Predicting the causes of crimes in the US

Suvam Patel

Problem Identification

- Hypothesis: There can be many reasons as to why crime occurs in US communities. It can be due to eviction, unemployment, ethnic background, age group, and/or education.
- Context: Crime occurrences can be validated because of a number of reasons. Many people are subject to hardships in the form of eviction and/or unemployment which makes them resort to crime. This study will analyze which variable is strongly correlated with crime occurrences.
- Criteria for Success: Identify which variable is strongly correlated with crime and if education brings down crime rates.
- Scope of solution space: Identify the variables that are strongly correlated with crime occurrences.
- Constraints within solution space: The data file which will be used for the analysis of this study is missing values which can lead to the skepticism of the accuracy of the study.
- Data Sources: UCI Machine Learning Repository and Kaggle.
- By using the data I was able to test Regression models on the dataset to help verify which variable is the best predictor of crime rates in the US.

Audience

- Such a study can help authorities working in the crime field in assessing crime-related factors and help alleviating socioeconomic circumstances in crime heavy communities.
- The analysis can also help authorities with safeguarding stratagems to help prevent crime.
- Crime rates can also be reduced by providing better educational tools to children living in low-income communities.
- Moreover, the purpose of this study is to provide an analysis to governmental institutions about the occurrence of crime rates and how such circumstances can be reduced or prevented.

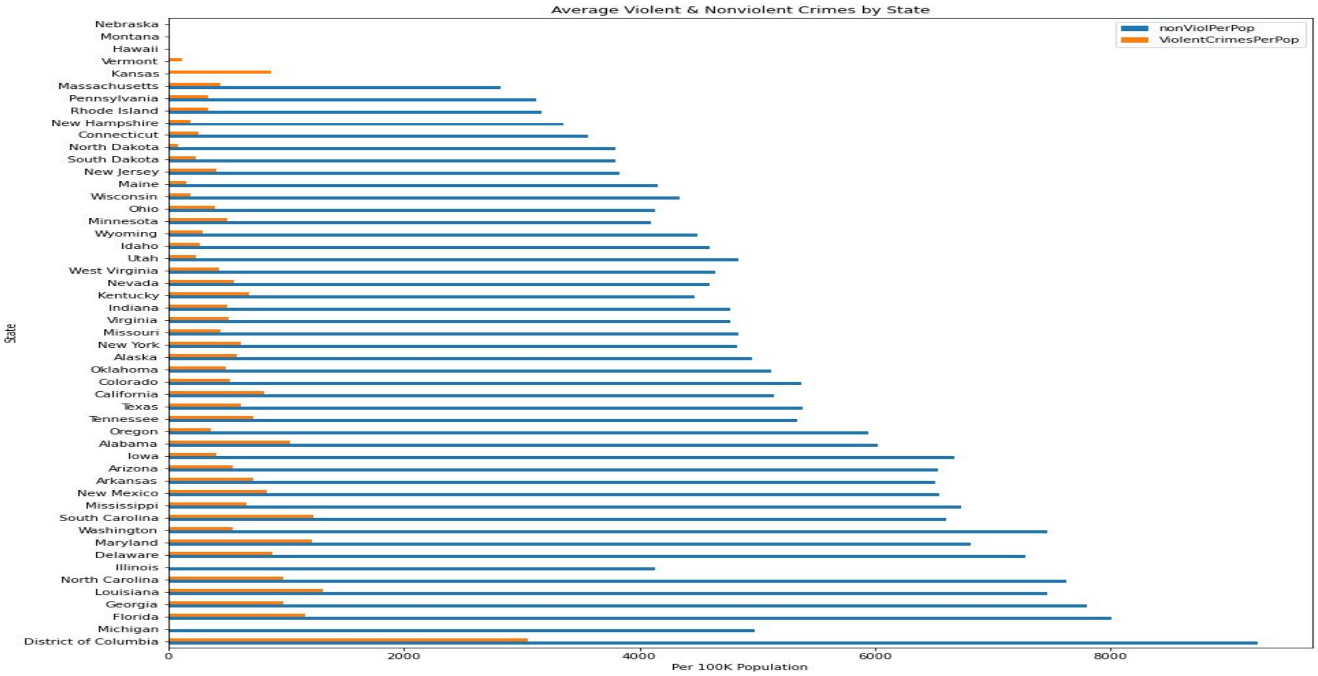
Step 1: Data Wrangling

- The dataset is originally from the UCI Machine Learning Repository and was prepared using real data from socio-economic data from 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime data from the 1995 FBI UCR [13]. UCI Machine Learning Repository contains a range of free datasets which can be used by anyone trying to hone their Machine Learning skills.
- To get a feel for the data, I started to explore the dataset to see the shape of the dataset, amount of missing values, and the column names. This dataset contains a total number of 147 attributes and 2216 instances. After thoroughly exploring the dataset by seeing the amount of missing values per column, 15 columns were chosen to be appropriate for the study. The variables were statistically observed by checking their mean, min, max, percentile, and std. The variables were missing values from as little as 3 rows to 227 rows. Instead of dropping the entire row, only the missing values were dropped.

