

Gun Violence Data- An Analytical Approach

Team RNRS

Riley Marfin, Nick Allen, Roke Mendiola, Stefen Ramirez

Introduction

- From the inception of this project, our team analyzed both the Gun Violence Archive (GVA) dataset and a city income dataset (2013-2017).
- The GVA dataset contains records such as 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.
- 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.

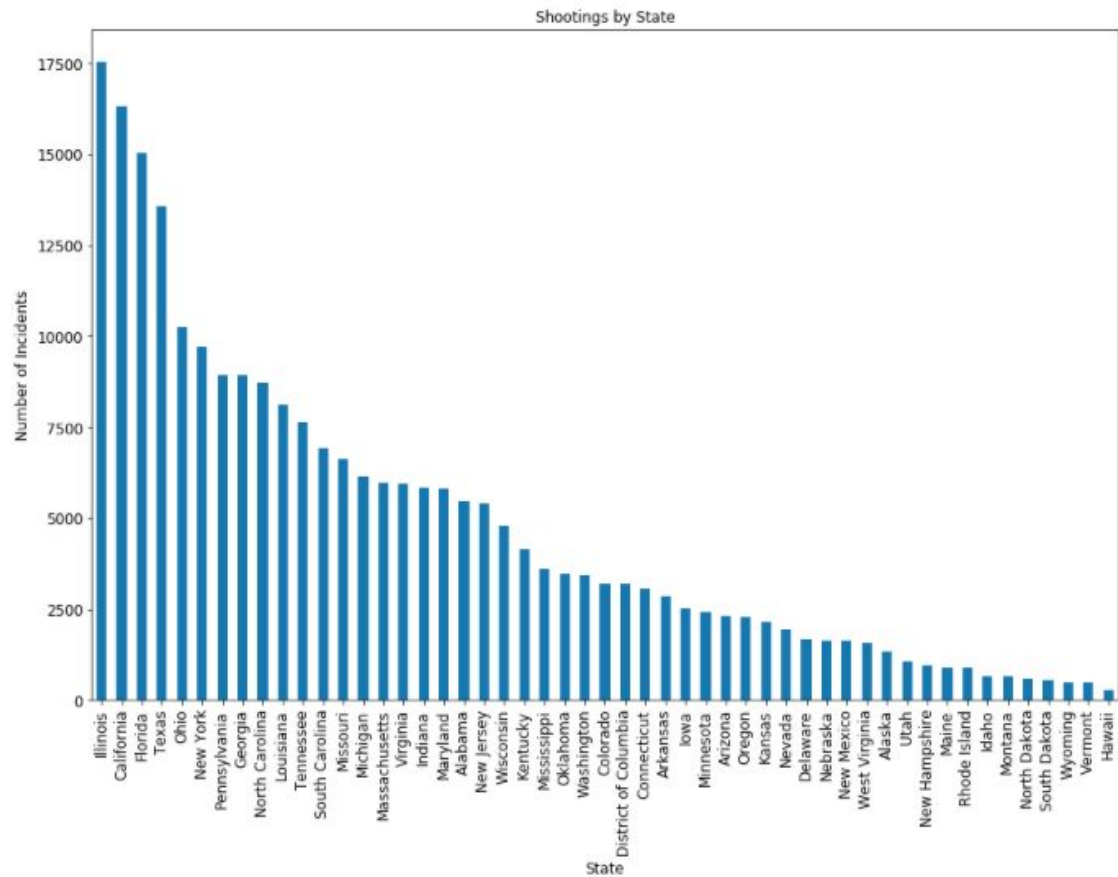
Obtaining Datasets

- Dataset found on a github repository from user jamesqo^[1].
- Lacked some national news involving mass shootings that involved an exorbitant amount of victims, i.e. 10+ victims (helped with outliers)
- Dataset correlated against the income database, data from the 2010 census, which was built from gathering info from the website datausa.io^[2].
- Used income API to build the income dataset from the geocodes that we found per city.
- After initial cleaning the two datasets were easily merged by the city and state columns.

