Universiteit van Amsterdam

Data Analysis and Visualization

Artificial Intelligence

Gun-violence in the U.S.A.

Authors:
Jelle Bosscher
Serge van Haag
River Vaudrin
Sarah Yeguez

Supervisor:
Willemijn Beks

June 28, 2018

1 Introduction

Gun violence has been a heavily covered topic in the United States for the past years. Currently, with the rising amount of gun violence incidents and mass-shootings [5], the debate about gun-control is a recurring theme across all of the mainstream media outlets.

The following series of visualizations show an analysis of the data about gun violence in the United States over the 2013-2018 period. The purpose of the study is to understand the underlying principles and patterns of gun violence. In order to achieve these goals various patterns and relations will be explored. First of all, there will be elaborated on the differences between states, cities and years to get a broader picture about the characteristics of this topic. Thereupon patterns are investigated and the ones that are worth to mention are described. In particular, the pattern between killer and victim between various incidents is captivating and examined in detail. Subsequently the correlation between mental health care and gun violence is examined since this might give one information about the significance of mental health care influencing the amount of gun violence related incidents. Moreover, to get a better view of the danger that inhabitants of each state are in with relation to gun violence, the amount of incidents and deaths per state are graphed. Finally, the correctness of a statement from Donald J. Trump, the current president of the United States, about the relationship between gun laws and gun violence is examined.

The broad hypotheses for this studies is partly in line with the research from Michael Rocque and Grant Duwe[5] (where an increase in gun violence incidents and mass-shootings is presented), expecting a continuous increase in gun violence related incidents in the United States. Also, with the inauguration of Donald J. Trump as president and the division that it caused amidst the people, there is expected to be an increase of clashes between communities and uncertainty among people.

2 Method

2.1 Data

The data was taken from the Gun Violence Archive[1], this is a corporation formed in 2013, that provides free access to accurate information about gun related incidents in the United States. The data provides a detailed overview of all the gun related incidents, such as numbers injured, numbers killed, type of gun used, amount of guns used, participant types, participant ages and the location details. The results found in the following pages are comprised of these 239.677 cases of gun violence in the United States from January 2013 up to and including March 2018

To answer all of the questions, a broader context was needed. Data about the population per state, the mental health care per state[3] and the gun laws per state[2] were all scraped off of various websites.

The data about mental health in the U.S per state was acquired from the 'Mental Health America' website[3]. The data shows a table with states ordered by ranking, from better to worse, alas the data consisted only of year 2018.

The data showing the gun laws per state were found on the website of the Stare Firearms Laws[2], it depicts a map of the number of gun laws per state. The data taken covered the same time span as the data about gun violence, 2013 to 2018.

The final dataset was taken off the US Census Bureau website[4]. The data shows a table of the population per state per year from 2010 to 2017. The data was then converted to estimate the population per state per month from January 2013 to and including December 2017. This data was used to calculate per capita, such as incidents per capita.

2.2 Data processing

To properly use the data it first had to be cleaned. As can be seen in figure 1, the data before was inconsistent. The data sheet only showed a maximum of five incidents for any day in 2013 even though there were more. Therefore, the analysis of the data has only been applied to 2014 and onward. Often data from 2014 up to and including 2017 were used, in order to have four full years. The data from 2018 was only complete through March.

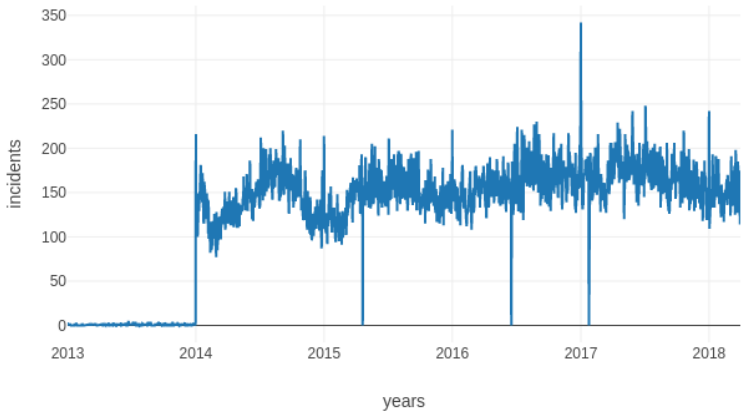

Figure 1: Plot of gun violence related incidents per day from 01-2013 through 03-2018.

The second step of cleaning the data was processing the missing values. The missing values for the data are plotted in the figure 2 and show certain columns with a lot of missing values. The highest missing values percentage was the participant_relationship column with over 93%. With that many missing values the best decision was to drop the whole column. However, since this column is was important factor in the research question on victim/killer relationships the column was kept as it was with the note that it may not be an accurate representation of the actual conclusion.

The other column with over half off the entries missing was location_description. None of the research in this study used that column so it was not taken into account in any analysis.

Any other columns with a high number of missing values the analysis was either only done on the values that were present or the time period was cut to get an accurate representation of the analysis.

