INF1340: Midterm Project

Wenxuan Li

Introduction

To investigate the population changes of a country or the whole world, the international migrant stock is an unavoidable topic. In this midterm project, we will focus on a given data file, which is denoted as UN data file, on the migrant stock to make it tidy and easy for further analysis.

The UN data set file contains 9 different sheets in total. The first and last sheets are descriptive sheets. The first sheet is denoted as content sheet, which describes the title of the data file, the detailed information of the file provider/creator, and the specific sheet titles for the following eight sheets. The last sheet is the notes made for the previous seven sheets (i.e., some vague points are marked by numbers and explained in this sheet). Besides the descriptive sheets, the middle seven sheets contain data that we are going to clean, process and combine. In the following context, we will introduce the principles we utilized to make the data file tidy, the results after the processing, the discussion, which is mainly about what the data set looks like, and a brief conclusion.

Methods and Results

To begin with, there are five principles that we need to be careful about when cleaning and processing the data file to make it tidy. They 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 one cell.

There are two initial steps to take before we jump to the specific sheets. The first thing is to load the file into Python in an organized manner. We choose to 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. Also, 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. The second step is to pre-process all data frames. If variable names are not correctly retrieved from the table, we manually add them.

Besides, if variable names are not in string type, we manually adjust them. This is required by the first principle that the column names should be informative.

After the pre-processing steps, for each sheet (or sub-table of the data file), we observe the following several common problems. The first problem is that in most sub-tables (except the last sub-table), there is one column that mixes the continent, the region, and the country together. The existence of such a column violates principle 2) because more than one variable is contained within one column. Therefore, this column is separated into multiple columns – country or area name, corresponding major area, region, whether it is a developed region, whether it is the least developed country, and whether it is located in the sub-Saharan Africa. To realize this step, we only kept the country or area information for Table 1 to Table 6, removing other rows, then, we join the processed table with the ANNEX table to add other information (e.g., corresponding major area, region, etc.).

The second problem is the gender problem – in Tables 1, 2, 3, 5, each table is divided into three major parts, by gender, as suggested by the title. This violates the principle 3) – the gender variable information is contained in rows instead of columns. To address this violation, a new column for the gender is created and each table is melted into the new format. In Table 4, since we only have part – the female data, to make the processed table consistent, we compute the male data (100-female data) and all sexes data (100) manually. If the female data is missing, we also set the corresponding male data as N/A.

The third problem is similar to the gender issue – within each gender category, there are multiple year columns. This indicates that the year information is contained in the rows instead of columns, which violates Principle 3). The solution is again to set up a new column for the year and melt the table into the new format.

The last issue is that Table 6 does not contain the extra gender information. Instead, it contains information from three different aspects. For this table, we set the gender as “all”.

After the processing of each table, they are combined using the country or area information as the key. Note that, the information of ANNEX table is combined into the final table via the processing mentioned in the first problem.

Above are the problems that I observed from the data set, which I can solve. There are several other observations that I cannot solve, or I am not sure whether my actions are correct or not. The first problem that I am not sure about is that in some tables, their year information is the rate within a year range. For this case, I align the year information to the smallest year. But I think maybe creating a new column to represent such information is more appropriate. The second problem is that some information is lost when we make the table tidy. For example, the last table, which is the descriptive one, is not incorporated into the final table. In the meantime, we do not incorporate the region data on the migrant stock, population, etc. into the final table. But fortunately, such information can be computed via available data.

Discussion

The final data set contains 4770 samples on 19 variables. Each row represents a sample, and each column represents a variable. The column names are informative – we can know what the column is about directly from its name. In other words, this is a tidy data set.

Besides the tidiness of the data set, there is one interesting observation that the processed data set is not as clear as the original data file – it has too many variables and too many samples within each table. Given a variable, it is harder to compare across different levels. However, such comparisons are quite easy in the original data file.

Conclusion

In the midterm project, we combine multiple tables together, separate one column that contains more than one variable into multiple columns, melt unnecessary columns, and name new columns with descriptive variable names. In short, we use the five tidy data set principals to clean a given data file. However, from the tidy data set, it is hard for us human beings to directly obtain some results – the tidy data set is for the machine only, not for human beings.