Course: inf1340
Name: Jiawei Liu
Student #: 1004803028

Introduction:

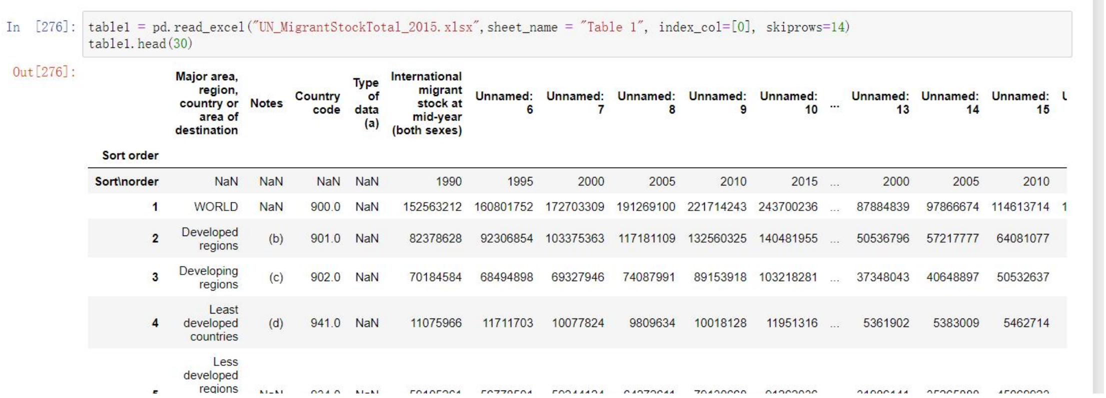
The midterm assignment is a project of tidy data, in other words, we need to clean the complicate document post by United Nations. The document includes 6 wide tables, each table has its own topic, but some are similar. I try to combine the related table and reduce the columns, melt the wide table to long table with more rows. After data cleaning, I will use tidy data principles to test the organized table.

Process

My opinion is to clean the table one by one firstly and then to seek if there is a possible way to merge some of tables.

Table 1 - International migrant stock at mid-year by sex and by major area, region, country or area, 1990-2015

The first thing to do was the preparation, I imported np and pandas, then checked the work directory. After that, reading the xlsx file from wd since I already downloaded to my wd. I discovered that the first 14 lines were useless, so I skipped them.

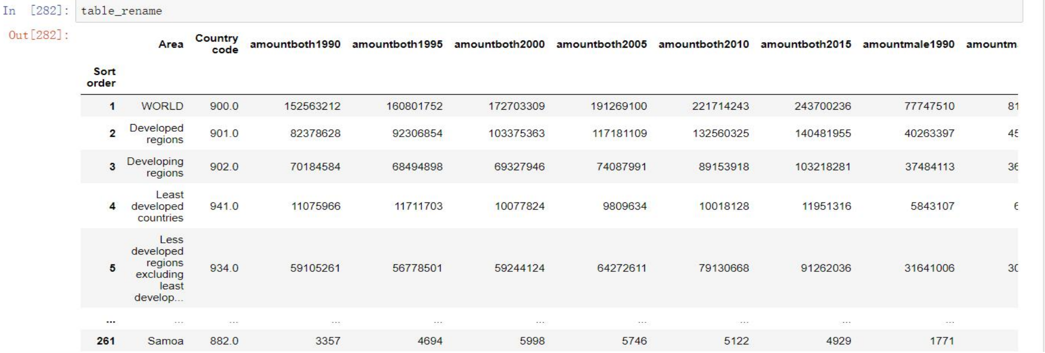


```
In  [276]: table1 = pd.read_excel("UN_MigrantStockTotal_2015.xlsx",sheet_name = "Table 1", index_col=[0], skiprows=14)
           table1.head(30)
```

Out[276]:

	Major area, region, country or area of destination	Notes	Country code	Type of data (a)	International migrant stock at mid-year (both sexes)	Unnamed: 6	Unnamed: 7	Unnamed: 8	Unnamed: 9	Unnamed: 10	...	Unnamed: 13	Unnamed: 14	Unnamed: 15	U
Sort order															
Sort\norder	NaN	NaN	NaN	NaN	1990	1995	2000	2005	2010	2015	...	2000	2005	2010	
1	WORLD	NaN	900.0	NaN	152563212	160801752	172703309	191269100	221714243	243700236	...	87884839	97866674	114613714	1
2	Developed regions	(b)	901.0	NaN	82378628	92306854	103375363	117181109	132560325	140481955	...	50536796	57217777	64081077	
3	Developing regions	(c)	902.0	NaN	70184584	68494898	69327946	74087991	89153918	103218281	...	37348043	40648897	50532637	
4	Least developed countries	(d)	941.0	NaN	11075966	11711703	10077824	9809634	10018128	11951316	...	5361902	5383009	5462714	
	Less developed regions														

The column names was weird, too many unnamed columns, therefore, I renamed all the names, meanwhile combined all the year in the first row to the names. Next, I dropped the first row. Besides, some columns are not necessary, for example "Notes" and "Type of data (a)", might prevent users from getting information. Hence, I dropped them.

