Analysis of the Global Terrorism Database

Submitted towards partial fulfillment of the criteria for the award of

PGPDSE

by **Great Lakes Institute of Management**

Submitted by:

Group 4

Batch: January 2020

Group Members:

Ankita Epari

Meenakshi K. Jha

Kallu Chandrakanth Reddy

Uppu Pavan

Mentored by:

Jatinder Bedi

CERTIFICATE OF COMPLETION

I hereby certify that the project titled "**Analysis of the Global Terrorism Database"** was undertaken and completed under my supervision by **Ankita Epari, Meenakshi K. Jha, Kallu Chandrakanth Reddy & Uppu Pavan** of Post Graduate Program in Data Science and Engineering (PGPDSE).

Jatinder Bedi

Date: 31/08/2020

Place: Hyderabad

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to our mentor **Jatinder Bedi** for providing his invaluable guidance, comments and suggestions throughout the course of this project. We value the assistance of Great Learning, Hyderabad campus. Learning from them helped us to become passionate about our research topic.

We will be failing in our duty if each one of us don't express our gratitude for other team members, for the valuable contributions during course of this project.

ABSTRACT

This research is conducted in order to determine if an event in the Global terrorism Database can be classified as exclusively terrorist or other forms of crime. Compared to most types of criminal violence, terrorism poses special challenges to a nation and exhausts all of its resources in prevention of it including the loss of life. While the human cost is devastating, the economic impact may be larger than most realize. Terrorism is one of the parameters that tourists check for, before visiting a country and hence if a Nation has more prevalent terrorism, chances are, despite its fascinating tourist attractions, it might end up in little to no Tourism. In response, there has been growing interest in researching about terrorism, their motives and most vulnerable target groups that are attacked. One thing that is infrequent is one common definition of Terrorism throughout the world. This is why the Global Terrorism Database has also included those events here which don't confirm to global inclusion criteria for terrorism but are identified as terrorist events by the locals.

Our aim in this project was to classify the events as terrorist or other forms of crime based on the Global inclusion parameters so that we can help the various intelligence agencies to drill down their study exclusive to Terrorism, avoiding any ambiguity that would come with the raw data(G.T.D).This would help them in formulation of the accurate responses to the same. The extracted features from the data are fed to the machine learning classification methods to build a model. Feature selection pre-processing steps are used to enhance the performance and scalability of the classification methods. As of now the results show that classification model has a good fit compared to other ensemble methods with the model being overfit.

Methods/Exploratory Data Analysis/Statistical analysis: When classifying the events, the inputs are as follows-Number of fatalities (terrorist and targets), No of Wounded (terrorist and targets)Terrorists captures, longitude, latitude, extended events, three inclusion criteria, Multiple incidents, country region, city, specificity, attack type, success of attack, Suicide attacks, Weapons used, Target type, Nationality of target groups, Individual or Group attack, Terrorist Group Name, Claimed attacks, Property damage, Doubtful terrorism or proper. Trends Globally have been studied over a period of 1970 to 2017(excluding 1993). Data for this study was obtained from the Global Terrorism Database.

Findings: According to the indicators mentioned above, the Terrorist activities were mainly classified on the latitude and longitude of events and the location, the 3 criteria for inclusion were also, an important metric and if the three were satisfied the event is definitely an act of terrorism, Besides this the nationality of target, weapons, type of attack, success were important too.Keywords: Global Terrorism Database, Fatalities, Property Damage, Perpetrators, Targets, Weapons, Criteria for Inclusion, Year of Attack.

Techniques:Supervised Learning Classification, Unsupervised Learning.

Tools: Python, Tableau

Domain: Crime.

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Background & need for study: Terrorist acts can cause ripple effects through the economy that have negative impacts. It leads to the Direct economic destruction of property and lives. It indirectly affects the economy by creating market uncertainty, xenophobia, loss of tourism, and increased insurance claims. In order to deal with terrorism and to cope with the effects, Countries must perform subsequent research on the factors that are contributing to it and fueling it, so that they can come up with a solution for this. The Global Terrorism Database has a brief information of attacks over the years and this study will help us in knowing the patterns on a global level. The analysis on Global Terrorism Database, will help in Identifying the groups that are responsible for the attacks and the regions that are adversely affected by the aftermath of terrorism.

Scope & Objectives: Once an analysis is done and the terrorist activities are identified there is a clear view of how much needs to be spend on security of the country and what will be the potential threat.

A clear view of the Terrorist groups, frequency of attacks, internal Marxist groups also will be identified. The findings from the research can be used by the National Consortium for the Study of Terrorism and Responses to Terrorism (START) at the University of Maryland, College Park in the United States for their study.

Approach & Methodology: After processing the dataset. Various classification algorithms are used to classify the Attacks based on set of independent variables like : Number of fatalities (terrorist an d targets), No of Wounded (terrorist and targets)Terrorists captures, longitude, latitude, extended ev ents, three inclusion criteria, Multiple incidents, country region, city, specificity, attack type, succes s of attack, Suicide attacks, Weapons used, Target type, Nationality of target groups, Individual or G roup attack, Terrorist Group Name, Claimed attacks, Property damage. Various statistical tests are u sed to understand which independent features influence the target: Doubtful terrorism or proper and removed the features that do not influence the target: Doubtful terrorism or proper. Precision, Recall and Specificity, are three major performance metrics used to evaluate the model. Base model is impl emented using the logistic regression algorithm, and looking forward to apply other classification te chniques for a better model.

INTRODUCTION

In order to do our research, we have made use of the Global Terrorism Database.

The **Global Terrorism Database** (**GTD**) is a database of incidents of terrorism from 1970 onward. As of July 2017, the list extended through 2016, with an incomplete data of 1993 due to issues with that year.

The database is maintained by the National Consortium for the Study of Terrorism and Responses to Terrorism (START) at the University of Maryland, College Park in the United States.

It is also the basis for other terrorism-related measures, such as the Global Terrorism Index (GTI) published by the Institute for Economics and Peace.

The GTD describe itself as the "most comprehensive unclassified data base on terrorist events in the world" and includes over 190,000 terrorist attacks in 2019 version. The GTD includes more than 83,000 bombings. It also includes more than 18000 assassinations and more than 11000 kidnappings.

Problem Statement: This research is conducted in order to determine if an event can be classified as exclusively terrorism or other form of crime.

Shape: The Global Terrorism database has 181691 rows and 135 Columns.

Dataset: In this project we are making use of the Global Terrorism Database has information of attacks from 1970 to 2017,few important details about out dataset is given below:

Geography: Worldwide

Time period: 1970-2017, except 1993

Unit of analysis: Attack

Variables: >100 variables on location, tactics, perpetrators, targets, and outcomes

Sources: https://www.start.umd.edu/research-projects/global-terrorism-database-gtd

➢ Project Justification

The dataset that we are dealing with in our research is the Global Terrorism Database

Compared to most types of criminal violence, terrorism poses special data collection challenges. In response, there has been growing interest in open source terrorist event data bases. One of the major problems with these data bases in the past is that they have been limited to a single definition of terrorism.

Definition of terrorism:"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."

But in certain cases, there may be some uncertainty whether an incident meets all of the criteria for inclusion into terrorism. In these ambiguous cases, where there is a strong possibility, but not certainty, that an incident represents an act of terrorism, the incident is included in GTD under a feature – "Doubt Terrorism Proper?".

This research is conducted in order to determine if an event can be classified as exclusively terrorism or other form of crime.

We have decided to choose this criterion for inclusiveness of events in the dataset because it was recognized at the outset that researchers and public officials ascribe to varying definitions of terrorism.

Therefore, the approach that was adopted was to collect and structure data such that it would be useful to as broad an audience as possible. The method chosen to achieve this was to err on the side of inclusiveness in our criteria, but to include in the database filtering mechanisms through which users can truncate the data set according to the definition of terrorism that meets their needs.

In such scenarios, the user can filter the data according to specific components of established definitions of terrorism.

The findings from our research can be used by the Center for terrorism and Intelligence, (United States), to research and exclusively work on events pertaining to terrorism and find repeating patterns across the globe.

It will also help in classifying the event into other types of crime if it is not exclusively terrorist.

Problem Statement: to determine if an event in the G.T.D can be classified as exclusively terrorism or other forms of crime.

Target variable: doubtterr (Categorical Variable)

1 = "Yes" There is doubt as to whether the incident is an act of terrorism.

0 = "No" There is essentially no doubt as to whether the incident is an act of terrorism.

Project Outcome:

Academic Value:

- To help the terrorism research department at the John Hopkins University with our contribution towards GTD analysis & visualizations.
- This work can be used by curious civilians, security related policy-makers, international organizations hosting worldwide events, foreign investors and academic researchers for the purpose of understanding terrorism and its nature.

Social Value:

- To help the investigation process regards to a crime activity by firstly classifying the form of crime occurred.
- To help the National Consortium for the Study of Terrorism and Responses to Terrorism which hosts the GTD dataset, with the inclusion of only those exclusively terrorist activities into the dataset.

Complexity Involved:

- The size of the dataset is relatively large, thereby leading to high processing time during our project work
- The terminologies involved are domain specific. So, familiarizing ourselves with the meaning of the feature columns & general awareness was time consuming.
- Since the data involves a lot of pre-processing & Data cleaning, we couldn't directly start with Tableau.
- The data in itself is unsupervised in nature and hence we had to frame the problem statement by studying the data.

DATASET and DATACLEANING

➢ Data Dictionary

Our dataset has 135 columns and a few important columns are mentioned below:

- eventid- this is the Unique ID of the event. (Numeric Variable)
- iyear-year of the incident (Numeric Variable)
- imonth-month of the incident (Numeric Variable)
- iday – Day of the incident (Numeric Variable)
- extended- Extended Incident (Categorical Variable)
- summary -Incident Summary (Text Variable)
- crit1, crit2, crit3-Inclusion Criteria (Categorical Variables)
- multiple - part of multiple incident (Categorical Variable)
- related - related incidents (Text Variable)
- country, country_txt- country (Categorical Variable)
- region, region_txt- region (Categorical Variable)
- city - city (Text Variable)
- latitude – latitude (Numeric Variable)
- longitude - longitude (Numeric Variable)
- specificity -Geo-Coding Specificity (Categorical Variable)
- attacktype1, attacktype1_txt – Attack Type (Categorical Variable)
- success -Successful Attack (Categorical Variable
- suicide -Suicide Attack (Categorical Variable)
- weaptype1, weaptype1_txt -Weapon Type (Categorical Variable)
- targtype1, targtype1_txt -Victim Type (Categorical Variable)
- corp1: Name of the Entity (Text Variable)
- natlty1; natlty1_txt -Nationality of the Victim (Categorical Variable)
- individual -Unaffiliated Individuals (Categorical Variable)
- gname -Perpetrator Group Name (Text Variable)

- nperps- Number of the Perpetrators (Numeric Variable)
- claimed -Claim of Responsibility (Categorical Variable)
- motive - Motive (Text Variable)
- Nkill - Total Fatalities - (Numeric Variable)
- nkillter - perpetrator Fatalities (Numeric Variable)
- nwound -Total Injured (Numeric Variable)
- nwoundte - Perpetrators Injured (Numeric Variable)
- property - property damage (Categorical Variable)
- propvalue- Value of the Property Damage (in usd) (Numeric Variable)
- nhostkid- Total No. Of Hostages / Victim Kidnaps (- Numeric Variable)
- ransomamt -Total Ransom Amount Demanded (Numeric Variable)
- ransompaid -Total Ransom Amount Paid (Numeric Variable)
- INT_IDEO – Attack of an International Ideology (Categorical Variable)

Target Variable:

doubtterr - Doubt Terrorism Proper (Categorical Variable)

Information: Multiclass classification (0: Terrorism,1: doubtful, -9: unknown)

➢ Variable Categorization

The Global Terrorism database has 181691 rows and 135 Columns.

