INFO201 lab: cleaning and reshaping

May 20, 2024

Instructions

This is a graded lab. It asks you to reshape data. Please:

- Do it in rmarkdown.
- Include any images, screenshots etc as images in markdown
- submit both html and .rmd on canvas!

Good luck!

1 Reshape manually

Here are a few small data frames. We ask you to do the reshaping *manually*, ie do not use computer code to do it. You can either do it on paper (and include the resulting image), use markdown tables, or write the reshaped data frame in a spreadsheet (like Excel) and include the screenshot here as an image.

As a refresher: markdown table for the soccer results below can be written as

```
| club | opponent | result |
| -----| ---------| ------ |
| P    | AM       | 1:0    |
| P    | I        | 2:1    |
| AM   | P        | 0:0    |
...
```

1.1 Soccer clubs

A tournament between soccer 3 clubs–*Palmeiras* (P), *Atlético Mineiro* (AM) and *Internacional* (I) gave the following results:

club	opponent	result
Palmeiras	Atlético Mineiro	1:0
Palmeiras	Internacional	2:1
Atlético Mineiro	Palmeiras	0:0
Atlético Mineiro	Internacional	1:1
Internacional	Palmeiras	2:3
Internacional	Atlético Mineiro	1:2

1. Is the data in long or wide form? Explain why do you think so!
2. Convert it into the other form!

1.2 Publicly traded businesses

Consider data about businesses:

Name	NASDAQ	established	revenue ($B)	net income	assets
Amazon	AMZN	1994.00	574.00	30.40	528.00
Google	GOOG	1998.00	307.00	79.80	402.00
...					

1. Are these data in a wide form or a long form? Explain!

   Note: this dataset contains 3 id-variables!
2. Transform it to the other form.

---

2 Reshape

Here we use UAH lower troposphere data, wide form *UAH-lower-troposphere-wide.csv.bz2*.

1. Load the dataset, and ensure it is good.

   We only use variables *year*, *month*, *globe*, *nopol*, *sopol* below, just drop all the other columns. This is in order to make the output better visible.

   The first few lines of the dataset should look like

```
## # A tibble: 3 x 5
##    year month globe nopol sopol
##   <dbl> <dbl> <dbl> <dbl> <dbl>
## 1  1978    12 -0.48 -0.39 -0.46
## 2  1979     1 -0.48 -0.46 -0.16
## 3  1979     2 -0.44 -2.01 -0.8
```

2. Is this in a wide form or in a long form? Explain!
3. Now reshape the dataset into a long form: collect all values into a column *temperature* and the regions *globe*, *nopol*, *sopol* into a column *region*.

   Store the long form dataset into a variable.
4. Now take your long form dataset and reshape it back into the wide form. The result should look like the original data!
