Capstone #1 Project - Data Wrangling Report

This analysis aims to analyze the cause of war in 19th and 20th century and use the Correlates of War Project datasets to predict future conflicts and to prevent war before it happens. Other sources are used in these project are Peace Research Institute Oslo's armed conflict data, Stockholm International Peace Research Institute's war data and Our World in Data websites.

The data from the Correlates of War Projects website has data from 1816 to 2018 but not all datasets has input for all years. So it is expected that there are null values throughout the datasets. This must be kept in mind throughout the entire process. The datasets in other sources are also missing values.

After identification of data sources below list gives the details for each data source.

1. Correlates of War has the best datasets of religion, interstate war, dyadic interstate wars, territorial data, alliance, military data. The site offers csv files of the each dataset. This can then be read into a pandas data frame very easily.
2. Peace Research Institute Oslo's armed conflict data is in xls and pdf. Conflict data is between 1946 and 2008.
3. Stockholm International Peace Research Institute's war data is in xls format and allow you to download countries separately.
4. Our World in Data provides data in csv format that can be read into pandas data frame very easily.

To concatenate correct tables all datasets are diagnosed for the inconsistent column names, missing data, duplicate rows , untidy and unexpected data values. All datasets were explored with pandas methods such as `.head()`, `.info()`, and `.describe()`, and DataFrame attributes like `.columns` and `.shape` . Needed datasets are concatenated and a single dataframe is created and saved for later analysis.

I spent most of the time cleaning and ensuring data quality. After concatenated the datasets missing data and duplicate rows appeared. I spent considerable time looking at what data was missing and how best to fill it. One way is using additional datasets.
I also use the left join to merge the two dataframes. This is because not all countries involved in a conflict. Any country missing in one dataset mean they might not involved, and thus have null values in new dataset. After combining columns, a more complete dataframe is now forming. The last step is to fill in the null data throughout the dataset. There are various ways to do this. I fill those values are set to -1, though this may change later with using other resources.

Among the many columns in the datasets I have combined, I only added one that describe the reason of the war based on conflicts, involved parts and type of the war. Additional columns may also be added relating to war, state and territory in the future.
Some changes may have to be made later on in the analysis process, but for now this is the first draft of the dataframe. This cleaned dataframe is saved for the further analyze.

Citation

COW: Sarkees, Meredith Reid and Frank Wayman (2010). Resort to War: 1816 - 2007. Washington DC: CQ Press.

Territorial change data : Tir, Jaroslav, Philip Schafer, Paul Diehl, and Gary Goertz. 1998. "Territorial Changes, 1816-1996: Procedures and Data"Conflict Management and Peace Science 16:89-97.

Alliance: Gibler, Douglas M. 2009. International military alliances, 1648-2008. CQ Press.

WRD: Zeev Maoz and Errol A. Henderson. 2013. “The World Religion Dataset, 1945-2010: Logic, Estimates, and Trends.” International Interactions, 39: 265-291.

National material capabilities: Singer, J. David, Stuart Bremer, and John Stuckey. (1972). "Capability Distribution, Uncertainty, and Major Power War, 1820-1965." in Bruce Russett (ed) Peace, War, and Numbers, Beverly Hills: Sage, 19-48.

Military Interstate dispute : Palmer, Glenn, Vito D'Orazio, Michael Kenwick, and Matthew Lane. 2015. "The Mid4 Dataset, 2002–2010: Procedures, Coding Rules and Description." Conflict Management and Peace Science 32: 222-42.