VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

A Data Analytics (CSE3505) Final project report

"Visual Analytics of Terrorism and its counter measuring"

Submitted to:

Dr. Tulasi Prasad Sariki

Submitted by:

Reg No: 19BCE1670 Name: SAMARTH SINHA

Reg No: 19BCE1447 Name: KUNAL SUDHIR MISHRA

In partial fulfillment for the award of the degree of

Bachelor of Technology

in

Computer Science and Engineering

DECEMBER 2021

DECLARATION BY THE CANDIDATE

I hereby declare that the report titled "**Visual analysis of Terrorism and its counter measuring"** submitted by me to VIT Chennai is a record of bona-fide work undertaken by me under the supervision of **Dr. Tulasi Prasad Sariki, Associate Professor, SCOPE, Vellore Institute of Technology, Chennai.**

Signature of the Candidate

ACKNOWLEDGEMENT

We wish to express our sincere thanks and deep sense of gratitude to our project guide, **Prof. Tulasi Prasad Sariki,** School of Computer Science and Engineering for hIS consistent encouragement and valuable guidance offered to us throughout the course of the project work.

We are extremely grateful to **Dr. R. Ganesan, Dean,** School of Computer Science and Engineering (SCOPE), Vellore Institute of Technology, Chennai, for extending the facilities of the School towards our project and for his unstinting support.

We express our thanks to our **Head of the Department** for his support throughout the course of this project.

We also take this opportunity to thank all the faculty of the School for their support and their wisdom imparted to us throughout the courses.

We thank our parents, family, and friends for bearing with us throughout the course of our project and for the opportunity they provided us in undergoing this course in such a prestigious institution.

BONAFIDE CERTIFICATE

Certified that this project report entitled "**Visual analysis of Terrorism and its counter measuring"** is a bona-fide work of **Samarth Sinha(19BCE1670),Kunal Sudhir Mishra(19BCE1447)** carried out the "J-Components" Project work under my supervision and guidance for **CSE3505-Foundation of Data Analytics**.

Dr. Tulasi Prasad Sariki

SCOPE

TABLE OF CONTENTS

ABSTRACT

In this work, I present analytical results obtained by data mining on the START (Study of Terrorism and Response to Terrorism) dataset. The main objective is to visualize terrorism data and make it available to users in an easy to understand format. A website is designed which contains a collection of various analyses and visualizations to interpret patterns and trends in it. The website also contains a visualization tool that provides the user with dataset exploration capabilities.

Lack of understanding and awareness about global terrorism leads to diverse opinions and common misconceptions among civilians. In this age of globalization, sufficient information about this topic can help strengthen our counter-terrorism strategies, improvise security concerns, regulate better economic policies and enhance the knowledge base of civilians.

The primary dataset for this project is provided by START Consortium which contains data of

terrorist events since 1970. Performing various data mining and data visualization techniques to interpret the nature of terrorism to better understand its trends and patterns in over 45 years of its recorded history.

INTRODUCTION

World peace was one of the core reasons for forming the United Nations organization. Terrorism is the biggest hurdle to world peace. Terrorism is commonly ignored by the civilians who are not affected directly by the dangers. For the most part, terrorism is considered unpredictable and unfortunate calamity that strikes some parts of the world more than others. Based on the location of the events, people at large have very limited information about any such event happening in other parts of the world and hence react differently. In this project, we focus on terrorism by analyzing the dataset provided by START (Study of Terrorism and Response to Terrorism) Consortium to explore meaningful patterns and statistics.

Terrorism is an unsettled term. Currently, the General Assembly of the United Nations is unable to agree on a single definition of terrorism. Because of this difficulty, different governments and organizations define terrorism in their own way. This confusion creates multiple conflicts about which events are considered under terrorism and which are not. Different organizations construct their own definition of terrorism and operate accordingly. As a result, there could be a reasonable difference in the contents of terrorism-related datasets collected by independent organizations . Hence, analyses and results provided in this study might vary with the similar analyses done on a different dataset.

BACKGROUND

a. Misconception about Terrorism

Terrorism is sporadic, widespread and inconsistent with time and nature. Because of these characteristics, international terrorism is difficult to summarize all aspects as a single conclusive solution and make this information available to be easily understood by most people. Exploring this dataset can provide an insight into how different parameters are correlated with each other, which can help identify unknown hidden patterns. This exploration will also assert enough facts to provide justifications for some common misconceptions regarding terrorism.

One of misconceptions is that more military can suppress and control terrorism. However, using the instrumental variable approach, studies show that counter-terrorism solutions like more military spending is not enough to control terrorism and is also dependent on other factors like economy and national politics.

Another popular misconception is that terrorism only affects the individuals directly involved in any terrorist event. Terrorism adversely impacts not only the economy of the victim country but also the countries financially associated with international terrorism. Empirical evidence shows, the effects of terrorism concerning the attack type have a strong correlation with stock markets of countries, especially in the SAARC (South) Asian Association for Regional Cooperation) region. Multiple regression on stock market data with terrorism as a control variable directly provides a strong connection among both.

a). Factors Affecting Terrorism

Identifying dependent factors of terrorism is one of the goals of this project. There are parameters like religion or nationalism which are not defined in the dataset but have a major influence on contemporary terrorism. Religion has been a very controversial topic among researchers about whether religion influences terrorism or not and if it does, up to what extent. Conclusive evidence shows how religious idealization or belief can shape and transform terrorism [12]. Religious idealization has been one of the major motivating factors leading to fanaticism and in turn, evolve into terrorism. Classification mining is done using a C4.5 algorithm with 10-fold cross-validation to generate a classification tree model resulting in an accuracy of 93.53% predicting the correct set of events; Religious event being the major dependent variable. Hence religion’s contribution in terrorism is an interesting subject to explore.

Dataset Challenges

START terrorism dataset has a marginally low occurrence of events occurred at the same geolocation. Most of the events are not consistent or do not occur frequently. Hence difficulty arise in making quantitative projections with varying degrees of similar events. As a result, different classification techniques provide different results. In this case, Lazy Classifier IBK, Linear NN search and Filtered Neighbor Search techniques provide higher accuracy on dataset compared to Naïve Bayes, Multiclass Classifier and Multilayer perceptron. This helps in understanding which techniques and methodologies are more effective for similar analysis on this dataset.

Another major challenge while working on this dataset is that individual studies lead to different conclusions. Current shortcomings and limitations in data collection techniques, definition debates, irregularity in coding and analysis give rise to disagreements among researchers and in turn ruling out their conclusions. An acceptable level of theoretical and empirical analysis is required to prove a heuristic casual model showing links between globalization and terrorism. One of the issues is critical disagreement over the definitional debates around various terrorist events exerts a detrimental influence on this field's development. This issue demands to exercise a need for common grounds that can be accepted by most experts and concerning authorities to agree on what could be the standard norms and procedure to be considered as a legitimate piece of information on terrorism on which appropriate researches can be done.

\

SYSTEM DESIGN

a) Project Structure

The problem statement here is to build a tool that can present processed information in the form of intuitive visual representation of analyzed data. Implementation of this project involves system design, backend design, visual design, and user interface.

System design includes the overall design plan of the whole project system which explains how each individual module is correlated with others.

Backend design contains a series of data preprocessing steps to transform the raw dataset into a more meaningful and focused collection required for this project. This design module also includes scripts for analyses and other factual information derived from the dataset.

Visual design mostly consists of analyses and visualization techniques to construct different graphics representing the end results in an easy-to-interpret format.

Project Design - Solution Approach

Most of the operations on the dataset are done by R Studio. R is used for data preprocessing, data modeling, analyses, and visualization. Anaconda is used as an open-source python distribution for handling R based dependencies and provide a environment for code development.

Data preprocessing:

Data preprocessing is the first step to be done after collecting data. It is a set of operations performed on the START (Study of Terrorism and Response to Terrorism) dataset to modify ambiguous data which can be a bottleneck to analytical results. Raw data is simply a collection of related information put together. Raw data is often unorganized and contains a lot of information which is irrelevant to the project requirements. Data preprocessing methodology helps in converting this raw data into a more meaningful, focused, interpretable and readable format.

Available START dataset from the Global Terrorism Database is incomplete, inconsistent, contains many errors, missing attributes values, contains outliers, incorrect tags, and duplicate entries. Data preprocessing can help resolve these discrepancies. The following are the steps used in this project as a part of data preprocessing methodology:-

Data cleaning is a process of filling missing values, removing outliers and handle inconsistencies in data. In terrorism dataset, there are numerous fields like 'motives' or 'responsible organizations' which are missing either due to information not available or that field was not relevant for that specific event. Fields like 'summary', 'claim_mode', 'claimmode_txt', 'guncertain', 'nperps' etc. are removed since they are not relevant to the analysis of this project. Fields like 'weapsubtype2', and 'weaptype3_txt' have more missing values than valid entries. Hence such fields are also removed to reduce complexity.

Data integration: In this step, conflicts among data are resolved. Different representations of the same data such as multiple subcategories of weapon type (weapsubtype1, weapsubtype2, weapsubtype3) are put together to avoid confusion and duplications. Fields with one to one correspondence like ‘country_code’ and ‘country_txt’ are mapped to avoid any conflicts.

Data transformation: Here data aggregation, generalization, and normalization are performed. Dataset has multiple target/victim subtypes. All those subtypes were aggregated to represent one value by summation of all similar subtypes. This technique reduces the total number of attributes in the dataset and hence reducing the variability in the data. There are multiple categorical attributes present in this dataset belonging to the same superset. For example, weapon Sub-Type has 4 different attributes which can contain one of the 27 different values. Those 4 attributes were generalized into one weapon Sub-Type and 27 different categorical values were generalized into 12 domains. Values like a grenade, landmine, dynamite, etc. were classified under the ‘explosives’ tag.

Dimensionality reduction: START dataset has high data sparsity which increases its overall dimensionality. This method reduces the effectiveness of density related operations like clustering and outlier detection. There are multiple fields having more missing or null values than valid ones. Some of them are ‘Mode_for_claim_of_responsibility’, ‘divert’, ‘kidhijcountry’ etc. which are of less significance to our project. Such attributes are removed to reduce dataset processing time, avoid the curse of dimensionality and ease of data visualization. However, some missing values in attributes like property_damage and motives which are of high importance cannot be removed. In such cases, missing values of property_damage and motives were replaced by the mean value of corresponding attributes associated with the ‘responsible group’ attributes.

Methodology

The problem statement here is to build a tool that can present processed information in the form of intuitive visual representation of analyzed data. Implementation of this project involves system design, backend design, visual design, and user interface. System design includes the overall design plan of the whole project system which explains how each individual module is correlated with others. Backend design contains a series of data preprocessing steps to transform the raw dataset into a more meaningful and focused collection required for this project. This design module also includes scripts for analyses and other factual information derived from the dataset. Visual design mostly consists of analyses and visualization techniques to construct different graphics representing the end results in an easy-to-interpret format.

Most of the operations on the dataset are done by R Studio. R is used for data preprocessing, data modeling, analyses, and visualization. Anaconda is used as an open-source python distribution for handling R based dependencies and provide a environment for code development. R is a high-level interpreted language that supports different platforms like Windows, R studio, etc. It can be used for high level data analysis. Tableau Desktop software application is used as a data visualization tool for raw data simplification in an easy to understand format. Some of Tableau’s popular features include data collaboration, analysis of real-time data and data blending

Related Works (Literature Review)

START terrorism dataset has a marginally low occurrence of events occurred at the same geolocation. Most of the events are not consistent or do not occur frequently. Hence difficulty arise in making quantitative projections with varying degrees of similar events. As a result, different classification techniques provide different results. In this case, we tried Lazy Classifier IBK, Linear NN search and Filtered Neighbor Search techniques provide higher accuracy on dataset compared to Naïve Bayes, Multiclass Classifier and Multilayer perceptron. This helps in understanding which techniques and methodologies are more effective for similar analysis on this dataset.

The problem statement here is to build a tool that can present processed information in the form of intuitive visual representation of analyzed data. Implementation of this project involves system design, backend design, visual design, and user interface. System design includes the overall design plan of the whole project system which explains how each individual module is correlated with others. Backend design contains a series of data preprocessing steps to transform the raw dataset into a more meaningful and focused collection required for this project. This design module also includes scripts for analyses and other factual information derived from the dataset.

Most of the operations on the dataset are done by R Studio. R is used for data preprocessing, data modeling, analyses, and visualization. Anaconda is used as an open-source python distribution for handling R based dependencies and provide a environment for code development. Visual design mostly consists of analyses and visualization techniques to construct different graphics representing the end results in an easy-to-interpret format.

Visual Analysis

This section consists of details regarding the visual results for the website.

a) Animation of Terrorist Activities

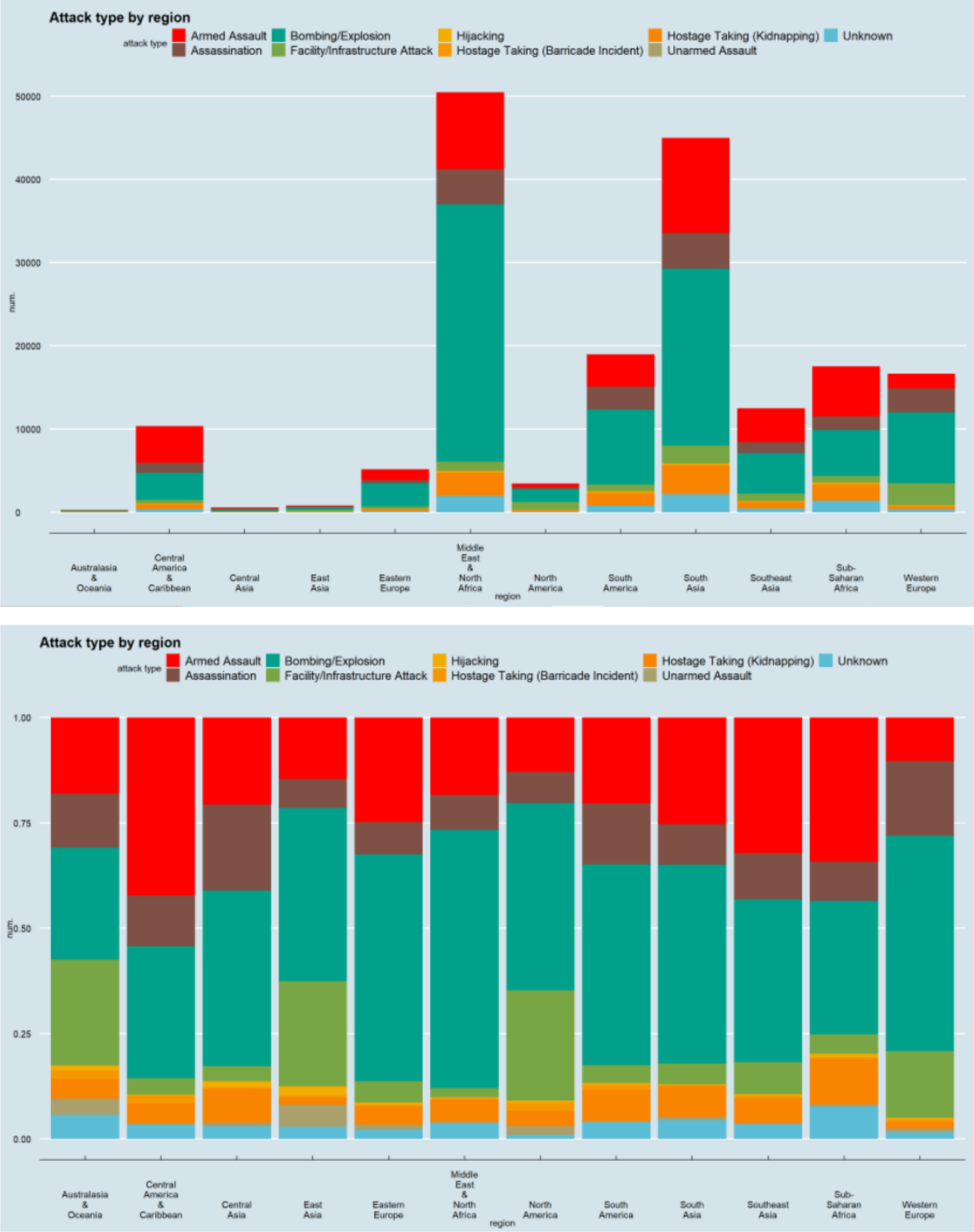


Figure 3. Attacking Methods by Terrorist

INFERENCES:

Different types of weapons and methods have been used by attackers. There are 8 categorical values for the defined attack type. They are unarmed assault, Infrastructure attack, kidnapping, barricade incident, hijacking, bombing/Explosion, armed assault and assassination. These attributes can explain which are the most often used means of attack. Figure 3 does reveal the potential target or focus of the attacker. For instance, unarmed assault attacks are usually focusing against specific individuals or a group of small people. Explosives and bombings are targeted towards a larger audience. Hijacking aims to achieve some sort of ransom in return. Here the graph pictures the total number of kill counts with respect to specific attacking methods used. Figure 3 uses the nine most used means of attacking based on the causalities caused. Explosions are the most common followed by armed assaults, assassinations, hostage and so on. Here the total number of casualties by explosive weapons is almost double than the next most attack which is armed assault. This observation indicates that most of the attacks were intended to civilians for the purpose of spreading terror among widespread targets.

