

R: A Hitchhikers Guide to Reproducible Research

- Take a parachute and jump (into the tidyverse)

Brendan Palmer,

Clinical Research Facility - Cork &

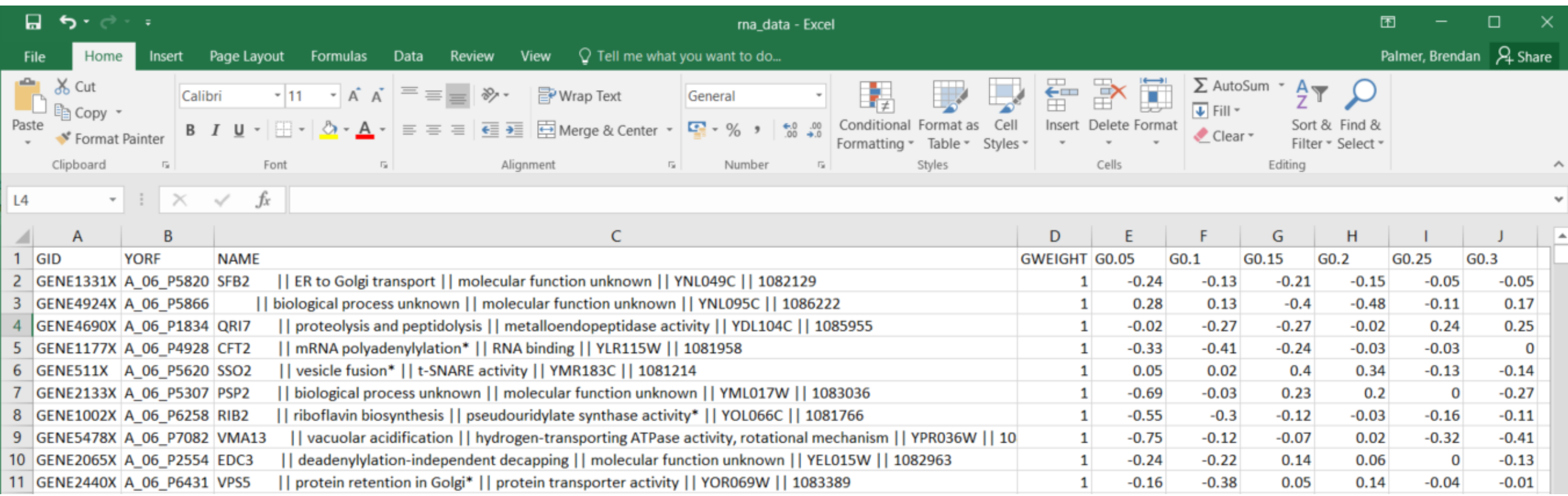
School of Public Health

 @B_A_Palmer

Tidyverse works best with tidy data

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

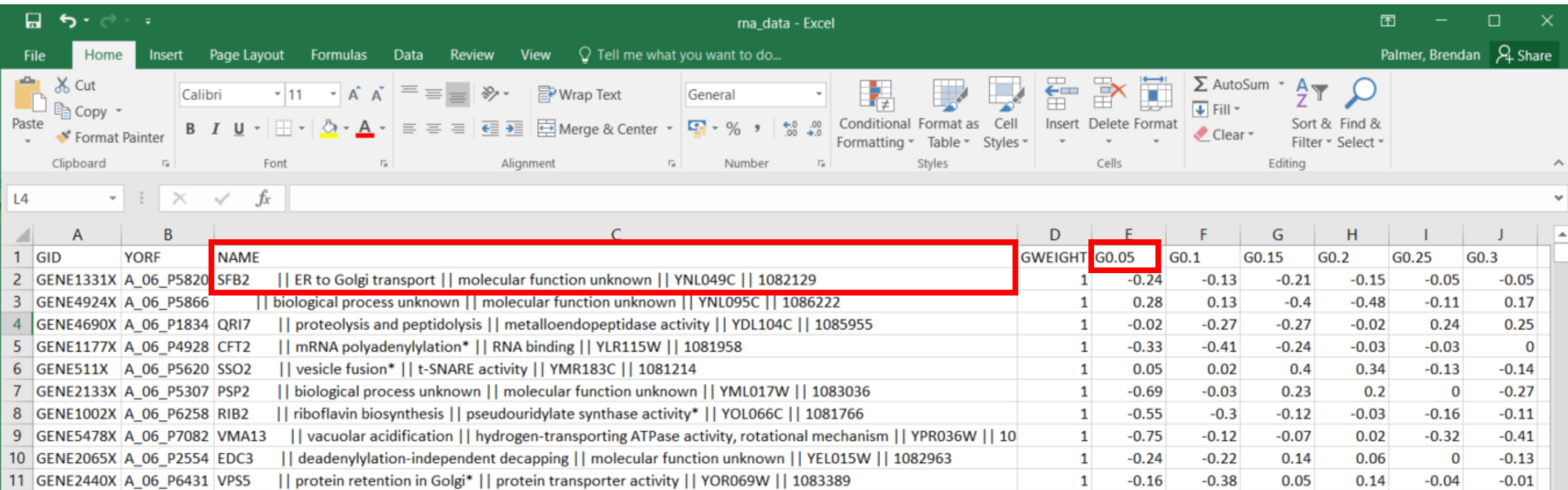
Problems with the example RNA data set...



The screenshot shows an Excel spreadsheet titled 'rna_data - Excel'. The ribbon includes tabs for File, Home, Insert, Page Layout, Formulas, Data, Review, and View. The Home tab is active, showing options for Clipboard, Font, Alignment, Number, Styles, Cells, and Editing. The spreadsheet has columns labeled A through J. Column A contains GID, B contains YORF, C contains NAME, D contains GWEIGHT, E contains G0.05, F contains G0.1, G contains G0.15, H contains G0.2, I contains G0.25, and J contains G0.3. The data is organized into rows, with each row representing a different gene and its associated measurements.

	A	B	C	D	E	F	G	H	I	J
1	GID	YORF	NAME	GWEIGHT	G0.05	G0.1	G0.15	G0.2	G0.25	G0.3
2	GENE1331X	A_06_P5820	SFB2 ER to Golgi transport molecular function unknown YNL049C 1082129	1	-0.24	-0.13	-0.21	-0.15	-0.05	-0.05
