Examining the Characteristics and Trends of Mass School Shootings through Empirical Analysis

DS490

By: Jake Klingler & Ryan Lee

Abstract

The goal of this project is to utilize various data science techniques to acquire and clean data related to school safety, which includes both specific school data, as well as surrounding city characteristics, and optimize machine learning models utilizing the data to determine if there are any trends or predictors in the events that threaten school safety.

Background

- First USA School Shooting: Pontiac's Rebellion school - July 26, 1764
- Columbine High School School Shooting 1999
  - Brought huge attention because of the area it happened
- From 2001-2022 there were 1,375 school shootings
  - Resulting in 515 deaths and 1,161 injuries
- School shootings jumped 124% between the 2020-21 and 2021-22 school years
- Security Theatre - “Feeling of improved security”

Research Question

- Can we use Data Science to help make schools safer?
  - Can data be helpful in incident prevention?
- What features should be the target variables to answer such a question?
- We hypothesize that the target variable "Shooter Killed" may indicate different types of attacks, distinguishing between those where the perpetrator attempts to escape and those where the perpetrator is on a suicide mission.

Analysis Approach

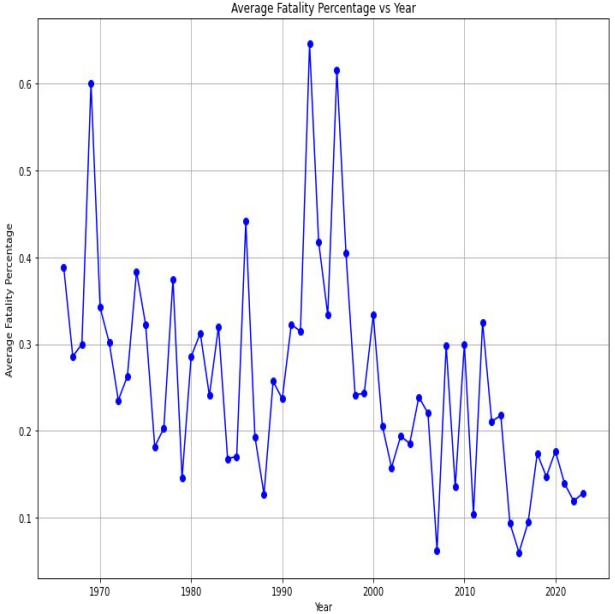
- Scrape multiple websites to get more data
- Clean/organize data
- Conduct exploratory analysis
- Split data into training test/train datasets
- Create ML models(both classification and regression)
- Optimize models
- Analyze results

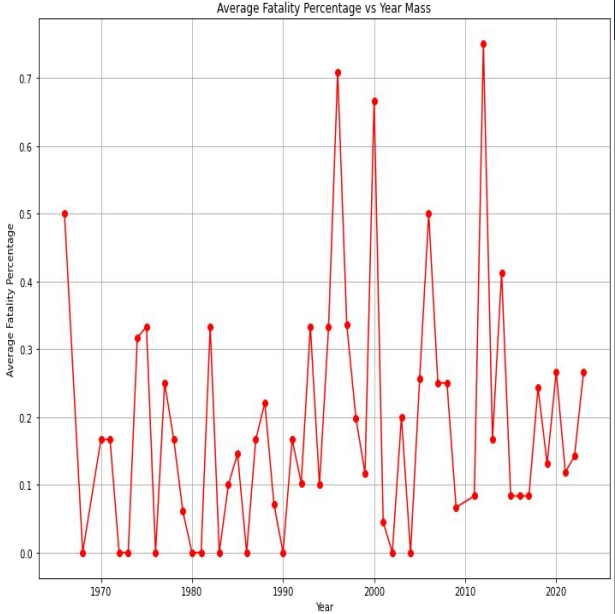
Our Data

- Over 1,000 columns
  - Started with 40
- Consists of:
  - Riedman's dataset - 2,585 incidents
    - Target Variables
  - School data (NCES)
  - Zip code data (Zip Atlas)

