Gun Violence Data

Data Source

Source of the dataset:
The dataset was originally obtained using web scraping techniques from the website of Gun Violence Archive (GVA) http://www.gunviolencearchive.org/ . Gun Violence Archive (GVA) is a not for profit corporation formed in 2013 to provide free online public access to accurate information about gun-related violence in the United States.

The open dataset is available on Kaggle.com under the following link:
Gun Violence Data (kaggle.com)

Collection:

As stated on the GVA website, gun violence incidents are collected/validated from 7,500 sources daily – Incident Reports and their source data are found at the gunviolencearchive.org website.

According to the "methodology" page on the website of the GVA, the data was collected using a mixed-method approach, which includes both automated and manual processes:

- Automated Queries: The Gun Violence Archive (GVA) utilized automated queries to gather initial data. This likely involves programmed searches and data scraping techniques to collect information from various databases and websites.
- Manual Research: In addition to automated queries, manual research was conducted through a broad range of sources, which includes local and state police reports, media outlets, data aggregates, government sources, and more. This step ensures a comprehensive collection of data and helps in validating the incidents reported by the automated systems.
- Verification Process: Each incident underwent a verification process involving initial researchers and secondary validation procedures to ensure the accuracy of the data.

Contents:

The dataset file encompasses a record of more than 260k gun violence incidents, with detailed information about each incident, available in CSV format.

The CSV file contains data for all recorded gun violence incidents in the US between January 2013 and March 2018, inclusive.

The original dataset contains the following columns:

Column Name	Description
incident_id	Unique identifier for the gun violence incident
date	Date when the incident occurred
state	State where the incident took place
city_or_county	City or county of the incident
address	Specific address of the incident
n_killed	Number of people killed in the incident
n_injured	Number of people injured in the incident
incident_url	URL to the detailed report of the incident
source_url	URL to the source of the incident report
incident_url_fields_missing	Indicator if any fields in the incident report are missing
congressional_district	Congressional district in which the incident occurred
gun_stolen	Information about whether the gun was stolen
gun_type	Type of gun(s) involved in the incident
incident_characteristics	Characteristics and nature of the incident
latitude	Latitude coordinate of the incident location
location_description	Description of the incident location
longitude	Longitude coordinate of the incident location
n_guns_involved	Number of guns involved in the incident
notes	Additional notes regarding the incident
participant_age	Age(s) of participant(s) in the incident
participant_age_group	Age group(s) of participant(s)
participant_gender	Gender(s) of participant(s)
participant_name	Name(s) of participant(s)
participant_relationship	Relationship(s) between participant(s)
participant_status	Status (e.g., arrested, injured) of participant(s)
participant_type	Role (e.g., victim, subject) of participant(s) in the incident
sources	Source(s) providing information about the incident
state_house_district	State house district where the incident occurred
state_senate_district	State senate district where the incident occurred

Why I have chosen this dataset

I think this dataset offers a comprehensive view of gun violence in the USA as it is rich in variables for robust analysis. It presents challenges like missing data and requires preprocessing, which are valuable skills to showcase. Its real-world significance and potential to inform policy add depth to the analytical exercise, making it an excellent choice for demonstrating data manipulation and visualization in applied analytics.

Data Profile

Cleaning and Consistency Checks:

The dataset contains several columns which I will not be using for my analysis. Therefore, to have better overview of the columns while working in Jupyter and to reduce the required memory needed for executing the codes, I will start by creating an alternative dataframe in Jupyter and calling it **df_gv**. The alternative dataframe will **exclude** the following columns: 'participant_relationship','location_description','participant_name','gun_stolen','n_guns_involved','gun_type','notes','state_house_district','state_senate_district','address','sources','source_url','incident_url_fields_missing','incident_url'

Before proceeding with data analysis and visualization, it is crucial to ensure the data is clean and consistent. Here are the basic data cleaning and consistency checks that I recommend performing on the df_gv dataframe in a Jupyter environment:

Missing Values:

- Identify any columns with missing values.
- Decide on a strategy for dealing with these missing values, such as imputing them with a central tendency measure (mean, median) or a constant value, or dropping the rows/columns altogether.

Columns with missing values	Action taken and reason
congressional_district: 11,944 missing values	No action taken for now. Most likely I will not be using this column in my analysis.
incident_characteristics: 326 missing values	No action taken for now. Most likely I will not be using this column in my analysis.
latitude: 7,923 missing values	My Analysis will focus on state-level insights. Thus, I will use the location values from the state column. I do not need longitude of latitude data.. No action required as I will not use this column for now
longitude: 7,923 missing values	My Analysis will focus on state-level insights. Thus, I will use the location values from the state column. I do not need longitude of latitude data.. No action required as I will not use this column for now
participant_age: 92,298 missing values	No actoin taken here. My analysis will focus on values from participant_age_group values

participant_age_group: 42,119 missing values	The proportion of missing values for the participant_age_group column is approximately 17.57%. The values in this column appear to be coded in a specific format, where each entry may consist of one or more age group designations, each prefixed with an identifier (like 0::, 1::, etc.), which may correspond to different participants. Thus, it would be very tricky to try to imputate the missing values. My approach would consider grouping the data into 'Known Age Group' and 'Unknown Age Group' to retain the records for certain types of analysis.
participant_gender: 36,362 missing values	No action taken for now. Most likely I will not be using this column in my analysis.
participant_status: 27,626 missing values	No action taken for now. Most likely I will not be using this column in my analysis.
participant_type: 24,863 missing values	No action taken for now. Most likely I will not be using this column in my analysis.

