Mengyan Wu

US School Shooting

Motivation

School shooting is one of the most serious social problems in America. The loss of innocent lives brought pains and concerns to the whole society. While from 1970 to 2020, the frequency of shootings happened in schools is still raising. In the past ten months of 2021, a total of 176 shooting occurred in K-12 schools, which is already a 50% increase from last year[1]. As an international student, I am curious about the gun situations here. I would like to find out the factors related to the frequency of school shootings, in order to help to predict and prevent school shootings.

My specific research questions are:

I. How does school shooting frequency vary across the county in the US?
II. What time is more likely for school shootings to take place?
III. Who is more likely to be the target of school shootings?

Data Sources

I used two main data sources for the research:

1. The K-12 School Shooting Database from Naval Postgraduate School's Center for Homeland Defense and Security (CHDS): https://www.chds.us/ssdb/
   This dataset have multiple sub-datasets. The *SSDB_Raw_Data.csv* includes school shooting information (state USPS code, city name, date, school name) for shootings happened in K-12 schools from 1970 to 2020 and has 1853 raw data records in total. The *victim.csv* includes the information of victims and has 2844 rows in total. The sub-datasets could be linked through the *incident_id* field.

[1] The school shooting happened 113 times in 2020. Data source: https://www.chds.us/ssdb/

2. US Census Demographic Data on Kaggle, based on 2015 American Community Survey 5-year Estimates. https://www.kaggle.com/muonneutrino/us-census-demographic-data/version/3?select=acs2015_county_data.csv

   This dataset includes county FIPS code (unique for each county), county name, state name, economic information (median household income, unemployment rate, percentage of types of employment, etc.) for counties across the US. It has 3220 rows in total.

I used two additional data sources to link the two main data sources:

3. US place and county relationship file which links the state USPS codes and the city names with the corresponding county names: https://www2.census.gov/geo/docs/reference/codes/PLACElist.txt (stack overflow helped me to find this data source https://gis.stackexchange.com/questions/279123/find-mapping-of-place-codes-to-county-codes)
4. US state list from Github which links the state names with the state USPS codes: https://gist.github.com/dantonnoriega/bf1acd2290e15b91e6710b6fd3be0a53/

Methods

I. Analysis 1-3: How does school shooting frequency vary across the county in the US?

   ***** I re-used the data cleaning code of project part 1 in this part.***

   Since the two main datasets do not share any same column, I need to link the city name in the shooting dataset to the corresponding county FIPS code. Note that there are cities in different states but sharing the same names, so in the join process, I joined on city name and state USPS code to avoid duplicates. Besides, since there are cities in different counties but sharing the same name, I used county FIPS code which is uniquely assigned to each county as the identifier, instead of the county name.

   I dropped the cities which were shared by multiple counties because they are too complicated to deal with and only made a very small portion of the data. I also dropped the data which I could not find the corresponding counties or county FIPS codes.

   As a result, the final cleaned shooting dataset, which is exported and saved as *fips_shooting.tsv,* has 1233 records and is labeled by the county FIPS code.

Finally, I grouped the data by the county code and calculated the number of school shooting cases from 1970 to 2020 for each county. This table was exported and saved as *freq_shooting_by_fips.tsv.*

To prepare for the analysis, I used pandas to join the *freq_shooting_by_fips.tsv* dataset with the census 2015 dataset *acs* on county FIPS code. For counties with no previous shooting cases, I filled the number_of_shooting field with 0.

II. Analysis 4-7: What time is more likely for school shootings to take place?

I used the raw shooting dataset *SSDB_Raw_Data.csv.* I used pandas to drop the null data as well as the rows with incorrectly formatted date field. Then, I used the *to_datetime* method in pandas to extract year, month, day of the week and hour from the date field. I added these information as new columns to the dataset. Finally, I saved the dataframe as *time_of_shooting.*

III. Analysis 8-9: Who is more likely to be the target of school shooting?

I joined the raw shooting dataset *SSDB_Raw_Data.csv* with the raw victim dataset *victim.csv* by *incident_id.* I trimmed the blank spaces of the *Situation* field. I also dropped the null data. Then, I iterated through the data frame to categorize the the age field. (<=13: Child, <18:Teen, >=18:Adult).

