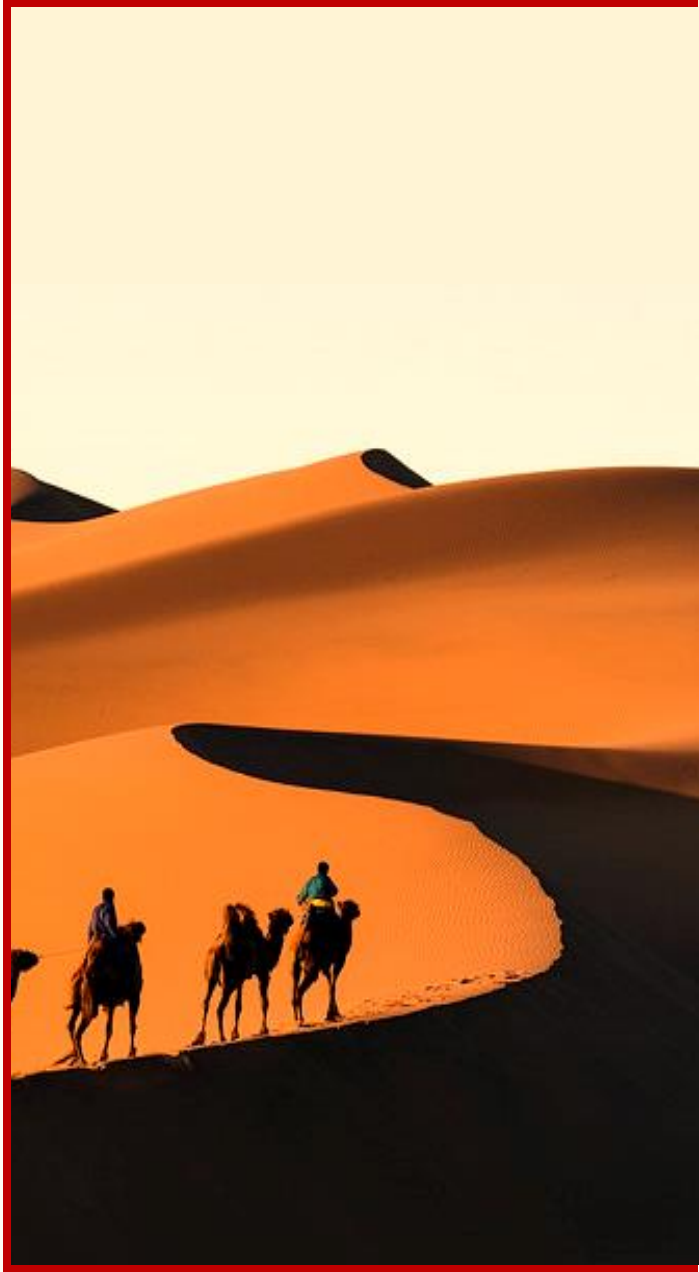
Global Terrorism Data Analysis



DS 5110

Yi-Tang Chou

Manaswini Nagaraj

Juhi Paliwal

Sushma Suresh

Introduction

Summary

Global Terrorism Dataset (GTD) is an open-source dataset that documents terrorist events around the world. GTD is a collaborative work of Pinkerton Global Intelligence Services, Study of Violent Group, Central for Terrorism and Intelligence Services and National Consortium for the Study of Terrorism and Responses to Terrorism. GTD defines a terrorist attack as "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." [1]

GTD contains information on terrorist attacks from 1970 until 2017, with 181691 records and 135 variables. “eventid” identifies each observation uniquely, whereas columns like “latitude” and “longitude” values of the location where the attack has taken place. Three criteria columns are provided, and at least two criteria must be fulfilled for an incident to be recorded. The three criteria are: (1) the incident was performed to attend social, economic, religious or political goals, (2) the act was trying to send messages to a larger audience, (3) the incident was outside the context of legitimate warfare activities. The success column identified the attacks that were attempted but ultimately failed. The "nkills" and "nwounds" columns identified the number of victims from an attack and were also used to generate a new computed column "ncasualties". At first glance, the size of the dataset seems quite substantial, however, upon further analysis, it can be observed that most of the columns are either categorical columns computed into binary numerical columns or text columns defining certain categorical columns.

