Global Terrorism DB Project

Evan Harley

April 13, 2015

First Steps

The very first things that I did were load the appropriate libraries and the database that I would be working with.

Data Wrangling

So the first thing that I noticed about this data set is the fact that there are rather a lot of variables in the data set that repeat the same information in. So, I decided that I would work with a subset of the variables. I chose the variables that made the most sense to me.

I chose to only keep the first attack type, target type, and weapon type, because the majority of the entries that I looked at did not have values. I chose to drop many of the kidnapping/hostage/hijacking specific values because they applied to a subset of the values that I wasn't particularly interested in.

After taking a look at the values in the Weapon Type 1 text variable I noticed that the vehicle variable value was long enough to obscure the value of a count, so I subset the data and changed the value to just Vehicle

Just to ensure that there aren't any date values that don't make sense I ran a tally of all of the day and month variables

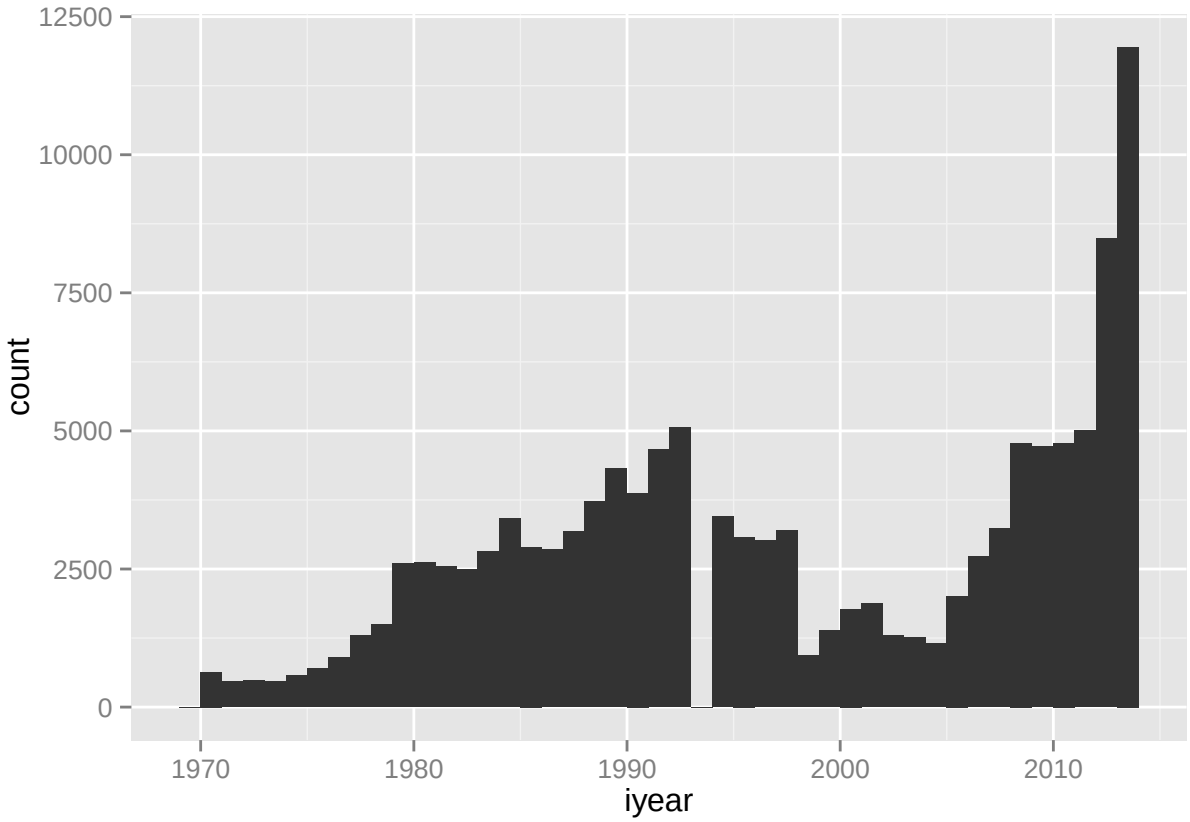
Finding some wrangling need, I continued to tally variables looking for values not accounted for in the codebook document.

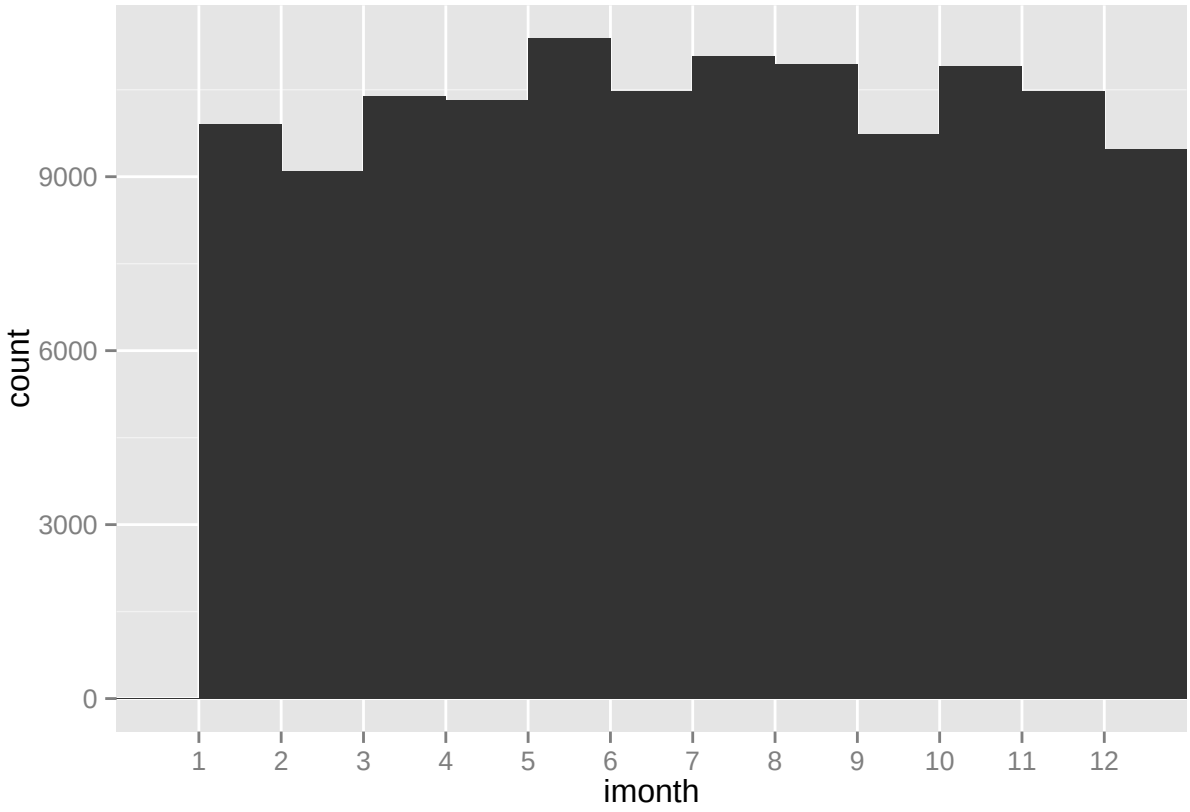
The only values that took me aback are the values between whole numbers in the nkill, nkillter, nwound, and nwoundte variables. These are explained in the literature as averages.

