STAT 440: Final Project Summary Report

Project Title: Analysis of Terror Attacks

Group Members: Alex Wang, Jace Goetsch, Siddharth Ahuja, Waseem Ebrahim

i. **Introduction**

Due to the fact that governments can only allocate a limited amount of resources to prevent terrorist attacks from occurring, it is not possible to use a brute force approach where all possible avenues of terrorism are removed. It is necessary to use a more efficient approach. Priority must be given to the avenues of terrorism which have the greatest chance of resulting in a successful terrorist attack. In order to determine where a country's limited resources must be used it is necessary to collect and analyse data on terrorist attacks. The purpose of this project is to use SAS in order to prepare data that has been collected on terrorist attacks.

There are two datasets that have been used for this project. The first dataset is called the 'Global Terrorism Database' (GTD) (Link: https://www.start.umd.edu/gtd/contact/). This is a regularly maintained database which is annually updated by the University of Maryland. This data file contains data from terrorist attacks since 1970.

The second dataset is called the 'Chicago Project on Security and Threats' (CPOST) (Link: http://cpostdata.uchicago.edu/search_new.php). This is a regularly maintained database which is annually updated by the University of Chicago. Unlike, GTD which contains data on all types of terrorist attacks, CPOST only contains data for suicide attacks. The variables and other information on these datasets have been further discussed in the 'Methods' section of this report.

ii. **Methods**

The GTD dataset contains a total 170,000 observations and 109 variables in total. The CPOST dataset contains a total of 50,000 observations and 7 variables in total.

First in order to validate the data, invalid data was printed out by checking for factors such as invalid dates. The table for the the GTD dataset can be seen in Figure 3 in the Appendix

(There is no table printed out for the CPOST dataset in Figure 3 because the invalid data table was empty). Additionally, frequency tables and N-levels tables were created to confirm that the observations under each variable were inputted into the datasets as expected (Refer to Figure 4 in the Appendix).

There were several steps taken to clean the data. The first step was to merge the iyear, imonth and iday variables in the GTD dataset into a single 'Date' variable. Secondly, the variables which were not important for the project were dropped from the dataset (the variables which were kept have been discussed further below). Appropriate labels were created for these variables. Next, using the frequency tables from the validation step, observations which contained invalid data were excluded from the project data set. Additionally, eight different formats were created for several variables to convert numeric data into more readable character data. These formats were also created with the help of the frequency tables from the validation step as these told us how the numeric data was organized. The table that was created after performing all of these cleaning steps will be discussed in the Results section of this report.

Further data preparation was performed to create tables with descriptive statistics (Figure 7 and Figure 8 in Appendix) and frequency tables for the Number of Deaths in terrorist attacks (Figure 9 and 10 in Appendix). To create the frequency table an additional format was created in order to make the frequency tables more readable. This format shortened the frequency table from several hundred rows to eight rows of observations while still providing the important information. The analysis of these descriptive statistics and frequency tables has been provided in the Results section.

The important variables from the GTD data set are country, nkill, nperps, nwound, property, ransom, region, success, suicide, targtype1, weaptype1, attacktype1, iyear, imonth and iday. The important variables in the CPOST dataset are number_killed, number_wounded, location and attack_date. In order to see what each of these variables mean and what type is refer to Figure 5 and 6 in the Appendix (which are the contents procedures of the cleaned GTD and CPOST datasets) where the variable label and type can be seen. These important variables are the ones that were kept in the SAS data set for this project. These were deemed to be important because they give a good idea about the patterns within terrorist attacks, which is the purpose of this project.

iii. Results

(Only the first 10 observations from this table were printed out into this report to conserve space)

iv. Sources

Global Terrorism Database (GTD). 2016. [Data File]. Retrieved from https://www.start.umd.edu/gtd/contact/

Chicago Project on Security and Terrorism (CPOST). 2016. Suicide Attack Database (October 12, 2016 Release). [Data File]. Retrieved from http://cpostdata.uchicago.edu/

v. Appendix

(Only the first 10 observations from this table were printed out into this report to conserve space)