A Clustering and Analysis of Terrorist Attacks

Ryan M. Allen

February 23, 2020

Regis Univerisity Practicum_1

by Ryan Allen

This project is analyzing the Globabl Terrorism Dataset found on Kaggle (https://www.kaggle.com/START-UMD/gtd (https://www.kaggle.com/START-UMD/gtd)).

My project will consist of a the following steps: - Initial exploratory analysis - Imputation of missing values - K means clustering - Hierarchical clustering - Exploratory analysis of the found clusters

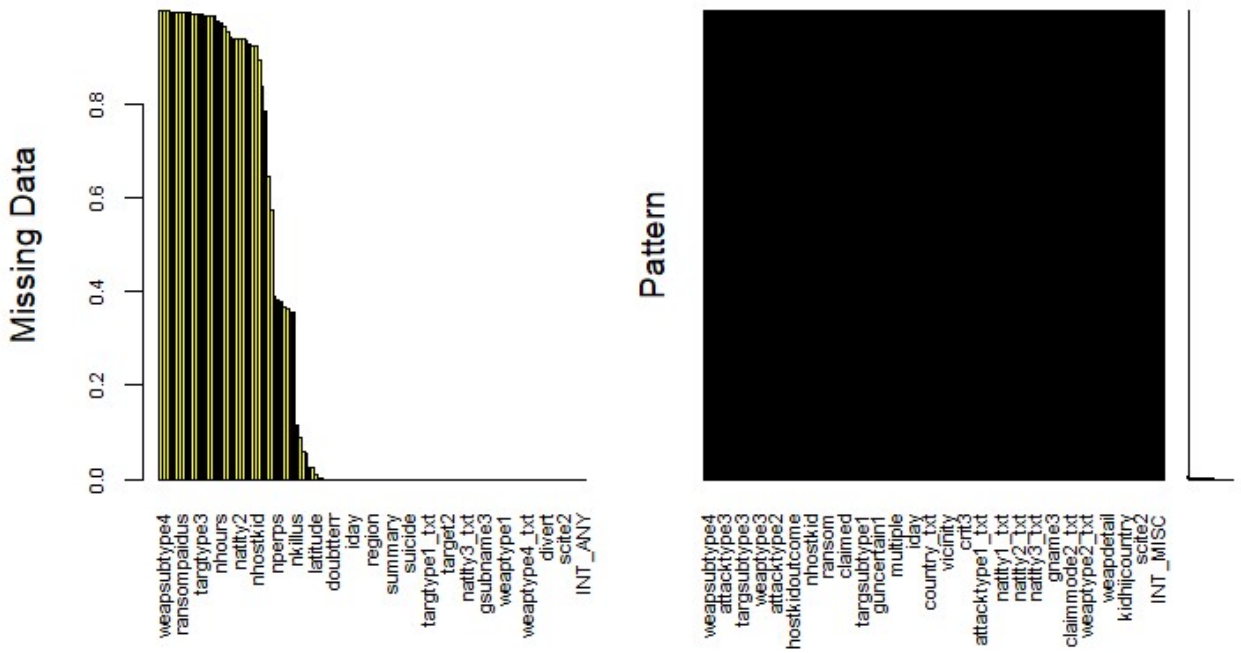
The driving question behind this analysis is what terrorist attacks are similar? Is is the weapon type, the type of attack, perhaps it is the group that carries out the attack or the group that is attacked. Maybe geographically there is something similar in these attacks. Perhaps regionally, certain weapons are favored for attacks, or maybe certain regions are deadlier than other regions. These are the types of questions that I am hoping to answer with this project. I will use the two clustering methods to create similar clusters of these attacks and wthen will analyze the groups.

About the data

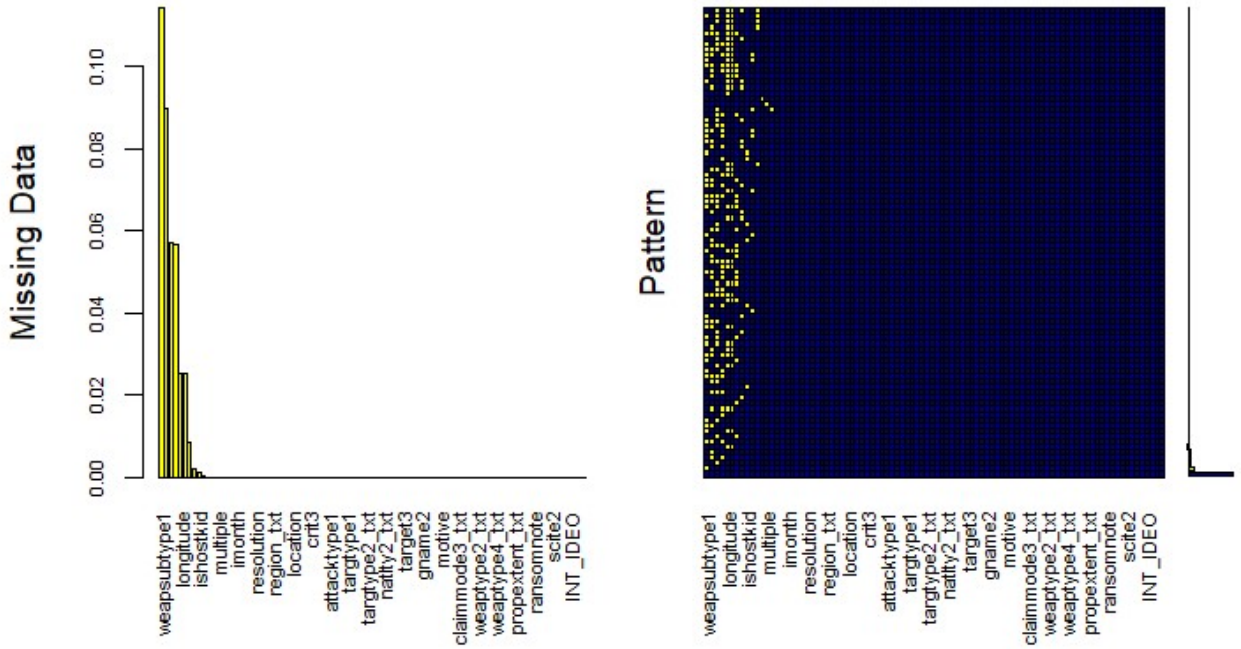
This data has been collected since 1970 until 2017. The dataset is maintained by the Study of Terrorism and Responses to Terrorism (START) at the University of Maryland.The dataset has 135 columns with over 180,00 values. Most of the data is categorical, one data column is a numeric field that relates to a corresponding text field. For the exploratory data analysis, I kept only the numeric field, but will rejoin the text fields when it comes time to do the conclusions.

Below are two plots from the MICE package that show the number of missing values in each column. The first plot is the original dataset, and the second is after I have filtered out any columns that have less than 25% missing values or NAs. Part of the dataset is that some columns are contingent on other columns, so if there is a 0 in one column, then the values in the next column will be blank. Doubterr is one of those columns.

MICE plot of all data



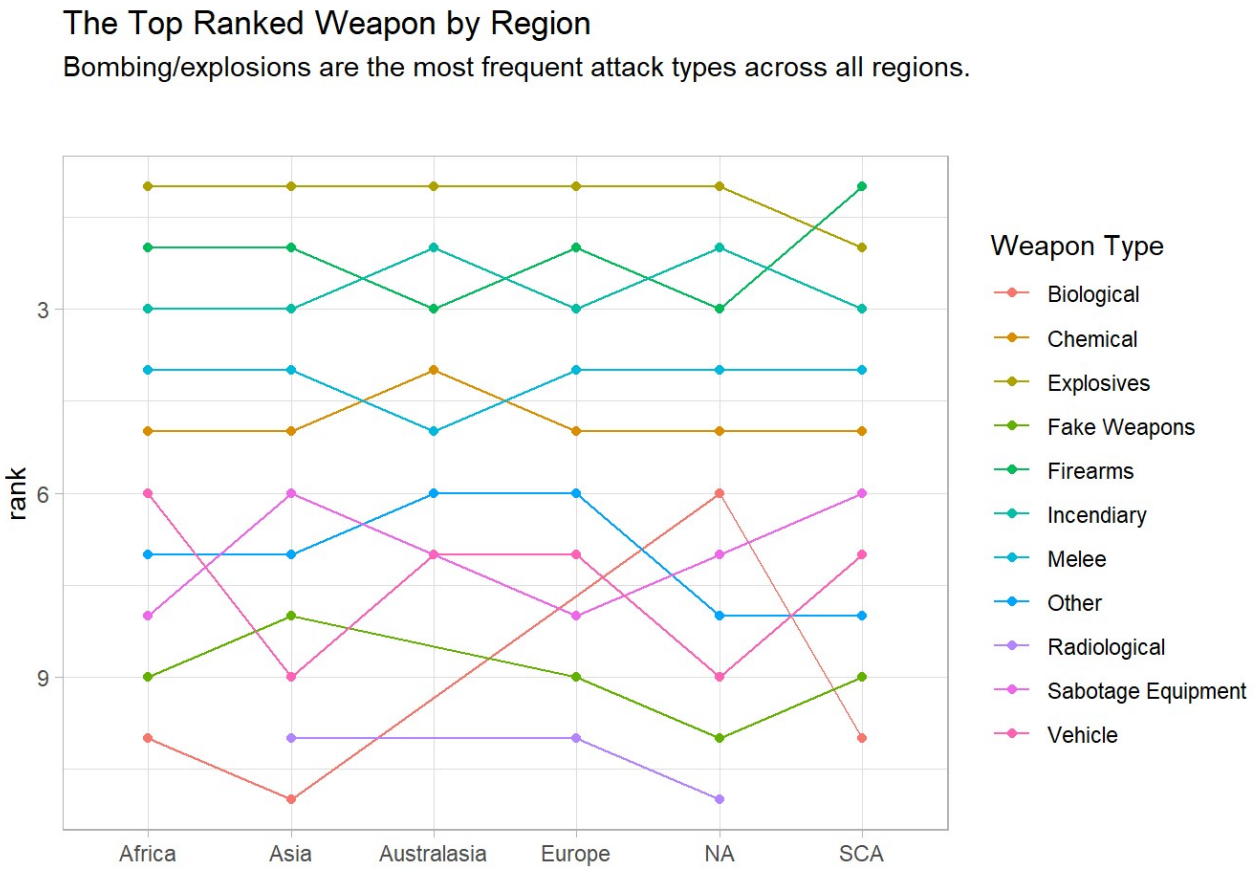
MICE plot of data with less than 25% missing values that was kept.



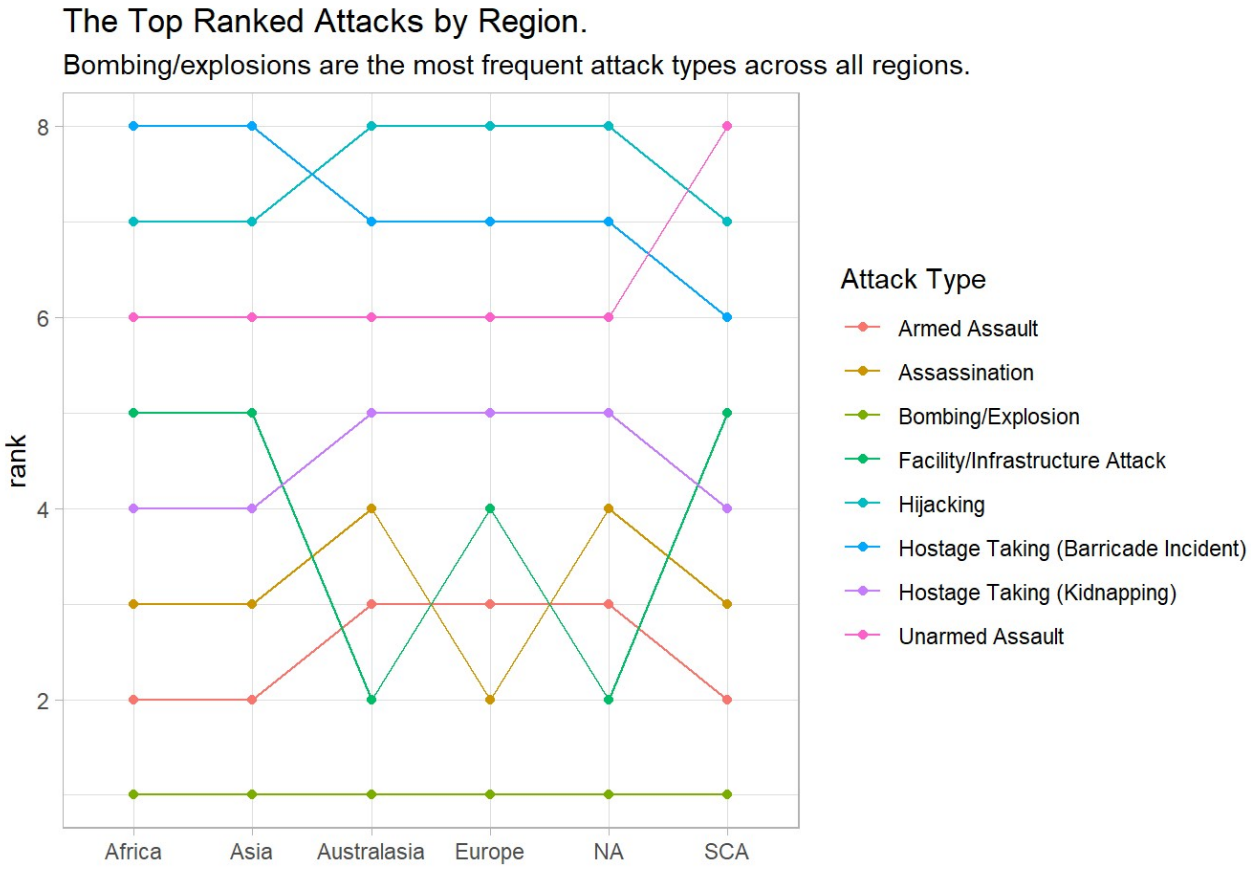
Exploratory Data Analysis

For exploratory analysis I focused mainly on weapontype, attacktype, crit1, crit2, crit3, geographic information (lat and long mainly), suicide and nkill. These values to me were the features that seem to make the most sense about terrorist attacks. For the first section I have kept all the data as is, with the exception of filtering on longitude as I had one value that was erroneous. In this section of the report I will focous on asking and answering questions about my data.

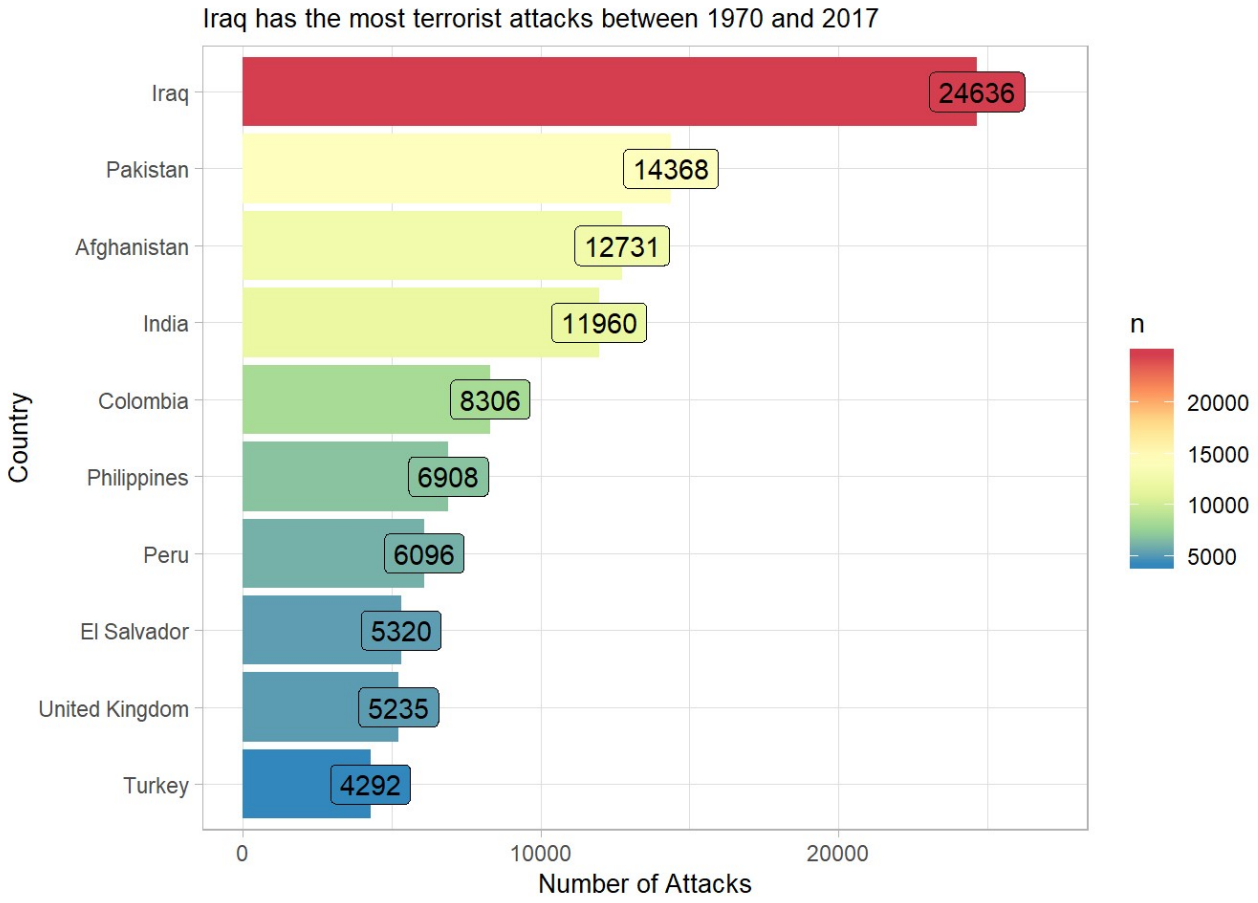
By region, what was the most popular weapon type?



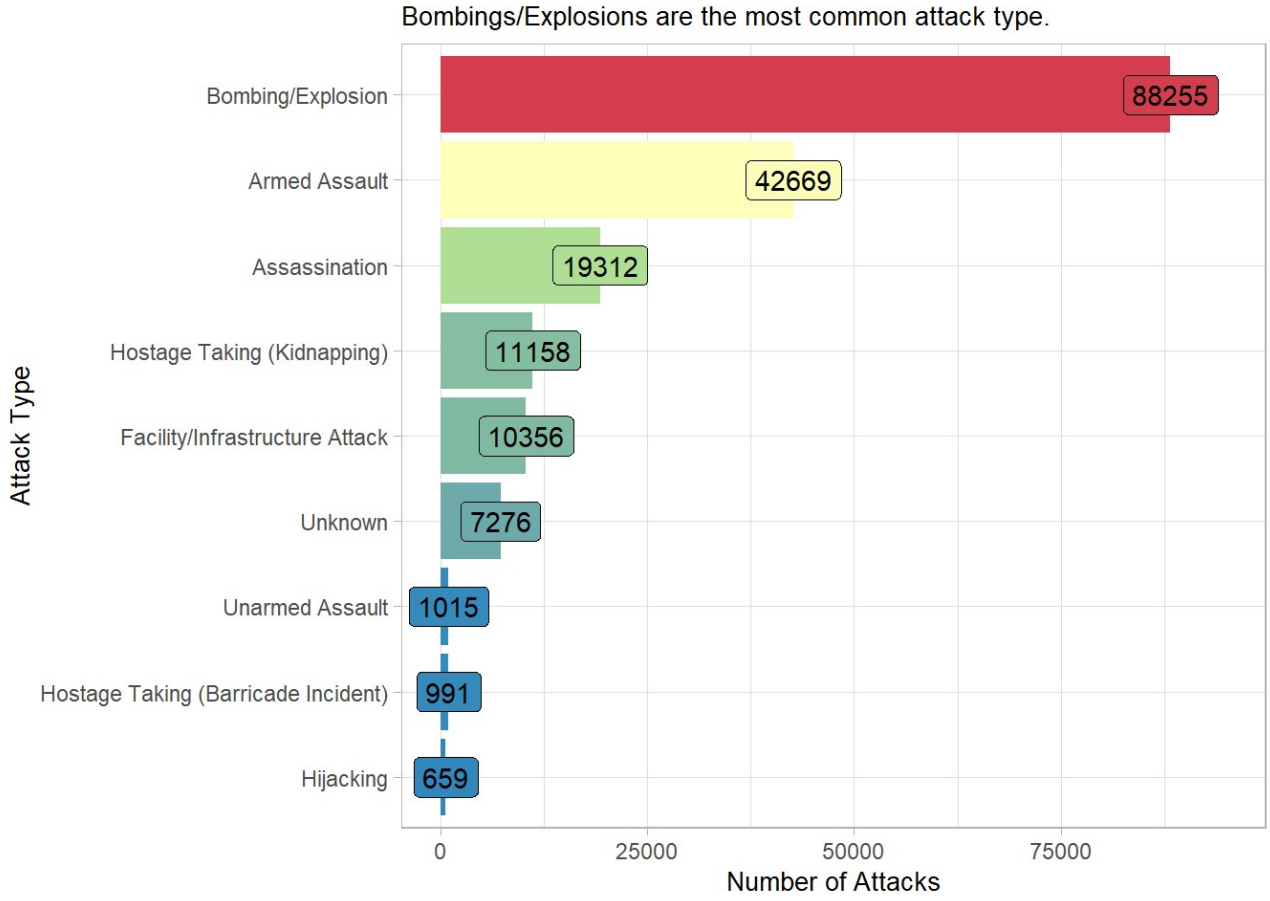
By region, what was the most popular attack type?



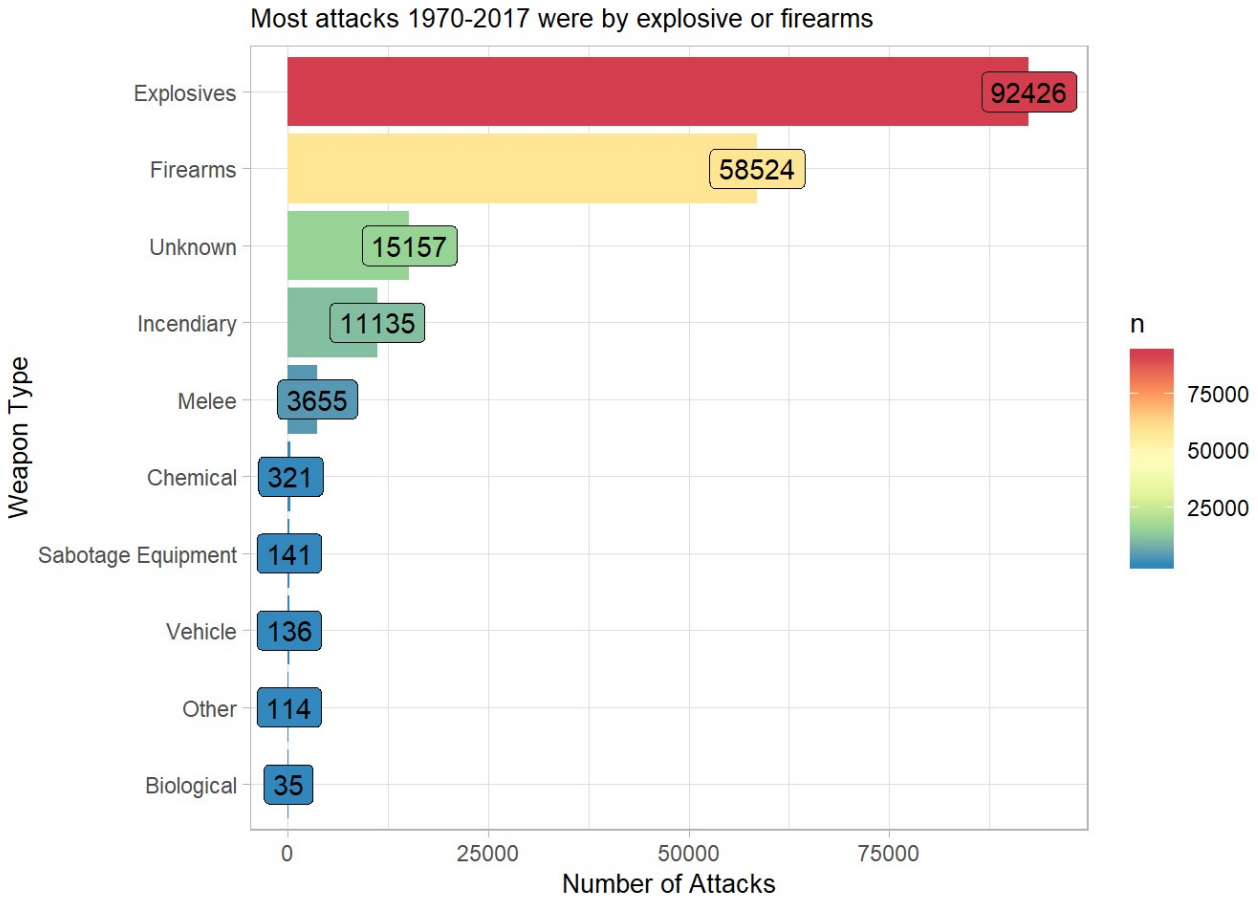
What countries had the most terrorist attacks?



