

R: The tidyverse and beyond

1-day workshop
Afternoon lecture

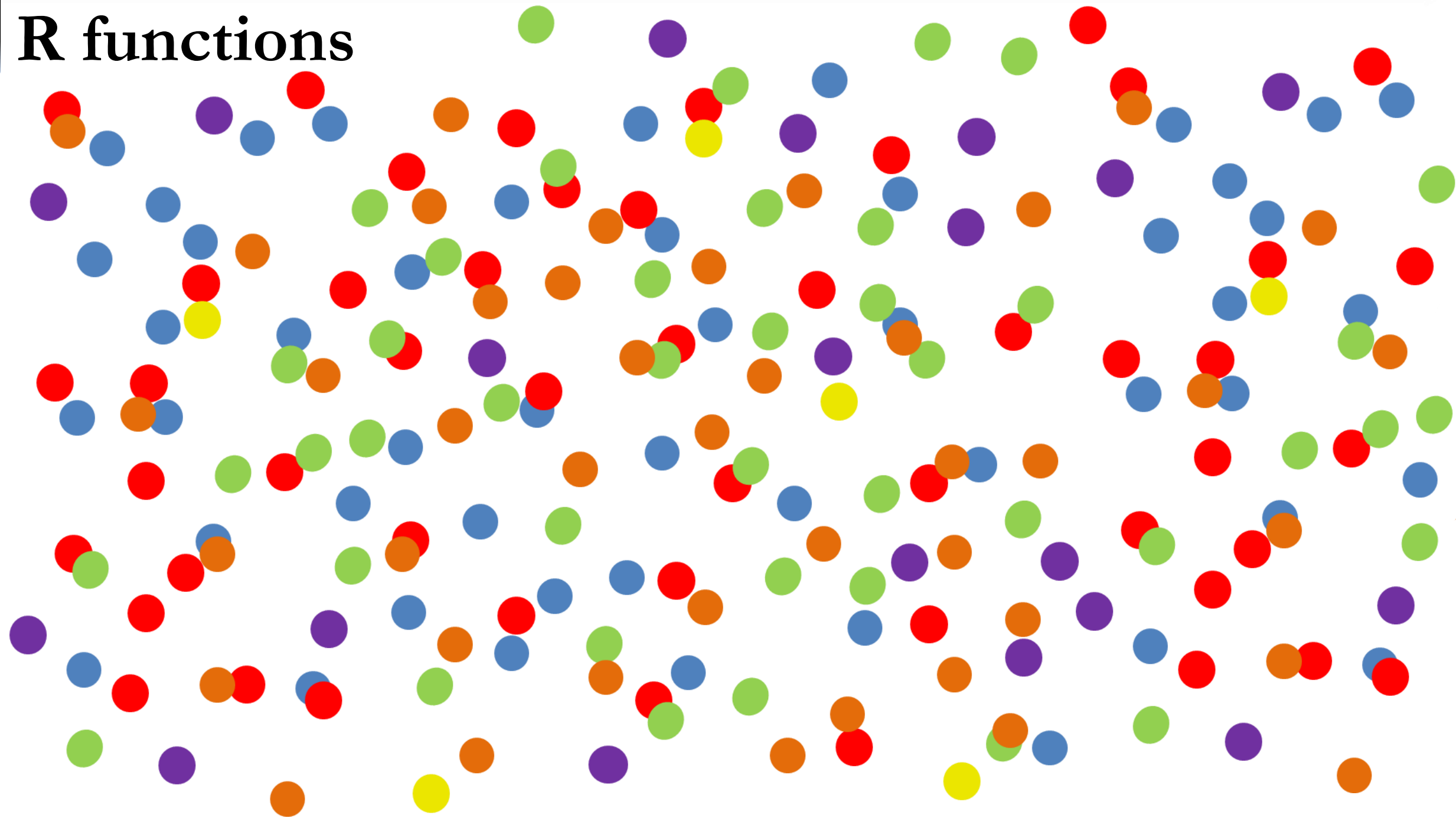


Brendan Palmer,
Statistics & Data Analysis Unit,
Clinical Research Facility - Cork

