

Data-driven Counter Terrorism Support: A Machine Learning Based
Approach to Achieve Actionable Intelligence

A Thesis
Presented to
The Division of Business & Economics
Berlin School of Economics and Law

In Partial Fulfillment
of the Requirements for the Degree
Master of Science (M.Sc.)
In
Business Intelligence & Process Management

Pranav Pandya

Immatriculation Number: 552590

July 2018



Hochschule für
Wirtschaft und Recht Berlin
Berlin School of Economics and Law

Approved for the Division
(Business Intelligence & Process Management)

Prof. Dr. Markus Loecher

Prof. Dr. Markus Schaal

Sworn Declaration

I, Pranav Pandya hereby formally declare that the work presented herein is genuine work done originally by me and has not been published or submitted elsewhere for the requirement of any degree programme or for any purpose. I am aware that the use of quotations, or of close paraphrasing, from books, magazines, newspapers, the internet or other sources, which are not marked as such, will be considered as an attempt at deception, and that the thesis will be graded as a fail.

I confirm that any literature, data, or works done by others and cited within this report has been given due acknowledgement and listed in the reference section.

Pranav Pandya
Berlin, July 2018

Acknowledgements

I want to express my deep sense of gratitude to my supervisor Prof. Dr. Markus Loecher (Berlin School of Economics & Law). Words are inadequate in offering my thanks to him for his encouragement and cooperation in carrying out this research project. His able guidance and useful suggestions helped me in completing the project work, in time.

Finally, yet importantly, I would like to express my heartfelt thanks to my beloved mother for her blessings, encouragement and wishes for the successful completion of this research project.

Table of Contents

Introduction	1
Definition of Terrorism	1
Problem Statement	2
Research Design and Data	3
Policy and Practice Implications	4
Chapter 1: Literature review	5
1.1 Intelligence Disciplines	5
1.2 OSINT and Data Relevance	7
1.2.1 Open-source Databases on Terrorism	7
1.3 What's Important in Terrorism Research Analysis?	8
1.3.1 Primary vs Secondary Sources	8
1.3.2 Use of Statistical Analysis	9
1.4 Overview of Prior Research	9
1.4.1 Harsh Realities	10
1.4.2 Review of Relevant Literature	11
1.4.3 GTD and Machine Learning in Previous Research	13
1.5 Literature Gap and Relevance	14
Chapter 2: Impact Analysis	16
2.1 Data Preparation	16
2.2 Global Overview	19
2.3 The Top 10 Most Active and Violent Groups	22
2.4 The Major and Minor Epicenters	26
Chapter 3: Statistical Hypothesis Testing	31
3.1 Data Preparation	31
3.2 Correlation Test	31
3.3 Hypothesis Test: Fatalities vs Groups	33
3.3.1 ANOVA Test	35
3.3.2 PostHoc Test	35
3.3.3 Interpretation	36
Chapter 4: Pattern Discovery	40
4.1 Data preparation	41

4.2	Islamic State (ISIL)	41
4.2.1	Explanation of key terms	42
4.2.2	Apriori model summary	42
4.2.3	Top 5 patterns (ISIL)	43
4.2.4	Network graph (ISIL)	44
4.3	Taliban	46
4.3.1	Apriori model summary	46
4.3.2	Top 5 patterns (Taliban)	46
4.3.3	Network graph (Taliban)	48
4.4	Boko Haram	48
4.4.1	Apriori model summary	48
4.4.2	Top 5 patterns (Boko Haram)	49
4.4.3	Network graph (Boko Haram)	51
Chapter 5: Time-series Forecasting	52
5.1	Afghanistan (Predict future attacks)	52
5.1.1	Data preparation	52
5.1.2	Seasonality analysis	53
5.1.3	Correlation test	56
5.1.4	Modelling	57
5.1.5	Evaluating models' Performance	60
5.1.6	Ensemble	62
5.1.7	Forecast future number of attacks	63
5.2	Iraq (Predict future fatalities)	66
5.2.1	Data preparation	66
5.2.2	Seasonality analysis	67
5.2.3	Correlation test	69
5.2.4	Modelling	69
5.2.5	Ensemble	71
5.2.6	Forecast future fatalities	71
5.3	SAHEL Region (Predict future attacks)	75
5.3.1	Data preparation	75
5.3.2	Seasonality analysis	75
5.3.3	Correlation test	77
5.3.4	Modelling	78
5.3.5	Ensemble	79
5.3.6	Forecast future attacks	80
Chapter 6: Classification Approach	84
6.1	Overview of target variables	84
Conclusion	85
Appendix A: The First Appendix	86

Appendix B: The Second Appendix, for Fun	88
References	89

List of Tables

2.1	Short description of important variables	18
2.2	Terrorism Epicenters in North America and Eastern Europe	26
3.1	Posthoc test (lsd, scheffe, bonf)	38
3.2	Post hoc test with Tukey HSD for Pair of Groups	39
5.1	Scope of Analaysis	52
5.2	Performance comparison of all models (Afghanistan)	61
5.3	Table of Predicted Future Attacks in Afghanistan	66
5.4	Performance comparison of all models (Iraq)	71
5.5	Table of Predicted Future Fatalities in Iraq	74
5.6	Performance comparison of all models (SAHEL Regioin)	79
5.7	Table of Predicted Future Attacks in SAHEL Region	83

List of Figures

1	Terrorist attacks around the world between 1970-2016	2
1.1	Use of statistics in terrorism research between 2007 to 2016	10
2.1	Attack Frequency by Year and Region	19
2.2	Trend in type of attack in all incidents globally	20
2.3	Trend in type of weapon used in all incidents globally	21
2.4	Trend in intended targets in all incidents globally	22
2.5	Top 10 Most Active and Violent Groups	24
2.6	Attack Frequency from Top 10 Groups	25
2.7	Characteristics of top 10 groups	25
2.8	The Major and Minor Epicenters of Terrorism	28
2.9	Terrorist Group and Impacted Cities	30
3.1	Correlation web plot	33
3.2	Boxplot: Group vs Fatalities	34
4.1	Association Rules in ISIL Group	44
4.2	Network Graph of Discovered Patterns- ISIL Group	45
4.3	Association Rules in Taliban Group	47
4.4	Network Graph of Discovered Patterns- Taliban Group	48
4.5	Association Rules in Boko Haram Group	50
4.6	Network Graph of Discovered Patterns- Boko Haram Group	51
5.1	Attack Frequency by Year- Afghanistan	53
5.2	Seasonal Pattern Within Year- Afghanistan	54
5.3	Seasonal Pattern (boxplot)- Afghanistan	55
5.4	Time-series Decomposition- Afghanistan	55
5.5	Correlation Test	56
5.6	Auto Arima: Residuals	58
5.7	Auto Arima: Actual vs Fitted vs Forecasted	58
5.8	Neural Net: Actual vs Fitted vs Forecasted	59
5.9	TBATS: Actual vs Fitted vs Forecasted	60
5.10	ETS: Actual vs Fitted vs Forecasted	60
5.11	Predicted Number of Attacks in Afghanistan (plots)	63
5.12	Predicted Number of Attacks in Afghanistan (plots)	64

5.13 Predicted Number of Attacks in Afghanistan (plots)	64
5.14 Predicted Number of Attacks in Afghanistan (plots)	65
5.15 Fatalities Frequency by Year- Iraq	68
5.16 Seasonality Plots - Iraq	68
5.17 Correlation Test	69
5.18 Predicted Number of Fatalities in Iraq (plots)	72
5.19 Predicted Number of Fatalities in Iraq (plots)	72
5.20 Predicted Number of Fatalities in Iraq (plots)	73
5.21 Predicted Number of Fatalities in Iraq (plots)	73
5.22 Attack Frequency by Year- SAHEL Region	76
5.23 Seasonal Pattern (heatmap) - SAHEL Region	76
5.24 Seasonality Pattern (boxplot) - SAHEL Region	77
5.25 Correlation Test	78
5.26 Predicted Number of Attacks in SAHEL Region (plots)	80
5.27 Predicted Number of Attacks in SAHEL Region (plots)	81
5.28 Predicted Number of Attacks in SAHEL Region (plots)	81
5.29 Predicted Number of Attacks in SAHEL Region (plots)	82

Abstract

This research project uses historical data of terrorist attacks that took place around the world between 1970 to 2016 from open-source Global Terrorism Database (GTD) and aims to make sense of observed patterns by translating it into actionable intelligence. The research is split into three distinct categories in order to evaluate cases of over 170,000 terrorist attacks. The first part is exploratory data analysis which is intended to examine patterns in terrorist attacks from various perspectives and with statistical analysis. Based on the analysis and findings from first exploratory data analysis, the next parts in this research makes use of machine learning algorithms to derive actionable insights. An observation from previous research in this field suggests that forecasting methods and classification models were mostly applied on the world level and on yearly data however little/no research is available to forecast future attacks or fatalities based on seasonal patterns (i.e. monthly and quarterly) and on country level.

As an extension to existing research, this research makes use of time series forecasting models to predict the future attacks and fatalities on micro level by establishing methodology to prepare and evaluate seasonal components from data. Similarly in the classification modelling part, previous research lacks use of algorithms that are recently developed and that (practically) out performs traditional algorithms such as random forest, native gbm or xgboost. In the third part of this research, cutting edge gradient boosting algorithm i.e. lightgbm is used to predict the class probability of an attack. In simple words, the classification model is trained on historical data, evaluated on validation data and then used to predict class probability on test data. The important insights from this part is to find causal variables for a particular attack for chosen target variable such as whether or not an attack will be a suicide attack, extended attack, successful attack, part multiple attacks or classifying political-economical-social-religious goal behind the attack. Apart from causal variables, this research also makes use of explainer objects from the model to justify reason behind each prediction/decision from the trained algorithm. In line with this scientific research, a shiny app is developed to make this reproducible research interactive and handy. This app is in fact the vital part of this research and is intended to provide interactive platform for all three parts of this research project (i.e. exploratory data analysis, time-series forecasting and classification models) to target audience such as individual researchers, counter terrorism agencies and government authorities.

Dedication

I dedicate this thesis to two people who means a lot to me. First and foremost, to my mother Anjana P. Pandya who has been constant source of inspiration for me. I am thankful to you for your constant support and blessings which helps me achieve set goals of my life.

Secondly, my maternal grandfather late Shri Upendrabhai M. Joshi who always believed in my ability. You made a garden of heart and planted all the good things which gave my life a start. You encouraged me to dream by fostering and nurturing the seeds of self-esteem. You taught me the difference between right and wrong and made pathway which will last a lifetime long. You have gone away forever from this world but your memories are and will always be in my heart.

Introduction

Today, we live in the world where terrorism is becoming a primary concern because of the growing number of terrorist incidents involving civilian fatalities and infrastructure damages. The ideology and intentions behind such attacks is indeed a matter of worry. Living under the constant threat of terrorist attacks in any place is no better than living in jungle and worrying about which animal will attack you and when. An increase in number of radicalized attacks around the world is a clear indication that terrorism transitioning to from a place to an idea however existence of specific terror group and their attack characteristics over the period of time can be vital to fight terrorism and to engage peace keeping missions effectively. Having said that number terrorist incidents are growing these days, availability of open-source data containing information of such incidents, recent developments in machine learning algorithms and technical infrastructure to handle large amount of data open ups variety of ways to turn information into actionable intelligence.

Definition of Terrorism

Terrorism in broader sense includes state sponsored and non-state sponsored terrorist activities. Scope of this research is limited to **non-state sponsored** terrorist activities only. Non-state actors in simple words mean entities that are not affiliated, directed or funded by the government and that exercise significant economic, political or social power and influence at a national and international level upto certain extent (NIC, 2007). An example of non-state actors can be NGOs, religious organizations, multi national companies, armed groups or even a online (Internet) community. ISIL is the prime example of non-state actor which falls under armed groups segment.

Global Terrorism Database (National Consortium for the Study of Terrorism and Responses to Terrorism (START), 2016) defines terrorist attack as a threatened or actual use of illegal force and violence by a non-state actor to attain a political, economic, religious or social goal through fear, coercion or intimidation.

This implies that three of the following attributes are always present in each events of our chosen dataset:

- The incident must be intentional – the result of a conscious calculation on the

part of a perpetrator.

- The incident must entail some level of violence or immediate threat of violence including property violence, as well as violence against people.
- The perpetrators of the incidents must be sub-national actors.

Problem Statement

Nowadays, data is considered as the most valuable resource and machine learning makes it possible to interpret complex data however most use cases are seen in business context such as music recommendation, predicting customer churn or finding probability of having cancer. With recent development in machine learning algorithms and access to open source data and software, there are plenty of opportunities to correctly understand historical terrorist attacks and prevent the future conflicts. In the last decade, terrorist attacks have been increased significantly as shown in the plot below:

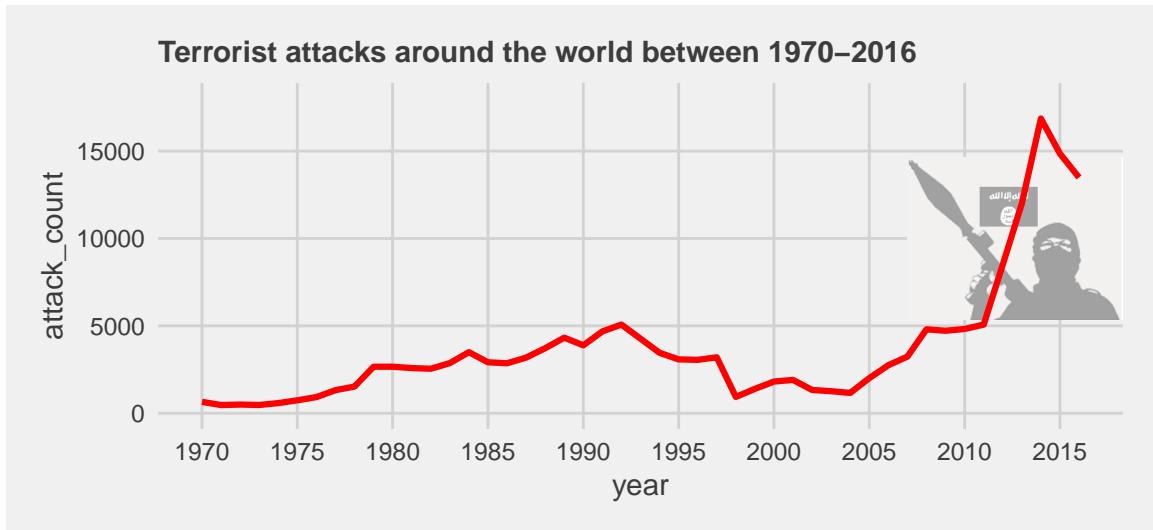


Figure 1: Terrorist attacks around the world between 1970-2016

After September 2001 attacks, USA and other powerful nations have carried out major operations to neutralize the power and spread of known and most violent terrorist groups within targeted regions such as in Afghanistan, Iraq and most recently in Syria. It's also worth mentioning that United Nations already have ongoing peacekeeping missions in conflicted regions around the world for a long time. However number of terror attacks continues to rise and in fact, it is almost at peak in the last 5 years. This leads to a question why terrorism is becoming unstoppable despite the continued efforts. I argue that the cause of problem is group's ideology and political motivations itself. Understanding and interpreting the attack characteristics of relevant groups in line with their motivations to do so can reflect bigger picture. An extensive research by

(Heger, 2010) supports this argument and suggests that a group's political intentions are revealed when we examine who or what it chooses to attack.

Research Design and Data

This research employs quantitative research methodology and focuses on reproducibility. A shiny app is developed in line with the analysis to allow target audience to have free choice of sampling from population (data). Scope of work for this research limited to non-state actors only. In the initial part, research focuses on impact analysis on global level to identify vulnerable regions and corresponding patterns in type of attack, type of weapon and type of target over the period of time. Within this exploratory data analysis part, we also determine and examine characteristics of top ten most active and deadliest groups. The major part of this research is based on machine learning algorithms; specifically time-series forecasting models and classification algorithms in order to achieve actionable intelligence toward counter terrorism support as primary objective of this research.

According to (Samuel, 1959), A well-known researcher in the field of artificial intelligence who coined the term “machine learning”, defines machine learning as a “field of study that gives computers the ability to learn without being explicitly programmed”. It is subset of artificial intelligence which enables computers to learn from experience in order to create inference over a possible outcome used later to take a decision.

Second part of this research begins with time-series analysis where objective is to determine future number of attacks and number of fatalities by seasonality i.e. months and quarters for chosen country/ countries. At first we examine seasonal patterns and perform correlation analysis. In the next part, we use different forecasting models and evaluate their performance on out of fold set i.e. validation data with various metrics. Within the context of time-series analysis, we generate forecasts for future periods for each model and then use ensemble method to make final predictions. Third part of this research makes use of gradient boosting algorithm lightgbm (which is recently developed by Microsoft and open-source) for classification task. The underlying idea is to predict class probability for chosen response variable. For example, what is the possibility of attack being a suicide attack, whether or not an attack will last longer than 24 hours or finding possibility of attack being part of multiple attacks. For each response variable, we find the most important features from our trained model. In the last part, we make use of explainer object to validate trustworthiness of our model. This is particularly helpful in understanding reasons behind predictions.

Data

This research project uses historical data of terrorist attacks that took place around the world between 1970 to 2016 from open-source Global Terrorism Database (GTD) as a primary source of data. It is currently the most comprehensive unclassified database on terrorist events in the world and contains information on over 170,000

terrorist attacks. It contains information on the date and location of the incident, the weapons used and nature of the target, the number of casualties and the group or individual responsible if identifiable. Total number of variables are more than 120 in this data. One of the main reason for choosing this database is because 4,000,000 news articles and 25,000 news sources were reviewed to prepare this data from 1998 to 2016 alone (National Consortium for the Study of Terrorism and Responses to Terrorism (START), 2016).

Main data is further enriched with country and year wise socio-economical conditions, arms import/export details and migration details from World Bank Open Data to get multi-dimensional view for some specific analysis. This additional data falls under the category of early warning indicators (short term and long term) and potentially linked to the likelihood of violent conflicts as suggested by the researcher (Walton, 2011) and (Stockholm International Peace Research Institute, 2017).

An important aspect of this research is use of open-source data and open-source software i.e R. The reason why media-based data source is chosen as primary source of data and is because journalists are usually the first to report and document such incidents and in this regard, first hand information plays significant role in quantitative analysis. Since the source of data is from publically available sources, the term “intelligence” refers to open-source intelligence (OSINT) category.

Policy and Practice Implications

This research project is aimed to provide data-driven counter terrorism support and contributes positively to the counter terrorism policy. Outcome of used machine learning algorithms can serve as an early warning system to address the rare events related to armed conflict, civil war and other political violence around the world from non-state actors. Research findings and insights will serve as an actionable intelligence and it will help policy makers or authorities to take necessary steps in time to prevent such situations. Alongwith time-series forecasting, the key take away from this research is significance of causal variables depending on group's motivation/intentions behind attacks.

Chapter 1

Literature review

I use structured approach to assess theoretical framework for the research context in order to narrow down and examine relevant literature. Terrorism research in broad context suggests that intelligence toward counter terrorism support comes in many forms. The primary objective of this research is achieve actionable intelligence through machine learning approach so it is important to identify the type of intelligence. In this chapter, first we distinguish between intelligence disciplines, justify reliability and relevance of chosen data and then review the relevant literature in counter terrorism research within machine learning context.

1.1 Intelligence Disciplines

An extensive research by (Tanner, 2014) suggests that establishing methodologies for collecting intelligence is important for authorities/ policy makers to combat terrorism. The Intelligence Officer's Bookshelf from CIA¹ recognizes Human Intelligence (HUMINT), Signals Intelligence (SIGINT), Geospatial Intelligence (GEOINT), Measurement and Signature Intelligence (MASINT) and Open Source Intelligence (OSINT) as five main disciplines of intelligence collection (Lowenthal & Clark, 2015).

Human Intelligence (HUMINT)

As the name suggests, HUMINT comes from human sources and remains synonymous with espionage and clandestine activities. This is one of the oldest intelligence techniques which uses covert as well as overt individuals to gather information. Examples of such individuals can be diplomats, special agents, field operatives or captured prisoners (The Interagency OPSEC Support Staff, 1996). According to (CIA, 2013),

¹<https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/csi-studies/studies/vol-60-no-1/pdfs/Peake-I0-Bookshelf-March-2016.pdf>

human intelligence plays vital role in developing and implementing U.S. national security policy and foreign policy to protect U.S. interests.

Signals Intelligence (SIGNIT)

SIGNIT is derived from electronic transmissions such as by intercepting communications between two channels/ parties. In the US, National Security Agency (NSA) is primarily responsible for signals intelligence (Groce, 2018). An example of SIGNIT is NSAs mass surveillance program PRISM which is widely criticized due to dangers associated with it in terms of misuse.

Edward Snowden, a former NSA contractor and source of the Guardian's investigation on systematic data trawling by the US government, suggests that, "The reality is this: if an NSA, FBI, CIA, DIA [Defence Intelligence Agency], etc analyst has access to query raw SIGINT [signals intelligence] databases, they can enter and get results for anything they want. Phone number, email, user id, cell phone handset id (IMEI), and so on – it's all the same. The restrictions against this are policy based, not technically based, and can change at any time." (Siddique, 2013)

Geospatial Intelligence (GEOINT)

GEOINT makes use of geo-spatial analysis and visual representation of activities on the earth to examine suspicious activities. This is usually carried out by observation flights, UAVs, drones and satellites (Brennan, 2016).

Measurement and Signature Intelligence (MASINT)

MASINT is comparatively less known methodology however it's becoming extremely important when concerns about WMDs (Weapons of Mass Destruction) are growing. This approach performs analyses of data from specific sensors for the purpose of identifying any distinctive features associated with the source emitter or sender. This analysis serves as a scientific and technical intelligence information. An example of MASINT is FBI's extensive forensic work that helps detecting traces of nuclear materials, chemical and biological weapons (Groce, 2018).

Open Source Intelligence (OSINT)

OSINT is relatively new approach that focuses on publicly available information and sources such as newspaper articles, academic records and open-source data made available to public from government or researchers. The key advantage of open source intelligence is accessibility and makes it possible for individuals researchers to contribute positively toward counter terrorism support as a part of community. It is important to note that reliability of data source can be complicated and thus requires review in order to be a use to policy makers (Groce, 2018; Tanner, 2014).

Focus and scope of work for this research is limited to Open Source Intelligence only.

1.2 OSINT and Data Relevance

Despite the huge (and technically limitless) potential for counter terrorism support, the reason as to why open source intelligence is often reviewed and analysed before it can be used by policy makers is because of complications related to authenticity of data source and methodology used to compile data for hypothesis testing by a researcher. In simple words what it means is, it is extremely important for policy makers to ensure that there is no selection bias or cherry-picking from a researcher to claim the success of particular theory or results (Brennan, 2016). A research paper from (Geddes, 1990/ed) namely “How the Cases You Choose Affect the Answers You Get: Selection Bias in Comparative Politics” explains the danger of biased conclusions when the cases that have achieved the outcome of interest are studied. This clearly forms the need for reproducible research and allows authorities to set the standard/mechanism to safe guard against selection bias. This is particularly important in terrorism research. This critical issue can be taken care by codes/ scripts shared through git repositories. Nowadays, making use of tools such as rmarkdown and bookdown to deliver reproducible research (Bauer, 2018; Xie, 2016) makes it even more easy to identify selection bias.

In one of the most recent article which reviews research methodologies and data in terrorism between 2007-2016, researcher (Schuurman, 2018) argues that the tendency to design research based on available data rather than gathering data required to address the research question is a matter of concern in terms of quality of quantitative research being conducted.

1.2.1 Open-source Databases on Terrorism

In the context of terrorism research, there are many databases available for academic research. Such databases extract and compile information from variety of sources (mainly open-source/ publicly available sources such as news articles) on regular interval and makes it easy to use for research. Some of the well-known databases that are open-source and widely used in academic research for counter terrorism support are as below:

1. Global Terrorism Database (GTD)²

- Currently the most comprehensive unclassified database on terrorist events in the world
- maintained by researchers at the National Consortium for the Study of Terrorism and Responses to Terrorism (START), headquartered at the University of Maryland in the USA

2. Armed Conflict Location and Event Data Project (ACLED)³

²<http://www.start.umd.edu/gtd/about/>

³<https://www.acleddata.com/data/>

- provides realtime data on all reported political violence and protest events however limited to developing countries i.e. Africa, South Asia, South East Asia and the Middle East

3. UCDP/PRIO Armed Conflict Database⁴

- a joint project between the UCDP and PRIO that records armed conflicts from 1946–2016
- maintained by Uppsala University in Sweden

4. SIPRI Databases⁵

- provides databases on military expenditures, arms transfers, arms embargoes and peacekeeping operations
- maintained by Stockholm International Peace Research Institute

In order to address the research objective, I find the Global Terrorism Database most relevant and it is the primary source of data for this research. As mentioned in Research Design and Data section, main data is further enriched with world development indicators for each countries by year from World Bank Open Data.⁶

1.3 What's Important in Terrorism Research Analysis?

Aim of any research can be seen as an effort toward creating new knowledge, insights or a perspective. In this regard, careful selection of data source and corresponding statistical analysis based on research objective is extremely important. Equally important aspect is to share the data and codes so that research claims or findings can be reproduced. This also forms the basis for the trustworthiness and usefulness of the research outcome.

1.3.1 Primary vs Secondary Sources

The term “sources” refers to data or materials used in research and has two distinct categories. The primary sources provide first hand information about an incident. Secondary sources are normally based on primary sources and provides interpretive information about an incident (Indiana University Libraries, 2007). For example, propaganda video/ speech released by ISIL or any other terrorist group is a primary source whereas newspaper article that publishes journalist’s interpretation of that speech becomes secondary source. Researcher (Schuurman, 2018) suggests that, in such scenarios, the difference is not always distinguishable because it depends on the

⁴<https://www.prio.org/Data/Armed-Conflict/UCDP-PRIO/>

⁵<https://www.sipri.org/databases>

⁶<https://data.worldbank.org/>

type of question being asked. Contrary to popular belief, newspaper or media articles are considered a secondary source of information about terrorism and terrorists. However news or media articles can be considered as primary source of information when the research focuses on how media reports on terrorism (Schuurman, 2018). In our case, the main source of data is through news and media articles about reported terrorist incidents and fits the category of primary source of data based on research objective.

1.3.2 Use of Statistical Analysis

In most areas of scientific analysis, statistics is often considered as an important and accepted way to ensure that claims made by researchers meet defined quality standards (Ranstorp, 2006). To be specific, descriptive statistics helps describing variables within data and often used to perform initial data analysis in most research. On the other hand, inferential statistics helps drawing conclusions/ decisions based on observed patterns (Patel, 2009).

A prominent researcher (Andrew Silke, 2004), in his book “Research on Terrorism: Trends, Achievements and Failures”, explains why inferential statistics is significantly important in terrorism research context. The author suggests that inferential statistics is useful to introduce element of control into research. In an experimental research, control is usually obtained by random assignment of research subjects to experimental and control groups however it’s difficult achieve in real world research. As a result, lack of control element raises doubt any relations between variables which the research claims to find. As a solution, inferential statistics can help to introduce recognized control element within research and so that less doubt and more confidence can be achieved over the veracity of research outcome.

1.4 Overview of Prior Research

Scientific research in the field of terrorism is heavily impacted by research continuance issue. According to (Gordon, 2007), there is indeed a growing amount of literature in terrorism field but majority of contributors are one-timers who visit and study this field, contribute few articles, and then move to another field. Researcher (Schuurman, 2018) points out another aspect that the terrorism research has been criticized for a long time for being unable to overcome methodological issues such as high dependency on secondary sources, corresponding literature review methods and relatively insufficient statistical analyses. This argument is further supported number of prominent researchers in this field. Compared to other similar fields such as Criminology, terrorism research suffers a lot due to complications in data availability, reliability and corresponding analysis to make the research useful to policy makers (Brennan, 2016).

