UNIVERSITY OF TORONTO

Faculty of Information

Data Cleaning on Trends in International Migrant Stock

Jinze Li (1002098829)

Master of Information

INF1340: Programming for Data Science

Midterm

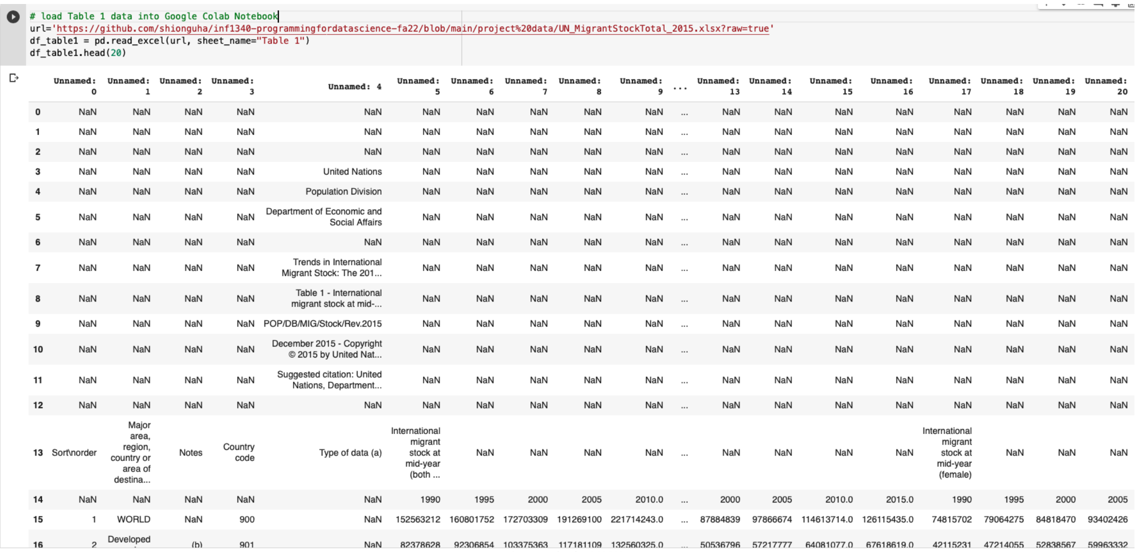
Nov 16th, 2022

Introduction

This document in general reflects the trends in international migrant stock ranging from 1990-2015. It measures the data in different categories including location and gender. It has six excel sheets which contains information including international migrant stock, total population, international migrant stock as a percentage of the total population, female migrants as a percentage of the international migraine stock, annual rate of change of the migrant stock, and estimated refugee stock. The objective of this project is to clean these sets of data following the tidy data principles. The end result should be ready for data visualization.

Methods and Results

Table 1 to 6 are different measurements on this topic. The process and methods used in cleaning these six tables are similar. Therefore, this report will focus on explaining the logic behind data cleaning on table 1. After loading table one into Google Colab Notebook, as shown in figure 1, many cells are showing 'NaN'. According to tidy data principle #2, that column names need to be informative, I decided to get started by deleting column 'Notes' and 'Type of data' as they are not relevant or informative and the measurements of those are not variables. Following the same idea, the first 13 rows were also deleted. The results are shown in figure 2.

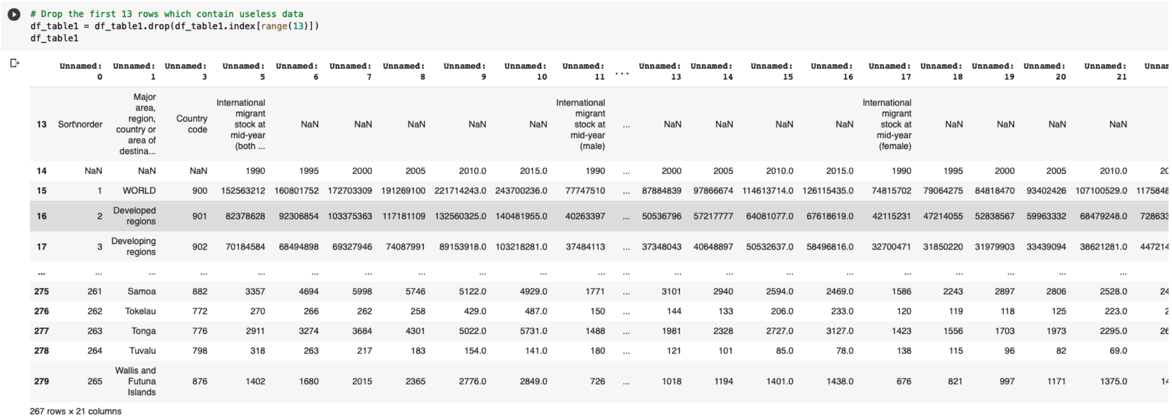


```
# load Table 1 data into Google Colab Notebook
url='https://github.com/shionguha/inf1340-programmingfordatascience-fa22/blob/main/project%20data/UN_MigrantStockTotal_2015.xlsx?raw=true'
df_table1 = pd.read_excel(url, sheet_name="Table 1")
df_table1.head(20)
```

