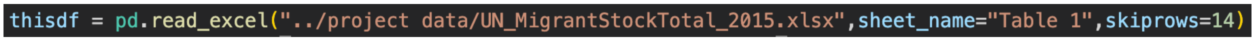
Intro:

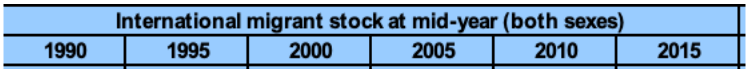
This is the process of clean “Trends in International Migrant Stock”.
There are 7 notebooks in the file. table_1,2,3,4,5,6 are the notebooks that clean each table. MergeTable is the notebook that merge the clean results of table_1,2,3,4,5. Because I have different notebooks, for simple export and import, I will let table_1,2,3,4,5 write excel files that can be read for MergeTable.

Table 1:



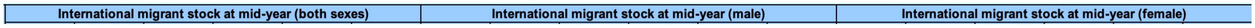
```
thisdf = pd.read_excel("../project data/UN_MigrantStockTotal_2015.xlsx",sheet_name="Table 1",skiprows=14)
```

Read the table 1 and skip the first 14 rows. Because the first 14 rows are irrelated to the content of the data.



International migrant stock at mid-year (both sexes)					
1990	1995	2000	2005	2010	2015

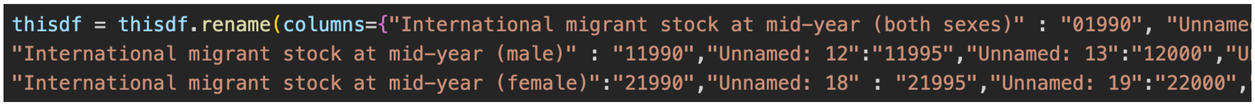
The second row of this table is the years. According to tidy data principle #3: variables need to be in cells, not rows and columns. (Years are the columns under the category gender)



International migrant stock at mid-year (both sexes)	International migrant stock at mid-year (male)	International migrant stock at mid-year (female)

The first row shows that the table shows that the table 1 sperate the international migrant stock by gender. I can treat the gender category as values. According to tidy data principle #1: Column names need to be informative, variable names and not values

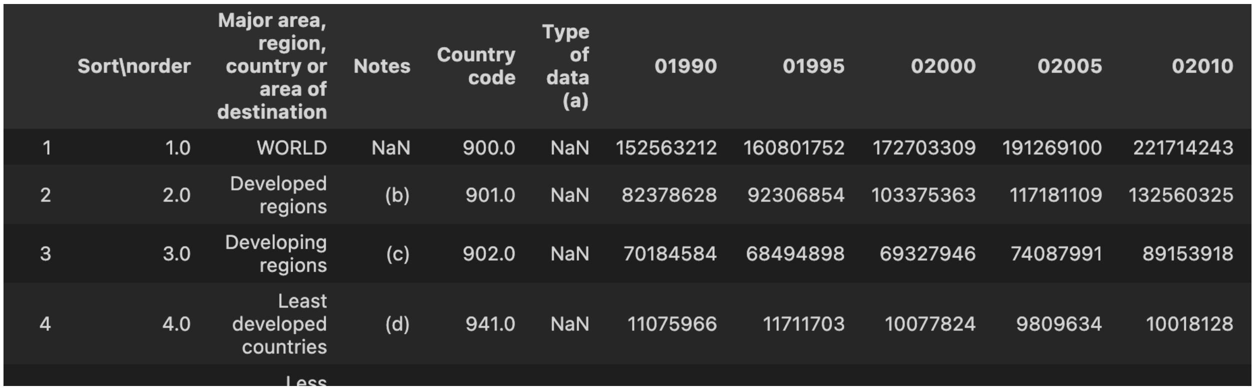
So, I combine the first two principle together and decide to rename the table as below. (The screen can only show the part of the code, please see the complete code in the code file)



```
thisdf = thisdf.rename(columns={"International migrant stock at mid-year (both sexes)" : "01990", "Unname
"International migrant stock at mid-year (male)" : "11990","Unnamed: 12":"11995","Unnamed: 13":"12000","U
"International migrant stock at mid-year (female)":"21990","Unnamed: 18" : "21995","Unnamed: 19":"22000",
```

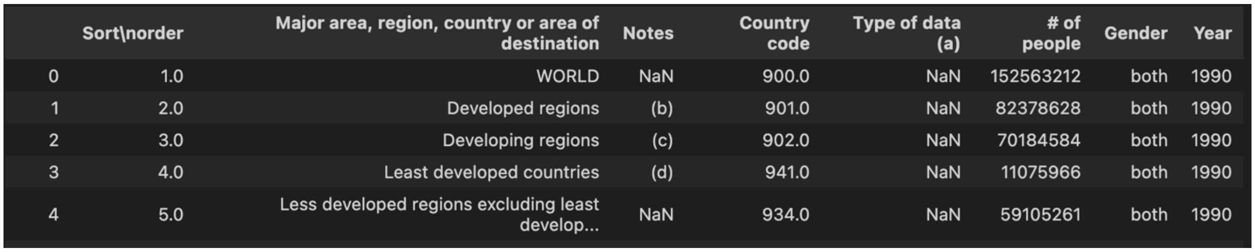
As the screen shot shows, I combine the two information to form new variables. The first character of the variables represents the gender. 0 is the symbol of “both sexes”, 1 is the symbol of “male”. 2 is the symbol of “female”.

The new table looks like:



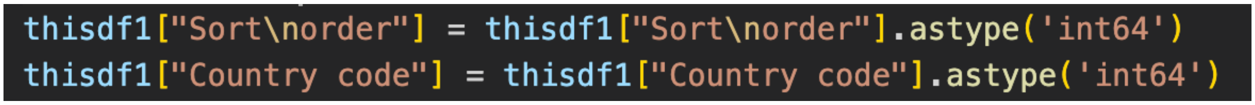
	Sort\norder	Major area, region, country or area of destination	Notes	Country code	Type of data (a)	01990	01995	02000	02005	02010
1	1.0	WORLD	NaN	900.0	NaN	152563212	160801752	172703309	191269100	221714243
2	2.0	Developed regions	(b)	901.0	NaN	82378628	92306854	103375363	117181109	132560325
3	3.0	Developing regions	(c)	902.0	NaN	70184584	68494898	69327946	74087991	89153918
4	4.0	Least developed countries	(d)	941.0	NaN	11075966	11711703	10077824	9809634	10018128
		Less								

As tidy data principle #2: each column needs to consist of one and only one variable. The table needs further cleaning. I used melt() and assign() function to create two separate new columns called “gender” and “sex” to store these two variables.



