USA Gun Violence 2013-2018

by

Bailey Wang

Abstract

Gun violence in the United States of America has become more common. Information regarding the victim and perpetrator, number of casualties or murders, types of firearms and the amount can all be provided to the user for each individual incident. This report will help provide the reader with information regarding the statistics of gun violence.

This topic is important as the rise in violence has started occurring again, and government officials will want to restrict the second amendment (right to bear arms). This report will help provide statistics to support whether such restrictions will be beneficial or harmful in stopping gun violence.

At the end of the project, an interactive dashboard will be available to present in-depth analytics and drill down methods to display the information. The visualization will follow proper use of color, data-ink ratio, and whitespace usage.

Introduction

Every day, many Americans are directly or indirectly involved in firearm incidents. The government has issued a bill to restrict and regulate the usage of assault weapons. However, the restriction of assault weapons may not be enough to keep Americans safe. The government's plan to restrict the access of automatic weapons may not necessarily provide the most protection against firearms. Thus, this report will analyze the growing number of firearm incidents and deduce whether the government should mandate stricter firearm policies.

This project's objective is to familiarize the reader with the implications of the severity of firearm incidents. It will also provide the reader with information regarding which locations contain the highest prevalence of firearm incidents. Within each of the firearm incidents, information regarding the relationship between the perpetrator and victim(s), the type of firearm used, or how many firearms were used will be recorded for each incident.

This report attempts to analyze the trends of firearm incidents and to provide statistics of the firearm usages.

Intended Audience

The intended audience are those who are interested in the growing cases of firearm violence as well as people who are interested in learning about firearms and how effective government policies would be when propagated.

This information will also be useful toward government officials. Government officials can use this information to create a more robust bill regarding access or restriction of firearms. It can also help government officials understand whether their policies will help create a safer environment for the citizens.

Project Deliverables

- Presentation:
https://docs.google.com/presentation/d/1jvhk0dgTLGiB8pFyVypB55c3NT1pie_rR4OzXrzi8XU/edit?usp=sharing
- GitHub:
https://github.com/misterbwang/data230_project
https://github.com/sjsu-data230/baileywang
- Report
- Tableau Dashboard
https://public.tableau.com/views/final_16393592742370/Story1?:language=en-US&publish=yes&:display_count=n&:origin=viz_share_link

Data Source

The dataset is provided on Kaggle.com which was created by a user James Ko. This dataset was created by using HTML web scraping on the website gunviolencearchive.org. James Ko created a web scape script that was run in 2018 which queried all the information at the time and compiled it all into a CSV file (Ko, 2018).

Gunviolencearchive is a non-profit organization with the goal of providing the most reliable information and data on crimes. The organization uses queries on local and state police, media, and government resources to create this free database. The database also catalogs detailed information regarding location of the incident, specifies the victim and perpetrator, defines the incident characteristics, and where the incident is sourced from. All incidents will be recorded from 2013 onwards (Gun Violence Archive).

Data Cleaning

The data considers a lot of information regarding the incidents. Majority of the data cleaning is done in Python using different data cleaning tools such as dropping unnecessary columns, or applying lambda functions and using regex to filter the data out.

A common occurrence for the data to appear as is with special characters separating the participants from each other. For example, “participant_age” column contains the datapoint as “0::25||1::31||2::33||3::34||4::33”. This shows the number of participants within each case. In this case since there are five values, it can be concluded that there are five participants in the overall incident. This also provides a meaningful challenge to separate the redundant information from the data. Applying regex as a method to clean the data can easily remove all the unnecessary information; however, it still leaves the data as a list. The results would look something like

“[25,31,33,34,33]“. This is much better compared to the raw input of the information, but it still contains issues in creating usable graphs to convey the information.

For certain columns such as “participant_gender” or “participant_status”, there are a few unique values for each of those columns. For “participant_gender”, there are only two values: male or female. These values make separating the two into two distinct categories easier; the ability to separate the two values into two separate columns with their counts is one way to interpret this situation. By applying lambda functions, the count function can be partitioned through the whole dataset and account for the number of male and female people involved in the incident. A similar methodology can be applied to “participant_status” and create multiple columns containing the count of each unique value.

Some basic information is provided within “incident_id”, “date”, “state”, “city_or_county”, “address”, “n_killed”, and “n_injured”. The “incident_id” provides the website’s unique identifier for the incident. Some of the common information is the date which provides when the incident occurred. The state, county, and addresses are all provided which would make creating a geographic map easier as the longitude and latitude can be generated from that information. Finally, the number of killed and injured is used to calculate the total number of victims by adding the two together.

The Participant Relationship between the subject and the victim mostly falls under “armed robbery”, “family”, or “significant other”. This means that the most offending cases are done at local businesses.

The next two groups have a domestic relationship, since the “family” and “significant other” groups are roughly similar. For the “significant other” category, the majority of incidents involved are adults whereas, the “family” category contains both adults and children in the incident.

Interestingly, despite constant media coverage, both “gang involvement” and “mass shooting” are considered extremely low numbers of incidents.

The dataset also provides additional information regarding the participants including the names; however, as this information is more personal and does not provide any usefulness, this column will be ignored.

The participant gender is similar to the structure in “participant_age”, where the gender is provided within a nested list. As mentioned earlier, this information will be required to be cleaned in order to provide meaningful information.

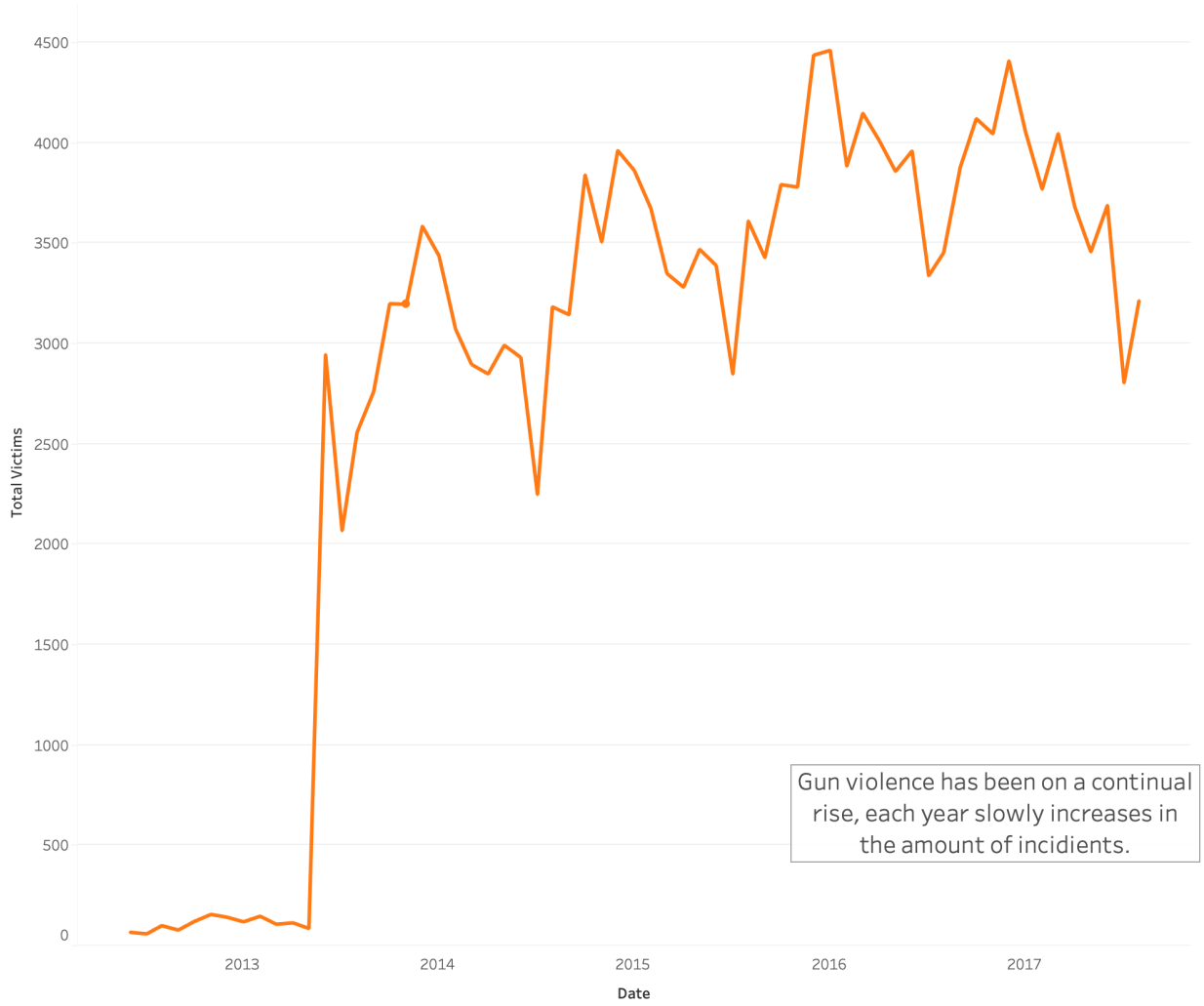
The participant age is provided as well as in a separate column which groups them into ordinal values, separating the age groups into “children”, “teen”, and “adult”. This provides a generalization of the age distributions within each incident; however, this would lose information in the overall distribution of the specific ages.

