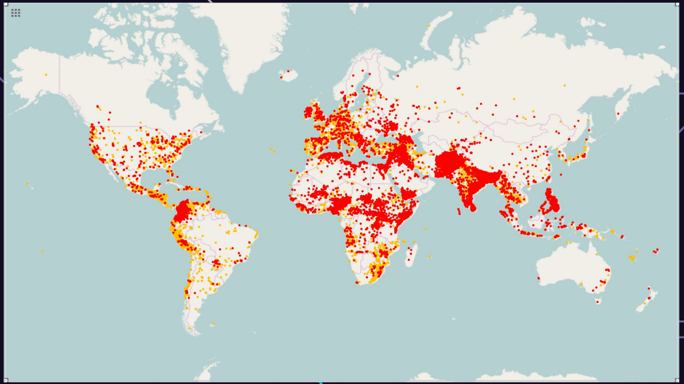
IST 707 Group Project

Kent Roller, Kevin Hansen, Ben Tisinger

Data Set



Global Terroism Database - Kaggle

This is a database of over 170+ global terrorist attacks that occured during 1970-2016. Many of the columns have been reworked and altered to create a streamlined version. There are over 58 Columns of information.

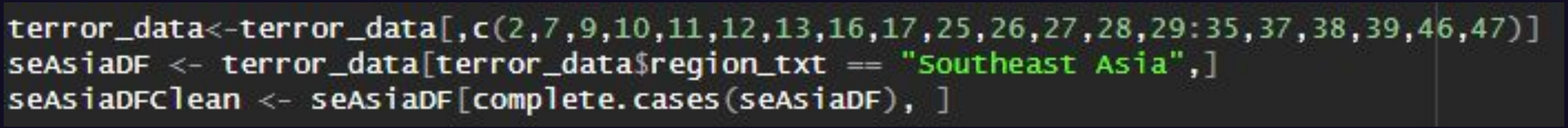
Main Table:

- Global_Terrorism_MDB

Columns Include:

- EventID
- YearID
- Country
- Latitude
- Longitude
- Summary
- AttackType
- Motive
- WeapType
- CompClaim

Code Clean Steps



```
terror_data<-terror_data[,c(2,7,9,10,11,12,13,16,17,25,26,27,28,29:35,37,38,39,46,47)]
seAsiaDF <- terror_data[terror_data$region_txt == "Southeast Asia",]
seAsiaDFClean <- seAsiaDF[complete.cases(seAsiaDF), ]
```

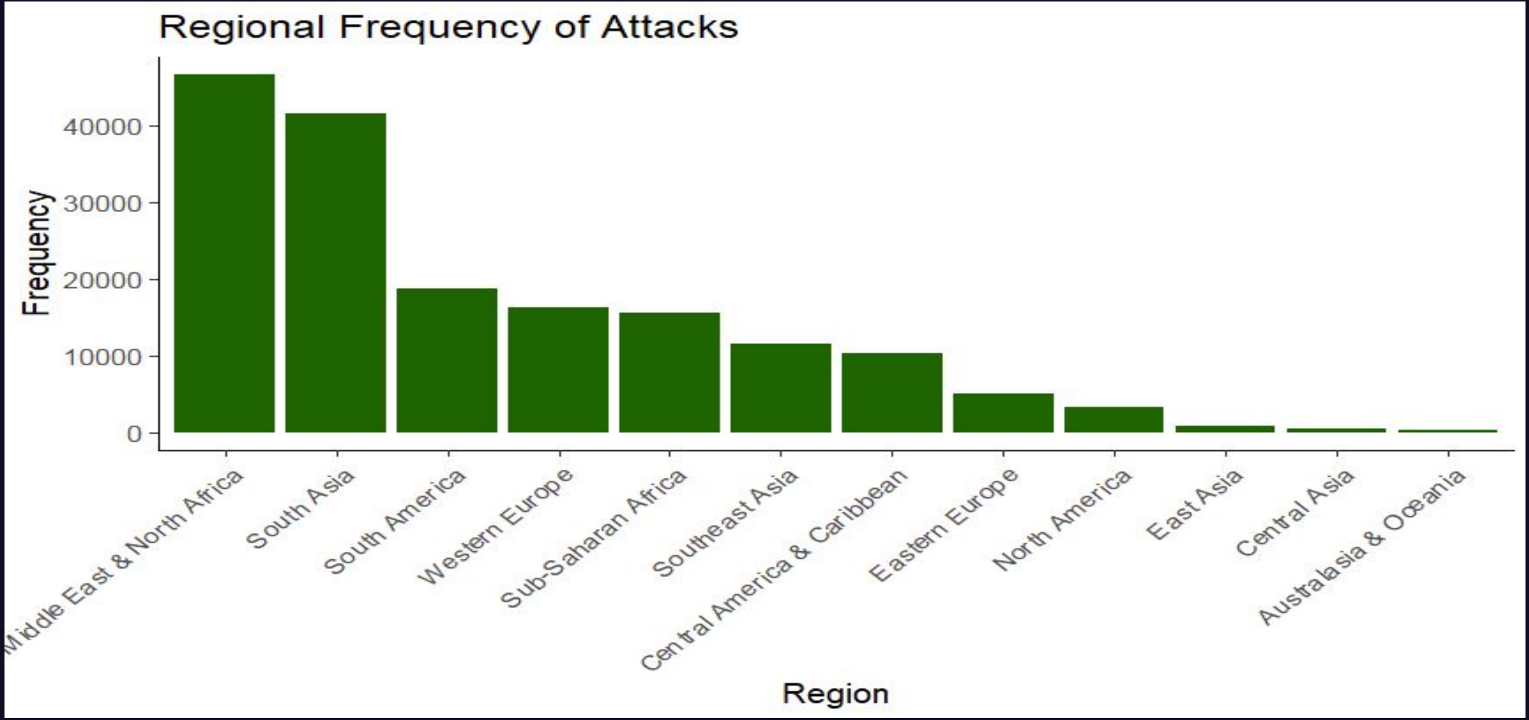
Steps Included:

- The initial terrorism dataset included many variables not required for the analysis intended to be performed.

- The revised terror dataset cuts the original 58 variable dataset down to 29.

