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

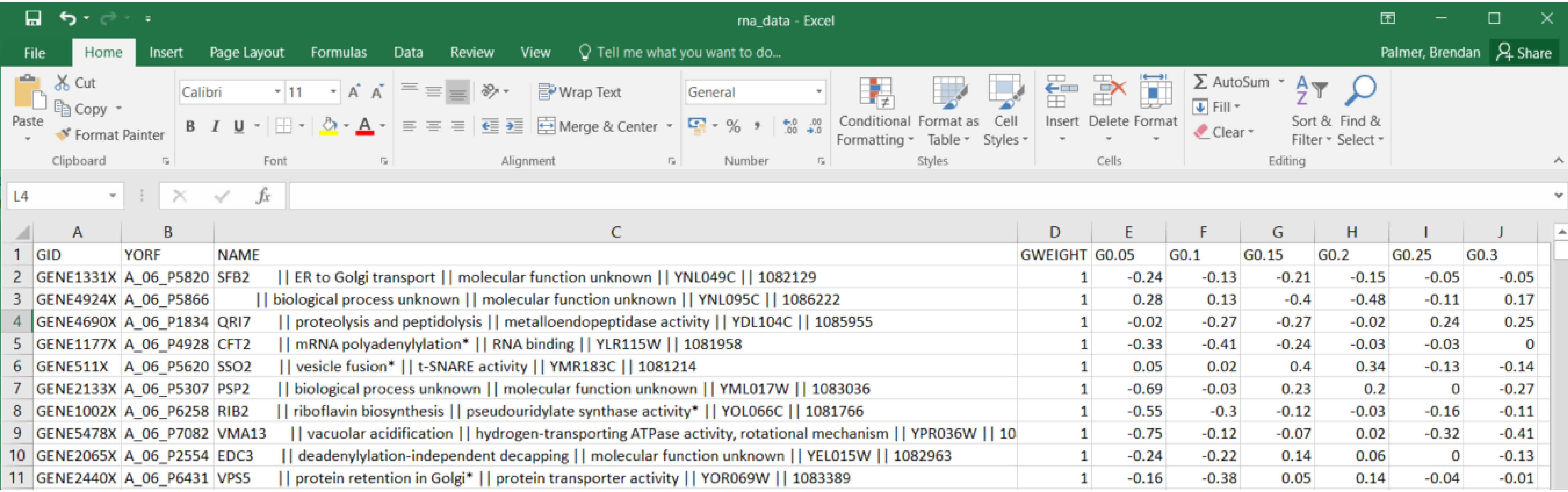
CRF-C HRB Clinical Research Facility Cork

UCC University College Cork, Ireland Coláiste na hOllscoile Corcaigh | School of Public Health

Tidyverse works best with tidy data

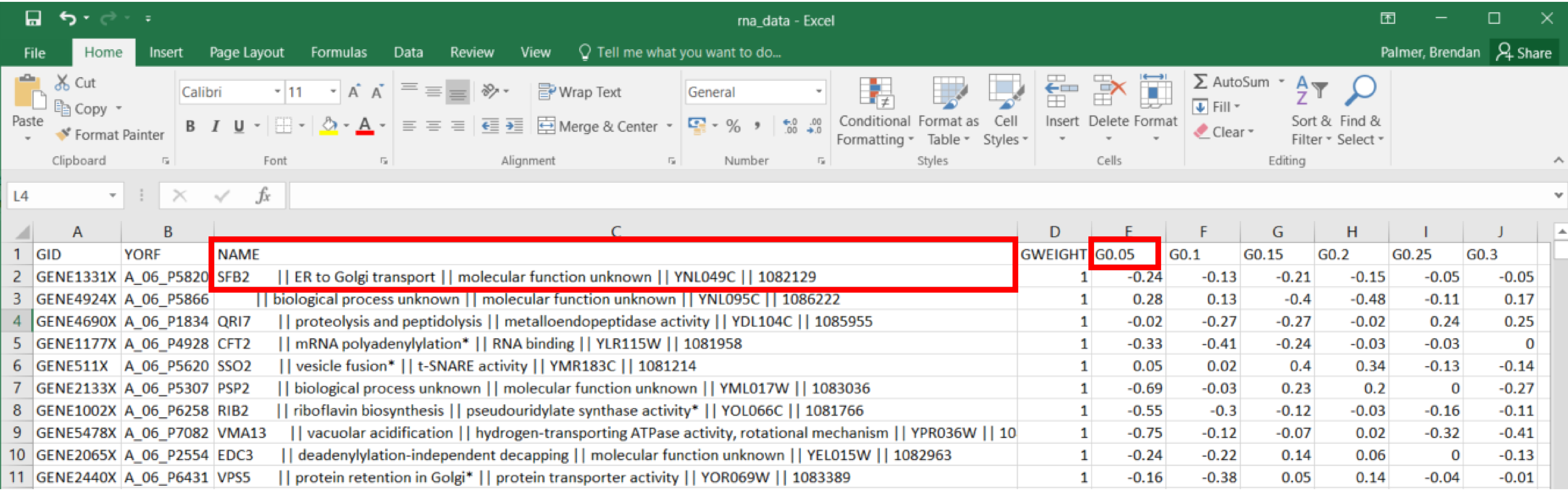
- Each variable forms a column
- Each observation forms a row

Problems with the example RNA data set…



	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 polyadenylylation* || 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 || pseudouridylate 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 || deadenylylation-independent decapping || molecular function unknown || YEL015W || 1082963	1	-0.24	-0.22	0.14	0.06	0	-0.13
11	GENE2440X	A_06_P6431	VPS5 || 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

