INF1340: Midterm Project

Wenxuan Li

This document is to explain the detailed steps taken to clean the UN data set to make it tidy. The whole process strictly follows 5 tidy data rules, which are:

1. Column names need to be informative and should not be values
2. Each column needs to consist of one and only one variable
3. Variables need to be in cells, not in rows or columns
4. Each table column needs to have a singular data type
5. A single observational units must be in 1 table

To start the data cleaning, the first thing to do is to understand the data file. The UN data set file contains 9 different sheets in total. The first and last sheets are descriptive sheets. Other 7 sheets contain data that should be cleaned and combined. Among 7 sheets, the first 6 sheets contain the immigrants' data from various perspectives and the other sheet contain the country/area information. We will integrate all information into one table via following steps.

- Read different sheets separately into Python.
  - One sheet is corresponding to one data frame. This will help future data cleaning to avoid calling the same data frame again and again at different cleaning stages.
  - Only important data are read into Python. Rows for titles of sheets are skipped, but variable names are partially kept. Only columns with data are read.
- Pre-process all data frames.
  - If variable names are not correctly retrieved from the table, we manually add them.
  - If variable names are not in string type, we manually adjust them. (Principal 1)
- Clean the table containing the country/area information.
  - We do not want to keep all variables. For example, we throw `sort order` out.
  - We rename some variables to clarify their meanings. For example, we rename the code of region into region code to distinguish it with country/area code.

After the pre-processing part is done, we start to clean the table one after another. Detailed steps are list as follows.

- Table 1
  - Separate the big table into three sub-tables according to the gender type (Principal 3). For each separated table, we only take the `Major area, region, ..` column plus the columns named after years.
  - Melt each sub-table into (country, year, international migrant stock) (Principal 2).
  - Add the gender information to each sub-table
  - Stack three sub-tables together into a big table
  - Merge into the country table by left join (Principal 3). This step removes the rows like "world", "developed region", "Africa", etc. Such information is summarized in the country table.
- Table 2-3
  - The processes of cleaning these two tables are like Table 1 but are dealing with different information on migrants.
  - The only difference is the last step – the table is merged into the tidy table obtained from the last section, and is no longer via left join, but outer join.
- Table 4

    - Table 4 is slightly different from Table 1-3 because it only contains the proportion of female immigrants.
    - Fortunately, we could compute the values for all gender (which should be 100) and the male (which should be 100-female proportion).
    - In the computation process, we use N/A to represent any cells that do not have an appropriate value
    - Other steps are quite similar.
- Table 5
    - Table 5 is again different from Table 1-3 because it talks about the information within a period rather than a specific year.
    - We are not quite sure how to make it align with the previous tables, thus, we choose to make the period information to be the as the information for the starting year
    - Other steps are quite similar.
- Table 6
    - Table 6 is also slightly different from previous tables because it does not contain information from different genders, but different information types (i.e., refugee stock, refugee proportion, and refugee rate).
    - For the gender, we all set it as "all".
    - For other information, we process it as we previously describe.
    - We merge the table for each sub-table once we finish clean it, instead of stacking sub-tables together and merge only once.

Above is the detailed step we took to make the UN data set tidy.

The observed untidiness that could not be incorporated appropriately include:

- as discussed above, we are not sure how to integrate the data points that under "year" with the data points that under "period";
- the notes information and the data type information are not incorporated in the cleaned data set.