Big Data Analytics
Lab 7

Preparing Dataset

WHO.csv (uploaded on moodle) is a data set of tuberculosis (TB) reported between 1995 and 2013 sorted by country, age and gender. The data comes from 2013 WHO Global Tuberculosis Report.

The data set is messy, especially the columns. The most unique feature of this data set is its coding system. Columns five through sixty encode four separate pieces of information in their column names:

1. 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.

2. 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 extra-pulmonary TB
   - `sn` - stands for cases of pulmonary TB that could not be diagnosed by a pulmonary smear
   - `sp` - stands for cases of pulmonary TB that could be diagnosed by a pulmonary smear

3. The sixth letter describes the sex of TB patients: `m` for males and `f` for females.

4. The remaining numbers describe the age group of patients:

- `014` - 0 to 14
- `1524` - 15 to 24
- `2534` - 25 to 34
- `3544` - 35 to 44
- `4554` - 45 to 54
- `5564` - 55 to 64
- `65` - 65 or older

Read the `tidy data article` uploaded on moodle. Make the WHO.csv dataset tidy in order to prepare it for performing analysis.