Final Project

Kevin Hansen, Benjamin Tisinger, Kent Roller

IST 707

09/20/24

Introduction

Terrorism was something most of the western world was unaware of prior to the events of 9/11. After the tragedy of 9/11 the western world at large was forced to reckon with something it had previously been able to largely ignore. The establishment of anti-terrorism groups both domestically and abroad soon unraveled the web of terror cells and organizations that spread across the globe. This newfound awareness of these groups' existence and proliferance throughout the world motivated the anti-terror groups to expand the monitoring and observation of these groups and their activities.

Due to the large number of terrorist organizations, the perpetrators behind unclaimed attacks can have many potential names. The need for being able to analyze these groups behaviors, their motives, their operation portfolios, locations of operation and more has grown exponentially. The ability to assign responsibility to unclaimed attacks allows for anti-terrorism groups to more efficiently allocate resources, both those of monetary type and man-hours, to areas that may be at high risk of attack or subsequent attack by unclaimed assailants.

Being able to categorize the behavior of these groups will further allow anti-terrorism organizations to act preemptively instead of on a reactionary basis to the actions of these terror groups. When a likely group can be assigned their tactics and operations will likely already be familiar to localized task groups which can allow for more effective hunting of unknown terror cells in the area as well as working to subvert or cut off funds and equipment moving through known channels.

Analysis and Models

The first steps in the analysis was uploading the data set and preparing it for analysis. 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. The data was further broken down to only attacks that occurred in the Southeast Asia region.

Now that the dataset is ready to be analyzed some simple data exploration was performed. This can be seen in the figures below that give a categorical breakdown of everything that is happening in the Southeast Asia region.Figure 1.1 shows that the most

attacked countries are the Philippines and Thailand. Figure 1.2 displays that most attacks were done by armed assault or bombing. Figure 1.3 the number of attacks has grown significantly in the last 10 years. Lastly, figure 1.4 shows who is the target of the attacks when they occur and most of the time it is the government or the military.

