# Tidy Data

The concept of *tidy data* was coined by Hadley Wickham in his paper Journal of Statistical Software (2014). In it, he defined the basic principles for what is needed to make data tidy. Too often, data recorded in a spread sheet is made to look nice but not be easily used by other programs or people for analysis. Tidy data files allow for better access to data by programs and visualizations.

#### Tidy Data

1. Each variable recorded should be in one column.
2. Each different observation of that variable should be in a different row.
3. There should be a table (sheet) for each *"kind"* of variable.
4. If you have multiple tables (sheets), they should include a column in the table that allows the tables to be linked.

#### Other important items

* Include in the first row the vaiable names for each column of data.
* Make variable names human readable
  + mnt => month
* Sometimes better to save each table (sheet) as its own file.
* File formats should follow open formats rather than proprietary.
  + Comma separated variables (CSV) rather than MS Excel (xlsx)

#### Common Mistakes

* Column headers are values, not variable names.
* Multiple variables are stored in one column.

|  |  |
| --- | --- |
| Gender/Age |  |
| M34 |  |
| F25 |  |
| F34 |  |

Should be

|  |  |
| --- | --- |
| Age | Gender |
| 34 | M |
| 25 | F |
| 34 | F |

* Variables are stored in both rows and columns.
* Multiple kinds of observational units are stored in the same table.

***Reference***

Wickham, Hadley. “Tidy Data.” Journal of Statistical Software 59, no. 1 (2014): 1–23. doi:10.18637/jss.v059.i10.

***Contact***

Mark Laufersweiler

Email: laufers@ou.edu

Calendar: http://libcal.ou.edu/appointment/9321