3	GENE4924X	A_06_P5866	biological process unknown molecular function unknown YNL095C 1086222	1	0.28	0.13	-0.4	-0.48	-0.11	0.17
4	GENE4690X	A_06_P1834	QRI7 proteolysis and peptidolysis metalloendopeptidase activity YDL104C 1085955	1	-0.02	-0.27	-0.27	-0.02	0.24	0.25
5	GENE1177X	A_06_P4928	CFT2 mRNA polyadenylation* RNA binding YLR115W 1081958	1	-0.33	-0.41	-0.24	-0.03	-0.03	0
6	GENE511X	A_06_P5620	SSO2 vesicle fusion* t-SNARE activity YMR183C 1081214	1	0.05	0.02	0.4	0.34	-0.13	-0.14
7	GENE2133X	A_06_P5307	PSP2 biological process unknown molecular function unknown YML017W 1083036	1	-0.69	-0.03	0.23	0.2	0	-0.27
8	GENE1002X	A_06_P6258	RIB2 riboflavin biosynthesis pseudouridylyl synthase activity* YOL066C 1081766	1	-0.55	-0.3	-0.12	-0.03	-0.16	-0.11
9	GENE5478X	A_06_P7082	VMA13 vacuolar acidification hydrogen-transporting ATPase activity, rotational mechanism YPR036W 10	1	-0.75	-0.12	-0.07	0.02	-0.32	-0.41
10	GENE2065X	A_06_P2554	EDC3 deadenylation-independent decapping molecular function unknown YEL015W 1082963	1	-0.24	-0.22	0.14	0.06	0	-0.13
11	GENE2440X	A_06_P6431	VP55 protein retention in Golgi* protein transporter activity YOR069W 1083389	1	-0.16	-0.38	0.05	0.14	-0.04	-0.01

