Global Terrorism Database

Max Herrera, Joshua Herrera, Patrick McDevitt, and Sunna Quazi

MSDS 7330 File Organization and Database Management

Section 405, Group A

**Abstract:** The following is an analysis of the of the Global Terrorisim Database (“GTD) dataset, which enumerates terrorist attacks from 1970 – 2015 and can be located at <https://www.kaggle.com/START-UMD/gtd>. A crucial component of initiating a data analysis is to frame the data in a manner that relational attirbutes can be clearly noted and easily manipulated, which is why Structured Query Language (“SQL”) will be used for this study. The realtional structure of SQL will be essential in analyzing the multiple subtypes of information: geographic, forensic, demographic, political, etc. that are used to describe each terror attack. Diving into this dataset, we hope discern if there is geographic clustering of terror attacks and if the burst concentration of those terror attacks trend for an average length of time. If an origin of a burst is discernable, then we may be able to determine external causality from corresponding historical events and possibly a common causality by aggregating the data. We would also want to determine if there are cumulative insights on the types of terror attacks that were used over time and the methodologies in which they were carried out.

**Related Work:** The GTD is analyzed by multiple users on Kaggle. A notable contribution is from Abigail Larion using python, which was last accessed 8 months ago. Her analysis is entitled “Terrorist Attacks in the United States” where she highlights the attacks on American soil by longitude and latitude, focusing specifically on which resulted in fatalities and injuries. Our analysis deviates from Abigail’s as it analyzes the dataset on a global scale to understand causation rather than ramification.

Another Kaggle contribution is from Pranav Pandya entitled “Visualizing Terrorist Attacks on India” , which was last run five months ago. The analysis uses R, which may explain why the visualzations are aesthetically more pleasing than the previous analysis.The premier visualzation shows a map of India with each terrorist attack as dot. This allows you to see a trend of attacks that starts predominantly on the northern border of India and down to the western border, ending at Visakhapatnam. The analysis goes on to break down which terrorist/rebel group was the aggressor by year. From a historical standpoint, the analysis is invaluable to understanding the chronology of conflict in India; however it differs from our analysis in that it does not include a predictive trend, and does not include a global overview. Also, the analysis focuses on the transgressing party, whereas our analysis that follows does not.

1. **Introduction**

In its most basic form, “terrorism” can be defined as using violence to coerce a party into action. With this definition, many events could be construed as terrorism. In order to analyze this dataset, it is critical to understand its specific definitions and bright lines for determining terror: "The threatened or actual use of illegal force and violence by a non-state actor to attain a political, economic, religious, or social goal through fear, coercion, or intimidation." Some historical background is needed to understand the full breadth of this data set. If to qualify as terrorism, any act that uses violence as means to an end by any non-government entity would mean that any failed coup in this history of humanity would be in this dataset. Noting that no data is included that exists before 1970, infers that this dataset is exclusively defined by what historians would call “modern terrorism”.

**Project File Location:** Our GitHub site with a majority of the files used for analysis along with the presentation and term paper is: <https://github.com/bici-sancta/global_terrorism>. All the scripts that were created to wrangle, and analyze the data can be found here. <https://smu.box.com/s/zhbm2idikopnuhxfvcxtmc880nud5pxa>. There were 22 scripts written to wrangle and profile the data. The SQL dump generated to scripts, which can be found in the “importdataintomysql” folder. The stored procedures and SQL used is for both Microsoft SQL server and MySQL.

**Background/Existing Work:** The GTD is the most comprehensive unclassified database on terrorist events in the world. Over 4,000,000 news articles and 25,000 news sources were reviewed to collect incident data from 1998 to 2016 alone. The dataset contains 137 attributes for 156,772 terrorist attacks, so the transactional speed of SQL is also advantageous.

Hardware Specifications: We used multiple devices shared by our group to conduct this analysis. The most basic of hardware specifications is as follows: an HP EliteBook Folio 9480m laptop with an Intel® Core™ i5-4310U CPU @ 2.00 GHz and 7.88 GB of usable RAM on a 64-bit Windows 7 operating system. On each of our machines we used MySQL Workbench operating on our individual local machines on root. This allowed us to create databases, which we could not do on the IBM Bluemix server.

