Gun Violence Data- An Analytical Approach

Team RNRS

Riley Marfin, Nick Allen, Roke Mendiola, Stefen Ramirez

Abstract

In this project, our team researched, processed, and analyzed the Gun Violence Dataset from the Gun Violence Archive (GVA) and correlated the results against a city income dataset. The gun violence dataset contains detailed information such as location, injuries/fatalities, and other statistical information, which we used to calculate and plot gun incidents by city, state, gender, and age. We also researched the number of gun laws per state and compared it to the number of incidents per state using a scatter plot, and created a linear regression based on our findings in order to forecast future incidents. Furthermore, we used the number of incidents per city, the total population of each city, and the central limit theorem to determine if a specific city had an abnormally high amount of fatalities. Finally, we used the location data to plot a density map of incidents across the United States, which gave us a valuable bird's eye view and revealed the scope of our dataset. Overall, we discovered many interesting characteristics about the data, some obvious and others not so obvious, and gained valuable insight into the practice and implementation of essential data science concepts.

1 Introduction

From the inception of this project, our team analyzed both the Gun Violence Archive (GVA) dataset and a city income dataset. The GVA dataset contains detailed records of gun incidents including, but not limited to, state, city/county, total fatalities, total injuries, participant ages, and gender. The city income dataset contains records of cities and their population, average income, age, wage, etc. Our team focused on finding interesting correlations with income, gender, gun laws, age, and population in comparison to their respective number of gun incidents. The goal of our project was to visually represent our results through histograms and scatter plots to provide a clear understanding of our discoveries.

2 Obtaining Dataset

The gun violence dataset, date ranging from 2013-2017, was found on a github repository from user jamesqo[1]. The dataset lacked some national news involving mass shootings that involved an exorbitant amount of victims, i.e. 10+ victims. This helped with reducing outliers that could have skewed the dataset further. The dataset correlated against the income database, data from the 2010 census, which was built from gathering info from the website datausa.io[2]. From their website we used their API to build the income dataset from the geocodes that we found per city. From there we filled NA values by averaging close city and towns. We found their distance by the latitude and longitude provided in the gun violence dataset. Then from the distance we were able to use the euclidean distance to average the areas that were within 1 degree of distance. After initial cleaning the two datasets were easily merged by the city and state columns.

3 Preprocessing

The most pertinent items we needed to address were handling missing values in the Gun Violence Dataset and the city income dataset. This proved to be more difficult than originally planned due to the location based nature of our dataset since data could be skewed. With missing values in the participant_age column, we took the closest x rows based on Euclidean distance from the latitude and longitude, found the mean, and used that to fill the current missing value. We then filled in the missing values of participant_age_group according to the values in the participant_age column. The missing values of the participant_gender column were a bit trickier since the values only consist of "Male" or "Female". In order to account for this binary type column, we decided to replace the missing value with the mode (most common gender) of the corresponding cities/counties. Lastly, we replaced the missing values of participant_status and participant_type with "Unknown" because there was we found no valid way to properly predict/forecast these data types. We believe this will help prevent skewing the data before analyzing it. From the Income dataset, we alleviated the city income using a similar approach as the participant_age column- we used the Euclidean distance from the latitudes and longitudes to find local cities with similar populations and find their income, without adding more noise to the data

