How to avoid dodging a bullet: US-gun data’s Abstract

Olivia (Olive) Kirk [i6 153 141] & Karim Abdel Kader [i6 131 210].

Questions

1. Which states, counties and congressional districts have the highest rates of shootings, both fatal and non-fatal?
2. What is the relationship between the number injured and/or killed and the "incidental_characteristics" (correlations between multiple factors)?
3. What trends change over time? (particularly the frequency of characteristics and the severity of shootings)

Approaches

The original dataset contains all sorts of incidents related to guns, including those lacking explicit violence such as threatened violence during a burglary.

Many features of the original dataset were deemed useless such as the “notes” column, and thus were discarded to help improve the computational efficiency of such a large dataset.

Approach 1

We utilised most of the original dataset for this, due to it having the required features in it to produce heat maps.

After aggregating the data into states, new features were added such as “entries” for said state as well as the proportion of entries that involve a given variable e.g. injuries.

Approach 2

Due to numerous missing or “NaN” values, a lot of rows were dropped which affects the integrity of our results. The dataset was more than halved in size, the only solace being that it is enormous, and as long as the stratification remained, then there as no issue (which while not something we can 100% prove, we suspect is not the case).

A lot of extrapolating had to be done from the initial strings of features from the original dataset. The most troublesome was incident characteristics as they had to be placed into 150 additional columns, making it very computationally heavy, even for the dropped dataset. The additional columns of extrapolated features were all ‘binary column’ that held either a 0 or a 1 value, as opposed to boolean values to ensure that correlations could be properly calculated later.

Again, due to the large amount of features, visualising the correlation matrix was not practical, and more ugly lists were conjured up in their place.

After finding values that shared high (absolute) correlations, how much each value coincided with the other was inspected by taking their conditional probabilities (the primarily used measure).

Many characteristical relations were explored as viewable in the notebook, but many had to be discarded due to not being able to draw anything original from them e.g. men commit suicide more than women and more violently so, or due to the data being inherently limiting e.g. the proportion of tragic children’s cases was absurdly high, since only incidents that had bad endings involving those under 11 would be reported.

Approach 3

The third approach bootstraps off of the previous two, combining both aggregation based on discrete values and the largely processed dataset.

While mild time series exploration was performed on the raw dataset e.g. it was found to be non-stationary which was remedied by a differencing of order 1, it was dropped in favour of general EDA, especially given the unusual trend (which may largely be explained by the ability to obtain reports changing over the years due to legislation, but not much could be explicitly found).

Limitations

Each and every data point is a verifiable incident via the links provided in the original data frame. However, that means that the dataset is inherently an incomplete view of reality, with the website authors stating something to the effect of “if it wasn’t important enough to report to the police, then the incident wasn’t very severe and doesn’t belong”, a while detestably lazy sentiment also a very understandable one from a practical perspective.

It should be noted that since all incidents are verified, the time of occurrence is somewhat irrelevant from a collection standpoint, as a verifiable instance is a verifiable instance.