1.4.1 Harsh Realities

One of the harsh realities in terrorism research is that the use of statistical analysis is fairly uncommon. In late 80s, (Jongman, 1988) in his book “Political Terrorism: A New Guide To Actors, Authors, Concepts, Data Bases, Theories, And Literature” identified serious concerns in terrorism research related to methodologies used by the researcher to prepare data and corresponding level of analysis. (A. Silke, 2001) reviewed the articles in terrorism research between 1995 and 2000 and suggests that key issues raised by (Jongman, 1988) remains unchanged in that period as well. Their research findings indicates that only 3% of research papers involved the use of inferential analysis in the major terrorism journals. Similar research was carried out by (Lum, Kennedy, & Sherley, 2006) on quality of research articles in terrorism research and their research finding suggests that, much has been written on terrorism between 1971 to 2003 and around 14,006 articles were published however the research that can help/support counter terrorism strategy was extremely low. This study also suggests that only 3% of the articles were based on some form of empirical analysis, 1% of articles were identified as case studies and rest of the articles (96%) were just thought pieces.

Very recently, researcher (Schuurman, 2018) also conducted an extensive research to review all the articles (3442) published from 2007 to 2016 in nine academic journals on terrorism and provides an insight on whether or not the trend (as mentioned) in terrorism research continues. Their research outcome suggests upward trend in the use of statistical analysis however major proportion is related to descriptive analysis only. They selected 2552 articles for analysis and their findings suggests that:

- only **1.3%** articles made use of inferential statistics
- 5.8% articles used mix of descriptive and inferential statistics
- 14.7% articles used descriptive statistics and
- 78.1% articles did not use any kind of statistical analysis

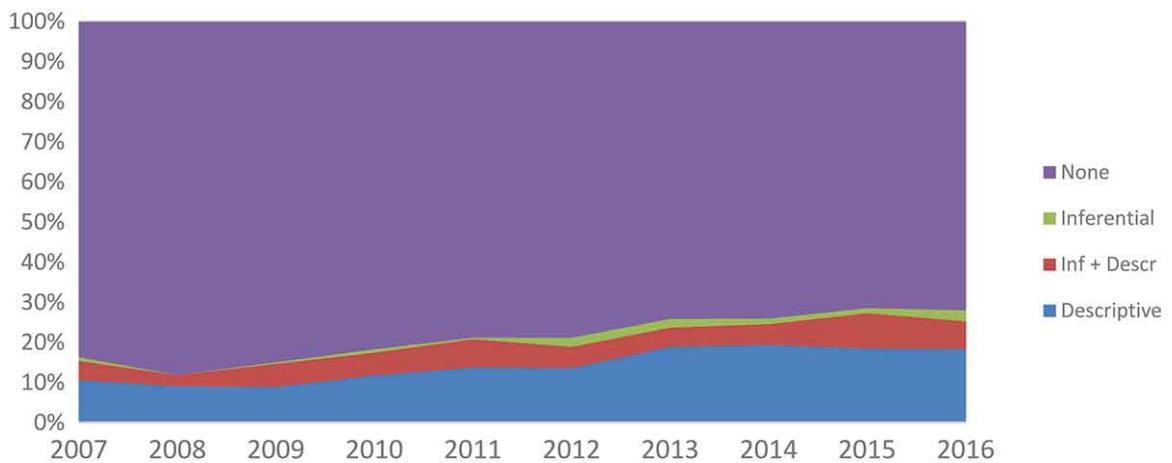


Figure 1.1: Use of statistics in terrorism research between 2007 to 2016

(Schuurman, 2018)

1.4.2 Review of Relevant Literature

In this section, we take a look at previous research that is intended toward counter terrorism support while making sure that the chosen research article/ literature contains at least some form of statistical modelling.

Simple linear regression was one of the approach for prediction models in early days but soon it was realized that such models are weak in capturing complex interactions. Emergence of machine learning algorithms and advancement in deep learning made it possible to develop fairly complex models however country-level analysis with resolution at year level contributes majority of research work in conflict prediction (Cederman & Weidmann, 2017).

(Beck, King, & Zeng, 2000) carried out an research to stress the important of the causes of conflict. Researchers claims that empirical findings in the literature global conflict are often unsatisfying, and accurate forecasts are unrealistic despite availability immense data collections, notable journals and complex analyses. Their approach uses a version of neural network model and argues that their forecasts are significantly better than previous effort.

In a study to investigate the factors that explain when terrorist groups are most or least likely to target civilians, researcher (Heger, 2010) examines why terrorist groups need community support and introduces new data on terrorist groups. The research then uses logit analysis to test the relationship between independent variables and civilian attacks between 1960-2000.

In a unique and interesting approach, a researcher from ETH Zürich (Chadefaux, 2014) examines a comprehensive dataset of historical newspaper articles and introduces weekly risk index. This new variable is then applied to a dataset of all wars reported since 1990. Outcome of this study suggests that the number of conflict-related news items increases dramatically prior to the onset of conflict. Researcher claims that the onset of a war within the next few months could be predicted with up to 85% confidence using only information available at the time. Another researcher (Cederman & Weidmann, 2017) supports the hypothesis and suggests that news reports are capable to capture political tension at a much higher temporal resolution and so that such variables have much stronger predictive power on war onset compared to traditional structural variables.

One of the notable (and publicly known) research in terrorism predicted the military coup in Thailand 1 month before its actual occurrence on 7 May 2014. In a report commissioned by the CIA-funded Political Instability Task Force, researchers (Ward Lab, 2014) forecasted irregular regime changes for coups, successful protest campaigns and armed rebellions, for 168 countries around the world for the 6-month period from April to September 2014. Researchers claims that Thailand was number 4 on their forecast list. They used an ensemble model (Ensemble Bayesian Model Averaging) that combines seven different split-population duration models.

Researchers (Fujita, Shinomoto, & Rocha, 2016) uses high temporal resolution data across multiple cities in Syria and time-series forecasting method to predict future

event of deaths in Syrian armed conflict. Their approach uses day level data of death tolls from Violations Documentation Center (VDC) in Syria. Using Auto-regression (AR) and Vector Auto-regression (VAR) models, their study identifies strong positive auto-correlations in Syrian cities and non-trivial cross-correlations across some of them. Researchers suggests that strong positive auto-correlations possibly reflects a sequence of attacks within short periods triggered by a single attack, as well as significant cross-correlation in some of the Syrian cities implies that deaths in one city were accompanied by deaths at another city.

Within a pattern recognition context, researchers (Klaussen, Marks, & Zaman, 2016) from MIT Sloan developed a behavioral model to predict which Twitter users are likely belonged to the Islamic state group. Using data of approximately 5,000 Twitter users who were linked with Islamic state group members, they created dataset of 1.3 million users by associating friends and followers of target users. At the same time, they monitored Twitter over few months to identify which profiles are getting suspended. Researchers claims that they were able to train a machine learning model that matched suspended accounts with the specifics of the profile and creating a framework to identify likely members of ISIS.

A similar research from (Ceron, Curini, & Iacus, 2018) examines over 25 million tweets in Arabic language when Islamic State was at its peak strength (between Jan 2014 to Jan 2015) and was expanding regions unders its control. Researchers assessed the share of support from online Arab community toward ISIS and investigated time time-granularity of tweets while linking the tweet opinions with daily events and geo location of tweets. Outcome of their research finds relationship between foreign fighters joining ISIS and online opinions across the regions.

One of the research evaluates the targeting patterns and preferences of 480 terrorist groups that were operational between 1980 and 2011 in order to find the impact of longetivity of terrorist groups based on their lethality. Based on group-specific case studies on the Afghan and Pakistani Taliban and Harmony Database from Combat Terrorism Center, researcher (Nawaz, 2017) uses Bivariate Probit Model to assess endogenous relationship and finds significant correlationship between negative group reputation and group mortality. Researcher also uses Cox Proportional Hazard Model to estimate longetivity of group.

(Colaresi & Mahmood, 2017) carried out a research to identify and avoid the problem of overfitting sample data. Researchers used the models of civil war onset data and came up with tool (R package: ModelCriticism) to illustrate how machine learning based research design can improve out of fold forecasting performance. Their study recommends making use of validation split along with train and test split to benefit from iterative model critisism.

Researchers (Muchlinski, Siroky, He, & Kocher, 2016/ed) uses The Civil War Data (1945-2000) and compared the performance of Random Forests model with three different versions of logistic regression. Outcome of their study suggests that Random Forests model provides significantly more accurate predictions on the occurrences of rare events in out of sample data compared to logistic regression models on chosen dataset. However in an experimental research to reproduce this claims, (Neunhoeffer

& Sternberg, 2018) ran re-analysis and finds problematic usage of cross-validation strategy. They contests the claim and suggests that there is no evidence of impressive predictive performance of random forest as claimed by the original authors. Research from (Neunhoeffer & Sternberg, 2018) also illustrates the importance of having access to replication code in order to measure the quality and/or claims of any research paper.

1.4.3 GTD and Machine Learning in Previous Research

Addressing the issue of rare events, researchers (Clauset & Woodard, 2013) comes up with statistical modelling approach to estimate future probability of large scale terrorist attack. Using the data from GTD and RAND-MIPT database between 1968-2007, and three different models i.e power law, exponential distributions and log normal, researchers estimates likelihood of observing 9/11 sized attack between 11-35%. Using the same procedure, researchers then makes a data-driven statistical forecast of at least one similar event over the next decade.

In a study to identify determinants of variation in country compliance with financial counter terrorism, researcher (Lula, 2014) used dataset on financial counter terrorism for the period 2004-2011 along with Global Terrorism Database. Researcher employs both quantitative and qualitative analysis in their approach and uses regression analysis (ordered logit model) to estimate statistical significance of independent variables on target variable i.e. compliance rates. Outcome of this study suggests that intensity and magnitude of terror threat, rate of international terror attacks, rate of suicide (terror) attacks, and military capability variable does not have statistically significant effect on country compliance with financial counter terrorism. Based on research findings, author suggests that many of the assumptions made in previous study in financial counter terrorism are incorrect.

A research from (Brennan, 2016) uses machine learning based approach to investigate terrorist incidents by country. This study makes use of regression techniques, Hidden Markov model, online time series detection algorithms such as twitter outbreak detection algorithm and Netflix's SURUS algorithm, as well as medical syndromic surveillance algorithms i.e EARSC based method and Farrington's method to detect change in behaviour (in terms of terrorist incident or fatalities). Outcome of their study suggests that time-series aberration detection methods were highly interpretable and generalizable compared to traditional methods (regression and HMM) for analysing time series data.

Researcher (Block, 2016) carried out a study to identify characteristics of terrorist events specific to aircrafts and airports and came up with situation crime prevention framework to minimize such attacks. In particular, researcher uses GTD data (2002-2014) specific to attacks involving airports/ aircraft that contains terrorist events related to 44 nations. In this study, Logistic Regression model is used to evaluate variables that are significantly associated with such attacks. Their research findings suggests that the likelihood of attacks against airports is mostly related to with domestic terrorists groups and, explosives and suicide attacks as a type of attack.

In contrast, attacks against aircraft is more associated with international terrorists groups.

In an effort to improve accuracy of classification algorithms, researchers (Mo, Meng, Li, & Zhao, 2017) uses GTD data and employs feature selection methods such as Minimal-redundancy maximal-relevancy (mRMR) and Maximal relevance (Max-Relevance). In this study, researchers uses Support Vector Machine (SVM), Naive Bayes (NB) and Logistic Regression (LR) algorithms and evaluates performance of each model through classification precision and computational time. Their research find suggests that feature selection methods improves the accuracy of the model and comparatively, Logistic Regression model with seven optimal feature subset achieves a classification precision of 78.41%.

A research from (Ding, Ge, Jiang, Fu, & Hao, 07AD–2017) also uses classification technique to evaluate risk of terrorist incident at global level using GTD and several other datasets. In particular, data comprising terror incidents between 1970 to 2015 was used to train and evaluate neural network (NNET), support vector machine (SVM), and random forest (RF) models. For performance evaluation, researchers used three-quarters of the randomly sampled data as training set, and the remaining as test set. Outcome of their study predicted the places where terror events might occur in 2015, with a success rate of 96.6%.

Similar research within classification context and addressing the issue of class unbalance in order to predict rare events i.e. responsible group behind terror attack, researchers (Gundabathula & Vaidhehi, 2018) employs various classification algorithms in line with sampling technique to improve the model accuracy. In particular, this study was narrowed down to terrorist incidents in India and data used from GTD was between 1970-2015. Researchers uses J48, IBK, Naive Bayes algorithms and ensemble approach using vote for classification task. Findings from their study points out importance of using sampling technique which improves the accuracy of base models and suggests that ensemble approach improves overall accuracy of base models.

1.5 Literature Gap and Relevance

Review of recent and relevant literature suggests that use of historical data from open source databases, and statistical modelling using time-series forecasting and classification algorithms is commonly used approach to address the research questions related to “when and where”. A trend can be seen in research study with variety of new approaches such as feature selection, sampling technique, validation split etc to achieve better accuracy in classification algorithms. This is one of the most relevant aspect for this research project.

While some approach argues that prediction is contentious issue and focuses on finding causal variables while neglecting model fit, there is an upward trend in an approach that uses diverse models, and out of fold method which also allows to evaluate and

compare model performance. Similarly, single model philosophy based on Occam's Razor principle is visible in majority of research in the past however ensemble philosophy to make use of weak but diverse models to improve the overall accuracy is gaining popularity amongst research nowadays.

It is also observed that use of gradient boosting algorithms is not popular in scientific research despite the availability and practical use cases of highly efficient and open-source algorithms such as XGBoost and LightGBM which are widely used in machine learning competitions such as Kaggle. In contrast, traditional algorithms such as Random Forests, Logistic Regression, Naive Bayes, J48 etc. are often used in majority of research.

One important observation from literature review is that code sharing is quite uncommon. Out of all the reviewed articles, only few provided codes or links to code repositories such as github. Replication crisis is a major issue in scientific research. Despite availability of number of open source tools for reproducible research such as Jupyter notebook, rmarkdown or a code repositories such as github, majority of research papers lacks code sharing aspect.

Chapter 2

Impact Analysis

This part of the research uses descriptive statistics to explore and understand terrorist events from various perspectives. This is essential to examine characteristics of attacks and responsible groups over the period of time. Findings and insights from this analysis is eventually helpful to select appropriate data for the statistical modelling part.

2.1 Data Preparation

The primary data file `globalterrorismdb_0617dist.xlsx` used in this research contains over 170,000 terrorist attacks between 1970-2016 (excluding the year 1993). This file can be downloaded by filling up a form on START Consortium's website.¹ This file contains total of 135 variables categorized by incident ID and date, incident information, attack information, weapon information, target/victim information, perpetrator information, casualties and consequences, and additional information. Out of 135 variables, I have selected total of 38 variables from each categories that are relevant to research objective. During data cleaning process, I have made following changes (corrective steps) to original data to make it ready for analysis:

- renaming of some variables (such as `gname` to `group_name`, `INT_LOG` to `int1_logistical_attack`) to keep the analysis and codes interpretable to wider audience.
- replacing 2.7% NAs in latitude and longitude with country level or closest matching geocodes. Note that most NAs refers to either disputed territories such as Kosovo or countries that no longer exists such as Czechoslovakia.
- 5% NAs in `nkill` (number of people killed) and 9% NAs in `nwound` (number of people wounded) variable replaced with 0. GTD reference manual suggests that “Where there is evidence of fatalities, but a figure is not reported or it is too vague to be of use, this field remains blank.”
- NAs in regional variables i.e `city` and `provstate` replaced with “unknown”

¹Accessing GTD data: <https://www.start.umd.edu/gtd/contact/>

GTD data is further enriched with country and year wise indicators from World Bank Open Data to get multi-dimensional view and for modelling part. This data is also open-source and can be accessed through R library `WDI`.² Following is the illustration of code used to query `WDI` api and to merge it with primary data.

```
# install WDI package in R and query the api with keyword
WDIsearch('conflict')

# create an index of selected variables based on search result
ind = c("arms_export" = "MS.MIL.XPRT.KD",
       "arms_import" = "MS.MIL.MPRT.KD",
       "population" = "SP.POP.TOTL",
       "gdp_per_capita" = "NY.GDP.PCAP.KD",
       "refugee_origin" = "SM.POP.REFG.OR",
       "refugee_asylum" = "SM.POP.REFG",
       "net_migration" = "SM.POP.NETM",
       "n_peace_keepers" = "VC.PKP.TOTL.UN",
       "conflict_index" = "IC.PI.CIR")

countries_vec <- as.vector(unique(df$ISO)) # countries in gtd dataset

#Extract selected data by specifying start and end year
wdi_data <- WDI(indicator = ind,
                  start = 1970,
                  end = 2016,
                  extra = TRUE) %>%
  select(year, ISO = iso3c, arms_export, arms_import, population,
         gdp_per_capita, refugee_origin, refugee_asylum, net_migration,
         n_peace_keepers, conflict_index) %>%
  drop_na(ISO) %>%
  filter(ISO %in% countries_vec) %>%
  replace_na(list(arms_export = 0, arms_import = 0, population = -1,
                 gdp_per_capita = 0, refugee_origin = 0,
                 refugee_asylum = 0, net_migration = 0,
                 n_peace_keepers = 0, conflict_index = -1))

# merge it with gtd data (by ISO code and year)
# (Not to run, already merged in data preparation step)
# df <- df %>% left_join(wdi_data)
```

Note that above mentioned external data is merged by country and year only, and missing values are replaced as shown in the code. Below are the variables (with short description where necessary) from clean and prepared dataset and that are used in data analysis part. Detailed information and explanation about each variable can be found GTD codebook.

²Searching and extracting data from the World Bank's World Development Indicators. : <https://cran.r-project.org/web/packages/WDI/WDI.pdf>

Table 2.1: Short description of important variables

Name of the Variable	description
eventid	a 12-digit Event ID
year	year in which the incident occurred
month	month
day	day
country	country
region	world region
provstate	an administrative division or unit of a country
city	city
latitude	latitude
longitude	longitude
attack_type	method of attack (reflects the broad class of tactics used)
weapon_type	type of weapon used in the incident
target_type	type of target/victim
target_nalty	nationality of the target that was attacked
group_name	name of the group that carried out the attack
nkill	number of total confirmed fatalities for the incident
nwound	number of confirmed non-fatal injuries
extended	whether or not an incident extended more than 24 hours
crit1_pol_eco_rel_soc	political, economic, religious, or social goal
crit2_publicize	intention to coerce, or publicize to larger audience
crit3_os_intl_hmn_law	action from the incident is outside intl humanitarian law
part_of_multiple_attacks	whether an incident being part of multiple attacks
attack_success	suicide attack
suicide_attack	whether an incident was successful
individual_attack	whether an attack carried out by unaffiliated Individual(s)
intl_logistical_attack	cross border incident
intl_ideological_attack	attack on target of a different nationality
ISO	ISO code for country
date	Approx. date of incident
arms_export	Arms exports (SIPRI trend indicator values)
arms_import	Arms imports (SIPRI trend indicator values)
population	Population, total
gdp_per_capita	GDP per capita (constant 2010 US\$)
refugee_origin	Refugee population by country or territory of origin
refugee_asylum	Refugee population by country or territory of asylum
net_migration	Net migration
n_peace_keepers	Presence of peace keepers
conflict_index	Extent of conflict of interest regulation index (0-10)

2.2 Global Overview

A quick look at region level number attacks suggests that situation is becoming worst in Middle East & North Africa followed by South Asia, Sub-suharan Africa and South-east Asia where exponential growth in number of attacks can be observed specifically from years 2010 to 2016.

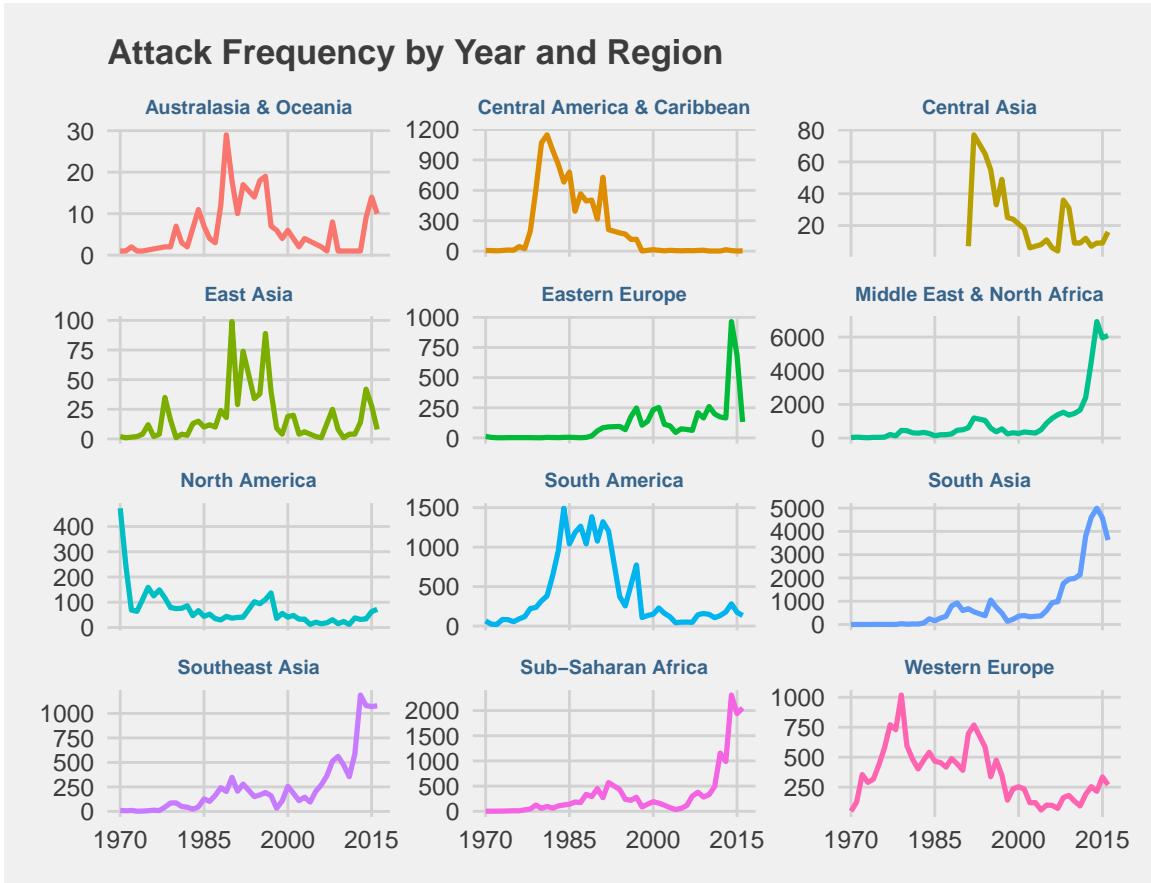


Figure 2.1: Attack Frequency by Year and Region

An interesting observation is in Eastern Europe region where sudden increase in number of attacks can be observed during 2014-2015 and then sudden decrease in 2016. Within the most impacted regions, nearly similar trend of gradual increase in number of attacks after 2010 and peak during 2014-2015 is visible. It's worth mentioning that in June 2014, Islamic State announced establishment of "Caliphate" while declaring Abu Bakr al-Baghdadi as "leader of Muslims everywhere" and urging other groups to pledge allegiance (Al Jazeera, 2014). Islamic State was at its peak strength during Jan 2014 to Jan 2015 (Ceron et al., 2018).

To understand the attack characteristics, let's take a look at Frequency of attack type and type of weapon used by terrorist groups.

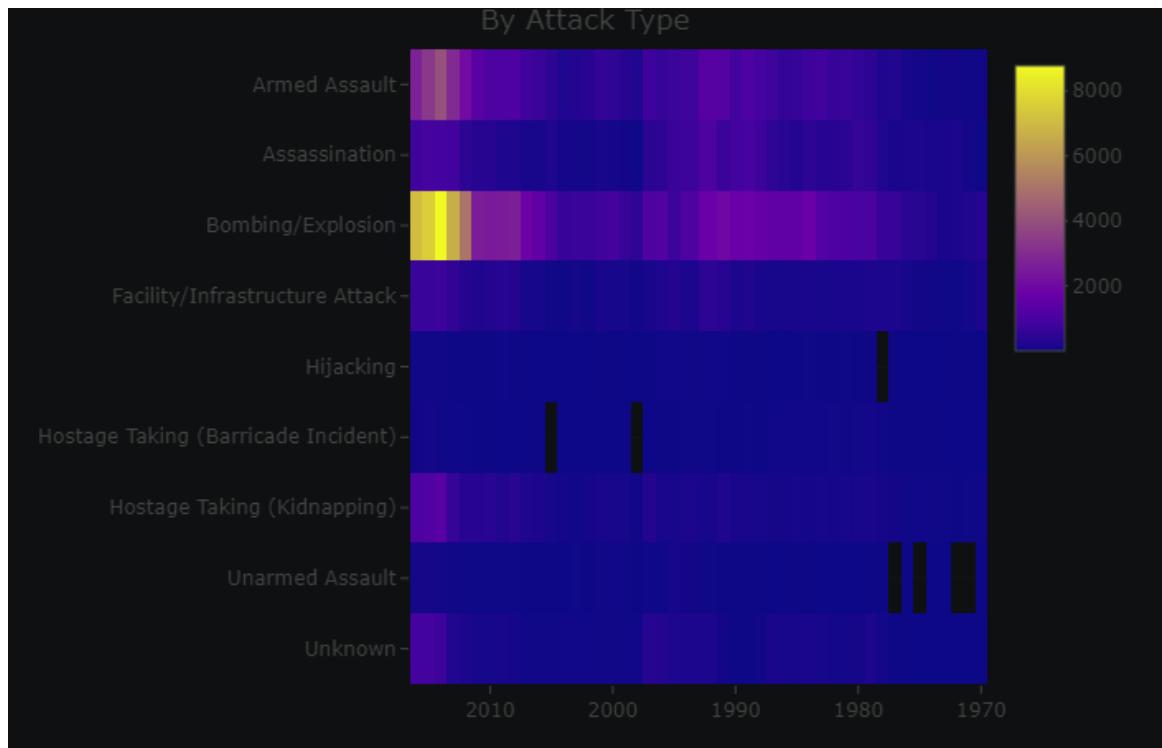


Figure 2.2: Trend in type of attack in all incidents globally

The heat signatures indicate Bombing/Explosive as one of the frequently used techniques by terrorist groups. Although the pattern in this tactic is visible throughout all the year, while rising during late 80s and early 90s however it has now increased to nearly 7 times since 2006. Similar pattern (with lower magnitude) can be observed in Armed Assault followed by Hostage Taking and Assassination technique.

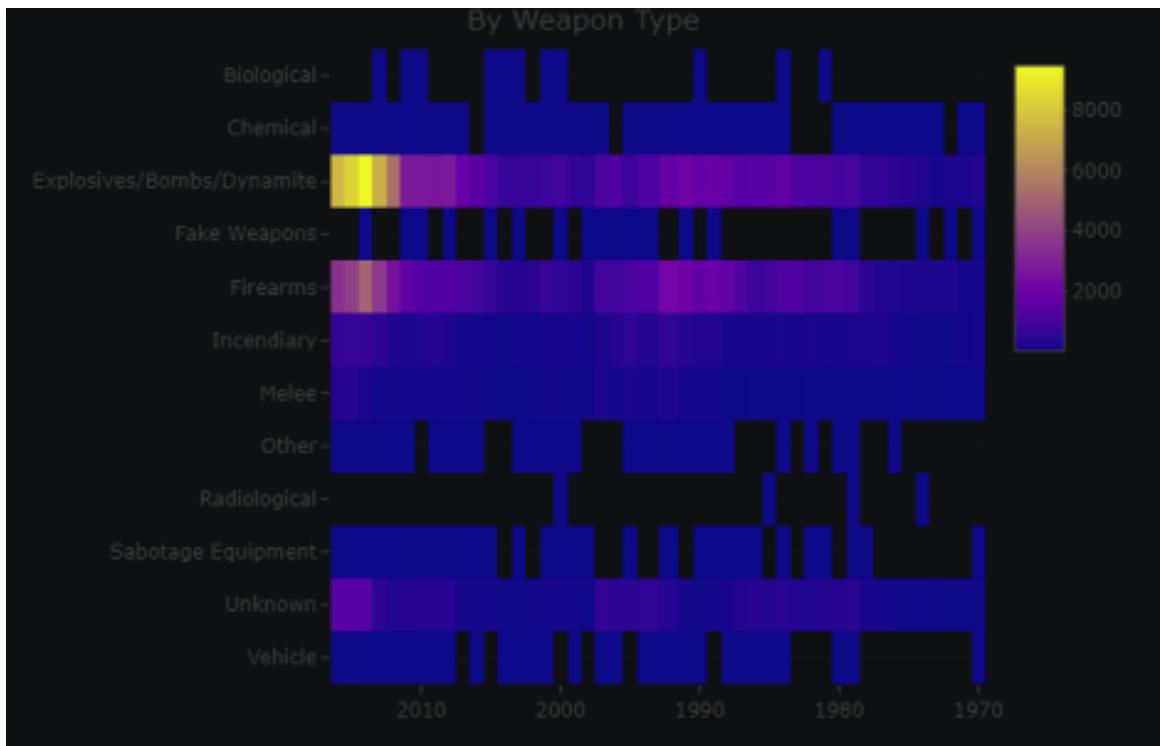


Figure 2.3: Trend in type of weapon used in all incidents globally

Upon examining the trends in type of weapon used in all terrorist incidents globally as shown in the figure 2.3, it is visible that use of Explosives/Bomb/Dynamites and Firearms is extremely high since 2011 and compared to other weapon types. Use of vehicles as weapon type was relatively low until 2013 however it was on peak in 2015 with total 34 number number of attacks.