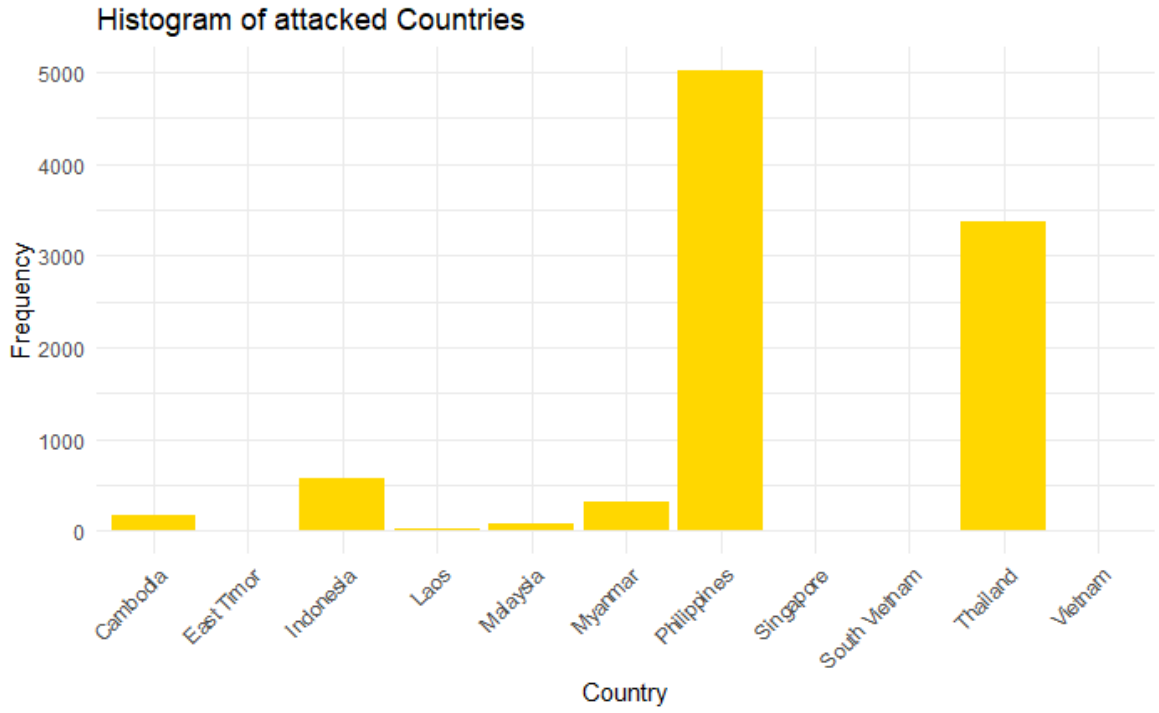

Figure 1.1

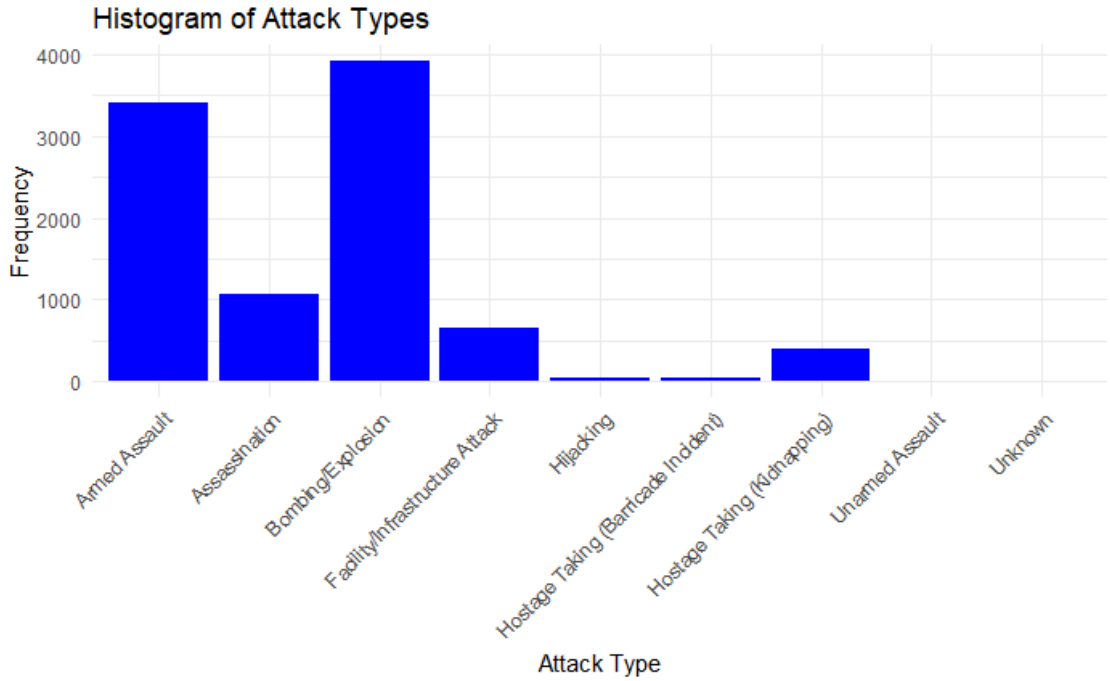

Figure 1.2

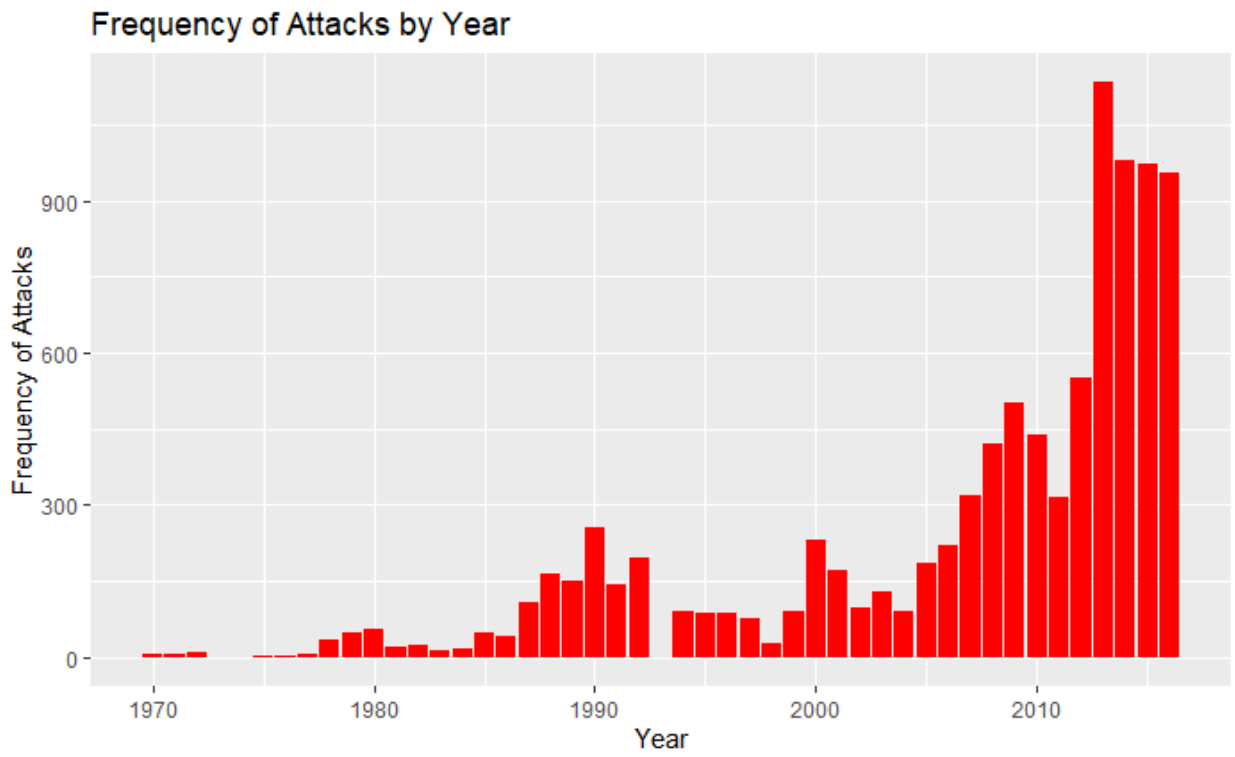

Figure 1.3

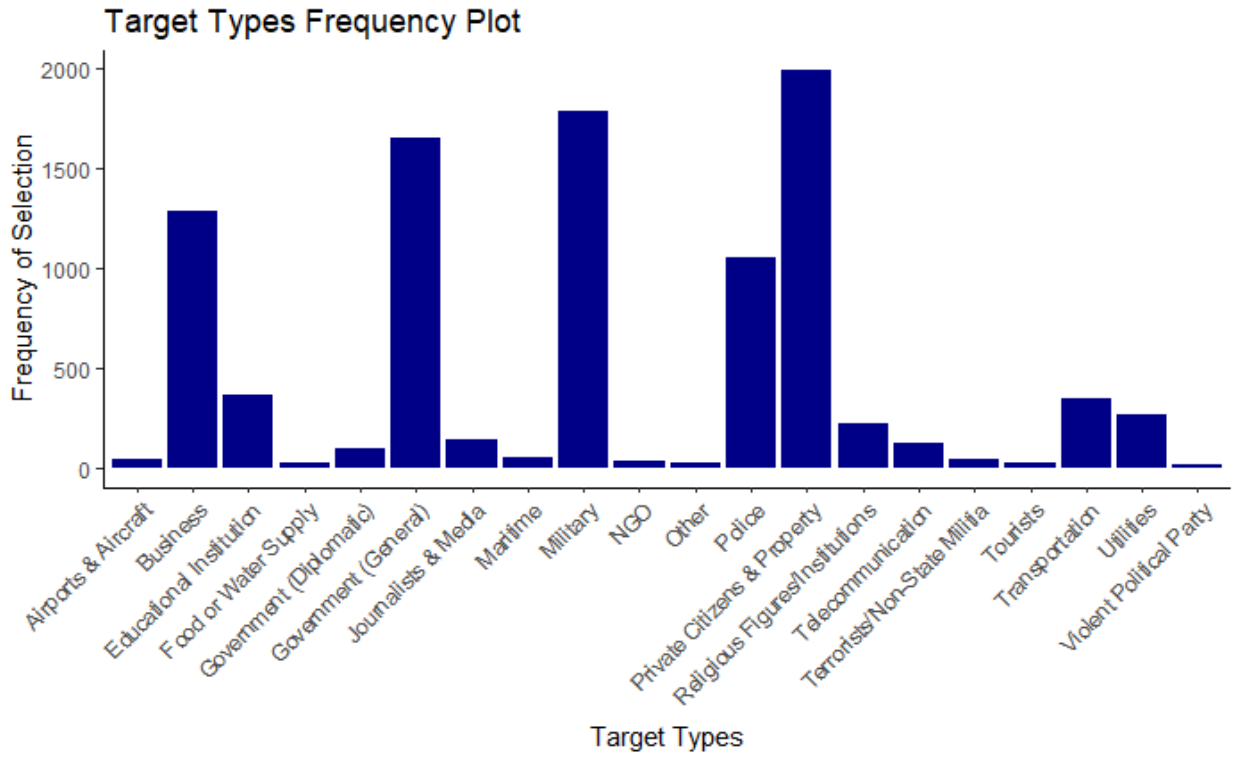

Figure 1.4

Now the part of the dataset that sparked interest was the category that belonged to the organization that takes responsibility for the attacks. As can be seen in Table 1.1, there are over 5,000 attacks where it is unknown who committed these atrocities. Our objective was to develop a very accurate model in order to determine who carried out these attacks.

