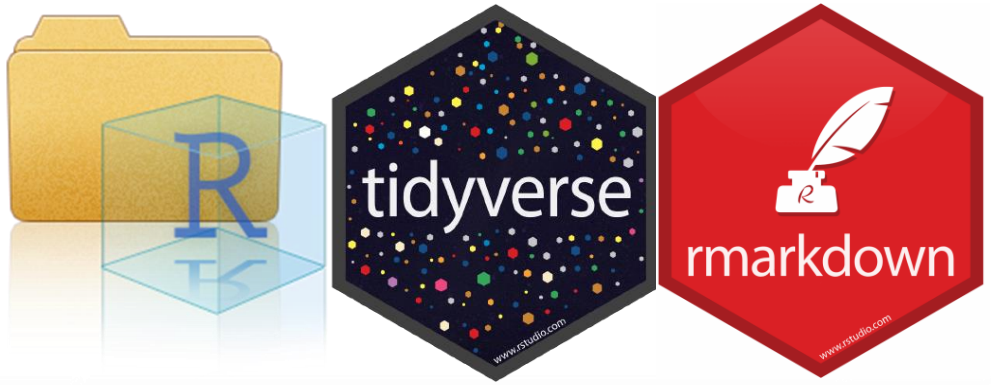
R-eproducible workflows

Half-day workshop

(after 2.5 days of intense R training!!)



Brendan Palmer, University College Cork
Adam Kane, University College Dublin
Enrico Pirotta, Washington State University

BEES
School of Biological, Earth & Environmental Sciences at University College Cork, Ireland

CRF-C
HRB Clinical Research Facility Cork

UCD DUBLIN
University College Dublin
Ireland's Global University

WASHINGTON STATE UNIVERSITY

Putting the pieces together

- Project structure
    - Naming conventions
        - Scripted workflows
            - R Markdown
                - Reproducible research

What is reproducible research?

		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	Generalisable

How is research presented?

Papers

Journal of Virology

Network Analysis of the Chronic Hepatitis C Virome Defines Hypervariable Region 1 Evolutionary Phenotypes in the Context of Humoral Immune Responses

Brendan A. Palmer,[a] Daniel Schmidt-Martin,[a] Zoya Dimitrova,[b] Pavel Skums,[b] Orla Crosbie,[c] Elizabeth Kenny-Walsh,[c] Liam J. Fanning[a]

Molecular Virology Diagnostic & Research Laboratory, Department of Medicine, University College Cork, Cork, Ireland[a]; Division of Viral Hepatitis, Centers for Disease Control and Prevention, Atlanta, Georgia, USA[b]; Department of Hepatology, Cork University Hospital, Cork, Ireland[c]

ABSTRACT

Hypervariable region 1 (HVR1) of hepatitis C virus (HCV) comprises the first 27 N-terminal amino acid residues of E2. It is classically seen as the most heterogeneous region of the HCV genome. In this study, we assessed HVR1 evolution by using ultradeep pyrosequencing for a cohort of treatment-naive, chronically infected patients over a short, 16-week period. Organization of the sequence set into connected components that represented single nucleotide substitution events revealed a network dominated by highly connected, centrally positioned master sequences. HVR1 phenotypes were observed to be under strong purifying (stationary) and strong positive (antigenic drift) selection pressures, which were coincident with advancing patient age and cirrhosis of the liver. It followed that stationary viromes were dominated by a single HVR1 variant surrounded by minor variants comprised from conservative single amino acid substitution events. We present evidence to suggest that neutralization antibody efficacy was diminished for stationary-virome HVR1 variants. Our results identify the HVR1 network structure during chronic infection as the preferential dominance of a single variant within a narrow sequence space.

IMPORTANCE

HCV infection is often asymptomatic, and chronic infection is generally well established in advance of initial diagnosis and subsequent treatment. HVR1 can undergo rapid sequence evolution during acute infection, and the variant pool is typically seen to diverge away from ancestral sequences as infection progresses from the acute to the chronic phase. In this report, we describe HVR1 viromes in chronically infected patients that are defined by a dominant epitope located centrally within a narrow variant pool. Our findings suggest that weakened humoral immune activity, as a consequence of persistent chronic infection, allows for the acquisition and maintenance of host-specific adaptive mutations at HVR1 that reflect virus fitness.

Hepatitis C virus (HCV) infection is a global health issue and is recognized as a major etiological agent of liver-related diseases (1). It has been estimated that the current prevalence of HCV represents approximately 2% of the global adult (15 years of age and older) population (2). Following transmission, HCV infection may remain asymptomatic for decades, resulting in the majority of infections initially passing undetected (3). It is estimated that up to 4 million Americans are living with the virus, the majority of whom became infected prior to the isolation and identification of the virus (4, 5). Consequently, the U.S. Centers for Disease Control and Prevention now recommend that Americans born from 1945 to 1965 be screened for the presence of the virus notwithstanding the presence of clinical symptoms (3, 5).

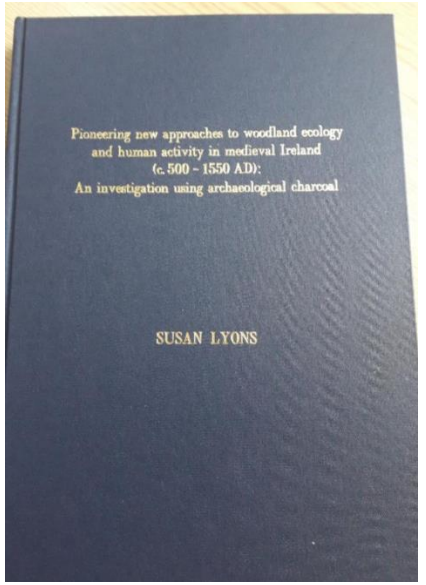
HCV is a single-stranded positive-sense RNA virus of considerable genomic heterogeneity. A recent reclassification defined the HCV global distribution into 7 genotypes and 67 subtypes, with genotypes 1 and 3 accounting for the majority of infections worldwide (6, 7). An error-prone RNA-dependent RNA polymerase, together with an inherent tolerance of defined hypervariable regions (HVR), accounts for much of this variability. Three HVRs are located within the envelope glycoprotein E2. The greatest heterogeneity has been identified at the 27-amino-acid HVR1 (residues 384 to 410 of the H77 reference strain), located at the amino-terminal end of the E2 glycoprotein (8). Recent studies indicated that the central region of E2 (residues 456 to 656) is globular and surprisingly compact, whereas the first 80 amino acids (including HVR1) lack this structural rigidity (9). This observation is consistent with a region that is proposed to shield conserved neutralizing epitopes and to participate in high-density lipoprotein enhancement of infection via scavenger receptor class B type I (SRBI) interactions and is itself targeted by neutralizing antibodies (nAb) (10–16).

Mutational flexibility at HVR1 was characterized soon after the initial identification of HCV (8, 17). Rapid mutational change of HVR1 has been documented over weeks during the acute phase of infection, where HVR1 evolution is governed predominantly by strong selective pressures, with fixation of beneficial mutations (11, 18, 19). Reports examining samples collected over years to decades have documented the emergence of convergent HVR1

Received 25 November 2015 Accepted 22 December 2015
Accepted manuscript posted online 30 December 2015
Citation Palmer BA, Schmidt-Martin D, Dimitrova Z, Skums P, Crosbie O, Kenny-Walsh E, Fanning LJ. 2016. Network analysis of the chronic hepatitis C virome defines hypervariable region 1 evolutionary phenotypes in the context of humoral immune responses. J Virol 90:3318–3329. doi:10.1128/JVI.02995-15.
Editor M. S. Diamond
Address correspondence to Liam J. Fanning, l.fanning@ucc.ie.
B.A.P. and D.S.-M. contributed equally to this article.

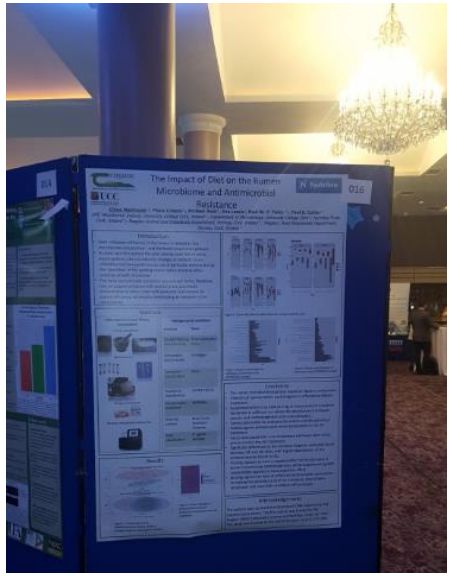
Books

Theses



Talks

Posters



But what does it look like under the bonnet?

Link to the video on YouTube

Disclaimer

Jenny Bryan
@JennyBryan

Software engineer @rstudio, humane #rstats, adjunct prof @UBC where I created @STAT545, part of @ropensci

Hadley Wickham
@hadleywickham

R, data, visualisation.

Houston, TX

hadley.nz

HADLEY WICKHAM TEACHING CODE PERSONAL

I also teach in person workshops from time-to-time; see the RStudio workshops page for more details.

CODE

Most of my work is in the form of open source R code, which you can find on my github. You can roughly divide my work into three categories: tools for data science, tools for data import, and software engineering tools.

DATA SCIENCE

- ggplot2 for visualising data.
- dplyr for manipulating data.
- tidyr for tidying data.
- stringr for working with strings.
- lubridate for working with date/times.

DATA IMPORT

- readr for reading .csv and fwf files.
- readxl for reading .xls and .xlsx files.
- haven for SAS, SPSS, and Stata files.
- httr for talking to web APIs.
- rvest for scraping websites.
- xml2 for importing XML files.

SOFTWARE ENGINEERING

- devtools for general package development.
- roxygen2 for in-line documentation.
- testthat for unit testing

David Robinson
@drob

Chief Data Scientist at @DataCamp, #rstats fan/evangelist

New York, NY

varianceexplained.org

STAT 545 Home FAQ Syllabus Topics People

Data wrangling, exploration, and analysis with R

UBC STAT 545A and 547M

Learn how to

- explore, groom, visualize, and analyze data
- make all of that reproducible, reusable, and shareable
- using R

VARIANCE EXPLAINED

ABOUT ME POSTS LEARN R TEXT MINING IN R INTRODUCTION TO EMPIRICAL BAYES

David Robinson

Chief Data Scientist at DataCamp, works in R and Python.

Email
Twitter
Github
Stack Overflow

This is the homepage and blog of David Robinson, Chief Data Scientist at DataCamp. For more about me, see here.

Recent Posts

Exploring college major and income: a live data analysis in R *October 16, 2018*
A live screencast of an exploratory data analysis from the Tidy Tuesday series. This one explores college major and income data from 538.

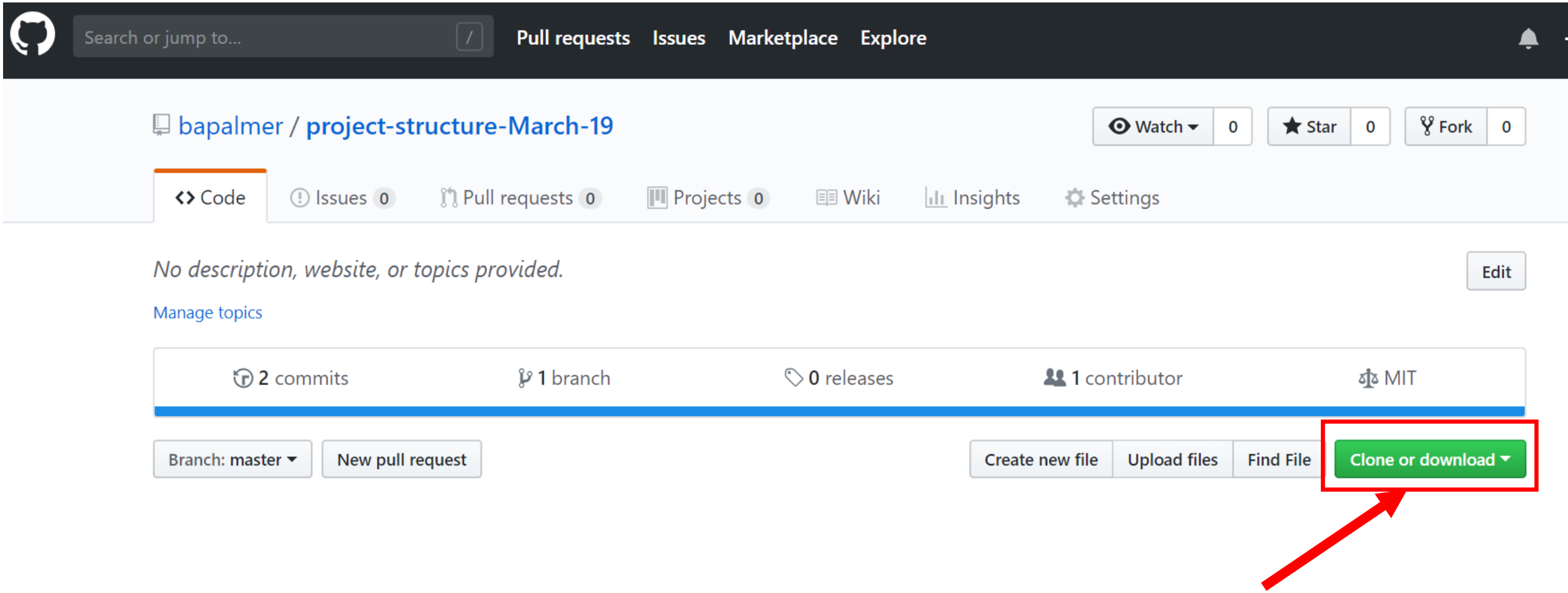
Who wrote the anti-Trump New York Times op-ed? Using tidytext to find document similarity *September 06, 2018*
An analysis of an anonymous op-ed in the New York Times, using document similarity metrics to match it to Twitter accounts.