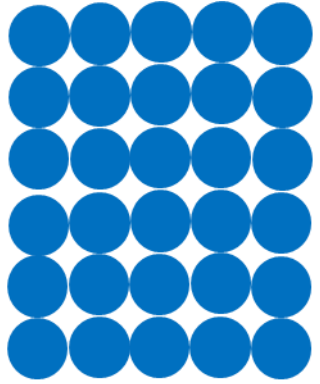
To understand R, remember the following

- Everything that exists is an object
- Everything that happens is a function

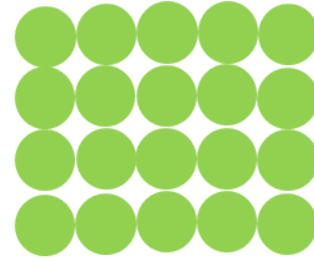
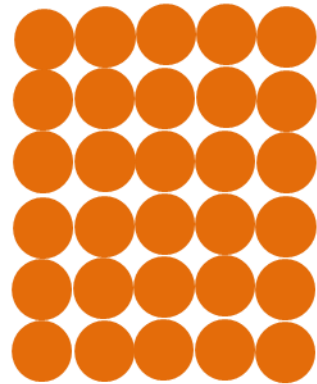
R functions



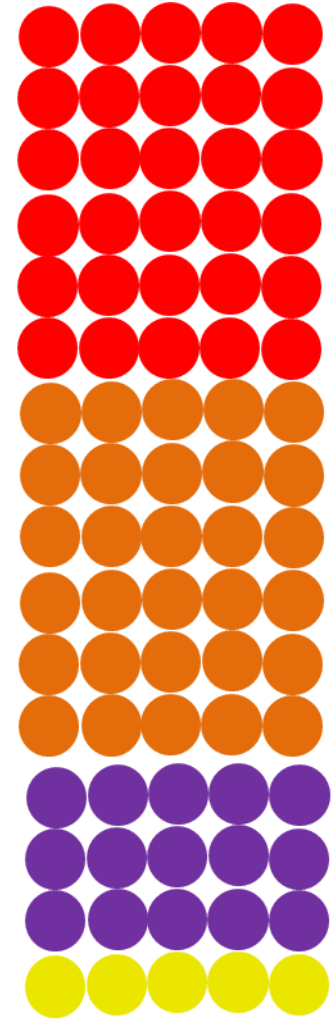
R packages



Base R:
Comes
pre-
loaded

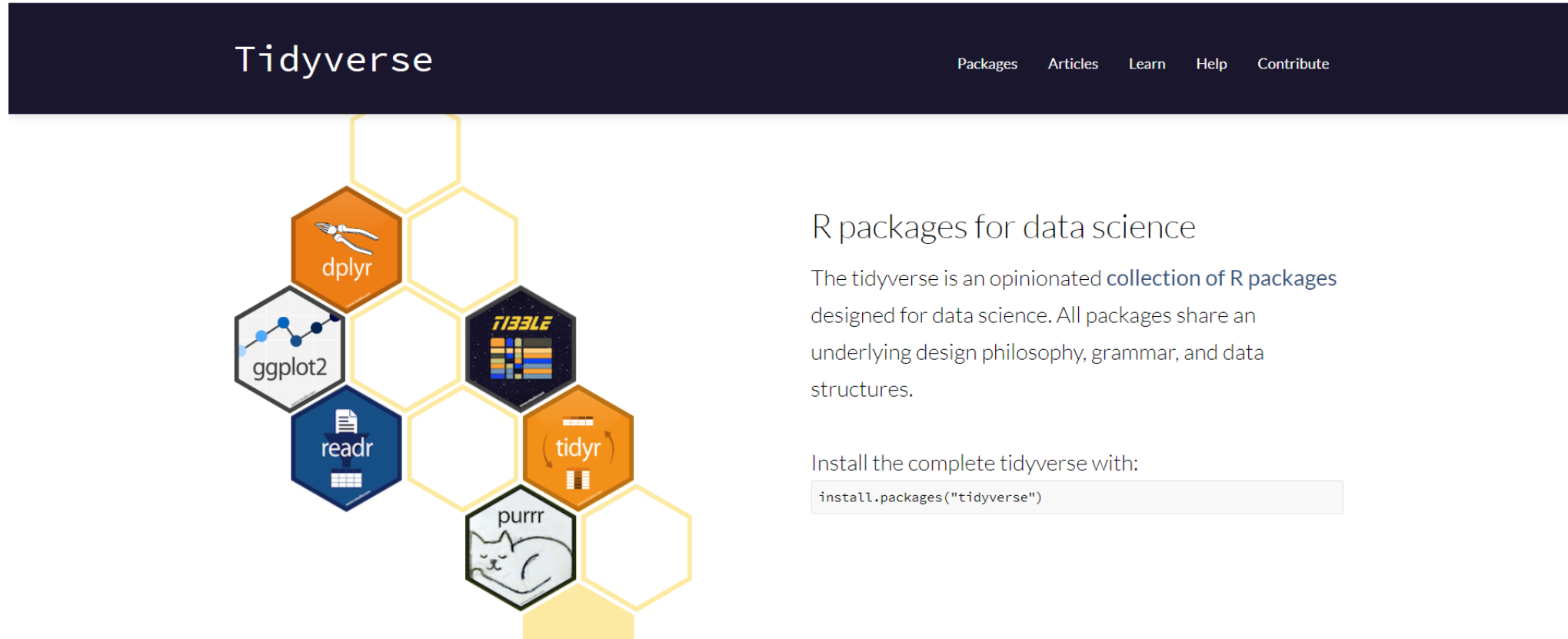


Other packages:
Install once
Update regularly
Load each session



core
tidyverse

What is the tidyverse?

A screenshot of the Tidyverse website. The header is dark blue with the word "Tidyverse" in white. To the right of the header are links for "Packages", "Articles", "Learn", "Help", and "Contribute". Below the header is a large graphic of a honeycomb grid. Several hexagons are filled with icons and package names: "dplyr" (orange, top left), "ggplot2" (grey, middle left), "readr" (blue, bottom left), "tidyr" (orange, middle right), "purrr" (grey, bottom right), and "TIBBLE" (dark blue, top right). To the right of the honeycomb graphic, the text "R packages for data science" is followed by a paragraph: "The tidyverse is an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures." Below this is the text "Install the complete tidyverse with:" followed by a code block containing the command `install.packages("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")
```

The tidyverse Oct 2017

