

R-reproducible workflows

1-day workshop
Morning practical session



Brendan Palmer,
Statistics & Data Analysis Unit,
Clinical Research Facility - Cork

A: Is too much choice good or bad?

Blue Horizon SW 6497	Sky High SW 6504	Snowdrop SW 6511	Ski Slope SW 6518	Rarified Air SW 6525
Byte Blue SW 6498	Atmospheric SW 6505	Balmy SW 6512	Hinting Blue SW 6519	Icelandic SW 6526
Stream SW 6499	Vast Sky SW 6506	Take Five SW 6513	Honest Blue SW 6520	Blissful Blue SW 6527
Open Seas SW 6500	Resolute Blue SW 6507	Respite SW 6514	Notable Hue SW 6521	Cosmos SW 6528
Manitou Blue SW 6501	Secure Blue SW 6508	Leisure Blue SW 6515	Sporty Blue SW 6522	Scanda SW 6529
Loch Blue SW 6502	Georgian Bay SW 6509	Down Pour SW 6516	Denim SW 6523	Revel Blue SW 6530
Bosporus SW 6503	Loyal Blue SW 6510	Regatta SW 6517	Cammodore SW 6524	Indigo SW 6531

Inconsistent function names, inconsistent syntax

- R is a very versatile language
 - Sometimes it can be too versatile
 - Do you want to use...

`row.names` or `rownames`

`rowSums` or `rowsum`

`Sys.time`, `system.time`

- Should it be written as...

`newobject` or `new.Object`

`x = 5` or `x <- 5`

`mapping=aes(x,y)` or `mapping = aes(x, y)`

Variable selection

```
summary(starwars$name)
```

```
summary(starwars$"name")
```

```
summary(starwars["name"])
```

```
summary(starwars[, "name"])
```

```
summary(starwars[1])
```

```
summary(starwars[, 1])
```

```
summary(starwars[[1]])
```

- Open the script `am_too_much.choice.R`

Motivation to move on from poorly written code

```
am_bad_habits.R x
Source on Save
21 sites1<-as.list(unique(RL6.7$Var1))
22 sites2<-as.list(unique(RL6.7$Var2))
23
24 sites<-as.data.frame(t(merge(sites1,sites2)))
25 colnames(sites)[1]<-"Position"
26
27 for(i in 1:nrow(sites)){
28   ans<-(sites$Position[i]<=65)
29   sites$E1[i]<-ans
30 }
31
32 # Start building network
33 RL6.7_topology<-subset(RL6.7[2:3])
34 g2<-graph.data.frame(RL6.7_topology,vertices=sites,directed=FALSE)
35 V(g2)$color<-ifelse(V(g2)$E1==TRUE,"white","grey")
36 V(g2)$color<-ifelse(V(g2)$E1==TRUE,"white","grey")
37 plot(g2,vertex.label.color="black",vertex.size=20,edge.color="black",edge.width=1.5)
38
```



Lack of annotation

Poor naming conventions

Poor readability

Spacing absent

- Open the script am_bad_habits.R

A screenshot of the RStudio Environment pane. It shows a 'Global Environment' with a list of objects. The objects include 'ans' (logical), 'df' (spec_tbl), 'g2' (igraph), 'i' (integer), and several 'RL' objects (tbl_df) representing different parts of a dataset. The objects are listed with their names, types, lengths, sizes, and values. The list is long and cluttered, illustrating the problem of having too many intermediate objects in the environment.

Name	Type	Length	Size	Value
ans	logical	1	56 B	TRUE
df	spec_tbl...	4	5.9 KB	39 obs. of 4 variables
g2	igraph	10	2.4 KB	List of 10
i	integer	1	56 B	4L
RL1.2	tbl_df	4	1.4 KB	3 obs. of 4 variables
RL2.3	tbl_df	4	1.4 KB	3 obs. of 4 variables
RL3.4	tbl_df	4	1.3 KB	1 obs. of 4 variables
RL4.5	tbl_df	4	1.8 KB	9 obs. of 4 variables
RL5.6	tbl_df	4	1.9 KB	20 obs. of 4 variables
RL6.7	tbl_df	4	1.3 KB	2 obs. of 4 variables
RL6.7_topolo...	tbl_df	2	1016 B	2 obs. of 2 variables
sites	data.fra...	2	1.1 KB	4 obs. of 2 variables
sites1	list	2	176 B	List of 2
sites2	list	2	176 B	List of 2



