INF 1340 Writeup
Jingyu Jia

The main idea of this data cleaning project is to meet all the tidy data principles and try to merge all tables into one by setting them into the same format individually and combining them in the end.

When I go through table ANNEX, I find out each major area, region, country, or area has Country Code as a unique identifier. Then, I decide to use Country Code to represent them.

Function: update_tables

After looking at all the tables, I find a common mistake that violates principle #1 and principle #3 for table 1, table 2, table 3, table 5. The header shouldn't include years and sexes and those two variables should be stored in cells. So, I try to add another two columns to store them. Instead of cleaning those tables using the same procedures for 4 times, I decide to define a function called update_tables to deal with it. Those four tables share the same frame about the sexes. I separate them into three categories as "both sexes", "male" and "female". I specifically set header in the format as ['Country Code', 1990, 1995, 2000, 2005, 2010, 2015]. For "years" problem, I melt them into cells. After all, I concat all the data from all genders. Also, I find out there are some '..'in the table, and I replace them into NaN.

Function: update_one_table

For table 4 and table 6, the same violation as other tables exist but don't have sex information as other tables have. Therefore, I define another function called update_one_table to deal with it. I set header in the format as ['Country Code', 1990, 1995, 2000, 2005, 2010, 2015]. It's the same format as I set for all the other tables. By doing that, I can merge all the tables in one at the end. As above, I melt "years" into cell and replace '..' into NaN.

For table 1:
I called **update_tables** and set header as 'International migrant stock at mid-year'.

For table 2:
I called **update_tables** and set header as ' Total population at mid-year'.

For table 3:
I called **update_tables** and set header as ' International migrant stock as a percentage of the total population'.

For table 4:
I called **update_one_table** and set header as 'Migrants as a percentage of the international migrant stock', 'female'.

For table 5:

Since annual rates are percentages for a period, I use the last year to represent the five-year period. For keeping along with the same data frame as other tables have, 1990 has be on the header and I put NaN on it because there are no annual rate change prior to 1990. Then, I called **update_tables** to separate sexes and set header as 'Annual rate of change of the migrant stock for last five years'.

For table 6:

There are three sections I have to deal with.

For the first sub-table(table6_1), I called **update_one_table** and set header as 'Estimated refugee stock at mid-year', 'both sexes'.

For the second sub-table(table6_2), I rename the header to be recognized by **update_one_table** function and called the function as well as set header as ' 'Refugees as a percentage of the international migrant stock', 'both sexes''.

For the third sub-table(table6_3), I name the header by using the last year to represent the five-year period and set 1990 to be NaN. Then I called **update_one_table** 'Annual rate of change of the refugee stock for last five years', 'both sexes' .

Merge Tables

I merge all tables into one table.