School Crime and Safety

Katherine Lee and Mark Rauschkolb

December 19, 2021

Contents

1 Executive Summary

Problem. Sandy Hook Elementary School and Santa Fe High School are synonymous with school shootings and violence. While national trends show that violence and victimization has declined from 1992 to 2017, there is a sentiment that the number of school shootings is increasing, mainly due to a steep increase in multiple-victim homicides (as opposed to single-victim homicides), revealing that the topic of school violence is still very relevant. For our final project, we decided to look into various crimes and incidents of violence across US schools throughout 2015. While only analyzing 2015 school trends inherently limits the generalizability of our results, we believe the analysis draws high-level analyses about school crime trends.

Data. Our datasets came from a variety of sources. First, we obtained school-level data from the Urban Institute Education Portal to see what school-specific features may be predictive of school crime. Using this portal, we were able to access data from: National Center for Education Statistics' Common Core of Data (CCD), the Civil Rights Data Collection (CRDC), the US Department of Education's EDFacts, and IPUMS' National Historical Geographic Information System (NHGIS). Specifically, we pulled: school-level directory / geographic data, school finance data, teacher and staff data, discipline instances, and criminal incidences. Secondly, we also obtained county-level data to analyze any macro-trends that predict school crime. This county level data came from US Economic Research Service (for unemployment and poverty estimates), US Quarterly Census of Employment and wages(for wage data), Census Bureau's American Community Survey (for education data), and Census Bureau's Demographic data (for general demographic data).

All macro-data was at the county-level, each row having a unique County FIPS code. This FIPS code was used to match to each shcool's FIPS code.

Analysis. Before running any analyses, we split our data into a training dataset and a testing dataset. Because of the skewed nature of our response variable, total crime incidents for a specific school, we log transformed the response, which resulted in more symmetric data. Then, we explored our data to make understand correlations between variables and the response. To understand the predictors of total crime and make a predictive model, we built six different cross-validated models, with the end goal of selecting the best model with the best performance. Specifically, we built a ridge regression, Lasso regression, pruned regression decision tree, random forest, and boosting model. Both penalized regression models gave similar RMSEs, with the Lasso method yielding slightly better predictions. Of the tree-based models, the random forest and boosting model greatly outperformed the regular regression decision tree. The Random Forest had the lowest test error of the tree-based models, and served as the best predictive model overall.

Conclusions.

1. As seen by our regression tree, random forest, and boosting model, school-specific features were more important predictors than county-level features. Although many macro variables (unemployment, poverty, low adult education) did correlate with school crime, these factors became less relevant when accounting for school-specific variables. I believe this is a sign that although schools may be placed in a more impoverished county, good funding, staff etc. in a school can help the school beat its precarious circumstances. Counselors and teachers can help create a safe environment for students irrelevant of the environment the school is placed in.

2. Lasso penalized regression revealed two county-level variables were % of a county that identifies as being black, and the % of 5-17 year old children in poverty. I think this signals two key aspects to local government: Firstly, all kinds data repeatedly shows that marginalized minorities are afflicted by a variety of issues the most; although it is already known, we believe local government needs to remember to recognize inequalities in race when allocating funds. Next, we believe it is important to track teenage poverty specifically - helping families that are sustaining younger children and adolescents could reduce crime and improve school systems.

3. How much a school is suspending its students is the most important variable in our regression tree, random forest, and boosting models. We believe the department of education should investigate this phenomenon more. There may be reverse causation (a lot of crime results in higher suspension rates) however, after some research we found a many analyses that corroborate that suspensions may actually be causing worse behavior such as this National Association of School Psychologists article, or

this National Education Association article, both of which explains how suspension creates a vicious cycle: antagonizing students, sending them away school to school which often lead to long-periods of unsupervised time, more crime upon returning to school, and a subsequent suspension. Reform on this traditional reprimand method could be a cost-free method of greatly reducing crime rates.

4. Investments in teachers and counselors seem to be the most important. Variables such as security guards, law enforcement, administration, and nonpersonnel expenditures did not prove to be important in any of our models. On the other hand, larger investment in teachers and counselors (those who directly interact with students) did show importance. I believe this is extremely important for school boards to consider when allocating budgets.

