INF 1340 Midterm Project

Yidan Chen
Due Date: Nov 16, 2022

Introduction

Data cleaning, often one of the first and the most important step for a project, is the process of improving the quality of the dataset and setting up the data for further analysis. For this project, I adopted the tidy data method and implement the five principles of tidy data to clean and structure the dataset to facilitate analysis. The dataset used for cleaning was collected by the United Nations, the population division, and the department of economics and social affairs. The entire dataset includes 6 tables, and it contains the data and trends in international migrant and refugee stock from 1990-2015. The programming language used for cleaning was Python.

Method and Process

Table 1

To start with, I read the dataset on google collab and import the pandas, numpy, matplotlib, seaborn, plotly libraries for data exploration and cleaning. Beginning with table 1, I noticed that the actual data starts at row 17, and both row 15 and row 16 is the column name. Instead of un-merged the two rows, I decided to skip the first 16 rows and rename the column name, I also noticed that the missing values are represented by "..", so I assigned all the ".." as "na" value for future use.

Problem 1: Multiple types of data in one table
By observing the table, I noticed that there are different types of data in 1 table, for example, the "both sexes" category means the summation of both female and male populations, and the female and male populations represent the data for each category. So this violates the **tidy data principle #4**: "each table needs to have a singular data type".
Problem 2: Each row is not a unique observation
Meanwhile, this table also violates the **tidy data principle 1:** "rows are unique observations, each row represents a unique element", by further observing the dataset, I noticed that each row does not contain a unique element of the dataset, it includes the data of "both sexes", "female", and "male" categories. So because of the violations of two principles, we need to first split the main table into three sections for "both sexes", "female", and "male" categories.

```
table1_both_sexes.head(10)
```

	Major area, region, country or area of destination	Notes	Country Code	Type of data(a)	1990	1995	2000	2005	2010	2015
15	WORLD	NaN	900	NaN	152563212	160801752	172703309	191269100	221714243.0	243700236.0
16	Developed regions	(b)	901	NaN	82378628	92306854	103375363	117181109	132560325.0	140481955.0
17	Developing regions	(c)	902	NaN	70184584	68494898	69327946	74087991	89153918.0	103218281.0
18	Least developed countries	(d)	941	NaN	11075966	11711703	10077824	9809634	10018128.0	11951316.0
19	Less developed regions excluding least develop...	NaN	934	NaN	59105261	56778501	59244124	64272611	79130668.0	91262036.0
20	Sub-Saharan Africa	(e)	947	NaN	14690319	15324570	13716539	13951086	15496764.0	18993986.0

Table 1: International migrant stock at mid-year (both sexes)

```
table1_male.head()
```

	Major area, region, country or area of destination	Notes	Country Code	Type of data(a)	1990	1995	2000	2005	2010	2015
15	WORLD	NaN	900	NaN	77747510	81737477.0	87884839.0	97866674.0	114613714.0	126115435.0
16	Developed regions	(b)	901	NaN	40263397	45092799.0	50536796.0	57217777.0	64081077.0	67618619.0
17	Developing regions	(c)	902	NaN	37484113	36644678.0	37348043.0	40648897.0	50532637.0	58496816.0
18	Least developed countries	(d)	941	NaN	5843107	6142712.0	5361902.0	5383009.0	5462714.0	6463217.0
19	Less developed regions excluding least develop...	NaN	934	NaN	31641006	30501966.0	31986141.0	35265888.0	45069923.0	52033599.0

Table 2: International migrant stock at mid-year (male)

	Major area, region, country or area of destination	Notes	Country Code	Type of data(a)	1990	1995	2000	2005	2010	2015
15	WORLD	NaN	900	NaN	74815702	79064275.0	84818470.0	93402426.0	107100529.0	117584801.0
16	Developed regions	(b)	901	NaN	42115231	47214055.0	52838567.0	59963332.0	68479248.0	72863336.0
17	Developing regions	(c)	902	NaN	32700471	31850220.0	31979903.0	33439094.0	38621281.0	44721465.0
18	Least developed countries	(d)	941	NaN	5236216	5573685.0	4721920.0	4432371.0	4560536.0	5493028.0
19	Less developed regions excluding least develop...	NaN	934	NaN	27464255	26276535.0	27257983.0	29006723.0	34060745.0	39228437.0

Table 3: International migrant stock at mid-year (female)

Problem 3: Variables are stored in columns, not cells

Problem 4: Column names are values(1990,1995,2000...), not informative

By observing the first three tables, I noticed that the variables "1990,1995,2000…2015" are stored as column names, instead of cells, so it firstly violates **tidy data principle 3**, which is "variables need to be in cells, not rows and columns".