Observing trends in target type over the period of time is also a useful way to understand characteristics and ideology amongst terrorist incidents. As shown in the plot below, the heat signature indicates the top five most frequently attacked target types as Private Citizens & Property followed by Military, Police, Government and Business.

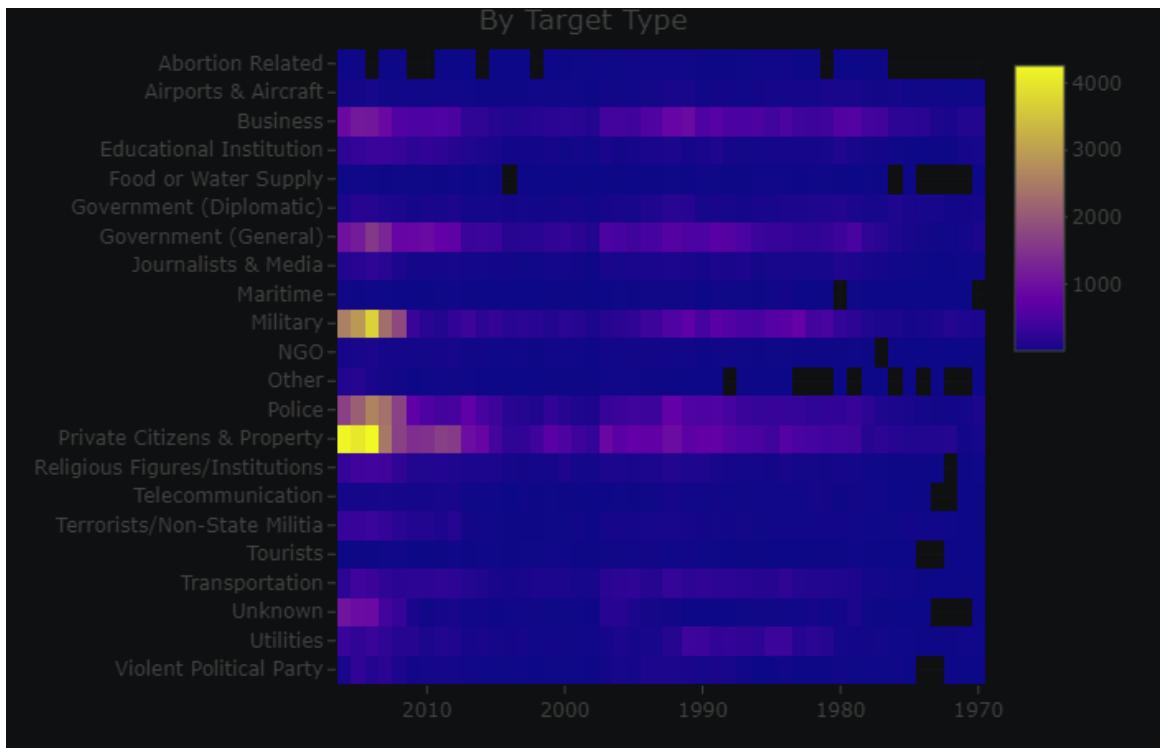


Figure 2.4: Trend in intended targets in all incidents globally

According to GTD codebook, Private Citizens & Property category includes attack on individuals, public in general or attacks in highly populated areas such as markets, commercial streets, busy intersections and pedestrian malls. In a study to investigate when terrorist groups are most or least likely to attack civilians, researcher (Heger, 2010) find a relationship with group's political motivation and suggests that terror groups pursuing a nationalist agenda are more likely to attack civilians. A relatively lower magnitude trend but with gradual increase in recent years is also visible on Religious Figures/Institution and Terrorist / Non-state Militia category. The inclusion criteria for Terrorist/ Non-state Militia category refers to terrorists or members of terrorist groups (that are identified in GTD) and broadly defined as informants for terrorist groups excluding former or surrendered terrorists.

2.3 The Top 10 Most Active and Violent Groups

Findings from exploratory data analysis at region level indicates that number of attacks have increased significantly from year 2010 and nearly at the same pace in Middle East & North Africa, South Asia, Sub-suharan Africa and Southeast Asia region. Trends in attack type, weapon type and target type over the same period of time (from 2010) suggests that bombings and explosions as a choice of attack type is growing exponentially while use of explosives & firearms and attacks on civilians is at alarming high level.

This part of the research identifies and examines the top ten most violent and active terrorist groups based on number of fatalities and number of people injured. GTD codebook suggests that when an attack is a part of multiple attacks, sources sometimes provide a cumulative fatality total for all of the incidents rather than fatality figures for each incident.

In order to determine top ten most active and violent groups based on fatalities and injured while preserving statistical accuracy, first I filter the dataset for the events that took place from 2010 onwards and remove the incidents where group name is not known. The new variable `impact` is sum of fatalities and number of people injured. Wherever an attack is observed as a part of multiple attacks, and reported figures are different, I use the figure which is maximum amongst all the reported figures while ensuring that reported incidents are distinct and grouped by month, year, region and name of the group as shown in the code below:

```
by_groups <- df %>%
  filter(group_name != "Unknown" & year >= 2010) %>%
  replace_na(list(nkill = 0, nwound = 0)) %>%
  select(group_name, region, year, month, nkill, nwound,
         part_of_multiple_attacks) %>%
  group_by(group_name, region, year, month) %>%
  filter(if_else(part_of_multiple_attacks == 1,
                 nkill == max(nkill) & nwound == max(nwound),
                 nkill == nkill & nwound == nwound)) %>%
  distinct(group_name, region, year, month, nkill, nwound,
           part_of_multiple_attacks) %>%
  mutate(impact = nkill + nwound) %>%
  group_by(group_name) %>%
  summarise(total = sum(impact)) %>%
  arrange(desc(total)) %>%
  head(10)

# create a vector of top 10 groups for later analysis
top10_groups <- as.vector(by_groups$group_name)

ggplot(by_groups, aes(x= reorder(group_name, -total), y= total)) +
  geom_bar(stat = "identity", fill = "tomato3") +
  scale_x_discrete(labels = function(x) str_wrap(x, width = 10)) +
  ggtitle("Top 10 Most Active and Violent Groups") +
  xlab("Name of the group") +
  ylab("Total fatalities + injured") +
  theme(axis.title = element_text(size=9),
        axis.text = element_text(size = 8),
        plot.title = element_text(size=12), legend.position = "none")
```

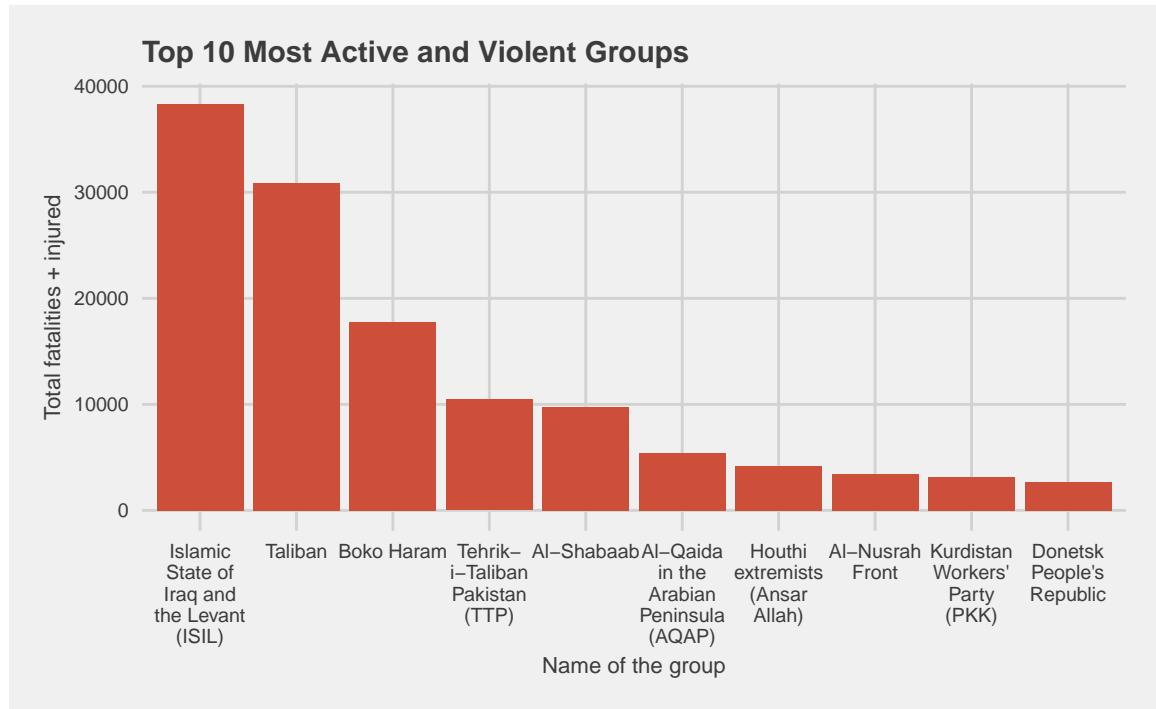


Figure 2.5: Top 10 Most Active and Violent Groups

Based on cumulative number of fatalities and injured people, we can see that ISIL and Taliban, followed by Boko Haram are the most violent groups that are currently active. To better understand their activity over the period of time, we take a look at attack frequency.

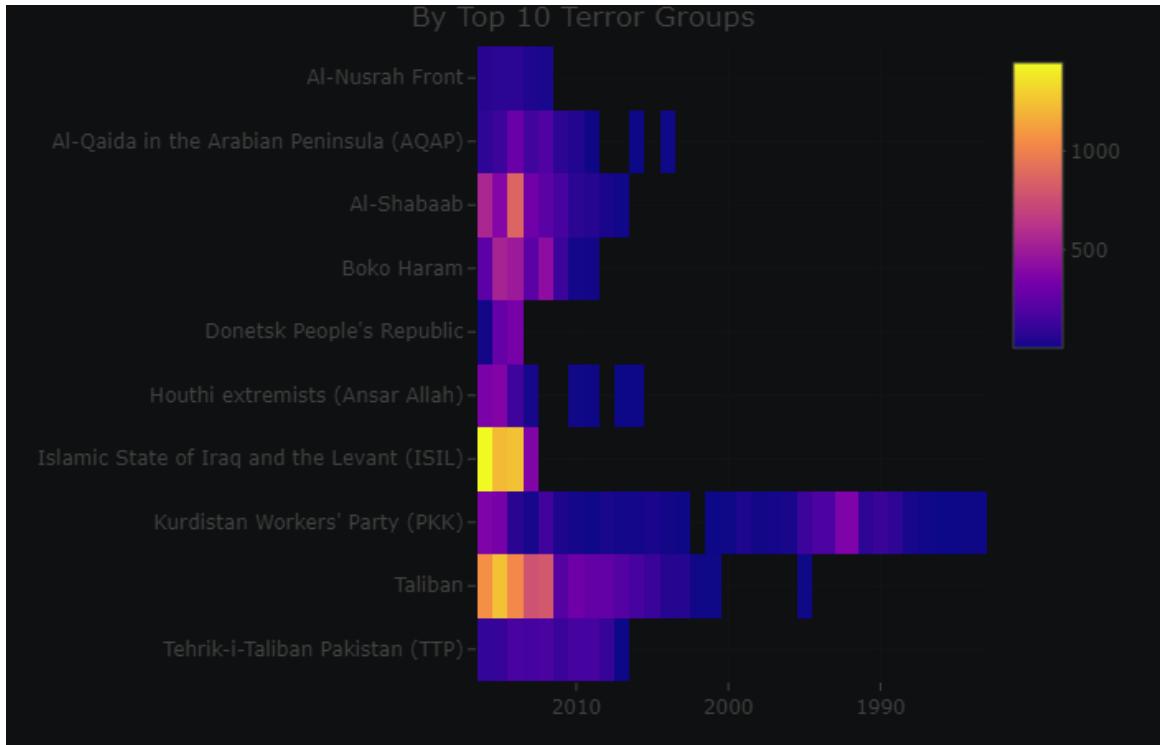


Figure 2.6: Attack Frequency from Top 10 Groups

It's interesting to see that majority of this most violent terrorist groups (6 out of 10) were formed after 2006 only. Particularly, number of attacks from ISIL can be seen increasing rapidly within shortest period of time (4 years) and a gradual increase in attacks from Taliban (reaching peak at 1249 in year 2015). Attack characteristics for all 10 groups (cummulative) indicates Military as the most frequent target (27.5%) followed by civilians (27.3%). Similarly, Bombing/Explosions and Armed assault as a most frequent attack tactics accounts for 70.4% of all the attacks as shown in the plots below.

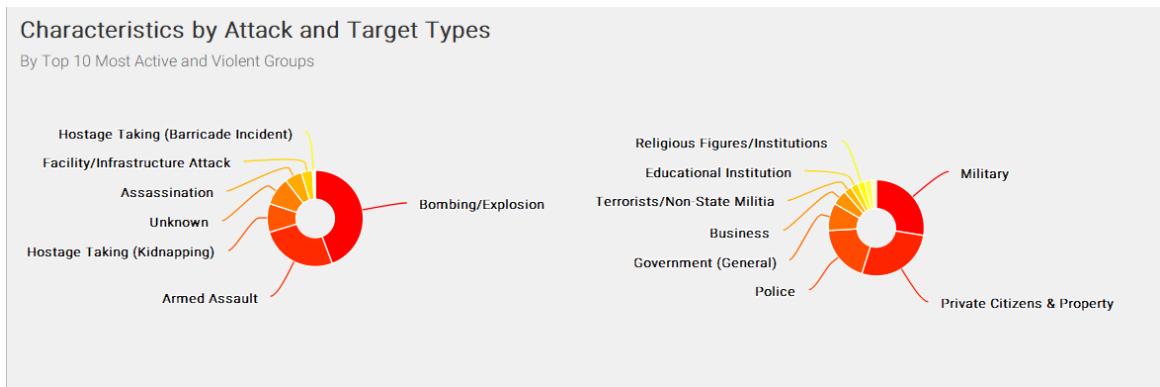


Figure 2.7: Characteristics of top 10 groups

To summarize, we identified the top 10 most lethal groups that are active between

2010 to 2016 and examined their characteristics behind attacks. We looked at the trend in type of attack and corresponding number of attacks over the period of time, which upto certain extent, indicates easy access to firearms and explosive devices either through illegal arms trade or through undisclosed support from powerful nation/s. We also examined pattern in target type in which, 46.7% attacks accounts for Military and Police category and 27.3% attack counts toward civilians.

2.4 The Major and Minor Epicenters

The term “Epicenter” used here refers to the location that is impacted by terrorist incidents from top 10 groups as defined. To examining the geographical spread and intensity of incidents from top 10 terroist groups at country level, I use cummulative sum of number of people killed and number of people wounded as measurement. Below is the code used to prepare the data for this analysis.

```
tmp <- df %>%
  filter(group_name %in% top10_groups) %>%
  replace_na(list(nkill = 0, nwound = 0)) %>%
  group_by(group_name, region, year, month) %>%
  filter(if_else(part_of_multiple_attacks == 1,
                nkill == max(nkill) & nwound == max(nwound),
                nkill == nkill & nwound == nwound)) %>%
  ungroup() %>%
  distinct(group_name, region, country, year, month, nkill,
          nwound, part_of_multiple_attacks) %>%
  group_by(country, region) %>%
  summarise(attack_count = n(), nkill_plus_nwound = sum(nkill + nwound))

tbl <- tmp %>% filter(region %in% c("North America", "Eastern Europe"))
knitr::kable(tbl, booktabs = TRUE,
             caption = "Terrorism Epicenters in North America and Eastern Europe") %>%
  kable_styling(latex_options = "hold_position")
```

Table 2.2: Terrorism Epicenters in North America and Eastern Europe

country	region	attack_count	nkill_plus_nwound
Russia	Eastern Europe	2	6
Ukraine	Eastern Europe	170	2695
United States	North America	2	2

As shown in table 2.2, North America and Eastern Europe region seems to have concentration on mainly one country across whole region (except Russia with comparatively minor impact). Next, we take a look at remaining regions.

```
tmp <- tmp %>% filter(!region %in% c("North America", "Eastern Europe"))
tmp %>%
  ggplot(aes(area = attack_count, fill = nkill_plus_nwound, label = country)) +
  geom_treemap() +
  geom_treemap_text(fontface = "italic", colour = "white",
                    place = "centre", grow = TRUE) +
  labs(title = "The Major and Minor Epicenters of Terrorism",
       subtitle = "Based on Top 10 Most Active and Violent Groups") +
  scale_fill_viridis(discrete = FALSE, begin = 0.1, end = 0.8) +
  facet_wrap(~ region, ncol = 2) +
  guides(fill = guide_colorbar(barwidth = 12))+
  theme(plot.title = element_text(size=12))
```



Figure 2.8: The Major and Minor Epicenters of Terrorism

While Afghanistan facing the largest impact from terrorist incidents in terms of fatalities and number of people injured followed by Iraq, it is also observed that spread across multiple countries in most regions. In Sub-Saharan Africa, the major epicenters of violent terrorist incidents can be seen mostly in Somalia and Nigeria. Although number of fatalities and injured is relatively less in Central Asia and Southeast Asia however major Epicenters can be identified as Philippines, Turkmenistan and Georgia.

In case Western Europe, it is surprising to see Germany and France as major epicenters and relatively high number of minor epicenters. Although number of fatalities and

injured is relatively low but spread across many countries implies greater threat.

To better understand exactly where this incidents belongs to, let us narrow down the analysis to city level for each groups.

```

tmp <- df %>%
  filter(group_name %in% top10_groups) %>%
  replace_na(list(nkill = 0, nwound = 0)) %>%
  group_by(group_name, region, year, month) %>%
  filter(if_else(part_of_multiple_attacks == 1,
                 nkill == max(nkill) & nwound == max(nwound),
                 nkill == nkill & nwound == nwound)) %>%
  ungroup() %>%
  distinct(group_name, region, country, city, year, month,
          nkill, nwound, part_of_multiple_attacks) %>%
  group_by(city, group_name) %>%
  summarise(attack_count = n(),
            nkill_plus_nwound = sum(nkill + nwound)) %>%
  filter(nkill_plus_nwound >= 100 & city != "Unknown")

tmp %>%
  ggplot(aes(area = attack_count,
             fill = nkill_plus_nwound,
             label = city)) +
  geom_treemap() +
  geom_treemap_text(fontface = "italic", colour = "white",
                    place = "centre", grow = TRUE) +
  labs(title = "Most Frequently Attacked Cities per Group",
       subtitle = "Based on Top 10 Most Active and Violent Groups") +
  scale_fill_viridis(discrete = FALSE, begin = 0.1, end = 0.8) +
  facet_wrap(~ group_name, ncol = 3) +
  guides(fill = guide_colorbar(barwidth = 12)) +
  theme(plot.title = element_text(size=12),
        strip.text = element_text(size = 8, face = "bold"))

```

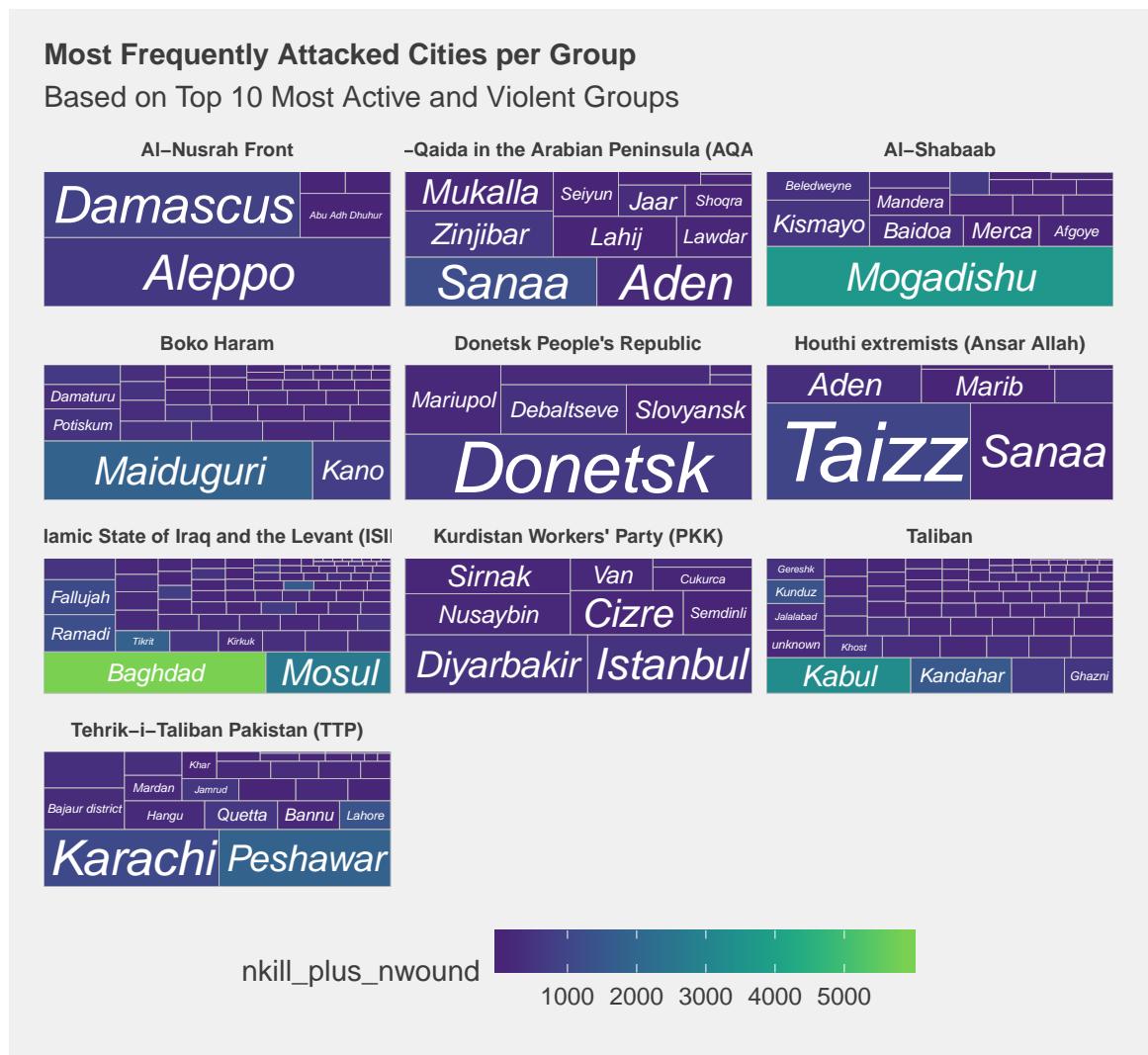


Figure 2.9: Terrorist Group and Impacted Cities

Chapter 3

Statistical Hypothesis Testing

This chapter is an extension to chapter 2 (Impact Analysis) and performs correlation test on numeric variables and hypothesis testing between groups and fatalities on data specific to top 10 most active and violent terrorist groups that were identified earlier.

3.1 Data Preparation

```
dfh <- df %>%
  filter(group_name %in% top10_groups) %>% # filter data by top 10 groups
  replace_na(list(nkill = 0, nwound = 0)) # replace NAs

# Shorten lengthy group names
dfh$group_name[dfh$group_name == "Kurdistan Workers' Party (PKK)"] <- "PKK"
dfh$group_name[dfh$group_name == "Al-Qaida in the Arabian Peninsula (AQAP)"] <-
dfh$group_name[dfh$group_name == "Houthi extremists (Ansar Allah)"] <- "Houthi_Ex
dfh$group_name[dfh$group_name == "Tehrik-i-Taliban Pakistan (TTP)"] <- "TTP"
dfh$group_name[dfh$group_name == "Al-Nusrah Front"] <- "Al-Nusrah"
dfh$group_name[dfh$group_name == "Islamic State of Iraq and the Levant (ISIL)"]
dfh$group_name[dfh$group_name == "Donetsk People's Republic"] <- "Donetsk_PR"
```

3.2 Correlation Test

Let's begin with correlation test to understand relationship between variables. I use pairwise complete observations method to compute correlation coefficients for each pair of numerical variables. Missing values are replaced appropriately to preserve statistical accuracy between variables.

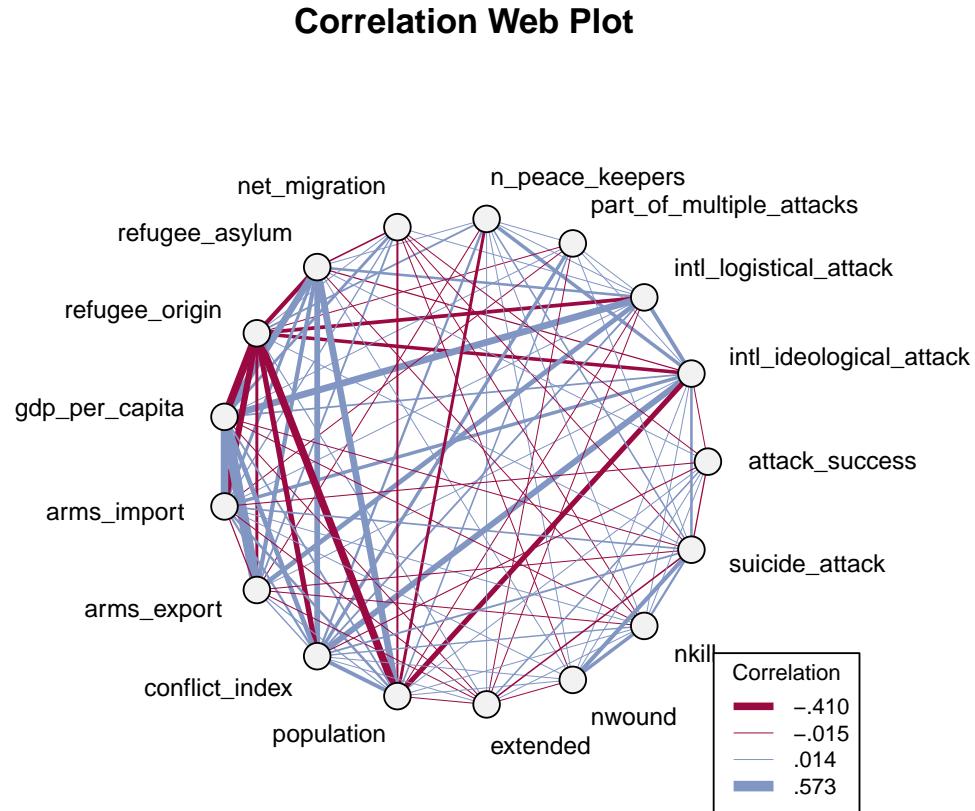
```
tmp <- dfh %>%
  select(intl_ideological_attack, intl_logistical_attack,
         part_of_multiple_attacks, n_peace_keepers, net_migration,
         refugee_asylum, refugee_origin, gdp_per_capita, arms_import,
         arms_export, conflict_index, population, extended,
         nwound, nkill, suicide_attack, attack_success)

# get the correlation matrix
m <- cor(tmp, use="pairwise.complete.obs")

# Get rid of all non significant correlations
ctest <- PairApply(tmp, function(x, y) cor.test(x, y)$p.value,
                    symmetric=TRUE)

m[ctest > 0.05] <- NA    # Replace p value > 0.05 with NAs

PlotWeb(m, lwd = abs(m[lower.tri(m)] * 10),
        cex.lab = 0.85, pt.bg = "#f2f2f2",
        args.legend = list(x = "bottomright",
                           cex = 0.75, bty = "0",
                           title = "Correlation"),
        main="Correlation Web Plot")
```



Pranav Pandya/2018-07-12

Figure 3.1: Correlation web plot

3.3 Hypothesis Test: Fatalities vs Groups

The objective behind this hypothesis is to determine whether or not means of the top 10 groups with respect to average fatalities are same. If at least one sample mean is different to others then we determine which pair of groups are different.