	Unnamed: 0	Unnamed: 1	Unnamed: 2	Unnamed: 3	Unnamed: 4	Unnamed: 5	Unnamed: 6	Unnamed: 7	Unnamed: 8	Unnamed: 9	...	Unnamed: 13	Unnamed: 14	Unnamed: 15	Unnamed: 16	Unnamed: 17	Unnamed: 18	Unnamed: 19	Unnamed: 20
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	NaN	NaN	NaN	NaN	United Nations	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	NaN	NaN	NaN	NaN	Population Division	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5	NaN	NaN	NaN	NaN	Department of Economic and Social Affairs	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
6	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
7	NaN	NaN	NaN	NaN	Trends in International Migrant Stock: The 201...	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
8	NaN	NaN	NaN	NaN	Table 1 - International migrant stock at mid-...	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9	NaN	NaN	NaN	NaN	POP/DB/MIG/Stock/Rev.2015	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
10	NaN	NaN	NaN	NaN	December 2015 - Copyright © 2015 by United Nat...	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
11	NaN	NaN	NaN	NaN	Suggested citation: United Nations, Department...	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
12	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
13	Sort\norder	Major area, region, country or area of destina...	Notes	Country code	Type of data (a)	International migrant stock at mid-year (both ...	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	International migrant stock at mid-year (female)	NaN	NaN	NaN
14	NaN	NaN	NaN	NaN	NaN	1990	1995	2000	2005	2010.0	...	2000	2005	2010.0	2015.0	1990	1995	2000	2005
15	1	WORLD	NaN	900	NaN	152563212	160801752	172703309	191269100	221714243.0	...	87884839	97866674	114613714.0	126115435.0	74815702	79064275	84818470	93402426
16	2	Developed	(b)	901	NaN	82378628	92306854	103375363	117181109	132560325.0	...	50536796	57217777	64081077.0	67618619.0	42115231	47214055	52838567	59963332

✓ 0s completed at 8:09 PM

Figure 1. Loading Data



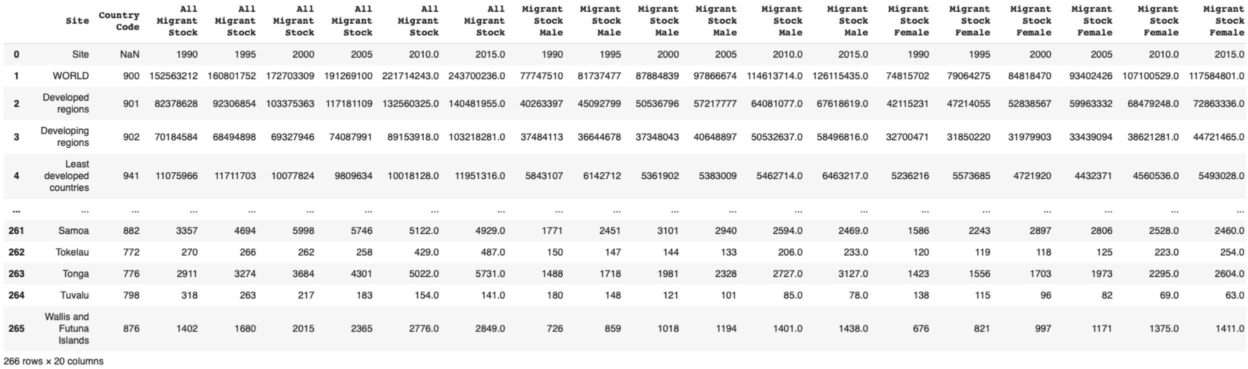
```
# Drop the first 13 rows which contain useless data
df_table1 = df_table1.drop(df_table1.index[range(13)])
df_table1
```