3.3.2 Events by Year

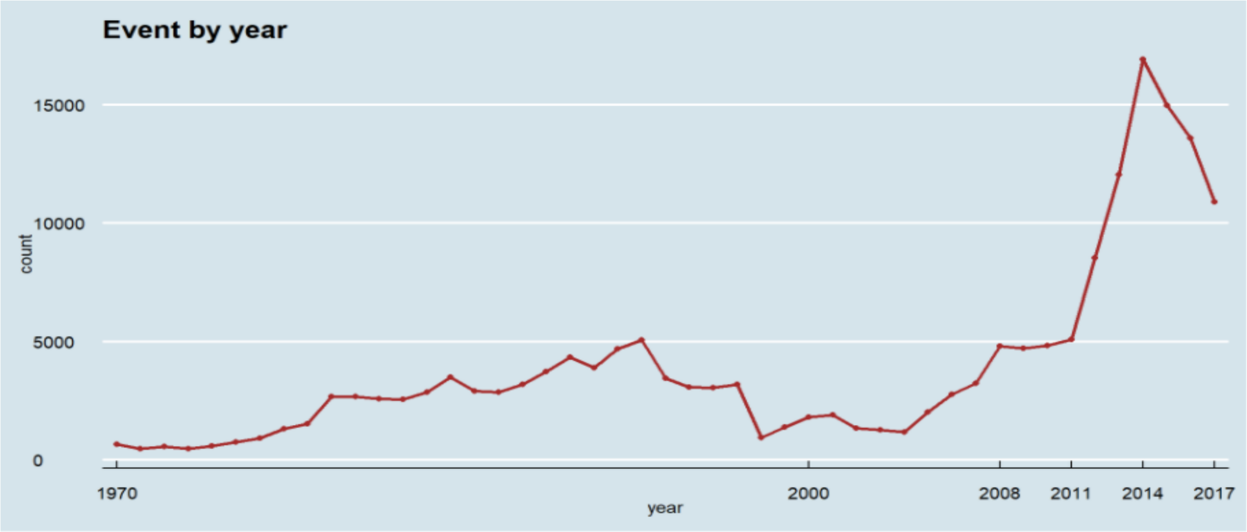


Figure 4. Events by Year

There is a rapid increase in terrorist event since year 2000. We’ll separately observe the trend by the region.

3.3.3 Correlation Matrix

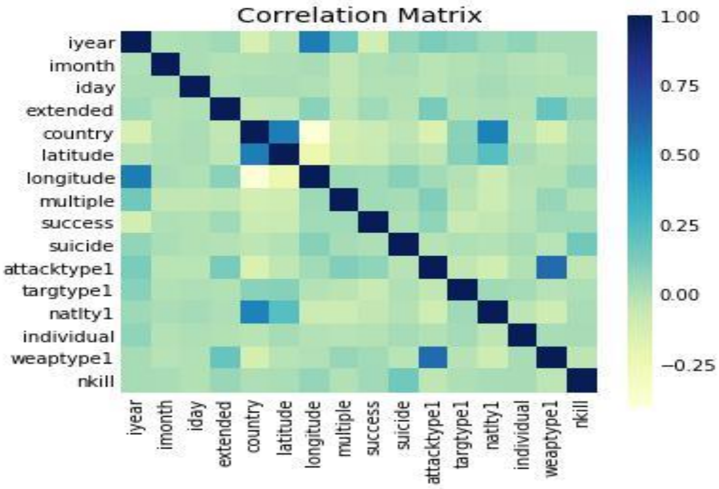


Figure 5. Correlation Matrix

Finding dependencies among the various parameters in the dataset can reveal a key pattern about the nature of terrorism. Out of 120 variables, we have selected the most significant 16 for this map. Some of those parameters are the day, year, country, latitude, longitude, success rate of attack, type of attack, target type, number of kills, etc. Forming a correlation map can provide us with one-to-one correspondence of each variable with rest. Figure 5 shows correlation matrix where darker the shade of the block, more the attributes are correlated proportionately. Here we can see that country and latitude are correlated which is expected. Values of neither of those two parameters change and hence they show a strong relation. Another relation we can see is among '*natlty1*' and '*country*'.'*natlty1*' defines the nationality of the attacker and '*country*' defines the country where the attack took place. This observation shows that most of the attacks are done by the citizen of their own country. Such a relation provides an interesting insight into how to perceive international terrorism as the proportion of international terrorism is significantly less in comparison with domestic terrorism. Attack type and weapon used in the attack also hold close ties with each other as attack type is defined based on the weapons used in that incident. Strangely '*success*' which represents the rate of success of any attack, shows no significant connections with any other listed parameters. The block representing year and success has a darker shade which means that both these parameters are inversely related to each other. So, over time, the rate of success of any attack has reduced. This is a noteworthy observation that in an era of growing terrorism, counter-terrorist forces can restrict the success factor of attacks more than they used to.

3.3.4 Terrorist Attacks by Each Year

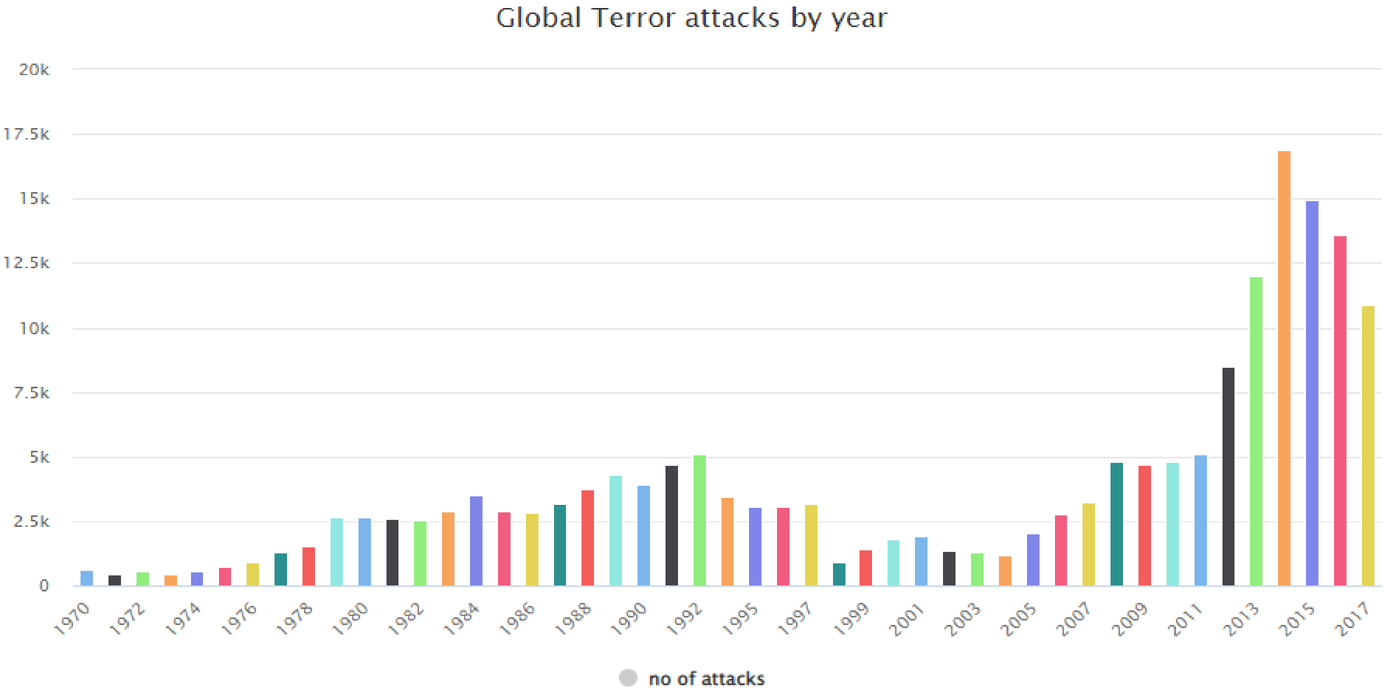


Figure 6. Terrorist Activities Each Year

Summarizing all the terrorist attacks over the years can provide us an idea about how terrorism has evolved and what rate has it impacted the world each year. Figure 6 shows data from 1970 to 2016 for the total number of attacks happened each year. Terrorist attacks were quite low in numbers in the decade of 1970. Terrorism then had a fairly rise in the 1980s and early 1990s and was considerably low in the next decade but then terrorism rose from early the 2000s topping the charts like never before in the history. Hostile environment and global tension have increased because of the number of attacks in recent years. This observation can help investigate factors that adversely impacted the sudden rise in number of attacks.

3.3.5 Total No of Attacks by Countries

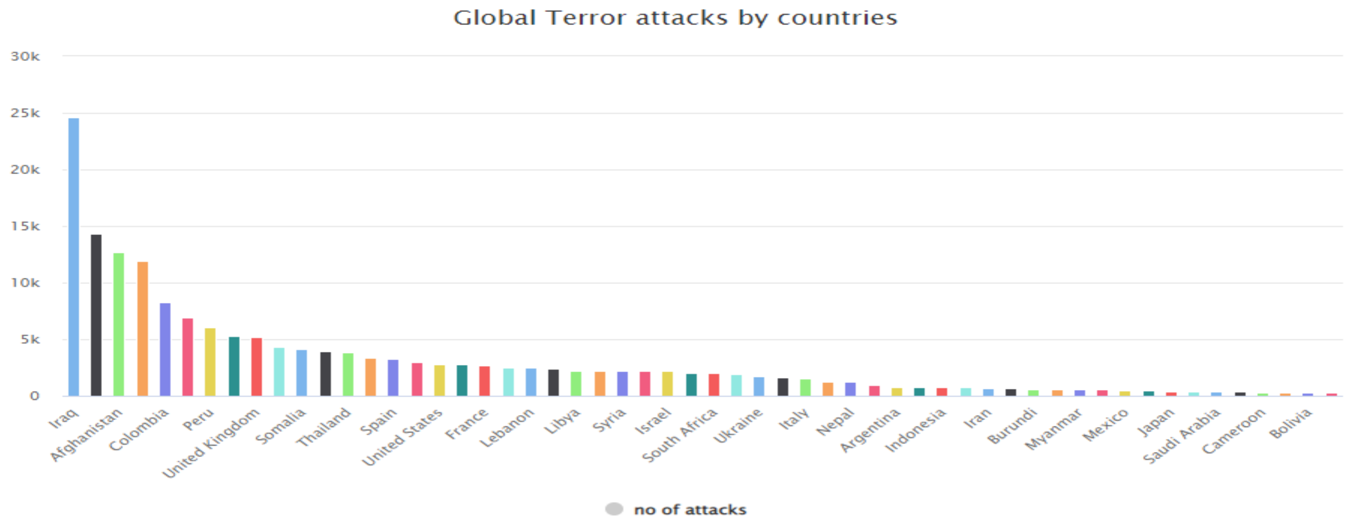


Figure 7. Number of Attacks by each Countries

Iraq and Pakistan have the highest number of attacks followed by Afghanistan and India and least in Ireland.

3.3.6 Terrorist Attacks by Region

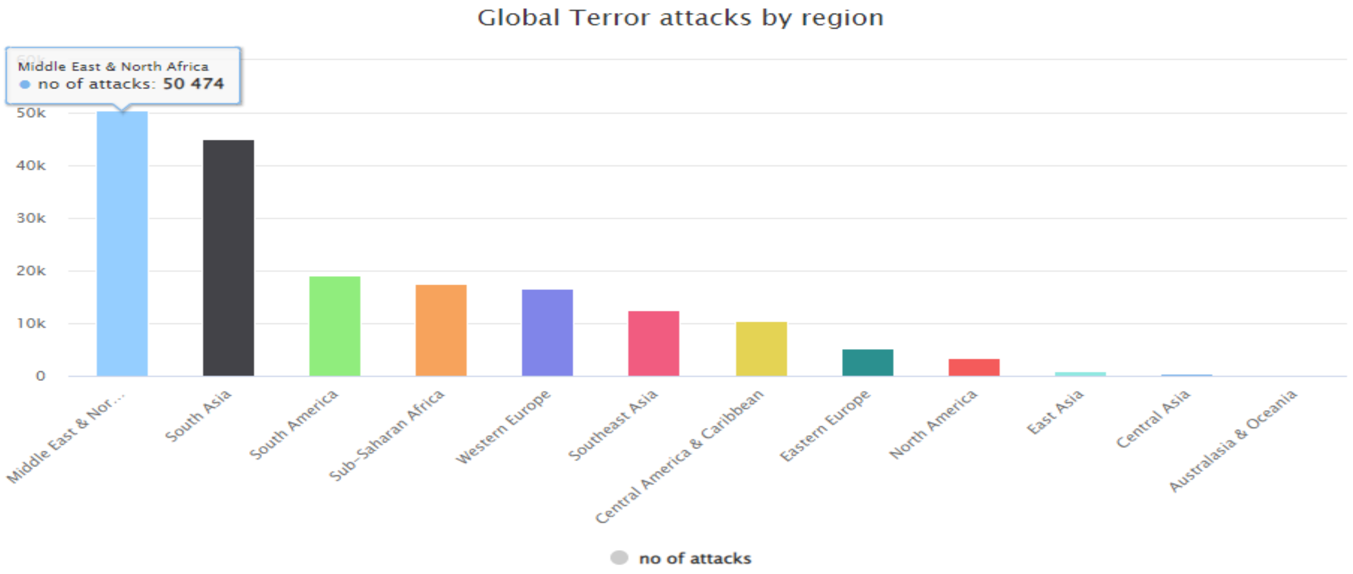


Figure 8. Terrorist Attacks in Each Region

Based on the geographic location of countries, they have been subcategorized into twelve regions to compare the rate of terrorism in each one of them as shown in Figure .Middle east and north Africa have the highest number of attacks followed by South Asia and South America. Terrorism here does not show an equal distribution among all regions. As a result, based on the number of attacks, different level of attention is required for each individual region.

3.3.7 Causalities in selected countries

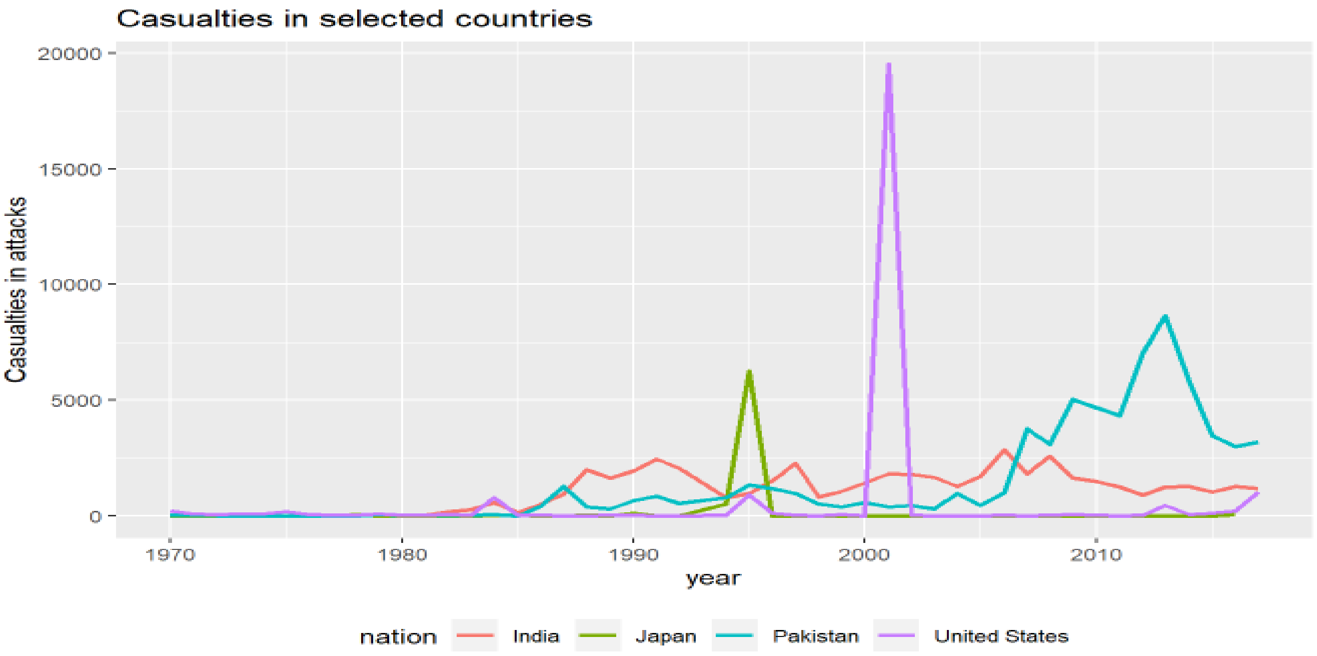


Figure 9. Causalities in selected countries

Casualities in attack is as high as in United states and steep rise in 2000.This association can help differentiate the counter-terrorism strategies in different countries resulting in a different number of kills for the same number of attacks. This comparison among countries can be taken into considerations while devising new tactics against terrorism.

3.3.8 Killed in USA by Terror Attack

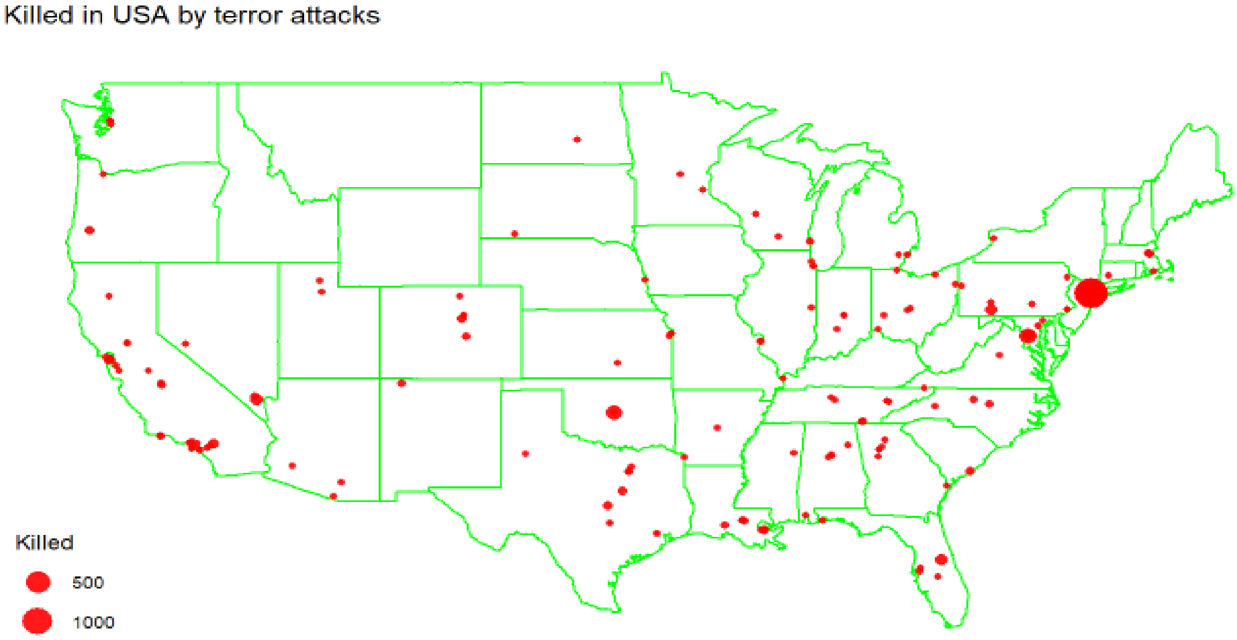


Figure 10. Top Affected Countrie

Figure 10 shows the size of the marked circle activity depends on the total number of causalities that happened in that attack. Bigger the circle more the casualty happened. We can get information of that specific event along with country, city,states and number of people wounded in that attack and bigger the circle it clearly represent large number people have been wounded by means of any type of attack.

