Police Encounters & Casualties

Group 24

Rasim Alam | Morgan Goetz | Emil Ghitman Gilkes | Benjamin Manning

CS 209 - Data Science 1
School Of Engineering & Applied Sciences | Harvard University
Fall 2020

Table of Contents

Introduction:

Policing involved casualties - although they have been an especially germane topic in the news cycle with the 2020 killings of individuals such as Breonna Taylor, George Floyd, and many others, are not a novel problem. Compared to many countries in the world with similar socioeconomic demographics to the United States, the rates of police casualties are substantially higher. Specifically, violence and discrimination against specific minority groups has been extremely problematic. Our group finds these societal problems to be depressing, and thusly worth attempting to tackle using advanced statistical methods.

So, we decided to dive into publicly available data from several different sources (which we describe in future sections) to look at some of the many factors that lead to police casualties. Before our exploratory data analysis or our literature, we knew that race and other classical demographic factors would be intriguing to assess in terms of the victims of said incidences. However, we wanted to find out more than just the classical indicators for police violence, so we tried to look for the factors within departments and underlying factors in the communities that police casualties occur to accrue our initial data. Through the process of acquiring our data, surveying previous literature, and group discussion, we concluded with several questions that began our research process.

1. Is there a correlation between community engagement and police violence?
2. What are some of the key variables in predicting police involved casualties?
3. What are the underlying attributes of communities associated with predicting police involved casualties?
4. What are the makeups of the police departments?

However, after moving forward with our analysis, as can be seen through the rest of this paper, our questions evolved.

Literature Review:

Data science and policing have become an incisive intersection in recent years as to the problems with racial bias, determining sentencing, the proliferation of law-enforcement propagated violence, and many other factors. As discussed in this project, we are looking from a slightly different angle than many of the public discussions. I.e., not who is perpetrating crimes, but what are the factors that lead to incidents of police violence. Whether this be the makeup of a community's demographics, the information on a specific person, or the employment demographics of the policemen themselves.

Previous research has focused on a variety of areas such as the incidence of any violent event occurring and using gradient boosting in order to predict future occurrences of such events and using feature importance methods in order to ascertain the most prominent predictors in such

violent situations[1]. Other work has focused on more specific factors that led to incidences of police violence – such as mental health, demonstrating that a wide variety of causes can be attributed to the underpinnings of injustice by law enforcement. In fact, factors involving mental health and police violence appear to be particularly magnified for racial, ethnic, and sexual minorities[2]. These disparities of police violence and machine learning have actually been cited in a variety of contexts and research using big data and has lead not only to understanding the intricacies of police violence and minorities, but the targeting of minorities from algorithms used to predict generalized incidences of violence[3].

Our project is very cognizant of the challenging material that we are choosing to wrestle with. It was therefore important to also review literature about the impacts of research surrounding police violence and predictive algorithms. That is, it is well documented through analysis such as programs like COMPAS that there are extremely problematic implementations of algorithms that discriminate against people of color[4]. However, there is also researching assessing the feedback loops that iterate after the process of discrimination has occurred. That is, not only do algorithms determining where police should focus their time and efforts tend to suffer for racial bias and discrimination, but they also reinforce these biases – often exaggerating the outcome. Algorithmic based policing decisions can cause a dangerous runaway effect, specifically hurting disparate communities[5]. As we move forward with our work and research, we will make sure to consider all these effects and be conscious of the implications of our work in a societally charged and socially important area.

Data Acquisition

Due to the complexity of our data questions and the topic we are choosing to analyze, we surveyed a plethora of sources and ended up aggregating our information from multiple data sets. This involved a series of merges and joins, allowing us to perform original analysis. We have provided a slightly updated description of data acquisition from previous iterations of our project below.