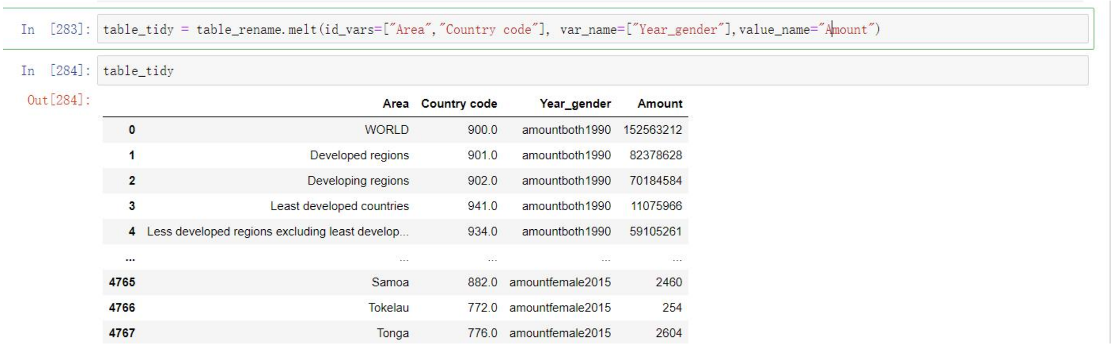


```
In  [282]: table_rename
```

Out[282]:

	Area	Country code	amountboth1990	amountboth1995	amountboth2000	amountboth2005	amountboth2010	amountboth2015	amountmale1990	amountm
Sort order										
1	WORLD	900.0	152563212	160801752	172703309	191269100	221714243	243700236	77747510	81
2	Developed regions	901.0	82378628	92306854	103375363	117181109	132560325	140481955	40263397	45
3	Developing regions	902.0	70184584	68494898	69327946	74087991	89153918	103218281	37484113	36
4	Least developed countries	941.0	11075966	11711703	10077824	9809634	10018128	11951316	5843107	6
5	Less developed regions excluding least develop...	934.0	59105261	56778501	59244124	64272611	79130668	91262036	31641006	30
...	...	...	...	...	...	...	...	...	...	
261	Samoa	882.0	3357	4694	5998	5746	5122	4929	1771	

The next step is melting the wide data to long format since wide format is difficult to do the data analysis, for example, creating pivot table and data visualization. The number of columns now is 4, they are “Area”, “country code”, “Year_gender” and “Amount”



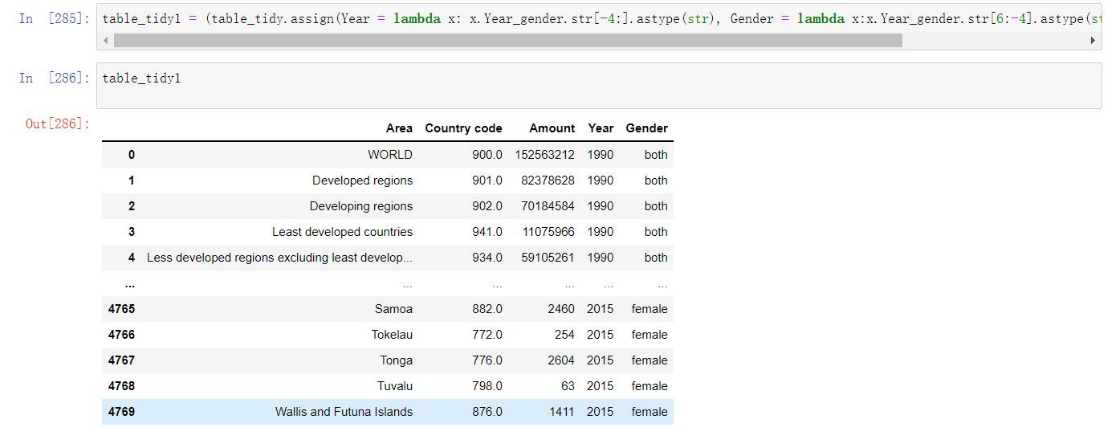
```
In  [283]: table_tidy = table_rename.melt(id_vars=["Area","Country code"], var_name=["Year_gender"],value_name="Amount")
```

```
In  [284]: table_tidy
```

Out[284]:

	Area	Country code	Year_gender	Amount
0	WORLD	900.0	amountboth1990	152563212
1	Developed regions	901.0	amountboth1990	82378628
2	Developing regions	902.0	amountboth1990	70184584
3	Least developed countries	941.0	amountboth1990	11075966
4	Less developed regions excluding least develop...	934.0	amountboth1990	59105261
...	...	...	...	...
4765	Samoa	882.0	amountfemale2015	2460
4766	Tokelau	772.0	amountfemale2015	254
4767	Tonga	776.0	amountfemale2015	2604

As we can see “Year” and “gender” are two variables but they are in the same columns, which violate the tidy data principle2 (#tidy data principle #2: each column needs to consist of one and only one variable). As a result, I split them.



