Background

This semester was the first time I have ever used python and pandas. It was also the first time I came across the concept of *tidy data*. As a result, this assignment was quite challenging for me, but I did my best to logically organize the data according to the principles of tidy data outlined in the article *Tidy Data* by Hadley Wickham. I also used an online article titled *Pythonic Data Cleaning with Pandas and NumPy* by Malay Agarwal which I found on realpython.com.

Successful Changes

Initially, I felt it was important to take a glance at dataset in excel to make viewing it easier. There were many issues with this dataset related that went against *tidy data* principles. For me to understand these issues I had to understand that the characteristics are variables, numerical values are observations My goal was to ensure every column is a variable, every row is an observation, and every cell has a single value. To make the process of *tidying* the data simpler, I decided to clean the data by each individual sheet. Therefore, I converted each sheet into its own .csv file and uploaded it to my *jupyter notebook.* I then imported pandas and NumPy as *pd* and *np*. Using df.head() I realized I needed to remove the blank and unnecessary rows. I did this using the function df1 = pd.read_csv('Table1.csv', header=14). I then wanted to remove another row so that the row with the years becomes the "row of columns" and rename them accordingly.

Roadblocks

As I mentioned earlier, I'm a complete beginner when it comes to programming. My next steps would've been to *lengthen* the tables by *melting* the data so that columns are variable names instead of values. For example, I would've made a column named "Year" instead of having columns "1990, … 2015". I attempted using the pd.melt function but was unable to yield my desired results. I also noticed numerous observational units in one table such as country code and population counts or percentage changes. I presume I would have to split these using the pd.iloc function. Overall, there wasn't much I could change to the datasets.