DATA WRANGLING

DATA COLLECTION AND DATA PROCESSING

LEARNING OBJECTIVES

Become familiar with the tidy data format

Increase familiarity of data wrangling functions

Identify R packages that facilitate data processing

DATA WRANGLING

A fair amount of time (up to 80%, perhaps) must be spent on data processing (both cleaning and manipulation).

The main goals of **data wrangling** are to:

- make the data useable by a specific piece of software
- reveal pre-analysis insights in the data

TIDY DATA

Tidy data has a specific structure:

- each variable is a column
- each observation is a row
- each type of observational unit is a table

Country	2011	2012	2013
FR	7000	6900	7000
DE	5800	6000	6200
US	15000	14000	13000

vs.

Country	Year	n
FR	2011	7000
DE	2011	5800
US	2011	15000
FR	2012	6900
DE	2012	6000
US	2012	14000
FR	2013	7000
DE	2013	6200
US	2013	13000

FUNCTIONALITY

Data wrangling functions should allow the analyst to:

- extract a subset of variables from the data frame
- extract a subset of observations from the data frame
- sort the data frame along any combination of variables in increasing or decreasing order
- to create new variables from existing variables
- to create (so-called) pivot tables, by observation groups
- database functionality (joins, etc.)
- etc.

FUNCTIONALITY

In R, this can be achieved in various ways. Current favoured packages include:

- `tidyr`
- `dplyr` (data transformation)
- `lubridate` (dates and times)
- `stringr` (string manipulation)
- `purrr` (functions)
- `readr` (data import)

For equivalent Python modules, consult Kazil & Jarmul's *Data Wrangling with Python*.

EXERCISES

What would the following dataset look like in a tidy format?

storm	stat	value
Alex	wind	68
Alex	pressure	130
Allison	wind	55
Allison	pressure	121
Bobbie	wind	72
Bobbie	pressure	118

EXERCISES

How would you go from the table on the left to the table on the right?

storm	stat	value
Alex	wind	68
Alex	pressure	130
Allison	wind	55
Allison	pressure	121
Bobbie	wind	72
Bobbie	pressure	118

stat	mean	std dev
wind	65	8.9
pressure	123	6.2

EXERCISES

Run section 9 of the notebook `CSPS 04 R Basics.ipynb` to explore how the packages `tidyr` and `dplyr` help the process of data wrangling in R.

EXERCISES

Turn the data found in `cities.txt` into a tidy dataset.