Predicting Outcome of Terrorist Attacks

STAT 471/571/701, Fall 2018

STAT 471/571/701 Modern Data Mining

12/12/2018

Contents

Executive Summary

Acts of intentional violence at a sub-national level have occurred since the development of civilizations. From religious acts of terror to violence with political intent, terrorism has manifested itself in various forms throughout history. The term itself was developed in the 1790s to describe Maximilien Robespierre's Jacobin regime as the "Reign of Terror", but it was popularized following the 1983 Beirut barracks bombings and the 2001 World Trade Center attacks. These acts of violence, regardless of motive, if successful, often claim the lives of the innocent and bystanders.[1]

This leads to the question - how can people use historical data to predict the outcome of terrorist activity? The answer to this lies in statistical analysis of terrorist incidents. Using location data, attack type, group names, target data, and other past terrorism data, we can apply statistical models to best predict the outcome of events - success, number of killed/wounded, total property damage, etc. With a better understanding of what factors can lead to a foiled or successful attempt, people can better try to prevent such attacks in the future.

SUMMARY OF SOME RESULTS<

Data Summary / EDA

Data Origins

The origins of the data is the Global Terrorism Database (GTD). The GTD was developed by the National Consortium for the Study of Terrorism and Responses to Terrorism (START) at the University of Maryland, College Park, in Maryland, USA. The database contains incidents of terrorism from 1970 to 2017, and is still under development. There are over 181,000 incidents in the database and 135 factors, including a few main factors listed below:

- `iyear`, `imonth`, `iday`: incident year, month, and day
- `country_txt`, `region_txt`, `provstate`, `city`: country, region, providence/state, city names
- `crit1`, `crit2`, `crit3`: which of the three criterion the incident satisfies (see below)
- `attacktype1_txt`: a text descriptor for the attack type; there are other variables regarding the type of attack
- `targtype1_txt`, `targsubtype1_txt`, `natlty1_txt`: a text descriptor for the target type, subtype, and nationality; there are other variables regarding the type of targets
- `gname`, `gnucertain1`, `individual`: group name, and indicator variables for presence of guns and if individual attack
- `weaptype1_txt`, `weapsubtype1_txt`: type of weapon used in attack
- `success`, `nkill`, `nwound`, `propextent_txt`: indicates if the incident was successful, the number of killed and wounded, and the extent of property damage (respectively)

To be included in the study, an incident must qualify with three fields: * 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. The database does not include acts of state terrorism.

Additionally, it must satisfy two of the following three criterion: * Criterion 1: The act must be aimed at attaining a political, economic, religious, or social goal. * Criterion 2: There must be evidence of an intention to coerce, intimidate, or convey some other message to a larger audience (or audiences) than the immediate victims. * Criterion 3: The action must be outside the context of legitimate warfare activities.

In general, the GTD does not include plots that are not enacted or attempted. For an incident to be considered, the attackers must be "out the door", or en route to execute the attack. This means, according to their handbook, "in general if a bomb is planted but fails to detonate; if an arsonist is intercepted by authorities before igniting a fire; or, if an assassin attempts and fails to kill his or her intended target, the attack is considered for inclusion in the GTD, and marked success=0."[2]

Goal of the study

The goal of the study is to utilize data on terrorist attacks and identify which factors can be best used to predict the outcome of such attacks. In this study, we will be analyzing various outcomes, from success, number of wounded, number of killed, and total property damage.

EDA

First, we read in the data given in csv format. There are 181,691 observations and 135 total variables. However, this must be futher cleaned. There were three main steps in the data cleaning process for eliminating variables:

1) There were many variables for multiple groups; for instance, there are 3 groups for target (target1, targettype1, targetsubtype1, corp1, target1, natlty1, etc.), 3 groups for attack type, 3 groups for claim,

and 3 groups for weapon types. These are present in the case that multiple groups stage an attack, or multiple targets are targeted. However, for the most parts of the dataset, the second and third group for most predictors were NA, and thus were dropped.

2) We filtered variables that were just encodings of other variables. For instance, there were two variables `country` and `country_txt`. The former is a number encoding for a country, while the latter is the name of the country. For purposes of easier readability, we kept the text description.

3) We finally dropped variables that contained too many NA's. This included number of killed US citizens, group that claimed the incident, etc. Considering the number of US wonded/kill/perp is quite specific, it makes sense that many incident do not report this. Since these caused our models to fail to run, we ended up removing this from the overall data set for the rest of the study.