1. **Loading the Data into SQL**

**2.1 Trial and Error**

The best way to determine what to do is to determine what not to do. Initially, we experienced complications in loading the data. We started off by loading the data via a SQL script However, the data actual data from the Global Terrorism table (“GT) was not imported. The schema that resulted with no table available and queries would result had no useful data

**Concerns:** We needed to make each column varchar(MAX) initially. Each column was surrounded by triple quotes, so we cleaned up the data.

**2.2 Successful Attempts**

To remedy this issue, we imported the data into a Microsoft SQL Server database. Afterwards, we wrote scripts to determine the max length of each column, and altered the column width for each of the 137 columns accordingly. We outputted the distribution of values per column into one table, DISTRIBUTION\_TABLE. Subsequently, we connected the SQL server to MYSQL to transfer the data using SSIS via an ODBC connection. After the data was in MYSQL, we performed a MYSQL dump of the data for the GT (I.e., the global terrorism table) and the DISTRIBUTION\_TABLE and uploaded it into box so that it could be loaded by the other teammates. The resulting file increase the previously 29mb data to a 201mb SQL file, but it allowed the team to start with the same dataset.

**2.3 Data Profiling**

Unfortunately, the method of uploading the entire dataset in one table hindered us by inundating us with data. We were bombarded with all 137 attributes, most of which were sparsely populated or NULL. We had to remove data in order to focus on key attributes that may be able to help us in determining correlation and causation. The criterion for sparseness was that 80% or more of the column was NULL or contained ‘.’. We then created alter tables for the sparse columns and dropped them. The final result left us with 66 attributes and almost a 52% reduction in inutile attributes.

At this point, the table was ready for normalization.

**2.4 Data Normalization**

The data was profiled using SQL Server Data Profiler.

The profiler suggested that nine tables could be extracted from the large GT table. Individual tables were created using the columns from GT. These nine individual tables were validated by joining back to GT. The redundant columns were dropped form GT because they had been moved to the individual tables. Primary keys were created on the nine tables. From there, indices were created on the join columns in the GT table. Finally, a nine-way inner join was performed with the GT table to validate that everything has worked correctly.

1. **Data Analysis**

**3.1 Excel**

We first investigated the dataset using MS excel. We completed this by performing a text analysis on the 200 most common phrases located in the file. See Figure 1. 1970 
1971 
augus t 
police 
firearm 
ex Jos Ve 
bomb 
1972 
auto matic 
ex 
firearm 
automatic 
1975 
1974 
ex 
rue 
bomb 
catholic 
residence 
ex 
automatic 
s tation 
1978 
civil 
residence 
christian 
democratic 
ex 
automatic 
s tation 
vehicle 
1981 
de 
ex 
automatic 
station 
electrical 
ants 
rocket 
ex 
automatic 
electrical 
station 
bank 
vehicle 
1985 
ex 
automatic 
line 
s tation 
electrical 
mne 
center 
re made 
branch 
central 
ex 
automatic 
s tation 
bomb 
car 
residence 
rocket 
branch 
school 
ex 
automatic 
s tation 
branch 
residence 
home 
ex 
automatic 
residence 
branch 
firearm 
automatic 
s tation 
firearm 
rue 
s tation 
1999 
october 
november 
international 1976 
embass 
disorders 1973 
o rations black 
embass 
rpetrators 
bomb 
a List 
s ember 
december 
november 
p be 
attacks 
motive 
inc dent 
october 
consulate 
terrorism 
violence 
s tation 
october 
catholic s 
s tation 
consulate 
offices 
buildi 
international 
house 
rocket 
rocket 
residence 
residence automatic 
sus 
rebels 
ted 
cas uatties 
unvers 
violence 
office 
zebra 
32 
mne 
buildi 
school 
c atholic 
school 
activists 
1997 
rocket 
candidate 
attac ked 
cobn-,bia 
december 
30 
wounded 
15 
18 
sus 
shot 
ted 
bomb 
civilians 
nov ember 
inc ident 
sus ted 
attacks 
c obmbia 
bus 
december 
bomb 
franc e 
claimed 
hdad 
kill 
oc tober 
attacks 
sus ted 
wounded 
suicide 
11 
bomb 
claimed 
wounded 
kill 
claim 
attac k 
national 
incidents 
track' 
attac k 
2011 
ac ademic 
lexis nexis 
terrorism 
counterterror ex 
re 
rted 
suicide 
lexis nexis 
academic 
ed 
wounded 
dec ember 
franc e 
november 
reuters 
rovised 
south 
secur 
watc h 
civilians 
wounded 
res 
nov 
sec 

