A Tool for Visualizing Gun Violence Data

By Mike Browne, Nina Illeva, Lexi Selldorff, Fabian Wikström
I. Introduction/Problem Statement

Guns were the cause of 12,942 deaths in America in 2015. Most Americans underestimate the number of gun-related incidents in their country. In the last year, guns have killed more people than drug and alcohol overdoses, Parkinson’s Disease, and war. We built a website that allows users to look at specific incidents or graphs of gun violence, helping them grasp the breadth of America’s gun addiction. Our solution uses data from news articles because the US government has no comprehensive record of gun violence incidents. These articles include details that are not covered in most analyses of gun violence. Our interactive website allows users to create and manipulate their own graphs enabling them to unlock insights that were previously unreachable.

II. Approach

Data

The data we used was collected by Professor Elie Pavlick, Chris Callison-Burch, and their team of researchers at the University of Pennsylvania. Google search results for the phrases “man shot” and “woman shot” were scraped and uploaded to Amazon’s Mechanical Turk for annotation. Mechanical Turk workers were then tasked with (1) determining if the article related to gun violence and (2) if so, the parts of article that contained information about the victim, shooter, and shooting.

Due to the nature of human annotation, our data had to be standardized to predetermined values in order to be analyzed. We wrote python scripts to edit the data and add new fields to it. First, we went through the data to look for different values that carry the same meaning. As an example, possible entries for the gender field included: “f”, “female”, “woman”, “w”, “m”, “male”, “man”, null, unknown, none, other, na, n/a or an empty string. The script set the values to one of three options: “f”, “m” or “null”. We also edited fields like the date field to give them a uniform format. Next, we added fields to help with our analysis. Giving every shooting a latitude and longitude value is an example of some of the fields we added to the dataset. To do this we used the Google Geocoding API to get the needed information based off of the already-entered location information.

Lastly, we had to scan the data for duplicate entries and multiple annotations of the same article. We first sorted the data based on the URL they were annotating. For each field we looked at all possible values (excluding nulls) from the doubly-tagged articles. We then created a new JSON element containing the value that was cited by the most entries. If the values were encountered the same amount of time, we assigned a random value to the new element. Lastly, we replaced the duplicate entries with the new JSON element.
**Web Application**

We decided to build a web application so that our project could be accessible to many users. Since our project focused on visualizations, we used D3.js and Plottable.js for the graphs, maps and calendars. These tools have a wide range of graphs and charts that you can customize, which made these technologies a good fit for our project. We only utilized the bar graph, line graph, pie chart, map and calendar visualizations, but there are many more possible visualizations should we wish to expand on our current offerings. The bar graphs are used to show histograms of the data, while the pie chart is used to categorize the data. Our line charts are primarily to help show gun violence statistics over time. Part of the appeal of the graphing libraries we used was that they can easily highlight data points on the graph. For example, we realized that when graphing many categories on the x-axis, it was hard to tell which bar went with which label. To mitigate this, we used Plottable tools to highlight a bar as the cursor passed over it and show the label on the bottom, clarifying which category the bar represented. In the case of the line chart, we used Plottable tools to zoom in or out and select certain time frames while looking at the data.

To build the web application itself, we used a MEAN (MongoDB, Express, Angular and Node) stack. In the beginning, we set up our project to use a Mongo database but we decided later that it was not necessary. We did not have enough data to warrant a database, as keeping it in a JSON file and parsing the file was plenty fast enough. However, if we get more data (on the level of 6 figures) in the future, we have built some of the functionality to store the data in the database. Express is used in conjunction with Node to host the files and route the urls. We used Angular to integrate the front end and our data enabling us to use AJAX to update the graphs instead of having separate pages. We chose Angular because it makes it easier to manipulate and display our data. For the front end design, we used Bootstrap because it’s simple responsive. Since Bootstrap made the application look too plain, we used a custom theme on top of it. We decided to create a dashboard that would give our users a quick overview of our gun violence dataset. The dashboard shows how many people have died and how many people have gotten injured from gun violence in the US in 2016. The dashboard also has two bar graphs, one that displays the race of the shooters and one that displays the number of shots.

In order to give our users an idea of where the shootings are happening we used D3 to build a map visualization that shows the location of each shooting. The data-parser simply goes through the JSON file and extracts the longitude and latitude of each shooting. Each shooting is then represented as a dot on a map of the United States. The map allows users to zoom in on each state to get a more detailed view.

We decided to use Plottable’s calendar heatmap to display when shootings usually occur. A calendar heatmap is best used when one wants to illustrate how some quantity varies depending on the day of the week, or how it trends over time. We wanted to give our user the
opportunity to see when most shootings occur. The calendar heatmap shows a matrix with the
dates on the x-axis and the day of the week on the y-axis. Each entry in the matrix is then
shaded depending on the number of shootings that happened on that day. The darker the
shade, the greater the number of shootings. The heatmap was fed data from our data-parser
which extracted the day of each shooting from our JSON file and summed the number of
shootings for each date. Our users could use our filtering system to customize their own
calendar heatmap visualizations.

