Gun Violence in the Lower 48 States

Contents

Research Questions

1. How has the amount of gun violence changed between 2013 and the early part of 2018?

   The number of incidents and gun deaths have steadily increased until the end of 2016, when an abrupt decrease began.

2. What does it mean to have an unusually large amount of gun violence?

   Due to the unusually skewed distribution of people who get hurt or killed, it is only in 10% of counties in the U.S. where any people have gotten hurt or killed at all, and any more than 4 or 5 people hurt or killed in a given county over the entire time of our study would be unusual.

3. Where is gun violence most evident?

   It is clearly most evident where population density is the highest, places like L.A., Seattle, Detroit, Chicago, and Washington D.C.

4. What can the data tell us about the causes of gun violence?

   It establishes very little other than the fact that gun violence is correlated with population density.

5. How closely is legal gun ownership related to gun violence?

   Our data indicates that legal gun ownership is independent of the amount of gun violence. Gun violence increases when the number of legal gun retailers per unit area increases, but this is likely due to a lurking variable in population density. When adjusted for population, the correlation is both negative and extremely weak.

Key Takeaways:

There was very little that we could say conclusively, other that the fact that gun violence appears to be strongly correlated to population density. Gun violence per person appears to be somewhat independent of population density, so as the population density grows, the amount of gun violence will increase per unit area.

Gun Violence in the U.S. is very far from uniform. Rather than being evenly spread out, gun violence is concentrated in areas with the highest populations, such as southern California, Washington D.C., Miami, and much of the east coast.

Motivation and Background

Mass shootings are always a tragedy, but the way they are covered, they seem to be the most prevalent and violent occurrences of gun violence. However, we believe that any death due to gun violence is an equal tragedy, and that this view of only a slice of gun violence trivializes other incidents. Given that gun violence and the second amendment are so hotly debated, it would also be beneficial to see the bigger picture. We wanted to see how gun violence is concentrated, since the common claim is that, because Americans own more guns, there is a higher likelihood of people getting shot or killed. We wanted to know if this is the case, and, whether it is or is not, give some more background about the reality of gun violence and as much about the causes of it as we can, as certain politicos have tried to use mass shootings as an excuse to widen gun laws.

We therefore wanted to find out as much as we could about the causes and concentration of gun violence as it stands in the U.S. This means looking at concentration with regards to area, personal income, and legal gun ownership (as far as it can be measured).

Datasets

1. Gun violence in the US
    a. A great database about gun violence – whether the gun was stolen, location, perp, etc.
    b. https://github.com/jamesqo/gun-violence-data
    c. .csv files are in the /intermediate/ folder
2. Geospatial data of US States
    a. Map of US States and territories, with county data
    b. https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html
3. Personal income data
    a. https://apps.bea.gov/regional/downloadzip.cfm
    b. We are using the second drop-down menu, second choice, "Annual Personal Income by State", under "Personal Income, State and Local"
4. Gun licensees
    a. It is illegal to keep track of gun ownership, we track gun licensees
    b. https://www.kaggle.com/doj/federal-firearm-licensees/version/3
    c. This may require a Kaggle account, but it is free and easy to set up with Google.
5. State Population Data
    a. We restricted this to the '2015' column, as we reasoned the population differences from year to year wouldn't be too dramatic.
    b. https://www.census.gov/data/datasets/time-series/demo/popest/2010s-state-total.html
6. State FIPS codes and Names
    a. This is mostly common knowledge, but this acted as a translator of sorts so that we could go between state name, stat FIPS code, and state abbreviation easily.
    b. https://gist.github.com/dantonnoriega/bf1acd2290e15b91e6710b6fd3be0a53