I also noticed that there was no consistant date variable. So, I added one

---

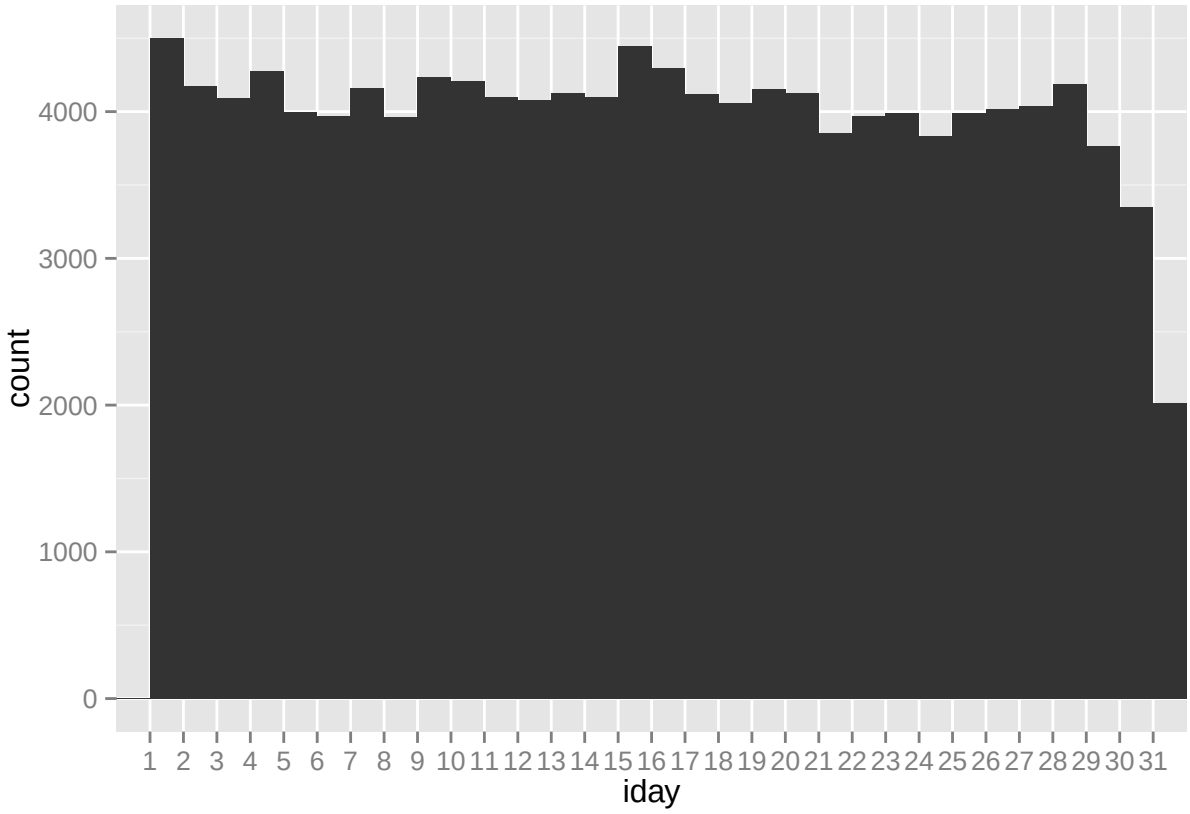
Single Variable Explorations





The first plot that I plotted was a histogram of years which shows an almost linear increase in incidence of terrorist attacks up until 1992 where there is a gap in the data set. This is due to a loss of data in the data set, which resulted in a total number of incidents that only totaled 15% of the previous estimate of incident numbers. While there is not specific data there is an estimate in the explanatory document of 4954 incidents in that year. This would make sense given the decrease in number of incidents in 1994. There is a significantly lower number of incidents between 1998 and 2004, before spiking again in what looks like an exponential increase.

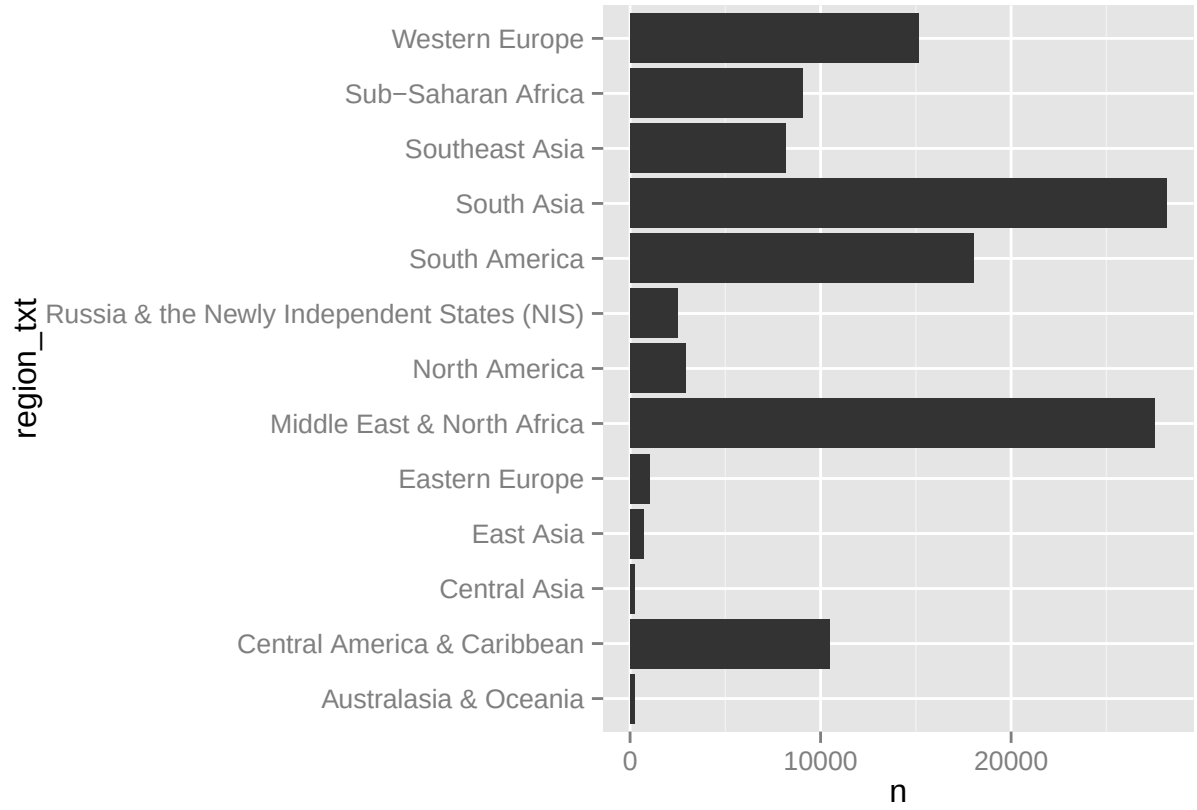
There doesn't seem to be much variance in the month variable.



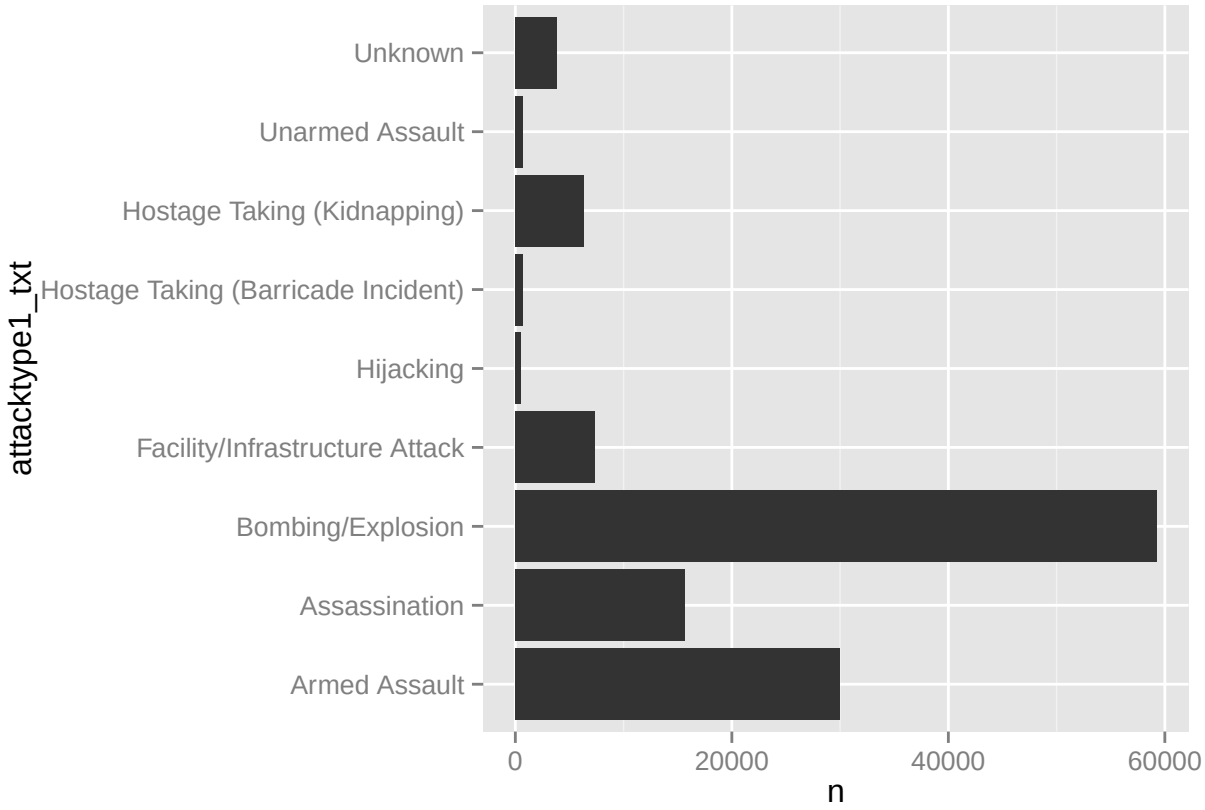
When I looked at the tally for days, there does not seem to be much variance

