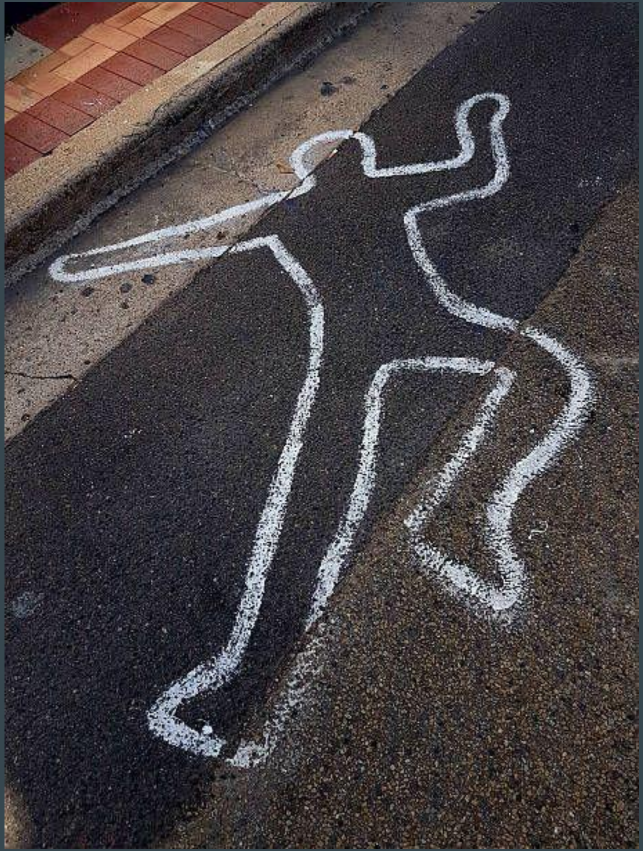
Murder
in the United States

…

Predicting Clearance Rate

with Machine Learning



Adam Nunley - 2021

More common than you think

1 in 20,000 people died from intentional homicide in 2019

5th highest cause of death for Americans 15-49

14th for all ages

94th highest homicide rate of 230 countries with UN data

Cost to Society

To some degree, we are all victims of murder:

Grief felt by loved ones

Loss of sense of security faced by anyone who knew the victim

Paranoia experienced by anyone who learns about a murder

Loss of faith in institutions when murders go unsolved

We are all stakeholders.

The objective of this project is to **predict the probability of a murder being solved** via machine learning.

Law enforcement agencies can use such a model to allocate investigative resources and **solve more murders.**

We can all use such a model to assess our law enforcement’s strengths and weaknesses.

The Data

Murder Accountability Project ©

www.murderdata.org

- Nonprofit that aggregates United States murder data
  - No federal equivalent
- SPSS and CSV formats
- Used with SQL queries to find unknown serial killers

804, 751

Murders in the US spanning 1976 - 2019

Variables

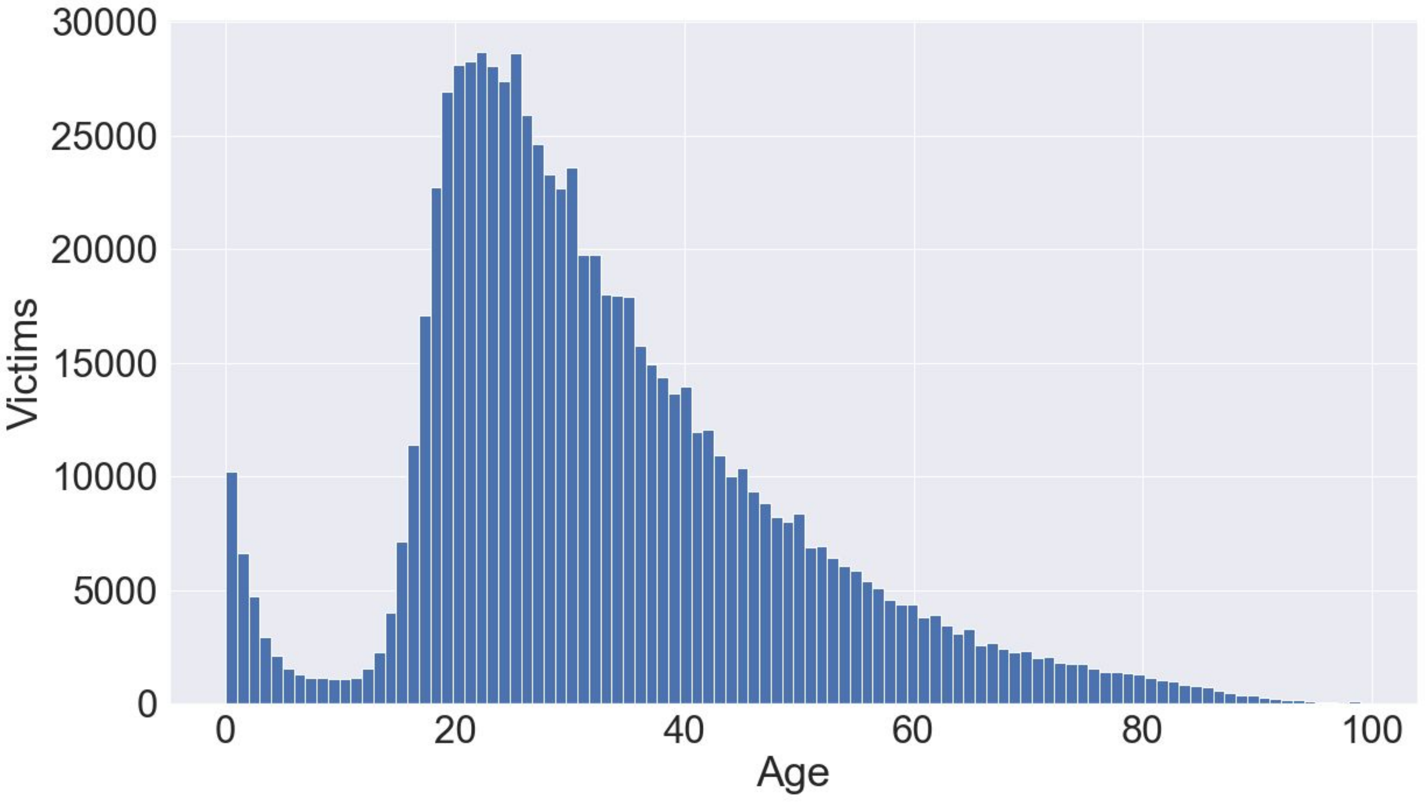
Date	
Location	
Victim	
Offender	Crime Characteristics
Investigator	Clerical

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

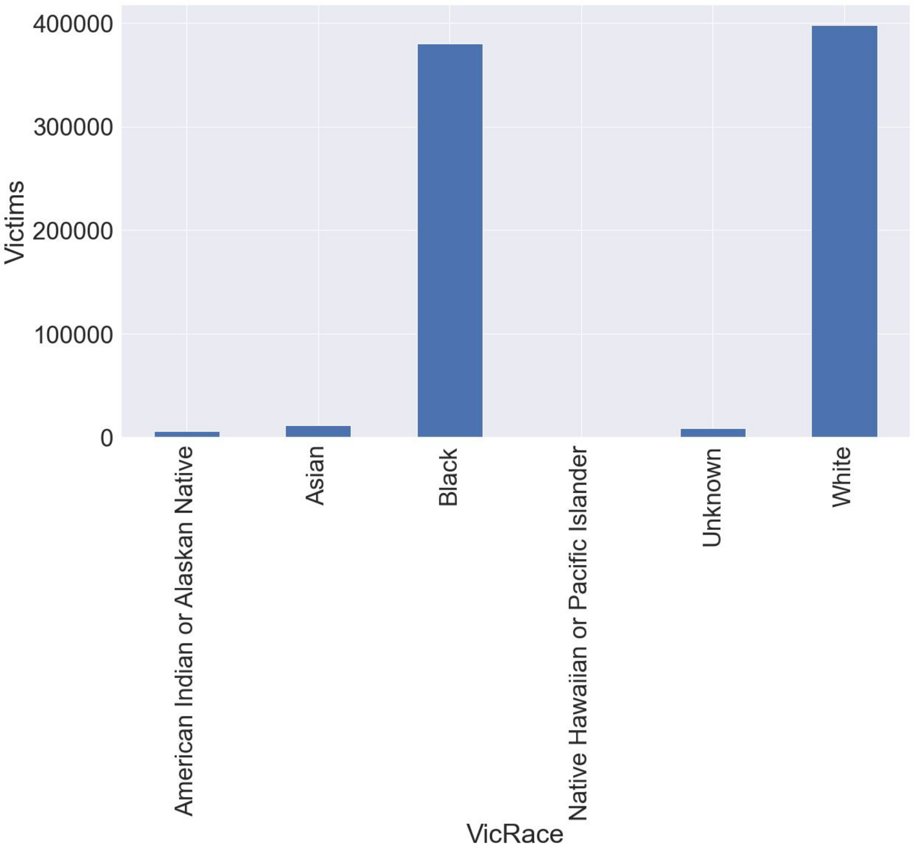
Name	Description	Type
SOLVED	**Crime Clearance Status**	**Categorical - Nominal**
HOMICIDE	Murder of Negligence Flag	Categorical - Nominal
SITUATION	Single/Multiple Victim(s)/Offender(s) Description	Categorical - Nominal
WEAPON	Murder Weapon Type	Categorical - Nominal
RELATIONSHIP	Offenders' Relationship 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