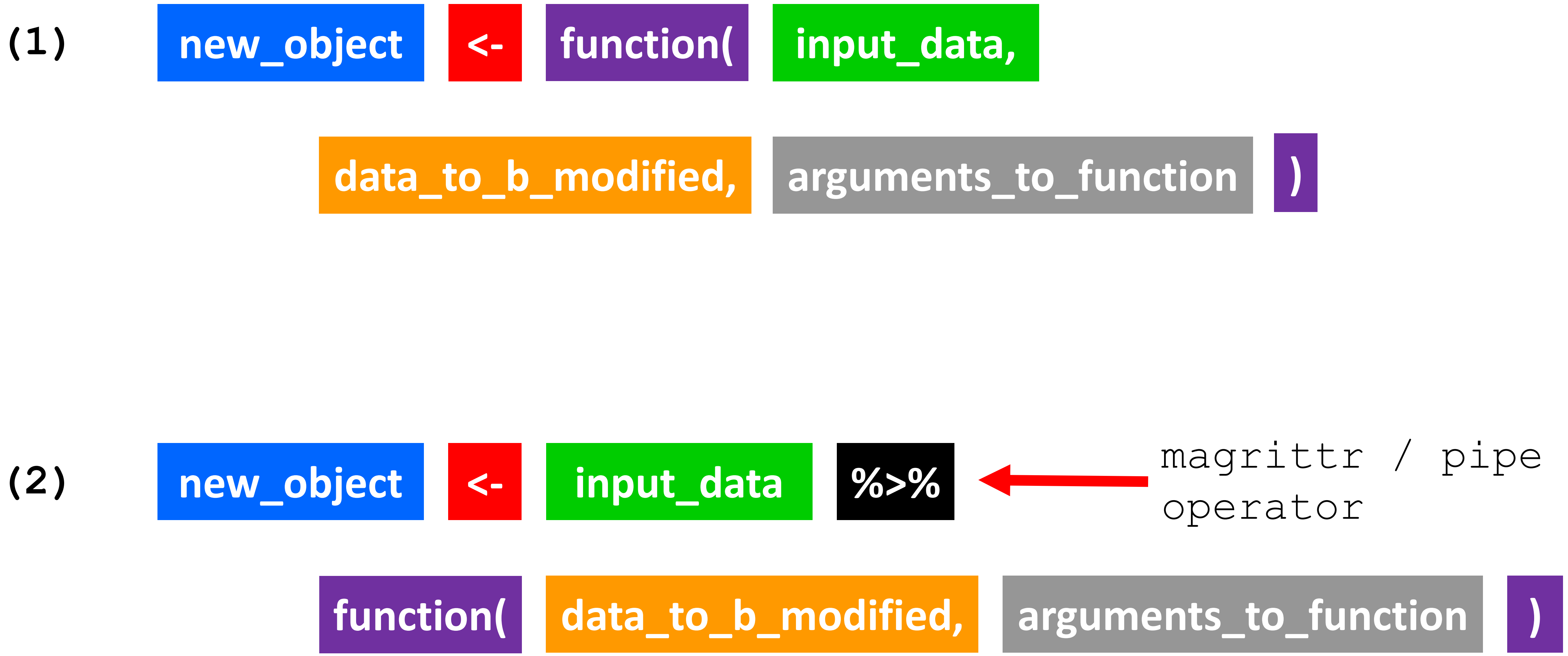


	A	B	C	D	E	F	G	H	I	J
1	GID	YORF	NAME	GWEIGHT	G0.05	G0.1	G0.15	G0.2	G0.25	G0.3
2	GENE1331X	A_06_P5820	SFB2 || ER to Golgi transport || molecular function unknown || YNL049C || 1082129	1	-0.24	-0.13	-0.21	-0.15	-0.05	-0.05
3	GENE4924X	A_06_P5866	|| biological process unknown || molecular function unknown || 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 polyadenylylation* || 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 || pseudouridylate 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	VPS5 || 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,

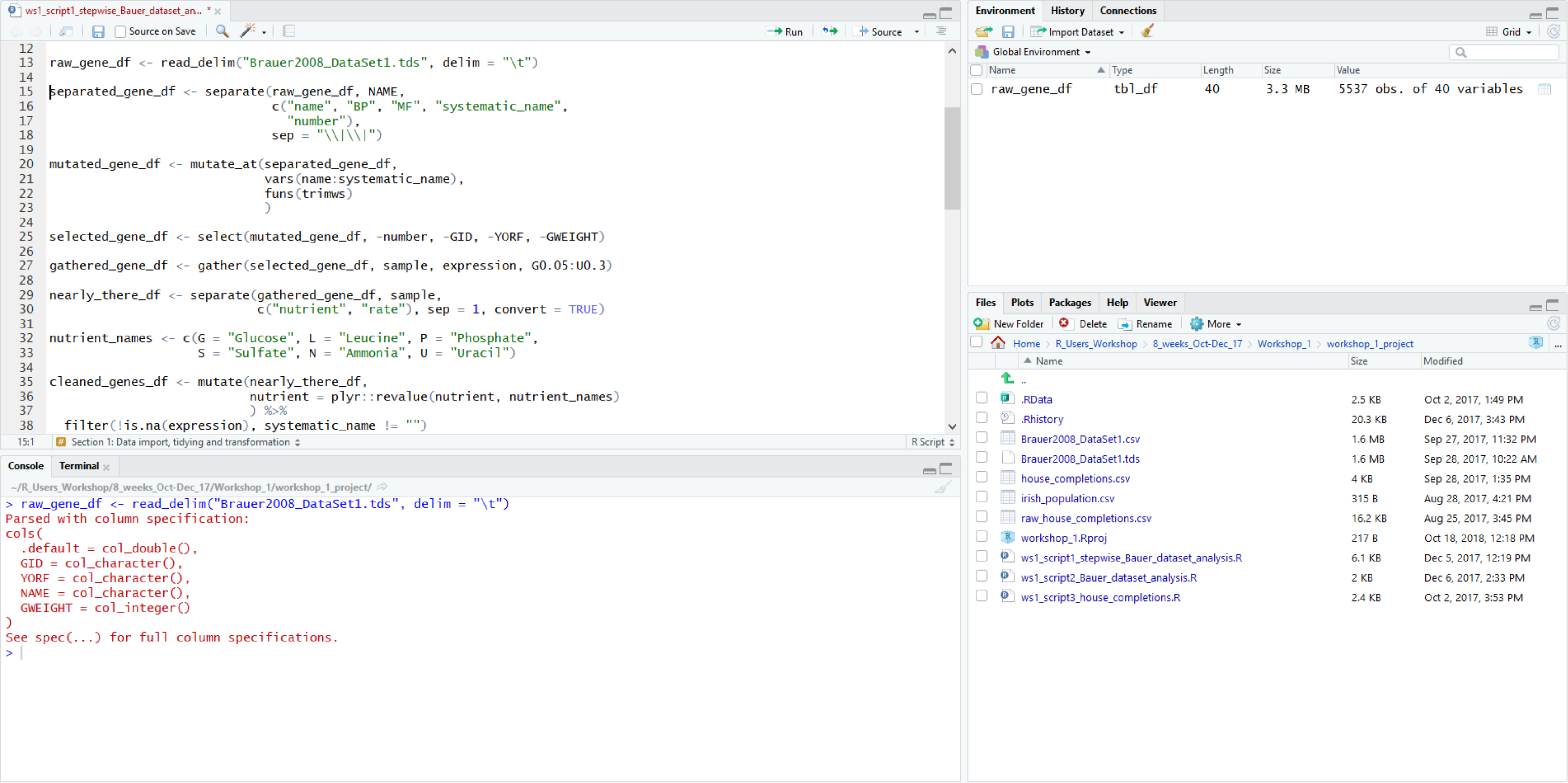
data_to_b_modified, arguments_to_function)

(2) new_object <- input_data %>%

function(data_to_b_modified, arguments_to_function)

magrittr / pipe operator

Line by line


ws1_script1_stepwise_Bauer_dataset_an... *
Source on Save
Run
Source
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/
Environment
History
Connections
Import Dataset
Grid
Global Environment
Files
Plots
Packages
Help
Viewer
New Folder
Delete
Rename
More
Home › R_Users_Workshop › 8_weeks_Oct-Dec_17 › Workshop_1 › workshop_1_project