Methodology

Numbered list correlates to research question

1. Gun Violence over time was a reasonably straightforward question to answer. We organized the gun violence database by year, and it was already broken up into months. We simply counted the total number of incidents and took a sum of all the people hurt or killed, and plotted them, by number of months after 2013, on a scatterplot with a regression line drawn through them. We also plotted the residuals of the same scatterplots to make any pattern more apparent.
2. The question of an unusual amount of gun violence was tricky to answer. We did so by aggregating data from the entire timeline we had by county, then plotting the counts of people hurt or killed for each county on a histogram. Since this showed an incredibly skewed distribution, we also filtered out those counties where nobody was hurt or killed at all (almost 3000 counties) and plotted that on a separate histogram. This allowed us to see how violence was distributed and say what would be an ‘unusual’ amount of gun violence. Further statistical analysis was deemed unnecessary, as the data was skewed enough to give results without detailed analysis. We initially planned to do a case study of Washington State, but then found county data for the U.S. and population data for each state, and decided that we might get more general results if we were to only consider the U.S.
3. This question was answered by plotting a map of the lower 48 states and DC, colored by the amount of gun violence. As this was somewhat difficult to see, we also came up with a different plot. Reasoning that, in 3108 counties in the lower U.S., there were 310 counties in the top 10%, 155 in the top 5%, and 78 in the top 2.5%, we aggregated gun violence by county, then plotted the top 10% yellow, the top 5% orange, and the top 2.5% red, and plotted them on an empty plot of the U.S. These plots together show the most dangerous areas of the United States: highly populated regions like southern California, Seattle, Chicago, Washington D.C., etc.
4. Statistical tests turned out to be unnecessary for this question, as regression analysis provided everything we needed. We did regressions for each of three predictors of gun violence for which we had data: population density, gun licensee density (the density of registered gun retailers), and personal income per capita. We plotted regressions for each of these variables, calculating density with area as the denominator and with population as the denominator, to give us information as to how violence scales with area and with population (explained below). The correlation coefficient was also calculated for each plot so the strength of the correlations could be gauged. After that, we trained a machine learning model on combinations of the variables above to see whether combinations of the variables were significantly more accurate than others. We took the average mean squared error over 100 trials for each model to reduce variation.
5. This was answered as a part of Question 4.

Results

How has the amount of gun violence changed between 2013 and the early part of 2018?

The plot of gun violence over time (at right) shows that gun violence has been steadily increasing, peaking in the late months of 2016. However, the residuals indicate a pattern, particularly with the number of incidents over time. The large gap around month 40 indicates that there is a pattern to the regression that is not entirely captured by the regression. The residuals indicate that there is a decrease that begins at roughly month 40, which is the early part of 2017. This coincides most nearly with the 2016 election, but it is impossible to say what causes it. One possibility is the economy that improved after the election, but there is no way to say for sure.

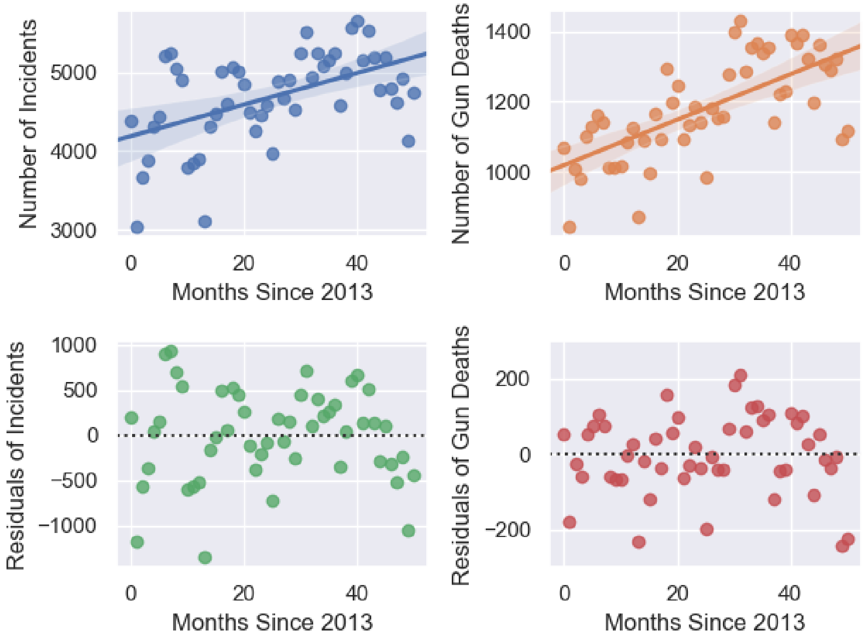

Figure 1: Gun Crime Over Time

What does it mean to have an unusually large amount of gun violence?

Gun violence is not occurring everywhere in the United States, in fact the vast majority of counties have less than three or four incidents over the five-year span of the data. The distribution of gun violence is highly clustered is not normally distributed. The plot below (left) shows the overall distribution of casualties by county, although it is very difficult to see anything. What is clear is that most counties have very little in the way of gun violence, while some have very high numbers. The plot on the right shows distributions from 2013 only, which illustrate the same trend over a smaller scale. It shows that close to 3,000 of the 3108 counties for which we have data had less than 4 or 5 people hurt by guns, if any. The plot on the bottom right is a tiny subsection of the plot on the bottom center, which shows the incredible way in which the data is skewed. It is therefore unusual for very many people to get hurt at all, so any county with more than four or five injuries in a year is unusual.

