UN_MigrantStockTotal_2015 Data Cleaning

Name: Danni Yang (student# 1009535558)

Introduction

This UN_MigrantStockTotal_2015 dataset collects migrants stock numbers, gender percentage, and annual changing rates among the year of 1990 to 2015. There are 9 sub tables in the whole dataset which includes 6 datasets, a content table, a notes table, and a classification table.

The main purpose of this project is to use python to catch the messy data problems and clean the 6 datasets as per the "tidy data" principles we learned from the textbook.

Methods

The three main principles of tidy data are that: each variable forms a column, each observation forms a row, and each type of observational unit forms a table (Wickham 2014). However, messy problems are various much more than these 3 principles.

Firstly, the data cleaning process start with identifying data we need to deal with and remove the irrelevant information. Thus, I removed introduction text from row 1 to row 15 as well as column "Notes", "Country code" and "Type of data (a)".

Secondly, I use "df.rename()" function to rename all the headers in a "gender letter + year" format. That would make it easier for the column split in the next step.

Principle 1

Take table 1 as an example, gender data and year data are formed in one column. Thus we use "df.melt()" function to pivot the table in a better structure. Set sort order and area as "id_vars" and set the mixed column as variable, the international migrant stock as the value. Then I use "df.assign()" function to split the mixed information column to 2 independent columns.

Principle 2

Since each observation should only form a row. The UN table does well in this area that, there is no row break this principle.

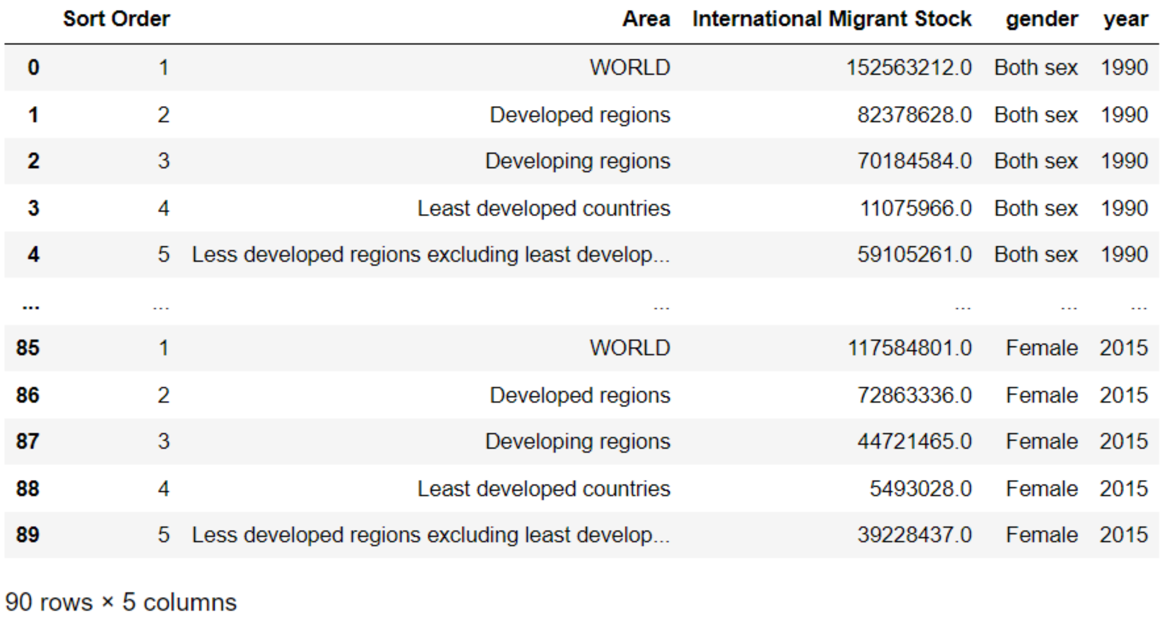
Principle 3

I found the most challenge thing is to identify observational units. Take table 1 as an example and then we can find different types of observation units formed altogether. Different types of data overlapped thus I decide to use "df.iloc()" function to locate specific rows and columns and divide the whole table into 3 sub tables. The base of classification for the first sub data set is "region developing status", and the rest of the data set classification base are "regions" and "countries".

During the data cleaning process, another big problem I find is the missing data. I use "df.replace()" function to change the string ".." to nmpy "NaN" and then use "df.fillna()" function to change "NaN" to value "0". All the missing data are due to the lack of data collection which can not be filled by connecting sub tables and calculating. However, directly removing may cause the loss of the related important data. Besides, I believe for future statics calculations or mathematic analysis, value 0 could be better to use than the "NaN" or "..".

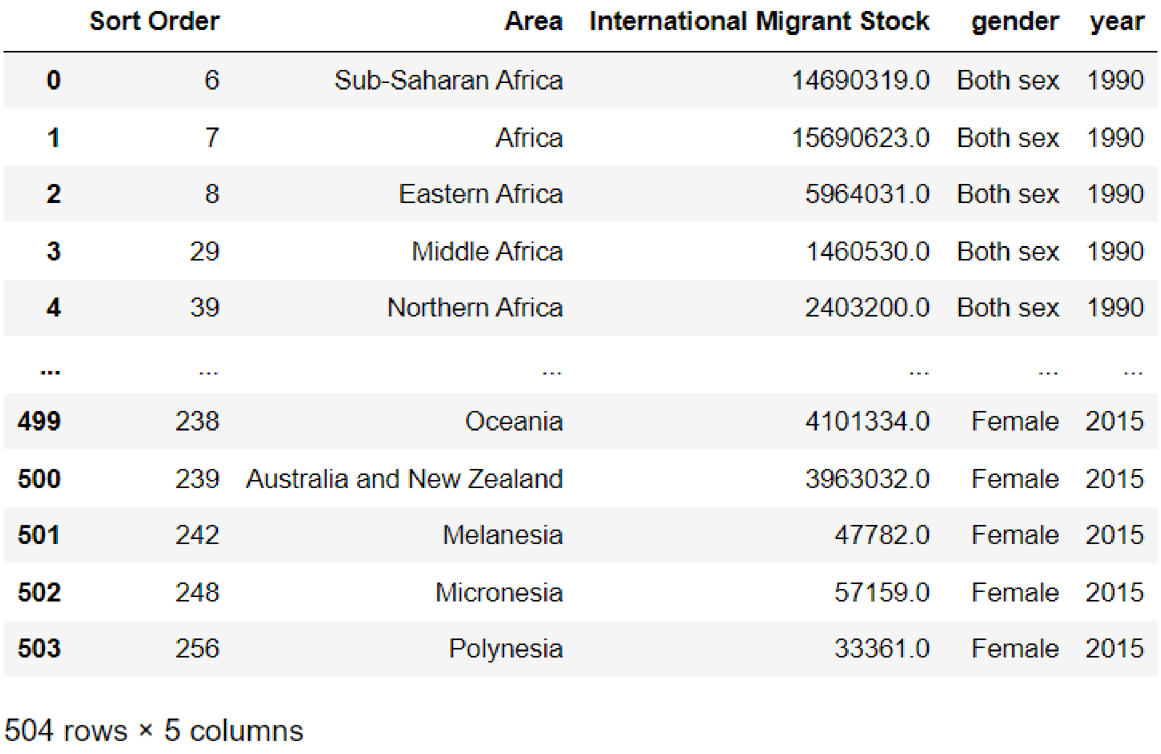
Results

The final tidy data version of the original excels file UN_MigrantStockTotal_2015 is shown as following:



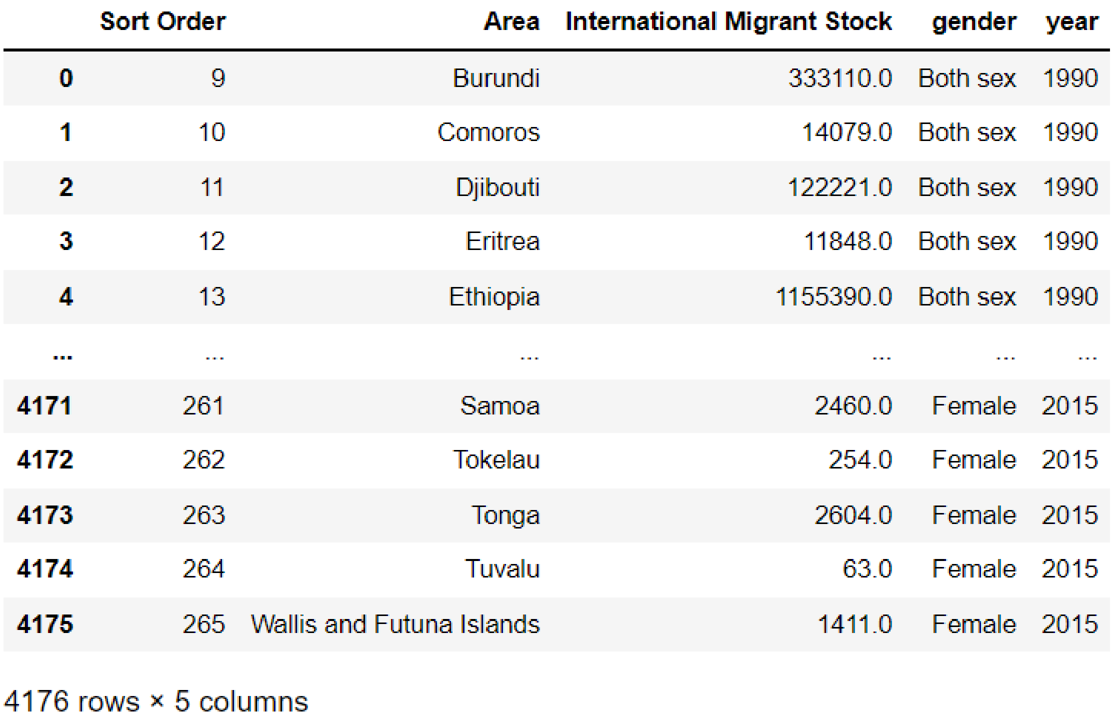
	Sort Order	Area	International Migrant Stock	gender	year
0	1	WORLD	152563212.0	Both sex	1990
1	2	Developed regions	82378628.0	Both sex	1990
2	3	Developing regions	70184584.0	Both sex	1990
3	4	Least developed countries	11075966.0	Both sex	1990
4	5	Less developed regions excluding least develop...	59105261.0	Both sex	1990
...	...	...	...	...	...
85	1	WORLD	117584801.0	Female	2015
86	2	Developed regions	72863336.0	Female	2015
87	3	Developing regions	44721465.0	Female	2015
88	4	Least developed countries	5493028.0	Female	2015
89	5	Less developed regions excluding least develop...	39228437.0	Female	2015

90 rows × 5 columns



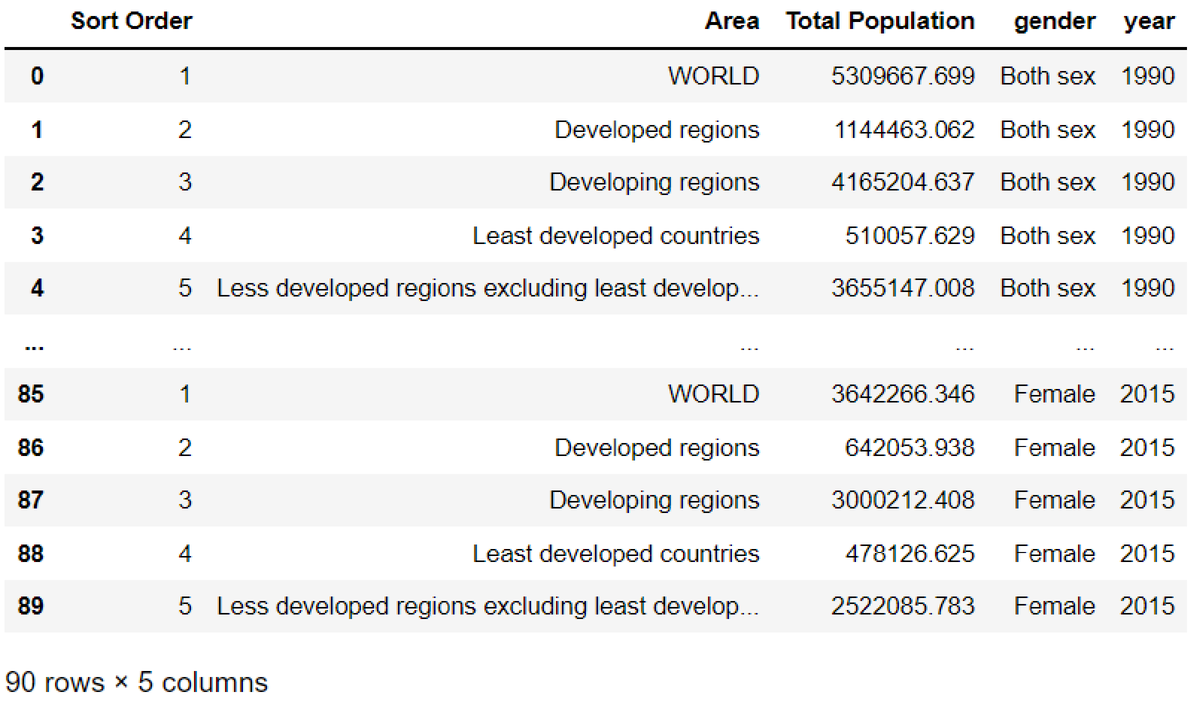
	Sort Order	Area	International Migrant Stock	gender	year
0	6	Sub-Saharan Africa	14690319.0	Both sex	1990
1	7	Africa	15690623.0	Both sex	1990
2	8	Eastern Africa	5964031.0	Both sex	1990
3	29	Middle Africa	1460530.0	Both sex	1990
4	39	Northern Africa	2403200.0	Both sex	1990
...	...	...	...	...	...
499	238	Oceania	4101334.0	Female	2015
500	239	Australia and New Zealand	3963032.0	Female	2015
501	242	Melanesia	47782.0	Female	2015
502	248	Micronesia	57159.0	Female	2015
503	256	Polynesia	33361.0	Female	2015

