How Victim's Gender, Race and Location Might Affect their Chance of Gunshot-Incurred Death with Presence of Police Force in United States

Zichen Gong

2023-04-05

Abstract

The increasing influence of media has brought police violence incidents into the public eye. This study examines data from print and online media sources, focusing primarily on fatal shootings involving police officers. The findings suggest that men are more likely to be fatally shot than women and transgender individuals. States with larger male populations in the United States have a higher probability of experiencing fatal shootings. In terms of ethnicity, Native Americans/Alaskans are more likely to be fatally shot compared to European-Americans/Whites. States with larger African-American/Black and Hispanic/Latino populations have a higher likelihood of fatal shootings. Although the study also explores non-fatal shooting incidents, the overall performance of the models is not as robust as those for fatal shootings.

Introduction

Police forces are established to maintain law and order while safeguarding people's lives and property. However, there have been instances where deaths have occurred in the presence of police officers, which were not directly the result of intentional police actions but rather due to police or officer presence at the scene.[2] Such violence involving police presence has increasingly attracted public attention in recent years, both within and beyond the United States. For instance, Canada's Winnipeg police recently disclosed details of male suicide in police presence.[4]

Many economists believe socioeconomic status (SES) is crucial in determining access to resources, including education and healthcare. In their book Socioeconomic Inequality and Educational Outcomes, Broer et al. mentioned that scholars and policymakers have long studied the relationship between SES and educational achievement, striving to reduce the educational outcome gaps caused by SES disparities.[1] The prevailing view in the Trends in International Mathematics and Science Study (TIMSS) is that there is a positive relationship between SES and educational outcomes. Building on this premise, SES differences arising from potential gender/race discrimination could further impact some individuals' educational outcomes, subsequently affecting their well-being and risk of victimization. Regarding health problems, Steptoe and Zaninotto found that SES is a decisive factor in many health issues for older populations, with less affluent individuals aging faster than their privileged counterparts.[7] Rahman et al. discovered that in the context of an unexpected large-scale health and economic crisis like COVID-19, low-income populations in Sichuan Province, China, are more vulnerable to poverty due to their reliance on agricultural income and government transfer payments.[6] The existing SES disparities, likely attributable to gender/race inequality, may exacerbate inequality in education, healthcare, and other outcomes, as well as differing resilience, potentially leading to increased crime or violence.

Various factors, including officer pursuits, suicides, and homicides, contribute to these deadly incidents. In this study, we will examine the likelihood of one specific type of deadly force that resulted in these fatalities: gunshots. We aim to determine whether some or all of the indicators, including the subject's gender, race, and the state of the incident, may relate to gunshot-related deaths. Preliminary data summaries and visualizations reveal significant variations in gunshot case counts across U.S. states, genders, and races.

A higher number of police-involved deaths were observed in some West coast states. However, when adjusted for state population, the proportion of gunshot-related deaths significantly diverges from the results based on raw death case counts. States in the central U.S. exhibit the highest proportion of deaths per population after adjusting for state population. Furthermore, states with higher average per-capita income tend to have a lower proportion of gunshot-related deaths, but the trend is not visible enough. Male victims overwhelmingly outnumber female victims, and white victims constitute a significant proportion. Nevertheless, this observation becomes less clear when compared to the population distribution of each race in the United States. After adjusting for race and ethnicity, the African-American/Black population is more likely to experience gunshot-related deaths in incidents involving police violence.

The findings above indicate the necessity of employing mathematical and statistical tools to verify the associations between gunshot-related fatalities and factors such as gender, race, and region rather than relying solely on empirical images and tables. In this study, we have established two basic regression models. The first model is a logistic regression model, which takes the binary variable of whether a victim dies from a gunshot in the presence of the police as the dependent variable and uses the victim's gender and race as two factor variables for independent variables. The second model is a linear regression model that replaces the individual level with the state level, using the total number of victims who died from gunshots/non-gunshots in a U.S. state as the dependent variable and incorporating the state's racial distribution data, gender distribution data, and per-capita income data as independent variables. Additionally, we briefly transformed both models into regression trees for some basic machine learning prediction. The study results indicate that a certain degree of gender and racial inequality exists in such violent incidents, while regional differences are not as pronounced and do not align with the findings depicted in the maps. This discrepancy may arise from the varying gender and population compositions of different states, contributing to the differences in regional distribution displayed on the maps. Income disparities do not exhibit high significance in either model.

As Kevin Lang and Ariella Kahn-Lang Spitzer mentioned, the issue of inequality is fundamentally an economic problem.[5] Focusing on areas prone to producing real or perceived inequalities can help guide the development of anti-discrimination policies. They suggest that discrimination may manifest as "taste-based" and "statistical." The former reflects personal biases or preferences, while the latter is based on empiricism. In the context of this study, it is possible that, in the presence of law enforcement, some government officials or police officers may allow personal biases to contribute to tragic outcomes, representing the first form of discrimination. Alternatively, heightened vigilance by police officers in response to specific demographics with historically higher crime rates could also play a role, illustrating the second form of discrimination.

Data

Summary Statistics

Fatal Encounters, a team of researchers from the University of Southern California, has been tracking atypical deaths during interactions with police since January 1, 2001, using social media and police reports as sources.[2] While media reports may only sometimes be accurate and information provided by the police can be incomplete, this team has diligently compiled this dataset, transforming it into a searchable national resource. However, the dataset's completeness varies across years, with data from some years needing to be completed more. Consequently, this inconsistency may impede time series analysis, necessitating statistical inferences based on cumulative case results without considering time.

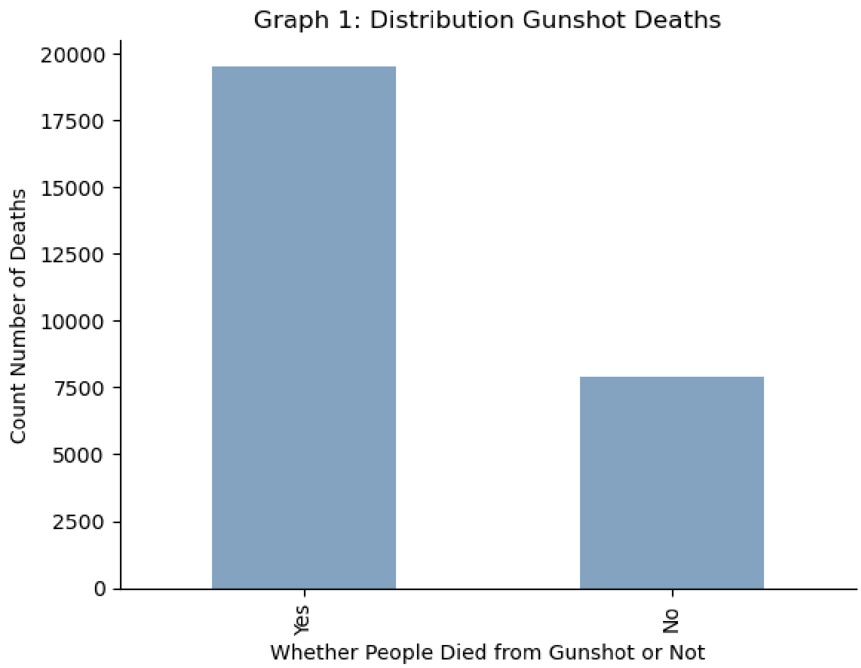
Table 1: Count of Words with Highest Frequency in Death Scene Description

Keywords	Count
	735914
man	5261
car	5154
gun	3985
vehicle	3869
fired	3459
killing	2814
home	2632
chase	2117
crashed	1962
driver	1947
armed	1827
pulled	1784
traffic	1776
fled	1693
p.m.	1691
struck	1684
ran	1680
driving	1650
hit	1537
shooting	1509
suspect	1504
a.m.	1498
woman	1433
stolen	1358
refused	1350
knife	1224
shots	1173
truck	1108
drove	1082

In Table 1, we found that the frequency of "men" in scene descriptions is much higher than that of "women." In the time description, "p.m." appears more frequently than "a.m." At the same time, some words that reveal violent methods, such as "gun," "vehicle", "car", and "shooting", or some intense or relatively harmful verbs, such as "chase", "crashed," and "fled" also appears frequently, as if showing us that the scene where death appears is not peaceful.