Exploring the Data

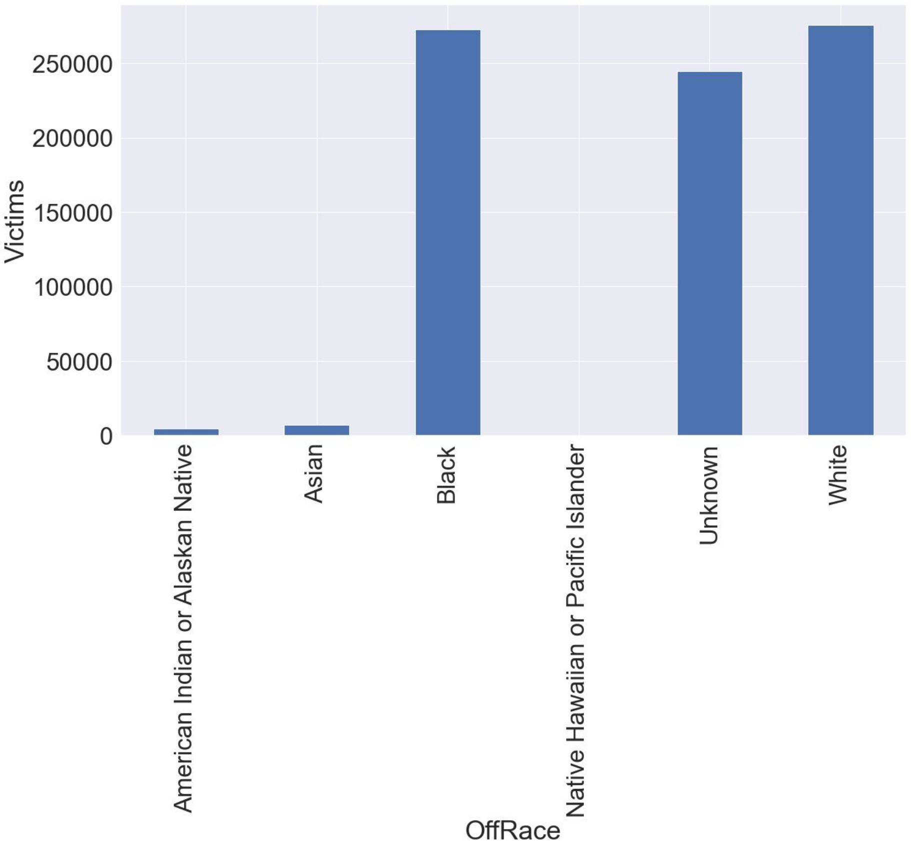
Victims by Age - Histogram



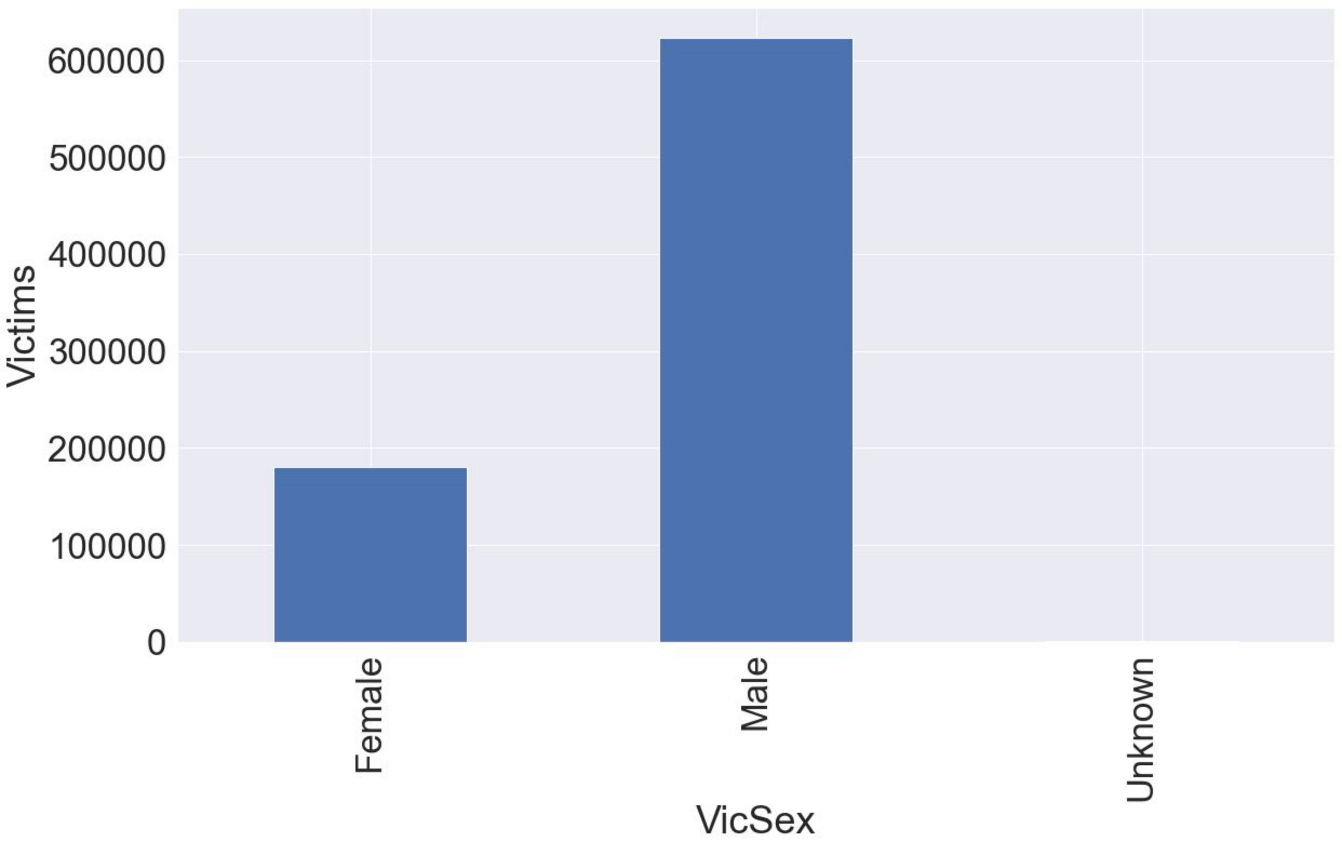
Victim Race - Bar Chart



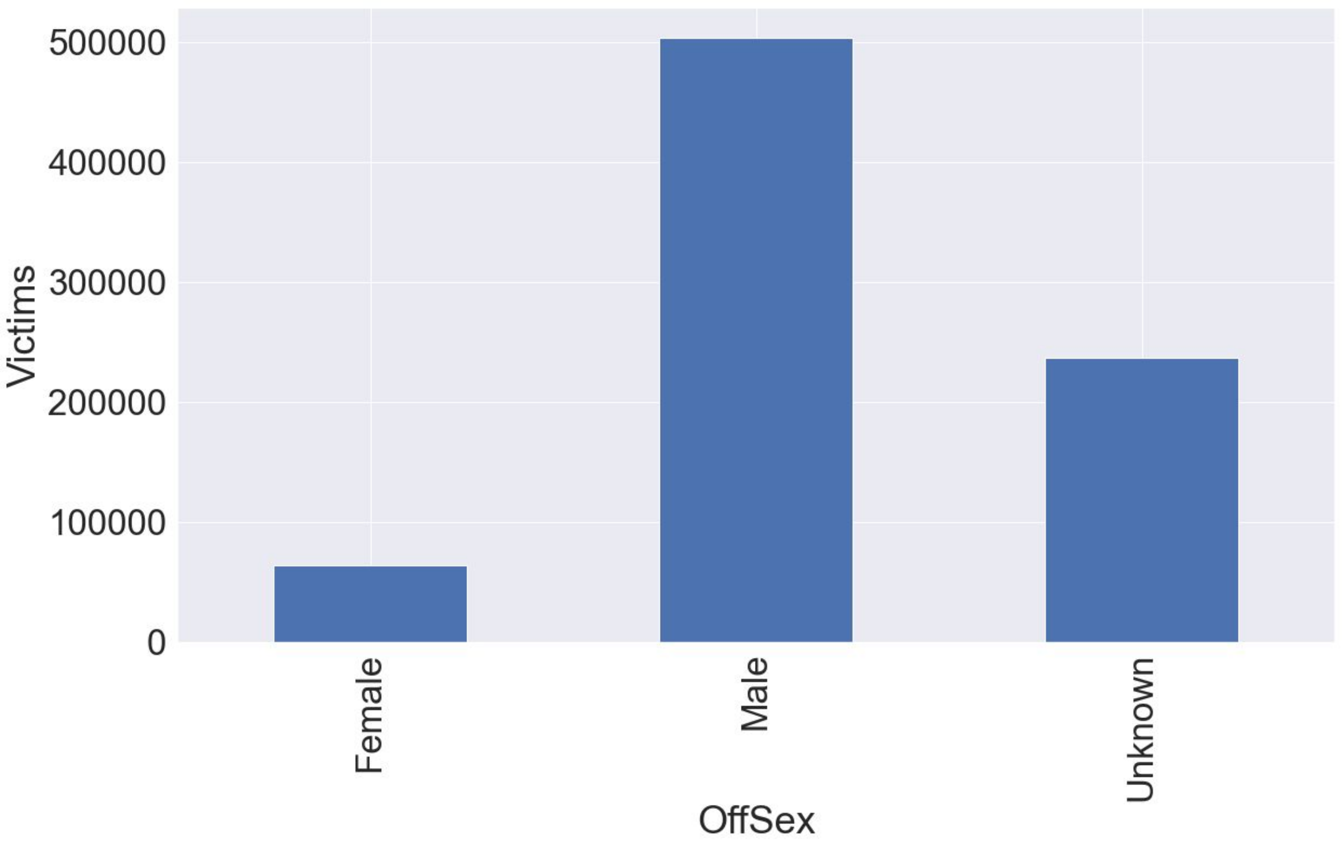
Offender Race - Bar Chart



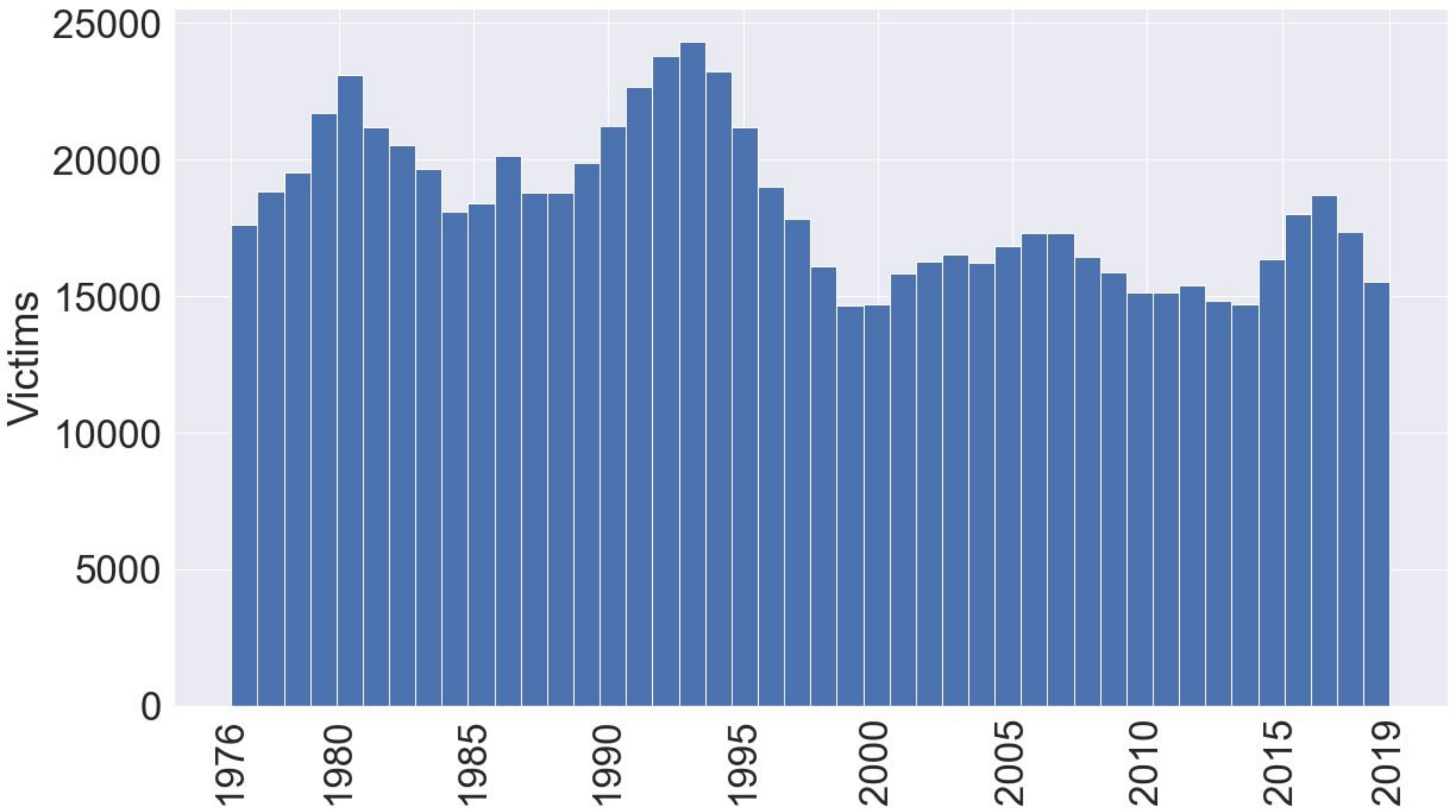
Victim Sex - Bar Chart



Offender Sex - Bar Chart

