Problem Statement:

According to the Academy for Critical Incident Analysis, between 2000 and 2010, there were 57 incidents of school violence worldwide that had two or more victims. Twenty eight (almost half!) of those occurred in the United States alone. Over the past few decades, students and teachers have expressed a growing concern about safety at schools, specifically related to these incidents. There are a variety of opinions on how best to reduce or protect against these events: increase in security, frequent drills to prepare, mental health outreach, etc. However, for each suggestion there is always the question of how feasible it is in terms of resources (time and money) to implement nationwide.
This question lead to my analysis of school shooting incidents in the United States from 2010-2018. It is my purpose through this study to determine if there are early warning signs or common characteristics of areas with school violence in the United States. More specifically, I would be interested in the following questions:

1. Which variables have the largest positive or negative impact on the likelihood of school violence?
2. Is school violence more likely to be clustered in similar areas/ times, or more uniformly distributed?
3. What are the (if any) common characteristics of the schools or counties in which you find the majority of recent school violence?
4. What relationship (if any) is there between a county's overall adult/teen health and the prevalence of school shootings?

I believe that a number of clients would find the results of this analysis particularly interesting regardless of what they may believe is the best approach to reducing the occurrences of school violence and protecting students. By identifying early warning signs or predictive variables for school violence, these programs may be able to more efficiently utilize their resources to have a greater impact on the country as a whole. It is my hope that someone could use my investigation to more accurately target proactive interventions in areas flagged as having the highest risk for school violence.

Data Acquisition and Cleaning:

My analysis is based on multiple datasets from two sources:

1. The Washington Post's dataset records information on school violence since Columbine in 1998. It is easily downloadable as a .csv file from the link as well as clear documentation on the variables included.

2. County Health Rankings records information and ranks counties in the United States based on various characteristics from 2010 to current. It is also easily downloadable as a .csv file from the link with clear documentation on the variables included.

While the county health rankings had a variety of variables for which information was recorded, I opted to select a smaller sample for my initial investigation: average mentally unhealthy days per week, percent who report binge drinking, teen birth rate, percent uninsured, high school graduation rate, percent of adults with some college education, unemployment rate, percent of children living in poverty, percent of single-parent households, and the primary care physician ratio to the county's population. There were other variables that I was interested in including (such as access to healthy food, mental health support, etc.) but these were not consistently reported throughout the years.

The school shooting and the county health datasets were merged together using FIPS county codes which are unique to each county. Then, the following steps were performed to clean and prepare the data for analysis:

1. Dropped information:
    a. Columns: age_shooter2, gender_shooter2, race_ethnicity_shooter2, shooter_relationship2, shooter_deceased2, deceased_notes2, deceased_notes1, weapon_source
        i. The columns about a second shooter had only 2 entries, leaving 105 empty values in each.
        ii. Didn't feel that deceased_notes1 (how the shooter died) would add much value to later analysis
        iii. Weapon_source (where the student got the weapon) was mostly missing and I don't believe it will add a lot to later analysis
    b. Row for the school shooting at Vereen School on November 5, 2015
        i. This row was responsible for the only missing value in many columns (school demographics, shooter information, and time).
2. Missing values:
    a. 22 missing values for the age of the shooter
        i. Replaced with the median age of shooters
        ii. Chose median because there is a wide range of ages (from 6 to 53)
    b. Various numbers of missing values in columns state, school_type, shooting_type, gender_shooter1, race_ethnicity_shooter1, ulocale, day_of_week, and city
        i. Replaced all missing entries with string 'unknown' in preparation for converting to categorical variables.
    c. 2 missing times of the incidents
        i. For Stellar Leadership Academy, I replaced the missing value with 1:30 PM. I was able to find this rough approximation in a local news article written about the event.

ii. For Redland Middle School, I replaced the missing value with 11:00 AM. There was no time reported in any of the news articles, but they all mentioned that it happened during a student's class. So I chose the middle of the typical school day as the time.