Tidyverse works best with tidy data



The screenshot shows an Excel spreadsheet with the following data:

	A	B	C	D	E	F	G	H	I	J
1	GID	YORF	NAME	GWEIGHT	G0.05	G0.1	G0.15	G0.2	G0.25	G0.3
2	GENE1331X	A_06_P5820	SFB2 ER to Golgi transport molecular function unknown YNL049C 1082129	1	-0.24	-0.13	-0.21	-0.15	-0.05	-0.05
3	GENE4924X	A_06_P5866	biological process unknown molecular function unknown YNL095C 1086222	1	0.28	0.13	-0.4	-0.48	-0.11	0.17
4	GENE4690X	A_06_P1834	QRI7 proteolysis and peptidolysis metalloendopeptidase activity YDL104C 1085955	1	-0.02	-0.27	-0.27	-0.02	0.24	0.25
5	GENE1177X	A_06_P4928	CFT2 mRNA polyadenylation* RNA binding YLR115W 1081958	1	-0.33	-0.41	-0.24	-0.03	-0.03	0
6	GENE511X	A_06_P5620	SSO2 vesicle fusion* t-SNARE activity YMR183C 1081214	1	0.05	0.02	0.4	0.34	-0.13	-0.14
7	GENE2133X	A_06_P5307	PSP2 biological process unknown molecular function unknown YML017W 1083036	1	-0.69	-0.03	0.23	0.2	0	-0.27
8	GENE1002X	A_06_P6258	RIB2 riboflavin biosynthesis pseudouridylyl synthase activity* YOL066C 1081766	1	-0.55	-0.3	-0.12	-0.03	-0.16	-0.11
9	GENE5478X	A_06_P7082	VMA13 vacuolar acidification hydrogen-transporting ATPase activity, rotational mechanism YPR036W 1081214	1	-0.75	-0.12	-0.07	0.02	-0.32	-0.41
10	GENE2065X	A_06_P2554	EDC3 deadenylation-independent decapping molecular function unknown YEL015W 1082963	1	-0.24	-0.22	0.14	0.06	0	-0.13
11	GENE2440X	A_06_P6431	VP55 protein retention in Golgi* protein transporter activity YOR069W 1083389	1	-0.16	-0.38	0.05	0.14	-0.04	-0.01

- Multiple variables are stored in one column
 - e.g. column "NAME" contains values such as;

G0.05 - letter identifies a compound

- number is the concentration of that compound

Tidyverse code structure has two main forms

(1)

new_object

<-

function(

input_data,

data_to_b_modified,

arguments_to_function

)

(2)

new_object

<-

input_data

%>%

magrittr / pipe
operator

function(

data_to_b_modified,

arguments_to_function

)

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 the variable `raw_gene_df` of type `tbl_df`, length 40, size 3.3 MB, and 5537 observations of 40 variables.
- Files Pane:** Shows the file explorer for the project directory `workshop_1_project`, listing various files including `Brauer2008_DataSet1.tds` and `workshop_1.Rproj`.

```
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

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:

```
> 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

```

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 selected_gene_df <- select(mutated_gene_df, -number, -GID, -YORF, -GWEIGHT)
24
25 gathered_gene_df <- gather(selected_gene_df, sample, expression, G0.05:U0.3)
26
27 nearly_there_df <- separate(gathered_gene_df, sample,
28                             c("nutrient", "rate"), sep = 1, convert = TRUE)
29
30 nutrient_names <- c(G = "Glucose", L = "Leucine", P = "Phosphate",
31                    S = "Sulfate", N = "Ammonia", U = "Uracil")
32
33 cleaned_genes_df <- mutate(nearly_there_df,
34                            nutrient = plyr::revalue(nutrient, nutrient_names)
35                            ) %>%
36 filter(!is.na(expression), systematic_name != "")
37
38 
```

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 an R script editor on the left, a console at the bottom, and a file explorer on the right.