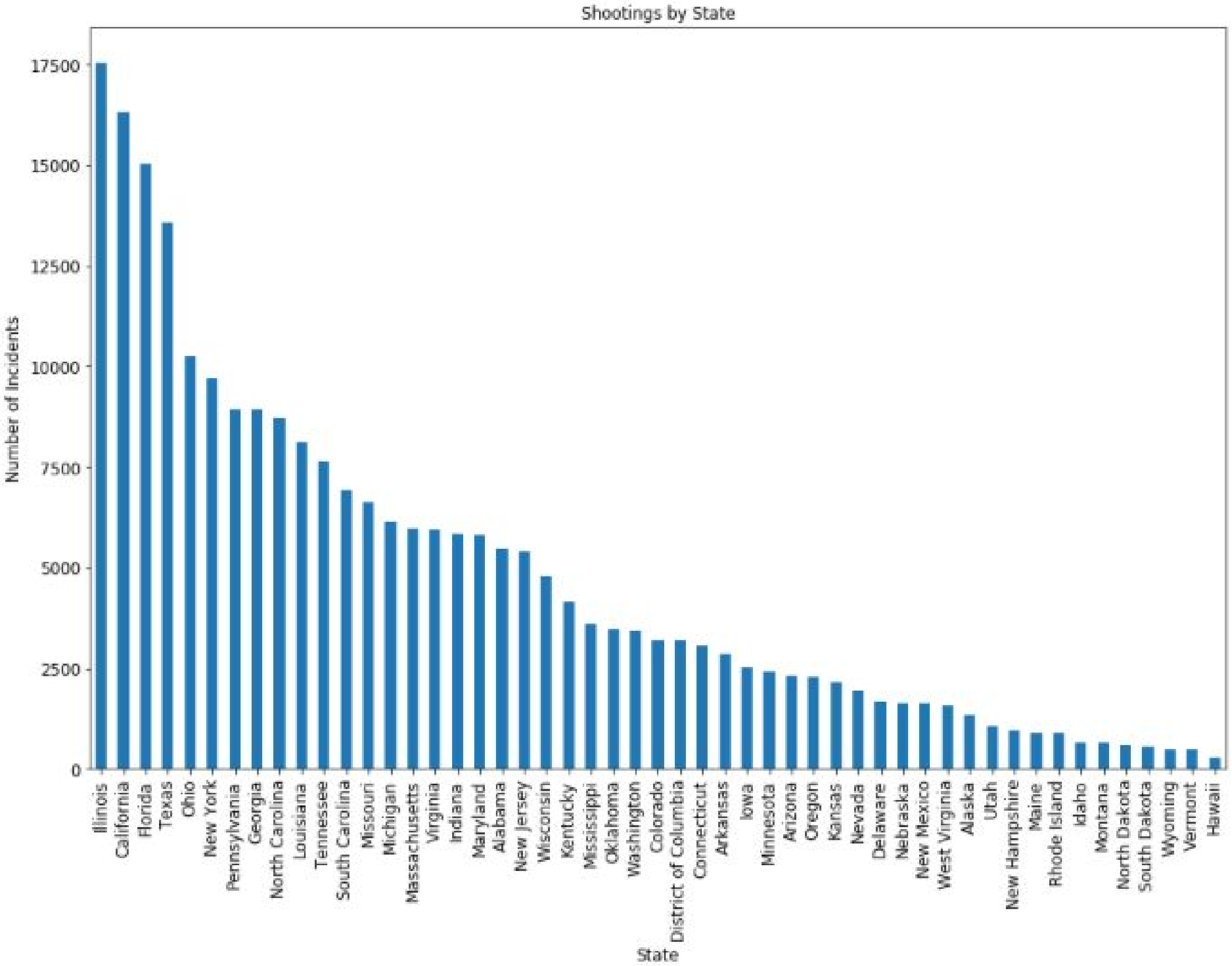
4 Analysis

Over the course of our analysis, we developed and tested python code in order to discover as much as we could about our data. After examining the dataset in its entirety, we were curious to see if a correlation existed between the data itself and numerous social and economical factors. Specifically, we compared our data against gun laws throughout the United States, gender, location, population, and income, to name a few. We encountered many interesting discoveries when we dissected our data, discovered interesting trends, and tested numerous hypotheses that we developed prior to our processing. Some of the most notable findings are described in detail below.