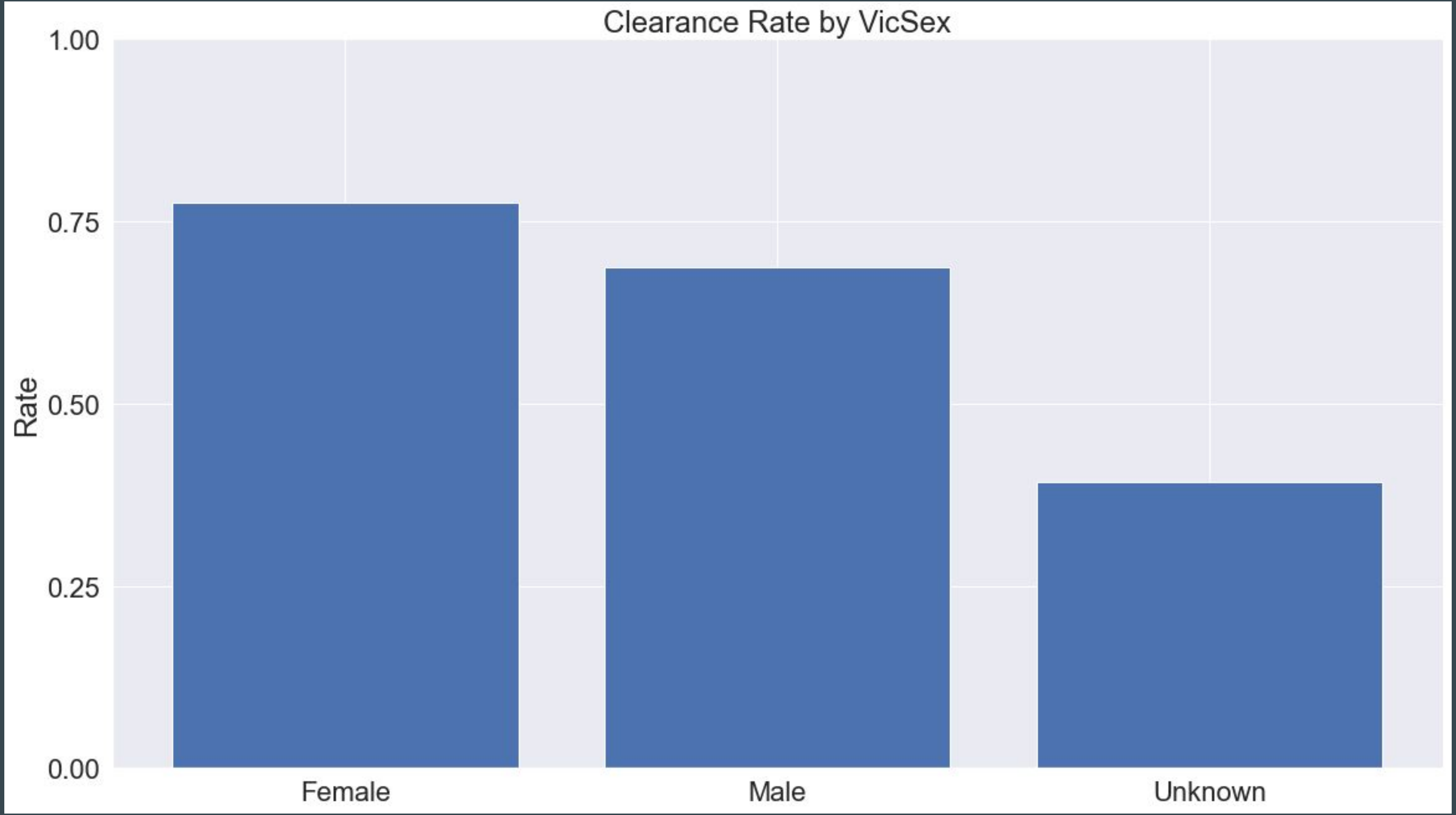


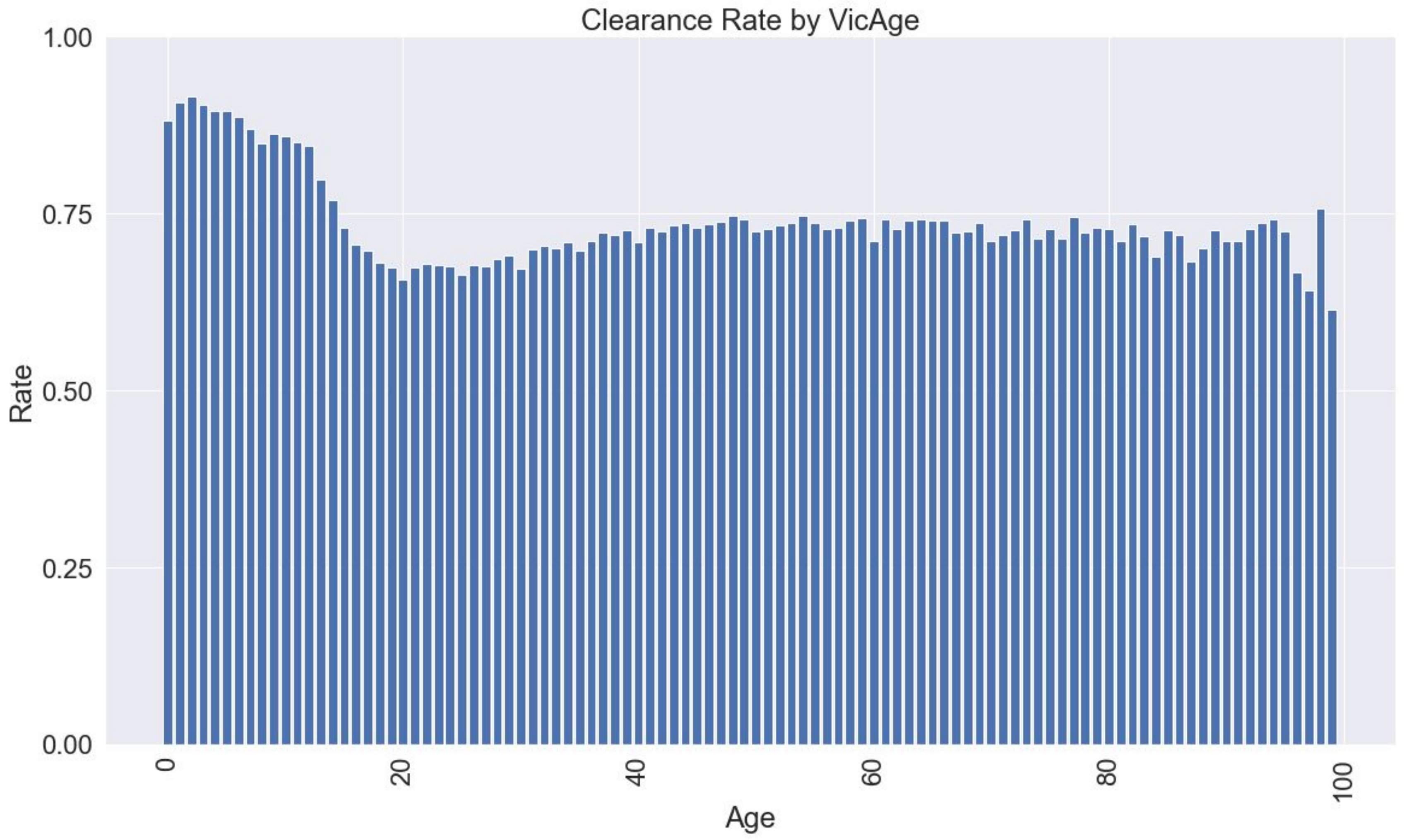
Murders By Year - Bar Chart

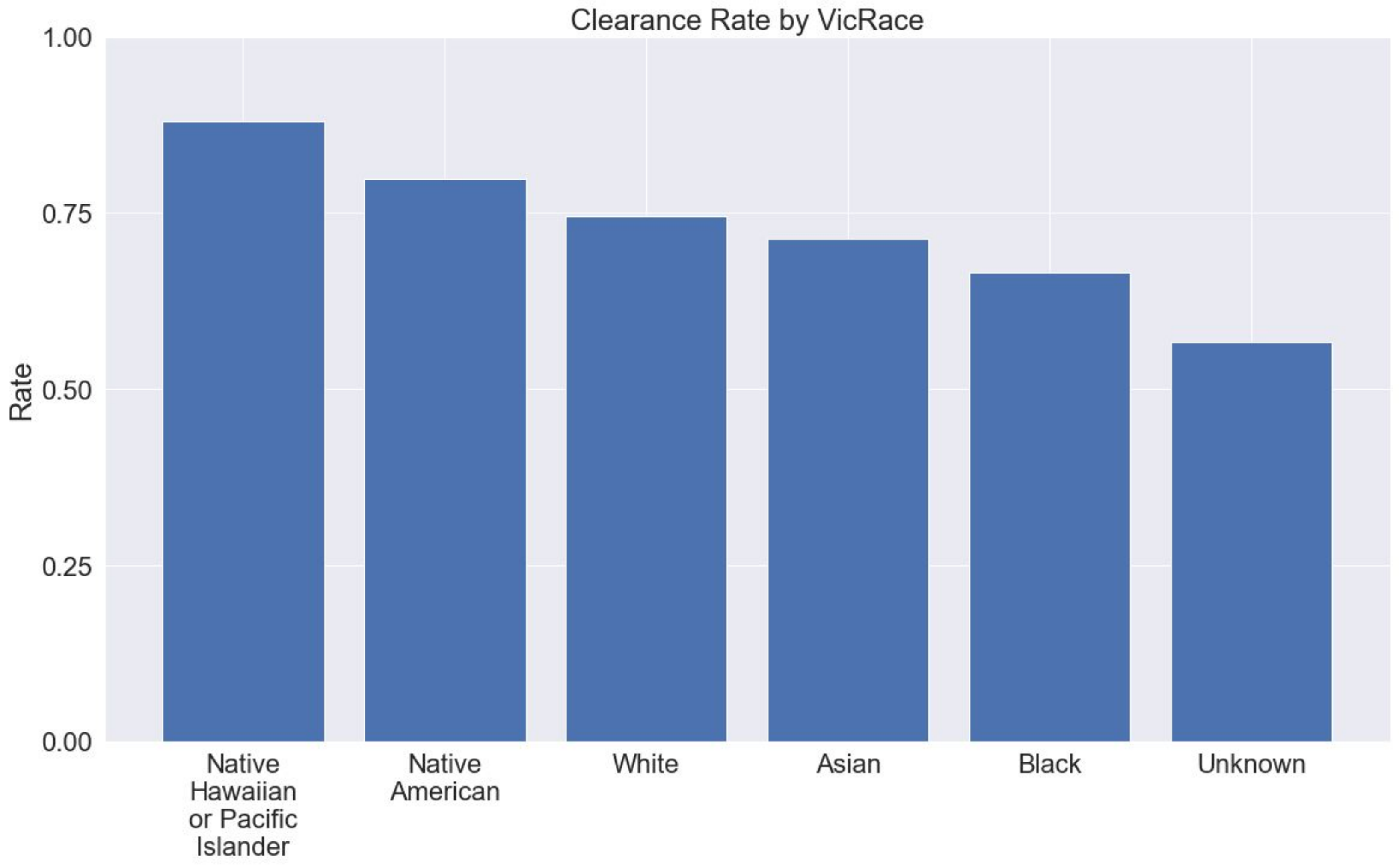


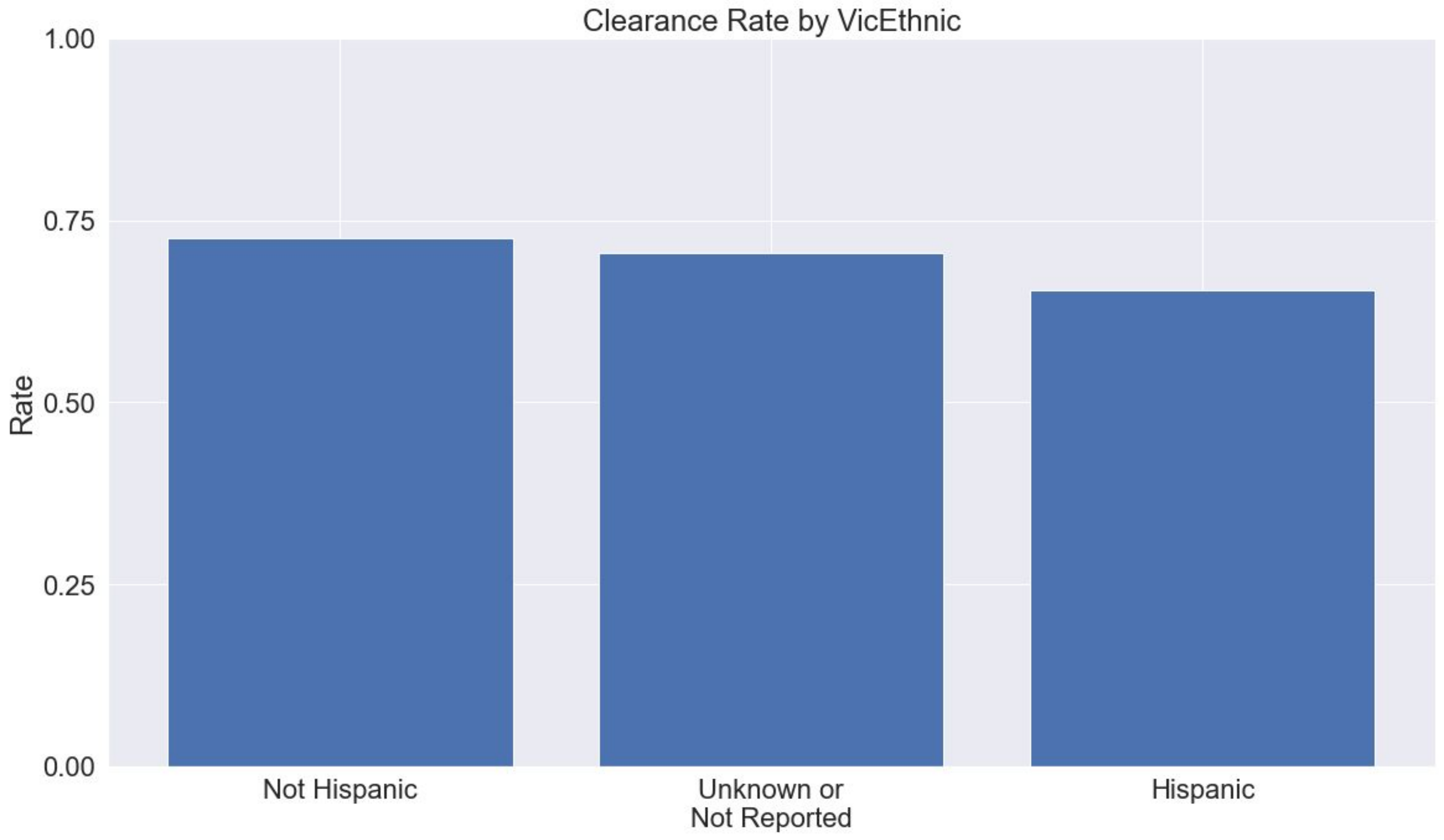
Clearance Rate

Solved murders / Total Murders

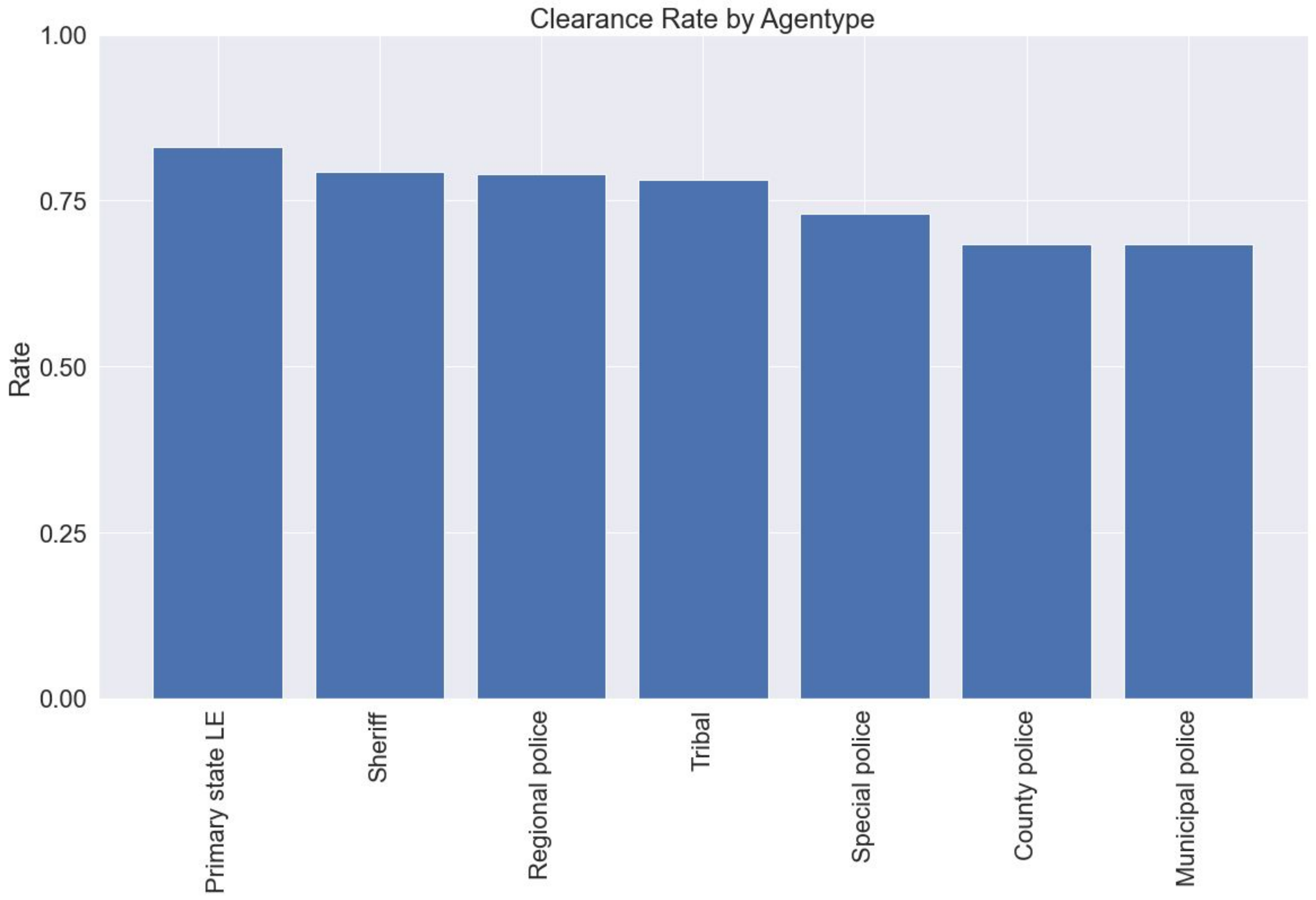

Clearance Rate by VicSex
Rate
1.00
0.75
0.50