Cluttered environment

Intermediate objects

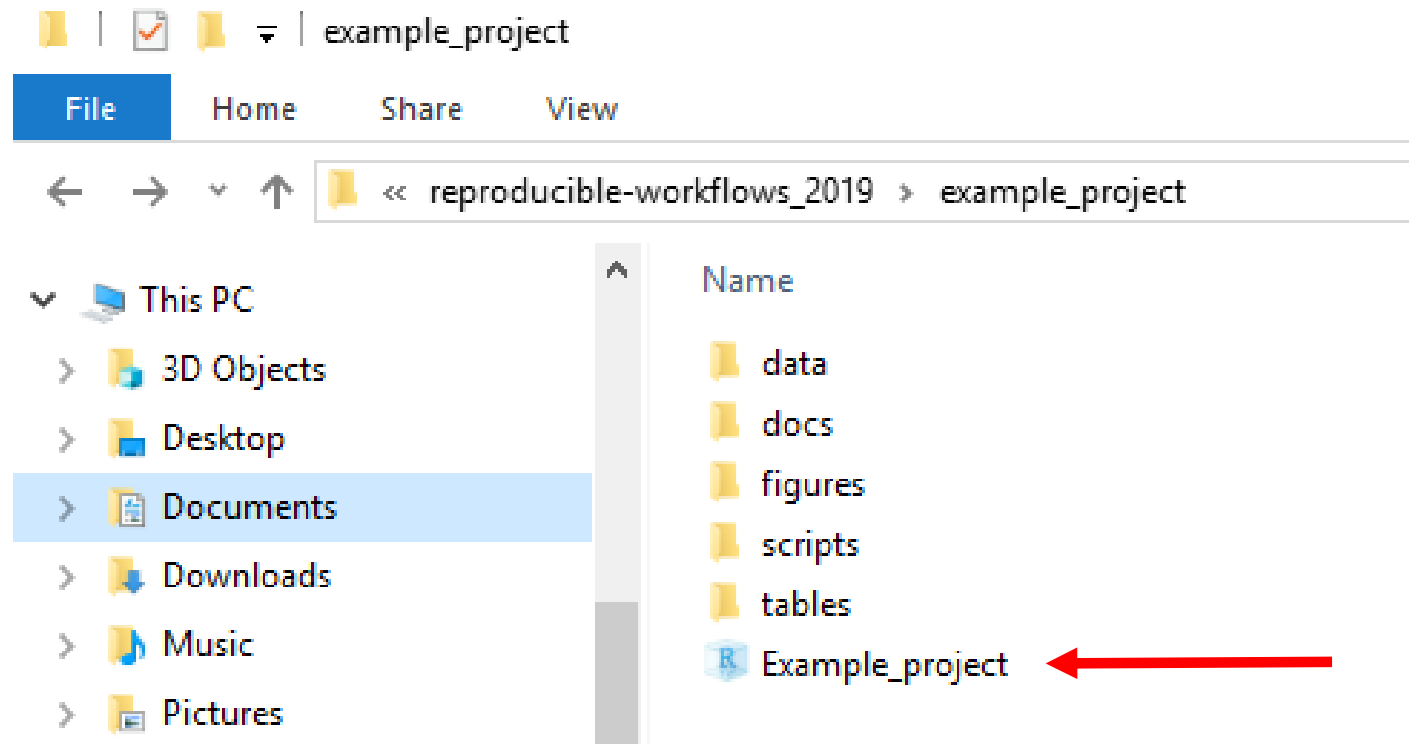
Writing clearer code

- Annotation
- Object names
 - should use only lowercase letters, numbers, and “_”
- Spacing
 - Put a space before and after =
 - Put a space after a ,
 - Operators should be surrounded by spaces e.g. ==, <-, +
- For a more complete list visit
 - <http://style.tidyverse.org/syntax.html>
- Open the script am_good_habits.R

Navigating RStudio – some useful tips

- Open the script `am_rstudio_ide_tricks.R`

B: R-projects - everything in its right place



- Open the scripts `01_eg_clean_data.R`, `02_eg_figures.R` and `03_eg_analysis.R`

Define a generic project structure



- STEP 1: Give your research projects a shared structure

File Home Share View			
← → ↕ ↑ This PC > Documents > Projects > generic.project > analysis			
	Name	Date modified	Type
Quick access	.Rproj.user	09/04/2018 20:28	File folder
	data	26/04/2017 06:43	File folder
	docs	26/04/2017 06:43	File folder
	plots	26/04/2017 06:43	File folder
	scripts	09/04/2018 20:28	File folder
	tables	26/04/2017 06:43	File folder
	.Rhistory	22/03/2018 14:09	RHISTORY File
	generic	25/08/2017 15:46	RMD File
	genericProject	22/03/2018 14:05	R Project
	style.1	06/07/2017 13:33	Microsoft Word D...
Desktop			
Downloads			
Documents			
Pictures			
Projects			
Google Drive			
House			
Google Drive File Stream (G:)			
FAIR_workshop			

Give your files informative names

- STEP 2: Include metadata in the file names











› This PC › Documents › R_Users_Workshop › lunchtime_sessions-master › R-projects › data

Name	Date modified
 raw_data	14/11/2018 23:02
 2018-11-04_clean_who_tb_data	04/11/2018 15:15













Come back to what you know

- STEP 3: Make you file names machine readable, human readable and work with default ordering

