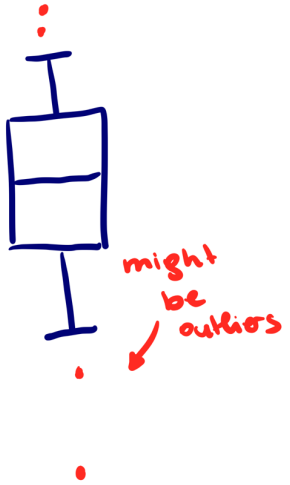
Data Handling

- Before an AI model can be trained, the data must be prepared
- This task usually takes the most time and is the most important part, else the result will be incorrect/biased (“garbage in, garbage out”)
- Data must be analysed first to detect any errors: summary statistics, distribution plot, box plot, etc.
- Additionally, one might ask an expert to determine erroneous data entries (implausible values)
- Errors include:
  - missing data and mixed NULL-values
  - outliers
  - inconsistent data entries, especially for categorical attributes
  - invalid data: character values in numerical variable and v.v.
  - inconsistent delimiters: semi-colon instead of comma
  - different decimal points: comma instead of dot
  - columns not correctly split
  - additional head- and footnote
  - duplicated rows
  - wrong datatype
  - switched month and date



Missing data, Class imbalance, Outliers

- Outliers can be detected by visually analysing a distribution plot or box plot, calculating the quantiles and consulting with an expert
- Missing, invalid data entries and outliers might be:
  - Investigated: analyse data source to fill in or correct the values
  - Removed: complete data entry or column
  - Imputed: median/mean impute, kNN Imputer
- Class Imbalance: If one target class is underrepresented in data, it can lead to a biased model - possible solutions:
  - Increase number of data entries for minority class by:
    - collecting more data
    - using the bootstrapping: draw random sample with replacement from each class
    - SMOTE method: artificially create same sized samples per class
  - Decrease number of data entries for majority class by using only subsample