	7	A	D	27		266	C	S	34
10	8	A	D	27		267	C	S	34
11	9	A	D	27		268	C	S	34
12	10	A	D	27		269	C	S	34
13	101	B	D	27		270	C	S	34
14	102	B	D	27		61	A	S	34
15	103	B	D	27					
16	104	B	D	27					
17	105	B	D	27					
18	106	B	D	27					
19	107	B	D	27					
20									
21									
22									
23									



DATA_FIXb | Hoja1

LISTO | 100%

DATA_FIXb - Microsoft Excel

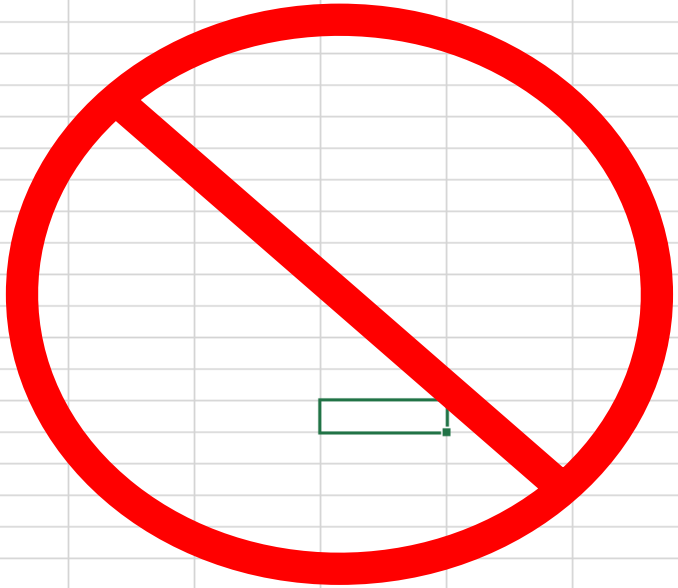
ARCHIVO | INICIO | INSERTAR | DISEÑO DE PÁGINA | FÓRMULAS | DATOS | REVISAR | VISTA | Iniciar sesión

Pegar | Calibri | 11 | Ajustar texto | General | Combinar y centrar | Formato condicional | Dar formato como tabla | Estilos de celda | Insertar | Eliminar | Formato | Ordenar y filtrar | Buscar y seleccionar | Firme y Codifique

Portapapeles | Fuente | Alineación | Número | Estilos | Celdas | Modificar | Privacidad

M15

	A	B	C	D
1	Inicial			
2	ID	IND	POS	TRT
3	1	A	D	27
4	2	A	D	27
5	3	A	D	27
6	4	A	D	27
7	5	A	D	27
8	6	A	D	27
9	7	A	D	27
10	8	A	D	27
11	9	A	D	27
12	10	A	D	27
13	101	B	D	27
14	102	B	D	27
15	103	B	D	27
16	104	B	D	27
17	105	B	D	27
18	106	B	D	27
19	107	B	D	27
20	Final			
21	ID	IND	POS	TRT
22	170	B	S	34
23	261	C	S	34



DATA_FIXb | Hoja1

LISTO | 100%

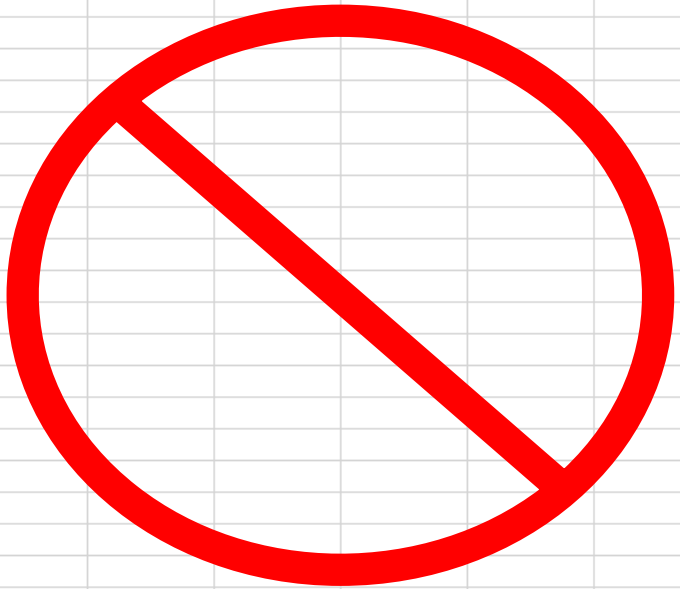
DATA_FIXb - Microsoft Excel

ARCHIVO | INICIO | INSERTAR | DISEÑO DE PÁGINA | FÓRMULAS | DATOS | REVISAR | VISTA | Iniciar sesión

Pegar | Portapapeles | Calibri | 11 | Fuente | Ajustar texto | Combinar y centrar | Alineación | General | Número | Formato condicional | Dar formato como tabla | Estilos de celda | Estilos | Insertar | Eliminar | Formato | Celdas | Ordenar y filtrar | Buscar y seleccionar | Modificar | Firme y Codifique | Privacidad