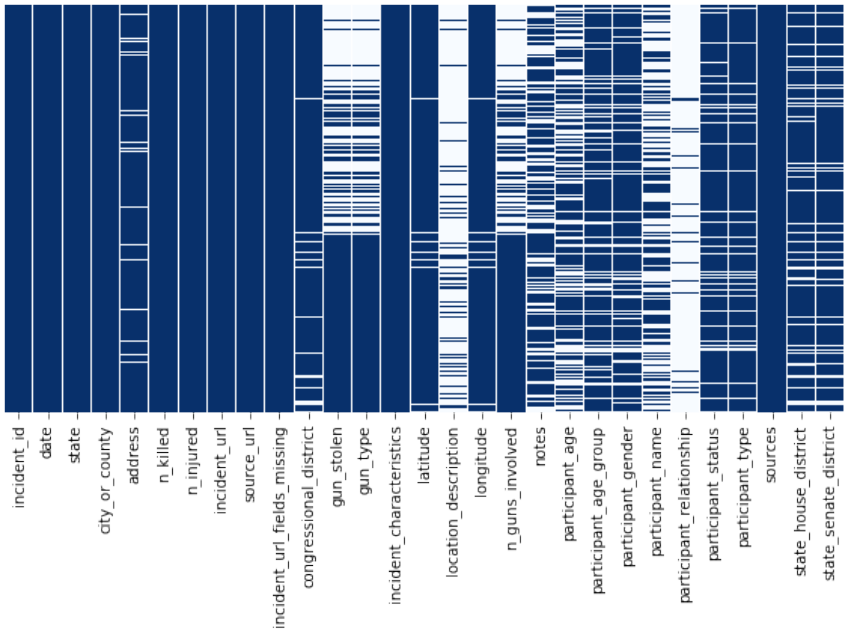

Figure 2: The missing values from each of the columns in the Gun Violence Dataset[1].

The only exceptions were the latitude and longitude columns. There were 7923 missing values in each of these columns and both columns had the exact same missing rows. Because the columns for state and county_or_city both had 0 missing values and the address column only had about 7% missing values, the missing values were calculated for those columns. This was done by using Google's GeoCoding API, where supplying a list of at least state and city_or_county values and if possible the address value yielded a result. The result was a JSON object containing, among other things, the latitude and longitude for the specific entry. A plot of all incidents from January 2013 up to and including March 2018 can be seen in figure 3.

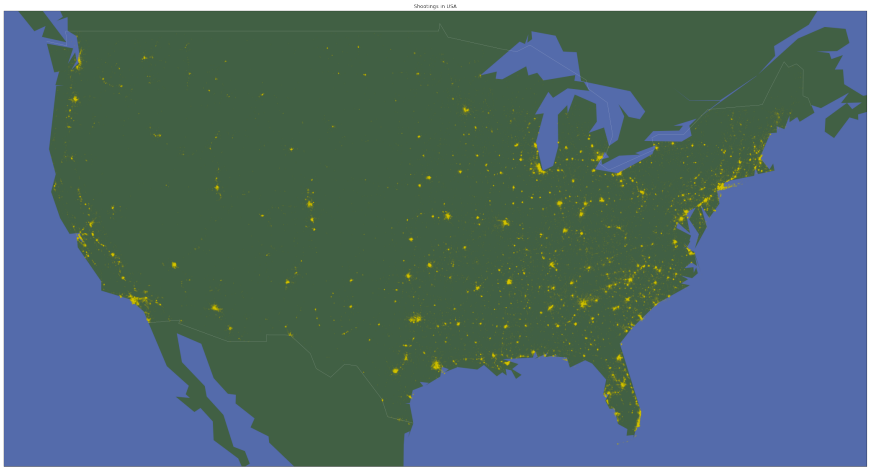

Figure 3: Plotted topographical location of every data point.

2.3 Techniques

In order to search for patterns, an implementation of a k-means algorithm was chosen to generate multi-dimensional data. The sample generators used were included in scikit. Columns that seemed interesting to analyze were put into a new csv. Since k-means clustering only accepts input that are integers, non-numerical values in columns were converted to a unique (integerger) value After choosing which columns - 'n_injured', 'n_killed', 'victim', 'suspect', 'unknown', 'female', 'male' - will be grouped together to analyze, the data set was divided into two parts: a training set (80%) and a testing set (20%).

Thereafter the KMeans command was used from the sklearn.cluster package. It only requires to pass the number of clusters (n_clusters). The 'trained' system was then tested with the 'training' set to examine it's accuracy. The number of clusters was chosen by examining which quantity of clusters in a trained system had the highest accuracy. Visualizations in higher than two dimensions is tricky, since anything above two dimensions would be difficult to read and plot, thus the clusters were therefore represented in a table.

3 Results

3.1 Differences between states

Since 2014 up to and including 2017, the most gun-violence incidents were reported in the state Illinois, followed up by California and Florida respectively. This has been displayed in Figure 4. Illinois is the state with the most gun violence partly due to the fact that city Chicago is in Illinois. The last mentioned city is the city with the most gun violence of the United States. This

can be seen in Figure 7 and in the next section there will be a further elaboration.