```
> library(tidyverse)
```

```
Loading tidyverse: ggplot2
```

```
Loading tidyverse: tibble
```

```
Loading tidyverse: tidyr
```

```
Loading tidyverse: readr
```

```
Loading tidyverse: purrr
```

```
Loading tidyverse: dplyr
```

Data visualisation

Modern version of data frames

Data tidying

Data import

Functional programming

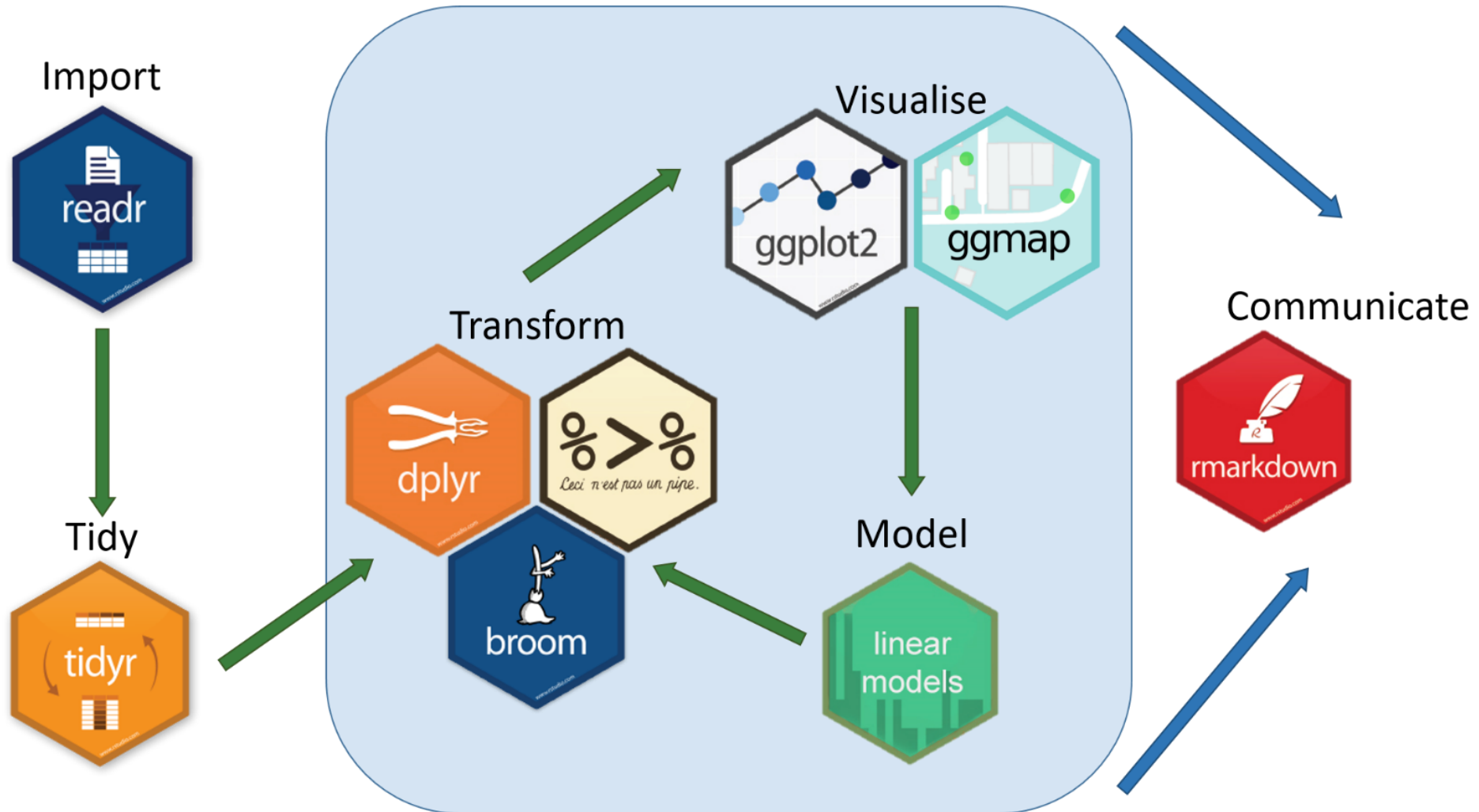
Data manipulation

The tidyverse Oct 2018

```
> library("tidyverse")
-- Attaching packages ----- tidyverse 1.2.1 --
v ggplot2 3.0.0      v purrr  0.2.5
v tibble  1.4.2      v dplyr  0.7.6
v tidyr   0.8.1      v stringr 1.3.1
v readr   1.1.1      v forcats 0.3.0
-- Conflicts ----- tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()     masks stats::lag()
> |
```


Putting the pieces together

- Data analysis in a tidyverse nutshell



Tidyverse works best with tidy data

- Each variable forms a column
- Each observation forms a row

Problems with Brauer et al., data...

Column headers contain values

Multiple variables are stored in one column

e.g. column "NAME" contains values such as;

SFB2 || ER to Golgi transport || molecular function unknown || YNL049C || 1082129

These need to be split up

- G0.05 - letter identifies a compound
- number is the concentration of that compound

Code structure v1

