

A Data Analytics (CSE3505) Final project report

**Visual analysis of Terrorism and its counter measuring**

**Submitted to:**

#### Dr. Rajalakshmi R

**Submitted by:**

Reg No: 19BCE1129 Name: Sheryas

Reg No: 19BPS1095 Name: Tarun Sairam

Reg No: 19BPS1124 Name: Sai mouli

In partial fulfilment for the award of the degree of

#### Bachelor of Technology

in

#### Computer Science and Engineering

*November 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. R. Rajalakshmi, 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, **Dr. R. Rajalakshmi,** School of Computer Science and Engineering for her 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 **Sheryas(19BCE1129), Tarun sairam(19BPS1095), Sai mouli(19BPS1124)** carried out the “J”-Project work under my supervision and guidance for **CSE3505-Foundation of Data Analytics**.

**Dr. R. Rajalakshmi**

SCOPE

**TABLE OF CONTENTS**

Page

1. ABSTRACT 7
2. Introduction 8
3. Background 9
4. Misconceptions about Terrorism
5. Factors affecting Terrorism
6. Dataset Challenges
7. System Design 11
8. Methodology 14
9. Related Works (Literature Review) 15
10. Analysis 16
11. Experiments and Results 24
12. Experimental setup,
13. Results and Discussion
14. Conclusion 51
15. References 52

**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 [5]. Hence, analyses and results provided in this study might

vary with the similar analyses done on a different dataset.

**BACKGROUND**

1. **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 marketdata with terrorism as a control variable directly provides a strong connection among both.

1. **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.

1. **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**

1. **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’ attribute.

**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 likeWindows, R studio, etc. It can be used for high level data analysis. Tableau Desktop software application is used as a data visualization tool for rawdata 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.

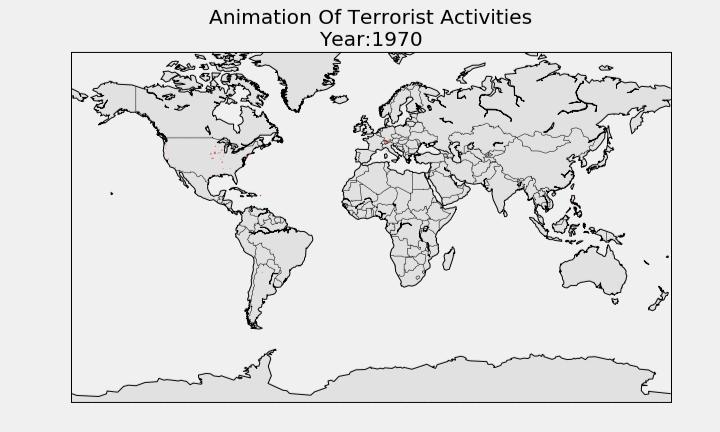
The problem statement here is to build a tool that can present processed information in the form of intuitive visual representation of analyzed data. Implementation of this project involves system design, backend design, visual design, and user interface. System design includes the overall design plan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

**Analysis**

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

1. **Animation of Terrorist Activities**

****

**Figure 3. Attacking Methods by Terrorist**

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

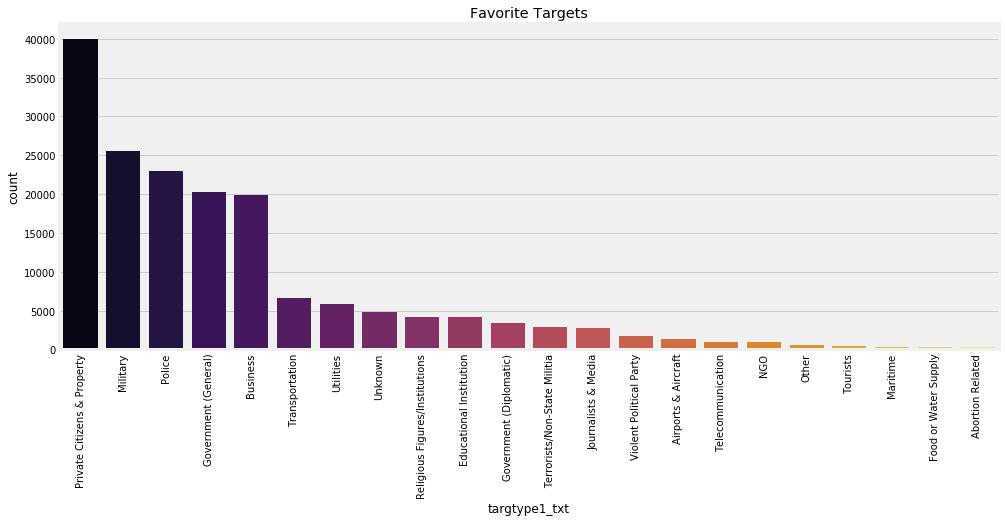
followed by armed assaults, assassinations, hostage and so on. Here the total number of

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.

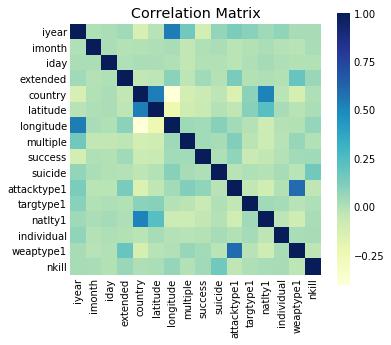
**3.3.2 Favorite Targets**

****

**Figure 4. Favorite Targets**

The attacker always tries to make an impact by targeting their victims. Analyzing the type of target will help understand their objective and most likely their motives. Terrorism is driven by an ideology that tries to make a change or impose an ideology. Looking into the most commonly targeted attributes will signify the attacker’s objectives and terrorism in general. There are more than 100 distinct target type. These target types are generalized into 22 categorizes. Here Figure 4 shows that citizens, military, government, and police are the most common targets. This graph explains that terrorist groups or individuals have a dislike towards the authority of the state or the nation. Their main focus is either to make a change at the political level or force their ideology on government forces by retaliating against their authority.

**3.3.3 Correlation Matrix**

****

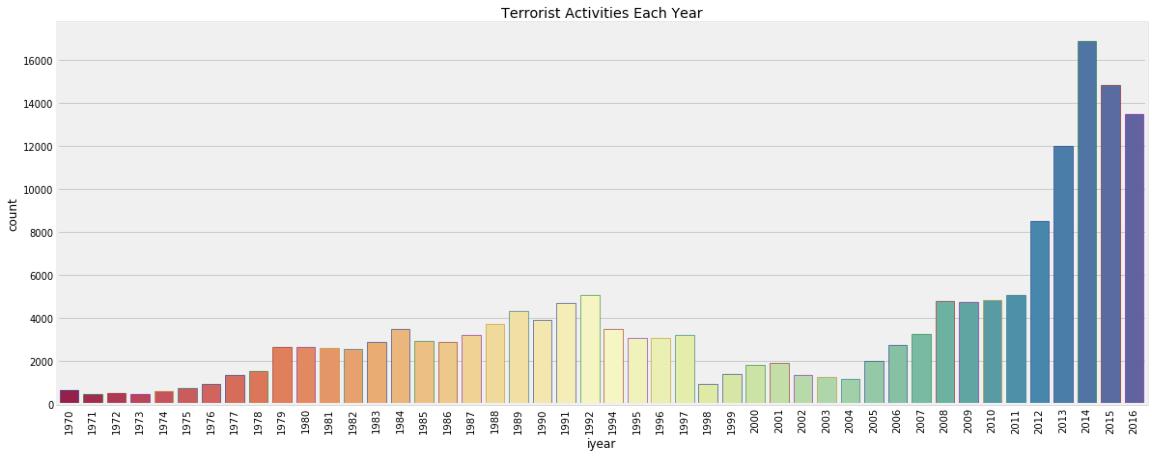
**Figure 5. Correlation Matrix**

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.

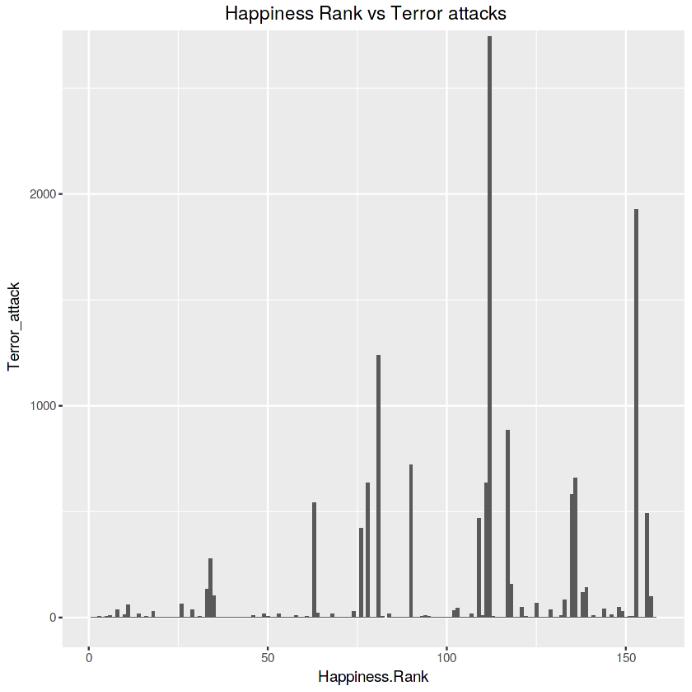
**3.3.4 Terrorist Activities Each Year**

****

**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.

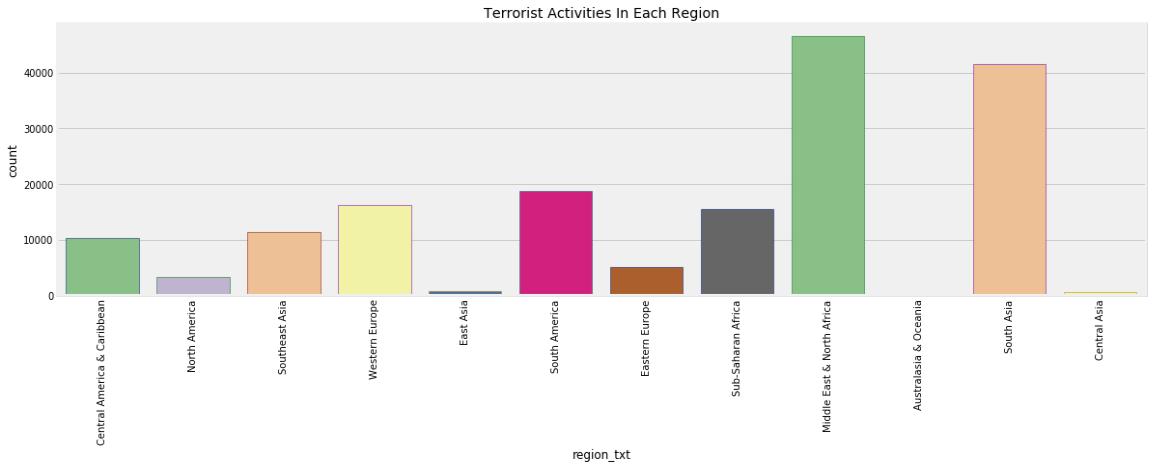
**3.3.5 Happiness Rank vs Terror Attacks**

****

**Figure 7. Happiness Rank vs Terror Attacks**

The Happiness Index dataset [29] provides information regarding the happiness index of almost 170 countries for 2014, 2015, and 2016. This dataset is provided by Sustainable Development Solutions Network [30]. Comparing these dataset ranks of countries with the total number of terrorist attacks associated with them in the same year can explain if terrorism impacts country’s happiness rank. Taking the mean of all these 3 years to obtain the rank of countries to examine rank index against the total number of attacks those countries have faced in the same 3 years period. Consider a case in which terrorism does affect the happiness rank of a given country. For that scenario, we should expect a skewed graph where the rank of the country should increase with the increase in the total number of attacks. But according to Figure 7, bar distribution is not skewed. This observation means there are countries that are adversely impacted by terrorism but still have competitive ranks and on the far side of the chart, there are countries that have a relatively low impact of terrorism and yet are far behind in the happiness rank. Of course, there are multiple factors other than terrorism which contribute to these ranks, but terrorism is not the major contributor in this happiness index scenario.

