Zewail City of Science and Technology

[CIE 457]
4th Year – fall
2022

Course:Statistical Inference and Data Analysis

Business Report

CIE 457

Project

Analyzing U.S. Crime Data

Submitted To: Dr. Mahmoud Abdelaziz,

Eng. Asmaa Ismail, Eng. Anhar Abdelmotaleb

Name:	Mohamed Helmy
ID:	201900859
Name:	Youssef Mahmoud Mohamed
ID:	201901093
Name:	Hossam Ashraf
ID:	201901898

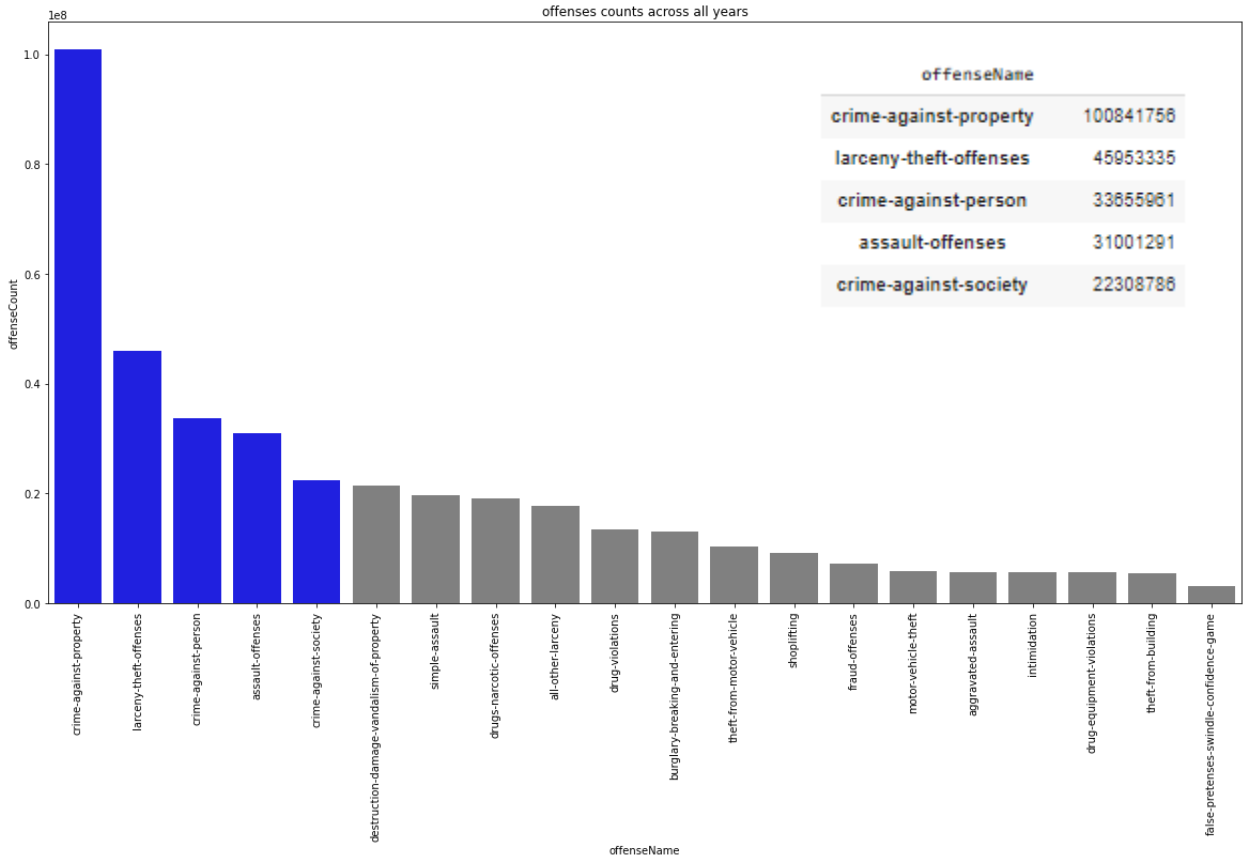
Introduction:

Analyzing crime data based on four different data sets which are the national crime victimization survey (NCVS),NIBRS Reported offense count data , Recidivism data for the state of Georgia [2013-2015] ,and Firearm laws per state. Following the steps of data cleaning , and understanding the features of each data set and choosing the suitable ones, questions about the data are managed to be answered. In addition, hypothesis testing and regression analysis are applied.Also, Neural network models are applied to predict the likelihood of recidivism within 3 years of release.

PART 2: Exploratory Analysis:

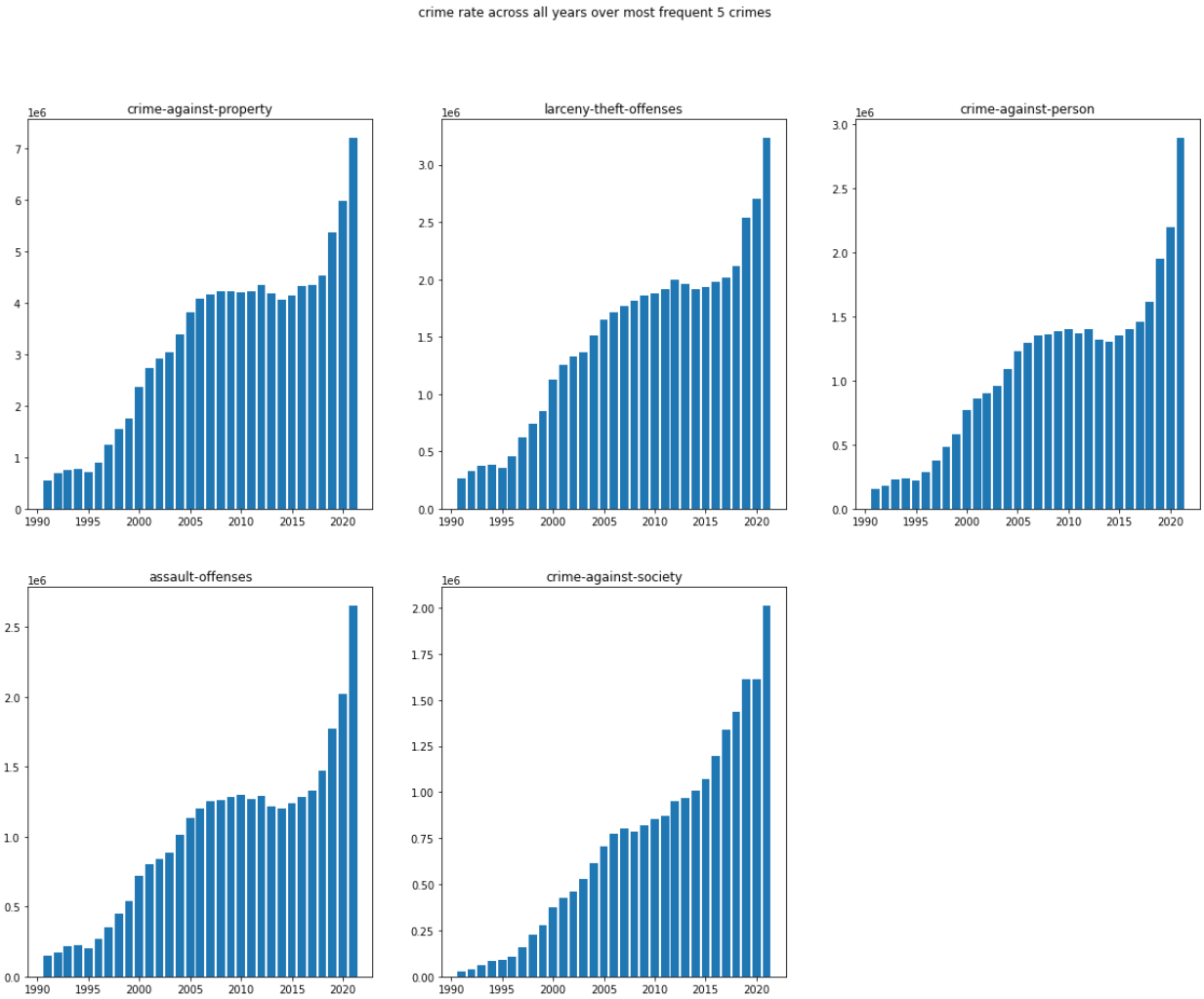
1. National criminal offense rates per year across all available years for the top five most frequent offense categories.

The most frequent offenses categories as show in the graph :



offenseName	
crime-against-property	100841756
larceny-theft-offenses	45953335
crime-against-person	33655961
assault-offenses	31001291
crime-against-society	22308786

For each one of the most frequent crime, the trend over the years as shown:

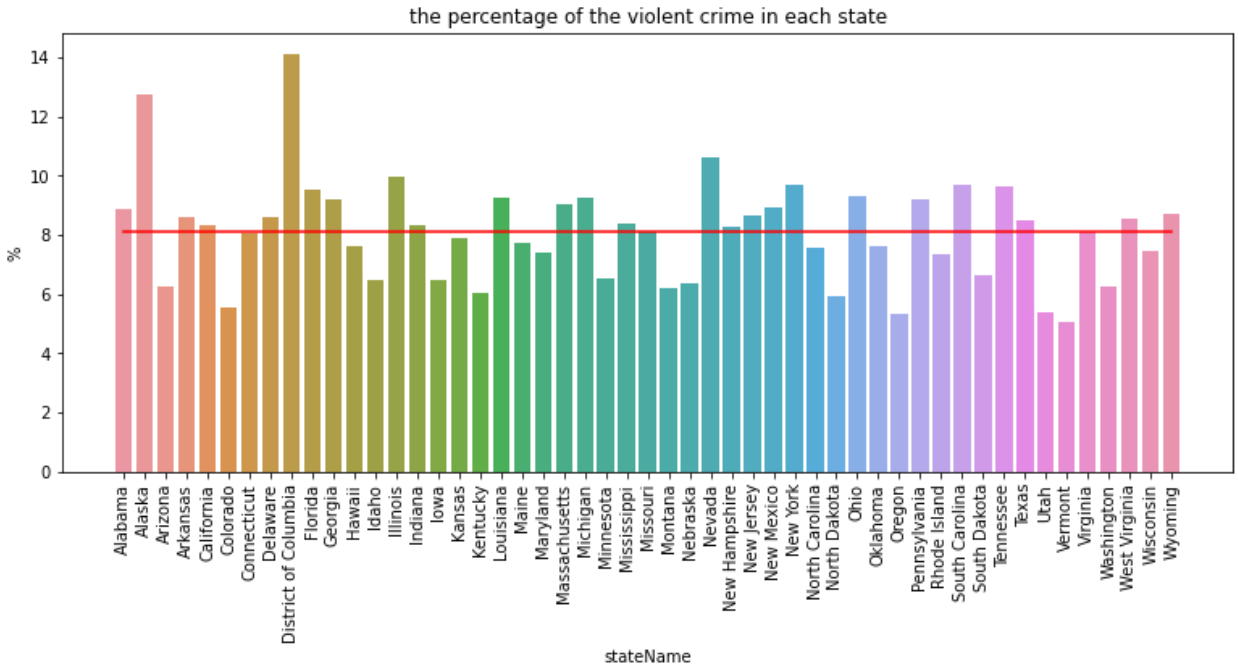


All of the offenses is increasing over the years by a semi linear curve

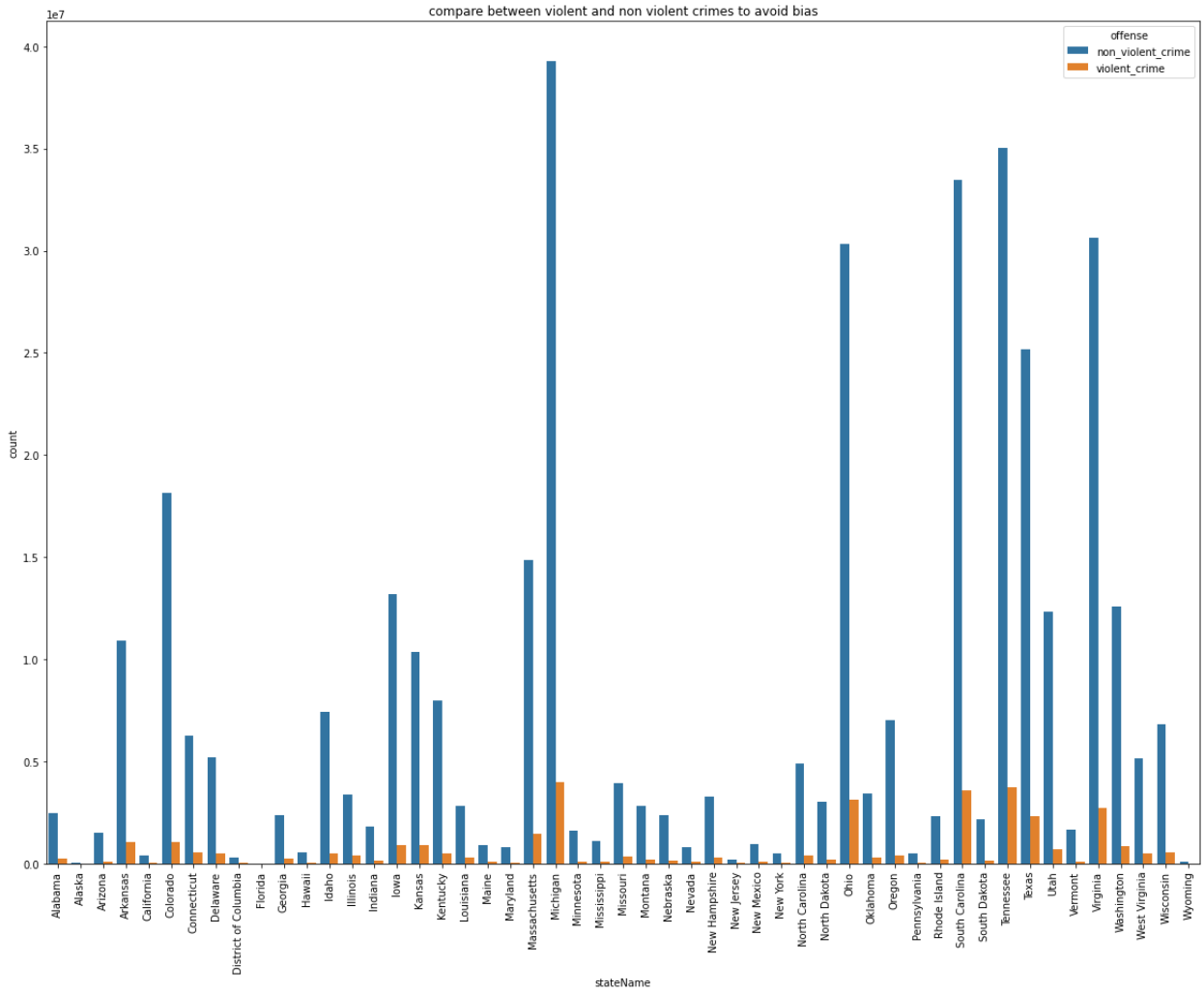
2. The average percentage of violent crimes relative to total crime per state over all available years.

Assumption : over all available years as the comparison between years is negligible

The percentage of the violent crime in each state over all years:



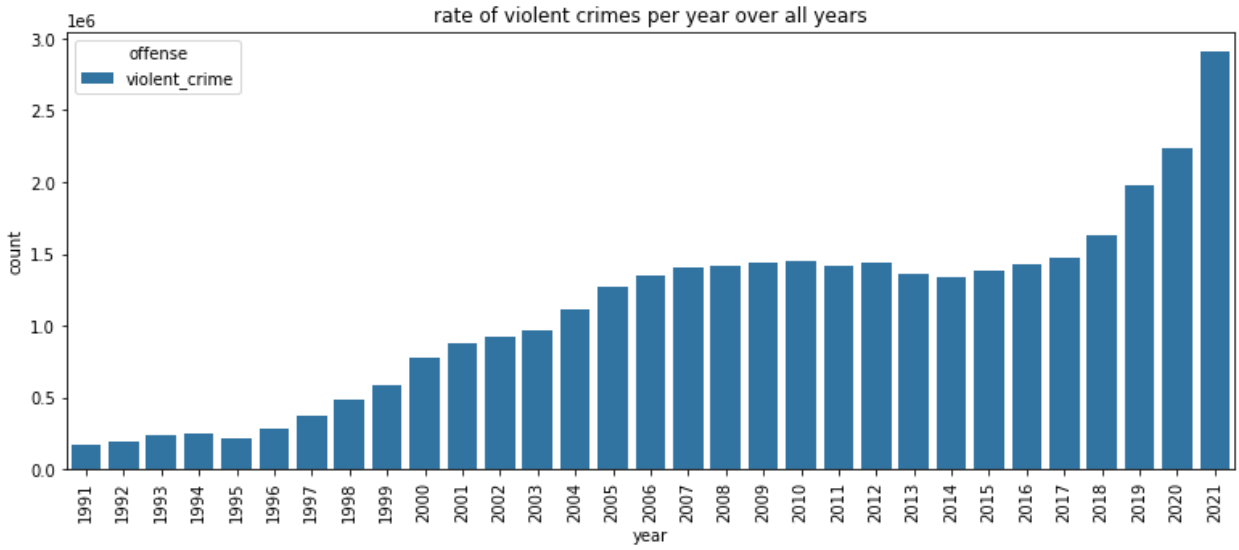
It is clear that the District of Columbia is the most violent crime state.However, the percentages of all states are near and all about 8% of the total crime.



To avoid bias in the previous graph this is to see the relation between violent and non violent crime for each state.

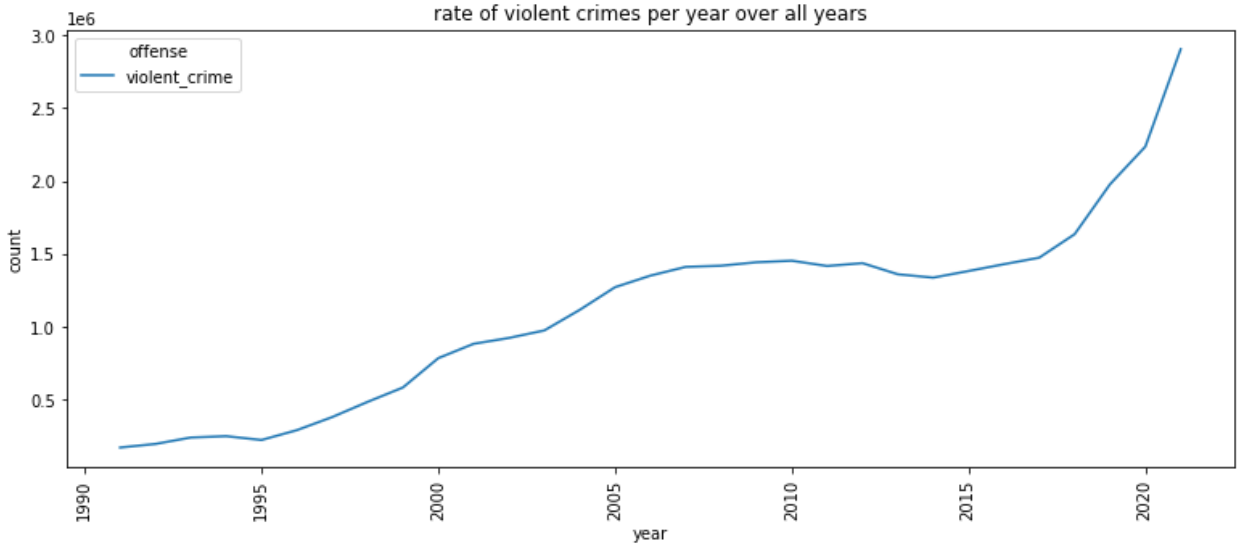
3. National homicide rates, as well as total violent crime rates per year over all years.

Violent crime rate per year over all years:

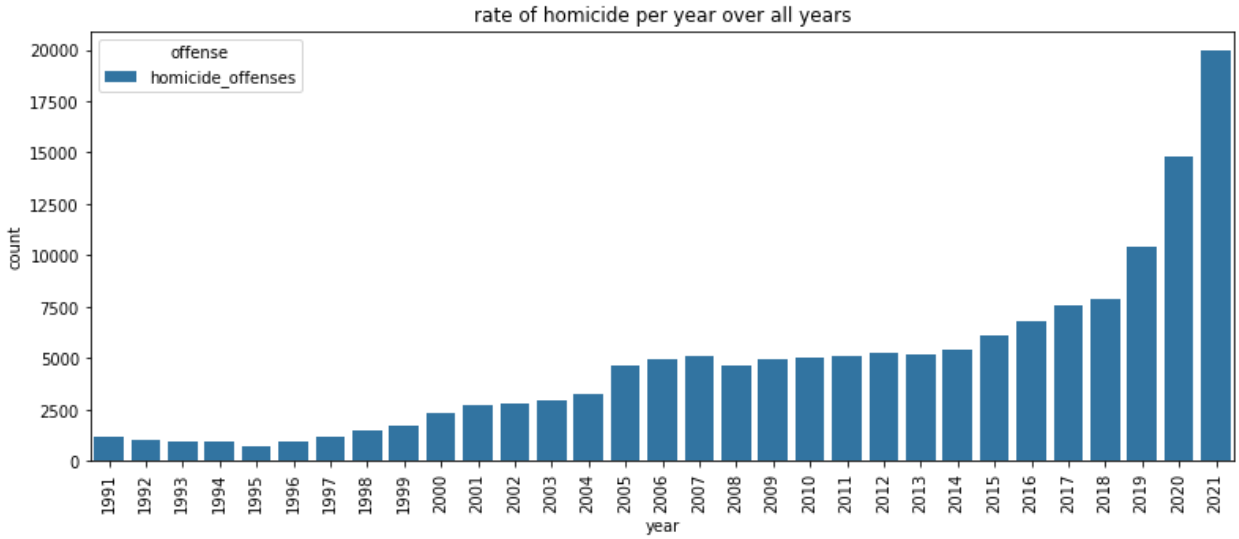


It seems that violent crimes are increasing with some relation over the years.

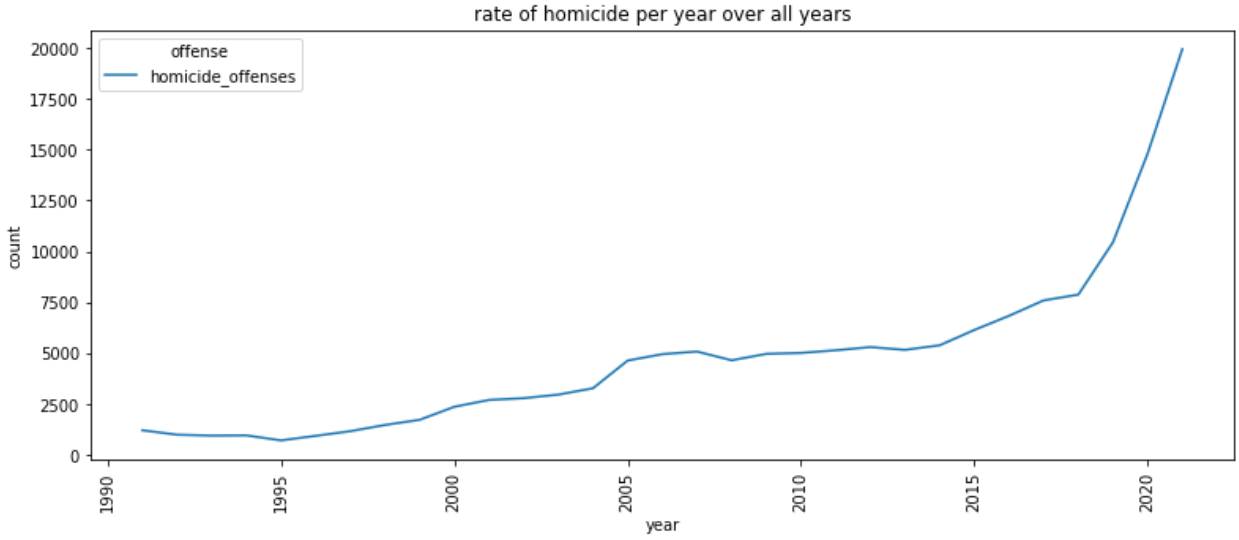
The trend of rate change as following:



Homicide rate per year over all years:



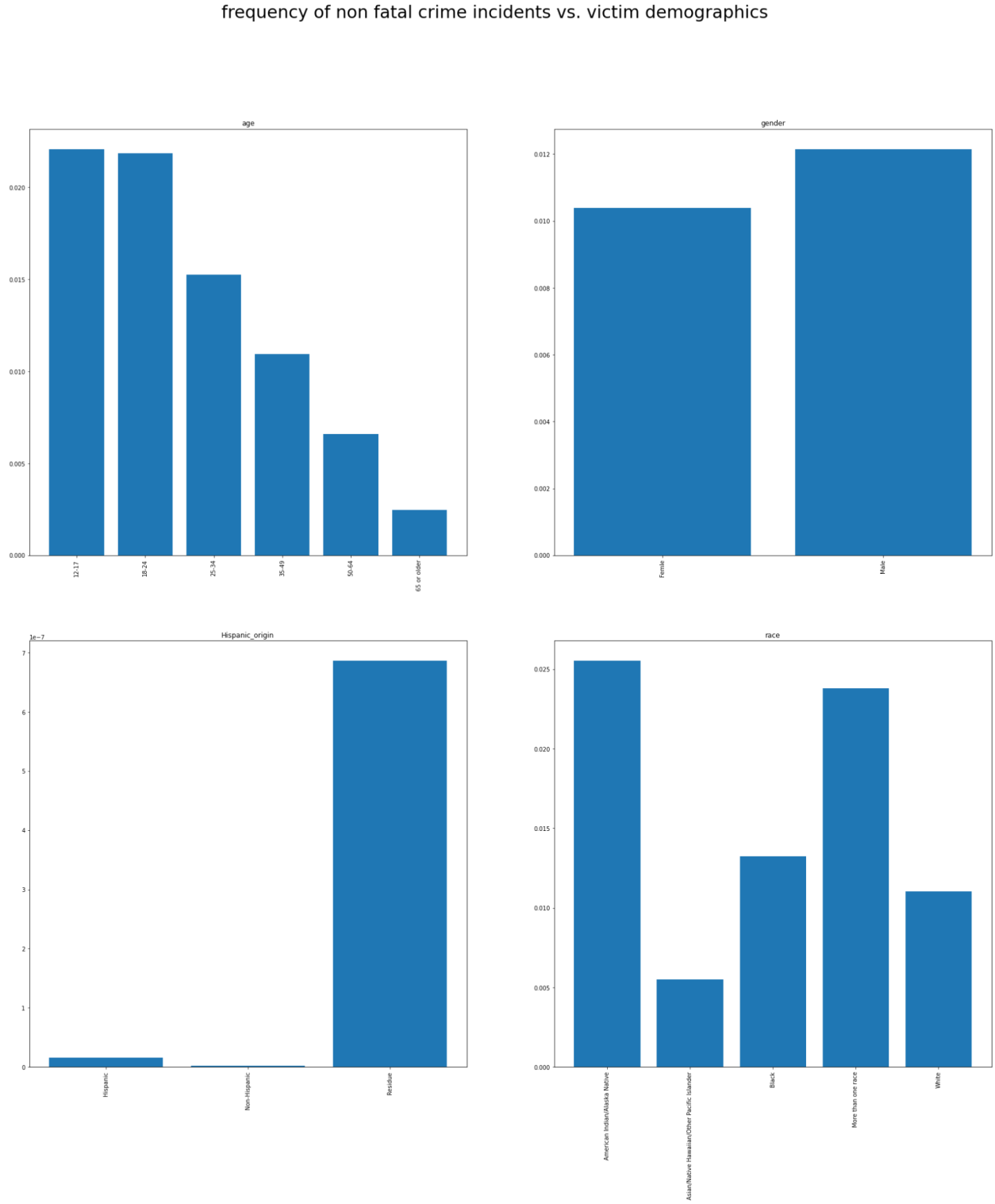
The trend that homicide rate follows:



It seems that the rate of the homicide offenses follow some increasing exponential curve over the years.

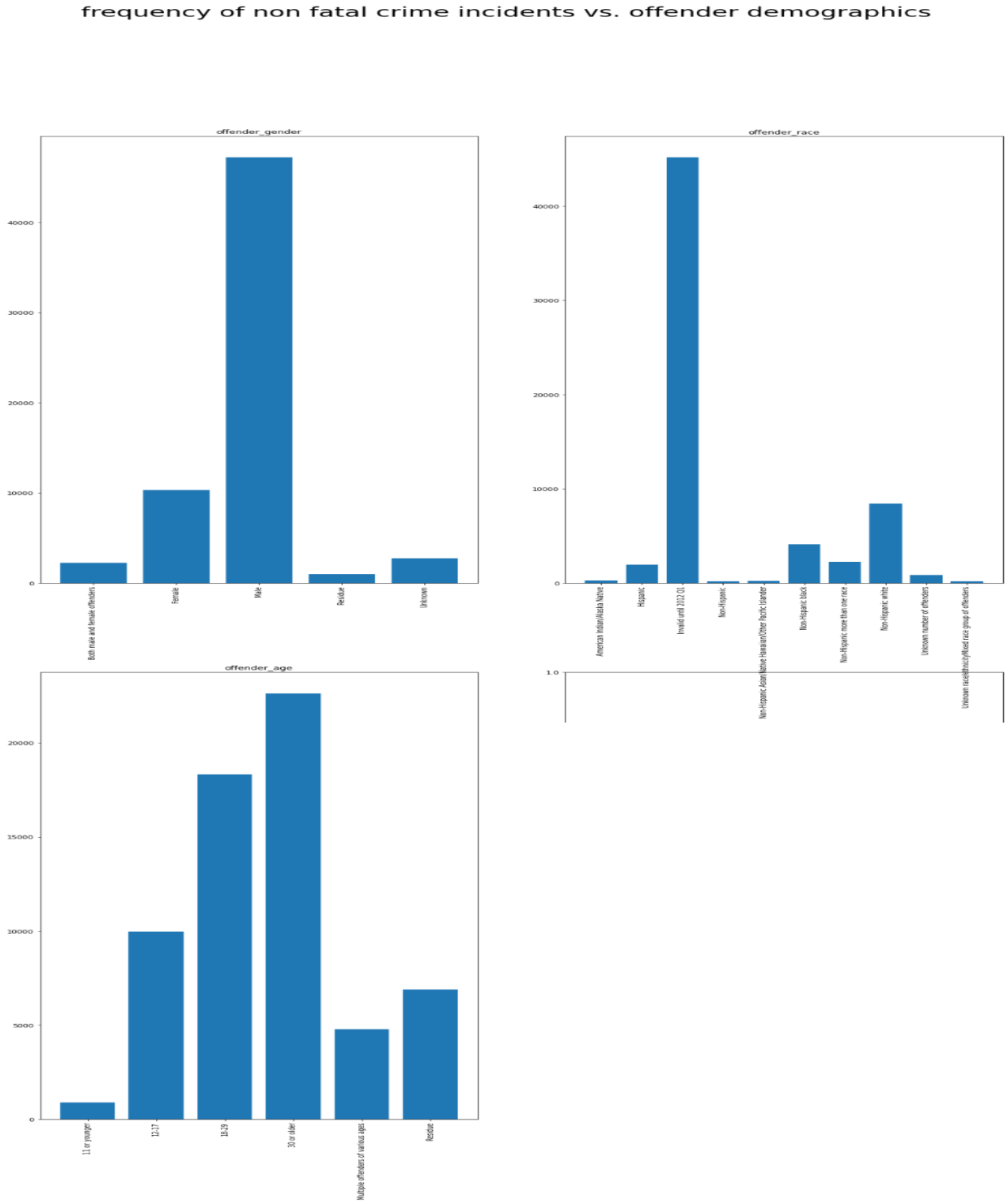
The last years have had the most rate of homicide offense and violent crime rate.

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



- **It seems that getting older is being less in risk**
- **Male face more non fatal crimes than females however the proportion is near in both.**
- **Most crimes are faced by residues who we don’t know his race, but with the proportion we know hispanic is the most.**
- **American Indian / Alaska Native faces more non-fatal crimes than others. The Asian Native are the least at risk.**

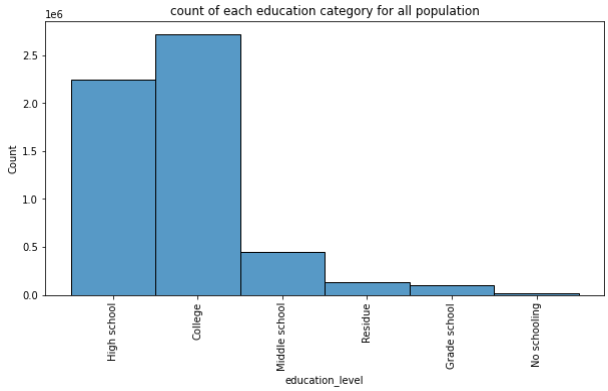
5. The frequency of non-fatal crime incidents in relation to offender demographics.

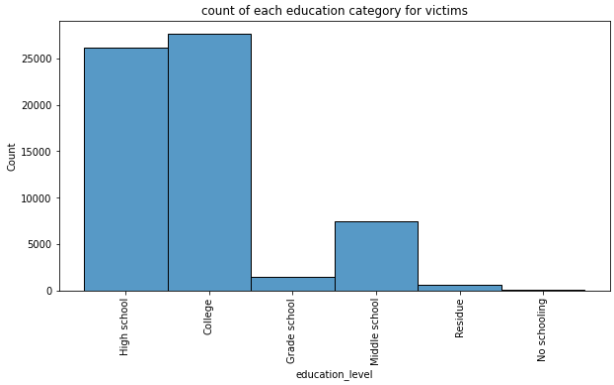


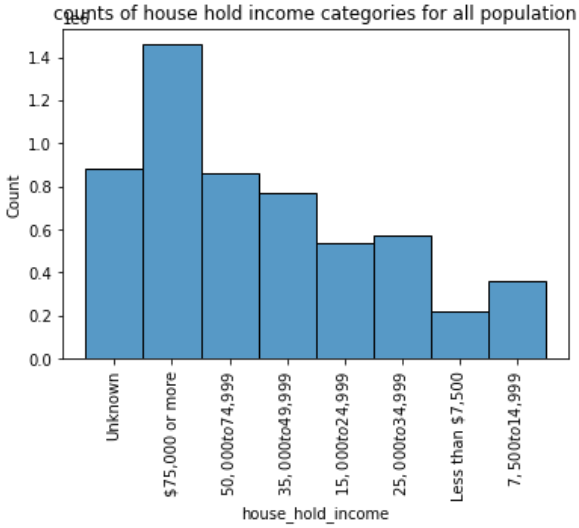
- **Males are the most offenders**

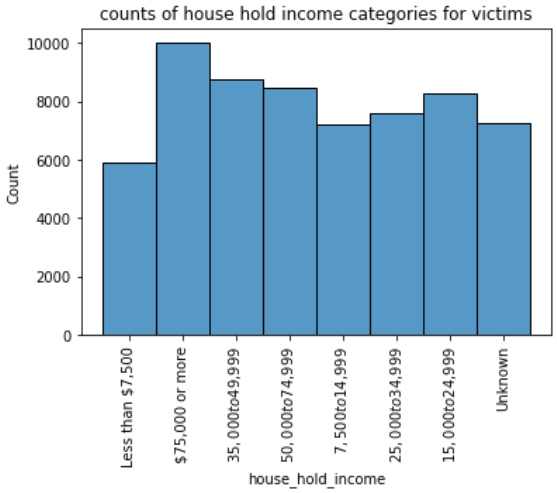
- **Non hispanic are the most offenders from 2012 till now.**
  It is a limitation from the data set to know before 2012 as shown in the curve and they are the most number of offenders.
- **30 or older are the most offenders**
  It is also a limitation to be sure about the result as there is a large part of data who we don't know their age