R Script Editor:

```
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 != "")
```

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)
```

Environment Panel:

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 Panel:

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 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 starts by reading a tab-separated file, separating it into columns, mutating the systematic names to remove whitespace, selecting specific columns, gathering sample and expression data, separating by nutrient and rate, and finally filtering 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: This pane shows the objects created in the global environment. It lists several data frames and their characteristics.

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: This pane shows the file structure of the project. It lists various files including data sets, completions, and R scripts.

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: The console shows the execution of the script, including the output of the `col_types` function and the first few lines of the `separated_gene_df` data frame.

```
> col_types(
+   .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)
>
```


Line by line

The image shows an RStudio session with R code for data processing. The code is as follows:

```
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 != "")
```

The RStudio interface includes a console at the bottom with the following output:

```
~/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)
>
>
```

The right pane shows the Environment tab with the following data frames:

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

The Files tab shows the project structure:

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

```
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
<input type="checkbox"/> gathered_gene_df	tbl_df	6	9.8 MB	199332 obs. of 6 variables
<input type="checkbox"/> mutated_gene_df	tbl_df	44	3.5 MB	5537 obs. of 44 variables
<input type="checkbox"/> nearly_there_df	tbl_df	7	11.3 MB	199332 obs. of 7 variables
<input type="checkbox"/> nutrient_names	character	6	984 B	Named chr [1:6] "Glucose" "
<input type="checkbox"/> raw_gene_df	tbl_df	40	3.3 MB	5537 obs. of 40 variables
<input type="checkbox"/> selected_gene_df	tbl_df	40	2.4 MB	5537 obs. of 40 variables
<input type="checkbox"/> separated_gene...	tbl_df	44	3.6 MB	5537 obs. of 44 variables

