Analyzing U.S. Crime Data

Business Report

CIE 457

Khaled Elbastawisy 201-800-029

Mohamed Ahmed 201-801-898

Moataz Mohamed 201-801-903

Introduction

The accurate analysis of crime data is crucial for understanding and addressing the complex issues surrounding crime in the United States. By using statistical methods to examine crime trends and patterns, we can gain valuable insights into the factors that contribute to crime and develop more effective strategies for crime prevention and intervention. In this project, we delve into the rich data on crime in the US to uncover hidden trends and relationships, with the goal of informing policy and decision-making at the local, state, and national levels. The importance of this work cannot be overstated, as it has the potential to make a real impact on public safety and the well-being of communities across the country.

Exploratory Data Analysis Results

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

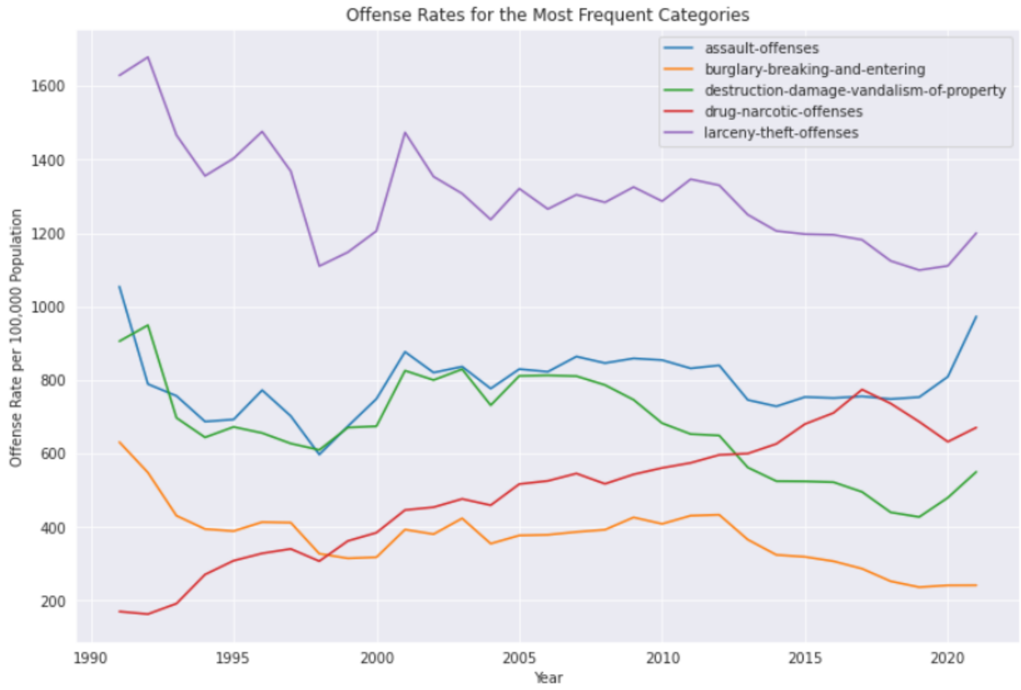


Figure 1

Commentary

- The line plot of offense rates for the top five most frequent categories over time provides valuable insights into the trends in criminal offenses in the United States. The plot shows that all the offense categories, with the exception of drug/narcotics offenses, have generally decreasing trends over the years, with noticeable variability in their rate of change.
- One notable trend is the decrease in larceny/theft offense rates over the past 30 years. Although the rates for this category were alarmingly high in the early 1990s, they have decreased by almost 25% over time. It is worth noting that larceny/theft offenses show a significant difference relative to all other offenses and have consistently had the highest rates among all categories.
- Assault and vandalism offenses also showed a decrease in rates in the early 1990s, followed by a relatively stable trend over time. However, it is important to note that assault offenses have recently seen a spike in rates, putting them at a close rate to larceny/theft. It is important to understand the factors driving this trend and the potential impact on communities and law enforcement resources.
- On the other hand, destruction/vandalism offenses have seen a significant decrease in rates between 2006 and 2018. This could be due to a variety of factors, such as improvements in economic conditions, social conditions, or law enforcement practices. Further analysis of the data and consideration of the specific context and conditions in the affected areas will be necessary to fully understand the reasons behind this trend.
- Narcotics offenses (out of all other categories) have seen a significant increase over the years, overtaking vandalism offenses in terms of rates. It is worth noting that the recent lockdown may have played a role in this trend, as people may be more likely to turn to illicit substances for recreational purposes or to cope with stress.
- Burglary/breaking and entering offenses rates went significantly down over time. It appears that efforts have been made recently to make people feel safer at home and implement sophisticated alarm systems. This has led to a noticeable decrease in burglary rates in recent years, suggesting that these efforts may be effective in reducing this type of offense.

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

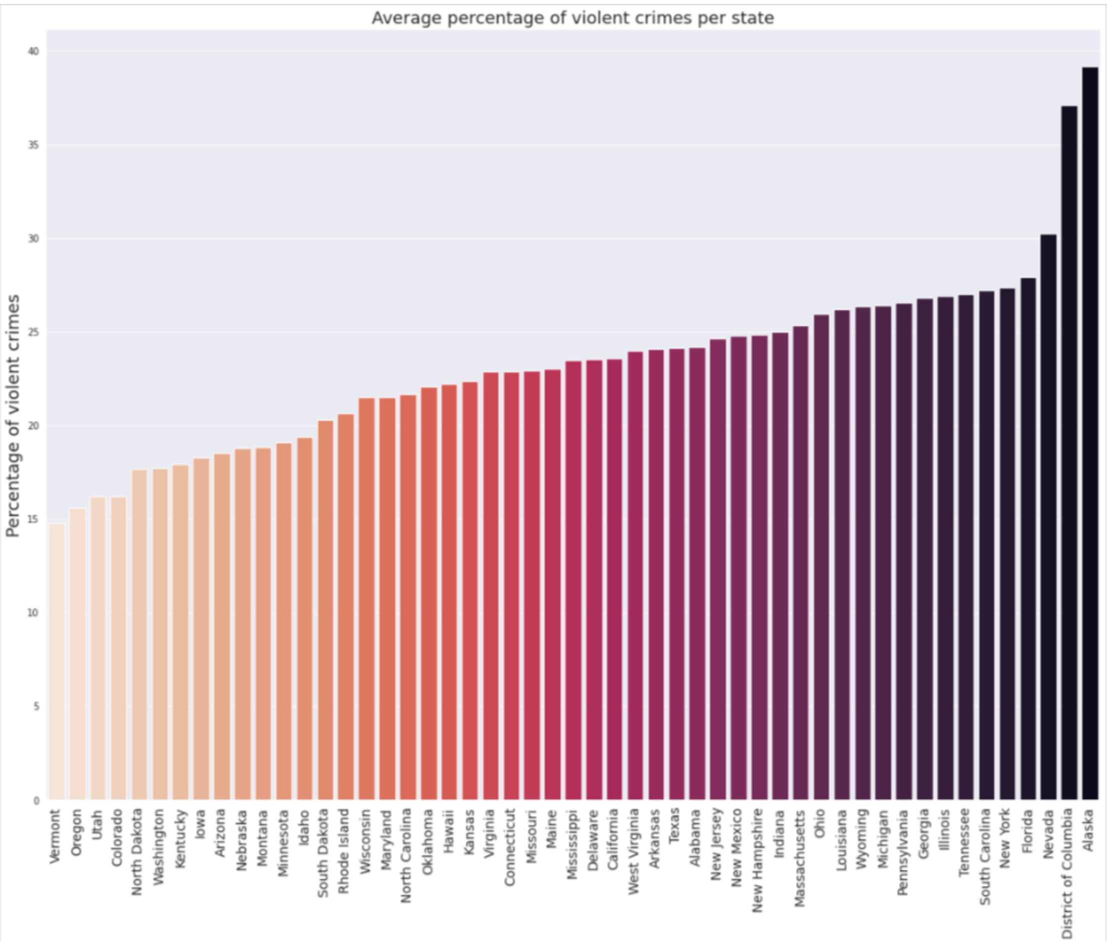


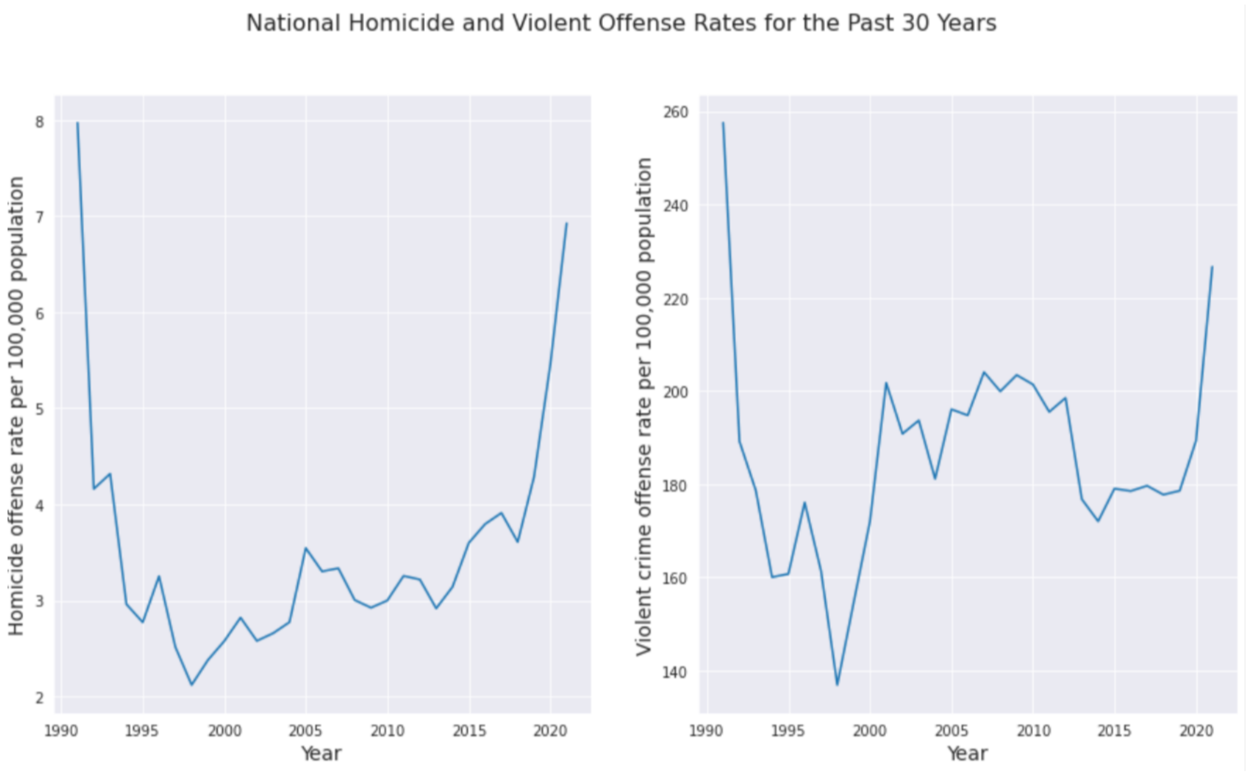
Figure 2

Commentary

- Based on our analysis of crime data for all US states, we have identified several key trends and patterns in the rates of violent crime.
- First, we found that Alaska has the highest average percentage of violent crimes, at almost 40%. This is significantly higher than the overall mean of violent crime percentages, which is around 23%. According to the Daily Mail, there are several potential reasons for this high rate of violent crime in Alaska, including the fact that men significantly outnumber women in the state, high levels of alcohol consumption, and the fact that law enforcement is spread thin across the vast territory. More details here.