6. The relationship between the victim's education level, their gross household income, and their rate of victimization.
Assumption : the interest is the population of victims not all the population, so , no need to normalize the data.
But,limitation: the data is not perfectly correlated with the population dataset as shown:



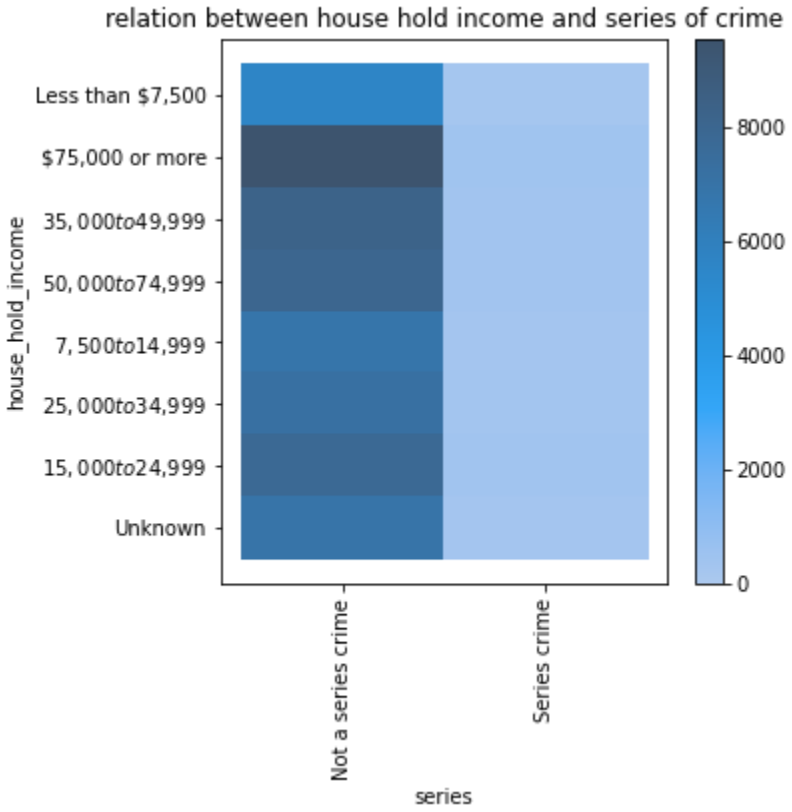






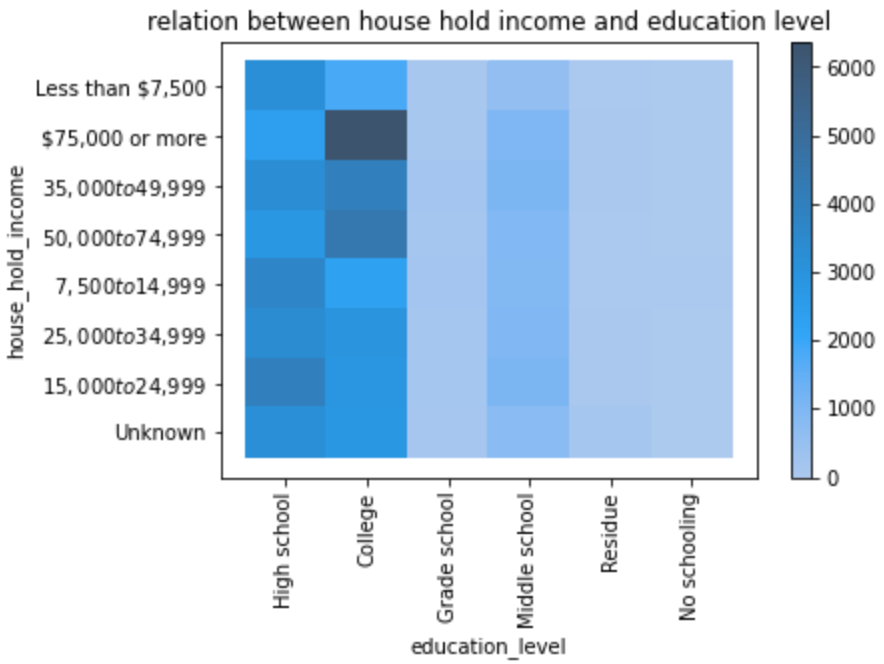
Assumed the two data sets follow the same trend:

Relationship between household income and the series of crime :



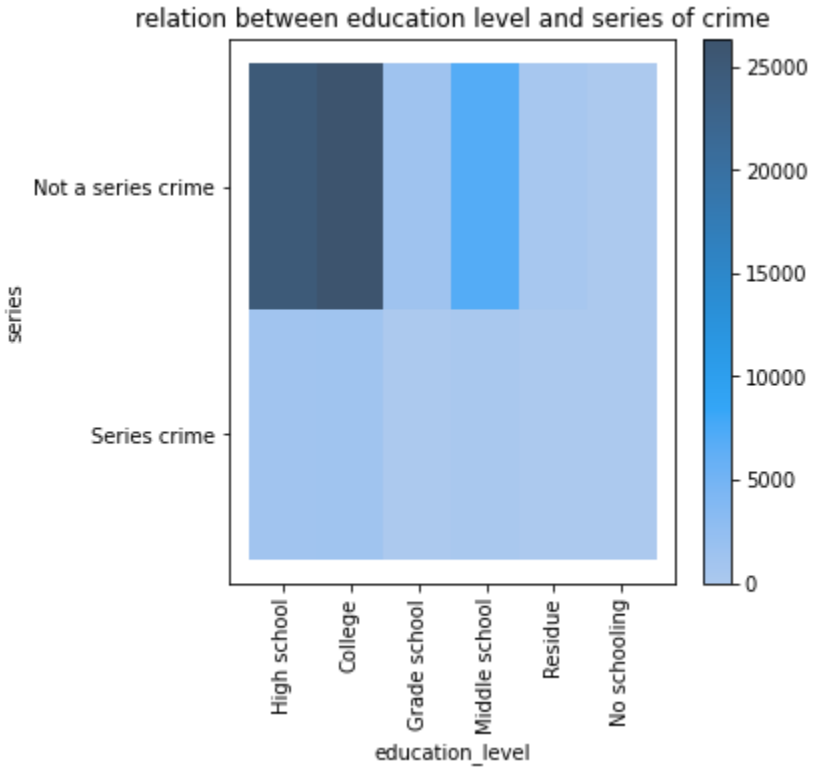
Series crimes are less than the non series in general and the large household income has probability not to have more than a crime.

Relationship between household income and the education level :



Colleges with large income are the most counts in the dataset, it seems with higher education level, your probability to have money is higher.

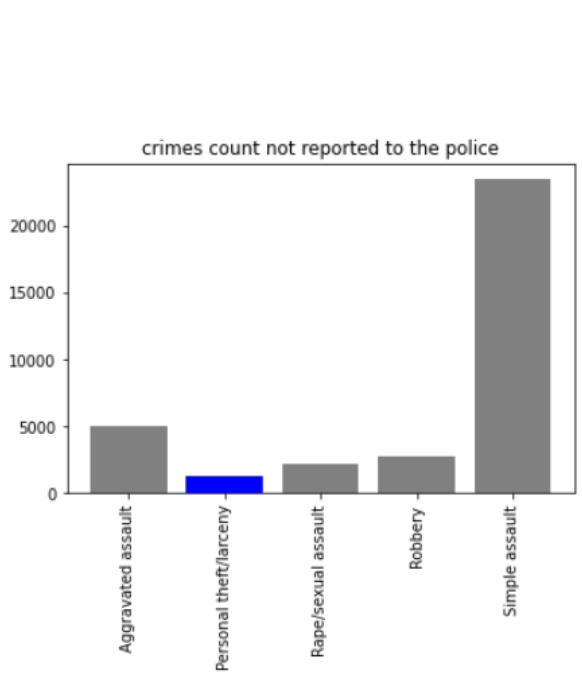
Relationship between crime rate and the education level :

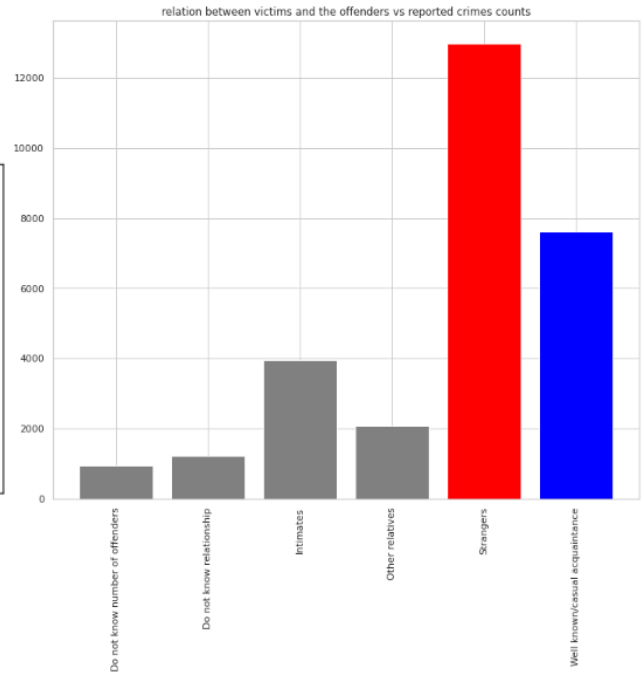


Colleges with non-series crimes have higher rates than any other one.

PART 3: Answering Questions:

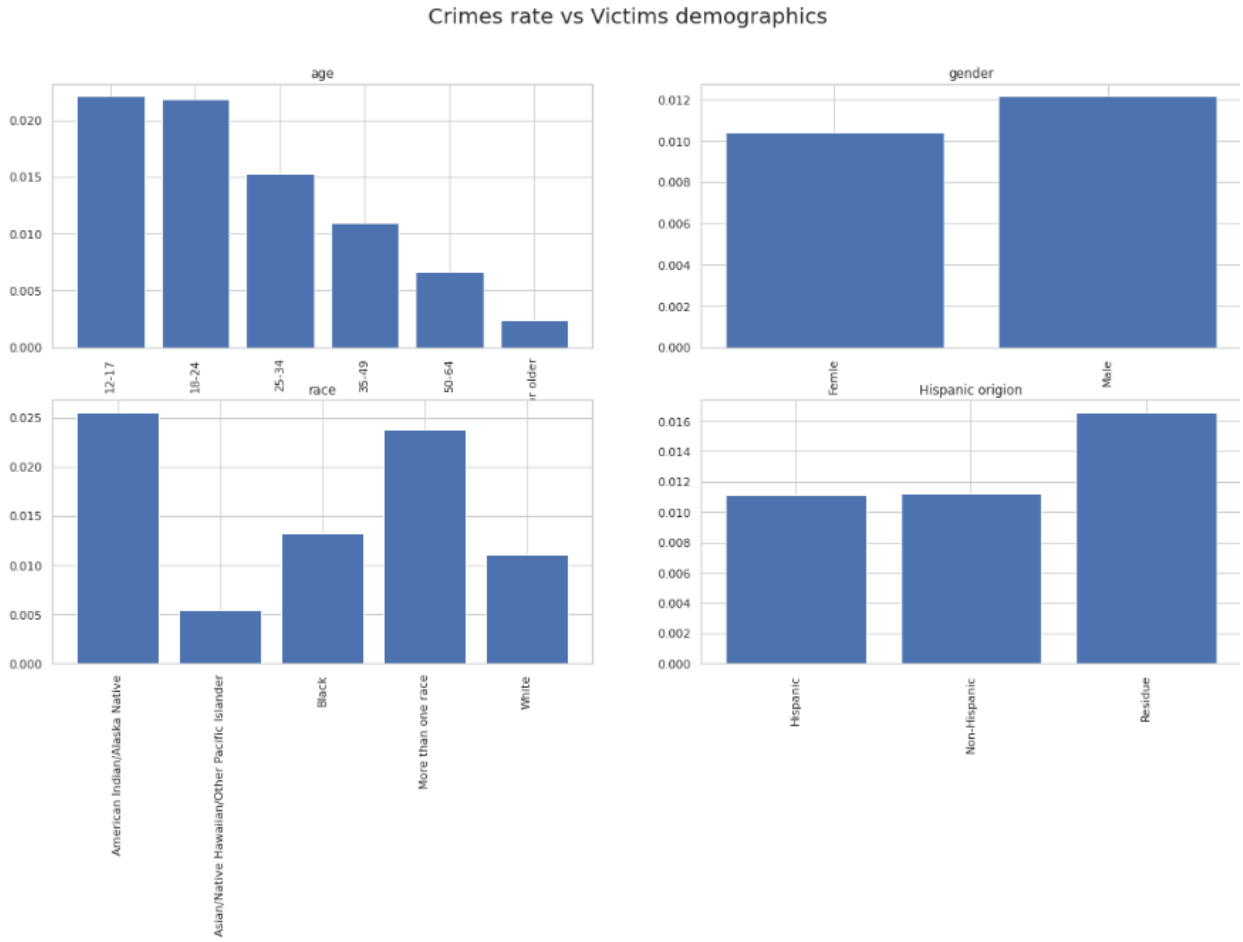
1. Which type of non-fatal crime is the most under-reported? Is there an association between the offender-victim relationship and the likelihood of a crime being reported? (reported: ie, police notified at time of occurrence)





