1340 Midterm Project Write Up

Keying Zhu

1． Introduction:

This data set is about the trends in international migrant stock provided by the United Nations, Department of Economic and Social Affairs, Population Division (2015). All data are stored in an excel file, there are nine separate sheets in total: one sheet for the introduction of contexts, six sheets are displayed six tables and two sheets are data frames. ANNEX explains country codes corresponding to countries, major areas, regions, whether is a developed region, whether is Least developed country, and whether in Sub-Saharan Africa. **NOTE:** Detailed explanation of specific words in the table.

To keep the data organized and available during future analysis. We need to clean and restructure the data for further analysis in the python environment. Through the above analysis of the content of each sheet, this tidy data will be carried out only on six tables. In this data cleaning, we will mainly be following5 principles of data cleaning, they are:

Principle 1: Column names need to be informative, variable names, and not values.
Principle 2: Each column needs to consist of one and only one variable.
Principle 3: Variables need to be in cells, not rows and columns.
Principle 4: Each table column needs to have a singular data type.
Principle 5: A single observational unit must be in one table.

2.Tidy data:

Firstly, we should import the excel file into jupyter notebook by using the *pd.read_excel()* function. Since this excel contains more than one sheet, I installed *openpyxl* to help load the excel file.

14	TABLE	TITLE
15	Table 1	International migrant stock at mid-year by sex...
16	Table 2	Total population at mid-year by sex and by maj...
17	Table 3	International migrant stock as a percentage of...
18	Table 4	Female migrants as a percentage of the interna...
19	Table 5	Annual rate of change of the migrant stock by ...
20	Table 6	Estimated refugee stock at mid-year by major a...
21	ANNEX	Classification of countries and areas by major...
22	NOTES	NOTES

There is an introduction in each table. Thus, when cleaning each table, my first step is to delete the introduction part and rename each column. There are values like "..", I replaced them with "NaN". 0 is meaningful in the tables, so I didn't replace ".." with 0. In addition, for the convenience of future analysis, I deleted "Sort order" and "Notes" eventually in each table.

2.1 Tidy Table 1

To make all the principles are followed, I tried two methods in the cleaning of the table1, the first method will follow principles 1, 2, and 3.

2.1.1 Method 1:

In table 1, column names are values, not variable names.

	Sort order	Area Destination	Notes	Country code	Type of data	1990B	1995B	2000B	2005B	2010B	...	2000M	2005M	2010M	
15	1	WORLD	NaN	900	NaN	152563212.0	160801752.0	172703309.0	191269100.0	221714243.0	...	87884839.0	97866674.0	114613714.0	12611
16	2	Developed regions	(b)	901	NaN	82378628.0	92306854.0	103375363.0	117181109.0	132560325.0	...	50536796.0	57217777.0	64081077.0	6761
17	3	Developing regions	(c)	902	NaN	70184584.0	68494898.0	69327946.0	74087991.0	89153918.0	...	37348043.0	40648897.0	50532637.0	5849

I solved this problem by following principle 1: Column names need to be informative, variable names, and not values. By using the *.melt* function, I created a new column named 'year' to store years' values. Switch "year" variables from row to column.

	Sort order	Area Destination	Notes	Country code	Type of data	Year	Count
0	1	WORLD	NaN	900	NaN	1990B	152563212.0
1	2	Developed regions	(b)	901	NaN	1990B	82378628.0
2	3	Developing regions	(c)	902	NaN	1990B	70184584.0

Now a new problem arises, variables "year" and "gender" are stored in one column. I solved this problem by following principle 2: Each column needs to consist of one and only one variable. By applying *.assign* function, I spat "year" and "gender" into two different columns.

	Area Destination	Country code	Year	Count	Gender
0	WORLD	900	1990	152563212.0	B
1	Developed regions	901	1990	82378628.0	B
2	Developing regions	902	1990	70184584.0	B
3	Least developed countries	941	1990	11075966.0	B
4	Less developed regions excluding least develop...	934	1990	59105261.0	B

After styling the context, I completed the full name of the gender. Now a new problem arises, variables are stored in both rows and columns.

	Area Destination	Country code	Year	Count	Gender
0	WORLD	900	1990	152563212.0	Both sexes
1	Developed regions	901	1990	82378628.0	Both sexes

I solved this problem by following principle 3: variables need to be in cells, not rows and columns. By using the *.pivot_table* function, I split the gender column into three columns by different gender groups.