---

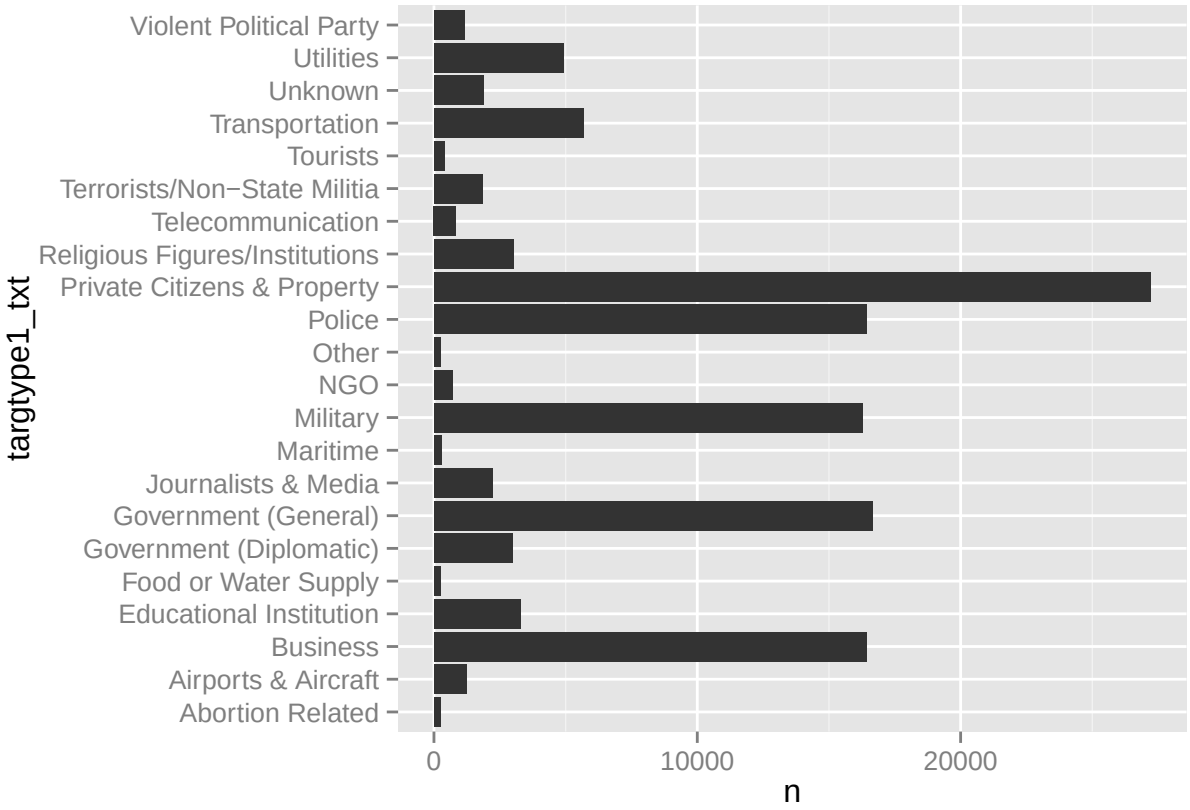
Now I am going to take a look at the categorical variables



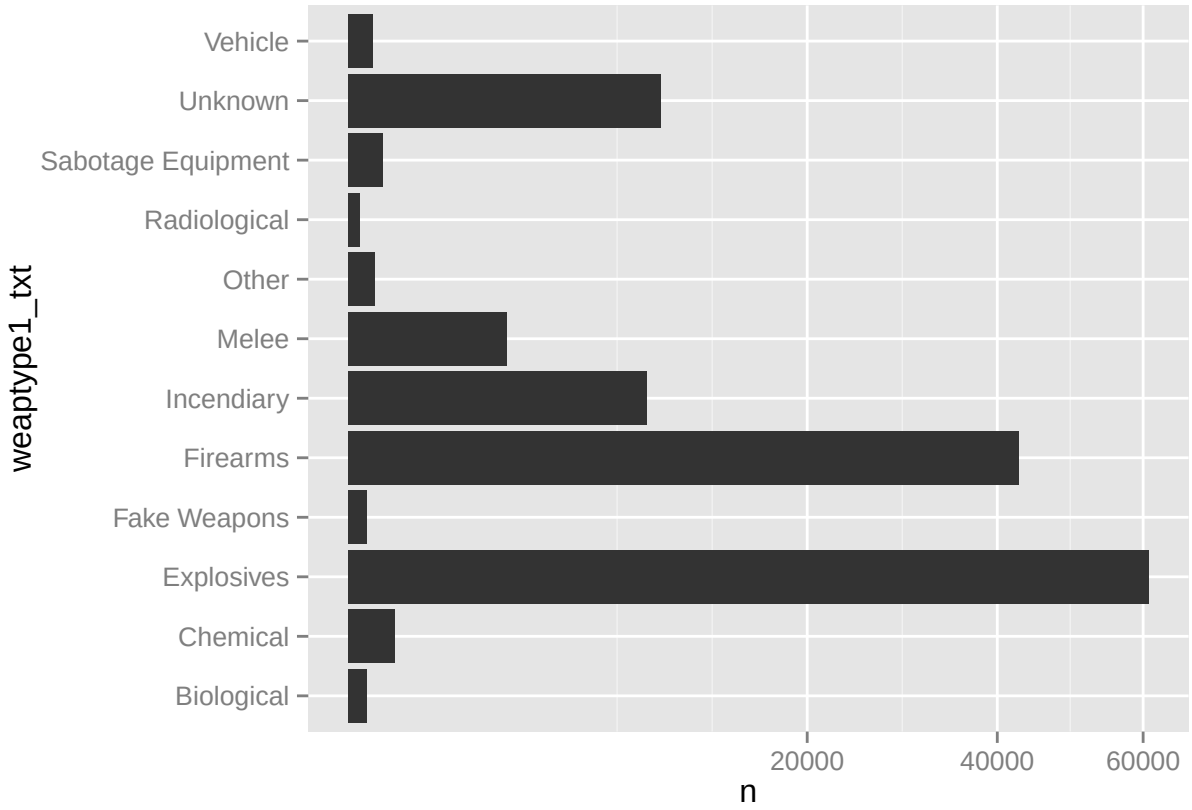
The next plot is of the region variable. This shows a low number in North America, East Asia, Central Asia, Eastern Europe, Russia, and Australasia. With high numbers in Central America, South America, Western Europe, and the Middle East.



This graph shows that the three most commonly reported terorist incidents are Assassinations, Armed Assaults, and Bombings. It aslo shows that the least common terrorist actions are Barricade Incidents, Hijackings, and Unarmed Assaults.



The bar chart is displaying the distribution of target types shows that there are 5 most common target types with the most common being Private Citizens, and the other four being businesses, government, police, and military buildings.



Firearms and explosives are the the most common weapon types, as shown in the plot

---

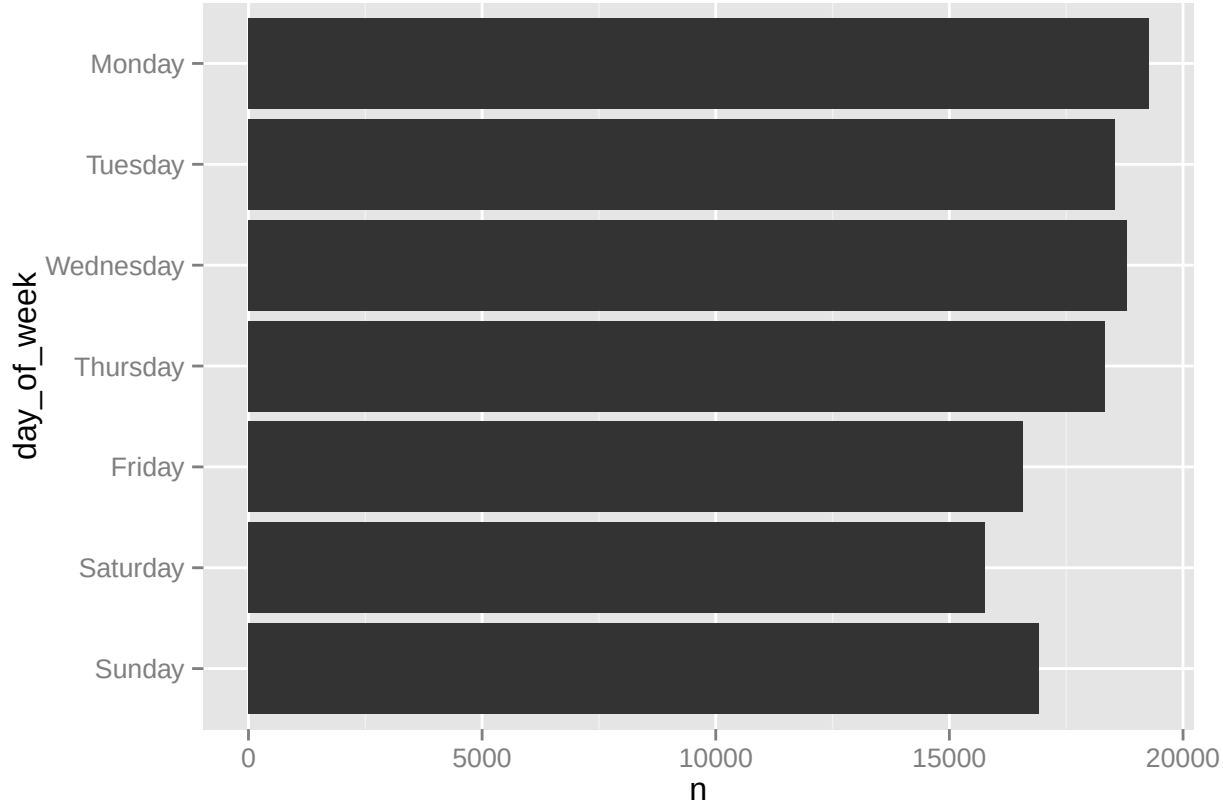
```
## Source: local data frame [209 x 2]
##
##            country_txt    n
## 1          Afghanistan 5931
## 2              Albania   71
## 3              Algeria 2660
## 4              Andorra    1
## 5               Angola  474
## 6  Antigua and Barbuda    2
## 7            Argentina  797
## 8              Armenia   20
## 9            Australia   73
## 10             Austria  105
## ..                 ...  ...
```

```
## Source: local data frame [27,859 x 2]
##
##                                  city n
## 1                     'A'Ishah Bakkar 1
## 2                             'Alayh 1
## 3                      'Ayta al-Sha'b 1
## 4     (Finca Los Manzanos) Las Palmas 1
```

```
## 5  (Lesotho) at Harakolo in Kulinyama 1
## 6                   (Mar Tagla) Beirut 1
## 7                                    * 6
## 8                 10 mi S. of Kampala 1
## 9                   15 September Dam 1
## 10    15 km from South African border 1
## ..                                 ... .
```

