Overview

Murder is pervasive in the United States. In 2019, 1 in 20,000 Americans died from intentional homicide. Murder is the 14th most common cause of death in the US and the **6th** most common cause for people aged 15 - 49. Of 230 countries about which the United Nations collects crime data, the United States has the 94th highest homicide rate, placing it well above the median. Combined with the country's large population, that rate results in Americans composing 4% to 5% of all murders that occur worldwide; US police are responsible for the investigation of tens of thousands of murders yearly, usually ranking around 6th largest total number of homicides worldwide.

Murder Accountability Project

Despite this enormous share of all homicides that occur worldwide, there exists at the federal level in the United States no comprehensive repository for all homicide-related data. States, counties, and municipalities collect crime data following their own chosen guidelines and typically are not obligated to submit what they collect to a central authority.

In order to fill this data gap, the Murder Accountability Project (MAP) was founded in 2015. The nonprofit organization collects and aggregates homicide data from multiple entities at all levels of government, including state, federal, and tribal. The aggregated data is then made publicly available on MAP's website to be used freely for analysis for anyone interested.

Project Goals

The primary purpose of the MAP dataset has traditionally been to pinpoint clusters of murders that are likely to be related in order to make law enforcement aware of the existence of currently-unknown serial killers. However, because the dataset includes clearance status for all observed murders, it can also be used to analyze the differences between cases that are cleared, i.e. solved, and cases that are not. **This project aims to create a machine learning-based predictive model that provides a clearance probability for any given murder.** With such a resource, investigative agencies could make better informed, more precise decisions about how to best distribute available resources in order to increase the national clearance rate and provide justice to more victims. **Furthermore, the statistical models generated may provide insight into factors that influence the probability of a murder being solved.**

Exploratory Data Analysis

The Murder Accountability Project's dataset in CSV format can be found here:

https://www.dropbox.com/s/ye37woe6et05qgs/SHR76_19.csv.zip?dl=1

Features

Date
Location
Victim
Offender
Investigator
Crime Characteristics
Clerical

Name	Description	Type
YEAR	Year of Murder	Numerical - Discrete
MONTH	Month of Murder	Numerical - Discrete
CNTYFIPS	County	Categorical - Nominal
STATE	State	Categorical - Nominal
STATENAME	Name of State	Categorical - Nominal
FSTATE	Numerical State Identifier	Categorical - Nominal
MSA	Name of Metro Area of Crime	Categorical - Nominal
VICAGE	Victim Age	Numerical - Discrete
VICSEX	Victim Sex	Categorical - Nominal
VICRACE	Victim Race	Categorical - Nominal
VICETHNIC	Victim Hispanic Identification	Categorical - Nominal
OFFAGE	Offender Age	Numerical - Discrete
OFFSEX	Offender Sex	Categorical - Nominal
OFFRACE	Offender Race	Categorical - Nominal
OFFETHNIC	Offender Hispanic Identification	Categorical - Nominal
ORI	Investigating Agency Number	Categorical - Nominal
AGENCY	Investigating Agency Name	Categorical - Nominal
AGENTYPE	Investigating Agency Type	Categorical - Nominal
SOLVED	**Crime Clearance Status**	**Categorical - Nominal**
HOMICIDE	Murder of Negligence Flag	Categorical - Nominal
SITUATION	Single/Multiple Victim(s)/Offender(s) Description	Categorical - Nominal
WEAPON	Murder Weapon Type	Categorical - Nominal
RELATIONSHIP	Offenders' Relationship to Victim	Categorical - Nominal
CIRCUMSTANCE	Circumstances Surrounding Crime	Categorical - Nominal
SUBCIRCUM	Secondary Circumstances Surrounding Crime	Categorical - Nominal
VICCOUNT	Number of Victims in Entire Related Incident	Numerical - Discrete
OFFCOUNT	Number of Offenders	Numerical - Discrete
ID	Unique Identifier	Numerical - Discrete
SOURCE	Source of data	Categorical - Nominal
INCIDENT	Alternative Identifier	Categorical - Nominal
ACTIONTYPE	Nature of Report (Original or Update)	Categorical - Nominal
FILEDATE	Date Record Added to Dataset	Date

The above

Data Collection and Wrangling

Again, the dataset was provided by the Murder Accountability Project. It can be found here:

www.murderdata.org/p/data-docs.html

The following are the steps I took to prepare the dataset for analysis:

- I first loaded the dataset into a pandas dataframe and viewed its info. Every column except one appeared to have no missing values. The column with missing values is likely unimportant to the analysis.

- I then checked the number of unique values in each column to see if anything strange would appear.

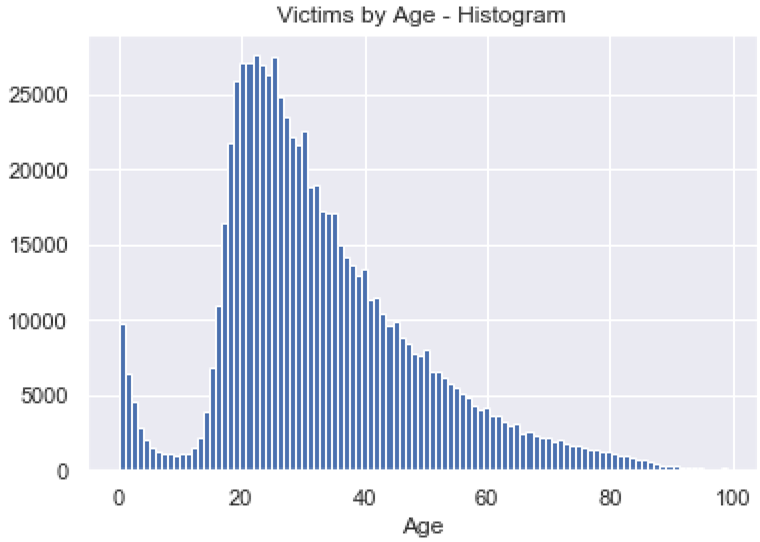
- I noticed that there are 51 states, so I checked to see if the 51st was DC. It was.