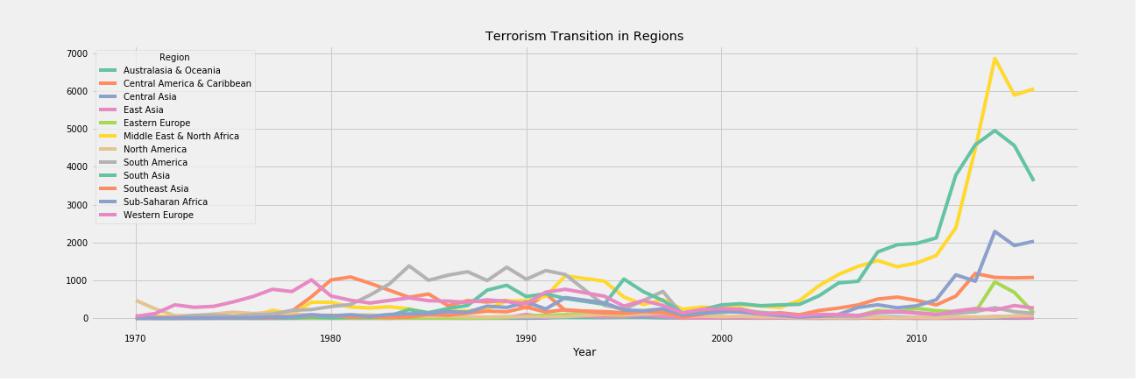
**3.3.6 Terrorist Activities in Each Region**

****

**Figure 8. Terrorist Activities 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 10. 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.

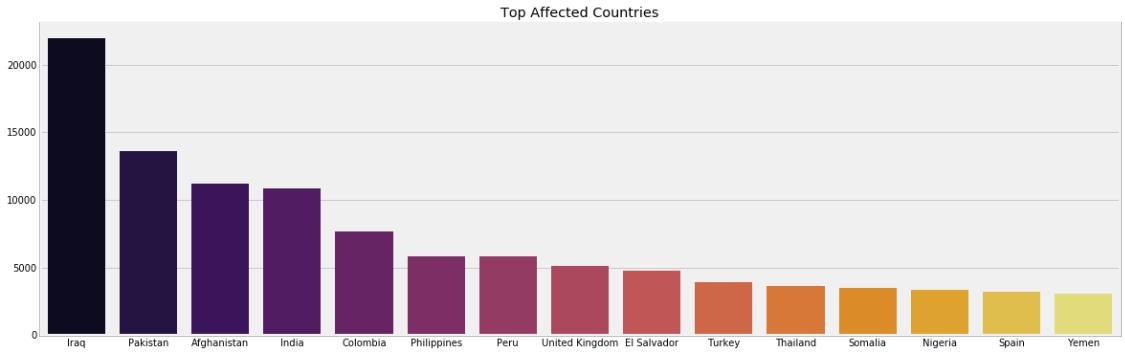
**3.3.7 Terrorism Transition in Regions**

****

**Figure 9. Terrorism Transition in Regions**

Figure 11 provides more details about how terrorism has aged in various regions. Observing this pattern, we can see that there has been a noticeable rise and falls of the graph for all regions. From the Figure 10, we know that South America has been the third-highest affected region in the world after North Africa & the Middle East and South Asia. But from Figure 11, we see South America has no significant contribution to the current trending terrorism. South America was impacted by terrorism during the early 1980s to mid-1990s. Since then, terrorism has been relatively low. So even though South America region shows an overall high rate of total attacks compared to other regions overall, this region does not add much to the current global terrorism situation. Opposite of this is true for the Middle East and North Africa region. Middle East and North Africa has shown no noticeable rise in terrorism other than the early 2000s. This change in terrorism has been sudden and steep. But otherwise, there is no previous history of such high terrorist activities in the past. Figure 11 emphasizes that there has been a rise and fall in almost all regions. So even though terrorism has prevailed throughout history, there is no region that shows consistent involvement of terrorism over time.

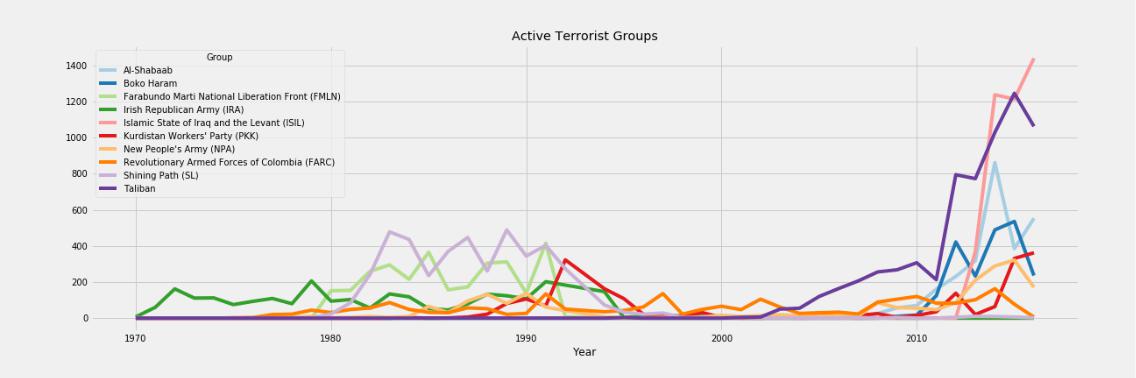
**3.3.8 Top Affected Countries**

****

**Figure 10. Top Affected Countries**

Figure 12 shows some of the most affected countries are Iraq, Pakistan, Afghanistan, and India based on the total number of attacks. Peru is also one of the most affected countries according to Figure12 but from previous results of Figure 11 we observed that Peru does not represent its current terrorism situation based on the total number of attacks. But Figure 12 graph does explain how some countries are prone to violent actions and difference in an ideology which can lead to extreme terrorism.

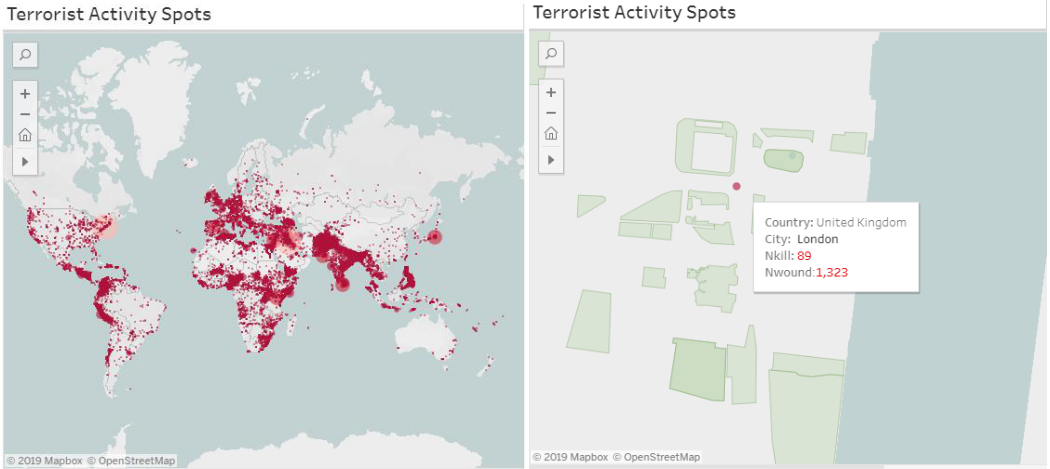
**Active Terrorist Groups**

****

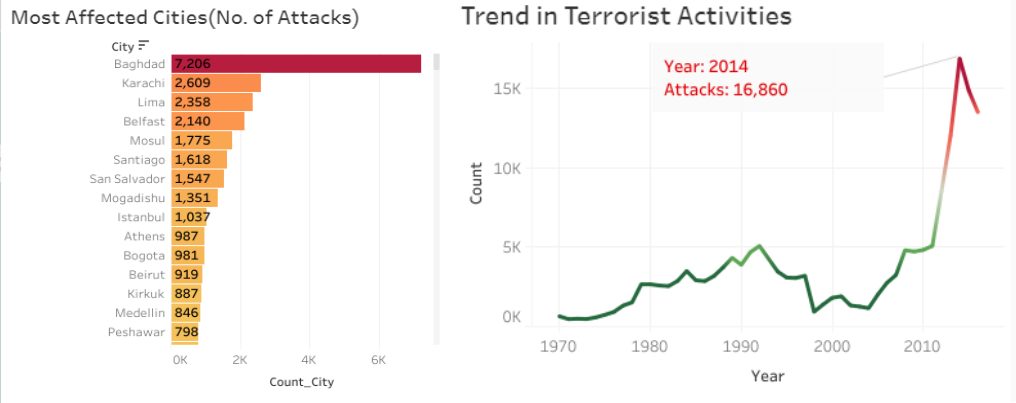
**Figure 11. Active Terrorist Groups**

Figure 13 describes some of the most notorious terrorist groups and their active years. Groups like SL (Shining Path) and FMLN (Farabundo Marti National Liberation Front) were highly proactive for almost over a decade from the 1980s to 1990s but after that no trace of those groups in following years. The same goes for the Boku Haram group for a different time interval. The majority of these groups have not made a comeback since their fall. New groups, however, rise with a different set of reasons and try to spread their ideology. Hence different strategies need to be played while tackling diverse groups. The same strategies used over time cannot be applied to different groups. Taliban and ISIL are the major contributors in the recent rise in attacks, especially in North Africa and the Middle East regions. Today there are more active terrorist groups spreading violence and terror than ever before. Almost all groups show multiple crests and troughs during their existence. One of the reasons for this observation could be the resistance terrorist groups have faced from multiple counter-terrorism forces over time. Such patterns are quite regular in the majority of the groups.

**3.3.10: Terrorist Activity Spots**

****

The left image in Figure 13 shows five options to select the region user wants to explore. Three of them are the different geometric ways to overlay the region. On the right image, in Figure 15 showing the result of a region selected by the user. Events contained in that selected region are highlighted over the rest. Hovering over any of the highlighted spot shows statistics related to that event along with the sum of the total casualties and events happened for that selected region.



**Figure 14. Affected Cities and Trends in Terrorism**

Most affected cities (Figure 14: image on left) list the results based on the filters selected. User

can toggle to specific year to see the number of attacks for each city. The image on the right in Figure 16, shows the overall attacks happened over each decade signifying its peak in 2014.

**APPENDIX**

**Implementation / Code**

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

1. **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 dataset

1. **Results and Discussion**

## 1.1 Load libraries

**library**(tidyverse)

**library**(data.table)

**library**(lubridate)

**library**(RColorBrewer)

**library**(gridExtra)

**library**(plotly)

**library**(ggthemes)

**library**(wesanderson)

**library**(leaflet)

**library**(VIM)

## 1.2 Load data

dt <- as.tibble(fread("globalterrorismdb\_0221dist.csv",

na.strings = c("", "NA")))

dt

## # A tibble: 201,183 x 135

## eventid iyear imonth iday approxdate extended resolution country country\_txt

## <int64> <int> <int> <int> <chr> <int> <chr> <int> <chr>

## 1 197000000001 1970 7 2 <NA> 0 <NA> 58 Dominican ~

## 2 197000000002 1970 0 0 <NA> 0 <NA> 130 Mexico

## 3 197001000001 1970 1 0 <NA> 0 <NA> 160 Philippines

## 4 197001000002 1970 1 0 <NA> 0 <NA> 78 Greece

## 5 197001000003 1970 1 0 <NA> 0 <NA> 101 Japan

## 6 197001010002 1970 1 1 <NA> 0 <NA> 217 United Sta~

## 7 197001020001 1970 1 2 <NA> 0 <NA> 218 Uruguay

## 8 197001020002 1970 1 2 <NA> 0 <NA> 217 United Sta~

## 9 197001020003 1970 1 2 <NA> 0 <NA> 217 United Sta~

## 10 197001030001 1970 1 3 <NA> 0 <NA> 217 United Sta~

## # ... with 201,173 more rows, and 126 more variables: region <int>,

## # region\_txt <chr>, provstate <chr>, city <chr>, latitude <dbl>,

## # longitude <dbl>, specificity <int>, vicinity <int>, location <chr>,

## # summary <chr>, crit1 <int>, crit2 <int>, crit3 <int>, doubtterr <int>,

## # alternative <int>, alternative\_txt <chr>, multiple <int>, success <int>,

## # suicide <int>, attacktype1 <int>, attacktype1\_txt <chr>, attacktype2 <int>,

## # attacktype2\_txt <chr>, attacktype3 <int>, attacktype3\_txt <chr>, ...

##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…

There are 135 variables in the original data.