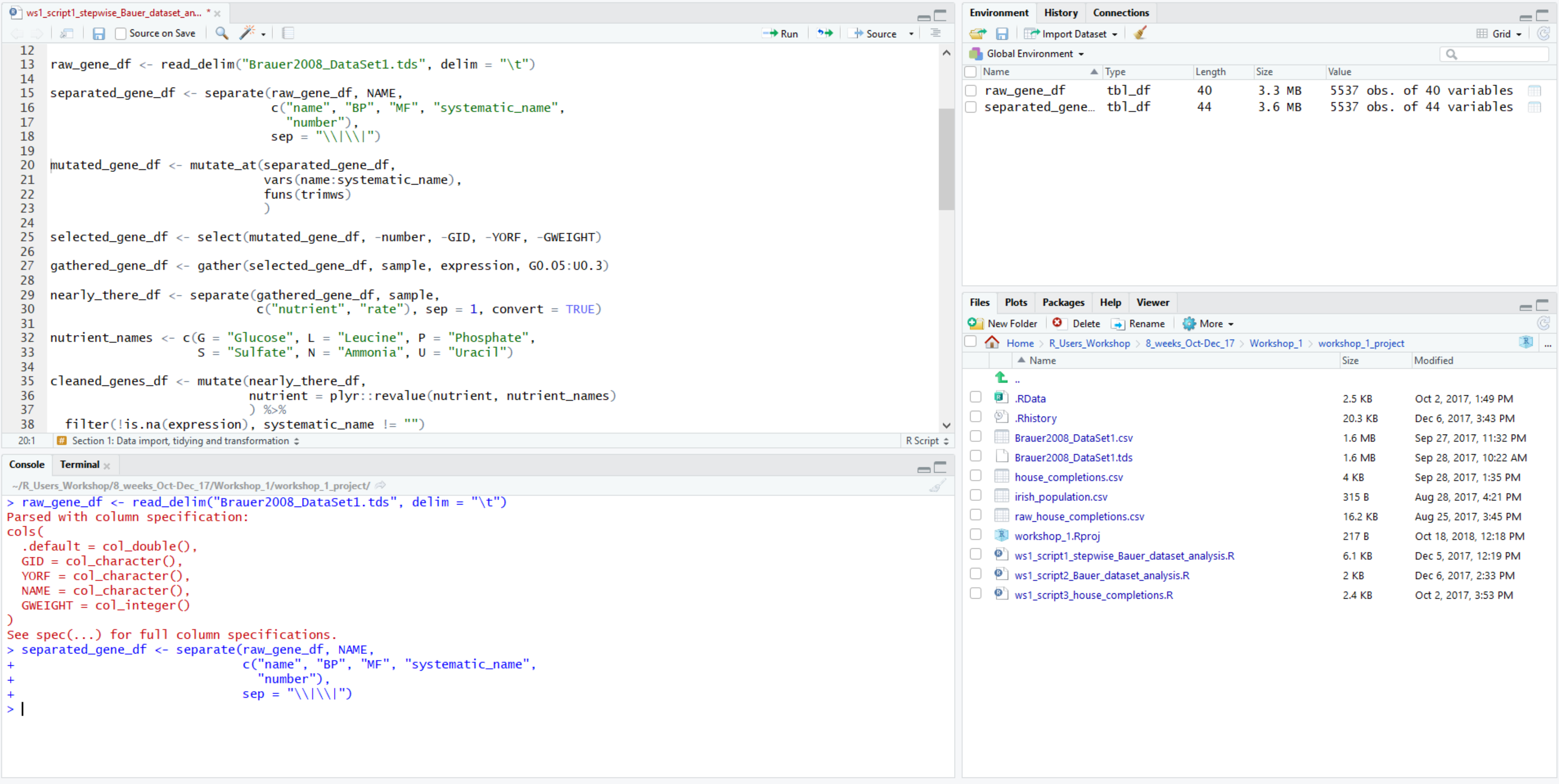
```r
raw_gene_df <- read_delim("Brauer2008_DataSet1.tds", delim = "\t")

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 != "")
```

```
> 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.
> 
```

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

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... *
Source on Save
Run
Source
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/
Environment
History
Connections
Import Dataset
Grid
Global Environment
Files
Plots
Packages
Help
Viewer
New Folder
Delete
Rename
More
Home › R_Users_Workshop › 8_weeks_Oct-Dec_17 › Workshop_1 › workshop_1_project

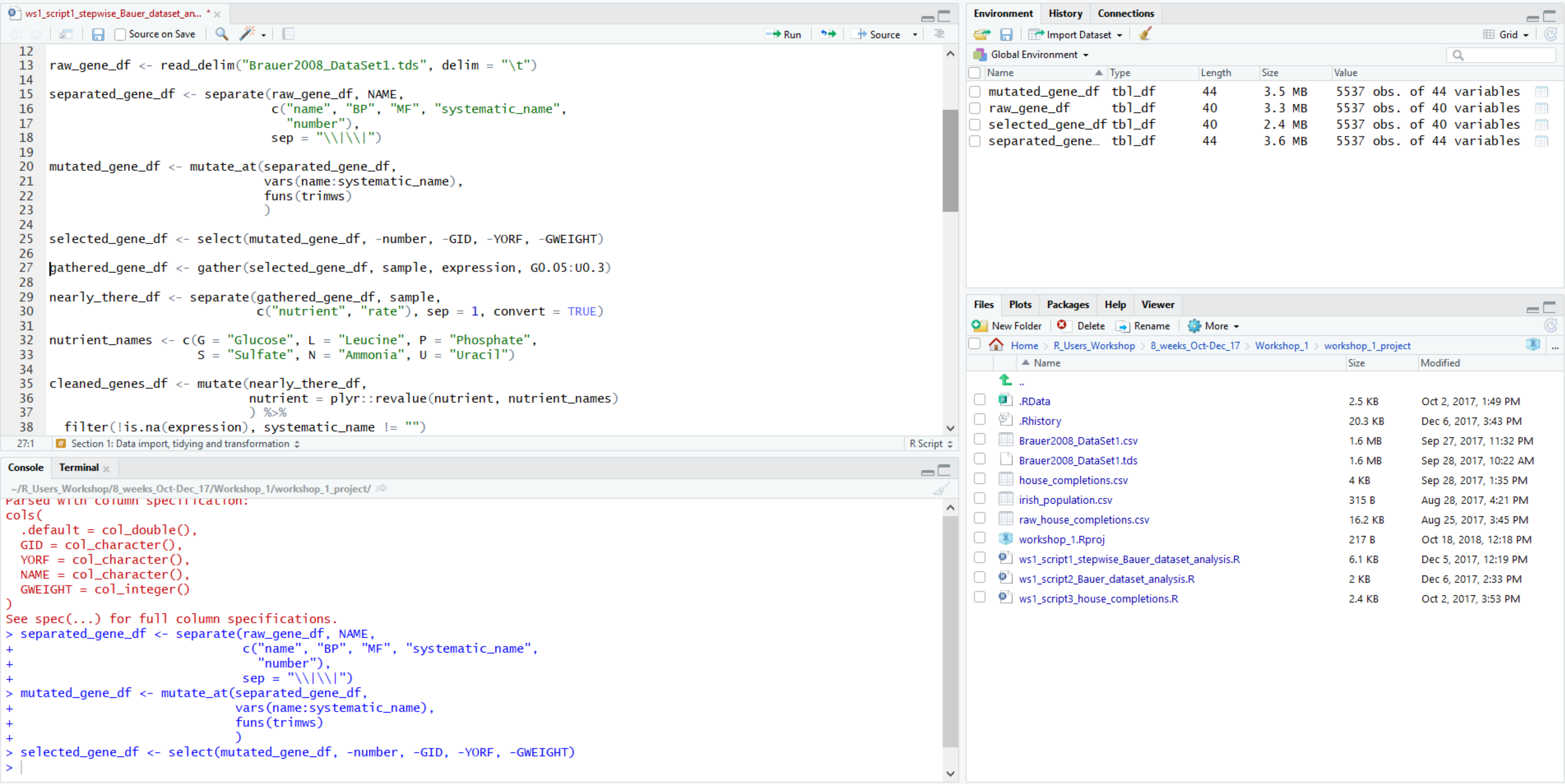
```r
raw_gene_df <- read_delim("Brauer2008_DataSet1.tds", delim = "\t")

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 != "")
```

```
> 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 = "\\|\\|")
> 
```