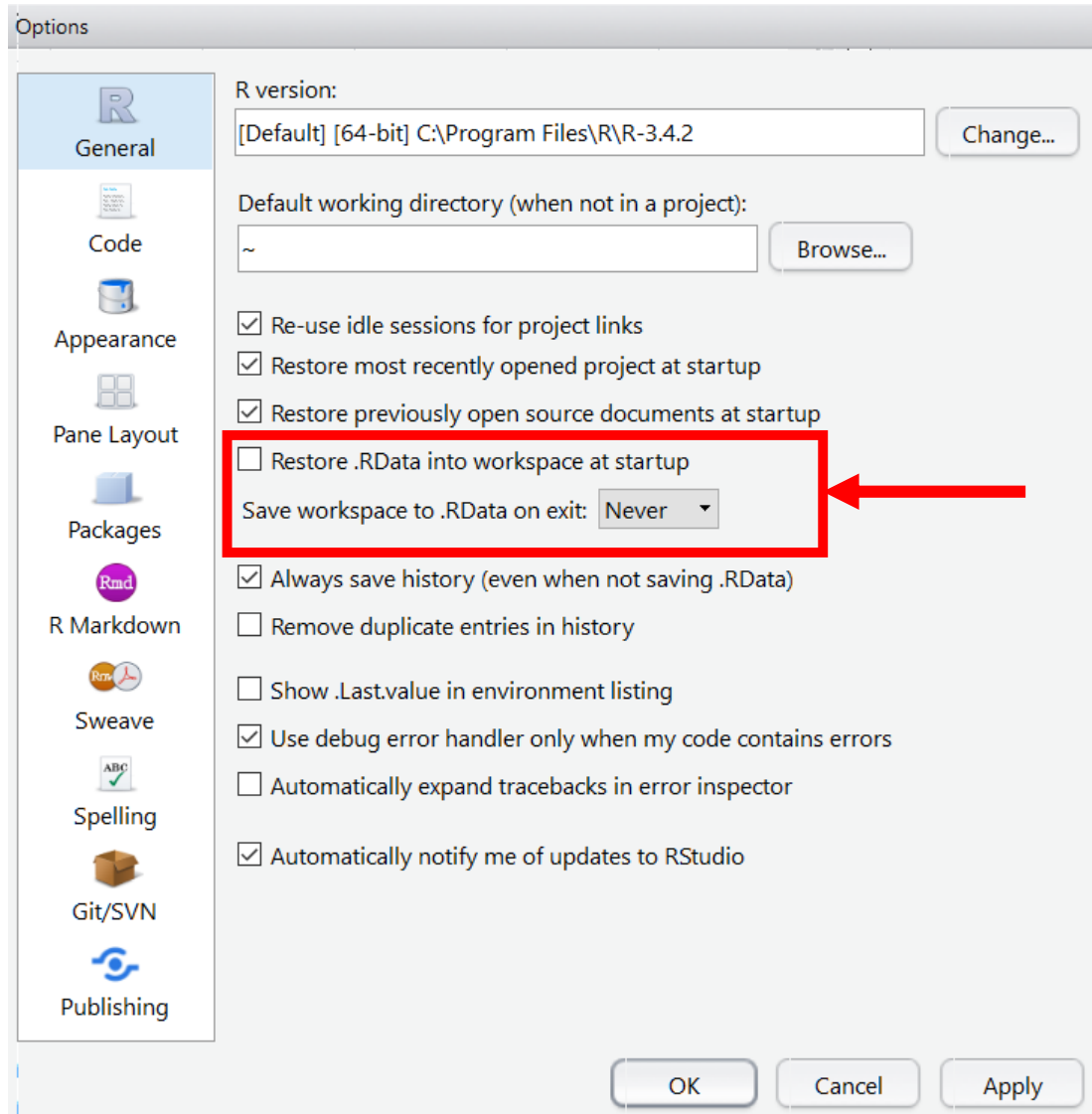
NO

 Epistatic_change
 Epistatic_change_match_discovery
 Epistatic_change_match_discovery_fig_2_point_1
 Epistatic_change_v2
 epistatic_codon_change_tracking
 Epistatic_connection_network
 Heatmap_for_epistatic_syn
 Heatmap1_for_epi_site_co-change
 Heatmap2_for_epi_fdr_adjusted_p-value
 Heatmap2_for_epi_p-value

Yes

Documents > Projects > 18.04.27-WP3_Feeding_Trial > analysis > scripts		
	Name	Date modified
	 01_data_import_and_tidying_master_file	02/10/2018 18:51
	 02_data_import_and_tidying_nutritics_grouped	19/10/2018 19:47
	 03_figures	17/10/2018 16:40
	 04_tables	22/05/2018 12:26
	 <u>05_study_overview</u>	19/10/2018 23:06
	 functions	13/05/2018 23:13

Other points to note



- You might consider your environment as "real"
- If you continue to use R, it is better for you to consider your R scripts as "real", as these should recreate the environment
- You may suffer short term pain
- This will prevent long term agony