Ryerson University
Student: Emil Ibrahim
Student Id: 500862456

Date: 01-Apr-2019
Course: CKM136XJ0 Capstone
Supervisor: Dr. Can Kavaklioglu

Global Terrorism Data Analysis

Introduction:

Terrorism is one of main problems in the 21st century.
Terrorism has continuously been an ongoing danger all over the world.
Analysing global terrorism historical dataset will help to understand the insights of the problem and be prepared in future to prevent or mitigate the number or impact of attacks.

Context:

Information on more than 180,000 Terrorist Attacks.
The Global Terrorism Database (GTD) is an open-source database including information on terrorist attacks around the world from 1970 through 2017. The GTD includes systematic data on domestic as well as international terrorist incidents that have occurred during this time period and now includes more than 180,000 attacks.

Content:

Geography: Worldwide.
Time period: 1970-2017, *except 1993.*
Variables: >135 variables on location, tactics, perpetrators, targets, and outcomes ,….etc

Research question:

Can a success of a terrorism attack be predicted by knowing attack features?

Literature Review:

Suspicious human activity recognition [1]:

From surveillance video this paper used image processing and computer vision to monitor the human activities in real-time and categorize them as usual and unusual activities.
This paper consists of six abnormal activities such as abandoned object detection, theft detection, fall detection, accidents and illegal parking detection on road, violence activity detection, and fire detection.
The classifier used various models such as SVM, Random forest, K-NN.

The Psychology of Suicide Terrorism [2]:

The paper reviewed the understanding of the psychology of suicide terrorism, the study shows that the outstanding common characteristics of a terrorists is their normality.
The paper examines various psychological models that may contribute to the expansion of suicide bombing.

Suicide Terrorists: Are They Suicidal? [3]

The study found many discrepancies uncovered between suicide terrorists and other suicides.

The paper suggest that the terrorists are not truly suicidal and should not be viewed as a subgroup of the general suicide population.

The paper also covered the motivations for suicide vs suicide terrorism.

Terrorism and the internet: a double-edged sword [4]:

The purpose of this paper was to highlight the extent of the use of the internet by terrorist organisations to achieve their strategic and operational objectives.
The study showed that one significant enabler has been the internet, which enables anonymous communication, aids recruitment, encourages the sharing of knowledge, as well as playing a significant role in the spreading of propaganda.
The understanding of the use of internet-based technologies, not only as a potential target for terrorist attack, but as a tool to achieve its ideological and operational goals, remains a neglected area of study requiring further focus.
The paper focused on Al-Qaeda's use of the internet to plan operations for 9/11 was brought to light when computers seized in Afghanistan revealed that al-Qaeda was collecting intelligence on targets and sending encrypted messages via the internet, as well as using internet-based phones. Further investigation revealed the existence of various al-Qaeda web sites.

PATHWAY TO JIHADIST TERRORISM [5]: A CRITICAL LITERATURE REVIEW

The paper addressed a comparison of 10 studies, it identifies commonly seen characteristics of individuals engaged in neo-jihadist terrorism in the United States.
Due to the diverse demographic factors of the individuals that engage in jihad, it is difficult to identify who will become radicalized and engage in acts of terrorism. The number of foreign fighters has grown through recruitment largely from al-Qaeda and ISIS. The third-wave of jihad has also introduced a new threat from homegrown extremists through self-radicalization.

Terrorism Risk Forecasting by Ideology [6]:

The paper presented a study to produce foresights on terrorism.
The paper suggests that target selection and the places where terrorist attacks occur are related to a group's strategies, and violent terrorist acts vary with respect to their surroundings in the same jurisdiction when the ideology is the point of comparison.
The analysis begins with a comparison of targets and risk factors by the ideology of the perpetrators.
After identifying high-risk locations, tested the predictive validity of risk terrain modeling.
The study results showed that context and influence the risky areas of terrorism vary by the nature of the ideology in the jurisdiction.
Terrorists have always added geography into the equation to make a rational choice of their attacks, and the choice-making is ideology-dependent
The predictive model was done using logistic regression and used risk terrain modeling (RTM) to explore the relationship between terrorism, ideology, and geography.

Explaining religious terrorism [7]: A data-mined analysis

The paper examined the relationship between religious liberty and faith-based terrorism by looking specifically at terrorism motivated by a religious imperative and a country's level of religious liberty.
Using classification data mining, paper tested a unique dataset on religious terrorism in order to discover the characteristics that contribute to a country experiencing religiously motivated terrorism. The analysis finds that religious terrorism is indeed a product of a dearth of religious liberty.
The data mining process involves both human and software resources. Data mining is used not only to predict the outcome of a future event but also to provide knowledge about the structure and interrelationships between data. Classification mining algorithms are used to create models that describe existing data and relationships within that dataset. The resulting model is expressed as a classification tree.

This article has made a simple but important claim: the denial of religious freedom increases the likelihood of violent religious forms of political engagement; paradoxically, the best way to combat religious terrorism is not by restricting religious practices but rather by safeguarding their legitimate manifestations.

The analytical study of terrorism [8]: Taking stock

This article presents a review of the analytical study of terrorism that views all agents as rational decisionmakers.

The papers used the Global Terrorism Database (GTD), these include analyses of terrorist attack trends, the economic consequences of terrorism, the study of counterterrorism effectiveness, the causes of terrorism, and the relationship of terrorism and liberal democracies.

New developments in the field focused on distinguishing key differences between domestic and transnational terrorism. Other major developments involved the study of networked terrorists and the role of counterterrorism foreign aid.

Causality between terrorism and economic growth [9]:

This article analyzes the causal relationship between terrorism and economic growth, running a series of tests for a maximum of 160 countries from 1970 to 2007. The authors find that the causal relationship between terrorism and growth is heterogeneous over time and across space.

In order to examine the hypotheses regarding the terrorism–economy the data compiled on terrorism and economic growth for a maximum of 160 countries for the 1970–2007 period.

Exploring the relationship between global terrorist ideology and attack methodology [10]:

The paper used cluster analysis to test the hypothesis that a visual model could be developed to identify patterns in the last 45 years of global terrorism.

Terrorist attack methodologies of assassinations, armed assaults and bombings figured prominently in most of the segments, but national culture and ideology were also present as factors in some clusters.

Logistic regression was then applied to test the hypothesis that combinations of terrorist attack methodology and ideology could estimate the risk of a terrorist strike being successful.

The paper concluded that the odds ratio was statistically significant but the results were theoretically weak in supporting the hypothesis. Hijacking was 1.131 times more likely to be successful as compared with other types of global terrorist methodologies, and the relative risk of a terrorist strike being committed by foreign immigrants was 1.013 higher as compared with domestic terrorists.

Poverty, Political Freedom, and the Roots of Terrorism [11]:

The paper used a new dataset on terrorist risk world-wide, It failed to find a significant association between terrorism and economic variables such as income once the effect of other country characteristics is taken into account.

The estimates suggest that political freedom has a non-monotonic effect on terrorism. This result is consistent with the observed increase in terrorism for countries in transition from authoritarian regimes to democracies. The results also show that certain geographic characteristics may favor the presence of terrorist.

ISIL's Execution Videos: Audience Segmentation and Terrorist Communication in the Digital Age [12]:

This article offers a bottom-up understanding of the media strategy employed by the Islamic State of Iraq and the Levant (ISIL) as it relates to the production and dissemination of its hostage execution videos.

Through an empirical analysis of sixty-two videos of executions produced by ISIL in the year following its establishment as the "Islamic State" in 2014, this study examines the videos as a major component of ISIL's media strategy. Through these media products, ISIL seeks to spread a political message aimed at both local and global, ingroup and outgroup consumption through audience segmentation, while striving to influence both local and global audiences through the use and production of graphic violence. This article also discusses the strategy governing the production and release of ISIL's execution videos; how it relies on the global media to transmit its intertwined political and religious agenda in the digital media age.

The paper concluded that ISIL's use of execution video productions provides important context to the ways in which an emergent terrorist group frames its execution videos in terms of narrative, images, and ideology in order to set its political and religious agenda by relying on global and digital media to be spread and accessed. This empirical study has shed light on the complexity of the relationship between terrorism and the digital media age.

Terrorism Financing with Virtual Currencies [13]: Can Regulatory Technology Solutions Combat This?

This article considers the terrorism financing risk associated with the growth of Financial Technology innovations and in particular, focuses on virtual currency products and services. The ease with which cross-border payments by virtual currencies are facilitated, the anonymity surrounding their usage, and their potential to be converted into the fiat financial system, make them ideal for terrorism financing and therefore calls for a coordinated global regulatory response.

The paper concluded that the growth of the FinTech industry worldwide signals huge opportunities for businesses and consumers, it also introduces challenges to the global financial industry.

The challenge considered in this article is the potential of VC, a key FinTech innovation, to be used to finance terrorism activities.

This article has argued that, while financial regulatory regimes worldwide are at different stages of development and the robustness of law enforcement regimes worldwide vary, the threat of terrorism financing remains real—more so as terrorism financing is likely to shift to jurisdictions with weaker regimes.

Terrorism and Corruption [14]:

Alternatives for Goal Attainment Within Political Opportunity Structures

This study on the connection between corruption and political violence. It attempts to uncover whether domestic terrorism as a specific form of political violence.

The models required the measurement of the following key concepts:

political violence, perceived corruption, government coercion, regime repressiveness, cultural fractionalization, inequality, economic development, "heritage" of rebellion, transnational corporate penetration, level of education, and population.

The study used The Global Terrorism Database (GTD).

Analyses demonstrated that corruption and terrorist violence exhibit an inverse relationship, which supports the theoretical perspective that they are shared avenues within an extralegal opportunity structure, demonstrating that where the avenue of corruption has been restricted, countries experience greater rates of terrorist violence. This initial attempt to quantify the relationship between terrorist violence and corruption requires further analysis in order to identify under which conditions, in which regions, and over which time periods this relationship may vary.

Applying Analytical Methods to Study Terrorism [15]:

This paper study domestic and transnational terrorism insights.

For empirical applications, the paper focuses on the study of trend, cycles, and forecasting.

For forecasting Terrorism is to relate statistically the current number of terrorist incidents to their past values, time, and potential shocks or interventions (events such as a policy change or the rise in fundamentalism). There are some important forecasting insights. First, observed patterns in the time series can be used for forecasting purposes. Second, short-run forecasts are more accurate than long-run forecasts. Third, a time series of terrorist incidents can be forecast without knowing precisely why the numbers of incidents change in a particular regular pattern. Fourth, the ability to forecast patterns of, say, skyjackings does not mean that one can predict a particular catastrophic event such as 9 ⁄ 11.

PSYCHIATRY AND BEHAVIORAL SCIENCES [16]:

On the Radicalization Process

This study aimed to provide an in-depth description of the radicalization process. The author analysis is first based on the author's experience in the psychological evaluation of terrorist behavior and second on an exhaustive review of the current literature. The search terms "terrorism," "radicalization," "social psychology," and "psychopathology" were used to identify relevant studies.

The study performed with a focus on several aspects, such as radicalization risk factors, brainwashing, the role of the media, and finally, in de-radicalization programs.

The paper concluded that the radicalization process is an increasing and complex phenomenon. It implies several aspects that are multidimensional (on the individual and societal levels) and heterogeneous, such as some individual risk factors, the brainwashing, and cognitive modifications (i.e. the role of the media in general).
In terms of counter-radicalization programs, have shown mixed results. The most successful efforts in Britain have been the efforts of the so-called Channel program, which is part of the British government's counter-terrorism strategy, to divert young people from extremism. Such efforts, which involve the police, social services, and local authorities working together, draw on methods used to help young people leave gangs.
On a more practical note, it is clear there are currently not enough detailed case studies of terrorists to inform psychological analyses or even to conduct comprehensive reviews of the literature.

Dataset:

The Global Terrorism Database (GTD) is an open-source database including information on terrorist attacks around the world from 1970 through 2017. The GTD includes systematic data on domestic as well as international terrorist incidents that have occurred during this time period and now includes more than 180,000 attacks.

Content:

Geography: Worldwide.
Time period: 1970-2017, *except 1993.*
Variables: >135 variables on location, tactics, perpetrators, targets, and outcomes ,….etc

The source of the dataset:

https://www.kaggle.com/START-UMD/gtd#globalterrorismdb_0718dist.csv

GTD Definition of Terrorism and Inclusion Criteria:

The GTD defines a terrorist attack as *the threatened or actual use of illegal force and violence by a non-state actor to attain a political, economic, religious, or social goal through fear, coercion, or intimidation*. In practice this means in order to consider an incident for inclusion in the GTD, *all three* of the following attributes must be present:

- ***The incident must be intentional*** – the result of a conscious calculation on the part of a perpetrator.
- ***The incident must entail some level of violence or immediate threat of violence*** -including property violence, as well as violence against people.
- ***The perpetrators of the incidents must be sub-national actors.*** The database does not include acts of state terrorism.

In addition, *at least two* of the following three criteria must be present for an incident to be included in the GTD:

- **Criterion 1: The act must be aimed at attaining a political, economic, religious, or social goal**. In terms of economic goals, the exclusive pursuit of profit does not satisfy this criterion. It must involve the pursuit of more profound, systemic economic change.
- **Criterion 2*:* There must be evidence of an intention to coerce, intimidate, or convey some other message to a larger audience (or audiences) than the immediate victims**. It is the act taken as a totality that is considered, irrespective if every individual involved in carrying out the act was aware of this intention. As long as any of the planners or decision-makers behind the attack intended to coerce, intimidate or publicize, the intentionality criterion is met.
- **Criterion 3: The action must be outside the context of legitimate warfare activities**. That is, the act must be outside the parameters permitted by international humanitarian law (particularly the prohibition against deliberately targeting civilians or non-combatants).

Additional Filtering Mechanism: “Doubt Terrorism Proper?”

- The inclusion criteria above are evaluated for each case to determine if it should be added to the GTD; however, there is often definitional overlap between terrorism and other forms of crime and political violence, such as insurgency, hate crime, and organized crime. Likewise, for many cases there is insufficient or conflicting information provided in source documents to allow coders to make a clear determination regarding whether or not the inclusion criteria are met. Such uncertainty, however, was not deemed to be sufficient to disqualify the incident from inclusion in the GTD. Users of the GTD can further govern the parameters of their search results by employing an additional terrorism definitional filter.
- The “Doubt Terrorism Proper” field records reservation reported in source materials that the incident in question is *exclusively* terrorism. Furthermore, such a determination of doubt is subsequently coded by GTD analysts as conforming to one of five possible alternative designations: 1) Insurgency/Guerilla Action; 2) Other Crime Type; 3) Intra/Inter-group conflict; 4) Lack of Intentionality; or 5) State Actor. As is the case with the criteria filters outlined above, the “Doubt Terrorism Proper” filter is available for use on the GTD’s Advanced Search Page. Note that the “Doubt Terrorism Proper” determination was only routinely made for incidents that occurred after 1997 and the “State Actor” was only routinely made for incidents that occurred after 2012.

Data label / Class (the dependent variable)

- *(success)*
- *Categorical Variable*
- Success of a terrorist strike is defined according to the tangible effects of the attack. Success is *not* judged in terms of the larger goals of the perpetrators. For example, a bomb that exploded in a building would be counted as a success even if it did not succeed in bringing the building down or inducing government repression.
- The definition of a successful attack depends on the type of attack. Essentially, the key question is whether or not the attack type took place. If a case has multiple attack types, it is successful if any of the attack types are successful, with the exception of assassinations, which are only successful if the intended target is killed.
- 1 = "Yes" The incident was successful.
- 0 = "No" The incident was not successful.

Example of Success of a terrorist strike

- **ASSASSINATION**
- In order for an assassination to be successful, the target of the assassination must be killed. For example, even if an attack kills numerous people but not the target, it is an unsuccessful assassination.
- **ARMED ASSAULT**
- An armed assault is determined to be successful if the assault takes place and if a target is hit (including people and/or property). Unsuccessful armed assaults are those in which the perpetrators attack and do not hit the target. An armed assault is also unsuccessful if the perpetrators are apprehended on their way to commit the assault. To make this determination, however, there must be information to indicate that an actual assault was imminent.
- **BOMBING/EXPLOSION**
- A bombing is successful if the bomb or explosive device detonates. Bombings are considered unsuccessful if they do not detonate. The success or failure of the bombing is not based on whether it hit the intended target.
- **HIJACKING**
- A hijacking is successful if the hijackers assume control of the vehicle at any point, whereas a hijacking is unsuccessful if the hijackers fail to assume control of the vehicle. The success or failure of the hijacking is not based on whether the vehicle reached the intended destination of the hijackers.
- **HOSTAGE TAKING (BARRICADE INCIDENT)**
- A barricade incident is successful if the hostage takers assume control of the individuals at any point, whereas a barricade incident is unsuccessful if the hostage takers fail to assume control of the individuals.
- **HOSTAGE TAKING (KIDNAPPING)**
- A kidnapping is successful if the kidnappers assume control of the individuals at any point, whereas a kidnapping is unsuccessful if the kidnappers fail to assume control of the individuals.

- **FACILITY / INFRASTRUCTURE ATTACK**
- A facility attack is determined to be successful if the facility is damaged. If the facility has not been damaged, then the attack is unsuccessful.
- **UNARMED ASSAULT**
- An unarmed assault is determined to be successful there is a victim that who has been injured. Unarmed assaults that are unsuccessful are those in which the perpetrators do not injure anyone. An unarmed assault is also unsuccessful if the perpetrators are apprehended when on their way to commit the assault. To make this determination, however, there must be information to indicate that an assault was imminent.

Attributes:

GTD ID and Date	eventid	Numeric
	iyear	Numeric
	imonth	Numeric
	iday	Numeric
	approxdate	Text
	extended	Categorical
	resolution	Numeric Date
Incident Information	summary	Text
	crit1	Categorical
	crit2	Categorical
	crit3	Categorical
	doubtterr	Categorical
	alternative	Categorical
	alternative_txt	Categorical
	multiple	Categorical
	related	Text
Incident Location	country	Categorical
	country_txt	Categorical
	region	Categorical
	region_txt	Categorical
	provstate	Text
	city	Text
	vicinity	Categorical
	location	Text
	latitude	Numeric
	longitude	Numeric
	specificity	Categorical
Attack Information	attacktype1	Categorical
	attacktype1_txt	Categorical
	attacktype2	Categorical
	attacktype2_txt	Categorical
	attacktype3	Categorical
	attacktype3_txt	Categorical
	success	Categorical
	suicide	Categorical
Additional Information & Sources	addnotes	Text
	INT_LOG	Categorical
	INT_IDEO	Categorical
	INT_MISC	Categorical
	INT_ANY	Categorical
	scite1	Text
	scite2	Text
	scite3	Text
	dbsource	Text
Weapon Information	weaptype1	Categorical
	weaptype1_txt	Categorical
	weapsubtype1	Categorical
	weapsubtype1_txt	Categorical
	weaptype2	Categorical
	weaptype2_txt	Categorical
	weapsubtype2	Categorical
	weapsubtype2_txt	Categorical
	weaptype3	Categorical
	weaptype3_txt	Categorical
	weapsubtype3	Categorical
	weapsubtype3_txt	Categorical
	weaptype4	Categorical
	weaptype4_txt	Categorical
	weapsubtype4	Categorical
	weapsubtype4_txt	Categorical
	weapdetail	Text
Target/Victim Information	targtype1	Categorical
	targtype1_txt	Categorical
	targsubtype1	Categorical
	targsubtype1_txt	Categorical
	corp1	Text
	target1	Text
	natlty1	Categorical
	natlty1_txt	Categorical
	targtype2	Categorical
	targtype2_txt	Categorical
	targsubtype2	Categorical
	targsubtype2_txt	Categorical
	corp2	Text
	target2	Text
	natlty2	Categorical
	natlty2_txt	Categorical
	targtype3	Categorical
	targtype3_txt	Categorical
	targsubtype3	Categorical
	targsubtype3_txt	Categorical
	corp3	Text
	target3	Text
	natlty3	Categorical
	natlty3_txt	Categorical
Perpetrator Information	gname	Text
	gsubname	Text
	gname2	Text
	gsubname2	Text
	gname3	Text
	gsubname3	Text
	guncertain1	Categorical
	guncertain2	Categorical
	guncertain3	Categorical
	individual	Categorical
	nperps	Numerical
	nperpcap	Numerical
	claimed	Categorical
	claimmode	Categorical
	claimmode_txt	Categorical
	compclaim	Categorical
	claim2	Categorical
	claimmode2	Categorical
	claimmode2_txt	Categorical
	claim3	Categorical
	claimmode3	Categorical
	claimmode3_txt	Categorical
	motive	Text
Casualties and Consequences	nkill	Numeric
	nkillus	Numeric
	nkillter	Numeric
	nwound	Numeric
	nwoundus	Numeric
	nwoundte	Numeric
	property	Categorical
	propextent	Categorical
	propextent_txt	Categorical
	propvalue	Numeric
	propcomment	Text
	ishostkid	Categorical
	nhostkid	Numeric
	nhostkidus	Numeric
	nhours	Numeric
	ndays	Numeric
	divert	Text
	kidhijcountry	Text
	ransom	Categorical
	ransomamt	Numeric
	ransomamtus	Numeric
	ransompaid	Numeric
	ransompaidus	Numeric
	ransomnote	Text
	hostkidoutcome	Categorical
	hostkidoutcome_txt	Categorical
	nreleased	Numeric

Attributes Descriptions:

Please refer to the table in the link below:

(if the embedded link did not work the excel file is attached separately).

attributes.xlsx

Approach Introduction:

- I Build my approach based on the above Literature Review.
- The classifiers in the reviews above used various models such as SVM, Random forest, K-NN and some other used logistic regression classifier or decision tree.
- Some papers used different datasets on the same subject of study and other used the same dataset I am using here.
- Also the difference between paper is selecting the target variables (success of the attack or causalities, ..etc)
- And the scope of the study (Global, or certain country or region or certain terrorism group).
- The selection of features also different between papers based on the aim of the study.
- I decided to try different classification model and select the top performance one that fits my classification problem.
- And below is my approach diagram.