Meanwhile, I also noticed that it also violates **tidy data principle 2**, which is "Column names need to be informative, variable names and not values", so we need to use the melt function and unpivot the dataframe from wide to long format, putting all the individual variables into cells, and change the column name as "Year". To use the melt function, I decided on some elements that keep unchanged, and set as them as id_vars1, which includes 'Major area, region, country or area of destination',' Notes', 'Country Code', 'Type of data(a)'; the var_name would be 'year', and the value name is 'International migrant stock'. Then after melt function, we come up with three tables as follows.

	Major area, region, country or area of destination	Notes	Country Code	Type of data(a)	Year	International migrant stock at mid-year (both sexes)
0	WORLD	NaN	900	NaN	1990	152563212
1	Developed regions	(b)	901	NaN	1990	82378628
2	Developing regions	(c)	902	NaN	1990	70184584
3	Least developed countries	(d)	941	NaN	1990	11075966
4	Less developed regions excluding least develop...	NaN	934	NaN	1990	59105261
5	Sub-Saharan Africa	(e)	947	NaN	1990	14690319

Table 4: Tidy_Year_International migrant stock at mid-year (both sexes)

	Major area, region, country or area of destination	Notes	Country Code	Type of data(a)	Year	International migrant stock at mid-year(male)
0	WORLD	NaN	900	NaN	1990	77747510
1	Developed regions	(b)	901	NaN	1990	40263397
2	Developing regions	(c)	902	NaN	1990	37484113
3	Least developed countries	(d)	941	NaN	1990	5843107
4	Less developed regions excluding least develop...	NaN	934	NaN	1990	31641006

Table 5: Tidy_Year_International migrant stock at mid-year (male)

	Major area, region, country or area of destination	Notes	Country Code	Type of data(a)	Year	International migrant stock at mid-year(female)
0	WORLD	NaN	900	NaN	1990	74815702
1	Developed regions	(b)	901	NaN	1990	42115231
2	Developing regions	(c)	902	NaN	1990	32700471
3	Least developed countries	(d)	941	NaN	1990	5236216
4	Less developed regions excluding least develop...	NaN	934	NaN	1990	27464255

Table 6: Tidy_Year_International migrant stock at mid-year (female)

Problem 5: One type of observational unit in different tables.
According to **tidy data principle 5**, each type of observational unit should be stored in its own table, we can consider female/male as the same data type(individual sex data), which is different from both sexes(the sum of male and female individual sex data), so we can combine the data type with the same observational unit into the same table. The next step is to merge table 5 and table 6 together into a new table 7.

	Major area, region, country or area of destination	Notes	Country Code	Type of data(a)	Year	International migrant stock at mid-year(male)	International migrant stock at mid-year(female)
0	WORLD	NaN	900	NaN	1990	77747510	74815702
1	Developed regions	(b)	901	NaN	1990	40263397	42115231
2	Developing regions	(c)	902	NaN	1990	37484113	32700471
3	Least developed countries	(d)	941	NaN	1990	5843107	5236216
4	Less developed regions excluding least develop...	NaN	934	NaN	1990	31641006	27464255

Table 7: Tidy1_Year_International migrant stock at mid-year (sex1)

Problem 6: There are two observations in one row
After merging two tables together, I noticed that table 7 violates **tidy data principle 1:** "rows are unique observations, each row represents a unique element", so I used the melt function to change the table from wide to long format. As Professor taught during class, gender and political opinions are hard to categorize, so researchers usually use sex instead of gender. Therefore, I renamed the new column "sex".

	Major area, region, country or area of destination	Notes	Country Code	Type of data(a)	Year	Gender	International migrant stock at mid-year
0	WORLD	NaN	900	NaN	1990	Male	77747510.0
1	Developed regions	(b)	901	NaN	1990	Male	40263397.0
2	Developing regions	(c)	902	NaN	1990	Male	37484113.0
3	Least developed countries	(d)	941	NaN	1990	Male	5843107.0
4	Less developed regions excluding least develop...	NaN	934	NaN	1990	Male	31641006.0

Table 8: Tidy1_Year_International migrant stock at mid-year (sex2)