- I then checked the unique value names of each column that did not contain very large numbers of unique values.

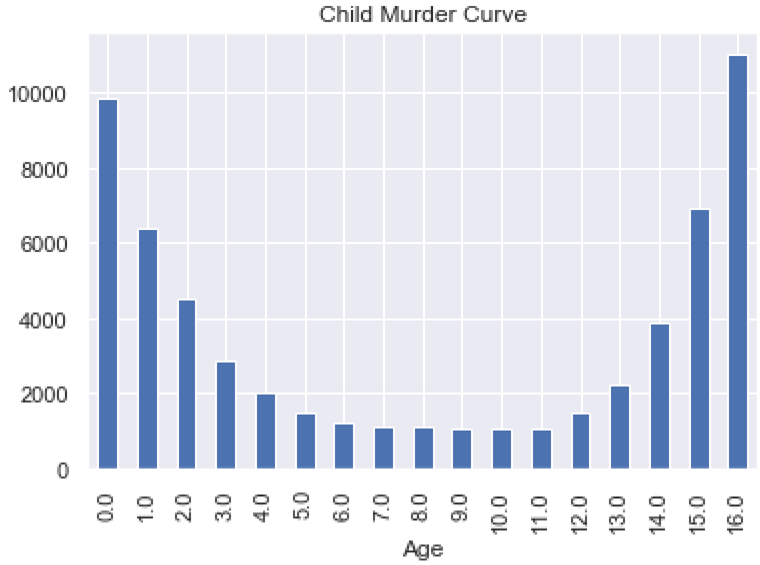
- All unique values looked reasonable except for the use of 999 in the place of unknown victim and offender ages.

- So I generated a dictionary of values (999) to associate with missing data for those two columns and re-loaded the dataset via pandas.

Exploratory Data Analysis

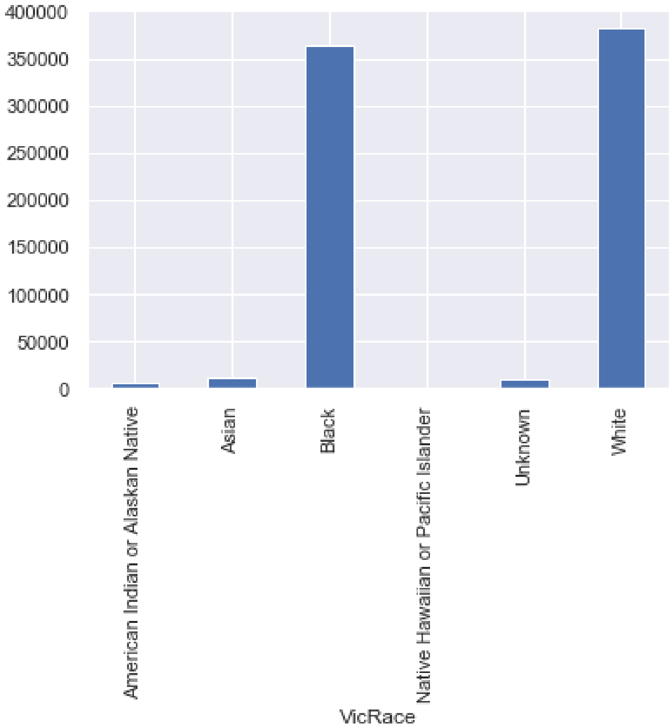


The downward curve post 27 years of age is likely tied to the age of the general population. Also, I notice that there are a lot of very young children who are murder victims.

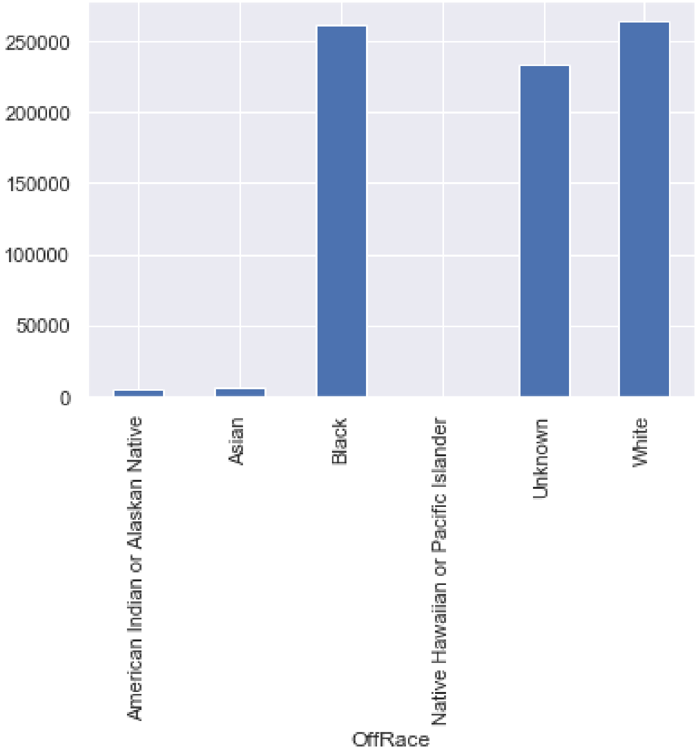


This is depressing, to say the least. It is possible that this results from the data collectors using zero in the place of unknown values, but the smoothness of the curve suggests to me that this may not be the case.