3.3.9 Killed in USA by Terror Attack

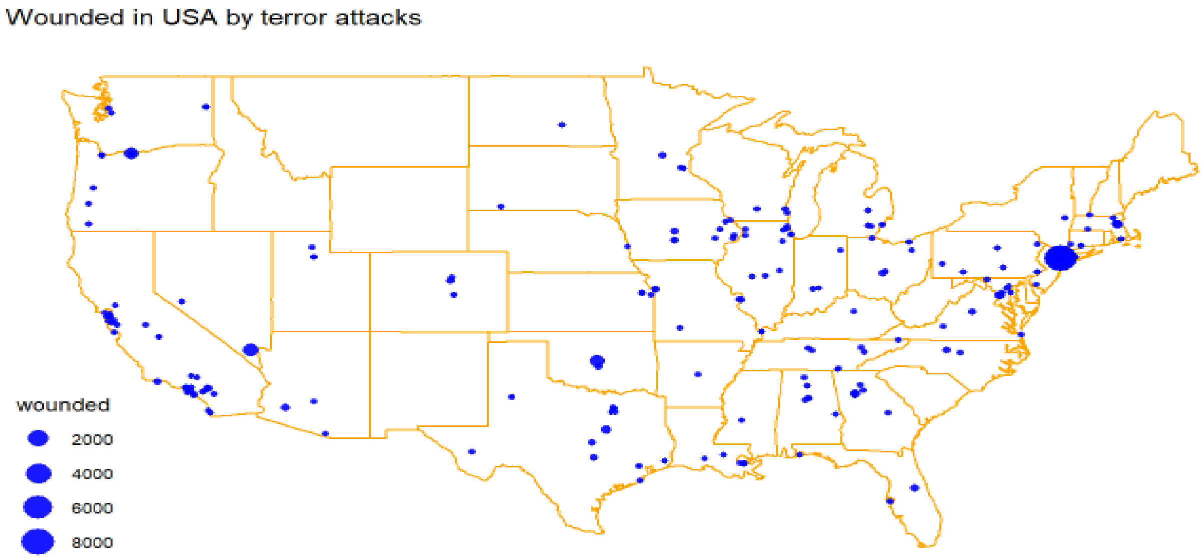
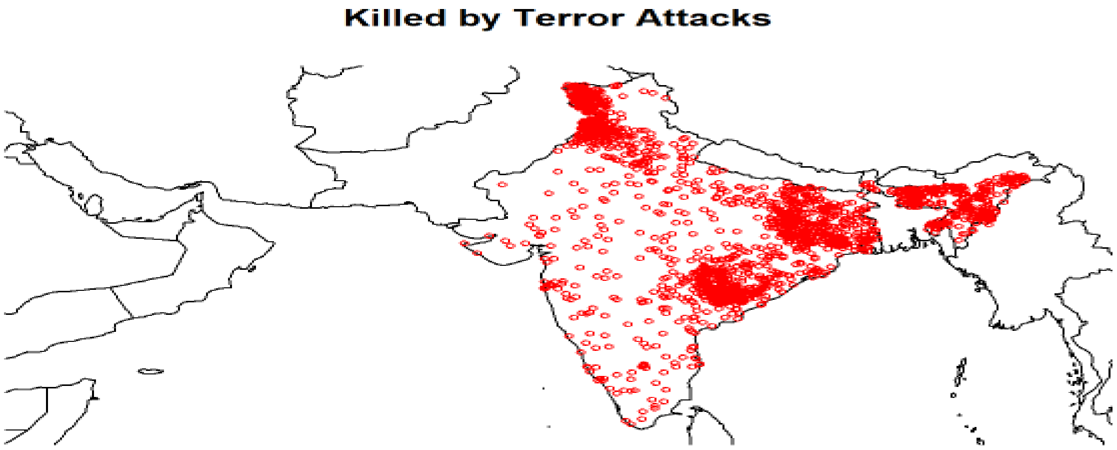


Figure 11. Active Terrorist Groups

From the figure we can see that We can see that North Eastern States and Jammu and Kashmir along the LOC have been high percent of wounded by terror attacks.

3.3.10: Killed by terror attack in India



From the attack we can clearly see that northern zone and eastern zone has large percentage of killed by terror attack followed by north eastern zone.

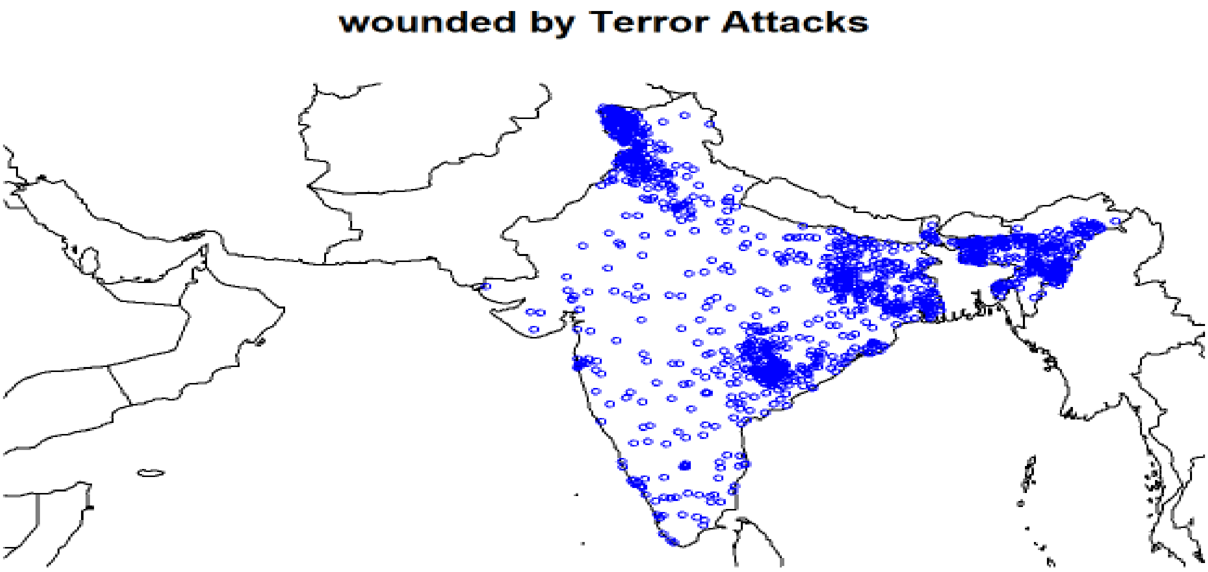


Figure 14. Regions in India wounded by Terror Attack

We can see that North Eastern States and Jammu and Kashmir along the LOC have been high percent of wounded by terror attacks.

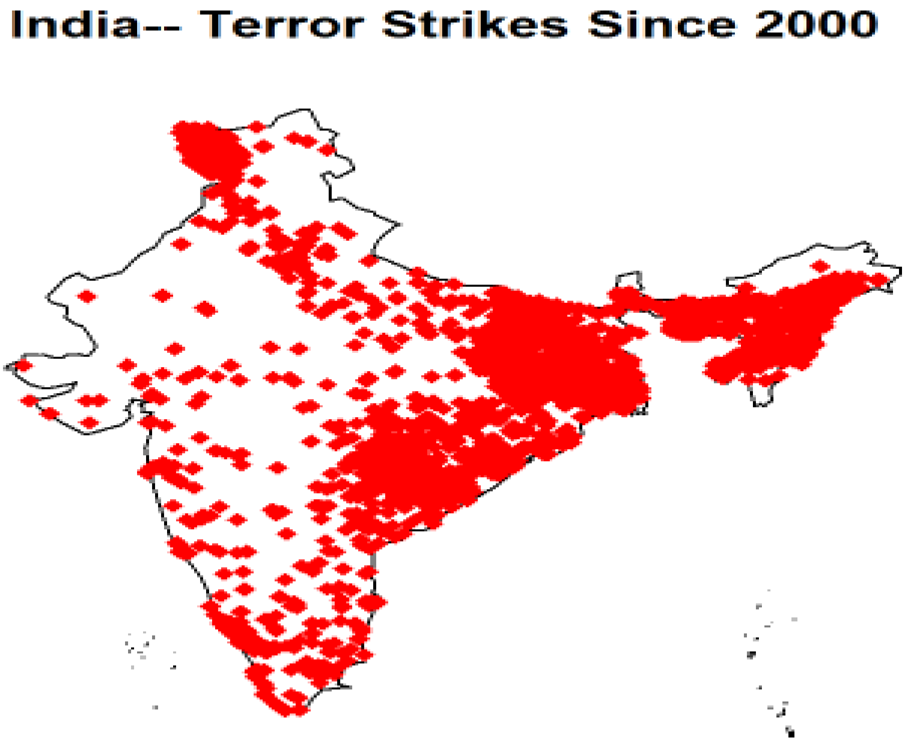


Figure 15. Terror Attack In India Since 2000

We can see that North Eastern States and Jammu and Kashmir along the LOC have been highly terror infested.

PROMINENTS WEAPONS USED:

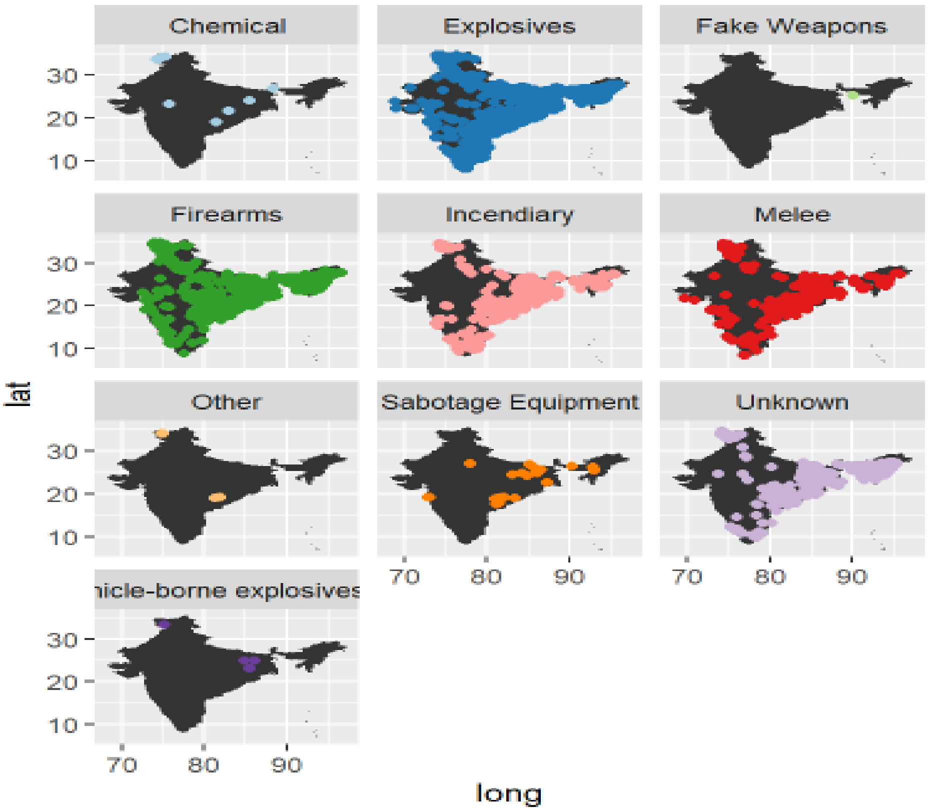


Figure 16. Weapons Used

Explosives and Firearms are used all over the country in terrorist attacks. J&K along LOC is most susceptible to be attacked by most kinds of weapons. Attacks targeting people rather than property (Melee) are more common along West Bengal and Orissa (eastern part of India)

What Kind of People are Targeted By the Terror Outfits?

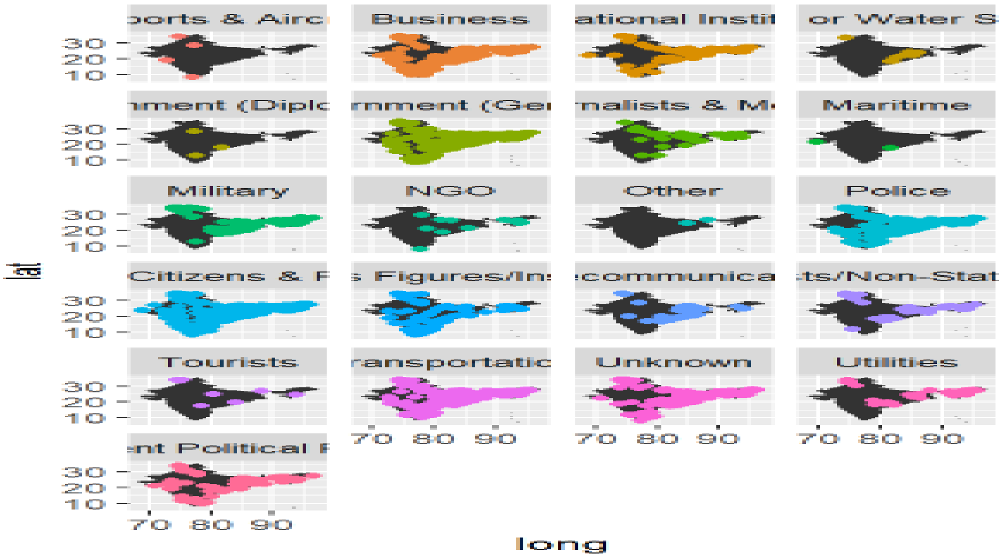


Figure 16. What Kind of People Are Attack.

Along the Eastern India, Individuals or Organisation Involved in Commercial Activity for their livelihood has been attacked. Orissa, Delhi and Bengaluru embassy has been attacked. Telecommunications has been attacked in Bihar, West Bengal and Orissa. As expected, Military is attacked in Jammu Kashmir and in North Eastern India.

Overall trend in each region

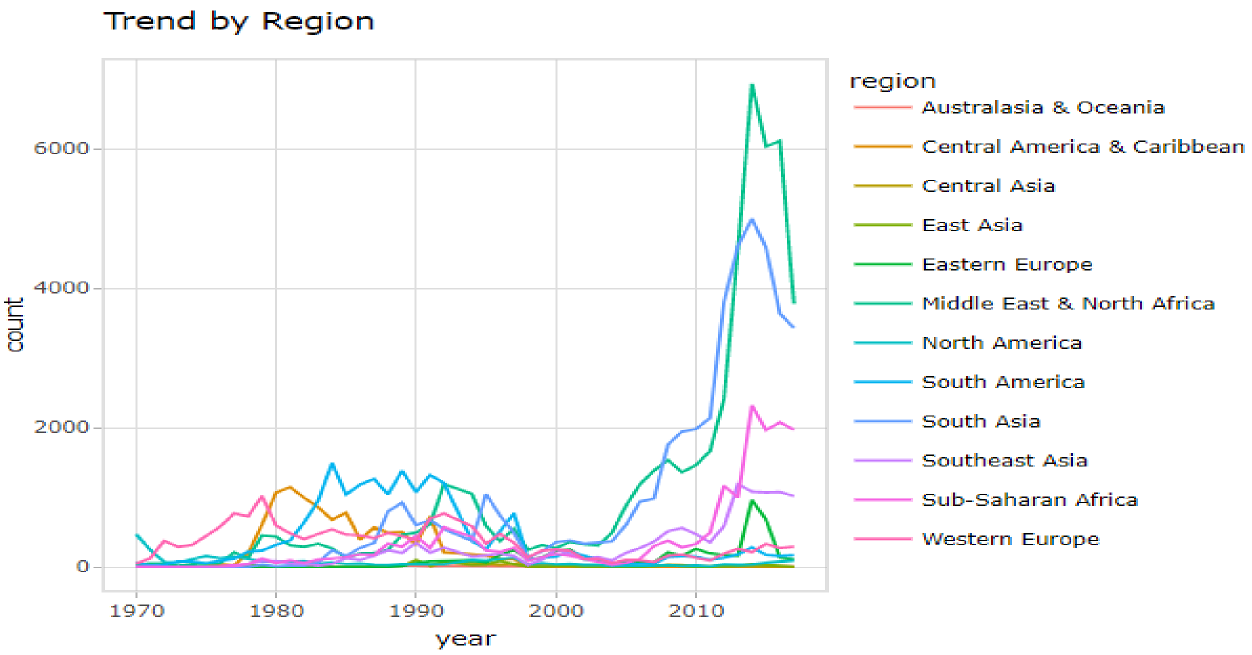


Figure 17. Overall Trend in Each Region

Hovering over the plot to see region label Middle East & North Africa and South Asia are the regions mainly responsible for the spike in data.

Events & num. of kills by region

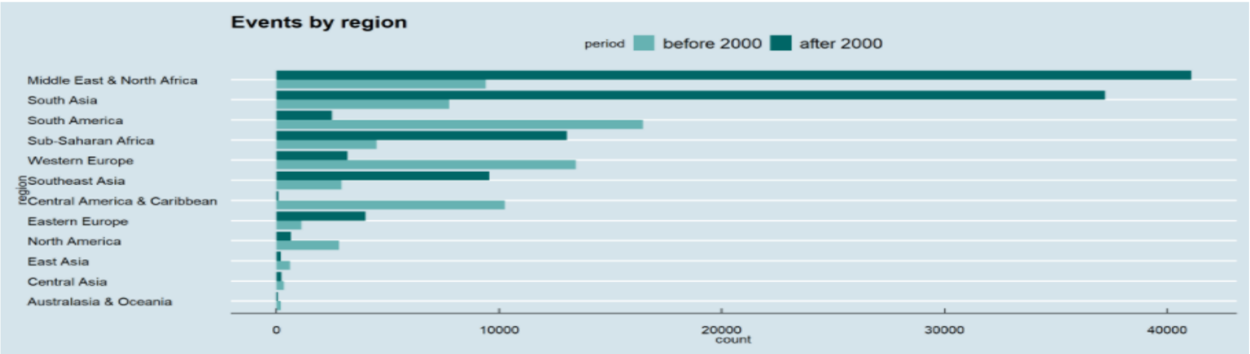


Figure 18. Event by Each Region

The region with the most terrorist attack bacame “Middle East & North Africa” after 2000. (“South America” before 2000).“South Asia” saw the largest increase in terrorism since the 70s.