Problem 7: There are multiple variables(major area, region, country) stored in 1 column
By observing table 8, I noticed a new problem that violates **tidy data principle 2:** "each column needs to consist of one and only one variable"**.** Therefore, we need to split table 8 into three sub tables by country, region, and major area. To do that, we need to firstly refer to the "Annex" table and organize all country names, region names, and major area names. Because there are duplicates in the region names and major area names column in table "Annex", I dropped the duplicates and make sure that each name is not repeated.

```
#orgnize all country names
country=df_an['Country or area']
#orgnize all major area names
maj_area=df_an['Major area']
maj_area.drop_duplicates(keep='first',inplace=True)
#orgnize all region names
region=df_an['Region']
region.drop_duplicates(keep='first',inplace=True)
```

Process 1: code to organize all the categories in table “Annex”

In order to match each category to an element in table 8, I decided to use the **select row** function and use the **“country” “maj_area” and “region”** as the indicator and select the rows that match the category names from table 8. Because this function will be used repeatedly in the future, for efficiency purposes, I defined the “select_row” function as below

```
#define selectrow function for future use
def selectrow(table,column_name,row_name):
  new_tabl=table[table[column_name].isin(row_name)]
  return new_tabl
```

Process 2: code to select rows for the new table if the column name matches the row name

Rename column: I also rename the new columns accordingly as “Country”, “Major area” and “Region” insead of the original table 'Major area, region, country or area of destination'.

```
#The first section, the international migrants stock at mid-year by Gender(male/female) in different countries
tidy_gender_country=selectrow(tidy_gender1,column_name="Major area, region, country or area of destination", row_name=country)
tidy_gender_country1=tidy_gender_country.rename(columns={"Major area, region, country or area of destination": "Country"}, inplace=True)
```

Process 3: code to rename the column

Reset Default: I noticed that the default index starts with 0 instead of 1, so I reset the index to start with 1.

```
#Problem: the index section needs to be reset
#Define the reset index function for future use
def re_index(table):
  table=table.reset_index()
  table.index=table.index+1
  table=table.drop(columns=['index'])
  return table
```

Process 4: code to reset the index and start with 1

Drop missing value: For missing data, I decided to drop the row if there is any missing value, for example, the total population for both sexes in 1990 is missing for South Sudan because there will be no meaning if we still keep the row if there are no values in the last column, I decided to drop the entire row if the value is missing in the “International migrant stock at mid-year” column.

Drop “Type of data” for Major area and region: By observing the original table 1, I noticed that both major area and region data don’t have any data types, so I decided to drop the entire column for the table of major area and region and keep the column only for the country table.

```
#Drop missing data
tidy_gender_maj_area1 = tidy_gender_maj_area.dropna(subset=["International migrant stock at mid-year"])
#Drop type of data for major area
tidy_gender_maj_area1=tidy_gender_maj_area1.drop(columns=['Type of data(a)'])
```

Process 5: code to drop missing value and “type of data” for region

<u>**Now the three tables in the following are tidy datasets for table 1, male and female:**</u>

- Columns are unique measurements
- Rows that are unique observations
- Rows*columns = observation unit

	Country	Notes	Country Code	Type of data(a)	Year	Sex	International migrant stock at mid-year
1	Burundi	NaN	108	B R	1990	Male	163267.0
2	Comoros	NaN	174	B	1990	Male	6717.0
3	Djibouti	NaN	262	B R	1990	Male	64242.0
4	Eritrea	NaN	232	I	1990	Male	6228.0
5	Ethiopia	NaN	231	B R	1990	Male	607284.0

Table 8: Tidy1_Country_International migrant stock at mid-year (sex)

	Major Area	Notes	Country Code	Year	Sex	International migrant stock at mid-year
1	Africa	NaN	903	1990	Male	8279564.0
2	Asia	NaN	935	1990	Male	26011875.0
3	Europe	NaN	908	1990	Male	23946673.0
4	Latin America and the Caribbean	NaN	904	1990	Male	3597037.0
5	Northern America	NaN	905	1990	Male	13497319.0

Table 9: Tidy1_Maj_area_International migrant stock at mid-year (sex)

	Region	Notes	Country Code	Year	Sex	International migrant stock at mid-year
1	Eastern Africa	NaN	910	1990	Male	3071189.0
2	Middle Africa	NaN	911	1990	Male	744494.0
3	Northern Africa	NaN	912	1990	Male	1230643.0
4	Southern Africa	NaN	913	1990	Male	840899.0

Table 10: Tidy1_counrty_International migrant stock at mid-year (sex)

After we tidy data cleaning for both the female and male categories, we need to proceed with the same cleaning process for the “both sexes” category. The following three tables are tidy datasets for table1, the sum of both sexes