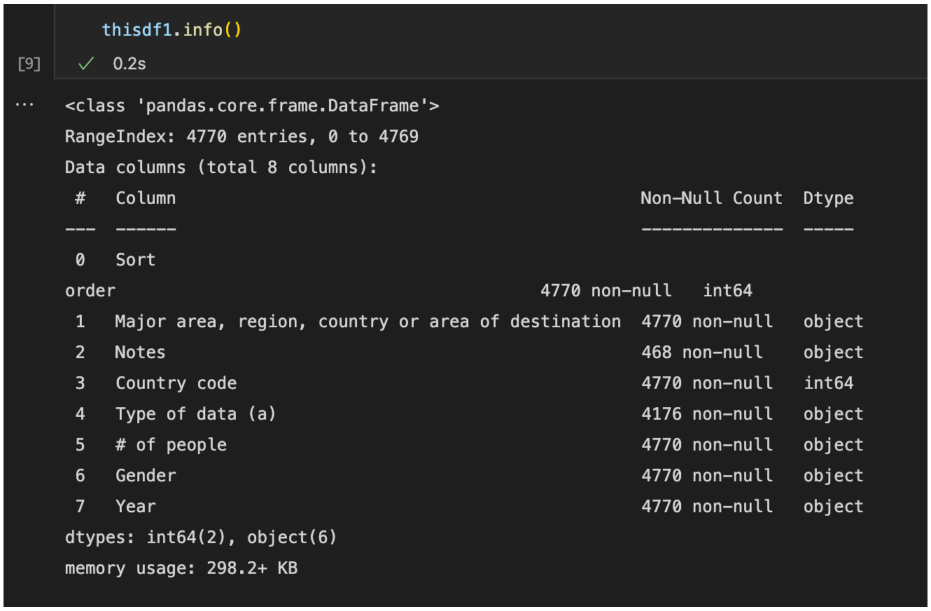
	Sort\norder	Major area, region, country or area of destination	Notes	Country code	Type of data (a)	# of people	Gender	Year
0	1.0	WORLD	NaN	900.0	NaN	152563212	both	1990
1	2.0	Developed regions	(b)	901.0	NaN	82378628	both	1990
2	3.0	Developing regions	(c)	902.0	NaN	70184584	both	1990
3	4.0	Least developed countries	(d)	941.0	NaN	11075966	both	1990
4	5.0	Less developed regions excluding least develop...	NaN	934.0	NaN	59105261	both	1990

As the content of “Sort\norder” and “Country code”, the variables of these two column should be integer type. I use the astype() function to change the type of variables.



```
thisdf1["Sort\norder"] = thisdf1["Sort\norder"].astype('int64')
thisdf1["Country code"] = thisdf1["Country code"].astype('int64')
```

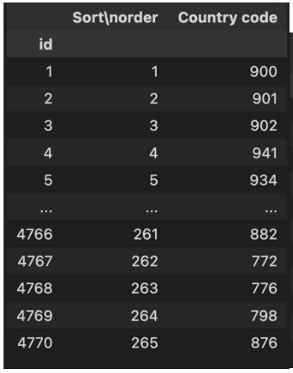
Now check the data types of all variables.



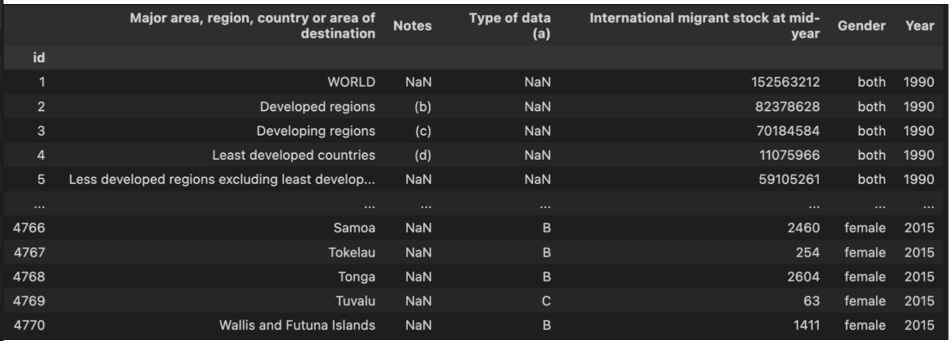
```
thisdf1.info()
[9]  ✓  0.2s

...  <class 'pandas.core.frame.DataFrame'>
     RangeIndex: 4770 entries, 0 to 4769
     Data columns (total 8 columns):
      #   Column                                         Non-Null Count  Dtype
     ---  ------                                         --------------  -----
      0   Sort
     order                                        4770 non-null   int64
      1   Major area, region, country or area of destination  4770 non-null   object
      2   Notes                                          468 non-null    object
      3   Country code                                   4770 non-null   int64
      4   Type of data (a)                               4176 non-null   object
      5   # of people                                    4770 non-null   object
      6   Gender                                         4770 non-null   object
      7   Year                                           4770 non-null   object
     dtypes: int64(2), object(6)
     memory usage: 298.2+ KB
```

This shows that table contains different types of variables (integer and object). As tidy data principle #4: each table column needs to have a singular data type. I separated the table into two sub tables (one table contains only object and one table contains only integer). The final tables look like: (They correspond with each other by id)

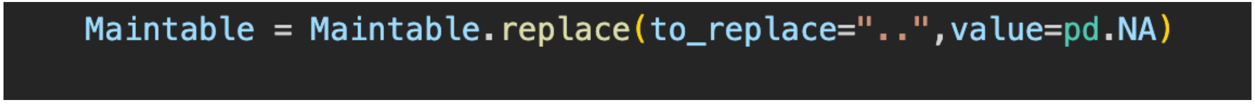


	Sort\norder	Country code
id		
1	1	900
2	2	901
3	3	902
4	4	941
5	5	934
...	...	...
4766	261	882
4767	262	772
4768	263	776
4769	264	798
4770	265	876



	Major area, region, country or area of destination	Notes	Type of data (a)	International migrant stock at mid-year	Gender	Year
id						
1	WORLD	NaN	NaN	152563212	both	1990
2	Developed regions	(b)	NaN	82378628	both	1990
3	Developing regions	(c)	NaN	70184584	both	1990
4	Least developed countries	(d)	NaN	11075966	both	1990
5	Less developed regions excluding least develop...	NaN	NaN	59105261	both	1990
...	...	...	...	...	...	...
4766	Samoa	NaN	B	2460	female	2015
4767	Tokelau	NaN	B	254	female	2015
4768	Tonga	NaN	B	2604	female	2015
4769	Tuvalu	NaN	C	63	female	2015
4770	Wallis and Futuna Islands	NaN	B	1411	female	2015

There are also missing values in the data frame shown as “..”. Change them to NA for better handling when using the data frame later.

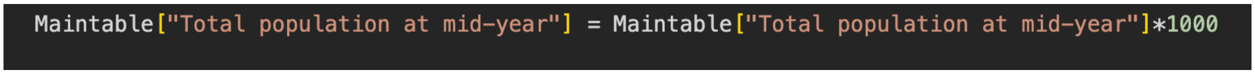


```
Maintable = Maintable.replace(to_replace="..",value=pd.NA)
```

Table 2:

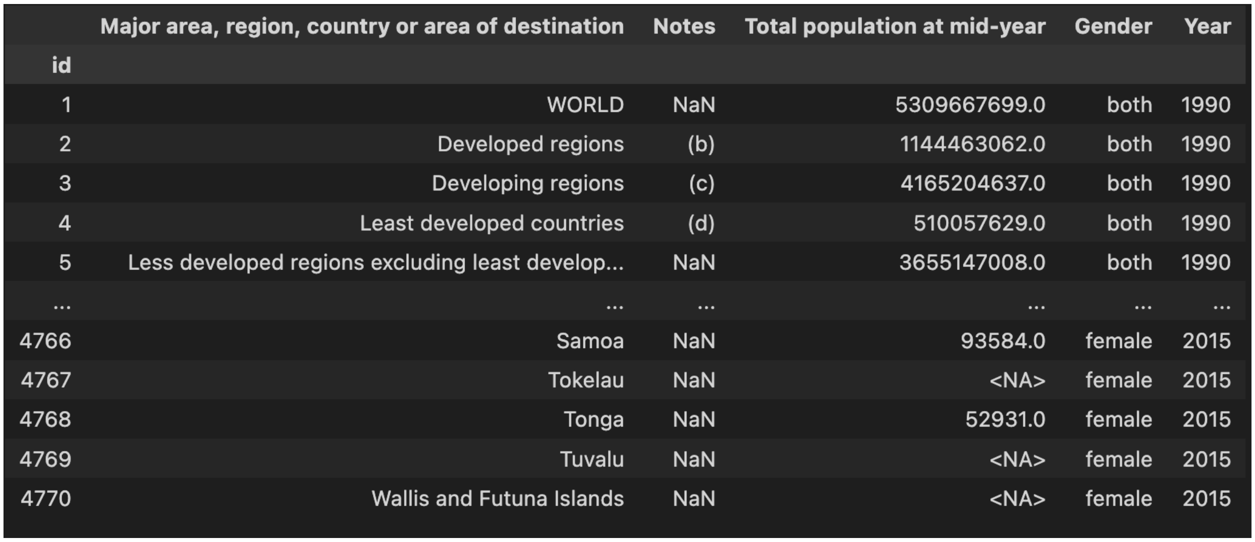
Can be cleaning the same way as Table 1.

I change the unit of total population to one instead of thousand.



```
Maintable["Total population at mid-year"] = Maintable["Total population at mid-year"]*1000
```

The final table looks like: (Sort\norder and Country code table is same as Table 1, will not be shown here)

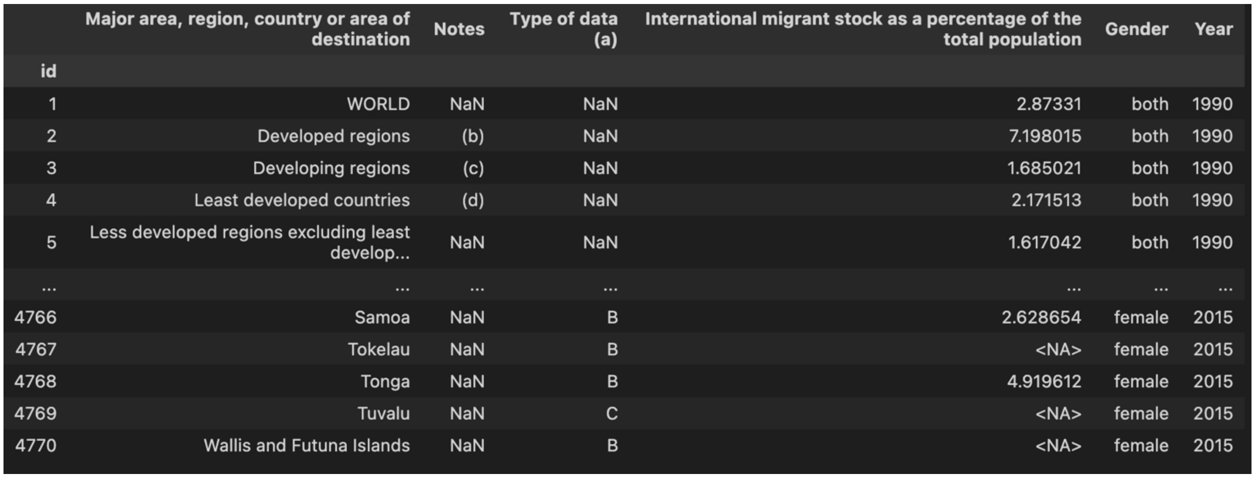


	Major area, region, country or area of destination	Notes	Total population at mid-year	Gender	Year
id					
1	WORLD	NaN	5309667699.0	both	1990
2	Developed regions	(b)	1144463062.0	both	1990
3	Developing regions	(c)	4165204637.0	both	1990
4	Least developed countries	(d)	510057629.0	both	1990
5	Less developed regions excluding least develop...	NaN	3655147008.0	both	1990
...	...	...	...	...	...
4766	Samoa	NaN	93584.0	female	2015
4767	Tokelau	NaN	<NA>	female	2015
4768	Tonga	NaN	52931.0	female	2015
4769	Tuvalu	NaN	<NA>	female	2015
4770	Wallis and Futuna Islands	NaN	<NA>	female	2015

Table 3:

Can be cleaning the same way as Table 1.

The final table looks like: (Sort\norder and Country code table is same as Table 1, will not be shown here)

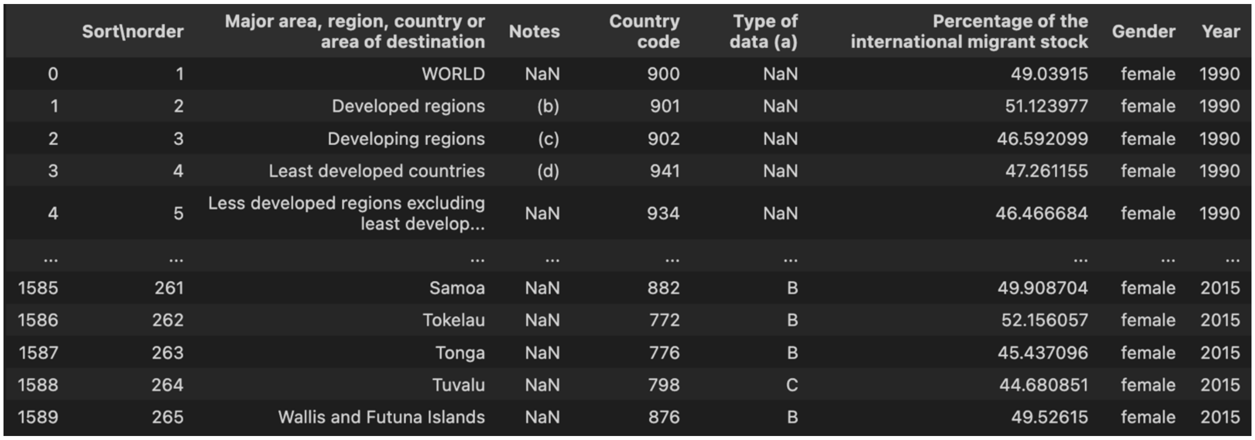


	Major area, region, country or area of destination	Notes	Type of data (a)	International migrant stock as a percentage of the total population	Gender	Year
id						
1	WORLD	NaN	NaN	2.87331	both	1990
2	Developed regions	(b)	NaN	7.198015	both	1990
3	Developing regions	(c)	NaN	1.685021	both	1990
4	Least developed countries	(d)	NaN	2.171513	both	1990
5	Less developed regions excluding least develop...	NaN	NaN	1.617042	both	1990
...	...	...	...	...	...	...
4766	Samoa	NaN	B	2.628654	female	2015
4767	Tokelau	NaN	B	<NA>	female	2015
4768	Tonga	NaN	B	4.919612	female	2015
4769	Tuvalu	NaN	C	<NA>	female	2015
4770	Wallis and Futuna Islands	NaN	B	<NA>	female	2015

Table 4:

First of all, it can be cleaning the same way as Table 1.