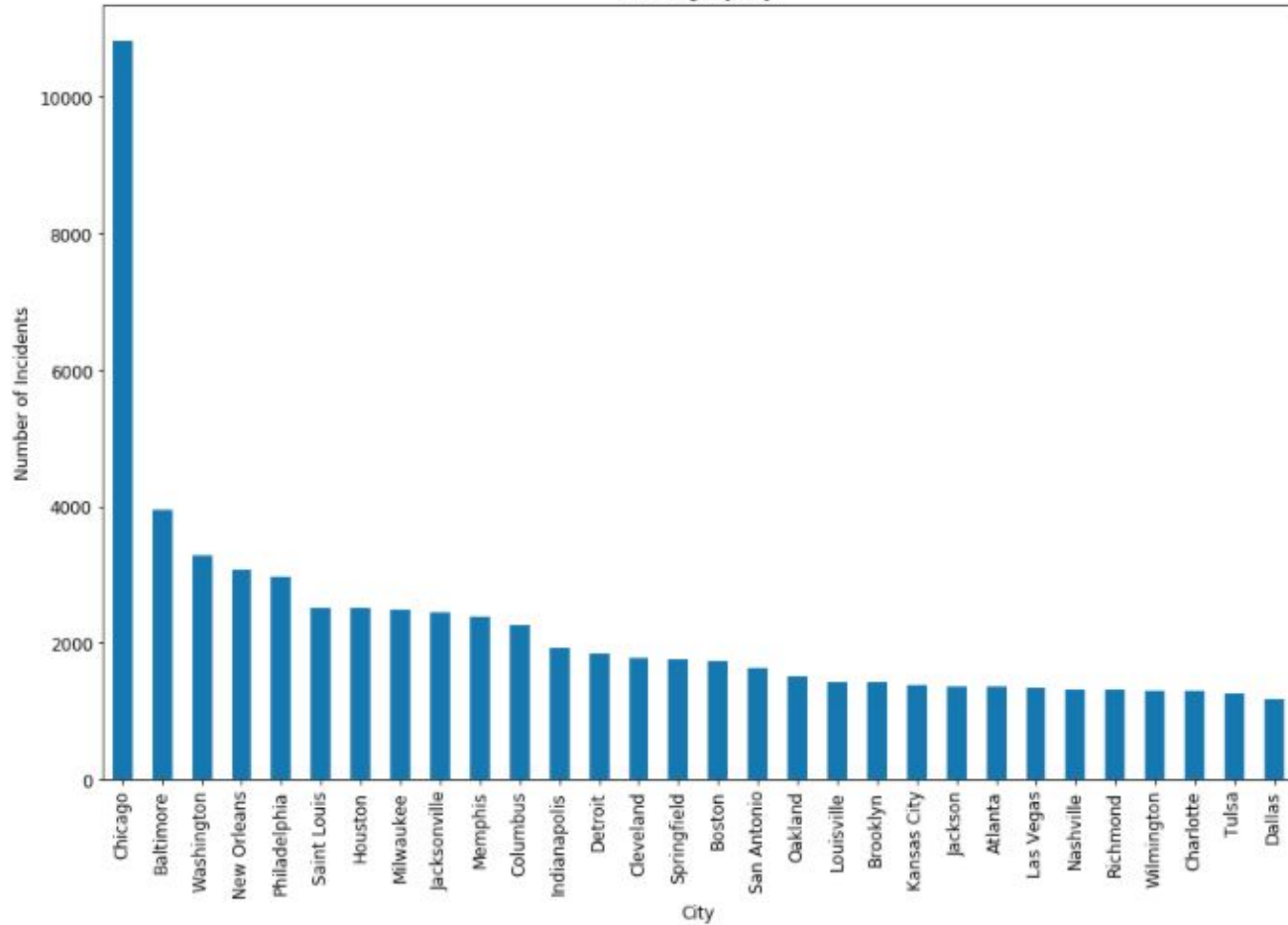
Preprocessing

- The most pertinent items we needed to address were handling missing values in the Gun Violence Dataset and the city income dataset.
- 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.
- The missing values of the participant_gender column were a bit trickier (binary column)
 - Replaced 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” (we found no valid way to properly predict/forecast these data types)
- From the Income dataset, 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

