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

List of Tables

List of Figures

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 anaylysis, 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 reasearch 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 reasearch and is intended to provide interactive platform for all three parts of this reasearch 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 incresed significantly as shown in the plot below:

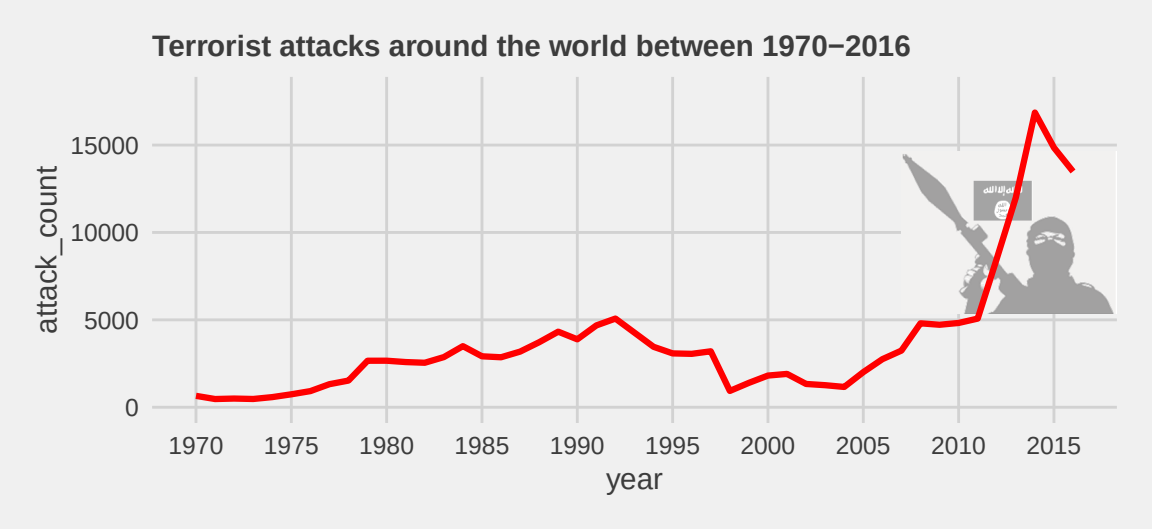

Figure 1: Terrorist attacks around the world between 1970-2016

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 targetted region 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 on 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 reporoducibility. 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 gardient 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 likelyhood of vilonet 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 machined 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 aseess theoritical framework for the reserach 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 form. The primary objective of this research is achieve actionable intelligence through machine learning apporach so it is important 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 CIA[1] 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 technique which uses covert as well as overt individuals to gather information. Example 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-IO-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 reprsentation 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 becomeing extremely important when concerns about WMDs (Weapons of Mass Destruction) are growing. This approach peforms 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, checmial 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 recods 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 sets 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 extracts 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 Database[4]

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

4. SIPRI Databases[5]

- 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 futher 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 statndards (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 failry 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 qulity 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 on 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

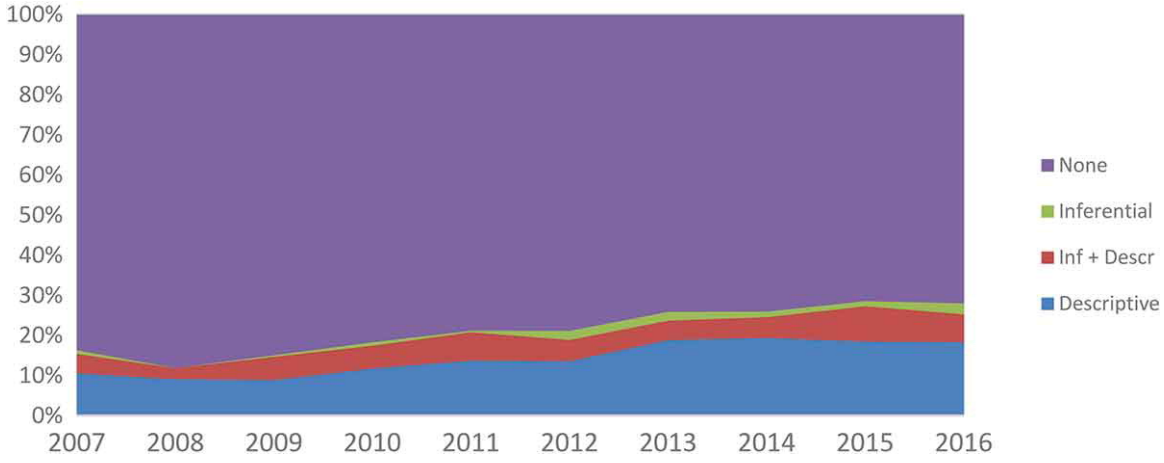

Figure 1.1: Use of statistics in terrorism research between 2007 to 2016

(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 confict. 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 (Klausen, 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 finanicial 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 terrrorist 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 involiving ariports/ 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 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 of base models and suggests that 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 `intl_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"

[1] Accessing 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 explaination about each variable can be found GTD codebook.

[2]Searching 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.

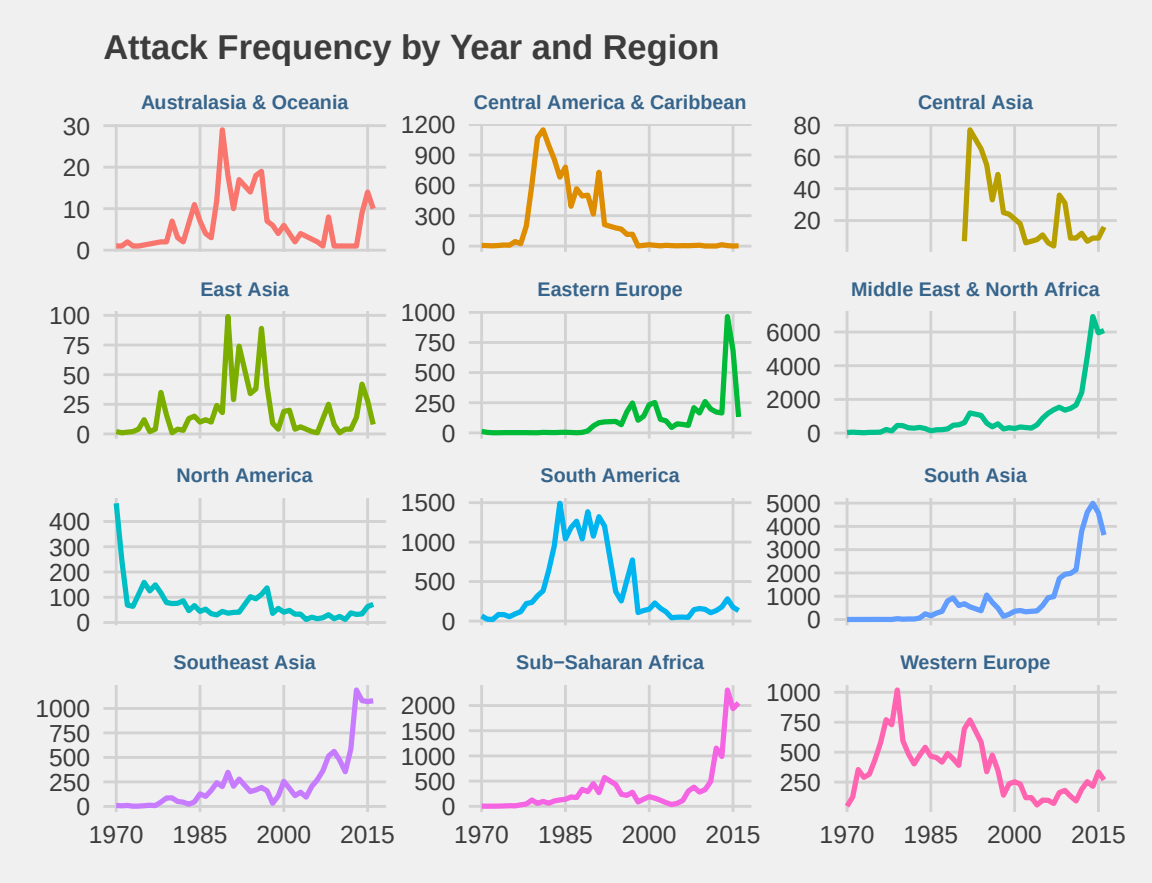

Figure 2.1: Attack Frequency by Year and Region

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.