Average Fatality Percentage Over Time

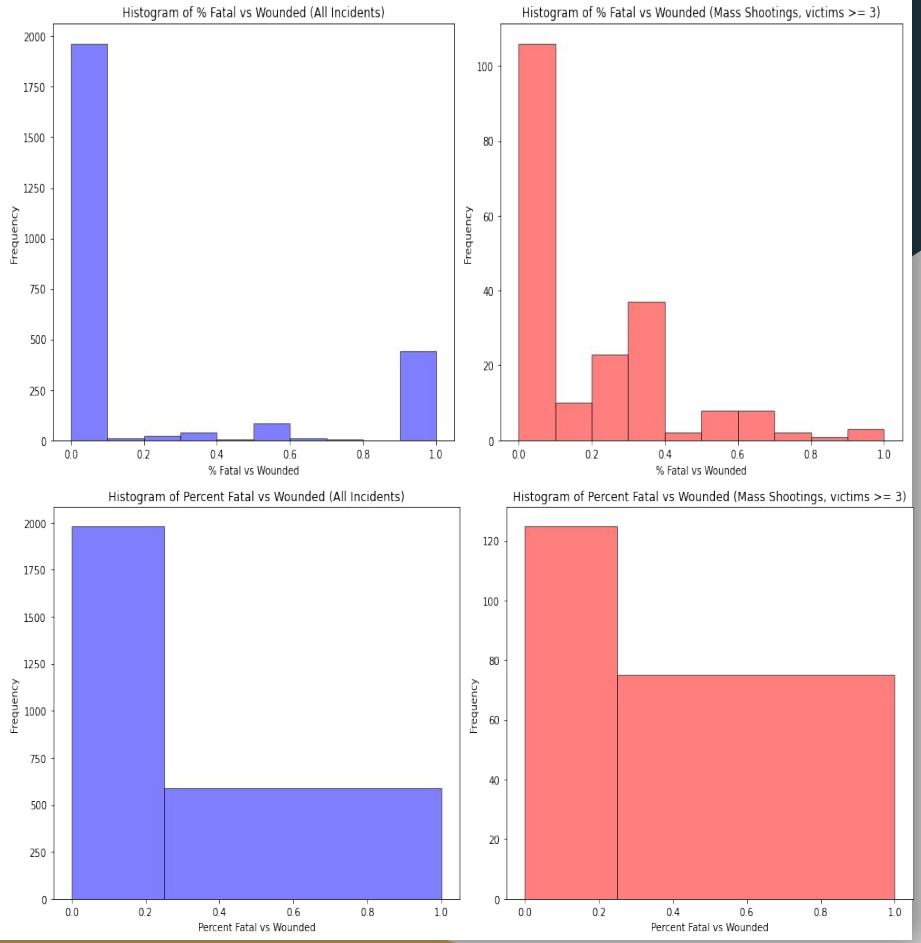
- General decrease over time in first plot
- Very inconsistent
  - Hard to say there is constant trend
- No trend in Mass plot
  - Spikes have relation to a large incident that specific year
  - I.e. 2007 VA Tech Shooting
  - 2012 Sandy Hook





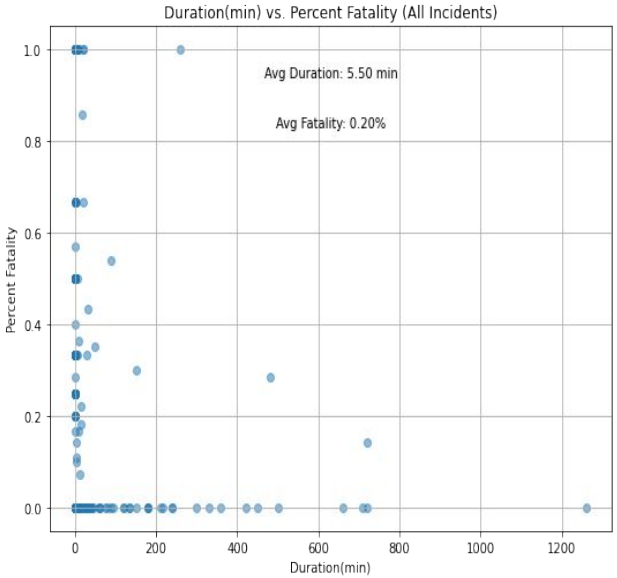
Exploratory Data Analysis Cont.

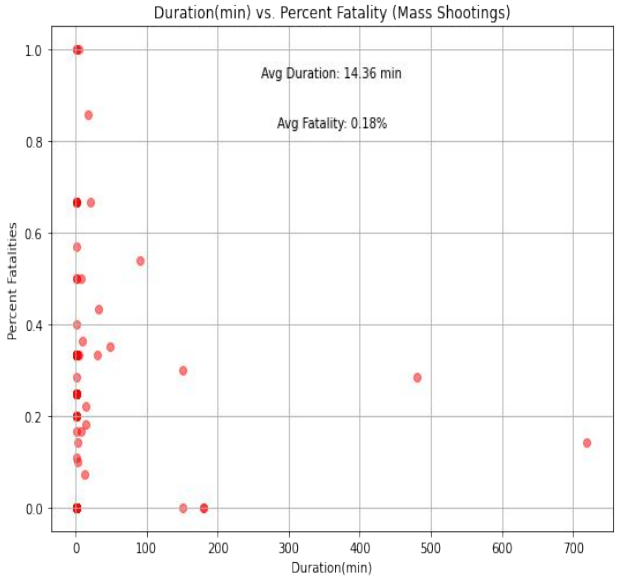
- % Fatal vs Wounded
- Changed the bin sizes
  - Type 1: Fatality % <= 10
  - Type 2: Fatality % >= 25