When aggregating data, we first found county-level demographics and general health data from County Health Rankings and Roadmaps, a database created and funded by the University of Wisconsin and the Robert Wood Johnson Foundation. Data from 2016 for each state was merged, which yielded myriad descriptive variables for every county to provide context for

---

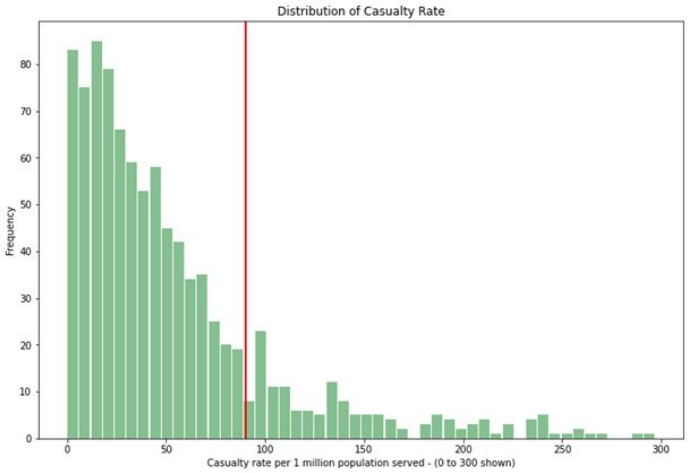
[1] Berk, Richard A., and Susan B. Sorenson. "Algorithmic Approach to Forecasting Rare Violent Events." *Wiley Online Library*, John Wiley & Sons, Ltd, 16 Dec. 2019, onlinelibrary.wiley.com/doi/full/10.1111/1745-9133.12476.
[2] DeVylder, Jordan E. "Association of Exposure to US Police Violence With Prevalence of Mental Health Symptoms." *JAMA Network Open*, JAMA Network, 21 Nov. 2018, jamanetwork.com/journals/jamanetworkopen/article-abstract/2715611.
[3] https://heinonline.org/HOL/LandingPage?handle=hein.journals/geolr52&div=8&id=&page=
[4] ProPublica. "COMPAS Recidivism Risk Score Data and Analysis." *ProPublica Data Store*, 19 Mar. 2019, www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis.
[5] Ensign, Danielle, et al. "Runaway Feedback Loops in Predictive Policing." *PMLR*, PMLR, 21 Jan. 2018, proceedings.mlr.press/v81/ensign18a.html.

police violence. We also collected the demographic and force size information of police departments across this country from the the Law Enforcement Management and Administrative Statistics (LEMAS) survey, which is administered every four years (most recent was 2016)[6]. We selected the relevant information on officer race, gender, supervisor race/gender, and collective bargaining rights.

Next, we combined three data sets from fatalenounters.org, mappingpoliceviolence.org, and the Washington Post. This data set had individual observations of police encounters that resulted in civilian death, along with a variety of information about many of the encounters. Our final data set was created via a series of joins involving counties, states, and Federal Information Processing Standards (FIPS) codes. The final outcome left us with several thousand data points representing comprehensive demographic information about communities and police departments around a given fatal encounter with a police officer.

Data and Methodology:

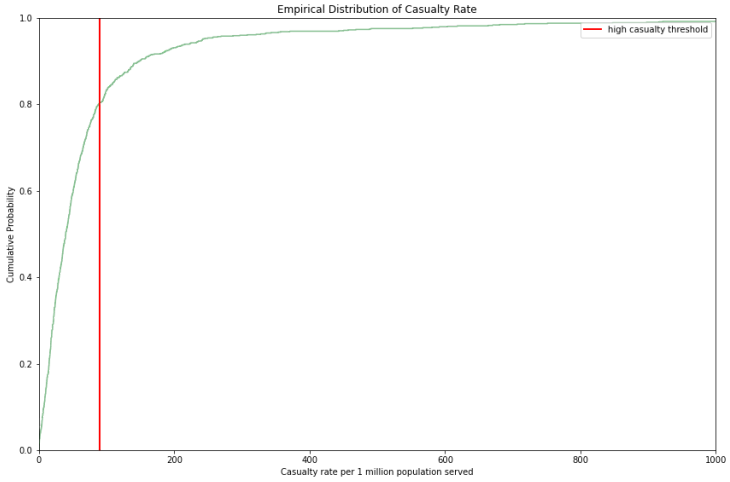
We determined the outcome variable by aggregating the number of casualties by police department and the dividing by the number of people served by the police department. Doing so gives us a "rate of casualties per population served" estimate. We use the population served by the department, and not the population of the county the police department operates in, because departments may have overlapping jurisdictions between different police departments in the same county. We then explored the distribution of our outcome variable:



The distribution is not bi-modal, and therefore we did not pick the median casualty rate for our cut off point for labelling our "high casualty" police departments. We therefore chose the point at which the trend of casualties tapers off, and we marked this cutoff to study the outlier departments that had higher than usual casualty rates: 90 casualties per one million population served. Since we controlled for population, the "high-casualty" police department indicator lets

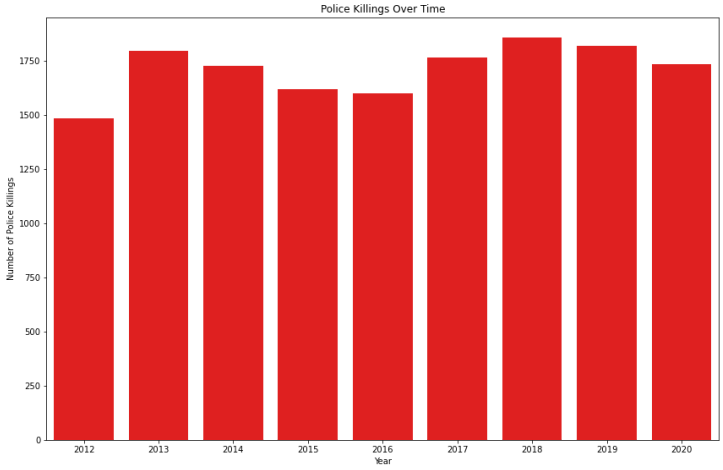
[6] The 2020 LEMAS survey is currently underway, but is yet to be completed. So, we used the best available data.

us study the factors that are associated with police departments that are responsible for high rates of casualties in police encounters.

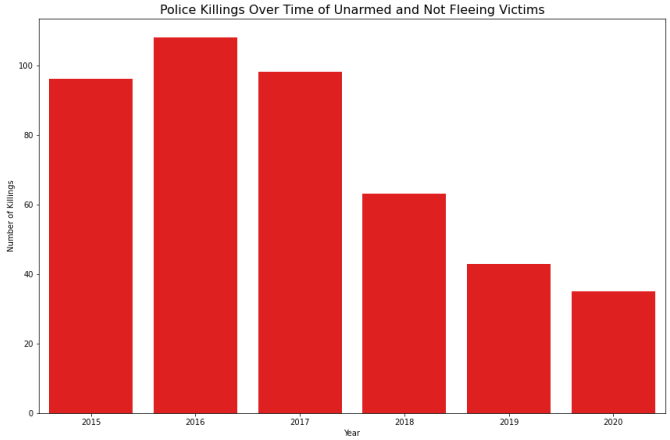


The empirical distribution plot below shows the proportion of the data that falls above and below the high casualty threshold: about 80% of observations fall in the low casualty class. Given that the classes are severely imbalanced, when building our models we used random downsampling in order to create more balanced classes. We chose not to use upsampling because creating "fake" data might be problematic given the sensitive nature of the data.