<u>Now the three tables in the following are tidy datasets for table one, both sexes:</u>

- Columns are unique measurements
- Rows that are unique observations
- Rows*columns = observation unit

	Country	Notes	Country Code	Type of data(a)	Year	International migrant stock at mid-year (both sexes)
1	Burundi	NaN	108	B R	1990	333110
2	Comoros	NaN	174	B	1990	14079
3	Djibouti	NaN	262	B R	1990	122221
4	Eritrea	NaN	232	I	1990	11848
5	Ethiopia	NaN	231	B R	1990	1155390

Table 11: *Tidy1_Country_International migrant stock at mid-year (sum_sex)*

	Major Area	Notes	Country Code	Year	International migrant stock at mid-year (both sexes)
1	Africa	NaN	903	1990	15690623
2	Asia	NaN	935	1990	48142261
3	Europe	NaN	908	1990	49219200
4	Latin America and the Caribbean	NaN	904	1990	7169728
5	Northern America	NaN	905	1990	27610542

Table 12: Tidy1_maj_area_International migrant stock at mid-year (sum_sex)

	Region	Notes	Country Code	Year	International migrant stock at mid-year (both sexes)
1	Eastern Africa	NaN	910	1990	5964031
2	Middle Africa	NaN	911	1990	1460530
3	Northern Africa	NaN	912	1990	2403200
4	Southern Africa	NaN	913	1990	1392359
5	Western Africa	NaN	914	1990	4470503

Table 13: Tidy1_maj_area_International migrant stock at mid-year (sum_sex)

Table 2

By observing table 2, we noticed that the only difference between table 1 and table 2 is that table 2 records the total population at mid-year by sex and by major area, region, country or area, 1990-2015 (thousands) and table1 only covers the international migrant stock. So we will follow the same tidy data cleaning step for table 2 and the following are 6 tables according to tidy data principles.

Total Population at mid-year for male and female in Country, Major Area, and Regions

	Country	Notes	Country Code	Year	Sex	Total population at mid-year (thousands)
1	Burundi	NaN	108	1990	Male	2755.028
2	Comoros	NaN	174	1990	Male	208.212
3	Djibouti	NaN	262	1990	Male	295.933
4	Eritrea	NaN	232	1990	Male	1558.486
5	Ethiopia	NaN	231	1990	Male	23965.647

Table 14: Tidy2_Country_Total Population at mid-year (sex)

	Major Area	Notes	Country Code	Year	Sex	Total population at mid-year (thousands)
1	Africa	NaN	903	1990	Male	315071.378
2	Asia	NaN	935	1990	Male	1634734.677
3	Europe	NaN	908	1990	Male	347356.281
4	Latin America and the Caribbean	NaN	904	1990	Male	221989.776
5	Northern America	NaN	905	1990	Male	137757.875

Table 15: Tidy2_Maj_area_Total Population at mid-year (sex)

	Region	Notes	Country Code	Year	Sex	Total population at mid-year (thousands)
1	Eastern Africa	NaN	910	1990	Male	98208.646
2	Middle Africa	NaN	911	1990	Male	35035.128
3	Northern Africa	NaN	912	1990	Male	70480.841
4	Southern Africa	NaN	913	1990	Male	20760.4
5	Western Africa	NaN	914	1990	Male	90586.363

Table 16: Tidy2_Region_Total Population at mid-year (sex)

Total Population at mid-year for sum of both sexes in Country, Major Area, and Regions

	Country	Notes	Country Code	Year	Total population of both sexes at mid-year (thousands)
1	Burundi	NaN	108	1990	5613.141
2	Comoros	NaN	174	1990	415.144
3	Djibouti	NaN	262	1990	588.356
4	Eritrea	NaN	232	1990	3139.083
5	Ethiopia	NaN	231	1990	48057.094

Table 17: Tidy2_Country_Total Population at mid-year (sum_sexes)

	Major Area	Notes	Country Code	Year	Total population of both sexes at mid-year (thousands)
1	Africa	NaN	903	1990	631614.304
2	Asia	NaN	935	1990	3202474.692
3	Europe	NaN	908	1990	721086.311
4	Latin America and the Caribbean	NaN	904	1990	446888.767
5	Northern America	NaN	905	1990	280633.063

Table 18: Tidy2_Maj_area_Total Population at mid-year (sum_sexes)