Exploratory Data Analysis

For the first EDA, a worldwide heatmap was constructed, which provided an insight how geography may affect attack attempts. It may be concluded that the middle east and southern Asia are most prone to terrorist attacks. Furthermore, the top 5 countries with the highest attack counts are: Iraq, Afghanistan, Pakistan, India, and Syria.

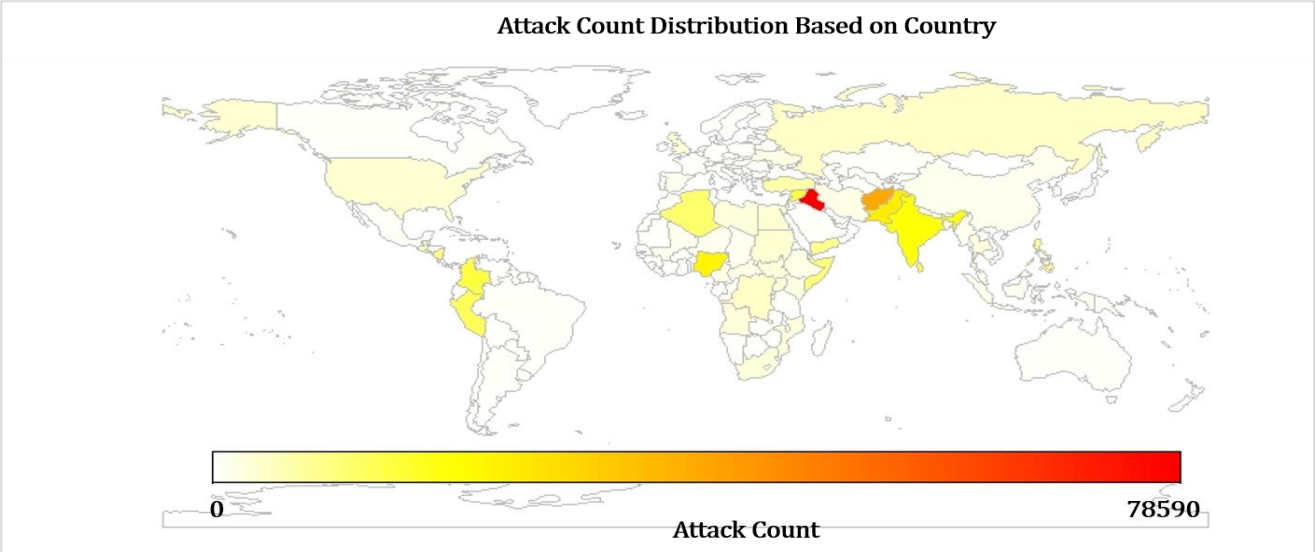


Figure 1 Attack Count Distribution Based on Country

After exploring the relationship between geography and attack count, exploratory analysis was also done on time, where attack count was plotted according to each year. As shown in the figure below:

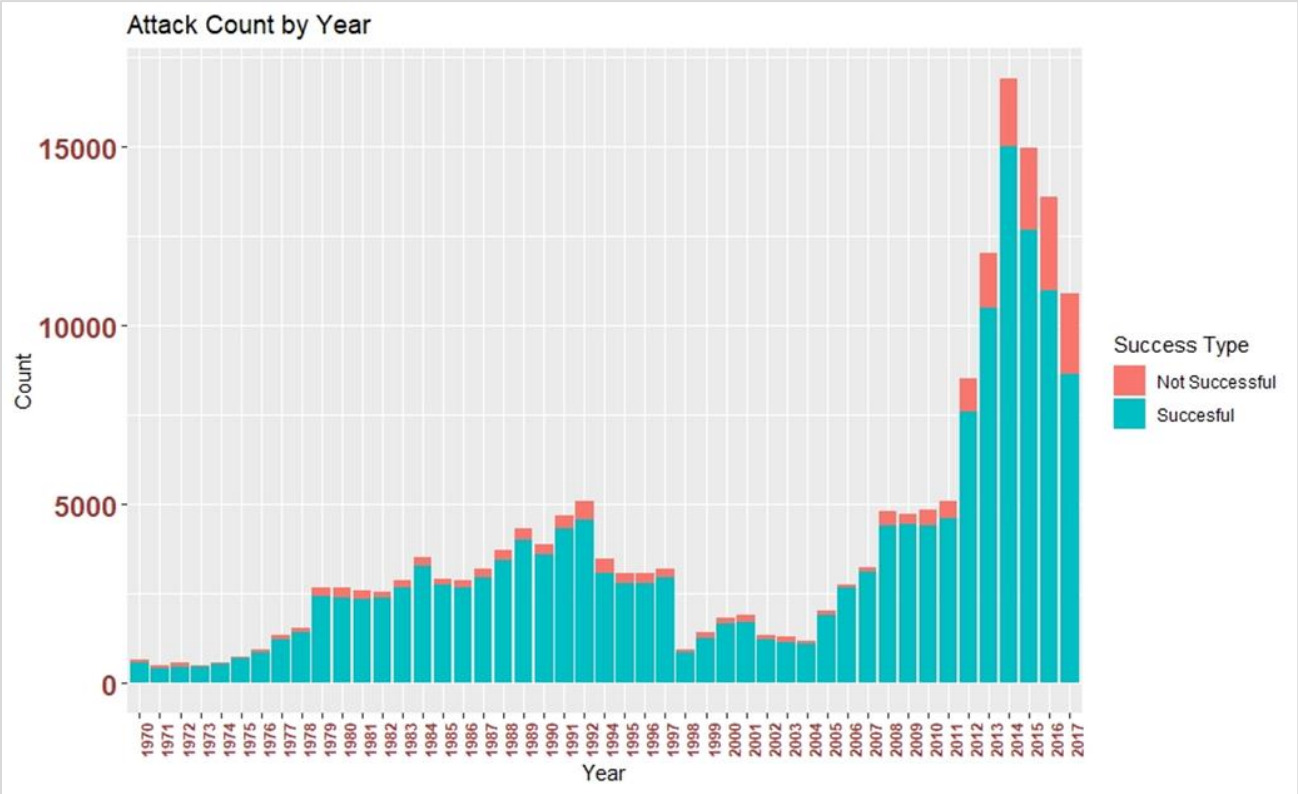


Figure 2 Attack Count Distribution Based on Year

It was through this analysis that the team realized a major feature of GTD. Due to the unpredictable nature of a terrorist attack, most of the events documented had a success outcome of 1 (successful), which is shown in the figure above, meaning the dataset is very skewed towards one type of outcome. This analysis helped the team anticipate certain roadblocks in terms of prediction tool construction that will need to be overcome.

Objective

This project had 3 major objectives. First, to predict if an attack will have victim presence, utilizing Random Forest classifiers to predict a binary result. Second, if the output of the first model is a 1 (yes), the second model is utilized, to predict the number of casualties with a XGBoost algorithm. Third, the last model calls in the predicted casualties from model 2 to predict if a attack will be successful or not.

Data Cleaning

The data cleaning process took two major parts. Firstly, the numerical columns were cleaned by replacing cells that were empty or contain “NA” with the mean of the entire column. Categorical columns were cleaned by first replace “NA” cells with the number -9. This is the reason because -9 was the number that represented Null in the code book for all categorical values, however, the dataset wasn’t cleaned properly, and there were both -9’s and null cells, so in essence, the team completed the cleaning process by unifying all null values into -9’s

Methodology & Results

Model 1: Random Forest Classifiers

Just as the name specifies, the random forest consists of many numbers of decision trees that work in an ensemble. In a random forest, each tree returns a prediction of a class and the class with the most predictions becomes the final prediction of the model.

This classifier predicts whether or not there are any victims after a terror attack has taken place. For analysis, the dependent variable is a column termed 'ncasualities' which was created after adding the column 'nkills', of people killed, and 'nwound', people wounded after every terror incident. Since result of the classification is yes or no, one hot encoding techniques were used to categorise the numbers in this column to '1' if the sum was greater than 0(implying that there were victims after the attack) and '0' if the sum was 0(implying that the attack resulted in no victims).