Duration vs Percent Fatality

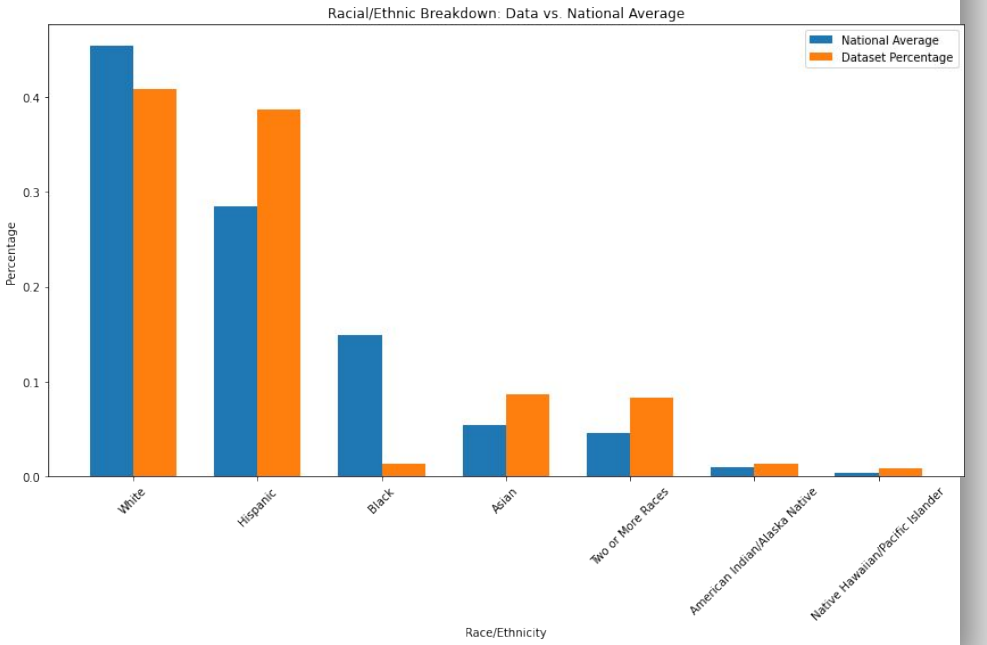
- Mass has longer average duration but lower average % Fatality
- Both plots have similar shapes