- Second, the District of Columbia (also known as Washington D.C.) also has a high percentage of violent crimes, exceeding 35%. This has been a long-standing problem in the capital city, and efforts have been implemented to combat violent crime through gun control regulations. More details in this article by the Washington Post.
- Third, Vermont, Oregon, and Utah have relatively low percentages of violent crimes, with rates around 15%. These states are often ranked among the safest places to live in the U.S.
- Finally, we observed that some of the larger states, such as New York, Florida, and Nevada, have relatively high percentages of violent crimes, with rates around 30%. This suggests that there may be a relationship between state size and rates of violent crime.

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



National Homicide and Violent Offense Rates for the Past 30 Years

Figure 3

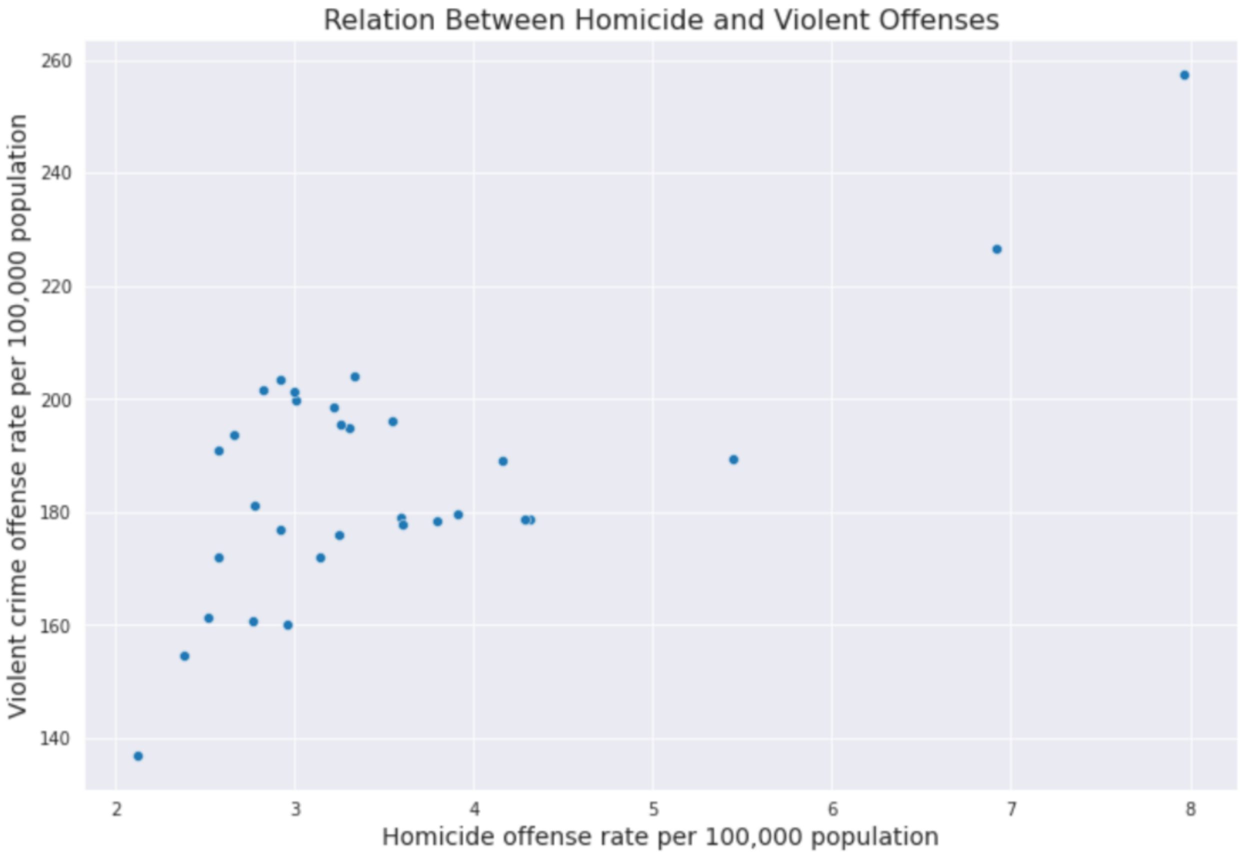


Figure 4

Commentary

- Based on the analysis of the homicide and violent crime offense rates over the past 30 years, it appears that both types of offenses have shown a similar trend over time, with a significant decrease in the early 1990s and then a more stable period for several years. However, the rates themselves are quite different, with the mean homicide offense rate being significantly lower than the mean violent offense rate.
- One potential reason for this difference in rates could be the various types of crime that are included in the definition of "violent crime." Homicide is just one type of crime that falls under this category, and it generally has a lower rate of occurrence than other types of violent crime such as assault, robbery, and sexual assault. This may explain why the overall violent crime rate is higher than the homicide rate.

- It is also worth noting that the violent crime rate appears to have spiked in the early 2000s, potentially due to a variety of social, economic, and other factors. In recent years, both homicide and violent crime rates have been on the rise again, which could be cause for concern for businesses and individuals in the community.
- One possible explanation for the increasing trend in both types of offenses could be a lack of resources or funding for law enforcement and crime prevention efforts. It could also be related to underlying social issues such as poverty, unemployment, and lack of access to education and opportunity.
- The scatter plot of the homicide and violent crime offense rates shows that there is a correlation between the two types of offenses, with a varying deviation. This means that as one type of offense rate increases, the other tends to increase as well, although the relationship is not necessarily a perfect one. It is important to note that the correlation between homicide and violent crime rates does not necessarily imply causation.

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

A. Occurrences of different crimes by victim sex

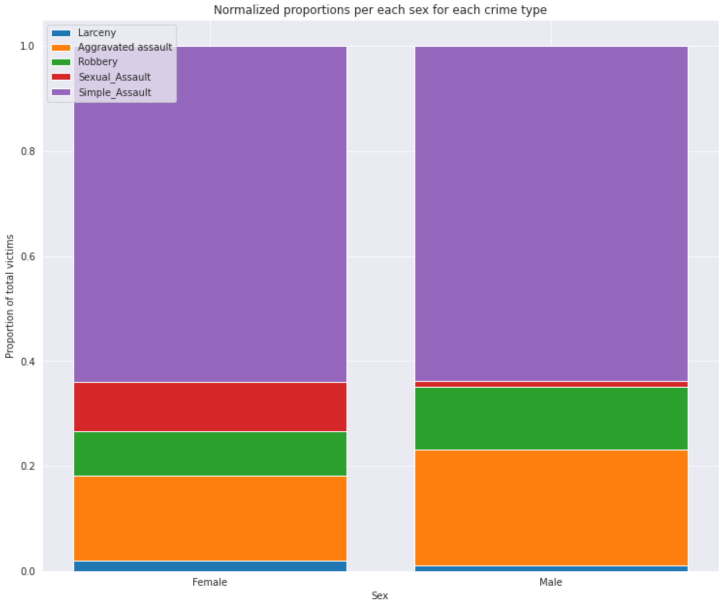


Figure 5

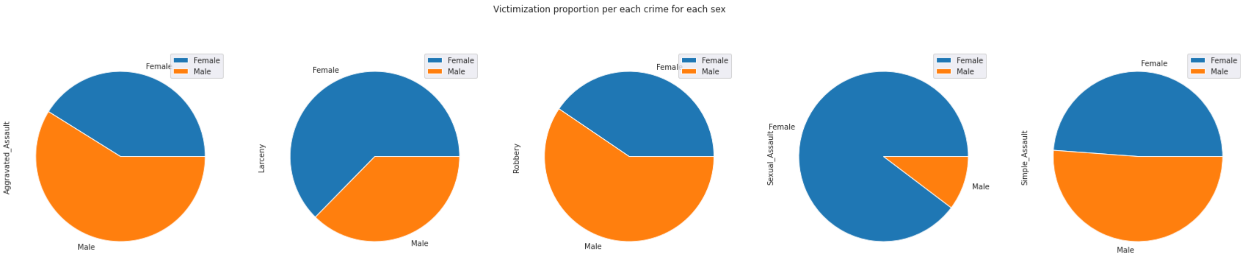


Figure 6

Commentary:

- More sexual assault victims are female by proportion and counts
- A higher proportion of male victims were subjected to aggravated assault
- Simple assault proportions were almost equal for both sexes
- Female larceny victims represented a slightly higher proportion than their male counterparts, with a majority count.
- Most robbery victims were male
- Simple assault constitutes the majority of victimization reports for both sexes.

B. Occurrences of different crimes by victim race

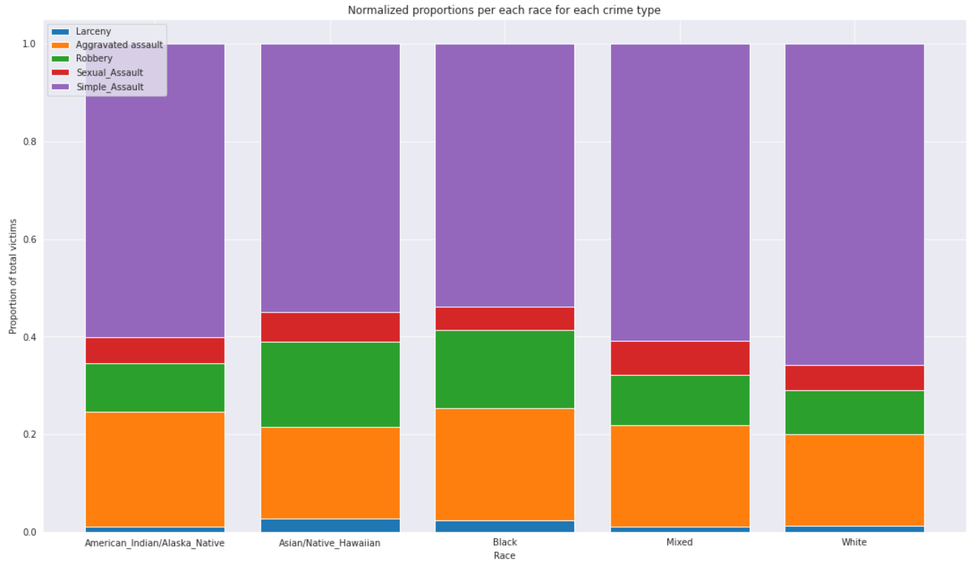


Figure 7

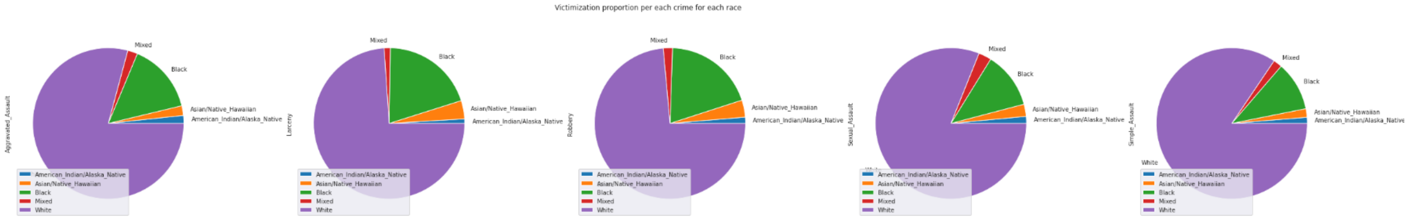


Figure 8

Commentary:

- White people represent the reported majority of victims for each crime, followed closely, the proportions of their victimization with respect to each crime was in line with the rest of the races. This leads us to the conclusion that the data collected was not equally representative for all races.
- A higher proportion of Native Americans were victims of larceny compared to other races
- Sexual Assault victims contained the largest white majority when compared to other crimes, excluding simple assault.
- proportion of sexual assault victims to the total victim count for all races was roughly equal
- Simple assault constitutes the majority of victimization reports for all races.