After computing a new column, feature selection was performed on 135 variables in the dataset. A feature selection technique called Extra Tree Classifier was used to quantify the strengths and narrow down on the independent variables.

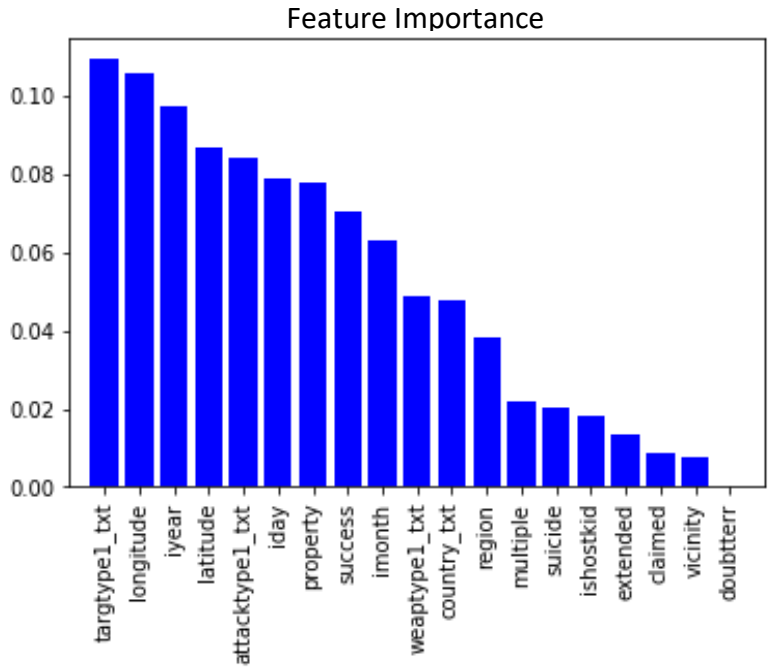


Figure 3 Feature contribution towards prediction

Extra Tree Classifier is an ensemble learning model that is based on Decision Trees. It randomizes certain decisions and subsets the data minimize overfitting of the model. It's working is very similar to random forest with the key difference being that nodes are split randomly as opposed to the best fits in a traditional random forest.

Figure 3 depicts the significance of each feature in predicting whether there will be victims or not. There seems to be a dominant presence of Spatio Temporal variables. To prevent the model from overfitting, only those features with an accuracy of 0.06 and higher were consiered. The variable 'iyear' has been removed as the event is not-recurring.

Model 1: Results

The dataset has been split in the ratio of 70:30 for training and testing respectively. Upon testing, the model achieves an accuracy of 86%. Figures 3 depicts the ROC curve generated. In order to ensure better validation, we also performed Cross 20-Fold Validation which resulted in an accuracy of 66%.