2 Introduction

Background. School violence has always been a part of the United States' education system. From 2017-2018, 80% of public schools recorded one or more incidents of violence, translating to a crime rate of 29 incidents per 1,000 students enrolled.[1] Not all recorded incidents of violence and crimes were reported to the police, suggesting that the crime rate in America's schools is even higher. While the total victimization rate and the rates of specific crimes have declined since 1992, crime rates are still high, suggesting it is of vital importance to utilize various data sources to understand the highest risk factors for school crimes. Furthermore, a thorough analysis of school crime rates and predictive factors may help inform strategies to improve school safety precautions. Establishing reliable indicators of school crime and safety are important in ensuring the safety of schools across America.

Prior research and articles have shown that school crime are influenced by a variety of factors. For example, research has shown that higher use of police in schools is associated with more school crimes.[2] Moreover, trends show that students residing in rural areas had higher rates of total victimization than students residing in suburban areas, suggesting crime rate differences across school geographies.[3] Despite past research, there is much more to be learned and there is insufficient research regarding school crime rates and school-specific features.

Analysis goals. Given school crime can be attributed to a variety of factors, including both school-specific variables and demographic variables, we explored how total crime rates (total crime per 1000 enrolled students) are affected by various school and demographic features. Specifically, we were interested in what kind of factors, either school-specific or demographic-specific, and which specific variables are most predictive in school crime rates.

Significance. Our analysis will contribute to research regarding school crime risk factors and inform government officials of what factors are predictive of school crime. This will help support efforts to minimize school crime and improve school safety.

3 Data

3.1 Data sources

Our dataset was merged from data from the following sources: 1) Urban Institute Education Portal, 2) US Economic Research Service, 3) US Quarterly Census of Employment and wages, 4) Census Bureau's American Community Survey, and 5) Census Bureau's Demographic data. Each data source includes multiple years of data, and we chose to focus on 2015 data since there were the most number of available features for that year.

The data regarding school-specific features came from the Urban Institute Education Portal. The data from the portal was pulled from a variety of sources, including the National Center for Education Statistics' CCD,

[1] School Crime: Fast Facts. (n.d.). https://nces.ed.gov/fastfacts/display.asp?id=49.

[2] Gottfredson, D., Na, C., (2011). Police Officers in Schools: Effects on School Crime and the Processing of Offending Behaviors

[3] National Center for Education Statistics. (n.d.). Indicators of School Crime and Safety: 2015. https://bjs.ojp.gov/content/pub/pdf/iscs15.pdf

the CRDC, the US Department of Education's EDFacts, and IPUMS' NHGIS. In order to ensure a wide range of explanatory variables in our dataset, we pulled 2015 variables across the following categories: school directory, demographic, finance, teacher and staff, and discipline data, in addition to our response variable, which came from the criminal data. After pulling these datasets, we merged the variables into a final dataset by each school's National Center for Education Statistics (NCES) identification number (a unique school identifier).

Because we were also interested in broader demographic data, we drew from various government and census datasets, including: US Economic Research Service, US Quarterly Census of Employment and Wages [mark this link is broken], Census Bureau's American Community Survey, and Census Bureau's Demographic data. The US Economic Research Service provided county-level unemployment and poverty estimates. The unemployment rates were recorded from 2000 to 2020, and the only poverty estimates available were from 2019. Because the rest of our data is of 2015, we assume that poverty estimates from 2015 are reasonably similar to those from 2019. We extracted data from 2015 in the unemployment data and the 2019 poverty estimates for our data analysis. The US Quarterly Census of Employment and Wages [ ]. From the CEnsus Bureau's American Community Survey, we obtained US education figures [links seem off this link looks like it's up until 2014 but the analysis has 2015-2019 data]????

3.2 Data cleaning

For data cleaning, we started by examining each dataset in depth to determine which features and observations to keep.

In the school-specific datasets, we first transformed all negative values to NA values, since negative values imply that the data was missing/not reported or not applicable, per the Education Data Portal. Then we dropped all features that had more than 85% NA values, and dropped repeat observations and observations with any NA feature values. We dropped NA values for consistency purposes, as many of the data mining methods we employed require that all variables be populated with non-NA fields. In the directory dataset, one important feature is the degree of urbanization (urban-centric locale) of the school, which was coded with 12 levels, so we collapsed the categorical variable into 4 broader levels, including city, suburb, town, and rural. After cleaning each individual dataset, we merged the school-specific sets by the schools' unique NCES identification number.