A1 | Final

	A	B	C	D
1	Final			
2	IND	B	C	A
3	POS	S	D	S
4	TRT	34	34	34
5	FV	X	X	X



DATA_FIXb | Hoja1

LISTO | 100%

Tidy data!

ID	IND	POS	TRT	Medicion	FV
1	A	D	27	F	
2	A	S	27	F	
3	B	D	34	I	
4	B	S	34	I	
5	C	D	38	F	

Observation

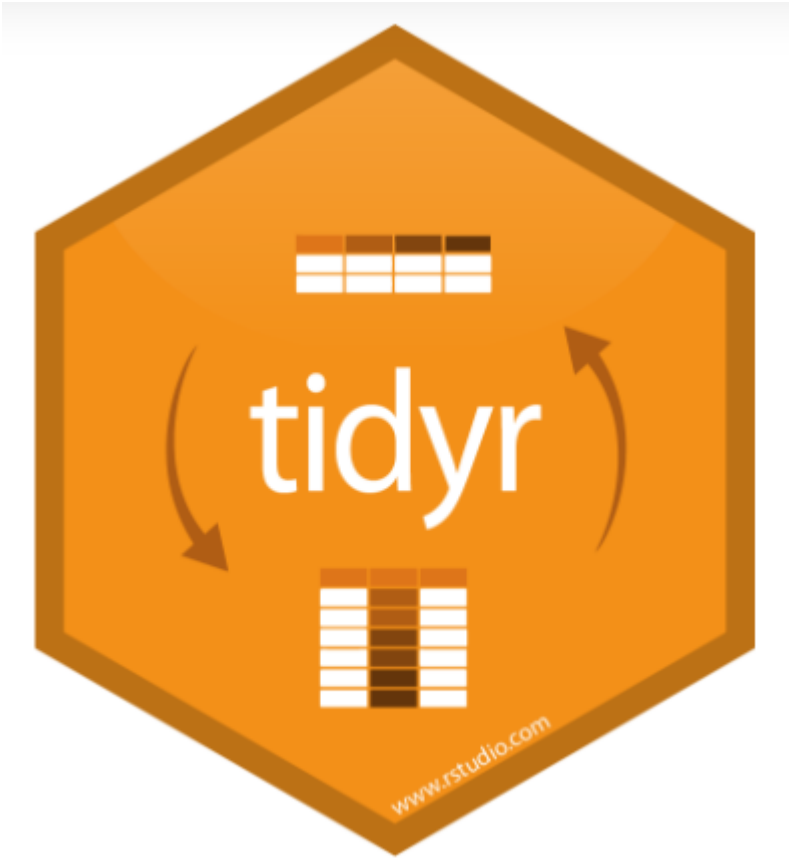
Variable / Attribute

-Each variable must have its own column.
-Each observation must have its own row.
-Each value must have its own cell.

Wickham, H. (2014). Tidy Data. *Journal of Statistical Software*, *59*(i10).

Four verbs in tidyr:

1- gather ( )
2- spread ( )
3- separate( )
4- unite ( )



