https://github.com/rudrakshsimlote/gun-violence-dmg

Sandhya Ganesh sg1790

Aaryaa Shah ams1004

Rudraksh Simlote rs2096

The United States is infamous for the high frequency and severity of gun-related violent incidents. While this report cannot solve this problem, we can shed light on possible causes and make predictions that would aid the future prevention of such events. We compared legislation that protects against firearm violence by state with the number of shootings in each state. The first database we utilized has different legislations introduced by state to protect against firearm violence. Our second database has around 240,000 data points on what day and in which state a firearm incident occurred, along with the number of people injured and killed. Next, we have two data sets focused on mass shootings and shooters, which were analyzed separately to get a better understanding of what causes more severe incidents. The data on mass shootings contain information on the day, state, and number of victims and has information on the shooter’s age and gender. Our data on mass shooters goes into more depth on the past criminal record and mental state of the shooter. Finally, the last data set simply has the death rate by state caused by firearm incidents.

Gun violence is a critical public health and safety issue that impacts countless lives. This project is important because it looks at trends over time and explores the risk factors behind mass shootings and general gun violence. It also examines how legislation plays a role. By predicting which states are most at risk based on current and new laws, we can help inform prevention strategies. We're also particularly interested in seeing how the recent change in political leadership might impact gun violence trends in the country. This project has the potential to provide valuable insights for creating safer communities.

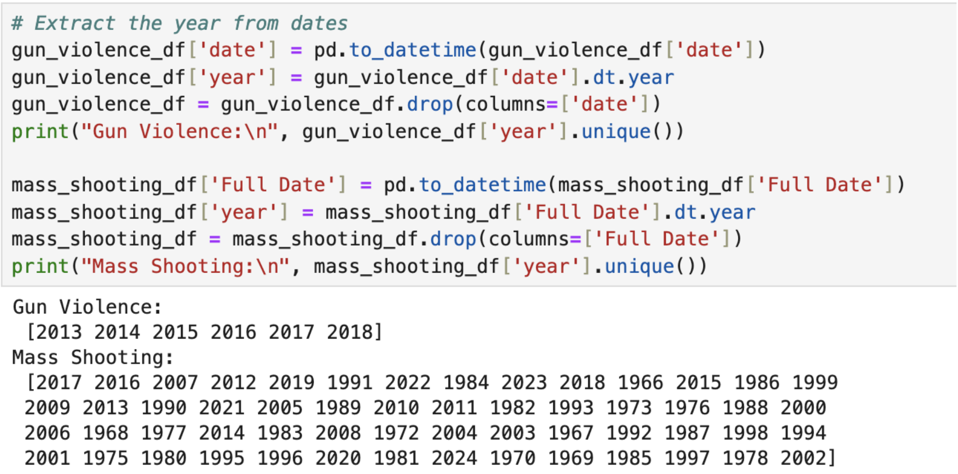
We are excited about this project because it focuses on an important issue that impacts lives nationwide. This semester has given us the opportunity to use the skills we obtained in Data Management for Data Science to discover insights that can shape existing policies or bring awareness to patterns that influence prevention that have not yet been recognized. We are also specifically motivated by the idea that our work could contribute to solutions that can make the communities we live in safer. It would be great to give our neighbors the right to live without worrying about being threatened with the harm of a firearm within country borders.

Most Americans all across the country acknowledge that the increasing number of shootings is a major concern, and therefore, it has been studied in the past. Similar projects that focus on different aspects of gun violence include studying the psychological impact of shootings on victims or identifying which locations have the highest crime rates to warn residents or tourists. Our project is concentrated on how the severity of legislation against firearm usage impacts the amount of shooting incidents in the corresponding state. As 2024 was an election year, we are particularly interested in finding out which policies, if any, are effective in bringing down the number of shootings and what the scale of impact for each one is.

