Homework 3

Group Members: James Willson, Kun Li, Sean Kugele

I. Objectives and Activities

Objective: Our objective is to create statistical models from the Global Terrorism Database data set for the purposes of prediction and inference. Specific research goal are outlined in the Section II, along with the principal data scientist assigned to each. Proposed solutions and methods will be discussed in the context of each research goal.

Data Set: The *Global Terrorism Database* contains over 180,000 observations and 135 variables. Each observation corresponds to a terrorism incident that occurred between 1970 and 2017. The type of information available for each incident is very diverse, including (but not limited to) the following:

1. *Date Variables* (Year, Month, Day, etc.)
2. *Geospatial Variables* (Lat/Long, Region, Country, City, Province, etc.)
3. *Incident Descriptive Variables* (Attack Type, Duration of Incident, Success/Failure, Weapons Used, Targets, etc.)
4. *Perpetrator Descriptive Variables* (Terrorist group membership, # Perpetrators, etc.)
5. *Casualty and Damage Variables* (# Fatalities, # Injured, etc.)

The diversity of the variables in this data set lends itself to a variety of research directions, and supports the creation of numerous regression and classification models.

II. Research Questions and Proposed Approaches with Individual Assignments

The sub-sections below identify the agenda and responsibilities of each data scientist. Each data scientist is expected to take ownership of their agenda, including implementation of solutions, final write-up, and presentation. Sean Kugele will serve as team lead, facilitating group collaboration and assisting in the resolution of any impediments that jeopardize the fulfillment of the group's research goals.

James Willson (Data Scientist)

Goal 1: Predict if an attack will be successful based on a variety of different factors.

Proposed Solutions: Since this is a classification problem, various models, such as Linear Regression or Decision Trees, will be attempted. Several potential issues might arise; for instance, backwards selection may not be possible as the data is so sparse that using all the variables to begin with would leave no rows to model from. Also, certain features, while being good predictors might not make much sense in the model. An example would be *Fatalities*; while it would clearly be a good predictor of a successful attack, it would be almost useless in practice as you would never know the fatalities until after the attack had occurred and its success status was already known.

Goal 2: Estimate the number of casualties in a successful terrorist attack.

Proposed Solutions: Given the nature of the problem, a method such as linear regression will probably be the best choice. Many of the same issues regarding the selection of features will apply to this problem as well. Again, many otherwise relevant features might excluded based on the intended use of the model, that being attempting to find ways to minimize casualties *before* attacks occur and/or find terrorist groups or other features that seem to be more dangerous as they lead to higher casualty numbers.

Kun Li (Data Scientist)

Goal 1: Predict the extent/dollar-amount of property damage from any given attack.

Proposed Solutions: Most likely will try to use a linear regression to model the damage extent. In order to find the best predictors, it would be helpful to test all the factors. One could do so efficiently using the stepwise regression and recursive feature elimination. The concern is definitely the data quality, and one might have to remove many variables before conducting the regressions.

Goal 2: Identify factors that could predict the target/victim type in an attack.

Proposed Solutions: One could use linear regression and decision trees to approach this multiclass classification problem. A potential problem that one could face is that the target types are centered around a few specific categories, such that there isn't sufficient data for the less common target types to train the model.

Sean Kugele (Data Scientist, Team Lead)

Goal 1: Predict the terrorist group responsible for perpetrating a terrorist attack.

Proposed Solutions: A variety of classification methods will be attempted and compared including logisitic regression, lda, qda, knn, decision trees, support vector machines, etc. to identify the model that can most accurately predict the terrorist group responsible for terrorist attacks in a "test" dataset. The models will be further analyzed to identify the most significant features contributing to the prediction of responsible terrorist groups.

Goal 2: Estimate the probability of an attack based on temporal and geo-spatial variables.

Proposed Solutions: This data set has a rich set of temporal and geo-spatial variables that we will use to develop one or more models predicting the probability of a terrorist attack. Classification models will be created (for example, using logistic regression, lda, and knn) to identify a usable model. Heat maps, or other relevant visualizations, will be created to visually identify regions and time periods (e.g., months, holidays, days of week, etc.) by level of attack risk.

Goal 3: Identify *clusters* of incidents that reveal interesting patterns in the data.

Proposed Solutions: Given the large number of variables in this dataset, it seems advantageous to apply a unsupervised learning algorithm (like K-means) to identify patterns of regularity in the data. The features included in this analysis may need to be adjusted to identify interesting patterns. Clusters will be visualized and interpreted.

III. Methods of Evaluation

All models will be evaluated based on appropriate measures of model quality (e.g., adjusted-R^2, AIC, BIC, etc.) and ROC curves will be created for classification algorithms. The data set will be partitioned into training and testing data sets, and model preditions will be evaluated based on the testing data set, which will only contain observations that were withheld from model training. Cross-validation will be also be applied to determine model quality.

IV. Potential (General) Difficulties and Concerns

Preliminary analysis of the dataset has revealed that the data is sparsely populated and will require a significant amount of data cleansing before it will be usable. Many of the variables are also redundant; therefore, the actual number of usable variables is likely closer to 90.

The data set spans over four decades of observations, and the data was collected by multiple agencies. As a result, the data may be inconsistent in quality, and the data collection methodology that was used may have changed over time. These factors will likely introduce sources of irreducible errors into models generated from this data.

According to the documentation, many of the variables were introduced after 1997; therefore, we will need to make a case-by-case decision about the inclusion of these older observations, depending on the nature of the research question and the needed features.