Scientific debt *May 10, 2018*
Introducing an analogy to 'technical debt' for data scientists.

R projects

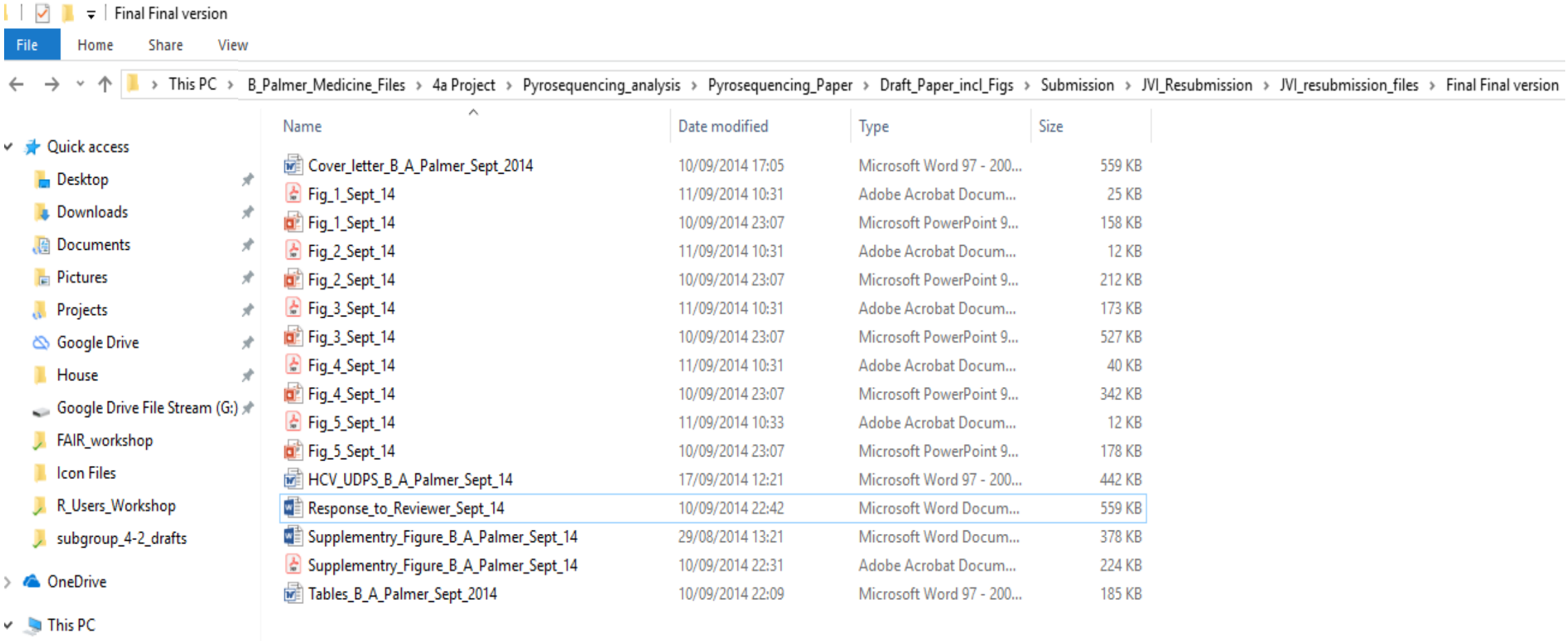
R projects

- Here's one I made earlier……

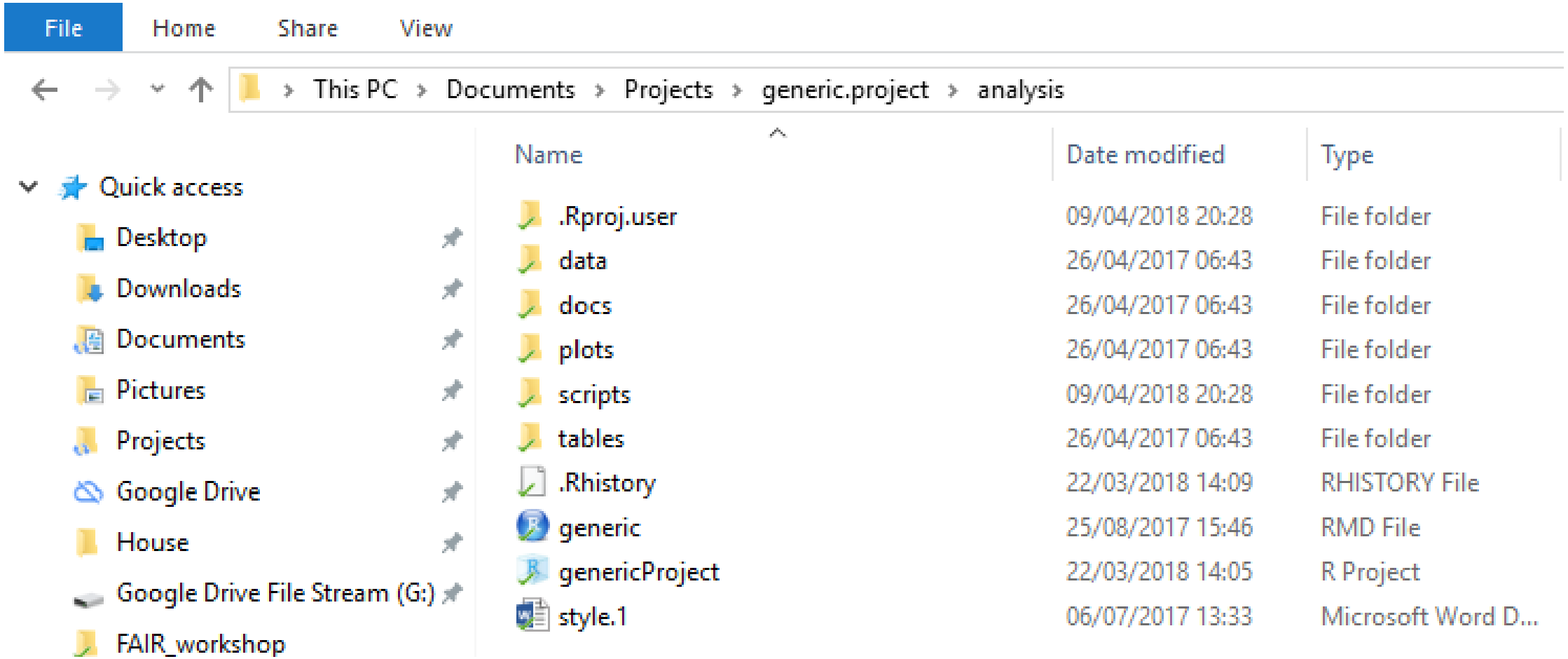


Still haven't found what I'm looking for

- **Help your future-self**

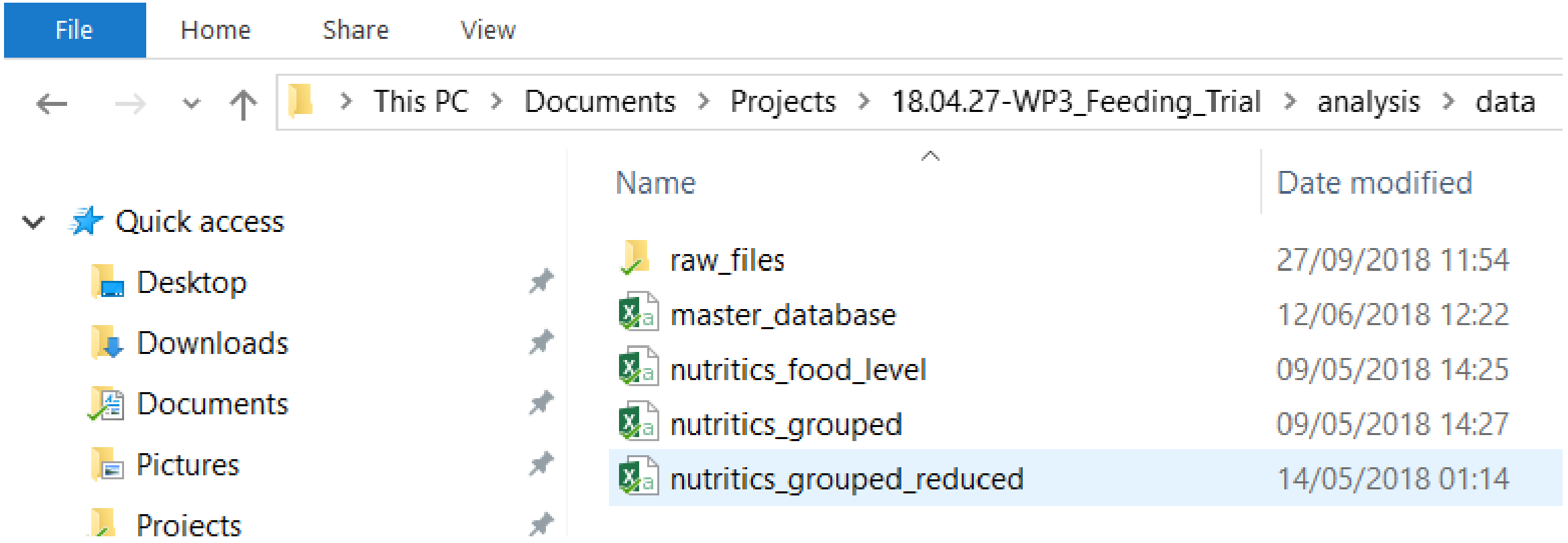


Step 1: Define a generic project structure



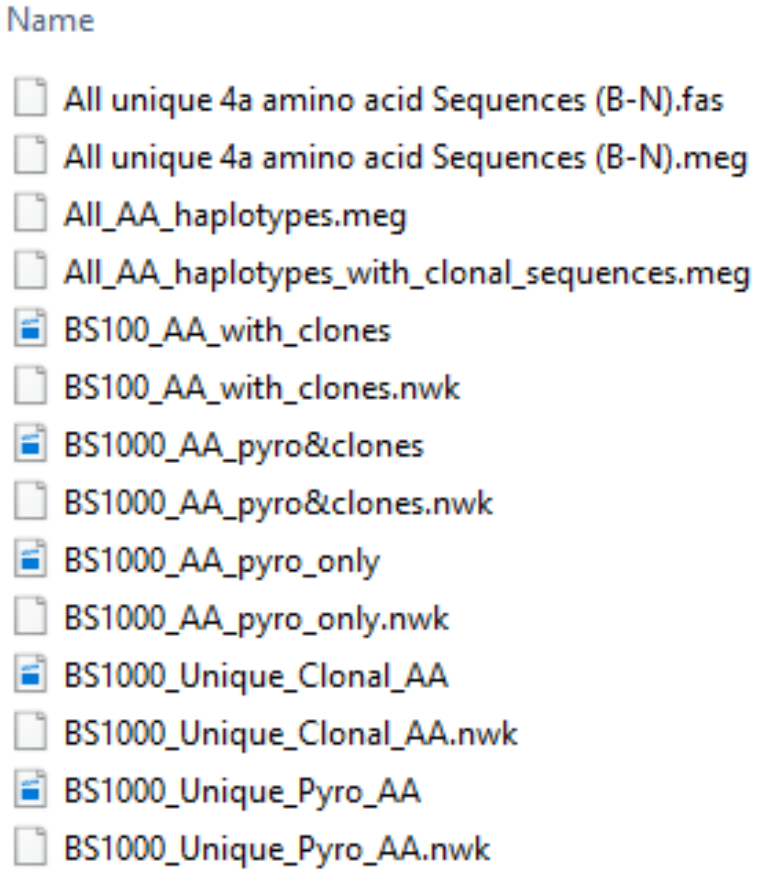
Step 2: Give your files informative names

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

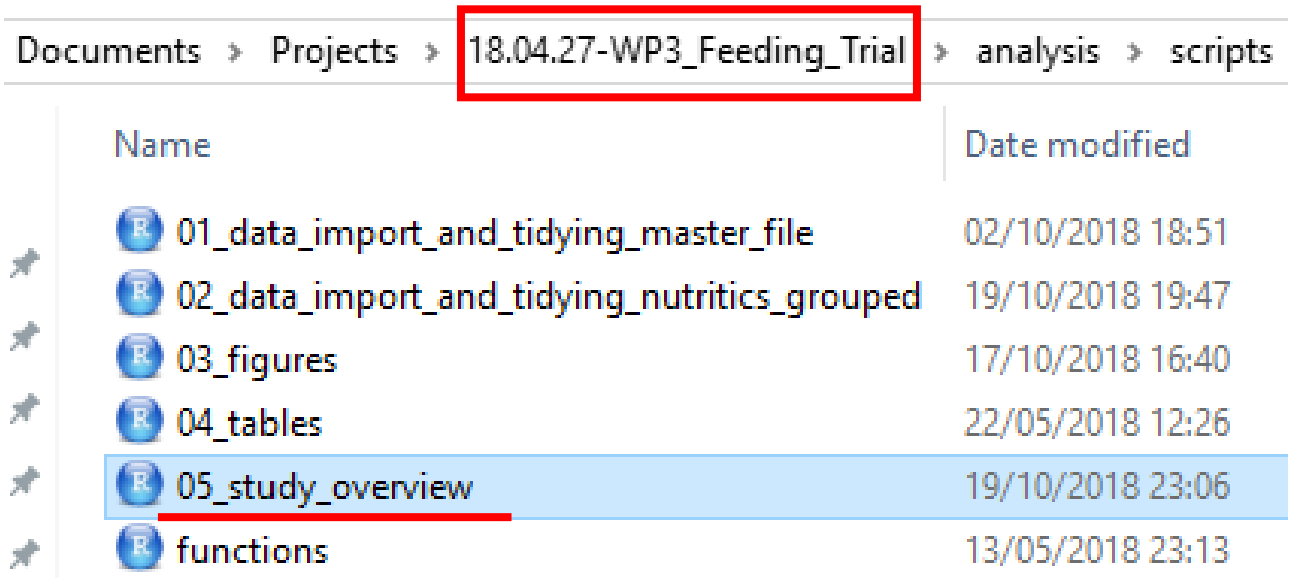


Step 3: Everything in its right place

NO



Yes



Outline a file naming convention

Machine readable:

- Inherent order
- Avoid spaces
- Avoid punctuation
- Remove case-sensitivity

Human readable:

- Contains info on content
- Avoid spaces
- Avoid punctuation
- Remove case sensitivity

Metadata:

Separate with underscores ("_")

- Avoid punctuation
- Remove case-sensitivity

```
01_marshal-data.r

02_pre-dea-filtering.r

03_dea-with-limma-voom.r

04_explore-dea-results.r

90_limma-model-term-name-fiasco.r

helper01_load-counts.r

helper02_load-exp-des.r

helper03_load-focus-statinf.r

helper04_extract-and-tidy.r
```

Image from Data Carpentry webpage

Outline a file naming convention

Chronological order:

```
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H01.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H02.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H03.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_platefile.csv
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A01.csv
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A02.csv
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A03.csv
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A04.csv
```

Logical order:

```
01_marshal-data.r
02_pre-dea-filtering.r
03_dea-with-limma-voom.r
04_explore-dea-results.r
90_limma-model-term-name-fiasco.r
helper01_load-counts.r
helper02_load-exp-des.r
helper03_load-focus-statinf.r
helper04_extract-and-tidy.r
```

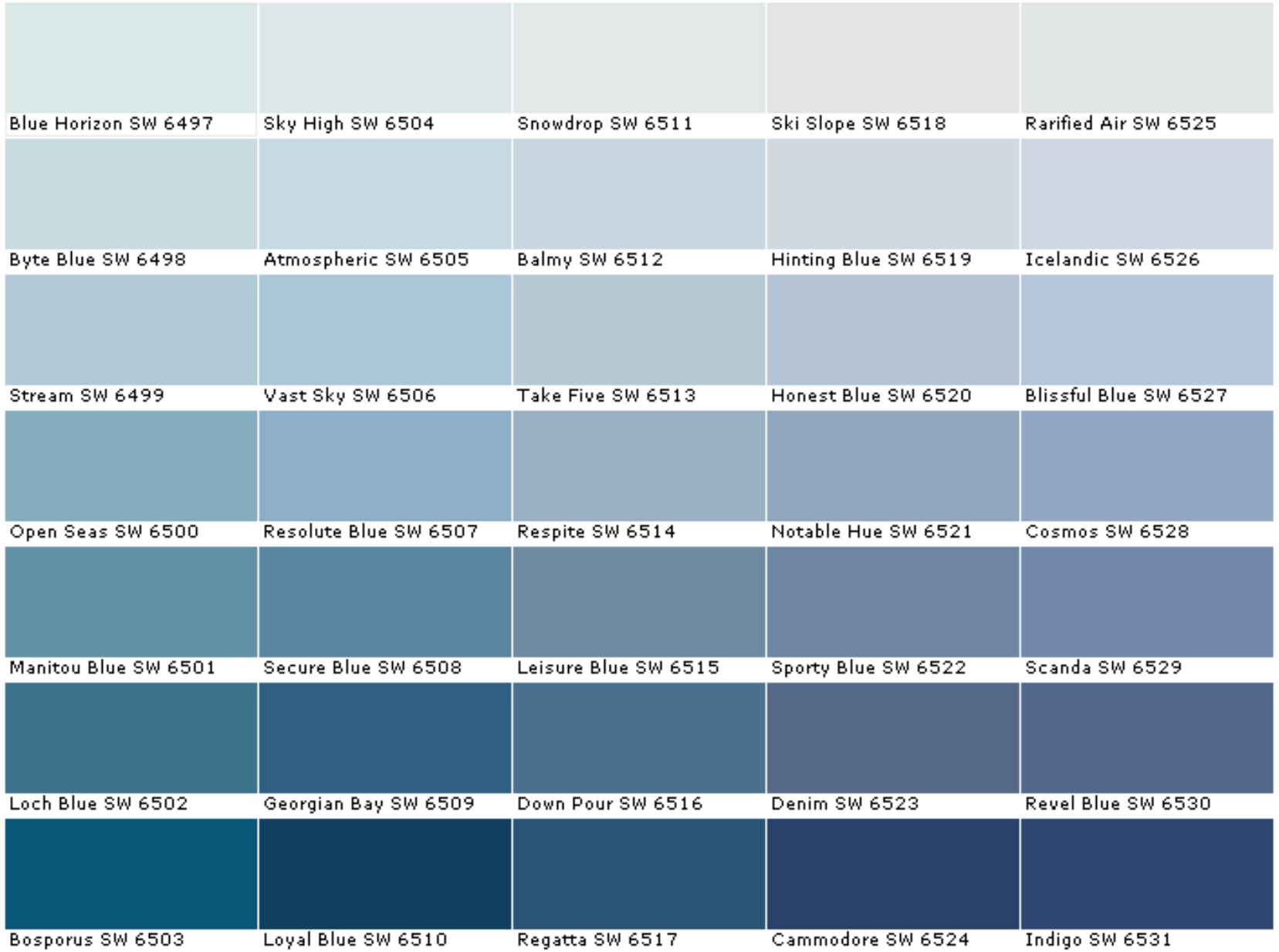
Image from Data Carpentry webpage

Joined up thinking

- The R scripts you generate should be human readable
    - Annotate the code
    - Break up the scripts into dedicated tasks
    - Interlink with other within project scripts

```
1  # Data ----
2  # Eight tibbles returned from the 01_data_import_and_tidying_master_file.R
3  # 1. fgf23_data => FGF23 readings from study centres 01-03
4  # 2. food_level_data => Food diary entries
5  # 3. grouped_data => Dialysis and nondialysis diary entries by component
6  # 4. k_data => Serum potassium
7  # 5. master_data_clean => all the clean master file data if required
8  # 6. p_data => Serum phosphate
9  # 7. pth_data => Parathyroid hormone readings
10 # 8. pulses_nuts_data
11
12 source("scripts/01_data_import_and_tidying_master_file.R")
```

Is too much choice good or bad?



Inconsistent function names, inconsistent syntax

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

```
Names or colnames
row.names or rownames
rowSums or rowsum
Sys.time, system.time
```

  - Is it 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 01_pm_too_much_choice.R

Writing clearer code

- Annotation

- Object names
  - should use only lowercase letters, numbers, and "_"

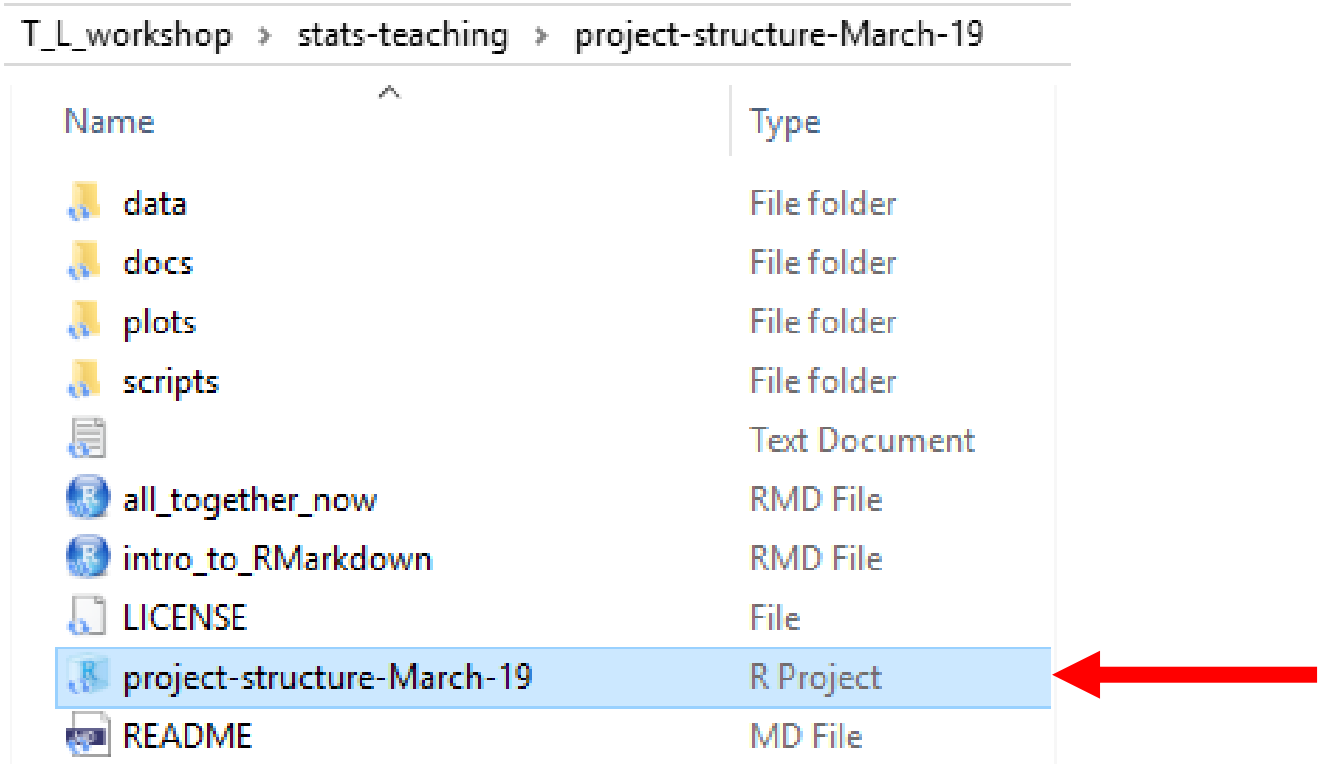
- Spacing
  - Put a space before and after =
  - Put a space after a ,
  - Operators should be surrounded by spaces e.g. ==, <-, +

- For a more complete list visit
  - http://style.tidyverse.org/syntax.html

- Open the script 02_pm_good_habits.R

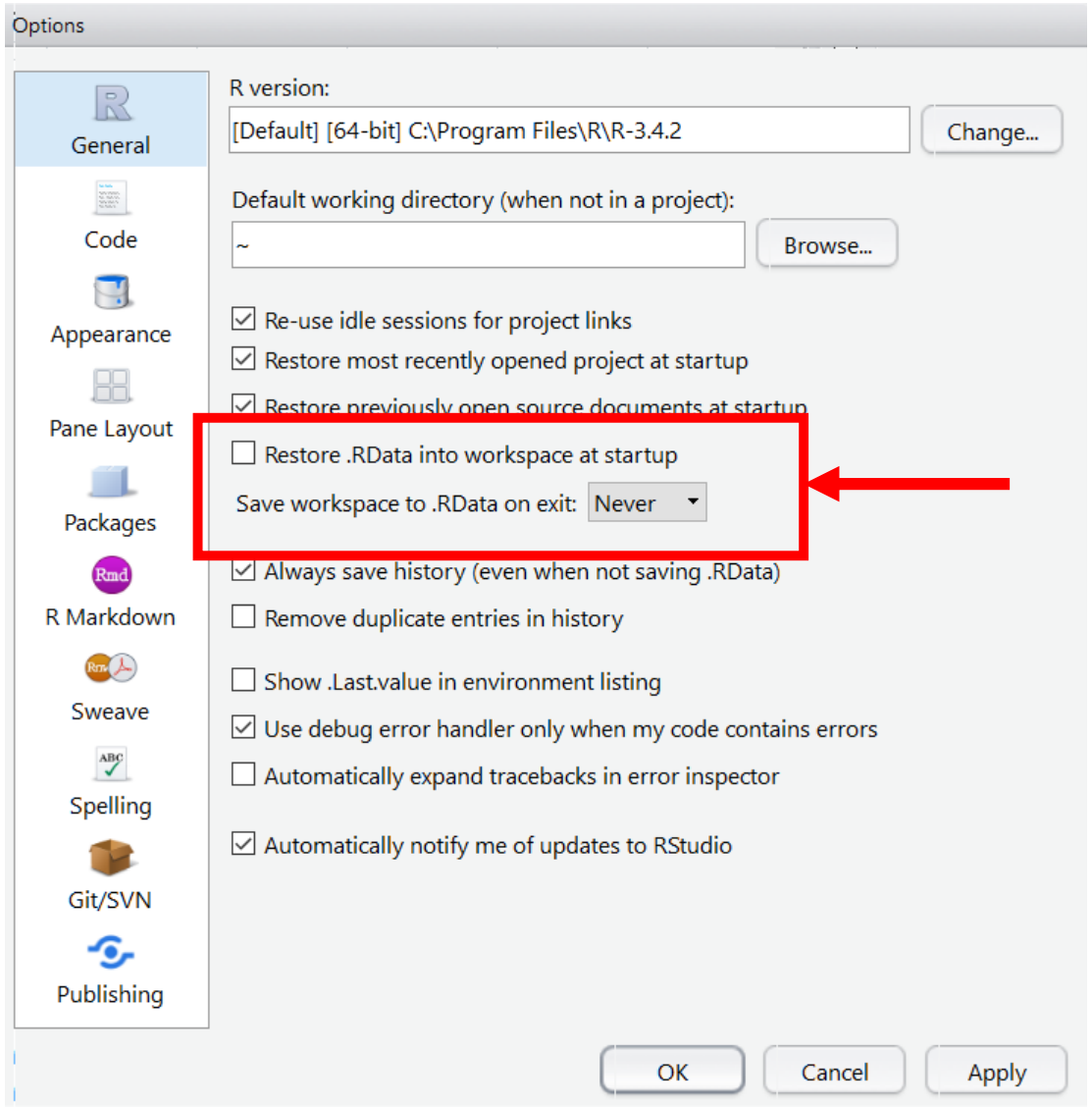
Everything in its right place

- benefits of using R projects for data analysis tasks



- Open the script 03_pm_clean_data.R
- Open the script 04_pm_plots.R
- Open the script 05_pm_analysis.R