Table 2: Summary Statistics for Distribution of Gunshot Death

Death by Gunshot	count
Yes	19528
No	7927

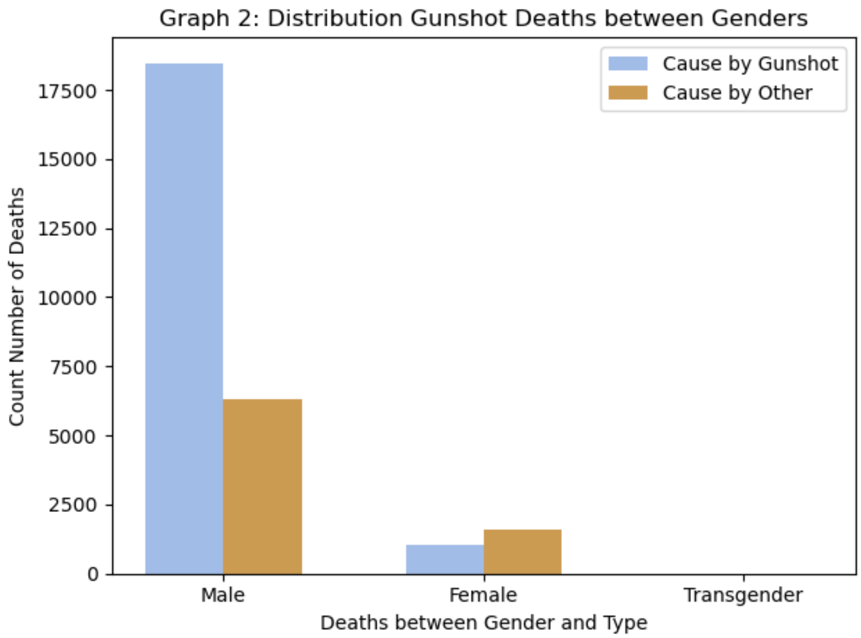


The response variable of our model is a binary one, indicating whether one victim died from a gunshot. In Table 2, gunshots cause the most deaths, and the number of gunshot deaths is more than twice as many as the number of other deaths.

Graph 1 is a visualization of Table 2. Similar to what we read from Table 2, the number of people who died from a gunshot is more than twice as many as people who died from other reasons. This binary variable is our response variable.

Table 3: Summary Statistics for Distribution of Death by Gender and by Type

Gender	Cause by Gunshot	Cause by Other
Male	18468	6332
Female	1049	1587
Transgender	11	8



According to Table 3, no matter whether it is caused by gunshot or other factors, the death count is highest for males than for females and the transgender population. Besides the fact that the transgender population is a smaller group compared with the male and female groups, this table might indicate that the male population is more likely to become victims in encounters with the police. However, we should remember that the death statistics also include officer deaths. In the year 2021, as published by Statista showed that among U.S. law enforcement officers, the gender ratio is highly skewed, with a predominantly high proportion of 72.2% as male employees.[3] As a result, a certain proportion of male deaths might come from officers as well, which makes the count of male deaths much more than the number of female deaths.

From the previous graph, we know that the number of deaths caused by gunshots is much more than the number of deaths caused by other factors. In Graph 2, we see that for females, the number of deaths caused by other factors is slightly more than those caused by gunshots. At the same time, the overall trend fits that from Table 3, where the number of male deaths is much more than the other two categories. The variation of death counts between genders shows that there might be association between gender of victims and types of deaths.

Table 4: Summary Statistics for Distribution of Death by Race and by Type

Race	Cause by Gunshot	Cause by Other
European-American/White	9855	3729
African-American/Black	5205	2486
Hispanic/Latino	3331	1263
Race unspecified	489	243
Asian/Pacific Islander	393	136
Native American/Alaskan	214	62
Middle Eastern	41	8

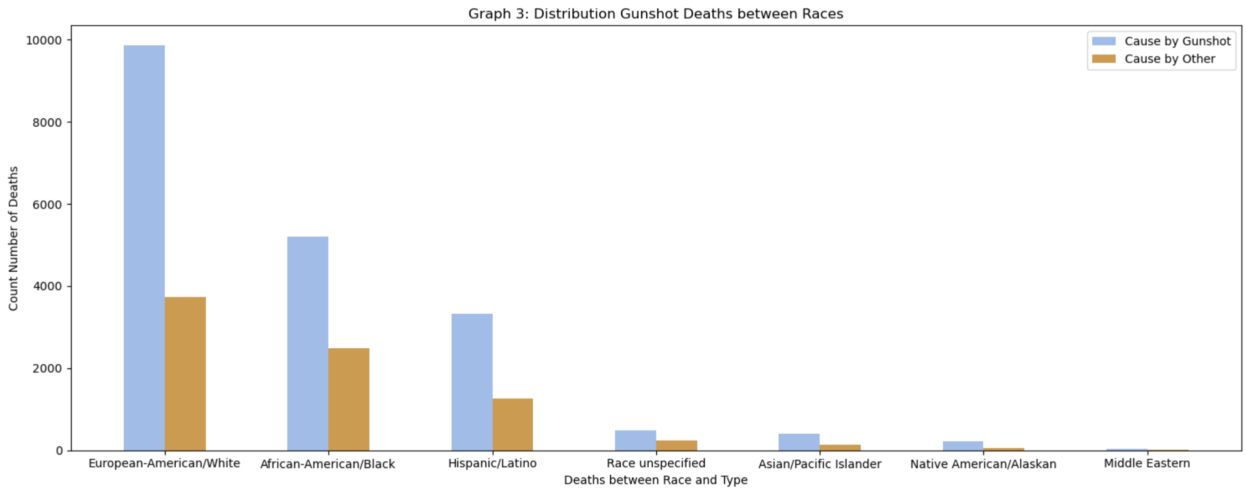
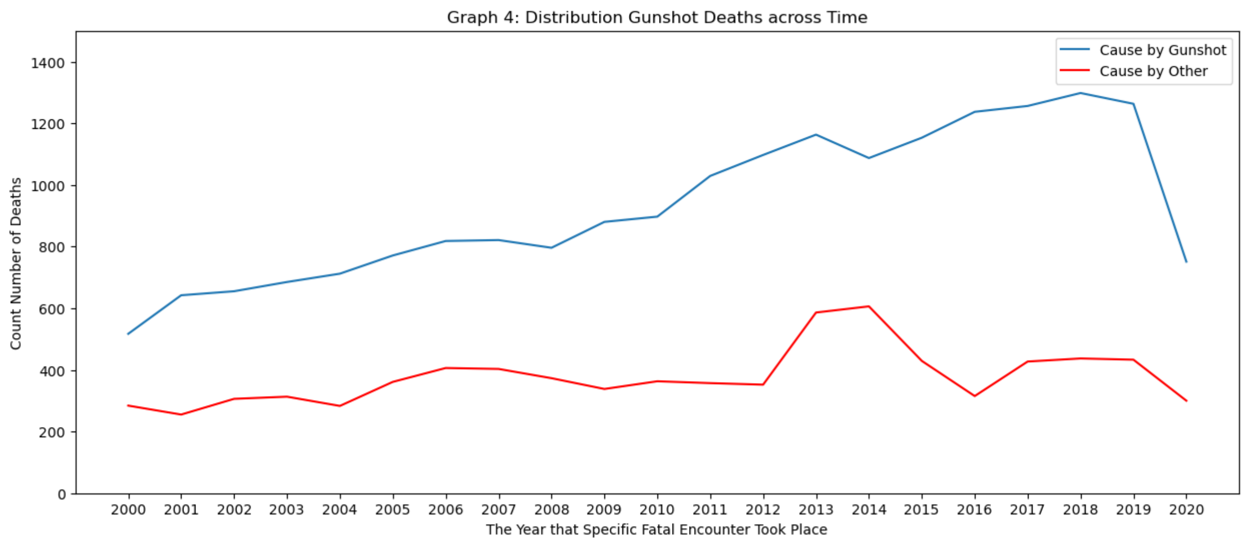


Table 4 lists the death count for different races, and these counts are categorized into two groups indicating a gunshot death or other deaths. Many white people who died compared to other races seem to come from the primarily white population proportion in the U.S. However, compared with the proportion of the white population in 2021 U.S. census data, the white population takes up to approximately 75.8% percent, which is five times more than the black population.[8] While in Table 4, the number of white deaths is not even twice as many as black deaths. If we consider the white and black populations, the proportion of white people who died is much smaller than that of black people.

