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-5-Modeling-Data

Outcome – The hardware that was used for this phase was R Studio.56.5% of the dataset has missing variables in it, so a lot of information we, as a group, could not get. Some sets that have more than 25-30% missing values were automatically taken out. The acknowledgement that the rest of the missing variables got for the remainder of the dataset were based on the information that we initially got. Many variables have types of data such as 'int' but are places as categorical variables. The algorithms for Decision Tree were selected as it enabled better classification results. Since it is better for classification results, we used it because it was based on GINI value and entropy not like the regression model, as that is more for probability. All information, graphs, charts, etc. can be seen on the PDF document in our GitHub repository.