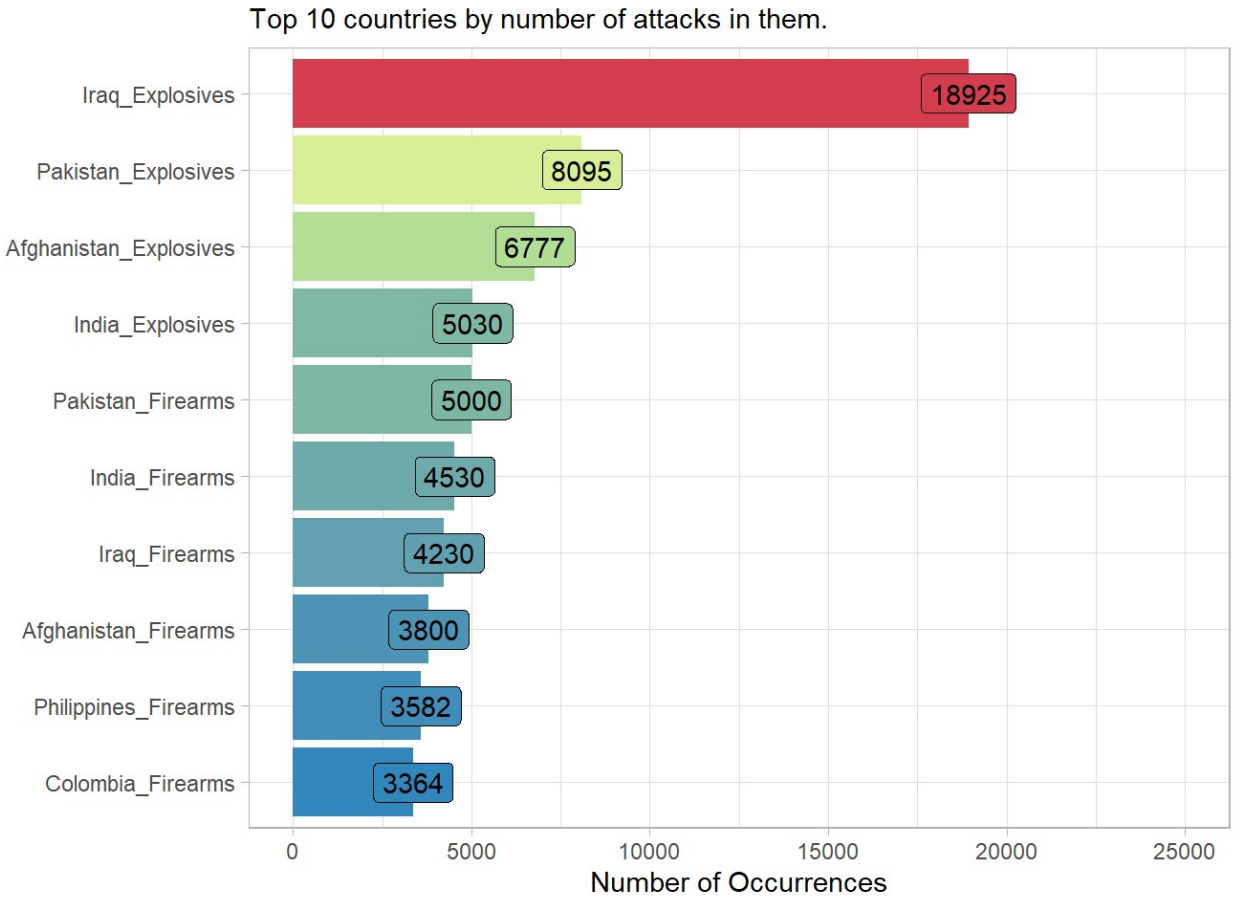
What attack types were most common?



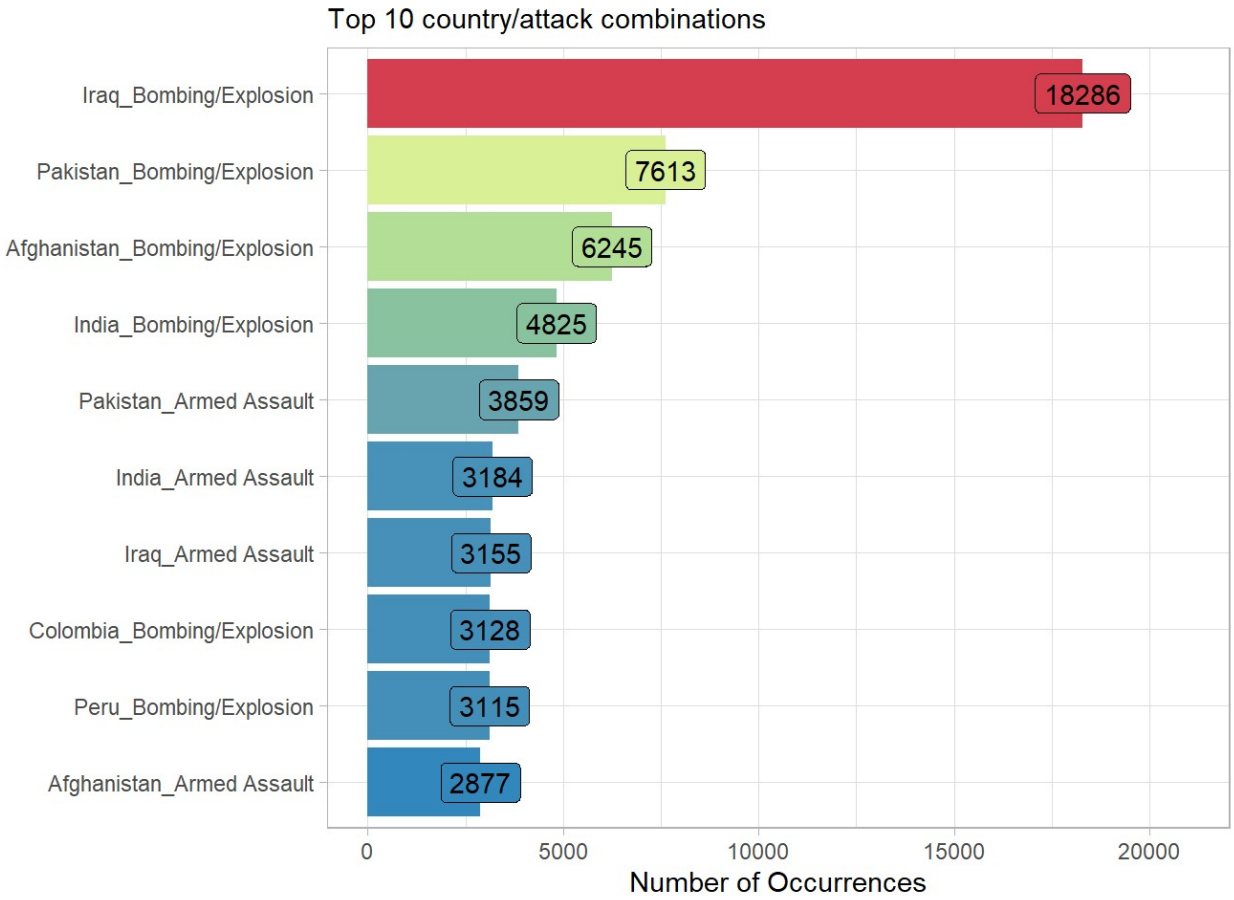
What weapons were used most often?



What were the most common weapon/country combinations?

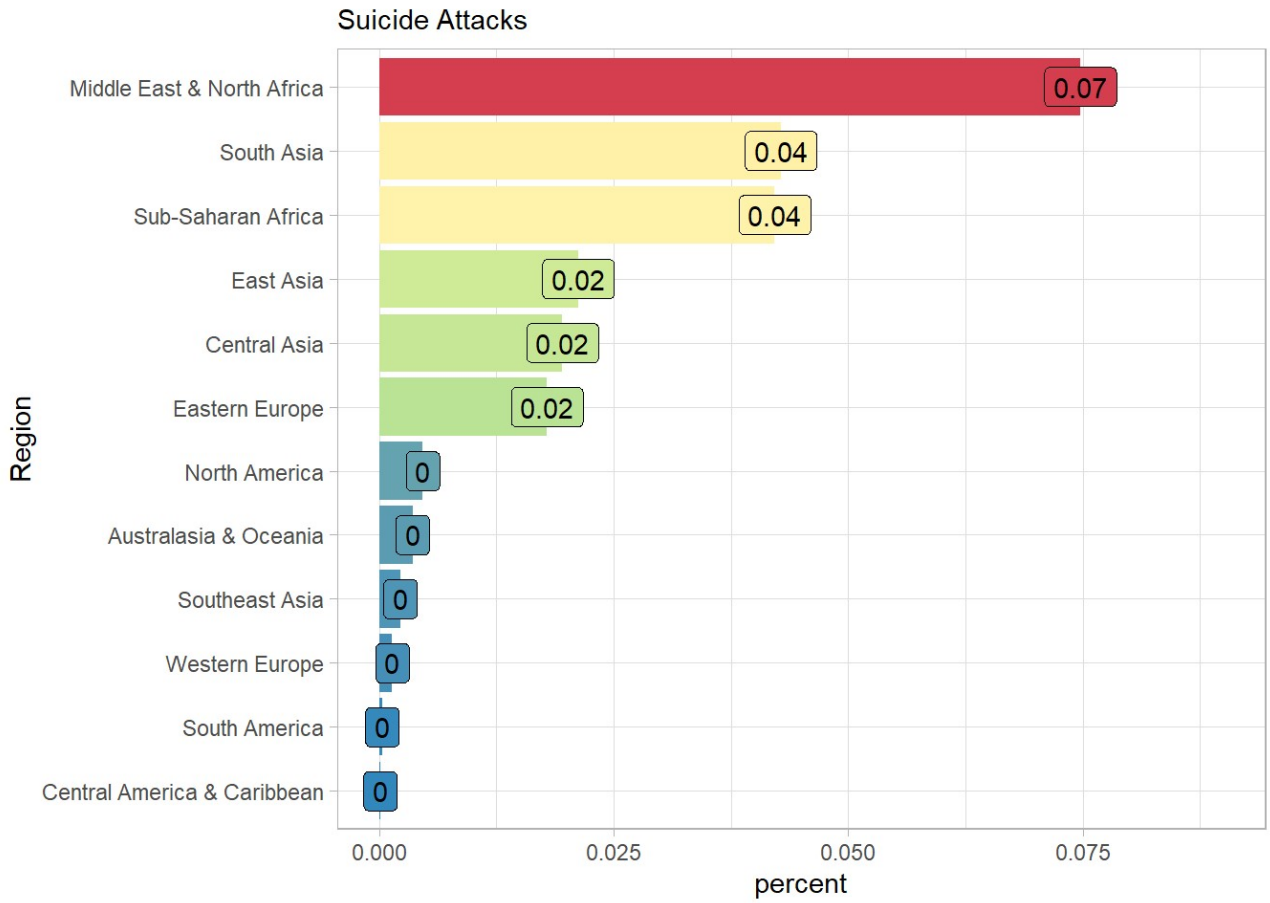


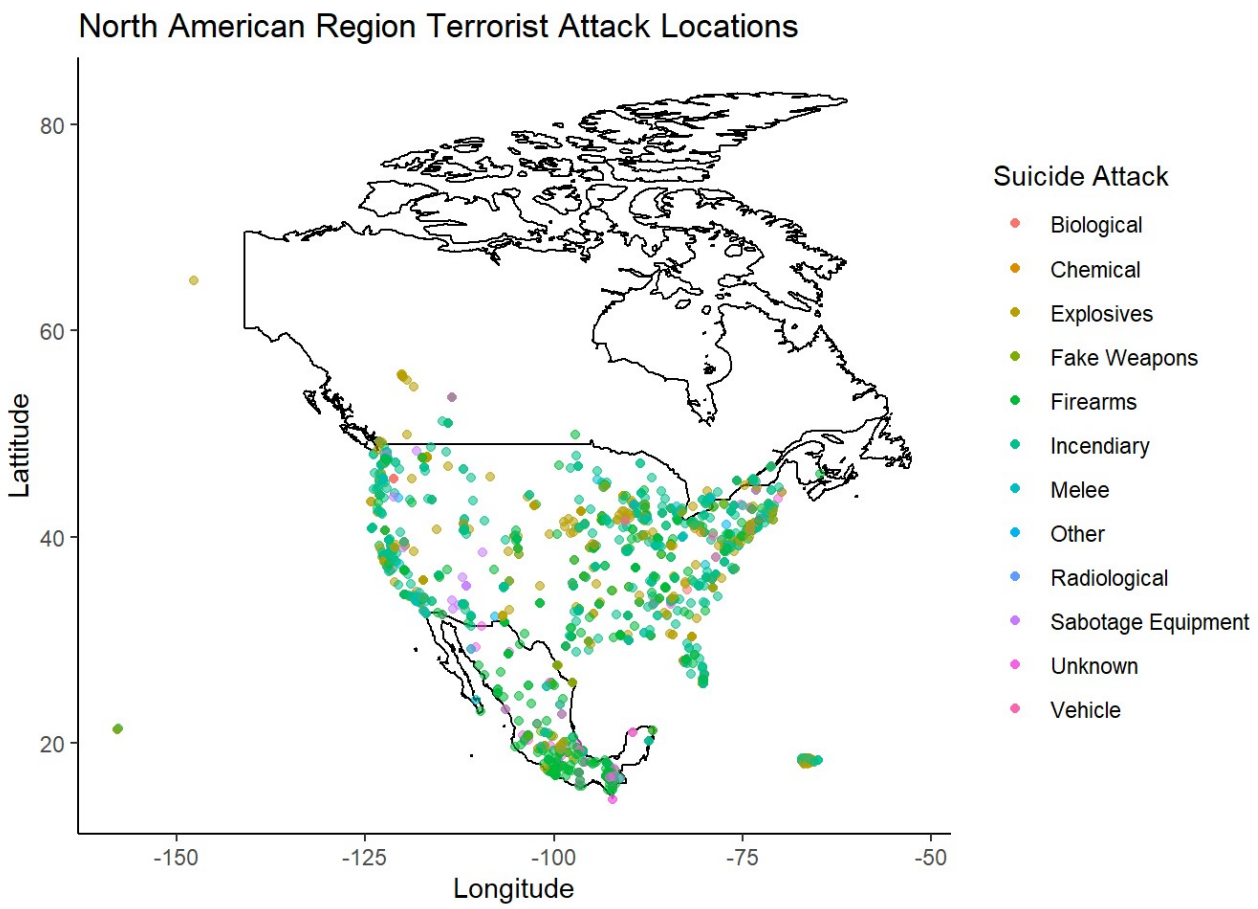
What were the most common attack type/country combinations?

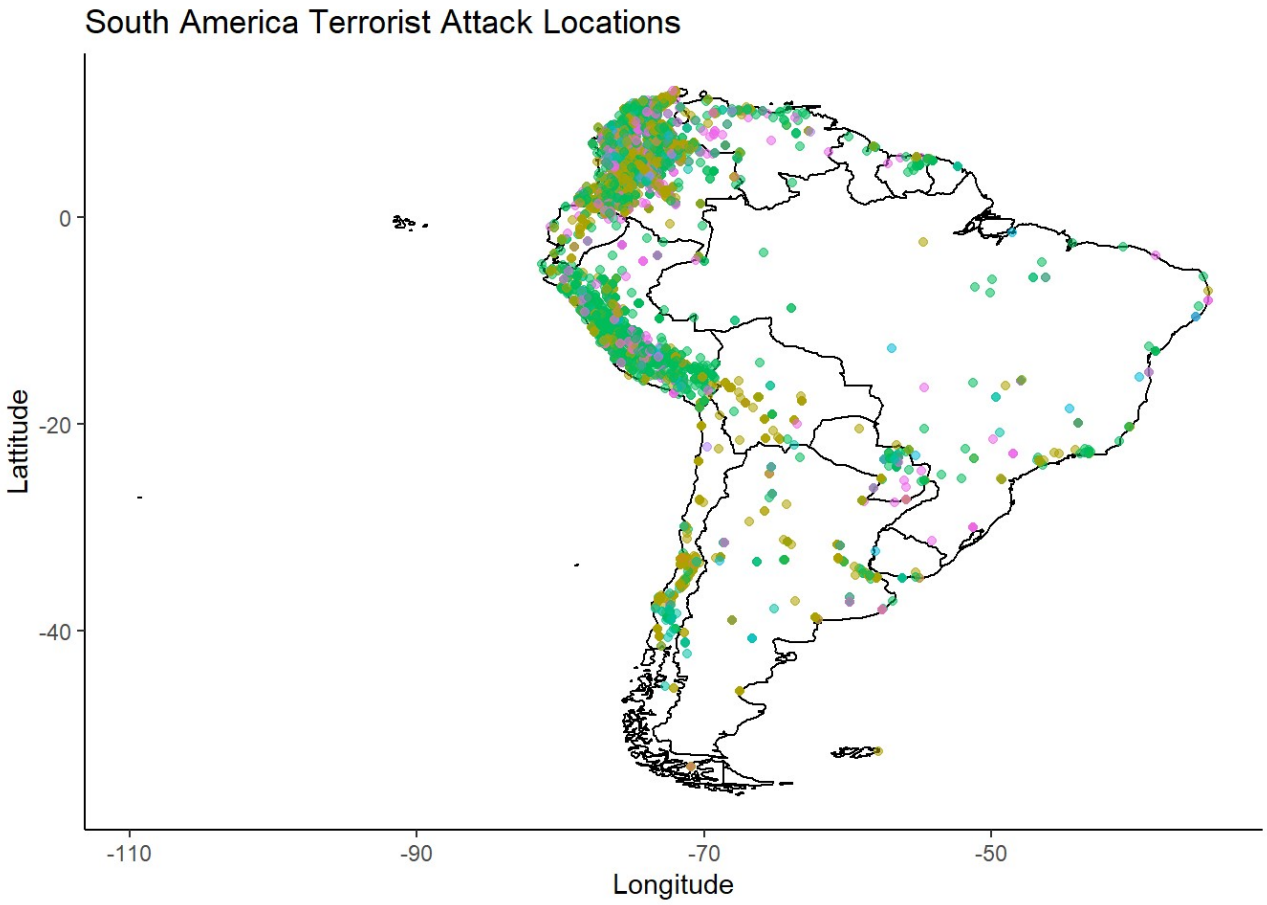


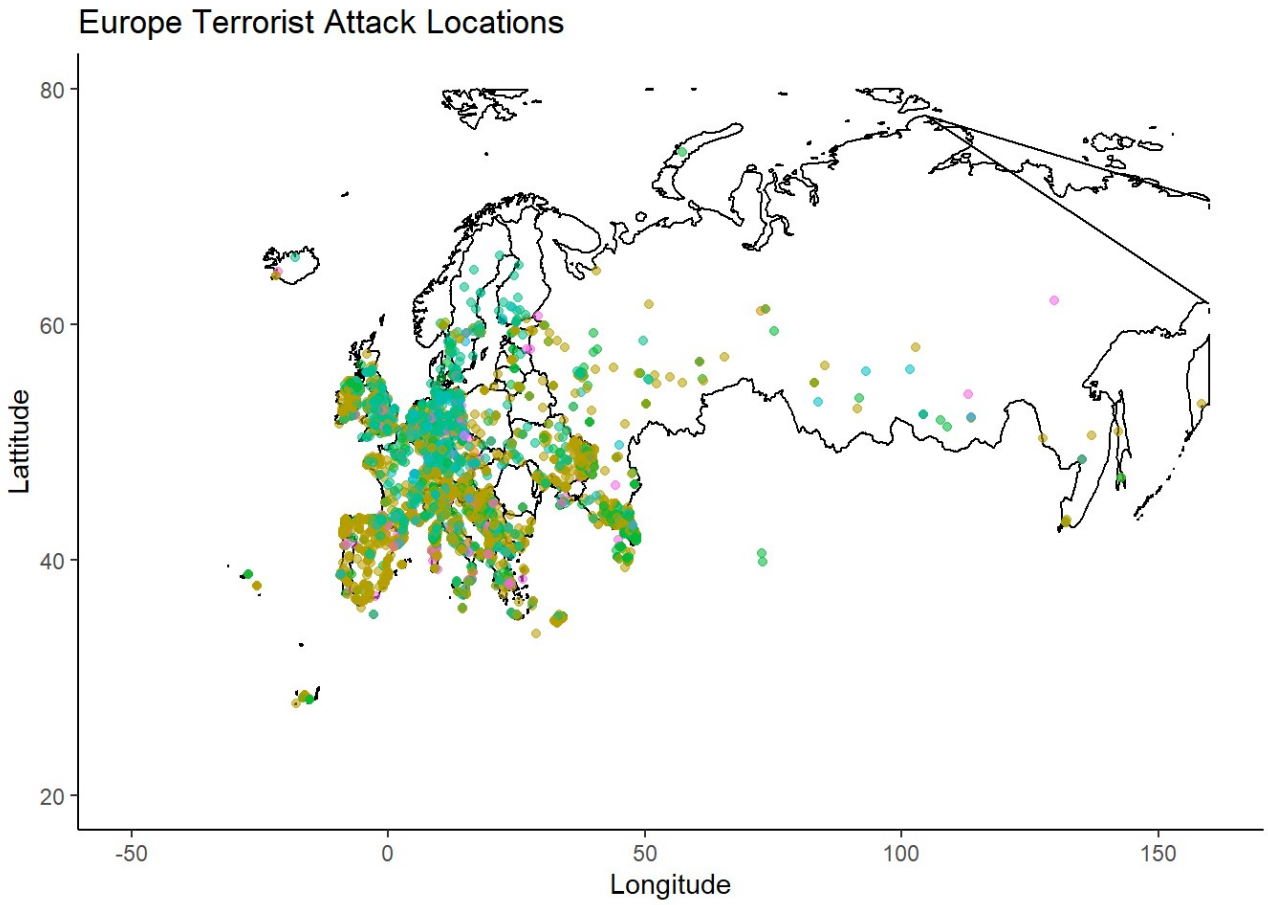
What regions had the highest percent of attacks be suicide attacks?

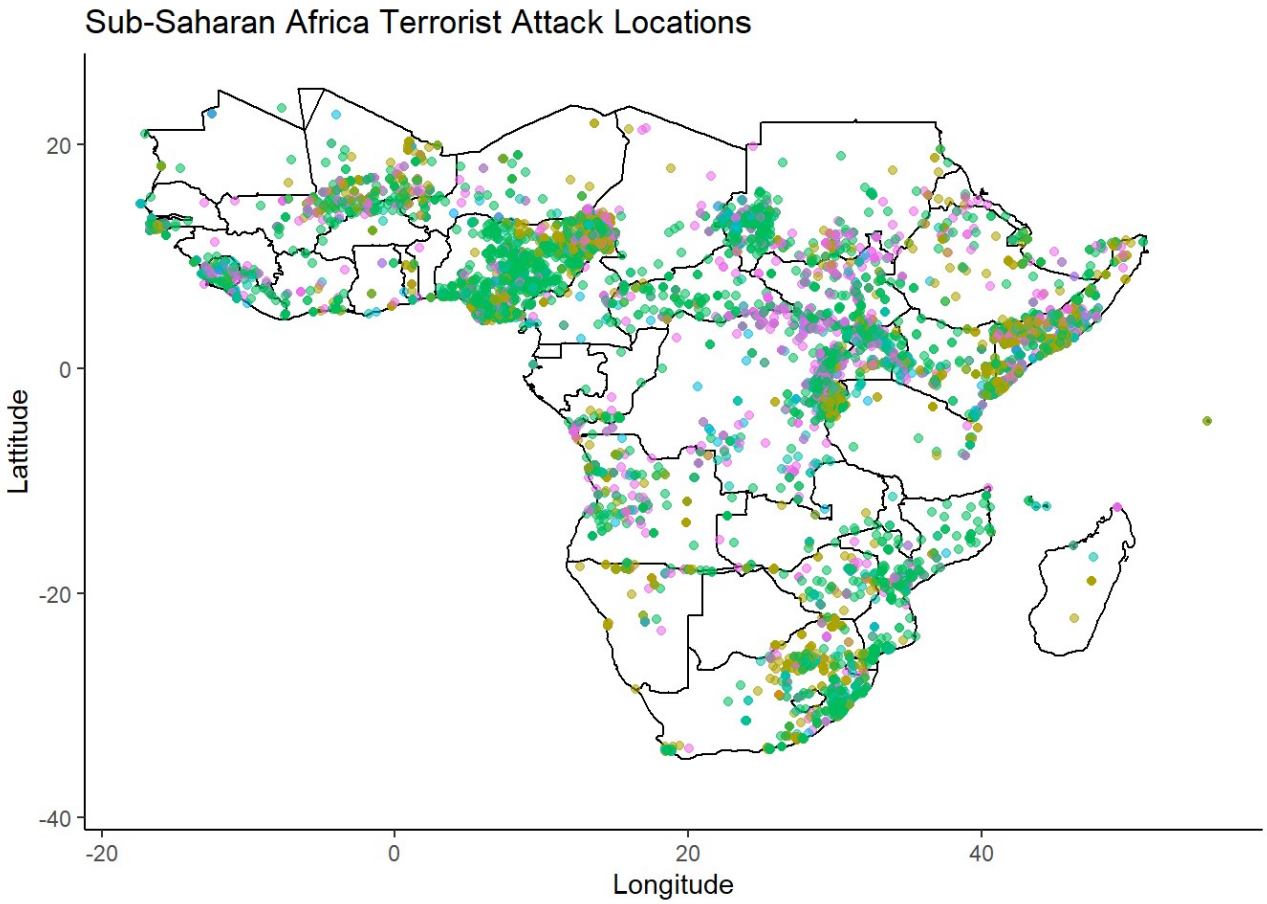
```
## # A tibble: 12 x 4
## # Groups:   region_txt [12]
##    region_txt                  suicidecount Attacks   percent
##    <fct>                              <int>   <int>     <dbl>
##  1 Middle East & North Africa          3772   50474 0.0747
##  2 South Asia                          1926   44974 0.0428
##  3 Sub-Saharan Africa                   740   17550 0.0422
##  4 East Asia                             17     802 0.0212
##  5 Central Asia                          11     563 0.0195
##  6 Eastern Europe                        92    5144 0.0179
##  7 North America                         16    3456 0.00463
##  8 Australasia & Oceania                  1     282 0.00355
##  9 Southeast Asia                        28   12485 0.00224
## 10 Western Europe                        23   16639 0.00138
## 11 South America                          6   18978 0.000316
## 12 Central America & Caribbean            1   10344 0.0000967
```