According to the 2021 U.S. census, the proportion of the Hispanic population is approximately 5 percent more than the proportion of the African-American population, and the white population dominates at around 75.8 percent.[8] In this sense, if we look at Graph 3, the number of deaths that happened to African-Americans is almost half that of European-Americans. In addition, the number of African-American deaths also exceeds the number of Hispanic deaths, even though the Hispanic population is, in fact, more significant than the

African-American population. The variation of unproportionate distributed death counts between races shows that there might be an association between victims' races and types of deaths.



We can extract from Graph 4 that there seems to be no significant rising of deaths for non-gunshot deaths, but indeed there is a rising trend for gunshot deaths. However, what is included in this graph is incomplete. First, we should be aware that prior to the year 2013, as indicated by Fatal Encounters researchers, their record of races was "spotty".[2] In this case, many earlier data that needed significantly more information might have already been eliminated during our data cleaning steps. Second, all data collected was from news media. Around twenty years ago, news pressing's ability to collect data and researchers' methods of collecting data were limited. Both factors might contribute to the seemingly upward trend of death cases increase. As a result, I still decide to use the data cumulatively rather than further divide it by time.

Visualizations

Because the primary purpose of this research is to investigate any association between one person's chance of dying from a gunshot police fatal encounter and this person's gender, race, and location, an ideal scraped dataset is likely to supplement information regarding this topic. Right now, based on my original dataset, what I lack significantly is information related to the situation of each state.

I incorporate state as one of my predictors because geographical location can represent an approximate wealth situation, race/ethnicity group population distribution, climate situation, production structure, and so on. However, if I want to investigate the influence of income on the chance of facing violence directly, I would better include a dataset about income in my analysis.

In addition, as I have emphasized above, my maps and graphs related to the total number of gunshot deaths and gunshot deaths distributed among different races can reveal a biased result because many cases can be pure because of the high proportion of a particular population. For example, suppose the white population takes up to 70% of the U.S. population. In that case, there is no doubt that most cases will occur among white people unless there are extreme cases in which one or more specific races face mistreatment and discrimination during police actions.

For this part, I will first use HTML-based web scraping to acquire a dataset related to average state incomes.[10] In particular, I will focus on per-capita income, which can avoid influence from population factors. This can be either a supplement or a substitute for the indicator state. It focuses on one specific aspect that geographic information can reflect rather than a mix of multiple factors.

In order to use this scraped dataset for further analysis, I will merge this income dataset with the original fatal encounter dataset on the variable "State." We can consider using the average per-capita income to represent the income of each victim and add this variable to our regression model. Although this is not the perfect approach because victim's income might not be accurately reflected upon by the state's average income, we have no better choice. The Fatal Encounters did not collect any salary data.

Suppose we see any significant association between the chance of facing a gunshot and victims' income. In that case, it can also help to conclude that victims' location can significantly influence their chance of facing violence, as we use the average income level in each state. However, to avoid multicollinearity issues, I will not simultaneously add the indicator state and the indicator income to my regression model. I will run two separate regressions, one with each indicator. The model with the variable state can have a general appearance of geographical influence, and the model with the variable income focuses specifically on geographic wealth distribution.

Technically speaking, the web scraping process was smooth sailing for me during this attempt. I could efficiently access the source code of websites, and Wikipedia permits the use of information obtained through web scraping for personal use. A more challenging aspect for me was determining the accuracy of the information I obtained from Wikipedia. This website is typically deemed unsuitable for citation in scholarly writing due to its editable nature by any user.

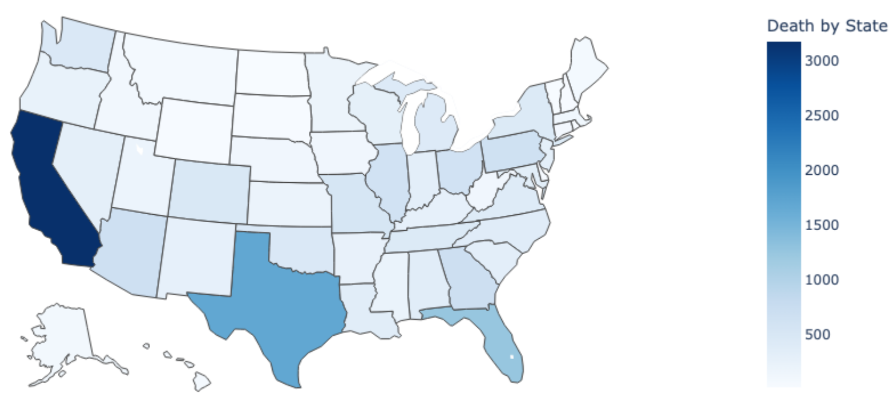
Fortunately, Wikipedia was one of many websites I searched during my information-hunting process. I initially discovered a primary source from the U.S. Census Bureau. Subsequently, I found that the Wikipedia source reproduced the census data, presenting it in a more organized and accessible tabular format. As a result, having located the primary source material, I proceeded to scrape the Wikipedia table as it is in a tidy version that is easier for data cleaning and merging and incorporate it into my research.

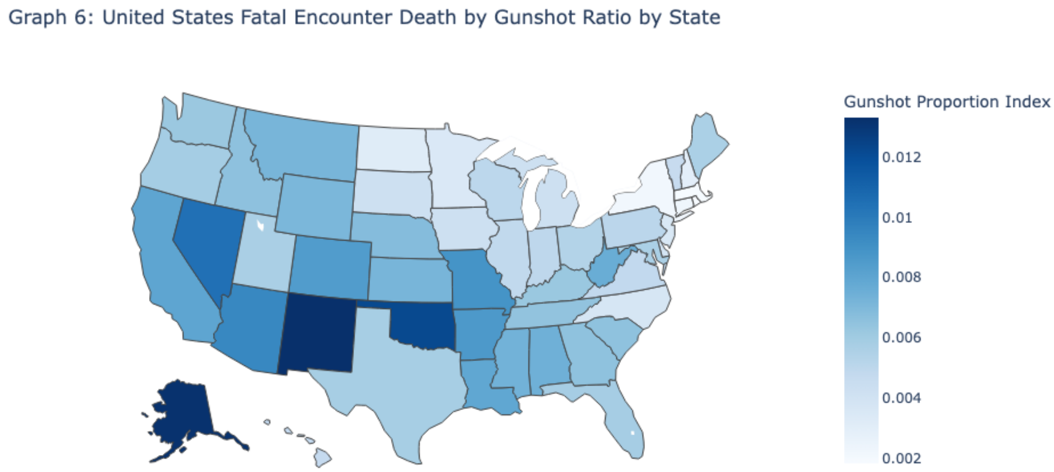
One issue with the fatal encounter dataset is the lack of information regarding the specific states where the incidents occurred. It may be worthwhile to consider incorporating descriptive information about the states to supplement my data analysis. Through preliminary data visualization, we can observe that most deaths due to police shootings are concentrated in certain states along the West coast of the United States. However, this may be attributed to the significantly larger populations in some states than others, which could result in a higher number of overall cases even if the proportion of deaths is not high.