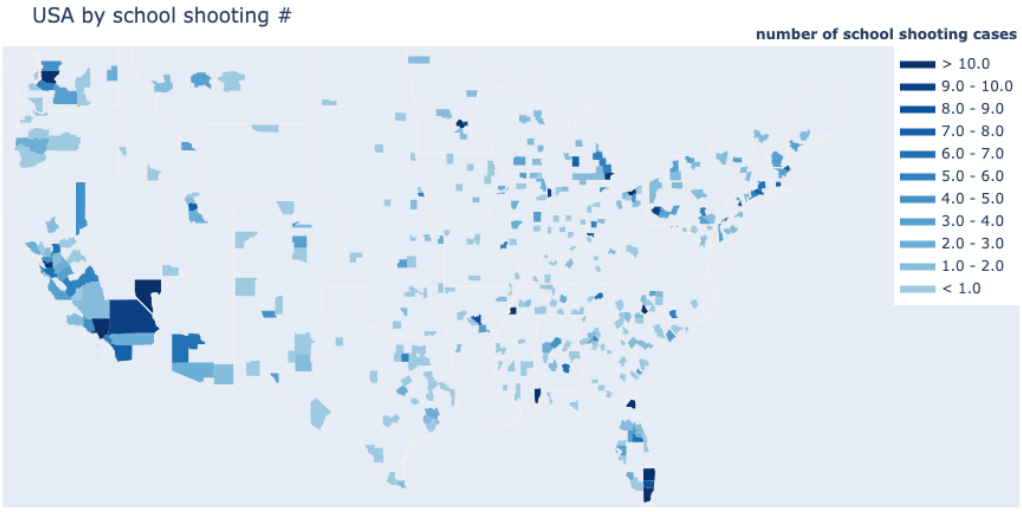
Analysis and Results

I. Analysis 1-3: How does the school shooting frequency vary by county in the US?

1. What is the distribution of school shooting cases by county?

   First, I sorted the *freq_shooting_by_fips.tsv* dataset by the number of school shootings in the past 50 years. The top 10 counties are Wayne, Los Angeles, Philadelphia, Duval, Shelby, Miami-Dade, Alameda, Hennepin, Cuyahoga, Mobile. I was surprised that the Wayne county, MI (where the Detroit city is located) had the highest frequency of school shooting across the country.

Then, to visualize the geographical distribution of school shootings, I continued to use *freq_shooting_by_fips.tsv* and plotted a geographic heat-map of school shootings by county with *geopandas* in python. (https://www.python-graph-gallery.com/choropleth-map-geopandas-python) The code is in the *geographic_heatmap.py* file.



The generated heat-map is shown above. Not all of the counties have shooting history, so there are blanks on this graph.

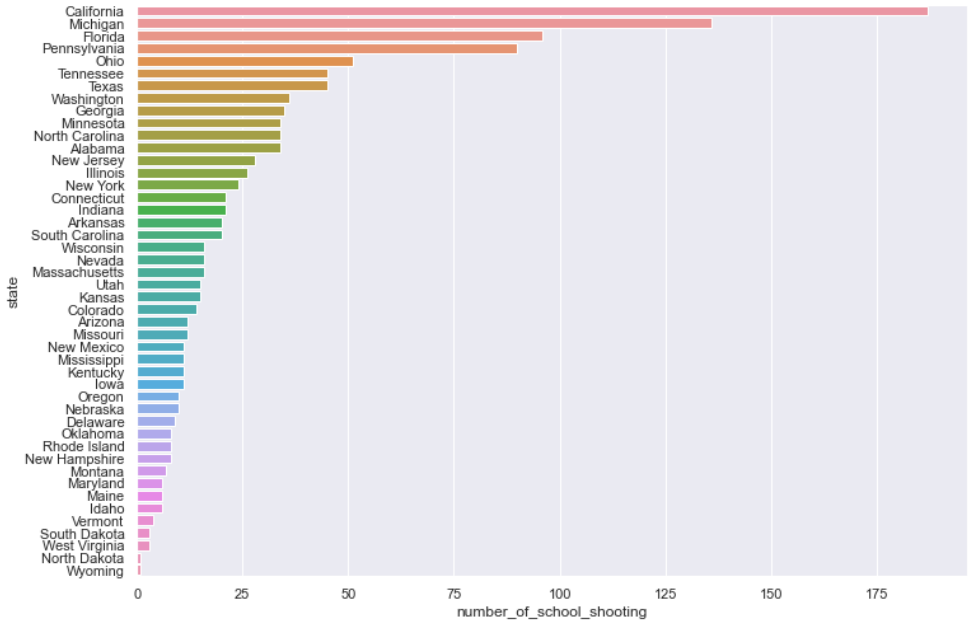
From the graph, we could notice that, coastal areas and the Great Lakes Region have more school shooting cases, while the middle-west area have less school shooting cases. By estimation, California, Washington and Florida seem to have more school shooting cases than other states from the graph, while I do want to analyze the distribution of school shooting cases by states in the following analysis.