The status of each firearm is very skewed to the “unknown” category. Some speculations as to why the “unknown” category contains the largest value is that the firearm would be more difficult to trace back to the owner. Another speculation is that the police report was unable to identify the firearm and whether its origin was stolen or not and classified it as “unknown”. Another interesting point is that there are a small number of registered firearms. It could be assumed that for some of those owned firearms, the incidents could have been suicide.

Some columns that only provide backend which are “incident_url”, “source_url”, and “incident_url_fields_missing”. These three columns provide information on the news website that reported on the incident. As these do not contain any meaningful extractions, these columns will be removed.

Analysis

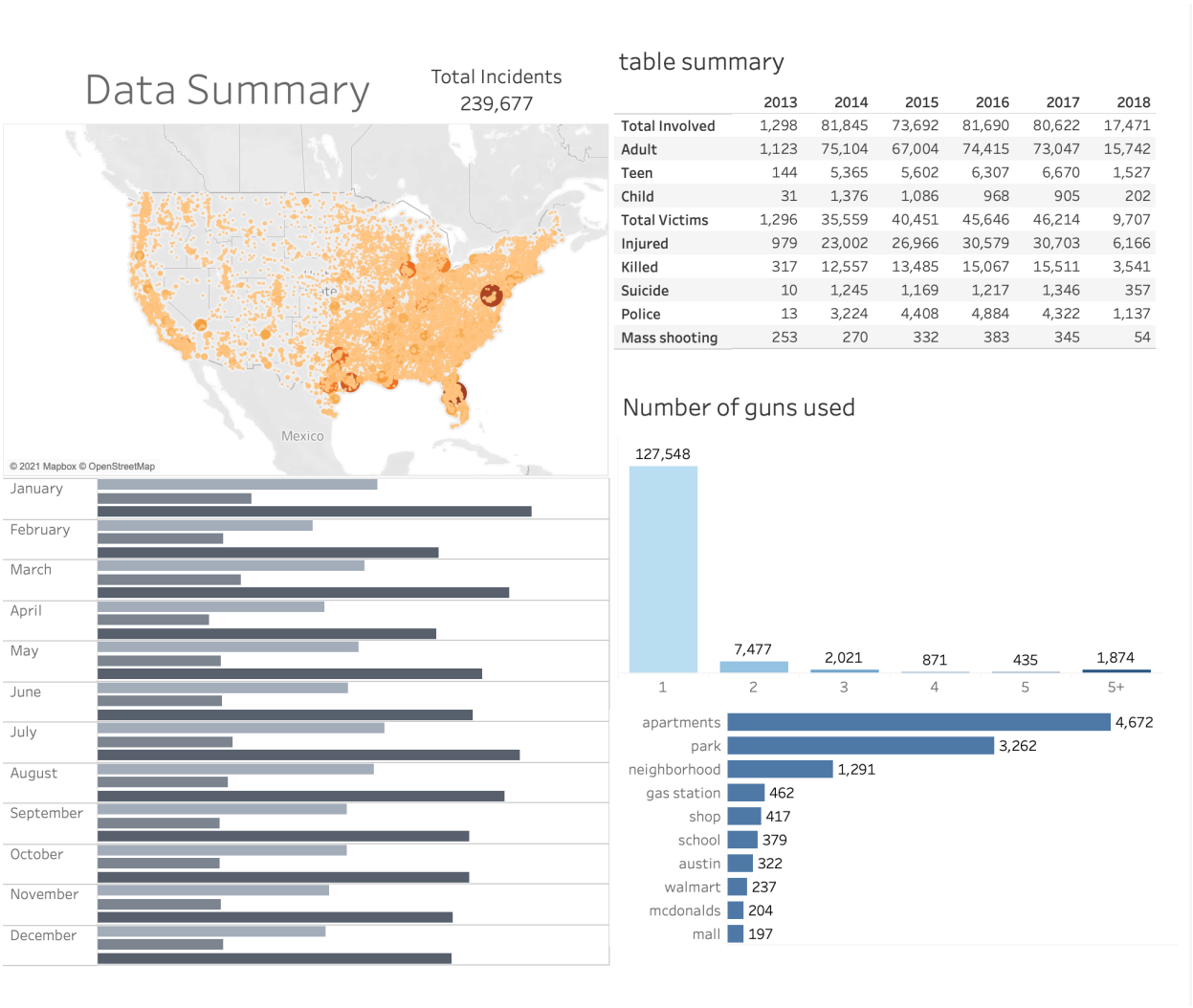
The introduction slide is very simple and contains the information on which the topic will be discussed for the presentation. Since there are no visualizations on this page, it will mainly be focused on introducing the topic and explaining the importance.



This slide shows the yearly trend of gun violence; based on the information (and where the data has come from), the amount of violence per year has a steady increase. The color red will be used to denote the violence which is represented in the incident.

Based on the year, it appears that 2014 had a numerous number of firearm incidents and 2015 had a slight decrease. Then the number of incidents increased again as the years continued. This trend seems to reoccur often.