C. Occurrences of different crimes by victim age group

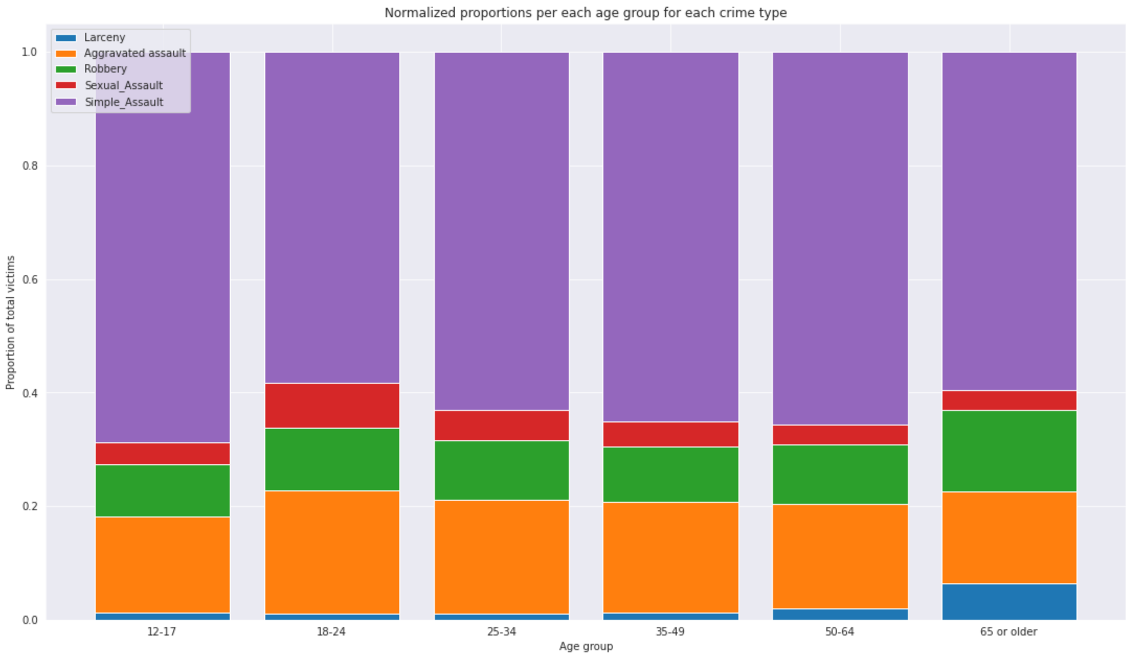


Figure 9

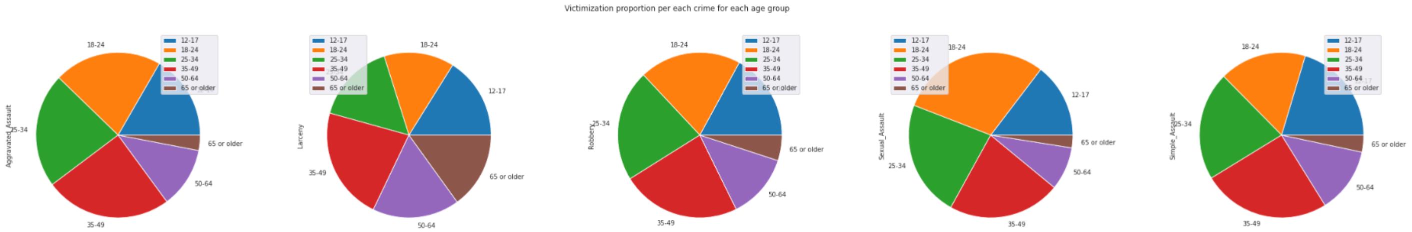


Figure 10

Commentary:

- For all crime types, almost half the victims were between 25-50.
- Ages 18-24 were the most victimized group through sexual assault.
- Sexual Assault victims contained the largest white majority when compared to other crimes, excluding simple assault.
- The proportion of 65+ cases of robbery and larceny in proportion to the total number of 65+ victims was higher than for other age groups.
- Simple assault constitutes the majority of victimization reports for all age groups.

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

A. Occurrences of different crimes by offender sex

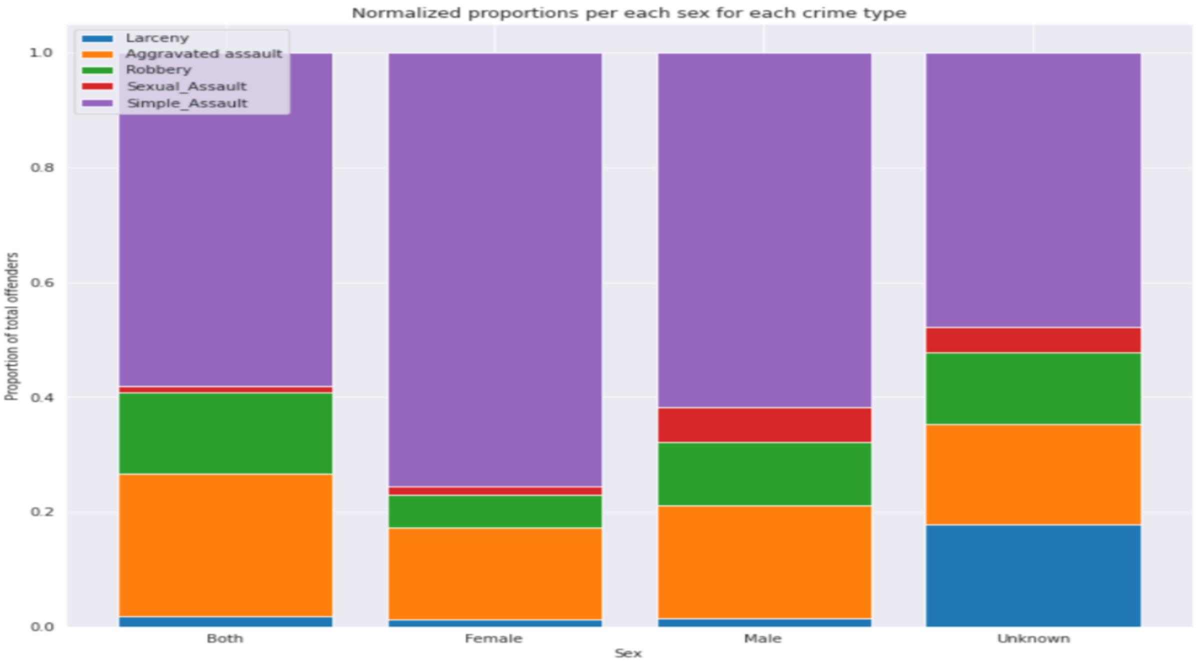


Figure 11

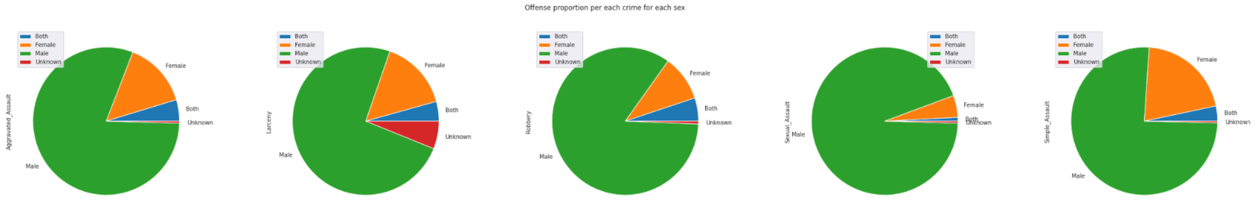


Figure 12

Commentary:

- Most offenders for all crimes are male.
- Male offenders have the highest proportion of sexual assault crimes to total crimes committed.
- Female offenders have the lowest proportion of robbery crimes to total crimes committed.
- Female larceny victims represented a slightly higher proportion than their male counterparts, with a majority count.