2. What is the distribution of school shooting cases by states?

   I loaded the *fips_shooting.tsv* file with pandas and grouped the data by state to count the number of school shooting cases for each state. I used a bar-plot to visualize this relationship.

   From the graph I noticed that, the result is not exactly the same as my previous estimation from the heat-map. The reason might be that, the area of counties are different and counties with larger area naturally draw more attention from the reader in the heat-

map.



As the result, the top 10 states of school shootings are: California, Michigan, Florida, Pennsylvania, Ohio, Tennessee, Texas, Washington, Georgia and Minnesota.

3. What are the possible factors related to the variation of the school shooting frequency by county?

   A. Is median household income related to school shooting frequency?

      I used Ordinary Least Squares Regression to model the relationship between school shooting frequency and median household income. I set *number_of_cases* as the dependent variable, *income* as the independent variable. The OLS regression results shown that, the coefficient of the independent variable is -0.2677 and the p-value = 7.32e-05. Since the p-value << 0.05, we could conclude that counties with higher income tends to have less school shooting cases, and this relationship is statistically significant at 0.05 level.

   B. Is unemployment rate related to school shooting frequency?

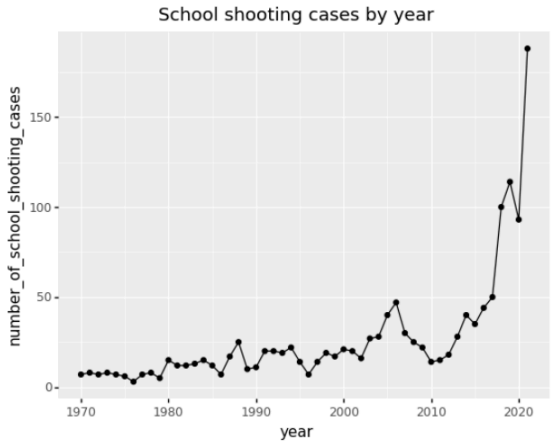
      I continued to use OLS to model the relationship between school shooting frequency and unemployment rate. the coefficient of the independent variable is 0.1346 and the

p-value = 0.00624. Since the p-value << 0.05, we could conclude that counties with higher unemployment rate tends to have higher school shooting cases, and this relationship is statistically significant at 0.05 level.

II. Analysis 4-7: What time is more likely for school shootings to take place?

4. school shooting by year

First, I used pandas to group the *time_of_shooting* dataset by year and to count the total number of school shooting cases for each year. Then I visualized the relationship with ggplot(plotnine) in python.



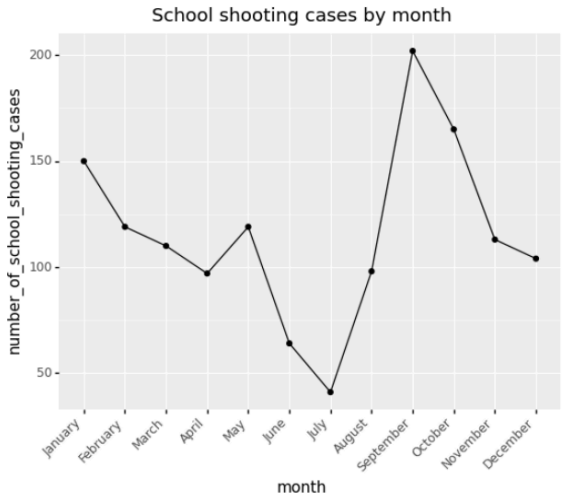
Generally speaking, there is a considerable upward trend in the number of school shooting cases in the past 50 years.

There is a drastically increase from 2017 till now. Although we have not yet come to the end of 2021, the number of school shooting cases this year has already been 2 times as many as the number of cases in last year.