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: 201,183

## Columns: 17

## $ eventid <int64> 197000000001, 197000000002, 197001000001, 197001000002, ~

## $ 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> "National", "Federal", "Tarlac", "Attica", "Fukouka", "Ill~

## $ 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~

## 1.3 some sample values

head(gbtr)

## # A tibble: 6 x 17

## eventid iyear imonth iday country\_txt region\_txt provstate city latitude

## <int64> <int> <int> <int> <chr> <chr> <chr> <chr> <dbl>

## 1 197000000001 1970 7 2 Dominican ~ Central A~ National Sant~ 18.5

## 2 197000000002 1970 NA NA Mexico North Ame~ Federal Mexi~ 19.4

## 3 197001000001 1970 1 NA Philippines Southeast~ Tarlac Unkn~ 15.5

## 4 197001000002 1970 1 NA Greece Western E~ Attica Athe~ 38.0

## 5 197001000003 1970 1 NA Japan East Asia Fukouka Fuko~ 33.6

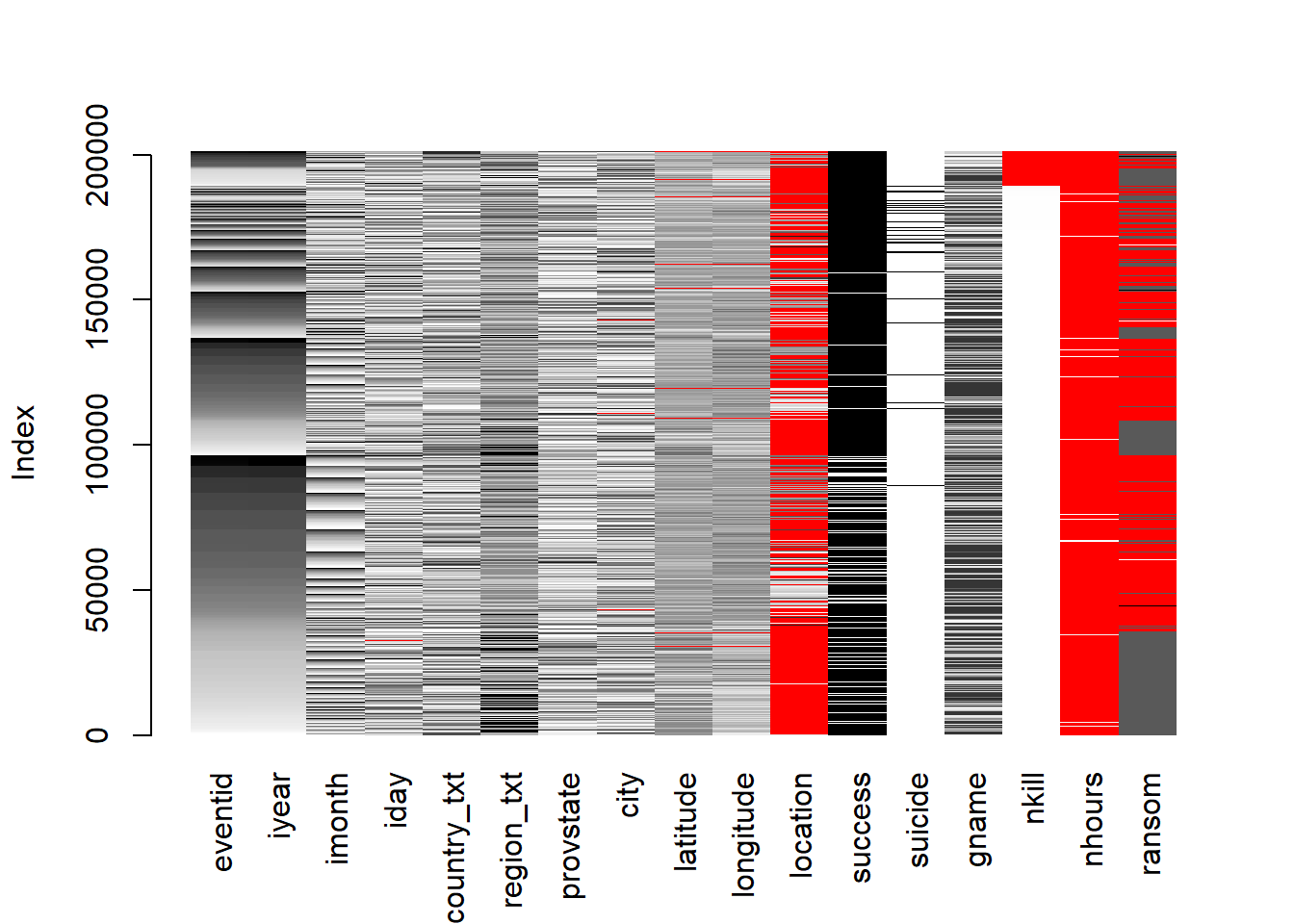
## 6 197001010002 1970 1 1 United Sta~ North Ame~ Illinois Cairo 37.0

## # ... with 8 more variables: longitude <dbl>, location <chr>, success <int>,

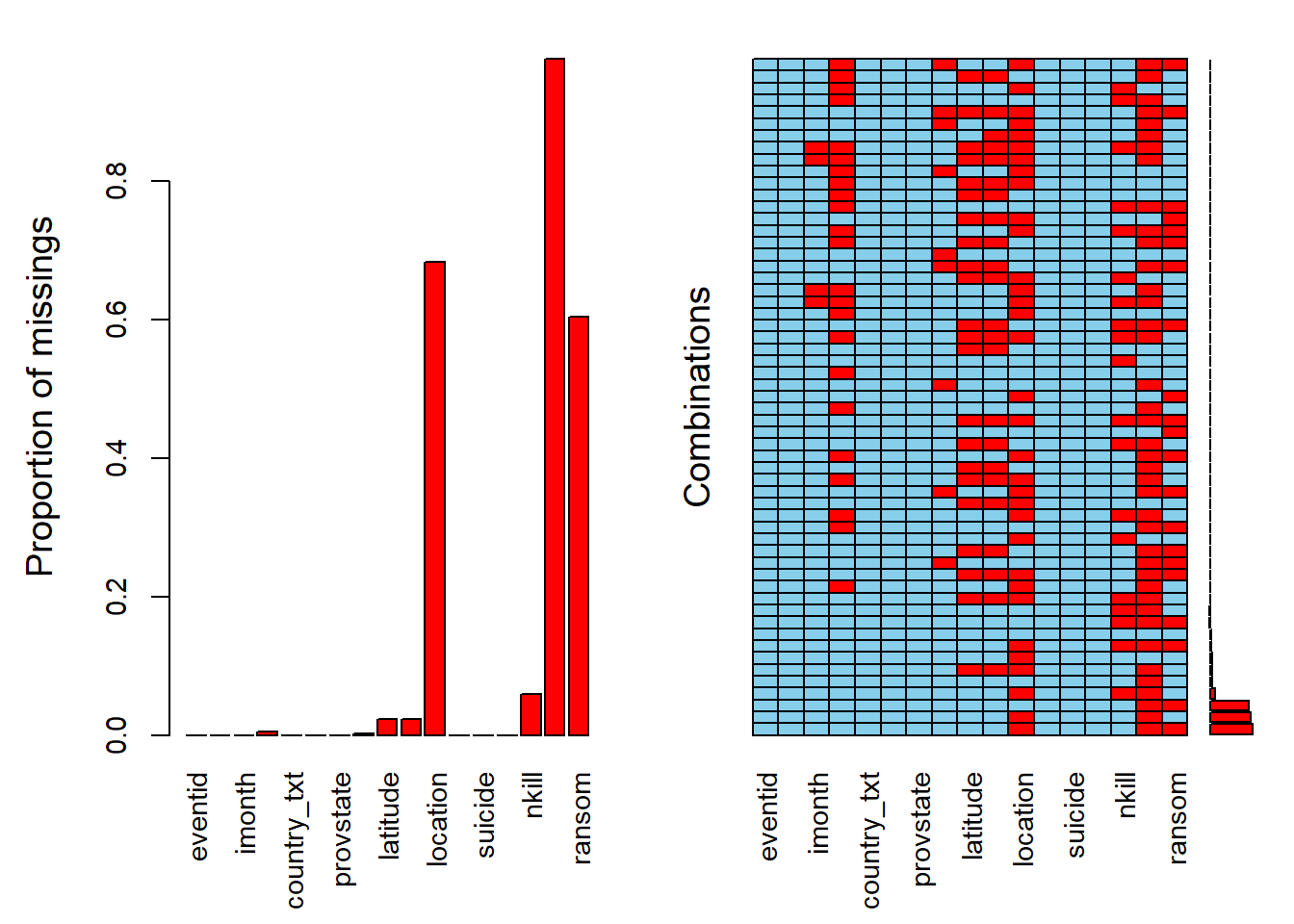
## # suicide <int>, gname <chr>, nkill <int>, nhours <dbl>, ransom <int>

## 1.4 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 valid entries/values.EDA with these variables will be avoided for further reduction of complexity.

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

## 2.1 Events by year

p <- gbtr %>% mutate(iyear=as.factor(iyear)) %>%

group\_by(iyear) %>% count() %>%

ggplot(aes(x=iyear,y=n,group=1)) +

geom\_line(size=1, color="brown")+

geom\_point(color="brown") +

scale\_x\_discrete(

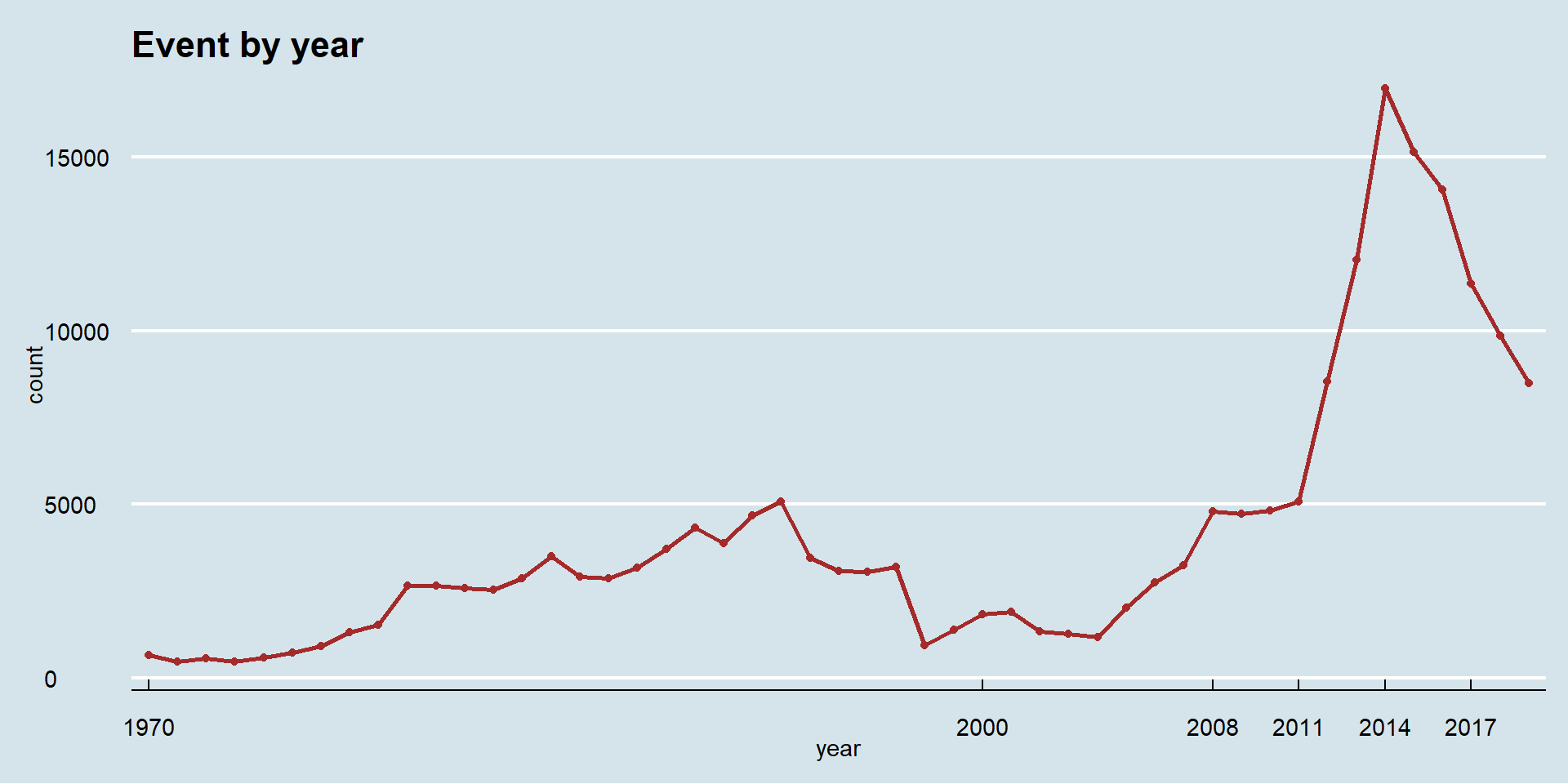
breaks=c("1970", "2000","2008", "2011", "2014","2017")

) +

labs(title = "Event by year", x = "year", y = "count")+

theme\_economist()

p

****

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

## 2.2 Overall trend in each region

p4 <- gbtr %>% count(region\_txt, iyear) %>%

ggplot(aes(iyear, n,color=region\_txt)) +

geom\_line(aes(group=region\_txt)) +

labs(title = "Trend by Region", x="year", y="count", color="region")+

theme\_light()

ggplotly(p4)