B. Occurrences of different crimes by offender race

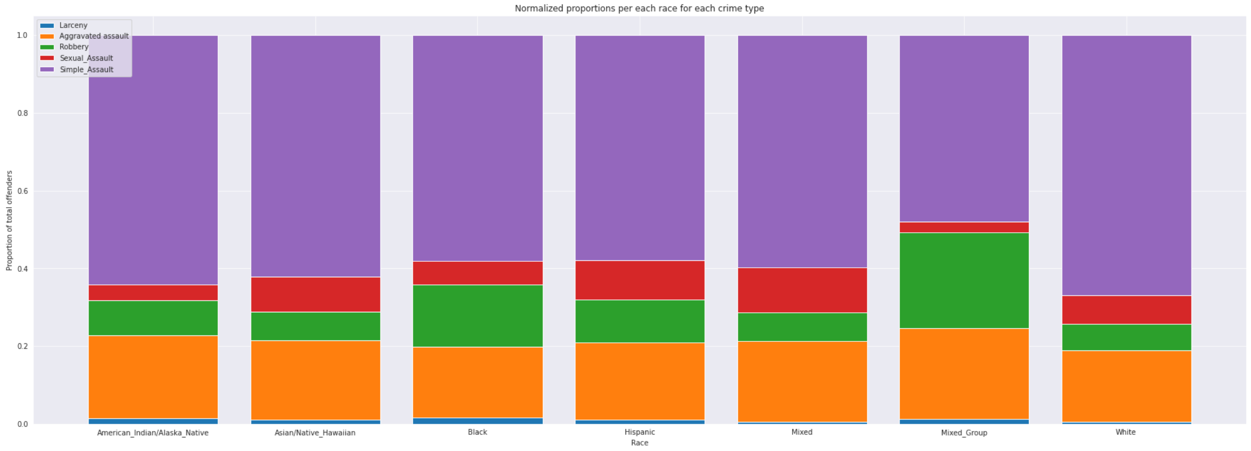


Figure 13

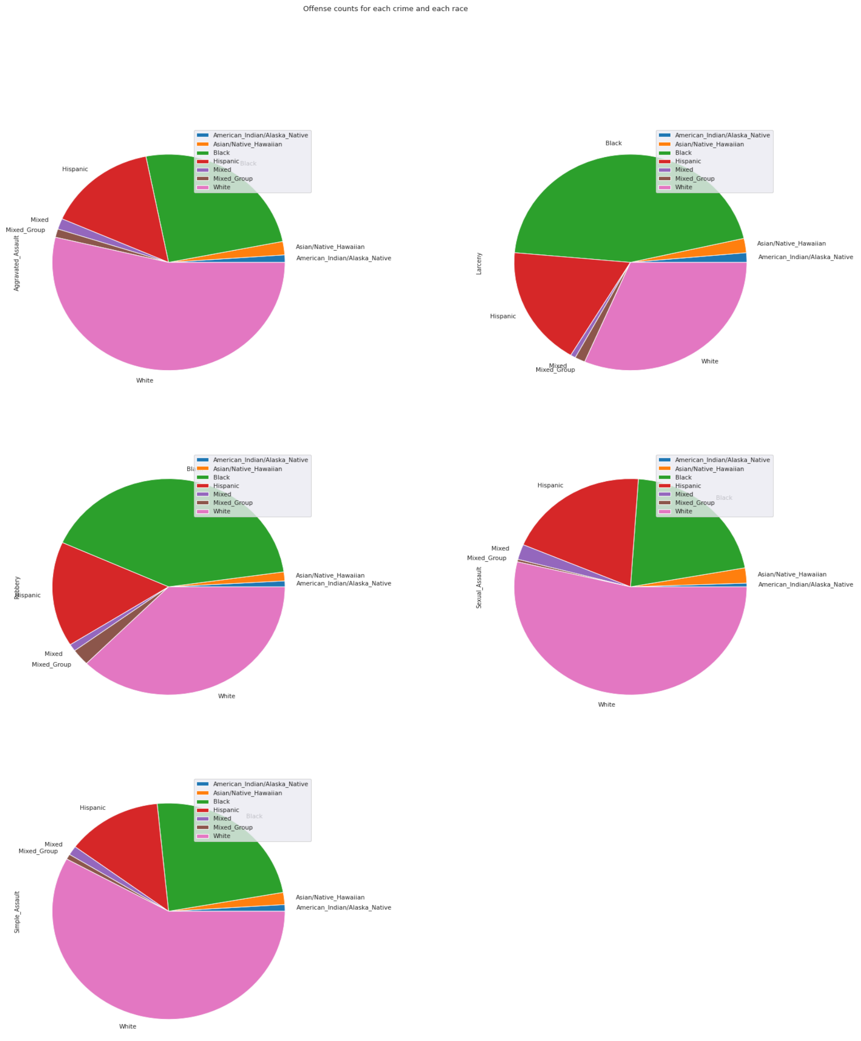


Figure 14

Commentary:

- Most offenders for all crimes are White. This reflects the same issue of inconsistent representation of all races. The proportions of committed crimes within the white race does not diverge much from the values of other races, nonetheless
- Male offenders have the highest proportion of sexual assault crimes to total crimes committed.
- Mixed race offender groups have the highest proportion of sexual assault crimes to total crimes committed.
- Female larceny victims represented a slightly higher proportion than their male counterparts, with a majority count.
- "Invalid Until 2021Q1", "Unknown", and "Unknown Group" do not give us any information regarding demographics and they contain a large number of records. Therefore, they were removed.

C. Occurrences of different crimes by offender age group

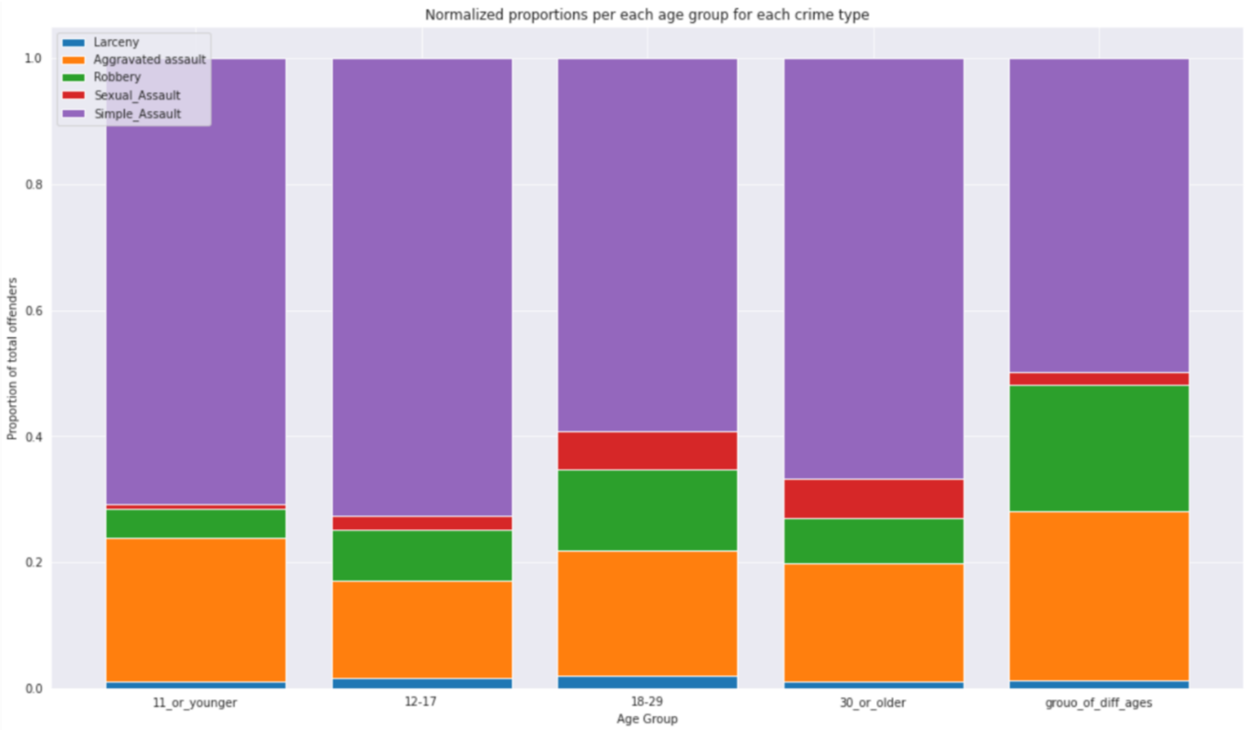


Figure 15

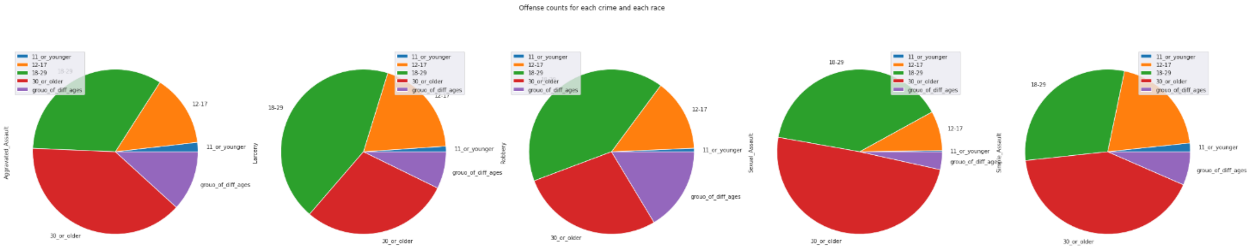


Figure 16

Comments:

- Majority of crimes were committed by the 18-29 age group.
- Almost half of sexual assault crimes were committed by men aged 30 or more.
- Groups of different ages had a higher proportion of robbery and larceny offenses than other groups.

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

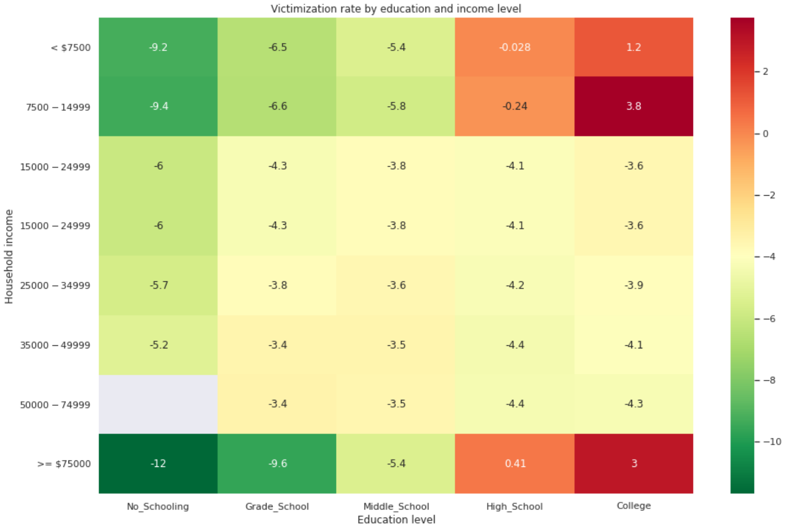


Figure 17

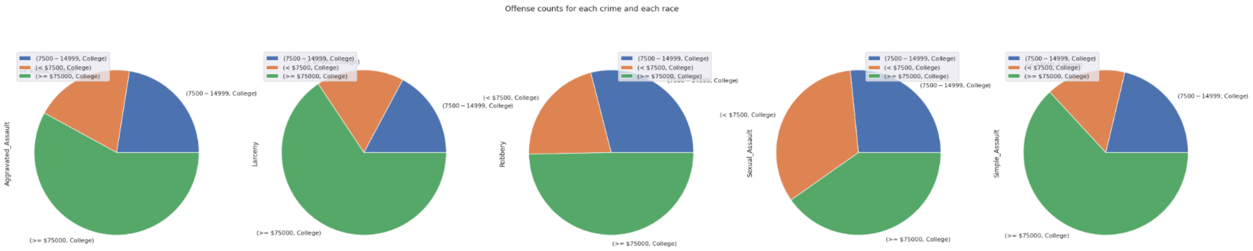


Figure 18

Comments:

- For the highest and lowest income levels, as educational levels increase, the risk of victimization increases.
- No records of uneducated individuals exist for the 50k-75k USD.
- Middle-Class income tiers show relatively consistent victimization levels
- The college educated, 75k+ USD group was the most targeted for all crimes.
- The highest proportion for the college educated sub 7.5k USD was in sexual assault victimizations.
- The highest proportion for the college educated 7.5k - 14.9k USD was in robbery victimizations.
- Unknown income column was removed as it provides no information regarding this comparison.

Answering Questions

Which type of non-fatal crime is the most under-reported?