504 rows × 5 columns



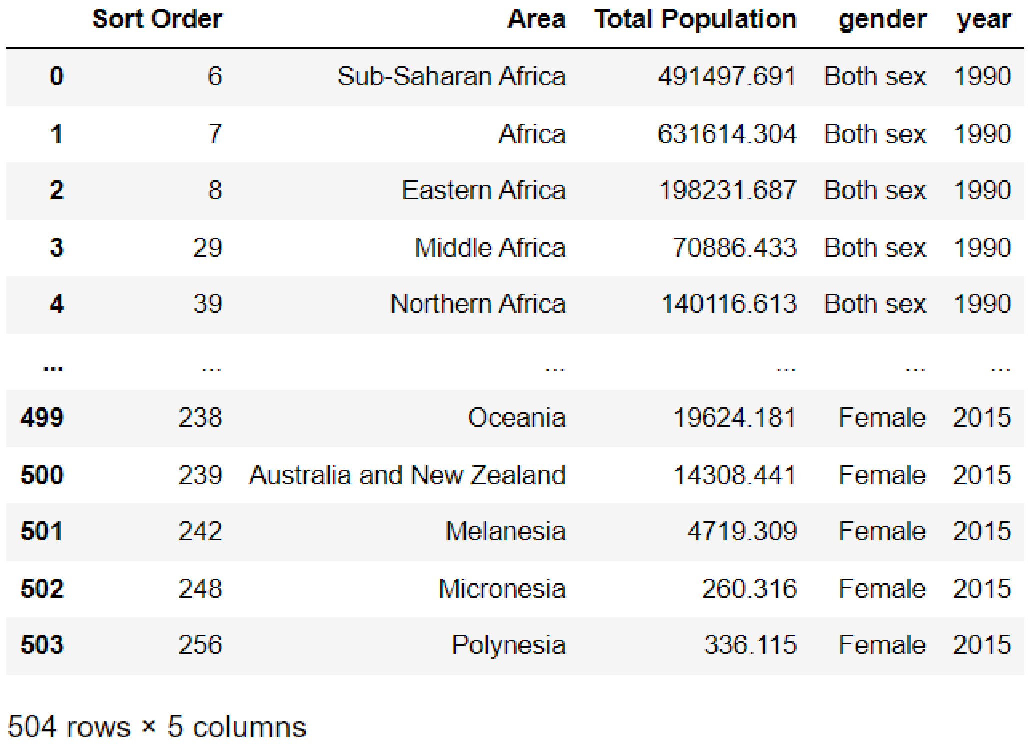
	Sort Order	Area	International Migrant Stock	gender	year
0	9	Burundi	333110.0	Both sex	1990
1	10	Comoros	14079.0	Both sex	1990
2	11	Djibouti	122221.0	Both sex	1990
3	12	Eritrea	11848.0	Both sex	1990
4	13	Ethiopia	1155390.0	Both sex	1990
...	...	...	...	...	...
4171	261	Samoa	2460.0	Female	2015
4172	262	Tokelau	254.0	Female	2015
4173	263	Tonga	2604.0	Female	2015
4174	264	Tuvalu	63.0	Female	2015
4175	265	Wallis and Futuna Islands	1411.0	Female	2015

4176 rows × 5 columns



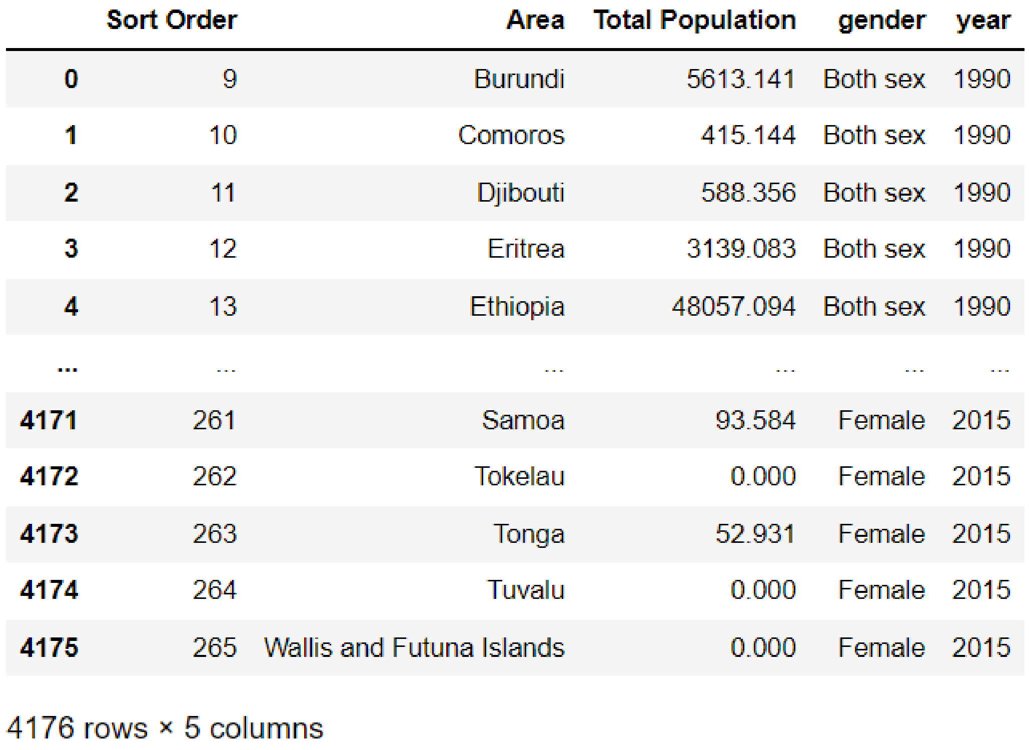
	Sort Order	Area	Total Population	gender	year
0	1	WORLD	5309667.699	Both sex	1990
1	2	Developed regions	1144463.062	Both sex	1990
2	3	Developing regions	4165204.637	Both sex	1990
3	4	Least developed countries	510057.629	Both sex	1990
4	5	Less developed regions excluding least develop...	3655147.008	Both sex	1990
...	...	...	...	...	...
85	1	WORLD	3642266.346	Female	2015
86	2	Developed regions	642053.938	Female	2015
87	3	Developing regions	3000212.408	Female	2015
88	4	Least developed countries	478126.625	Female	2015
89	5	Less developed regions excluding least develop...	2522085.783	Female	2015