*Figure 1*

From this analysis we were able to make several conclusions on the theme for each decade:

1. 1970s: Catholics vs. Protestants, US, Puerto Rico
2. 1980s: Abortion, Israel
3. 1990s: Turkey, Israel, India

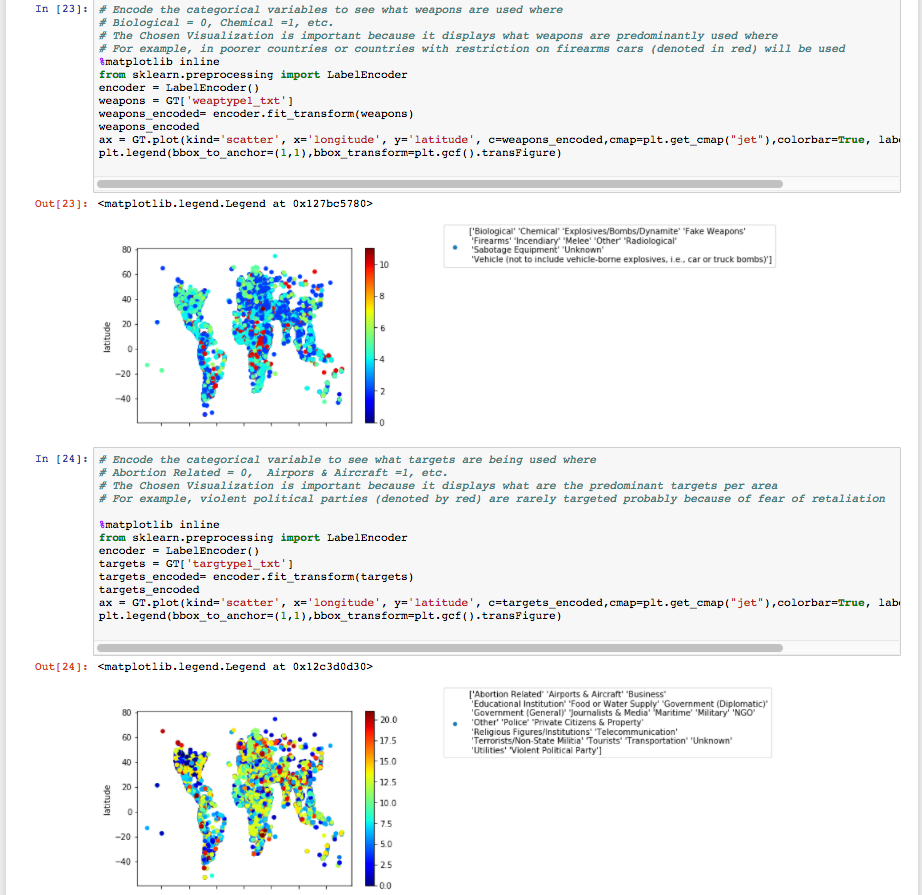
The first conclusion confirms our assertion that the data set is based exclusively off of modern terrorism. It is commonly understood that the inception of modern terrorism originated with in Northern Ireland in the pursuite to disolve the Anglican Church of Ireland. As the dataset starts with a high volume of these attacks, it can only be concluded that the dataset revolves around the modern definition of terrorism.

