Police Brutality in Communities Across California

Dylan Mather, Layth Zubaidi, Nasser Mohieddin

How's My Police California Team

San Diego State University

San Diego, CA 92182

1.0 Abstract

Police brutality has always been a problem in the United States and has become a more controversial topic recently. There have been multiple instances of police going beyond the call of duty and killing individuals when the situation did not necessarily call for such actions. Regardless of how one would view the decisions to be made in many of these complex situations, we shall also consider the situation that police officials find themselves in with suspected offenders. While this has been a prominent issue throughout America's history, there has been greater awareness of it due to a series of police killings of many individuals, especially African Americans such as George Floyd. This has brought backlash from the public on a national scale, bringing mass protest and calls of accountability of the general police force. Action against the specific officers responsible for these incidents, and even calls for defunding many police departments. The goal for this project is to look into data records developed from police reports on the individuals killed in police custody. There shall be analysis through visualization, as well as the creation of Machine Learning Models to examine the full dataset by using certain parameters to develop predictors and find any accurate pattern that can provide more insight on the issue. The main questions on this topic are as follows: is there an actual rise in the amount of people killed by police over the years?, is there a connection between the death rates and the individual's socio-economic background?, is there a specific part in the process when bringing suspected individuals into custody that appear to be problematic? As part of this study we will be looking into data recorded by the Department of Justice (DOJ) Criminal Justice Statistics Center (CJSC) to develop tools to track and monitor police brutality rates. This database collects information on deaths in custody (DIC) in California that are reported by law enforcement agencies in compliance with California Government Code (GC) section 12525.

This topic has come into line with the Sustainable Development Goals produced by the United Nations, as a way to provide resources and reporting for those who wish to assist in developing solutions for these goals. This topic falls under three of the seventeen goals of of the SDG's; that being goal 10 'Reduced Inequalities' which aims to reduce inequality among the different populations residing within each country, goal 11 'Sustainable Cities and Communities' which aims to make cities and human settlements inclusive, safe, resilient and sustainable, and lastly goal 16 'Peace, Justice, and Strong Institutions' which aims to build resilient infrastructure, promotes inclusive and sustainable industrialization and foster innovation.

2.0 Introduction

Police brutality has recently been a hot topic in the United States. Recent events have resulted in backlash and protest fueled by a sense that certain minority groups were particularly targeted and police brutality was generally on the rise. The issue of police brutality in the United States is not new; there have been many studies looking into this issue and ways to solve this dilemma. A study on the" Risk of Police-Involved Death by Race/Ethnicity and Place, United States, 2012–2018" showed that " police were responsible for about 8% of all homicides with adult male victims between 2012 and 2018" (Edwards). However, few studies have examined police brutality with respect to disadvantaged communities beyond racial inequality. Furthermore, most existing studies are conducted at the country level. In this study, we look into incidents of police brutality in California with respect to various community demographics to identify current trends and connections, if any, between disadvantaged communities and higher levels of police brutality. We examine the impact of poverty, education levels, and race/ethnicity on the levels of police violence in a community. This will be presented in various forms of

visualizations that will help raise awareness and educate californians.The study will propose a possible solution for dealing with police brutality and restoring trust between the general public and law enforcement agencies.

We can take a brief look into the datasets used for this project, that being the CalEnviroScreen 4.0 Data provided by the California Office of Environmental Health Hazard Assessment and Death in Custody & Arrest-Related Deaths dataset from CA.gov . The CalEnviroScreen dataset is a recent dataset from 2021 that lists different health and environmental related issues recorded in different areas of California measured by the general population, separated by ethnicity and age. This is listed by the county level and uses coordinates to pinpoint areas within the county. Similarly, the Death in Custody dataset spans a time from 2015 to 2021 and also covers all of California on a county level. The CalEnviroScreen dataset holds about 8035 data entries for both the health and environment hazard data points as well as the demographic data set with a large amount of column attributes, about 58 columns. The Death in Custody data set has 12,488 entries listed, with 18 columns of attributes. We will be using these datasets side by side to create a more in depth analysis on the proposed issue.

3.0 Literature Review

This study focuses on examining police brutality in California by looking at death in custody as a general indicator of police excess use of force. The main question is has police brutality been on the rise in California in recent years, and how is that influenced by a community's demographics and other socioeconomic factors. Police brutality has long been a pressing issue in California and the United States as a whole. Thus, there have already been many studies conducted on the topic. As part of this project, we looked into studies conducted in

California and the country as a whole. Most studies tend to focus on rates of police violence with respect to ethnicity, gender, race, metal health, and homelessness. Studies have also focused on the impact of COVID-19 on police violence and arrests, as well as distress in the population.

Sirry Alang

We first looked at research involving police brutality in the United States. One of the first papers we looked at was "The More Things Change, the More Things Stay the Same: Race, Ethnicity, and Police Brutality" by Sirry Alang. In this review the author examines a study on police brutality with respect to race and location, and discusses how the findings of current inequality stems from historic and institutionalized racism (Alang 2018). The author makes the argument for a more upstream approach that relies on ending harmful policies and building a connection and trust between local communities and police.

Center of Juvenile and Criminal Justice

A Blog released by the Center on Juvenile and Criminal Justice, "Who Are Police Killing?", analyzes police killing in the U.S and looks into the demographics of those killed (Males 2014). Males states that California ranks sixth from the top when looking at states where a person is most likely to be killed by law enforcement (Males 2014). Native Americans were the racial group most likely to be killed by law enforcement, followed by African Americans, Latinos, Whites, and Asian Americans. Latinos are 30% more likely than average to be victims of police killings and 1.9 times more likely than non-Latino whites. 25% of people killed by police are under the age of 25. The author relied on data from the Centers for Disease Control

and Prevention (CDC), and concluded that police killings today appear to be much lower than in the past, especially when compared to the 1960s.

Frank Edwards

In their paper, "Risk of being killed by police use of force in the United States by age, race–ethnicity, and sex", Frank, Hedwig, and Michael analyze existing study and data on police brutality with respect to multiple demographics (Edwards 2019). They found that black men and Native American men were most at risk of being killed by police violence, followed by Hispanic and white men. They also noted that the risk of dying from police violence was highest in men and women aged 20 to 35, and that risk was significantly higher in men than in women.

Amnesty International USA

A report by Amnesty International USA, "Deadly Force: Police Use of Lethal Force in the United States", offers a comprehensive analysis of the use of lethal force by U.S. police (DEADLY 2015). Utilizing official statistics, media reports, and interviews with survivors, the report views police brutality as a systemic problem in the United States that disproportionately affects marginalized communities. It is concluded that there is an excessive level of police brutality in the U.S, and that African Americans, Native Americans, and other communities of color are disproportionately affected. It also indicates that there is a general lack of accountability and transparency due current laws and policies.

All five sources reviewed on the topic of police brutality in the United States reached similar conclusions with respect to the lack of data on police use of force, certain communities and minorities being at elevated risk of falling victim to police brutality, and lack of adequate process for reporting and accountability when killings occur. While most studies looked into

demographics behind police killings, none of the studies looked into indicators beyond race, age, and gender.

Further studies looked into other aspects of police brutality. In “POLICING THE POLICE: THE IMPACT OF "PATTERN-OR-PRACTICE" INVESTIGATIONS ON CRIME”, Tanya and Ronald provide an empirical research on the impact on crime and policing that federal and state “pattern of practice” investigations can have (Devi 2020). The authors concluded that when an investigation was initiated without being preceded by a viral incidence of deadly force use, led to significant reduction of homicides and total crime compared to investigations initiated in response to viral incidents. Those investigations resulted in an increase in homicides and total crime due to increase in police interaction. The paper also discusses how lower community trust and increased tension and aggressiveness can be influencing those rates.

Xenia Bion

In the study, “Racism Fuels Double Crisis: Police Violence and COVID-19 Disparities”, Xenia looks into how both police brutality and COVID-19 have disproportionately affected communities of color in the US. Discussing how each event has impacted communities of color, the author concludes that systemic racism and inequality in the criminal justice system and the healthcare system represent the source of the disparities observed (Bion 2022). In both cases, the disproportionate impact on those communities is stemming from structural racism. Thus, combating structural racism and reforming government institutions are key toward reducing disparity in both cases. Furthermore, the author examines psychological and physical effects on communities of color that experience or witness events of police violence. Communities that experience or witness incidents of violence may often struggle with a wide-spread sense of fear and anxiety, putting members of the community at a greater disadvantage.

Understanding how preceding events, such as “pattern of practice” investigations, impact levels of police brutality provides an important insight on how successful are current practices. Furthermore, results can be helpful to implement changes to how current incidents of police brutality are handled. In addition, it is interesting to see how other unrelated events, such as the impact of COVID-19 on disadvantaged communities, can be stemming from the same causes. This allows us to deal with the root causes of the problem. While police brutality is a general problem in the United States, our research has shown that certain disadvantaged communities are at greater risk. Xenia’s research shows how racism has been a root cause for this issue, and thus any solution should be aimed at the structural racism in U.S institutions (Bion).

