WHO - TB example

- The who data set in the DSR package contains cases of tuberculosis (TB) reported between 1995 and 2013 sorted by country, age, and gender.
- The data comes in the 2014 World Health Organization Global Tuberculosis Report, available for download at

  `www.who.int/tb/country/data/download/en/`
- The data provides a wealth of epidemiological information, but it would be difficult to work with the data as it is.

Tidy Data with R

To see the data in its raw form, load DSR with library(DSR) then run
View(who)
A subset of the who data frame displayed with View().
who provides a realistic example of tabular data in the wild. It contains redundant columns, odd variable codes, and many missing values. In short, who is messy.

TIP
The View() function opens a data viewer in the RStudio IDE. Here you can examine the data set, search for values, and filter the display based on logical conditions. Notice that the View() function begins with a capital V.

Tidy Data with R

The most unique feature of who is its coding system. Columns five through sixty encode four separate pieces of information in their column names:
The first three letters of each column denote whether the column contains new or old cases of TB. In this data set, each column contains new cases.

The next two letters describe the type of case being counted. We will treat each of these as a separate variable.

- rel stands for cases of relapse
- ep stands for cases of extrapulmonary TB
- sn stands for cases of pulmonary TB that could not be diagnosed by a pulmonary smear (smear negative)
- sp stands for cases of pulmonary TB that could be diagnosed be a pulmonary smear (smear positive)

The sixth letter describes the sex of TB patients. The data set groups cases by

The remaining numbers describe the age group of TB patients. The data set groups cases into seven age groups:

- 014 stands for patients that are 0 to 14 years old
- 1524 stands for patients that are 15 to 24 years old
- 2534 stands for patients that are 25 to 34 years old
- 3544 stands for patients that are 35 to 44 years old
- 4554 stands for patients that are 45 to 54 years old
- 5564 stands for patients that are 55 to 64 years

Tidy Data with R

- Notice that the who data set is untidy in multiple ways.
- First, the data appears to contain values in its column names.
- We can move the values into their own column with gather().
- This will make it easy to separate the values combined in each code.

```
who <- gather(who, "code", "value", 5:60)
```

Tidy Data with R

- We can separate the values in each code with two passes of `separate()`.
- The first pass will split the codes at each underscore.

```
who <- separate(who, code, c("new", "var",
```

- The second pass will split sexage after the first character to create a sex column and an age column.

```
who <- separate(who, sexage, c("sex", "age"
```

Tidy Data with R

- Finally, we can move the rel, ep, sn, and sp keys into their own column names with spread().

```
who <- spread(who, var, value)
```

- The who data set is now tidy. It is far from sparkling (for example, it contains several redundant columns), but it will now be much easier to work with in R.
- We will continue to work with this tidy version of who in Section 3.7, where we will remove the redundant columns and calculate new variables.