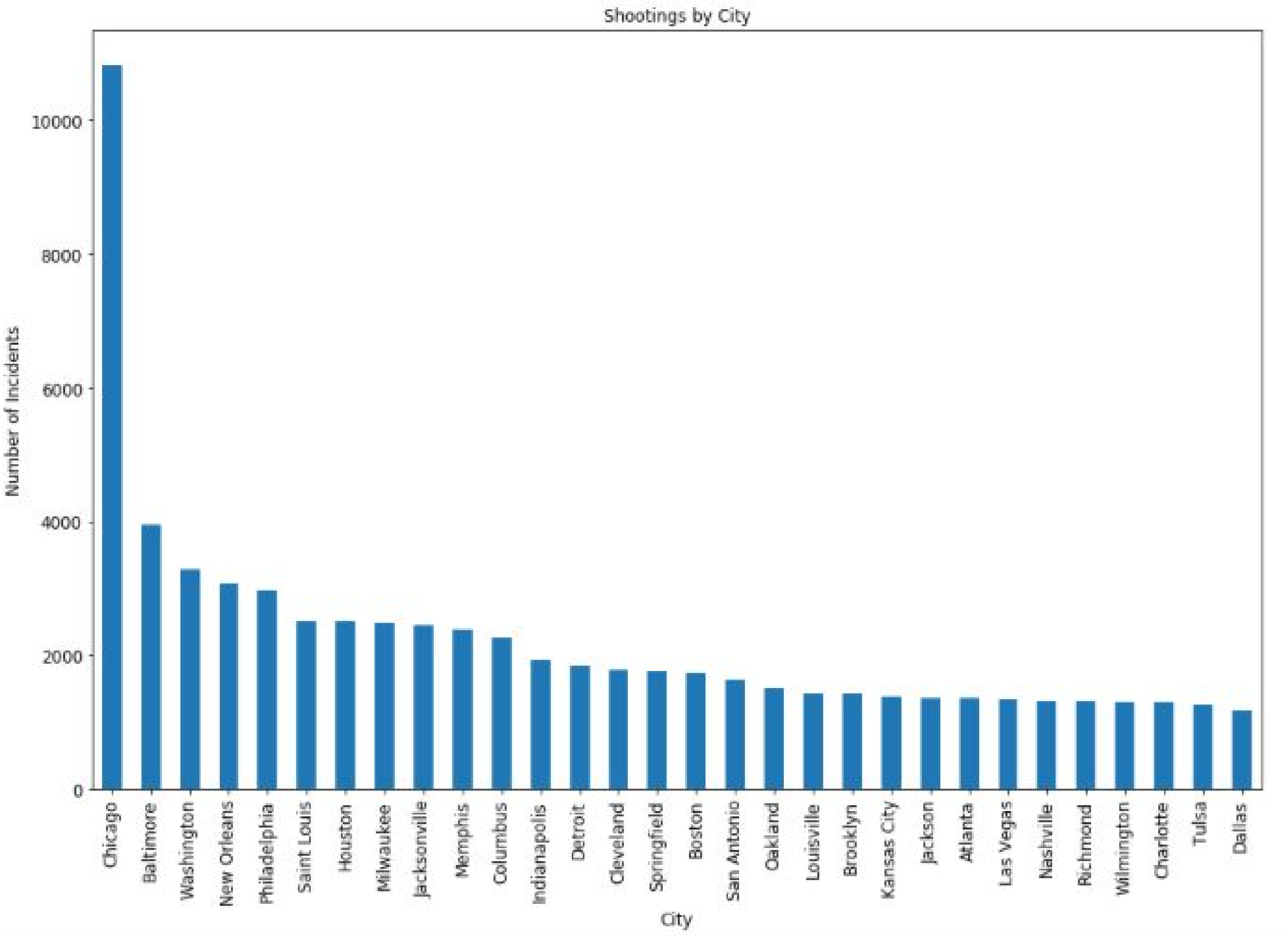
There are a few major functions we used to analyze our data. These functions are located in our python notebook and process different aspects of the data in different ways. After our preprocessing functions (described in the preprocessing section), we began our analysis by calculating basic statistical information on the different categories or columns of data, such as mode and average. We created a function called caclulateModes(df, columnList) which takes in the main dataframe and a list of one or more columns and returns numerical information about these columns. This functions displays both individual column information and grouped column information using the groupby function. Next, we used the value_counts() method to obtain the total incidents for each state from the ['state'] attribute in our DataFrame, and plotted this data in the bar graph using pyplot from the matplotlib library. The graph entitled “Shootings by State” below displays our results.

We found that Illinois held a little over 17,500 gun incidents, the most out of all the states we observed. Our top four states, including Texas, reported over 12,500 gun incidents, while the other states reported 10,000 or less gun incidents. After the top four states, we noticed a significant drop of gun incident reportings per state.

A very interesting thing we found was that Chicago was accountable for over 10,000 gun incidents, more than half of the total incidents in Illinois. This was abnormally high compared to any of the top 30 cities we analyzed with the most gun incidents. San Antonio ranked 17th on our list with a little under 2,000 gun reports.