For example, our analysis reveals that in most states, white individuals are the largest racial group among gunshot victims. However, considering the previously mentioned factors, although the United States is a melting pot, the white population constitutes the majority. Consequently, upon calculating the fatalities, the group with the highest number of deaths may not necessarily be white individuals, even though they represent the largest number of victims. This is why it is crucial to incorporate data on state populations and the number of individuals from various racial groups in each state.[11]

Similarly, the data is sourced from the U.S. Census Bureau. I scraped the neatly formatted table on Wikipedia for convenience in subsequent data merging, ensuring the data source remains credible.

Graph 5: United States Fatal Encounter Death by Gunshot Count by State

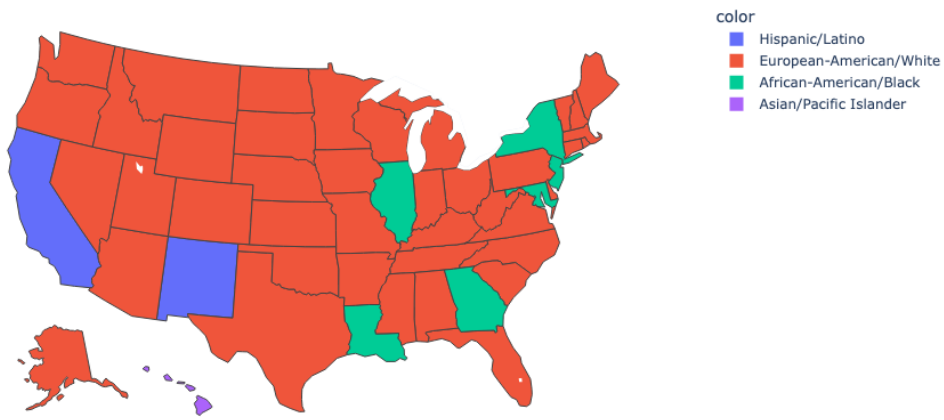




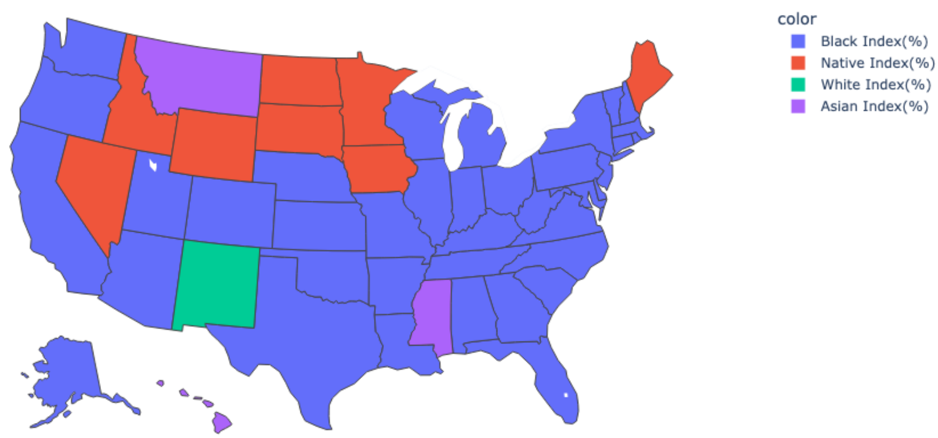
Now we only take a look at gunshot death distribution in the U.S in Graph 5. The three states with the highest number of deaths are California, Texas, and Florida. This does not seem to provide much new information as these states have the highest population.[9] A state with a higher population and a higher number of fatal encounters might not simultaneously have the highest rate of deaths. Nevertheless, besides California, the other two states do not have a high enough population density, and the number of fatal encounters does not seem to correspond to population density.

Graph 6 displays the proportion of deaths resulting from police-involved shootings in each state. Unlike Graph 5, which shows the highest number of fatalities in California, Texas, and Florida, when the population data of each state is incorporated to calculate the death rate (i.e., the number of deaths per capita), the highest mortality rates are primarily concentrated in several states in the central United States, as well as Alaska. This underscores the importance of considering the state as an indicator. The varying population sizes across states can significantly impact the likelihood of an individual falling victim to lethal police violence.

Graph 7: United States Races which Faced the Greatest Number of Gunshot Death by State

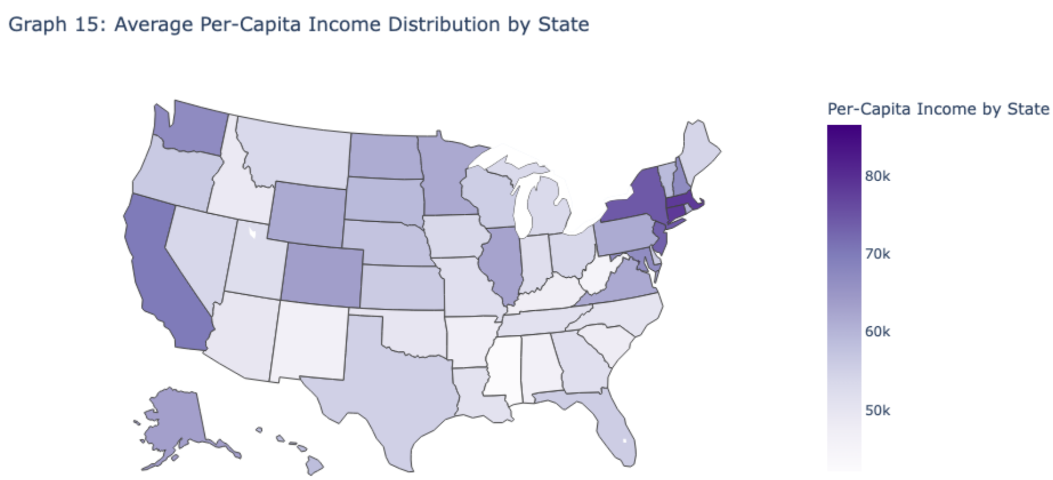


Graph 8: United States Races with the Greatest Proportion of Gunshot Death by State



Graph 7 illustrates the race with the highest number of gunshot deaths with police presence in each state. For most states, the races of victims seem irrelevant to the state. This is because the white population is more than twice as many as any other race population, so they should not surprise us when we see that most states have white people to be killed the most. However, there are several deviates. West states, California and New Mexico have the most Hispanic/Latino people killed. For New York, Maryland, New Jersey, Illinois, Georgia, and Louisiana, the number of African-American/Black victims outweighs other race populations. These are all states on the east side of the U.S. Finally, in Hawaii, Asians/Pacific Islanders have the most got killed.

Graph 8 highlights the importance of incorporating population data into the analysis. As previously emphasized, most of the United States population consists of white individuals, which could likely explain why in Graph 7, the highest fatalities are among white individuals in many states. However, upon calculating the death rate for each racial group, it becomes evident in Graph 8 that the highest mortality rates are predominantly among black individuals in the vast majority of states. In some northern states, Native Americans exhibit the highest death rates, whereas only in New Mexico do white individuals have the highest mortality rate. Therefore, after adjusting for the population sizes of each racial group, a more apparent association between the victims' race and fatalities resulting from police-involved shootings can be observed.



We can observe that states with higher incomes are primarily concentrated along the East and West coasts of the United States, with these regions generally exhibiting darker shades on the map. Notably, the higher number of fatalities along the West Coast and in the South does not appear to be directly correlated with income levels. Assuming that higher-income areas experience less violence, the West Coast region would exhibit the highest violent incidents.

