INF1340
Xinxuan Lin

# Midterm Assignment Write-up

## 1. Introduction

In this assignment, I clean the UN Migrant Stock Total file that contains six tables based on the Tidy data principles. Table 1, 2, 3, and 5 have similar violations of Tidy data principles, therefore their cleaning processes are similar. The details of the cleaning steps for Table 1 would be discussed in section *Table 1*, and the process for Table 2, 3, 5 are discussed in section *Table 2, 3, 5*. The screenshots of the dataset are included after the explanation of the cleaning steps to provide a glance at the data structure. The procedure of cleaning Table 4 and 6 are discussed in sections *Table 4* and *Table 6*, respectively. Based on the Tidy data principle, I merge the cleaned tables together, and the reason and process of doing so would be discussed in the section *Table manipulation in merging (section 2.5)*. The challenges in cleaning process are discussed in section *Discussion (section 3)*.

## 2. Method/Result

All tables have multi index headers, therefore, when import these datasets, I combine the first two rows as a new header, then rename each column based on the associated characteristics including the sex categories and topic of interest. For example, table 1 contains the international migrant stock number for three sex categories: both sexes, male and female between 1990 and 2015. Each category has six columns which are the stock data collected every five-year from 1990 to 2015. The header names are changed by adding the associated sex category to each column of years. Therefore, the columns of 1990 to 2015 under the international migrant stock for both sexes became “IMS Both 1990”, “IMS Both 1995, and etc.

### 2.1 Table 1

After renaming the headers of the international migrant stock columns (1990 to 2015 for each category of sex), the first problem is that the column headers are values instead of variable names for these 18 columns (Figure 1). This violates the Tidy data principle 1 that every variable must have its own column. So, I pivot the table into a longer format by melting the headers as one column named *IMS* and saving all the associated numeric values in another variable called *International migrant stock at mid-year*. Thus, eighteen-column headers and their values are melted into two columns (Figure 2).

| | Sort order | Major area, region, country or area of destination | Notes | Country code | Type of data (a) | IMS Both 1990 | IMS Both 1995 | IMS Both 2000 | IMS Both 2005 | IMS Both 2010 | ... |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | WORLD | NaN | 900 | NaN | 152563212 | 160801752 | 172703309 | 191269100 | 221714243 | ... |
| 1 | 2 | Developed regions | (b) | 901 | NaN | 82378628 | 92306854 | 103375363 | 117181109 | 132560325 | ... |
| 2 | 3 | Developing regions | (c) | 902 | NaN | 70184584 | 68494898 | 69327946 | 74087991 | 89153918 | ... |
| 3 | 4 | Least developed countries | (d) | 941 | NaN | 11075966 | 11711703 | 10077824 | 9809634 | 10018128 | ... |
| 4 | 5 | Less developed regions excluding least develop... | NaN | 934 | NaN | 59105261 | 56778501 | 59244124 | 64272611 | 79130668 | ... |

Figure 1: Data frame structure after merging the multi-index headers.

| | Sort order | Major area, region, country or area of destination | Notes | Country code | Type of data (a) | IMS | International migrant stock at mid-year |
|---|---|---|---|---|---|---|---|
| **0** | 1 | WORLD | NaN | 900 | NaN | IMS Both 1990 | 152563212 |
| **1** | 2 | Developed regions | (b) | 901 | NaN | IMS Both 1990 | 82378628 |
| **2** | 3 | Developing regions | (c) | 902 | NaN | IMS Both 1990 | 70184584 |
| **3** | 4 | Least developed countries | (d) | 941 | NaN | IMS Both 1990 | 11075966 |
| **4** | 5 | Less developed regions excluding least develop... | NaN | 934 | NaN | IMS Both 1990 | 59105261 |
| **5** | 6 | Sub-Saharan Africa | (e) | 947 | NaN | IMS Both 1990 | 14690319 |

Figure 2: Data frame structure after melting the column headers and associated values.