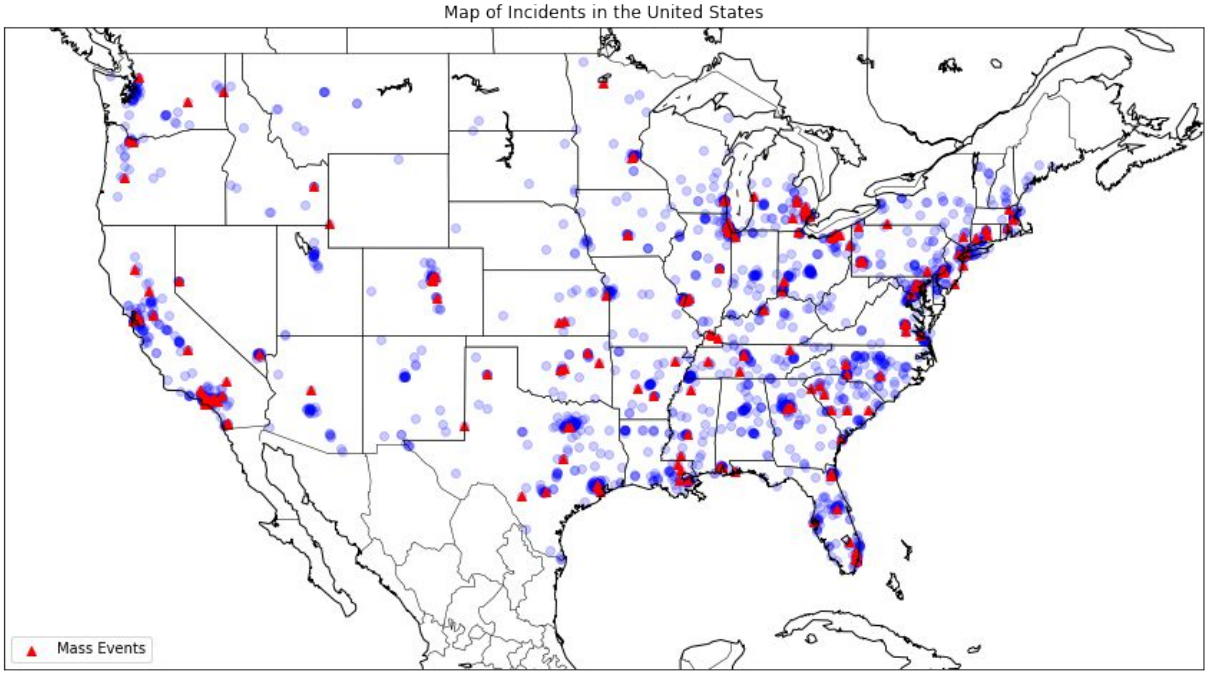
Racial Breakdown

- See that our dataset had higher rates of Hispanic, Asian, and 2 or more races
- Biggest difference is percentage of Black students