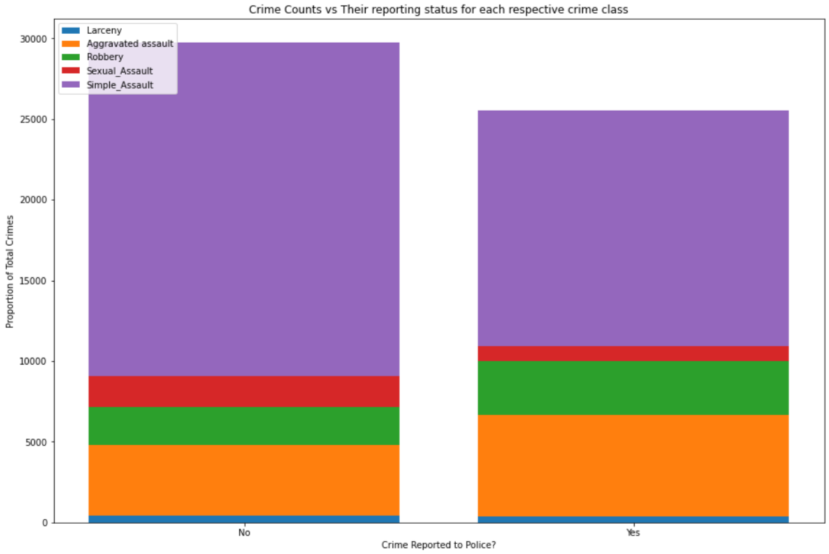


Figure 19

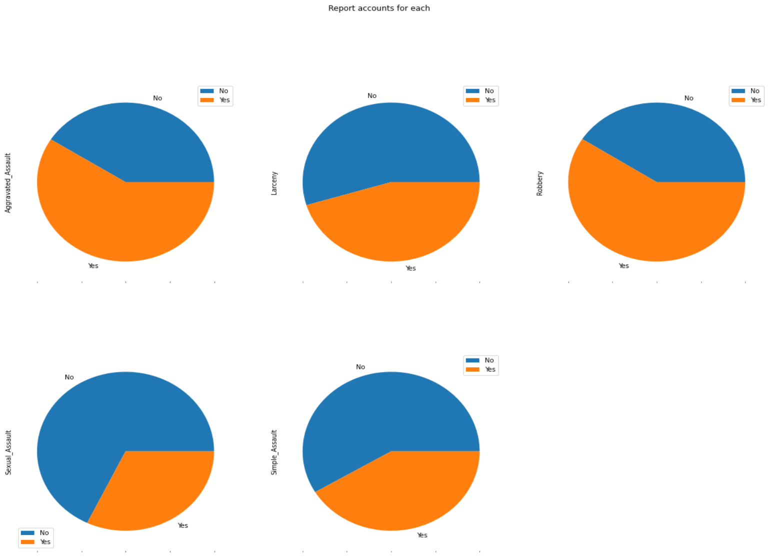


Figure 20

Answer

The Most underreported crime in sheer numbers is Simple Assault. However, if we normalize for the total number of offenses for each crime type, we can see that Sexual Assault has the highest probability of being unreported.

Is there an association between the offender-victim relationship and the likelihood of a crime being reported?

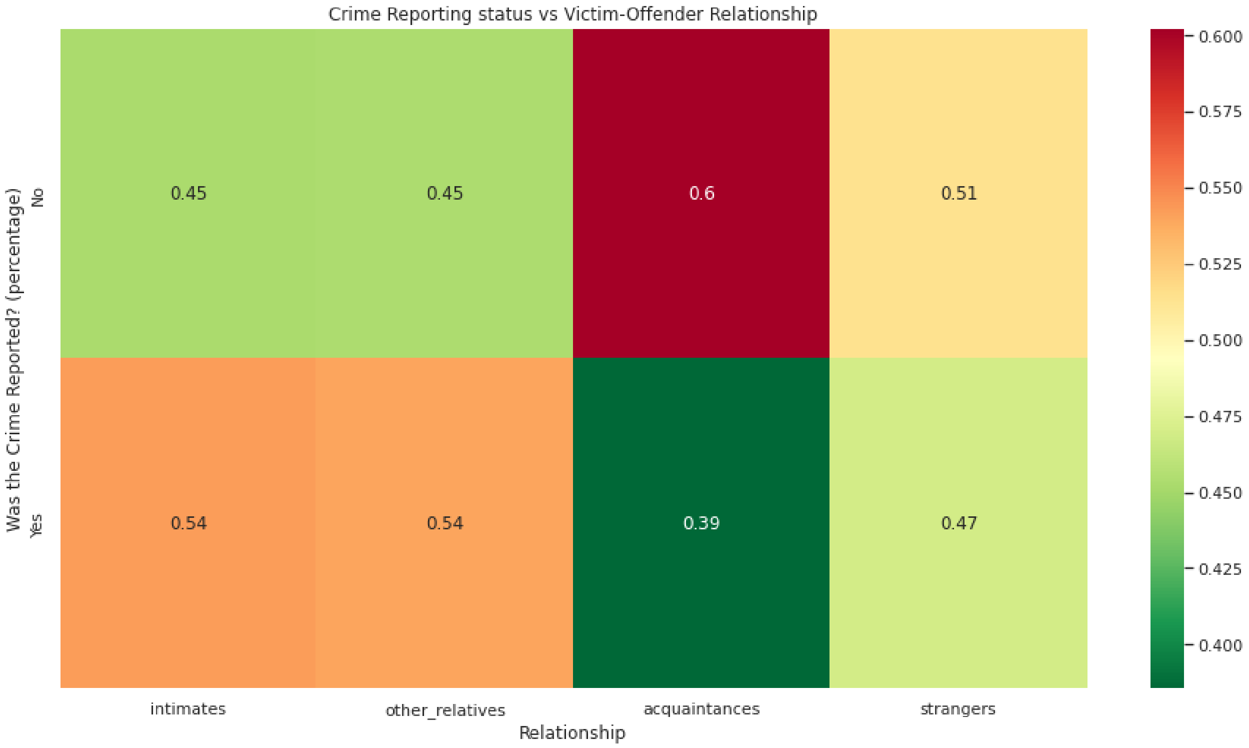


Figure 21

Answer

- *Strangers' and Acquaintance unreported crimes percentage is higher than that of relatives and intimates. This goes against our initial expectation of stronger relationships correlating with higher unreported crimes rate. The percentage of unreported crimes for both intimates and relatives is actually lower than the percentage of reported crimes.*
- *Crimes with acquainted victims and offenders seem to be the most unreported of all categories, followed by crimes of estranged victims and offenders.*