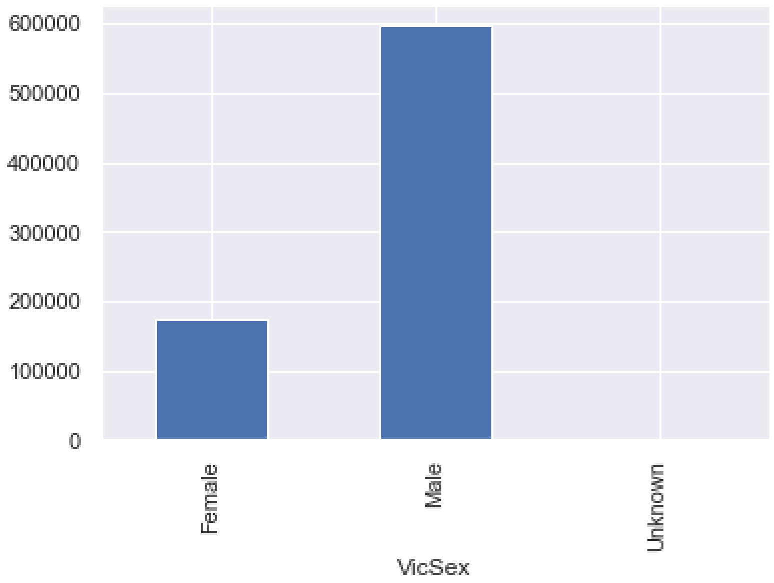
Bar chart of the racial demographics and victims:



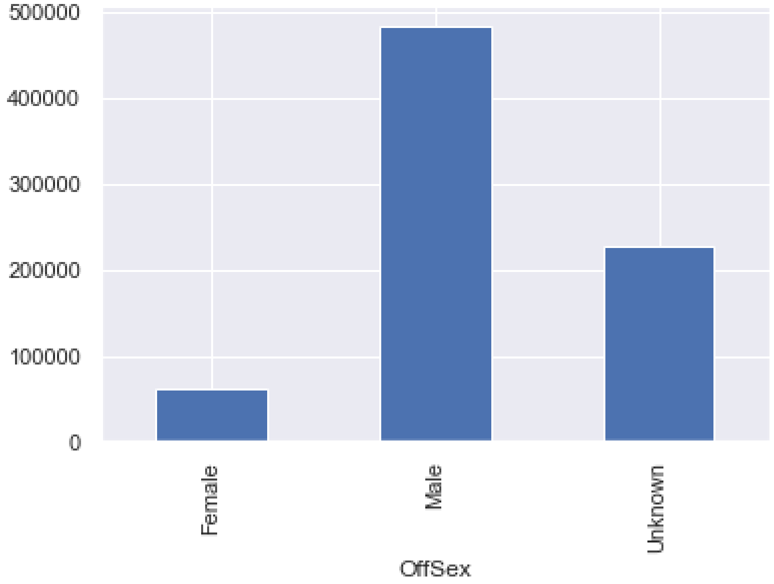
And a bar chart of the racial demographics of murderers. It is very similar except for the unknown group that is associated with unsolved cases.



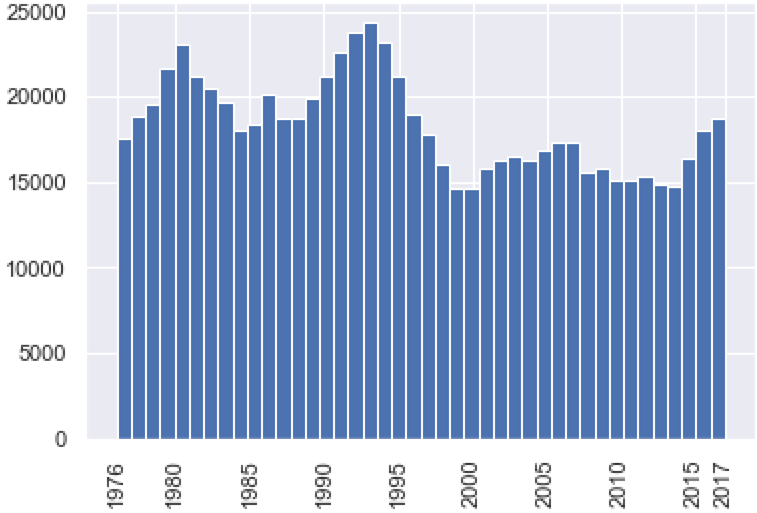
A lot more men are murdered than are women:



And a lot more men are murderers as well:

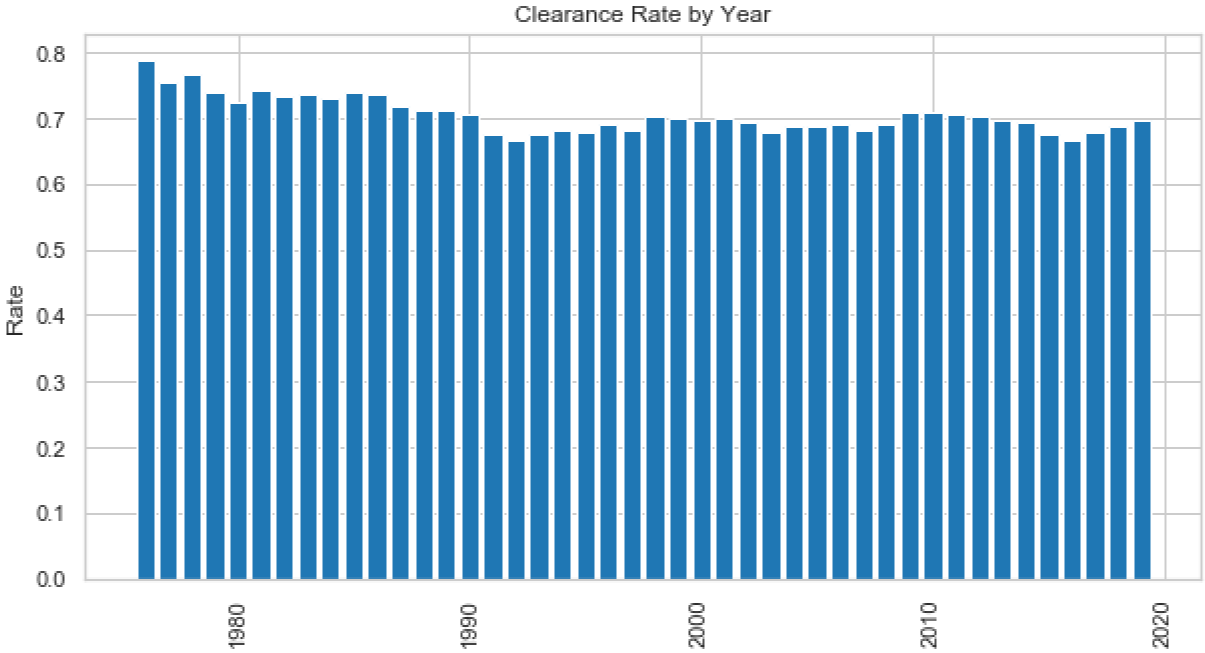


The total number of murders committed each year. On a positive note, there appear to be fewer now than in the past:



Statistical Relationships

In the United States, **30% of murders are unsolved**. And the clearance rate of murders has been getting worse over the past several decades.

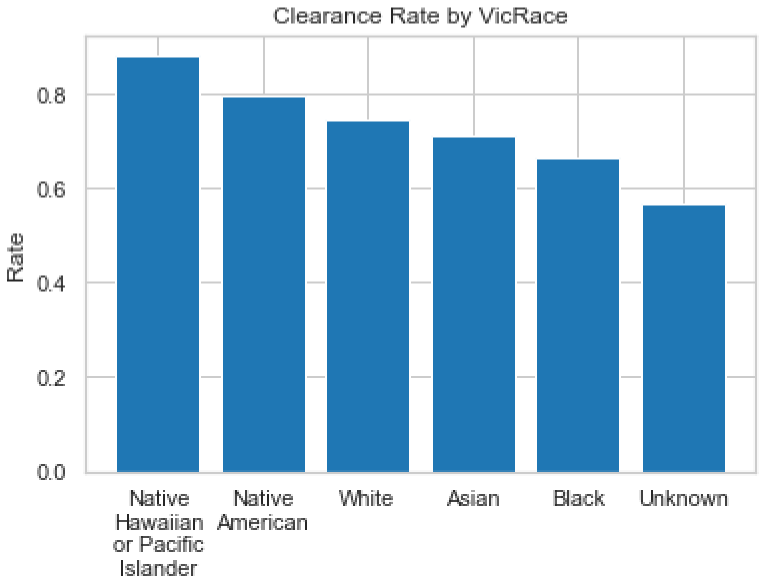


The primary question that I am hoping to answer with this project is:

Is it possible to use the Murder Accountability Project's dataset containing murders in the United States from 1976 to 2019 to predict the likelihood of a murder being solved?

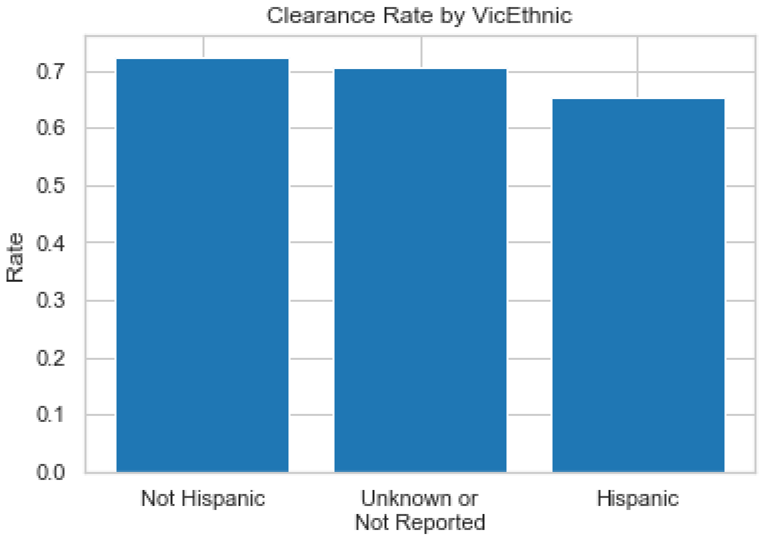
Before fitting the data to various machine learning models, I investigated the relationship that various features have with the predictor variable, a column indicating whether the case was solved or not. Specifically, I calculated the clearance rate (the number of solved cases divided by the total number of cases) for each category within each feature and charted the distribution. I found some results that were surprising and some that were not.

First, I investigated the relationship between the demographics of the victims and the clearance rate. I expected there to be significant differences especially when examining race because of the well-documented institutional and economic racism that exists in the United States.