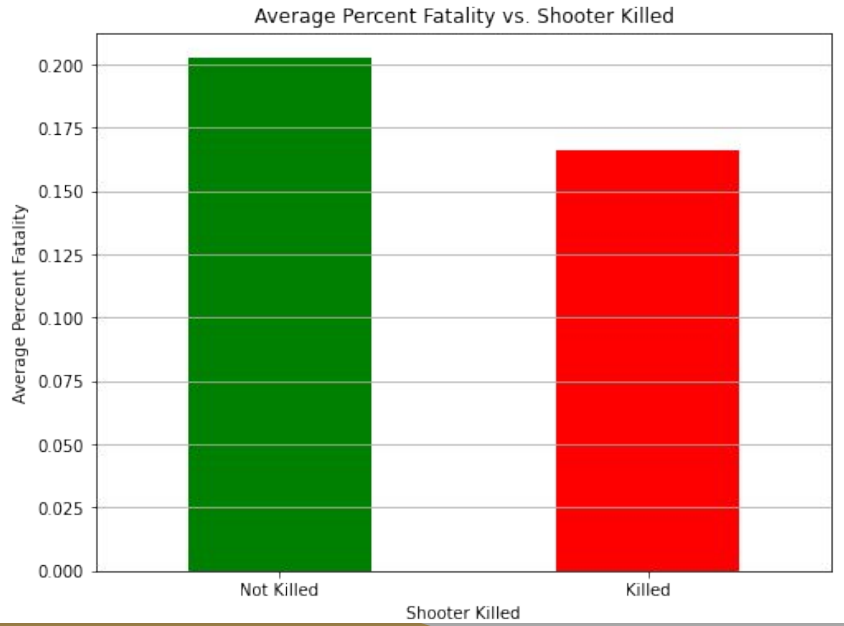
Map plot

- Shows how the data is distributed throughout US
- Alpha = 0.2 to show density

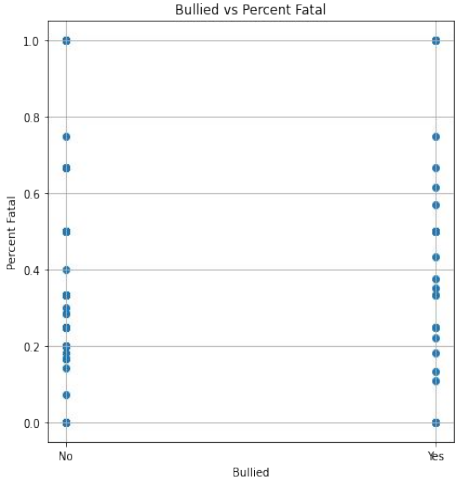


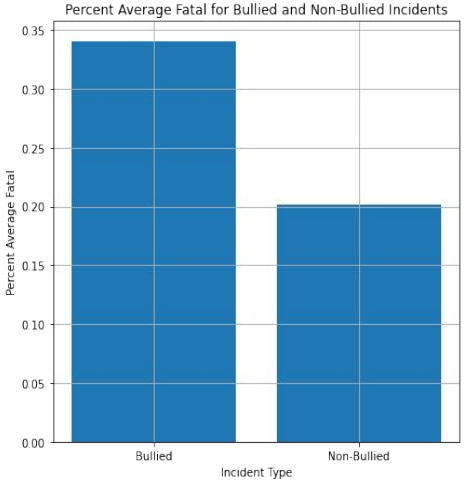
Average Percent Fatal vs Shooter Killed

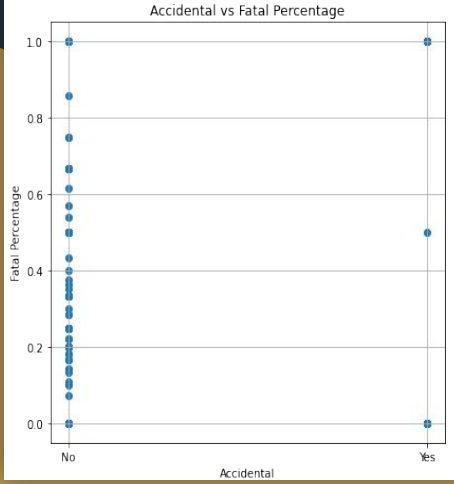
- A look at our two target variables and how they may be related
- Would have expected the “Killed” bar to be greater