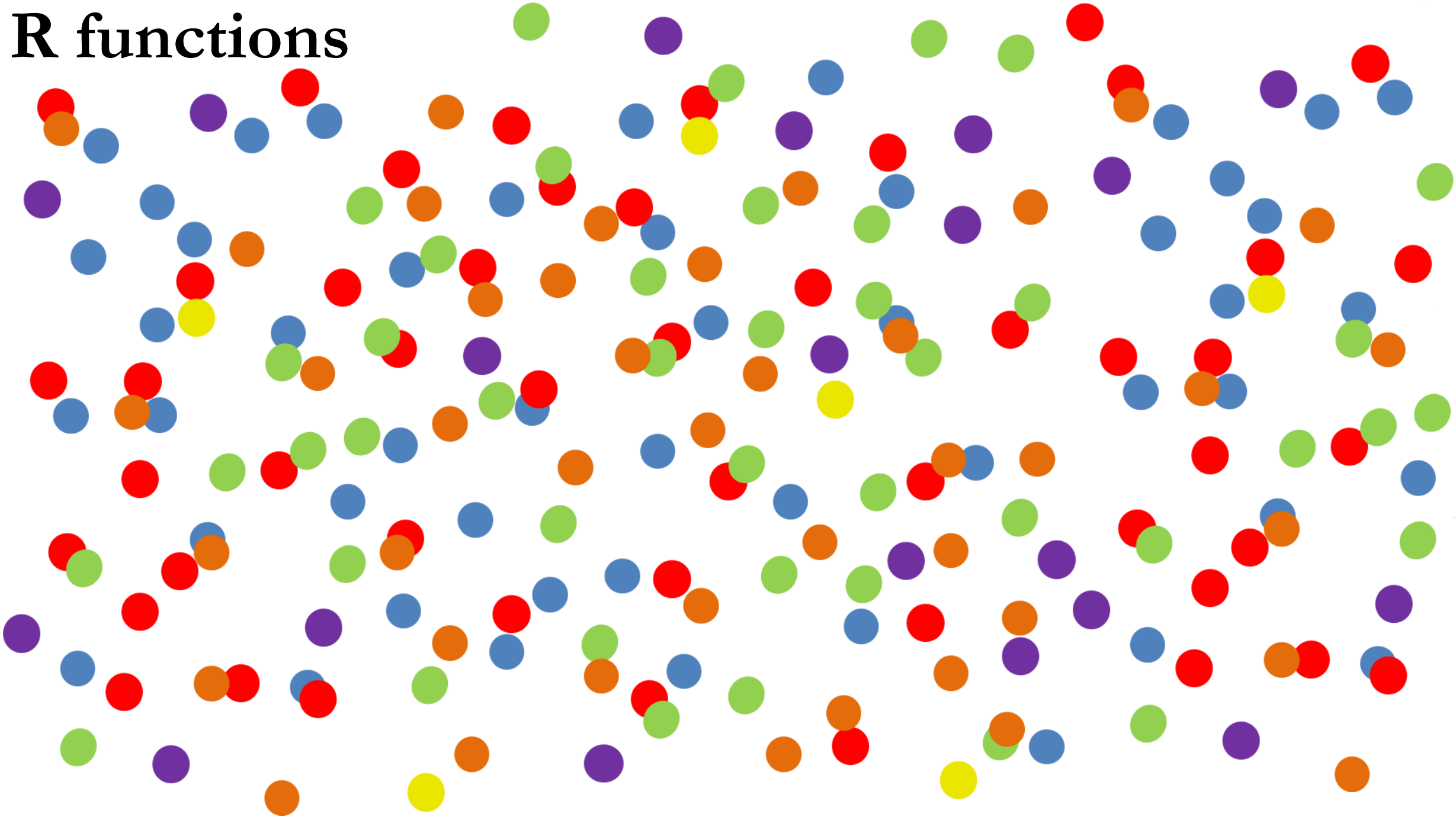
Other points to note



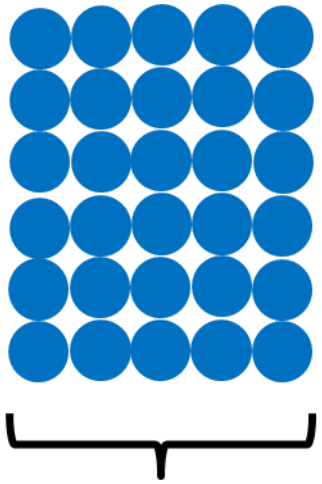
- You might consider your environment as "real"
- If you continue to use R, it is better for you to consider your R scripts as "real", as these should recreate the environment
- You may suffer short term pain
- This will prevent long term agony

Tidyverse

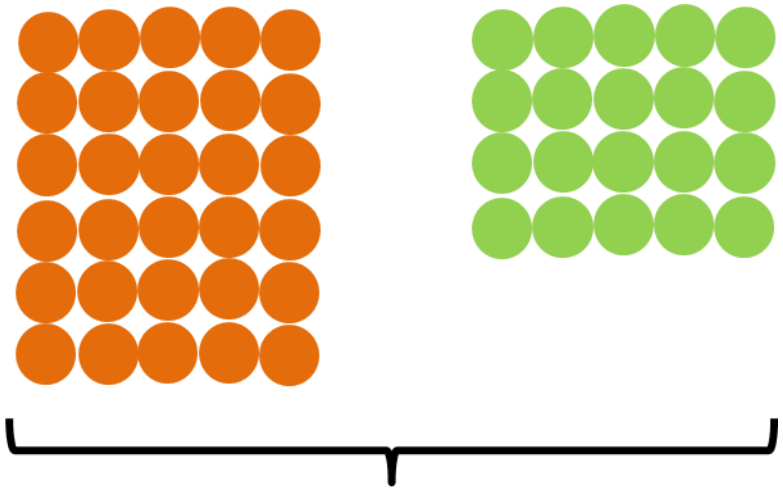
R functions



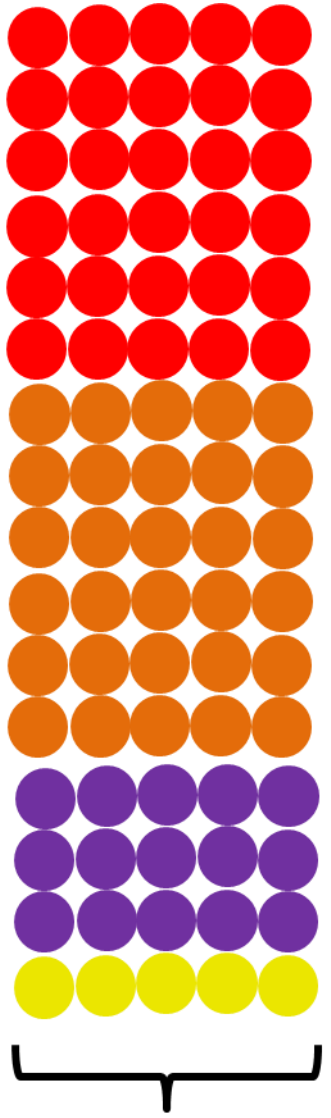
R packages



Base R:
Comes
pre-
loaded

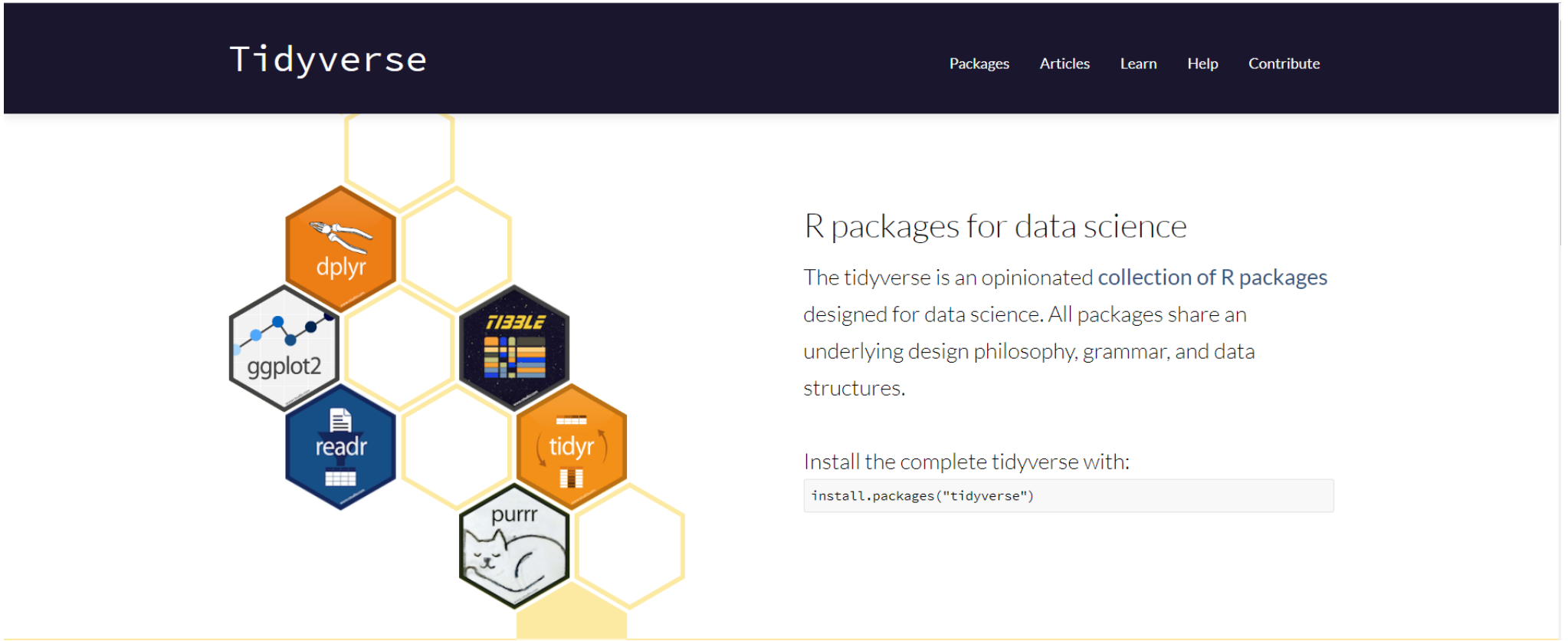


Other packages:
Install once
Update regularly
Load each session



core
tidyverse

What is the tidyverse?



