EARIN Miniproject 2 Data mining

Preliminary version

Ernest Pokropek

2019

# Problem formulation

Gun violence is a huge problem especially for countries with availability of possessing firearms. Understanding some factors behind it, it is easier to predict and later prevent such drastic events and decrease the number of casualties.

# Aim

The aim of the data mining is to provide following information:

* Where gun violence is mostly occurring
* What is the magnitude of gun violence in those places
* How is gun violence correlated with factors such as age, place of occurrence
* How number of injures is correlated to number of killed and does it differ by any circumstances

# Dataset

Gun Violence Data - <https://www.kaggle.com/jameslko/gun-violence-data>

It is worthy to mention, that most of the data is from United States of America, thus only this country will be thoroughly considered and examined.

Dataset attributes and description:

* incident\_id – unique id of incident
* date – date of the crime
* state – state where crime has happened
* city\_or\_country – city or country where crime has happened
* address – address of the location of the crime
* n\_killed – number of people killed
* n\_injured – number of people injured
* incident\_url - URL regarding the incident
* source\_url - Reference to the reporting source
* incident\_url\_fields\_missing - TRUE if the incident\_url is present, FALSE otherwise
* congressional\_district - Congressional district id
* gun\_stolen - Status of guns involved in the crime (i.e. Unknown, Stolen, etc...)
* gun\_type - Typification of guns used in the crime
* incident\_characteristics - Characteristics of the incidence
* latitude - Location of the incident
* longitude - Location of the incident
* location\_description – description of the location of the crime
* n\_guns\_involved – number of guns involved in crime
* notes – additional information about the crime
* participant\_age – Age of participant(s) at the time of crime
* participant\_age\_group - Age group of participant(s) at the time crime
* participant\_gender - Gender of participant(s)
* participant\_name - Name of participant(s) involved in crime
* participant\_relationship - Relationship of participant to other participant(s)
* participant\_status - Extent of harm done to the participant
* participant\_type - type of participant
* sources - Participants source
* state\_house\_district - Voting house district
* state\_senate\_district - Territorial district from which a senator to a state legislature is elected.

# Data preparing

Following attributes has been removed (because of given reasons):

* incident\_id – unnecessary for mining
* address – unnecessary given geo coordinates
* incident\_url – unnecessary for mining
* source\_url - unnecessary for mining
* incident\_url\_fields\_missing – unnecessary given removal of two previous attributes
* congressional\_district – unnecessary given state
* gun\_stolen – huge lack of data
* incident\_characteristics – too many different factors, lack of data
* location\_description – unnecessary for mining
* notes – too many different factors
* participant\_age – unnecessary given age group, too specific
* participant\_name – unnecessary for mining
* sources – lack of constancy and recurrence
* state\_house\_district - unnecessary given state
* state\_senate\_district - unnecessary given state

Also, following changes have been made:

* removed data from before year 2014, due to lack of information compared to other years, and compared to whole dataset the number of removed rows is negligible

# Data visualization

|  |
| --- |
|  |
| Figure 1) Number of gun violence occurrences to date with regression line |

|  |
| --- |
|  |
| Figure 2) Number of gun violence occurrences for each US state |

|  |
| --- |
|  |
| Figure 3) Number of killed in gun violence occurrences to number of injured in same event |