Extract, Transform, Load

Gun Violence and Gun Lobbying

# Introduction

While looking for data to use for this project, I stumbled across this dataset on Kaggle: <https://www.kaggle.com/jameslko/gun-violence-data>, an archive of gun violence incidents in the US from 2013 to 2018. James Ko compiled this data by scraping the Gun Violence Archive. The Archive does allow downloads, but it limits the number of results that can be returned from any specific query, and also limits the fields that can be downloaded.

From there, I found the Center for Responsive Politics’ site [OpenSecrets.org](http://OpenSecrets.org), which tracks lobbying and donations. Specifically, I was interested in which members of Congress had received donations from the NRA. I settled on this page, <https://www.opensecrets.org/industries/summary.php?ind=Q13&recipdetail=H&sortorder=A&mem=Y&cycle=2018>, which lists members of the US House of Representatives who have received pro-gun donations and the amount they received in 2018.

I wanted to connect the donations to the gun violence, so I decided to use the Sunlight Foundation’s congressional API, which allows you to search for a congressional district with the congressperson’s name. Unfortunately, this API was taken over piecemeal by a number of other organizations in 2017 and lost this functionality in the process. So I investigated various other APIs, such as Geocodio and the Google Civic Information API, but I eventually settled on the ProPublica Congress API (the direct successor of the Sunlight Foundation API).

# Extract

* The Kaggle data was in CSV format already, so I just read it into a Pandas dataframe.
* The [OpenSecrets.org](http://OpenSecrets.org) website rejected Pandas’s ‘read\_html’ command, so I had to use Splinter to navigate to the site and download the HTML . Then I was able to read it into Pandas.
* The ProPublica API returned data in JSON format, so I looped through the data with a list comprehension to create a list of tuples and and then converted to a dataframe.

# Transform

* The Kaggle data was mostly clean. I converted the ‘date’ column to datetimes, and I used the uszipcode package to convert the full states’ names into state abbreviations.
* The OpenSecrets data was in tabular format, but there were multiple fields strung together in the first column. I split that column using multiple delimiters, stripped trailing whitespace, and removed unnecessary symbols.
* The ProPublica data was mostly clean once it was turned into a dataframe.

# Load

* I created a ‘gun\_db’ database in MySQL and created three tables: ‘gun\_violence’, ‘rep\_gun\_donations’ and ‘rep\_district’. Each table had a primary key, and the ‘rep\_district’ table had a foreign key that referenced the ‘rep\_gun\_donations’ table.
* I had trouble with the encoding when uploading the Kaggle data, so I followed a Stack Overflow post to strip non-ASCII characters from the table. Then it uploaded with no problem.
* The other two tables loaded into MySQL with no hiccups.