Capstone Project 1 Milestone 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 Conflict Project goal is to create a better model that screen the red conflict zones with daily updates, and reports the weekly connected events analysis. Possible interested parties are international relation based think-tank's, international relation departments in universities, defense industries, human rights organizations, political groups.

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

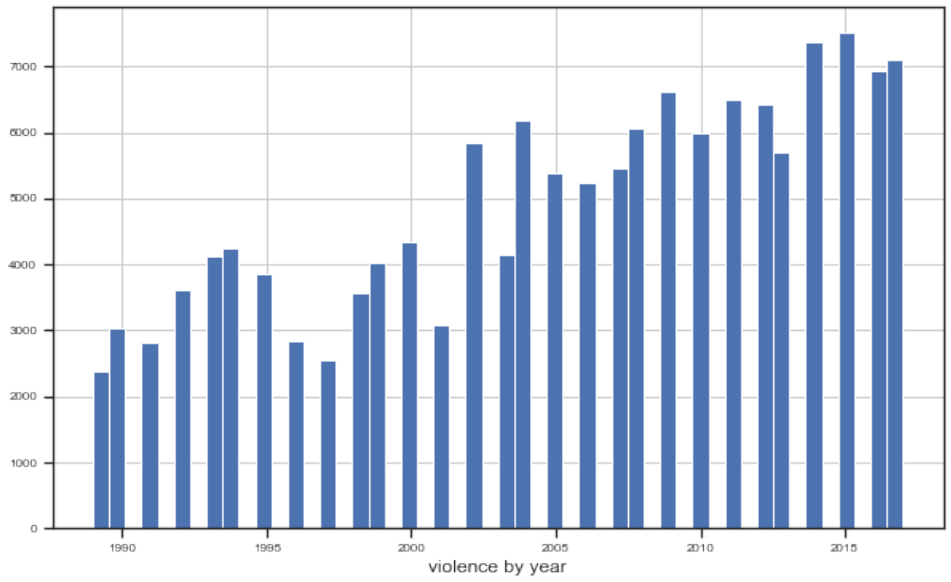
DataFrame attributes like .columns and .shape . Needed datasets are concatenated and a single dataframe is created and saved for later analysis.

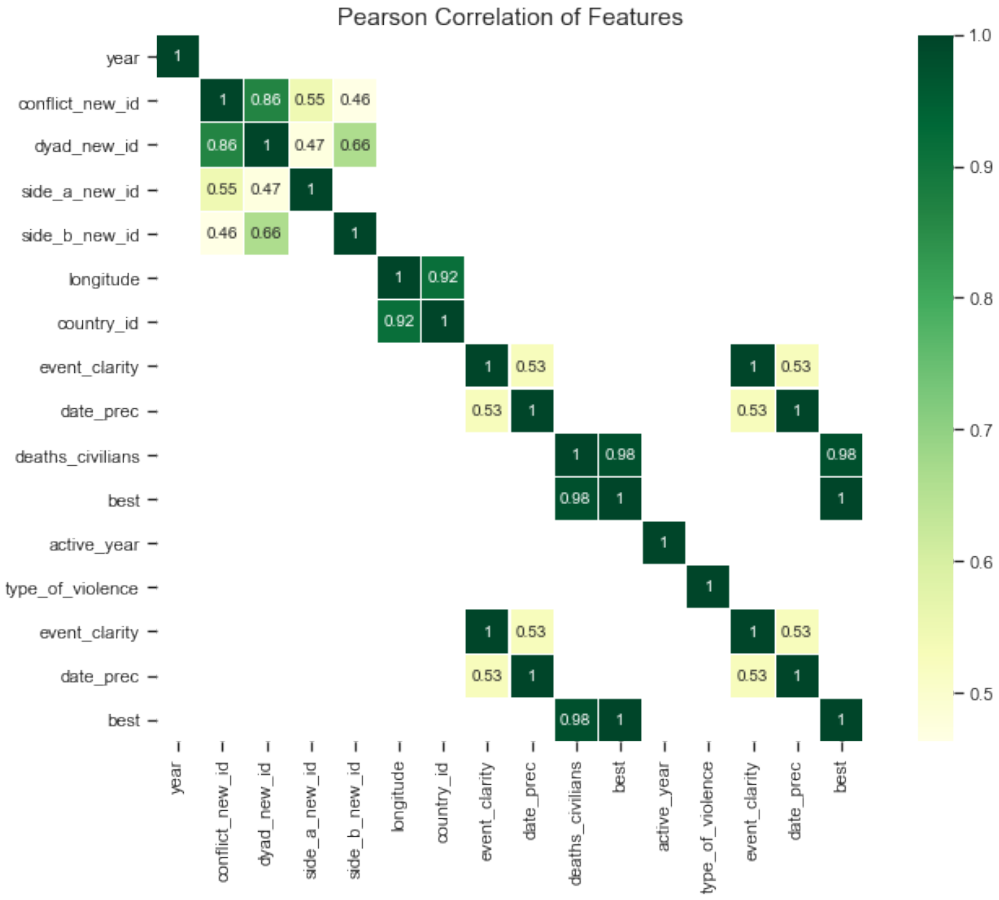
I spent most of the time cleaning and ensuring data quality. After concatenated the datasets missing data and duplicate rows appeared. I spent considerable time looking at what data was missing and how best to fill it. One way is using additional datasets.
I also use the left join to merge the two dataframes. 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.