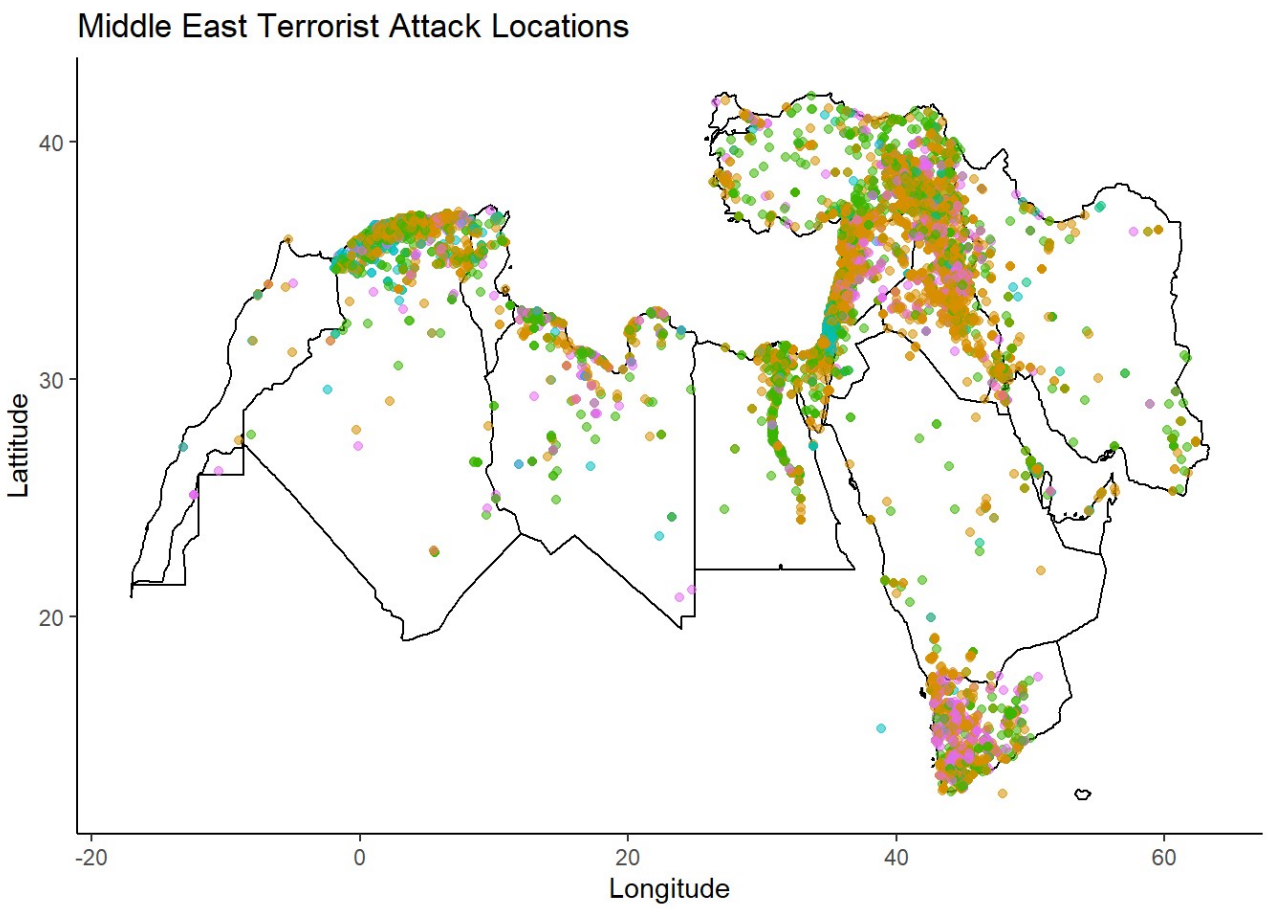


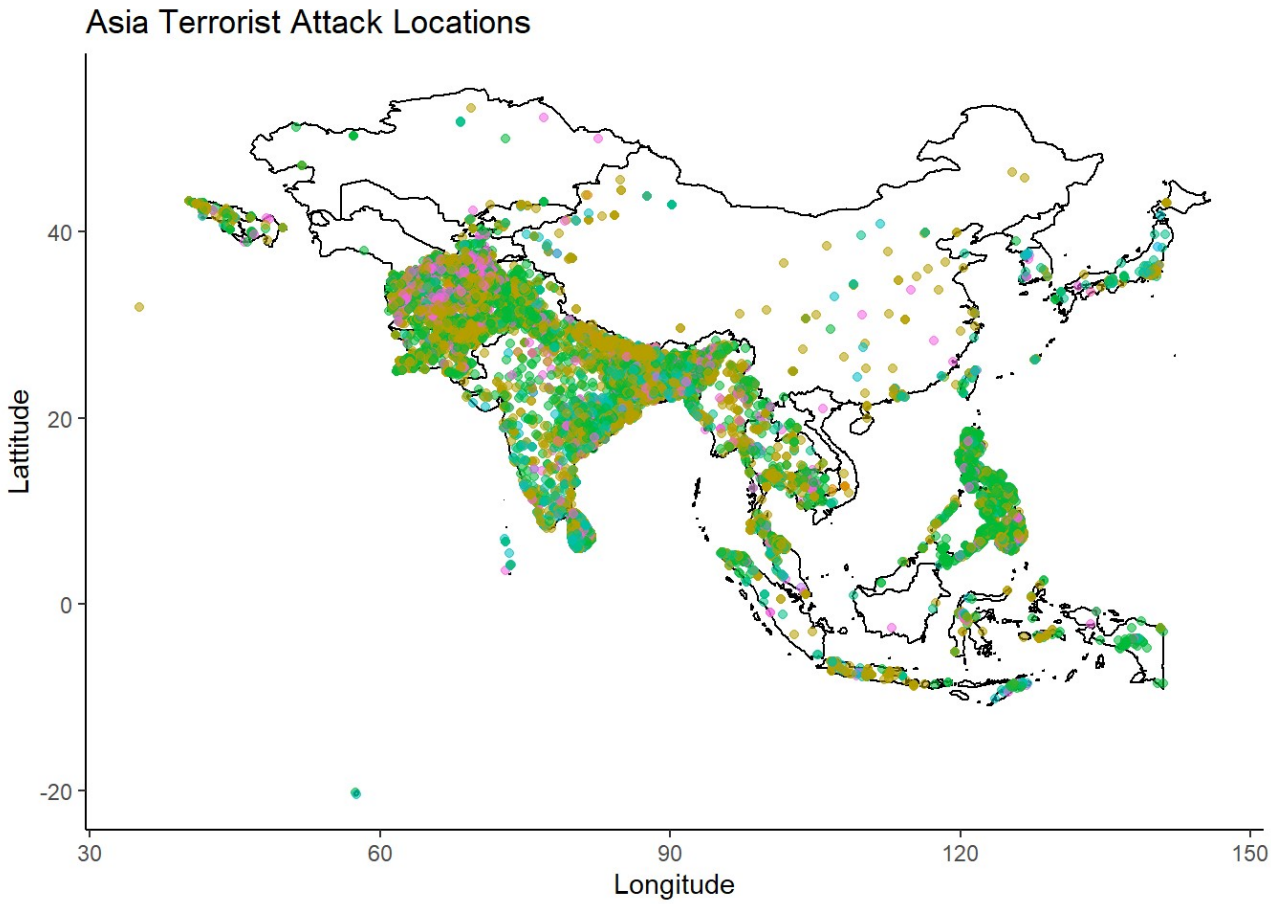

North American Region Terrorist Attack Locations
Lattitude
80
60
40
20
-150
-125
-100
-75
-50
Longitude
Suicide Attack
Biological
Chemical
Explosives
Fake Weapons
Firearms
Incendiary
Melee
Other
Radiological
Sabotage Equipment
Unknown
Vehicle

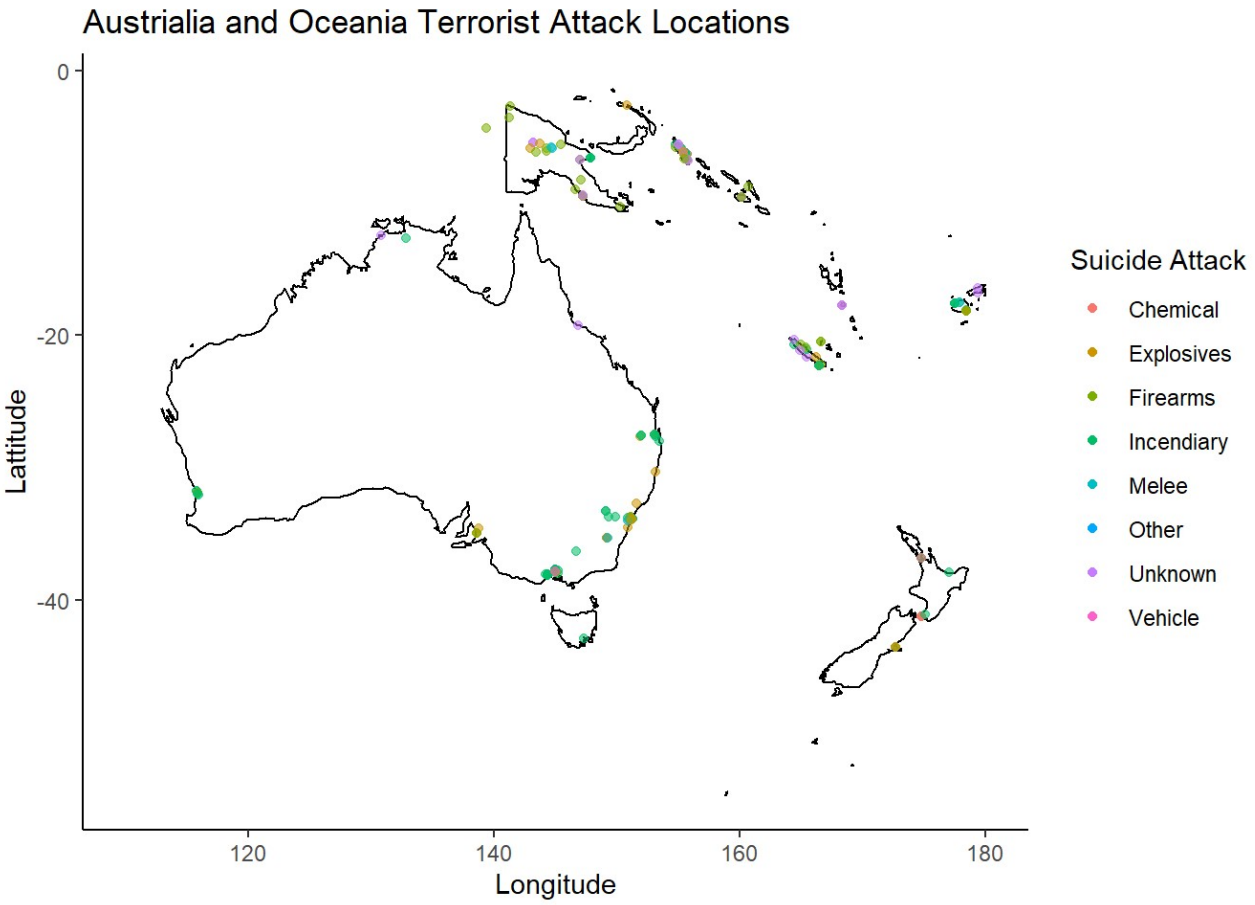

South America Terrorist Attack Locations
0
-20
-40
Lattitude
-110
-90
-70
-50
Longitude


Europe Terrorist Attack Locations
80
60
40
20
Lattitude
-50
0
50
100
150
Longitude


Sub-Saharan Africa Terrorist Attack Locations
Lattitude
20
0
-20
-40
-20
0
20
40
Longitude

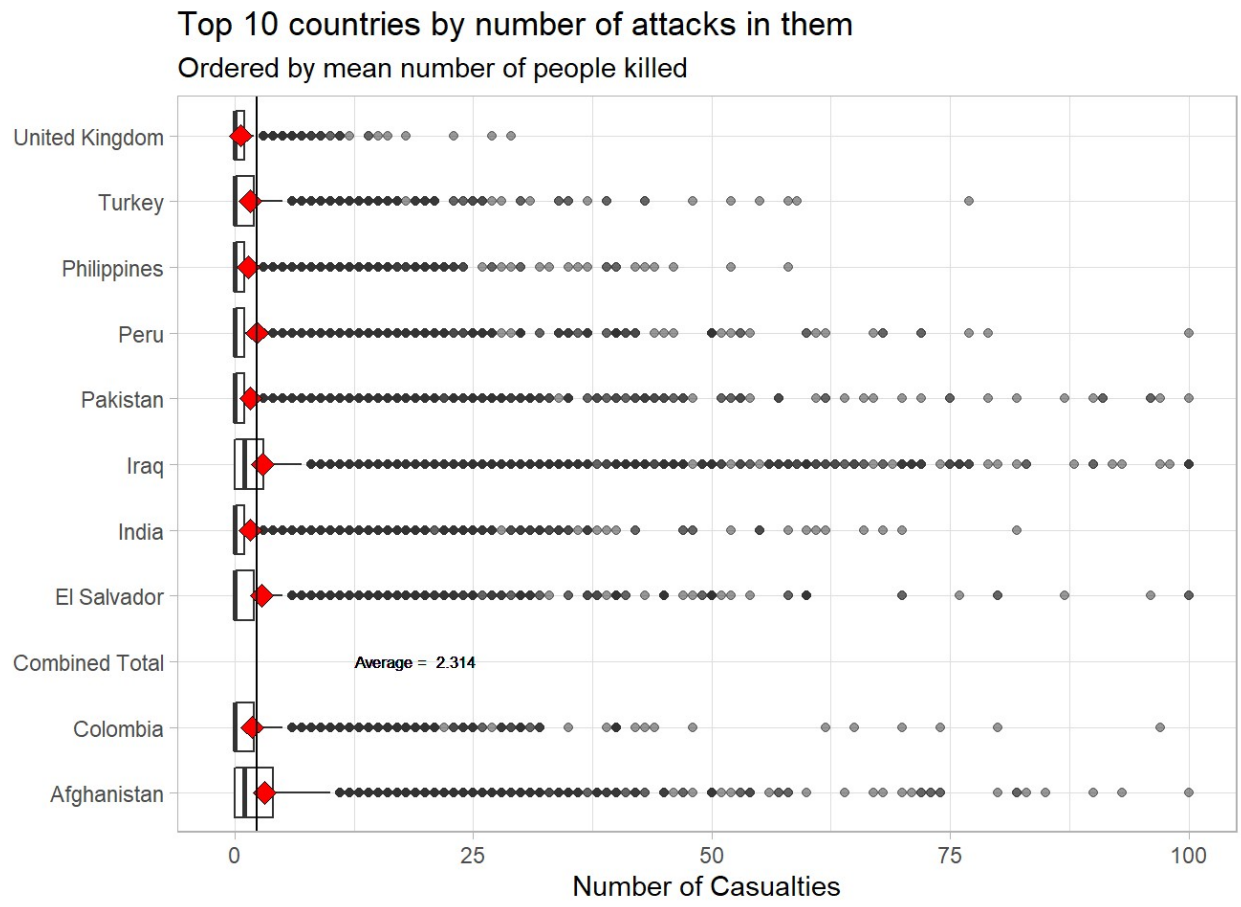

Middle East Terrorist Attack Locations
Lattitude
40
30
20
-20
0
20
40
60
Longitude

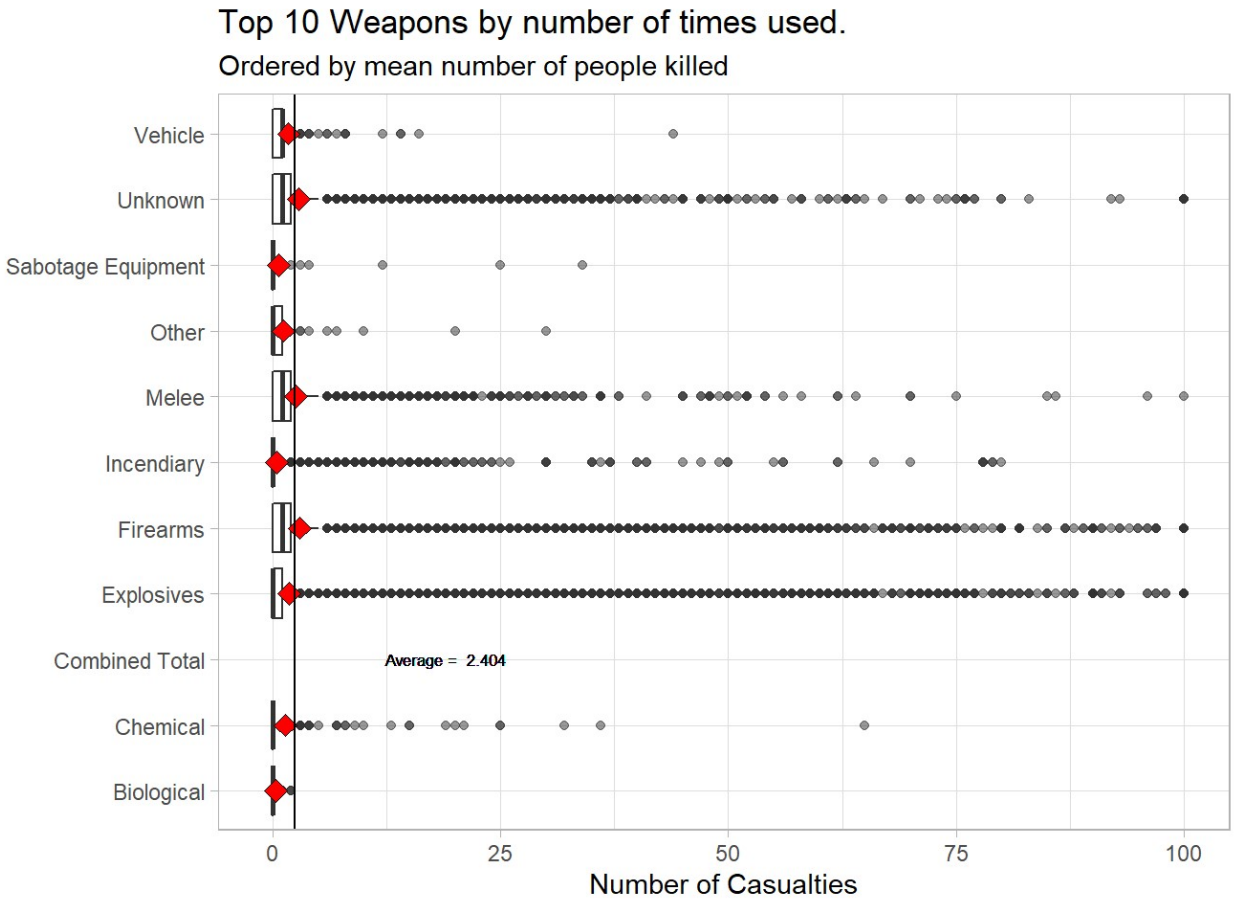

Asia Terrorist Attack Locations
Lattitude
40
20
0
-20
30
60
90
120
150
Longitude

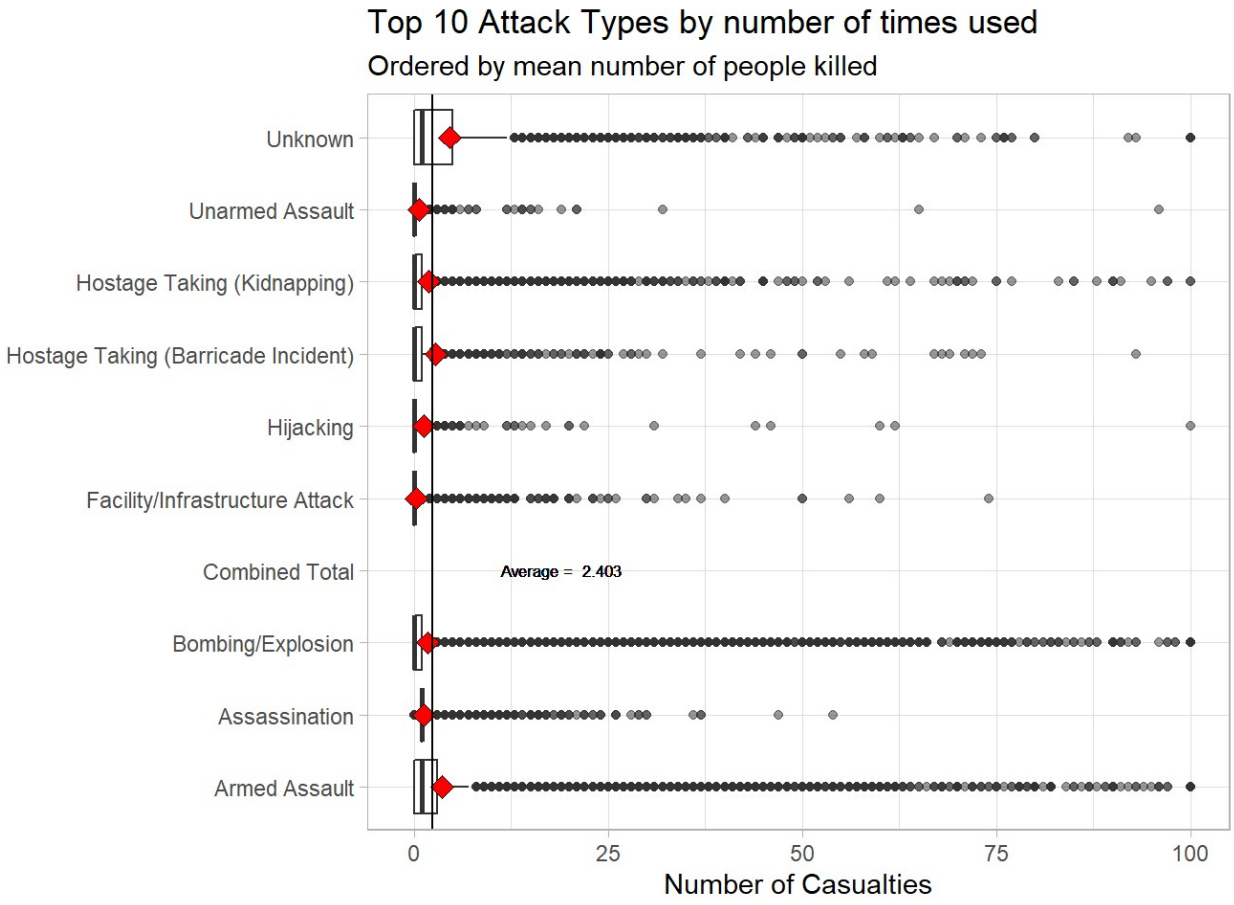

Austrialia and Oceania Terrorist Attack Locations
0
-20
-40
Lattitude
120
140
160
180
Longitude
Suicide Attack
Chemical
Explosives
Firearms
Incendiary
Melee
Other
Unknown
Vehicle

Distributions of number of people killed by region, weapon type and attack type

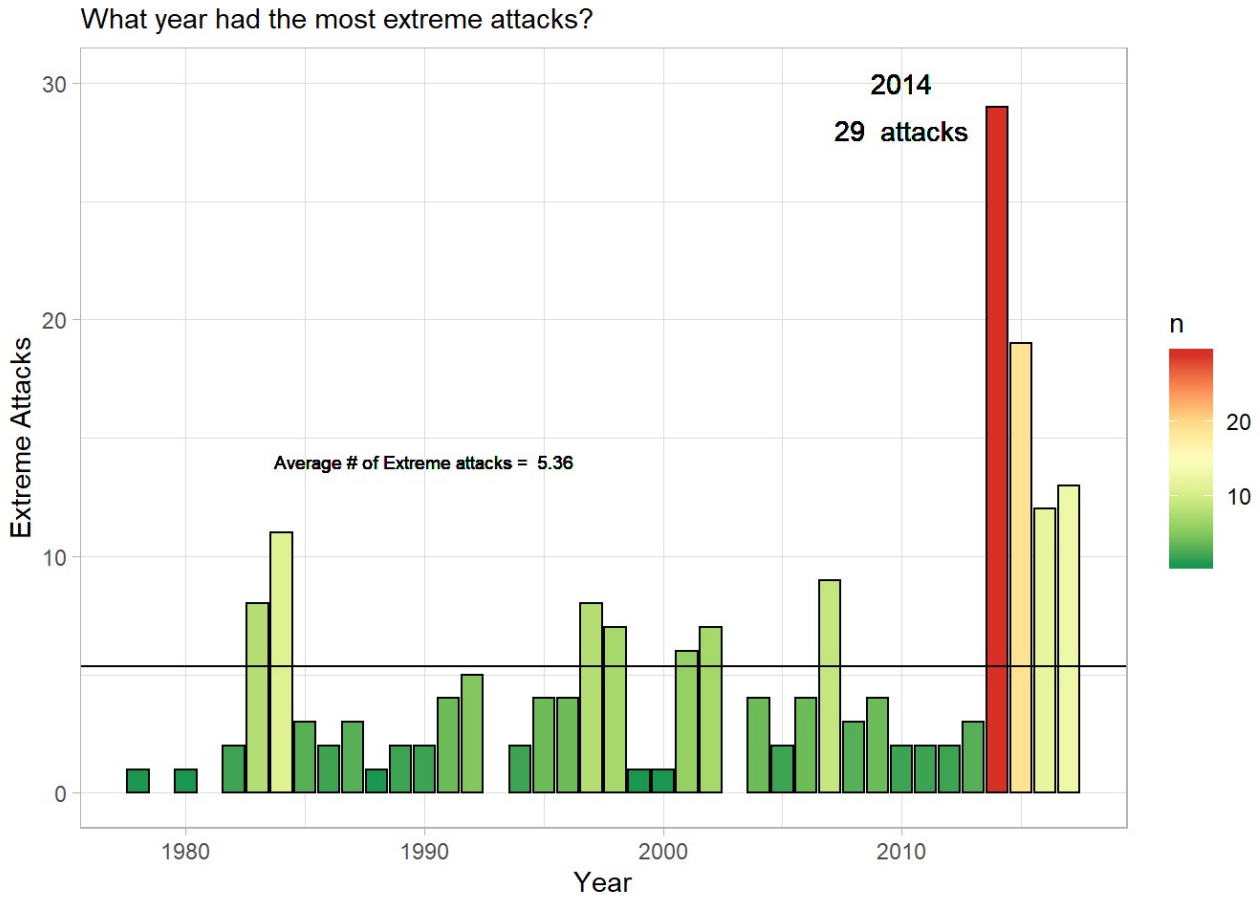
Filtered to not show number of attacks with 100+




Top 10 Weapons by number of times used.
Ordered by mean number of people killed
Vehicle
Unknown
Sabotage Equipment
Other
Melee
Incendiary
Firearms
Explosives
Combined Total
Chemical
Biological
Average = 2.404
0
25
50
75
100
Number of Casualties

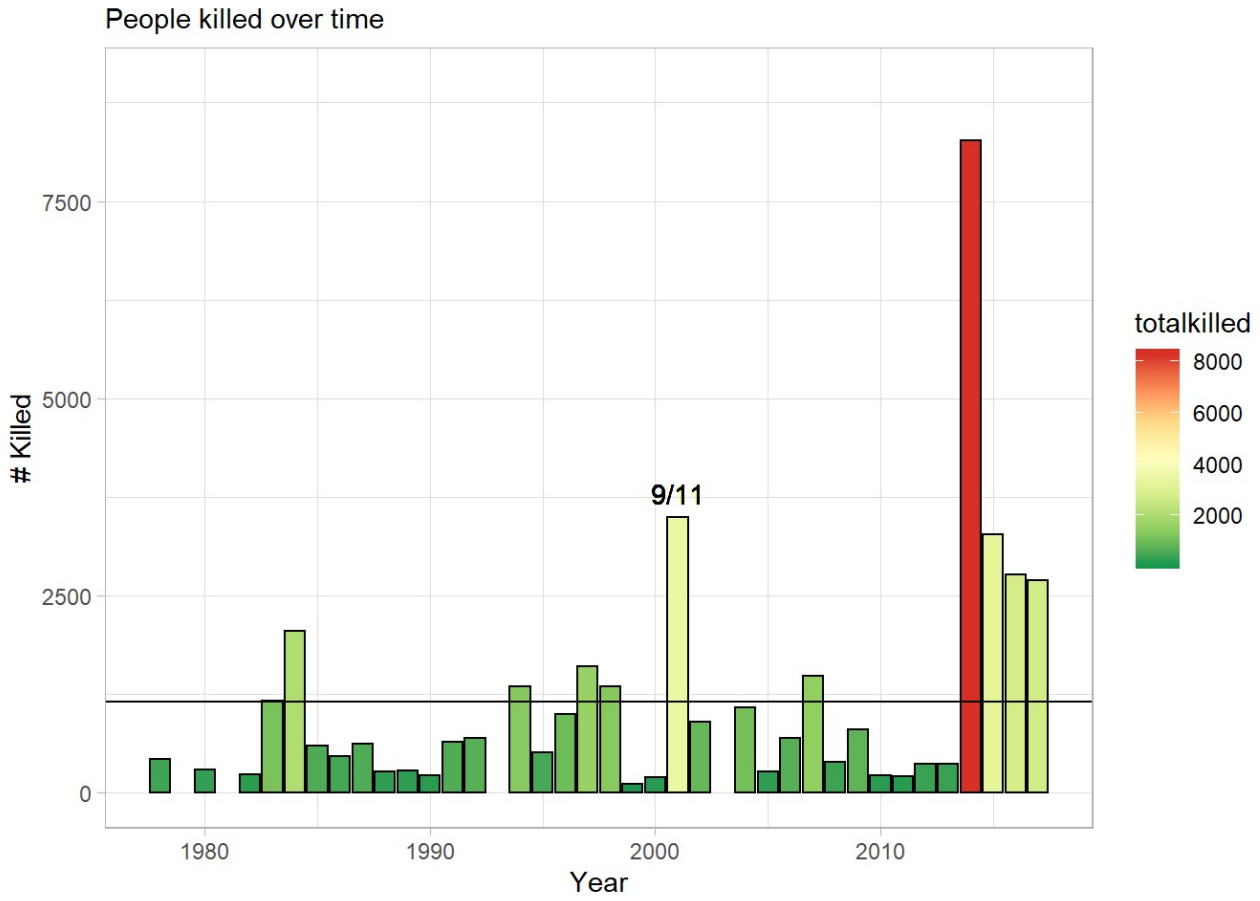

Top 10 Attack Types by number of times used
Ordered by mean number of people killed
Unknown
Unarmed Assault
Hostage Taking (Kidnapping)
Hostage Taking (Barricade Incident)
Hijacking
Facility/Infrastructure Attack
Combined Total
Bombing/Explosion
Assassination
Armed Assault
Average = 2.403
0
25
50
75
100
Number of Casualties