For the macro-level county datasets, we first filtered the data to include only 2015 data (except poverty estimates because 2015 data was not available). We renamed variables for clarity and computed various race metrics as a percentage of population for comparability purposes. Each school is associated with a county-specific FIPS code and the macro-level county data includes observations for each county, which allowed us to merge in the macro-level county data in with the school-specific data, resulting in our final dataset that we used for analysis.

3.3 Data description

3.3.1 Observations

Our cleaned and merged final dataset has a total of 66694 observations, corresponding to each of the schools included in our analysis.

3.3.2 Response Variable

Our response variable is the number of crimes per 1000 students enrolled (square root transformed). In the criminal incident dataset, we created our response variable by summing rape incidents, sexual battery incidents, robberies and attacks to get a high level variable that captured crime rates. We also transformed the total crime figure to be per 1000 students enrolled (by dividing by the school's enrollment and then multiplying by 1000), to improve comparability of schools on a 'per student enrolled' basis as it inherently controls for school size variability. Then we square root transformed the response to obtain a more "symmetric" variable.

3.3.3 Features

In our dataset, we included 46 explanatory variables in our analysis, which fall into two broad categories: school-specific factors and macro/county-level factors. For a detailed specification of these variables, refer to Appendix A.

3.4 Data allocation

After data cleaning, we split our dataset into a training dataset and a testing dataset. We used an 80-20 split, such that the training set consists of 80% of the observations and the testing set consists of the other 20% of the observations. The training dataset was used for building our predictive models and the testing dataset was used for model evaluation. subsets: a training dataset used for building our predictive models and a test dataset used for evaluating our models.

3.5 Data exploration

3.5.1 Response

We first explored the response variable's distribution. As seen in the histogram of total crimes per 1000 students variable (Figure 1). The histogram is very right-skewed, indicating that the dataset contains a number of outliers with extremely high crime rates. The mean is 19.3 crimes per 1000 enrolled students. We proceeded to determine which schools had extreme response rates by looking at the sorted data. The sorted data (Table 1) shows that Maple Lane School had the highest crime rate by far, more than double the second highest ranked school. Further research shows that Maple Lane School was a juvenile corrections facility that closed in 2010, indicating that the observation should not be in the dataset to begin with.[4]

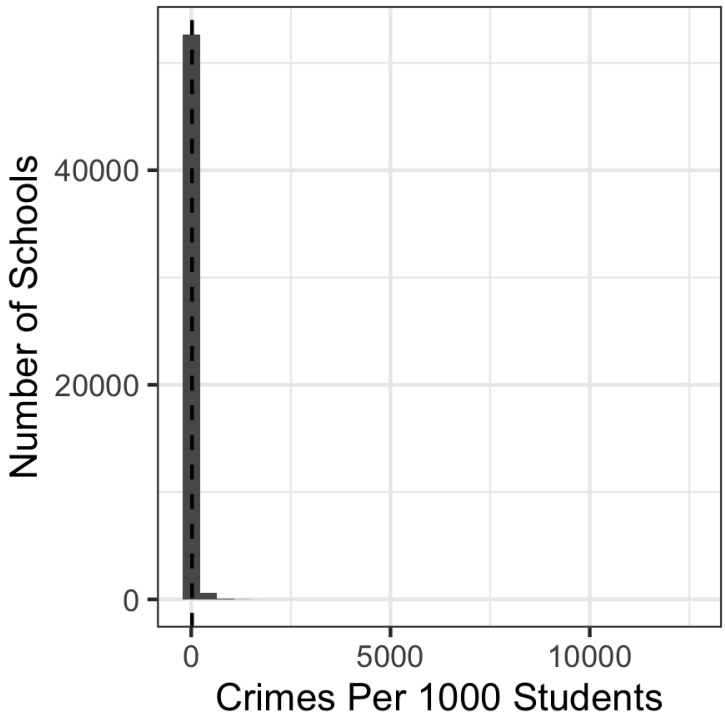

Figure 1: Distribution of Crime Rate; vertical dashed line indicates the mean.

[4]Turn Maple Lane into prison center? Allen, Marqise. https://www.theolympian.com/news/local/article25274248.html

Table 1: Top ten schools by crime rate (expressed as per 1000 students).

School	Crimes per 1000 Students
Maple Lane School	12458
LITTLE SCHOOL	4967
Carver Middle School	2990
Tangipahoa Alternative Solutions Program	2400
DELLA LAMB @ WALLACE	2396
Monroe Area High School	2103
EASTERN WRIGHT PROGRAM	2000
Murrell School	1709
Luther E Ball (Chjcf)	1600
DELLA LAMB @ WOODLAND	1527

3.5.2 Features

[Omitted from template.]