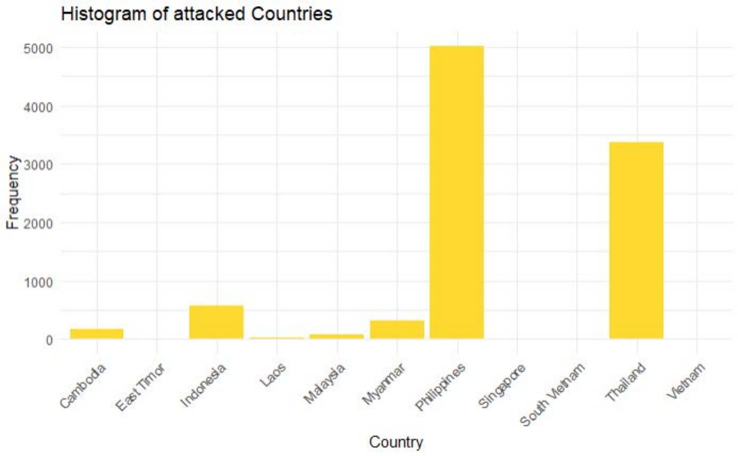
- We only wanted to analyze the attacks that took place in the Southeast Asia Region

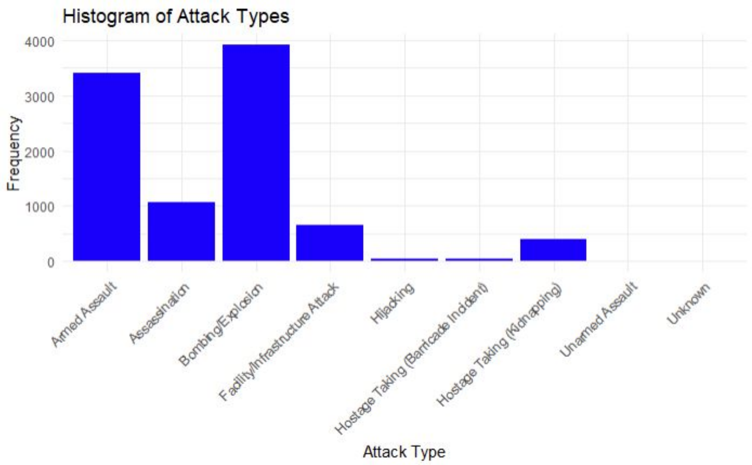
Exploratory Analysis

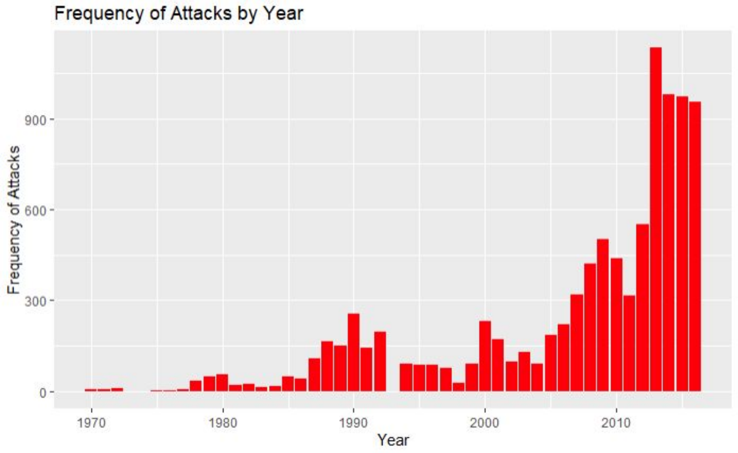


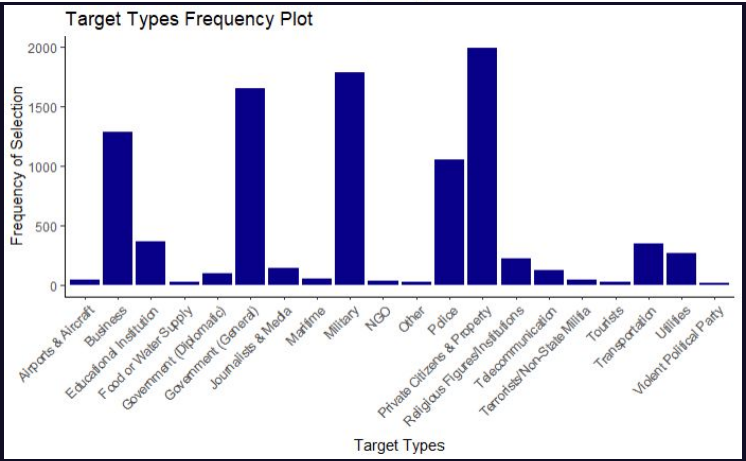
The histogram displays the frequency of attacks subdivided by region. The Middle East & North Africa being the area with the highest frequency, however, we will be focusing on the **Southeast Asia Region**