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

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
raw_gene_df <- read_delim("Brauer2008_DataSet1.tds", delim = "\t")

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 != "")
```

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

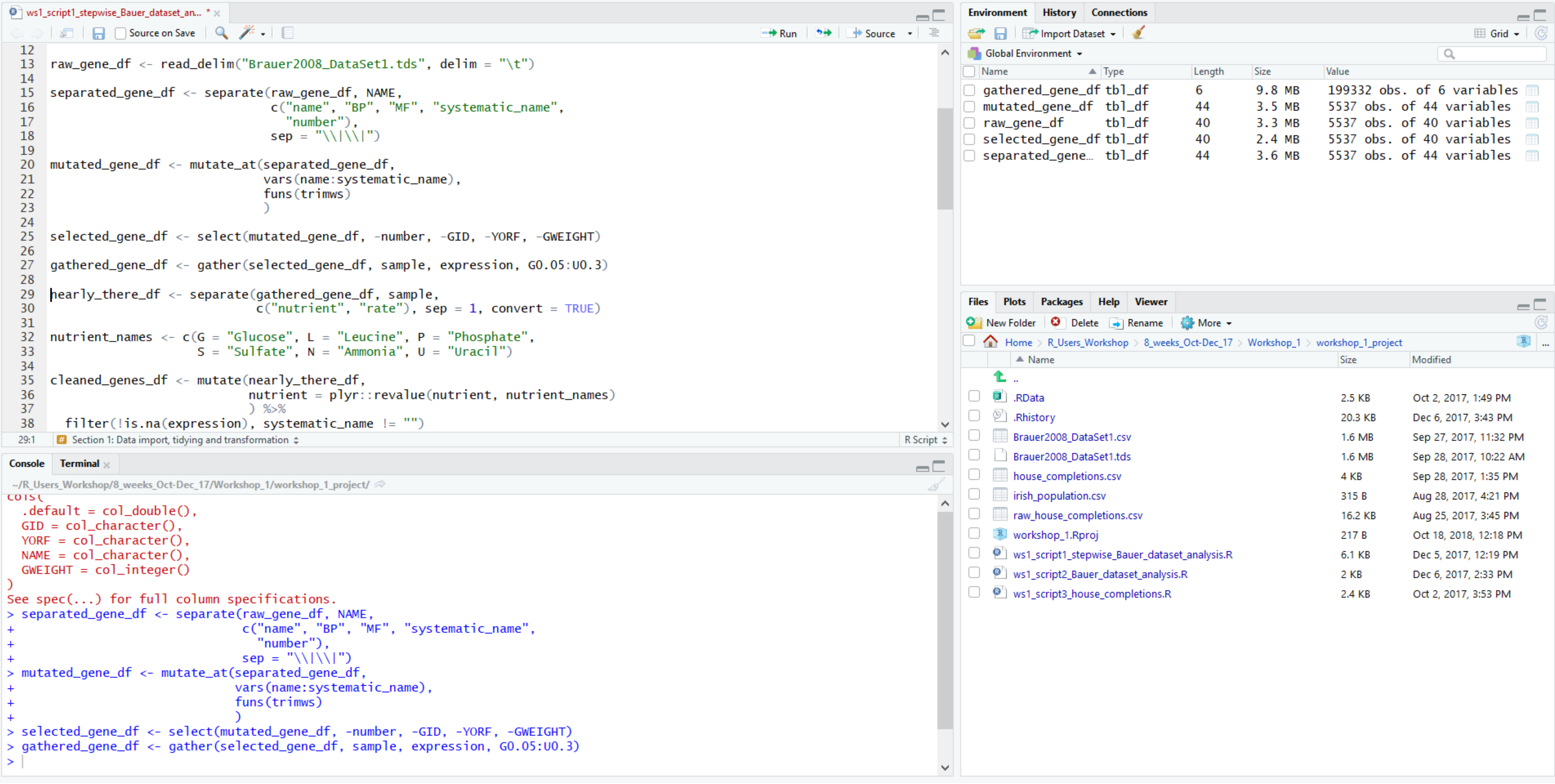
```
Parsed with column specification:
cols(
  .default = col_double(),
  GID = col_character(),
  YORF = col_character(),
  NAME = col_character(),
  GWEIGHT = col_integer()
)
See spec(...) for full column specifications.
> separated_gene_df <- separate(raw_gene_df, NAME,
+                               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)
> 
```

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

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



ws1_script1_stepwise_Bauer_dataset_an... *
Source on Save
Run
Source
Environment
History
Connections
Import Dataset
Grid
Global Environment
Files
Plots
Packages
Help
Viewer
New Folder
Delete
Rename
More
Home › R_Users_Workshop › 8_weeks_Oct-Dec_17 › Workshop_1 › workshop_1_project
29:1
Section 1: Data import, tidying and transformation
R Script
Console
Terminal

```r

raw_gene_df <- read_delim("Brauer2008_DataSet1.tds", delim = "\t")