Counts of the numerical and Categorical variables in the Global Terrorism Database are listed below:

```
terror1.dtypes.value_counts()
```

```
object     58
float64    55
int64      22
dtype: int64
```

In the dataset there are 58 Object Columns or Categorical columns,55 float columns and 22 integer columns.

➢ Missing or Null Values:

Treatment of Missing Values: In any real-world data set, there are usually few null values. It doesn't really matter whether it is regression, classification or any other kind of problem no model can handle these NULL or NaN values on its own so we need to intervene. First of all we need to check whether we have null values in our dataset or not. We can do that using the isnull () method. There are various ways for us to handle this problem. The easiest way to solve this problem is by dropping the rows or columns that contain null values.

Null Value imputation is risky and might lead to misleading results and interpretations: For example: the column 'nkill' has 10313 null values. If we impute these with mean, median or k nearest neighbors imputer it will depict that at every instance of an attack, there were fatalities. While in reality there were no fatalities reported in that particular event.

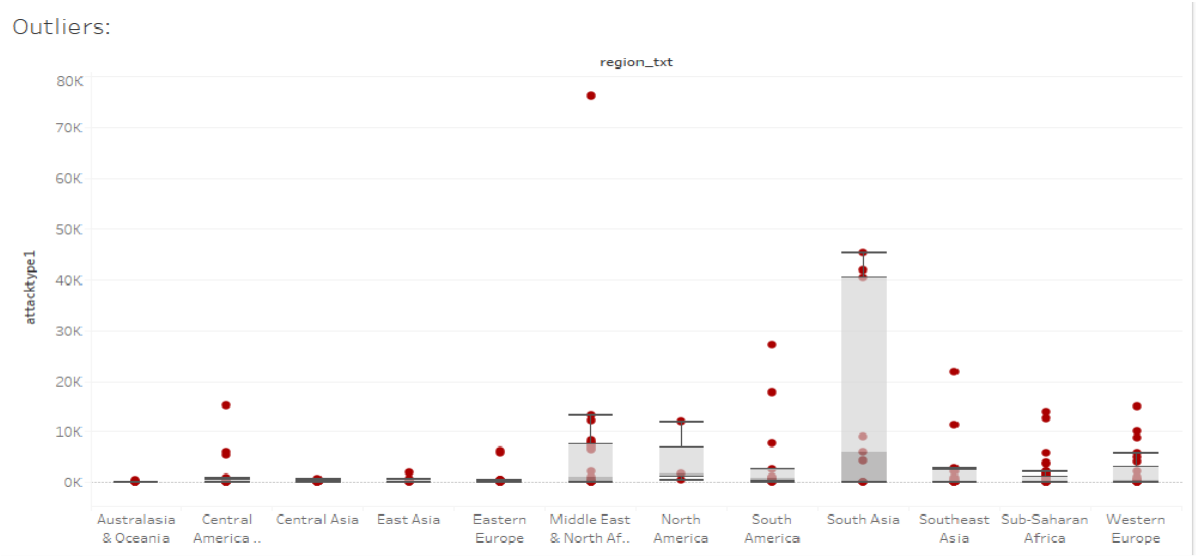
In our case its best to drop the values that are null, but since we are not sure about that impact it would create, we have built models; both by imputation of the above missing values as well as after dropping the missing values.The columns with null values are listed below:

1. city 434
2. latitude 4556
3. longitude 4557
4. multiple 1
5. corp1 42550
6. nkill 10313
7. nkillter 66958
8. nwound 16311
9. nwoundte 69143
10. summary 66129
11. target1 636
12. targsubtype1 10373
13. natlty1 1559
14. nperps 71115
15. nperpcap 69489
16. motive 131130
17. related 156653
18. propvalue 142702
19. propextent 117626

20. the nhostkid 168119
21. ransompaid 180917
22. ransomamt 180341
23. doubtterr 1

Initial dataset consists of 167906 and 38 columns, after removing the null values the size is 91686 ro ws and 38 columns. After removing the redundant text columns the dataset size is 91686 rows and 2 7 columns.

➢ Presence of Outliers



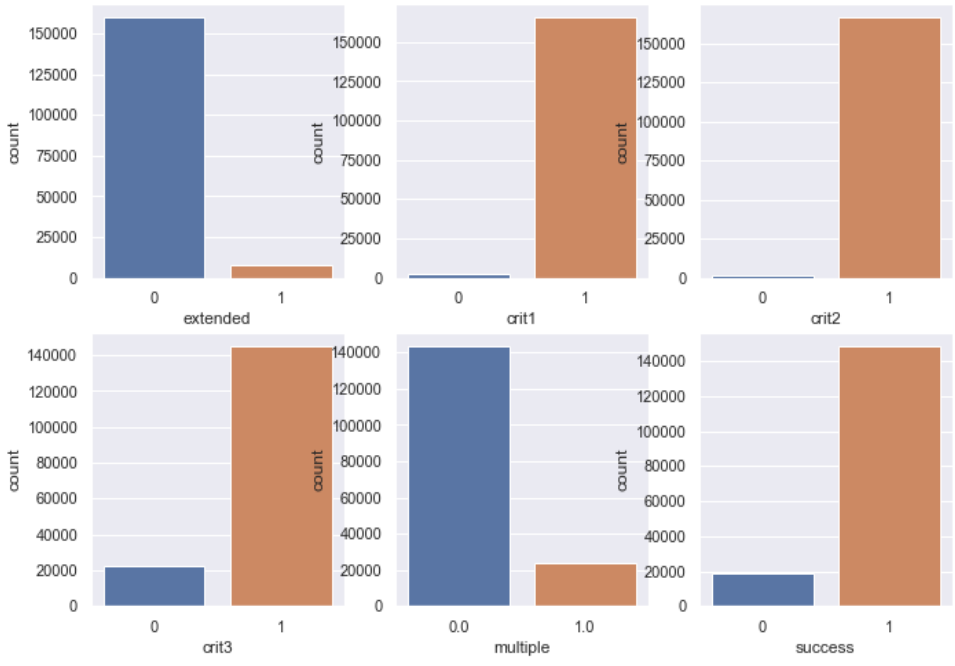
In the Global terrorism database, we do have data points that differ significantly from other observations. However, this alone is not enough to call them bad data as this is important information and dropping or treating them might lead to loss of important information.

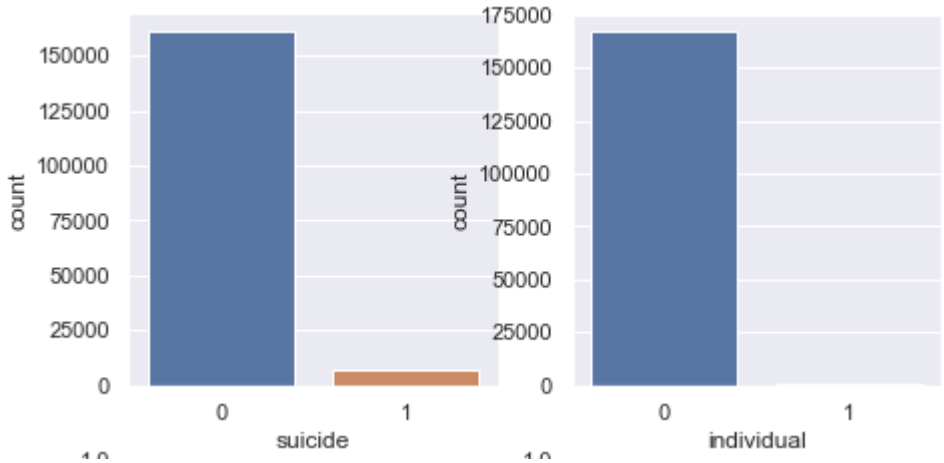
The above is a boxplot representation of outliers in number of attacks globally in the top 15 Regions. The outliers here are the Regions that were adversely affected by the attacks. Treating this data or dropping it might hamper our end result and hence in our domain treatment of outliers is not required.

EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) is an open-ended process where we make plots and calculate statistics in order to explore our data. The purpose is to find anomalies, patterns, trends, or relationships. These may be interesting by themselves (for example finding a correlation between two variables) or they can be used to inform modeling decisions such as which features to use. In short, the goal of EDA is to determine what our data can tell us! EDA generally starts out with a high-level overview, and then narrows in to specific parts of the dataset once as we find interesting areas to examine.

- Univariate analysis of Features(Categorical):Categorical variables in the dataset: Extended, Criteria1,2,3, Multiple, success, suicide, Individual, Doubtterr (Target).



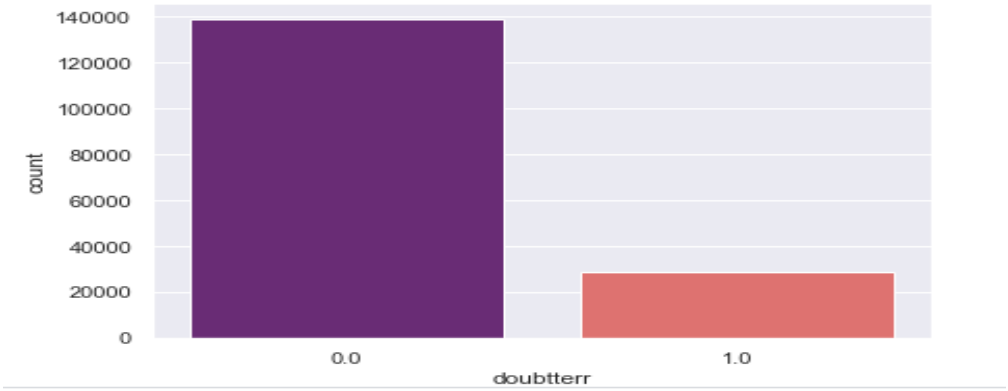


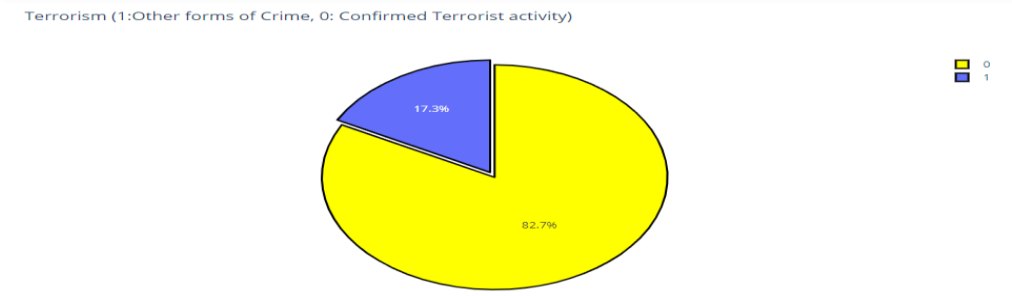
- ## Univariate Analysis(Dependent Feature):

0: Confirm terrorist activity

1:Doubtful/Other forms of Crime

```
0.0    82.727836
1.0    17.272164
Name: doubtterr, dtype: float64
```