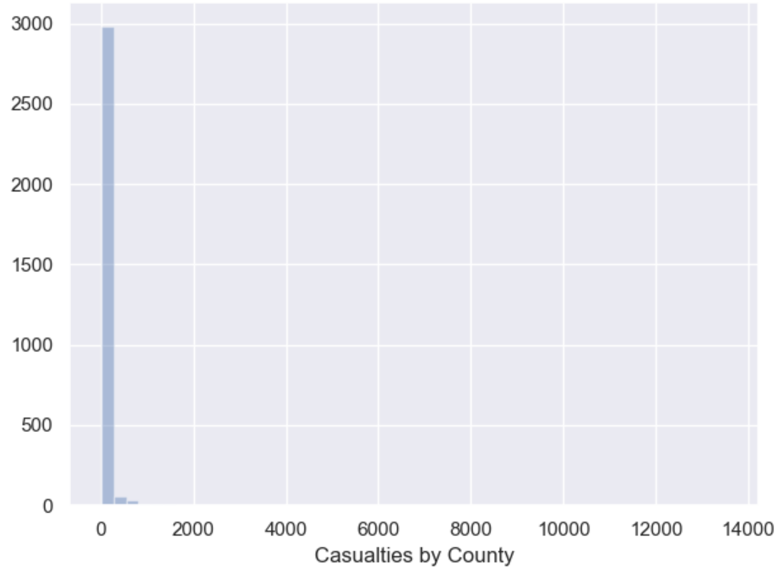

Figure 2: Histogram of Entire Timeline

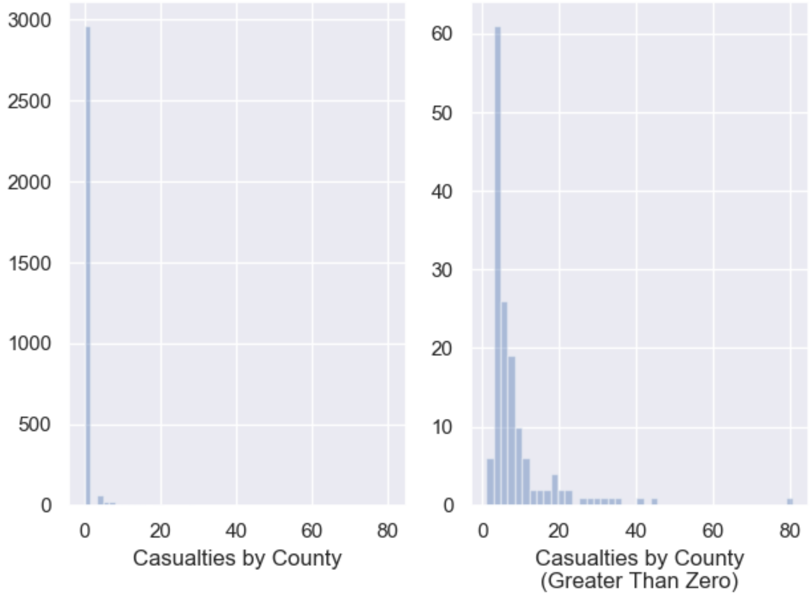

Figure 3: Histograms of 2013 Only

Where is gun violence most evident?

Gun violence is most prevalent in the larger cities in the United States. Los Angeles, Chicago, Detroit, Miami, New Orleans, Philadelphia, Houston, and the Brooklyn borough of New York. The areas highlighted below are the most violent counties in the U.S., highlighted by degree. The highlighted regions seem to match the most populated areas of the U.S.; southern California, Miami, Seattle, the Twin Cities, Detroit, Chicago, and other densely populated areas are all represented below.