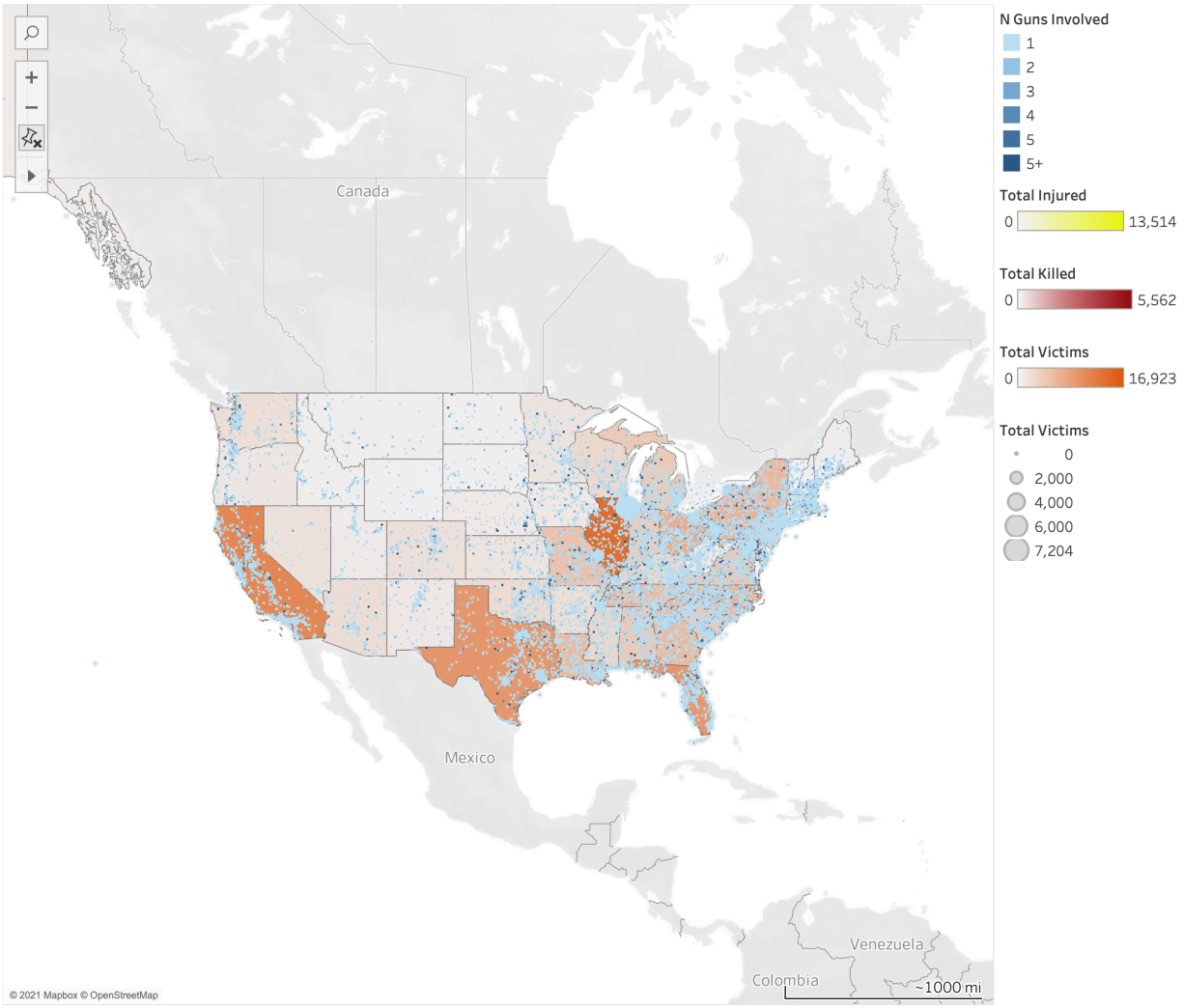
Here is the data cleaning aspect of the project. The main data cleaning aspect was to use regex, since the majority of the data is text based. Here are two examples of the before and after where it represents the text data. As the data shows, there are a variety of special characters that need to be filtered out of the text in order to use the data more efficiently. After cleaning the special characters out of the data, the data will remain in a list format. In order to extract the information, the data will be counted and split into new columns to denote the number of counts. Although this will make the dataframe larger, it will be useful when trying to apply actions on the dashboard for later.



The data summary after all the data cleaning has been completed. The map shows information when hovering over the point and includes state, address, city or county, and amount of victims in the incident. This will greatly help the viewer understand the frequency of incidents as well as the severity based on the color. The "Number of guns used" graph provides information on the amount of firearms used per case, here the most amount of firearms used would be one. Right below the "number of guns used" graph, the location graph provides the commonly occurring location for firearm violence to occur. This will be useful information later, as shown in the relationship between suspect and victims, but it is interesting to note that the highest number of incidents occur in apartments (residential locations). The frequency of firearm incidents per month shows the viewer which months have the highest number of incidents.

It is interesting to note when comparing the firearm incidents between months, that January contains the most incidents. This is followed by July, March, and August. Regarding January and July, these are the time periods during holiday. March is when the weather starts to

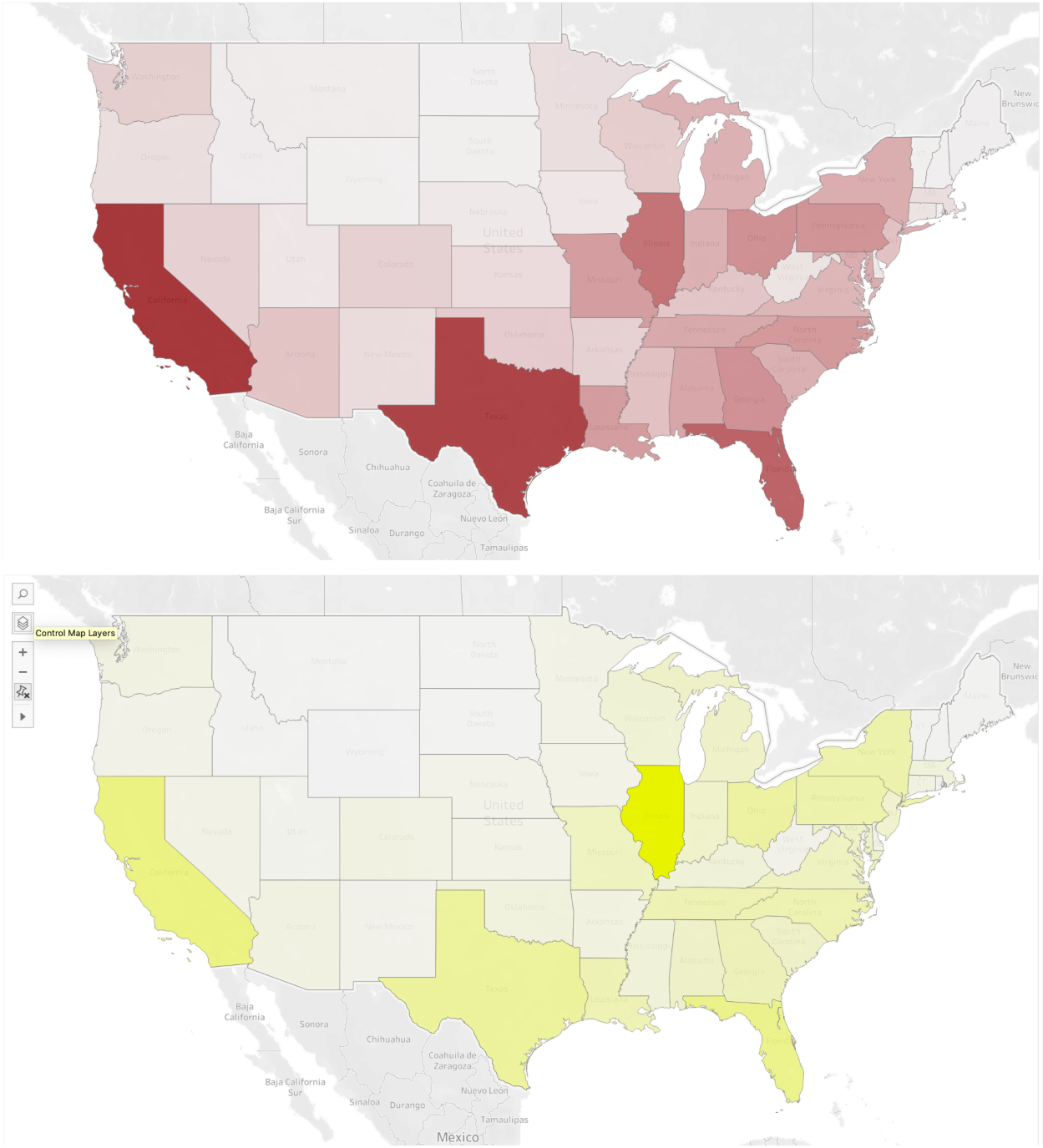
warm up, and August is during the period where schools are opening up again. However, these observations are not a correlation for why there is a rise in firearm cases during the months.