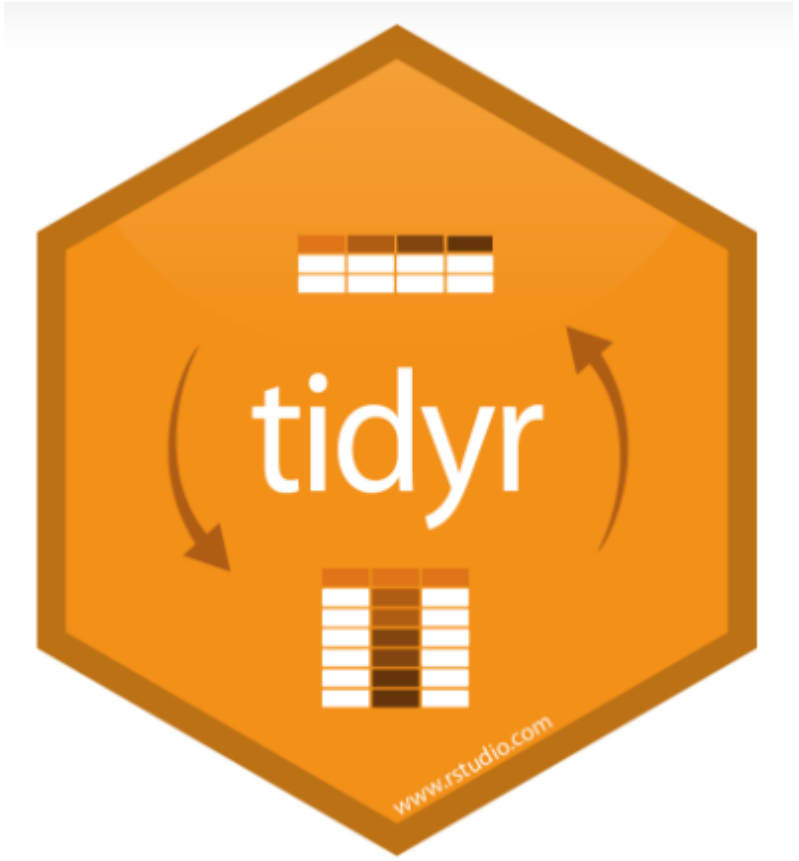
How to deal with not tidy data?

What is the problem with this dataset?

data

Filter

	Country	Y1980	Y1981	Y1982	Y1983
1	Afghanistan	21.48678	21.46552	21.45145	21.43822
2	Albania	25.22533	25.23981	25.25636	25.27176
3	Algeria	22.25703	22.34745	22.43647	22.52105
4	Andorra	25.66652	25.70868	25.74681	25.78250
5	Angola	20.94876	20.94371	20.93754	20.93187
6	Antigua and Barbuda	23.31424	23.39054	23.45883	23.53735
7	Argentina	25.37913	25.44951	25.50242	25.55644
8	Armenia	23.82469	23.86401	23.91023	23.95649
9	Australia	24.92729	25.00216	25.07660	25.14938
10	Austria	24.84097	24.88110	24.93482	24.98118
11	Azerbaijan	24.49375	24.52584	24.56064	24.60150



How to deal with not tidy data?

What is the problem with this set of data?

data_fix

Filter

	Country	year	bmi
1	Afghanistan	Y1980	21.48678
2	Albania	Y1980	25.22533
3	Algeria	Y1980	22.25703
4	Andorra	Y1980	25.66652
5	Angola	Y1980	20.94876
6	Antigua and Barbuda	Y1980	23.31424
7	Argentina	Y1980	25.37913
8	Armenia	Y1980	23.82469

`gather(data, key, value, …)`

`data`: a data frame

`key`: bare name of new key column

`value`: bare name of new value column

`…`: bare names of columns to gather (or not)

>Check_the_code::

tidyr

How to deal with not tidy data?

data_unfix

Filter

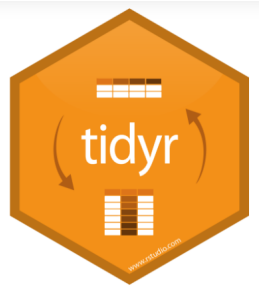
	Country	Y1980	Y1981	Y1982	Y1983
1	Afghanistan	21.48678	21.46552	21.45145	21.43822
2	Albania	25.22533	25.23981	25.25636	25.27176
3	Algeria	22.25703	22.34745	22.43647	22.52105
4	Andorra	25.66652	25.70868	25.74681	25.78250
5	Angola	20.94876	20.94371	20.93754	20.93187
6	Antigua and Barbuda	23.31424	23.39054	23.45883	23.53735
7	Argentina	25.37913	25.44951	25.50242	25.55644
8	Armenia	23.82469	23.86401	23.91023	23.95649
9	Australia	24.92720	25.00216	25.07660	25.14038

```
spread(data, key, value)
```

data: a data frame

key: bare name of column containing keys

value: bare name of column containing values



How to deal with not tidy data?