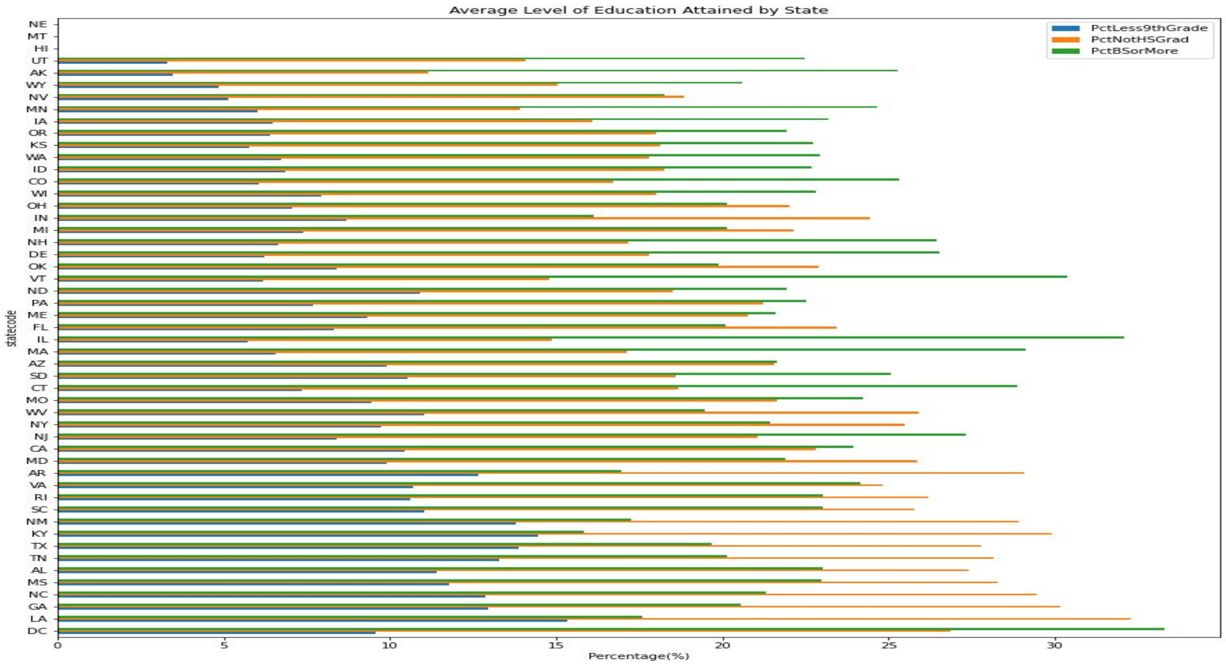
Data Wrangling (Continued)

- The dataset’s categorical attributes were checked for unique values and any duplicate values under 'communityname'. Communities with the same names were checked which states they are from and were validated that the duplicates belonged to different states. Then the attributes of interest were averaged and grouped by state to see how each attribute differed from state to state. The distributions were also visualized through horizontal bar graphs. Boxplots were also used to visualize the distribution of Violent and NonViolent Crimes for each state.
- Scatter plots were created for the target features to analyze the correlation between the socioeconomic variables and crime variables, in order to gain a premature understanding of which attribute is highly correlated with the occurrence of crime. I am really interested in seeing how all the variables are correlated with crime, especially education.

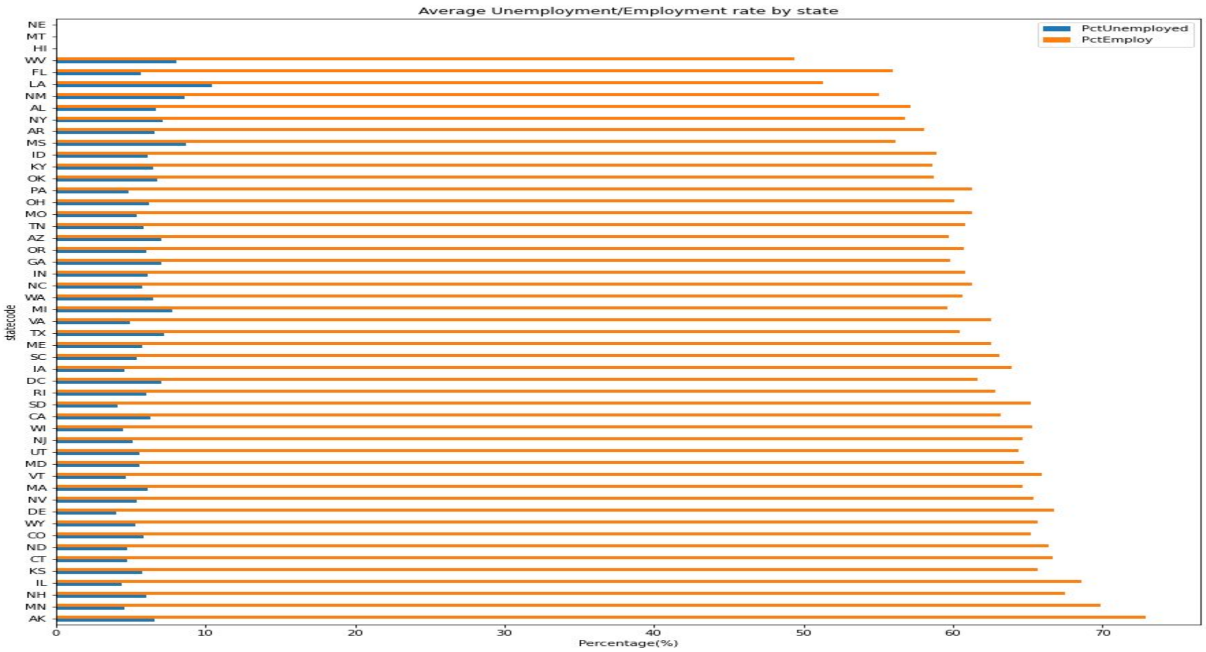
Average Violent & Nonviolent Crimes by State (Bar Graph)



