ETL Project: Gun Violence

By Alex Chao, Chris Glessner, Evan Johnson, Joe Strawinski

### Overview

For this project, we used three datasets found on the online communities of Kaggle and data.world. From these datasets we extracted details on over 100 mass shootings, including number of injuries and fatalities, as well as location data, gun type and details on the shooter (name and age). In conjunction with the first dataset, we were able to exploit a much larger database of over 250,000 gun-related crimes in the U.S. to detect patterns around mass shootings. Additionally, we used city-level population data paired with the incidences of mass shootings and gun violence, to help form a representative depiction explaining a pattern of occurrences.

### Data Sources

1. Mass shootings in the U.S. from 1982-2019 (<https://data.world/awram/us-mass-shootings>)

*A collection of indiscriminate rampages in public places resulting in four or more victims killed by the attacker. It excludes shootings stemming from more conventional crimes such as armed robbery or gang violence.*

1. Gun violence data from 2013-2018 (<https://www.kaggle.com/jameslko/gun-violence-data#gun-violence-data_01-2013_03-2018.csv>)

*This project aims to change that; we make a record of more than 260k gun violence incidents, with detailed information about each incident, available in CSV format. The CSV file contains data for all recorded gun violence incidents in the US between January 2013 and March 2018, inclusive.*

1. City population from 2015 (<https://data.world/gmoney/us-city-populations/workspace/file?filename=US+City+Populations.xlsx>)

*Data includes every city name (over 6k), its population and state. Numbers are from August 2015. Source:* [*http://www.city-data.com/*](http://www.city-data.com/)

### Extract

Once we chose our topic for this project and found the data sources above, our first step was to extract the raw data from our downloaded files, two of which were in CSV format while the mass shootings data was an Excel file. Reading the respective CSV/Excel files, we input the data into Pandas dataframes to be further transformed.

We performed an initial review of the attributes and available data within each of the raw data files to determine the relevant information that we wanted to keep. Below illustrates the data structure that we had within each file and attributes to be extracted.

1. Mass shootings in the U.S. from 1982-2019 (<https://data.world/awram/us-mass-shootings>)
   1. Date
      1. summary
      2. extract name from summary (to be completed at a later time)
   2. Location
      1. extract city
      2. extract state
   3. fatalities
   4. injured
   5. weapon\_details
   6. age\_of\_shooter

1. Gun violence data from 2013-2018 (<https://www.kaggle.com/jameslko/gun-violence-data#gun-violence-data_01-2013_03-2018.csv>)
   1. date
   2. participant\_name
   3. city
   4. state
   5. N\_killed (renamed to fatalities)
   6. N\_injured (renamed to injured)
   7. gun\_type
   8. participant\_age
2. City population from 2015 (<https://data.world/gmoney/us-city-populations/workspace/file?filename=US+City+Populations.xlsx>)
   1. city
   2. Population

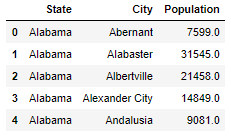
### Transform

Once we determined which data was relevant to be included in our database, we were able to begin cleaning the datasets to remove irrelevant variables. Our data transformation also included some additional steps to get the data in a consistent city/state format across datasets. Within the Mass Shootings dataset, this step included parsing location data using a split string method to break out city and state into two separate columns. One additional step included abbreviating state names for consistency.

Another step we took to transform the data included joining the cities dataframe to the state abbreviated dataframe which added the city populations. Moving from Figure 1 to Figure 2 required the addition of state abbreviations, which we added by appending a dictionary of abbreviations to the city population dataframe in Pandas. This allowed us to attach a city population column to the shooting databases, which we use to construct per capita gun violence measures.

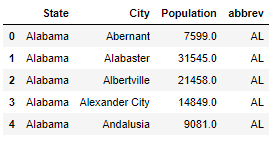
To generate per capita data: the gun violence dataset was grouped by City and State combination to find the number of incidents associated with each pair of City and State. This count data was merged onto the city population data set. Then, per capita = number of incidents / population.

#### **Figure 1: City Population with State**

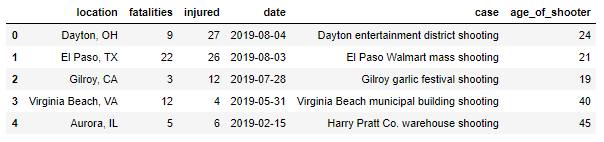


#### 

#### **Figure 2: City Population with State abbreviation**



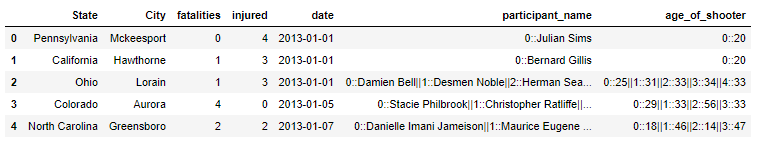
#### **Figure 3: Mass Shooting database**



#### **Figure 4:** **Mass Shooting database parse City / State**



#### **Figure 5: Gun Violence database**



1. Mass shootings in the U.S. from 1982-2019
   1. Location → parse city, state, convert state to full name
   2. Summary → parse participant name
   3. Date → Convert to string from TimeStamp object
2. Gun violence data from 2013-2018
   1. Participant\_name → parse first and last name

### Load

The last step in our project was to transfer our final output into a DataBase. We created the final tables and database to match the columns from the pandas.DataFrame using PostgreSQL Each respective final DataFrame was then uploaded to the database using SQLAlchemy to load the results. We laid out the structure of our database by defining a schema within the file “ETL schema.sql”.

### Future Work and Improvements

Heat maps by geographic location

Identify records in Mass Shootings and Gun Violence datasets that represent the same event. We attempted to merge these datasets on date and participant name, however we were unable to achieve this goal due to time constraints.

Utilizing the larger gun violence database to create aggregate measures for gun violence at the city level joining by city to the mass shootings data on date and participant name, we could then be able to tie in some other basic city level stats like population (per capita) to create a database to further analyze occurences of gun violence and mass shootings.