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

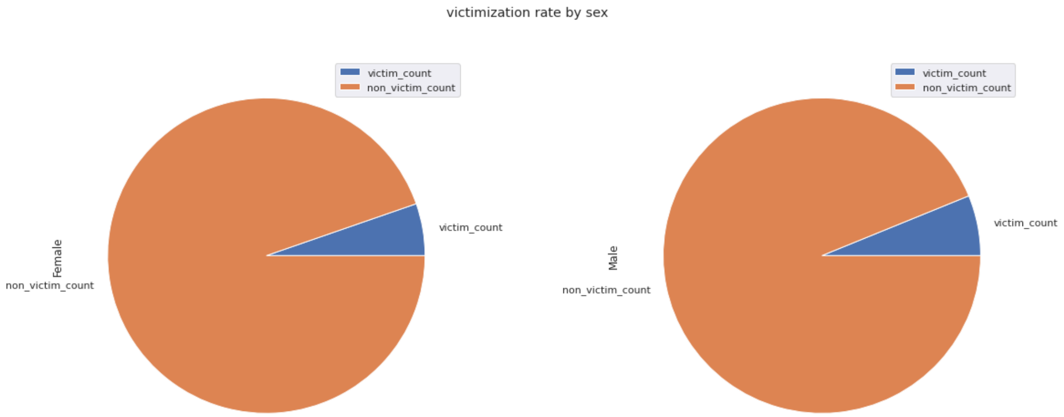


Figure 21

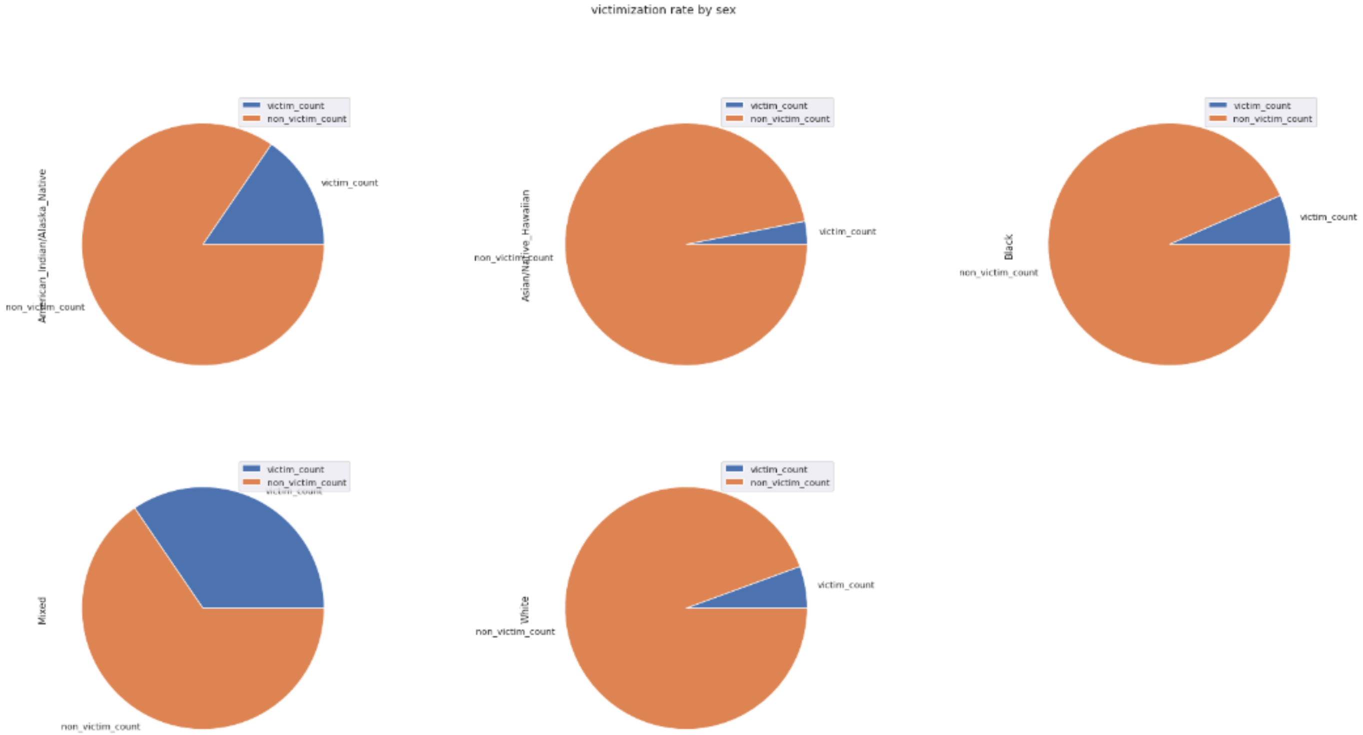


Figure 22

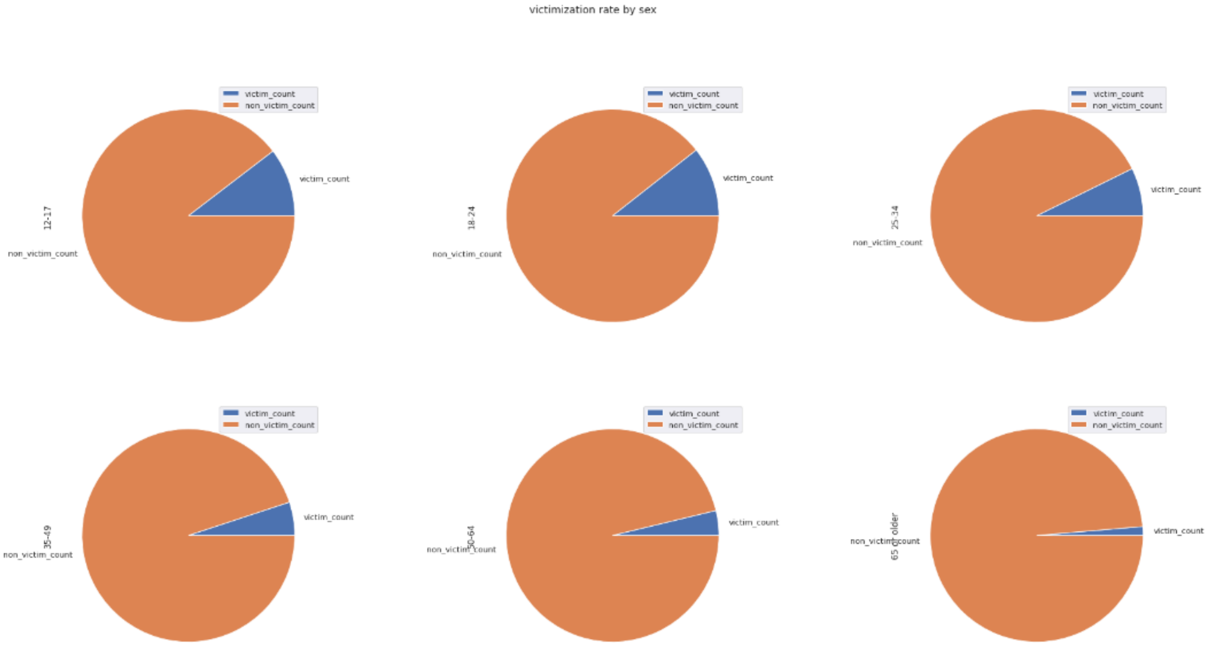


Figure 23

Answer

Sex has very minimal effect on the victimization rate. this leaves us with the following conclusions:

- *Mixed race individuals between 12 and 24 are the most likely to be victimized*
- *Native Americans between 12 and 24 are the most non-mixed-race individuals likely to be victimized*
- *Asian individuals over 65 years old are the least likely to be victimized*

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?

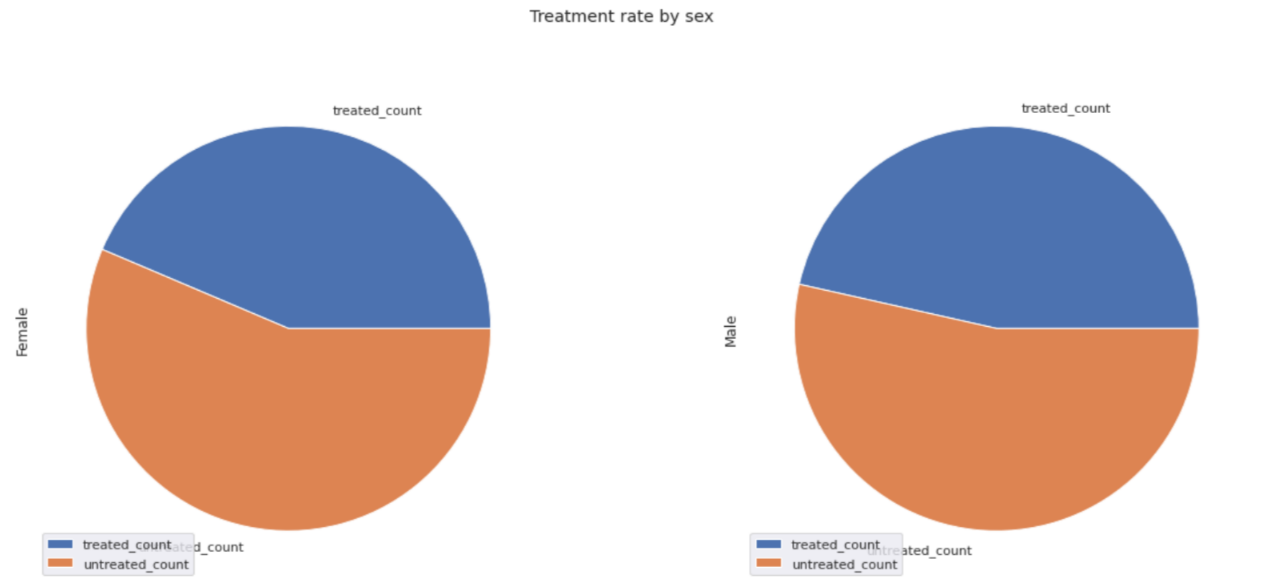


Figure 24

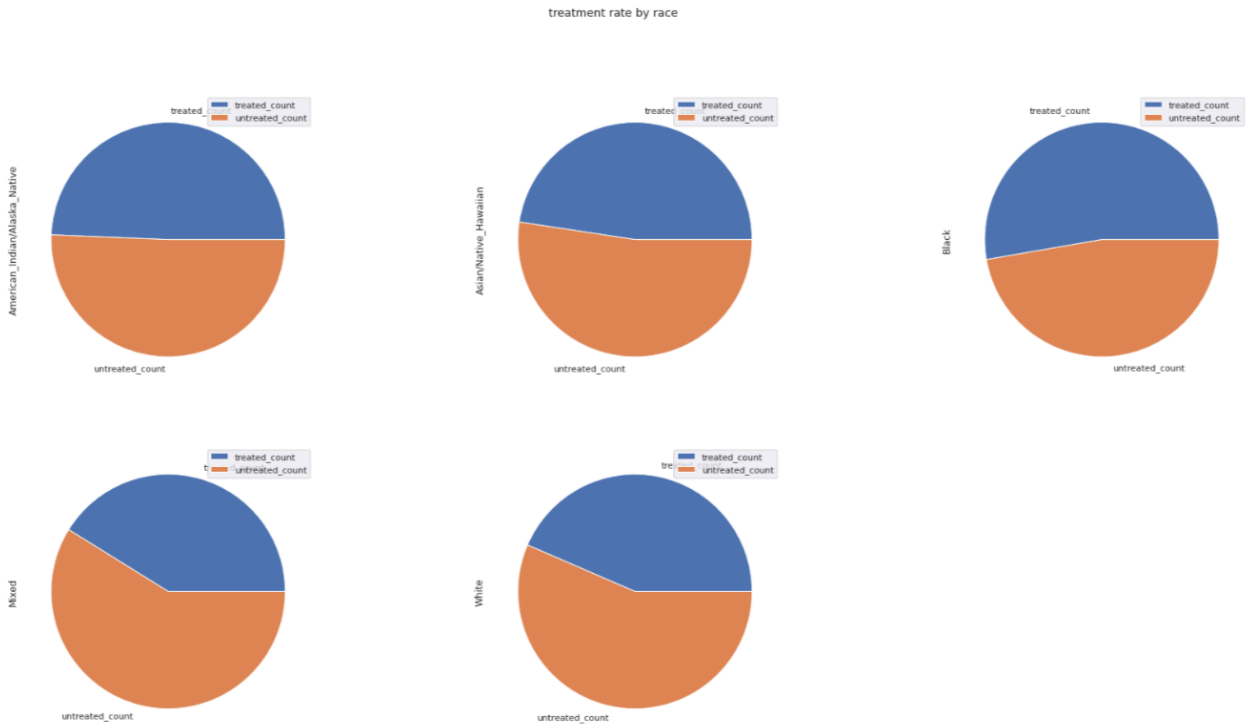


Figure 25

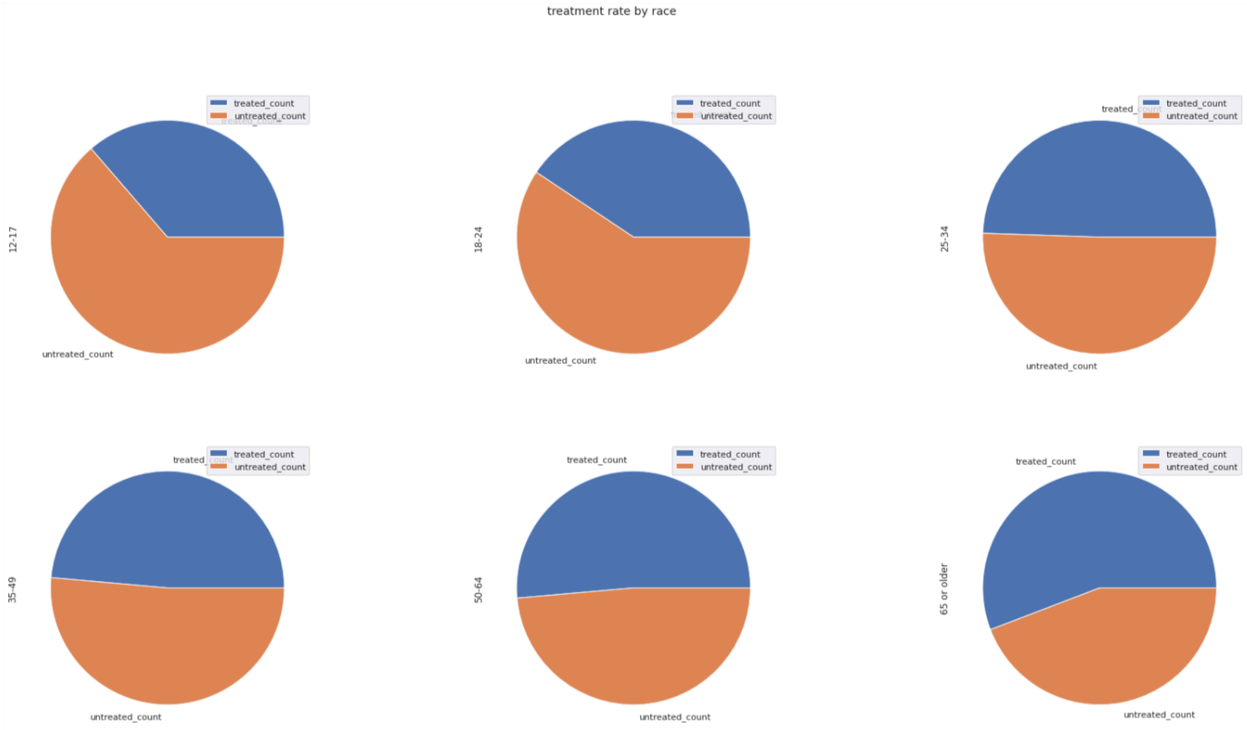


Figure 26

Answer

<u>Most treated demographic:</u>

- *Males have a higher probability of receiving treatment*
- *Black individuals are the most likely of all races to receive treatment*
- *Victims aged 65+ are the most probable age group to receive treatment*
- *This leads us to conclude that 65+ year old black males are the most probable to receive treatment*

<u>Least treated demographic:</u>

- *Females have a lower probability of receiving treatment*
- *Mixed race individuals are the least likely of all races to receive treatment*
- *White individuals are the least likely of all non-mixed races to receive treatment*
- *children aged 12-17 are the least probable age group to receive treatment*
- *This leads us to conclude that 12-17 year old mixed race females are the least probable to receive treatment*
- *This also leads us to conclude that 12-17 year old white females are the least probable non-mixed race group to receive treatment*

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?