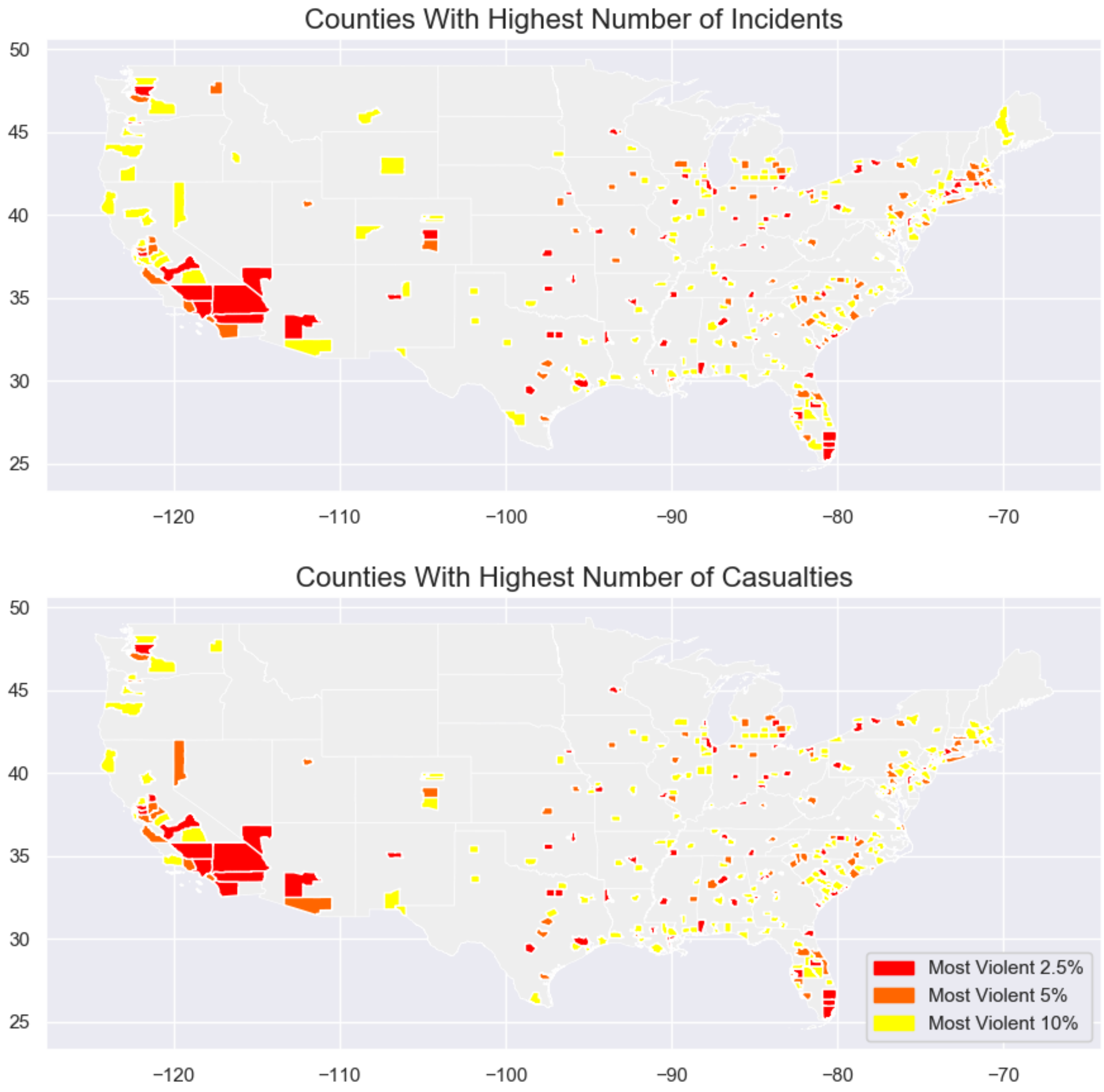

Figure 4: Most Dangerous Counties

What can the data tell us about the causes of gun violence?

We plotted regressions for each predictor, with the only conclusive predictor being population density. The first plot shown at right is a plot of gun crimes per unit area over population density, which increases. This plot does have a lot of high outliers, so we filtered out any states with populations over 1.5 million, allowing a more conclusive plot to be drawn. Both plots show gun crimes per unit of area. For all of these plots, D.C. is excluded, as it is an incredibly high outlier that makes the rest of the data invisible.