Number of deaths and number of events

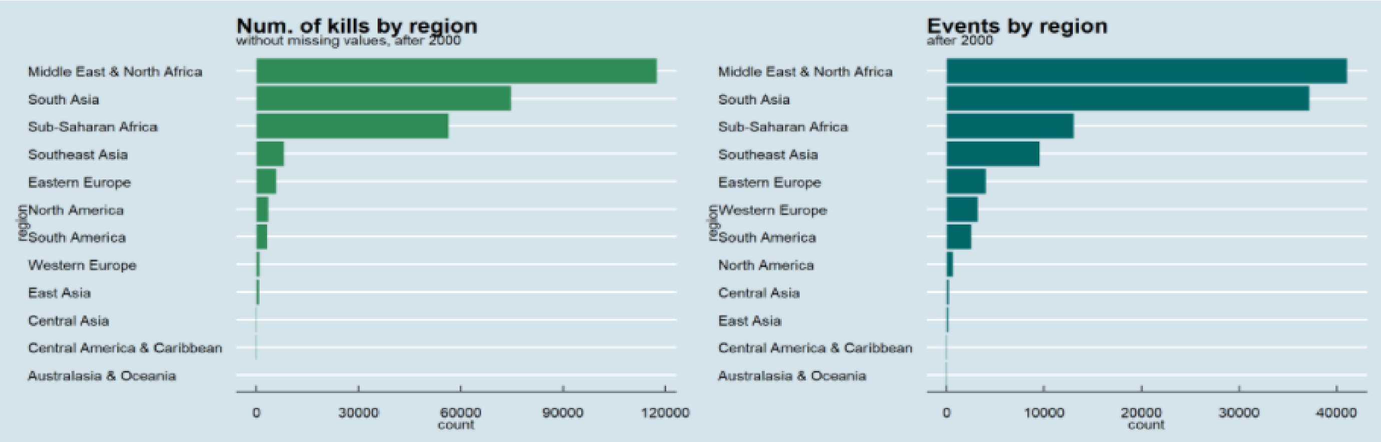


Figure 18. Events and Kills by Region

- South Asia has the largest num. of kills (other than “Sub-Saharan Africa”, “Middle East & North Africa”) despite the missing values.
- North America has higher number of kills than Western Europe and South America, even though there is less attacks.

Events by country

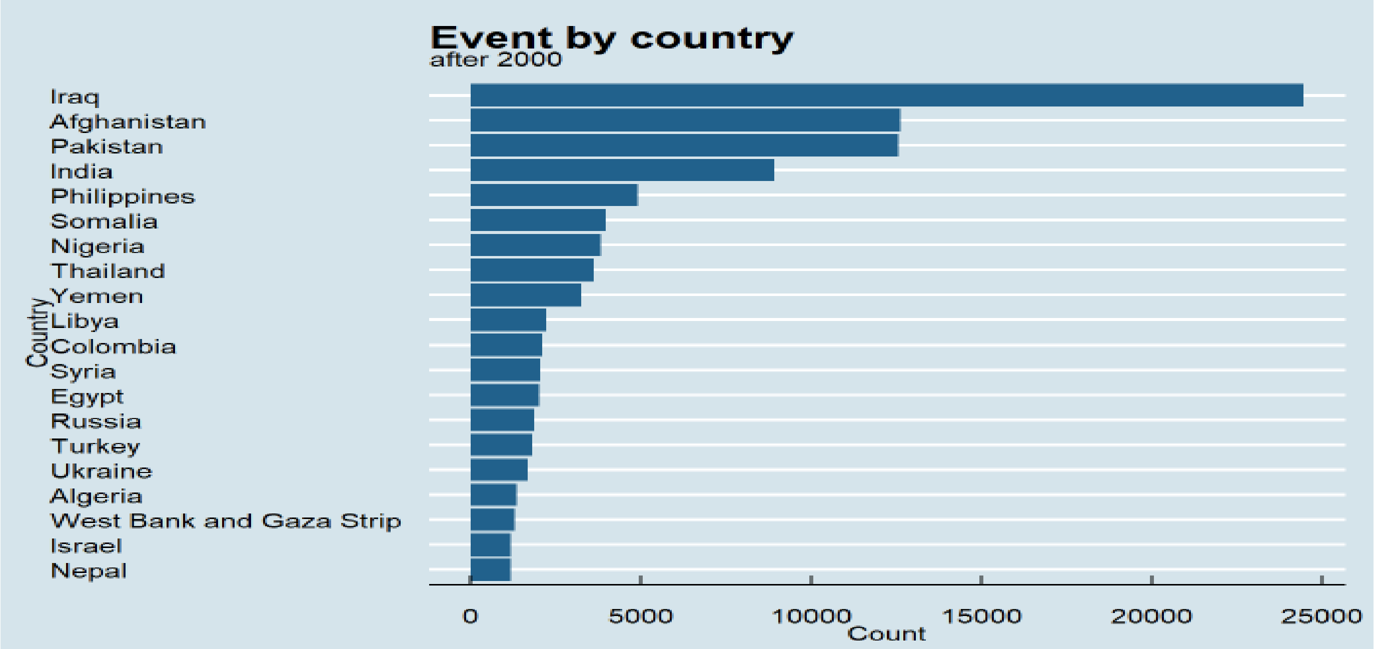


Figure 19. Events by Country

Iraq ,Pakistan , Afghanistan has maximum number of Events.

Suicide attack

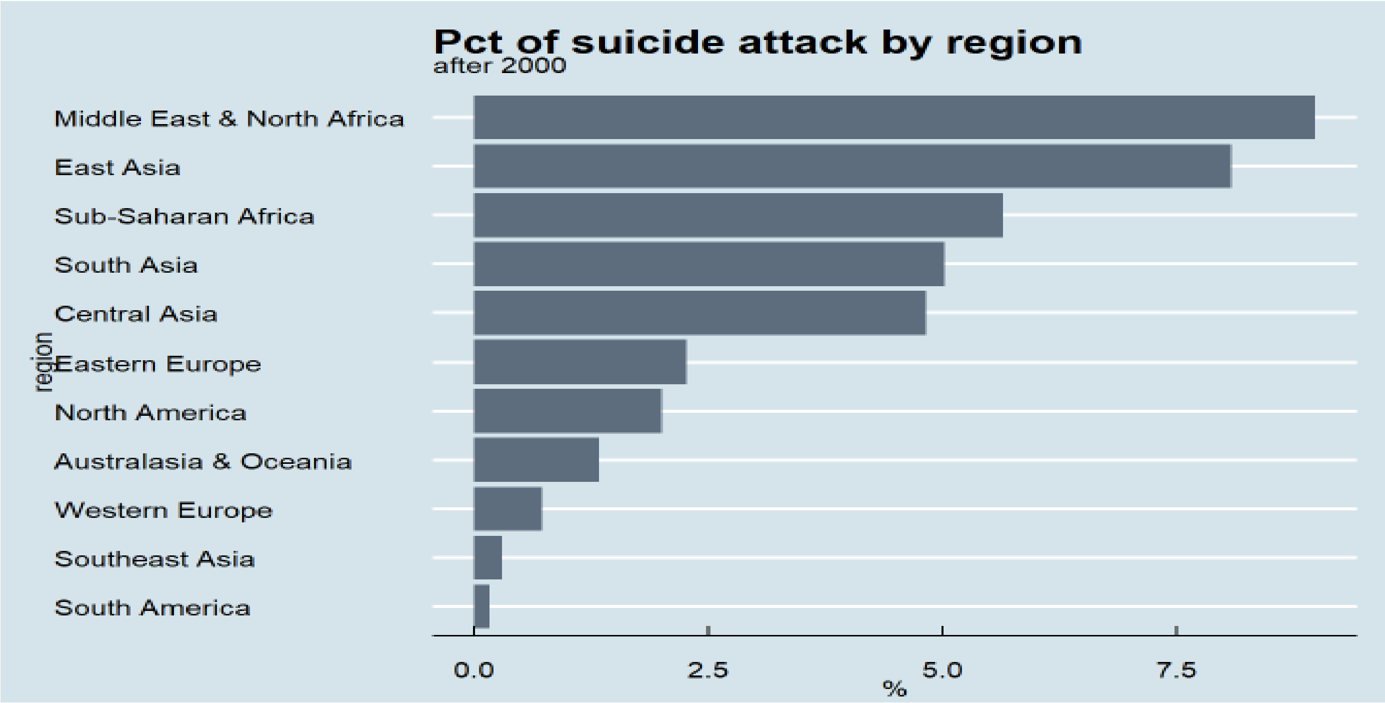


Figure 20. Attacks by region

Groups, attacks, and suicide

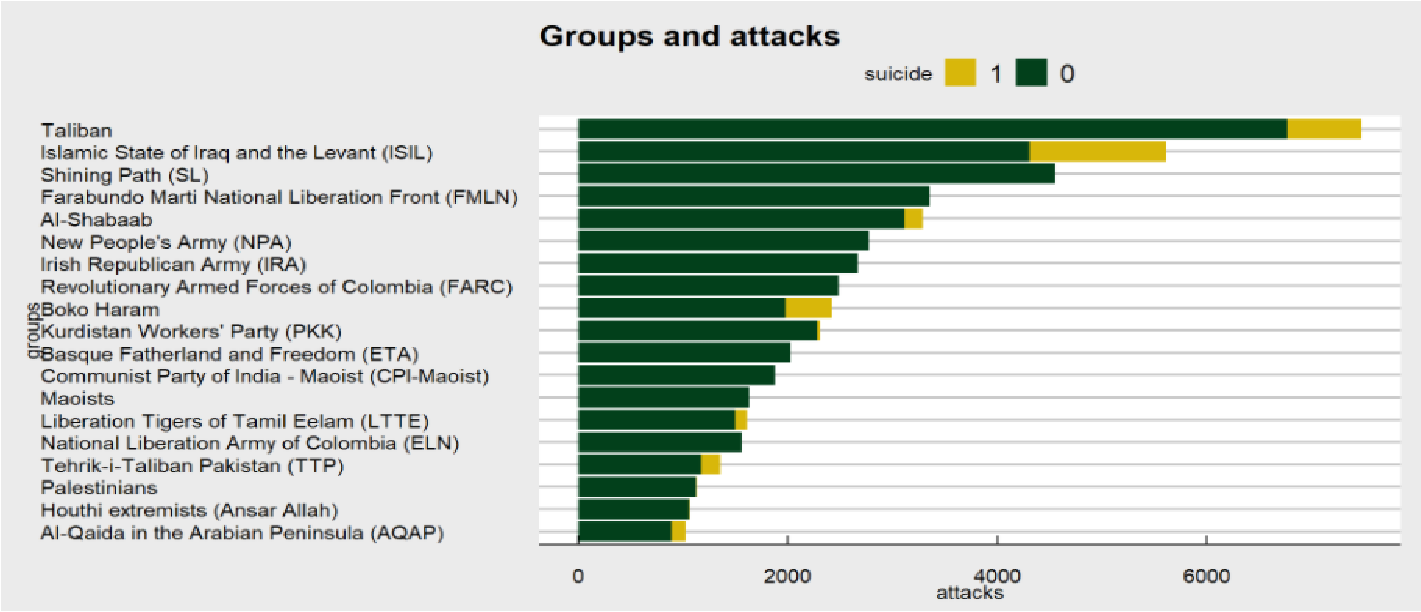


Figure 20. Groups, Attacks and Suicide by region

Attack type by group

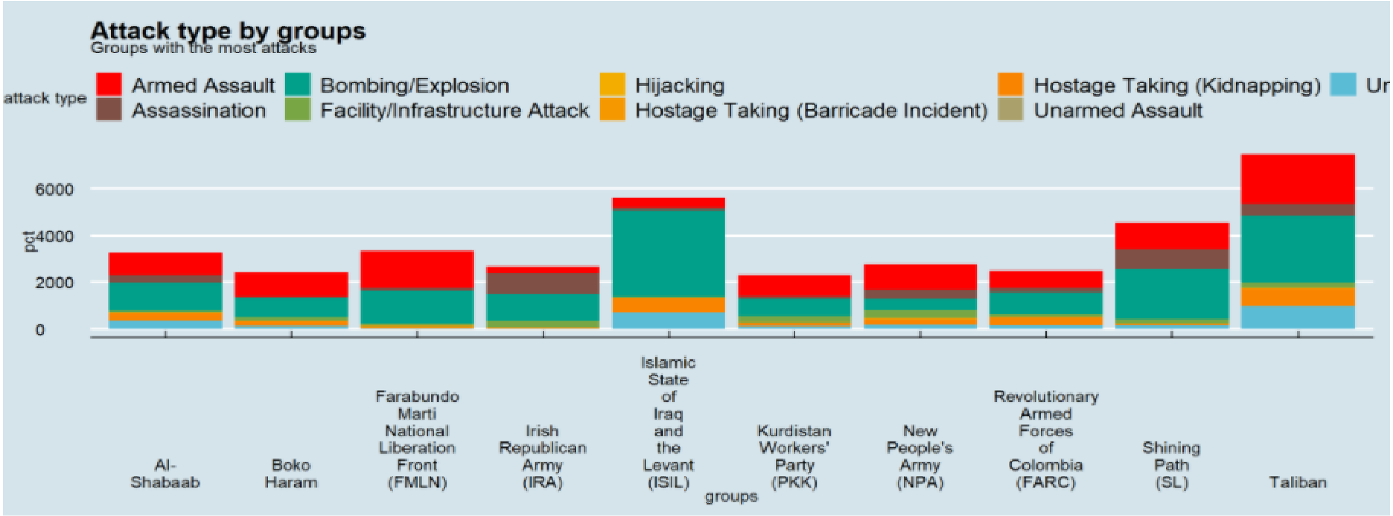


Figure 21. Attack Type by Group

Different groups might prefer different types of attack method. There are 3537 groups in the data. We'll look at the groups with the most attacks. Armed assault is common in most groups except IRA which prefers assassination next to bombing. Bombing is the most used attack type by ISIL.34% of ISIL's bombing attack is suicide attack.

Number of death by attack type and region

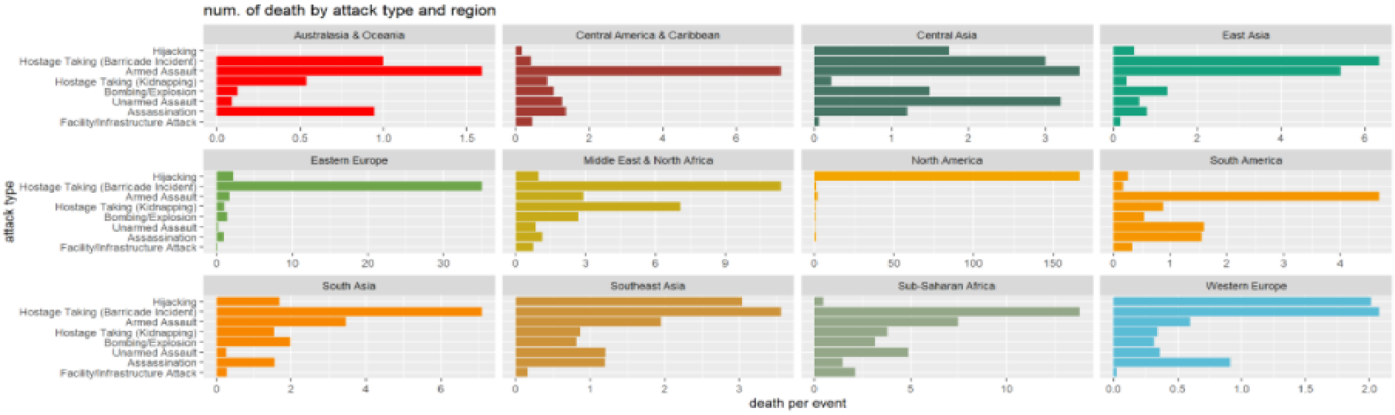


Figure 22 .Number of Attack Type by Death and Region

Types of attack that cause the most death/attack is drastically different from region to region. Bombing (to my surprise) isn't responsible for the most death/attack. Instead it's armed assault and hostage taking in most region. Hostage taking has the most death/attack in East Asia, Eastern Europe, Middle East & North Africa, South Asia, Southeast Asia, Sub-Saharan Africa and Western Europe. North America's extreme data reflects 9/11 attacks on 2001, with nearly 3,000 recorded deaths in 4 attacks.

APPENDIX

Implementation / Code

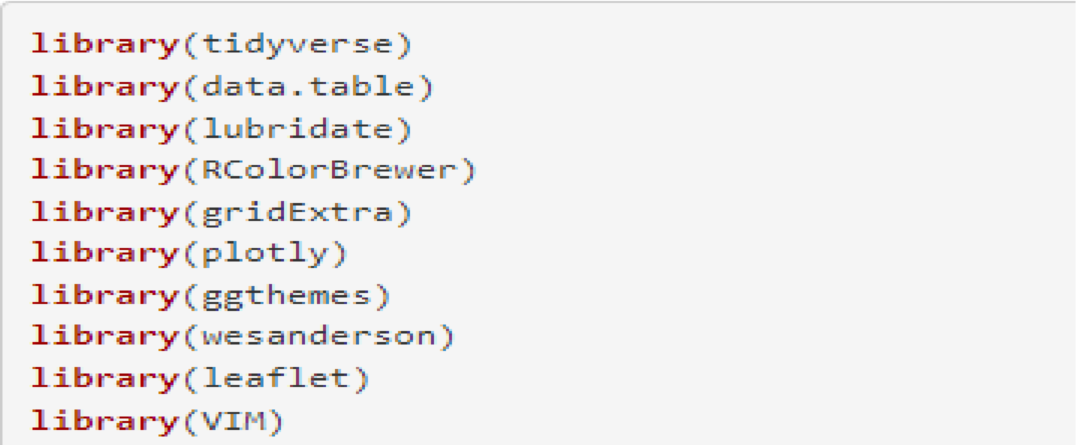
This chapter discusses about the code and implementation phase of the project. All scripts are written in Rstudio. Primary source of data is START dataset. All operations are performed on START dataset to generate visualizations.