This showed me some interesting information namely that there are about 28,000 city values, and that many of the city values actually hold a value describing the location of the incident. So, I will not likely use that feature in my visualizations. There were too many values in the country and city variables to plot effectively

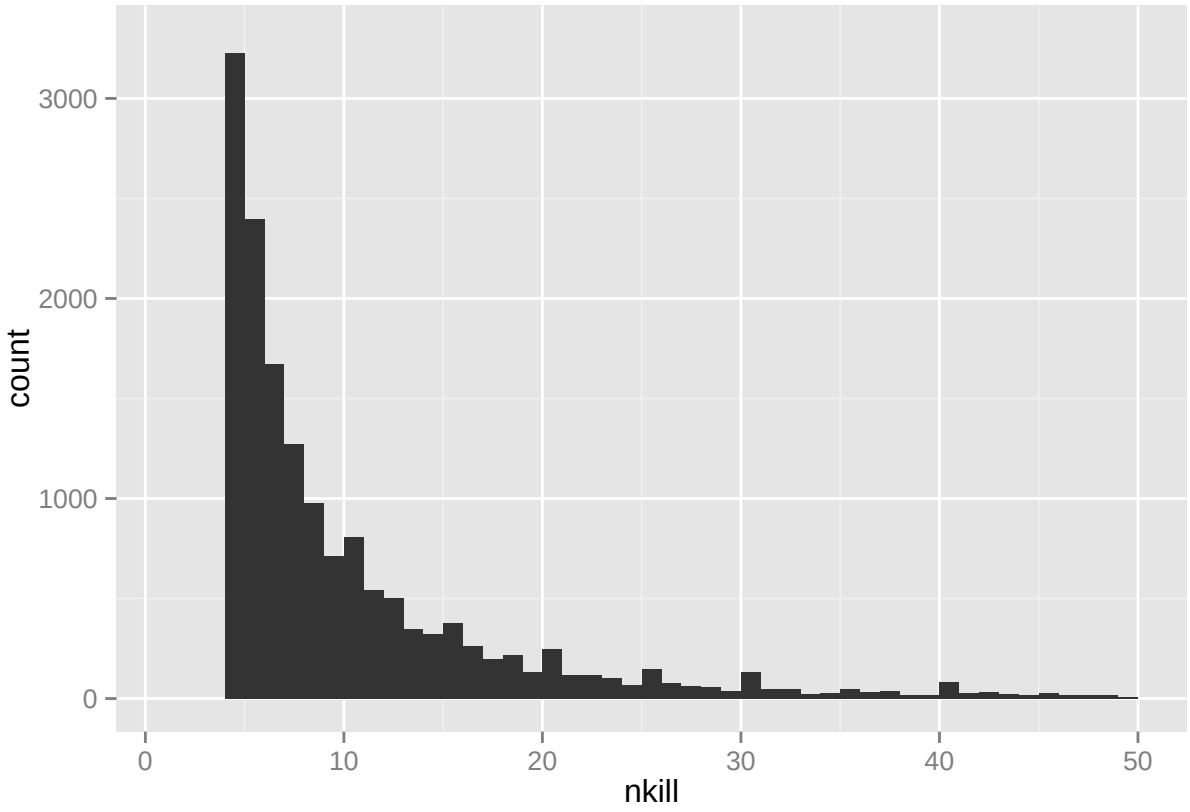
---



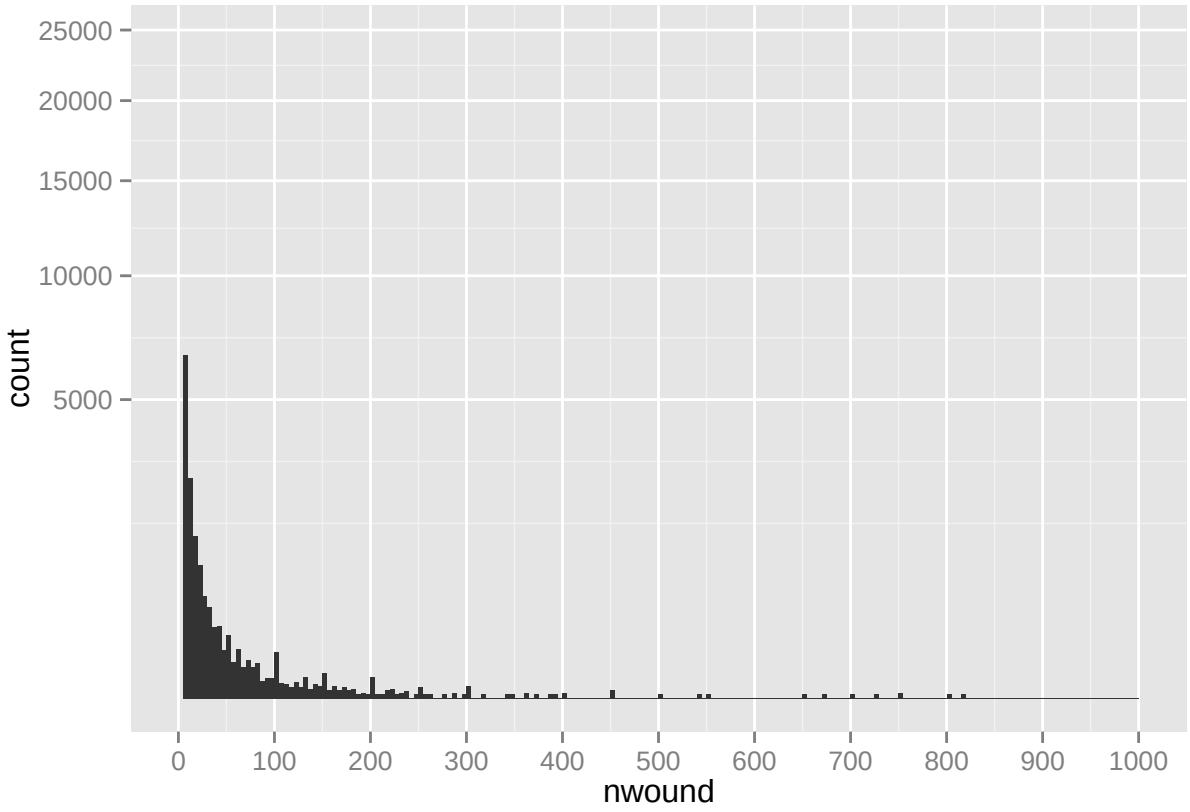
There seems to be a decrease through the latter half of the week, with Friday, Saturday, and Sunday having significantly fewer incidents than there are on Monday, Tuesday, Wednesday, and Thursday

---

So, the next thing that I wanted to take a look at was the distribution of number killed. In this histogram I noticed that there was a long tail, and that the number of attacks with less than ten kiled was a very large number. So, I made another histogram looking at less than 50 killed, and set the binwidth to 1 so that I could see what the most common number killed was. It turned out to be 0.



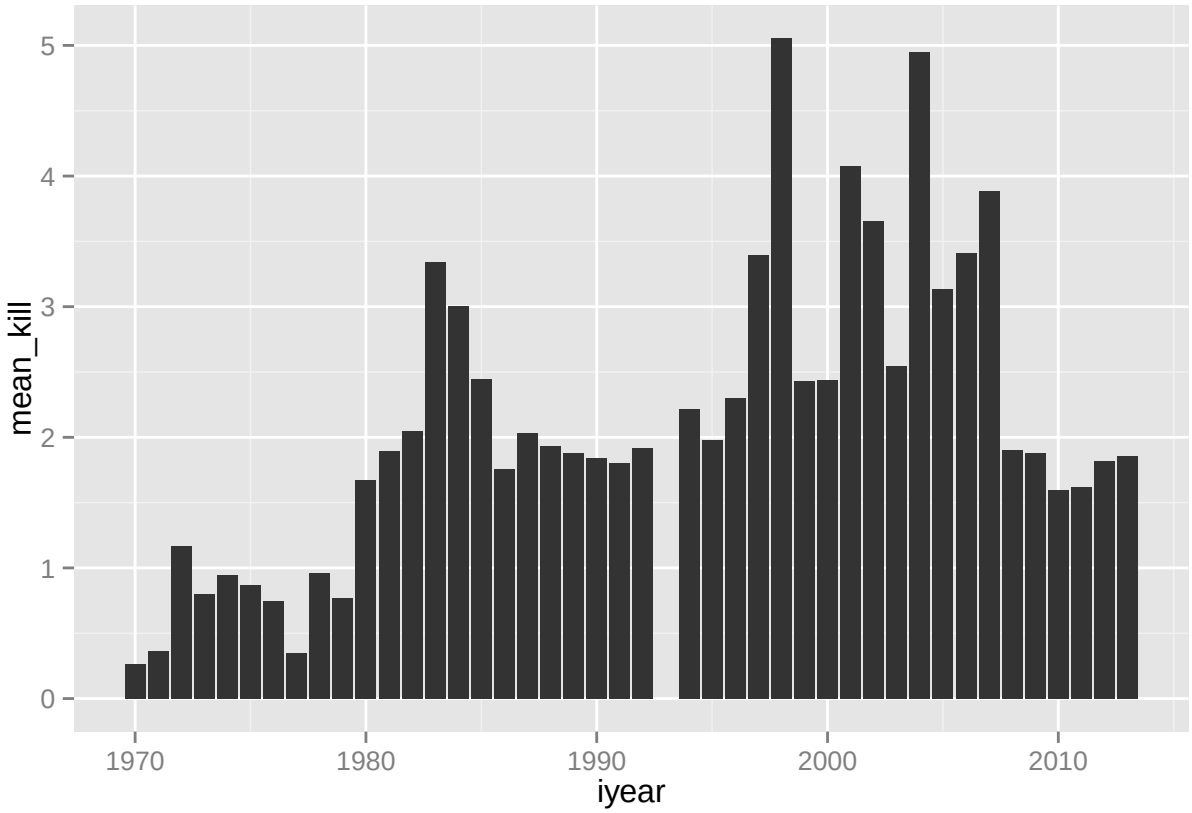
Having looked at the number killed, next logical step was to look at the number wounded.



