Tidy Data

Prof. Eric A. Suess

September 14, 2020

Tidy Data

Tidy data is *long* and *narrow*.

Each row is an *observation* or *case*.

Each column is a *variable*.

Examples:

1. gapmider
2. babynames

Tidy Data

Being *tidy* does not mean neat.

All columns need to contain data from the variable in that column only.

All rows contain data from that specific observation in that row only.

Variables

Each column in a tidy dataset contains a *variable*.

Each variable is either *categorical* or *numeric*.

Categorical variables are often stored as *factors* in R.

Codebooks

Codebooks were common when variable names were short due to computer memory restrictions.

Now days variable names can be as long as you need. So codebooks are less important.

```
> ??gapmider
```

```
> ?babynames
```

Reshaping data

mutate()

spread()

gather()

tidyr 1.0

These functions have been updated to the new functions.

See the tidyr website.

pivot_longer()

pivot_wider()

Some examples

1. finance.yahoo.com What is the current value of Ford? Google? Apple? Is the data in a tidy format? How to download the data? Is the downloaded .csv file tidy?
2. SF Open Data Pick a topic of interest. I looked at Transportation and searched for Parking data. Is the data in a tidy format? Click on View Data. Is the data in a tidy format?

Basically all data that is available through and API is in a tidy format.

R Style Guides

1. Tidyverse style guide
2. Advanced R style guilde
3. Google's R style guide

Take a look at the examples in R

Take a look at the variable names in the R datasets.

1. Is the gapmider dataset tidy? What do you think of the variable names?
2. Is the babynames dataset tidy? What do you think of the variable names? Note that *n* is used as one of the variable names. It represents *counts* not sample size. My suggestion, don't use *n* as a variable name, it can be confusing.