There are many publicly available datasets with information on gun-related deaths. Different datasets record varied information on incidents across the country, including the location, number of victims, and the shooter's background. We focused on finding datasets that deal with location and legislation. We did not create any sample data points, but did generate sample response values based on explanatory variables. For example, using the model to predict the gun-related death rate given a location point or identifying the state with the highest likelihood of a mass shooting. Though morbid, the use of our test data can inform states on how to improve their policies to prevent shootings. For this project, we used one-hot encoding for the categorical data and created a new field summarizing the "strictness" of legislation. Additionally, we built a prediction model using various predictors and analyzed the severity of the single predictor legislation. We also made graphs to better visualize and understand the data.

We searched for data frames from official government sites to obtain accuracy. Some existing issues in current data management practices include inconsistent or incomplete data. Another issue is duplicate records in the database, which leads to inefficiencies. The data also includes the characteristics of the shooters, which is subjective and could be biased. To protect against the weaknesses of the data, we cleaned the table using different methodologies to fill or delete null data and replace outliers as we saw fit. The data sets we found came from the Centers for Disease Control (CDC), Gun Violence Archive, and Department of Justice. After researching, we found that all the organizations were trustworthy and provided us with credible data.

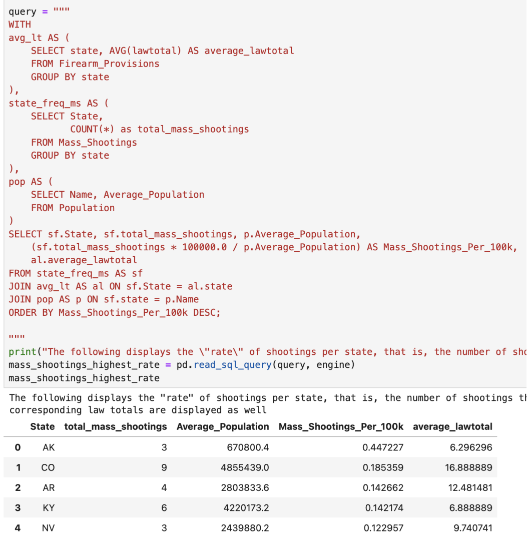
For the data cleaning portion of the project, we started by checking for null and values and found that there were none. To make the data uniform, we decided to change all of the state names to the same abbreviations for the Gun Violence and Provisions dataframe and the 'District of Columbia' was the only value that had to be changed in the Mortality dataset. We also only kept the year in which an incident occurred to study the impact of policies over the decades which meant extracting the year from dates or deleting columns with the month and day. Next we checked for bad values and found that the numeric data in the Mortality dataframe was not stored at integers because of the commas for the thousand separators. The process of changing the data to numeric included the removal of commas from numbers.



```
# Extract the year from dates
gun_violence_df['date'] = pd.to_datetime(gun_violence_df['date'])
gun_violence_df['year'] = gun_violence_df['date'].dt.year
gun_violence_df = gun_violence_df.drop(columns=['date'])
print("Gun Violence:\n", gun_violence_df['year'].unique())

mass_shooting_df['Full Date'] = pd.to_datetime(mass_shooting_df['Full Date'])
mass_shooting_df['year'] = mass_shooting_df['Full Date'].dt.year
mass_shooting_df = mass_shooting_df.drop(columns=['Full Date'])
print("Mass Shooting:\n", mass_shooting_df['year'].unique())
```

```
Gun Violence:
 [2013 2014 2015 2016 2017 2018]
Mass Shooting:
 [2017 2016 2007 2012 2019 1991 2022 1984 2023 2018 1966 2015 1986 1999
 2009 2013 1990 2021 2005 1989 2010 2011 1982 1993 1973 1976 1988 2000
 2006 1968 1977 2014 1983 2008 1972 2004 2003 1967 1992 1987 1998 1994
 2001 1975 1980 1995 1996 2020 1981 2024 1970 1969 1985 1997 1978 2002]
```