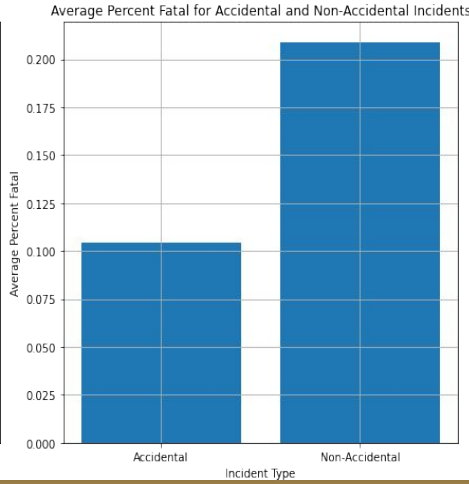


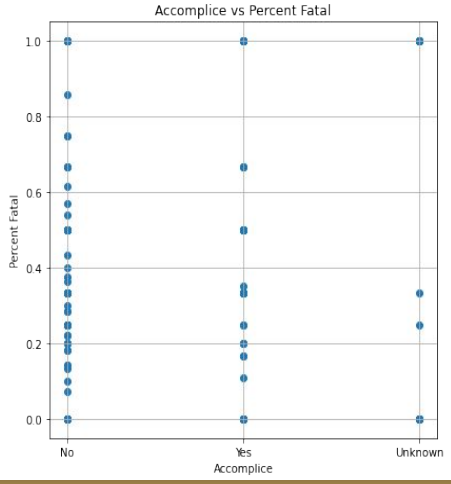
Features vs. Percent Fatal

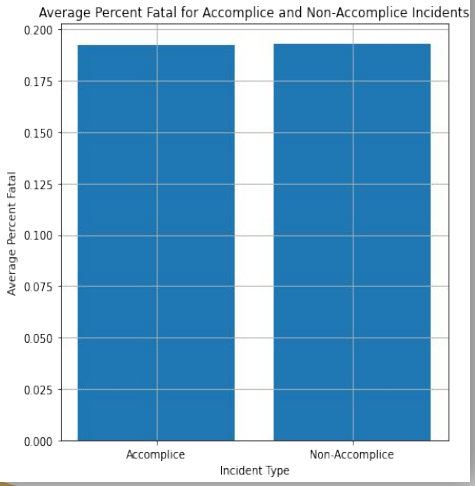






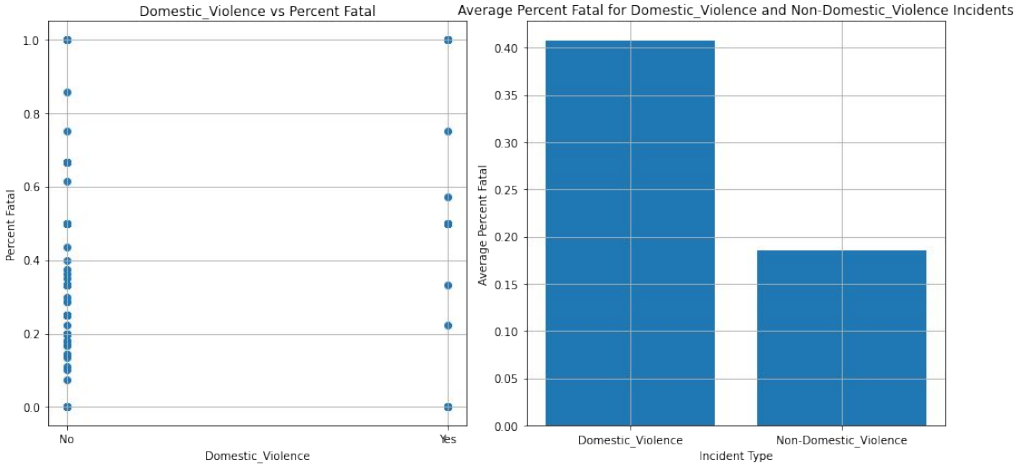


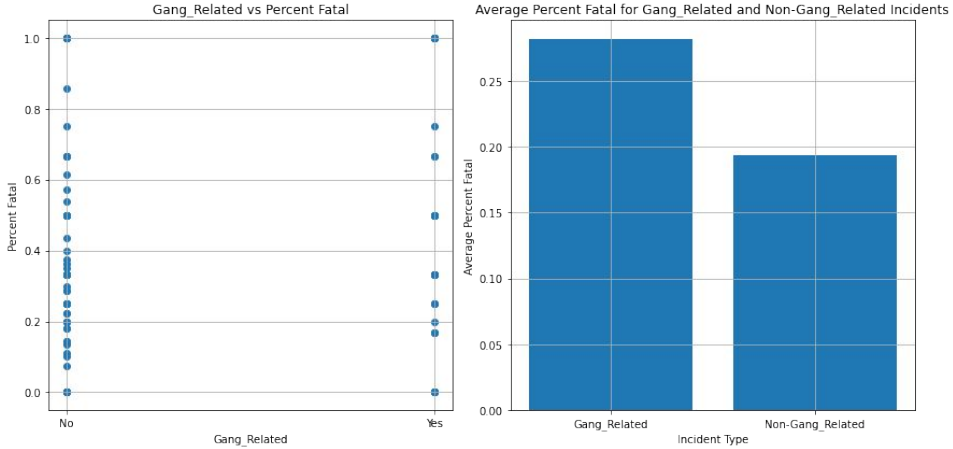




Features vs Percent Fatal Cont.

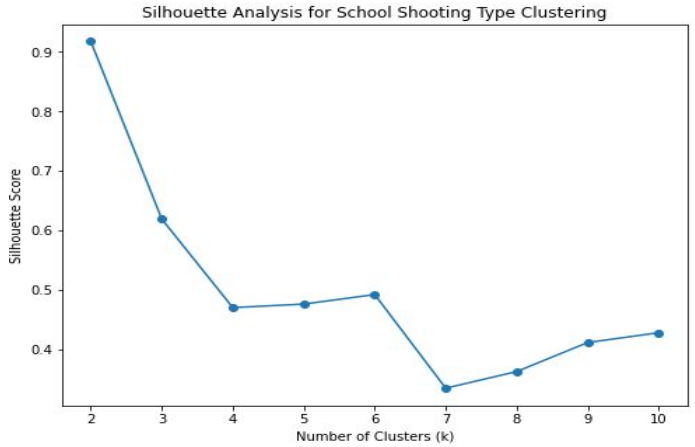
- Domestic Violence feature has biggest difference than any other feature
- Accomplice plot was surprising

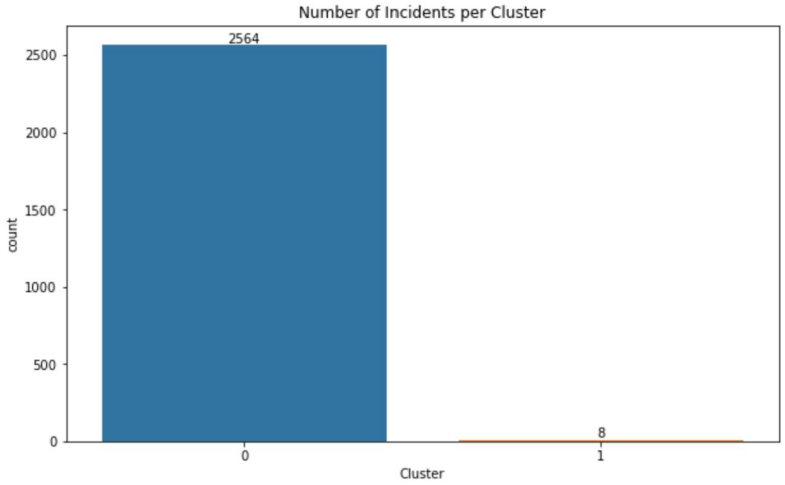




K-Means Clustering

Cluster Number	Number of Incidents	Shooter Killed	Victims Killed	Victims Wounded	Number of Victims	Income Overview Characteristic Per Capita Income Measure
0	2564	235	685	1944	2629	35764
1	8	6	94	201	295	42206