Area Destination	Country code	Year	Both sexes	Female	Male
Afghanistan	4	1990	57686.0	25128.0	32558.0
Afghanistan	4	1995	71522.0	32417.0	39105.0
Afghanistan	4	2000	75917.0	33069.0	42848.0

2.1.2 Method 2:

Unlike method 1, in the first step, Method 2 spat three different data frames based on gender grouping(Both sexes, Male and Female).

	Area Destination	Country code	Type of data	1990	1995	2000	2005	2010	2015
15	WORLD	900	NaN	152563212.0	160801752.0	172703309.0	191269100.0	221714243.0	243700236.0
16	Developed regions	901	NaN	82378628.0	92306854.0	103375363.0	117181109.0	132560325.0	140481955.0
17	Developing regions	902	NaN	70184584.0	68494898.0	69327946.0	74087991.0	89153918.0	103218281.0
18	Least developed countries	941	NaN	11075966.0	11711703.0	10077824.0	9809634.0	10018128.0	11951316.0
19	Less developed regions excluding least develop...	934	NaN	59105261.0	56778501.0	59244124.0	64272611.0	79130668.0	91262036.0

The table above is one of the examples that shows the international migrant stock of both sexes. There is a problem with column names are values not variable names in the three data frames, I used *.melt* function three times to tidy data.

International Migrant Stock of both sexes:

	Area Destination	Country code	Type of data	Year	International Migrant Stock of Both Sexes
0	WORLD	900	NaN	1990	152563212.0
1	Developed regions	901	NaN	1990	82378628.0
2	Developing regions	902	NaN	1990	70184584.0
3	Least developed countries	941	NaN	1990	11075966.0
4	Less developed regions excluding least develop...	934	NaN	1990	59105261.0

International Migrant Stock of male:

	Area Destination	Country code	Type of data	Year	International Migrant Stock of Male
0	WORLD	900	NaN	1990	77747510.0
1	Developed regions	901	NaN	1990	40263397.0
2	Developing regions	902	NaN	1990	37484113.0
3	Least developed countries	941	NaN	1990	5843107.0
4	Less developed regions excluding least develop...	934	NaN	1990	31641006.0

International Migrant Stock of female:

	Area Destination	Country code	Type of data	Year	International Migrant Stock of Female
0	WORLD	900	NaN	1990	74815702.0
1	Developed regions	901	NaN	1990	42115231.0
2	Developing regions	902	NaN	1990	32700471.0
3	Least developed countries	941	NaN	1990	5236216.0
4	Less developed regions excluding least develop...	934	NaN	1990	27464255.0

Then I merged three data frames into one table on the first 4 columns. Thus, table_1 after cleaning will have 7 columns and 1590 rows.

	Area Destination	Country code	Type of data	Year	International Migrant Stock of Female	International Migrant Stock of Both Sexes	International Migrant Stock of Male
0	WORLD	900	NaN	1990	74815702.0	152563212.0	77747510.0
1	Developed regions	901	NaN	1990	42115231.0	82378628.0	40263397.0
2	Developing regions	902	NaN	1990	32700471.0	70184584.0	37484113.0
3	Least developed countries	941	NaN	1990	5236216.0	11075966.0	5843107.0
4	Less developed regions excluding least develop...	934	NaN	1990	27464255.0	59105261.0	31641006.0
...	...	...	...	...	...	...	...
1585	Samoa	882	B	2015	2460.0	4929.0	2469.0
1586	Tokelau	772	B	2015	254.0	487.0	233.0
1587	Tonga	776	B	2015	2604.0	5731.0	3127.0
1588	Tuvalu	798	C	2015	63.0	141.0	78.0
1589	Wallis and Futuna Islands	876	B	2015	1411.0	2849.0	1438.0

1590 rows × 7 columns

2.2 Tidy Table 2

Tidy table 2 using the similar method used for table 1 method 2. First I spat table 2 into three data frames by different gender groups(both sexes, male and female). There is a problem with column names are values not variable names in the three data frames, I used *.melt* function three times to tidy data.