Exploratory Analysis - Southeast Asia

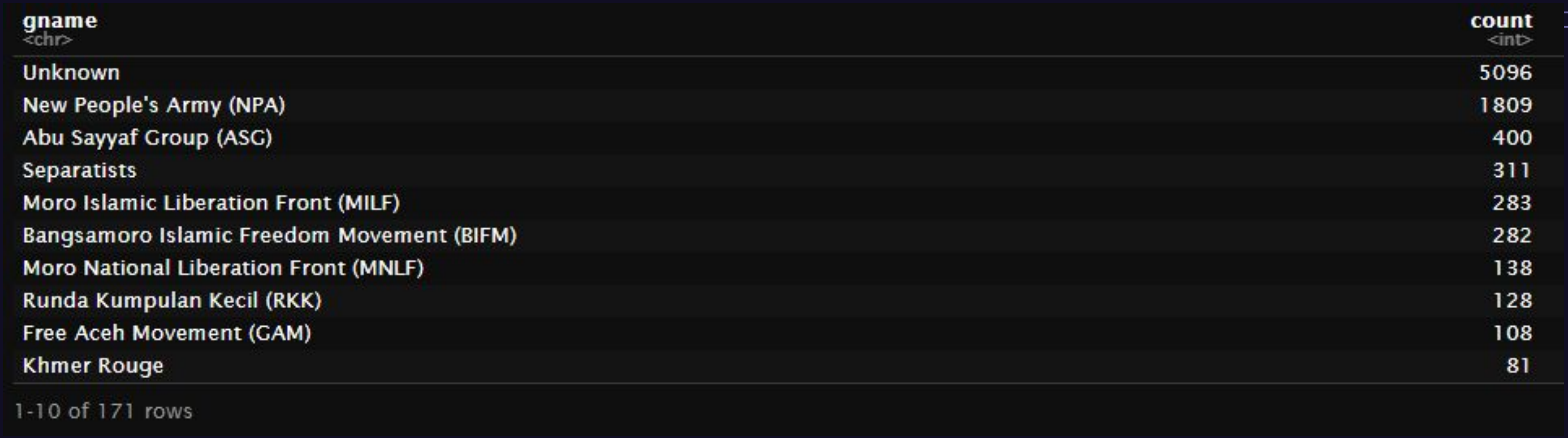

Histogram of attacked Countries


Histogram of Attack Types


Frequency of Attacks by Year


Target Types Frequency Plot

Terror Organization Prediction

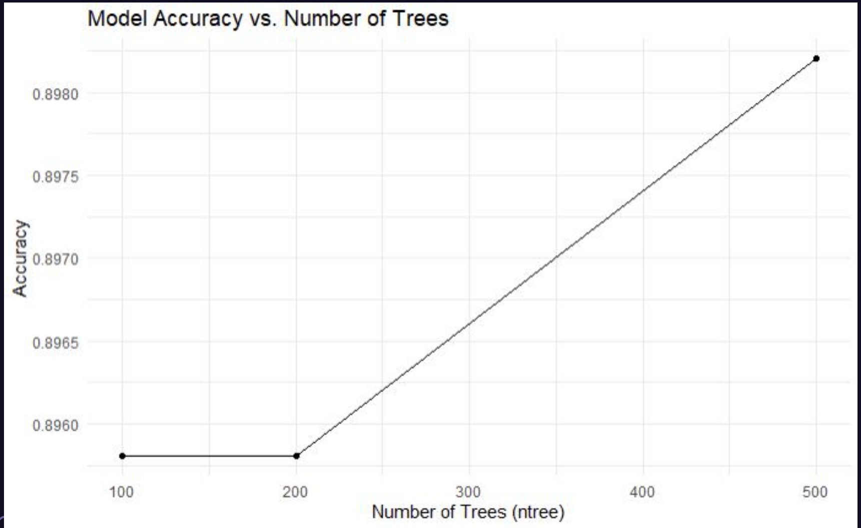


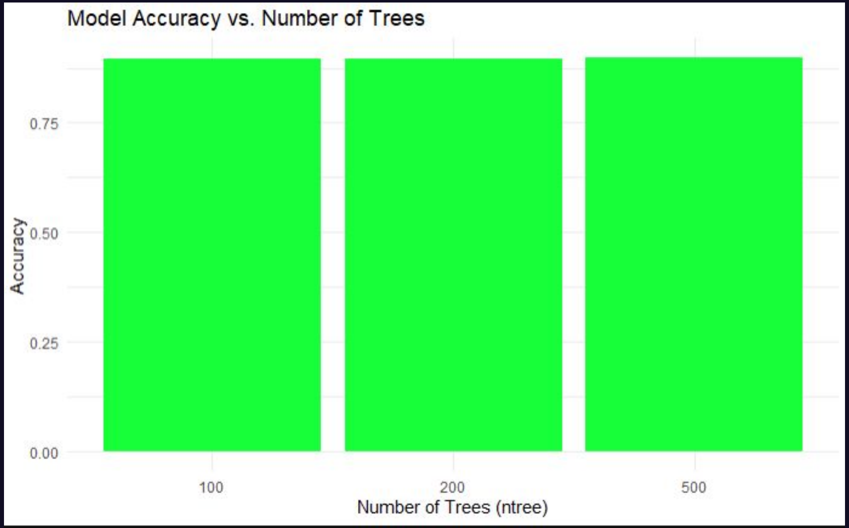
gname	count
<chr>	<int>
Unknown	5096
New People's Army (NPA)	1809
Abu Sayyaf Group (ASG)	400
Separatists	311
Moro Islamic Liberation Front (MILF)	283
Bangsamoro Islamic Freedom Movement (BIFM)	282
Moro National Liberation Front (MNLF)	138
Runda Kumpulan Kecil (RKK)	128
Free Aceh Movement (GAM)	108
Khmer Rouge	81

1-10 of 171 rows

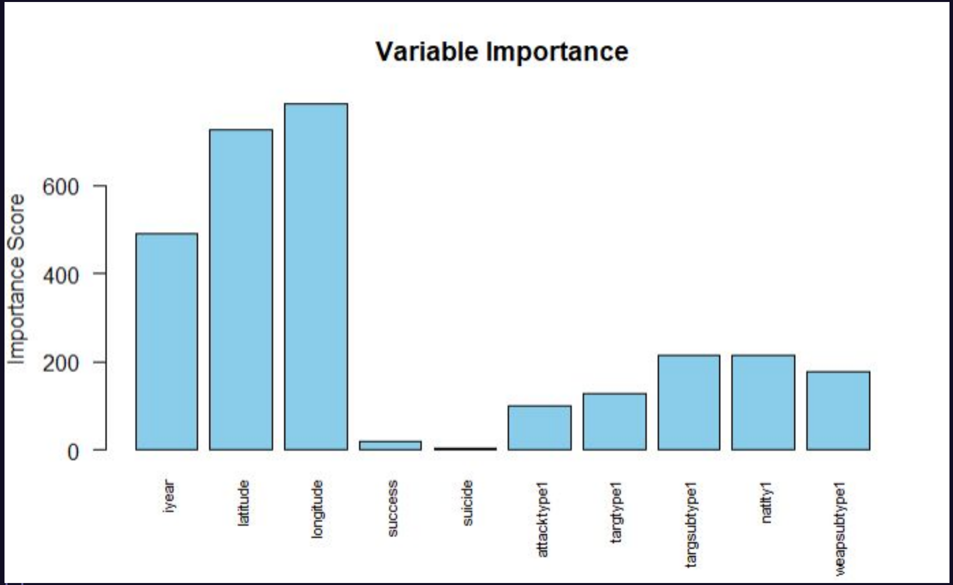
- Table displaying number of attacks that were attributed to each “terrorist organization.”
- 5096 attacks where it is unknown who committed them
- Can we develop a highly accurate machine learning model that can tell us who committed the attacks?

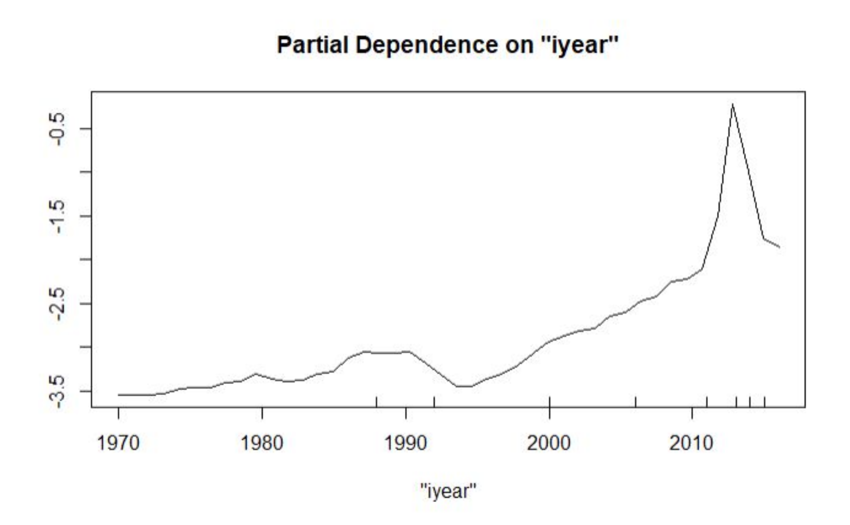
Random Forest Model - 3 Models (ntree = 100, 200, 500)



