My COLAB LINK:
https://colab.research.google.com/drive/1n-6Q3XdOatINyA3SxNeMGb8noKBLIWRz?usp=sharing

Introduction

In this data analysis project, we need to do the data wrangling and cleaning for the dataset "Trends in international migrant stock: the 2015 Revision", published by the United Nations - Population Division's Department of Economic and Social Affairs. The main objective of this project is to program the entire data set in a format that is more suitable for data analysis based on the principles and standards of tidy data while preserving the original data and improving its computer readability in preparation for future visualization processes.

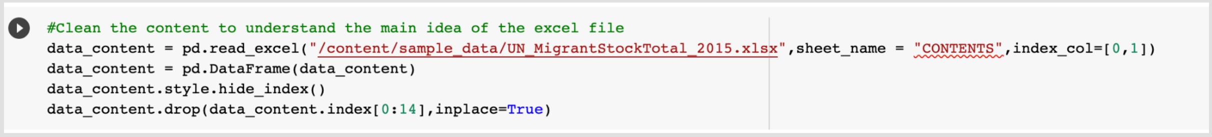
The overall data set includes data for 265 regional countries (Major area, region, country, or area of destination), and the data are available every five years（mid of each year） from 1990 to 2015. This data set includes a total of six main tables from table1-table6, which are "*International migrant stock at mid-year by sex and by major area, region, country or area, 1990-2015*"; "*Total population at mid-year by sex and by major area, region, country or area, 1990-2015 (thousands)*"; "*International migrant stock as a percentage of the total population, 1990-2015*"; "*Female migrants as a percentage of the international migrant stock by major area, region, country or area, 1990-2015*"; "*Annual rate of change of the migrant stock by sex and by major area, region, country or area, 1990-2015 (percentage)*"; "*Estimated refugee stock at mid-year by major area, region, country or area, 1990-2015*".

These six tables are messy data because their structure, the arrangement is not in line with the tidy data principles, and the structure of these six tables is generally similar. Therefore I will use similar techniques to clean and wrangle these data. In the next section, I will detail my specific methods and motivations/reasons for the process of cleaning up this data.

Methods and Result

- ***Content***

For the table of contents, although he did not contain valuable data and content, I imported him into my Colab Notebook as a first step to understanding the general content of the whole data set and enhancing the logic and integrity of the whole project.



```
#Clean the content to understand the main idea of the excel file
data_content = pd.read_excel("/content/sample_data/UN_MigrantStockTotal_2015.xlsx",sheet_name = "CONTENTS",index_col=[0,1])
data_content = pd.DataFrame(data_content)
data_content.style.hide_index()
data_content.drop(data_content.index[0:14],inplace=True)
```

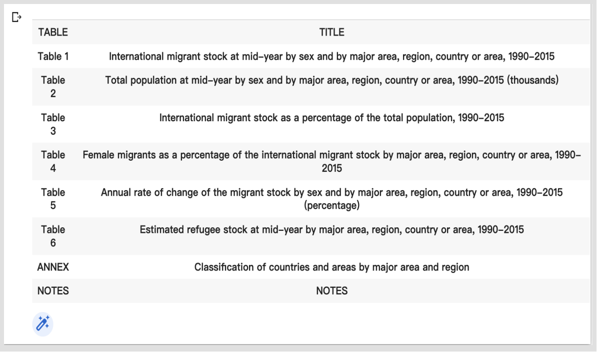


TABLE	TITLE
Table 1	International migrant stock at mid–year by sex and by major area, region, country or area, 1990–2015
Table 2	Total population at mid–year by sex and by major area, region, country or area, 1990–2015 (thousands)
Table 3	International migrant stock as a percentage of the total population, 1990–2015
Table 4	Female migrants as a percentage of the international migrant stock by major area, region, country or area, 1990–2015
Table 5	Annual rate of change of the migrant stock by sex and by major area, region, country or area, 1990–2015 (percentage)
Table 6	Estimated refugee stock at mid–year by major area, region, country or area, 1990–2015
ANNEX	Classification of countries and areas by major area and region
NOTES	NOTES

- ***Table categories***

Since the structure of the six tables is roughly the same, many of the methods and steps in the process are similar. So in this section, I will focus on the differences in the processing of the six tables without repeating the same processing steps and methods.

Before starting this project, I skimmed through the entire data set in general, understanding the structure and specifics of each table separately and the similarities and differences between them. I find that all tables contain some of the same columns of country/region information, such as "Sort\order," "Major area, region, country or area of destination," "Notes," "Country code," "Type of data (a)," I define these columns as background information columns.