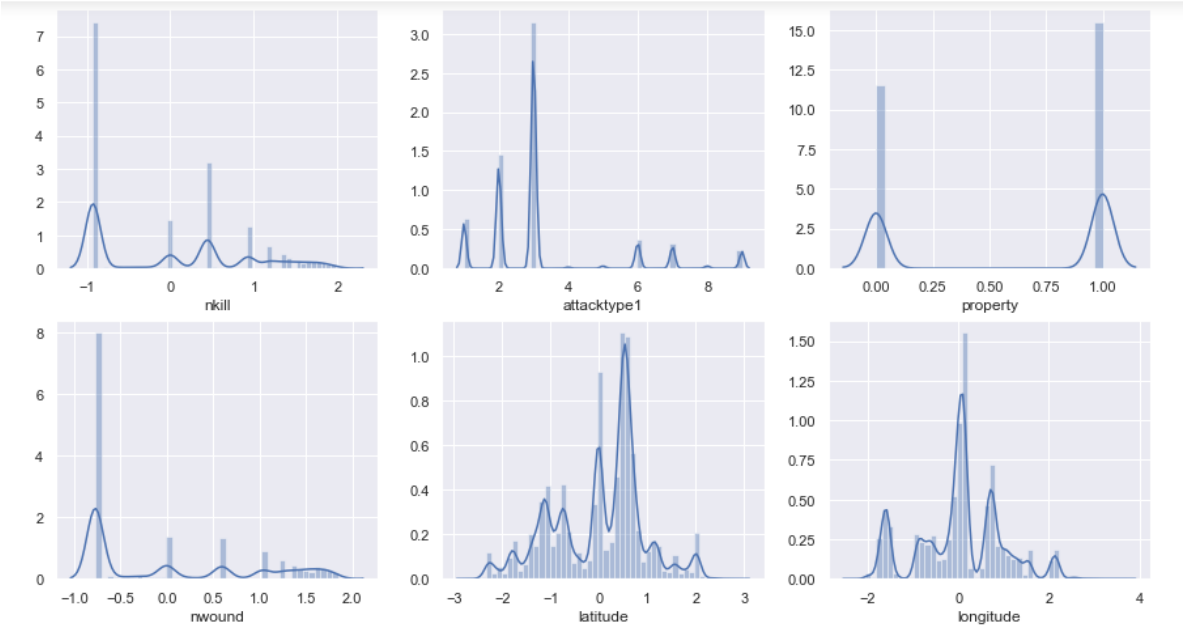




From this we are Clear that the terrorist events are more in number and there is ambiguity associated few events.

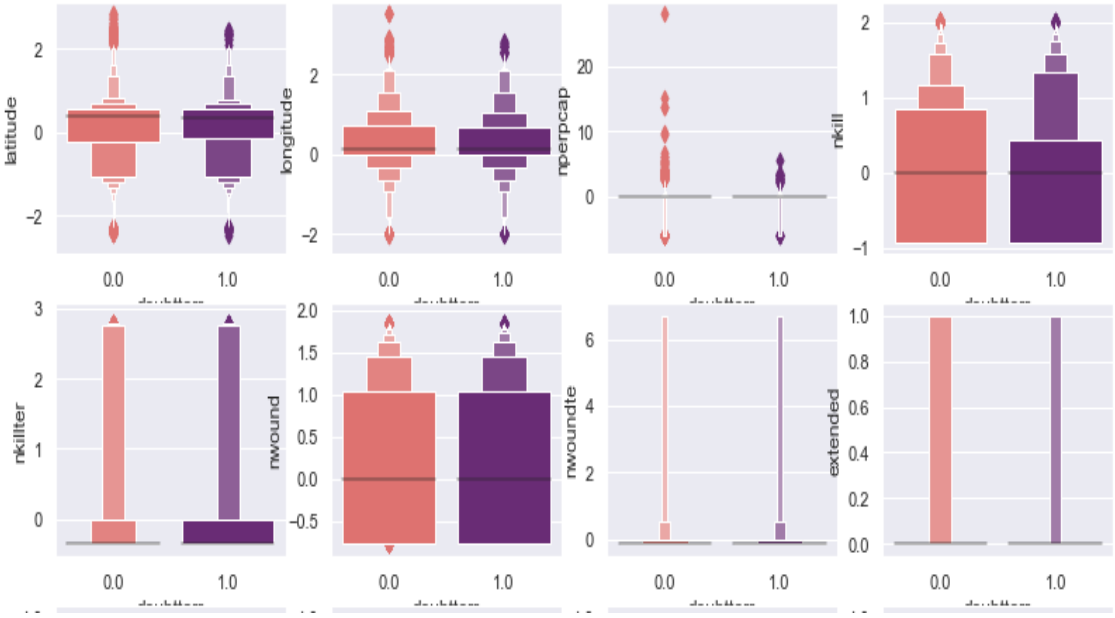
➢ Univariate analysis of Features (Continuous):

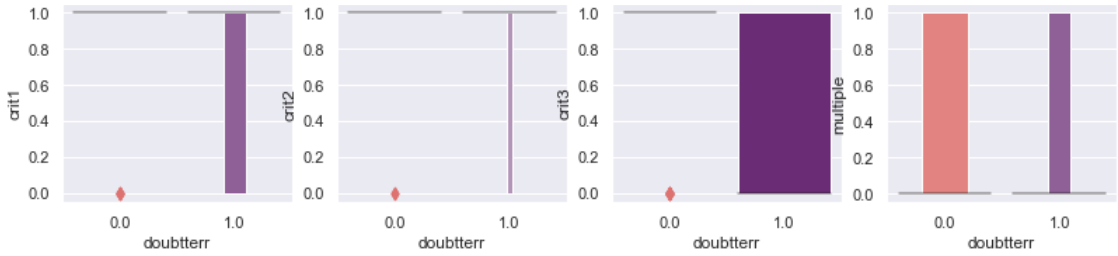
Numerical Features-Latitude, longitude, attacks, wounded, property damage, Fatalities.



As we can see from the graph above that none of the continuous features are normally distributed, Our data ranges from Nearly normal to Substantially positively skewed. Overall Distribution is Positively skewed or Right Skew.

➢ Bivariate analysis (continuous v/s categorical):

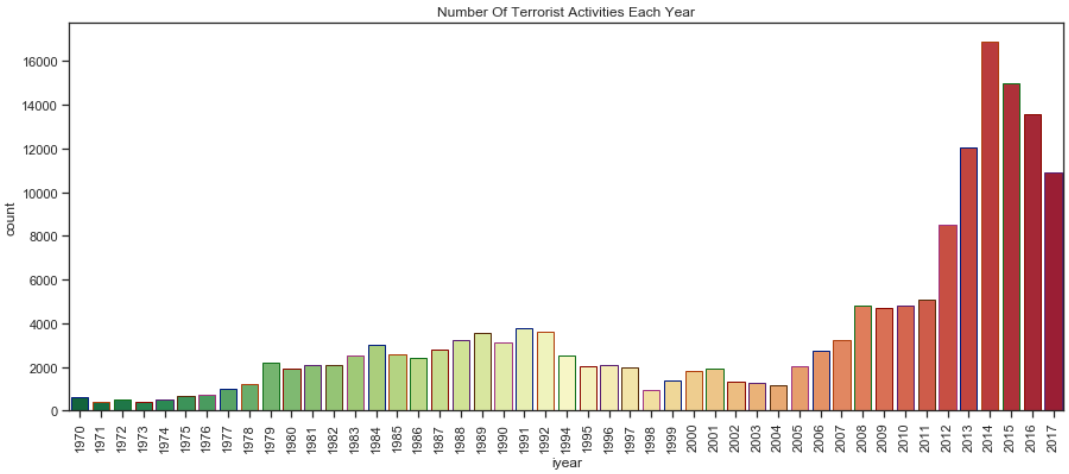




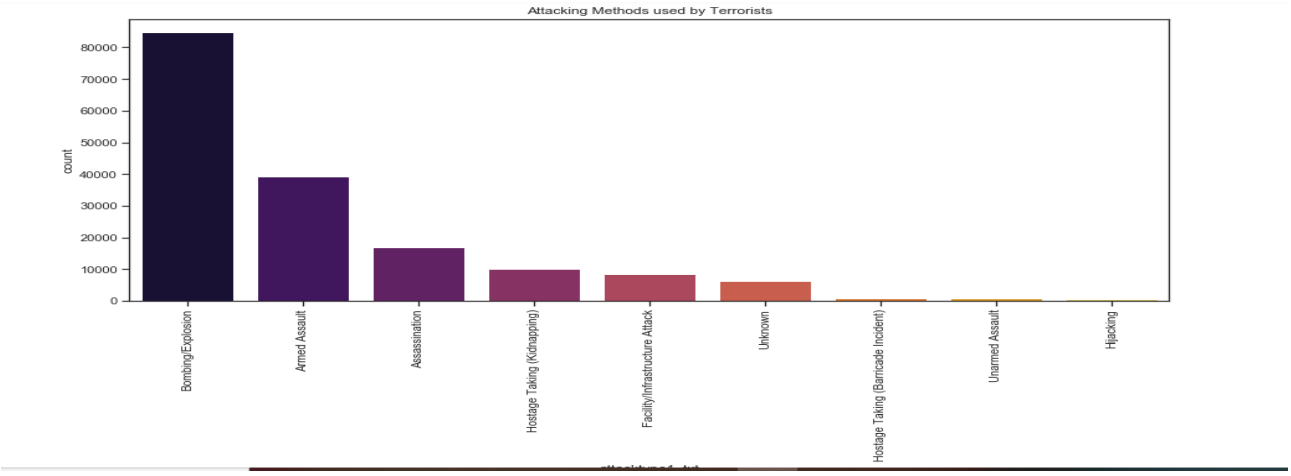
From above plots,we can tell that latitude, longitude, number of perpetrators captured, number of kills, number of people/terrorists wounded has outliers, we must be accepting the outlier's as they are the important events that recorded extreme values in deaths and the number of wounded individuals. Though the feature's crit1,2,3, multiple have outliers but they are discrete number's and we will consider them as categorical feature.

➢ Bivariate analysis using Bar plots and Line plots:

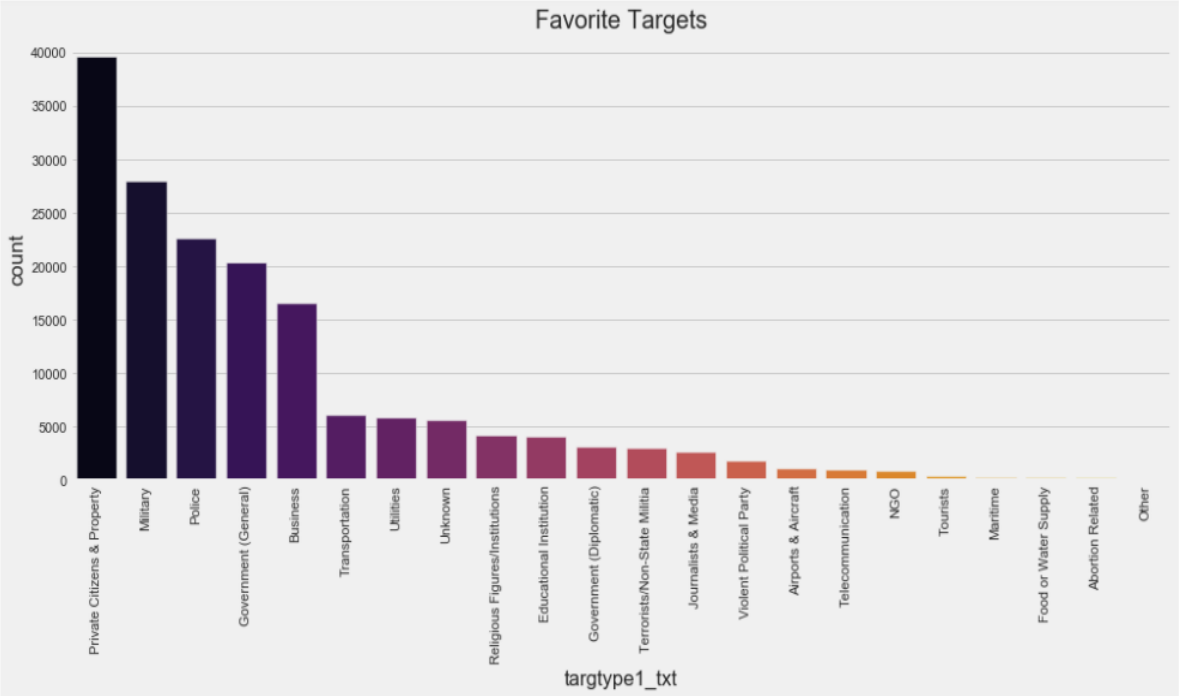
From the below graph we know that, Highest attacks were recorded in 2014.