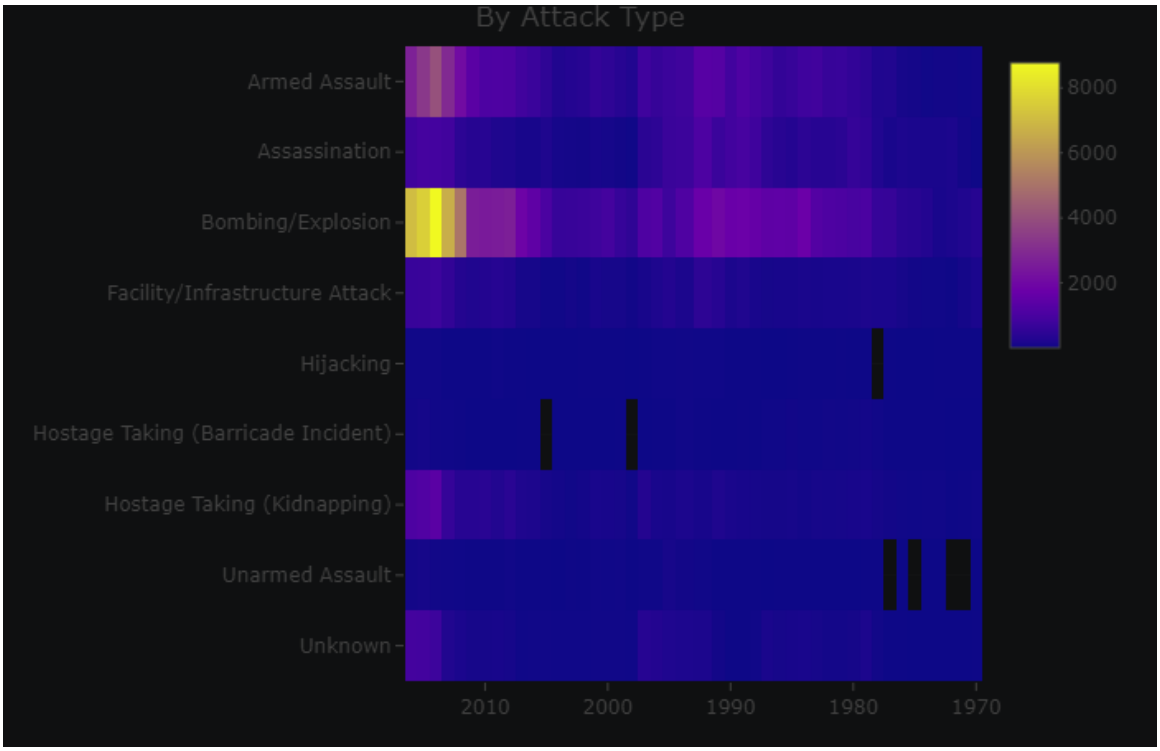

Figure 2.2: Trend in type of attack in all incidents globally

The heat signatures indicates Bombing/Explosive as one of the frequently used techniques by terrorist groups. Although the pattern in this tactic is visiable 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.

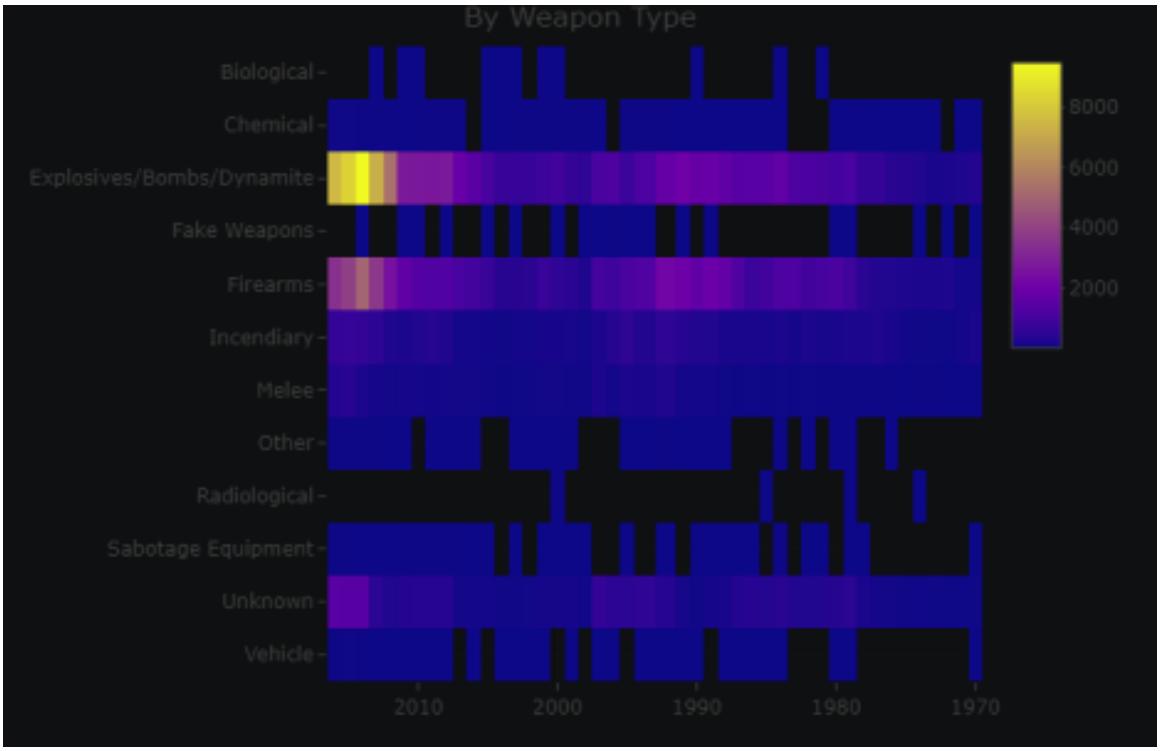

Figure 2.3: Trend in type of weapon used in all incidents globally

Upon examining the trends in type of weapon used in all terrorist incidents gloablly as shown in the figure below, 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 most frequently attacked target types as Private Citzens & Property followed by Military, Police, Government and Business.

