INF1340: Midterm assignment write-up

Ruolan Zhang
1004711010

Measurement plan

Carefully cleaning the messy dataset facilitates our future data analysis process and makes it easy to apply complicated statistical methods. As I was browsing around 6 tables in this excel, UN_MIgrantStockTotal, I observed that the biggest challenge in the data cleaning process is tidying data. Every table in this excel provides analysts with distinct statistics about international migrant stocks. According to the five principles stated in our class, all those tables require a few steps of manipulation to become tidy data. I will first check whether there exist null values in each table before passing on to tidy the dataset.

The most frequent tidy data principles violated are Principles 1 and 2. All the tables violate Principle 1, which means their column headers are named by values instead of the variable names. Principle 2 is violated in most of the tables(table1,2,3,5,6) after eliminating the violation of Principle 1 using ‘melt’ and lambda function. Since both variables, gender and year, are stored in one column. These two violations may easily confuse people who have a first glimpse of the dataset, making the data in this dataset hard to access, and this is the reason why I would like to correct them. Violation of Principle 4 appears in table 6, which stores multiple data types in one table. So I decided to split this table into separate ones.

Introduction

Prior to tidying the dataset, I did a series of data manipulations. I read the excel table starting from the 15th row because 1 to 14 rows are merged into the title and provide no useful information. Since I observed that the column, “Major area, region, country or area of destination”, complies with stratified levels of districts together into one column, I was thinking about separating observations of major areas(eg. Asia), regions(eg. Southern Asia) and countries(eg. Afghanistan) and store them into different datasets. I stratified the original column and tidied those datasets respectively. In this way, future analysts who get access to the cleaned datasets are able to filter needed information more efficiently and have clear minds about the overall condition for every stratified level. In addition to that, the first 6 rows of every table group each dependent variable according to another criterion, developing or developed regions, which is a completely different grouping method. Therefore, I also filter out those rows and store them in a separate single data frame.

Basically, I separated those new datasets according to the information provided in the column “Type of data(a)”. Firstly, observations with values inside this column are stratified into the country level, while the rest of the observations are the collection of area and region levels. For further stratification, I used the index of “Country code” of areas and regions in the ANNEX worksheet to match with that in the above dataset. Observations that have country codes in the list of area country codes will be classified as area observation. Similarly, the rest are region observations.

During the process of tidying the dataset, I renamed column names into meaningful abbreviations with a combination of gender and year, and then used the melt function to eliminate the violation of Principle 1. Secondly, I applied the lambda function to assign two new columns to the dataset using the abbreviation I set before. Finally, I styled the resulting data frame by replacing the abbreviation of gender. This is a summary of dealing with violations of principles 1 and 2 in general. More detailed operations will be provided in the next section.

Results

Table 1 & 2 & 3

The first 3 tables in the UN_MIgrantStockTotal excel face pretty much similar problems although their focuses are different from each other, so I decided to discuss them together in this part. Those three tables contain statistics from male, female, and also has a section for both sexes which is the sum of the two genders.

Checking null values is a necessary step before we implement any further cleaning jobs. Selects all statistics and uses isna() to produce boolean values for each cell of observations. Since false stands for 0 and true stands for 1 in the python language, calculating the sum of all boolean values of cells can inform us whether there exist any missing values in this table. All tables 1-3 pass this null value check and we can push forward our cleaning process. After that, I renamed each column in this table with a more informative name to provide us with better references.

Firstly, discuss how I dealt with the violation of Principle 1. These three tables name a few headers using 1990, 1995, 2000, 2005, and 2015 respectively, instead of the variable name “year”. Also, I have to specify the gender type for each observation when tidying the dataset. I renamed the original columns in the combination of gender and year(eg. “b1990” stands for both sexes in 1990). The resolution for this violation is to use the ”melt” function in a way that the specific gender-year combination would not be headers of the column but rather be the values under the header, “demographic”. Therefore, we successfully unpivot the dataset from wide to long format. Till now, we have solved the violation of Principle 1 and we are going to move on to Principle 2.