Our study has also looked at existing studies on police brutality in California. In their report “Police Use of Force and Misconduct in California”, Deepak Premkumar, Alexandria Gumbs, Shannon McConville, and Renee Hsia offer an analysis of police misconduct and use of force in California (Premkumar. The report draws on data from a variety of sources, including official statistics, agency investigation orders, law enforcement agencies, and media reports. While incidents of police brutality in California are relatively low, certain communities, especially African Americans and Hispanic, are disproportionately affected. The authors concluded that police misconduct has been a significant problem, many of the complaints filed going uninvestigated. Thus, increasing transparency and accountability are key to addressing the problem of police brutality in California.

Public Policy Institute of California

The Public Policy Institute of California (PPIC) provides another article "Assessing the Impact of COVID-19 on Arrests in California" (Premkumar 2023). In this article, the authors analyze the COVID-19 pandemic’s impact on arrests in California using data from California

Department of Justice's Criminal Justice Statistics Center and explore the reasons for changes in patterns . The pandemic has resulted in a general decrease in arrests that can be attributed to decreased activity and changes caused by social distancing guidelines. However, the decrease was not equal across all offenses. Violent crimes saw a lower decrease, while domestic violence has actually increased slightly. Those insights can be linked to the impact of the COVID-19 pandemic on psychological and mental health, as well as, the events that followed the death of George Floyd.

Based on our literature review, we have concluded that while there have been numerous studies on the issue of police brutality. Most of these studies are limited in scope, focusing on demographics such as race and gender. There are even less studies focusing on police brutality in the state of California. Furthermore, no studies currently utilizes the deaths in custody data as a way of looking into police brutality. Thus, we see an opportunity to expand on existing studies by using the deaths in custody dataset from the California Department of Justice's Criminal Justice Statistics Center as a measure of police brutality, and analyzing results with respect to various demographics and socio-economic factors such as poverty, unemployment, linguistic isolation, and pollution. In Addition, our literature review revealed a limitation in available data as well as an inadequate process of reporting and lack of accountability. In this project we aim to create publicly available tools to track and monitor police brutality rates as a way of increasing social accountability.

4.0 Team Members and Responsibilities

Our team consists of three members who all have performed their own assigned task as well as helping one another for other tasks if needed. Before looking into our main project,

How's My Police California', we should mention that we had originally started work on a different project, Histopathologic Cancer Detection which utilized computer vision for detecting tumors. In the beginning we had initially used a Kaggle dataset that was used for a topic like this, but Layth and Nasser would proceed to look for other datasets similar to this one and approach it in a manner of cleaning the data and preparing it for our own work. Dylan would begin to look for proper machines that could run the massive size of any of these datasets as well as looking for a proper environment that could allow collaborative coding and save our work possibly to the respective cloud service. We had all begun looking into some options for the environment as there were many hurdles to get around. After some more attempts of working with a school machine that Dylan was able to acquire through extensive conversation with different individuals a part of the GIS department in SDSU, we had to construct an alternative plan and so Layth and Nasser would need to look for other datasets and a proper data analytics topic to possibly focus on.

We shall continue more in-depth to the roles of our final research project. As mentioned, while Dylan was continuing some final attempts to make the code work from the previous project, two more datasets of the alternative project's topic were formed. Finding raw data collected from the Department of Justice for California's Criminal Justice on the deaths of those in police custody as well as finding more datasets to accompany the current dataset. We have found another dataset that concerns socio-economic aspects of different geo-locations across California that can easily match with our current dataset due to sharing essentially the same county information. Of course in later sections we mention how we do this, but essentially this provides more information for the people affected by these killings, via their home environments and the level of different socio-economic factors. From there we began cleaning data and inserting geographic information

for the main dataset for usage by Nasser in constructing data visualizations, via mapping in ArcGIS and developing our main dashboard. Dylan and Layth began looking more into the data and writing different Python scripts in Google Colab Notebook. There was development of manipulating the data to find Death Rates for the dataset per 100,000 people per population of each county. Creating additional datasets of this new information for Nasser to use in more important geographic visualizations. Then going into the main chunk of script for manipulating the data to make it appropriate for different machine learning models. Dylan had a heavy hand in applying the manipulated data into specific machine learning algorithms that were appropriate for this type of dataset we have. This also included both coders manipulating the columns of the additional dataset including the socio-economic details that we decided provided the most important information for describing the environment many of these people could be surrounded by. From there Nasser began creating a very informative website, presentation slides, and began on the video. We all came together to record our work for the video and insert different information we all came across to insert in our website.

5.0 Data and Methods

Overview of the data

We can first look into our data source which was found in the State of California Department of Justice database. They have their own website under the title of OpenJustice that provides multiple different datasets. Our dataset is titled ‘Death in Custody & Arrest-Related Deaths’. The data revolves on both state and local law enforcement agencies, as well as correctional facilities that report information on deaths that occur in custody or during the process of arrest. We must keep in mind the source of the data, despite the laws and codes that

are made for law enforcement responsibility. In particular to our case Section 12525 of the California Government Code is what these agencies base their responsibilities for these reports and are also subject to revision as reports are received by the California Department of Justice. This dataset contains information on deaths related to police custody in California from 2005 and 2021. Each row represents a death and includes basic demographic information (age, race, gender) as well as some other features including: location, custodial responsibility, manner and means of death, custody status and the offense the criminal committed. The first step we took was to deal with the NaN values. Thankfully this data set only had 3 NA values which represented a small portion of the 12k+ dataset. Two rows didn't have the "race" value filled in so we classified this as "other". The other NaN was for "facility_death_occured" so we matched that with the "custodial_responsibilty_at_time_of_death". A table of all parameters provided in the dataset's file will be provided in the appendix for better understanding of what each column entails. It should be noted that two new columns were created; the 'Longitude' and 'Latitude' attributes. They were created, inserted, and saved into the file from a section of our code that was developed for the purpose of creating geographical based visualizations of our data, such as heat maps and geo-based cluster points. They would also combine well with the other additional dataset with social and environmental levels of the population together to produce greater insight and analysis. The code was implemented to create two new and empty columns and then require the user to insert longitude and latitude points based off of the specific agency that the data point belonged to. From there it would update and save the dataset's file. These are the first two assumptions we made for this dataset.

We first included a dataset that has information of various statistics of each county in California in the year 2021. This is called the CalEnviroScreen 4.0 Data provided by the

California Office of Environmental Health Hazard Assessment as we mentioned before, this dataset includes different environmental and social descriptors of different populations located in different latitude and longitude points. While a lot of this information is very useful, we focused on a small handful of data points that can be reasonably be related to aspects of life that may affect an individual to have a run in with law enforcement that could possibly lead to any one of these deaths associated in our main data set. We chose the columns that we surmised would be the most correlated with the death data. These columns and their definitions being the 'Ozone Percentile' (Amount of daily maximum 8 hour Ozone concentration), 'Toxic Air Release' (Toxicity-weighted concentrations of modeled chemical releases to air from facility emissions and off-site incineration from RSEI), 'Solid Waste' (Sum of weighted solid waste sites and facilities (SWIS) within buffered distances to populated blocks of census tracts), 'Pollution Burden' (Average of percentiles from the Pollution Burden indicators with a half weighting for the Environmental Effects indicators), 'Education Score' (Percent of population over 25 with less than a high school education), 'Linguistic Isolation' (Percent limited English speaking households), 'Poverty' (Percent of population living below two times the federal poverty level), 'Unemployment' (Percent of the population over the age of 16 that is unemployed and eligible for the labor force), and 'Housing Burden' (Percent housing burdened low income households). We have also included some extra columns from the demographic aspect of the Environmental Screen data set that are measured by the 2019 ACS population estimates in census tracts. Our second assumption for this data is that these statistics were similar in 2021 as they were for the 6 previous years. In order to merge the two datasets, we had to average the metrics of all the zipcodes in the same county to be able to match up to the deaths in custody. It should be mentioned that demographic and screen scores are separated from each other and had to be

imported individually and then merged, which isn't a huge issue as they both share the same ID numbers when numerically ordered. Now when we want to merge the death in custody and California environmental screen together we had to keep in mind that the custody dataset is recorded on an individual basis, as each person is a data point while the other dataset is recorded by a single geolocation point of the county they are in which can include any number of individuals within that geopoint. So we decided to average the values for each of the counties within the calEnviornScreen data set to get them into a format where we can include them into our DeathCustody dataset that we will be using for our machine learning models later on.

Preprocessing

Now that we combined both our datasets together, it was time to begin the preprocessing of the data to prepare it for the machine learning models. The first set is to choose which columns we want to be our target variables. These would be columns that we can predict either using classification or regression methods, and would provide deeper insight on what we could glean from our data. We narrowed down our selection to race, age, and reporting agency.

We chose to attempt to predict race since it would show us which features showed the highest predictability so we could identify any discrepancies in how people were treated in the justice system based on their ethnic background. This would shed some light on areas police departments could focus their efforts to help fight racial injustices in their communities. We also wanted to try to predict the age of the deceased to see if there were any other trends that came up based on how old the person was when they died. Finally we chose to try to predict the reporting agency to show if there were similarities in the deaths compared to which agency they were arrested by. We are trying to make this prediction to help us find out what aspects of each of

these agencies make them different from the other and if there is anything that one agency could improve in their arresting process to help reduce instances of police brutality.