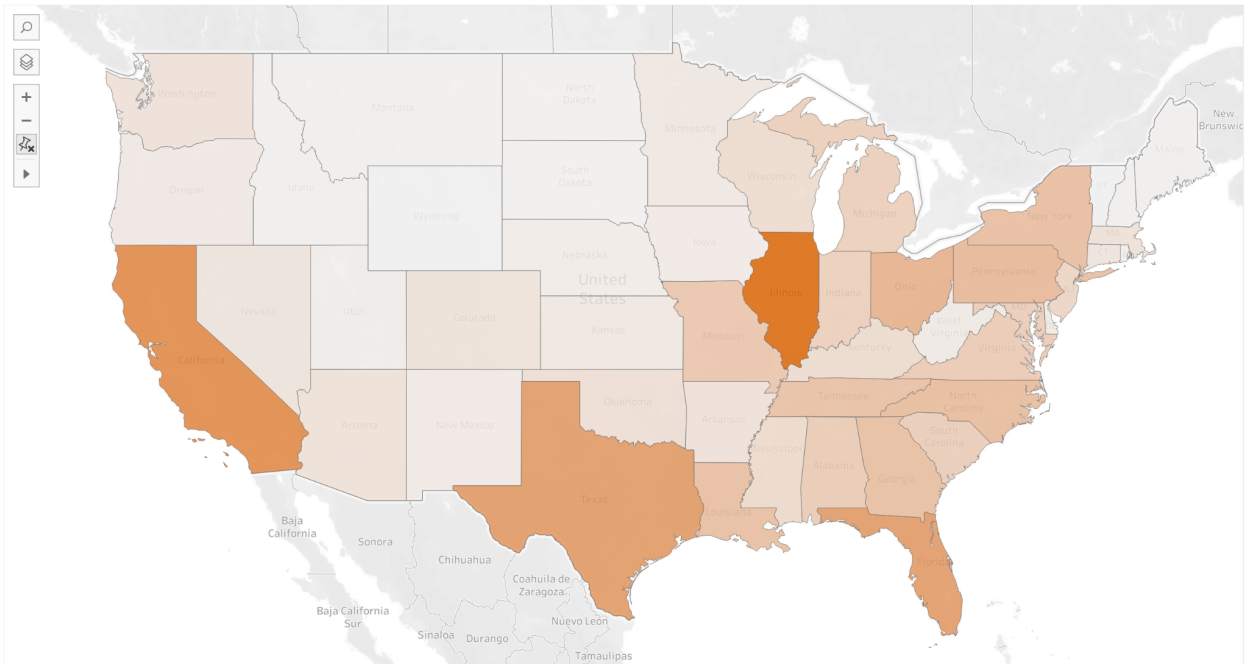


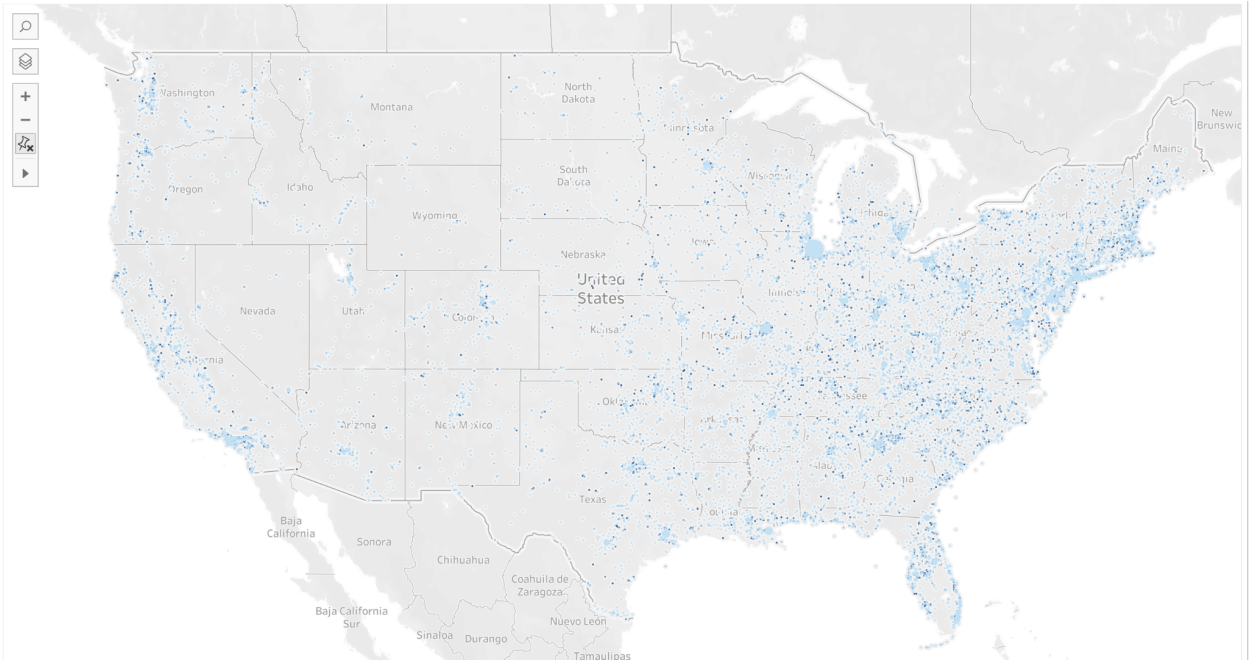
The first visualization to represent the whole picture of what's happening in the US. Here the map is a data layer that shows different information. The “Orange” color is used to describe the “Total Victims” per incident; this color is a combination of “Yellow” which represents people injured and “Red” which represents people killed. The blue represents the amount of firearms in each case and the size of the dot is the total number of victims per case. This map is very integral to the visualization as there is also the ability to toggle the states to drill down other information in the dashboard.

