Team Name: ***Another Day, Another Data***

Team Members:
Joseph Veltri (Team Leader)
jvelt1@unh.newhaven.edu

Khadija Khorakiwala
kkhor1@unh.newhaven.edu

Akhila Pitla
apitl1@unh.newhaven.edu

Research Question: How does the level of risk for each country correlate to the terrorism that each country endures.

Data Set: The data set that we have selected contains data from the time of 1970-2017 that includes the type of terrorism event, the exact date it happened (month, day, and year), the exact location it happened (city and country), longitude and latitude, a summary of the event, the type of attack, location it was targeting (what kind of organization, business, etc), and who the attack was from.

A list of the techniques that we used are:

- Problem Definition
- Data Preparation
- Data Exploration
- Evaluation

Data Exploration is about describing the data by means of statistical and visualization techniques. We explore data to bring important aspects of that data into focus for further analysis.

Problem Definition – Our problem that the group was trying to solve, and the objective is to clean (clear missing values) the dataset and to visualize the data to find the level of risk for each country correlate to the terrorism that each country endures. As expected, the objective was achieved during our analysis.

Data Preparation –
We used the ETL method in the preparation of our data and we had some complications in the process. There were some missing values in the dataset (which we resolved) that prolonged our preparation and analysis. The ETL method is Extraction, Transformation, and Loading which basically extracts data from sources, then cleans and converts the data, and then provides the ability to load data into a different database via update, insert, or delete.

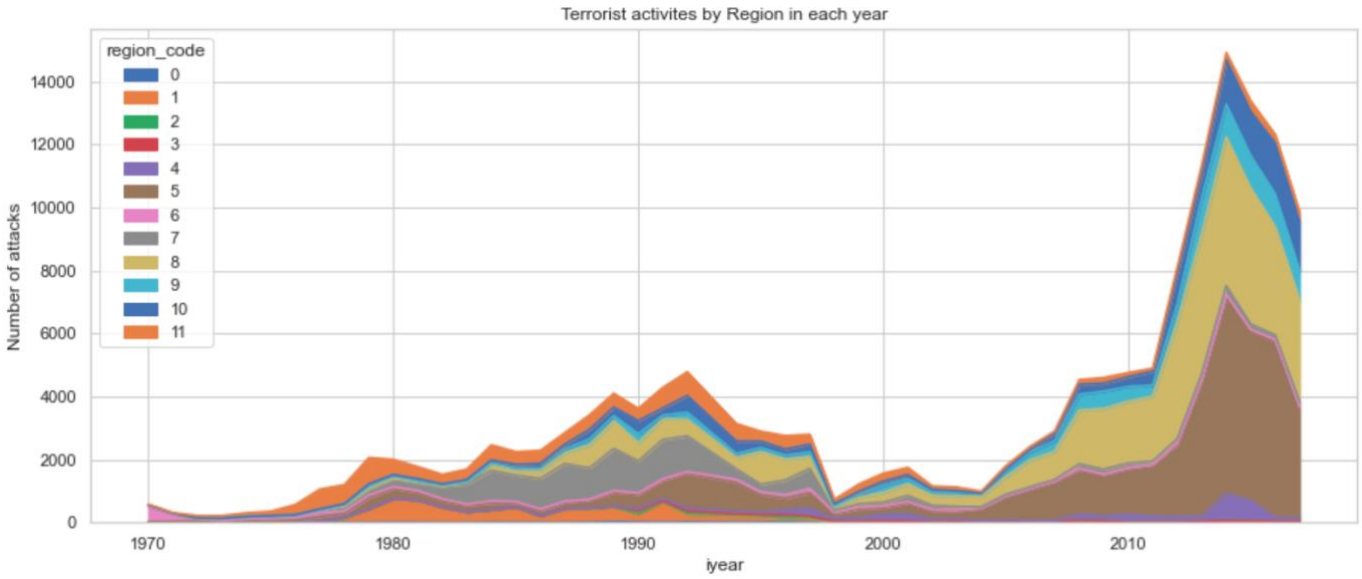
Data Exploration –
As shown in the code screenshots and graphs below, a method that was used was the Bivariate Analysis. In each of our graphs and charts, we used all the different variables in the dataset and examined the relationship

between them all to see which are significant and which are outliers. The type of Bivariate Analysis that was used was Numerical & Categorical because some data types were numerical as dates, number of attacks, longitude and latitude and some were categorical with type of attack, country, and geographical connection.

