Achievement 6 Project Information

**Data Source**

Dataset: [Gun Violence Data](https://www.kaggle.com/datasets/jameslko/gun-violence-data)

Data Collection Source: <https://www.gunviolencearchive.org/methodology>

* **Data Source:** The data set is an external data source as it is data collected from [gunviolencearchive.org](http://gunviolencearchive.org/), a non-profit organization that aims to provide public, accurate information about gun-related violence in the US.
* **Data Collection:** The data set was generated using web scraping techniques with Python for each date between 1/1/2013 and 3/31/2018. The incident data from GVA is collected using a combination of automated queries and manual research through over 7,500 sources such as local and state police, media, data aggregates, etc.
* **Data Contents:** The data set includes one entry per incident and includes the following information: date, location, number of people killed, number of people injured, number of guns involved, participant age and age group, participant gender, participant name, participant relationship to other participants, participant status, state house and senate districts, and the GVA URL providing further details as well as a reporting source URL.
* **Limitations:** Each incident is annotated to its associated cause (hate crimes, domestic violence, robbery, etc.) within the incident URL, but those causes are not a clear variable in the actual dataset. In addition, GVA’s definition of gun violence is “intended to be fully inclusionary of disparate elements of gun related incidents…in that, all types of shootings are included, whether OIS, accidental, children shooting themselves, murders…and everything else”. This means that the incidents can only be analyzed as overall gun-related violence rather than specific types of gun crime.

I have chosen this data set as I have done previous research on gun violence, specifically school shootings, and am passionate about the need for stricter gun control laws in the US. Although I am not specifically seeking out a position as an analyst for this topic, I am interested in the use of data analysis within the political/social sphere.

**Data Profile**

Cleaning the Dataset:

* Basic cleaning and consistency checks in Jupyter:
  + Removed irrelevant columns, columns containing PII, and those with unclear or a lot of missing values
  + Checked for mixed type data and duplicates, none found
  + Removed records with missing values
* Changing format of variables into flags:
  + Changed the following columns for readability: gun\_stolen, incident\_characteristics, participant\_age, participant\_age\_group, participant\_gender, participant\_type

Understanding the Dataset:

* Reviewing variables and performing basic descriptive statistical analysis
  + Basic statistical analysis conducted in Jupyter
* Data Profile of Raw Data

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Column** | **Column Description** | **Time Variant** | **Structured/ Unstructured** | **Qualitative/Quantitative** | **Nominal/Ordinal or Discrete/Continuous** |
| incident\_id | unique code assigned to incident | no | structured | qualitative | ordinal |
| date | date of incident | no | structured | qualitative | ordinal |
| state | state in which incident occurred | no | structured | qualitative | nominal |
| city\_or\_county | city or county in which incident occurred | no | structured | qualitative | nominal |
| address | address at which incident occurred | no | structured | qualitative | nominal |
| n\_killed | number of victim killed in incident | no | structured | quantitative | discrete |
| n\_injured | number of victims injured in incident | no | structured | quantitative | discrete |
| incident\_url | GVA url of incident | no | structured | qualitative | nominal |
| source\_url | url of reporting source | no | structured | qualitative | nominal |
| incident\_url\_fields\_missing | boolean indicating if there are any fields missing within the GVA ur | no | structured | qualitative | binary |
| congressional\_district | congressional district number in which incident occurred | no | structured | qualitative | ordinal |
| gun\_stolen | free text indicating if each gun involved in incident was stolen, not stolen, or unknown | no | structured | qualitative | nominal |
| gun\_type | free text describing type of gun | no | unstructured | qualitative | nominal |
| incident\_characteristics | free text describing categories of incident | no | unstructured | qualitative | nominal |
| latitude | latitude of incident | no | structured | qualitative | ordinal |
| location\_description | description of the location of the incident | no | unstructured | qualitative | nominal |
| longitude | longitude of incident | no | structured | qualitative | ordinal |
| n\_guns\_involved | number of guns involved in incident | no | structured | quantitative | discrete |
| notes | free text with any additional notes | no | unstructured | qualitative | nominal |
| participant\_age | free text listing the ages of each participant | no | unstructured | qualitative | ordinal |
| participant\_age\_group | free text listing the age groups of each participant | no | unstructured | qualitative | nominal |
| participant\_gender | free text listing the gender of each participant | no | unstructured | qualitative | nominal |
| participant\_name | free text listing the name of each participant | no | unstructured | qualitative | nominal |
| participant\_relationship | free text listing the relationship between each participant | no | unstructured | qualitative | nominal |
| participant\_status | free text describing the status of each participant | no | unstructured | qualitative | nominal |
| participant\_type | free text describing the type of each participant (victim or subject/suspect) | no | unstructured | qualitative | nominal |
| sources | url of any additional sources | no | structured | qualitative | nominal |
| state\_house\_district | state house district number of incident | no | structured | qualitative | ordinal |
| state\_senate\_district | state senate district number of incident | no | structured | qualitative | ordinal |

