To do:

Frame the success vs fatality thing when introducing the problem

Do I need to revisit inferential statistics?

Machine learning + conclusions

**Capstone Project 1**

Josh You

**Problem**

Terrorism is a major social problem because of the direct damage caused by terrorist attacks and because of its broader negative side-effects to society, such as widespread fear and political instability. This project seeks to analyze what factors predict whether a given terrorist attack, or attempted terrorist attack, will be successful.

**Client**

The client for this project would be government agencies and security officials that are seeking to prevent acts of terrorism. The results of this analysis may help them make better decisions about which sources of potential terrorist attacks to focus on (for example, whether to focus on individuals or organizations), and which types and means of attack are most likely to succeed. The analysis may be relevant for both broad regulatory responses to terrorism (such as weapons control), and how governments should respond to specific threats of suspected terrorist plots.

**Data**

The data for this project will come from the Global Terrorism Dataset compiled by the National Consortium for the Study of Terrorism and Responses to Terrorism (START), which is a research center based at the University of Maryland. The dataset contains information on 170,000 terrorist attacks between 1970 and 2016. The dataset is publicly available [here](https://www.kaggle.com/START-UMD/gtd).

The dataset contains over 100 variables describing various characteristics of each attack. The main outcome of interest is success, which is a binary categorical variable, indicating whether the attack achieved a tangible outcome. For example, a bombing attack is successful if a bomb detonates, though the bombing may not necessarily lead to fatalities or achieve the perpetrators' intended long-run goals. The dataset also contains information on the number of people killed and injured in the attack, and the extent of property damage. Other variables of interest include the region where the attack occurred, attack type (which includes categories such as assassination, armed assault, bombing, hijacking, and others), weapon type, and characteristics of the target.

There are several other organizations who compile data on terrorist attacks, and these datasets would be potential alternatives to the dataset I use. The [Chicago Project on Security and Terrorism (CPOST)](http://cpostdata.uchicago.edu/search_new.php) has compiled a database of all suicide attacks from 1974 through June 2016. While this data is restricted to suicide attacks, it contains additional biographical information on the individuals who carried out the attacks. The [RAND Corporation](https://www.rand.org/nsrd/projects/terrorism-incidents.html) maintains a dataset of terrorist attacks from 1968 through 2009.

**Data Cleaning and Wrangling**

The first step I took in cleaning my dataset was identifying which columns were relevant to my project. The original dataset had over 100 columns. I narrowed down the variables I am interested in to year, month, region, attack type, target type, weapon type, success (whether the attack was successful), number killed, latitude, and longitude.

Next, I dealt with missing values - The dataset did not contain values for every column I was interested in. Many values were missing in that the fields were blank:

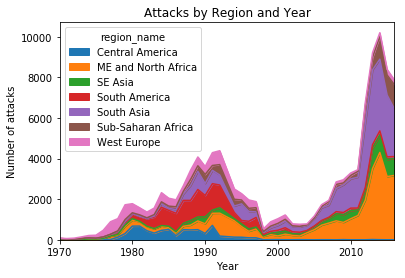
The dataset also encoded missing values using specific codes indicating that the value for a column is unknown. For example, attack type is encoded by a number between 1 and 9, where 1 through 8 represent types of attacks, while 9 indicates that the type is unknown. I removed data points with missing values, though I had to be careful in which columns I used to filter. In this dataset, attack type, target type, and weapon type are represented by several columns because some events fall under multiple types of attack, target, and weapon. However, most rows have only one type for each of these sets of columns, so I only filtered based on whether the first column had a missing value (if the first column has a missing value, the second and third will have missing values too).

Cleaning the dataset to remove missing values brought the number of data points down to 140928, compared to the original 170,350.