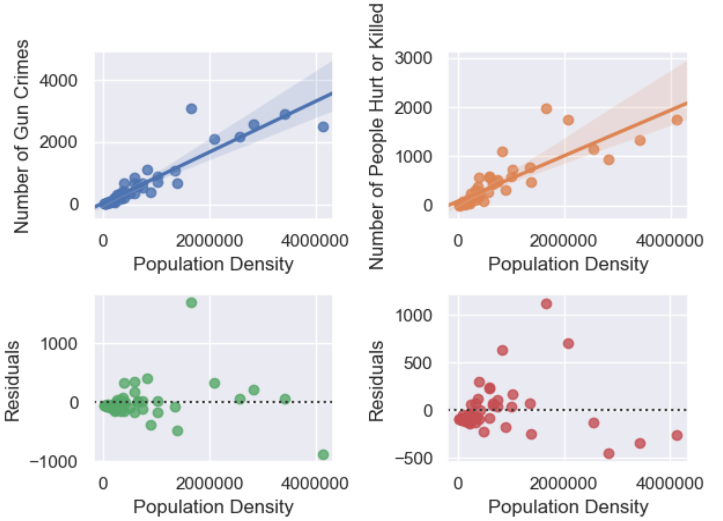

Figure 5: Gun Crimes per Unit Area over Population, without high outliers

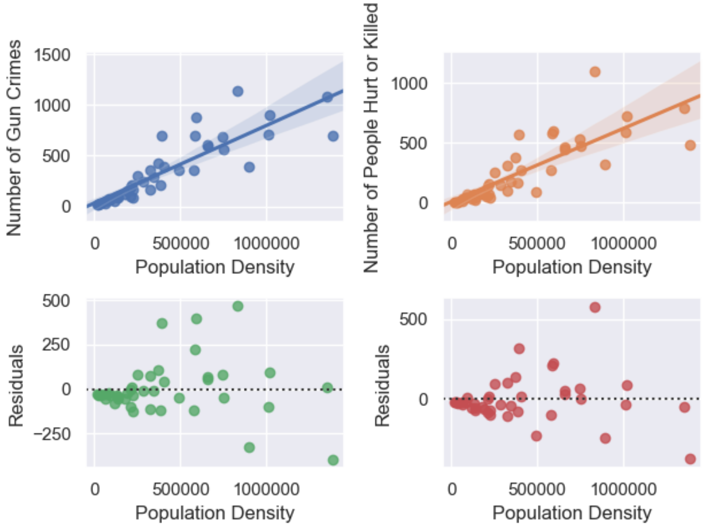

Figure 6: Gun Crimes per Unit Area over Population, without D.C.

The correlation of both plots is over 0.8, indicating a moderate to strong correlation. This was a much stronger correlation than any of the other plots. However, both residual plots show a lot of scatter, meaning the correlation becomes less strong as population increases, indicating that highly populated areas vary tremendously in their crime, but lower populated areas are somewhat universally low in crime. The machine learning data tells a similar story. With both combinations including population density, the mean squared error is similar with all combinations that include population density, but all the other predictors seem to have negligible effect.

Combination	Error for Gun Crimes	Error for Gun Casualties
Population Density and Licensee Density	215404.6327	125708.9404
Population Density and Income Per Capita	193064.4215	129106.275
Licensee Density and Income Per Capita	530695.2632	243943.2094
All Three	185942.7932	128396.9008

Table 1: Machine Learning MSE from data without Washington D.C.

This, however, is not the whole story. When plotting gun violence over licensee density per unit area, there appears to be a strong correlation. However, this is likely due to a lurking variable, in that there will generally be more gun retailers where there is higher population density.

When plotting gun crimes per person over population density, there is almost no correlation This means that the number of gun crimes per person is independent of population density. Therefore, if the number of gun crimes per person is constant, then the number of gun crimes per unit area will increase proportionally to the number of people per unit area.

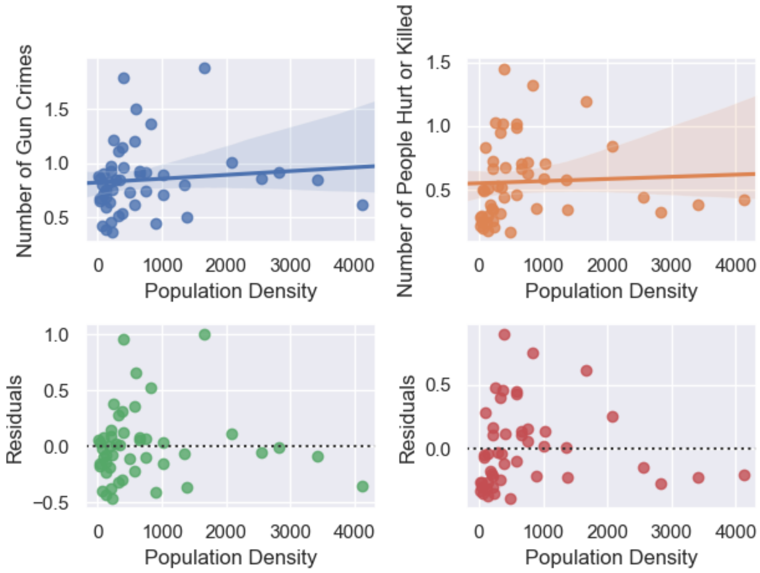

Figure 7: Gun Crime per Person over Population

How closely is legal gun ownership related to gun violence?

