**Exercise 3 for Data Oriented Programming Paradigms**

Group 3

Question 17: With which means of transport do people move around in cities (modal split)? How has this changed over time? How has this changed in various countries? Are there specific characteristics of countries/cities that can be shown to correlate with modal split and its evolution?

Links to the datasets:

(Dataset 1) Modal split of passenger transport:

<https://ec.europa.eu/eurostat/databrowser/view/tran_hv_psmod/default/table?lang=en>

(Dataset 3) Population on 1 January by NUTS 2 region:

<https://ec.europa.eu/eurostat/databrowser/view/tgs00096/default/table?lang=en>

(Dataset 4) Volume of passenger transport relative to GDP: <https://ec.europa.eu/eurostat/databrowser/view/tran_hv_pstra/default/table?lang=en>

(Add. Dataset) Total length of railway lines:

<https://ec.europa.eu/eurostat/databrowser/view/ttr00003/default/table?lang=en>

(Add. Dataset) Railtransport of passengers:

<https://ec.europa.eu/eurostat/databrowser/view/ttr00015/default/table?lang=en>

Notes of Fardokht and Mehdin, 04.01.2022

You will find several folders that each contain the data for a certain topic, and the script that was used to create it in the main folder. For example, the modal split dataset (Dataset 1) has a dedicated folder “ModalSplit” that contains the preprocessed data, and a python script “modal\_split.py”, which was used to create it.

**Question regarding missing values:**

In each of the dataset obtained from EUROSTAT, some countries are missing information because e.g. they did not have certain transportation types up to a certain date or did not record it up to a certain time point. For example, some countries are missing transportation data in the Modal Split dataset from e.g. 1990-1994, or 1990-1999, or in extreme cases like Malta or Cyprus up until 2009.

We thought that maybe we could consider the following two scenarios when doing our analysis:

1. For studying long-term effects of modal split, we can take the list of countries that have the complete data available.
2. For questions/investigations that anyway are concerned with shorter time frames, we could consider solely the last ten years (2009-2019), as all countries will have data to that time point (depends on the dataset).

Imputing data is very difficult in this regard, because it was either not recorded due to reasons that are specific for a country (maybe the country’s infrastructure wasn’t good enough to record it yet), or because a country maybe does not offer a certain type of transportation (like a train infrastructure).

For visualization purposes, we could include them as they will simply not appear in the plots, but for aggregations (means, sums, correlations, etc.) we would have to be careful to separately consider only the countries with full data.

What do you think?