1970198019902000201020200200040006000

Australasia & OceaniaCentral America & CaribbeanCentral AsiaEast AsiaEastern EuropeMiddle East & North AfricaNorth AmericaSouth AmericaSouth AsiaSoutheast AsiaSub-Saharan AfricaWestern EuropeTrend by Regionyearcountregion

**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.

## 2.3 Events & num. of kills by region

Since there is a steep upward trend since aproximately year 2000, we’ll inspect the period before and after 2000 seperately.

p2 <- gbtr %>% mutate(pd=ifelse(iyear<2000,"before 2000", "after 2000")) %>%

mutate(pd = factor(pd, levels = c("before 2000", "after 2000")))%>%

group\_by(region\_txt, pd) %>% count() %>%

ggplot(aes(x=reorder(region\_txt, n), y=n))+

geom\_bar(aes( fill=pd), stat= "identity", position = "dodge")+

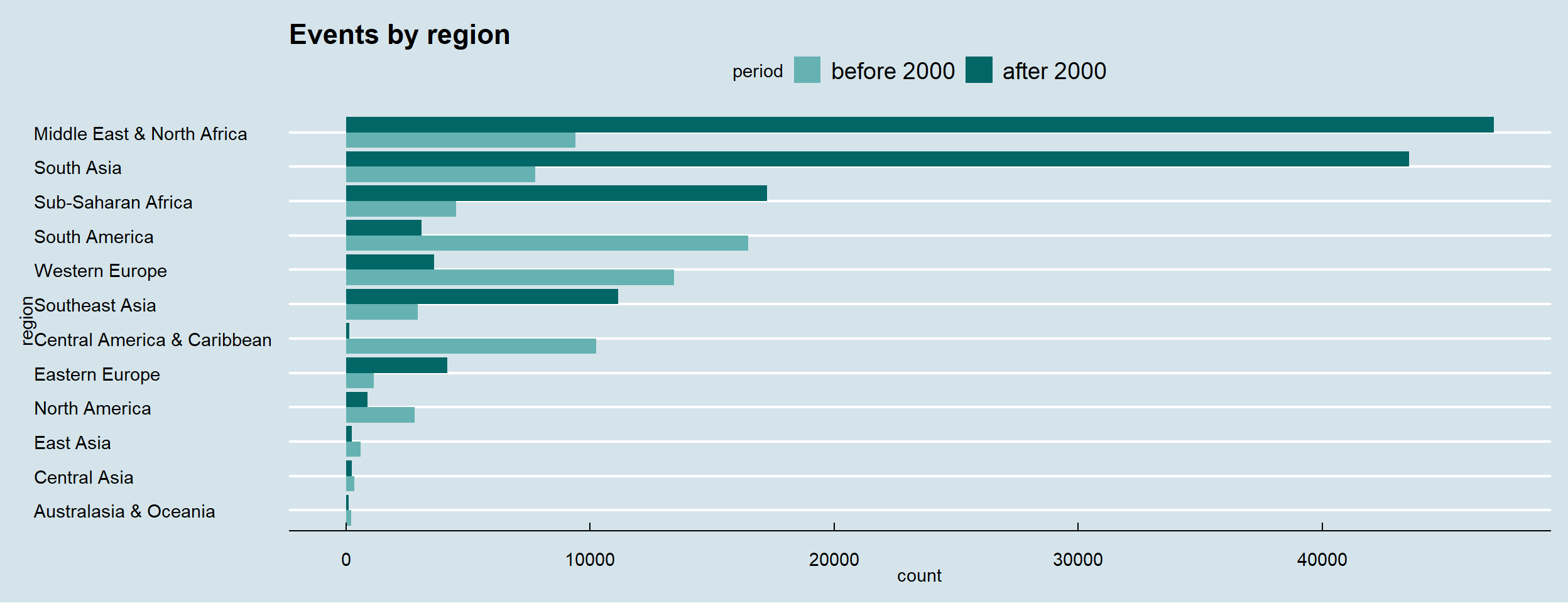
labs(title = "Events by region", x = "region", y = "count", fill = "period")+

theme\_economist()+

scale\_fill\_manual(values = c("#66b2b2","#006666")) +

coord\_flip()

p2

****

* 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.

## 2.4 Number of deaths and number of events

pkr <- gbtr2k %>% filter(!is.na(nkill)) %>% group\_by(region\_txt) %>%

summarise(ksum=sum(nkill)) %>%

ggplot(aes(reorder(region\_txt,ksum), ksum))+

geom\_bar(stat = "identity", fill="#2E8B57")+

coord\_flip()+

labs(title = "Num. of kills by region", subtitle = "without missing values, after 2000", x="region", y="count")+

theme\_economist()

per <- gbtr2k %>% group\_by(region\_txt) %>% count() %>% top\_n(10,n) %>%

ggplot(aes(x=reorder(region\_txt, n), y=n))+

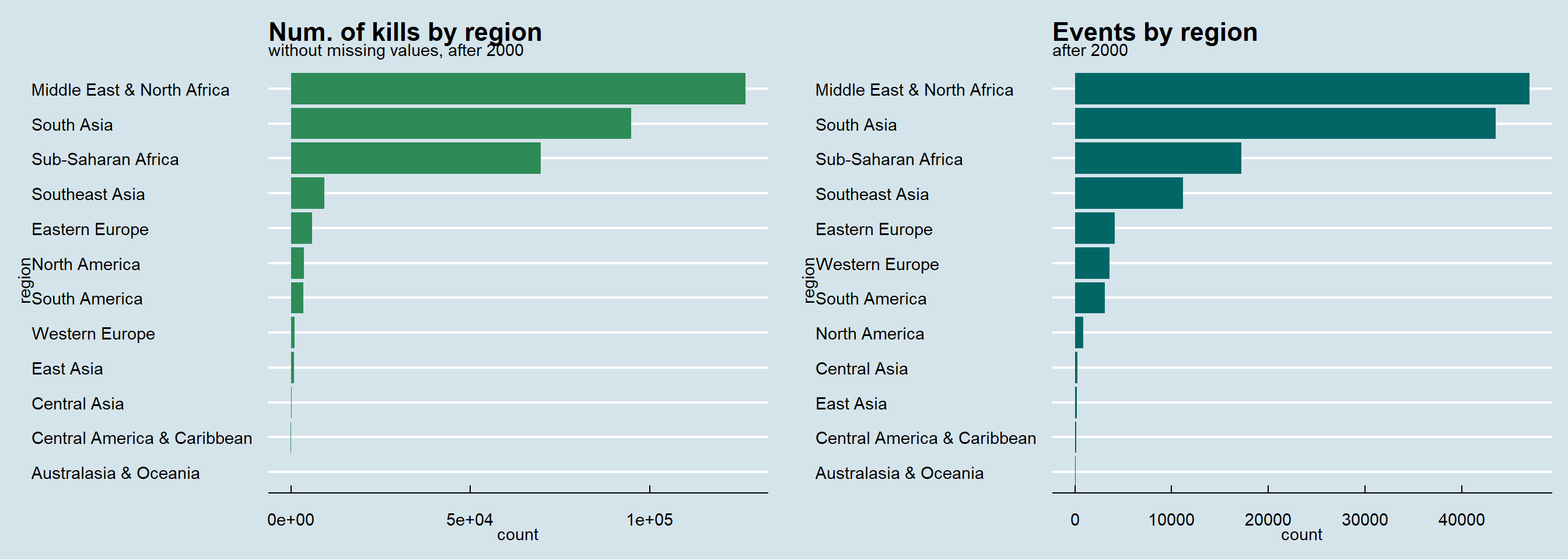
geom\_bar(stat= "identity", fill="#006666")+

labs(title = "Events by region",subtitle = "after 2000", x = "region", y = "count")+

theme\_economist()+

coord\_flip()

grid.arrange(pkr,per,ncol=2)

****

* 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.

## 2.5 Events by country

We’ll look at data after year 2000

pec <- gbtr2k %>% group\_by(country\_txt) %>% count() %>% ungroup() %>%

top\_n(n=20,wt = n) %>%

ggplot(aes(reorder(country\_txt, n), n))+

geom\_bar(stat = "identity", fill="#21618C") +

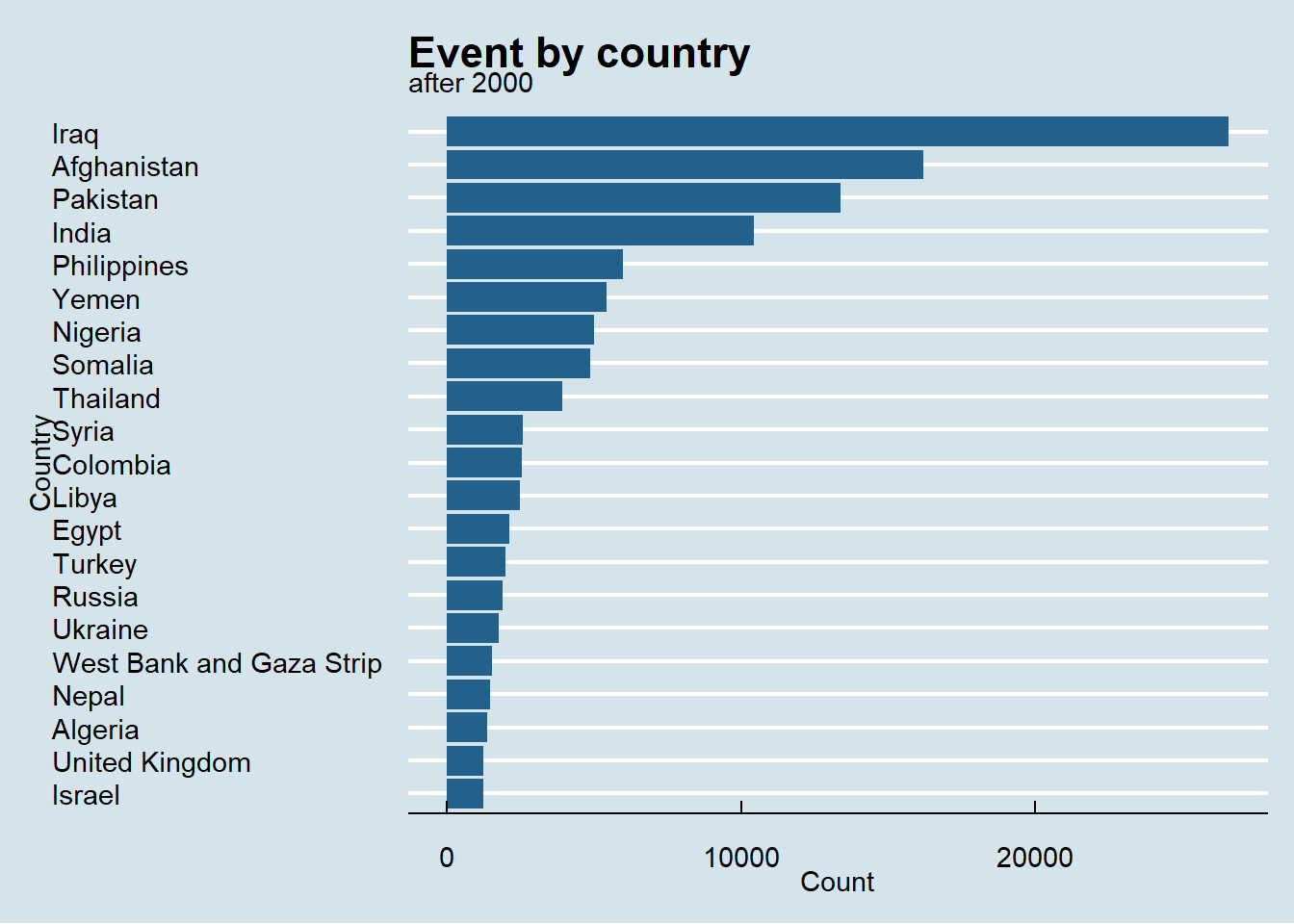
labs(title = "Event by country", subtitle = "after 2000", x = "Country", y = "Count") +

theme\_economist() +

scale\_fill\_manual(values = wes\_palette(n=4,"Cavalcanti1"))+

coord\_flip()

pec

****

## 2.6 Suicide attack

dtscd <- gbtr2k %>% filter(!is.na(suicide)) %>% group\_by(region\_txt, suicide) %>% count() %>%

ungroup() %>% group\_by(region\_txt) %>% mutate(pct=n/sum(n)) %>% filter(suicide==1)