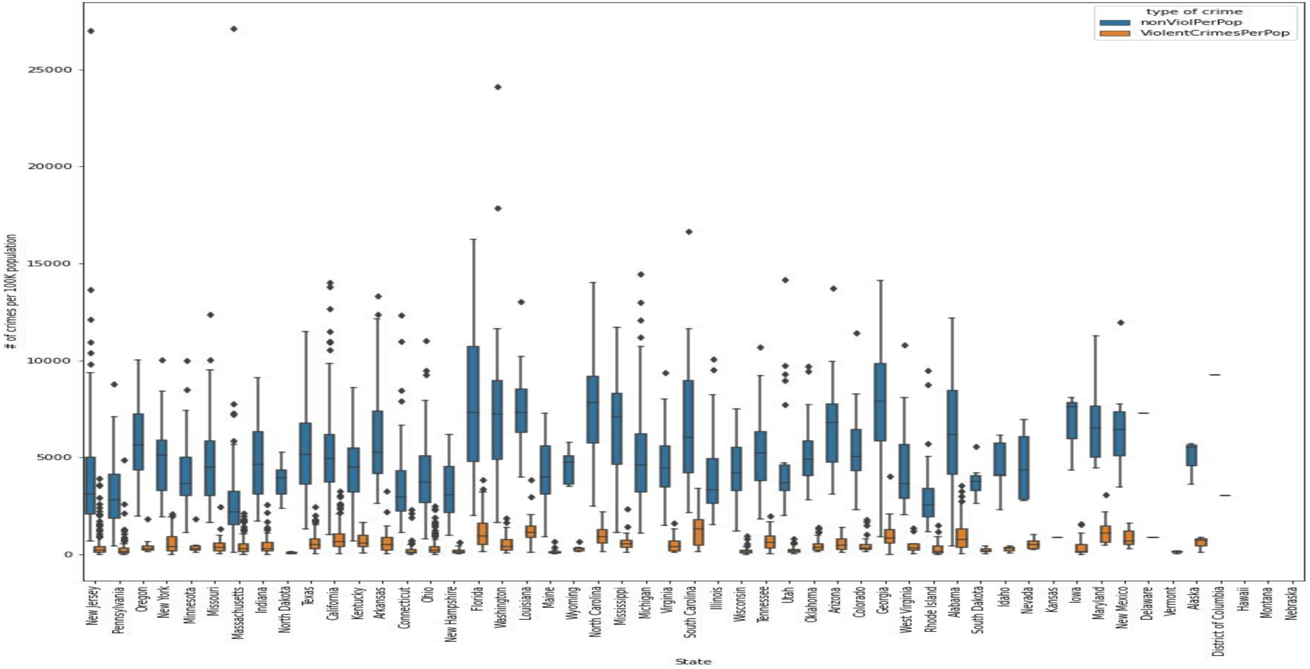
Average Level of Education Attained by State (Bar Graph)



Average Unemployment/Employment Rate by State (Bar Graph)



Average Violent & Nonviolent Crimes by State (Boxplot)



Step 2: EDA

- Step 1: Created bar graphs plotting the distribution for violent crimes, non-violent crimes, and income levels of each race.
- Step 2: Regression Plots were used to visualize the correlation between target variables and crime variables. Calculated the Correlation Coefficients for each target.
  - Non-Violent Crimes:
    - Education:

```
Correlation coefficient for PctLess9thGrade:  0.28784927687473005
Correlation coefficient for PctNotHSGrad:  0.36650015753649645
Correlation coefficient for PctBSorMore:  -0.27101682578840325
```

    - Employmen/Unemployment:

```
Correlation coefficient for PctUnemployed:  0.3920850019155378
Correlation coefficient for PctEmploy:  -0.30471049193594246
```

    - Vacancy:

```
Correlation coefficient for PctHousOccup:  -0.3039032395515144
Correlation coefficient for PctHousOwnOcc:  -0.4622358628933084
Correlation coefficient for PctVacantBoarded:  0.32367867144782136
Correlation coefficient for PctVacMore6Mos:  -0.0430259621892053
```

    - Race:

EDA (Continued)

```
Correlation coefficient for pctraceblack:  0.4743247060336021
Correlation coefficient for pctRaceWhite:  -0.4765791610681369
Correlation coefficient for pctRaceAsian:  -0.03474179713723831
Correlation coefficient for pctRaceHisp:  0.17462237036514378
```

- Violent Crimes:
  - Education:

```
Correlation coefficient for PctLess9thGrade:  0.37080716309505024
Correlation coefficient for PctNotHSGrad:  0.46651461611308775
Correlation coefficient for PctBSorMore:  -0.2992900545785156
```

  - Employment/Unemployment:

```
Correlation coefficient for PctUnemployed:  0.4749680398078534
Correlation coefficient for PctEmploy:  -0.31226118672258435
```

  - Vacancy:

```
Correlation coefficient for PctHousOccup:  -0.25554595819128334
Correlation coefficient for PctHousOwnOcc:  -0.46069357769159813
Correlation coefficient for PctVacantBoarded:  0.47510410552705856
Correlation coefficient for PctVacMore6Mos:  0.017526764073398652
```

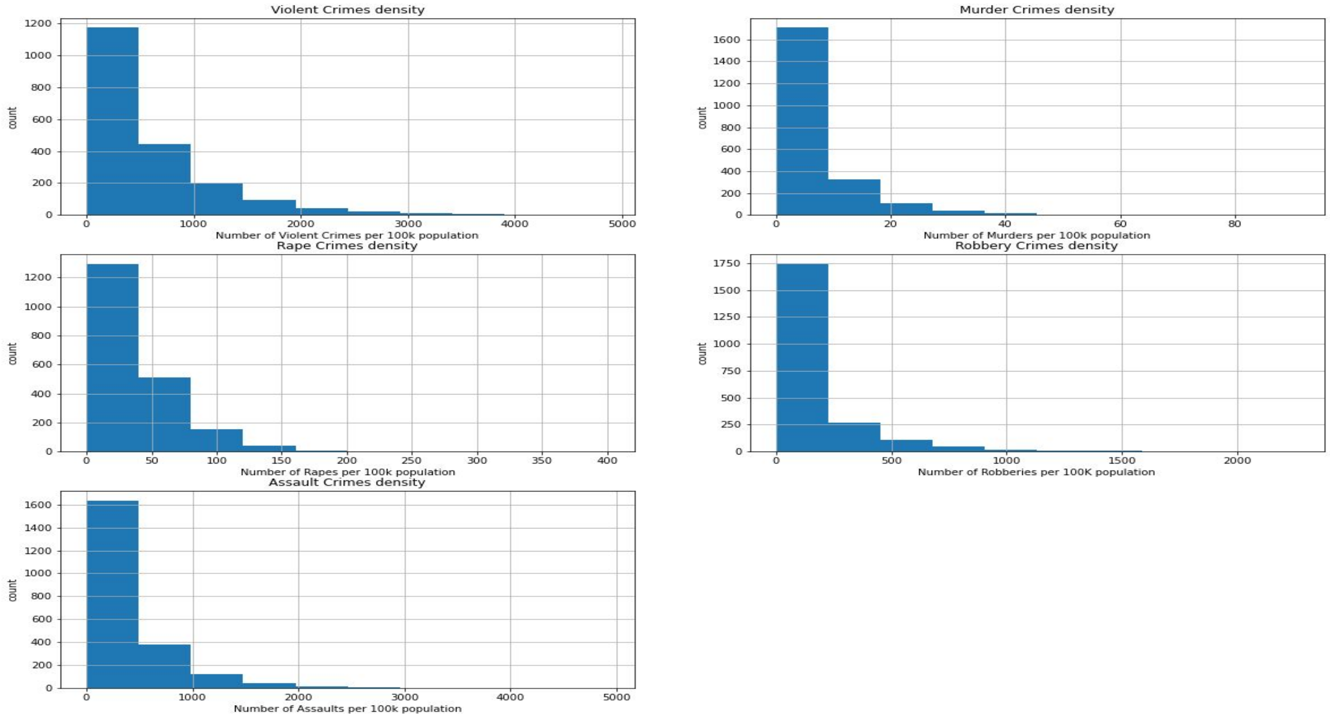
  - Race:

```
Correlation coefficient for pctraceblack:  0.6238334896507505
Correlation coefficient for pctRaceWhite:  -0.676357463352348
Correlation coefficient for pctRaceAsian:  0.03604447688047008
Correlation coefficient for pctRaceHisp:  0.26451715732322045
```