Approach:

initial Analysis

- **Univariate Analysis** (detailed data dictionary / list all variables and understand what each variable mean and represent)(what is the source of each variable directly measured or calculated based on other variables or does the variable has a time domain associated) - (Decide on the dependent (target) variable) (Assigning the correct data types and appropriate column names)(for numeric attributes check the 5 number summary min 1sQ mean 3rdQ max)(for categorical attributes check and decide on the levels , frequency) (check and deal with data inconsistencies , missing values , errors , duplicates , outliers (boxplots) , numeric signs , upper and lower cases , spaces or special characters in striges) (check distributions of the variables : Normal distribution?)(Low variance filter)(check the imbalance in the dependent variable) (check time variable) (Univariate visualizations)
- **BiVariate Analysis** (pairwise relations) (pairwise visuliazations like scatter plots) (correlation analysis spearman or pearson).
- **Multivariate Analysis** (relations between more than 2 variables)(statistical tools such as one way ANOVA analysis or rank or to compare the means.

Exploratory Analysis

- Normalizing / scaling
- Sub-setting the data
- Decision rules , association rules , n grams
- Clustering such as K-mean
- Hypothesise analysis
- Correlation analysis
- NLP analysis for some text attributes

Dimensionality Reduction

- remove attributes with too many missing values
- remove attributes with zero or very low variance
- remove one of the attributes with high correlations with other - prefer the one with more missing values or lower variance
- Feature selection(decide on the importance of the attribute using statistical measures like information gain or Gini index) (Forward selection and backward elimination)

Experimental Desgin

- Randomizing, Splitting the data into training and test sets to training, validation and testing
- Treatment for imbalance (under-sampling the majority class and over-sampling the minority class)
- Cross validation such as 10 folds

Modeling

- building the Classification Models
- Train the model
- Validate the model

Evaluation

- Classification (Confusion matric , ROC (receiver operating characteristics) , Accuracy , Recall , Precision)
- compare performance between different models (contingency tables , multivariate analysis of variance)

Improving the Model

- deal with very high accuracy(Fixation variables , overfitting)
- deal with very low performance (many need new variables or more observations)
- iterate to improve the model performance

Conclusions

- Inferences , dissuasions
- Threats to validity (Internal , External , construct) and propose the solution that may mitigate these threats
- Next/ future steps

Project Github link:
https://github.com/emilkaram/CKM136XJ0-Global-Terrorism-Data-Analytics-Capstone

Initial Analysis: Univariant Analysis:

Checking data shape:

```
mydata.shape
```

```
(181691, 135)
```

Checking data frame info:

```
mydata.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 181691 entries, 0 to 181690
Columns: 135 entries, eventid to dbsource
dtypes: float64(56), int64(21), object(58)
memory usage: 187.1+ MB
```

Checking data description:

```
obj1.describe()
```

	approxdate	resolution	summary	alternative_txt	related	country_txt	region_txt	provstate	city
count	9239	2220	115562	29011	25038	181691	181691	181270	181257
unique	2244	1859	112492	5	14306	205	12	2855	36674
top	September 18-24, 2016	8/04/98	09/00/2016: Sometime between September 18, 201...	Insurgency/Guerilla Action	201612010023, 201612010024, 201612010025, 2016...	Iraq	Middle East & North Africa	Baghdad	Unknowr
freq	101	18	100	23410	80	24636	50474	7645	9775

Checking data column:

```
mydata.columns
```

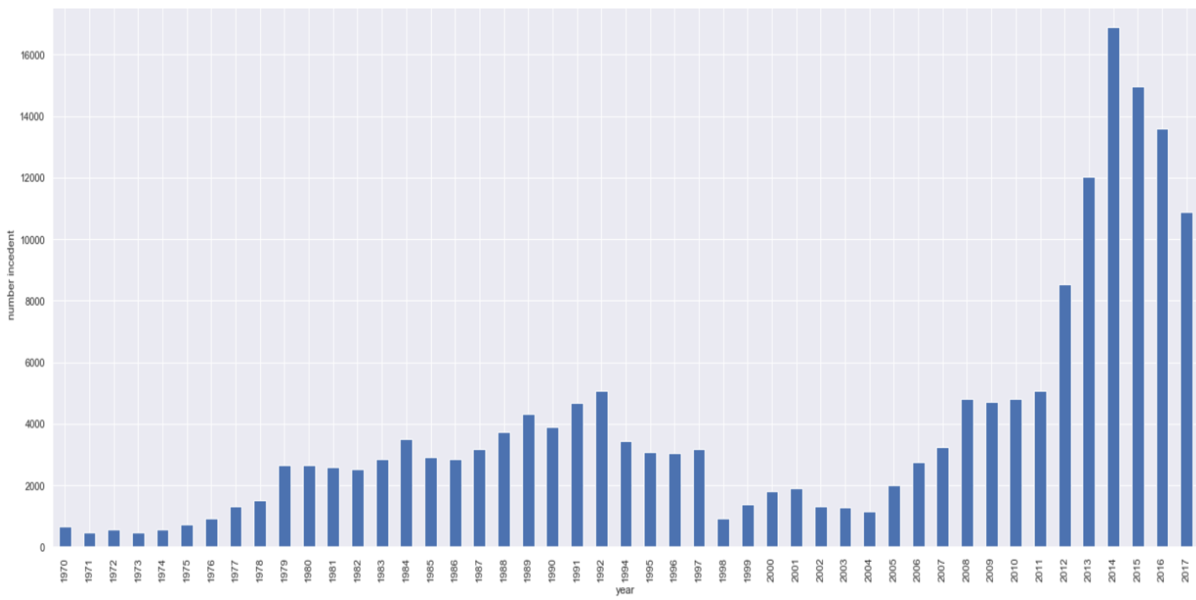
```
Index(['eventid', 'iyear', 'imonth', 'iday', 'approxdate', 'extended',
       'resolution', 'summary', 'crit1', 'crit2',
       ...
       'nreleased', 'addnotes', 'INT_LOG', 'INT_IDEO', 'INT_MISC', 'INT_ANY',
       'scite1', 'scite2', 'scite3', 'dbsource'],
      dtype='object', length=135)
```

Checking 5 numbers summary:

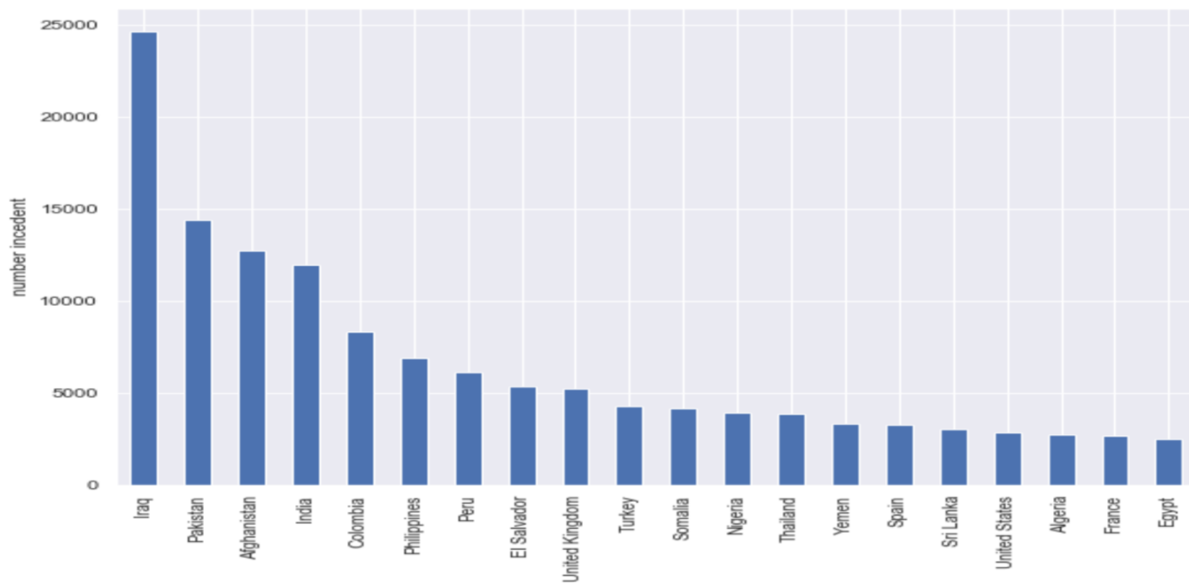
```
mydata.describe()
```

	eventid	iyear	imonth	iday	extended	crit1	crit2
count	1.816910e+05	181691.000000	181691.000000	181691.000000	181691.000000	181691.000000	181691.000000
mean	2.002704e+11	2002.638997	6.467277	15.505644	0.045346	0.988530	0.993093
std	1.325955e+09	13.259430	3.388303	8.814045	0.208063	0.106483	0.082823
min	1.970000e+11	1970.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	1.991020e+11	1991.000000	4.000000	8.000000	0.000000	1.000000	1.000000
50%	2.009020e+11	2009.000000	6.000000	15.000000	0.000000	1.000000	1.000000
75%	2.014080e+11	2014.000000	9.000000	23.000000	0.000000	1.000000	1.000000
max	2.017120e+11	2017.000000	12.000000	31.000000	1.000000	1.000000	1.000000

Checking number of attacks yearly distribution:



Checking number of attacks per country distribution:



Exploring categorial data levels:

```
#list the levels
mydata['country_txt'].unique()
```

```
array(['Dominican Republic', 'Mexico', 'Philippines', 'Greece', 'Japan',
       'United States', 'Uruguay', 'Italy', 'East Germany (GDR)',
       'Ethiopia', 'Guatemala', 'Venezuela', 'West Germany (FRG)',
       'Switzerland', 'Jordan', 'Spain', 'Brazil', 'Egypt', 'Argentina',
       'Lebanon', 'Ireland', 'Turkey', 'Paraguay', 'Iran',
       'United Kingdom', 'Colombia', 'Bolivia', 'Nicaragua',
       'Netherlands', 'Belgium', 'Canada', 'Australia', 'Pakistan',
       'Zambia', 'Sweden', 'Costa Rica', 'South Yemen', 'Cambodia',
       'Israel', 'Poland', 'Taiwan', 'Panama', 'Kuwait',
       'West Bank and Gaza Strip', 'Austria', 'Czechoslovakia', 'India',
       'France', 'South Vietnam', 'Brunei', 'Zaire',
       "People's Republic of the Congo", 'Portugal', 'Algeria',
       'El Salvador', 'Thailand', 'Haiti', 'Sudan', 'Morocco', 'Cyprus',
       'Myanmar', 'Afghanistan', 'Peru', 'Chile', 'Honduras',
       'Yugoslavia', 'Ecuador', 'New Zealand', 'Malaysia', 'Singapore',
       'Botswana', 'Jamaica', 'Chad', 'North Yemen', 'Andorra', 'Syria',
       'South Korea', 'United Arab Emirates', 'South Africa', 'Kenya',
```

Exploring categorial data value counts:

```
mydata['gname'].value_counts()
```

```
Unknown                                              82782
Taliban                                               7478
Islamic State of Iraq and the Levant (ISIL)           5613
Shining Path (SL)                                     4555
Farabundo Marti National Liberation Front (FMLN)      3351
Al-Shabaab                                            3288
New People's Army (NPA)                               2772
Irish Republican Army (IRA)                           2671
Revolutionary Armed Forces of Colombia (FARC)         2487
Boko Haram                                            2418
Kurdistan Workers' Party (PKK)                        2310
Basque Fatherland and Freedom (ETA)                   2024
Communist Party of India - Maoist (CPI-Maoist)        1878
Maoists                                               1630
Liberation Tigers of Tamil Eelam (LTTE)               1606
National Liberation Army of Colombia (ELN)            1561
Tehrik-i-Taliban Pakistan (TTP)                       1351
Palestinians                                          1125
Houthi extremists (Ansar Allah)                       1062
Al-Qaida in the Arabian Peninsula (AQAP)              1020
```

Exploring categorial data summary:

```
mydata['success'].describe()
```

```
count    181691.000000
mean          0.889598
std           0.313391
min           0.000000
25%           1.000000
50%           1.000000
75%           1.000000
max           1.000000
Name: success, dtype: float64
```

```
mydata['country_txt'].describe()
```

```
count     181691
unique       205
top         Iraq
freq       24636
Name: country_txt, dtype: object
```

```
# Frequency
mydata['country_txt'].value_counts()
```

```
Iraq                                  24636
Pakistan                              14368
Afghanistan                           12731
India                                 11960
Colombia                               8306
Philippines                            6908
Peru                                   6096
El Salvador                            5320
United Kingdom                         5235
Turkey                                 4292
Somalia                                4142
Nigeria                                3907
Thailand                               3849
Yemen                                  3347
```

Cleaning data:

Select only records that meet the terrorism attack criteria:

- Criterion 1: The act must be aimed at attaining a political, economic, religious, or social goal.
- Criterion 2: There must be evidence of an intention to coerce, intimidate, or convey some other message to a larger audience (or audiences) than the immediate victims.
- Criterion 3: The action must be outside the context of legitimate warfare activities.

```
mydata['doubtterr'] = mydata['doubtterr'].replace(-9,0)
```

```
#mydata.doubtterr
```

```
mydata = mydata[(mydata.crit1 == 1) & (mydata.crit2 == 1) & (mydata.crit3 == 1) & (mydata.doubtterr == 0)]
```

```
mydata.shape
```

```
(152622, 135)
```

Dealing with null values:

Removing attributes with more than 75% Null values:

```
# find Null value , calculate % of missing value from total , sort de-ascenc
list1 =100*mydata.isnull().sum()/mydata.shape[0]
list1.sort_values(ascending=False)
```

```
alternative_txt      99.996069
alternative          99.996069
gsubname3            99.994103
weapsubtype4_txt     99.961342
weapsubtype4         99.961342
weaptype4_txt        99.958722
weaptype4            99.958722
claimmode3           99.929892
claimmode3_txt       99.929892
gsubname2            99.911546
guncertain3          99.836197
claim3               99.836197
gname3               99.833576
divert               99.803436
attacktype3_txt      99.741191
attacktype3          99.741191
ransomnote           99.701223
claimmode2_txt       99.678945
claimmode2           99.678945
ransompaidus         99.676325
ransomamtus          99.669772
ransompaid           99.549213
corp3                99.384754
targsubtype3_txt     99.340200
targsubtype3         99.340200
```

```python
#droping more than 75% Null values attributes
mylist = []
for i in mydata:
    if 100*mydata[i].isnull().sum()/mydata.shape[0] > 75:
        mylist.append(i)
```

```python
mydata = mydata.drop(mylist,axis=1)
```

```python
mydata.shape
```

```
(152622, 65)
```

Filling null values:

```python
#fill null vaules
mydata['weapsubtype1_txt'].fillna('No Record', inplace=True)
mydata['natlty1_txt'].fillna('unknown', inplace=True)
mydata['target1'].fillna('unknown', inplace=True)
mydata['city'].fillna('unknown', inplace=True)
mydata['provstate'].fillna('unknown', inplace=True)

mydata['country_txt'].fillna('unknown', inplace=True)
mydata['region_txt'].fillna('unknown', inplace=True)
mydata['attacktype1_txt'].fillna('unknown', inplace=True)
mydata['targtype1_txt'].fillna('unknown', inplace=True)
```

```python
#replaceve unk with unknown
mydata.target1 = mydata.target1.replace('unk','unknown')
```

```python
# fill missing value for nkill and nwound with the median
mydata.nkill = np.round(mydata.nkill.fillna(mydata.nkill.median())).astype(int)
mydata.nwound = np.round(mydata.nwound.fillna(mydata.nwound.median())).astype(int)
```

```python
mydata.weaptype1_txt.replace(
    'Vehicle (not to include vehicle-borne explosives, i.e., car or truck bombs)',
    'Vehicle', inplace = True)
```

Text data set to lower case:

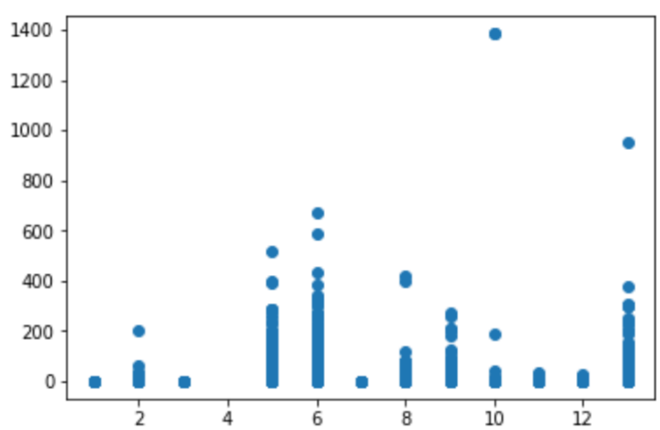
```
# set text to lower case
mydata.target1 = mydata.target1.str.lower()
mydata.gname = mydata.gname.str.lower()
mydata.summary = mydata.summary.str.lower()
mydata.city = mydata.city.str.lower()
mydata.weapsubtype1_txt = mydata.weapsubtype1_txt.str.lower()
mydata.natlty1_txt = mydata.natlty1_txt.str.lower()
mydata.provstate = mydata.provstate.str.lower()

mydata.country_txt = mydata.country_txt.str.lower()
mydata.region_txt =  mydata.region_txt.str.lower()
mydata.attacktype1_txt  = mydata.attacktype1_txt.str.lower()
mydata.targtype1_txt   = mydata.targtype1_txt.str.lower()
```

Initial Analysis: Bivariate Analysis

Scatter plot showing weapon type vs number of killed:

```
plt.scatter(mydata['weaptype1'] , mydata['nkill'])
```

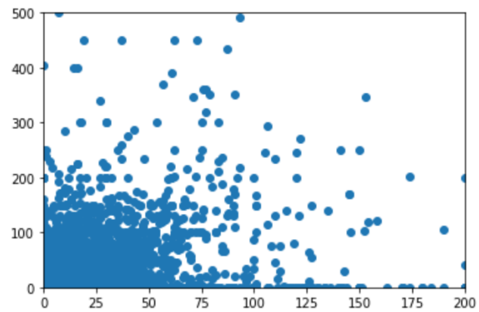
```
<matplotlib.collections.PathCollection at 0x1dcae84b780>
```



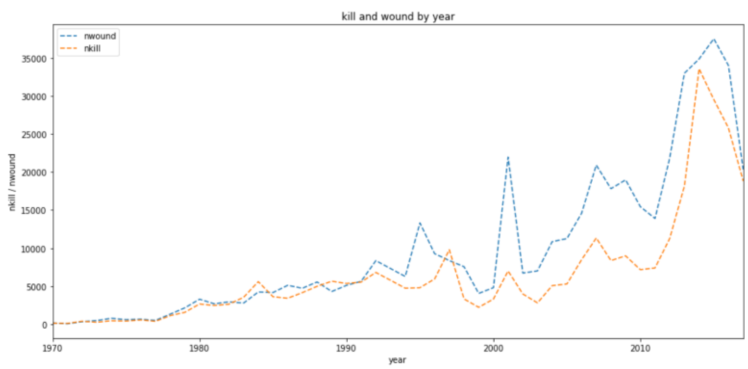
Scatter plot showing number of wounded vs number of killed:

```
ax =plt.subplot(1,1,1)
ax.scatter(mydata['nkill'] , mydata['nwound'])
ax.set_xlim([0,200])
ax.set_ylim([0,500])
```

```
(0, 500)
```



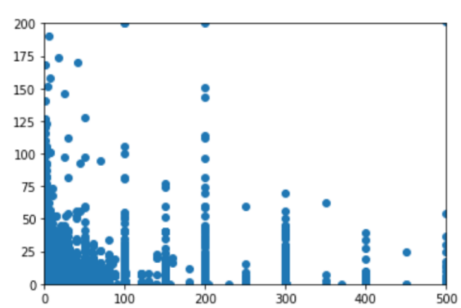
Line graph showing number of wounded and killed:



Scatter plot showing number of perpetrators vs number of killed:

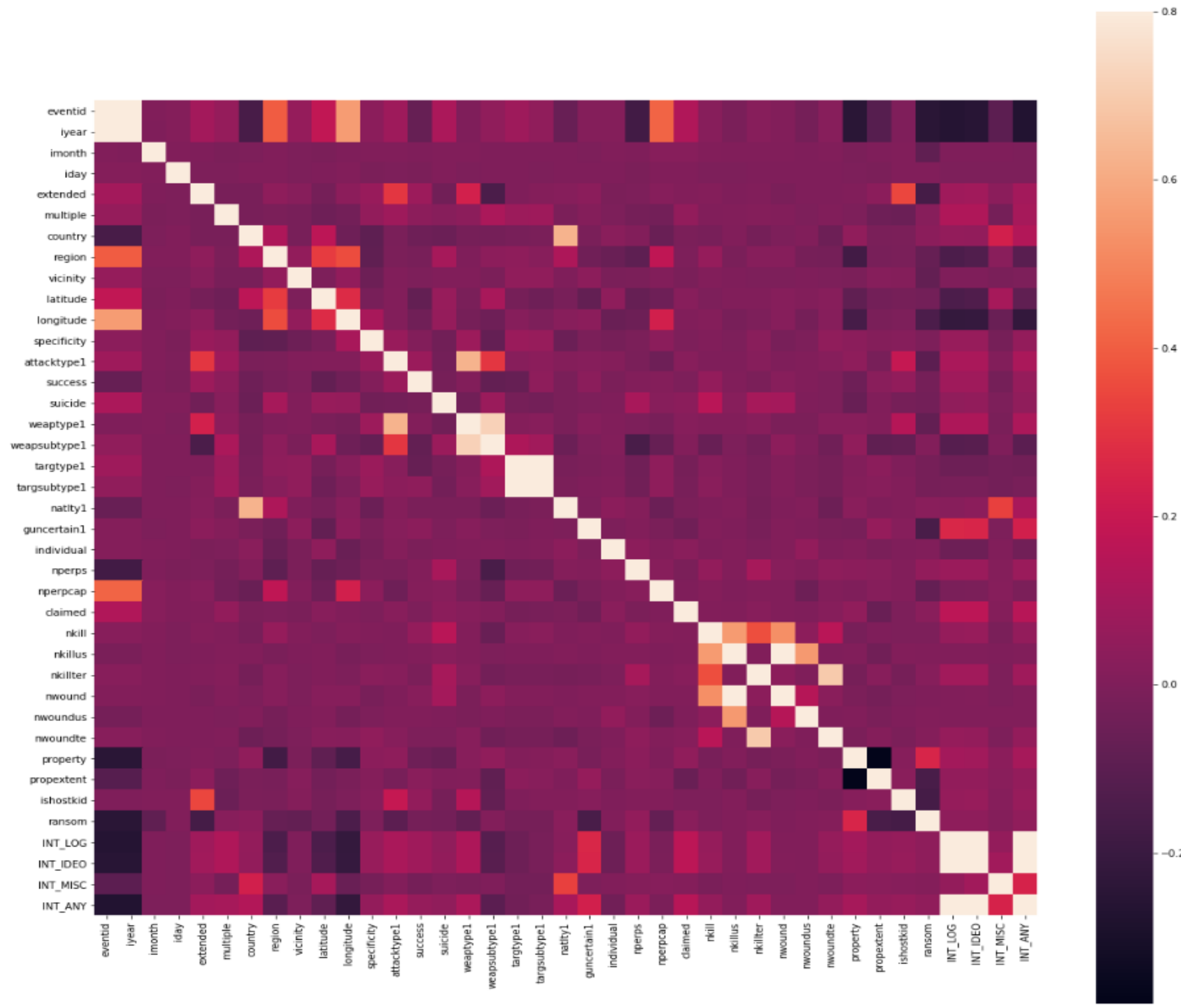
```
ax =plt.subplot(1,1,1)
ax.scatter(mydata['nperps'] , mydata['nkill'])
ax.set_xlim([0,500])
ax.set_ylim([0,200])
```

```
(0, 200)
```



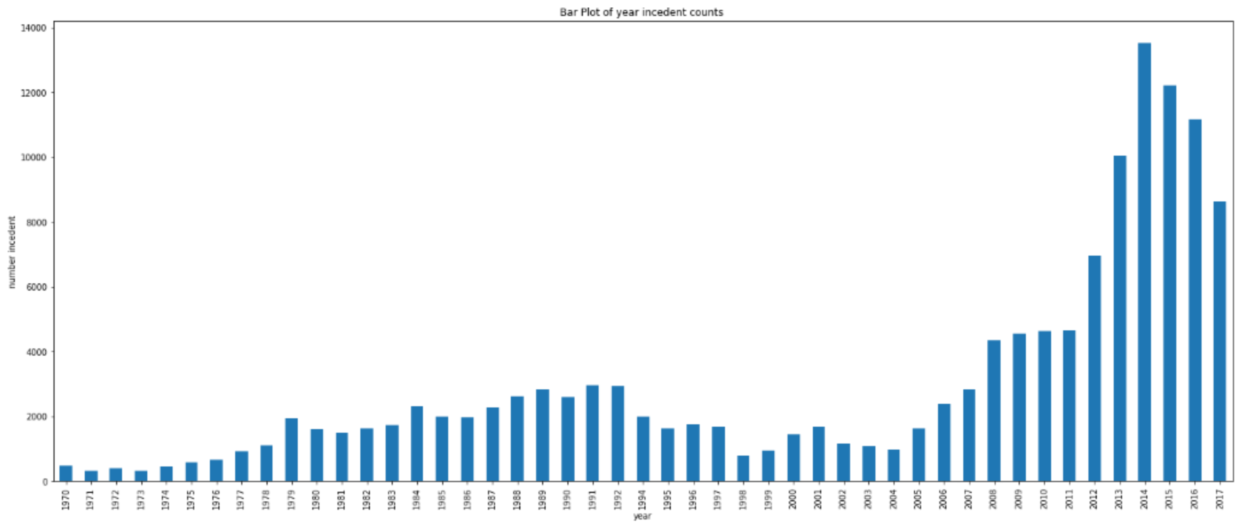
Initial analysis: Multivariate Analysis:

Correlation graph showing correlation levels between attributes:

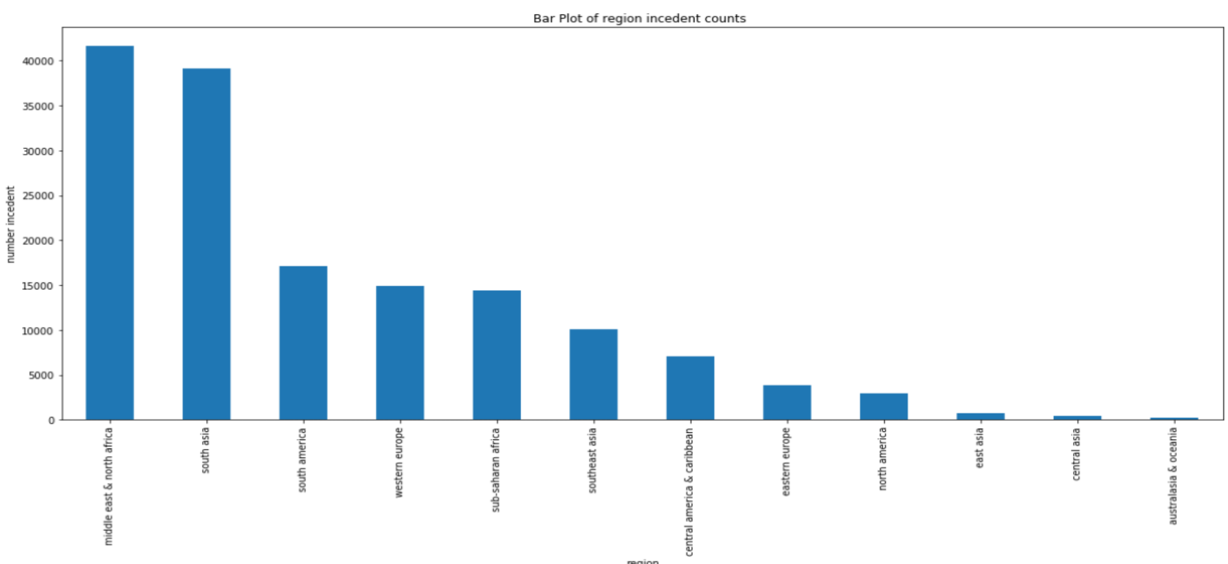


Exploratory Analysis:

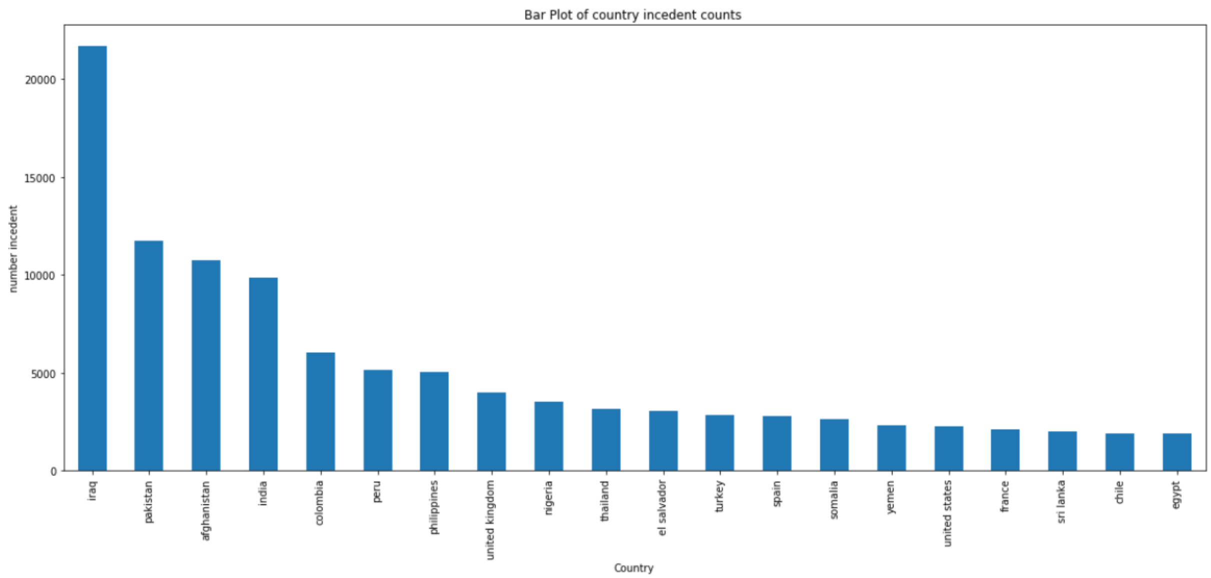
Number of attacks yearly distribution: sudden jump post year 2007 and trending to increase.



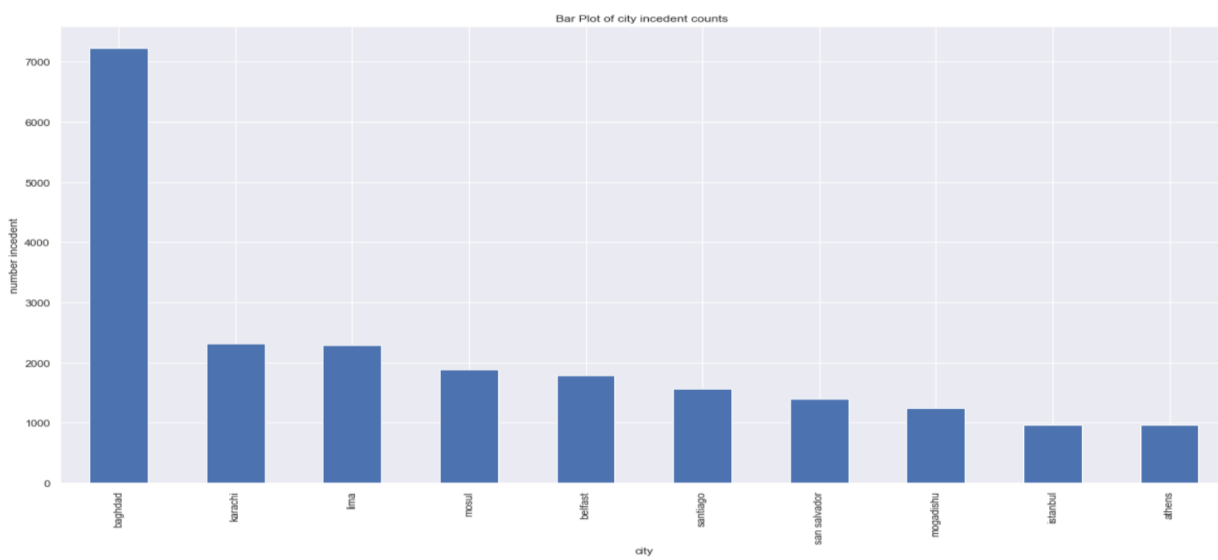
Number of attacks region distribution: Top 2 regions are Middle East and South Asia.



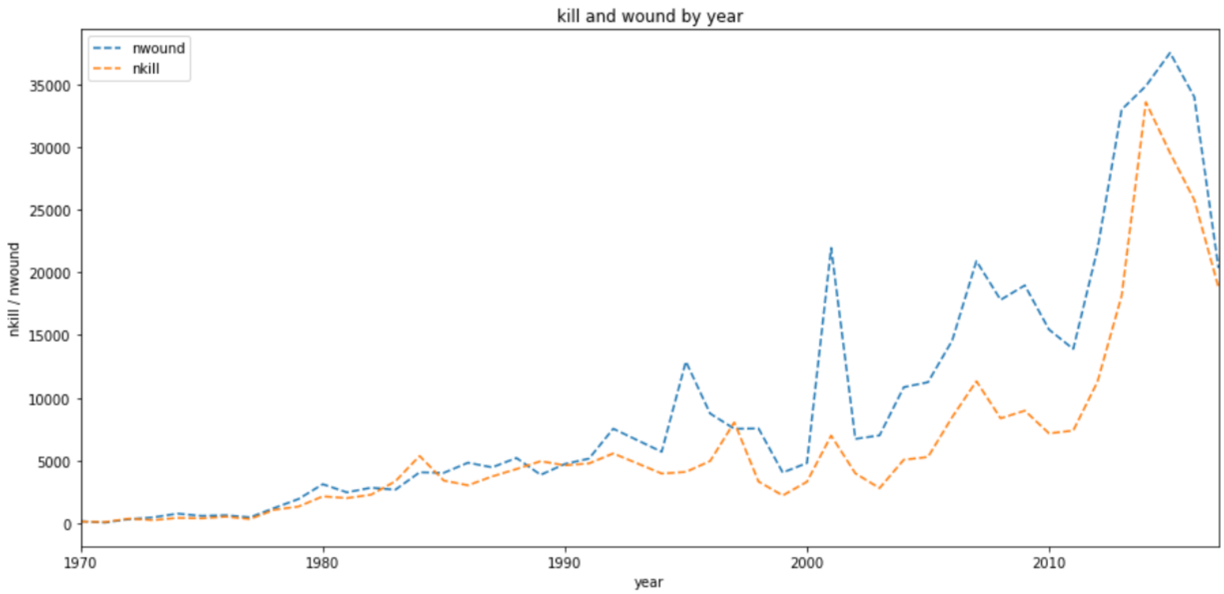
Number of attacks country distribution: Top 3 countries are Iraq, Pakistan and Afghanistan.



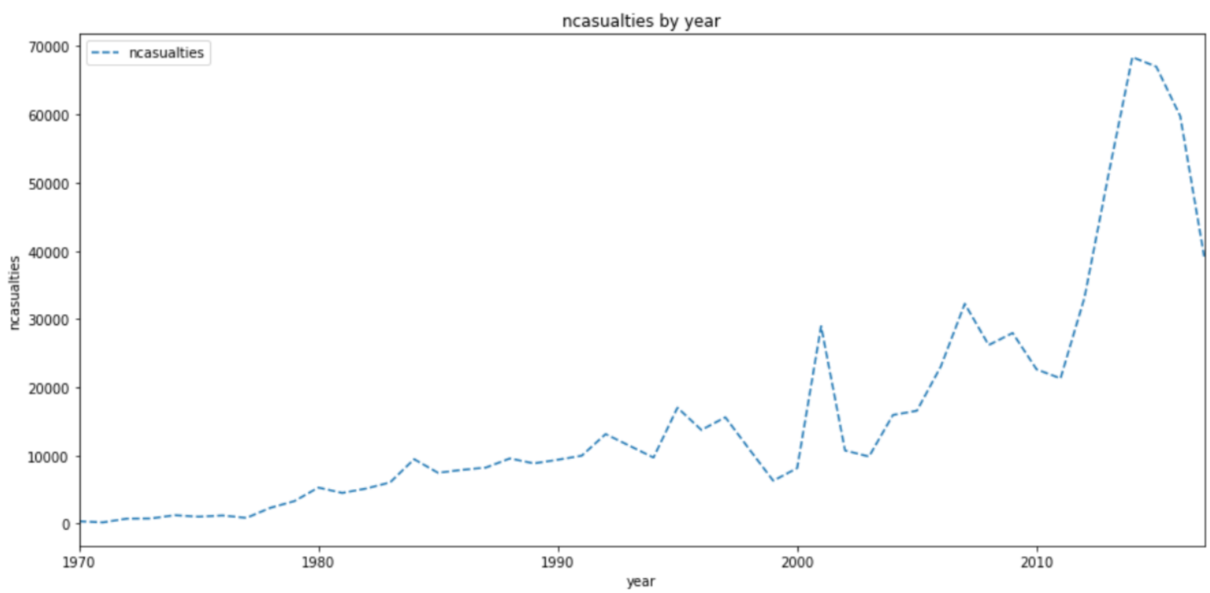
Number of attacks city distribution: Top 2 cities are Baghdad, Karachi.



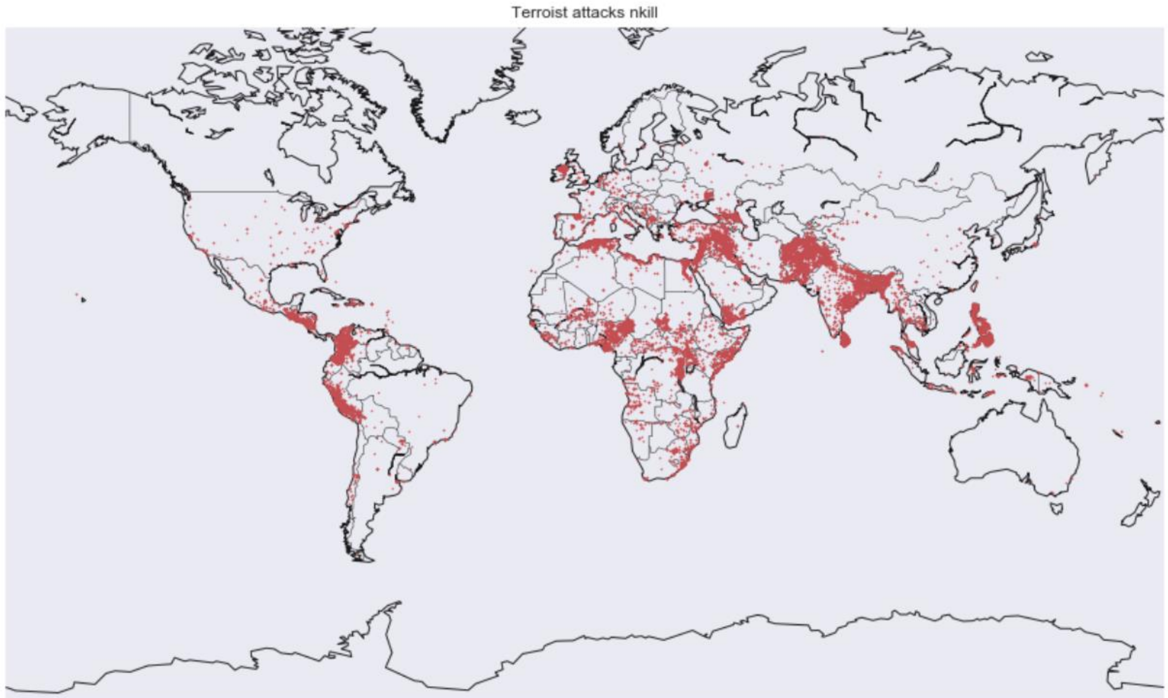
Line graph for Wounded and killed over the years:



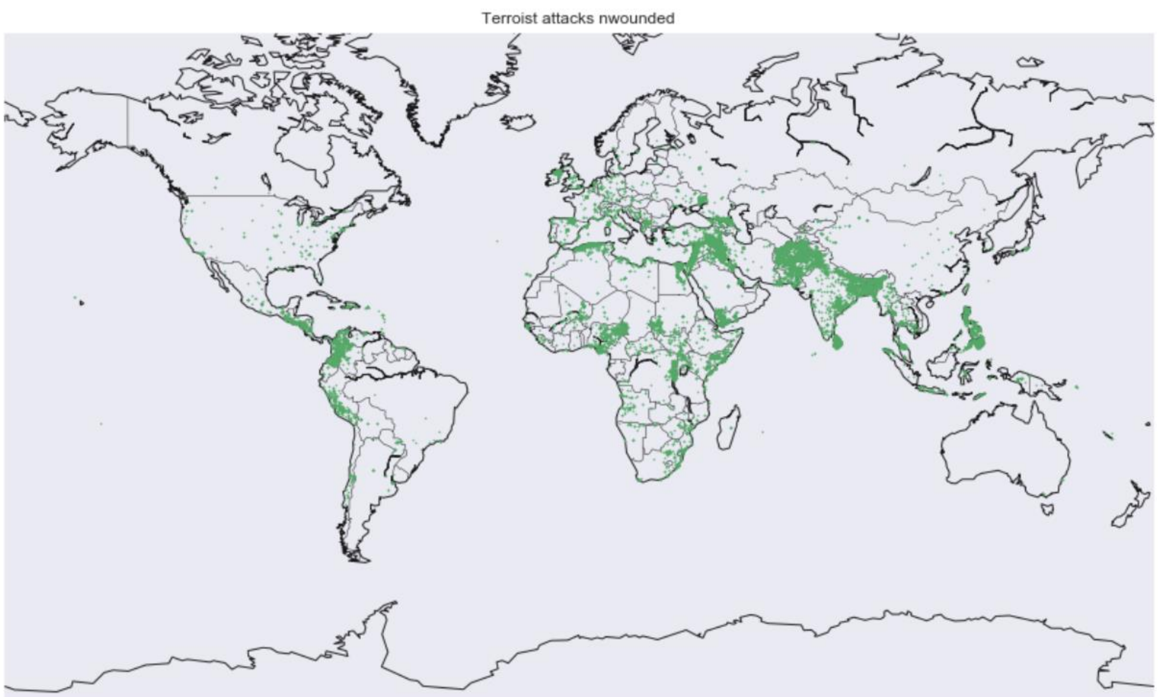
Line graph for casualties = wounded + killed over the years:



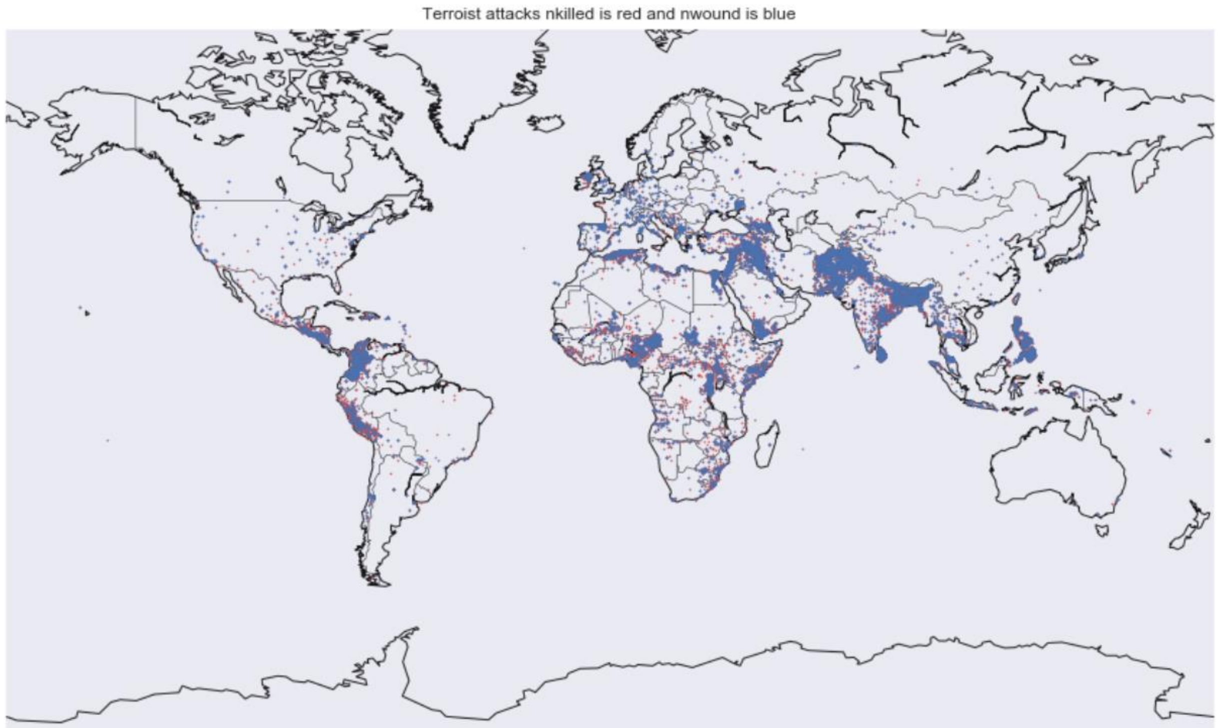
Number of killed world distribution from 1970 to 2017:



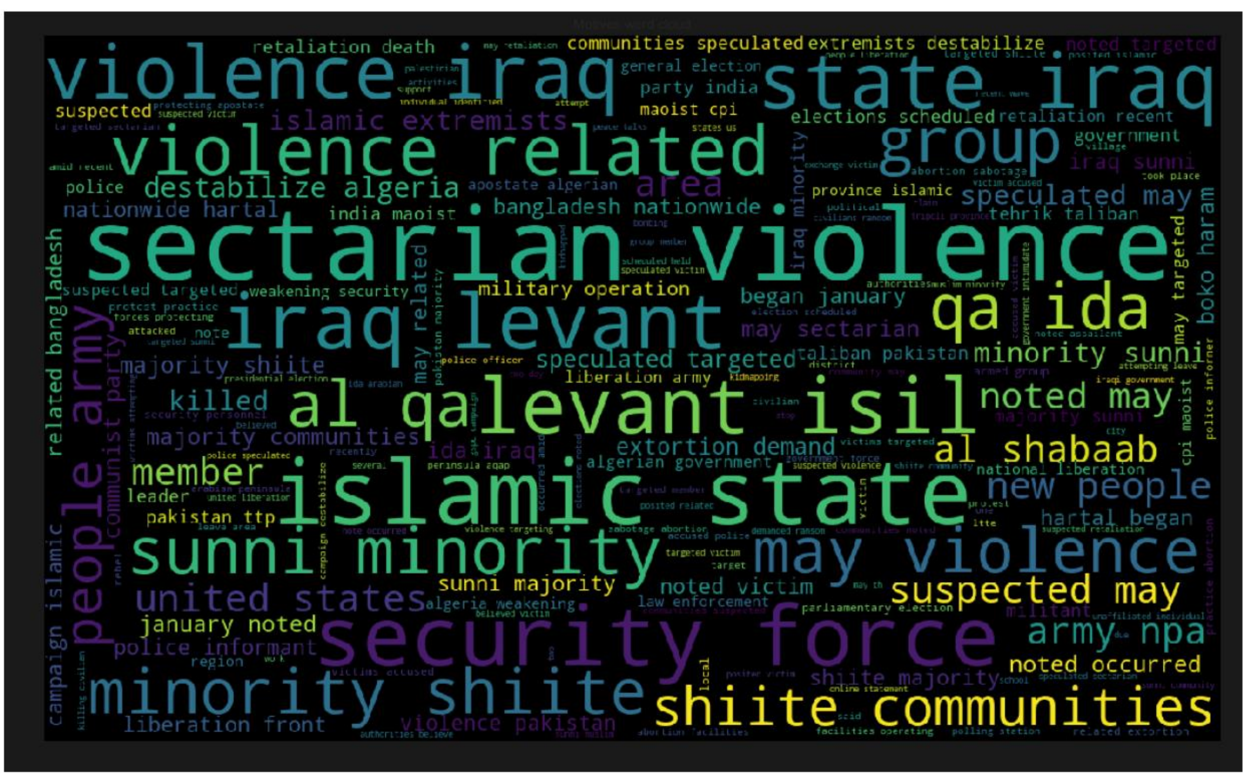
Number of wounded world distribution from 1970 to 2017:



Number of wounded and killed world distribution from 1970 to 2017:

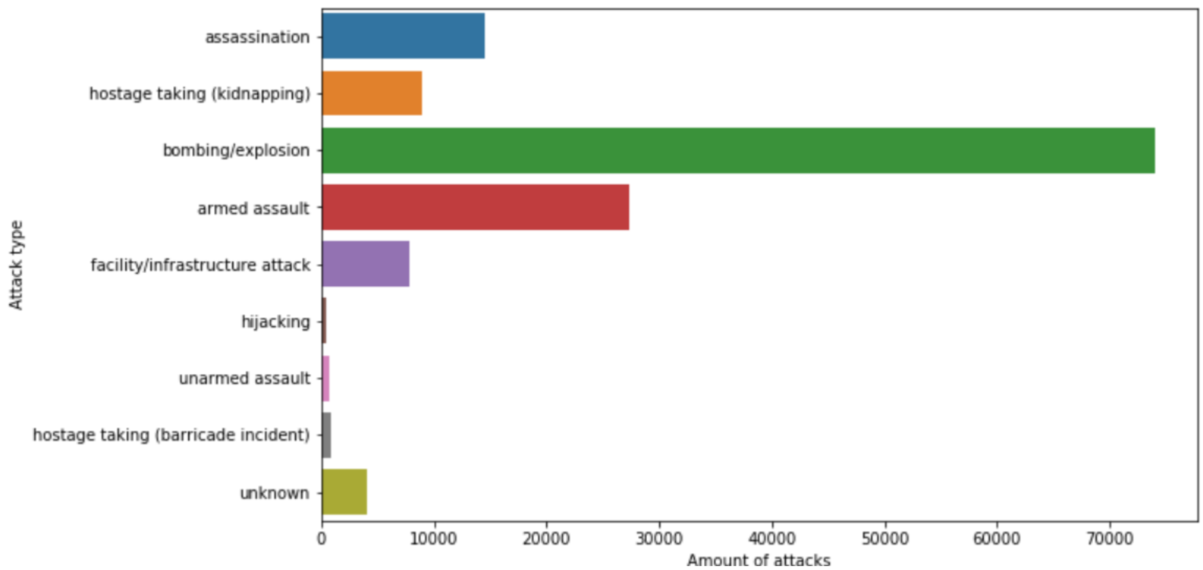


NLP analysis: Word cloud for terrorism motivation:



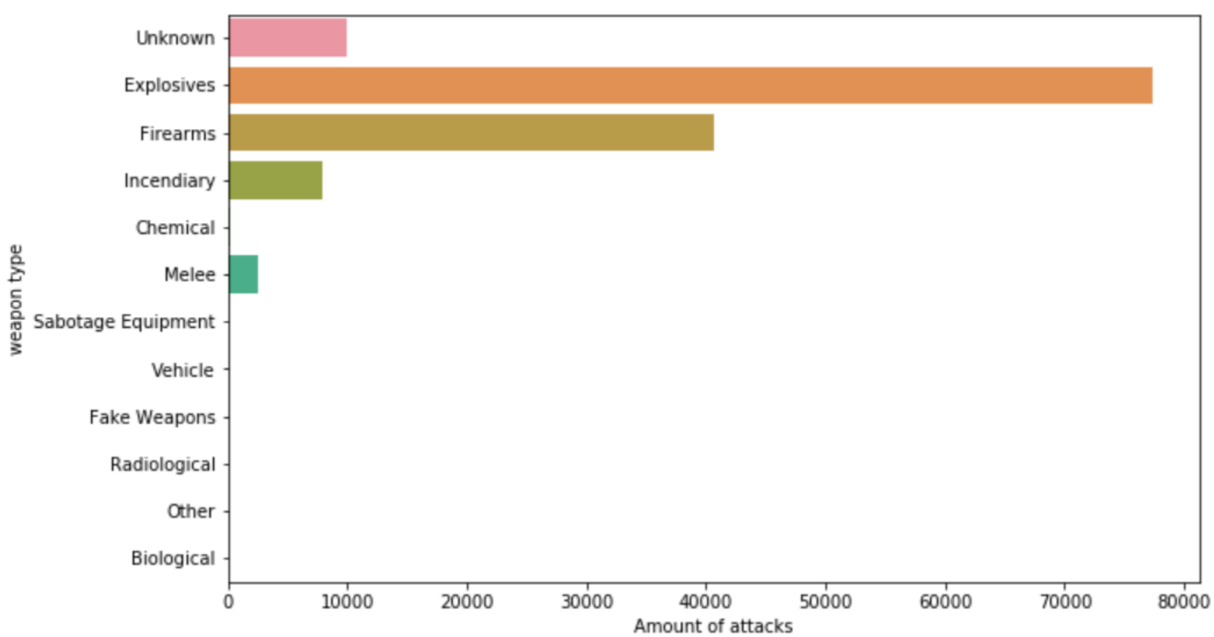
Attack type distribution:

Bombing and armed assault are the most common attack types used by terrorism groups.



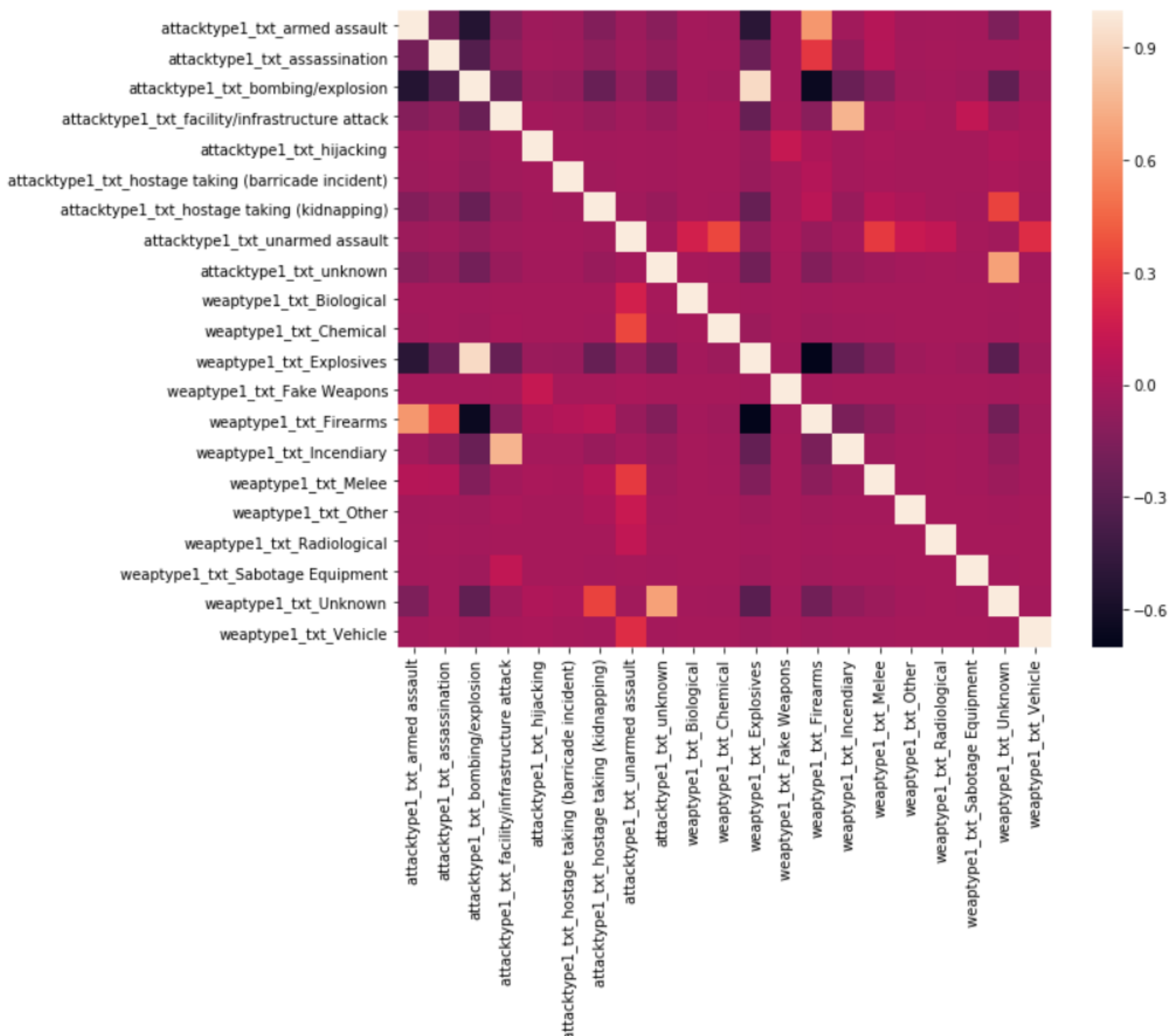
Weapon type distribution:

Explosive and firearms are the most common weapon type used by terrorism groups.

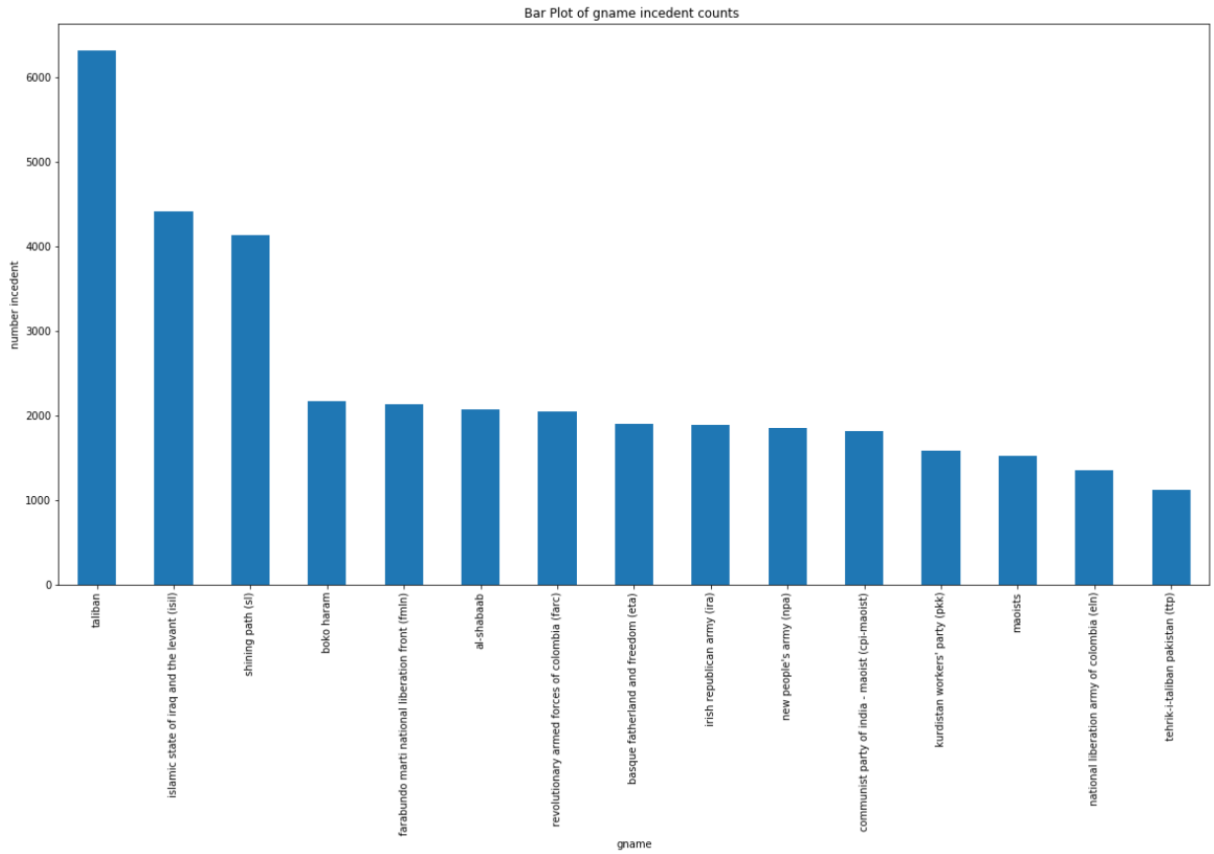


Correlation chart showing correlation between different attributes:

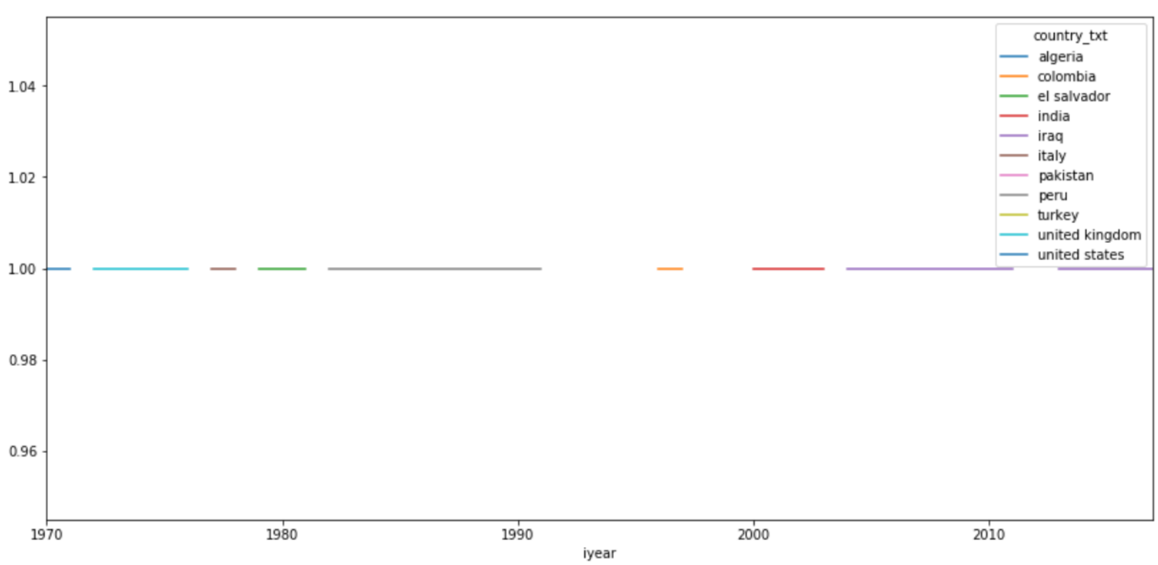
For example, assassination attack type is highly correlated with fire arms weapon type.