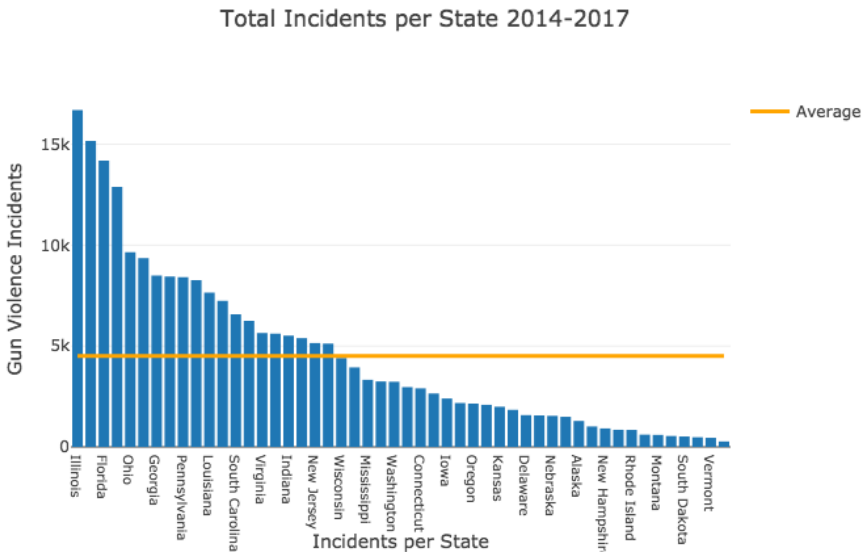

Figure 4: A histogram of the total number of incidents per state

The next analysis shows the different guns used in the top three states with the most incidents. This is displayed in a pie chart, see Figure 5. It is evident that handguns are the most popular in gun violence in Illinois, California and Florida. Handguns are indicated with the big blue plane in the pie chart below. However, it should be noted that a 9mm, 40 Smith and Wesson and a 45 auto are examples of handguns as well, yet they are displayed separately. This is because the different types of guns are not always recognized and so they are often generalized.

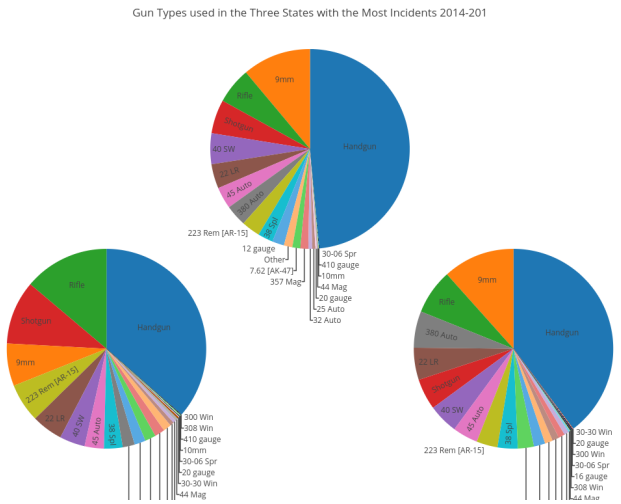

Figure 5: Pie chart of the guns used in all incidents in three states. (from left to right California, Illinois, Florida)

Figure 6 shows that the most gun violence per capita happened in the state Alaska; there is one incident reported per every 574 inhabitants in the state. This is remarkable since Alaska is one of the states with a low number of reported incidents compared to the other states in the United States. It is even ranked twelfth place, if you take a look at the states with the least reported incidents.

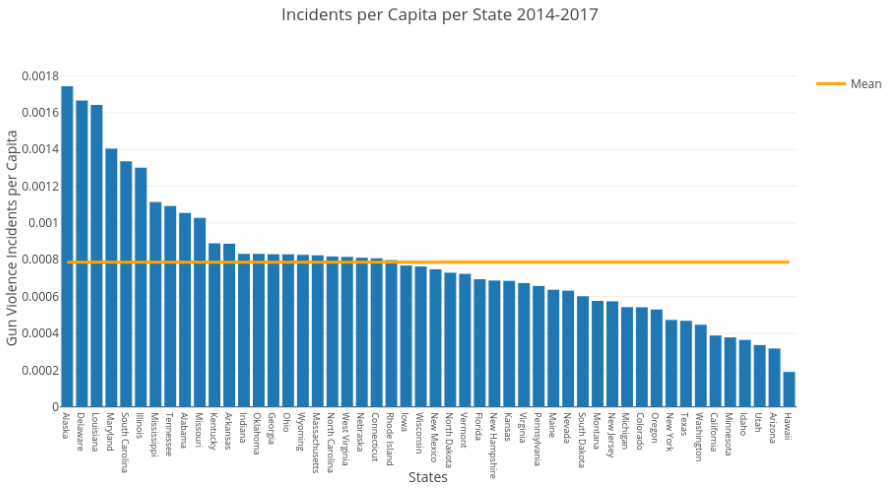

Figure 6: Histogram of the incidents per capita per state from 2014 through 2017.

3.2 Differences between cities

As stated before, Chicago is the city with by far the most gun violence in the United States in regard to the years 2014, 2015, 2016 and 2017. It has been displayed in Figure 7, the histogram shows that Chicago has a lot more incidents of gun violence than the second place Baltimore, 2.78 times as much.

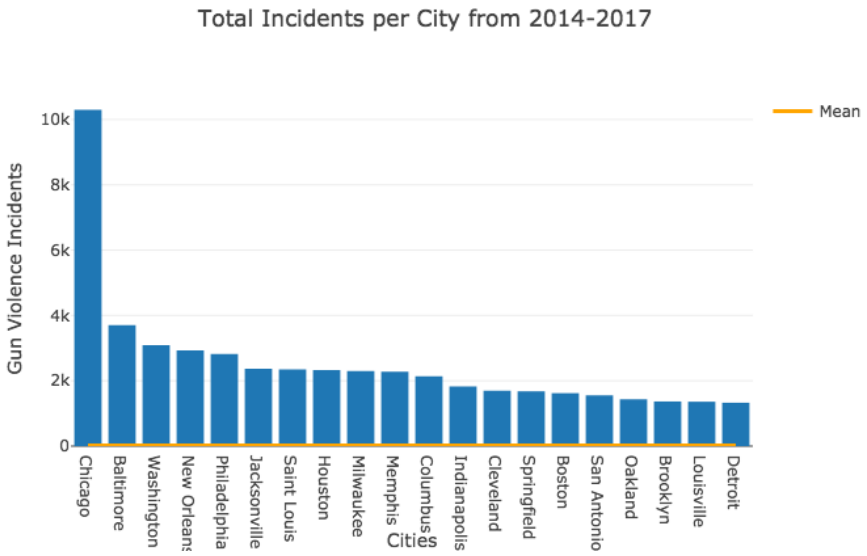

Figure 7: A histogram of the top 20 cities with the most incidents in it from 2014-2017. The yellow line displays the mean of all cities in the United States, which is 17.9.