Regardless of 2017-2021, in the plot, 1987-1989, 2005-2007 are all the local maximum. It is worth noting that these time periods are 1 year before or at the year of economy crises.

5. school shooting by month

The steps were basically the same except that I grouped and counted the data by month in this analysis.

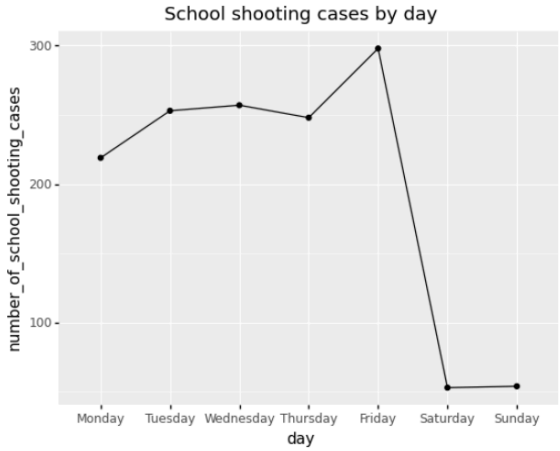


It is natural to have the minimum of cases in June and July, as these two months are the Summer Break and students are not at school. However, I notice that there is an increase of cases from July to August, and the case number reaches the maximum in September. Similarly, December has fewer cases as this month has the Christmas Break. However, immediately after that, January is the local maximum of cases.

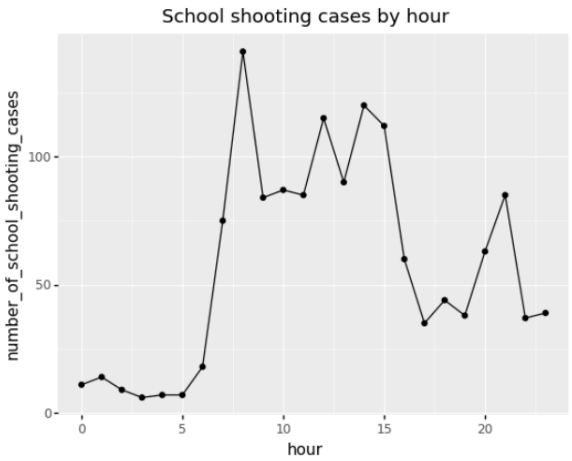
Both September and January are the months when students return to school from vacation to start a new semester. Sadly, these months are the most likely periods of school shooting cases. This pattern warns us that, at the beginning of the new semesters, we must pay extra attention to campus safety and students' mental health problems.

6. school shooting by day of the week

School shootings are less likely to take place on Monday and more likely to take place on Friday. Saturday and Sunday naturally have few cases as there is no class.



7. school shooting by hour of the day



Generally, school shootings are more likely to happen during the day(7am - 4pm).

The beginning of school days, 8 am, seems to be the most likely period of school shootings.

Another local maximum of the curve is 9 pm, indicating the most possible time period of school shooting cases at night.

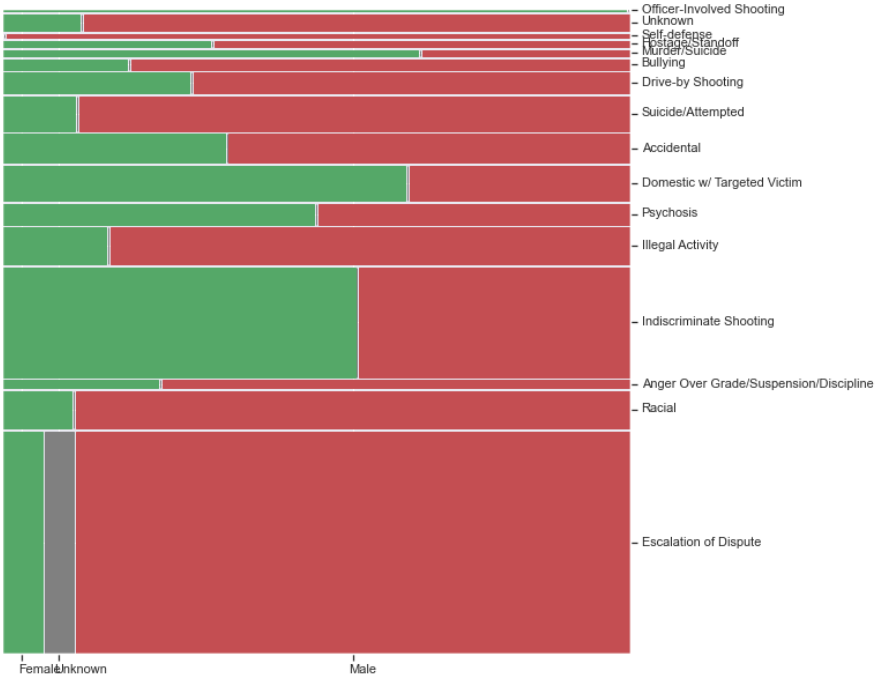
III. Analysis 8-9: Who is more likely to be the target of school shooting?

