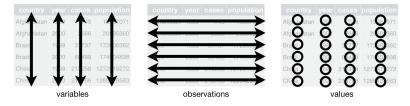
Tutorial - R - Data Wrangling

Psychology Tutorial Series - Emily Towner

May 10, 2021

What is tidy data?

- ▶ A way to organize your data that is consistent across datasets
- ► Each variable is a column, observation is a row, and cell is a value



"Tidy datasets are all alike, but every messy dataset is messy in its own way." – Hadley Wickham

What is tidyverse?

There are several ways to wrangle your data into a tidy format, but *tidyverse* has some great/easy functions.

The *tidyverse* is a coherent system of packages for data manipulation, exploration and visualization.

Includes:

- ggplot2 graphics and plots
- dplyr data manipulation
- tidyr to create tidy data format
- & others

Filter

- ▶ Use *filter* to include only observations that meet a specific criteria.
- ► For example, if I want a dataset consisting only of females, or only of those individuals with depression above the average depression score in this sample.
- ► This is useful when your data contains many waves/arms.

Filter

```
head(data_female[,c(1,2,4)])
   # A tibble: 6 \times 3
##
     participant group
                                     sex
##
           <dbl> <chr>
                                     <chr>>
                1 Early-Adolescence Female
## 1
               3 Mid-Adolescence
                                    Female
##
               7 Late-Adolescence Female
## 3
## 4
                 Early-Adolescence Female
## 5
               10 Early-Adolescence Female
## 6
               13 Early-Adolescence Female
```

Filter

```
mean(data$depression_1, na.rm = T)
## [1] 55.71905
head(data_above_average[,c(1,2,4,9)])
## # A tibble: 6 x 4
##
                                            depression_1
     participant group
                                    sex
##
           <dbl> <chr>
                                    <chr>>
                                                   <dbl>
## 1
               6 Mid-Adolescence
                                    Male
                                                    64.7
## 2
               7 Late-Adolescence
                                    Female
                                                    66.6
              10 Early-Adolescence Female
##
                                                    55.9
##
              11 Late-Adolescence
                                    Male
                                                    70.5
## 5
              13 Early-Adolescence Female
                                                    79.1
              14 Mid-Adolescence
                                                    56.5
##
                                    Female
```

Select

- Use select to include only the variables of interest
- ► For example, in subset 1 if I wanted data that only contains "age, sex, anxiety and early life stress" variables
- Or, use a selection helper like *contains*, to subset data with all variables that contain the string "depression"
- Other helpers include starts_with, ends_with, and matches
- ► This is useful when subsetting and scoring individual measures that are named with a convention (e.g. bdi_1, bdi_2, etc.)

Select

```
head(data_subset_1)
```

```
# A tibble: 6 x 5
##
     participant
                    age sex
                               anxiety early_life_stress
##
           <dbl> <dbl> <chr>
                                  <dbl>
                                                     <dbl>
## 1
                     10 Female
                                  54.9
                                                      42.8
## 2
                2
                     10 Male
                                  74.4
                                                      27.2
               3
## 3
                     16 Female
                                  77.1
                                                      48.5
                                                      60.1
## 4
               4
                     17 Male
                                  73.0
               5
## 5
                     23 Male
                                  58.5
                                                      48.6
                6
                                                      57.1
## 6
                     15 Male
                                  93.1
```

head(data_subset_2)

5

6

Select

5

6

```
# A tibble: 6 \times 4
##
##
     participant depression_1 depression_2 depression_3
            <dbl>
##
                          <dbl>
                                         <dbl>
                                                       <dbl>
## 1
                           43.8
                                          36.5
                                                        24.3
## 2
                           52.1
                                         43.4
                                                        28.9
                3
                                          26.3
## 3
                           26.4
                                                        24.7
                                                        17.5
## 4
                4
                           41.6
                                          29.7
```

45.6

64.7

45.6

46.2

46.3

27.2

Mutate

Use mutate to add new variables while preserving existing ones

```
## # A tibble: 6 x 4
##
     participant group
                                     anxiety anxiety_norm
##
           <dbl> <chr>
                                       <dbl>
                                                    <dbl>
## 1
                1 Early-Adolescence
                                        54.9
                                                    0.730
                                       74.4
## 2
               2 Early-Adolescence
                                                    0.989
               3 Mid-Adolescence
                                       77.1
                                                    1.03
## 3
## 4
               4 Mid-Adolescence
                                       73.0
                                                    0.971
                                        58.5
                                                    0.777
## 5
               5 Late-Adolescence
## 6
               6 Mid-Adolescence
                                        93.1
                                                     1.24
```

BONUS

Normal Syntax

Can be difficult to read when things get complicated (focus on nouns, must read from inside out).

▶ In an object called data_new, mutate a variable using data.

Pipe Operator (technically from magrittr package)

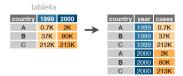
- ► Focuses on verbs, easier to read
- ▶ In an object called data_new, take my data and then mutate it.
- Can be used with all functions

Pivot

Convert data from long to wide format and vice versa

pivot_longer(data, cols, names_to = "name",
 names_prefix = NULL, names_sep = NULL,
 names_pattern = NULL, names_totypes = list(),
 names_transform = list(), names_repair =
 "check_unique", values_to = "value", values_drop_na =
 FALSE, values_ptypes = list(), values_transform =
 list(), ...)

pivot_longer() pivots **cols** columns, moving column names into a **names_to** column, and column values into a **values to** column.



pivot_longer(table4a, cols = 2:3,
 names_to = "year", values_to = "cases")

pivot_wider(data, id_cols = NULL, names_from = name, names_prefix = "", names_sep = "_", names_glue = NULL, names_sort = FALSE, names_repair = "check_unique", values_from = value, values_fill = NULL, values_fn = NULL, ...)

pivot_wider() pivots a names_from and a values_from column into a rectangular field of cells.



pivot_wider(table2, names_from = type, values_from = count)

Separate

```
separate(data, col, into, sep = "[^[:alnum:]]
+", remove = TRUE, convert = FALSE,
extra = "warn", fill = "warn", ...)
```

Separate each cell in a column to make several columns.

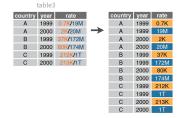


country	year	rate		country	year	cases	рор
Α	1999	0.7K/19M		Α	1999	0.7K	19M
Α	2000	2K/20M	\rightarrow	Α	2000	2K	20M
В	1999	37K/172M		В	1999	37K	172
В	2000	80K/174M		В	2000	80K	174
С	1999	212K/1T		С	1999	212K	1T
С	2000	213K/1T		С	2000	213K	1T

separate(table3, rate, sep = "/",
into = c("cases", "pop"))

separate_rows(data, ..., sep = "[^[:alnum:].]
+", convert = FALSE)
Separate each cell in a column to make

several rows.

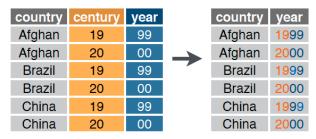


separate_rows(table3, rate, sep = "/")

My favorite tidyr functions
Unite

Collapse cells across several columns to make a single column.

table5



Cheatsheets and More

► Find cheatsheets and more for each package at: https://tidyverse.tidyverse.org/

Cheat Sheet



- Now for a data cleaning tutorial!