Principle 2 of tidy data demonstrates that there cannot be multiple variables stored in one column. As I mentioned in the last step, under the column “demographic”, there are gender-year combinations stored for every observation. This violates principle 2 and we have to convert them into tidy data structures. Therefore, I used the lambda function to separate this compiled column and used string operations to assign two new columns “Year” and “Gender”. Then “b”, “m”(stands for male), and “f”(stands for female) are assigned as values under the gender column, and specific year values are mapped under the “Year” columns. We have to drop the original “demographic” column at the same time since we don’t need this column as a reference anymore. At this point, we have eliminated the violation for Principle 2, but we can make some more modifications to style the value under columns. For example, values under the gender column are still abbreviations, so I replace those abbreviations with their full spellings.

After dealing with the violation of Principles 1 and 2, we have converted the messy table into a tidy dataset. No other principles are violated, so this table is ready for future analysis. As I mentioned that I separated the country, region, and area datasets, we should repeat this process on those three datasets as well as the one that groups observations by developed or developing countries. To be noticed, for the country dataset, I created two new columns that stated the region and area of every specific country using the “join” function. I left-joined the country dataset with the information from the ANNEX table to enable quick searching for more information.

Final datasets for table 1

Figure 1.1: World dataset:

Sort order	Country or area of destination	Notes	Country code	Type of data (a)	International migrant stock at mid-year	Gender	Year
1	WORLD	NaN	900	NaN	152563212	Both	1990
2	Developed regions	(b)	901	NaN	82378628	Both	1990
3	Developing regions	(c)	902	NaN	70184584	Both	1990
4	Least developed countries	(d)	941	NaN	11075966	Both	1990
5	Less developed regions excluding least develop...	NaN	934	NaN	59105261	Both	1990
...	...	...	...	...	...	...	...
3	Developing regions	(c)	902	NaN	44721465	Female	2015
4	Least developed countries	(d)	941	NaN	5493028	Female	2015
5	Less developed regions excluding least develop...	NaN	934	NaN	39228437	Female	2015
6	Sub-Saharan Africa	(e)	947	NaN	8894500	Female	2015
7	Africa	NaN	903	NaN	9526134	Female	2015

126 rows × 7 columns

Figure 1.2: Major area dataset:

Sort order	Major area	Notes	Country code	Type of data (a)	International migrant stock at mid-year	Gender	Year
7	Africa	NaN	903	NaN	15690623	Both	1990
71	Asia	NaN	935	NaN	48142261	Both	1990
127	Europe	NaN	908	NaN	49219200	Both	1990
180	Latin America and the Caribbean	NaN	904	NaN	7169728	Both	1990
232	Northern America	NaN	905	NaN	27610542	Both	1990
...	...	...	...	...	...	...	...
71	Asia	NaN	935	NaN	31530709	Female	2015
127	Europe	NaN	908	NaN	39873338	Female	2015
180	Latin America and the Caribbean	NaN	904	NaN	4650938	Female	2015
232	Northern America	NaN	905	NaN	27902348	Female	2015
238	Oceania	NaN	909	NaN	4101334	Female	2015

108 rows × 7 columns

Figure 1.3: Region dataset:

Sort order	Region	Notes	Country code	Type of data (a)	International migrant stock at mid-year	Gender	Year
8	Eastern Africa	NaN	910	NaN	5964031	Both	1990
29	Middle Africa	NaN	911	NaN	1460530	Both	1990
39	Northern Africa	NaN	912	NaN	2403200	Both	1990
47	Southern Africa	NaN	913	NaN	1392359	Both	1990
53	Western Africa	NaN	914	NaN	4470503	Both	1990
...	...	...	...	...	...	...	...
217	South America	NaN	931	NaN	2965406	Female	2015
239	Australia and New Zealand	NaN	927	NaN	3963032	Female	2015
242	Melanesia	NaN	928	NaN	47782	Female	2015
248	Micronesia	NaN	954	NaN	57159	Female	2015
256	Polynesia	NaN	957	NaN	33361	Female	2015

378 rows × 7 columns

Figure 1.4: Country dataset:

Sort order	Country or area of destination	Notes	Country code	Type of data (a)	Major area	Region	International migrant stock at mid-year	Gender	Year
9	Burundi	NaN	108	B R	Africa	Eastern Africa	333110	Both	1990
10	Comoros	NaN	174	B	Africa	Eastern Africa	14079	Both	1990
11	Djibouti	NaN	262	B R	Africa	Eastern Africa	122221	Both	1990
12	Eritrea	NaN	232	I	Africa	Eastern Africa	11848	Both	1990
13	Ethiopia	NaN	231	B R	Africa	Eastern Africa	1155390	Both	1990
...	...	...	...	...	...	...	...	...	...
261	Samoa	NaN	882	B	Oceania	Polynesia	2460	Female	2015
262	Tokelau	NaN	772	B	Oceania	Polynesia	254	Female	2015
263	Tonga	NaN	776	B	Oceania	Polynesia	2604	Female	2015
264	Tuvalu	NaN	798	C	Oceania	Polynesia	63	Female	2015
265	Wallis and Futuna Islands	NaN	876	B	Oceania	Polynesia	1411	Female	2015