```
separated_gene <- separate(raw_gene, NAME, c("name", "BP", "MF", "systematic_name", "number"), sep = "\\|\\|")
```

- | | |
|---|---------------------------------------|
| separated_gene | - the new tibble you will create |
| <- | - the assign operator |
| separate | - the function you are calling on |
| (raw_gene, | - the tibble to be used |
| NAME, | - the column to be altered |
| c("name", "BP", "MF", "systematic_name", "number"), | - new columns IDs for the new columns |
| sep = "\\ \\ ") | - identify the separator to be used |

| Filter | | | | | | | | | | | | |
|--------|-----------|------------|---|---------|-------|-------|-------|-------|-------|-------|--|--|
| | GID | YORF | NAME | GWEIGHT | G0.05 | G0.1 | G0.15 | G0.2 | G0.25 | G0.3 | | |
| 1 | GENE1331X | A_06_P5820 | SFB2 ER to Golgi transport molecular function unknown YNL049C 108... | 1 | -0.24 | -0.13 | -0.21 | -0.15 | -0.05 | -0.05 | | |
| 2 | GENE4924X | A_06_P5866 | biological process unknown molecular function unknown YNL095C 1... | 1 | 0.28 | 0.13 | -0.40 | -0.48 | -0.11 | 0.17 | | |
| 3 | GENE4690X | A_06_P1834 | QRI7 proteolysis and peptidolysis metalloendopeptidase activity YDL104... | 1 | -0.02 | -0.27 | -0.27 | -0.02 | 0.24 | 0.25 | | |
| 4 | GENE1177X | A_06_P4928 | CFT2 mRNA polyadenylation* RNA binding YLR115W 1081958 | 1 | -0.33 | -0.41 | -0.24 | -0.03 | -0.03 | 0.00 | | |
| 5 | GENE511X | A_06_P5620 | SSO2 vesicle fusion* t-SNARE activity YMR183C 1081214 | 1 | 0.05 | 0.02 | 0.40 | 0.34 | -0.13 | -0.14 | | |
| 6 | GENE2133X | A_06_P5307 | PSP2 biological process unknown molecular function unknown YML01... | 1 | -0.69 | -0.03 | 0.23 | 0.20 | 0.00 | -0.27 | | |
| 7 | GENE1002X | A_06_P6258 | RIB2 riboflavin biosynthesis pseudouridylate synthase activity* YOL066C... | 1 | -0.55 | -0.30 | -0.12 | -0.03 | -0.16 | -0.11 | | |
| 8 | GENE5478X | A_06_P7082 | VMA13 vacuolar acidification hydrogen-transporting ATPase activity, rota... | 1 | -0.75 | -0.12 | -0.07 | 0.02 | -0.32 | -0.41 | | |
| 9 | GENE2065X | A_06_P2554 | EDC3 deadenylation-independent decapping molecular function unkno... | 1 | -0.24 | -0.22 | 0.14 | 0.06 | 0.00 | -0.13 | | |
| 10 | GENE2440X | A_06_P6431 | VPS5 protein retention in Golgi* protein transporter activity YOR069W ... | 1 | -0.16 | -0.38 | 0.05 | 0.14 | -0.04 | -0.01 | | |
| 11 | GENE4180X | A_06_P6220 | biological process unknown molecular function unknown YOL029C 1... | 1 | -0.22 | -0.18 | 0.27 | 0.18 | 0.03 | -0.04 | | |
| 12 | GENE5247X | A_06_P1410 | AMN1 negative regulation of exit from mitosis* protein binding YBR158... | 1 | 0.18 | 0.61 | 1.55 | 1.34 | 0.23 | -0.03 | | |
| 13 | GENE2121X | A_06_P2983 | SCW11 cytokinesis, completion of separation glucan 1,3-beta-glucosidas... | 1 | -0.67 | -0.47 | 1.16 | 1.05 | -0.18 | -0.68 | | |
| 14 | GENE1985X | A_06_P3720 | DSE2 cell wall organization and biogenesis* glucan 1,3-beta-glucosidase ... | 1 | -0.59 | -0.17 | 1.17 | 0.85 | -0.12 | -0.61 | | |
| 15 | GENE4728X | A_06_P2774 | COX15 cytochrome c oxidase complex assembly* oxidoreductase activity,... | 1 | -0.28 | -0.81 | -0.39 | 0.24 | 0.01 | 0.01 | | |
| 16 | GENE3153X | A_06_P4597 | SPE1 pantothenate biosynthesis* ornithine decarboxylase activity YKL18... | 1 | -0.19 | 0.24 | 0.03 | 0.17 | 0.00 | -0.01 | | |
| 17 | GENE3704X | A_06_P5667 | MTF1 transcription from mitochondrial promoter S-adenosylmethionine-... | 1 | -0.42 | -0.43 | -0.36 | -0.12 | 0.05 | 0.24 | | |
| 18 | GENE2141X | A_06_P3260 | KSS1 invasive growth (sensu Saccharomyces)* MAP kinase activity YGR... | 1 | -0.76 | -0.32 | -0.05 | -0.27 | -0.31 | -0.01 | | |
| 19 | GENE2978X | A_06_P3607 | biological process unknown molecular function unknown YHR036W 1... | 1 | -0.91 | -0.43 | -0.05 | -0.09 | -0.27 | -0.45 | | |
| 20 | GENE1203X | A_06_P5929 | biological process unknown molecular function unknown YNL158W 1... | 1 | -0.47 | -0.43 | -0.15 | 0.08 | -0.26 | -0.25 | | |

Try to limit “uninformative” data

“GWEIGHT” contains the same information in every cell

- This isn’t going to add to our analysis

“GID” and “YORF” appear to be study specific IDs

“NAME” column contains a lot of information

Going back to the previous example;