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 != "")
```

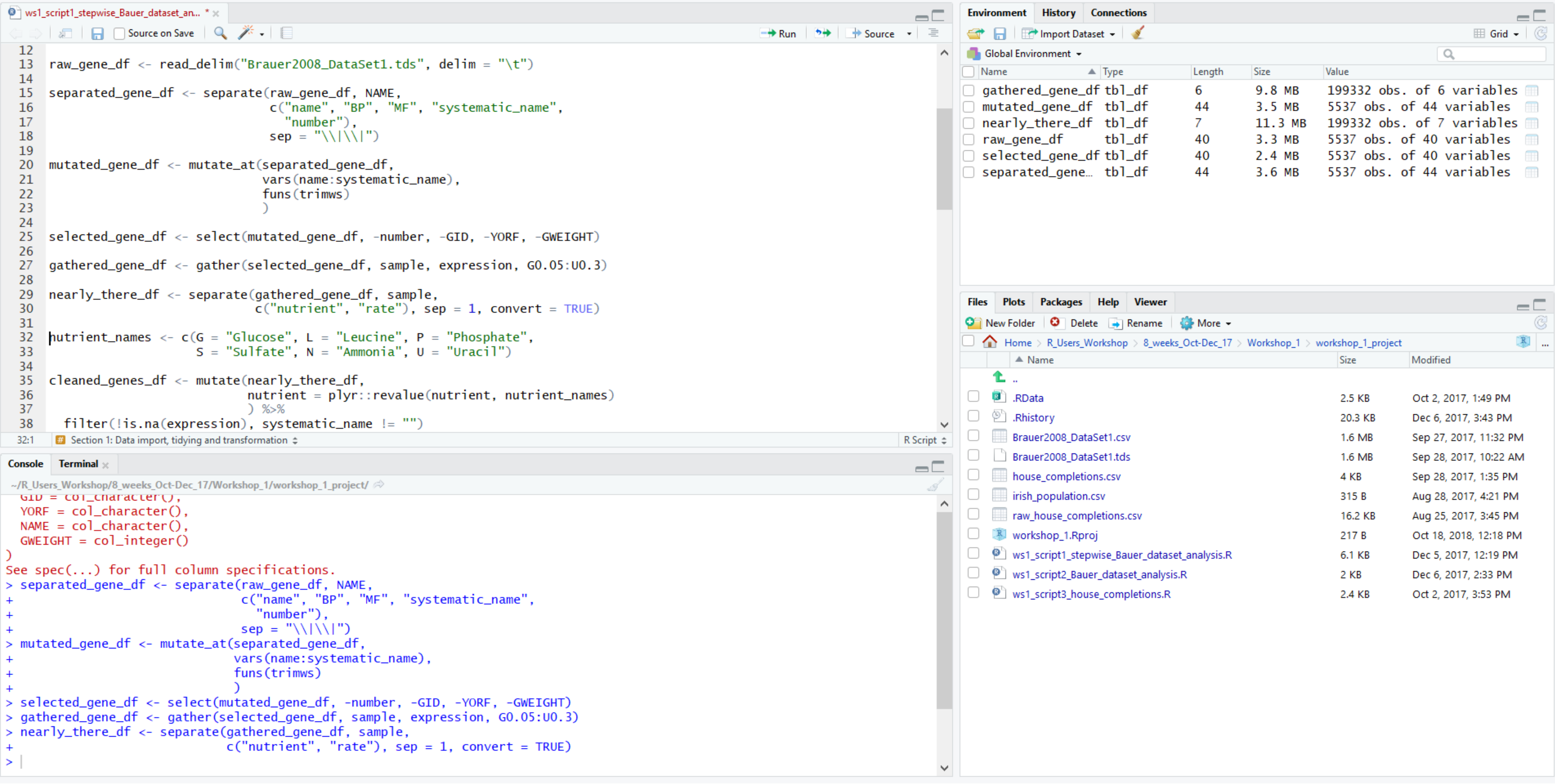
~/R_Users_Workshop/8_weeks_Oct-Dec_17/Workshop_1/workshop_1_project/

```
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)
> 
```

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

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... *
Source on Save
Run
Source
Environment
History
Connections
Import Dataset
Grid
Global Environment
Files
Plots
Packages
Help
Viewer
New Folder
Delete
Rename
More
Home › R_Users_Workshop › 8_weeks_Oct-Dec_17 › Workshop_1 › workshop_1_project
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/

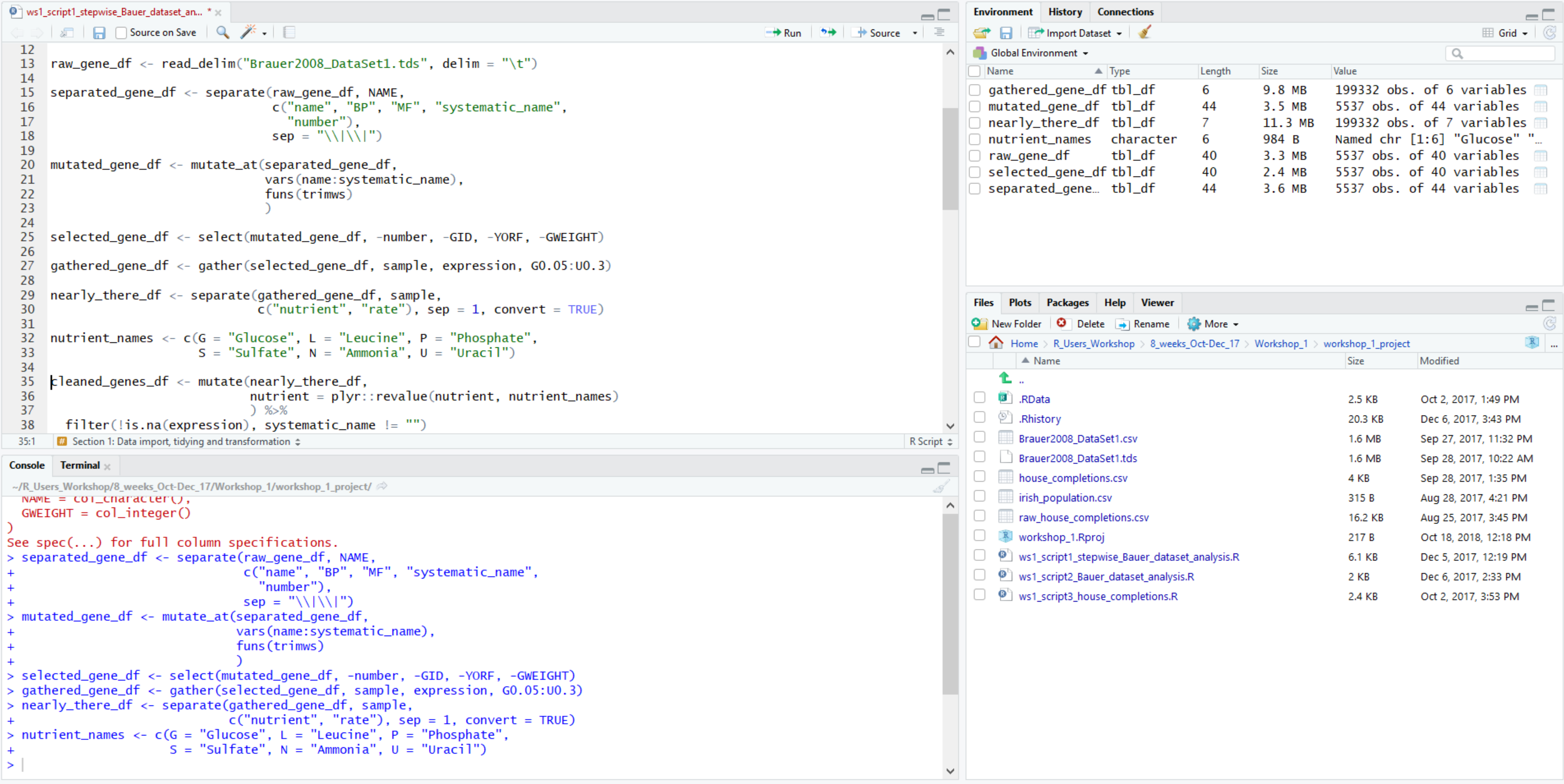
```r

raw_gene_df <- read_delim("Brauer2008_DataSet1.tds", delim = "\t")

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 != "")
```

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

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

```
  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



```
raw_gene_df <- read_delim("Brauer2008_DataSet1.tds", delim = "\t")

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 != "")
```