Figure 8 shows the ratio of number of people killed and number of incidents per city. In the figure you can see that 80 people are killed per 100 incidents in Compton. On average, 22 people are killed per 100 incidents over the entire United States. That means that Compton is more than three and a half times as deadly regarding gun violence than the average. As to the relationship between killer and victim in Compton, there were only seven data points; four of them were gang-related and the other three were neighbours, family and robbery.

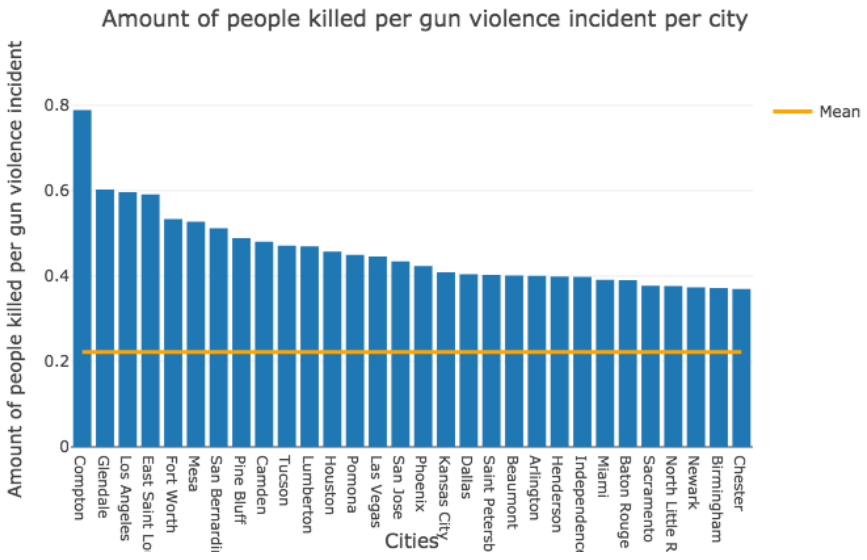

Figure 8: A histogram of the top 30 cities with the most people killed in it per incident from 2014-2017. The yellow line displays the mean of all cities in the United States, which is 0.22.

3.3 Differences between years

In Figure 9 you can see that in the year 2014, 2015 and 2016 there is a general increase of gun violence incidents as the year passes. In contrast, a general decrease of incidents in the year 2017. The underlying cause is that incidents sky rocketed on average in the end of 2016 and in the beginning of 2017. The year 2016 ended with an average of ± 175 incidents per day and on the first day of 2017 there were more incidents than ever, that is to say 342 reported incidents. This caused the average trend line in 2017 to start off very high and it restored slowly to a more 'normal' amount of an average of ± 160 incidents at the end of 2017. This is quite similar to the average incidents at the end of 2015. The amount of incidents each day through the four years can be seen more clear in Figure 10, which shows a polynomial regression line. You could see that there are 156 incidents reported on average at the end of 2015 (day 730) and this amount is the same on the last day of 2017 (day 1460).

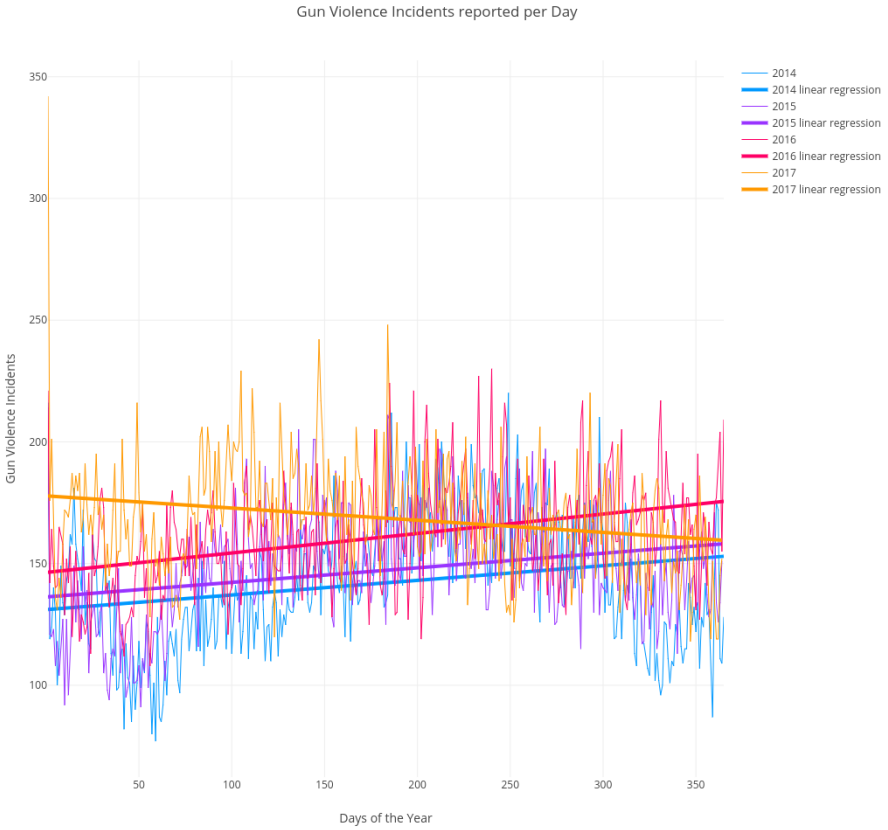

Figure 9: A plot of the amount of incidents on each day for the years 2014 through 2017 with the applied linear regression.