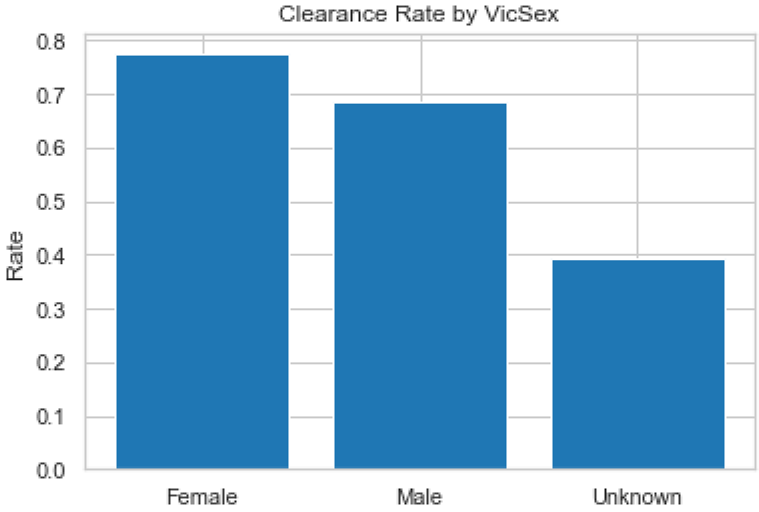


As I expected, murders involving white victims are roughly 10% more likely to be solved than those involving black victims. I was very surprised to see a higher clearance rate for Native Americans - “American Indian or Alaskan Native” in the dataset - and I am skeptical of it. Perhaps this results from the particularities of the specific data that was accessible to the nonprofit group that compiled the dataset. The category for “Native Hawaiian or Pacific Islander” is likely erroneous, because the entire dataset (of 804,751 victims) only contains 92 victims within that racial category.

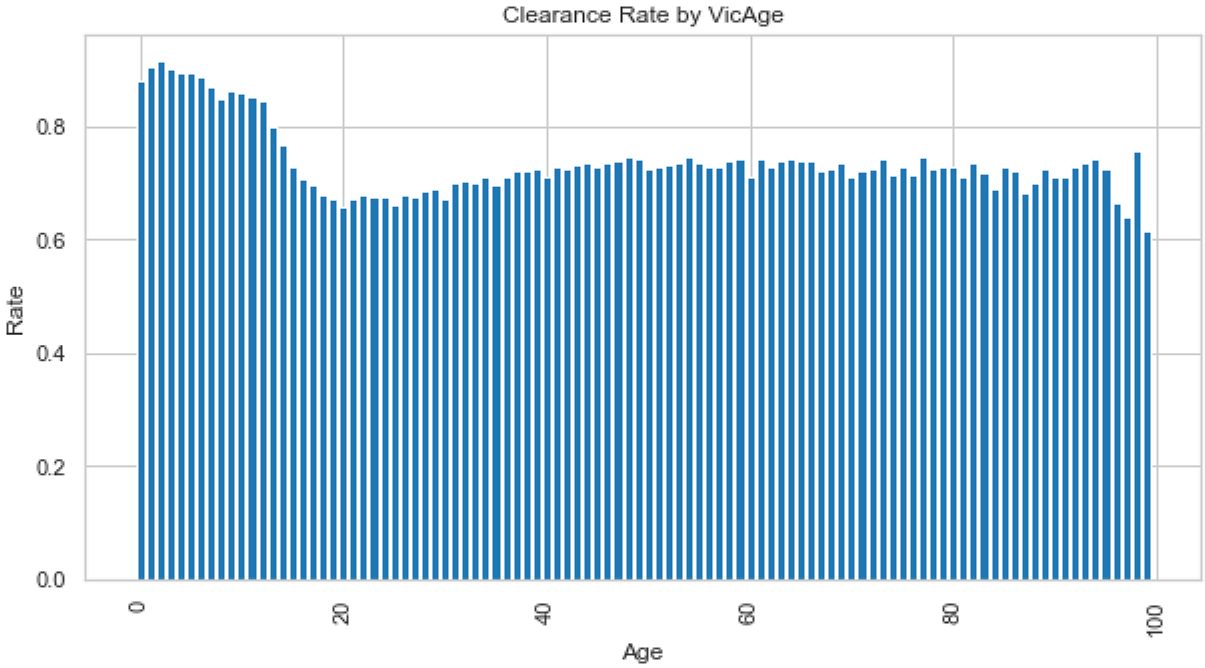
There was a similar relationship suggesting some kind of institutional and/or economic bias for the category of Hispanic vs non-Hispanic origin, which is (correctly) organized as a feature separate from the victim's race.