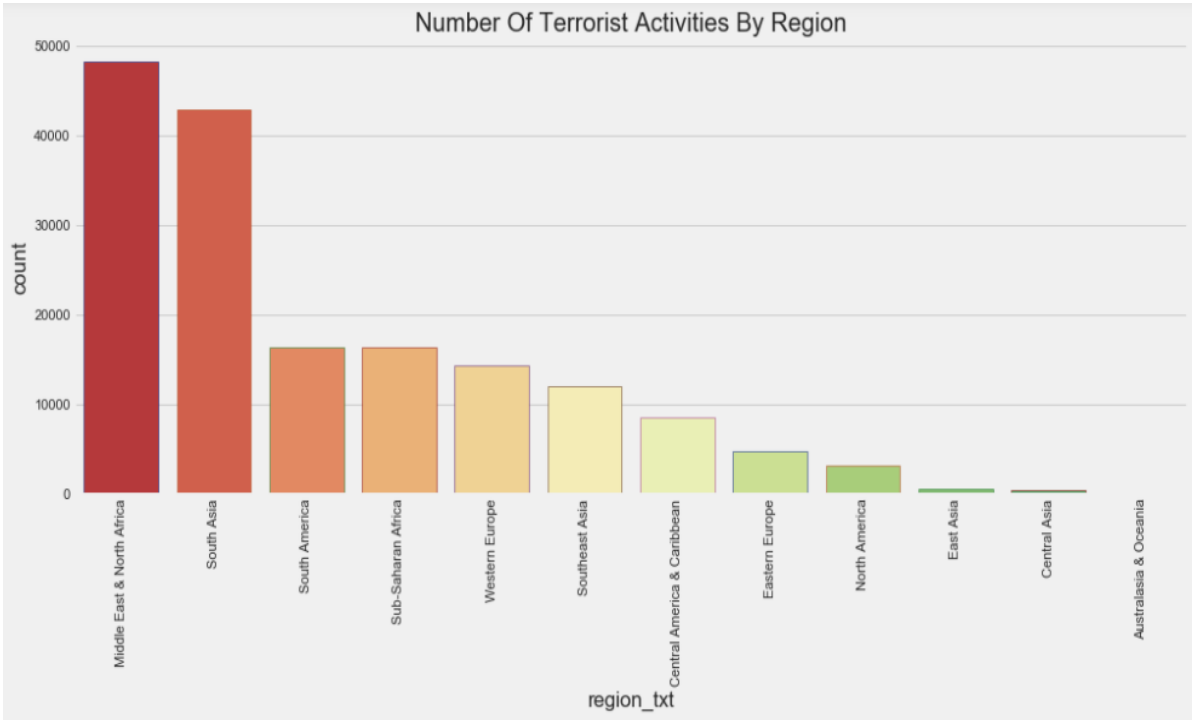


The graph below represents the attack types, most common were bombing and explosion.

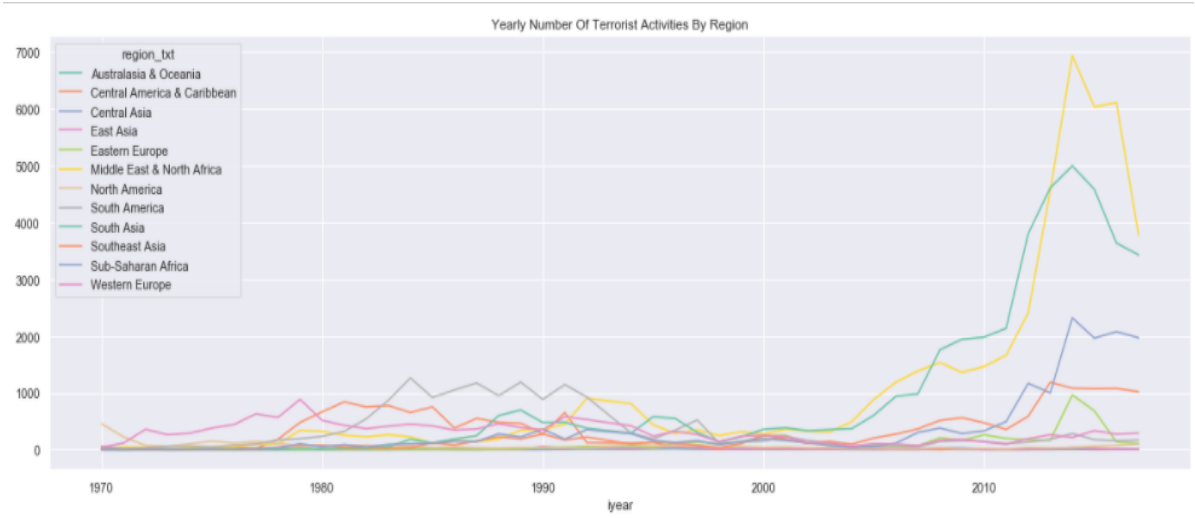


From the below graph, we can say that the Favourite Targets of the perpetrators are: Private Citizens and Property, Military and the Police.

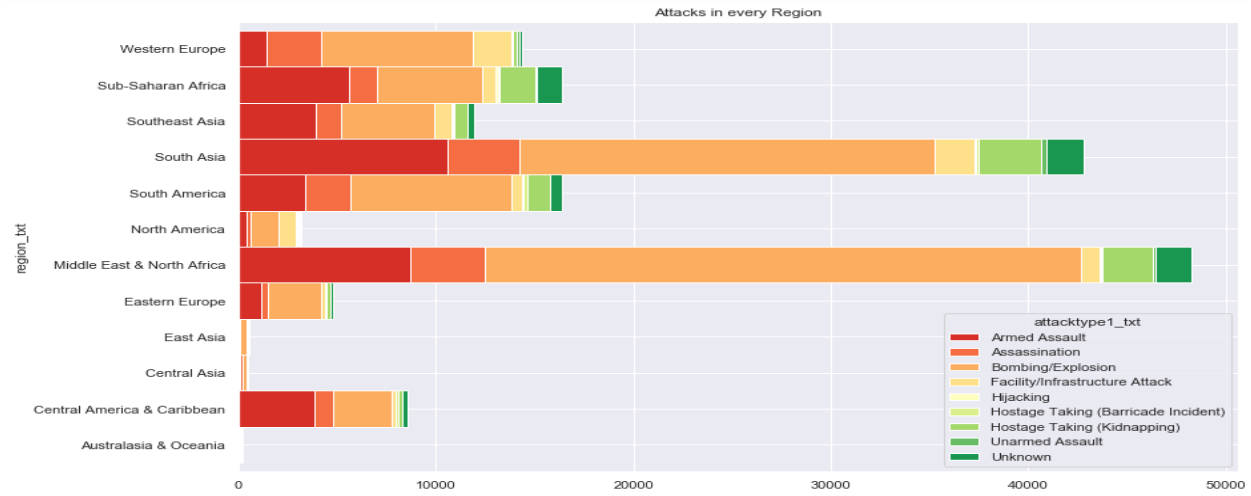




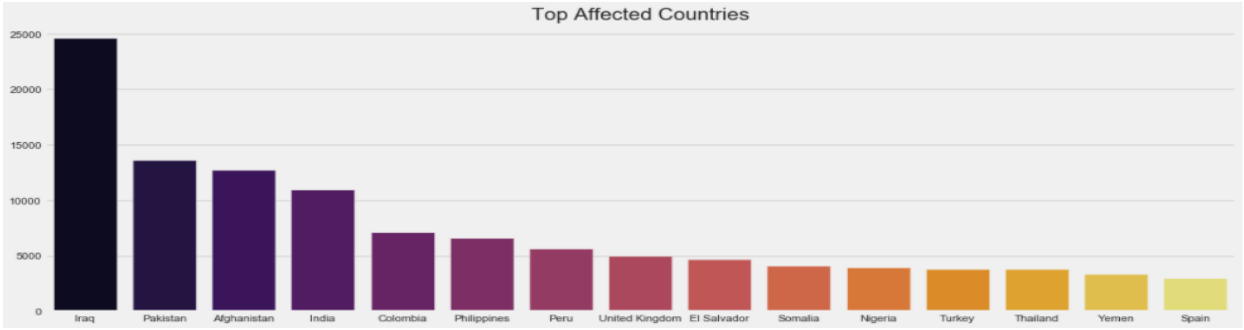
The above graph shows that: Most number of insurgencies are in: Middle East and North Africa, South Asia, South America.



The figure above shows that the region with highest Attacks throughout years are in Middle East and North Africa.



Attacks are the most in Middle East and North Africa and the most common attack type was bombing and explosion.



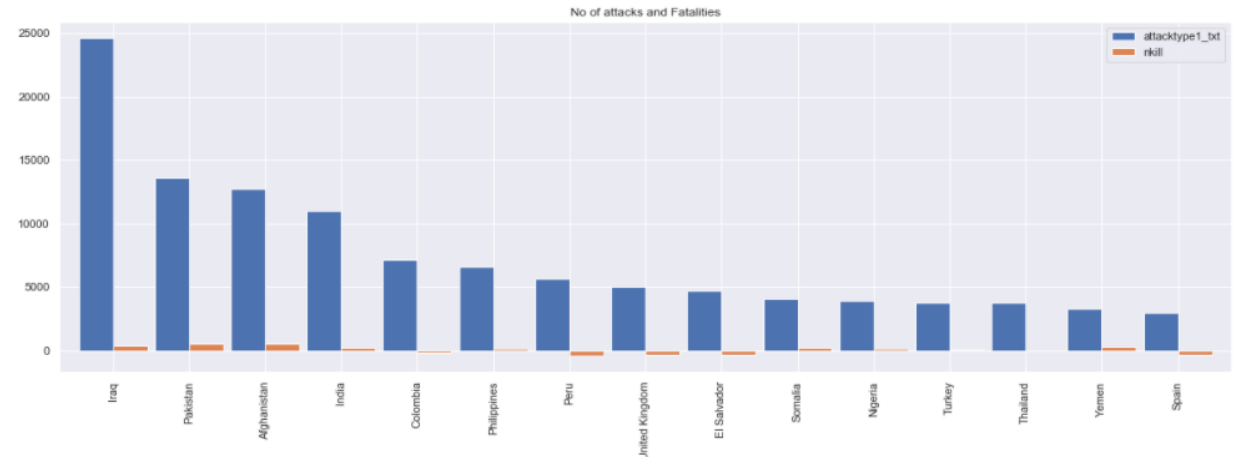
Most affected Countries are: Iraq, Pakistan, Afghanistan as seen from the graph above, reasons are discussed below:

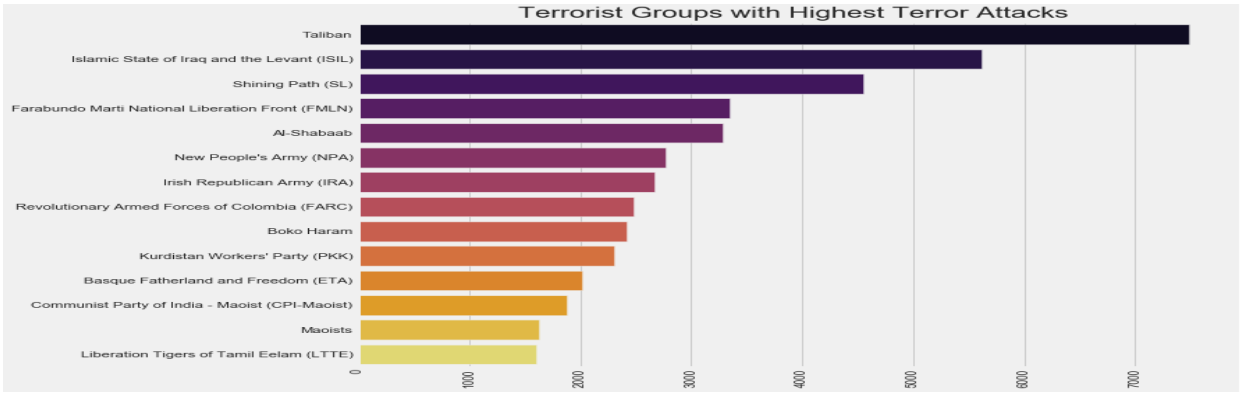
IRAQ: Most attacks are due to the Iraq War which was **a** protracted armed conflict that began in 2003 with **the** invasion of **Iraq** by **a** United States-led coalition that overthrew **the** government of Saddam Hussein. **The** conflict continued for much of **the** next decade as an insurgency emerged to oppose **the** occupying forces and **the** post-invasion **Iraqi** government.

PAKISTAN: The current wave of terrorism in Pakistan is believed to have started in 2000 and peaked during 2009. Since then it has drastically declined as result of military operations conducted by the Pakistan Army in response to these attacks. The attacks are mostly by Tehrik-i-Taliban Pakistan which is a Pakistani terrorist organization that acts against the government.

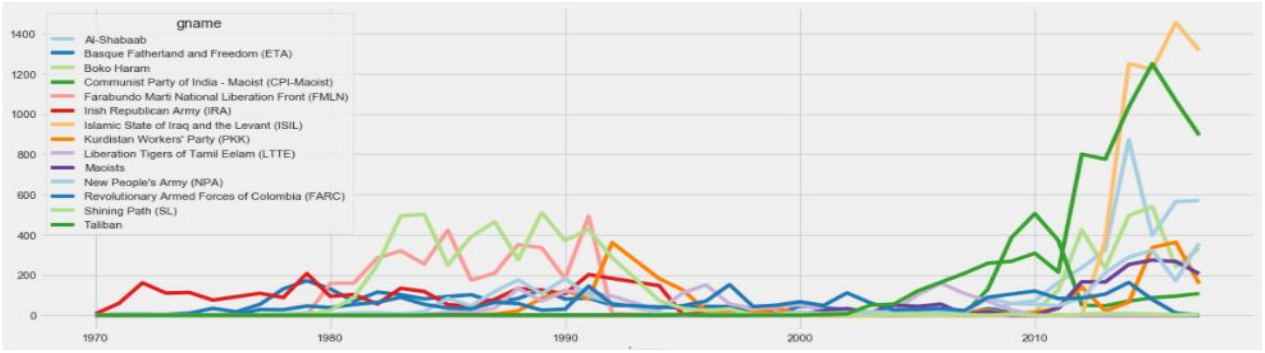
AFGHANISTAN: the Afghanistan it is due to the international conflict in Afghanistan beginning in 2001 that was triggered by the September 11 attacks and consisted of three phases. The first phase—toppling the Taliban (the ultraconservative political and religious faction that ruled Afghanistan and provided sanctuary for al-Qaeda, perpetrators of the September 11 attacks)—was brief, lasting just two months. The second phase, from 2002 until 2008, was marked by a U.S. strategy of defeating the Taliban militarily and rebuilding core institutions of the Afghan state. The third phase, a turn to classic counterinsurgency doctrine, began in 2008 and accelerated with U.S. Pres. Barack Obama's 2009 decision to temporarily increase the U.S. troop presence in Afghanistan.

The graph below represents the attacks and no of fatalities relative to the top 15 countries.



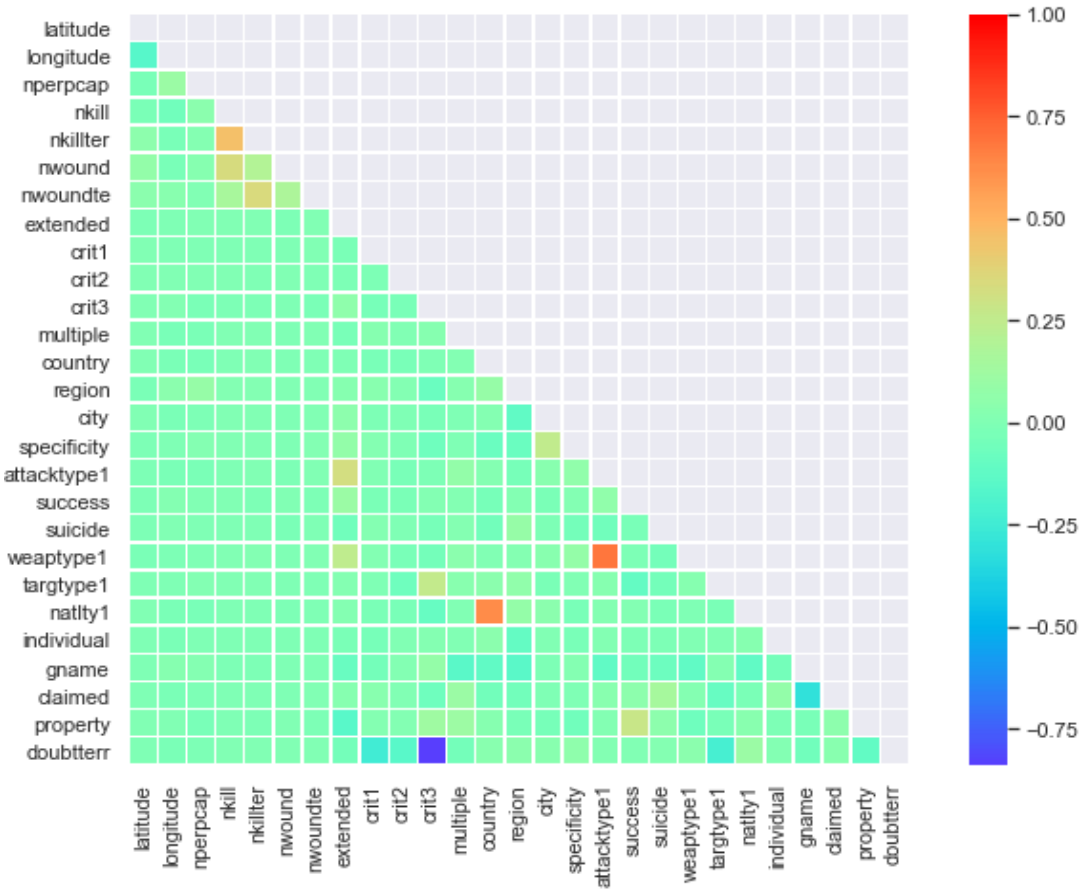


The terrorist Groups that have led to the greatest number of terrorist activities are: Taliban, Islamic State of Iraq and Levant and Shining Path.



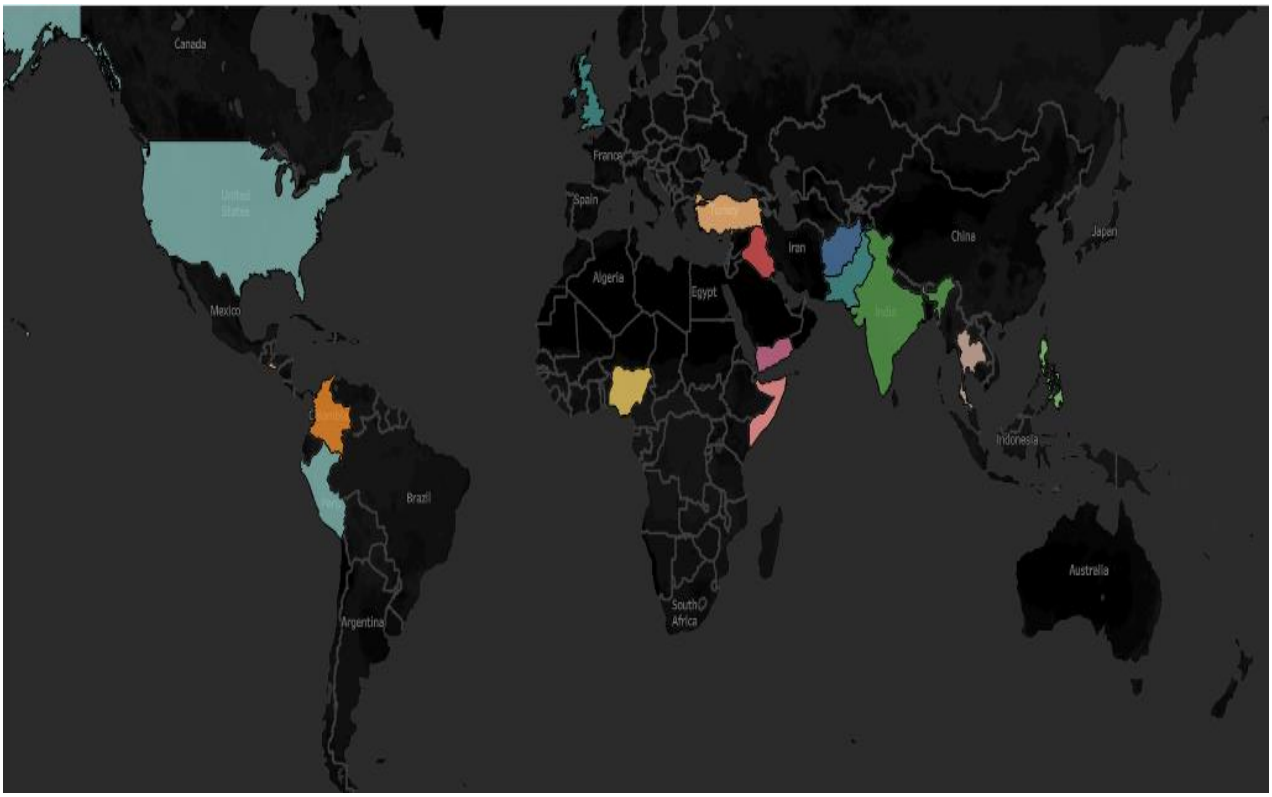
The above graph represents the groups and their attacks over the years.