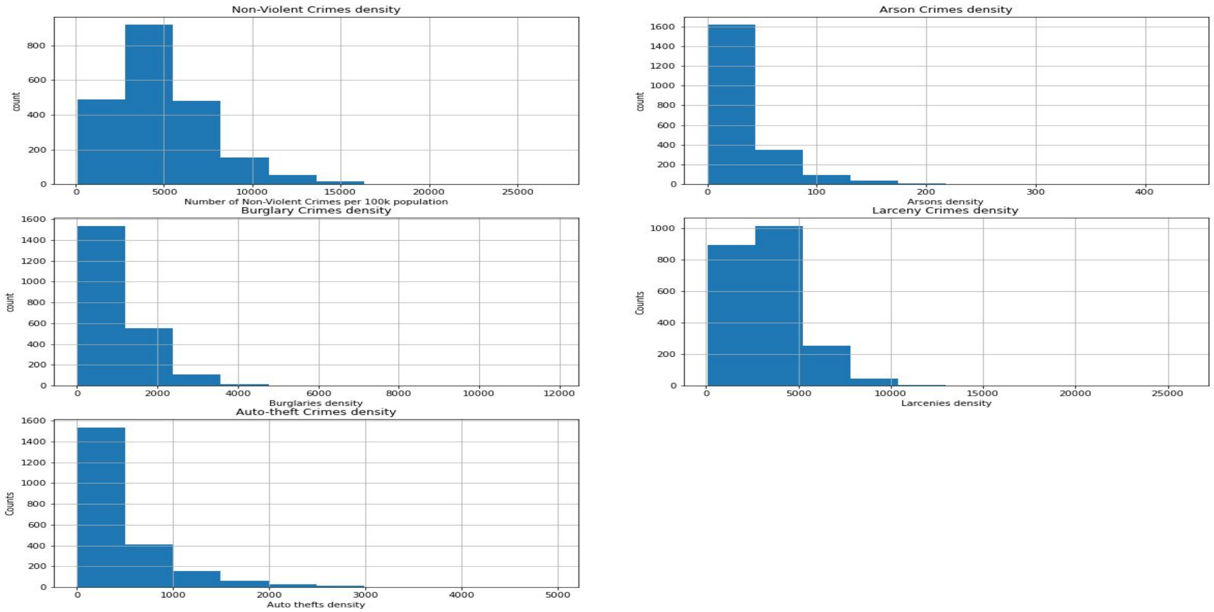
EDA (Continued)

- Step 3: A heatmap was used to provide a concise way of visualizing the correlation coefficient between all variables.

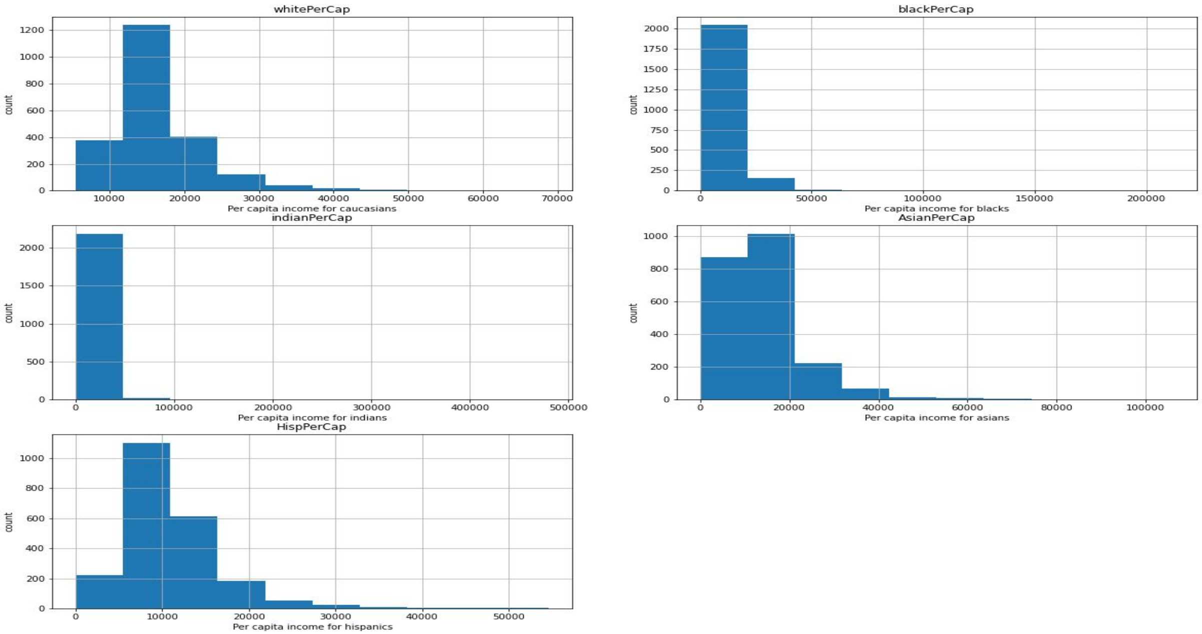
Violent Crimes Density (Bar Graph)



