INF1340 Midterm Writeup

Zhou Laien

2022/11/6

Table 1:

Step 1: Read the table, drop meaningless lines and reallocate column names

note that the column names are allocated as a string "gender+year" for further steps.

Step 2: melt the data to make all columns be variables

According to the principle 1, column names need to be informative and variable names should not be values.

Step 3: Split genderyear column to two column with variables gender and year separately

Step 4: Change the names to proper form and reset the index

Table 2:

The steps of tiding table 2 data are the same as the steps of table 1.

Table 3:

The steps of tiding table 3 data are the same as the steps of table 1.

Table 4:

Step 1: lines and reallocate column names

Step 2: Complete the other half of the information (male) in the table

This step is done because i think the percentage of male migrants is also a meaningful information and it is beneficial to the further combined table.

Step 3 - Step 5 is the same as Step 2 -Step 4 of table 1

Table 5:

The Steps here is the same as table 1 except the step marked in the code:

The operation here aims to turn year period into specific year, for example, 1995-2000 is changed to 2000 as well as the annual rate is now calculated based on the past 5 years so it does not change the meaning. Without such step, it may cause errors to variable "year" when all tables are further combined together.

Table 6:

Step 1: Read the table, drop meaningless lines and reallocate column names

The column names here are strings in the form of "value_types+gender+year", for example, str[0] = "e" indicates that the value is from "Estimated refugee stock at mid-year" variable, and similar for the other two alphabets.

Marked Step 1.5: This step aims to add a new empty column to represent the Annual rate of change of the refugee stock in past 5 years of 1990 (it dose not has values according to our dataset). The reason to do so is the same as the marked step of table 5.

Step 2: melt the data to make all columns be variables

According to the principle 1, column names need to be informative and variable names should not be values.

Step 3: Split typegenderyear column to three column with variables variable types, gender and year seperately. Then change the names to proper form and reset the index.

Step 4: Convert the three variables in the "type" column to three distinct columns with one variable.

According to principle 4, each table column needs to have a singular data type.

Combined Table:

I combine all seperate table to a large final table since every of these tables consists of 6 columns that are totally the same, which I considered to be a violation to the principle 5: a single observational units must be in one table. And on the other hand, the combined table dose not violate any principles, so I regard it as a proper result.