	Region	Notes	Country Code	Year	Total population of both sexes at mid-year (thousands)
1	Eastern Africa	NaN	910	1990	198231.687
2	Middle Africa	NaN	911	1990	70886.433
3	Northern Africa	NaN	912	1990	140116.613
4	Southern Africa	NaN	913	1990	42049.013
5	Western Africa	NaN	914	1990	180330.558

Table 18: Tidy2_region_Total Population at mid-year (sum_sexes)

Table 3

By observing table 3, we noticed that table 3 represents the International migrant stock as a percentage of the total population by major area, region, country, or area from 1990-2015. It is the result of dividing table 1 to table 2. However, the nature of the datasets is still the same as table 1 and table 2, so we will execute the same tidy data cleaning steps for table 3 and the following are 6 tables after tidy data cleaning.

International migrant stock as a percentage of the total population for males and females in Country, Major Area, and Regions

	Country	Notes	Country Code	Type of data(a)	Year	Sex	International migrant stock as a percentage of the total population
1	Burundi	NaN	108	B R	1990	Male	5.926147
2	Comoros	NaN	174	B	1990	Male	3.226039
3	Djibouti	NaN	262	B R	1990	Male	21.708292
4	Eritrea	NaN	232	I	1990	Male	0.399619
5	Ethiopia	NaN	231	B R	1990	Male	2.533977

Table 19: Tidy3_Country_Interational as % of Total Population at mid-year (sex)

	Major Area	Notes	Country Code	Year	Sex	International migrant stock as a percentage of the total population
1	Africa	NaN	903	1990	Male	2.627838
2	Asia	NaN	935	1990	Male	1.591199
3	Europe	NaN	908	1990	Male	6.89398
4	Latin America and the Caribbean	NaN	904	1990	Male	1.620362
5	Northern America	NaN	905	1990	Male	9.797857

Table 20: Tidy3_Maj_area_Interational as % of Total Population at mid-year (sex)

	Region	Notes	Country Code	Year	Sex	International migrant stock as a percentage of the total population
1	Eastern Africa	NaN	910	1990	Male	3.127208
2	Middle Africa	NaN	911	1990	Male	2.124993
3	Northern Africa	NaN	912	1990	Male	1.746067
4	Southern Africa	NaN	913	1990	Male	4.050495
5	Western Africa	NaN	914	1990	Male	2.640948

Table 21: Tidy3_Region_Interational as % of Total Population at mid-year (sex)

International migrant stock as a percentage of the total population for the sum of both sexes in Country, Major Area, and Regions

	Country	Notes	Country Code	Type of data(a)	Year	International migrant stock as a percentage of the total population(both sexes)
1	Burundi	NaN	108	B R	1990	5.934467
2	Comoros	NaN	174	B	1990	3.391353
3	Djibouti	NaN	262	B R	1990	20.773307
4	Eritrea	NaN	232	I	1990	0.377435
5	Ethiopia	NaN	231	B R	1990	2.404203

Table 22: Tidy3_Country_Interation as % of Total Population at mid-year (sum_sex)

	Major Area	Notes	Country Code	Year	International migrant stock as a percentage of the total population(both sexes)
1	Africa	NaN	903	1990	2.48421
2	Asia	NaN	935	1990	1.503283
3	Europe	NaN	908	1990	6.825702
4	Latin America and the Caribbean	NaN	904	1990	1.604365
5	Northern America	NaN	905	1990	9.838663

Table 23: Tidy3_Maj_area_Interation as % of Total Population at mid-year (sum_sex)

	Region	Notes	Country Code	Year	International migrant stock as a percentage of the total population(both sexes)
1	Eastern Africa	NaN	910	1990	3.008616
2	Middle Africa	NaN	911	1990	2.06038
3	Northern Africa	NaN	912	1990	1.715143
4	Southern Africa	NaN	913	1990	3.311276
5	Western Africa	NaN	914	1990	2.47906

Table 24: Tidy3_Region_Interational as % of Total Population at mid-year (sum_sex)

Table 4

By observing table 4, we can see that it is the dataset that represents the female migrants as a percentage of the international migrant stock from 1990-2015. This is the result of dividing table 6(see table above), which is the International migrant stock for females, by table 4(see table above), which is the sum of International migrant stock for both sexes. Still, the nature of this dataset is still the same as previous tables, so we will use the same logic to proceed with the tidy data cleaning steps.