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

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

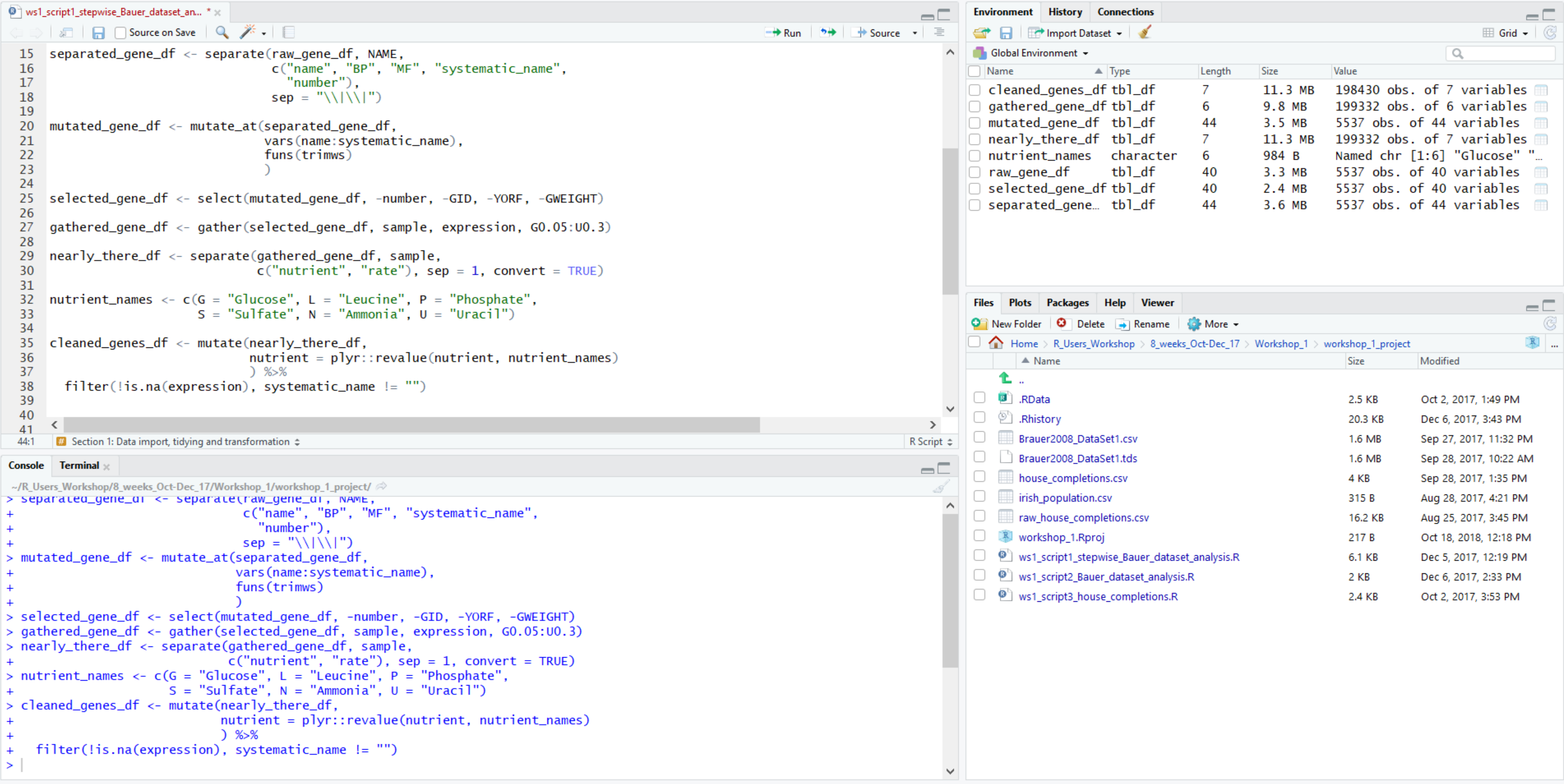
```
  NAME = col_character(),
  GWEIGHT = col_integer()
)
See spec(...) for full column specifications.
> separated_gene_df <- separate(raw_gene_df, NAME,
+                               c("name", "BP", "MF", "systematic_name",
+                                 "number"),
+                               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")
> 
```

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

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



ws1_script1_stepwise_Bauer_dataset_an... *
Source on Save
Run
Source
Environment
History
Connections
Import Dataset
Grid
Global Environment
Files
Plots
Packages
Help
Viewer
New Folder
Delete
Rename
More
Home › R_Users_Workshop › 8_weeks_Oct-Dec_17 › Workshop_1 › workshop_1_project
Console
Terminal
~/R_Users_Workshop/8_weeks_Oct-Dec_17/Workshop_1/workshop_1_project/

```r
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 != "")
```

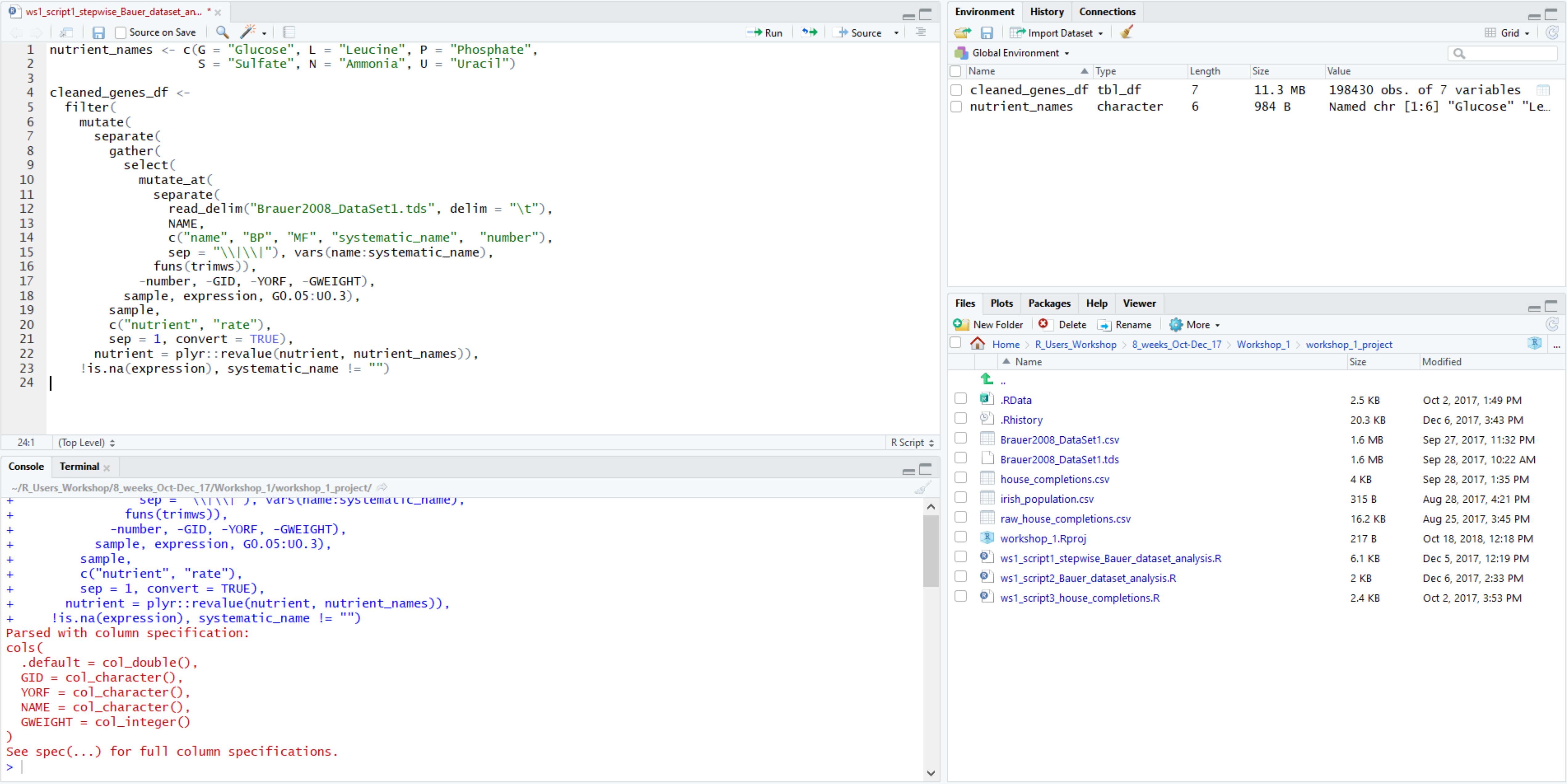
44:1 Section 1: Data import, tidying and transformation R Script

