Comparative Analysis of Firearm Ownership and Crime Rates Across U.S. States

Ching Sze, Fung – 2313 7788

Introduction

Firearm ownership is always a controversial topic in the United States, where the right to bear arms is protected by the Constitution. However, individual states vary greatly in the regulations they impose on firearm ownership.

This project analyzes the correlation between firearm ownership, regulation strictness, and crime rates across various U.S. states. States like California, Massachusetts and New York have strict regulations, while Texas, Alaska and Wyoming are more lenient. By comparing data from both types of states, we aim to identify patterns and assess how regulation differences impact crime rates, offering insights into the relationship between gun ownership and crime.

Question

How does the level of firearm ownership and the stringency of restrictions correlate with crime rates across different states in the United States, and what differences can be observed between states with stricter vs. more lenient gun laws?

Data Sources

Data Source 1: FBI NICS Firearm Background Checks

Description: The dataset is sourced from Kaggle and originally provided by the National Instant Criminal Background Check System (NICS). It contains the number of FBI NICS firearm background checks by month, state, and type from November 1998 to 2023. This dataset provides insights into the stringency of firearm restrictions across different states, as it includes mandatory background checks for citizens wishing to purchase firearms.

Structure and Quality: The data is organized in a fixed tabular format, with columns representing different permit types and activities, collected monthly by state. The quality of the data is generally consistent and suitable for analyzing firearm background checks.

Licenses: The MIT License (MIT) , which is a permissive open-source license.

The data consists of FBI records that are available for public review on the FBI's electronic FOIA Library. To comply with the license obligations, appropriate credit to the FBI and a link to the license will be provided in this project.

Metadata URL: Kaggle - FBI NICS Firearm Background Checks

Data URL: FBI NICS Firearm Background Checks Raw Data

Data Source 2: Firearm Mortality by State

Description: The dataset is sourced from the Centers for Disease Control and Prevention (CDC). It provides firearm mortality statistics at the state level across the United States, including annual data on deaths resulting from firearm-related incidents. The data includes information such as the number of firearm-related deaths, categorized by state and year. This dataset is a crucial source for understanding trends in firearm-related mortality.

Structure and Quality: The data is organized in a fixed tabular format with five columns: year, state, rate, deaths, and url. The quality of the data is straightforward but includes all the necessary information for analysis.

Licenses: The dataset is licensed under Section 308(d) of the Public Health Service Act [42 U.S.C. 242m(d)] and the Confidential Information Protection and Statistical Efficiency Act

(CIPSEA) [44 U.S.C. 3561-3583]. Data from the National Center for Health Statistics (NCHS) may be used only for statistical reporting and analysis, which is suitable for this project. To comply with the license obligations, appropriate credit to the CDC and a link to the license will be provided in this project.
Metadata URL: CDC - Firearm Mortality by State

Data Source 3: FBI Crime Data
Description: The dataset is sourced from the FBI's Centers for Disease Control and Prevention (CDC). It contains crime statistics across the United States, including detailed factors associated with the crimes, such as “Type of Weapon Involved by Offense”, “Offense Linked to Another Offense”, and others. This dataset provides valuable insights into crime rates across different U.S. states and helps to identify the factors contributing to various types of offenses.
Structure and Quality: Although the dataset is organized in a fixed tabular format, the data is divided across 13 separate CSV files, which need to be combined during the transformation step. The quality of the data is incomplete and inconsistent, especially in the "Type of Weapon Involved by Offense" and "Offense Linked to Another Offense" columns, as they are not recorded on a monthly or yearly basis, unlike the estimated crime CSV file.
Licenses: The dataset is available for public review under the FBI's Freedom of Information Act (FOIA) Library. The FBI provides public access to this data for transparency purposes, and it can be used for analysis and academic purposes, following the general terms of the FBI's privacy policy. To comply with license obligations, appropriate credit to the FBI will be given, and a link to the privacy policy and terms of use will be provided in the project documentation.
Metadata URL: FBI Crime Data Explorer

Data Pipeline

The ETL(Extract, Transform, Load) pipeline is implemented using python to handle both data sources, each downloaded as a CSV directory within a zip archieve. This process involves extracting the right CSV file, transforming, and saving it as CSV format and SQLite databases.

Extraction: The dataset source 1 is downloaded using the Kaggle API. For the dataset source 2, `Selenium` is used to automate the download of the CSV file from the CDC website. Then, multiple CSV files are downloaded directly for the dataset source 3. After extraction, the loaded data is processed using `pandas` for further transformation.
Transformation: The transformation process varies based on the specific data sources:

Data Source 1: FBI NICS Firearm Background Checks	Data Source 2: Firearm Mortality by State	Data Source 3: FBI Crime Data
Filter Data: The rows are filtered by 6 specific states, California, Massachusetts, New York, Texas, Alaska and Wyoming.		
Date Conversion: The "month" column was converted to extract only the year using `pd.to_datetime()`.	**Dropping Irrelevant Columns**: The "URL" column was dropped as it was deemed irrelevant to the analysis.	**Combined Dataset**: The "Offense Linked to Another Offense Data" and "Weapon Type Involved by Offense Data" datasets both

Grouping and Summarizing Data: After extracting the year, the data was grouped by "year" and "state" and aggregated using `.sum()`, which helps in summarizing the data to observe yearly trends.	**Data Type Conversion:** The "DEATHS" column was initially in a string format that included commas, which made it unsuitable for numerical operations. To address this, commas were removed using `.replace()` and the column was converted to integer type using `.astype(int).`	contain columns named "Key" and "Value." First, data from both datasets is filtered to include only six specific states, "Weapon Law Violations," and specific weapon types such as "Handgun," "Rifle," "Shotgun," etc. After filtering, the relevant CSV files are combined, and the number of cases is summed.