In addition to showing premade visualizations, we wanted the user to be able to
manipulate the graphs to explore the data more thoroughly. To implement this, we added
filters so that the user could filter the data shown on the graph based on various victim, shooter
and shooting details. This allows the user to easily hone in on parts of the data that he/she find
interesting. To take this one step further, we added a “custom graph” tool where the user can
choose what data they want to display on their graph. They can use this in conjunction with the
filters to create novel graphs that will provide them with new insights. The filters are especially
useful when the user is interested in reading articles on specific types of incidents. We added a
list page that displays each incident, victim, or shooter (and their respective details) along with
a link to the article that the data came from. The ideal workflow for the user would then be to
explore the data on the graphs, maps or calendars then, once having narrowed down the data
to an interesting subset, read the relevant articles to draw further conclusions. This workflow is
what we kept in mind as we built our project.

III. Results

Visualizations

Our final application had bar graphs, line graphs, pie charts, maps, calendars and lists.
We felt that the visualizations that we made were sufficient for a basic, but detail-oriented,
analysis of the dataset which was our goal. We tried to balance usability with complexity of the
possible analysis and felt like the final product had maximum functionality possible with an
interface that was still easy to navigate.

The bar graph visualization showed the data clearly for most subsets with many
categories while the pie chart was useful for those with few categories. We did not restrict the
subsets that could be used for each visualization or graph but instead relied on the user to
intuit which graphs to use for each subset once they tried them. Since we only offered a few
visualization options, we feel confident that the user should be able to easily find one that suits
each subset. The line graphs and calendars were both focused on visualizing the data over time
however, the calendar showed the data on a more granular level (per day) than the line graph.
The list was useful for the most detailed level of analysis while the map gave a good overview of
incidents per location. We felt that we covered most of the important basics that could lead to
a successful analysis. As our data was incomplete, our tool was designed to start a discussion
that would warrant further research which we felt it accomplished. The calendar heatmap design does a very good job of letting users see when most shootings happen. However, since we had a limited dataset drawing conclusions from the visualizations we provide was difficult. As more data is added to the dataset the heatmap will provide a better overview of when shootings happen in the US. The map visualization provided our users with a great overview of the location of gun violence in America. The map visualization could have worked even better if our original dataset contained coordinate information (as opposed to us getting coordinates from the city names). This would have meant that we could give our users more precision about the location of the shootings.

**Further Data**

In addition to visualizations that we overlooked, there are also plenty of pieces of information that would have helped make our analyses more interesting. The ability to normalize our graphs as a function of the number of victims (i.e. two shootings with the same number of shots but a different number of victims have different shot/victim ratios.) and population density. One of our judges rightfully asked us if the concentration of gun violence near urban areas was a function of those areas just having more people or if it was actually representative of a level of violence that we do not see in other, less populated, parts of the country. Another piece of information that some of our judges asked about was the variability in gun laws across states. It would have been interesting to overlay the severity/stringency of each state’s gun codes on top of that state’s gun violence incidents.

**IV. Ethical/Privacy Considerations**

**Limited Information**

One concern we had was drawing conclusions based on a limited dataset. We sourced our data from news articles around the web, which is inherently biased. Media coverage tends to be more extensive in urban areas. Articles also tend to cover more “interesting” stories rather than all incidents equally. Furthermore, crimes in areas under the poverty line tend to be covered less in the news. We believe our final product should be a tool for analyzing data but the end user should be aware of the bias built into our dataset. Without paying heed to this, users could draw harmful conclusions from our visualizations, especially our list functionality. In fact, even if our dataset were complete, we would be worried that the use of identifying details could be used to stereotype the person most likely to commit gun violence, without considering the differing circumstances of each incident. We would like our tool to encourage users to explore trends they find in the data and not make judgments based on individual incidents.

This issue led us to discuss whether or not we should create our own analysis of the dataset as we know the dataset is not complete. We decided to draw our own conclusions from the dataset as it is now with the disclaimer that this is incomplete data and does not necessarily
represent the whole story behind gun violence in the US. We wanted to analyze the dataset as it is both interesting and novel, however we did not want anyone to take our conclusions as fact. We decided to present the project as a tool for analyzing gun violence data first and foremost but with some brief insights of the dataset as is.

Specific Names & Details

Another privacy concern for us was revealing the specific names, personal information and location of those involved in the shootings. We would not want either the victims’ or shooters’ personal information to be misused if revealed. However, since all of our information came from news articles, the relevant information was already public so we felt confident it was ok to aggregate and display it in one place. If we were to integrate a dataset that had information which was not already public, we would need to readdress this issue and make sure that revealing it is allowed and ethical.