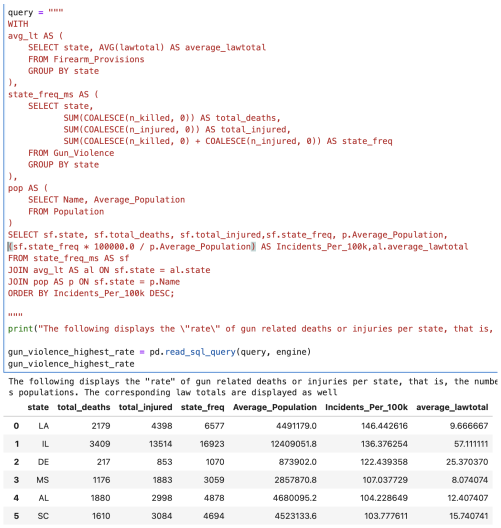
After the data had been cleaned, we connected to a new database and wrote the csv files in as tables. For this portion, we used a new dataset that contained population by state. First, we identified the states that had the highest and lowest amount of firearm provisions each year. Then, we joined the provisions table on the state and year columns to the other tables to display different violence statistics alongside the total number of firearm provisions. We found the states with the highest mortality rate for each year, then found the rate of shootings by state and simply plotted the results. These queries required three joins to account for using the population, obtaining the incident data, and calculating the average law totals for the state across all available years. The rate of mass shootings per 100k people and gun related incidents were calculated separately, and required a join to the population table on the state column. This is because we need to account for the impact of population on the direct number of incidents–a state with a higher population will always have more incidents, but we want to see if a state has a *disproportionate* number of incidents. If this project were to be taken further, we could refine our analysis by accounting for the existence of urban centers and population density as well.



```
query = """
WITH
avg_lt AS (
    SELECT state, AVG(lawtotal) AS average_lawtotal
    FROM Firearm_Provisions
    GROUP BY state
),
state_freq_ms AS (
    SELECT State,
           COUNT(*) as total_mass_shootings
    FROM Mass_Shootings
    GROUP BY state
),
pop AS (
    SELECT Name, Average_Population
    FROM Population
)
SELECT sf.State, sf.total_mass_shootings, p.Average_Population,
    (sf.total_mass_shootings * 100000.0 / p.Average_Population) AS Mass_Shootings_Per_100k,
    al.average_lawtotal
FROM state_freq_ms AS sf
JOIN avg_lt AS al ON sf.State = al.state
JOIN pop AS p ON sf.state = p.Name
ORDER BY Mass_Shootings_Per_100k DESC;

"""
print("The following displays the \"rate\" of shootings per state, that is, the number of sho
mass_shootings_highest_rate = pd.read_sql_query(query, engine)
mass_shootings_highest_rate
```

```
The following displays the "rate" of shootings per state, that is, the number of shootings th
corresponding law totals are displayed as well
```

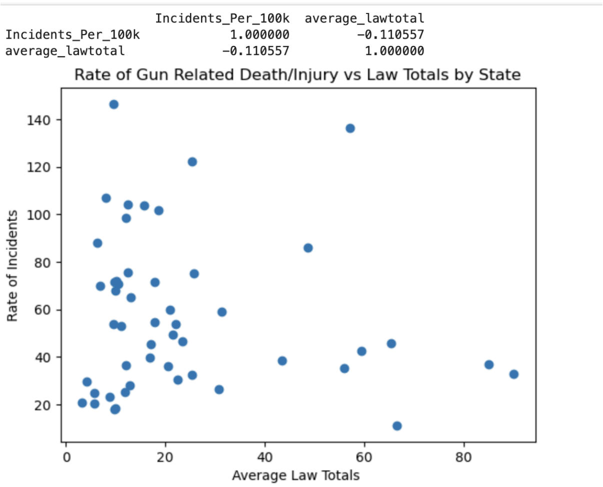
	State	total_mass_shootings	Average_Population	Mass_Shootings_Per_100k	average_lawtotal
0	AK	3	670800.4	0.447227	6.296296
1	CO	9	4855439.0	0.185359	16.888889
2	AR	4	2803833.6	0.142662	12.481481
3	KY	6	4220173.2	0.142174	6.888889
4	NV	3	2439880.2	0.122957	9.740741