gname <chr>	count <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 1.1

Results

Naïve Bayes Model

For the next model naïve bayes was selected. As with the previous model, the dataset used is preprocessed from the original dataset. Naïve Bayes is a classification method, which means that the data will need to be transformed in order to use this type of mode. More specifically, the variables currently stored as characters in the data set will need to be converted to factors. The first step in this process is then converting these chr into factors. After this initial conversion takes place an additional conversion is made, whereby all factors are then converted to numeric variables. Next the response variable is removed from the data and placed in a variable in named response, in this instance the gname or group name is the variable of interest. Next, the predictors are then placed in their own variable named accordingly. Normalization must then be applied to the data set before PCA can be performed. After normalizing the

dataset,Principal Component Analysis was used to determine the variables of interest

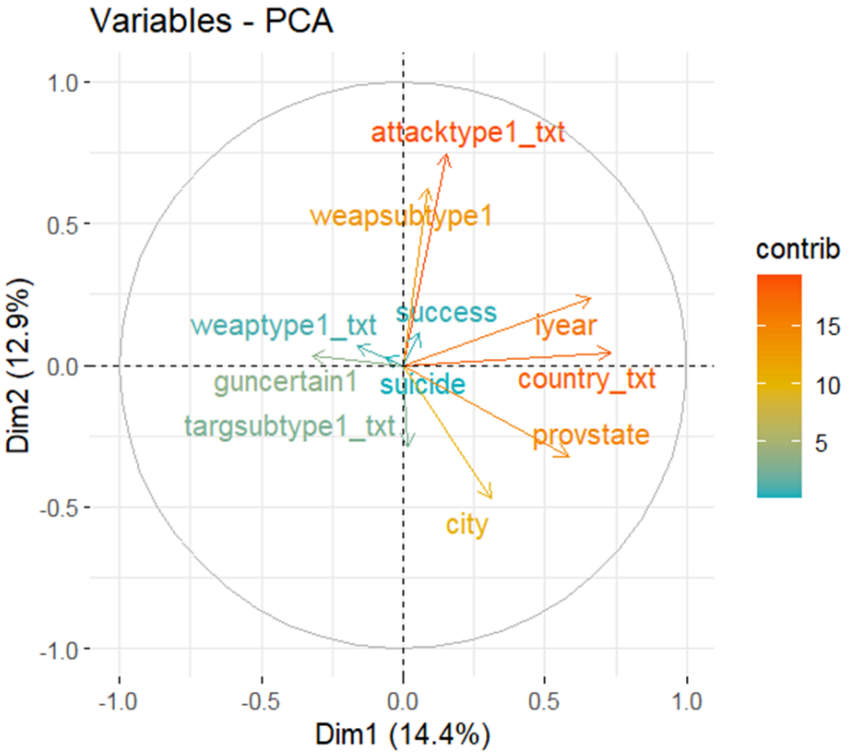

Figure 2.1

Figure 2.1 displays the PCA variables plot, where according to the plot the first component explains 14.4% of the variation in the data, and the second component explains about 12.9% of the variation in the data. The plot also indicates which variables are of the most interest via color coding of the text in the plot. Here it can be observed that the attacktype, country, provstate, and iyear are the most significant contributors to the explanation of the variance in the data. Additional visualizations of the component selection can be found by looking at the scree plot and barplot of the component selection.

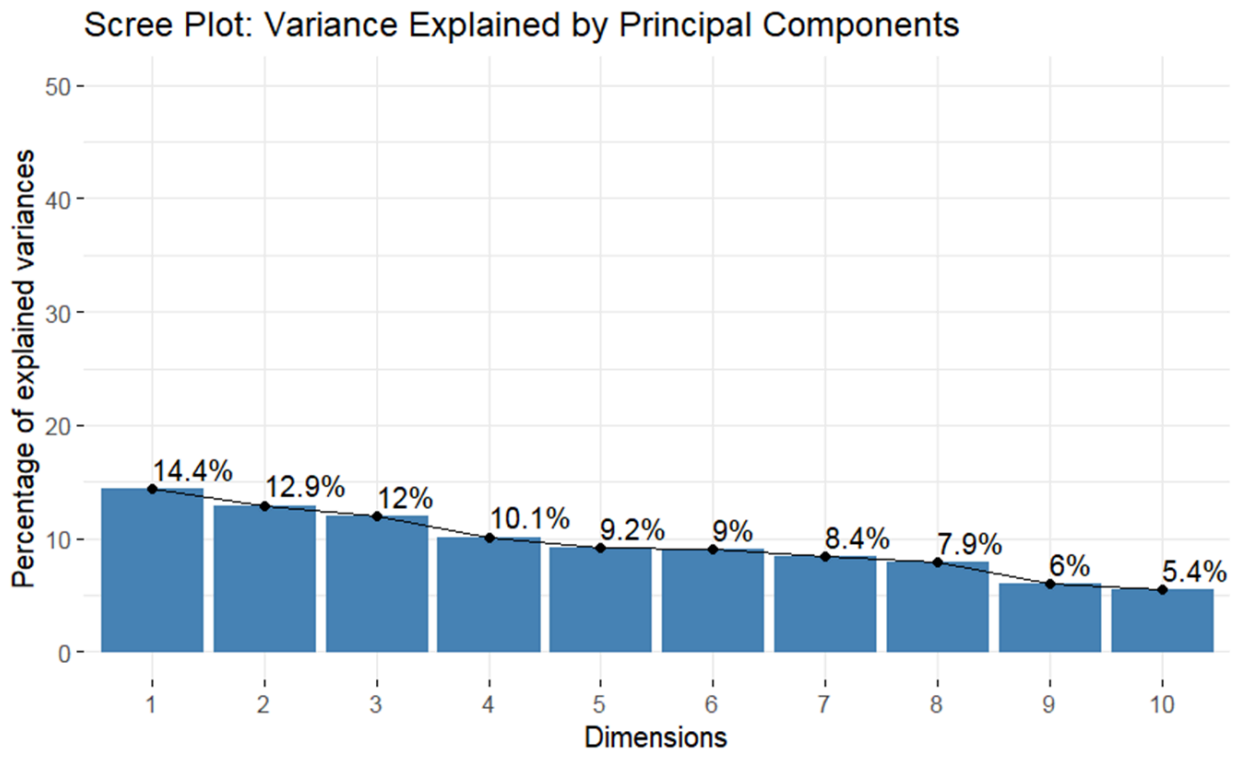

Figure 1.2

Figure 2.2 displays the scree plot for the principal component selection. The scree plot reflects the findings from figure 2.1, where component 1 is shown to explain 14.4% of the variation in the data. Component 2 likewise, is shown to explain 12.9% of the variation as shown in figure 2.1. However, the additional components can be seen here, where component 3 also has a decent amount of percentage. Components 4,5 and 6 show a decent amount of percentage as well, with the real drop coming at component 9. As such, models with up to 8 components should have the highest explanation of the variance in the model, and as a result, should perform the best in terms of accuracy of predictions by the model.