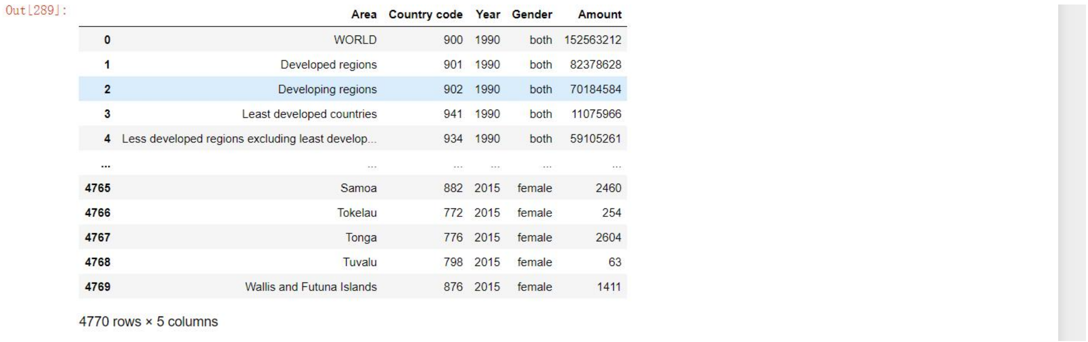
```
In  [285]: table_tidy1 = (table_tidy.assign(Year = lambda x: x.Year_gender.str[-4:].astype(str), Gender = lambda x:x.Year_gender.str[6:-4].astype(st
```

```
In  [286]: table_tidy1
```

Out[286]:

	Area	Country code	Amount	Year	Gender
0	WORLD	900.0	152563212	1990	both
1	Developed regions	901.0	82378628	1990	both
2	Developing regions	902.0	70184584	1990	both
3	Least developed countries	941.0	11075966	1990	both
4	Less developed regions excluding least develop...	934.0	59105261	1990	both
...	...	...	...	...	...
4765	Samoa	882.0	2460	2015	female
4766	Tokelau	772.0	254	2015	female
4767	Tonga	776.0	2604	2015	female
4768	Tuvalu	798.0	63	2015	female
4769	Wallis and Futuna Islands	876.0	1411	2015	female

It is a long format table now with only five columns, which is good for data analysis. To make it easier to read, I switched the float to int under “country code” and also rearrange the order of the columns. The “Amount” was moved to the last column. The following is the final vision for the table1.



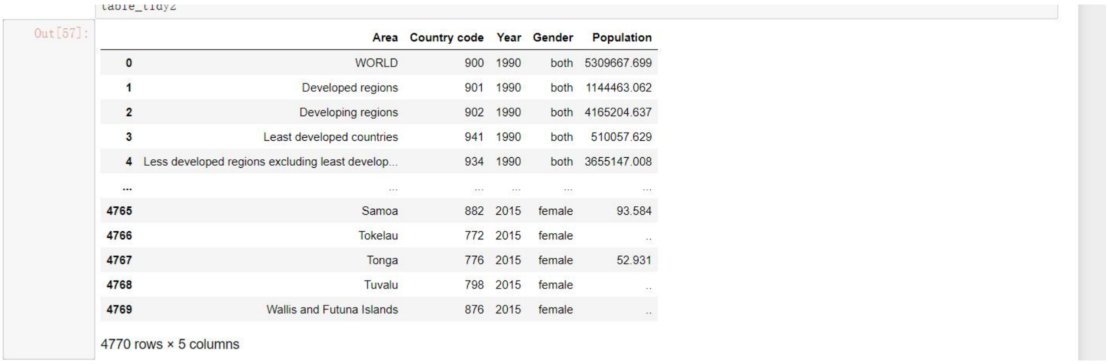
Out[289]:

	Area	Country code	Year	Gender	Amount
0	WORLD	900	1990	both	152563212
1	Developed regions	901	1990	both	82378628
2	Developing regions	902	1990	both	70184584
3	Least developed countries	941	1990	both	11075966
4	Less developed regions excluding least develop...	934	1990	both	59105261
...	...	...	...	...	...
4765	Samoa	882	2015	female	2460
4766	Tokelau	772	2015	female	254
4767	Tonga	776	2015	female	2604
4768	Tuvalu	798	2015	female	63
4769	Wallis and Futuna Islands	876	2015	female	1411