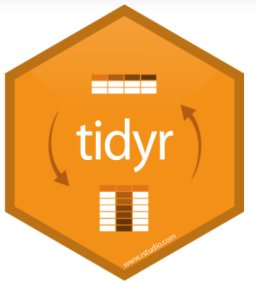
How to solve this?

fotos_sep

Filter

	i	plot	MIG.grey	Aniso.grey	Timestamp	Date	Time	sitio	camara
1	1	LADERA	0.2220831	0.9363847	21/06/2016 11:47	2015:05:21	13:21:50	Piro	1
2	2	LADERA	0.2334211	0.9371384	21/06/2016 11:47	2015:05:21	14:21:50	Piro	1
3	3	LADERA	0.1685896	0.9400231	21/06/2016 11:48	2015:05:22	09:00:01	Piro	1
4	4	LADERA	0.1554022	0.9261470	21/06/2016 11:48	2015:05:22	10:00:01	Piro	1
5	5	LADERA	0.2006901	0.9033331	21/06/2016 11:48	2015:05:22	11:00:01	Piro	1

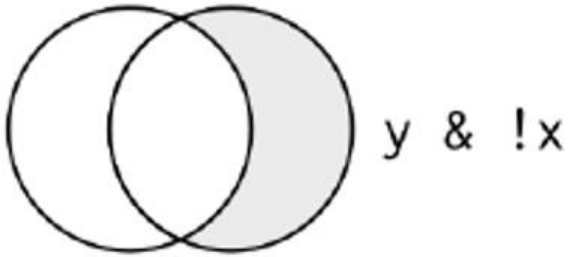
>Check_the_code::



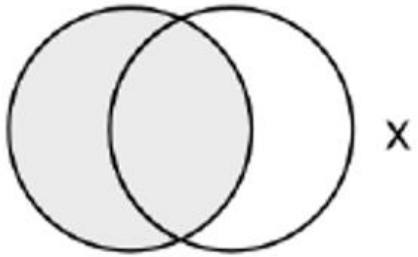
Five verbs in dplyr:

1- select ( )
2- filter ( )
3- arrange ( )
4- mutate ( )
5- summarise ( )
    *- group_by ( )

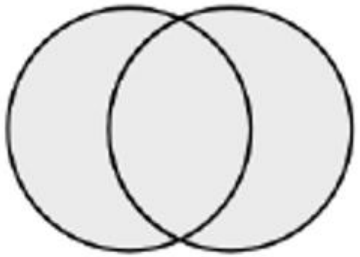
Logical operators:



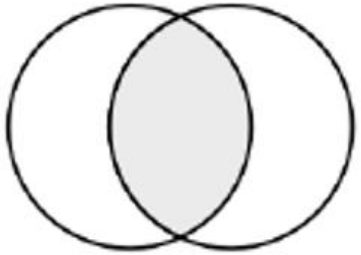
y & !x



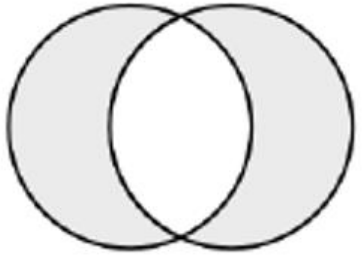
x



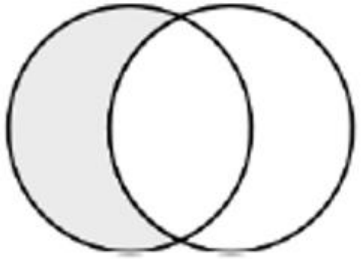
x | y



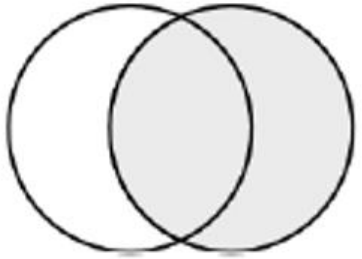
x & y



xor(x, y)



x & !y



y

>Check_the_code::

ggplot2:

```
ggplot(data = <DATA>) +
        <GEOM_FUNCTION> (
                mapping = aes(<MAPPINGS>),
                stat = <STAT>,
                position = <POSITION>
        ) +
        <COORDINATE_FUNCTION> +
        <FACET_FUNCTION>
```

>Check_the_code::