Total population at mid-year of both sexes:

	Sort order	Area Destination	Notes	Country code	Year	Total population at mid-year of Both Sexes
0	1	WORLD	NaN	900	1990	5309667.699
1	2	Developed regions	(b)	901	1990	1144463.062
2	3	Developing regions	(c)	902	1990	4165204.637

Total population at mid-year of male:

	Sort order	Area Destination	Notes	Country code	Year	Total population at mid-year of Male
0	1	WORLD	NaN	900	1990	2670423.701
1	2	Developed regions	(b)	901	1990	555255.626
2	3	Developing regions	(c)	902	1990	2115168.075

Total population at mid-year of female:

	Sort order	Area Destination	Notes	Country code	Year	Total population at mid-year of Female
0	1	WORLD	NaN	900	1990	2639243.998
1	2	Developed regions	(b)	901	1990	589207.436
2	3	Developing regions	(c)	902	1990	2050036.562

Since table 2 does not have "Type of data", I merged three data frames into one table on the first 3 columns. Thus, table 2 after cleaning will have 6 columns and 1590 rows.

	Area Destination	Country code	Year	Total population at mid-year of Female	Total population at mid-year of Both Sexes	Total population at mid-year of Male
0	WORLD	900	1990	2639243.998	5309667.699	2670423.701
1	Developed regions	901	1990	589207.436	1144463.062	555255.626
2	Developing regions	902	1990	2050036.562	4165204.637	2115168.075
3	Least developed countries	941	1990	256015.073	510057.629	254042.556
4	Less developed regions excluding least develop...	934	1990	1794021.489	3655147.008	1861125.519
...	...	...	...	...	...	...
1585	Samoa	882	2015	93.584	193.228	99.644
1586	Tokelau	772	2015	NaN	1.250	NaN
1587	Tonga	776	2015	52.931	106.170	53.239
1588	Tuvalu	798	2015	NaN	9.916	NaN
1589	Wallis and Futuna Islands	876	2015	NaN	13.151	NaN

1590 rows × 6 columns

2.3 Tidy Table 3

By applying the same method used in Table 1 Method 2. First I spat table 3 into three data frames by different gender groups (both sexes, male and female). After using the three times *.melt* function, merged them into one table on the first 4 columns. Thus, table 3 after cleaning will have 7 columns and 1590 rows.

	Area Destination	Country code	Type of data	Year	International migrant stock as a percentage of Female	International migrant stock as a percentage of Both Sexes	International migrant stock as a percentage of Male
0	WORLD	900	NaN	1990	2.834740	2.873310	2.911430
1	Developed regions	901	NaN	1990	7.147777	7.198015	7.251326
2	Developing regions	902	NaN	1990	1.595116	1.685021	1.772158
3	Least developed countries	941	NaN	1990	2.045276	2.171513	2.300050
4	Less developed regions excluding least develop...	934	NaN	1990	1.530877	1.617042	1.700101
...	...	...	...	...	...	...	...
1585	Samoa	882	B	2015	2.628654	2.550873	2.477821
1586	Tokelau	772	B	2015	NaN	38.960000	NaN
1587	Tonga	776	B	2015	4.919612	5.397947	5.873514
1588	Tuvalu	798	C	2015	NaN	1.421944	NaN
1589	Wallis and Futuna Islands	876	B	2015	NaN	21.663752	NaN

1590 rows × 7 columns

2.4 Tidy Table 4

Table 4 is only about the female migrants as a percentage of the international migrant stock. I used the *.melt* function to move the year values from rows to columns. Thus, table 4 after cleaning will have 5 columns and 1590 rows.

	Area Destination	Country code	Type of data	Year	Female migrant presentage
0	WORLD	900	NaN	1990	49.039150
1	Developed regions	901	NaN	1990	51.123977
2	Developing regions	902	NaN	1990	46.592099
3	Least developed countries	941	NaN	1990	47.261155
4	Less developed regions excluding least develop...	934	NaN	1990	46.466684
...	...	...	...	...	...
1585	Samoa	882	B	2015	49.908704
1586	Tokelau	772	B	2015	52.156057
1587	Tonga	776	B	2015	45.437096
1588	Tuvalu	798	C	2015	44.680851
1589	Wallis and Futuna Islands	876	B	2015	49.526150

1590 rows × 5 columns

2.5 Tidy Table 5

By applying the same method used in Table 1 Method 2. First I spat table 5 into three data frames by different gender groups (both sexes, male and female). After using the three times *.melt* function, merged them into one table on the first 4 columns by using *.merge* function. Thus, table 5 after cleaning will have 7 columns and 1325 rows.