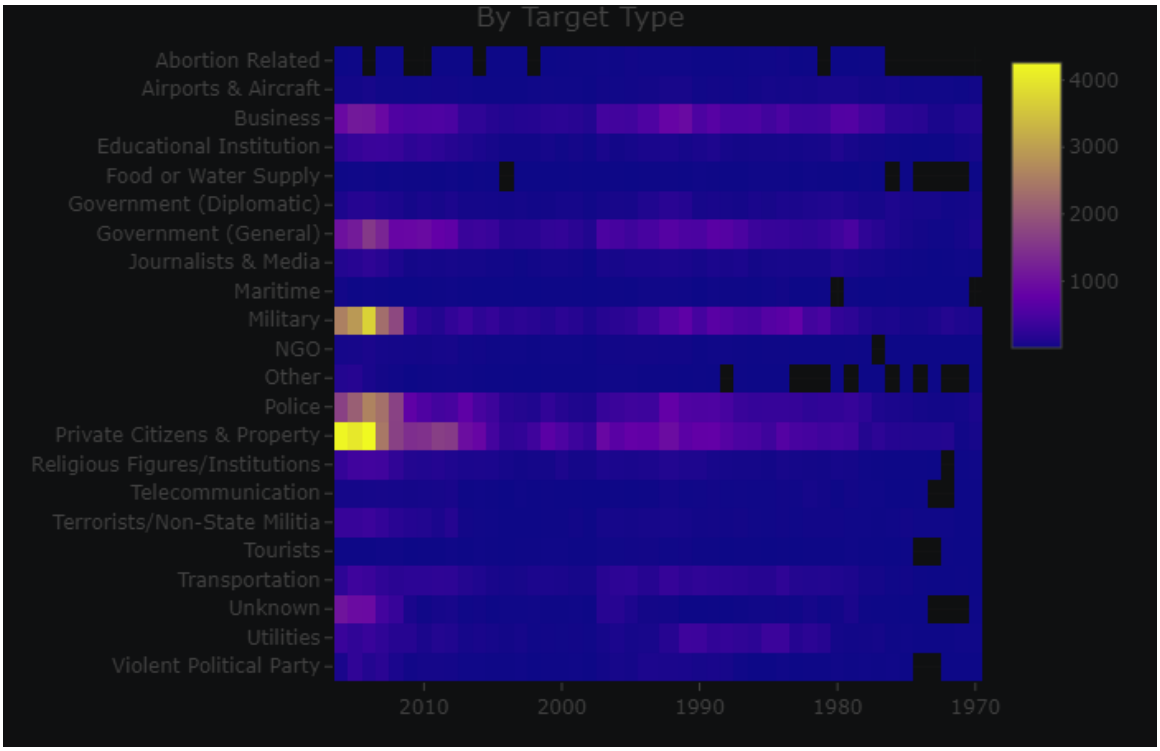

Figure 2.4: Trend in intended targets in all incidents globally

According to GTD codebook, Private Citzens & 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")
```

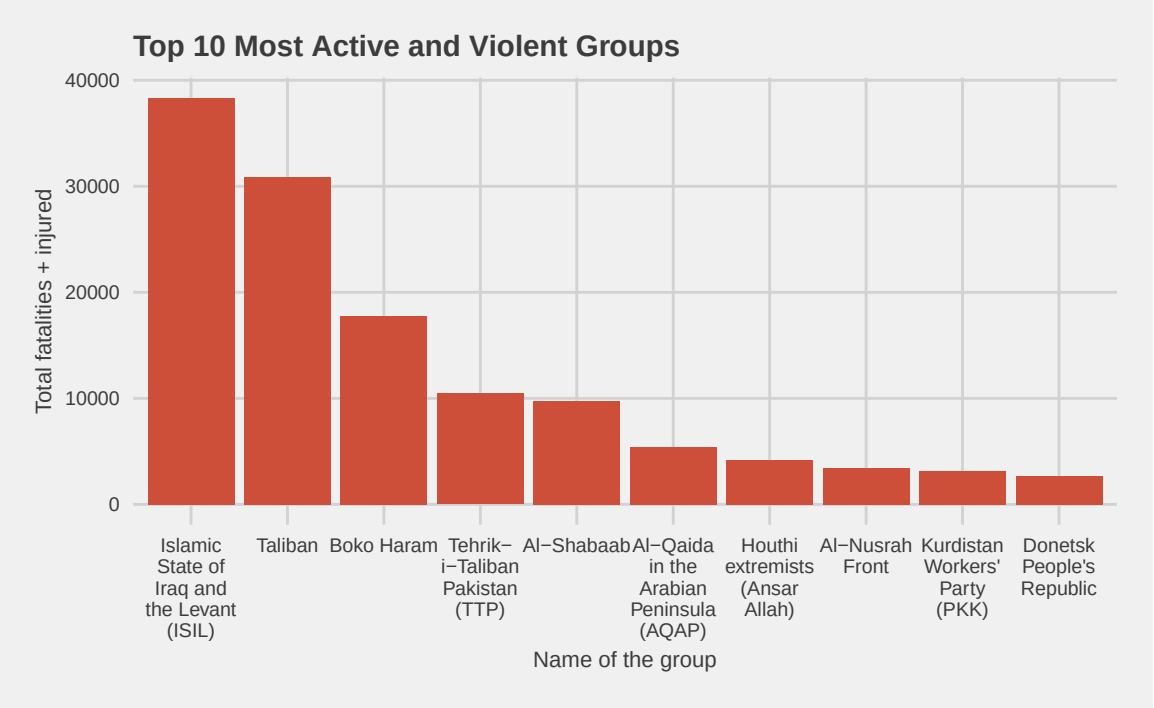

Figure 2.5: Top 10 Most Active and Violent Groups

Based on cummulative 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.

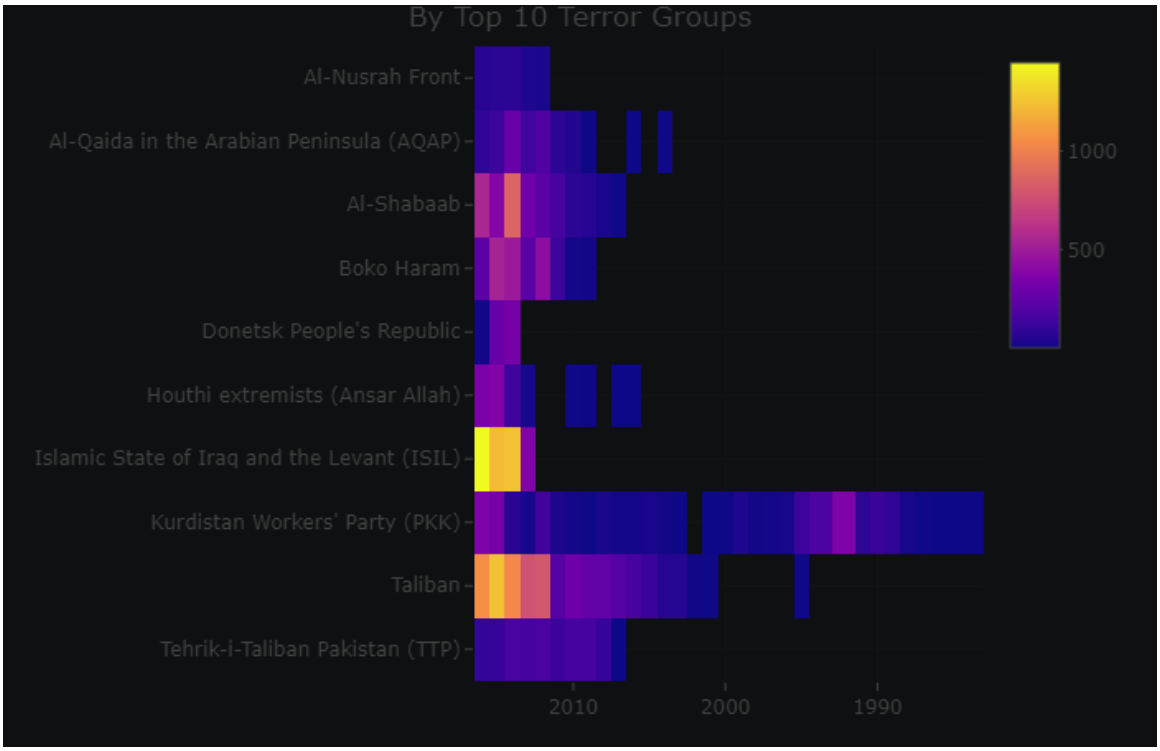

Figure 2.6: Attack Frequency from Top 10 Groups

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.

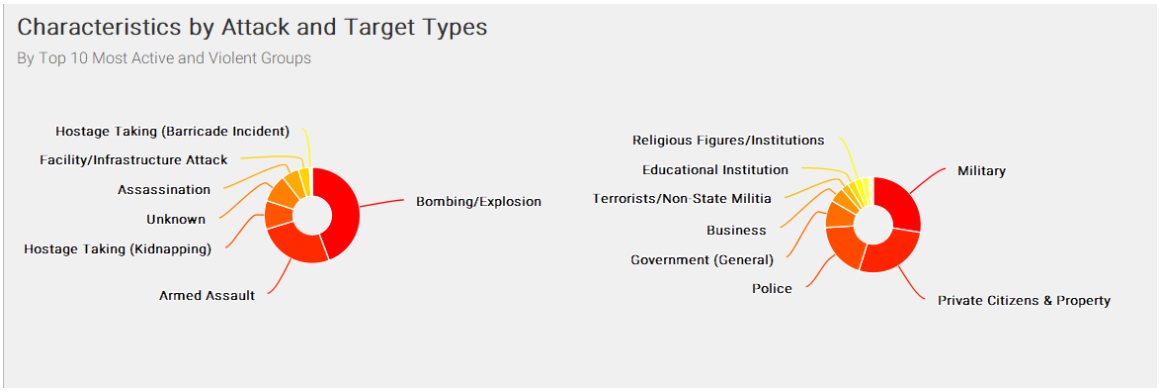

Figure 2.7: Characteristics of top 10 groups

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 examing 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-Suharan 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"))
```