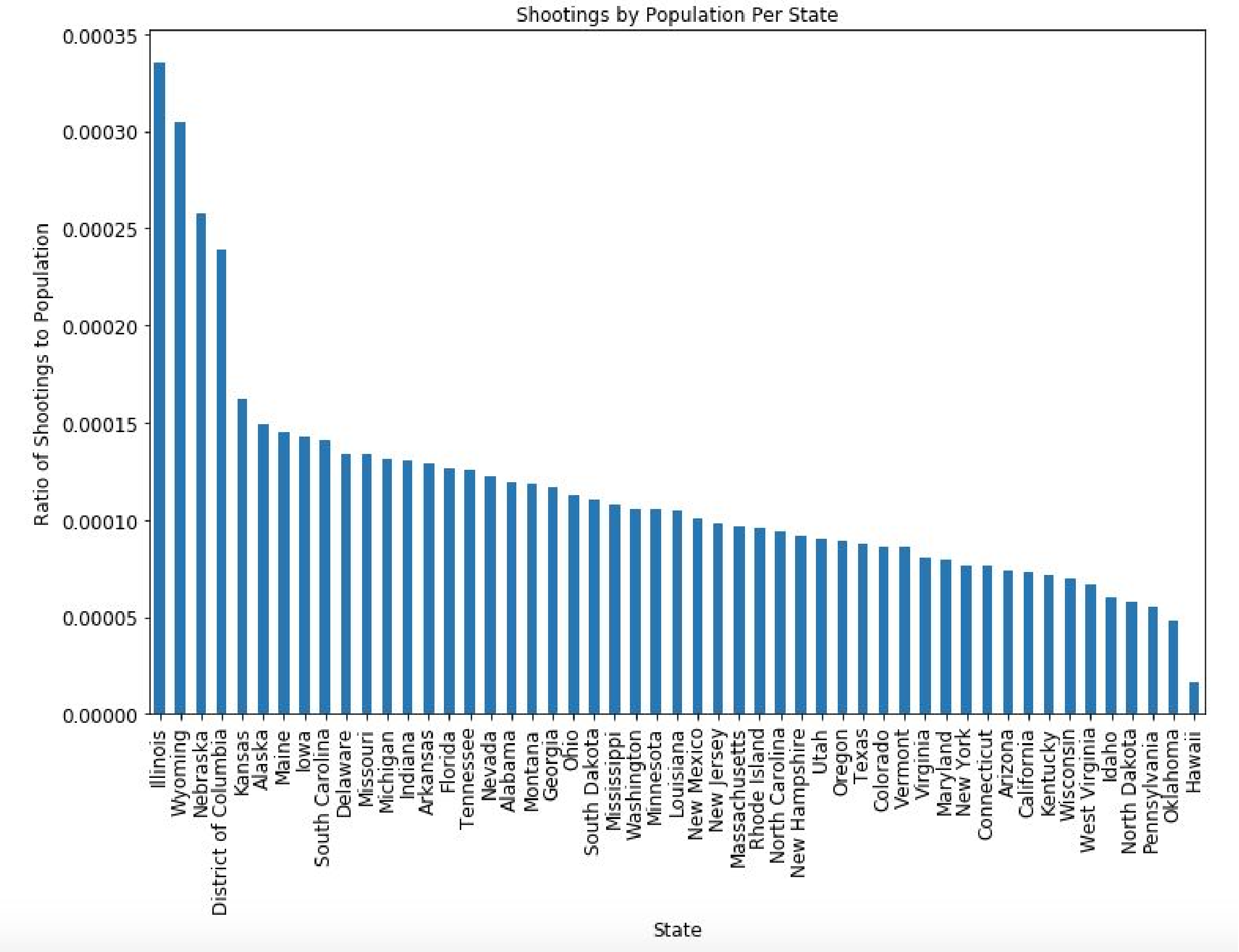


Similarly, we used value_counts() on the [‘city_or_county’] attribute in our DataFrame to get the total incidents per city and graphed the top 30 cities using .nlargest(). The reason for taking this approach was due to the fact that there were too many cities to meaningfully graph in one visual, and many of these cities had extremely low rates relative to the cities with the most shootings. So, we decided to go with the top 30 cities because these are the locations used in many other aspects of our analysis. An easy manipulation of the parameter for the nlargest() function will convey more information if desired. The bar graph below entitled “Shootings by City” depicts our results after applying this function.