4176 rows × 9 columns

Final datasets for table 2

Figure 2.1: World dataset:

Sort order	Country or area of destination	Notes	Country code	Total population at mid-year	Gender	Year
1	WORLD	NaN	900	5309667.699	Both	1990
2	Developed regions	(b)	901	1144463.062	Both	1990
3	Developing regions	(c)	902	4165204.637	Both	1990
4	Least developed countries	(d)	941	510057.629	Both	1990
5	Less developed regions excluding least develop...	NaN	934	3655147.008	Both	1990
...	...	...	...	...	...	...
3	Developing regions	(c)	902	3000212.408	Female	2015
4	Least developed countries	(d)	941	478126.625	Female	2015
5	Less developed regions excluding least develop...	NaN	934	2522085.783	Female	2015
6	Sub-Saharan Africa	(e)	947	481234.301	Female	2015
7	Africa	NaN	903	592723.652	Female	2015

126 rows × 6 columns

Figure 2.2: Major area dataset:

Sort order	Major area	Notes	Country code	Total population at mid-year	Gender	Year
7	Africa	NaN	903	631614.304	Both	1990
71	Asia	NaN	935	3202474.692	Both	1990
127	Europe	NaN	908	721086.311	Both	1990
180	Latin America and the Caribbean	NaN	904	446888.767	Both	1990
232	Northern America	NaN	905	280633.063	Both	1990
...	...	...	...	...	...	...
71	Asia	NaN	935	2146310.075	Female	2015
127	Europe	NaN	908	382275.191	Female	2015
180	Latin America and the Caribbean	NaN	904	320877.844	Female	2015
232	Northern America	NaN	905	180455.403	Female	2015
238	Oceania	NaN	909	19624.181	Female	2015

108 rows × 6 columns

Figure 2.3: Region dataset:

Sort order	Region	Notes	Country code	Total population at mid-year	Gender	Year
8	Eastern Africa	NaN	910	198231.687	Both	1990
29	Middle Africa	NaN	911	70886.433	Both	1990
39	Northern Africa	NaN	912	140116.613	Both	1990
47	Southern Africa	NaN	913	42049.013	Both	1990
53	Western Africa	NaN	914	180330.558	Both	1990
...	...	...	...	...	...	...
217	South America	NaN	931	211982.336	Female	2015
239	Australia and New Zealand	NaN	927	14308.441	Female	2015
242	Melanesia	NaN	928	4719.309	Female	2015
248	Micronesia	NaN	954	260.316	Female	2015
256	Polynesia	NaN	957	336.115	Female	2015

378 rows × 6 columns

Figure 2.4: Country dataset:

Sort order	Country or area of destination	Notes	Country code	Major area	Region	Total population at mid-year	Gender	Year
9	Burundi	NaN	108	Africa	Eastern Africa	5613.141	Both	1990
10	Comoros	NaN	174	Africa	Eastern Africa	415.144	Both	1990
11	Djibouti	NaN	262	Africa	Eastern Africa	588.356	Both	1990
12	Eritrea	NaN	232	Africa	Eastern Africa	3139.083	Both	1990
13	Ethiopia	NaN	231	Africa	Eastern Africa	48057.094	Both	1990
...	...	...	...	...	...	...	...	...
261	Samoa	NaN	882	Oceania	Polynesia	93.584	Female	2015
262	Tokelau	NaN	772	Oceania	Polynesia	..	Female	2015
263	Tonga	NaN	776	Oceania	Polynesia	52.931	Female	2015
264	Tuvalu	NaN	798	Oceania	Polynesia	..	Female	2015
265	Wallis and Futuna Islands	NaN	876	Oceania	Polynesia	..	Female	2015