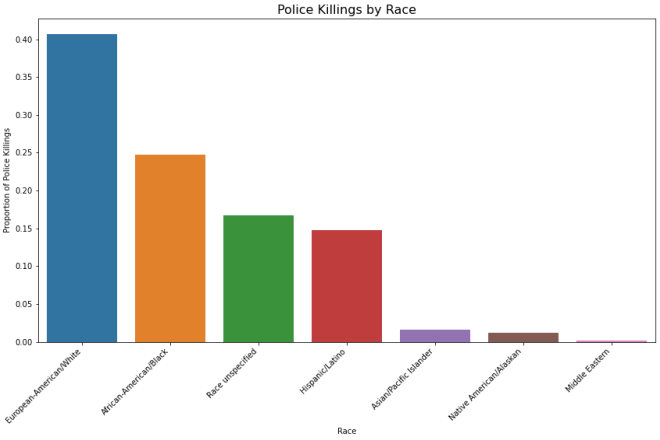
Upon determining our response variable, our next step in exploring the data was to illustrate the state of police killings in the United States in order to reaffirm why this is an issue that should be investigated. First, we explored how the number of police killings per year has changed over time. The plot below shows that the number of police killings is higher in 2020 than in 2012, which is the earliest year of police killing in our data.



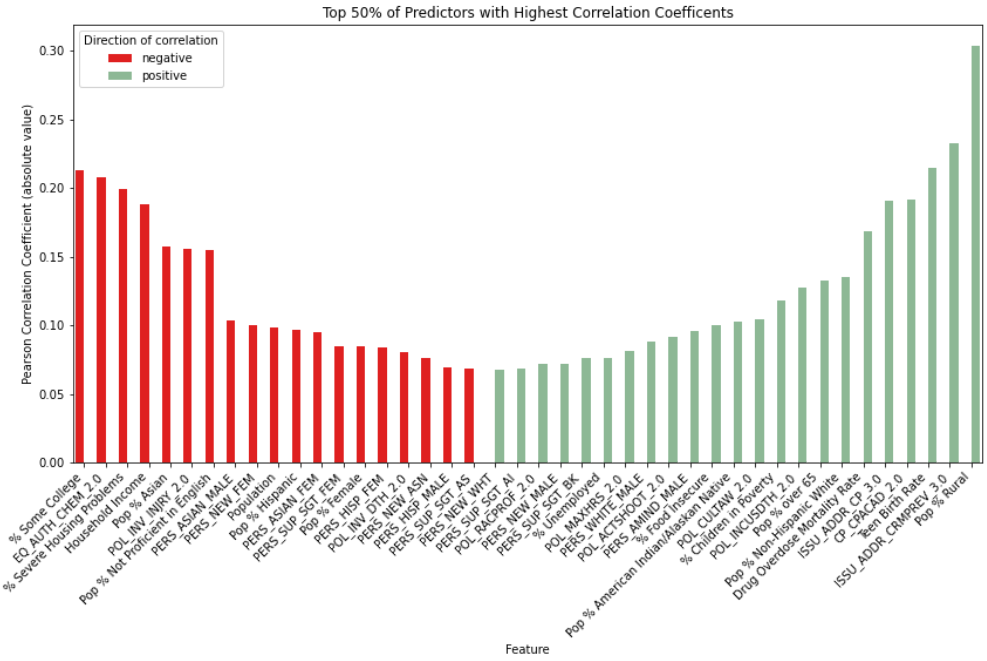
Diving deeper, we explored the number of police killings of victims who were unarmed and not fleeing. We thought this was important to explore given the legal guidelines that police are given in order to determine when the use of deadly force is "reasonable," and given that advocates for police reform claim that police use deadly force in situations when they could have used less-lethal methods. The plot below shows that the number of police killings of victims who were unarmed and not fleeing has decreased since 2012, but there are still a significant number of police killings of unarmed victims who were not fleeing. Although being unarmed and not fleeing is not conclusive evidence that the victim was not a threat to the police officer or others, it could suggest the police officer could have used less-than-lethal force. Killing another person should be avoided at all costs, and police officers should be trained on how to safely deescalate high-risk situations.



Another narrative put forth by advocates for police reform is that people of color are killed at disproportionately higher rates than Caucasian people. The plot below supports this claim. Though Caucasians do make up the largest proportion of victims, Black or African-Americans make up about 25% of police casualty victims, which is almost twice the proportional makeup of the entire U.S. population that identifies as Black or African-American.