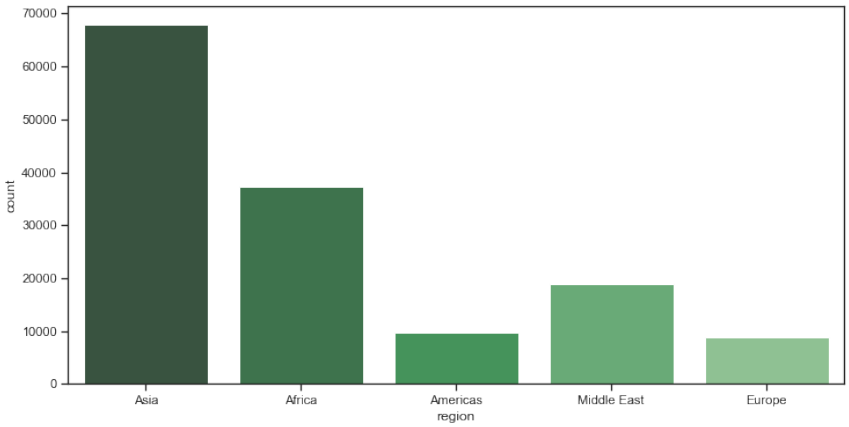
Exploratory Data Analysis

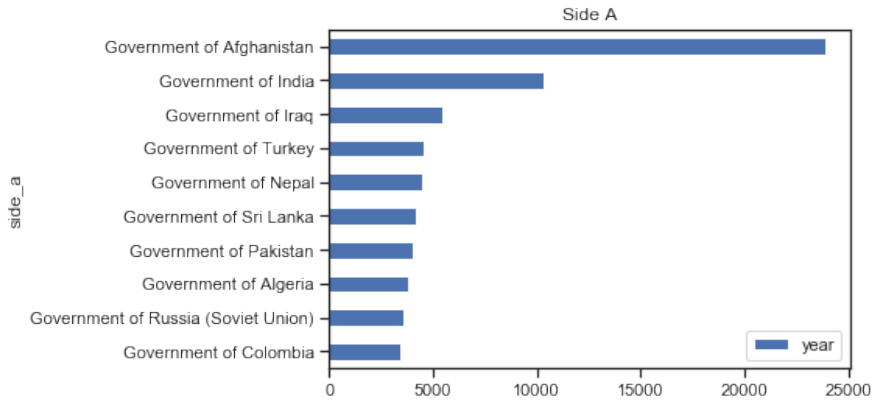
As seen in below graphic over the years there is an increase in violence. There might be many reasons that caused an increase in violence. One and basic reason might be the better reporting of the event and better communication channels now a days. Our Pearson Correlation graph shows that there is a strong relation between the events clarity and reporting time.