90 rows × 5 columns



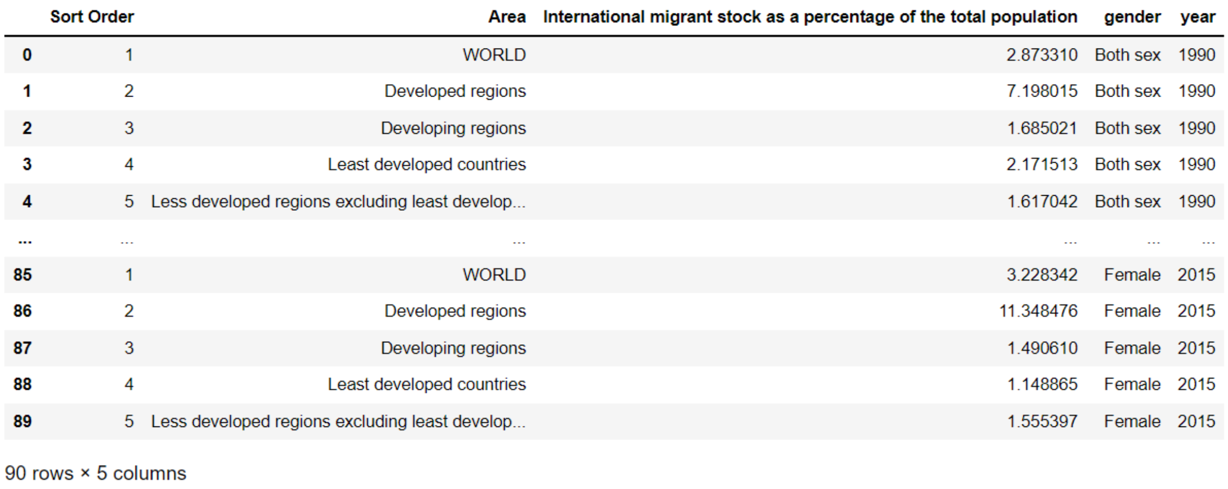
	Sort Order	Area	Total Population	gender	year
0	6	Sub-Saharan Africa	491497.691	Both sex	1990
1	7	Africa	631614.304	Both sex	1990
2	8	Eastern Africa	198231.687	Both sex	1990
3	29	Middle Africa	70886.433	Both sex	1990
4	39	Northern Africa	140116.613	Both sex	1990
...	...	...	...	...	...
499	238	Oceania	19624.181	Female	2015
500	239	Australia and New Zealand	14308.441	Female	2015
501	242	Melanesia	4719.309	Female	2015
502	248	Micronesia	260.316	Female	2015
503	256	Polynesia	336.115	Female	2015

504 rows × 5 columns



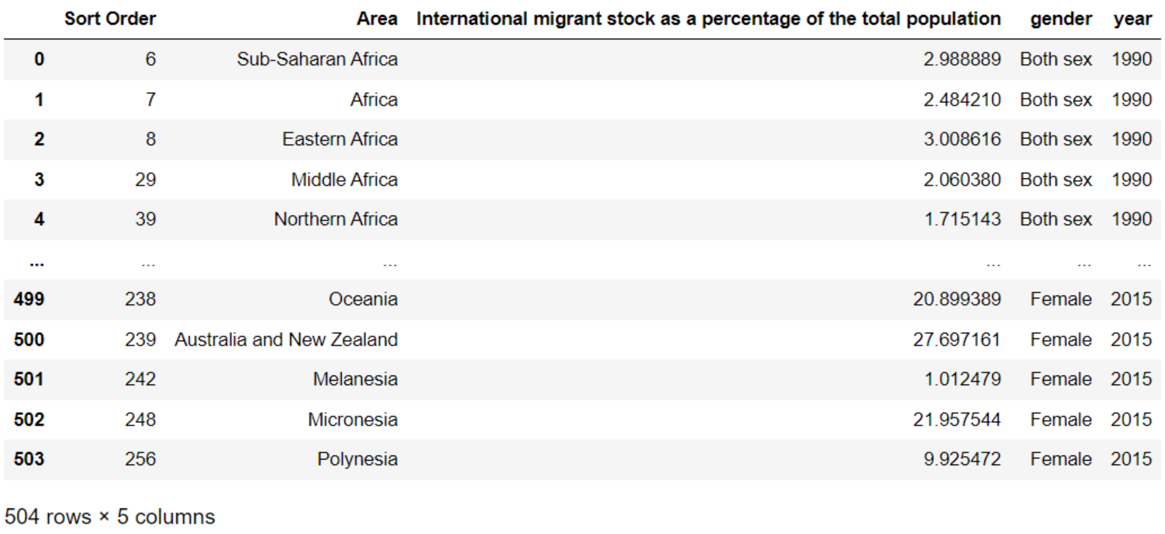
	Sort Order	Area	Total Population	gender	year
0	9	Burundi	5613.141	Both sex	1990
1	10	Comoros	415.144	Both sex	1990
2	11	Djibouti	588.356	Both sex	1990
3	12	Eritrea	3139.083	Both sex	1990
4	13	Ethiopia	48057.094	Both sex	1990
...	...	...	...	...	...
4171	261	Samoa	93.584	Female	2015
4172	262	Tokelau	0.000	Female	2015
4173	263	Tonga	52.931	Female	2015
4174	264	Tuvalu	0.000	Female	2015
4175	265	Wallis and Futuna Islands	0.000	Female	2015