First, I divided the six tables into three groups based on their structure. The structures of Table 1,2,3,5 are the same, so I applied the same techniques and methods in the processing of cleaning and wrangling. These four tables contain not only the background information columns but also statistics on the different genders of a topic (both sex, female, male)

Secondly, Table 4 only includes statistics for females. Therefore the structure is relatively more uncomplicated, so the cleaning process will also be relatively short.

Thirdly, Table 6 is more complicated because this table contains three different statistics. Therefore, I divided Table 6 into three sub-tables to clean it up better.

- ***Review tidy data principle***

Since this table is based on the tidy data principle to tidy and clean up the data, I would like to briefly review and summarize the main contents of the tidy data principle before explaining my steps/methods and motivation to enhance the logic and clarity.

The tidy data principle emphasizes:

- *Every column is a variable*

- *Every row is an observation.*
- *Every cell is a single value.*

The tidy data principle:

tidy data principle #1: column names need to be informative, variable names and not values
tidy data principle #2: each column needs to consist of one and only one variable
tidy data principle #3: variables need to be in cells, not rows and columns
tidy data principle #4: each table column needs to have a singular data type
tidy data principle #5: a single observational unit must be in 1 table

- ***Explanation of steps - Table 1, Table 2, Table 3, Table5***

Table1-International migrant stock at mid-year by sex and by major area, region, country or area, 1990-2015

Table2- Total population at mid-year by sex and by major area, region, country or area, 1990-2015 (thousands)

Table3- International migrant stock as a percentage of the total population, 1990-2015

Table5 - Annual rate of change of the migrant stock by sex and by major area, region, country or area, 1990-2015 (percentage)

1. Import- I start by importing excel-tables to Colabs Notebook. Since this data set contains many tables, I need to be careful to make sure that the paths and table names are accurate.

2. Clean up - The upper part of each of the four tables contains a fixed template introduction, they do not contain any valuable data or information, so I used the ".drop" function to remove the unwanted rows.

3. Set new head row- I set a new header row for the new data frame since, in the previous step, I deleted all the rows above the table that I didn't need, so now I need to redefine the head.

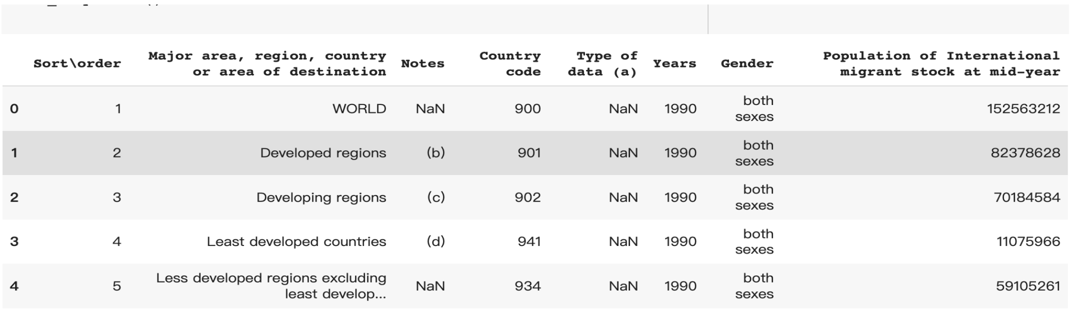
4. Renaming - Since we lost some vital information during the table extraction: the year, I decided to rename the header to include this information back in the table. The new header now includes gender, year, and statistic title.

5. Identify issues - At this point, we have extracted the original table into the Colab notebook. However, we can see that the newly extracted table needs to be revised with the tidy data principle. For example, they have included variables in the rows, multiple variables in one column, etc. So I will use the techniques I learned to adjust the whole table to ensure the final table complies with the tidy data principle.

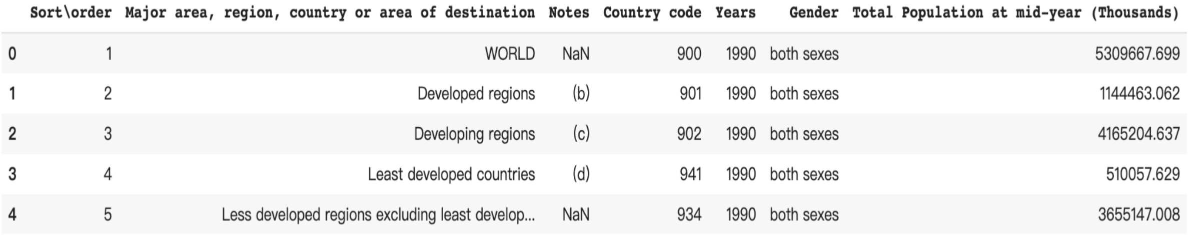
6. Melt function - I use the melt function to keep the background information columns of the table and then turn the rows containing the “table’s name,” “gender,” “years,” and all variables of the table into columns (in other words, the purpose of melt is to turn a relatively wide table into a relatively long table).