Conversely, if higher-income areas experience more violence, the East Coast region would exhibit darker shades in Graph 5. Consequently, it is evident that in addition to income, other factors are encapsulated within the state indicator that influences the number of fatalities resulting from police shootings.

Methods

We do not need to assume a linear data distribution in logistic regression models. However, we do assume a linear relationship between the logit of the probability of the outcome and the predictor variables. In the context of this study, the outcome is whether the victim died from a gunshot, a binary variable, and the predictor variables are Gender, Race, and State (which can be replaced by the Population of each state and the average per-capita Income of each state). The logit function helps transform a logistic regression problem into a linear one.

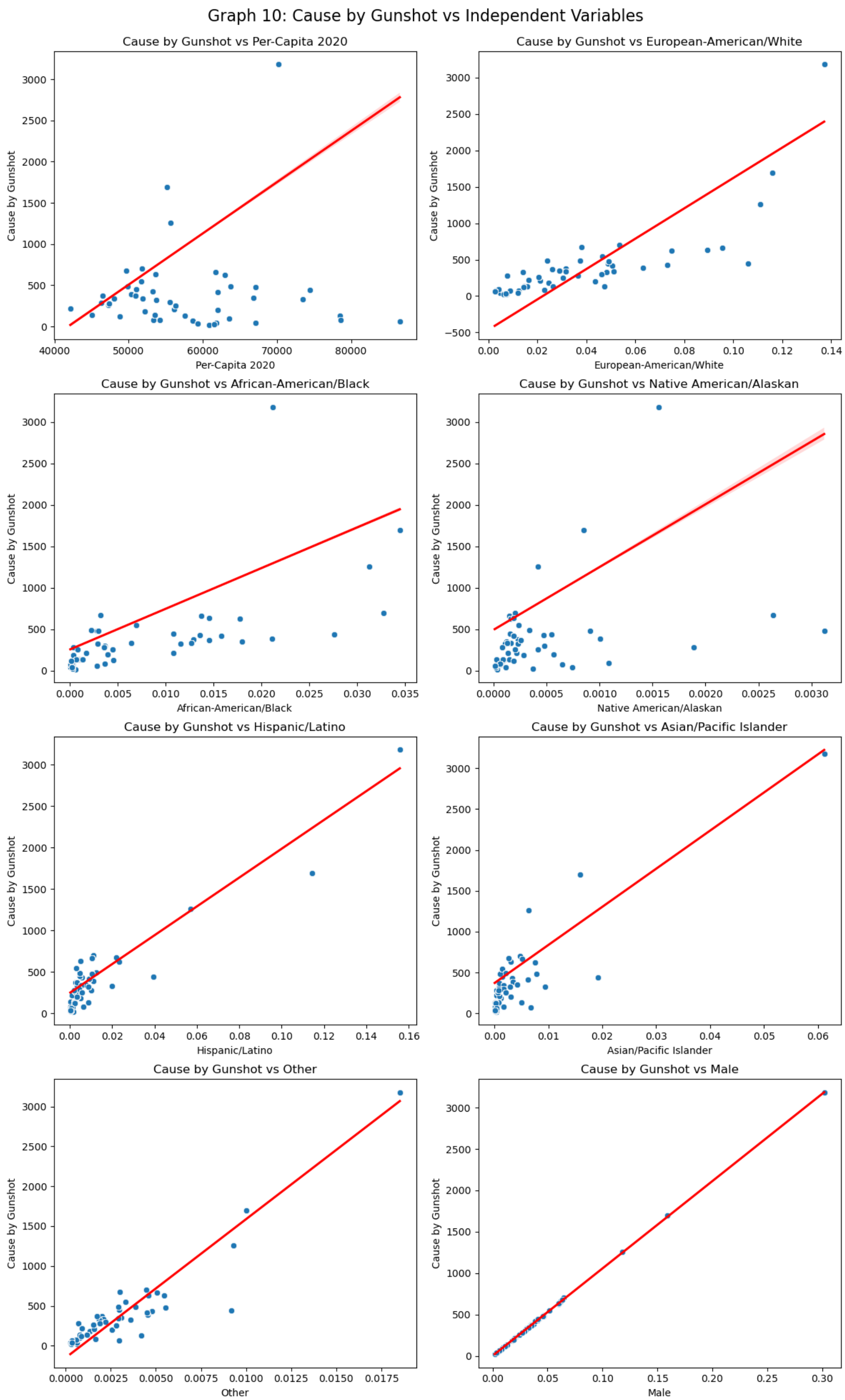
Subsequently, I conducted another secondary research study examining the relationships between the number of fatalities in police-involved shootings and other police-involved fatal incidents across all U.S. states and various demographic and socioeconomic factors. These factors include the average per-capita income in the year 2020, the populations of European-American/White, African-American/Black, Native American/Alaskan, Hispanic/Latino, Asian/Pacific Islander, other race/ethnicity groups, and the male population. This study employed Ordinary Least Squares (OLS) regression models to analyze these relationships. The existing literature has already demonstrated the linear relationships between each independent variable and dependent variable pair. However, the primary difference between the previous model and the current one is the unit of observation. The former model focused on each victim as an individual observational unit. In contrast, the current model aggregates the number of victims per state and utilizes each state as an observational unit. Consequently, the methodology transitioned from a logistic regression model to a linear regression model.

As Lang and Spitzer mentioned, discriminatory attitudes from employers, colleagues, and potential customers could result in unequal market outcomes.[5] These factors might make it more challenging for specific gender/race groups to succeed in the labor market or result in lower income levels. Lower-income individuals may be more likely to become involved in crime or live in more dangerous neighborhoods. They also mentioned that discrimination could exist in the criminal justice system, including police actions and court decisions. As the subjects of this study have all passed away, this might have led to a particular bias in the distribution of gender/race within the population. Furthermore, as a more severe form of mortality, gunshot deaths may also differ across genders and races.

When comparing the compositions of the independent variables between the two research models, the first logistic regression model utilizes a binary variable representing whether each victim was fatally shot as the response variable. The long model includes the victim's gender, race, the state in which they died, and the average per-capita income in the year 2020 for their state as a proxy for their income (recognizing that this may have some discrepancy with the actual income of each victim, but this is due to the lack of income data in the original dataset). Furthermore, it accounts for the state's overall population and the populations of European-American/White, African-American/Black, Native American/Alaskan, Hispanic/Latino, Asian/Pacific Islander, and other race/ethnicity groups. A final model is obtained by comparing the variable coefficients' p-values and examining the variance inflation factor (VIF) data to address multicollinearity issues.

In contrast with the first research model, the second model's response variable is no longer at the individual level but at the state level. Besides this change, the model includes the aforementioned racial and ethnic populations for each state, the average per-capita income in 2020, and the number of male residents in each state.

Determining the relationship between the Cause of Death and Gender, Race, and State through qualitative/empirical studies alone is challenging and might be unreliable. Further research and statistical analysis are necessary to understand and validate the relationship between these variables.



Graph 10 illustrates the relationships between the total number of deaths caused by a gunshot in each U.S. state (the dependent variable) and various independent variables. The independent variables presented on the x-axis are: 'Per-Capita 2020' (average per-capita income in the year 2020 in each state), 'European-American/White' (European-American/White population divided by 10,000 for each state), 'African-American/Black' (African-American/Black population divided by 10,000 for each state), 'Native American/Alaskan' (Native American/Alaskan population divided by 10,000 for each state), 'Hispanic/Latino' (Hispanic/Latino population divided by 10,000 for each state), 'Asian/Pacific Islander' (Asian/Pacific Islander population divided by 10,000 for each state), 'Other' (Other race/ethnicity groups population divided by 10,000 for each state), and 'Male' (Male population divided by 10,000 for each state).