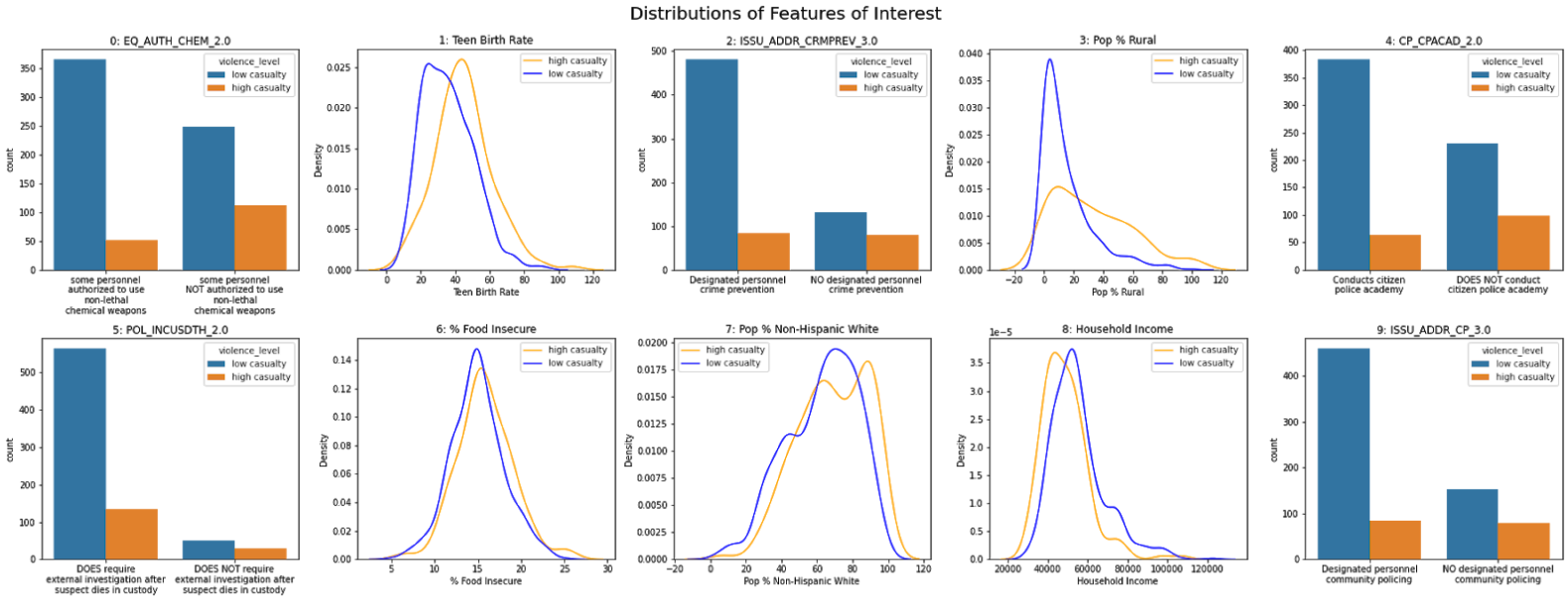
Diving into our final dataset that contains the total number of police killings across 970 police departments, we set out to explore our predictors and their correlations with police casualty rates. The plot below illustrates the pearson correlation coefficients of the top 50% of predictors with the highest absolute correlations.