7. Now we find that even though all the variables have become one column, the “years,” “gender,” and “table’s name” still appear in the same column named “Years/Sexes” due to the renaming I just did so all I have to do now is to separate the “gender,” “year” and the “table’s name” into three separate columns.

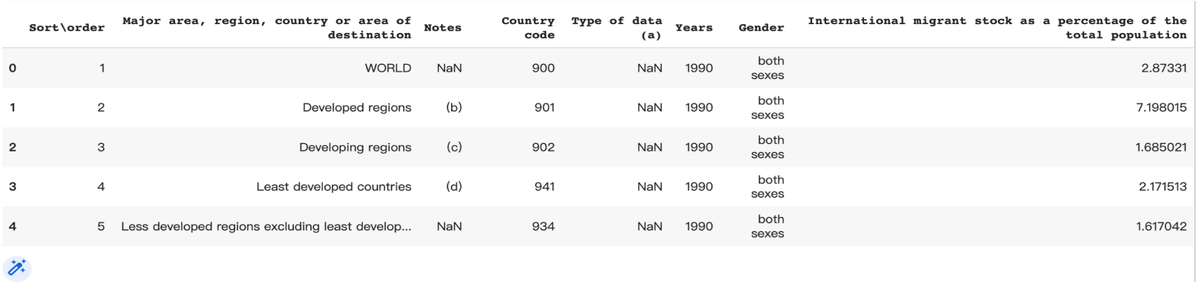
8. Separate - I use the “str. split” function to separate the “year,” “table’s name,” and “gender,” and then use the drop function to remove the duplicate columns. In the following, I have shown that the table organized with a melt function includes only the “year,” “gender,” and “all variables” in a separate column, in addition to the background information columns.



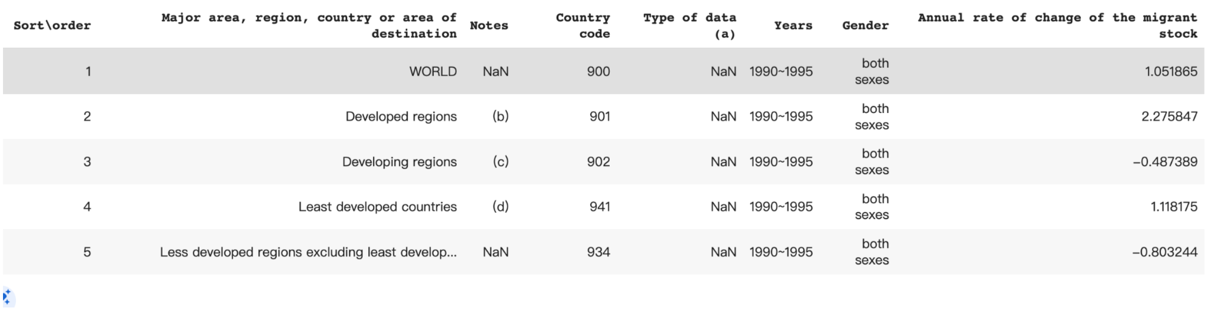
	Sort\order	Major area, region, country or area of destination	Notes	Country code	Type of data (a)	Years	Gender	Population of International migrant stock at mid-year
0	1	WORLD	NaN	900	NaN	1990	both sexes	152563212
1	2	Developed regions	(b)	901	NaN	1990	both sexes	82378628
2	3	Developing regions	(c)	902	NaN	1990	both sexes	70184584
3	4	Least developed countries	(d)	941	NaN	1990	both sexes	11075966
4	5	Less developed regions excluding least develop...	NaN	934	NaN	1990	both sexes	59105261



	Sort\order	Major area, region, country or area of destination	Notes	Country code	Years	Gender	Total Population at mid-year (Thousands)
0	1	WORLD	NaN	900	1990	both sexes	5309667.699
1	2	Developed regions	(b)	901	1990	both sexes	1144463.062
2	3	Developing regions	(c)	902	1990	both sexes	4165204.637
3	4	Least developed countries	(d)	941	1990	both sexes	510057.629
4	5	Less developed regions excluding least develop...	NaN	934	1990	both sexes	3655147.008



