In the following, there is a description on the preprocessing that was done.

Generally, there were six datasets analyzed for now about the following topics:

1. Average CO2 emissions per km from new passenger cars (https://ec.europa.eu/eurostat/databrowser/view/sdg\_12\_30/default/table?lang=en),
2. Modal split of passenger transport (https://ec.europa.eu/eurostat/databrowser/view/tran\_hv\_psmod/default/table?lang=en),
3. Volume of passenger transport relative to GDP (https://ec.europa.eu/eurostat/databrowser/view/tran\_hv\_pstra/default/table?lang=en),
4. population by NUTS 2 region (https://ec.europa.eu/eurostat/databrowser/view/tgs00096/default/table?lang=en),
5. total length of railway lines(https://ec.europa.eu/eurostat/databrowser/view/ttr00003/default/table?lang=en),
6. and railway transport of passengers (https://ec.europa.eu/eurostat/databrowser/view/ttr00015/default/table?lang=en)

The preprocessing script for each dataset is kept seperate in the main folder of our DOPP repository (e.g. modal\_split.py, railway\_length.py, etc.) and there is a seperate folder for each dataset where the preprocessed CSV files are stored.

All folders follow the same structure / contain the same preprocessed information.

1. Each file has a prefix relative to the topic. For example, all preprocessed files in "ModalSplit" have their own prefix "MODSPLIT\_", the same is true for "Population"->"POP\_", and the others.
2. The most important file, which actually contains the fully processed and transformed data, always lists the columns seperated by underscore. For "Population", the main file would be "POP\_unit\_location\_year\_POP\_full\_time\_period.csv". The "full\_time\_period" denotes that this file does not contain missing values. In case you want to work with all countries, even if they are missing some of the years, choose "POP\_unit\_location\_year\_POP\_full\_time\_period\_all\_countries.csv".  
     
     
   Other examples of the file equivalents would be:
   1. CO2Emissions: CO2\_location\_year\_CO2\_full\_time\_period.csv and CO2\_location\_year\_CO2\_full\_time\_period\_all\_countries.csv
   2. ModalTransportToGDP: GDP\_unit\_location\_year\_gdp\_full\_time\_period.csv and GDP\_unit\_location\_year\_gdp\_full\_time\_period\_all\_countries.csv
   3. and so on ...

The file itself should be ready for plotting, because it will only have 3-4 columns, depending on how much data is in the original dataset. For example, the "Population" dataset records the population count in a year at January 1st. The final dataset has four columns:

* + - "Unit": in every dataset there is a unit, but its usually only a single value (like in this case its "NR" and stands for "Number"),
    - "Location": is the country,
    - "Year": is the year, and
    - "Count": is the frequency

1. The file from point 2 is basically the information which is extracted from the dataset. Nevertheless, there is some other information that could be useful for our analysis. For example, before knowing how to deal with missing values, we had to first do some missing value analysis to understand how many missing values we have, how they are spread, etc. This information is recorded in other files, like: "GDP\_missing\_data\_per\_country\_per\_unit.csv" records the missing data per country and per unit, and the "GDP\_missing\_data\_per\_location\_per\_year.csv" would give similar information but per country/location and per year. The idea of keeping those file was that maybe you want to do some plotting on the missing values, too, so it is not strictly necessary to use them, only if you are interested in it.
2. The files that contain the phrase "\*\_all\_countries\_full\_time\_period\_\*" are simply the original dataset, which means that the columns are not transformed.

If you have any more questions, please just send a message right away and we will answer it as soon as possible.