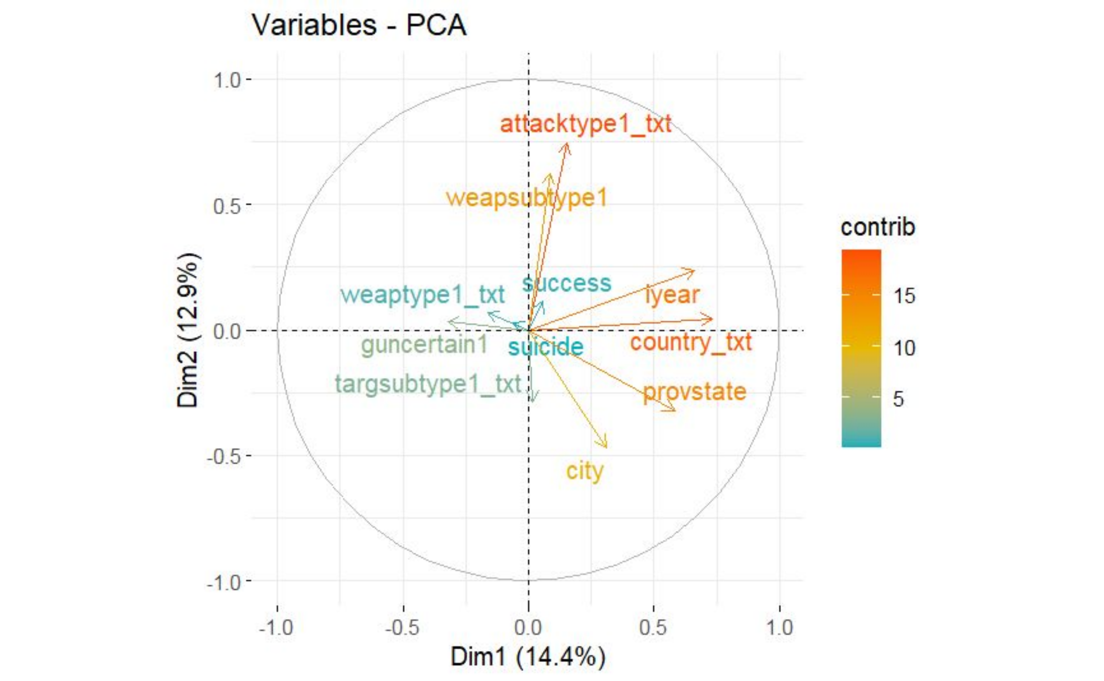


▶ Random Forest: Output Influence





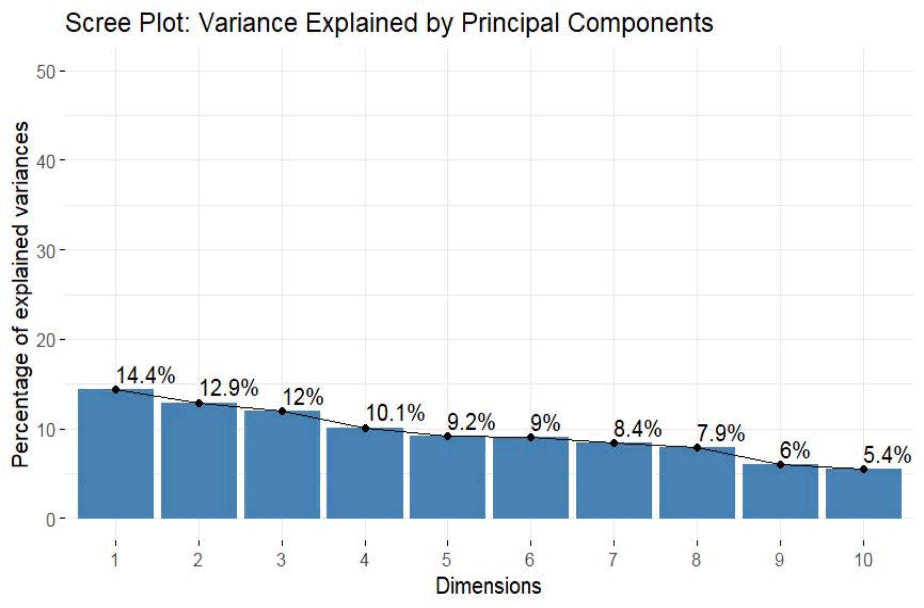
PCA and Naive Bayes



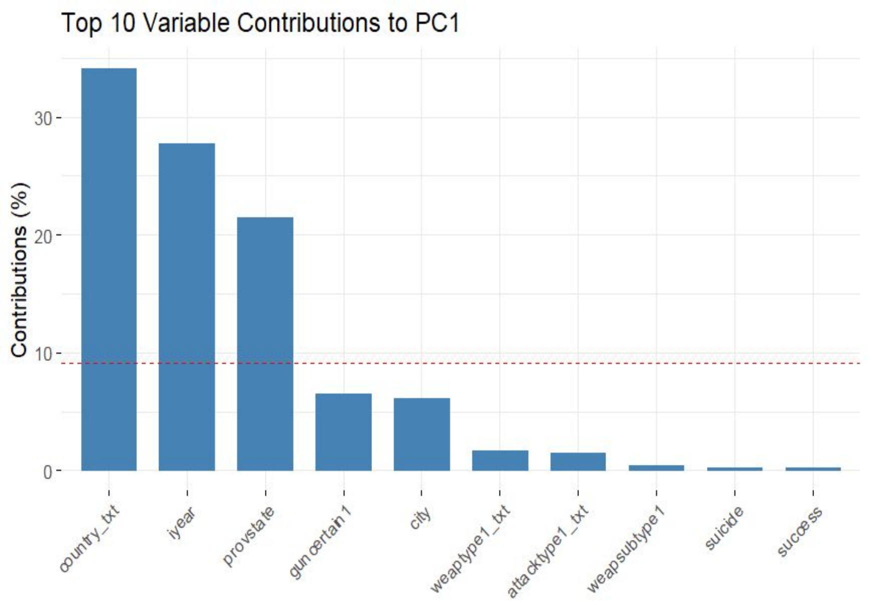
Here the visualization of the PCA can be seen where the variables are color coded based on importance (percentage of explained variance)

PCA and Naive Bayes II

The scree plot for the PCA and a barplot showing the contribution by component are shown below.



Scree Plot: Variance Explained by Principal Components



Top 10 Variable Contributions to PC1

Naive Bayes Model Performance

The Naive Bayes model was run 4 times, adjusting the number of principal components used in the analysis each time in order to find optimal number.

Model 1: PCA=default	Model 2: PCA=11	Model 3: PCA=6	Model 4: PCA=8
Accuracy: 61.20%	Accuracy: 64.18%	Accuracy: 61.20%	Accuracy: 66.17%

▶ SVM pre-binarization and post-binarization

SVM was used next in order to establish if it would offer better performance with the data.

The initial accuracy obtained via this model was: 65.90%.

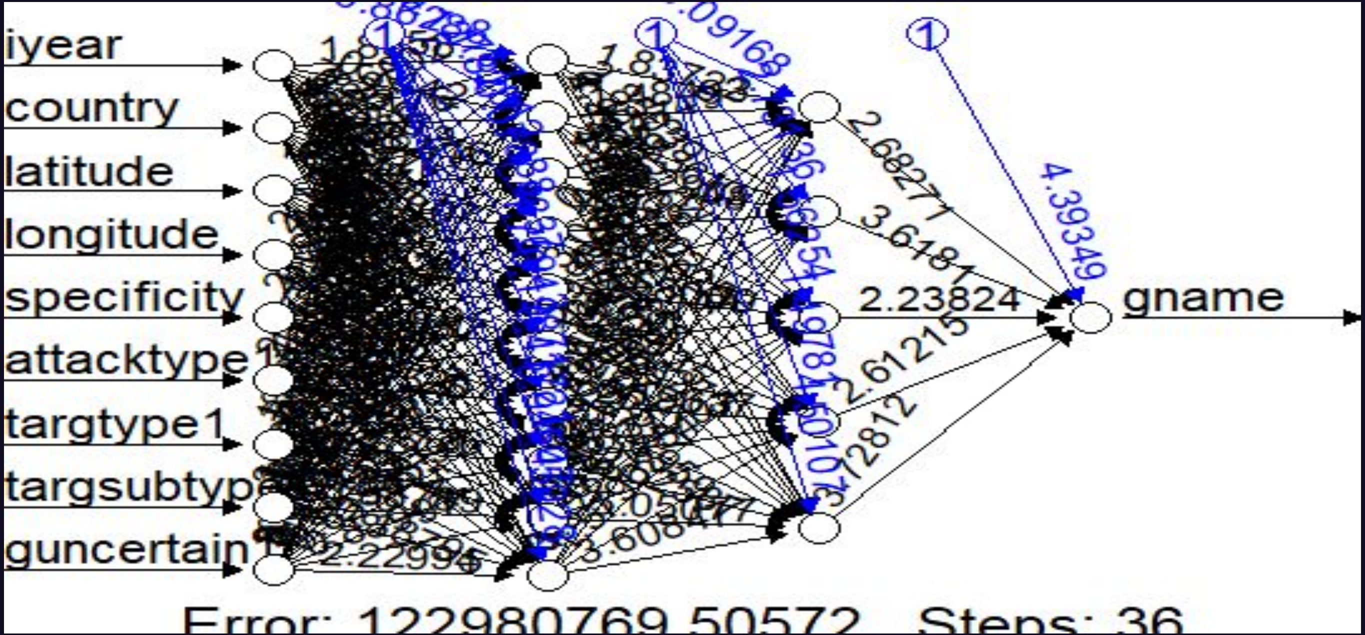
This accuracy is not a substantial improvement over the accuracy obtained with Naive Bayes.