```
SFB2 || ER to Golgi transport || molecular function unknown || YNL049C || 1082129
```

SFB2: Gene names, but not present in all cases

ER to Golgi transport: Biological process

molecular function unknown: Molecular function

YNL049C: Gene ID listed on public repositories

1082129: Another identifier that does not appear to be useful

Line by line

The screenshot displays the RStudio environment with three main panes:

- Source Editor:** Contains R code for data processing, line by line.
- Environment:** Shows the Global Environment with a table of objects.
- File Explorer:** Shows the project directory structure.

Source Editor Code:

```
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 != "")
```

Environment Pane:

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

File Explorer:

| 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 |

Console:

```
> 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.
> |
```

Line by line

[illegible]

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

The screenshot displays the RStudio environment with three main panes:

- Source Editor:** Contains R code for data processing, line by line.
- Environment Pane:** Shows the Global Environment with a table of data frames.
- Files Pane:** Shows the project file structure.

Source Editor Code:

```
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 != "")
```

Environment Pane:

| 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 Pane:

| 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 |

Console:

```
~/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)
```


Line by line

The screenshot displays the RStudio environment with a script editor on the left and the Environment pane on the right.

Script Editor: The script defines data processing steps for the Brauer2008 dataset. It reads a TSV file, separates columns, mutates the systematic names, selects specific columns, gathers data, and filters out missing values.

```
12 raw_gene_df <- read_delim("Brauer2008_DataSet1.tds", delim = "\t")
13
14 separated_gene_df <- separate(raw_gene_df, NAME,
15                               c("name", "BP", "MF", "systematic_name",
16                                 "number"),
17                               sep = "\\|\\|\\|\\|\\|")
18
19 mutated_gene_df <- mutate_at(separated_gene_df,
20                               vars(name:systematic_name),
21                               funs(trimws))
22
23
24 selected_gene_df <- select(mutated_gene_df, -number, -GID, -YORF, -GWEIGHT)
25
26 gathered_gene_df <- gather(selected_gene_df, sample, expression, G0.05:U0.3)
27
28 nearly_there_df <- separate(gathered_gene_df, sample,
29                              c("nutrient", "rate"), sep = 1, convert = TRUE)
30
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 != "")
```

Environment Pane: Shows the Global Environment with the following objects:

| 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 Pane: Shows the project structure for '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 |

Console: Shows the execution of the script, including column specifications and the start of the data processing steps.

```
> 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)
```

Line by line

The screenshot displays the RStudio environment with three main panes:

- Source Editor:** Contains R code for data processing, line by line.
- Environment:** Shows the Global Environment with a list of data frames and their characteristics.
- Files:** Shows the project file structure.

Source Editor Code:

```
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 != "")
```

Environment Pane:

| 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 Pane:

| 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 |

Console:

```
~/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)
>
```

Line by line

```
ws1_script1_stepwise_Bauer_dataset_an... *  
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

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 |
| .Rhistry | 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

The image shows the RStudio IDE interface. The main editor pane displays R code for data processing. The code defines several data frames: `separated_gene_df`, `mutated_gene_df`, `selected_gene_df`, `gathered_gene_df`, `nearly_there_df`, `nutrient_names`, `cleaned_genes_df`, and `filtered_genes_df`. The code uses functions like `separate`, `mutate_at`, `select`, `gather`, `mutate`, and `filter` from the `dplyr` package, and `revalue` from the `plyr` package.

```
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
```

The Global Environment pane on the right shows the following data frames:

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

The Files pane on the right shows the project structure:

- Home > R_Users_Workshop > 8_weeks_Oct-Dec-17 > Workshop_1 > workshop_1_project
- Files: `..`, `.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).