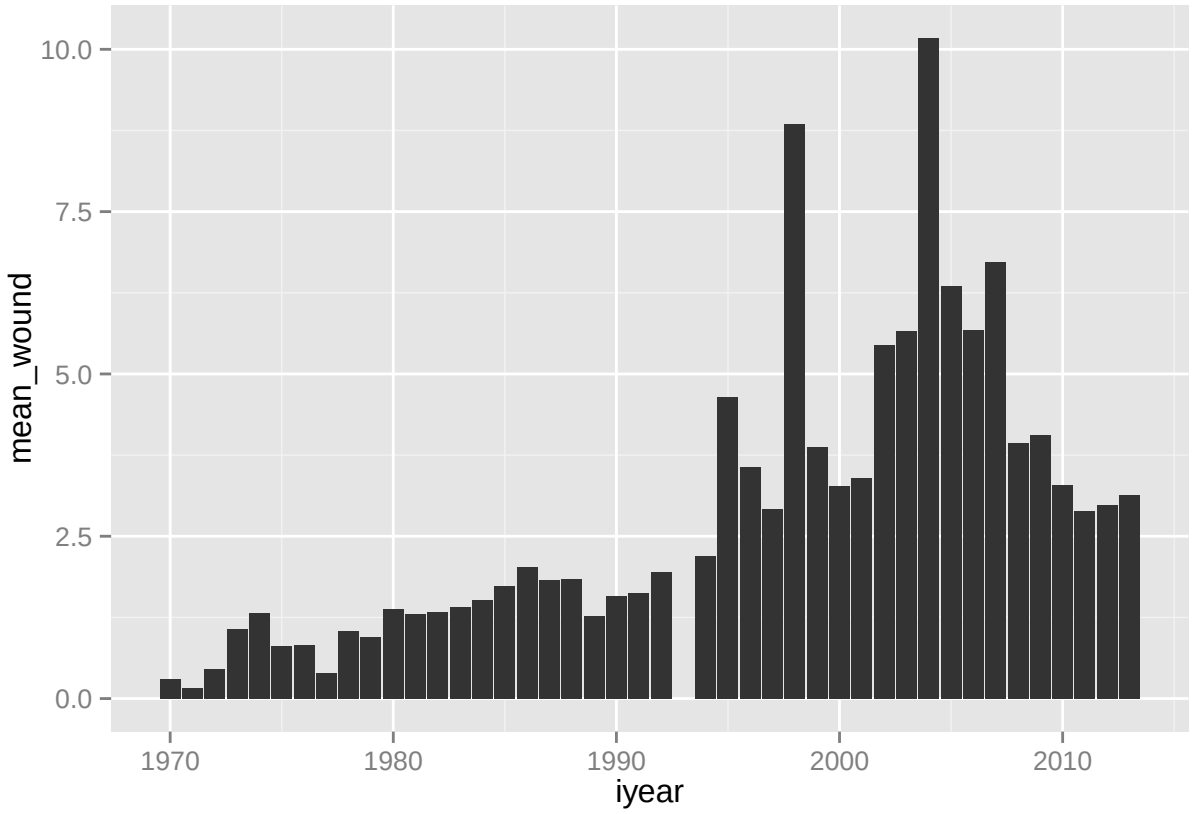
This histogram shows that, as above with the number killed variable, the majority of incidents have no wounded.

Two Variable Explorations

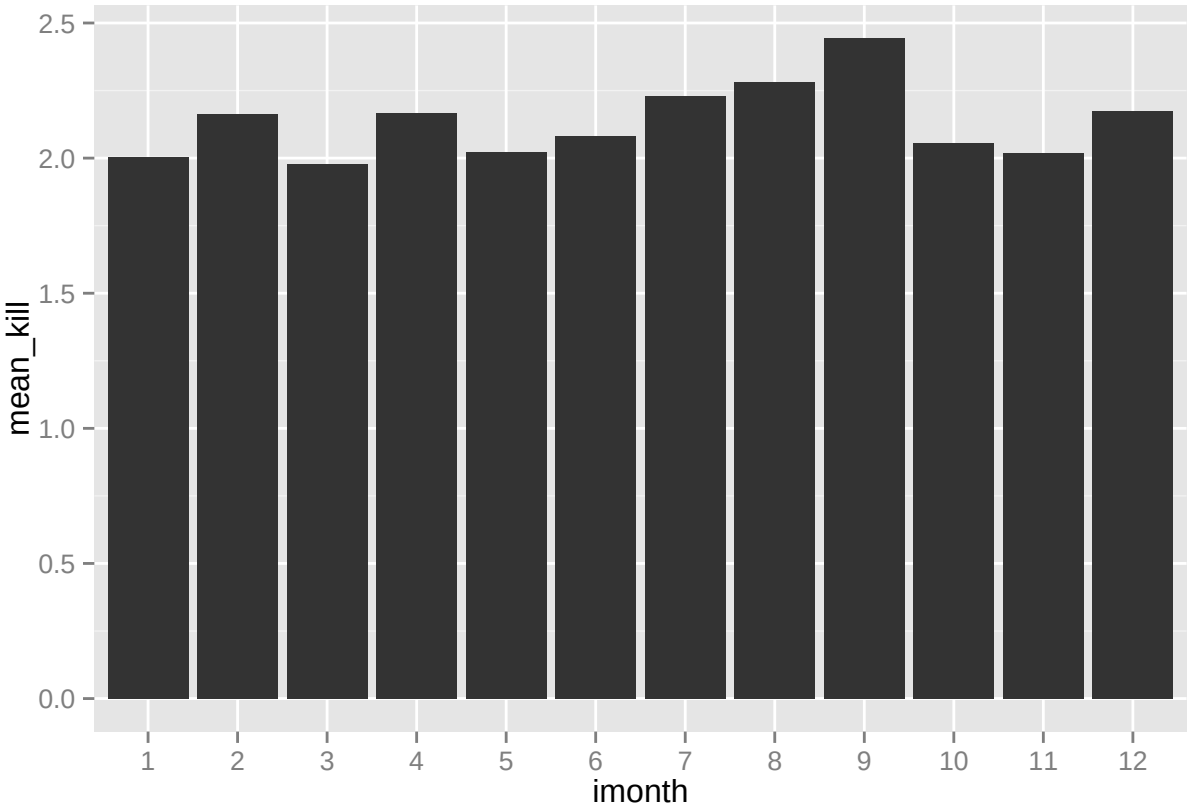
Now I am going to take a look at the relationship between two variables



The first visualization that I decided to plot was taking a look at the relationship between year and number of people killed in incidents. This shows fairly clearly that while there are many terrorist attacks per year with significant number of people killed the average number of people killed is still rather small.



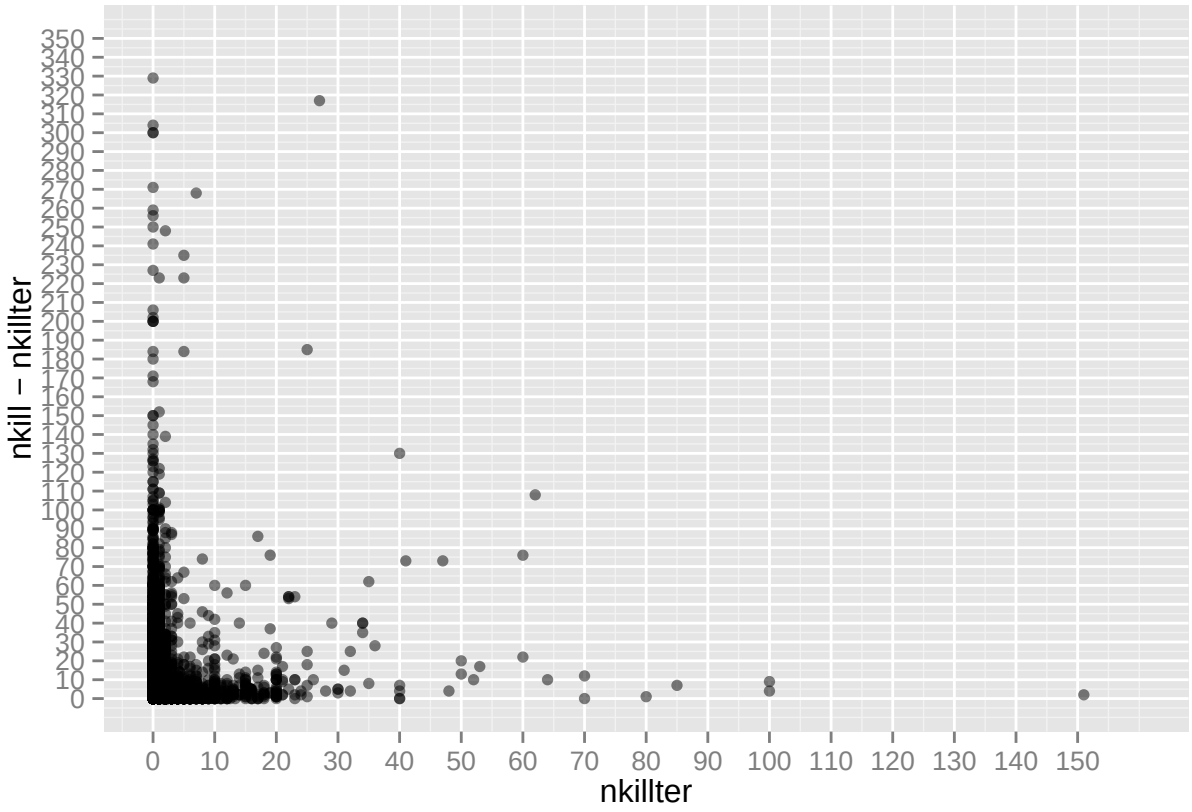
I thought it might be interesting to take a look at the number of people injured by year. The mean value per year is a fair bit higher than the number number of people killed per year.