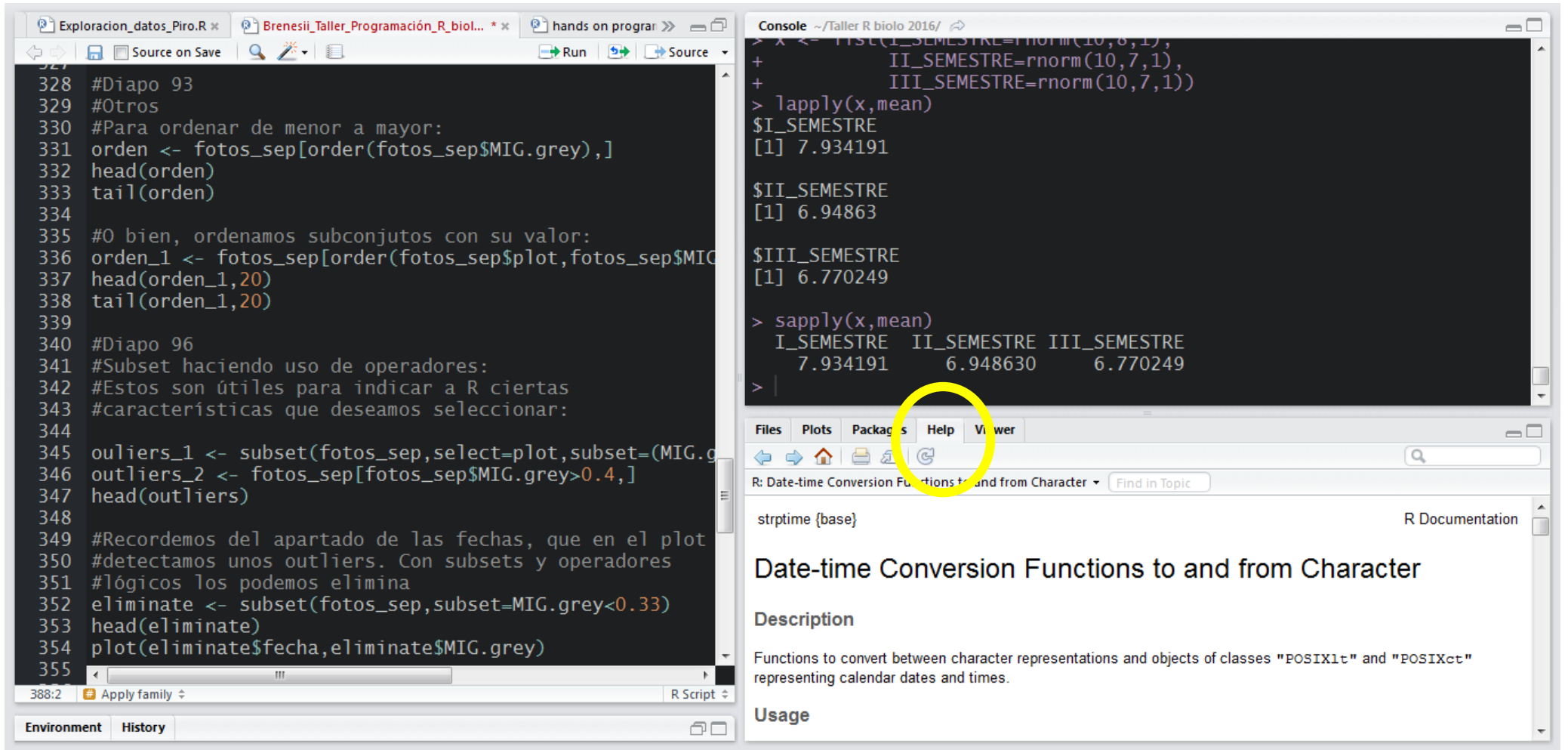
Save R objects for the next session:

```
save(my_data, file = “my_data.Rdata)

saveRDS(mtcars, “my_data.rds”)

readRDS(file = “my_data.rds”)
```