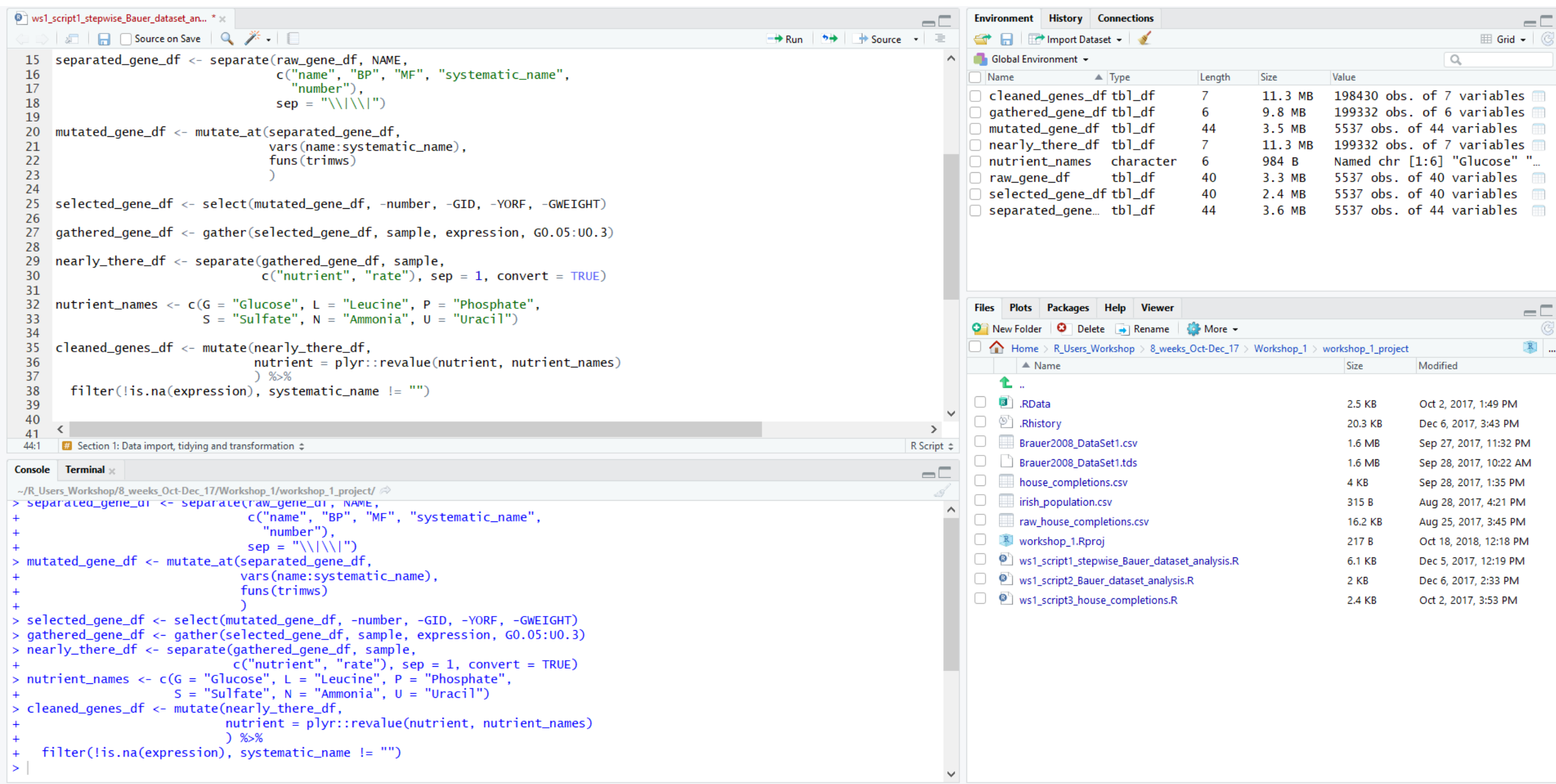
Files **Plots** **Packages** **Help** **Viewer**

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 screenshot displays the RStudio IDE interface with the following components:

- Script Editor:** Contains R code for data processing, with line numbers 15 to 41 visible. The code performs the following steps:
 - 15: `separated_gene_df <- separate(raw_gene_df, NAME,`
 - 16: `c("name", "BP", "MF", "systematic_name",`
 - 17: `"number"),`
 - 18: `sep = "\\|\\|\\|\\|")`
 - 19: (blank line)
 - 20: `mutated_gene_df <- mutate_at(separated_gene_df,`
 - 21: `vars(name:systematic_name),`
 - 22: `funs(trimws)`
 - 23: `)`
 - 24: (blank line)
 - 25: `selected_gene_df <- select(mutated_gene_df, -number, -GID, -YORF, -GWEIGHT)`
 - 26: (blank line)
 - 27: `gathered_gene_df <- gather(selected_gene_df, sample, expression, G0.05:U0.3)`
 - 28: (blank line)
 - 29: `nearly_there_df <- separate(gathered_gene_df, sample,`
 - 30: `c("nutrient", "rate"), sep = 1, convert = TRUE)`
 - 31: (blank line)
 - 32: `nutrient_names <- c(G = "Glucose", L = "Leucine", P = "Phosphate",`
 - 33: `S = "Sulfate", N = "Ammonia", U = "Uracil")`
 - 34: (blank line)
 - 35: `cleaned_genes_df <- mutate(nearly_there_df,`
 - 36: `nutrient = plyr::revalue(nutrient, nutrient_names)`
 - 37: `) %>%`
 - 38: `filter(!is.na(expression), systematic_name != "")`
 - 39: (blank line)
 - 40: (blank line)
 - 41: (blank line)
- Environment:** A table showing the state of the R environment after each line of code is executed.

