Write-up

Introduction

This project is about wrangling the data of trends in international migrant stock revised in 2015 based on the tidy data principles. The excel data file includes 8 sheets in total: tables 1 to 6 are sheets containing the data related to the migrant stock, the annex sheet is a table for classification of countries and areas by major area and region, and the notes sheet is a footnote table that explains some information in table 1 to 6. The target goal of wrangling this data in tidyverse is to have a more readable and clear dataset to be analyzed. My desired result is a table that contains its key values in the exact year which means the value of migrant information collected in each table, and also includes a country's identification information such as a country's major area, region, and development situation.

Method & Result

<u>Steps of tidying table 1 to 3, and 5 data</u>

Step 1: Load the data from excel file

Step 2: Drop the meaningless columns -- sort order, notes

Step 3: Use a for-loop assign the major area, region, developed, least developed, and Sub-Saharan Africa columns from annex table to the current table regarding the equal country name.

Step 4: Use melt function let gender-year and target values to be two column names in order to satisfy the tidy data principle 1 which is letting column names to be informative, variable names and not values.

Step 5: Use assign function and lambda to separate year and gender to two variables based on

principle 2(each column needs to consist of one and only one variable).

Step 6: Split the dataset by unique year, since I want to analyze the data regarding the migrant stock in different years.

Steps of tidying table 4 data

Step 1: Load the data from excel file

Step 2: Drop the meaningless columns -- sort order, notes

Step 3: Use a for-loop assign the major area, region, developed, least developed, and Sub-Saharan Africa columns from annex table to the current table regarding the equal country name.

Step 4: Use melt function let year and target values to be two column names in order to satisfy the tidy data principle 1 which is letting column names to be informative, variable names and not values.

Step 5: Split the dataset by unique year, since I want to analyze the data regarding the migrant stock in different years.

Steps of tidying table 6 data

Step 1: Load the data from excel file

Step 2: Drop the meaningless columns -- sort order, notes

Step 3: Use a for-loop assign the major area, region, developed, least developed, and Sub-Saharan Africa columns from annex table to the current table regarding the equal country name.

Step 4: Split dataset into three tables regarding the different target variables

Step 5: Use three melt functions let year and target values of those three tables to be two column names in order to satisfy the tidy data principle 1 which is letting column names to be

informative, variable names and not values.

Conclusion

Overall, I achieved my target goal which is to have clear identification information for each country and the migrant stock variables. However, I did not work on the missing and zero values since they are not part of tidy data principles. Moreover, I append the classification of countries and areas by major area and region from the annex sheet instead of building functions and classifying the areas and regions based on the data from the original tables, but I do not know if this follows the tidy data rules.