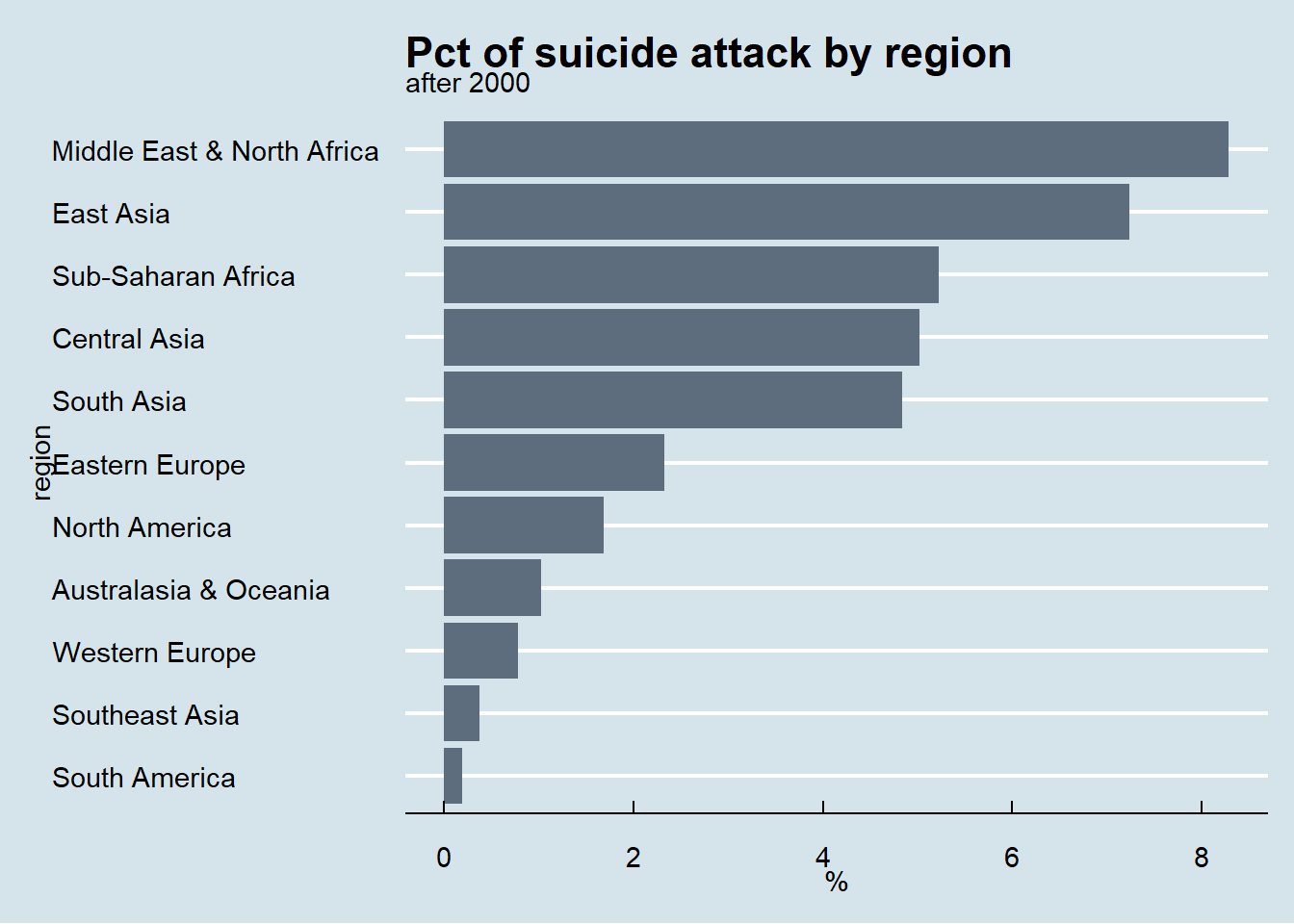
ggplot(dtscd, aes(reorder(region\_txt, pct), pct\*100)) +

geom\_bar(stat = "identity", fill="#5D6D7E")+

coord\_flip()+

labs(title = "Pct of suicide attack by region", subtitle = "after 2000", x="region",y="%")+

theme\_economist()

****

## 2.7 Groups, attacks, and suicide

gbtr %>%filter(gname!="Unknown") %>% group\_by(gname,suicide) %>% summarise(n=n()) %>%

ungroup() %>% group\_by(gname) %>% mutate(sum=sum(n)) %>% ungroup() %>% top\_n(30,sum) %>%

ggplot(aes(x=reorder(gname,sum),n, fill=factor(suicide, levels = c(1,0)))) +

geom\_bar(stat = "identity") +

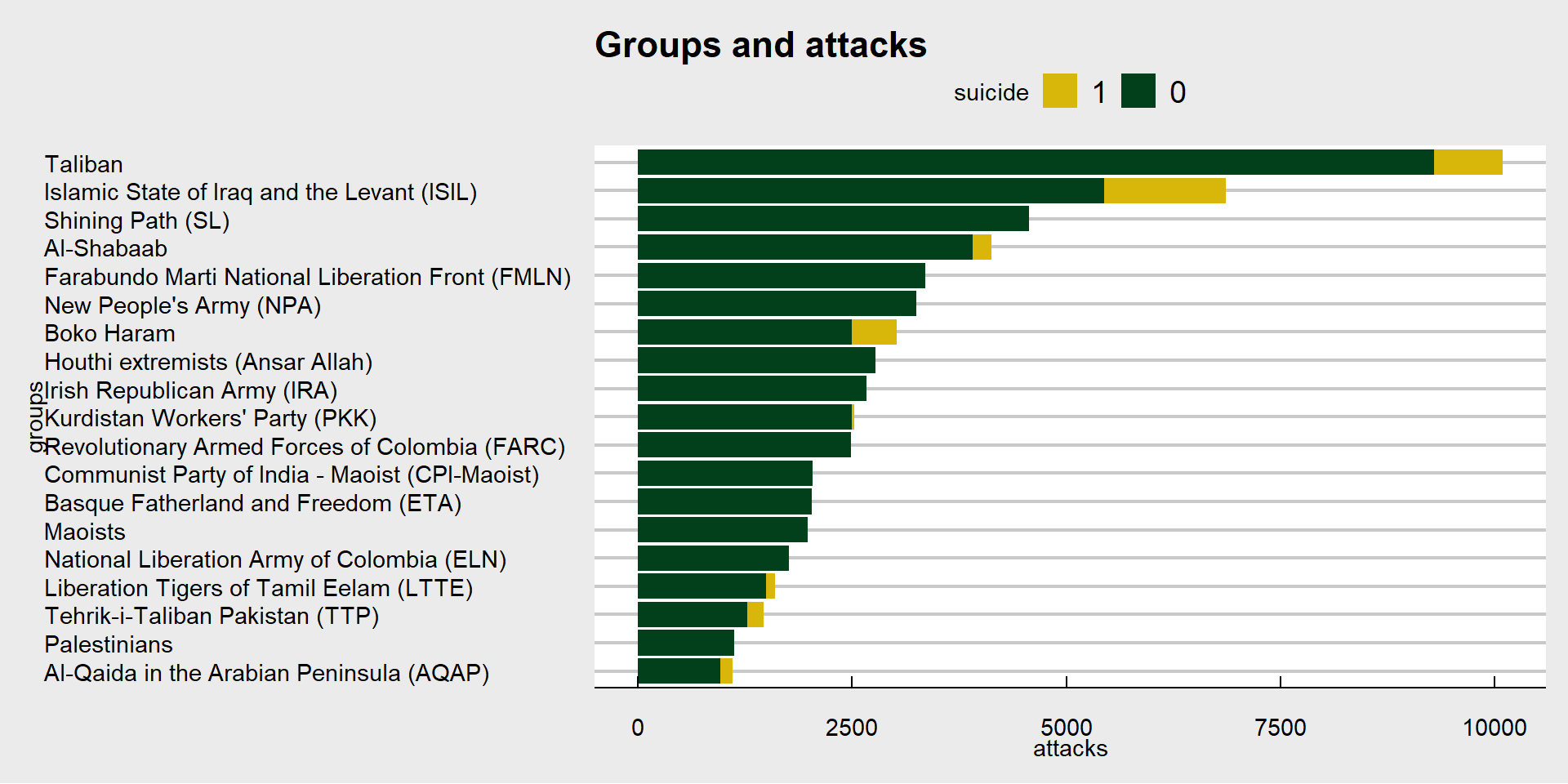
coord\_flip() +

labs(title = "Groups and attacks", x="groups", y="attacks", fill="suicide") +

theme\_economist\_white() +

scale\_fill\_manual(values = wes\_palette(n=2, "Cavalcanti1"))

## `summarise()` has grouped output by 'gname'. You can override using the `.groups` argument.

**** Disregarding the “Unknown” groups

* Taliban, ISIL, SL is reponsible for the most attacks.
* ISIL carried out the most suicide attacks.(23%)

## 2.8 Attack type by region

wp <- dt %>% select(1,2,3,4,9,11,13,14,15,27,28,30,59,83,85,99,102,117)

wp$imonth[wp$imonth==0] <- NA

wp$iday[wp$iday==0] <- NA

patkrg<- wp %>% group\_by(region\_txt, attacktype1\_txt) %>% count() %>%

ggplot(aes(region\_txt, n, fill=attacktype1\_txt)) +

geom\_bar(stat = "identity",position = "stack")+

scale\_fill\_manual(values = wes\_palette("Darjeeling1" ,n=9, type="continuous"))+

theme\_economist()+

scale\_x\_discrete(labels = **function**(x) stringr::str\_wrap(x, width = 0.8))+

labs(title = "Attack type by region",

x="region", y="num.", fill="attack type")

patkrg2<- wp %>% group\_by(region\_txt, attacktype1\_txt) %>% count() %>%

ggplot(aes(region\_txt, n, fill=attacktype1\_txt)) +

geom\_bar(stat = "identity",position = "fill")+

scale\_fill\_manual(values = wes\_palette("Darjeeling1" ,n=9, type="continuous"))+