However, the second problem is that the *IMS* column contain both the sex and year categories, which again violates principle 1 as each column needs to contain a unique variable. Therefore, the sex and year values in the *IMS* column are separated into two new columns called *Sex* and *Year.* For convivence in reading, I move *Year* and *Sex* columns to the front of *International migrant stock at mid-year* and save it as a new data frame called *un1_order* (Figure 3).

| | Sort order | Major area, region, country or area of destination | Notes | Country code | Type of data (a) | Year | Sex | International migrant stock at mid-year |
|---|---|---|---|---|---|---|---|---|
| **0** | 1 | WORLD | NaN | 900 | NaN | 1990 | Both | 152563212 |
| **1** | 2 | Developed regions | (b) | 901 | NaN | 1990 | Both | 82378628 |
| **2** | 3 | Developing regions | (c) | 902 | NaN | 1990 | Both | 70184584 |
| **3** | 4 | Least developed countries | (d) | 941 | NaN | 1990 | Both | 11075966 |
| **4** | 5 | Less developed regions excluding least develop... | NaN | 934 | NaN | 1990 | Both | 59105261 |

Figure 3: Data frame structure after separating the year and sex into two columns

Thirdly, the international stock column contains missing value in ".." form. To keep the data type of variable consistent, I change the missing values into 0.

Fourthly, in Figure 3, we can see that all the countries, regions, continents, and other larger ranges of destinations are collected as one variable named *Major area, region, country or area of destination*. Despite the fact that they all indicate the destination, they are different types of location, which also violates principle 1 as there are more than one variable in one column. Therefore, I separate the destination column into 7 different location columns: *Total, Nation, Separation in developing regions, Separation in Africa, Major area, Region, Country or area.* In order to do so, I filter the destination rows based on the characteristics of the locations, and then save them as a new data frame. For example, I filter out all the continent rows and save them as a dataset called *Main area* (Figure 4).

| | Sort order | Major area | Notes | Country code | Type of data (a) | Year | Sex | International migrant stock at mid-year |
|---|---|---|---|---|---|---|---|---|
| **6** | 7 | Africa | NaN | 903 | NaN | 1990 | Both | 15690623 |
| **70** | 71 | Asia | NaN | 935 | NaN | 1990 | Both | 48142261 |
| **126** | 127 | Europe | NaN | 908 | NaN | 1990 | Both | 49219200 |
| **179** | 180 | Latin America and the Caribbean | NaN | 904 | NaN | 1990 | Both | 7169728 |
| **231** | 232 | Northern America | NaN | 905 | NaN | 1990 | Both | 27610542 |

Figure 4: Data frame that only contains main area locations

After the filtering, I concatenate all seven datasets together to have all information in one table. Since the variables *Notes, Type of data (a)* do not provide important information about the observational unit, they are removed from the dataset. *Country code* is dropped from the data table because both this and *Sort order* provide the unique ID for each destination (Figure 5).

| | Sort order | Total | Nation | Separation in developing regions | Separation in Africa | Major area | Region | Country or area | Year | Sex | International migrant stock at mid-year |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | WORLD | NaN | NaN | NaN | NaN | NaN | NaN | 1990 | Both | 152563212 |
| 1 | 1 | WORLD | NaN | NaN | NaN | NaN | NaN | NaN | 1995 | Both | 160801752 |
| 2 | 1 | WORLD | NaN | NaN | NaN | NaN | NaN | NaN | 2000 | Both | 172703309 |
| 3 | 1 | WORLD | NaN | NaN | NaN | NaN | NaN | NaN | 2005 | Both | 191269100 |
| 4 | 1 | WORLD | NaN | NaN | NaN | NaN | NaN | NaN | 2010 | Both | 221714243 |
| 5 | 1 | WORLD | NaN | NaN | NaN | NaN | NaN | NaN | 2015 | Both | 243700236 |
| 6 | 1 | WORLD | NaN | NaN | NaN | NaN | NaN | NaN | 1990 | Male | 77747510 |
| 7 | 1 | WORLD | NaN | NaN | NaN | NaN | NaN | NaN | 1995 | Male | 81737477 |
| 8 | 1 | WORLD | NaN | NaN | NaN | NaN | NaN | NaN | 2000 | Male | 87884839 |
| 9 | 1 | WORLD | NaN | NaN | NaN | NaN | NaN | NaN | 2005 | Male | 97866674 |
| 10 | 1 | WORLD | NaN | NaN | NaN | NaN | NaN | NaN | 2010 | Male | 114613714 |
| 11 | 1 | WORLD | NaN | NaN | NaN | NaN | NaN | NaN | 2015 | Male | 126115435 |
| 12 | 1 | WORLD | NaN | NaN | NaN | NaN | NaN | NaN | 1990 | Female | 74815702 |
| 13 | 1 | WORLD | NaN | NaN | NaN | NaN | NaN | NaN | 1995 | Female | 79064275 |
| 14 | 1 | WORLD | NaN | NaN | NaN | NaN | NaN | NaN | 2000 | Female | 84818470 |
| 15 | 1 | WORLD | NaN | NaN | NaN | NaN | NaN | NaN | 2005 | Female | 93402426 |
| 16 | 1 | WORLD | NaN | NaN | NaN | NaN | NaN | NaN | 2010 | Female | 107100529 |
| 17 | 1 | WORLD | NaN | NaN | NaN | NaN | NaN | NaN | 2015 | Female | 117584801 |
| 18 | 2 | NaN | Developed regions | NaN | NaN | NaN | NaN | NaN | 1990 | Both | 82378628 |
| 19 | 3 | NaN | Developing regions | NaN | NaN | NaN | NaN | NaN | 1990 | Both | 70184584 |