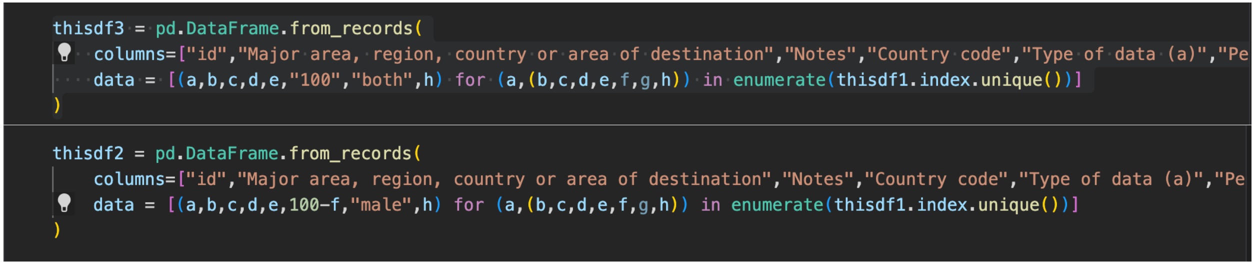
The table looks like this before principle 4:



	Sort\norder	Major area, region, country or area of destination	Notes	Country code	Type of data (a)	Percentage of the international migrant stock	Gender	Year
0	1	WORLD	NaN	900	NaN	49.03915	female	1990
1	2	Developed regions	(b)	901	NaN	51.123977	female	1990
2	3	Developing regions	(c)	902	NaN	46.592099	female	1990
3	4	Least developed countries	(d)	941	NaN	47.261155	female	1990
4	5	Less developed regions excluding least develop...	NaN	934	NaN	46.466684	female	1990
...	...	...	...	...	...	...	...	...
1585	261	Samoa	NaN	882	B	49.908704	female	2015
1586	262	Tokelau	NaN	772	B	52.156057	female	2015
1587	263	Tonga	NaN	776	B	45.437096	female	2015
1588	264	Tuvalu	NaN	798	C	44.680851	female	2015
1589	265	Wallis and Futuna Islands	NaN	876	B	49.52615	female	2015

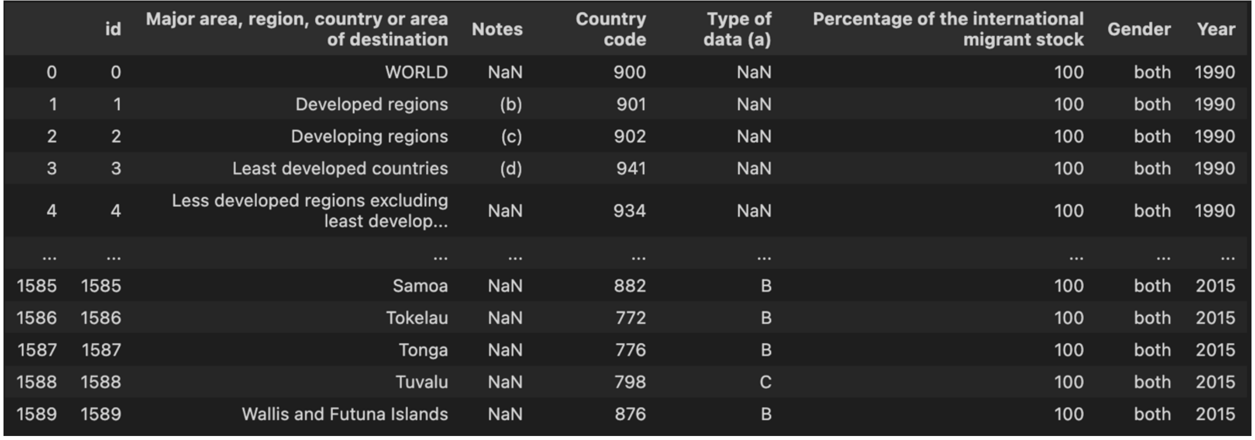
This data frame only contains female information. From the original excel table, I know that gender only have two options (male and female). So, I try to use female information to calculate male information. I do this because I have seen a pattern that most of these table are related to each other. This step can help me in the later merge step.

I set both gender's percentage to 100.

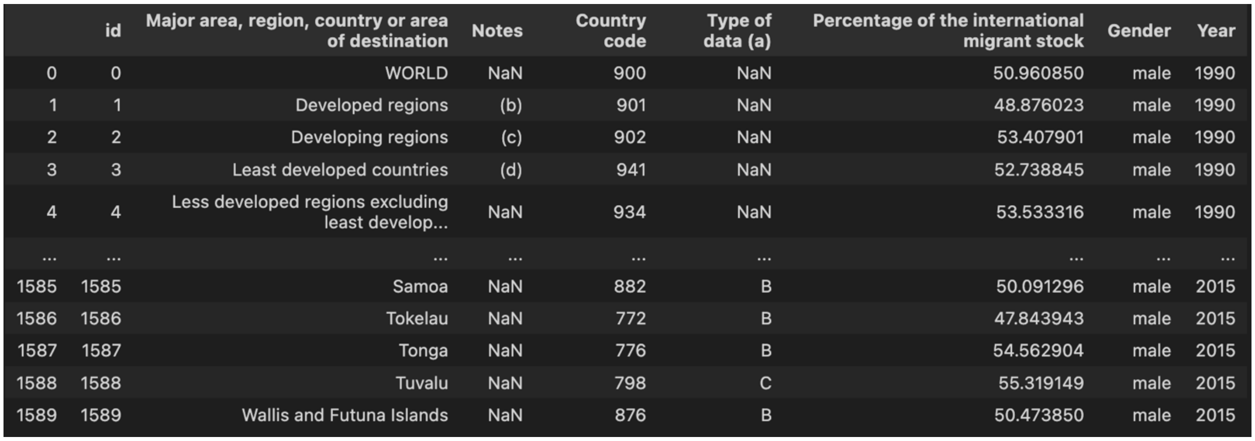


```
thisdf3 = pd.DataFrame.from_records(
    columns=["id","Major area, region, country or area of destination","Notes","Country code","Type of data (a)","Pe
    data = [(a,b,c,d,e,"100","both",h) for (a,(b,c,d,e,f,g,h)) in enumerate(thisdf1.index.unique())]
)

thisdf2 = pd.DataFrame.from_records(
    columns=["id","Major area, region, country or area of destination","Notes","Country code","Type of data (a)","Pe
    data = [(a,b,c,d,e,100-f,"male",h) for (a,(b,c,d,e,f,g,h)) in enumerate(thisdf1.index.unique())]
)
```

