Nov 21st 2021 INF1340: Introduction to Programming for Data Science Michael Ou

Midterm 1: Tidy Data write up

1. Introduction:

When approaching this dataset, I noticed that the excel sheet was not properly formatted for the read_excel function. One of the glaring issues is that the first sheet in the file was the contents page. This page isn’t necessary and did not provide any value to data analysis. To address this issue, I focused on each table individually and ignored the contents/annex/notes sheets. Secondly, the each of the tables in the excel sheet contain a large gap for the UN logo. To address the issue, it was important for me to skip this section via header = 15. This allowed me to get the necessary columns to begin my data wrangling process.

For each table, I noticed that there are columns that did not have any significance. These columns note, type of data, country code and sort order. I drop these columns as I felt that they were unnecessary.

2. Methods for each table:

a. Table 1: International migrant stock at mid-year by sex and by major area, region, country, or area

   Based on the principles of tiny data, I noticed two obvious violations. The violations are based on the principle of “Columns names need to be informative variable names, not values “ and “Each column needs to consist of one and only one variable”. The variable International migrant stock at mid-year (both sexes) contains both the variable of sex and year. These values provide insight and should be in the observation rows.

   To bring the year and sex into the rows, I first had to rename each column to the year + the certain gender (example: 1990_both sex). I also removed the “mid-year“ part as it isn’t informative since the years column will already show this information. Then using the melt function, I can bring these columns into a long format with the melt function under a column named variable. Using string manipulation, I separate based on the “_” for each string to get the year and sex. I separate these values and fill them in two new columns called year and sex. Lastly, I rename the first column to region to help simplify the long name that contain multiple variables (country, area, area of destination and region) .

b. Table 2:
   Table 2 has many of the same glaring problems as Table 1. There are violations of “Columns names need to be informative variable names, not values“ and “Each column needs to consist of one and only one variable”.

   To solve these issues, I would be doing the same methods as done for table 1.

c. Table 3:
Table 3 also has many of the same problems as Table 1. There are violations of "Columns names need to be informative variable names, not values" and "Each column needs to consist of one and only one variable".

To solve these issues, I would be doing the same methods as done for table 1.

d. Table 4:
Table 4 also has many of the same problems as Table 1. There are violations of "Columns names need to be informative variable names, not values" and "Each column needs to consist of one and only one variable".

This table only had one gender which was female. I decided to leave it as is but pivoted the years into long format.

e. Table 5:
Table 5 also has many of the same problems as Table 1. There are violations of "Columns names need to be informative variable names, not values" and "Each column needs to consist of one and only one variable".

Unfortunately, I wasn't able to get the range of the dates right during the string manipulation.

f. Table 6:
One issue of table 6 is that the table combines multiple types of data. There are data measuring the refugees as a percentage of international migrant stock and annual rate of change of the refugee stock. It is best to separate these types of data to ensure tidy data.

I separate each different types of data into separate data frames with two different functions. The first data frame used the iloc function combined with indexing to obtain only estimated refugee stock at mid-year (both sexes). The second and third data frame just obtains the necessary columns for Refugees as a percentage of the international migrant stock and Annual rate of change of the refugee stock.

After tidying each data frame, I wasn't sure if I should have concat to combine df1 and df2 since they do provide similar information with the same time ranges. However, even trying to concat df1/df2, I get NAN values.

3. Conclusion/final notes:

   I felt that my dataset is as clean as possible based on my coding abilities. There are areas of improvement which I would have like to have done. One example of that is I would have love to filter out values based on geographical terms within the "Major area, religion, country or area of destination" to solve the multiple variables problem better. To do this, I would have had a separate column for region that would include non-variables like developing region and developed region. I would also filter out continents since there were non-variables like Africa to create a column called continents. Lastly, I would have also created a country column to fill with individual countries. This would have it easier to compare different regions, continents, and countries.