**DATA PREPROCESSING**:

In this section, we describe the procedures we took to process data for later analysis. We used the PANDAS library and two of our own modules to complete this task. The ‘datapreprocessign.py’ module holds some general and dataset-specific functions and the ‘preprocessing\_test.py’ uses the first module and PANDAS to fully process our three data sets.

Our first step was to create various lists and dictionaries to hold both key attributes for later analysis, and attributes that could be removed from the datasets. These largely consisted of dictionaries to map categorical datatype attributes to its more intuitive text representation. Due to the high amount of column attributes, a list was created to store all irrelevant attributes for subsequent removal.

Once the necessary data structures were created, we were able to reduce dimensions for the data sets in a few different ways. First was the straightforward removal of irrelevant, redundant, and unnecessary attributes. For instance, some of the attributes were regarding a unique ID number for each entry. Some were date approximations for incidents when the dataset already provided the actual dates. The second types of attributes that were removed were because we had the dictionaries such that we only needed either the key or the text representation to access the other type. The third types of attributes that were removed were redundant attributes with discrepancies in the value representations. For instance all 3 datasets had a column attribute for country, but some datasets used historical over the modern names to represent specific countries. In this sense we had to manually look over where territories match/overlap to find the best matches.

After removing these column attributes, we were able to manage discrepancies of how null data was represented. For our project purposes, Numpy’s NaN value was most compatible for the other libraries we were using. Other common discrepancies were date-numbers outside the range of a month or day, and similar attributes with conflicting values. For instance, with the terrorism data set, there were 3 attributes that represented media sources reporting the terrorism incident and whether or not it was speculative or dubious. In this case there would be conflicting reports at times.

After cleaning the datasets, the datasets were split into smaller data frames and put into a pipeline to prepare for integration. In this step, key common attributes were identified as the source to merge the different datasets. The datasets were modified in shape and index representations. This was to better align the attributes and data. For instance, the poverty/health datasets had to be transposed in order for the date attribute to face the same direction as the terrorism dataset. Once all data frames were modified, they were re-appended.

In the last stage of preprocessing, all three datasets were merged using country and year as the common keys to map them together.