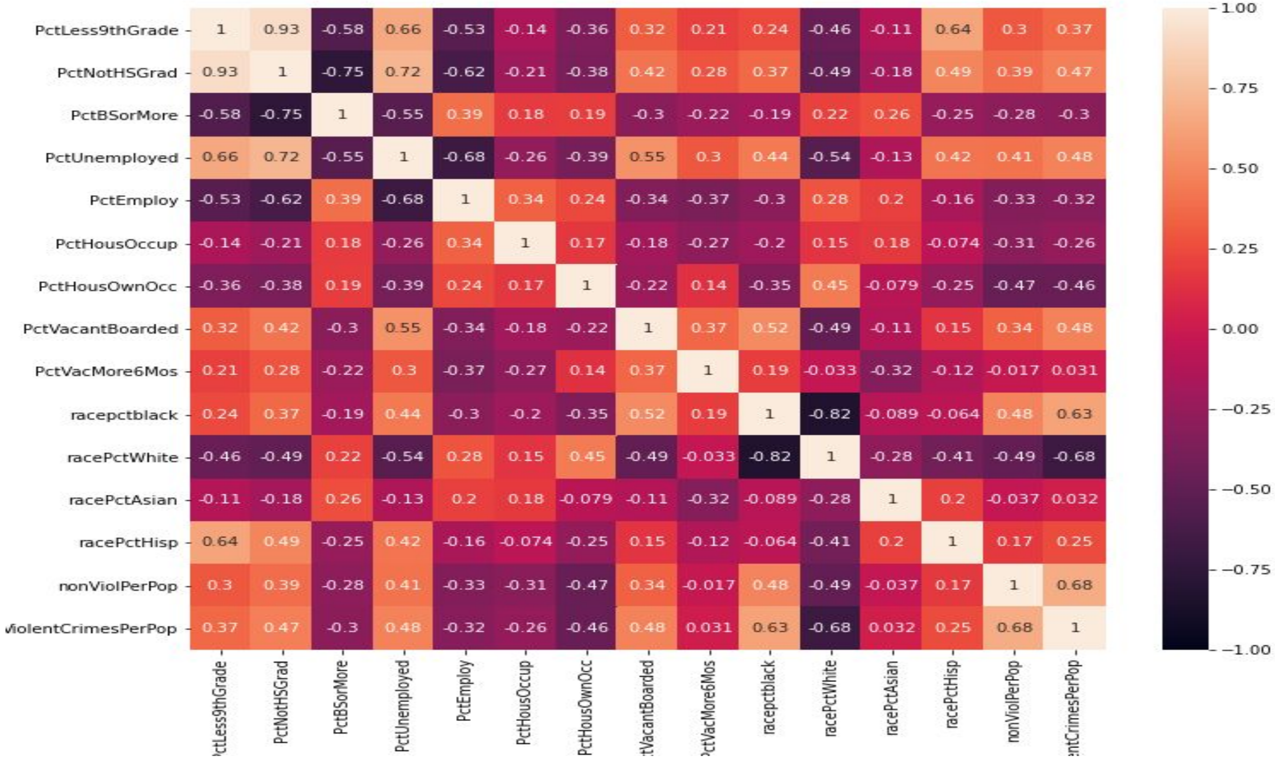
Non Violent Crimes Density (Bar Graph)



Income Density (Bar Graph)



Heatmap



PctLess9thGrade 1 0.93 -0.58 0.66 -0.53 -0.14 -0.36 0.32 0.21 0.24 -0.46 -0.11 0.64 0.3 0.37
PctNotHSGrad 0.93 1 -0.75 0.72 -0.62 -0.21 -0.38 0.42 0.28 0.37 -0.49 -0.18 0.49 0.39 0.47
PctBSorMore -0.58 -0.75 1 -0.55 0.39 0.18 0.19 -0.3 -0.22 -0.19 0.22 0.26 -0.25 -0.28 -0.3
PctUnemployed 0.66 0.72 -0.55 1 -0.68 -0.26 -0.39 0.55 0.3 0.44 -0.54 -0.13 0.42 0.41 0.48
PctEmploy -0.53 -0.62 0.39 -0.68 1 0.34 0.24 -0.34 -0.37 -0.3 0.28 0.2 -0.16 -0.33 -0.32
PctHousOccup -0.14 -0.21 0.18 -0.26 0.34 1 0.17 -0.18 -0.27 -0.2 0.15 0.18 -0.074 -0.31 -0.26
PctHousOwnOcc -0.36 -0.38 0.19 -0.39 0.24 0.17 1 -0.22 0.14 -0.35 0.45 -0.079 -0.25 -0.47 -0.46
PctVacantBoarded 0.32 0.42 -0.3 0.55 -0.34 -0.18 -0.22 1 0.37 0.52 -0.49 -0.11 0.15 0.34 0.48
PctVacMore6Mos 0.21 0.28 -0.22 0.3 -0.37 -0.27 0.14 0.37 1 0.19 -0.033 -0.32 -0.12 -0.017 0.031
racepctblack 0.24 0.37 -0.19 0.44 -0.3 -0.2 -0.35 0.52 0.19 1 -0.82 -0.089 -0.064 0.48 0.63
racePctWhite -0.46 -0.49 0.22 -0.54 0.28 0.15 0.45 -0.49 -0.033 -0.82 1 -0.28 -0.41 -0.49 -0.68
racePctAsian -0.11 -0.18 0.26 -0.13 0.2 0.18 -0.079 -0.11 -0.32 -0.089 -0.28 1 0.2 -0.037 0.032
racePctHisp 0.64 0.49 -0.25 0.42 -0.16 -0.074 -0.25 0.15 -0.12 -0.064 -0.41 0.2 1 0.17 0.25
nonViolPerPop 0.3 0.39 -0.28 0.41 -0.33 -0.31 -0.47 0.34 -0.017 0.48 -0.49 -0.037 0.17 1 0.68
iolentCrimesPerPop 0.37 0.47 -0.3 0.48 -0.32 -0.26 -0.46 0.48 0.031 0.63 -0.68 0.032 0.25 0.68 1
ctLess9thGrade
PctNotHSGrad
PctBSorMore
PctUnemployed
PctEmploy
PctHousOccup
PctHousOwnOcc
tVacantBoarded
ctVacMore6Mos
racepctblack
racePctWhite
racePctAsian
racePctHisp
nonViolPerPop
entCrimesPerPop
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Step 3: Preprocessing