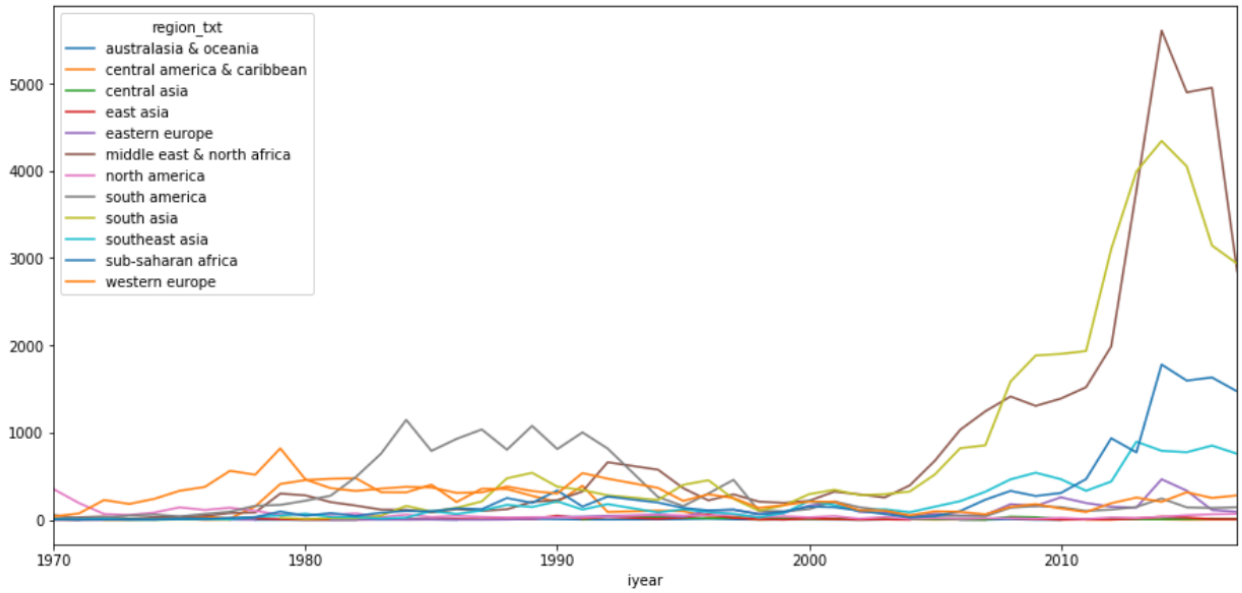
Number of attacks terrorism group distribution: Top 2 groups are Taliban, Isil.



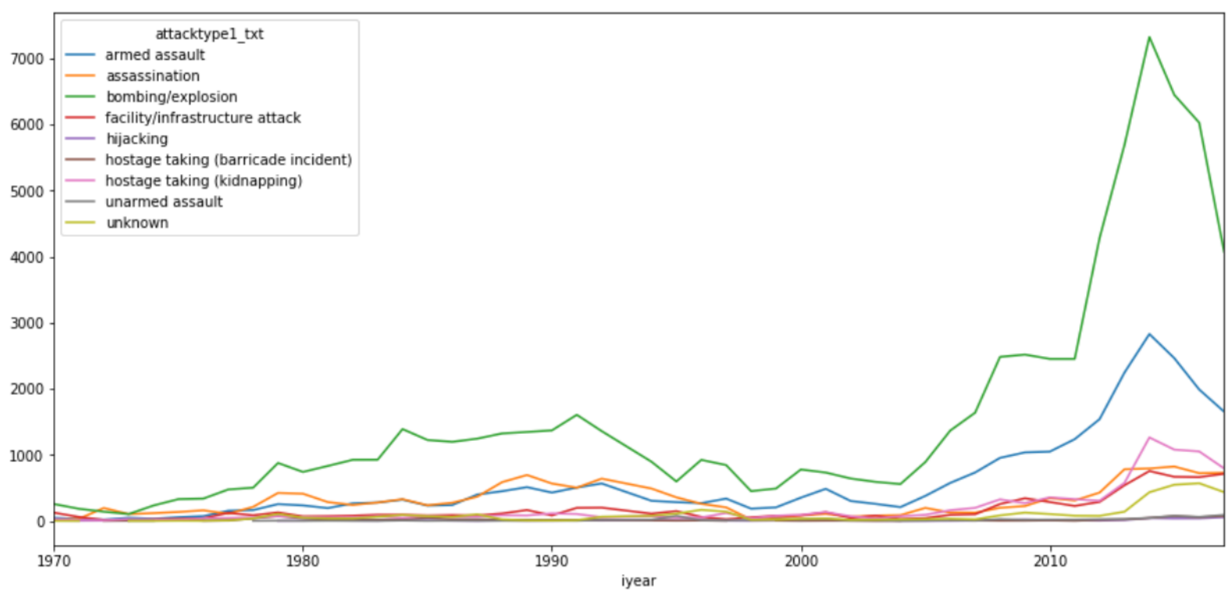
Top countries distribution over years.



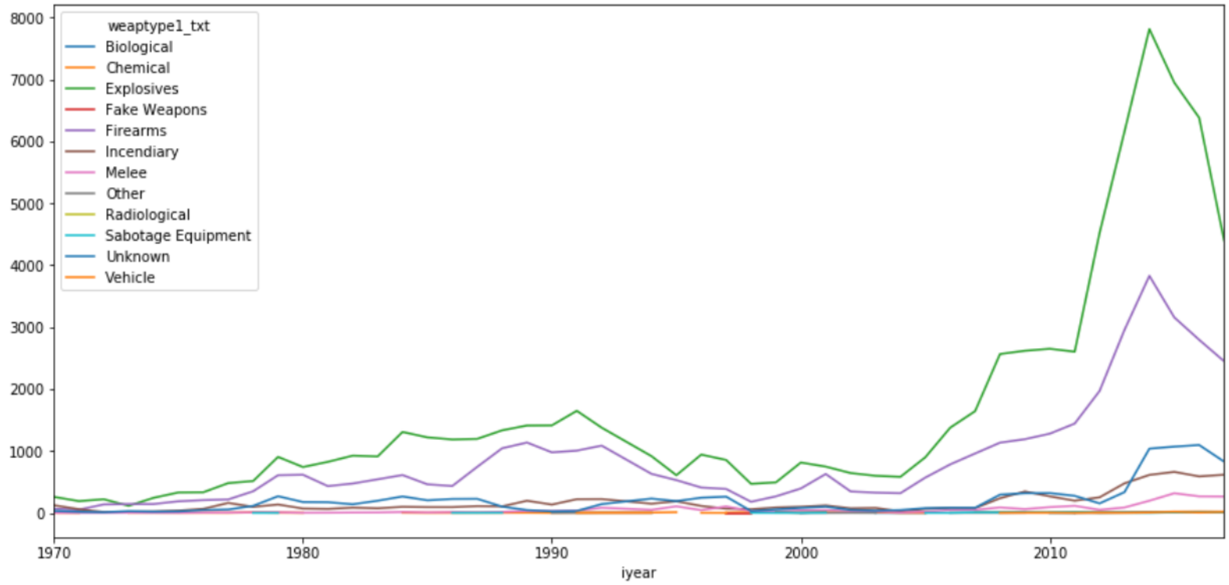
Multi-Line graph showing number of terrorism attacks per region over years:



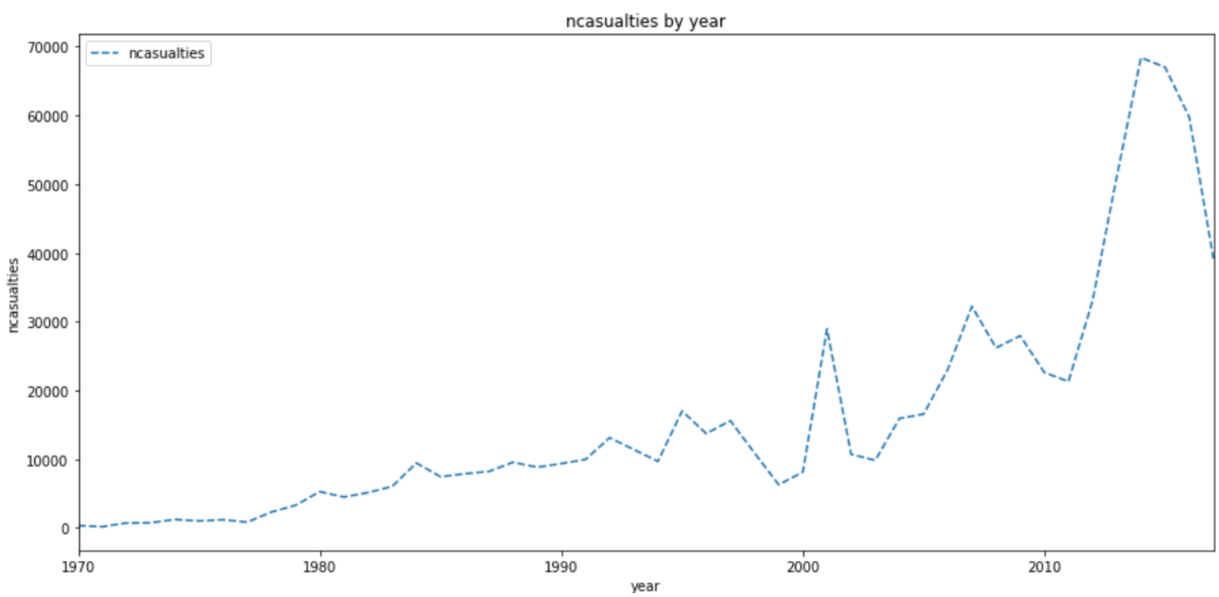
Multi-Line graph showing number of terrorism attacks per attack type over years:



Multi-Line graph showing number of terrorism attacks per weapon type over years:



Line graph showing number of casualties over years:



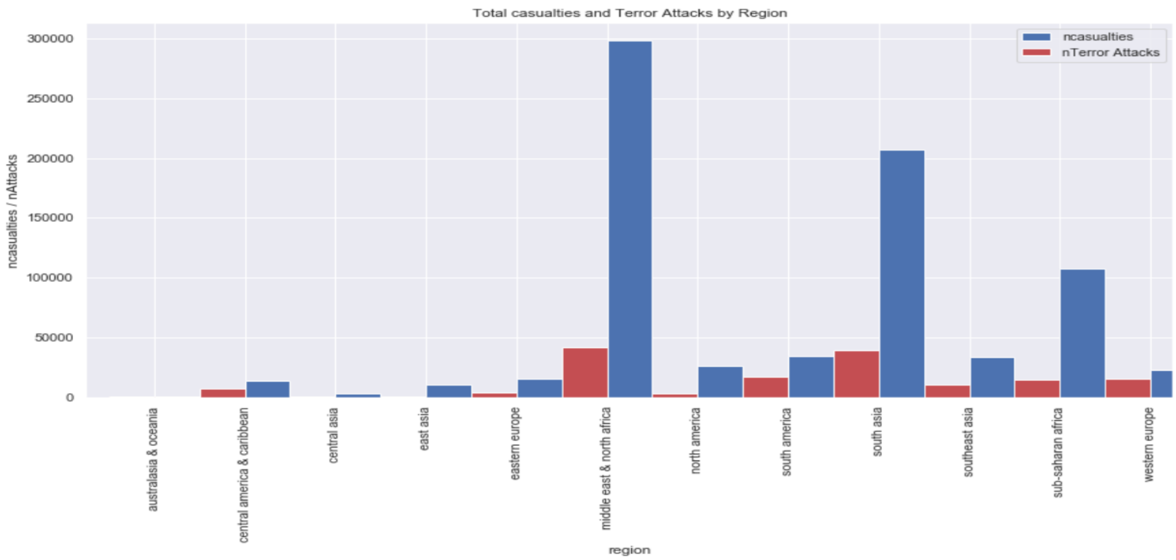
Looking at the above graph there are some interesting peaks for example in 2001 the peak due to 9/11 attack, the tables below show some statistics for the incident happened on those peaks.

iyear	imonth	iday	extended	summary	scite1	scite2	scite3	dbsource	ncasualties
2001	9	11	0	09/11/2001: this was one of four related attac...	United States Government, The 9/11 Commission ...	Lindsay Kines, □United States on high alert af...	Joe Frolick, □Hijackers Ram Two Airliners Into...	CETIS	9574

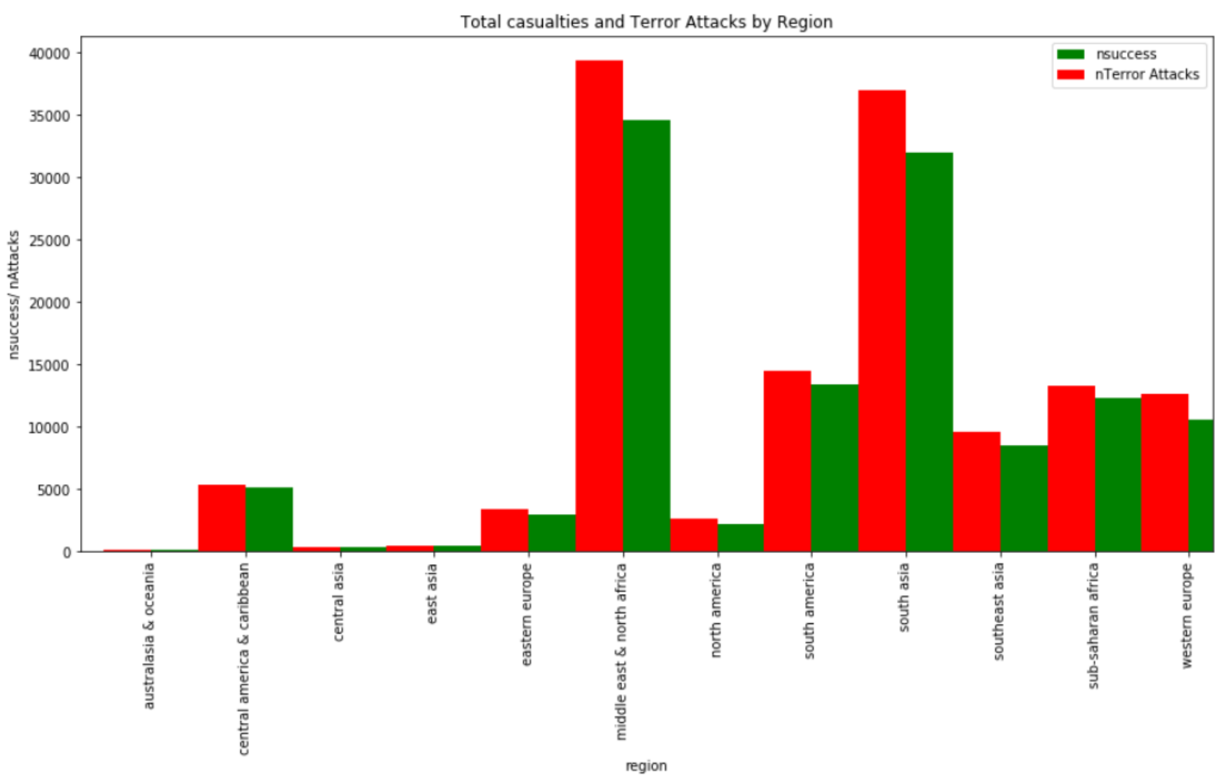
iyear	imonth	iday	extended	summary	scite1	scite2	scite3	dbsource	ncasualties
2014	8	3	1	08/03/2014: assailants attacked yizidi civilia...	"Freed From ISIS, Yazidi Women Return in 'Seve...	"Dozens of Yazidi women 'sold into marriage' b...	"WATCH: Yazidi sisters reunite after three yea...	START Primary Collection	953
2014	4	15	0	04/15/2014: assailants attacked a mosque being...	"UN confirms hundreds slaughtered in S.Sudan a...	"S.Sudan Rebels Slaughter 'Hundreds' In Ethnic...	"UN Reveals 353 Killed in South Sudan Attacks,...	START Primary Collection	687
2014	6	10	0	06/10/2014: assailants stormed badush prison i...	"Jihadists seize Iraq's Nineveh province," Age...	"Rebels seize control of Iraq's Nineveh," Al J...	"UN condemns mass executions by Islami...	START Primary Collection	670
2014	8	15	1	08/15/2014: assailants attacked yizidi civilia...	"Freed From ISIS, Yazidi Women Return in 'Seve...	"ISIS strikes Iraq village, kills Yazidis, off...	"Islamic State group kills 80 men," Himalayan ...	START Primary Collection	400
2014	11	28	0	11/28/2014: two suicide bombers and a roadside...	"Nigeria: Kano mosque blasts death toll above ...	"Nigerian mosque attack death toll climbs ...	"At Least 120 Dead In Nigeria Mosque Suicide A...	START Primary Collection	392

Distribution of number of terrorism attacks and number of casualties per region:

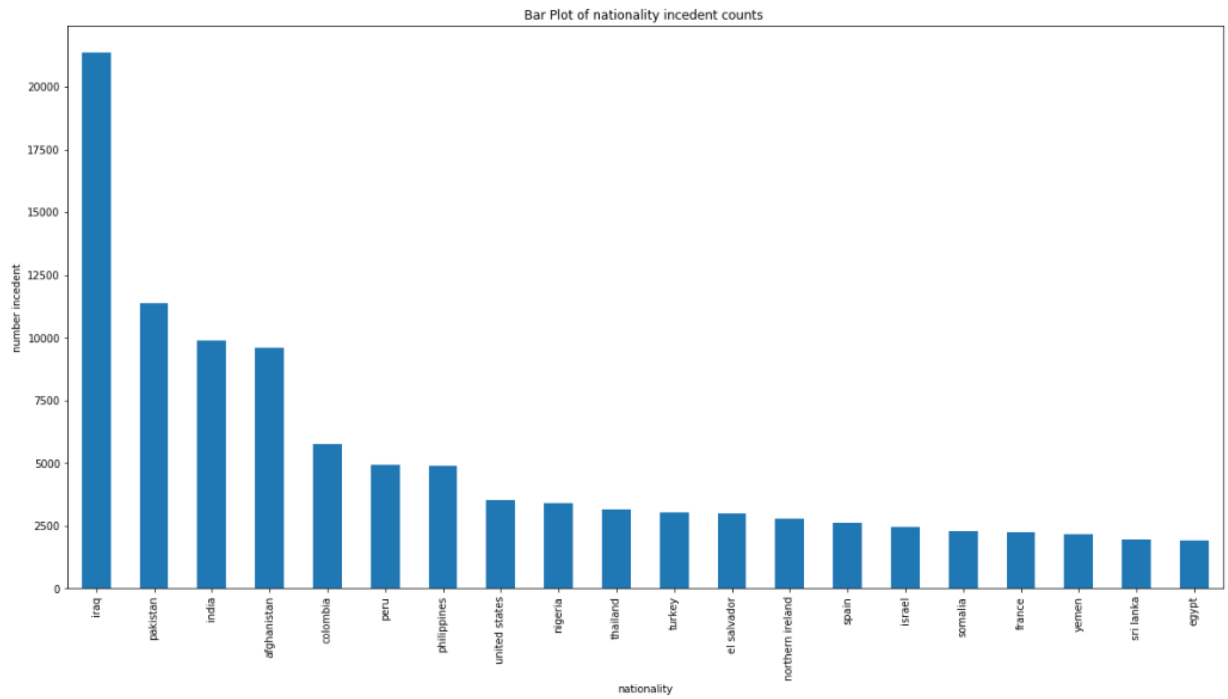
Middle east and north Africa are the top regions.