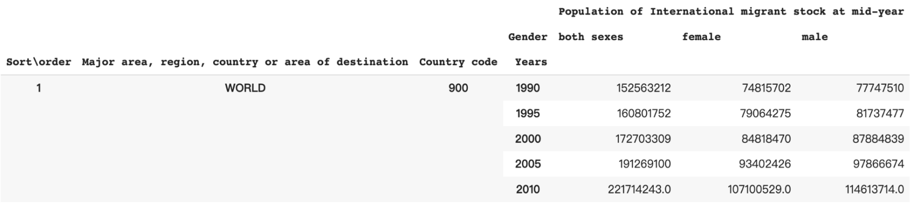
	Sort\order	Major area, region, country or area of destination	Notes	Country code	Type of data (a)	Years	Gender	International migrant stock as a percentage of the total population
0	1	WORLD	NaN	900	NaN	1990	both sexes	2.87331
1	2	Developed regions	(b)	901	NaN	1990	both sexes	7.198015
2	3	Developing regions	(c)	902	NaN	1990	both sexes	1.685021
3	4	Least developed countries	(d)	941	NaN	1990	both sexes	2.171513
4	5	Less developed regions excluding least develop...	NaN	934	NaN	1990	both sexes	1.617042



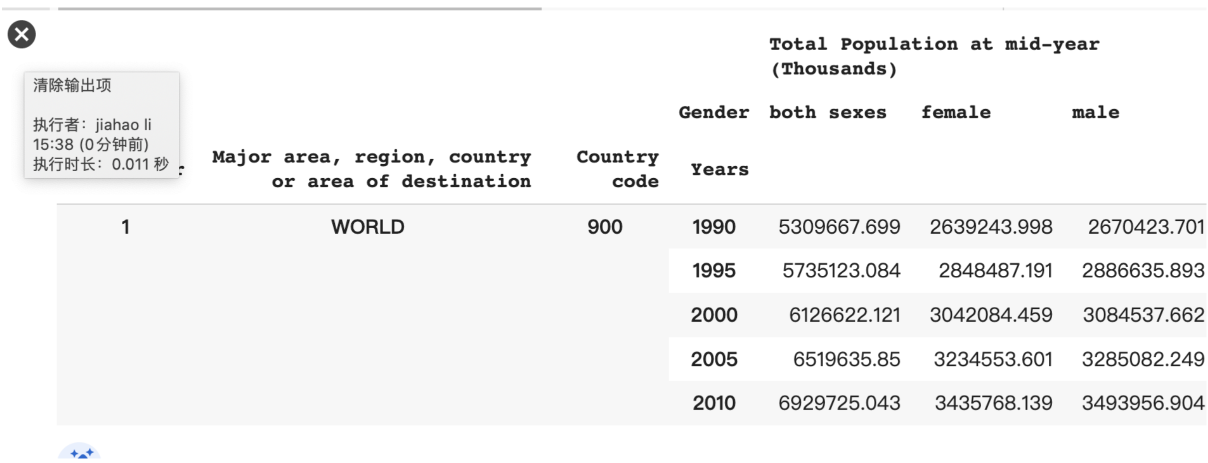
Sort\order	Major area, region, country or area of destination	Notes	Country code	Type of data (a)	Years	Gender	Annual rate of change of the migrant stock
1	WORLD	NaN	900	NaN	1990~1995	both sexes	1.051865
2	Developed regions	(b)	901	NaN	1990~1995	both sexes	2.275847
3	Developing regions	(c)	902	NaN	1990~1995	both sexes	-0.487389
4	Least developed countries	(d)	941	NaN	1990~1995	both sexes	1.118175
5	Less developed regions excluding least develop...	NaN	934	NaN	1990~1995	both sexes	-0.803244

9. Discuss the tables after "melt function."- as of now, our table has met the requirements of the tidy data principle, such as Column names are informative variable names such as gender and year, and each column consists of only one variable. Every cell is a single value etc.

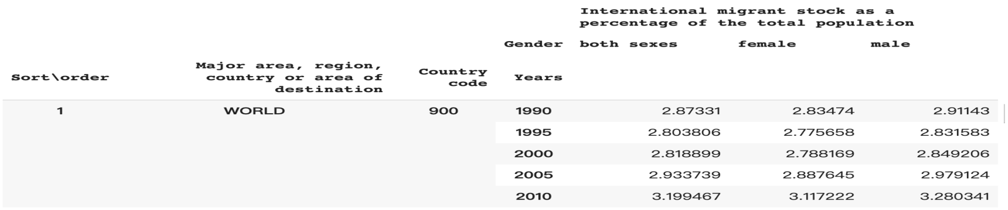
10. "pivot table function" - After using the melt function, I applied the pivot table function to the data frame because I realized it would be more straightforward for the analysis. The pivot table takes simple column-wise data as input and groups the entries into a two-dimensional table that provides a multidimensional summarization of the data. Therefore, in this case, I only kept the key basic information column, sort/order destination and country number. The relevant data for a particular destination (region/country) are shown together, whether it is to find a specific year or a particular gender; it is evident and intuitive, as I show below the final structure (results) of Tables 1-3, and 5, respectively.