H_0 : The means of the different groups are the same
 $(ISIL) = (Taliban) = (AQAP) = (Al - Shabaab) =$
 $(Al - Shabaab) = (TTP) = (BokoHaram) =$
 $(Al - Nusrah) = (DonetskPR) = (HouthiExtrem)$

H_a : At least one sample mean is not equal to the others

```
ggplot(dfh, aes(group_name, nkill, fill = group_name)) +
  geom_boxplot(outlier.stroke = 0.1) +
  geom_jitter(alpha = 0.03, aes(color = group_name)) +
  theme_minimal() + coord_flip() + scale_y_log10() +
  ggtitle("Boxplot of Groups vs Fatalities") +
  xlab("Name of the group") + ylab("Total fatalities (log10)") +
  theme(plot.title = element_text(size=12), legend.position = "none")
```

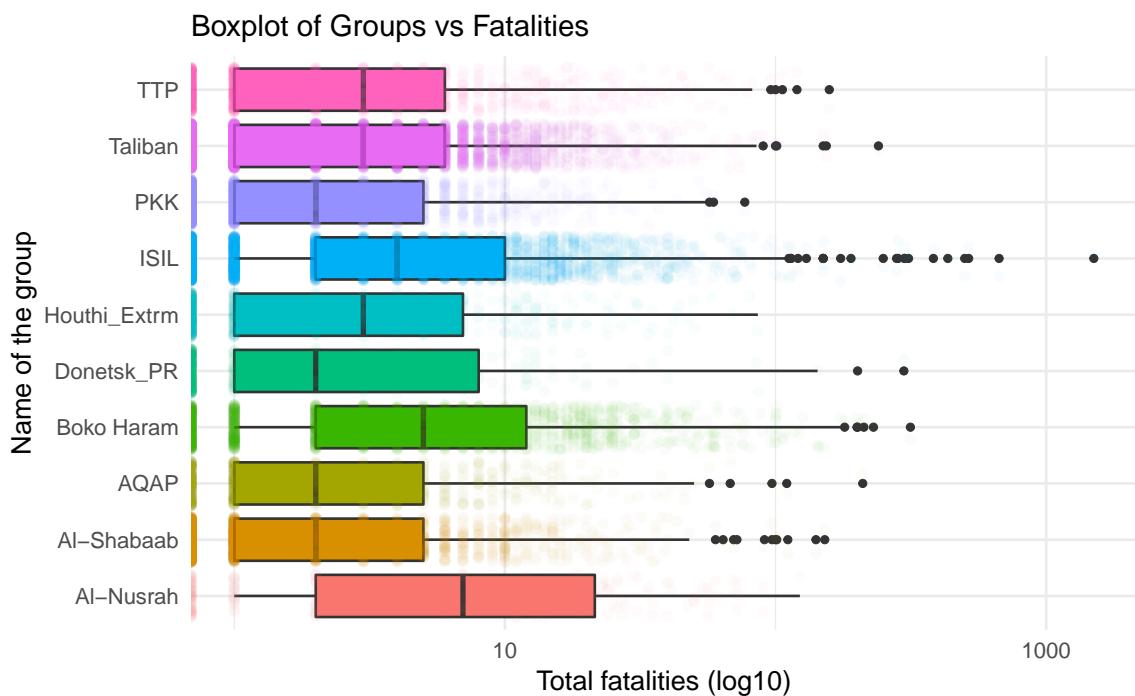


Figure 3.2: Boxplot: Group vs Fatalities

In statistical terms, we have some extreme outliers i.e. $nkill \sim 1500$ in ISIL group so X axis is log transformed for visualization purpose.

3.3.1 ANOVA Test

The ANOVA model computes the residual variance and the variance between sample means in order to calculate the F-statistic. This is the first step to determine whether or not means are different in pair of groups.

$$F - \text{statistic} = (S_{\text{between}}^2 / S_{\text{within}}^2)$$

```
# Compute the analysis of variance
r.aov <- aov(nkill ~ group_name , data = dfh)
# Summary of the analysis
summary(r.aov)

Df  Sum Sq Mean Sq F value      Pr(>F)
group_name     9 111070   12341    40.7 <0.0000000000000002 ***
Residuals    21770 6597154       303
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model summary provides us F value and Pr(>F) corresponding to the p-value of the test. As we can see that p-value is < 0.05 , which means there are significant differences between the groups. In other words, we reject the null hypothesis. From this test, we identified that some of the group means are different. The next step is to identify which pair of groups are different.

3.3.2 PostHoc Test

PostHoc test is useful to determine where the differences occurred between groups. For this test, I use several different methods for the comparison purpose. This methods can be classified as either conservative or liberal approach. Conservative methods are considered to be robust against committing Type I error as they use more stringent criterion for statistical significance. First we compare results from The Fisher LSD (Least Significant Different), Scheffe and Dunn's (Bonferroni) test.

```
# compare p-values:
posthoc1 <- as.data.frame(cbind(
  lsd= PostHocTest(r.aov, method="lsd")$group_name[, "pval"],          # The Fisher
  scheffe= PostHocTest(r.aov, method="scheffe")$group_name[, "pval"], # Scheffe
  bonf=PostHocTest(r.aov, method="bonf")$group_name[, "pval"]))        # Bonferroni

posthoc1 <- rownames_to_column(posthoc1, var = "Pair of groups") %>%
  arrange(desc(scheffe))

if( knitr:::is_latex_output() ) {
  knitr:::kable(posthoc1, caption = "Posthoc test (lsd, scheffe, bonf)",
  row.names = FALSE, booktabs = T) %>%
```

```

    kable_styling(font_size = 11, full_width = F, latex_options = "hold_position")
} else {
  # split lengthy table into two tables
  tbl1 <- head(posthoc1, 23)
  tbl2 <- tail(posthoc1, 22)
  knitr::kable(list(tbl1, tbl2), caption = "Posthoc test (lsd, scheffe, bonf)",
              row.names = FALSE, booktabs = T) %>%
    kable_styling(font_size = 11, full_width = F)
}

```

The Fisher LSD (Least Significant Different) test is the most liberal in all the post hoc test. The Scheffe test is the most conservative in all the post hoc test and protects against Type I error. On the other hand, Dunn's (Bonferroni) test is extremely conservative (Andri Signorell et mult. al., 2018). Out of all the possible combination of pairs (45), 16 pair of groups indicates p adj value > 0.9 based on Scheffe test. In statistical terms, it means 16 pairs of groups as shown in the table above have non-significantly different means in number of fatalities.

Next, I use Tukey HSD (Honestly Significant Difference) method which is the most common and preferred method.

```

# extract only p-values by setting conf.level to NA
hsd <- PostHocTest(r.aov, method = "hsd", conf.level=NA)
# convert to data frame and round off to 3 digits
hsd <- as.data.frame(do.call(rbind, hsd)) %>% round(4) # round off to 4 digits

if( knitr:::is_latex_output() ) {
  knitr::kable(hsd, booktabs = T, "latex",
               caption = "Post hoc test with Tukey HSD for Pair of Groups") %>%
    kable_styling(font_size = 7, full_width = F,
                  position = "left", latex_options = "hold_position")
} else {
  knitr::kable(hsd, booktabs = T,
               caption = "Post hoc test with Tukey HSD for Pair of Groups") %>%
    kable_styling(font_size = 10, full_width = F, position = "left")
}

```

3.3.3 Interpretation

The pairs of groups with adj p-value near or equals to 1 represents non-significantly different means in number of fatalities such as Boko Haram - Al-Nusrah, Al-Qaida in Arabian Peninsula (AQAP)- Al-Shabaab, Houthi Extremist- PKK, Taliban- Tehrik-i-Taliban etc.

Similarly, pair of groups with adjusted p-value near zero indicates significant different means in number of fatalities such as pairs of ISIL with all the remaining groups,

Taliban - Al-Nusrah, PKK - Boko Haram, Donetsk_PR - Al-Nusrah etc.

Table 3.1: Posthoc test (lsd, scheffe, bonf)

Pair of groups	lsd	scheffe	bonf
Donetsk_PR-Al-Shabaab	0.9191	1.0000	1.0000
Houthi_Extrm-Al-Shabaab	0.7934	1.0000	1.0000
Houthi_Extrm-Donetsk_PR	0.7797	1.0000	1.0000
Taliban-AQAP	0.6811	1.0000	1.0000
PKK-Donetsk_PR	0.5800	1.0000	1.0000
Houthi_Extrm-AQAP	0.4850	1.0000	1.0000
Donetsk_PR-AQAP	0.3615	0.9997	1.0000
PKK-Houthi_Extrm	0.3152	0.9994	1.0000
PKK-Al-Shabaab	0.3021	0.9993	1.0000
AQAP-Al-Shabaab	0.2561	0.9984	1.0000
Taliban-Houthi_Extrm	0.1928	0.9954	1.0000
TTP-AQAP	0.1508	0.9904	1.0000
Taliban-Donetsk_PR	0.1476	0.9898	1.0000
TTP-Taliban	0.1253	0.9846	1.0000
Boko Haram-Al-Nusrah	0.0851	0.9656	1.0000
PKK-AQAP	0.0610	0.9406	1.0000
TTP-Houthi_Extrm	0.0324	0.8694	1.0000
TTP-Donetsk_PR	0.0278	0.8481	1.0000
Taliban-Al-Shabaab	0.0135	0.7301	0.6094
TTP-Al-Shabaab	0.0024	0.4187	0.1088
ISIL-Al-Nusrah	0.0008	0.2574	0.0354
Taliban-PKK	0.0005	0.2071	0.0226
ISIL-Boko Haram	0.0002	0.1338	0.0097
TTP-PKK	0.0002	0.1172	0.0076
TTP-ISIL	0.0000	0.0072	0.0001
TTP-Al-Nusrah	0.0000	0.0006	0.0000
ISIL-AQAP	0.0000	0.0000	0.0000
ISIL-Donetsk_PR	0.0000	0.0000	0.0000
AQAP-Al-Nusrah	0.0000	0.0000	0.0000
Donetsk_PR-Al-Nusrah	0.0000	0.0000	0.0000
Houthi_Extrm-Al-Nusrah	0.0000	0.0000	0.0000
Taliban-Al-Nusrah	0.0000	0.0000	0.0000
ISIL-Houthi_Extrm	0.0000	0.0000	0.0000
TTP-Boko Haram	0.0000	0.0000	0.0000
Al-Shabaab-Al-Nusrah	0.0000	0.0000	0.0000
PKK-Al-Nusrah	0.0000	0.0000	0.0000
Donetsk_PR-Boko Haram	0.0000	0.0000	0.0000
Boko Haram-AQAP	0.0000	0.0000	0.0000
Houthi_Extrm-Boko Haram	0.0000	0.0000	0.0000
Taliban-ISIL	0.0000	0.0000	0.0000
ISIL-Al-Shabaab	0.0000	0.0000	0.0000
PKK-ISIL	0.0000	0.0000	0.0000
Taliban-Boko Haram	0.0000	0.0000	0.0000
Boko Haram-Al-Shabaab	0.0000	0.0000	0.0000
PKK-Boko Haram	0.0000	0.0000	0.0000

Table 3.2: Post hoc test with Tukey HSD for Pair of Groups

	Al-Nusrah	Al-Shabaab	AQAP	Boko Haram	Donetsk_PR	Houthi_Extrm	ISIL	PKK	Taliban
Al-Shabaab	0.0000	NA	NA	NA	NA	NA	NA	NA	NA
AQAP	0.0000	0.9811	NA	NA	NA	NA	NA	NA	NA
Boko Haram	0.7832	0.0000	0.0000	NA	NA	NA	NA	NA	NA
Donetsk_PR	0.0000	1.0000	0.9961	0.0000	NA	NA	NA	NA	NA
Houthi_Extrm	0.0000	1.0000	0.9995	0.0000	1.0000	NA	NA	NA	NA
ISIL	0.0272	0.0000	0.0000	0.0082	0.0000	0.0000	NA	NA	NA
PKK	0.0000	0.9904	0.6872	0.0000	0.9999	0.9921	0.0000	NA	NA
Taliban	0.0000	0.2852	1.0000	0.0000	0.9119	0.9534	0.0000	0.0180	NA
TTP	0.0000	0.0731	0.9158	0.0000	0.4568	0.4992	0.0001	0.0065	0.8792

Chapter 4

Pattern Discovery

This part of analysis is based on unsupervised machine learning algorithm and makes use of association rules to discover patterns in terrorist incidents from Islamic State, Taliban and Boko Haram group that were identified in top 5 most active and violent groups.

Mining of association rules is widely used method in retail and ecommerce environment and commonly known as Market Basket Analysis using Apriori algorithm. The theory behind this approach is that if a customer buys a certain group of products then they are more or less likely to buy another group of products (Karthiyayini & Balasubramanian, 2016).

- **Pseudocode of the Apriori algorithm:**

Apriori(T, ϵ)

$L_1 \leftarrow \{\text{large 1 - itemsets}\}$

$k \leftarrow 2$

while $L_{k-1} \neq \emptyset$

$C_k \leftarrow \{a \cup \{b\} \mid a \in L_{k-1} \wedge b \notin a\} - \{c \mid \{s \mid s \subseteq c \wedge |s| = k-1\} \not\subseteq L_{k-1}\}$

for transactions $t \in T$

$D_t \leftarrow \{c \mid c \in C_k \wedge c \subseteq t\}$

for candidates $c \in D_t$

$\text{count}[c] \leftarrow \text{count}[c] + 1$

$L_k \leftarrow \{c \mid c \in C_k \wedge \text{count}[c] \geq \epsilon\}$

$k \leftarrow k + 1$

return $\bigcup_k L_k$

As the goal of this algorithm is to determine set of frequent items among the candidates, this methodology can also be applied to discover patterns within terrorism context. The idea is to understand attack habits from terrorist groups by finding association and correlation between different attacks that were carried out in the

past. It's important to note that output from this algorithm is a list of association rules (frequent patterns) and provides descriptive analysis only. The real value of such unsupervised learning is in the insights we can take away from algorithm's finding.

4.1 Data preparation

For this analysis, I have chosen specific variables that are not highly correlated with chosen groups i.e. target type, weapon type, attack type, suicide attack and number of fatalities while excluding the observations where value is "Unknown".

```
tmp <- dfh %>%
  select(group_name, target_type, weapon_type, attack_type, suicide_attack, nkill)
  filter(target_type != "Unknown" & weapon_type != "Unknown" & attack_type != "Unknown")
  mutate(nkill = if_else(nkill == 0, "0",
                        if_else(nkill >= 1 & nkill <= 5, "1 to 5",
                               if_else(nkill > 5 & nkill <= 10, "6 to 10",
                                      if_else(nkill > 10 & nkill <= 50, "11 to 50", "more than 50"))))

#shorten lengthy names for visualization purpose
tmp$weapon_type[tmp$weapon_type == "Explosives/Bombs/Dynamite"] <- "Explosives"
tmp$attack_type[tmp$attack_type == "Facility/Infrastructure Attack"] <- "Facility"
tmp$target_type[tmp$target_type == "Private Citizens & Property"] <- "Civilians"
tmp$target_type[tmp$target_type == "Terrorists/Non-State Militia"] <- "Non-State"
tmp$target_type[tmp$target_type == "Religious Figures/Institutions"] <- "Religious"

#convert everything to factor
tmp[] <- lapply(tmp, factor)
str(tmp)

'data.frame': 18088 obs. of 6 variables:
 $ group_name : Factor w/ 10 levels "Al-Nusrah","Al-Shabaab",...: 8 8 8 8 8 8 8 ...
 $ target_type : Factor w/ 20 levels "Airports & Aircraft",...: 10 10 2 3 3 3 3 3 ...
 $ weapon_type : Factor w/ 8 levels "Chemical","Explosives",...: 2 3 3 3 3 3 3 3 ...
 $ attack_type : Factor w/ 8 levels "Armed Assault",...: 3 3 4 3 3 1 1 2 1 1 ...
 $ suicide_attack: Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ nkill       : Factor w/ 5 levels "0","1 to 5","11 to 50",...: 2 2 1 3 4 3 2 2 ...
```

Next, let's run the analysis by extracting data for chosen group.

4.2 Islamic State (ISIL)

Model Parameters

```
# set cut-off (threshold) values for model
params <- list(support = 0.001, confidence = 0.5, minlen = 2)
group_ISIL <- list(rhs='group_name=ISIL', default="lhs")
```

4.2.1 Explaination of key terms

The Apriori algorithm has three main measures namely Support, Confidence and Lift. These three measure are used to decide the relative strength of the rules. lhs refers to frequent pattern that is observed and rhs refers to the group selected (in this case ISIL).

Support indicates how interesting a pattern is. In the algorithm, I have set the threshold to 0.001 which means a pattern must have appeared atleast $0.001 * \text{nrow}(\text{tmp}) = 18$ times.

Confidence value i.e 0.5 (set as threshold in model params) means that in order to be included in the results, the rule has to be correct at least 50 percent of the time. This is particularly helpful to eliminate the most unreliable rules.

Lift indicates probability (support) of the itemset (pattern) over the product of the probabilities of all items in the itemset (Hahsler et al., 2018).

In general, high confidence and good lift are the standard measures to evaluate importance of a particular rule/ association however not all the rules are useful. This rules normally falls into three categories i.e. actionable, trivial(useless) and inexplicable (Klimberg & McCullough, 2017). Example of useless rule can be an association that is obvious and thus not worth mentioning.

4.2.2 Apriori model summary

```
# run model with parameters as defined above
rules <- apriori(data = tmp, parameter= params, appearance = group_ISIL)
```

Apriori

Parameter specification:

confidence	minval	smax	arem	aval	originalSupport	maxtime	support	minlen	
0.5	0.1	1	none	FALSE		TRUE	5	0.001	2
maxlen	target	ext							
10	rules	FALSE							

Algorithmic control:

filter	tree	heap	memopt	load	sort	verbose
0.1	TRUE	TRUE	FALSE	TRUE	2	TRUE

```
Absolute minimum support count: 18
```

```
set item appearances ...[1 item(s)] done [0.00s].
set transactions ...[53 item(s), 18088 transaction(s)] done [0.00s].
sorting and recoding items ... [49 item(s)] done [0.00s].
creating transaction tree ... done [0.02s].
checking subsets of size 1 2 3 4 5 6 done [0.00s].
writing ... [56 rule(s)] done [0.00s].
creating S4 object ... done [0.00s].
```

```
rules <- rules[!is.redundant(rules)] # Remove redundant rules if any
```

4.2.3 Top 5 patterns (ISIL)

```
# Extract top 5 patterns based on confidence
subrules <- head(sort(rules, by="confidence"), 5)

if( knitr:::is_latex_output() ) {
  inspect(subrules, ruleSep = ":")
} else {
  knitr:::kable(x = inspect(subrules, ruleSep = ">"), booktabs = TRUE,
                caption = "Five Most Important Patterns (ISIL)") %>%
    kable_styling(latex_options = "hold_position")
}

lhs                                rhs          support confidence lift
[1] {weapon_type=Chemical,
     attack_type=Bombing/Explosion} : {group_name=ISIL} 0.001050   0.9048 4.838
[2] {target_type=Non-State Militia,
     attack_type=Bombing/Explosion,
     nkill=6 to 10}                 : {group_name=ISIL} 0.001050   0.7308 3.907
[3] {target_type=Non-State Militia,
     attack_type=Bombing/Explosion,
     suicide_attack=1}              : {group_name=ISIL} 0.003428   0.6526 3.489
[4] {target_type=Military,
     suicide_attack=1,
     nkill=11 to 50}                : {group_name=ISIL} 0.007961   0.6457 3.453
[5] {target_type=Non-State Militia,
     suicide_attack=1}              : {group_name=ISIL} 0.003483   0.6238 3.335

if( knitr:::is_latex_output() ) {
  plot(rules)
} else {
  plotly_arules(rules, jitter = 5,
                marker = list(opacity = .5, size = 10),
                colors = viridis(10, end = 0.9, option = "D")) %>%
```

```
    layout(title = "Association Rules in ISIL Group")
}
```

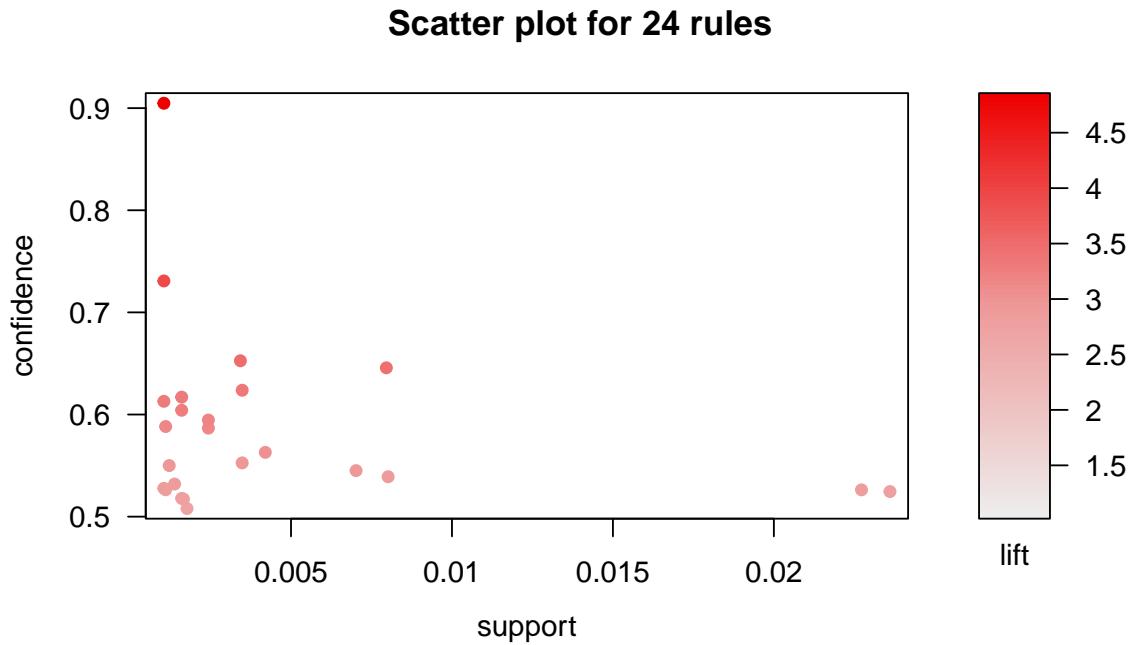


Figure 4.1: Association Rules in ISIL Group

4.2.4 Network graph (ISIL)

```
#work around for pdf and html output
if( knitr:::is_latex_output() ) {
  plot(rules, method="graph", verbose = FALSE,
       control=list(nodeCol="orange", edgeCol="#9cb7f4"))
} else {
  ig_df <- get.data.frame(
    plot(rules, method="graph", verbose = FALSE,
         control=list(nodeCol="orange", edgeCol="#9cb7f4)), what = "both")

  nodes = data.frame(
    id = ig_df$vertices$name,
    value = ig_df$vertices$support, # get the nodes by support
    title = ifelse(ig_df$vertices$label == "", ig_df$vertices$name, ig_df$vertices$label)
    ig_df$vertices)

  visNetwork(nodes, edges = ig_df$edges) %>%
  visEvents() %>%
```

```

visNodes(size = 5, color = "#9cb7f4") %>%
visLegend() %>%
visEdges(smooth = TRUE, color = "#ffd596" ) %>%
visOptions(highlightNearest = TRUE, nodesIdSelection = TRUE) %>%
visEdges(arrows = 'from') %>%
visPhysics(solver = "barnesHut", maxVelocity = 35,
            forceAtlas2Based = list(gravitationalConstant = -6000))
}

```

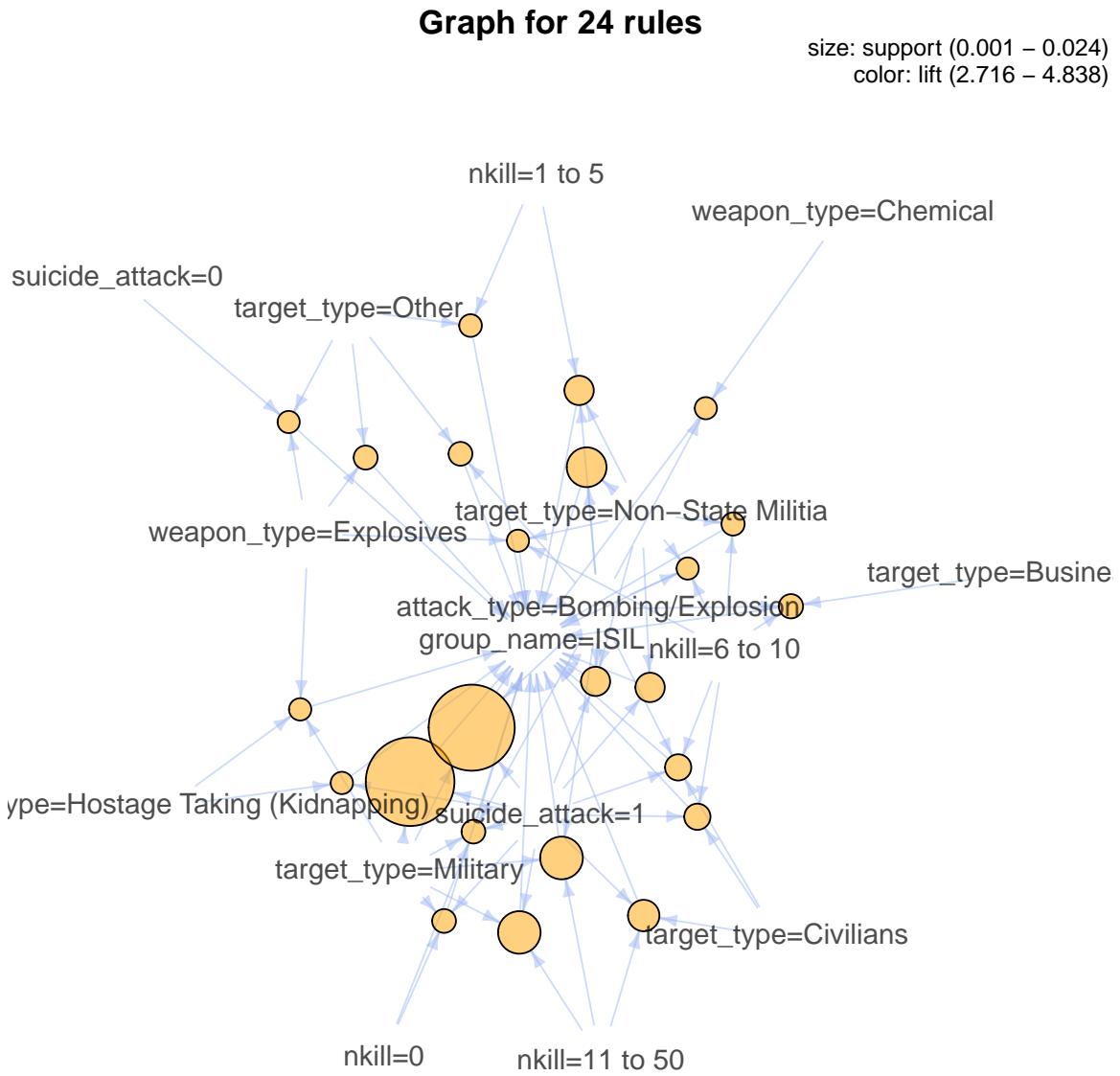


Figure 4.2: Network Graph of Discovered Patterns- ISIL Group

4.3 Taliban

4.3.1 Apriori model summary