	Unnamed: 0	Unnamed: 1	Unnamed: 3	Unnamed: 5	Unnamed: 6	Unnamed: 7	Unnamed: 8	Unnamed: 9	Unnamed: 10	Unnamed: 11	...	Unnamed: 13	Unnamed: 14	Unnamed: 15	Unnamed: 16	Unnamed: 17	Unnamed: 18	Unnamed: 19	Unnamed: 20	Unnamed: 21	Unnam
13	Sort\norder	Major area, region, country or area of destina...	Country code	International migrant stock at mid-year (both ...	NaN	NaN	NaN	NaN	NaN	International migrant stock at mid-year (male)	...	NaN	NaN	NaN	NaN	International migrant stock at mid-year (female)	NaN	NaN	NaN	NaN	
14	NaN	NaN	NaN	1990	1995	2000	2005	2010.0	2015.0	1990	...	2000	2005	2010.0	2015.0	1990	1995	2000	2005	2010.0	20
15	1	WORLD	900	152563212	160801752	172703309	191269100	221714243.0	243700236.0	77747510	...	87884839	97866674	114613714.0	126115435.0	74815702	79064275	84818470	93402426	107100529.0	1175848
16	2	Developed regions	901	82378628	92306854	103375363	117181109	132560325.0	140481955.0	40263397	...	50536796	57217777	64081077.0	67618619.0	42115231	47214055	52838567	59963332	68479248.0	728633
17	3	Developing regions	902	70184584	68494898	69327946	74087991	89153918.0	103218281.0	37484113	...	37348043	40648897	50532637.0	58496816.0	32700471	31850220	31979903	33439094	38621281.0	447214
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
275	261	Samoa	882	3357	4694	5998	5746	5122.0	4929.0	1771	...	3101	2940	2594.0	2469.0	1586	2243	2897	2806	2528.0	24
276	262	Tokelau	772	270	266	262	258	429.0	487.0	150	...	144	133	206.0	233.0	120	119	118	125	223.0	2
277	263	Tonga	776	2911	3274	3684	4301	5022.0	5731.0	1488	...	1981	2328	2727.0	3127.0	1423	1556	1703	1973	2295.0	26
278	264	Tuvalu	798	318	263	217	183	154.0	141.0	180	...	121	101	85.0	78.0	138	115	96	82	69.0	
279	265	Wallis and Futuna Islands	876	1402	1680	2015	2365	2776.0	2849.0	726	...	1018	1194	1401.0	1438.0	676	821	997	1171	1375.0	14

267 rows × 21 columns

Figure 2. Table after deleting irrelevant rows and columns

Another issue as shown in figure 2 is that when the data was loaded from excel to google colab notebook, the merged cells got separated and the column headers became a mess. And the headers as shown in figure 2 appeared as unnamed. Therefore, I got started by assigning names to the header of each column so that I can delete the first row which was originally the header of each column. The result is shown in figure 3.