There seems between May and September to be an increase in the average number of people killed. But that seems to be the only trend in the data

---

I wanted to see if there was any correlation between the number of terrorists killed in/after an incident and the number of people killed in the incident.

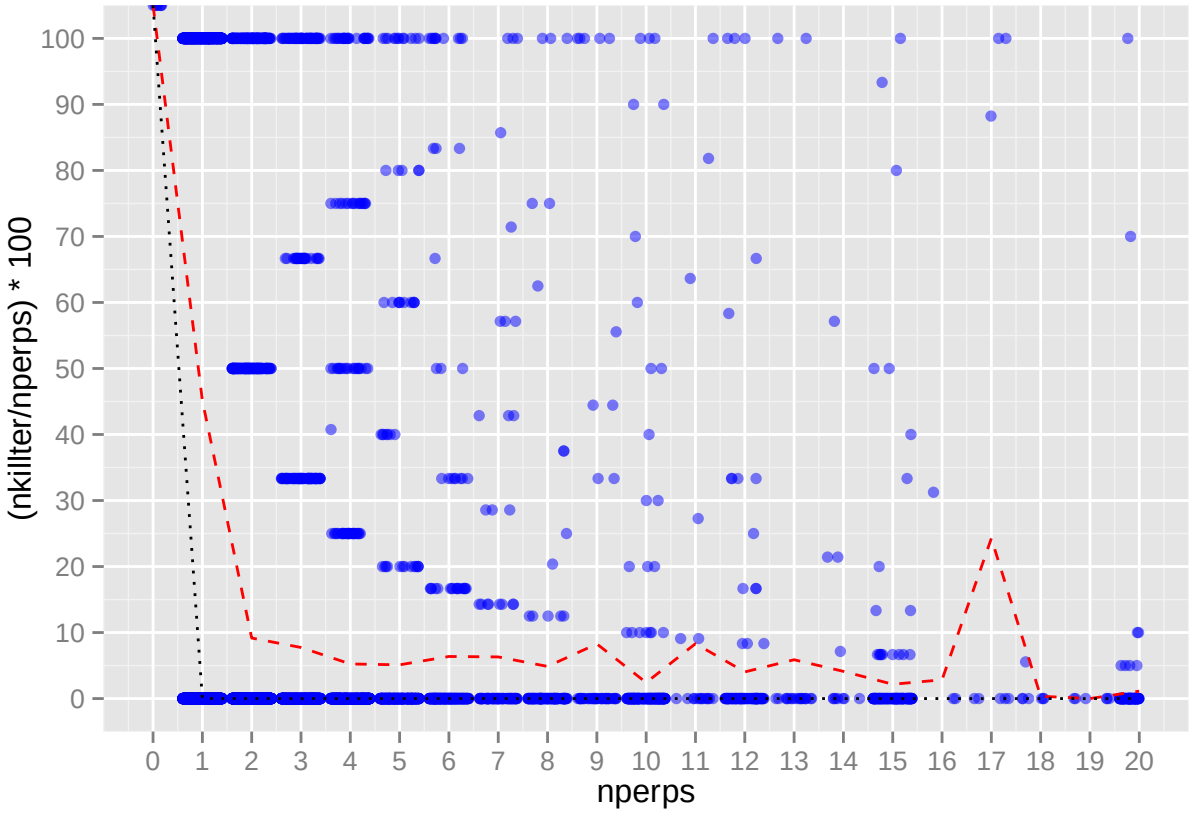



```
##
##  Pearson's product-moment correlation
##
## data:  data$nkillter and civilian_fatalities
## t = 16.33, df = 124189, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.04073 0.05183
## sample estimates:
##     cor
## 0.04628
```

This plot shows a rather sharp clear line where nkillter = 0. I subtracted number of terrorists killed from the total number killed because the number of terrorists killed is included in the total number killed. The correlation cooeficient of 0.046 suggests a very small correlation between the two fatality numbers

---

After looking at that I wanted to see what the common percentage of terrorists involved in incidents that were killed. I decided to graph the percentage against the number of terrorists involved.



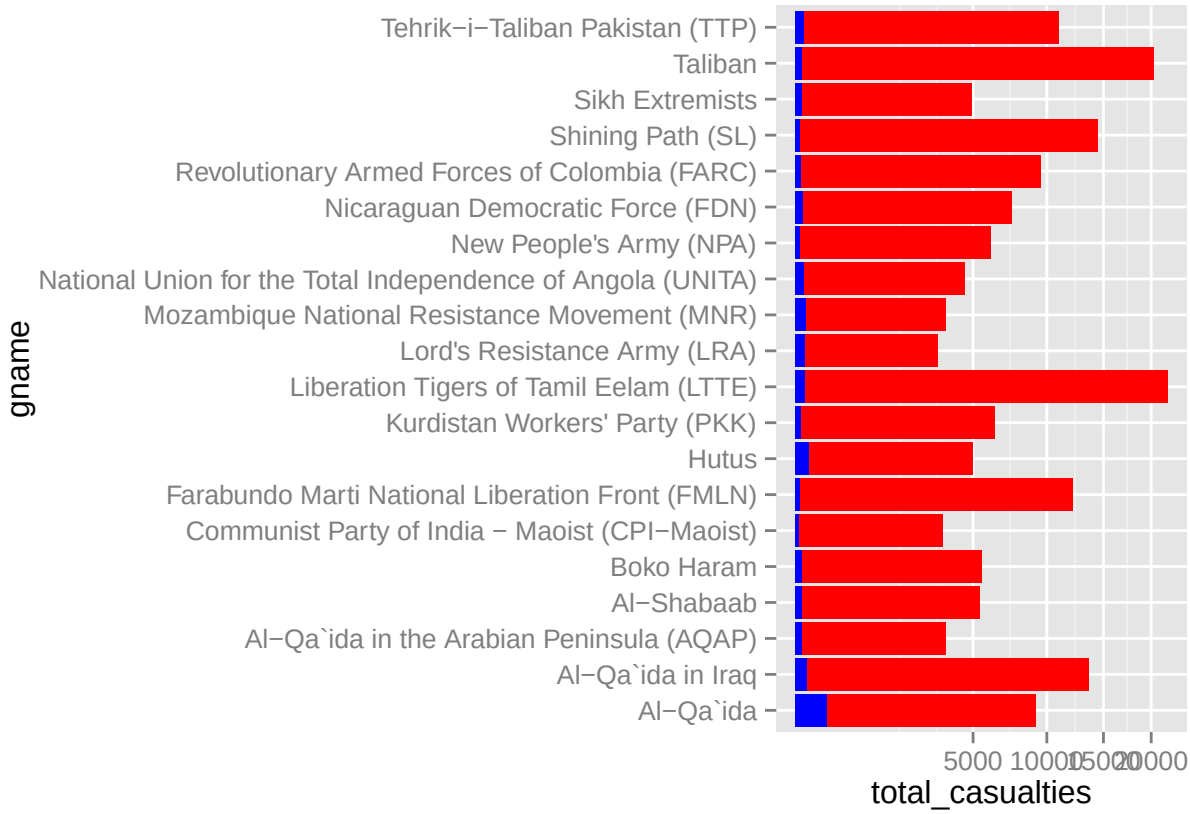
This plot compares the total number of perpetrators involved in terrorist acts to the number of terrorists killed. I notice that most of the data points are clustered in the lower left corner, so I plotted the lower left corner of the graph, and that there are some rediculous outliers with 25,000 perpetrators, etc. That seemed a little extreme to me so I replotted the graph (several times). There was also an outlier with a percentage terrorist killed of over 1500% so I set the y scale limits to between 0 and 100, as it represents a percentage

It looked as though numbers over about 50 perps involved are estimaes, as they seem for the most part to fall on numbers divisible by 50. I am going to plot one more plot based on this looking at the lower left corner of the graph yet again. This time zeroing in on the sub-20 perpetrator corner.