4 Modeling

4.1 Regression-based methods

4.1.1 Ordinary least squares

[Omitted from template.]

4.1.2 Penalized regression

Despite the ordinary least squares method seeming to work well, we realized that fitting a linear model with so many explanatory variables might incur a large cost in variance and lead to suboptimal predictions. Hence, we decided to build and evaluate shrinkage models with the hopes of getting a more parsimonious and interpretable model. We ran three cross-validated regressions for which optimal values of lambda were chosen according to the one-standard-error rule: ridge, LASSO (Least Absolute Shrinkage and Selection Operator), and elastic net.

For the lasso, Figure **??** shows the CV plot, Figure **??** shows the trace plot, and Table 2 shows the selected features and their coefficients.

[Interpretation of lasso omitted from this template. Other penalized regression methods omitted from this template.]

Following are the results for tree based methods:

```
read_tsv("../results/model-evaluation-regression.tsv") %>%
  kable(format = "latex", row.names = NA,
        booktabs = TRUE, digits = 2,
        col.names = c("Method", "RMSE (Log)", "RMSE"),
        caption = "RMSE summary for Regression based methods") %>%
  kable_styling(position = "center")
```

```
## Rows: 3 Columns: 3
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (1): Method
```

Table 2: Standardized coefficients for features in the lasso model based on the one-standard-error rule.

Feature	Coefficient
threats_w_weapon_incidents	19.75
threats_no_weapon_incidents	8.76
urban_centric_localecity	4.89
black_pop	4.75
school_type2	4.57
teachers_fte_crdc	-3.91
threats_w_firearm_incidents	-3.46
free_lunch	3.39
urban_centric_localerural	-2.84
Civilian_labor_force_2015	-2.61
teachers_current_sy	-2.54
teachers_absent_fte	2.34
social_workers_fte	2.24
percent_school_age_children_poverty	2.09
teachers_first_year_fte	2.00
possession_firearm_incidents	1.99
labor_force_participation	1.86
law_enforcement_ind	1.78
school_type1	-1.73
hisp_pop	-1.67
reduced_price_lunch	-1.47
security_guard_fte	1.45
teachers_uncertified_fte	-1.42
salaries_teachers	-1.35
percent_some_college	1.08
urban_centric_localetown	-1.06
avg_wkly_wage	-0.77
administration_fte	-0.72
charter0	0.64
teachers_previous_sy	-0.63
virtual0	0.62
salaries_total	-0.48
psychologists_fte	0.28
Poverty_all_population	0.21
avg_suspensions	0.21
asian_pop	-0.19
counselors_fte	-0.19
salaries_administration	0.17
percent_only_highschool	-0.11
teachers_second_year_fte	0.09
salaries_instructional_aides	-0.06
Unemployment_rate_2015	0.06
expenditures_nonpersonnel	-0.06
charter1	0.00

Table 3: RMSE summary for Regression based methods

Method	RMSE (Log)	RMSE
Ridge	3.59	89.6
Lasso	3.62	89.8
Elnet	3.50	88.3

Table 4: RMSE summary for Tree based methods

Method	RMSE (Log)	RMSE
Regression Tree	3.30	88.1
Random Forest	3.00	85.0
Boosting	3.05	84.2

```
## dbl (2): Test RMSE (Sqrt Transformed), Test RMSE

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

4.2 Tree-based methods

[Omitted from template.]

4.2.1 Random forest

4.2.2 Boosting

```
read_tsv("../results/model-evaluation-tree.tsv") %>%
  kable(format = "latex", row.names = NA,
        booktabs = TRUE, digits = 2,
        col.names = c("Method", "RMSE (Log)", "RMSE"),
        caption = "RMSE summary for Tree based methods") %>%
  kable_styling(position = "center")
```

```
## Rows: 3 Columns: 3

## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (1): Method
## dbl (2): Test RMSE (Sqrt Transformed), Test RMSE

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

5 Conclusions

5.1 Method comparison

Table 5 shows the test RMSE for all the methods considered. Except for the OLS, the random forest and the boosted model have the lowest test errors. This is reasonable given these models' tendencies to have high

Table 5: RMSE summary for all data minig methods used

Method	RMSE
Ridge	89.6
Lasso	89.8
Elnet	88.3
Tree	88.1
Random Forest	85.0
Boosting	84.2

predictive accuracy. Between the two, the boosted model has the lowest test error, with a mean squared error of 0.000139, but it is closely followed by random forest, which has a mean squared error of 0.00141. Notably, however, the ridge, LASSO, and elastic net regressions perform about as well, with test MSEs of 0.000158, 0.000164, and 0.000161, respectively. Although OLS has the lowest training and test error, its adjusted R-squared value was only about 0.3, and there were too many features given the number of observations.