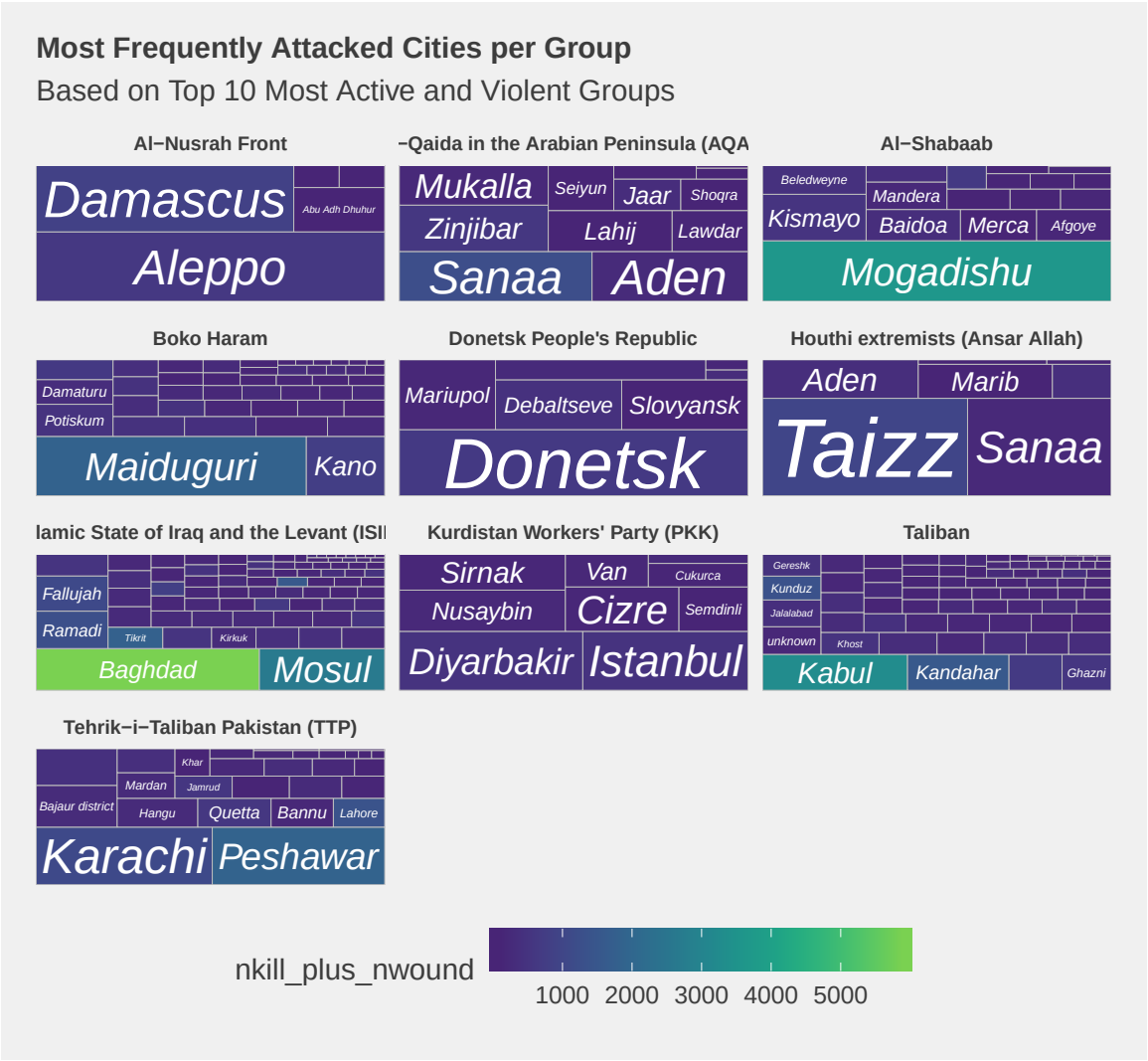

Figure 2.9: Terrorist Group and Impacted Cities

Chapter 3

Statistical Hypothesis Testing

This chapter is an extension to chapter 2 (Impact Analysis) and peforms 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")
```

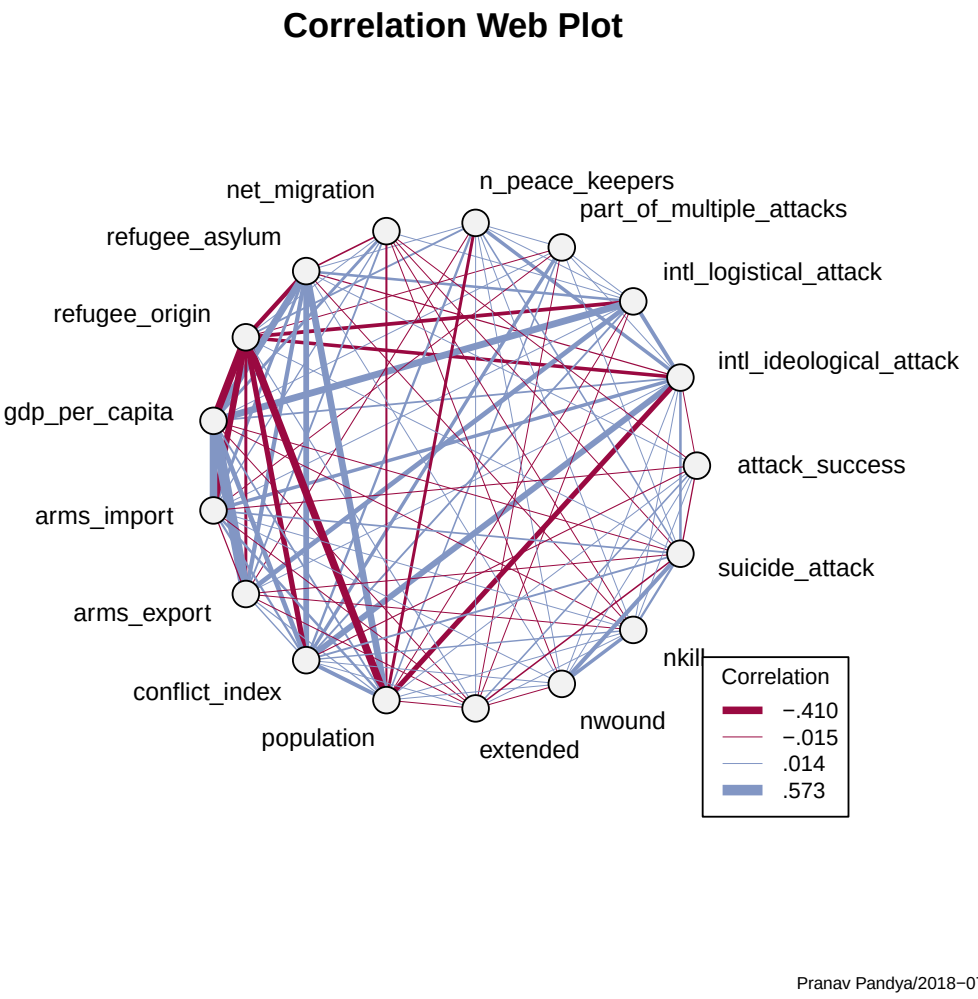

Figure 3.1: Correlation web plot

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

$$(ISIL) = (Taliban) = (AQAP) = (Al - Shabaab) = (Al - Shabaab) = (TTP) = (BokoHaram) = (Al - Nusrah) = (Donetsk_PR) = (Houthi_Extrm)$$

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")
```