- Joined up collection of packages for data analysis
    - Consistent functions
    - Uses (tidy) data
    - Supports end-to-end workflows

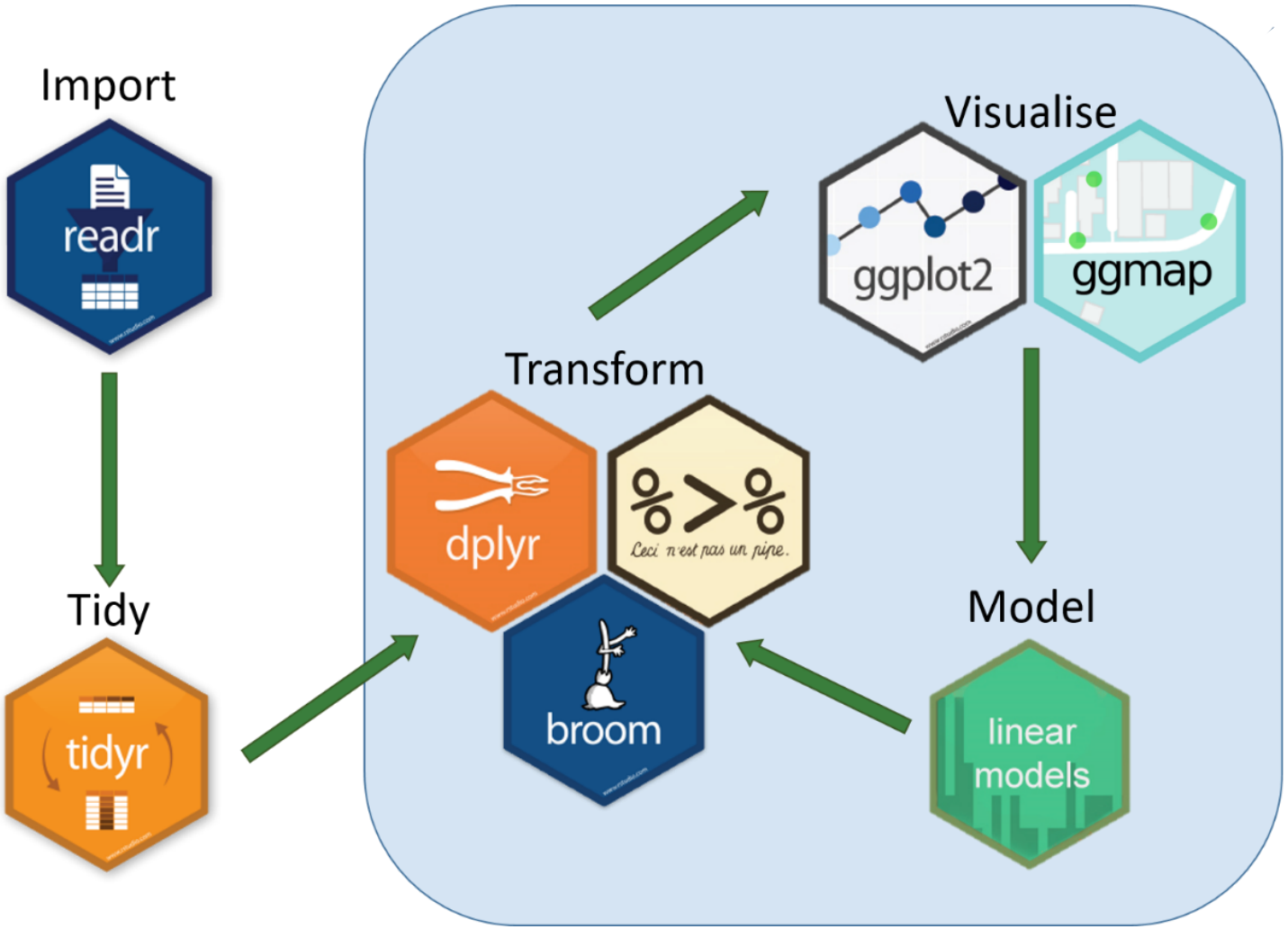
What is the tidyverse?

```
> install.packages(c("broom", "cli2", "crayon",
"dbplyr", "dplyr", "forcats", "ggplot2", "haven",
"hms", "httr", "jsonlite", "lubridate",
"magrittr", "modelr", "pillar", "purrr", "readr",
"readxl", "reprex", "rlang", "rstudioapi",
"rvest", "stringr", "tibble", "tidyr", "xml2")
```

```
> install.packages("tidyverse")
```

Putting the pieces together

- Data analysis in a tidyverse nutshell



Line by line

~/R_Users_Workshop/8_weeks_Oct-Dec_17/Workshop_1/workshop_1_project - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | workshop_1_project — 8_weeks_Oct-Dec_17

ws1_script1_stepwise_Bauer_dataset_an... *

Source on Save | Run | Source

```
12
13  raw_gene_df <- read_delim("Brauer2008_DataSet1.tds", delim = "\t")
14
15  separated_gene_df <- separate(raw_gene_df, NAME,
16                                c("name", "BP", "MF", "systematic_name",
17                                  "number"),
18                                sep = "\\|\\|")
19
20  mutated_gene_df <- mutate_at(separated_gene_df,
21                               vars(name:systematic_name),
22                               funs(trimws)
23                               )
24
25  selected_gene_df <- select(mutated_gene_df, -number, -GID, -YORF, -GWEIGHT)
26
27  gathered_gene_df <- gather(selected_gene_df, sample, expression, G0.05:U0.3)
28
29  nearly_there_df <- separate(gathered_gene_df, sample,
30                              c("nutrient", "rate"), sep = 1, convert = TRUE)
31
32  nutrient_names <- c(G = "Glucose", L = "Leucine", P = "Phosphate",
33                      S = "Sulfate", N = "Ammonia", U = "Uracil")
34
35  cleaned_genes_df <- mutate(nearly_there_df,
36                             nutrient = plyr::revalue(nutrient, nutrient_names)
37                             ) %>%
38    filter(!is.na(expression), systematic_name != "")
```

15:1 | Section 1: Data import, tidying and transformation | R Script

Console Terminal

~/R_Users_Workshop/8_weeks_Oct-Dec_17/Workshop_1/workshop_1_project/

```
> raw_gene_df <- read_delim("Brauer2008_DataSet1.tds", delim = "\t")
Parsed with column specification:
cols(
  .default = col_double(),
  GID = col_character(),
  YORF = col_character(),
  NAME = col_character(),
  GWEIGHT = col_integer()
)
See spec(...) for full column specifications.
>
```

Environment History Connections

Import Dataset | Grid

Global Environment

Name	Type	Length	Size	Value
raw_gene_df	tbl_df	40	3.3 MB	5537 obs. of 40 variables

Files Plots Packages Help Viewer

New Folder | Delete | Rename | More

Home > R_Users_Workshop > 8_weeks_Oct-Dec_17 > Workshop_1 > workshop_1_project

Name	Size	Modified
..		
.RData	2.5 KB	Oct 2, 2017, 1:49 PM
.Rhistory	20.3 KB	Dec 6, 2017, 3:43 PM
Brauer2008_DataSet1.csv	1.6 MB	Sep 27, 2017, 11:32 PM
Brauer2008_DataSet1.tds	1.6 MB	Sep 28, 2017, 10:22 AM
house_completions.csv	4 KB	Sep 28, 2017, 1:35 PM
irish_population.csv	315 B	Aug 28, 2017, 4:21 PM
raw_house_completions.csv	16.2 KB	Aug 25, 2017, 3:45 PM
workshop_1.Rproj	217 B	Oct 18, 2018, 12:18 PM
ws1_script1_stepwise_Bauer_dataset_analysis.R	6.1 KB	Dec 5, 2017, 12:19 PM
ws1_script2_Bauer_dataset_analysis.R	2 KB	Dec 6, 2017, 2:33 PM
ws1_script3_house_completions.R	2.4 KB	Oct 2, 2017, 3:53 PM

Line by line

~/R_Users_Workshop/8_weeks_Oct-Dec_17/Workshop_1/workshop_1_project - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | workshop_1_project — 8_weeks_Oct-Dec_17

ws1_script1_stepwise_Bauer_dataset_an... *

Source on Save | Run | Source

```
12
13  raw_gene_df <- read_delim("Brauer2008_DataSet1.tds", delim = "\t")
14
15  separated_gene_df <- separate(raw_gene_df, NAME,
16                                c("name", "BP", "MF", "systematic_name",
17                                  "number"),
18                                sep = "\\|\\|")
19
20  mutated_gene_df <- mutate_at(separated_gene_df,
21                               vars(name:systematic_name),
22                               funs(trimws)
23                               )
24
25  selected_gene_df <- select(mutated_gene_df, -number, -GID, -YORF, -GWEIGHT)
26
27  gathered_gene_df <- gather(selected_gene_df, sample, expression, G0.05:U0.3)
28
29  nearly_there_df <- separate(gathered_gene_df, sample,
30                              c("nutrient", "rate"), sep = 1, convert = TRUE)
31
32  nutrient_names <- c(G = "Glucose", L = "Leucine", P = "Phosphate",
33                      S = "Sulfate", N = "Ammonia", U = "Uracil")
34
35  cleaned_genes_df <- mutate(nearly_there_df,
36                             nutrient = plyr::revalue(nutrient, nutrient_names)
37                             ) %>%
38    filter(!is.na(expression), systematic_name != "")
```

20:1 | Section 1: Data import, tidying and transformation | R Script

Console Terminal

~/R_Users_Workshop/8_weeks_Oct-Dec_17/Workshop_1/workshop_1_project/

```
> raw_gene_df <- read_delim("Brauer2008_DataSet1.tds", delim = "\t")
Parsed with column specification:
cols(
  .default = col_double(),
  GID = col_character(),
  YORF = col_character(),
  NAME = col_character(),
  GWEIGHT = col_integer()
)
See spec(...) for full column specifications.
> separated_gene_df <- separate(raw_gene_df, NAME,
+                               c("name", "BP", "MF", "systematic_name",
+                                 "number"),
+                               sep = "\\|\\|")
>
```

Environment History Connections

Import Dataset | Grid

Global Environment

Name	Type	Length	Size	Value
raw_gene_df	tbl_df	40	3.3 MB	5537 obs. of 40 variables
separated_gene...	tbl_df	44	3.6 MB	5537 obs. of 44 variables

Files Plots Packages Help Viewer

New Folder | Delete | Rename | More

Home > R_Users_Workshop > 8_weeks_Oct-Dec_17 > Workshop_1 > workshop_1_project

Name	Size	Modified
..		
.RData	2.5 KB	Oct 2, 2017, 1:49 PM
.Rhistory	20.3 KB	Dec 6, 2017, 3:43 PM
Brauer2008_DataSet1.csv	1.6 MB	Sep 27, 2017, 11:32 PM
Brauer2008_DataSet1.tds	1.6 MB	Sep 28, 2017, 10:22 AM
house_completions.csv	4 KB	Sep 28, 2017, 1:35 PM
irish_population.csv	315 B	Aug 28, 2017, 4:21 PM
raw_house_completions.csv	16.2 KB	Aug 25, 2017, 3:45 PM
workshop_1.Rproj	217 B	Oct 18, 2018, 12:18 PM
ws1_script1_stepwise_Bauer_dataset_analysis.R	6.1 KB	Dec 5, 2017, 12:19 PM
ws1_script2_Bauer_dataset_analysis.R	2 KB	Dec 6, 2017, 2:33 PM
ws1_script3_house_completions.R	2.4 KB	Oct 2, 2017, 3:53 PM

Line by line

~/R_Users_Workshop/8_weeks_Oct-Dec_17/Workshop_1/workshop_1_project - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | workshop_1_project — 8_weeks_Oct-Dec_17

ws1_script1_stepwise_Bauer_dataset_an... *

Source on Save | Run | Source

```
12
13  raw_gene_df <- read_delim("Brauer2008_DataSet1.tds", delim = "\t")
14
15  separated_gene_df <- separate(raw_gene_df, NAME,
16                                c("name", "BP", "MF", "systematic_name",
17                                  "number"),
18                                sep = "\\|\\|")
19
20  mutated_gene_df <- mutate_at(separated_gene_df,
21                               vars(name:systematic_name),
22                               funs(trimws)
23                               )
24
25  selected_gene_df <- select(mutated_gene_df, -number, -GID, -YORF, -GWEIGHT)
26
27  gathered_gene_df <- gather(selected_gene_df, sample, expression, G0.05:U0.3)
28
29  nearly_there_df <- separate(gathered_gene_df, sample,
30                              c("nutrient", "rate"), sep = 1, convert = TRUE)
31
32  nutrient_names <- c(G = "Glucose", L = "Leucine", P = "Phosphate",
33                      S = "Sulfate", N = "Ammonia", U = "Uracil")
34
35  cleaned_genes_df <- mutate(nearly_there_df,
36                             nutrient = plyr::revalue(nutrient, nutrient_names)
37                             ) %>%
38    filter(!is.na(expression), systematic_name != "")
```

27:1 | Section 1: Data import, tidying and transformation | R Script

Console Terminal

~/R_Users_Workshop/8_weeks_Oct-Dec_17/Workshop_1/workshop_1_project/

```
Parsed with column specification:
cols(
  .default = col_double(),
  GID = col_character(),
  YORF = col_character(),
  NAME = col_character(),
  GWEIGHT = col_integer()
)
See spec(...) for full column specifications.
> separated_gene_df <- separate(raw_gene_df, NAME,
+                               c("name", "BP", "MF", "systematic_name",
+                                 "number"),
+                               sep = "\\|\\|")
> mutated_gene_df <- mutate_at(separated_gene_df,
+                              vars(name:systematic_name),
+                              funs(trimws)
+                              )
> selected_gene_df <- select(mutated_gene_df, -number, -GID, -YORF, -GWEIGHT)
>
```

Environment History Connections

Import Dataset | Grid

Global Environment

Name	Type	Length	Size	Value
mutated_gene_df	tbl_df	44	3.5 MB	5537 obs. of 44 variables
raw_gene_df	tbl_df	40	3.3 MB	5537 obs. of 40 variables
selected_gene_df	tbl_df	40	2.4 MB	5537 obs. of 40 variables
separated_gene...	tbl_df	44	3.6 MB	5537 obs. of 44 variables

Files Plots Packages Help Viewer

New Folder | Delete | Rename | More

Home > R_Users_Workshop > 8_weeks_Oct-Dec_17 > Workshop_1 > workshop_1_project

Name	Size	Modified
..		
.RData	2.5 KB	Oct 2, 2017, 1:49 PM
.Rhistory	20.3 KB	Dec 6, 2017, 3:43 PM
Brauer2008_DataSet1.csv	1.6 MB	Sep 27, 2017, 11:32 PM
Brauer2008_DataSet1.tds	1.6 MB	Sep 28, 2017, 10:22 AM
house_completions.csv	4 KB	Sep 28, 2017, 1:35 PM
irish_population.csv	315 B	Aug 28, 2017, 4:21 PM
raw_house_completions.csv	16.2 KB	Aug 25, 2017, 3:45 PM
workshop_1.Rproj	217 B	Oct 18, 2018, 12:18 PM
ws1_script1_stepwise_Bauer_dataset_analysis.R	6.1 KB	Dec 5, 2017, 12:19 PM
ws1_script2_Bauer_dataset_analysis.R	2 KB	Dec 6, 2017, 2:33 PM
ws1_script3_house_completions.R	2.4 KB	Oct 2, 2017, 3:53 PM