0.25
0.00
Female
Male
Unknown

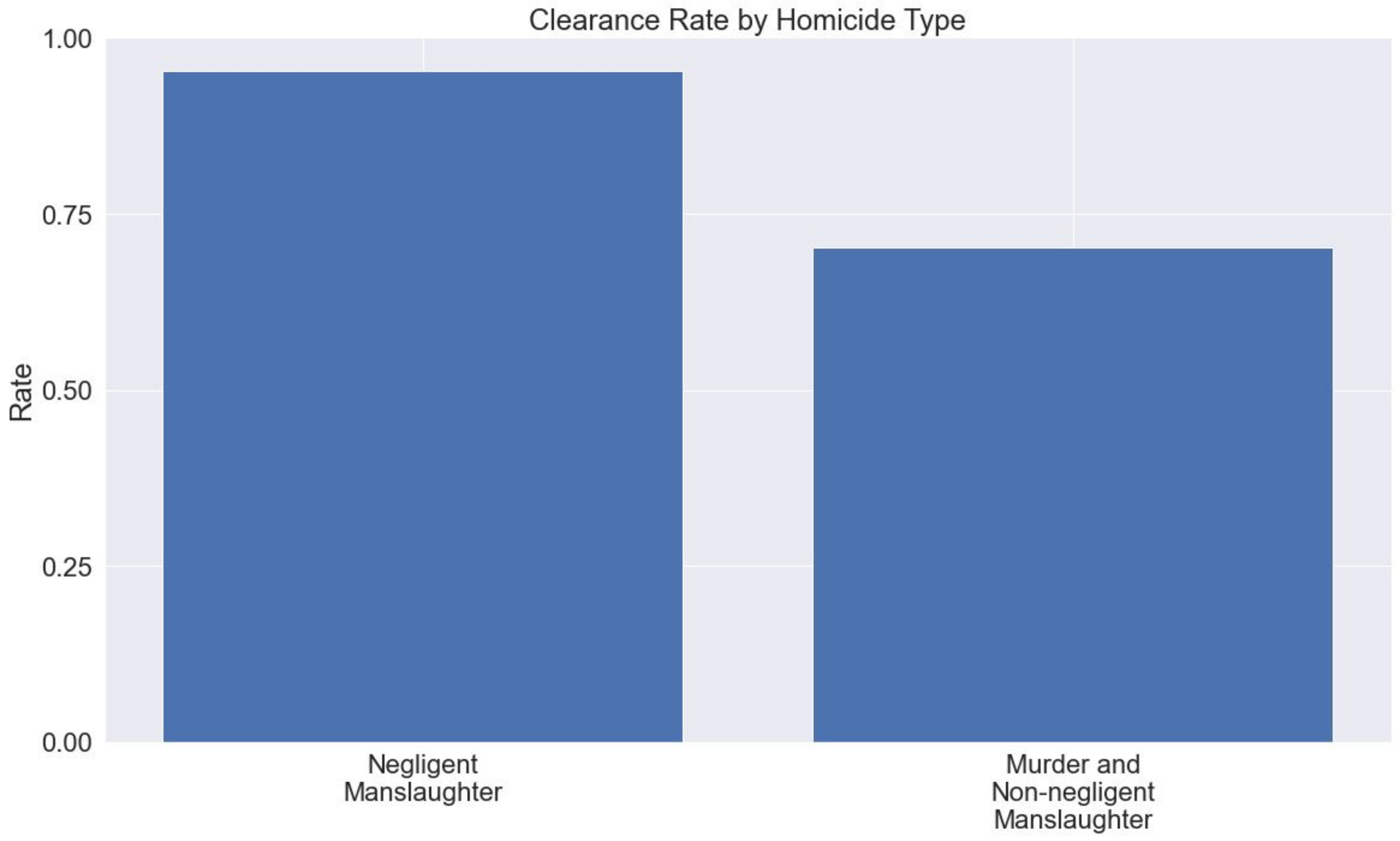

Clearance Rate by VicAge
1.00
0.75
0.50
0.25
0.00
Rate
0
20
40
60
80
100
Age

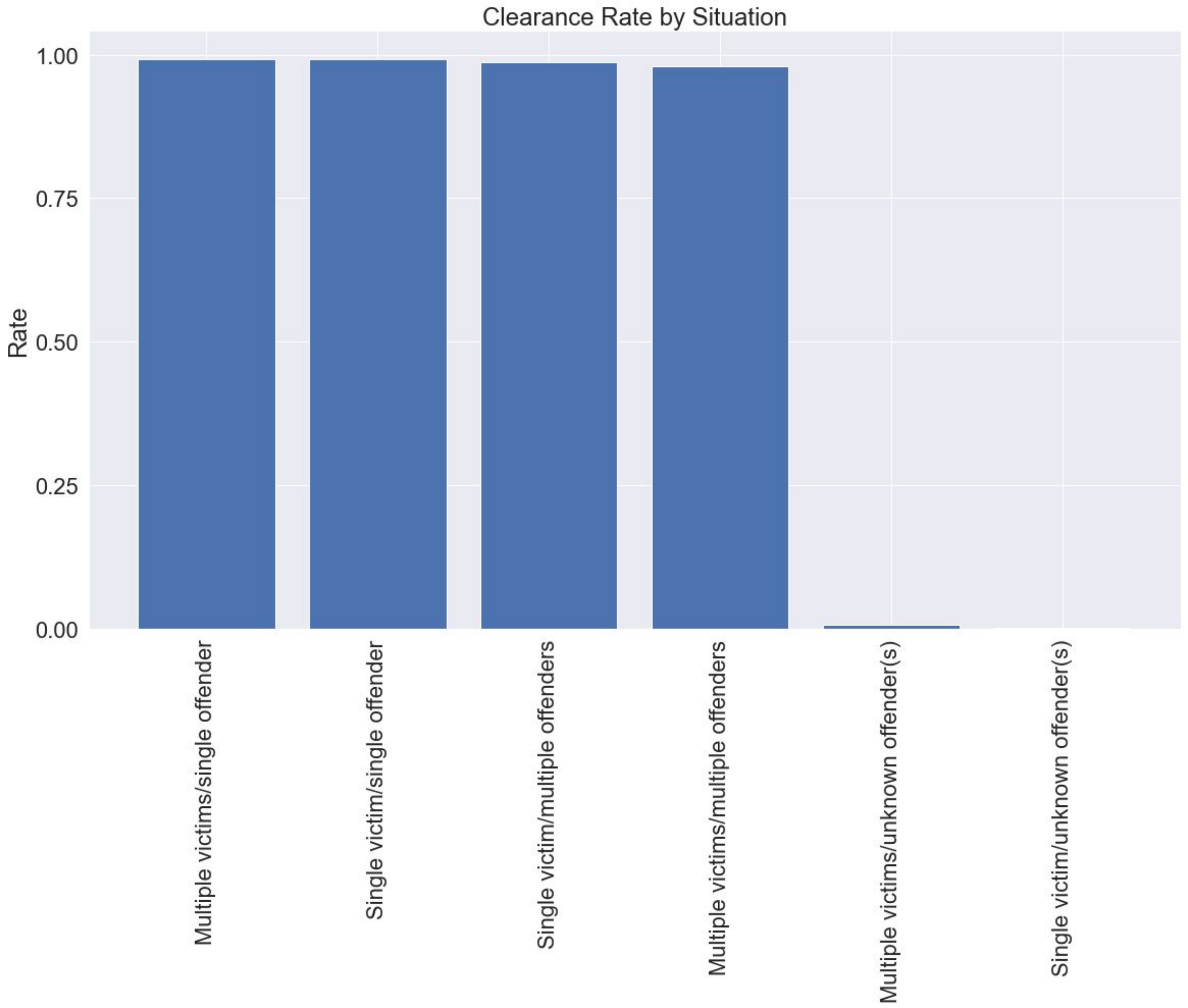

Clearance Rate by VicRace
Rate
1.00
0.75
0.50
0.25
0.00
Native Hawaiian or Pacific Islander
Native American
White
Asian
Black
Unknown