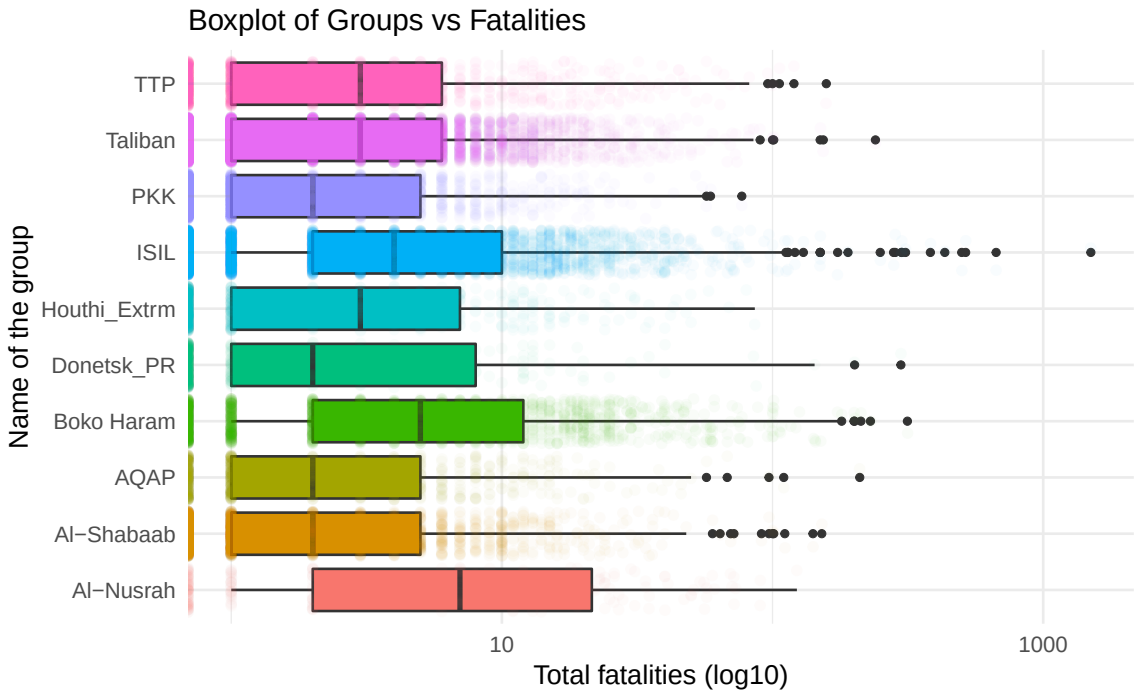

Figure 3.2: Boxplot: Group vs Fatalities

In statistical terms, we have some extreme outliers i.e. nkill ~ 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.

$$F - statistic = (S^2_{between} \ / S^2_{within})$$

```
# 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")$group_name[,"pval"],          # The Fishe
  scheffe= PostHocTest(r.aov, method="scheffe")$group_name[,"pval"], # Scheffe
  bonf=PostHocTest(r.aov, method="bonf")$group_name[,"pval"]))       # Bonferron

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 = 10, 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 signicant different means in number of fatalities such as pairs of ISIL with all the remanining 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

Discovering Patterns in Top 10 Groups

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 10 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).

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, nkil
  filter(target_type != "Unknown" & weapon_type != "Unknown" & attack_type != "U
```

```
  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"] <- "Religiou

#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 8
 $ target_type   : Factor w/ 20 levels "Airports & Aircraft",..: 10 10 2 3 3 3 3 6
 $ weapon_type   : Factor w/ 8 levels "Chemical","Explosives",..: 2 3 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 1
```

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 * nrow(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 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)
} else {
  knitr::kable(x = inspect(subrules), booktabs = TRUE,
               caption = "Five Most Important Patterns (ISIL)")
}
```

```
    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")
}
```

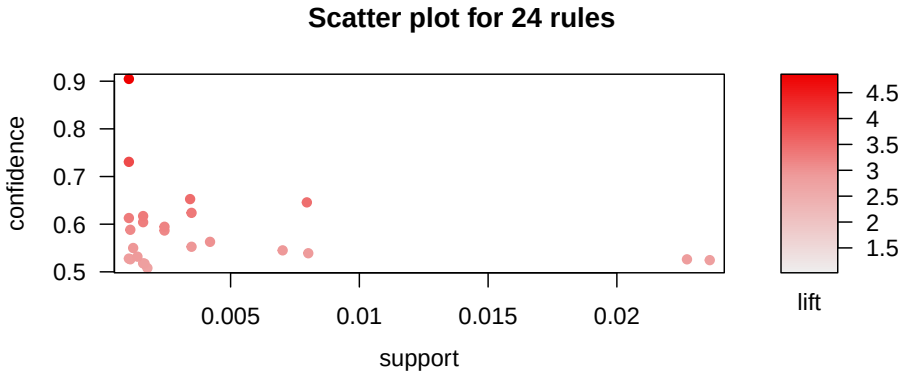

Figure 4.1: Association Rules in ISIL Group

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$vertic
    ig_df$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))
}
```