Attacks over time killing over 100 people

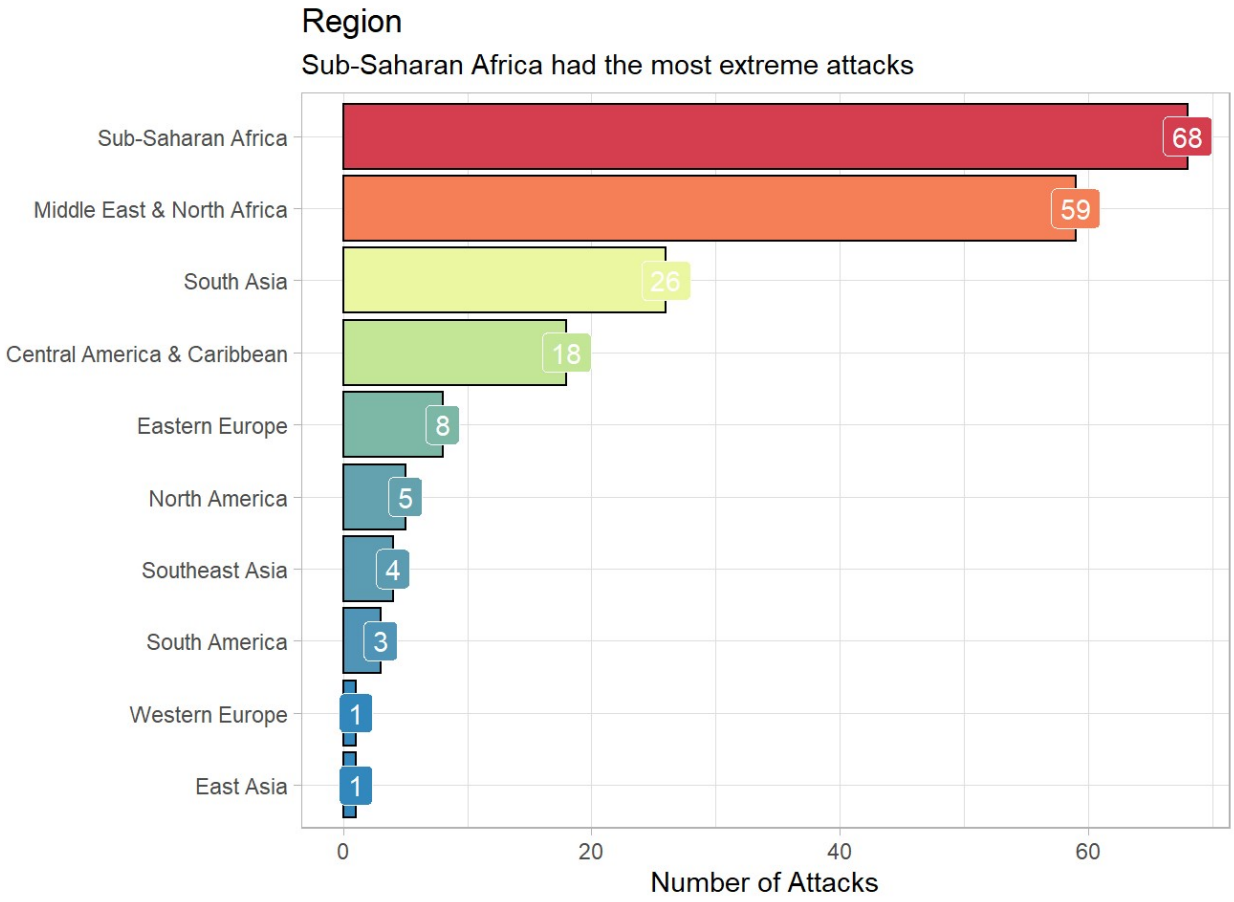


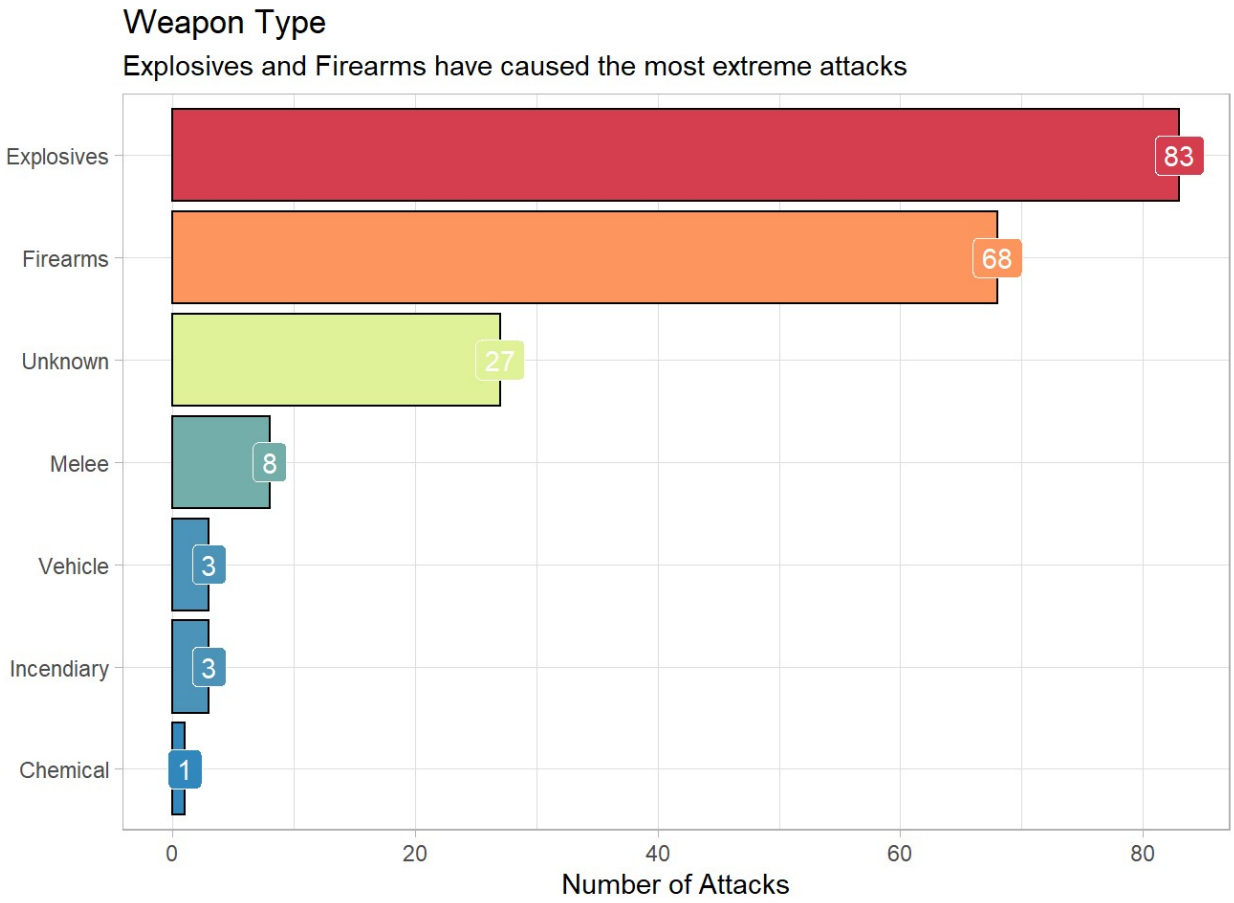
Attacks over time

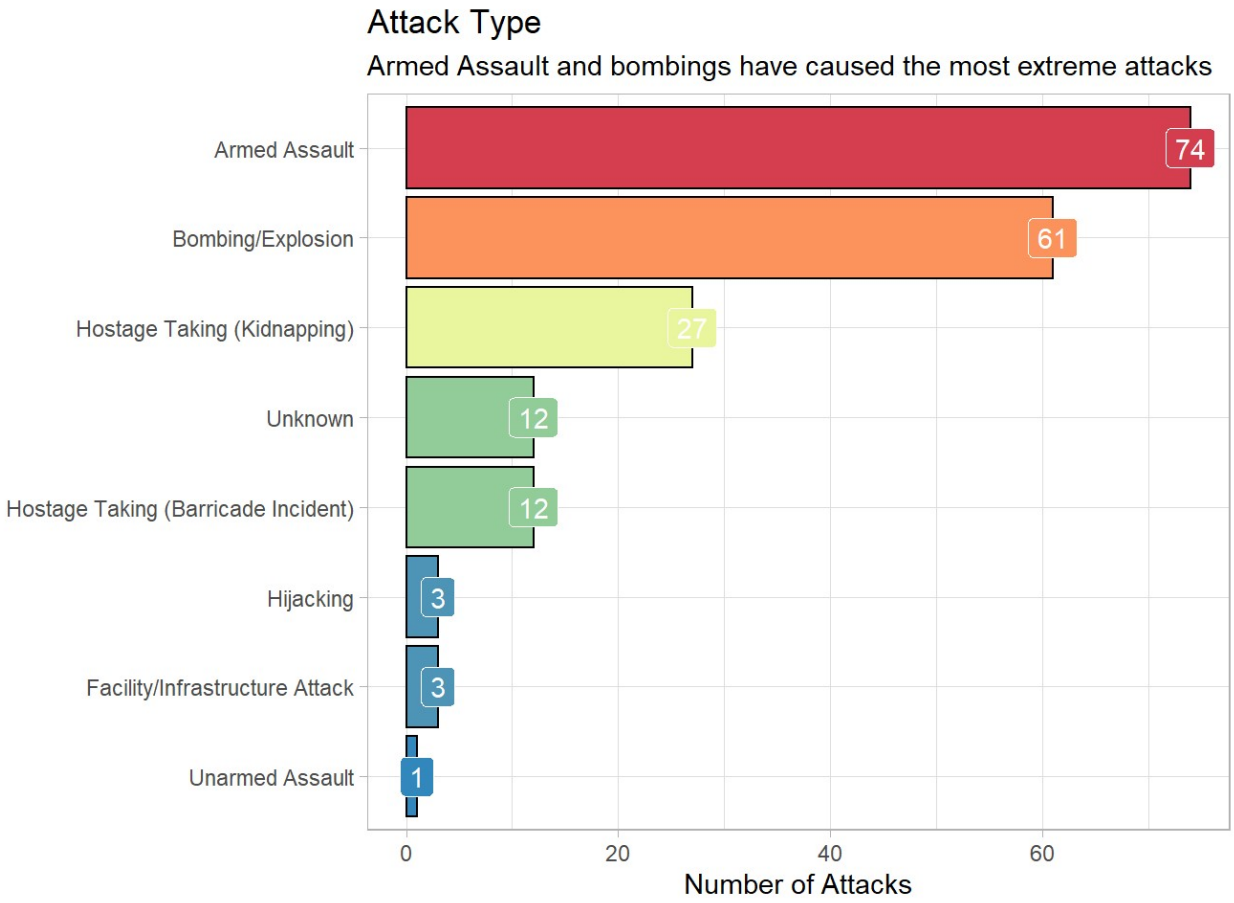
Number of people killed over time.



What regions, weapontypes, and attack types had the most “Extreme Attacks”?

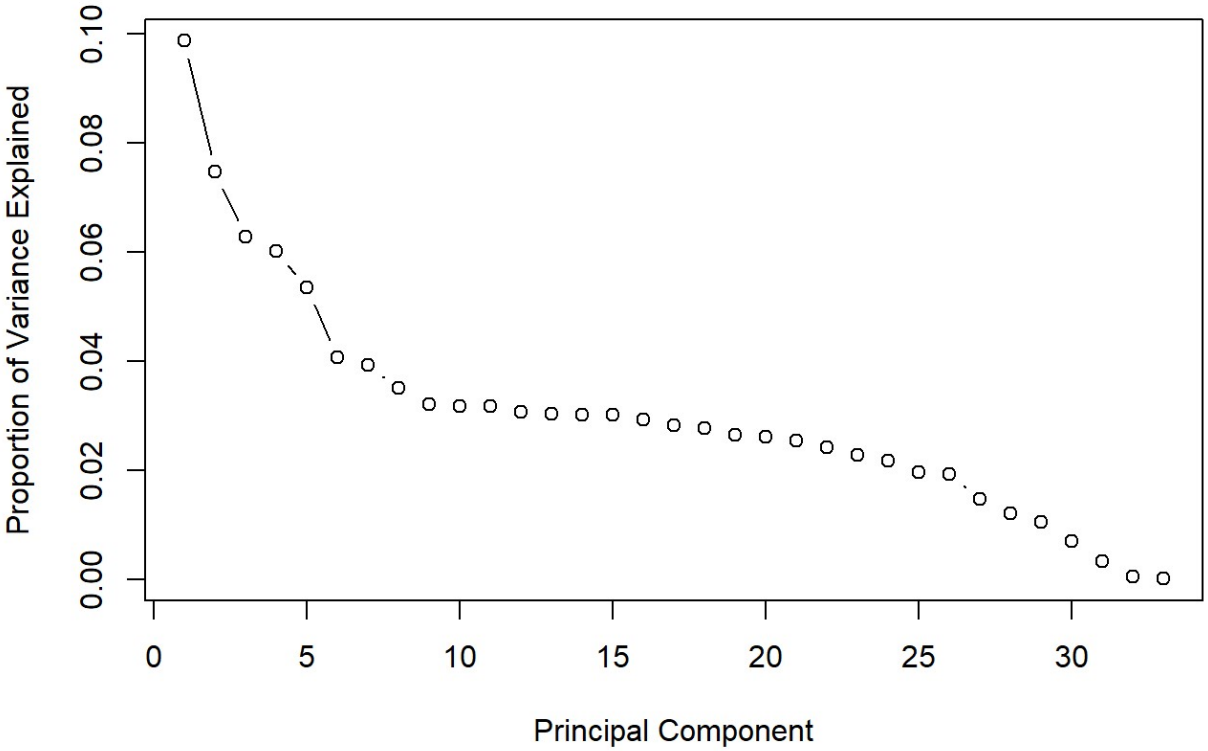


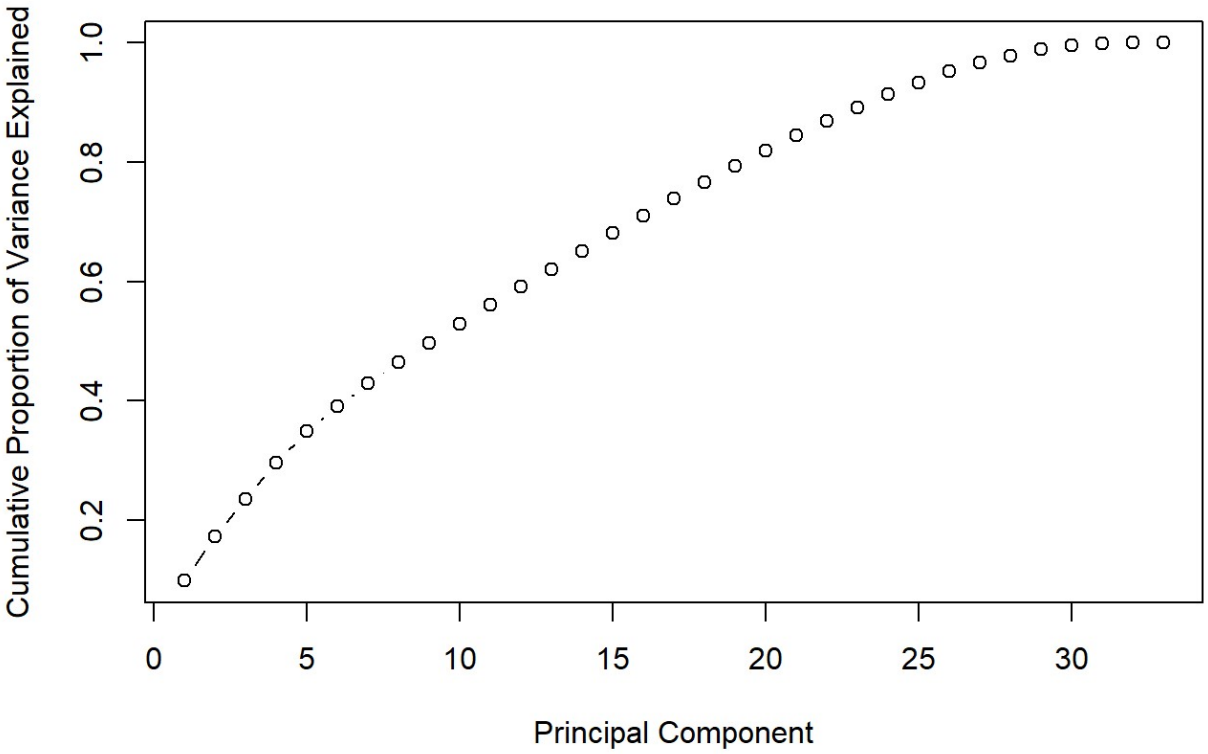

Weapon Type
Explosives and Firearms have caused the most extreme attacks
Explosives
Firearms
Unknown
Melee
Vehicle
Incendiary
Chemical
83
68
27
8
3
3
1
0
20
40
60
80
Number of Attacks



Clustering

I performed two types of clustering examples on this data. One was kmeans clustering and the other was hierarchical. Before I could do this on either of them, I imputed the missing data values. I used the imputation method in the MICE package which can perfrom separate models for imputation. I used the imputation method of “cart” which is classification and regression trees. I had found someone else who had performed a similar impuation using cart. Once I had an imputed dataset, I then used this imputed dataset to for my principal componenent analysis (PCA). I was still dealing with a dataset that did not need to have as many columns as it did. After I performed my PCA, I found that most of my variance was being explained by about 28-29 of my columns. Because of other data decisions, (filtering on <25% missing) I ended up only excluding a handful of columns with the PCA. I had to filter initially first though, because I was running into performance issues without the filtering step.


Proportion of Variance Explained
0.00
0.02
0.04
0.06
0.08
0.10
0
5
10
15
20
25
30
Principal Component



After the imputation and PCA, I then ran the function, clusGap to determine the best number of clusters. I also did a visual check of this with the dendogram that was created by the hclust, and I determined that 8 clusters was the best option. Below is my code for clustering my data and creating two new datasets, one for the kmeans clustered data and one for the hclust data.

Kmeans Clustering

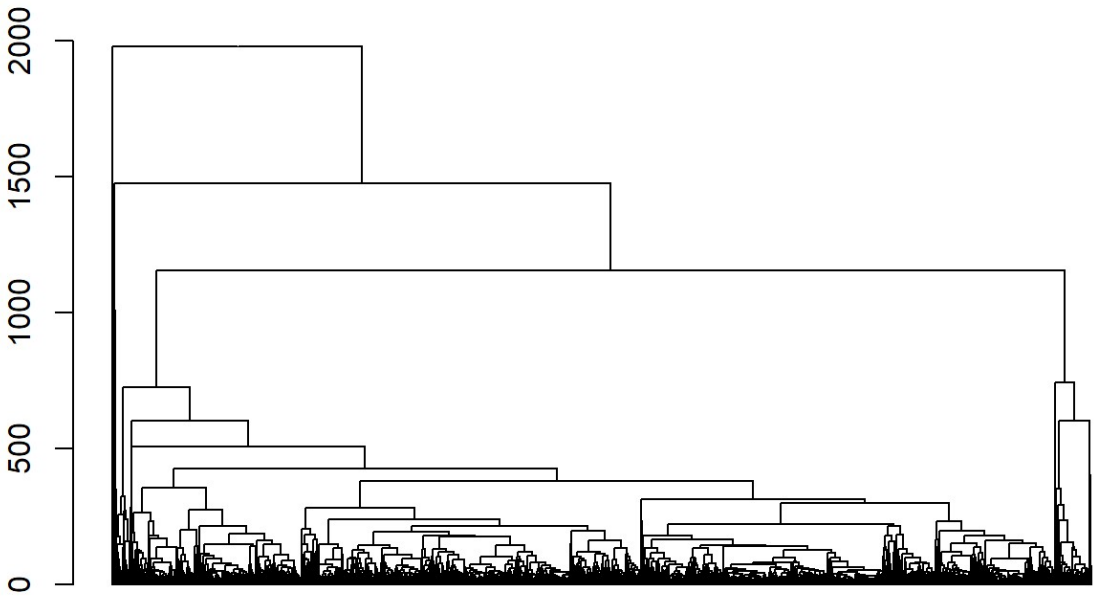
```
sampleTrain <- trainData %>% sample_n(., 20000, replace = TRUE)
sampleTrainNum <- trainData %>% sample_n(., 2000, replace = TRUE)
scaledTrain <- scale(sampleTrain)
scaledTrainNum <- scale(sampleTrainNum)
km.out=kmeans(scaledTrain, 8, nstart = 20)


head(km.out$cluster)
```

```
## [1] 5 7 5 3 7 5
```

```
# fviz_cluster(km.out, data = scaledTrain, geom = "point", pointsize = .5)

# Kmeans groups in dataset
kmeansData <- sampleTrain

kmeansData$Cluster <- km.out$cluster

# Creating a new grouped dataset (hclustData and kmeansData) for data analysis and tra
ining purposes
write.csv(kmeansData, "C:/Users/Ryan Allen/Documents/Regis/Classes/Practicum_I/Data/Da
ta/kmeansData.csv")
```

Hierarchical Clustering

```
trainData <- imputed1[,1:28]
sampleTrain <- trainData %>% sample_n(., 20000, replace = TRUE)
dist <- dist(sampleTrain, method = "euclidean")

hcl <- hclust(dist, method = "complete")

dnd <- as.dendrogram(hcl)
plot(dnd, leaflab = 'none')
```




```
dnd_cut <- cutree(hcl, 8)



hclustData <- sampleTrain
hclustData$Cluster <- dnd_cut
hclustData %>% select(Cluster) %>% add_count(Cluster) %>% unique() %>% arrange(desc
(n))
```

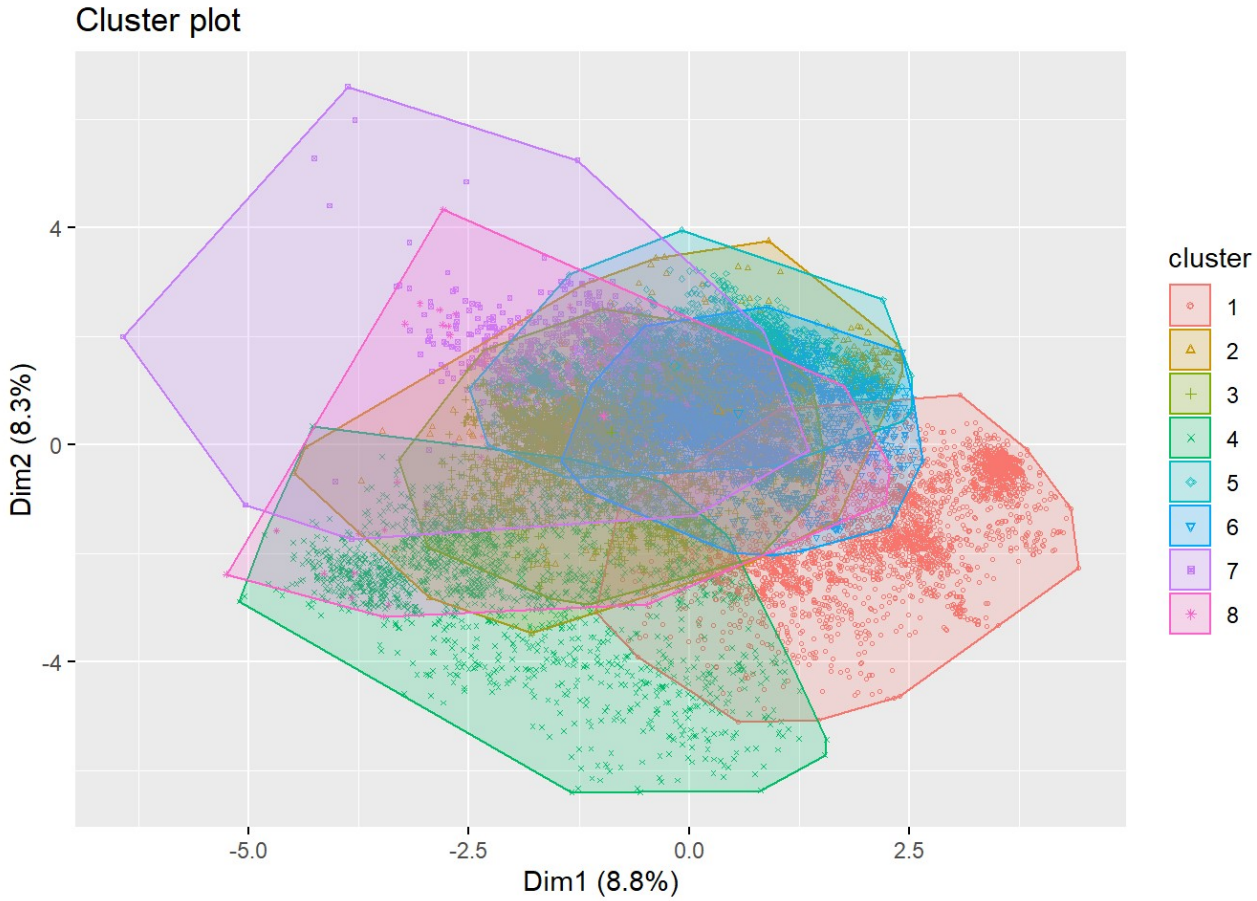
```
## # A tibble: 8 x 2
##   Cluster     n
##     <int> <int>
## 1       1 18873
## 2       2   694
## 3       3   297
## 4       4    41
## 5       7    41
## 6       5    36
## 7       6    17
## 8       8     1
```

```
# Creating a new grouped dataset (hclustData and kmeansData) for data analysis and tra
ining purposes
# write.csv(hclustData, "C:/Users/Ryan Allen/Documents/Regis/Classes/Practicum_I/Data/
Data/hclustData.csv")
```