Female migrants as a percentage of the international migrant stock

	Country	Notes	Country Code	Type of data(a)	Year	Female migrants as a percentage of the international migrant stock
1	Burundi	NaN	108	B R	1990	50.987061
2	Comoros	NaN	174	B	1990	52.290646
3	Djibouti	NaN	262	B R	1990	47.437838
4	Eritrea	NaN	232	I	1990	47.434166
5	Ethiopia	NaN	231	B R	1990	47.439047

Table 25: Tidy4_Country_Female migrants as a percentage of the international migrant stock

	Major Area	Notes	Country Code	Year	Female migrants as a percentage of the international migrant stock
1	Africa	NaN	903	1990	47.232408
2	Asia	NaN	935	1990	45.96873
3	Europe	NaN	908	1990	51.346887
4	Latin America and the Caribbean	NaN	904	1990	49.830217
5	Northern America	NaN	905	1990	51.115342

Table 26: Tidy4_Maj_area_Female migrants as a percentage of the international migrant stock

	Region	Notes	Country Code	Year	Female migrants as a percentage of the international migrant stock
1	Eastern Africa	NaN	910	1990	48.504812
2	Middle Africa	NaN	911	1990	49.025765
3	Northern Africa	NaN	912	1990	48.791486
4	Southern Africa	NaN	913	1990	39.606165
5	Western Africa	NaN	914	1990	46.486134

Table 27: Tidy4_Region_Female migrants as a percentage of the international migrant stock

Table 5

By observing table 5, we noticed that table 5 documented the annual rate of change of the migrant stock from 1990-2015. Because it captured the rate of change, the year column for table 5 is a range instead of a specific year since it captures the percentage during the five-year time. However, the nature of the datasets is still the same as in previous tables, so we will execute the same tidy data cleaning steps for table 5 and below are the 6 tables that follow the tidy data principles after cleaning.

The annual rate of change of the migrant stock for female and male in Country, Major Area, and Regions

	Country	Notes	Country Code	Type of data(a)	Year	Sex	Annual rate of change of the migrant stock
1	Burundi	NaN	108	B R	1990-1995	Male	-5.475511
2	Comoros	NaN	174	B	1990-1995	Male	-0.30906
3	Djibouti	NaN	262	B R	1990-1995	Male	-4.046026
4	Eritrea	NaN	232	I	1990-1995	Male	0.983754
5	Ethiopia	NaN	231	B R	1990-1995	Male	-7.179744

Table 28: Tidy5_Country_Annual rate of change of the migrant stock(sex)

	Major Area	Notes	Country Code	Year	Sex	Annual rate of change of the migrant stock
1	Africa	NaN	903	1990-1995	Male	0.798774
2	Asia	NaN	935	1990-1995	Male	-0.63615
3	Europe	NaN	908	1990-1995	Male	1.363213
4	Latin America and the Caribbean	NaN	904	1990-1995	Male	-1.424381
5	Northern America	NaN	905	1990-1995	Male	3.898245

Table 29: Tidy5_Maj_area_Annual rate of change of the migrant stock(sex)

	Region	Notes	Country Code	Year	Sex	Annual rate of change of the migrant stock
1	Eastern Africa	NaN	910	1990-1995	Male	-3.446375
2	Middle Africa	NaN	911	1990-1995	Male	11.845106
3	Northern Africa	NaN	912	1990-1995	Male	-2.13921
4	Southern Africa	NaN	913	1990-1995	Male	-3.266338
5	Western Africa	NaN	914	1990-1995	Male	3.611415

Table 30: Tidy5_Region_Annual rate of change of the migrant stock(sex)

Annual rate of change of the migrant stock for the sum of both sexes in Country, Major Area, and Regions

	Country	Notes	Country Code	Type of data(a)	Year	Annual rate of change of the migrant stock (both sexes)
1	Burundi	NaN	108	B R	1990-1995	-5.355717
2	Comoros	NaN	174	B	1990-1995	-0.199873
3	Djibouti	NaN	262	B R	1990-1995	-4.058465
4	Eritrea	NaN	232	I	1990-1995	0.910748
5	Ethiopia	NaN	231	B R	1990-1995	-7.179771

Table 31: Tidy5_Country_Annual rate of change of the migrant stock (sum_sex)

	Major Area	Notes	Country Code	Year	Annual rate of change of the migrant stock (both sexes)
1	Africa	NaN	903	1990-1995	0.826734
2	Asia	NaN	935	1990-1995	-0.673431
3	Europe	NaN	908	1990-1995	1.420702
4	Latin America and the Caribbean	NaN	904	1990-1995	-1.37121
5	Northern America	NaN	905	1990-1995	3.771892