	Area Destination	Country code	Type of data	Year	Annual Rate of Change in Female	Annual Rate of Change in Both Sexes	Annual Rate of Change in Male
0	WORLD	900	NaN	1990-1995	1.104667	1.051865	1.000922
1	Developed regions	901	NaN	1990-1995	2.285643	2.275847	2.265595
2	Developing regions	902	NaN	1990-1995	-0.526904	-0.487389	-0.452980
3	Least developed countries	941	NaN	1990-1995	1.249146	1.118175	1.000073
4	Less developed regions excluding least develop...	934	NaN	1990-1995	-0.884180	-0.803244	-0.733256
...	...	...	...	...	...	...	...
1320	Samoa	882	B	2010-2015	-0.545343	-0.768177	-0.987758
1321	Tokelau	772	B	2010-2015	2.603250	2.536144	2.463246
1322	Tonga	776	B	2010-2015	2.526318	2.641235	2.737439
1323	Tuvalu	798	C	2010-2015	-1.819436	-1.763854	-1.718849
1324	Wallis and Futuna Islands	876	B	2010-2015	0.516899	0.519140	0.521340

1325 rows × 7 columns

2.6 Tidy Table 6

Table 6 has a problem in that there are multiple types of data stored in one table. By following tidy data principle 4: Each table column needs to have a singular data type. I spat table 3 into three data frames by estimated refugee stock at mid-year (both sexes), refugees as a percentage of the international migrant stock and the annual rate of change of the refugee stock.

Estimated refugee stock at mid-year (both sexes):

	Area Destination	Country code	Type of data	Year	Estimated refugee stock at mid-year (both sexes)
0	WORLD	900	NaN	1990	18836571.0
1	Developed regions	901	NaN	1990	2014564.0
2	Developing regions	902	NaN	1990	16822007.0
3	Least developed countries	941	NaN	1990	5048391.0
4	Less developed regions excluding least develop...	934	NaN	1990	11773616.0
...	...	...	...	...	...
1585	Samoa	882	B	2015	0.0
1586	Tokelau	772	B	2015	0.0
1587	Tonga	776	B	2015	0.0
1588	Tuvalu	798	C	2015	0.0
1589	Wallis and Futuna Islands	876	B	2015	0.0

1590 rows × 5 columns

Refugees as a percentage of the international migrant stock:

	Area Destination	Country code	Type of data	Year	Refugees as a percentage of the international migrant stock
0	WORLD	900	NaN	1990	12.346732
1	Developed regions	901	NaN	1990	2.445494
2	Developing regions	902	NaN	1990	23.968236
3	Least developed countries	941	NaN	1990	45.565880
4	Less developed regions excluding least develop...	934	NaN	1990	19.919743
...	...	...	...	...	...
1585	Samoa	882	B	2015	0.000000
1586	Tokelau	772	B	2015	0.000000
1587	Tonga	776	B	2015	0.000000
1588	Tuvalu	798	C	2015	0.000000
1589	Wallis and Futuna Islands	876	B	2015	0.000000

1590 rows × 5 columns

Annual rate of change of the refugee stock:

	Area Destination	Country code	Type of data	Year	Annual rate of change of the refugee stock
0	WORLD	900	NaN	1990-1995	-2.123497
1	Developed regions	901	NaN	1990-1995	9.388424
2	Developing regions	902	NaN	1990-1995	-2.839417
3	Least developed countries	941	NaN	1990-1995	-0.680327
4	Less developed regions excluding least develop...	934	NaN	1990-1995	-4.383600
...	...	...	...	...	...
1320	Samoa	882	B	2010-2015	NaN
1321	Tokelau	772	B	2010-2015	NaN
1322	Tonga	776	B	2010-2015	NaN
1323	Tuvalu	798	C	2010-2015	NaN
1324	Wallis and Futuna Islands	876	B	2010-2015	NaN

1325 rows × 5 columns

3. Merge tables and save datasets

Since a single observational unit must be in one table, following tidy data principle 5 a single observational unit must be in one table. I merged six tables into 2 .csv files to help with future analysis.

3.1 dataset_1

In Table 1, Table 2, Table 3, Table 4, Table estimated refugee stock at mid-year (both sexes), and Table refugees as a percentage of the international migrant stock the first five columns are the same, and the number of rows is the same. Although Table 2 in the original file does not have "Type of data", a complete table can still be generated by merging the tables. By using *.merge* function 5 times, the dataset_1 will have 16 columns and 1590 rows. In the end, I use *.to_csv* to create a new .csv file to store this table.