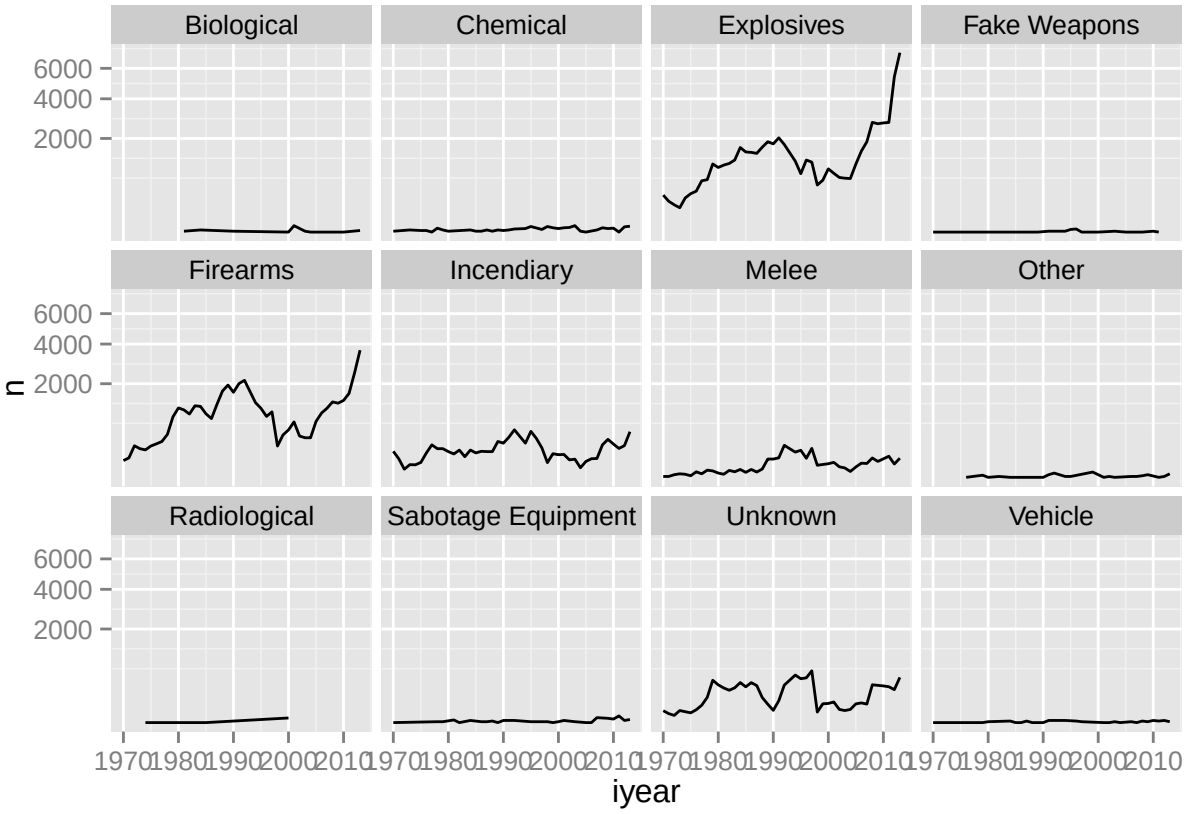
In the final iteration of the plot, the black dotted line shows the median value, and the red dashed line shows the mean value. After 1 on the x axis the median dropps to 0 which is expected, as most terrorist acts don't seem to have any fatalities (civilian or otherwise). But, the mean seems to hover around 10% of the total number of pepetrators. I find that interesting.

The plot has a horizontal bar at zero that stretches all the way across the plot. This shows that across all number of perpetrators there are many incidents with no resulting deaths to the terrorists involved.

Multivariate Plots

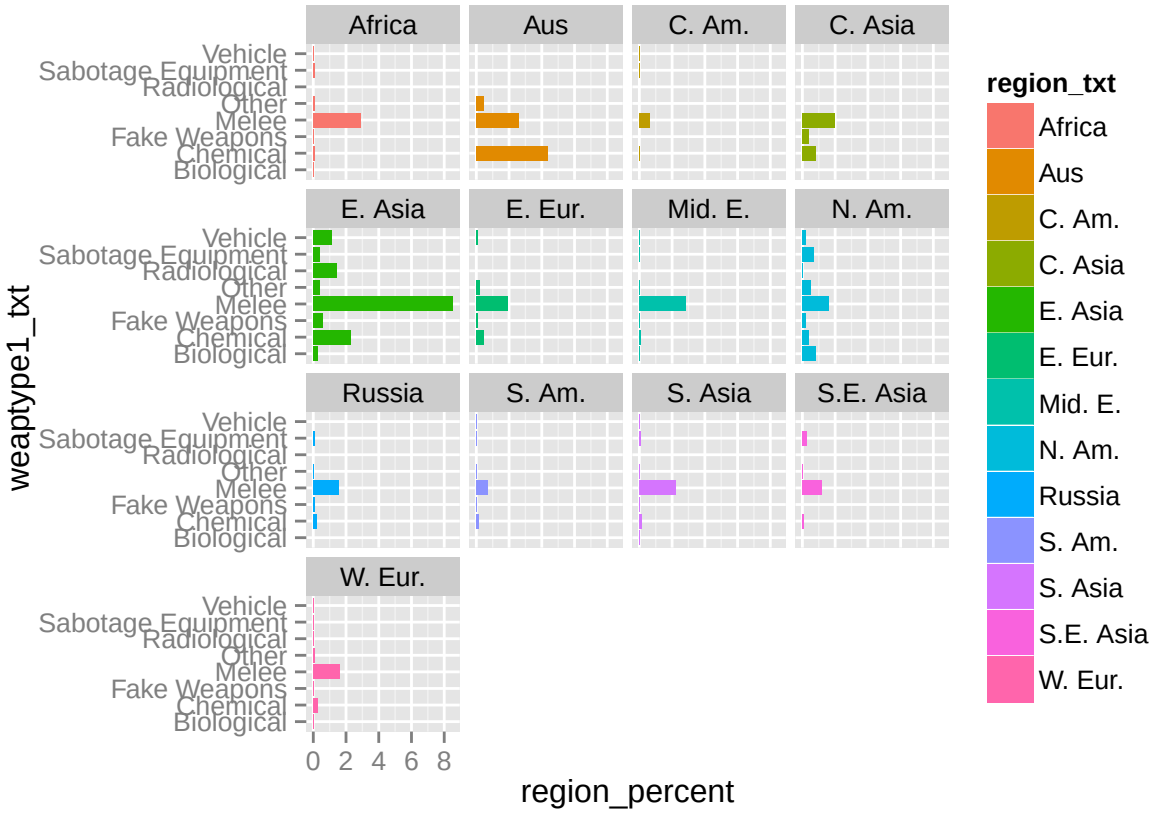


This is a visualization showing the top 20 terrorist groups as shown by the number of casualties. I find it interesting that a goup that I have never heard of has the most casualties. In this graph the red bars show the total number of casualties of each group's attacks, and the blue bar show's the average number.



The above plot shows the trend in the incidence of specific weapon types over the time period the data set covers. It shows that there are only 5 commonly used weapon types, even though all have been used. Those which are not common are almost flat distributions.

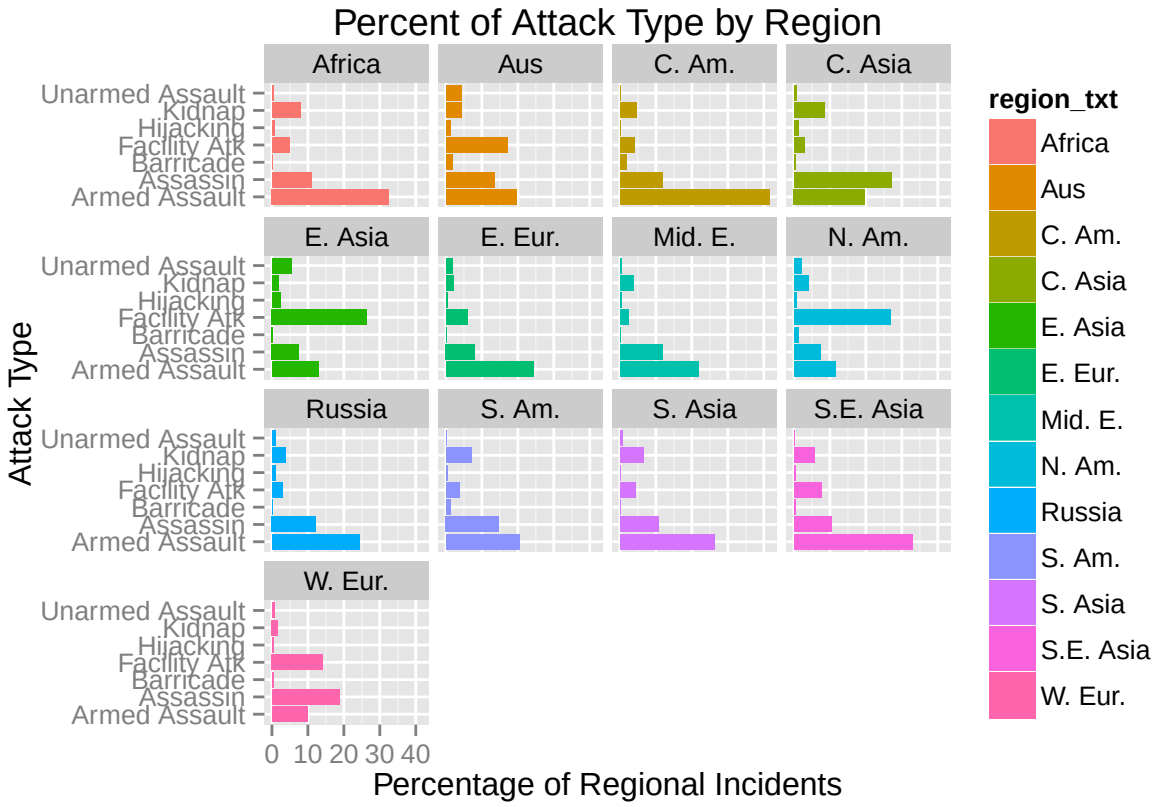
I wondered if there were regional specialties in Weapon Type. So, the plot below demonstrateswhich regions use what weapon types most often. I renamed the regions so that it would be a bit clearer in the plot. And I removed the four most common weapon types world wide (Firearms, Incendiary, Explosives, and Unknown).