4176 rows × 8 columns

Final datasets for table 3

Figure 3.1: World dataset

Sort order	Country or area of destination	Notes	Country code	Type of data (a)	International migrant stock as a percentage of the total population	Gender	Year
1	WORLD	NaN	900	NaN	2.873310	Both	1990
2	Developed regions	(b)	901	NaN	7.198015	Both	1990
3	Developing regions	(c)	902	NaN	1.685021	Both	1990
4	Least developed countries	(d)	941	NaN	2.171513	Both	1990
5	Less developed regions excluding least develop...	NaN	934	NaN	1.617042	Both	1990
...	...	...	...	...	...	...	...
3	Developing regions	(c)	902	NaN	1.490610	Female	2015
4	Least developed countries	(d)	941	NaN	1.148865	Female	2015
5	Less developed regions excluding least develop...	NaN	934	NaN	1.555397	Female	2015
6	Sub-Saharan Africa	(e)	947	NaN	1.848268	Female	2015
7	Africa	NaN	903	NaN	1.607180	Female	2015

126 rows × 7 columns

Figure 3.2: Major area dataset

Sort order	Major area	Notes	Country code	Type of data (a)	International migrant stock as a percentage of the total population	Gender	Year
7	Africa	NaN	903	NaN	2.484210	Both	1990
71	Asia	NaN	935	NaN	1.503283	Both	1990
127	Europe	NaN	908	NaN	6.825702	Both	1990
180	Latin America and the Caribbean	NaN	904	NaN	1.604365	Both	1990
232	Northern America	NaN	905	NaN	9.838663	Both	1990
...	...	...	...	...	...	...	...
71	Asia	NaN	935	NaN	1.469066	Female	2015
127	Europe	NaN	908	NaN	10.430532	Female	2015
180	Latin America and the Caribbean	NaN	904	NaN	1.449442	Female	2015
232	Northern America	NaN	905	NaN	15.462185	Female	2015
238	Oceania	NaN	909	NaN	20.899389	Female	2015

108 rows × 7 columns

Figure 3.3: Region dataset

Sort order	Region	Notes	Country code	Type of data (a)	International migrant stock as a percentage of the total population	Gender	Year
8	Eastern Africa	NaN	910	NaN	3.008616	Both	1990
29	Middle Africa	NaN	911	NaN	2.060380	Both	1990
39	Northern Africa	NaN	912	NaN	1.715143	Both	1990
47	Southern Africa	NaN	913	NaN	3.311276	Both	1990
53	Western Africa	NaN	914	NaN	2.479060	Both	1990
...	...	...	...	...	...	...	...
217	South America	NaN	931	NaN	1.398893	Female	2015
239	Australia and New Zealand	NaN	927	NaN	27.697161	Female	2015
242	Melanesia	NaN	928	NaN	1.012479	Female	2015
248	Micronesia	NaN	954	NaN	21.957544	Female	2015
256	Polynesia	NaN	957	NaN	9.925472	Female	2015

378 rows × 7 columns

Figure 3.4: Country dataset

Sort order	Country or area of destination	Notes	Country code	Type of data (a)	Major area	Region	International migrant stock as a percentage of the total population	Gender	Year
9	Burundi	NaN	108	B R	Africa	Eastern Africa	5.934467	Both	1990
10	Comoros	NaN	174	B	Africa	Eastern Africa	3.391353	Both	1990
11	Djibouti	NaN	262	B R	Africa	Eastern Africa	20.773307	Both	1990
12	Eritrea	NaN	232	I	Africa	Eastern Africa	0.377435	Both	1990
13	Ethiopia	NaN	231	B R	Africa	Eastern Africa	2.404203	Both	1990
...	...	...	...	...	...	...	...	...	...
261	Samoa	NaN	882	B	Oceania	Polynesia	2.628654	Female	2015
262	Tokelau	NaN	772	B	Oceania	Polynesia	..	Female	2015
263	Tonga	NaN	776	B	Oceania	Polynesia	4.919612	Female	2015
264	Tuvalu	NaN	798	C	Oceania	Polynesia	..	Female	2015
265	Wallis and Futuna Islands	NaN	876	B	Oceania	Polynesia	..	Female	2015

4176 rows × 9 columns

Table 4

Only Principle 1 is violated in this table. It specifically focuses on the change of female migrants as the percentage of international migrant stock between 1990 and 2015. Having a similar problem as previous tables, it puts specific years as names of headers, which violates the structure of the tidy dataset. The melt function is applied here to generate a new column called “year” and pulls all specific years in its values. Different from table 1-3, it is totally fine not to use the lambda function since gender is not one of our variables in this focus. Performing this process on each of the datasets split, we can successfully structure tidy datasets.