4176 rows × 5 columns



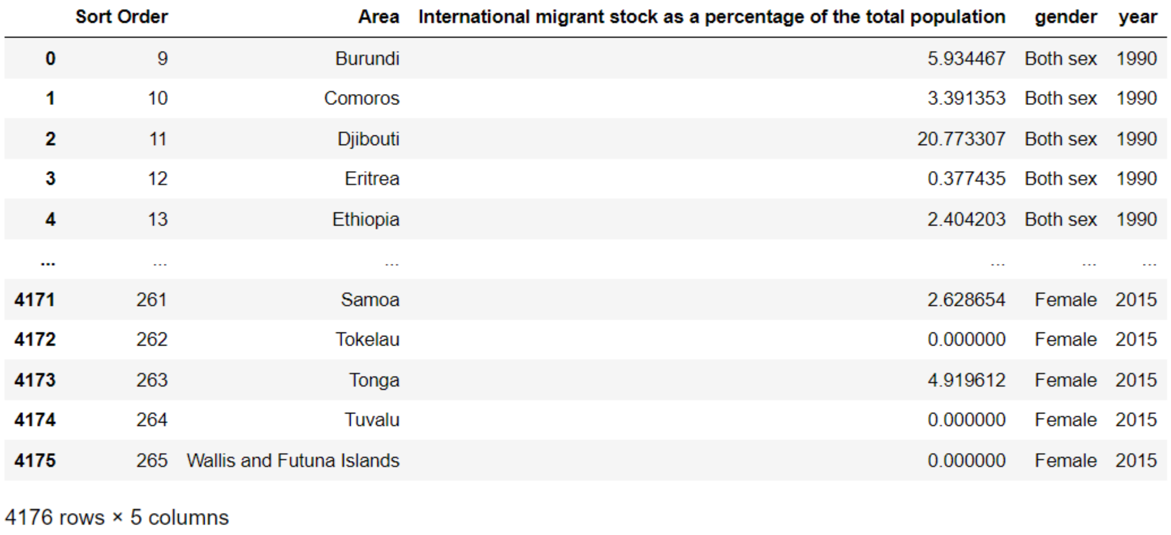
	Sort Order	Area	International migrant stock as a percentage of the total population	gender	year
0	1	WORLD	2.873310	Both sex	1990
1	2	Developed regions	7.198015	Both sex	1990
2	3	Developing regions	1.685021	Both sex	1990
3	4	Least developed countries	2.171513	Both sex	1990
4	5	Less developed regions excluding least develop...	1.617042	Both sex	1990
...	...	...	...	...	...
85	1	WORLD	3.228342	Female	2015
86	2	Developed regions	11.348476	Female	2015
87	3	Developing regions	1.490610	Female	2015
88	4	Least developed countries	1.148865	Female	2015
89	5	Less developed regions excluding least develop...	1.555397	Female	2015

90 rows × 5 columns



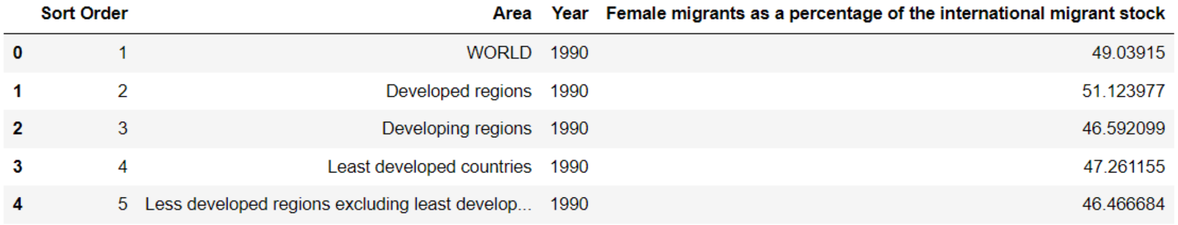
	Sort Order	Area	International migrant stock as a percentage of the total population	gender	year
0	6	Sub-Saharan Africa	2.988889	Both sex	1990
1	7	Africa	2.484210	Both sex	1990
2	8	Eastern Africa	3.008616	Both sex	1990
3	29	Middle Africa	2.060380	Both sex	1990
4	39	Northern Africa	1.715143	Both sex	1990
...	...	...	...	...	...
499	238	Oceania	20.899389	Female	2015
500	239	Australia and New Zealand	27.697161	Female	2015
501	242	Melanesia	1.012479	Female	2015
502	248	Micronesia	21.957544	Female	2015
503	256	Polynesia	9.925472	Female	2015

504 rows × 5 columns

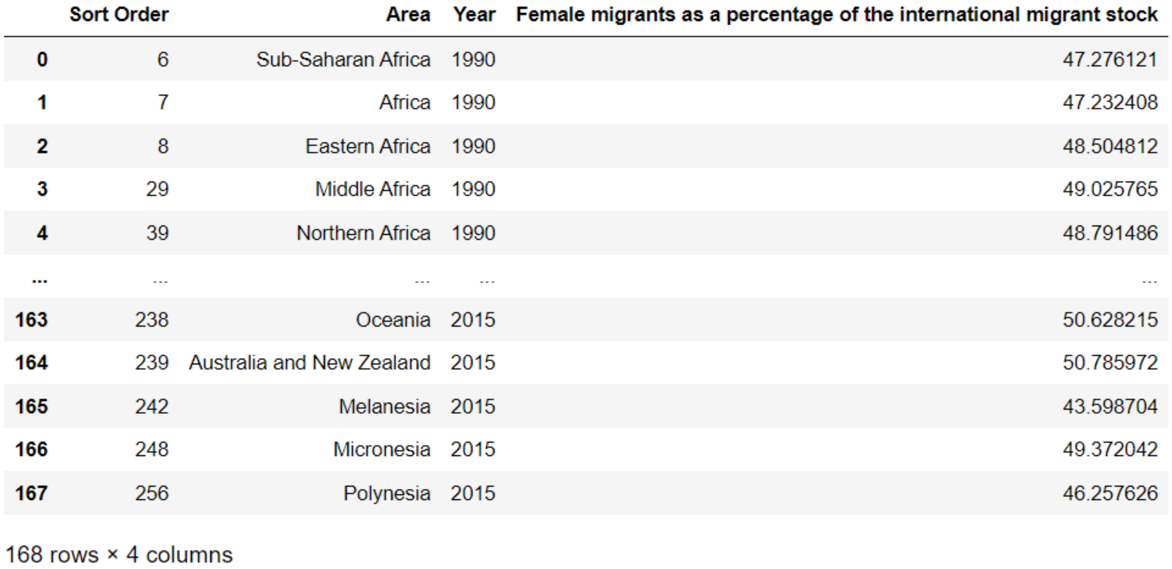


	Sort Order	Area	International migrant stock as a percentage of the total population	gender	year
0	9	Burundi	5.934467	Both sex	1990
1	10	Comoros	3.391353	Both sex	1990
2	11	Djibouti	20.773307	Both sex	1990
3	12	Eritrea	0.377435	Both sex	1990
4	13	Ethiopia	2.404203	Both sex	1990
...	...	...	...	...	...
4171	261	Samoa	2.628654	Female	2015
4172	262	Tokelau	0.000000	Female	2015
4173	263	Tonga	4.919612	Female	2015
4174	264	Tuvalu	0.000000	Female	2015
4175	265	Wallis and Futuna Islands	0.000000	Female	2015

