Data Immersion

Exercise 6.1

Summary of Dataset

The dataset comes from the Gun Violence Archive. It contains ~239,000 rows and 29 columns with incident level details such as:

- Incident info: ID, date, state, city/county, address.
- Casualties: number killed, number injured.
- Weapons: number of guns involved, type, stolen/not stolen.
- Location: latitude, longitude, congressional/state districts.
- Participants: age, gender, role (victim/suspect), relationship, status.
- Textual info: incident characteristics, notes, and source links.

This data provides both event-level context (e.g., location, weapon count) and participant-level demographics (e.g., age, gender, role).

Why This Dataset Was Chosen

This dataset was chosen because it:

1. Real-world relevance: Gun violence is a critical social issue in the U.S. and this dataset captures detailed, large-scale information.
2. Supports diverse analysis: It allows exploration of incident severity, demographic patterns, temporal and geographic trends.
3. Policy relevance: The dataset can provide insights useful for public safety strategies and prevention programs.

Ethical considerations:

1. Participant info (age, gender, relationship) raises privacy concerns if misused even though it is anonymized.
2. Bias risk: The dataset only reflects reported cases in the U.S. so unreported or misclassified incidents are excluded.

Limitations:

1. Many **missing values**: e.g., participant age, gun type, relationship.
2. Some columns contain **outliers or unrealistic values** (e.g., extreme ages, 400+ guns in one event).
3. Highly **skewed distributions**: most incidents involve zero or one death/injury with few extreme outliers.

4. Text-heavy columns (incident characteristics, notes) require preprocessing to be usable.

Basic Descriptive Analysis

1. Incidents: ~239,000 total.
2. Casualties: Average ≈ 0.25 killed, 0.49 injured per incident; most events result in 0 or 1 casualty.
3. Weapons: Majority involve only 1 gun very few incidents involve more than 5.
4. Geography: 51 states represented
5. Participants: Ages range widely with most falling between teen to adult categories. Gender is mostly male, but many missing values.

Questions to Explore

1. Which states record the highest number of gun violence incidents overall?
2. Which states have the highest total deaths and highest total injuries?
3. When adjusted for population which states are the most dangerous in terms of gun violence per capita?
4. Are there regional differences (South vs Northeast vs Midwest vs West) in the frequency and severity of incidents?
5. Do certain states show higher proportions of mass shootings (multiple deaths/injuries)?
6. How do state-level patterns change over time (yearly trends in incidents, deaths and injuries)?
7. What is the distribution of casualties across incidents?
8. Is there a relationship between the number of people killed and the number injured?

Data cleaning summary

1. Handling Missing Values: Many NaNs (e.g., participant info, gun_type). Strategy: keep “Unknown” categories for categorical variables; drop high-missing columns like participant_relationship.
2. Data Type Conversions: Converted date to datetime
3. Outlier Detection: Extreme values in n_killed, n_injured, n_guns_involved. For analysis, these will be flagged rather than removed as they reflect rare but important mass events