In an attempt to remedy the poor performance the dataset was binarized.

Unfortunately, the accuracy of the model was further degraded using this method with accuracies ranging from 43-45%.

Attempting to Use Neural Networks

Given the poor performance of the Naive Bayes model and SVM, we decide instead to try to use Neural Networks to tackle the problem.



Neural Networks Outcome

Unfortunately errors were encountered during the normalization process.

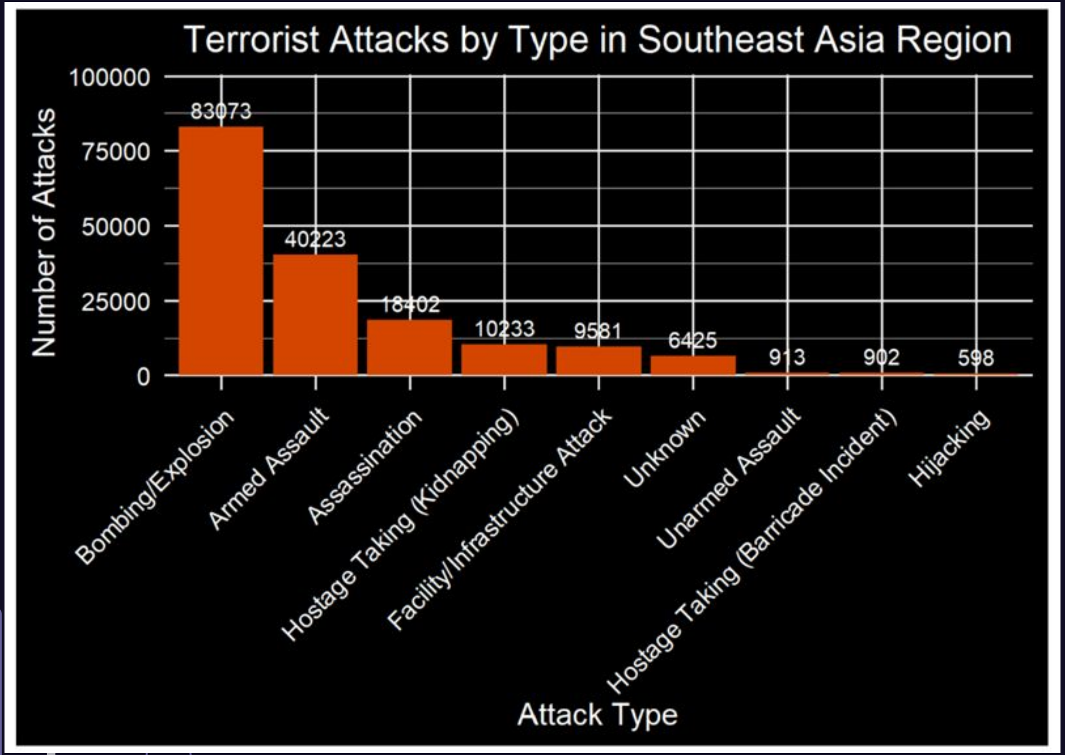
In the interest of time all factor variables were removed from the dataset except the response and the model was rerun.

Initially we had success with the model with output being generated and the model appeared to work.

However, upon trying to determine the accuracy of the model ran into factor conversion errors and then mismatched column types.

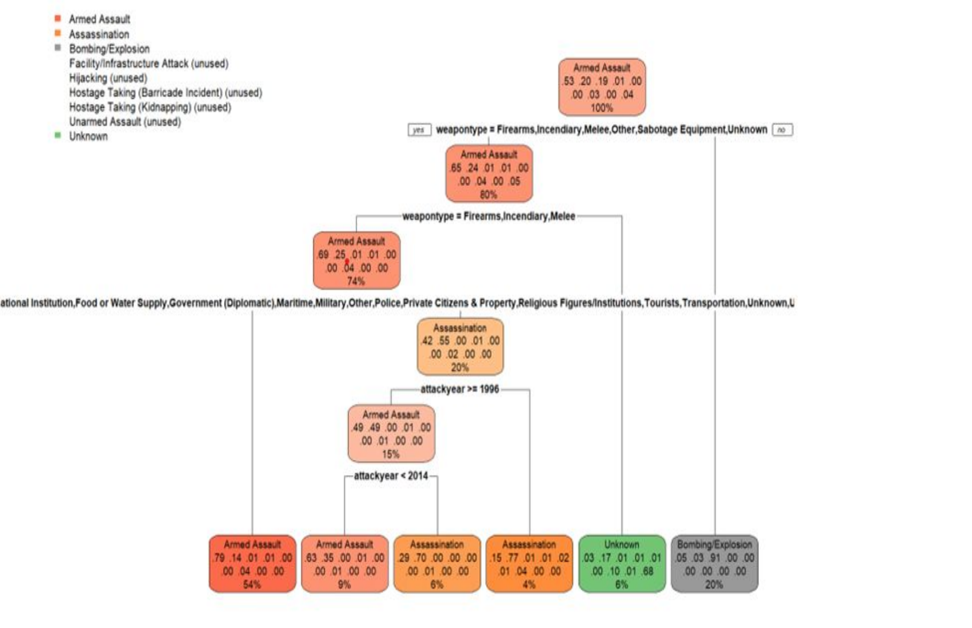
This shows some promise of being a viable alternative to Random Forest if more time was allotted to circumvent the errors encountered.