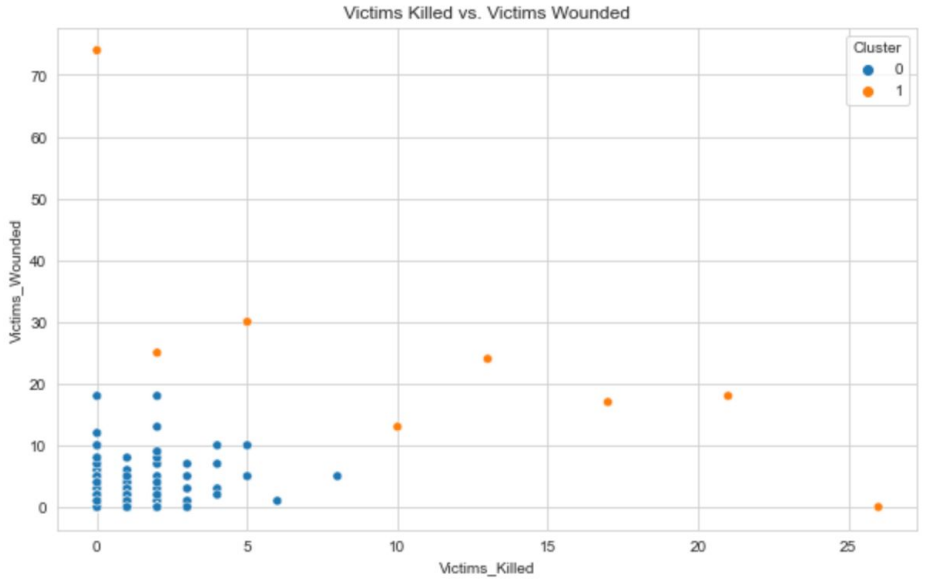




Cluster 0 and Cluster 1 Victims: Killed vs Wounded

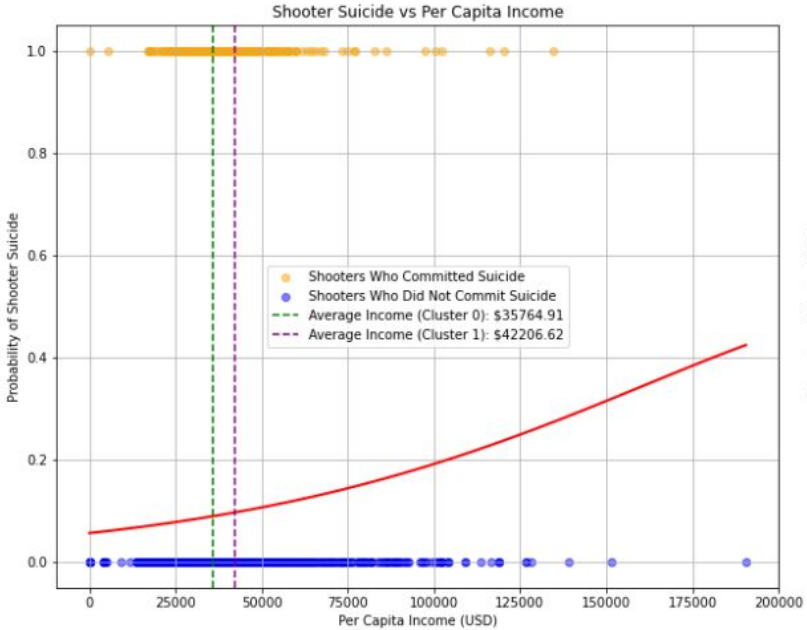
- Cluster 0: Lower Income ~ $35,764
- Cluster 1: Higher Income ~ $42,206

