Data Wrangling - Session 2

Data Preparation

Content

1. Data preparation
2. Encoding
3. Feature Scaling
4. Hands-On:

Data preparation

Data preparation deals with different issues in data such as:

- Missing values (nulls, NaN)
- Duplicates
- Different data types (e.g., object, float, int, str,)
- Heterogeneity (e.g. ‘Guadalajara’, ‘GUADALAJARA’)

Encoding

Encoding consists in converting data types to convenient numeric variables.
The data can come from:

- Categorical variables
- Text
- Ordinal variables
- Boolean variables

Feature scaling

Finally, keeping in mind that we want to train a ML model, different algorithms have different performance depending on the input scale. Scaling could be:

- Standardization
- Normalization
- Robust Scaling (median and quartiles)
- MinMax Scaling

Hands-On