Data Cleaning Process

1. **Import data**
   - Check for importing issues such as broken lines when importing .csv files
   - Make sure you have unique IDs
   - lowercase all the column names and replace spaces with _
2. **Inspection:**
   Detect unexpected, incorrect, and inconsistent data
   - Profiling
     - create summary statistics
     - How many values are missing?
     - How many unique values in a column?
     - Check the distribution of features and target variable
   - Visualization (when necessary)
3. **Cleaning:**
   fix or remove the anomalies discovered
   - Drop Irrelevant features (e.g., phone number when analyzing health data)
   - Remove Duplicates (e.g., the same article was scrapped twice)
   - Type conversion
     - Numbers as numerical data types
     - date as date object
   - Syntax errors
     - Remove white spaces (e.g., `" hello world "` -> `"hello world"`
     - Pad strings (e.g., `313` -> `000313`)
   - Transformation using a function or mapping
     - map different gender values to either `"male"` or `"female"` **p.200**
     - replace values (e.g., negative values with `NaN`) p.200
     - Renaming column / index p.201
     - Discretization and Binning p.203
   - Handle missing values
     - Filter out / Drop p.193
     - Fill in / Impute p.195
     - Flag
   - Standardize
     - Strings must be either in lower or upper case
     - Numerical values have the same measurment unit (kg or bound)
     - Dates: USA version vs. European version
   - Scaling
     - put the data within a specific scale such as `0-1` or `0-100` (e.g., convert exam scores from `0-5` scale to `0-100`)
   - Normalization
     - rescale the values into a range of `0-1` (normy distributed)
   - Detect and Filter Outliers
   - Computing Dummy Variables p.208
   - String Manipulation p.211
4. **Verifying:**
   After cleaning, the results are inspected to verify correctness.
5. **Reporting:**
   A report about the changes made and the quality of the currently sotred data is recorded