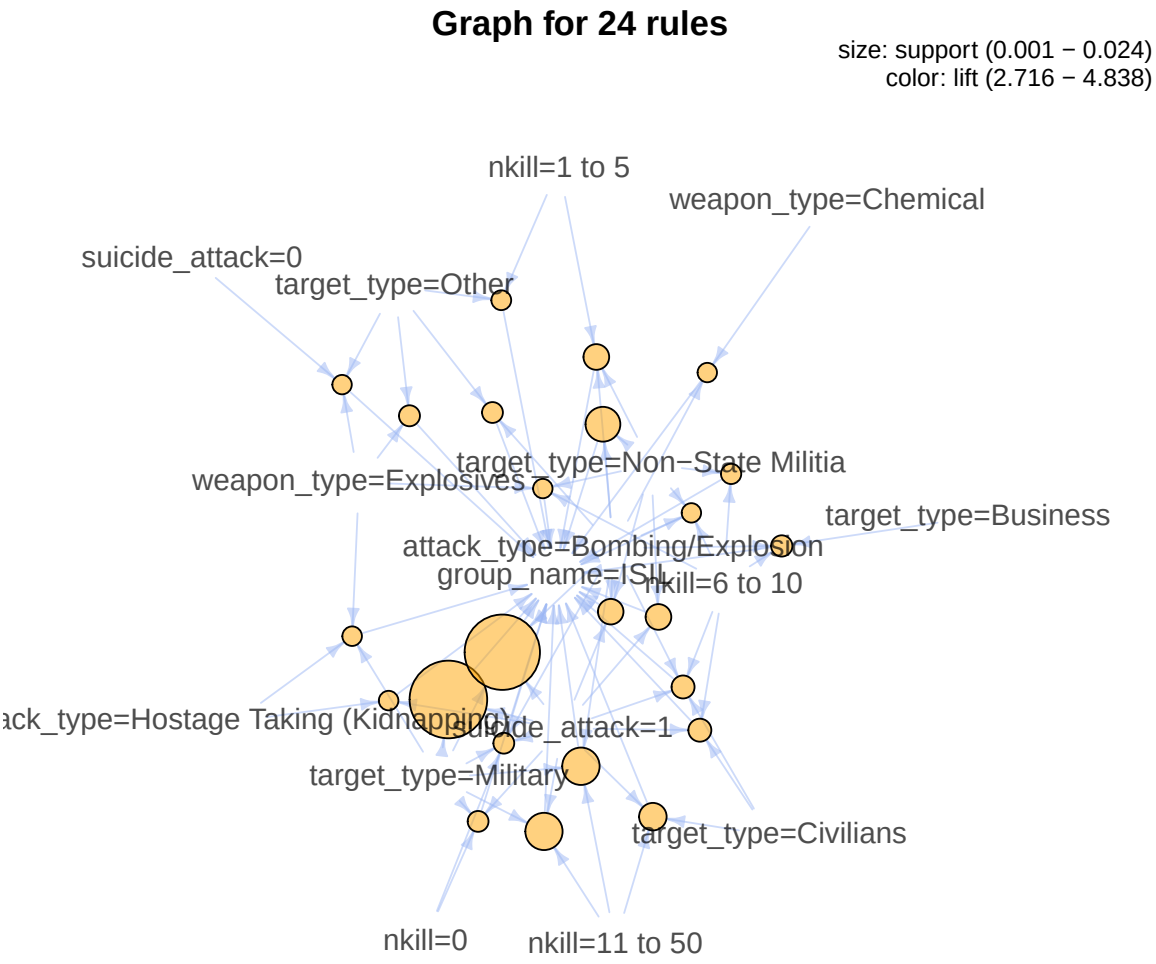

Figure 4.2: Network Graph of Discovered Patterns- ISIL Group

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 [0.02s].
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].
```

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

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
```

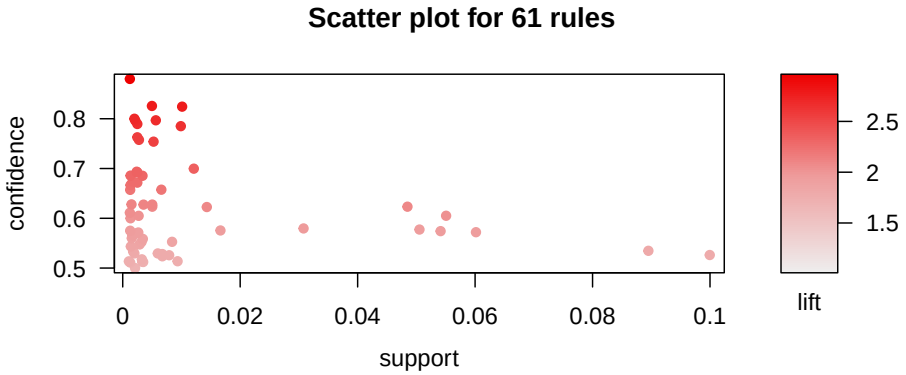

Figure 4.3: Association Rules in Taliban Group

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.

4.3.2 Network graph (Taliban)

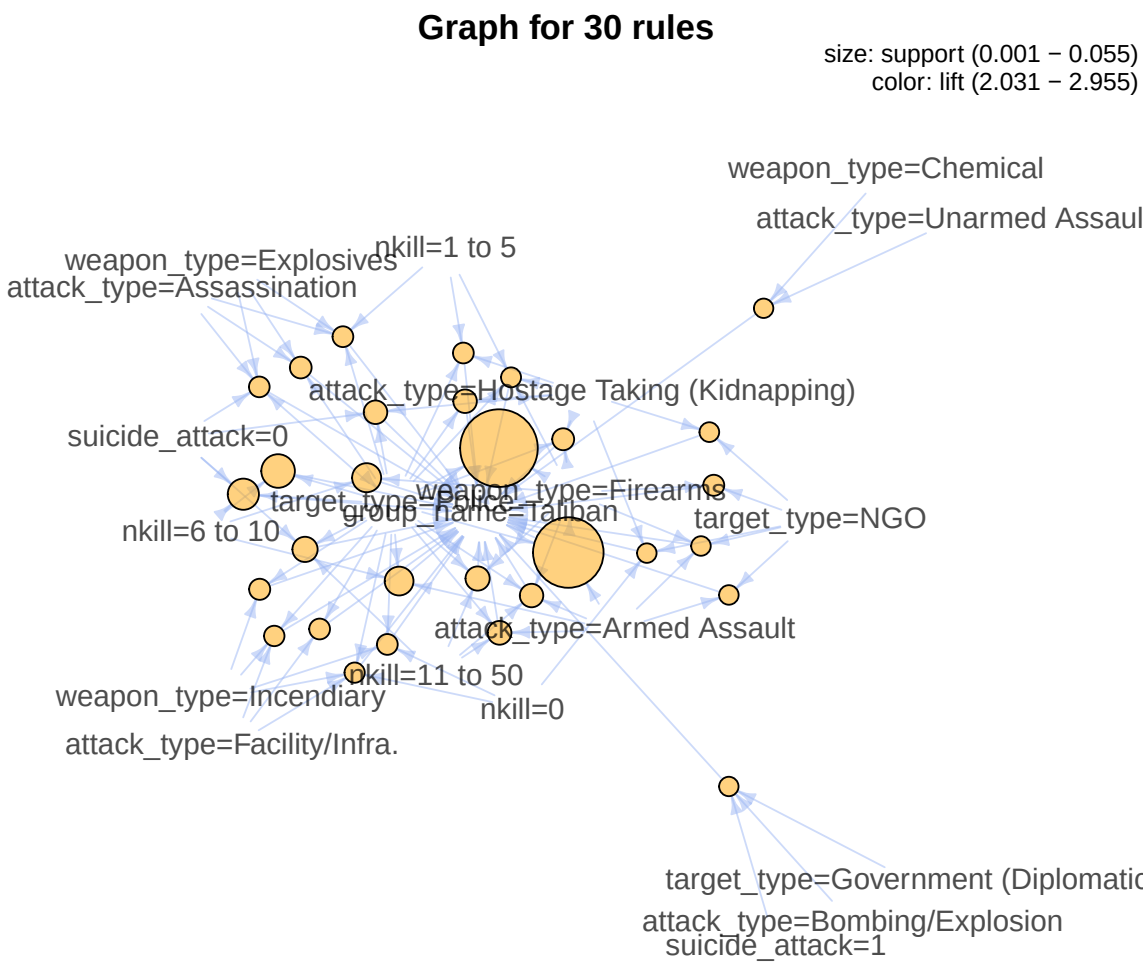

Figure 4.4: Network Graph of Discovered Patterns- Taliban Group

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 [0.00s].
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].
```

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

4.4.2 Top 5 Patterns (Boko Haram)

