

C4M Homework: Phase II: Project 1 Preparation Exercise 2, Part B

The World Health Organization provides data on tuberculosis burden in different countries at this website:

<http://www.who.int/tb/country/data/download/en/>

Download the CSV file that you will be working with from:

<https://extranet.who.int/tme/generateCSV.asp?ds=estimates>

and the corresponding data dictionary (i.e., a table that explains what each column means) from:

<https://extranet.who.int/tme/generateCSV.asp?ds=dictionary>

The CSV file is fairly large and cumbersome. One way that scientists make working with large datasets more manageable is by writing a Python program that reads in an entire CSV file, and then creates a new file that contains the information they want (that file can be read using Excel). This is what you will do for this exercise.

Write a function with the signature `write_est_incidence_in_year(src_filename, dst_filename, year)` that reads the data from a file named `src_filename` (the name of the file that you downloaded from the WHO website), and then writes a CSV file named `dst_filename` that contains only two columns: country name, and the estimated incidence of TB (in all forms) per 100,000 population. Only write one row per country — the estimated incidence for the year that was provided to your function as a parameter. You'll need to use the data dictionary that you downloaded to work out which columns in the original file hold the data that you need.

Test your function by running it and then opening the newly-written file in Excel.

Hint: you might like to review what you did for anonymizing data in Exercise 1

Submit your work on MarkUs.