- The Preprocessing step focuses on cleaning the dataset to be used for the Modelling portion of the Capstone. Dummy variables were created for the ‘State’ variable but were not used for the modelling portion. All missing values were dropped rather than filled. A few regression models were tried on the dataset to see whether the models were working or not. The cleaned dataset was uploaded into a new csv file for modelling.

Step 4: Modelling

- The Modelling step purely focuses on Machine learning. For this process, different Regression models were used on the dataset. The models consist of: Linear Regression, Gradient Boosting, Random Forest, Lasso Regression, Ridge Regression, K Nearest Neighbors, and SVM. The models were fit and tested. Each model was tested twice, once for Violent Crimes and Non Violent Crimes. The scores were compared to see which model did the best. The models were also fit with a cross-validation score to avoid overfitting and estimate the skill of the model on the new data.

Modelling (Continued)

- Violent Crimes

	Algorithm	Model accuracy score
0	Linear Regression	0.560155
1	Gradient Boosting	0.442684
2	Random Forest	0.580870
3	Ridge Regression	0.556205
4	Lasso Regression	0.555650
5	KNN	0.500767
6	SVM	-0.035707

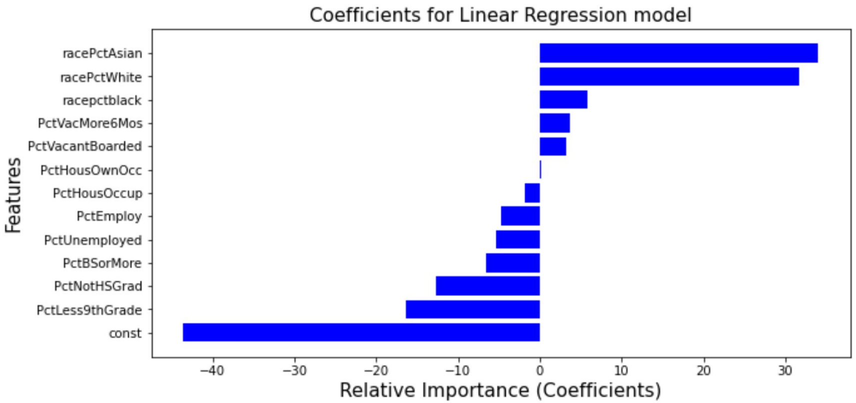
- Non Violent Crimes

	Algorithm	Model accuracy score
0	Linear Regression	0.421480
1	Gradient Boosting	0.218447
2	Random Forest	0.368297
3	Ridge Regression	0.414873
4	Lasso Regression	0.420396
5	KNN	0.273717
6	SVM	-0.023574

Modelling (Violent Crimes)

- After being fit with optimal hyperparameters, the cross validation score and r^2 score for Linear Regression is 0.587 and 0.545, respectively. The feature importance is formatted as a horizontal bar graph below. The feature importance shows that racePctAsian is the best predictor with an importance score of 34.03.

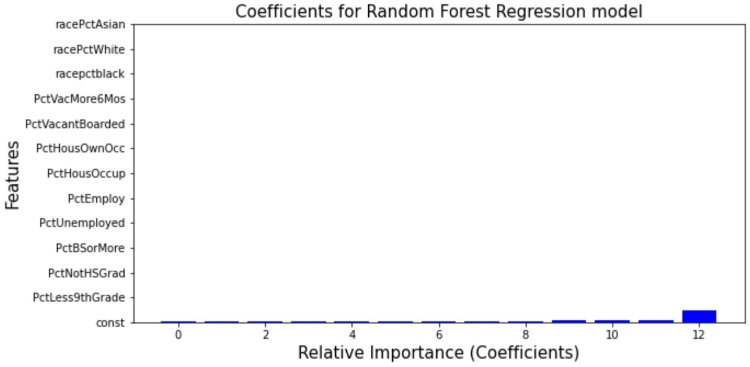
	Features	Importance scores (Coefficients)
0	const	-43.675972
1	PctLess9thGrade	-16.367967
2	PctNotHSGrad	-12.792997
3	PctBSorMore	-6.552172
4	PctUnemployed	-5.380249
5	PctEmploy	-4.751642
6	PctHousOccup	-1.900757
7	PctHousOwnOcc	0.143150
8	PctVacantBoarded	3.143726
9	PctVacMore6Mos	3.661888
10	racepctblack	5.872149
11	racePctWhite	31.751488
12	racePctAsian	34.036442



Modelling (Violent Crimes)

- After being fit with optimal hyperparameters, the cross validation score and r^2 score for Random Forest Regression is 0.582 and 0.539, respectively.The feature importance is formatted as a bar graph below. The feature importance, again shows that racePctAsian is the best predictor with an importance score of 0.49.

