### Processing and Cleaning for a Single Reuse

Within a project, once external data have been identified and acquired there may need to be extensive data processing in order to make the data suitable for the project’s purpose. In some cases this might mean cleaning operations like making sure the values for variables fall into the defined value domains. In other cases a variable might not be in the correct form for input into an analytic procedure. Similarly, the structure of a dataset might need to be rearranged, such as a transposition from a tall structure to a wide structure.

A research group might benefit by developing a library of transformation operations, either as a set of descriptions or as a set of machine actionable scripts. For our example study we looked through a script of about 2000 lines of Python code used to clean DHS data for a machine learning study. We attempted to classify these operations as follows. Some of these operations are generic, being applicable to many uses. Others are specific to the needs of the machine learning application of our use case.

### Sample operations from the DHS use case

#### Generic

* Variables with a constant value are dropped
* Variables that completely duplicate another variable are dropped
* Cases with specific (substantive or sentinel) values are dropped
* Codes with misspellings are mapped to a single value
* Case of codes is regularized
* Variables are renamed with semantic names (Identifying and representing top and bottom coding (Number of injections in the last 12 months = »90+ »)
* Identifying and dealing with outliers

#### Specific purpose driven

* Some nominal variables are dummy coded
* Variables with more than some percentage missing are dropped
* Top and bottom coding is transformed using a fixed value or reference value (e.g. person’s age)
* Nuanced sentinel values are mapped to a common value (e.g. Stata missing values to R, Python NaN or NA)
* Some ordinal variables are transformed into interval variables
* Collapse categories for some variables (e.g. « catholic, roman Catholic)  « Catholic »
	+ This is best done using some international classification scheme
* Value mappings may follow complex rules, for example if fewer than five variables in a set are originally missing (NA) then recode to .1, but if all variables in the set are NA then do not recode)
* Dropping variables (survey paradata may not be relevant)
* Drop variables with low variance ( this is a choice depending on analysis requirements)
* Imputing missing values

We noted that a library of these operations could be built over time by a research group and used to train staff or even to automate at least some of the data processing.

### Mappings between value domains

One of the issues that this exercise revealed was that the set of country codes used by DHS did not match the ISO 3166-1 country codes. In particular as seen in the table below Burundi and Namibia have different codes in the two sets. Good practice would be for data shared from our example to recode the DHS codes to ISO codes. It should be noted that these codes would change over time as countries change. The set of codes to be shared should be time stamped.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **DHS** |  | **ISO 3166-1** |
| Angola | AO | --> | AO |
| Burundi | BU | --> | BI |
| Ethiopia | ET | --> | ET |
| Lesotho | LS | --> | LS |
| Malawi | MW | --> | MW |
| Mozambique | MZ | --> | MZ |
| Namibia | NM | --> | NA |
| Rwanda | RW | --> | RW |
| Zambia | ZM | --> | ZM |
| Zimbabwe | ZW | --> | ZW |