Experimental setup:-

Most of the operations on the dataset are done by R Studio. R is used for data preprocessing, data modeling, analyses, and visualization. Anaconda is used as an open-source python distribution for handling R based dependencies and provide a environment for code development. Visual design mostly consists of analyses and visualization techniques to construct different graphics representing the end results in an easy-to-interpret format.Data preprocessing methodology helps in converting this raw data into a more meaningful, focused, interpretable and readable format and it helps us to resolve the discrepancies. Data cleaning is a process of filling missing values, removing outliers and handle inconsistencies in data. Data integration in this step, conflicts among data are resolved. Data transformation: Here data aggregation, generalization, and normalization are performed. Dataset has multiple target/victim subtypes. All those subtypes were aggregated to represent one value by summation of all similar subtypes. This technique reduces the total number of attributes in the datasets.

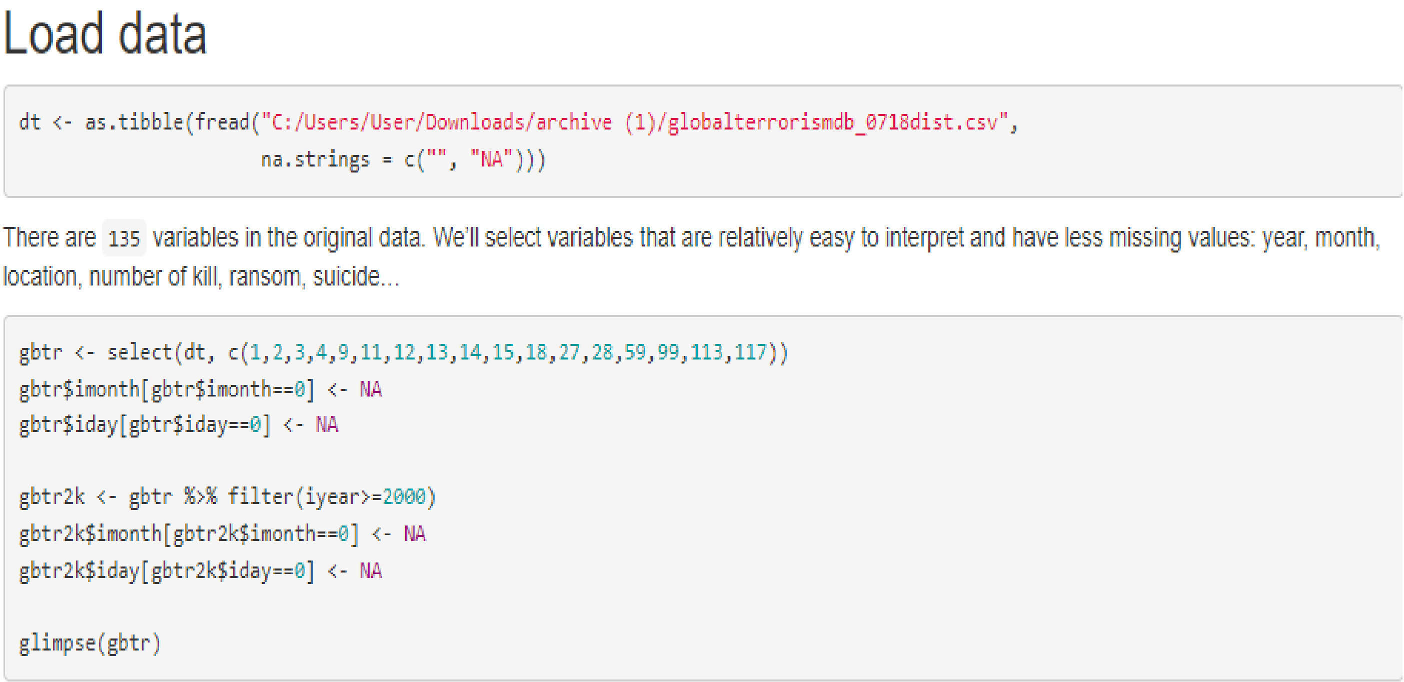
Data Preparation

LOADING LIBRARIES



```
library(tidyverse)
library(data.table)
library(lubridate)
library(RColorBrewer)
library(gridExtra)
library(plotly)
library(ggthemes)
library(wesanderson)
library(leaflet)
library(VIM)
```

LOAD DATA:



Load data

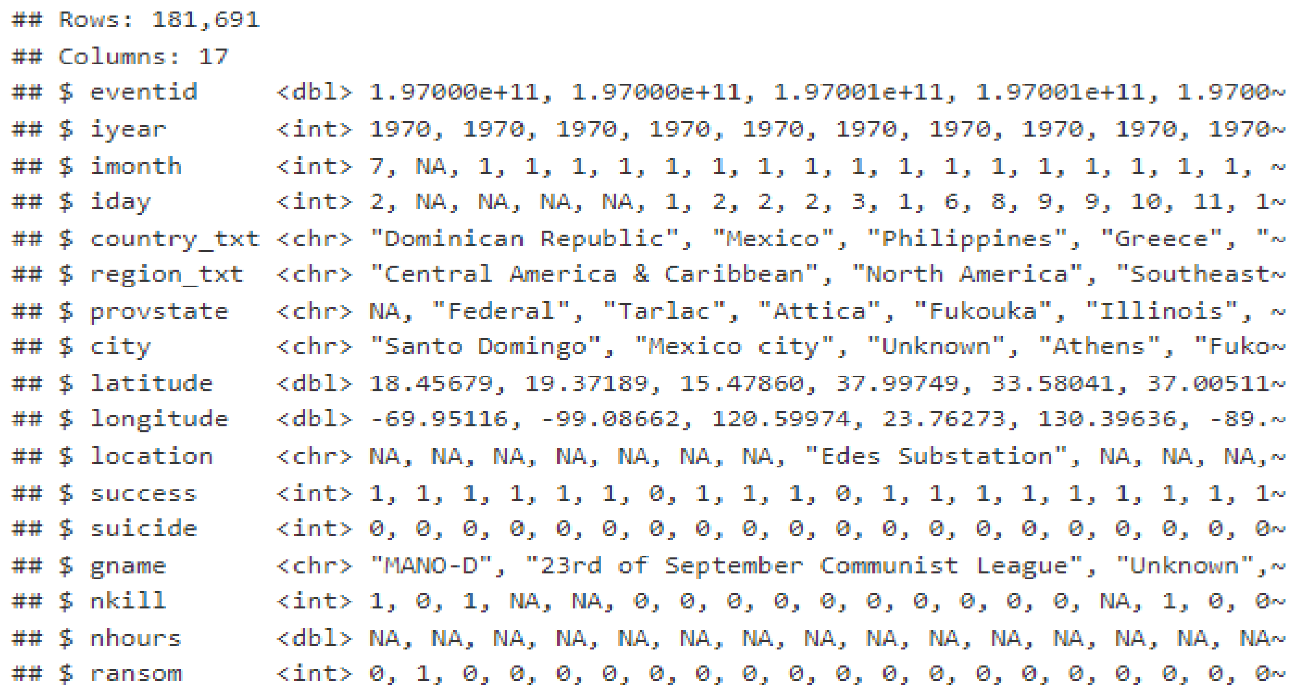
```
dt <- as.tibble(fread("C:/Users/User/Downloads/archive (1)/globalterrorismdb_0718dist.csv",
                      na.strings = c("", "NA")))
```

There are `135` variables in the original data. We'll select variables that are relatively easy to interpret and have less missing values: year, month, location, number of kill, ransom, suicide...

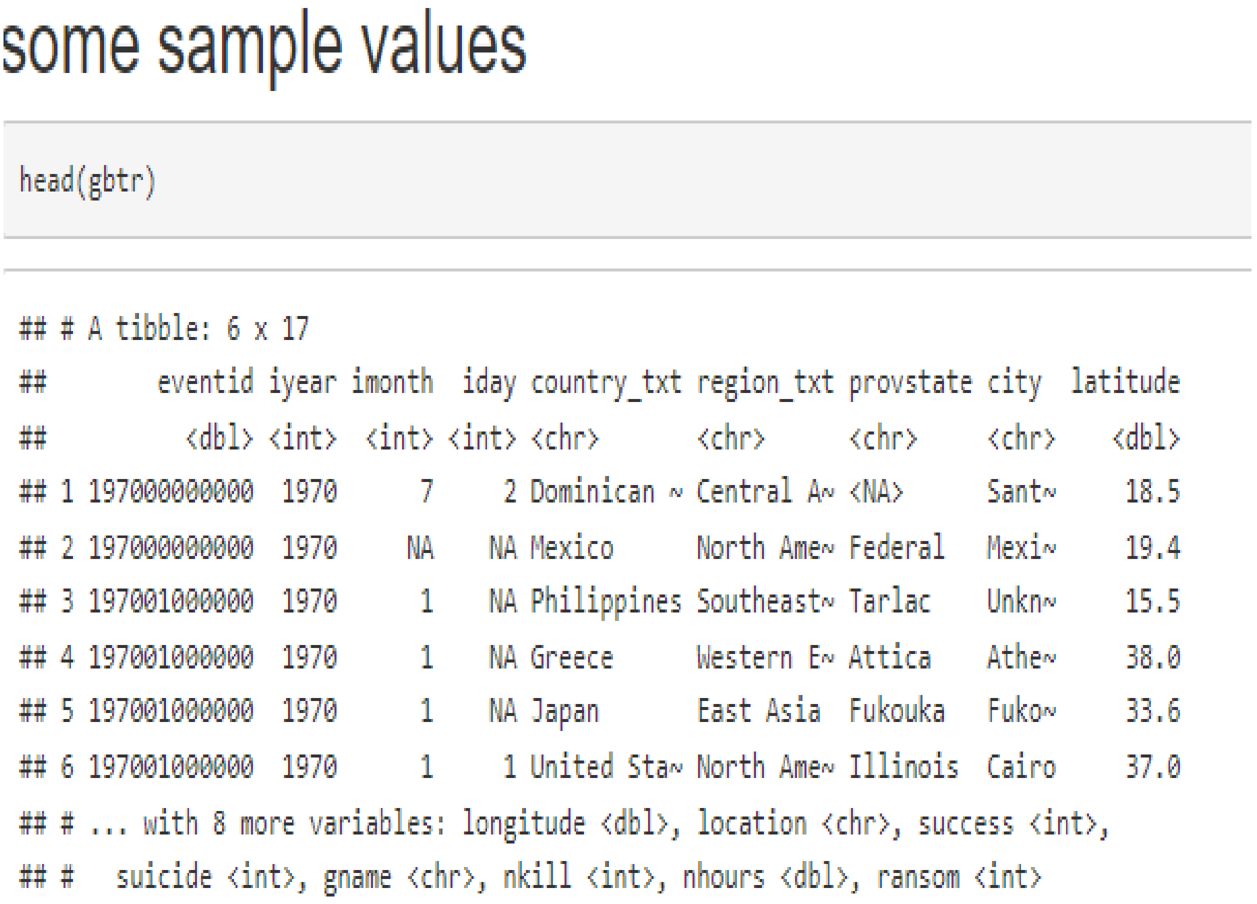
```
gbtr <- select(dt, c(1,2,3,4,9,11,12,13,14,15,18,27,28,59,99,113,117))
gbtr$imonth[gbtr$imonth==0] <- NA
gbtr$iday[gbtr$iday==0] <- NA

gbtr2k <- gbtr %>% filter(iyear>=2000)
gbtr2k$imonth[gbtr2k$imonth==0] <- NA
gbtr2k$iday[gbtr2k$iday==0] <- NA

glimpse(gbtr)
```



```
## Rows: 181,691
## Columns: 17
## $ eventid     <dbl> 1.97000e+11, 1.97000e+11, 1.97001e+11, 1.97001e+11, 1.9700~
## $ iyear       <int> 1970, 1970, 1970, 1970, 1970, 1970, 1970, 1970, 1970, 1970~
## $ imonth      <int> 7, NA, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ iday        <int> 2, NA, NA, NA, NA, 1, 2, 2, 2, 3, 1, 6, 8, 9, 9, 10, 11, 1~
## $ country_txt <chr> "Dominican Republic", "Mexico", "Philippines", "Greece", "~
## $ region_txt  <chr> "Central America & Caribbean", "North America", "Southeast~
## $ provstate   <chr> NA, "Federal", "Tarlac", "Attica", "Fukouka", "Illinois", ~
## $ city        <chr> "Santo Domingo", "Mexico city", "Unknown", "Athens", "Fuko~
## $ latitude    <dbl> 18.45679, 19.37189, 15.47860, 37.99749, 33.58041, 37.00511~
## $ longitude   <dbl> -69.95116, -99.08662, 120.59974, 23.76273, 130.39636, -89.~
## $ location    <chr> NA, NA, NA, NA, NA, NA, NA, "Edes Substation", NA, NA, NA,~
## $ success     <int> 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ suicide     <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ gname       <chr> "MANO-D", "23rd of September Communist League", "Unknown",~
## $ nkill       <int> 1, 0, 1, NA, NA, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, NA, 1, 0, 0~
## $ nhours      <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ ransom      <int> 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
```

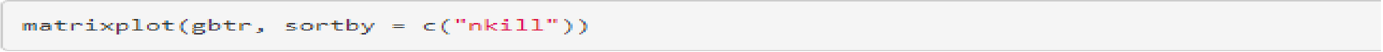


some sample values

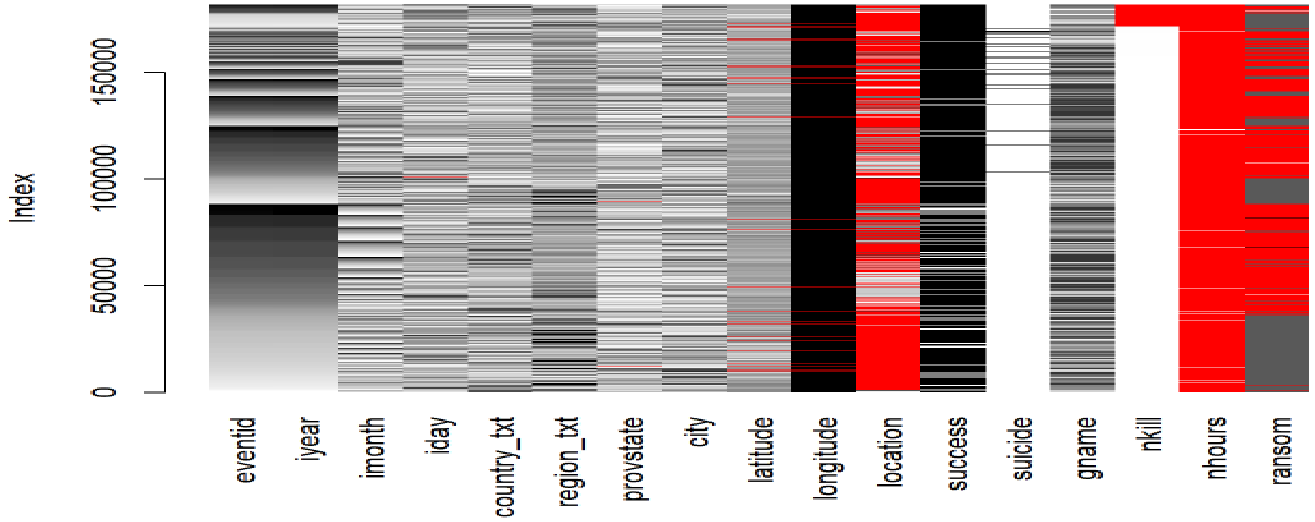
```
head(gbtr)
```

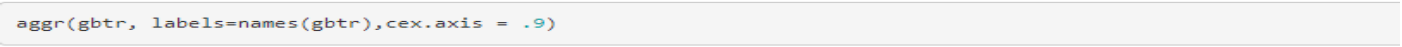
```
## # A tibble: 6 x 17
##        eventid iyear imonth  iday country_txt region_txt provstate city  latitude
##          <dbl> <int>  <int> <int> <chr>       <chr>      <chr>     <chr>    <dbl>
## 1 197000000000  1970      7     2 Dominican ~ Central A~ <NA>      Sant~     18.5
## 2 197000000000  1970     NA    NA Mexico      North Ame~ Federal   Mexi~     19.4
## 3 197001000000  1970      1    NA Philippines Southeast~ Tarlac    Unkn~     15.5
## 4 197001000000  1970      1    NA Greece      Western E~ Attica    Athe~     38.0
## 5 197001000000  1970      1    NA Japan       East Asia  Fukouka   Fuko~     33.6
## 6 197001000000  1970      1     1 United Sta~ North Ame~ Illinois  Cairo     37.0
## # ... with 8 more variables: longitude <dbl>, location <chr>, success <int>,
## #   suicide <int>, gname <chr>, nkill <int>, nhours <dbl>, ransom <int>
```

Visualization of missing value

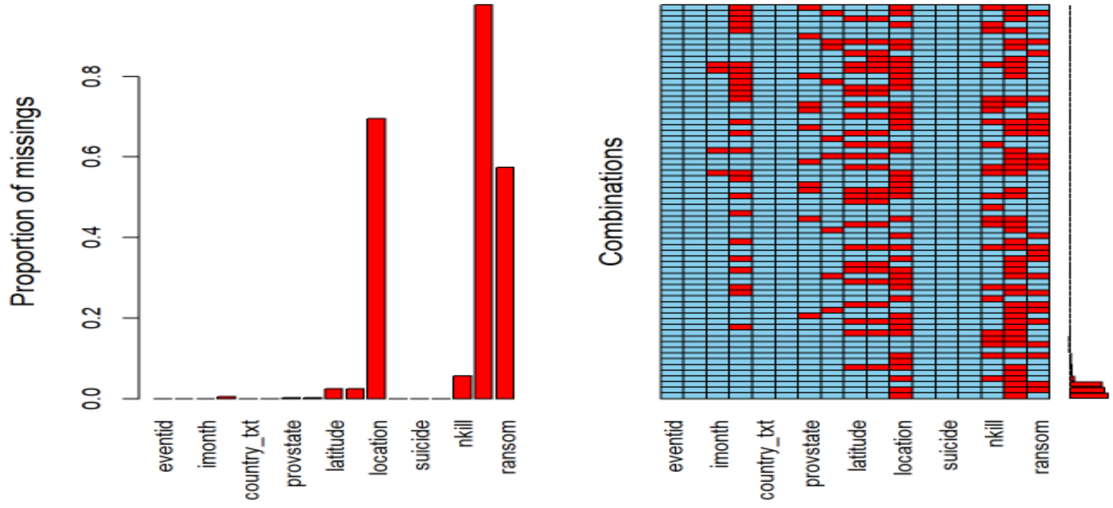


```
matrixplot(gbtr, sortby = c("nkill"))
```