Loading: The cleaned datasets are saved as CSV files and into an SQL database for future analysis.

Problems Encountered, Solutions,Meta-Quality Measures and Error Handling:

File Download Automating: Since Selenium is adopted for downloading CSV files from the CDC website, there are instances where the download takes longer, and no direct link to the CSV file can be fetched. To handle this, a timeout mechanism (time.sleep()) was implemented to allow sufficient time for the download to complete.

However, for dataset source 3, there are multiple download buttons (more than 5 to 6), and the CSV files are separated based on different crime characteristics. Using Selenium to automate these downloads was not feasible. Therefore, the CSV files were downloaded manually for data processing.

Irregular Data Formatting: Some columns, such as the "value" column in Dataset Source 3, contained non-numeric entries, which were problematic for analysis. To address this, the entries were converted to numeric using pd.to_numeric(). Any non-convertible entries were coerced to NaN and subsequently replaced with 0 to avoid errors.

Result and Limitations

Output Data: The final output includes four transformed and cleaned CSV files, as well as an SQLite database. All datasets are focused on six specific states: California, Massachusetts, and New York (representing states with strict firearm regulations), and Texas, Alaska, and Wyoming (representing states with more lenient restrictions). These datasets capture firearm background checks on permits and various types of firearms by state and year. Besides, two other datasets provide information on the firearm mortality rate, the number of deaths, and categorized crime rates by state and year, including the number of weapon types involved in the crimes. (See Tables 1-4)

Data Structure, Quality and Format: All cleaned datasets are organized in a tabular schema with defined data types for each attribute, such as Year, State, and specific features related to background checks, mortality rates, crime, and weapon types. After going through the transformation pipeline, the datasets were cleaned and consistent by removing irrelevant columns, normalizing data formats, and converting non-numeric values. Therefore, the dataset fulfills the criteria of completeness, consistency, timeliness, and relevance, as it contains the required information for further analysis and captures over 20 years of data. CSV files and an SQLite database were chosen for storage because they are easily accessible and allow for efficient querying and further analysis.

Reflection and Potential Issues: Although the datasets have been cleaned and transformed, the granularity might not fully capture a comprehensive comparison between states with strict and lenient firearm regulations. Specifically, for Dataset 3, the weapon types involved in other offenses are not captured by year, making them inconsistent with the yearly totals of estimated crimes. This inconsistency could affect the accuracy and reliability of the analysis. Furthermore, social crime rates can be influenced by various factors, such as current social issues, the economic situation of the state, and more. As a result, the stringency of firearm regulations and the level of firearm ownership may be just one of the factors relevant to crime rates.

Table 1: FirearmBackgroundCheck_cleaned_dataset.csv

Year	State	Permit	Permit_recheck	handgun	Long_gun	…	totals
2023	Wyoming	4495	122	18945	17979	…	47238
2023	Texas	239276	0	495329	262526	…	1153813
2023	New York	45951	24893	94475	100566	…	286031
2023	Massachusetts	81549	4	44781	24282	…	165835
2023	California	240772	128351	312780	197309	…	1064943
2023	Alaska	1994	210	25414	21142	…	55684
2022	Wyoming	5142	257	29273	27529	…	70276

Table 2: FirearmMortalitybyState_cleaned_dataset.csv

Year	State	Rate	Deaths
2022	CA	8.6	3484
2022	NY	5.3	1044
2022	WY	20.4	124
2022	TX	15.3	4630
2022	AK	22.4	164
2022	MA	3.7	263
2021	CA	9	3576

Table 3: EstimateCrimes_cleaned_dataset.csv

Year	State_abbr	State_name	Population	Violent_crime	Homicide	…	Totals
2023	AK	Alaska	733,406	5,327	62	…	9218
2023	CA	California	38,965,193	198,036	1,929	…	331905
2023	MA	Massachusetts	7,001,399	21,998	146	…	39526
2023	NY	New York	19,571,216	76,298	595	…	127118
2023	TX	Texas	30,503,301	123,856	1,845	…	211615
2023	WY	Wyoming	584,057	1,116	18	…	1851
2022	AK	Alaska	733,583	5,627	70	…	9684

Table 4: CombinedData_Weapon_cleaned_dataset.csv

Key	AK	CA	MA	NY	TX	WY	…
Weapon Law Violations	294	6122	11635	11836	9025	36	…

Violation of National Firearm Act of 1934	0	0	0	0	0	0	…
Handgun	520	17073	24346	6084	156970	285	…
Shotgun	114	403	865	263	6958	23	…