```
params <- list(support = 0.001, confidence = 0.5, minlen = 2)
group_Taliban <- list(rhs='group_name=Taliban', default="lhs")
rules <- apriori(data = tmp, parameter= params, appearance = group_Taliban)
```

Apriori

Parameter specification:

confidence	minval	smax	arem	aval	originalSupport	maxtime	support	minlen	
0.5	0.1	1	none	FALSE		TRUE	5	0.001	2
maxlen	target	ext							
10	rules	FALSE							

Algorithmic control:

filter	tree	heap	memopt	load	sort	verbose
0.1	TRUE	TRUE	FALSE	TRUE	2	TRUE

Absolute minimum support count: 18

```
set item appearances ...[1 item(s)] done [0.00s].
set transactions ...[53 item(s), 18088 transaction(s)] done [0.00s].
sorting and recoding items ... [49 item(s)] done [0.00s].
creating transaction tree ... done [0.01s].
checking subsets of size 1 2 3 4 5 6 done [0.00s].
writing ... [139 rule(s)] done [0.00s].
creating S4 object ... done [0.00s].
```

```
rules <- rules[!is.redundant(rules)] # Remove redundant rule if any
```

4.3.2 Top 5 patterns (Taliban)

	lhs	rhs	support	confidence	lift
[1]	{weapon_type=Chemical, attack_type=Unarmed Assault} : {group_name=Taliban}		0.001216	0.8800	2.955
[2]	{target_type=Police, weapon_type=Firearms, attack_type=Armed Assault, nkill=11 to 50}	: {group_name=Taliban}	0.004976	0.8257	2.772
[3]	{target_type=Police, weapon_type=Firearms, nkill=6 to 10}	: {group_name=Taliban}	0.010117	0.8243	2.768

```
[4] {target_type=Police,
     weapon_type=Incendiary,
     attack_type=Facility/Infra.,
     nkill=0} : {group_name=Taliban} 0.001990 0.8000 2.686
[5] {target_type=Police,
     weapon_type=Firearms,
     nkill=11 to 50} : {group_name=Taliban} 0.005639 0.7969 2.676
```

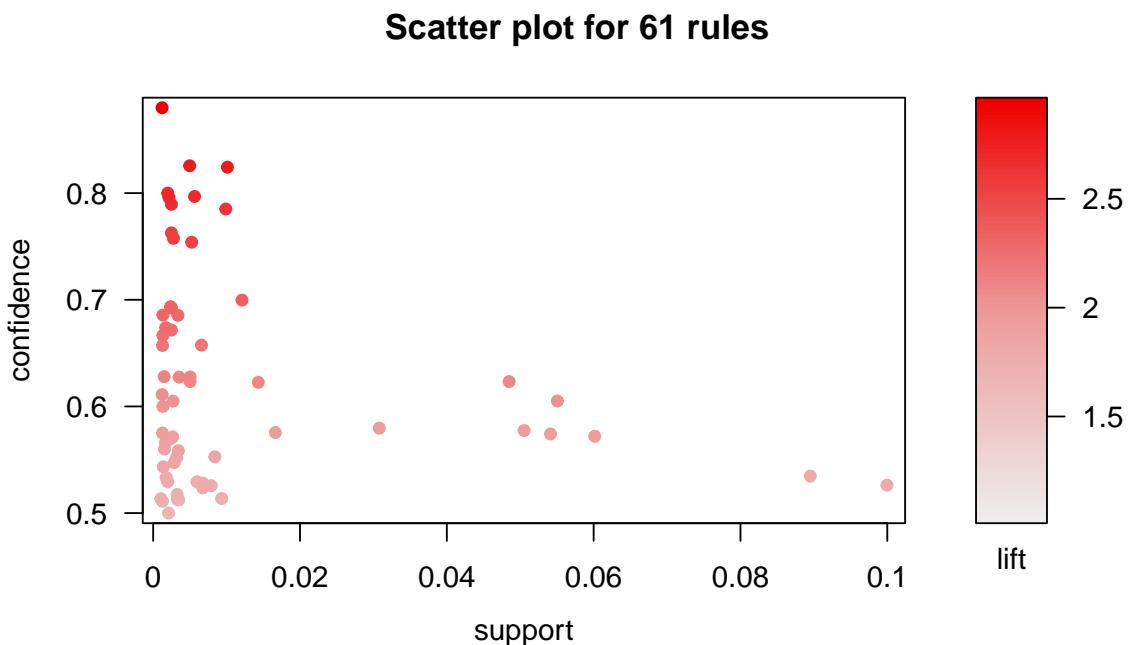


Figure 4.3: Association Rules in Taliban Group

In case of Taliban, we can see many interesting patterns with confidence above 0.5. For the visualization purpose, let's narrow down to most interesting rules only by setting confidence threshold to 0.6.

4.3.3 Network graph (Taliban)

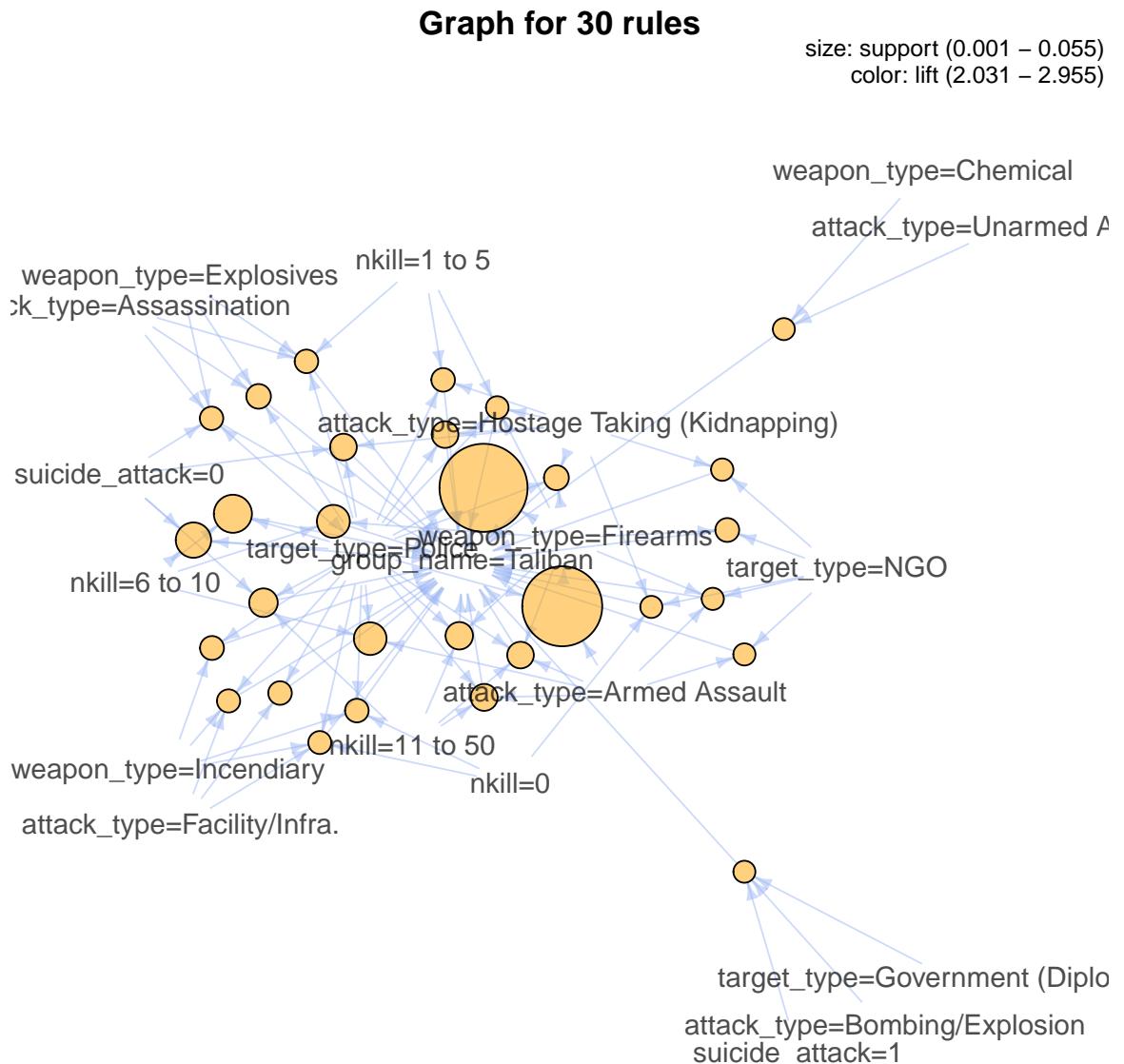


Figure 4.4: Network Graph of Discovered Patterns- Taliban Group

4.4 Boko Haram

4.4.1 Apriori model summary

```

params <- list(support = 0.001, confidence = 0.5, minlen = 2)
group_Boko_Haram <- list(rhs='group_name=Boko Haram', default="lhs")
rules <- apriori(data = tmp, parameter= params, appearance = group_Boko_Haram)

Apriori

Parameter specification:
confidence minval smax arem aval originalSupport maxtime support minlen
          0.5    0.1     1 none FALSE           TRUE      5  0.001     2
maxlen target   ext
       10  rules FALSE

Algorithmic control:
filter tree heap memopt load sort verbose
  0.1 TRUE TRUE FALSE TRUE    2    TRUE

Absolute minimum support count: 18

set item appearances ...[1 item(s)] done [0.00s].
set transactions ...[53 item(s), 18088 transaction(s)] done [0.01s].
sorting and recoding items ... [49 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 2 3 4 5 6 done [0.00s].
writing ... [63 rule(s)] done [0.00s].
creating S4 object ... done [0.00s].
rules <- rules[!is.redundant(rules)] # Remove redundant rule if any

```

4.4.2 Top 5 patterns (Boko Haram)

	lhs	rhs	support	confidence	lift
[1]	{target_type=Civilians, weapon_type=Explosives, suicide_attack=0, nkill=more than 50}	: {group_name=Boko Haram}	0.001106	0.8000	7.689
[2]	{target_type=Civilians, weapon_type=Explosives, attack_type=Armed Assault, nkill=11 to 50}	: {group_name=Boko Haram}	0.001106	0.7692	7.393
[3]	{target_type=Civilians, attack_type=Armed Assault, nkill=more than 50}	: {group_name=Boko Haram}	0.001548	0.7568	7.273
[4]	{target_type=Civilians, weapon_type=Explosives, attack_type=Armed Assault,				

```
nkill=6 to 10} : {group_name=Boko Haram} 0.001382 0.7353 7.06  
[5] {target_type=Civilians,  
    weapon_type=Incendiary,  
    attack_type=Armed Assault} : {group_name=Boko Haram} 0.001050 0.6786 6.52
```

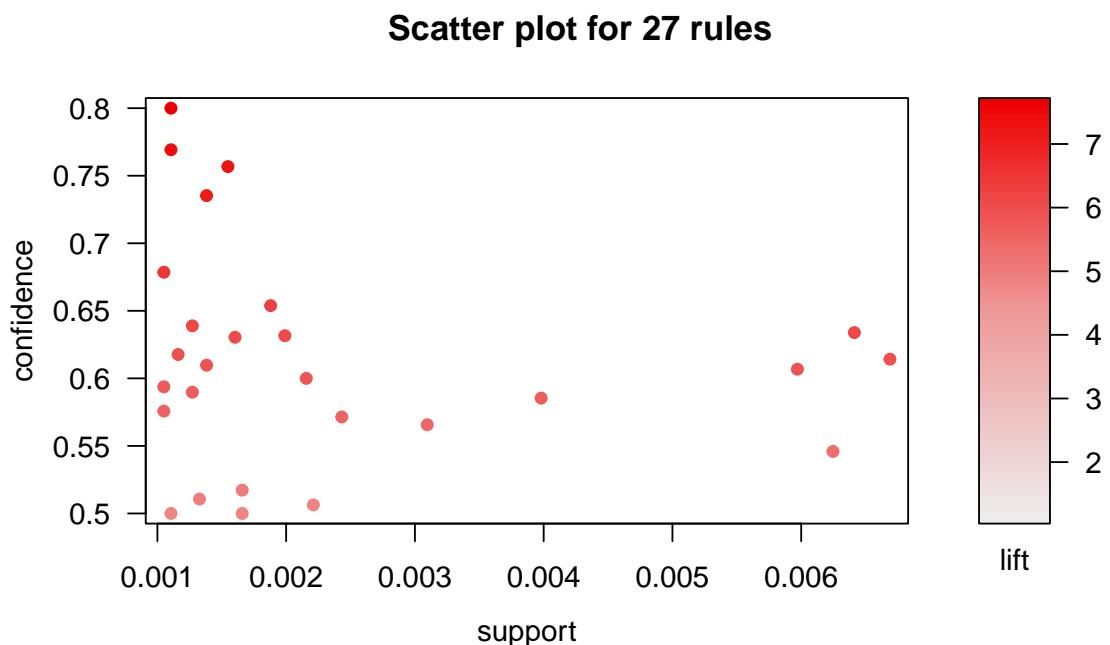


Figure 4.5: Association Rules in Boko Haram Group

4.4.3 Network graph (Boko Haram)

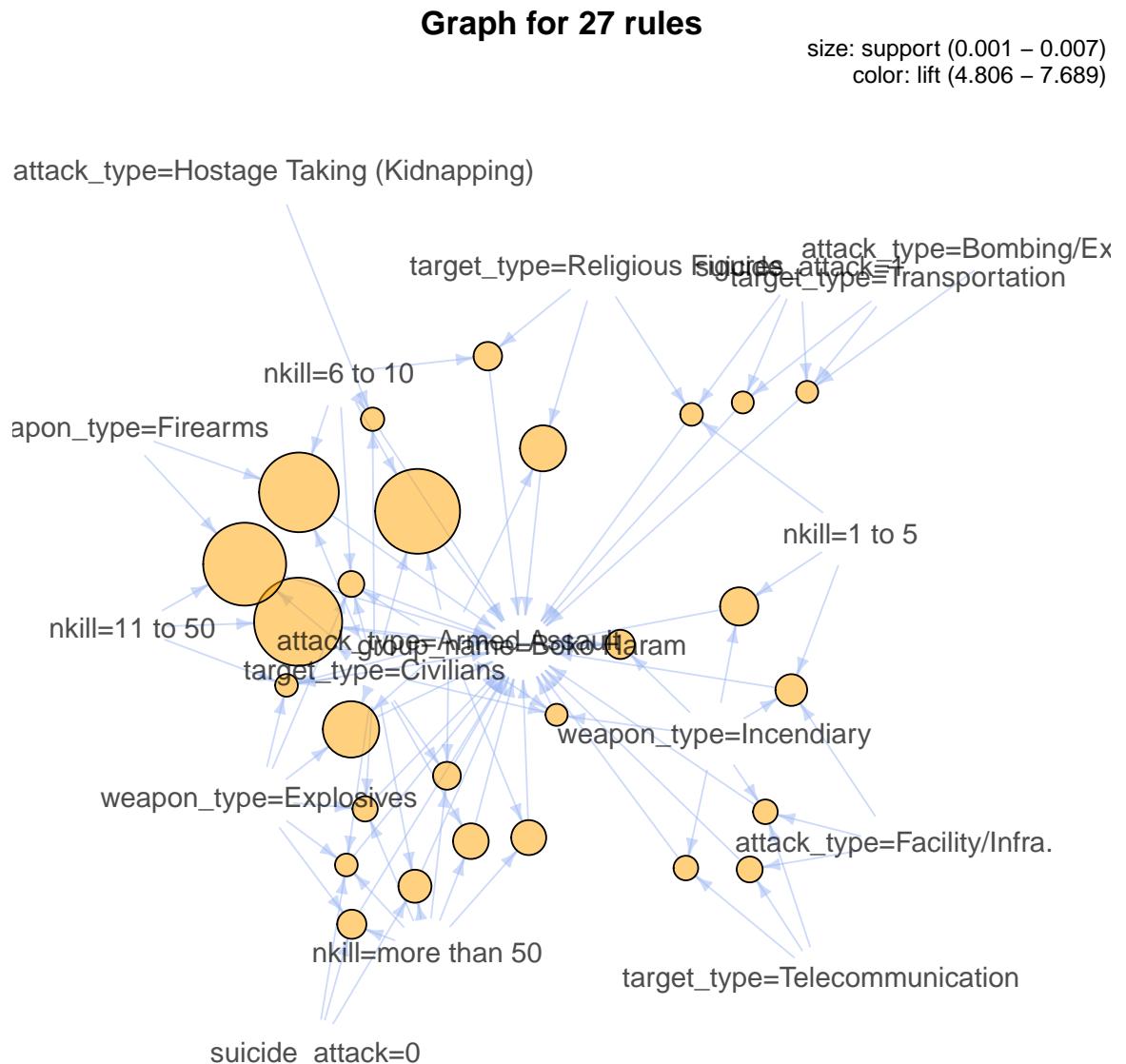


Figure 4.6: Network Graph of Discovered Patterns- Boko Haram Group

Chapter 5

Time-series Forecasting

Time-series forecasting is a supervised machine learning approach that uses historical data to predict future occurrences. This is particularly helpful in terrorism context for long term strategic planning. For this analysis, the forecasting goal and corresponding data is chosen as below: For each analysis, first we select the appropriate data,

Table 5.1: Scope of Analysis

Forecasting_Goal	Frequency	Chosen_Country
Predict future number of attacks	By Months	Afghanistan
Predict future number of fatalities	By Months	Iraq
Predict future number of attacks	By Months	SAHEL region + Somalia

examine seasonal components and then split the data in training and test set to evaluate performance of Auto Arima, Neural Network, TBATS and ETS models with seven different metrics. To examine whether an ensemble predictions can improve the overall accuracy, we take the average of all the predictions and compute Theil's U statistic. In the last part of the analysis, we use all the data points (train + test) to make forecast for chosen future period.

5.1 Afghanistan (Predict future attacks)

5.1.1 Data preparation

Based on exploratory data analysis, it is observed that the number attacks with visible pattern began from year 2000 so the data is selected between year 2000 to 2016. To get the time-series frequency by months for all the years, I add missing months and assign zero as shown in the code below:

```

dft <- df %>%
  filter(year >= 2000 & country == "Afghanistan") %>%
  group_by(year, month) %>%
  summarise(total_count = n()) %>%
  ungroup() %>%
  group_by(year) %>%
  # Add missing months and assign 0 where no occurrences
  tidyr::complete(month = full_seq(seq(1:12), 1L), fill = list(total_count = 0))
  ungroup()

dft <- dft %>%
  mutate(month_year = paste(year, month, sep="-"),
         month_year = zoo::as.yearmon(month_year)) %>%
  select(month_year, total_count)

# Create a ts object
dft <- ts(dft[, 2], start = Year(min(dft$month_year)),
          frequency = 12) # 1=annual, 4=quarterly, 12=monthly
dft <- na.kalman(dft)

```

5.1.2 Seasonality analysis

```

if( knitr:::is_latex_output() ) {
  include_graphics(path = "figure/afg_line_plot.png")
} else {
  ts_plot(dft, line.mode = "line", Xtitle = "Year", Ytitle = "Attack Counts",
          title = "Attack Frequency (Afghanistan)", color = "red")
}

```

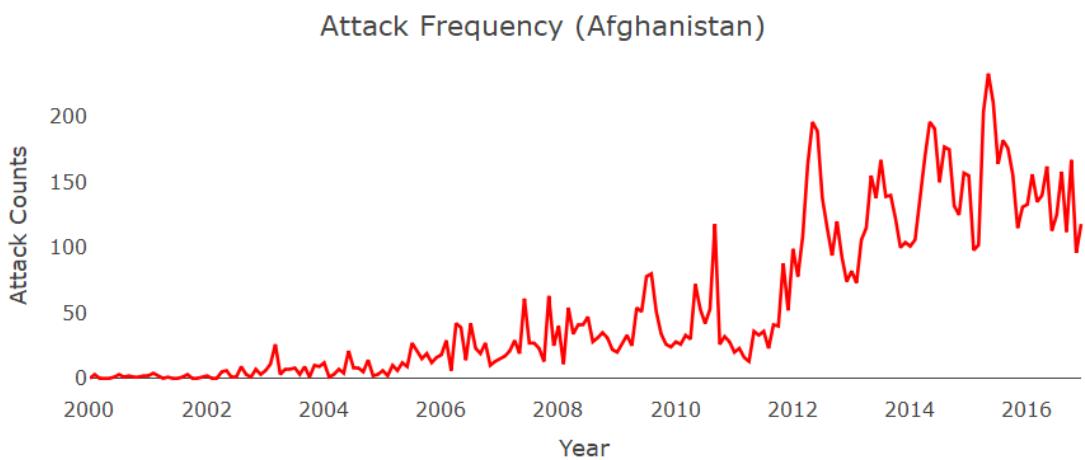


Figure 5.1: Attack Frequency by Year- Afghanistan

```

if( knitr:::is_latex_output() ) {
  include_graphics(path = "figure/afg_seasonal.png")
} else {
  ts_seasonal(dft, type = "cycle", Ygrid = T, Xgrid = T,
              title = "Seasonality Plot (Afghanistan)")
}

```

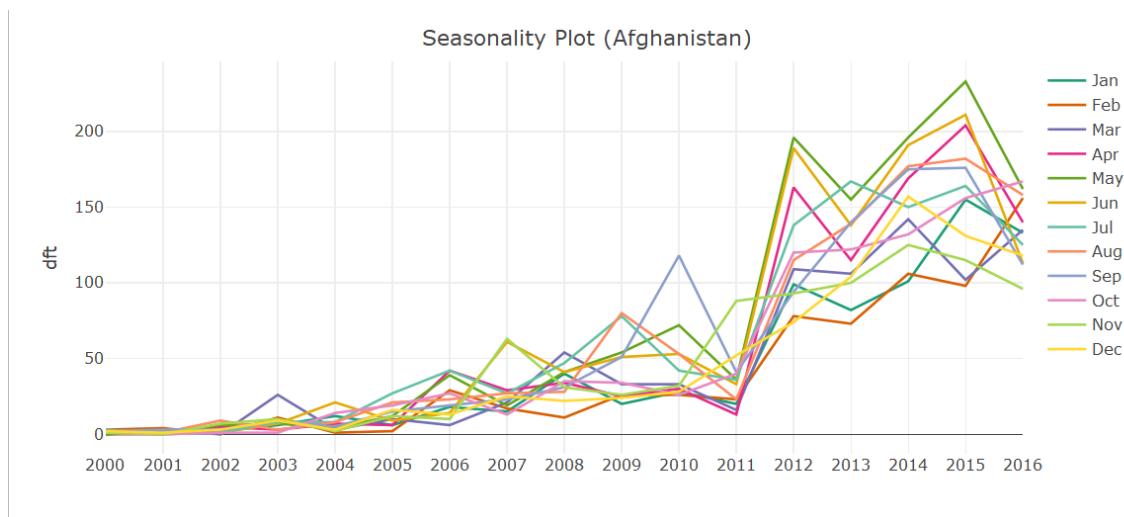


Figure 5.2: Seasonal Pattern Within Year- Afghanistan

From the seasonal patterns within year as shown in the plot above, we can see that year 2015 (followed by 2012) was the deadliest year in terms of number of terror attacks. In both years, spike is visible in May month.

```

if( knitr:::is_latex_output() ) {
  include_graphics(path = "figure/afg_box.png")
} else {
  ts_seasonal(dft, type = "box", Ygrid = T, Xgrid = T,
              title = "Seasonality Plot (Afghanistan)")
}

```

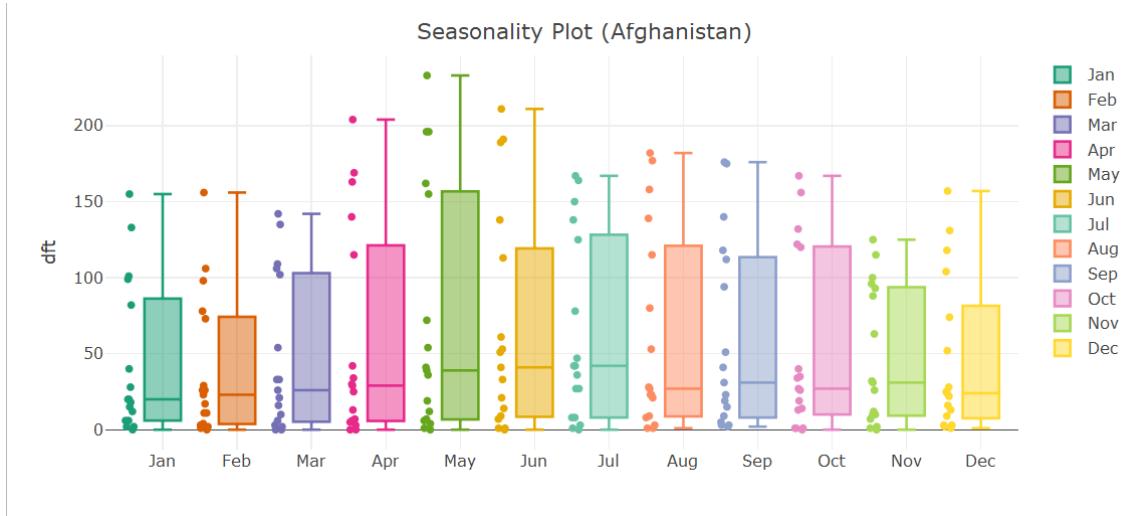


Figure 5.3: Seasonal Pattern (boxplot)- Afghanistan

From the boxplot, we can confirm that the May month contributes the most in terms of terrorist incidents throughout all the years (2000-2016) in Afghanistan. It's also worth mentioning that the Ramadan (holy month in Islamic calendar) usually starts in May and ends in June every year. From the boxplot above, we see the upward trend in number of attacks starting from February and reaching peak in Ramadan months.

```
if( knitr:::is_latex_output() ) {
  include_graphics(path = "figure/afg_decompose.png")
} else {
  ts_decompose(dft, type = "both")
}
```

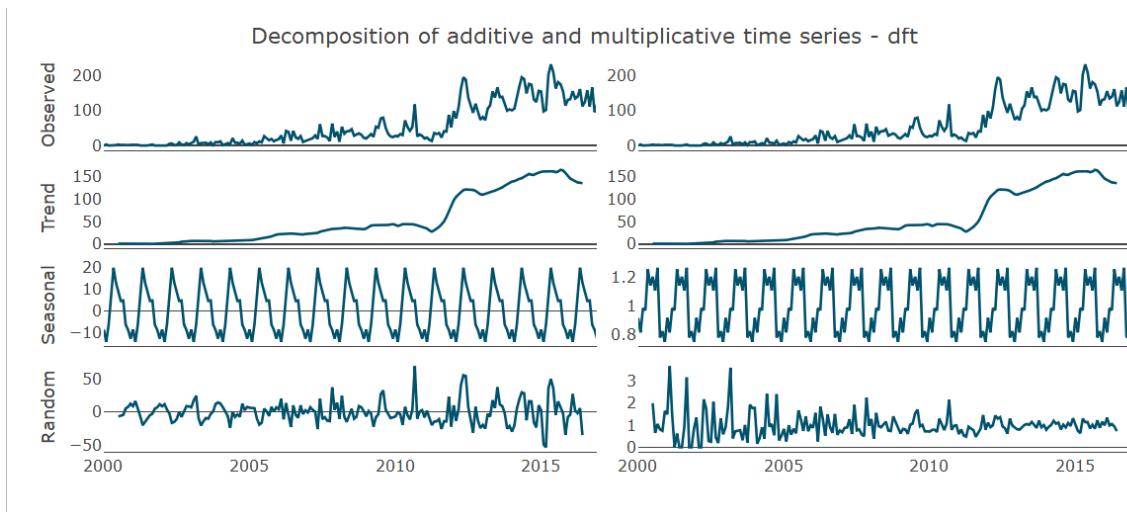


Figure 5.4: Time-series Decomposition- Afghanistan

Decomposition by additive and multiplicative time-series is helpful to understand anomalies in data. Based on observed patterns, this decomposition comprises three components: a seasonal component, a trend-cycle component (containing both trend and cycle), and a remainder component (containing anything else in the time series) (Hyndman & Athanasopoulos, 2018).

In simple words, seasonal component means pattern that occur frequently within fixed period of time. Random component is also called noise/ remainder and it represents residuals of the original time-series after removing seasonal and trend component (Anomaly.io, 2015).

5.1.3 Correlation test

There are several methods to identify correlation between series and lags such as ACF, PACF and lag plots. I use lag plots method for this analysis which allows us to quickly visualize outliers and randomness and auto-correlation.

