In some process models Data Cleansing is a separate task, it is closely tied to Feature Creation but also draws findings from the Initial Data Exploration task. The actual data transformations are implemented in the Feature Creation asset deliverable; therefore, Data Cleansing is part of the Feature Creation task in this process model.

While tuning machine learning models, this deliverable asset is touched on a regular basis anyway because features need to be transformed to increase model performance. In such iterations, often issues with data are detected and therefore need to be corrected/addressed here as well.

The following none exhaustive list gives you some guidelines:

- Data types
  Are data types of columns matching their content? E.g. is age stored as integer and not as string?

- Ranges
  Does the value distribution of values in a column make sense? Use stats (e.g. min, max, mean, standard deviation) and visualizations (e.g. box-plot, histogram) for help

- Emptiness
  Are all values non-null where mandatory? E.g. client IDs

- Uniqueness
  Are duplicates present where undesired? E.g. client IDs

- Set memberships
  Are only allowed values chosen for categorical or ordinal fields? E.g. Female, Male, Unknown

- Foreign key set memberships
  Are only allowed values chosen a field? E.g. ZIP code

- Regular expressions
  Some files need to stick to a pattern expressed by a regular expression. E.g. a lower-case character followed by 6 digits

- Cross-field validation
  Some fields can impact validity of other fields. E.g. a male person can't be pregnant

Please transform your data set accordingly and add all code to the Feature Creation asset deliverable. Please comply with the naming convention documented in the process model.