Milestone Report

Problem Statement

The United States has notoriously high gun violence and mass shootings incidents, and the problem seems to be getting worse over time. Often people say that gun control laws wouldn't affect gun violence rates because gun homicides are committed by criminals who don't follow laws. However, I wanted to see if analyzing the data would give help to confirm or reject this claim.

The primary goal of this project is to examine gun violence trends in the United States and find correlations with gun control laws. Here are some of the questions we will be answering:

- How have gun violence rates changed over the years?
  Which states have seen the biggest increases in gun violence? the lowest?
- How does gun control in states with low gun violence rates compare with gun control in states with high gun violence rates?
- Which gun control laws have the most correlation with reduced gun violence?
- Are there categories of gun control laws that perform better than others? (e.g. background checks vs. banning assault weapon?)

We are also interested to see if we can find any relationship between gun violence and features such as income, substance abuse, and other crime. Finally, we will be creating a model to help predict whether gun violence will increase in the next year for each state.

Through this project, I hope to find useful insights on gun control provisions that may help policymakers make better decisions in interest of reducing gun violence. Using data to analyze laws allows us to justify policies with evidence and understand which laws work and which ones don't. Furthermore, by predicting gun violence increases in each state, we can take countermeasures to prevent them such as budgeting for law enforcement or implementing new gun control policies.

Data Wrangling

I. Summary of the Data Used

The data used in this project comes from many different sources. The primary goal of the data wrangling was to create CSV files that consolidate all of the useful features to be used during the visualization and modeling stages.

Here are the following datasets used listed alphabetically. Each entry includes the name of the file in the './data/raw' directory, the source, and descriptions of each dataset:

National Institute on Alcoholism and Alcohol Abuse
Alcohol consumption per capita for each state from 1977-2016.
./data/raw/alcohol.csv

GunPolicy.org
Annual gun homicides data per state from 2000 to 2013.
./data/raw/annual_gun_deaths.csv

Gun Violence Archive
Gun violence incidents from January 2014 to March 2018. This dataset was downloaded from Kaggle, although it was originally scraped from the Gun Violence Archive.
**Note: I added one entry for the Las Vegas shooting in October 2017, as it appeared to be missing*.
./data/raw/incidents.csv.gz

Disaster Center

Annual crime factors from 2010 to 2016. The Disaster Center collected the data from the FBI UCS Annual Crime Reports.
./data/raw/crime.csv

US Census Bureau
Two different population CSVs were downloaded from the US Census Bureau, and merged together in a simpler file ./data/raw/population.csv via a Python script merge-populations.py. This merged CSV contains annual populations for each state from 2000 to 2017.
./data/raw/population_2000_2010.csv
./data/raw/population_2010_2017.csv
./data/raw/population.csv

Kaggle
Gun control provisions as annual entries for each state from 1991 to 2017. This dataset was generated from several sources, including Thomson Reuters Westlaw legislative database and data from Everytown for Gun Safety and Legal Science, LLC. Each gun provision is encoded as a shortened codename. The details about each provision and its shortened name can be found in ./data/raw/codebook.xlsx
./data/raw/provisions.csv

270 to Win
Election results from 2000 to 2016 for each state. This dataset simply lists the percentage of votes each part received in each state.
./data/raw/election_results.csv

Bureau of Economic Analysis
Personal income data from 2009-2017, listed annually and for each state.
./data/raw/income.csv

Bureau of Alcohol, Tobacco, and Firearms
Annual gun registration data for each state from 2011-2017.
./data/raw/registrations.csv

SAMHSA
Substance use for each state from survey results from 2012-2016. 6 features were selected from the survey results. They are:
marijuana - Use of marijuana in the last year
cocaine - Use of cocaine in the last year
tobacco - Use of tobacco within the last month
alcohol - Alcohol abuse or dependency
mental - Any mental illness
depression - Serious depression episode within the last year
More details of the criteria of these features can be found from the SAMHSA website.