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

```
> 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 != "")
> 
```

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



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

cleaned_genes_df <-
  filter(
    mutate(
      separate(
        gather(
          select(
            mutate_at(
              separate(
                read_delim("Brauer2008_DataSet1.tds", delim = "\t"),
                NAME,
                c("name", "BP", "MF", "systematic_name",  "number"),
                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 != "")
```

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

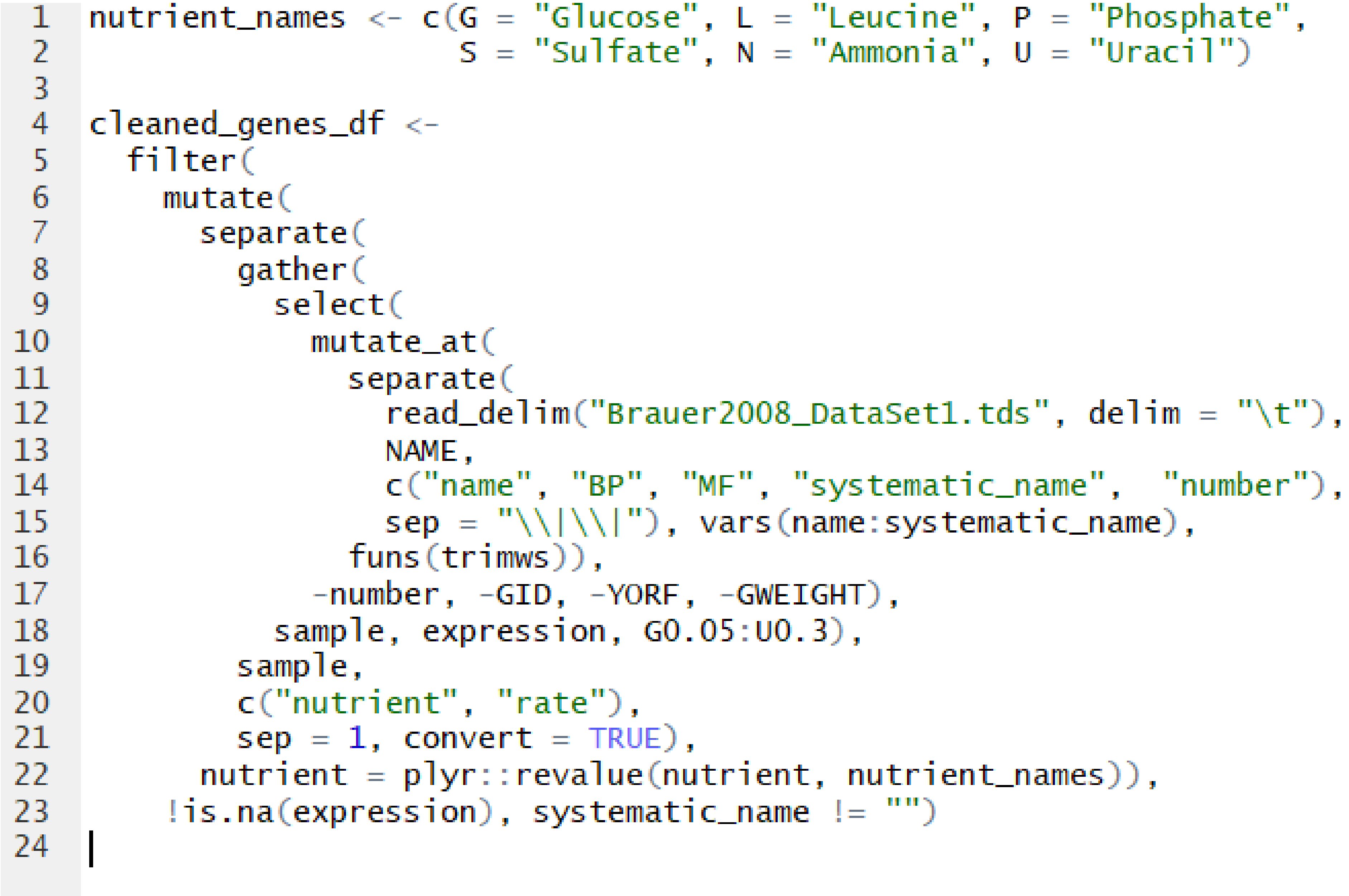
```
+                sep = "\\|\\|"), vars(name:systematic_name),
+              funs(trimws)),
+            -number, -GID, -YORF, -GWEIGHT),
+          sample, expression, G0.05:U0.3),
+        sample,
+        c("nutrient", "rate"),
+        sep = 1, convert = TRUE),
+      nutrient = plyr::revalue(nutrient, 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.
> 
```

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…

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



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

cleaned_genes_df <-
  filter(
    mutate(
      separate(
        gather(
          select(
            mutate_at(
              separate(
                read_delim("Brauer2008_DataSet1.tds", delim = "\t"),
                NAME,
                c("name", "BP", "MF", "systematic_name",  "number"),
                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 != "")