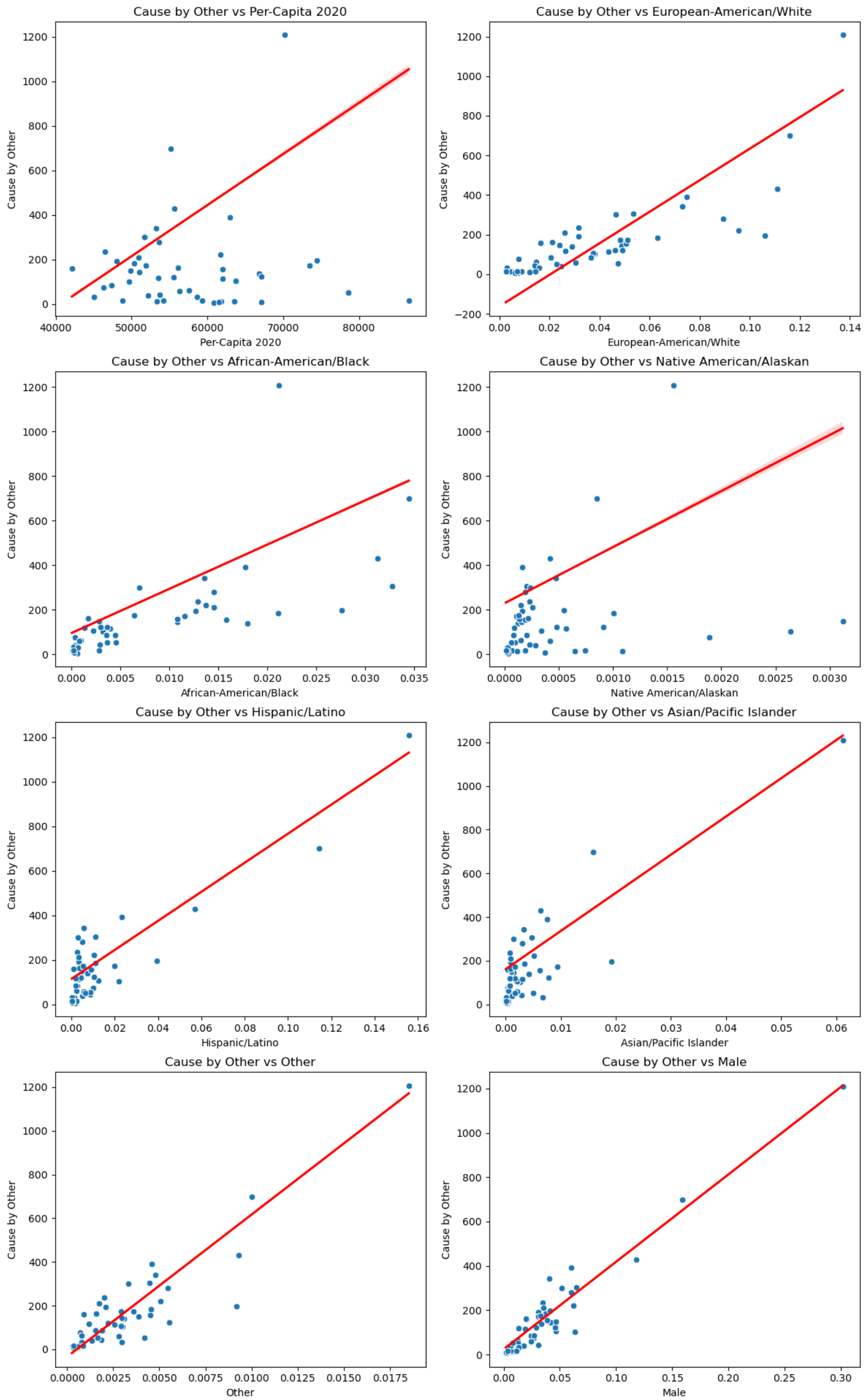
In the graph, each panel represents the relationship between the total number of gunshot-related deaths in each state and one of the independent variables. The scatterplot in each panel displays the observed data points, while the solid line represents the best-fitting linear regression line.

Visually, there appears to be a positive relationship between the total number of gunshot-related deaths and the European-American/White population, African-American/Black population, Hispanic/Latino population, Other race/ethnicity groups population, and Male population. This suggests that as these population groups increase, the total number of gunshot-related deaths in a state may also increase.

On the other hand, the relationship between the total number of gunshot-related deaths and the Native American/Alaskan population and Asian/Pacific Islander population seems to be less evident.

Regarding the 'Per-Capita 2020' variable, it is difficult to discern a strong or consistent relationship with the total number of gunshot-related deaths from the graph alone. Further statistical analysis would be required to determine the significance and strength of the associations between each independent variable and the total number of gunshot-related deaths in each state.

Graph 11: Cause by Other vs Independent Variables



Graph 11 illustrates the relationships between the number of deaths caused by factors other than gunshots in each U.S. state (the dependent variable) and various independent variables. Similar to Graph 10, the independent variables presented on the x-axis are: 'Per-Capita 2020', 'European-American/White,' 'African-American/Black,' 'Native American/Alaskan,' 'Hispanic/Latino,' 'Asian/Pacific Islander,' 'Other,' and 'Male.'

In the graph, each panel represents the relationship between the total number of non-gunshot-related deaths in each state and one of the independent variables. The scatterplot in each panel displays the observed data points, while the solid line represents the best-fitting linear regression line.

Visually, there appears to be a positive relationship between the total number of non-gunshot-related deaths and the European-American/White population, African-American/Black population, Hispanic/Latino population, Other race/ethnicity groups population, and Male population. These relationships are similar to those observed in Graph 18, which focused on gunshot-related deaths.

On the other hand, the relationship between the total number of non-gunshot-related deaths and the Native American/Alaskan population and Asian/Pacific Islander population seem to be less clear, as observed in Graph 10 as well.

Regarding the 'Per-Capita 2020' variable, the relationship with the total number of non-gunshot-related deaths is still not discernible from the graph alone. This is also consistent with the findings in Graph 10.

While there are similarities between Graph 10 and Graph 11 in terms of the relationships between the dependent variables and the independent variables, further statistical analysis would be required to determine the significance and strength of the associations, as well as any potential differences in the relationships when focusing on gunshot-related versus non-gunshot-related deaths.

Results

Regression Results

Table 5: Logistic Regression Result

Variable	Estimate	Standard Error
Intercept	-0.34***	0.04
Male	1.50***	0.04
Transgender	0.79*	0.47
African-American/Black	-0.28***	0.03
Asian/Pacific Islander	0.13	0.10
Hispanic/Latino	-0.06	0.04
Middle Eastern	0.65	0.40
Native American/Alaskan	0.32**	0.15
Race unspecified	-0.32***	0.08

Table 5 presents the results of a logistic regression model after model selection. The dependent variable is a binary variable indicating whether a victim died from a gunshot or other causes. The model selection process has led to the retention of several key independent variables, including Gender, Race, and their respective levels. The model is presented as:

$log\frac{p(Y=1)}{1-p(Y=1)} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

In the given model, Y represents a binary dependent variable, where 1 indicates that the victim died due to gun violence, and 0 indicates that the victim died due to other causes. X_1 is a categorical independent variable representing different genders, with the base level being female. X_2 is another categorical independent variable representing various races, with the base level being European-American/White.

In the final model, the Male level of the Gender variable has a positive and highly significant association with the likelihood of dying from a gunshot, with an estimated coefficient of 1.50 and a p-value of less than 0.01. This suggests that, compared to females (the base level), males have higher odds of dying from a gunshot, holding all other variables constant.

The Transgender level of the Gender variable exhibits a positive association with the likelihood of dying from a gunshot, with an estimated coefficient of 0.79 and a p-value of 0.1. This indicates that, compared to females, transgender individuals have higher odds of dying from a gunshot, although the relationship is less statistically significant.