GitHub Link: https://github.com/joeveltri/Phase-6-Optimization.git

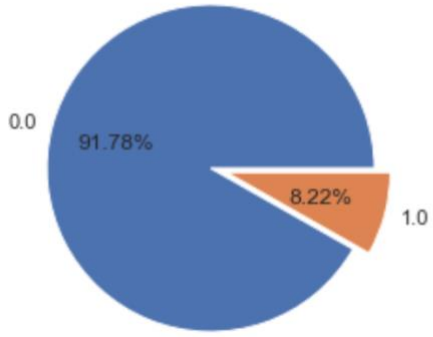
Conclusion – The method that was used in this phase were Gradient Boost Classifier and Decision Tree classifier. A decision tree combines some decisions, whereas a random forest combines several decision trees. Thus, it is a long process, yet slow. We used Decision Tree because it is easier to understand and interpret. These methods increased the accuracy percentage and helped the group by the data preprocessing. After all the data Cleaning and removing the null values and separating and clearing all the numerical and categorical values are big factors why the data went from a 59% accuracy rate to 94%. For example, there's an attribute called iyear. This value can't be 0, so the group removed all the values that has zeros in this attribute. Same way was done with all the important attributes (numerical and categorical). Outlier detection was also used in this process. The purpose of optimization is to achieve the "best" design relative to a set of prioritized criteria or constraints. These include maximizing factors such as productivity, strength, reliability, longevity, efficiency, and utilization. That is why all techniques that the group used in order to enhance the accuracy percentage was necessary. Each technique that was used may not have been the best in terms of performance metric, but the group was able to accomplish the task at hand.

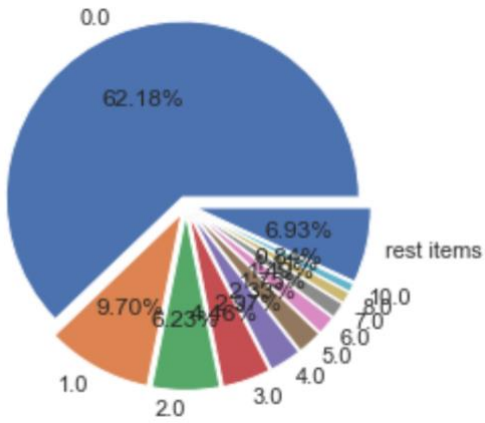
Some graphs and charts are listed below:

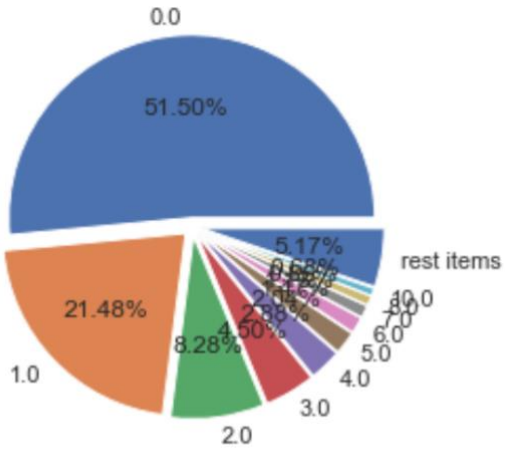


These graphs are when the victims are being counted and when 0 is taken out of the data because iyear cannot be 0. Since we have hundreds of nulls in these 7 features in the graphs, the group filled them with 'other' to make them ready for label encoder. The group labeled encoded all of the categorical features and dropped the original features.

```
Number of unique vaure are 3 which are :
 [0.0, 1.0, nan]
```






0.0
51.50%
rest items
21.48%
1.0
8.28%
2.0
3.0
4.0
5.0
6.0
7.0
8.0
10.0