Sourcing of Data

Summary

The Gun Violence dataset comprises records of over 260,000 gun violence incidents in the US from January 2013 to March 2018. The project aimed to provide detailed information about each incident in CSV format, intending to facilitate study and informed predictions about future trends by data scientists and statisticians. The dataset is intended to address the current lack of large and easily accessible detailed data on gun violence, and it is expected to contribute to a better understanding of the issue and the development of effective preventive measures.

Source

The data was collected from gunviolencearchive.org.

Collection

The dataset was collected using web scraping techniques due to limitations in the number of incidents that could be obtained from a single query, and the absence of crucial fields in the website's "Export to CSV" functionality. The data collection process involved three stages:

1. A Python script was used to query all incidents for each date between January 1, 2013, and March 31, 2018. The data was then scraped and written into monthly CSV files.
2. Each entry was enhanced with additional data not directly viewable from the query results page, such as participant information and geolocation data.
3. The entries were sorted by date and merged into a single CSV file. The dataset aims to provide detailed information about gun violence incidents in the US to facilitate study and informed predictions about future trends by data scientists and statisticians

Contents

There is a single dataset. Table:'**gun-violence-data_01-2013_03-2018'**

Column	Description
incident_id	ID number for each incident
date	Date xxxx-xx-xx
state	USA State
city_or_county	USA City or County
address	Street address of incident

n_killed	Number of people killed because of the incident
n_injured	Number of people injured because of the incident
incident_url	Link to gunviolencearchive.org log of the incident
source_url	Link to local news outlet/coverage of the incident
incident_url_fields_missing	(TRUE/FALSE) If the ‘incident_url’ column is blank
congressional_district	Numeric ID for State
gun_stolen	Was the firearm stolen or not
gun_type	Identification of firearm
incident_characteristics	List of criminal activities resulting from the incident
latitude	Latitude
location_description	Description of shooting location (business name, apartment name, etc)
longitude	Longitude
n_guns_involved	Number of firearms involved in incident
notes	Any additional information provided
participant_age	(0::0) Participant::Age
participant_age_group	(0::Age Group) Participant:: Age Group
participant_gender	(0::Gender) Participant:: Gender
participant_name	(0::Name) Participant::Name
participant_relationship	(0::Relationship) Participant::Relationship
participant_status	(0::Outcome) Participant::Outcome
participant_type	(0::Involvement) Participant::Involvement
sources	Additional source link
state_house_district	Numeric ID for State
state_senate_district	Numeric ID for State

Limitations

The dataset is incredibly dense with information. Unfortunately, there are a handful of columns that have little to no use to us, as they are essentially duplicates, or are just links to external sources. (ie ‘incident_url’, ‘source_url’, ‘sources’, etc). There is no information provided before the incident that helps us understand the shooter. For example, are they in school? Employed? Married? Additionally, there is no information on whether the shooters are convicted or not.

Ethics

This data breaks many ethics violations, as many columns identify people by name, age, gender, and location. The data will need to be cleaned in such a way that prevents any of this information from getting out.

Clean Your Data

Column	Issue	Action
address	Displayed addresses that may break the code of Ethics	Remove column
incident_url	Displayed addresses, names, ages, etc, that may break the code of Ethics	Remove column
source_url	Displayed addresses, names, ages, etc, that may break the code of Ethics	Remove column
sources	Displayed addresses, names, ages, etc, that may break the code of Ethics	Remove column
participant_name	Displays a person’s name, and breaks the code of ethics	Remove column
participant_age	Displays the person’s age, and breaks the code of ethics	Remove column
participant_gender	Displays the person’s gender, and breaks the code of ethics	Remove column
participant_relationship	Displays multiple people’s relation to other people involved in the incident. This breaks the code of ethics.	Remove Column
latitude	Provides exact location details, thus breaking the code of ethics	Removing column
longitude	Provides exact location details, thus breaking the code of ethics	Removing column