```
if( knitr:::is_latex_output() ) {
  include_graphics(path = "figure/afg_lags.png")
} else {
  ts_lags(dft, lag.max = 9)
}
```

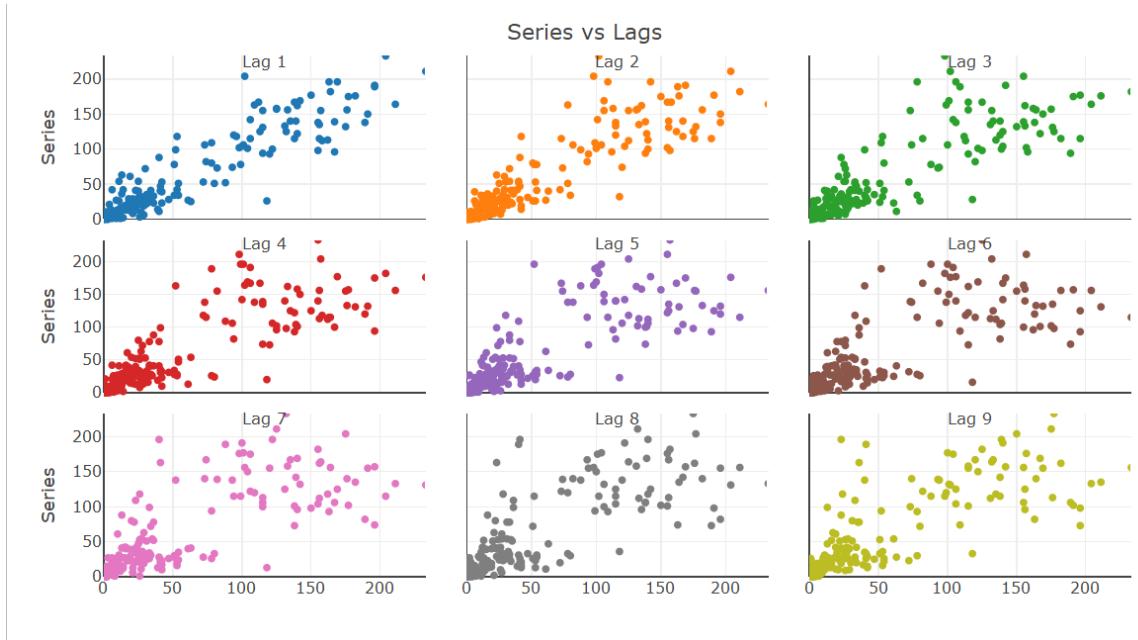


Figure 5.5: Correlation Test

Although we can spot outliers in all nine lags however linear pattern is visible. To further explain this, we can see positive linear trend (going upward from left to right) which is an indication that positive auto-correlation is present. Specifically, lags 1, 2,

3 and 9 shows strong positive auto-correlation. Presence of auto-correlation can be problematic for some models.

5.1.4 Modelling

In this part of the analysis, I split the data in training and test set in order to evaluate performance of four different models before making the actual forecasts.

Train-Test Split

```
set.seed(84)
# horizon (look ahead period)
horizon <- 12
# create split for train and test set
data <- ts_split(dft, sample.out = horizon)
# Split the data into training and testing sets
train <- data$train
test <- data$test
```

Auto Arima

```
fit_arima <- auto.arima(train)
# plot the residuals
if( knitr:::is_latex_output() ) {
  include_graphics(path = "figure/afg_residuals.png")
} else {
  check_res(fit_arima)
}
```

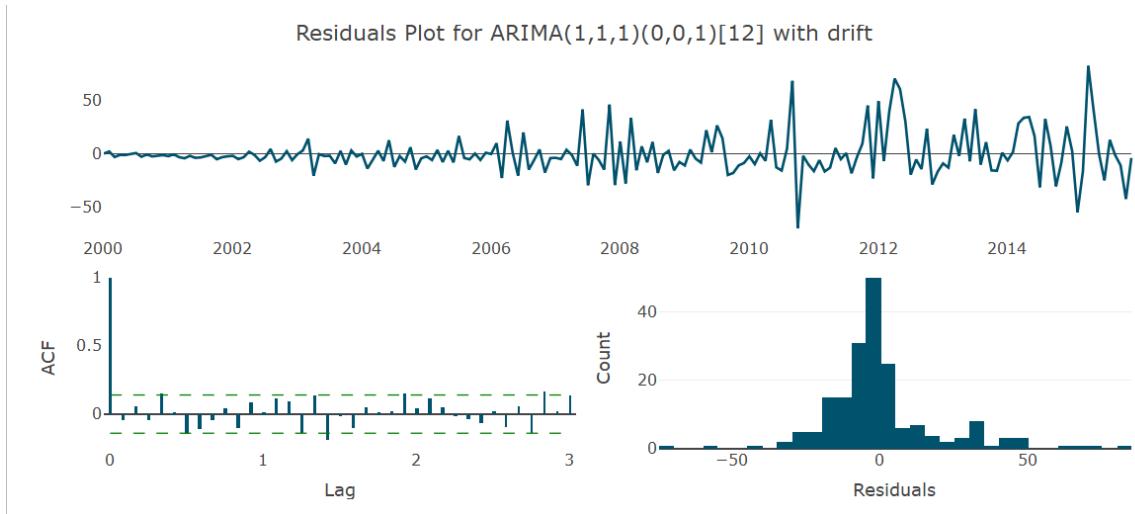


Figure 5.6: Auto Arima: Residuals

A quick look at residuals from Auto Arima suggests that the mean of residuals is very close to zero however from the histogram, we can see that residuals doesn't follow the normal distribution. Specifically, the right tail is little too long. What this means is, forecasts from this method will probably be quite good but prediction intervals computed assuming a normal distribution may be inaccurate (Hyndman & Athanasopoulos, 2018).

```
# Accuracy check/ Forecast evaluation
fc_arima <- forecast(fit_arima, h = horizon)
#plot actual vs fitted and forecasted
if( knitr:::is_latex_output() ) {
  include_graphics(path = "figure/afg_arima_fitted.png")
} else {
  test_forecast(actual = dft, forecast.obj = fc_arima, test = test) %>%
  layout(legend = list(x = 0.1, y = 0.9))
}
```

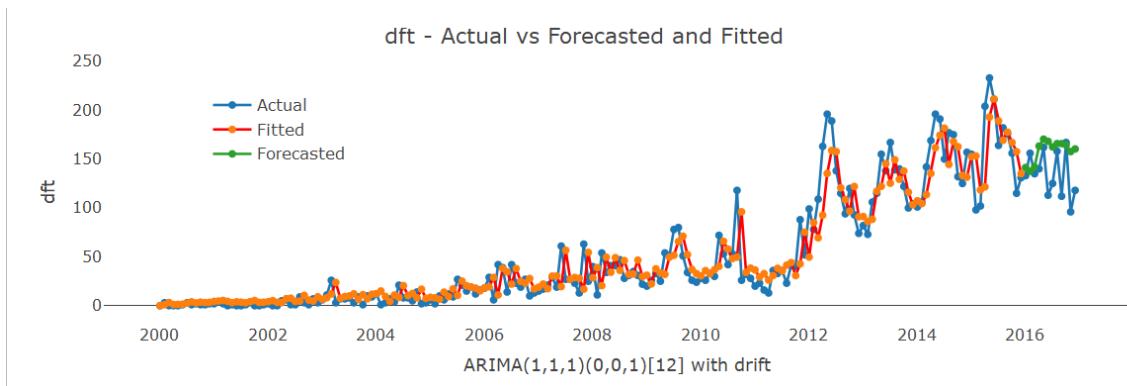


Figure 5.7: Auto Arima: Actual vs Fitted vs Forecasted

From the plot above, it is observed that Auto Arima model nearly captures fitted values based on training data but forecasted values little bit apart from actual values (test data). Next, we observe the pattern in actual vs fitted and forecasted values for remaining three models.

Neural Network

```
fit_nn <- nnetar(train, repeats = 5)
# Accuracy check/ Forecast evaluation
fc_nn <- forecast(fit_nn, h = horizon)
#plot actual vs fitted and forecasted
if( knitr:::is_latex_output() ) {
  include_graphics(path = "figure/afg_nn_fitted.png")
} else {
  test_forecast(actual = dft, forecast.obj = fc_nn, test = test) %>%
  layout(legend = list(x = 0.1, y = 0.9))
}
```

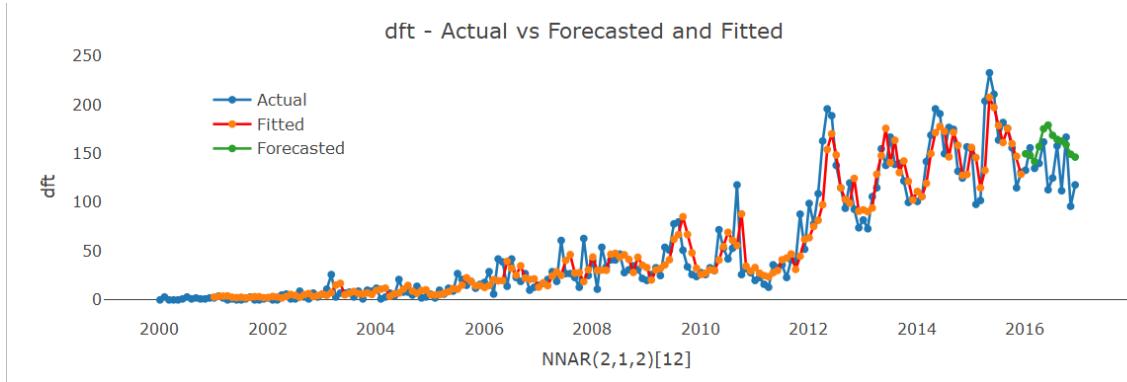


Figure 5.8: Neural Net: Actual vs Fitted vs Forecasted

TBATS

```
fit_tbats <- tbats(train)
# Accuracy check/ Forecast evaluation
fc_tbats <- forecast(fit_tbats, h = horizon)
#plot actual vs fitted and forecasted
if( knitr:::is_latex_output() ) {
  include_graphics(path = "figure/afg_tbats_fitted.png")
} else {
  test_forecast(actual = dft, forecast.obj = fc_tbats, test = test) %>%
  layout(legend = list(x = 0.1, y = 0.9))
}
```

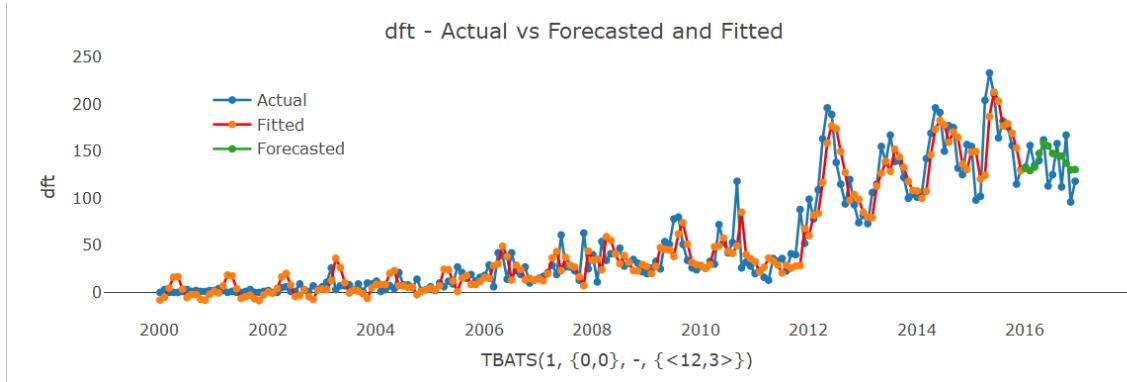


Figure 5.9: TBATS: Actual vs Fitted vs Forecasted

ETS

```
fit_ets <- ets(train)
# Accuracy check/ Forecast evaluation
fc_ets <- forecast(fit_ets, h = horizon)
#plot actual vs fitted and forecasted
if( knitr:::is_latex_output() ) {
  include_graphics(path = "figure/afg_ets_fitted.png")
} else {
  test_forecast(actual = dft, forecast.obj = fc_ets, test = test) %>%
    layout(legend = list(x = 0.1, y = 0.9))
}
```

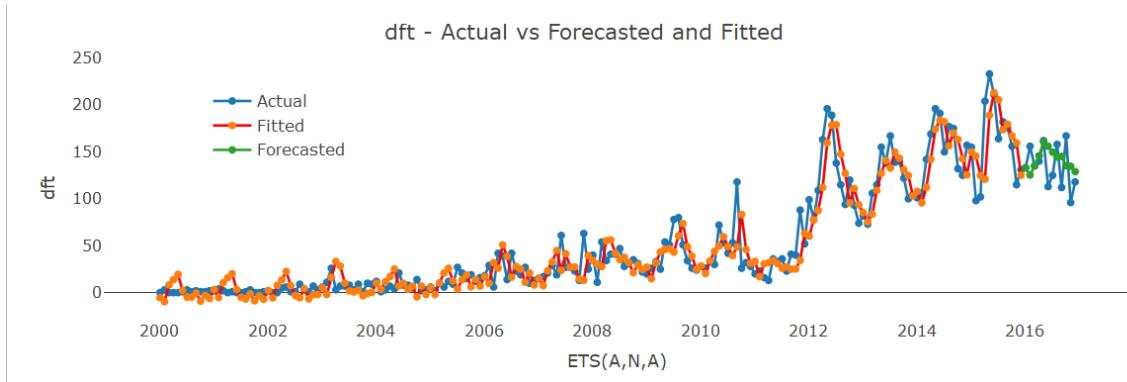


Figure 5.10: ETS: Actual vs Fitted vs Forecasted

5.1.5 Evaluating models' Performance

To compare the performance of all four models on test data, I have extracted mean accuracy from each model and have arranged the models by MAPE metric which is

most commonly used. We will also look at six other metrics to get better idea about model's performance.

Out of all the seven metrics as shown in the table below, ME (Mean Error), RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error) are scale-dependent error. Whereas MPE (Mean Percentage Error) and MAPE (Mean Absolute Percent Error) are percentage errors and ACF stands for first-order correlation. Researchers (Hyndman & Athanasopoulos, 2018) suggest that percentage errors have the advantage of being unit-free, and so are frequently used to compare forecast performances between data sets.

```
metrics <- rbind(as.data.frame(round(accuracy(fc_arima$mean, test), 3)),
                  as.data.frame(round(accuracy(fc_nn$mean, test), 3)),
                  as.data.frame(round(accuracy(fc_tbats$mean, test), 3)),
                  as.data.frame(round(accuracy(fc_ets$mean, test), 3))) %>%
  add_column(models = c("Auto Arima", "NeuralNet", "TBATS", "ETS"),
             .before = "ME") %>% arrange(MAPE)

if( knitr:::is_latex_output() ) {
  knitr::kable(metrics, booktabs = TRUE,
               caption = "Performance comparison of all models (Afghanistan)") %>%
    kable_styling(full_width = F, latex_options = "hold_position",
                  font_size = 12)
} else {
  knitr::kable(metrics,
               caption = "Performance comparison of all models (Afghanistan)") %>%
    kable_styling(bootstrap_options = c("striped", "hover"), full_width = F,
                  font_size = 13) %>%
    column_spec(1, bold = T, color = "black") %>%
    column_spec(2:8, color = "black", background = "#dee2ed") %>%
    column_spec(6, color = "black", background = "#c7cfe5") %>%
    row_spec(1, background = "#c7cfe5")
}
```

Table 5.2: Performance comparison of all models (Afghanistan)

models	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
TBATS	-6.374	23.34	18.93	-7.424	15.29	-0.303	0.645
ETS	-6.615	24.75	19.50	-7.781	15.79	-0.315	0.684
NeuralNet	-24.098	33.63	26.68	-21.038	22.64	-0.170	0.899
Auto Arima	-23.698	34.26	27.21	-20.991	23.22	-0.064	0.953

Based on MAPE metrics, we can see that TBATS and ETS models achieves the higher accuracy (~ 15) and out performs Auto Arima and Neural Network models. TBATS (Exponential Smoothing State Space Model With Box-Cox Transformation) and ETS (Exponential Smoothing State Space Model) both uses exponential smoothing method. Specifically, TBATS modelling approach offers several key advantages

such as handling of typical non linear features and allowing any auto-correlation in the residuals to be taken into account (Livera, Hyndman, & Snyder, 2011).

In addition to MAPE metric which is chosen to identify the best model, we also look at Theil's U statistic to estimate how good or bad the model is. In simple words, Theil's U-statistic compares the performance of model with naïve/ random walk model($U=1$). If Theil's U statistic value equals one, it means that the model forecasting method is as good as naïve model (guessing). Value greater than one means the forecasting method is even worst than guessing. Similarly, value less than 1 indicates that forecasting method is better than naïve model and worth considering (Oracle, n.d.).

From the comparison, we can see that all four models have Theil's U score less than one while TBATS and ETS models having comparatively good score of 0.6 compared to Neural Network at 0.95.

5.1.6 Ensemble

As stated in literature review, many research focuses on single model approach or using the best single model out of all the models. Instead of throwing out weak models, I employ simple ensemble approach (averaging predictions of all four models) to improve the overall accuracy on test set. This is one of the well-known approach used in machine learning competitions such as on Kaggle (Jacob van Veen, Nguyen, Dat, & Segnini, 2015). Following is the code used to extract predictions from all four models and then new column “ensemble” is added which take the average of all models. Next, we calculate Theil's U score on ensemble predictions using a simple function in DescTools package by supplying actual observations and predicted observations as shown below:

```
# extract predictions from all four models and get average
ensemble <- rowMeans(
  cbind(fc_arima$mean, fc_nn$mean, fc_tbats$mean, fc_ets$mean))
# Compute Theil's U statistic (a = actual values, p= predicted values)
TheilU(a = test, p = ensemble)
```

[1] 0.2043

Although TBATS model is our best single model however ensemble predictions by averaging forecasts of other weak models is even better. We can see that the ensemble approach significantly improves the overall accuracy as measured by Theil's U score of 0.2. The most recent theoretical framework also supports the ensemble approach in time-series forecasting. Researchers (Hyndman & Athanasopoulos, 2018), in their book “Forecasting: Principles and Practice”, suggests that using several different methods on the same time-series data and then averaging the results of forecast often guarantees better performance than any single best models.

To summarize, it is possible that TBATS model may not be the best model on other data however use of ensemble approach and corresponding Theil's U score can be used

in time-series forecasting to improve the accuracy and justify the reliability of final predictions.

5.1.7 Forecast future number of attacks

As we have evaluated performance of all four models, the next step of process is to generate forecast using all the data points i.e 2000-2016. The forecast horizon can be changed based on business requirement and by observing the predictions. For this part, I use 18 months period as a look ahead period. As shown in the code chunk below, first we will generate forecasts from all four models and then we will visualize the results with plots.

```
# look ahead period
f_horizon <- 18
# run model on full data i.e dft (2000-2016)
fore_arima <- forecast(auto.arima(dft), h = f_horizon, level = c(80, 95))
fore_nn <- forecast(nnetar(dft, repeats = 5), h = f_horizon, level = c(80, 95),
fore_tbats <- forecast(tbats(dft), h = f_horizon, level = c(80, 95))
fore_ets <- forecast(ets(dft), h = f_horizon, level = c(80, 95))

# combine plots for latex output
if( knitr:::is_latex_output() ) {
  plot(fore_arima)
  plot(fore_nn)
  plot(fore_tbats)
  plot(fore_ets)
}
```

Forecasts from ARIMA(1,1,2)(0,0,1)[12] with drift

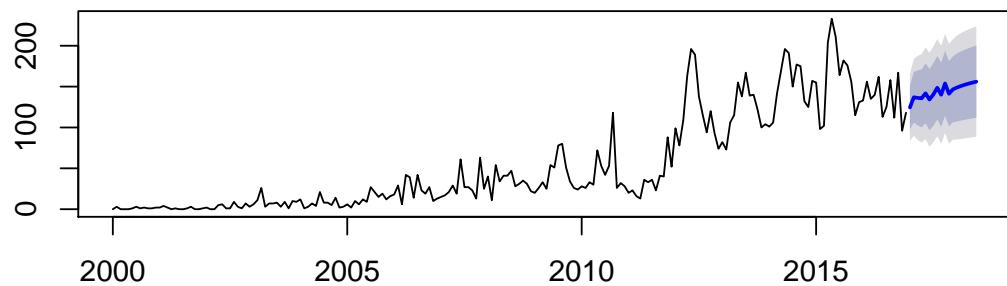


Figure 5.11: Predicted Number of Attacks in Afghanistan (plots)

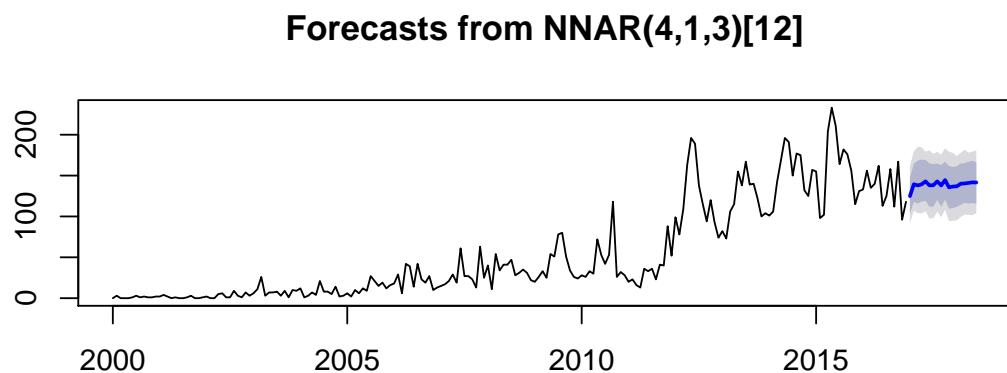


Figure 5.12: Predicted Number of Attacks in Afghanistan (plots)

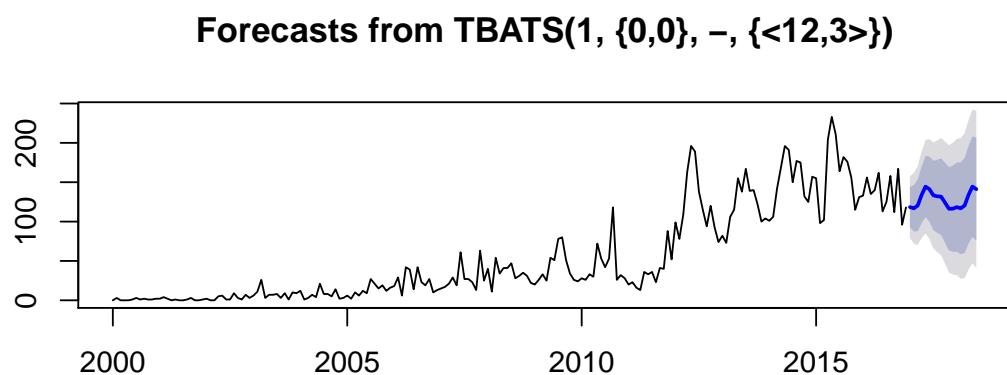


Figure 5.13: Predicted Number of Attacks in Afghanistan (plots)

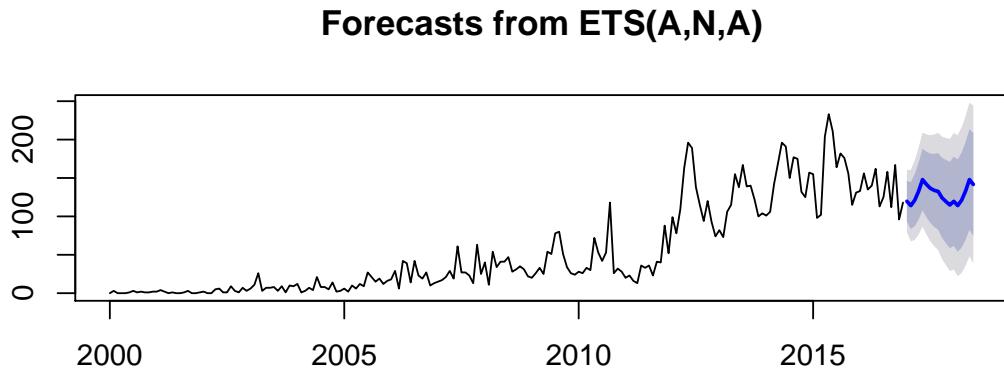


Figure 5.14: Predicted Number of Attacks in Afghanistan (plots)

The forecasting results are often represented by mean value and by confidence interval of 80% and 95%. The line in orange represents the mean value of forecast and considered as final forecasting value. Next, we extract forecasts for chosen horizon and add the ensembled predictions to be used as insights as predicted future attacks in Afghanistan as shown in the code chunk below:

```
tbl_arima    <- timetk::tk_tbl(round(fore_arima$mean))
tbl_nn       <- timetk::tk_tbl(round(fore_nn$mean))
tbl_tbats   <- timetk::tk_tbl(round(fore_tbats$mean))
tbl_ets      <- timetk::tk_tbl(round(fore_ets$mean))

tbl <- tbl_arima %>%
  left_join(tbl_nn, by = "index") %>%
  left_join(tbl_tbats, by = "index") %>%
  left_join(tbl_ets, by = "index")

names(tbl) <- c("Time_period", "Arima", "NN", "TBATS", "ETS")
tbl$Ensemble <- round(rowMeans(tbl[,2:5]))

if( knitr:::is_latex_output() ) {
  knitr::kable(tbl, booktabs = TRUE,
  caption = "Table of Predicted Future Attacks in Afghanistan") %>%
  kable_styling(full_width = F, latex_options = "hold_position",
  font_size = 12)
} else {
  knitr::kable(tbl, caption = "Table of Predicted Future Attacks in Afghanistan")
  kable_styling(bootstrap_options = c("striped", "hover"), full_width = F,
  font_size = 13, position = "left") %>%
  column_spec(1:5, color = "black", background = "#dee2ed") %>%
```

```
column_spec(1, bold = T) %>%
  column_spec(6, bold = T, color = "black", background = "#c7cfec")
}
```

Table 5.3: Table of Predicted Future Attacks in Afghanistan

Time_period	Arima	NN	TBATS	ETS	Ensemble
Jan 2017	124	125	118	120	122
Feb 2017	137	139	117	114	127
Mar 2017	136	138	120	121	129
Apr 2017	136	139	134	133	136
May 2017	142	143	144	148	144
Jun 2017	134	138	141	142	139
Jul 2017	140	138	133	137	137
Aug 2017	149	143	132	134	140
Sep 2017	140	138	132	133	136
Oct 2017	154	145	124	124	137
Nov 2017	141	136	116	119	128
Dec 2017	147	137	116	115	129
Jan 2018	149	137	118	120	131
Feb 2018	151	140	117	114	130
Mar 2018	152	140	120	121	133
Apr 2018	154	141	134	133	140
May 2018	155	142	144	148	147
Jun 2018	156	142	141	142	145

5.2 Iraq (Predict future fatalities)

For this analysis, we use the exact same approach as before to estimate the number of fatalities in Iraq.

5.2.1 Data preparation

I have selected the data between 2004 to 2016 to make it appropriate for the modelling. Wherever an incident is part of multiple attacks, we have different reported figures from different sources. To overcome this issue, I group the data on specific variable and then take the maximum reported value as shown in the code chunk below:

```
dft <- df %>%
  filter(year >= 2004 & country == "Iraq") %>%
  replace_na(list(nkill = 0)) %>%
```

```

group_by(group_name, region, year, month) %>%
filter(if_else(part_of_multiple_attacks == 1,
               nkill == max(nkill), nkill == nkill)) %>%
ungroup() %>%
distinct(group_name, region, country, year, month, nkill,
         nwound, part_of_multiple_attacks) %>%
group_by(year, month) %>%
summarise(total_count = sum(nkill)) %>%
ungroup() %>%
group_by(year) %>%
# Add missing months and assign 0 where no occurrence
tidyr::complete(month = full_seq(seq(1:12), 1L), fill = list(total_count = 0))
ungroup()

dft <- dft %>%
mutate(month_year = paste(year, month, sep="-"),
       month_year = zoo::as.yearmon(month_year)) %>%
select(month_year, total_count)

# Create a ts object
dft <- ts(dft[, 2], start = Year(min(dft$month_year)),
           frequency = 12) # 1=annual, 4=quarterly, 12=monthly
dft <- na.kalman(dft)

```

5.2.2 Seasonality analysis

```

if( knitr:::is_latex_output() ) {
  include_graphics(path = "figure/iraq_line_plot.png")
} else {
  ts_plot(dft, line.mode = "line", Xtitle = "Year", Ytitle = "Fatalities Counts"
          title = "Fatalities Frequency (Iraq)", color = "red")
}

```

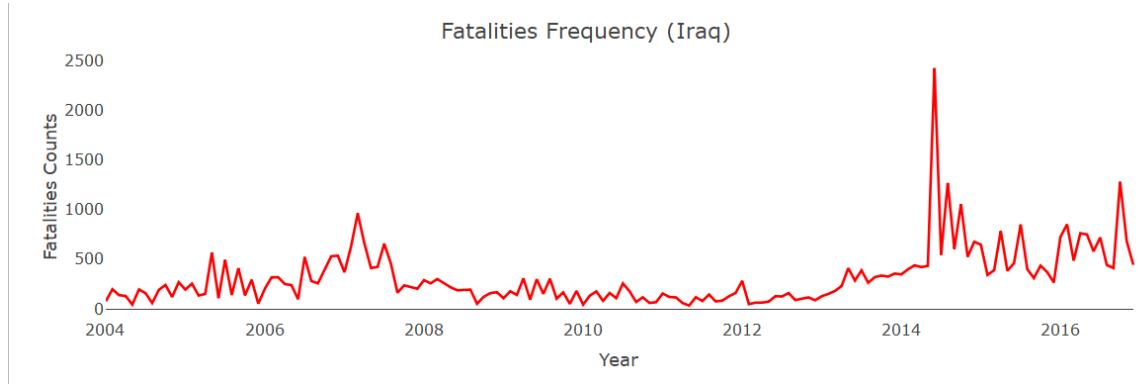


Figure 5.15: Fatalities Frequency by Year- Iraq

An unusual spike indicating 2426 deaths in June 2014 refers to the major incidents from ISIL where 1500 were reportedly killed in a single incident followed by another single incident involving 600 deaths.