Preliminary Stats - Before Decision Trees



- Preliminary Stats for the Southeast Asia Region highlight a high level of bombing/explosion and armed assaults
- Hijacking and Hostage are very low in this Region

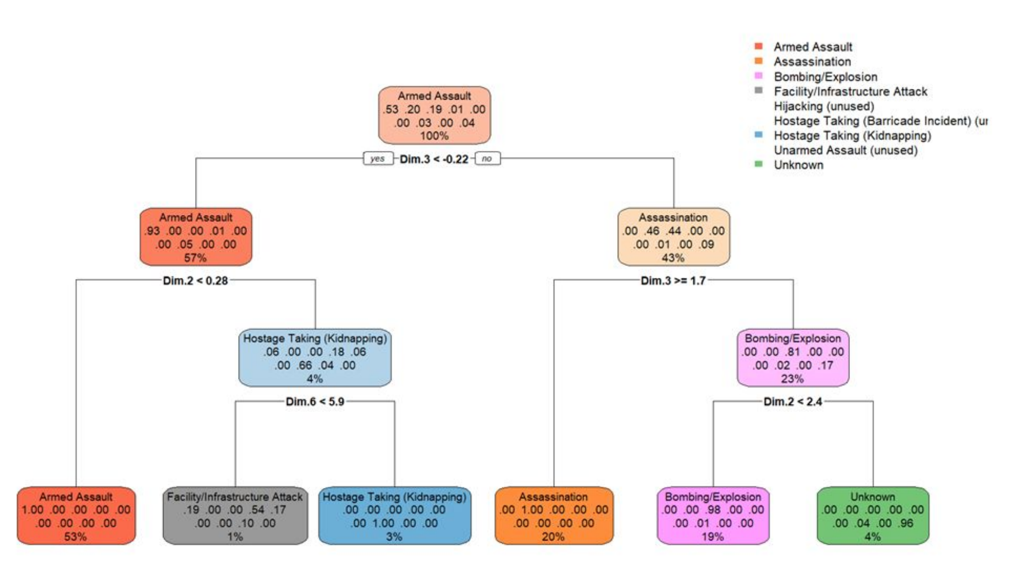
Decision Trees



Armed Assault
Assassination
Bombing/Explosion
Facility/Infrastructure Attack (unused)
Hijacking (unused)
Hostage Taking (Barricade Incident) (unused)
Hostage Taking (Kidnapping) (unused)
Unarmed Assault (unused)
Unknown
Armed Assault .53 .20 .19 .01 .00 .00 .03 .00 .04 100%
yes
weapontype = Firearms,Incendiary,Melee,Other,Sabotage Equipment,Unknown
no
Armed Assault .65 .24 .01 .01 .00 .00 .04 .00 .05 80%
weapontype = Firearms,Incendiary,Melee
Armed Assault .69 .25 .01 .01 .00 .00 .04 .00 .00 74%
ational Institution,Food or Water Supply,Government (Diplomatic),Maritime,Military,Other,Police,Private Citizens & Property,Religious Figures/Institutions,Tourists,Transportation,Unknown,U
Assassination .42 .55 .00 .01 .00 .00 .02 .00 .00 20%
attackyear >= 1996
Armed Assault .49 .49 .00 .01 .00 .00 .01 .00 .00 15%
attackyear < 2014
Armed Assault .79 .14 .01 .01 .00 .00 .04 .00 .00 54%
Armed Assault .63 .35 .00 .01 .00 .00 .01 .00 .00 9%
Assassination .29 .70 .00 .00 .00 .00 .01 .00 .00 6%
Assassination .15 .77 .01 .01 .02 .01 .04 .00 .00 4%
Unknown .03 .17 .01 .01 .01 .00 .10 .01 .68 6%
Bombing/Explosion .05 .03 .91 .00 .00 .00 .00 .00 .00 20%

- Filtered for Region #5 - Southeast Asia
- Removed Unknown and NA Values
- Removed No Casualties
- Used Values such as Attack Type, Target Type, Year and Success to help Predict which weapons would be of Primary Use
- We learned that Armed Assaults are most occurring and after the year 2014 - 54% of all attacks will be Armed Assaults
- Confusion Matrix - Accuracy of 77%
- Kappa 0.63

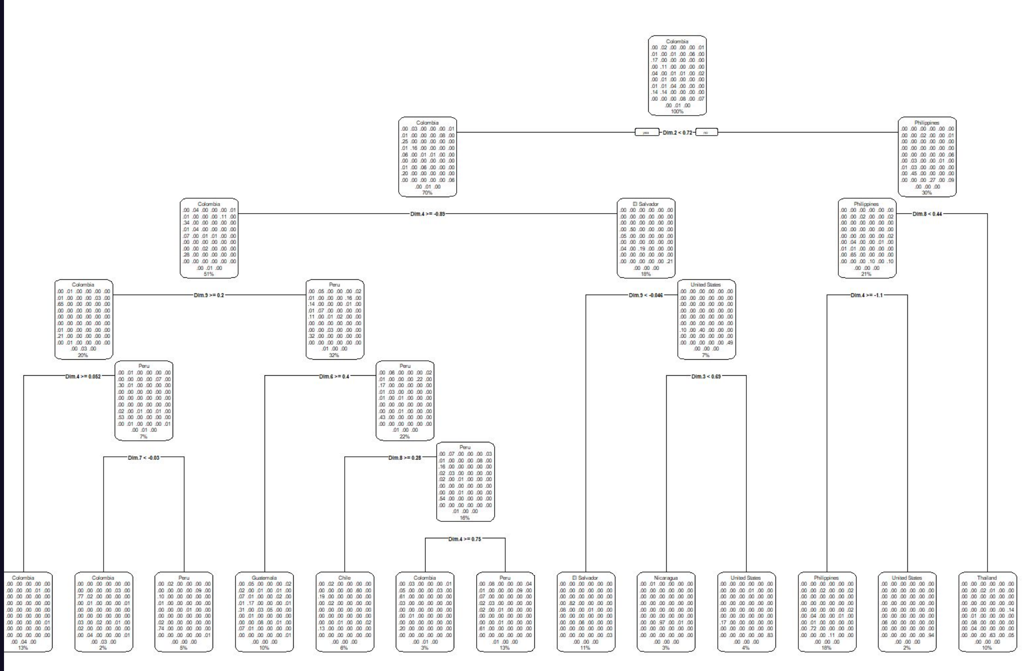
Decision Trees



Armed Assault
Assassination
Bombing/Explosion
Facility/Infrastructure Attack
Hijacking (unused)
Hostage Taking (Barricade Incident) (ur
Hostage Taking (Kidnapping)
Unarmed Assault (unused)
Unknown
Armed Assault .53 .20 .19 .01 .00 .00 .03 .00 .04 100%
yes
Dim.3 < -0.22
no
Armed Assault .93 .00 .00 .01 .00 .00 .05 .00 .00 57%
Assassination .00 .46 .44 .00 .00 .00 .01 .00 .09 43%
Dim.2 < 0.28
Dim.3 >= 1.7
Hostage Taking (Kidnapping) .06 .00 .00 .18 .06 .00 .66 .04 .00 4%
Bombing/Explosion .00 .00 .81 .00 .00 .00 .02 .00 .17 23%
Dim.6 < 5.9
Dim.2 < 2.4
Armed Assault 1.00 .00 .00 .00 .00 .00 .00 .00 .00 53%
Facility/Infrastructure Attack .19 .00 .00 .54 .17 .00 .00 .10 .00 1%
Hostage Taking (Kidnapping) .00 .00 .00 .00 .00 .00 1.00 .00 .00 3%
Assassination .00 1.00 .00 .00 .00 .00 .00 .00 .00 20%
Bombing/Explosion .00 .00 .98 .00 .00 .00 .01 .00 .00 19%
Unknown .00 .00 .00 .00 .00 .00 .04 .00 .96 4%