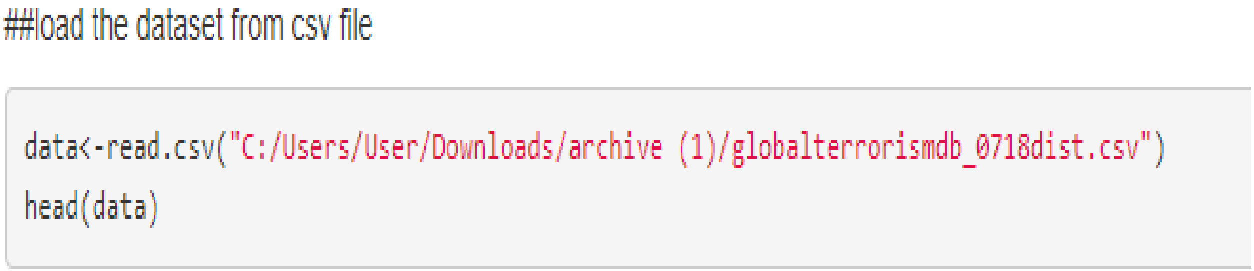


```
aggr(gbtr, labels=names(gbtr),cex.axis = .9)
```



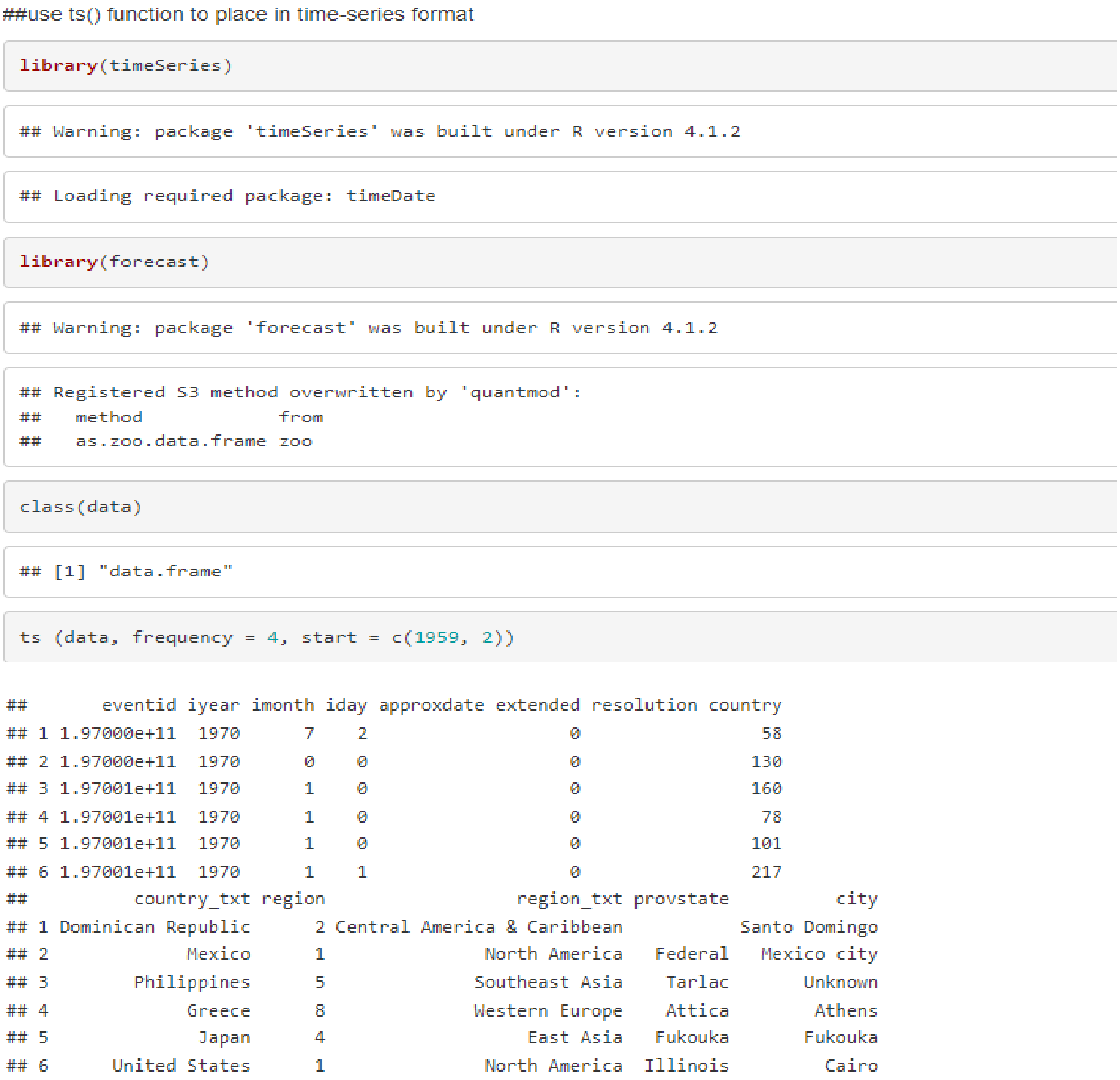
Variables such as location, nhours, and ransom has large number of missing values. EDA with thses variables will be avoided.

TIME SERIES ANALYSIS USING ARIMA-MODEL



##load the dataset from csv file

```
data<-read.csv("C:/Users/User/Downloads/archive (1)/globalterrorismdb_0718dist.csv")
head(data)
```



##use ts() function to place in time-series format

```
library(timeSeries)
```

```
## Warning: package 'timeSeries' was built under R version 4.1.2
```

```
## Loading required package: timeDate
```

```
library(forecast)
```

```
## Warning: package 'forecast' was built under R version 4.1.2
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
class(data)
```

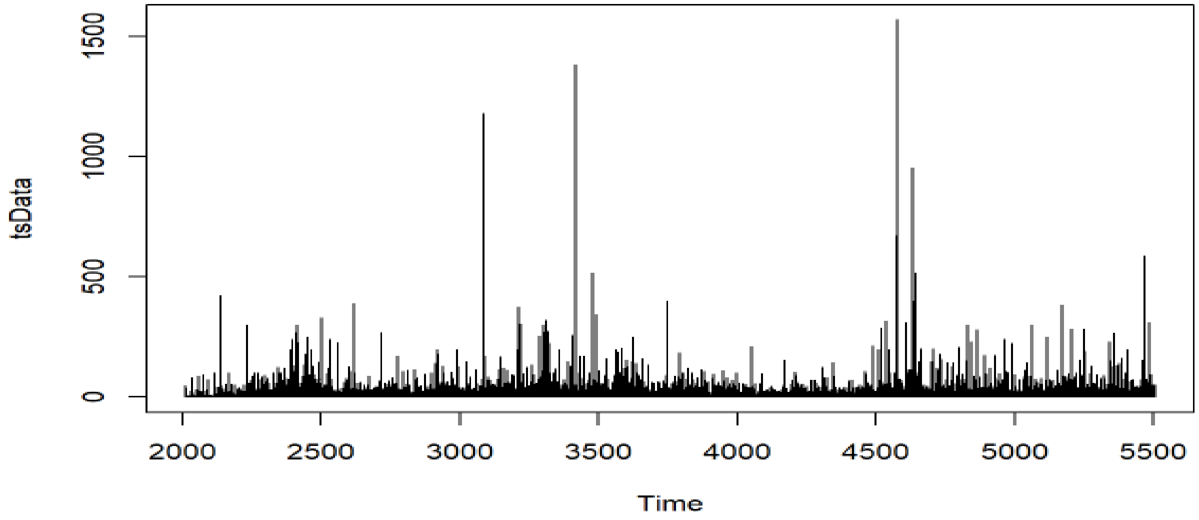
```
## [1] "data.frame"
```

```
ts (data, frequency = 4, start = c(1959, 2))
```

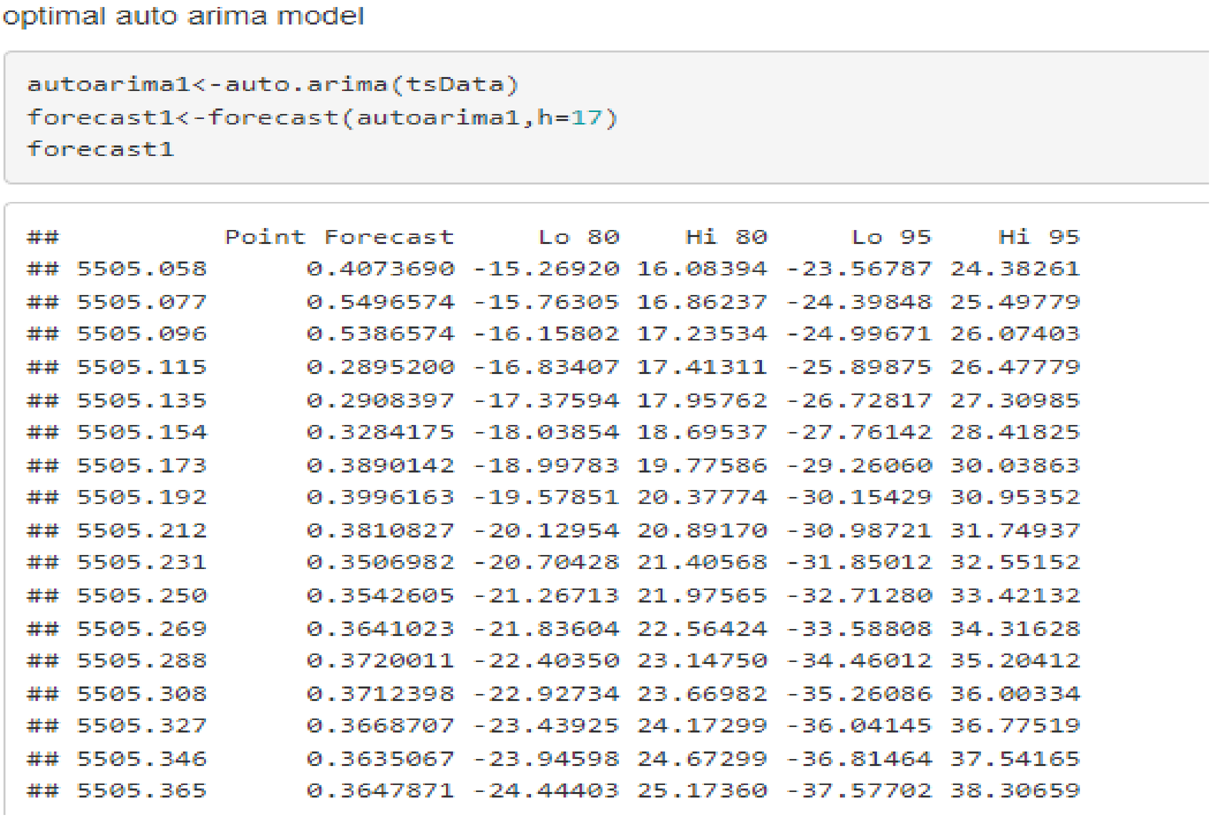
```
##       eventid iyear imonth iday approxdate extended resolution country
## 1 1.97000e+11  1970      7    2                   0                 58
## 2 1.97000e+11  1970      0    0                   0                130
## 3 1.97001e+11  1970      1    0                   0                160
## 4 1.97001e+11  1970      1    0                   0                 78
## 5 1.97001e+11  1970      1    0                   0                101
## 6 1.97001e+11  1970      1    1                   0                217
##            country_txt region                  region_txt provstate          city
## 1 Dominican Republic      2 Central America & Caribbean           Santo Domingo
## 2             Mexico      1               North America   Federal   Mexico city
## 3        Philippines      5              Southeast Asia    Tarlac       Unknown
## 4             Greece      8              Western Europe    Attica        Athens
## 5              Japan      4                   East Asia   Fukouka       Fukouka
## 6      United States      1               North America  Illinois         Cairo
```

##use timeseries formatted data

```
plot(tsData)
```



BUILDING ARIMA-MODEL AND VISUALISING IT



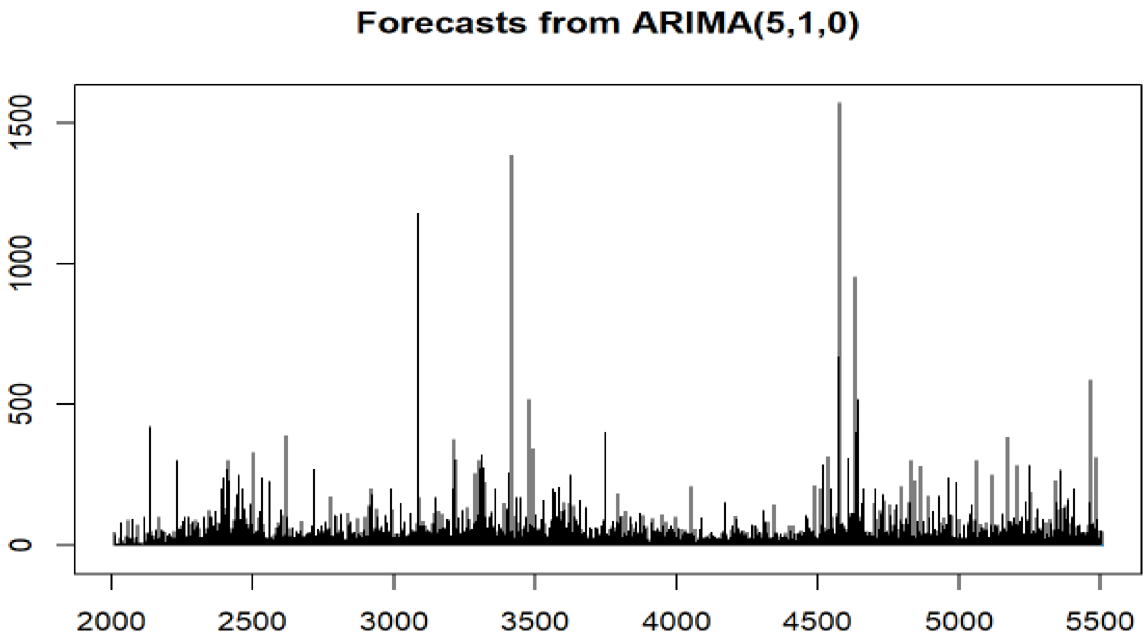
optimal auto arima model

```
autoarima1<-auto.arima(tsData)
forecast1<-forecast(autoarima1,h=17)
forecast1
```

```
##          Point Forecast     Lo 80    Hi 80     Lo 95    Hi 95
## 5505.058      0.4073690 -15.26920 16.08394 -23.56787 24.38261
## 5505.077      0.5496574 -15.76305 16.86237 -24.39848 25.49779
## 5505.096      0.5386574 -16.15802 17.23534 -24.99671 26.07403
## 5505.115      0.2895200 -16.83407 17.41311 -25.89875 26.47779
## 5505.135      0.2908397 -17.37594 17.95762 -26.72817 27.30985
## 5505.154      0.3284175 -18.03854 18.69537 -27.76142 28.41825
## 5505.173      0.3890142 -18.99783 19.77586 -29.26060 30.03863
## 5505.192      0.3996163 -19.57851 20.37774 -30.15429 30.95352
## 5505.212      0.3810827 -20.12954 20.89170 -30.98721 31.74937
## 5505.231      0.3506982 -20.70428 21.40568 -31.85012 32.55152
## 5505.250      0.3542605 -21.26713 21.97565 -32.71280 33.42132
## 5505.269      0.3641023 -21.83604 22.56424 -33.58808 34.31628
## 5505.288      0.3720011 -22.40350 23.14750 -34.46012 35.20412
## 5505.308      0.3712398 -22.92734 23.66982 -35.26086 36.00334
## 5505.327      0.3668707 -23.43925 24.17299 -36.04145 36.77519
## 5505.346      0.3635067 -23.94598 24.67299 -36.81464 37.54165
## 5505.365      0.3647871 -24.44403 25.17360 -37.57702 38.30659
```

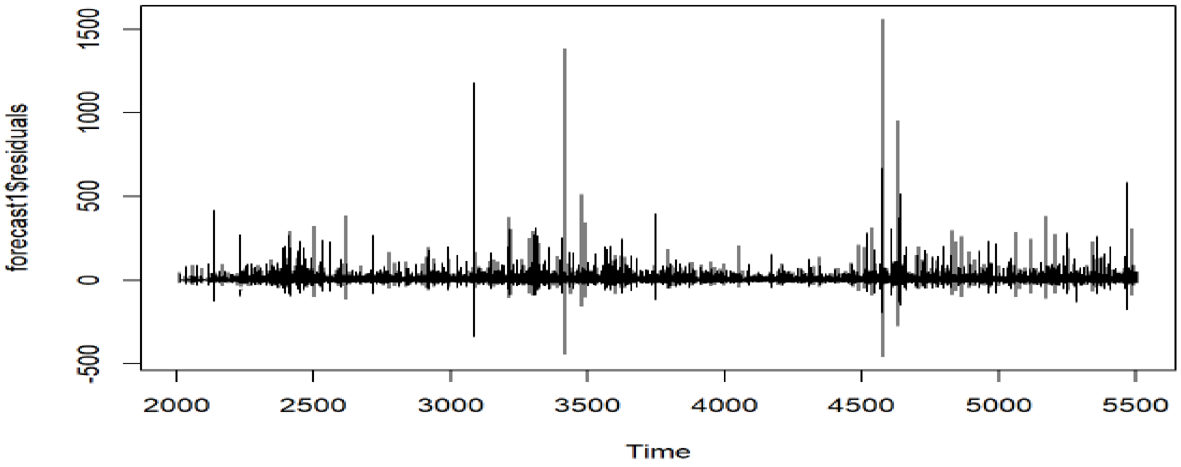
##plot forcasted data from auto-arima model

```
plot(forecast1)
```