Cluster Exploratory Data Analysis

Clusters visualized and means of each column grouped by cluster

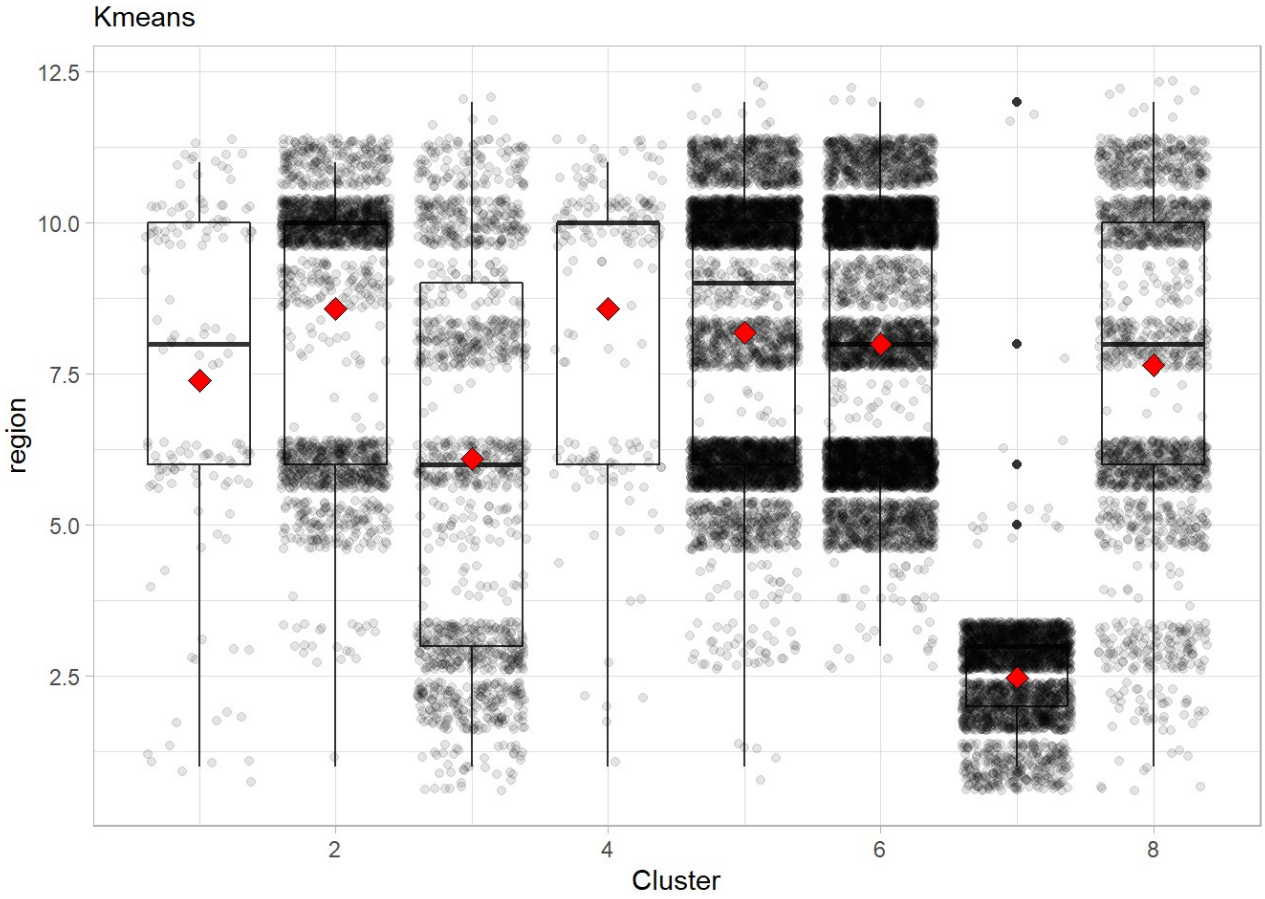
My goal with this section is to determine what separates each cluster from others.

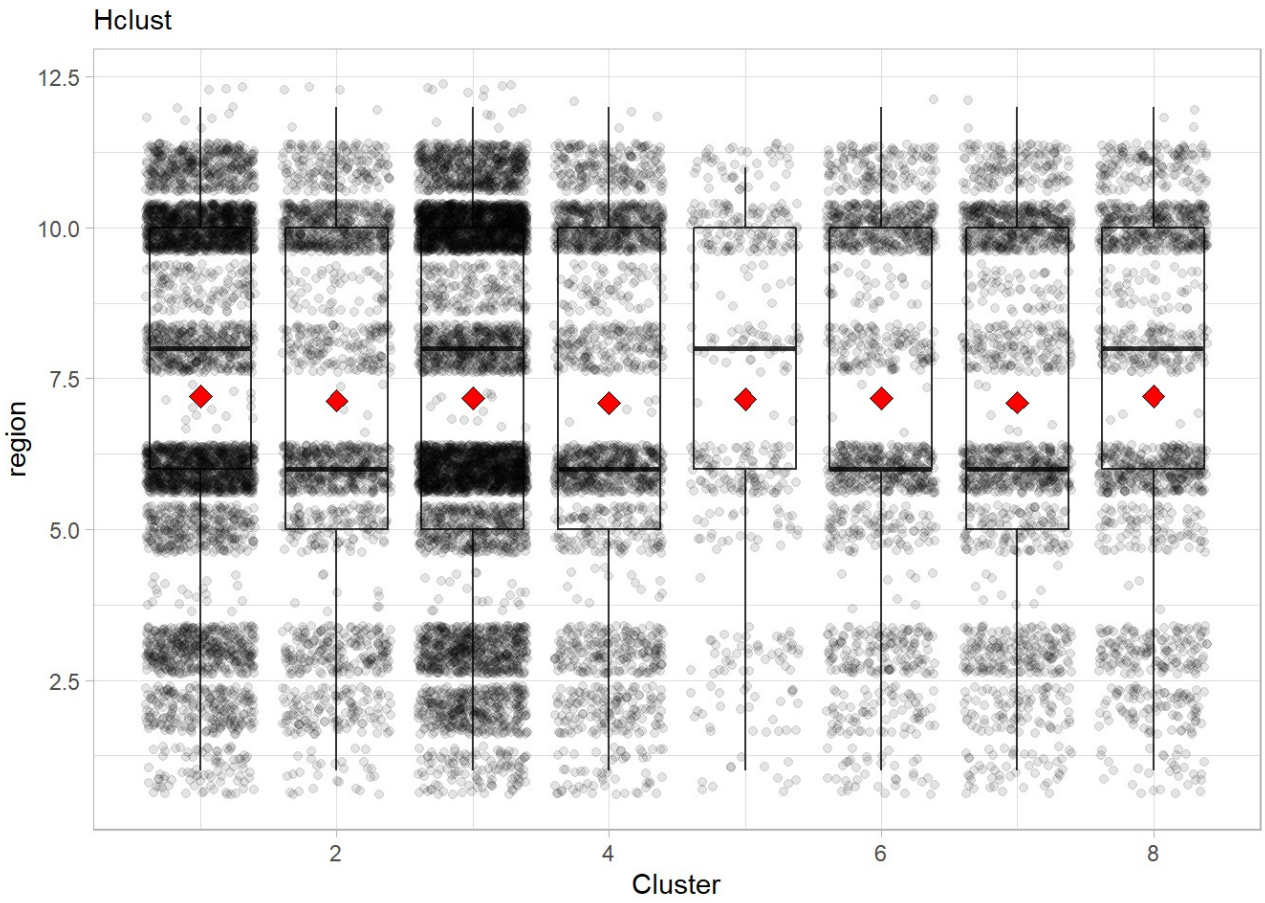


```
##   Group.1         X    iyear   imonth     iday    extended   country
## 1       1  9898.511 1999.460 6.179856 16.90647 0.035971223 172.05755
## 2       2  9956.888 2011.730 6.591954 15.80364 0.014846743 124.74521
## 3       3 10163.430 1988.445 6.706091 15.80099 0.018413598 120.25850
## 4       4  9445.616 2004.198 6.610169 15.61017 0.067796610 122.05085
## 5       5  9928.510 2007.055 6.511540 15.62595 0.040588938 140.96637
## 6       6 10088.475 2004.918 6.425459 15.50342 0.030034236 140.05633
## 7       7 10004.223 1987.878 6.528282 14.90857 0.009605123  99.88332
## 8       8  9875.265 2004.850 6.402915 15.61791 0.202498699 138.05102
##     region  latitude   longitude specificity   vicinity     crit1 crit2
## 1 7.388489 23.107764  35.2503305    1.575540 0.06474820 1.0000000     0
## 2 8.578065 27.690345  51.8645684    1.328544 0.09195402 0.9976054     1
## 3 6.093484 19.287013  -0.8076664    1.308782 0.04815864 0.9985836     1
## 4 8.576271 27.874747  42.2658853    1.254237 0.06214689 0.9943503     1
## 5 8.187823 27.757829  49.1713995    1.461600 0.08734580 0.9775169     1
## 6 7.995176 28.439099  51.0792388    1.432306 0.07454093 0.9940865     1
## 7 2.482746  3.449397 -79.0574337    1.456421 0.03379580 0.9875489     1
## 8 7.642374 24.017244  39.6411256    1.707444 0.06350859 0.9906299     1
##       crit3    doubtterr   multiple   success    suicide attacktype1
## 1 1.0000000  1.000000000 0.10071942 0.9352518 0.02158273    3.223022
## 2 0.8682950  0.161877395 0.16522989 0.9588123 0.06896552    2.698276
## 3 1.0000000 -9.000000000 0.07082153 0.9305949 0.00000000    3.351275
## 4 0.9548023  0.005649718 0.18644068 0.9830508 0.46327684    3.231638
## 5 1.0000000  0.036211699 0.14862714 0.8267012 0.03342618    2.952646
## 6 0.7695300  0.256924992 0.10846561 0.8750389 0.05073140    2.583256
## 7 0.8203486  0.205976521 0.16115261 0.9345429 0.00000000    2.791889
## 8 0.8542426 -0.277980219 0.18219677 0.8927642 0.00000000    7.181676
##   targtype1 targsubtype1  natlty1 guncertain1   individual weaptype1
## 1 16.683453     88.96403 147.1942  0.07913669 0.0000000000  6.870504
## 2  8.138410     46.66523 130.0014  0.10009579 0.0000000000  5.748084
## 3  7.637394     43.19830 118.0907  0.01274788 0.0000000000  7.068697
## 4  8.966102     51.25424 129.4520  0.12429379 0.0000000000  6.169492
## 5 15.870076     83.75249 126.5895  0.09132511 0.0023875846  5.868484
## 6  3.207127     24.54373 137.4143  0.07998755 0.0006224712  5.656240
## 7  8.123444     46.79189 104.4842  0.08182142 0.0124510850  5.821060
## 8  7.124414     42.04060 135.1223  0.10255075 0.0010411244 11.663196
##   weapsubtype1      nkill     nwound    property Cluster
## 1     10.55396  3.0143885  3.0071942 -0.48201439       1
## 2     11.09148  2.4693487  4.5043103 -9.00000000       2
## 3     13.20963  0.9582153  0.6458924  0.68130312       3
## 4     13.38418 48.3559322 89.4576271 -0.02259887       4
## 5     12.26641  2.0005969  2.5662555  0.51153999       5
## 6     10.34329  1.7324930  2.2637722  0.54699658       6
## 7     10.65564  2.2426183  0.9807898  0.75880470       7
## 8     21.81780  2.1832379  1.1244144 -0.37844872       8
```

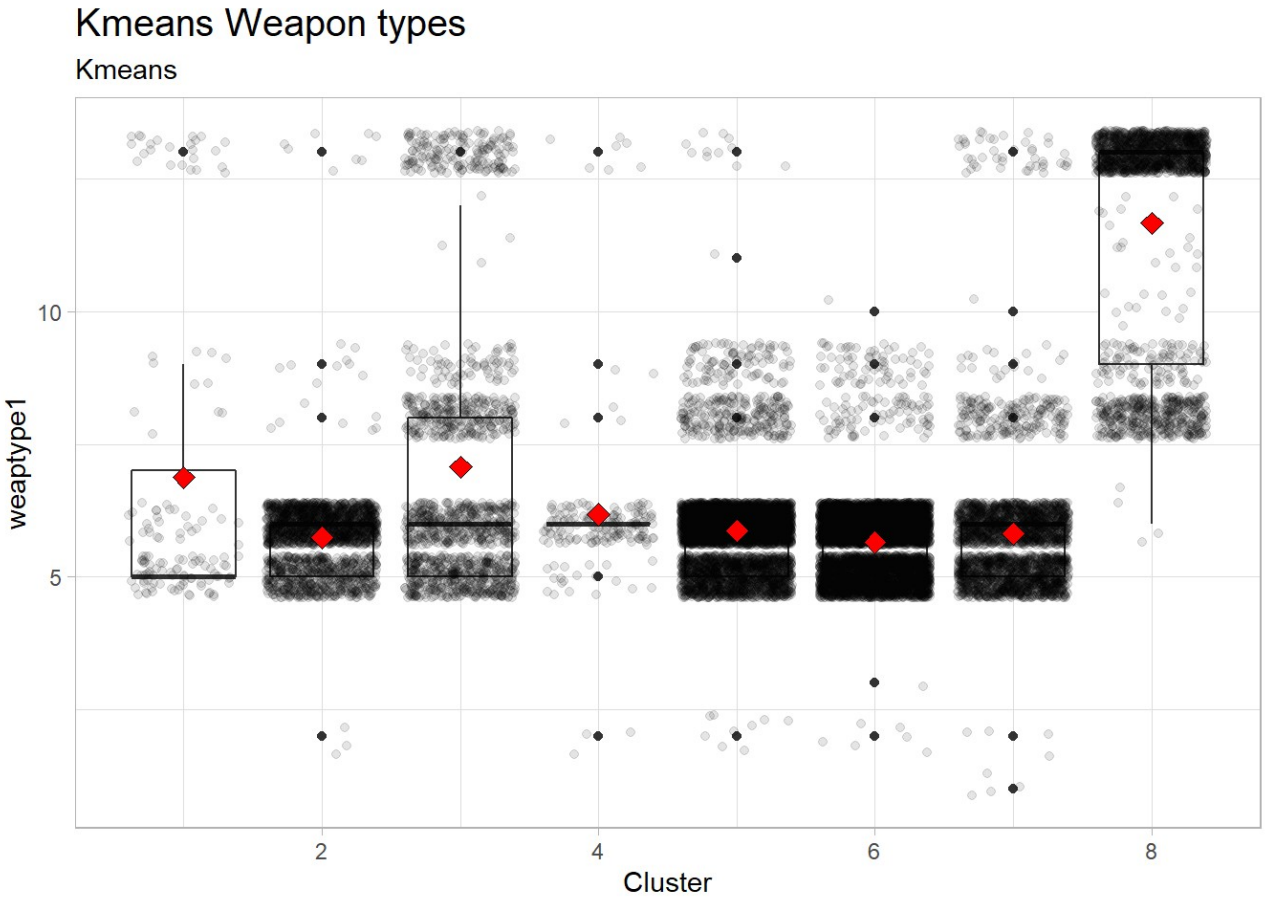
```
##   Group.1       X.1        X      eventid    iyear   imonth     iday
## 1       1  9965.588 72514.00 200265774729 2002.592 6.436733 15.33988
## 2       2  9835.646 72678.03 200271766347 2002.652 6.400936 15.72439
## 3       3 10047.309 72535.67 200268376681 2002.617 6.439182 15.37925
## 4       4  9860.722 71474.30 200228358186 2002.217 6.451923 15.69048
## 5       5  9525.329 73218.58 200278650886 2002.720 6.495556 15.52889
## 6       6 10338.738 73561.72 200296619006 2002.899 6.524983 15.37175
## 7       7  9878.094 72971.62 200279891933 2002.732 6.403670 15.60608
## 8       8 10270.562 70933.28 200210581792 2002.041 6.325828 15.10066
##     extended  country   region latitude longitude specificity   vicinity
## 1 0.04440973 133.9645 7.205311 23.53660  28.28886    1.446775 0.07141263
## 2 0.04420177 131.8794 7.134685 22.78333  26.57818    1.499220 0.05668227
## 3 0.04108265 132.5172 7.179314 23.53483  27.27061    1.455615 0.06895441
## 4 0.04349817 129.5325 7.090201 23.19336  26.06551    1.452839 0.06959707
## 5 0.05111111 128.4778 7.164444 24.35419  26.65515    1.482222 0.07777778
## 6 0.04063957 136.2738 7.171885 23.82034  28.21011    1.468354 0.06062625
## 7 0.03899083 135.7391 7.102638 23.69529  26.47525    1.397936 0.06651376
## 8 0.03907285 129.8828 7.211921 23.55871  25.15713    1.416556 0.06953642
##       crit1     crit2     crit3  doubtterr  multiple   success    suicide
## 1 0.9877260 0.9915198 0.8792680 -0.4896229 0.1330060 0.8895336 0.03793796
## 2 0.9901196 0.9921997 0.8772751 -0.4440978 0.1393656 0.8902756 0.03952158
## 3 0.9890446 0.9945223 0.8751410 -0.4976639 0.1400032 0.8804575 0.03351055
## 4 0.9935897 0.9903846 0.8859890 -0.6396520 0.1456044 0.8827839 0.02793040
## 5 0.9866667 0.9911111 0.8622222 -0.3466667 0.1333333 0.8822222 0.04222222
## 6 0.9913391 0.9960027 0.8767488 -0.4503664 0.1372418 0.8847435 0.03397735
## 7 0.9902523 0.9919725 0.8755734 -0.4615826 0.1353211 0.8887615 0.04185780
## 8 0.9894040 0.9920530 0.8715232 -0.5397351 0.1317881 0.8927152 0.03774834
##   attacktype1 targtype1 targsubtype1  natlty1 guncertain1  individual
## 1    3.255077  8.445436     48.62062 129.5653  0.07565276 0.002454809
## 2    3.241290  8.327093     48.09048 126.7171  0.07488300 0.002080083
## 3    3.208635  8.306267     47.93185 127.3963  0.08474303 0.001933301
## 4    3.343864  8.584707     49.28663 129.9913  0.07829670 0.003663004
## 5    3.377778  8.486667     48.61778 126.4244  0.07555556 0.000000000
## 6    3.224517  8.208528     47.55230 129.9487  0.07794803 0.003997335
## 7    3.193234  8.280963     47.64966 130.7385  0.08142202 0.004013761
## 8    3.254967  8.628477     49.66623 128.2219  0.07615894 0.001986755
##   weaptype1 weapsubtype1    nkill Cluster
## 1  6.466860     12.38630 2.437626       1
## 2  6.427457     12.12845 2.322413       2
## 3  6.442726     12.32447 2.444659       3
## 4  6.580586     12.65751 1.974359       4
## 5  6.591111     12.62000 2.740000       5
## 6  6.435710     12.12791 2.306462       6
## 7  6.400803     12.38360 2.168578       7
## 8  6.417881     12.26755 2.227152       8
```

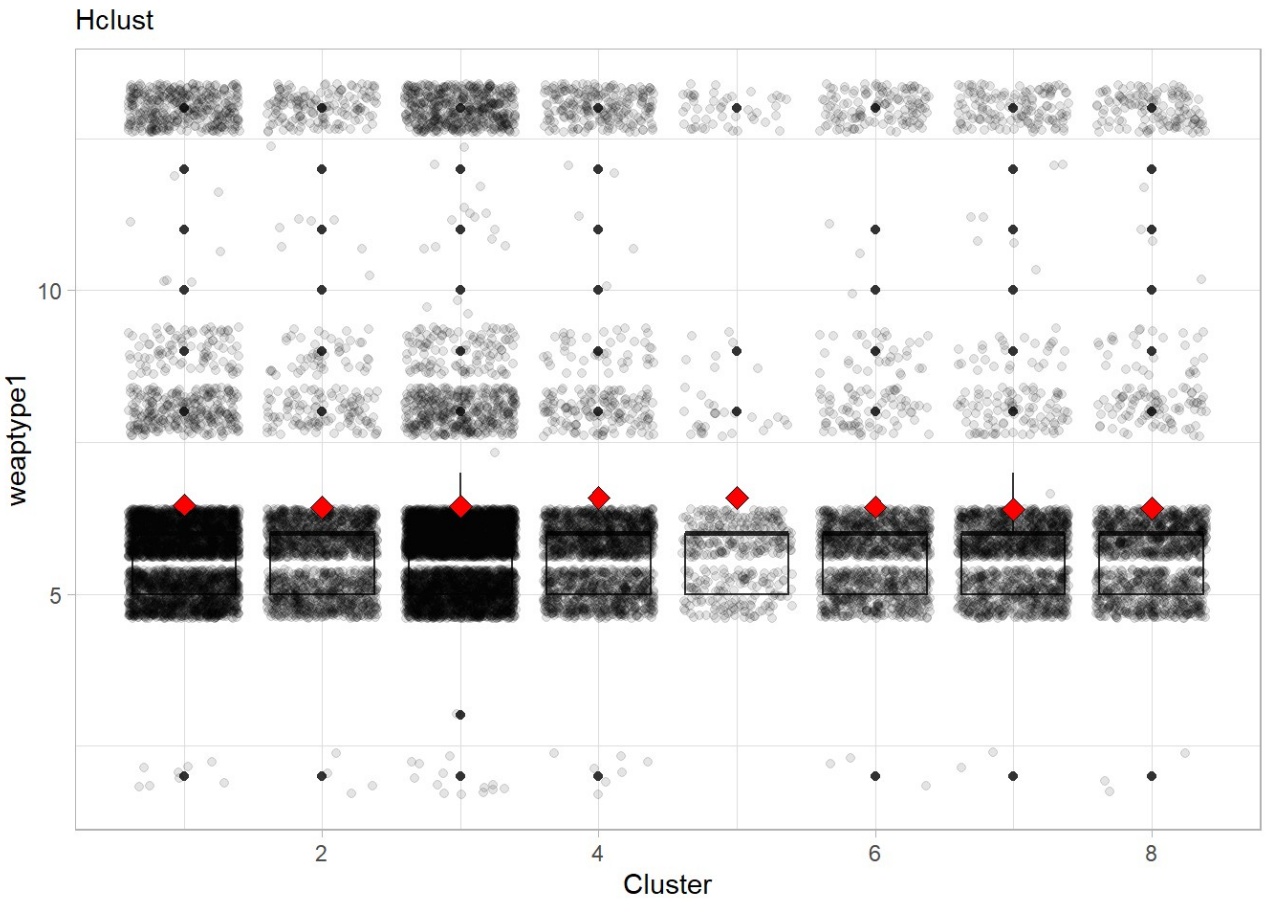
Distribution of Regions in the Clusters



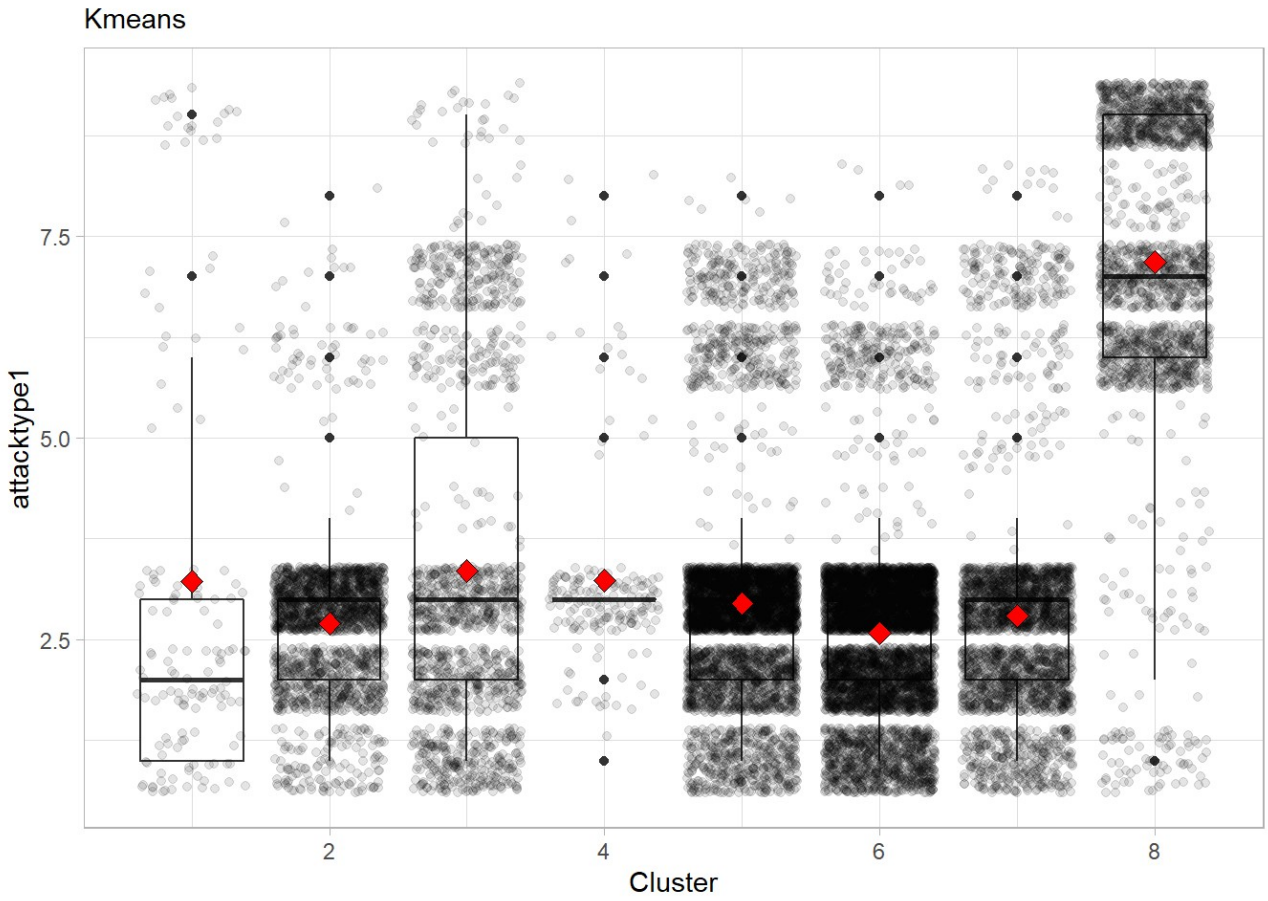

Hclust
12.5
10.0
7.5
5.0
2.5
region
2
4
6
8
Cluster

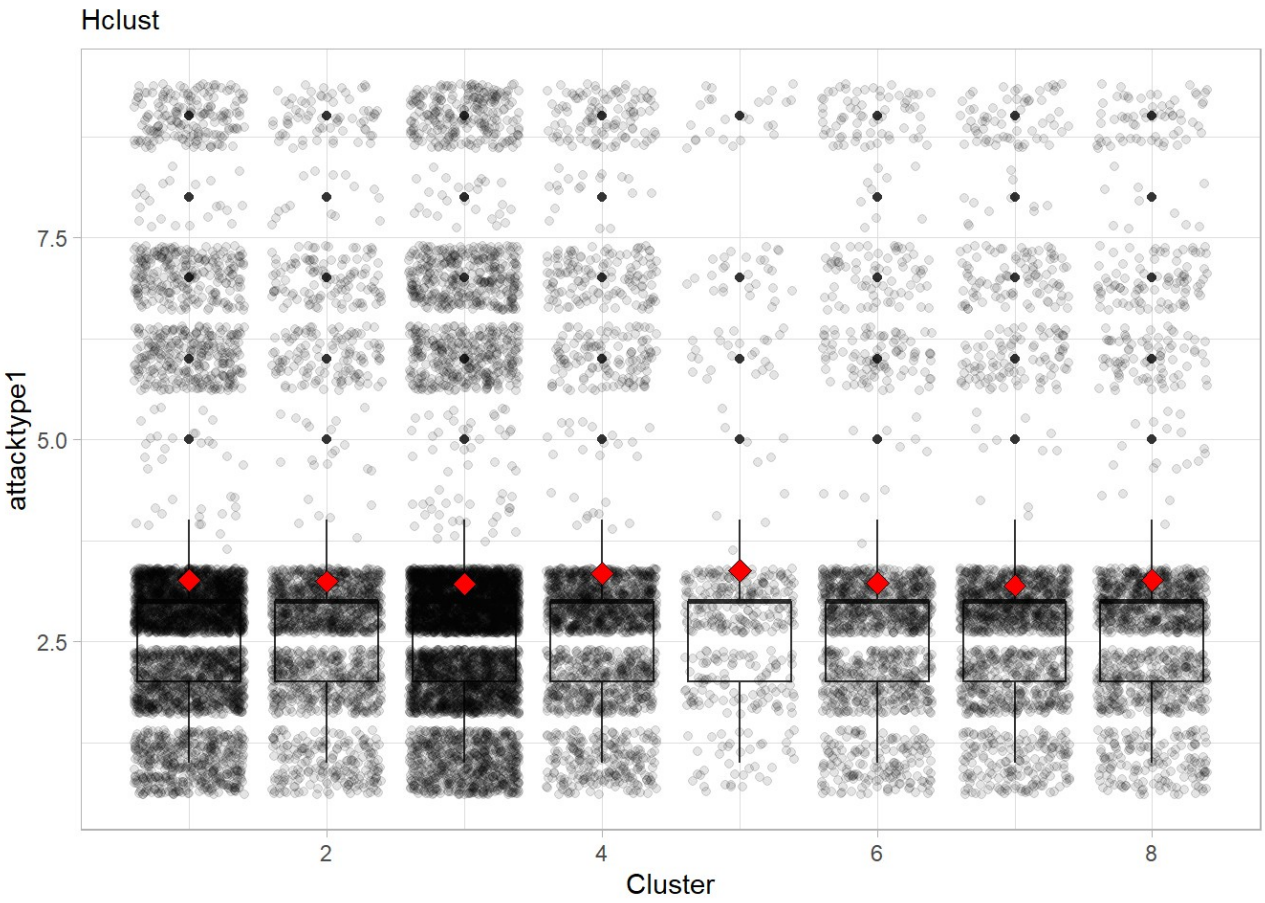
Distribution of Weapon Types in the Clusters



