Table of Contents

Project Plan

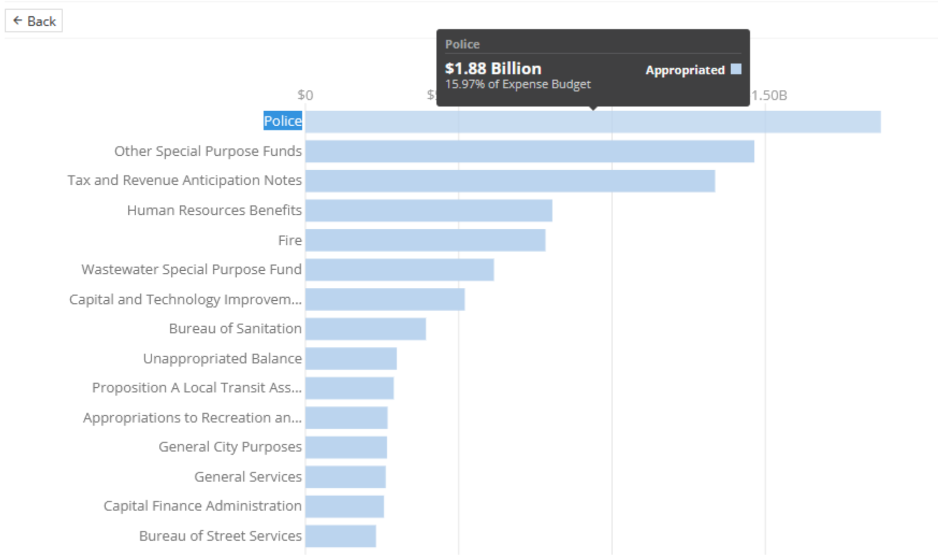
Los Angeles Police Department (LAPD)

Formed: December 13, 1869

https://web.archive.org/web/20141217130309/http://www.laphs.org/history.html

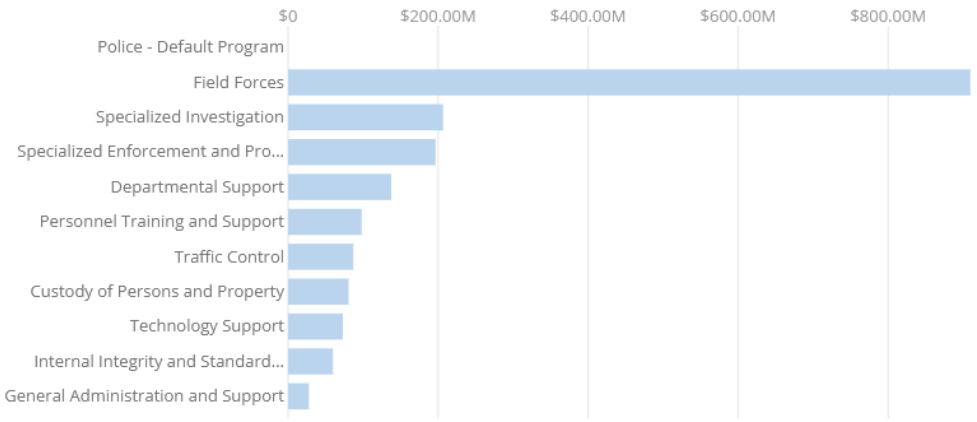
Employees: 12,000

Budget: $1.88 Billion or 15.97% City of Los Angeles Expenses (2023)



(City of Los Angeles, 2023)

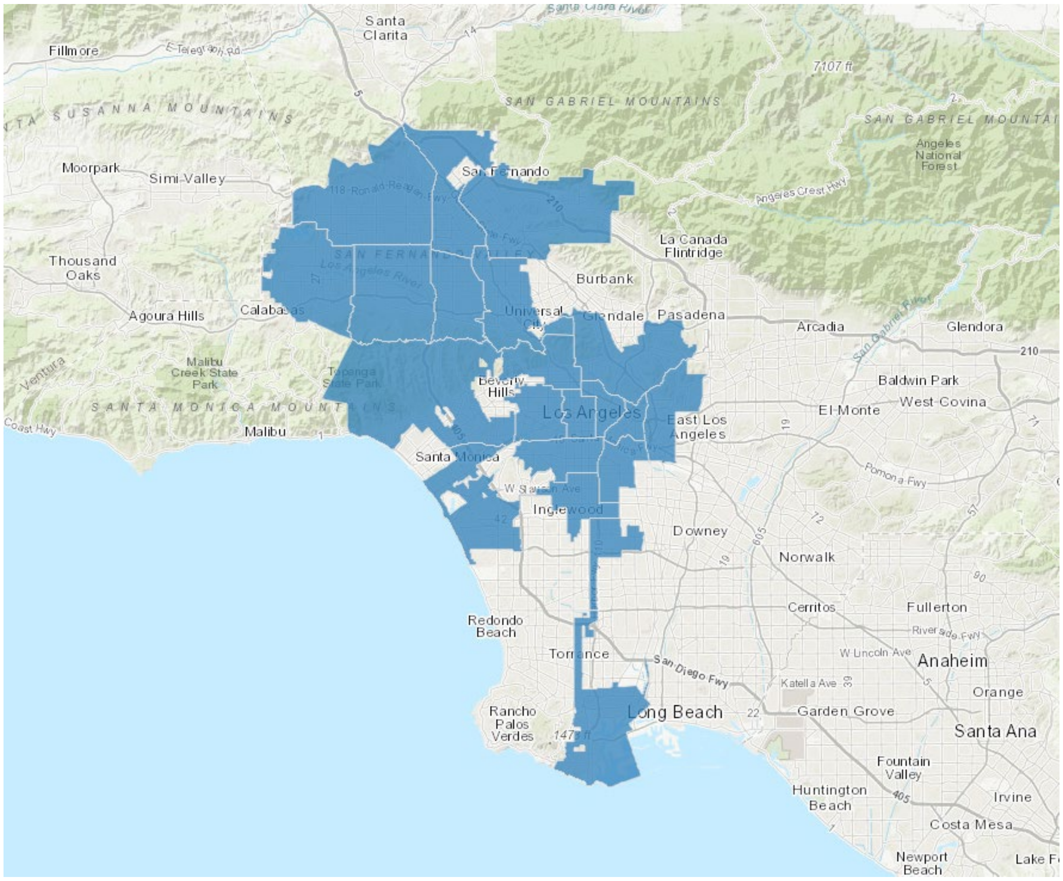
Budget Breakdown:



(City of Los Angeles, 2023)

LA Police Department Description:

Law enforcement is the organized effort “to enforce the law by discovering, deterring, rehabilitating, or punishing people who violate the rules and norms governing society” (New Law Journal, 1974). This data is from the LAPD where their jurisdiction area is shown below.



(Julia, 2021)

According to the LAPD 2022 Homicide Report, there are 3 general categories of homicide that occur: 1) domestic-related, 2) homeless-related, and 3) gang-related. These are useful dimensions to consider when analyzing the nature of crime (Moore, 2022).

Hypotheses

RQ1: What is the status of race and crime today, and has the nature of racism in law enforcement changed?

In a time of rapid technological development with many new phenomena such as the public attention on law enforcement practices and crime, the analysis of both law enforcement practices and crime in a more wholistic and data-centric methods has become increasingly important. Not to mention, there are some other larger trends such as increasing income inequality, the struggle for housing, and the use of drugs. With a data-centric approach, each of these dimensions can be used to evaluate the explanatory power of the respective dimensions to understand the nature of crime and how it has changed over time.

Analyzing the frequency and the type of crime, as well as geographic location will help us answer this question.

RQ2: How has the nature of the acts of crime changed overtime?

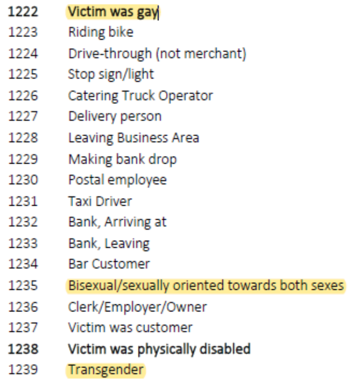
The nature of crime is important to understand for law enforcement to be able to properly discover, deter, rehabilitate, and punish people who violate rules. This analysis can inform the public which may enable their safety or extra precaution. It may also reveal trends in ongoing social justice issues such as racism, discrimination against LGBTQ people, discrimination against homeless people, or other forms of discrimination against certain groups of people. Lastly, it can also reveal trends in law enforcement behavior and practices by comparing with other departments, countries, and periods of time.

RQ3: How have economic factors influenced crime overtime?

The economic factors are very important dimensions to consider because it explains two of the three major categories of homicide: gang-related and homeless-related. Although this is a much more vague category, it opens up the variety of possible trends that can be used to explain crime in a combinatorial way such as educational funding, access to healthcare, food, transportation, etc.

Data

The dataset includes 686,436 observations and 28 attributes. Some useful attributes include the area of the crime, victim age, victim sex, victim descent (White, Black, Hispanic, and many subcategories of Asian), the weapon used, crime code, crime description. There are around 4000 crime codes and each code has unique descriptions that can further be used for the analysis of many specific categories such as discrimination against homeless people, certain races, mentally disabled people, or the LGBQT community. Here is a sample of the codes:



1222	**Victim was gay**
1223	Riding bike
1224	Drive-through (not merchant)
1225	Stop sign/light
1226	Catering Truck Operator
1227	Delivery person
1228	Leaving Business Area
1229	Making bank drop
1230	Postal employee
1231	Taxi Driver
1232	Bank, Arriving at
1233	Bank, Leaving
1234	Bar Customer
1235	Bisexual/sexually oriented towards both sexes
1236	Clerk/Employer/Owner
1237	Victim was customer
1238	**Victim was physically disabled**
1239	Transgender

(Los Angeles Police Department, 2023)

Literature Review

Los Angeles Police Department Homicide Report 2022

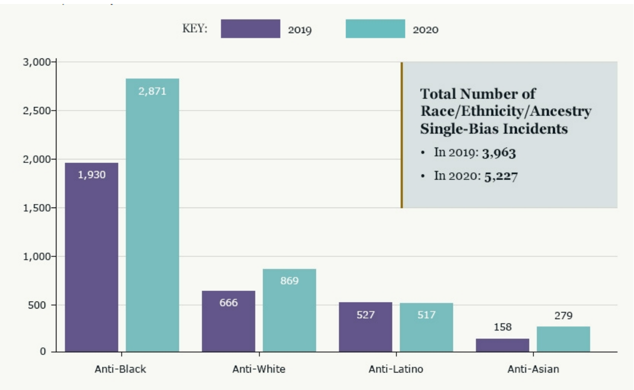
The LAPD publishes a homicide report every year. In this document includes their mission statement and statistics of “year-end at a glance”. Some statistics in this section include total number of homicides by each decade, ethnic makeup and size of the population, the improvements compared to last year (20 less victims than 2021), firearms seized, and a list of firearm-based homicide motives (gang, dispute/revenge, robbery, etc.) (Moore, 2022).

