Arianna Lang Wang
Springboard Jan 2nd, 2018 Cohort

Global Terrorism Data Wrangling Step

Feb 9th, 2018

DATA IMPORTATION

I downloaded my raw data as a CSV file from the Kaggle website https://www.kaggle.com/START-UMD/gtd/data. Initially, I tried to import the data with Pandas package as df = pd.read_csv('global_terrorism_db_0617dist.csv') alone, but I received an error message "UnicodeDecodeError: 'utf-8' codec can't decode ..." After searching on Stack Overflow and trying a couple of different encoding methods, df = pd.read_csv('global_terrorism_db_0617dist.csv', encoding="ISO-8859-1") allowed me to import the data successfully.

DATA CLEANING

The data frame had 170,351 rows and 135 columns. After examining all 135 columns, I decided to keep 22 columns for later analysis.

The first three columns were 'year', 'month' and 'day' respectively. Each of the 'month' and 'day' columns has one missing value. After examining the missing value, I dropped them both. There are 20 zeros in the 'month' column and 891 zeros in the 'day' column. Zeros are meaningless for these two columns since you cannot have month 'zero' or day 'zero'; thus, they are considered 'missing values'. After some careful consideration, I decided to impute a random number from 1 to 12 for the missing month and a random number from 1 to 28 for the missing day. After that, I converted the type of values in all three columns from 'string' to 'int'. Finally, I used datetime.strptime to convert these three columns into a DateTime object and set it as the index.

I further cleaned the data by converting some columns' data types from float to int since these columns are ID's and thus do not need the decimal points. I also converted the 'success' and 'suicide' columns to categorical data type since both have binary data. Furthermore, I renamed some column names and dropped duplicate data by using Pandas' drop_duplicates() method. At this point, there are 158,285 rows left in the data frame.

Another crucial step during data cleaning is to check outliers. Columns 'N_killed' and 'n_wound’ are the only two columns in the data frame that have meaningful numeric data. After using the describe() method, I realized that both have very large maximum values, which is an indication for potential outliers. I further examined the potential outliers by drawing a boxplot on both columns and by subsetting the data frame with some conditionals. At this stage, I’m aware of the potential outliers but I can’t say for sure that they are real outliers.