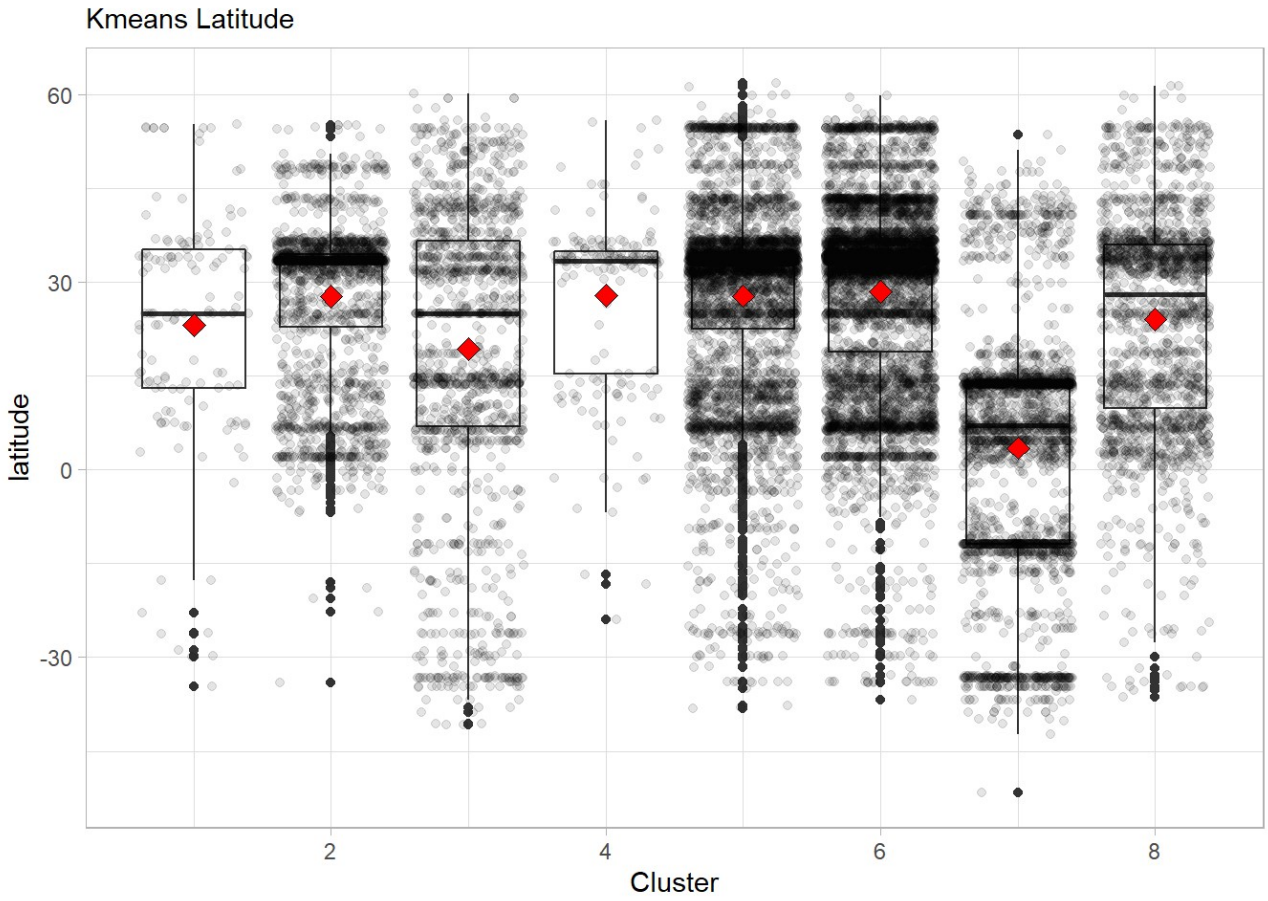

Hclust
weaptype1
10
5
2
4
6
8
Cluster

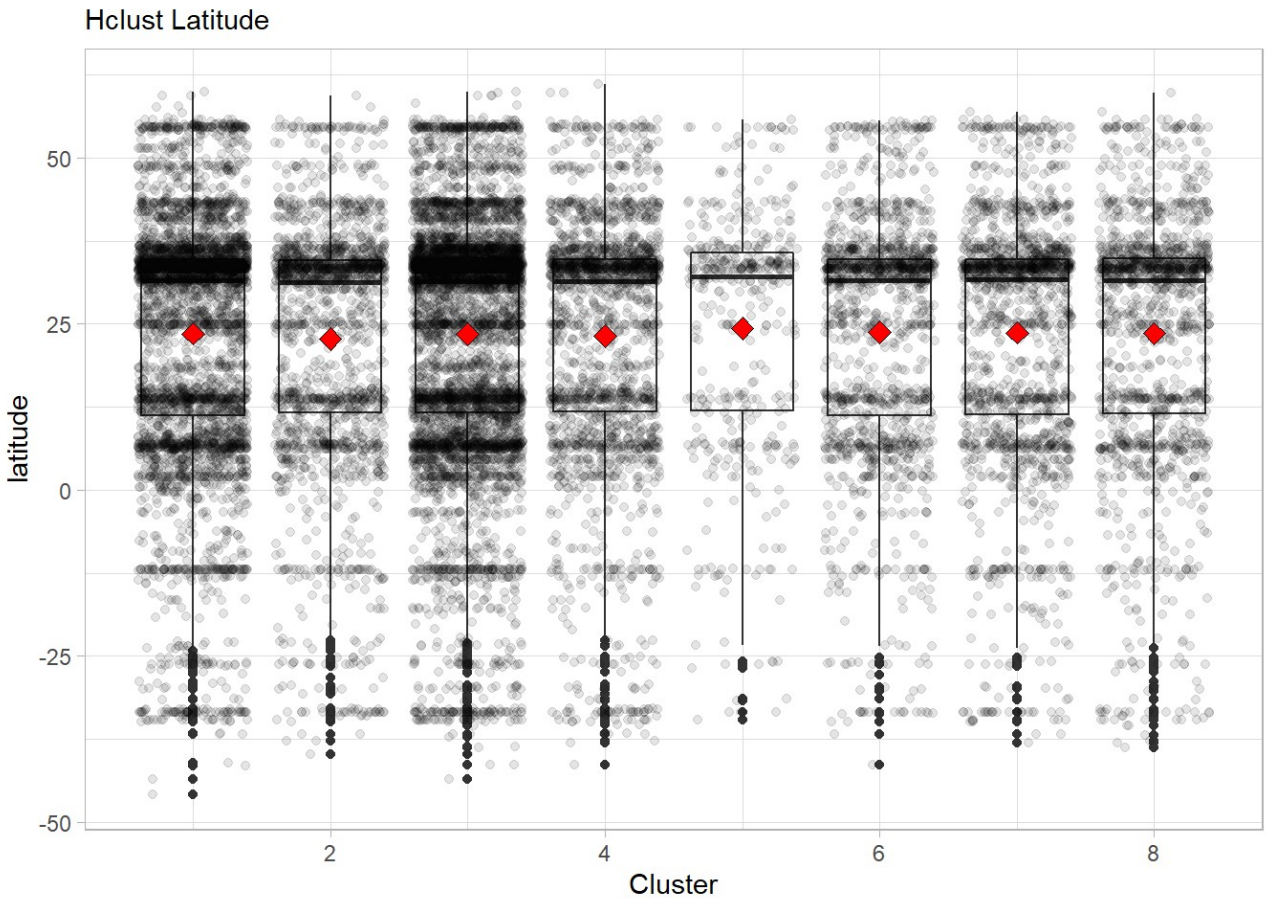
Distribution of Attack Types in the Clusters

