# Exploring World Bank data (Ch. 1; Ex. 5)

The World Bank data repository was used to gather 3 CSV files each representing time-series data of every single recognized country in the world with respect to one indicator (such as GDP) throughout the years (this varies sometimes starting from the late 1960s) up until 2019. The website of the World Bank contains a search function where the research is able to browse through various indicators as tracked or aggregated by the World Bank repository and data teams. The 3 chosen indicators for all the countries were:

|  |  |  |
| --- | --- | --- |
| Access to electricity (% of population) | Educational attainment, at least completed primary, population 25+ years, total (%) | Urban population (% of total population) |

The local downloads (CSVs) of the data turned out to be quite numerous:

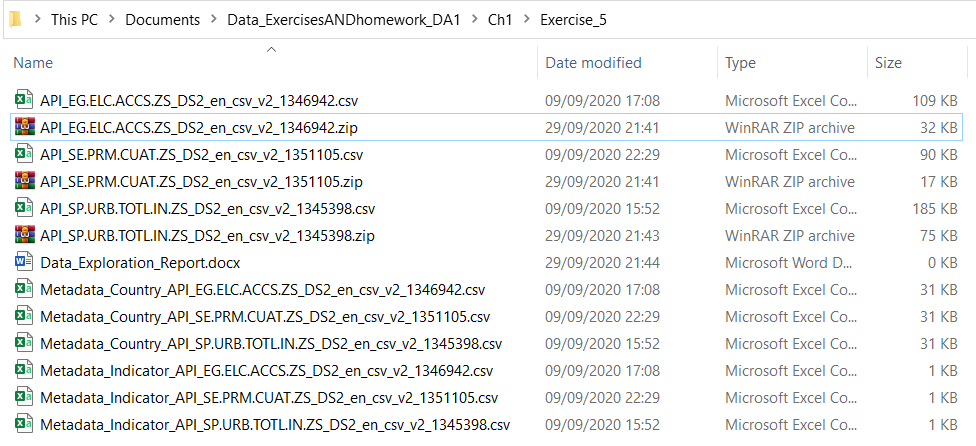


Figure 1: Each indicator's CSV came with 2 supplementary CSVs of metadata, one explaining country codes, denominations and other conventions, while the other explained the values of the indicator variables and how it was calculated.

While the World Bank repository has an easy interface to understand and provides a dashboard to see the progression of each indicator over the years using the time-series data and aggregating all the results year-by-year, the CSV files contain a lot of missing values for each country, ranging from entire decades of missing/untracked data to fewer years of missing data.

The list of countries spans 264 in total, however a fair amount of observations (which are supposed to each represent one country) also contain aggregated data such as “World”, “Upper Middle Income” (referring to countries with upper middle income levels as opposed to the rest), etc. This makes the dataset a lot harder to work with, as the researcher would have to manually erase all observations that are not single states for the purpose of this specific exercise. Furthermore, the columns of years span from 1960 until 2019, with most values for all the observations missing for the first ~30 years of measurements (especially for these indicators). As such, the columns spanning until the 1990s most often do not bring any contribution to the researcher as they are completely empty.

Overall, a lot of difficulties were encountered in understanding the data and I can foresee difficulties in preparing and analyzing the data, as thorough cleaning and merging needs to happen. Seeing as the number of columns is very high, it would triple once all the 3 datasets would be merged if no prior cleaning were done. Even if only the years with non-empty observation values, the final data table would have too many columns to consider. A choice of one year needs to be made in order to compare all the countries on the same indicator. Furthermore, the observations that correspond to aggregated data need to be dealt with either by removal or by acknowledged inclusion. The merger of these 3 time-series data tables into one cross-section time-series data table would require much more time than a simple exploratory descriptive analysis of the values (with averages, medians, and distributions).