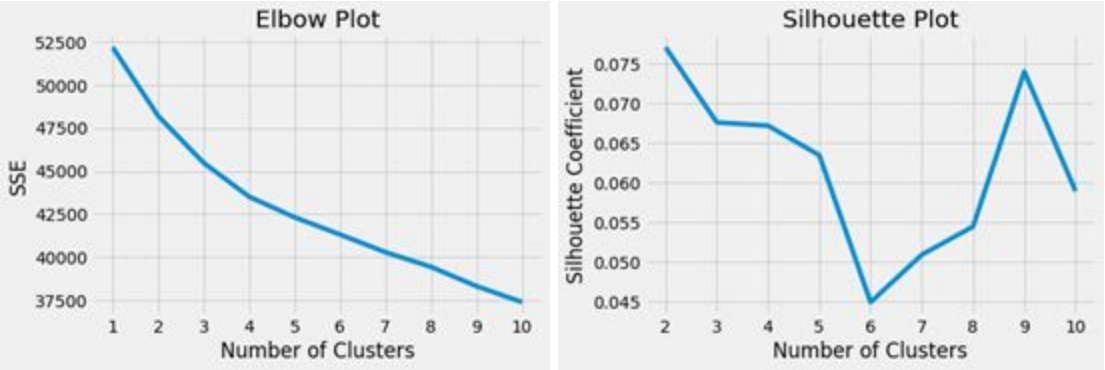
We then teased out 10 features from the plot above that we thought were interesting given our primary research questions: to explore the correlates between police policies and high casualty police departments (PDs), and the correlates between population demographics and high casualty PDs. Observations from each plot below:

1. **EQ_AUTH_CHEM_2.0:** Higher frequency of low casualty PDs among those that authorize some officers to use non-lethal chemical weapons, as opposed to not authorizing any officers to use them.
2. **Teen Birth Rate:** The density of high casualty PDs is greater in counties that have higher teen birth rates
3. **ISSU_ADDR_CRMPREV_3.0:** The frequency of low casualty PDs is higher among those that have personnel designated for crime prevention.
4. **Pop % Rural:** The density of high casualty PDs is greater in more rural counties.
5. **CP_CPACAD_2.0:** The frequency of high casualty PDs is higher among those that do not conduct a citizen police academy.
6. **% Food Insecure:** The density of high casualty PDs is greater in counties that have higher percentages of food insecure people.

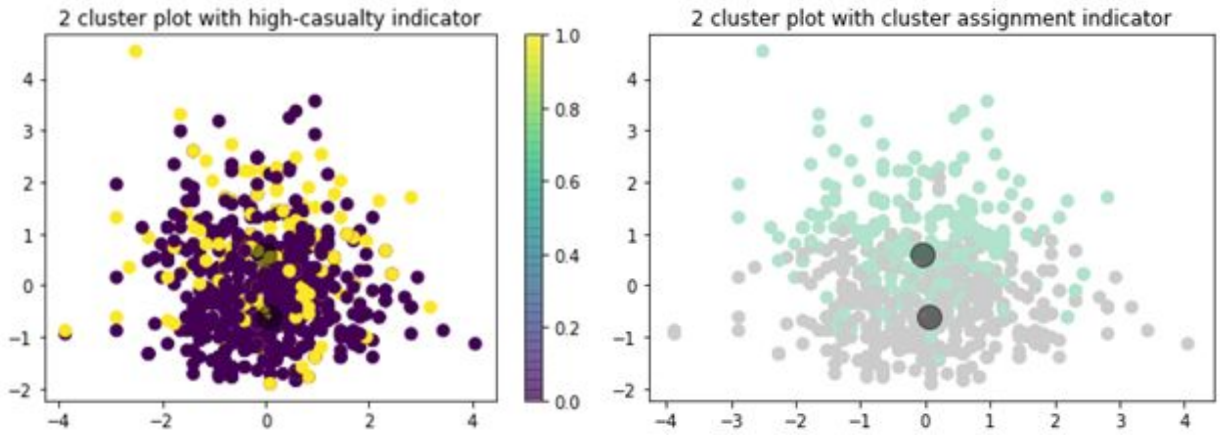
7. **Pop % Non-Hispanic White:** The density of high casualty PDs is greater in counties where the populations have extremely high percentages of Non-Hispanic Whites (possibly because our data contains many more observations of Caucasian victims).
8. **Household Income:** The density of low violence PDs is greater in counties that have very high average household incomes.
9. **ISSU_ADDR_CP_3.0:** The frequency of low casualty police departments is higher in counties that have designated personnel to address community policing.



