Section A01
CDS 101 *Introduction to Computational and Data Sciences*

Class 10: Tidy data

June 4, 2018

General

Announcements

- Homework 2 due on June 6th @ 11:59pm:
  http://summer18.cds101.com/assignments/homework-2/

`tidyr` package continued

`tidyr` verbs

- `gather()`: transforms wide data to narrow data
- `spread()`: transforms narrow data to wide data
- `separate()`: make multiple columns out of a single column
- `unite()`: make a single column out of multiple columns

Simple examples from textbook

Follow along in RStudio

spread() example

Untidy data frame stored in `table2`

country	year	type	count
Afghanistan	1999	cases	745
Afghanistan	1999	population	19987071
Afghanistan	2000	cases	2666
Afghanistan	2000	population	20595360
Brazil	1999	cases	37737
Brazil	1999	population	172006362
Brazil	2000	cases	80488
Brazil	2000	population	174504898
China	1999	cases	212258
China	1999	population	1272915272
China	2000	cases	213766
China	2000	population	1280428583

spread() example

```
table2 %>%
  spread(key = type, value = count)
```

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

`separate()` example

Untidy data frame stored in `table3`

country	year	rate
Afghanistan	1999	745/19987071
Afghanistan	2000	2666/20595360
Brazil	1999	37737/172006362
Brazil	2000	80488/174504898
China	1999	212258/1272915272
China	2000	213766/1280428583

separate() example

```
table3 %>%
  separate(
    col = rate,
    into = combine("cases", "population")
  )
```

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

separate() example

```
table3 %>%
  separate(
    col = rate,
    into = combine("cases", "population"),
    sep = "/",                  # Set the separating symbol
  )
```

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

separate() example

```
table3 %>%
  separate(
    col = rate,
    into = combine("cases", "population"),
    sep = "/",                  # Set the separating symbol
    convert = TRUE              # Convert data types, ensures that
  )                             # cases and population cols are numeric
```

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

`unite()` example

Untidy data frame stored in `table5`

country	century	year	rate
Afghanistan	19	99	745/19987071
Afghanistan	20	00	2666/20595360
Brazil	19	99	37737/172006362
Brazil	20	00	80488/174504898
China	19	99	212258/1272915272
China	20	00	213766/1280428583

unite() example

```
table5 %>%
  unite(new, century, year)
```

country	new	rate
Afghanistan	19_99	745/19987071
Afghanistan	20_00	2666/20595360
Brazil	19_99	37737/172006362
Brazil	20_00	80488/174504898
China	19_99	212258/1272915272
China	20_00	213766/1280428583

unite() example

```
table5 %>%
  unite(new, century, year, sep = "")
```

country	new	rate
Afghanistan	1999	745/19987071
Afghanistan	2000	2666/20595360
Brazil	1999	37737/172006362
Brazil	2000	80488/174504898
China	1999	212258/1272915272
China	2000	213766/1280428583

unite() example

```
table5 %>%
  unite(new, century, year, sep = "") %>%
  mutate(new = as.integer(new)) %>%   #  Change data type to integer
  rename(year = new)                  #  Rename column to year
```

country	year	rate
Afghanistan	1999	745/19987071
Afghanistan	2000	2666/20595360
Brazil	1999	37737/172006362
Brazil	2000	80488/174504898
China	1999	212258/1272915272
China	2000	213766/1280428583

Class activity

Tidy gradebook dataset exercise

Download the Github Classroom repo linked in channel `#su18-a01-activities` on Slack and complete the following exercises:

1. Make the dataset tidy using either `gather()` or `spread()`. The tidy gradebook should have one observation per row, which gives all the grades a student has received for the different assignments in the semester.

2. Use the tidy gradebook and create a histogram that answers the question, "What was the grade distribution for the Midterm Exam?"

Remember to commit and push your work!

Introduce the midterm project

Midterm project instructions

Follow along on printed handout.

Credits

- Examples in the section tidyr package continued taken from Chapter 12 of *R for Data Science* written by Garrett Grolemund and Hadley Wickham and made available under the CC BY-NC-ND 3.0 license.