Regardless of these differences in test MSE, the methods overlap significantly in their identification of important variables from the larger set. For instance, the elastic net regression selects the following variables, which are also selected by LASSO and deemed significant in the OLS model: other providers ratio, unemployment, income inequality, housing overcrowding, residential segregation—non-White/White, homeownership, and physical inactivity. The random forest and boosting models both include low birthweight percentage, median income, and unemployment percentage in the top 10 most important variables, as measured by their contributions to node purity.

5.2 Takeaways

Our results point to a few key determinants of health that, given their impact on COVID-19 deaths per cases rates in 2020, policymakers should consider when aiming to improve factors that would improve health overall but also potentially mitigate the mortality risk of another pandemic. The boosted model, which had the strongest predictive performance, suggests that residential segregation between non-white and white residents is the most important variable in predicting a county's COVID-19 deaths per cases rate. The unemployment rate, availability of physical exercise opportunities, and measures of home ownership burden variables were also highly important in this model. These variables are identified across all models suggesting these relationships are robust. Residential segregation, unemployment, and home ownership burden are socioeconomic factors that affect an individual's ability to access and pay for healthcare. In that regard, it is unsurprising that these factors would have a greater ability to predict differences in COVID-19 case fatalities across different counties in the United States. While some behavioral variables are also found to be significant (including STI incidence, high school completion, and others identified in the elastic net and lasso regressions), the variable importance ranking from the boosted model provides a highly interpretable hierarchy of the most influential factors from the greater set.

Given that socioeconomic factors were the strongest predictors of deaths per cases, it appears that on the county level, COVID-19 rates are most associated with community healthcare burdens. That is, COVID-19 outcomes appear to reflect the reality of healthcare accessibility across counties; those with high percentages of uninsured citizens, or with high degrees of segregation and inequality, might feature division of healthcare resources that reflects these disparities. It is thus reasonable that deaths per case rates would be higher when significant groups of a population have lesser access to healthcare resources and treatment. Notably, if our conclusions are indeed correct, this effect would likely be pronounced in situations of scarcity such as the early months of the pandemic studied here, when many hospitals faced shortages of ventilators and other resource shortages. Given the progression of COVID-19 since 2021 as well as the inherent complexity of fatality incidence, we are hesitant to make any assertive claims about the true predictive capacity of any of the top factors we identified. Nonetheless, these results can help inform policies directed toward improving various determinants of important health outcomes in counties across the US.

As the world shifts towards herd immunity as vaccines are made more widely available, it is important to reflect upon how the pandemic has asymmetrically impacted different counties across the country. Our results suggest that structural vulnerabilities can be captured by measures of inequality and poverty; the identification of counties high on these factors should serve as a warning for future vulnerability to health crises. These analyses can serve to protect already vulnerable communities from suffering disproportionately in the future.

5.3 Limitations

5.3.1 Dataset limitations

As it is detailed in the Frequently Asked Questions page of County Health Rankings & Roadmaps program website,[5] all of the variables are from 2019 or earlier. Thus, it is possible that the values of the variables assessed were different in 2020, meaning that the interpretation of our analysis may need to be taken with a grain of salt. Regardless, given that we analyzed data on the county level, it is unlikely that any county experienced enough drastic change over the course of the year to significantly affect our analysis. Furthermore, our dataset has a large number of observations to account for potential variability, although notably, many observations had to be removed as some of the R packages used to build some of our models required that no NA values were present in the data. In other words, since each observation represents a different US county, many counties were left out of our analysis. Another limitation is that, as described in the Exploratory Data Analysis section, there is evidence of correlation amongst some of our explanatory variables. This means that some variables can be confounding variables, which mask or distort the relationship between measured variables. Also, variables selected in the LASSO regression and elastic net regression as well as variables marked important in the tree methods might be misleading in that, given how variable selection works, it is possible that some selected variables are simply representative of a larger group of correlated variables.