After merging our Death in Custody and Enviroscreen datasets, we had a total of 42 features and 12,329 rows. We merged these datasets together on the county variable. Since the Enviroscreen dataset had metrics listed on the zipcode level, we had to average these columns to the county level for the merging process. After this averaging process these columns were all in continuous data format so they needed no further initial preprocessing to be implemented in our models. Since we have preselected which columns form the enviroscreen, we also need to narrow down which columns we want to use from the Death in Custody dataset. We decided to remove, “agency_full_name”, "agency_number", "California County", and 'location_where_cause_of_death_occurred', since these columns were either redundant or did not provide any relevant information for predicting our chosen target variables.

There were many columns in our dataset that held nominal values. All of these features needed to be run through one hot encoding, so they could be parsed by our machine learning models. Nominal data, unlike ordinal data, cannot have each of their values simply replaced by numbers for preprocessing them for ML models since this would suggest to the models that have a ranking between them. Running one hot encoding makes a new column for each variable in a feature so there isn't this ranking misconception. However, making a new column for each of our features as they are would add way too many columns which can introduce high multicollinearity. Therefore, we needed to group some of the values of the features together so they did not have as many unique values, reducing the number of new columns one hot encoding would create. It's important to note that the grouping of values in these columns were only used for machine learning purposes and in our visualizations we kept the original categories.

The first columns where we grouped the variables were "facility_death_occurred" and “custodial_responsiblity_at_time_of_death” . These columns give the name of where the person died and who had custody of them when they died respectively. Most of these people died in custody of California Detention Centers or California Rehabilitation Centers (CDC/CRC) which is synonymous with prison. However more than half of the deaths occurred in local hospitals, most likely because these prisons didn't have their own medical area to treat inmates. These columns also had values that only accounted for less than 10 values each, so we grouped them into the larger categories. Our final categories for both columns were CDC/CRC, Jail, Crime Scene, Hospital, and Other. These two columns had the same categories so we also wanted to check how many of these matched, so we made a new feature to check this and see if its predictive.

The next column we processed was the race column. In the original dataset there were 18 unique races included in the dataset. In order to reduce the unique variables and to create a closer balance between the race categories we grouped some of these together as well. After aggregation we ended up with these four race categories for our machine learning purposes: White, Hispanic, Black, and Other. We did a similar aggregation for the reporting agency column, ending up with Local Police (Police), State Police (State) and Sheriff. Since these are both also target variables, this grouping decreased the complexity of our models to give us a higher likelihood of good performance.

The last four columns we did our grouping were custody status, custody offense, manner of death, and means of death. For custody status and custody offense we simply grouped them into sentenced vs not sentenced and felony vs other respectively in order to try get the best split between the data, avoiding unbalanced columns. For means of death, we grouped these into:

weapons of all manner, hanging/strangulation, drug overdose, other and not applicable. For the means, the final groups were natural, suicide, accidental, homicide justified and willful and other. These were the last columns that needed to be preprocessed for the one hot encoding. After this last step we had 58 columns.

Machine Learning

Now that all of our data has been prepared for our machine learning models, we are ready to take a deeper look into what our data can show us. Our first visualization, in figure 1 below, we create is the correlation matrix of all the features to see the relationship between our different columns. At a glance there are two sections of the data that show a high multicollinearity. These sections are the Calenviroscreen that we merged in by the county, and all the one hot encoded variables. Both of these occurrences of high multicollinearity are justified. The first section of the Calenviroscreen has such a high collinearity since the data was averaged for the counties and since the data was only from 2021 and we assumed that the metrics would not change much over the course of the 6 year span of the death in custody dataset. One hot encoded section showing high multicollinearity also is justified since that is a byproduct of splitting these columns up using this method. Moving forward we need to either drop some of these columns and or choose ML models that perform well even with highly collinear data.

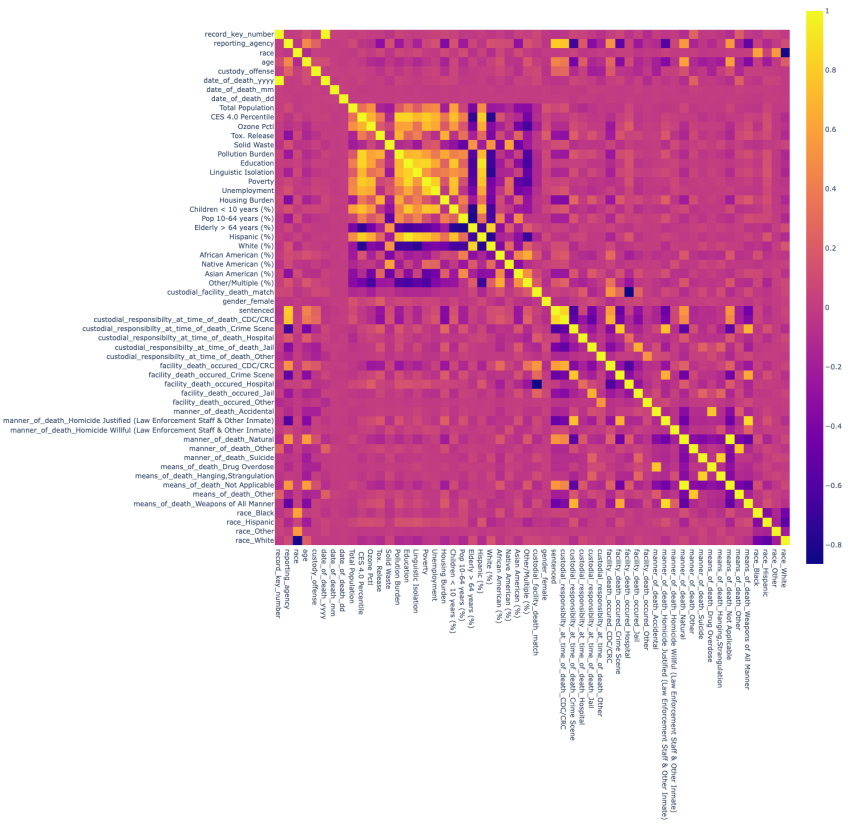


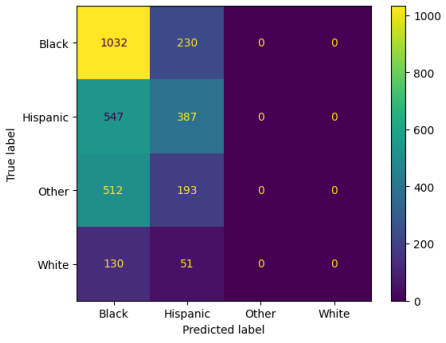
Figure 1. Correlation heatmap of machine learning data

We had three features that we chose to be our target variables for prediction using machine learning. The first of which we tried to predict was age, our only discrete data type. This means we need to use regression models to try to predict our age values. The four models we used were Linear Regression, Decision Tree Regression, Random Forest Regression, and Support Vector Regression. None of these models had great scores and decision trees had the worst performance shown in Table 1. This is likely due to the multicollinearity of the data. The mean squared error of around 146 for our best performing model of SVR shows that on average, the models were more than 12 years off of the actual age of the person. To improve the scores of this model we would need to remove the colinearity of the data through more feature engineering and perhaps more relevant data.

	Name	Mean Squared Error	R-Squared Score	CV Score - MSE	CV Score R-Squared
0	Random Forest Regressor	168.395499	0.379792	-164.912765	0.297939
1	Linear Regressor	146.251781	0.461348	-150.279640	0.359237
2	Decision Tree Regressor	289.538611	-0.066384	-313.171755	-0.335544
3	Support Vector Regression	146.251781	0.461348	-148.036536	0.367906

Table 1. Dataframe of different machine learning models outputs

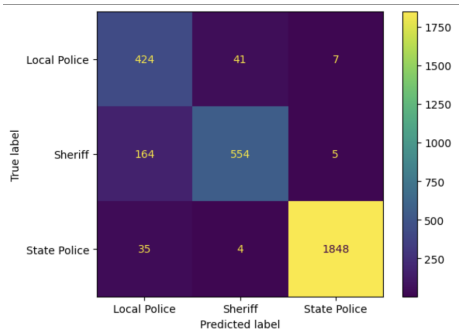
We also did two multiclass classification predictions for the Race and the Agency features. The classification models we used for these predictions in order of increasing performance were: K-Nearest Neighbors, Gradient Boosted Classifier, Naive Bayes Classifier, Support Vector Classifier, and the best performer, Random Forest Classifier. The multicollinearity of the data did have a great impact on predicting the race column. This was shown by the high dependence on the training dataset sampling causing a wide variation in which race was predicted the most. The RFC did the best with the multicollinearity since it uses subset selection methods to choose the most predictive columns for its overall prediction. For most of our runs, as shown on the confusion matrix, black and hispanic were the most predicted race since they made up most of the population normalized values. This led to an accuracy score of 46% which is about twice as good as random guessing. The models seemed to choose two of the four races and did a good job at predicting the right one between those two. Again these metrics could be improved with more relevant data on the deaths.