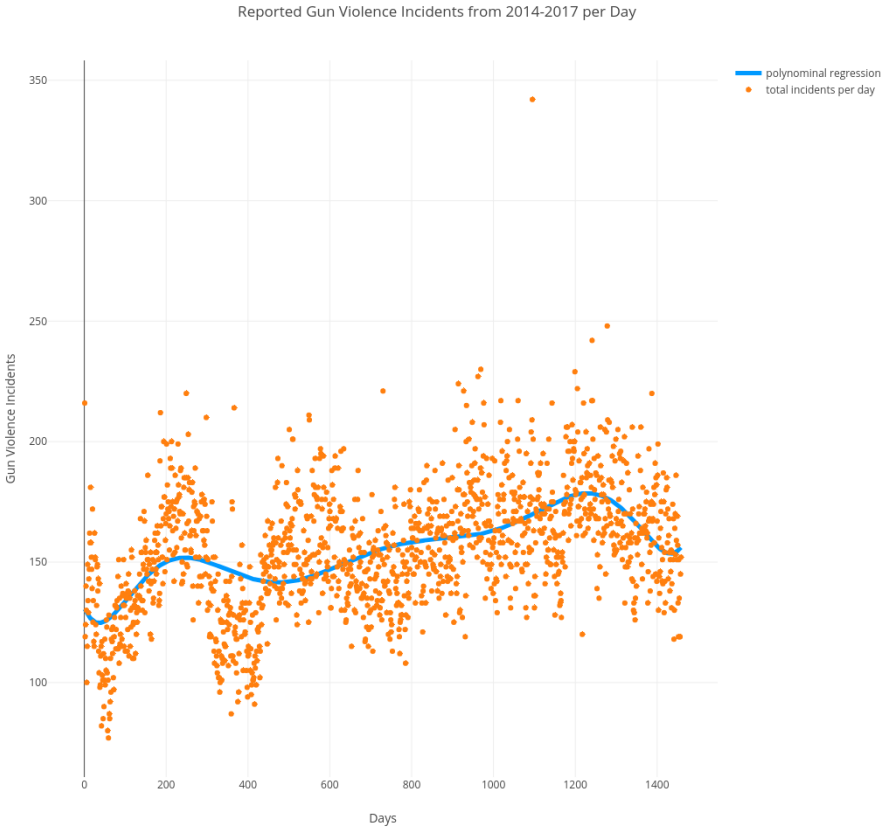

Figure 10: A scatter plot of the amount of incidents from 2014 through 2017 with polynomial regression.

3.4 Interesting patterns

The following tables tell us what the differences are between the clusters and shine light to some patterns. It shows mean values of the attribute per each cluster.

By observing table 1 it looks like the entries of cluster 0 are cases that predominantly consists of killings in which participants are mostly male. Cluster 1 are cases that only consist of injuries and male participants. Cluster 2 are cases that predominantly consist of injuries in which participants are mostly male. Cluster 3 are cases that have fairly more injuries than killing with slightly more females than males.

Table 1: Patterns between killings, injuries, female and male participants.

clusters	n_killed	n_injured	female	male
0	1.132826	0.099556	0.085094	1.61359
1	0.000000	0.364279	0.000000	0.807268
2	0.055466	1.227119	0.053248	2.677199
3	0.279255	0.635759	1.156974	1.03045

By observing table 2 it looks like cluster 0 are cases that consist mostly of injuries with a close even ratio between suspect and victim mainly consisting of male participants. Cluster 1 are cases that are composed mainly of killings with participants consisting mostly of male suspects. Cluster 2 are cases that only contain injuries with participants predominantly being male victims and suspects

Table 2: Patterns between killings, injuries, victims, suspects, female and male participants.

clusters	n_killed	n_injured	female	male	victim	suspect
0	0.186241	1.225540	0.454487	2.824093	1.673687	1.921459
1	1.124958	0.091105	0.223925	1.400510	1.069254	0.633239
2	0.000000	0.449644	0.103755	0.903563	0.525479	0.642006

By observing table 3 it looks like cluster 0 are cases that have fairly more kills than injuries participants and consists of mainly suspect. Cluster 1 are cases that predominantly contain injuries and consists of mainly victims.

Table 3: Patterns between killings, injured, suspects, victims and unknown (no data of participant type).

clusters	n_killed	n_injured	unknown	suspect	victim
0	0.360616	0.004046	0.177418	0.948258	0.444797
1	0.099780	1.183807	0.000000	0.666834	1.313324

In figure 11, one can observe that the number of gun violence increases through the years 2014 and 2017 and thereafter decreases. The blue outline starting in April 2018 is a prediction made. One can also observe that gun violence mainly occur on Saturday and Sunday, between the months April and November with December and January having a lower number of incidents and March having nearly none.