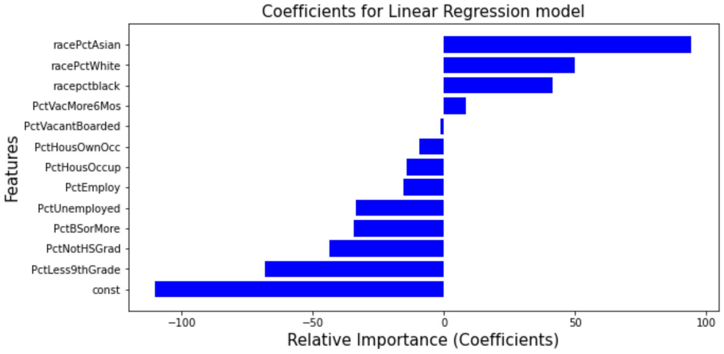
	Features	Importance scores
0	const	0.021564
1	PctLess9thGrade	0.027850
2	PctNotHSGrad	0.031117
3	PctBSorMore	0.031263
4	PctUnemployed	0.031810
5	PctEmploy	0.032788
6	PctHousOccup	0.035632
7	PctHousOwnOcc	0.038366
8	PctVacantBoarded	0.052973
9	PctVacMore6Mos	0.068381
10	racepctblack	0.068527
11	racePctWhite	0.069180
12	racePctAsian	0.490550



Modelling (Non Violent Crimes)

- After being fit with optimal hyperparameters, the cross validation score and r^2 score for Linear Regression is 0.399 and 0.417, respectively. The feature importance is formatted as a horizontal bar graph below. The feature importance shows that racePctAsian is the best predictor with an importance score of 94.39.

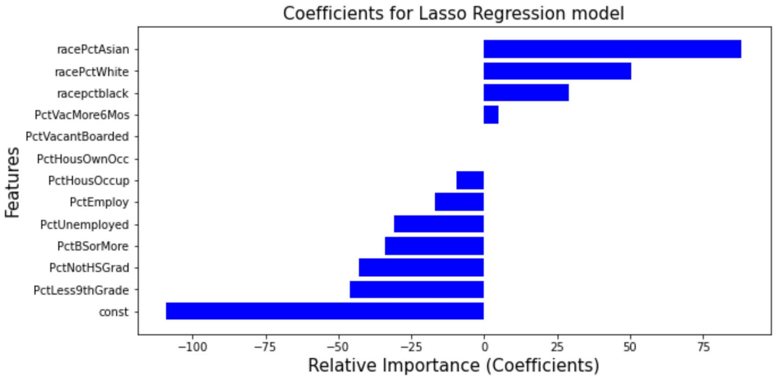
	Features	Importance scores (Coefficients)
0	const	-110.176435
1	PctLess9thGrade	-68.279661
2	PctNotHSGrad	-43.435640
3	PctBSorMore	-34.192700
4	PctUnemployed	-33.402981
5	PctEmploy	-15.338687
6	PctHousOccup	-14.203891
7	PctHousOwnOcc	-9.323151
8	PctVacantBoarded	-1.160332
9	PctVacMore6Mos	8.249100
10	racepctblack	41.629904
11	racePctWhite	50.087039
12	racePctAsian	94.388425



Modelling (Non Violent Crimes)

- After being fit with optimal hyperparameters, the cross validation score and r^2 score for Lasso Regression is 0.402 and 0.416, respectively. The feature importance is formatted as a horizontal bar graph below. The feature importance shows that racePctAsian is the best predictor with an importance score of 88.21.

	Features	Importance scores (Coefficients)
0	const	-109.064763
1	PctLess9thGrade	-45.890463
2	PctNotHSGrad	-42.894850
3	PctBSorMore	-33.929250
4	PctUnemployed	-30.828892
5	PctEmploy	-16.819440
6	PctHousOccup	-9.668281
7	PctHousOwnOcc	-0.000000
8	PctVacantBoarded	0.000000
9	PctVacMore6Mos	4.927079
10	racepctblack	29.151054
11	racePctWhite	50.440241
12	racePctAsian	88.210876



Conclusion

- After fitting the best performing models with optimal hyperparameters and by calculating the feature importance of each model, it seems that the top three predictors are racePctAsian, racePctWhite, and racepctblack.
- The percentage of Asian and White population could be the best predictor of crime rates since when it comes to income distribution, they have higher incomes than other races. This could indicate that non-violent crimes (burglaries, auto thefts, larcenies, etc.) occur at communities/neighborhoods which are predominantly white or asian, since they are more affluent.
- Such results should not be correlated with the current consensus, since this data is from the US Census of 1990. A better approach would be to update the data every year for more accurate results.