Regarding the Race variable, African-American/Black, Native American/Alaskan, and Race Unspecified levels show significant associations with the likelihood of dying from a gunshot. African-American/Black individuals have a negative and highly significant association, with an estimated coefficient of -0.28 and a p-value of less than 0.01. This suggests that, compared to European-American/White individuals (the base level), African-American/Black individuals have lower odds of dying from a gunshot, controlling for other factors.

Native American/Alaskan individuals have a positive and significant association, with an estimated coefficient of 0.32 and a p-value of 0.05, indicating higher odds of dying from a gunshot than European-American/White individuals. Race Unspecified individuals have a negative and highly significant association, with an estimated coefficient of -0.32 and a p-value of less than 0.01, suggesting lower odds of dying from a gunshot than European-American/White individuals.

The levels of Asian/Pacific Islander, Hispanic/Latino, and Middle Eastern Race variables are not statistically significant in the final model.

The model's intercept has an estimated coefficient of -0.34 and a p-value of less than 0.01. The model has an overall sample size of 27,455 observations.

Table 6 presents the results of two linear regression models that examine the relationship between various demographic factors and the number of deaths caused by gunshots (Model 2) and other causes (Model 3) in each U.S. state. The independent variables in both models include Per-Capita Income in 2020, populations of various racial/ethnic groups (European-American/White, African-American/Black, Native American/Alaskan, Hispanic/Latino, Asian/Pacific Islander, Other), and Male population. All population variables are divided by

	Table 6: Linear Regression Result	
Variable	Caused by Gunshot	Caused by Other
Intercept	1.740***	2771.682***
African-American/Black	512.225***	7526.350***
Asian/Pacific Islander	-561.275***	1450.091***
European-American/White	-126.766***	301.433***
Hispanic/Latino	57.606***	3206.052***
Male	10548.182***	-25227.632***
Native American/Alaskan	-1650.594***	-25635.245***
Other	1840.083***	-0.003***
Per-Capita 2020	-0.000***	173.163***
Observations	27,359	27,359
R2	1.000	0.983
Adjusted R2	1.000	0.983
Residual Std. Error	4.448 (df=27350)	50.329 (df=27350)
F Statistic	183593571.882*** (df=8; 27350)	202136.602*** (df=8; 27350)

10,000 for each state. The model is presented as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8$$

In the given model, X_1 represents the population of African-American/Black individuals in each state, X_2 represents the population of Asian/Pacific Islander individuals in each state, X_3 represents the population of European-American/White individuals in each state, X_4 represents the population of Hispanic/Latino individuals in each state, X_5 represents the population of Native American/Alaskan individuals in each state, and X_6 represents the population of other unspecified race/ethnicity groups in each state. X_7 represents the male population in each state, while X_8 represents the average per-capita income in each state for 2020. The dependent variable Y in Model 2 represents the population of individuals who died due to gun violence in each state's fatal counters, while in Model 3, it represents the population of individuals who died due to non-gun-related causes in each state.

Model 2 has an adjusted R-squared value of 1.000, indicating that the independent variables explain virtually all the variation in the number of deaths caused by gunshots across states.

Model 3 has an adjusted R-squared value of 0.983, suggesting that the independent variables explain approximately 98.3% of the variation in the number of deaths caused by other factors across states.

In both models, the coefficients for African-American/Black, Asian/Pacific Islander, European-American/White, Hispanic/Latino, Male, Native American/Alaskan, and Other are statistically significant at the 1% level. The coefficients for Per-Capita Income in 2020 are also statistically significant at the 1% level in both models.

In Model 2, the coefficients for African-American/Black, Hispanic/Latino, Male, and Other populations are positive, indicating that an increase in these population groups is associated with increased gunshot-related deaths. Conversely, the coefficients for Asian/Pacific Islander, European-American/White, and Native American/Alaskan populations are negative, suggesting that an increase in these population groups is associated with decreased gunshot-related deaths.

In Model 3, all population groups, except Native American/Alaskan and Other, have positive coefficients, indicating that an increase in these population groups is associated with increased deaths caused by other factors. The coefficients for Native American/Alaskan and Other populations are negative, suggesting that an increase in these population groups is associated with decreased deaths caused by other factors.

The coefficient for Per-Capita Income in 2020 is negative and statistically significant in both models, indicating that higher per-capita income is associated with fewer deaths caused by gunshots and other factors.

Model Evaluation

In this study, we sought to investigate the relationship between the cause of death during fatal encounters and the factors of gender and race. The dependent variable, Cause of Death, is a binary variable where 1 represents death by gunshot and 0 indicates death

Table 7: Model 1 Performance Evaluation (Confusion Matrix)

Actual / Predicted	Predicted 0	Predicted 1
Actual 0	452	1884
Actual 1	312	5560

by other means. Our independent variables are Gender and Race, with Gender consisting of three levels: Female (base level), Male, and Transgender; and Race comprising seven levels: European-American/White (base level), African-American/Black, Hispanic/Latino, Asian/Pacific Islander, Race unspecified, Middle Eastern, and Native American/Alaskan.

In the confusion matrix, the rows represent the actual class labels, and the columns represent the predicted class labels. The matrix can be interpreted as follows:

1. True Negative (TN): There are 452 instances where the model correctly predicted that the actual cause of death was not due to gunshots (i.e., it was due to other means).

2. False Positive (FP): There are 1,884 instances where the model falsely predicted the cause of death to be due to gunshots, while in reality, the cause was other means.

3. False Negative (FN): There are 312 instances where the model falsely predicted that the actual cause of death was not due to gunshots, while in reality, the cause was due to gunshots.

4. True Positive (TP): There are 5,560 instances where the model correctly predicted that the actual cause of death was due to gunshots.

Based on this confusion matrix, we can calculate various performance metrics for the logistic regression model:

1. Accuracy: $\frac{TN+TP}{TN+FP+FN+TP} = \frac{452+5560}{452+1884+312+5560} = 0.729$

2. Sensitivity/Recall: $\frac{TP}{TP+FN} = \frac{5560}{5560+312} = 0.947$

3. Specificity: $\frac{TN}{TN+FP} = \frac{452}{452+1884} = 0.193$

4. Precision: $\frac{TP}{TP+FP} = \frac{5560}{5560+1884} = 0.747$

The table demonstrates the model's performance in predicting the cause of death in fatal encounters based on the factors of gender and race. The model has a relatively high sensitivity (0.947), indicating that it correctly identifies a large proportion of actual gunshot-related deaths. However, the specificity is relatively low (0.193), suggesting the model has difficulty correctly identifying deaths caused by other means. The model's overall accuracy is 0.729, and the precision is 0.747.

Table 8: Model 2 and 3 Performance Evaluation

Model	MSE	RMSE	MAE
Model 2	2.063345e+06	1436.434796	777.677981
Model 3	1.756230e+08	13252.283386	8982.602359

The table presents three performance metrics for Models 2 and 3, designed to assess the relationship between the total number of deaths caused by gunshots or other factors in police fatal encounters across U.S. states and various independent variables. These independent variables include average per-capita income in 2020, race/ethnicity group populations divided by 10,000 for each state, and male population divided by 10,000 for each state.

The performance metrics displayed in the table are Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE):

For Model 2, the Mean Squared Error $2.063345e+06$, the average squared difference between the predicted and actual values for the number of deaths caused by gunshots. The Root Mean Squared Error for Model 2 is 1436.434796, the square root of MSE, providing a measure of the prediction error in the same unit as the dependent variable. The Mean Absolute Error for Model 2 is 777.677981, the average absolute difference between the predicted and actual values for the number of deaths caused by gunshots.

For Model 3, the Mean Squared Error is $1.756230e+08$, the average squared difference between the predicted and actual values for the number of deaths caused by other factors.

The Root Mean Squared Error for Model 6 is 13252.283386. The Mean Absolute Error for Model 6 is 8982.602359, the average absolute difference between the predicted and actual values for the number of deaths caused by other factors.