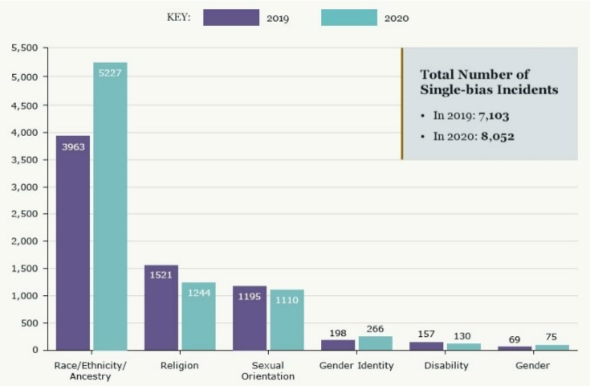
Other major sections include:

- Clearances (related to the solving of a crime)
- Victim and Arrestee Profiles
- Domestic-Related
- Homeless-Related
- Gang-Related
- Operations
- Strategies

Hate Crime Statistics 2020 by the United States Justice Department

The FBI also releases yearly Hate Crime Statistics, and the United States Justice Department has a brief overview of the findings (The United States Department of Justice, 2020).





Additionally, there are more datasets from the FBI and Bureau of Justice Statistics, some of the notable ones are listed below:

- National Incident-Based Reporting System
- National Crime Victimization Survey: Identity Theft Supplement
- Law Enforcement Officers Killed and Assaulted
- The National Use-of-Force Data Collection
- Crime Data Explorer
- Law Enforcement Suicide Data Collection
- Supplementary Homicide Reports

- Survey of Prison Inmates
- Homicides in Chicago, 1965-1995

Learning to Detect Crime

MIT doctoral student, a crime analyst from Chicago Police Department (CPD) and Assoc. Prof. of Statistics at MIT, a Lieutenant from CPD, and Cynthia Rudin (winner of the Squirrel AI Award for Artificial Intelligence for the Benefit of Humanity, "the most prestigious award in the field of artificial intelligence") authored a paper called "Learning to Detect Crime" in which "a pattern detection algorithm called Series Finder", "in order to predict crime, anticipate criminal activity and prevent it" (Rudin & Wang, 2013).

The goal of this algorithm is to significantly reduce the time spent by a typical crime analyst whose job it is to investigate a recent crime in relation to previous crimes that are similar. Specifically, "the algorithm searches through the database looking for similarities between crimes in a growing pattern and in the rest of the database looking for similarities between crimes in a growing pattern and in the rest of the database and tries to identify the modus operandi (M.O.) of the particular offender" (Rudin & Wang, 2013).

The paper lists a few literatures itself:

- "Use of a semi-supervised clustering algorithm to detect crime patterns"
- "A cascaded network of Kohonen neural networks followed by heuristic processing of the network outputs"
- "Association rule mining"
- "Classification"
- "Clustering"
- "Studying hotspots"

Research Questions

Data

Link: https://catalog.data.gov/dataset/crime-data-from-2020-to-present

Secondary Dataset: https://catalog.data.gov/dataset/crimes-2001-to-present

Tertiary Dataset: https://catalog.data.gov/dataset/low-wage-high-violation-industries-324ff

Link to Description: https://dev.socrata.com/foundry/data.lacity.org/2nrs-mtv8

Dimensions

- Number of instances: 686,436
- Number of columns: 28

Attributes

- Date Reported - text
- Date Occurred – floating_timestamp
- Time Occurred – floating_timestamp
- Area – text (according to LAPD 21 Community Police Stations referred to as Geographic Areas within the department)
- Area Name - text
- Reported Distance Number – text (a four-digit code that represents a sub-area within a Geographic Area)
- Part 1-2 (unknown) - number
- Crime Committed - text
- Description of Crime Committed - text
- Modus Operandi – text (activities associated with the suspect in commission of the crime)
- Victim Age - text
- Victim Sex - text
- Victim Descent – text (ethnicity)
- Premise Code – number (the type of structure, vehicle, or location where the crime took place)
- Premise Description - text
- Weapon used - text
- Weapon description - text
- Status – text (code for the status of the case)
- Status description – text (description of the status code)
- Crime Committed 1 – text (code 1 is the primary and most serious one while 2, 3, and 4 are respectively less serious crimes)
- Crime Committed 2 - text
- Crime Committed 3 - text
- Crime Committed 4 - text

- Location – text (street address of crime rounded to the nearest hundred block to maintain anonymity)
- Cross street – text (of rounded addresses)
- Latitude - number
- Longitude – number

Questions

- What is the status of race and crime today, and has the nature of racism in law enforcement changed?
- How has the nature of the acts of crime changed overtime?
- How have economic factors influenced crime overtime?

Comments

In order to answer my 3rd research question, some (such as the 3rd data source listed above) that describes economic hardship need to be associated linked together, perhaps by geographic location.

Abstract and Executive Summary

Abstract

Criminal behavior is a very sophisticated problem that involves an innumerable number of variables. The behaviors behind these criminal acts are also always evolving and often reflect the trends of society that we all feel such as COVID-19 and economic hardship, and at times, may reveal a portrait of how a crime takes place. Specifically, the first topic that is explored is how crime affects certain races (victims). The data reveals some types of crimes, premises, and locations that certain races are disproportionally affected by. The second topic that is explored is how the nature of crime has changed over time. Crimes that have been affected by COVID-19 strongly reflect the economic conditions related to the pandemic. Additionally, a few crimes that have been steadily on the rise have been identified.

Executive Summary

The findings suggest interesting characteristics of the influence of race on crime. It suggests that racism happens more domestically or with someone that they know, rather than with strangers. Also, the data suggests that African Americans endure more racism, even though the population of Los Angeles is more Hispanic.

Unlike racism, the data suggests that crimes that involve an “aimed gun” often do not involve people that the victims know. These types of crimes are characterized by demanding money or property, and the act of taking that property.

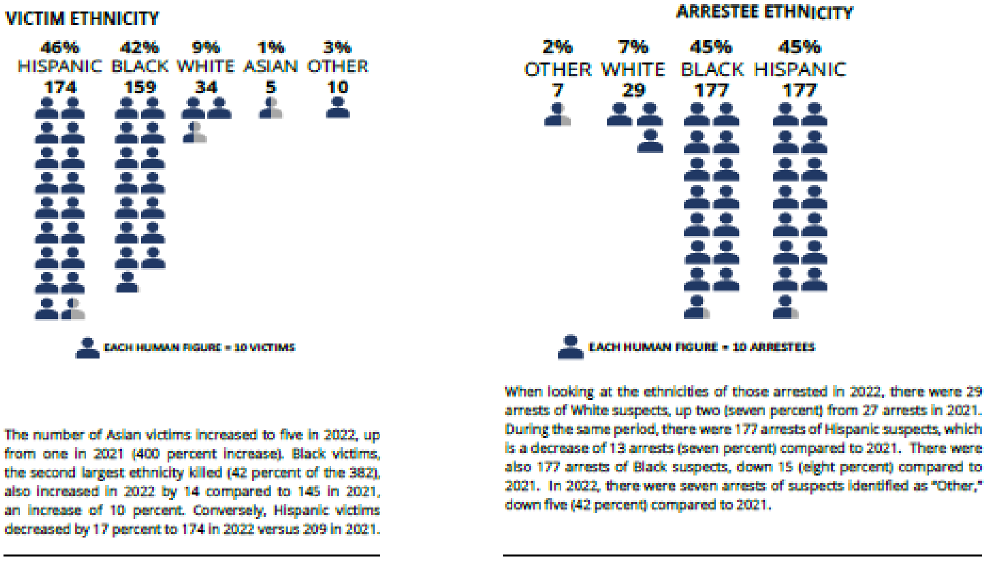
Methodology

Research Question 1 - What is the nature of the acts of crime?

Before performing logistic regression, the data will need to be encoded. One-hot encoding will be used for each m.o. code, but because there are so many unique m.o. codes, only m.o. codes that appear at a frequency of above 100 will be considered.

This question can be answered by using logistic regression and by using the m.o. code of “victim targeted based on Race/Ethnicity/Ancestry” as the target variable. Further analysis might bring insights into the relationship between the nature of crime and race, the location of crime and race, a combination of income/wealth-related attributes and race, or the sheer population density of crime and race. Unfortunately, because of limited computational resources, only other m.o. codes will be considered (there around 200 unique m.o. codes in this data set).

The following is an example of how to visualize the distribution of ethnicities. Another dimension to consider would be the ethnicity profile of the population so that a fairer comparison can be made. Yet another dimension might be to provide several graphs of this overtime to model trends.



The number of Asian victims increased to five in 2022, up from one in 2021 (400 percent increase). Black victims, the second largest ethnicity killed (42 percent of the 382), also increased in 2022 by 14 compared to 145 in 2021, an increase of 10 percent. Conversely, Hispanic victims decreased by 17 percent to 174 in 2022 versus 209 in 2021.

When looking at the ethnicities of those arrested in 2022, there were 29 arrests of White suspects, up two (seven percent) from 27 arrests in 2021. During the same period, there were 177 arrests of Hispanic suspects, which is a decrease of 13 arrests (seven percent) compared to 2021. There were also 177 arrests of Black suspects, down 15 (eight percent) compared to 2021. In 2022, there were seven arrests of suspects identified as "Other," down five (42 percent) compared to 2021.

(Moore, 2022)

RQ2 - How has the nature of the acts of crime changed over time?

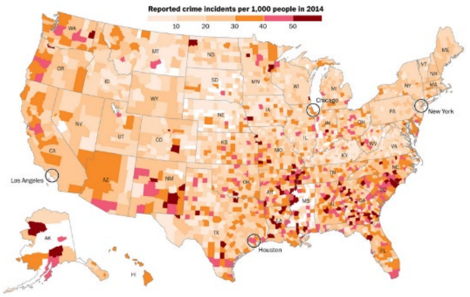
One way to answer this question is to explore the explanatory ability of certain attributes (such as the premise description code which could be a sidewalk, a house, or a parking lot) to predict the type of weapon used. This can also be done by performing a logistic regression with the type of weapon as the target variable and then subsequently ranking the coefficients from greatest to least. Another angle to consider is the model accuracy.

