Hands on Lab – Combining Data from Multiple Sources

Your task is to create an R dataframe that shows rates of tuberculosis infection by country. You should use the information from the `tb` database and the attached `population.csv` file. Your R dataframe should have the following columns:

```
Country, Year, Rate
```

Where Rate is defined as Cases/Population.

You should provide all of the code (SQL and/or R) needed, along with documentation. An R Markdown file (or an IPython Notebook for the adventurous) would be a great place for at least the documentation and the R code. Your code should all be in your GitHub repository, with a link to the GitHub repo in your assignment submission.

For you to receive full credit, I should be able to run all of your code, as provided, on my machine.

This assignment is intentionally open ended. As in real world assignments of even moderate complexity, there are reasonable alternative choices for where you do the data preparation work (in SQL, in R, in a combination of the two environments), how you combine the data (and handle any messy data), how you handle NULLs, any necessary simplifying assumptions, etc.

Not required, but to consider: what kind of downstream data analysis and reporting might you want to do on this infection rate data? What kinds of questions could this data help answer?