**3.2 SQL**

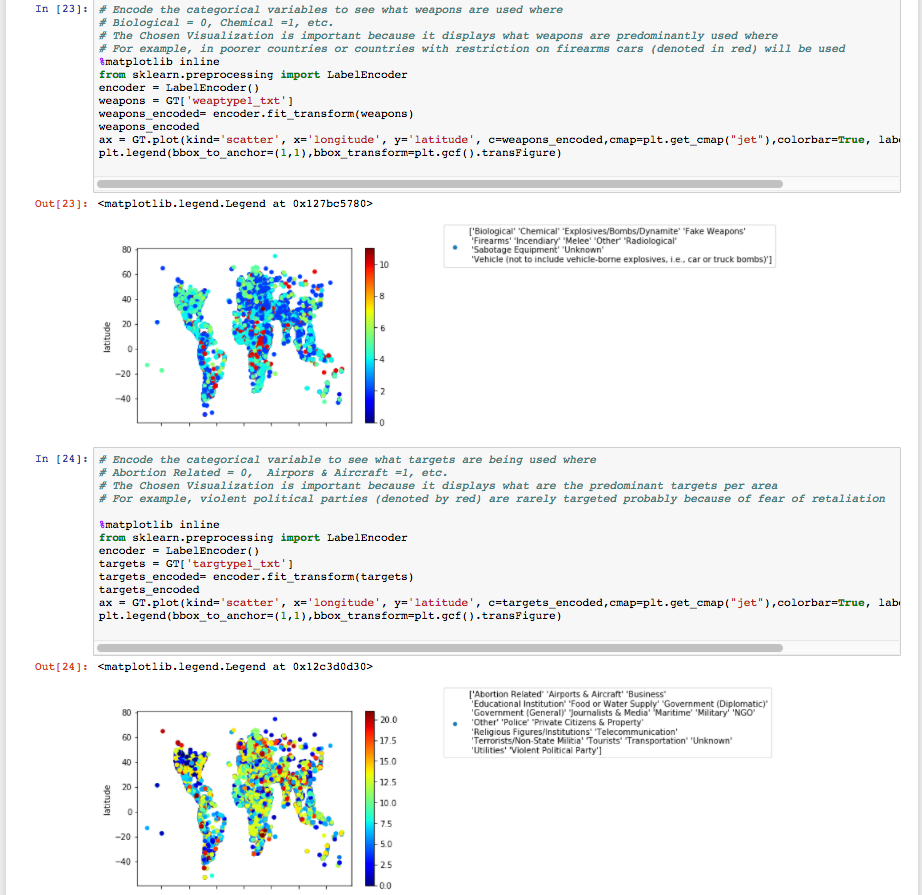
After loading the data in SQL, we first tried to run some basic queries to get a general idea of the dataset. We determined that the countries with the largest number of attacks total were Iraq, Pakistan, India, and Afghanistan. Despite narrowing our query to only attacks that were categorized as bombings, the same countries still held the highest attacks. We then constricted our analysis even further to limit attacks to pre 1990 and the countries with the highest attacks were El Salvador, Peru, United Kingdom, and Colombia. This allowed us to discern that there was a distinct point in time where trends in terrorist attacks shifted.

**3.3 Python**

Unfortunately SQL is limited in its ability to perform visualizations. To append our analysis, we used python and were able to determine from the resulting visualizations (See Figure 14 and 15) that the weapon of choice in indigent countries is the motor vehicle. The favorite target of terrorists are abortion related centers and airports.