	precision	recall	f1-score	support
Black	0.46	0.82	0.59	1262
Hispanic	0.45	0.41	0.43	934
Other	0.00	0.00	0.00	705
White	0.00	0.00	0.00	181
accuracy			0.46	3082
macro avg	0.23	0.31	0.26	3082
weighted avg	0.33	0.46	0.37	3082

Table 2. Confusion Matrix of race prediction metrics for the Random Forest Model

Agency prediction on the other hand had a 92% accuracy. Our most predictive column for this prediction was who had custodial responsibility at the time of death. Most of the State police arrests had inmates who died in the custody of the CDC/CRC (Prisons) so this is likely what led to the higher prediction accuracy. All of our models except KNN had over 90% prediction accuracy. KNN does not perform well with high dimensional datasets especially when there is a high rate of multicollinearity. The Random forest did well again here because of its ability to retain high performance with these types of datasets. With mixed results in our machine learning models, we move onto visualizing our data and analyzing what they showed us.



	precision	recall	f1-score	support
Local Police	0.69	0.84	0.76	447
Sheriff	0.90	0.83	0.86	766
State Police	0.99	0.97	0.98	1869
accuracy			0.92	3082
macro avg	0.86	0.88	0.87	3082
weighted avg	0.92	0.92	0.92	3082

Table 3. Confusion Matrix of agency prediction metrics for the Random Forest Model

6.0 Results and Discussion

Python Constructed Graphs

After much overview of our dataset and conducting different approaches of analyzing and processing the data we have begun to draw some conclusions to our topic. From here we can start to see our work through different tools of visualization and analysis has developed different statistical charts that break down our dataset to its unique attributes.

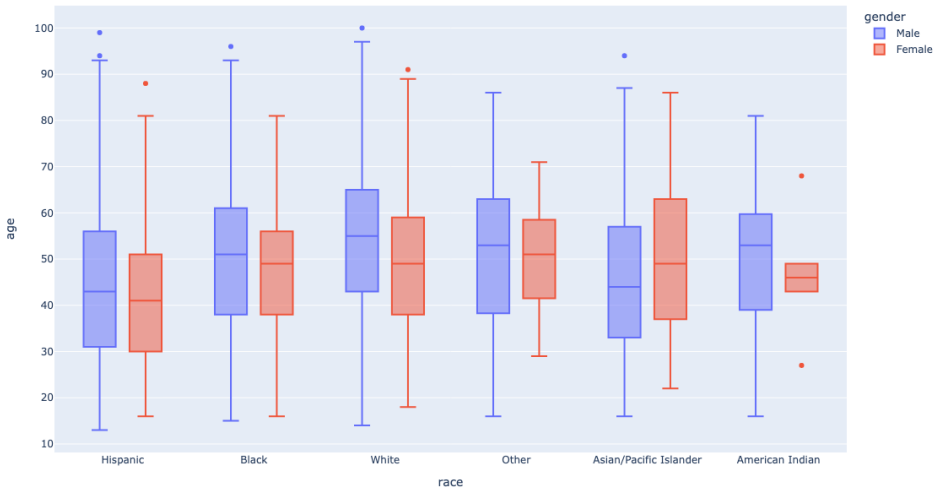


Figure 2. Multi-Variable Box Plot consisting of Age, Gender, and Race

Figure 2 shows the distribution of age for individuals in different racial groups (White, Black, Hispanic, Asian/Pacific Islander, American Indian, and Other), with each group represented by a boxplot. The gender of the individuals is represented by different colors of the boxes. The position of the box on the y-axis represents the median value of the age distribution, while the height of the box represents the interquartile range (IQR) of the data. The whiskers extending from the box indicate the range of the data, and any points beyond the whiskers are considered outliers. The resulting plot is meant to provide an informative visual representation of the distribution of age for individuals of different race and gender in the dataset, allowing for easy comparison between the groups. It can provide insights into any differences or similarities in the age distribution across different racial groups and genders. From here we can see half of these instances feature outliers of individuals who have died in custody, as well as almost every box plot having a median range between 40 and 55 years of age. The typical majority of point clusters for deaths by age are 35 to 60 years of age, which can be expected as older adults make up a larger count of general arrests than those outside the range of much younger and older.

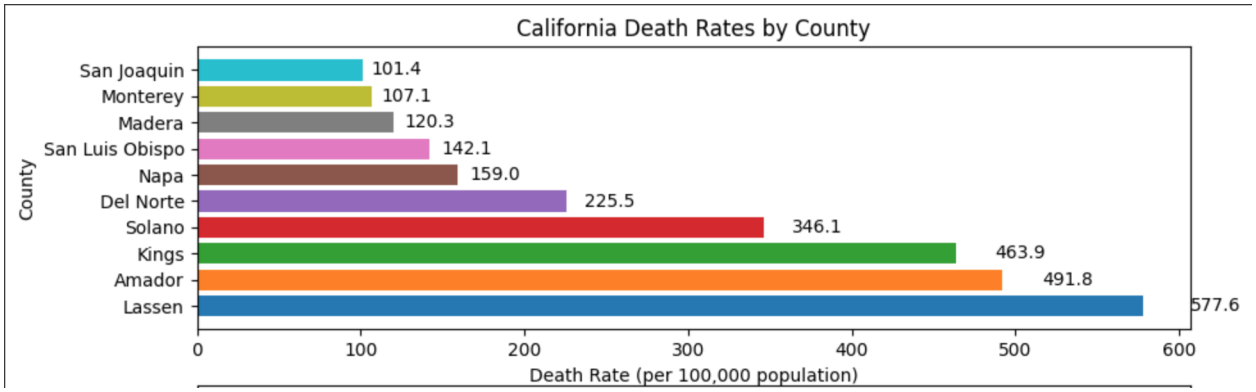


Figure 3. Bar Graph of Death Rates by County

The bar graph in figure 3 represents Death Rates in different counties. We see the Death Rates of every County. The counties are sorted by their death rate, and each bar represents a county, with the height of the bar indicating the death rate. The plot is divided into multiple subplots, with each subplot containing a group of 10 counties, here we took the top 10 of the list. The x-axis of each subplot represents the death rate (per 100,000 population), and the y-axis shows the name of the county. The highest death rate is at the top of the plot, and the lowest death rate is at the bottom, while also doing the reverse in each individual box of 10 Counties. The values of death rates are also displayed next to each bar. This plot is useful for identifying the counties with the highest and lowest death rates and comparing the death rates across different counties.

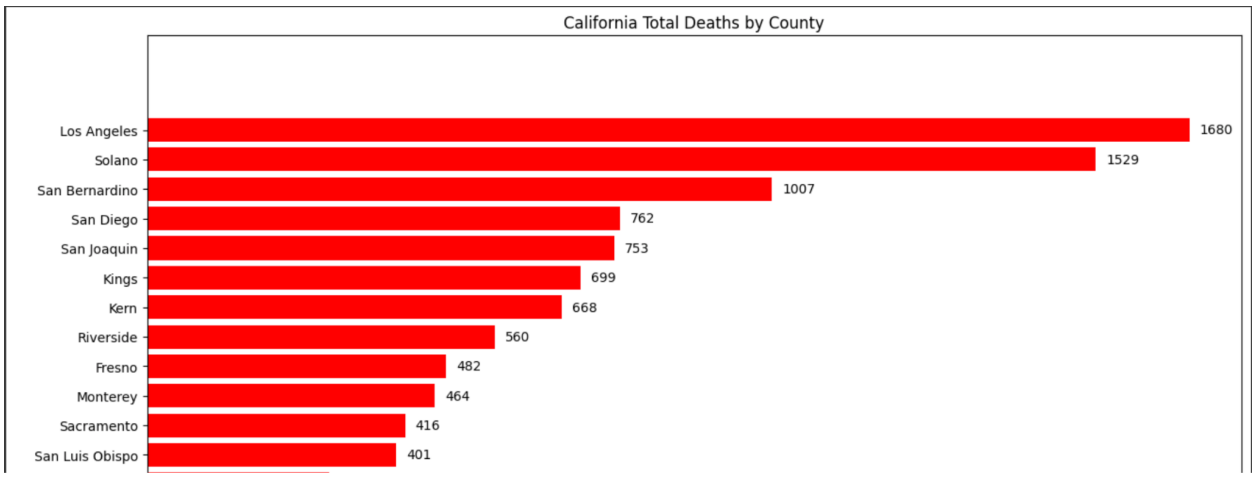


Figure 4. Bar Graph of Total Deaths by County

In figure 4 we see the total deaths that each county had. This chart is sorted by the total number of deaths in each county, and it shows the county name on the y-axis and the total

number of deaths on the x-axis. For this case we have posted the top 12 Counties with the total most deaths. Each bar represents the total number of deaths in a particular county. Text labels on top of each bar show the exact value of the total number of deaths in each county. The plot is informative and provides an easy-to-read summary of the total number of deaths in each county in California. It allows for easy comparison between the counties and provides insights into which counties have higher or lower total numbers of deaths. When comparing these two recent bar graphs together we can see in some instances that while certain Counties may have a larger number of deaths, in comparison to their Total Population doesn't produce a high death rate, like Los Angeles for example.