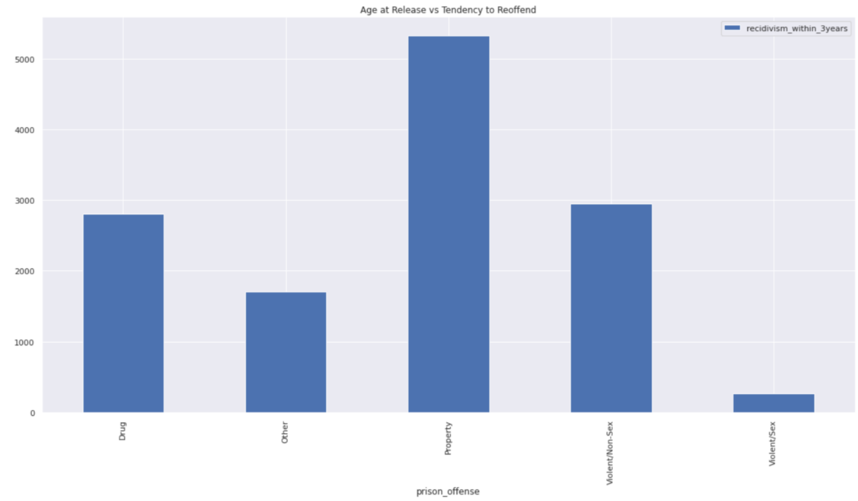
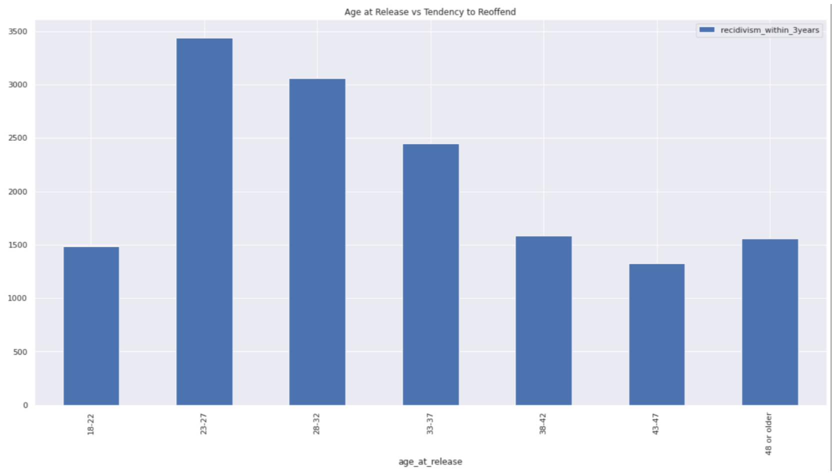


Figure 27

Answer

- *Property offenders are the most likely to reoffend within the three year window, followed by Drug and Violent/Non-Sex offenders.*
- *Sexual Assault offenders are the least likely to commit a similar offense within the 3-year time window*

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



Answer

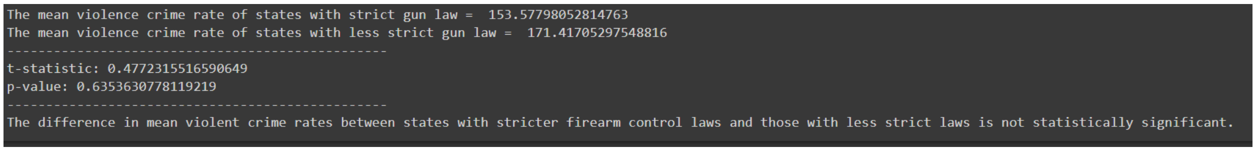
- *Property offenders are the most likely to reoffend within the three year window, followed by Drug and Violent/Non-Sex offenders.*
- *Sexual Assault offenders are the least likely to commit a similar offense within the 3-year time window*

Hypothesis Testing

In this section, we will test the Claim that "U.S. states that implement stricter firearm control laws, have lower violent crime rates on average" using two-tailed t-test and linear regression analysis test.

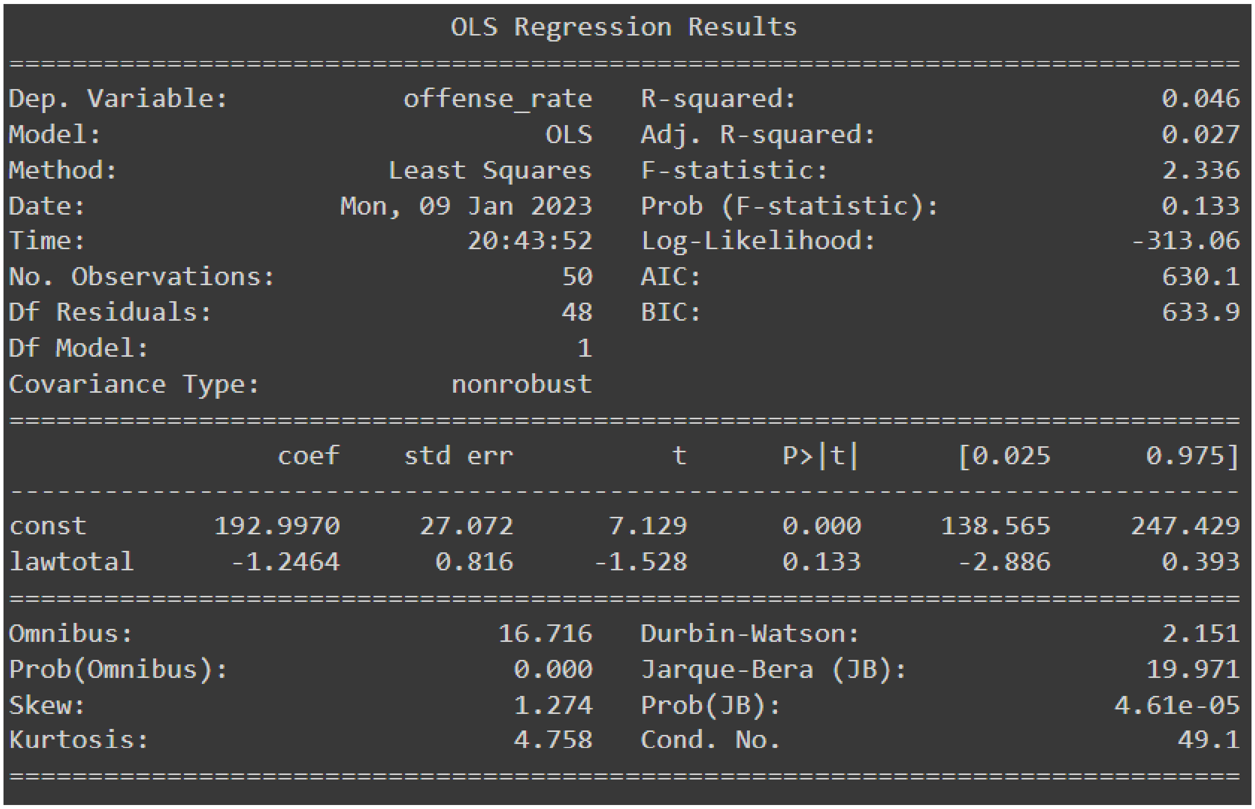
Hypothesis test 1:

Two tailed t-test results:



```
The mean violence crime rate of states with strict gun law =  153.57798052814763
The mean violence crime rate of states with less strict gun law =  171.41705297548816
------------------------------------------------
t-statistic: 0.4772315516590649
p-value: 0.6353630778119219
------------------------------------------------
The difference in mean violent crime rates between states with stricter firearm control laws and those with less strict laws is not statistically significant.
```

Linear regression test



```
                            OLS Regression Results
==============================================================================
Dep. Variable:           offense_rate   R-squared:                       0.046
Model:                            OLS   Adj. R-squared:                  0.027
Method:                 Least Squares   F-statistic:                     2.336
Date:                Mon, 09 Jan 2023   Prob (F-statistic):              0.133
Time:                        20:43:52   Log-Likelihood:                -313.06
No. Observations:                  50   AIC:                             630.1
Df Residuals:                      48   BIC:                             633.9
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        192.9970     27.072      7.129      0.000     138.565     247.429
lawtotal      -1.2464      0.816     -1.528      0.133      -2.886       0.393
==============================================================================
Omnibus:                       16.716   Durbin-Watson:                   2.151
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               19.971
Skew:                           1.274   Prob(JB):                     4.61e-05
Kurtosis:                       4.758   Cond. No.                         49.1
==============================================================================
```

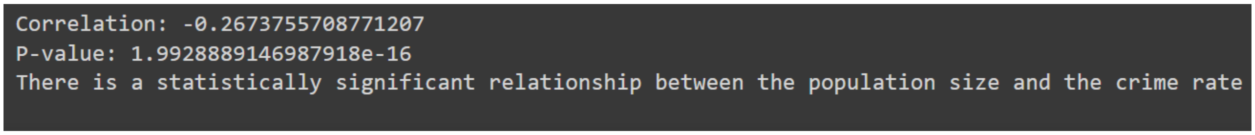
Discussion

- There is not enough evidence to reject the null hypothesis for both tests which found no significant relationship between the number of firearm control laws in a state and the violent crime rate in that state.
- There could be a number of reasons for this result. It is possible that the relationship between firearm control laws and violent crime rates is more complex than the claim suggests, and other factors (such as economic conditions, demographics, and social policies) may have a greater influence on violent crime rates.
- Our results are confirmed by the Brady gun fact and violent crime campaign. It was stated in their study that California which ranks #1 in gun control and has strongest gun control laws and Arizona, the state which ranks #50 with almost no gun control at all have nearly identical violent crime rates. California has 423 violent crimes per every 100,000 residents and Arizona has 429. Not a dime's worth of difference.
- The short answer then is "no, gun control strictness does not lead to lower crime."
- In the political context, this result could be seen as undermining the argument for stricter firearm control laws as a means of reducing violent crime. It is possible that politicians and advocacy groups on both sides of the issue could use this result to support their positions.
- Additionally, the result suggests that the relationship between firearm control laws and violent crime rates may not be straightforward, and that addressing violent crime may require a more nuanced and multifaceted approach. Overall, the p-value being too large does not necessarily say anything about the data, political situation, or society in general, but rather about the strength of the relationship between the variables being studied (in this case, firearm control laws and violent crime rates).

In this section, we will test the Claim that "U.S states with larger populations tend to have more crime rate" using Pearson's correlation test.