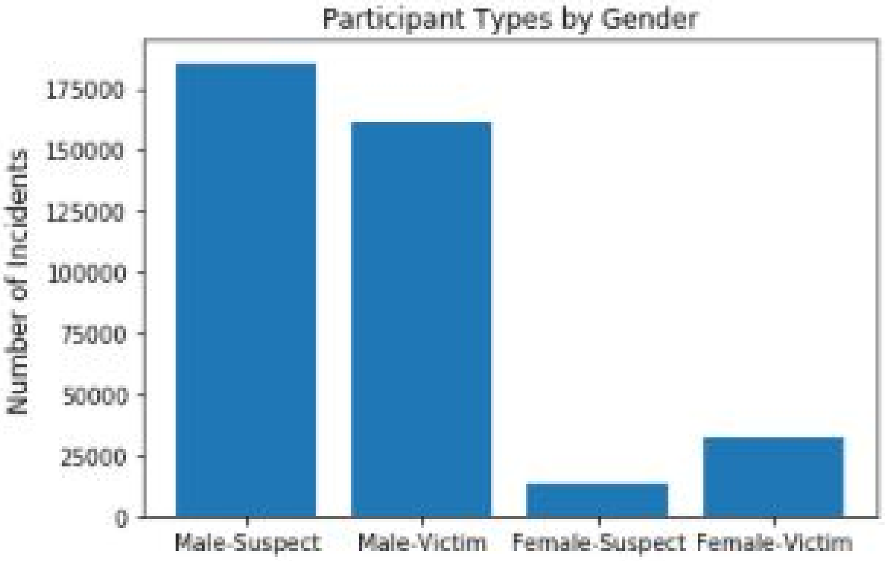
Based on the information above, we can see that Chicago had the most shootings in the time frame of which the data was collected. Followed by Baltimore, Washington, and so on. This follows a historical trend which has been observed in the past. Also, we can see that Houston, Dallas, and San Antonio both appear in the top 30 cities in the US with the most shootings. San Antonio ranking above Dallas by a significant amount of places while Houston ranks 7th in our list.

After graphing the number of instances we wanted to take a look at the data from another perspective. To accomplish this we created a ratio of the number of shootings to population to see which states had a large number of shootings per capita. This is seen in the python notebook as the createStateAveragesDict(df) function. This function will take the number of shootings for a city and divide it by the total population. Then for each state, it will take each ratio for each city within the state and average it out to create the states ratio. The resulting dictionary is then converted to a dataframe and is depicted in the graph below titled “Shootings by Population Per State.” As you can see, Illinois remains number one, but the runner ups vary greatly as they are no longer California and Florida. Instead, the runner ups become Wyoming, Nebraska, and DC.

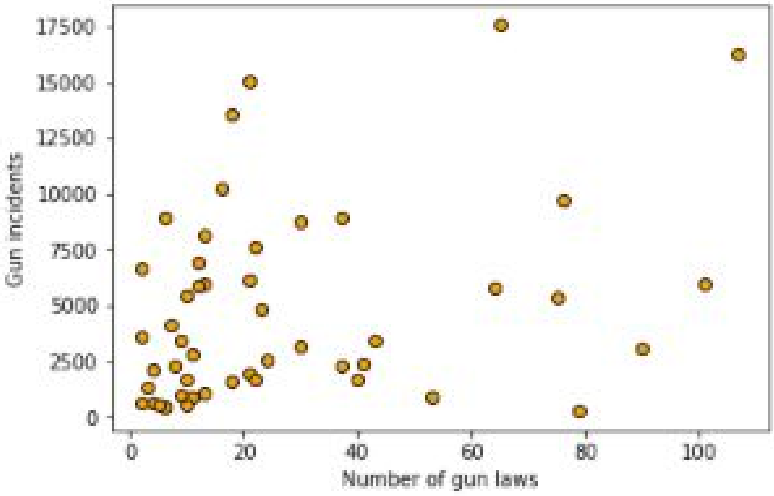


Another discovery was the amount of suspects and victims reported for each gender. Records that did not specify a participant’s type and whether they were a victim or a suspect in a gun incident were ignored to prevent skewing of the data. We found that a large portion of our dataset were reported “Male.” Male suspects in particular are reported the most, shortly being trailed by male victims. As for females, we found the complete opposite in their reports. There were more reports of female victims in our dataset than female suspects. There was also tremendously less reports of females in our dataset than males. One thing to note, incidents with reportings of suicides labeled the participant as both a suspect and victim. This however did not skew our data.

We obtained the total incidents for each participant type by gender using a dictionary. We used a for loop of complexity O(n) to traverse each record in our dataset to find the participant's gender and whether they were a suspect or victim. After collecting this information, it would increment its corresponding entry in our dictionary which we used to create our graph below, titled "Participant Types by Gender." We used the dictionary's keys to represent the x-axis and their values to represent the y-axis.