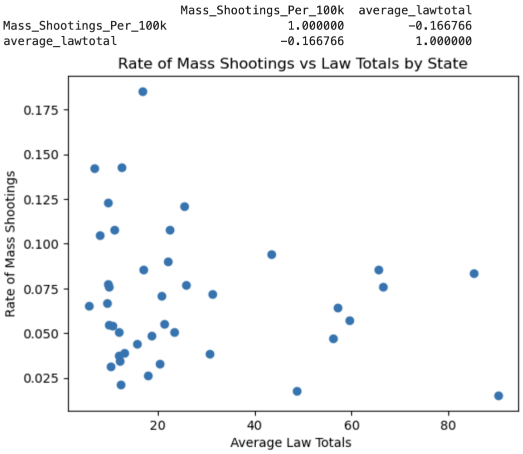
```
query = """
WITH
avg_lt AS (
    SELECT state, AVG(lawtotal) AS average_lawtotal
    FROM Firearm_Provisions
    GROUP BY state
),
state_freq_ms AS (
    SELECT state,
           SUM(COALESCE(n_killed, 0)) AS total_deaths,
           SUM(COALESCE(n_injured, 0)) AS total_injured,
           SUM(COALESCE(n_killed, 0) + COALESCE(n_injured, 0)) AS state_freq
    FROM Gun_Violence
    GROUP BY state
),
pop AS (
    SELECT Name, Average_Population
    FROM Population
)
SELECT sf.state, sf.total_deaths, sf.total_injured,sf.state_freq, p.Average_Population,
(sf.state_freq * 100000.0 / p.Average_Population) AS Incidents_Per_100k,al.average_lawtotal
FROM state_freq_ms AS sf
JOIN avg_lt AS al ON sf.state = al.state
JOIN pop AS p ON sf.state = p.Name
ORDER BY Incidents_Per_100k DESC;

"""
print("The following displays the \"rate\" of gun related deaths or injuries per state, that is,

gun_violence_highest_rate = pd.read_sql_query(query, engine)
gun_violence_highest_rate
```

```
The following displays the "rate" of gun related deaths or injuries per state, that is, the numbe
s populations. The corresponding law totals are displayed as well
```

	state	total_deaths	total_injured	state_freq	Average_Population	Incidents_Per_100k	average_lawtotal
0	LA	2179	4398	6577	4491179.0	146.442616	9.666667
1	IL	3409	13514	16923	12409051.8	136.376254	57.111111
2	DE	217	853	1070	873902.0	122.439358	25.370370
3	MS	1176	1883	3059	2857870.8	107.037729	8.074074
4	AL	1880	2998	4878	4680095.2	104.228649	12.407407
5	SC	1610	3084	4694	4523133.6	103.777611	15.740741



```
                   Incidents_Per_100k  average_lawtotal
Incidents_Per_100k            1.000000         -0.110557
average_lawtotal             -0.110557          1.000000
```

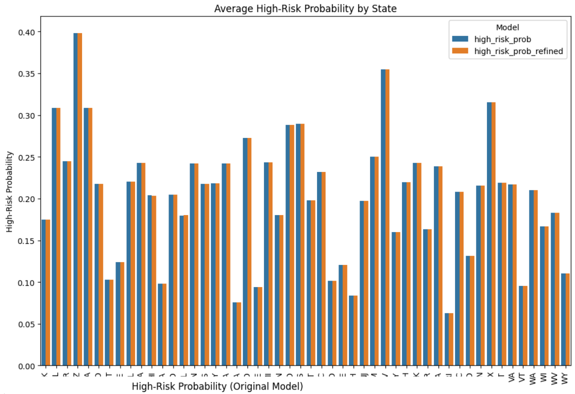


```
                       Mass_Shootings_Per_100k  average_lawtotal
Mass_Shootings_Per_100k                1.000000         -0.166766
average_lawtotal                      -0.166766          1.000000
```