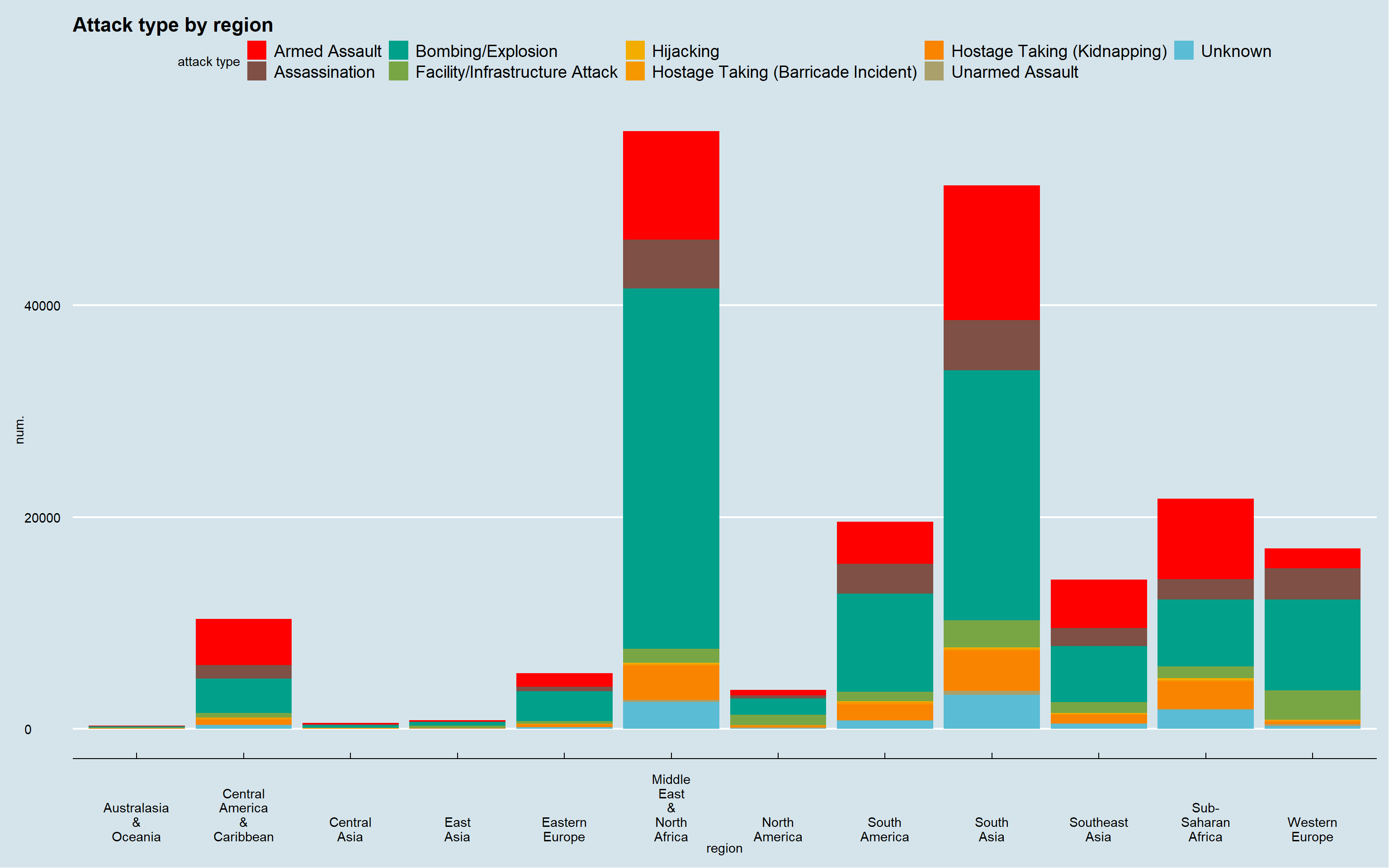
theme\_economist()+

scale\_x\_discrete(labels = **function**(x) stringr::str\_wrap(x, width = 0.8))+

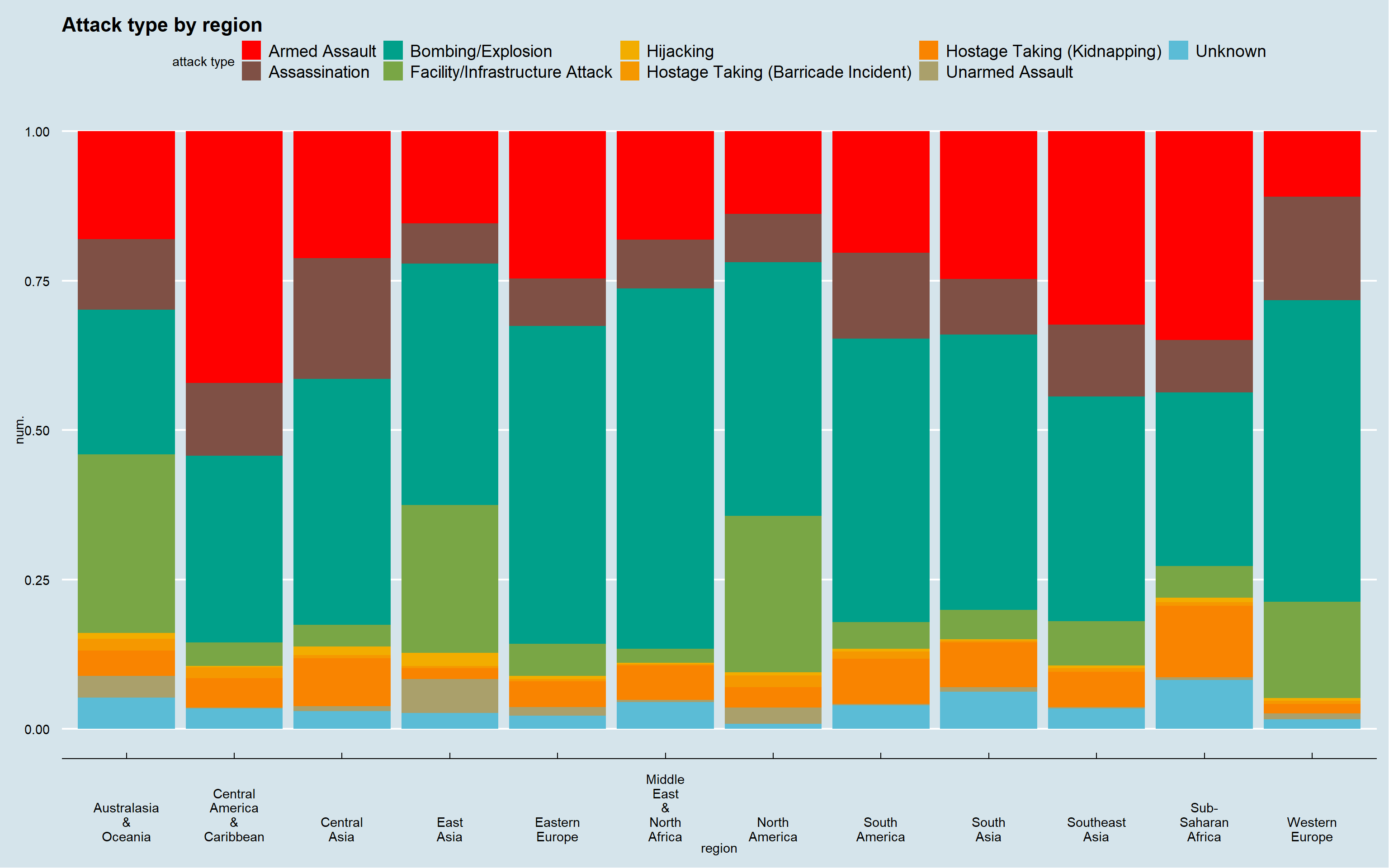
labs(title = "Attack type by region",

x="region", y="num.", fill="attack type")

patkrg

****

patkrg2

****

* Western Europe has more “Facility/Infrastructure Attack” (in number) than any other region.
* Bombing/Explosion is the most common attack type in Middle East & North Africa.

## 2.9 Attack type by group

Different groups might prefer different types of attack method. There are 3671 groups in the data. We’ll look at the groups with the most attacks.

wp %>% filter(gname %**in**% grp$gname)%>%

group\_by(gname, attacktype1\_txt) %>% count() %>%

ggplot(aes(gname, n, fill= attacktype1\_txt))+

geom\_bar(stat = "identity",position = "stack")+

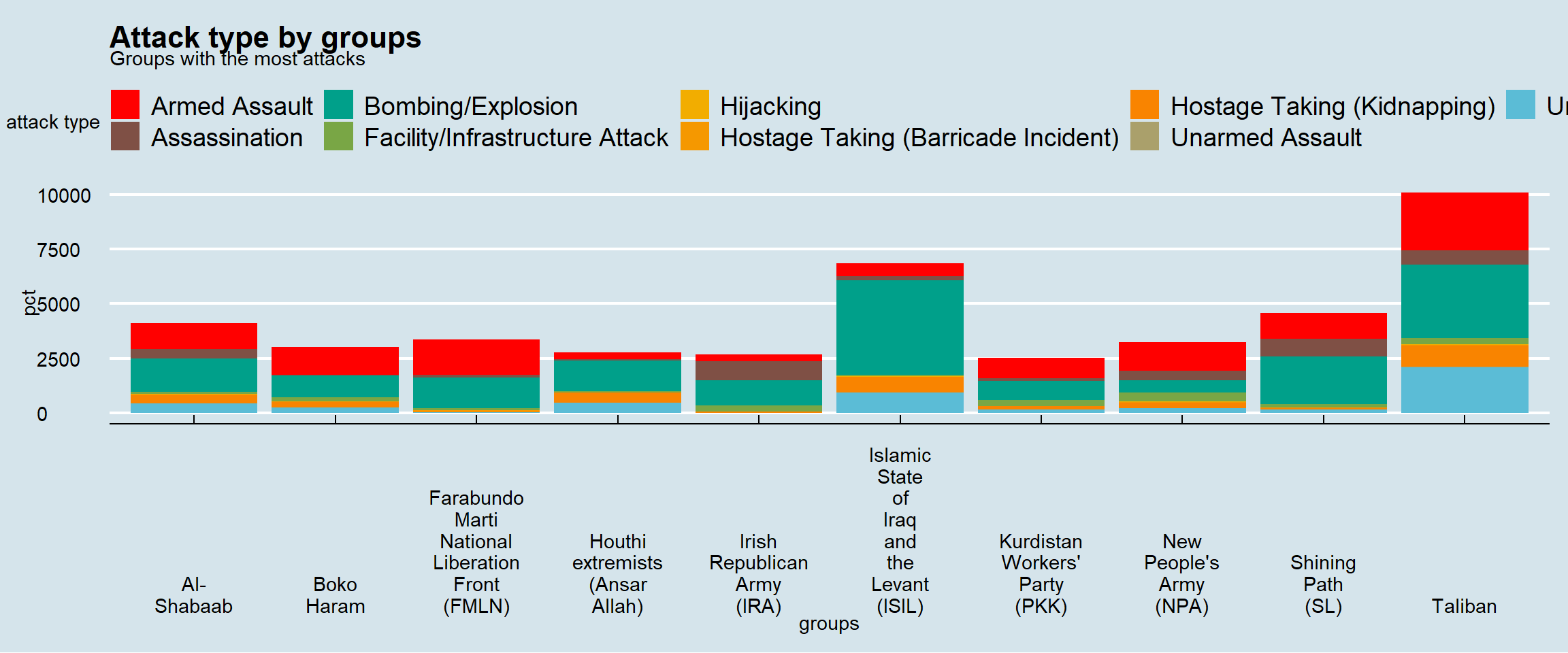
scale\_fill\_manual(values = wes\_palette("Darjeeling1",n=9, type="continuous"))+

theme\_economist()+

scale\_x\_discrete(labels = **function**(x) stringr::str\_wrap(x, width = 5))+

labs(title = "Attack type by groups", subtitle = "Groups with the most attacks",

x="groups", y="pct", fill="attack type")

****

* 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.

wp %>% filter(attacktype1\_txt=="Bombing/Explosion" & gname %**in**% grp$gname ) %>%

group\_by(gname, suicide) %>% count() %>% ungroup() %>% group\_by(gname) %>% mutate(pct=n/sum(n)) %>% filter(suicide==1) %>% arrange(desc(pct))

## # A tibble: 6 x 4

## # Groups: gname [6]

## gname suicide n pct

## <chr> <int> <int> <dbl>

## 1 Boko Haram 1 510 0.527

## 2 Islamic State of Iraq and the Levant (ISIL) 1 1369 0.317

## 3 Taliban 1 727 0.216

## 4 Al-Shabaab 1 192 0.125

## 5 Kurdistan Workers' Party (PKK) 1 27 0.0310

## 6 Houthi extremists (Ansar Allah) 1 2 0.00142

## 2.10 Number of death by attack type and region

wp %>% filter(!is.na(nkill)&attacktype1\_txt!="Unknown") %>%

group\_by(region\_txt,attacktype1\_txt) %>%

summarise(sumk=sum(nkill), event=n(), kperattack=sum(nkill)/n()) %>%

ggplot(aes(reorder(attacktype1\_txt, kperattack), kperattack))+

geom\_bar(aes(fill=region\_txt), stat = "identity")+

coord\_flip()+

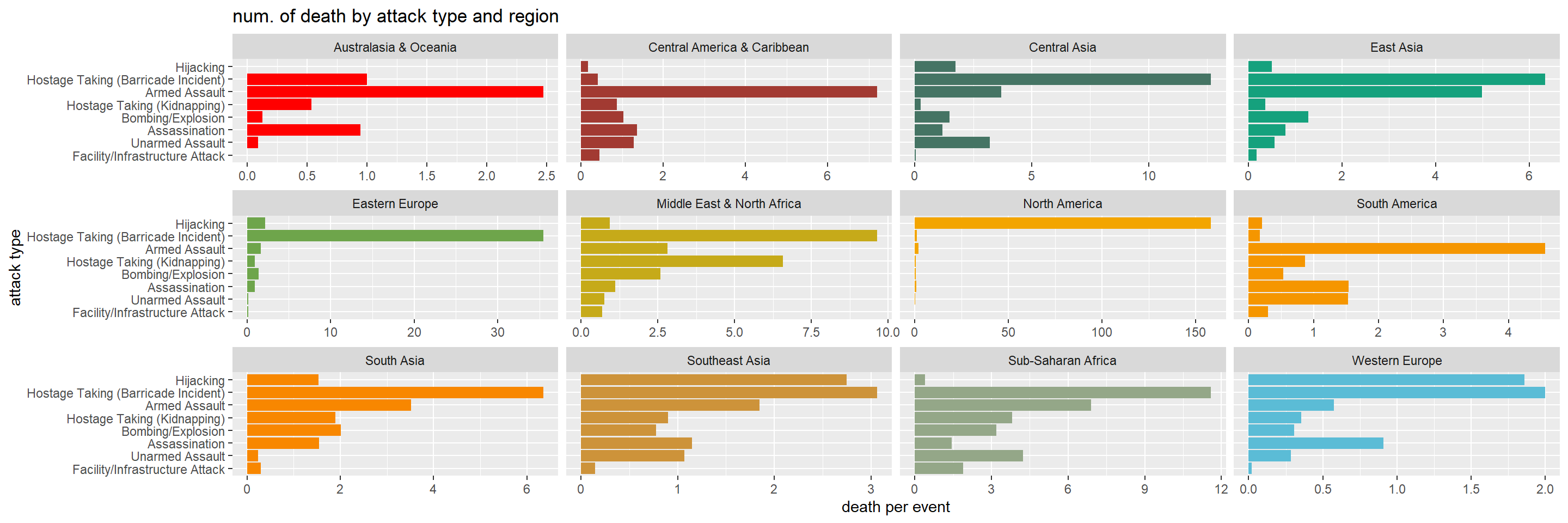
facet\_wrap(.~ region\_txt, ncol = 4, scales = "free\_x")+

labs(title = "num. of death by attack type and region", x="attack type", y="death per event")+

scale\_fill\_manual(values = wes\_palette("Darjeeling1", n=12, type = "continuous"))+

theme(legend.position = "none")