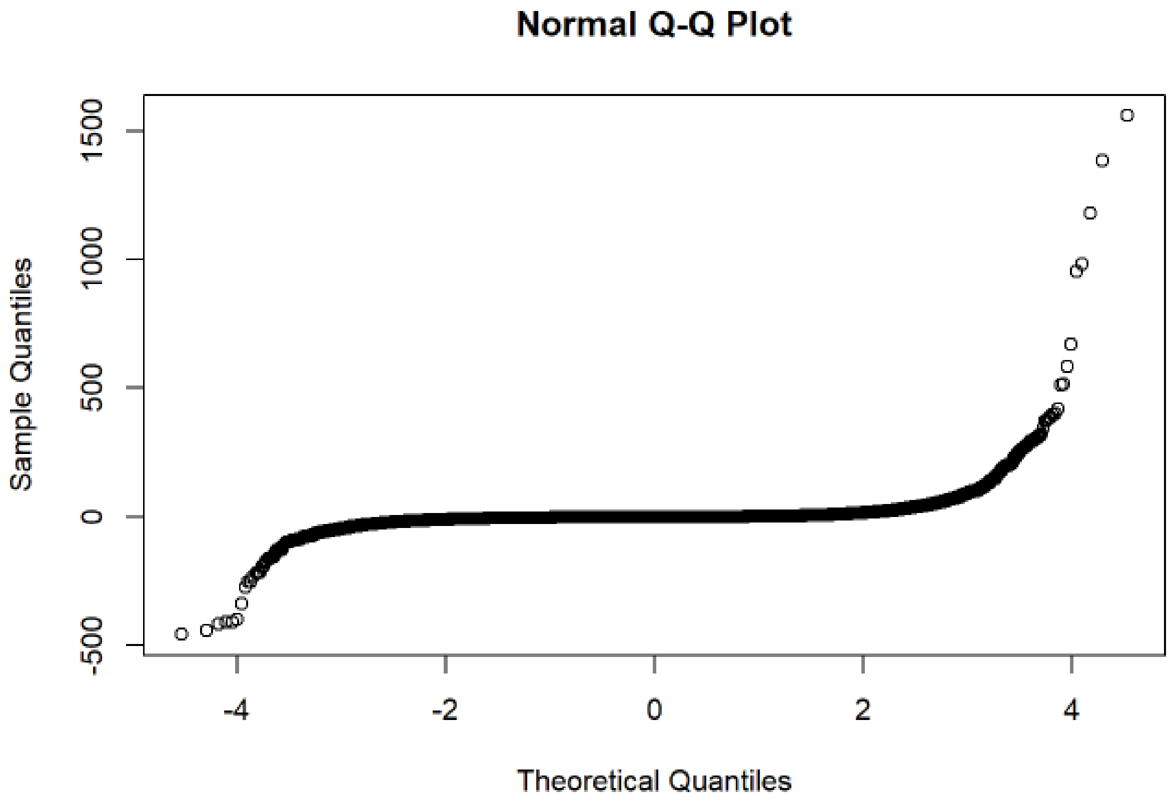
variance

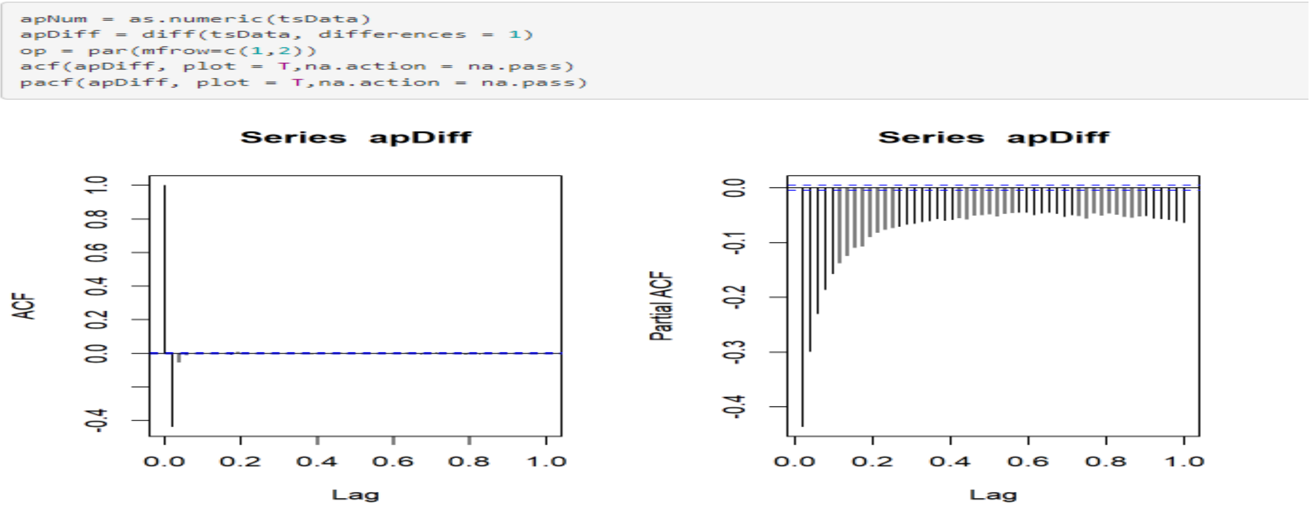
```
plot(forecast1$residuals)
```



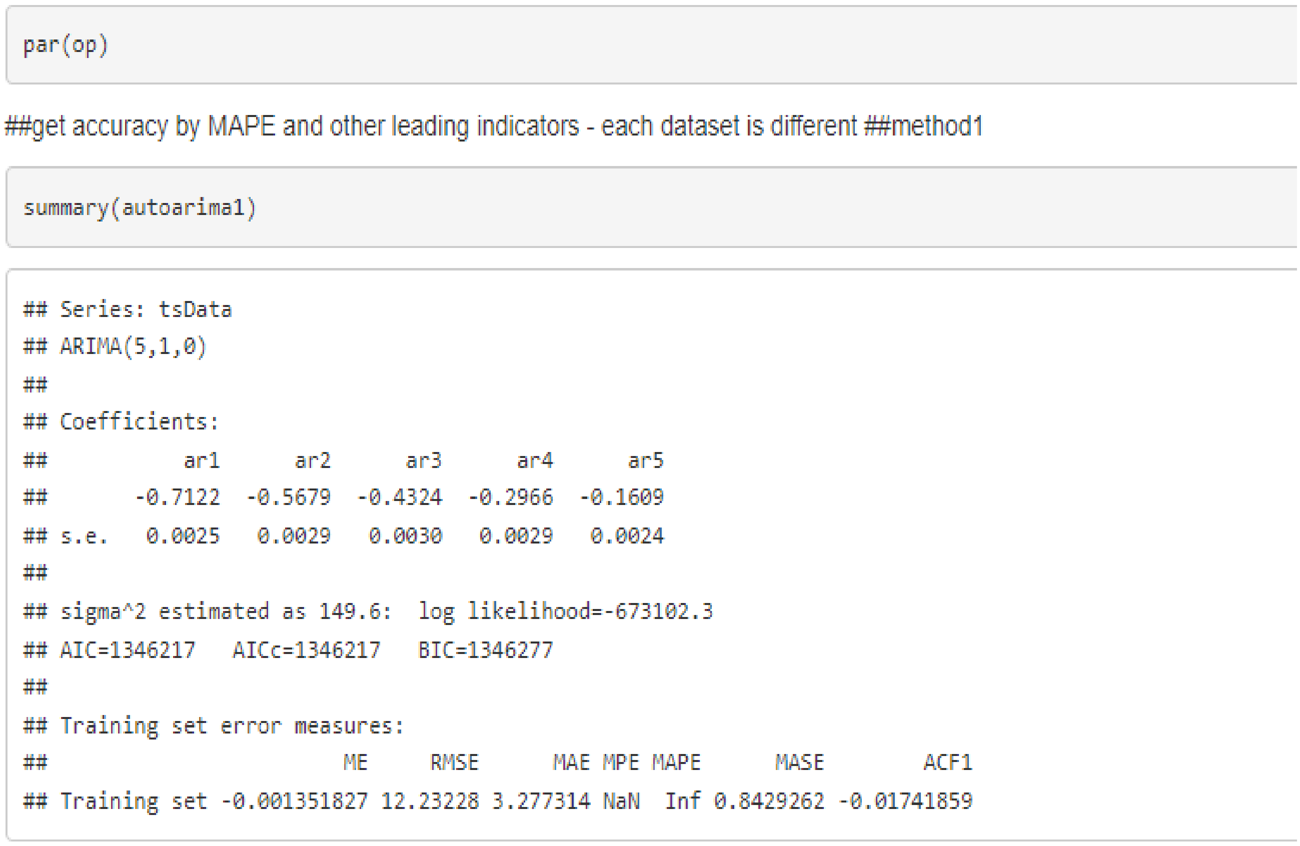
theoritical)

```
qqnorm(forecast1$residuals)
```





```
apNum = as.numeric(tsData)
apDiff = diff(tsData, differences = 1)
op = par(mfrow=c(1,2))
acf(apDiff, plot = T,na.action = na.pass)
pacf(apDiff, plot = T,na.action = na.pass)
```

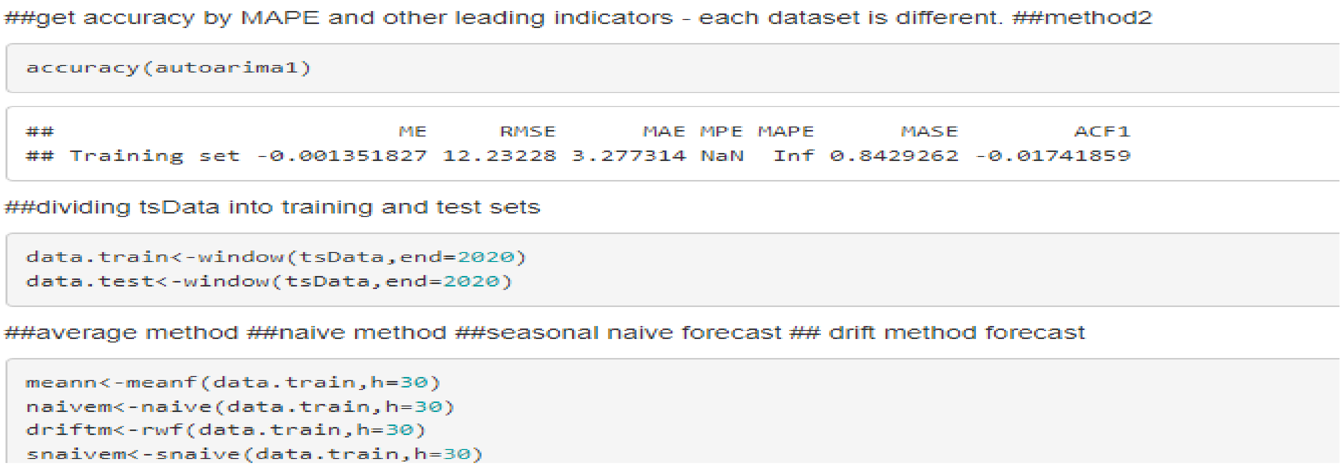


```
par(op)
```

##get accuracy by MAPE and other leading indicators - each dataset is different ##method1

```
summary(autoarima1)
```

```
## Series: tsData
## ARIMA(5,1,0)
##
## Coefficients:
##           ar1      ar2      ar3      ar4      ar5
##       -0.7122  -0.5679  -0.4324  -0.2966  -0.1609
## s.e.   0.0025   0.0029   0.0030   0.0029   0.0024
##
## sigma^2 estimated as 149.6:  log likelihood=-673102.3
## AIC=1346217   AICc=1346217   BIC=1346277
##
## Training set error measures:
##                        ME     RMSE      MAE MPE MAPE      MASE        ACF1
## Training set -0.001351827 12.23228 3.277314 NaN  Inf 0.8429262 -0.01741859
```



##get accuracy by MAPE and other leading indicators - each dataset is different. ##method2

```
accuracy(autoarima1)
```

```
##                        ME     RMSE      MAE MPE MAPE      MASE        ACF1
## Training set -0.001351827 12.23228 3.277314 NaN  Inf 0.8429262 -0.01741859
```

##dividing tsData into training and test sets

```
data.train<-window(tsData,end=2020)
data.test<-window(tsData,end=2020)
```

##average method ##naive method ##seasonal naive forecast ## drift method forecast

```
meann<-meanf(data.train,h=30)
naivem<-naive(data.train,h=30)
driftm<-rwf(data.train,h=30)
snaivem<-snaive(data.train,h=30)
```

##visualisation for all method

```
plot(meann,plot.conf=F,main="")
```

```
## Warning in plot.window(xlim, ylim, log, ...): "plot.conf" is not a graphical
## parameter
```

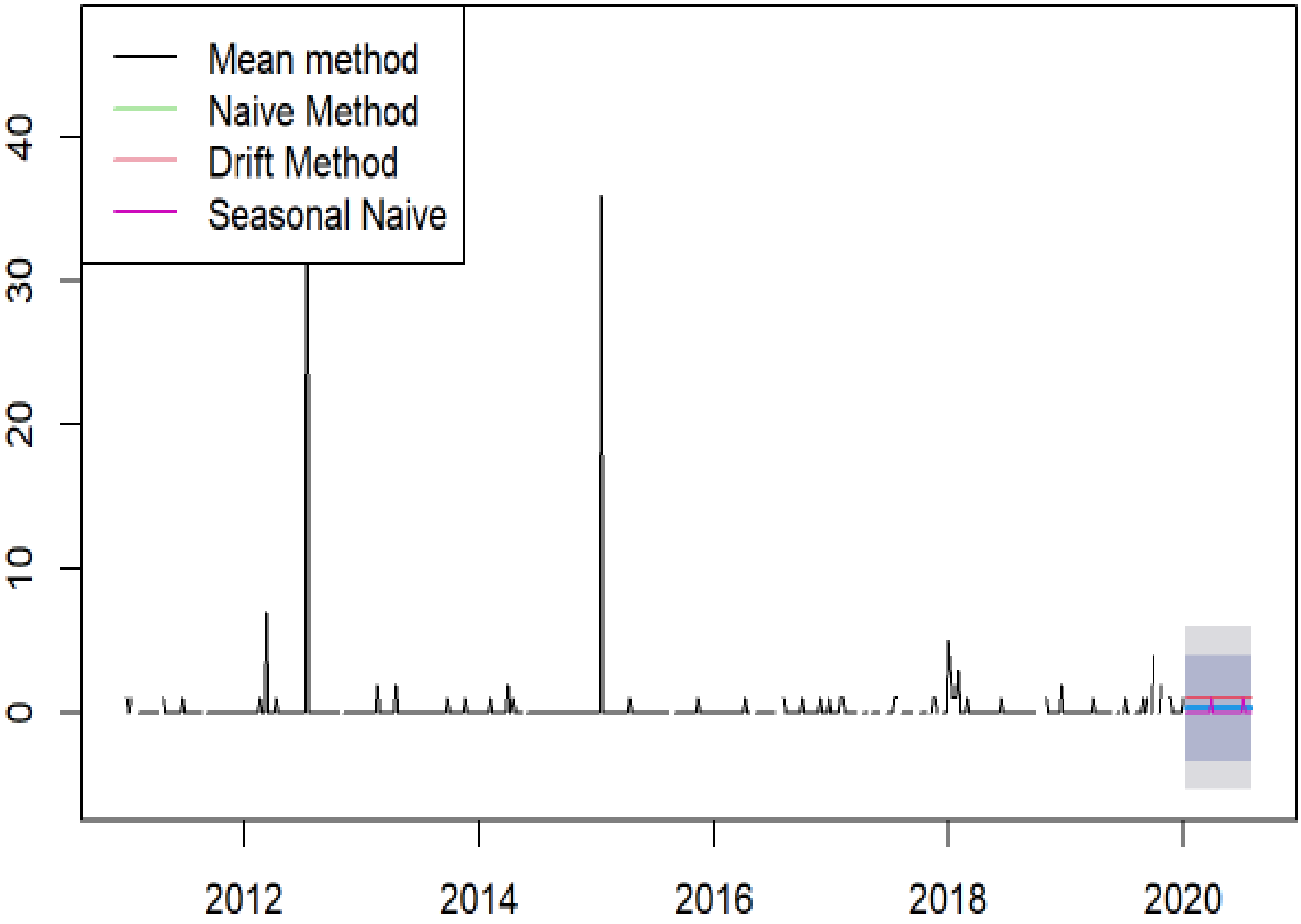
```
## Warning in title(main = main, xlab = xlab, ylab = ylab, ...): "plot.conf" is not
## a graphical parameter
```

```
## Warning in axis(1, ...): "plot.conf" is not a graphical parameter
```

```
## Warning in axis(2, ...): "plot.conf" is not a graphical parameter
```

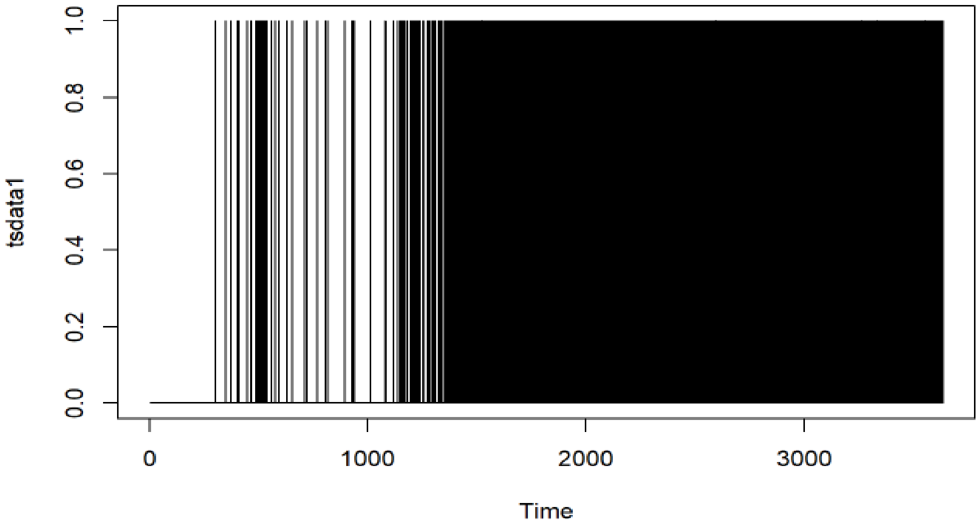
```
## Warning in box(...): "plot.conf" is not a graphical parameter
```