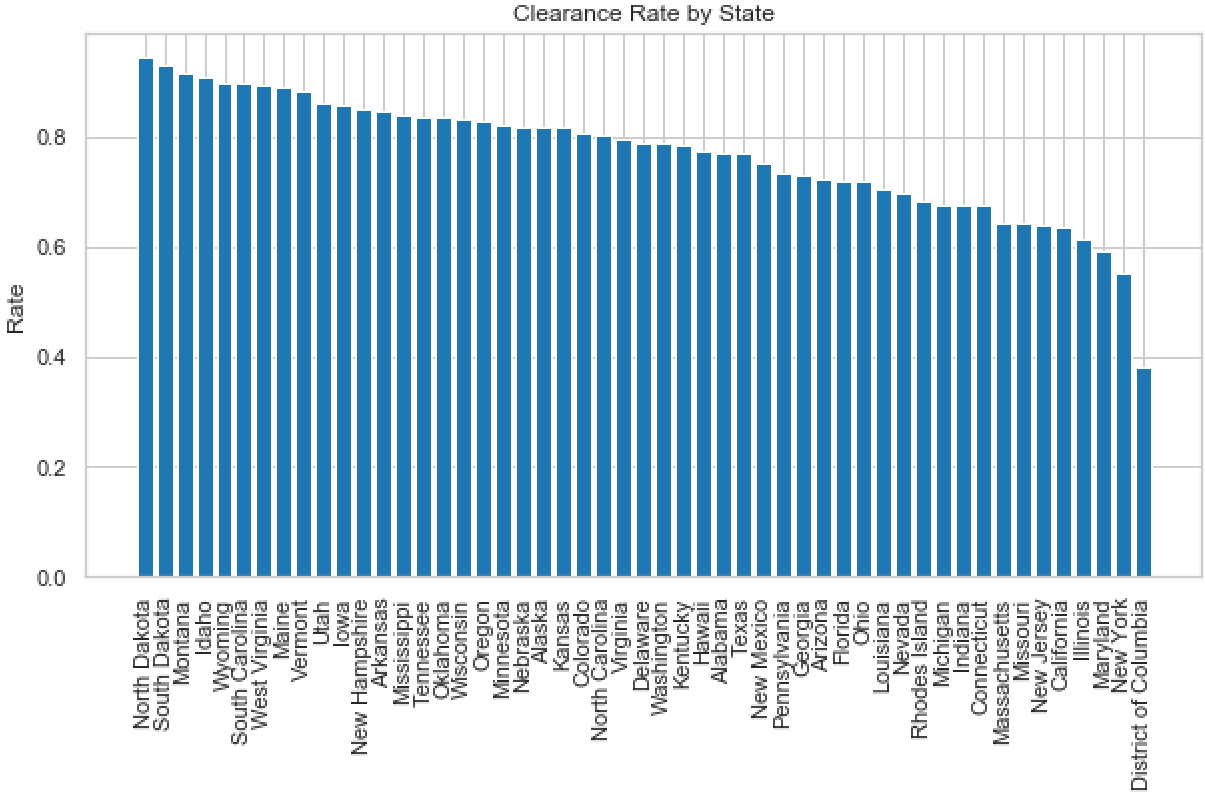
Continuing with demographics, I was surprised to find that the gender of the victim appears to have a significant effect on clearance rate.



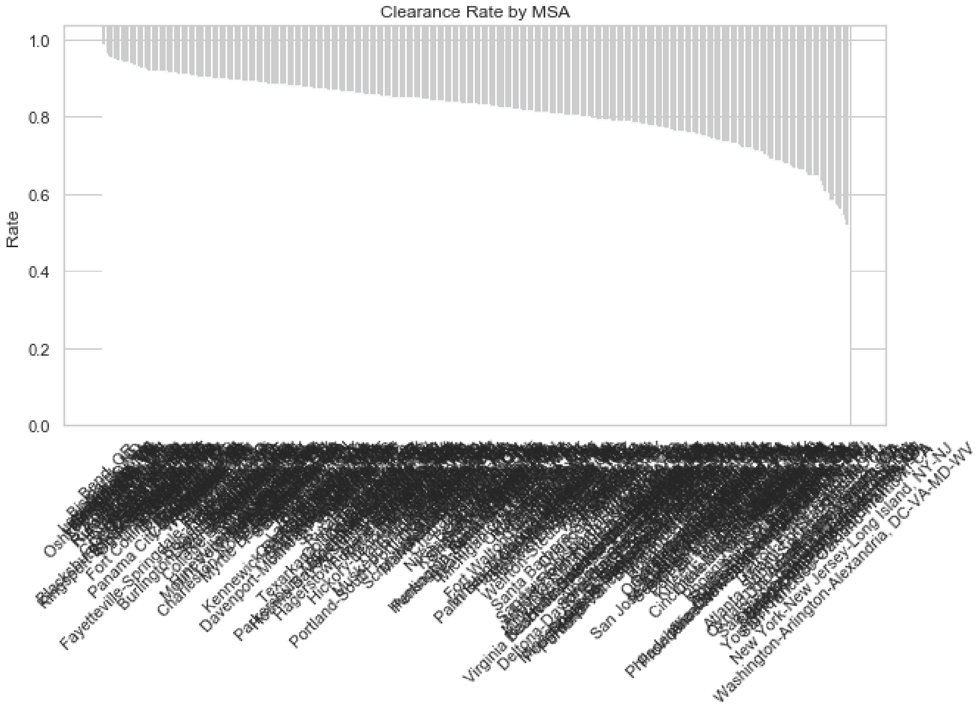
Age also has a clear effect, with the murders of young children being the most likely to be solved while teens and young adults are the least likely.

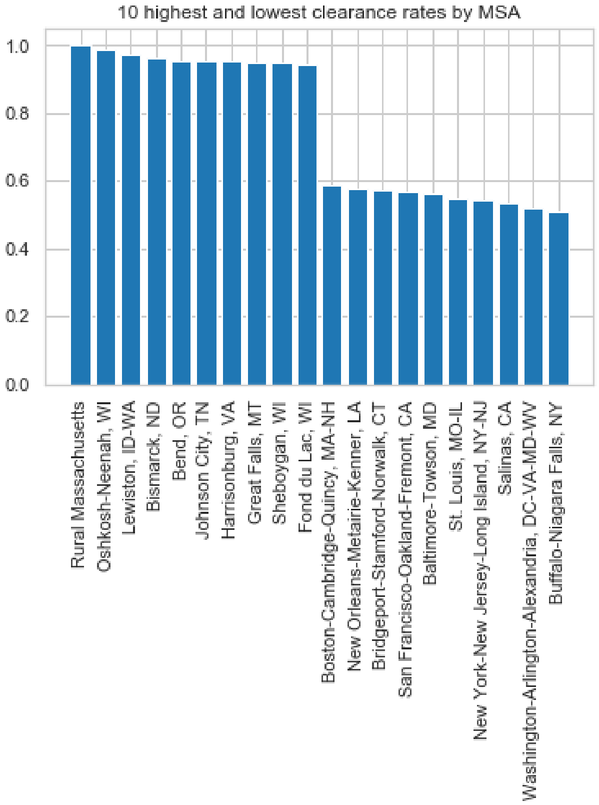


Beyond demographics, I found interesting relationships between the clearance rate and the location of the murder. There is a wide spread between the highest and lowest rates:

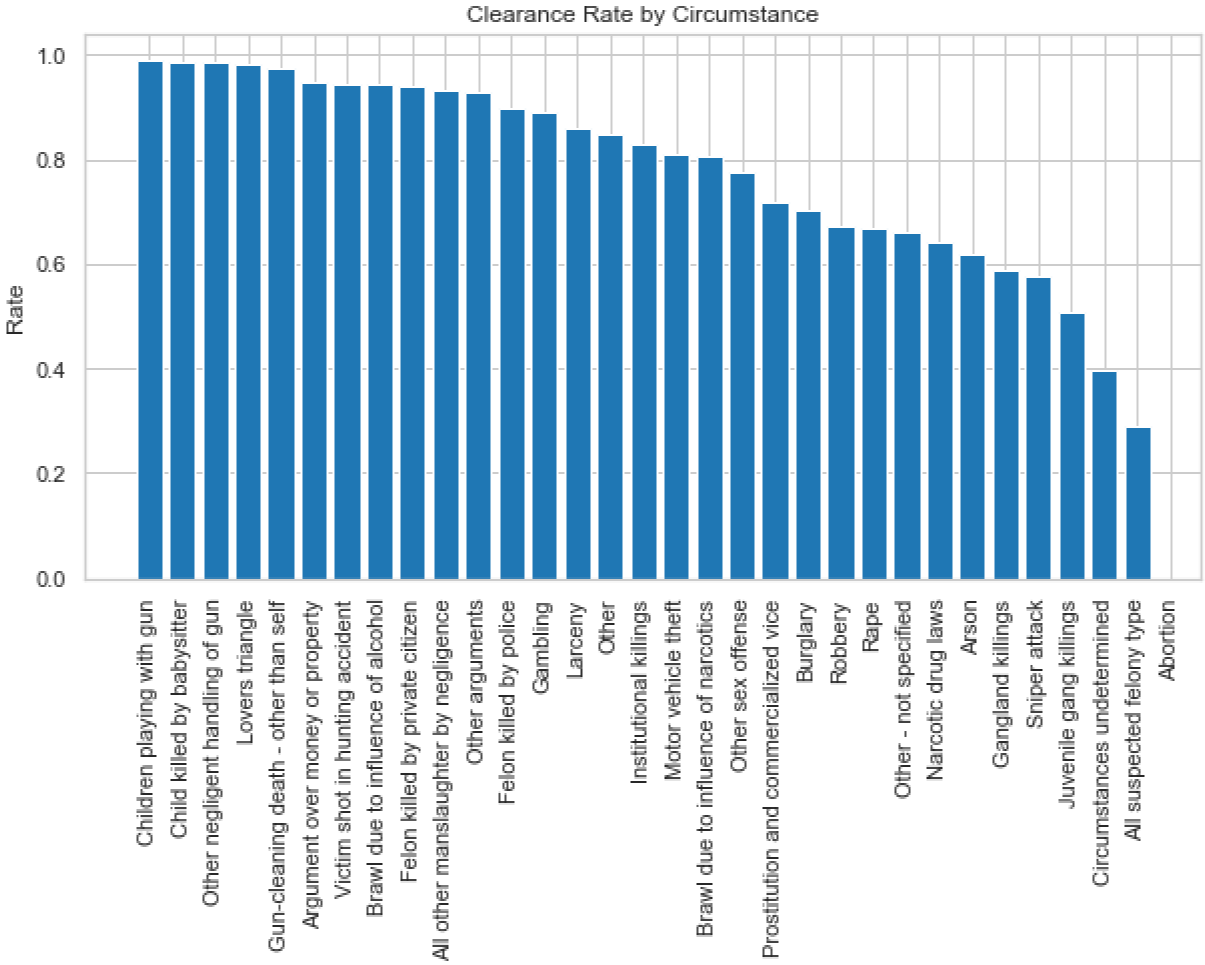


Similarly, different Metropolitan Statistical Areas (MSA) have dramatically different clearance rates. The specific names are cluttered in the below chart, but the disparity of clearance rates is apparent:

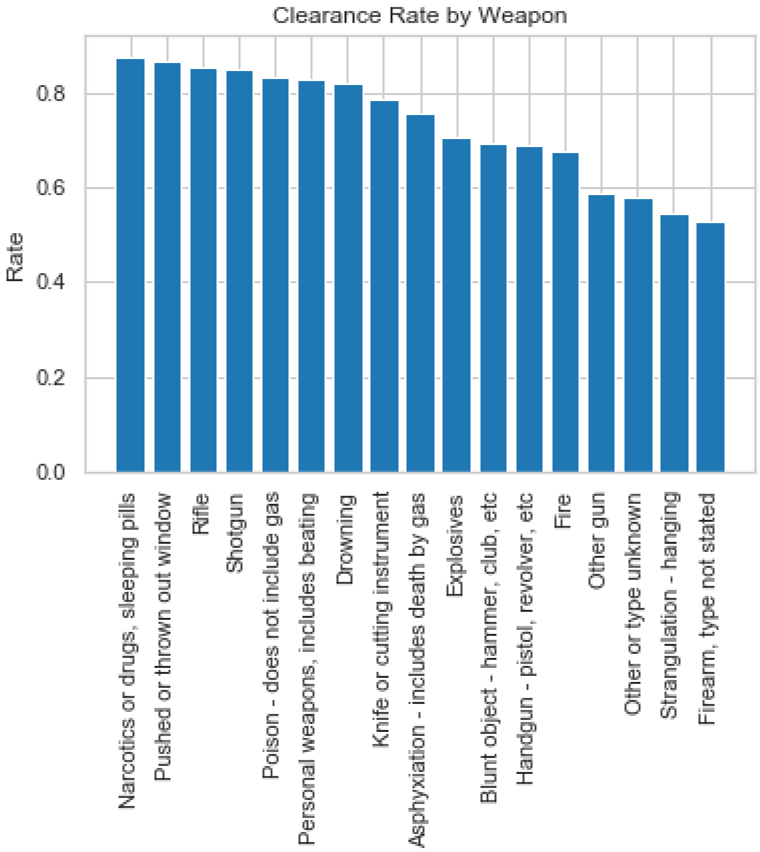




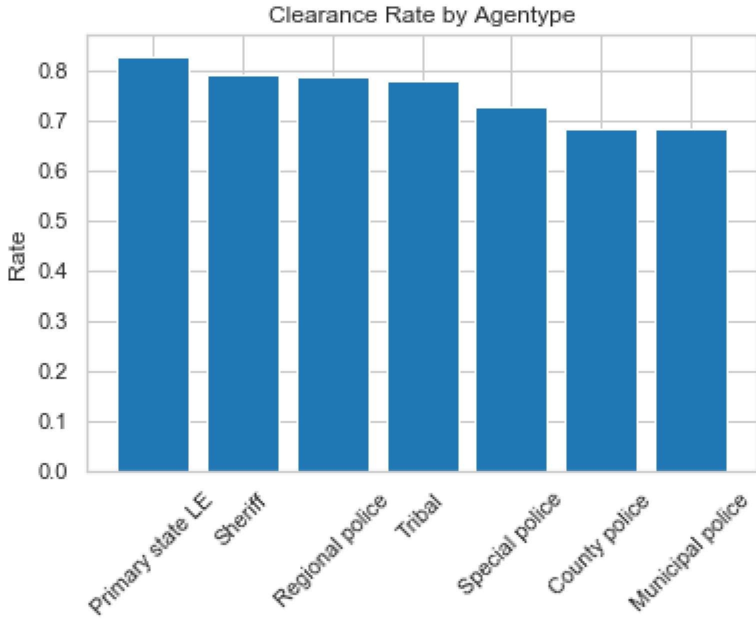
The circumstances surrounding the murder also have a major effect:



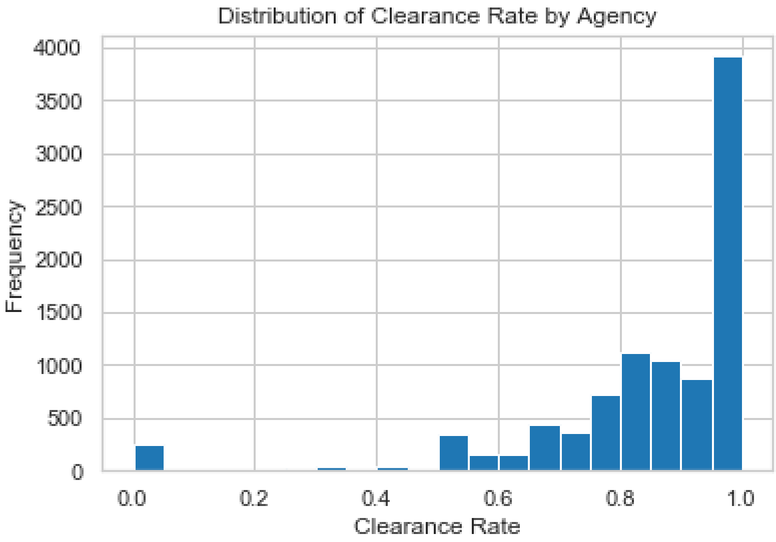
And, interestingly, the murder weapon also seems to have a large effect on the likelihood of a case being solved. For example, strangulations have around a 50% clearance rate, whereas beatings, poisonings, and drownings all have clearance rates greater than 80%.



Also of note is that the type of police agency investigating the crime seems to have a significant effect:



Similarly, the range of clearance rates among specific police agencies, of which the dataset contains 9,606 unique entries, varies widely.



Modeling Process

Algorithms

Dataset Balancing

Encoding and Scaling

Balanced and Unbalanced Test Sets

ML Scoring Metrics

Requirements

Common Metrics

Custom Metric - Binned Sum of Squared Residuals

ML Results

Logistic Regression

Naive Bayes

Random Forest

Extremely Randomized Trees

Gradient Boosting (or something else)

Comparison

Insights from Optimal Model

Characteristics of High Clearance Probability Cases

Characteristics of Low Clearance Probability Cases

Real-Life Model Use

Implementation

Recent Year Simulation

Future Improvements and Projects

New Features (Jurisdiction demographics, investigating agency funding)

More Data from Law Enforcement (new features and national uniformity)

Unknown Victim Clearance Predictor Model

Offender Demographic Predictor Model

Sources

Max Roser and Hannah Ritchie (2013) - "Homicides". *Published online at OurWorldInData.org.* Retrieved from: 'https://ourworldindata.org/homicides' [Online Resource]*

Victims of intentional homicide, 1990-2018. (2020). [Dataset]. United Nations Office on Drugs and Crime. https://dataunodc.un.org/content/data/homicide/homicide-rate

*revised 2019