Course of Study Bachelor Computer Science	Exercises Statistics WS 2021/22
Sheet III - Voluntary and Additional	

Case Study Tidying Data

The following exercise is from "R for Data Science", Garrett Grolemund and Hadley Wickham, chapter 12.6. It contains a case study to apply the methods of tidyr for tidying data.

The tidyr::who dataset contains tuberculosis (TB) cases broken down by year, country, age, gender, and diagnosis method. The data comes from the 2014 World Health Organization Global Tuberculosis Report, available at `http://www.who.int/tb/country/data/download/en/`.

The dat set contains redundant columns, columns which are not variables, a lot of missing values, etc.. It is a typical example for a messy data set. We now want to apply commands from the tidyr package to clean the data set

1. Load the data set tidyr::who and inspect the columns of the dat set.
2. Clean the data set.
   (a) Identify columns that are not variables.
   (b) Inspect the gather() command and apply the command to gather together all the columns from new_sp_m014 to newrel_f65. Since we do not know what the values represent, give them the generic name "key". The cells represent the count of cases, therefore use the variable cases. Remove the missing values in the current representation using na.rm.
   (c) Count the values in the new "key" column.
   (d) The values of the new column "key" have the following structure:
      - The first three letters of each column denote whether the column contains new or old cases of TB. In this dataset, each column contains new cases.
      - The next two letters describe the type of TB:
        - 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 gives the sex of TB patients. The dataset groups cases by males (m) and females (f).
- The remaining numbers gives the age group. The dataset groups cases into seven age groups:
  - 014 = 0 - 14 years old
  - 1524 = 15 - 24 years old
  - 2534 = 25 - 34 years old
  - 3544 = 35 - 44 years old
  - 4554 = 45 - 54 years old
  - 5564 = 55 - 64 years old
  - 65 = 65 or older

Unfortunately the names are slightly inconsistent because instead of new_rel we have newrel. Use str_replace() command to replace the characters "newrel" with "new_rel". Separate the values of the column "key".

- Split the codes at each underscore
- Separate the values of sexage after the first character

(e) Remove the redundant columns new, iso2 and iso3.

3. Write all steps to clean the data using pipes.

4. After tidying the data we want to have a table containing for every year the country, the population and the number of all infections. Use the function tally/count().