## `summarise()` has grouped output by 'region\_txt'. You can override using the `.groups` argument.

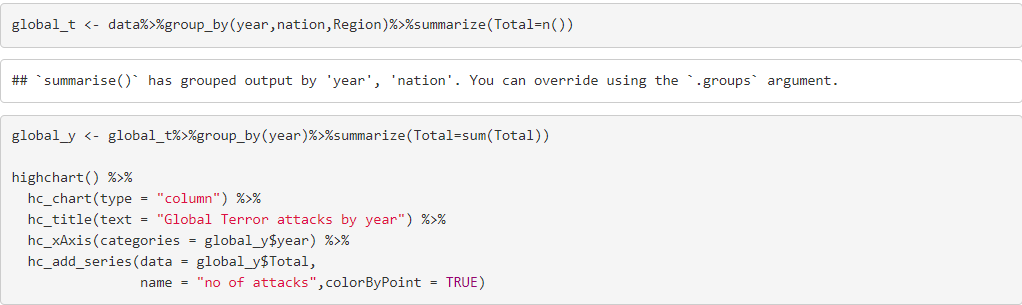
****

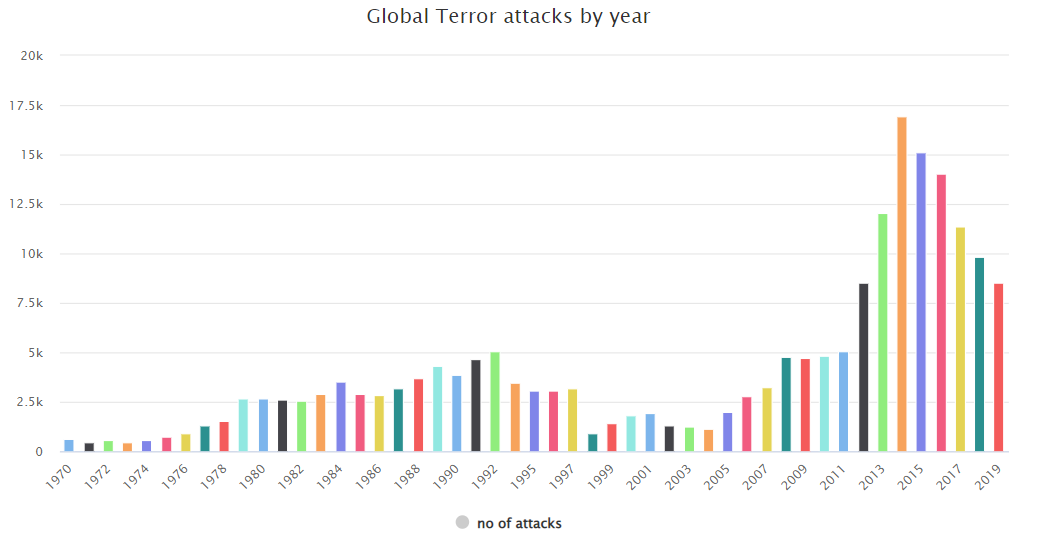
* Types of attack that cause the most death/attack is drastically different from region to region.
* Bombing (to my surprise) isn’t responsible for the most death/attack. Instead it’s armed assault and hostage taking in most region.
* Hostage taking has the most death/attack in East Asia, Eastern Europe, 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.