To further explore the associations with high casualty police departments, we also used K-means clustering to evaluate whether there is a clear distinction between the two casualty levels in our clusters. We used an elbow plot and silhouette coefficient measures to study the optimal number of clusters.



The elbow plot does not give us a clear idea of the optimal number of clusters. However, the silhouette plot suggests a 2-cluster model, or a 9-cluster model, will produce clusters where any data point is closer to its own cluster than other clusters. Since we are interested in high-casualty police departments, we performed cluster analysis with two clusters.



The cluster analysis shows that our two clusters are not clearly differentiable in two-dimensions. However, there may be some separation between the two classes along the y-axis (Dimension 2) as indicated by the yellow-colored dots. This is being reflected in the pale green cluster shown in the second plot. Although in two-dimensions we may not be able to clearly distinguish this by eye, algorithms such as support vector machines with soft margins (to account for the relative lack of separation between the two classes) may distinguish between the two classes. We decided to use GAM models because it is a powerful prediction method and it allows predictor-level analysis for inference. Since we are interested in both social parameters and department level parameters, ceteris paribus, using GAMs, we can hold other variables to their means and explore a single instrumental predictor's association with high-casualty police departments.

Our EDA indicated that there may be a relationship between the police policies and high-casualty police departments, while taking into account the impacts of county level data. Due to this, we wanted to look at something more malleable to change, like a policy within a police department, as opposed to county demographic and socio-economic features where change is more enduring. This resulted in us revising our final problem statement to be: what kinds of police policies contribute to predicting incidences of police caused casualties?

Model, Results and Discussion

As stated previously, we ultimately used a logistic GAM model due to the ability to interpret the partial relationship between the response variable and the predictors, it's ability to regularize. However, while trying GAM modeling we also looked at other models as well including: random forest, SVM, and logistic regression with lasso regularization. We tried KNN

imputing the police department personnel level information, for which many variables had missing values. Using the imputed personnel data made our model less accurate. Since we were not confident about fair imputation of sensitive data, we used the non-imputed data, leaving out the personnel level information. After removing the observations with missing variables, we downsampled our data to balance the classes, ending up with 537 total observations. Despite the attempts at various models, the GAM model gave us the highest accuracy for the overall model at 72.98%. More importantly, both the positive and negative response variables predicted with similar accuracies: positive variables predicting with 71.93% accuracy and negative variables predicting with 73.25% accuracy. Additionally, the false positive rate for this model was low at 26.75%.

Our goal of this model was to not only predict the violence level of a police department with good accuracy and a low false positive rate, but also to observe which police policies are potentially positively influencing the high casualty rate. In order to do this we looked more deeply into the relationship between the predictor and response variable in the model. We viewed the partial dependence plots of the model for each feature to observe this relationship and we tried ELI5 which drops the variable in question to record the change in accuracy. We found that our demographic predictors had more explanatory power in our model than the police policies, as seen in the partial dependence plots. Interestingly, although the partial dependence plots show a small marginal impact of the police policies on the response variables, the ELI5 feature importance measure shows that police policies are important features for our model. This may be due to how the two measures are created. ELI5 removes the variable from the model and evaluates the model's accuracy with it gone, while GAM partial dependence marginalizes the other features. One plausible interpretation is that the police department level policies work instrumentally with other variables, demographic or otherwise, to explain the variations in the model, but are not correlated with the outcome variable by itself.

The most notable finding is that violent crime rate is not associated with high-casualty police departments. Neither in our ELI5 measure or our partial dependence measure shows any evidence that police violence is associated with violent crime rates. High casualty police departments were positively associated with more uninsured persons under the age of 65, higher unemployment rates, and higher teen birth rates. This is in accordance with research that shows police violence is more common in poor minority neighborhoods where these issues are more salient.[7] For police policies, the department's policy on crime prevention was an important predictor in our ELI5 measure. Police departments that did not have designated personnel for crime prevention are more likely to be classified as high casualty police departments.