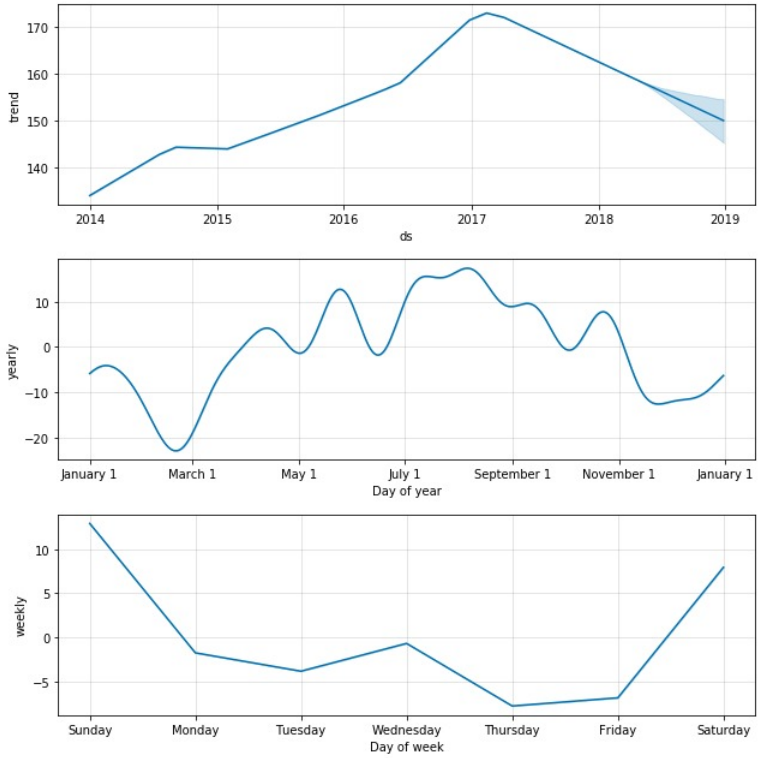

Figure 11: The plotted components for the forecast of the remains 9 months of 2018. With a yearly and weekly trend.

3.5 The pattern between killer/victim

The relation of killer/victim is plotted in a scatter plot with linear regression as seen in Figure 12. The scatter markers and lines of the same color belong to the same relation. To answer the question with the data set it has to be noted that from the 239.677 rows of data there are 223.903 missing values, leaving only 15.774 data points. The visualized data is from January 2014 up to and including March 2018, plotted per month showing the percentage of the seven most significant relations. Where significance is determined by percentage of the total amount of relations that occurred per month.

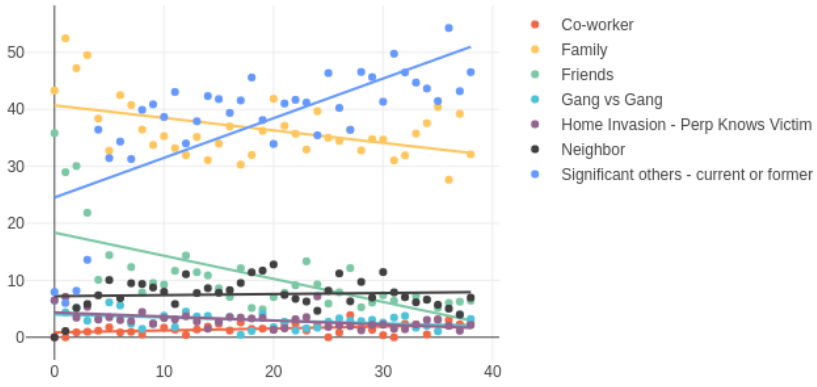

Figure 12: Scatter plot of the percentage of incidents per month for each relation and the associated linear regression.

3.6 The relation between mental-health and gun-violence

To answer this question, it required the data about the mental health care ranking through all the states and the analysis about gun violence per capita. Figure 13 shows the states in the order of best accessibility to mental health care to worst, to find the data points for each state the total amount of incidents per state were divided by the population of that state. The blue line shows a polynominal regression line. Unfortunately, the data was ranked from 'best state to worst state' instead of an actual score, since the latter would probably have given more insight to the relation.

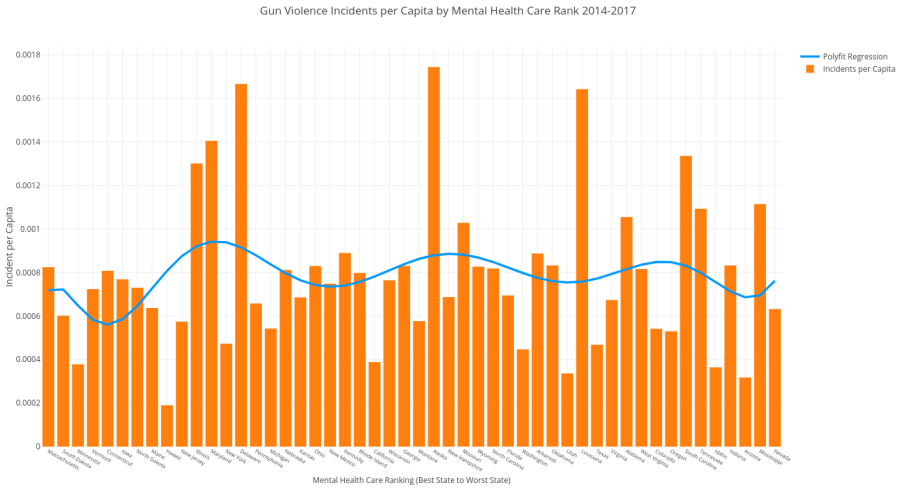

Figure 13: Histogram of the number of incidents of gun-violence per state per capita. The states are ordered by the accessibility to mental health care ranking.

