Predicting Counties with Higher/Lower Risk of a School Shooting Incident

Alex Bass, Connie Cui, Tulsi Ratnam

1. **Data**
    i. The dataset, created by the Center for Homeland Defense and Security, is a compilation of data obtained from different sources and consists of over 2000 school shootings from 1970 to present (2020). https://github.com/acbass49/school-shootings/blob/main/data_w_county_match.csv
    ii. Combine with census data. Our data will have features such as percent with Bachelor's or higher, percent greater than $75K household income, and percent by race, gender, and age. https://github.com/acbass49/school-shootings/blob/main/county_level_data_raw/co-est2021-alldata.csv
    iii. Combine law data related to gun laws by state, gun ownership / crime rate data. These will add features like percent of gun owners, strict or loose gun laws, etc. https://www.kaggle.com/datasets/mikejohnsonjr/united-states-crime-rates-by-county
    b. Because we are creating the dataset by researching and joining in other sources, it adds rarity and uniqueness which may give us a unique look into a complex problem. We see our dataset consisting of just over 2000 records of school shootings and plan to have 15 or more features of county level variables.
2. **Motivation**
    a. We intend to create predictions for the likelihood of school shootings by county. Recently gun violence has taken the stage in national news media attention and Congress is forming a bipartisan gun bill. There are many paths that could be explored in this data, but we are most interested in deepening our understanding of the types of counties in which gun violence is prevalent. Our report aims to answer questions like: Are rich or poor counties more susceptible to gun violence? Are gun violence incidents concentrated in particular regions in America? If so, which? How do state laws play a role in deterring gun violence? In the end, our motivation can be summed up in providing more context, information, and understanding in a major debate in American politics today.
3. **Foresight**
    a. First "bottleneck" is getting all of the data wrangled and joined in with the time constraints of the class. There is a lot of county-level data available, but which is the best to use and will we be able to obtain it and join it for our analysis.
    b. One consideration that might hold us back is our geospatial granularity. Counties can be large and there can be many different environments that exist in the same county. This analysis is not as useful, say, as a study that yields city or census tract level results which can parse out socio-economic and other differences more distinctly.
    c. We also may run into issues with some of our variables having missing data though I am hopeful we will learn about tools later in the class.