Final datasets for table 4

Figure 4.1: World dataset

```
UN_4_world.head()
```

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

Figure 4.2: Major area dataset

```
UN_4_area.head()
```

Sort order	Major area	Notes	Country code	Type of data (a)	Year	Female migrants as a percentage of the international migrant stock
7	Africa	NaN	903	NaN	1990	47.232408
71	Asia	NaN	935	NaN	1990	45.96873
127	Europe	NaN	908	NaN	1990	51.346887
180	Latin America and the Caribbean	NaN	904	NaN	1990	49.830217
232	Northern America	NaN	905	NaN	1990	51.115342

Figure 4.3: Region dataset

Sort order	Region	Notes	Country code	Type of data (a)	Year	Female migrants as a percentage of the international migrant stock
8	Eastern Africa	NaN	910	NaN	1990	48.504812
29	Middle Africa	NaN	911	NaN	1990	49.025765
39	Northern Africa	NaN	912	NaN	1990	48.791486
47	Southern Africa	NaN	913	NaN	1990	39.606165
53	Western Africa	NaN	914	NaN	1990	46.486134
...	...	...	...	...	...	...
217	South America	NaN	931	NaN	2015	50.895754
239	Australia and New Zealand	NaN	927	NaN	2015	50.785972
242	Melanesia	NaN	928	NaN	2015	43.598704
248	Micronesia	NaN	954	NaN	2015	49.372042
256	Polynesia	NaN	957	NaN	2015	46.257626

126 rows × 6 columns

Figure 4.4: Country dataset

Sort order	Country or area of destination	Notes	Country code	Type of data (a)	Major area	Region	Year	Female migrants as a percentage of the international migrant stock
9	Burundi	NaN	108	B R	Africa	Eastern Africa	1990	50.987061
10	Comoros	NaN	174	B	Africa	Eastern Africa	1990	52.290646
11	Djibouti	NaN	262	B R	Africa	Eastern Africa	1990	47.437838
12	Eritrea	NaN	232	I	Africa	Eastern Africa	1990	47.434166
13	Ethiopia	NaN	231	B R	Africa	Eastern Africa	1990	47.439047
...	...	...	...	...	...	...	...	...
261	Samoa	NaN	882	B	Oceania	Polynesia	2015	49.908704
262	Tokelau	NaN	772	B	Oceania	Polynesia	2015	52.156057
263	Tonga	NaN	776	B	Oceania	Polynesia	2015	45.437096
264	Tuvalu	NaN	798	C	Oceania	Polynesia	2015	44.680851
265	Wallis and Futuna Islands	NaN	876	B	Oceania	Polynesia	2015	49.52615

1392 rows × 8 columns

Table 5

This table displays the annual rate of change of the migrant stock by sex between 1990 and 2015. From my perspective, although columns are named with time intervals rather than specific time points, it essentially violates the same principles(Principles 1&2) as we stated in the first three tables. After renaming columns with gender and year-interval combinations, perform similar cleaning steps by using the melt function, lambda function, and string operations. The resulting tables are the final tidy version.

Final datasets for table 5

Figure 5.1: World dataset

Sort order	Country or area of destination	Notes	Country code	Type of data (a)	Annual rate of change of the migrant stock	Gender	Year
1	WORLD	NaN	900	NaN	1.051865	Both	1990-1995
2	Developed regions	(b)	901	NaN	2.275847	Both	1990-1995
3	Developing regions	(c)	902	NaN	-0.487389	Both	1990-1995
4	Least developed countries	(d)	941	NaN	1.118175	Both	1990-1995
5	Less developed regions excluding least develop...	NaN	934	NaN	-0.803244	Both	1990-1995
...	...	...	...	...	...	...	...
3	Developing regions	(c)	902	NaN	2.933003	Female	2005-2015
4	Least developed countries	(d)	941	NaN	3.720790	Female	2005-2015
5	Less developed regions excluding least develop...	NaN	934	NaN	2.825127	Female	2005-2015
6	Sub-Saharan Africa	(e)	947	NaN	3.928769	Female	2005-2015
7	Africa	NaN	903	NaN	3.996510	Female	2005-2015