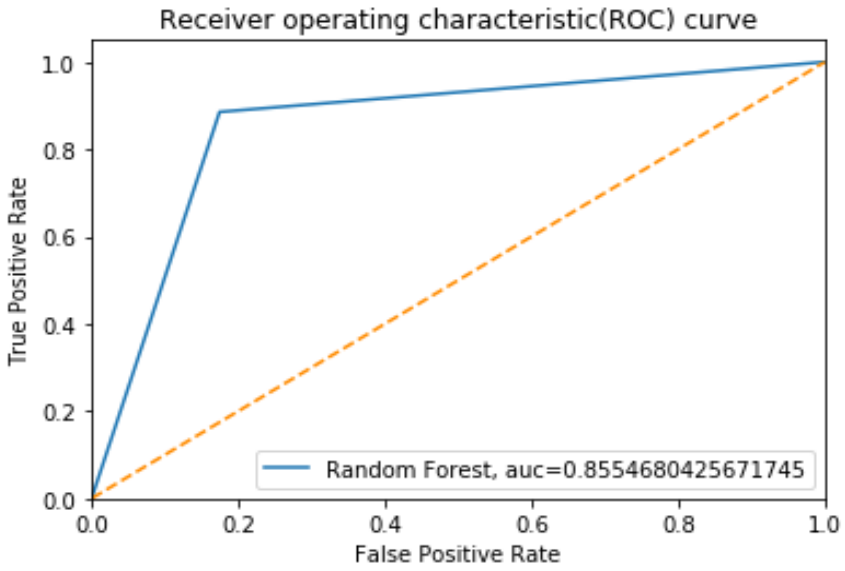


Figure 4 ROC Curve for Model 1

Model 2: XGBoost

XGBoost stands for eXtreme Gradient Boosting, it is an implementation of gradient boosting machines. It is an open-source software library that belongs to a collection of tools under the Distributed Machine Learning Community (DMLC).

XGBoost is a decision tree-based ensemble machine learning algorithm that uses a gradient boosting framework. Boosting is an ensemble technique that corrects the errors made by previous models while building the new models, it is called gradient boosting when the boosting technique uses gradient descent algorithm to minimize the loss that occurs while building a new model during each iteration. The models built are sequentially added until no further improvement is seen to obtain one final model. [2]

The tree-based approach for XGBoost first builds a decision tree with weak learners, then a new model is built to predict the residuals or errors of the previous model. This process is done stage-wise and the final prediction model is in the form of an ensemble of weak prediction models.

Figure 5 Tree-based Approach of XGBoost

Model 2: Results

After the initial training process, it was noted that the model can be improved by tuning the parameters, hence a 10 fold cross-validation was used to tune the parameters such as nrounds (number of rounds), max_depth (maximum depth of the decision tree built), gamma (Lagrangian multiplier), alpha (L1- regularization, to avoid overfitting), eta (learning rate) and min_child_weight (minimum number of instances in each node).

After the training, the model achieved a final train RMSE of 16.849 and test RMSE of 19.290 at 200 rounds and a max_depth of 25, with the learning rate at 0.05 along with an L1 regularization applied.

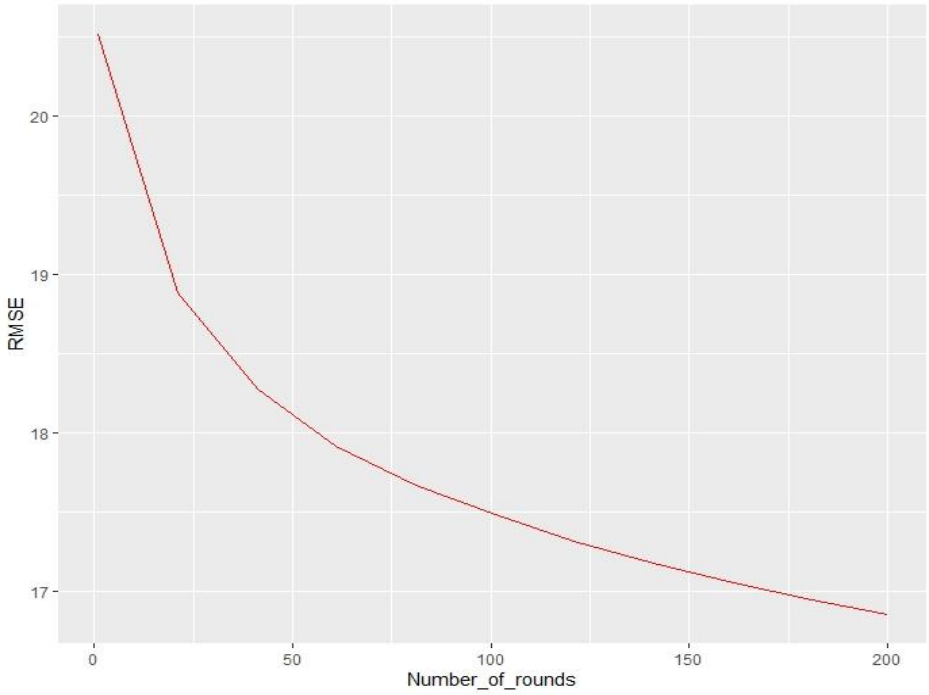


Figure 6 RMSE curve depicting the decrease in train RMSE as the number of rounds is increased.