Comparing Models 2 and 3, we can observe that Model 2 has lower values for all three performance metrics (MSE, RMSE, and MAE) than Model 3. This suggests that Model 2, which predicts the number of deaths caused by gunshots, has better predictive accuracy than Model 3, which predicts the number of deaths caused by other factors. The lower values in Model 2 indicate that the model has fewer prediction errors on average compared to Model 3.

In the first logistic regression model, we observe that gender (Male), race (African-American/Black), and race (unspecified) are the three most statistically significant factors among all factors. Among these, males are more likely than females to encounter fatal shootings when police are present. Compared to the European-American/White racial group, African-American/Black and unspecified racial groups are less likely to experience fatal shootings. Furthermore, gender (Transgender) and race (Native American/Alaskan) are statistically significant factors. Transgender individuals are more likely than females to encounter fatal shootings when police are present, and Native American/Alaskan individuals are more likely to experience fatal shootings than European-American/White racial groups.

In the second linear regression model, with the number of fatal shootings encountered in each state in the United States under police presence as the dependent variable, the number of males in each state is the most influential independent variable regarding economic significance. This is followed by the number of Native American/Alaskan and Other race groups, and finally, the number of African-American/Black and Asian/Pacific Islander individuals. The numbers of European-American/White and Hispanic/Latino individuals are relatively less significant in magnitude. However, the degree of significance of these gender and race-related data is not significantly different when measured by statistical significance. Among these, African-American/Black, Hispanic/Latino, Male, and Other race group numbers positively correlate with the number of fatal shooting victims.

In the third linear regression model, with the number of non-shooting-related fatalities encountered in each state in the United States under police presence as the dependent variable, the numbers of Native American/Alaskan and Other race groups are the most influential in terms of economic significance. This is followed by Asian/Pacific Islander, European-American/White, African-American/Black, and Male individuals. Hispanic/Latino individuals have a relatively larger impact in this model compared to the previous one. However, their influence is less evident than the other variables in this model. However, statistically speaking, the degree of statistical significance of these independent variables is roughly similar. Among these, African-American/Black, Asian/Pacific Islander, European-American/White, Hispanic/Latino, and Male individuals correlate positively with the number of non-shooting-related fatalities.

Conclusion

Our analysis of the binary variable concerning whether a death in police presence is attributable to a gunshot, alongside the state where the incident occurred, the gender, and the victims' race, reveals substantial variations across categories. This suggests an association between each pair of indicators and the response variable. While certain states report higher death counts, adjusting for state population significantly alters the proportion of gunshot-related fatalities. Although gunshots account for most deaths, females are less likely to be killed by gunfire, and males exhibit a higher death rate than other genders. Furthermore, despite the notable scale of European-American/White fatalities during fatal encounters, adjusting for population sizes across racial and ethnic groups reveals that the white population is less likely to experience police violence.

Importantly, there is no discernible association between the three indicator variables—gender, race, and location of victims—suggesting that it is improbable that these indicators exert a collective influence on the response variable. This is advantageous for drawing causal inferences using statistical models.

In this research, we developed two fundamental regression models to investigate the factors contributing to gunshot-related fatalities. The first, a logistic regression model, examines the influence of gender and race on the likelihood of a victim's death caused by gunshot in police presence using a binary dependent variable. The second model, a linear regression, shifts focus from individual to state level, employing data on racial and gender distribution and per-capita income as independent variables while considering the total number of gunshot and non-gunshot fatalities as the dependent variable. Both models were briefly adapted into regression trees for elementary machine-learning predictions. Our findings reveal the presence of gender and racial disparities in such violent incidents, though regional differences were less pronounced, deviating from the patterns observed in geographic representations. This incongruity may be derived from the diverse gender and population compositions across states, influencing the displayed regional distributions. Notably, income disparities demonstrated low significance in both models.

The map reveals that most cases are dominated by Black individuals when calculating the mortality rates for different racial groups by combining the number of gunshot deaths with the total population of each racial group in each state. However, the regression model indicates that Black individuals are, in fact, less likely to die from gunshots than White individuals. This discrepancy may be due to the model's inability to consider the individual victim's death with the overall mortality rate for their racial group within the state through control variables. The significantly larger White population still influences the model in each state. In future research, we should consider incorporating mortality rates rather than the number of deaths to better account for this discrepancy and further enhance our understanding of the factors at play.

The Fatal Encounters dataset is continually evolving. We hope that, in the future, this dataset will encompass more societal sources, enabling us to examine the relationship between victims and deaths beyond incidents that have already occurred. As more defining traits of victims are incorporated into the data, this can also enhance our machine learning models, reducing the mean squared error and Gini impurity. Developing a theoretical framework to guide police training to reduce the frequency of deaths resulting from unprofessional actions will ultimately enhance the safety of both officers and citizens alike.

References

1. Broer, Markus., Yifan Bai, and Frank Fonseca. "A Review of the Literature on Socioeconomic Status and Educational Achievement." IEA Research for Education 5 (May 16, 2019): 7–17. https://doi.org/10.1007/978-3-030-11991-1_2.

2. Burghart, Brian. "Fatal Encounters: A Step toward Creating an Impartial, Comprehensive and Searchable National Database of People Killed during Interactions with Police." Fatal Encounters. University of Southern California, Eunice Kennedy Shriver National Institute of Child Health and Human Development, September 18, 2020. https://fatalencounters.org/.

3. Duffin, Erin. "Gender Distribution of Full-Time U.S. Law Enforcement Employees 2021." Statista, October 11, 2022. https://www.statista.com/statistics/195324/gender-distribution-of-full-time-law-enforcement-employees-in-the-us/.

4. Friesen, Corwyn. "IIU Concludes Investigation into Man's Death in Police Presence." mySteinbach, January 14, 2023. https://www.mysteinbach.ca/news/12054/iiu-concludes-investigation-into-mans-death-in-police-presence/.

5. Lang, Kevin., and Ariella Kahn-Lang Spitzer. 2020. "Race Discrimination: An Economic Perspective." Journal of Economic Perspectives, 34 (2): 68-89. DOI: 10.1257/jep.34.2.68.

6. Rahman, Imran U., Jian Deng, Junrong Liu, and Mohsin Shafi. "Socio-Economic Status, Resilience, and Vulnerability of Households under Covid-19: Case of Village-Level Data in Sichuan Province." PLOS ONE 16, no. 4 (April 29, 2021). https://doi.org/10.1371/journal.pone.0249270.

7. Steptoe, Andrew., and Paola Zaninotto. "Lower Socioeconomic Status and the Acceleration of Aging: An Outcome-Wide Analysis." Proceedings of the National Academy of

Sciences 117, no. 26 (2020): 14911–17. https://doi.org/10.1073/pnas.1915741117.

8. "U.S. Census Bureau Quickfacts: United States." U.S. Census Bureau QuickFacts: United States. Population Estimates, American Community Survey, Census of Population and Housing, Current Population Survey, Small Area Health Insurance Estimates, Small Area Income and Poverty Estimates, State and County Housing Unit Estimates, County Business Patterns, Nonemployer Statistics, Economic Census, Survey of Business Owners, Building Permits., July 1, 2021. https://www.census.gov/quickfacts/fact/table/US/PST045221.

9. "U.S. State Population by Rank (Update for 2023!)." Infoplease. Infoplease, February 13, 2023. https://www.infoplease.com/us/states/state-population-by-rank.

10. Wikipedia contributors. (2022, January 30). List of U.S. states and territories by income. In Wikipedia, The Free Encyclopedia. Retrieved March 22, 2023, from https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_income.

11. Wikipedia contributors. (2023, March 14). List of U.S. states and territories by race/ethnicity. In Wikipedia, The Free Encyclopedia. Retrieved March 22, 2023, from https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_race/ethnicity.