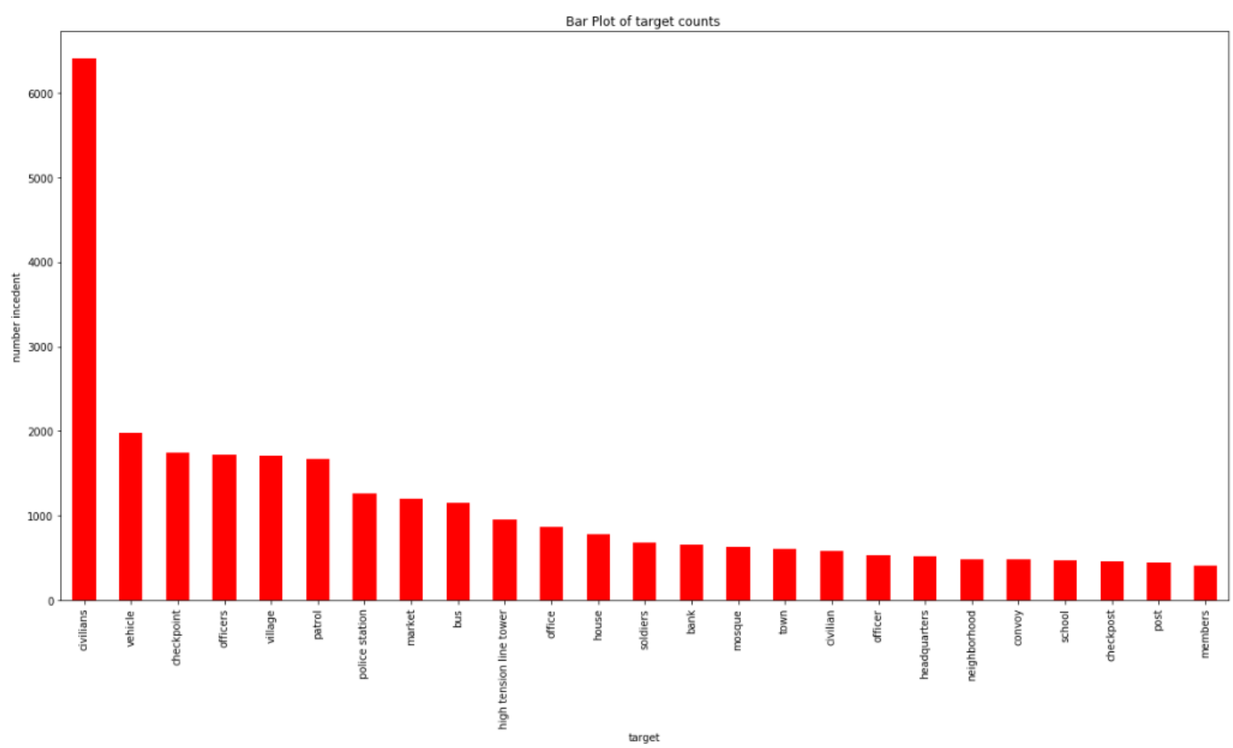
Distribution of number of terrorism attacks and number succeeded attacks per region:



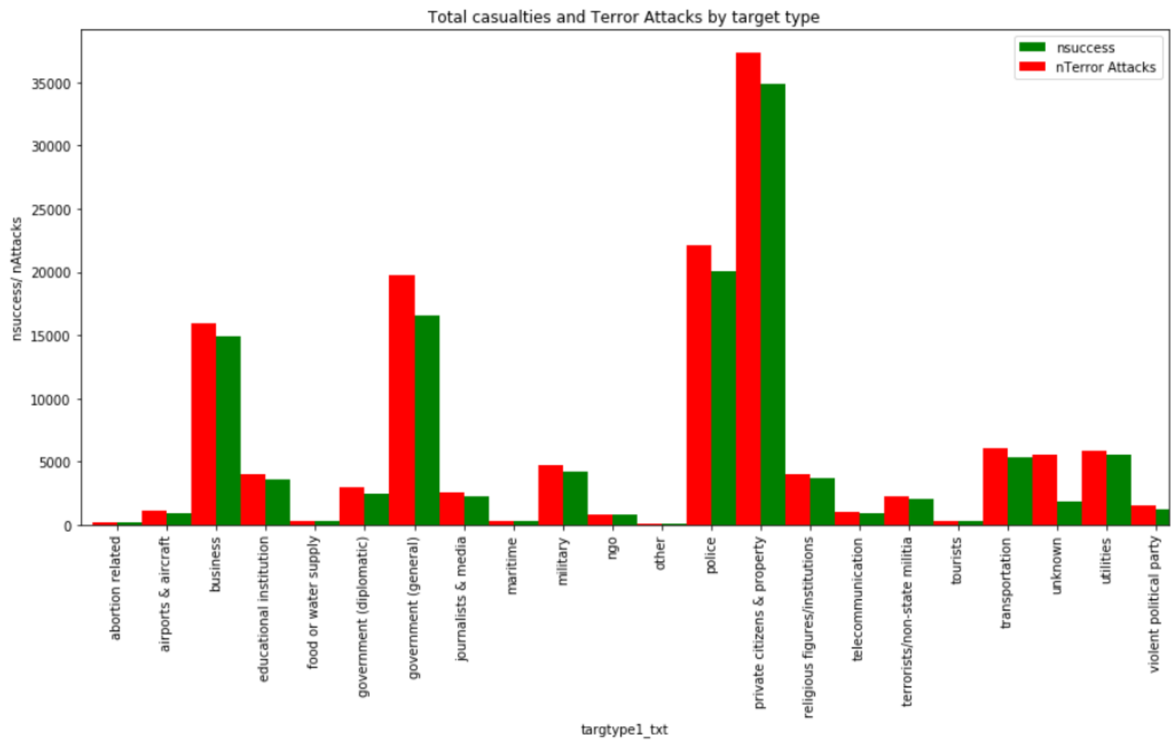
Distribution of number of terrorism attacks and target nationalities:



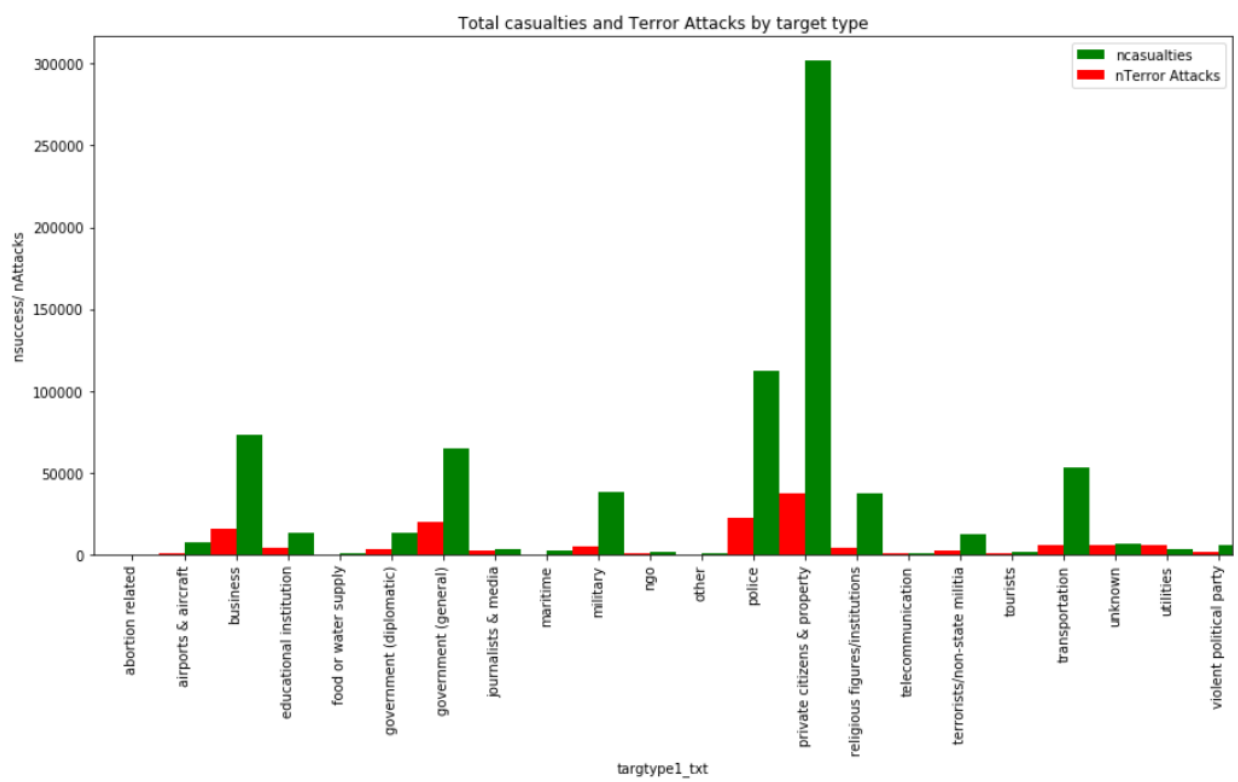
Distribution of number of terrorism attacks and targets: Civilians are the most targeted.



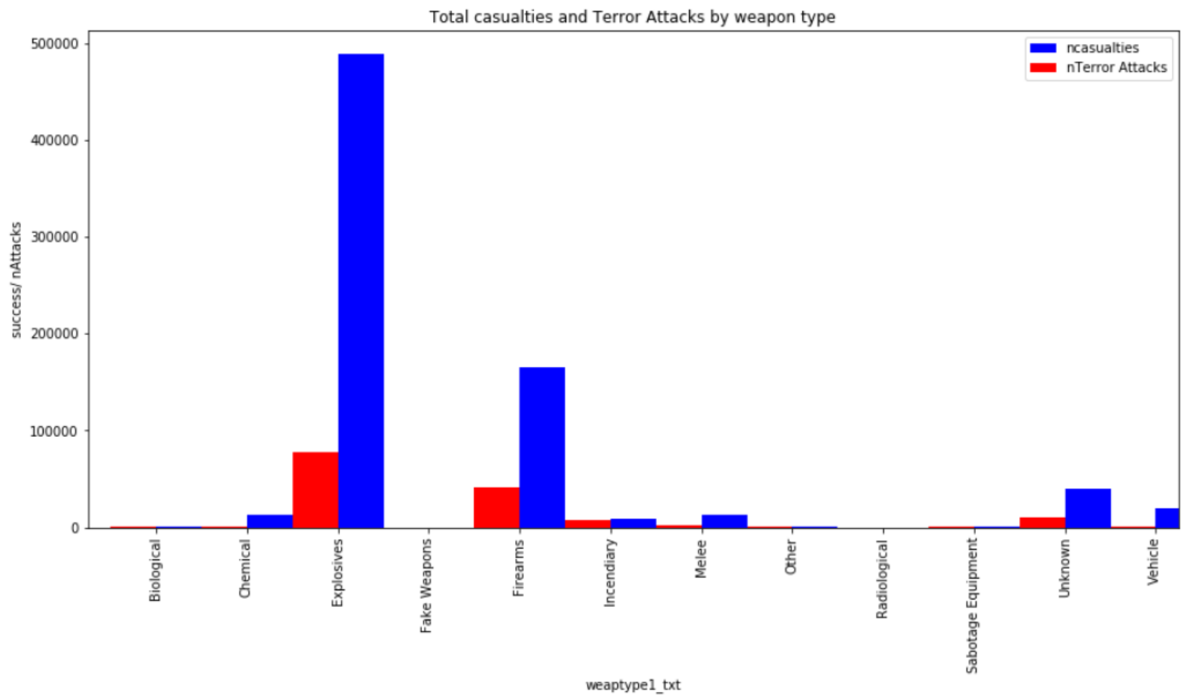
Distribution of number of terrorism attacks and number succeeded attacks per target type:

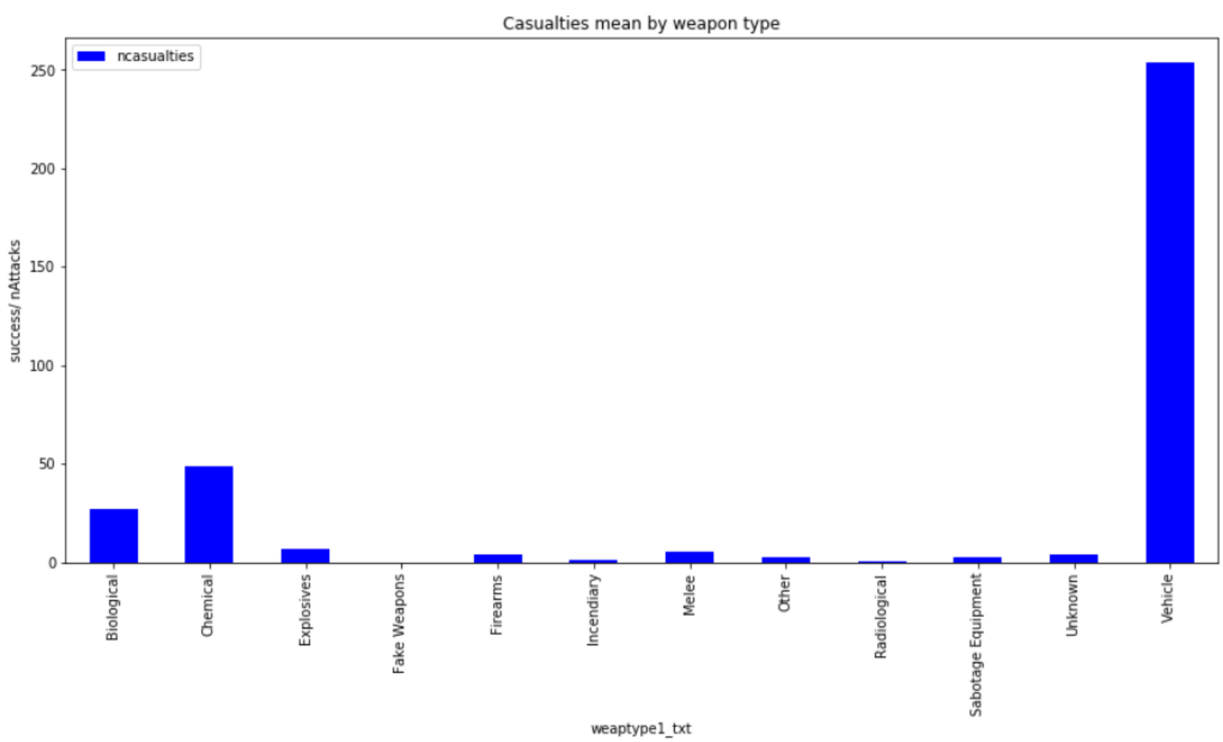


Distribution of number of terrorism attacks and number of casualties per target type:



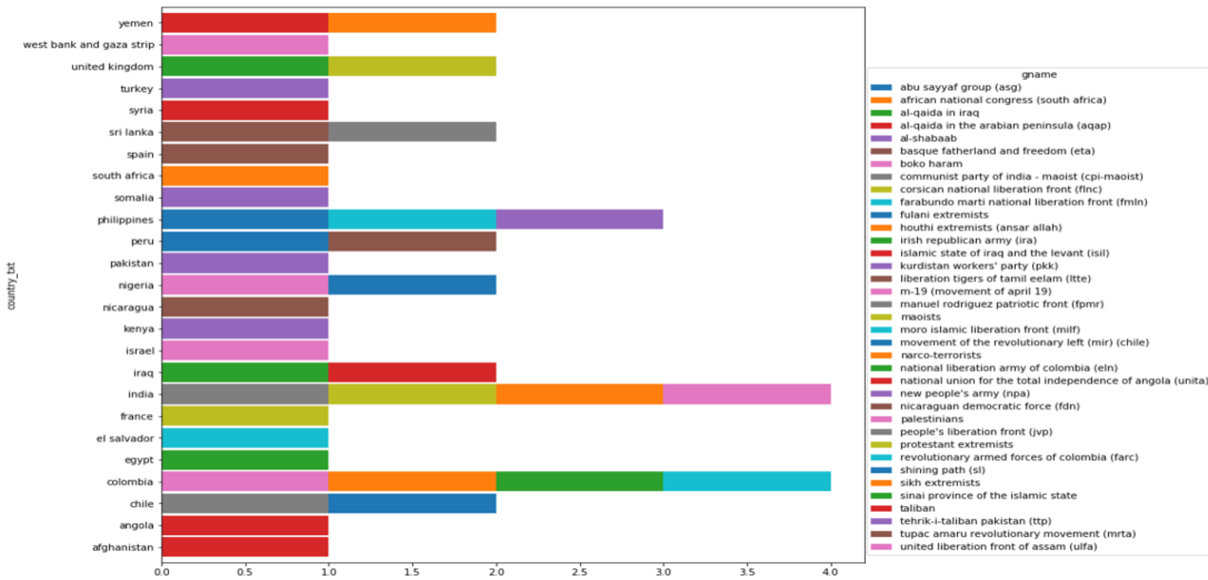
Distribution of number of terrorism attacks and number of casualties per weapon type:



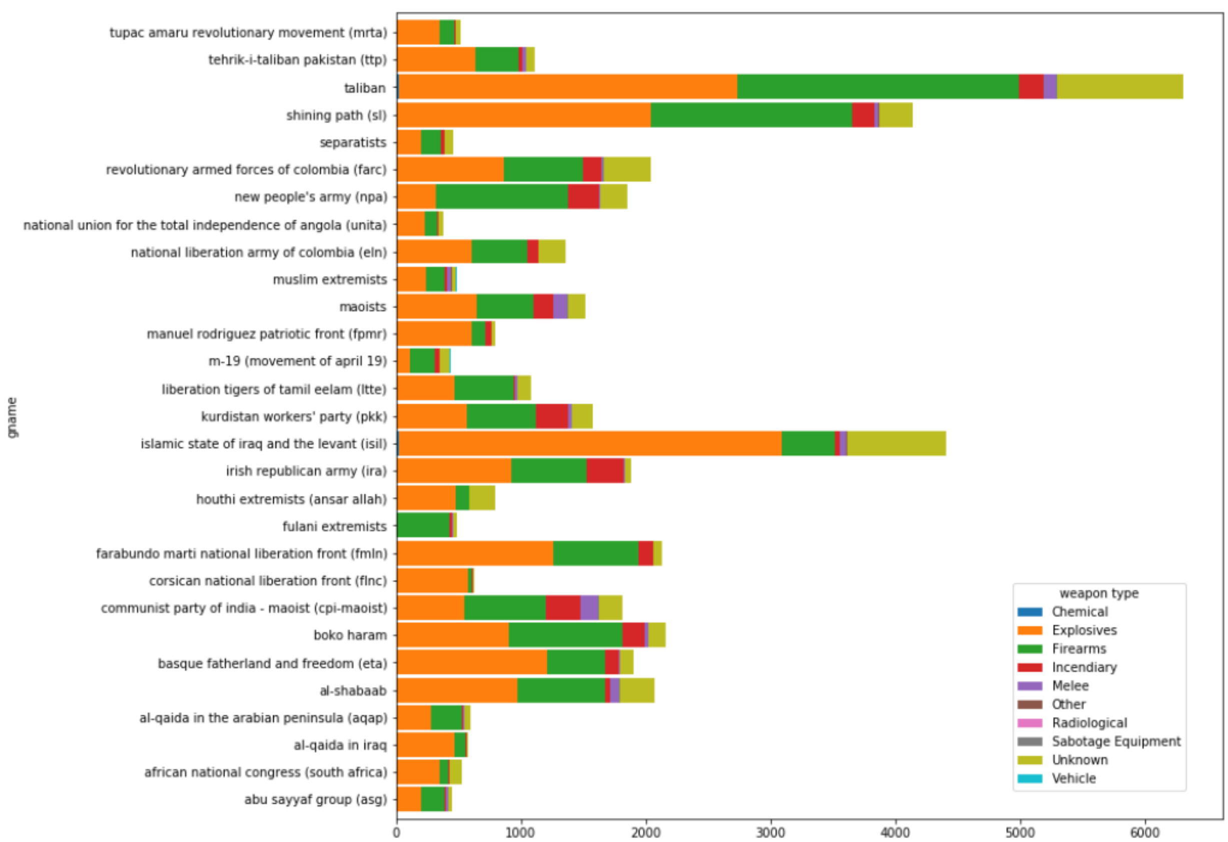


Terrorism group signature:

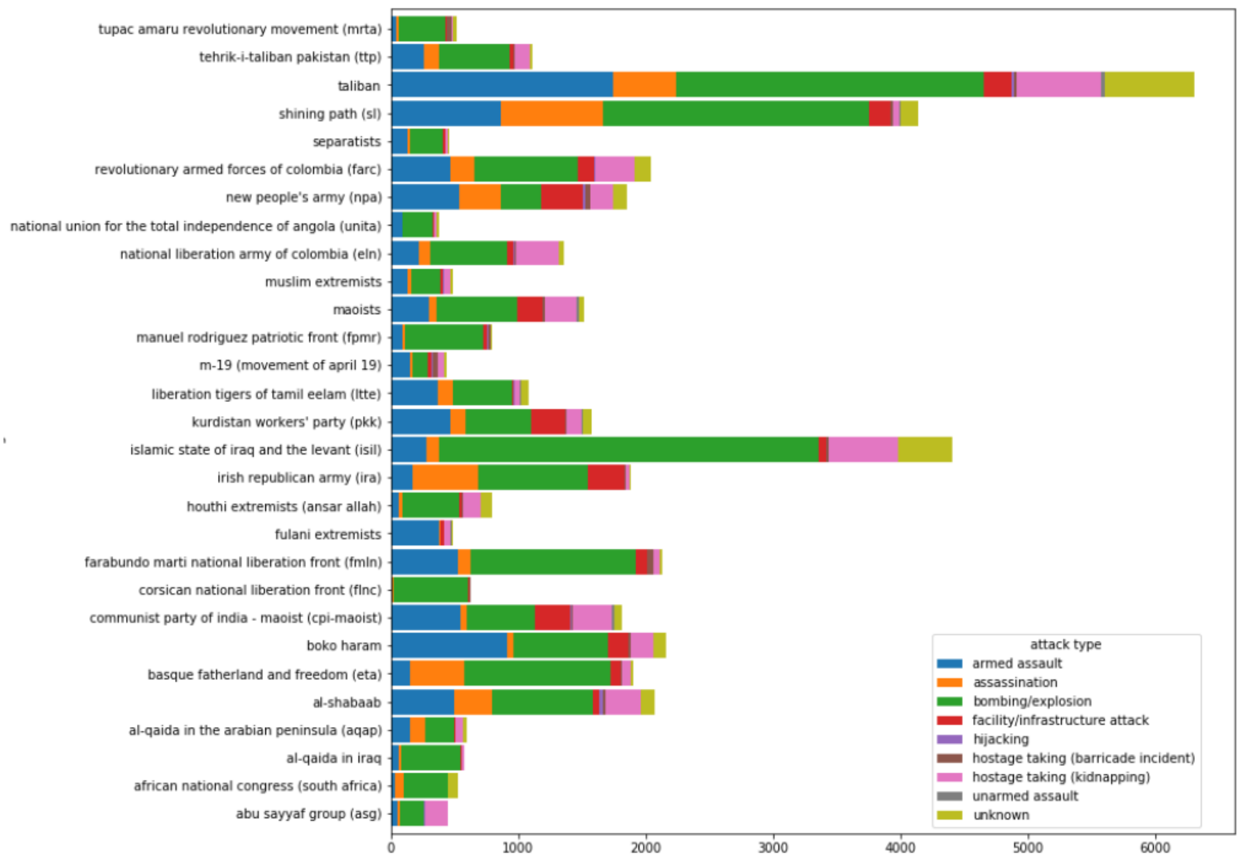
The plot shows each terrorism group where it is active in which country:



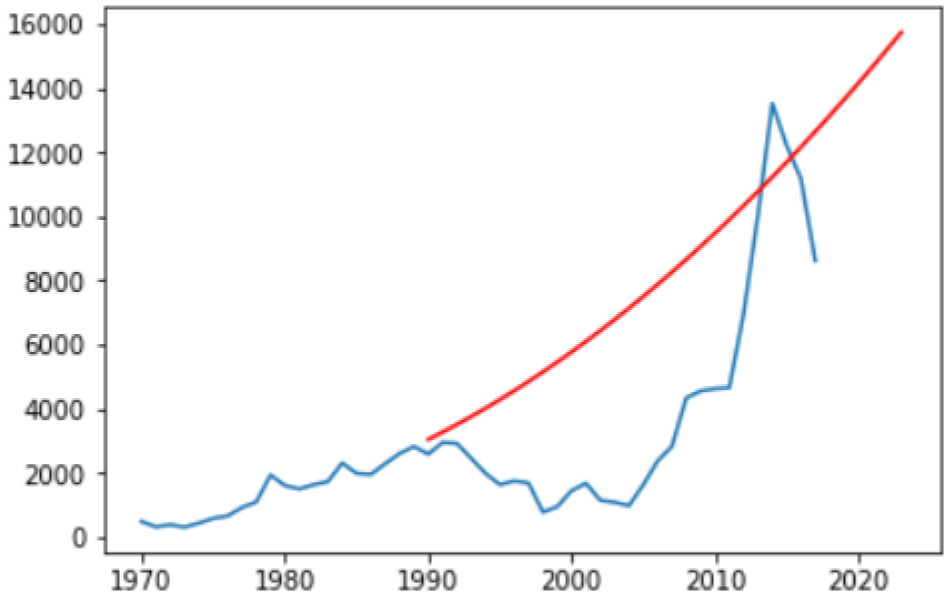
The plot shows what is the popular weapon types are used by each terrorism group:



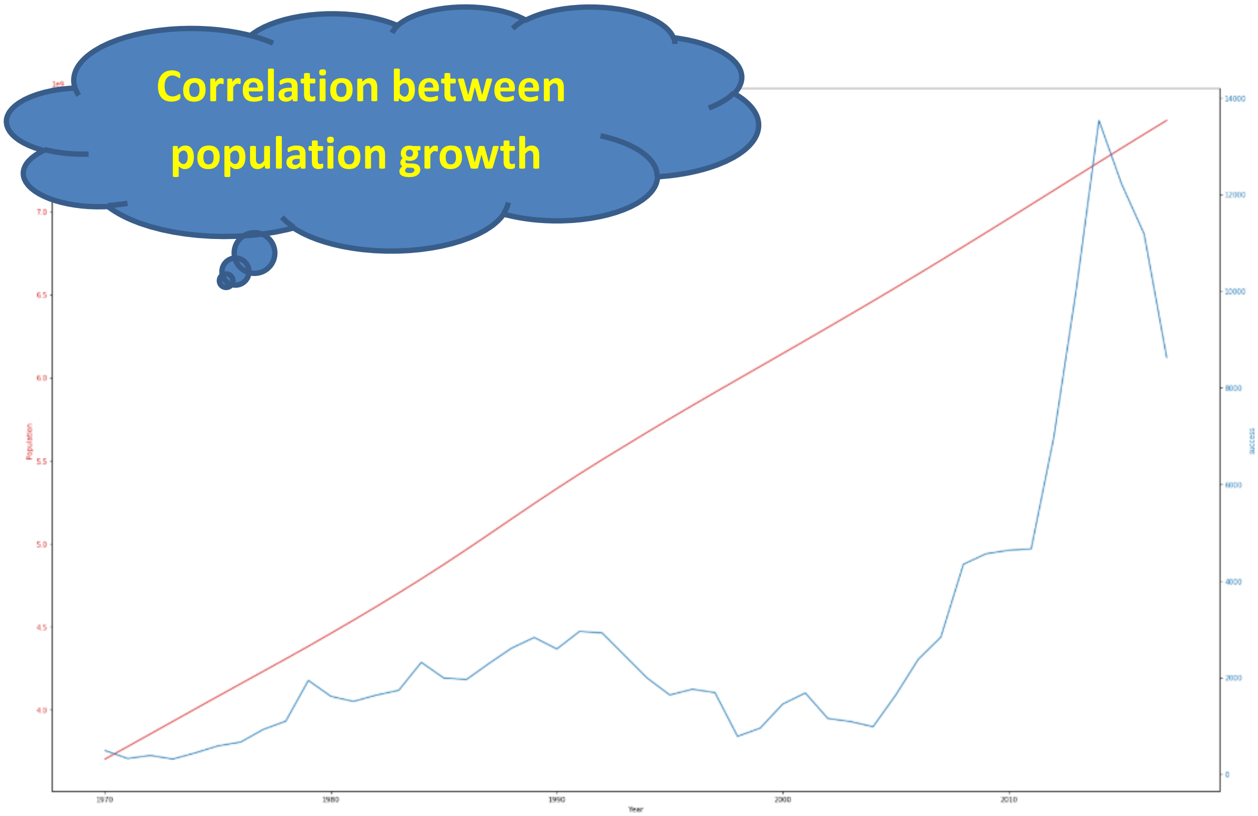
The plot shows what is the popular attack types are by each terrorism group:



Exploratory Analysis: Time Series analysis using ARIMA packages show the trend of number of terrorism attacks will rise in future.

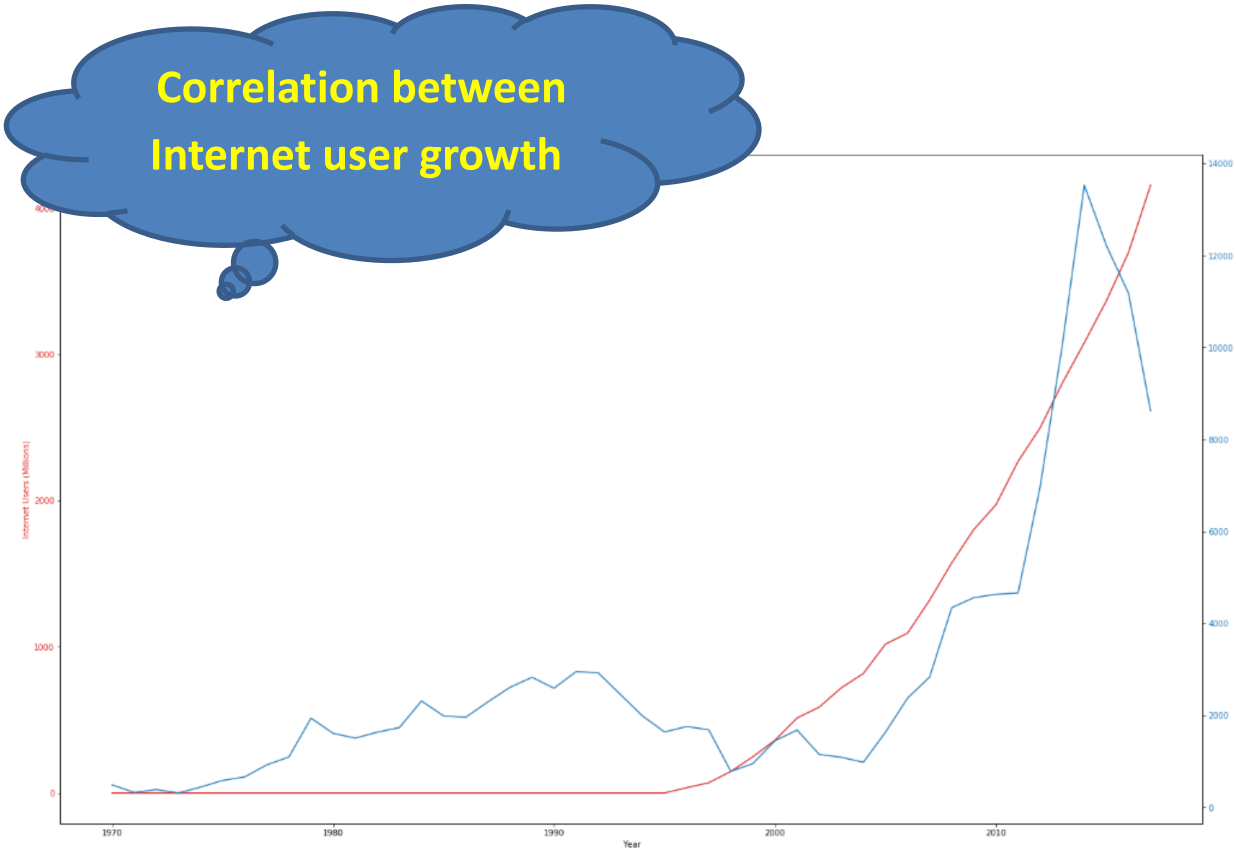


Exploratory Analysis: Correlation analysis using external datasets




```
from scipy.stats import pearsonr
corr, p = pearsonr(pop['World_Population'] , TS['success'])
corr
```

```
0.6996405274523172
```




```
from scipy.stats import pearsonr
corr, p = pearsonr(inter['users_millions'] , TS['success'])
corr
```

0.8917587261182954

Dimensionality Reduction: Remove low or No Variance, and none relevant attributes & Remove high correlated attributes.

```
mydata['crit1'].describe()
```

```
count    152622.0
mean          1.0
std           0.0
min           1.0
25%           1.0
50%           1.0
75%           1.0
max           1.0
Name: crit1, dtype: float64
```

```
mydata['crit2'].describe()
```

```
count    152622.0
mean          1.0
std           0.0
min           1.0
25%           1.0
50%           1.0
75%           1.0
max           1.0
Name: crit2, dtype: float64
```

```
mydata['crit3'].describe()
```

```
count    152622.0
mean          1.0
std           0.0
min           1.0
25%           1.0
50%           1.0
75%           1.0
max           1.0
Name: crit3, dtype: float64
```

```
# remove attributes USA spesefic which will not be used in this study
# 'nkillus' , 'nwoundus' , 'nhostkidus' , 'ransomamtus' , 'ransompaidus'
list1 = ['nkillus' , 'nwoundus' ]
```

```
# remove the additional infromation attributes which will not add value in this study
#'addnotes' , 'INT_LOG' , 'INT_IDEO' , 'INT_MISC', 'INT_ANY' , 'scite1' , 'scite2', 'scite3' ,'dbsource'
list1 = ['INT_LOG' , 'INT_IDEO' , 'INT_MISC', 'INT_ANY' , 'scite1' , 'scite2', 'dbsource']
```

```
# Save the clean dataframe
mydata.to_csv('../code/mydata_clean2.csv' , index= False)
```

```
#Here start to  Read the clean dataframe
mydata = pd.read_csv('../code/mydata_clean2.csv', encoding='ISO-8859-1')
```

```
mydata.shape
```

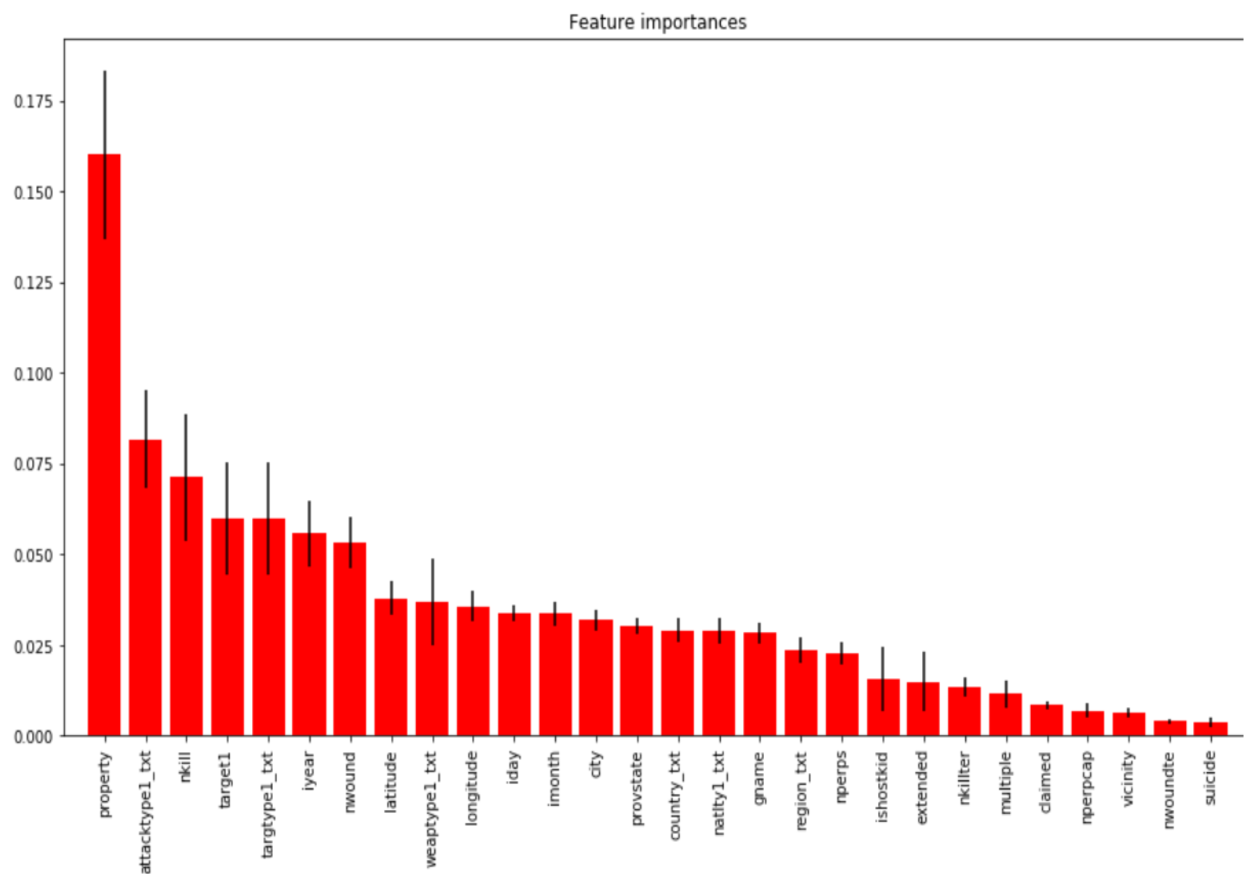
```
(152622, 30)
```

Dimensionality Reduction: feature selection / importance.

```
lb = LabelEncoder()
```

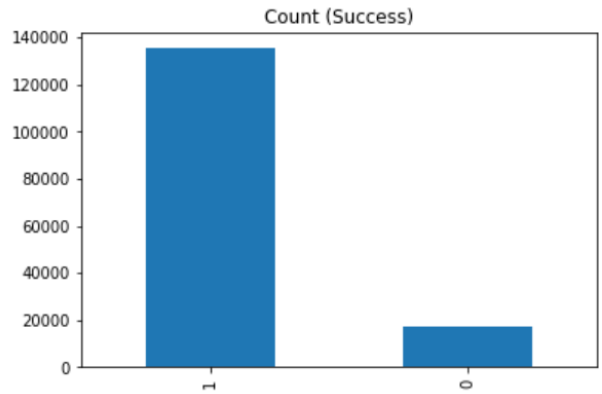
```
mydata['country_txt'] = lb.fit_transform(mydata['country_txt'])
mydata['region_txt'] = lb.fit_transform(mydata['region_txt'])
mydata['city'] = lb.fit_transform(mydata['city'])
mydata['provstate'] = lb.fit_transform(mydata['provstate'])
mydata['attacktype1_txt'] = lb.fit_transform(mydata['attacktype1_txt'])
mydata['targtype1_txt'] = lb.fit_transform(mydata['targtype1_txt'])
mydata['weaptype1_txt'] = lb.fit_transform(mydata['weaptype1_txt'])
mydata['natlty1_txt'] = lb.fit_transform(mydata['natlty1_txt'])
mydata['gname'] = lb.fit_transform(mydata['gname'])
mydata['target1'] = lb.fit_transform(mydata['target1'])
```

```
importances = forest.feature_importances_
std = np.std([tree.feature_importances_ for tree in forest.estimators_],axis=0)
```



Experimental design: Class Imbalance treatment:

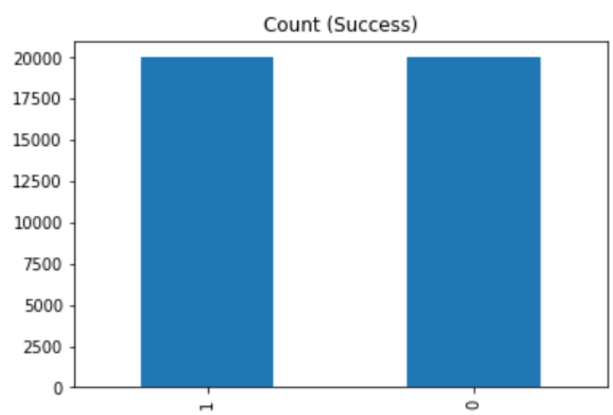
```
Class 0: 17335
Class 1: 135287
Proportion: 0.13  : 1
```




```python
# Downsample majority class
df_majority_downsampled = resample(df_majority,
                                   replace=False,
                                   n_samples=20000,
                                   random_state=123)
```

```python
# Upsample minority class
df_minority_upsampled = resample(df_minority,
                                 replace=True,
                                 n_samples=20000,
                                 random_state=123)
```

```
Class 0: 20000
Class 1: 20000
Proportion: 1.0  : 1
```



Experimental Design: Train Validation test split.

```
target_col =  target_col
```

```
X = mydata[feature_cols].fillna(0)
y = mydata[target_col]
```

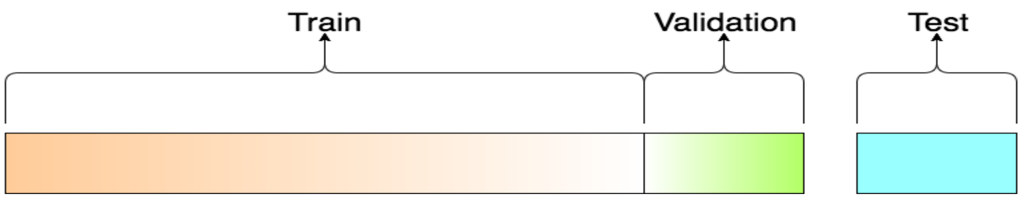
```
X.shape
```

```
(40000, 22)
```

```
y.shape
```

```
(40000,)
```

```
X_train, X_vald_test, y_train, y_vald_test = train_test_split(X, y, test_size=0.4 , shuffle = True)
X_vald, X_test, y_vald, y_test = train_test_split(X_vald_test, y_vald_test, test_size=0.5 , shuffle = True
)
```




```
X_train.shape
```

```
(24000, 22)
```

```
X_vald.shape
```

```
(8000, 22)
```

```
X_test.shape
```

```
(8000, 22)
```

```
y_train.shape
```

```
(24000,)
```

```
y_vald.shape
```

```
(8000,)
```

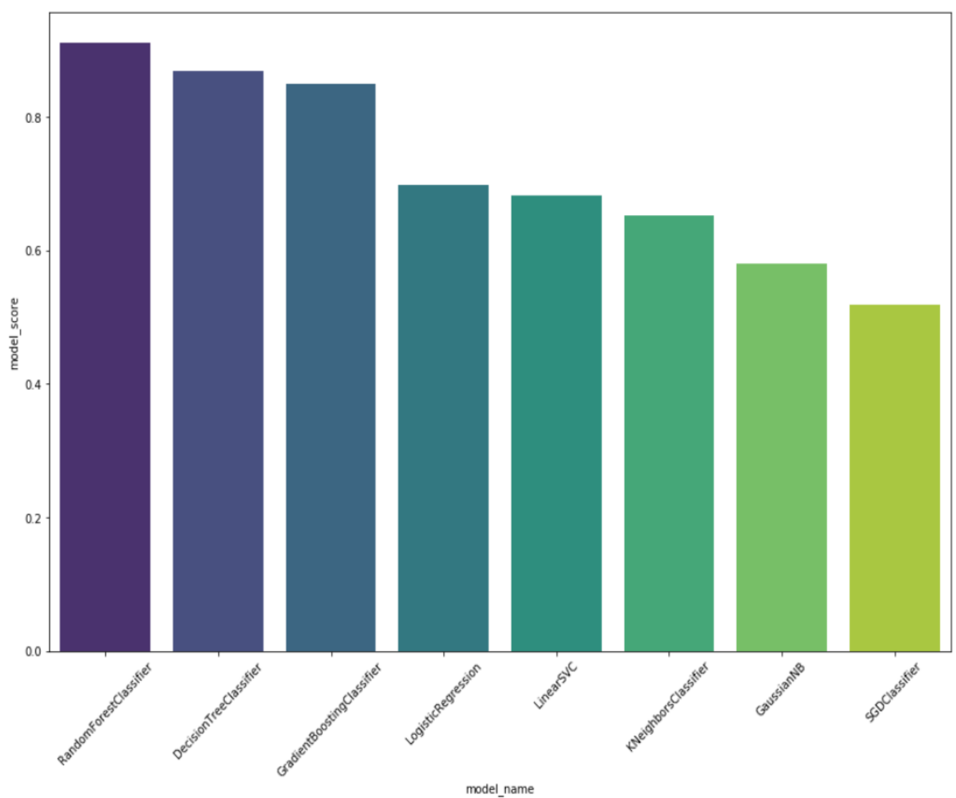
```
y_test.shape
```

```
(8000,)
```

Experimental Design: Model Selection.