4770 rows × 5 columns

Table 2 - Total population at mid-year by sex and by major area, region, country or area, 1990-2015 (thousands)

The step of cleaning table 2 is same to the table1. This is the final version for table2.



table_tidy2

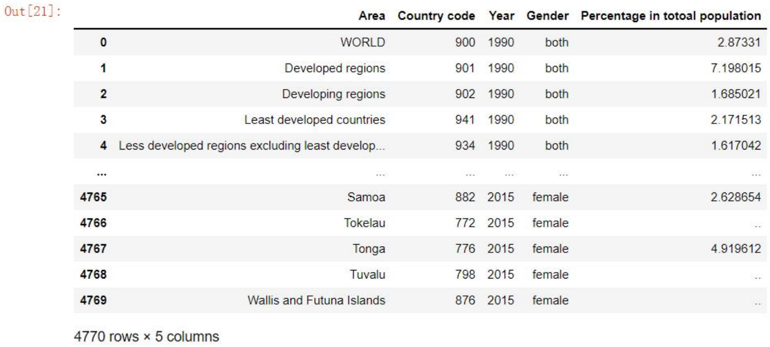
Out[57]:

	Area	Country code	Year	Gender	Population
0	WORLD	900	1990	both	5309667.699
1	Developed regions	901	1990	both	1144463.062
2	Developing regions	902	1990	both	4165204.637
3	Least developed countries	941	1990	both	510057.629
4	Less developed regions excluding least develop...	934	1990	both	3655147.008
...	...	...	...	...	...
4765	Samoa	882	2015	female	93.584
4766	Tokelau	772	2015	female	..
4767	Tonga	776	2015	female	52.931
4768	Tuvalu	798	2015	female	..
4769	Wallis and Futuna Islands	876	2015	female	..

4770 rows × 5 columns

Table 3 - International migrant stock as a percentage of the total population by sex and by major area, region, country or area, 1990-2015

The step of cleaning table3 is same to the table1. This is the final version for table3.



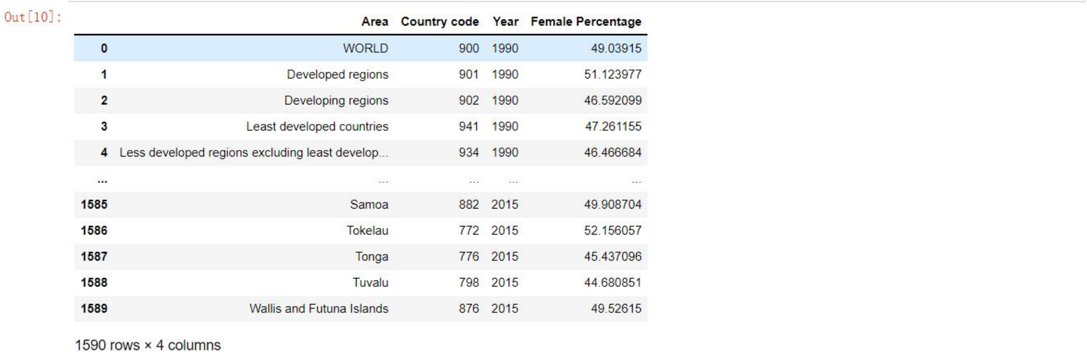
Out[21]:

	Area	Country code	Year	Gender	Percentage in totoal population
0	WORLD	900	1990	both	2.87331
1	Developed regions	901	1990	both	7.198015
2	Developing regions	902	1990	both	1.685021
3	Least developed countries	941	1990	both	2.171513
4	Less developed regions excluding least develop...	934	1990	both	1.617042
...	...	...	...	...	...
4765	Samoa	882	2015	female	2.628654
4766	Tokelau	772	2015	female	..
4767	Tonga	776	2015	female	4.919612
4768	Tuvalu	798	2015	female	..
4769	Wallis and Futuna Islands	876	2015	female	..

4770 rows × 5 columns

Table 4 - Female migrants as a percentage of the international migrant stock by major area, region, country or area, 1990-2015

Also the same idea to clean.