3.7 Most dangerous states taking population into consideration

One could say that the number of incidents in a state defines how threatening or dangerous a state is. However, the number of deaths per shooting might define the real danger of a state. In Figure 14 and 15 these two are plotted in a histogram, in both figures the population size was considered. In Figure 14 Alaska came up as the state with the most incidents per capita and in Figure 15 it was Louisiana.

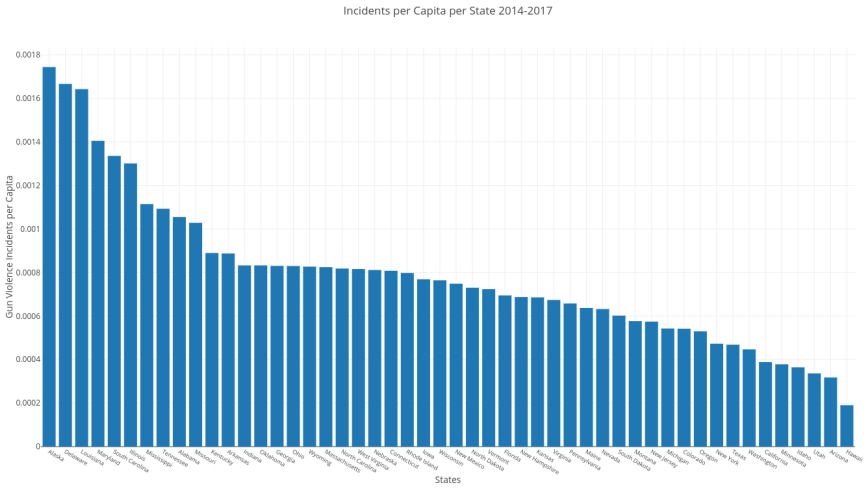

Figure 14: Histogram of the number of incidents of gun-violence per capita per stare

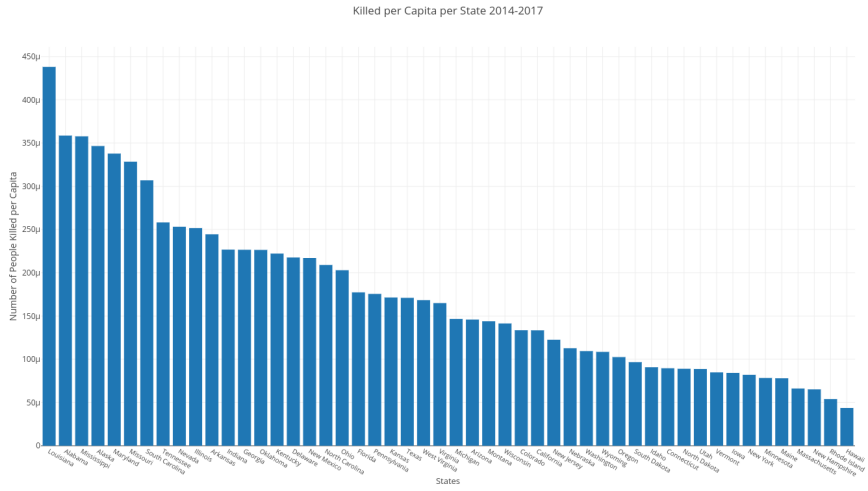

Figure 15: Histogram of the number of people killed by gun-violence per capita per state

3.8 The relation between gun-laws and gun-violence

To find the relation between the amount of gun-laws and gun-violence, there was taken a look at the amount of incidents in Chicago and the amount of gun-laws in Illinois. Figure 16 is a histogram mapping the total amount of incidents per city, and unsurprisingly Chicago the city with the most incidents.

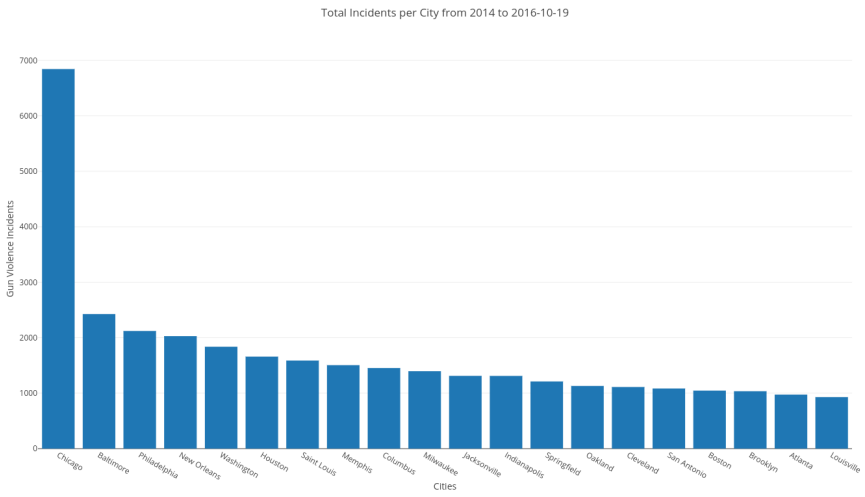

Figure 16: Histogram of the total amount of incidents per city

Table 4 gives the ranking of the first seven states, ordered by their average number of gun-laws over the 2014-2016 period. Illinois ranks as the seventh highest state in the United States when it comes to gun-laws. The amount of gun-laws in Illinois were still a considerable amount less than in California and Massachusetts.