105 rows × 7 columns

Figure 5.2: Major area dataset

Sort order	Major area	Notes	Country code	Type of data (a)	Annual rate of change of the migrant stock	Gender	Year
7	Africa	NaN	903	NaN	0.826734	Both	1990-1995
71	Asia	NaN	935	NaN	-0.673431	Both	1990-1995
127	Europe	NaN	908	NaN	1.420702	Both	1990-1995
180	Latin America and the Caribbean	NaN	904	NaN	-1.371210	Both	1990-1995
232	Northern America	NaN	905	NaN	3.771892	Both	1990-1995
...	...	...	...	...	...	...	...
71	Asia	NaN	935	NaN	2.583965	Female	2005-2015
127	Europe	NaN	908	NaN	1.121519	Female	2005-2015
180	Latin America and the Caribbean	NaN	904	NaN	2.288607	Female	2005-2015
232	Northern America	NaN	905	NaN	1.292130	Female	2005-2015
238	Oceania	NaN	909	NaN	2.679989	Female	2005-2015

90 rows × 7 columns

Figure 5.3: Region dataset

Sort order	Region	Notes	Country code	Type of data (a)	Annual rate of change of the migrant stock	Gender	Year
8	Eastern Africa	NaN	910	NaN	-3.435412	Both	1990-1995
29	Middle Africa	NaN	911	NaN	11.885810	Both	1990-1995
39	Northern Africa	NaN	912	NaN	-2.872903	Both	1990-1995
47	Southern Africa	NaN	913	NaN	-3.114352	Both	1990-1995
53	Western Africa	NaN	914	NaN	3.817706	Both	1990-1995
...	...	...	...	...	...	...	...
217	South America	NaN	931	NaN	2.510638	Female	2005-2015
239	Australia and New Zealand	NaN	927	NaN	2.776495	Female	2005-2015
242	Melanesia	NaN	928	NaN	0.648292	Female	2005-2015
248	Micronesia	NaN	954	NaN	-0.191872	Female	2005-2015
256	Polynesia	NaN	957	NaN	-0.186769	Female	2005-2015

315 rows × 7 columns

Figure 5.4: Country dataset

```
UN_5.head()
```

Sort order	Country or area of destination	Notes	Country code	Type of data (a)	Major area	Region	Annual rate of change of the migrant stock	Gender	Year
9	Burundi	NaN	108	B R	Africa	Eastern Africa	-5.355717	Both	1990-1995
10	Comoros	NaN	174	B	Africa	Eastern Africa	-0.199873	Both	1990-1995
11	Djibouti	NaN	262	B R	Africa	Eastern Africa	-4.058465	Both	1990-1995
12	Eritrea	NaN	232	I	Africa	Eastern Africa	0.910748	Both	1990-1995
13	Ethiopia	NaN	231	B R	Africa	Eastern Africa	-7.179771	Both	1990-1995

Table 6

Before eliminating violations in Table 6, I first replaced all “..” values with “No Value” for clarification. Table 6 violates the first, second, and fourth principles of tidy data. Firstly, talking about the violation of Principle 4, this single table contains multiple sections of data, therefore it would be better to break this table into separate ones. I split the last section “Annual rate of change of refugee stock” from the first two since it has column names in time intervals instead of specific time points.

Final dataset for table 6

Figure 6.1: Area or destination dataset part1

	Sort order	Area or destination	Notes	Country code	Type of data (a)	Year	Estimated refugee stock at mid-year for both sexes	Refugees as a percentage of the international migrant stock
0	1	WORLD	NaN	900	NaN	1990	18836571	12.346732
1	2	Developed regions	(b)	901	NaN	1990	2014564	2.445494
2	3	Developing regions	(c)	902	NaN	1990	16822007	23.968236
3	4	Least developed countries	(d)	941	NaN	1990	5048391	45.56588
4	5	Less developed regions excluding least develop...	NaN	934	NaN	1990	11773616	19.919743
...	...	...	...	...	...	...	...	...
193	238	Oceania	NaN	909	NaN	2015	54610	0.674124
194	239	Australia and New Zealand	NaN	927	NaN	2015	49408	0.63316
195	242	Melanesia	NaN	928	NaN	2015	4812	4.390711
196	248	Micronesia	NaN	954	NaN	2015	390	0.336869
197	256	Polynesia	NaN	957	NaN	2015	0	0.0