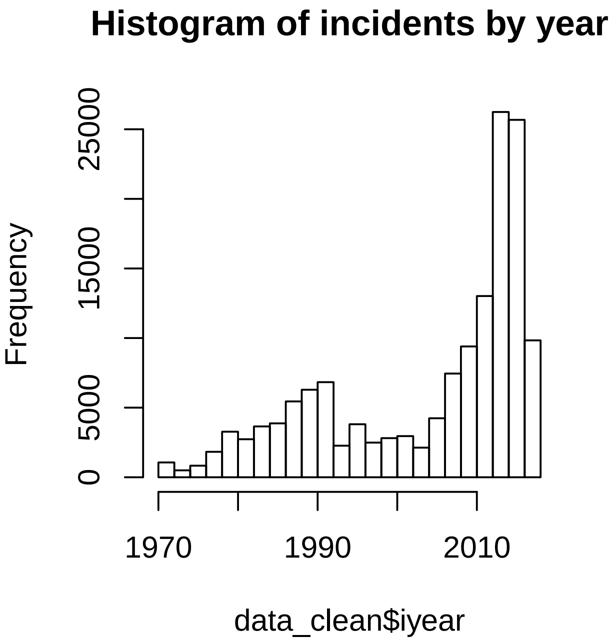
In general, this final cleaned dataset was used as a baseline for each of our subset analysis. See the appendix for the full list of removed variables. Unless otherwise noted, each subset will be derived from these 31 remaining variables (see below). Additionally, the -9 and -99s that were encoded for missing variables were coded into NA in R. We omitted these NA's from the cleaned dataset due to the large number of examples we had from the database already.

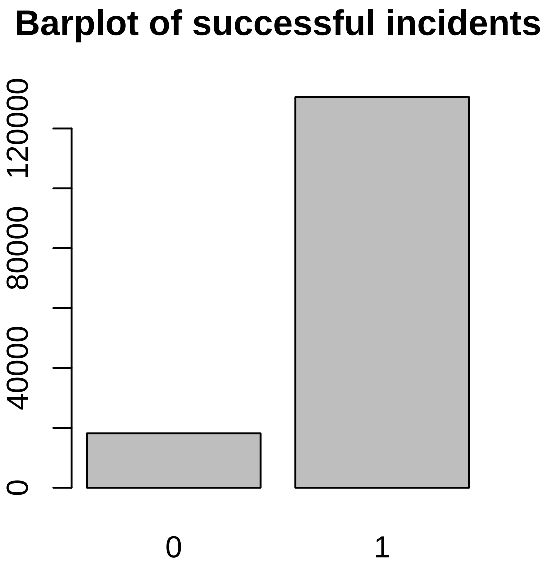
```
##  [1] "iyear"            "imonth"           "iday"            
##  [4] "extended"         "country_txt"      "region_txt"      
##  [7] "provstate"        "city"             "latitude"        
## [10] "longitude"        "specificity"      "vicinity"        
## [13] "crit1"            "crit2"            "crit3"           
## [16] "doubtterr"        "multiple"         "success"         
## [19] "suicide"          "attacktype1_txt"  "targtype1_txt"   
## [22] "targsubtype1_txt" "natlty1_txt"      "gname"           
## [25] "guncertain1"      "individual"       "weaptype1_txt"   
## [28] "weapsubtype1_txt" "nkill"            "nwound"          
## [31] "propextent_txt"
```

Let's first get a sense of the (cleaned) dataset as a whole. Through the summary of the dataset (see Appendix for full summary), we can elucidate a few key insights from the data. Dropping the NA's yielded 148,627 observations on 31 variables.

```
## [1] 148627     31
```

The years range from 1970 to 2017, with a huge left skew in data, meaning there are a lot more reported incidents in the recent years, which makes sense given the development of the Internet. Additionally, there are significantly more successful than unsuccessful incidents, likely due to the fact that only incidents where the perpetrators were "out of the door" were recorded, as aforementioned.





