Introduction

Working With “Flat” Data Files

Will Doyle

Flat Data Files

Flat data has the following characteristics

- Each row is a case
- Rows may or may not be labeled
- Each column is a variable
- Columns are almost always labeled

When we say “dataset” this is what we almost always mean

Types of Flat Data Files

- Delimited: comma-separated (csv), tab-separated
- Fixed width (fwf)
- Other languages: Stata, SPSS, SAS, and so on

Keys for Organizing Flat Data

- No extraneous information (metadata) included **inside** dataset
- Clear column (variable) names
- Clear value labels
- Understanding units of analysis

Flat Data vs. Excel

- Excel is organized in rows and columns.
- **But** there are no constraints.
- Cells may contain data and/or calculations.
- In our work, we always separate data and results of data analysis.
- Excel spreadsheets can be organized into datasets, but it’s work.

VANDERBILT
PEABODY COLLEGE

Databases and Flat Files

Will Doyle

Databases

A database:

- Contains multiple tables
- Contains tables that are linked to one another via certain ids
- Can contain many different units of analysis
- Can cover multiple domains

Datasets

Datasets:

- Have unique ids for each case
- Have single units of analysis
- Cover a single conceptual domain

A Dataset Is Not a Database

- Databases can be used to generate single flat files.
- Each flat file can be a dataset.
- But don’t get databases and datasets confused!
- What’s needed for statistical analysis is a dataset.

VANDERBILT
PEABODY COLLEGE

Introduction to “Tidy Data”

Will Doyle

Tidy Data

Concept is due to Hadley Wickham. In tidy data:

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

This structure makes life MUCH easier for the analyst.

Messy Data

Any time data aren’t tidy, they’re messy. What makes a messy dataset?

- Column headers are values, not variable names (instead of month, column header is “July”).
- Multiple variables are stored in one column.
- Multiple types of units are stored in the same table (including both score and rank of score).
- A single observational unit is stored in multiple tables.

Messy Data

- Much of the time, the data we get will not be ready for analysis.
- It takes time and care to restructure data to be “tidy.”
- Think **carefully** about the desired data structure before you begin wrangling.

VANDERBILT
PEABODY COLLEGE