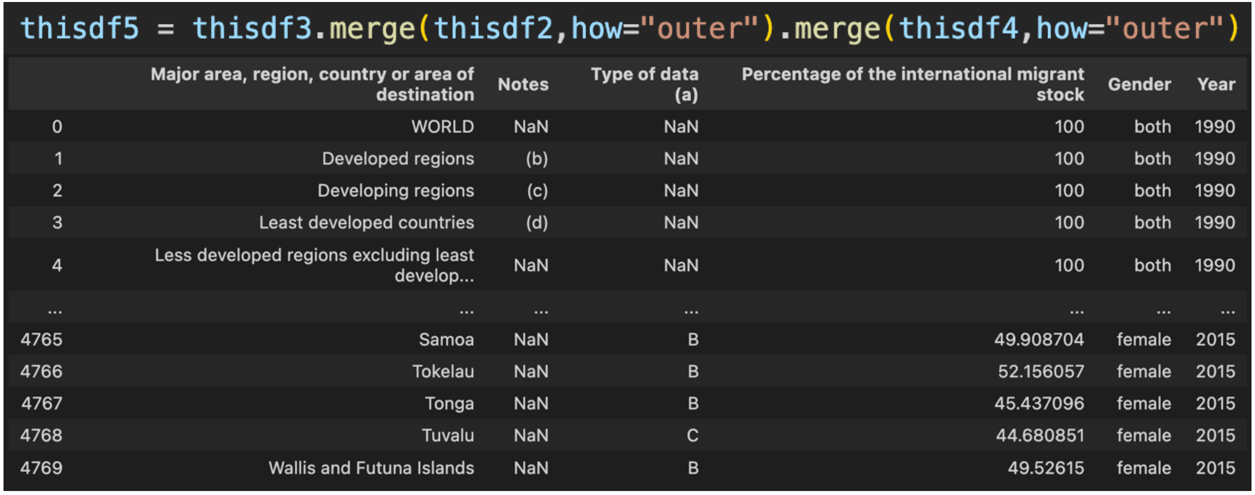


	id	Major area, region, country or area of destination	Notes	Country code	Type of data (a)	Percentage of the international migrant stock	Gender	Year
0	0	WORLD	NaN	900	NaN	100	both	1990
1	1	Developed regions	(b)	901	NaN	100	both	1990
2	2	Developing regions	(c)	902	NaN	100	both	1990
3	3	Least developed countries	(d)	941	NaN	100	both	1990
4	4	Less developed regions excluding least develop...	NaN	934	NaN	100	both	1990
...	...	...	...	...	...	...	...	...
1585	1585	Samoa	NaN	882	B	100	both	2015
1586	1586	Tokelau	NaN	772	B	100	both	2015
1587	1587	Tonga	NaN	776	B	100	both	2015
1588	1588	Tuvalu	NaN	798	C	100	both	2015
1589	1589	Wallis and Futuna Islands	NaN	876	B	100	both	2015



	id	Major area, region, country or area of destination	Notes	Country code	Type of data (a)	Percentage of the international migrant stock	Gender	Year
0	0	WORLD	NaN	900	NaN	50.960850	male	1990
1	1	Developed regions	(b)	901	NaN	48.876023	male	1990
2	2	Developing regions	(c)	902	NaN	53.407901	male	1990
3	3	Least developed countries	(d)	941	NaN	52.738845	male	1990
4	4	Less developed regions excluding least develop...	NaN	934	NaN	53.533316	male	1990
...	...	...	...	...	...	...	...	...
1585	1585	Samoa	NaN	882	B	50.091296	male	2015
1586	1586	Tokelau	NaN	772	B	47.843943	male	2015
1587	1587	Tonga	NaN	776	B	54.562904	male	2015
1588	1588	Tuvalu	NaN	798	C	55.319149	male	2015
1589	1589	Wallis and Futuna Islands	NaN	876	B	50.473850	male	2015

Finally, I merge all the percentages into one data frame

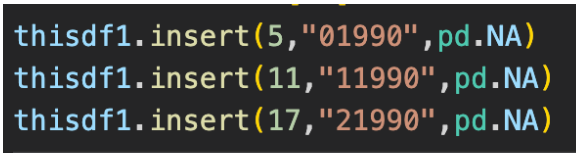


```
thisdf5 = thisdf3.merge(thisdf2,how="outer").merge(thisdf4,how="outer")
```

	Major area, region, country or area of destination	Notes	Type of data (a)	Percentage of the international migrant stock	Gender	Year
0	WORLD	NaN	NaN	100	both	1990
1	Developed regions	(b)	NaN	100	both	1990
2	Developing regions	(c)	NaN	100	both	1990
3	Least developed countries	(d)	NaN	100	both	1990
4	Less developed regions excluding least develop...	NaN	NaN	100	both	1990
...	...	...	...	...	...	...
4765	Samoa	NaN	B	49.908704	female	2015
4766	Tokelau	NaN	B	52.156057	female	2015
4767	Tonga	NaN	B	45.437096	female	2015
4768	Tuvalu	NaN	C	44.680851	female	2015
4769	Wallis and Futuna Islands	NaN	B	49.52615	female	2015

Table 5:

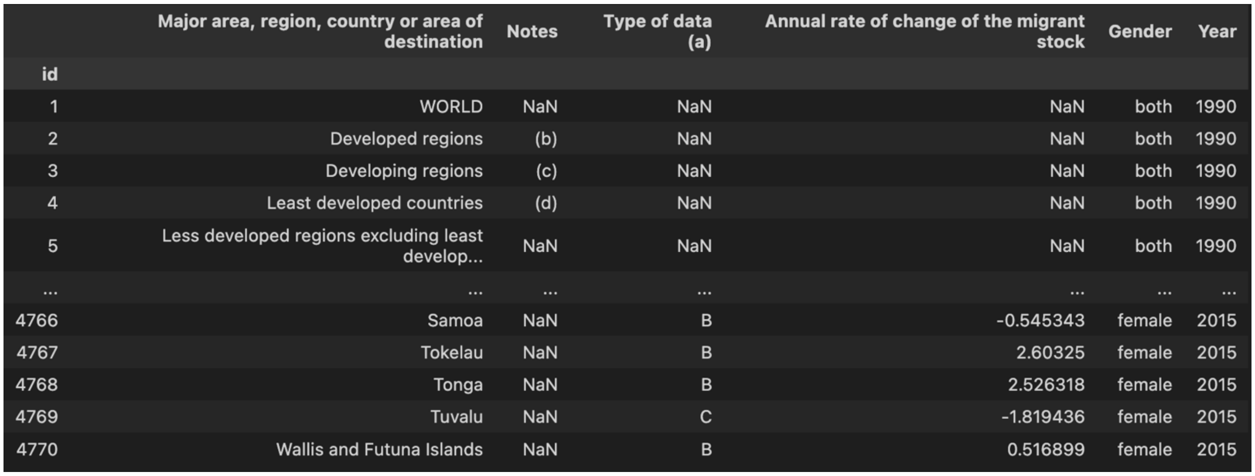
Table 5 is the annual rate of change of the migrant stock from 1995 to 2015. We do not have the 1990’s information. So, I put three columns of NA for 1990’s. This will help me when doing the final merge step.



```
thisdf1.insert(5,"01990",pd.NA)
thisdf1.insert(11,"11990",pd.NA)
thisdf1.insert(17,"21990",pd.NA)
```

Other steps are the same as Table 1.

Final data frame:

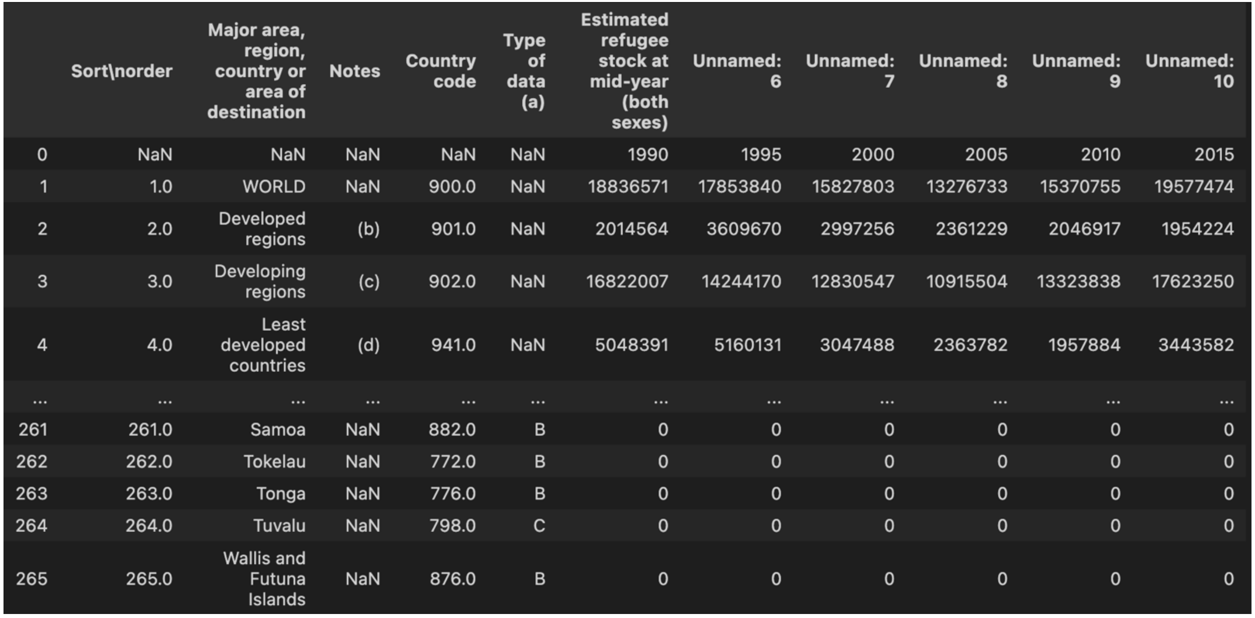


	Major area, region, country or area of destination	Notes	Type of data (a)	Annual rate of change of the migrant stock	Gender	Year
id						
1	WORLD	NaN	NaN	NaN	both	1990
2	Developed regions	(b)	NaN	NaN	both	1990
3	Developing regions	(c)	NaN	NaN	both	1990
4	Least developed countries	(d)	NaN	NaN	both	1990
5	Less developed regions excluding least develop...	NaN	NaN	NaN	both	1990
...	...	...	...	...	...	...
4766	Samoa	NaN	B	-0.545343	female	2015
4767	Tokelau	NaN	B	2.60325	female	2015
4768	Tonga	NaN	B	2.526318	female	2015
4769	Tuvalu	NaN	C	-1.819436	female	2015
4770	Wallis and Futuna Islands	NaN	B	0.516899	female	2015

Table 6:

Table 6 is different than other tables. It only contains both gender category. Also, it contains three different variables in the same table (Estimated refugee stock at mid-year, Refugees as a percentage of the international migrant stock and Annual rate of change of the refugee stock). I cannot use melt() function directly. Thus, I will first separate the data frame into three data frames. Clean each one and merge back together.

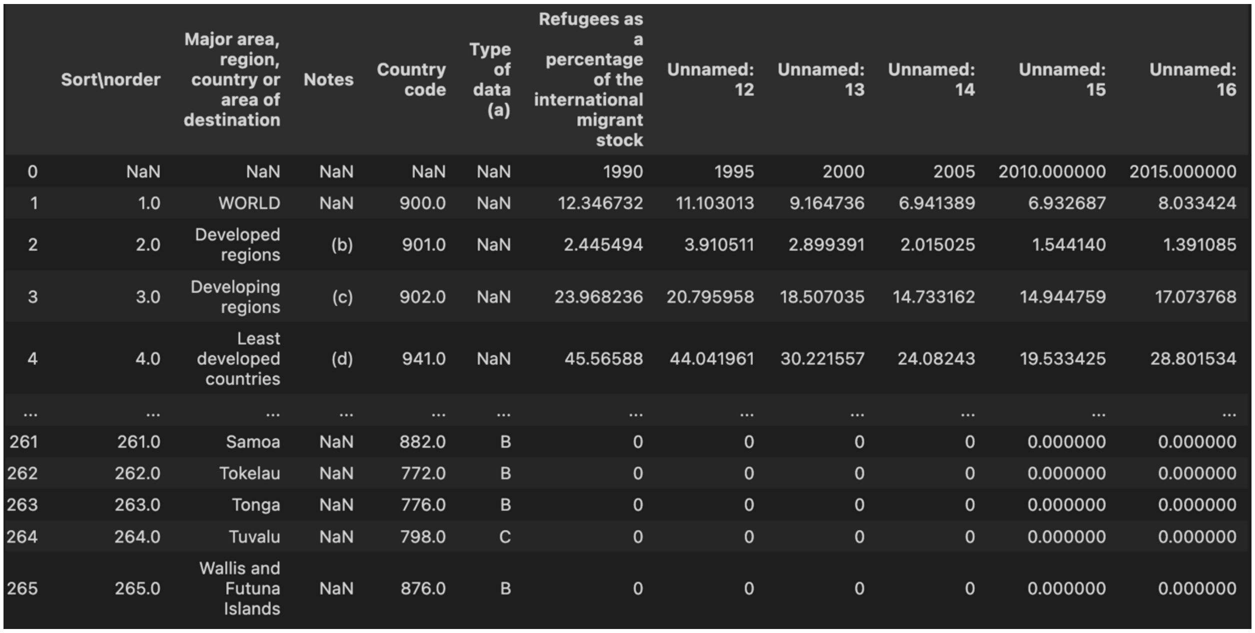
A picture of sub table 1:



	Sort\norder	Major area, region, country or area of destination	Notes	Country code	Type of data (a)	Estimated refugee stock at mid-year (both sexes)	Unnamed: 6	Unnamed: 7	Unnamed: 8	Unnamed: 9	Unnamed: 10
0	NaN	NaN	NaN	NaN	NaN	1990	1995	2000	2005	2010	2015
1	1.0	WORLD	NaN	900.0	NaN	18836571	17853840	15827803	13276733	15370755	19577474
2	2.0	Developed regions	(b)	901.0	NaN	2014564	3609670	2997256	2361229	2046917	1954224
3	3.0	Developing regions	(c)	902.0	NaN	16822007	14244170	12830547	10915504	13323838	17623250
4	4.0	Least developed countries	(d)	941.0	NaN	5048391	5160131	3047488	2363782	1957884	3443582
...	...	...	...	...	...	...	...	...	...	...	...
261	261.0	Samoa	NaN	882.0	B	0	0	0	0	0	0
262	262.0	Tokelau	NaN	772.0	B	0	0	0	0	0	0
263	263.0	Tonga	NaN	776.0	B	0	0	0	0	0	0
264	264.0	Tuvalu	NaN	798.0	C	0	0	0	0	0	0
265	265.0	Wallis and Futuna Islands	NaN	876.0	B	0	0	0	0	0	0

This table can be clean the same way as Table1.