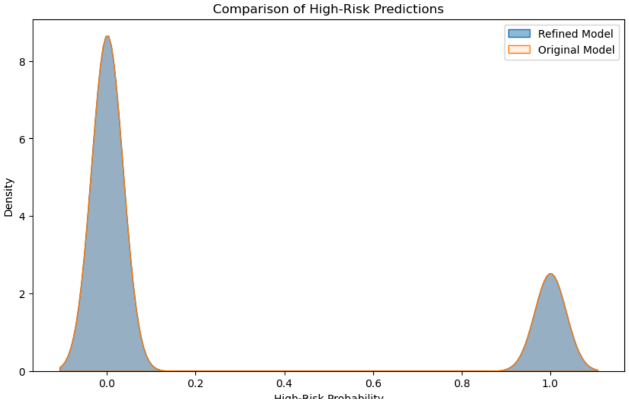
We are primarily studying how state legislation impacts the number of shootings by state. Still, it is important to note that other factors may also be related to the higher crime statistics in a state. The other factor we will consider is the time period of when the shootings take place. For example, whether or not shootings are more likely to take place during times of economic recessions or pandemics. We will also study shooter profiles to see whether mental illnesses are more prevalent in a state than others.

To make the data more presentable, we would use melt, pivot, and concat and join the databases by common locations to consolidate all of the information by state. For preliminary analysis, we would create charts and graphs to make some initial predictions on which legislation has the largest impact on shootings. After gaining more insight, we will use prediction models to train and test the data to confirm or reject our initial hypothesis. If our model is accurate, we would further proceed by using the model to predict where future shootings will take place based on current and new legislation that will be introduced in the coming years.

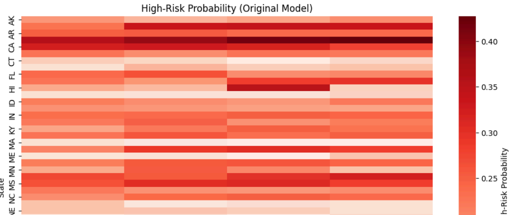
We can test if our model is accurate by separating out some validation data and comparing our generated output to the real values. Further, we will evaluate the accuracy of our model using metrics like accuracy, precision, and R-squared. We can compare predictions against real data and use statistical tests to confirm the relationships we find.



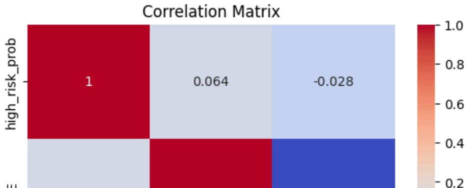
Average High-Risk Probability by State
Model
high_risk_prob
high_risk_prob_refined
High-Risk Probability
VA
VT
WA
WI
WV
WY



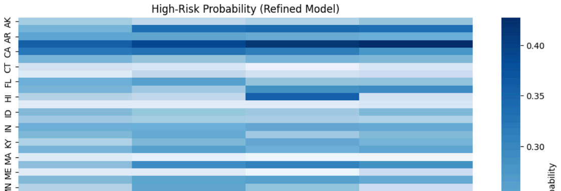
Comparison of High-Risk Predictions
Refined Model
Original Model
Density
High-Risk Probability



High-Risk Probability (Original Model)
State
AK
AR
CA
CT
FL
HI
ID
IN
KY
MA
ME
MN
MS
NC
High-Risk Probability



Correlation Matrix
high_risk_prob



High-Risk Probability (Refined Model)
AK
AR
CA
CT
FL
HI
ID
IN
KY
MA
ME
MN