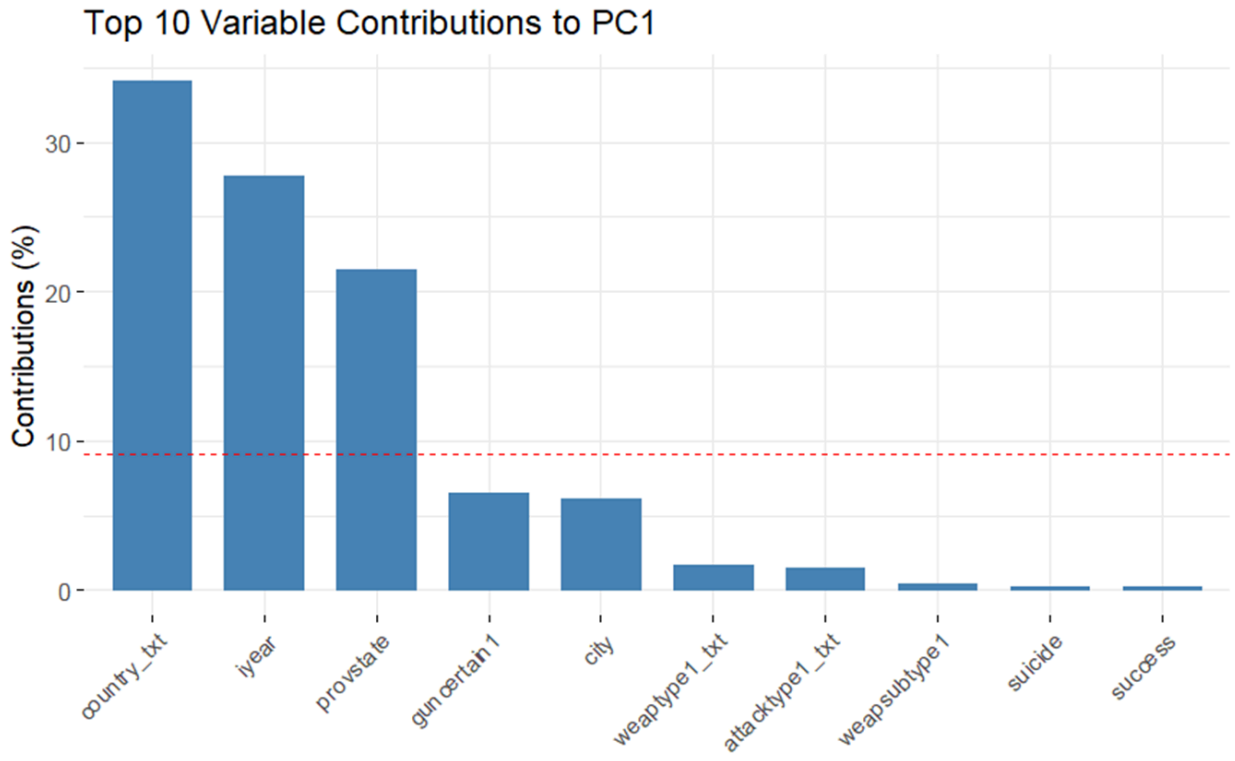

Figure 2.3

Figure 2.3 displays the barplot produced by examining the top contributing variables to component 1. Here it can be observed that the country of attack played the most significant contributing role to determining the group name, which in itself is not very surprising. Year is next, which also makes some logical sense as demographics and political attitudes change over time, so a group that may be highly active for one decade may come into power, may dissolve, disband, or fade/combine into other factions to make a new group. Guns are slightly more interesting in terms of unique unexpected outcomes. This indicates that certain groups operate with a preference for certain types of attacks, those who use guns or do not use guns can provide a good indication of who is behind an unclaimed attack. This value is of course below the cutoff for contribution to the component as indicated by the red line running horizontally across the plot.

Model 1: PCA=Default Value	Accuracy: 61.20%
Model 2: PCA=11	Accuracy: 64.18%
Model 3: PCA=6	Accuracy: 61.20%
Model 4: PCA=8	Accuracy: 66.17%
Model 5: PCA= 4	Accuracy: 61.70%

Figure 2.4

Figure 2.4 shows a table with the results of each of the Naïve Bayes models run. As can be seen in the table, the performance is fairly poor across the board, with the best performance found by using 8 components netting an accuracy of 66.17%. This indicates that Naïve Bayes may not be the best method for trying to obtain predictions with this dataset. Additional performance may be found by additional parameter tuning, but for now, the next model will be run to see if better performance can be found elsewhere.

Support Vector Machine Model

The next model selected is the Support Vector Machine model. This model, like the naïve bayes model used before it, will require some additional pre-processing of the data before it can be used. It should be noted that the PCA selection of 11 was used for the creation of the training and test data sets used in the SVM model. This was done to ensure that the largest number of predictor variables was used with the dataset, however, this may have also caused performance issues with the model. As such, given more time, the model would have been run a second time with independent pre-processing performed on the data before being used with SVM. This may have resulted in better performance for the model. Once the data was split the SVM model was run on the data as is to establish a baseline of performance. The baseline performance of the model was 65.90%. This is roughly the same performance that one could expect from using the naïve bayes model. In order to gain more accuracy, the dataset was binarized and then the SVM model was reapplied. Unfortunately, this did not help performance, in fact, it degraded the performance of the model substantially. The accuracy of the model after binarizing the dataset dropped to 45.13%. Repeated model creation using various kernels did little to improve the performance of the model, where sigmoid and radial kernels performed as poorly as the default method. In hindsight, a polynomial kernel should have been used on the non binarized data to establish more of a baseline for this method. Due to the poor performance of the SVM with this data, this model can not be recommended, that is, without further considerations being mentioned at the beginning of this section.

Neural Network Model

As a last ditch hail mary to obtain a model that gave reasonable performance, a neural network model was quickly cobbled together. However, due to the rapid nature of the deployment of this model, pre-processing errors and incorrect model application may be present. As such, this section should be regarded only as a proof of concept and not as a conclusive statement about the effectiveness of Neural Networks with this type of data. With more time additional pre-processing and model creation steps would have been taken in order to try and obtain optimal performance from the model. The first issue arose when trying to use variables that were factors in the dataset. After several different attempted solutions were applied to no avail, the data set was stripped of the factor variables and only left with the numeric variables with the exception of the predictor. The dataset was then normalized with only the numeric variables present, and then split into training and test datasets.