Nested

```
ws1_script1_stepwise_Bauer_dataset_an... * x
Source on Save
Run
Source

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

24:1 (Top Level) R Script

```
~/R_Users_Workshop/8_weeks_Oct-Dec_17/Workshop_1/workshop_1_project/
+       sep = "\\|\\|\\|\\|", vars(name:systematic_name),
+       funs(trimws)),
+       -number, -GID, -YORF, -GWEIGHT),
+       sample, expression, G0.05:U0.3),
+       sample,
+       c("nutrient", "rate"),
+       sep = 1, convert = TRUE),
+       nutrient = plyr::revalue(nutrient, nutrient_names)),
+       !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

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

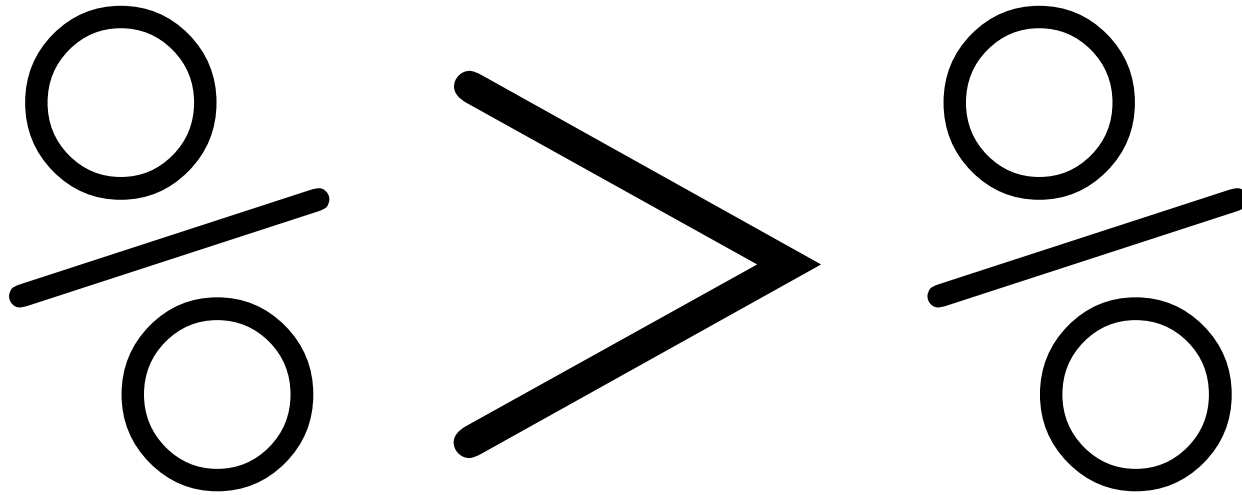
Home > R_Users_Workshop > 8_weeks_Oct-Dec_17 > Workshop_1 > workshop_1_project

| 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 |

Nested

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

Putting the pieces together



Code structure v2

```
separated_gene <- raw_gene %>%
```

Here the input data is outside the function

```
  separate(NAME,
```

First argument is no longer the data

```
    c("name", "BP", "MF", "systematic_name", "number"),
```

```
    sep = "\\|\\|\\|"
```

```
)
```


Piped

The screenshot displays the RStudio environment with three main panes:

- Source Pane:** Contains an R script for data cleaning and transformation using the `dplyr` package. The script defines `nutrient_names`, reads a TSV file, and performs a series of `select`, `mutate`, and `filter` operations.
- Console Pane:** Shows the execution output of the script, including the `col_types` specification for the imported data.
- Environment Pane:** Lists the objects in the global environment, including `cleaned_genes_df` and `nutrient_names`.

```
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   select(-number, -GID, -YORF, -GWEIGHT) %>%
14
15   gather(sample, expression, G0.05:U0.3) %>%
16
17   separate(sample, c("nutrient", "rate"), sep = 1, convert = TRUE) %>%
18
19   mutate(nutrient = plyr::revalue(nutrient, nutrient_names)) %>%
20
21   filter(!is.na(expression), systematic_name != "")
22
23
24
25
26
27
```