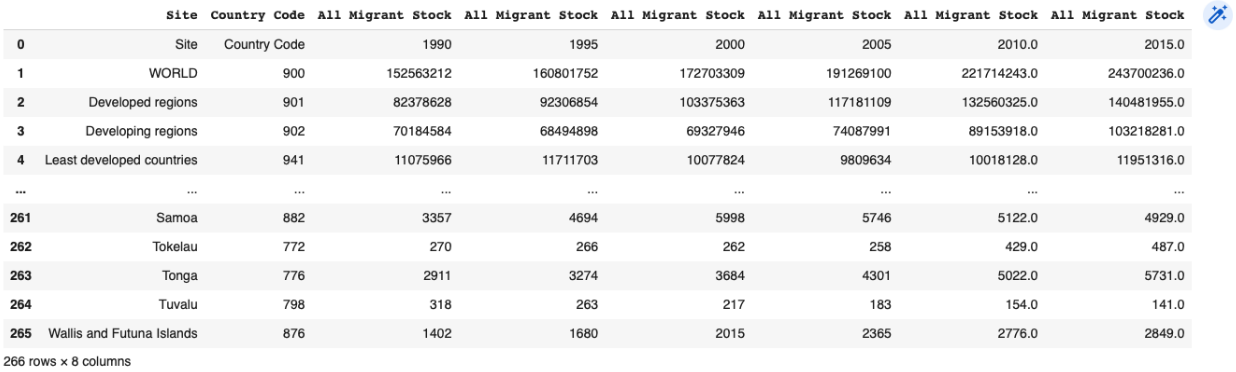


	Site	Country Code	All Migrant Stock	All Migrant Stock	All Migrant Stock	All Migrant Stock	All Migrant Stock	All Migrant Stock	Migrant Stock Male	Migrant Stock Male	Migrant Stock Male	Migrant Stock Male	Migrant Stock Male	Migrant Stock Male	Migrant Stock Female	Migrant Stock Female	Migrant Stock Female	Migrant Stock Female	Migrant Stock Female	Migrant Stock Female
0	Site	NaN	1990	1995	2000	2005	2010.0	2015.0	1990	1995	2000	2005	2010.0	2015.0	1990	1995	2000	2005	2010.0	2015.0
1	WORLD	900	152563212	160801752	172703309	191269100	221714243.0	243700236.0	77747510	81737477	87884839	97866674	114613714.0	126115435.0	74815702	79064275	84818470	93402426	107100529.0	117584801.0
2	Developed regions	901	82378628	92306854	103375363	117181109	132560325.0	140481955.0	40263397	45092799	50536796	57217777	64081077.0	67618619.0	42115231	47214055	52838567	59963332	68479248.0	72863336.0
3	Developing regions	902	70184584	68494898	69327946	74087991	89153918.0	103218281.0	37484113	36644678	37348043	40648897	50532637.0	58496816.0	32700471	31850220	31979903	33439094	38621281.0	44721465.0
4	Least developed countries	941	11075966	11711703	10077824	9809634	10018128.0	11951316.0	5843107	6142712	5361902	5383009	5462714.0	6463217.0	5236216	5573685	4721920	4432371	4560536.0	5493028.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
261	Samoa	882	3357	4694	5998	5746	5122.0	4929.0	1771	2451	3101	2940	2594.0	2469.0	1586	2243	2897	2806	2528.0	2460.0
262	Tokelau	772	270	266	262	258	429.0	487.0	150	147	144	133	206.0	233.0	120	119	118	125	223.0	254.0
263	Tonga	776	2911	3274	3684	4301	5022.0	5731.0	1488	1718	1981	2328	2727.0	3127.0	1423	1556	1703	1973	2295.0	2604.0
264	Tuvalu	798	318	263	217	183	154.0	141.0	180	148	121	101	85.0	78.0	138	115	96	82	69.0	63.0
265	Wallis and Futuna Islands	876	1402	1680	2015	2365	2776.0	2849.0	726	859	1018	1194	1401.0	1438.0	676	821	997	1171	1375.0	1411.0

266 rows × 20 columns

Figure 3. Table after renaming the columns

At this step, I have a more cleaned table set. However, I realized that I have two rows of columns which is unacceptable as each cell should be a single measurement. The first row contains only three measurements that are all migrant stock, migrant stock male and migrant stock female. The second row of headers contains year information which is not acceptable as 1990-2015 are variables, but column names can not be values. 1990-2015 needs to be in the cell and the column name needs to be year. Ideally, I should have year as one column, and all migrant stock count, migrant stock male count and migrant stock female count as the other three columns. To make things easier for myself, I decided to separate the table into three tables for now so that I could fix the year column first and combine the three tables at the end. Therefore, as shown in figure 4, I get started by fixing the all migrant stock count table.