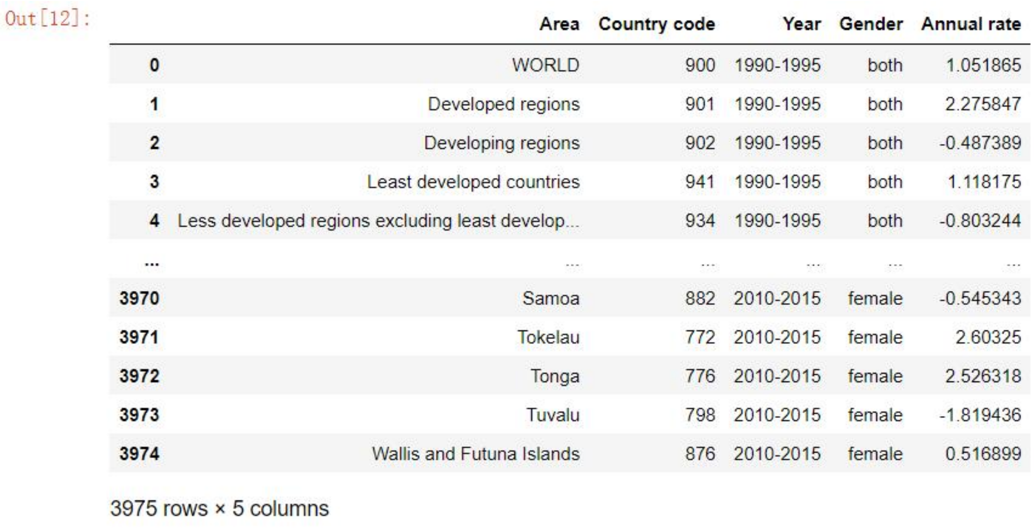
Out[10]:

	Area	Country code	Year	Female Percentage
0	WORLD	900	1990	49.03915
1	Developed regions	901	1990	51.123977
2	Developing regions	902	1990	46.592099
3	Least developed countries	941	1990	47.261155
4	Less developed regions excluding least develop...	934	1990	46.466684
...	...	...	...	...
1585	Samoa	882	2015	49.908704
1586	Tokelau	772	2015	52.156057
1587	Tonga	776	2015	45.437096
1588	Tuvalu	798	2015	44.680851
1589	Wallis and Futuna Islands	876	2015	49.52615

1590 rows × 4 columns

Table 5 - Annual rate of change of the migrant stock by sex and by major area, region, country or area, 1990-2015 (percentage)

Same idea.



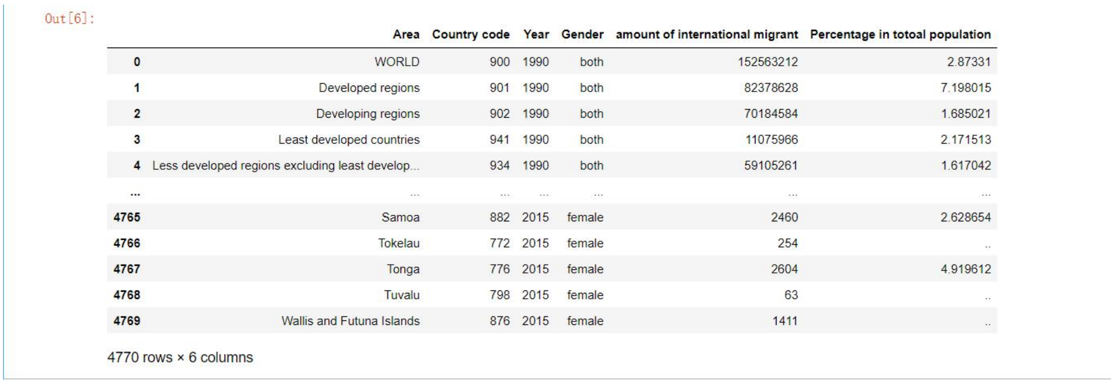
Out[12]:

	Area	Country code	Year	Gender	Annual rate
0	WORLD	900	1990-1995	both	1.051865
1	Developed regions	901	1990-1995	both	2.275847
2	Developing regions	902	1990-1995	both	-0.487389
3	Least developed countries	941	1990-1995	both	1.118175
4	Less developed regions excluding least develop...	934	1990-1995	both	-0.803244
...	...	...	...	...	...
3970	Samoa	882	2010-2015	female	-0.545343
3971	Tokelau	772	2010-2015	female	2.60325
3972	Tonga	776	2010-2015	female	2.526318
3973	Tuvalu	798	2010-2015	female	-1.819436
3974	Wallis and Futuna Islands	876	2010-2015	female	0.516899

3975 rows × 5 columns

Merging table 1 and table3

The first five table all talks about international migrants, table 1,2,3,5 are highly related. Therefore, I tried to combine that 4 tables. However, I discovered that the number of rows of table4 was 3975, while the other three were 4770. The reason is the value of year is different, the table4 shows data about annual rate, therefore, under the variable "Year", it is a period of time. That was the challenge I faced during merging and I did not find a way to merge table4 with table 1,2,3 successfully. Unfortunately, when I merged table 1,2,3, I found that table two was about total population, if combine with table1 and 3, might violet tidy data principle5 (a single observational unit must be in 1 table.) It seems that they are two kind of observational units, the total population and immigrants. Besides, table three is about the percentage of international immigrants in the total population, it is unnecessary to have an extra column to show the amount of total population at that year. The following table is the one after merging.



