Task 6.1: Sourcing Open Data
Ronda Austin

Data Source

The Data Source Information

The Data set is *Gun Violence Data*. The dataset was obtained on Kaggle.com.
Gun Violence Data (kaggle.com)

The data is all recorded gun violence incidents in the US between January 2013 and March 2018.

Data Collection

The data was downloaded from gunviolencearchive.org, by James Ko, using scraping methods.
The License is Data files © Original Authors

From the organization's description:

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. GVA will collect and check for accuracy, comprehensive information about gun-related violence in the U.S. and then post and disseminate it online.

Data Contents

There are 29 Variables in the dataset. These include:

Incident_id: number identifying the crime

Date: the date of the crime

State, city or county, address: The physical location of the crime

n_killed, n_injured: the number of persons killed or injured

Incident_url: The url regarding the incident

Source_url: The reference to the reporting source

incident_url_fields_missing: TRUE if the incident_url is present, FALSE otherwise

congressional_district: congressional district id

gun_stolen: status of guns involved in the crime (ie, unknown, stolen, etc)

gun_type: Typification of guns used in the crime

incidents_characteristics: Characteristics of the incidence

latitude: Location of the incident

location_description:

longitude: Location of the incident

n_guns_involved: Number of guns involved in incidentnotes: Additional information of the crime

participant_age: Age of participant(s) at the time of crime

participant_age_group: Age group of participant(s) at the time crime

participant_gender: Gender of participants involved

participant_name: Name of participants involved

participant_relationship: Relationship of participant to other participant(s)

participant_status: Extent of harm done to the participant

participant_type: Type of participant

sources: Participants source

state_house_districts: Voting house district

state_senate_districts: Territorial district from which a senator to a state legislature is elected.

Data Choice

I chose this data because gun violence and gun control are of interest to me. I was born in the US into a family of avid gun owners and collectors. Being well versed and competent in the handling of firearms was a requisite of living in my father's home. I repeatedly balked at the presence of guns and am the only person in my family who does not believe in possessing firearms. After living for several decades in a European country, where people do not often own or possess guns, I am more certain that education is needed in the US, to provide an understanding and prevention of gun violence.

Data Profile

Data Cleaning

- I inspected the data. I created a new dataframe without 16 columns that I deemed unnecessary for analysis. The columns are address, incident_url, incident_url_fields_ missing, source_url, congressional_district, gun_stolen, gun_type, latitude, location_description, longitude, notes, sources, participants_name, participants_relationship, state_house_district, state_senate_district.

- I looked for missing values and found all columns necessary for analysis contained no missing values. There were missing values in incident characteristics, but I found leaving them in the data set will not affect analysis.

- I addressed mixed data types and converted the following variables to string: Incident_id, incident_characteristics, participant_age, participant_age_group, participant_gender, participant__status, participant_type, n_guns_involved.

- I found that there were no duplicates in the data.

- I looked for outliers. I found no obvious outliers. The number_guns_involved had many instances of a large number of guns involved. I will not be using this data in my analysis as it contained no deaths and no injuries and no shootings.

- I changed the name of the following columns: n_killed changed to number_killed, n_injured changed to number_injured, and n_guns_involved to number_guns_involved.

- I checked the dataframe and exported the csv file to Prepared Data.

Descriptive Statistics

	number_killed	number_injured	number_guns_involved
count	239677	239677	140226
mean	0.252290	0.494007	1.372442
std	0.521779	0.729952	4.678202
min	0.000000	0.000000	1.000000
25%	0.000000	0.000000	1.000000
50%	0.000000	0.000000	1.000000
75%	0.000000	1.000000	1.000000
max	50.000000	53.000000	400.000000

Variable Profile

incident_id	Invariant	Qualitative	Nominal	structured
date	Invariant	Qualitative	Ordinal	structured
state	Invariant	Qualitative	Nominal	structured
city_or_county	Invariant	Qualitative	Nominal	structured
number_killed	Invariant	Quantitative	Discrete	structured
number_injured	Invariant	Quantitative	Discrete	structured
incident_characteristics	Invariant	Qualitative	Nominal	unstructured
number_guns_involved	Invariant	Quantitative	Discrete	structured
participant_age	Invariant	Quantitative	Discrete	structured
participant_age_group	Invariant	Qualitative	Ordinal	structured
participant_gender	Invariant	Qualitative	Binary/Nominal	structured
participant_status	Invariant	Qualitative	Ordinal	structured
participant_type	Invariant	Qualitative	Nominal	structured

Data Ethics and Limitations

The data could be prone to human errors because it was collected manually. Errors and missing data could also occur from scraping used for collection purposes.
For ethical reasons, for the sake of privacy to persons involved, I removed the names and addresses of participants involved in gun violence data.

Questions to Explore

- What are the total number of people killed and injured in the US from gun violence?
- Which states have the most gun violence.
- Which city/counties have the most gun violence? How does this compare to the states with the most violence from guns?
- What age ranges are involved in gun violence?
- Which months and days of the week have the highest rate of gun violence?