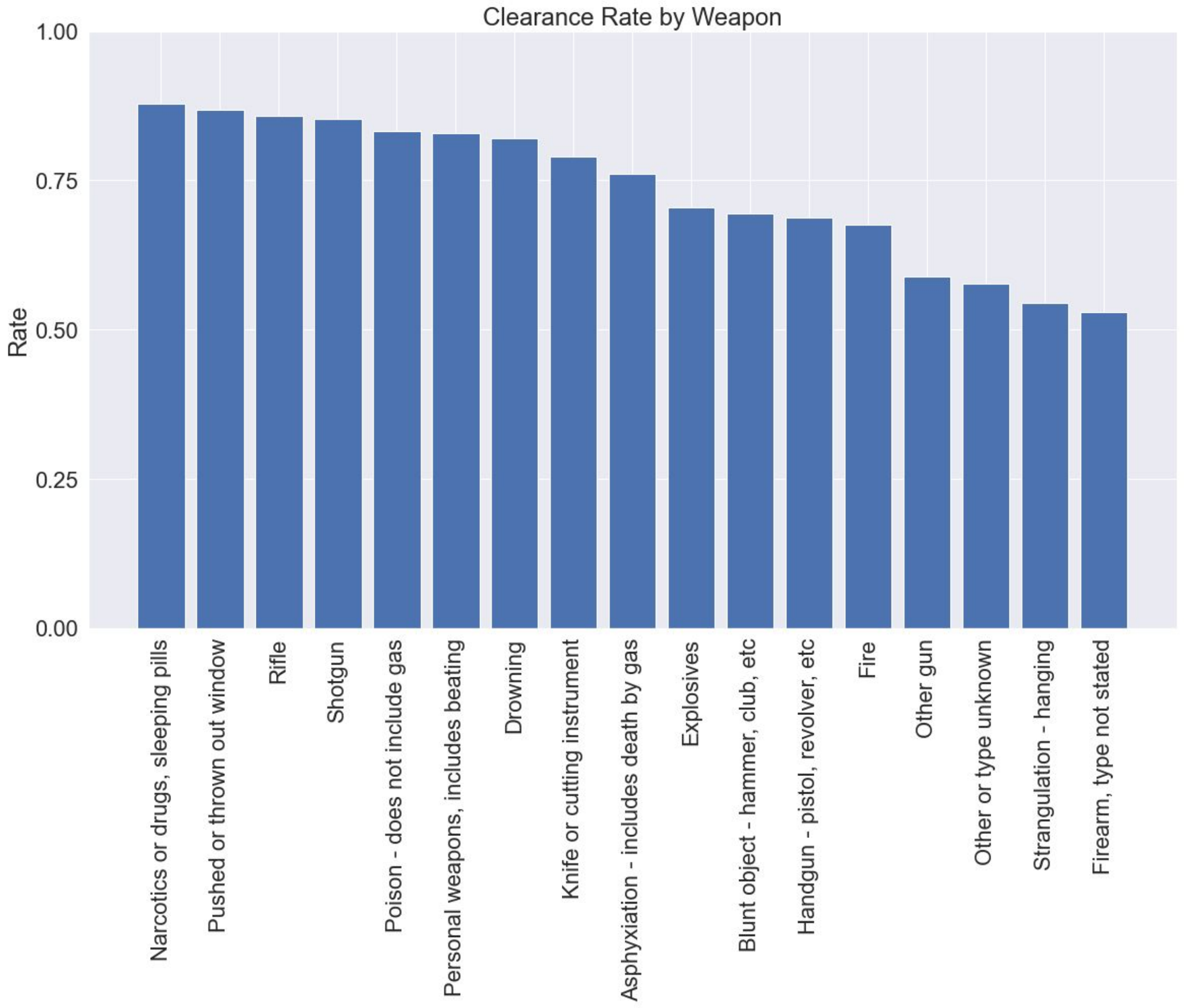


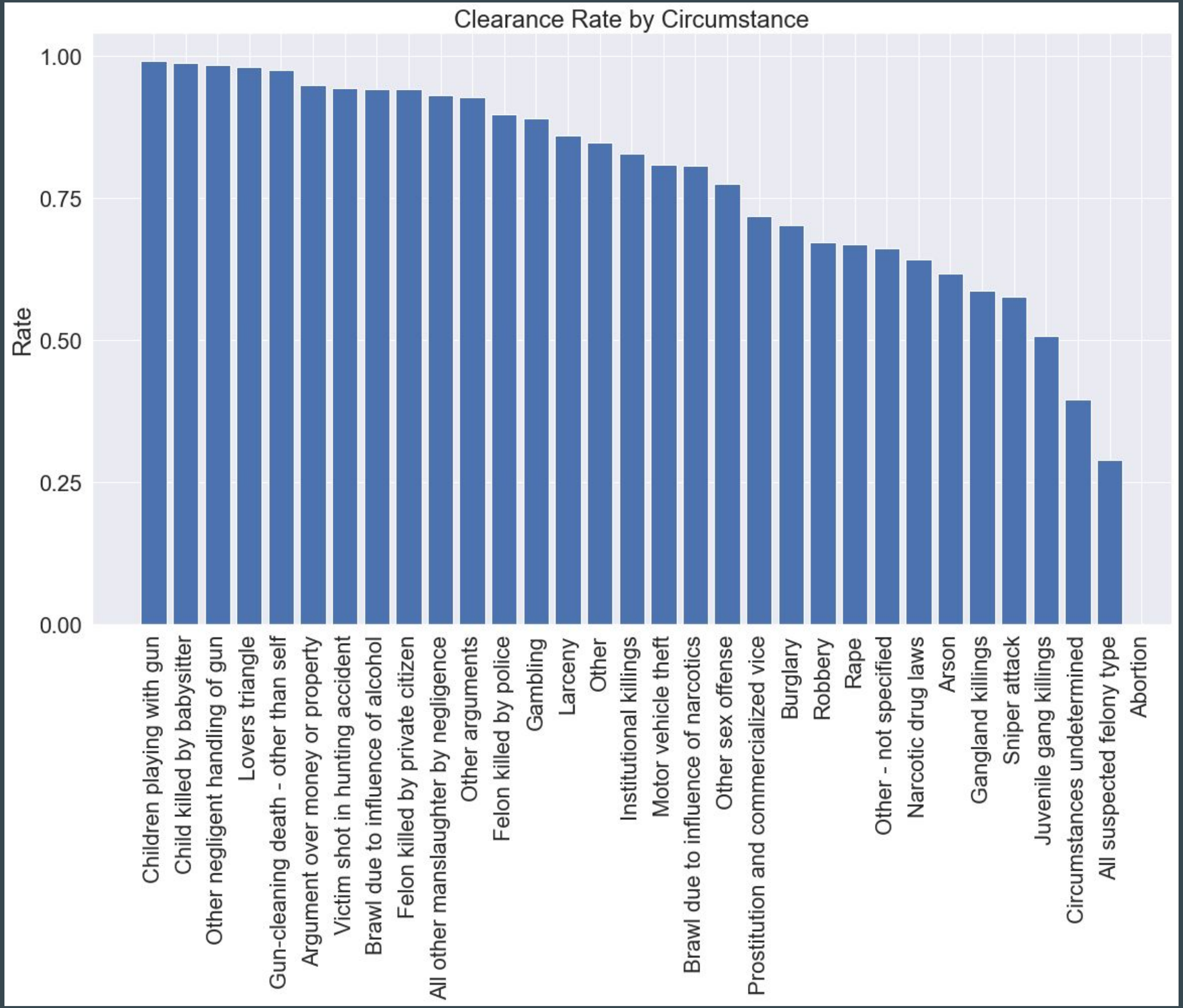
Clearance Rate by VicEthnic

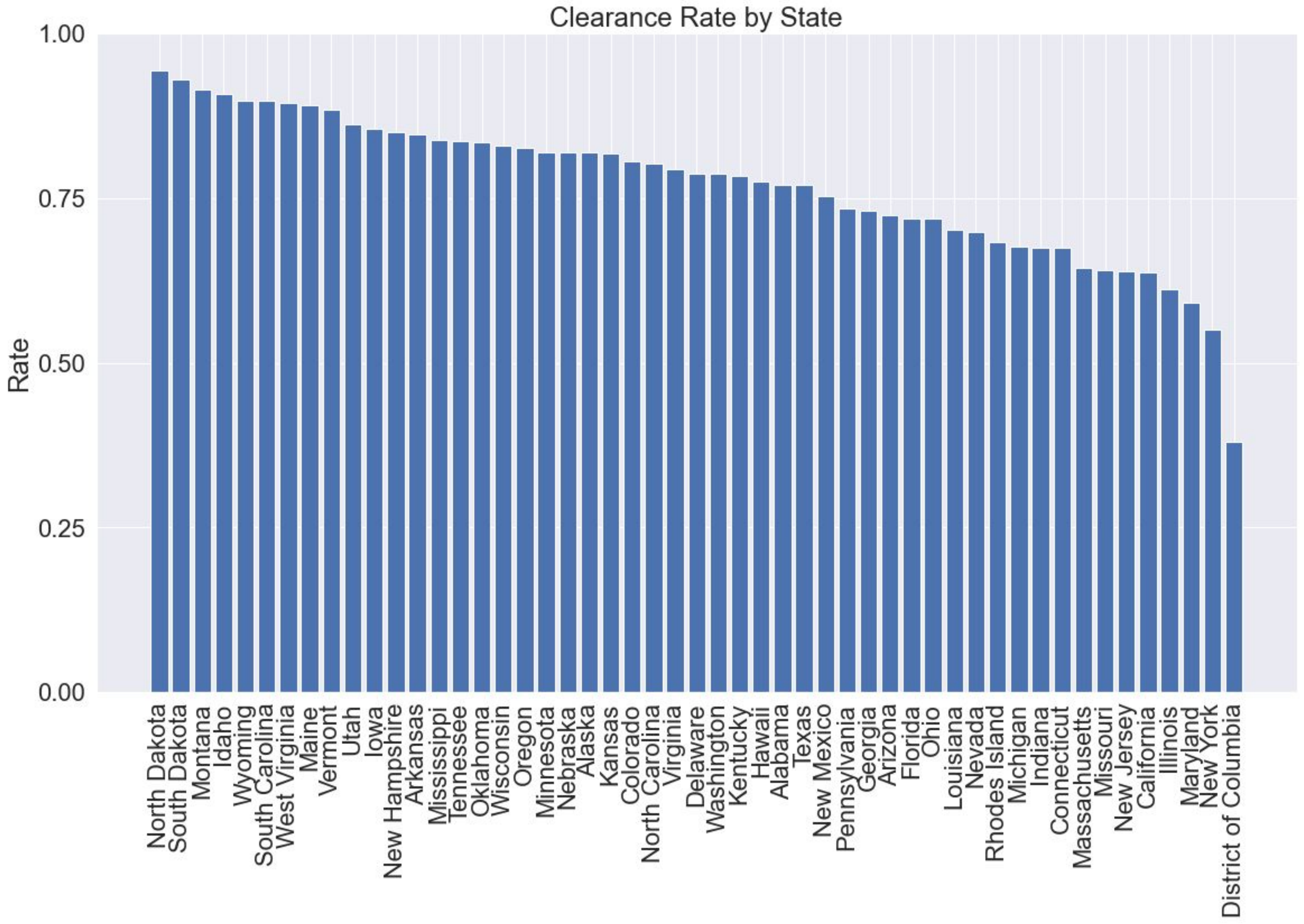

Clearance Rate by Agentype
Rate
1.00
0.75
0.50
0.25
0.00
Primary state LE
Sheriff
Regional police
Tribal
Special police
County police
Municipal police

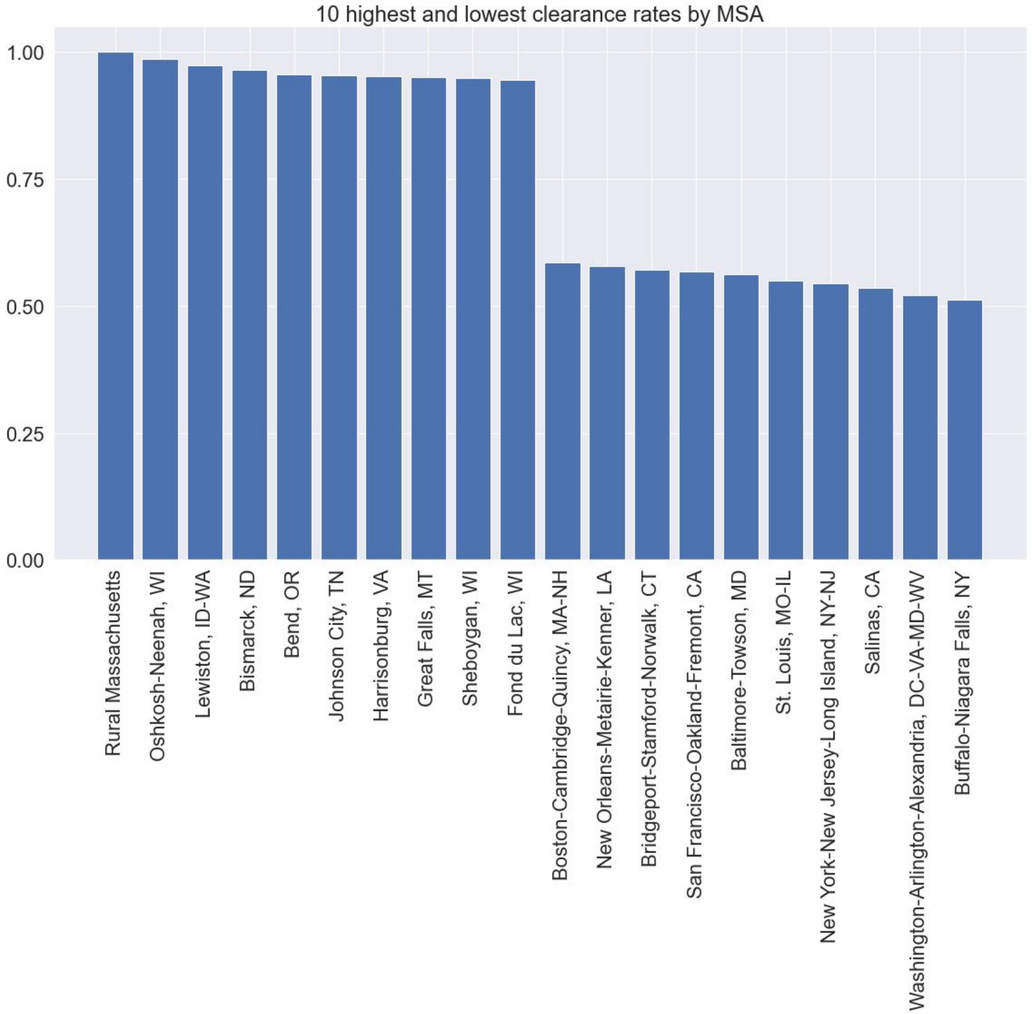

Clearance Rate by Homicide Type
Rate
1.00
0.75
0.50
0.25
0.00
Negligent Manslaughter
Murder and Non-negligent Manslaughter


Clearance Rate by Situation
Rate
1.00
0.75
0.50
0.25
0.00
Multiple victims/single offender
Single victim/single offender