Line by line

~/R_Users_Workshop/8_weeks_Oct-Dec_17/Workshop_1/workshop_1_project - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | workshop_1_project — 8_weeks_Oct-Dec_17

ws1_script1_stepwise_Bauer_dataset_an... *

Source on Save | Run | Source

```
12
13  raw_gene_df <- read_delim("Brauer2008_DataSet1.tds", delim = "\t")
14
15  separated_gene_df <- separate(raw_gene_df, NAME,
16                                c("name", "BP", "MF", "systematic_name",
17                                  "number"),
18                                sep = "\\|\\|")
19
20  mutated_gene_df <- mutate_at(separated_gene_df,
21                               vars(name:systematic_name),
22                               funs(trimws)
23                               )
24
25  selected_gene_df <- select(mutated_gene_df, -number, -GID, -YORF, -GWEIGHT)
26
27  gathered_gene_df <- gather(selected_gene_df, sample, expression, G0.05:U0.3)
28
29  nearly_there_df <- separate(gathered_gene_df, sample,
30                              c("nutrient", "rate"), sep = 1, convert = TRUE)
31
32  nutrient_names <- c(G = "Glucose", L = "Leucine", P = "Phosphate",
33                      S = "Sulfate", N = "Ammonia", U = "Uracil")
34
35  cleaned_genes_df <- mutate(nearly_there_df,
36                             nutrient = plyr::revalue(nutrient, nutrient_names)
37                             ) %>%
38    filter(!is.na(expression), systematic_name != "")
```

29:1 | Section 1: Data import, tidying and transformation | R Script

Console Terminal

~/R_Users_Workshop/8_weeks_Oct-Dec_17/Workshop_1/workshop_1_project/

```
cols(
  .default = col_double(),
  GID = col_character(),
  YORF = col_character(),
  NAME = col_character(),
  GWEIGHT = col_integer()
)
See spec(...) for full column specifications.
> separated_gene_df <- separate(raw_gene_df, NAME,
+                               c("name", "BP", "MF", "systematic_name",
+                                 "number"),
+                               sep = "\\|\\|")
> mutated_gene_df <- mutate_at(separated_gene_df,
+                              vars(name:systematic_name),
+                              funs(trimws)
+                              )
> selected_gene_df <- select(mutated_gene_df, -number, -GID, -YORF, -GWEIGHT)
> gathered_gene_df <- gather(selected_gene_df, sample, expression, G0.05:U0.3)
>
```

Environment History Connections

Import Dataset | Grid

Global Environment

Name	Type	Length	Size	Value
gathered_gene_df	tbl_df	6	9.8 MB	199332 obs. of 6 variables
mutated_gene_df	tbl_df	44	3.5 MB	5537 obs. of 44 variables
raw_gene_df	tbl_df	40	3.3 MB	5537 obs. of 40 variables
selected_gene_df	tbl_df	40	2.4 MB	5537 obs. of 40 variables
separated_gene...	tbl_df	44	3.6 MB	5537 obs. of 44 variables

Files Plots Packages Help Viewer

New Folder | Delete | Rename | More

Home > R_Users_Workshop > 8_weeks_Oct-Dec_17 > Workshop_1 > workshop_1_project

Name	Size	Modified
..		
.RData	2.5 KB	Oct 2, 2017, 1:49 PM
.Rhistory	20.3 KB	Dec 6, 2017, 3:43 PM
Brauer2008_DataSet1.csv	1.6 MB	Sep 27, 2017, 11:32 PM
Brauer2008_DataSet1.tds	1.6 MB	Sep 28, 2017, 10:22 AM
house_completions.csv	4 KB	Sep 28, 2017, 1:35 PM
irish_population.csv	315 B	Aug 28, 2017, 4:21 PM
raw_house_completions.csv	16.2 KB	Aug 25, 2017, 3:45 PM
workshop_1.Rproj	217 B	Oct 18, 2018, 12:18 PM
ws1_script1_stepwise_Bauer_dataset_analysis.R	6.1 KB	Dec 5, 2017, 12:19 PM
ws1_script2_Bauer_dataset_analysis.R	2 KB	Dec 6, 2017, 2:33 PM
ws1_script3_house_completions.R	2.4 KB	Oct 2, 2017, 3:53 PM

Line by line

~/R_Users_Workshop/8_weeks_Oct-Dec_17/Workshop_1/workshop_1_project - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | workshop_1_project — 8_weeks_Oct-Dec_17

ws1_script1_stepwise_Bauer_dataset_an... *

Source on Save | Run | Source

```
12
13  raw_gene_df <- read_delim("Brauer2008_DataSet1.tds", delim = "\t")
14
15  separated_gene_df <- separate(raw_gene_df, NAME,
16                                c("name", "BP", "MF", "systematic_name",
17                                  "number"),
18                                sep = "\\|\\|")
19
20  mutated_gene_df <- mutate_at(separated_gene_df,
21                               vars(name:systematic_name),
22                               funs(trimws)
23                               )
24
25  selected_gene_df <- select(mutated_gene_df, -number, -GID, -YORF, -GWEIGHT)
26
27  gathered_gene_df <- gather(selected_gene_df, sample, expression, G0.05:U0.3)
28
29  nearly_there_df <- separate(gathered_gene_df, sample,
30                              c("nutrient", "rate"), sep = 1, convert = TRUE)
31
32  nutrient_names <- c(G = "Glucose", L = "Leucine", P = "Phosphate",
33                      S = "Sulfate", N = "Ammonia", U = "Uracil")
34
35  cleaned_genes_df <- mutate(nearly_there_df,
36                             nutrient = plyr::revalue(nutrient, nutrient_names)
37                             ) %>%
38    filter(!is.na(expression), systematic_name != "")
```

32:1 | Section 1: Data import, tidying and transformation | R Script

Console Terminal

~/R_Users_Workshop/8_weeks_Oct-Dec_17/Workshop_1/workshop_1_project/

```
  GID = col_character(),
  YORF = col_character(),
  NAME = col_character(),
  GWEIGHT = col_integer()
)
See spec(...) for full column specifications.
> separated_gene_df <- separate(raw_gene_df, NAME,
+                               c("name", "BP", "MF", "systematic_name",
+                                 "number"),
+                               sep = "\\|\\|")
> mutated_gene_df <- mutate_at(separated_gene_df,
+                              vars(name:systematic_name),
+                              funs(trimws)
+                              )
> selected_gene_df <- select(mutated_gene_df, -number, -GID, -YORF, -GWEIGHT)
> gathered_gene_df <- gather(selected_gene_df, sample, expression, G0.05:U0.3)
> nearly_there_df <- separate(gathered_gene_df, sample,
+                             c("nutrient", "rate"), sep = 1, convert = TRUE)
> |
```

Environment History Connections

Import Dataset | Grid

Global Environment

Name	Type	Length	Size	Value
gathered_gene_df	tbl_df	6	9.8 MB	199332 obs. of 6 variables
mutated_gene_df	tbl_df	44	3.5 MB	5537 obs. of 44 variables
nearly_there_df	tbl_df	7	11.3 MB	199332 obs. of 7 variables
raw_gene_df	tbl_df	40	3.3 MB	5537 obs. of 40 variables
selected_gene_df	tbl_df	40	2.4 MB	5537 obs. of 40 variables
separated_gene...	tbl_df	44	3.6 MB	5537 obs. of 44 variables

Files Plots Packages Help Viewer

New Folder | Delete | Rename | More

Home > R_Users_Workshop > 8_weeks_Oct-Dec_17 > Workshop_1 > workshop_1_project

Name	Size	Modified
..		
.RData	2.5 KB	Oct 2, 2017, 1:49 PM
.Rhistory	20.3 KB	Dec 6, 2017, 3:43 PM
Brauer2008_DataSet1.csv	1.6 MB	Sep 27, 2017, 11:32 PM
Brauer2008_DataSet1.tds	1.6 MB	Sep 28, 2017, 10:22 AM
house_completions.csv	4 KB	Sep 28, 2017, 1:35 PM
irish_population.csv	315 B	Aug 28, 2017, 4:21 PM
raw_house_completions.csv	16.2 KB	Aug 25, 2017, 3:45 PM
workshop_1.Rproj	217 B	Oct 18, 2018, 12:18 PM
ws1_script1_stepwise_Bauer_dataset_analysis.R	6.1 KB	Dec 5, 2017, 12:19 PM
ws1_script2_Bauer_dataset_analysis.R	2 KB	Dec 6, 2017, 2:33 PM
ws1_script3_house_completions.R	2.4 KB	Oct 2, 2017, 3:53 PM

Line by line

~/R_Users_Workshop/8_weeks_Oct-Dec_17/Workshop_1/workshop_1_project - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | workshop_1_project — 8_weeks_Oct-Dec_17

ws1_script1_stepwise_Bauer_dataset_an... *

Source on Save | Run | Source

```
12
13  raw_gene_df <- read_delim("Brauer2008_DataSet1.tds", delim = "\t")
14
15  separated_gene_df <- separate(raw_gene_df, NAME,
16                                c("name", "BP", "MF", "systematic_name",
17                                  "number"),
18                                sep = "\\|\\|")
19
20  mutated_gene_df <- mutate_at(separated_gene_df,
21                               vars(name:systematic_name),
22                               funs(trimws)
23                               )
24
25  selected_gene_df <- select(mutated_gene_df, -number, -GID, -YORF, -GWEIGHT)
26
27  gathered_gene_df <- gather(selected_gene_df, sample, expression, G0.05:U0.3)
28
29  nearly_there_df <- separate(gathered_gene_df, sample,
30                              c("nutrient", "rate"), sep = 1, convert = TRUE)
31
32  nutrient_names <- c(G = "Glucose", L = "Leucine", P = "Phosphate",
33                      S = "Sulfate", N = "Ammonia", U = "Uracil")
34
35  cleaned_genes_df <- mutate(nearly_there_df,
36                             nutrient = plyr::revalue(nutrient, nutrient_names)
37                             ) %>%
38    filter(!is.na(expression), systematic_name != "")
```

35:1 | Section 1: Data import, tidying and transformation | R Script

Console Terminal

~/R_Users_Workshop/8_weeks_Oct-Dec_17/Workshop_1/workshop_1_project/

```
  NAME = col_character(),
  GWEIGHT = col_integer()
)
See spec(...) for full column specifications.
> separated_gene_df <- separate(raw_gene_df, NAME,
+                               c("name", "BP", "MF", "systematic_name",
+                                 "number"),
+                               sep = "\\|\\|")
> mutated_gene_df <- mutate_at(separated_gene_df,
+                              vars(name:systematic_name),
+                              funs(trimws)
+                              )
> selected_gene_df <- select(mutated_gene_df, -number, -GID, -YORF, -GWEIGHT)
> gathered_gene_df <- gather(selected_gene_df, sample, expression, G0.05:U0.3)
> nearly_there_df <- separate(gathered_gene_df, sample,
+                             c("nutrient", "rate"), sep = 1, convert = TRUE)
> nutrient_names <- c(G = "Glucose", L = "Leucine", P = "Phosphate",
+                     S = "Sulfate", N = "Ammonia", U = "Uracil")
> |
```

Environment | History | Connections

Import Dataset | Grid

Global Environment

Name	Type	Length	Size	Value
gathered_gene_df	tbl_df	6	9.8 MB	199332 obs. of 6 variables
mutated_gene_df	tbl_df	44	3.5 MB	5537 obs. of 44 variables
nearly_there_df	tbl_df	7	11.3 MB	199332 obs. of 7 variables
nutrient_names	character	6	984 B	Named chr [1:6] "Glucose" "...
raw_gene_df	tbl_df	40	3.3 MB	5537 obs. of 40 variables
selected_gene_df	tbl_df	40	2.4 MB	5537 obs. of 40 variables
separated_gene...	tbl_df	44	3.6 MB	5537 obs. of 44 variables

Files | Plots | Packages | Help | Viewer

New Folder | Delete | Rename | More

Home > R_Users_Workshop > 8_weeks_Oct-Dec_17 > Workshop_1 > workshop_1_project

Name	Size	Modified
..		
.RData	2.5 KB	Oct 2, 2017, 1:49 PM
.Rhistory	20.3 KB	Dec 6, 2017, 3:43 PM
Brauer2008_DataSet1.csv	1.6 MB	Sep 27, 2017, 11:32 PM
Brauer2008_DataSet1.tds	1.6 MB	Sep 28, 2017, 10:22 AM
house_completions.csv	4 KB	Sep 28, 2017, 1:35 PM
irish_population.csv	315 B	Aug 28, 2017, 4:21 PM
raw_house_completions.csv	16.2 KB	Aug 25, 2017, 3:45 PM
workshop_1.Rproj	217 B	Oct 18, 2018, 12:18 PM
ws1_script1_stepwise_Bauer_dataset_analysis.R	6.1 KB	Dec 5, 2017, 12:19 PM
ws1_script2_Bauer_dataset_analysis.R	2 KB	Dec 6, 2017, 2:33 PM
ws1_script3_house_completions.R	2.4 KB	Oct 2, 2017, 3:53 PM

