Grammar of Data Wrangling for Anoamly Detection in Water Quality Data

Priyanga Talagala

03/12/2021

,

Objective : A more **general framework for approaching data cleaning** which is rooted in a desire to represent data both accurately and fully.

- This framework informs decisions about an ideal order in which data cleaning should be conducted
- Types of data and steps of cleaning
  - Compiling data (Downloading/collecting, Merging, Appending, Reshaping)
  - Define an anomaly
  - Labeling and naming variables
    - Examining data
    - Identifying obvious anomalies (out of range values)
    - Manual labelling (with water quality experts)
    - Identifying special situations (eg: wiper anomalies)
  - Altering variables (Recoding variables, Transforming variables)
  - New variables (Scale construction, Substantive variable combination)
  - Examining missing values (regular/ irragular)
  - Re-configuring data for specific purpose
  - Re-examining data
  - Documentation and presentation

- "You need to clean the data" what exactly does that mean?
- What steps are involved and in what order?
- How do we decide what needs to be done?

Data cleaning involves all the steps that occur between data collection and analysis (e.g., merging, appending, labeling, data analytics, cross-validation, constructing/re-constructing variables for analysis, identifying missing data).