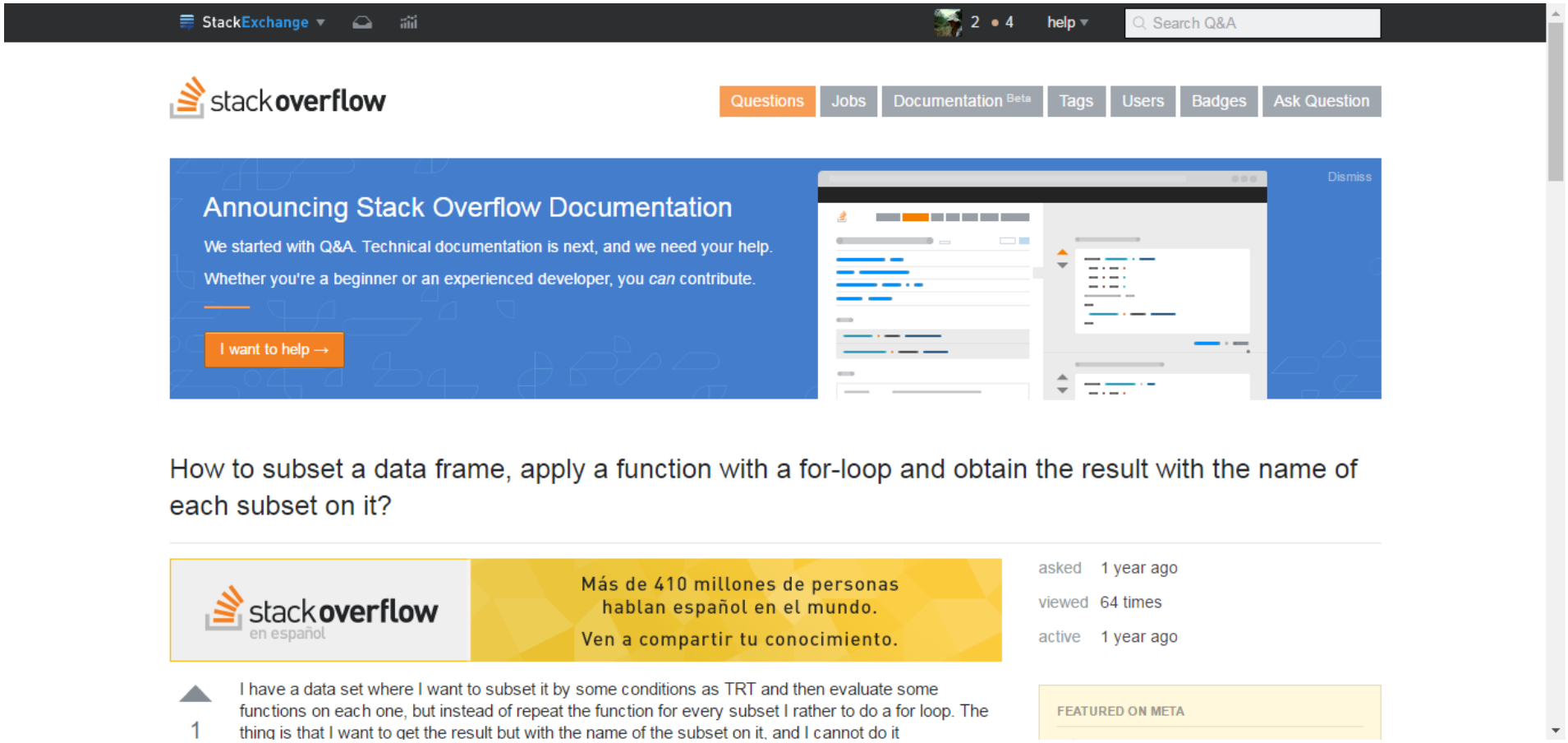
Repositories:

- CRAN, http://cran.r-project.org
- Bitbucket, https://bitbucket.org
- Bioconductor, http://www.bioconductor.org
- GitHub, https://github.com
- Gitorious, http://www.gitorious.com

Help:


Exploracion_datos_Piro.R
Brenesii_Taller_Programación_R_biol...
hands on prograr
Source on Save
Run
Source
328 #Diapo 93
329 #Otros
330 #Para ordenar de menor a mayor:
331 orden <- fotos_sep[order(fotos_sep$MIG.grey),]
332 head(orden)
333 tail(orden)
334
335 #O bien, ordenamos subconjutos con su valor:
336 orden_1 <- fotos_sep[order(fotos_sep$plot,fotos_sep$MIC
337 head(orden_1,20)
338 tail(orden_1,20)
339
340 #Diapo 96
341 #Subset haciendo uso de operadores:
342 #Estos son útiles para indicar a R ciertas
343 #características que deseamos seleccionar:
344
345 ouliers_1 <- subset(fotos_sep,select=plot,subset=(MIG.g
346 outliers_2 <- fotos_sep[fotos_sep$MIG.grey>0.4,]
347 head(outliers)
348
349 #Recordemos del apartado de las fechas, que en el plot
350 #detectamos unos outliers. Con subsets y operadores
351 #lógicos los podemos elimina
352 eliminate <- subset(fotos_sep,subset=MIG.grey<0.33)
353 head(eliminate)
354 plot(eliminate$fecha,eliminate$MIG.grey)
355
388:2
Apply family
R Script
Environment
History
Console ~/Taller R biolo 2016/
+ II_SEMESTRE=rnorm(10,7,1),
+ III_SEMESTRE=rnorm(10,7,1))
> lapply(x,mean)
$I_SEMESTRE
[1] 7.934191
$II_SEMESTRE
[1] 6.94863
$III_SEMESTRE
[1] 6.770249
> sapply(x,mean)
I_SEMESTRE II_SEMESTRE III_SEMESTRE
7.934191 6.948630 6.770249
>
Files
Plots
Packages
Help
Viewer
R: Date-time Conversion Functions to and from Character
Find in Topic
strptime {base}
R Documentation
Date-time Conversion Functions to and from Character
Description
Functions to convert between character representations and objects of classes "POSIXlt" and "POSIXct" representing calendar dates and times.
Usage