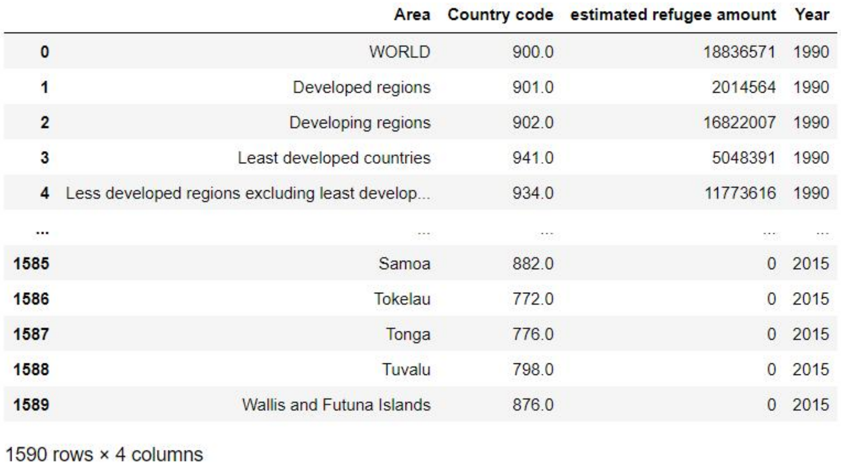
Out[6]:

	Area	Country code	Year	Gender	amount of international migrant	Percentage in totoal population
0	WORLD	900	1990	both	152563212	2.87331
1	Developed regions	901	1990	both	82378628	7.198015
2	Developing regions	902	1990	both	70184584	1.685021
3	Least developed countries	941	1990	both	11075966	2.171513
4	Less developed regions excluding least develop...	934	1990	both	59105261	1.617042
...	...	...	...	...	...	...
4765	Samoa	882	2015	female	2460	2.628654
4766	Tokelau	772	2015	female	254	..
4767	Tonga	776	2015	female	2604	4.919612
4768	Tuvalu	798	2015	female	63	..
4769	Wallis and Futuna Islands	876	2015	female	1411	..

4770 rows × 6 columns

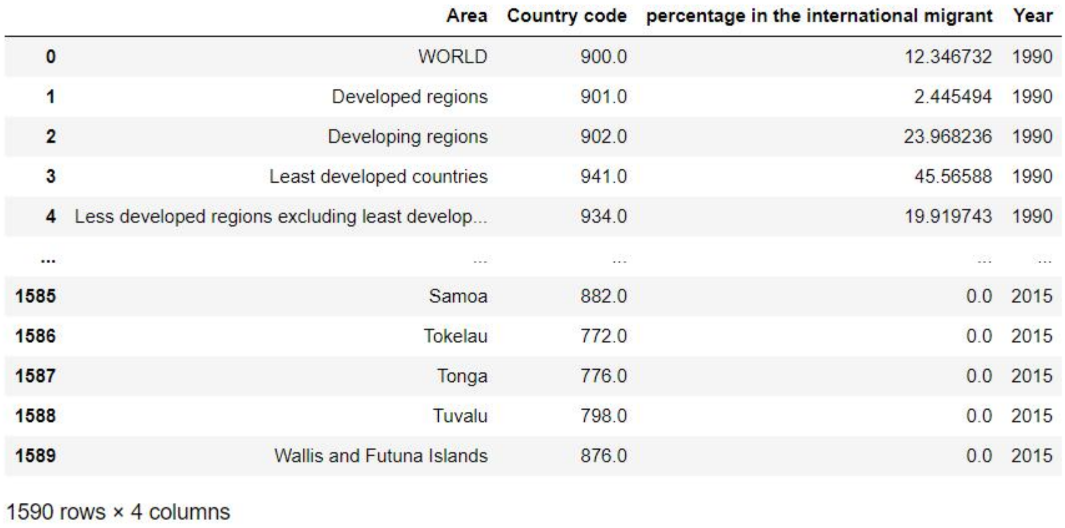
Table 6 - Estimated refugee stock at mid-year by major area, region, country or area, 1990-2015

After reading the table6, the first thing I did was separate it to three tables, and then did the initial clean for each table, which was similar to table1 process.