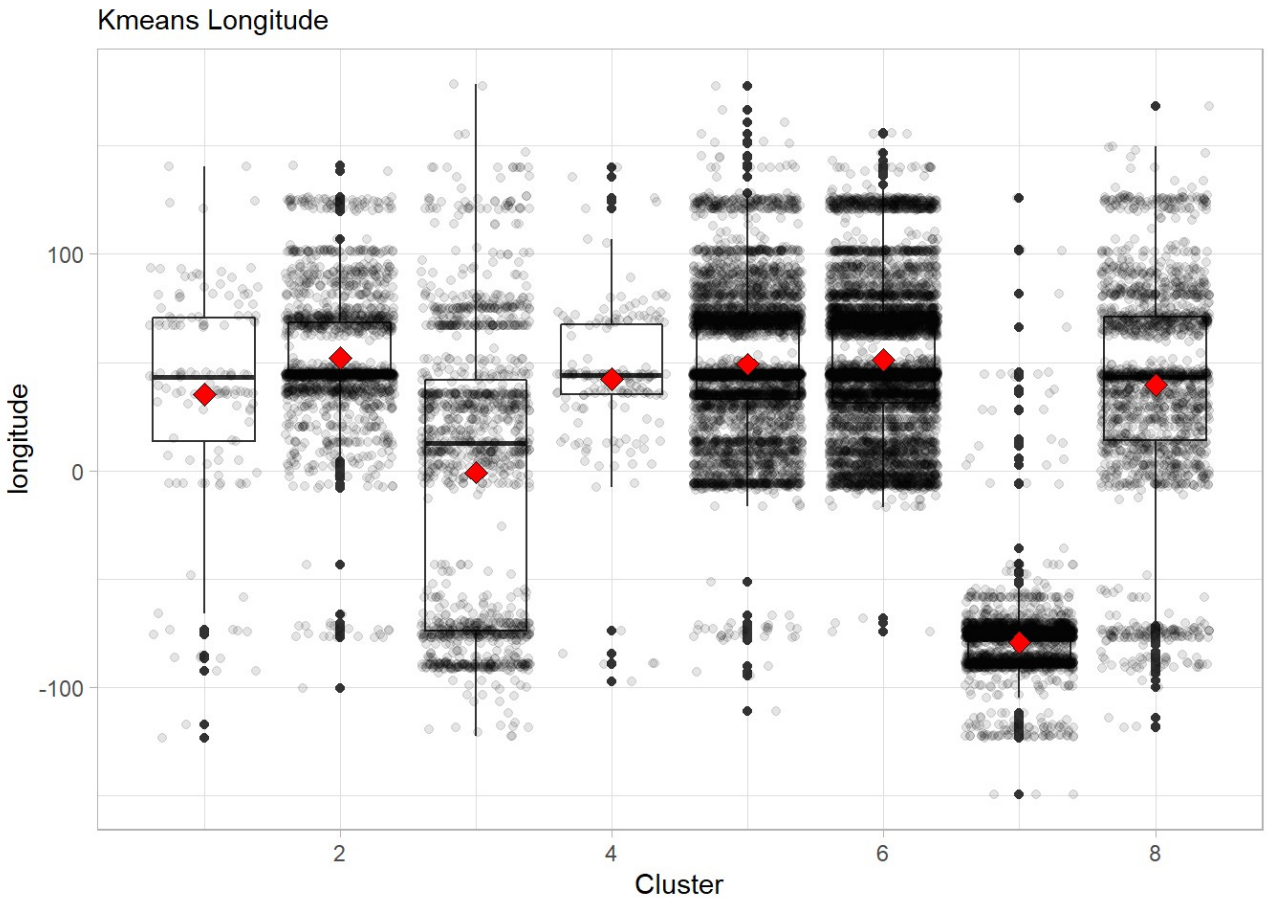


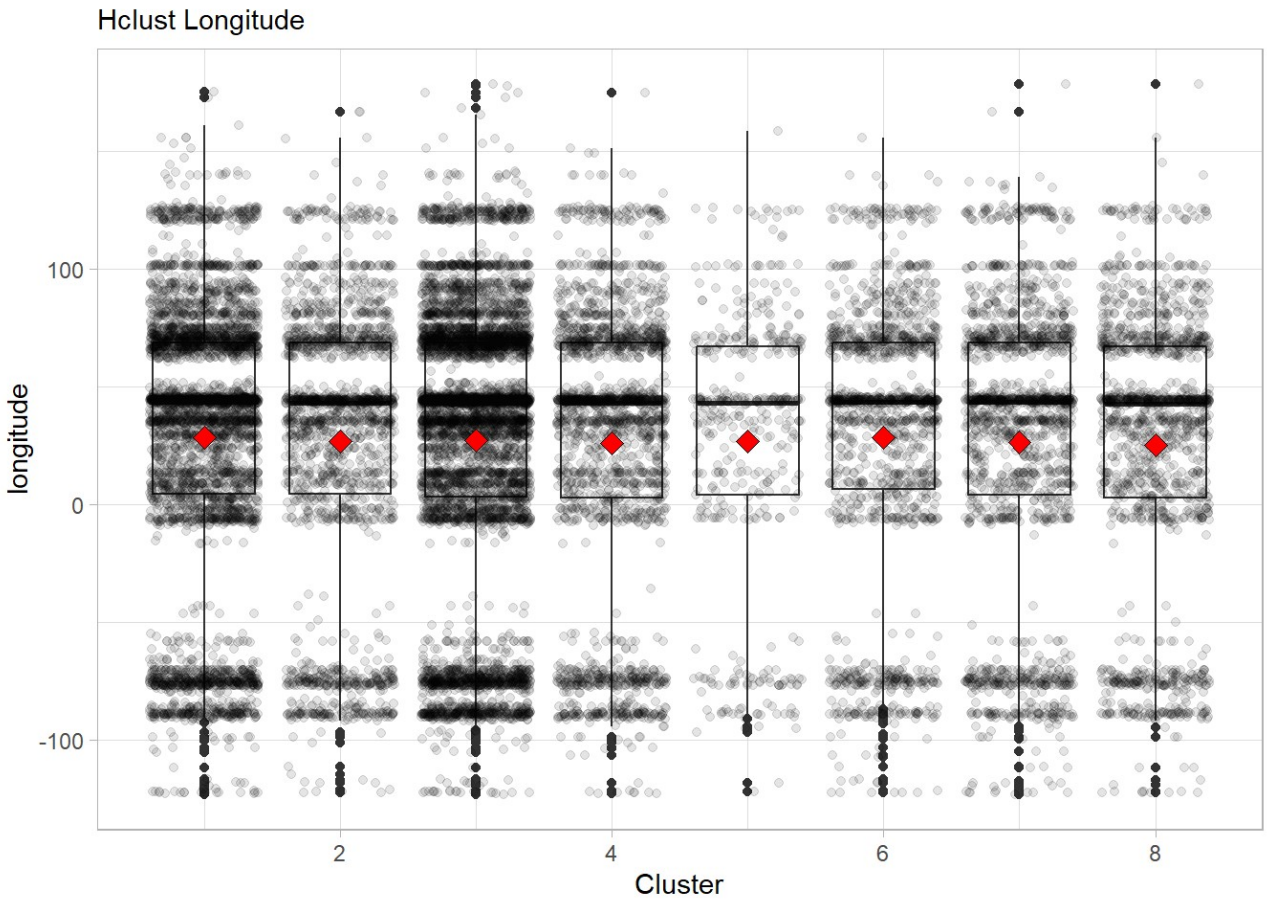

Hclust
attacktype1
7.5
5.0
2.5
2
4
6
8
Cluster

Distribution of Latitude and Longitude by Cluster

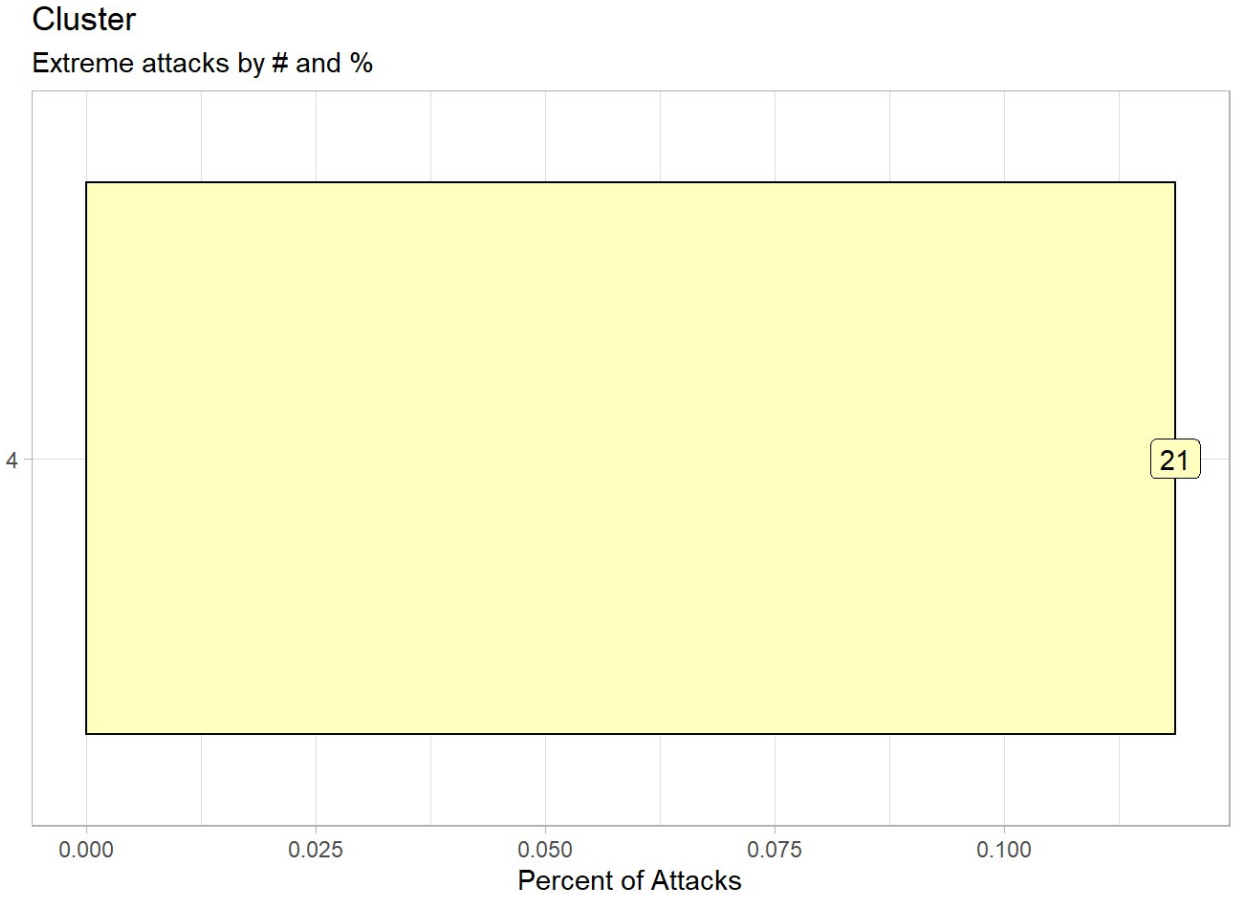



Hclust Latitude
latitude
50
25
0
-25
-50
2
4
6
8
Cluster

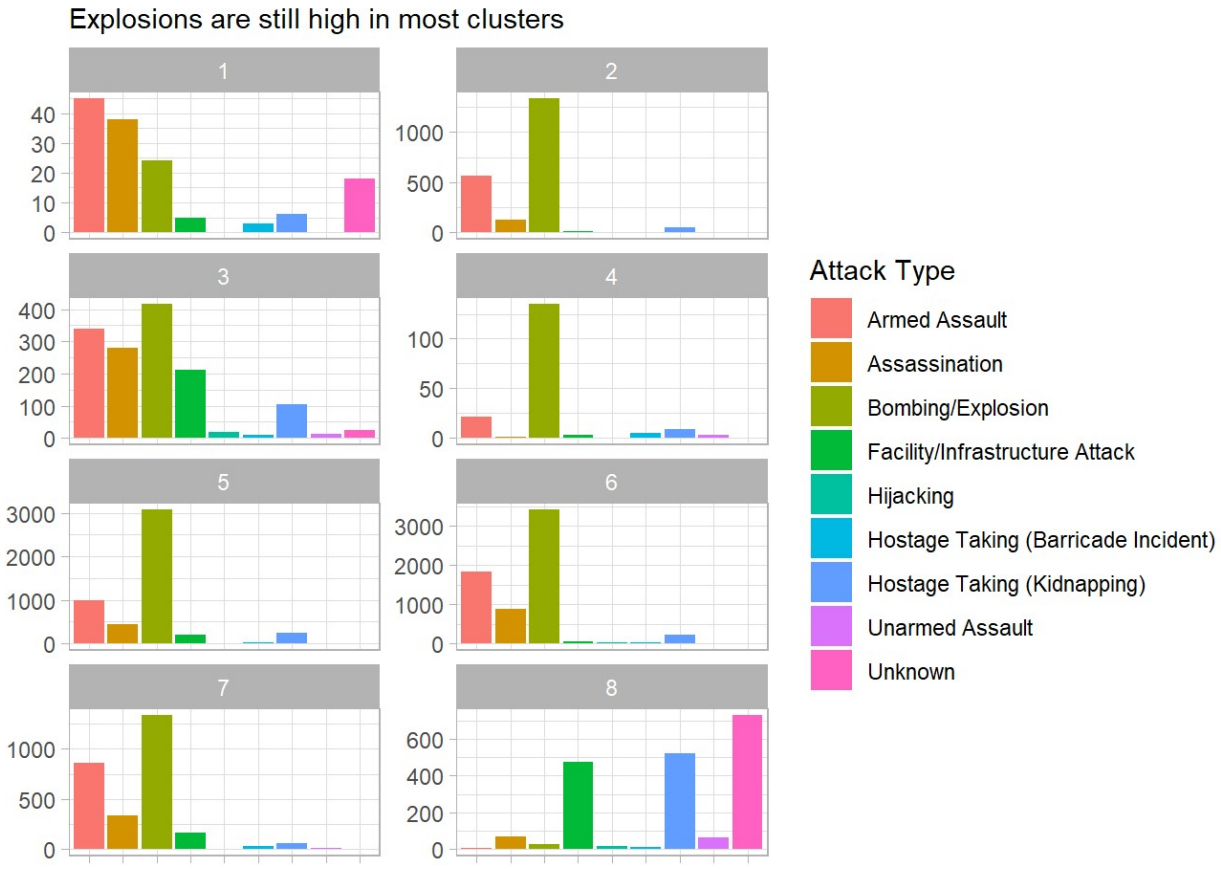

Kmeans Longitude
longitude
100
0
-100
2
4
6
8
Cluster


Hclust Longitude
longitude
100
0
-100
2
4
6
8
Cluster

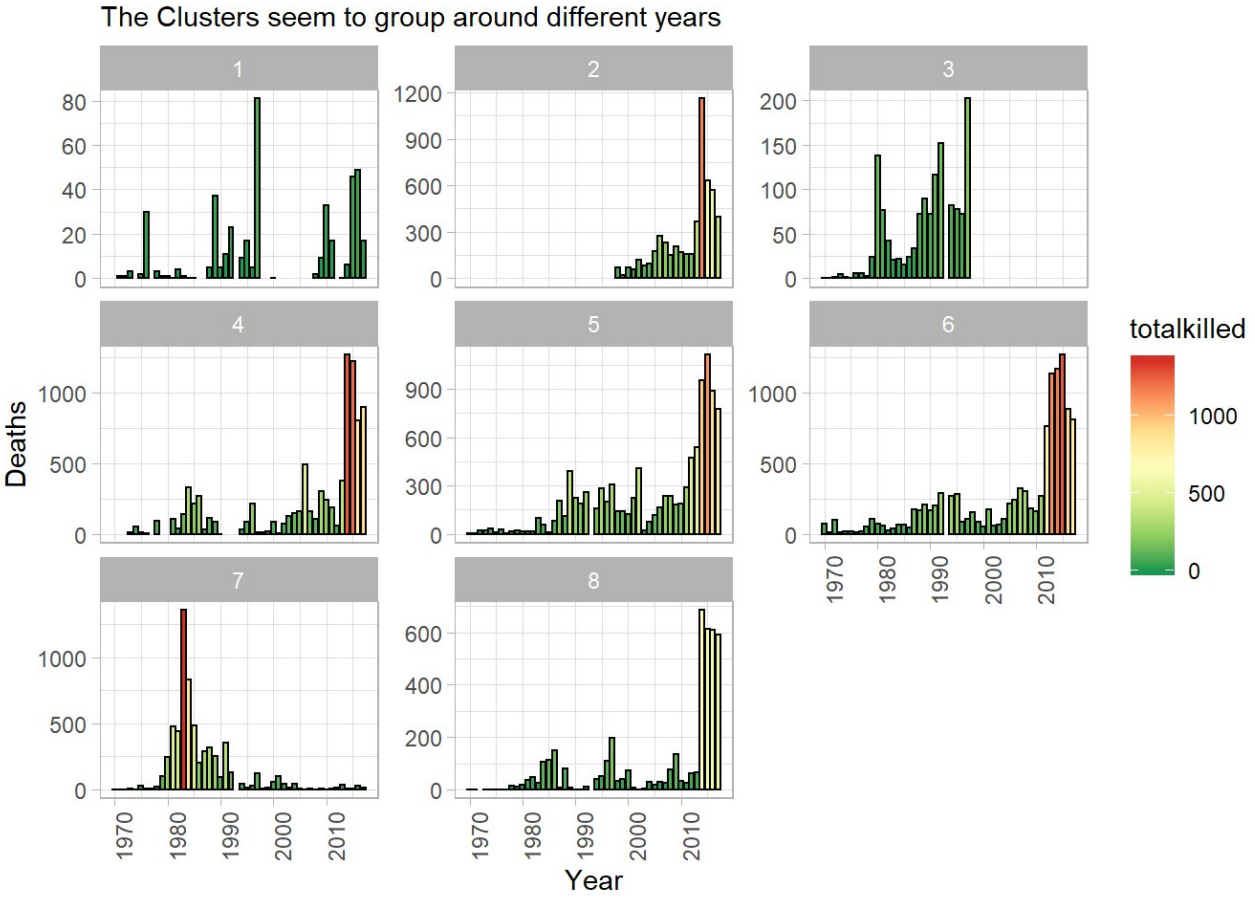
Extreme Attacks by Cluster

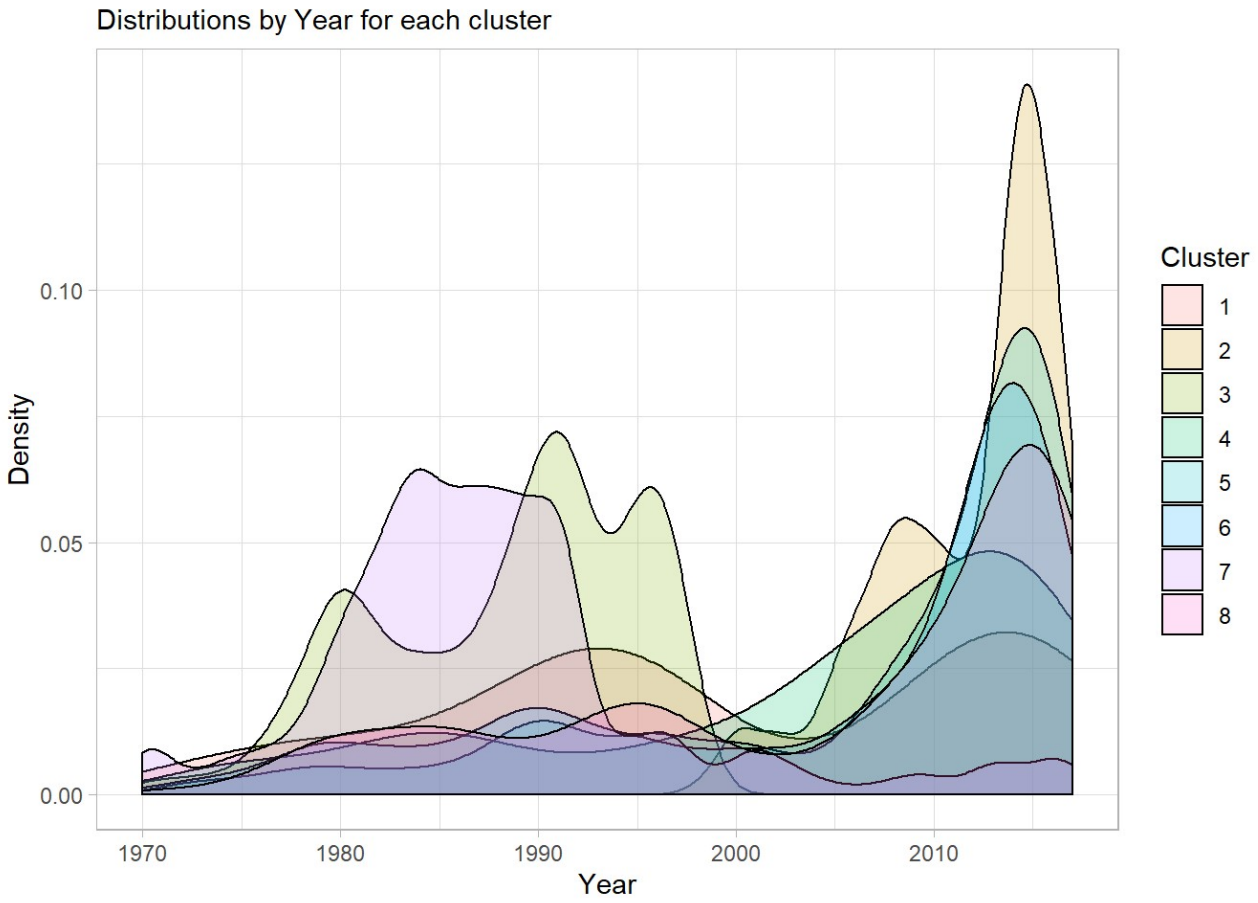


Distribution of Attack type by Cluster

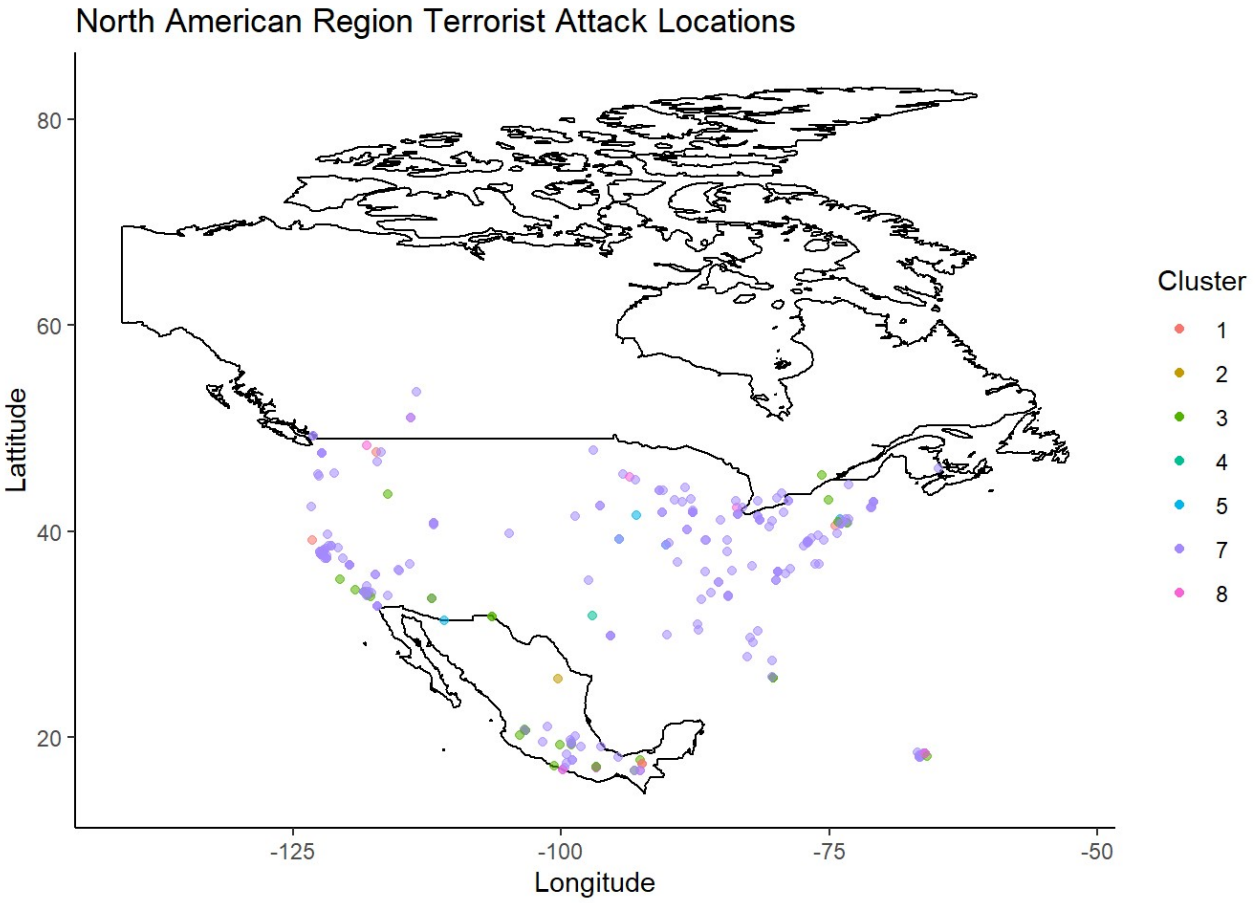


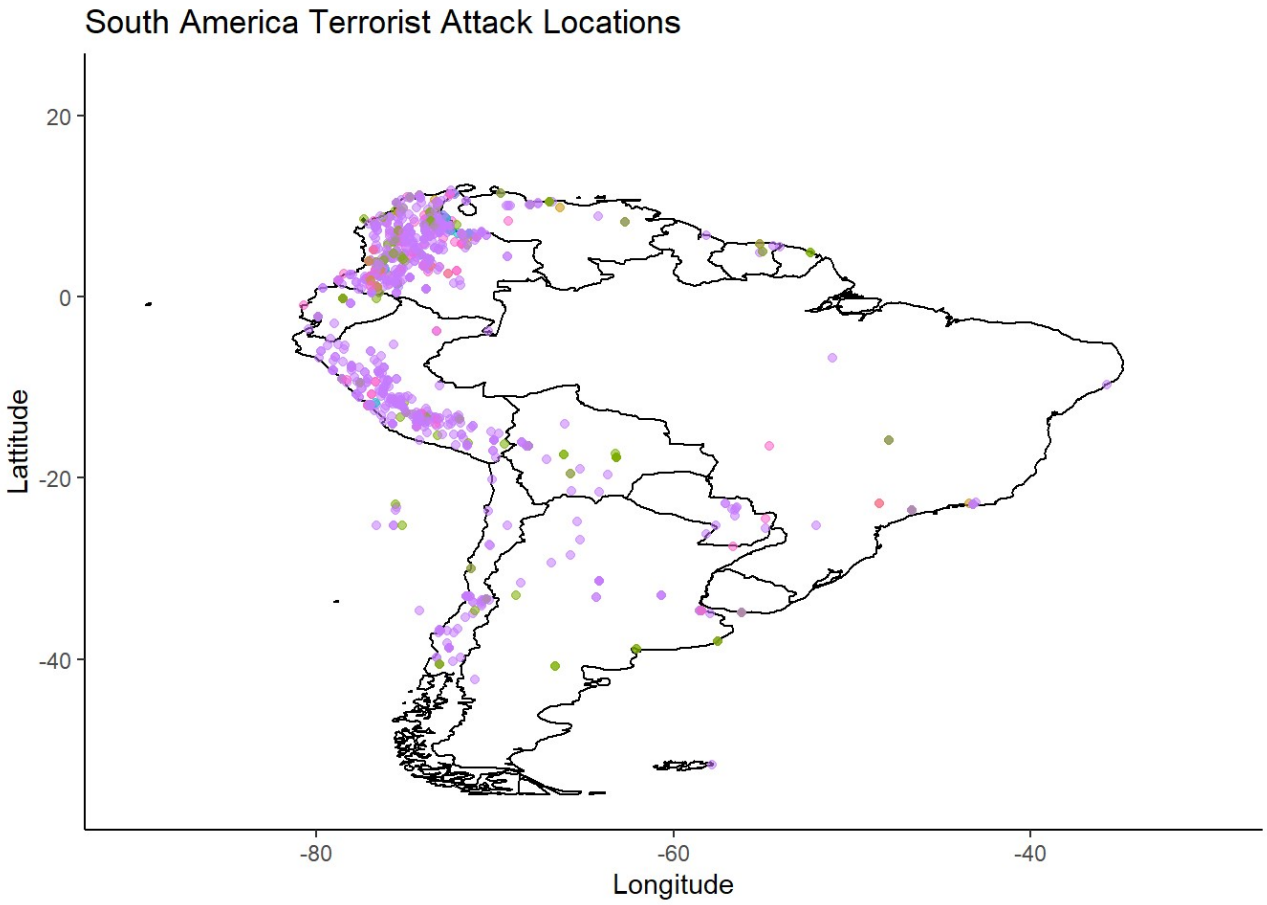
How do the distribution of deaths over time differ by cluster?

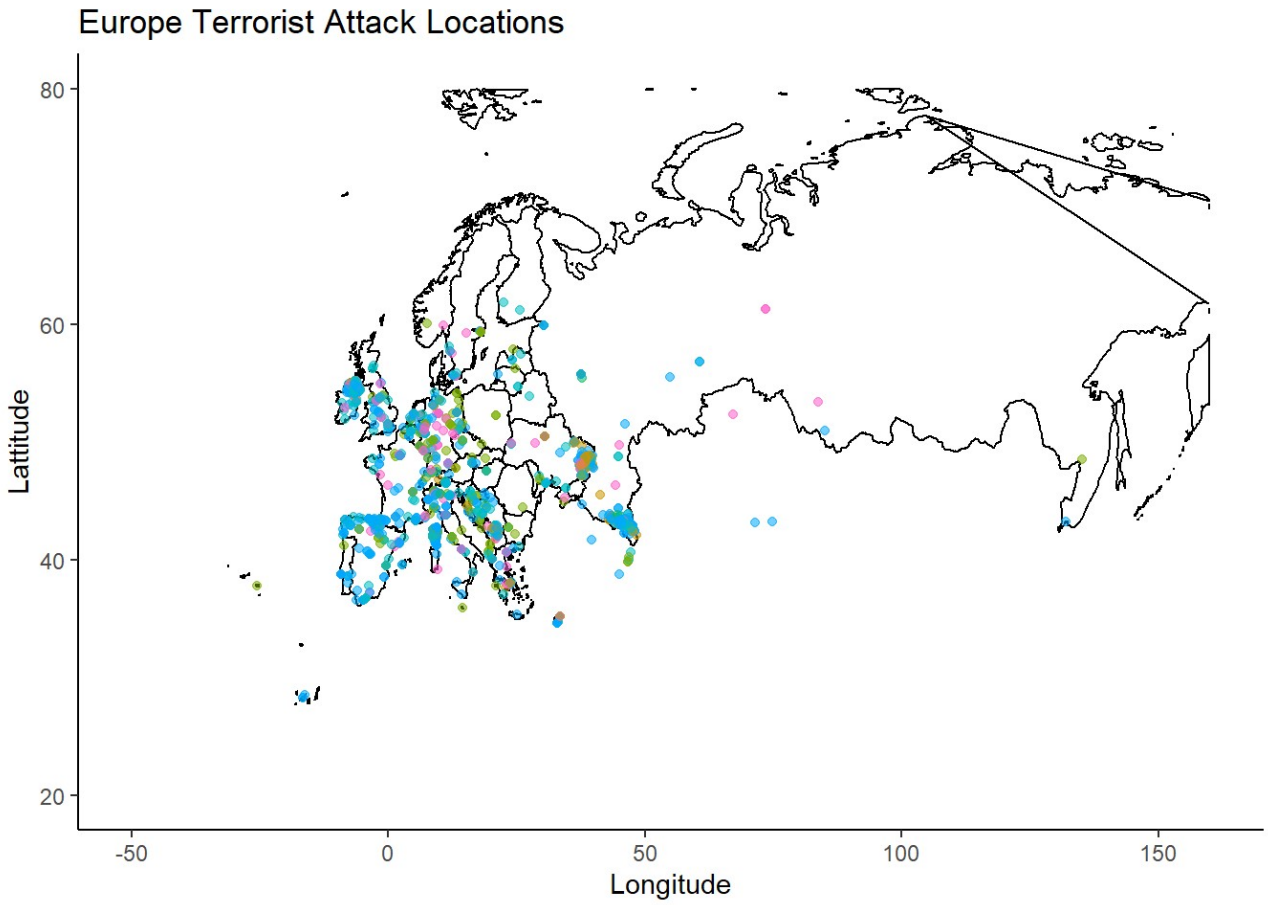


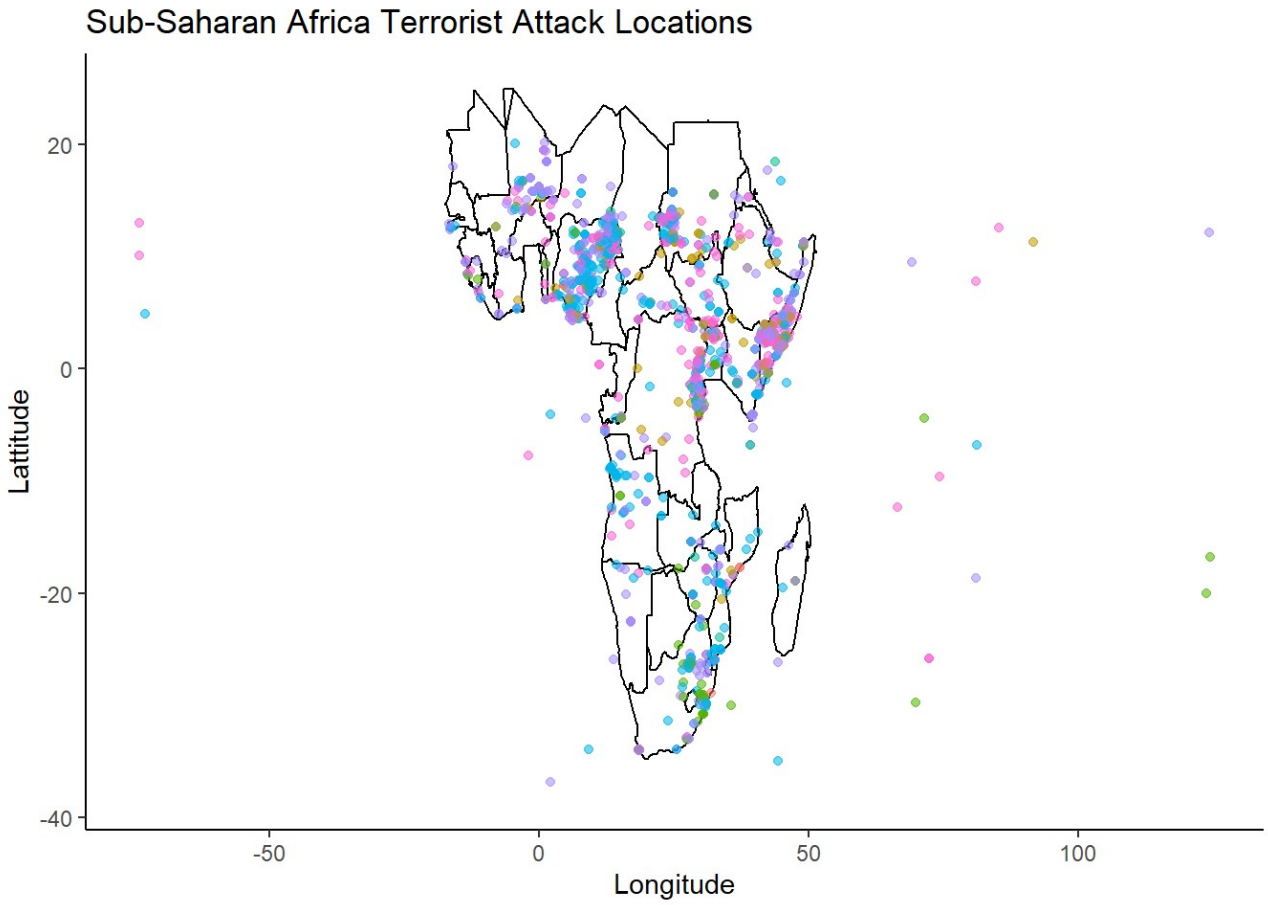

Distributions by Year for each cluster
Density
0.10
0.05
0.00
1970
1980
1990
2000
2010
Year
Cluster
1
2
3
4
5
6
7
8