notes	Names shooters/victims directly, plus other personal information. This breaks the code of ethics.	Removing column
incident_url_fields_missing	Provides nothing to the analysis	Removing column
location_description	Provides exact location details, thus breaking the code of ethics	Removing column
state_house_district	Provides information we have in another column	Removing the column for clarity
state_senate_district	Provides information we have in another column	Removing the column for clarity
congressional_district	Long and confusing name	Renamed to 'state_code'
gun_stolen	Very few columns with information entered, or mostly 'unknown'	Removing column
gun_type	Very few columns with information entered, or mostly 'unknown'	Removing column
incident_characteristics	Very long 'string' input answers that would be impossible to sort with, or cross with another column. Information deemed dispensable for this project.	Removing column
n_guns_invovled	Variables are wildly inaccurate, and we're unable to use the information for our project.	Removing column

Continued Cleaning & Aggregation of the Data

Column	Issue	Action
adult	Created from participant_age_group, uses 'TRUE/FALSE' to indicate presence of someone from this age group in incident	Created new column
teen	Created from participant_age_group, uses 'TRUE/FALSE' to indicate presence of someone from this age group in incident	Created new column
child	Created from participant_age_group, uses 'TRUE/FALSE' to indicate presence of someone from this age group in incident	Created new column
participant_age_group	With creation of 'adult', 'teen', and 'child' columns, this column is no longer needed for analysis.	Removing column

killed	Created from participant_status column, uses ‘TRUE/FALSE’ to indicate presence of death from incident	Created new column
injured	Created from participant_status column, uses ‘TRUE/FALSE’ to indicate the presence of injury from incident	Created new column
arrested	Created from participant_status column, uses ‘TRUE/FALSE’ to indicate presence of an arrest from incident	Created new column
unharmed	Created from participant_status column, uses ‘TRUE/FALSE’ to indicate presence of an unharmed person from incident	Created new column
participant_status	With creation of ‘killed’, ‘injured’, ‘arrested’, and ‘unharmed’, this column is no longer needed for our analysis	Removing column
victims	Created from participant_type column, counts number of victims and produces a single numeric value	Created a new column
suspects	Created from participant_type column, counts number of ‘Subject-Suspect’ and produces a single numeric value	Created a new column
participant_type	With creation of ‘victims’, and ‘suspects’, this column is no longer needed for our analysis	Removing column

- 52748 rows had ‘NULL’ values in the data. I chose to remove these from the data set. We are left with 186929 rows, and 17 columns, of data.
    - The removal was roughly 22% of the data set

Understanding the Data

Column	Qualitative/Quantitative	Discrete/Continuous	Nominal/Ordinal/Binary
incident_id	Qualitative	Discrete	Nominal
date	Qualitative	Discrete	Nominal
state	Qualitative	Discrete	Nominal
n_participants	Quantitative	Discrete	
n_killed	Quantitative	Discrete	
n_injured	Quantitative	Discrete	

victims	Quantitative	Discrete	
suspects	Quantitative	Discrete	
adult	Qualitative	Discrete	Binary
teen	Qualitative	Discrete	Binary
child	Qualitative	Discrete	Binary
killed	Qualitative	Discrete	Binary
injured	Qualitative	Discrete	Binary
arrested	Qualitative	Discrete	Binary
unharmed	Qualitative	Discrete	Binary

Cleaned Dataset - Ready for Analysis

Column	Description
incident_id	The ID number for each incident
date	Date (xxxx-xx-xx)
state	USA State where the incident took place
n_participants	The Number of Participants involved in the incident
n_killed	The Number of people killed in this incident
n_injured	The Number of people injured from this incident
victims	The amount of people involved in the incident that are victims
suspects	The amount of people involved in the incident that are the suspects
adult	Was an Adult involved in this incident
teen	Was a Teen involved in this incident
child	Was a Child involved in this incident
killed	Qualitative
injured	Were people killed during this incident
arrested	Was someone arrested as a result of this incident
unharmed	Did anyone leave unharmed from this incident

Define Questions to Explore

- Has gun violence increased/decreased from 2013-2018
- What states are leading in incidents, deaths, participants
- What percentage of these shootings involve Teens and Children
- How common are arrests?
    - Are there any trends we discover from this?
- How many of these shootings are leading to deaths?