Data Types:

- Check the data types of each column to ensure they are appropriate for the data they contain.
- Convert data types if necessary, such as transforming `date` columns to datetime objects, or categorical columns to 'category' dtype.

Based on the data types reviewed in Jupyter, the data type in the 'date' column is currently object. This needs to be corrected to datetime64. I will perform this change in Jupyter.

Duplicates:

- Check for and remove any duplicate rows to prevent skewed analysis: As per my check in Jupyter, no duplicates found.

Unique Values:

- Examine the columns 'state', 'n_killed', 'n_injured' for unique values to spot any anomalies or unusual entries (e.g., a state name that is misspelled). I focused on these columns for now as I expect them to be key for my analysis: No issues found.

Understanding the Data:

Variables of the dataframe df_gv:

Variables	Time-variant / -invariant	Structured / Unstructured	Qualitative / Quantitative	Qualitative: Nominal / Ordinal Quantitative: Discrete / Continuous
incident_id	Time-invariant	Structured	Quantitative	Discrete
date	Time-variant	Structured	Quantitative	Continuous
state	Time-invariant	Structured	Qualitative	Nominal
city_or_county	Time-invariant	Structured	Qualitative	Nominal
n_killed	Time-variant	Structured	Quantitative	Discrete
n_injured	Time-variant	Structured	Quantitative	Discrete
congressional_district	Time-invariant	Structured	Quantitative	Discrete
incident_characteristics	Time-variant	Unstructured	Qualitative	Nominal
latitude	Time-invariant	Structured	Quantitative	Continuous
longitude	Time-invariant	Structured	Quantitative	Continuous
participant_age	Time-variant	Structured	Quantitative	Discrete
participant_age_group	Time-variant	Structured	Qualitative	Ordinal
participant_gender	Time-variant	Structured	Qualitative	Nominal
participant_status	Time-variant	Structured	Qualitative	Nominal
participant_type	Time-variant	Structured	Qualitative	Nominal
age_group_classification	Time-variant	Structured	Qualitative	Nominal

Descriptive statistics of the dataframe:

```
# descriptive statistics of the dataframe
df_gv.describe()
```

	incident_id	n_killed	n_injured	congressional_district	latitude	longitude
count	2.396770e+05	239677.000000	239677.000000	227733.000000	231754.000000	231754.000000
mean	5.593343e+05	0.252290	0.494007	8.001265	37.546598	-89.338348
std	2.931287e+05	0.521779	0.729952	8.480835	5.130763	14.359546
min	9.211400e+04	0.000000	0.000000	0.000000	19.111400	-171.429000
25%	3.085450e+05	0.000000	0.000000	2.000000	33.903400	-94.158725
50%	5.435870e+05	0.000000	0.000000	5.000000	38.570600	-86.249600
75%	8.172280e+05	0.000000	1.000000	10.000000	41.437375	-80.048625
max	1.083472e+06	50.000000	53.000000	53.000000	71.336800	97.433100

Considering limitations and ethics:

Limitations:

The original dataset contains detailed records of gun violence incidents in the United States from January 2013 to March 2018, with 239,677 entries and 29 columns. However, several limitations are apparent from the preliminary review:

- Missing Data: There are numerous columns with missing values, such as `address`, `gun_stolen`, `gun_type`, `location_description`, `participant_age`, `participant_relationship`, and others. This can limit the comprehensiveness of analysis performed on the dataset.
- Inconsistency in Data Format: The columns like `participant_age`, `participant_gender`, `participant_status`, and `participant_type` contain data in a concatenated string format (e.g., "0::Male||1::Male"), which would require significant preprocessing for analysis.
- Reliability and Verification: Since the data is collected from various sources, including media, the accuracy of each incident's details depends on the original reporting source, which may have varied standards of verification and completeness.
- Potential Underreporting: Defensive gun uses not reported to the police are not included. Therefore, the dataset may underrepresent the actual number of defensive gun use incidents.

When analyzing this dataset, one must factor in these limitations.

Ethics:

The original dataset in question involves sensitive information pertaining to gun violence incidents, which can include personal details of individuals involved in these events, especially under the columns: 'participant_name', 'notes', 'incident_url', and 'source_url'.

From a data ethics and privacy standpoint, some concerns could arise regarding Personal Identifiable Information (PII). The original dataset contains names or other identifiers specially in the columns above mentioned. Even though, this information has been published by US authorities and it complies with local data privacy regulations, I would consider dropping the respective columns from the dataset in Jupyter as this PII is not really relevant to my data analysis.

Questions to Explore

Below is some brainstorming of questions that I will consider to explore the data. These questions serve as a comprehensive starting point for data analysis, each potentially leading to actionable insights. The questions are not final and will be adjusted as I go through the analysis:

1. How many incidents of gun violence are recorded per year, and what is the trend over time?
2. What are the states with the highest number of gun violence incidents?
3. How do the numbers of killed and injured vary across different incidents?
4. Are there common characteristics among the most violent incidents (high number of casualties)?
5. Are there particular times of the year when gun violence peaks, and what might be the reasons for this?