Since there are many reasonably valid target variables to consider, it might also be interesting to investigate models with the highest accuracy score when changing the target variable. Additionally, there could be a column with random values to explore the model's tendency to overfit the data and use that as a baseline to explore the randomness of other target variables.

Because this question is inherently vague, there needs to be a specific target variable that is interesting, useful, and relatively common (or easy to understand). The model used to predict the target variable will be a Decision Tree Classifier. Additionally, it can be pruned to adjust for overfitting of the model.

RQ3 - How have economic factors influenced crime overtime?

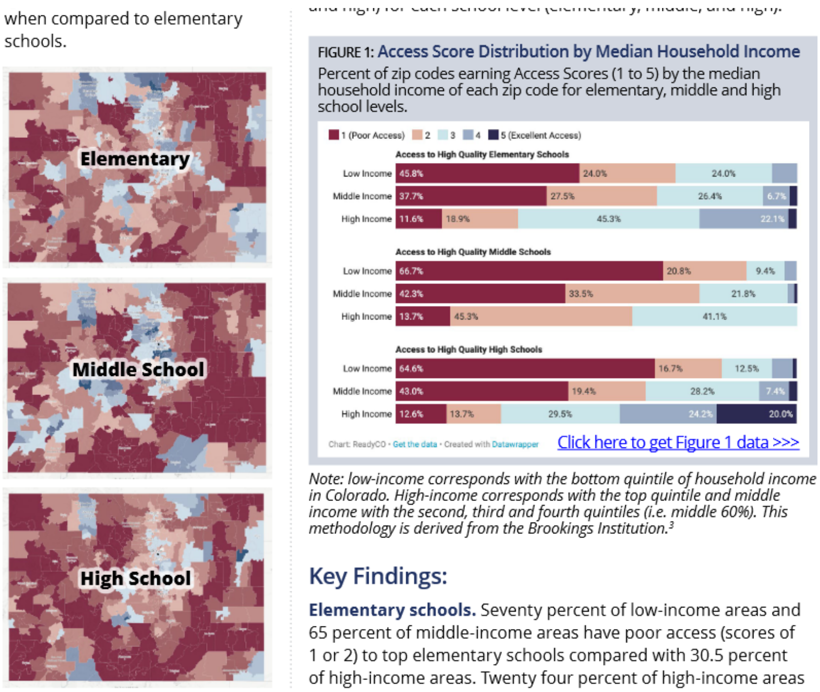
We can create a heat map or a map with labels representing the type of crime and the intensity of the heat by frequency of the crime. Another layer of the map might include different gang-related crimes, domestic-related crime, and homeless-related crime. Another dataset (ArcGIS, 2021) that predicts the safety of walkable areas can be used in combination with neighborhoods that are considered safe for walking to predict the safety of a neighborhood based on the 28 attributes of the crime dataset (although this is based on reports and should be used carefully). Although this would require some processing of the data, the walkability of the areas could be used as the target variable in a logistic regression model.



A heat map similar to this one, but for California would allow us to see the comparison between crime and

(Keating & Lu, 2016)

An example of the type of heat map that can be used where household income can be used to understand to the economic status of the locations.


when compared to elementary schools.

Elementary
Middle School
High School

and high) for each school level (elementary, middle, and high).

FIGURE 1: Access Score Distribution by Median Household Income

Percent of zip codes earning Access Scores (1 to 5) by the median household income of each zip code for elementary, middle and high school levels.

1 (Poor Access)
2
3
4
5 (Excellent Access)
Access to High Quality Elementary Schools
Low Income 45.8% 24.0% 24.0%
Middle Income 37.7% 27.5% 26.4% 6.7%
High Income 11.6% 18.9% 45.3% 22.1%
Access to High Quality Middle Schools
Low Income 66.7% 20.8% 9.4%
Middle Income 42.3% 33.5% 21.8%
High Income 13.7% 45.3% 41.1%
Access to High Quality High Schools
Low Income 64.6% 16.7% 12.5%
Middle Income 43.0% 19.4% 28.2% 7.4%
High Income 12.6% 13.7% 29.5% 24.2% 20.0%

Chart: ReadyCO • Get the data • Created with Datawrapper

Click here to get Figure 1 data >>>

Note: low-income corresponds with the bottom quintile of household income in Colorado. High-income corresponds with the top quintile and middle income with the second, third and fourth quintiles (i.e. middle 60%). This methodology is derived from the Brookings Institution.[3]

Key Findings:

Elementary schools. Seventy percent of low-income areas and 65 percent of middle-income areas have poor access (scores of 1 or 2) to top elementary schools compared with 30.5 percent of high-income areas. Twenty four percent of high-income areas

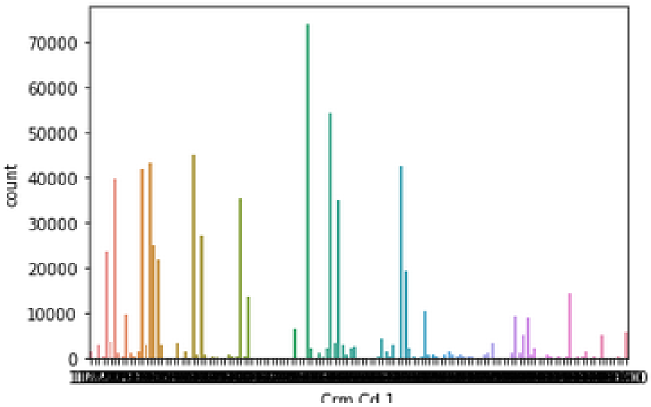
(Ready Colorado, 2022)

Exploratory Data Analysis

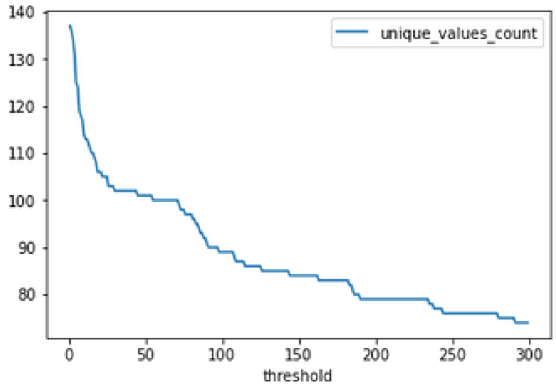
I first wanted to investigate the distribution of the types of crime by looking at the frequency of each type of crime. Because there were so many crime committed codes, I suspected that many of them were not used or very rarely used. Indeed, there were only 141 unique codes for the "Crm Cd 1" or Primary Crime Committed column.

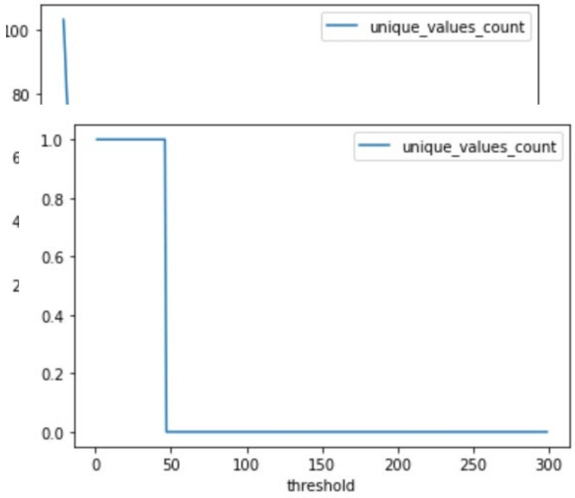
The 'crime committed' were separated into 4 categories, where code 1 is the primary and most serious one and 2, 3, and 4 are respectively less serious crimes. The second, third, and fourth crimes are associated with the same instance or row (person).

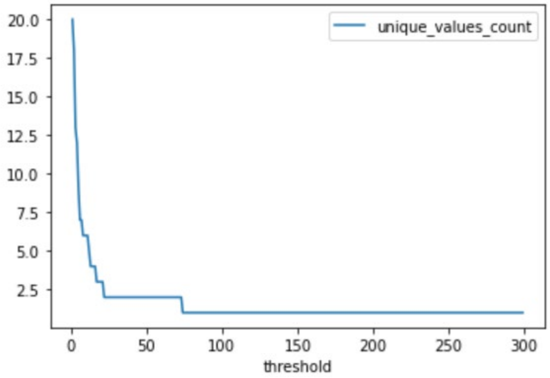
To get a general idea of the diversity and frequency of the crimes, I plotted the count of all 141 unique crimes committed.



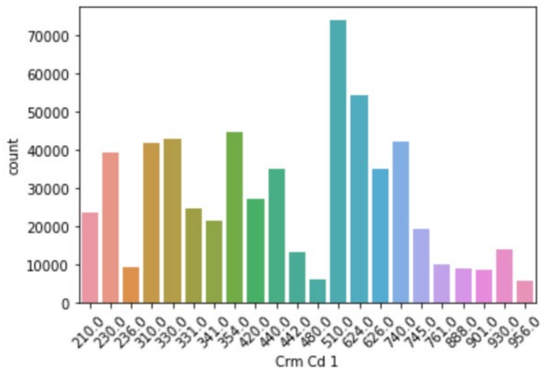
Even within the 141 unique categories, I noticed that there were a significant number of crimes that were only committed a few times (as far as I can see on the graph). So, I plotted the number of unique crime codes committed based on a certain threshold. Specifically, the threshold filters out crime codes if the count of the crime codes is less than 1, 2, 3, and so on to 300. Notice that there is a significant drop-off where the crime codes' frequency is less than 50 and really starts to flatten out above 200. Expanding this graph to test thresholds of up to 10,000 or 20,000 would give a more complete picture but would take significant computational power and time. Below are graphs of the changing number of unique values when an increasing threshold is applied. The top left is associated with 'Cr Cd 1' and top right is associated with 'Cr Cd 2'.