Model 3: Naïve Bayes' Classifer

Naïve Bayes' Classifier was the final model utilized for the 3-step prediction model. Utilizing the computed column, "casualties" along with the rest of the subset data, the model will be able to predict the likely hood of success of a potential attack. This classifying method was chosen based a few of its strength's over other methods. First, Naïve Bayes' algorithm generally performs well under a small set of variables, and due to the nature of a terrorist attack, data availability before an event maybe very limited, and the algorithm must perform with the information available to it. Secondly, this classifier also provides a priori probability, providing the exact probability of a successful or non-successful event happening. This is important because in the case of an event that has a successful event classified with a priori probability of 49.9% and that of a non-successful event of 50.1%, law-enforcement will be able be aware that because a non-successful event is only 0.5% more likely to happen, they should also be fully prepared for a successful attack.

Model 3: Results

As mentioned in the summary section, the data subset utilized by this model is very skewed towards successful events, so the team anticipated the function will only perform well in predicting successful events. The following is the confusion matrix of the classifier:

Trained Data Sensitivity: 35.5 Specificity: 85.3 Validation Data Sensitivity: 34.1 Specificity: 85.3		Predicted	
		0	1
Actual	0	8291 5478	43120 28758
	1	3750 2540	53853 35901

The skewed performance of the model is expected. If more available data on non-successful attacks, the classifier would be able to pick up the features of 0's and hence increase the sensitivity quite substantially. Another feature of this model is the apriori probability mentioned before. The table on the right demonstrates that of two specific events. Both events are examples where the probabilities of 0 and 1's are so close that both outcomes should be prepared for, meaning if an attack like eventid 197005060001 were to happen, anti-terrorist agencies should be prepared for a successful attack even though the prediction is classified as non-successful

Eventid	Apriori Probability	
	1	0
197004110002	0.4979589211	0.5020411
197005060001	0.50124304	0.4987570

Discussion

Future Works

In the case of this project's future development, the first major direction that must be ventured is to combine all 3 models into one singular environment. This project aimed to prepare society against a terrorist attack in a step by step manner. As mentioned in the objectives, Model 1 will first predict if there will be casualties, if the prediction is yes, model 2 is used to predict how many casualties will be present, lastly, model 3 is used to predict if this attack will reach it's final goal and succeed at it's target by inputting the casualty prediction into the predicting variables . Currently, all 3 models are built, however, the hierarchy of each model has not been established. This hierarchy is essential because in real-life application, when an attack is anticipated to happen, casualty data won't be available, which will deem model 3 useless.

A second major direction that needs to be look into is the data of this prediction model. This model is attempting to perform prediction analysis on post event data. In a real-world scenario, post event data would not be available. Currently, this team suggests some type of clustering method that can cluster potential attacks, provided that there is already certain information known about it, based on the likeliness of historical events.

The third major improvement that can be made to this model is the data partitioning process. The team suggests a process that can help models pick up features of minorities of the data such as events classified as not successful.

The final improvement is specific to the Naïve Bayes' classifier. The team has attempted to overcome the skewedness of the data utilizing Laplace smoothing or selective data partitioning, where instead of

the traditional 0.6 0.4 partitioning on the entire data set, non-successful were intentionally up-sampled in hopes of the function to capture more of it's features. However, the accuracy rate came down to around 47%, with high number of False negatives. This specific variation of the model was deemed inappropriate by the team as dealing matters like a terrorist attack, it is better to anticipate it then not have it happening, instead of the other way around. Furthermore, the team also attempted to utilize different compositions of subset to predict success, especially on the nkill, nwound, casualties columns. It was found that excluding columns nwound and casualties resulted to a model that performed better on predicting non-successful attacks, concluding the properties of the subset that includes only nkill should be explored.

Statement of Contribution

The completion process of this project was extremely collaborative. From the inception of the project where topics were evaluated for the project, to the end where the power-point presentation and report were being finalized, each and every member participated equally. The team truly believes that it was due to the equal efforts of each and every member that 3 models were able to be generated, and even under the scope of time of the project, the team was just a few steps away from achieving the objective of synchronizing all 3 models.