	Area Destination	Country code	Type of data	Year	International Migrant Stock of Female	International Migrant Stock of Both Sexes	International Migrant Stock of Male	Total population at mid-year of Female	Total population at mid-year of Both Sexes	Total population at mid-year of Male	International migrant stock as a percentage of Female	International migrant stock as a percentage of Both Sexes	Internationa migrar stock as percentag of Mal
0	WORLD	900	NaN	1990	74815702.0	152563212.0	77747510.0	2639243.998	5309667.699	2670423.701	2.834740	2.873310	2.91143
1	Developed regions	901	NaN	1990	42115231.0	82378628.0	40263397.0	589207.436	1144463.062	555255.626	7.147777	7.198015	7.25132
2	Developing regions	902	NaN	1990	32700471.0	70184584.0	37484113.0	2050036.562	4165204.637	2115168.075	1.595116	1.685021	1.77215
3	Least developed countries	941	NaN	1990	5236216.0	11075966.0	5843107.0	256015.073	510057.629	254042.556	2.045276	2.171513	2.30005
4	Less developed regions excluding least develop...	934	NaN	1990	27464255.0	59105261.0	31641006.0	1794021.489	3655147.008	1861125.519	1.530877	1.617042	1.70010
...	...	...	...	...	...	...	...	...	...	...	...	...	.
1585	Samoa	882	B	2015	2460.0	4929.0	2469.0	93.584	193.228	99.644	2.628654	2.550873	2.47782
1586	Tokelau	772	B	2015	254.0	487.0	233.0	NaN	1.250	NaN	NaN	38.960000	Nal
1587	Tonga	776	B	2015	2604.0	5731.0	3127.0	52.931	106.170	53.239	4.919612	5.397947	5.87351
1588	Tuvalu	798	C	2015	63.0	141.0	78.0	NaN	9.916	NaN	NaN	1.421944	Nal
1589	Wallis and Futuna Islands	876	B	2015	1411.0	2849.0	1438.0	NaN	13.151	NaN	NaN	21.663752	Nal

1590 rows × 16 columns

3.2 dataset_2

In Table 5 and Table annual rate of change of the refugee stock, the first five columns are the same and the number of rows is the same. By using *.merge* function, the dataset_2 will have 8 columns and 1325 rows. In the end, I used *.to_csv* to create a new .csv file to store this table.

	Area Destination	Country code	Type of data	Year	Annual Rate of Change in Female	Annual Rate of Change in Both Sexes	Annual Rate of Change in Male	Annual rate of change of the refugee stock
0	WORLD	900	NaN	1990-1995	1.104667	1.051865	1.000922	-2.123497
1	Developed regions	901	NaN	1990-1995	2.285643	2.275847	2.265595	9.388424
2	Developing regions	902	NaN	1990-1995	-0.526904	-0.487389	-0.452980	-2.839417
3	Least developed countries	941	NaN	1990-1995	1.249146	1.118175	1.000073	-0.680327
4	Less developed regions excluding least develop...	934	NaN	1990-1995	-0.884180	-0.803244	-0.733256	-4.383600
...	...	...	...	...	...	...	...	...
1320	Samoa	882	B	2010-2015	-0.545343	-0.768177	-0.987758	NaN
1321	Tokelau	772	B	2010-2015	2.603250	2.536144	2.463246	NaN
1322	Tonga	776	B	2010-2015	2.526318	2.641235	2.737439	NaN
1323	Tuvalu	798	C	2010-2015	-1.819436	-1.763854	-1.718849	NaN
1324	Wallis and Futuna Islands	876	B	2010-2015	0.516899	0.519140	0.521340	NaN

1325 rows × 8 columns

4.What I have learned

This is my first time dealing with such an amount of data. Firstly, before tidying the data I should know this data set very well, go through all the tables in the file and understand each table and variable's meaning. Secondly, be very careful with missing and invalid data. For example, I replaced all ".." with 0 at the beginning, which is a wrong cause in this data set 0 has meaning. I learn to be very familiar with the 5 principles and familiar with the application of different functions.

5. Conclusion

The dataset was originally about trends in the stock of international migrants, and it

started with 6 tables for recording and classification reasons. Since some tables have the variable "year" and some have variables "range of year", I store the data into two datasets. After I cleaned and tidied the data, I finally integrated 6 excel sheets into 2 csv files. The purpose of this is to make it more convenient for future data analysis, I can use single .csv which is easier in python environment instead of .xlsx with multiple sheets.