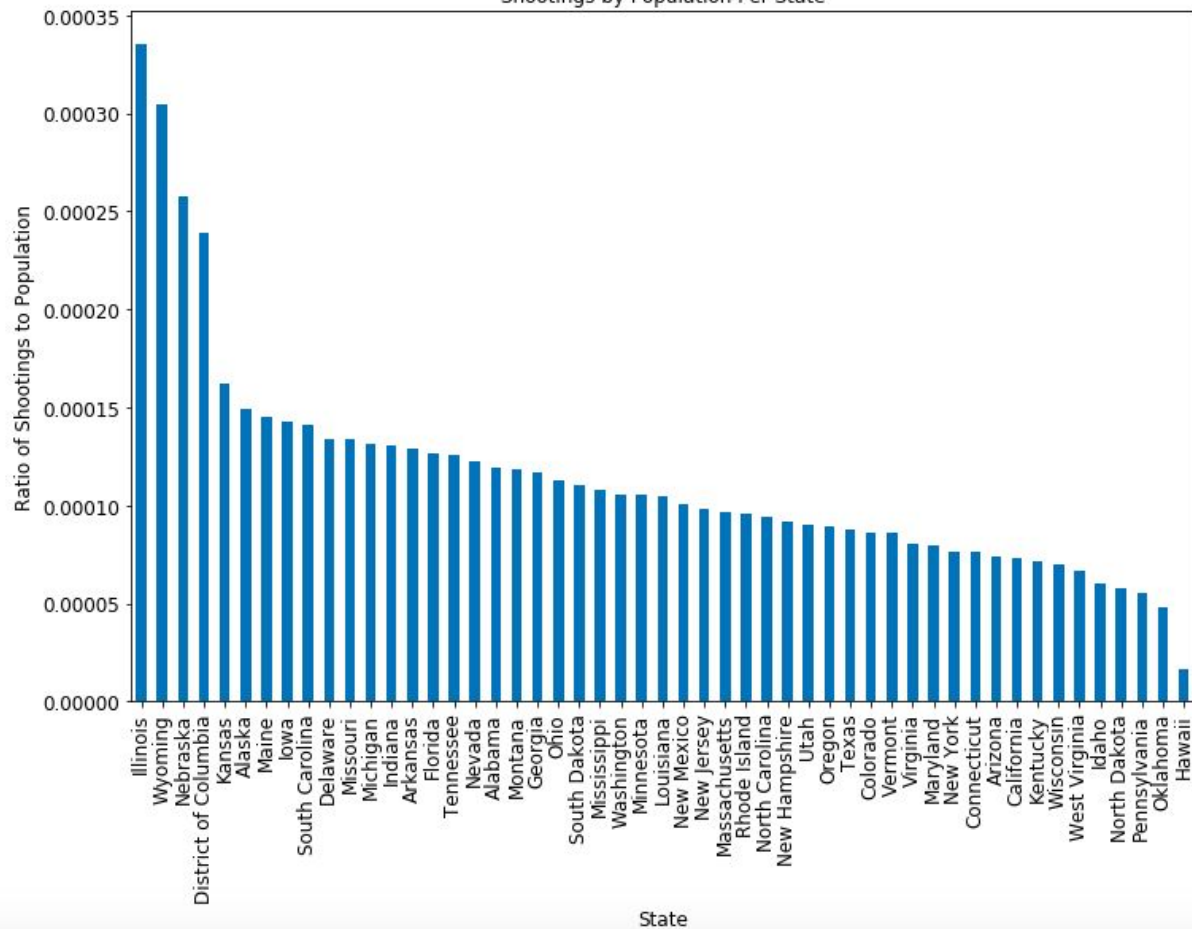
Shootings by State



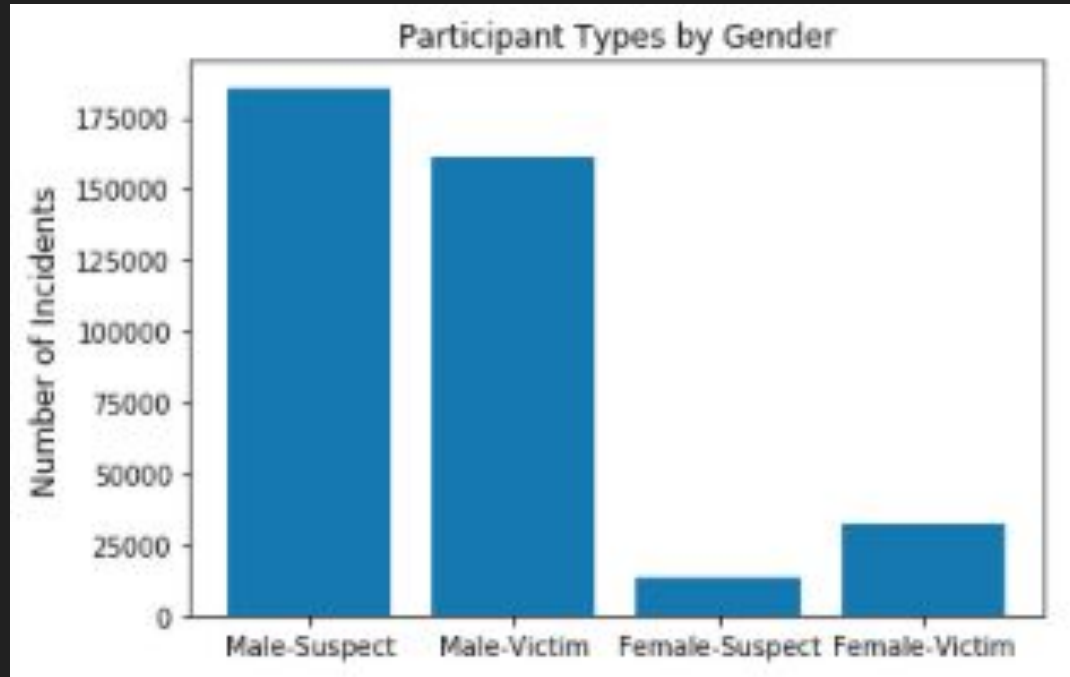
Shootings by City



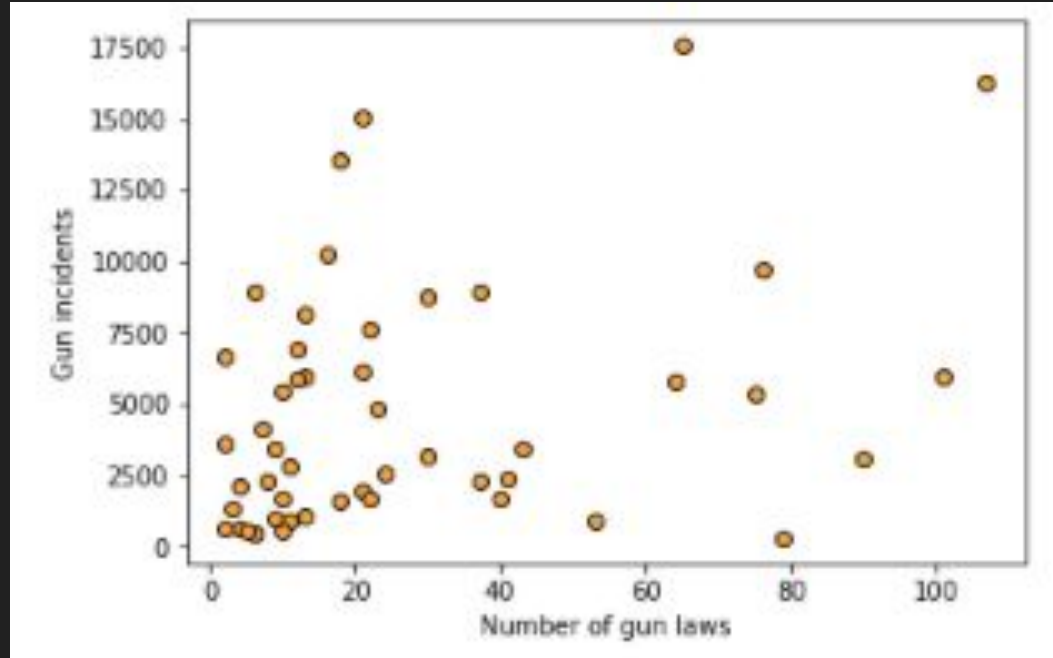
Shootings by Population Per State



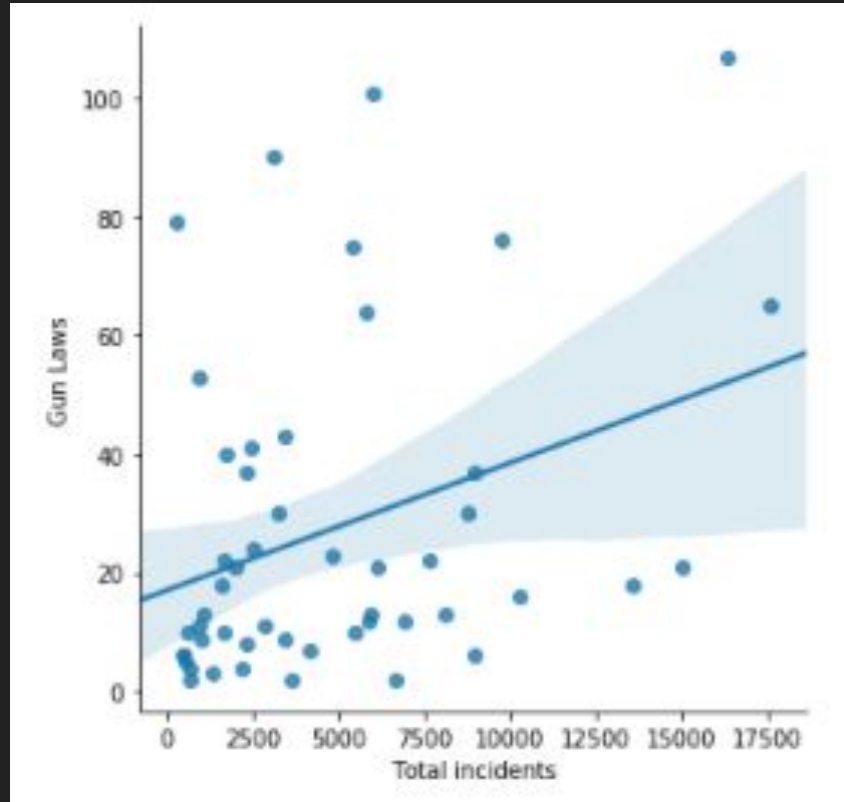
Suspects and Victims by Gender