Rank	State	Mean amount of Gun Laws 2014-2016
1	California	102.7
2	Massachusetts	101.0
3	Connecticut	86.7
4	Hawaii	78.3
5	New York	76.0
6	New Jersey	67.0
7	Illinois	65.0

Table 4: Ranking of the seven states with the highest amount of gun-laws over the 2014-2016 period

4 Discussion

The different guns used in the top three states with the most incidents -Illinois, followed by California and Florida respectively- were analyzed. It is evident that handguns are the most popular when it comes to gun violence in Illinois, California and Florida.

However, it should be noted that a 9mm, 40 Smith Wesson and a 45 auto are examples of handguns as well, yet they are displayed separately. This is because the different types of guns are not always recognized and so they are

often generalized.

Most incidents per capita happened in Alaska; there is one incident reported per 574 inhabitants in the state. This is remarkable since Alaska is one of the states with a low number of reported incidents compared to the other states in the United States, ranked twelfth in place, and having the least amount of gun laws.

On average, 22 people are killed per 100 incidents in all cities in the United States while in Compton 80 people are killed per 100 incidents. Making Chicago the city with the most gun violence in the United States regarding the years 2014 to 2017.

There is a general increase of gun violence in the years 2014 to 2016 and a decrease in the year 2017. The underlying cause is that incidents sky rocketed on average at the end of 2016 and at the beginning of 2017. The year 2016 ended with an average of ± 175 incidents per day, and on the first day of 2017 there were more incidents than ever with 342 reported incidents. This caused the average trend line in 2017 to start off very high and slowly turned to a more 'normal' amount with an average of ± 160 incidents at the end of 2017. This is quite similar to the average incidents at the end of 2015, which had an average of 156 incidents reported at the end of 2015 (day 730), this amount is similar to the last day of 2017 (day 1460).

So it seems that victims in overall get mostly injured and suspects get mostly killed. Moreover if an incidents does not have any killings, one can assume that no females were involved but if females mainly take part in incidents where there are more injuries than killings.

Furthermore, gun violence report during the holiday seasons in December and January are far lower than other months. Holidays around this time tend to make people more caring and compassionate towards each other. What's more, people tend to have more free time during the weekend rather than during the week. Which explains why incidents are more active on Saturday and Sunday.

Lastly, in line with the hypotheses we expected the amount of incidents would increase over the last 9 months of 2018, although prediction states it will decrease. This might be owed to the increasing attention of gun violence in the U.S

Because the results from the analysis of the relation between killer and victim are based on such a small part of the dataset, the results are not a correct representation of the answer to the question. Nonetheless, graphs have been analyzed and have yielded some substantial, though negligible in context, results. The trends of significant others and family related incidents are clearly visible, with family related incidents decreasing over the years and significant others related incidents increasing. Incidents related to friends went from about 20% to 3% over the course of 39 months.

Mental Health America provided mental health care ranking of states. Unfortunately, the data was ranked from 'best state to worst state' instead of an actual score, the latter would probably be more precise. The obtained results were inconclusive, a direct correlation was not apparent or non-existent.

According to the analysis done on two datasets[1][4] about gun violence and population, the most dangerous state is Alaska. However, Alaska and most of the other states in the top 10 most 'dangerous' states all have relatively low population. The only outlier seems to be Illinois. Going from the assumption that a low population in a state skews the results, Illinois should be ranked higher. For future research, a different technique might not favor states with low population.

The quote from Donald Trump was interpreted as a statement stating that even though Chicago has by far the most gun laws of any city, it still has the most gun related incidents. According to the analysis of the two datasets[1][2] this is considered false. There does not seem to be an obvious correlation between the amount of gun laws and the amount of incidents per state. It has to be noted that used dataset is the amount of gun laws was per state and not per city. Although most cities do not differ from the states in the terms of gun laws, some do. If the gun laws would be found per city and the analysis would be performed again, it might yield different results.

The results of the research question about the pattern between killer/victim are not conclusive. This has everything to do with the lacking data available for the analysis. The solution to improve this is to acquire more data. This could be done by either finding a different data source with a complete view on the victim/killer relation for all the events or possibly by filling in the missing data from the sources of the entries in the Gun Violence Dataset[1].

The forecast of the trend for the remaining nine months of 2018 does not line up with the hypotheses nor does it line up with the research from Michael Rocque and Grant Duwe[5]. The amount of incidents in the remainder of 2018 does not seem to increase according to the forecast although the hypotheses is that it would. This could be due to a faulty hypotheses for this question or an incorrect method to calculate the forecast. By using an algorithm that mostly focuses on yearly seasonality for the time series the result may be different.

There also was some uncertainty around the regression of certain data plots. To get a better visual approximation of polynomial regression lines, it is necessary to approximate such a line on a training data set and then measure the performance on a testing set by observing the mean error. The polynomial regression line that performs best on the testing set should then be used. In this study, the degree of the polynomial regression was increased until the line visually fitted right on the scattered data points. In the future, using the described method with training and testing sets the result are expected to be more precise.

Finally, it would be interesting to analyze the correlation between the number of incidents per state and the accessibility to obtain guns. Also, because of the controversy around race or color, it would have been interesting data to work with. Apart from age and sex that was not much demographic data to work with.

References

[1] Gun violence data in the US. https://github.com/jamesqo/gun-violence-data. Accessed: 06-2018.

[2] Firearm laws in the US. https://www.statefirearmlaws.org. Accessed: 06-2018.

[3] Mental health rankings in the US for 2017. http://www.mentalhealthamerica.net. Accessed: 06-2018.

[4] Population in the US for 2010 through 2017. https://factfinder.census.gov. Accessed: 06-2018.

[5] Michael Rocque and Grant Duwe. Rampage shootings: an historical, empirical, and theoretical overview. 2017.