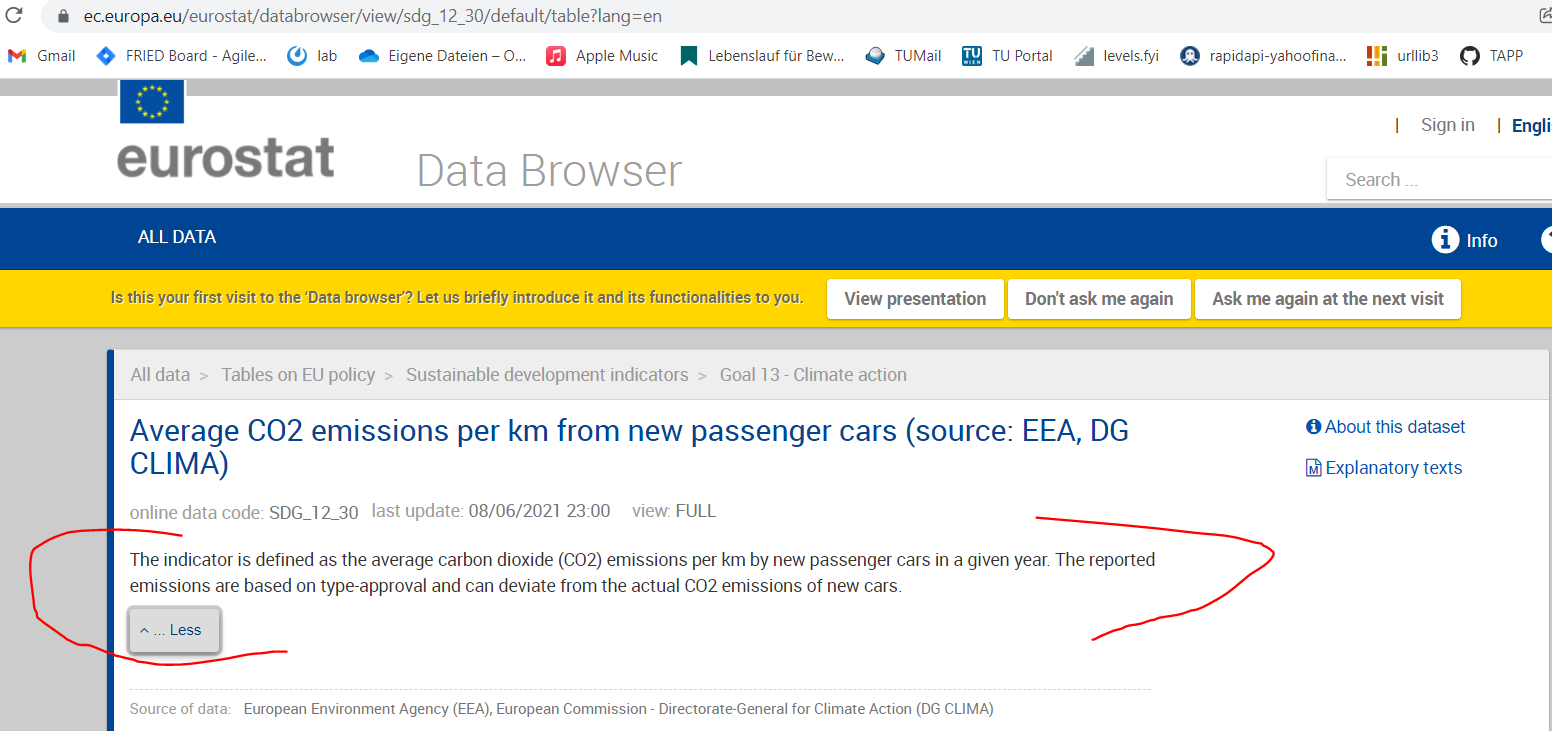
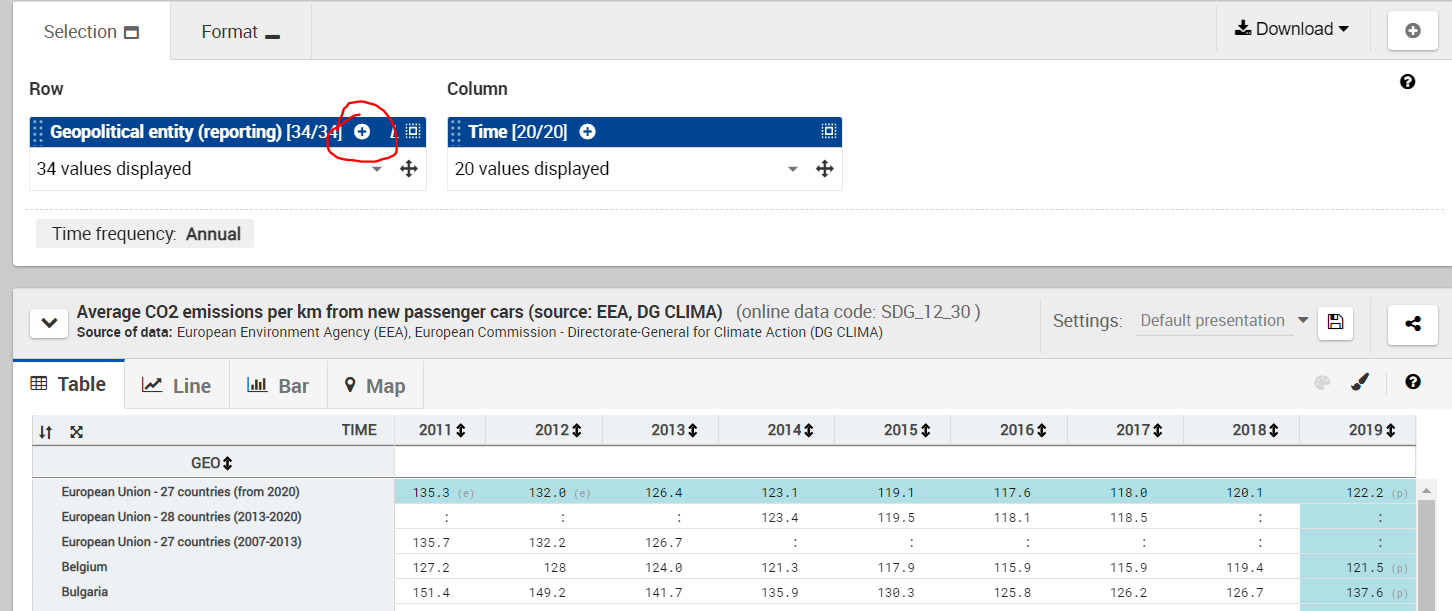
The Eurostat datasets are well structured. For every topic (Railway length, CO2 emission, etc.) it records only a single “value”/”measurement” column, in addition to descriptive columns such as country code, year, or any kind of type.