4176 rows × 5 columns

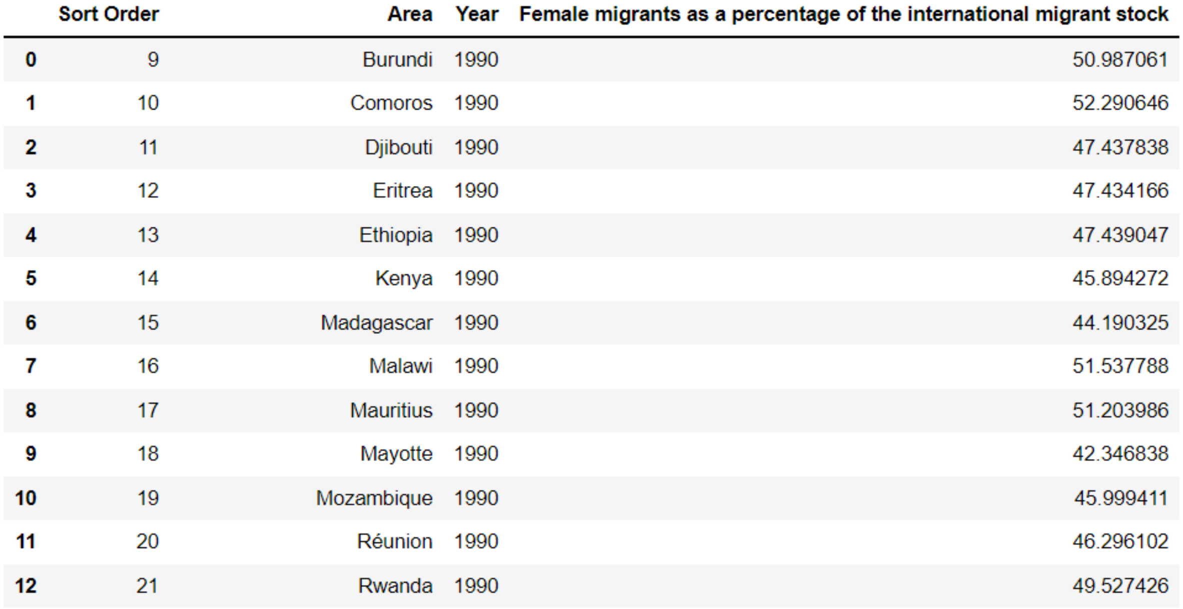


	Sort Order	Area	Year	Female migrants as a percentage of the international migrant stock
0	1	WORLD	1990	49.03915
1	2	Developed regions	1990	51.123977
2	3	Developing regions	1990	46.592099
3	4	Least developed countries	1990	47.261155
4	5	Less developed regions excluding least develop...	1990	46.466684

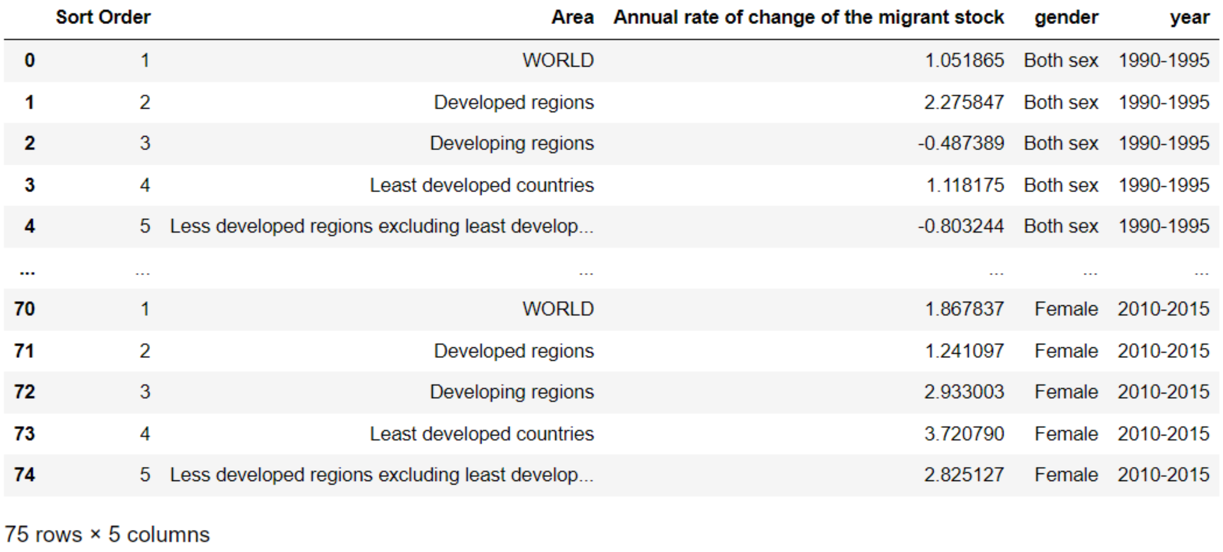


	Sort Order	Area	Year	Female migrants as a percentage of the international migrant stock
0	6	Sub-Saharan Africa	1990	47.276121
1	7	Africa	1990	47.232408
2	8	Eastern Africa	1990	48.504812
3	29	Middle Africa	1990	49.025765
4	39	Northern Africa	1990	48.791486
...	...	...	...	...
163	238	Oceania	2015	50.628215
164	239	Australia and New Zealand	2015	50.785972
165	242	Melanesia	2015	43.598704
166	248	Micronesia	2015	49.372042
167	256	Polynesia	2015	46.257626

168 rows × 4 columns

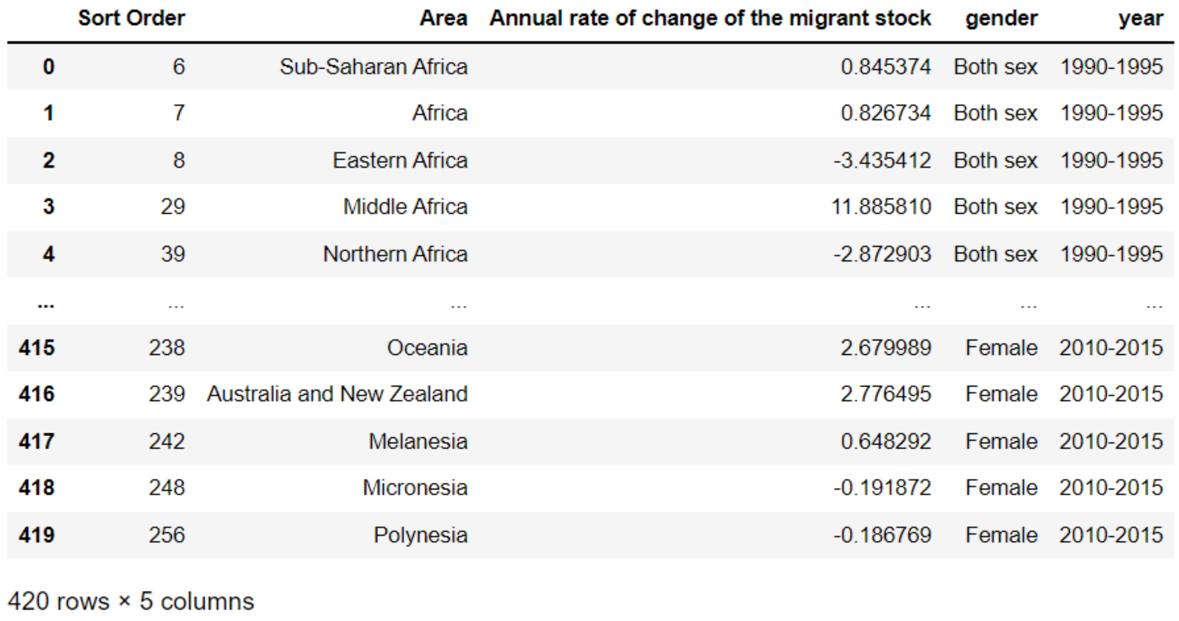


	Sort Order	Area	Year	Female migrants as a percentage of the international migrant stock
0	9	Burundi	1990	50.987061
1	10	Comoros	1990	52.290646
2	11	Djibouti	1990	47.437838
3	12	Eritrea	1990	47.434166
4	13	Ethiopia	1990	47.439047
5	14	Kenya	1990	45.894272
6	15	Madagascar	1990	44.190325
7	16	Malawi	1990	51.537788
8	17	Mauritius	1990	51.203986
9	18	Mayotte	1990	42.346838
10	19	Mozambique	1990	45.999411
11	20	Réunion	1990	46.296102
12	21	Rwanda	1990	49.527426



	Sort Order	Area	Annual rate of change of the migrant stock	gender	year
0	1	WORLD	1.051865	Both sex	1990-1995
1	2	Developed regions	2.275847	Both sex	1990-1995
2	3	Developing regions	-0.487389	Both sex	1990-1995
3	4	Least developed countries	1.118175	Both sex	1990-1995
4	5	Less developed regions excluding least develop...	-0.803244	Both sex	1990-1995
...	...	...	...	...	...
70	1	WORLD	1.867837	Female	2010-2015
71	2	Developed regions	1.241097	Female	2010-2015
72	3	Developing regions	2.933003	Female	2010-2015
73	4	Least developed countries	3.720790	Female	2010-2015
74	5	Less developed regions excluding least develop...	2.825127	Female	2010-2015

75 rows × 5 columns



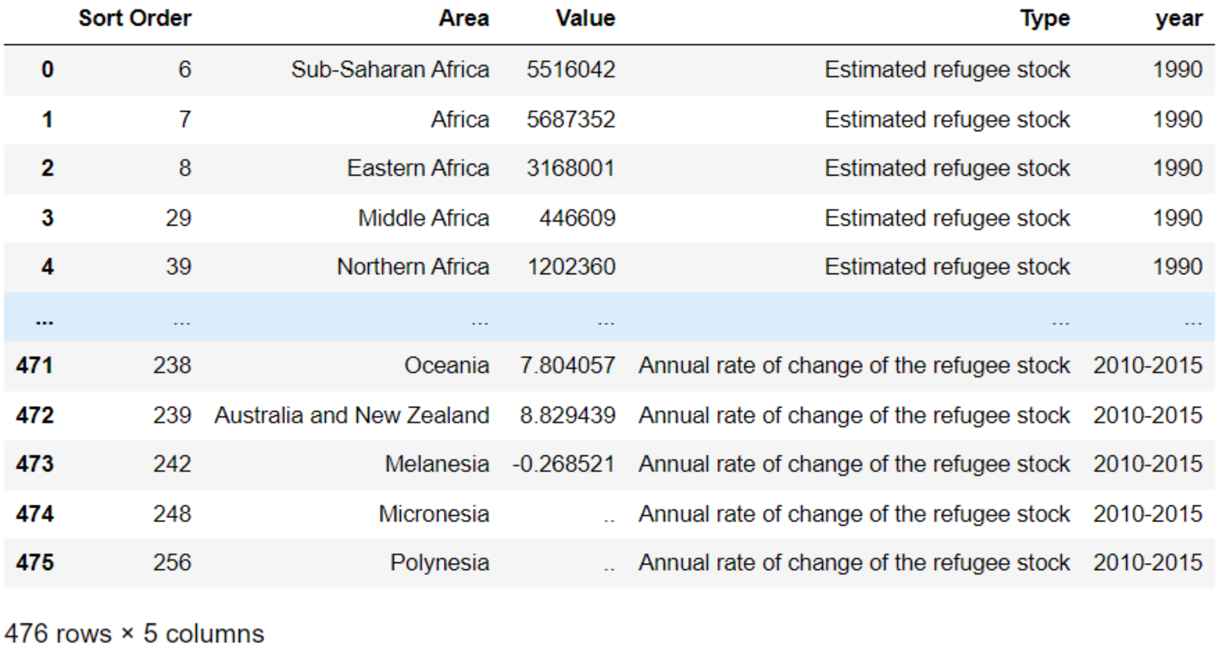
	Sort Order	Area	Annual rate of change of the migrant stock	gender	year
0	6	Sub-Saharan Africa	0.845374	Both sex	1990-1995
1	7	Africa	0.826734	Both sex	1990-1995
2	8	Eastern Africa	-3.435412	Both sex	1990-1995
3	29	Middle Africa	11.885810	Both sex	1990-1995
4	39	Northern Africa	-2.872903	Both sex	1990-1995
...	...	...	...	...	...
415	238	Oceania	2.679989	Female	2010-2015
416	239	Australia and New Zealand	2.776495	Female	2010-2015
417	242	Melanesia	0.648292	Female	2010-2015
418	248	Micronesia	-0.191872	Female	2010-2015
419	256	Polynesia	-0.186769	Female	2010-2015

420 rows × 5 columns

	Sort Order	Area	Annual rate of change of the migrant stock	gender	year
0	9	Burundi	-5.355717	Both sex	1990-1995
1	10	Comoros	-0.199873	Both sex	1990-1995
2	11	Djibouti	-4.058465	Both sex	1990-1995
3	12	Eritrea	0.910748	Both sex	1990-1995
4	13	Ethiopia	-7.179771	Both sex	1990-1995
5	14	Kenya	14.659568	Both sex	1990-1995
6	15	Madagascar	-2.433476	Both sex	1990-1995
7	16	Malawi	-30.811478	Both sex	1990-1995
8	17	Mauritius	14.588616	Both sex	1990-1995
9	18	Mayotte	10.939512	Both sex	1990-1995
10	19	Mozambique	6.374959	Both sex	1990-1995
11	20	Réunion	5.982790	Both sex	1990-1995

