# Home assignment - Data Analytics 1 (ITLB352, MIB)

## Task description

Yours task is to read the datafile you are given into a pandas dataframe and follow the instructions.

* Read in the data into a pandas dataframe.
* Display the usual basic information (datatypes, names, non-null values for columns) about the dataframe.
* Display the dataframe.
* Find out what values occur under "Indicator Name".
* Set the name of the columns to the only value you found under „Indicator Name” (the columns and their labels remain the same!). (Apply the operation to your dataframe!)
* Create a new dataframe dfcode which contains the "Country Code" and "Country Name" columns of your original dataframe. You will continue working with your original dataframe, in what follows, though.
* Remove the "Country Code", "Indicator Name", "Indicator Code" columns. (Apply the operation to your dataframe!)
* Make the column "Country Name" the index of your dataframe. (Apply the operation to your dataframe!)
* Get rid of columns which contain no values. (Apply the operation to your dataframe!)
* Display the basic statistics by year, incl. minimal, maximal, mean, median, standard deviation values. (Make sure you display these values for all the years!)
* What can you conclude from the yearly descriptives? Answer in a markdown cell.
* Display the basic statistics by country, incl. minimal, maximal, mean, median, standard deviation values. (Make sure you display these values for all the countries!)
* Display the basic statistics of your previous by country-statistics (i.e., the mean of country means, etc.).
* Display the first thirty rows of your original dataframe after sorting by the values for the **last year** in **descending** order. (Do not apply sorting to your original dataframe, just display!)
* What are some of the conclusions you can draw from the data you displayed (e.g., about the "countries" in the data)?
* Calculate the difference between the (0-filled) values for the **years 2010 and 2000** by country, sort it in a **descending** order, and display the **first 30 rows**. What are some of the conclusions you could draw from what you see?
* Display the **mean** values by year as a percentage of the **maximal** values for that year (e.g., if the mean value for 1965 were 200, and the maximal for 1965 were 400, then for 1965, you should be displaying 50).
* Display the countries (not the subdataframe, just the countries!) where the values for all years are missing.
* Display the subdataframe with the countries whose final year's values are bigger than the **final** year values' mean.
* Load the "capital.json" file, and arrive at a dataframe capitaldf containing the **name, capital, iso2, iso3** information (these three should end up as the columns) for every country in it.
* Find the subdataframe of the dfcode dataframe you created earlier on where the "Country Code" value is **not** in the "iso3" column of capitaldf. Hint: look up and use the .isin pandas Series method. Store this subdataframe in the variable dfnocode. Display dfnocode and explain in a markdown cell what you see and what this could be used for.
* Now display the subdataframe of your main dataframe with the yearly data where the index (the Country Name) is **not** in the "Country Name" column of dfnocode.

Please observe the following:

* You must use a single standalone Jupyter Notebook to solve the task, and submit to Moodle *both*
  + the **.ipynb file**, and
  + a **.pdf file** generated from the notebook.

Note for those working on Google Colab: a link to your notebook will not suffice: you have to download and submit the file itself to Moodle.

* Follow the principle of **literate programming**, and make use of the markdown cells of the notebook.
* Even if you cannot solve a necessary step using Python code, specify **in pseudocode or detailed text** how the relevant step should be solved. This can count towards your score on the „Specification fulfilment” criterion (though NOT on the Coding criterion). Important: simply rephrasing the task description is NOT sufficient: your pseudocode or description must show you aware of how the relevant task should be approached, you are just not able to code it.
* Even if you cannot solve a subtask using automated methods, you can solve it using manual means. This can still count towards your score on the „Specification fulfilment” criterion (though NOT on the Automation criterion).

## Deadline

Please refer to the Moodle page of the module.

## Assessment

The assignment will be assessed based on the following criteria (see the grid on Moodle):

* Specification fulfilment (20%)
* Literate programming and markdown cells (20%)
* Automation of subtasks (20%)
* Reflection (20%)
* Coding (20%)

The resit arrangement for the assignment is the same as above; you may resubmit the same paper, with corrections, that you submitted by the original deadline. The resubmission deadline will be specified on Moodle after the grades for the original submission are published.

Upload all your files to Moodle.

## Academic conduct notice

Where the Academic Conduct Officer has reason to suspect that a piece of work submitted by a student was wholly or in part written by someone other than the student who submitted it, and this has not been disclosed by the student, they may call for the student to defend the work in a viva or a written comprehension test. The burden of proof in such a viva or test will be upon the student to demonstrate to the examination panel’s satisfaction his/her full comprehension of the work s/he has submitted. Failure to appear without satisfactory explanation will result in immediate failure of that assessment, with consequences of academic misconduct and application of sanctions.