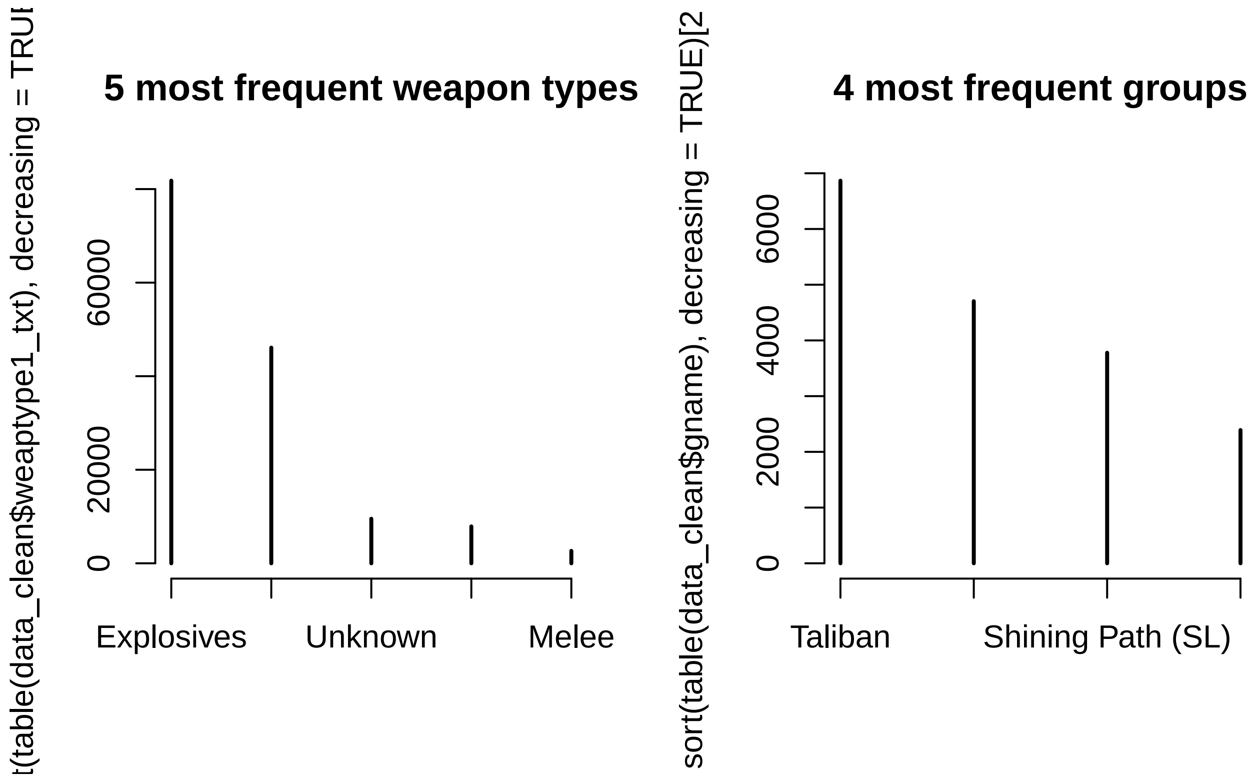
Looking at the summary of the number of number of killed and wounded (respectively), we see that on avreage there are 2.192 killed and 3.408 wounded, but the medians are both 0. The maximum of these incidents were both the tragic attack on the World Trade Center on 9/11/2001.

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    0.000    0.000    0.000    2.192    2.000 1384.000
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    0.000    0.000    0.000    3.408    2.000 8191.000
```

Looking at the types of attacks, the top 5 weapons of choice range from explosives and firearms down to melee incidents. The most frequent 4 groups (the top factor was "Unknown") were Taliban, ISIL, Shining Path, and New People's Army.

```
par(mfrow = c(1, 2))
plot(sort(table(data_clean$weaptype1_txt), decreasing=TRUE)[1:5], main = "5 most frequent weapon types"
plot(sort(table(data_clean$gname), decreasing=TRUE)[2:5], main="4 most frequent groups")
```



To view

Findings / Analysis

Future Work

nperps would be nice to use, but it is also quite sparse for some reason, we can mention that in the write up.

Conclusion

Appendix

Works Cited

[1] https://books.google.com/books?id=6qSjk2C9x6wC&pg=PA161#v=onepage&q&f=false [2] https://www.start.umd.edu/gtd/downloads/Codebook.pdf

Removed variables

```
# filter data
data_clean <- data %>% select(-c(approxdate, resolution, location, summary, alternative, alternative_tx

# remove text or IDs
data_clean <- data_clean %>% select(-c(country, region, attacktype1, targtype1, targsubtype1, natlty1,

# re-code -9 or -99 to NA
data_clean <- data_clean %>% select(-c(nkillus, nkillter, nwoundus, nwoundte, nperps, nperpcap, claimed
data_clean$vicinity[data_clean$vicinity < 0] <- NA
data_clean$doubtterr[data_clean$doubtterr < 0] <- NA
data_clean <- na.omit(data_clean)
```

Summary of cleaned data

