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

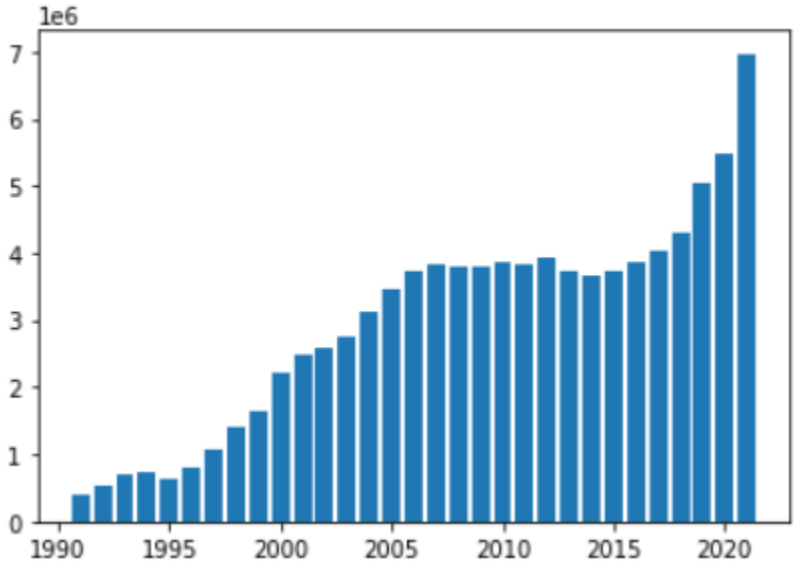
I. Introduction

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 of Part Two

A. Question 1

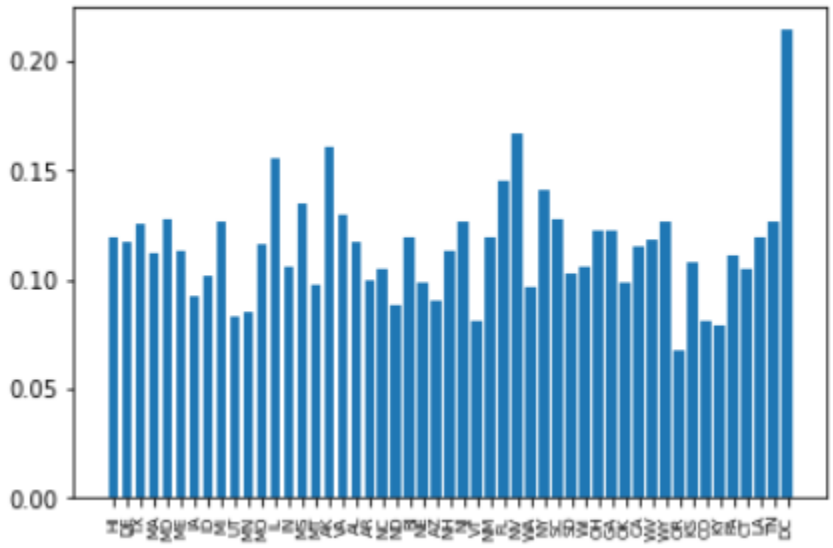
National criminal offense rates per year across all available years for the top five most frequent offense categories.



The bar plot shows the required rates of the five top most offense rates, it can be concluded from the plot that the rates had significantly increased over the years, peaking at year 2021, the peak can be due to the fact that older data are more difficult to collect or in general less reliable. That is just an intuitive assumption or concern that is to be put into consideration through further analysis.

B. Question Two

The average percentage of violent crimes relative to total crime per state over all available years.



Shown in the figure above is the percentage of violent crimes represented in 'assault-offenses','homicide-offenses', 'robbery','kidnapping_abduction' and 'sexual_assault'. It can be interpreted that the highest rate is present in DC: District of Columbia.

C. Question Three

National homicide rates, as well as total violent crime rates per year over all years.

```
<BarContainer object of 31 artists>
```

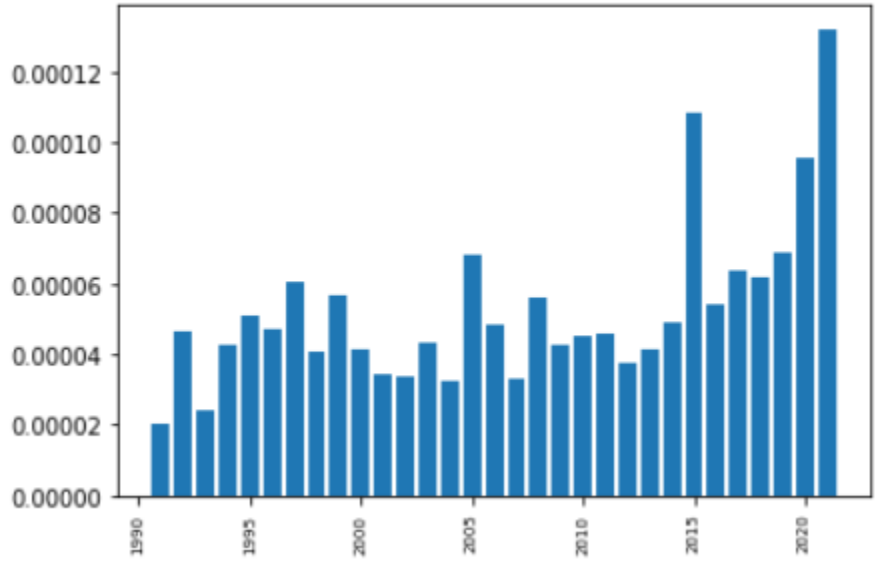


Fig. (a)

Figure (a) represents the rates of homicide rates per year for all years through all states. It can be concluded that years 2021, 2015 and 2005 have the highest rates of homicide.

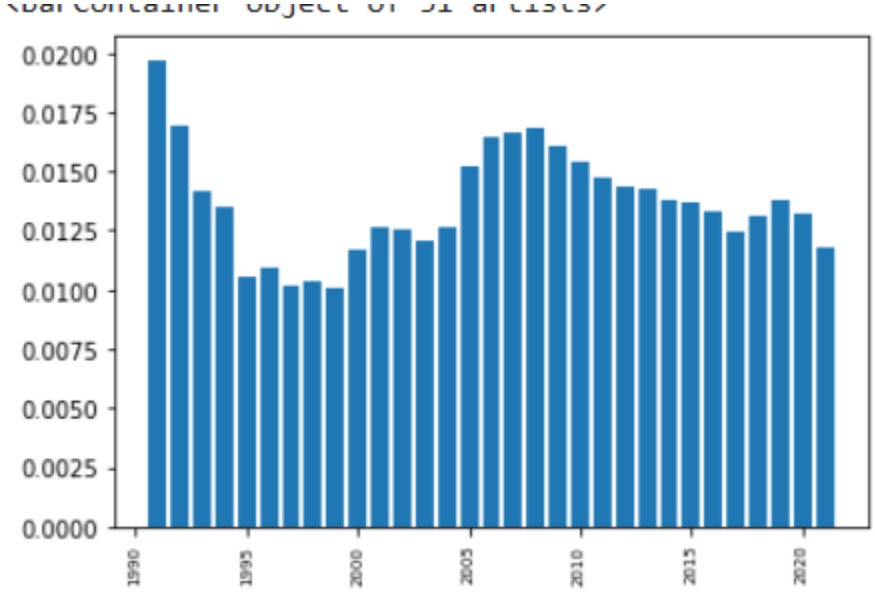


Fig. (b)

Shown in figure (b) rates of violent criminals in general compared to its rates of the whole population. Year 1991 is showin the highest rate of violent crimes. However, the trend tends to decay temporarily only to peak again in 2009.

D. Question Four

The frequency of non-fatal crime incidents in relation to victim demographics

	newoff
('12-17', 'Female', 'Hispanic', 'Simple assault')	0.644172
('12-17', 'Female', 'Hispanic', 'Aggravated assault')	0.18865
('12-17', 'Female', 'Hispanic', 'Robbery')	0.0736196
('12-17', 'Female', 'Hispanic', 'Rape/sexual assault')	0.0690184
('12-17', 'Female', 'Hispanic', 'Personal theft/larceny')	0.0245399
('12-17', 'Female', 'Non-Hispanic American Indian/Alaska Native', 'Simple assault')	0.744186
('12-17', 'Female', 'Non-Hispanic American Indian/Alaska Native', 'Aggravated assault')	0.162791
('12-17', 'Female', 'Non-Hispanic American Indian/Alaska Native', 'Rape/sexual assault')	0.0697674
('12-17', 'Female', 'Non-Hispanic American Indian/Alaska Native', 'Robbery')	0.0232558
('12-17', 'Female', 'Non-Hispanic Asian/Native Hawaiian/Other Pacific Islander', 'Simple assault')	0.602941
('12-17', 'Female', 'Non-Hispanic Asian/Native Hawaiian/Other Pacific Islander', 'Aggravated assault')	0.161765
('12-17', 'Female', 'Non-Hispanic Asian/Native Hawaiian/Other Pacific Islander', 'Personal theft/larceny')	0.102941
('12-17', 'Female', 'Non-Hispanic Asian/Native Hawaiian/Other Pacific Islander', 'Rape/sexual assault')	0.102941
('12-17', 'Female', 'Non-Hispanic Asian/Native Hawaiian/Other Pacific Islander', 'Robbery')	0.0294118
('12-17', 'Female', 'Non-Hispanic black', 'Simple assault')	0.611268
('12-17', 'Female', 'Non-Hispanic black', 'Aggravated assault')	0.18169
('12-17', 'Female', 'Non-Hispanic black', 'Rape/sexual assault')	0.0957746
('12-17', 'Female', 'Non-Hispanic black', 'Robbery')	0.084507
('12-17', 'Female', 'Non-Hispanic black', 'Personal theft/larceny')	0.0267606
('12-17', 'Female', 'Non-Hispanic more than one race', 'Simple assault')	0.698113
('12-17', 'Female', 'Non-Hispanic more than one race', 'Robbery')	0.103774
('12-17', 'Female', 'Non-Hispanic more than one race', 'Aggravated assault')	0.0943396
('12-17', 'Female', 'Non-Hispanic more than one race', 'Rape/sexual assault')	0.0754717
('12-17', 'Female', 'Non-Hispanic more than one race', 'Personal theft/larceny')	0.0283019
('12-17', 'Female', 'Non-Hispanic white', 'Simple assault')	0.710843
('12-17', 'Female', 'Non-Hispanic white', 'Aggravated assault')	0.12249
('12-17', 'Female', 'Non-Hispanic white', 'Rape/sexual assault')	0.101071
('12-17', 'Female', 'Non-Hispanic white', 'Robbery')	0.0475234
('12-17', 'Female', 'Non-Hispanic white', 'Personal theft/larceny')	0.0180723
('12-17', 'Male', 'Hispanic', 'Simple assault')	0.587773
('12-17', 'Male', 'Hispanic', 'Aggravated assault')	0.208734
('12-17', 'Male', 'Hispanic', 'Robbery')	0.171179
('12-17', 'Male', 'Hispanic', 'Personal theft/larceny')	0.0262009

Fig. (a)

Figure a is showing the frequencies of different combinations of demographic features compared to the different non-fatal crimes shown in a tabular form. For further exploration for such question following is a bar plot of the top five most frequent demographics vs. crime:

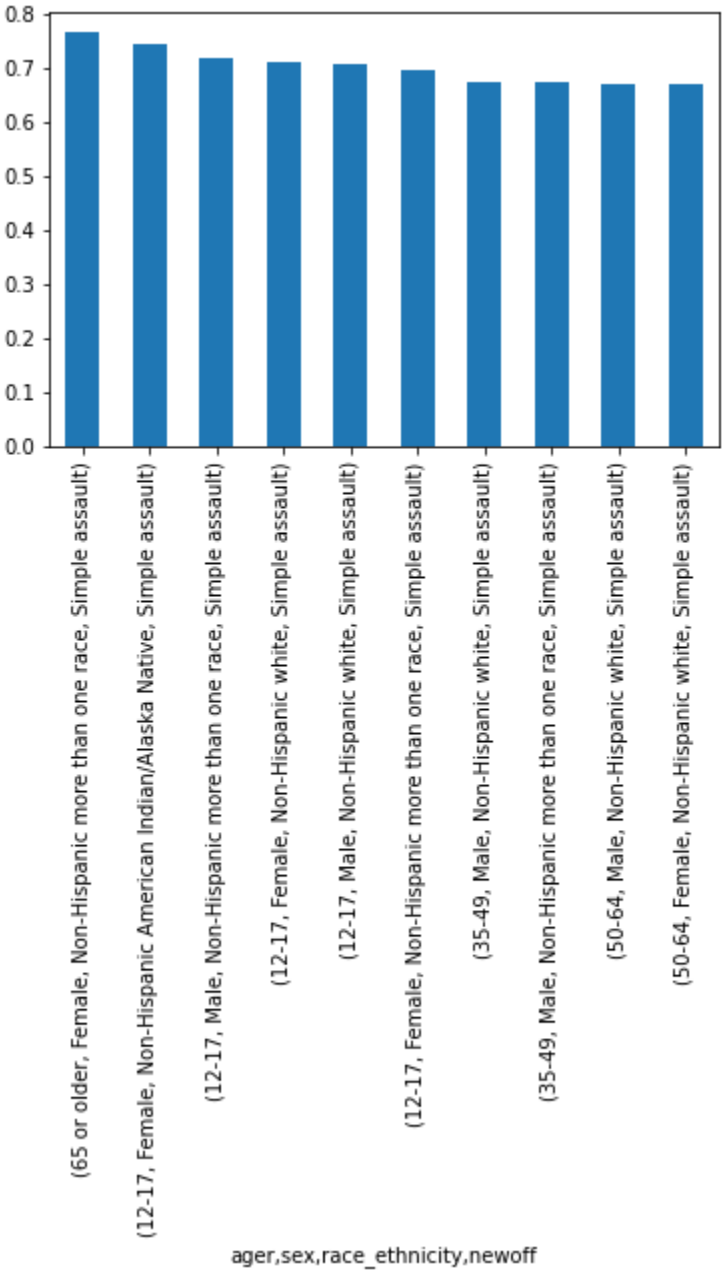


Fig. (b)

It is shown that the most population prone to simple assault are Non-Hispanic, Females whose ages are 65 or older.

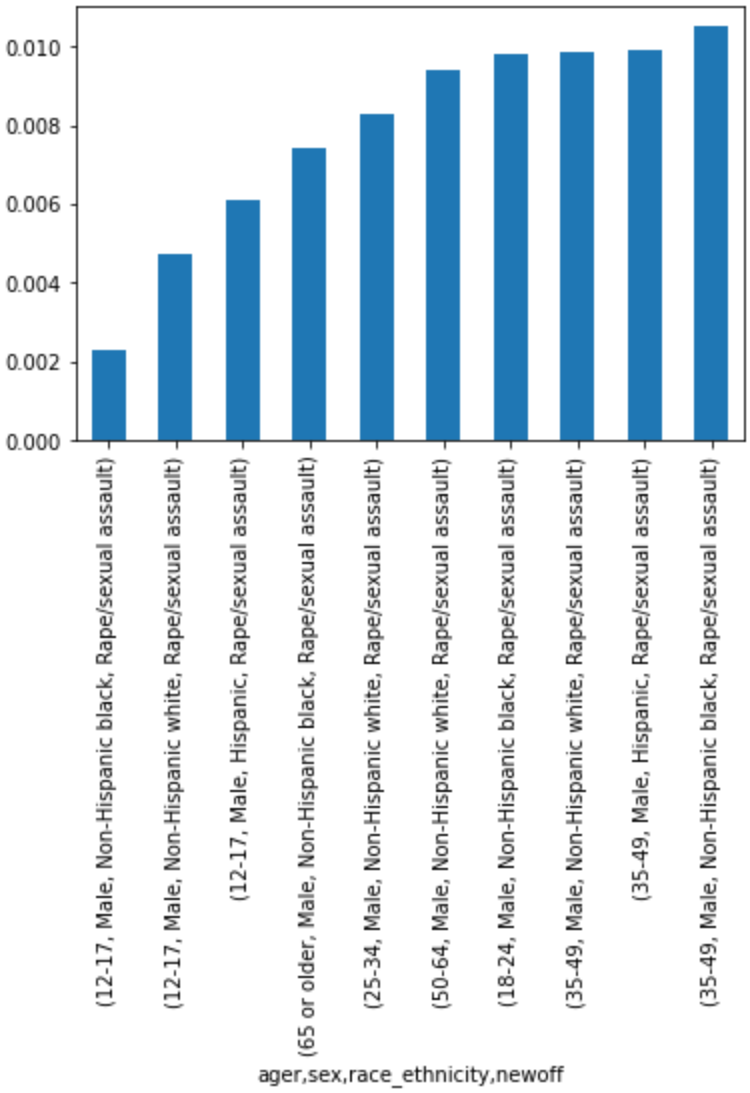


Fig. (c)

Figure (c) shows the least combinations of demographic aspects versus non-fatal crimes.

Non-Hispanic, black males whose age varies from 12 to 17 years ols are the least prone to Rape/sexual assault according to the dataset under investigation.

E. Question Five

The frequency of non-fatal crime incidents in relation to offender Demographics.

Similar to the last question, the same approach for assessing the frequency was taken; presenting a table that shows all combinations and their frequencies, plotting the highest and lowest combinations in a bar plot:

	newoff
('11 or younger', 'Both male and female offenders', 'Invalid until 2012 Q1', 'Simple assault')	0.705882
('11 or younger', 'Both male and female offenders', 'Invalid until 2012 Q1', 'Aggravated assault')	0.294118
('11 or younger', 'Both male and female offenders', 'Non-Hispanic white', 'Simple assault')	1
('11 or younger', 'Female', 'Hispanic', 'Simple assault')	0.75
('11 or younger', 'Female', 'Hispanic', 'Aggravated assault')	0.25
('11 or younger', 'Female', 'Invalid until 2012 Q1', 'Simple assault')	0.65
('11 or younger', 'Female', 'Invalid until 2012 Q1', 'Aggravated assault')	0.2625
('11 or younger', 'Female', 'Invalid until 2012 Q1', 'Robbery')	0.075
('11 or younger', 'Female', 'Invalid until 2012 Q1', 'Personal theft/larceny')	0.0125
('11 or younger', 'Female', 'Non-Hispanic Asian/Native Hawaiian/Other Pacific Islander ', 'Simple assault')	1
('11 or younger', 'Female', 'Non-Hispanic black ', 'Simple assault')	0.818182
('11 or younger', 'Female', 'Non-Hispanic black ', 'Aggravated assault')	0.0909091
('11 or younger', 'Female', 'Non-Hispanic black ', 'Robbery')	0.0909091
('11 or younger', 'Female', 'Non-Hispanic more than one race', 'Simple assault')	1
('11 or younger', 'Female', 'Non-Hispanic white', 'Simple assault')	0.6
('11 or younger', 'Female', 'Non-Hispanic white', 'Aggravated assault')	0.333333
('11 or younger', 'Female', 'Non-Hispanic white', 'Robbery')	0.0666667
('11 or younger', 'Female', 'Unknown race/ethnicity', 'Simple assault')	1
('11 or younger', 'Male', 'Hispanic', 'Simple assault')	0.75
('11 or younger', 'Male', 'Hispanic', 'Aggravated assault')	0.1875
('11 or younger', 'Male', 'Hispanic', 'Rape/sexual assault')	0.03125
('11 or younger', 'Male', 'Hispanic', 'Robbery')	0.03125
('11 or younger', 'Male', 'Invalid until 2012 Q1', 'Simple assault')	0.703093
('11 or younger', 'Male', 'Invalid until 2012 Q1', 'Aggravated assault')	0.230928
('11 or younger', 'Male', 'Invalid until 2012 Q1', 'Robbery')	0.043299
('11 or younger', 'Male', 'Invalid until 2012 Q1', 'Personal theft/larceny')	0.014433
('11 or younger', 'Male', 'Invalid until 2012 Q1', 'Rape/sexual assault')	0.00824742
('11 or younger', 'Male', 'Mixed race group of offenders', 'Aggravated assault')	1
('11 or younger', 'Male', 'Non-Hispanic American Indian/Alaska Native', 'Simple assault')	0.6
('11 or younger', 'Male', 'Non-Hispanic American Indian/Alaska Native', 'Aggravated assault')	0.4
('11 or younger', 'Male', 'Non-Hispanic Asian/Native Hawaiian/Other Pacific Islander ', 'Simple assault')	1
('11 or younger', 'Male', 'Non-Hispanic black ', 'Simple assault')	0.775862

Fig. (a): table showing frequencies

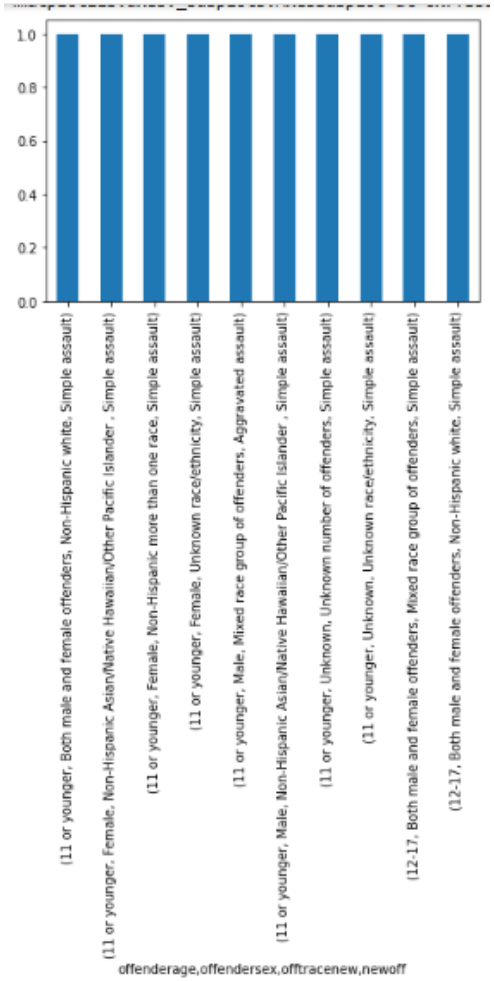


Fig. (b): Bar chart showing top 10 combinations

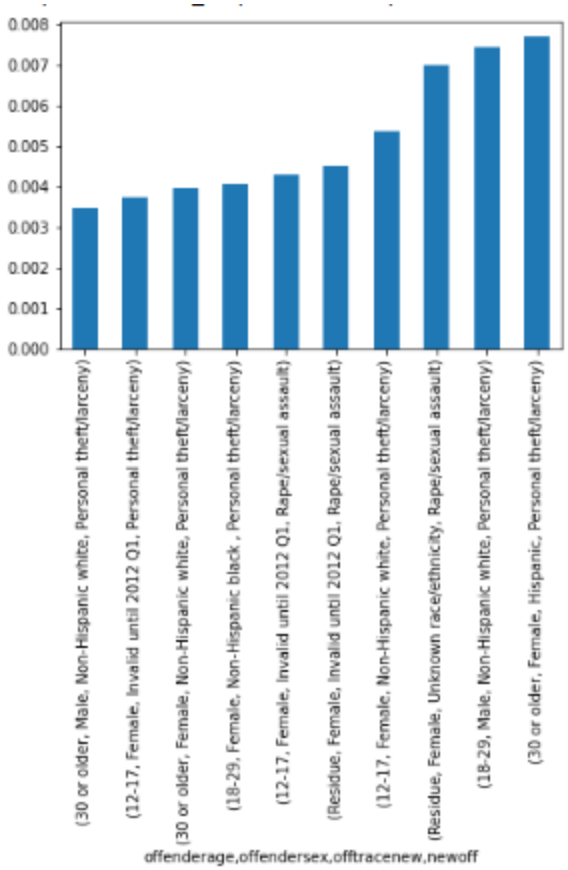


Fig. (c): Bar plots showing the least frequent 10 combinations

N.B. Further improvements in discovering this statistical question can be through dropping unknown or undefined responses that may affect the bias of the results. However, the previously shown results assume they're relevant.

F. Question 6

The relationship between the victim's education level, their gross household income, and their rate of victimization.

The answer to this question shows how would either the education level or the gross household income would affect the rate of victimization. The rate of victimization is basically the number of victims divided by the number of population respondents on the survey.

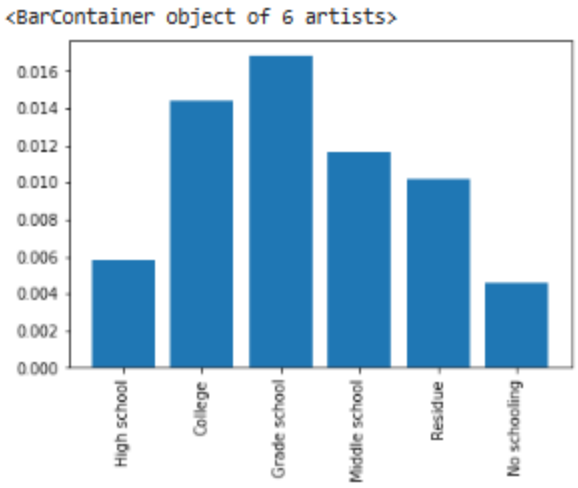


Fig. (a)

Figure (a) shows each education level and its rate of victimization. It can be interpreted that students in Grade School are the most prone to victimization.

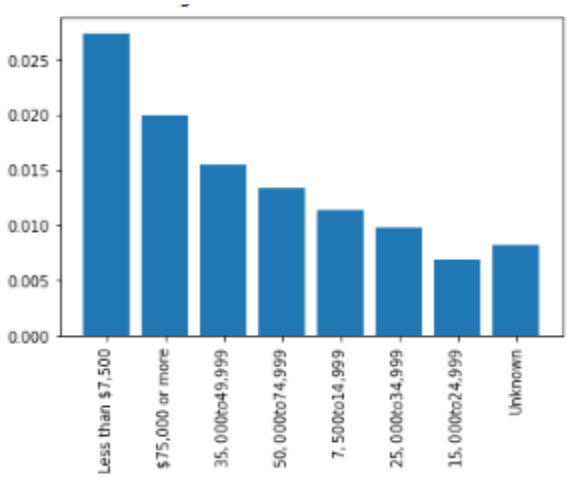


Fig. (b)

Figure (b) shows each segment of household gross income versus the rate of victimization, it can be concluded that people who make less that $7500 have the highest rate of victimization.

III. Answering Questions of Part Three

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:

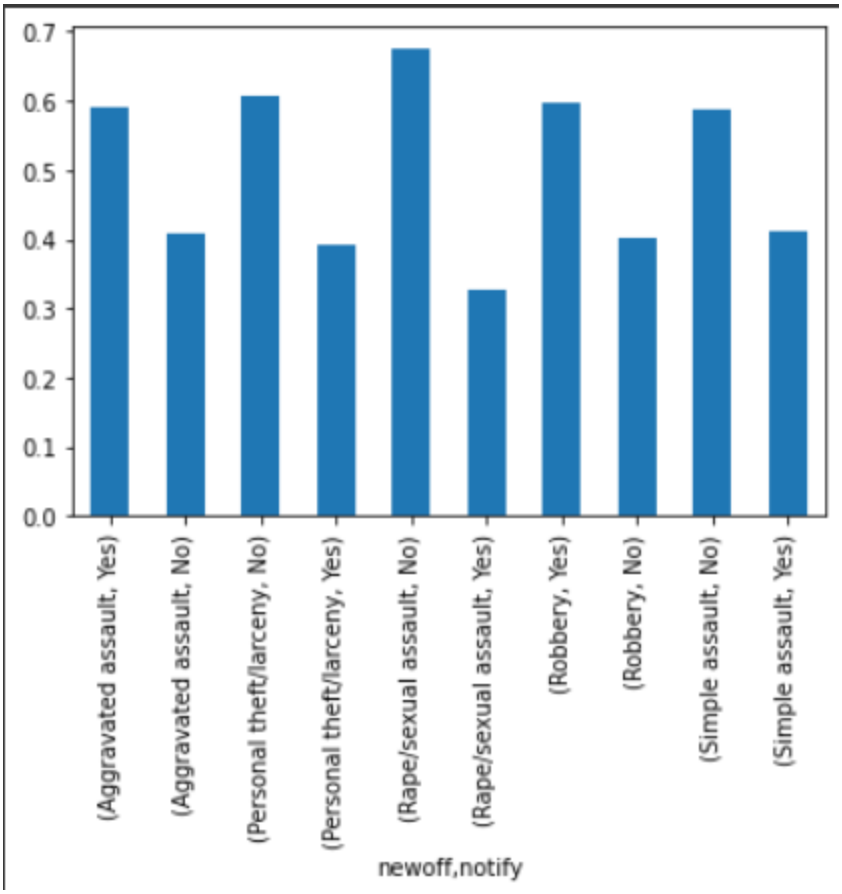


Figure 1: Bar plot between all crimes and reporting ratio.

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:

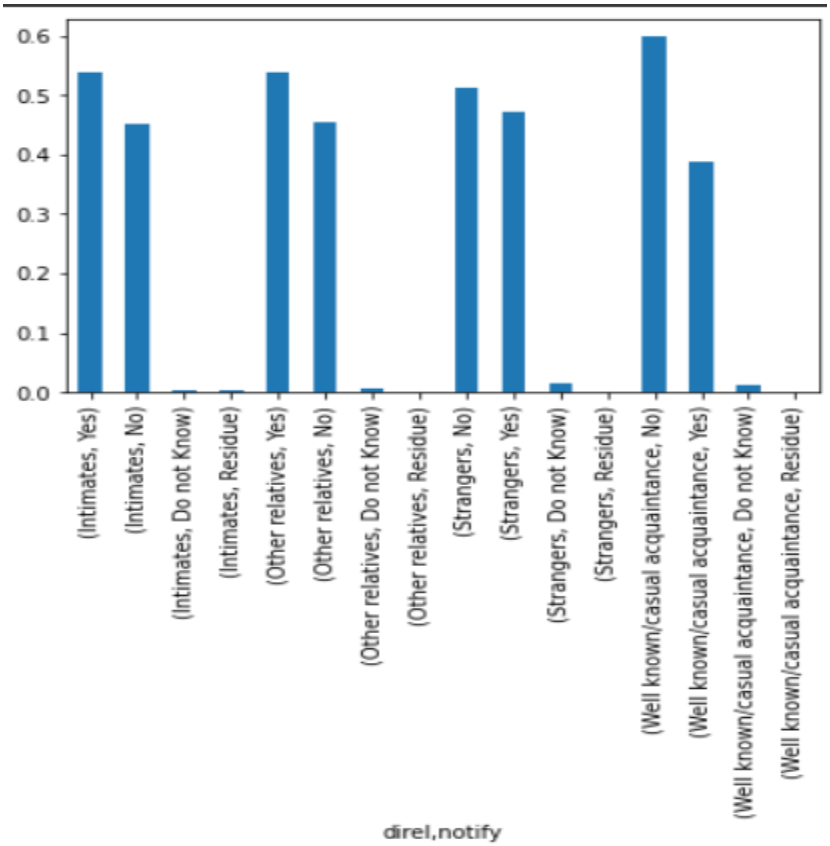


Figure 3: Bar plot of reporting ratio and victim-offender relationship

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:

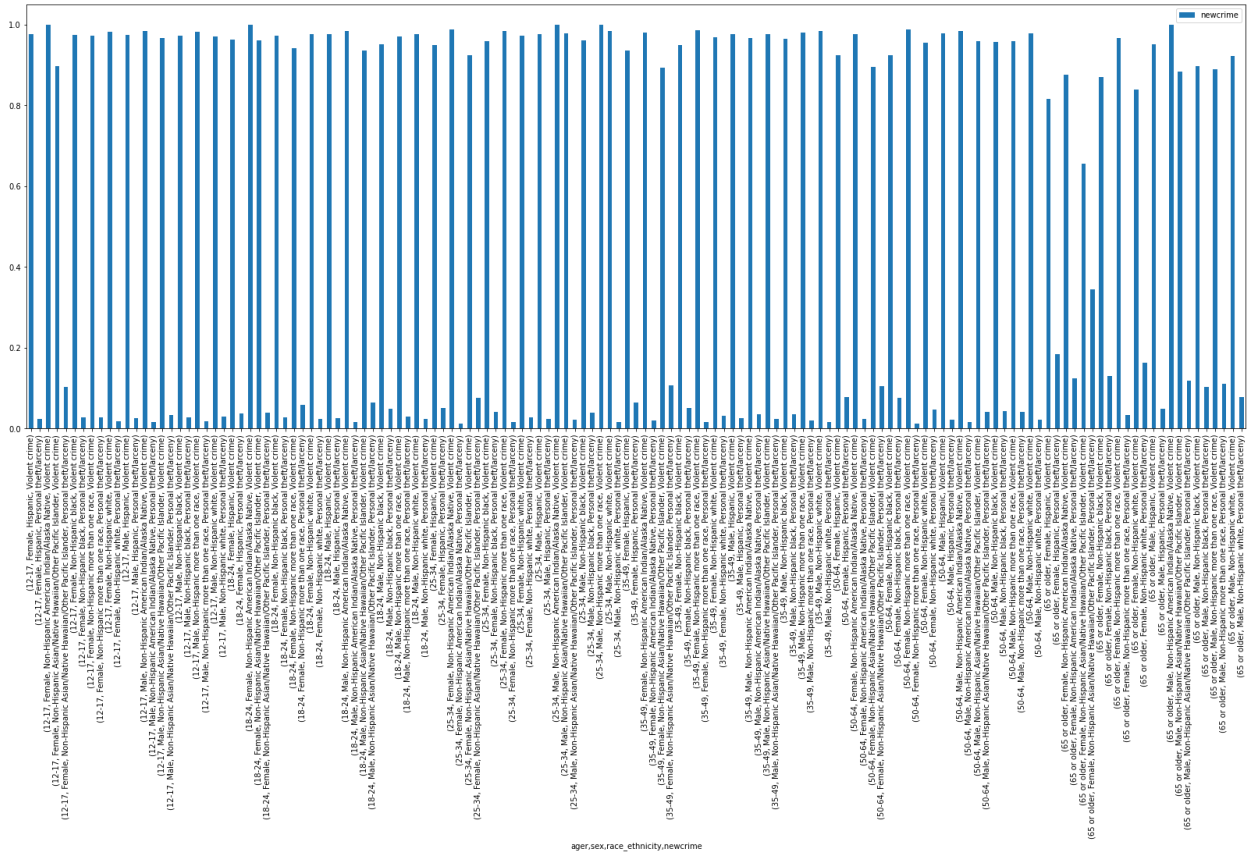


Figure 5: Bar plot for demographic vs violent.

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:

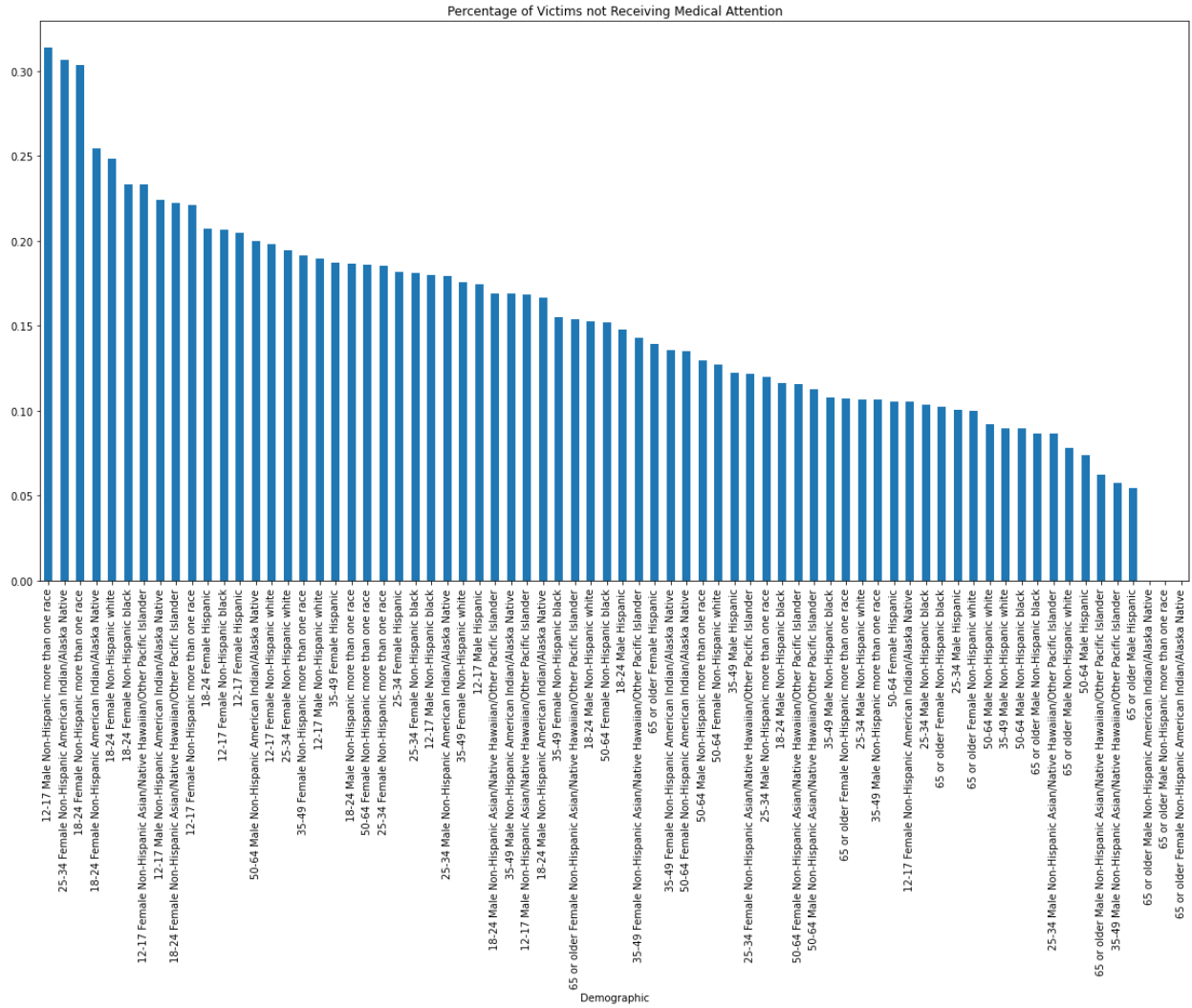


Figure 7: Bar plot for demographic vs non treated.

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:

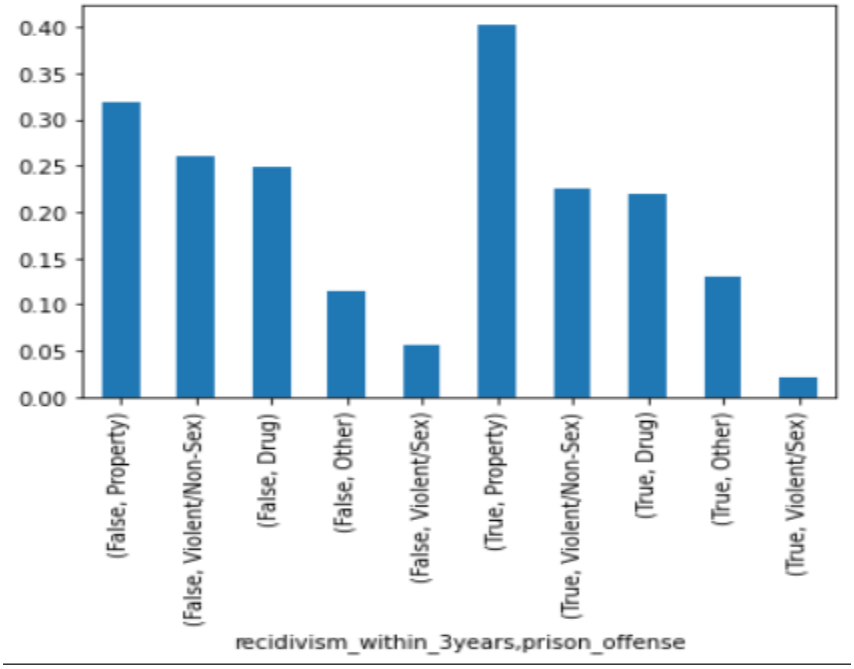


Figure 9: Bar plot for occurrence of crimes at Georgia state [False and True means reoffended]

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:

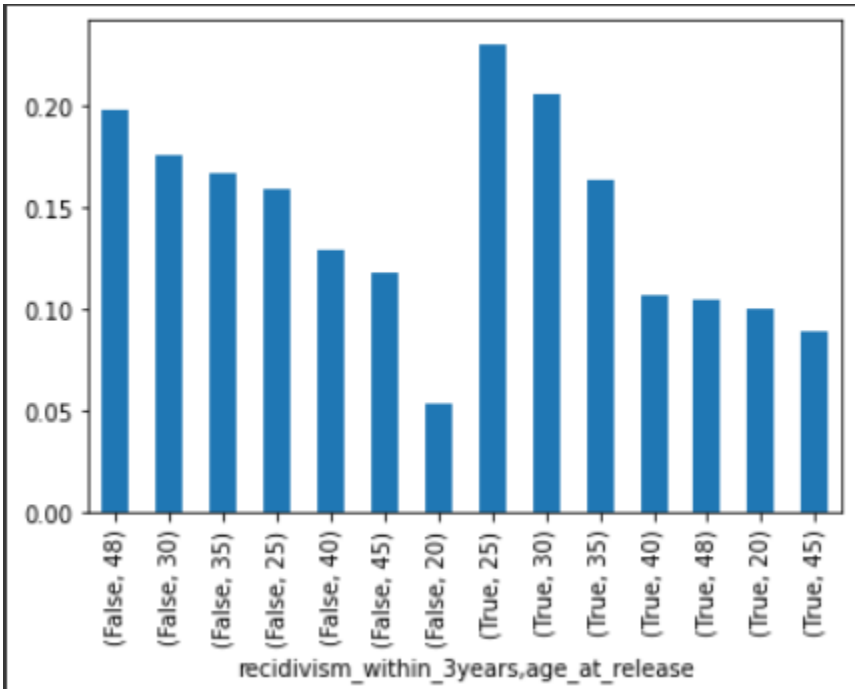


Figure 10: Bar plot for occurrence of being reoffended and age

IV. 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.

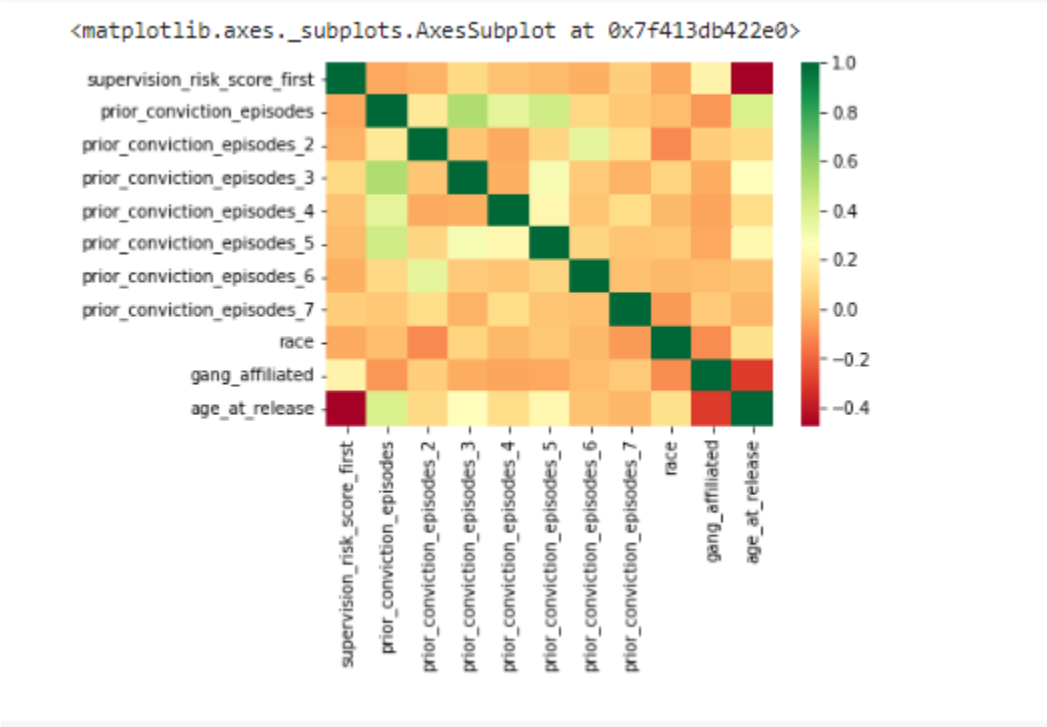
V. Regression analysis

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

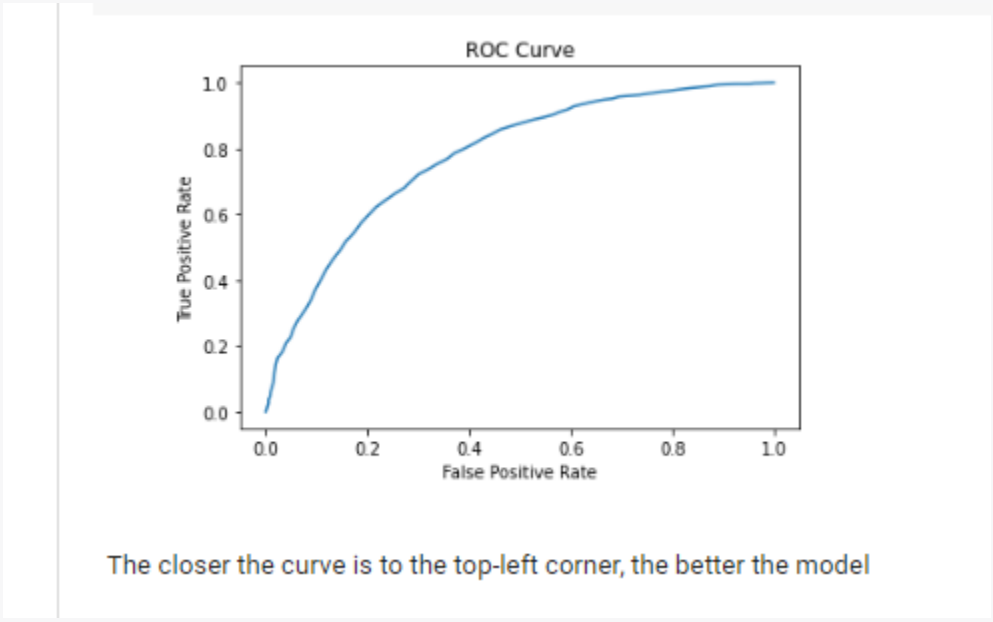
Our regression model performance:

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

VI. 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.

VII. 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.