Answer 6.1

Data Source

This dataset, sourced from Kaggle, documents more than 260,000 gun violence incidents across the United States from 2013 to 2018. It provides detailed information for each case, including location, date, casualties, and contributing factors. Designed to support research and analysis, the dataset serves as a valuable resource for data scientists and statisticians to uncover patterns, investigate root causes, and develop data-driven forecasts regarding future trends in gun violence.

The data can be access here:

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

Data Collection

This dataset compiles information from publicly accessible records of gun violence incidents in the United States, drawing from sources such as police reports, news articles, and government data to provide a reliable and comprehensive account of events. It includes details such as the date, location, number of casualties, and demographic information about those involved, where available. The dataset also classifies incidents by categories like type of gun violence (e.g., mass shootings, domestic violence) and contributing factors. Its meticulous collection process ensures a wide-ranging and accurate resource, ideal for analyzing patterns and trends in gun violence over a five-year span.

Data Limitations

Although this dataset offers a detailed record of gun violence incidents in the United States from 2013 to 2018, it has several limitations:

- **Incomplete Coverage**: Some incidents might be excluded due to differences in data collection practices, leading to potential underrepresentation of certain cases.
- **Data Gaps and Inconsistencies**: Variables such as demographics or contributing factors are sometimes incomplete or inconsistently recorded, which could impact the precision of in-depth analyses.
- **Geographic Imbalance**: The data may disproportionately reflect areas with greater media attention or more robust reporting systems, possibly introducing bias.
- **Limited Context**: While the dataset includes incident details, it does not account for broader contextual factors such as socioeconomic conditions, local legislation, or historical trends, which are essential for a deeper understanding.
- **Time Restriction**: Covering incidents only through 2018, the dataset may not reflect recent trends or serve as a basis for analyzing future patterns.

Why this data

I chose this dataset because gun violence is a critical and complex issue that impacts countless lives. As someone passionate about leveraging data to address real-world challenges, I was drawn to this topic to uncover meaningful insights and contribute to potential solutions.

With over 260,000 recorded incidents, the dataset provides a rare opportunity to analyze trends and patterns in gun violence, helping to better understand its causes and effects. Its detailed variables, such as location, date, and contributing circumstances, enable a nuanced exploration of the factors behind these incidents.

This dataset also aligns with my personal mission to use data analysis to address pressing societal issues. By studying this information, I aim to contribute to conversations about public safety and policymaking while developing my skills as a data analyst.

Ethical Considerations

Analyzing sensitive data, such as records of gun violence incidents, requires a thoughtful and responsible approach to uphold ethical standards. Key considerations include:

- Protecting Privacy: Even though the dataset is publicly accessible, it may contain information about individuals involved in these incidents. Care must be taken to avoid revealing identifiable details or using the data in ways that could harm individuals or communities.
- Ensuring Accurate Representation: Data should not be misinterpreted or manipulated to promote biased narratives. Maintaining objectivity and accuracy is essential to prevent the spread of misinformation or reinforcement of harmful stereotypes.
- Communicating with Sensitivity: Gun violence is an emotionally charged subject. Findings from the analysis should be shared with respect and compassion, recognizing the potential impact on affected individuals and communities.
- Purpose-Driven Analysis: The data should be used to support meaningful discussions, research, or policymaking, rather than for sensationalism or exploitative purposes.
- Transparency About Limitations: Ethical analysis involves clearly communicating the dataset's limitations to avoid overstating results or making claims that the data does not fully support.

Questions to Explore

Temporal Trends:

- How have gun violence incidents evolved from 2013 to 2018?
- Are certain months or seasons associated with higher rates of incidents?

Geographic Patterns:

- Which states or cities experience the highest and lowest levels of gun violence?

- Are there identifiable regional trends or concentrated hotspots of gun violence?

Incident Characteristics:

- What are the most frequent types of gun violence, such as domestic violence or mass shootings?
- How do casualty numbers differ across various types of incidents?

Demographic Insights:

- Are there observable trends related to the ages, genders, or races of victims or perpetrators?
- How do demographic factors relate to the severity of incidents?

Correlating Factors:

- Is there a relationship between socioeconomic factors (e.g., poverty, unemployment) and gun violence?
- Does proximity to specific locations, such as schools or urban centers, influence gun violence trends?

Policy and Prevention:

- Did legislative changes during this time period impact the frequency or severity of incidents?
- Can insights from the data help shape targeted strategies for prevention?

Data Cleaning Summary

1. **Handling Missing Data:**

- **Categorical/Text Columns:** Missing values in columns like address, source_url, incident_characteristics, notes, participant_status, and participant_type were replaced with "Unknown" to preserve as much information as possible for analysis.
- **Numerical Columns:** Missing values in fields such as congressional_district, state_house_district, and state_senate_district were filled with -1 as a placeholder to indicate unavailable data.

2. **Removing Irrelevant Data:**

- The participant_age column was dropped due to extensive missing data and the complexity it presented, which made it less relevant for the current analysis.

3. **Optimizing Data Types:**

- Columns with categorical data, such as state, city_or_county, and gun_stolen, were converted to the category type to enhance storage efficiency and processing speed.
- Numerical columns were appropriately typed to minimize memory usage.

4. **Ensuring Data Integrity:**

- Verified that there were no remaining missing values in key columns.

- Confirmed that the dataset was well-structured and ready for analysis.

5. **Saving the Processed Dataset:**

- Stored the cleaned dataset in a designated directory, ensuring it is well-organized and easily accessible for further exploration and analysis.