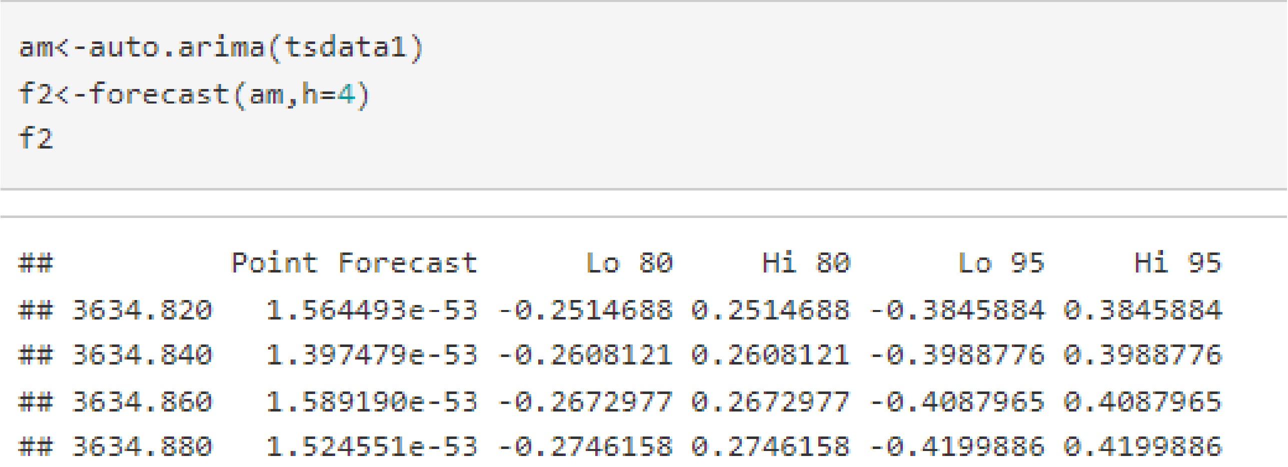
```
lines(naivem$mean,col=3,lty=1)
lines(driftm$mean,col=2,lty=1)
lines(snaivem$mean,col=6,lty=1)
legend("topleft",lty=1,col=c(1,3,2,6),legend=c("Mean method","Naive Method","Drift Method","Seasonal Naive"))
```



for suicide attributes

```
tsdata1<-ts(data$suicide,frequency=50,start = c(1,1))
plot(tsdata1)
```



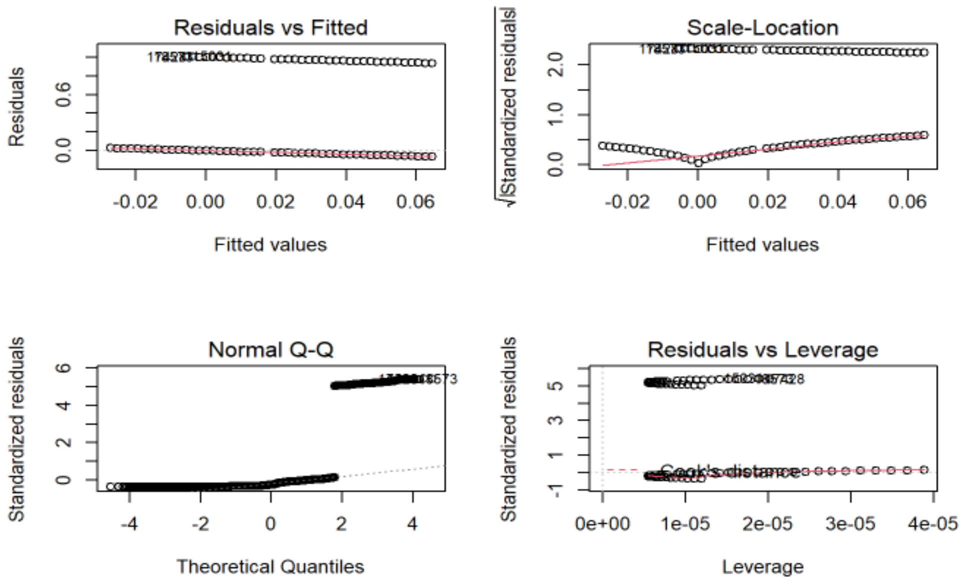


```
am<-auto.arima(tsdata1)
f2<-forecast(am,h=4)
f2
```

```
##          Point Forecast      Lo 80     Hi 80      Lo 95     Hi 95
## 3634.820   1.564493e-53 -0.2514688 0.2514688 -0.3845884 0.3845884
## 3634.840   1.397479e-53 -0.2608121 0.2608121 -0.3988776 0.3988776
## 3634.860   1.589190e-53 -0.2672977 0.2672977 -0.4087965 0.4087965
## 3634.880   1.524551e-53 -0.2746158 0.2746158 -0.4199886 0.4199886
```

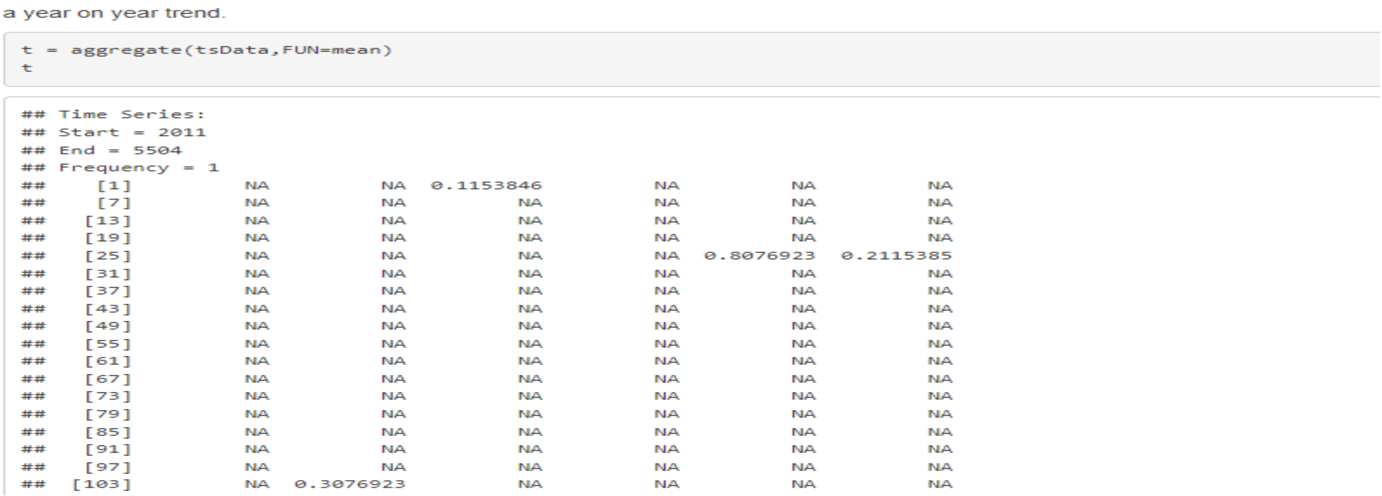
fit in a Linear Model (Intercept & Slope), and plot the line

```
layout(matrix(1:4,2,2))
plot(lm(data$suicide~data$iyear))
```



aggregate the cycles and display

a year on year trend.

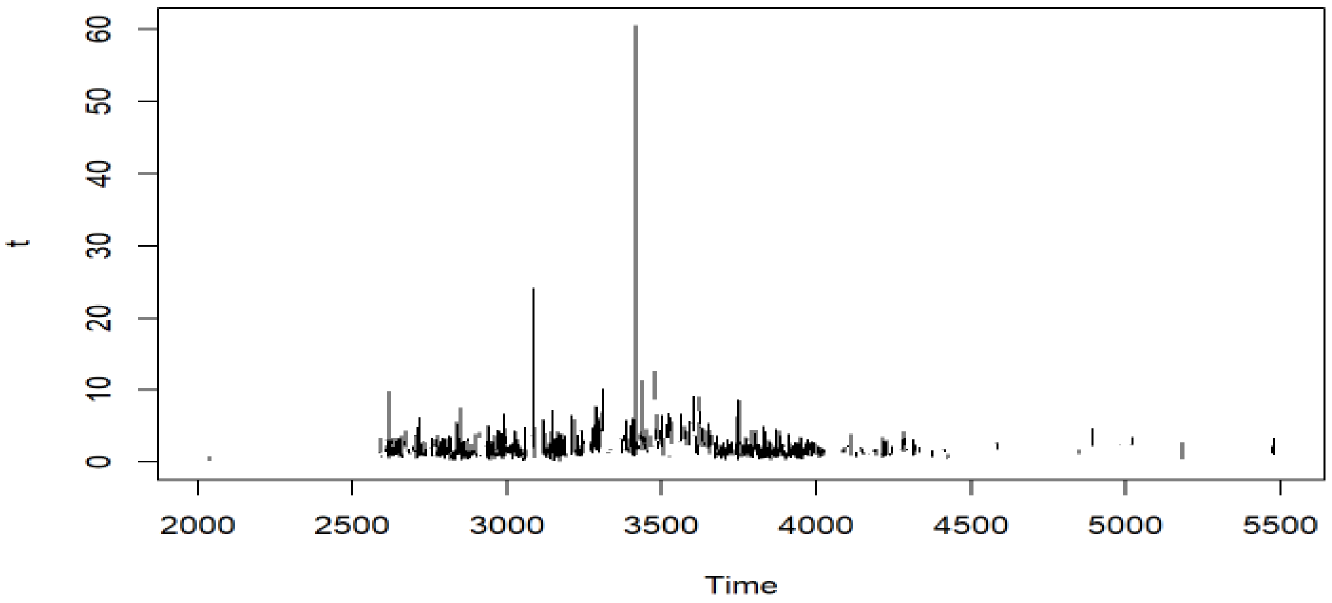


a year on year trend.

```
t = aggregate(tsData,FUN=mean)
t
```

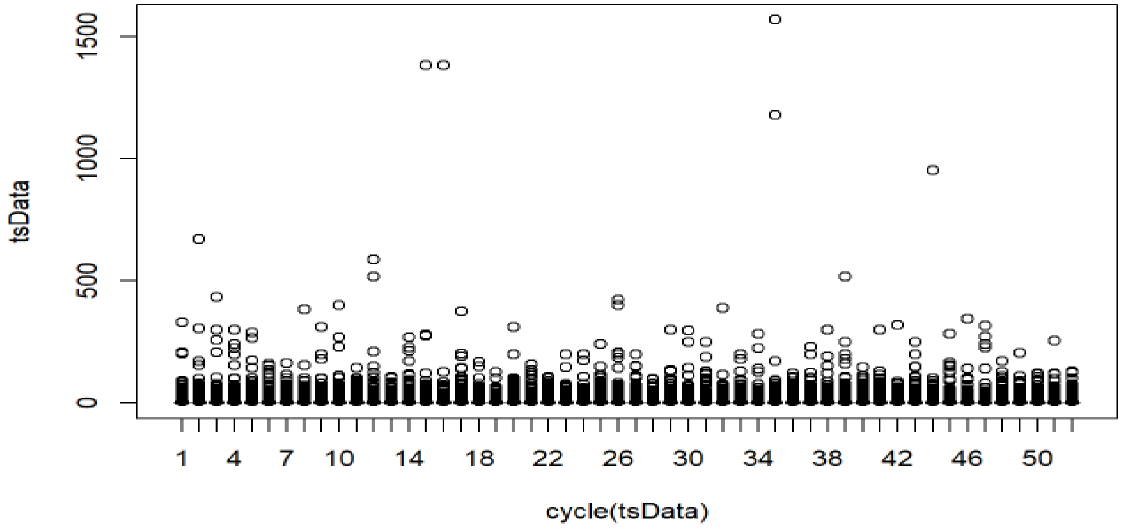
```
## Time Series:
## Start = 2011
## End = 5504
## Frequency = 1
##     [1]         NA         NA  0.1153846         NA         NA         NA
##     [7]         NA         NA         NA         NA         NA         NA
##    [13]         NA         NA         NA         NA         NA         NA
##    [19]         NA         NA         NA         NA         NA         NA
##    [25]         NA         NA         NA         NA  0.8076923  0.2115385
##    [31]         NA         NA         NA         NA         NA         NA
##    [37]         NA         NA         NA         NA         NA         NA
##    [43]         NA         NA         NA         NA         NA         NA
##    [49]         NA         NA         NA         NA         NA         NA
##    [55]         NA         NA         NA         NA         NA         NA
##    [61]         NA         NA         NA         NA         NA         NA
##    [67]         NA         NA         NA         NA         NA         NA
##    [73]         NA         NA         NA         NA         NA         NA
##    [79]         NA         NA         NA         NA         NA         NA
##    [85]         NA         NA         NA         NA         NA         NA
##    [91]         NA         NA         NA         NA         NA         NA
##    [97]         NA         NA         NA         NA         NA         NA
##   [103]         NA  0.3076923         NA         NA         NA         NA
```

```
plot(t)
```

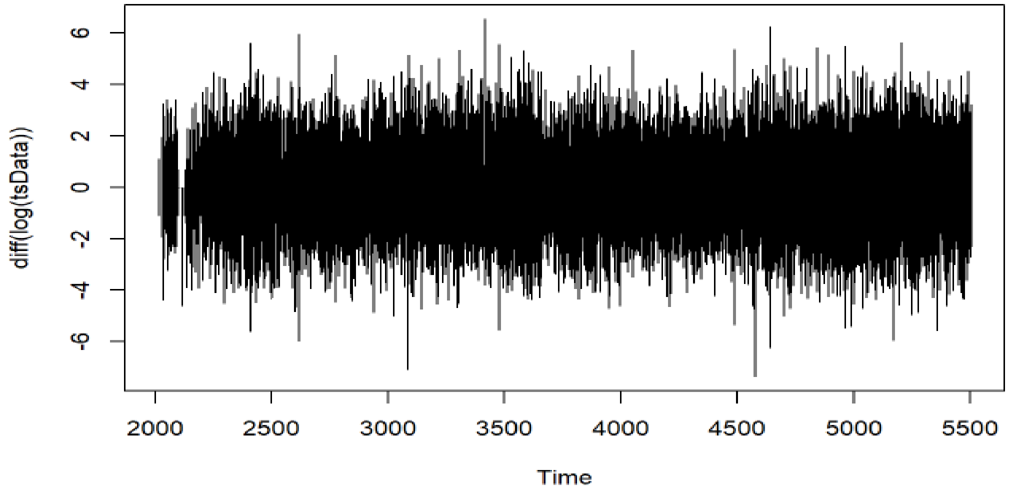


seasonal effect

```
boxplot(tsData ~ cycle(tsData))
```



```
plot(diff(log(tsData)))
```

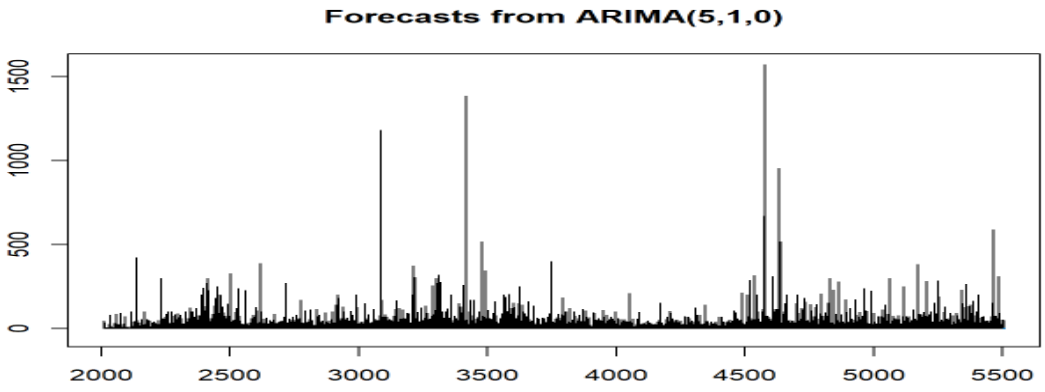


```
autoArimaModel = auto.arima(tsData, d = 1)
autoArimaModel
```

```
## Series: tsData
## ARIMA(5,1,0)
##
## Coefficients:
##          ar1      ar2      ar3      ar4      ar5
##      -0.7122  -0.5679  -0.4324  -0.2966  -0.1609
## s.e.  0.0025   0.0029   0.0030   0.0029   0.0024
##
## sigma^2 estimated as 149.6:  log likelihood=-673102.3
## AIC=1346217   AICc=1346217   BIC=1346277
```

##predict and test

```
autoPred = forecast(autoArimaModel, h=25)
plot(autoPred)
```



Conclusion

The goal of this project was to build a tool which helps users to understand and interpret the nature of terrorism. Users can perceive the START dataset through visual designs. A visualization which can be used to calculate the total number of attacks, total kill counts and location based on the selected region and year provides interactive interface to explore this dataset. Users can understand various patterns, trends and correlation in terrorism through visual interpretation and its provided explanation. Users can also explore START dataset and other terrorism related sources for additional research purposes provided in this tool. 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.

Future Work

Following is a list of directions which can enhance the quality and quantity of this current project work:

1. **Improve dataset quality**: Dataset needs to be populated more by adding the missing values. Many historical terrorist events are yet to be documented because of conflict in information from multiple sources or lack of credibility from the

source providing information. Resolving this conflict will increase the scope of analysis of new attributes that are mostly sparse at the moment.

2. **Prediction**: Different prediction methodologies can be used to make a system that can predict various parameters like attack count, rate of a successful attack, prospective casualties, type of attack, types of weapons used, etc. Currently, prediction models are difficult to achieve high accuracy because of the relatively small dataset size.
3. **Enhance current work**: More techniques can be added in this project like classification and regression. Design sophisticated patterns like how terrorist groups act and react over the years. Add more visualisations to make user interface more interactive.
4. **Connections with other datasets**: Exploring impacts of terrorism on other fields like country's Development index, stock market, international investments, happiness rating, tourism, etc. can reveal new patterns and relationships among them. These correlations will help understand how terrorism influences other domains.

References

[1] United Nations, “Chapter-1 Purposes and Principles,” [Online]. Available: https://www.un.org/en/sections/un-charter/chapter-i/index.html [Accessed: May 2019].

[2] A. Z. Borda, “Why we react differently to terror attacks depending on where they happen,” [Online]. Available: http://theconversation.com/why-we-react-differently-to-terror-attacks-depending-on-where-they-happen-57389 [Accessed: May 2019].

[3] START organization, “Global Terrorism Database,” [Online]. Available: https://www.start.umd.edu/gtd/about/ [Accessed: May 2019].

[4] Kaggle, “Global Terrorism Database,” [Online]. Available: https://www.kaggle. com/ash316 [Accessed: August 2019].

[5] START Consortium Organization, “Global Terrorism Index,” [Online]. Available: http://visionofhumanity.org/app/uploads/2017/02/Global-Terrorism-Index-2016. pdf [Accessed: October 2019].