./data/raw/substances.csv

II. Cleaning and Consolidating the Data

Each dataset has a different range of years for which it is available. Here are the important things to note about the availability of the data:

- **Daily gun homicide data is available for 2014-2017**, whereas only **annual gun homicide data is available for 2000-2013**.
- **All data is available for at least 2013-2016**, allowing us to use every feature in a model predicting gun violence for 2014-2017 (we make predictions for at least one year into the future).

To make the data easier to work with, I decided to combine the data into several CSV files, each structured to make specific tasks easier later on. Here is a summary of these CSV files:

./data/cleaned/annual.csv
This CSV contains all of the features* for each state from 2000 - 2017, with entries where feature was not available as the only null values. This will be the primary CSV used for visualizations, as it is easy to interpret and manipulate.

./data/cleaned/feature.csv
This CSV contains monthly homicide rates for each state from 2014-2017. It is organized as gun homicide observations for each month and state, paired with all of the annual features* from the previous year. This will be the CSV we use for modeling, as it has no null values. We will be using observations in 2014-2016 for training to predict monthly gun homicide rate changes for each state in 2017.

./data/cleaned/by_date_total.csv
A CSV that contains the number of daily deaths for each state (as columns) from January 2014 to March 2018. I decided to make this CSV because it is the only CSV that will contain daily homicide rates. These daily homicide rates will be resampled and used for visualizations as well as time series analysis.

./data/cleaned/location.csv
A CSV that contains latitude and longitude coordinates for each incident. It also includes location information for each incident. Since we have a lot of missing values for these features and we don't care about individual incidents for our modeling, I decided to put them in a separate CSV file for a few quick visualizations later.

* For the annual and feature files, provisions data was excluded. The provisions data will be added accordingly during visualization and modeling. The reasoning is explained below.

III. Null Values, Outliers, and Provisions Data

Null Values

There are several null values to be noted among the datasets:

- **'Suppressed' values for annual gun homicide totals for states with very low annual gun homicides**. There are 6 states with these suppressed values (Hawaii, New Hampshire, North Dakota, South Dakota, Vermont, and Wyoming). According to the source GunPolicy.org, these values are cases with fewer than 10 homicides. Since these states have very low gun violence incidents, I imputed them with the mean during visualization. During modeling, states with very low average gun homicide rates are excluded, including these 6 states.
- **Missing values for District of Columbia for provisions data**. There are no entries for the District of Columbia for gun control provisions. This is unfortunate as the District of Columbia is an outlier for high gun violence rates. Ultimately, the District of Columbia will have to be dropped from the analysis, as provisions data is extremely important to this project.
- **Missing latitude and longitude values for around 40% of incidents**. There seems to be more missing values the more recent the incident is. However, this is not a big problem, as these values will only be used for an fun visualization.

Outliers

In our data, the notable outliers are states like Hawaii, which has an extremely low gun violence rate, and the District of Columbia, which has an extremely high gun violence rate. These outliers are useful in analysis and is part of our exploration on the factors that affect gun violence rates. Deciding whether to remove these outliers is an important for the modeling phase, and will be decided when tweaking the features before training.

Provisions

Furthermore, there are so many provisions (133 in total), that I decided to add them later as needed for visualization and modeling. The provisions data is not limited by availability; it contains years from 1991 to 2017, so it will be easy to selectively choose the provisions to use later. This will give us more flexibility in visualizations and modeling, and make the data cleaner overall.

Exploratory Data Analysis and Hypothesis Testing

I. Provision Efficiency

Single Provisions

Under the “Analyzing the Effects of Provisions” portion of my project, I attempt to find gun control provisions that correlate most with reduced gun violence. I examine the differences in gun violence for states that have a certain gun control provision in effect compared with states that don’t have that provision for a given year.