Gun crime appears to correlate closely with gun licensee when considering both as a function of area, but this is most likely due to the lurking variable of population, which causes both to increase together. However, as evidenced by the machine learning data, there is no substantial benefit to having data on gun licensees when predicting gun violence. When gun violence is plotted over licensee density per person, there is no substantial correlation. This seems to indicate that legal gun ownership has no effect on gun violence. There is even a slight negative correlation, but that is likely due to the high outliers on the right-hand side, which have high leverage over the regression.

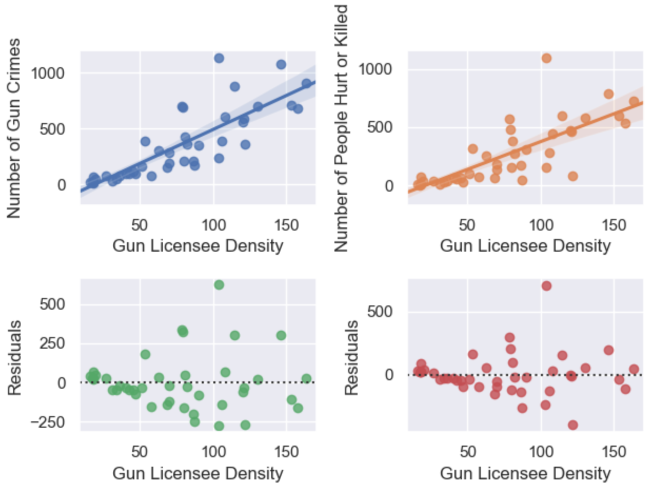

Figure 8: Gun Violence per Unit Area over Gun Licensee per Unit Area

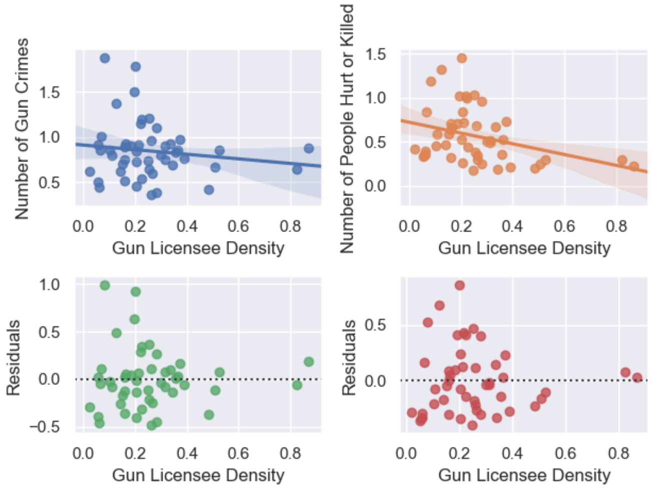

Figure 9: Gun Violence per Person over gun Licensees per Person

Why is Washington D.C. Not in any of the Regressions?

Washington D.C. has a small enough area, high enough population, and disproportionate enough amounts of gun violence that the plots that include it are useless. The outlier in all following plots is D.C. In all the plots above, D.C. is excluded as an outlier.

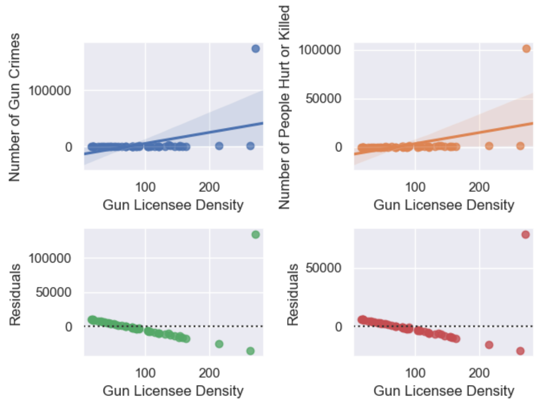

Figure 11: Gun Crimes over Gun Licensee Density, with D.C.