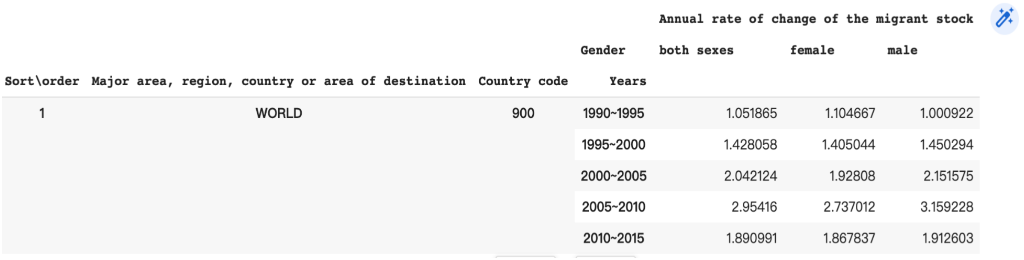
				Population of International migrant stock at mid-year		
			Gender	both sexes	female	male
Sort\order	Major area, region, country or area of destination	Country code	Years			
1	WORLD	900	1990	152563212	74815702	77747510
			1995	160801752	79064275	81737477
			2000	172703309	84818470	87884839
			2005	191269100	93402426	97866674
			2010	221714243.0	107100529.0	114613714.0



				Total Population at mid-year (Thousands)		
			Gender	both sexes	female	male
	Major area, region, country or area of destination	Country code	Years			
1	WORLD	900	1990	5309667.699	2639243.998	2670423.701
			1995	5735123.084	2848487.191	2886635.893
			2000	6126622.121	3042084.459	3084537.662
			2005	6519635.85	3234553.601	3285082.249
			2010	6929725.043	3435768.139	3493956.904



				International migrant stock as a percentage of the total population		
			Gender	both sexes	female	male
Sort\order	Major area, region, country or area of destination	Country code	Years			
1	WORLD	900	1990	2.87331	2.83474	2.91143
			1995	2.803806	2.775658	2.831583
			2000	2.818899	2.788169	2.849206
			2005	2.933739	2.887645	2.979124
			2010	3.199467	3.117222	3.280341



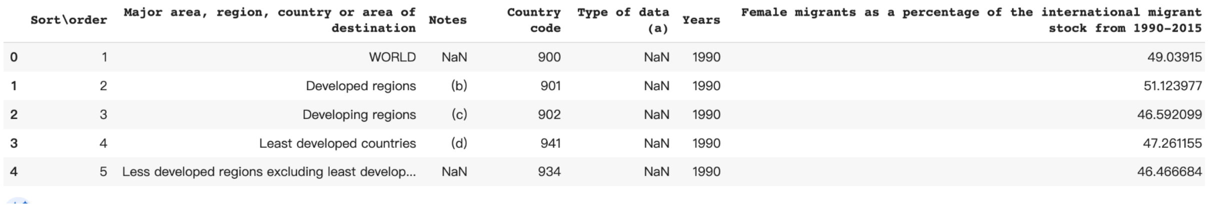
				Annual rate of change of the migrant stock		
			Gender	both sexes	female	male
Sort\order	Major area, region, country or area of destination	Country code	Years			
1	WORLD	900	1990~1995	1.051865	1.104667	1.000922
			1995~2000	1.428058	1.405044	1.450294
			2000~2005	2.042124	1.92808	2.151575
			2005~2010	2.95416	2.737012	3.159228
			2010~2015	1.890991	1.867837	1.912603

- ***Explanation of steps - Table 4***

Female migrants as a percentage of the international migrant stock by major area, region, country or area, 1990-2015

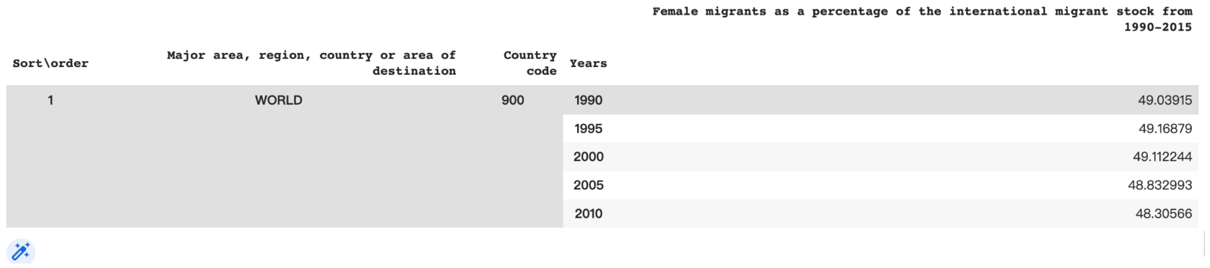
After reviewing Table 4, it contains less content than the other tables because it includes only the percentage of female immigrants as a percentage of global immigrants for each of the five years 1990-2015, in addition to the basic information column. Although there is no difference in the general methods from Table 1 - 3, Table 4 does not need to consider gender disaggregation because it only records data on females.