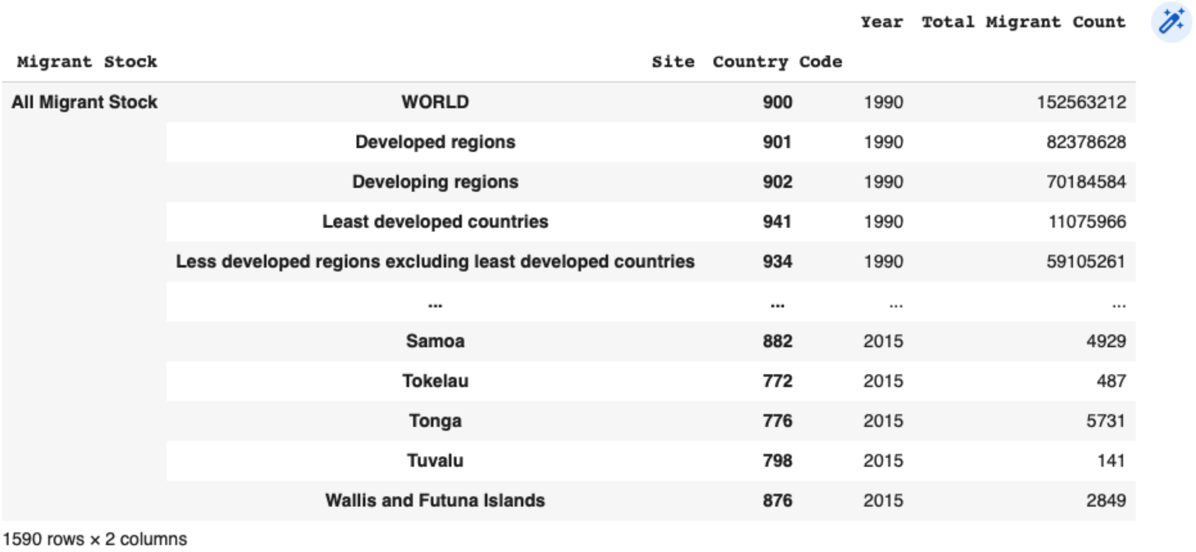


	Site	Country Code	All Migrant Stock	All Migrant Stock	All Migrant Stock	All Migrant Stock	All Migrant Stock	All Migrant Stock
0	Site	Country Code	1990	1995	2000	2005	2010.0	2015.0
1	WORLD	900	152563212	160801752	172703309	191269100	221714243.0	243700236.0
2	Developed regions	901	82378628	92306854	103375363	117181109	132560325.0	140481955.0
3	Developing regions	902	70184584	68494898	69327946	74087991	89153918.0	103218281.0
4	Least developed countries	941	11075966	11711703	10077824	9809634	10018128.0	11951316.0
...	...	...	...	...	...	...	...	...
261	Samoa	882	3357	4694	5998	5746	5122.0	4929.0
262	Tokelau	772	270	266	262	258	429.0	487.0
263	Tonga	776	2911	3274	3684	4301	5022.0	5731.0
264	Tuvalu	798	318	263	217	183	154.0	141.0
265	Wallis and Futuna Islands	876	1402	1680	2015	2365	2776.0	2849.0

266 rows × 8 columns

Figure 4. Separate table for all migrant stock

After separating the table, things are getting easier, the first row contains the same information so that I could delete it and promote the number of years as the temporary headers. After formatting the data, I moved the number of years from columns to the cells. The table after styling is shown in figure 5.