```
if( knitr:::is_latex_output() ) {
  include_graphics(path = "figure/iraq_seasonal.png")
} else {
  ts_seasonal(dft, type = "all", Ygrid = T, Xgrid = T, title = "Seasonality Plot")
}
```

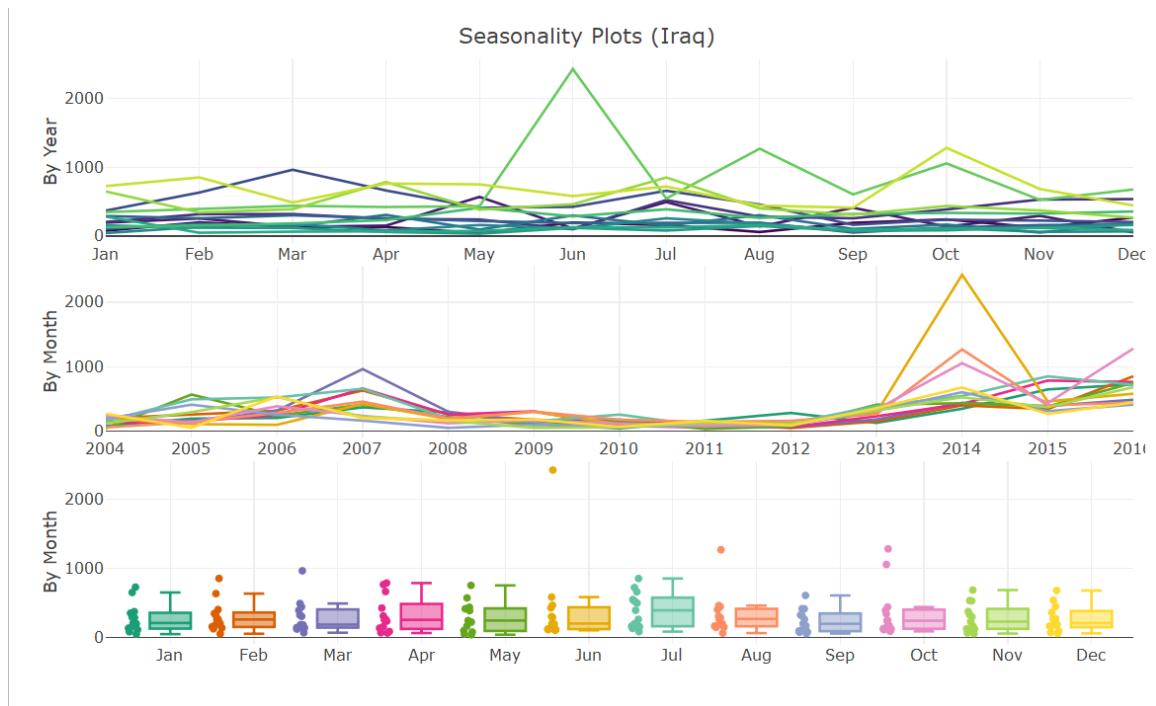


Figure 5.16: Seasonality Plots - Iraq

Seasonal components within year indicates high number of fatalities in July followed

by April and May.

5.2.3 Correlation test

```
if( knitr:::is_latex_output() ) {
  include_graphics(path = "figure/iraq_lags.png")
} else {
  ts_lags(dft, lag.max = 9)
}
```

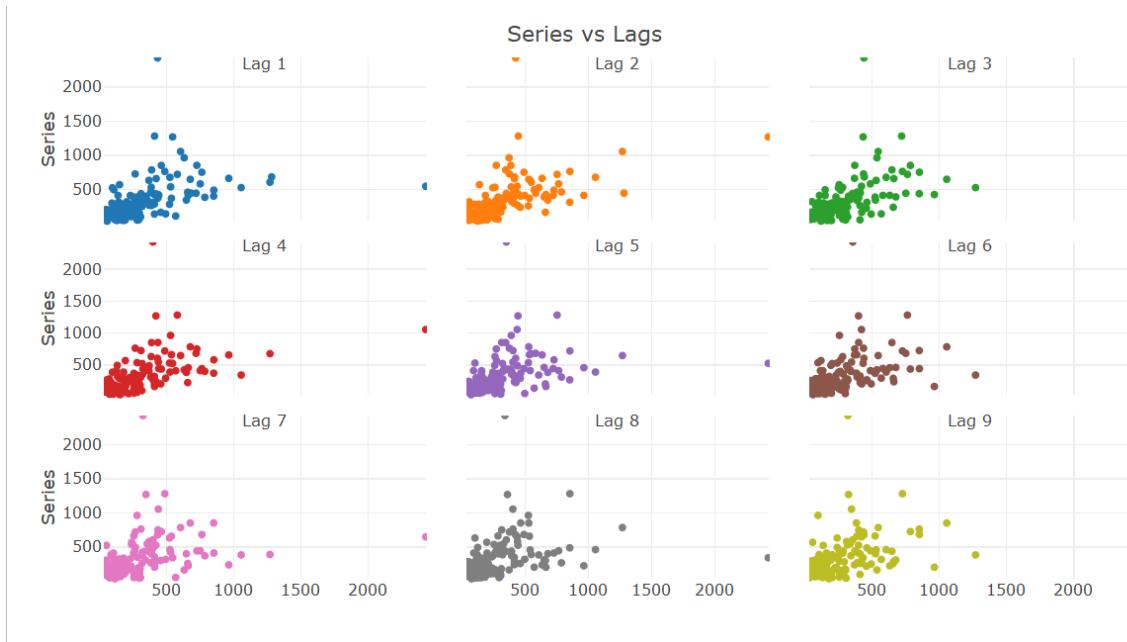


Figure 5.17: Correlation Test

In statistical terms, correlation means the extent of a linear relationship between two variables. Same way, autocorrelation means the linear relationship between lagged values of a time series as shown in the plot above for nine lags. We can spot few outliers in all nine lags as well as positive linear trend indicates the presence of autocorrelation.

5.2.4 Modelling

```
set.seed(84)
# horizon (look ahead period)
horizon <- 18
```

```

# Create split for train and test set
data <- ts_split(dft, sample.out = horizon)
# Split the data into training and testing sets
train <- data$train
test <- data$test

# Run models
fit_arima <- auto.arima(train)
fit_nn <- nnetar(train, repeats = 5)
fit_tbats <- tbats(train)
fit_ets <- ets(train, lambda = BoxCox.lambda(train))

#Get validation forecasts
fc_arima <- forecast(fit_arima, h = horizon)
fc_nn <- forecast(fit_nn, h = horizon)
fc_tbats <- forecast(fit_tbats, h = horizon)
fc_ets <- forecast(fit_ets, h = horizon)

metrics <- rbind(as.data.frame(round(accuracy(fc_arima$mean, test), 3)),
                  as.data.frame(round(accuracy(fc_nn$mean, test), 3)),
                  as.data.frame(round(accuracy(fc_tbats$mean, test), 3)),
                  as.data.frame(round(accuracy(fc_ets$mean, test), 3))) %>%
  add_column(models = c("Auto Arima", "NeuralNet", "TBATS", "ETS"),
             .before = "ME") %>% arrange(MAPE)

if( knitr:::is_latex_output() ) {
  knitr::kable(metrics, booktabs = TRUE,
               caption = "Performance comparison of all models (Iraq)") %>%
    kable_styling(full_width = F, latex_options = "hold_position",
                  font_size = 12)
} else {
  knitr::kable(metrics,
               caption = "Performance comparison of all models (Iraq)") %>%
    kable_styling(bootstrap_options = c("striped", "hover"), full_width = F,
                  font_size = 13) %>%
    column_spec(1, bold = T, color = "black") %>%
    column_spec(2:8, color = "black", background = "#dee2ed") %>%
    column_spec(6, color = "black", background = "#c7cfe5") %>%
    row_spec(1, background = "#c7cfe5")
}

```

From the model comparison based on MAPE metric, we can see that Auto Arima model performs better on this data. The corresponding Theil's U score is ~ 0.8 for all the models which means forecasts from chosen model are better than random guessing.

Table 5.4: Performance comparison of all models (Iraq)

models	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Auto Arima	116.42	271.3	201.8	5.843	31.34	0.054	0.822
ETS	103.37	266.5	201.7	3.261	32.16	0.052	0.810
TBATS	102.52	266.1	201.7	3.096	32.21	0.052	0.809
NeuralNet	-10.82	255.1	195.8	-18.967	38.47	-0.018	0.817

Next we calculate Theil's U score on ensembled predictions to see how much improvement can be achieved compared to best single model.

5.2.5 Ensemble

```
# extract predictions from all four models
preds <- as.data.frame(cbind(fc_arima$mean, fc_nn$mean, fc_tbats$mean, fc_ets$mean))
preds$ensemble <- rowMeans(preds)

# Compute Theil's U statistic (a = actual values, p= predicted values)
cat(paste("Theil's U score on Ensemble: ",
          round(TheilU(a = test, p = preds$ensemble), 3)))
```

Theil's U score on Ensemble: 0.399

As expected, we can see the significant improvement in forecasting accuracy by averaging predictions from all four models. Just to re-iterate, Theil's U score less than 1 means predictions are better than random guess (naive model).

5.2.6 Forecast future fatalities

In the validation part, data was into train and test in order to evaluate performance of different models. For the forecast, we run the models all the data points i.e. from 2000-2016.

```
# look ahead period
f_horizon <- 12
# run model on full data i.e dft (2000-2016)
fore_arima <- forecast(auto.arima(dft), h = f_horizon, level = c(80, 95))
fore_nn <- forecast(nnetar(dft, repeats = 5), h = f_horizon, level = c(80, 95))
fore_tbats <- forecast(tbats(dft), h = f_horizon, level = c(80, 95))
fore_ets <- forecast(ets(dft, lambda = BoxCox.lambda(dft)),
                      h = f_horizon, level = c(80, 95))

# combine plots for latex output
if( knitr:::is_latex_output() ) {
```

```
plot(fore_arima)
plot(fore_nn)
plot(fore_tbats)
plot(fore_ets)
}
```

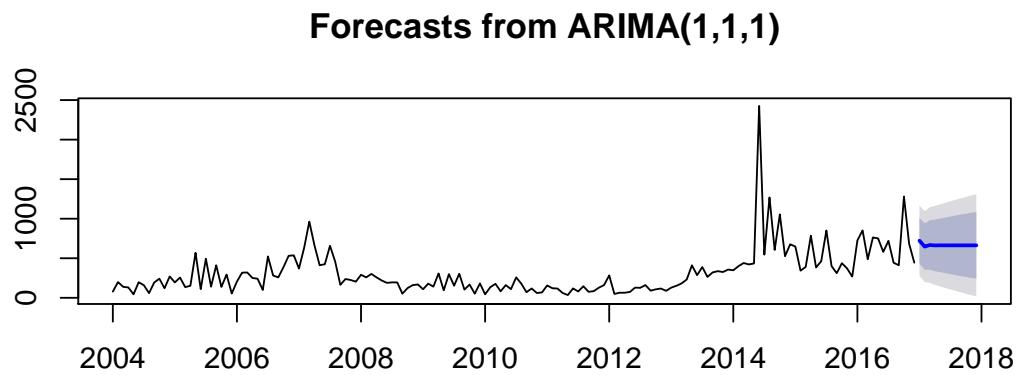


Figure 5.18: Predicted Number of Fatalities in Iraq (plots)

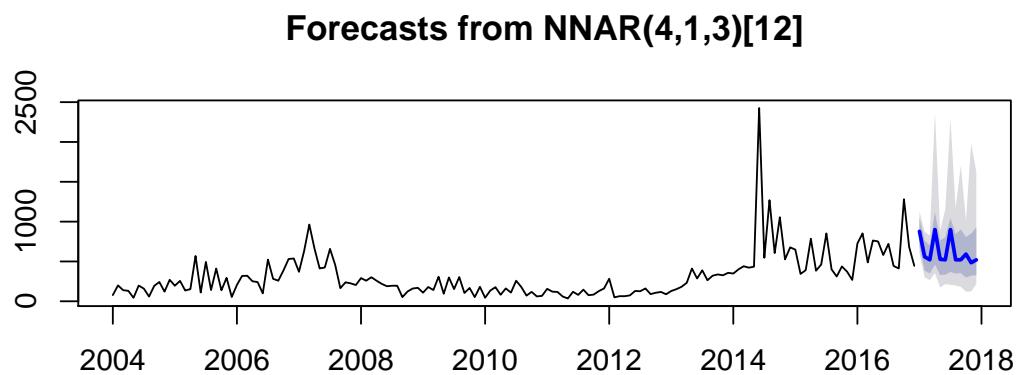


Figure 5.19: Predicted Number of Fatalities in Iraq (plots)

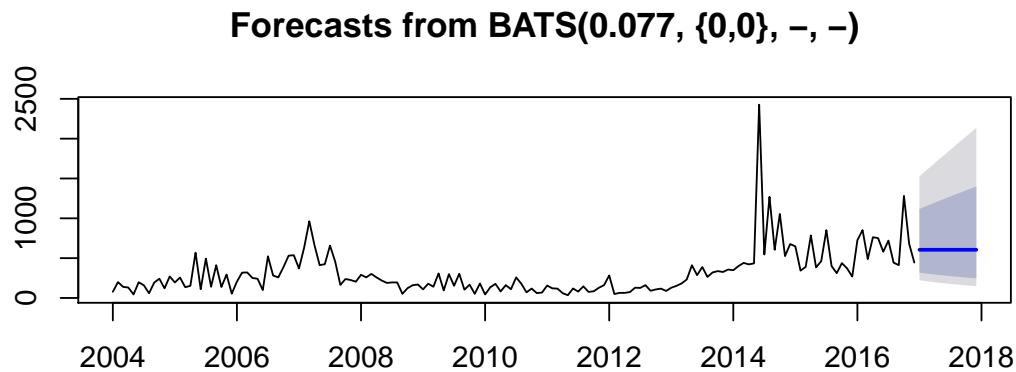


Figure 5.20: Predicted Number of Fatalities in Iraq (plots)

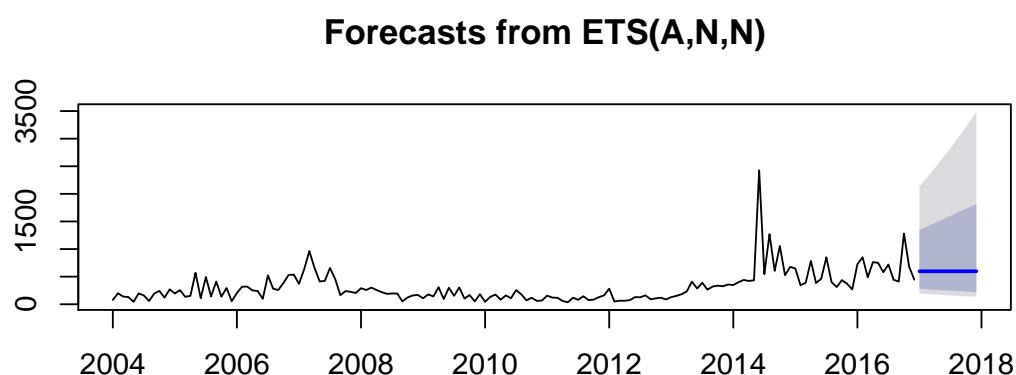


Figure 5.21: Predicted Number of Fatalities in Iraq (plots)

```

tbl_arima   <- timetk::tk_tbl(round(fore_arima$mean))
tbl_nn      <- timetk::tk_tbl(round(fore_nn$mean))
tbl_tbats   <- timetk::tk_tbl(round(fore_tbats$mean))
tbl_ets     <- timetk::tk_tbl(round(fore_ets$mean))

tbl <- tbl_arima %>%
  left_join(tbl_nn, by = "index") %>%
  left_join(tbl_tbats, by = "index") %>%
  left_join(tbl_ets, by = "index")

names(tbl) <- c("Time_period", "Arima", "NN", "TBATS", "ETS")

```

```

tbl$Ensemble <- round(rowMeans(tbl[,2:5]))

if( knitr:::is_latex_output() ) {
  knitr::kable(tbl, booktabs = TRUE,
    caption = "Table of Predicted Future Fatalities in Iraq") %>%
    kable_styling(full_width = F, latex_options = "hold_position",
      font_size = 12)
} else {
  knitr::kable(tbl,
    caption = "Table of Predicted Future Fatalities in Iraq") %>%
    kable_styling(bootstrap_options = c("striped", "hover"), full_width = F,
      font_size = 13, position = "left") %>%
    column_spec(1:5, color = "black", background = "#dee2ed") %>%
    column_spec(1, bold = T) %>%
    column_spec(6, bold = T, color = "black", background = "#c7cfef")
}

```

Table 5.5: Table of Predicted Future Fatalities in Iraq

Time_period	Arima	NN	TBATS	ETS	Ensemble
Jan 2017	722	878	605	597	700
Feb 2017	647	559	605	597	602
Mar 2017	668	521	605	597	598
Apr 2017	662	900	605	597	691
May 2017	664	527	605	597	598
Jun 2017	663	521	605	597	596
Jul 2017	663	900	605	597	691
Aug 2017	663	521	605	597	596
Sep 2017	663	521	605	597	596
Oct 2017	663	592	605	597	614
Nov 2017	663	484	605	597	587
Dec 2017	663	521	605	597	596

We can see flat forecast in ETS and TBATS model on this data which means that the trend and seasonality is insufficient to allow the future observations to have different conditional means. In that case, both models return the last observed value. However, as we have computed the Theil's U score for ensemble on test set which is less than 1, the ensembled forecasts as shown above can still be considered as approximate estimate better than random walk model.

5.3 SAHEL Region (Predict future attacks)

The Sahel region in Africa stretches from east to west across African continent. At present, this region draws huge political attention due to the indications of possible geographical expansion of ISIL (Liautaud, 2018). To estimate the future number of attacks in this region, I have selected data from year 2000 and filtered by eight countries that falls within sahel region as shown in the data preparation step.

5.3.1 Data preparation

```
sahel_region <- c("Mauritania", "Mali", "Burkina Faso",
                  "Niger", "Nigeria", "Chad", "Sudan", "Eritrea")

dft <- df %>%
  filter(year >= 2000 & country %in% sahel_region) %>%
  group_by(year, month) %>%
  summarise(total_count = n()) %>%
  ungroup() %>%
  group_by(year) %>%
  tidyr::complete(month = full_seq(seq(1:12), 1L), fill = list(total_count = 0))
ungroup()

dft <- dft %>%
  mutate(month_year = paste(year, month, sep="-"),
         month_year = zoo::as.yearmon(month_year)) %>%
  select(month_year, total_count)

# Create a ts object
dft <- ts(dft[, 2], start = Year(min(dft$month_year)),
           frequency = 12) # 1=annual, 4=quarterly, 12=monthly
dft <- na.kalman(dft)
```

5.3.2 Seasonality analysis

```
if( knitr:::is_latex_output() ) {
  include_graphics(path = "figure/sahel_line_plot.png")
} else {
  ts_plot(dft, line.mode = "line", Xtitle = "Year", Ytitle = "Attack Counts",
          title = "Attack Frequency (SAHEL Region)", color = "red")
}
```

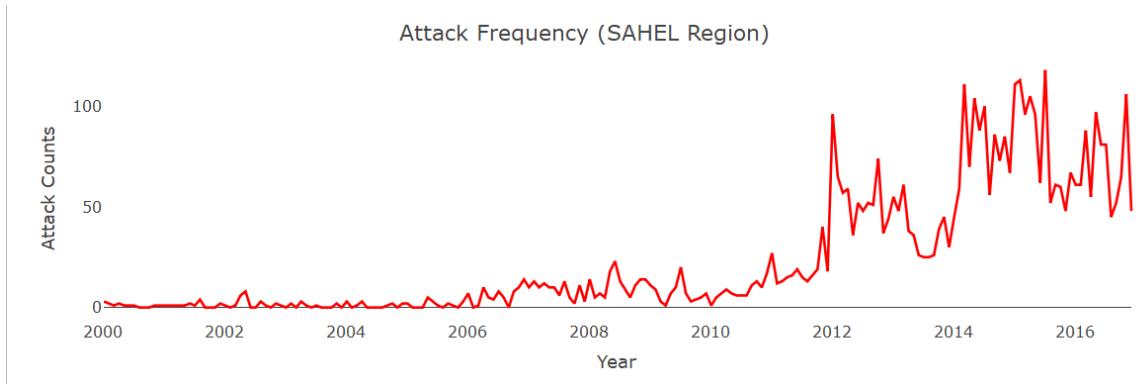


Figure 5.22: Attack Frequency by Year- SAHEL Region

From the attack frequency by year, it is observed that number of attacks have increased exponentially in the last decade and reaching peak during year 2014-2015. Several researchers (Crone, 2017; Onuoha & Oyewole, 2018) have indicated that Boko Haram affiliated itself with Islamic State in 2015 as well as large number of small groups from entire region have also declared their affiliation with Islamic State.

```
if( knitr:::is_latex_output() ) {
  include_graphics(path = "figure/sahel_heatmap.png")
} else {
  ts_heatmap(dft)
}
```

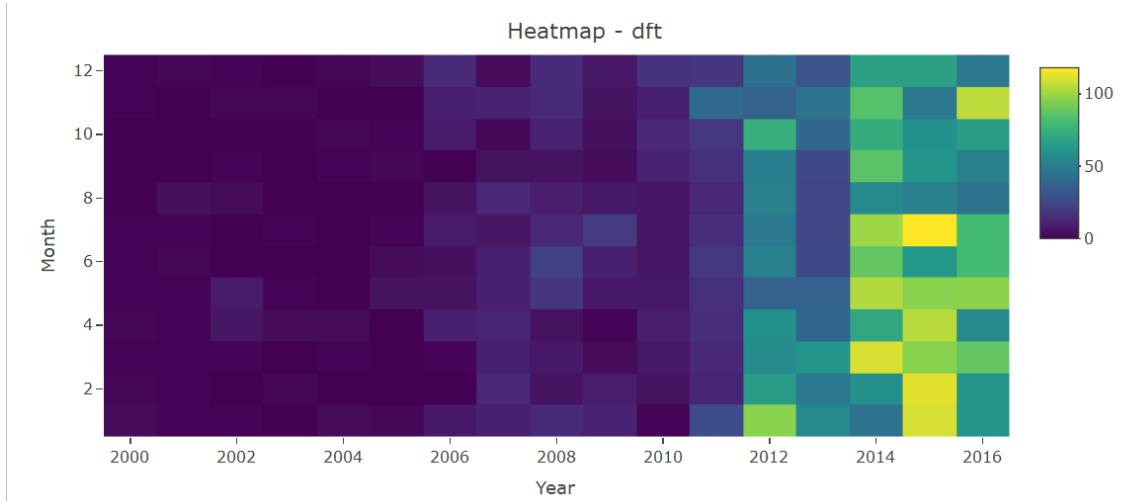


Figure 5.23: Seasonal Pattern (heatmap) - SAHEL Region

From the heatmap above, we can see sudden increase in number of attacks from year 2012 and more than 50 attacks a month on average. Let's have a look at seasonal components to see if there is any pattern by cycles.