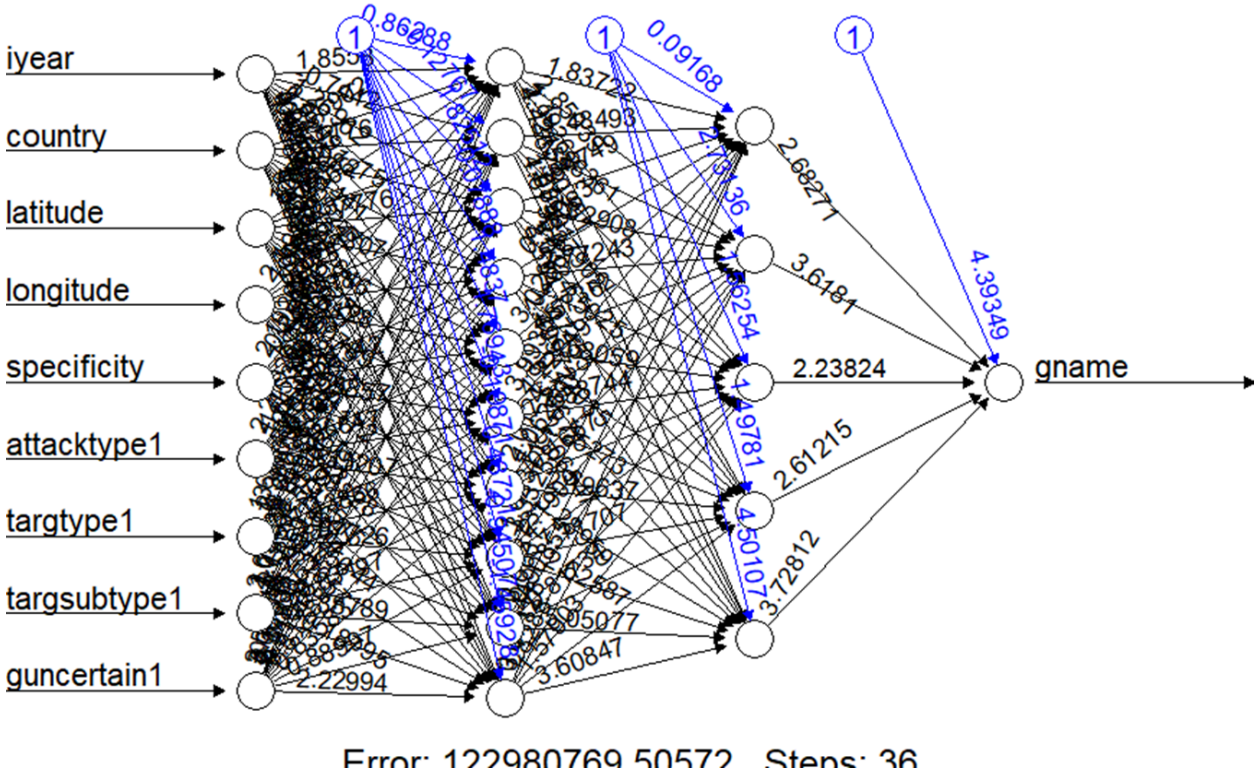

Figure 2.5

Figure 2.5 shows the neural network plot. Once more it should be noted that the variables listed are only numeric variables as all other variables were removed. Obtaining this output resulted in a short lived celebration as immediately problems were encountered in trying to obtain the accuracy of the model. Unfortunately, despite obtaining a model, issues were present in trying to convert the factor response back to its original state for comparison. This is most assuredly operator error, and as a result of this operator error the results of the Neural Network are inconclusive. As a result the model is considered void for this analysis.

Random Forest Model

The Random Forest model was chosen because of the scalability of the algorithm. It is suitable for large datasets with many features, making it a powerful tool for complex problems. It also has higher accuracy than a decision tree model because instead of a single tree it is taking the average of several trees which can be a huge advantage. This model took advantage of that by developing three different models using 100, 200, and 500 trees.

Figure 3.1 shows the difference in accuracy between the three models. The difference between 100 and 200 is nonexistent but even the difference between them and 500 is only 0.002%. Overall, the model is decently accurate at just below 90%. Figure 3.2 shows the different variables and which were most important when making the prediction. Location is the most important factor to decide which group performed the attack followed closely by the year.