198 rows × 8 columns

Figure 6.2: Area or destination dataset part 2

	Sort order	Area or destination	Notes	Country code	Type of data (a)	Year	Annual rate of change of the refugee stock
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
...	...	...	...	...	...	...	...
160	238	Oceania	NaN	909	NaN	2010-2015	7.804057
161	239	Australia and New Zealand	NaN	927	NaN	2010-2015	8.829439
162	242	Melanesia	NaN	928	NaN	2010-2015	-0.268521
163	248	Micronesia	NaN	954	NaN	2010-2015	No Value
164	256	Polynesia	NaN	957	NaN	2010-2015	No Value

165 rows × 7 columns

Figure 6.3: Country dataset part1

	Sort order	Country	Notes	Country code	Type of data (a)	Major area	Region	Year	Estimated refugee stock at mid-year for both sexes	Refugees as a percentage of the international migrant stock
0	9	Burundi	NaN	108	B R	Africa	Eastern Africa	1990	267929	80.43259
1	10	Comoros	NaN	174	B	Africa	Eastern Africa	1990	0	0
2	11	Djibouti	NaN	262	B R	Africa	Eastern Africa	1990	54508	44.597901
3	12	Eritrea	NaN	232	I	Africa	Eastern Africa	1990	0	0
4	13	Ethiopia	NaN	231	B R	Africa	Eastern Africa	1990	741965	64.21771
...	...	...	...	...	...	...	...	...	...	...
1387	261	Samoa	NaN	882	B	Oceania	Polynesia	2015	0	0.0
1388	262	Tokelau	NaN	772	B	Oceania	Polynesia	2015	0	0.0
1389	263	Tonga	NaN	776	B	Oceania	Polynesia	2015	0	0.0
1390	264	Tuvalu	NaN	798	C	Oceania	Polynesia	2015	0	0.0
1391	265	Wallis and Futuna Islands	NaN	876	B	Oceania	Polynesia	2015	0	0.0

1392 rows × 10 columns

FIgure 6.4: Country dataset part2

	Sort order	Country	Notes	Country code	Type of data (a)	Major area	Region	Year	Annual rate of change of the refugee stock
0	9	Burundi	NaN	108	B R	Africa	Eastern Africa	1990-1995	-3.390926
1	10	Comoros	NaN	174	B	Africa	Eastern Africa	1990-1995	No Value
2	11	Djibouti	NaN	262	B R	Africa	Eastern Africa	1990-1995	-9.763426
3	12	Eritrea	NaN	232	I	Africa	Eastern Africa	1990-1995	No Value
4	13	Ethiopia	NaN	231	B R	Africa	Eastern Africa	1990-1995	-5.505717
...	...	...	...	...	...	...	...	...	...
1155	261	Samoa	NaN	882	B	Oceania	Polynesia	2010-2015	No Value
1156	262	Tokelau	NaN	772	B	Oceania	Polynesia	2010-2015	No Value
1157	263	Tonga	NaN	776	B	Oceania	Polynesia	2010-2015	No Value
1158	264	Tuvalu	NaN	798	C	Oceania	Polynesia	2010-2015	No Value
1159	265	Wallis and Futuna Islands	NaN	876	B	Oceania	Polynesia	2010-2015	No Value

1160 rows × 9 columns

Discussion

By observing the final tidy datasets, we can perform easy analysis without implementing any further data analysis methodology. Selecting the Observation named “World” in the “UN_1_world” dataset can present us with a better overview of the change in the number of international migrants in the world between 1990 and 2015, in which the total number of the international migrant stock rose from 152,563,212 in 1990 to 243,700,236 in 2015. Now the dataset is ready for performing data featuring and visualizations. In this project, I will stop here and talk more about what I learned from this cleaning process.

I figure out that most messy datasets violate the first two tidy data principles that are associated with the use of column names. Usually, a single “melt” function combined with a lambda function and string operations can eliminate these violations. That is the reason why we repetitively perform similar steps throughout this assignment.

Conclusion

I would like to conclude that tidying the dataset is a tedious but necessary process that every data scientist should perform on the raw dataset before applying further analysis methods. Similar to this assignment, it can be a repetitive process and requires the use of the same function under different conditions, but we should expect everyone’s resulting dataset to be somehow different according to their data cleaning preferences. Therefore, as data scientists, we cannot find the most correct method to do data cleaning but only the most proper one that provides convenience for your further analysis.