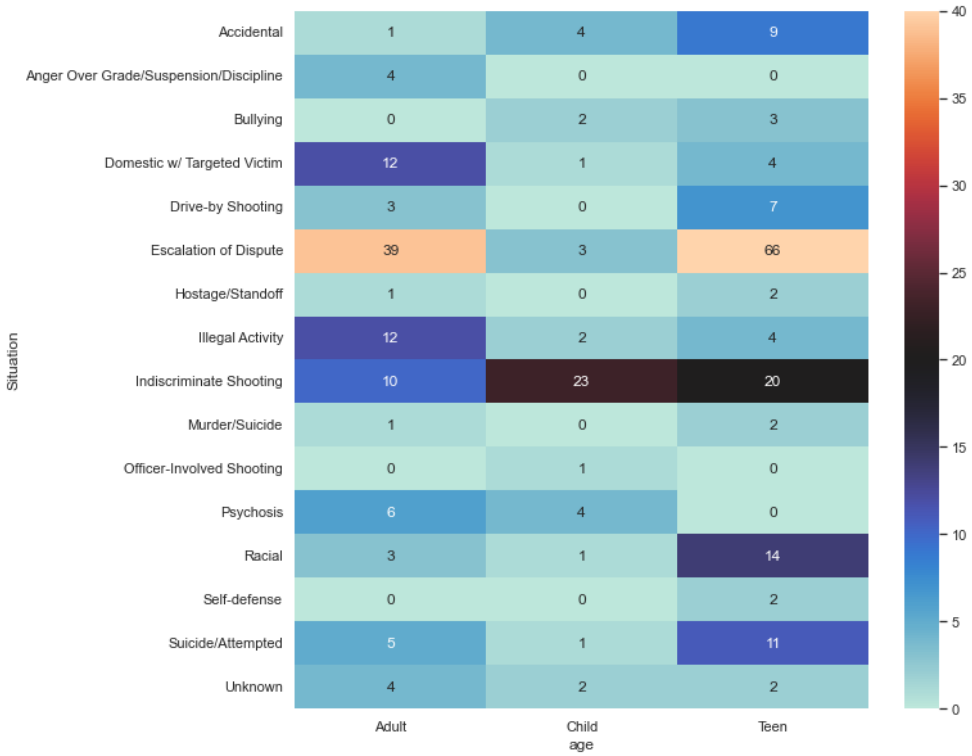
8. analysis of victims by sex under different situations (Accidental, Escalation of dispute, Drive-by shooting, etc.)

   I visualize the data of victims of different sex group under different situations with a mosaic plot.

   From the graph, I noticed that, females are more likely to be the victims of Officer-Involved Shooting, Murder/Suicide, Domestic/Targeted Victim and Indiscriminate Shooting; while males are more likely to be the victims of Self-defense, Hostage/ Standoff, Bullying, Drive-by Shooting, Suicide/Attempted, Accidental, Illegal Activity, Anger Over Grade/Suspension/Discipline, Racial and Escalation of Dispute.

Overall, males are more likely to be the target of school shootings.





9. analysis of victims by age group under different situations (Accidental, Escalation of dispute, Drive-by shooting, etc.)

   I visualize the data of victims of different sex group under different situations with a heat-map using seaborn.

   Children and teens are more likely to be the victims of indiscriminate shooting. Teens and adults are more likely to be the victims of targeted shooting, drive-by shooting, escalation of dispute and suicide. Teens are also likely to be the victims of accidental and racial shootings.

   Overall, teens are more likely to be the victims of school shootings.

Reference

1. I re-used the data cleaning code of project part 1 for analysis 1-3.