Key concepts for working with data

Author: Nicholas G Reich

Foundational concepts for data

Unit of observation

- What uniquely defines a row of your dataset.

Sampling Frame

- "a list of the items or people forming a population from which a sample is taken" (Oxford Languages via Google).

Tidy data

- a set of principles about how to store data in a standardized way, for easier interoperability across data wrangling, visualization, and modeling tools (hence, the 'tidyverse'!)

Unit of observation

The minimum set of features that uniquely define a row of your data frame.

Sometimes this is simple: a single song, a single person, a geographic location.

Unit of observation

The minimum set of features that uniquely define a row of your data frame.

But sometimes it is more complicated, like when you have multiple row for a person, or a song.

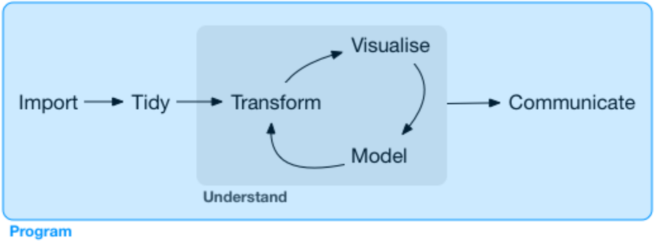
Sampling frames and populations

Three features dictate what conclusions can be drawn.

- **Population**: “set of all possible units which might have been included” (Daniel Kaplan)
- **Sampling frame**: “a list of the items or people forming a population from which a sample is taken” (Oxford/Google)
- **Sample**: “a selection of cases from the population“ (Daniel Kaplan)

For the Sad Songs analysis:

Tidy data is a promise for downstream tools



Central principles

- Each variable forms a column.
- Each observation forms a row.
- Each type of observational unit forms a table.

Breakout rooms

Revisit “Sad songs” and SAT analyses in the context of these concepts.