Single victim/multiple offenders
Multiple victims/multiple offenders
Multiple victims/unknown offender(s)
Single victim/unknown offender(s)

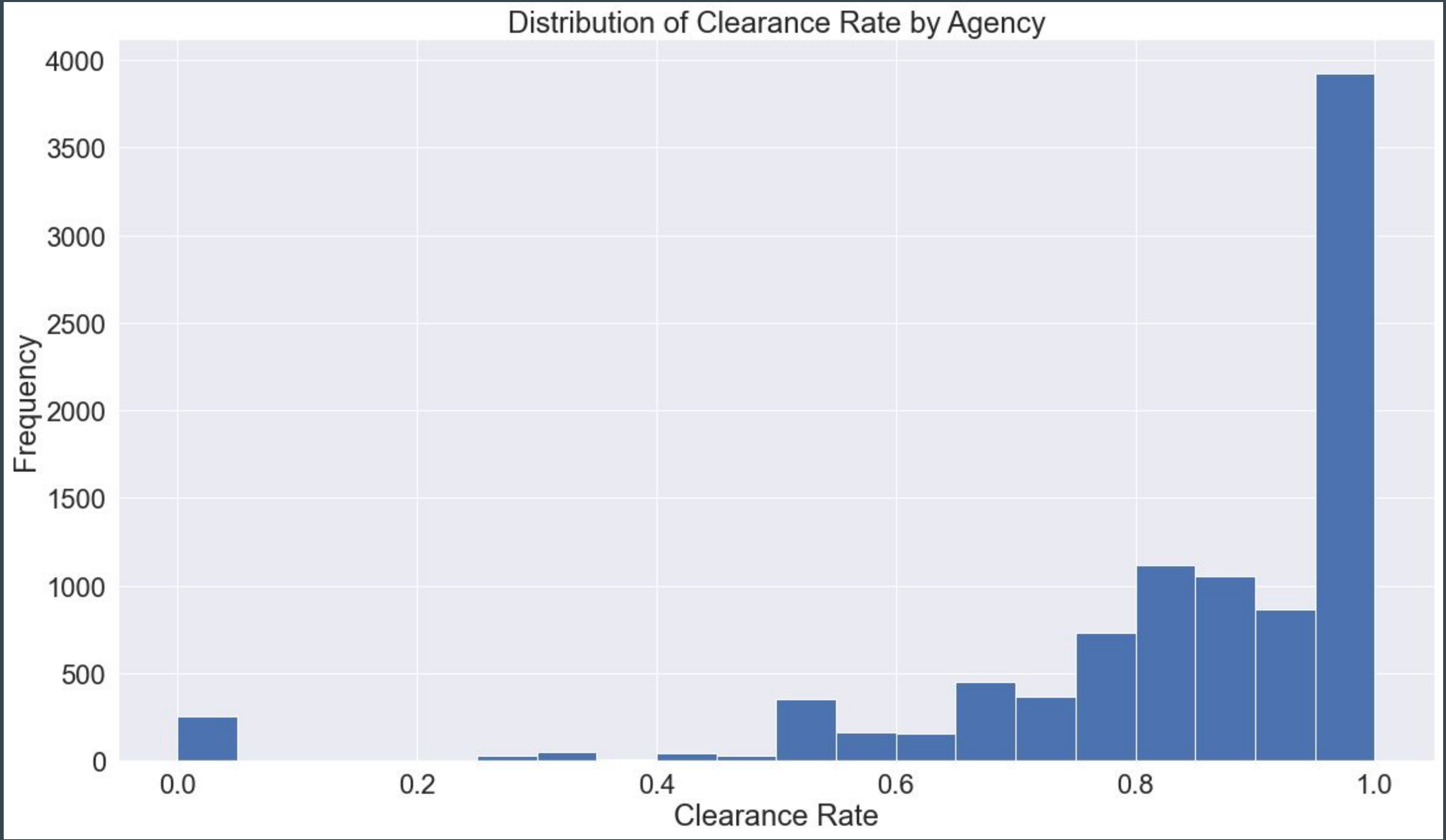

Clearance Rate by Weapon
Rate
1.00
0.75
0.50
0.25
0.00
Narcotics or drugs, sleeping pills
Pushed or thrown out window
Rifle
Shotgun
Poison - does not include gas
Personal weapons, includes beating
Drowning
Knife or cutting instrument
Asphyxiation - includes death by gas
Explosives
Blunt object - hammer, club, etc
Handgun - pistol, revolver, etc
Fire
Other gun
Other or type unknown
Strangulation - hanging
Firearm, type not stated