*Figure 14*

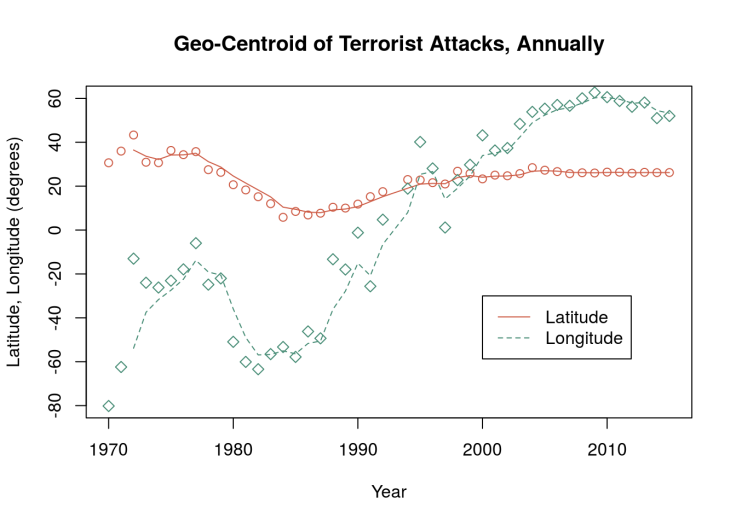


*Figure 15*

**3.4 R**

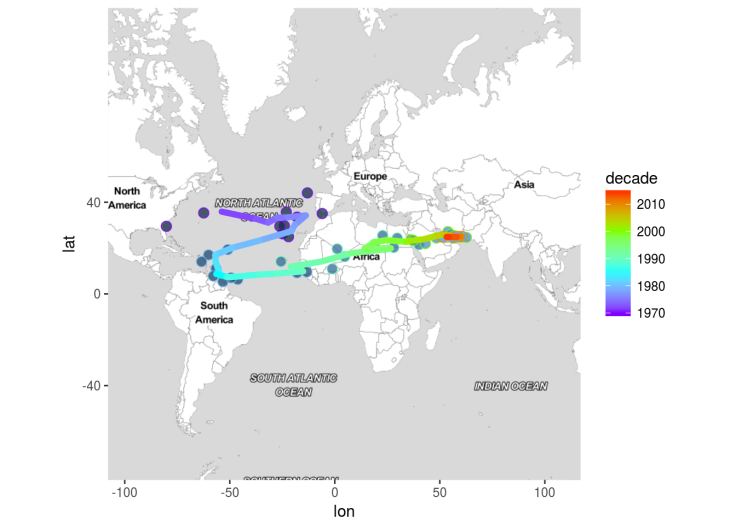
We elaborated on our analysis by performing visualization in R. To understand the geographic changing nature of terrorism over this period, we determine the geographic center of all the events that occurred in each year. That is to say, for all of the terror events that occurred in 1970, and for each subsequent year, the geo-centroid of the all of the latitude and longitude coordinates was determined.

The result of this is depicted in Figure 16. We can observe that in 1970 the “center” of terrorist activity occurred at latitude, longitude (30.649873, -80.178993), which corresponds to a location in the North Atlantic Ocean, approximately 40 miles east of Jacksonville, Florida. This means that of the 651 reported terrorist events reported in 1970, the geographic center was just off of the east coast of the United States. Furthermore, we can observe that during the subsequent 10 years of reported terrorist activity that the center of activity consistently moved eastward towards the European coast, then reverted back towards the Americas in the mid-1980s, and then made a steady progression, year by year eastward to arrive in the Persian Gulf about 20 miles NE of Qatar, at latitude, longitude (26.204019, 51.978154).



*Figure 16*

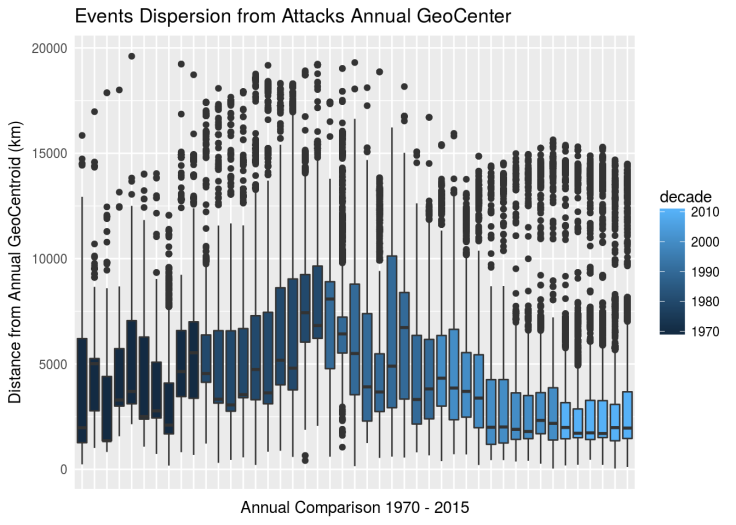
This can be more easily visualized on an earth map, as shown in Figure 16. Here, a dot is presented that represents the geo-centroid for each year’s reported terrorist activity, along with a connecting line to show the steady progression from west to east during this 45 year time period. This is likely a view of the history of terrorism either never known (for everyone under 50) or no longer remembered by most in the United States.



*Figure 16*

Another approach to visualize the trends is not just with the geographic center, but also by the dispersion from the center and the distributional characteristics. Figure 17 presents box plots for the annual distribution of terrorist events for the 45 year period. This visualization presents the distribution of the distance of each event from the geo-centroid of all events for that year. In other words, from the annual center, how closely (or disparately) spaced are all of the events for that year. We can make the following observations :

* 1970s - large majority of events occurred within a 5000 km distance from the geocentroid
* 1980s - the average distance increases slightlyto 5000 km from the centers (now located between Western Europe or northern South America) and a clear bi-modal trend begins to emerge. There is a population of events that appear consistently at more than 15000 km from the annual center.
* 1990s - slight reduction in population distance from the annual mean, and also a return to a primarily uni-model distribution, as the overall centroid begins to march eastward across Africa
* 2000s - much stronger concentration of events to the annual center - now generally within 4000 km of the annual center and a reappearance of a bi-modal distribtion of events
* 2010s - least dispersion in the overall population difference from the annual centers. Seventy-five percent or more of the events in each year are contained within 2000 km of the annual center, but also, the development of three population distributions, one located at 7000 km from the centers and the other at 13000 km from the primary location. Thus, in the current era, there is a very strong concentration of terrorist events in the Mid-East, but also two other locations affected by 25% of the events.