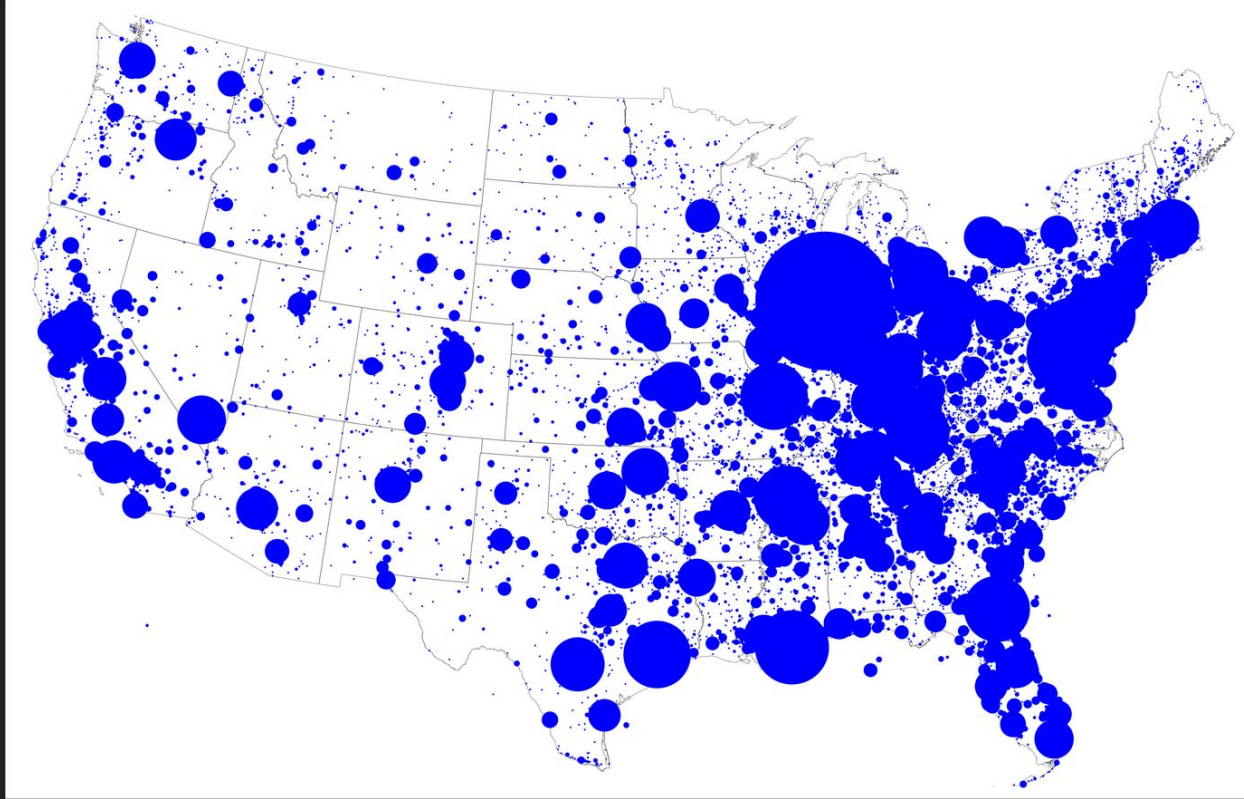
Gun Laws vs Total Incidents Per State



Linear Regression Model



Density Map of Shootings Across the US



Conclusion

- Overall, found both obvious and not so obvious results
- One interesting find is the lack of correlation between gun laws and shootings per state
- The density map showed a large majority of shootings on the East/Northeast side of the United States
- Abnormal fatalities per city didn't necessarily correlate to shootings per city (ex- Chicago vs Illinois)
- Acquired important data science concepts which will aid us in future endeavors