➢ Correlation Plot:



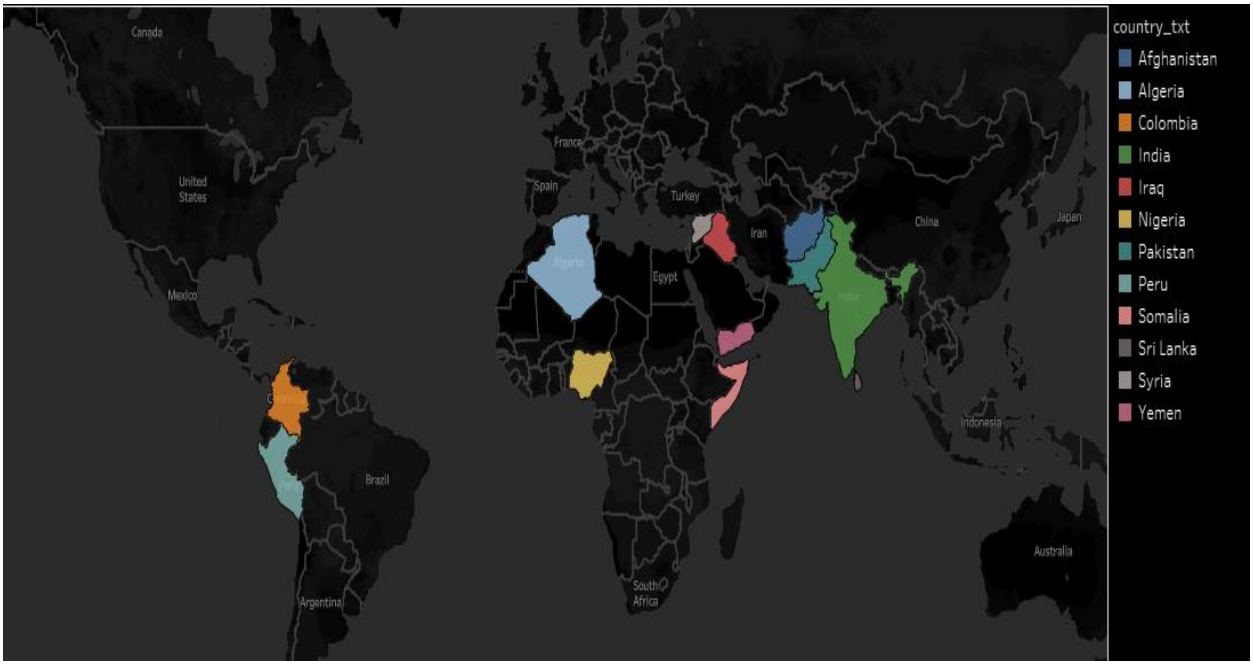
Correlation plot is used to investigate the dependence between multiple variables at the same time and to highlight the most **correlated** variables in a data table. In this visual, **correlation** coefficients are colored according to the value. We can see that: nkillter, nwound, nwoundte are having high correlation amongst them, hence we can call them multicollinear. We will explore that in the multicollinearity section of this report.

Representation of countries with most Attacks:



The countries are India, Pakistan, Afghanistan, Iraq, Nigeria, Peru, Philippines, Colombia, Somalia, Thailand, Turkey, Yemen, El Salvador, United Kingdom, United States.

Representation of countries with most Fatalities:



➢ Multicollinearity

Multicollinearity occurs when two or more independent variables are highly correlated with one another in a linear model.

The variance inflation factor (**VIF**) quantifies the extent of correlation between one predictor and the other predictors in a model. We have used it for diagnosing collinearity/multicollinearity.

	VIF	Feature
4	1.371536	nkillter
3	1.346065	nkill
5	1.173043	nwound
6	1.142016	nwoundte
7	1.074537	gname_map
1	1.060858	longitude
0	1.036410	latitude
8	1.030928	related_to_count
2	1.015546	nperpcap

A value of 1 means that the predictor is not correlated with other variables. The higher the value, the greater the correlation of the variable with other variables. Here we can clearly see that the predictors don't possess any multicollinearity in the Global Terrorism Database as all values are near to 1.

➢ Distribution of the Variables

Variables in our dataset are positively skewed or Right skewed. That's because there is a long tail in the positive direction on the number line.

➢ Class Imbalance

- Our target column 'doubtterr' consists 82.72% data of one class ('0') and 17.27% data of another class ('1').
- Here the majority class is ('0'): Exclusively terrorist activity and the minority class is the ('1'): Doubtful activities.
- We can deal with these using the SMOTE (Synthetic Minority Oversampling Technique).

STATISTICAL ANALYSIS

➢ Test of Significance for Categorical Variables with the Target

For this dataset, we have performed separate Statistical tests for the Categorical and Continuous variables.

Categorical Columns: For checking the significance of the categorical columns with Target variable, Chi-square test of Independence has been used.

Chi-square test of Independence:

The Chi-Square test of independence is used to determine if there is a significant relationship between two nominal (categorical) variables. The frequency of each category for one nominal variable is compared across the categories of the second nominal variable. The data can be displayed in a contingency table where each row represents a category for one variable and each column represents a category for the other variable.

Null Hypothesis: The null hypothesis for this test is that there is no significant relationship between doubtterr and other categorical variables.

Alternate Hypothesis: The alternative hypothesis is that there is a significant relationship between doubtterr and other categorical variables.

The critical value for the chi-square statistic is determined by the level of significance (typically .05) and the degrees of freedom. The degrees of freedom for the chi-square are calculated using the following formula: $df = (r-1)(c-1)$ where r is the number of rows and c is the number of columns.

1	iyear	99.569756	0.000000e+00
2	imonth	99.569756	1.067818e-03
3	iday	99.569756	5.517781e-04
4	extended	99.569756	1.249763e-52
5	related	99.569756	0.000000e+00
6	crit1	99.569756	0.000000e+00

7	crit2	99.569756	0.000000e+00
8	crit3	99.569756	0.000000e+00
9	multiple	99.569756	2.334324e-169
10	country	99.569756	0.000000e+00
11	country_txt	99.569756	0.000000e+00
12	region	99.569756	0.000000e+00
13	region_txt	99.569756	0.000000e+00
14	city	99.569756	0.000000e+00

From the above snippet of the 14 p values out of the total 30, we conclude that:

- The pvalues of all the categorical variables are subsequently lower than the critical value (0.05).
- This implies that we can reject the null hypothesis.
- The observed chi-square test statistic in our case is greater than the critical value (0.05), hence the null hypothesis can be rejected.
- Hence, we conclude that all the categorical variables are significant.

➢ Test of significance of Numerical Variables with the Target Variable

Numerical Columns: For testing the significance of numerical columns with Target variable, the following tests have been performed:

1. ANOVA
2. Barlett test
3. Mann-Whitney test
4. Two-Sample Independent Ttest

1.ANOVA (Analysis of variance):

One-way ANOVA is used for the analysis of independent and dependent variable. It is a technique that can be used to compare means of two or more samples

Null hypothesis: μ1=μ2= μ3

there is no significant difference among the group means.

Alternative hypothesis: μ1≠μ2= μ3 or **μ1=μ2≠ μ3** or **μ1=μ3≠ μ2** or **μ1≠μ3≠ μ2**

There is at least a pair of means that are significantly different among the groups.

Here in our dataset, we notice that for latitude, longitude, number of perpetrators captured, number of individuals killed, number of terrorists killed, number of people/terrorists wounded have pvalues less than the critical value (0.05).

pvalue < **α ,**we reject Ho or null hypothesis, which means at least one pair of means are different for these variables.

2.Barlett's Test:

Bartlett's test is an inferential statistic used to assess the equality of variance in different samples. Some common statistical procedures assume that variances of the populations from which different samples are drawn are equal. Bartlett's test assesses this assumption.

Null hypothesis: $\sigma^2 = \sigma^2 = \sigma^2$

The population variances are equal for the Continuous features and the target variable.

Alternative hypothesis: $\sigma^2 i \neq \sigma^2 j$ *for at least one pair(i, j)*

The population variances are unequal for at least a pair of groups.

Here in our dataset, we notice that for latitude, longitude, number of perpetrators captured, number of individuals killed, number of terrorists killed, number of people/terrorists wounded have pvalues less than the critical value (0.05) for the barletts test.

pvalue < **α ,** we reject Ho or null hypothesis, which means at least one pair of variances are different for these variables across the Population.

3. Mann-Whitney Test:

The Mann Whitney U test, sometimes called the Mann Whitney Wilcoxon Test or the Wilcoxon Rank Sum Test, is used to test whether two samples are likely to derive from the same population (i.e., that the two populations have the same shape).This is a non-parametric test.

Null Hypothesis: the distributions of both populations are equal

Alternate Hypothesis: the distributions of both populations are unequal

pvalue < **α ,** we reject Ho or null hypothesis, which means the distributions of both populations are unequal**.**

4.Two Sample Independent T test:

The independent t-test, also called the two sample t-test, independent-samples t-test or student's t-test, is an inferential statistical test that determines whether there is a statistically significant difference between the means in two unrelated groups. In context to our dataset, the mean of the number of fatalities(nkill: independent variable) and Doubterr(Target Column: Categorical).

Null Hypothesis: $u_1 = u_2$

The null hypothesis for the independent t-test is that the population means from the two unrelated groups (doubter and nkill) are equal.

Alternate Hypothesis: $u_1 \neq u_2$

the population means are not equal.

pvalue < ***α,*** we reject Ho or null hypothesis, which implies that the means of both populations are unequal**.**

Below, we put all the significant features of all the tests in a data frame:

num_col_name	barlett_stat	manwhit_stat	anova_stat	ttest_stat	barlett_p_values	manwhit_p_values	anova_p_values	ttest_p_values
latitude	60.781805	1.955118e+09	7.053666	-2.655874	6.376456e-15	1.757008e-15	7.246321e-16	7.246321e-16
longitude	43.556248	1.910070e+09	7.053666	-2.655874	4.119469e-11	4.597577e-44	5.063706e-38	5.063706e-38
nperpcap	8.442822	2.008757e+09	7.053666	-2.655874	3.664870e-03	2.045357e-01	6.390757e-11	6.390757e-11
nkill	1.294807	2.008975e+09	7.053666	-2.655874	2.551640e-01	2.329752e-01	4.313236e-01	4.313236e-01
nkillter	7.132721	1.949783e+09	7.053666	-2.655874	7.568978e-03	2.225152e-29	1.010059e-02	1.010059e-02
nwound	13.939722	2.003850e+09	7.053666	-2.655874	1.887670e-04	6.928156e-02	1.329268e-03	1.329268e-03
nwoundte	9.018777	1.957084e+09	7.053666	-2.655874	2.672201e-03	1.553596e-35	7.911066e-03	7.911066e-03

Here in our dataset, we notice that for latitude, longitude, number of perpetrators captured, number of individuals killed, number of terrorists killed, number of people/terrorists wounded have pvalues less than the critical value (0.05) and are significant in our analysis of classifying if an event is terrorist or other forms of crime.

FEATURE ENGINEERING

Basically, all machine learning algorithms use some input data to create outputs. This input data comprises features, which are usually in the form of structured columns. Algorithms require features with some specific characteristic to work properly. Here, the need for feature engineering arises. We think feature engineering efforts mainly have two goals:

1. Preparing the proper input dataset, compatible with the machine learning algorithm requirements.
2. Improving the performance of machine learning models.
3. The features you use influence more than everything else the result. No algorithm alone, to my knowledge, can supplement the information gain given by correct feature engineering.

List of Techniques Used:

1. Feature Selection.
2. Transform.
3. Label Encoding.

➢ Feature Selection:

It is the process of selecting a subset of relevant attributes to be used in making the model in machine learning. Effective feature selection eliminates redundant variables and keeps only the best subset of predictors in the model which also gives shorter training times. Besides this, it avoids the curse of dimensionality and enhance generalization by reducing overfitting.