V. Discussion

Dataset - Lack of Centralized Database

Our first major challenge was getting exhaustive data on gun violence incidents. Surprisingly, despite the recent political unrest on gun regulations in the US, there is no centralized dataset on the topic. This is because the CDC was defunded by congress (for gun violence research) in 1996. We then explored the possibility of using the FBI’s homicide reports which are available to the public. However, we found that they were not exhaustive and had some rather serious limitations. We found (1) that the reports often contain a lot of errors and (2) reports on police shootings were often not included. Other databases we considered using were aggregated using crowdsourcing but would have needed to have their information verified or didn’t contain the level of detail we were looking for. As a result we decided to use information available in online articles which were parsed by Mechanical Turk workers.

Dataset - Human Error

Another challenge we encountered was the human error in the data we received. When we would have duplicate entries (articles that talked about the same shooting, etc.), these entries often contained conflicting information. For example, one entry would say that the victim was killed while another would say they were merely hospitalized. In the cases where we had many overlapping entries describing the same incident, we would aggregate the details that appeared in a majority of the entries into one entry. In the few cases where there was great parity in the details submitted, we looked at the article and picked the value ourselves. Another contributing factor to the potential inaccuracy of the data was malformed data entry. Some of the entries had typos that often made the information hard to parse. We did our best to ferret out these typos but those that weren’t found created detail gaps in our data.
These data issues mean that it isn’t possible to get a thorough and accurate look at gun violence trends in the US. We have a lot of missing data and details which could lead to inaccurate conclusions. Since the dataset is also drawn from online media articles, there is a drawback that the incidents that get the most media coverage are more likely to be in our database creating a bias towards incidents with a lot of media coverage. Another interesting point of contention is whether or not articles are more likely to include details about a particular incident if they fit the nation’s stereotype of gun violence. Simply put, articles might be more likely to include racial and gender details if those details happen to be black and male respectively. Regardless, we have a novel dataset that is interesting to interpret as is. It still represents a portion of gun violence incidents covered by the media in the US so it can potentially be used to make a hypothesis that can be confirmed with more data.

Selection of Visualizations

Switching to the technical side, we felt we made good use of D3.js and Plottable.js to construct informative graphs. However, we could have done more complicated and fun visualizations. We felt the ones we implemented were the most appropriate but if we had more time, we would expand our tool to include more fun, but perhaps less practical, visualizations. There are also more interactions and animations that we could have added to the graphs. Although these may have added more unnecessary noise, some may have been useful. We added zoom, labels, data selection, and legends to the graphs themselves as well as the filter function to be able to create subsets of the data. These interactions allow the user to view and manipulate the data clearly. However, we would have liked to show more details when hovering over certain parts of the graph or have an option to highlight certain parts of the graph by clicking. These proved to be more complicated than we originally thought and were thus not made a priority. It would have been nice to allow the user to alter the appearance of the visualizations or to save them to show other people but these features were also too time-consuming to implement.

Filters

The filters we built could be used to both match on certain details and hide certain details. This allowed the user to easily show only the part of the data that they were interested in. For example, they could show only incidents with male victims and all age groups except those over 60. These filters were for most victim, shooter and shooting details. We originally wanted to implement date and location filters but ran into formatting and display issues. Instead we added the line graph and calendar to display date oriented data and the map to display location oriented data. If we had more time, we would have liked to add these filters to the rest of the graphs to add another dimension to them.
Another drawback of the filters we built are that they are not logically checked on the back end. That is, it is up to the user to set them in a logical way. For example, it is possible to both show only male victims and not show any male victims on the same graph. This obviously produces a blank graph but may not be intuitive if the user doesn’t realize that the filters are set this way. If we had more time, we would use the bootstrap disable tag for form data to disable the user from making any filter combinations that wouldn’t produce any data.

Data Parsing

Although parsing the JSON file is relatively quick, it takes a while to do so and then render the list visualization. This makes for a 1-2 second delay to display the list or to re-render the graphs. With more time, some of the filtering would be done in advance so that it wouldn’t be so slow to re-render the list and graphs. With the amount of data we have, it is not a huge problem but if we had more data, this would be a significant bottleneck.

Additional Data Categories

We decided to present the raw data as-is without normalizing it or doing any calculations. This is because we felt like we had an incomplete dataset so it didn’t make sense to do any calculations as if it was complete or conclusive. This tool was built to explore specific instances with certain details or observe potential trends than find accurate final conclusions. However, if we did have a complete and accurate dataset, we would have added some visualizations with interesting calculations. For example, normalizing the graph of number of incidents by age of the victim. As there is not an even distribution of ages in the US, the total number of incidents could be normalized by the amount of people in that age group to make a more accurate representation of the typical number of incidents per age group. Another thing we could have done was account for population density in each area where the incidents occurred. That way the number of incidents per capita would be more obvious and accurate.