10 folds cross validation used to rank the models scores:

model_name	model_score
RandomForestClassifier	0.911375
DecisionTreeClassifier	0.869791
GradientBoostingClassifier	0.849875
LogisticRegression	0.697792
LinearSVC	0.682292
KNeighborsClassifier	0.652416
GaussianNB	0.580125
SGDClassifier	0.518292



Statistical tests between top 3 models:

		comparison						
		1 data Set	2 data sets		> 2 data sets		Correlation	Regression
			unpaired	Paired	unpaired	Paired		
Large Sample Size	Normal distribution (mean)	Z-score for mean	Z-score of mean Diff	??	one-way ANOVA	two-way ANOVA	Pearson	L. regression
Small Sample size	Normal distribution (mean)		T-test	Paired T-test	one-way ANOVA	two-way ANOVA	Pearson	L. regression
	Non-Normal distribution -Rank (Median)		Wilcoxon Rank-Sum	Wilcoxon Signed rank	Kruskal-Wallis	Friedman test	Spearman	Non-Parametric
	Dichotomous (Categorical)		Chi-square test / Fisher extract test	McNeman test	Chi-square test	Q ?	Contingency Coefficient	Logistic

model_name
RandomForestClassifier
DecisionTreeClassifier
GradientBoostingClassifier

```
scores1
```

```
array([0.92958333, 0.91833333, 0.92666667, 0.92458333, 0.92541667,
       0.9275    , 0.92083333, 0.925     , 0.92458333, 0.93041667])
```

```
scores2
```

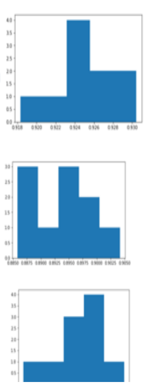
```
array([0.90416667, 0.89      , 0.89375   , 0.88625   , 0.89916667,
       0.89666667, 0.89458333, 0.88916667, 0.88583333, 0.89875   ])
```

```
scores3
```

```
array([0.86041667, 0.8425    , 0.8575    , 0.85125   , 0.85291667,
       0.86041667, 0.85833333, 0.855     , 0.86833333, 0.85958333])
```

```
# One Way ANOVA
#H0 the null hypothesis: u1=u2=u3  (u is the mean).
#Ha the alternative hypothesis is : at least one u is diffrent.
```

check normal distribution?



```
#statistic:The computed F-value of the test & pvalue:The associated p-value from the F-dist
stats.f_oneway(scores1 , scores2 , scores3)
```

F_onewayResult(statistic=366.72655969344885, pvalue=2.6849289583133914e-20)

```
#statistic:The computed F-value of the test & pvalue:The associated p-value from the F-distribu
stats.f_oneway(scores1 , scores2)
```

F_onewayResult(statistic=199.33267280568782, pvalue=3.535364163200079e-11)

```
#statistic:The computed F-value of the test & pvalue:The associated p-value from the F-distribution
stats.f_oneway(scores1 , scores3)
```

F_onewayResult(statistic=780.1830833067391, pvalue=2.8265598862131117e-16)

```
#statistic:The computed F-value of the test & pvalue:The associated p-value from the F-distribution
stats.f_oneway(scores2 , scores3)
```

F_onewayResult(statistic=166.3469926990382, pvalue=1.566720886717172e-10)

Reject the null hypothesis H0, U1 != U2 !=U3 there are statistical significant differences between the groups.

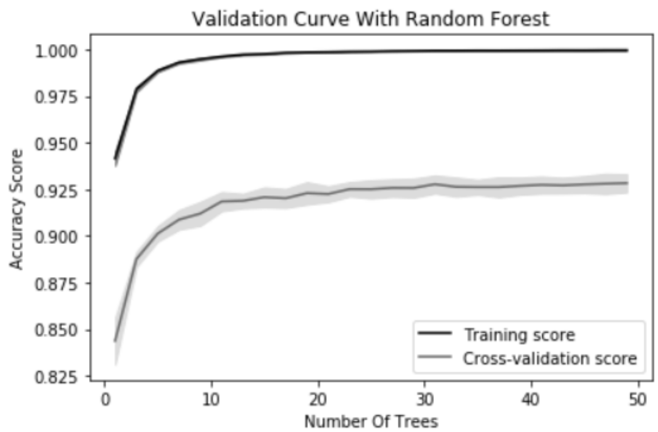
Reject the null hypothesis H0, and U1 != U2 !=U3

there are statistically significant differences between the groups.

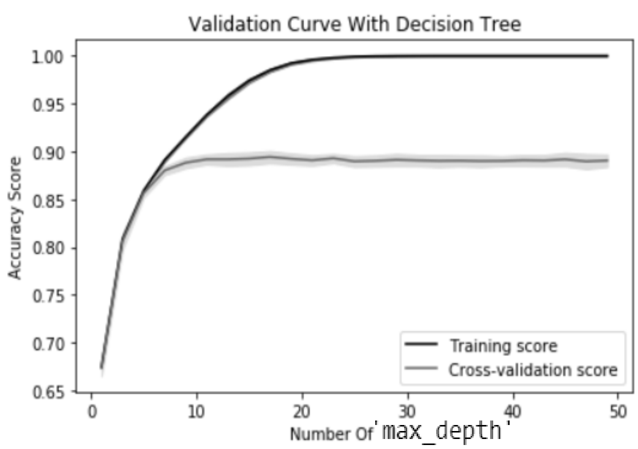
Modeling & Evaluation: Improving the Models:

Model Evaluation, calibration and hyperparameters tuning

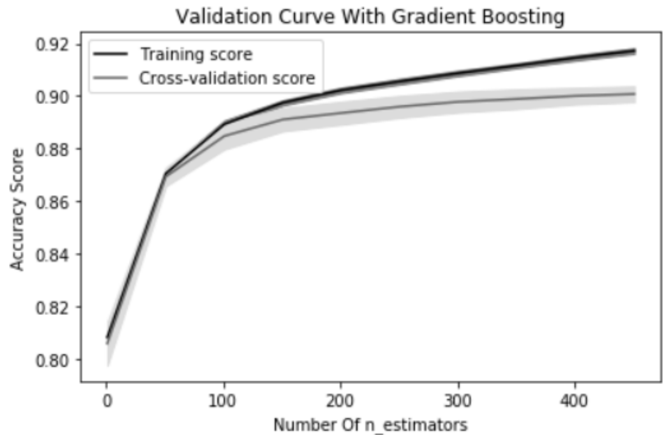
“GridSearchCV” and “validation_curve” are used to tune the model hyper paraments:




```
0.928214757894134
{'n_estimators': 27}
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
            max_depth=None, max_features='auto', max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, n_estimators=27, n_jobs=None,
            oob_score=False, random_state=None, verbose=0,
            warm_start=False)
```



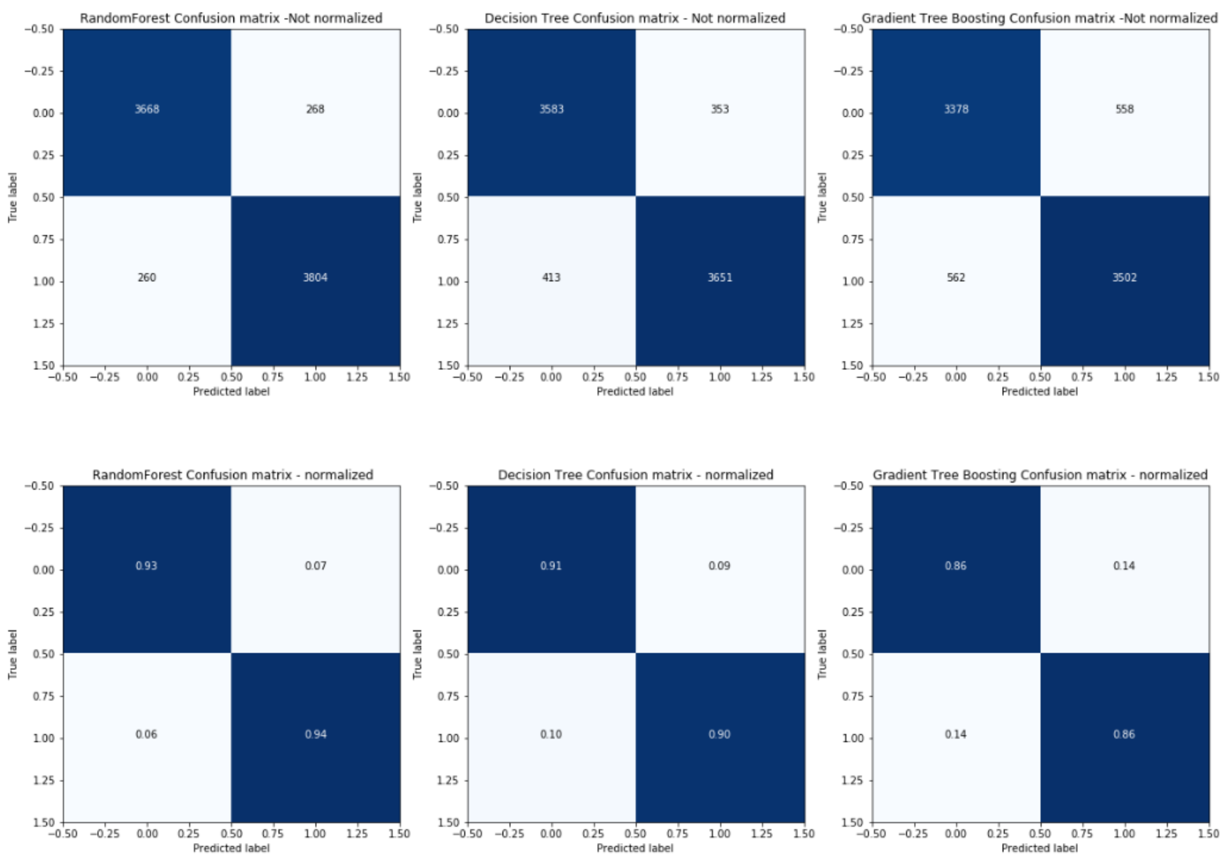

```
0.8926379379937673
{'max_depth': 11}
DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=11,
            max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, presort=False, random_state=None,
            splitter='best')
```



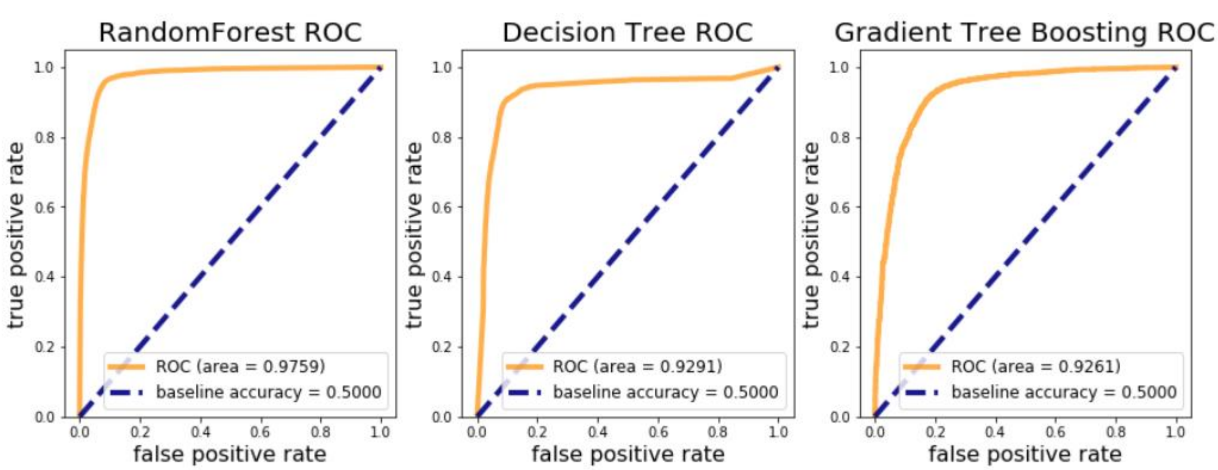

```
0.8562625997601467
{'n_estimators': 210}
GradientBoostingClassifier(criterion='friedman_mse', init=None,
              learning_rate=1.0, loss='deviance', max_depth=1,
              max_features=None, max_leaf_nodes=None,
              min_impurity_decrease=0.0, min_impurity_split=None,
              min_samples_leaf=1, min_samples_split=2,
              min_weight_fraction_leaf=0.0, n_estimators=210,
              n_iter_no_change=None, presort='auto', random_state=0,
              subsample=1.0, tol=0.0001, validation_fraction=0.1,
              verbose=0, warm_start=False)
```

Final Models Evaluation on Test set:

Confusion Matrix and table is used as evaluation tool:
Ransom forest was the best classifier with > 90% precision, recall, F1-score.
The second model was the decision tree with almost 90% precision, recall, F1-score.
And the third model was Gradient Tree boosting with about 86% precision, recall, F1-score.
Also, ROC Curve is used for evaluation:
Ransom forest the area under the curve is about 97%.
The second model was the decision tree the area under the curve is about 92%.
And the third model was Gradient Tree boosting the area under the curve is about 92%.



```
------------------- RandomForest -------------------
              precision    recall  f1-score   support

           0       0.93      0.93      0.93      3936
           1       0.93      0.94      0.94      4064

   micro avg       0.93      0.93      0.93      8000
   macro avg       0.93      0.93      0.93      8000
weighted avg       0.93      0.93      0.93      8000

------------------- Decision Tree -------------------
              precision    recall  f1-score   support

           0       0.90      0.91      0.90      3936
           1       0.91      0.90      0.91      4064

   micro avg       0.90      0.90      0.90      8000
   macro avg       0.90      0.90      0.90      8000
weighted avg       0.90      0.90      0.90      8000

--------------- Gradient Tree Boosting ---------------
              precision    recall  f1-score   support

           0       0.86      0.86      0.86      3936
           1       0.86      0.86      0.86      4064

   micro avg       0.86      0.86      0.86      8000
   macro avg       0.86      0.86      0.86      8000
weighted avg       0.86      0.86      0.86      8000
```



Conclusions

Inferences, discussion: -

- Post 2005 there are significant increase in the number of the terrorism attacks and majority of them are happening in Middle East and South Asia.

- The best classification model was random forest and the model was able to predict the success of an attack with high accuracy > 90% by knowing the attack features. This prediction can help in future to prevent/apprehended or mitigate terrorism attacks.

- The study also explored the terrorism groups active in each country and their signature (attack type, weapon type, target type) which can help to mitigate and control future attacks.

Threats to validity and Solutions to mitigate these threats:

- The analysis based on long period of time 1970 – 2017 about 47 years and the pattern, countries and terrorism group signatures keep changing, future deeper analysis will have to analysis based on each decade or based on a certain country or city. Also 1993-year data is missing.

- Source Data Legacy issues: The GTD now includes incidents of terrorism from 1970 to 2017, however several new variables were added to the database beginning with the post-1997 data collection effort. Wherever possible, values for these new variables were retroactively coded for the original incidents. This mean some fields are presently only systematically available with incidents occurring after 1997.

- Some additional important attributes were not considered in the original data like political or economical status for the country. For accurate future analysis may need to add more external datasets.

- Only one hyper-parameter per model was tuned. Future tuning for more hyper-parameters may improve the model performance.

Next / Future steps:

- Deeper analysis on one county or a city over decade periods.

- Deeper analysis on one terrorism group to understand their signature in weapon types, attack types and motives.

- More statistical tests and analysis using external dataset (mental illness, education, GDP/economic, war status, unemployment, internet growth / technology, healthcare, weather, virtual currencies such as crypto coin).

- Deeper Time series analysis to understand the future trend for the terrorism attacks.

References:

1. Rajesh Kumar Tripathi, Anand Singh Jalal and Subhash Chand Agrawal [Feb 2017]. **Suspicious human activity recognition.** Retrieved from: **https://link-springer-com.ezproxy.lib.ryerson.ca/article/10.1007%2Fs10462-017-9545-7**

2. Jerrold M; Ali, Farhana; Henderson, Schuyler W; Shanfield, Stephen; Victoroff, Jeff; [Spring 2009]. **The Psychology of Suicide Terrorism.** Retrieved from: https://search-proquest-com.ezproxy.lib.ryerson.ca/docview/220656477?pq-origsite=summon

3. Townsend, Ellen.**Suicide & Life - Threatening Behavior; Washington** Vol. 37, Iss. 1, [Feb 2007]. **Suicide Terrorists: Are They Suicidal?.** Retrieved from: https://search-proquest-com.ezproxy.lib.ryerson.ca/docview/224869508?pq-origsite=summon

4. Keene, Shima D.**Journal of Money Laundering Control; London** Vol. 14, Iss. 4, (2011): **Terrorism and the internet: a double-edged sword.** Retrieved from: https://search-proquest-com.ezproxy.lib.ryerson.ca/docview/900083145?pq-origsite=summon

5. Bussel, Breanna Renee.William James College, ProQuest Dissertations Publishing, [2017]. **PATHWAY TO JIHADIST TERRORISM: A CRITICAL LITERATURE REVIEW.** Retrieved from: https://search-proquest-com.ezproxy.lib.ryerson.ca/docview/1966517098?pq-origsite=summon

6. Ismail Onat 1 & Zakir Gul [Dec 2018]. **Terrorism Risk Forecasting by Ideology**. Retrieved from: https://link-springer-com.ezproxy.lib.ryerson.ca/article/10.1007%2Fs10610-017-9368-8

7. Nilay Saiya, Anthony Scime **[Nov 2015]. Explaining religious terrorism: A data-mined analysis.** Retrieved from: https://journals-scholarsportal-info.ezproxy.lib.ryerson.ca/details/07388942/v32i0005/487_ertada.xml

8. Todd Sandler [March 2014]. **The analytical study of terrorism: Taking stock.** Retrieved from: https://journals-scholarsportal-info.ezproxy.lib.ryerson.ca/details/00223433/v51i0002/257_tasotts.xml

9. Daniel Meierrieks & Thomas Gries [January 2013]. Causality between terrorism and economic growth. Retrieved from: https://journals-scholarsportal-info.ezproxy.lib.ryerson.ca/details/00223433/v50i0001/91_cbtaeg.xml

10. Kenneth David Strang [May 2015]. **Exploring the relationship between global terrorist ideology and attack methodology**. Retrieved from: https://www-jstor-org.ezproxy.lib.ryerson.ca/stable/43695457?pq-origsite=summon&seq=1#metadata_info_tab_contents

11. Alberto Abadie [May 2006]. **Poverty, Political Freedom, and the Roots of Terrorism.** Retrieved from: https://www-jstor-org.ezproxy.lib.ryerson.ca/stable/30034613?pq-origsite=summon&seq=1#metadata_info_tab_contents

12. Andrew Barra and Alexandra Herfroy-Mischlerb [March 2017]. ISIL's Execution Videos: Audience Segmentation and Terrorist Communication in the Digital Age. Retrieved from: https://www-tandfonline-com.ezproxy.lib.ryerson.ca/doi/abs/10.1080/1057610X.2017.1361282

13. Iwa Salami [May 2017]. **Terrorism Financing with Virtual Currencies: Can Regulatory Technology Solutions Combat This?** Retrieved from: https://www-tandfonline-com.ezproxy.lib.ryerson.ca/doi/abs/10.1080/1057610X.2017.1365464

14. Matthew Simpson [July 2014]. **Terrorism and Corruption.** Retrieved from: https://journals-scholarsportal-info.ezproxy.lib.ryerson.ca/details/00207659/v44i0002/87_tac.xml

15. Todd Sandler; Walter Enders [2007]. **Applying Analytical Methods to Study Terrorism.** Retrieved from: https://onlinelibrary-wiley-com.ezproxy.lib.ryerson.ca/doi/abs/10.1111/j.1528-3585.2007.00290

16. Samuel J. Leistedt [July 2016]. **PSYCHIATRY AND BEHAVIORAL SCIENCES.** Retrieved from: https://onlinelibrary-wiley-com.ezproxy.lib.ryerson.ca/doi/full/10.1111/1556-4029.13170