- **Personal theft/larceny is the most under-reported non-fatal crime**
- **No association between the offender-victim relationship and the likelihood of a crime being reported**

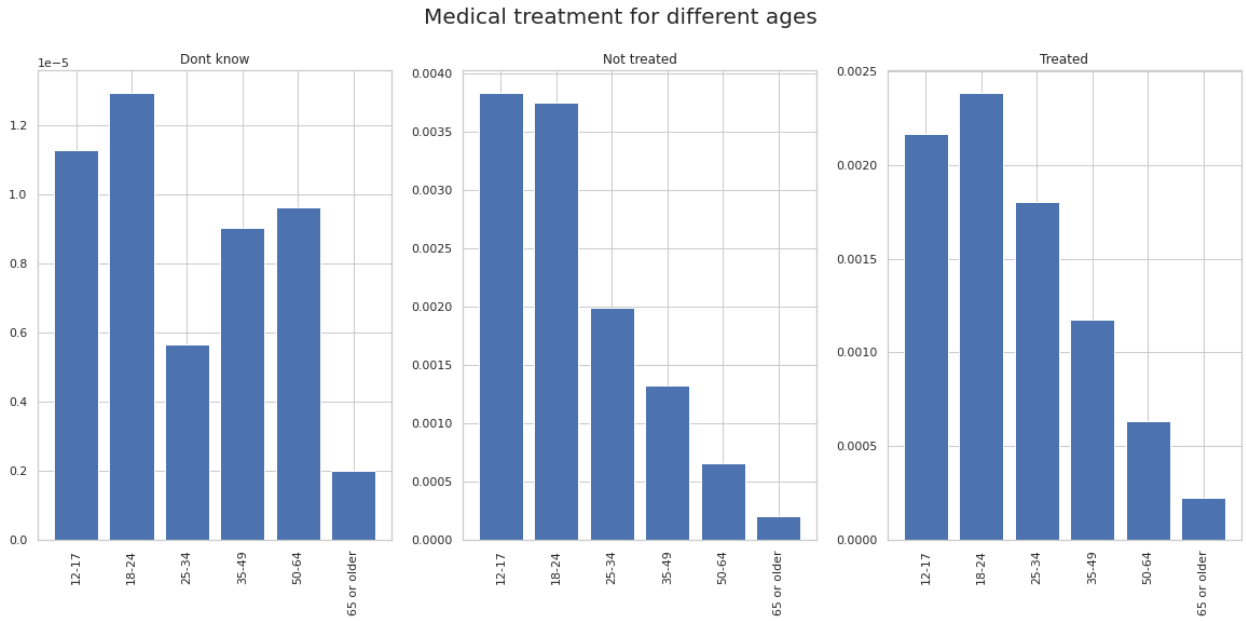
2. Who are the people (the demographic segment) that appear to be most at risk of violent victimization? Who is the least at risk?

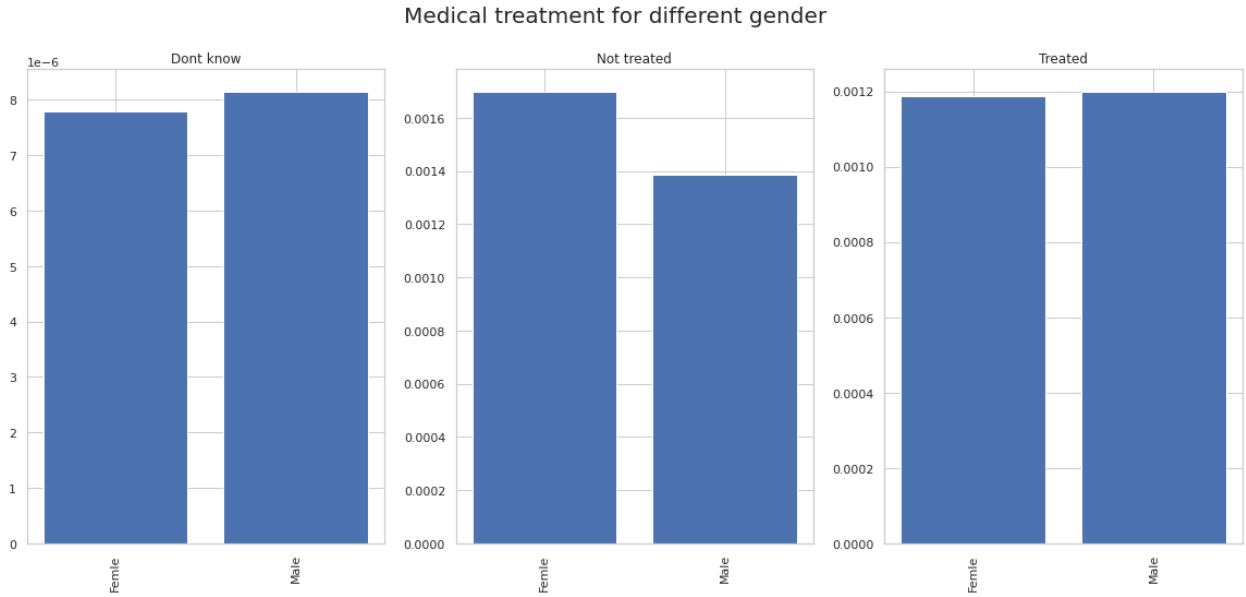


Most risk is found with:

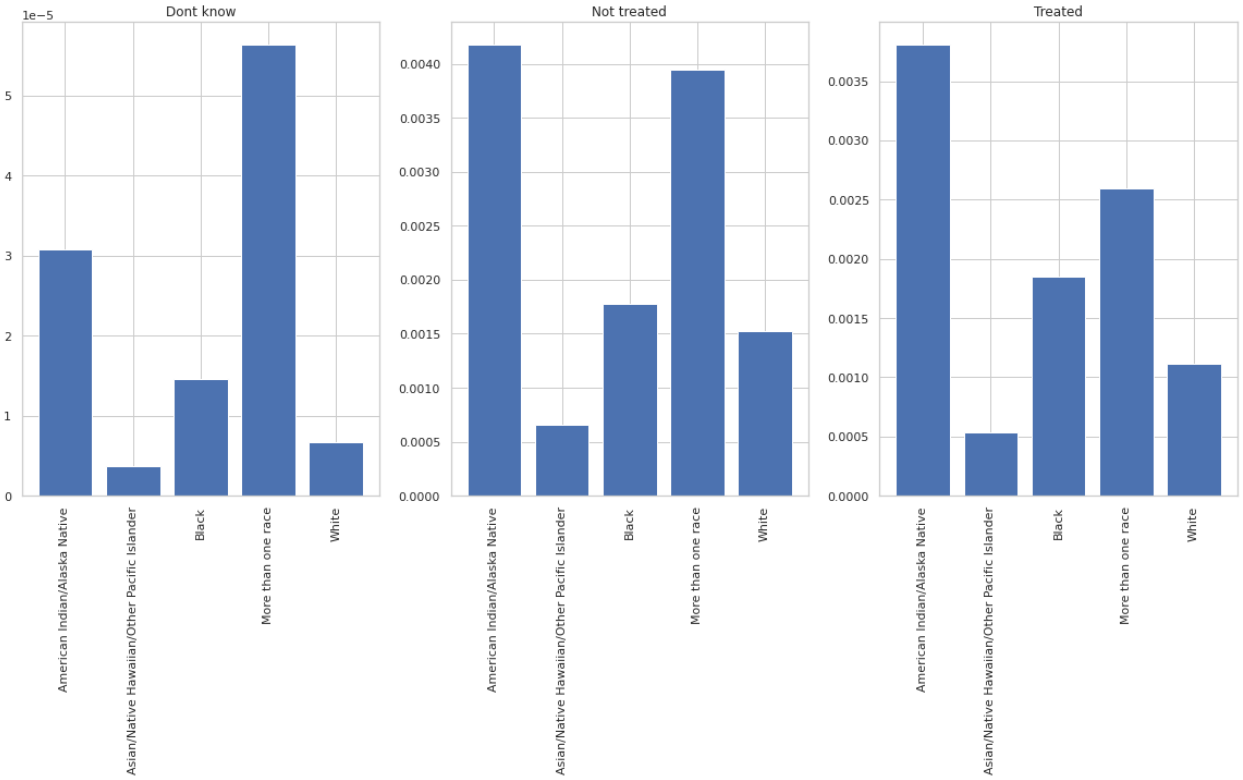
- **Younger people (age)**
- **American/indian (race)**
- **Hispanic category is not affecting the victimization risk**
- **Gender is not affecting the victimization risk**

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 like

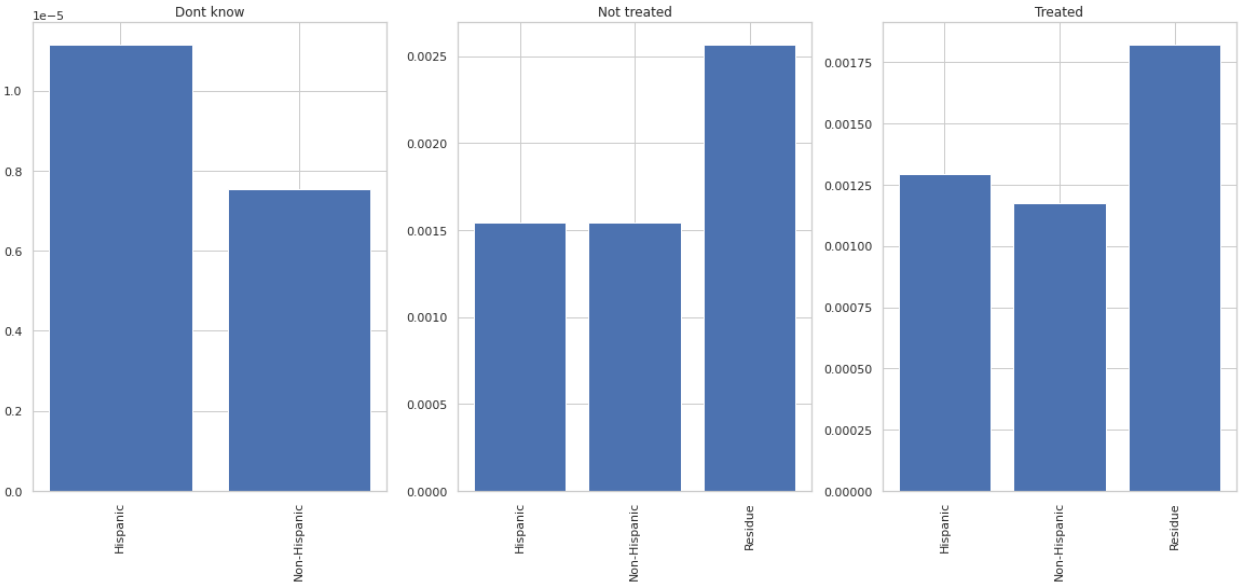




Medical treatment for different race



Medical treatment for different hispanic origion



Age:

- **‘18-24’** are the most treated category
- **‘65 or older’** are the least treated category

Gender:

- There is a balance between males and females in treatment

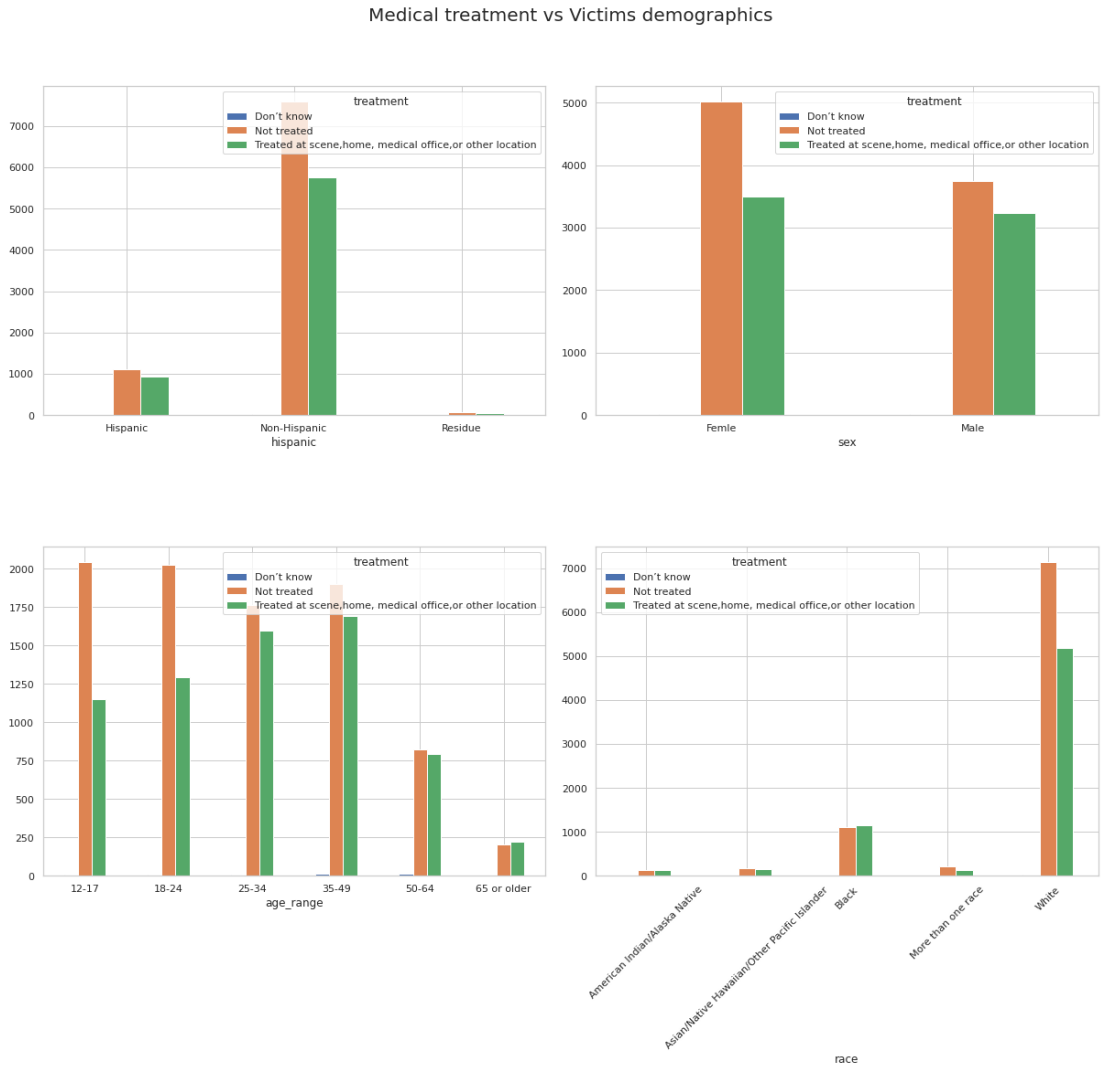
Race:

- **American indian/Alaska native’** are the most treated category
- **‘Asian/native Hwaiian/Other pacific islander ’** are the least treated category

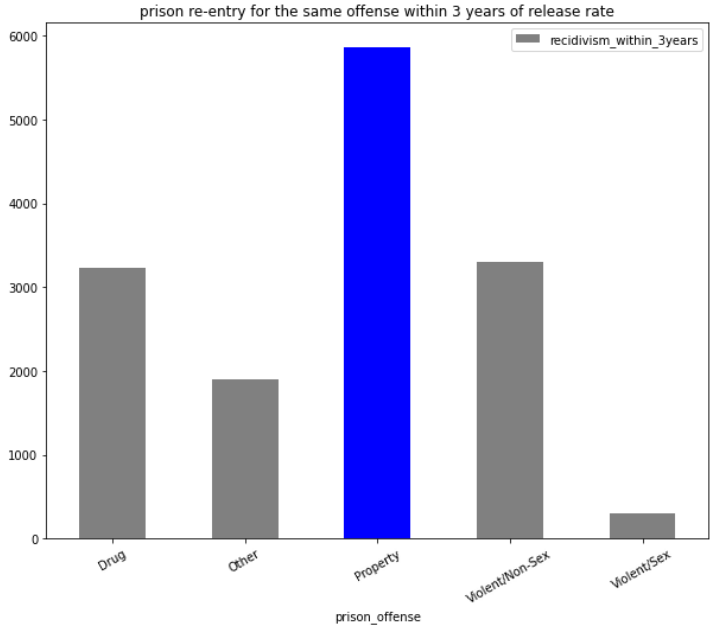
Hispanic Origin:

- **‘Hispanic ’** are the most treated category
- **‘Non Hispanic** are the least treated category

To see the ration between the treatment and un treatment in each major

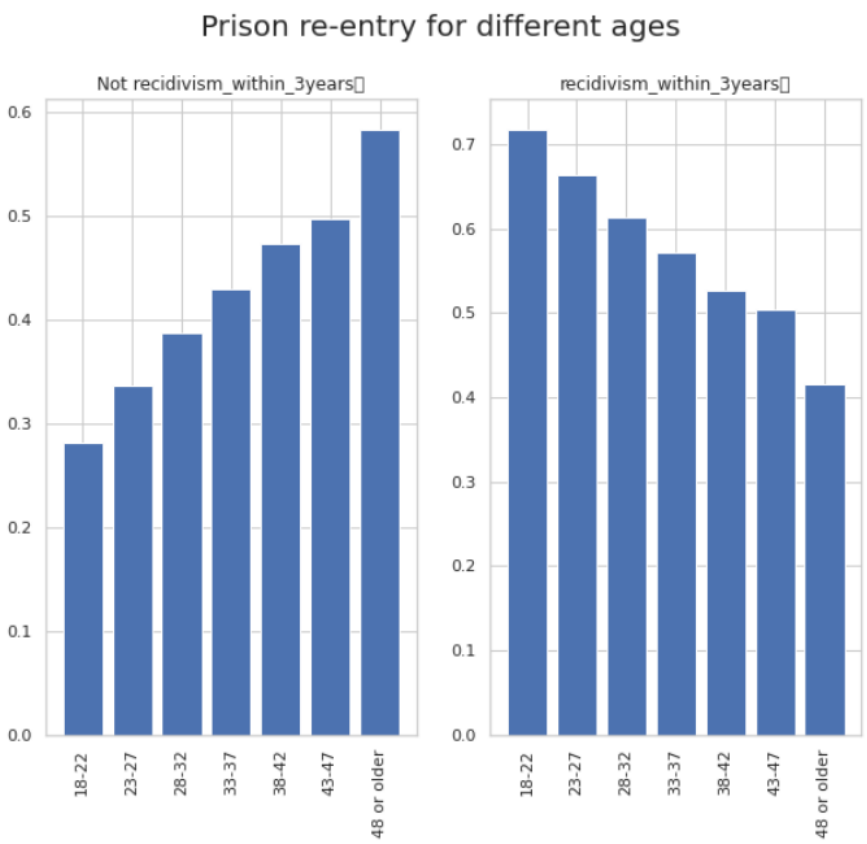


4. Which class of crimes is associated with the highest rate of same-offense-recidivism; i.e. prison re-entry for the same offense within 3 years of release?



Property offense prisoners were found to re-enter the prison within 3 years from the release more frequently than other offenses

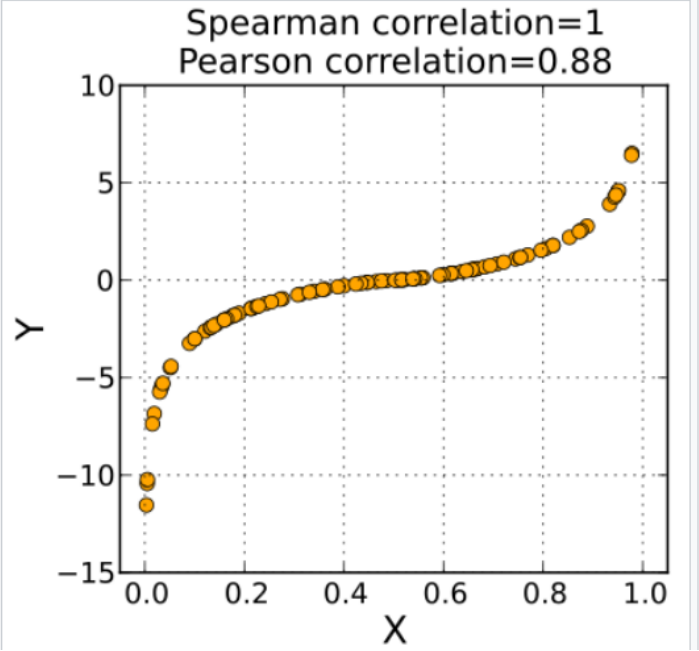
5. Are prisoners who are younger at the time of release more or less likely to reoffend than those who are older?



Prisoners who are younger at the time of release more likely to reoffend than those who are older

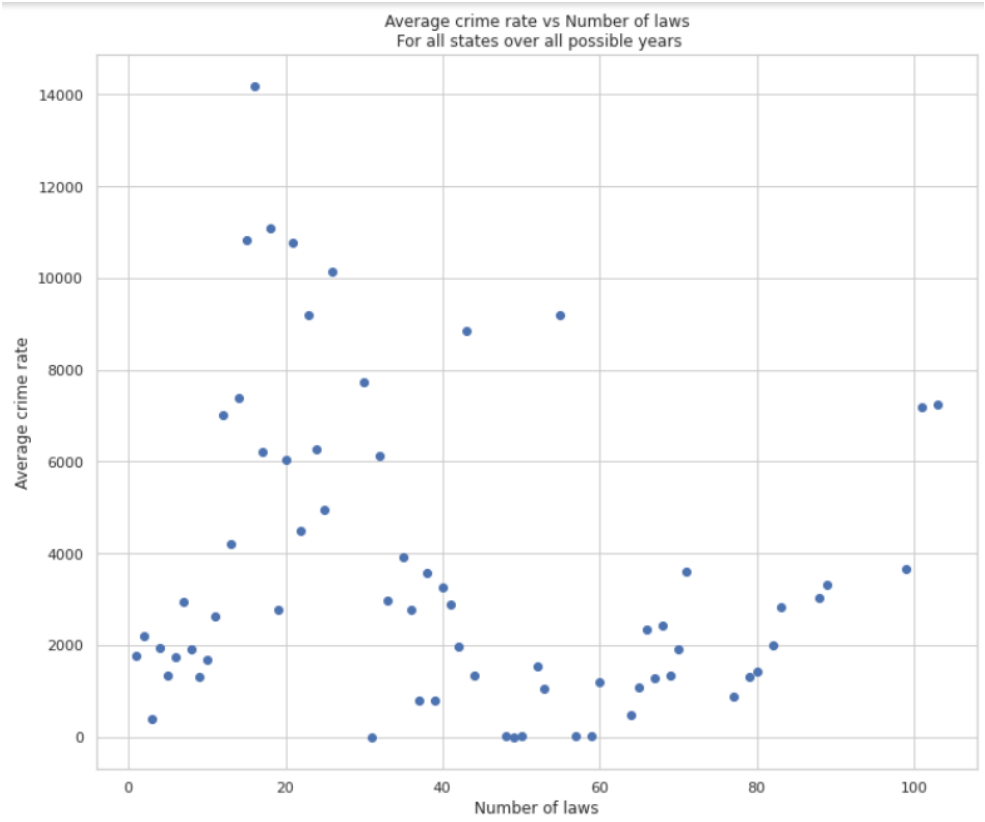
PART 4: Hypothesis Testing:

- **The statistical test we used to measure the correlation *Spearman correlation coefficient*.**
- **The main difference between Spearman and Pearson correlation is that Pearson is limited to measuring typical linear relationship between variables, while Spearman measures how far could the relationship between the variables be described by a monotonic function, hence it tests the monotonicity in general not just the linear one.**
- **Here we can find that spearman indicates perfect positive correlation (ρ=1) while pearson correlation (ρ=0.88) due to non-linearity at tails.**



- **<u>Claim 1:</u> "U.S. states that implement stricter firearm control laws, have lower crime rates on average"**
- **H0: "U.S. states that implement stricter firearm control laws, have the *same or larger* violent crime rates on average"**
- **Ha: "U.S. states that implement stricter firearm control laws, have *lower* violent crime rates on average"**

```
Correlation=-0.18, p-value=0.073
```

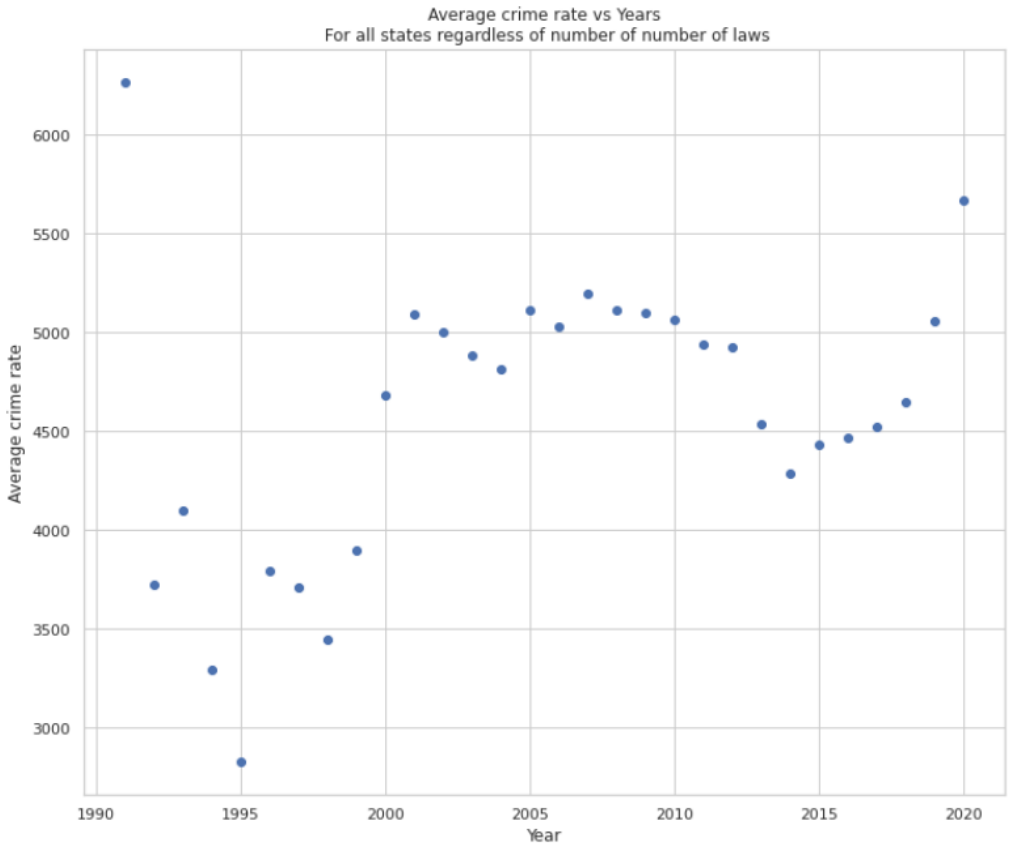


- **From the results we found that p-value=0.0727 >0.05 hence we cannot reject the null-hypothesis. Moreover, the correlation value itself is weak=-0.18 and the scatter plot emphasises this weak correlation.**

  Hence we cannot reject the null-hypothesis i.e the claim is invalid.

- **<u>Claim 2</u>:"U.S. states encounter an *increasing* crime rates on average over all the years of the study regardless of how strict is the applied laws"**
- **H0: "U.S. states encounter the *same or decreasing* crime rates on average over all the years of the study regardless of how strict is the applied laws"**
- **Ha: "U.S. states encounter an *increasing* crime rates on average over all the years of the study regardless of how strict is the applied laws"**

```
Correlation=0.36, p-value=0.025
```



- **From the results we found that p-value=0.025 < 0.05, hence we can safely reject the null-hypothesis. We can colnclude that perhaps this increase or crime rates over time is the reason why the crime rates does not decline even when appplying more strict laws.**

PART 5: Regression Analysis:

predicts the offender's supervision risk score based on :

- All prior convictions.

- Offender's race.

- Offender's gang affiliation.

- Offender's age at release.

ONE HOT ENCODED FEATURES:

OLS Regression Results

Dep. Variable:	supervision_risk_score_first	R-squared:	0.311
Model:	OLS	Adj. R-squared:	0.311
Method:	Least Squares	F-statistic:	477.1
Date:	Sun, 08 Jan 2023	Prob (F-statistic):	0.00
Time:	20:26:31	Log-Likelihood:	-53260.
No. Observations:	25360	AIC:	1.066e+05
Df Residuals:	25335	BIC:	1.068e+05
Df Model:	24		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
intercept	5.8493	0.079	73.716	0.000	5.694	6.005
race_BLACK	-0.1001	0.026	-3.874	0.000	-0.151	-0.049
age_at_release_18-22	4.0326	0.059	67.981	0.000	3.916	4.149
age_at_release_23-27	3.5568	0.047	75.921	0.000	3.465	3.649
age_at_release_28-32	2.8511	0.045	62.822	0.000	2.762	2.940
age_at_release_33-37	2.0972	0.046	45.640	0.000	2.007	2.187
age_at_release_38-42	1.4709	0.050	29.705	0.000	1.374	1.568
age_at_release_43-47	0.9282	0.051	18.215	0.000	0.828	1.028
gang_affiliated_False	-0.4103	0.034	-11.973	0.000	-0.477	-0.343
prior_conviction_episodes_2_False	-0.2609	0.030	-8.656	0.000	-0.320	-0.202
prior_conviction_episodes_5_False	-0.3408	0.033	-10.474	0.000	-0.405	-0.277
prior_conviction_episodes_6_False	0.2594	0.050	5.218	0.000	0.162	0.357
prior_conviction_episodes_7_False	-0.3546	0.037	-9.552	0.000	-0.427	-0.282
prior_conviction_episodes_4_0	-0.6633	0.037	-18.136	0.000	-0.735	-0.592
prior_conviction_episodes_4_1	-0.4031	0.037	-10.983	0.000	-0.475	-0.331
prior_conviction_episodes_3_0	-1.5986	0.043	-36.762	0.000	-1.684	-1.513
prior_conviction_episodes_3_1	-1.0279	0.043	-23.908	0.000	-1.112	-0.944
prior_conviction_episodes_3_2	-0.6192	0.046	-13.321	0.000	-0.710	-0.528
prior_conviction_episodes_1_0	0.6435	0.045	14.193	0.000	0.555	0.732
prior_conviction_episodes_1_1	0.5011	0.043	11.583	0.000	0.416	0.586
prior_conviction_episodes_1_2	0.4245	0.044	9.706	0.000	0.339	0.510
prior_conviction_episodes_1_3	0.2067	0.047	4.401	0.000	0.115	0.299
prior_conviction_episodes_0	-0.0063	0.048	-0.132	0.895	-0.100	0.087
prior_conviction_episodes_1	-0.1021	0.043	-2.388	0.017	-0.186	-0.018
prior_conviction_episodes_2	-0.1040	0.042	-2.471	0.013	-0.187	-0.022

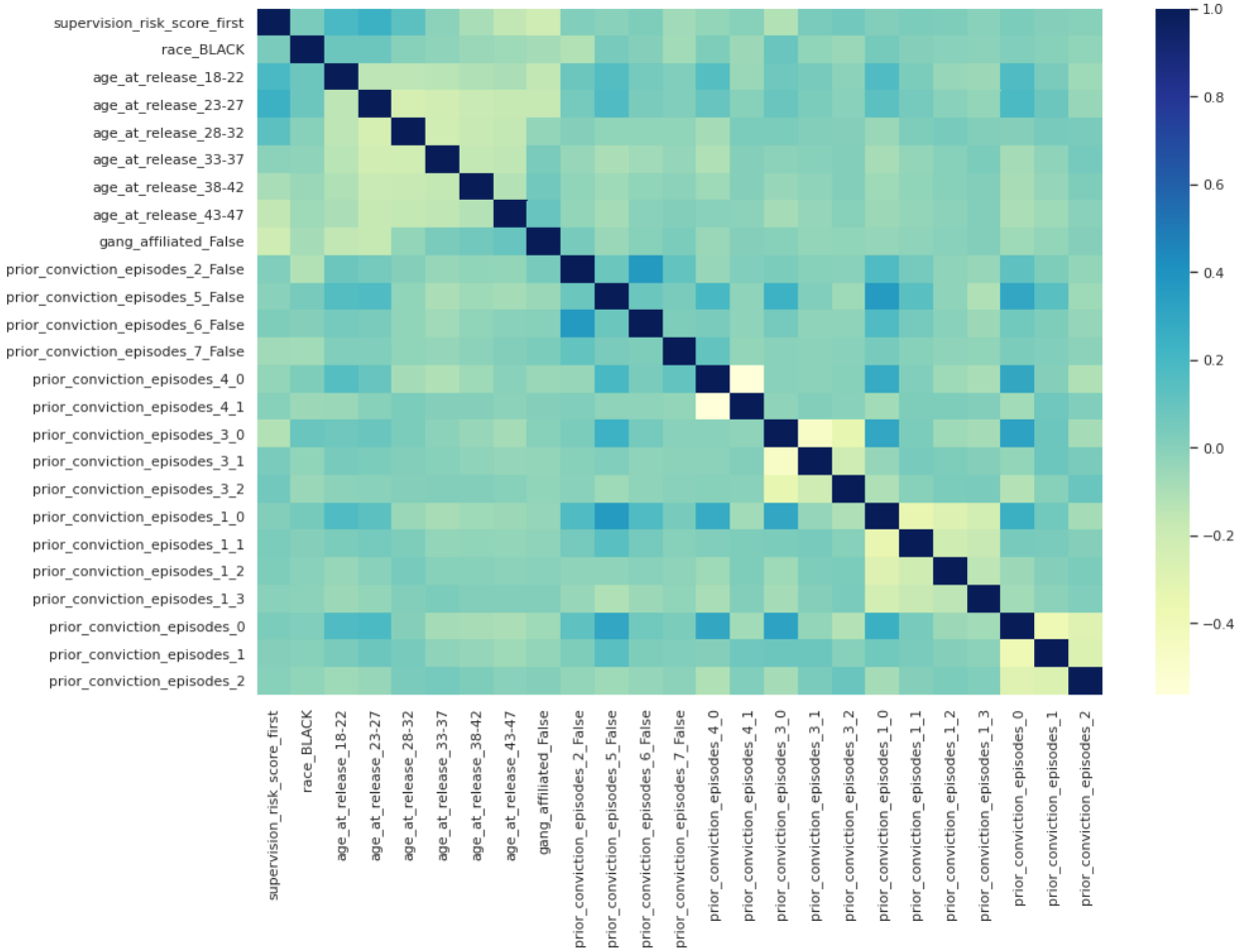
Omnibus:	122.618	Durbin-Watson:	1.928
Prob(Omnibus):	0.000	Jarque-Bera (JB):	173.008
Skew:	0.025	Prob(JB):	2.70e-38
Kurtosis:	3.401	Cond. No.	20.6

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

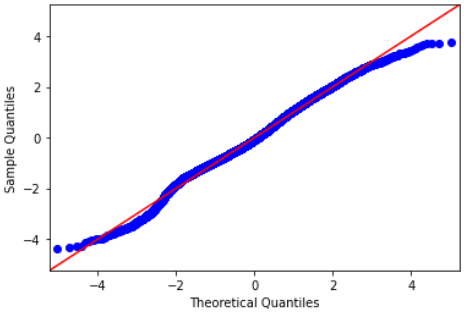
- **All parameters are statistically significant except prior_convection_episoide_0, P_value > 0.05**

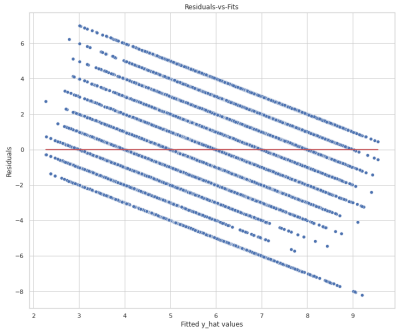
Also, the value of its beta is slightly small which is logical to us. All are good predictors except this variable.

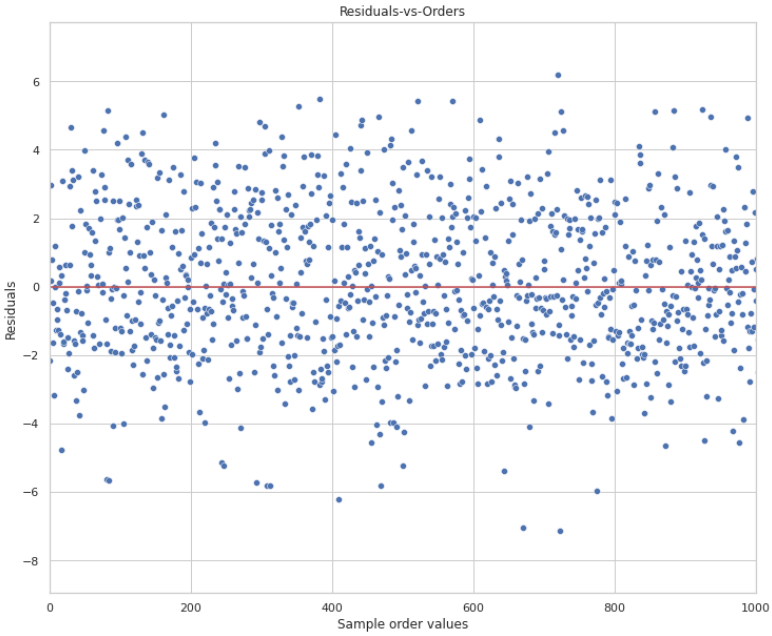
- **There is no high correlation between the variables as shown:**



The model is not good for prediction as the r_squared are about 0.3.However the model follows all of our assumptions of homocidicity with 0 mean and constant variance, but the residual with y_hat is not uniform distribution around the mean , but there is no correlation between the errors.


Sample Quantiles
Theoretical Quantiles


Residuals-vs-Fits
Residuals
Fitted y_hat values


Residuals-vs-Orders
Residuals
Sample order values

LABEL ENCODED FEATURES:

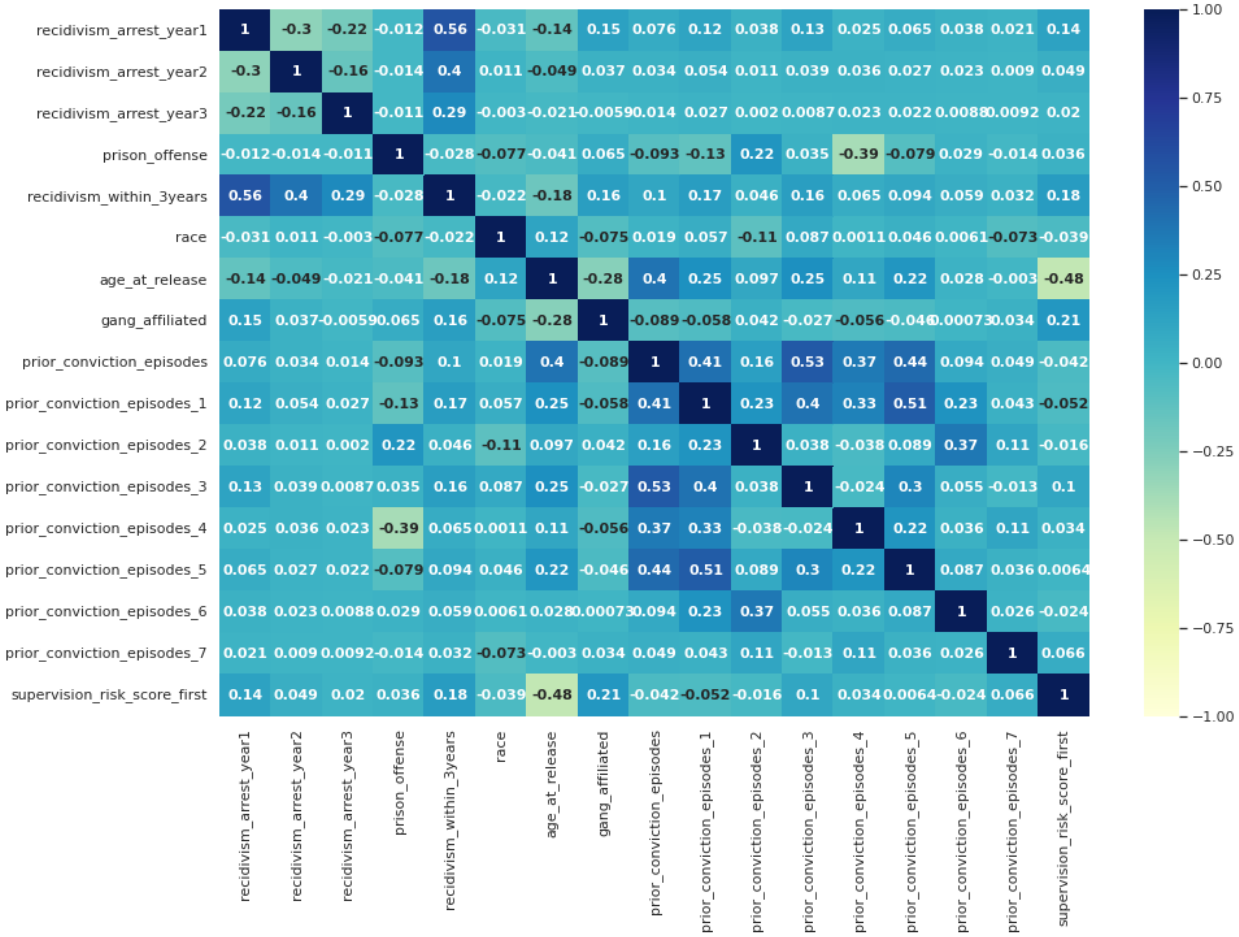
OLS Regression Results

Dep. Variable:	supervision_risk_score_first	R-squared:	0.309
Model:	OLS	Adj. R-squared:	0.309
Method:	Least Squares	F-statistic:	1031.
Date:	Sun, 08 Jan 2023	Prob (F-statistic):	0.00
Time:	20:26:34	Log-Likelihood:	-53300.
No. Observations:	25360	AIC:	1.066e+05
Df Residuals:	25348	BIC:	1.067e+05
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
intercept	7.1567	0.031	231.350	0.000	7.096	7.217
race	0.0980	0.026	3.810	0.000	0.048	0.148
age_at_release	-0.6799	0.008	-89.407	0.000	-0.695	-0.665
gang_affiliated	0.4063	0.034	11.868	0.000	0.339	0.473
prior_conviction_episodes	-0.0004	0.015	-0.023	0.982	-0.031	0.030
prior_conviction_episodes_1	-0.1562	0.011	-14.460	0.000	-0.177	-0.135
prior_conviction_episodes_2	0.2620	0.030	8.690	0.000	0.203	0.321
prior_conviction_episodes_3	0.5302	0.014	38.292	0.000	0.503	0.557
prior_conviction_episodes_4	0.3288	0.018	18.408	0.000	0.294	0.364
prior_conviction_episodes_5	0.3485	0.032	10.767	0.000	0.285	0.412
prior_conviction_episodes_6	-0.2600	0.050	-5.232	0.000	-0.357	-0.163
prior_conviction_episodes_7	0.3557	0.037	9.583	0.000	0.283	0.428

We got about the same results as the one hot encoding , the prior convection_eposodes are the one who is a bad variable.

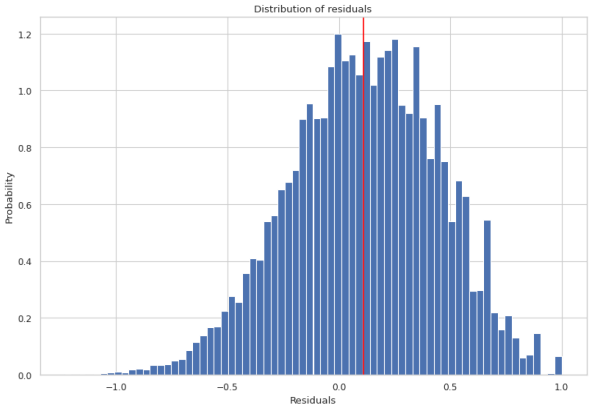
The correlation between the variables

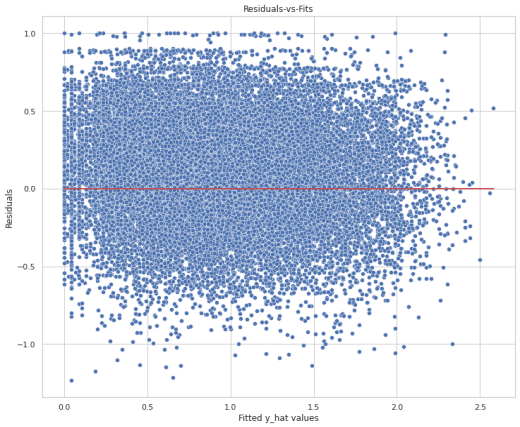


The model curves are the same as the one hot encoding

Trying to scale the variables to see the response of the model:

- **Model prediction is better → r_squared = 0.6**
  And the error curve with y_hat are normally distributed around the mean, but the model is biased to 0.1





PART 6: Bonus Task:

We fit a neural network model with 2 neurons and one layer

```
interest =
['employment_exempt','prior_arrest_episodes_drug','prior_arrest_episodes_violent','prior_arrest_episodes_misd','program_attendances','violations_instruction','gender','dependents','education_level','prison_years','prison_offense','recidivism_within_3years','race', 'age_at_release',
'gang_affiliated'
,'prior_conviction_episodes','prior_conviction_episodes_1','prior_conviction_episodes_2','prior_conviction_episodes_3','prior_conviction_episodes_4'
,'prior_conviction_episodes_5','prior_conviction_episodes_6','prior_conviction_episodes_7','supervision_risk_score_first']
```

The accuracy we got to predict the `recidivism_within_3years` is 67%

Conclusion:

From the analysis to the dataset some improvements could be added to the US institutes which care about crime. For example,additional care to the states which have more crimes, and they could predict who will return to the crime or not depending on his information. Modify the rules based on the crime rates.