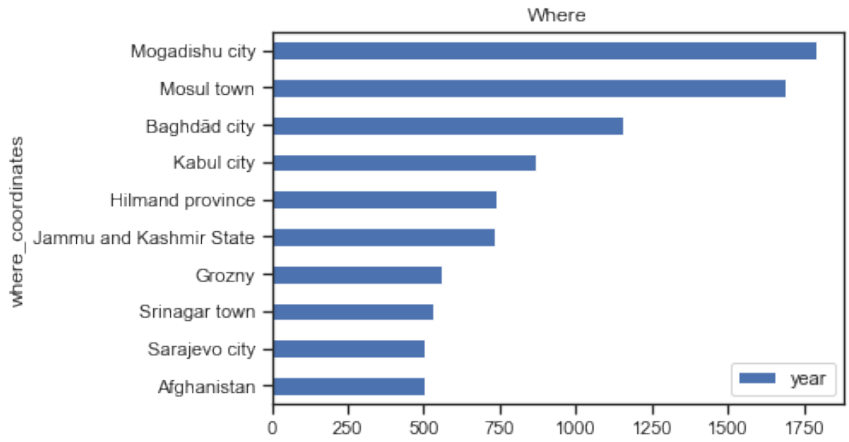

7000
6000
5000
4000
3000
2000
1000
0
1990
1995
2000
2005
2010
2015
violence by year


Pearson Correlation of Features
year
conflict_new_id
dyad_new_id
side_a_new_id
side_b_new_id
longitude
country_id
event_clarity
date_prec
deaths_civilians
best
active_year
type_of_violence
event_clarity
date_prec
best
1.0
0.9
0.8
0.7
0.6
0.5

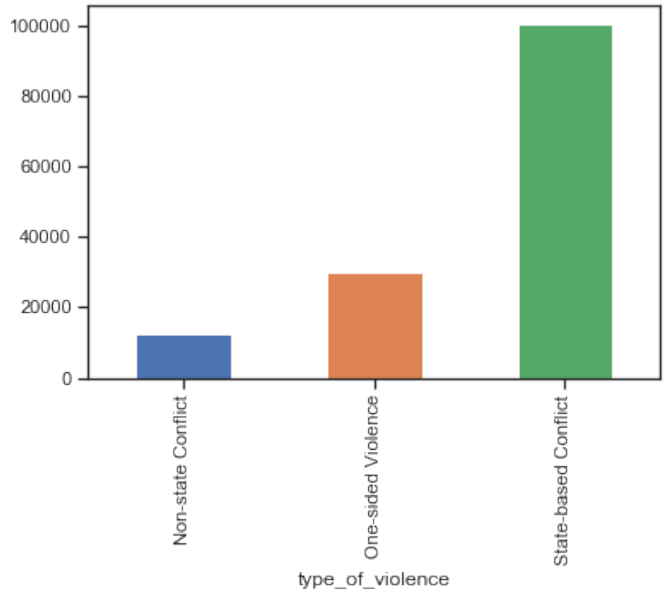
When I did exploratory data analysis with the help of Pearson Correlation I found that there is a strong correlations between country and its longitude. The country and longitude are particularly significant in terms of explaining the answer to my project question. Now I need to add another question to my research. How geographic position of a country effected people's relation and behavior? As seen in below graphic 8 out of first 10 countries that involved in a conflict are in Asia. What are the main reason behind the Asian conflicts? Is it economic reasons, religion or something else?