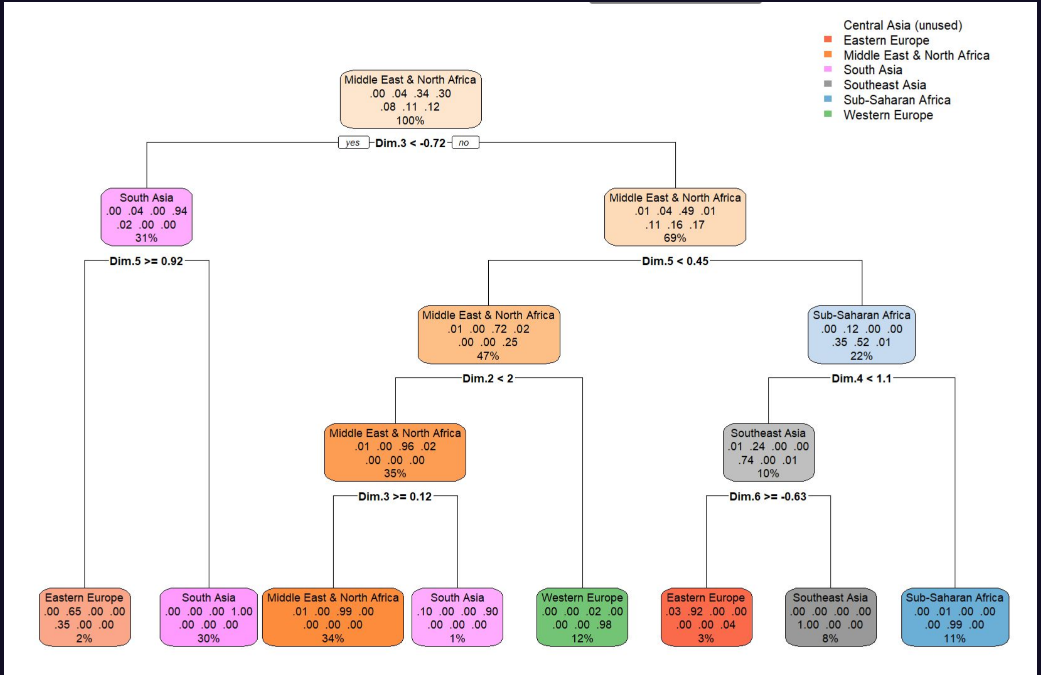
- Filtered for Region #5 - Southeast Asia
- Removed Unknown and NA Values
- Removed No Casualties
- Used Values such as Attack Type, Target Type, Year and Success to help Predict which weapons would be of Primary Use
- Used PCA to gather better testing results
- The model excels in predicting attacks of Armed Assaults, Assisnation and Bombing or Explosions.
- With PCA Confusion Matrix Accuracy increased to 95%
- Kappa - 0.97

Decision Trees



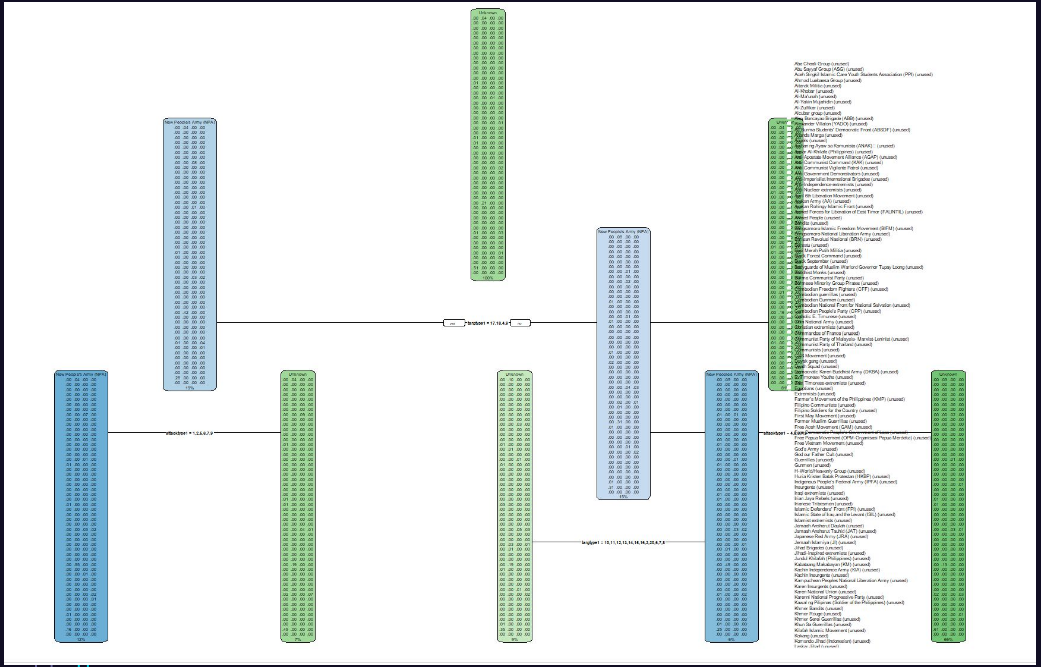
- Filtered for Regions 1-5
- Removed Unknown and NA Values
- Used Values such as Attack Type, Region, Target Type, Year, Success, Weapon to try and predict future countries of Attack
- Used PCA to gather better testing results
- The model excels in predicting attacks for huge countries such as Colombia, Chile and Mexico. Struggles with smaller countries like Antigua and Barbuda
- With PCA Confusion Matrix Accuracy increased to 70%
- Kappa - 0.65

Decision Trees



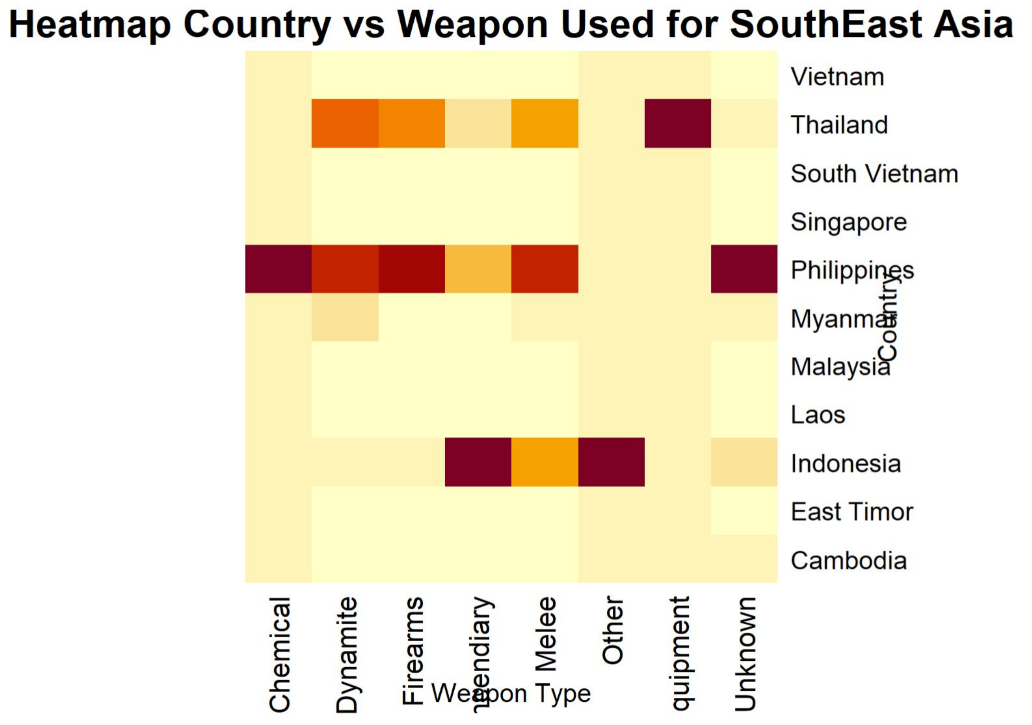
- Filtered for Regions 5-11
- Removed Unknown and NA Values
- Used Values such as Attack Type, Region, Target Type, Year, Success, Weapon to try and predict future Regions of Attack
- Used PCA to gather better testing results
- Southeast Asia accounts for 31% while Middle East accounts for 69%
- With PCA - Confusion Matrix Accuracy increased to 98%
- Kappa - 0.97