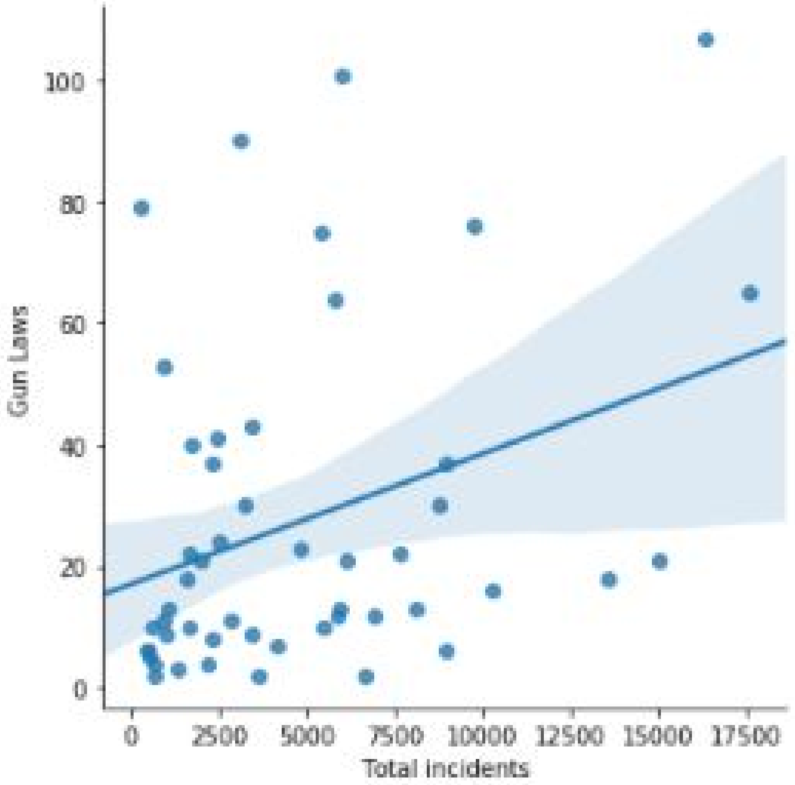


We also graphed a scatter plot (see right image) comparing the number of gun incidents in a state vs. its respective number of gun laws[4]. We found a large cluster of states reported both a low number of gun incidents and a low number of gun laws. The states that had more than 50 gun laws seemed to be spread out more on our graph. We initially thought states with lower gun laws would report a higher number of gun incidents while states with higher gun laws would report a lower number of gun incidents. We believe states with low gun incidents did not need to enforce additional gun laws, which is why a lot of our data points is clustered in the bottom left corner of our graph. This is an interesting point to observe because there is much political controversy today regarding whether there should be stricter gun laws throughout the country. Perhaps this is an indication that there is simply no correlation



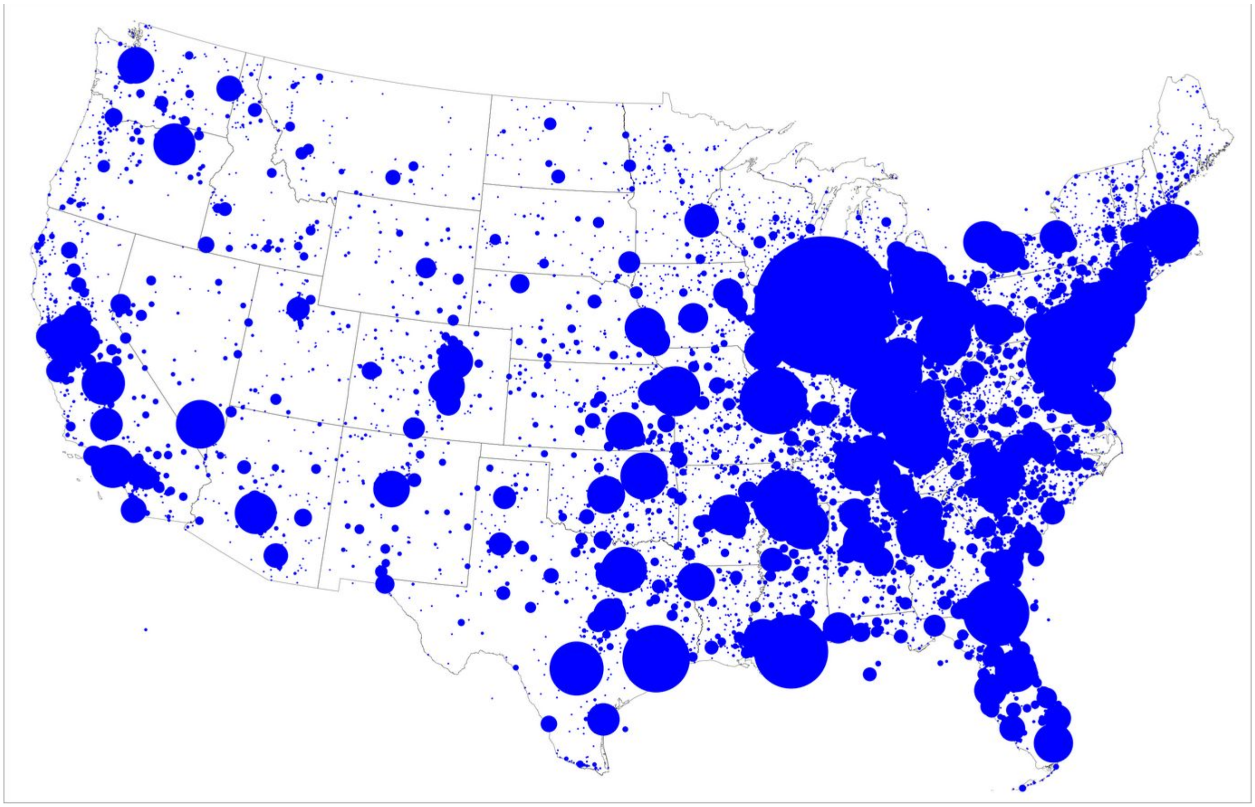
between gun laws and amount of shootings, or perhaps the data needs further examination and wider scope in order to truly discover some sort of meaningful pattern here. What we would have liked to accomplish with more time was to compare states' gun incidents with their respective number of gun laws per year. We would be able to see how the graph changed over time as gun laws were added and how that affected the state's gun reportings, and this would have been able to provide more context to our scatterplot, which seems to have an unpredictable correlation. Also, we would have liked to have our data points labeled with their respective state's abbreviations to provide a clearer understanding of our graph.

