INF1340
Xinxuan Lin

Midterm Assignment Write-up

There are six tables in the UN Migrant Stock Total file. The procedure of cleaning each table would be discussed in the following subsections.

All tables have multi index headers, therefore, when import these datasets, I combine the first two rows as a new header, then rename each column based on the associated characteristics including the sex categories and topic of interest. For example, table 1 contains the international migrant stock number for three sex categories: both sexes, male and female between 1990 and 2015. Each category has six columns which are the stock data collected every five-year from 1990 to 2015. The header names are changed by adding the associated sex category to each column of years. Therefore, the columns of 1990 to 2015 under the international migrant stock for both sexes became “IMS Both 1990”, “IMS Both 1995, and etc.

Table 1

After renaming the headers of the international migrant stock columns (1990 to 2015 for each category of sex), the first problem is that the column headers are values instead of variable names for these 18 columns. This violates the Tidy data principle 1 that every variable must have its own column. So, I pivot the table into a longer format by melting the headers as one column named *IMS* and saving all the associated numeric values in another variable called *International migrant stock at mid-year*. Thus, eighteen-column headers and their values are melted into two columns.

However, the second problem is that the *IMS* column contain both the sex and year categories, which again violates principle 1 as each column needs to contain a unique variable. Therefore, the sex and year values in the *IMS* column are separated into two new columns called *Sex* and *Year.* For convivence in reading, I move *Year* and *Sex* columns to the front of *International migrant stock at mid-year* and save it as a new data frame called *un1_order*.

Thirdly, all the countries, regions, continents, and other larger ranges of destinations are collected as one variable named *Major area, region, country or area of destination*. Despite the fact that they all indicate the destination, they are different types of location, which also violates principle 1 as there are more than one variable in one column. Therefore, I separate the destination column into 7 different location columns: *World, Nation, Separation in developing regions, Separation in Africa, Major area, Region, Country or area.* In order to do this, I filter the destination rows based on the characteristics of the locations, and then save them as a new data frame. For example, I filter out all the continent rows and save them as a dataset called *Main area*. After the filtering, I concatenate all seven datasets together to have all information in one table. Since the variables *Notes, Type of data (a)* do not provide important information about the observational unit, they are removed from the dataset. *Country code* is dropped from the data table because both this and *Sort order* provide the unique ID for each destination.

Fourthly, the final data table contains the international migrant stock of each destination in each sex category for every 5-year. Thus, the dataset has three observational units which are the destination, year and sex. This violates principle 3 that a table should have only one type of observational unit. Therefore, the table should be separated into three based on the types of the observational unit, one contains only the destination and *Sort order* which is the unique identification variable for the destination, one contains *Sort order, Year,* and the associated international migrant stock, and the other one that has *Sort order, Sex*, and the associated international migrant stock. But I fail in separating the year and sex into two tables, so I only create two new tables, one for the destination observational unit, and one that contains *Sort order*, *Year*, *Sex* and *International migrant stock.*

Table 2, Table 3 and Table 5

All Table 2, 3 and 5 of the UN data files have the same data structure as Table 1, therefore, the violations of the Tidy data principles and the cleaning process are similar as Table 1. Instead of *international migrant stock*, Table 2, Table 3 and Table 5 have *Total population at mid-year (thousands)*, *International migrant stock as a percentage of the total population*, and *Annual rate of change of the migrant stock*, respectively. Each of Table 2, 3 and 5 are supposed to separate into 3 tables based on the types of the observational unit, but due to the difficulty in separation, the dataset ends with two separations instead of three, which is the same as Table 1 case.

Table 4

Table 4 also has similar violations of the Tidy data principles and the cleaning process as table 1. The only difference is that instead of three sex categories, this dataset only contains data for female migrants, which means it has one measurement per destination every 5-year. In other words, there are two observational units: the destination and the year of data. Therefore, according to the Tidy data principle 3, the data table is separated into two tables: one contains *Sort order* and the destination, and the other one contains *Sort order*, *Year,* and *percentage of the international migrant stock* for female migrants.

Table 6

Table 6 contains no sex category and three topics of interest: the estimated refugee stock at mid-year for both sexes, the percentage of the international migrant stock for refugees, and the annual rate of change of the refugee stock. The data of the annual rate of change of the refugee stock were collected in 4 5-year intervals between 1990 and 2015. For convenience in formatting, the values of the year interval are renamed into the upper bound of the interval. For example, “1990-1995” is changed into “1995”.

Similar to Table 1, I rename the multi-index column headers into a single level based on the corresponding topic of interest. Instead of melting the column headers directly as I did in the Table 1 case, I separate the table into three based on the topics of interest. Then I melt the column headers into a column within each table due to its violation of Tidy data principle 1. Each table has the *Sort order,* the destination, *Notes*, *Country code*, *Type of data (a*), *Year*, and the corresponding topic of interest. After that, I combine all three tables based on the destination to have a complete table that includes all the topics. Due to the same reason as table 1 case, the variables *Notes*, *Type of data(a)*, and *Country code* are removed from the dataset.

The missing data in the dataset were recorded in different forms: “NaN”, “..”, and 0. To keep the format

consistent, I replace the missing values with 0.

The destination column in this dataset also violates the Tidy data principle 1, which is the same as the Table 1 case. Thus, this column is also separated into different ones based on the geographic characteristics. The procedure of doing this is the same as the Table 1 case.

Since for each destination, we measure the three topics of interest every five years, the observational units are the destination and the year. Therefore, due to the violation of principle 3 that each table should have only one type of observational unit, the dataset is separated into two. One is the location table that contains the sorting order and the destination; the other contains *Sort order*, *Year*, and the three topics of interest.

For Table 1 to 6, one of the observational units is the destination. As Tidy data principle 3 states that one type of observational unit should be in one table, the destination table for each sheet should be combined into one. I concatenate all six destination tables and remove the duplicate rows. The final destination table for all the UN data sheets contains 265 rows.