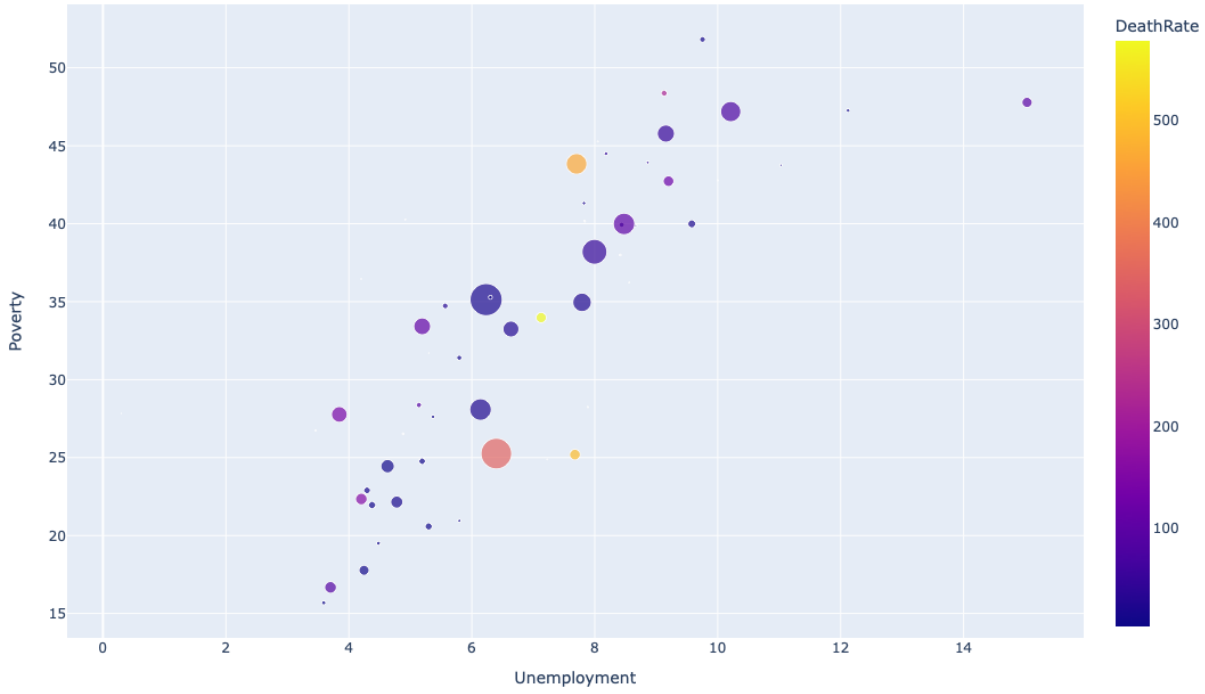


Figure 5. Scatter Plot of Socio-Economic attributes and Death Rates

Figure 5 shows the relationship between the Unemployment rate and Poverty rate of different counties, with the Death Rate represented by different colors and the Deaths represented by the

size of the markers. Additional information appears when the user hovers over a data point on the plot, this includes the Toxic Air Release, Linguistic Isolation, Education, and California County variables. The resulting plot is meant to provide an informative visual representation of the relationship between these different variables and how they relate to the Death Rate and total Deaths.

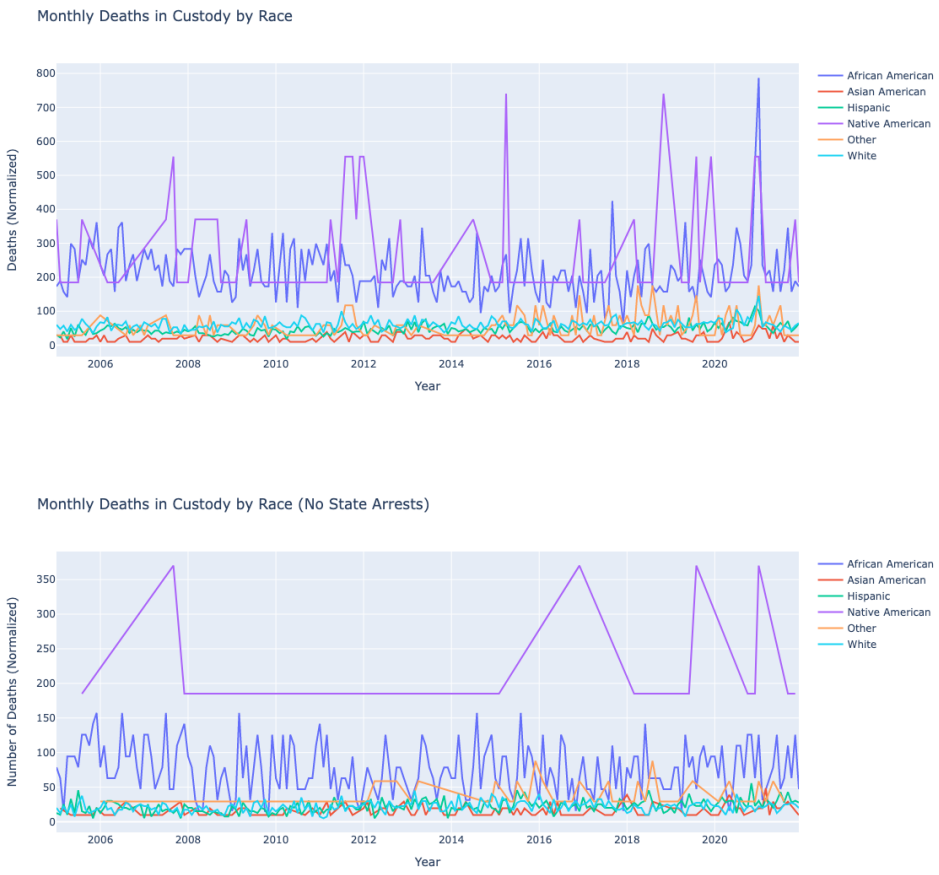


Figure 6. Time Series Graph of Monthly Deaths by Race (With and Without State Arrests)

The two graphs depicted in figure 6 above focus on deaths of different races in Police Custody through a monthly time scale. These two graphs below are comparable as the second graph is the same as the first graph but doesn't include State Arrests. When looking at the results produced by the graph we can see that when taking out State Arrests, the large spike that you

would've originally seen on the top graph for the year 2020 doesn't appear anymore. This can give an insight of the conditions for prisoners facing the Covid Pandemic and resulting in a large spike of deaths when located in State Correctional Facility. As stated by the Bureau of Justice Statistics "From March 2020 to February 2021, nearly 2,500 incarcerated people in state and federal prisons died of COVID-19, ... This translates to a death rate[2] of about 1.5 deaths per 1,000 incarcerated people from the virus" (USAFacts).

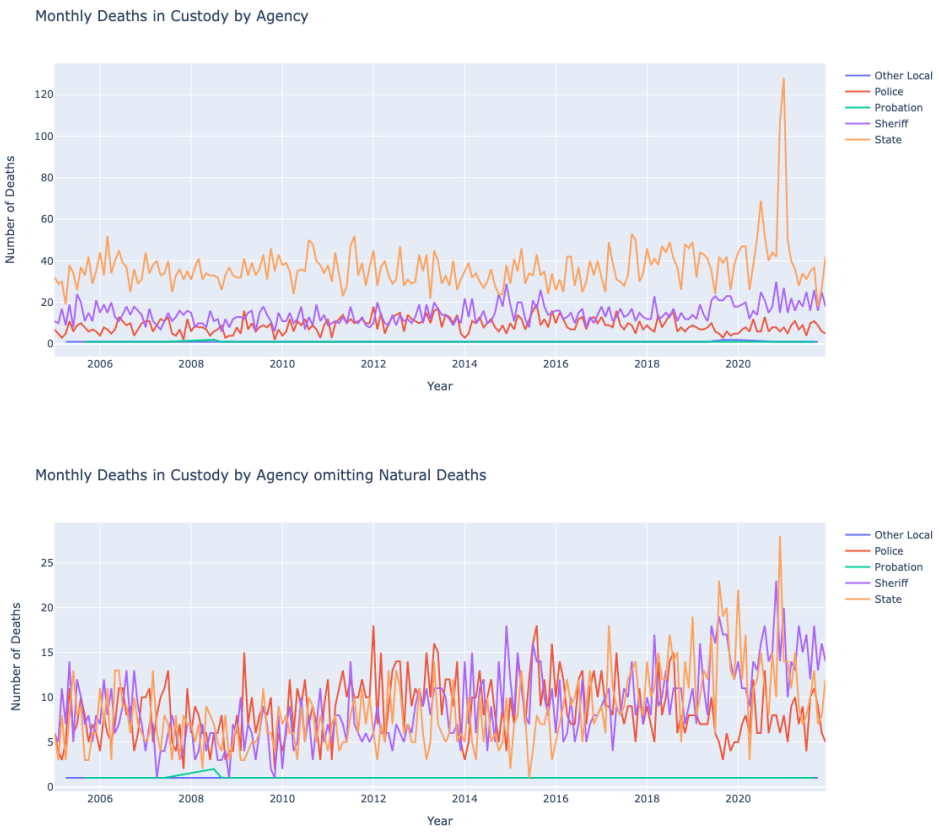


Figure 7. Time Series Graph of Monthly Deaths by Agency (With and Without Natural Deaths)

Here, we have also taken a look at the number of Deaths for each of the different Agencies that made arrests. The three main agencies who made the most arrests were Local Police (Police), Sheriffs, and State Police (State). As clearly shown in figure 7, the State value had the most cases and was the only agency that had a spike of deaths by the end of 2020. When

comparing it to the fourth figure, which omits natural deaths, there is still a spike in deaths. These deaths are more likely to be related to police brutality.

