Toolkit 3: DATA CLEANING

What "cleaning the data" means:

In most cases, previous datasets have more information than the researcher needs. Most data are unstructured or unorganised. Hence, it is necessary to reorganise the data by eliminating missing values, irrelevant information or transforming variables. This process is called "data cleaning".

> Example: dataset might provide absolute numbers whereas the researcher might be interested in analysing percentages. In this case, it would be necessary to transform whole number into proportions (%).

Sometimes, it is necessary to create new variables.

Why data cleaning is relevant:

Datasets might have been designed for purposes other than those aimed by researchers. Hence, every researcher or data analyst needs to adapt them to answer their specific research questions.

Further tips for building/cleaning a dataset:

- Make it simple. Using short names with no special characters or spaces for variables usually facilitates the analysis. For example, R programming language requires the usage of brackets to deal with variables with composed names separated by spaces. This makes coding more complicated.
- Transform your variables according to your research questions.
- Choose the adequate software. Some tools are not able to deal with large datasets.
- Maintain only the variable that might be useful for the study.
- Always write a codebook explaining what the variables mean and how they were measures in addition to clarifying the sources of the data.

References and further readings:

Tableau. Guide for Data Cleaning. Available at

Data Cleaning: Definition, Benefits, And How-To | Tableau