Table 32: Tidy5_Maj_area_Annual rate of change of the migrant stock (sum_sex)

	Region	Notes	Country Code	Year	Annual rate of change of the migrant stock (both sexes)
1	Eastern Africa	NaN	910	1990-1995	-3.435412
2	Middle Africa	NaN	911	1990-1995	11.88581
3	Northern Africa	NaN	912	1990-1995	-2.872903
4	Southern Africa	NaN	913	1990-1995	-3.114352
5	Western Africa	NaN	914	1990-1995	3.817706

Table 33: Tidy5_Region_Annual rate of change of the migrant stock (sum_sex)

Table 6

By observing table 6, I found that table 6 has three types of data, which are 1. Estimated refugee stock at mid-year (both sexes), 2. Refugees as a percentage of the international migrant stock and 3. The annual rate of change of the refugee stock from 1990-2015. Because they are different data types, we need to separate them into three parts, then separate each part by country, major are,a and region. Therefore, after executing the same tidy data cleaning logic, there will be 9 tidy datasets as follows.

Keep "0" value: I also noticed that different from previous tables, table 6 has many cells that has the value of "0", I choose not to drop these "0" values because 0 does have meaning. For example, in table 34, the estimated refugee stock for the country Comoros is 0 in the year of 1990, the value 0 means that there might be no refugees in Comoros in 1990, but if this cell is "..", that means the number may not be available, in that way, I will choose to drop the value.

Estimated refugee stock at mid-year (both sexes)

	Country	Notes	Country Code	Type of data(a)	Year	Estimated refugee stock at mid-year (both sexes)
1	Burundi	NaN	108	B R	1990	267929
2	Comoros	NaN	174	B	1990	0
3	Djibouti	NaN	262	B R	1990	54508
4	Eritrea	NaN	232	I	1990	0
5	Ethiopia	NaN	231	B R	1990	741965

Table 34: Tidy6_Coutry_Estimated refugee stock at mid-year (both sexes) (sum_sex)

	Major Area	Notes	Country Code	Year	Estimated refugee stock at mid-year (both sexes)
1	Africa	NaN	903	1990	5687352
2	Asia	NaN	935	1990	9937007
3	Europe	NaN	908	1990	1321884
4	Latin America and the Caribbean	NaN	904	1990	1197198
5	Northern America	NaN	905	1990	583450

Table 35: Tidy6_Maj_area_Estimated refugee stock at mid-year (both sexes) (sum_sex)

	Region	Notes	Country Code	Year	Estimated refugee stock at mid-year (both sexes)
1	Eastern Africa	NaN	910	1990	3168001
2	Middle Africa	NaN	911	1990	446609
3	Northern Africa	NaN	912	1990	1202360
4	Southern Africa	NaN	913	1990	135525
5	Western Africa	NaN	914	1990	734857

Table 36: Tidy6_Region_Estimated refugee stock at mid-year (both sexes) (sum_sex)

Refugees as a percentage of the international migrant stock

	Country	Notes	Country Code	Type of data(a)	Year	Refugees as a percentage of the international migrant stock
1	Burundi	NaN	108	B R	1990	80.43259
2	Comoros	NaN	174	B	1990	0
3	Djibouti	NaN	262	B R	1990	44.597901
4	Eritrea	NaN	232	I	1990	0
5	Ethiopia	NaN	231	B R	1990	64.21771

Table 37: Tidy6_Country_Refugees as a percentage of the international migrant stock

	Major Area	Notes	Country Code	Year	Refugees as a percentage of the international migrant stock
1	Africa	NaN	903	1990	36.246821
2	Asia	NaN	935	1990	20.640923
3	Europe	NaN	908	1990	2.685708
4	Latin America and the Caribbean	NaN	904	1990	16.697956
5	Northern America	NaN	905	1990	2.113142

Table 38: Tidy6_Maj_area_Refugees as a percentage of the international migrant stock

	Region	Notes	Country Code	Year	Refugees as a percentage of the international migrant stock
1	Eastern Africa	NaN	910	1990	53.118453
2	Middle Africa	NaN	911	1990	30.578557
3	Northern Africa	NaN	912	1990	50.031625
4	Southern Africa	NaN	913	1990	9.733481
5	Western Africa	NaN	914	1990	16.437904

Table 39: Tidy6_Region_Refugees as a percentage of the international migrant stock