For example, the original “CO2Emissions” dataset contains for every European country per year a single measurement value. If you want to know what exactly that measurement value contains, it would be best to click on the dataset link (it is in the “Description\_on\_preprocessing.docx” word-file):

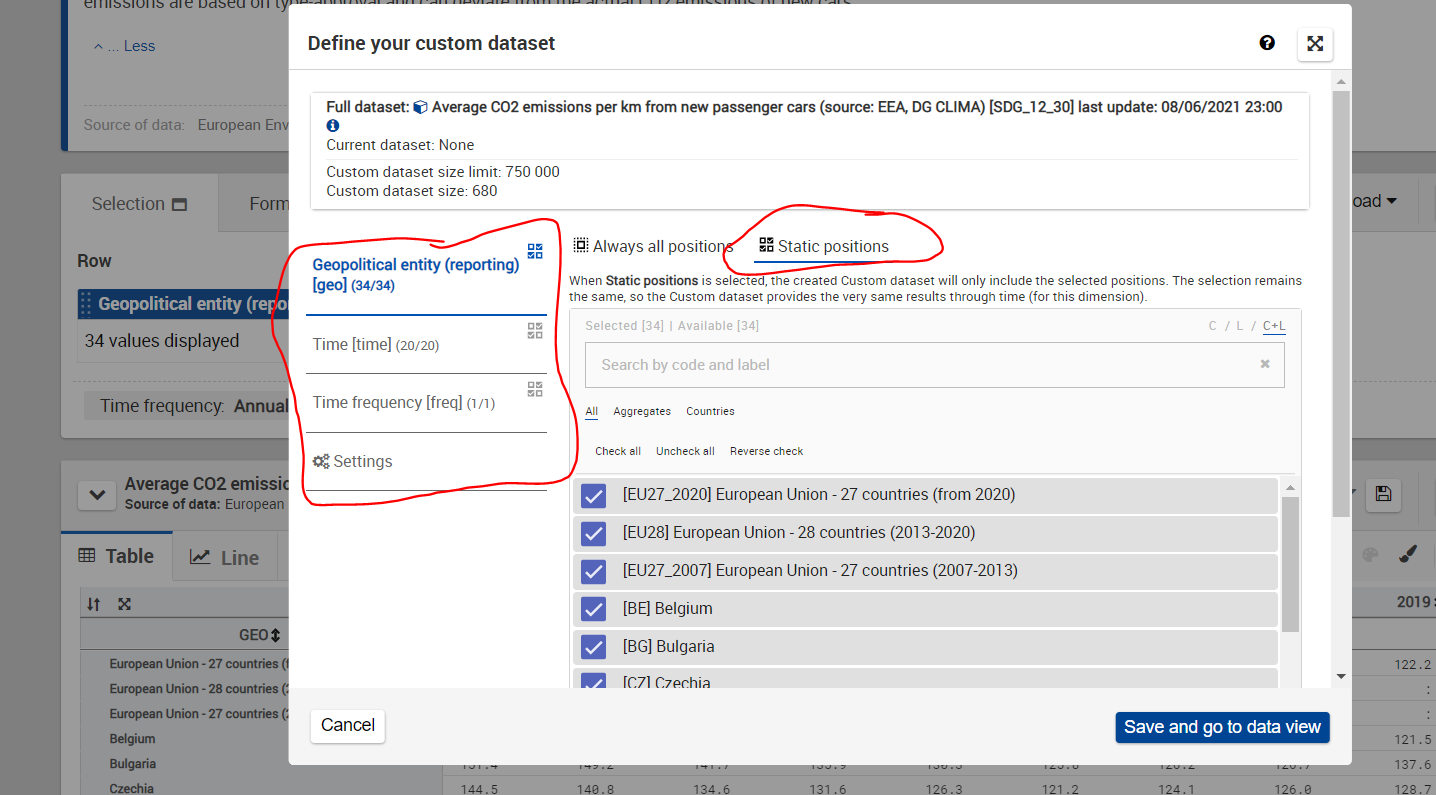


There is a short description for each dataset that should explain what the measurement value contains.

If you want to know more about each column, you can scroll down on the same page and click here:



Which will open:



On the left hand side, you can select each column name. After you do that, you can choose “Static positions”, which will list you explicitly each unique value in that column. Many times, there will be very important information written for each individual row value. Like in the example above, you will now know that EU27\_2020 means that it encodes the European union from 2020, where only 27 countries where in the EU.

The same logic can be applied to all the other datasets.

Additionally, there is a description in the comments in each of the preprocessing scripts that explain why certain steps where taken. For example for the modal\_split.py file:

