Analyzing U.S. Crime Data | Technical Report

Muhammed Khalid	201901493
Samaa Khair	201901481
Eman Allam	201900903
Abdelmonem Ali	201800276

University of Science and Technology at Zewail City

2022

Supervised by:
Dr. Mahmoud Abdelaziz

I. Introduction

The aim of this report is to document the results and analysis of the project. In this report, we will show all results and all of the comments to such results, and finally we will list some limitations that affected the results of the project.

II. Answering Questions

A. Question 1)

Which type of non-fatal crime is the most under-reported? Is there an association between the offender-victim relationship and the likelihood of a crime being reported?

Most under reported crime is Rape/Sexual Assault. The bar plot for all crimes was as follow:

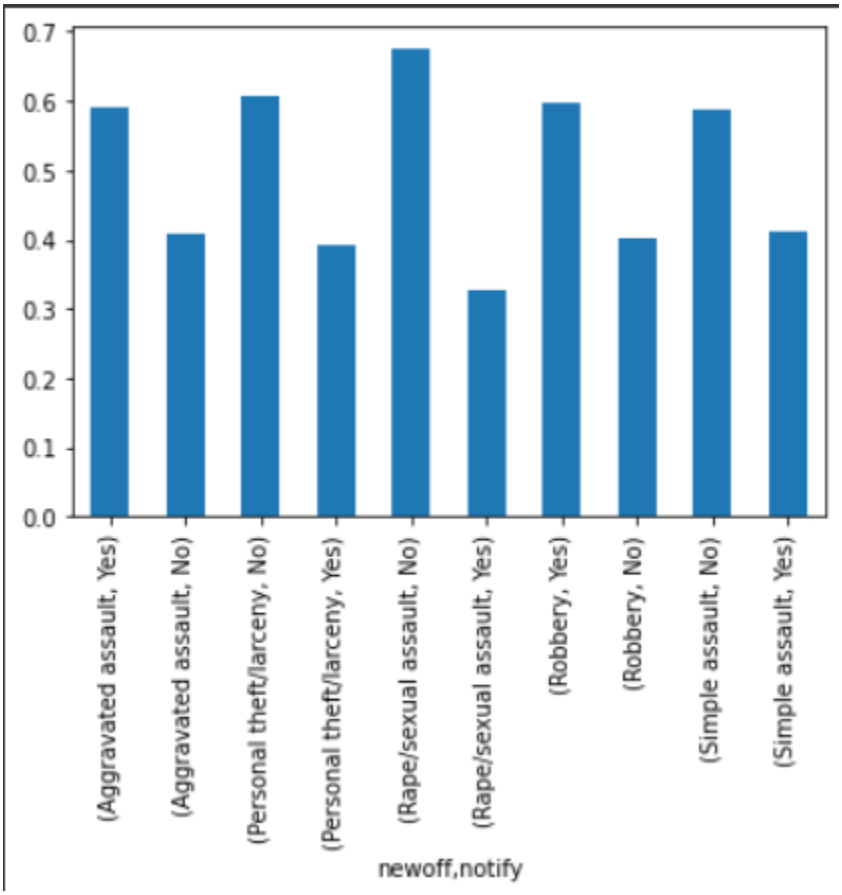


Figure 1: Bar plot between all crimes and reporting ratio.

The exact numbers are:

```
newoff                  notify
Aggravated assault      Yes       0.591993
                        No        0.408007
Personal theft/larceny  No        0.608374
                        Yes       0.391626
Rape/sexual assault     No        0.674792
                        Yes       0.325208
Robbery                 Yes       0.596823
                        No        0.403177
Simple assault          No        0.588360
                        Yes       0.411640
Name: notify, dtype: float64
```

Figure 2: Reporting ratio as numbers for all crimes.

For the second part of the question (association between the offender-victim relationship and the likelihood of a crime being reported) The results showed that there is some association. The bar plot result is below:

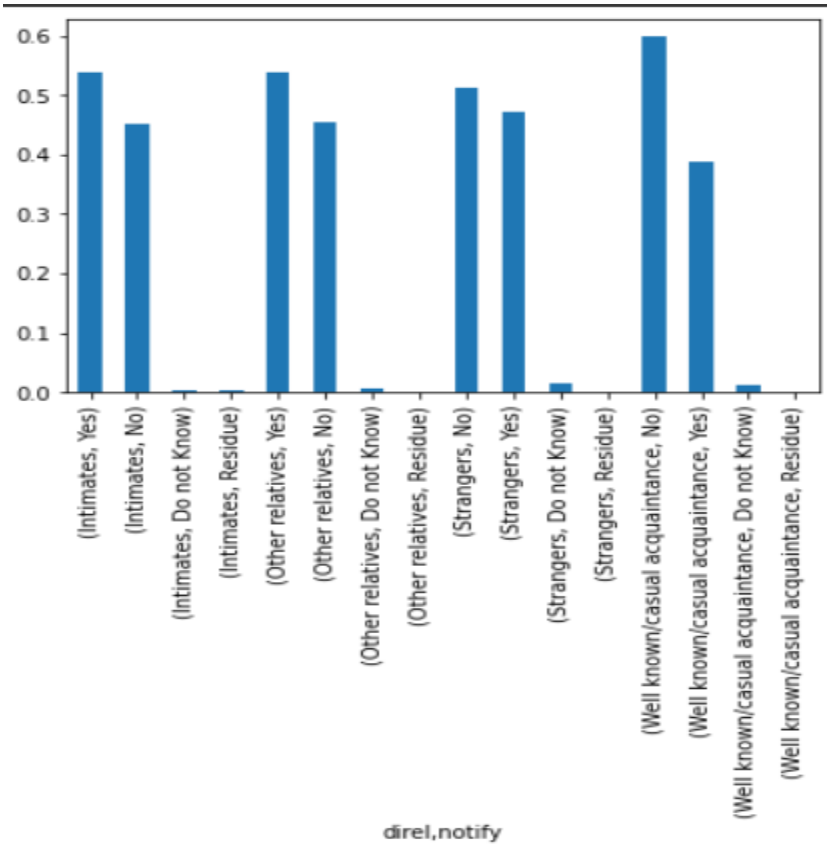


Figure 3: Bar plot of reporting ratio and victim-offender relationship

The result suggests that if the offender is a well known/casual acquaintance, then the crime is most likely under-reported. The exact numbers are below:

```
direl                          notify
Intimates                      Yes           0.538682
                               No            0.452843
                               Do not Know   0.004784
                               Residue       0.003691
Other relatives                Yes           0.537607
                               No            0.454123
                               Do not Know   0.007495
                               Residue       0.000775
Strangers                      No            0.512672
                               Yes           0.470723
                               Do not Know   0.016352
                               Residue       0.000254
Well known/casual acquaintance No            0.598776
                               Yes           0.387863
                               Do not Know   0.012596
                               Residue       0.000765
Name: notify, dtype: float64
```

Figure 4: All ratios as numbers of reporting ratio vs victim-offender relationship

B. Question 2)

Who are the people (the demographic segment) that appear to be most at risk of violent? Who is the least at risk?

People at most risk:

```
('12-17', 'Female', 'Non-Hispanic American Indian/Alaska Native', 'Violent crime')

 ('18-24', 'Female', 'Non-Hispanic American Indian/Alaska Native', 'Violent crime')


 ('25-34', 'Male', 'Non-Hispanic American Indian/Alaska Native', 'Violent crime')

('25-34', 'Male', 'Non-Hispanic more than one race', 'Violent crime')

('65 or older', 'Male', 'Non-Hispanic American Indian/Alaska Native', 'Violent crime')
```

People at least risk:

```
('65 or older', 'Female', 'Non-Hispanic American Indian/Alaska Native', 'Personal theft/larceny')
```

The Bar plot:

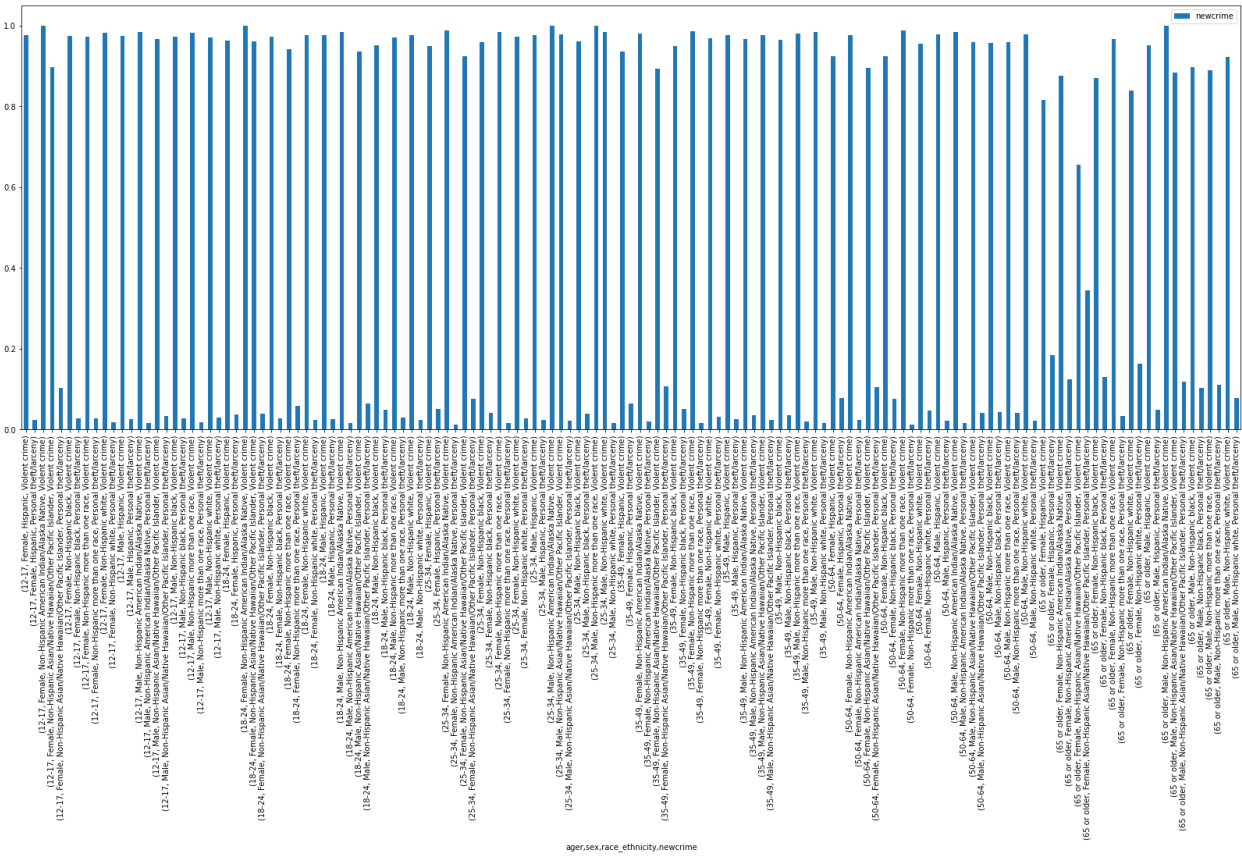


Figure 5: Bar plot for demographic vs violent.

Some of the numbers:

	newcrime
('12-17', 'Female', 'Hispanic', 'Violent crime')	0.97546
('12-17', 'Female', 'Hispanic', 'Personal theft/larceny')	0.0245399
('12-17', 'Female', 'Non-Hispanic American Indian/Alaska Native', 'Violent crime')	1
('12-17', 'Female', 'Non-Hispanic Asian/Native Hawaiian/Other Pacific Islander', 'Violent crime')	0.897059
('12-17', 'Female', 'Non-Hispanic Asian/Native Hawaiian/Other Pacific Islander', 'Personal theft/larceny')	0.102941
('12-17', 'Female', 'Non-Hispanic black', 'Violent crime')	0.973239
('12-17', 'Female', 'Non-Hispanic black', 'Personal theft/larceny')	0.0267606
('12-17', 'Female', 'Non-Hispanic more than one race', 'Violent crime')	0.971698
('12-17', 'Female', 'Non-Hispanic more than one race', 'Personal theft/larceny')	0.0283019
('12-17', 'Female', 'Non-Hispanic white', 'Violent crime')	0.981928
('12-17', 'Female', 'Non-Hispanic white', 'Personal theft/larceny')	0.0180723
('12-17', 'Male', 'Hispanic', 'Violent crime')	0.973799
('12-17', 'Male', 'Hispanic', 'Personal theft/larceny')	0.0262009
('12-17', 'Male', 'Non-Hispanic American Indian/Alaska Native', 'Violent crime')	0.984127
('12-17', 'Male', 'Non-Hispanic American Indian/Alaska Native', 'Personal theft/larceny')	0.015873
('12-17', 'Male', 'Non-Hispanic Asian/Native Hawaiian/Other Pacific Islander', 'Violent crime')	0.966667
('12-17', 'Male', 'Non-Hispanic Asian/Native Hawaiian/Other Pacific Islander', 'Personal theft/larceny')	0.0333333
('12-17', 'Male', 'Non-Hispanic black', 'Violent crime')	0.971526
('12-17', 'Male', 'Non-Hispanic black', 'Personal theft/larceny')	0.0284738
('12-17', 'Male', 'Non-Hispanic more than one race', 'Violent crime')	0.981982
('12-17', 'Male', 'Non-Hispanic more than one race', 'Personal theft/larceny')	0.018018
('12-17', 'Male', 'Non-Hispanic white', 'Violent crime')	0.970334

Figure 6: Some numbers for demographic vs violent

C. Question 3)

Of all victims of non-fatal crimes who suffer an injury, which demographic is the most likely to receive medical attention at the scene? Which is the least likely?

Most likely:
65 or older - Male - Hispanic
Least likely:
12-17 - Male - Non-Hispanic more than one race
Bar Plot:

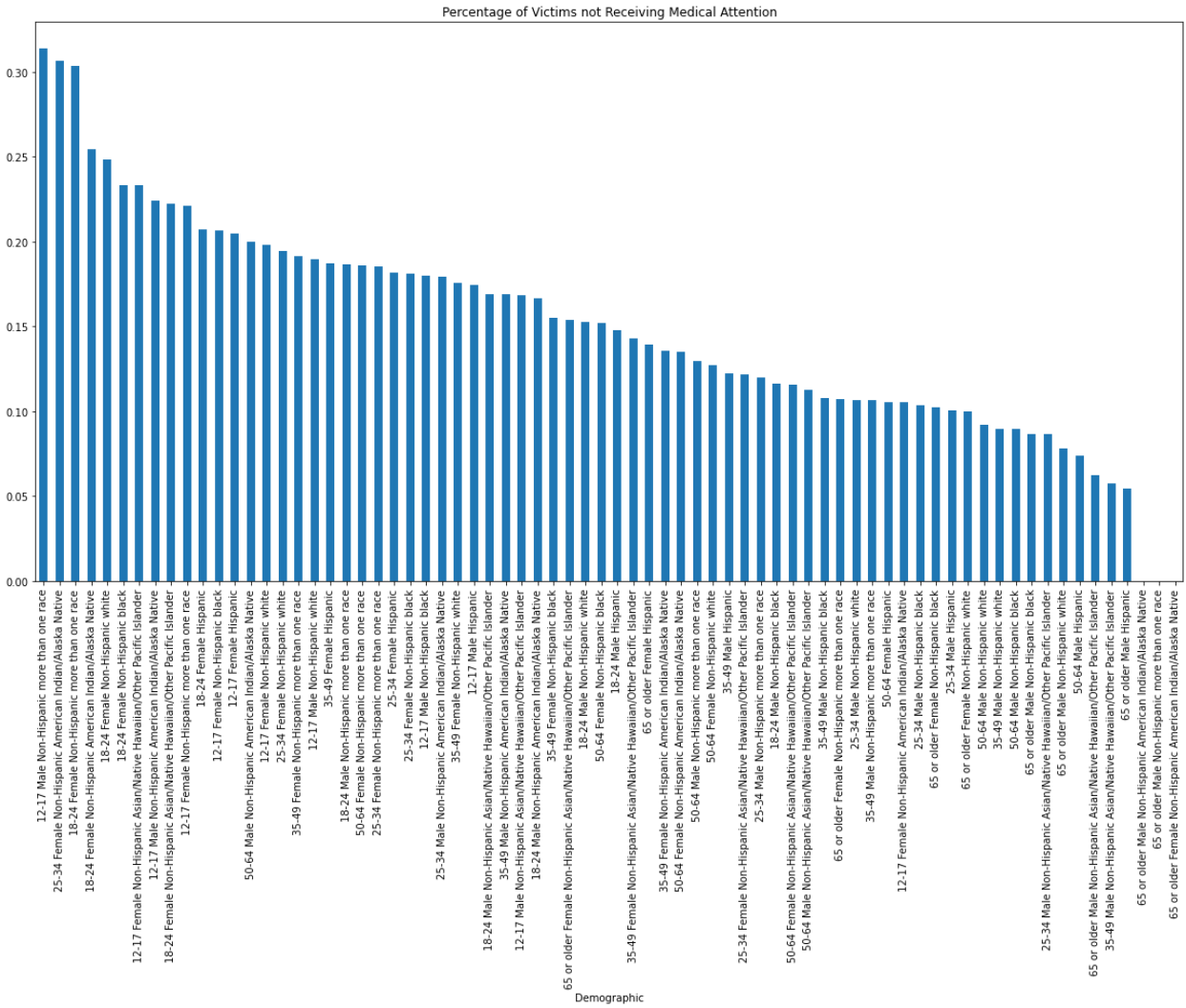


Figure 7: Bar plot for demographic vs non treated.

Some Numbers:

	treatment
('12-17', 'Female', 'Hispanic', 'Not injured')	0.697853
('12-17', 'Female', 'Hispanic', 'Not treated')	0.179448
('12-17', 'Female', 'Hispanic', 'Treated at scene,home,medical office or other location')	0.122699
('12-17', 'Female', 'Non-Hispanic American Indian/Alaska Native', 'Not injured')	0.790698
('12-17', 'Female', 'Non-Hispanic American Indian/Alaska Native', 'Treated at scene,home,medical office or other location')	0.116279
('12-17', 'Female', 'Non-Hispanic American Indian/Alaska Native', 'Not treated')	0.0930233
('12-17', 'Female', 'Non-Hispanic Asian/Native Hawaiian/Other Pacific Islander', 'Not injured')	0.676471
('12-17', 'Female', 'Non-Hispanic Asian/Native Hawaiian/Other Pacific Islander', 'Not treated')	0.205882
('12-17', 'Female', 'Non-Hispanic Asian/Native Hawaiian/Other Pacific Islander', 'Treated at scene,home,medical office or other location')	0.117647
('12-17', 'Female', 'Non-Hispanic black', 'Not injured')	0.685915
('12-17', 'Female', 'Non-Hispanic black', 'Not treated')	0.178873
('12-17', 'Female', 'Non-Hispanic black', 'Treated at scene,home,medical office or other location')	0.133803
('12-17', 'Female', 'Non-Hispanic black', 'Do not know')	0.00140845
('12-17', 'Female', 'Non-Hispanic more than one race', 'Not injured')	0.698113
('12-17', 'Female', 'Non-Hispanic more than one race', 'Not treated')	0.198113
('12-17', 'Female', 'Non-Hispanic more than one race', 'Treated at scene,home,medical office or other location')	0.103774
('12-17', 'Female', 'Non-Hispanic white', 'Not injured')	0.72992
('12-17', 'Female', 'Non-Hispanic white', 'Not treated')	0.180388
('12-17', 'Female', 'Non-Hispanic white', 'Treated at scene,home,medical office or other location')	0.0893574
('12-17', 'Female', 'Non-Hispanic white', 'Do not know')	0.000334672
('12-17', 'Male', 'Hispanic', 'Not injured')	0.739738
('12-17', 'Male', 'Hispanic', 'Not treated')	0.156332
('12-17', 'Male', 'Hispanic', 'Treated at scene,home,medical office or other location')	0.103057
('12-17', 'Male', 'Hispanic', 'Do not know')	0.000873362
('12-17', 'Male', 'Non-Hispanic American Indian/Alaska Native', 'Not injured')	0.714286
('12-17', 'Male', 'Non-Hispanic American Indian/Alaska Native', 'Not treated')	0.206349
('12-17', 'Male', 'Non-Hispanic American Indian/Alaska Native', 'Treated at scene,home,medical office or other location')	0.0793651
('12-17', 'Male', 'Non-Hispanic Asian/Native Hawaiian/Other Pacific Islander', 'Not injured')	0.741667
('12-17', 'Male', 'Non-Hispanic Asian/Native Hawaiian/Other Pacific Islander', 'Not treated')	0.15
('12-17', 'Male', 'Non-Hispanic Asian/Native Hawaiian/Other Pacific Islander', 'Treated at scene,home,medical office or other location')	0.108333
('12-17', 'Male', 'Non-Hispanic black', 'Not injured')	0.726651
('12-17', 'Male', 'Non-Hispanic black', 'Not treated')	0.159453
('12-17', 'Male', 'Non-Hispanic black', 'Treated at scene,home,medical office or other location')	0.113895
('12-17', 'Male', 'Non-Hispanic more than one race', 'Not injured')	0.630631

Figure 8: Some numbers for demographic vs non treated ratio.

D. Question 4)

Which class of crimes is associated with the highest rate of same-offense-recidivism; i.e. prison re-entry for the same offense within 3 years of release?

The crime associated with the highest rate of same-offense-recidivism is property.

Bar Plot:

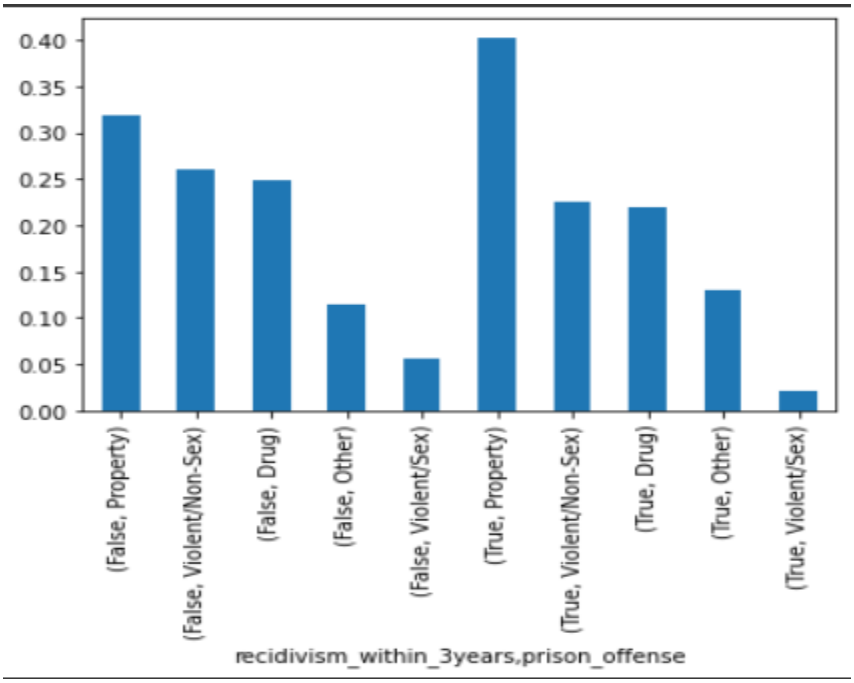


Figure 9: Bar plot for occurrence of crimes at Georgia state [False and True means reoffended]

E. Question 5)

Are prisoners who are younger at the time of release more or less likely to reoffend than those who are older?

Yes, with a higher percentage than being older.

Bar plot:

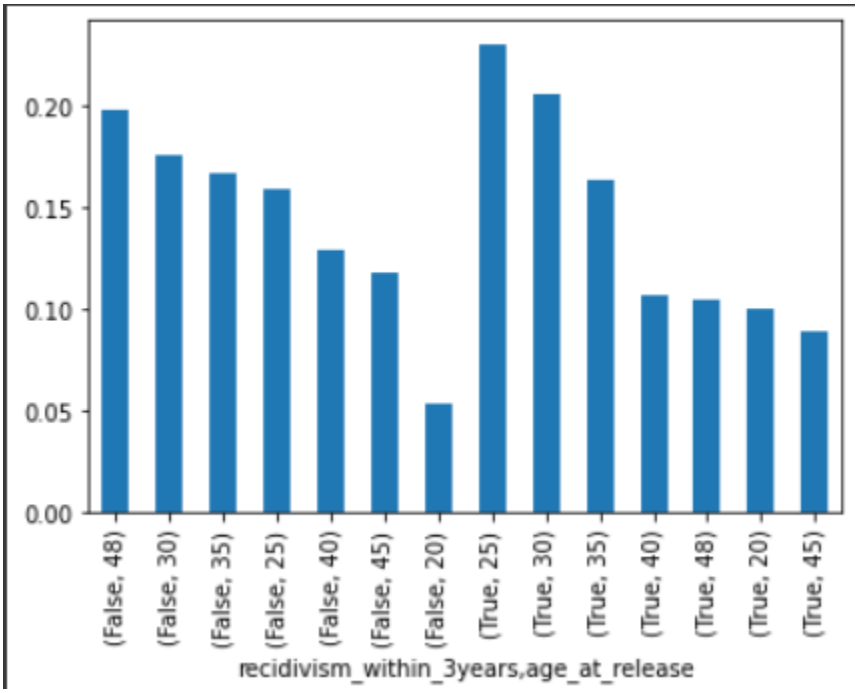


Figure 10: Bar plot for occurrence of being reoffended and age

III. Hypothesis Testing:

Claim: "U.S. states that implement stricter firearm control laws, have lower violent crime rates on average"

The testing results are pending the FBI data extraction.

IV. Regression analysis

Our goal is to predicts the

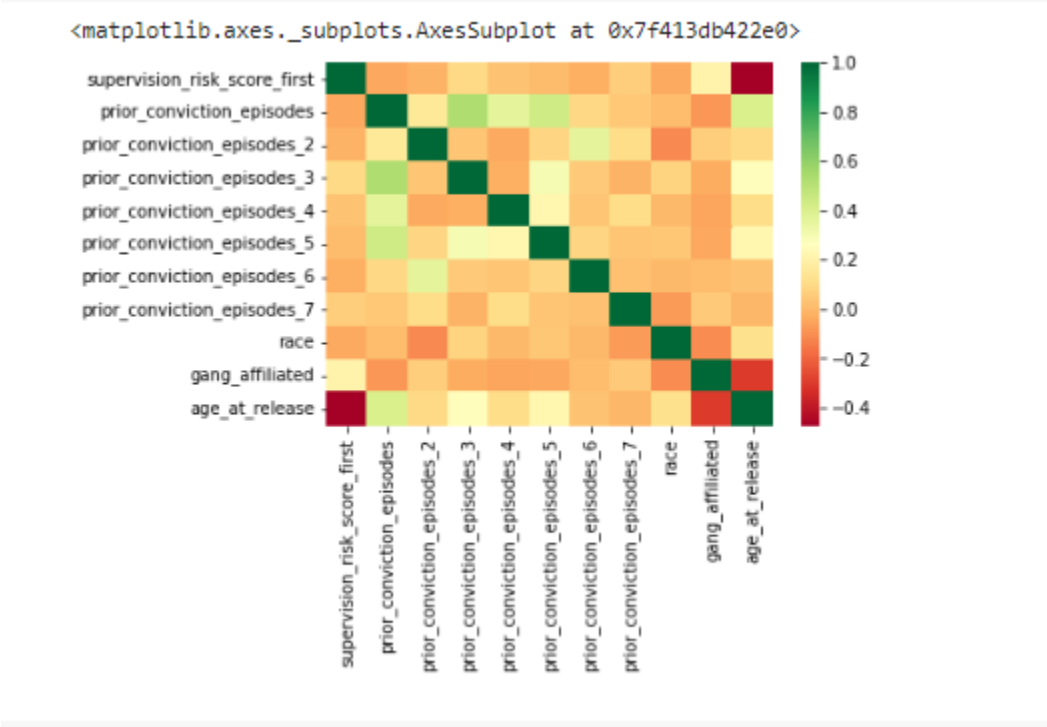
Offender's supervision risk score based on :

- All prior convictions.
- Offender's race.
- Offender's gang affiliation.
- Offender's age at release.

Based on our analysis and regression to find which these variables are good predictors of the variabilities in the target

We found that good predictors are all variables are good predictors except prior_conviction_episodes

Some of these predictors correlated with each other based on our correlation analysis as shown in this heat map



Figure[11]. Heat map of correlated variables

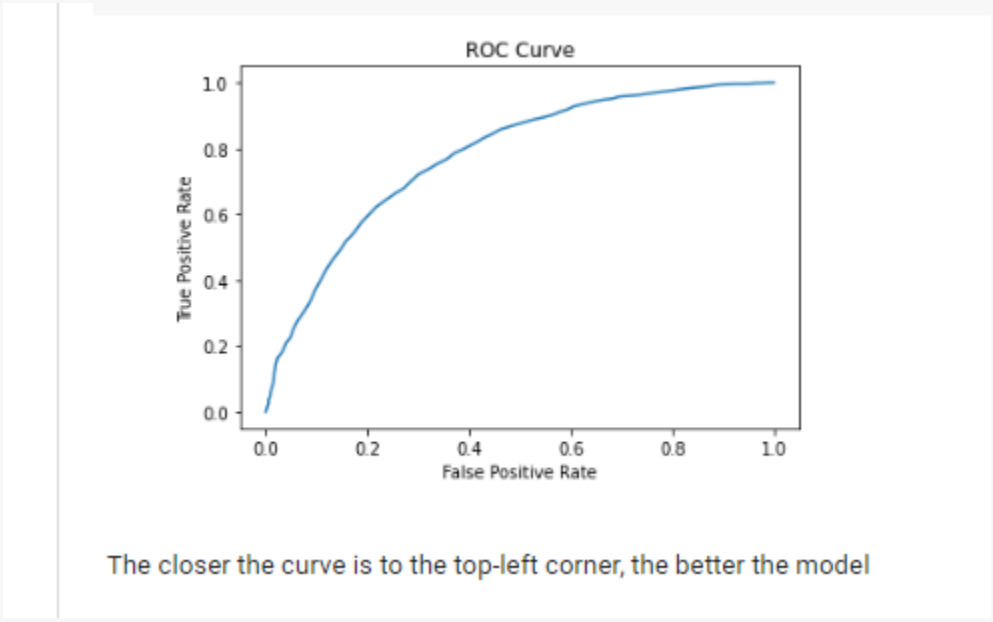
Our regression model performance:

The R-squared value of a linear regression model is a measure of the amount of variance in the target variable that is explained by the model. It takes on a value between 0 and 1, where a higher value indicates a better fit. In our case, an R-squared value of 0.301 indicates that about 30.1% of the variance in the target variable is explained by the model. This means that the model is able to capture some of the variability in the target, but not all of it

V. Machine Learning Model

Data set: The recidivism in Georgia dataset
Required: classifier to predict the likelihood of recidivism
within 3 years of release based on the state of Georgia recidivism records.

Based on our analysis and classification model we reached accuracy equal 0.73
As shown in our ROC curve the model's performance is acceptable



Figure[12] ROC of our curve

Future work:
A test set accuracy of 0.73 for a random forest model indicates that the model is able to correctly predict the target class for about 73% of the test set examples.

To improve the model's test set accuracy, we can try the following:

- Collect more data and use it to train the model.
- Use different hyperparameters for the model.
- Try different feature engineering techniques to create more relevant features for the model.
- Use a different model altogether.

VI. Limitations

Limitations to our regression model :

There could be several reasons why the model has a low R-squared value. Some possible reasons include:

- The relationship between the predictors and the target is not linear.
- There is a lot of noise in the data, making it difficult to accurately predict the target.
- The model is underfitting or overfitting the data.
- There is multicollinearity between the predictors

Limitations to our Georgia Dataset :

- There were a lot of nulls which affected the results for sure.

Limitations to our all datasets :

- There are a lot of mapping which may exhibit a human error.