Yi-Tang Chou
chou.yi@husky.neu.edu


Manaswini Nagaraj
ngaraj.m@husky.neu.edu


Juhi Paliwal
paliwal.j@husky.neu.edu


Sushma Suresh
kalkunte.s@husky.neu.edu

References

[1] Codebook: https://start.umd.edu/gtd/downloads/Codebook.pdf

[2] XGBoost: https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/

[3] Global Terrorism Dataset: https://www.kaggle.com/START-UMD/gtd/data# =

Appendix:

```
# A tibble: 181,691 x 35
   eventid iyear imonth  iday country country_txt region region_txt city
     <dbl> <dbl>  <dbl> <dbl>   <dbl> <chr>        <dbl> <chr>      <chr>
 1 1.97e11  1970      7     2      58 Dominican ~      2 Central A~ Sant~
 2 1.97e11  1970      0     0     130 Mexico           1 North Ame~ Mexi~
 3 1.97e11  1970      1     0     160 Philippines      5 Southeast~ Unkn~
 4 1.97e11  1970      1     0      78 Greece           8 Western E~ Athe~
 5 1.97e11  1970      1     0     101 Japan            4 East Asia  Fuko~
 6 1.97e11  1970      1     1     217 United Sta~      1 North Ame~ Cairo
 7 1.97e11  1970      1     2     218 Uruguay          3 South Ame~ Mont~
 8 1.97e11  1970      1     2     217 United Sta~      1 North Ame~ Oakl~
 9 1.97e11  1970      1     2     217 United Sta~      1 North Ame~ Madi~
10 1.97e11  1970      1     3     217 United Sta~      1 North Ame~ Madi~
# ... with 181,681 more rows, and 26 more variables: provstate <chr>,
#   latitude <dbl>, longitude <dbl>, location <chr>, summary <chr>,
#   crit1 <dbl>, crit2 <dbl>, crit3 <dbl>, doubtterr <dbl>,
#   suicide <dbl>, attacktype1 <dbl>, attacktype1_txt <chr>,
#   targtype1 <chr>, targtype1_txt <chr>, gname <chr>, weaptype1 <dbl>,
#   weaptype1_txt <chr>, nkill <dbl>, nwound <dbl>, motive <chr>,
#   vicinity <dbl>, success <dbl>, individual <dbl>, nperps <dbl>,
#   property <dbl>, casualties <dbl>
```