Figure 5: Dataset that contains all 7 different location types.

Fifthly, Figure 5 contains the international migrant stock of each destination in each sex category for every 5-year. Thus, the dataset has three types of observational units which are the destination, year and sex. This violates principle 3 that a table should have only one type of observational unit. Therefore, the table should be separated into three based on the types of the observational unit, one contains only the destination and *Sort order* which is the unique identification variable for the destination, one contains *Sort order, Year,* and the associated international migrant stock, and the other one that has *Sort order, Sex*, and the associated international migrant stock. But due to the difficulty in separating the year and sex into two tables, so I only create two new tables, one for the destination observational unit (Figure 6), and one that contains *Sort order*, *Year*, *Sex* and *International migrant stock* (Figure 7).

| | Sort order | Total | Nation | Separation in developing regions | Separation in Africa | Major area | Region | Country or area |
|---|---|---|---|---|---|---|---|---|
| 0 | 1 | WORLD | NaN | NaN | NaN | NaN | NaN | NaN |
| 1 | 2 | NaN | Developed regions | NaN | NaN | NaN | NaN | NaN |
| 2 | 3 | NaN | Developing regions | NaN | NaN | NaN | NaN | NaN |
| 3 | 4 | NaN | NaN | Least developed countries | NaN | NaN | NaN | NaN |
| 4 | 5 | NaN | NaN | Less developed regions excluding least develop... | NaN | NaN | NaN | NaN |

Figure 6: Data table for destination.

| | Sort order | Year | Sex | International migrant stock at mid-year |
|---|---|---|---|---|
| 0 | 1 | 1990 | Both | 152563212 |
| 1 | 1 | 1995 | Both | 160801752 |
| 2 | 1 | 2000 | Both | 172703309 |
| 3 | 1 | 2005 | Both | 191269100 |
| 4 | 1 | 2010 | Both | 221714243 |

Figure 7: Data table for the year and sex.

**2.2 Table 2, Table 3 and Table 5**

All Table 2, 3 and 5 of the UN data files have the same data structure as Table 1, therefore, the violations of the Tidy data principles and the cleaning process are similar as Table 1. Instead of *international migrant stock*, Table 2, Table 3 and Table 5 have *Total population at mid-year (thousands)*, *International migrant stock as a percentage of the total population*, and *Annual rate of change of the migrant stock*, respectively. Similar as Table 1, each of Table 2, 3 and 5 are supposed to be separated into 3 tables based on the types of the observational unit, but due to the difficulty in separation, the dataset ends with two separations instead of three, which is the same as Table 1 case.