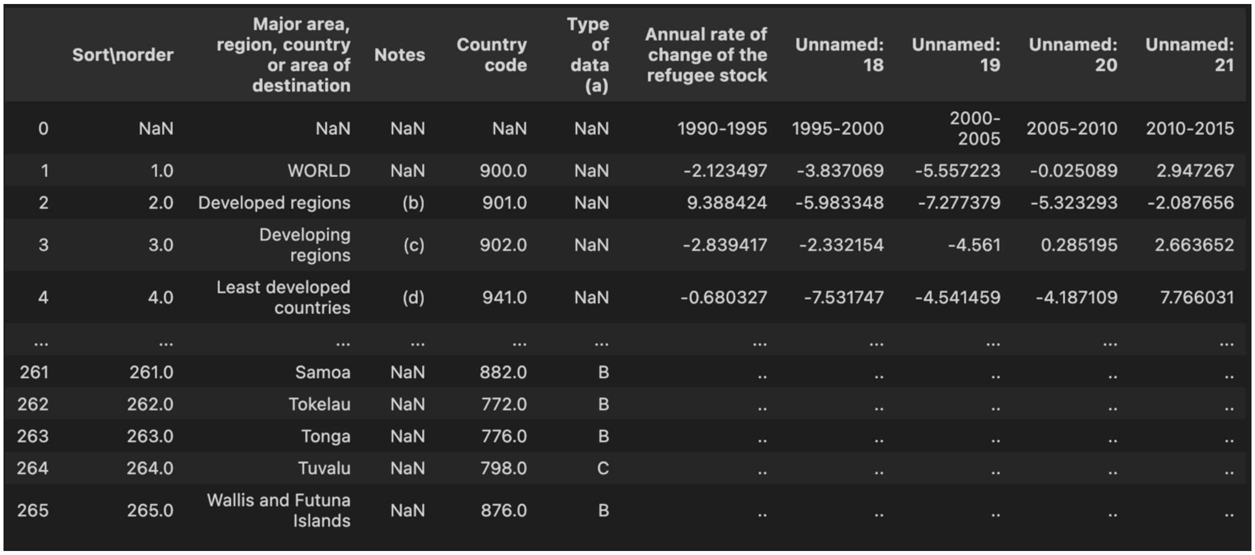
A picture of sub table 2:



	Sort\norder	Major area, region, country or area of destination	Notes	Country code	Type of data (a)	Refugees as a percentage of the international migrant stock	Unnamed: 12	Unnamed: 13	Unnamed: 14	Unnamed: 15	Unnamed: 16
0	NaN	NaN	NaN	NaN	NaN	1990	1995	2000	2005	2010.000000	2015.000000
1	1.0	WORLD	NaN	900.0	NaN	12.346732	11.103013	9.164736	6.941389	6.932687	8.033424
2	2.0	Developed regions	(b)	901.0	NaN	2.445494	3.910511	2.899391	2.015025	1.544140	1.391085
3	3.0	Developing regions	(c)	902.0	NaN	23.968236	20.795958	18.507035	14.733162	14.944759	17.073768
4	4.0	Least developed countries	(d)	941.0	NaN	45.56588	44.041961	30.221557	24.08243	19.533425	28.801534
...	...	...	...	...	...	...	...	...	...	...	...
261	261.0	Samoa	NaN	882.0	B	0	0	0	0	0.000000	0.000000
262	262.0	Tokelau	NaN	772.0	B	0	0	0	0	0.000000	0.000000
263	263.0	Tonga	NaN	776.0	B	0	0	0	0	0.000000	0.000000
264	264.0	Tuvalu	NaN	798.0	C	0	0	0	0	0.000000	0.000000
265	265.0	Wallis and Futuna Islands	NaN	876.0	B	0	0	0	0	0.000000	0.000000

This table can be clean the same way as Table 1.

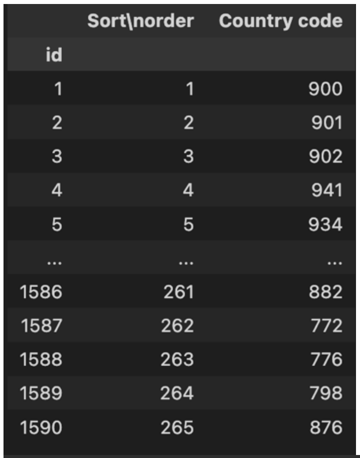
A picture of sub table 3:



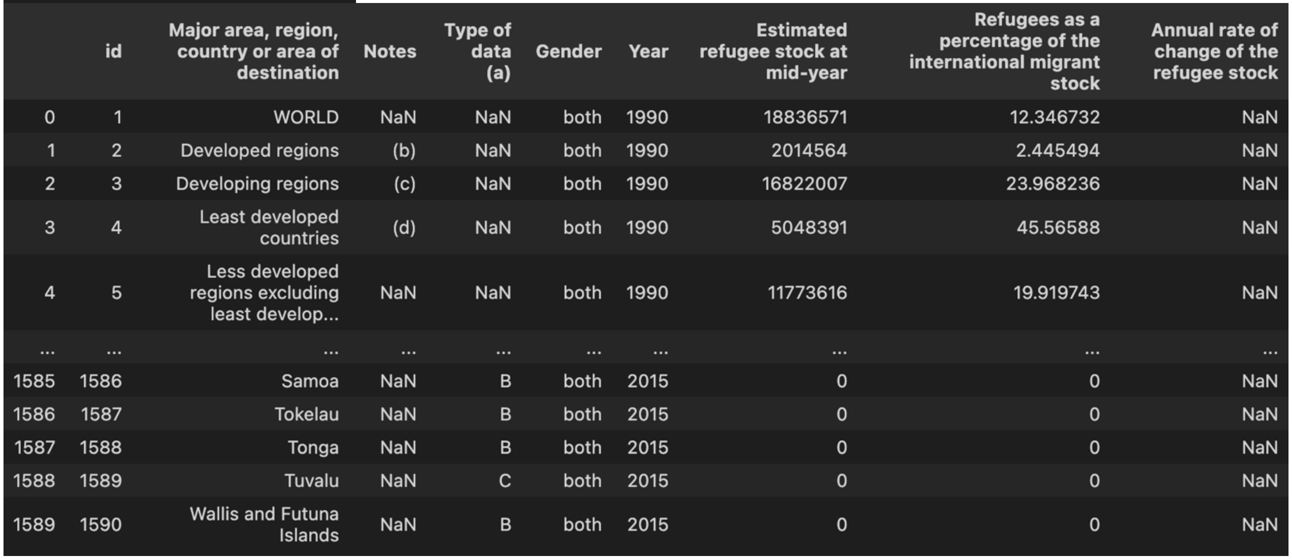
	Sort\norder	Major area, region, country or area of destination	Notes	Country code	Type of data (a)	Annual rate of change of the refugee stock	Unnamed: 18	Unnamed: 19	Unnamed: 20	Unnamed: 21
0	NaN	NaN	NaN	NaN	NaN	1990-1995	1995-2000	2000-2005	2005-2010	2010-2015
1	1.0	WORLD	NaN	900.0	NaN	-2.123497	-3.837069	-5.557223	-0.025089	2.947267
2	2.0	Developed regions	(b)	901.0	NaN	9.388424	-5.983348	-7.277379	-5.323293	-2.087656
3	3.0	Developing regions	(c)	902.0	NaN	-2.839417	-2.332154	-4.561	0.285195	2.663652
4	4.0	Least developed countries	(d)	941.0	NaN	-0.680327	-7.531747	-4.541459	-4.187109	7.766031
...	...	...	...	...	...	...	...	...	...	...
261	261.0	Samoa	NaN	882.0	B	..	..	..	..	..
262	262.0	Tokelau	NaN	772.0	B	..	..	..	..	..
263	263.0	Tonga	NaN	776.0	B	..	..	..	..	..
264	264.0	Tuvalu	NaN	798.0	C	..	..	..	..	..
265	265.0	Wallis and Futuna Islands	NaN	876.0	B	..	..	..	..	..

This table can be clean the same way as Table 5.