In order to control for sample size and outliers, I used a minimum sample size threshold and discarded provisions where the number of samples in either group was below the threshold. I created bar chart visualizations, but I wanted to conduct a t-test to confirm the differences seen in the bar chart.

The chart below shows the p-values:

Provision	Gun Violence Rates in States Without Provision	Gun Violence Rates in States With Provision	Difference	P-value
nosyg	6.024715	3.466635	2.558080	0.000893
immunity	5.629389	3.030228	2.599161	0.001455
violentpartial	5.359422	3.441462	1.917959	0.024187
statechecksh	5.336470	3.490234	1.846237	0.030354
mcdvdating	5.349055	3.574407	1.774649	0.034754
cap14	5.286320	3.596802	1.689518	0.048442

Here’s a brief description of each provision codename:

nosyg - Use of deadly force is not allowed to be a first resort in public. This is sometimes referred to as a "stand your ground" law.

immunity - No law provides blanket immunity to gun manufacturers or prohibits state or local lawsuits against gun manufacturers.

violentpartial - Firearm possession is prohibited for people who have committed a violent misdemeanor punishable by more than one year of imprisonment.

mcdvdating - All people convicted of a misdemeanor crime of domestic violence are prohibited from possessing firearms.

statechecksh - State conducts separate background checks, beyond the federal NICS check, for handguns.
cap14 - Criminal liability for negligent storage applies to access by children less than 14 years old.

For a better explanation of what each provision means, the codenames can be looked up in the ./data/raw/codebook.xlsx

Pairs of Provisions

I also wanted to see whether a pair of provisions would have a significant correlation with reduced gun violence.

Once again, a t-test is used to get a p-value for each of the differences:

Provision	Gun Violence Rates in States Without Provision	Gun Violence Rates in States With Provision	Difference	P-value
age21handgunsale permith	5.235497	2.698168	2.537328	0.015916
statechecks universalpermith	5.330041	2.913307	2.416734	0.021859
statechecks universalpermit	5.187620	2.913307	2.274313	0.022568
age21handgunsale universalpermith	5.336773	2.982114	2.354659	0.026356
age21handgunsale mcdv	5.443077	3.293261	2.149815	0.028332

Here are the descriptions of the provisions above:
statechecks - State conducts separate background checks, beyond the federal NICS check, for all firearm.
universalpermith - Background checks conducted through permit requirement for all handgun sales (or universal background checks).
universalpermit - Background checks conducted through permit requirement for all firearm sales (or universal background checks).

age21handgunsale - Purchase of handguns from licensed dealers and private sellers restricted to age 21 and older.
mcdv - People convicted of a misdemeanor crime of domestic violence against a spouse, ex-spouse, or cohabitating partner are prohibited from possessing firearms.