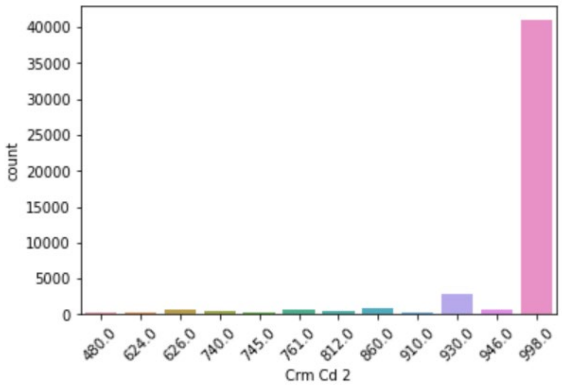


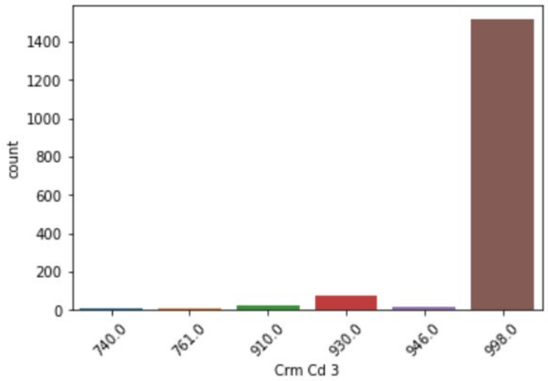


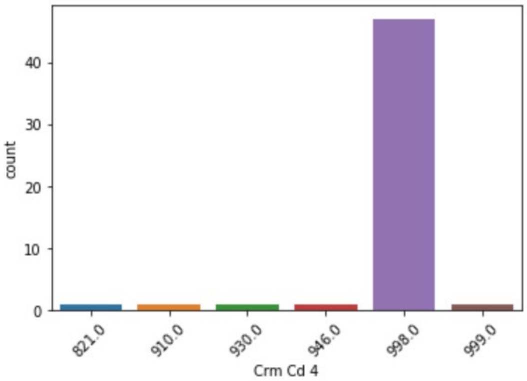


Applying the threshold values to the frequency graphs allows for much more visibility for the type of crime on the horizontal axis. The frequency bar graphs for the other secondary, tertiary, and quaternary crimes are worth looking at as well. There seems to be patterns associated with cases that involve multiple crime sprees. The secondary crime column has a crime committed code that is way higher than the others.



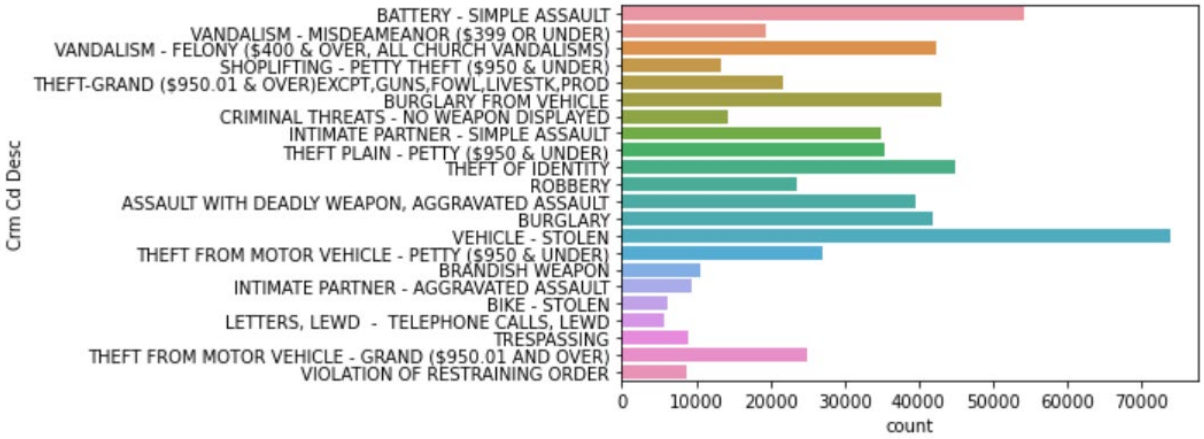




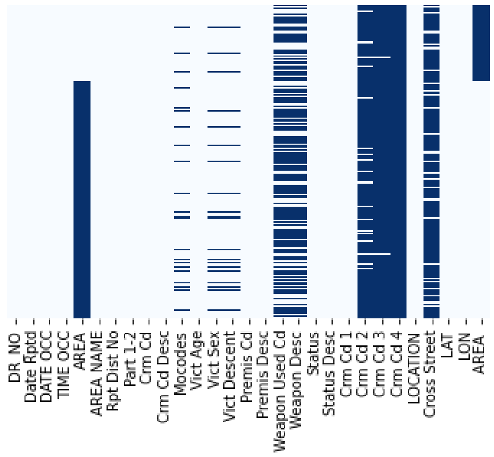


In cases where there is a quaternary crime, there are significantly less unique values. Code 998 is associated with the crime description which states 'brandish weapon'.

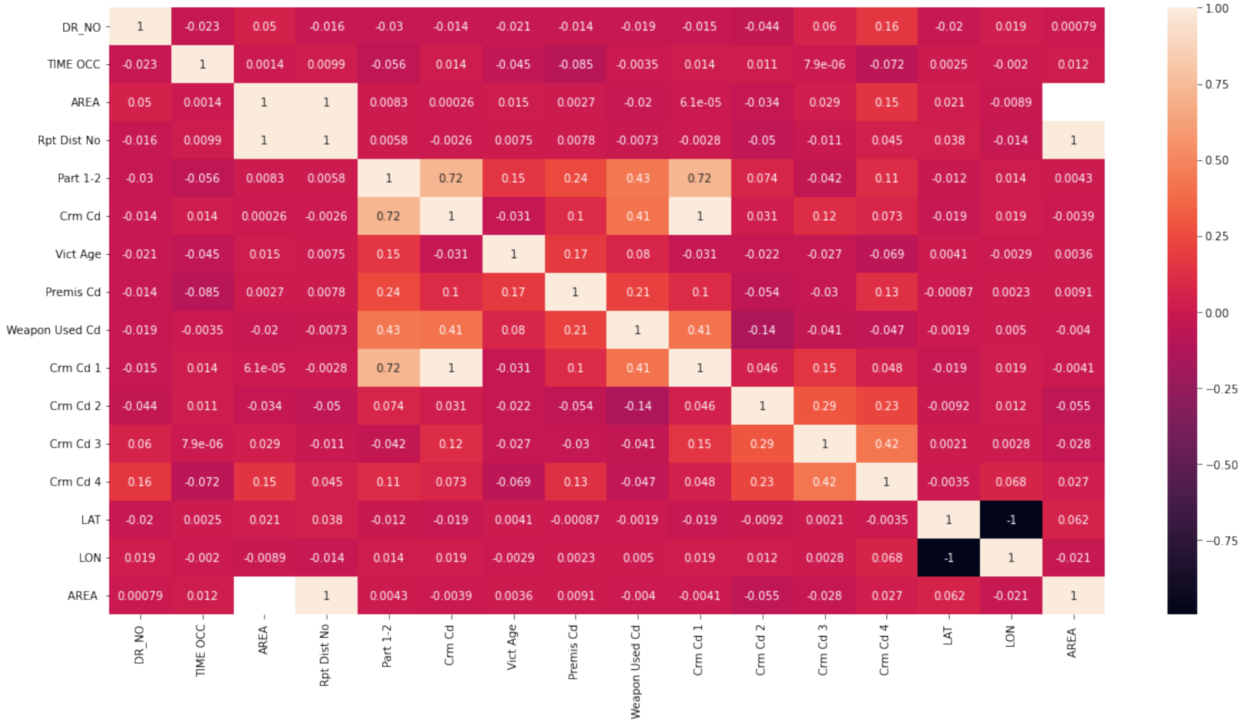
The crime codes are obviously difficult to decipher, so using the "Cr Cd Desc" or Crime Code Description column is much more suitable:



A heatmap of the missing values is showing below:



A heat map of the correlation between each of the categories is shown below.



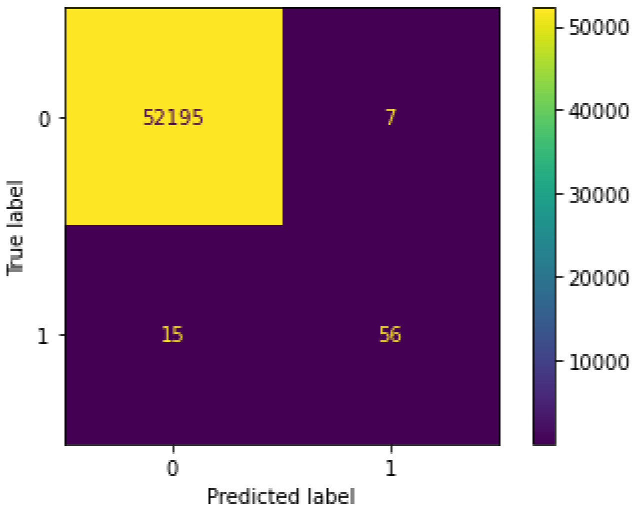
	DR_NO	TIME OCC	AREA	Rpt Dist No	Part 1-2	Crm Cd	Vict Age	Premis Cd	Weapon Used Cd	Crm Cd 1	Crm Cd 2	Crm Cd 3	Crm Cd 4	LAT	LON	AREA
DR_NO	1	-0.023	0.05	-0.016	-0.03	-0.014	-0.021	-0.014	-0.019	-0.015	-0.044	0.06	0.16	-0.02	0.019	0.00079
TIME OCC	-0.023	1	0.0014	0.0099	-0.056	0.014	-0.045	-0.085	-0.0035	0.014	0.011	7.9e-06	-0.072	0.0025	-0.002	0.012
AREA	0.05	0.0014	1	1	0.0083	0.00026	0.015	0.0027	-0.02	6.1e-05	-0.034	0.029	0.15	0.021	-0.0089	
Rpt Dist No	-0.016	0.0099	1	1	0.0058	-0.0026	0.0075	0.0078	-0.0073	-0.0028	-0.05	-0.011	0.045	0.038	-0.014	1
Part 1-2	-0.03	-0.056	0.0083	0.0058	1	0.72	0.15	0.24	0.43	0.72	0.074	-0.042	0.11	-0.012	0.014	0.0043
Crm Cd	-0.014	0.014	0.00026	-0.0026	0.72	1	-0.031	0.1	0.41	1	0.031	0.12	0.073	-0.019	0.019	-0.0039
Vict Age	-0.021	-0.045	0.015	0.0075	0.15	-0.031	1	0.17	0.08	-0.031	-0.022	-0.027	-0.069	0.0041	-0.0029	0.0036
Premis Cd	-0.014	-0.085	0.0027	0.0078	0.24	0.1	0.17	1	0.21	0.1	-0.054	-0.03	0.13	-0.00087	0.0023	0.0091
Weapon Used Cd	-0.019	-0.0035	-0.02	-0.0073	0.43	0.41	0.08	0.21	1	0.41	-0.14	-0.041	-0.047	-0.0019	0.005	-0.004
Crm Cd 1	-0.015	0.014	6.1e-05	-0.0028	0.72	1	-0.031	0.1	0.41	1	0.046	0.15	0.048	-0.019	0.019	-0.0041
Crm Cd 2	-0.044	0.011	-0.034	-0.05	0.074	0.031	-0.022	-0.054	-0.14	0.046	1	0.29	0.23	-0.0092	0.012	-0.055
Crm Cd 3	0.06	7.9e-06	0.029	-0.011	-0.042	0.12	-0.027	-0.03	-0.041	0.15	0.29	1	0.42	0.0021	0.0028	-0.028
Crm Cd 4	0.16	-0.072	0.15	0.045	0.11	0.073	-0.069	0.13	-0.047	0.048	0.23	0.42	1	-0.0035	0.068	0.027
LAT	-0.02	0.0025	0.021	0.038	-0.012	-0.019	0.0041	-0.00087	-0.0019	-0.019	-0.0092	0.0021	-0.0035	1	-1	0.062
LON	0.019	-0.002	-0.0089	-0.014	0.014	0.019	-0.0029	0.0023	0.005	0.019	0.012	0.0028	0.068	-1	1	-0.021
AREA	0.00079	0.012		1	0.0043	-0.0039	0.0036	0.0091	-0.004	-0.0041	-0.055	-0.028	0.027	0.062	-0.021	1