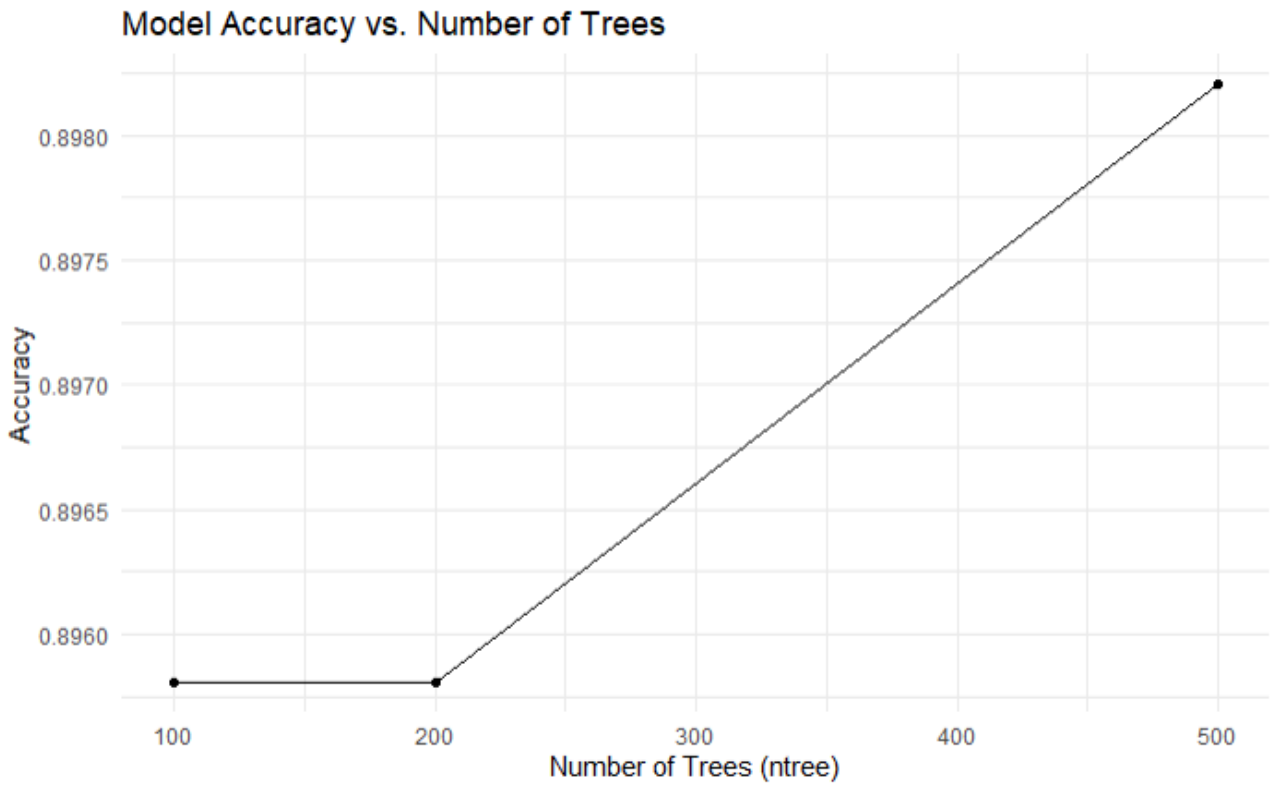

Figure 3.1

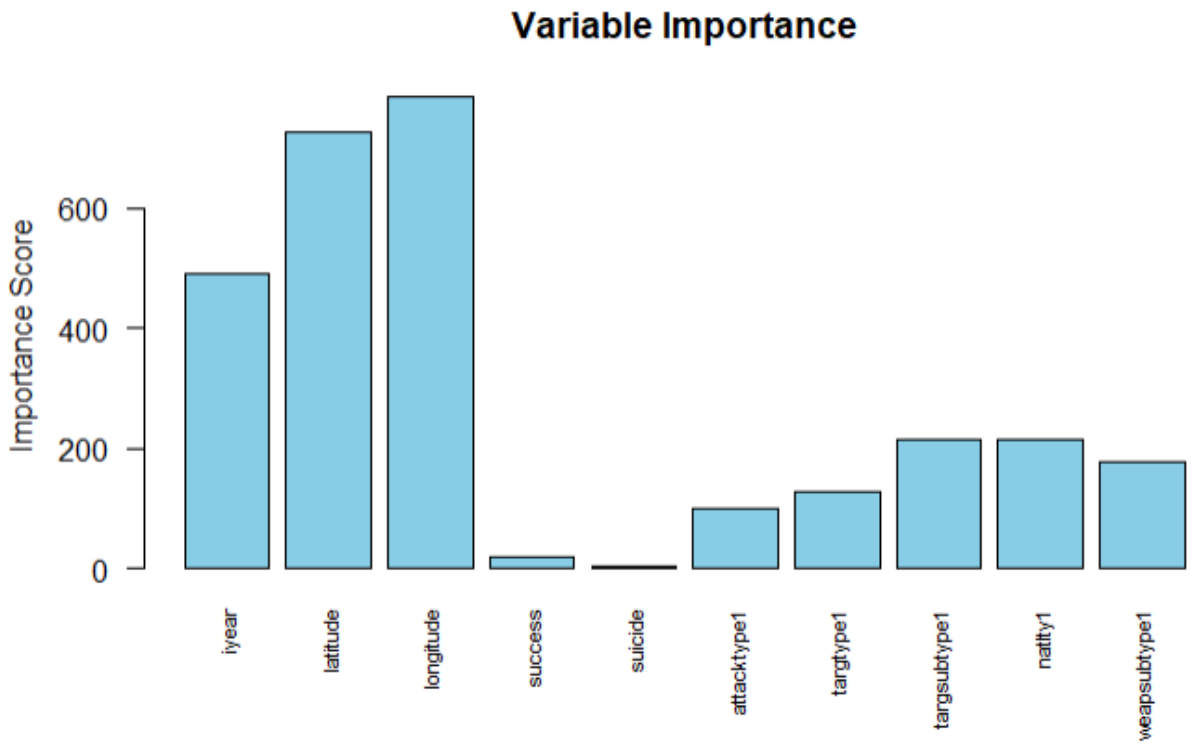

Figure 3.2

Decision Tree Model

In this analysis, I began by loading a filtered dataset focused on region #5. Further refinements were made by excluding cities classified as "unknown" and removing rows where attacks resulted in zero casualties. Key columns used in the decision tree analysis included Attack Type, Target Type, Year, Success, and Weapon Type. These variables were renamed and converted into factors to ensure compatibility with the decision tree model.

To build the model, the dataset was divided into training and testing sets following a conventional 70/30 split. The decision tree revealed that armed assaults represent the most frequent attack type. When integrating temporal data, it became apparent that after 2014, the decision tree predicts that 54% of attacks will likely be armed assaults.

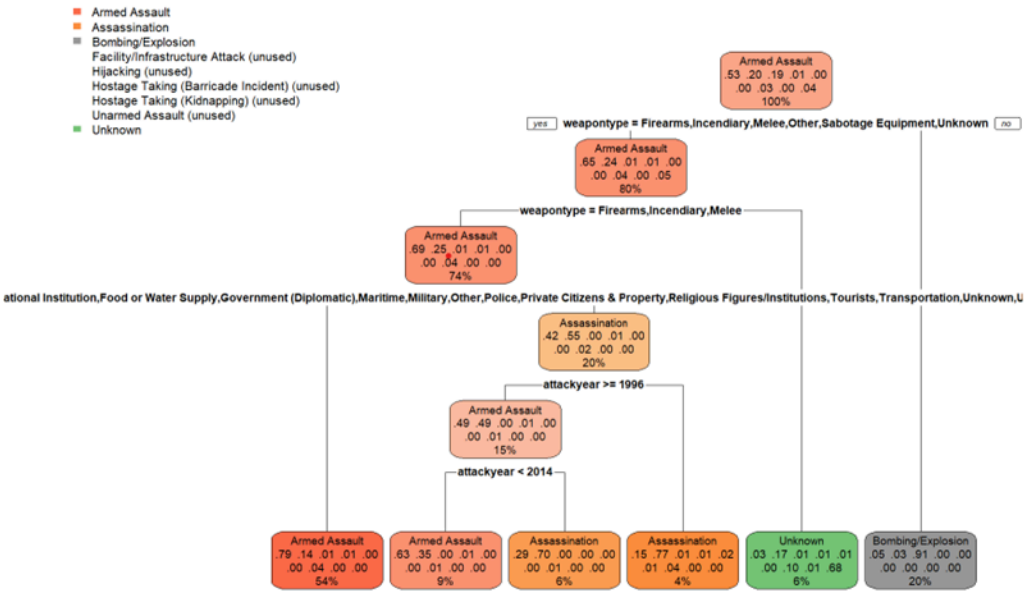

Figure 3.3

In the subsequent phase of our analysis, we developed another decision tree after filtering out instances of unknown attacks and rows with no casualties, similar to the previous process. The same key columns—Attack Type, Target Type, Year, Success, and Weapon Type—were used. This time, we applied Principal Component Analysis (PCA) to transform the data and then converted these columns into factors.