The preliminary steps of Table 4 are similar to Table1,2,3,5 (refer to ***Explanation of steps Table 1, Table 2, Table 3, Table5***), including the importing step, setting a new header row to rename the header, etc. The only difference is that I don't need to consider keeping the gender column separately in the melt step, because all the data in the whole table is related to females. Below I show Table 4 after melting.



	Sort\order	Major area, region, country or area of destination	Notes	Country code	Type of data (a)	Years	Female migrants as a percentage of the international migrant stock from 1990-2015
0	1	WORLD	NaN	900	NaN	1990	49.03915
1	2	Developed regions	(b)	901	NaN	1990	51.123977
2	3	Developing regions	(c)	902	NaN	1990	46.592099
3	4	Least developed countries	(d)	941	NaN	1990	47.261155
4	5	Less developed regions excluding least develop...	NaN	934	NaN	1990	46.466684

Similarly, I ended up using a pivot_table function to make the whole table look more organized. As I mentioned above, this allows the related data for a particular region to be summarized together, making it easier to analyze later.



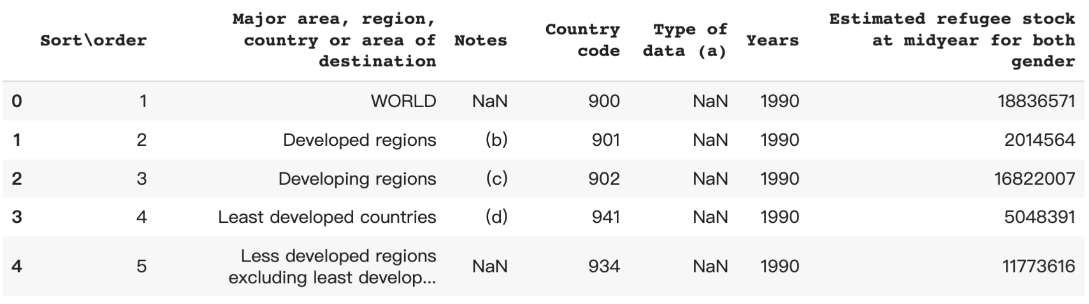
				Female migrants as a percentage of the international migrant stock from 1990-2015
Sort\order	Major area, region, country or area of destination	Country code	Years	
1	WORLD	900	1990	49.03915
			1995	49.16879
			2000	49.112244
			2005	48.832993
			2010	48.30566

Explanation of stepsTable 6

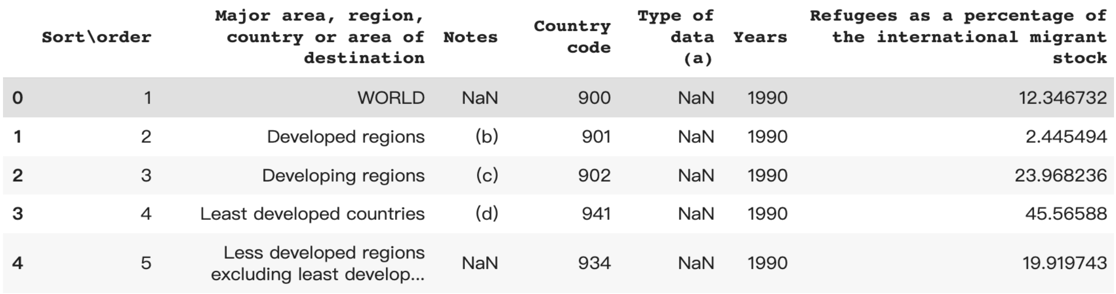
Table 6 is the most complex table in the entire dataset, because the structure of the remaining five tables is similar, and each table contains only one type of statistics, while 6 contains three different statistics which are "*Estimated refugee stock at mid-year", "Refugees as a percentage of international migrant stocks", and "Annual rate of change refugee stocks".* Therefore, Table 6 cannot be treated as a whole like the other tables, but it is necessary to discuss the three statistics contained in Table 6 separately.