Analysis and Data Visualizations

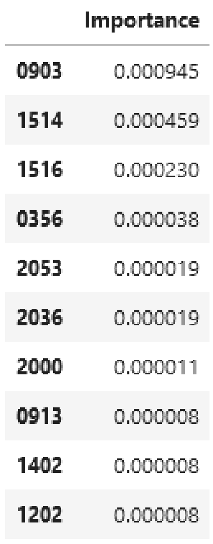
Research Question 1 - What is the influence of race on crime?

Logistic Regression

A logistic regression model was used to predict the associated m.o. codes with the target m.o. code of "victim targeted based on Race/Ethnicity/Ancestry". The accuracy of the model was rounded to 100% on the test set, but is likely due to overfitting. Additionally, the labels that do not have m.o. code based on race (m.o. code 2055) heavily outnumbered the ones that do as seen below in the confusion matrix.



According to the logistic regression model, the most important m.o. codes that predicts the m.o. code related to race, "victim targeted based on Race/Ethnicity/Ancestry" are the following m.o. codes:



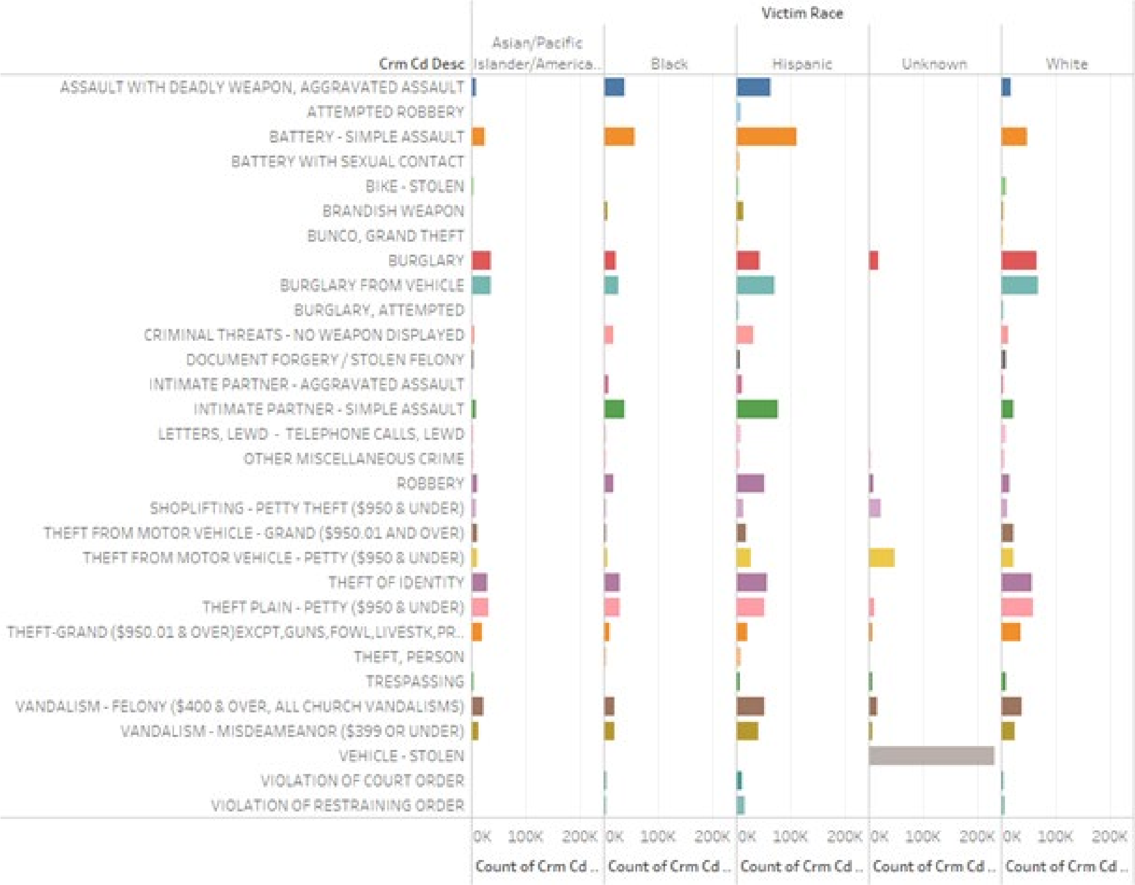
	Importance
0903	0.000945
1514	0.000459
1516	0.000230
0356	0.000038
2053	0.000019
2036	0.000019
2000	0.000011
0913	0.000008
1402	0.000008
1202	0.000008

- 0903 – Hatred/Prejudice
- 1514 - Bias: Anti-Black or African American
- 1516 - Bias: Anti-Hispanic or Latino
- 0356 - Suspect spits on victim
- 2053 - Victim targeted based on religion
- 2036 - Hate-related language
- 2000 - Domestic violence
- 0913 - Victim knew Suspect
- 1402 - Evidence Booked (any crime)
- 1202 - Victim was aged (60 & over) or blind/physically disabled/unable to care for self

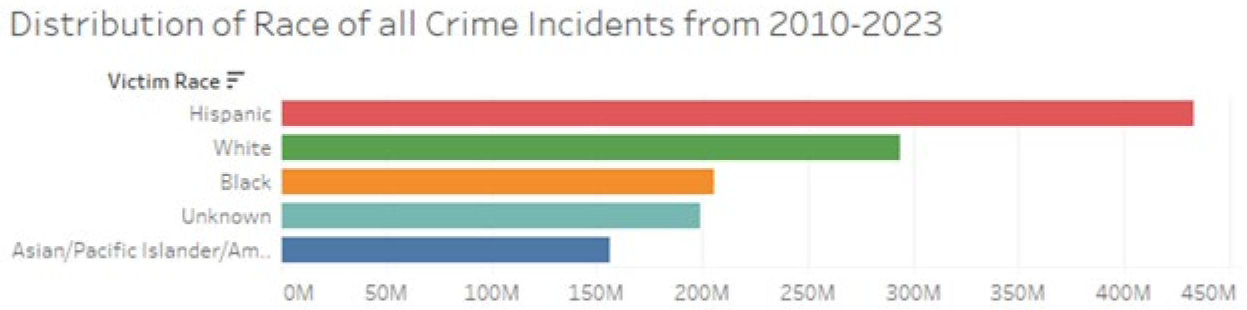
One notable aspect of this data is that is anti-black bias is stronger than anti-hispanic bias even though the Black population is much smaller. Another notable characteristic is the “domestic violence” and “victim knew suspect”. This is in contrast with the target variable “aimed gun” where “stranger” was an associated m.o. code. This may be an important hint at the nature of racism and the context it arises.

Data Visualizations

This figure shows the categories of the top 30 most common crimes by the race of the victims. This type of visualization was chosen to detect any anomalistic distributions where a certain race of a victim might have particularly high frequency.



To show an accurate comparison of the race of the victims, the distribution of the race of victims of all crimes from 2010 to 2022 is displayed below:



For the vast majority of the many types of different crimes, the distribution of the races of victims is roughly proportional to the overall distribution of all crimes in general. There were only a few notable cases where the distribution seemed to be noticeably different.

"Disproportionally affected" (below) in this case means visually being significantly higher than the expected value (based off the distribution of the race of victims of all crime). Some categories should or shouldn't be included because of lack of precision. A future goal would be to have a more accurate (computational) filtering of anomalistic distributions.

Crimes where white victims are disproportionally affected:

- Disturbing the peace
- Pickpocket
- Resisting arrest (the victim in this case is not obvious)
- Unauthorized computer access

Crimes where Hispanic victims are disproportionally affected:

- Shots fired at inhabited dwelling
- Extortion
- Criminal homocide
- Child pornography

Crimes where black victims are disproportionally affected:

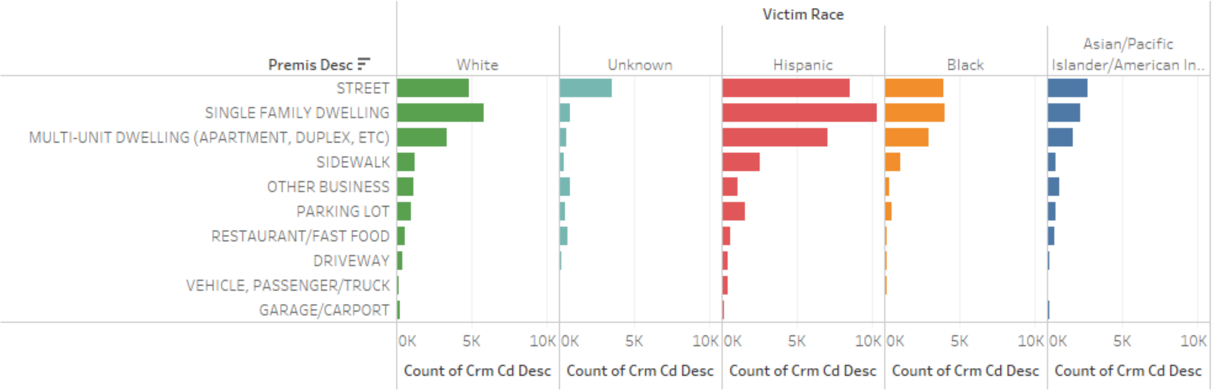
- Criminal homicide
- Pickpocket
- Pimping
- Oral copulation
- Human trafficking
- Credit cards, fraud use (950 and under)
- Child stealing
- Child pornography