Line by line

~/R_Users_Workshop/8_weeks_Oct-Dec_17/Workshop_1/workshop_1_project - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | workshop_1_project — 8_weeks_Oct-Dec_17

ws1_script1_stepwise_Bauer_dataset_an... *

Source on Save | Run | Source

```
15 separated_gene_df <- separate(raw_gene_df, NAME,
16                               c("name", "BP", "MF", "systematic_name",
17                                 "number"),
18                               sep = "\\|\\|")
19
20 mutated_gene_df <- mutate_at(separated_gene_df,
21                              vars(name:systematic_name),
22                              funs(trimws)
23                              )
24
25 selected_gene_df <- select(mutated_gene_df, -number, -GID, -YORF, -GWEIGHT)
26
27 gathered_gene_df <- gather(selected_gene_df, sample, expression, G0.05:U0.3)
28
29 nearly_there_df <- separate(gathered_gene_df, sample,
30                              c("nutrient", "rate"), sep = 1, convert = TRUE)
31
32 nutrient_names <- c(G = "Glucose", L = "Leucine", P = "Phosphate",
33                     S = "Sulfate", N = "Ammonia", U = "Uracil")
34
35 cleaned_genes_df <- mutate(nearly_there_df,
36                            nutrient = plyr::revalue(nutrient, nutrient_names)
37                            ) %>%
38   filter(!is.na(expression), systematic_name != "")
39
40
41
```

44:1 | Section 1: Data import, tidying and transformation | R Script

Console Terminal

~/R_Users_Workshop/8_weeks_Oct-Dec_17/Workshop_1/workshop_1_project/

```
> separated_gene_df <- separate(raw_gene_df, NAME,
+                               c("name", "BP", "MF", "systematic_name",
+                                 "number"),
+                               sep = "\\|\\|")
> mutated_gene_df <- mutate_at(separated_gene_df,
+                              vars(name:systematic_name),
+                              funs(trimws)
+                              )
> selected_gene_df <- select(mutated_gene_df, -number, -GID, -YORF, -GWEIGHT)
> gathered_gene_df <- gather(selected_gene_df, sample, expression, G0.05:U0.3)
> nearly_there_df <- separate(gathered_gene_df, sample,
+                             c("nutrient", "rate"), sep = 1, convert = TRUE)
> nutrient_names <- c(G = "Glucose", L = "Leucine", P = "Phosphate",
+                     S = "Sulfate", N = "Ammonia", U = "Uracil")
> cleaned_genes_df <- mutate(nearly_there_df,
+                            nutrient = plyr::revalue(nutrient, nutrient_names)
+                            ) %>%
+   filter(!is.na(expression), systematic_name != "")
>
```

Environment History Connections

Import Dataset | Grid

Global Environment

Name	Type	Length	Size	Value
cleaned_genes_df	tbl_df	7	11.3 MB	198430 obs. of 7 variables
gathered_gene_df	tbl_df	6	9.8 MB	199332 obs. of 6 variables
mutated_gene_df	tbl_df	44	3.5 MB	5537 obs. of 44 variables
nearly_there_df	tbl_df	7	11.3 MB	199332 obs. of 7 variables
nutrient_names	character	6	984 B	Named chr [1:6] "Glucose" "...
raw_gene_df	tbl_df	40	3.3 MB	5537 obs. of 40 variables
selected_gene_df	tbl_df	40	2.4 MB	5537 obs. of 40 variables
separated_gene...	tbl_df	44	3.6 MB	5537 obs. of 44 variables

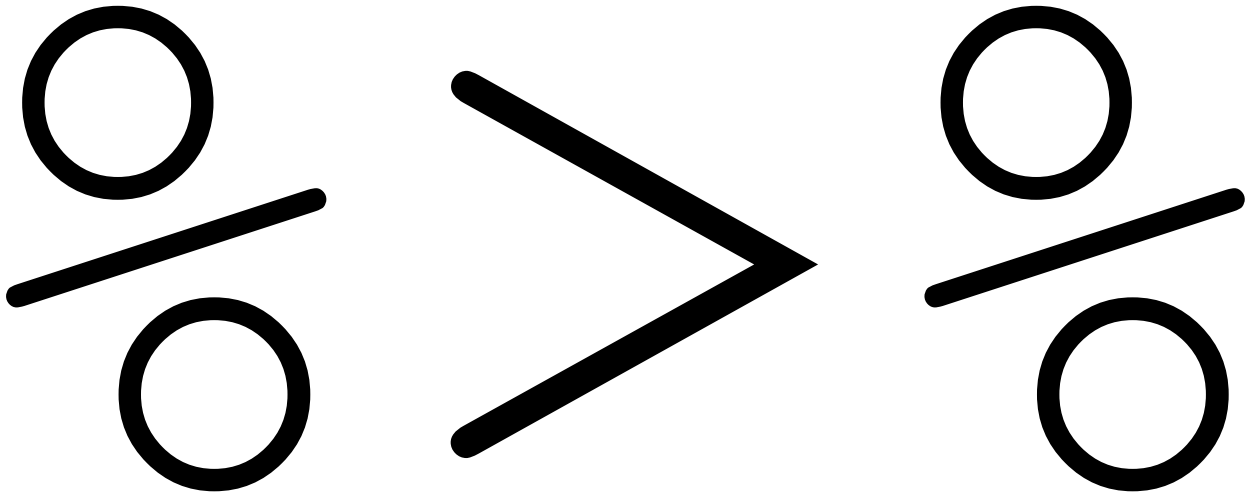
Files Plots Packages Help Viewer

New Folder | Delete | Rename | More

Home > R_Users_Workshop > 8_weeks_Oct-Dec_17 > Workshop_1 > workshop_1_project

Name	Size	Modified
..		
.RData	2.5 KB	Oct 2, 2017, 1:49 PM
.Rhistory	20.3 KB	Dec 6, 2017, 3:43 PM
Brauer2008_DataSet1.csv	1.6 MB	Sep 27, 2017, 11:32 PM
Brauer2008_DataSet1.tds	1.6 MB	Sep 28, 2017, 10:22 AM
house_completions.csv	4 KB	Sep 28, 2017, 1:35 PM
irish_population.csv	315 B	Aug 28, 2017, 4:21 PM
raw_house_completions.csv	16.2 KB	Aug 25, 2017, 3:45 PM
workshop_1.Rproj	217 B	Oct 18, 2018, 12:18 PM
ws1_script1_stepwise_Bauer_dataset_analysis.R	6.1 KB	Dec 5, 2017, 12:19 PM
ws1_script2_Bauer_dataset_analysis.R	2 KB	Dec 6, 2017, 2:33 PM
ws1_script3_house_completions.R	2.4 KB	Oct 2, 2017, 3:53 PM

Putting the pieces together



Tidyverse code structure

new_object - assign the output to a new object
<- - the assign operator
input_data - data to be manipulated
%>% - the magrittr/pipe operator
function - the function you are calling on
data_ - elements of the input data to use
arguments_ - how you want to apply the function

Piped

~/R_Users_Workshop/8_weeks_Oct-Dec_17/Workshop_1/workshop_1_project - RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function | Addins | workshop_1_project — 8_weeks_Oct-Dec_17

ws1_script1_stepwise_Bauer_dataset_an... | ws1_script2_Bauer_dataset_analysis.R*

Source on Save | Run | Source

```
1  nutrient_names <- c(G = "Glucose", L = "Leucine", P = "Phosphate",
2                      S = "Sulfate", N = "Ammonia", U = "Uracil"
3                      )
4
5  cleaned_genes_df <- read_delim("Brauer2008_DataSet1.tds", delim = "\t"
6                                 ) %>%
7
8    separate(NAME, c("name", "BP", "MF", "systematic_name", "number"), sep = "\\|\\|"
9             )  %>%
10
11   mutate_at(vars(name:systematic_name), funs(trimws)
12             ) %>%
13
14   select(-number, -GID, -YORF, -GWEIGHT
15          ) %>%
16
17   gather(sample, expression, G0.05:U0.3
18          ) %>%
19
20   separate(sample, c("nutrient", "rate"), sep = 1, convert = TRUE
21            ) %>%
22
23   mutate(nutrient = plyr::revalue(nutrient, nutrient_names)
24          ) %>%
25
26   filter(!is.na(expression), systematic_name != ""
27          )
```

9:18 (Top Level) R Script

Console | Terminal

~/R_Users_Workshop/8_weeks_Oct-Dec_17/Workshop_1/workshop_1_project/

```
+
+   separate(sample, c("nutrient", "rate"), sep = 1, convert = TRUE
+            ) %>%
+
+   mutate(nutrient = plyr::revalue(nutrient, nutrient_names)
+          ) %>%
+
+   filter(!is.na(expression), systematic_name != ""
+          )
Parsed with column specification:
cols(
  .default = col_double(),
  GID = col_character(),
  YORF = col_character(),
  NAME = col_character(),
  GWEIGHT = col_integer()
)
See spec(...) for full column specifications.
>
```

Environment | History | Connections

Import Dataset | Grid

Global Environment

Name	Type	Length	Size	Value
cleaned_genes_df	tbl_df	7	11.3 MB	198430 obs. of 7 variables
nutrient_names	character	6	984 B	Named chr [1:6] "Glucose" "Le...

Files | Plots | Packages | Help | Viewer

New Folder | Delete | Rename | More

Home > R_Users_Workshop > 8_weeks_Oct-Dec_17 > Workshop_1 > workshop_1_project

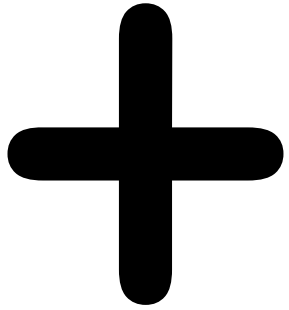
Name	Size	Modified
..		
.RData	2.5 KB	Oct 2, 2017, 1:49 PM
.Rhistory	20.3 KB	Dec 6, 2017, 3:43 PM
Brauer2008_DataSet1.csv	1.6 MB	Sep 27, 2017, 11:32 PM
Brauer2008_DataSet1.tds	1.6 MB	Sep 28, 2017, 10:22 AM
house_completions.csv	4 KB	Sep 28, 2017, 1:35 PM
irish_population.csv	315 B	Aug 28, 2017, 4:21 PM
raw_house_completions.csv	16.2 KB	Aug 25, 2017, 3:45 PM
workshop_1.Rproj	217 B	Oct 18, 2018, 12:18 PM
ws1_script1_stepwise_Bauer_dataset_analysis.R	6.1 KB	Dec 5, 2017, 12:19 PM
ws1_script2_Bauer_dataset_analysis.R	2 KB	Dec 6, 2017, 2:33 PM
ws1_script3_house_completions.R	2.4 KB	Oct 2, 2017, 3:53 PM

Piped

```
nutrient_names <- c(G = "Glucose", L = "Leucine", P = "Phosphate",
                    S = "Sulfate", N = "Ammonia", U = "Uracil"
                    )

cleaned_genes_df <- read_delim("Brauer2008_DataSet1.tds", delim = "\t"
                               ) %>%

  separate(NAME, c("name", "BP", "MF", "systematic_name", "number"), sep = "\\|\\|"
           )  %>%

  mutate_at(vars(name:systematic_name), funs(trimws)
            ) %>%

  select(-number, -GID, -YORF, -GWEIGHT
         ) %>%

  gather(sample, expression, G0.05:U0.3
         ) %>%

  separate(sample, c("nutrient", "rate"), sep = 1, convert = TRUE
           ) %>%

  mutate(nutrient = plyr::revalue(nutrient, nutrient_names)
         ) %>%

  filter(!is.na(expression), systematic_name != ""
         )
```

The moral of the story…….

You can go from this

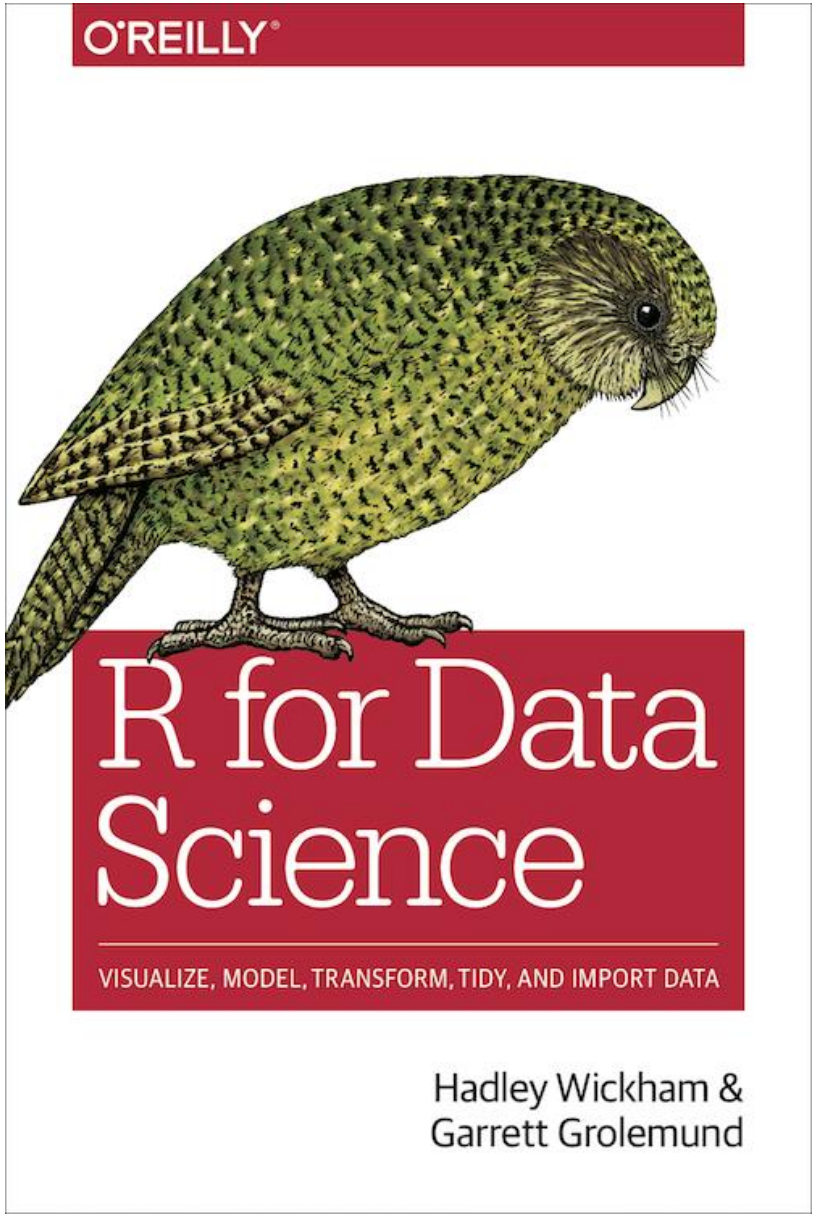


To this!!