**2.3 Table 4**

Table 4 also has similar violations of the Tidy data principles and the cleaning process as table 1. The process of cleaning the dataset are similar. The only difference is that instead of three sex categories, this dataset only contains data for female migrants (Figure 8), which means it has one measurement per destination every 5-year. In other words, there are two observational units: the destination and the year of data. Therefore, according to the Tidy data principle 3, the data table is separated into two tables: one contains *Sort order* and the destination (Figure 9), and the other one contains *Sort order*, *Year,* and *percentage of the international migrant stock* for female migrants (Figure 10).

| | Sort order | Major area, region, country or area of destination | Notes | Country code | Type of data (a) | Year | Female migrants as a percentage of the international migrant stock |
|---|---|---|---|---|---|---|---|
| 0 | 1 | WORLD | NaN | 900 | NaN | 1990 | 49.03915 |
| 1 | 2 | Developed regions | (b) | 901 | NaN | 1990 | 51.123977 |
| 2 | 3 | Developing regions | (c) | 902 | NaN | 1990 | 46.592099 |
| 3 | 4 | Least developed countries | (d) | 941 | NaN | 1990 | 47.261155 |
| 4 | 5 | Less developed regions excluding least develop... | NaN | 934 | NaN | 1990 | 46.466684 |

Figure 8: Dataset structure after pivoting the column headers down for Table 4.

| | Sort order | Total | Nation | Separation in developing regions | Separation in Africa | Major area | Region | Country or area |
|---|---|---|---|---|---|---|---|---|
| 0 | 1 | WORLD | NaN | NaN | NaN | NaN | NaN | NaN |
| 1 | 2 | NaN | Developed regions | NaN | NaN | NaN | NaN | NaN |
| 2 | 3 | NaN | Developing regions | NaN | NaN | NaN | NaN | NaN |
| 3 | 4 | NaN | NaN | Least developed countries | NaN | NaN | NaN | NaN |

Figure 9: Dataset for destination

| | Sort order | Year | Female migrants as a percentage of the international migrant stock |
|---|---|---|---|
| 0 | 1 | 1990 | 49.039150 |
| 1 | 1 | 1995 | 49.168790 |
| 2 | 1 | 2000 | 49.112244 |
| 3 | 1 | 2005 | 48.832993 |
| 4 | 1 | 2010 | 48.305660 |

Figure 10: Dataset for the other type of observational unit.

**2.4 Table 6**

Table 6 contains no sex category and three topics of interest: the estimated refugee stock at mid-year for both sexes, the percentage of the international migrant stock for refugees, and the annual rate of change of the refugee stock. Similar to Table 1, I rename the multi-index column headers into a single level based on the corresponding topic of interest. Instead of melting the column headers directly as shown in Figure 2, I separate the table into three based on the topics of interest. Then I melt the column headers into a

column within each table due to its violation of Tidy data principle 1. Each table has the *Sort order,* the destination, *Notes*, *Country code*, *Type of data (a*), *Year*, and the corresponding topic of interest (Figure 11).

| | Sort order | Major area, region, country or area of destination | Notes | Country code | Type of data (a) | Year | Estimated refugee stock at mid-year |
|---|---|---|---|---|---|---|---|
| 0 | 1 | WORLD | NaN | 900 | NaN | 1990 | 18836571 |
| 1 | 2 | Developed regions | (b) | 901 | NaN | 1990 | 2014564 |
| 2 | 3 | Developing regions | (c) | 902 | NaN | 1990 | 16822007 |
| 3 | 4 | Least developed countries | (d) | 941 | NaN | 1990 | 5048391 |
| 4 | 5 | Less developed regions excluding least develop... | NaN | 934 | NaN | 1990 | 11773616 |

| | Sort order | Major area, region, country or area of destination | Year | Refugees as a percentage of the international migrant stock |
|---|---|---|---|---|
| 0 | 1 | WORLD | 1990 | 12.346732 |
| 1 | 2 | Developed regions | 1990 | 2.445494 |
| 2 | 3 | Developing regions | 1990 | 23.968236 |
| 3 | 4 | Least developed countries | 1990 | 45.56588 |
| 4 | 5 | Less developed regions excluding least develop... | 1990 | 19.919743 |

Figure 11: Melting the column headers that contains the values of year in Refugee stock dataset and percentage of the international migrant stock for refugees' dataset, respectively.

