Tiffani Barnett

Achievement 6

6.1 Sourcing Open Data

November 15, 2025

Data Source

This project uses an open dataset from Kaggle that documents more than 260,000 gun-related incidents in the United States between 2013 and 2018. Each record includes detailed information about the event, such as the date, location, number of people harmed, and related circumstances.

The dataset is widely used by researchers because it offers enough depth to identify patterns in gun violence, examine possible contributing factors, and build models to forecast trends. It serves as a strong foundation for analytical work focused on public safety.

Dataset link: https://www.kaggle.com/datasets/jameslko/gun-violence-data?resource=download

Data Collection

This dataset was created using information from public records in the United States. These records come from police reports, news stories, and government sources, which helps make the data trustworthy and well rounded.

It includes details like the date, location, number of victims, and any available information about the people involved. It also groups incidents by type, such as mass shootings or domestic violence, and notes any known circumstances.

Because the information is collected carefully from many reliable sources, the dataset gives a strong overview of gun violence over five years and makes it easier to study patterns and trends.

Data Limitations

Even though the dataset is extensive, several limitations need to be kept in mind when analyzing it:

1. **Time constraints.** The dataset ends in 2018, so it may not reflect more recent trends or legislative changes
2. **Not all incidents are captured.** Reporting standards differ across states and agencies, so some cases may be missing.
3. **Certain fields are incomplete.** Demographic information or contributing factors may be blank or inconsistent.

4. **Limited context.** The dataset does not include detailed socioeconomic information or policy context that may influence patterns.

Why This Dataset Was Chosen

I chose this gun violence dataset because it is one of the most complete and reliable sources available. It includes important details like dates, locations and the number of people involved in each incident. Since the data is well organized and easy to verify, it allows for accurate analysis without guessing or making assumptions.

The dataset also covers several years, which makes it possible to see patterns over time. It helps show where gun violence happens most, when it increases and what types of incidents occur. This supports my goal of using data to better understand real social problems and to take part in conversations about safety and prevention since it is an ongoing issue in the United States today

Ethical Considerations

Because this dataset includes sensitive incident information, responsible handling is essential:

1. **Protecting privacy.** Even though the data is public, it is important not to expose or misuse information about individuals involved in these events.
2. **Accurate representation.** Findings must be presented objectively to avoid misinterpretation or reinforcing biases.
3. **Careful communication.** Gun violence is a sensitive subject, so analysis should be shared respectfully and thoughtfully.
4. **Purpose-driven use.** The data should serve research, safety discussions, and constructive analysis, not sensationalism.
5. **Transparency about limitations.** It is important to acknowledge where the data cannot fully explain underlying causes or patterns.

Key Questions for Exploration

1. How did the number of gun-violence incidents change each year between 2013 and 2018?
2. Which states experienced the highest and lowest number of gun-related incidents during 2013 and 2018?
3. Are gun-violence incidents more common during specific months or seasons?
4. What types of gun-violence events occur most frequently (mass shooting, domestic disputes, armed robberies, accidental shootings, etc)?
5. How do casualty counts(injuries and fatalities) differ across states or incident types?
6. Are there particular cities or counties that emerge as persistence hotspots for gun violence?
7. What demographic patterns can be observed among participants (age, gender, role as victim or suspect)?

8. Is there a relationship between the number of firearms used in an incident and the severity of causalities?
9. Do incidents involving stolen firearms result in higher casualty counts or different incident types?
10. Are there noticeable correlations between incident locations (e.g., schools, business, homes) and outcomes such as the number of victims or type of violence?

Data Cleaning Summary

1. **Managing Missing Values**
    - Text-based fields such as (address, incident characteristics, participant status and participant type) were filled with "Unknown" to avoid losing records.
    - Numeric political district fields that were missing values were filled with unknown for cleaner visuals.
    - If a value was missing in the columns gun stolen, gun type, participant age group, and participant gender it was also filled in as unknown.
2. **Removing Columns**
    - The columns (participant relationships, location description, participant name, participant age) was dropped due to extensive missing values and limited usefulness for this stage of analysis.
3. **Optimizing Data Types**
    - The date column was converted to a proper datetime format
4. **Checking Data Integrity**
    - Verified that essential columns no longer contained missing values and that the structure was consistent.
5. **Saving the Cleaned Data**
    - The cleaned dataset was stored in a dedicated directory to maintain organization for subsequent analysis.