Course
CS-E4002
Course materials
Your points
Code Vault

This course has already ended.

« Structure and Cleaning Course materials Preparing Data for Analysis »

CS-E4002 / Structure and Cleaning / Erroneous Data

Erroneous Data

Once a DataFrame is created from the file containing the data we want to use, we wan get quite a comprehensive view of the data by using the method head(). Peeking into the file, we often notice that the data **cannot be used in this raw state**. This can be cause by problems with encoding, duplicates, and incorrect types.

1. Problems with encoding:

- There are problems with **different conventions** for writing **dates**: depending on the country, the day can be first, or the month. The year can be written with 2 or 4 digits (sometimes ranges of dates can be given instead of a starting date, brackets can be used around years, ...). If we want to use the data and compare different dates from different countries, we will need to **choose one of the conventions** and convert everything sto that convention. Alternatively, we can remove the data that doesn't fit our conventions.
- There can also be problems with **decimal numbers**: in some countries, a comma is used instead of a point to separate the decimal part from the integer part of a number. Also, in other countries, a comma can be used to separate thousands from hundreds... This needs to be **standardized** as well if we want to be able to work with all the numbers without making any mistakes due to the difference in encoding.
- There can be issues with **language encodings**. Depending on the language, different accents and letters are used, which can give rise to encoding issues but mainly interfere with a good understanding of the data. Before using some data, the language it uses must be unified and all potential difficulties due to encoding of accents and letters handled.

2. Problems with duplicates:

- There could be **duplicates** that should be found and removed in order not to keep redundant data.

3. Working on elements that have the right type:

- To work on the data, we will want to compute means, minimums, maximums, ... But when the dataframe is created, the raw data is often just represented by strings. It is important to convert the data into regular python types (int, double, ...) to be able to do computations on it.

Once the data is **clean**, it has a uniform encoding for all different values and words used, and all the values in it can be compared, there is still one issue to take care of: the amount of data available. Indeed, very often, databases contain more data than we need or want to analyse, even once the duplicates have been removed. Keeping all the data can make it difficult to work on a project and make everything slower, especially when handling a huge amount of unused data in our programs. That is why it is important to **filter the data** before using it: keeping only the data we want to analyse and work with.

« Structure and Cleaning Course materials Preparing Data for Analysis »