DAI 6.1 Sourcing Open Data

Data Source:

The dataset is sourced from Kaggle and provides a detailed record of over 260,000-gun violence incidents in the United States, spanning from 2013 to 2018. Each entry contains comprehensive information about individual incidents, including variables such as location, date, casualties, and contributing factors.

The primary purpose of this dataset is to facilitate research and analysis by data scientists and statisticians, enabling them to explore patterns, understand underlying causes, and make data driven predictions about future trends in gun violence.

The data can be access here:

https://www.kaggle.com/datasets/jameslko/gun-violence-data/data

Data Collection:

The data in this dataset was collected from publicly available records of gun violence incidents in the United States. These records include police reports, news articles, and official government sources, ensuring a reliable and diverse representation of incidents

The dataset aggregates information such as the date, location, number of casualties, and specific details about the individual involved, including demographics where available. It also categorizes incidents by factors such as type of gun violence (e.g., mass shooting, domestic violence) and contributing circumstances.

The rigorous collection methodology ensures a broad and accurate dataset, making it a valuable resource for exploring patterns and trends in gun violence over a five-year period.

Data Limitations:

While this dataset provides a comprehensive overview of gun violence incidents in the United States from 2013 to 2018, it is not without limitations. These include:

1. **Incomplete Reporting:** Some incidents may not have been reported or included due to variations in data collection methods across different sources, potentially leading to underrepresentation of specific cases.
2. **Missing or Inconsistent Data:** Certain variables such as demographic details or contributing factors, may be incomplete or inconsistently reported, which could affect the accuracy of detailed analysis.

3. **Geographic Bias:** Data may be disproportionately collected from areas with higher media coverage or more detailed reporting infrastructure, potentially skewing results.
4. **Lack of Context:** While the dataset captures incident details, it may lack broader contextual information such as socioeconomic factors, local laws, or historical trends, which are critical for a more nuanced analysis.
5. **Temporal Limitations:** The dataset covers incidents only up to 2018, which may limit its applicability for understanding current trends or predicting future incidents.

Why This Data:

Chose data as it was interesting to find out how much gun violence is in the world and how to address the gun problem. It is a sensitive topic since it is both a cause of violence but also a protector of said violence.

Ethical Considerations:

When working with sensitive data like gun violence incidents, it is essential to approach the analysis with care and responsibility to ensure ethical standards are upheld. Key ethical considerations include:

1. **Respect for Privacy:** Although the dataset is publicly available, it may contain details about individuals involved in gun violence incidents. It is crucial to avoid sharing identifiable information or using the data in ways that could further victimize individuals or communities.
2. **Avoiding Misrepresentation:** Data can be misinterpreted or manipulated to support biased narratives. Ensuring accurate, objective analysis is critical to avoid spreading misinformation or perpetuating stereotypes.
3. **Avoiding Misrepresentation:** Data can be misinterpreted or manipulated to support biased narratives. Ensuring accurate, objective analysis is critical to avoid spreading misinformation or perpetuating stereotypes.
4. **Purposeful Use of Data:** This data should be analysed with the intention of contributing to constructive discussions, research, or policymaking. It should not be used for sensationalism or exploitation.
5. **Acknowledging Limitations:** Ethical responsibility also involves transparently communicating the dataset's limitations to avoid overstating findings or drawing conclusions that the data cannot fully support.

Questions to Explore:

Temporal Trends:

1. How have gun violence incidents changed over time from 2013 to 2018?
2. Are there specific months or seasons with higher frequencies of incidents?

Geographic Patterns:

1. Which states or cities have the highest and lowest rates of gun violence?

2. Are there regional trends or hotspots of gun violence?

Incident Characteristics:

1. What are the most common types of gun violence (e.g., domestic violence, mass shootings)?
2. How do the number of casualties vary across different types of incidents?

Demographic Insights:

1. Are there trends in the ages, genders, or races of victims or perpetrators?
2. How do demographic factors correlate with the severity of incidents?

Correlating Factors:

1. Are there correlations between socioeconomic factors (e.g., poverty rates, unemployment) and gun violence?
2. Does gun violence correlate with proximity to certain types of locations (e.g., schools, urban centres)?

Policy and Prevention:

1. Did specific legislative changes during this period affect the frequency or severity of incidents?
2. Can patterns in the data inform targeted prevention strategies?

<u>Data Cleaning Summary:</u>

1. **Missing Values:**
    a. Categorical/Text Columns: Filled missing values in columns like address, source_url, incident_characteristics, notes, participant_status, and participant_type with "Unknown" to retain as much information as possible for analysis.
    b. Numerical Columns: For columns like congressional_district, state_house_district, and state_senate_district, replace missing values with -1 as a placeholder to indicate unavailable data.
2. **Optimizing Data Types:**
    a. Converted categorical columns (e.g., state, city_or_county, gun_stolen) to category type for efficient storage and processing.
    b. Ensured numerical columns were properly typed, reducing memory usage
3. **Saved Cleaned Dataset**