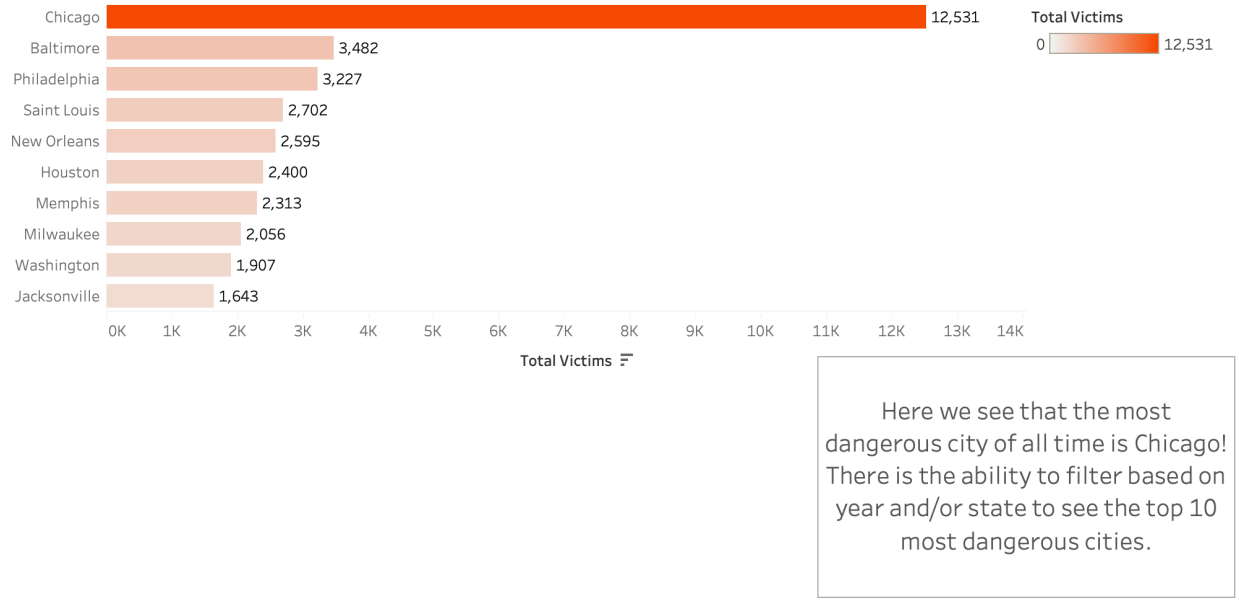
Here are examples of the drill down of the data layer for this geomap. Each layer provides information which is not found in any of the other maps. As previously mentioned, “Orange” describes the “Total Victims”, “Yellow” describes the number of people injured, “Red”

describes the number of people killed, and “Blue” represents the amount of firearms in each incident.

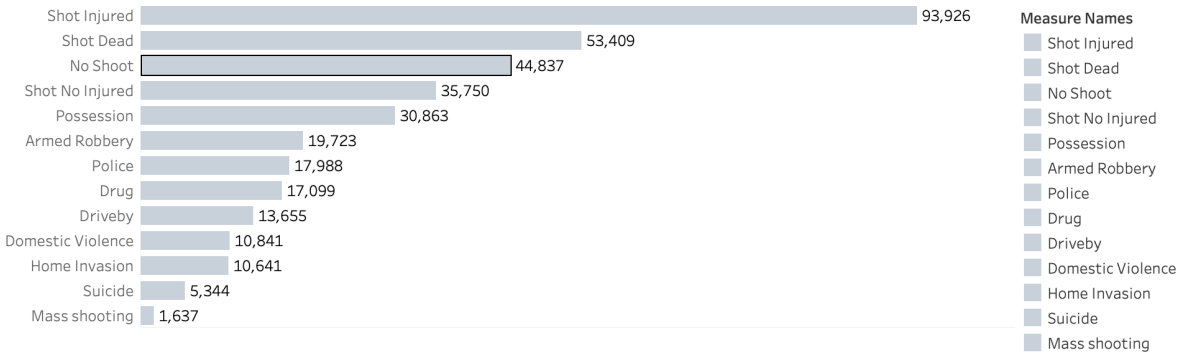



Washington
Montana
North Dakota
Minnesota
Oregon
Idaho
South Dakota
Wyoming
Nebraska
Iowa
Wisconsin
Michigan
New York
Pennsylvania
Nevada
Utah
Colorado
Kansas
Missouri
Illinois
Indiana
Ohio
California
Arizona
New Mexico
Oklahoma
Arkansas
Tennessee
Kentucky
Mississippi
Alabama
Georgia
Texas
Louisiana
Florida
Baja California
Sonora
Chihuahua
Coahuila de Zaragoza
Nuevo Leon
Baja California Sur
Sinaloa
Durango
Tamaulipas

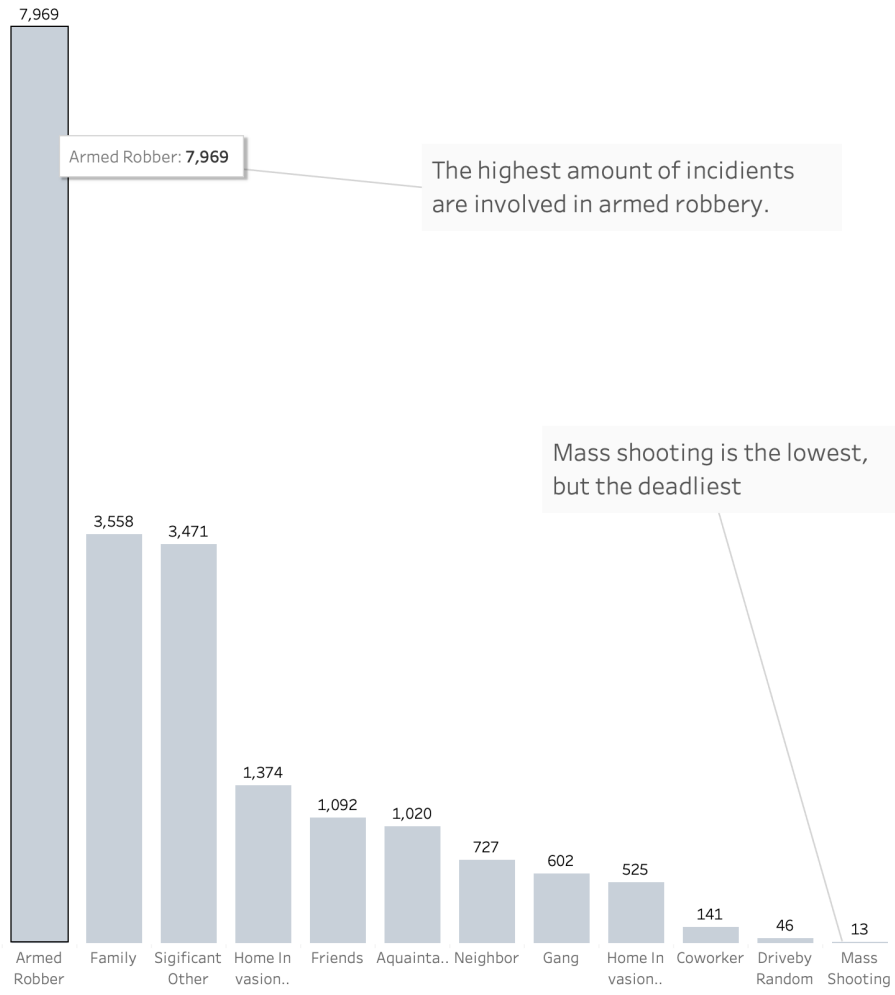

Washington
Montana
North Dakota
Oregon
Idaho
South Dakota
Wyoming
Nebraska
Nevada
Utah
Arizona
New Mexico
Texas
Baja California
Sonora
Chihuahua
Coahuila de Zaragoza
Nuevo Leon
Baja California Sur
Sinaloa
Durango
Tamaulipas