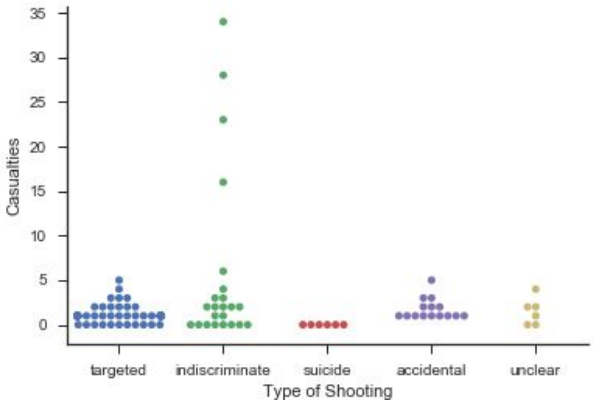
d. Various number of missing values for racial/ethnic demographics of the school
   i. Replaced missing counts for Rebound High School and Success Academy based on online reports.
   ii. Assuming all remaining missing ethnic counts (hawaiian native and two or more ethnicities) are actually 0

e. Missing values in staffing and lunch
   i. Replaced all missing values in staffing and lunch with 999 and 9999 respectively
      1. These represent unknown values while maintaining the columns as numeric dtypes

3. Correcting dtype of DataFrame columns
   a. Categorical
      i. Converted state, school_type, shooting_type, gender_shooter1, race_ethnicity_shooter1, shooter_relationship1, weapon, ulocale, day_of_week, and city
         1. shooter_relationship1 - There were 20 unique responses in this column that I re-grouped into student, former student, student relation, not a student, and unknown
         2. weapon - There were 20+ unique responses in this column as well so I regrouped them into generic type of gun (revolver, handgun (semi-automatic), shotgun, rifle, unknown
         3. shooting_type - Consolidated responses into fewer categories (targeted, accidental, indiscriminate, suicide, and unclear)
   b. Datetime
      i. Combined date and time columns into a new column labeled date_time. Then converted date_time column to a datetime type.
   c. Numeric
      i. Replaced kindergarten(KG) and pre-kindergarten(PK) in low_grade column to numeric values 0.5 and 0 respectively.
      ii. Then converted both low_grade and high_grade to float type.

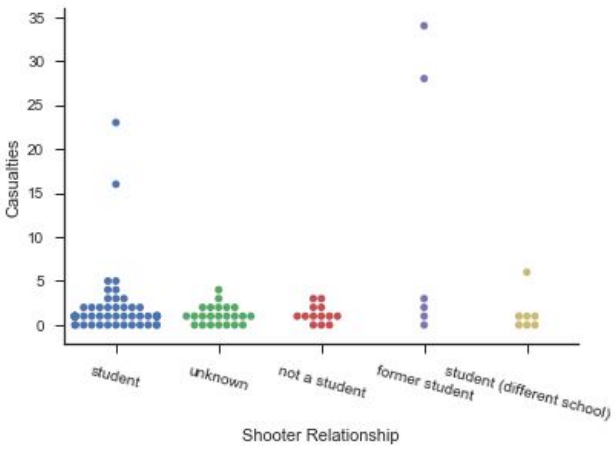
Initial Findings - Targeted vs. Indiscriminate

I began my analysis by investigating the breakdown of incidents into their type of school shooting: targeted, indiscriminate, suicide, accidental, or unclear/unknown. In the figure to the right, we can see that the targeted shooting types comprised the majority of incidents (with each dot representing a



single incident). While targeted shooting incidents were the most common, they had relatively low casualty rates per incident. Indiscriminate shooting types, while less common, had the most devastating impact with 4 of the incidents having more than 15 casualties. In fact, the indiscriminate shootings had 129 casualties altogether while the other types combined had 99. While the other types of shootings are equally important in terms of prevention and safety, I chose to focus on targeted and indiscriminate shooting types as they contained the vast majority of incidents as well as casualties.