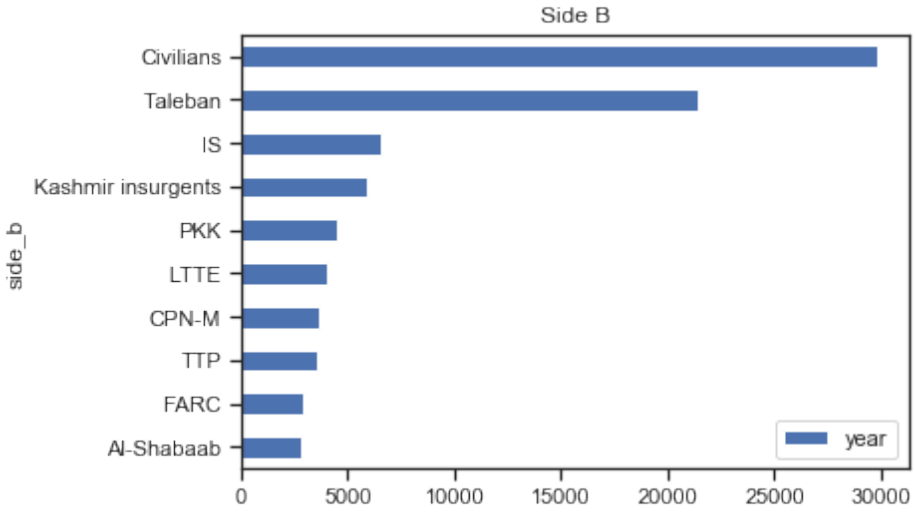




When we look at the type of the violence that occurred we see that state-based conflict (Any non-governmental group of people having announced a name for their group and using armed force against a government) is the one mostly happened.



Civilians are the largest group that used the armed force against government. Since it is the sum of all civilians that involved in a conflict we need to do a separate analyze to examine the reasons, sides and effective years. The following groups are internationally recognized as terrorist organizations.


Side B
side_b
Civilians
Taleban
IS
Kashmir insurgents
PKK
LTTE
CPN-M
TTP
FARC
Al-Shabaab
year
0
5000
10000
15000
20000
25000
30000

The variables included in the UCDP GED version 18.1:

Variable name	Content	Type
id	A unique numeric ID identifying each event.	integer
year	The year of the event	integer
active_year	1: if the event belongs to an active conflict/dyad/actor- Integer year 0: otherwise	integer
type_of_violence	1: state-based conflict 2: non-state conflict 3: one-sided violence	integer
conflict_new_id	A unique conflict identification code for each individual conflict in the dataset.	integer
conflict_name	Name of the UCDP conflict to which the event belongs. For non-state conflicts and one-sided violence this is the same as the dyad name.	string(9999)
dyad_new_id	A unique conflict identification code for each individual dyad in the dataset.	integer
dyad_name	Name of the conflict dyad creating the event. A dyad is the pair of two actors engaged in violence (in the case of one-sided violence, the perpetrator of violence and civilians).	string(9999)
side_a_new_id	A unique ID of side A.	integer
gwnoa	The Gleditsch and Ward number for Side A if the side is a state. Empty if Side A is not a state.	string(9999)
side_a	The name of Side A in the dyad. In state-based conflicts always a government. In one-sided violence always the perpetrating party.	string(9999)
side_b_new_id	A unique ID of side B.	integer
gwnob	The Gleditsch and Ward number for Side B if the side is a state. Empty if Side B is not a state.	string(9999)
side_b	The name of Side B in the dyad. In state-based always the rebel movement or rivalling government. In one- sided violence always “civilians”.	integer
number_of_sources	Number of total sources containing information for an event that were consulted.	
source_article	References to the names, dates and titles of the source material from which information on the event is gathered. This variable is highly streamlined for information collected since 2013, and is less so for older data. For such older data, abbreviations are sometimes used for source agencies. The most frequent are: R: Reuters News, BBC: BBC Monitoring AP: Associated Press Newswires AFP: Agence France Presse, X: Xinhua DOW: Dow Jones Wires	text
source_office	The name of the organizations publishing the source materials.	text
source_date	The dates the source materials were published on.	text
source_headline	The titles of the source materials.	text
source_original	The name or type of person or organization from which the information about the event originates in the original report.	string(9999)