```
##      iyear          imonth            iday          extended
##  Min.   :1970   Min.   : 0.000   Min.   : 0.00   Min.   :0.00000
##  1st Qu.:1994   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:0.00000
##  Median :2011   Median : 6.000   Median :15.00   Median :0.00000
##  Mean   :2005   Mean   : 6.468   Mean   :15.54   Mean   :0.03228
##  3rd Qu.:2014   3rd Qu.: 9.000   3rd Qu.:23.00   3rd Qu.:0.00000
##  Max.   :2017   Max.   :12.000   Max.   :31.00   Max.   :1.00000
##
##       country_txt                         region_txt
##  Iraq       :23159   Middle East & North Africa:43663
##  Pakistan   :13113   South Asia                :40537
##  Afghanistan:11844   South America             :14080
##  India      :10414   Sub-Saharan Africa        :12777
##  Colombia   : 6282   Western Europe            :11971
##  Philippines: 6001   Southeast Asia            :11021
##  (Other)    :77814   (Other)                   :14578
##              provstate              city           latitude
##  Baghdad           :  7426   Baghdad  :  7370   Min.   :-53.15
##  Balochistan       :  3563   Unknown  :  6273   1st Qu.: 12.10
##  Saladin           :  3203   Mosul    :  2112   Median : 31.66
##  Al Anbar          :  3029   Karachi  :  2028   Mean   : 24.04
##  Khyber Pakhtunkhwa:  3018   Lima     :  1883   3rd Qu.: 34.53
##  Nineveh           :  2969   Mogadishu:  1333   Max.   : 74.63
##  (Other)           :125419   (Other)  :127628
##    longitude         specificity        vicinity            crit1
##  Min.   :-157.86   Min.   :1.000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:  11.26   1st Qu.:1.000   1st Qu.:0.00000   1st Qu.:1.0000
##  Median :  44.19   Median :1.000   Median :0.00000   Median :1.0000
## Mean   :  32.18   Mean   :1.373   Mean   :0.07357   Mean   :0.9872
## 3rd Qu.:  69.42   3rd Qu.:1.000   3rd Qu.:0.00000   3rd Qu.:1.0000
## Max.   : 179.37   Max.   :4.000   Max.   :1.00000   Max.   :1.0000
##
##     crit2            crit3           doubtterr          multiple
## Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
## 1st Qu.:1.0000   1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:0.0000
## Median :1.0000   Median :1.0000   Median :0.0000   Median :0.0000
## Mean   :0.9932   Mean   :0.8774   Mean   :0.1603   Mean   :0.1414
## 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
## Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##
##    success          suicide                            attacktype1_txt
## Min.   :0.0000   Min.   :0.00000   Bombing/Explosion            :77923
## 1st Qu.:1.0000   1st Qu.:0.00000   Armed Assault                :33845
## Median :1.0000   Median :0.00000   Assassination                :15294
## Mean   :0.8778   Mean   :0.04106   Facility/Infrastructure Attack: 7865
## 3rd Qu.:1.0000   3rd Qu.:0.00000   Hostage Taking (Kidnapping)   : 6754
## Max.   :1.0000   Max.   :1.00000   Unknown                       : 5029
##                                    (Other)                       : 1917
##                    targtype1_txt
## Private Citizens & Property:34273
## Military                   :22956
## Police                     :20804
## Government (General)       :18780
## Business                   :14893
## Transportation             : 5500
## (Other)                    :31421
##                                                    targsubtype1_txt
## Unnamed Civilian/Unspecified                              :  9886
## Police Security Forces/Officers                           :  9180
##                                                           :  8491
## Military Personnel (soldiers, troops, officers, forces):  6614
## Military Unit/Patrol/Convoy                               :  6378
## Government Personnel (excluding police, military)      :  5945
## (Other)                                                   :102133
##      natlty1_txt                                            gname
## Iraq       :22738   Unknown                                     :67999
## Pakistan   :12723   Taliban                                     : 6867
## India      :10549   Islamic State of Iraq and the Levant (ISIL): 4704
## Afghanistan:10163   Shining Path (SL)                           : 3779
## Colombia   : 5995   New People's Army (NPA)                     : 2391
## Philippines: 5840   Al-Shabaab                                  : 2375
## (Other)    :80619   (Other)                                     :60512
##  guncertain1       individual           weaptype1_txt
## Min.   :0.0000   Min.   :0.000000   Explosives:81793
## 1st Qu.:0.0000   1st Qu.:0.000000   Firearms  :46106
## Median :0.0000   Median :0.000000   Unknown   : 9523
## Mean   :0.0901   Mean   :0.003398   Incendiary: 7882
## 3rd Qu.:0.0000   3rd Qu.:0.000000   Melee     : 2655
## Max.   :1.0000   Max.   :1.000000   Chemical  :  276
##                                     (Other)   :  392
##                                  weapsubtype1_txt     nkill
## Unknown Explosive Type                   :37776   Min.   :   0.000
##  Unknown Gun Type                         :27376   1st Qu.:   0.000
##                                           :12343   Median :   0.000
##  Automatic or Semi-Automatic Rifle        :12283   Mean   :   2.192
##  Vehicle                                  : 9150   3rd Qu.:   2.000
##  Projectile (rockets, mortars, RPGs, etc.): 8753   Max.   :1384.000
##  (Other)                                  :40946
##      nwound                                               propextent_txt
##  Min.   :   0.000                                                :94236
##  1st Qu.:   0.000   Catastrophic (likely >= $1 billion)          :    6
##  Median :   0.000   Major (likely >= $1 million but < $1 billion):  792
##  Mean   :   3.408   Minor (likely < $1 million)                  :38265
##  3rd Qu.:   2.000   Unknown                                      :15328
##  Max.   :8191.000
##
```