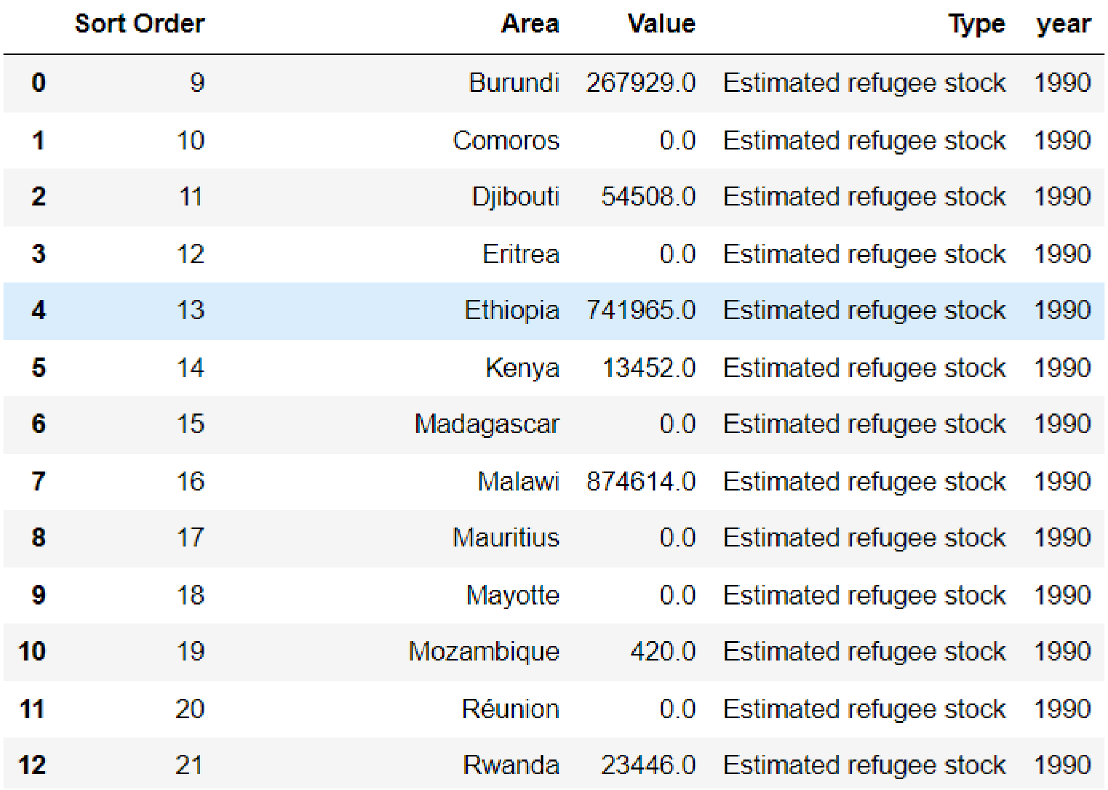
	Sort Order	Area	Value	Type	year
0	1	WORLD	1.883657e+07	Estimated refugee stock	1990
1	2	Developed regions	2.014564e+06	Estimated refugee stock	1990
2	3	Developing regions	1.682201e+07	Estimated refugee stock	1990
3	4	Least developed countries	5.048391e+06	Estimated refugee stock	1990
4	5	Less developed regions excluding least develop...	1.177362e+07	Estimated refugee stock	1990
...	...	...	...	...	...
80	1	WORLD	2.947267e+00	Annual rate of change of the refugee stock	2010-2015
81	2	Developed regions	-2.087656e+00	Annual rate of change of the refugee stock	2010-2015
82	3	Developing regions	2.663652e+00	Annual rate of change of the refugee stock	2010-2015
83	4	Least developed countries	7.766031e+00	Annual rate of change of the refugee stock	2010-2015
84	5	Less developed regions excluding least develop...	1.571047e+00	Annual rate of change of the refugee stock	2010-2015

85 rows × 5 columns



	Sort Order	Area	Value	Type	year
0	6	Sub-Saharan Africa	5516042	Estimated refugee stock	1990
1	7	Africa	5687352	Estimated refugee stock	1990
2	8	Eastern Africa	3168001	Estimated refugee stock	1990
3	29	Middle Africa	446609	Estimated refugee stock	1990
4	39	Northern Africa	1202360	Estimated refugee stock	1990
...	...	...	...	...	...
471	238	Oceania	7.804057	Annual rate of change of the refugee stock	2010-2015
472	239	Australia and New Zealand	8.829439	Annual rate of change of the refugee stock	2010-2015
473	242	Melanesia	-0.268521	Annual rate of change of the refugee stock	2010-2015
474	248	Micronesia	..	Annual rate of change of the refugee stock	2010-2015
475	256	Polynesia	..	Annual rate of change of the refugee stock	2010-2015

476 rows × 5 columns



	Sort Order	Area	Value	Type	year
0	9	Burundi	267929.0	Estimated refugee stock	1990
1	10	Comoros	0.0	Estimated refugee stock	1990
2	11	Djibouti	54508.0	Estimated refugee stock	1990
3	12	Eritrea	0.0	Estimated refugee stock	1990
4	13	Ethiopia	741965.0	Estimated refugee stock	1990
5	14	Kenya	13452.0	Estimated refugee stock	1990
6	15	Madagascar	0.0	Estimated refugee stock	1990
7	16	Malawi	874614.0	Estimated refugee stock	1990
8	17	Mauritius	0.0	Estimated refugee stock	1990
9	18	Mayotte	0.0	Estimated refugee stock	1990
10	19	Mozambique	420.0	Estimated refugee stock	1990
11	20	Réunion	0.0	Estimated refugee stock	1990
12	21	Rwanda	23446.0	Estimated refugee stock	1990

Discussion

The most challenge thing I found in this data cleaning project is to draw a mind map about how to clean the data set and what should an idea tidy data set look like in this case. It took me almost 80% of time and effort to do so. On the country, the python methods can be easily found in the lab records and google websites.

I learned a lot from the data cleaning project that, real-world data may be structured in a way that is not conducive to analysis, thus we should think about how to deal with the collected dataset according to the tidy data principles before we start to do the analysis.

Resource

UN_MigrantStockTotal_2015.xlsx

Jupyter notebook

Python 3.4