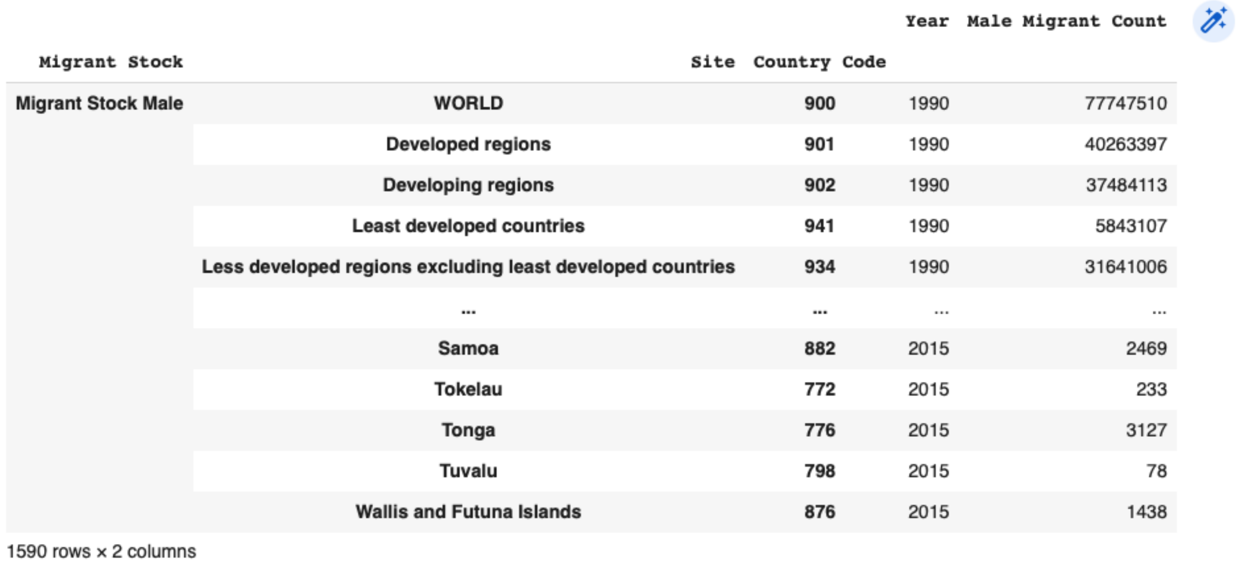


			Year	Total Migrant Count
Migrant Stock	Site	Country Code		
All Migrant Stock	WORLD	900	1990	152563212
	Developed regions	901	1990	82378628
	Developing regions	902	1990	70184584
	Least developed countries	941	1990	11075966
	Less developed regions excluding least developed countries	934	1990	59105261
	...	...	...	...
	Samoa	882	2015	4929
	Tokelau	772	2015	487
	Tonga	776	2015	5731
	Tuvalu	798	2015	141
	Wallis and Futuna Islands	876	2015	2849

1590 rows × 2 columns

Figure 5. Finished table for total migrant count

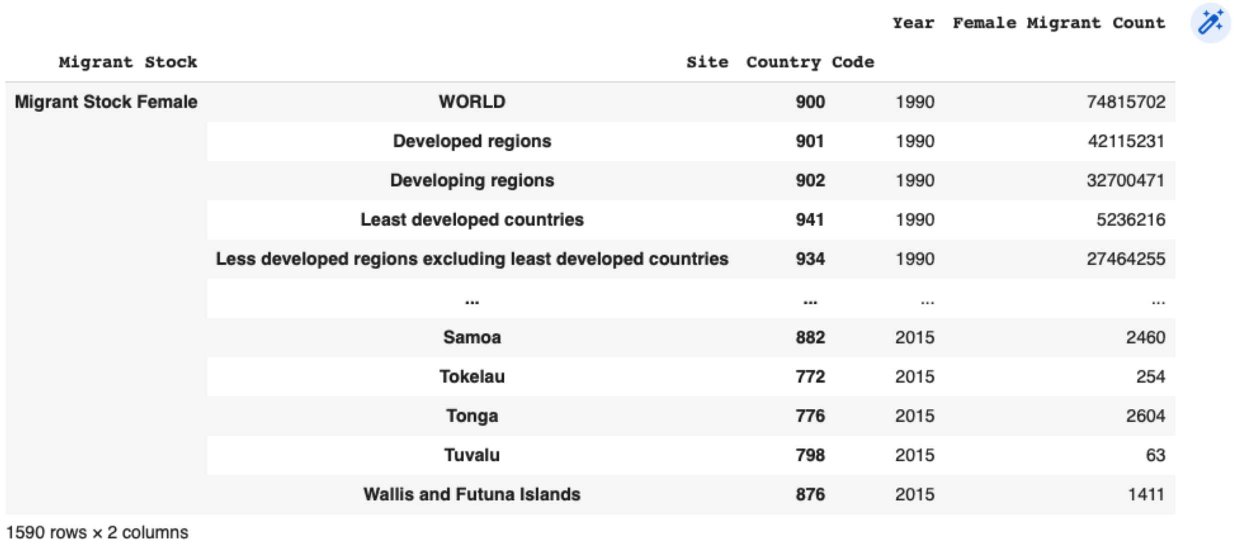
After successfully cleaning the table for total migrant count, I followed the same process and methods on cleaning female migrant count and male migrant count. The finished table for male migrant count and female migrant count are shown in figure 6 and 7.



			Year	Male Migrant Count
Migrant Stock	Site	Country Code		
Migrant Stock Male	WORLD	900	1990	77747510
	Developed regions	901	1990	40263397
	Developing regions	902	1990	37484113
	Least developed countries	941	1990	5843107
	Less developed regions excluding least developed countries	934	1990	31641006
	...	...	...	...
	Samoa	882	2015	2469
	Tokelau	772	2015	233
	Tonga	776	2015	3127
	Tuvalu	798	2015	78
	Wallis and Futuna Islands	876	2015	1438