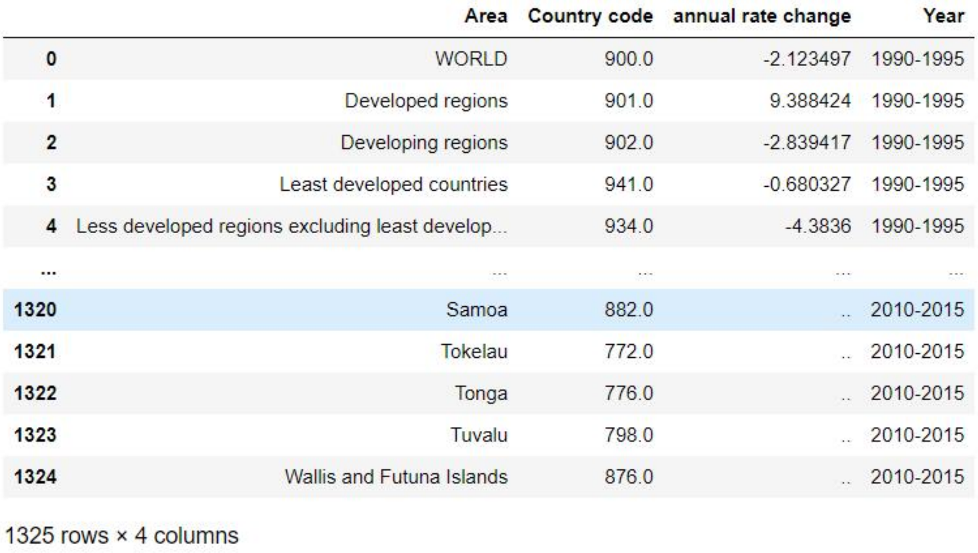
	Area	Country code	estimated refugee amount	Year
0	WORLD	900.0	18836571	1990
1	Developed regions	901.0	2014564	1990
2	Developing regions	902.0	16822007	1990
3	Least developed countries	941.0	5048391	1990
4	Less developed regions excluding least develop...	934.0	11773616	1990
...	...	...	...	...
1585	Samoa	882.0	0	2015
1586	Tokelau	772.0	0	2015
1587	Tonga	776.0	0	2015
1588	Tuvalu	798.0	0	2015
1589	Wallis and Futuna Islands	876.0	0	2015

1590 rows × 4 columns



	Area	Country code	percentage in the international migrant	Year
0	WORLD	900.0	12.346732	1990
1	Developed regions	901.0	2.445494	1990
2	Developing regions	902.0	23.968236	1990
3	Least developed countries	941.0	45.56588	1990
4	Less developed regions excluding least develop...	934.0	19.919743	1990
...	...	...	...	...
1585	Samoa	882.0	0.0	2015
1586	Tokelau	772.0	0.0	2015
1587	Tonga	776.0	0.0	2015
1588	Tuvalu	798.0	0.0	2015
1589	Wallis and Futuna Islands	876.0	0.0	2015

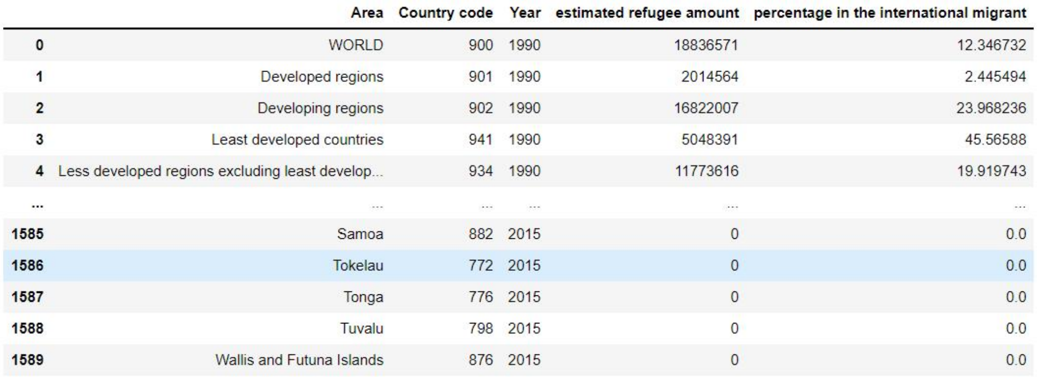
1590 rows × 4 columns



	Area	Country code	annual rate change	Year
0	WORLD	900.0	-2.123497	1990-1995
1	Developed regions	901.0	9.388424	1990-1995
2	Developing regions	902.0	-2.839417	1990-1995
3	Least developed countries	941.0	-0.680327	1990-1995
4	Less developed regions excluding least develop...	934.0	-4.3836	1990-1995
...	...	...	...	...
1320	Samoa	882.0	..	2010-2015
1321	Tokelau	772.0	..	2010-2015
1322	Tonga	776.0	..	2010-2015
1323	Tuvalu	798.0	..	2010-2015
1324	Wallis and Futuna Islands	876.0	..	2010-2015