Annual rate of change of the refugee stock from 1990-2015

	Country	Notes	Country Code	Type of data(a)	Year	Annual rate of change of the refugee stock
1	Burundi	NaN	108	B R	1990-1995	-3.390926
3	Djibouti	NaN	262	B R	1990-1995	-9.763426
5	Ethiopia	NaN	231	B R	1990-1995	-5.505717
6	Kenya	NaN	404	B R	1990-1995	42.521055
8	Malawi	NaN	454	B R	1990-1995	-104.307376

Table 40: Tidy6_Country_Annual rate of change of the refugee stock from 1990-2015

	Major Area	Notes	Country Code	Year	Annual rate of change of the refugee stock
1	Africa	NaN	903	1990-1995	0.076037
2	Asia	NaN	935	1990-1995	-3.819461
3	Europe	NaN	908	1990-1995	13.2017
4	Latin America and the Caribbean	NaN	904	1990-1995	-23.096408
5	Northern America	NaN	905	1990-1995	1.917003

Table 41: Tidy6_Maj_area_Annual rate of change of the refugee stock from 1990-2015

	Region	Notes	Country Code	Year	Annual rate of change of the refugee stock
1	Eastern Africa	NaN	910	1990-1995	-5.30801
2	Middle Africa	NaN	911	1990-1995	12.964162
3	Northern Africa	NaN	912	1990-1995	-3.456178
4	Southern Africa	NaN	913	1990-1995	-1.954547
5	Western Africa	NaN	914	1990-1995	8.717581

Table 42: Tidy6_Region_Annual rate of change of the refugee stock from 1990-2015

Discussion and Reflection

From this project, by applying the five tidy data principles into a real messy dataset, I began to understand the nature of tidy data. To sum up, I believe it has three core concepts: 1. Each unique variable is a column, 2. each observation is a row, and by combing them together, 3. each type of observational unit forms a table. By following these key frameworks, it becomes much easier to observe, clean, and analyze a dataset in a more structural way.

In detail, I learned how to link the structure of a dataset with its semantics and incorporate it into a standardized way. For example, there are multiple variables(major area, region, country) stored in 1 column and I need to divide them into three tables, I first referred to the “Annex” table and organized all country names, region names, and major area names, then I applied the “select row” function, using the country names, region names, major area names as the unique column indicators that linked two datasets together. In this way, if the name in the original table matched the name in the Annex table, then the row in the original table will be selected and together, form the new tables.

Meanwhile, I also utilized different ways to deal with missing values. For example, I decided to drop the row if the value is missing in the last column. Because the entire row will have no meaning if there are no values in the last column, therefore they can be safely removed. However, I decided to keep all the value = ‘0’ in the last column because sometimes 0 does have meaning, it might mean that the data is actually equal to 0, instead of not available or missing, so it is important to keep them. Moreover, there are also lots of missing values in the “Notes” column and the “Type of data(a)” column, I decided to delete all the “Type of data(a)” columns for regions and major areas because these they don’t have any values for the type of data(a), only countries have the values of the type of data(a). Also, I choose to keep all the "Notes" columns for all six tables because notes occurred for all three categories, so it may cause potential issues or confusion if we delete the entire column or row with the missing "Notes".

Limitations and potential issues

There are certain limitations and potential issues for the current tidy datasets. Firstly, I ended up with 36 tidy datasets, even though I followed all the tidy data principles, I still think there are too many small sub-tables that I cleaned from the 6 main tables, and I believe there might be potential ways to combine some of the tables together in a more logic way while not violating any of the tidy data principles.

The other potential issue that comes to my attention is that I used the melt function to unpivot some tables from wide to long format. Taking *Table 8: Tidy1_Country_International migrant stock at mid-year (sex)* as an example, each country in each year for each sex would need its own row, and metadata like 'country name', 'note', and 'country code', and 'type of data would need to be repeated many times, this may not be the most concise and efficient way to sort this dataset, but if we choose to delete any of the repeated columns(like country code, or notes), then the entire row may not be complete because some information is missing. Therefore, there might be some potential ways to deal with this repeated, duplicated situation and sort them in a more logical, manageable way.

Conclusions

Coming from a marketing background, most of my previous experience with data cleaning and data analysis is either using Excel or SPSS, and this was my first time using Python to conduct data cleaning. By using the logic of tidy data, I was able to develop a consistent tidy data structure for all 6 tables of the UN datasets and apply tidy data tools in a more efficient manner. Even though there might be many flaws in the tidy datasets that I conducted, I am eager to learn more knowledge in this realm and rebuild the datasets for further analysis.