Next, we created a new DataFrame using the states as the index and created columns for their gun laws and how many incidents in our original DataFrame included their state. After gathering the proper values, we simply used pyplot.scatter() to plot the data. From there, we applied a Linear Model to the same graph with the axes swapped to see if we could find any interesting results using seaborn and .lmplot() function. The first thing we noticed was a majority of our data points were outside the confidence interval. Because of this we did not believe the quality of the graph was sufficient.

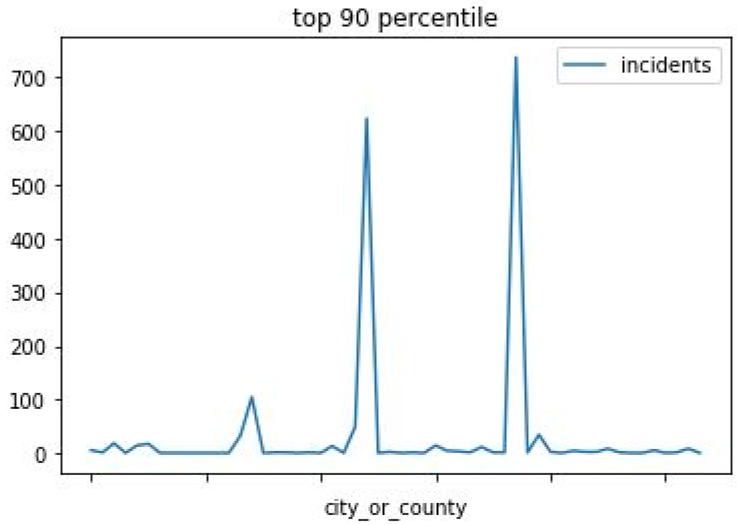


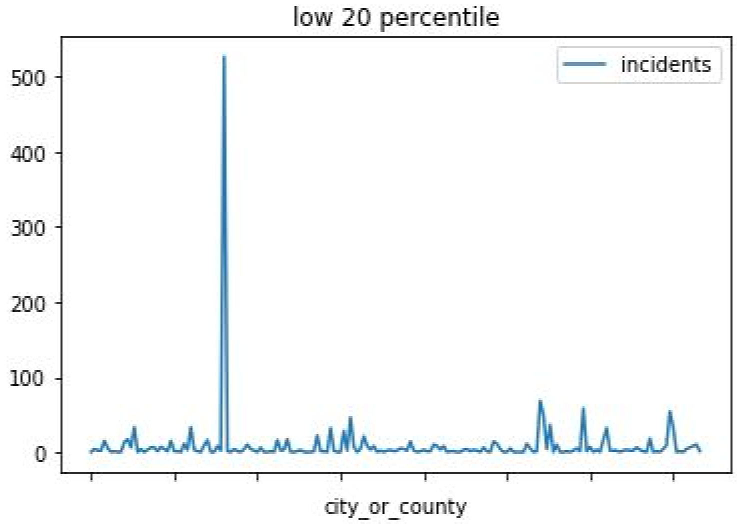
Another subject we wanted to analyze was the resulting fatalities of the gun violence. We created a function that used the central limit theorem to determine if a certain city had an abnormal amount of fatalities in comparison to the states average. In the python notebook the createStateDict(df) function prepares two dictionaries: the averageStateFatalities and the stateDict. The stateDict dictionary has the average fatalities for each city. The averageStateFatalities dictionary has average fatalities for each city within each state and averages it out to create the states ratio. Then there is the determineAbnormalAmountOfShootings(avgStateFatalities, stateDict, state, city) which utilizes the dictionaries produced from the previous function, a state, and a city. This function uses the central limit theorem to produce a value between 0 and 1. The closer the value is to 0 means that city has less fatalities compared to the average of the state and the closer you are to 1 that city has more fatalities than that of the rest of the state. For instance, here are some results of the function: Los Angeles CA provides a value of 1, Houston TX provides a value of 0.807, San Antonio TX provides a value of 0.0018, and Chicago IL provides a value of nearly 0.