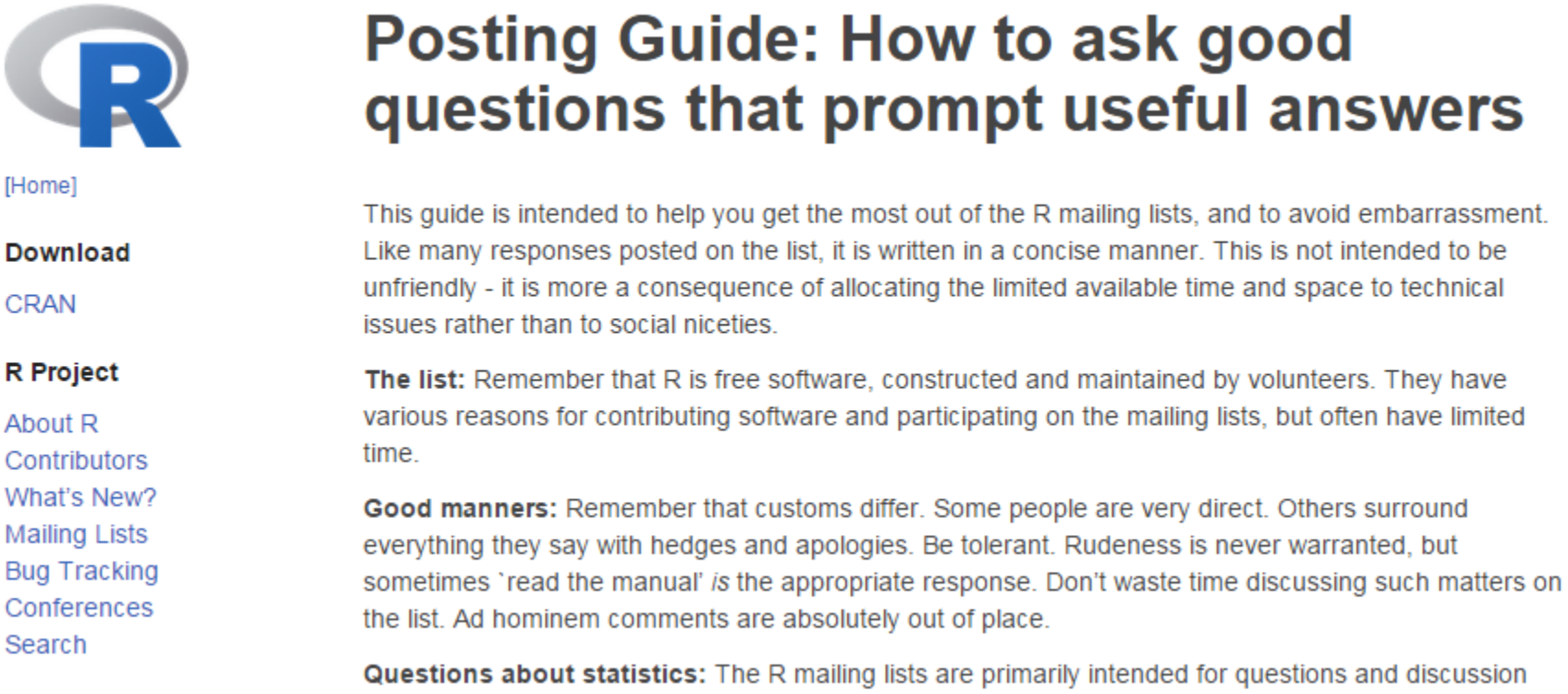
Help:



- http://stackoverflow.com

Help:

- Recomendaciones de foros: http://www.r-project.org/posting-guide.html
  - No hacer preguntas repetidas
  - A buenas preguntas, buenas respuestas



[Home]