In this project we make the use of Statistical tests and Backward elimination in order to improve the classification performance and/or scalability of the system. An alternative of feature selection is the use a feature extraction technique such as Principal Component Analysis for dimensionality reduction.

```
def back_feature_elem (data_frame,dep_var,col_list):

    while len(col_list)>0 :
        model=sm.Logit(dep_var,data_frame[col_list])
        result=model.fit(disp=0)
        largest_pvalue=round(result.pvalues,3).nlargest(1)
        if largest_pvalue[0]<(0.05):
            return result
            break
        else:
            col_list=col_list.drop(largest_pvalue.index)

result=back_feature_elem(df_constant,df_constant['doubtterr'],cols)
```

Dep. Variable:	doubtterr	**No. Observations:**	91686
Model:	Logit	**Df Residuals:**	91669
Method:	MLE	**Df Model:**	16
Date:	Fri, 28 Aug 2020	**Pseudo R-squ.:**	0.7393
Time:	18:06:19	**Log-Likelihood:**	-10559.
converged:	True	**LL-Null:**	-40497.
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	-2.1727	0.041	-52.709	0.000	-2.253	-2.092
latitude	0.0441	0.021	2.114	0.035	0.003	0.085
nperpcap	-0.1372	0.012	-11.553	0.000	-0.160	-0.114
nwound	0.0839	0.021	4.071	0.000	0.044	0.124
nwoundte	-0.0741	0.026	-2.880	0.004	-0.125	-0.024
crit1	-0.9299	0.048	-19.461	0.000	-1.024	-0.836
crit2	-0.6045	0.048	-12.608	0.000	-0.699	-0.511
crit3	-3.3971	0.098	-34.512	0.000	-3.590	-3.204
multiple	-0.1068	0.023	-4.562	0.000	-0.153	-0.061
country	0.0904	0.021	4.402	0.000	0.050	0.131
region	-0.1326	0.020	-6.547	0.000	-0.172	-0.093
success	0.2253	0.026	8.682	0.000	0.174	0.276
targtype1	-0.1251	0.021	-5.889	0.000	-0.167	-0.083
natlty1	0.1153	0.021	5.604	0.000	0.075	0.156
individual	0.1207	0.010	11.931	0.000	0.101	0.141
claimed	-0.1917	0.025	-7.657	0.000	-0.241	-0.143
property	-0.1058	0.022	-4.886	0.000	-0.148	-0.063

Backward elimination helped us in selecting important features that we can give as an input to our model.

➢ Transformation

The data has substantially positive skewness and hence we will have to use the transformations to make it near normal.

Transformations applied:

Scaling: After applying standard scalar we did not see much change in the skewness.

Yeo Johnson transformation: it is used to make the data more normal distribution-like, improve the validity of measures of association such as the Pearson correlation between variables and for other data stabilization procedures.

We applied it using the below code:

```
from sklearn.preprocessing import PowerTransformer

pt_yeo = PowerTransformer(method='yeo-johnson', copy=True)
td1_num_yeo = pt_yeo.fit_transform(td1[nume_cols])
td1_num_yeo = pd.DataFrame(td1_num_yeo, columns=nume_cols)
td1_num_yeo.head()
```

The Yeo Jhonson transform changed the skewness but it ended up increasing the pvalues of significant features and hence we do not transform our dataset using any of these two.

➢ Label Encoding:

When we convert the numerical feature's to categorical, our normal practice is label encoding for ordinal data and one hot for nominal data, but we can also use label encoding for ordinal data if there isn't any curse of dimensionality, so we will convert the categorical to numerical with label encoding.

In our data we have label encoded two categorical columns: City and Group name of the terrorists.

```
from sklearn import preprocessing
label_encoder = preprocessing.LabelEncoder()
df1['city']= label_encoder.fit_transform(df1['city'])
df1['gname']= label_encoder.fit_transform(df1['gname'])
```

MODELLING WITH NULL IMPUTATION

In our approach we built models both with and without null imputation.

➢ Base Model

Our Base Model is Logistic regression since we are modelling a binary dependent variable.

Below are the scores of the **Logistic Regression model**:

Training score--- 0.9813063934963223 Test score --- 0.9804061699720088

Accuracy – 98.04%

Below is the classification report:

Logistic Regression:

```
Logistic Regression:
              Precision    recall  f1-score   support

           0       0.98      1.00      0.99     27779
           1       1.00      0.89      0.94      5803

    Accuracy                           0.98     33582
   Macro avg       0.99      0.94      0.96     33582
Weighted avg       0.98      0.98      0.98     33582
```

This is a generalized model.

We have also built Decision Tree and Random Forest models and the scores are below:

Decision Tree:

Training score—99.9 Testing score—95.2

Accuracy –95.2

Random forest:

Training score—99.9 Testing score—95.2

Accuracy—97.8

➢ Scaling:

We scaled the data using Standard scalar, then performed logistic regression on the data again, the results are as follows:

We have got the accuracy of 98.1 (increased by 0.1%)

Training score—98.1%

Test score—98.12%

Logistic Regression:

```
Logistic Regression:
              Precision    recall  f1-score   support

           0       0.98      1.00      0.99     41757
           1       1.00      0.89      0.94      8615

    Accuracy                           0.98     50372
   Macro avg       0.99      0.95      0.97     50372
Weighted avg       0.98      0.98      0.98     50372
```

Accuracy, Precision, Recall and F1 are the metrics on which we will evaluate our model:

Accuracy : 98% percent of the models predictions were correct, which seems like a good number.

Precision: Precision answers the question, out of the number of times a model predicted positive, how often was it correct? Well in our case 98% of the time for terrorist activities and 100% for the other criminal activities.

Recall: the fraction of the total amount of relevant instances that were actually retrieved by the model in terms of (‘0’) is 100% and in terms of (‘1’) is 89%.

F1 score: The F-score is a way of combining the precision and recall of the model, and it is defined as the harmonic mean of the model’s precision and recall. For (‘0’) it is 99% and in terms of (‘1’) is 94%.

➢ Cross Validation

Cross validation is used to assess the predictive performance of the model and to judge how it performs on unseen data.

Using cross validation technique, we can verify error (Bias and Variance) performance of the model, thereby we will get to understand how consistent it is over entire dataset since we are validating entire training dataset via validating each subset.

Performance of Different Classification Models after Cross Validation:

Logistic Regression:

Train score – 98.120% Test score – 98.086%

Accuracy score – 98.0862% roc_auc score – 94.5254%

Performed 10-Fold cross validation (Scoring - F1 weighted):

Bias error - 1.92715% Variance error - 0.04473

Gradient Boosting Classifier:

Train score – 98.125% Test score – 98.098%

Accuracy score – 98.0981% roc_auc score – 94.482%

Performed 5-Fold cross validation (Scoring - F1 weighted):

Bias error - 1.91969% Variance error - 0.023116%

XGBoost Classifier:

Train score – 98.127% Test score – 98.094%

Accuracy score – 98.0941% roc_auc score – 98.488%

Performed 5-Fold cross validation (scoring - F1 weighted):

Bias error - 1.91944% Variance error - 0.02247%

- ## Principal Component Analysis on the model built after imputation:

After performing P.C.A on the data, we have observed that the scores have not improved. The scores:

Training Score-0.970034202868957 Test Score - 0.9689510045263241

Accuracy-96.8%

In P.C.A domain the variance explained by all the components is 70%, the rest 30 is almost explaining the same variance.

The P.C.A in this case is not producing any fruitful result and this is because, P.C.A should be used mainly for variables which are strongly correlated. If the relationship is weak between variables as in our case, P.C.A does not work well to reduce data.

In the dataset most of the correlation coefficients are smaller than 0.3, hence, PCA does not produce desired result.

MODELLING BY DROPPING NULLS

Null Value imputation is risky and might lead to misleading results and interpretations and hence we consider these models in our Domain that is built by dropping all the null values assuming that the values are null due to no data of that particular event.

Logit Model: a logit model is similar to Logistic Regression, just that it displays the coefficients instead of the Odds Ratio.

```
cols=df_constant.columns[:-1]
model=sm.Logit(df_constant['doubtterr'],df_constant[cols])
result=model.fit()
result.summary()
```

```
Optimization terminated successfully.
         Current function value: 0.115096
         Iterations 11
```

Logit Regression Results

Dep. Variable:	doubtterr	**No. Observations:**	91686
Model:	Logit	**Df Residuals:**	91659
Method:	MLE	**Df Model:**	26
Date:	Thu, 27 Aug 2020	**Pseudo R-squ.:**	0.7394
Time:	22:11:38	**Log-Likelihood:**	-10553.
converged:	True	**LL-Null:**	-40497.
Covariance Type:	nonrobust	**LLR p-value:**	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	24.3708	0.905	26.917	0.000	22.596	26.145
latitude	0.0658	0.029	2.288	0.022	0.009	0.122
longitude	0.0422	0.030	1.424	0.154	-0.016	0.100
nperpcap	-0.2304	0.020	-11.670	0.000	-0.269	-0.192
nkill	-0.0272	0.027	-1.027	0.305	-0.079	0.025
nkillter	-0.0079	0.027	-0.294	0.768	-0.060	0.045
nwound	0.0990	0.023	4.226	0.000	0.053	0.145
nwoundte	-0.0729	0.028	-2.629	0.009	-0.127	-0.019
extended	-0.0385	0.087	-0.440	0.660	-0.210	0.133
crit1	-8.7440	0.449	-19.466	0.000	-9.624	-7.864
crit2	-8.9604	0.710	-12.618	0.000	-10.352	-7.569
crit3	-10.4854	0.304	-34.494	0.000	-11.081	-9.890
multiple	-0.2843	0.065	-4.384	0.000	-0.411	-0.157

We are applying all the algorithms on Scaled data:

Logistic Regression: After fitting training data to the logistic regression, the accuracy is: 97.41%.

These are the precision recall and F1 scores.

	precision	recall	f1-score	support
0.0	0.97	1.00	0.98	23103
1.0	1.00	0.84	0.91	4403
accuracy			0.97	27506
macro avg	0.98	0.92	0.95	27506
weighted avg	0.97	0.97	0.97	27506

Cross Validation Scores: n folds (5)

Logistic Regression: 0.839340 (0.000447)

cross validation scores:[0.81465729 0.82416302 0.84704878 0.84179779 0.86903328]
Bias error: 16.065996795263125
variance error: 4.466834951634697

Decision Tree: After fitting data to the Decision Tree classifier the accuracy is: 95.40%.