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
- Files:** A file explorer showing the project directory structure.
 - Home > R_Users_Workshop > 8_weeks_Oct-Dec_17 > Workshop_1 > workshop_1_project
 - Files listed: `..`, `.RData` (2.5 KB), `.Rhistory` (20.3 KB), `Brauer2008_DataSet1.csv` (1.6 MB), `Brauer2008_DataSet1.tds` (1.6 MB), `house_completions.csv` (4 KB), `irish_population.csv` (315 B), `raw_house_completions.csv` (16.2 KB), `workshop_1.Rproj` (217 B), `ws1_script1_stepwise_Bauer_dataset_analysis.R` (6.1 KB), `ws1_script2_Bauer_dataset_analysis.R` (2 KB), `ws1_script3_house_completions.R` (2.4 KB).
- Console:** Shows the execution of the script line by line, with the same code as the script editor.

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

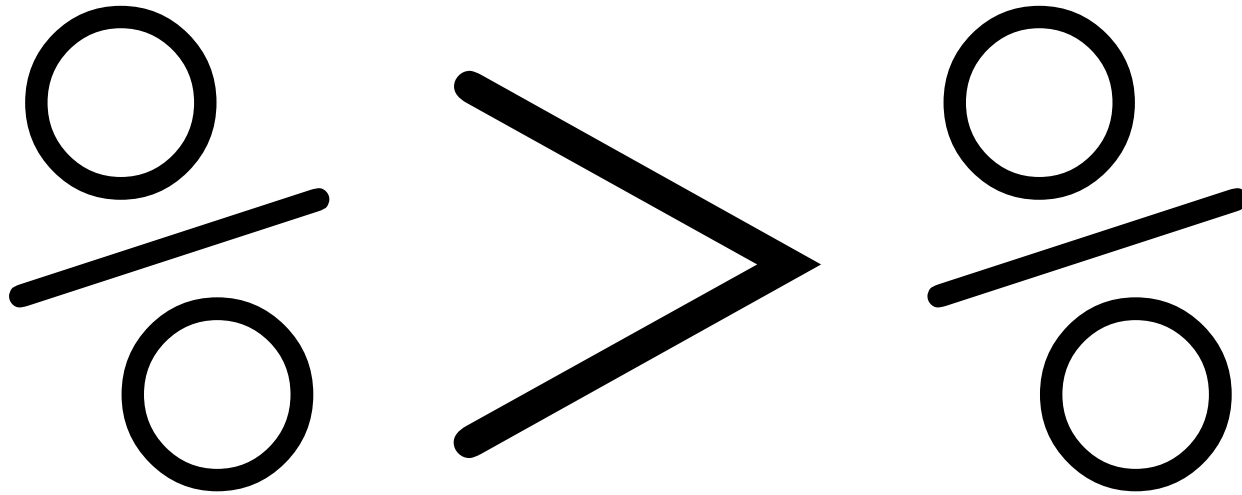
Home > R_Users_Workshop > 8_weeks_Oct-Dec_17 > Workshop_1 > workshop_1_project

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 has two main forms

(1)

new_object

<-

function(

input_data,

data_to_b_modified,

arguments_to_function

)

(2)

new_object

<-

input_data

function(

data_to_b_modified,

arguments_to_function

)

Piped

The screenshot displays the RStudio environment with three main panes:

- Source Pane:** Contains an R script using the `dplyr` package to process gene expression data. The script defines nutrient names, reads a TSV file, separates columns, trims whitespace, selects specific columns, filters out missing values, and separates sample information.
- Console Pane:** Shows the execution of the script. It displays the `separate` function call, the `mutate` operation, the `filter` operation, and the resulting column specifications for the `cols` function.
- Environment Pane:** Lists the objects in the global environment. It shows `cleaned_genes_df` as a `tbl_df` with 11.3 MB and 198,430 observations, and `nutrient_names` as a character vector of length 6.

```
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   )
```

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   )
```

Moral of the story...

You can go from this



+



=

To this!!



**Master
Builder!!**