Chapter 6: Tidy data

Many reasons to prefer the ‘long’ or ‘tall’ version of representing/working with data

-It is a more efficient way to store data

-more convenient for data analysis

-it is more scalable: columns can be added with ease unlike in a wide format dataset

\*always keep your original dataset and your analyses separate. Modify the original dataset with code in a reproducible way that can be traceable to the original file you have.

Tidy data: the arrangement of data in a simple pattern in systemically defined data tables with rows, columns in long format.

Two simple rules:

1. The rows specify a unique ‘case’: a person, named X in year Y. ‘cases’ refer to the attributes that make that row unique.
2. The columns specify the values that collectively make each row unique

‘When data are tidy, I is relatively straightforward to transform your data into arrangements that are more useful for answering interesting questions’

Transforming implicit information from a data table into a new data table that make the information explicit is called data wrangling.

Variables: a known quantity in a data table that can either be categorical (group) or quantitative (number)

Codebook: should include information about how the data were collected and what constitutes a cases.

‘Being tidy is not about shoving everything into one table’

While tidy prefers the long format, sometimes wide format is easier when doing certain operations, like using values in two columns to produce a new value in a new column.

Naming conventions:

-Can’t start a name with a number

-Best to avoid ‘.’ In names and stick with ‘\_’

-R is case sensitive, capital letters produce different names than lower case letters.

Tidyverse style guide:

<https://style.tidyverse.org>

[*Data cleaning*](https://en.wikipedia.org/w/index.php?search=Data%20cleaning) refers to taking the information contained in a variable and transforming it to a form in which that information can be used.

Advisable to store categorical data as factors unless you have a clear reason why they need to be factors. Always good to check this.