Estimated refugee stock at mid-year by major area, region, country or area, 1990-2015

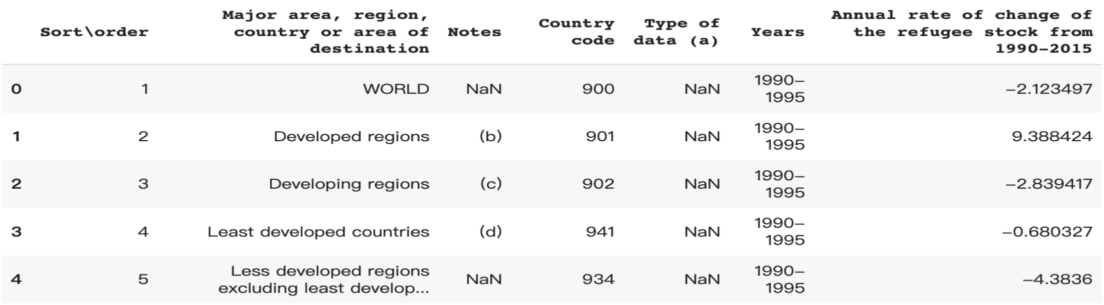
After splitting Table 6, I will look at the three tables as three separate tables, using the same method as the other tables (melt and pivot table) to make the format of these three tables unified, below I show the melt table and pivot table table respectively.



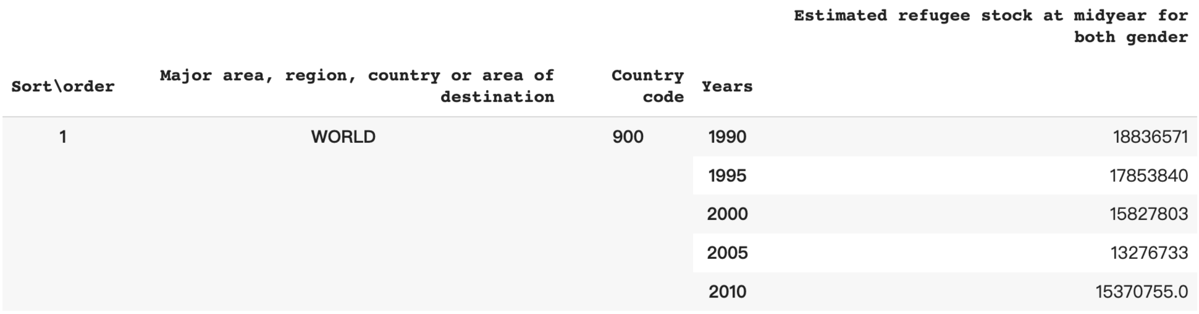
	Sort\order	Major area, region, country or area of destination	Notes	Country code	Type of data (a)	Years	Estimated refugee stock at midyear for both gender
0	1	WORLD	NaN	900	NaN	1990	18836571
1	2	Developed regions	(b)	901	NaN	1990	2014564
2	3	Developing regions	(c)	902	NaN	1990	16822007
3	4	Least developed countries	(d)	941	NaN	1990	5048391
4	5	Less developed regions excluding least develop...	NaN	934	NaN	1990	11773616



	Sort\order	Major area, region, country or area of destination	Notes	Country code	Type of data (a)	Years	Refugees as a percentage of the international migrant stock
0	1	WORLD	NaN	900	NaN	1990	12.346732
1	2	Developed regions	(b)	901	NaN	1990	2.445494
2	3	Developing regions	(c)	902	NaN	1990	23.968236
3	4	Least developed countries	(d)	941	NaN	1990	45.56588
4	5	Less developed regions excluding least develop...	NaN	934	NaN	1990	19.919743

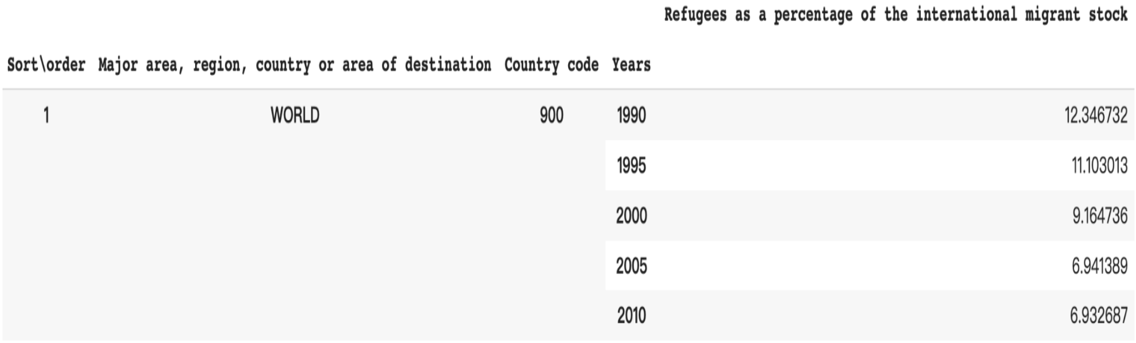


	Sort\order	Major area, region, country or area of destination	Notes	Country code	Type of data (a)	Years	Annual rate of change of the refugee stock from 1990-2015
0	1	WORLD	NaN	900	NaN	1990–1995	−2.123497
1	2	Developed regions	(b)	901	NaN	1990–1995	9.388424
2	3	Developing regions	(c)	902	NaN	1990–1995	−2.839417
3	4	Least developed countries	(d)	941	NaN	1990–1995	−0.680327
4	5	Less developed regions excluding least develop...	NaN	934	NaN	1990–1995	−4.3836