Figure 7. Tidy Dataset

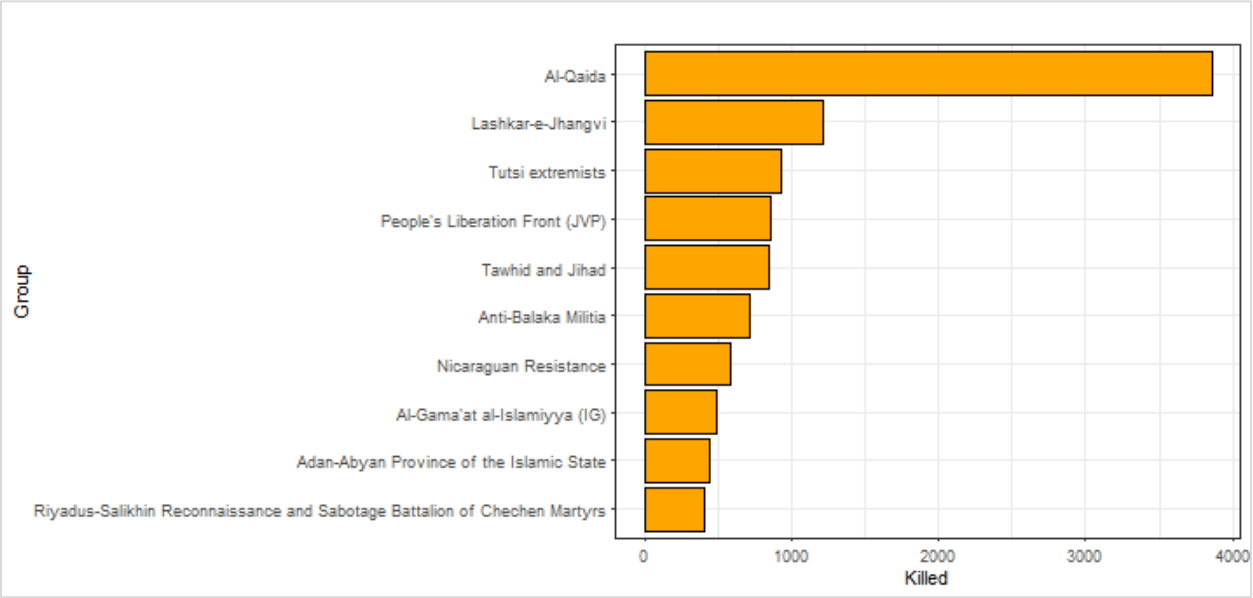


Figure 8 Top 10 Groups with highest number of attacks

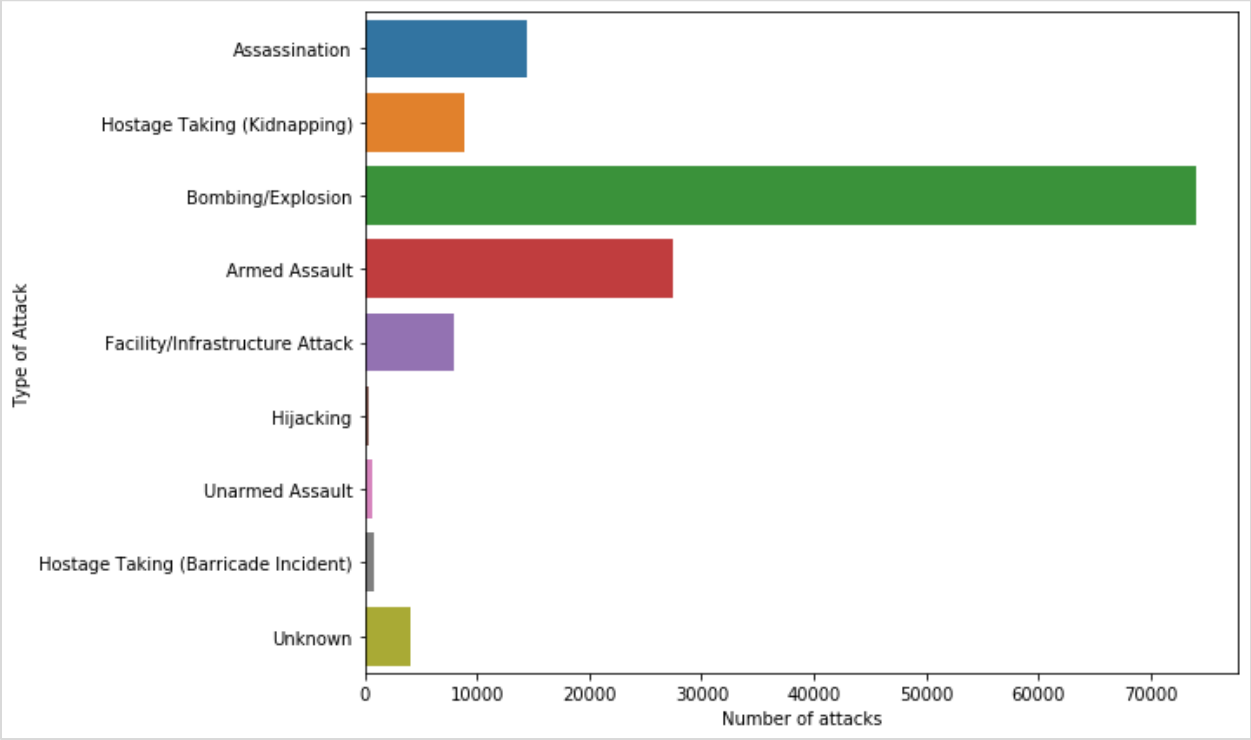


Figure 9. Count of type of terrorist attacks

Figure 10. Word Cloud of Summary Variables