```
    lhs                                rhs                      support confidence  li
[1] {target_type=Civilians,
     weapon_type=Explosives,
     suicide_attack=0,
     nkill=more than 50}        => {group_name=Boko Haram} 0.001106     0.8000 7.68
[2] {target_type=Civilians,
     weapon_type=Explosives,
     attack_type=Armed Assault,
     nkill=11 to 50}            => {group_name=Boko Haram} 0.001106     0.7692 7.39
[3] {target_type=Civilians,
     attack_type=Armed Assault,
     nkill=more than 50}        => {group_name=Boko Haram} 0.001548     0.7568 7.27
[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
```

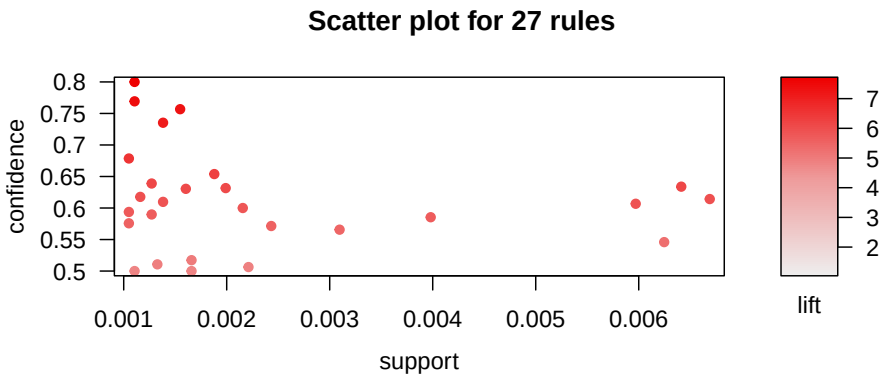

Figure 4.5: Association Rules in Boko Haram Group

In case of Boko Haram,

4.4.3 Network graph (Boko Haram)

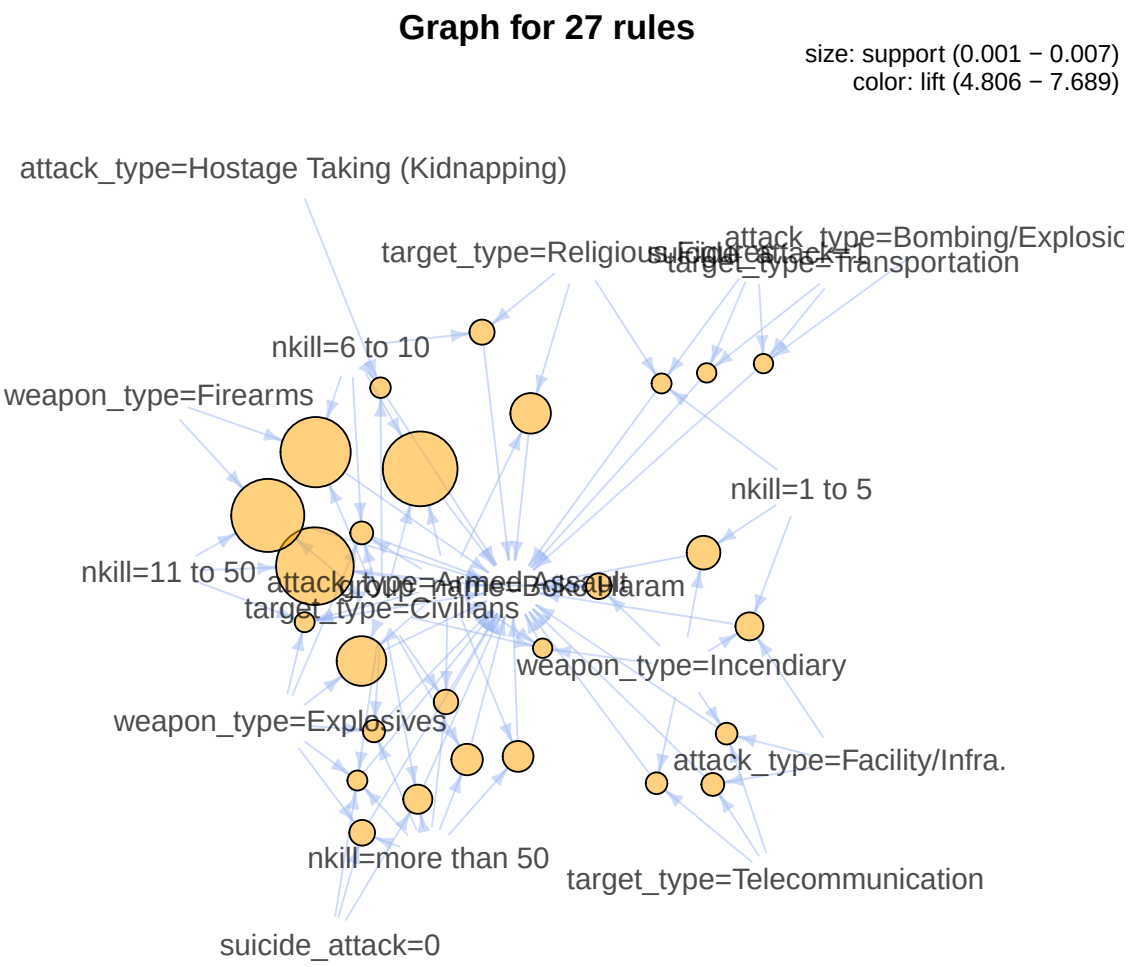

Figure 4.6: Network Graph of Discovered Patterns- Boko Haram Group

Chapter 5

Time-series Forecasting

5.1 Tables

In addition to the tables that can be automatically generated from a data frame in **R** that you saw in [R Markdown Basics] using the `kable` function, you can also create tables using *pandoc*. (More information is available at `http://pandoc.org/README.html#tables`.) This might be useful if you don't have values specifically stored in **R**, but you'd like to display them in table form. Below is an example. Pay careful attention to the alignment in the table and hyphens to create the rows and columns.

Table 5.1: Correlation of Inheritance Factors for Parents and Child

Factors	Correlation between Parents & Child	Inherited
Education	-0.49	Yes
Socio-Economic Status	0.28	Slight
Income	0.08	No
Family Size	0.18	Slight
Occupational Prestige	0.21	Slight

We can also create a link to the table by doing the following: Table 5.1. If you go back to [Loading and exploring data] and look at the `kable` table, we can create a reference to this max delays table too: Table **??**. The addition of the `(\#tab:inher)` option to the end of the table caption allows us to then make a reference to Table `\@ref(tab:label)`. Note that this reference could appear anywhere throughout the document after the table has appeared.

5.2 Figures

If your thesis has a lot of figures, *R Markdown* might behave better for you than that other word processor. One perk is that it will automatically number the figures accordingly in each chapter. You'll also be able to create a label for each figure, add a caption, and then reference the figure in a way similar to what we saw with tables earlier. If you label your figures, you can move the figures around and *R Markdown* will automatically adjust the numbering for you. No need for you to remember! So that you don't have to get too far into LaTeX to do this, a couple **R** functions have been created for you to assist. You'll see their use below.

In the **R** chunk below, we will load in a picture stored as `reed.jpg` in our main directory. We then give it the caption of "Reed logo", the label of "reedlogo", and specify that this is a figure. Make note of the different **R** chunk options that are given in the R Markdown file (not shown in the knitted document).

```
include_graphics(path = "figure/reed.jpg")
```

Figure 5.1: Reed logo

Here is a reference to the Reed logo: Figure 5.1. Note the use of the `fig:` code here. By naming the **R** chunk that contains the figure, we can then reference that figure later as done in the first sentence here. We can also specify the caption for the figure via the R chunk option `fig.cap`.

Below we will investigate how to save the output of an **R** plot and label it in a way similar to that done above. Recall the `flights` dataset from Chapter **??**. (Note that we've shown a different way to reference a section or chapter here.) We will next explore a bar graph with the mean flight departure delays by airline from Portland for 2014. Note also the use of the `scale` parameter which is discussed on the next page.