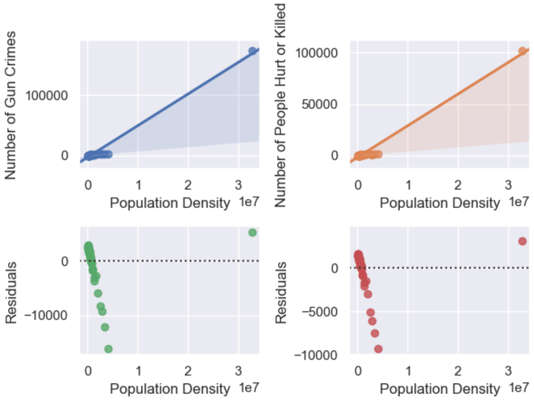

Figure 10: Gun Crimes over Population Density, with D.C.

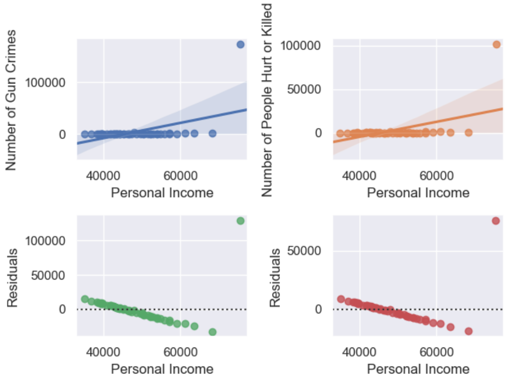

Figure 12: Gun Violence over Personal Income, with D.C.

Reproducing Results

data_parser.py was based off of a very sophisticated directory. It was written to work on any computer, but the directory must be intact for it to work. The directory should be organized as follows, with the exact same spelling and capitalization (gun violence .csv system is abbreviated for brevity, but the entire gun database must be sorted as it is below. To download the datasets, follow the hyperlinks on page 3 and the directions there.

- Master project folder
    - Databases
        - FIPS Codes
            - us-state-ansi-fips.csv
        - Firearm Licensees
            - Federal-firearm-licensees.csv
        - Geospatial Data
            - US County Map
                - cb_2018_us_county_500k.shp
        - Gun Violence
            - Stage 1
                - 2018
                    - stage1.01.2018.csv
                    - stage1.02.2018.csv
                    - stage1.03.2018.csv
                - 2017
                    - stage1.01.2017.csv
                    - stage1.02.2017.csv
                    - stage1.03.2017.csv
                    - stage1.04.2017.csv
                    - stage1.05.2017.csv
                    - stage1.06.2017.csv
                    - stage1.07.2017.csv
                    - stage1.08.2017.csv
                    - stage1.09.2017.csv

        - stage1.10.2017.csv
        - stage1.11.2017.csv
        - stage1.12.2017.csv
    - 2016
        - etc.
    - 2015
    - 2014
    - 2013
- Stage 2
    - 2018
        - stage2.01.2018.csv
        - stage2.02.2018.csv
        - stage2.03.2018.csv
    - 2017
        - stage2.01.2017.csv
        - stage2.02.2017.csv
        - stage2.03.2017.csv
        - stage2.04.2017.csv
        - stage2.05.2017.csv
        - stage2.06.2017.csv
        - stage2.07.2017.csv
        - stage2.08.2017.csv
        - stage2.09.2017.csv
        - stage2.10.2017.csv
        - stage2.11.2017.csv
        - stage2.12.2017.csv
    - 2016
        - etc.
    - 2015
    - 2014

        - 2013
    - Personal Income
        - Research Data
            - SAINC50_ALL_AREAS_1948_2018.csv
    - State Populations
        - state_populations.csv

For question 1, run question_1.py. For question_2-3.py, to run only on 2013 data, uncomment lines 53 and 53, and comment out line 50.

For question_4-5.py, to filter out Washington D.C., uncomment line 225. To filter out areas with population over 1,500,000, uncomment line 238.

To produce plots that aggregate based on ratios to area, change the 'Population' in lines 227, 229, 231 to 'Area' and comment out line 224, as this will make the plots visible.

After these changes, run question_4-5.py, and the appropriate r coefficients, plots, and machine learning data will appear in the project directory.