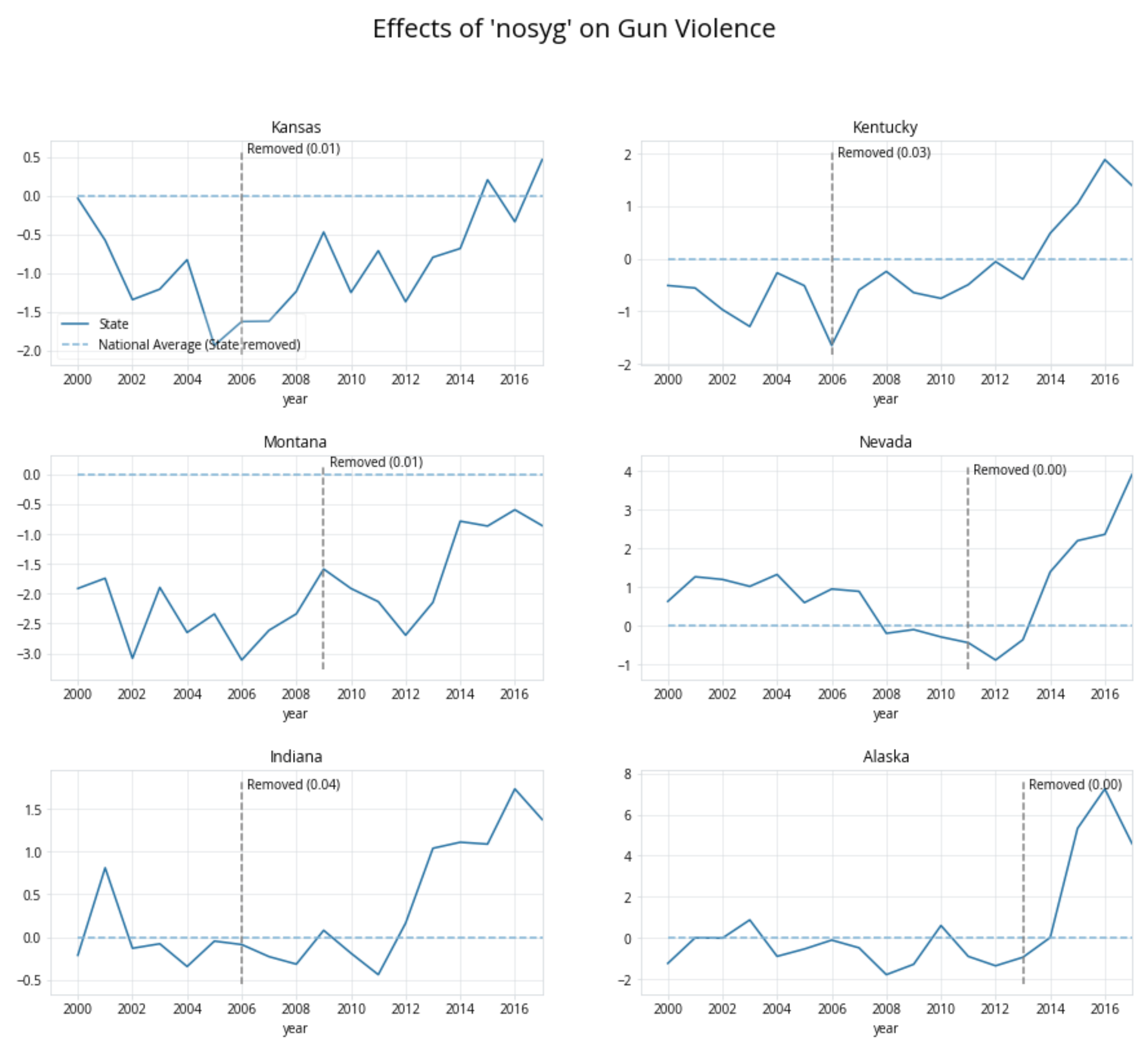
II. Testing for Structural Breaks

We looked at when provisions were either added or removed by a state to see if we could see any differences in gun violence after those changes were made. In order to quantify this change, we did a Chow test for structural break, an f-statistic defined as:

$$\frac{((RSS_{pooled} - (RSS_{before} + RSS_{after})) / k)}{(RSS_{before}+RSS_{after})/(N_{before} + N_{after} - 2k)}$$

where:
RSS_{pooled} - The residual sum of squares from regression on the entire time series
RSS_{before} - The residual sum of squares from regression on the time series before the structural break
RSS_{after} - The residual sum of squares from regression on the time series after the structural break
N_{before} - Number of samples in the time series before the structural break
N_{after} - Number of samples in the series after the structural break
k - Number of parameters estimated during the regression

To visualize this test, each time series is shown with a grey dotted line which indicates a change in provision. Next to it in parentheses is the p-value of the Chow test at that point. For conciseness, only one examples of these tests are shown (more can be found in visualization.ipynb):



This test gives us some confidence in making judgements about observed changes in gun violence rates after a provision is changed. However, there is a limited sample size for the Chow test, and a large number of factors that could affect gun violence. The structural break test results are to be interpreted with this in mind.