Hypothesis test 2

Pearson Test Result



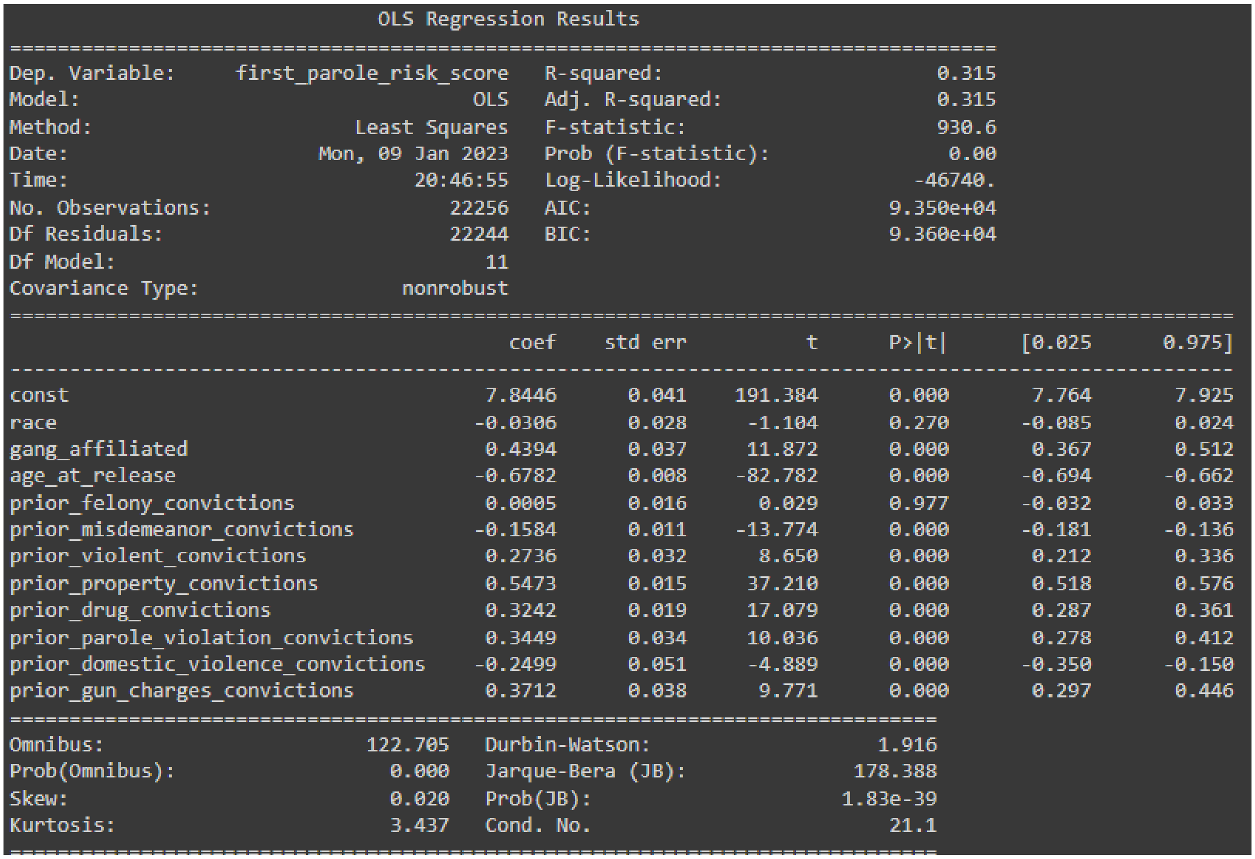
```
Correlation: -0.2673755708771207
P-value: 1.9928889146987918e-16
There is a statistically significant relationship between the population size and the crime rate
```

Discussion:

Based on the results of the Pearson's correlation coefficient test, with a p-value less than the significance level of 0.05, we can conclude that there is a significant relationship between crime rate and population size. However, a negative correlation coefficient indicates that there is a negative relationship between the crime rate and population. This means that as the population decreases, crime goes up and states with larger populations tend to have lower crime rates.

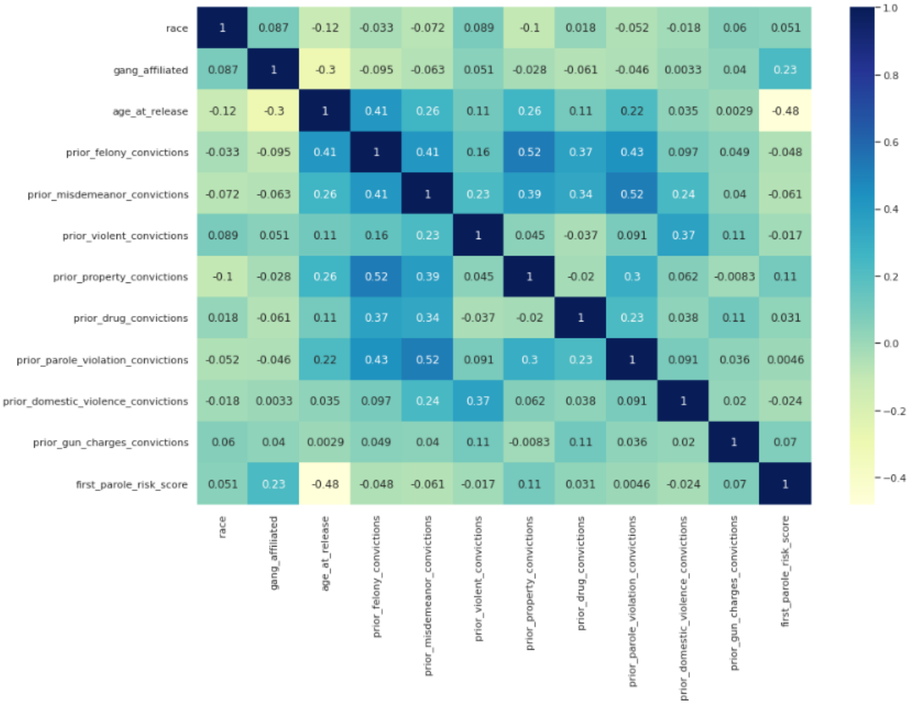
Regression Analysis

Model Results:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:     first_parole_risk_score   R-squared:                       0.315
Model:                                 OLS   Adj. R-squared:                  0.315
Method:                      Least Squares   F-statistic:                     930.6
Date:                     Mon, 09 Jan 2023   Prob (F-statistic):               0.00
Time:                             20:46:55   Log-Likelihood:                -46740.
No. Observations:                    22256   AIC:                         9.350e+04
Df Residuals:                        22244   BIC:                         9.360e+04
Df Model:                               11
Covariance Type:                 nonrobust
=======================================================================================================
                                          coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------
const                                   7.8446      0.041    191.384      0.000       7.764       7.925
race                                   -0.0306      0.028     -1.104      0.270      -0.085       0.024
gang_affiliated                         0.4394      0.037     11.872      0.000       0.367       0.512
age_at_release                         -0.6782      0.008    -82.782      0.000      -0.694      -0.662
prior_felony_convictions                0.0005      0.016      0.029      0.977      -0.032       0.033
prior_misdemeanor_convictions          -0.1584      0.011    -13.774      0.000      -0.181      -0.136
prior_violent_convictions               0.2736      0.032      8.650      0.000       0.212       0.336
prior_property_convictions              0.5473      0.015     37.210      0.000       0.518       0.576
prior_drug_convictions                  0.3242      0.019     17.079      0.000       0.287       0.361
prior_parole_violation_convictions      0.3449      0.034     10.036      0.000       0.278       0.412
prior_domestic_violence_convictions    -0.2499      0.051     -4.889      0.000      -0.350      -0.150
prior_gun_charges_convictions           0.3712      0.038      9.771      0.000       0.297       0.446
==============================================================================
Omnibus:                      122.705   Durbin-Watson:                   1.916
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              178.388
Skew:                           0.020   Prob(JB):                     1.83e-39
Kurtosis:                       3.437   Cond. No.                         21.1
==============================================================================
```

Correlation Matrix



	race	gang_affiliated	age_at_release	prior_felony_convictions	prior_misdemeanor_convictions	prior_violent_convictions	prior_property_convictions	prior_drug_convictions	prior_parole_violation_convictions	prior_domestic_violence_convictions	prior_gun_charges_convictions	first_parole_risk_score
race	1	0.087	-0.12	-0.033	-0.072	0.089	-0.1	0.018	-0.052	-0.018	0.06	0.051
gang_affiliated	0.087	1	-0.3	-0.095	-0.063	0.051	-0.028	-0.061	-0.046	0.0033	0.04	0.23
age_at_release	-0.12	-0.3	1	0.41	0.26	0.11	0.26	0.11	0.22	0.035	0.0029	-0.48
prior_felony_convictions	-0.033	-0.095	0.41	1	0.41	0.16	0.52	0.37	0.43	0.097	0.049	-0.048
prior_misdemeanor_convictions	-0.072	-0.063	0.26	0.41	1	0.23	0.39	0.34	0.52	0.24	0.04	-0.061
prior_violent_convictions	0.089	0.051	0.11	0.16	0.23	1	0.045	-0.037	0.091	0.37	0.11	-0.017
prior_property_convictions	-0.1	-0.028	0.26	0.52	0.39	0.045	1	-0.02	0.3	0.062	-0.0083	0.11
prior_drug_convictions	0.018	-0.061	0.11	0.37	0.34	-0.037	-0.02	1	0.23	0.038	0.11	0.031
prior_parole_violation_convictions	-0.052	-0.046	0.22	0.43	0.52	0.091	0.3	0.23	1	0.091	0.036	0.0046
prior_domestic_violence_convictions	-0.018	0.0033	0.035	0.097	0.24	0.37	0.062	0.038	0.091	1	0.02	-0.024
prior_gun_charges_convictions	0.06	0.04	0.0029	0.049	0.04	0.11	-0.0083	0.11	0.036	0.02	1	0.07
first_parole_risk_score	0.051	0.23	-0.48	-0.048	-0.061	-0.017	0.11	0.031	0.0046	-0.024	0.07	1

Discussion:

The model's coefficients and p-values are displayed in the model's statistics summary above. Based on the magnitude and p-value of the model's coefficients, the good predictors

By looking at the magnitude p-values of the parameters' coefficients, we can deduce if the predictor associated with each parameter accounts for the variability in the target variable (supervision risk score) in a statistically significant way or not.

From the statistics summary we can see that some predictors have relatively high p-values (with significance level 0.05) which would allow us to deduce that they are bad predictors

- race, p-value = 0.270
- prior_felony_convictions, p-value = 0.977

The good predictors typically have a p-value of zero which means there's statistically significant relationship between the predictor and the target and these are (with p-value = 0):

- gang_affiliated
- age_at_release
- prior_misdemeanor_convictions
- prior_violent_convictions
- prior_property_convictions
- prior_drug_convictions
- prior_parole_violation_convictions
- prior_domestic_violence_convictions
- prior_gun_charges_convictions
- By looking at the correlation heatmap of all model's features, we can observe that none of the model's predictors are correlated with each other.

To assess the model's quality we could look at the R^2 value in the model's statistics summary. An R^2 of 0.315 is generally considered to be a low to moderate amount of explained variance. This may indicate that the model is not a very good fit to the data, and that other factors may be influencing the dependent variable.

Training a Random Forest Classifier (Bonus Task)

The random forest classifier was trained on the state of Georgia recidivism records, which included information on various factors such as location, time of day, and type of crime. The classifier was able to reach an accuracy of 73% on the test set, indicating that it was able to correctly predict the type of crime for a large portion of the cases. This high level of accuracy was likely achieved by the random forest's ability to make use of the rich information provided in the dataset and learn from the patterns present in the data. Additionally, the use of multiple decision trees in the random forest helped to reduce overfitting and improve the generalizability of the model.

Conclusion

In conclusion, the data analysis performed on US crime datasets has revealed valuable insights into the patterns and trends present in crimes committed within the country despite the low volume of the data and various structural and missing value issues. By utilizing various data visualization techniques and machine learning algorithms, it was possible to gain a deeper understanding of the factors that may contribute to the occurrence of different types of crimes. Overall, the results of this study demonstrate the importance of data analysis in helping to shed light on complex social issues and inform the development of effective strategies for crime prevention and reduction.