

Data Cleaning and Preparation

A.Ederoclite



Ayuda ICTS-MRR-2021-03-CEFCA, financiada por:





Get to know your data

If your data is associated to a paper: check the paper

E.g. if it is a catalogue: how are sources identified/selected? If it is photometry, how was it performed?

If your data come as a fits file: check the header

E.g. which instrument? What exposure time? What kind of processing has been applied?



Let us explore an image with Python

Notebook `INPE_image.ipynb`



What about a catalog?

There are three things which can happen in a catalog:

- Non valid value in a column
- Missing value in a column
- Outlier or unexpected value



Not valid values

It is not uncommon to have upper limits in a catalogue.

In some cases, you have a series of numerical values and then... “<19”.

How do you deal with it?



“Null” is not “Null”

In some cases, some authors prefer to leave blank a non detection.

There are different approaches:

- Ignore the source
- Give the source a value which is the mean (or the mode or the median) of the sample
- Infer the value that that field is supposed to have (“imputation”)

Outliers and where to find them

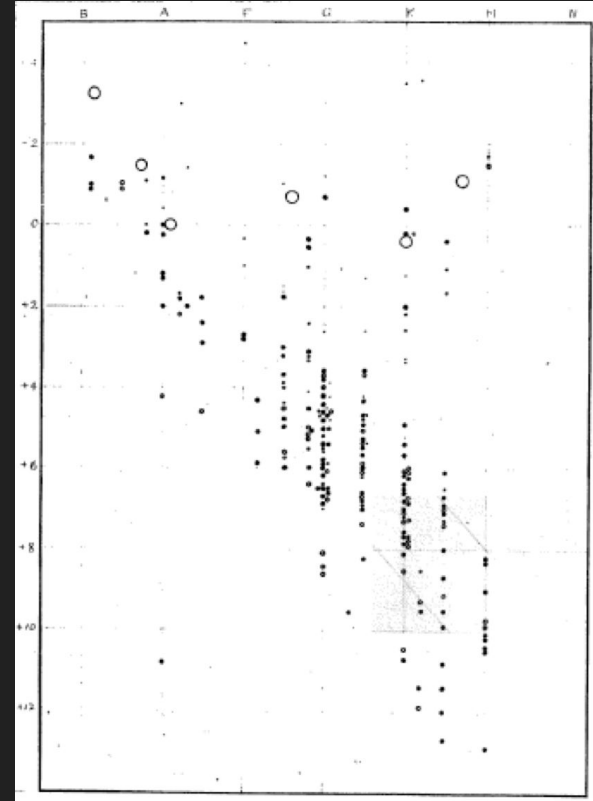
Not all outliers are bad data.

The first HR diagram shows a clear outlier on the lower left.

Naively, one would be tempted to mark it as an error.

It is the first known white dwarf.

Think and have good reasons to treat an outlier as a mistake.





Python vs R (vs IDL, Julia, C...)

Python and R “dominate” the field of data science.

None of them is an “astronomy oriented” language.

R is a statistics oriented language while Python is a generic language.

We focus on Python... because I use Python.

IDL is expensive.

C is inherently evil (although powerful).

Julia is rising.



Numpy vs Pandas

These are the two most used numeric Python modules for Python.

Numpy is better at some things and Pandas is better at others.

You want to use **both**.



Matplotlib vs Seaborn

Data visualization is a science (and an art) per se.

Visualising the data in the wrong way may be a danger.

In Python, the two main plotting modules are matplotlib and seaborn.

Seaborn uses matplotlib but it makes the experience smoother (in some cases).

I am a matplotlib user.



Let us explore a catalogue with Python

Notebook INPE_Demo.ipynb



Modifying data should be kept at minimum.

The programming language is a tool:

you need a compromise between comfort and efficiency.