Crimes where Asian/Pacific Islander/American Indian victims are disproportionally affected:

- Defrauding Inkeeper (950 or less)
- Document worthless ($200 and over)
- Failure to yield

- Bomb scare
- Pickpocket
- Sex offender registrant out of compliance

This graph excludes the top three locations to better view the distribution of less frequent premises. The top 10 premises is provided below:



A few premises where white victims are disproportionally affected:

- Driveway
- Hotel
- Restaurant/Fast Food

A few premises where hispanic victims are disproportionally affected:

- High school
- Sidewalk
- Park/Playground

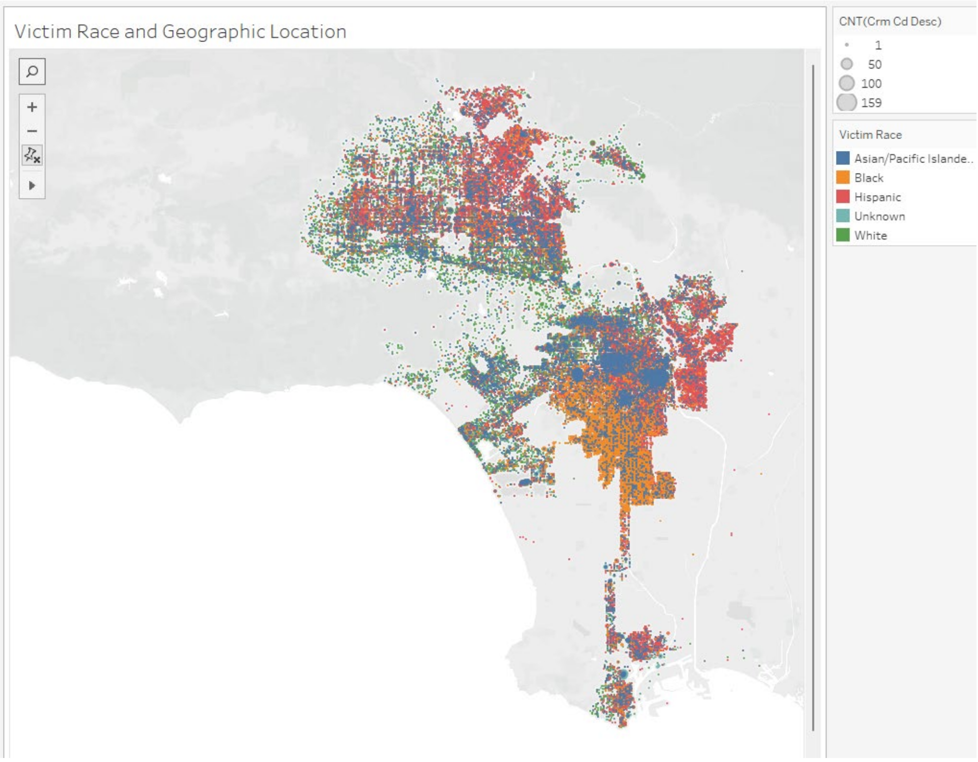
A few premises where black victims are disproportionally affected:

- Motel

A few premises where Asian/Pacific Islander/American Indian victims are disproportionally affected:

- Restaurant/Fast Food
- Other business
- Other store
- Airport
- Gas station

Lastly, a map of where the victims are typically located by race was explored:

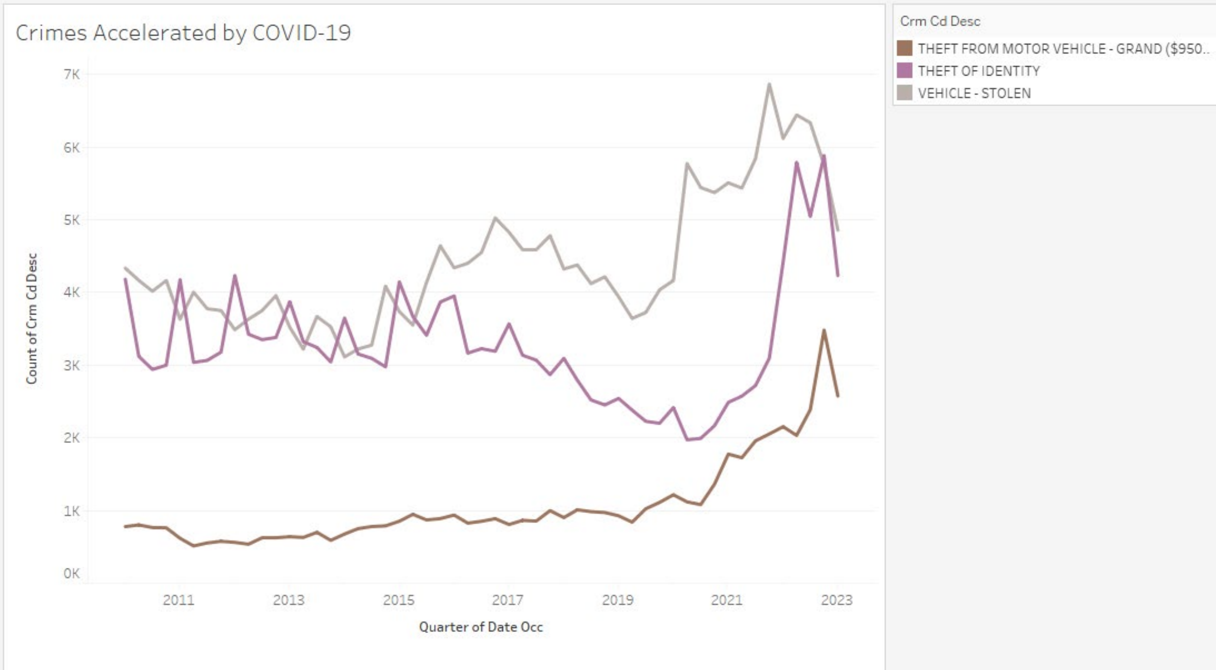


Victim Race and Geographic Location

Research Question 2 – What is the nature of the acts of crime?

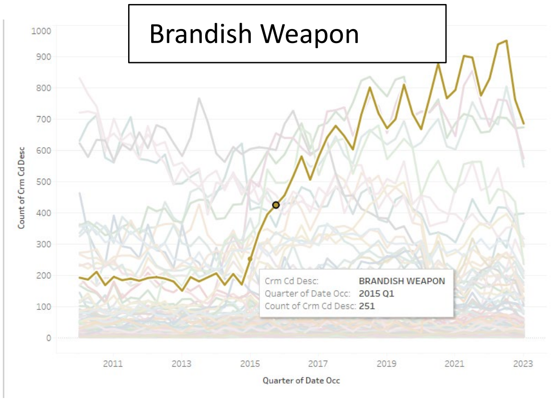
A very consistent trend among all types of crime is the effect of COVID on crime. Most crimes have had a significant drop at the late 2019 mark with the lowest point being 2020 (which is when COVID started in the United States), but there are a few crimes that have had a sharp rise since 2020.
This includes:

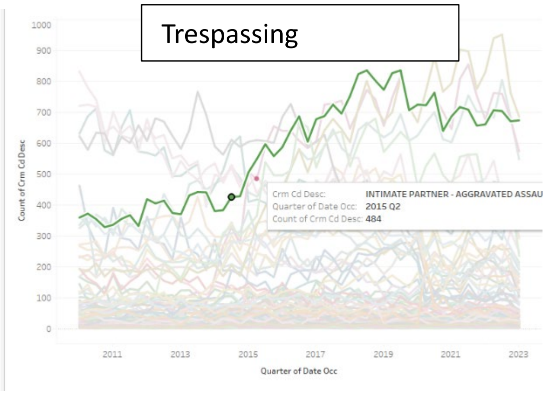
- Theft from motor vehicle ($950 and over)
- Stolen vehicle
- Theft from motor vehicle

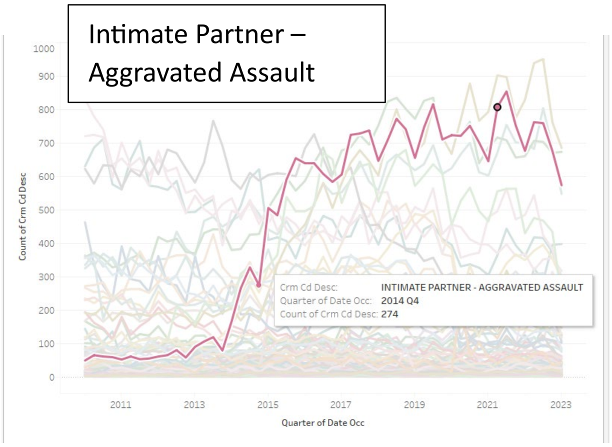


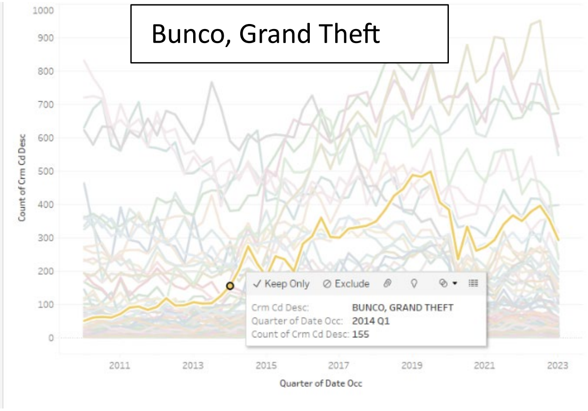
A few theories come to mind. The increased demand for transportation due to the halt of public transportation systems stimulated these crimes. Another contributor to this would be the halt in production of vehicles, which also contributed to the demand for transportation. Lastly, the general inability to access common stores during COVID-19 also increased the demand for personal vehicles. In fact, there is a significant drop-off for crimes in 2023 which is in alignment with the reduction of COVID-19 (although this data may be somewhat misleading because this data collected in 2023 only includes a portion of crimes in quarter 1).

Additionally, many crimes from 2011 to 2023 are either relatively consistent or have not increased in frequency significantly. However, there are a few exceptions in the following graphs. Note that the following graphs exclude crimes with over 50,000 occurrences in total to illustrate the trends of less frequent crimes.