5.3.2 Analysis limitations

While splitting the data into training and testing datasets allows for a more unbiased test of the models, we recognize that our conclusions inherently contain some randomness due to the random split of the data. In other words, splitting the data again using a different random seed may have yielded different p-values in the OLS regression, different selected variables in the shrinkage methods, and different variables selected as important in the tree-based methods. Next, although we provide different methods for robust interpretation of the variables, our analysis incorporates only a specific subset of health related variables. The results of the analysis might change dramatically if we were to incorporate other variables. For example, as mentioned in the Exploratory Data Analysis, states in the northeastern part of the United States suffered from a high rate of COVID-19 cases and deaths in 2020. This may not have been because these states performed poorly in the health variables mentioned above. Rather, it may be due to the fact that these states are densely populated and hence were more susceptible to disease spread at the outset of the COVID-19 pandemic. In other words, other factors like geographic or demographic variables can hugely impact case fatality rate.

5.4 Follow-ups

To compensate for the limitations mentioned above, more extensive analysis can be done as we acquire more data from 2020 and 2021. Not only can we extend our analysis by utilizing the most up-to-date datasets, but we can also examine how COVID-19 cases and deaths have affected various health factors of each county. In other words, the explanatory and response variables can be reversed to conduct more dynamic data analyses. Next, given that many observations needed to be omitted in our dataset as they contained NA fields, we recommend that our analyses be reconducted once the missing data is collected. Finally, future work on the social determinants of health in the context of COVID-19 might also look at different population levels such as states, bigger geographical regions in America, or even different countries.

[5] County Health Rankings & Roadmaps. (n.d.) Frequently Asked Questions. https://www.countyhealthrankings.org/explore-health-rankings/faq-page

A Appendix: Descriptions of features

Below are the 45 features we used for analysis. Words written in parentheses represent variable names. Unless noted otherwise, all variables are continuous.

School-Specific Variables:

- *Geography*
  - Urbanization (`urban_centric_locale`): Factor variable representing the degree of urbanization. Factors include schools located in: cities, rural areas, suburbs, and towns.
- *School Attributes*
  - School type (`school_type`): Factor variable representing the school type. Factors include: regular schools, special education schools, and vocational schools.
  - Charter school (`charter`): Binary variable representing whether the school is a charter school (1) or not a charter school (0).
  - Students eligible for free or reduced-price lunch (`free_reduced_lunch_per1000`): Number of students eligible for free or reduced-price lunch per 1000 enrolled students.
- *Teacher and Staff*
  - Full-time equivalent teachers (Civil Rights Data Collection) (`teachers_fte_crdc`): Number of full-time equivalent teachers.
  - Full-time equivalent certified teachers (`teachers_certified_fte`): Number of full-time equivalent certified teachers.
  - Full-time equivalent uncertified teachers (`teachers_uncertified_fte`): Number of full-time equivalent uncertified teachers.
  - Full-time equivalent first-year teachers (`teachers_first_year_fte`): Number of full-time equivalent first-year teachers.
  - Full-time equivalent second-year teachers (`teachers_second_year_fte`): Number of full-time equivalent second-year teachers.
  - Current school year teachers (`teachers_current_sy`): Number of current school year teachers.
  - Previous school year teachers (`teachers_previous_sy`): Number of previous school year teachers.
  - Full-time equivalent teachers absent more than 10 school days (`teachers_absent_fte`): Number of full-time equivalent teachers absent more than 10 school days
  - Full-time equivalent school counselors (`counselors_fte`): Number of full-time equivalent school counselors.
  - Full-time equivalent psychologists (`psychologists_fte`): Number of full-time equivalent psychologists.
  - Full-time equivalent social workers (`social_workers_fte`): Number of full-time equivalent social workers.
  - Full-time equivalent nurses (`nurses_fte`): Number of full-time equivalent nurses.
  - Full-time equivalent security guards (`security_guard_fte`): Number of full-time equivalent security guards.
  - Sworn law enforcement officers indicator (`law_enforcement_ind`): Indicator of whether a sworn law enforcement officer has been assigned to the school.
  - Teacher salaries (`salaries_teachers`): Personnel salaries at school level (teachers only) amount.
- *School Finance*
  - Non-personnel expenditures (`expenditures_nonpersonnel`): Amount of non-personnel expenditures. May include: professional development for teachers, computers, library books, and other learning materials.