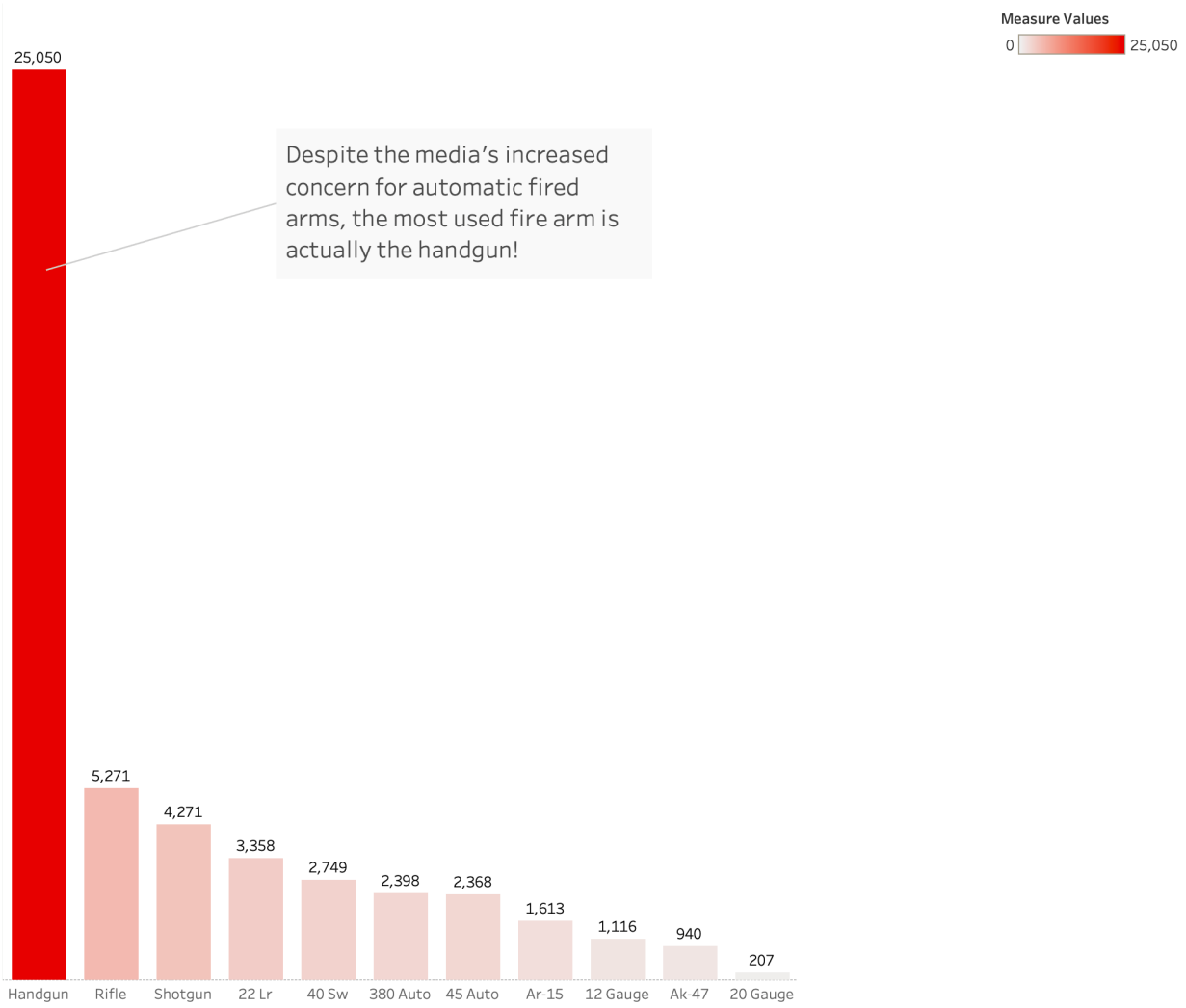
The barchart shows how many victims per city or county. This graphic complements the map visualization as selecting a state will drill down to only cities within that state as well. This is extremely meaningful in understanding areas or locations that contain the highest number of violent firearm incidents.



Here shows the types of incidents committed. The majority of the incidents can occur within a single incident. This basic information still provides the viewer information regarding the likely situations that occur within each incident.