Completely ordinary

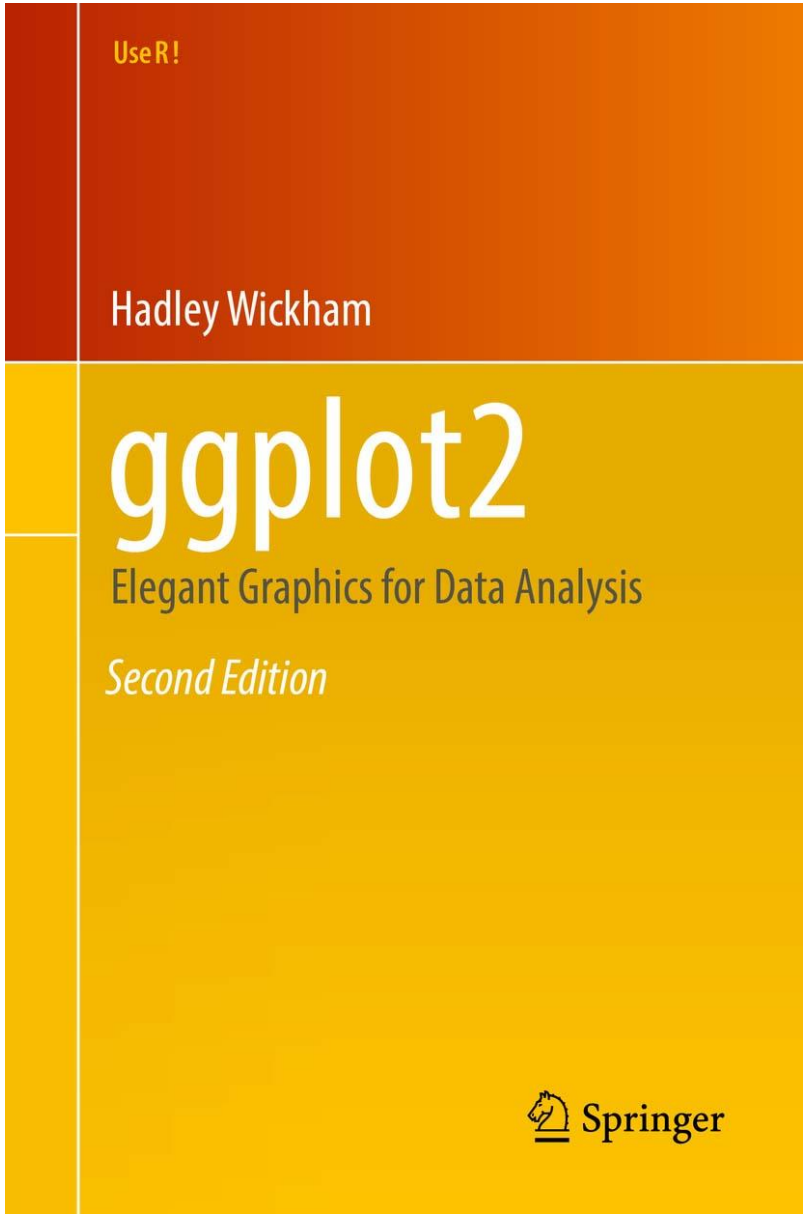
Master Builder!

You could write a book on that!!



R for Data Science webpage

And on this!!


Use R!
Hadley Wickham
ggplot2
Elegant Graphics for Data Analysis
Second Edition
Springer

Explore the tidyverse example scripts

- Open the scripts in the tidyverse folder
- Make sure you are using the correct R-project file

R Markdown

R Markdown

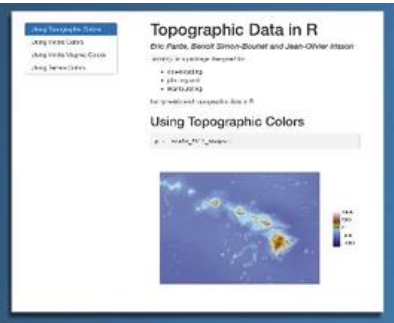
- R Markdown combines the code you wrote, the output produced and you own comments

- You can view it as a digital lab notebook, where you are both recording what you're doing, and what you were thinking while you were doing it!

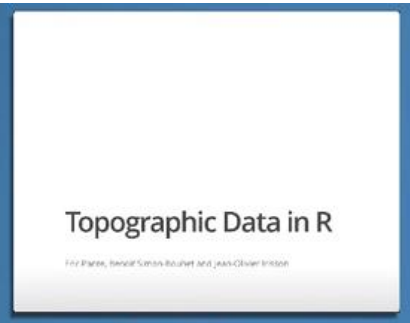
- R Markdown outputs can take many forms
    - Word documents, PDFs, slideshows etc.

- Once created the .Rmd file get sent to knitr, which executes the chunks of code and creates a new markdown document
    - this is then processed by pandoc which creates the finished file
        - knitr and pandoc are external websites

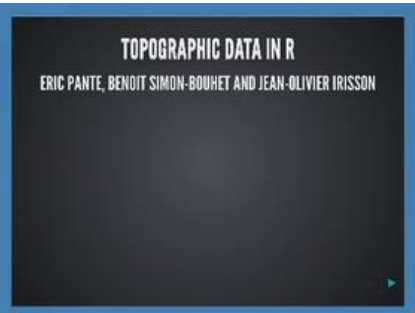
What has R Markdown ever done for us?



html

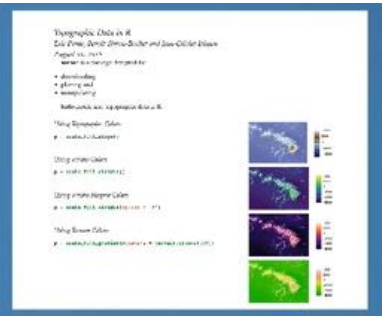


ioslides

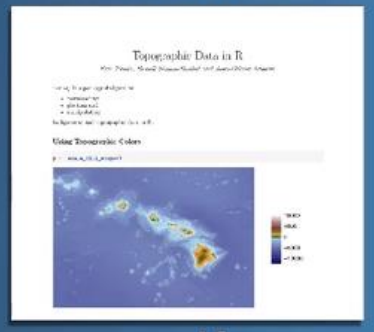


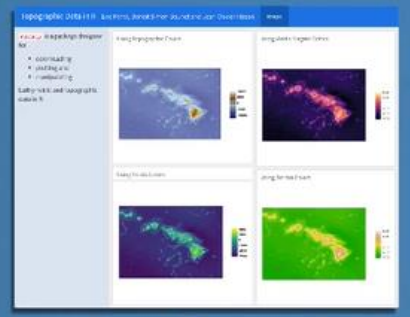
reveal.js


rtf

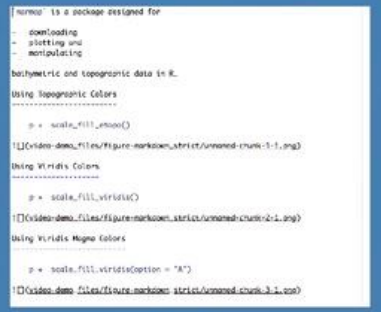

tufte handout

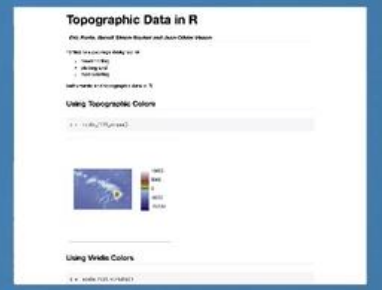
book


pdf

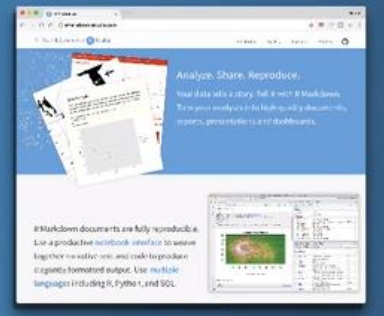

dashboard

slidy

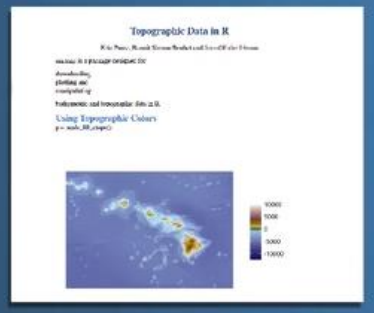

markdown

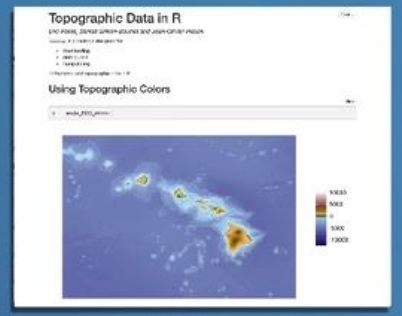


package vignette

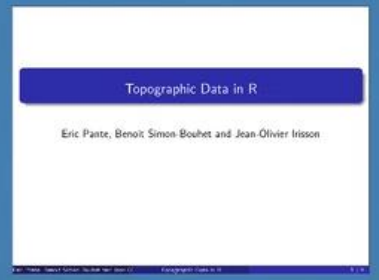


website


Word



notebook



beamer


latex

custom template


shiny app

R Markdown

YAML header

```
---
title: "Diamond sizes"
date: 2016-08-25
output: html_document
---
```

Chunks of code

````
```{r setup, include = FALSE}
library(ggplot2)
library(dplyr)
smaller <- diamonds %>%
filter(carat <= 2.5)
```
````

Plain text with data outputs from R code

````
We have data about `r nrow(diamonds)`
diamonds. Only
`r nrow(diamonds) - nrow(smaller)` are
larger than
2.5 carats. The distribution of the
remainder is shown below:
````

Chunks of code

````
```{r, echo = FALSE}
smaller %>%
ggplot(aes(carat)) +
geom_freqpoly(binwidth = 0.01)
```
````

Introduction to R Markdown

- We're now going to look at a R Markdown file that provides some of the tips and tricks you'll need yourselves
    - Code chunks
    - Formatting
    - Tables
    - Figures etc.

- Open the R Markdown file intro_to_RMarkdown.rmd

Where to next?

- Understanding basic statistical concepts
www.khanacademy.org

- Collection of YouTube videos describing statistics through R
http://rafalab.github.io/pages/harvardx.html

- You know what you want to do, but don't know how to do it
https://stats.stackexchange.com/

Structured training……

coursera

Catalog | statistics with r

Fo

Course Languages

- [ ] English 557
- [ ] Spanish 12
- [ ] Chinese (Simplified) 7

Show More

Subtitle Languages

- [x] English 586
- [ ] Chinese (Simplified) 62
- [ ] Spanish 60

Show More

All Topics

- [ ] Data Science 212
- [ ] Business 207
- [ ] Computer Science 162

Show More

You searched for **statistics with r.** 586 matches

Active filters: English ×

Courses and Specializations

Statistics with R

5-course Specialization · Duke University

The R Programming Environment

Johns Hopkins University

Basic Statistics

University of Amsterdam

Advanced Linear Models for Data Science 2: Statistical Linear Models

Johns Hopkins University

Viewing 42 results matching

Search:

"statistics r" ✖ CLEAR ALL

Refine your search

Availability

Current	20
Starting Soon	7
Upcoming	5
Self-Paced	24
Archived	8

Subjects

Biology & Life Sciences	9
Business & Management	3
Computer Science	12
Data Analysis & Statistics	

Statistics and R	HarvardX	Course	7/12/2017
Explore Statistics with R	KIx	Course	7/7/2015
Introduction to R for Data Science	Microsoft	Course	10/1/2017
Programming with R for Data Science	Microsoft	Course	10/1/2017
Analyzing Big Data with Microsoft R	Microsoft	Course	10/1/2017
Statistical Analysis in Bioinformatics	USMx	Course	10/23/2017

PAID COURSE

Introduction to the Tidyverse

Start Course For Free | ▷ Play Intro Video

4 hours | 16 Videos | 50 Exercises | 10,553 Participants | 4,150 XP

Course Description

This is an introduction to the programming language R, focused on a powerful set of tools known as the "tidyverse". In the course you'll learn the intertwined processes of data manipulation and visualization through the tools dplyr and ggplot2. You'll learn to manipulate data by filtering, sorting and summarizing a real dataset of historical country data in order to answer exploratory questions. You'll then learn to turn this processed data into informative line plots, bar plots, histograms, and more with the ggplot2 package. This gives a taste both of the value of exploratory data analysis and the power of tidyverse tools. This is a suitable introduction for people who have no previous experience in R and are interested in learning to perform data analysis.

David Robinson
Chief Data Scientist, DataCamp

meetup

Cork (Ireland) R-Users Group