1325 rows × 4 columns

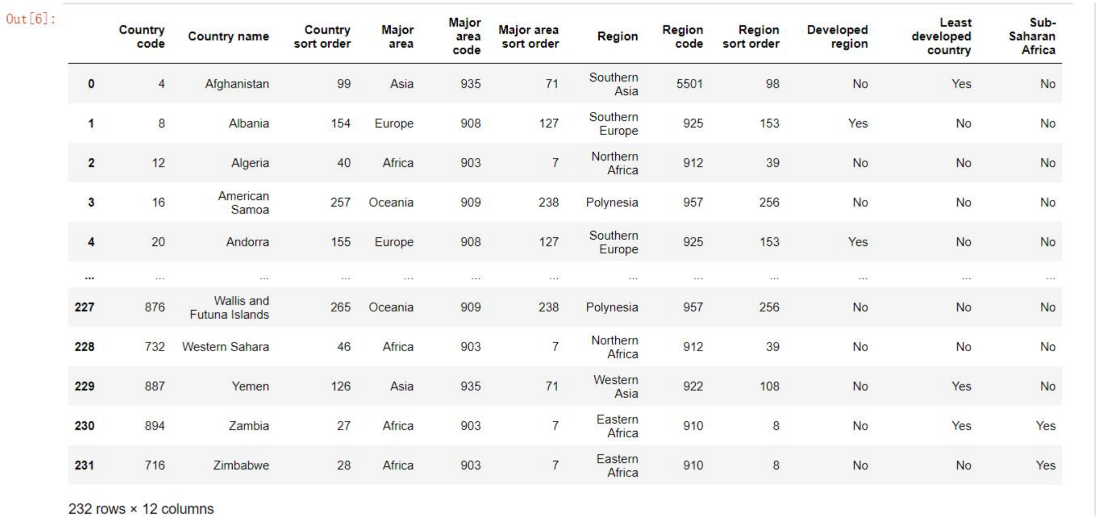
Next, I tried to merge this three, but the same problem I met again, the year for annual rate change was a period. I have no good idea right now to solve that. Therefore, I only merged two of the table, just like I did on merging table1 and table 3 before.



	Area	Country code	Year	estimated refugee amount	percentage in the international migrant
0	WORLD	900	1990	18836571	12.346732
1	Developed regions	901	1990	2014564	2.445494
2	Developing regions	902	1990	16822007	23.968236
3	Least developed countries	941	1990	5048391	45.56588
4	Less developed regions excluding least develop...	934	1990	11773616	19.919743
...	...	...	...	...	...
1585	Samoa	882	2015	0	0.0
1586	Tokelau	772	2015	0	0.0
1587	Tonga	776	2015	0	0.0
1588	Tuvalu	798	2015	0	0.0
1589	Wallis and Futuna Islands	876	2015	0	0.0

Table ANNEX. Classification of countries and areas by major area and region

I just renamed some columns, since the original one confused me. There are some columns have the same name in the original table, difficult for extracting data and analyzing later. The following one is the renamed table Annex.



Out[6]:

	Country code	Country name	Country sort order	Major area	Major area code	Major area sort order	Region	Region code	Region sort order	Developed region	Least developed country	Sub-Saharan Africa
0	4	Afghanistan	99	Asia	935	71	Southern Asia	5501	98	No	Yes	No
1	8	Albania	154	Europe	908	127	Southern Europe	925	153	Yes	No	No
2	12	Algeria	40	Africa	903	7	Northern Africa	912	39	No	No	No
3	16	American Samoa	257	Oceania	909	238	Polynesia	957	256	No	No	No
4	20	Andorra	155	Europe	908	127	Southern Europe	925	153	Yes	No	No
...	...	...	...	...	...	...	...	...	...	...	...	...
227	876	Wallis and Futuna Islands	265	Oceania	909	238	Polynesia	957	256	No	No	No
228	732	Western Sahara	46	Africa	903	7	Northern Africa	912	39	No	No	No
229	887	Yemen	126	Asia	935	71	Western Asia	922	108	No	Yes	No
230	894	Zambia	27	Africa	903	7	Eastern Africa	910	8	No	Yes	Yes
231	716	Zimbabwe	28	Africa	903	7	Eastern Africa	910	8	No	No	Yes

232 rows × 12 columns

Tidy principles check:

#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 units must be in 1 table

The following table is the result of checking each table's final version.

	Pinciple1	Pinciple2	Pinciple3	Pinciple4	Pinciple5
Table1	√	√	√	? √	√
Table2	√	√	√	? √	√
Table3	√	√	√	? √	√
Table4	√	√	√	? √	√
Table5	√	√	√	? √	√
Merging table1,3	√	√	√	? √	√
Table6	√	√	√	? √	√
Table annex	√	√	√	√	√

As we can see from the principles check table, there are many "?" under principle4. Because most of the table have some missing values, they are represented by "..", I am not sure whether the principle4 accept that. Accept the missing value, other values are the same data type. Deleting the ".." for each table might be better, and then the it will be printed as "NaN"

Conclusion:

For this project, I tried my best to clean the table by using tidy data principles though there are some problems that I still could not solve. I learned that “practice makes perfect”, doing practice to understand the code is quite important. Data cleaning is usually conducted before data analysis, and after the reading week, we will learn how to do data virtualization. I am now on the way to be a data analyst.