Based on our model, the police department policies included in our data were not significantly correlated with high casualty police departments. However, the pairwise pearson correlations and density plots in our plots did suggest that certain police policies (especially

---

[7] Feldman, Justin M, Gruskin, Sofia, Coull, Brent A, and Krieger, Nancy. "Police-Related Deaths and Neighborhood Economic and Racial/Ethnic Polarization, United States, 2015-2016." *American Journal of Public Health (1971)* 109, no. 3 (2019): E1-464.

around community policing and crime prevention policies) were correlated with high casualty police departments. One reason for these conflicting findings might be that the county demographic predictors accounted for more variability in the casualty rates of police departments, so the model relied upon those rather than the police department policies.

In order to better investigate the impact of police department policies on casualty rates, we would want to pull in data on police violence that did not result in death, as well as non-violent arrests. Collecting this data might help the model better distinguish police departments that use deadly force at higher rates. Moreover, we would want to collect data on clearance rates in order to account for how effective the police departments are at solving crime, especially violent crimes.

Conclusion:

In our analysis, we study high-casualty police departments and the police policies, demographic information, and socio-economic factors that influence these police departments. Our study is correlational, we do not claim any causal linkage between the factors we study and the outcome of this research. We study the racial composition of the population that each police department serves as well as the racial composition of the police department itself. The racial composition of the police department data was partially missing, and we imputed that information from other variables available to us. A second caveat is that we did not include the racial composition of the people killed by each police department because we do not have the counterfactual: the racial composition of the people who had non-fatal encounters with the police. As such, any results we have interpreted from our analysis consider these shortcomings of the study. From our analysis we find no correlation between violent crime rates and high casualty police departments. Furthermore, we find that higher casualty police departments are associated with poorer, minority communities. Lastly, most individual police policies were not bilaterally correlated with the outcome variable, although the policies added to our model's predictive power.

Impact Statement

The history of police violence and unqualified impunity for police officers is long standing in the United States. The extra-judicial killing of George Floyd this year has brought a much-needed reckoning in criminal justice reform, inspiring mass protests in many cities against police brutality and systemic racism. We recognize the sensitive social and historical context in which our research contributes and view our findings through the lens of this context. This study contributes to the literature on policing, and finds no evidence between violent crime rates and police violence. The study also contributes evidence of more violent policing in low-income communities. The development of this research affects police departments and the communities they serve.

As we strive to determine police policies that have a relationship to police violence, we understand this could impact how police departments evaluate their programs and how communities evaluate the efforts of their local police departments; therefore, there is great responsibility in how the model evaluates the policies. We do not endorse any causal inference taken from this study, and our finding does not include any directional relationships between police departments and casualty rates in low-income communities. The funding and training available to police departments in low-income communities may affect policing and casualty rates. Further study of this issue would benefit from independent, department-level policing data and counterfactual data regarding non-fatal police arrests.

Note: Without title page, table of contents, and charts, this paper is under 10 pages long.

Additional Citations

1. *County Health Rankings Reports*. (2016). Retrieved 2020, from https://www.countyhealthrankings.org/reports/county-health-rankings-reports.
2. United States Department of Justice. Office of Justice Programs. Bureau of Justice Statistics. Law Enforcement Management and Administrative Statistics Body-Worn Camera Supplement (LEMAS-BWCS), 2016. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2019-06-20. https://doi.org/10.3886/ICPSR37302.v1
3. Burghart, Brian. *Fatal Encounters*, 2018, fatalencounters.org/.
4. Tate, Julie. *Washington Post Police Shootings Data,* https://github.com/washingtonpost/data-police-shootings/blob/master/2015README.md
5. Sinyangwe, Samuel. *Mapping Police Violence*, 2020, mappingpoliceviolence.org/.