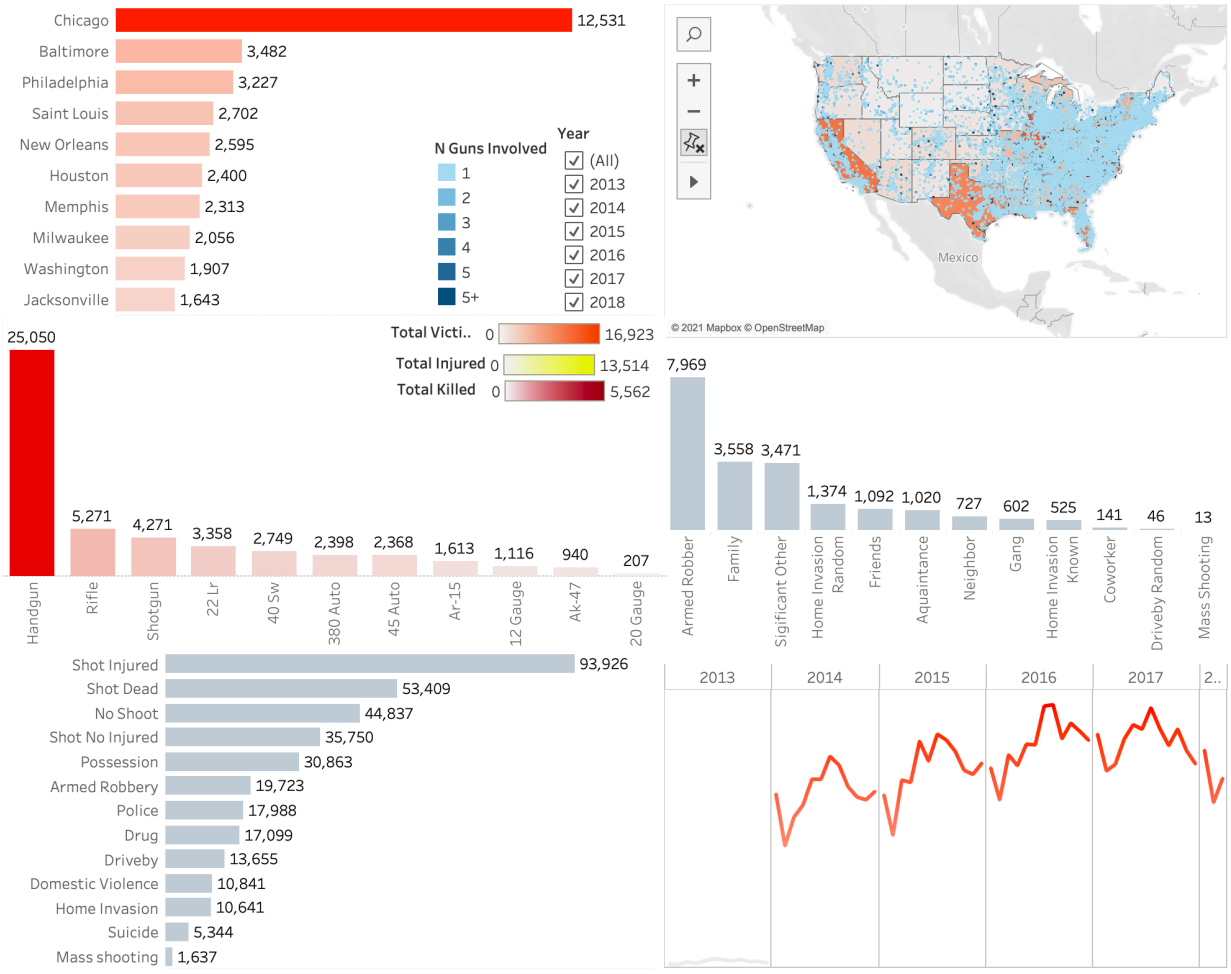


Surprisingly, the highest number of types of relationship between suspect and victim would be armed robbery. Given the information from the data summary, the viewer would see a high number of cases occurring at franchises or apartments; it could be considered that robbery and family members would be high in this case. Here it is also mentioned that mass shootings are the lowest number for type of relationships, despite being the most talked about within the news.

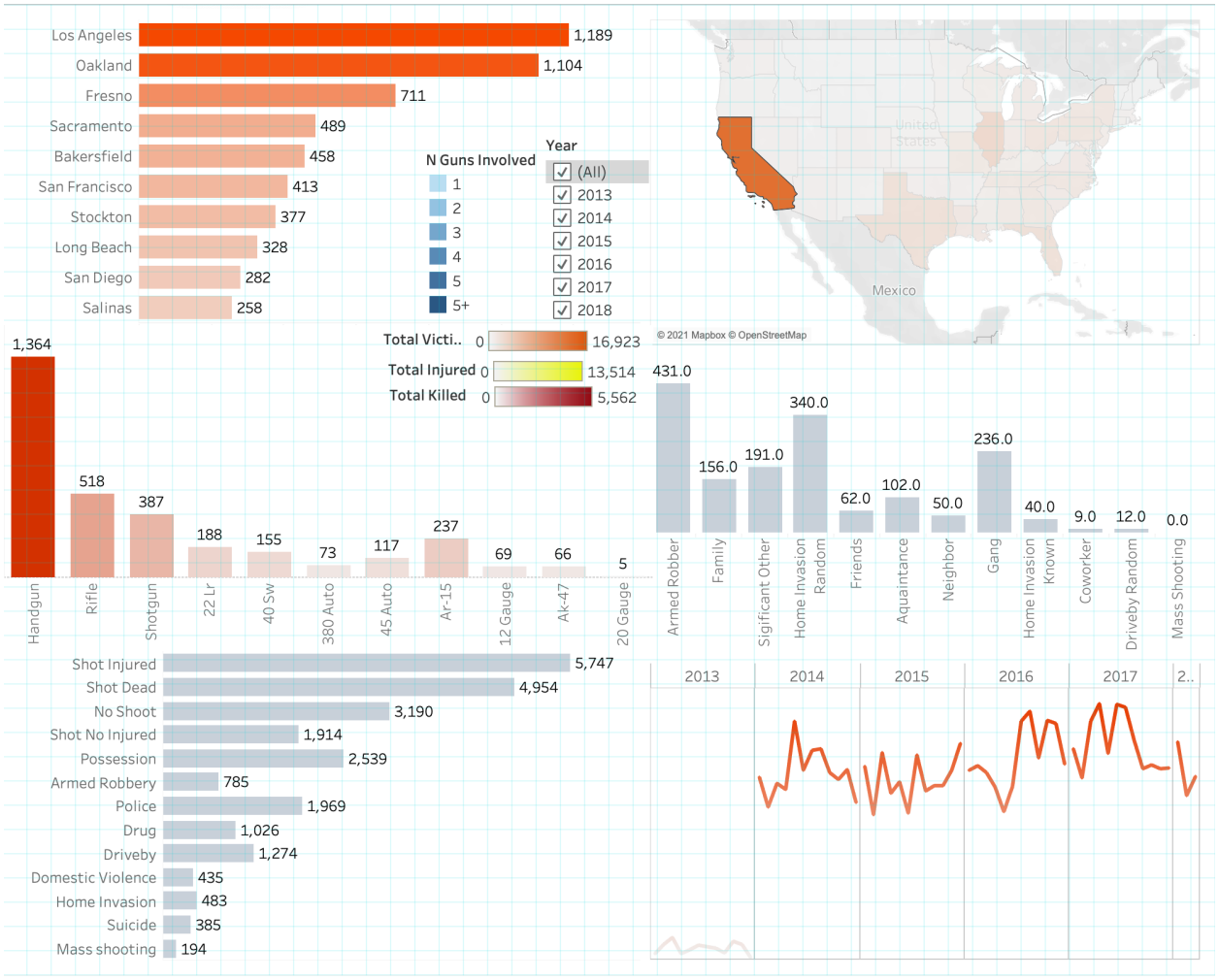


Here is the final bar chart, which uses the gun usage statistics. Returning back to an earlier point regarding the media's coverage for firearm ban, the information provided shows that the most commonly used firearm is actually a handgun rather than an automatic firearm. Majority of news networks would report that semi-automatic or automatic firearms should have stricter policies in place. However, based on the information provided within the data, the only automatic firearms are the "45 auto" and the "AK-47". Within the two classified as automatic firearms, the "45 auto" also falls under "handgun". The handgun accounts for around 25,000 cases, while the "45 auto" and "AK-47" only account for 2,300 and 900 respectively.