Console Output:

```
+ 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 Pane:

| Name | Type | Length | Size | Value |
|---|-----------|--------|---------|----------------------------------|
| <input type="checkbox"/> cleaned_genes_df | tbl_df | 7 | 11.3 MB | 198430 obs. of 7 variables |
| <input type="checkbox"/> nutrient_names | character | 6 | 984 B | Named chr [1:6] "Glucose" "Le... |

Files Pane:

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

Piped

```
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   )
```

Magrittr package

- Tidyverse packages automatically make `%>%` available
- Magrittr has other tools that you might find useful

`%$%`

`%<>%`

`%T>%`

The moral of the story.....

You can go from this

To this!!



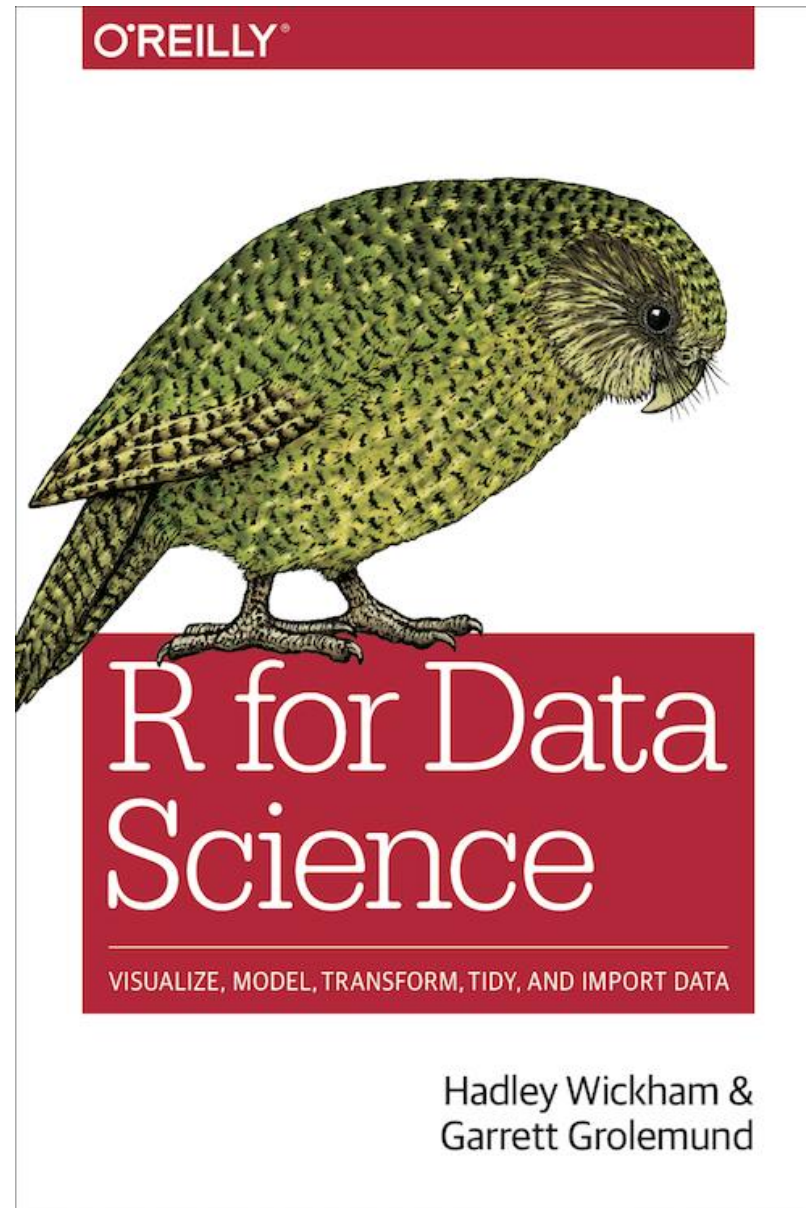
+



=



You could write a book on that!!



<https://r4ds.had.co.nz/>

And on this!!