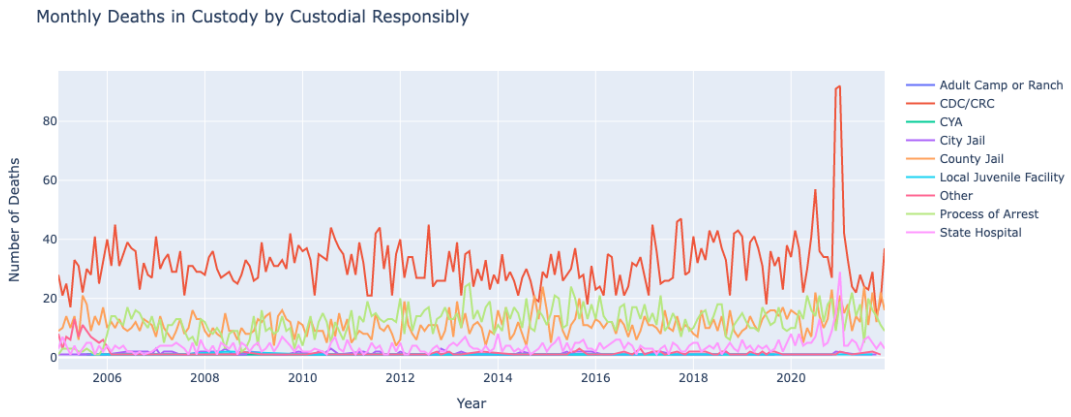


Figure 8. Time Series Graph of Monthly Deaths by Custodial Responsibility

This time series graph in figure 8 shows deaths sorted by who had custody of the deceased at the time of death. This dataset mostly contained cases where the deaths occurred at CDC/CRC. These stand for California Detention Center and California Rehabilitation center. This is the same as state prisons. This graph indicates that inmates who were housed here had worse living conditions than the other facilities listed in this dataset. This brings some insight of certain custodial locations that tend to have a negative view or interaction of those in their custody. This can bring more insight to these issues and can shine light in areas of authorial holding that neglect those they are responsible for. This can also include the living conditions of these individuals as around these times, there was the spread of Covid-19 which could've led to higher death rates due to a lack of safety precautions for those in holding.

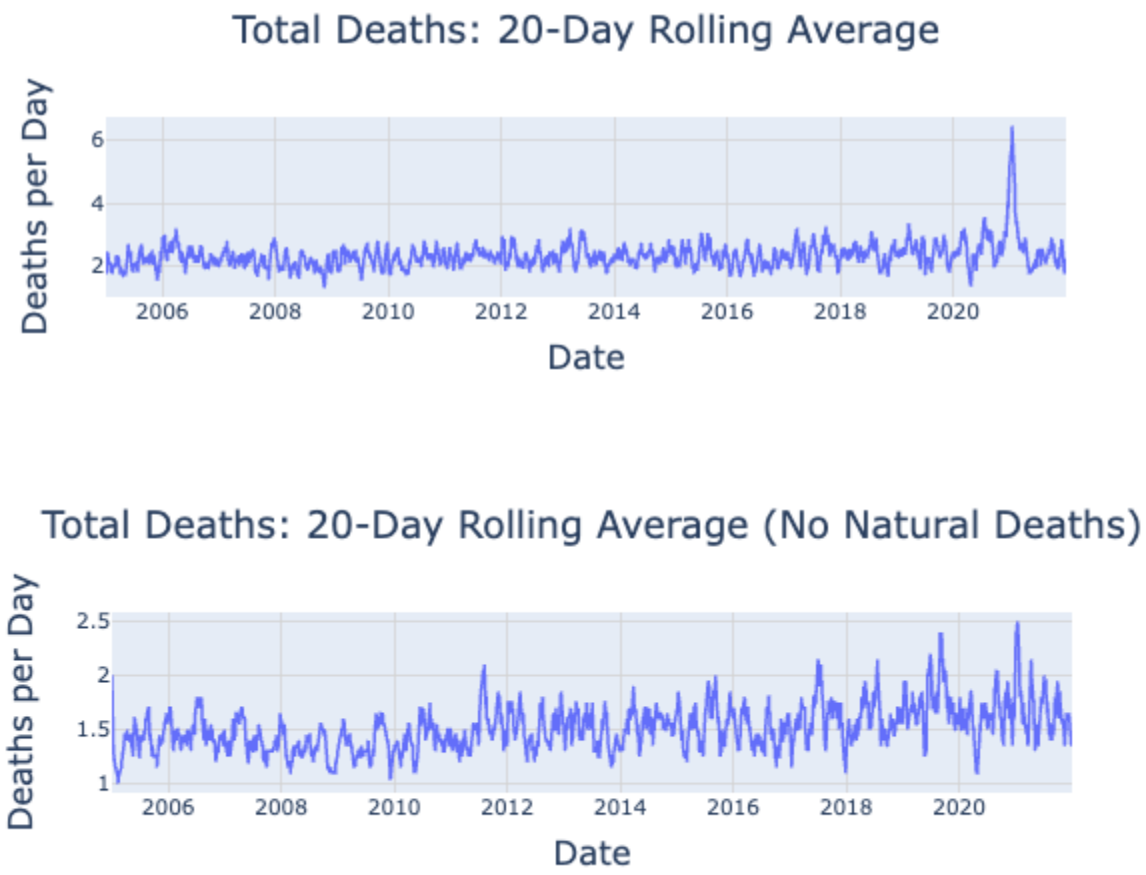


Figure 9. Time Series Graphs of the 20 Day Rolling Average of Deaths With and Without Natural Deaths

Our last time series shown in figure 9 shows the rolling average of deaths per day over the 6 year span of our data. Adding this rolling average and comparing the non-natural deaths to the total deaths. It again is clear to see how much the pandemic had an effect on the health of prisoners. The number of deaths per day more than tripled at its peak, due to the lack of safety precautions in the holding. While looking at the rolling average without natural deaths taken into consideration, there is a clear uptrend in the amount of deaths happening per day and the rolling average makes that fact easier to see. These non-natural deaths are much more likely to be linked to the effects of police brutality, whether from direct violence or the psychological effect, since the third most common manner of the recorded deaths were suicides.

ArcGIS Mapping and Visualizations

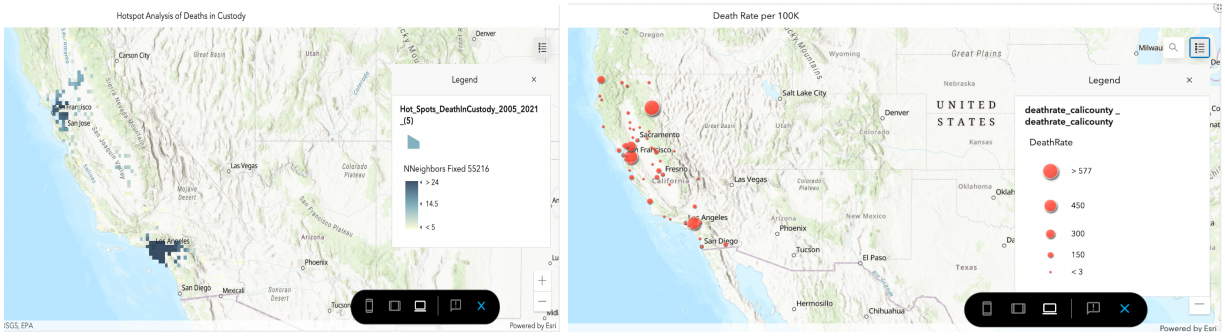


Figure 10. Hotspot Analysis & Clustering by Death Rate

The two main maps on the dashboard shown in figure 10, reflect the result of a hotspot analysis that was created through ArcGIS reflecting the hotspots based on the number of deaths per county. Based on that analysis, we can see that Los Angeles County is the hotspot where most deaths in custody in California tend to occur. This is to be expected as the county is the most populous in the state. The second map represents a closer examination of the data, showing clusters of death rates in each county per 100K of residents. While all the major clusters continued to appear in counties with the highest number of residents and largest urban centers, it was interesting to observe that Lassen, CA. Looking into the matter, it was observed that Lassen, CA is actually known as a prison town with more than 9 jails and prisons in the County, including California Correctional Center (Ccc) (Jails & Prisons).