Clearance Rate by Circumstance
Rate
1.00
0.75
0.50
0.25
0.00
Children playing with gun
Child killed by babysitter
Other negligent handling of gun
Lovers triangle
Gun-cleaning death - other than self
Argument over money or property
Victim shot in hunting accident
Brawl due to influence of alcohol
Felon killed by private citizen
All other manslaughter by negligence
Other arguments
Felon killed by police
Gambling
Larceny
Other
Institutional killings
Motor vehicle theft
Brawl due to influence of narcotics
Other sex offense
Prostitution and commercialized vice
Burglary
Robbery
Rape
Other - not specified
Narcotic drug laws
Arson
Gangland killings
Sniper attack
Juvenile gang killings
Circumstances undetermined
All suspected felony type
Abortion


Clearance Rate by State
Rate
1.00
0.75
0.50
0.25
0.00
North Dakota
South Dakota
Montana
Idaho
Wyoming
South Carolina
West Virginia
Maine
Vermont
Utah
Iowa
New Hampshire
Arkansas
Mississippi
Tennessee
Oklahoma
Wisconsin
Oregon
Minnesota
Nebraska
Alaska
Kansas
Colorado
North Carolina
Virginia
Delaware
Washington
Kentucky
Hawaii
Alabama
Texas
New Mexico
Pennsylvania
Georgia
Arizona
Florida
Ohio
Louisiana
Nevada
Rhodes Island
Michigan
Indiana
Connecticut
Massachusetts
Missouri
New Jersey
California
Illinois
Maryland
New York
District of Columbia



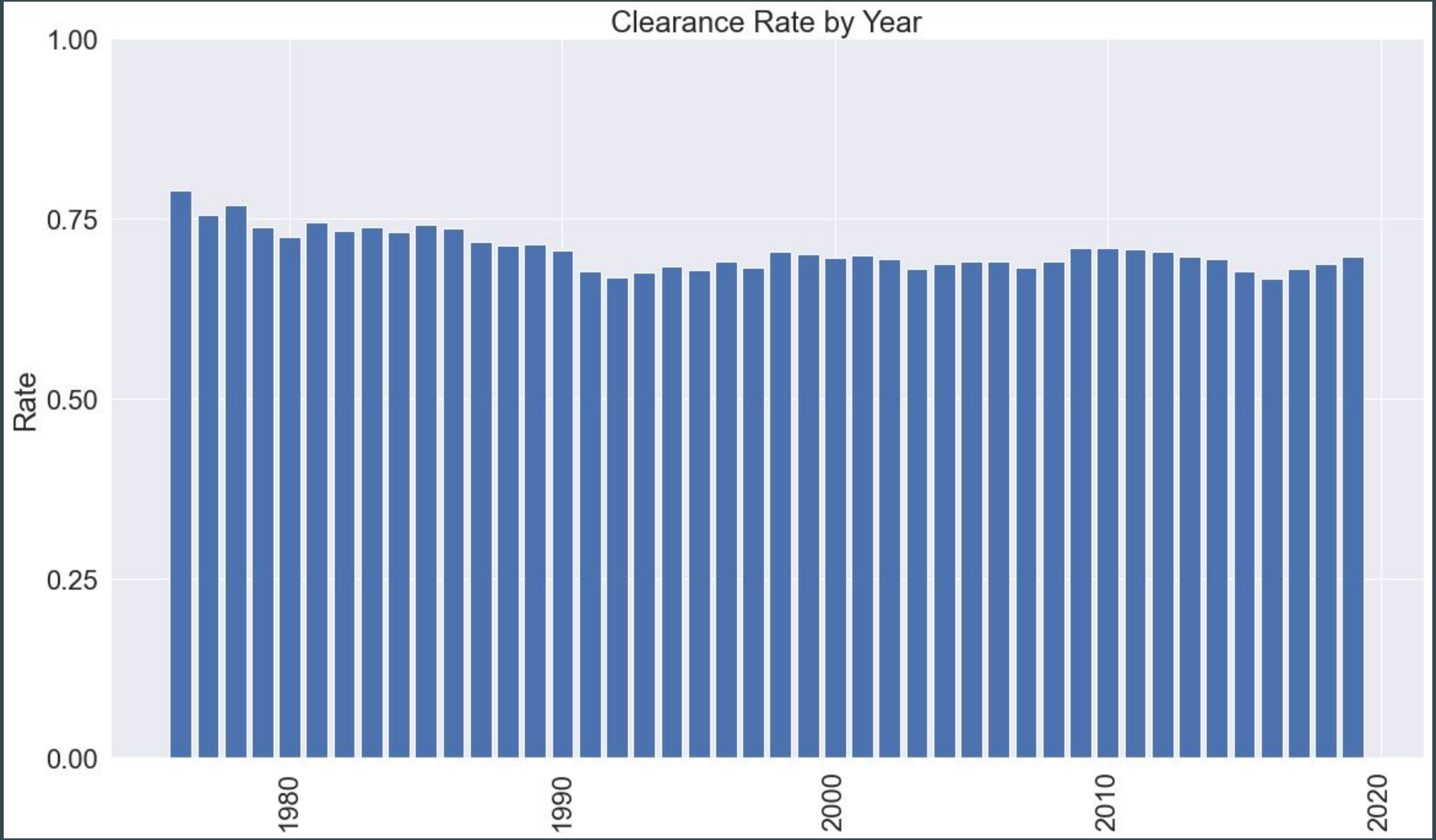
10 highest and lowest clearance rates by MSA
1.00
0.75
0.50
0.25
0.00
Rural Massachusetts
Oshkosh-Neenah, WI
Lewiston, ID-WA
Bismarck, ND
Bend, OR
Johnson City, TN
Harrisonburg, VA
Great Falls, MT
Sheboygan, WI
Fond du Lac, WI
Boston-Cambridge-Quincy, MA-NH
New Orleans-Metairie-Kenner, LA
Bridgeport-Stamford-Norwalk, CT
San Francisco-Oakland-Fremont, CA
Baltimore-Towson, MD
St. Louis, MO-IL
New York-New Jersey-Long Island, NY-NJ
Salinas, CA
Washington-Arlington-Alexandria, DC-VA-MD-WV
Buffalo-Niagara Falls, NY


Distribution of Clearance Rate by Agency
Frequency
4000
3500
3000
2500
2000
1500
1000
500
0
0.0
0.2
0.4
0.6
0.8
1.0
Clearance Rate

30%

Of murders in the United States go unsolved, and this rate has worsened since the 1970’s.

Victims, their families, and all of society deserve a better outcome.


Clearance Rate by Year
Rate
1.00
0.75
0.50
0.25
0.00
1980
1990
2000
2010
2020

Unknown Data Entries

Changing ages (999) to average age

Leaving ‘Unknown’ as a viable category for victim demographic data

Data Engineering

New Feature : White Victim Percent

New Feature : Total Agency Cases

Removing data leakage from ‘SITUATION’

Chosen Features for Modeling

Name	Description	Type
YEAR	Year of Murder	Numerical - Discrete
MONTH	Month of Murder	Numerical - Discrete
VICAGE	Victim Age	Numerical - Discrete
VICSEX	Victim Sex	Categorical - Nominal
VICRACE	Victim Race	Categorical - Nominal
VICETHNIC	Victim Hispanic Identification	Categorical - Nominal
AGENTYPE	Investigating Agency Type	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
VICCOUNT	Number of Victims in Entire Related Incident	Numerical - Discrete
WhiteVictimPercent	(White Victims)/(Total Victims) for Cases Handled by Investigating Agency	Numerical - Continuous
AgencyCases	Total Murder Cases Handled by Investigating Agency	Numerical - Discrete
SOLVED	**Crime Clearance Status**	**Categorical - Nominal**

Date

Victim

Investigator

Crime Characteristics

Engineered

Algorithm Requirements

Classifiers

Must Predict Probabilities (no SVM)

Must Handle Large Dataset (no KNN)

Algorithm Choices

Logarithmic Regression

Naive Bayes Classifier

Random Forest

Extremely Randomized Trees

Model Scoring

Metrics must account for entire range of probabilities

ROC AUC, Log Loss

New Metric: ‘Binned Sum of Squared Residuals’

Grid Searches

Logistic Regression

C

Random Forest, Extremely Randomized Trees

n_estimators

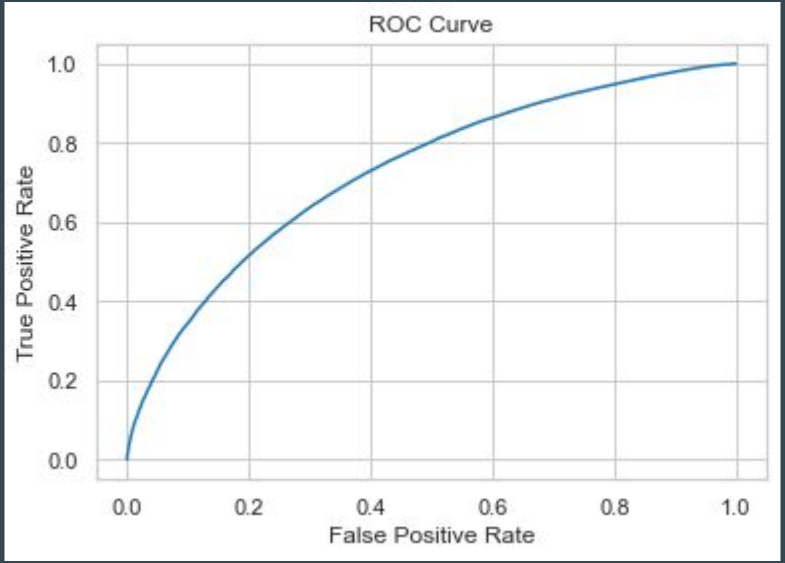
max_depth

min_samples_split

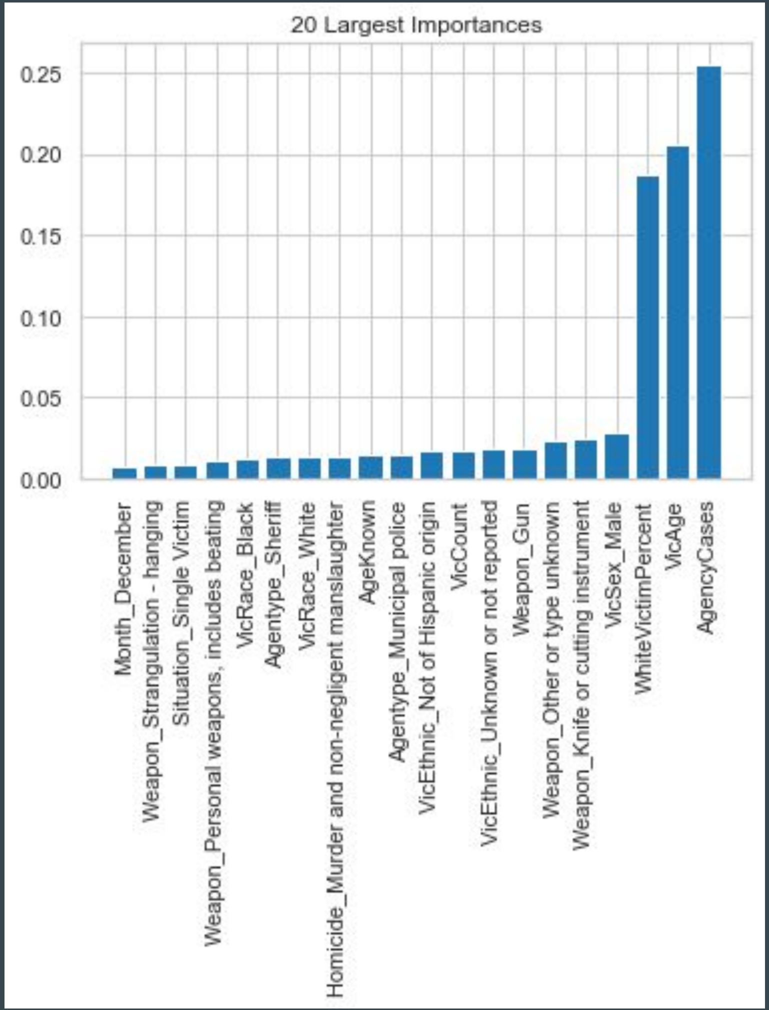
Final Grid Search Output

	BSSR Mean	ROC_AUC Mean	Precision Mean	Recall Mean	Accuracy Mean	F1 Mean
RandomForest balanced	-96.05	0.7185	0.6614	0.6607	0.6607	0.6603
ExtraTrees balanced	-164.29	0.7057	0.6492	0.6476	0.6476	0.6467
RandomForest unbalanced	-172.33	0.7247	0.7024	0.7285	0.7285	0.6832
ExtraTrees unbalanced	-178.45	0.7083	0.6974	0.7229	0.7229	0.6586
LogReg balanced	-255.99	0.6642	0.6226	0.6220	0.6220	0.6215
LogReg unbalanced	-695.70	0.6648	0.6646	0.7093	0.7093	0.6233
RandomForest model balanced	-3391.83	0.7201	0.7126	0.6665	0.6665	0.6796
ExtraTrees model balanced	-4195.71	0.7112	0.7077	0.6474	0.6474	0.6624
LogReg model balanced	-6304.65	0.6650	0.6857	0.6087	0.6087	0.6260
NaiveBayes balanced	-7516.91	0.6454	0.6187	0.5549	0.5549	0.4857
NaiveBayes unbalanced	-14680.93	0.6453	0.6790	0.5320	0.5320	0.5451