```

if( knitr:::is_latex_output() ) {
  include_graphics(path = "figure/sahel_boxplot.png")
} else {
  ts_seasonal(dft, type = "box", Ygrid = T, Xgrid = T,
              title = "Seasonality Plot (SAHEL Region)")
}

```

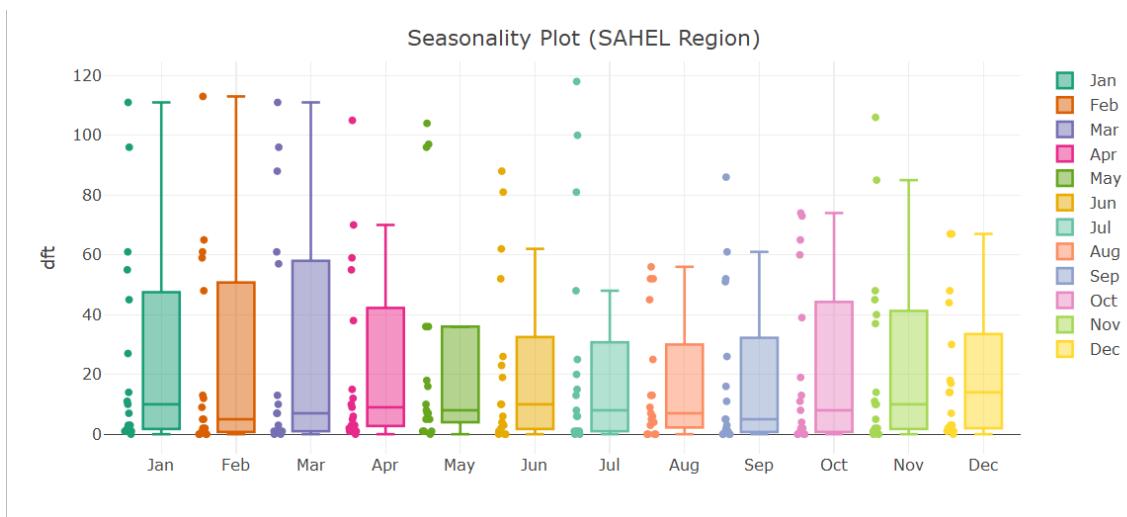


Figure 5.24: Seasonality Pattern (boxplot) - SAHEL Region

In a comparison to number of attacks in Afghanistan and number of fatalities in Iraq, we can see opposite trend in SAHEL region where months in the beginning and end of the year (Jan to Mar and Oct to Dec) indicates higher number of attacks through the period (2000-2016). In case of Afghanistan and Iraq, it was mostly observed in the months middle of year.

5.3.3 Correlation test

```

if( knitr:::is_latex_output() ) {
  include_graphics(path = "figure/sahel_lags.png")
} else {
  ts_lags(dft, lag.max = 9)
}

```

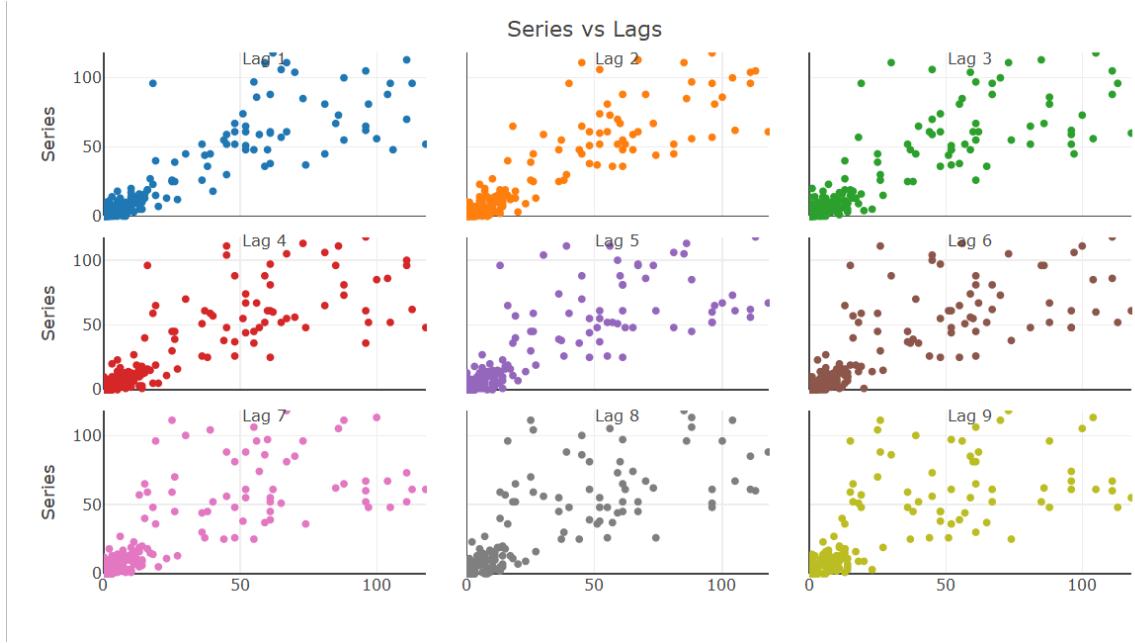


Figure 5.25: Correlation Test

Similar to correlation tests in Iraq and Afghanistan, a positive linear trend is visible in all nine lags while lag 1 and 2 suggesting strong autocorrelation.

5.3.4 Modelling

```

set.seed(84)
# horizon (look ahead period)
horizon <- 18

# create split for train and test set
data <- ts_split(dft, sample.out = horizon)
# Split the data into training and testing sets
train <- data$train
test <- data$test

# Run models
fit_arima <- auto.arima(train)
fit_nn <- nnetar(train, repeats = 5)
fit_tbats <- tbats(train)
fit_ets <- ets(train)

#Get validation forecasts
fc_arima <- forecast(fit_arima, h = horizon)
fc_nn <- forecast(fit_nn, h = horizon)

```

```

fc_tbats <- forecast(fit_tbats, h = horizon)
fc_ets <- forecast(fit_ets, h = horizon)

metrics <- rbind(as.data.frame(round(accuracy(fc_arima$mean, test), 3)),
                  as.data.frame(round(accuracy(fc_nn$mean, test), 3)),
                  as.data.frame(round(accuracy(fc_tbats$mean, test), 3)),
                  as.data.frame(round(accuracy(fc_ets$mean, test), 3))) %>%
  add_column(models = c("Auto Arima", "NeuralNet", "TBATS", "ETS"),
              .before = "ME") %>% arrange(MAPE)

if( knitr:::is_latex_output() ) {
  knitr::kable(metrics, booktabs = TRUE,
               caption = "Performance comparison of all models (SAHEL Regioin)") %>%
    kable_styling(full_width = F, latex_options = "hold_position",
                  font_size = 12)
} else {
  knitr::kable(metrics,
               caption = "Performance comparison of all models (SAHEL Regioin)") %>%
    kable_styling(bootstrap_options = c("striped", "hover"), full_width = F,
                  font_size = 13) %>%
    column_spec(1, bold = T, color = "black") %>%
    column_spec(2:8, color = "black", background = "#dee2ed") %>%
    column_spec(6, color = "black", background = "#c7cfe5") %>%
    row_spec(1, background = "#c7cfe5")
}

```

Table 5.6: Performance comparison of all models (SAHEL Regioin)

models	ME	RMSE	MAE	MPE	MAPE	ACF1	Theil's U
Auto Arima	-2.010	19.89	17.61	-10.75	26.46	-0.143	0.733
NeuralNet	-4.881	22.28	19.73	-16.01	30.62	-0.117	0.811
TBATS	-9.321	22.66	19.82	-22.25	32.11	-0.198	0.809
ETS	-11.439	23.67	21.11	-25.90	34.97	-0.178	0.901

From the model comparison based on MAPE metric, we can see that Auto Arima followed by Neural Network performs better on this data and all four models having Theil's U score below 1.

5.3.5 Ensemble

```

# extract predictions from all four models and get average
ensemble <- rowMeans(cbind(fc_arima$mean, fc_nn$mean, fc_tbats$mean, fc_ets$mean))
# Compute Theil's U statistic (a = actual values, p= predicted values)
TheilU(a = test, p = ensemble)

```

```
[1] 0.2995
```

An ensemble prediction further improves the accuracy as measured by Theil's U score.

5.3.6 Forecast future attacks

```
# look ahead period
f_horizon <- 18
# run model on full data i.e dft (2000-2016)
fore_arima <- forecast(auto.arima(dft), h = f_horizon, level = c(80, 95))
fore_nn <- forecast(nnetar(dft, repeats = 5), h = f_horizon, level = c(80, 95),
fore_tbats <- forecast(tbats(dft), h = f_horizon, level = c(80, 95))
fore_ets <- forecast(ets(dft), h = f_horizon, level = c(80, 95))

# combine plots for latex output
if( knitr:::is_latex_output() ) {
  plot(fore_arima)
  plot(fore_nn)
  plot(fore_tbats)
  plot(fore_ets)
}
```

Forecasts from ARIMA(2,1,1)(0,0,2)[12] with drift

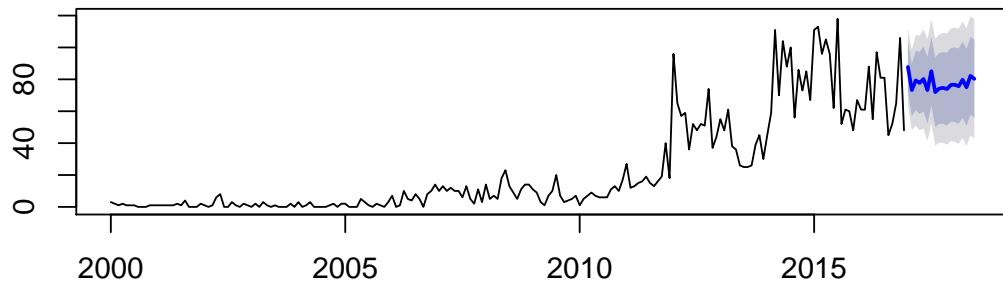


Figure 5.26: Predicted Number of Attacks in SAHEL Region (plots)

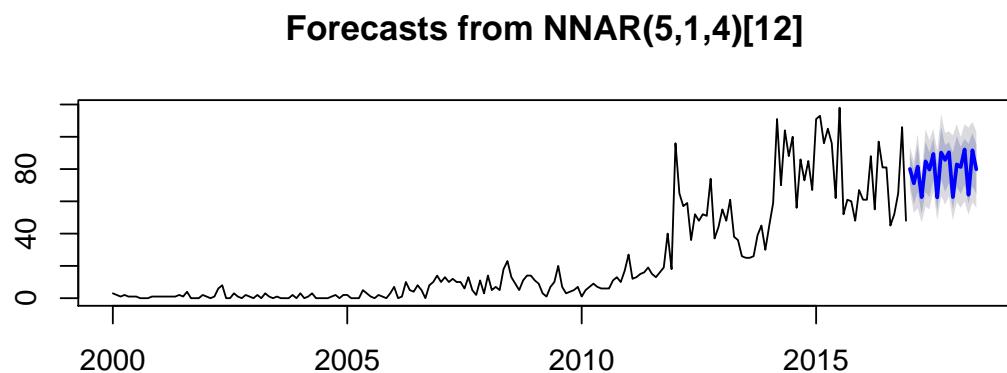


Figure 5.27: Predicted Number of Attacks in SAHEL Region (plots)

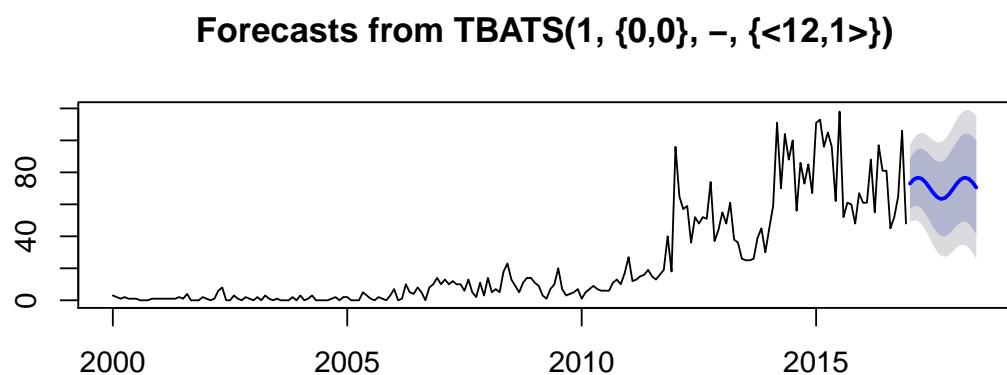


Figure 5.28: Predicted Number of Attacks in SAHEL Region (plots)

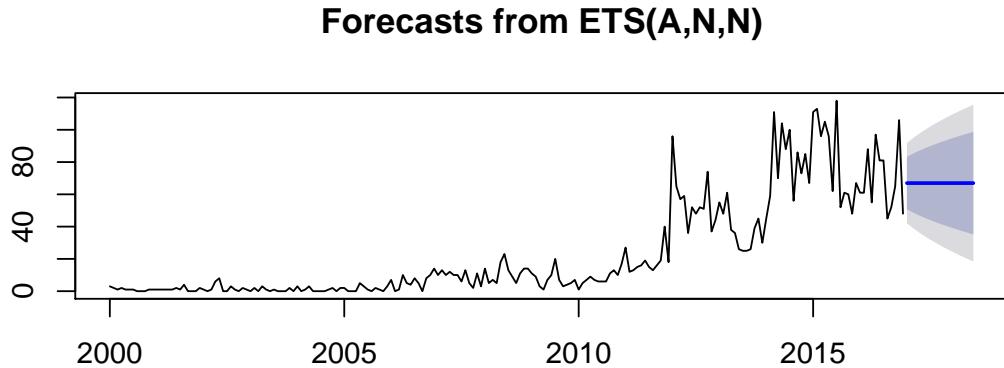


Figure 5.29: Predicted Number of Attacks in SAHEL Region (plots)

From the plots above, we can see that Auto Arima, Neural Network and TBATS model are able to capture observed trend whereas ETS model fails and generates flat predictions. However, ensemble approach will compensate the loss from any weak models as computed before.

```

tbl_arima    <- timetk::tk_tbl(round(fore_arima$mean))
tbl_nn       <- timetk::tk_tbl(round(fore_nn$mean))
tbl_tbats   <- timetk::tk_tbl(round(fore_tbats$mean))
tbl_ets      <- timetk::tk_tbl(round(fore_ets$mean))

tbl <- tbl_arima %>%
  left_join(tbl_nn, by = "index") %>%
  left_join(tbl_tbats, by = "index") %>%
  left_join(tbl_ets, by = "index")

names(tbl) <- c("Time_period", "Arima", "NN", "TBATS", "ETS")
tbl$Ensemble <- round(rowMeans(tbl[,2:5]))

if( knitr:::is_latex_output() ) {
  knitr::kable(tbl, booktabs = TRUE,
  caption = "Table of Predicted Future Attacks in SAHEL Region") %>%
  kable_styling(full_width = F, latex_options = "hold_position",
  font_size = 12)
} else {
  knitr::kable(tbl,
  caption = "Table of Predicted Future Attacks in SAHEL Region") %>%
  kable_styling(bootstrap_options = c("striped", "hover"), full_width = F,
  font_size = 13, position = "left") %>%
  column_spec(1:5, color = "black", background = "#dee2ed") %>
}

```

```

  column_spec(1, bold = T) %>%
  column_spec(6, bold = T, color = "black", background = "#c7cfec")
}

```

Table 5.7: Table of Predicted Future Attacks in SAHEL Region

Time_period	Arima	NN	TBATS	ETS	Ensemble
Jan 2017	88	80	73	67	77
Feb 2017	73	71	76	67	72
Mar 2017	79	82	77	67	76
Apr 2017	78	63	76	67	71
May 2017	80	85	74	67	76
Jun 2017	73	80	71	67	73
Jul 2017	85	89	67	67	77
Aug 2017	72	62	64	67	66
Sep 2017	74	90	63	67	74
Oct 2017	75	86	64	67	73
Nov 2017	74	90	66	67	74
Dec 2017	77	63	70	67	69
Jan 2018	77	83	73	67	75
Feb 2018	76	81	76	67	75
Mar 2018	80	92	77	67	79
Apr 2018	75	64	76	67	70
May 2018	82	92	74	67	79
Jun 2018	80	80	71	67	74

Chapter 6

Classification Approach

6.1 Overview of target variables

Conclusion

If we don't want Conclusion to have a chapter number next to it, we can add the `{-}` attribute.

More info

And here's some other random info: the first paragraph after a chapter title or section head *shouldn't be* indented, because indents are to tell the reader that you're starting a new paragraph. Since that's obvious after a chapter or section title, proper typesetting doesn't add an indent there.

Appendix A

The First Appendix

This first appendix includes all of the R chunks of code that were hidden throughout the document (using the `include = FALSE` chunk tag) to help with readability and/or setup.

In the main Rmd file

```
# This chunk ensures that the thesisdown package is
# installed and loaded. This thesisdown package includes
# the template files for the thesis.
if(!require(devtools))
  install.packages("devtools", repos = "http://cran.rstudio.com")
if(!require(thesisdown))
  devtools::install_github("ismayc/thesisdown")
library(thesisdown)

#load packages
if (!require("pacman")) install.packages("pacman")
pacman::p_load(data.table, DT, openxlsx, RCurl, stringr, stringi, reshape, knitr,
  DescTools, GGally, StandardizeText, scales, lubridate, countrycode,
  viridis, viridisLite, RColorBrewer, ggfortify, plotly, highcharter,
  arules, arulesViz, visNetwork, igraph,
  TSstudio, timetk, tidyquant, tidyr, zoo, forecast, tseries, impute,
  countrycode, WDI, purrr, igraph, visNetwork, randomcoloR, treemap,
  shiny, ggmap, maptools, maps, eply,
  # shinydashboard, shinythemes, shinyjs, shinyBS, shinyWidgets, s
  parallel, caret, pROC, lightgbm,
  bookdown, servr, ggthemes, tidyverse)

options(warn = -1, digits = 4, scipen = 999)
set.seed(84)

# load clean and prepared data (GTD)
```

```
setwd("C:/Users/Pranav_Pandya/Desktop/Thesis/gtd_eda/index")

# load clean data (GTD)
df <- readRDS("data/gtd_clean_v2.rds")

theme_set(theme_fivethirtyeight(base_size = 12))
```

In Chapter ??:

Appendix B

The Second Appendix, for Fun

References

- Al Jazeera. (2014). Sunni rebels declare new 'Islamic caliphate'. Retrieved from <https://www.aljazeera.com/news/middleeast/2014/06/isil-declares-new-islamic-caliphate-201462917326669749.html>
- Andri Signorell et mult. al. (2018). DescTools: Tools for Descriptive Statistics. Retrieved from <https://cran.r-project.org/package=DescTools>
- Anomaly.io. (2015, December). Extracting Seasonality and Trend from Data: Decomposition Using R. *Anomaly*. Retrieved from <https://anomaly.io/seasonal-trend-decomposition-in-r/>
- Bauer, P. (2018). *Writing a Reproducible Paper in R Markdown* (SSRN Scholarly Paper No. ID 3175518). Rochester, NY: Social Science Research Network. Retrieved from <https://papers.ssrn.com/abstract=3175518>
- Beck, N., King, G., & Zeng, L. (2000). Improving Quantitative Studies of International Conflict: A Conjecture. *American Political Science Review*, 94(1), 21–35. <http://doi.org/10.1017/S0003055400220078>
- Block, M. (2016). Applying situational crime prevention to terrorism against airports and aircrafts. *Electronic Theses and Dissertations*. <http://doi.org/10.18297/etd/2479>
- Brennan, P. (2016). *The detection of outbreaks in terrorist incidents using time series anomaly detection methods* (PhD thesis). Institute of Technology, Tallaght. Retrieved from https://github.com/brennap3/thesis_2/blob/master/thesis.pdf
- Cederman, L.-E., & Weidmann, N. B. (2017). Predicting armed conflict: Time to adjust our expectations? *Science*, 355(6324), 474–476. <http://doi.org/10.1126/science.aal4483>
- Ceron, A., Curini, L., & Iacus, S. M. (2018). ISIS at its apogee: The Arabic discourse on Twitter and what we can learn from that about ISIS support and Foreign Fighters. *arXiv:1804.04059 [Cs]*. Retrieved from <http://arxiv.org/abs/1804.04059>
- Chadefaux, T. (2014). Early warning signals for war in the news. *Journal of Peace Research*, 51(5), 611–624. <http://doi.org/10.1177/0022343314529302>

- Research*, 51(1), 5–18. <http://doi.org/10.1177/0022343313507302>
- CIA. (2013). INTelligence: Human Intelligence. Retrieved from <https://www.cia.gov/news-information/featured-story-archive/2010-featured-story-archive/intelligence-human-intelligence.html>
- Clauset, A., & Woodard, R. (2013). Estimating the historical and future probabilities of large terrorist events. *The Annals of Applied Statistics*, 7(4), 1838–1865. <http://doi.org/10.1214/12-AOAS614>
- Colaresi, M., & Mahmood, Z. (2017). Do the robot , Do the robot: Lessons from machine learning to improve conflict forecasting , Lessons from machine learning to improve conflict forecasting. *Journal of Peace Research*, 54(2), 193–214. <http://doi.org/10.1177/0022343316682065>
- Crone, M. (2017). Islamic State's Incursion into North Africa and Sahel: A Threat to al-Qaeda? *Connections*, 16(1), 63–76. Retrieved from <http://www.jstor.org/stable/26326471>
- Ding, F., Ge, Q., Jiang, D., Fu, J., & Hao, M. (07AD–2017). Understanding the dynamics of terrorism events with multiple-discipline datasets and machine learning approach. *PLOS ONE*, 12(6), e0179057. <http://doi.org/10.1371/journal.pone.0179057>
- Fujita, K., Shinomoto, S., & Rocha, L. E. C. (2016). Correlations and forecast of death tolls in the Syrian conflict. *arXiv:1612.06746 [Physics, Stat]*. Retrieved from <http://arxiv.org/abs/1612.06746>
- Geddes, B. (1990/ed). How the Cases You Choose Affect the Answers You Get: Selection Bias in Comparative Politics. *Political Analysis*, 2, 131–150. <http://doi.org/10.1093/pan/2.1.131>
- Gordon, A. (2007). Transient and continuant authors in a research field: The case of terrorism. *Scientometrics*, 72(2), 213–224. <http://doi.org/10.1007/s11192-007-1714-z>
- Groce, A. (2018). LibGuides: Intelligence Studies: Types of Intelligence Collection. Retrieved from <http://usnwc.libguides.com/c.php?g=494120&p=3381426>
- Gundabathula, V. T., & Vaidhehi, V. (2018). An Efficient Modelling of Terrorist Groups in India using Machine Learning Algorithms. *Indian Journal of Science and Technology*, 11(15). <http://doi.org/10.17485/ijst/2018/v11i15/121766>
- Hahsler, M., Buchta, C., Gruen, B., Hornik, K., Johnson, I., & Borgelt, C. (2018, April). Arules: Mining Association Rules and Frequent Itemsets. Retrieved from <https://CRAN.R-project.org/package=arules>
- Heger, L. L. (2010). *In the crosshairs : Explaining violence against civilians* (PhD thesis). UC San Diego. Retrieved from <https://escholarship.org/uc/item/>

6705k88s

- Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice*. OTexts. Retrieved from <https://otexts.org/fpp2>
- Indiana University Libraries. (2007, July). Identifying Primary and Secondary Sources. *Indiana University Bloomington*. Retrieved from <https://libraries.indiana.edu/identifying-primary-and-secondary-sources>
- Jacob van Veen, H., Nguyen, L., Dat, T., & Segnini, A. (2015). Kaggle Ensembling Guide | MLWave. *Kaggle Ensembling Guide*. Retrieved from <https://mlwave.com/kaggle-ensembling-guide/>
- Jongman, A. J. (1988). *Political Terrorism: A New Guide To Actors, Authors, Concepts, Data Bases, Theories, And Literature*. Transaction Publishers.
- Karthiyayini, R., & Balasubramanian, D. R. (2016). Affinity Analysis and Association Rule Mining using Apriori Algorithm in Market Basket Analysis, 6.
- Klausen, J., Marks, C., & Zaman, T. (2016). Finding Online Extremists in Social Networks. *arXiv:1610.06242 [Physics, Stat]*. Retrieved from <http://arxiv.org/abs/1610.06242>
- Klimberg, R., & McCullough, B. D. (2017). *Fundamentals of Predictive Analytics with JMP, Second Edition*. SAS Institute.
- Liautaud, A. (2018). U.S. military presence in Africa grew again, but “we’re not at war,” top U.S. commander says. *VICE News*. Retrieved from https://news.vice.com/en_us/article/j5b3pb/us-military-presence-in-africa-grew-again-but-were-not-at-war-top-us-commander-says
- Livera, A. M. D., Hyndman, R. J., & Snyder, R. D. (2011). Forecasting Time Series With Complex Seasonal Patterns Using Exponential Smoothing. *Journal of the American Statistical Association*, 106(496), 1513–1527. <http://doi.org/10.1198/jasa.2011.tm09771>
- Lowenthal, M. M., & Clark, R. M. (2015). *The Five Disciplines of Intelligence Collection*. SAGE.
- Lula, K. (2014). *Terrorized into compliance: Why countries submit to financial counterterrorism* (PhD thesis). Rutgers University - Graduate School - Newark. Retrieved from <https://rucore.libraries.rutgers.edu/rutgers-lib/42328/>
- Lum, C., Kennedy, L. W., & Sherley, A. J. (2006). THE EFFECTIVENESS OF COUNTER-TERRORISM STRATEGIES A Campbell Systematic Review.
- Mo, H., Meng, X., Li, J., & Zhao, S. (2017). Terrorist event prediction based on revealing data. In *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)*((pp. 239–244). <http://doi.org/10.1109/ICBDA.2017.>

- 8078815
- Muchlinski, D., Siroky, D., He, J., & Kocher, M. (2016/ed). Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data. *Political Analysis*, 24(1), 87–103. <http://doi.org/10.1093/pan/mpv024>
- National Consortium for the Study of Terrorism and Responses to Terrorism (START). (2016). Global Terrorism Database [Data file]. University of Maryland. Retrieved from <https://www.start.umd.edu/gtd>
- Nawaz, M. A. (2017). *How terrorism ends: The impact of lethality of terrorist groups on their longevity* (PhD thesis). Retrieved from <http://krex.k-state.edu/dspace/handle/2097/35788>
- Neunhoeffer, M., & Sternberg, S. (2018). How Cross-Validation Can Go Wrong and What to Do About it. | Marcel Neunhoeffer. *Forthcoming, Political Analysis*. Retrieved from http://www.marcel-neunhoeffer.com/publication/pa_cross-validation/
- NIC. (2007). Nonstate Actors: Impact on International Relations and Implications for the United States. National Intelligence Council. Retrieved from https://www.dni.gov/files/documents/nonstate_actors_2007.pdf
- Onuoha, F. C., & Oyewole, S. (2018). Anatomy of Boko Haram: The Rise and Decline of a Violent Group in Nigeria. *Al Jazeera*, 10. Retrieved from http://studies.aljazeera.net/mritems/Documents/2018/4/23/4f179351e3244e1882a6033e0bf43d89_100.pdf
- Oracle. (n.d.). Oracle Enterprise Performance Management Workspace, Fusion Edition User's Guide. Retrieved from https://docs.oracle.com/cd/E40248_01/epm.1112/cb_statistical/frameset.htm?ch07s02s03s04.html
- Patel, P. (2009). Introduction to Quantitative Methods. Retrieved from http://hls.harvard.edu/content/uploads/2011/12/quantitative_methods.pdf
- Ranstorp, M. (2006). *Mapping Terrorism Research: State of the Art, Gaps and Future Direction*. Routledge.
- Samuel, A. L. (1959). Some studies in machine learning using the game of Checkers. *Ibm Journal of Research and Development*, 71–105.
- Schuurman, B. (2018). Research on Terrorism, 20072016: A Review of Data, Methods, and Authorship. *Terrorism and Political Violence*, 0(0), 1–16. <http://doi.org/10.1080/09546553.2018.1439023>
- Siddique, H. (2013). Edward Snowden's live Q&A: Eight things we learned. *The Guardian*. Retrieved from <http://www.theguardian.com/world/2013/jun/18/edward-snowden-live-q-and-a-eight-things>
- Silke, A. (2001). The Devil You Know: Continuing Problems with Research on Terrorism. *Terrorism and Political Violence*, 13(4), 1–14. <http://doi.org/10.1080/08862605.2001.9518410>

- 1080/09546550109609697
- Silke, A. (2004). *Research on Terrorism: Trends, Achievements and Failures*. Routledge.
- Stockholm International Peace Research Institute. (2017). SIPRI Yearbook 2017, Summary. Retrieved from <https://www.sipri.org/sites/default/files/2017-09/yb17-summary-eng.pdf>
- Tanner, A. (2014). Examining the Need for a Cyber Intelligence Discipline. *Journal of Homeland and National Security Perspectives*, 1(1), 38–48. Retrieved from <https://journals.tdl.org/jhnsp/index.php/jhnsp/article/view/16>
- The Interagency OPSEC Support Staff. (1996). Operations Security Intelligence Threat Handbook. *Federation Of American Scientists*. Retrieved from <https://fas.org/irp/nsa/ioss/threat96/part02.htm>
- Walton, O. (2011). Early warning indicators of violent conflict: Helpdesk report. Retrieved from <https://researchportal.bath.ac.uk/en/publications/early-warning-indicators-of-violent-conflict-helpdesk-report>
- Ward Lab. (2014, May). The coup in Thailand and progress in forecasting. *Predictive Heuristics*. Retrieved from <https://predictiveheuristics.com/2014/05/22/the-coup-in-thailand-and-progress-in-forecasting/>
- Xie, Y. (2016). *Bookdown: Authoring Books and Technical Documents with R Markdown*. CRC Press.