Data Analytics Summary Sheet
Project name: Peace and Love

- Overview

We downloaded the dataset from Kaggle. This dataset contains information about almost all terrorism activities from 1970 to 2017 around the world. The purpose of this project is to understand the characteristics or patterns of global terrorism and present them by multiple data visualization methods. We have conducted the following analysis:

- ❖ Present the changes in numbers of attack in different time and in different region
- ❖ Identify the most attacked countries and the most active terrorist groups
- ❖ Illustrate the characteristics of terrorism target in United States

Besides, we used Logistic Regression, Random Forest and SVM models to predict the chances of success for a potential terrorist attack.

- Data Visualization

The highlight of our project lands on the various visualization tools that were utilized. We strived to look at data from different perspectives and present them in a way that help us better understand these terrorism activities. This process is driven by the questions that we wanted to ask as we discovered more while working.

Firstly, we looked at how the the total number of global terrorism activities has changed from 1970 to 2017. We drew a animation to show the changes and it's clear that the total number of attacks has increased and the range of region has expanded significantly. Then, we had a close look on the four most-attacked countries: Iraq, Pakistan, Afghanistan and India. These four countries have the same tendency in the increase of number and the expansion of region as the rest of the world.

We wanted to know the motives behind the attacks, so we did a word cloud on motives. It was of no surprise that most motives are political or religious factors.We also studied the relationship between regions and attack types, and found out that the most common attack types are bombing, armed assault and assassination.

There were two major questions for us to solve at this stage: what are the most active groups and what have they done; and what happened in these most-attacked countries? We draw eight graphs grouped by four most attacked countries, analyzing them separately with respect to the attack groups and attack types. Most attacks in Iraq and Syria are conducted by ISIL(ISIS), most attacks in Pakistan are conducted by Taliban. Most of the attacks occurred in Afghanistan are unknown, which indicates that there are no people or groups claim to be responsible for the attacks. We finally show two graphs to see the countries suffering most by ISIL and Taliban.

Most attacks of ISIS take place in Syria and Turkey. Besides, most attacks of Taliban occur in Pakistan.

Last but not least, we concentrated on the data in the United States. The relationship between target type, number of injuries and kills are revealed by dividing the terrorists activities into ten groups by their targets (e.g. schools, hospitals, businesses, etc.) and counting the number of wounds and kills in each group. This interactive graph shows us the total casualties for each target type. The result indicates that businesses are always on the list for terrorists and they also have very hgih number of total injuries and kills.

- Model Analysis

Logistic Regression, Random Forest and SVM models are used for prediction.

We first performed correlation matrix and found that most variables in our dataset are independent, and variables that are highly correlated are not chosen for our model thus pose no influence on the result.

In our models, 8 features are chosen as input, and the success is the outcome.

X ∼country, region, attack type, individual, suicide, weapon type, target type, gun certain

Y ∼success

The success variable is a binary variable, if the attack was successfully carried out, success=1, otherwise success=0.

We split our dataset into training set(80%*80%), validation set(80%*20%) and test set(20%). All models were trained on training set, best model was selected on validation set, and finally prediction was made on test set.

We choose Random Forest as our final prediction model because of its highest accuracy on validation set (0.908616), and then we refitted the model with both training set and validation set. The accuracy of Random Forest on test set is 0.905020.

Later we reviewed the feature importance for Random Forest. According to the chart, *gun certain, target type*, and *weapon type* are three most important factors for the result, followed by *suicide, individual, attack type, region* and *country*. Obviously, as guns are more portable and harder to detect than most other weapon choices, it is critical in helping one carry out an attack.

- Conclusion

We seem to be living in a more unsettling world, and we took this project as the opportunity to understand the reasons and patterns of previous terrorist attacks. We hope for stricter gun control, and a more peaceful world in the years to come.