* Data Profile of Cleaned Data:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Column** | **Column Description** | **Time Variant** | **Structured/ Unstructured** | **Qualitative/Quantitative** | **Nominal/Ordinal or Discrete/Continuous** |
| incident\_id | unique code assigned to incident | no | structured | qualitative | ordinal |
| date | date of incident | no | structured | qualitative | ordinal |
| state | state in which incident occurred | no | structured | qualitative | nominal |
| city\_or\_county | city or county in which incident occurred | no | structured | qualitative | nominal |
| latitude | latitude of incident | no | structured | qualitative | ordinal |
| longitude | longitude of incident | no | structured | qualitative | ordinal |
| no\_injuries\_or\_deaths | Boolean flag that is true if no injuries or deaths occurred in the incident | no | structured | qualitative | binary |
| injuries | Boolean flag that is true if there were 1 or more injuries in the incident | no | tructured | qualitative | binary |
| n\_injured | number of victims injured in incident | no | structured | quantitative | discrete |
| deaths | Boolean flag that is true if there were 1 or more deaths in the incident | no | structured | qualitative | binary |
| n\_killed | number of victim killed in incident | no | structured | quantitative | discrete |
| n\_guns\_involved | number of guns involved in incident | no | structured | quantitative | discrete |
| stolen\_gun\_involved | flag that reads ‘Stolen’ if all guns involved were stolen, ‘Not Stolen’ if all guns involved were not stolen, and ‘Unknown’ if it is not known for any of the guns involved | no | structured | qualitative | nominal |
| all\_adults | Boolean flag that is true if all participants were adults (18+) | no | structured | qualitative | binary |
| adult\_participant | Boolean flag that is true if 1 or more participants were adults (18+) | no | structured | qualitative | binary |
| teen\_participant | Boolean flag that is true if 1 or more participants were teens (12-17) | no | structured | qualitative | binary |
| child\_participant | Boolean flag that is true if 1 or more participants were children (0-11) | no | structured | qualitative | binary |
| all\_female\_participants | Boolean flag that is true if all participants were female | no | structured | qualitative | binary |
| all\_male\_participants | Boolean flag that is true if all participants were male | no | structured | qualitative | binary |
| mixed\_gender\_participants | Boolean flag that is true if there were a mix of male and female participants | no | structured | qualitative | binary |
| all\_victims | Boolean flag that is true if all participants were also victims of the incident | no | structured | qualitative | binary |
| all\_suspects | Boolean flag that is true if all participants were suspects | no | structured | qualitative | binary |
| suspects\_and\_victims | Boolean flag that is true if there are a mix of victims and suspects | no | structured | qualitative | binary |
| mass\_shooting\_flag | Boolean flag that is true if the incident qualifies as a mass shooting (four or more shot and/or killed in a single event, not including the shooter) | no | structured | qualitative | binary |
| drive\_by\_flag | Boolean flag that is true if the incident qualifies as a drive-by shooting | no | structured | qualitative | binary |
| domestic\_violence\_flag | Boolean flag that is true if the incident involved domestic violence | no | structured | qualitative | binary |
| armed\_robbery\_flag | Boolean flag that is true if the incident was an armed robbery | no | structured | qualitative | binary |
| gang\_involvement\_flag | Boolean flag that is true if the incident involves gang affiliation | no | structured | qualitative | binary |
| drug\_involvement\_flag | Boolean flag that is true if the incident involves drugs | no | structured | qualitative | binary |
| accidental\_flag | Boolean flag that is true if the incident was an accidental shooting | no | structured | qualitative | binary |
| self\_inflicted\_flag | Boolean flag that is true if the incident included a self-inflicted shooting (not a suicide or suicide attempt) | no | structured | qualitative | binary |

Limitations and Ethics:

* In order to have a more focused dataset and analysis, only incidents in which a gun was possessed by a felon or prohibited person was kept in the dataset. In addition, all non-shooting incidents were removed.
* Flag columns were created for the following situations:
  + Replacing the free-text ‘incident\_characteristics’ variable
    - mass\_shooting, drive\_by, domestic\_violence, armed\_robbery, gang\_involvement, drug\_involvement, accidental, and self\_inflicted flags were created for the types of incidents
    - no\_injuries\_or deaths, injuries, and deaths flags were created for the types of incidents
  + Replacing the free-text ‘participant\_age’ and ‘participant\_age\_group’ variable
    - all\_adults, adult\_participant, teen\_participant, and child\_participant flags were created
  + Replacing the free-text ‘gun\_stolen’ variable
    - stolen\_gun\_involved flag was created
  + Replacing the free-text ‘participant\_gender’ variable
    - all\_female\_participants, all\_male\_participants, and mixed\_gender\_participants flags were created
* All PII data was removed to prevent the identification of victims and suspects.
* Analyzing gun violence data can have ethical implications, particularly when discussing potential causes of increased gun violence in certain areas. When discussing the results of the analysis, it will be important to keep in mind the economic, political, and social climates that can contribute to increased violence, particularly in lower income areas.

**Questions to Explore**

* Which areas have the most gun violence incidents where the weapon was possessed by a felon or prohibited person?
* Is there a relationship between incident location and number of people injured/killed?
* Are certain types of incidents (drive-by, accidental, domestic violence, etc.) more common in certain areas?
* Are incidents involving younger participants (children and/or teens) more common in certain areas?