where_prec	The precision with which the coordinates and location assigned to the event reflects the location of the actual event. 1: exact location of the event known and coded. 2: event occurred within at maximum a ca. 25 km radius around a known point. The coded point is the known point. 3: only the second order administrative division where an event happened is known. That administrative division is coded with a point representing it (typically the centroid). 4: only the first order administrative division where an event happened is known. That administrative division is coded with a point representing it (typically the centroid). 5: the only spatial reference for the event is neither a known point nor a known formal administrative division, but rather a linear feature (e.g. a long river, a border, a longer road or the line connecting two locations further afield than 25 km) or a fuzzy polygon without defined borders (informal regions, large radiuses etc.). A representation point is chosen for the feature and employed. 6: only the country where the event took place in is known. 7: event in international waters or airspace.	integer
where_coordinates	Name of the location to which the event is assigned. Fully standardized and normalized.	string(9999)
adm_1	Name of the first order (largest) administrative division string(9999) where the event took place	string(9999)
adm_2	Name of the second order administrative division where the event took place	string(9999)
latitude	Latitude (in decimal degrees)	numeric(9,6)
longitude	Longitude (in decimal degrees)	numeric(9,6)
geom_wkt	An Open Geospatial Consortium textual representation of the location of each individual point. Formatted as OGC WKT (well known text) without SRID.	string(9999)
priogrid_gid	An Open Geospatial Consortium textual representation of the location of each individual point. Formatted as OGC WKT (well known text) without SRID.	integer
country	The PRIO-grid cell id (gid) in which the event took place. Compatibility with PRIO-grid (T ollefsen, 2012) is guaranteed for both PRIO-grid 1 and 2.	string(9999)
country_id	Gleditsch and Ward number of the country in which the event takes place.	integer
region	Region where the event took place. One of following: {Africa, Americas, Asia, Europe, Middle East}	string(9999)
event_clarity	1 (high) for events where the reporting allows the coder to identify the event in full. 2 (lower) for events where an aggregation of information was already made by the source material that is impossible to undo in the coding process.	integer
date_prec	How precise the information is about the date of an event. 1: exact date of event is known; 2: the date of the event is known only within a 2-6 day range. 3: only the week of the event is known 4: the date of the event is known only within an 8-30 day range or only the month when the event has taken place is known 5: the date of the event is known only within a range longer than one month but not more than one calendar year.	integer
date_start	The earliest possible date when the event has taken place.	Date YYYY-MM- DD
date_end	The last possible date when the event has taken place.	Date YYYY-MM- DD
deaths_a	The best estimate of deaths sustained by side a. Always 0 for one-sided violence events.	integer
deaths_b	The best estimate of deaths sustained by side b. integer Always 0 for one-sided violence events.	integer
deaths_civilians	The best estimate of dead civilians in the event. For non-state or state-based events, this is the number of collateral damage resulting in fighting between side a and side b. For one-sided violence, it is the number of civilians killed by side a.	integer
deaths_unknown	The best estimate of deaths of persons of unknown integer status.	integer
best_est	The best (most likely) estimate of total fatalities resulting from an event.	integer

high_est	It is always the sum of deaths_a, deaths_b, deaths_civilians and deaths_unknown.	integer
low_est	The lowest reliable estimate of total fatalities	integer
geom / geometry	An Open Geospatial Consortium / ESRI binary representation of each individual point. Contains the SRID (4326) where supported.	geometry (Point,4326)

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.