These are the precision recall and F1 scores.

	precision	recall	f1-score	support
0.0	0.98	0.97	0.97	23103
1.0	0.84	0.87	0.86	4403
accuracy			0.95	27506
macro avg	0.91	0.92	0.92	27506
weighted avg	0.95	0.95	0.95	27506

Cross Validation Scores: n folds (5)

Decision Tree: 0.952050 (0.000002)
cross validation scores: [0.95103279 0.95254007 0.95038394 0.95297378 0.95331998]
Bias error: 4.794988669204498
variance error: 0.01629210683022408

Random Forest: After fitting data to the Random Forest Classifier the accuracy is: 83.99%. These are the precision recall and F1 scores.

	precision	recall	f1-score	support
	0.0	0.97	1.00	0.98	23103
	1.0	0.99	0.85	0.91	4403
accuracy				0.97	27506
macro avg		0.98	0.92	0.95	27506
weighted avg		0.97	0.97	0.97	27506

Cross Validation Scores: nfolds (5)

Random Forest: 0.972995 (0.000001)
cross validation scores: [0.97227707 0.97282395 0.97393023 0.9715648 0.97437725]
Bias error: 2.700534139632231
variance error: 0.01343854658120796

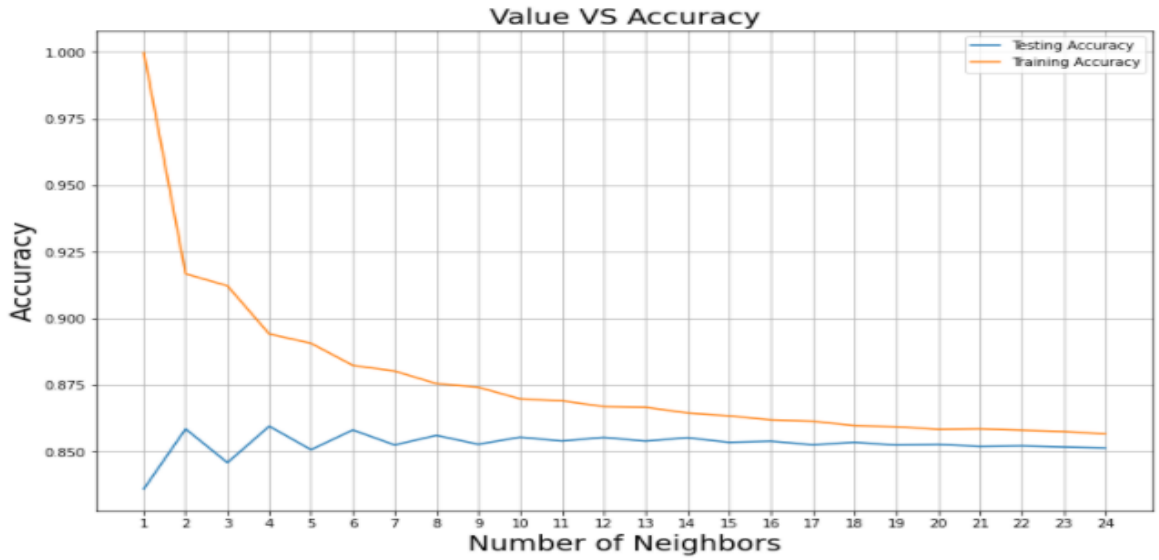
Naïve Bayes Classification: After fitting data to the Naïve Bayes Classifier the accuracy is: 95.75%.These are the precision recall and F1 scores.

	precision	recall	f1-score	support
	0.0	0.97	1.00	0.98	23103
	1.0	0.98	0.85	0.91	4403
accuracy				0.97	27506
macro avg		0.97	0.92	0.94	27506
weighted avg		0.97	0.97	0.97	27506

Cross Validation Scores: nfolds (5)

Naive bayes: 0.953808 (0.000010)
cross validation scores: [0.95219253 0.9518388 0.95687222 0.95048565 0.95764925]
Bias error: 4.619231051422441
variance error: 0.10417535036447967

K Nearest Neighbors: After fitting data to the K Nearest Neighbor Classifier the accuracy is: 84.58% with 3 neighbors. But model getting accuracy of 85.95% with 4 neighbors.




```
Best accuracy is 0.8595942703410165 with K = 4
```

Cross Validation Scores: nfolds (5)

KNN: 0.833394 (0.000032)
cross validation scores: [0.83074954 0.83300635 0.82660464 0.83479221 0.8418177]
Bias error: 16.660591045742002
variance error: 0.3153815243021222

- ## K Means Clustering

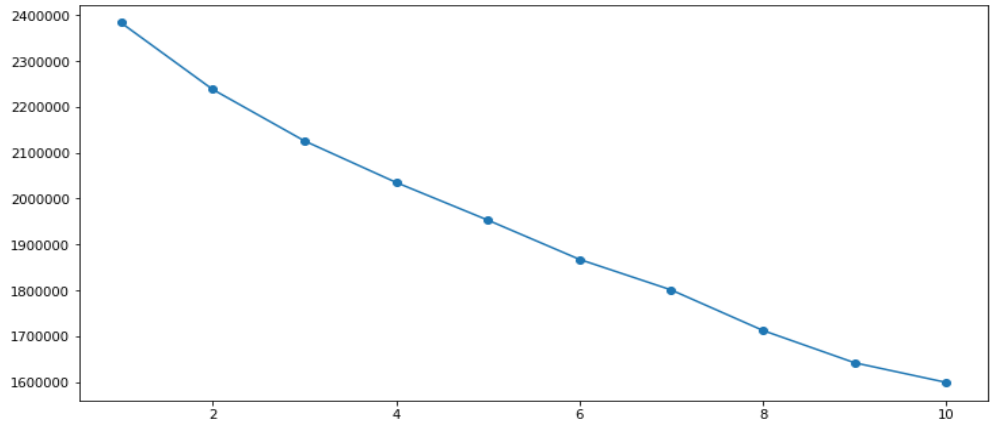
The k-means clustering algorithm is known to be efficient in clustering large data sets

```
cluster_range = range( 1, 11 )
wcss = []
for num_clusters in cluster_range:
  clusters = KMeans( num_clusters, n_init = 10 )
  clusters.fit(data_scaled)
  labels = clusters.labels_
  centroids = clusters.cluster_centers_
  wcss.append( clusters.inertia_ )
clusters_df = pd.DataFrame( { "num_clusters":cluster_range, "Inertia": wcss } )
clusters_df
```

	num_clusters	Inertia
0	1	2.383836e+06
1	2	2.238268e+06
2	3	2.126039e+06
3	4	2.035371e+06
4	5	1.953167e+06
5	6	1.867487e+06
6	7	1.800811e+06
7	8	1.712453e+06
8	9	1.641908e+06
9	10	1.599000e+06

We have small values of inertia, this means that points within the clusters are close to each other, this is ideal.

The basic idea behind partitioning methods, such as k-means clustering, is to define clusters such that the total intra-cluster variation [or total within-cluster sum of square (WSS)] is minimized.



From the above graph we can say that optimum number of clusters is 2 as we see a sharp fall after 2.

K means v/s Original Class:

Comparing original classes and K-Means Algorithm Classes:

```
print('Original Data Classes:')
print(df1['doubtterr'].value_counts())
print('-'*30)
print('Predicted Data Classes')
print(df_labeled['labels'].value_counts())
```

```
Original Data Classes:
0.0    76904
1.0    14782
Name: doubtterr, dtype: int64
------------------------------
Predicted Data Classes
1    75785
0    15901
Name: labels, dtype: int64
```

From above comparison of original classes and k-means algorithm classes, it is observed that our algorithm is classifying correctly for most of the classes.

- ## Principal Component Analysis after dropping all the nulls

After performing P.C.A on the data, we have observed that the scores have not improved, but stayed the same. The scores in the Logistic Regression model :

```
print('Logistic Regression:')
print(classification_report(ypca_test, ypca_pred))
```

```
Logistic Regression:
              precision    recall  f1-score   support

         0.0       0.97      1.00      0.98     23184
         1.0       1.00      0.84      0.91      4322

    accuracy                           0.97     27506
   macro avg       0.98      0.92      0.95     27506
weighted avg       0.97      0.97      0.97     27506
```

Training Score-97.2% Test Score – 97.3%

Accuracy-97.3%(accuracy is better for the model after dropping the nulls)

The P.C.A in this case is not producing any fruitful result and this is because, P.C.A should be used mainly for variables which are strongly correlated. If the relationship is weak between variables as in our case, P.C.A does not work well to reduce data.

We have observed that in both the cases, with the imputation of nulls or after dropping nulls, P.C.A does not improve the scores.

In the dataset most of the correlation coefficients are smaller than 0.3, hence, P.C.A does not produce desired result in both the cases.

- ## SMOTE on imbalanced data:

We have also performed Data Balancing technique: SMOTE **(Synthetic Minority Oversampling Technique)**

SMOTE (synthetic minority oversampling technique) is one of the most commonly used oversampling methods to solve the imbalance problem.
It aims to balance class distribution by randomly increasing minority class examples by replicating them.
SMOTE synthesises new minority instances between existing minority instances. It generates the **virtual training records by linear interpolation** for the minority class. These synthetic training records are generated by randomly selecting one or more of the k-nearest neighbours for each example in the minority class. After the oversampling process, the data is reconstructed and several classification models can be applied for the processed data.

SMOTE on unscaled data:

Accuracy—97.794%

Precision	recall	f1-score	support
0	0.98	0.99	41723
1	0.97	0.90	8649
Accuracy		0.98	50372
Macro avg	0.97	0.95	0.96	50372
Weighted avg	0.98	0.98	0.98	50372

SMOTE on scaled data:

Accuracy—72.224%

Precision	recall	f1-score	support
0	0.83	0.84	41757
1	0.17	0.16	8615
Accuracy		0.72	50372
Macro avg	0.50	0.50	0.50	50372
Weighted avg	0.71	0.72	0.72	50372

However, in our domain we do not strive to treat the class imbalance as it is evident that the G.T.D will have more terrorist activities and less events of other forms of crime.

SMOTE might tamper the results by making both classes alike, which we don't wish for in this Domain.

KEY INSIGHTS, RECOMENDATIONS AND VALUE ADDITION

Key Insights from models:

- In all the models are very good model apart from a few of the assumptions not going in favours of the linear model guidelines mainly normality.
- The Logistic Regression model and the Naive bayes Models have the best accuracy of 97% and 95.7% namely.

Recommendations and Value additions of the model:

- Latitude and Longitude act as important parameters to decide if an attack in a particular region can be called terrorist or not.
- Once we are aware of the Latitude and longitude of the regions that have a tendency of being attacked potentially, we can take the following steps:
    1. Government can mark that area in the public map as a Red Alert and restrict public movement in that area.
    2. The administration can also increase surveillance in that area and have more cameras both hidden and visible so that even if the visible cameras are dismantled, the footages can still be recorded from the hidden ones, this in turn will leads to an arrest of the suspicious individuals.
    3. More security troops should be sent in that area in undercover form.
    4. More hospitals and healthcare should be made available to individual's in that area so that even if there is an attack despite the aforementioned steps, the risk of death can be reduced by proactive medical treatment and support.
- Number of perpetrators captured is another parameter of great significance as shown by the model. Reason being, if the perpetrator is captured, it can be known clearly about the motive of attack and whether or not it meets the three criterions to call them terrorist. If it is terrorist, then the intelligence bureau can launch further investigations to find out what groups or countries are responsible for this.
- Direct economic advantage or the Risk Mitigation done by the model: Property damage is another significant feature identified by our model.
  This implies that the latitude and longitude where there are frequent property damages and the areas in the vicinity of it can be annually insured by the government for the property damage to minimize, monitor, and control the probability or impact of unfortunate events or to maximize the realization of opportunities.

LIMITATIONS

1.Null Value imputation is risky and might lead to misleading results and interpretations: For example: the column ‘nkill’ has 10313 null values. If we impute these with mean, median or k nearest neighbors imputer it will depict that at every instance of an attack, there were fatalities. While in reality there were no fatalities reported in that particular event.

2.Outlier Treatment might lead to information loss: We have extreme values in our data that significantly differ from the other values. However, in our case Outliers are very informative about the subject-area and data collection process.

For instance, in the boxplots representing the number of attacks, the outliers represent the Region with a high frequency of attack. In this case, removal of outlier will lead to loss of important data that is of great value to our analysis. Hence, it is essential to understand how outliers occur in this case and how it’s a normal part of this process or study area.

3. Data transformation is not significant in our case as our model is a binary classification model and data need not always be nearly normal in order to perform parametric tests for the same.

REFERENCES

- Wikipedia
- Analytics Vidya
- Towards Data Science
- Machine Learning using Python: *Mana Ranjan Pradhan and U Dinesh Kumar, Wiley Publication*