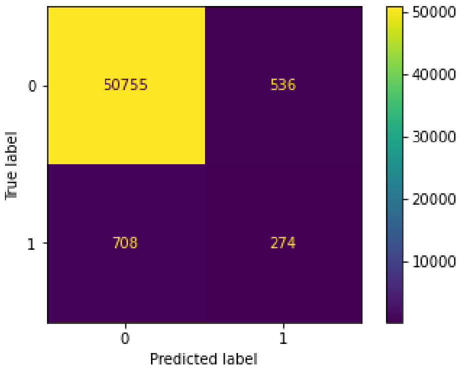




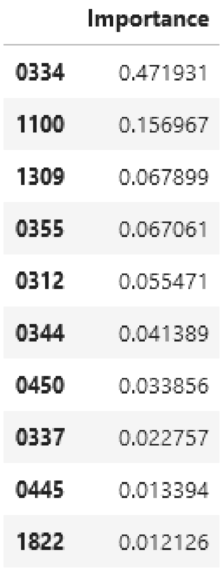


The following graph shows a symbol map of the weapons used. Because using strongarm (hands, fist, feet) as a weapon is so common, it is more representative of the geographic area of LA, rather than a reflection of some characteristic of the crime.

Decision Tree Classifier with Pruning



One important dimension of crime to consider is when guns are involved. Specifically, the m.o. code “aimed gun” is an important attribute to consider. When aimed gun is used as a target variable in a Pruned Decision Tree Classifier with the most significant m.o. codes are listed below. The accuracy score of the model on the training set was 0.996 and for the test set it was 0.976. Although the model still shows some signs of overfitting, it does a better job than the logistic regression model.



	Importance
0334	0.471931
1100	0.156967
1309	0.067899
0355	0.067061
0312	0.055471
0344	0.041389
0450	0.033856
0337	0.022757
0445	0.013394
1822	0.012126

- 0334 – Brandishes weapon
- 1100 – Shots fired
- 1309 – Susp uses vehicle
- 0355 – Demanded property other than money
- 0312 – Gun in waistband
- 0344 – Removes vict property
- 0450 – Suspect shot at victim (no hits)
- 0337 – Demands money
- 0445 – Suspect swung weapon
- 1822 – Stranger

Many of these characteristics are relatively obvious; however, there are a few characteristics that reveal the nature of the “aimed gun” m.o. code. Specifically, it is characterized by “demanding” money or property other than money.

Ethical Recommendations and Implications

Data Bias

Another consideration for the data associated with "Crime Data from 2020 to Present" is that it seems to reflect only reports that are coordinated with the victim and law enforcement (although the exact details are not known). This is implied by the absence of the description of the perpetrator. In conclusion, the data is biased towards victim and law enforcement perception of the crime.

Safety (Harm)

One way that the data can harm others is that it can inform behaviors of crime. One historically well-regarded theory of crime is that criminal behavior is learned (Sutherland & Cressey, 1966). By providing knowledge of the nature of crime (including the location, the weapon used, the victim description), these types of crimes may be learned. "Learning to be Deviant," by Edwin Sutherland and Donald Cressey, 1966. *Principles of Criminology,* pp. 78-83

Immutable Characteristics

Immutable characteristics are defined as "decisions that rest on immutable characteristics deny people that possess these characteristics the agency to achieve different outcomes from the decision-making process, effectively condemning all such people to adverse outcomes" (Fairness and Machine Learning, p.79). There are certain immutable characteristics of the study.

> **Ethnicity** The ethnic makeup of California is not something that components of the data collection process has agency over. For example, at times, there may be characterizations of certain races through frequency, but it is important to consider that overall population of California.
>
> **Laws** Another component of that the data collection process does have agency over is the laws of California. In Howard Becker's seminal justice theory of labeling, he illustrates how an innocent crime can be perceived as worse than it is by the third-party attention, and this can be a factor for both laws and outcomes of the verdict of the crime (Becker, 1963). Becker's theory helps identify the differences in jurisdiction and law enforcement practices of criminal behavior in certain districts, states, and countries. This is an important consideration because outcomes of this study may characterize crimes without much consideration of the differences in laws, law enforcement practices, and victim reaction on multiple levels (district, state, country). "Outsiders," by Howard Becker, 1963. pp. 8-14 and pp. 31-33. MacMillan publishing.

Challenges

Encoding Crime Code into Categories

Originally the crimes were supposed to be encoded into the following categories: 1) crimes against a person, 2) crimes against property, 3) inchoate crimes, 4) statutory crimes, and 5) financial crimes. However, because of the lack of knowledge in criminology, law, and law enforcement operations, this was not in my expertise to accurately determine.

Computing Anomalistic Distributions of Race of Victims for each Crime

Without a formal method to detect what comprises of a distribution that is “anomalistic”, the only way that I was able to categorize crimes/premises that certain races were more vulnerable to was through visual aids. This would be a significant challenge to overcome given its technical complexity and lack of support in Tableau.

Map Technical Challenges

Because there were so many instances, it was difficult to see overall trends. In addition, drawing out districts and having contextual clues (such as population makeup, residential/commercial areas, streets, etc.) would require even more data and most-likely more sophisticated processes and tooling.

Recommendations and Next Steps

Interpretation of Data

Law Enforcement

Certain races of crime are disproportionally affected by certain crimes. This knowledge in combination with saturations of certain races might aid in predicting where and when a crime might happen, and it may also aid in a more efficient investigation.

City of Los Angeles

By understanding where dangerous premises are, the city can use this information to build safer neighborhoods, parks, sidewalks, streets, and parking lots. It can also be aided to determine the most effective locations for surveillance systems.

Residents of Los Angeles

This data can help identify certain races that might be vulnerable to crimes such as investigating the combination of weapon, premise, and geographic location in order to better evaluate their safety.

Modeling Next Steps

General

Encoding Only the m.o. codes were encoded, and more encoding would be beneficial to describe if other factors (such as area, district, time of day, etc.) are significant when victims are targeted based on race or ethnicity or when considering the nature of the crime.

Geographic Data Mapping and analyzing data based off geographical data would be beneficial but would require a significant learning curve to overcome. Specialized tools would also need to be considered such as ArcGIS Pro.

Research Question 1 - What is the influence of race on crime?

Overfitting Compared to a Pruned Decision Tree Classifier, the technique to deal with overfitting data is not as elegant. More research into if there are any hyperparameters to adjust or if there are methods to prevent overfitting need to be researched.

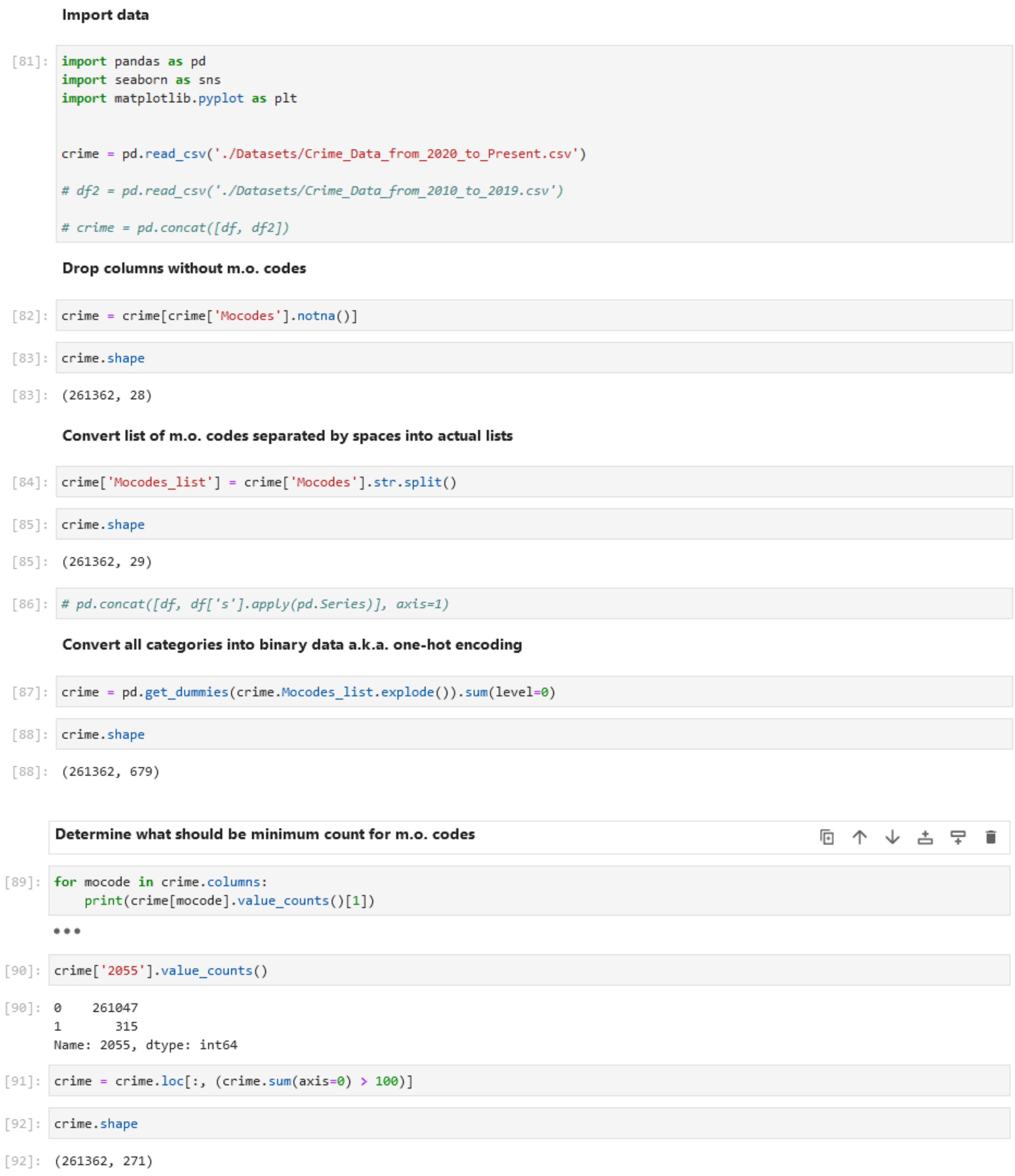
Research Question 2 - What is the nature of the acts of crime?

Hyperparameters Optimization for Overfitting For the Decision Tree Classifier used for predicting the presence of shooting, another step to determine the hyperparameter of depth or pruning level would help to reduce the overfitting of the model.