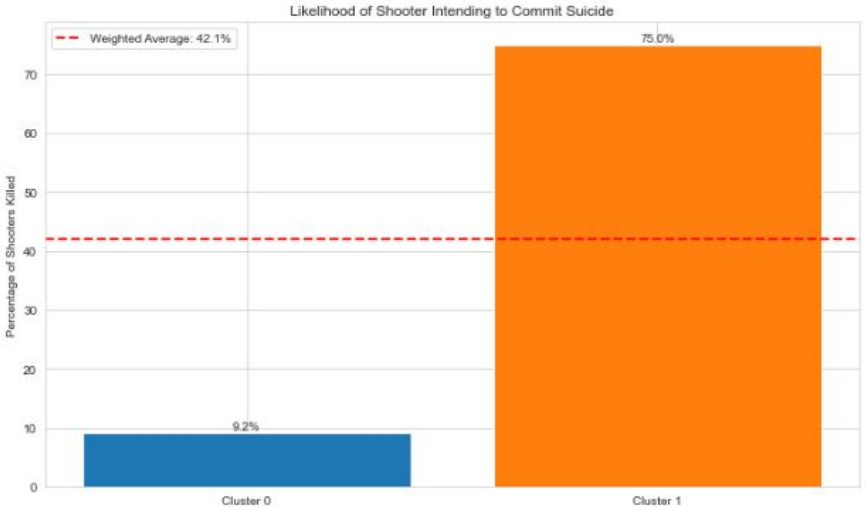
- Cluster 1 killed more people than Cluster 0



Logistic Regression

- Higher income areas come in with the intent to commit suicide.
- Weighted Average = 0.5





Models & Features

- ML Models Used
  - Neural Network
  - Multi-Linear/Logistic
  - Decision Tree
  - XGBoost
  - Random Forest
- Target Variables
  - Percent Fatality (Regression)
  - Shooter Killed (Classification)
- Features
  - Orig: Consists of the NCES and Riedman's K-12 School SHooting Database (21 Total Features)
  - Final: Consists of the "Orig" data as well as 29 additional features from the zipcode data (50 Total Features)

Results

Classification	NN	Multi-Logistic	Decision Tree	XGBoost	Random Forest
Accuracy(orig)	0.6868	0.4086	0.749	0.8852	0.9008
Accuracy(Final)	0.7023	0.5895	0.677	0.9047	0.8969
MSE(orig)		1.1342	0.2743	0.1206	0.1051
MSE(Final)		0.6206	0.3288	0.107	0.1089

Results

Regression	NN	Multi-Linear	Decision Tree	XGBoost	Random Forest
MSE (orig)	0.0175	0.0839	0.0003	0.0002	0.0009
MSE (Final)	0.0562	0.0843	0.0019	0.0008	0.001
R^2 (orig)	0.8689	0.3706	0.9977	0.9983	0.9931
R^2 (Final)	0.5788	0.3676	0.9856	0.9941	0.9927

What We Learned?

- Data Scraping really is 70%-80% of the project timeline
- XGBoost models performed the best in both regression and classification problems
- Most of the models performed very well
    - The Logistic classification model did not perform great
- Data on school characteristics and its surrounding area can be predictors in school safety incidents
    - Could possibly build off this project to work on prevention

Challenges Faced

- Collecting the data
  - Time consuming
  - Web Scraping
    - Original Zip Code Data
    - Fuzzy match schools
- Staying on track with the project
  - There were so many ideas and questions that we had
    - Tough to make sure we weren't trying to do too much
- Feature Selection
  - So many to choose from

Limitations

- Time
- Zip code data is current
  - Incidents happened in different years
- Computing power
  - Some scraping code took 12 hours

Future Ideas

- Gather more data
- Come up with a way to use our data and models to work on prevention

References

Abhishek Bagwan. (2023). School dataset csv-file. Kaggle.com.

https://www.kaggle.com/datasets/abhishekbagwan/school-dataset/data

THE ASSOCIATED PRESS. (2022, May 25). *List of deadliest US school shootings*. AP News; AP News.

https://apnews.com/article/list-of-deadliest-us-school-shootings-f25dad31e68c8acbdbcb952352df9249

Bankert, A. (2022, July 19). *School expert: We spend too much on "security theater."* NewsNation; NewsNation.

https://www.newsnationnow.com/morninginamerica/answersforamerica/school-expert-school-safety-security-theater/

Glavin, C. (2018, July 26). *History of School Shootings in the United States | K12 Academics*. K12academics.com.

https://www.k12academics.com/school-shootings/history-school-shootings-united-states

GPS Coordinates - Latitude and Longitude Finder. (2024). Gps-Coordinates.org. https://gps-coordinates.org/

National Center for Education Statistics (NCES) Home Page, part of the U.S. Department of Education. (2022). Ed.gov; National Center for Education Statistics. https://nces.ed.gov/

Riedman, David (2023). K-12 School Shooting Database

USAFacts. (2022, May 26). *The latest government data on school shootings*. USAFacts; USAFacts. https://usafacts.org/articles/the-latest-government-data-on-school-shootings/

Wikipedia Contributors. (2024, April 11). *Security theater*. Wikipedia; Wikimedia Foundation. https://en.wikipedia.org/wiki/Security_theater