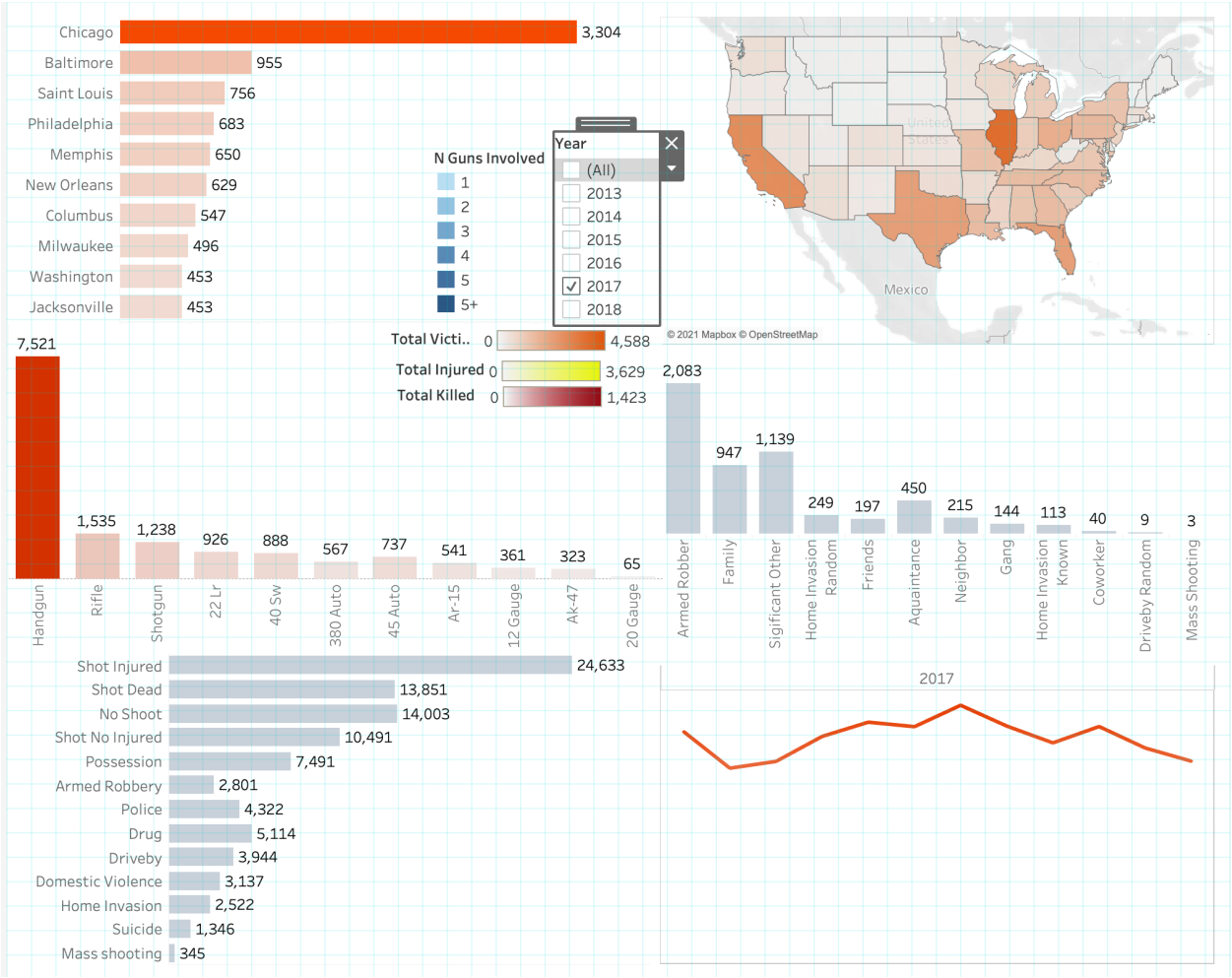
All of the bar charts have their grids removed, and their values displayed on top of the bars for easy readability. It also lets the viewer compare the values of each bar more easily.



Here is the dashboard that provides all the information. The key instrument for the dashboard is the map, as clicking a state will allow all the other graphics to filter to that state's information.

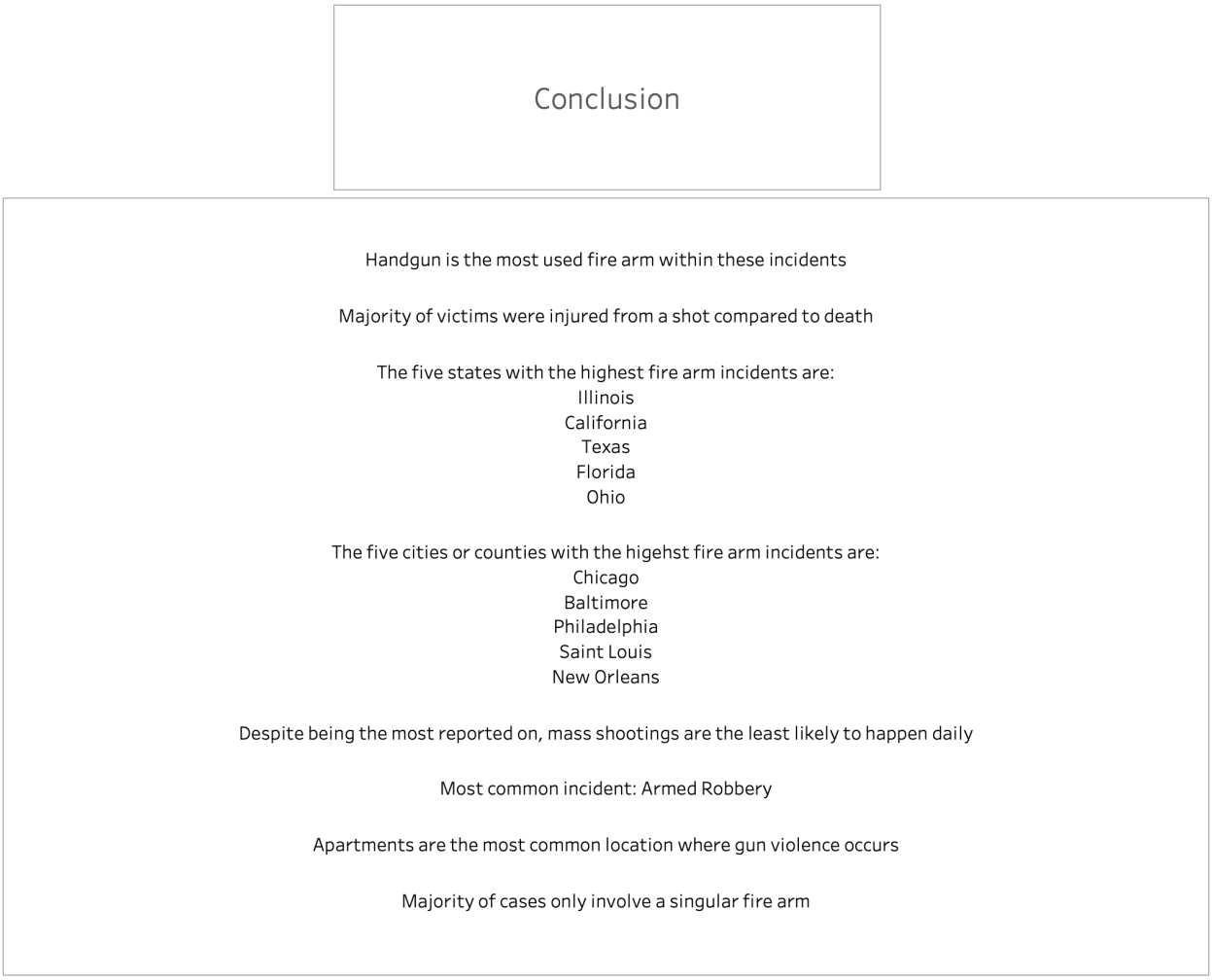


If we want to look at a specific state, just click on a state and it will filter all the information regarding that state. Here it shows for California that the cities or counties with the highest amount of firearm incidents are Los Angeles followed by Oakland. The highest incident type is shot with an injury. Finally it still shows that the highest relationship status is armed robbery.



In the legend, it is also possible to filter based on year. If we are only interested in the year 2017, we can only select “2017”. This can provide the viewer with more specific information regarding what has happened within the year 2017. As previously seen, Chicago is still considered the most dangerous location.

Conclusion



The conclusion slide provides a summary of information that was provided within the presentation. Here the viewer should remember these important aspects as key takeaways from the presentation.

Data Sources and Bibliography

Gun violence archive. Gun Violence Archive. (n.d.). Retrieved September 21, 2021, from https://www.gunviolencearchive.org/.

Ko, J. (2018, April 15). *Gun violence data*. Kaggle. Retrieved September 21, 2021, from https://www.kaggle.com/jameslko/gun-violence-data.