Target Variable There are other crimes that do not involve guns and targeting other outcomes such as the victim being hurt or murdered could reveal interesting

Code

Logistic Regression



Import data

```
[81]: import pandas as pd
      import seaborn as sns
      import matplotlib.pyplot as plt


      crime = pd.read_csv('./Datasets/Crime_Data_from_2020_to_Present.csv')

      # df2 = pd.read_csv('./Datasets/Crime_Data_from_2010_to_2019.csv')

      # crime = pd.concat([df, df2])
```

Drop columns without m.o. codes

```
[82]: crime = crime[crime['Mocodes'].notna()]

[83]: crime.shape

[83]: (261362, 28)
```

Convert list of m.o. codes separated by spaces into actual lists

```
[84]: crime['Mocodes_list'] = crime['Mocodes'].str.split()

[85]: crime.shape

[85]: (261362, 29)

[86]: # pd.concat([df, df['s'].apply(pd.Series)], axis=1)
```

Convert all categories into binary data a.k.a. one-hot encoding

```
[87]: crime = pd.get_dummies(crime.Mocodes_list.explode()).sum(level=0)

[88]: crime.shape

[88]: (261362, 679)
```

Determine what should be minimum count for m.o. codes

```
[89]: for mocode in crime.columns:
          print(crime[mocode].value_counts()[1])
      •••

[90]: crime['2055'].value_counts()

[90]: 0    261047
      1       315
      Name: 2055, dtype: int64

[91]: crime = crime.loc[:, (crime.sum(axis=0) > 100)]

[92]: crime.shape

[92]: (261362, 271)
```

```
[93]: crime['2055'].value_counts()
```

```
[93]: 0    261047
      1       315
      Name: 2055, dtype: int64
```

•••

```
[95]: X = crime.loc[:, crime.columns != '2055']
      y = crime.loc[:, crime.columns == '2055']
```

```
[96]: y.columns
```

```
[96]: Index(['2055'], dtype='object')
```

```
[97]: from sklearn.model_selection import train_test_split
      X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
```

```
[98]: import seaborn as sns
      import matplotlib.pyplot as plt
```

```
[99]: crime = crime.corr()
      display(crime)
```

•••

```
[102]: from sklearn.linear_model import LogisticRegression
       logreg = LogisticRegression()
       logreg.fit(X_train, y_train)
```

```
/apps/spack/anvil/apps/anaconda/2021.05-py38-gcc-8.4.1-vrzyh2x/lib/python3.8/site-packages/sklearn/utils/validation.py:63: Da
taConversionWarning: A column-vector y was passed when a 1d array was expected. Please change the shape of y to (n_samples,
), for example using ravel().
  return f(*args, **kwargs)
```

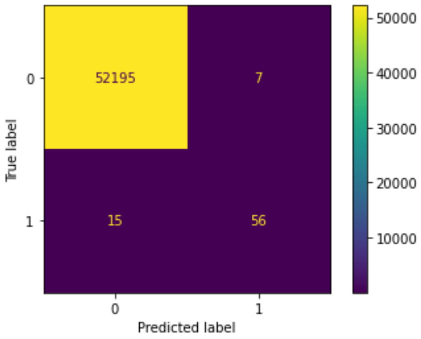
```
[102]: LogisticRegression()
```

```
[103]: y_pred = logreg.predict(X_test)
       y_pred_train = logreg.predict(X_train)
```

```
[104]: from sklearn import metrics
       cf_matrix = metrics.confusion_matrix(y_test, y_pred)
       cf_matrix
```

```
[104]: array([[52195,     7],
              [   15,    56]])
```

```
[105]: from sklearn.metrics import ConfusionMatrixDisplay
       import matplotlib.pyplot as plt
       disp = ConfusionMatrixDisplay(cf_matrix)
       disp.plot()
       plt.show()
```



```
[106]: print("Accuracy: {:.3f}".format(metrics.accuracy_score(y_test,y_pred)))
```

```
Accuracy: 1.000
```

```
[114]: importanceDF = pd.DataFrame(importance.importances_mean, index=X.columns, columns=['Importance'])
       importanceDF.sort_values(by="Importance", ascending=False)
```

[114]:

	Importance
0903	0.000945
1514	0.000459
1516	0.000230
0356	0.000038
2053	0.000019
...	...
0522	0.000000
0527	0.000000
0528	0.000000
0529	0.000000
9999	0.000000

270 rows × 1 columns

Decision Tree

```
[62]: import pandas as pd
      import seaborn as sns
      import matplotlib.pyplot as plt


      crime = pd.read_csv('./Datasets/Crime_Data_from_2020_to_Present.csv')

      # df2 = pd.read_csv('./Datasets/Crime_Data_from_2010_to_2019.csv')

      # crime = pd.concat([df, df2])
```

```
•[63]: # Solution inspired by https://stackoverflow.com/questions/13413590/how-to-drop-rows-of-pandas-dataframe-whose-value-in-a-cer
       crime = crime[crime['Mocodes'].notna()]
```

```
[64]: crime.shape

[64]: (261362, 28)
```

```
•[65]: # Solution inspired by https://stackoverflow.com/questions/51290134/using-pandas-how-do-i-split-based-on-the-first-space
       crime['Mocodes_list'] = crime['Mocodes'].str.split()
```

```
[66]: crime.shape

[66]: (261362, 29)
```

```
•[67]: # Solution inspired by https://stackoverflow.com/questions/61081729/pandas-list-of-values-to-binary-columns
       crime = pd.get_dummies(crime.Mocodes_list.explode()).sum(level=0)
```

```
[68]: crime.shape

[68]: (261362, 679)
```

```
[69]: crime['0302'].value_counts()

[69]: 0    256617
      1      4745
      Name: 0302, dtype: int64
```

```
[69]: crime['0302'].value_counts()
```

```
[69]: 0    256617
      1      4745
      Name: 0302, dtype: int64
```

```
[70]: crime = crime.loc[:, (crime.sum(axis=0) > 100)]
```

```
[71]: crime.shape
```

```
[71]: (261362, 271)
```

```
[72]: crime['0302'].value_counts()
```

```
[72]: 0    256617
      1      4745
      Name: 0302, dtype: int64
```

```
[73]: X = crime.loc[:, crime.columns != '0302']
      y = crime.loc[:, crime.columns == '0302']
```

```
[74]: y.columns
```

```
[74]: Index(['0302'], dtype='object')
```

```
[75]: from sklearn.model_selection import train_test_split
      X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
```

```
[76]: import seaborn as sns
      import matplotlib.pyplot as plt
```

```
[77]: from sklearn.tree import DecisionTreeClassifier
      tree = DecisionTreeClassifier(random_state=0)
      tree.fit(X_train, y_train)
```

```
[77]: DecisionTreeClassifier(random_state=0)
```

```
[78]: print("Accuracy on the training set:{:.3f}".format(tree.score(X_train, y_train)))
```

```
      Accuracy on the training set:0.996
```

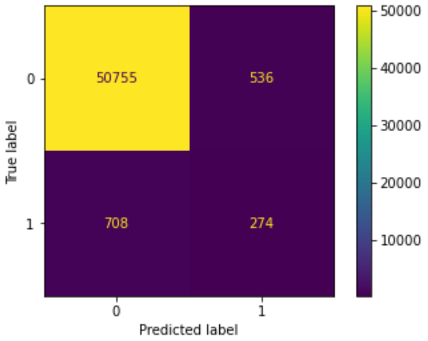
```
[79]: y_pred = tree.predict(X_test)
```

```
[80]: from sklearn.metrics import confusion_matrix
      cm = confusion_matrix(y_test, y_pred)
      print(cm)
```

```
      [[50755   536]
       [  708   274]]
```

```
[81]: from sklearn.metrics import plot_confusion_matrix
      plot_confusion_matrix(tree,X_test,y_test)
```

```
[81]: <sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x7f19238b41c0>
```



[82]:

```
from sklearn.metrics import accuracy_score
print("Accuracy on the test set: {:.3f}".format(accuracy_score(y_pred, y_test)))
```

```
Accuracy on the test set: 0.976
```

[83]:

```
tree_pruned = DecisionTreeClassifier(max_depth= 5, random_state=0)
tree_pruned.fit(X_train, y_train)
```

[83]:

```
DecisionTreeClassifier(max_depth=5, random_state=0)
```

[84]:

```
y_pruned_pred = tree_pruned.predict(X_test)
```

[85]:

```
print("Accuracy on the training set:{:.3f}".format(tree_pruned.score(X_train, y_train)))
print("Accuracy on the test set:{:.3f}". format(accuracy_score(y_pruned_pred, y_test)))
```

```
Accuracy on the training set:0.983
Accuracy on the test set:0.982
```

[94]:

```
importance = pd.DataFrame(tree_pruned.feature_importances_, index = X.columns, columns = ["Importance"])
importance.sort_values(by = "Importance", ascending = False).iloc[0:10,:]
```

[94]:

	Importance
0334	0.471931
1100	0.156967
1309	0.067899
0355	0.067061
0312	0.055471
0344	0.041389
0450	0.033856
0337	0.022757
0445	0.013394
1822	0.012126

References

ArcGIS. (2021, September 22). *Neighborhoods Considered Safe for Walking (2011 & 2015)*. Retrieved from ArcGIS: https://www.arcgis.com/home/item.html?id=54fa5a3e98eb48d88c67fd640a7c2bfe

City of Los Angeles. (2023). *Expense Budget 2023*. Retrieved from Open Budget: https://openbudget.lacity.org/#!/year/2023/operating/0/department_name

Julia, P. (2021, September 30). *LAPD Divisions*. Retrieved from LACity: https://geohub.lacity.org/datasets/031d488e158144d0b3aecaa9c888b7b3_0/explore?location=34.010805%2C-118.320870%2C10.00

Keating, D., & Lu, D. (2016, November 16). *Here's what crime rates by county actually look like*. Retrieved from The Washington Post: https://www.washingtonpost.com/graphics/national/crime-rates-by-county/

Los Angeles Police Department. (2023, April 5). *Crime Data from 2020 to Present*. Retrieved from LACity: https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8

Los Angeles Police Museum. (n.d.). *History of the LAPD*. Retrieved from Los Angeles Police Museum: https://web.archive.org/web/20141217130309/http://www.laphs.org/history.html

Moore, M. R. (2022). *Homicide Report.* Los Angeles: Los Angeles Police Department.

New Law Journal. (1974). *New Law Journal, 123*, 358.

Ready Colorado. (2022). *High Quality Schools: Heat Map*. Retrieved from Colorado School Map: https://coloradoschoolmap.com/wp-content/uploads/2021/01/ReadyCO-HQ-Schools-Heat-Map-Analysis-Final.pdf