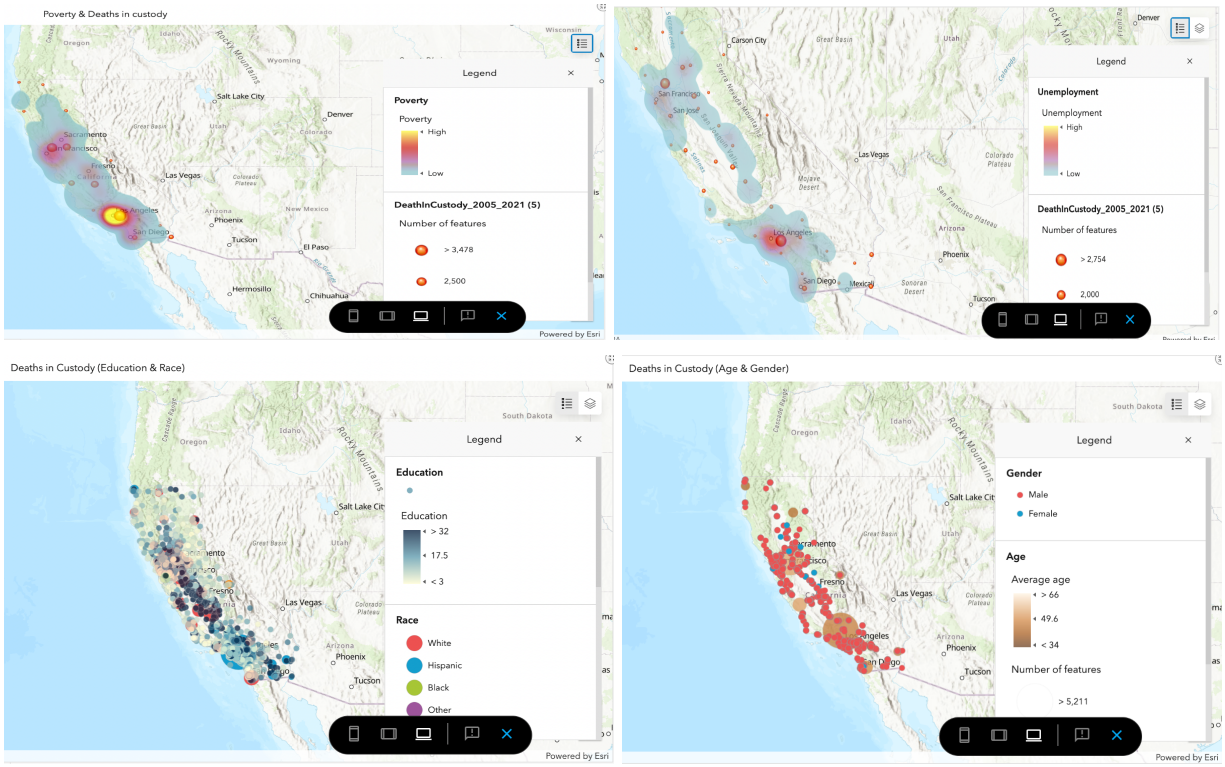


Figure 11. Visualizations with multiple socio-economic layers

Furthermore, the maps in figure 11 show the different socio-economic layers that can be applied on top of a cluster of deaths in custody across California. The maps provide a look into how certain socio-economic factors that disadvantaged communities usually experience tend to overlap with greater numbers of deaths in custody. One of the first maps created shows how a heat map of poverty in the state of california by county overlaps with death in custody per county. The map shows how areas with higher rates of poverty tend to have larger clusters of deaths in custody. Another layer examined Age and Gender, which shows the average age per cluster of death and the distribution of death, which was overwhelmingly males. We also looked into education and racial breakdown as other demographics that can be influencing rates of deaths. In the map, we see racial breakdown of death by county, in addition to levels of education. No correlation was observed between education rates in a county and deaths in custody in that county. We have also provided users with the ability to visualize additional layers

with respect to deaths in custody, including unemployment, linguistic isolation, housing burden, and pollution burden. Overall, it was observed from the maps that socio-economic factors experienced by disadvantaged communities correlates with higher rates of deaths in custody, which can indicate that those disadvantaged communities are at greater risk of falling victim to police brutality.

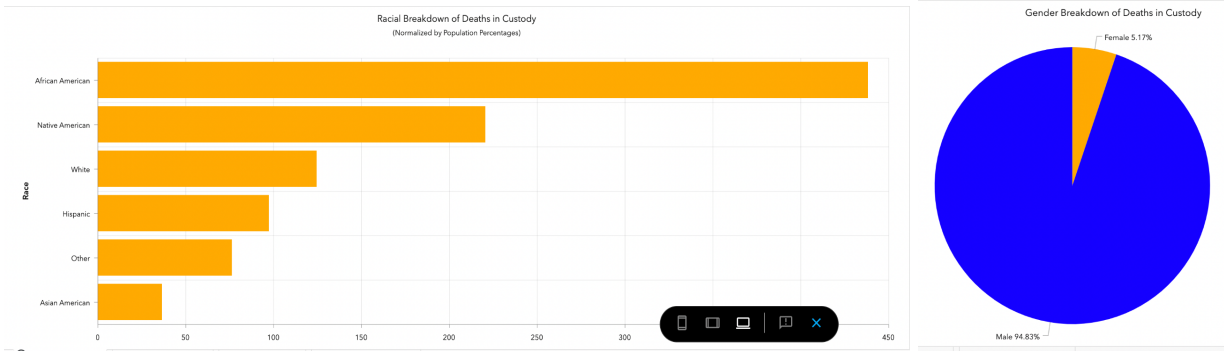


Figure 12. Deaths Breakdown by Race and Gender

The dashboard also provides supplemental figures showing important findings based on our study. Figure 12 shows two of the figures on the dashboard representing the racial and gender breakdown of the deaths in custody. A look into the racial/ethnic breakdown, normalized by population percentage in the state, shows that African Americans were more than three times as likely to experience death in custody. This is inline with previous studies that found that African Americans were three times more likely to fall victim to police brutality. The elevated risk for Native Americans was also inline with previous studies. Our result reflects how the deaths in custody datasets represent a high correlation with police brutality. Furthermore, the pie-chart of gender breakdown, shows that males are overwhelmingly impacted by the issue of police brutality.

7.0 SWOT Analysis

A S.W.O.T analysis was performed for this project based on our literature review, primary research , data analysis, and opinion. The S.W.O.T analysis is important to identify the strengths, weaknesses, opportunities, and threats associated with our project concerning police brutality in California and its impact on disadvantaged communities. This was a necessary step to better understand our project and capitalize its strengths.

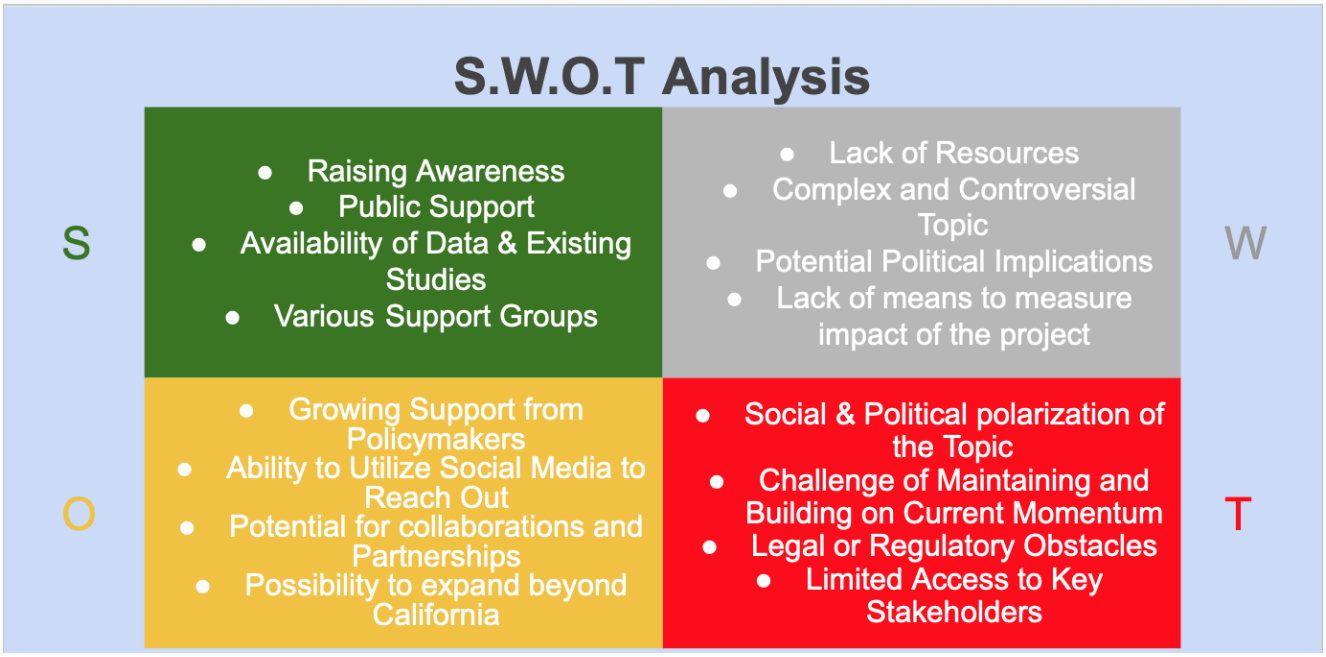


Figure 13. SWOT Analysis of the whole research project

Based on the S.W.O.T analysis, we identified multiple strengths including strong public awareness on an important issue that has an impact on the general public. In addition, given recent events, there has been increased public support for reforms to current policing policies and institutions. While data on this topic has been limited, there are still some publicly available datasets and existing research that can facilitate this study. Finally, since this has been a long

existing issue in the United States, there are various support groups that advocate for dealing with the issue of police brutality and provide support to the victims.

