Table of Contents

[data description - talks about source and type of data. (Sunna) 2](#_Toc490001990)

[Data Wrangling 2](#_Toc490001991)

[Data Profiling 2](#_Toc490001992)

[Data Normalization - show before and after of database schema (Patrick and Joshua) 2](#_Toc490001993)

[Data Analysis 3](#_Toc490001994)

[Summary - one paragraph recapping 1-5 above. (Sunna) 3](#_Toc490001995)

# data description - talks about source and type of data. (Sunna)

# Data Wrangling[[1]](#footnote-1)

Initially, the GT data would not import into MYSQL, so we imported the data into a Microsoft SQL Server database. We needed to make each column varchar(MAX) initially. Each column was surrounded by triple quotes, so we cleaned up the data. Afterwards, we wrote scripts to determine the max length of each column, and we altered the column width for each of the 127 columns accordingly. We then used sliding window functions to determine distribution of values for each of the 127 columns. We outputted the distribution of values per column into one table, DISTRIBUTION\_TABLE. Subsequently, we connected the SQL server to MYSQL to transfer the data using SSIS via an ODBC connection. After the data was in MYSQL, we performed a MYSQL dump of the data for the GT (I.e., the global terrorism table) and the DISTRIBUTION\_TABLE and uploaded it into box so that it could be loaded by the other teammates. This allowed the team to start with the same dataset.

# Data Profiling

Having moved the data to MYSQL, we ran scripts that determined which columns were sparse. The criterion for sparseness was that 80% or more of the column was NULL or contained ‘.’. We then created alter tables for the sparse columns and dropped them. At this point, the table was ready for normalization.

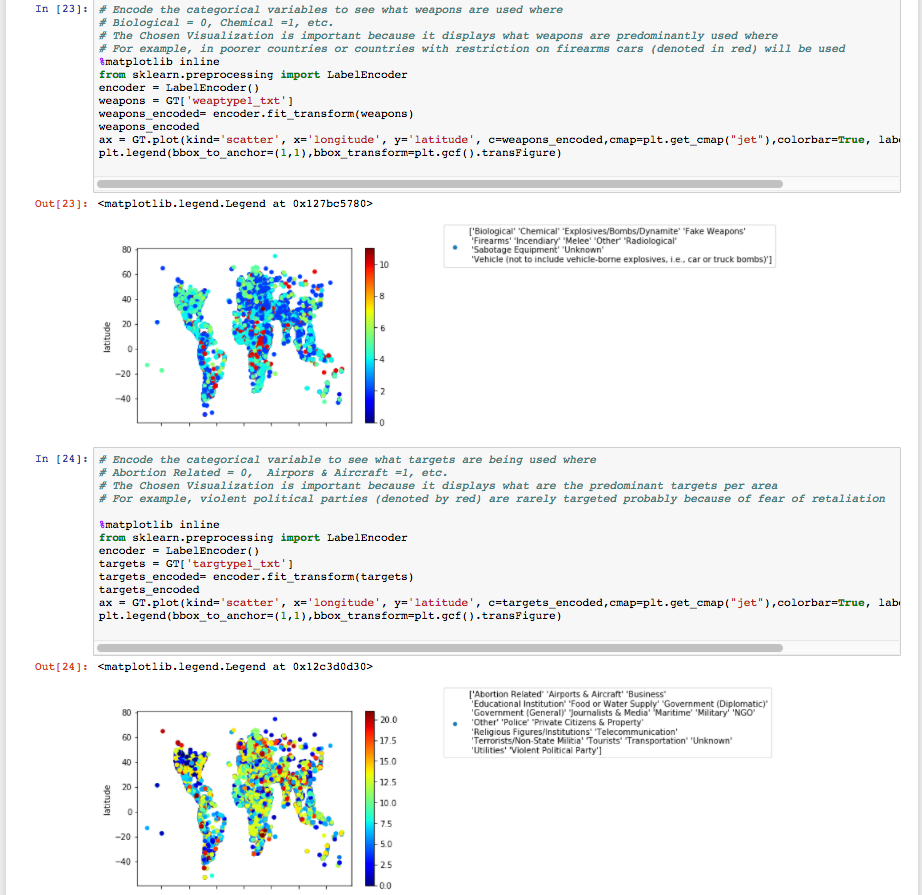
# Data Normalization - show before and after of database schema (Patrick and Joshua)

The data was profiled using SQL Server Data Profiler (see below). The profiler suggested that nine tables could be extracted from the large GT table. Individual tables were created using the columns from GT. These nine individual tables were validated by joining back to GT. The redundant columns were dropped form GT because they had been moved to the individual tables. Primary keys were created on the nine tables. From there, indices were created on the join columns in the GT table. Finally, a nine-way inner join was performed with the GT table to validate that everything has worked correctly.

|  |
| --- |
|  |
|  |

# Data Analysis

We used python to analyze the data, and found that weapon of choice in indigent countries is the motor vehicle. The favorite target of terrorists are abortion related centers and airports.. Below are the snippets.



# Summary - one paragraph recapping 1-5 above. (Sunna)

1. All the scripts that were created to wrangle, and analyze the data can be found here. <https://smu.box.com/s/zhbm2idikopnuhxfvcxtmc880nud5pxa>. There were 22 scripts written to wrangle and profile the data. The SQL dump generated to scripts, which can be found in the importdataintomysql folder. Professor, you have been given access to all of these scripts. They can be found on box. The stored procedures and sql used is for both microsoft sql server and mysql [↑](#footnote-ref-1)