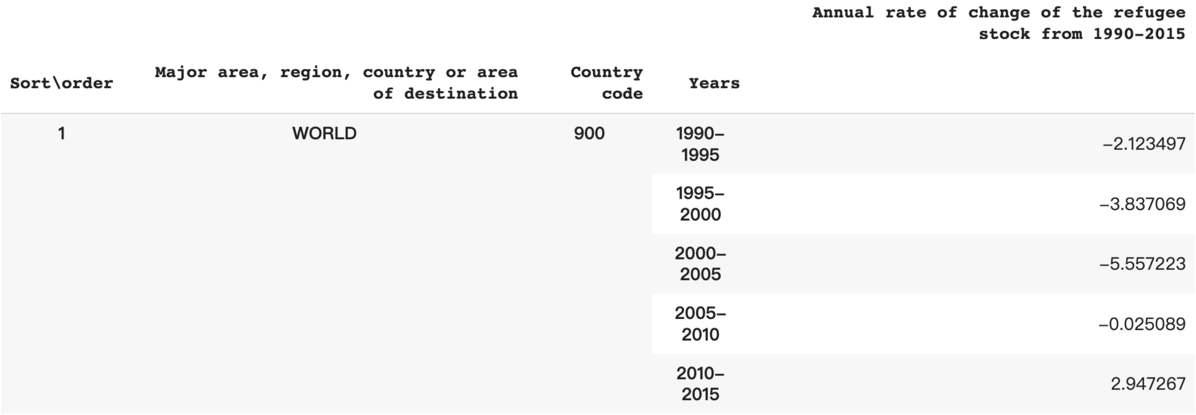


				Estimated refugee stock at midyear for both gender
Sort\order	**Major area, region, country or area of destination**	**Country code**	**Years**	
1	WORLD	900	1990	18836571
			1995	17853840
			2000	15827803
			2005	13276733
			2010	15370755.0

6.2 "Refugees as a percentage of the international migrant stock"



				Refugees as a percentage of the international migrant stock
Sort\order	Major area, region, country or area of destination	Country code	Years	
1	WORLD	900	1990	12.346732
			1995	11.103013
			2000	9.164736
			2005	6.941389
			2010	6.932687



				Annual rate of change of the refugee stock from 1990-2015
Sort\order	Major area, region, country or area of destination	Country code	Years	
1	WORLD	900	1990–1995	−2.123497
			1995–2000	−3.837069
			2000–2005	−5.557223
			2005–2010	−0.025089
			2010–2015	2.947267

Discussion

In this section, I would like to discuss the problems in the original data tables based on each tidy data principle, how I dealt with these problems and my motivations. To illustrate the thought process of cleaning up the information and demonstrate my analytical logic.

Before starting, I skimmed through the six tables, identifying some problems first, developing a method based on those problems and then cleaning up the data. Since the focus of our project was to organize the tables according to the tidy data principle, the structure of the tables was our focus rather than the content. The six tables have a similar structure, which means they also have the same issues.

The issues with the original header rows are undeniable; they need to be more precise and comply with tidy data principle #1. This is because they contain the variable's name of the year, gender, and variables in column names. After the adjustment, the header rows became very clear, including only the background information columns (order, country, country number, etc.) and gender, year, and statistical content (variables). The original data columns do not contain just one variable. Table 1, for example, included both year and gender in one column. After cleaning, the new table has only one variable per column. In other words, all these jumbled columns are separated into separate columns according to the new column names. The year becomes the header row in the original data, which also does not conform to the tidy data #3 formulation. After cleaning and tidying, the years become variables under the "Years" column.

For the tables after cleaning, the final structure of the tables presented is broadly similar, despite the slight differences in the structure and cleaning process of the six tables. After the melt function, they all contain the background information column ("Sort\order," "Major area, region, country or area of destination," "Notes," "Country code," "Type of data (a)"), followed by the year, gender, and what is being counted. For Table 4, I removed the gender column because the entire table's data are statistics for women. They contain a variable in each column and an observation in each row. This structure not only complies with the tidy data principles but also makes the table structure more clear and concise. After applying the pivot table, some of the same basic information is integrated; for example, for the same country data, we only show the country/region once. And related data about this region/country will be displayed together, laying the foundation for later analysis and visualization.