tidyR

The two most important properties of tidy data are:

- Each column is a variable.
- Each row is an observation.

Tidy Data with R

Why should that be?
R follows a set of conventions that makes one layout of tabular data much easier to work with than others. Your data will be easier to work with in R if it follows three rules

- Each variable in the data set is placed in its own column
- Each observation is placed in its own row
- Each value is placed in its own cell

Data that satisfies these rules is known as tidy data.

Principles of Tidy Data

- Tidy data was popularized by Hadley Wickham, and it serves as the basis for many R packages and functions.
- You can learn more about tidy data by reading Tidy Data a paper written by Hadley Wickham and published in the Journal of Statistical Software.
- Tidy Data is available online at www.jstatsoft.org/v59/i10/paper.

Principals of Tidy Data

- Wickhams idea leverages from ideas of relational databases and database normalization from computer science, although his audience is statisticians and data analysts.
- He starts off by defining terms, suggesting that talking about rows and columns is not rich enough:

- The data is a collection of values of a given type
- Every value belongs to a variable
- Every variable belongs to an observation
- Observations are variables for a unit (like an object or an event).

- Variables are columns, observations are rows and types of observations are tables.
- Classically, Wickham relates this to third normal form from relational database theory.
- He also describes types of variables as fixed and measured and suggests organizing fixed before measured in a table.