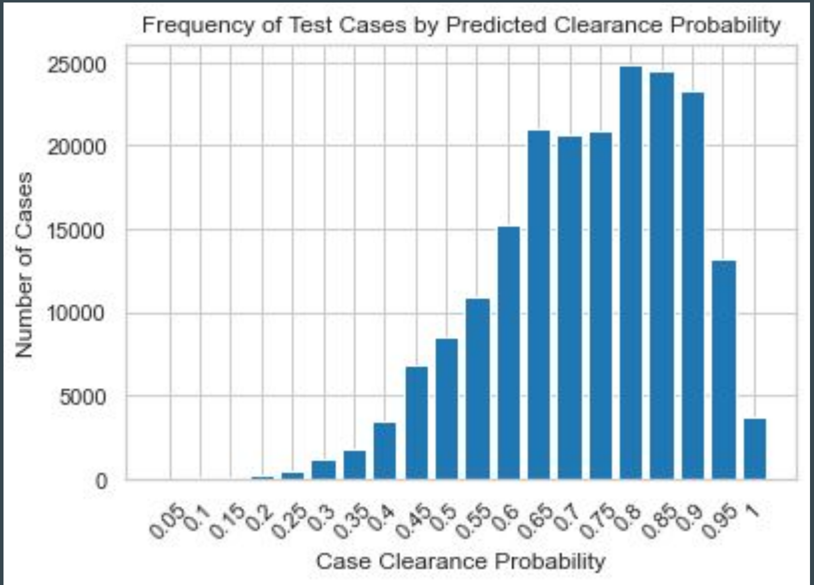
Optimal Model ROC Curve



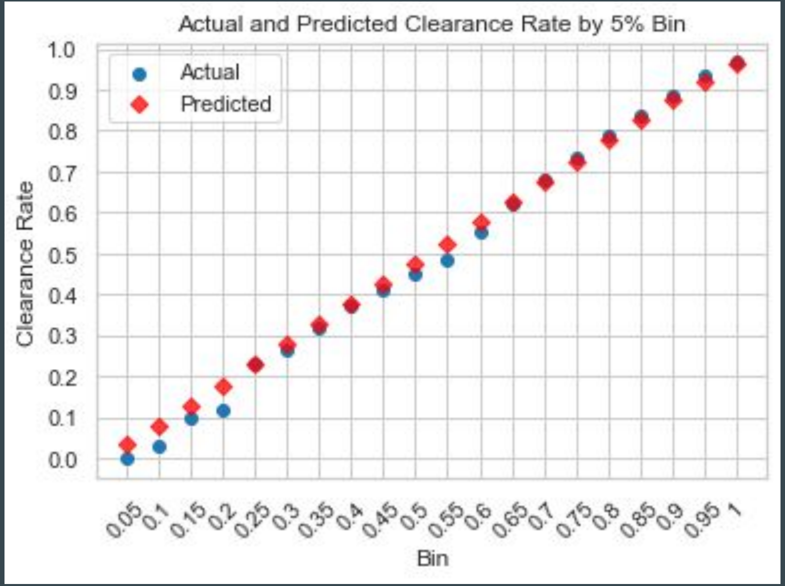
Optimal Model Feature Importances



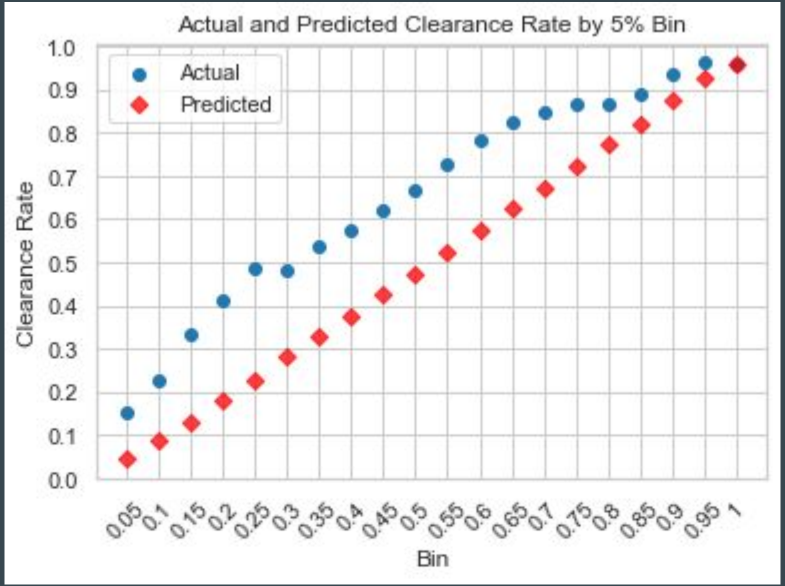
Optimal Model Test Set Probability Distribution



Optimal Model Binned Sum of Squared Residuals Chart

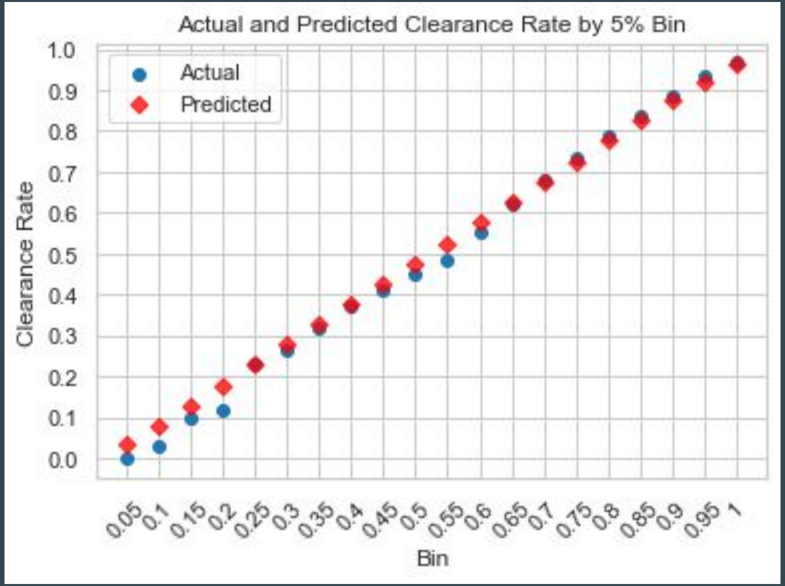


Sub-Optimal Model Binned Sum of Squared Residuals Chart



Can clearance rate be predicted?

Yes!


Actual and Predicted Clearance Rate by 5% Bin
Actual
Predicted
Clearance Rate
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
0.55
0.6
0.65
0.7
0.75
0.8
0.85
0.9
0.95
1
Bin

What should law enforcement do with this?

When a new murder occurs, input the features of the murder into the model.

Calculate the probability of the murder being solved. If low:

Send idle resources!

Send computing power!

Send the *best* investigators!

Send *more* investigators!

What information do you need for a prediction?

Victim demographics: age, race, gender, ethnicity (LatinX or not)

Investigating agency information

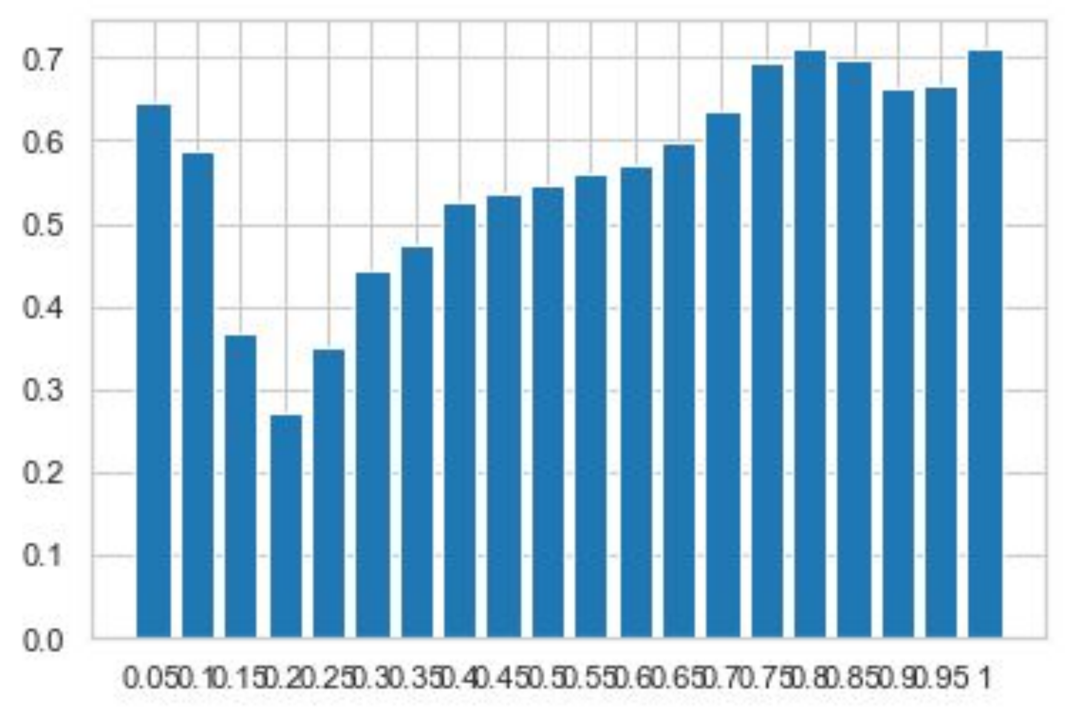
Date of homicide or discovery

Location of homicide

Type of weapon used

Total number of victims

‘WhiteMurderPercent’ by Probability



Potential Pitfalls of the Model

Reinforcing bias and inequality.

The model's output should be used as a method of determining which murders should have more resources directed toward their clearance, not as a filtering process that reinforces already-existing demographic or geographic inequality. "Ignore this murder, because it's unlikely to be solved anyway."

Complacency toward cases with high clearance probabilities

Future Work

Better, more computationally-intensive models

Census data about murder location as features

Requesting from law enforcement more features related to characteristics of the murder

Major data constraint: lacking time elapsed between murder and clearance, a likely useful feature

Activism. The federal government **must** enact uniform collection and reporting requirements for data relating to violent crime.