The goal of this decision tree was to improve understanding of the relationship between weapon types and specific geo locations within Southeast Asia. After running the model, the confusion matrix showed an accuracy of approximately 98%.

From this decision tree, we can still infer that Armed Assault is the most likely type of attack. Additionally, the decision tree and confusion matrix effectively predicted Armed Assault, Assassination, and Bombings/Explosions. However, the model encountered difficulties in predicting attack types such as Facility/Infrastructure Attacks and Hostage-Taking.

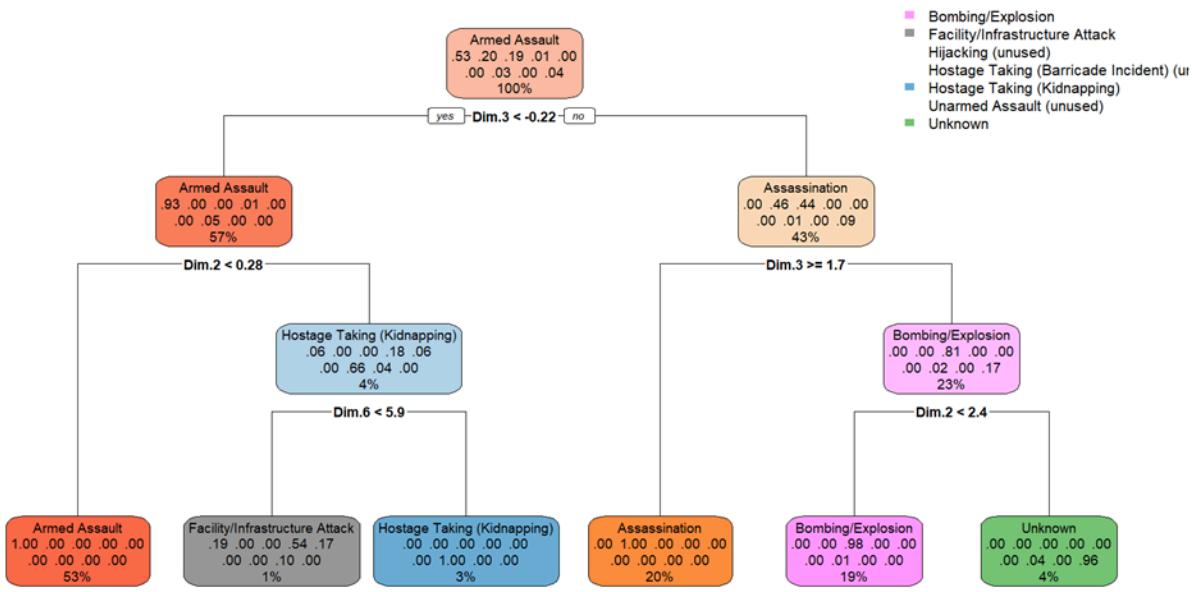

Figure 3.4

In our analysis focused on Region 5 (Southeast Asia), we aimed to predict responsibility for unknown attacks using data points such as Group Name, Attack Type, Target Type, and Property.

The model demonstrated limited predictive ability, with the confusion matrix showing an accuracy of around 55%, indicating that the model is not particularly effective. Specifically, it was either unable to identify the responsible group 81% of the time or predicted that the New People's Army was responsible in 19% of cases. This highlights significant challenges in accurately attributing responsibility for unknown attacks in the region.

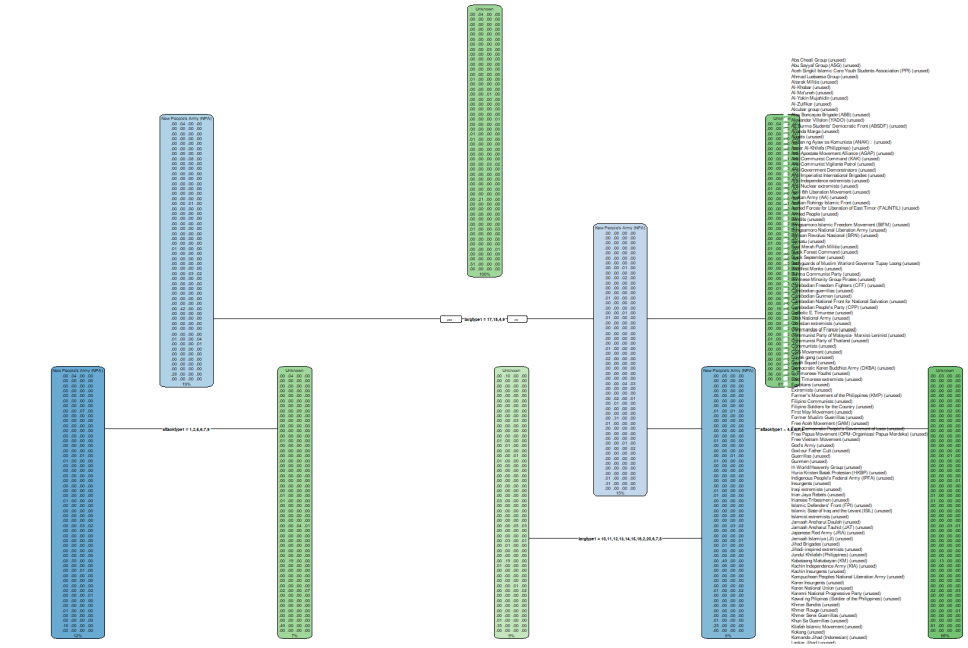

Figure 3.5

K-Means Clustering / Heatmap

For our analysis in Region 5 (Southeast Asia), we refined the dataset by removing instances where values were classified as "Unknown" or "NA" and excluding cases where there were no casualties. This filtering process allowed for a more focused dataset, ensuring that only relevant and complete records were included in the analysis. By eliminating these incomplete data points, the model's predictions could be more accurate and meaningful, improving the overall quality of insights drawn from the cluster charts and heatmaps.