Finally merge as one table. (It also has a sub table for different data type)



	Sort\norder	Country code
id		
1	1	900
2	2	901
3	3	902
4	4	941
5	5	934
...	...	...
1586	261	882
1587	262	772
1588	263	776
1589	264	798
1590	265	876

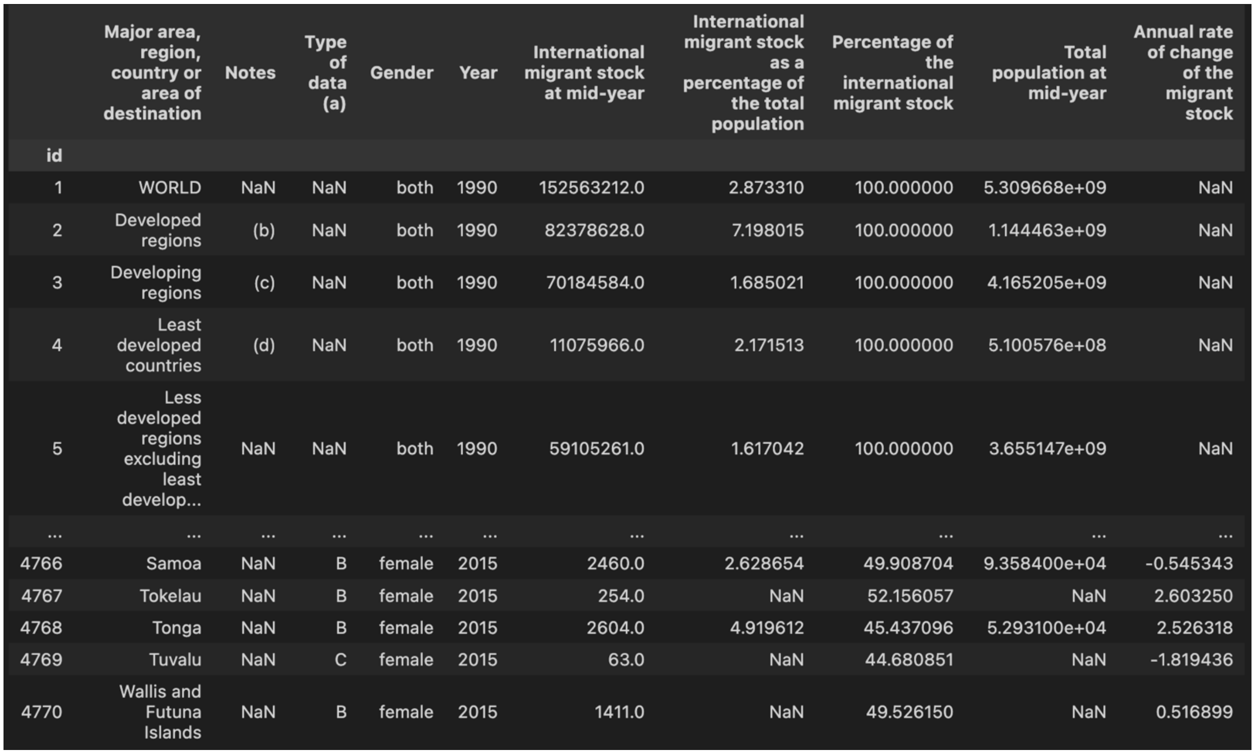


	id	Major area, region, country or area of destination	Notes	Type of data (a)	Gender	Year	Estimated refugee stock at mid-year	Refugees as a percentage of the international migrant stock	Annual rate of change of the refugee stock
0	1	WORLD	NaN	NaN	both	1990	18836571	12.346732	NaN
1	2	Developed regions	(b)	NaN	both	1990	2014564	2.445494	NaN
2	3	Developing regions	(c)	NaN	both	1990	16822007	23.968236	NaN
3	4	Least developed countries	(d)	NaN	both	1990	5048391	45.56588	NaN
4	5	Less developed regions excluding least develop...	NaN	NaN	both	1990	11773616	19.919743	NaN
...	...	...	...	...	...	...	...	...	...
1585	1586	Samoa	NaN	B	both	2015	0	0	NaN
1586	1587	Tokelau	NaN	B	both	2015	0	0	NaN
1587	1588	Tonga	NaN	B	both	2015	0	0	NaN
1588	1589	Tuvalu	NaN	C	both	2015	0	0	NaN
1589	1590	Wallis and Futuna Islands	NaN	B	both	2015	0	0	NaN

Final merge table:

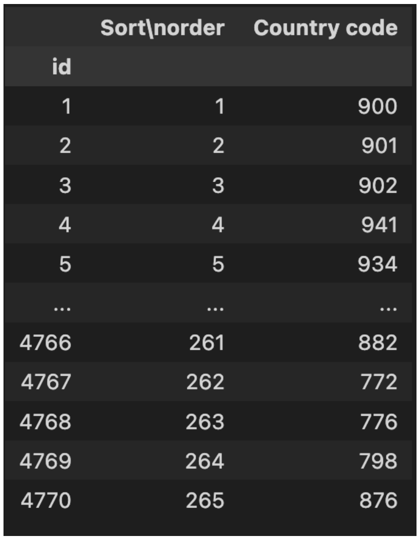
I plan to merge Table 1,2,3,4,5 together because they all a part for "Trends in International Migrant Stock". For Table 6, it is a little bit different because it only contains the information of both gender category. It is hard to merge into the final main table. And gender and year columns can make it more easier for the further analysis.

Final main table:



	Major area, region, country or area of destination	Notes	Type of data (a)	Gender	Year	International migrant stock at mid-year	International migrant stock as a percentage of the total population	Percentage of the international migrant stock	Total population at mid-year	Annual rate of change of the migrant stock
id										
1	WORLD	NaN	NaN	both	1990	152563212.0	2.873310	100.000000	5.309668e+09	NaN
2	Developed regions	(b)	NaN	both	1990	82378628.0	7.198015	100.000000	1.144463e+09	NaN
3	Developing regions	(c)	NaN	both	1990	70184584.0	1.685021	100.000000	4.165205e+09	NaN
4	Least developed countries	(d)	NaN	both	1990	11075966.0	2.171513	100.000000	5.100576e+08	NaN
5	Less developed regions excluding least develop...	NaN	NaN	both	1990	59105261.0	1.617042	100.000000	3.655147e+09	NaN
...	...	...	...	...	...	...	...	...	...	...
4766	Samoa	NaN	B	female	2015	2460.0	2.628654	49.908704	9.358400e+04	-0.545343
4767	Tokelau	NaN	B	female	2015	254.0	NaN	52.156057	NaN	2.603250
4768	Tonga	NaN	B	female	2015	2604.0	4.919612	45.437096	5.293100e+04	2.526318
4769	Tuvalu	NaN	C	female	2015	63.0	NaN	44.680851	NaN	-1.819436
4770	Wallis and Futuna Islands	NaN	B	female	2015	1411.0	NaN	49.526150	NaN	0.516899

Final sub table: (contain all the different data type according to principle 4)



	Sort\norder	Country code
id		
1	1	900
2	2	901
3	3	902
4	4	941
5	5	934
...	...	...
4766	261	882
4767	262	772
4768	263	776
4769	264	798
4770	265	876

Conclusion:

The final result contains four data frames: Main table, sub table, refugee’s main table and refugee’s sub table. Main table conclude all the information that from table_1,2,3,4,5. It is easier to get information from main table than 5 separate tables. All tables have used principle 1,2,3 in this case.

List of data frames:

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

Table 3 - International migrant stock as a percentage of the total population by sex and by major area, region, country or area, 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

Mergetable – the result combination of Table 1,2,3,4,5