Work Plan Evaluation

1. Loading and Parsing the data **5 hours**
    - Create a script that would retrieve the data, parse it, and convert it to a format that would be useful for the analysis.
    - This means having various formats for the gun database, as we needed it in DataFrame and GeoDataFrame format at different times. In some cases, we also needed it broken up into multiple DataFrames by month.
    - The rest are returned in their 'native' format, either DataFrame or GeoDataFrame. In some cases, this will mean joining it with the FIPS code dataset for readability's sake.
2. Data Manipulation and Calculated Columns **8 hours**
    a. Question 1: create a new DataFrame with rows analogous to months after 2013 by aggregating the amount of gun violence by month, as each .csv in the original database corresponds to one month.
    b. Question 2: Join the gun database with the US map and count casualties by county, then take the casualties column so that it can be plotted on a histogram. If necessary, do statistical analysis based on the histogram.

    c. Question 3: Join the overall gun database with the US map and aggregate by county, so that each county has a column that specifies how much gun violence took place in it. This will be done twice, once by absolute counts of incidents and once by the number of people hurt or killed.
    d. Question 4: Join the gun database, personal income dataset, gun licensee dataset, areas of the states from the US map dataset, and state population dataset into one table, aggregated by state. Drop any unnecessary columns and calculate ratio columns based on the area and population.
    e. Question 5: No extra work. Integrated into question 4.
3. Plotting **2 hours**
    a. Question 1: plot data over time for both number of casualties and number of incidents, include residual plots to highlight patterns.
    b. Question 2: plot the casualty counts by each county on a histogram. Perform further analysis as necessary.
    c. Plot the top 10% of counties on a map of the U.S. in yellow, 5% in orange, and 2.5% in red to show the most dangerous counties in the U.S.
    d. Plot gun violence over each of three predictors with residual plots: gun licensee density, population density, personal income per capita. Do once with density as a function of area and once as a function of population. Then train ML models on combinations of the data and compare Mean Squared Error for each.
    e. Question 5: Examine the gun licensee plots from question 4.
4. Report Writing **5 hours**
    a. Perform analysis as necessary, trying to find a pattern for the violence over time.

We were reasonably accurate in our work plan evaluations; the data manipulations and calculated columns took the most amount of time due to the datasets not having all the information that we needed. We also knew that this would require extra attention and care as this would determine the quality of the plots and analysis. Plotting ended up taking a bit more time than expected as they proved difficult to format and add legends to.

We ended up having to skip over the case study of Washington due to the formatting of the data. Different datasets had different names for everything, and we didn’t have enough understanding of the FIPS codes to join them properly. We also found data on individual U.S. counties, so we could do more in-depth analysis with that.

Testing

In general, we tested our code using a debugger and common sense. For question_2-3.py, we used only one file from the gun database, the one from 2013, and only plotted the top ten, five, and three counties for gun violence, counting them on the screen to make sure the right number of counties were appearing. We then compared the small number of counties that were showing up to what we already knew about what areas were particularly bad in the U.S., and we found the areas in our test plots to match nicely. We didn't know how to test out the regressions, so the only way we could was to try and use common sense to see if it made sense. Since much of this is unknown, we could only hope to capture a general trend, rather than specific numbers. Everything we saw made sense, so we passed it. Also, we proved the same thing multiple different ways, and all the results point in the same direction: gun crime per unit of area increases with population density.

It is important to recognize that this data proves nothing about the causes of gun violence. We examined some predictors, and it appears that gun ownership is independent of gun violence, but with a study like this, we can prove nothing.

Collaboration

The work was done by Alex Knowlton and Cody Murphy. Some clarifying questions were answered by course staff about the more complex aspects of matplotlib and seaborn, and there was consultation of tutorials and message boards concerning the various software, but no outside help was given with the data analysis.

Further Questions

Some digging showed that the activity of drug cartels was very similar to our map of the most violent areas, with heavy concentration in southern California, the coast of the Pacific Northwest, and much of the northeastern United States. This indicates that a possible correlation could exist between drug trade/use and gun violence, but we could not find data, so it was outside the scope of this project.

Another interesting study might be to examine how gun violence and political governance are related. Gun policy is very different across the aisle, so it might be interesting to see how policy, governance, and gun crime all fit together.

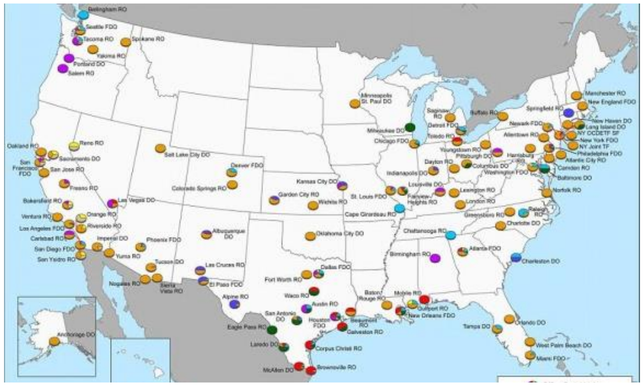

Figure 13: Major Drug Areas, source: https://geo-mexico.com/?tag=drugs