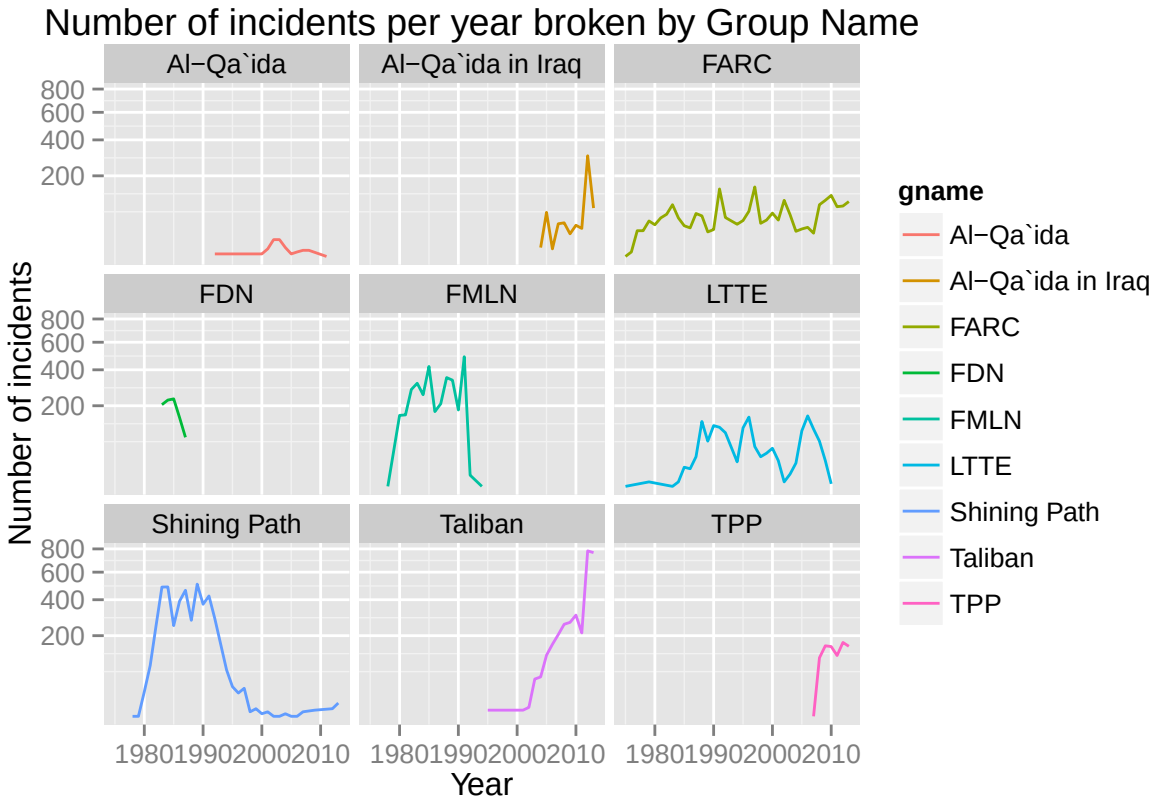
The plot shows that there is a fair bit of regional variance. With only East Asia having examples of all weapons being used. I calculated a percentage of the total incidents that each weapon and plotted those by region. Showing that Australasia and Oceania have a much higher level of Chemical attacks, and Central America seems to only have melee attacks when you take out the most common weapon types.

Final Visualizations and Summary

The final visualization that I wanted to take a look at in this section built on the previous one. I wanted to see if there was a preferred attack type by region. So I applied all of the above methods to the data set again, this time looking at attack types



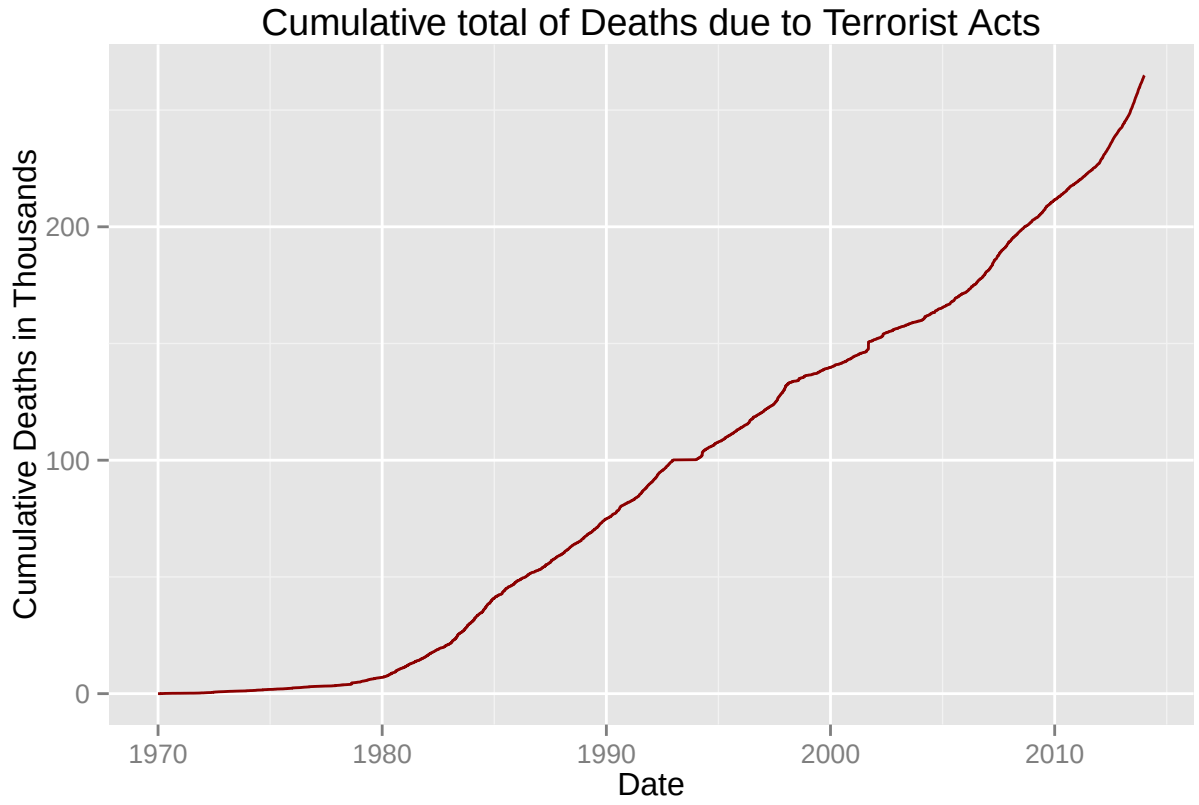
The above plot shows, as the one centred around weapon types did, regional variances in the attack type preferred by terrorist organisations in the area. Central America shows the greatest percentage of Armed Assaults across the regions, and Central Asia seems to prefer Assassinations.



This plot shows the attack number by year of the 9 most casualty causing terrorist organizations. I notice that only three of the organizations have distributions across the entire time span, with most of them being relatively short lived but prolific. Al-Qa'ida has an interesting distribution. They are 8th in terms of most Casualties caused and yet their number of incidents is very low. Even Al-Qa'ida in Iraq is showing a decreasing trend up to last year. There appear to only be two groups that are exhibiting a substantial increase in incident numbers. Shining path, who have been relatively quiet for most of the last decade, and the Taliban which has been climbing a great deal since they started their climb, shortly after 2000. The Taliban's current climb is very steep, and I will be looking forward to perusing the next release of the data set to see if the trend continues.

I first decided to take a look at this data set, because I first started looking at the material for the course Exploratory Data Analysis with R in September hoping to prepare for the Data Analyst Nanodegree Program. I work in the security industry, and was living in the US on September 11th 2001. At work in the weeks leading up to the anniversary of that terrible day, there was a lot of talk about that incident and terrorism in general. I found the data set, and did some very basic plots to show people at work. I found the data set again on my computer when I set out to start this project, and thought I would use it for the project.

My last visualization is a pretty stark reminder of how terrible this pattern is.



This last plot shows the cumulative rise in deaths attributed to terrorism over the last 43 years. The 70's showed a slow but steady climb, that shot up through the 80's and 90's, and became steeper still in 2005, and again in 2011. More than 260,000 lives have been lost to Terrorist acts since 1970.

Reflections

I found this dataset a challenge to use. There are a many more variables than I was interested in taking a look at. The first hard decision was trying to figure out what I was going to look at in this project. Many of the variables also have too many possible values to graph properly. I feel too that the subset of interesting values I have taken a look at are too narrow, and don't really show anything interesting about the data. I feel that I have had many successes in the look at the dataset that I have taken. The first of which being figuring out how to construct the date variable from the available information. Secondly the structuring group by calls to put together an analsis of group involvement brought me a great deal of joy. As for future analysis, I plan to do a much more in depth analysis of this data using all of the tools that I now have available to me, including python's data analysis tools, and MongoDB. I would like to take a look at Incident types by region, possibly country, to see if there are any patterns there. I would like to analyse the groups, and see if I can group them and do an analysis of their spread, and efficacy over time, and many others.