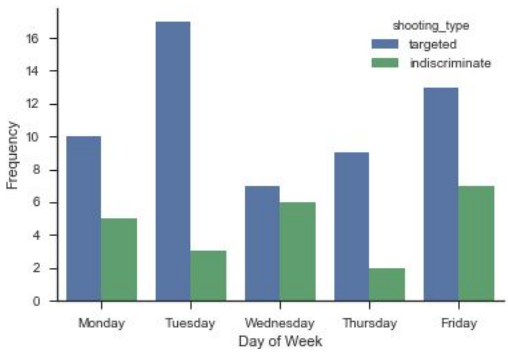
Shooter's Relationship to School

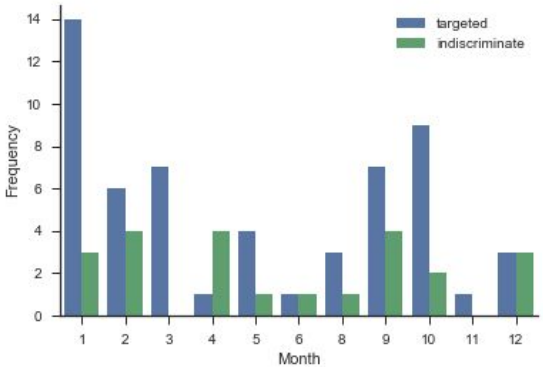


Next, I was interested in the relationship of the shooter to the school. As seen in the figure to the left, the majority of school shootings are perpetrated by a student, most of those by a current student of that school. And the incidents with the highest casualty rates are committed by a current or former student of that school. While any incident of school shooting is important and should be investigated thoroughly, it appears that the most common and devastating of these are perpetrated by students targeting an individual or individuals, or indiscriminately attacking individuals. If we look into these incidents more closely, what might we find?

Day and Month of Incident

There appears to be a difference between the two shooting types in relation to the day of the week that the shooting occurred on. Indiscriminate shootings appeared to occur independently of the day of the week, while the targeted shootings seemed to mostly favor Tuesday. However, after calculating a χ^2 goodness of fit statistic for both, there wasn't any significant statistical evidence to suggest that they were not evenly distributed. (Targeted: p-value = 0.2461, Indiscriminate: p-value = 0.4425)





There were, however, more substantial results when looking at the distribution of months for each type. For the most part, the results were as one would expect. There is a substantial dip during the summer months (June through August) while most schools are traditionally not in session. It also appears that

indiscriminate shootings (green bars) are fairly evenly distributed throughout the remaining months. However, it appears that targeted shootings are most likely to occur in January. In fact, almost 18% of all school shooting incidents during this time period occurred were targeted shootings in January!

There is statistically significant evidence through a χ^2 goodness of fit test to support the hypothesis that the months are not evenly distributed for targeted shooting types (Targeted: p-value = 0.0025, Indiscriminate: p-value = 0.8494). Upon further investigation, through a hypothesis test for proportions, there is also statistically significant evidence to support that January has an unusually higher proportion (z = 2.57, p-value = 0.0102) than you would expect the incidents were occurring independently of the month. Surprisingly, there were also two months that were substantially lower than expected: November (z = -4.824, p-value = 0.000001) and April (z = -4.824, p-value = 0.000001).

Clearly, there is something important about these months or the months leading up to it. Perhaps there is some relation to the traditionally big family holidays adding stress to a student's life, or the stress of grades being released over winter break?

County Health Statistics

Next, I searched for statistically significant differences between the various health characteristics per county. I continued on this track of comparing and contrasting targeted vs. indiscriminate shooting types to find that there may be some significant differences between the counties in which these occurred. Most notably, the counties with targeted school shootings had substantially higher teen birth rates, uninsured individuals, and the number of children in poverty.

Lastly, I compared counties with higher casualty rates (greater than 1) and counties with lower casualty rates (1 or lower) per each incident to find that there is little evidence to suggest they are different with respect to the variables I included. However, there was one substantial difference: self-reported rates of excessive drinking. I found it particularly interesting that there didn't appear to be a difference in the other variables that many people would attribute to an increase in drinking: mentally unhealthy days, unemployment, etc. What about heavy/binge drinking in parents could be related to a higher casualty rate in incidents of school shootings?