Teresa Ferrill  
DSC540-T301, Data Preparation  
Bellevue University

I have decided to work with data regarding gun violence, including shootings, murders, and suicides. The data sources I plan to use include:

* CSV data source – Gun Deaths in the US: 2012 – 2014
  + Link <https://www.kaggle.com/hakabuk/gun-deaths-in-the-us>
  + Total of 100,798 rows, 10 columns
  + Data elements include year of incident, month of incident, intent, whether or not police were involved in the shooting, sex of victim, age of victim, race of victim, race, Hispanic, place where shooting occurred, and education status of the victim
* Website data source – Gun Violence Archive
  + Link: <https://www.gunviolencearchive.org/>
  + Provides information on gun violence (GV) deaths, including total number of deaths (all causes), number deaths by homicide/murder, number of deaths by mass shootings and mass murders, number of children killed and injured, officer involved incident killed and injured, suspect-suspect killed or injured, defense use, unintentional shooting, murder/suicides incidents
  + Website offers the ability to drill down by state
* API – Gun Violence Data, January 2013 to March 2018
  + Through Kaggle API: <https://www.kaggle.com/duttadebadri/gun-violence-in-usa-insights-forecast>
  + Total of 239,678 rows and 29 columns
  + Data elements include incident date, state and city of incident, number of people killed, number of injured, whether or not the gun was stolen, the gun type, number of guns used, and incident characteristics which include whether or not it was a mass shooting, murder, gang attack, etc.

The relationships that already exist between the datasets includes location, number of victims, officer involvement, and the type of shooting. I will have to perform some grouping to put these records together by city and/or state. I expect to have to replace headers, format some of the data, find duplicates, fix inconsistent values, and identify outliers and bad data as I transform and cleanse the CSV file. As I work with the website data, I expect to have to replace headers, format the data into a more readable format, find duplicates, fix inconsistent values, and conduct fuzzy matching. As I work with the API data source, I expect to replace headers, format data into a more readable format, identify outliers and bad data, find duplicates, and fix inconsistent values.

Once I merge the datasets, my expectation for my analysis is to be able to identify those metropolitan areas that have more gun violence than others. According to the Security.org website, the top cities for gun incidents between 2014 and 2018 were Ocala, FL; Orlando FL; Tuscaloosa, AL; Tulsa, OK; and Greenville, SC (Security.org, 2019). My expectations are that there would be more violence in places like Los Angeles, CA; Chicago, IL; New York City, NY; and Washington D.C. I look forward to discovering additional relationships between the data sources along with the differences or disparities from location to location.

Reference

Security.org. (2019). *City Crimes Inolving Guns*. Retrieved from Security.org: https://www.security.org/resources/city-crimes-involving-guns/