To explore the visual analysis further we took to mapping each of the incidents into the location on the map of United States. We can immediately see that more generally populated areas of America give much more incidents. This is evident by the large blue circles on the map, i.e. locations like Detroit, Los Angeles, and Chicago. But the even spread shows closer analysis that this is not affected by the income of an area, but by population. This is obvious that a higher population will increase the likelihood of an incident occurring in that city or town.



To break down further analysis we took compared the lower percentile to the high percentile to see if there was a difference. We managed by a function called plot_income_percentile which mapped the percentile to the amount of gun incidents that occured for that city. From there we could do a visual analysis of the data.
Unfortunately the hypothesis that income could influence the gun incident was very unlikely show by the data. We saw very little change by income but by the visual analysis of the map. There was a much stronger correlation of population to gun incidents to the dataset's findings.





6 Summary and Comments

Our group will now share some of our experiences, team organization and work-load distributions.

6.1 Riley Marfin

Over the course of this project I learned a great deal from the dataset and the results produced from our analysis. Part of my job was creating some of the visuals and calculating various column statistics and functions which processed one or multiple columns of data at a time. I managed to discover correlations and relationships between the data and the current state of the United States today, which not only solidified my data science skills, but also gave me important life knowledge which will carry me far as a person. Because of this project, I have a deeper insight into the issues America is facing today, and I have factual evidence to back up my claims. This project has definitely spiked my interest in data science, and I enjoyed working with my group to accomplish our goals.

6.2 Nick Allen

Throughout the project I had the opportunity to learn how to visualize data with python and manipulate datasets to provide informative insight. Gun violence is a prevalent topic in today's society and I found it interesting to see the different types of analysis we could perform to validate or invalidate our initial thoughts. I helped with preprocessing some of the data and provided some functions to perform analysis that could be converted to graphs. This proved challenging, but also entertaining as I enjoyed experimenting with the data and the project forced me to really dive into and utilize the different functions learned in class.

6.3 Roke Mendiola

My experience with this project was truly insightful. With shootings and gun incidents being prevalent throughout the recent years, working with a dataset involving gun violence was interesting. I learned a lot more about python, specifically working with DataFrames and plotting attributes. I was satisfied with how our group coordinated the work properly and efficiently. I worked on collecting the suspects and victims for each gender and plotting a histogram to represent them. I also gathered information for each state's gun laws and plotted the comparison of their gun laws with their respective number of incidents. In addition, I worked on the Linear Regression Model to represent state gun laws in relation to their number of incidents, although the quality of the graph was not sufficient. Overall, I obtained substantial information regarding gun violence in the United States as well as data science concepts.

6.4 Stefen Ramirez

I quite enjoyed the analysis of some data and trying to pull some meaningful data out of the chosen datasets. Being able to pick and fully analyze a dataset with new tools that we learned during the semester is rewarding in its own right. But throughout the project I quite enjoyed being to do visual analysis since this gives the more human way of looking at the data you choose to analyze. Also being able to build a dataset and compare against another dataset gave me a clearer insight into choosing dataset more carefully and seeing more meaningful data. Overall, I learned more about visual analysis and how to show correlation between multiple dimensions of separate datasets.

Resources

[1] Gun-Violence-data:
https://github.com/jamesqo/gun-violence-data/tree/17dc3079a18c872c684da2e1fe051cc41f054334

[2] Datausa.io:
https://datausa.io/

[3] Data Science Concepts:
https://utsa.blackboard.com/bbcswebdav/pid-3314162-dt-content-rid-50942503_1/courses/CS-3753-001-22270-201910/index.html

[4] Gun Laws per State:
https://statefirearmlaws.org