The data of the annual rate of change of the refugee stock were collected in 4 5-year intervals between 1990 and 2015. For convenience in formatting, the values of the year interval are renamed into the upper bound of the interval. For example, "1990-1995" is changed into "1995" as shown in Figure 12.

| | Sort order | Major area, region, country or area of destination | Year | Annual rate of change of the refugee stock |
|---|---|---|---|---|
| 0 | 1 | WORLD | ARCRS 1990–1995 | –2.123497 |
| 1 | 2 | Developed regions | ARCRS 1990–1995 | 9.388424 |
| 2 | 3 | Developing regions | ARCRS 1990–1995 | –2.839417 |
| 3 | 4 | Least developed countries | ARCRS 1990–1995 | –0.680327 |
| 4 | 5 | Less developed regions excluding least develop... | ARCRS 1990–1995 | –4.3836 |

| | Sort order | Major area, region, country or area of destination | Year | Annual rate of change of the refugee stock |
|---|---|---|---|---|
| 0 | 1 | WORLD | 1995 | –2.123497 |
| 1 | 2 | Developed regions | 1995 | 9.388424 |
| 2 | 3 | Developing regions | 1995 | –2.839417 |
| 3 | 4 | Least developed countries | 1995 | –0.680327 |
| 4 | 5 | Less developed regions excluding least develop... | 1995 | –4.3836 |

Figure 12: The interval for year are changed into the upper bound of the range.

After that, I combine all three tables based on the destination to have a complete table that includes all the topics (Figure 13). Due to the same reason as table 1 case, the variables *Notes*, *Type of data(a)*, and *Country code* are removed from the dataset. The missing data in the dataset were recorded in different forms: "NaN", "..", and 0. To keep the format consistent, I replace the missing values with 0. In this case, the value for the annual rate of change of the refugee stock in 1990 would be 0, because the value of the year for 1990-1995 was changed into 1995.

| | Sort order | Major area, region, country or area of destination | Year | Estimated refugee stock at mid-year | Refugees as a percentage of the international migrant stock | Annual rate of change of the refugee stock |
|---|---|---|---|---|---|---|
| 0 | 1 | WORLD | 1990 | 18836571 | 12.346732 | 0.0 |
| 1 | 2 | Developed regions | 1990 | 2014564 | 2.445494 | 0.0 |
| 2 | 3 | Developing regions | 1990 | 16822007 | 23.968236 | 0.0 |
| 3 | 4 | Least developed countries | 1990 | 5048391 | 45.565880 | 0.0 |
| 4 | 5 | Less developed regions excluding least develop... | 1990 | 11773616 | 19.919743 | 0.0 |

Figure 13: Dataset that contains three topics of interest with the melted column headers.

The destination column in this dataset also violates the Tidy data principle 1, which is the same as the Table 1 case. Thus, this column is also separated into different ones based on the geographic characteristics. The procedure of doing this is the same as the Table 1 case.

Since for each destination, we measure the three topics of interest every five years, the observational units are the destination and the year. Therefore, due to the violation of principle 3 that each table should have only one type of observational unit, the dataset is separated into two. One is the location table that contains the sorting order and the destination (same as Figure 9); the other contains *Sort order*, *Year*, and the three topics of interest (Figure 14).

| | Sort order | Year | Estimated refugee stock at mid-year | Refugees as a percentage of the international migrant stock | Annual rate of change of the refugee stock |
|---|---|---|---|---|---|
| 0 | 1 | 1990 | 18836571 | 12.346732 | 0.000000 |
| 1 | 1 | 1995 | 17853840 | 11.103013 | -2.123497 |
| 2 | 1 | 2000 | 15827803 | 9.164736 | -3.837069 |
| 3 | 1 | 2005 | 13276733 | 6.941389 | -5.557223 |
| 4 | 1 | 2010 | 15370755 | 6.932687 | -0.025089 |

Figure 14: Dataset that contains only one type of the observational unit (year) and three topics of interest.

**2.5 Tables manipulation in merging**

For Table 1 to 6, one of the observational units is the destination. As Tidy data principle 3 states that one type of observational unit should be in one table, the destination table for each sheet should be combined into one. I concatenate all six destination tables and remove the duplicate rows. The final destination table for all the UN data sheets contains 265 locations, the structure and observations are the same as Figure 6.