- *Criminal Activity*
  - Threats of physical attack with a weapon (`threats_w_weapon_incidents`): Number of incidents of threats of physical attack with a weapon.
  - Threats of physical attack with a firearm or explosive device (`threats_w_firearm_incidents`): Number of incidents of threats of physical attack with a firearm or explosive device.
  - Threats of physical attack without a weapon (`threats_no_weapon_incidents`): Number of

incidents of threats of physical attack without a weapon.
  - Possession of a firearm or explosive device (`possession_firearm_incidents`): Number of incidents of possession of a firearm or explosive device.
- *School Discipline*
  - Suspensions (`avg_suspensions`): Number of student suspensions.

County-Level Variables: [mark to do] - *Access to Care* - Uninsured (`uninsured`): Percentage of population under age 65 without health insurance. - Primary care physicians (`primarycare_ratio`): Ratio of population to primary care physicians. - Dentists (`dentist_ratio`): Ratio of population to dentists. - Mental health providers (mentalhealth_ratio): Ratio of population to mental health providers. - Other primary care providers (`otherproviders_ratio`): Ratio of population to primary care providers other than physicians. - *Quality of Care* - Preventable hospital stays (`preventable_hospitalization`): Rate of hospital stays for ambulatory-care sensitive conditions per 100,000 Medicare enrollees. - Mammography screening (`mammogram_perc`): Percentage of female Medicare enrollees ages 65-74 that received an annual mammography screening. - Flu vaccinations (`flu_vaccine_perc`): Percentage of fee-for-service (FFS) Medicare enrollees that had an annual flu vaccination. - Teen births (`teen_births`): Number of births per 1,000 female population ages 15-19.

Social and economic factors:

- *Education*
  - High school completion (`HS_completion`): Percentage of adults ages 25 and over with a high school diploma or equivalent.
  - Some college (`some_college`): Percentage of adults ages 25-44 with some post-secondary education.
  - Disconnected youth (`disconnected_youth`): Percentage of teens and young adults ages 16-19 who are neither working nor in school.
- *Employment*
  - Unemployment (`unemployment`): Percentage of population ages 16 and older who are unemployed but seeking work.
- *Income*
  - Children in poverty (`children_poverty_percent`): Percentage of people under age 18 in poverty.
  - Income inequality (`income_inequality`): Ratio of household income at the 80th percentile to income at the 20th percentile.
  - Median household income (`median_income`): The income where half of households in a county earn more and half of households earn less.
  - Children eligible for free or reduced price lunch (`children_freelunches`): Percentage of children enrolled in public schools that are eligible for free or reduced price lunch.
- *Family & Social Support*
  - Children in single-parent households (`single_parent_households`): Percentage of children that live in a household headed by a single parent.
  - Social associations (`social_associations`): Number of membership associations per 10,000 residents.
  - Residential segregation—Black/White (`segregation_black_white`): Index of dissimilarity where higher values indicate greater residential segregation between Black and White county residents.
  - Residential segregation—non-White/White (`segregation_nonwhite_white`): Index of dissimilarity where higher values indicate greater residential segregation between non-White and White county residents.
- *Community Safety*
  - Violent crime rate (`Violent_crime`) Number of reported violent crime offenses per 100,000 residents.

Physical environment:

- *Air & Water Quality*
  - Air pollution - particulate matter (`air_pollution`): Average daily density of fine particulate matter in micrograms per cubic meter (PM2.5).

  - Drinking water violations (`water_violations`): Indicator of the presence of health-related drinking water violations. 1 indicates the presence of a violation, 0 indicates no violation.
- *Housing & Transit*
  - Housing overcrowding (`housing_overcrowding`): Percentage of households with overcrowding,
  - Severe housing costs (`high_housing_costs`): Percentage of households with high housing costs
  - Driving alone to work (`driving_alone_perc`): Percentage of the workforce that drives alone to work.
  - Long commute—driving alone (`long_commute_perc`): Among workers who commute in their car alone, the percentage that commute more than 30 minutes.
  - Traffic volume (`traffic_volume`): Average traffic volume per meter of major roadways in the county.
  - Homeownership (`homeownership`): Percentage of occupied housing units that are owned.
  - Severe housing cost burden (`severe_ownership_cost`): Percentage of households that spend 50% or more of their household income on housing.