```
flights %>% group_by(carrier) %>%
  summarize(mean_dep_delay = mean(dep_delay)) %>%
  ggplot(aes(x = carrier, y = mean_dep_delay)) +
  geom_bar(position = "identity", stat = "identity", fill = "red")
```

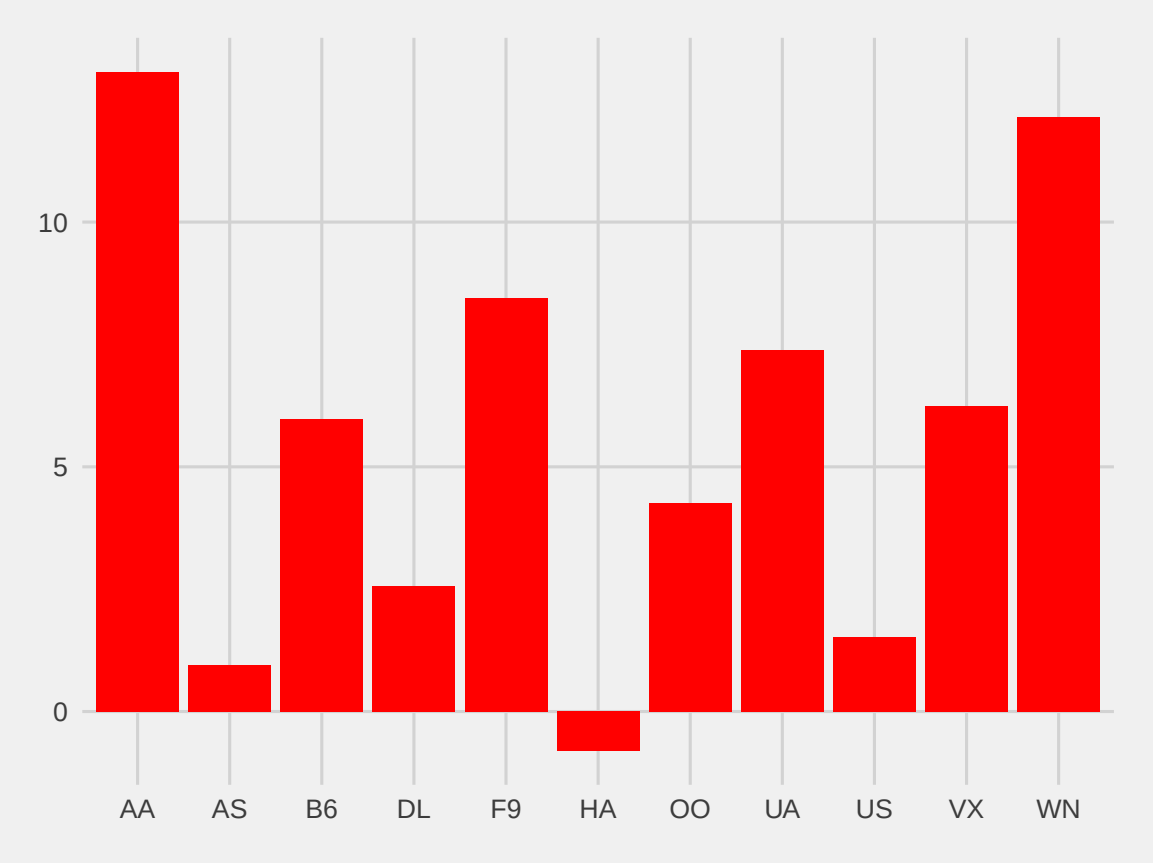

Figure 5.2: Mean Delays by Airline

Here is a reference to this image: Figure 5.2.

A table linking these carrier codes to airline names is available at `https://github.com/ismayc/pnwflights14/blob/master/data/airlines.csv`.

Next, we will explore the use of the `out.extra` chunk option, which can be used to shrink or expand an image loaded from a file by specifying `"scale= "`. Here we use the mathematical graph stored in the "subdivision.pdf" file.

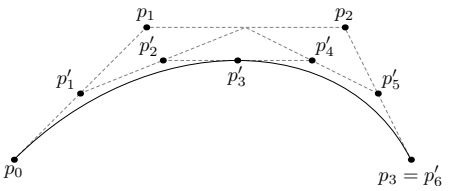

Figure 5.3: Subdiv. graph

Here is a reference to this image: Figure 5.3. Note that `echo=FALSE` is specified so that the **R** code is hidden in the document.

More Figure Stuff

Lastly, we will explore how to rotate and enlarge figures using the `out.extra` chunk option. (Currently this only works in the PDF version of the book.)

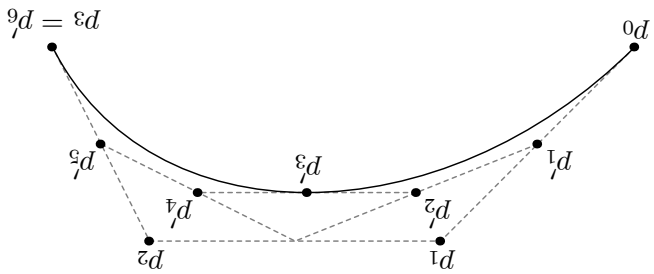

Figure 5.4: A Larger Figure, Flipped Upside Down

As another example, here is a reference: Figure 5.4.

5.3 Footnotes and Endnotes

You might want to footnote something.[1] The footnote will be in a smaller font and placed appropriately. Endnotes work in much the same way. More information can be found about both on the CUS site or feel free to reach out to `data@reed.edu`.

5.4 Bibliographies

Of course you will need to cite things, and you will probably accumulate an armful of sources. There are a variety of tools available for creating a bibliography

[1] footnote text

database (stored with the .bib extension). In addition to BibTeX suggested below, you may want to consider using the free and easy-to-use tool called Zotero. The Reed librarians have created Zotero documentation at `http://libguides.reed.edu/citation/zotero`. In addition, a tutorial is available from Middlebury College at `http://sites.middlebury.edu/zoteromiddlebury/`.

R Markdown uses *pandoc* (`http://pandoc.org/`) to build its bibliographies. One nice caveat of this is that you won't have to do a second compile to load in references as standard LaTeX requires. To cite references in your thesis (after creating your bibliography database), place the reference name inside square brackets and precede it by the "at" symbol. For example, here's a reference to a book about worrying: (**???**). This `Molina1994` entry appears in a file called `thesis.bib` in the `bib` folder. This bibliography database file was created by a program called BibTeX. You can call this file something else if you like (look at the YAML header in the main .Rmd file) and, by default, is to placed in the `bib` folder.

For more information about BibTeX and bibliographies, see our CUS site (`http://web.reed.edu/cis/help/latex/index.html`)[2]. There are three pages on this topic: *bibtex* (which talks about using BibTeX, at `http://web.reed.edu/cis/help/latex/bibtex.html`), *bibtexstyles* (about how to find and use the bibliography style that best suits your needs, at `http://web.reed.edu/cis/help/latex/bibtexstyles.html`) and *bibman* (which covers how to make and maintain a bibliography by hand, without BibTeX, at `http://web.reed.edu/cis/help/latex/bibman.html`). The last page will not be useful unless you have only a few sources.

If you look at the YAML header at the top of the main .Rmd file you can see that we can specify the style of the bibliography by referencing the appropriate csl file. You can download a variety of different style files at `https://www.zotero.org/styles`. Make sure to download the file into the csl folder.

Tips for Bibliographies

- Like with thesis formatting, the sooner you start compiling your bibliography for something as large as thesis, the better. Typing in source after source is mind-numbing enough; do you really want to do it for hours on end in late April? Think of it as procrastination.
- The cite key (a citation's label) needs to be unique from the other entries.
- When you have more than one author or editor, you need to separate each author's name by the word "and" e.g. `Author = {Noble, Sam and Youngberg, Jessica},`.
- Bibliographies made using BibTeX (whether manually or using a manager) accept LaTeX markup, so you can italicize and add symbols as necessary.
- To force capitalization in an article title or where all lowercase is generally used, bracket the capital letter in curly braces.
- You can add a Reed Thesis citation[3] option. The best way to do this is to use the phdthesis type of citation, and use the optional "type" field to enter "Reed

[2](**???**)
[3](**???**)

thesis" or "Undergraduate thesis."

5.5 Anything else?

If you'd like to see examples of other things in this template, please contact the Data @ Reed team (email `data@reed.edu`) with your suggestions. We love to see people using *R Markdown* for their theses, and are happy to help.

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 readibility 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, treemapi
               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 ??:

```
# This chunk ensures that the thesisdown package is
# installed and loaded. This thesisdown package includes
# the template files for the thesis and also two functions
# used for labeling and referencing
if(!require(devtools))
  install.packages("devtools", repos = "http://cran.rstudio.com")
if(!require(dplyr))
    install.packages("dplyr", repos = "http://cran.rstudio.com")
if(!require(ggplot2))
    install.packages("ggplot2", repos = "http://cran.rstudio.com")
if(!require(ggplot2))
    install.packages("bookdown", repos = "http://cran.rstudio.com")
if(!require(thesisdown)){
  library(devtools)
  devtools::install_github("ismayc/thesisdown")
  }
library(thesisdown)
flights <- read.csv("data/flights.csv")
```

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

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*(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

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 //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

Indiana University Libraries. (2007, July). Identifying Primary and Secondary Sources. *Indiana University Bloomington*. Retrieved from https://libraries.indiana.edu/identifying-primary-and-secondary-sources

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 Associa-

tion 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.

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

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/10.1080/09546550109609697

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.