Data Preprocessing

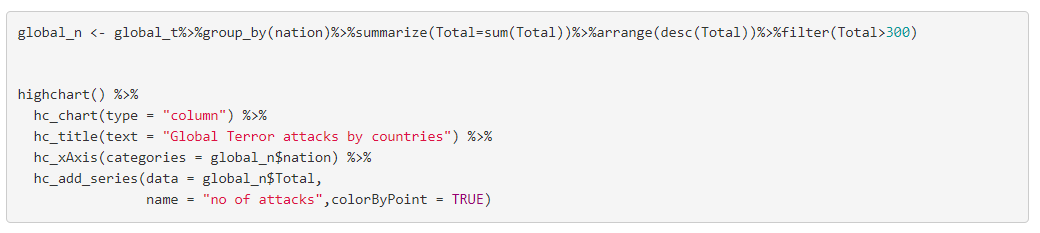


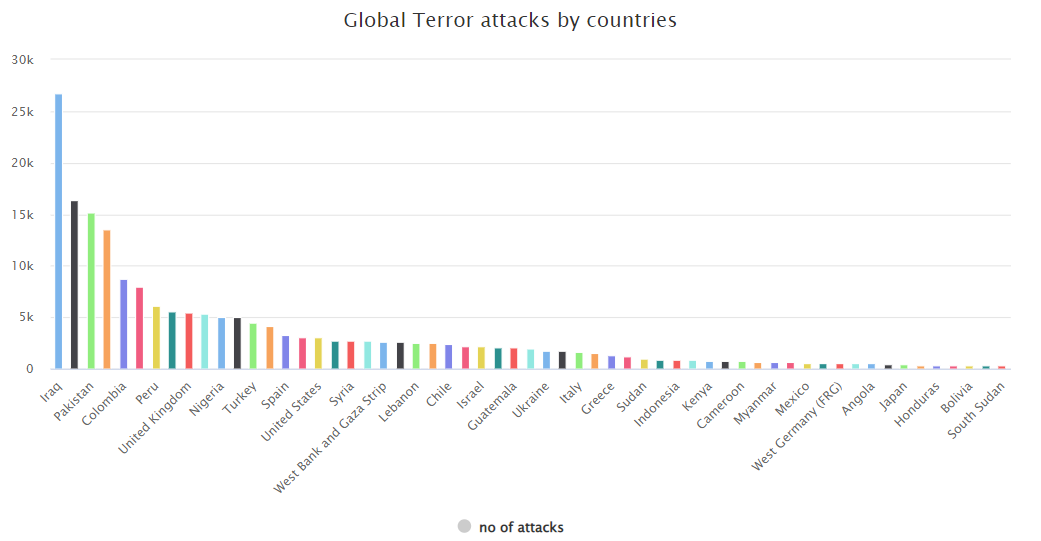
# Total No of Attacks by Year



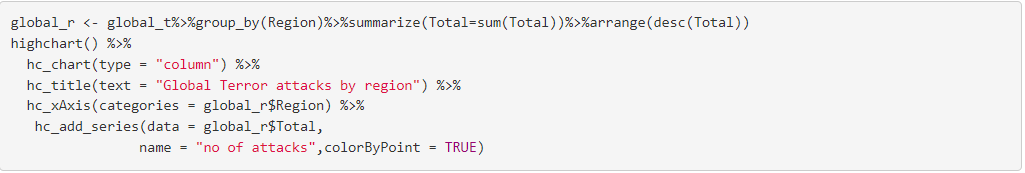


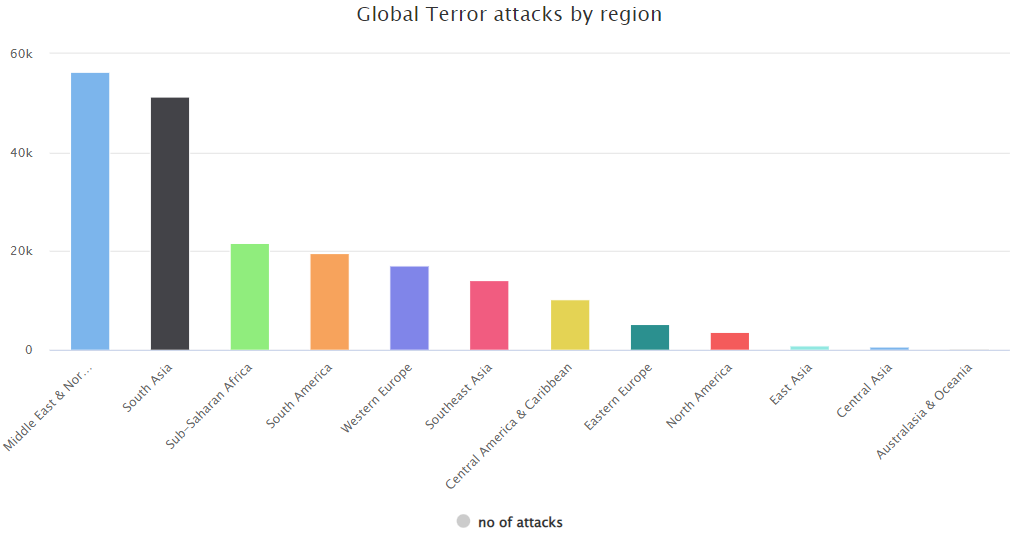
# Total No of Attacks by countries



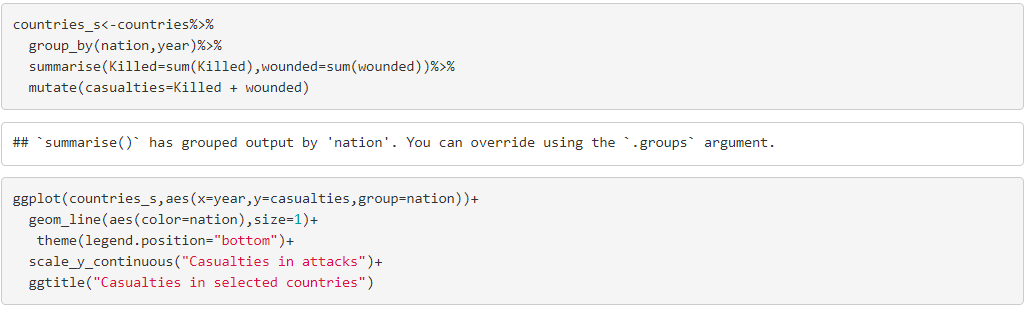


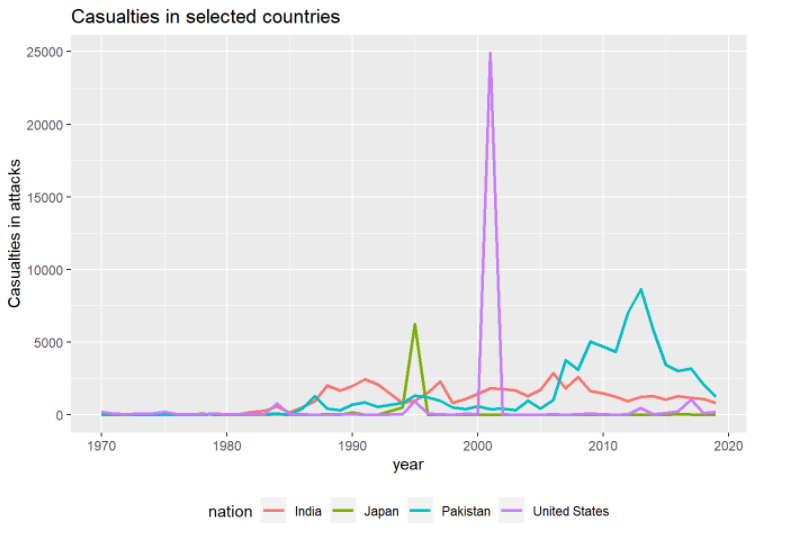
# Total No of Attacks by Region



****

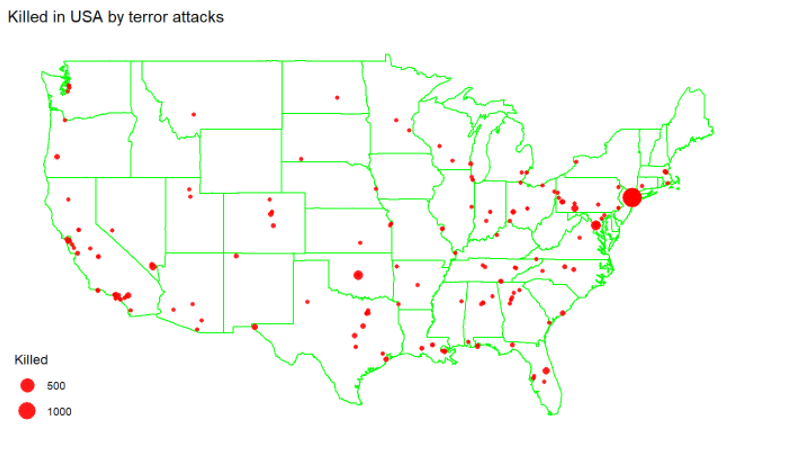
Casualties in attacks





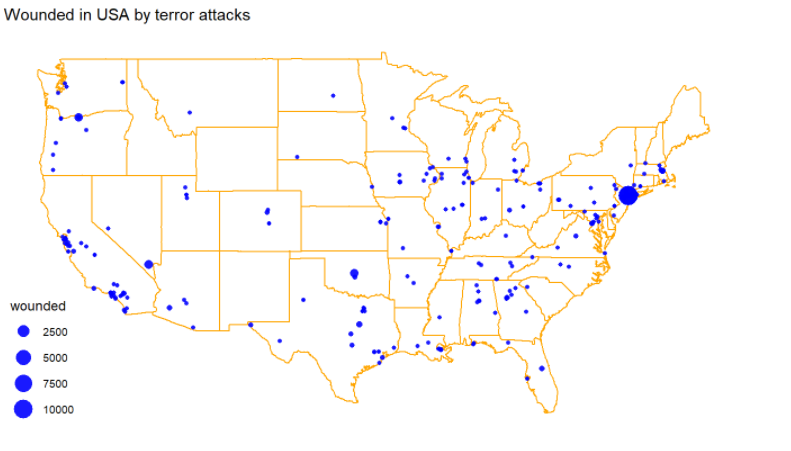
Killed in USA by terror attacks

****

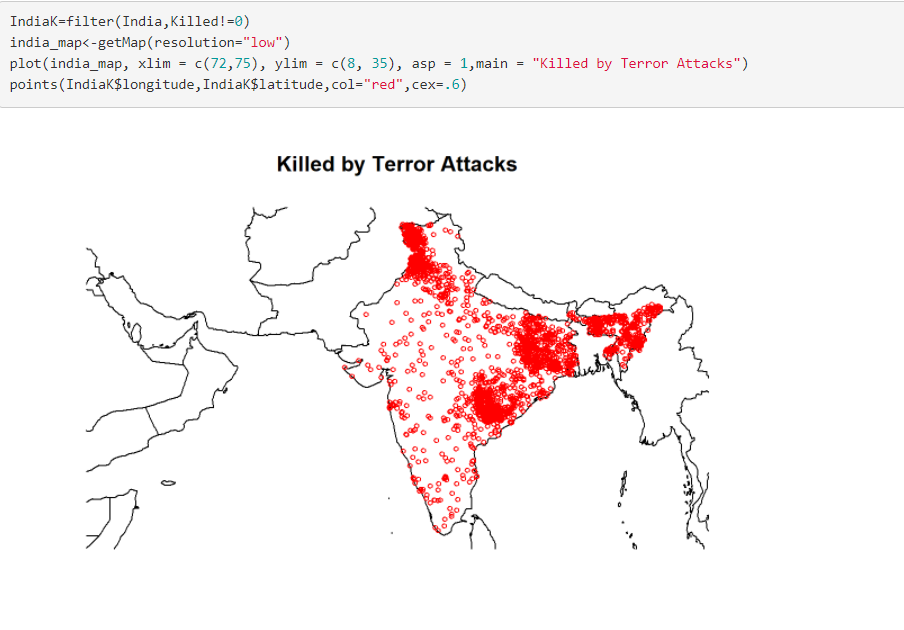
****

Wounded in USA by terror attacks

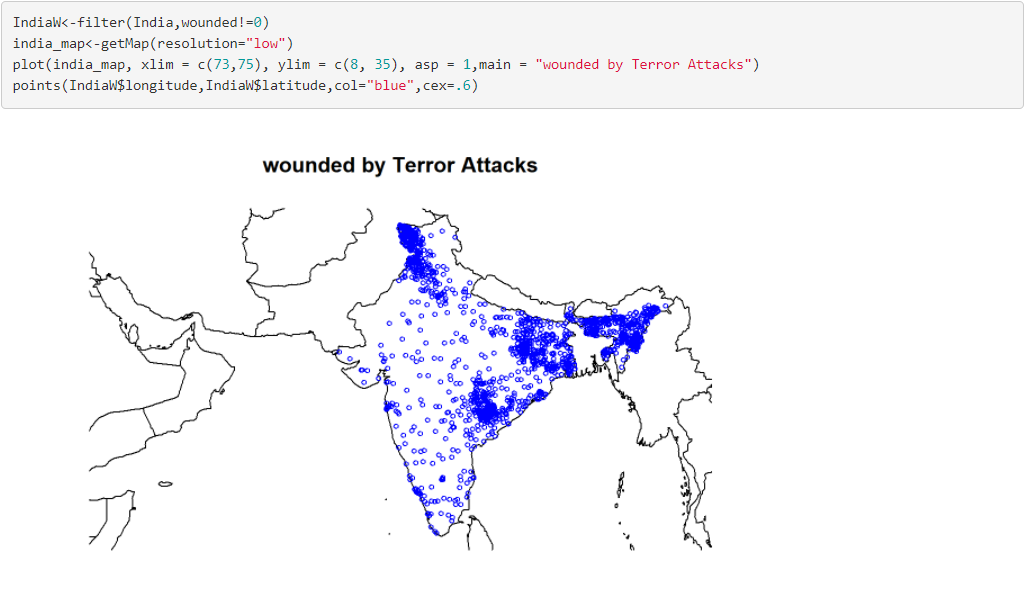
****



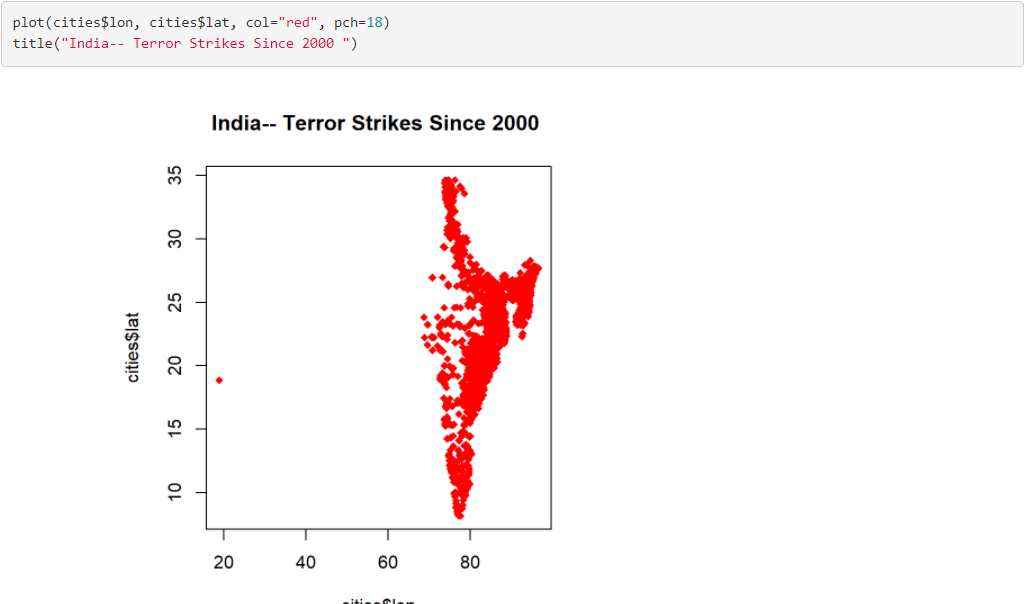
Killed by Terror attacks in india



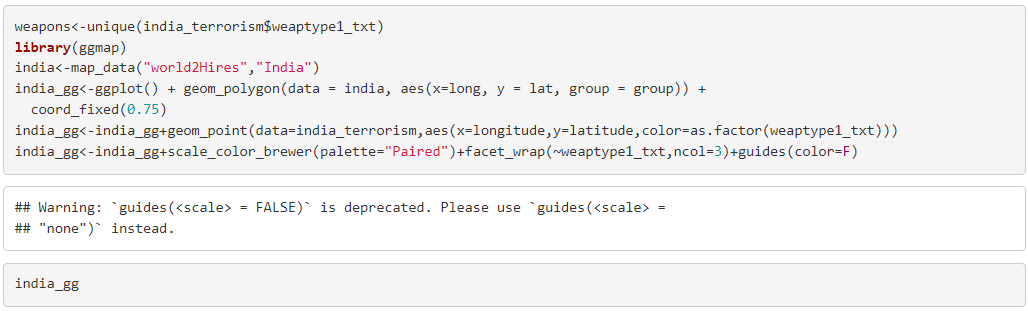
Wounded by Terrror attacks

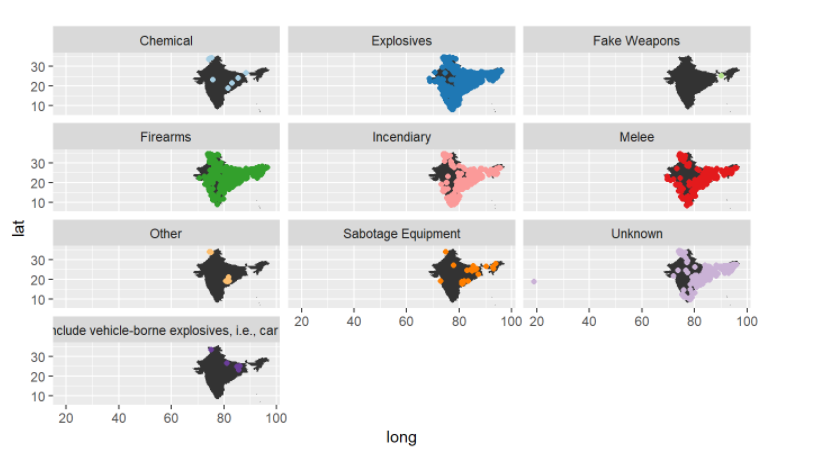


India-- Terror Strikes Since 2000

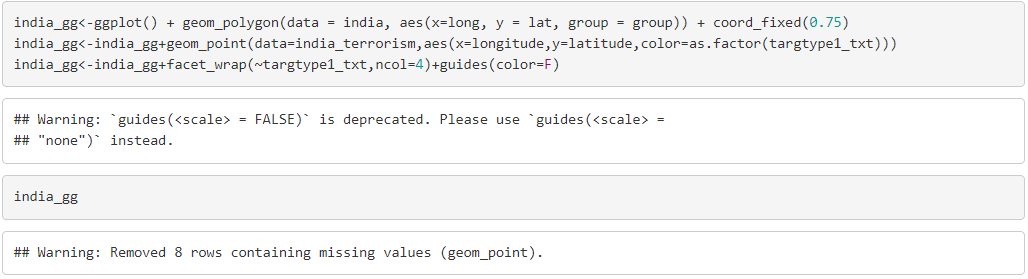


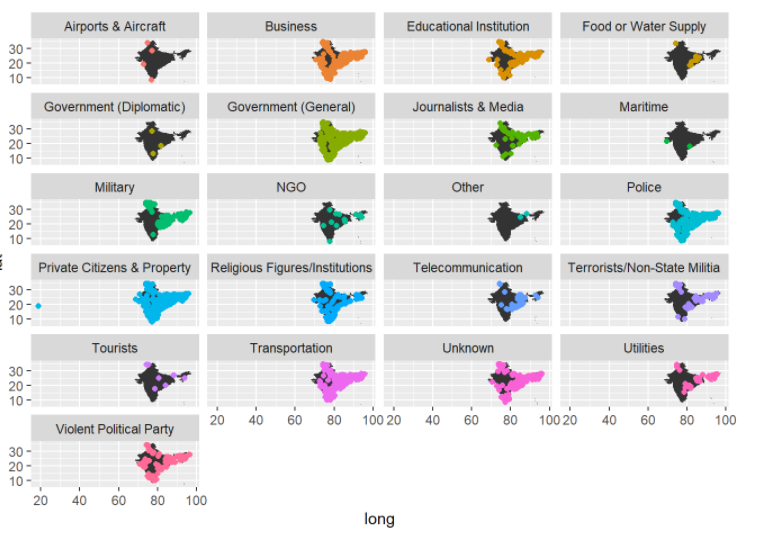
weapons used in india





# What Kind of People are Targeted By the Terror Outfits





**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 [28], security related policy-makers, international organizations hosting worldwide events, foreign investors and academic researchers for the purpose of understanding terrorism and its nature.

**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.](https://www.kaggle.com/ash316)

[com/ash316](https://www.kaggle.com/ash316) [Accessed: August 2019].

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

**Video Link:-**

[**https://drive.google.com/file/d/1AUViNK1ImWtOFc28N8h7ciAwZDgsHJDJ/view?usp=sharing**](https://drive.google.com/file/d/1AUViNK1ImWtOFc28N8h7ciAwZDgsHJDJ/view?usp=sharing)