Download

CRAN

R Project

About R
Contributors
What's New?
Mailing Lists
Bug Tracking
Conferences
Search

Posting Guide: How to ask good questions that prompt useful answers

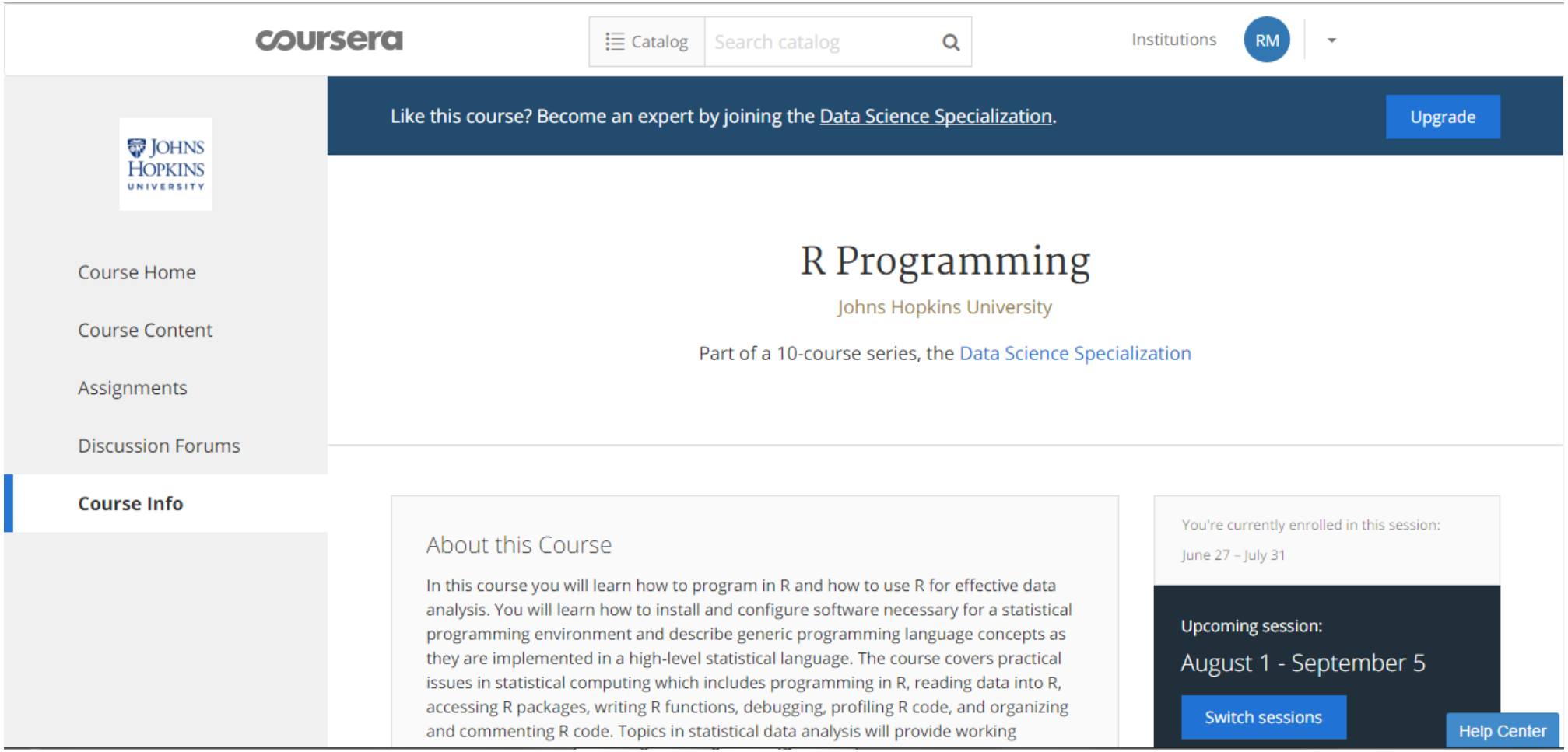
This guide is intended to help you get the most out of the R mailing lists, and to avoid embarrassment. Like many responses posted on the list, it is written in a concise manner. This is not intended to be unfriendly - it is more a consequence of allocating the limited available time and space to technical issues rather than to social niceties.

The list: Remember that R is free software, constructed and maintained by volunteers. They have various reasons for contributing software and participating on the mailing lists, but often have limited time.

Good manners: Remember that customs differ. Some people are very direct. Others surround everything they say with hedges and apologies. Be tolerant. Rudeness is never warranted, but sometimes `read the manual' *is* the appropriate response. Don't waste time discussing such matters on the list. Ad hominem comments are absolutely out of place.

Questions about statistics: The R mailing lists are primarily intended for questions and discussion

Learning resources:



- https://www.coursera.org

Learning resources:

- http://swirlstats.com/

References:

- **Hands on programming with R**
- Teetor, P. (2011). ***R cookbook***. " O'Reilly Media, Inc.".
- Adler, J. (2010). ***R in a nutshell: A desktop quick reference***. " O'Reilly Media, Inc.".
- R programming
- The R book
- R for Data Science