1590 rows × 2 columns

Figure 6. Finished table for male migrant count

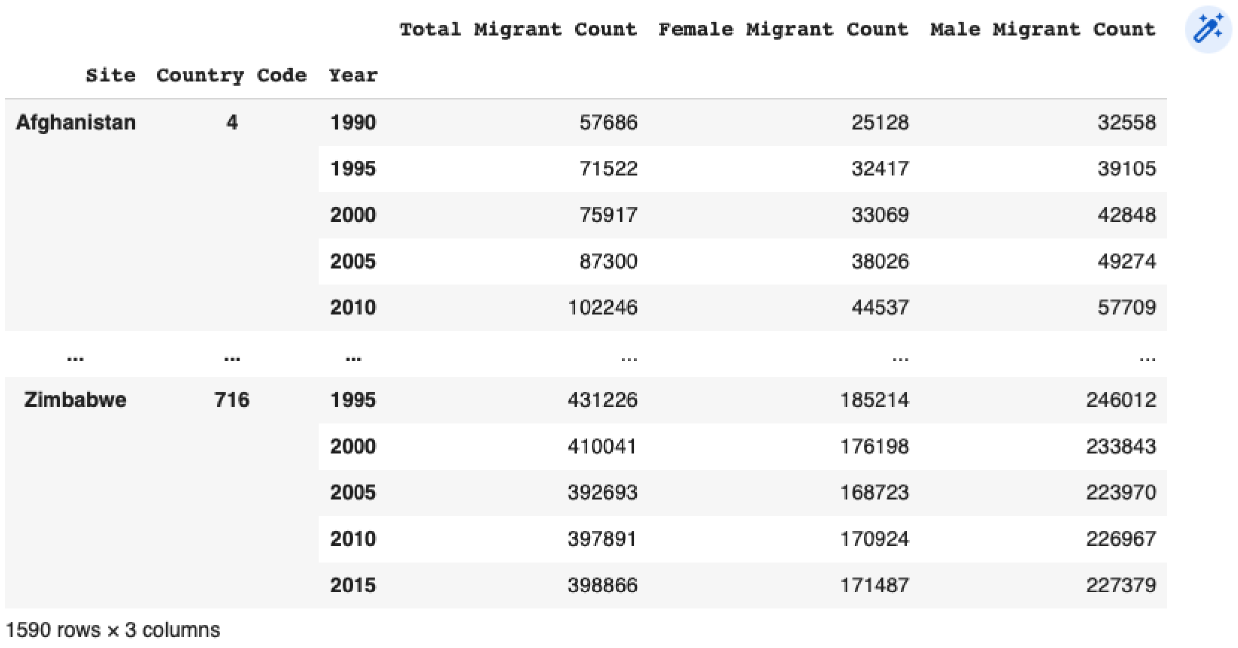


			Year	Female Migrant Count
Migrant Stock	Site	Country Code		
Migrant Stock Female	WORLD	900	1990	74815702
	Developed regions	901	1990	42115231
	Developing regions	902	1990	32700471
	Least developed countries	941	1990	5236216
	Less developed regions excluding least developed countries	934	1990	27464255
	...	...	...	...
	Samoa	882	2015	2460
	Tokelau	772	2015	254
	Tonga	776	2015	2604
	Tuvalu	798	2015	63
	Wallis and Futuna Islands	876	2015	1411

1590 rows × 2 columns

Figure 7. Finished table for female migrant count

After having three cleaned tables of data, I start pivoting each table to summarize the data for each site. After having readable and analyzable tables, I combined them as one table as shown in figure 8.



			Total Migrant Count	Female Migrant Count	Male Migrant Count
Site	**Country Code**	**Year**			
Afghanistan	**4**	**1990**	57686	25128	32558
		1995	71522	32417	39105
		2000	75917	33069	42848
		2005	87300	38026	49274
		2010	102246	44537	57709
...	**...**	**...**	...	...	...
Zimbabwe	**716**	**1995**	431226	185214	246012
		2000	410041	176198	233843
		2005	392693	168723	223970
		2010	397891	170924	226967
		2015	398866	171487	227379

1590 rows × 3 columns

Figure 8. Finished table 1

Following similar steps, tables 2 to 6 were also cleaned into similar formats as the one shown above.

Discussion

The end results are data sets that are easy to read and understand. These finished data sets are also ready for data visualization. The steps getting to the end result might not be the best ones, but they do the job. The next step will be to find ways of integrating these six tables. Also, separating the locations by developed and developing countries, etc. might provide us with more insights on the end result.