Data Exploration: Managerial Section

The data that is needed to conduct our analysis on gun violence in the United States of America was spread out across five datasets. Our main dataset included information on all the gun violence that occurred in the USA from 2013 to 2018. This dataset included 29 variables and 239,678 records. Our remaining datasets are related to socioeconomic factors like median household income, percentage of population below the poverty line, high school graduation percentage, and percentage of race per city. Each of the socioeconomic datasets included 3 variables and 29,322 records.

The way we had decided to conduct our analysis was with our data being at the city level; all our datasets contained this level of granularity. The main issue that we uncovered was that there was no clear link to join our socioeconomic datasets to our main gun violence dataset. The one common element between them was that our main gun violence dataset contained the city name, state, latitude, and longitude of each city. We had initially considered joining on city name; however, the city names were not consistent between each dataset which made it difficult to join on city name. As a result, we tried to join on latitude and longitude. Given that our socioeconomic datasets had the name of each city, we used the geography datatype in Excel to produce the latitude and longitude for each city. However, joining on both latitude and longitude was incredibly difficult. We used a left join, while using our gun violence dataset as our main dataset and the socioeconomic sets as the right. We decided to use a left join as we wanted to ensure that we included all the records from our main dataset and only wanted to include records from our socioeconomic dataset that matched. This join ended up creating 25 million rows of data, ultimately not working as intended. Next, we tried to join on city name and state name. However, this required extensive cleaning to try and use this as the join criteria. Once the data had been cleaned, we joined our datasets on city name and state; this yielded favourable results. While our join was successful, we were not able to join all our data together. At the end of this processes, we ended up with a dataset that contained all our original records and 213,671 records from our joined datasets.