The next data wrangling step was pivoting the categorical features. This meant separating variables with multiple possible values into a binary (dummy) variable for each of these values. Thus, each of the nine attack types was given its own column, with a value of 1 or 0 depending on whether a given attack had that type. There are two motivations for this. First, weapon type, attack type, and target type were represented by multiple columns, because a terrorist incident may involve multiple weapons and targets, and may be described using more than one attack category. However, the second, third and fourth columns for these categories had many missing values, so turning these columns into dummy variables allowed me to retain all of that information while removing missing values. Second, the machine learning libraries I used required that categorical variables be converted to dummy variables in order to be interpreted correctly. Without converting them to dummy variables, these variables would be interpreted as quantities. For example, an attack type of 8 would be treated as 8 of *something*, and as more similar to an attack type of 9 than an attack type of 2.

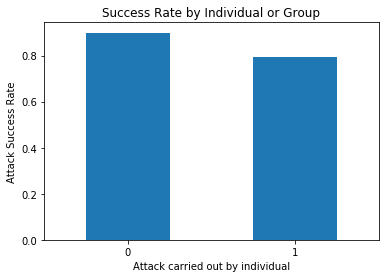
**Exploratory Data Analysis**

I conducted exploratory data analysis on the cleaned dataset, and found a few interesting patterns. Here is a stacked area chart showing the number of attacks per year in the seven regions with the highest number of attacks.

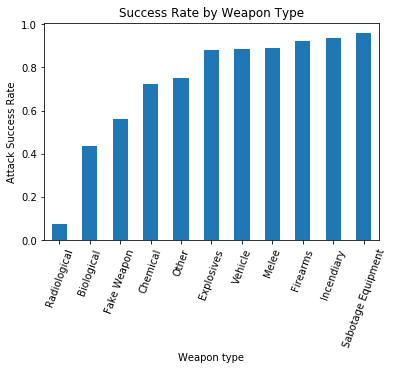


As this plot shows, there is a spike in terrorist attacks in the late 1980s and early 1990s in this dataset, and a larger spike occurring after 2010. There are clear differences in the distribution of attacks by region over time. South Asia and the Middle East and North Africa appear to be the main contributors to the recent spike in attacks. South America and Western Europe have seen large declines in both the absolute number of attacks, and their share in the world total, over time.

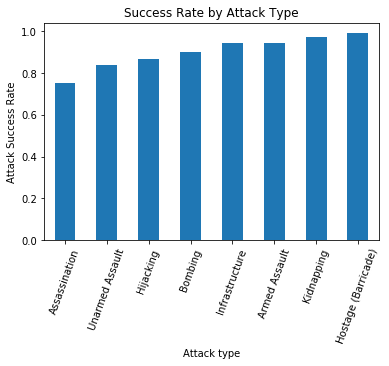
In addition, I found that terrorist attacks carried out by individuals had a lower success rate in this dataset, as shown in this graph.



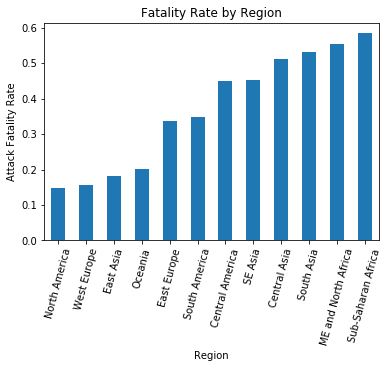
There were also large differences in success rates based on the type of weapon used. Radiological attacks rarely succeed, while attacks with sabotage equipment, incendiary weapons, and firearms usually succeed.



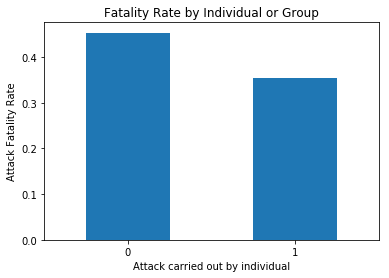
Success rates do not differ as much between attack types. However, assassinations are less successful than other types of attacks, while kidnappings and hostage attacks are usually successful.



