**Country Codes or Names?**

Examine the two csv files and decide what column in each is the best one to use to determine the country for an entry. Remember: one line is one record is one entry, and nothing permanent is being done to the original data. There are both numerical country-codes and textual country-names.

In no more than 250 words, reply to the following: Which is better for us to use in linking entries in one data file with entries in the other, and why? Type your response in the text field and submit it to the dropbox.

In this instance, if we would like to connect entries in one file with the other, we would have to use the “country name” column. The “MortalityFertilityIncome” dataset has very limited unique identifiers associated with any single entry (Country Name and Country Code), with Country Name being the only overlap with the Global Terrorism dataset. We can associate Country Code in the first dataset with the “country\_txt” column in the second dataset.

There would be instances where it wouldn’t be a direct association, such as “West Germany (FRG)” in the Global Terrorism file which is not in the MortalityFertilityIncome dataset. We could pretty easily do a comparison of these columns between the two files, figure out where any discrepancies arise, and either manually fix them or create a little code snippet that would adjust accordingly.

In general, it would likely be easier to associate numeric codes if they exist. This would avoid having to do any sort of preprocessing or search on the initial entries. For instance, correlating “USA” with “U.S.A” with “United States” etc. One thing to always be wary of, though, is numeric entries being interpreted as text format. It is important to ensure all entries are of the same datatype to ensure appropriate matches.