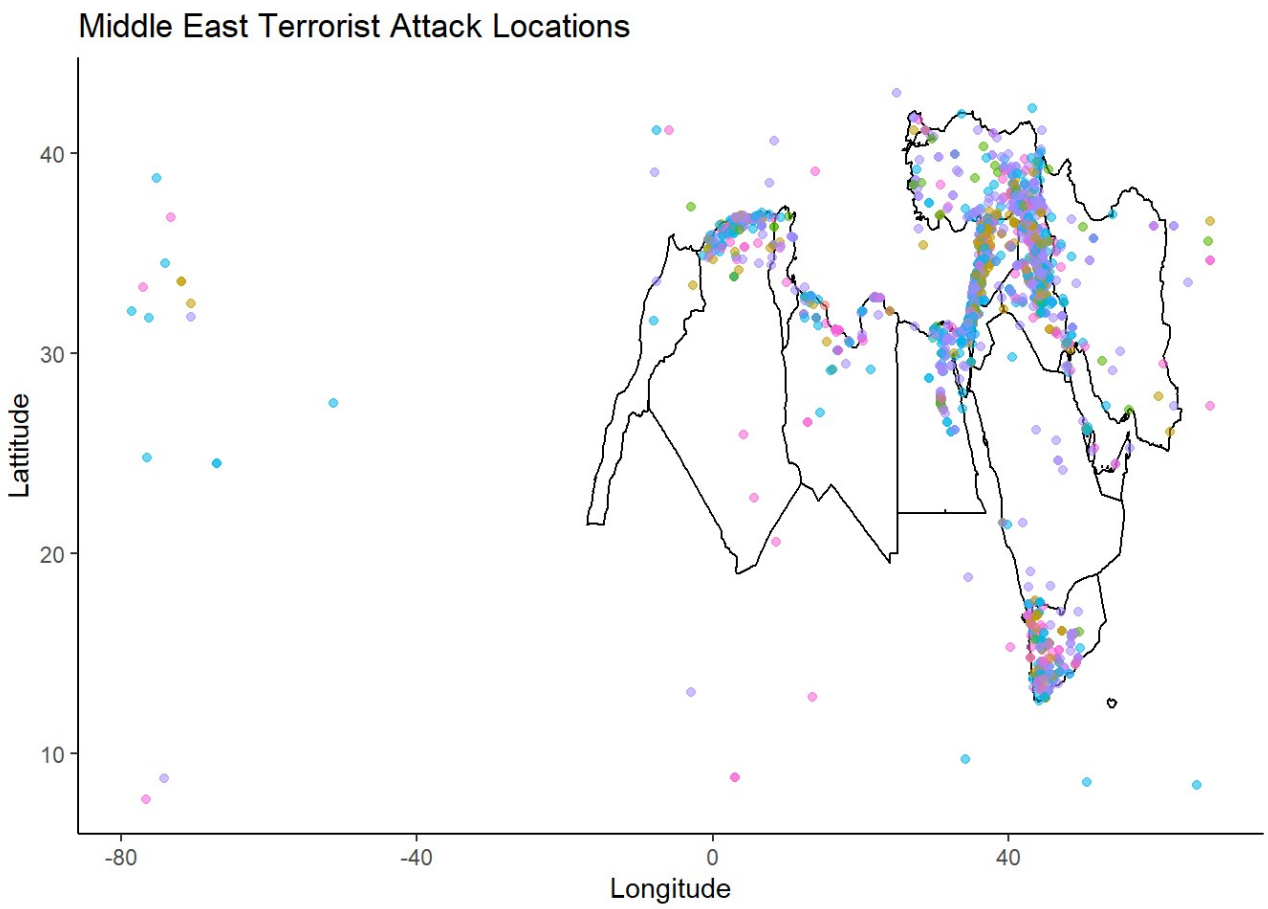
Mapping of Regions and Clusters

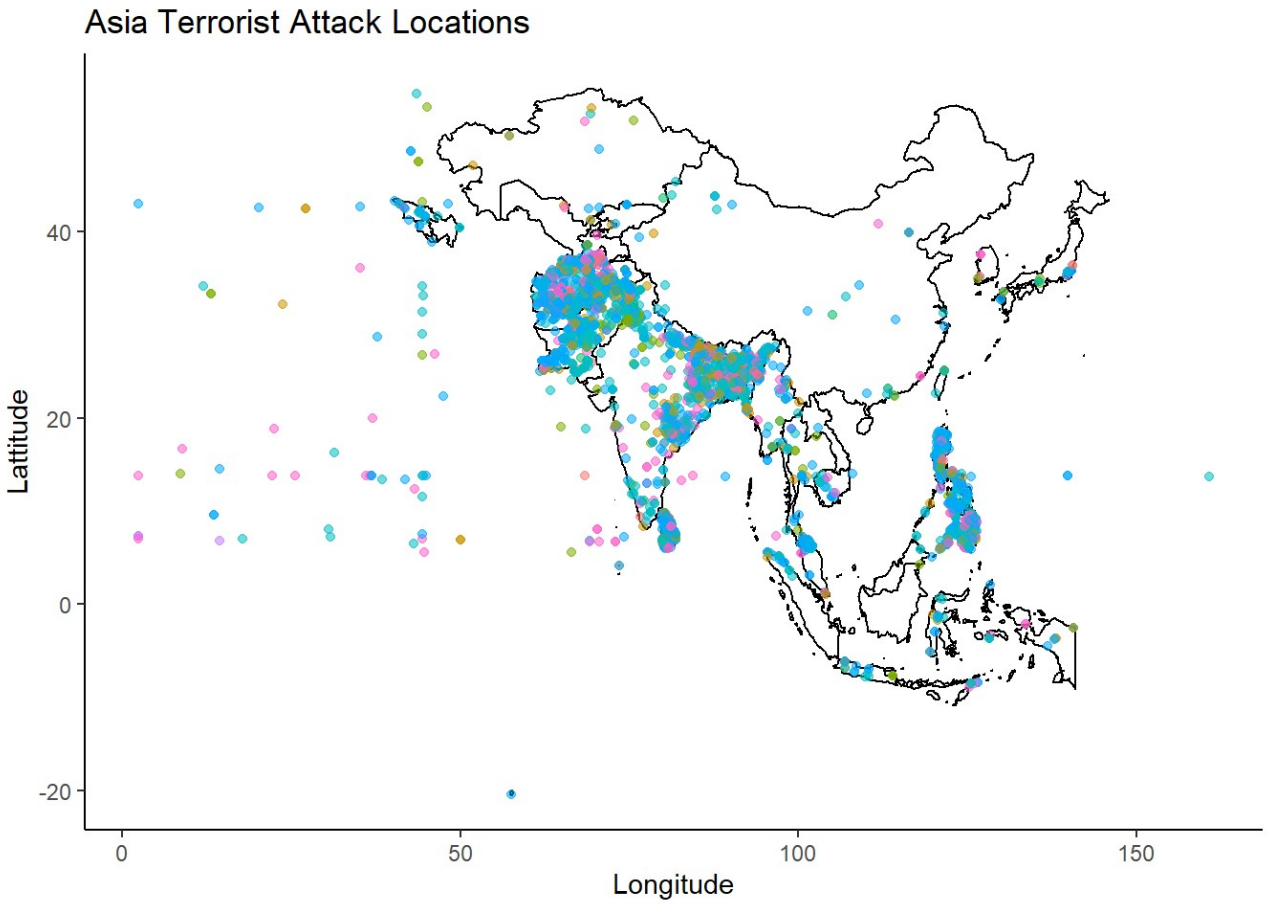


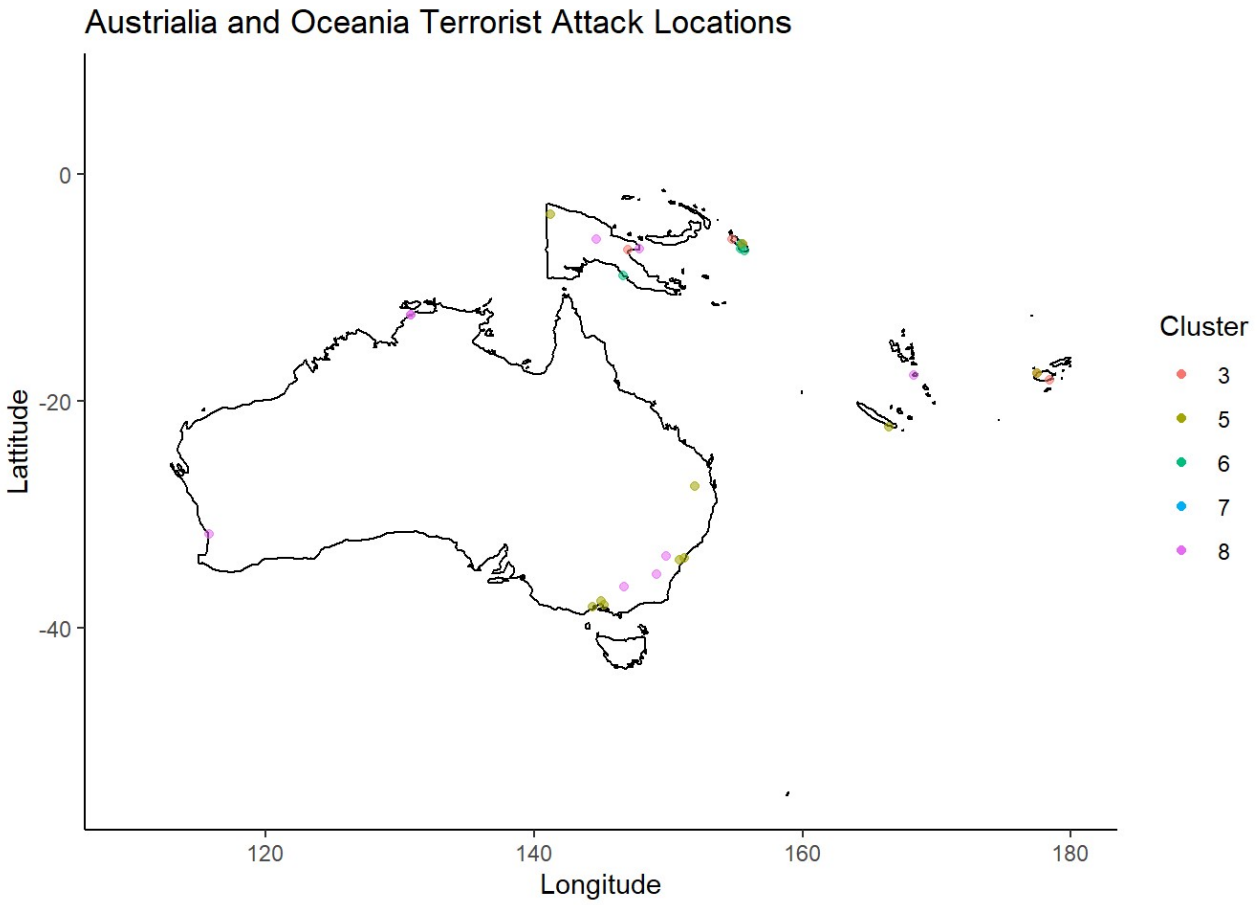

South America Terrorist Attack Locations
20
0
-20
-40
Lattitude
-80
-60
-40
Longitude


Europe Terrorist Attack Locations
80
60
40
20
Lattitude
-50
0
50
100
150
Longitude

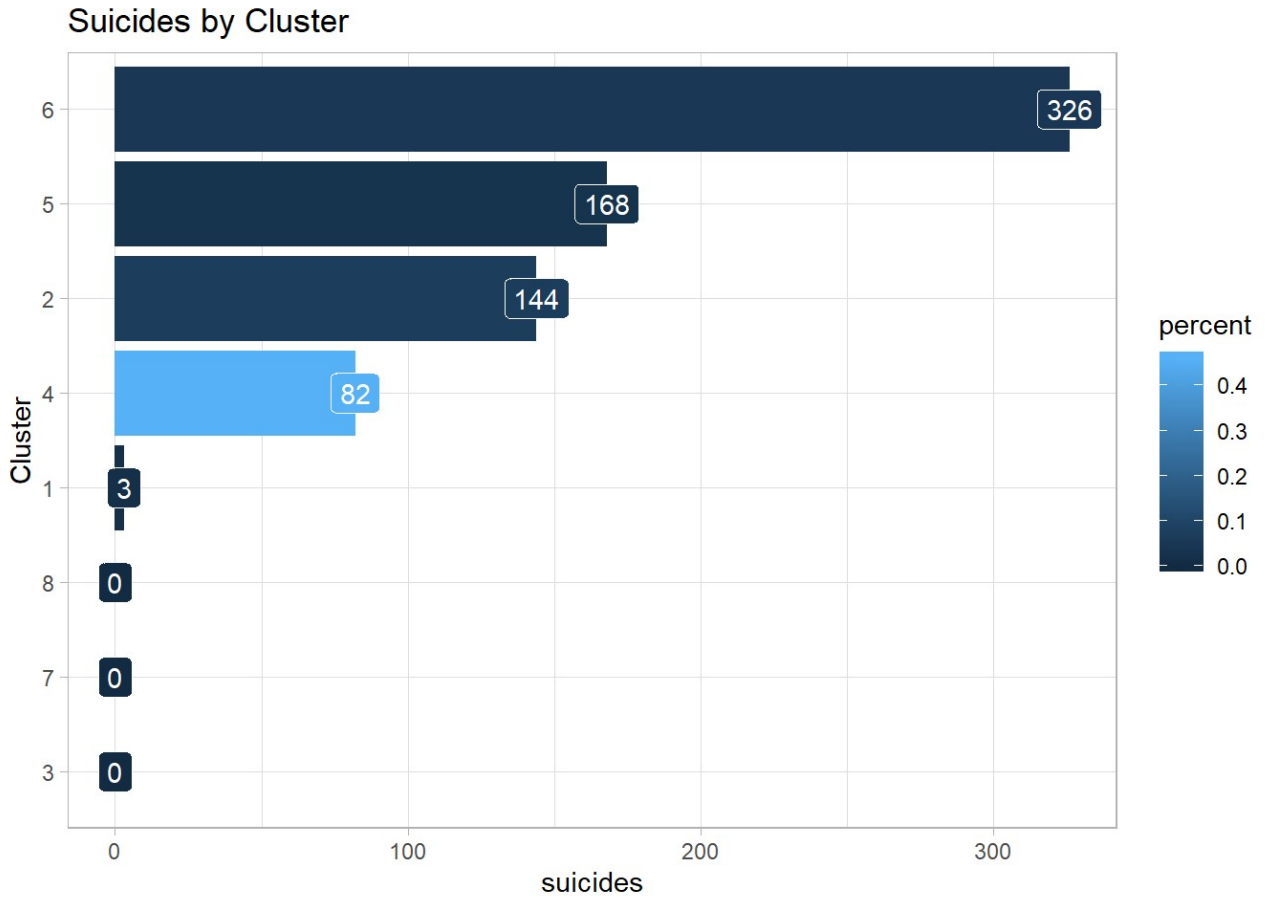

Sub-Saharan Africa Terrorist Attack Locations
Lattitude
20
0
-20
-40
-50
0
50
100
Longitude


Middle East Terrorist Attack Locations
Lattitude
40
30
20
10
-80
-40
0
40
Longitude


Asia Terrorist Attack Locations
Lattitude
40
20
0
-20
0
50
100
150
Longitude


Austrialia and Oceania Terrorist Attack Locations
0
-20
-40
Lattitude
120
140
160
180
Longitude
Cluster
3
5
6
7
8

What are the suicide rates for each cluster?

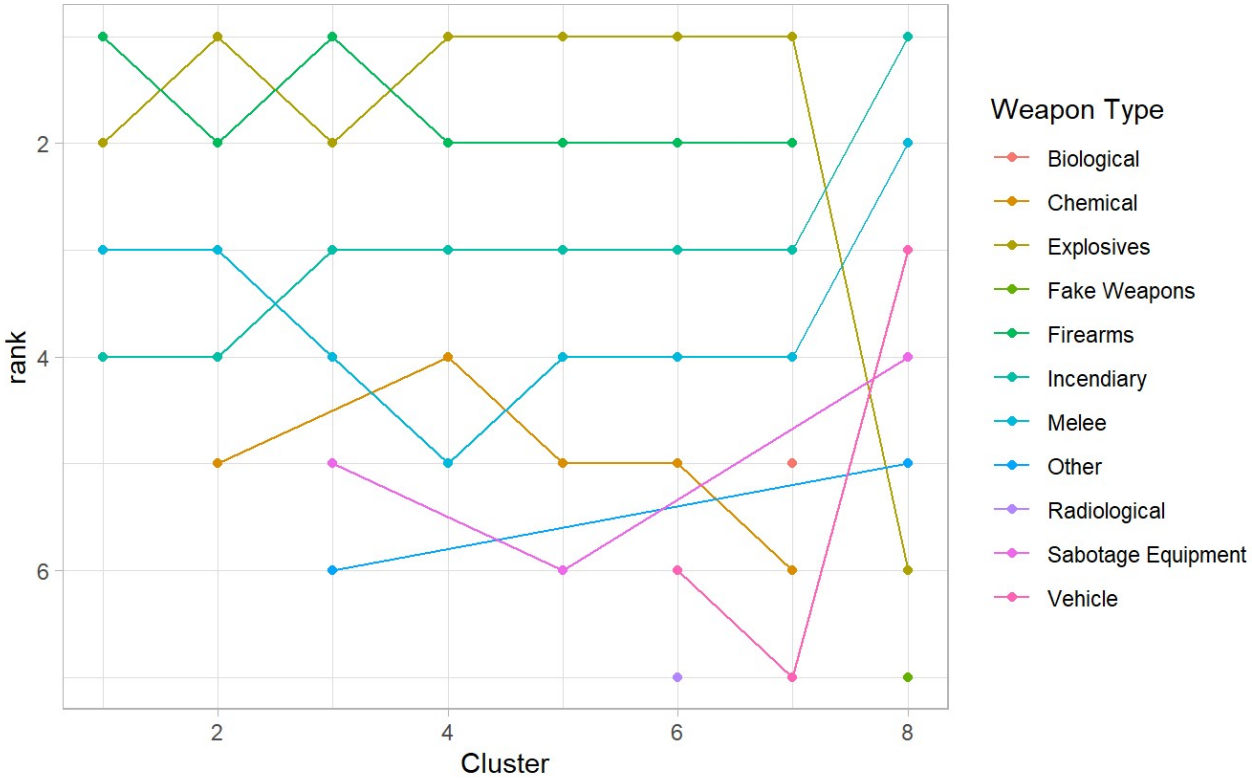


Even though Cluster 4 does not have the most it has a significantly higher percent of attacks than every other cluster.

Most Popular Weapon by Cluster

The Top ranked weapon types for each Cluster.

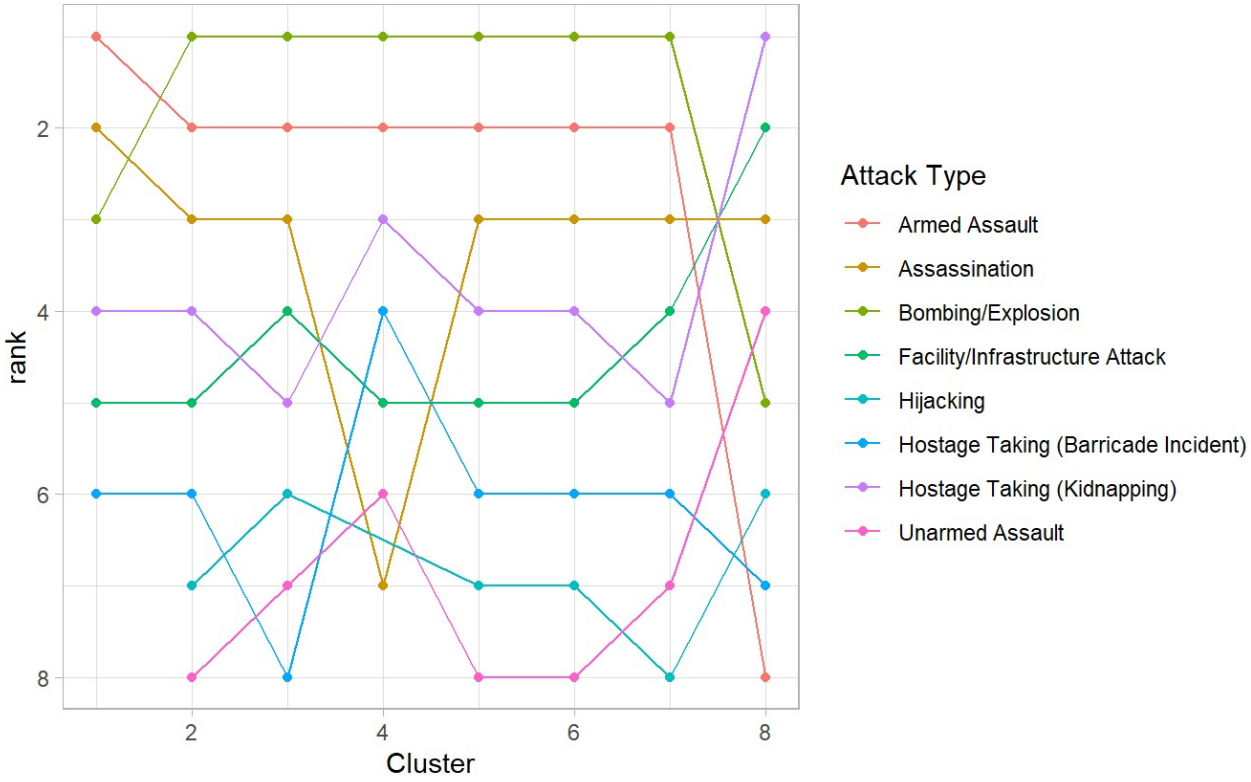
Explosives are the most frequent weapon types across most Clusters.



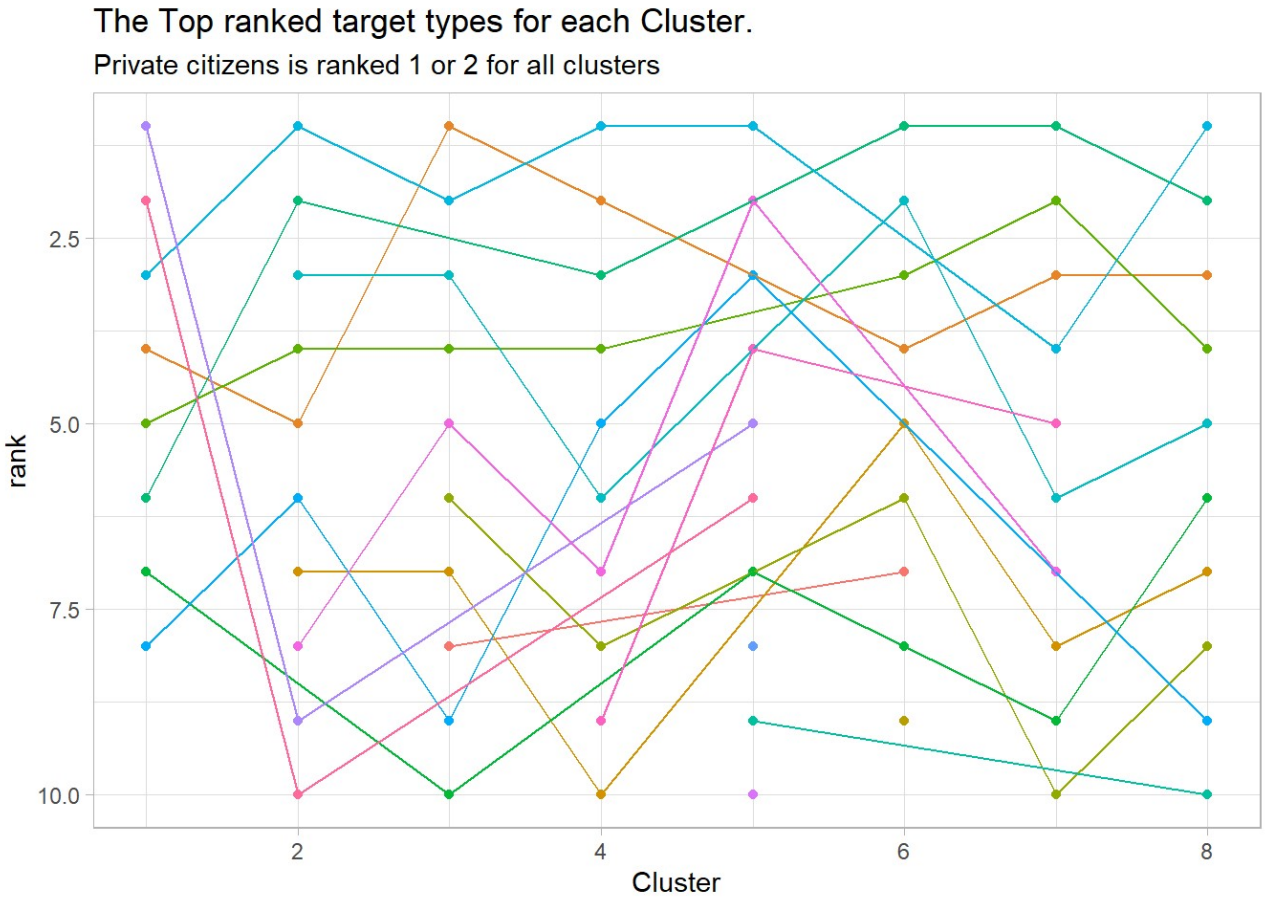
Most Popular Attack type by Cluster.

The Top ranked attack types for each Cluster.

Bombing/explosions are the most frequent attack types across all Clusters.



What target types were most popular in each cluster?




```
## # A tibble: 77 x 4
##    Cluster targtype1_txt                        n  rank
##      <int> <fct>                            <int> <int>
##  1       1 Terrorists/Non-State Militia        67     1
##  2       1 Violent Political Party             35     2
##  3       1 Private Citizens & Property         25     3
##  4       1 Business                             5     4
##  5       1 Government (General)                 3     5
##  6       1 Military                             1     6
##  7       1 Journalists & Media                  1     7
##  8       1 Religious Figures/Institutions       1     8
##  9       2 Private Citizens & Property        705     1
## 10       2 Military                           383     2
## # ... with 67 more rows
```

Conclusions and Future Work

As I have explored the relationships between the clusters I continued to look for ways to quanitfy the strength of my clustering I found a handful of statistics to look for. One was the gap statistic (this is determined to find the optimum number of clusters), NbClust(also used to find the optimum number of clusters), and silhouette width of each cluster.

When I did the gap statistic and the NbClust I actually got two different results. NbClust said that the best (majority vote) said that three clusters was the best and eight clusters was second. But the gap statistic suggested eight was best and at one point when I had run the NbClust, it suggested eight as well, but I had not set my seed, so I was not able to reproduce this. Because of all this, I chose to go with 8 clusters.

My goal was to cluster the data and look for trends or defining charachteristics within the clusters to see what set each cluster apart.

I believe that I have successful completed my goal of clustering the data into meaningful and distinct clusters. The rand index pointed me to the fact that my clusters did not just fall along the lines of my regions, weapon types, or attack types. My silhouette coefficients pointed to the fact that my clusters and assigned points were in good agreement with each other. And I was able to visually find defining characteristics of each cluster. Below are some of the mentionable and noticeable cluster findings.

Findings

- My rand index did not suggest that there was much of a similarity between my clusters and the regions, weapon type or attack type, which I was worried about. All of these values centered around 0
- Cluster 7 has a distinct distribution right around regions 2 and 3, that is the Central and South American region.
- Clusters 5 and 7 are heavy in the regions of 5, 6, 8, 10, and 11 (Southeast Asia, South Asia, Western Europe, Middle East and North Africa and Sub-Saharan Africa respectively)
- Cluster three has a mean of region 6, South Asia.
- Cluster 8 has a distinct difference in weapon type, centered on weapon type of Unknown and a grouping around Melee.
- Cluster 8 also has a distinct difference in attack type, more Unknown attacks, Hostages, and Infrastructure attacks.
- Cluster 4 had ALL of the extreme attacks. It also represented more than 10% of its attacks. More than any regions percentage.
- Cluster 7 has a mean around the equator, the rest of the clusters have northern hemisphere centers.
- Cluster 7 also has a significant difference in longitude, which makes sense if it has mostly Central and South America (different longitudes than most other regions).
- Cluster 8 has a different distribution than all the others for Attack Type
- Clusters 1 and 3 are alike, whereas 2, 4, 5, 6, and 7 all have higher armed assault numbers and bombing/explosions, but skip over the higher assassinations (comparatively).
- Cluster 1 spikes on Unknown attack types and Cluster 3 spikes Hostage (kidnapping).
- Cluster 3 centers around attacks in 1990.

- Cluster 7 clusters around attacks in 1980.
- Clusters 2, 4, 5, 6 and 8 all have the normal attack pattern (normal compared to the overall data's distribution).
- Cluster 1 has had very few people killed compared to all the other clusters.
- Clusters 1 and 4 have similar amount of attacks (139 and 177 respectively) but cluster 4 was way deadlier.
- Even though Cluster 4 does not have the most it has a significantly higher percent of attacks than every other cluster.
- The clusters had variability when it came to the rank of weapons, some of the clusters shared the top few, but after that there were significant mixes.
- Attack ranks experience more stability, especially amongst the top two attack types. I know these plots are messy, but they highlight the variability of the clusters. We want the clusters to be different, that extends to the attack and weapon types and how they ranked in each cluster.

Future work and considerations

One area of consideration for future work of this would be to investigate more on how many clusters are best. I had non-agreeing results with my NbClust and gap statistic. There could be further analysis in of the internal measures of the clusters to see how many clusters between 2-10 clusters would be best. Due to limited time and computing capacity I was not able to dedicate as much time to this section. Another area would be to investigate other ways of handling such large datasets. I sampled the data and only took 20,000 rows compared to the over 180,000 original. This again was due to the computing constraints of my machine. This leaves plenty of opportunity to research other ways of sampling the data, or methods for clustering with large datasets.

Future work from here:

- Assign data points to their clusters and use this as the prediction variable for future attacks. This could be used to help identify characteristics of future attacks and grouping them with known attacks.
- There could be other analyses to look for defining features and characteristics of the clusters. I broke them down into years, extreme attacks, weapon type, attack type, regions, countries, etc.. But with so many features, the are many other analyses that could be performed. Creativity in this part would be helpful. Thinking outside the box is what lead to the discovery of extreme attacks, which ended up being a defining feature of one of the clusters.
- Combining the data with other time-based data for insights. I think it would be interesting to join terrorism data with stock market performance to look for similarities or patterns. This could be used to see if one affects the other or predicts the other. Do terrorist attacks predate stock markets swings or vice versa?
- There was a description field that text analysis or natural language processing could have been used on.
- Combining text analysis with a deadly year by finding news articles and headlines either the year or month before and then after to see if there were trends in the sentiments of headlines.
- Web scraping Twitter to see if sentiments of tweets reflected the attacks or perhaps foretold of pending attacks. Since the data and tweet data both have geographic information you could use regional analysis to predict attacks based on text analysis insights.

- One area could be to gather more information about the perpetrating groups and seeing if there are insights that would lead to being able to predict which group was responsible for an attack before credit has been taken. This could lead to faster responses and justice.

The following helpful resources and what topics they contributed to in the my project:

Principal component analysis: (Kassambara, 2017) Mapping in R: (Godlee, n.d.) Imputation strategies: (Mekala, 2018) Optimal Clustering: (Oldach, 2019) Clustering using gap statistic: (Boehmke, 2017) Ggplot and visualizations (R Graphics Cookbook): (Chang, 2013) Practical Statistics for Data Scientists: (Bruce & Bruce, 2017)

References

Boehmke, B. (2017, March 03). K-Means Cluster Analysis. Retrieved from UC Business Analytics R Programming Guide: https://uc-r.github.io/kmeans_clustering (https://uc-r.github.io/kmeans_clustering) Bruce, P., & Bruce, A. (2017). Practical Statistics for Data Scientists. Sebastopol, CA: O'Reilly Media, Inc. Chang, W. (2013). R Graphics Cookbook. Sebastopol, CA: O'Reilly Media, Inc. Godlee, J. (n.d.). Spatial Data and Maps. Retrieved from Coding Club: https://ourcodingclub.github.io/tutorials/maps/ (https://ourcodingclub.github.io/tutorials/maps/) Kassambara, A. (2017, August 10). Principal Component Analysis in R: prcomp vs princomp. Retrieved from Statistcal tools for high-throughput data analysis: http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/118-principal-component-analysis-in-r-prcomp-vs-princomp/ (http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/118-principal-component-analysis-in-r-prcomp-vs-princomp/) Mekala, H. (2018, June 29). Dealing with Missing Data using R. Retrieved from Medium: https://medium.com/coinmonks/dealing-with-missing-data-using-r-3ae428da2d17 (https://medium.com/coinmonks/dealing-with-missing-data-using-r-3ae428da2d17) Oldach, M. (2019, Jan 27). 10 Tips for Choosing the Optimal Number of Clusters. Retrieved from towards data science: https://towardsdatascience.com/10-tips-for-choosing-the-optimal-number-of-clusters-277e93d72d92 (https://towardsdatascience.com/10-tips-for-choosing-the-optimal-number-of-clusters-277e93d72d92)