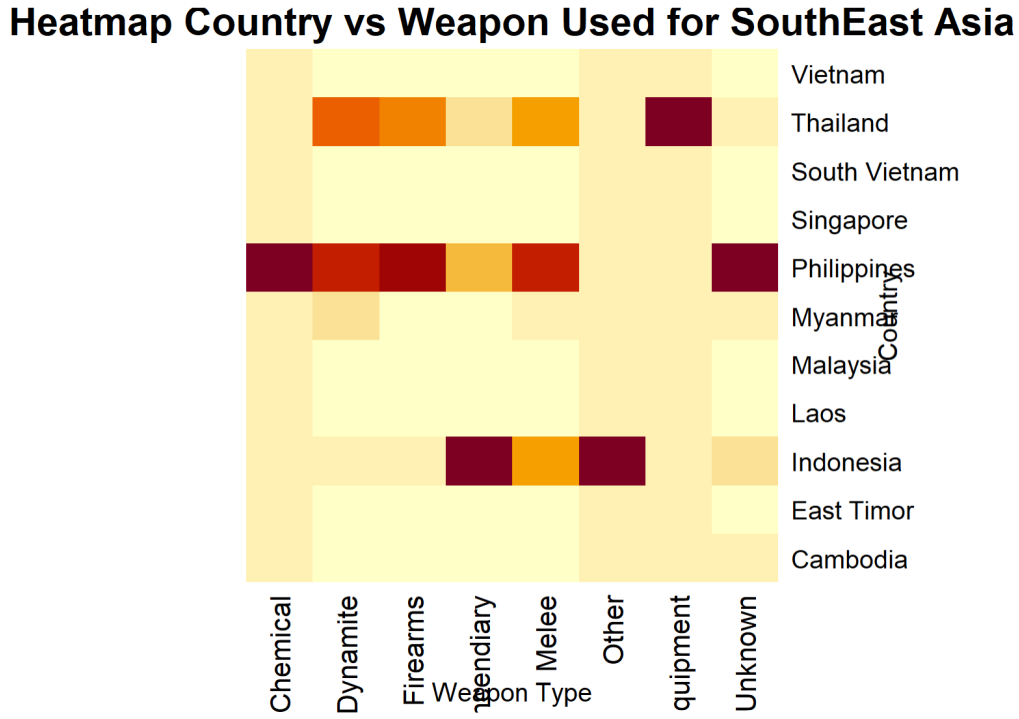

Figure 3.6

The map illustrates that firearms and dynamite are the most frequently used weapons across Southeast Asia, with certain countries showing a stronger preference for one over the other. Other weapons, such as chemical agents, incendiary devices, and melee weapons, are used far less often. Additionally, the map highlights significant differences in weapon usage across the region, reflecting the diverse conflict dynamics and security challenges faced by each country in Southeast Asia.

The chart identifies four distinct clusters: Cluster 1 is dominated by countries that primarily use firearms, Cluster 2 includes those that frequently deploy incendiary devices, Cluster 3 features countries that rely on explosives and sabotage equipment, and Cluster 4 consists of nations that predominantly use melee weapons and other unconventional arms. This clustering suggests that countries within each group face similar security challenges and could potentially benefit from coordinated efforts to address weapon-related concerns.

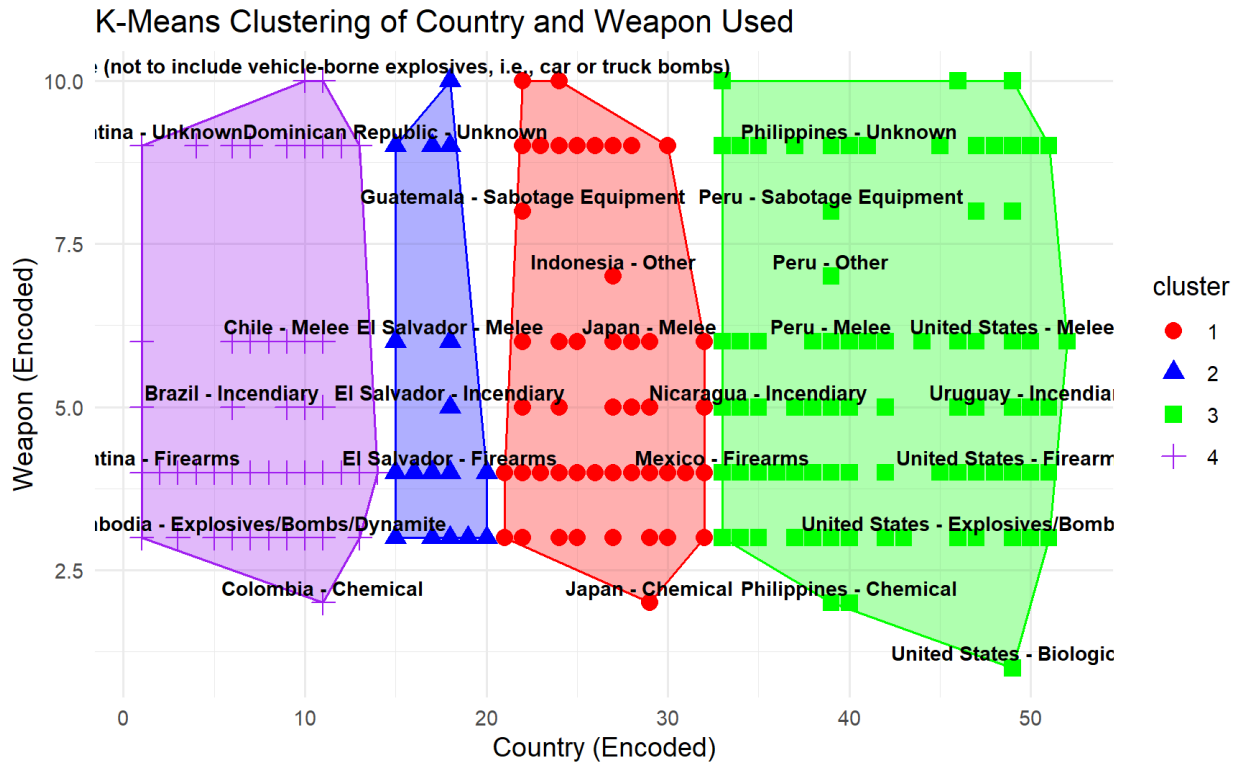

Figure 3.7

Conclusions

In conclusion, our analysis of terrorism in Southeast Asia, particularly in Region 5, has provided valuable insights into attack patterns, weapon usage, and the challenges of predicting responsibility for unknown attacks. By refining the dataset and applying various models such as decision trees, Naïve Bayes, Support Vector Machines, and Random Forests, we deduced key findings. Armed assaults emerged as the most frequent attack type, and the New People's Army was identified in a minority of unclaimed attacks, though the model's overall performance accuracy remained limited. The use of Principal Component Analysis further enhanced the model's effectiveness in understanding the relationships between attack types and specific regions, achieving a high accuracy of 98% in certain cases. Despite these successes, predicting

responsibility for unclaimed attacks remains a challenge, as demonstrated by the relatively low accuracy in some of our models.

The geographic clustering and weapon usage analysis revealed distinct patterns across Southeast Asia, with firearms and dynamite being the most common weapons, and differing conflict dynamics influencing political challenges in the region. Ultimately, the project analysis highlights the need for improved predictive models and better-targeted data to enhance efforts in concluding potential responsible parties.