As discussed in previous subsections, Table 1, 2, 3 and 5 have the same types of observational units which are the destination, the sex and the year. Other than combining the destination table, I also merge the tables for the other type of observational unit (Figure 15). Therefore, instead of having one destination table and 4 separated sex and year tables for each sheet, Table 1, 2, 3 and 5 are combined into two tables, one for the destination and the other for the sex and year.

| | Sort order | Year | Sex | International migrant stock at mid-year | Total population at mid-year (thousands) | International migrant stock as a percentage of the total population | Annual rate of change of the migrant stock |
|---|---|---|---|---|---|---|---|
| 0 | 1 | 1990 | Both | 152563212 | 5309667.699 | 2.873310 | 0.000000 |
| 1 | 1 | 1995 | Both | 160801752 | 5735123.084 | 2.803806 | 1.051865 |
| 2 | 1 | 2000 | Both | 172703309 | 6126622.121 | 2.818899 | 1.428058 |
| 3 | 1 | 2005 | Both | 191269100 | 6519635.850 | 2.933739 | 2.042124 |
| 4 | 1 | 2010 | Both | 221714243 | 6929725.043 | 3.199467 | 2.954160 |

Figure 15: Table contains the sex, year and the associated topics of interest in Table 1, 2, 3, and 5.

Similarly, except for the destination, both Table 4 and 6 also have the same type of observational unit – year. Therefore, the year tables for both cases are merged together.

| | Sort order | Year | Female migrants as a percentage of the international migrant stock | Estimated refugee stock at mid-year | Refugees as a percentage of the international migrant stock | Annual rate of change of the refugee stock |
|---|---|---|---|---|---|---|
| 0 | 1 | 1990 | 49.039150 | 18836571 | 12.346732 | 0.000000 |
| 1 | 1 | 1995 | 49.168790 | 17853840 | 11.103013 | −2.123497 |
| 2 | 1 | 2000 | 49.112244 | 15827803 | 9.164736 | −3.837069 |
| 3 | 1 | 2005 | 48.832993 | 13276733 | 6.941389 | −5.557223 |
| 4 | 1 | 2010 | 48.305660 | 15370755 | 6.932687 | −0.025089 |

Figure 16: Table contains the year and the topics of interest in Table 4 and 6

## 3. Discussion

According to three Tidy data principles, the six UN data tables are cleaned and merged into 3 datasets. Each table was originally recorded in a format that is easier for the human to read, but not trivial for computers. Therefore, the datasets are transformed into more efficient forms to do the analysis computationally based on the Tidy data principles. I realize that principle 3 –the same type of observational unit should be in the same table –is challenged to be fulfilled in this data file.

Instead of having all variables in one dataset by directly merging all six tables together, I separate the table based on the type of observational unit, and then merge the same type of observational unit into one table. As discussed in section 2.5, Table 1, 2, 3, and 5 have identical types of observational unit –destination, year and sex, thus, each of these tables should be separated into three datasets; Table 4 and 6 have two types of the observational unit –destination and year, so each table should be split into two datasets. One of the types of observational unit is the destination and the elements in the destination are the same, therefore, I decided to merge the destination tables together. Hence, there is one destination table for Table 1, 2, 3, 4 5, and 6.

However, due to the difficulties in separating the year and sex into two tables, these two variables are kept in the same dataset. Therefore, Table 1, 2, 3, and 5 are separated into two tables –one for the destination, one for the year, sex and the corresponding variables. Since the tables for year and sex for Table 1, 2, 3, 5 contain the same types of observational unit, these tables are merged into one dataset.

The other challenge in cleaning the datasets is that the destination column in the raw data contains the locations that have an inclusive relationship with other observations. For example, the countries and the

associated continent are stored in the same column. Since we know that a column should record the same feature of the dataset, we should eliminate this inclusive relationship. I separate the column into different columns based on the information provided in the annex file.

## 4. Conclusion

In conclusion, the UN dataset initially contains 6 different tables (Table 1 to 6). By cleaning the data based on the Tidy data principles, the missing values in the dataset are converted to 0, and the datasets end with 3 different tables, where one for the destination list, one for the year, sex and the associated measurements for Table 1,2,3 and 5, and the other one contains the year and associated measurement for Table 4, 6.