Midterm Write Up
Name: Zhen Wu
Date: 2022.11.17
Words counts: 745

For the midterm assignment, we are using the UN_MigrantStockTotal data, last modified in 2015. The datasheet is an excel document that contains nine sheets, including Contents, Table 1, Table 2, Table 3, Table 4, Table 5, Table 6, Annex, and notes. The contents sheet explained the datasheet sections and titles, which gave us a general understanding of the data sheet. The annex and notes sheet provides apparent data for us to reference. However, the Tables show very complex data, which are hard to use for further analysis. For this assignment, I will be using tidy data principles to clean the six tables. After the data cleaning process, we should have more precise access to our needed information.

Data Cleaning for Tables 1 & 2

Table 1 and Table 2 have similar data structures. Both tables contain data from 6 years' population in both sexes, male and female. One of the main problems of these two datasets is that they have column headers that are values, not variable names. Another problem is that we have multiple variables stored in one column. For example, we have years and sex stored in one column. Instead, we are supposed to separate the category of years and sex so each row can have its unique information for further access. On top of these two problems, we also see duplicated data pointing to the same information of country names.

To solve these problems, I started to use the drop function to drop the sort and major area name columns since it is duplicated with the index number and country code, which is provided on the annex sheet and explains country code and country names. I knew that I would need to separate the sex and year information, so I renamed the first row for more accessible coding later on. Then I used the melt function to ensure every column is a variable instead of a value. After that, we see the Group column still contains more than one piece of information. Therefore, I used the assign function and lambda to separate gender and year. Finally, I renamed the “short-cut” gender information I created in the first step back to the original value, which is both sex, male and female, so that people can have a better understanding of the data.

After this cleaning, we now have a better table containing only one variable in both rows and columns. We can now put the country code to call for their immigrant information from any year and any gender filter. We could also split the dataset into two since the country code, notes, and type of data are duplicated.

Data Cleaning for Tables 3 to 5

The data cleaning process for Tables 3, 4, and 5 is similar to Tables 1 and 2. I started with renaming and dropping the duplicated information and then used melt and assign functions to solve the problems. The main difference is that these three tables, it contains values as a percentage and rate changes. Therefore, the output of the data cleaning is a bit different. For Tables 1 and 2, the purpose was to easily access each country's immigrant number separated by year and gender. For these three tables, instead of calling for the exact number, we probably want to use a histogram to visualize the percentage of refugees in the country or the female percentage in the total population. For Table 5, we might want to use line charts to see the trends of the rate change.

Data Cleaning for Table 6

Table 6 has more complex data than the previous five tables, which comes with more problems when we consider the Tidy data principles. The table contains three types of information, including the estimated refugee number (counts), refugee percentage (rate), and annual rate change (values). There are potentially two ways to clean the data. The first is to split the dataset into three parts based on the three types of information. The second is to melt these three types of information into three columns with unique information. I choose to split the dataset because they share the primary key of the country code, and we can have a clear overview of each piece of information.

I started to drop the information and rename columns with the same process. Then, I split the three types of information using the loc function. Then for each split table, I used the melt and similar process to clean the table.