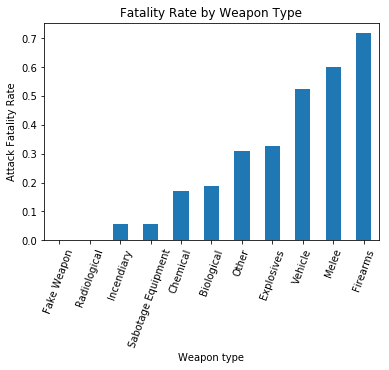
Next, I explored which factors affected the fatality rate of terrorist attacks (the proportion of attacks that result in more than one death). First, I show fatality rates by region:



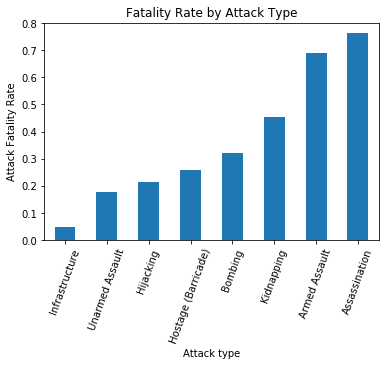
Most attacks lead to fatalities in Africa and the Middle East, while most attacks do not lead to fatalities in North America, Western Europe, and East Asia. There appears to be a relationship between the economic success and political stability of a region and a low fatality rate of attacks.



Attacks carried out by individuals have a lower fatality rate, in addition to having a lower success rate.



This plot shows large differences in fatality rates between weapon types. Unsurprisingly, firearms have a relatively high fatality rate. Attacks using fake weapons and radiological weapons do not lead to any fatalities in this data set.



Finally, are large differences in fatality rate based on attack type. There differences are largely as one would expect: most assassinations and armed assaults lead to fatalities. Most infrastructure attacks and unarmed assaults do not lead to any deaths.

**Inferential Statistics**

I conducted several hypothesis tests based on the patterns found in the exploratory data analysis.

First, I tested the null hypothesis that the success rate of attacks attempted by unaffiliated individualsis the same as the success rate of attacks attemped by groups, against the alternative hypothesis is that the success rate for attacks attempt by unaffiliated individuals is smaller than the success rate for group attacks. I used a z-test for the difference in proportions, resulting in a p-value of 2.004\*10−12. Thus, I reject the null hypothesis and conclude that group attacks have a higher success rate. In addition, I found that the 95% confidence interval for the difference in success rates was [0.067, 0.148], i.e. 6.7 percentage points to 14.8 percentage points.

Next, I used a chi-square test for independence to test the null hypothesis that the success rate does not differ by region, compared with the alternative hypothesis that success rates do differ by region. This resulted in a p-value of 4.96\*10−277, so I reject the null hypothesis and conclude that attack success rates do vary by region.

I also used a chi-square test for independence to test the null hypothesis that success rate does not change based on attack type against the alternative hypothesis that success rate does vary by attack type. The p-value was small enough to be reported as 0.0 by the software package I used, so I reject the null hypothesis and conclude that success rates do vary by attack type.

**Classification analysis**

I used several machine learning algorithms to build a classifier that will predict the outcome of a terrorist attack based on its characteristics.

First, [did stuff with success]. I tried using logistic regression and random forests. Because the data set is unbalanced (almost 90% of attacks in this dataset are successful), I had low recall for unsuccessful attacks. This meant that the classifier would tend to mark almost all attacks as successful, and could only classify at most around 20% of unsuccessful attacks as unsuccessful. I decided that I change the focus of this project from trying to predict the success of terrorist attacks. From a practical standpoint, there is no real benefit to trying to distinguish the small percentage of observable terrorist attacks that are unsuccessful - a 90% risk of a terrorist attack succeeding is not much different from a 100% risk. If I had a different dataset containing all the potential attacks in earlier planning stages that were not known to the public, then predicting the success of those plots would have greater practical value.

Thus, I turned to classifying whether a terrorist attack will lead to any fatalities. I first created a binary variable encoding whether the number of individuals that were killed was greater than zero, which would be the outcome variable.

Logistic regression:

Random forests:

Also I actually have to run it on the test set lol

**Discussion and Recommendations**