*Figure 17*

Lastly, we created a text analysis chart in R that showed the changing language used in the events summary to describe / characterize the events in each decade. This was constructed from text mining the event summaries 150,000+ events. For each decade, significant words from each decade were discovered from among the words that were not used in previous decades. This analysis allowed us to see emerging themes in Terror over time. In the 1980s, the words “Bethesda” “Ashley” and “Brockhoeft” were major results that conclude a new trend in abortion center attacks. In 2010, Saladin was a major result that refers to a 12th century Kurdish military leader. Although Saladin was not directly involved in 2010 attacks, it is a result as it is a common illusion used in the emerging trend of radical islamist terror as it discuses old world Muslim military campaign against the Crusader states in the Levant. He is referenced in the current era associated to the brutality of the methods by which he gained and retained power.

1. **Conclusion**

Every day, there is more and more extractable data available in every possible subject, whether mundane or critical to personal security. What separates real, tenable, insight from noise is the value that is extracted when analyzing that data. With the Global Terrorism Database, the goal of analysis is intuitive. Modern terrorism is a blight on humanity, as it not only claims lives, but also the livelihood, mental wellbeing of its survivors. Our end goal was to delve into this topic to find any information that could be used to prevent these heinous acts from occurring.

With the use of SQL, we were able to delve into a seemingly overwhelming amount of data. As data grows exponentially, it can be staggering to manipulate the amount of data that is generated in just a few years, much less decades. SQL allowed us to import an exceedingly large amount of data, perform transactions, and normalize the data so we are able to grasp the full spectrum of a very broad vastly encompassing set. It is also rare that there is data that reaches as far as the 1970s; before notions of Big Data were a thought, no one would have thought to save specifics regarding the emerging trend of modern terrorism. In this case, we were fortunate to have the data cultivated for us, but in doing this, relationships were pre-established in a tabular format. SQL allowed us to work with the data in its native form so we could focus on further trimming down erroneous variables and obtaining a concentrated set that would enable us to analyze our data and discover trends.

As our dataset indicates that the reasons for causing terror can be numerous, so prevention based off of deterring motive would be difficult. We are however to capitalize on similarities that can point to certain aspects as being indicators or potentially risky attributes of Terror. With our Python analysis, we were able to determine the primary utensil of these antagonists: motor vehicles. We were also able to discover that two unrelated locations share a common likelihood in being the setting of a terrorist attacks: Abortion Centers and Airports.

It is not enough to just evaluate what has retroactively occurred in the past. Yes, it is true that history has a tendency to repeat itself, but as modern terrorism is a relatively recent, 40 year old, phenomenon; the escalating nature of violence could give birth new threats that we are not yet aware of. With our R analysis, we were able to see a solid trend line regarding geographic location for seemingly unrelated events. This level of insight unlocks predictive capabilities in determining ways to safeguards from future insurgent and insurrectionist attacks.

### References

[1] National Consortium for the Study of Terrorism and Responses to Terrorism (START), “The Global Terrorism Database, ” *University of Maryland,* <https://www.kaggle.com/START-UMD/gtd>.

[2] Devon McCann, “Northern Ireland: The Origin of Modern Terrorism,” *Gerogetown University,* [https://apps.cndls.georgetown.edu/projects/borders/exhibits/show/northern-ireland-modern-terror/societal-division- and-the-ira/ira](https://apps.cndls.georgetown.edu/projects/borders/exhibits/show/northern-ireland-modern-terror/societal-division-%20and-the-ira/ira).

[3] GTD Global Terrorism Database ,“Codebook: Inclusion Criteria and Variables,” *START,* <http://start.umd.edu/gtd/downloads/Codebook.pdf>

[4] Bryan Burrough, “The Bombings of America That We Forgot” Sep 20, 2016, <http://time.com/4501670/bombings-of-america-burrough/>

**[5] SQL Scripts** <https://smu.box.com/s/zhbm2idikopnuhxfvcxtmc880nud5pxa>.

**[6] Group Research Files**

<https://github.com/bici-sancta/global_terrorism>

You have to refer to the following rubric to get the good score as the isntructor will evaluate your paper based on it.