```

Putting the pieces together

% > %

Code structure has two main forms

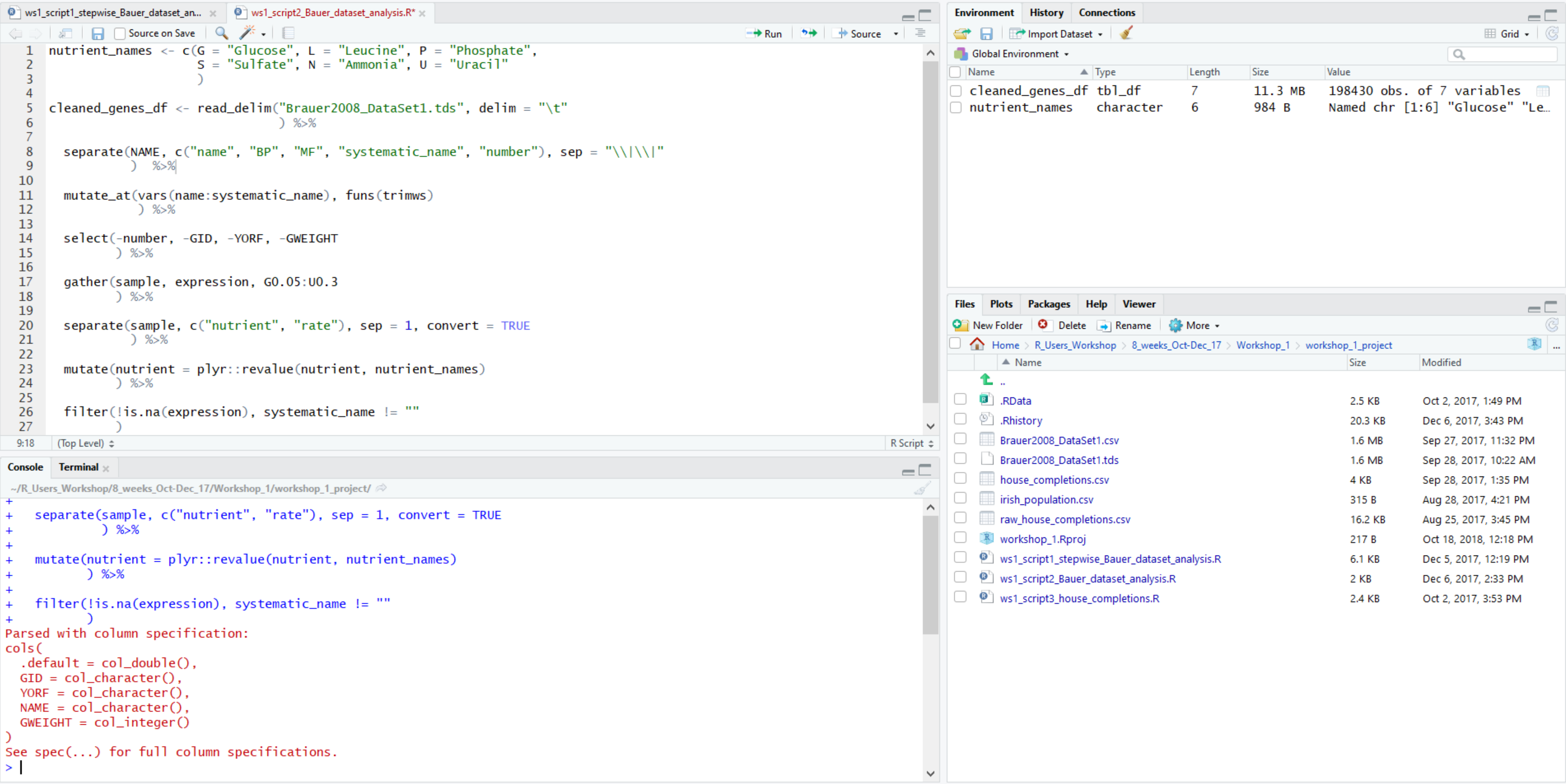
(1)

```
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



```
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 != ""
         )
```

9:18 (Top Level) R Script

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

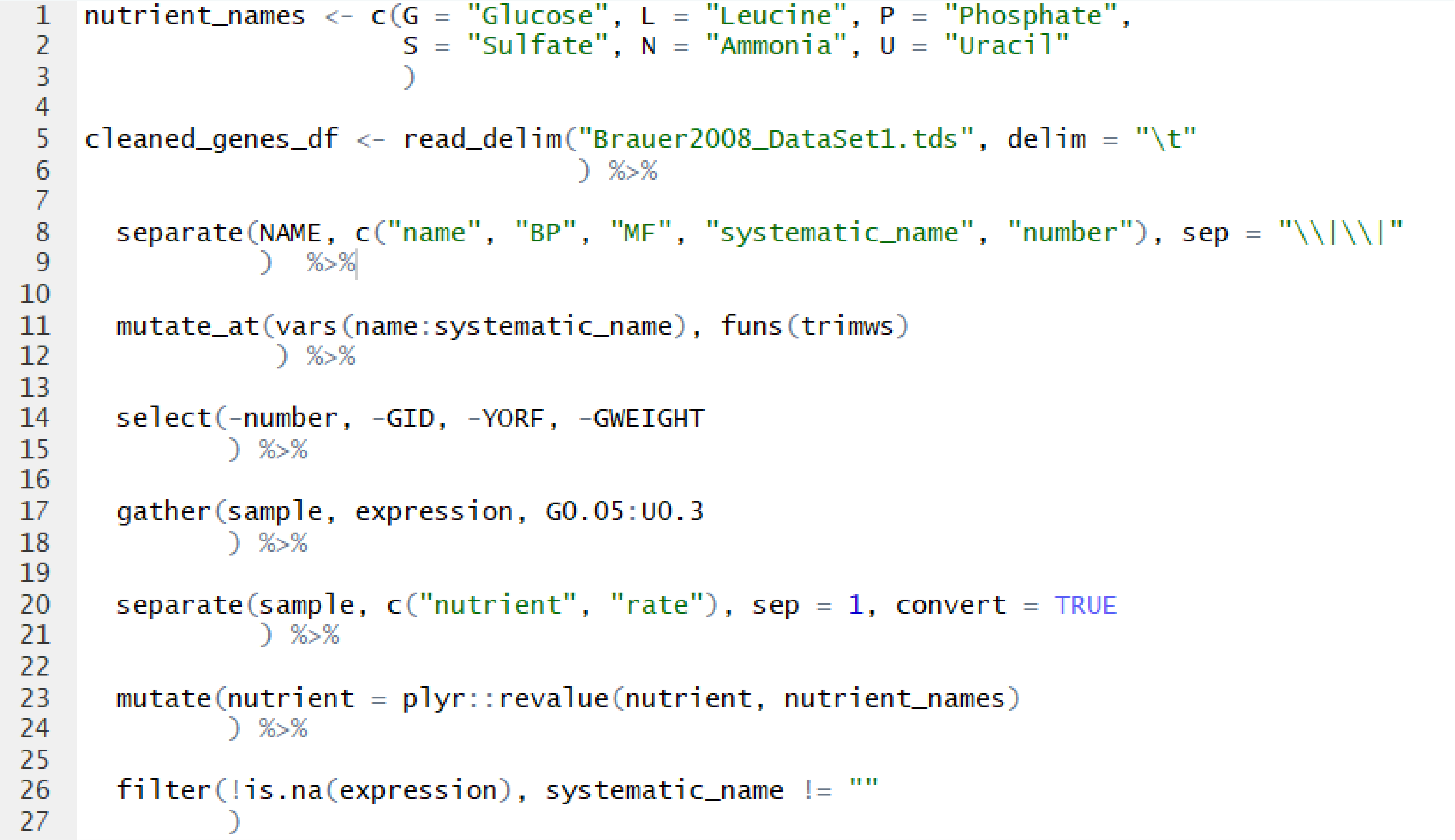
```
+
+   separate(sample, c("nutrient", "rate"), sep = 1, convert = TRUE
+            ) %>%
+
+   mutate(nutrient = plyr::revalue(nutrient, nutrient_names)
+          ) %>%
+
+   filter(!is.na(expression), systematic_name != ""
+          )
Parsed with column specification:
cols(
  .default = col_double(),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

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...

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 != ""
         )
```

Moral of the story…

You can go from this

+

=

To this!!

Master Builder!!