Some weaknesses of the project include lack of resources. Limited resources can make it difficult for agencies to implement reforms. Furthermore, police use of force tends to be a complex and controversial topic that has been increasingly polarized by recent events. Thus, it has developed into a political topic, which can generate backlash and resistance to reforms. In Addition, due to limited access to data, we do not have the means to measure the impact of our project.

Threats include socially or politically polarizing the topic, losing the current momentum that have built up due to recent events. In Addition, there may be potential legal or regulatory obstacles arising as a result of backlash from law enforcement agencies and unions. Finally, the project may be hindered by limited access to stakeholders which can make it difficult to implement reforms and build trustful connections with victims and their families.

Finally, the opportunities of this project lies in capitalizing on the growing public support to influence policymakers. Social media platforms can be utilized to reach out to more people and raise awareness. There can also be potential collaborations with existing support groups, and the scope of the project can be expanded beyond California.

8.0 Conclusion

After extensive analysis and thorough research of the many cases and instances of police brutality we have been able to come across some conclusions that come in line with previously done research. For one major thing there is the claim that in proportion to their population count in California, African Americans are around three times more likely to die within police custody,

which we have seen through our findings of this dataset to be true. Through different visualizations that have normalized the death and calculating the death rate of all ethnicities, we see this to be true. We have also seen that through the implementation of additional datasets that include socio-economic details of populations throughout the counties, that those living in instances of high poverty, low education, and poor living conditions (i.e. high levels of housing burden and different forms of pollution such as air toxicity and solid waste) that these areas tend to have higher cases of death via police custody.

During our research we found out that large scale prisons tend to have higher rates of death compared to local jails. This may be due to multiple factors, including the possibility that authorities in these larger scaled prison systems have lower standards when dealing with inmates under their custody compared to local authorities, especially considering the length of stay and type of charges. Furthermore, this conclusion can speak for the living conditions and the greater health risk to those imprisoned. One connection to consider, would be the fact that the rise of deaths in the year 2020 coincided with the Covid-19 pandemic during which people in congregate settings like prisons were at greater risk of sickness and death.

While our project has been focused on California, future studies can expand the scope to other states and the national level. In addition, more machine learning can be done, particularly, conducting a time based machine learning that can predict the number of deaths in custody. We can also further improve our models by optimizing our training set to more accurately reflect the racial composition of the population. Furthermore, the study can benefit from utilizing more dataset including those on gun violence and police killings.

9.0 References and Links

Alang, Sirry. “The More Things Change, the More Things Stay the Same: Race, Ethnicity, and Police Brutality.” American Journal of Public Health, U.S. National Library of Medicine, Sept. 2018, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6085007/.

Bion, Xenia Shih. “Racism Fuels Double Crisis: Police Violence and Covid-19 Disparities.” California Health Care Foundation, 8 June 2022, https://www.chcf.org/blog/racism-fuels-double-crisis-police-violence-covid-19-disparities/.

“DEADLY FORCE: POLICE USE OF LETHAL FORCE IN THE UNITED STATES.” Amnesty International USA, 17 June 2015, https://www.amnestyusa.org/reports/deadly-force-police-use-of-lethal-force-in-the-united-states/.

Devi, Tanaya, and Roland G. Fryer Jr. “Policing the Police: The Impact of ‘Pattern-or-Practice’ Investigations on Crime.” NBER, 8 June 2020, https://www.nber.org/papers/w27324.

Edwards, Frank. Risk of Being Killed by Police Use of Force in the United States ... - PNAS. 5 Aug. 2019, https://www.pnas.org/doi/10.1073/pnas.1821204116.

Jails & Prisons - Lassen County, CA (inmate Rosters & Records). County Office. (n.d.). Retrieved May 1, 2023, from https://www.countyoffice.org/ca-lassen-county-jails-prisons/

Males, Mike. “Who Are Police Killing?” Center on Juvenile and Criminal Justice, 26 Aug. 2014, https://www.cjcj.org/news/blog/who-are-police-killing-2.

Premkumar, Deepak, et al. “Assessing the Impact of COVID-19 on Arrests in California.” Public Policy Institute of California, Public Policy Institute of California, 1 Feb. 2023, https://www.ppic.org/publication/assessing-the-impact-of-covid-19-on-arrests-in-california/.

Premkumar, Deepak, et al. “Police Use of Force and Misconduct in California.” Public Policy Institute of California, Public Policy Institute of California, 1 Oct. 2021, https://www.ppic.org/publication/police-use-of-force-and-misconduct-in-california/.

USAFacts. “How Many People in Prisons Died of Covid-19?” *USAFacts*, USAFacts, 20 Sept. 2022, https://usafacts.org/articles/how-many-people-in-prisons-died-of-covid-19/#:~:text=From%20March%202020%20to%20February,incarcerated%20people%20from%20the%20virus.

10.0 Appendix

Data Element and Values Defined

Cell Location	Data Element	Description	Value
A	Record Key Number	System generated number that uniquely identifies the record in the system.	• Numeric
B	Reporting Agency	The type of agency reporting the death:	• Other • Local Police • Probation • Sheriff • State
C	Agency Number	The identifier assigned to the reporting agency.	• 6-Digit Alphabet Character/ Numeric
D	Agency Full Name	The literal name of the reporting agency.	• Alpha
E	County	The county the reporting agency resides in.	• Alameda – Yuba • In-State • Out-of-State
F	Race/Ethnicity	The subject's race/ethnicity:	• American Indian • Asian Indian • Black • Cambodian • Chinese • Filipino • Guamanian • Hawaiian • Hispanic • Japanese • Korean • Laotian • Other

			• Other Asian • Pacific Islander • Samoan • Vietnamese • White
G	Gender	The subject's gender.	• Male • Female
H	Age	The age of the subject at time of death.	• Numeric
I	Custody Status	The custody status of the subject immediately preceding death.	• Process of Arrest • In Transit • Awaiting Booking • Booked – No Charges Filed • Booked – Awaiting Trial • Sentenced • Out to Court • Other
J	Custody Offense	The group code for the offense the subject was detained for.	• 3-Digit Numeric
K	Date of Death Year	The Year the subject was pronounced dead by law enforcement or medical authorities.	• YYYY
L	Date of Death Month	The Month the subject was pronounced dead by law enforcement or medical authorities.	• MM
M	Date of Death Day	The day the subject was pronounced dead by law enforcement or medical authorities.	• DD

N	Custodial Responsibility at Time of Death	The agency or facility that has control, care, or custody of the subject immediately preceding death. (This data element should not be used to capture the subject's location where she/he was first diagnosed with, or contracted, a disease such as cancer or AIDS.)	• Process of Arrest • City Jail • County Jail • Adult Camp or Ranch • Local Juvenile Facility/Camp • CDC/CRC • CYA • State Hospital Other
O	Location Where Cause of Death Occurred	The subject's location at the time of an unexpected injury or medical condition that led to death.	• Not Applicable • Crime/Arrest Scene • Administrative • Booking • Living • Common • Holding • Medical Treatment • Other • Unknown
P	Facility of Death	LEA or facility where the subject died.	• Crime/Arrest Scene • Local Hospital • City Jail • County Jail • Adult Camp or Ranch • Local Juvenile • Facility/Camp • CDC/CRC • CYA • State Hospital • Other
Q	Manner of Death	Type of death based on available information.	• Pending Investigation • Natural • Accidental • Suicide • Homicide Willful (Law

			Enforcement Staff) ● Homicide Willful (Other Inmate) ● Homicide Justified (Law Enforcement Staff) ● Homicide Justified (Other Inmate) ● Execution ● Cannot be Determined ● Other
R	Means of Death	Instrument used to cause injuries which contributed to the subject's death.	● Pending Investigation ● Not Applicable (Natural) ● Firearm ● Handgun ● Rifle/Shotgun ● Club, Blunt Instrument ● Hands, Feet, Fists ● Knife, Cutting Instrument ● Hanging, Strangulation ● Drug Overdose ● Mandated Method ● Cannot be Determined ● Other ● Unknown