Decision Trees



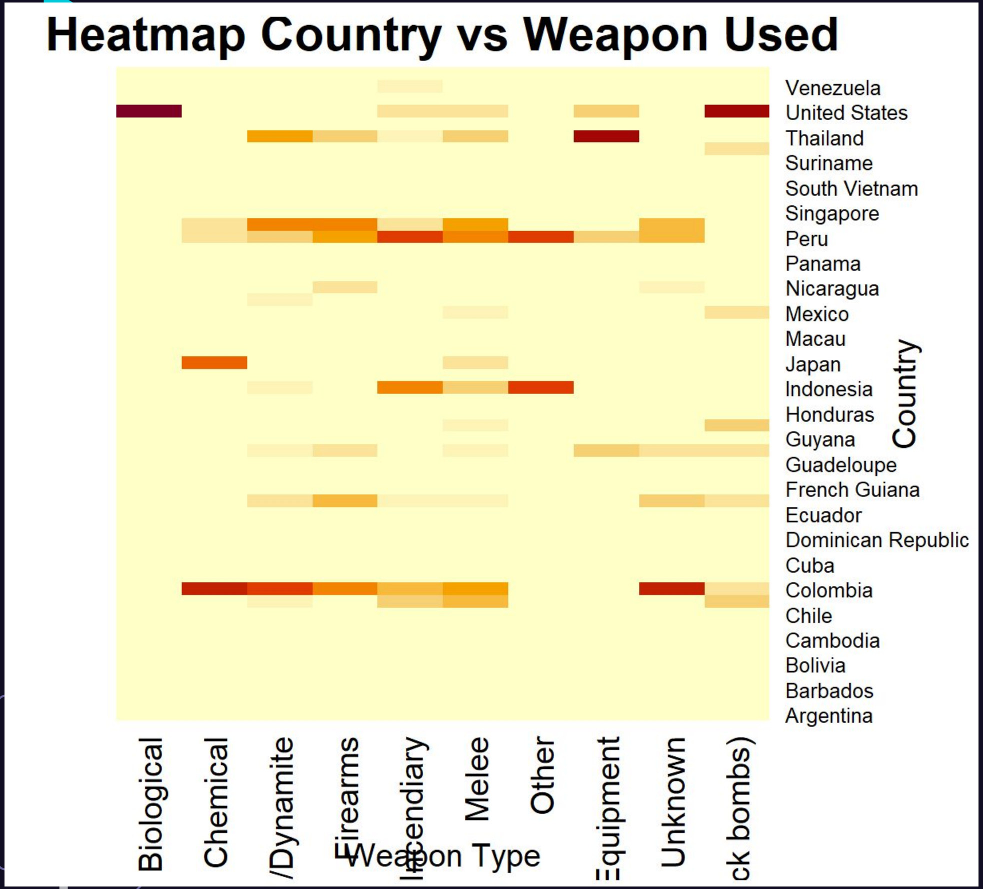
- Filtered for Regions 5
- Attempting to predict who is responsible for unknown attacks in the Southeast Asia Region
- Using data points such as Group Name, Attack Type, Target Type and Property
- The Models is either unable to predict who is responsible 81% of the time or predicts the New People's Army is responsible 19% of the time
- Confusion Matrix outputs an Accuracy of around 55% - Not very Effective

K Means Clustering / Heat Map



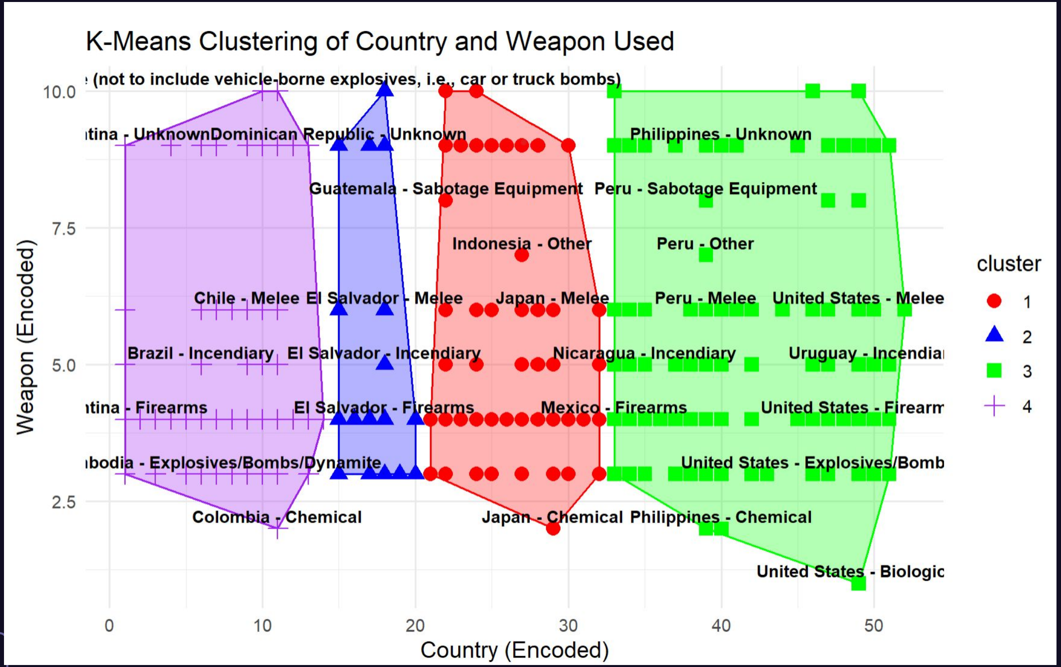
- Filtered for Region 5
- Removed Unknown and NA Values
- Removed Instances of no Casualties

K Means Clustering / Heat Map



- Filtered for Region 1-5
- Removed Unknown and NA Values
- Removed Instances of no Casualties

K Means Clustering / Heat Map



- Filtered for Region 1-5
- Removed Unknown and NA Values
- Removed Instances of no Casualties
- Clustering on Countries of Attack and Weapon Used

Thank You. That is the End of the Presentation

Any Questions?